Geometric constraints within tripeptides and the existence of tripeptide reconstructions

Timothée O'Donnell*and Frédéric Cazals†

June 21, 2022

Abstract

Designing movesets providing high quality protein conformations remains a hard problem, especially when it comes to deforming a long protein backbone segment, and a key building block to do so is the so-called tripeptide loop closure (TLC). Consider a tripeptide whose first and last segments ($N_1C_{\alpha;1}$ and $C_{\alpha;3}C_3$) are fixed, and so are all internal coordinates except the six $\{(\phi,\psi)\}_{i=1,2,3}$ dihedral angles associated to the three C_α carbons. Under these conditions, the TLC algorithm provides all possible values for these six dihedral angles–there exist at most 16 solutions. TLC moves atoms up to ~ 5Å in one step and retains low energy conformations, whence it pivotal role to design move sets sampling protein loop conformations.

In this work, we relax the previous constraints, allowing the last segment ($C_{\alpha;3}C_3$) to freely move in 3D space–or equivalently in a 5D configuration space. We exhibit necessary geometric constraints in this 5D space for TLC to admit solutions. Our analysis provides key insights on the geometry of solutions for TLC. Most importantly, when using TLC to sample loop conformations based on m consecutive tripeptides along a protein backbone, we obtain an exponential gain in the volume of the $5m$-dimensional configuration space to be explored.

1 Introduction

1.1 Molecular dynamics and the Tripeptide Loop Closure

Biomolecular motions are inherently multi-scale, spanning ~ 15 and ~ 4 orders of magnitude in time and amplitude respectively [1]. Despite intensive efforts over the past fifty years or so, developing methods able to exploit this multi-scale structure has remained elusive. Two broad classes of methods have been developed. The first one relies on Newton's equations, whose numerical solution uses time steps of the order of femto-seconds [2]. Alas, unless massive simulations are used [3] such tiny time steps are prohibitive for simulations with large systems , or systems undergoing large amplitude conformational changes. The second one, encompassing Monte Carlo based methods and basin-hopping like methods [2, 4], require *movesets* to propose novel conformations, which are accepted or not to further the simulation. While such methods are appealing since arbitrarily large spatial steps may be used, a shear difficulty is to avoid steric clashes and retain low (potential) energy conformations. This latter constraint is a strong incentive to work in internal coordinates, favoring dihedral angles which are *softer* coordinates than bond lengths and valence angles. In designing movesets, one classically distinguishes side chains and the protein backbone. To deal with side chains, various rotamer libraries have been used – see [5, 6, 7] and references therein. Dealing with the backbone is more challenging due to closure constraints arising when dealing with a loop squeezed in-between two *anchor* points.

A key building block in this context is the celebrated Tripeptide Loop Closure (TLC) algorithm [8, 9, 10]. In TLC, one consider three consecutive amino acids, with two types of constraints. The first one stipulates

*Centre Inria at Université Côte d'Azur, France

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

that the segments (also called *anchors* in the sequel) $N_1C_{\alpha;1}C_1$ and $N_2C_{\alpha;2}C_2$ are fixed in a given reference frame. The second one imposes that all internal coordinates are fixed, except the six rotatable bonds / dihedral angles $\{(\phi_i, \psi_i)\}_{i=1,2,3}$ found before / after the three C_α carbons. Remarkably, TLC admits at most 16 solutions. This property sorts of discretizes the search space when running a simulation, in a manner akin to rotameric degrees of freedom for side chains.

Using TLC in the context of backbone simulations raises a difficulty, though. To see which, consider $m(> 1)$ consecutive tripeptides. To use TLC on each tripeptide and mix their solutions if any, one needs to fix the position of all anchors. For a given tripeptide, this raises a novel question which is to study the existence of solutions to TLC when the second anchors moves relatively to the first one. This is precisely the problem addressed in this work, as specified now.

1.2 Contributions

We consider a tripeptide in which all internal coordinates (but the 6 dihedral angles) hold canonical values. Assuming that the two segments $N_1C_{\alpha;1}$ and $C_{\alpha;3}C_3$ are free to move with respect to one another, we aim at finding necessary conditions on these two segments for TLC to admit solutions. A tripeptide yielding solutions is termed *embeddable*. As we can assume without loss of generality that the first segment is fixed in a reference frame, this problem is posed in a five dimensional configuration space: the position of $C_{\alpha;3}$ enjoys 3 Cartesian coordinates, and that of C_3 two spherical coordinates w.r.t. $C_{\alpha;3}$ (Fig. 1). The question becomes to find out which positions of $C_{\alpha;3}$ and C_3 yielding embeddable tripeptides. To answer this question, our contributions are organized as follows:

- In Section 2, we present background material from [9].
- In Section 3, we derive C_α valence constraints at each $C_{\alpha;i}$ carbon to guarantee that the valence angle θ_i at this C_α is preserved. These constraints involve two angles denoted σ_{i-1} and τ_i.
- In Sec. 4, we exploit a constraint associated with each $C_{\alpha;i}C_{\alpha;i+1}$ edge, to derive a deep validity constraint of iterative depth on all σ_{i-1} and τ_i angles. This constraint is thus a necessary condition for the whole tripeptide
- Section 5 provides illustrations of our constructions, showing the sharpness of our constraints in the aforementioned five dimensional space.

These contributions hinge on several interval types for the various angles involved in TLC (Fig. 2): IVI for Initial Validity Intervals, RVI for Rotated Validity Intervals, DVI for Depth-n/Deep Validity Intervals, RDVI for Restricted Deep Validity Intervals. In a sense, this work aims to understand the geometry of solutions of TLC in terms of necessary conditions on the six dihedral angles involved, expressed using these interval types.

2 Background on the Tripeptide Loop Closure

In this section, we review in detail the solution to TLC from [9].

2.1 Rotations and constraints

We consider the tripeptide $(N_1C_{\alpha;1}C_1)(N_2C_{\alpha;2}C_2)(N_3C_{\alpha;3}C_3)$.

Geometry of the C_α triangle of the tripeptide. Consider first the following four consecutive atoms $C_{\alpha;1}C_1N_2C_{\alpha;2}$ along the backbone, with C_1N_2 the peptide bond. Since the ω_1 angle of the peptide bond is fixed, the distance $\|C_{\alpha;1}C_{\alpha;2}\|$ is constant (Fig. 3(A)). This observation holds for the other edge $C_{\alpha;2}C_{\alpha;3}$. Consequently, the atom $C_{\alpha;2}$ is restrained to the circle defined by the intersection of two spheres: $S_1(C_{\alpha;1}, \|C_{\alpha;1}C_{\alpha;2}\|)$ and $S_2(C_{\alpha;3}, \|C_{\alpha;2}C_{\alpha;3}\|)$.

Finally, consider the base of this triangle. By hypothesis, atoms $N_1, C_{\alpha;1}, C_{\alpha;3}, C_3$ are fixed, so that the length of the base is fixed.

The geometry of the triangle $C_{\alpha;1}C_{\alpha;2}C_{\alpha;3}$ is therefore fixed. However, one has one rotating rigid body attached to each of its edges (Fig. 3(B)):

- The movement of the atom C_1 (resp. N_2) can be modeled by a rotation of angle τ_1 (resp. σ_1) about the axis $C_{\alpha;1}C_{\alpha;2}$;
- The movement of the atom C_2 (resp. N_3) can be modeled by a rotation of angle τ_2 (resp. σ_2) about the axis $C_{\alpha;2}C_{\alpha;3}$;
- The movement of atoms $N_1C_{\alpha;1}C_{\alpha;3}C_3$ can be modeled by a rotation of angle τ_3 about the axis $C_{\alpha;1}C_{\alpha;3}$.

Remark 1 *Using the dihedral angle δ_i defined by the four atoms $C_{\alpha;i}, C_i, N_{i+1}, C_{\alpha;i+1}$ (Fig. 3(B)) one has the relationship:*

$$\sigma_i = \tau_i + \delta_i. \tag{1}$$

Local C_α frames. The TLC from [9] defines one local frame to handle the rotation angles τ_i and σ_i. Using these frames and Eq. 1, TLC reduces the problem to three variables and three constraints:

- The three variables: the angles $\tau_i \forall i \in \{1, 2, 3\}$ which can be rotated.
- The three constraints: the valence angles $\theta_i \forall i \in \{1, 2, 3\}$ each at the corresponding $C_{\alpha;i}$, which which must be kept constant.

It is the coupling introduced by the θ_i angles onto the rotation angles τ_i that yields a degree 16 polynomial [9, 10].

We have recalled above that in TLC, the atoms $N_1, C_{\alpha;1}, C_{\alpha;3}, C_3$ are fixed. But in using the rotation angles $\{(\tau_i, \sigma_i)\}$, the segments $N_1C_{\alpha;1}$ and $C_{\alpha;3}C_3$ are moving. Therefore, once the atomic positions have been obtained using the local frames, all atoms are rotated such that these four are back into their original positions in the main frame.

Remark 2 *The valence angles around C_i and N_i atoms do not play a role, since they are fixed and internal to the two segments being rotated.*

2.2 Local coordinate system at $C_{\alpha;i}$

Three **unit** vectors $\hat{\mathbf{Z}}_\mathbf{i}, \hat{\mathbf{Y}}_\mathbf{i}, \hat{\mathbf{X}}_\mathbf{i}$ are defined to form an orthonormal coordinate system (Fig. 3(C)). These vectors are:

- (Unit vector aligned with one side of the triangle) $\hat{\mathbf{Z}}_\mathbf{i}$ is the unit vector anchored at $C_{\alpha;i}$ aligned with the edge of the C_α triangle. Note that $C_{\alpha;3}$ points to $C_{\alpha;1}$.
- (Unit vector perpendicular to the plane of the triangle) $\hat{\mathbf{Y}}_\mathbf{i} = \hat{\mathbf{Z}}_{\mathbf{i+2}} \times \hat{\mathbf{Z}}_\mathbf{i} / \left\| \hat{\mathbf{Z}}_{\mathbf{i+2}} \times \hat{\mathbf{Z}}_\mathbf{i} \right\|$. Intuitively, note that $\hat{\mathbf{Y}}_\mathbf{i}$ is obtained by taking the cross product between two of the unit vectors just mentioned: that *entering* the C_α carbon and that *exiting* it. (Nb: the vector $\hat{\mathbf{Y}}_\mathbf{i}$ does not depend on the index i, which can be dropped.)
- (Reference unit vector to define the rotation angles) $\hat{\mathbf{X}}_\mathbf{i} = \hat{\mathbf{Y}}_\mathbf{i} \times \hat{\mathbf{Z}}_\mathbf{i} = (\hat{\mathbf{Z}}_\mathbf{i} \cdot \hat{\mathbf{Z}}_{\mathbf{i+2}})\hat{\mathbf{Z}}_\mathbf{i} - \hat{\mathbf{Z}}_{\mathbf{i+2}}$. (For the latter, we use the double cross product formula $u \times (v \times w) = (u.w)v - (u.v)w$.)

Using the local frames, one also defines the angles α_i, η_i and ξ_i as follows (Fig. 3(D)):

$$\begin{cases} \alpha_i & = \angle \hat{\mathbf{Z}}_\mathbf{i} \hat{\mathbf{Z}}_\mathbf{i-1}, \alpha_\mathbf{i} \in [\mathbf{0}, \pi) \\ \xi_i & = \angle - \hat{\mathbf{Z}}_\mathbf{i} \hat{\mathbf{r}}^\sigma_\mathbf{i}, \xi_\mathbf{i} \in [\mathbf{0}, \pi) \\ \eta_i & = \angle \hat{\mathbf{Z}}_\mathbf{i} \hat{\mathbf{r}}^\tau_\mathbf{i}, \eta_\mathbf{i} \in [\mathbf{0}, \pi) \end{cases} \tag{2}$$

These variables will be used to handle the conservation of the valence angle at $C_{\alpha;i}$, which requires considering three atoms: N_{i-1}, $C_{\alpha;i}$, and C_i.

Remark 3 *Indices $i \in \{1, 2, 3\}$ are counted modulo 3; that is, $i-1$ is equivalent to $i+2$: e.g., $\hat{\mathbf{Z}}_\mathbf{i+2} = \hat{\mathbf{Z}}_\mathbf{i-1}$.*

Remark 4 *The following equalities are used throughout all calculations [1]:*

$$\begin{cases} \langle \hat{\mathbf{Z}}_\mathbf{i}, \hat{\mathbf{Z}}_\mathbf{i-1} \rangle & = \cos \alpha_i \\ \langle \hat{\mathbf{X}}_\mathbf{i-1}, \hat{\mathbf{X}}_\mathbf{i} \rangle & = \langle \hat{\mathbf{Y}} \times \hat{\mathbf{Z}}_\mathbf{i-1}, \hat{\mathbf{Y}} \times \hat{\mathbf{Z}}_\mathbf{i} \rangle = \cos \alpha_\mathbf{i} \\ \langle \hat{\mathbf{Z}}_\mathbf{i}, \hat{\mathbf{Y}} \rangle & = 0 \\ \langle \hat{\mathbf{Z}}_\mathbf{i}, \hat{\mathbf{X}}_\mathbf{i-1} \rangle & = \langle \hat{\mathbf{Z}}_\mathbf{i}, \hat{\mathbf{Y}} \times \hat{\mathbf{Z}}_\mathbf{i-1} \rangle - \langle \hat{\mathbf{Z}}_\mathbf{i}, \times \hat{\mathbf{Y}} \rangle = -\langle \hat{\mathbf{Z}}_\mathbf{i} \times \hat{\mathbf{Z}}_\mathbf{i-1}, \hat{\mathbf{Y}} \rangle = -\langle -\sin \alpha_\mathbf{i} \hat{\mathbf{Y}}, \hat{\mathbf{Y}} \rangle = \sin \alpha_\mathbf{i} \\ \langle \hat{\mathbf{Z}}_\mathbf{i-1}, \hat{\mathbf{X}}_\mathbf{i} \rangle & = \langle \hat{\mathbf{Z}}_\mathbf{i-1}, \hat{\mathbf{Y}} \times \hat{\mathbf{Z}}_\mathbf{i} \rangle = -\langle \hat{\mathbf{Z}}_\mathbf{i-1}, \hat{\mathbf{Z}}_\mathbf{i} \times \hat{\mathbf{Y}} \rangle = -\langle \hat{\mathbf{Z}}_\mathbf{i-1} \times \hat{\mathbf{Z}}_\mathbf{i}, \hat{\mathbf{Y}} \rangle = -\sin \alpha_\mathbf{i}. \end{cases} \tag{3}$$

2.3 Rotations of N_i and C_i

Consider $\hat{\mathbf{r}}^\sigma_\mathbf{i-1}$ the $C_{\alpha;i}N_i$ unit vector and $\hat{\mathbf{Z}}_\mathbf{i}$ the $C_{\alpha;i}C_i$ unit vector. The rotations of C_i and N_i are described as follows (Fig. 3(C,D)):

- atom N_i, angle σ_{i-1}: rotation of $\hat{\mathbf{r}}^\sigma_\mathbf{i-1}$ about $\hat{\mathbf{Z}}_\mathbf{i-1}$
- atom C_i, angle τ_i: rotation of $\hat{\mathbf{r}}^\tau_\mathbf{i}$ about $\hat{\mathbf{Z}}_\mathbf{i}$

The vectors $\hat{\mathbf{r}}^\sigma_\mathbf{i-1}$ and $\hat{\mathbf{r}}^\tau_\mathbf{i}$ are easily obtained using the local frames (Fig. 3(E)):

$$\text{Frame}(\hat{\mathbf{X}}_\mathbf{i-1}, \hat{\mathbf{Y}}, \hat{\mathbf{Z}}_\mathbf{i-1}): \qquad \hat{\mathbf{r}}^\sigma_\mathbf{i-1} = -\cos \xi_{i-1} \hat{\mathbf{Z}}_\mathbf{i-1} + \sin \xi_\mathbf{i-1} (\cos \sigma_\mathbf{i-1} \hat{\mathbf{X}}_\mathbf{i-1} + \sin \sigma_\mathbf{i-1} \hat{\mathbf{Y}}) \tag{4}$$

$$\text{Frame}(\hat{\mathbf{X}}_\mathbf{i}, \hat{\mathbf{Y}}, \hat{\mathbf{Z}}_\mathbf{i}): \qquad \hat{\mathbf{r}}^\tau_\mathbf{i} = \cos \eta_i \hat{\mathbf{Z}}_\mathbf{i} + \sin \eta_\mathbf{i} (\cos \tau_\mathbf{i} \hat{\mathbf{X}}_\mathbf{i} + \sin \tau_\mathbf{i} \hat{\mathbf{Y}}) \tag{5}$$

Using the previous two equations and the equalities from Eq. (3), one obtains the dot product between $\hat{\mathbf{r}}^\sigma_\mathbf{i-1}$ and $\hat{\mathbf{r}}^\tau_\mathbf{i}$:

$$\begin{aligned} \langle \hat{\mathbf{r}}^\sigma_\mathbf{i-1}, \hat{\mathbf{r}}^\tau_\mathbf{i} \rangle = & -\cos \xi_{i-1} \cos \eta_i \cos \alpha_i \\ & + \cos \xi_{i-1} \sin \eta_i \cos \tau_i \sin \alpha_i \\ & + \cos \eta_i \sin \xi_{i-1} \cos \sigma_{i-1} \sin \alpha_i \\ & + \sin \xi_{i-1} \sin \eta_i (\cos \sigma_{i-1} \cos \tau_i \cos \alpha_i + \sin \sigma_{i-1} \sin \tau_i) \end{aligned} \tag{6}$$

The conservation of the valence angle θ_i imposes the following *valence angle constraint*: – see also Fig. S1:

$$\langle \hat{\mathbf{r}}^\sigma_\mathbf{i-1}, \hat{\mathbf{r}}^\tau_\mathbf{i} \rangle = \cos \theta_i. \tag{7}$$

In TLC only the τ_i and σ_{i-1} terms vary, the value of this dot product can thus be represented depending on those two angles.

Remark 5 *It should be noticed that the angles of Eq. 2 and the $\{\sigma, \tau\}$ angles serve complementary purposes: the angles ξ, η define cones for the rotating atoms N and C, while the angles $\{\sigma, \tau\}$ position the rotating atoms onto a circle orthogonal to the axis of each such cone.*

[1] Recall the triple product formula $\langle a, b \times c \rangle = \langle (a \times b), c \rangle$.

2.4 Internal coordinates and the 5D configuration space

As mentioned in Sec. 1.2, we study TLC when the position of the first two and last two atoms are given, and all internal coordinates (but ϕ, ψ angles) are fixed. (Note that fixing the position of the segment $C_{\alpha;3}C_3$ requires 5 parameters, whence a 5D configuration space.) With these two pieces of informations, we compute all angles which define a unique TLC problem. To obtain the $\{\alpha_i, \chi_i, \eta_i\, \delta_i\}_{i\in 1..3}$ values which are sufficient to solve TLC, we proceed as follows. We first note that η_3, ξ_3 and δ_3 are directly obtained from the positions of the anchors $N_1C_{\alpha;1}$ and $C_{\alpha;3}C_3$. Fixed internal coordinates yield the geometry of the C_α triangle, whence the α_i values. Finally, using the geometry of the C_α triangle and that of the peptide bodies $C_\alpha CNC_\alpha$, one obtains η_i, ξ_i, δ_i for $i \in 1..2$.

3 C_α valence angle constraints

In this section, we study constraints on σ_{i-1} and τ_i so as to guarantee that the valence angle θ_i is preserved. We first present derivation of intervals for σ_{i-1} and τ_i (Sec. 3.1), and proceed with the no-solution case, introducing a necessary condition for solutions to exist (Sec. 3.2).

3.1 Initial validity intervals for σ_{i-1} and τ_i

We first derive initial validity intervals, using boundary conditions at each C_α carbon.

3.1.1 Angle σ_{i-1}

We wish to define a *validity interval* for σ_{i-1}, namely $I_{\sigma_{i-1}} = [\sigma_{i-1;-}, \sigma_{i-1;+}] \subseteq [0, \pi]$.

Basic equation. We first wish to restrict the values of σ_{i-1} for which θ_i can be preserved, and use to this end the reference vector $\hat{\mathbf{Z}}_\mathbf{i}$. Using the expression of $\hat{\mathbf{r}}^\sigma_{\mathbf{i-1}}$ in the local frame of $C_{\alpha;i-1}$ (Eq. (4)), we get:

$$\langle \hat{\mathbf{r}}^\sigma_{\mathbf{i-1}}, \hat{\mathbf{Z}}_\mathbf{i} \rangle = \cos\sigma_{\mathbf{i-1}} \sin\xi_{\mathbf{i-1}} \sin\alpha_\mathbf{i} - \cos\xi_{\mathbf{i-1}} \cos\alpha_\mathbf{i}. \tag{8}$$

Independently from the angle σ_{i-1}, elementary spherical trigonometry based on Viète's law of cosines makes it possible to bound the angle $\angle \hat{\mathbf{r}}^\sigma_{\mathbf{i-1}} \hat{\mathbf{Z}}_\mathbf{i}$ as follows (Sec. S8.2, Fig. S2):

$$\langle \hat{\mathbf{r}}^\sigma_{\mathbf{i-1}}, \hat{\mathbf{Z}}_\mathbf{i} \rangle \in [\cos(\theta_\mathbf{i} + \eta_\mathbf{i}), \cos(\theta_\mathbf{i} - \eta_\mathbf{i})]. \tag{9}$$

Exploiting the bounds of the interval yields two extreme values for the angle σ, denoted $\sigma_{i-1;-}$ and $\sigma_{i-1;+}$.

Angle $\sigma_{i-1;-}$. The first limit case reads as:

$$\langle \hat{\mathbf{r}}^\sigma_{\mathbf{i-1};-}, \hat{\mathbf{Z}}_\mathbf{i} \rangle = \cos(\theta_\mathbf{i} - \eta_\mathbf{i}) \tag{10}$$

from which we obtain

$$\begin{cases} S^- = \frac{+\cos(\theta_i - \eta_i) + \cos\xi_{i-1}\cos\alpha_i}{\sin\xi_{i-1}\sin\alpha_i} \\ \sigma_{i-1;-} = \arccos S^- \end{cases} \tag{11}$$

When $S^- \to 1^-$ by properties of arccos, we have $\sigma_{i-1;-} \to 0^+$ (Fig. S3). Therefore, when

$$S^- \geq 1, \tag{12}$$

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

Angle $\sigma_{i-1;+}$. The second limit case reads as:

$$\langle \hat{\mathbf{r}}^{\sigma}_{\mathbf{i-1};+}, \hat{\mathbf{Z}}_{\mathbf{i}} \rangle = \cos(\theta_{\mathbf{i}} + \eta_{\mathbf{i}}) \tag{13}$$

from which we obtain

$$\begin{cases} S^+ = \frac{+\cos(\theta_i+\eta_i)+\cos\xi_{i-1}\cos\alpha_i}{\sin\xi_{i-1}\sin\alpha_i} \\ \sigma_{i-1;+} = \arccos S^+ \end{cases} \tag{14}$$

When $S^+ \to -1^+$, by properties of arccos, we have $\sigma_{i-1;+} \to \pi^-$. Therefore, when

$$S^+ \leq -1, \tag{15}$$

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

Illustration. When considering the dot product $\langle \hat{\mathbf{r}}^{\sigma}_{\mathbf{i-1}}, \hat{\mathbf{r}}^{\tau}_{\mathbf{i}} \rangle$ as a function of the two variables τ_i and σ_{i-1}, the angles $\sigma_{i-1;-}$ and $\sigma_{i-1;+}$ correspond to planes orthogonal to the σ_{i-1} axis (Fig. 4(B,C)).

3.1.2 Angle τ_i

We also wish to set a *validity interval* for τ_i, that is $I_{\tau_i} = [\tau_{i;-}, \tau_{i;+}] \subseteq [0, \pi]$.

Basic equation. We proceed mutatis mutandis for the vector $\hat{\mathbf{r}}^{\tau}_{\mathbf{i}}$, using vector $\hat{\mathbf{Z}}_{\mathbf{i-1}}$ as landmark. Using the expression of $\hat{\mathbf{r}}^{\tau}_{\mathbf{i}}$ in the local frame of $C_{\alpha;i}$ (Eq. (5)), one obtains:

$$\langle \hat{\mathbf{r}}^{\tau}_{\mathbf{i}}, \hat{\mathbf{Z}}_{\mathbf{i-1}} \rangle = -\cos\tau_{\mathbf{i}} \sin\eta_{\mathbf{i}} \sin\alpha_{\mathbf{i}} + \cos\eta_{\mathbf{i}} \cos\alpha_{\mathbf{i}} \tag{16}$$

Viète's law of cosines now yields the following for the angle $\hat{\mathbf{r}}^{\tau}_{\mathbf{i}}$:

$$\langle \hat{\mathbf{r}}^{\tau}_{\mathbf{i}}, \hat{\mathbf{Z}}_{\mathbf{i-1}} \rangle \in [-\cos(\theta_{\mathbf{i}} - \xi_{\mathbf{i-1}}), -\cos(\theta_{\mathbf{i}} + \xi_{\mathbf{i-1}})]. \tag{17}$$

Combining the previous two equations yields up to two solutions, namely $\tau_{i;-}$ and $\tau_{i;+}$.

Angle $\tau_{i;-}$. The first limit case reads as:

$$\langle \hat{\mathbf{r}}^{\tau}_{\mathbf{i};-}, \hat{\mathbf{Z}}_{\mathbf{i-1}} \rangle = -\cos(\theta_{\mathbf{i}} - \xi_{\mathbf{i-1}}) \tag{18}$$

from which we obtain

$$\begin{cases} T^- = \frac{+\cos(\theta_i-\xi_{i-1})+\cos\eta_i\cos\alpha_i}{\sin\eta_i\sin\alpha_i} \\ \tau_{i;-} = \arccos T^- \end{cases} \tag{19}$$

When $T^- \to 1^-$, by properties of arccos, we have $\tau_{i;-} \to 0^+$. Therefore, when

$$T^- \geq 1, \tag{20}$$

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

Angle $\tau_{i;+}$. The second limit case reads as:

$$\langle \hat{\mathbf{r}}^{\tau}_{\mathbf{i};-}, \hat{\mathbf{Z}}_{\mathbf{i-1}} \rangle = -\cos(\theta_{\mathbf{i}} + \xi_{\mathbf{i-1}}) \tag{21}$$

from which we obtain

$$\begin{cases} T^+ = \frac{+\cos(\theta_i+\xi_{i-1})+\cos\eta_i\cos\alpha_i}{\sin\eta_i\sin\alpha_i} \\ \tau_{i;+} = \arccos T^+ \end{cases} \tag{22}$$

When $T^+ \to -1^+$, by properties of arccos, we have $\tau_{i;+} \to \pi^-$. Therefore, when

$$T^+ \leq -1, \tag{23}$$

we set $\tau_{i;+} = \pi$, so that any value of $\tau_i \geq \tau_{i;-}$ is valid.

Illustration. When considering the dot product $\langle \hat{\mathbf{r}}^{\sigma}_{\mathbf{i-1}}, \hat{\mathbf{r}}^{\tau}_{\mathbf{i}} \rangle$ as a function of the two variables τ_i and σ_{i-1}, the angles $\tau_{i-1;-}$ and $\tau_{i-1;+}$ correspond to planes orthogonal to the τ_i axis (Fig. 4,(B),(C)).

3.2 Necessary conditions for σ_{i-1} and τ_i

In deriving the lower and upper bounds of the initial validity intervals for σ and τ, we already processed four limit cases for the dot products (Eqs. (12), (15), (20), (23)). The remaining four yield the following:

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

- *Angle* $\sigma_{i-1;-}$*: the condition* $\sigma_{i-1;-} < \sigma_{i-1;+}$ *requires*
$$S^- \geq -1. \tag{24}$$
- *Angle* $\sigma_{i-1;+}$*: the condition* $\sigma_{i-1;-} < \sigma_{i-1;+}$ *requires*
$$S^+ \leq 1. \tag{25}$$
- *Angle* $\tau_{i;-}$*: the condition* $\tau_{i;-} < \tau_{i;+}$ *requires*
$$T^- \geq -1. \tag{26}$$
- *Angle* $\tau_{i;+}$*: the condition* $\tau_{i;-} < \tau_{i;+}$ *requires*
$$T^+ \leq 1. \tag{27}$$

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

Summarizing, when the previous equations are not verified, no validity interval can be defined for σ_{i-1} and/or τ_i (Fig. 5(E)).

3.3 Symmetry around the C_α triangular plane, and C_α valence constraints

The previous angles are defined in $[0, \pi]$. Due to the symmetry of the tripeptide with respect to the C_α plane, these angles have counterparts in $[\pi, 2\pi]$. We therefore define the following symmetric intervals:

- The *symmetric validity interval* for σ_{i-1} is defined by
$$I^{'}_{\sigma_{i-1}} = [\sigma^{'}_{i-1;-}, \sigma^{'}_{i-1;+}] \stackrel{Def}{=} [2\pi - \sigma_{i-1;+}, 2\pi - \sigma_{i-1;-}]. \tag{28}$$
- The *symmetric validity interval* for τ_i is defined by
$$I^{'}_{\tau_i} = [\tau^{'}_{i;-}, \tau^{'}_{i;+}] \stackrel{Def}{=} [2\pi - \tau_{i;+}, 2\pi - \tau_{i;-}]. \tag{29}$$

Using these, we can finally specify the valid intervals for the angles σ_{i-1} and τ_i must belong to:

Definition. 2 *The* initial validity interval set *for* σ_{i-1} *is defined by:*
$$\mathcal{I}_{\sigma_{i-1}} = \left\{ I_{\sigma_{i-1}}, I^{'}_{\sigma_{i-1}} \right\} \tag{30}$$
Likewise, the initial validity interval set for τ_i *is defined by:*
$$\mathcal{I}_{\tau_i} = \left\{ I_{\tau_{i-1}}, I^{'}_{\tau_i} \right\}. \tag{31}$$

For σ_{i-1}, as long as the conditions of Eqs. (24) and (25) are satisfied, we have:
$$\sigma_{i-1} \in \mathcal{I}_{\sigma_{i-1}}. \tag{32}$$
For τ_i, as long as the conditions of Eqs. (26) and (27) are satisfied, we have:
$$\tau_i \in \mathcal{I}_{\tau_i}. \tag{33}$$

Combining the previous conditions yields the complete case analysis (Fig. 5).

4 Deep validity constraints associated with the C_α triangle

4.1 Exploiting the coherence along a $C_{\alpha;i}C_{\alpha;i+1}$ edge

The constraints presented in the previous section focus on the three C_α carbons independently. On the other hand, for a given tripeptide, the angles τ_i and σ_i and the dihedral angle δ_i defined by the four atoms $C_{\alpha;i}C_iN_{i+1}C_{\alpha;i+1}$ satisfy $\sigma_i = \tau_i + \delta_i$ (Eq. (1)).

Given an interval of values for τ_i, the previous formula can be used to infer a projected interval for σ_i, and vice-versa (Fig. 2(B)). Whence the following definitions which exploit the previous formula along the two edges of the C_α triangle incident on $C_{\alpha;i}$:

Definition. 3 *The* rotated validity interval set *for the angles σ_{i-1} and τ_i are defined by:*

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

Summarizing, for each of the σ_{i-1} and τ_i angles, we have obtained 4+4 intervals using the θ_i angle constraint at $C_{\alpha;i}$ (Eq. 6):

- Four for the σ_{i-1} angle: $I_{\sigma_{i-1}}, I'_{\sigma_{i-1}}, I_{\sigma_{i-1}|\delta}, I'_{\sigma_{i-1}|\delta}$
- Four for the τ_i angle: $I_{\tau_i}, I'_{\tau_i}, I_{\tau_i|\delta}, I'_{\tau_i|\delta}$

4.2 Deep validity Intervals: depth one

We are now in position to combine two pieces of information:

- The conditions on σ_{i-1} and τ_i inherent to the conservation of the valence angles (Eq. (7)).
- The conditions exploiting rotated validity intervals, stemming from Eq. (1)

Since the previous intervals define necessary conditions, intersections between intervals for a given angle must be non empty. We therefore combine them as follows $(I_{\cdot_{i-1}}, I'_{\cdot_{i-1}}) \times (I_{\cdot_{i-1}|\delta}, I'_{\cdot_{i-1}|\delta})$, which yields *depth 1 validity intervals*:

Definition. 4 *The* depth 1 validity interval set $\mathcal{J}^{(1)}_{\sigma_{i-1}}$ *for σ_{i-1} is:*

$$\mathcal{J}^{(1)}_{\sigma_{i-1}} = \left\{ \{I_{\sigma_{i-1}} \cap I_{\sigma_{i-1}|\delta}\}, \{I_{\sigma_{i-1}} \cap I'_{\sigma_{i-1}|\delta}\}, \{I'_{\sigma_{i-1}} \cap I_{\sigma_{i-1}|\delta}\}, \{I'_{\sigma_{i-1}} \cap I'_{\sigma_{i-1}|\delta}\} \right\} \tag{34}$$

Similarly, the depth 1 validity interval set $\mathcal{J}^{(1)}_{\tau_i}$ for τ_i:

$$\mathcal{J}^{(1)}_{\tau_i} = \left\{ \{I_{\tau_i} \cap I_{\tau_i|\delta}\}, \{I_{\tau_i} \cap I'_{\tau_i|\delta}\}, \{I'_{\tau_i} \cap I_{\tau_i|\delta}\}, \{I'_{\tau_i} \cap I'_{\tau_i|\delta}\} \right\} \tag{35}$$

Note that each depth 1 validity interval has up to four connected components.

Definition. 5 *The* depth 1 validity constraint *for σ_{i-1} is:*

$$\mathcal{J}^{(1)}_{\sigma_{i-1}} \neq \{\emptyset\} \tag{36}$$

The depth 1 validity constraint for τ_i is:

$$\mathcal{J}^{(1)}_{\tau_i} \neq \{\emptyset\} \tag{37}$$

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

As we shall see, this constraint is significantly more restrictive than the corresponding C_α valence constraint (Def. 1, Fig. 6)

The derivation of these validity constraints can be seen as a bootstrap process (Fig. 2). The initialization consists of computing the initial validity intervals, using the boundary conditions imposed by Equations (11) (14) (19) (22). The second step consists of exploiting the coherence along each C_α edges, as imposed by Eq. (1).

It should also be noticed that transposing the initial validity intervals from τ_i to σ_i using Eq. 1 is equivalent to transposing from σ_i to τ_i using the same equation. Therefore the *depth 1 validity constraint* for σ_i and the one for τ_i are redundant.

Remark 6 *The previous intersections naturally depend on the two anchor positions and the fixed internal coordinates.*

4.3 Deep Validity Intervals: arbitrary depth

The qualifier *depth 1* used in the previous section indicates that the dual process *specify validity intervals at each C_α* and *project along a C_α edge* can be repeated, moving from necessary conditions at depth j to necessary conditions at depth depth $j+1$.

To see how, we first note that an interval for σ_{i-1} or τ_i implies two intervals for the second angle – obtained by computing σ_{i-1} from τ_i, or vice versa. To see how, note that the dot product equation (6) can be written as:

$$K^{(1)}_{\sigma_{i-1}} \cos\sigma_{i-1} + K^{(2)}_{\sigma_{i-1}} \sin\sigma_{i-1} + K^{(3)}_{\sigma_{i-1}} = 0, \tag{38}$$

with

$$\begin{cases} K^{(1)}_{\sigma_{i-1}} & = \cos\tau_i \sin\xi_{i-1} \sin\eta_i \cos\alpha_i + \sin\alpha_i \cos\eta_i \sin\xi_{i-1} \\ K^{(2)}_{\sigma_{i-1}} & = \sin\xi_{i-1} \sin\eta_i \sin\tau_i \\ K^{(3)}_{\sigma_{i-1}} & = -\cos\xi_{i-1} \cos\eta_i \cos\alpha_i + \cos\xi_{i-1} \sin\eta_i \cos\tau_i \sin\alpha_i - \cos\theta_i \end{cases} \tag{39}$$

This latter equation makes it possible to obtain σ_{i-1} given τ_i. Dividing by $\sqrt{{K^{(1)}_{\sigma_{i-1}}}^2 + {K^{(2)}_{\sigma_{i-1}}}^2}$ and using the trigonometric identity $\cos(a-b) = \cos a \cos b + \sin a + \sin b$, the two values for σ_{i-1} $(\sigma^*_{i-1}, \sigma^{**}_{i-1})$ given τ_i are obtained as follows:

$$\begin{cases} \cos\left(\sigma^*_{i-1}(\tau_i)\right) & = \cos\left(\arccos\frac{K^{(1)}_{\sigma_{i-1}}}{\sqrt{{K^{(1)}_{\sigma_{i-1}}}^2 + {K^{(2)}_{\sigma_{i-1}}}^2}} + \arccos\frac{-K^{(3)}_{\sigma_{i-1}}}{\sqrt{{K^{(1)}_{\sigma_{i-1}}}^2 + {K^{(2)}_{\sigma_{i-1}}}^2}}\right) \\ \cos\left(\sigma^{**}_{i-1}(\tau_i)\right) & = \cos\left(2\pi + \arccos\frac{K^{(1)}_{\sigma_{i-1}}}{\sqrt{{K^{(1)}_{\sigma_{i-1}}}^2 + {K^{(2)}_{\sigma_{i-1}}}^2}} - \arccos\frac{-K^{(3)}_{\sigma_{i-1}}}{\sqrt{{K^{(1)}_{\sigma_{i-1}}}^2 + {K^{(2)}_{\sigma_{i-1}}}^2}}\right) \end{cases} \tag{40}$$

Remark 7 *Eq. (40) defines the cosine of $\sigma^*_{i-1}(\tau_i)$ and $\sigma^{**}_{i-1}(\tau_i)$. For each of them, two values are possible namely σ_{i-1} and $2\pi - \sigma_{i-1}$. Each time only one of those values validates the valence angle constraint of Eq. (7).*

With these, we define:

Definition. 6 *For each interval* $I \overset{Def}{=} [{\tau_i}^{\min}, {\tau_i}^{\max}] \in \mathcal{J}^{(j)}_{\tau_i}$*, consider the two intervals*

$$\begin{cases} I^*_\sigma(I) = [\min_{\tau_i \in I}(\sigma^*_{i-1}(\tau_i)), \max_{\tau_i \in I}(\sigma^*_{i-1}(\tau_i))], \\ I^{**}_\sigma(I) = [\min_{\tau_i \in I}(\sigma^{**}_{i-1}(\tau_i)), \max_{\tau_i \in I}(\sigma^{**}_{i-1}(\tau_i))]. \end{cases} \tag{41}$$

The depth j validity interval set *is defined by:*

$$\mathcal{K}^{(j)}_{\sigma_{i-1}} = \bigcup_{I \in \mathcal{J}^{(j)}_{\tau_i}} \{I^*_\sigma(I), I^{**}_\sigma(I)\}. \tag{42}$$

One proceeds mutatis mutandis to obtain the values and intervals for τ_i (τ^*_i, τ^{**}_i) given σ_{i-1}, as well as $\mathcal{K}^{(j)}_{\tau_i}$.

Given the depth 1 validity intervals for σ_{i-1}, a set of intervals can be defined for τ_i. A temporary set of intersections between this set and the depth 1 validity intervals for τ_i can then be defined. Finaly using intersections between temporary sets and rotated temporary sets (Eq. 1) we can obtain *depth 2 validity intervals* (Fig. 2(C)). This iterative process is summarized in Algo. 1.

Algorithm 1 Computing depth j validity intervals

Input: $\mathcal{J}^{(j)}_{\sigma_{i-1}}$ and $\mathcal{J}^{(j)}_{\tau_i}$, $i = 1, 2, 3$
Output: $\mathcal{J}^{(j+1)}_{\sigma_{i-1}}$ and $\mathcal{J}^{(j+1)}_{\tau_i}$, $i = 1, 2, 3$

for i $\in \{1, 2, 3\}$ **do**
  Step 1a Compute σ_{i-1} given τ_i and vice versa(Eq. (40))
  Step 1b Using $\mathcal{J}^{(j)}_{\sigma_{i-1}}$ and $\mathcal{J}^{(j)}_{\tau_i}$, define $\mathcal{K}^{(j)}_{\sigma_{i-1}}$ and $\mathcal{K}^{(j)}_{\tau_i}$ (Def. 6)
  Step 2a Define $\mathcal{J}^{(tmp)}_{\sigma_{i-1}}$, the intersections between $\mathcal{J}^{(j)}_{\sigma_{i-1}}$ and $\mathcal{K}^{(j)}_{\sigma_{i-1}}$.
  Step 2b Do the same for $\mathcal{J}^{(tmp)}_{\tau_i}$
  if $\mathcal{J}^{(tmp)}_{\sigma_{i-1}} = \{\emptyset\}$ or $\mathcal{J}^{(tmp)}_{\tau_i} = \{\emptyset\}$ **then**
    There will be no solutions for TLC
  Step 3: Project the $\mathcal{J}^{(tmp)}_{\sigma_{i-1}}$ and $\mathcal{J}^{(tmp)}_{\tau_i}$ along the edges of the C_α triangle using Eq. (1)
  Define $\mathcal{J}^{(j+1)}_{\sigma_{i-1}}$ as the intersections between $\mathcal{J}^{(tmp)}_{\sigma_{i-1}}$ and the projected $\mathcal{J}^{(tmp)}_{\tau_{i-1}}$.
  Do the same for $\mathcal{J}^{(j+1)}_{\tau_i}$ using $\mathcal{J}^{(tmp)}_{\tau_i}$ and the projected $\mathcal{J}^{(tmp)}_{\sigma_i}$
  if $\mathcal{J}^{(j+1)}_{\sigma_{i-1}} = \{\emptyset\}$ or $\mathcal{J}^{(j+1)}_{\tau_i} = \{\emptyset\}$ **then**
    There will be no solutions for TLC

Maximum number of deep validity intervals (DVI).

Lemma. 1 *Let* n_j *be the maximum number of intervals at depth* j *in* $\mathcal{J}^{(j)}_{\sigma_{i-1}}$ *or* $\mathcal{J}^{(j)}_{\tau_i}$*. This number satisfies*

$$n_1 = 4, n_{j+1} = 6n_j. \tag{43}$$

Proof. We first raise an observation for two sets of intervals $\mathcal{I}_1$ (of size n_1) and $\mathcal{I}_2$ (of size n_2) such that $I_1 \cap I_2 = \emptyset, \forall I_1 \in \mathcal{I}_1, \forall I_2 \in \mathcal{I}_1 \backslash I_1$. Assuming, without loss of generality, that $n_2 \geq n_1$, the maximum number of intervals determined by the intersections between intervals in these sets is $2 \times n_1 + n_2 - n_1$ for intervals on a circle. To build this worst-case, we stab two intervals in $\mathcal{I}_1$ with an interval in $\mathcal{I}_2$, and squeeze the remaining $n_2 - n_1$ intervals from $\mathcal{I}_2$ inside intervals of $\mathcal{I}_1$.

To establish the lemma, we follow the steps of Alg0. 1, starting with the set $\mathcal{J}^{(j)}_{\sigma_{i-1}}$. If needed, we redefine the intervals in this set so that they are disjoint – to meet the hypothesis on the two sets $\mathcal{I}_1$ and $\mathcal{I}_2$ used in the observation above.

Considering the n_j intervals making up $\mathcal{J}^{(j)}_{\sigma_{i-1}}$ (or $\mathcal{J}^{(j)}_{\tau_i}$), we now inspect the following steps of Algo. 1:

- **Step 1b**) projects $\mathcal{J}^{(j)}_{\tau_i}$ to yield $\mathcal{K}^{(j)}_{\sigma_{i-1}}$ (Def. 6). By definition this yields $2n_j$ restricted validity intervals.
- **Step 2a**) takes the intersection between $\mathcal{J}^{(j)}_{\sigma_{i-1}} \cap \mathcal{K}^{(j)}_{\sigma_{i-1}}$, which the observation above, yields $2n_j + (2n_j - n_j) = 3n_j$ intervals.
- **Step 3**) takes the intersection between the $3n_j$ intervals of the last step and the $3n_j$ from the δ projected intervals. This yields a maximum of $6n_j$ intervals.

□

5 C_α valence constraint and depth 1 validity interval constraints: illustrations

5.1 Material: dataset of random instances

Our experiments use standard internal coordinates for bond lengths and valence angles [9]. The canonical values are available in our TLC implementation user manual (https://sbl.inria.fr/doc/Tripeptide_loop_closure-user-manual.html)

General dataset. We place our tripeptide in the reference frame using the first segment (Fig. 1). In this frame, we randomly generate $C_{\alpha;3}$ between two spheres around the origin. The inner radius $r_1 = 2$Å is smaller than the smallest value $\|C_{\alpha;3} - C_{\alpha;1}\|$ found in our exhaustive database of ~ 2.5 million tripeptides extracted from the PDB [11]. The outer radius $r_2 = 2\,\|C_{\alpha;2} - C_{\alpha;1}\|$ respects the triangle inequality based on the distances between C_α carbons, and the canonical internal coordinates values. The position of $C_{\alpha;2}$ is generated uniformly in this volume. Atom C_3 is generated on sphere centered at $C_{\alpha;3}$, with a radius defined by the canonical bond length. The positions of $N_1C_{\alpha;1}$ and $C_{\alpha;3}C_3$ together with the canonical internal coordinate yield a TLC problem, which is fertile if embeddings/solutions are obtained. To each of those inputs correspond values for the input angles α_i, ξ_i, η_i for each $C_{\alpha;i}$.

Over $100,000$ instances, $24,076$ fertile ones were observed.

Planar dataset. A second similar dataset of the same size with $C_{\alpha;i+2}$ being positioned uniformly between two circles using the same radii. C_{i+2} is then also generated on a sphere around $C_{\alpha;i+2}$.

5.2 Validity intervals

Signatures. To assess the diversity of situations faced at C_α carbons, we introduce a signature based on the σ and τ interval types:

Definition. 7 *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 $-1 < cos(endpoint) < 1$.*

Among the sample of embeddable anchor positions for TLC, only seven different signatures are found (Fig. S4). Not all combinations are possible as the the first and third letter cannot be N (Eq. (26)). Similarly the second and fourth cannot be P (Eq. (27)). If the first and second are ZZ (resp. third and fourth) then their counterpart will be PN. This illustrates that if the two endpoints are defined in $]0, \pi[$ for σ_{i-1} then it means that the whole circle is possible for τ_i (and vice versa).

5.3 Necessary conditions and validity constraints

General dataset. Using the general dataset, plotting the positions of $C_{\alpha;3}$ associated with solutions yields a bell shaped distribution with a hole on top (Fig. 6(A)).

Planar dataset. Due to the symmetry around the $N_1C_{\alpha;1}$, we consider the aforementioned planar data set (Fig. 6(B)). Consider a set of conformation, some fertile (TLC admits solutions), and some sterile. Fertile conformations naturally satisfy the necessary conditions defined by both the C_α *valence constraints* (Def. 1) and the depth 1 validity constraints (Def. 5) for all σ and τ angles. However, we would like these constraints to be as tight as possible, retaining as few sterile configurations as possible. To assess this, for each constraint, we color code sterile configurations (Fig. 6(B), red points) using two colors: orange for sterile but failing the necessary test, and yellow for sterile passing the necessary test. Ideally, we would like as few yellow points as possible. It clearly appears that the *depth 1 validity constraints* (Fig. 6(D)) are much tighter than the C_α *valence constraints* (Fig. 6(C)).

In terms of proportions, 24.08% of the points are instances where TLC yields solutions. In total 41.35% fulfill the C_α valence constraints. Finally 28.12% fulfill the *depth one validity constraints*. They are included in the 41.35% and include the 24.08% just mentioned. The cases validating the depth one validity constraint therefore represent $\sim 2/3$ of the cases fulfilling the θ_i linked equations. For the sake of clarity, it should be stressed that the gap between the 24.08% of instances with solutions, and the 28.12% fulfilling the depth one validity constraints is not clearly visible using our projections in 3D and 2D (Fig. 6) since the whole configuration space is 5D.

Nevertheless, the of 2/3 gained can be used to scale the savings when it comes to explore the conformational space associated with m tripeptides defining a flexible region along a protein backbone. For various plausible values of m, one obtains: $m = 5 :\sim (3/2)^5 = 7.6$ $m = 10 :\sim (3/2)^{10} = 57.7$ $m = 15 :\sim (3/2)^{15} = 438$ $m = 20 :\sim (3/2)^{20} = 3325$. Therefore, when m increases, an exponential reduction is the size of the search space to be explored is gained. This strategy is used in the companion paper [12]. Using depth j validity intervals with $j > 1$ can further reduce the search space as illustrated with the diminishing number of false positives (Fig. 7). Starting at depth 2 we have 18.12% fulfilling the constraint, then 16.57% for depth 3. There is no false positive in the planar dataset for depth 4 and onwards.

6 Outlook

The compact nature of folded proteins makes the exploration of their conformational space especially challenging. One must indeed avoid steric clashes and optimizes interactions, while dealing with complex coupled sub-problems involving the backbone and side-chains. A critical need in this context are so-called movesets able to propose plausible (low energy) configurations given a starting pause. Such movesets are indeed a corner stone in Monte Carlo based simulations at large.

The design of backbone movesets is a problem in itself, due to the necessity to handle loop closure constraints. The tripeptide loop closure provides an optimal solution to this problem, and therefore plays a crucial role to develop move sets. However, for loops involving several tripeptides, the question of combining solutions yielded by the individual tripeptides remains a challenging problem. Greedy approaches incrementally concatenating tripeptides have been developed, but these break the symmetry between the individual peptides, as the degree of freedom of those near the endpoints enjoy a finer sampling. On the way to processing all tripeptides in a sequence on an equal footing, this work studies (tight) necessary conditions on the first and last segment (bond) of a tripeptide, for TLC to yield solutions. As illustrated by our experiments, our conditions are rather tight, and yield an exponential saving in terms of the conformational space to be explored, when pooling several tripeptides. We leave the problem of improving the tightness of our constraints (notably using their iterated versions) as an open problem. Application-wise, the direct use of our constraints for the design of backbone movesets is presented in a companion paper.

7 Artwork

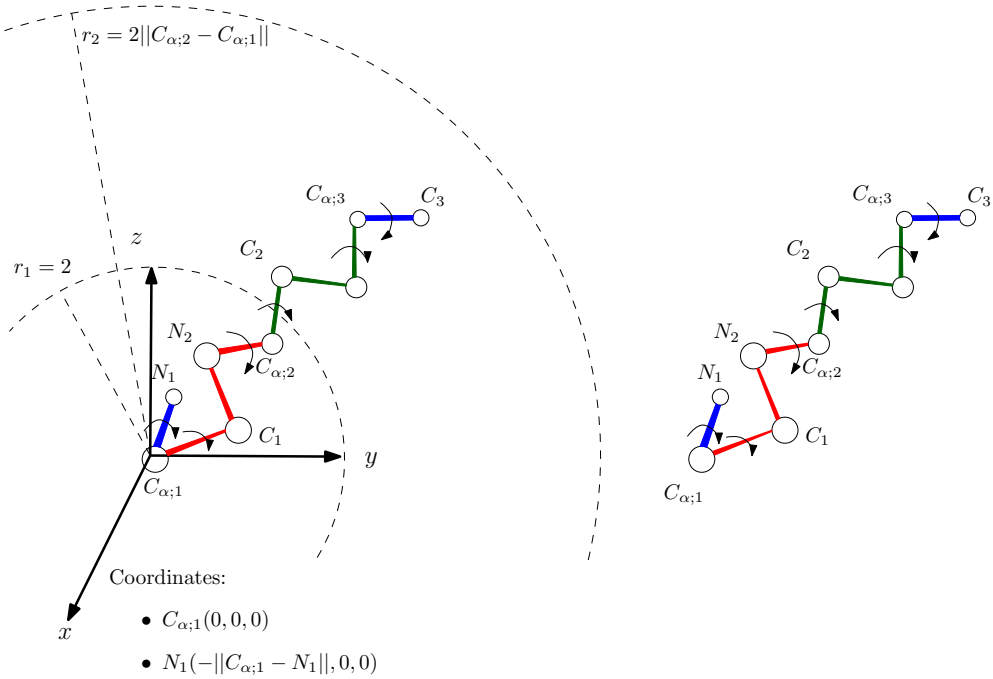

Figure 1: **Reference frame for tripeptide embeddings.** We consider a tripeptide whose internal coordinates are fixed, except the six $\{(\phi, \psi)\}$ dihedral angles associated with the three C_α carbons. We assume that the segment $N_1 C_{\alpha;1}$ (first red line line segment) is fixed *i.e.* $C_{\alpha;1}$ is placed at the origin, and N_1 is placed at $(-\|N_1 - C_{\alpha;1}\|, 0, 0)$. We then aim at characterizing necessary conditions on the position of the last segment *i.e.* $C_{\alpha;3}C_3$ for the Tripeptide Loop Closure (TLC) algorithm to hold solutions.

Acknowledgments. Théo Roncalli and Viraj Agashe are acknowledged for carefully re-reading the paper.

This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

References

[1] S.A. Adcock and A.J. McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews*, 106(5):1589–1615, 2006.

[2] D. Frenkel and B. Smit. *Understanding molecular simulation.* Academic Press, 2002.

[3] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

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

[5] Maxim V Shapovalov and Roland L Dunbrack Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858, 2011.

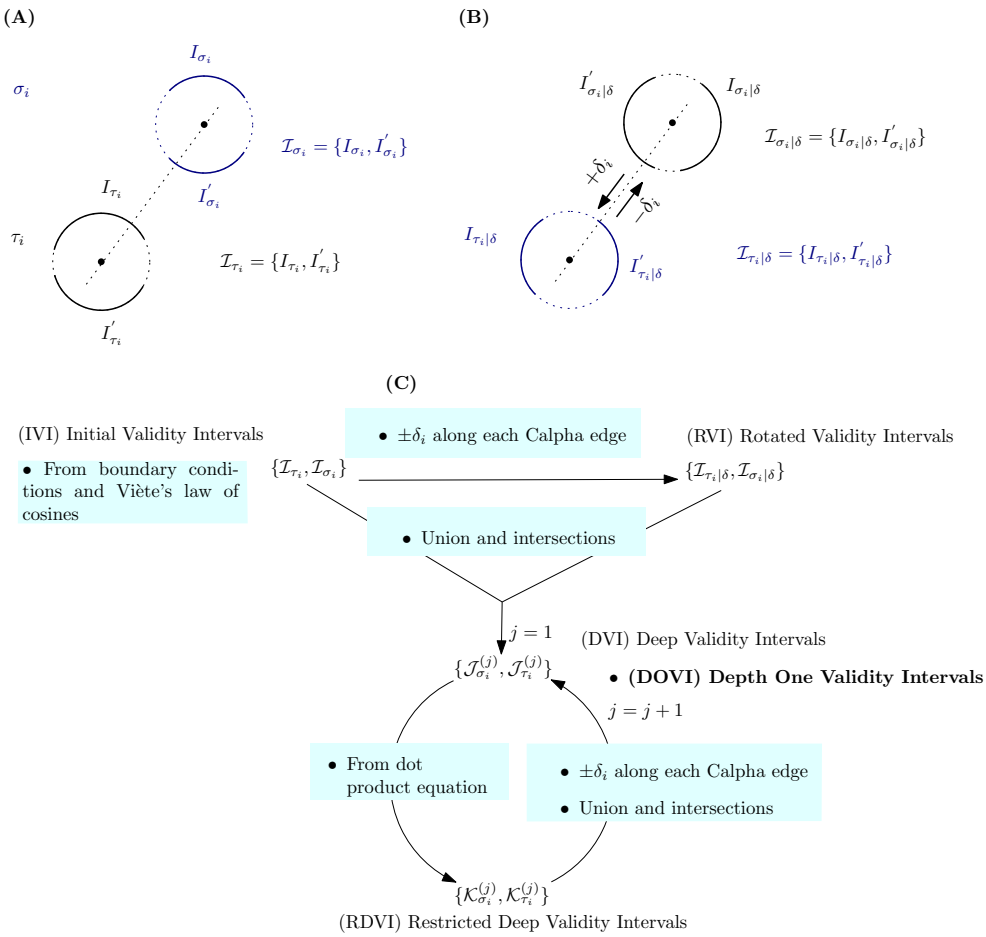

Figure 2: **Validity interval types and their relationships. (A) (IVI) Initial Validity Intervals.** See Def. 2. **(B) (TVI) Rotated validity intervals.** See Def. 3. Obtained from the initial validity intervals (**(A)**). **(C) Depth j/Deep Validity Intervals and their restrictions.** From $\mathcal{I}_\tau(i)$ and $\mathcal{I}_\sigma(i)$ we obtain $\mathcal{I}_{\tau|\delta}(i)$ and $\mathcal{I}_{\sigma|\delta}(i)$. From all of those we obtain intersections constituting $\mathcal{J}^{(1)}_{\tau_i}$ and $\mathcal{J}^{(1)}_{\sigma_i}$. This *depth one validity interval* set can be refined to depth n iteratively (Def. 4, Algo 1).

[6] Bradley J Hintze, Steven M Lewis, Jane S Richardson, and David C Richardson. Molprobity's ultimate rotamer-library distributions for model validation. *Proteins: Structure, Function, and Bioinformatics*, 84(9):1177–1189, 2016.

[7] Clare-Louise Towse, Steven J Rysavy, Ivan M Vulovic, and Valerie Daggett. New dynamic rotamer libraries: data-driven analysis of side-chain conformational propensities. *Structure*, 24(1):187–199, 2016.

[8] 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.

[9] 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.

[10] 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.

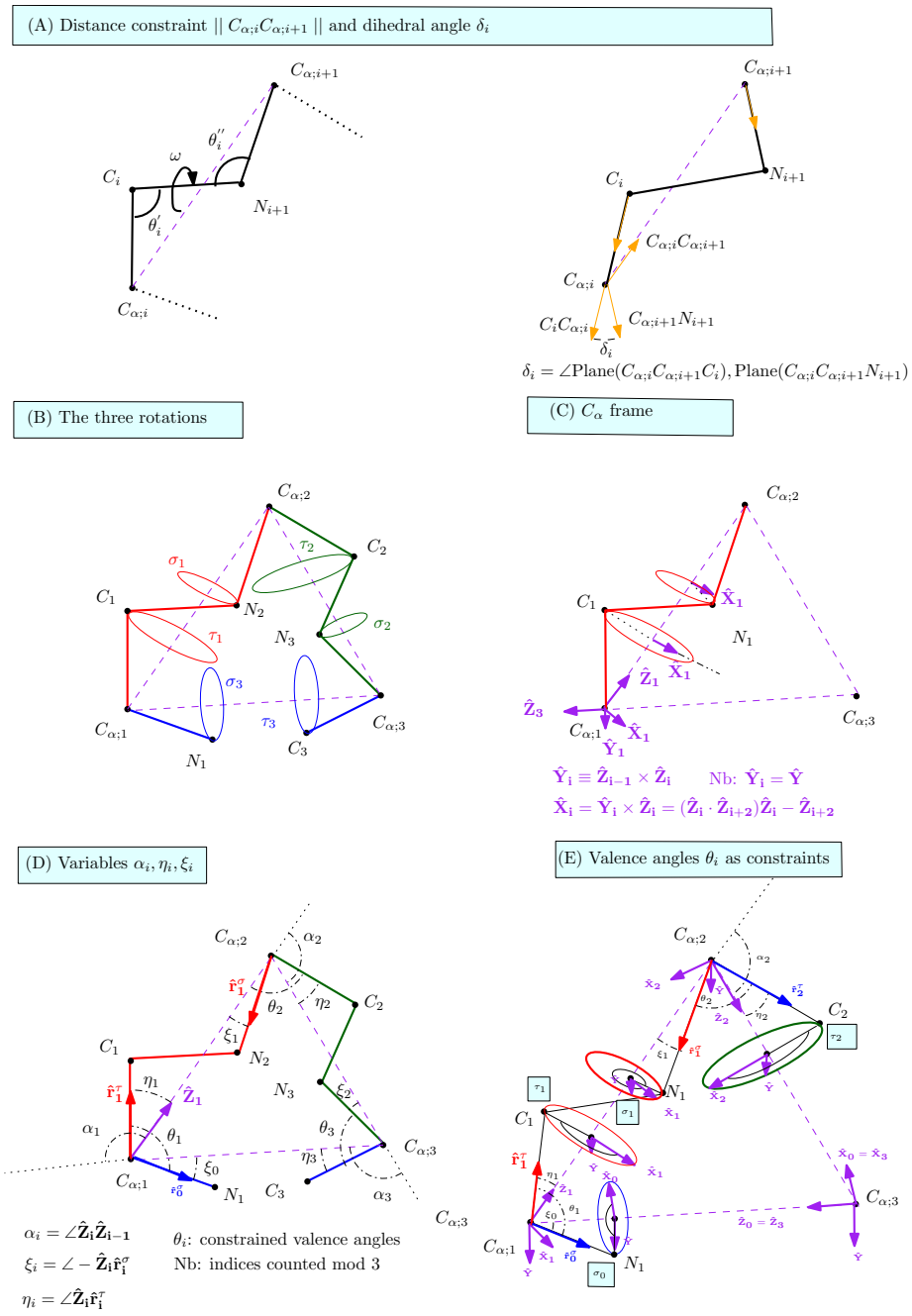

Figure 3: **Tripeptide Loop Closure: main steps of the construction.** Adapted from [9]. **(A)** Peptide bond linking two consecutive amino acids, and distance constraint induced on the line segment $C_{\alpha;i}C_{\alpha;i+1}$. The dihedral angle δ_i is defined by the three vectors $C_iC_{\alpha;i}, C_{\alpha;i}C_{\alpha;i+1}, C_{\alpha;i+1}N_{i+1}$. **(B)** The three rotations associated with the segments $C_{\alpha;1}C_{\alpha;2}$, $C_{\alpha;2}C_{\alpha;3}$ and $C_{\alpha;3}C_{\alpha;1}$. The rotation angles τ_i (resp. σ_i) concern atoms C_i (resp. N_i). But τ_i and σ_i satisfy $\sigma_i = \tau_i + \delta_i$. **(C)** Construction of the local orthonormal frame associated with $C_{\alpha;i}$ i.e. the C_α frame. **(D)** Introducing the variables α_i, η_i, ξ_i. **(E)** Modeling the constraint on valence angles at $C_{\alpha;i}$ carbons.

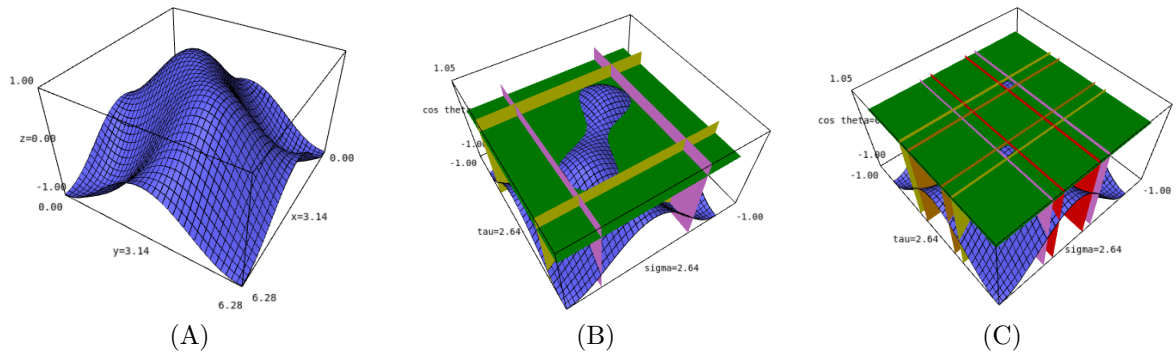

Figure 4: **Example dot product surface and extreme angles** $\sigma_{i-1;-}, \sigma_{i-1;+}, \tau_{i-1;-}, \tau_{i-1;+}$**.** TLC problem for the values $\alpha_i = 100, \chi_{i-1} = 50, \eta_i = 50$ **(A)** Whole surface **(B)** With horizontal plane $\cos\theta_i = \cos 9°$. Note the four vertical planes corresponding to the extreme angles. In this case, the intersection between the surface and the plane consists of a plane curve with two connected components. One component is enclosed by the four vertical planes. **(C)** With horizontal plane $\cos\theta_i = \cos 35°$. In this case, the intersection between the surface and the plane consists of a plane curve with one connected component.

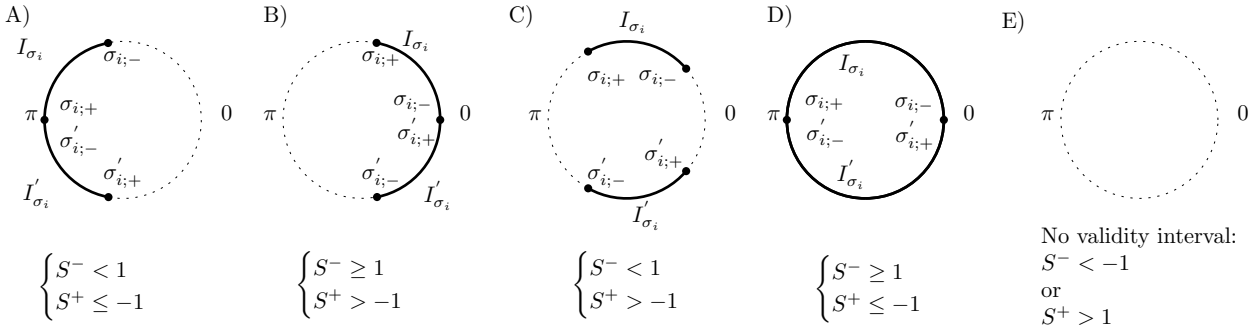

Figure 5: **The four possible types of *initial/rotated* validity interval types for angle** σ**.** The quantities of interest are defined in Eqs. (12) (15), (24), (25). Cases (A) to (D) stand for situations where $\sigma_{i;-}$ and/or $\sigma_{i;+}$ can be defined. In case (E), no validity interval can be defined.

[11] 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.

[12] T. O'Donnell and F. Cazals. Protein loop sampling based on a global parameterization of the backbone conformational space. Technical report, 2022.

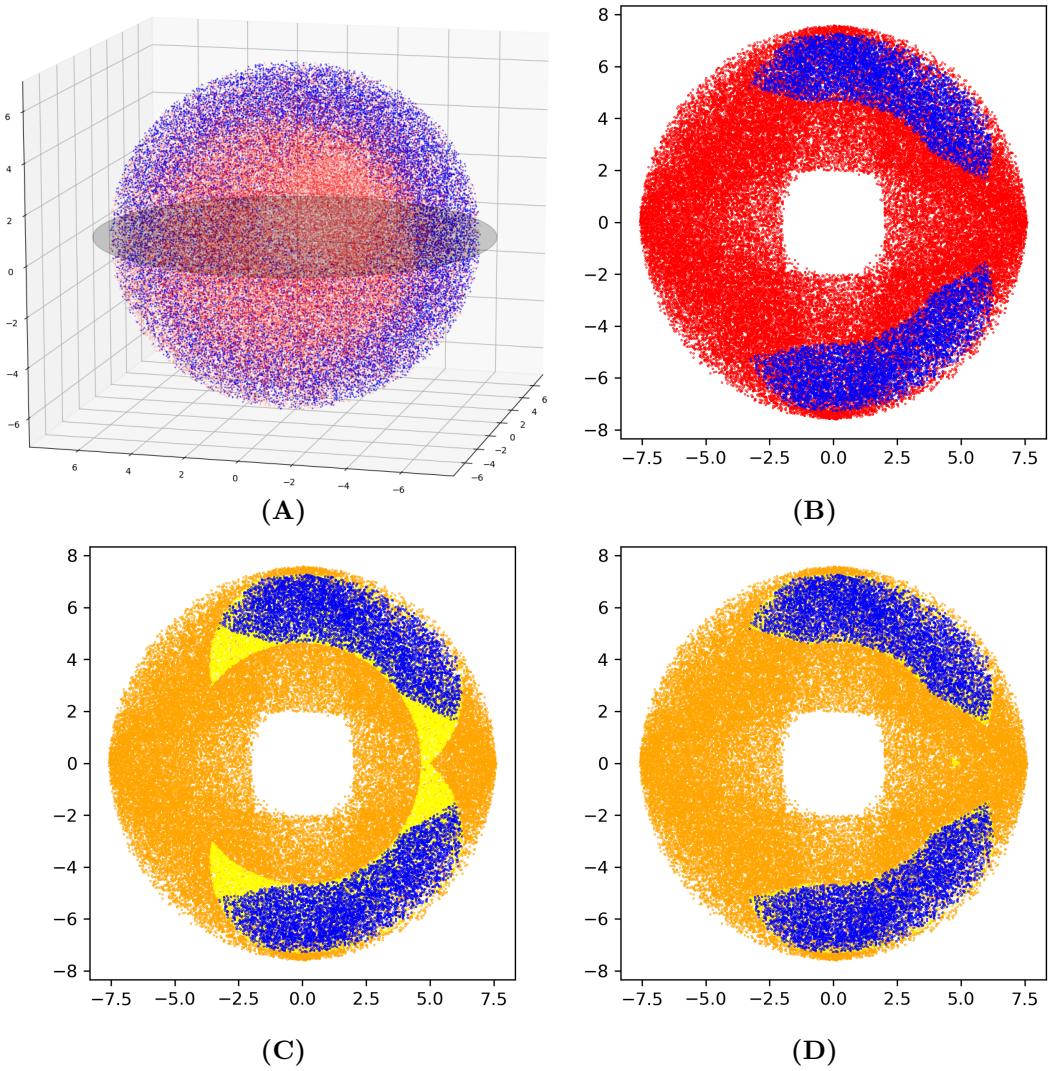

Figure 6: **Embeddable tripeptides and necessary conditions: stringency of C_α valence constraints (Def. 1) versus depth 1 validity constraints (Def. 5), illustrated on random instances projected into the reference frame of Fig. 1. (Nb: figures in 3D and 2D, while the configuration space is 5D.) (A)** Blue (resp. red) points represent positions of $C_{\alpha;3}$ in instances when TLC yields at least one solution (resp. yields no solution). **(B)** A similar dataset generated uniformly on the of the sphere–gray equator in (A), color code as in (A). **(C)** C_α **valence constraints.** The $C_{\alpha;3}$ positions are depicted using three colors: blue points as in (A,B); orange points: points failing the C_α *valence constraints*; yellow points: points satisfying the C_α *valence constraints*, but for which TLC admits no solution. **(D)** ***Depth 1* validity constraints.** Color code as in (C), using the depth 1 validity constraints instead of the C_α valence constraints. Note the reduction of the yellow region.

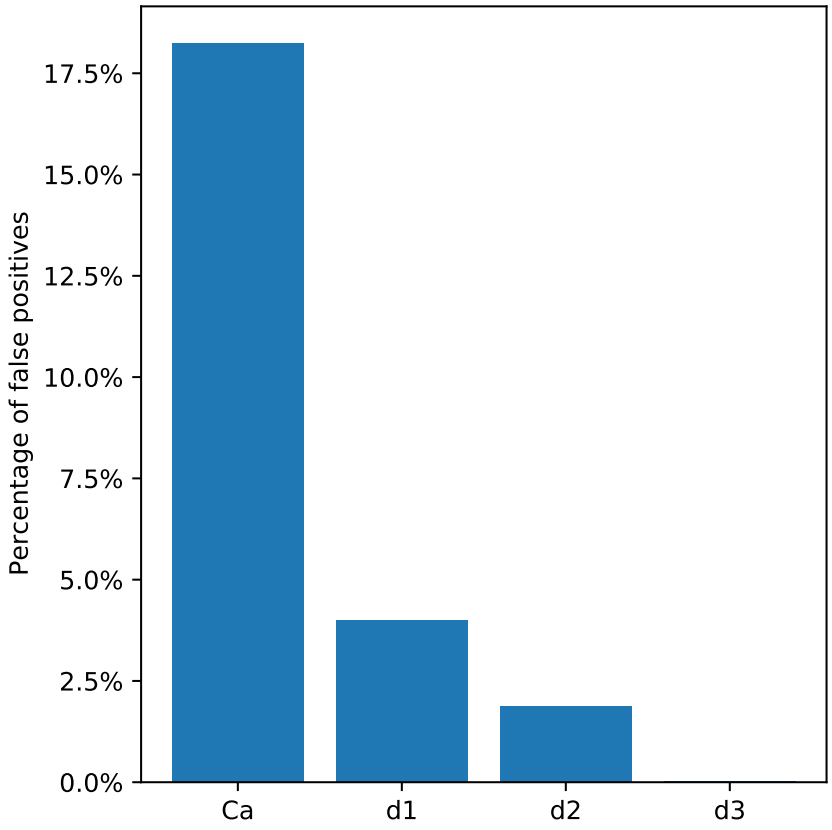

Figure 7: **Proportion of false positives for $\mathbf{C}_\alpha$ valence and depth j validity intervals with $j \in \{1, 2, 3\}$.** The proportion is defined as the number of false positives divided by the number instances when TLC yields no solution in the planar dataset (Sec. 5.1). The specific percentages are, C_α valence constraint: 18.24%, depth 1 validity constraint: 4.01%, depth 2 validity constraint: 1.89%, depth 3 validity constraint: 0.02%.