Enhanced conformational exploration of protein loops using a global parameterization of the backbone geometry

Timothée O'Donnell *and Frédéric Cazals†

June 21, 2022

Abstract

Flexible loops are paramount to protein functions, with action modes ranging from localized dynamics contributing to the free energy of the system, to large amplitude conformational changes accounting for the repositioning whole SSE or protein domains. However, generating diverse and low energy loops remains a difficult problem.

This work introduces a novel paradigm to sample loop conformations, in the spirit of the Hit-and-Run (HAR) Markov chain Monte Carlo technique. The algorithm uses a decomposition of the loop into tripeptides, and a novel characterization of necessary conditions for Tripeptide Loop Closure to admit solutions. Denoting m the number of tripeptides, the algorithm works in an angular space of dimension $12m$. In this space, the hyper-surfaces associated with the aforementioned necessary conditions are used to run a HAR-like sampling technique.

On classical loop cases up to 15 amino acids, our parameter free method compares favorably to previous work, generating more diverse conformational ensembles. We also report experiments on a 30 amino acids long loop, a size not processed in any previous work.

*Centre Inria at Université Côte d'Azur, France

†Centre Inria at Université Côte d'Azur, France. Correspondence: Frederic.Cazals@inria.fr

1 Introduction

Protein loops. Protein loops are structural components playing various roles in protein function, as illustrated by the following illustrative examples. Enzymes typically involve conformational changes of loops for the substrate (resp. product) to enter (resp. leave) the active site [1]. Membrane transporters implement complex efflux mechanisms resorting to loops changing the relative position of (essentially) rigid domains [2]. In the humoral immune response, the binding affinity of antibodies for antigens is modulated by the dynamics of loops called complementarity determining regions (CDRs) [3]. In G-Protein-Coupled Receptors, extracellular loops binding to ligands trigger signal transduction inside the cell [4].

From the experimental standpoint, these complex phenomena are studied using structure determination methods. However, the structural diversity of loops often results in a low signal to noise ratio, yielding difficulties to report complete polypeptide chains. As a matter of fact, a recent study on structures from the PDB showed that about 83% of structures solved at a resolution of 2.0Å or worse feature missing regions, which for 90% of them are located on loops or unstructured regions [5].

Loop modeling strategies. From the theoretical standpoint, loop mechanisms are best described in the realm of energy landscapes [6], which distinguishes between structure, thermodynamics, and dynamics. In terms of structure, one wishes to characterize active conformations and important intermediates in functional pathways. In assigning occupation probabilities to these states, one treats thermodynamics, while transitions between the states correspond to dynamics. While all atom simulations can naturally be used to explore the conformational variability of loops, their prohibitive cost prompted the development of simplified strategies, which we may ascribed to four tiers.

First, continuous geometric transformations can be used to deform loops, e.g. based on rotations of rigid backbone segments sandwiched between two C_α carbons. Such methods, which include `Crankshaft` [7] and `Backrub` [8, 9], proved effective to reproduce motions observed in crystal structures. However, they are essentially limited to hinge like motions.

Second, a loop may be deformed using loop closure techniques solving an inverse problem which consists in finding the geometric parameters of the loop so that its endpoints obey geometric constraints. Remarkably, various such methods have been developed at the interface of structural biology and robotics [10, 11, 12, 13, 14]. Using loop closure techniques, the seminal concept of *concerned rotations* was introduced long ago to sample loop conformations [15]: first, the prerotation stage changes selected internal degrees of freedom (dof) and brakes loop connectivity; second, the postrotation step restores loop closure using a second set of dof. While early such strategies used solely dihedral angles only [15], more recent ones use a combination of valence and dihedral angles [16, 17]. The latter angles indeed provide a finer control on the amplitude of angular changes in the postrotation stage, and therefore of atomic displacements. A specific type of loop closure playing an essential role is Tripeptide Loop Closure (TLC), where the gap consists of three amino, and loop closure is obtained using the six (ϕ, ψ) angles of the three C_α carbons [18, 19, 13, 20].

Third, considering a loop as a sequence of protein fragments stitched together, high resolution structures from the protein data bank (PDB) can be used to sample its conformations [21, 22]. These methods are greedy/incremental in nature, and the exponential growth of solutions results in a poorer sampling of residues in the middle of the loop. Also, they suffer from the bias inherent to the PDB structures, which favors metastable conformations. As a matter of fact, it has been shown recently using Ramachandran statistics that conformations found in the PDB are less diverse than those yielded by reconstructions in the rigid geometry model [23].

Finally, several classes of methods may be combined. For example, exploiting structural data to bias the choices of angles used to perform loop closure yields a marked improvement in prediction accuracy [24]. More recently, a method growing the two sides of a loop by greedily concatenating (perturbed) tripeptides, before closing the loop using TLC has been proposed [25].

Despite intensive research efforts, predicting large amplitude conformational changes, and/or predicting thermodynamic quantities for long loops, say beyond 12 amino acids, remains a challenge [26, 27]. These difficulties owe to the high dimensionality of loop conformational space, and also to the subtle biophysical constraints that must be obeyed.

Contribution. This work develops a new paradigm to explore the conformational space of flexible protein loops, able to deal with loop length that were out of reach. While our method relies on the tripeptide loop closure, it is, to the best of our knowledge, the first one exploiting a global continuous parameterization of the conformational space on the loop studied. This parameterization is based on the rigidity of peptide bodies (the four atoms $C_\alpha - C - N - C_\alpha$), which is used to define initial conditions for the individual TLC problems and couple them.

Our presentation is organized as follows: Sec. 2 provides a high-level description of the method; Sec. 3 introduces (mandatory) background material; Sec. 4 details the algorithms; Sec. 5 present experiments. Finally, Sec. 6 discusses future work.

Nb: Section S8.1 contains a compendium of the main notations used throughout the paper.

2 Algorithm overview

2.1 Geometric model and ingredients

We consider a loop L consisting of $M = 3 \times m$ amino acids, including one or two a.a. on the boundary of the loop if necessary to obtain a multiple of three. We work in the rigid geometry model [28], in which bond lengths, valence angles, and peptide bond dihedral angle are fixed. In this model, the internal geometry of each tripeptide is defined by 12 angles [18], whence an overall angular configuration space $\mathcal{A}$ of dimension $12m$ for the m tripeptides. (As we shall see later, this model can be relaxed, see Rmk. 8.)

Our algorithm uses a strategy similar to Hit-and-Run (HAR) [29] to sample a region $\mathcal{V} \subset \mathcal{A}$ (Fig. 1). The region $\mathcal{V}$ defines necessary conditions for the m TLC problems to admit solutions. This region is explored by shooting random rays, and intersections between the rays and the hyper-surfaces bounding $\mathcal{V}$ are used to generate configurations of the whole loop. Individual solutions to the m TLC problems are then obtained in a subset $\mathcal{S} \subset \mathcal{V}$. The Cartesian product of solutions for the m tripeptides defines the new conformations of the loop L. We now introduce these ingredients in turn.

Geometric model. The four atoms making up the peptide bond $(C_{\alpha;1}, C_1, N_2, C_{\alpha;2})$ form a rigid body termed the *peptide body* (Fig. S1). For the sake of exposure, we call the two segments $C_{\alpha;1} - C_1$ and $N_2 - C_{\alpha;2}$ the *legs* of the tripeptide, and the tripeptide minus its legs the *tripeptide core*. We model the loop as a sequence of peptide bodies P_k connecting tripeptides cores $T_k^{'}$ (Fig. 2):

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

(Nb: strictly speaking, P_0 and P_m contain each two atoms of the loop L.) The main idea to generate conformations of L is to sample the positions of peptide bodies independently using rigid motions, and then, to solve individual TLC problems. To describe this strategy more precisely, the following ingredients are needed.

Tripeptide loop closure. Tripeptide Loop Closure is a method computing all possible valid geometries of a tripeptide, under two types of constraints. First, the first and last two atoms of the tripeptide, *i.e.* its legs, are fixed. Second, all internal coordinates are fixed, except the six (ϕ, ψ) dihedral angles of the three C_α carbons.

TLC admits at most 16 solutions corresponding to the real roots of a degree 16 polynomial. These solutions have been shown to be geometrically diverse (atoms are moving up to 5Å), and low potential energy [23]. Solving TLC can be done using three rigid bodies associated with the three edges of the triangle involving the three C_α carbons. The rotations of these rigid bodies are described by three angles τ_1, τ_2, τ_3, two of which can be eliminated to yield the degree 16 polynomial. The coefficients of this polynomial depends on $3 \times 4 = 12$ angles describing the internal geometry of the tripeptide [18]. This 12 dimensional space is denoted $\mathcal{A}_k$ for the tripeptide T_k. Taking the Cartesian product of the individual angular spaces of the m tripeptides yields a $12m$ dimensional space denoted $\mathcal{A}$.

Necessary conditions for TLC to admit solutions. In the angular space $\mathcal{A}_k$, we have recently exhibited a region $\mathcal{V}_k$ defining necessary conditions for TLC to admit solutions [30]. For a given tripeptide, this region is defined from 24 implicit equations involving the 12 variables parameterizing TLC. The corresponding space for all tripeptides is denoted $\mathcal{V}$. This space contains the solution space $\mathcal{S} \subset \mathcal{V}$, such that each tripeptide admits solutions.

Identifying active constraints with Hit-and-Run. To sample $\mathcal{S}$, we use the Hit-and-Run (HAR) technique invented long ago to identify redundant hyperplanes in linear programs [29]. In a nutshell, given a starting point inside the polytope, HAR iteratively proceeds as follows: shoot a random ray inside the polytope and identify the nearest hyperplane intersected; generate a random point onto the segment defined by the starting and the intersection point; iterate. Since then, this algorithm has been modified to generate points following a Gaussian distribution, a key step in the computation of the volume of polytopes [31]. Other random walks serving similar purposes are billiard walk and Hamiltonian Monte Carlo [32, 33], as well as walks based on piecewise deterministic processes [34]. In the sequel, we use HAR to sample a high dimensional curved region.

2.2 Algorithm: wrapping up

Unmixed loop sampler. Similarly to HAR, our algorithm consists of consecutive steps, called embedding steps. Each step generates a conformation L' of the loop L by moving the peptide bodies. Given the internal coordinates of L', we solve TLC for each individual tripeptide, and take the Cartesian product of these solutions.

To see how the conformation L' is generated, let $SE(3)$ be the special Euclidean group representing rigid motions (translation+rotation) in 3D. The $m-1$ peptide bodies being rigid bodies, we move them in 3D space using rigid motions parameterized over the motion space $\mathcal{M} = (SE(3))^{m-1}$. We consider a (random) ray in $\mathcal{M}$, whose parameter t is called the *time*. Every point on this ray defines a rigid motion applied to each peptide body. Since the tripeptide legs are moving due to this motion, the 12 angular coordinates of each tripeptide become time dependent. We use the image of the ray in the angle space $\mathcal{A}$ to find intersections with the hyper-surfaces defining the aforementioned necessary conditions (Fig. 1). In a manner similar to HAR, these intersections are used to generate a random point in the validity space $\mathcal{V}$. Each such point encodes an internal geometry for each tripeptide, so that TLC can be solved for each individual tripeptide. The solutions to the individual TLC problems are then combined, retaining one at random or all of them. The ability to generate efficiently points in $\mathcal{S}$ depends on the stringency of necessary conditions defining $\mathcal{V}$, that is to say on the volume of the region $\mathcal{S} \backslash \mathcal{V}$.

Combining these steps yields algorithm $\mathbf{ULS}^{N_V;N_{OR}}_{One|All;N_{ES}}[p_0]$, whose parameters are as follows: $One|All$ 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), and p_0 the starting configuration.

Mixed loop sampler. To alleviate the constraint of fixed peptide bodies throughout the simulation, we also provide a two-step variant of the algorithm $\mathbb{MLS}^{N_V;N_{OR}}_{One|All;N_{ES}}[p_0]$. In short, every other HAR step, the loop is shorted by three residues (two a.a. on a random end, one on the other), and a HAR step is performed for this reduced model. One solution is then picked at random, and the updated positions of the peptide bodies used for the next HAR step.

3 Background and notations for peptides and TLC

3.1 Peptides and tripeptides

Peptides, peptide bonds, tripeptides, and protein loops. Atoms within a tripeptide are denoted as $C_{\alpha;3k-2}, C_{\alpha;3k-1}, C_{\alpha;3k}$, and likewise for the C and N atoms (Fig. 2). Note that with these notations, one has $A_{4k-3} = N_{3k-2}$, $A_{4k-2} = C_{\alpha;3k-2}$, $A_{4k-1} = C_{\alpha;3k}$ and $A_{4k} = C_{3k}$.

As noticed above, the two segments $N_{3k-2}C_{\alpha;3k-2}$ and $C_{\alpha;3k}C_{3k}$ form *legs* of the tripeptide, while the tripeptide minus its legs form the *tripeptide core* $T_k^{'}$. Note that for two consecutive tripeptides, the second leg of T_k and the first one of T_{k+1} form the peptide bond. Note also that in the decomposition of Eq. 1, $P_0 = A_1A_2$ and $P_m = A_{4m-1}A_{4m}$ play a special role: these two fixed segments are called *anchors*.

3.2 Tripeptide loop closure (TLC) with fixed legs

TLC uses constraints on the tripeptide legs and internal coordinates (See Sec. 2). We may also recall that TLC induces a partition of the nine atoms in the tripeptide T_k into two classes. On the one hand, the first two and the last two atoms, *i.e.* the legs, are fixed. On the other hand, the remaining five middle atoms are moving. When considering all solutions of TLC on an exhaustive database of tripeptides extracted from the PDB, these atoms move up to 5Å[23].

Solutions of TLC [18] rely on the following observations (Fig. 3(A,B) [1]):

- TLC involves three rigid bodies: the first two involve the five atoms in-between the first and third C_α carbons; the third one consists of the four atoms defining the legs of the tripeptide.
- The solution space of TLC can be modeled using rotation angles denoted $\{\sigma_{k,i}, \tau_{k,i}\}$ associated to the three rigid bodies. (Nb: the two angles associated with the $C_{\alpha;k,i}$ carbon are $\sigma_{k,i-1}$ and $\tau_{k,i}$.) Positions of the rigid bodies must respect the valence angles θ_i at the three C_α carbons. The rotation of a rigid body about its $C_\alpha - C_\alpha$ axis only impacts the valence angle constraints at its endpoints.
- Searching for solutions to the loop closure is akin to searching for rotation combinations of the angles $\{\sigma_{k,i}, \tau_{k,i}\}$ respectful of θ angles. $\sigma_{k,i-1}$ is the rotation angle of N_i atoms around their corresponding axis. $\tau_{k,i}$ is the rotation angle for C_i around its axis (Fig. 3(A,B)).

The geometry of the backbone can be used to define local frames at each C_α carbon ([18] and Fig. 3(B)), based on three vectors: $\mathbf{\hat{Z}_{k,i}}$ – unit vector along two consecutive C_α carbons, $\mathbf{\hat{r}^\tau_{k,i}}$ – to define the rotation of angle $\tau_{k,i}$, $\mathbf{\hat{r}^\sigma_{k,i}}$ – to define the rotation of angle $\sigma_{k,i}$. Using these local frames, one defines the angles $\alpha_{k,i}, \xi_{k,i}, \eta_{k,i}$, with indices $i = 1, 2, 3$ – counted modulo three, for the tripeptide T_k (Fig. S2):

$$
\begin{cases}
\alpha_{k,i} = \angle \mathbf{\hat{Z}_{k,i}} \mathbf{\hat{Z}_{k,i+2}}; & \alpha_{k,i} \in [0, \pi) \\
\xi_{k,i} = \angle - \mathbf{\hat{Z}_{k,i}} \mathbf{\hat{r}^\sigma_{k,i}}; & \xi_{k,i} \in [0, \pi) \\
\eta_{k,i} = \angle \mathbf{\hat{Z}_{k,i}} \mathbf{\hat{r}^\tau_{k,i}}; & \eta_{k,i} \in [0, \pi) \\
\delta_{k,i} = \angle C_{k,i}C_{\alpha;k,i}, C_{\alpha;k,i}C_{\alpha;k,i+1}, C_{\alpha;k,i+1}N_{k,i+1} & \delta_{k,i} \in [0, 2\pi)
\end{cases}
\tag{2}
$$

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

The corresponding 12-dimensional space is denoted $\mathcal{A}_k$.

[1] When talking of individual tripeptides i is used as an index with $i \in \{1, 2, 3\}$. These indices are counted mod 3, that is $i - 1 = i + 2$.

3.3 Tripeptide and necessary constraints for TLC

From now on, we assume that the peptide of interest is the k-th tripeptide in our loop, see Eq. (1).

In recent work [30], we have introduced necessary conditions for TLC to admit solutions. For each of the three angles $\tau_{k,i}$, these so-called *depth 1 inter-angular constraints* are based on intervals to which $\tau_{k,i}$ must belong. These intervals, which are parameterized by the angular representation of the peptide, are denoted as follows:

$$\begin{cases} \mathcal{I}_{\tau_{k,i}} = \{I_{\tau_{k,i}}\} \text{ with } I_{\tau_{k,i}} = [I_{\tau}^{\min}(\mathbf{A}_{k,i}), I_{\tau}^{\max}(\mathbf{A}_{k,i})] \\ \mathcal{I}_{\tau_{k,i}|\delta} = \{I_{\tau_{k,i}|\delta}\} \text{ with } I_{\tau_{k,i}|\delta} = [I_{\tau|\delta}^{\min}(\mathbf{A}_{k,i+1}), I_{\tau|\delta}^{\max}(\mathbf{A}_{k,i+1})] \end{cases} \tag{3}$$

There are two intervals of each type, and their pairwise intersection results in four so-called *depth one validity intervals* or DOVI. As established in [30], the bounds of these angles depend on the values

$$\arccos \frac{+\cos(\theta_i \pm \xi_{i-1}) + \cos\eta_i \cos\alpha_i}{\sin\eta_i \sin\alpha_i}. \tag{4}$$

For a given tripeptide, we may consider the mapping from its angular representation in the angle space $\mathcal{A}_k$ to the validity intervals:

$$\mathrm{DOVI}_{\tau_{k,i}}(\cdot) : \mathcal{A}_k \mapsto (\mathcal{I}_{\tau_{k,i}} \cap \mathcal{I}_{\tau_{k,i}|\delta})^4. \tag{5}$$

That is, upon fixing the angular representation of the tripeptide (Def. 1), we obtain up to four validity intervals, or the empty set if the four intersections are empty. As reported in the companion paper [30], our necessary conditions are rather tight.

Remark 1 *The function $DOVI_{\tau_{k,i}}$ is obtained using the interval $I_{\tau_{k,i}}$ whose definition requires the angles $\alpha_{k,i}, \eta_{k,i}, \xi_{k,i-1}$ for $I_{\tau_{k,i}}$, and the interval $I_{\tau_{k,i}|\delta}$ whose definition requires the angles $\alpha_{k,i+1}, \eta_{k,i+1}, \xi_{k,i}, \delta_{k,i}$. The number of parameters is thus seven. For the sake of conciseness, we use the supersets $\mathbf{A}_{k,i}$ and $\mathbf{A}_{k,i+1}$. See [30] for details.*

4 Algorithm: details

4.1 Tripeptides with moving legs

Moving peptides bodies. When considering the decomposition of Eq. (1), the $m-1$ peptide bodies move independently. The motion of one peptide body is parameterized by the special Euclidean group $SE(3)$, which combines one translation and one rotation. To be more specific, let S^2 be the sphere of directions in $\mathbb{R}^3$, and A a positive real number. The motion space $\mathcal{R}$ for one peptide body is defined via the motion space

$$\mathcal{R} : (S^2 \times [0, A)) \times (S^2 \times [0, 1/A)) \subset SE(3). \tag{6}$$

The term $S^2 \times [0, A)$ codes the translation defined by a unit vector and a real number in $[0, A)$, while the term $S^2 \times [0, 1/A)$ codes the rotation defined by an angle about a direction given by a unit vector on S^2 and a real number in $[0, 1/A)$. Therefore, specifying a random rigid motion for each peptide body requires $2(m-1)$ unit vectors. We pool these vectors into a $6(m-1)$-dimensional vector denoted V in the sequel. Each rigid body is simultaneously translated along the first unit vector and rotated around the second using a corresponding kinematic function. The value of A defines the speed of translation in such a function relative to the rotation speed, *i.e.* if $A = 0.5$ then $1/A = 2$ and the corresponding rigid body will rotate four times as fast as it translates. We use the default value $A = 1$, as we hardly noticed any incidence for this parameter (data not shown). Summarizing, the overall motion space for peptide bodies is the $6(m-1)$ dimensional space:

$$\mathcal{M} = \mathcal{R}^{m-1}. \tag{7}$$

Remark 2 *By the Mozzi–Chasles' theorem, our rigid motion can be modeled as a screw motion. The corresponding analytical form is used to find intersections with the surfaces defining the TLC necessary constraints.*

Using a 1-parameter family in the motion space. We restrict motions in $\mathcal{M}$ to a a 1-parameter family, performing the following linear interpolation defined by vector V:

$$\text{Ray}(V) = \{\gamma(t) = Id + tV, \text{ with } \gamma(0) = Id\}. \tag{8}$$

The restriction of this one parameter family to each peptide body defines a rigid transformation

$$\gamma_k : [0, 1] \mapsto SE(3), \gamma_k(0) = Id, \tag{9}$$

such that the position of the k-th peptide body $P_k(t)$ at time t satisfies

$$P_k(t) = \gamma_k(t)P_k(0). \tag{10}$$

The full equations for this motion are provided in the supplementary section 8.4.

4.2 Validity domain and overall configuration space $\mathcal{A}$

We now wish to use the depth one validity constraints for the m peptides, whose legs are moving as just explained. To this end, we concatenate the angular representations of the m tripeptides (Def. 1), and define:

Definition. 2 *(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:*

$$\mathcal{A} \overset{Def}{=} \prod_{k=1}^{m} \mathcal{A}_k. \tag{11}$$

Fixing the positions of the peptide bodies in Eq. (1) yields the angular representations of the m tripeptides. We therefore define a mapping from the motion space into the global angular space:

$$f_{\mathcal{M}\to\mathcal{A}} : \mathcal{M} \mapsto \mathcal{A} \tag{12}$$

Having discussed the depth one validity interval for one tripeptide– see Eq. (5), we can finally aggregate such conditions:

Definition. 3 *(Angular validity domain $\mathcal{V}$.) The* angular *validity domain $\mathcal{V}_k$ of the angle $\tau_{k,i}$ of the k-th tripeptide is the subset of $\mathcal{A}_k$ such that $DOVI_{\tau_{k,i}}(\cdot) \neq \emptyset$.*

The angular *validity domain of the loop L is the subset $\mathcal{V} \subset \mathcal{A}$ such that*

$$\forall k = 1, \ldots, m, \forall i = 1, \ldots, 3, \forall a \in \mathcal{V} : DOVI_{\tau_{k,i}}(a) \neq \emptyset.$$

Note that there are $3m$ individual angular validity domains since each tripeptide has 3 angles τ.

Points in $\mathcal{V}$ satisfy necessary conditions. However, for a point $p \in \mathcal{V}$, one or several tripeptide may not admit any valid geometry. We therefore define:

Definition. 4 *(Solution space $\mathcal{S}$) The* solution space *$\mathcal{S} \subset \mathcal{V}$ of the loop L is the subspace of $\mathcal{A}$ such that TLC admits at least one solution for each tripeptide. A point in $\mathcal{S}$ (resp. $\mathcal{V}\backslash\mathcal{S}$) is termed* fertile *(resp. sterile).*

Let s_k the number of solutions yielded by TLC for a point $p \in \mathcal{S}$. The Cartesian product of these sets yields a total number of embeddings, *i.e.* conformations, equal to $\Pi_{k=1,\ldots,m} s_k$.

Remark 3 *Note that the degrees of freedom are defined for rigid bodies in-between tripeptides while the constraints are defined within the tripeptides (Fig. 2).*

4.3 Kinetic validity intervals

We now wish to use our 1-parameter family of motions to explore the solutions space $\mathcal{S}$ via an exploration of the valid space $\mathcal{V}$.

The tripeptide legs move according to the motion imposed to the peptide bodies (Eq. 10). It is therefore possible to define a time dependent (aka kinetic) version of the angles $\mathbf{A}_{k,i}$:

$$\mathbf{A}_{k,i}(t) = (f^{(\alpha)}_{(k,i)}(t), f^{(\xi)}_{(k,i)}(t), f^{(\eta)}_{(k,i)}(t), f^{(\delta)}_{(k,i)}(t)), \tag{13}$$

with

$$\begin{cases} f^{(\alpha)}_{(k,i)}(t) & : \text{ function computing the angle } \alpha_{k,i} \text{ at time } t \\ f^{(\xi)}_{(k,i)}(t) & : \text{ function computing the angle } \xi_{k,i} \text{ at time } t \\ f^{(\eta)}_{(k,i)}(t) & : \text{ function computing the angle } \eta_{k,i} \text{ at time } t \\ f^{(\delta)}_{(k,i)}(t) & : \text{ function computing the angle } \delta_{k,i} \text{ at time } t \end{cases} \tag{14}$$

Once plugged into the intervals of Eq. (3), these functions make it possible to define a kinetic version of the four static validity intervals:

Definition. 5 *(Kinetic validity intervals) The* kinetic validity intervals *for a given angle $\tau_{k,i}$ of a tripeptide T_k are the validity intervals obtained for the time varying angles $\mathbf{A}_{k,i}(t)$:*

$$\begin{cases} I_{\tau_{k,i}}(t) = [I_{\tau}^{min}(\mathbf{A}_{k,i}(t)), I_{\tau}^{max}(\mathbf{A}_{k,i}(t))] \\ I_{\tau_{k,i}|\delta}(t) = [I_{\tau|\delta}^{min}(\mathbf{A}_{k,i+1}(t)), I_{\tau|\delta}^{max}(\mathbf{A}_{k,i+1}(t))] \end{cases} \tag{15}$$

Remark 4 *The time dependent angles are computed as follows (Fig. S2):*

- *The fixed internal coordinates within each tripeptide are sufficient to determine the value of $\eta_{k,1}, \xi_{k,1}, \eta_{k,2}$ and $\xi_{k,2}$. (Note that these are defined in the rigid bodies associated with $C_{\alpha;3k-2}C_{\alpha;3k-1}$ or $C_{\alpha;3k-1}C_{\alpha;3k}$.)*
- *The position of the* legs *are sufficient to define $\eta_{k,3}$ and $\xi_{k,3}$.*
- *The leg positions together with the fixed internal coordinates are sufficient to compute all three $\alpha_{k,i}, i \in \{1,2,3\}$ angles as these angles are defined by the C_α triangle.*

Remark 5 *The motions of consecutive rigid bodies is constrained by the triangle inequality between the three consecutive C_α atoms (Fig. 3). Indeed, these atoms must satisfy the following triangle inequality:*

$$\|C_{\alpha;3k-2}C_{\alpha;3k}\| \leq \|C_{\alpha;3k-2}C_{\alpha;3k-1}\| + \|C_{\alpha;3k-1}C_{\alpha;3k}\|. \tag{16}$$

Note that following the rigidity of peptide bodies, the two right hand side distances are fixed.

4.4 Sampling: one step

Sampling $\mathcal{V}$ with Hit-and-Run. We sample the validity domain $\mathcal{V}$ using the Hit-and-Run algorithm (Fig. 1 and [29]). For a ray $\text{Ray}(V)$ in the motion space (Eq. 8), consider the restriction of this ray to the valid space $\mathcal{V}$, that is

$$\text{Ray}_{\mathcal{V}}(V) = \{\gamma(t) \in \text{Ray}(V) \mid f_{\mathcal{M}\to\mathcal{A}}(\gamma(t)) \in \mathcal{V}\}. \tag{17}$$

The Hit-and-Run algorithm consists of iteratively sampling a new point on $\text{Ray}_{\mathcal{V}}(V)$, so that the restriction of the ray to the valid space $\mathcal{V}$ must be computed.

To see how, consider two kinetic intervals $I_{\tau_{k,i}}(t) \in \mathcal{I}_{\tau_{k,i}}$ and $I_{\tau_{k,i}|\delta}(t) \in \mathcal{I}_{\tau_{k,i}|\delta}$ as specified in Eq. (15). For these intervals, consider the limit conditions (Fig. 4):

$$\begin{cases} I_{\tau}^{\max}(\mathbf{A}_{k,i}(t)) = I_{\tau|\delta}^{\min}(\mathbf{A}_{k,i+1}(t)), \\ \text{ or } I_{\tau}^{\min}(\mathbf{A}_{k,i}(t)) = I_{\tau|\delta}^{\max}(\mathbf{A}_{k,i+1}(t)) \end{cases} \tag{18}$$

For a given $\tau_{k,i}$ angle, there are 8 such conditions, namely two (Eqs. (18)) for each of the depth one validity interval. And since there are three $\tau_{k,i}$ angles per tripeptide, we obtain 24 conditions.

With these ingredients, our algorithm operates as follows:

- Generate a random ray $\text{Ray}(V)$ in the motion space $\mathcal{M}$.
- (`get_tau_tmax`, Algorithm 1 and Sec. S8.4.5) For a given $\tau_{k,i}$ angle, find out the largest interval $[0, t_{\max}]$ such that the $\text{DOVI}_{\tau_{k,i}}$ is different from the $\emptyset$ on this interval (Nb: an upper bound on $t_{\max}$ is obtained from the triangle inequality applied to the C_α carbons, see remark 5.)
- (`LS_one_step`, Algorithm 2) Take the intersection of all such intervals for the $3m$ angles, generate a t value on the resulting interval, and apply the corresponding motions to the tripeptide legs. This yields a candidate conformation $L_{\text{cand.}} \in \mathcal{V}$ of the loop L.
- (`Loop_sampler`, Algorithm 3). Perform `LS_one_step` until $L_{\text{cand.}} \in \mathcal{S}$. Once obtained, start again from $L_{\text{cand.}}$ and iterate.

Remark 6 *In algorithm* `LS_one_step`*, it should be noted that taking the intersection ensures that all conditions hold. But one may have* $DOVI_{\tau_{k,i}} \neq \emptyset$ *on other segments defined by intersection between the ray and the 24 hyper-surfaces. In practice, preliminary tests did not show a significant improvement in tracking such segments.*

Remark 7 *In the* real random access memory model *(real RAM), which assumes exact calculations with real numbers, Algorithm* `LS_one_step` *is exact. In practice, our implementation uses multiprecision numbers and root finding routines provided by Maple [35]. Due to the cost of such operations, algorithm 2 can be further optimized, see algorithm S5.*

Leaving the realm of multiprecision, an approximate version has also been developed to strike a compromise between exactness and performances, see `LS_one_step_approx` *(Algorithm 4). This variant performs a regular sampling of the ray, from which* t_{max} *is estimated.* `LS_one_step_approx` *is the version used in the experiments thereafter.*

4.5 Sampling: combining several steps

We use the building block `Loop_sampler` to define two algorithms. In our Experiments, the loops assessed are those generated by these two algorithms, without any relaxation/energy minimization or post-processing.

Unmixed loop sampler. Combining steps of `Loop_sampler` yields algorithms $\mathbf{ULS}^{N_V;N_{OR}}_{One|All;N_{ES}}[p_0]$, whose parameters are as follows:

1. p_0: the starting point/conformation in space $\mathcal{S}$.
2. One|All: a point in the solution space $\mathcal{S}$ generates a total of $N_m = \Pi_{k=1,\ldots,m} s_k$ loop conformations, with s_k the number of TLC solutions for the tripeptide T_k. The flag One|All states whether we choose one embedding at random, or keep them all.
3. N_{ES}: for a given HAR trajectory, the number of embedding steps performed.
4. N_V: number of HAR trajectories started at p_0, each defined by a random vector defining a ray in the motion space $\mathcal{M}$.
5. N_{OR}: the output rate in the form $1/n$, with n the number of HAR steps performed along a HAR trajectory, before an *embedding step* is performed–as dictated by the flag One|All. An output rate of one means that all embeddings steps are exploited.

For example, $\mathbf{ULS}^{5;1/4}_{One;1000}$ uses five HAR trajectories with an output rate of 1/4, and 1000 embedding steps, each retaining a single embedding. Thus, the number of loop conformations returned is exactly 1250. On the other hand, $\mathbf{ULS}^{1;1}_{All;1000}$ uses a single HAR trajectory of 1000 steps with an output rate of one, retaining all solutions at each step. The number of loop conformations generated is at least 2000, and at most 1000×16^m.

Mixed loop sampler. In the previous version of the algorithm, peptide bodies remain rigid during the whole simulation. The two-step variant of the algorithm $\mathbb{MLS}^{N_V;N_{OR}}_{One|All;N_{ES}}[p_0]$ removes this limitation. Every other HAR step, three residues are removed from the loop endpoints (two a.a. on one end, one on the other, at random), and a HAR step is performed for this reduced model. One solution is then picked at random, and the updated positions of the peptide bodies used for the next HAR step.

Steric clashes and collision checking. A general post-processing strategy in loop generation consists of checking the absence of steric clash. Denoting R_i and R_j the van der Waals radii of two atoms i and j, the usual criterion consists of checking that $d_{ij}/(R_i + R_j) \geq r_{min}$, usually taken in the range $0.5 - 0.7$, see [20, 25]. Upon generating a conformation, we perform this check for all pairs of N, C_α, C atoms in the loop.

Remark 8 *We have recalled above the two types of constraints used by TLC: the legs' positions and internal coordinates. Practically, we use standard values for internal coordinates [18, 23]. These internal coordinates can be changed and sampled in the course of the algorithm, an option not used in our experiments.*

In using these standard coordinates, we assume that all tripeptides of the loop have angular parameters $\mathbf{A}_k \in \mathcal{S}$.

5 Experiments

5.1 Material and methods

Implementation. Our implementation is sketched in Sec. S8.3. Consider a loop together with a valid starting point p_0 – see below. First, the $12(m-1)$ Cartesian coordinates of the peptide bodies are extracted, together with the 12 Cartesian coordinates of the two loop anchors (4 points in total). Then, the steps are iteratively performed as described above for the unmixed and mixed versions of the loop sampler.

We compare our samplers against the state-of-the-art method `MoMA-LS` [25] discussed in Introduction. We note however that the comparison cannot be done on par for three reasons. On the one hand, `MoMA-LS` samples three ω angles in the loop before using tripeptide loop closure. Using a distribution learned from the data may restrict the conformational space explored. On the other hand, `MoMA-LS` also samples the ω angle preceding the first tripeptide of the loop is also sampled (Fig. S7); this degree of freedom induces a rotation of all atoms in the loop, including $C_{\alpha;1}$ which is fixed in our algorithm. Finally, we filter out conformations using backbone atoms only – the precise position of the C_β atoms depends on its chemical environment, while `MoMA-LS` uses a coarse grain model for the C_β. This filtering step is subtle, since removing conformations may reduce the conformational diversity, but may also push the system further away, fostering exploration.

Loops tested. Several loop datasets have been assembled, see e.g. [36, 37, 26, 25]. Note that a loop refers to a set of structures with the same sequence and anchor positions which can be superimposed via a rigid motion. Most of these loops comprise between 12 and 15 amino acids. In the sequel, we focus on three such loops.

• PTPN9-MEG2. The first one is a 12 a.a. long loop found in the in human protein tyrosine phosphatase PTPN9-MEG2 [38, 39], between residue 466 and 477. For this case, four conformations (aka landmarks) have been crystallized: L_0 : 4GE2.pdb/chain A, L_1:2PA5.pdb/chain A, L_2: 4GE6.pdb/chain B, L_3:4ICZ.pdb/chain A. Interestingly, three of these loops form a cluster (lRMSD < 0.1, Table S1), while L_3

is significantly different (lRMSD > 1.5). We choose L_0 as a starting point, since it is furthest away from L_3.

• CCP-W191G. The second loop is a 15 a.a. long loop found in cytochrome C peroxidase (CCP), a water-soluble heme-containing enzyme reducing hydrogen peroxide (H_2O_2) to water. CCP contains three cavities which are hydrophobic (cavity: L99A), slightly polar (cavity: L99A/M102Q), and anionic (cavity: W191G), the latter binding almost exclusively small monocations. Out of the several crystal structures reported [40], one of them features N-methyl-1-phenylmethanamine – N-Methylbenzylamine for short in the W191G cavity. This binding is of interest, as the aforementioned 15 a.a. long loop flips out by nearly 12Å, opening the cavity to the bulk solvent for the entry/exit of the ligand [40].

• CDR-H3-HIV. To illustrate the ability of our method to handle long loops as a whole, we process a 30 a.a. long complementarity-determining region (CDR H3) loop, one of the longest CDR observed in human antibodies [41]. Broadly neutralizing antibodies against the human immunodeficiency virus type of 1 (HIV-1) exhibit two typical features, namely an extensive affinity maturation (accomplished over long periods of time), and an exceptionally long heavy chain CDR.

5.2 Conformational diversity

To assess the conformational diversity of a set of conformations generated, we plot the root mean square fluctuations (RMSF) of the $3m$ heavy atoms $\{N, C_\alpha, C\}$ of the loop backbone, in the form of boxplots. (Recall that the RMSF of a given atom is the stdev of distances between its positions and their center of mass.)

Loop PTPN9-MEG2. We first analyze the RMSF values observed for the loop PTPN9-MEG2 (Fig. 6). A general observation is the bell shape traced by the RMSF median marks, which is expected since the middle of the loop incurs less steric constraints than its endpoints. To compare the methods, the RMSF plots for `MoMA-LS` converge rapidly. A median of $\sim 2-3$Å in the middle of the loop is obtained, with numerous extreme/outlier configurations. Our algorithm needs more steps to stabilize, reaching a stable distribution for 500 conformations. Overall, our methods generate RMSF fluctuations larger than those from `MoMA-LS`, with $\mathbf{ULS}^{1;1}_{One;\cdot}$ and $\mathbb{MLS}^{1;1}_{One;\cdot}$ yielding median RMSF values $\sim 5-6$Å and ~ 8Å respectively near the center of the loop.

Our plots also shed light on the various ingredients of our method. A marked difference is observed between $\mathbf{ULS}^{N_V;N_{OR}}_{\text{One}|\text{All};N_{ES}}$ and $\mathbb{MLS}^{N_V;N_{OR}}_{\text{One}|\text{All};N_{ES}}$. The RMSF plots of the former contain plateaus of length 4 corresponding to the atoms found in rigid peptide bodies. Those of the latter do not, a consequence of the shift shift along the backbone inherent to the removal of three amino acids.

Otherwise, an important point is the stability of our method with respect to the parameter One|All and to the number of vectors N_V. Beyond 500 conformations, little variation is actually observed (Fig. 6 versus Fig. S3 and Fig. S4).

CCP-W191G. The patterns for this slightly longer loop are similar to those observed for the previous one, so that we focus solely on the most striking point. Interestingly, despite the lack of sampling of the ω angle, our algorithms reach a max RMSF circa 7.5Å, while `MoMA-LS` culminates at about 3.7Å(Fig. 7 and Fig. S5).

The ability to generate such diverse ensembles is clearly an advantage over more classical methods such as Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) which fail from sampling conformations as diverse as 12Å[40].

CDR-H3-HIV. Loops beyond 15 a.a. are usually considered to be beyond reach [26, 27]. To illustrate the capabilities of our method, we process a 30 a.a. long loop CDR-H3-HIV (Fig. 8), one of the longest CDR observed in human antibodies [41]. The CDR3 resembles an axe, with a handle and a head (Fig. 8(A)). This CDR represents alone 42% of the surface area exposed by the CDRs [41].

Remarkably, compared to the two loops just discussed, `MoMA-LS` exhibits a much larger diversity (Fig. S6). Naturally, the longer the loop, the larger the benefits of also sampling the ω angle preceding the loop. The RMSF plots for our algorithm show a flattened bell shape curve $\mathbb{MLS}^{10;1/2}_{One;500}$, $\mathbb{MLS}^{10;1/2}_{One;5000}$,$\mathbb{MLS}^{10;1}_{One;500}$ and $\mathbb{MLS}^{10;1}_{One;5000}$, with a maximum RMSF near 12Å.

It has been speculated that the head of this CDR3 can substantially deform, possibly to maneuver into a recessed epitope [41]. Our simulations mitigates this intuition (Fig. 8(B,C,D)). On the one hand, while the *middle* of the head does deform substantially, in particular in the vertical direction, the front and the back appear quite rigid. On the other hand, the stem of the axe exhibits a substantial lateral flexibility. Naturally, these preliminary observations call for further structural analysis in the presence of the antigens.

5.3 Exploration of the conformational landscape

To assess the ability of the algorithm to explore a complex conformational landscape, we focus on loops for which several conformations have been obtained experimentally. Consider a set $\{L_j\}, j = 1, \ldots, J$ of J loop conformations, called *landmarks*. To assess the amount of conformational space explored, we generate conformations, and check the min and max lRMSD distances of these conformations to all landmarks.

Loop PTPN9-MEG2. These distances are of special interest in the context of the 2-cluster structure of the four conformations of PTPN9-MEG2 (Table S1).

Starting from L_0, we first study the ability to move away from the cluster $L_0/L_1/L_2$ (maxlRMSD values for columns L_1 and L_2, Table 1). For a fixed number of conformations (50/500/5000), the lRMSD observed for our algorithms are significantly larger than those obtained with the loops from `MoMA-LS`. Consistent with the analysis of RMSF, the variant $\mathbf{ULS}^{N_V;N_{OR}}_{\text{One}|\text{All};N_{ES}}$ outperforms all contenders.

Also starting from L_0, we next investigate the speed at which we approach the significantly different conformation L_3 (minlRMSD values for column L_3, Table 1). The values reported by our methods are slightly worse than those from `MoMA-LS` (Table 1): best `MoMA-LS`: 0.99Å; best $\mathbf{ULS}^{N_V;N_{OR}}_{\text{One}|\text{All};N_{ES}}$: 1.46Å; best $\mathbb{MLS}^{N_V;N_{OR}}_{\text{One}|\text{All};N_{ES}}$ 1.40Å. However, as noticed above, `MoMA-LS` also samples the ω angle preceding the loop. Inspecting ω values, one obtains: $\omega(L_0)$: $-177°$; $\omega(L_3)$: $-165°$; ω(best from `MoMA-LS`): $-167°$. It is therefore the sampling of this dihedral angle which favors `MoMA-LS`.

5.4 Failure rate, running time and steric clashes

Algorithm 2 fails as soon as one TLC does not admit any solution. This failure probability depends on the number of tripeptides, and naturally depends on the discrepancy between the two spaces $\mathcal{S}_k$ and $\mathcal{V}_k$, that is on the volume of the region $\mathcal{V}_k \backslash \mathcal{S}_k$. In turn, this failure naturally impacts the running time of algorithm `Loop_sampler`.

Calculations were run on a desktop DELL Precision 7920 Tower (Intel Xeon Silver 4214 CPU at 2.20GHz, 64 Go of RAM), under Linux Fedora core 32. Each HAR is processed on a single CPU core. For PTPN9-MEG2, there there is on average 0.69 failure per success when tested on $\mathbf{ULS}^{1;1}_{One;1000}[L_0]$ and 2.92 with $\mathbb{MLS}^{1;1}_{One;1000}[L_0]$.

The average time taken for one step by $\mathbf{ULS}^{1;1}_{One;1000}[L_0]$ is 0.04 seconds, and 0.17 for $\mathbb{MLS}^{1;1}_{One;1000}[L_0]$. The latter algorithm involves more operations than the former, and as just noticed, also incurs a higher failure rate. Whence the increased running time.

For the long loop CDR-H3-HIV, the average failure per success becomes 1.18 for $\mathbf{ULS}^{1;1}_{One;1000}$ and 6.09 for $\mathbb{MLS}^{1;1}_{One;1000}$. The average time per step in $\mathbf{ULS}^{1;1}_{One;1000}$ becomes 0.21 seconds, and 0.98 for $\mathbb{MLS}^{1;1}_{One;1000}[L_0]$.

To assess steric clashes (Sec. 4.5), we compute for a given algorithm and a set of solutions, the fraction of conformations featuring a clash. For PTPN9-MEG2, algorithms $\mathbb{MLS}^{10;1}_{All;1000}$ and $\mathbf{ULS}^{10;1}_{All;1000}$ generate 15.3% and 92.5% of clashes, respectively. For CCP-W191G, algorithm $\mathbb{MLS}^{10;1}_{All;1000}$ generates 0.9% of clashes.

For CDR-H3-HIV and algorithm $\mathbb{MLS}^{10;1}_{All;1000}$, the percentage is 81,7%. As expected, steric clashes increase with the loop length, and are also more frequent when the loop resembles a hairpin.

Remark 9 *Parameter One/All has no impact on failure rate since all solutions are computed in any case.*

6 Outlook

Method. Loops sampling methods raise difficult mathematical problems due to the high dimensionality of the parameter space, and the non linear interaction between the degrees of freedom (dof). Current state-of-the art methods belong to two main classes. The first one consists of methods relying on kinematic loop closure; such methods first perturb selected dof (the prerotation step), and proceed with loop closure (the postrotation step). However, a first difficulty is to balance the amplitude of changes incurred by pre and post dof, to avoid steric clashes during the loop closure step. Another difficulty lies in the non linear nature of the solution space. For systems involving n dof, such methods typically results in a solution space which is a $n-6$ dimensional manifold. Sampling this manifold is usually done via back-projection upon walking the tangent space, which is numerically challenging and imposes rather local changes. A second class of methods of utmost importance exploit structures from the Protein Data Bank, and possibly resort to loop closure too. However, such methods face a combinatorial explosion when the loop length increases. As a matter of fact, modeling as a whole loops beyond 15 amino acids is still considered out of reach.

Our work introduces a new paradigm for this problem, based on a global geometric parameterization of the loop relying on a decomposition into tripeptides. The method lies in the lineage of the Hit-and-Run algorithm, invented long ago to identify redundant constraints in a linear program. Since then, HAR and related techniques have proven essential to sample high dimensional distributions in bounded and unbounded domains, yielding effective polynomial time algorithms of low complexity to compute the volume of polytopes in hundreds of dimensions [32, 31, 33, 34]. The connexion between these algorithms and loop sampling is non trivial, as using HAR to generate loop conformations involves two new ingredients. The first one is a description of the loop sampling problem in a fully dimensional conformational space, as it is the absence of codimension which removes the constraint to follow a curved manifold. We achieve such a description using the intrinsic description of tripeptides. The second one is the design of necessary conditions for the individual tripeptide problems to admit solutions. These conditions can then be used in a manner akin to the hyperplanes of the polytope, to explore the region of interest and generate novel conformations.

Our results improve on those produced by a recent state-of-the-art method. On classical loop examples (12 to 15 a.a.), we show that our solutions enjoy wider RMSF fluctuations. We also show that our method copes easily with a 30 a.a. long loop as a whole, a loop length usually considered beyond reach. Last but not least, it should be stressed that our method is parameter free, as the generation process does not depend on any statistical or biophysical model.

Future work. Computational Structural Biology recently underwent a very significant progress with the advent of deep learning methods for structure prediction [42, 43]. However, such methods generally face difficulties for unstructured and/or highly flexible regions [44]. Also, they do not yield insights on the intrinsic complexity of the problem. In this context, our work opens new perspectives in structural modeling. In terms of structure, we anticipate several straightforward applications. The ability of our sampler to generate very diverse ensembles of conformations should prove key to investigate systems with highly flexible regions, including enzymes, membrane transporters, CDRs, and also intrinsically disordered proteins. The realm of thermodynamics appears more challenging. As discussed in Introduction, methods in the lineage of `Conrot` come with correction factors which, once incorporated into Metropolis-Hastings and Monte Carlo sampling, ensure that the correct distribution (typically canonical) is sampled. Our work primarily focuses on the geometric rather than thermodynamic setting. In fact, current sampling methods of choice are multiphase / adaptive sampling methods, including meta-dynamics, Wang-Landau, etc [45, 46]. A question of critical importance in future work will be to ensure that our exploration methods are suitable to sample NVE and/or NVT ensembles, via the calculation of densities of states (DoS). Along the way, the question of incorporating changes on internal coordinates other than dihedral angles naturally arises–but we note that changing such coordinates solely affects the conditioning of the individual TLC problems. The connexion with polytope volume calculations is a strong hint that this may indeed be the case, and that sampling micro-canonical ensembles may be possible. If so, our paradigm may eventually yield a definitive step towards effective structural and thermodynamic predictions. Meanwhile, our method can still be used in the context of global optimization and energy landscapes, which decouples structure, thermodynamics, and dynamics Upon discovering (deep) local minima, one can sample their basins [6] using classical MC methods.

Algorithm 1 `get_tau_tmax`**.** For a given angle $\tau_{k,i}$, find the largest value of $t_{\max}$ of t such that $\text{DOVI}_{\tau_{k,i}}(p(t)) \neq \emptyset$ on the segment $[0, t^{\Delta}_{max}]$.

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**
    $S = S \cup$ numerical solutions for Eqs. 18 $t \in [0, t^{\Delta}_{max}]$
Sort S by ascending order
Let t_l be the l-th element of S
$l = 1$
$u_l := \frac{t_l + t_{l+1}}{2}$
// Stop when no validity interval can be defined for $\tau_{k,i}$
while $\text{DOVI}_{\tau_{k,i}}(f_{\mathcal{M}\to\mathcal{A}}\gamma(u_l)) \neq \emptyset$ **do**
  $t_{max} = u_l$
  $l = l + 1$
return $\{t_{max}\}$

Algorithm 2 `LS_one_step`. Given a starting point $p_0 \in \mathcal{S}$ and a random direction V in the motion space $\mathcal{M}$, the algorithm finds the nearest intersection p_{near} of the image of the ray $\text{Ray}(V)$ (by the map $f_{\mathcal{M}\to\mathcal{A}}$) with a surface constraint, and generates a random value on the segment $[0, t_{\max}]$. Then, applies the corresponding motion to peptide bodies of the loop L.

Input: $p_0 \in \mathcal{S}$: starting point in the fertile space
Input: V: direction in motion space
Output: a point $p_{out} \in \mathcal{V}$
Var $t^{\Delta}_{\max}$: initialized using the smallest value of $t > 0$ breaking triangular inequality in a given tripeptide
V: Random direction (Eq. 8)
$S = \{t^{\Delta}_{\max}\}$
for k $\in \{1, ..., m\}$ **do**
  for i $\in \{1, 2, 3\}$ **do**
    $S = S \cup \text{get\_tau\_tmax}(\tau_{k,i})$
// Get the smallest value – most stringent condition
$t_{\max} = \min S$
// Output the next sample
$t_s \leftarrow \text{Uniform}(0, t_{\max})$
Apply the rigid transforms defined by t_s to the $m - 1$ peptide bodies
return Loop L with moved peptide bodies

Algorithm 3 `Loop_sampler`**.** Given a starting point $p_0 \in \mathcal{S}$, algorithm `Loop_sampler` iterates `LS_one_step` until $L_{\text{cand.}}$ yields solution(s) for all tripeptides in the loop. This process is then repeated iteratively from $L_{\text{cand.}}$.

Input: $p_0 \in \mathcal{V}$
$p_{tmp} = p_0$
$Sample = \emptyset$
while not done **do**
    $no\_clash = false$
    while not no_clash **do**
        $is\_in\_S = false$
        while not is_in_S **do**
            Generate random direction V
            $L_{\text{cand.}} \leftarrow$ LS_one_step(p_{tmp}, V)
            Solve individual TLC for the m peptide bodies
            if all m tripeptide have at least one solution **then**
                $is\_in\_S = true$
        if $\exists$ at least one conformation with no steric clash **then**
            $no\_clash = True$
    $p_{tmp} = L_{\text{cand.}}$
    if Stop condition met **then**
        done=true
    Combine the individual solutions obtained for the individual tripeptides

Algorithm 4 `LS_one_step_approx`. Given a starting point $p_0 \in \mathcal{S}$, a random direction V in the motion space $\mathcal{M}$, and a number of iteration X, the algorithm uniformly samples between 0 and $t_{\max}$, and finds the largest value u_l such that $\mathrm{DOVI}_{\tau_{k,i}}(f_{\mathcal{M}\to\mathcal{A}}(\gamma(u_l))) \neq \emptyset$. It then iterates between the step were it stopped and the one before it until $\mathrm{DOVI}_{\tau_{k,i}}(f_{\mathcal{M}\to\mathcal{A}}\gamma(u_l)) \neq \emptyset$. If $X \to \infty$ the $t_{\max}$ obtained using this algorithm corresponds to the one obtained using `LS_one_step`.

Input: $p_0 \in \mathcal{S}$: starting point in the fertile space
Input: V: direction in motion space
Input: X: max number of iteration to obtain approximate solution
Output: a point $p_{out} \in \mathcal{V}$
Var $t^{\Delta}_{\max}$: initialized using the smallest value of $t > 0$ breaking triangular inequality in a given tripeptide
V: Random direction (Eq. 8)
$u_l := 0$
$x = 1$
// Identify the first iteration failing the condition
in_validity_space=true
while in_validity_space **do**
  $u_l = (x/X)t_{\max}$
  for $k \in \{1..m\}$ **do**
    for $i \in \{1, 2, 3\}$ **do**
      if $\mathrm{DOVI}_{\tau_{k,i}}(f_{\mathcal{M}\to\mathcal{A}}\gamma(u_l)) = \emptyset$ **then**
        in_validity_space=false
  $x = x + 1$
// Slice the failing interval into X bits and iterate
$t_{min} = (x-1)/Xt_{\max}$
$x = 1$
in_validity_space=true
while in_validity_space **do**
  $u_l = t_{min} + \frac{t_{\max}(x)}{X^2}$
  for $k \in \{1..m\}$ **do**
    for $i \in \{1, 2, 3\}$ **do**
      if $\mathrm{DOVI}_{\tau_{k,i}}(f_{\mathcal{M}\to\mathcal{A}}\gamma(u_l)) = \emptyset$ **then**
        in_validity_space=false
  $x = x + 1$
$t_{\max} = t_{min} + \frac{t_{\max}(x-1)}{X^2}$
// Output the next sample
$t_s \leftarrow \mathrm{Uniform}(0, t_{\max})$
Apply the rigid transforms defined by t_s to the $m-1$ peptide bodies
return Loop L with moved peptide bodies

7 Artwork

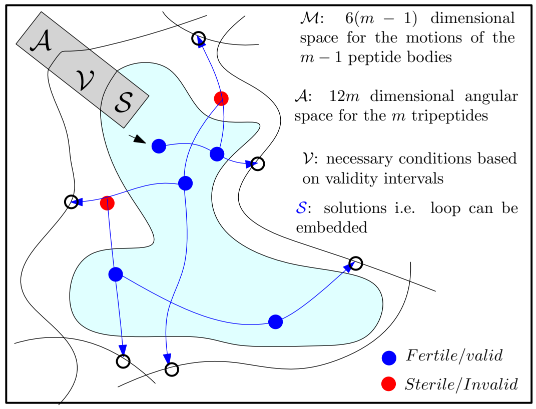


Figure 1: **Sampling a loop involving m tripeptides: algorithm overview.** Spaces used: $\mathcal{A}$: a $12m$ dimensional angular space coding the internal geometry of all tripeptides; $\mathcal{V} \subset \mathcal{A}$: a region characterized by necessary conditions for the m individual TLC problems to admit solutions; $\mathcal{S} \subset \mathcal{V}$ corresponds to individual geometries of the tripeptides such that TLC admits solutions for each tripeptide. The Hit-and-Run algorithm is started at the point indicated by an arrow. It is used to find intersection (empty bullets) between 1D trajectories (blue curves) in the angular space of the tripeptides, and hyper-surfaces bounding the regions defining necessary conditions for the m individual TLC problems to admit solutions. One point is then generated on the curve segment joining the staring point and the intersection point. This point is fertile if all TLC problems admit solutions, and sterile otherwise. The number of conformations obtained is the product of the individual numbers for the m tripeptides.

Acknowledgments. Jean-Pierre Merlet is acknowledged for providing a root finder based on ALIAS. Théo Roncalli and Viraj Agashe are acknowledged for carefully re-reading the paper.

References

[1] M Merced Malabanan, Tina L Amyes, and John P Richard. A role for flexible loops in enzyme catalysis. *Current opinion in structural biology*, 20(6):702–710, 2010.

[2] M. Simsir, I. Broutin, I. Mus-Veteau, and F. Cazals. Studying dynamics without explicit dynamics: a structure-based study of the export mechanism by AcrB. *Proteins: structure, function, and bioinformatics*, 89:259–275, 2021.

[3] A. Schmidt, H. Xu, A. Khan, T. O'Donnell, S. Khurana, L. King, J. Manischewitz, H. Golding, P. Suphaphiphat, A. Carfi, E. Settembre, P. Dormitzer, T. Kepler, R. Zhang, A. Moody, B. Haynes, H-X. Liao, D. Shaw, and S. Harrison. Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *PNAS*, 110(1):264–269, 2013.

[4] Daniel Hilger, Matthieu Masureel, and Brian K Kobilka. Structure and dynamics of gpcr signaling complexes. *Nature structural & molecular biology*, 25(1):4–12, 2018.

[5] Kristina Djinovic-Carugo and Oliviero Carugo. Missing strings of residues in protein crystal structures. *Intrinsically disordered proteins*, 3(1):e1095697, 2015.

[6] D. J. Wales. *Energy Landscapes*. Cambridge University Press, 2003.

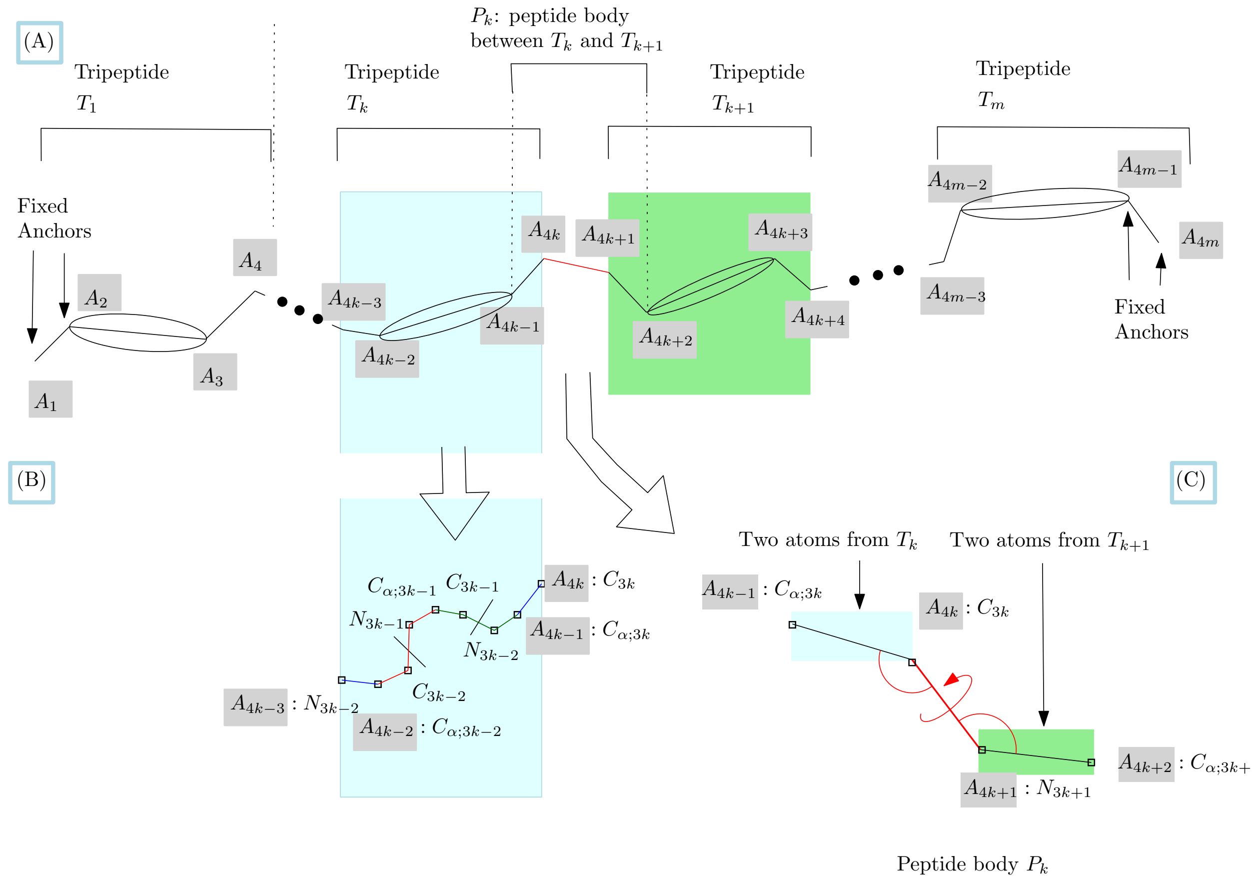


Figure 2: **Loop decomposition into tripeptides and peptide bodies, and associated geometric model.** **(A)** Each ellipsis and its two legs correspond to one tripeptides. In red, the peptide bond between the consecutive tripeptides T_k and T_{k+1}. The peptide body encompasses the peptide bond, as well as one atom to the left and the right. **(B)** Indexing of atoms within the k-th tripeptide. **(C)** Geometry of the peptide bond linking tripeptides T_k and T_{k+1} with constrained bond lengths, valence angles, and torsion angle – in red. These four atoms form the rigid body P_k.

[7] Marcos R Betancourt. Efficient monte carlo trial moves for polypeptide simulations. *The Journal of chemical physics*, 123(17):174905, 2005.

[8] Ian W Davis, W Bryan Arendall III, David C Richardson, and Jane S Richardson. The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure*, 14(2):265–274, 2006.

[9] Colin A Smith and Tanja Kortemme. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of molecular biology*, 380(4):742–756, 2008.

[10] Nobuhiro Go and Harold A Scheraga. Ring closure and local conformational deformations of chain molecules. *Macromolecules*, 3(2):178–187, 1970.

[11] Ioannis Z Emiris and Bernard Mourrain. Computer algebra methods for studying and computing molecular conformations. *Algorithmica*, 25(2):372–402, 1999.

[12] J. Cortés, T. Siméon, M. Remaud-Siméon, and V. Tran. Geometric algorithms for the conformational analysis of long protein loops. *Journal of computational chemistry*, 25(7):956–967, 2004.

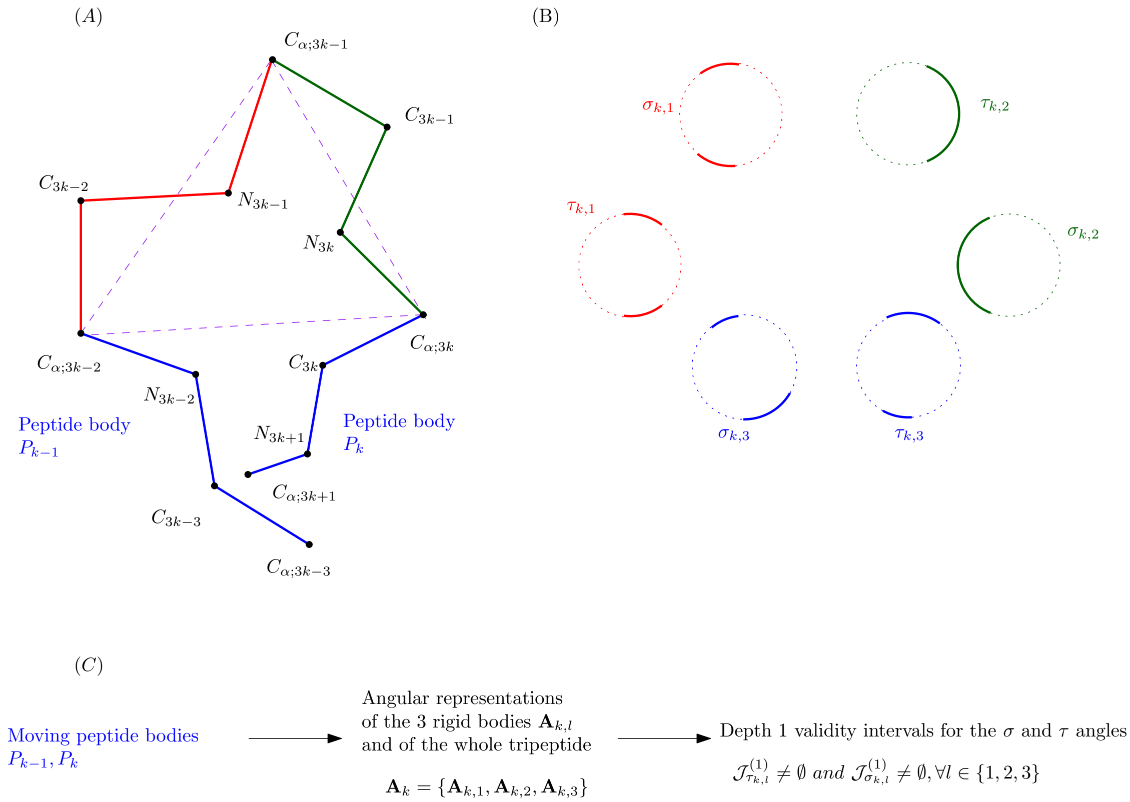


Figure 3: **Geometric model used for an individual tripeptide. (A)** Tripeptide with moving legs. Given internal coordinates and two rigid bodies around a tripeptide the C_α triangle can be defined together with $\{\alpha_i, \eta_i, \xi_i\}$ angles. **(B)** $\mathcal{J}^{(1)}_{\sigma_i}$ and $\mathcal{J}^{(1)}_{\tau_i}$. **(C)** Illustration of the relationship between rigid body positions, $\{\alpha_i, \eta_i, \xi_i\}$ angles and the *depth one inter-angular constraint*.

[13] Kimberly Noonan, David O'Brien, and Jack Snoeyink. Probik: Protein backbone motion by inverse kinematics. *The International Journal of Robotics Research*, 24(11):971–982, 2005.

[14] J. Porta, L. Ros, F. Thomas, F. Corcho, J. Cantó, and J. Pérez. Complete maps of molecular-loop conformational spaces. *Journal of computational chemistry*, 28(13):2170–2189, 2007.

[15] LR Dodd, TD Boone, and DN Theodorou. A concerted rotation algorithm for atomistic monte carlo simulation of polymer melts and glasses. *Molecular Physics*, 78(4):961–996, 1993.

[16] Jakob P Ulmschneider and William L Jorgensen. Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a gaussian bias. *The Journal of chemical physics*, 118(9):4261–4271, 2003.

[17] Sandro Bottaro, Wouter Boomsma, Kristoffer E. Johansson, Christian Andreetta, Thomas Hamelryck, and Jesper Ferkinghoff-Borg. Subtle Monte Carlo updates in dense molecular systems. *Journal of chemical theory and computation*, 8(2):695–702, 2012.

[18] Evangelos A Coutsias, Chaok Seok, Matthew P Jacobson, and Ken A Dill. A kinematic view of loop closure. *Journal of computational chemistry*, 25(4):510–528, 2004.

[19] Juan Cortés and Thierry Siméon. Sampling-based motion planning under kinematic loop-closure constraints. In *Algorithmic Foundations of Robotics VI*, pages 75–90. Springer, 2004.

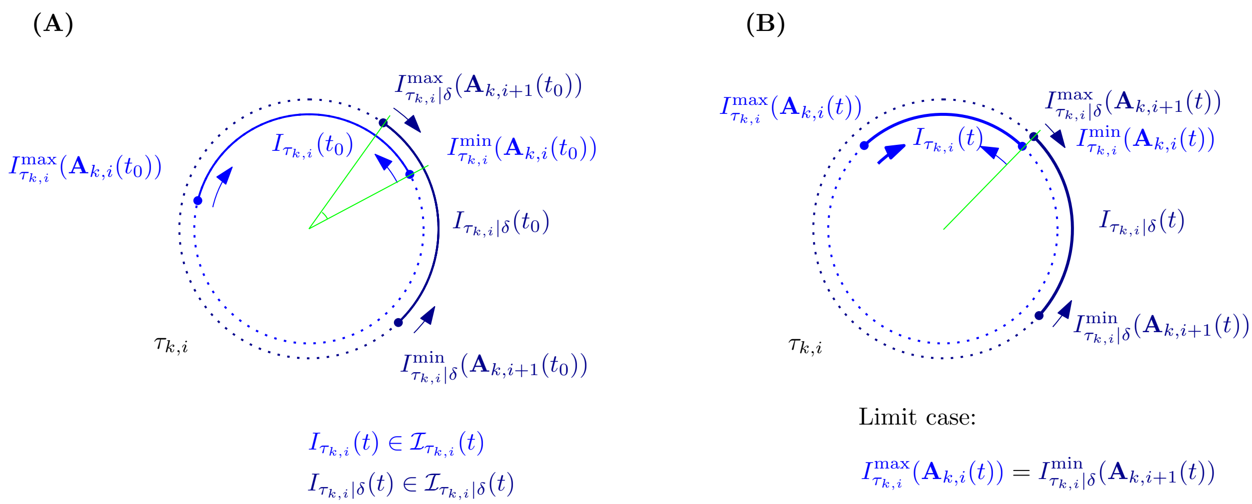


Figure 4: **Kinetic validity intervals.** We focus on a given interval pair $I_{\tau_{k,i}} \in \mathcal{I}_{\tau_{k,i}}$ and $I_{\tau_{k,i}|\delta} \in \mathcal{I}_{\tau_{k,i}|\delta}$ for the angle $\tau_{k,i}$ from tripeptide T_k. The legs of T_k are moving with P_{k-1} and P_k. These movements impact the positions of the interval endpoints via the angles $\mathbf{A}_{k,i}(t)$ and $\mathbf{A}_{k,i+1}(t)$. **(A)** The interiors of the two intervals intersect. **(B)** The intervals intersect on their boundary–a limit case. The arrow indicate the derivative of the endpoints of intervals with respect to time.

[20] D.J. Mandell, E.A. Coutsias, and T. Kortemme. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature Methods*, 6:551 – 552, 2009.

[21] T Alwyn Jones and Soren Thirup. Using known substructures in protein model building and crystallography. *The EMBO journal*, 5(4):819–822, 1986.

[22] R. Kolodny, L. Guibas, M. Levitt, and P. Koehl. Inverse kinematics in biology: The protein loop closure problem. *The International Journal of Robotics Research*, 24(2-3):151–163, 2005.

[23] 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.

[24] Amelie Stein and Tanja Kortemme. Improvements to robotics-inspired conformational sampling in rosetta. *PloS one*, 8(5):e63090, 2013.

[25] A. Barozet, K. Molloy, M. Vaisset, T. Simeon, and J. Cortés. A reinforcement-learning-based approach to enhance exhaustive protein loop sampling. *Bioinformatics*, 36(4):1099–1106, 2019.

[26] C. Marks, J. Shi, and C. Deane. Predicting loop conformational ensembles. *Bioinformatics*, 34(6):949–956, 2018.

[27] Amélie Barozet, Pablo Chacón, and Juan Cortés. Current approaches to flexible loop modeling. *Current Research in Structural Biology*, 3:187–191, 2021.

[28] Richard A Engh and Robert Huber. Accurate bond and angle parameters for x-ray protein structure refinement. *Acta Crystallographica Section A: Foundations of Crystallography*, 47(4):392–400, 1991.

[29] H. Berbee, C. Boender, A. Ran, C. Scheffer, R. Smith, and J. Telgen. Hit-and-run algorithms for the identification of nonredundant linear inequalities. *Mathematical Programming*, 37(2):184–207, 1987.

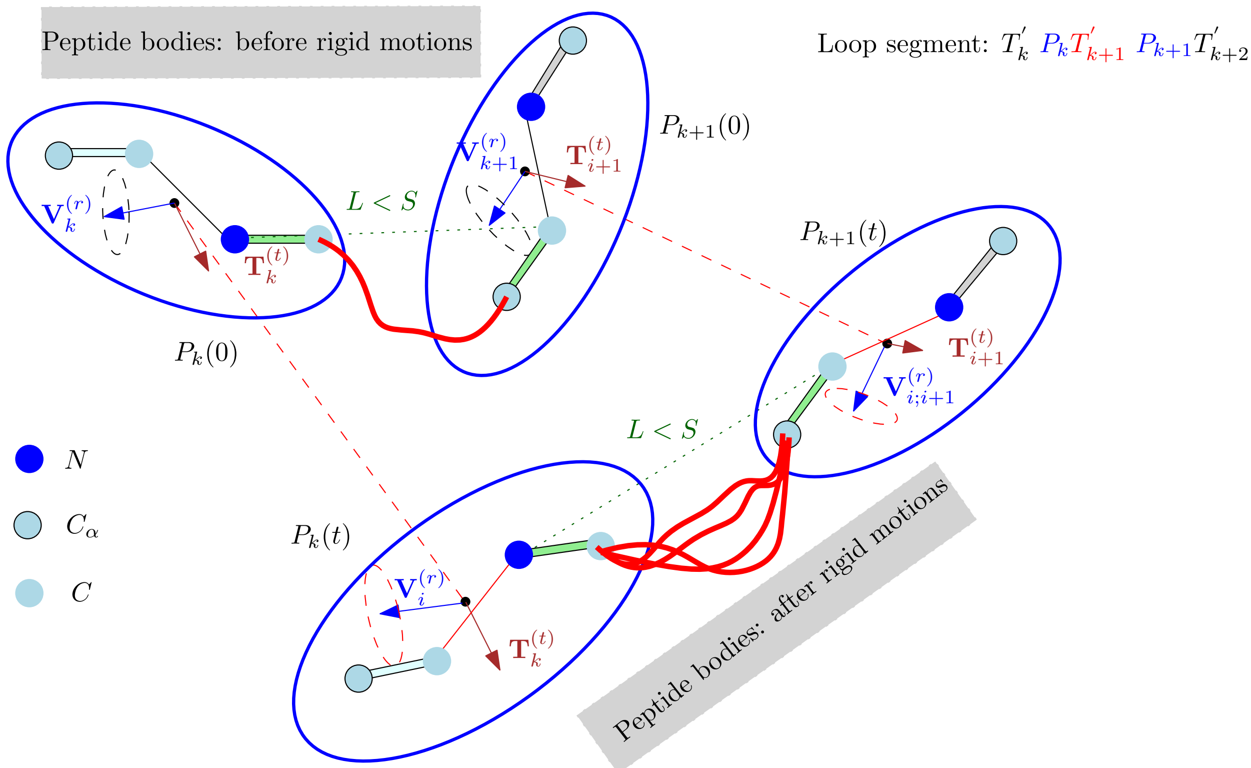


Figure 5: **Interpolation in the space of rigid motions $\mathcal{R}$ and associated transformations applied to rigid bodies.** The figure features two peptide bodies P_k and P_{k+1} in the loop segment $T_k^{'}\ P_k T_{k+1}^{'}\ P_{k+1} T_{k+2}^{'}$. The initial positions of the bodies are denoted $P_k(0)$ and $P_{k+1}(0)$ respectively; these bodies must satisfy a distance constraint materialized by the green line segment – *length* $< S$. Each rigid body undergoes a translation (unit vectors $\mathbf{T}_k^{(t)}$ and $\mathbf{T}_{k+1}^{(t)}$ respectively) composed with a rotation (unit vectors $\mathbf{V}_k^{(r)}$ and $\mathbf{V}_{k+1}^{(r)}$ respectively). The positions corresponding to time t are denoted $P_k(t)$ and $P_{k+1}(t)$ respectively. The distance between the last C_α of $P_k(t)$ and the first C_α of $P_{k+1}(t)$ is constrained by the triangular inequality (SI Sec. 8.4.5). This constraint is represented by the maximum length S on the figure.

[30] T. O'Donnell and F. Cazals. Geometric constraints within tripeptides and the existence of tripeptide reconstructions. Technical report, 2022.

[31] B. Cousins and S. Vempala. A practical volume algorithm. *Mathematical Programming Computation*, 8(2):133–160, 2016.

[32] Y.T. Lee and S. Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *STOC*, pages 1115–1121. ACM, 2018.

[33] 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.

[34] 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.

[35] Michael B Monagan, Keith O Geddes, K Michael Heal, George Labahn, SM Vorkoetter, James McCarron, and Paul DeMarco. Maple 9: Advanced programming guide. 2003.

[36] Matthew P Jacobson, David L Pincus, Chaya S Rapp, Tyler JF Day, Barry Honig, David E Shaw, and Richard A Friesner. A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 55(2):351–367, 2004.

[37] Suwen Zhao, Kai Zhu, Jianing Li, and Richard A Friesner. Progress in super long loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 79(10):2920–2935, 2011.

[38] Ying Qi, Runxiang Zhao, Hongxi Cao, Xingwei Sui, Sanford B Krantz, and Z Joe Zhao. Purification and characterization of protein tyrosine phosphatase ptp-meg2. *Journal of cellular biochemistry*, 86(1):79–89, 2002.

[39] Sheng Zhang, Sijiu Liu, Rongya Tao, Dan Wei, Lan Chen, Weihua Shen, Zhi-Hong Yu, Lina Wang, David R Jones, Xiaocheng C Dong, et al. A highly selective and potent ptp-meg2 inhibitor with therapeutic potential for type 2 diabetes. *Journal of the American Chemical Society*, 134(43):18116–18124, 2012.

[40] Alan P Graves, Devleena M Shivakumar, Sarah E Boyce, Matthew P Jacobson, David A Case, and Brian K Shoichet. Rescoring docking hit lists for model cavity sites: predictions and experimental testing. *Journal of molecular biology*, 377(3):914–934, 2008.

[41] Marie Pancera, Jason S McLellan, Xueling Wu, Jiang Zhu, Anita Changela, Stephen D Schmidt, Yongping Yang, Tongqing Zhou, Sanjay Phogat, John R Mascola, et al. Crystal structure of PG16 and chimeric dissection with somatically related PG9: structure-function analysis of two quaternary-specific antibodies that effectively neutralize HIV-1. *Journal of virology*, 84(16):8098–8110, 2010.

[42] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[43] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[44] Kiersten M Ruff and Rohit V Pappu. Alphafold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, 433(20):167208, 2021.

[45] T. Lelièvre, G. Stoltz, and M. Rousset. *Free energy computations: A mathematical perspective.* World Scientific, 2010.

[46] Rafael C Bernardi, Marcelo CR Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):872–877, 2015.

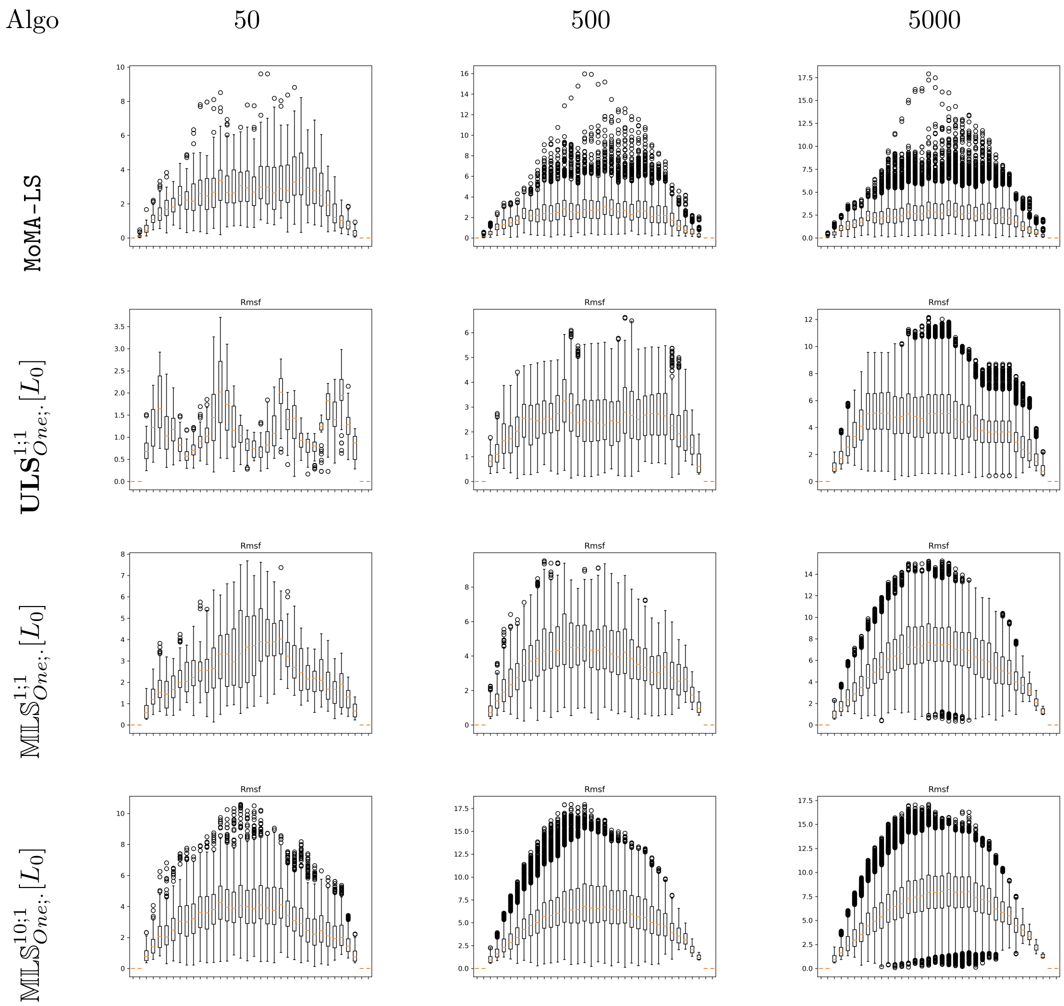


Figure 6: **Loop PTPN9-MEG2: Backbone RMSF for the 12 amino acid long loop PTPN9-MEG2.** Simulations started from the conformation/landmark L_o – see text. Each tick on the x-axis corresponds to a heavy atom of the loop – 36 in this case. For `MoMA-LS`, note that only one atom is fixed on the left hand side of the loop, since the ω angle preceding the loop is also sampled.

	L_1 min/maxlRMSD	L_2 min/maxlRMSD	L_3 min/maxlRMSD
`MoMA-LS`, 50	$1.00/3.81$	$1.03/3.80$	$1.38/4.11$
`MoMA-LS`, 500	$0.78/4.29$	$0.77/4.30$	$1.11/4.70$
`MoMA-LS`, 5000	$0.74/4.92$	$0.73/4.94$	$0.99/4.97$
$\mathbf{ULS}^{1;1}_{One;50}[L_0]$	$0.41/2.21$	$0.42/2.23$	$1.43/2.60$
$\mathbf{ULS}^{10;1}_{One;50}[L_0]$	$0.39/3.09$	$0.38/3.09$	$1.39/3.50$
$\mathbf{ULS}^{1;1/4}_{One;50}[L_0]$	$0.59/3.04$	$0.59/3.05$	$1.34/3.45$
$\mathbf{ULS}^{10;1/4}_{One;50}[L_0]$	$0.46/3.53$	$0.47/3.56$	$1.34/3.73$
$\mathbb{MLS}^{1;1}_{One;50}[L_0]$	$0.46/3.99$	$0.47/4.01$	$1.59/4.56$
$\mathbb{MLS}^{10;1}_{One;50}[L_0]$	$0.43/4.02$	$0.43/4.03$	$1.53/4.75$
$\mathbb{MLS}^{1;1/4}_{One;50}[L_0]$	$1.80/5.05$	$1.81/5.07$	$2.20/5.38$
$\mathbb{MLS}^{10;1/4}_{One;50}[L_0]$	$1.35/5.45$	$1.36/5.47$	$1.81/5.61$
$\mathbf{ULS}^{1;1}_{One;500}[L_0]$	$0.46/3.77$	$0.46/3.79$	$1.36/4.19$
$\mathbf{ULS}^{10;1}_{One;500}[L_0]$	$0.38/4.88$	$0.37/4.89$	$1.36/4.97$
$\mathbf{ULS}^{1;1/4}_{One;500}[L_0]$	$0.63/5.25$	$0.64/5.28$	$1.45/5.54$
$\mathbf{ULS}^{10;1/4}_{One;500}[L_0]$	$0.59/5.17$	$0.59/5.21$	$1.45/5.55$
$\mathbb{MLS}^{1;1}_{One;500}[L_0]$	$0.61/5.47$	$0.61/5.48$	$1.60/6.12$
$\mathbb{MLS}^{10;1}_{One;500}[L_0]$	$0.52/5.86$	$0.53/5.87$	$1.52/6.46$
$\mathbb{MLS}^{1;1/4}_{One;500}[L_0]$	$1.69/5.66$	$1.71/5.69$	$1.93/6.05$
$\mathbb{MLS}^{10;1/4}_{One;500}[L_0]$	$1.45/5.75$	$1.43/5.77$	$1.68/6.30$
$\mathbf{ULS}^{1;1}_{One;5000}[L_0]$	$0.48/5.26$	$0.49/5.29$	$1.42/5.51$
$\mathbf{ULS}^{10;1}_{One;5000}[L_0]$	$0.43/5.36$	$0.43/5.40$	$1.40/5.74$
$\mathbf{ULS}^{1;1/4}_{One;5000}[L_0]$	$0.56/5.19$	$0.56/5.22$	$1.45/5.58$
$\mathbf{ULS}^{10;1/4}_{One;5000}[L_0]$	$0.46/5.42$	$0.47/5.46$	$1.46/5.80$
$\mathbb{MLS}^{1;1}_{One;5000}[L_0]$	$0.71/5.83$	$0.72/5.86$	$1.56/6.22$
$\mathbb{MLS}^{10;1}_{One;5000}[L_0]$	$0.57/5.96$	$0.57/5.99$	$1.52/6.48$
$\mathbb{MLS}^{1;1/4}_{One;5000}[L_0]$	$1.82/5.88$	$1.83/5.89$	$1.66/6.32$
$\mathbb{MLS}^{10;1/4}_{One;5000}[L_0]$	$1.45/6.06$	$1.44/6.10$	$1.46/6.59$

Table 1: **Loop PTPN9-MEG2: exploration to reach landmark conformations.** Four conformations of loop PTPN9-MEG2 form two clusters: L_0, L_1, L_2 and L_3. For `MoMA-LS`, we compute min and max lRMSD distances to these landmarks. For $\mathbf{ULS}^{N_V;N_{OR}}_{\text{One}|\text{All};N_{ES}}$ and $\mathbb{MLS}^{N_V;N_{OR}}_{\text{One}|\text{All};N_{ES}}$, starting from L_0, we investigate the ability to get away from the cluster (maxlRMSD values) and to approach conformation L_3 (minlRMSD values).

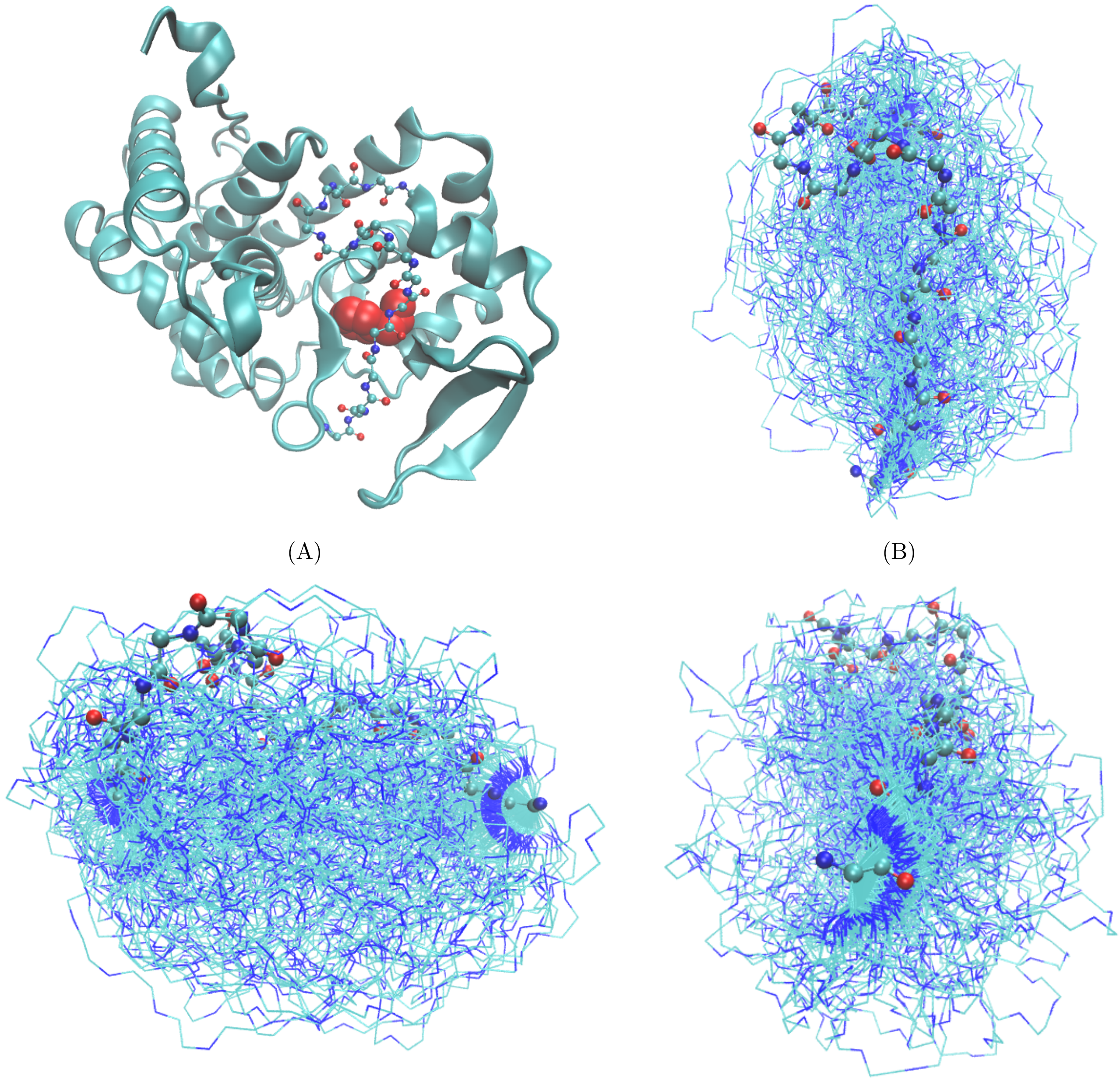
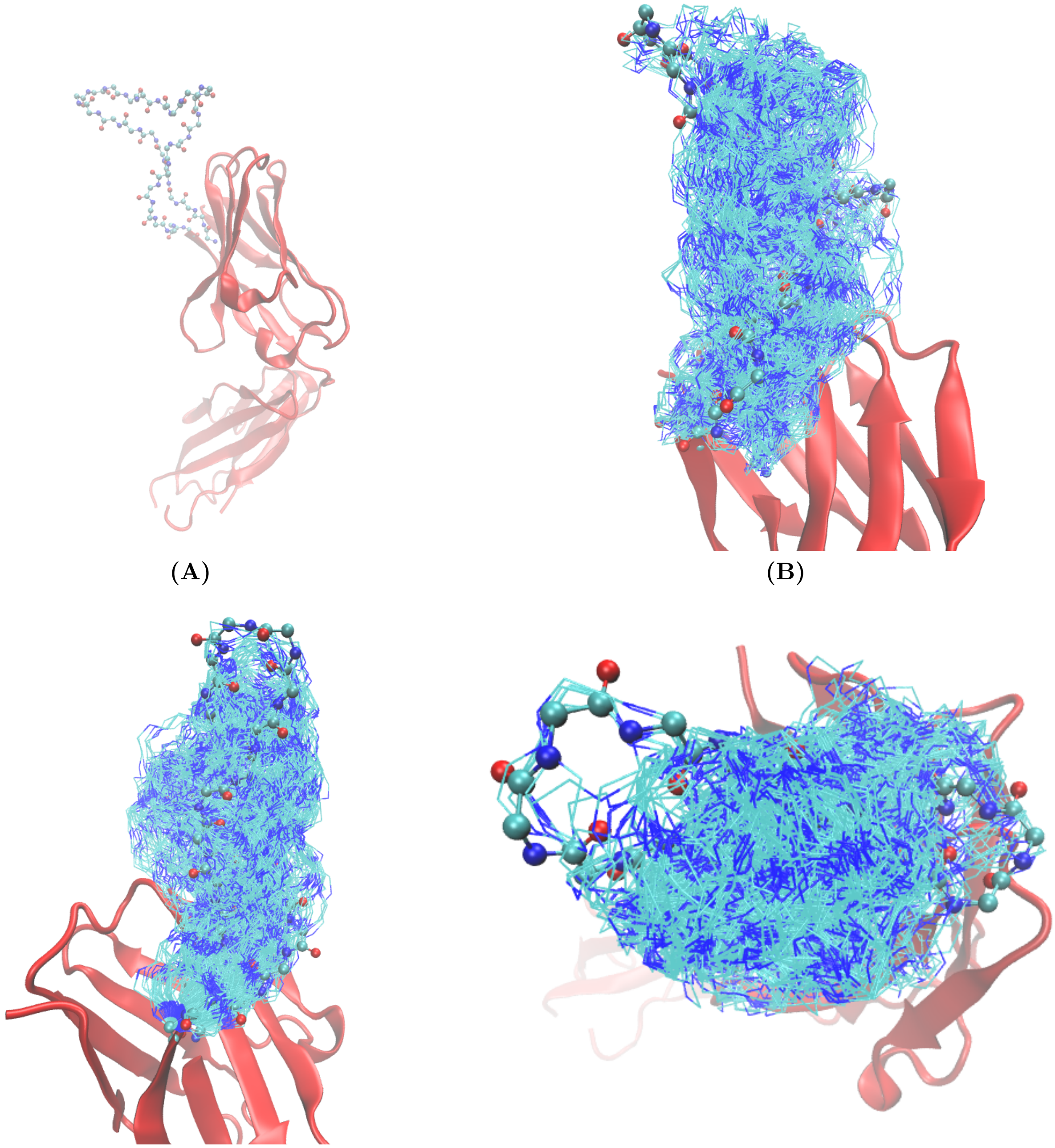


Figure 7: **Loop CCP-W191G, 15 amino acids.** Loop found in cytochrome C peroxidase (CCP). Loop specification: pdbid: 2rbt, chain X, residues 186-200. Conformations generated by algorithm $\mathbb{MLS}^{1;1}_{One;250}$. **(A)** Overview of the protein: cartoon mode: protein; CPK mode: loop; VDW representation: ligand N-Methylbenzylamine. **(B,C,D)** Top, side, front view of the loop conformations. Protein omitted for the sake of clarity.



(A) (B) (C) (D)

Figure 8: **CDR-H3-HIV, 30 amino acids.** The loop is a complementarity-determining region (CDR-H3) from PG16, an antibody with neutralization effect on HIV-1 [41]. Loop specification: 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,D)** Side/front/top view of 250 conformations .