Geometry, topology, and machine learning, with Applications in Computational Structural Biology

F. Cazals
Inria, Algorithms-Biology-Structure
https://team.inria.fr/abs/

Outline

- F. Cazals
  - Sampling biomolecular conformations
  - Dimensionality reduction and motion modeling
  - Statistical tests and applications
- J-D. Boissonnat
  - Triangulation of point clouds
  - Meshing non linear manifolds
- M. Carriere
  - Persistent Homology and applications to phylogeny and viral evolution
  - Mapper visualization and applications to single-cell inference
  - Multiparameter Persistent Homology and applications to immunohistology

Perspective and warning(s)

- Overarching goals:
  - Efficient (optimal) algorithms for the statistical physics of biomolecules (proteins, nucleic acids)
  - Help molecular modeling to move from an Art to a Science
- Warning(s):
  - Not a unique body of deep mathematics / computer science, but...
  - complex / high-dimensional modeling problems,...
  - borrowing upon/contributing to various topics
    - Theoretical biophysics and statistical physics
    - Geometric modeling
    - Randomized algorithms
    - Numerical probability theory
    - Numerics
    - Machine learning
    - Algebra
    - ...

Key players - collaborators

- Protein geometry

Timothee O'Donnell

- Volume of polytopes

Augustin Chevallier

Mining molecular flexibility: novel tools, novel insights

PART 1: Introduction to Structural Bioinformatics
PART 2: Protein structure and geometry
PART 3: Thermodynamics and the volume of polytopes
PART 4: Outlook

Mining molecular flexibility: novel tools, novel insights

Computational Structural Biology: what is a protein?

Protein functions: structure and dynamics

Computational Structural Biology: challenges

The importance of dynamics

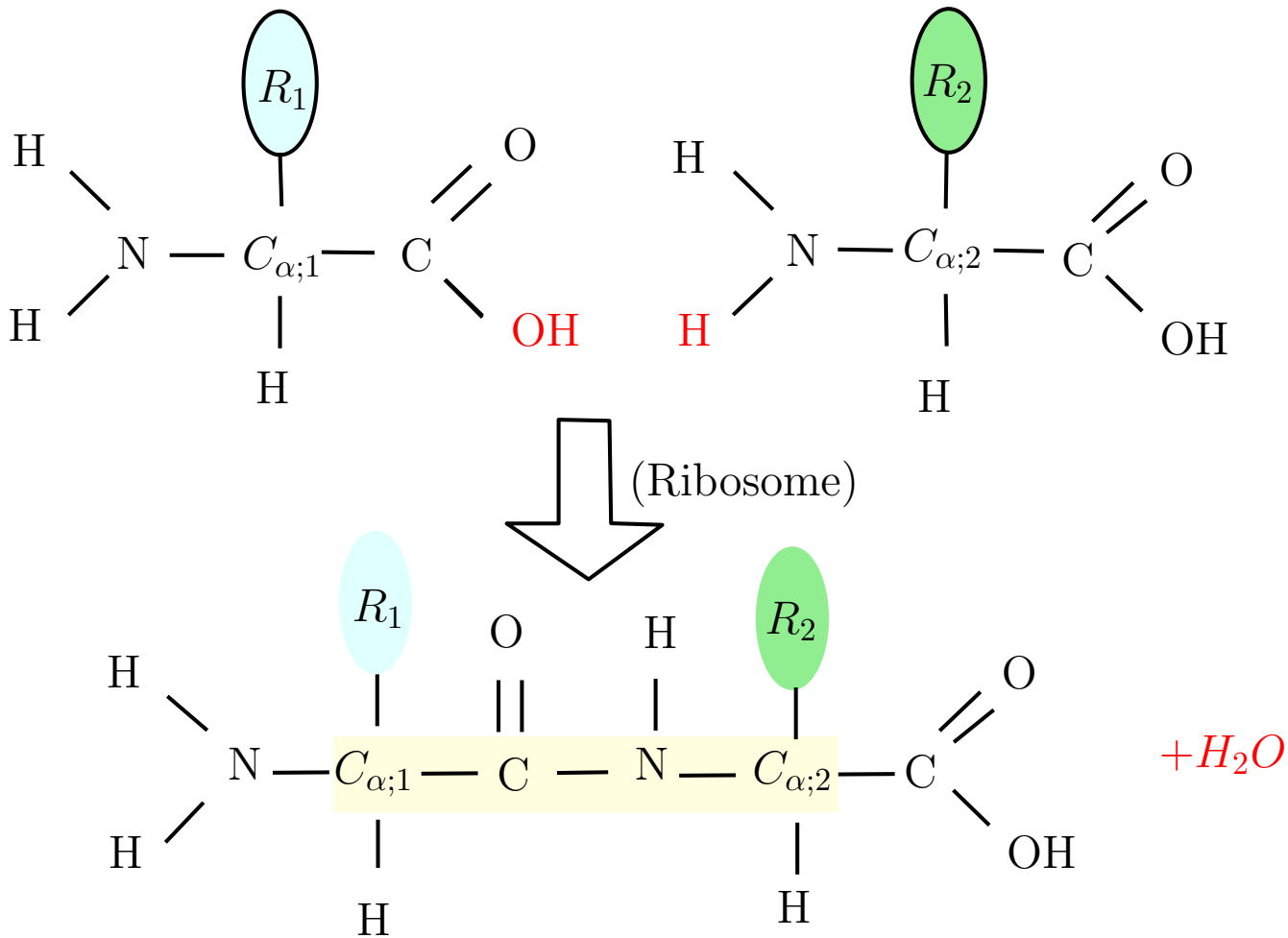
Main difficulties phrased in modern (geometric) terms

Amino acids and the peptide bond

- ▷ Natural amino acids and their side chains

Nb: 0 to 10 heavy atoms per side chain

- ▷ Peptide bond synthesis:

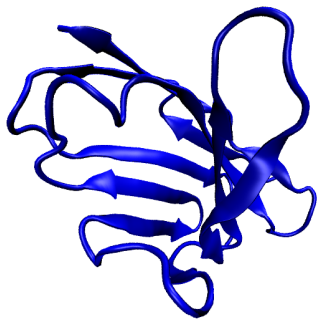


What is a protein?

- Primary structure: sequence of amino acids

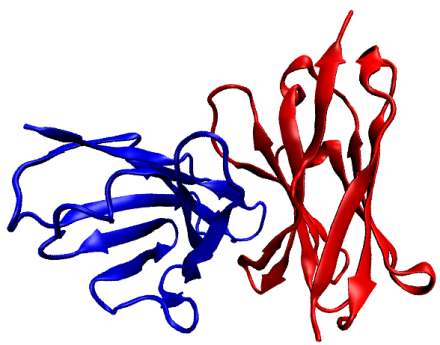
```
P69905 (HBB_HUMAN) MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GS 53
P68871 (HBB_HUMAN) MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN 58
P02144 (MYG_HUMAN) -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKAS 59
                    : *:  :   *   ****         * * *::   * *   *  *  *
```

- Polypeptide chain



- Protein - protein complex

- Heterodimeric protein



- Nb: median number of a.a. in a chain: ~ 400

Computational Structural Biology

- Goals: unveil the *structure-dynamics-function* conundrum for biomolecules (proteins and nucleic acids)
- Methods: biophysics (crystallography, NMR, cryo-microscopy) + modeling
- Nobel prizes as of 01/2019 [1]: related to molecular/structural biology
  - Chemistry or Physiology-medicine for structures and mechanisms: 64
  - Chemistry or Physics for Methods : 11
  - Chemistry 2013: Levitt, Karplus, Warshel for *the development of multiscale models for complex chemical systems*
- An extraordinary field
  - Technology driven: novel biophysical experiments,
  - Reveals the molecular foundations of biology and medicine,
  - Raises open mathematical / computational questions.

[1] https://pdb101.rcsb.org/learn/flyers-posters-and-other-resources/other-resource/structural-biology-and-nobel-prizes

Methods: molecular simulation

The Nobel Prize in Chemistry 2013
Martin Karplus, Michael Levitt, Arieh Warshel

The Nobel Prize in Chemistry 2013

© Harvard University
Martin Karplus

Photo: © S. Fisch
Michael Levitt

Photo: Wikimedia Commons
Arieh Warshel

The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel *"for the development of multiscale models for complex chemical systems"*.

Mining molecular flexibility: novel tools, novel insights

Computational Structural Biology: what is a protein?

Protein functions: structure and dynamics

Computational Structural Biology: challenges

The importance of dynamics

Main difficulties phrased in modern (geometric) terms

Molecular dynamics and protein functions: movies

▷ Selected (great) movies:

- Protein synthesis by the ribosome:
  `https://www.youtube.com/watch?v=TfYf_rPWUdY`
- Membrane fusion and infection by SARS-Cov-2:
  `https://youtu.be/e2Qi-hAXdJo`
- Molecular motors: `https://www.youtube.com/watch?v=X_tYrnv_o6A`

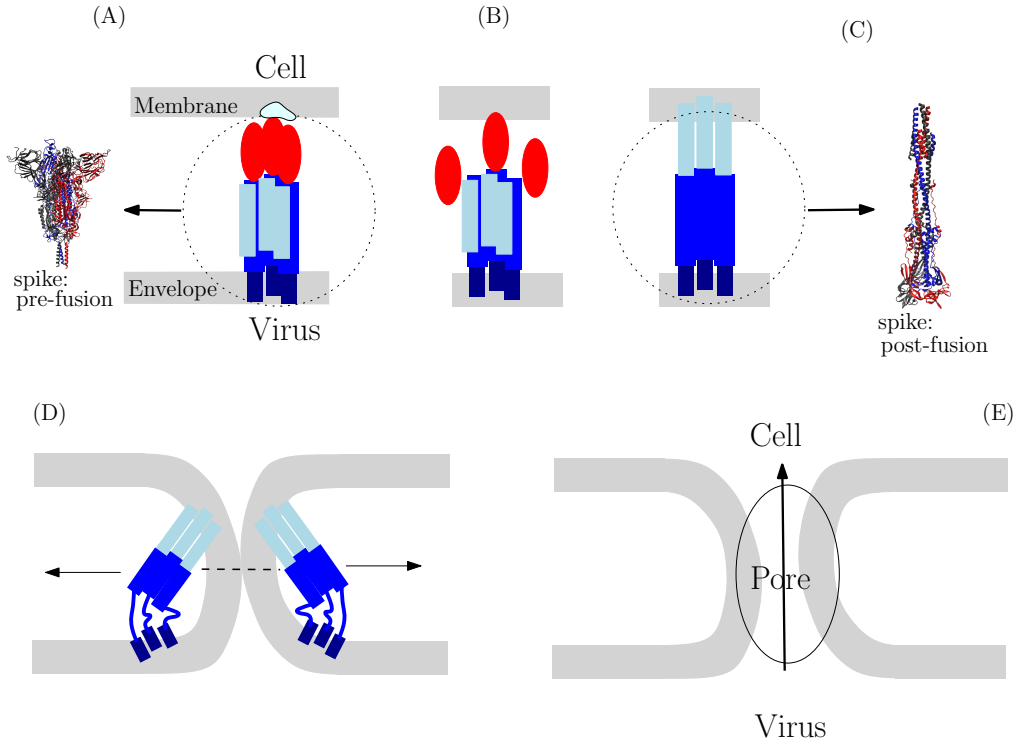
▷ Other videos of interest:

- Various phenomena in this movie:
  `https://www.youtube.com/watch?v=wJyUtbn005Y`
- More XVivo movies at
  `https://www.youtube.com/channel/UCAUL7Wl_lydKXI8q0oi4CUw`

▷ Rmk. Remarkable illustration of the aforementioned mechanisms can be found in the book [?]; see also the gallery on the PDB portal, at `https://pdb101.rcsb.org/sci-art/goodsell-gallery`.

SARS-Cov-2: cell entry mechanism

▷ SARS-Cov-2: cell entry mechanism via virus envelope - cell membrane fusion:



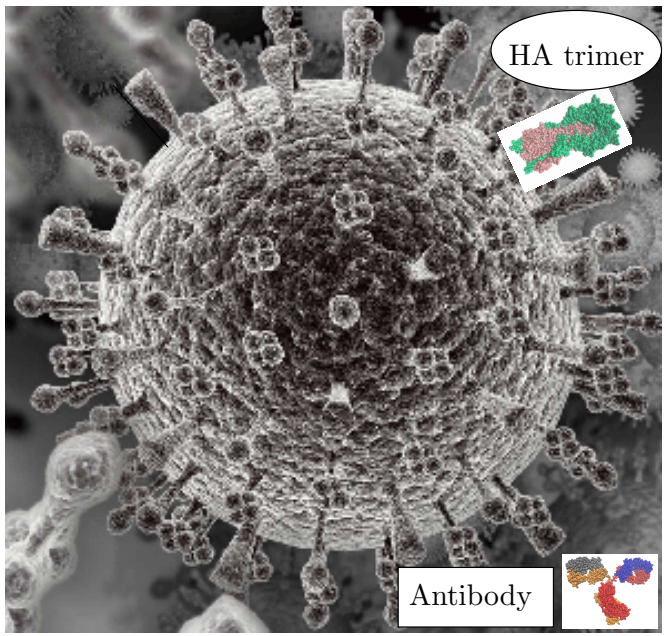
▷ Spike, the S1 and S2 domains: S1: the receptor binding domain (RBD, red ellipsis); S2: the fusion machinery (blue rectangles)

- **(A)** Attachment of the RBD to its receptor ACE2
- **(B)** Cleavage step removing the S1 subunit
- **(C)** Fusion machinery: refolding + membrane anchoring
- **(D,E)** Formation of the hemi-pore and pore

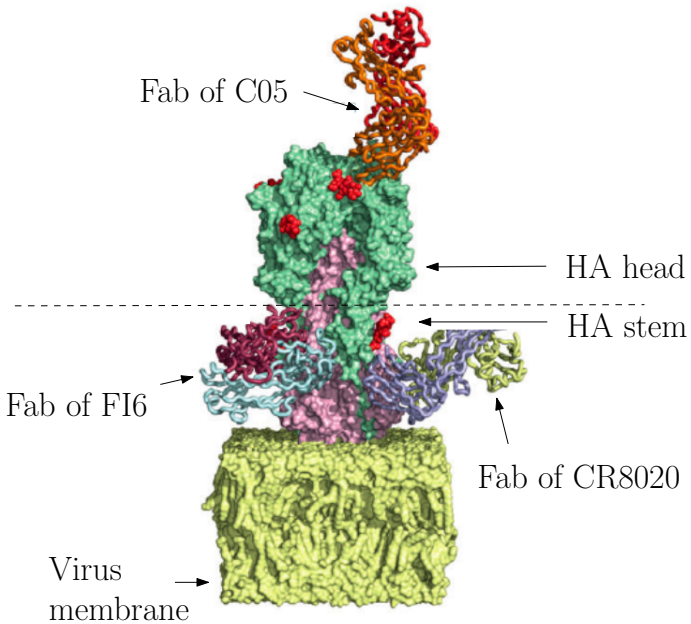
▷ Biophysics and biology of SARS-Cov-2/Omicron – Marc Gozlan:
`https://www.lemonde.fr/blog/realitesbiomedicales/2022/02/09/omicron-une-biologie-et-une-dynamique-virale-differentes-de-celles-observees-c`

Structure - dynamics - function: illustration on antibody - antigen complexes

▷ Influenza



▷ (Broadly) neutralizing antibodies



▷ Core questions – illustrated on on IG-Ag complexes

– Binding affinity: geometry (cf lock and key) + dynamics (entropy / free energy)
– Interaction specificity
– Multivalent binding: affinity - avidity - virus entry inhibition

Mining molecular flexibility: novel tools, novel insights

Computational Structural Biology: what is a protein?

Protein functions: structure and dynamics

Computational Structural Biology: challenges

The importance of dynamics

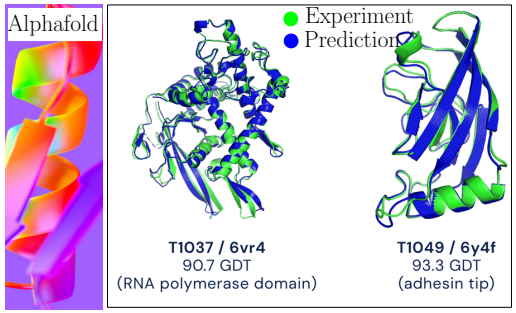
Main difficulties phrased in modern (geometric) terms

Challenge *Structure of proteins*: specification

- Input: sequences from genome sequencing projects

```
P69905 (HBB_HUMAN) MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GS 53
P68871 (HBB_HUMAN) MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN 58
P02144 (MYG_HUMAN) -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKAS 59
                    : *:  :   *   ****         * * *::   * *   *  *  *
```

- Output: plausible structures i.e. atomic coordinates $\{(x_i, y_i, z_i)\}$

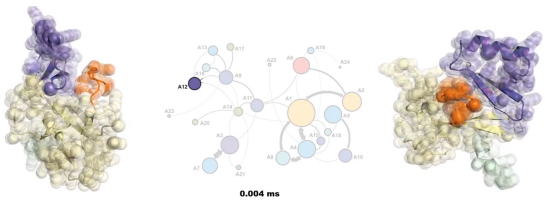


- Protein sequences versus structures: numbers
  - Num. sequences in `UniProtKB`: `TrEMBL` ($\sim 10^8$), `Swiss-Prot` ($\sim 10^5$)
  - Num. structures in the Protein Data Bank: $\sim 150,000$ structures
- Recent & notable: the Deepmind combined approach (DL, optimization)
  - Bias towards well folded structure – no disorder (IDP)
  - Structure only – neither thermodynamics nor kinetics
  - Predicting is not explaining

Challenge *Dynamics of proteins*: specification

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

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



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

Modeling dynamics: shear difficulties

- Three sources of difficulties
  - System size: $3n - 6$ degrees of freedom: typically $> 10^4$
  - Time scales: 15 orders of magnitude
  - Spatial scales: 3 orders of magnitude

Table 1. Characteristic Time Scales for Protein Motions

event	spatial extent (nm)	amplitude (nm)	time (s)	appropriate simulations
bond-length vibration	0.2−0.5	0.001−0.01	10^{-14}−10^{-13}	QM methods
elastic vibration of globular domain	1.0−2.0	0.005−0.05	10^{-12}−10^{-11}	conventional MD
rotation of solvent-exposed side chains	0.5−1.0	0.5−1.0	10^{-11}−10^{-10}	conventional MD
torsional libration of buried groups	0.5−1.0	0.05	10^{-11}−10^{-9}	conventional MD
hinge bending (relative motion of globular domains)	1.0−2.0	0.1−0.5	10^{-11}−10^{-7}	Langevin dynamics, enhanced sampling MD methods?
rotation of buried side chains	0.5	0.5	10^{-4}−1	enhanced sampling MD methods?
allosteric transitions	0.5−4.0	0.1−0.5	10^{-5}−1	enhanced sampling MD methods?
local denaturation	0.5−1.0	0.5−1.0	10^{-5}−10^{1}	enhanced sampling MD methods?
loop motions	1.0−5.0	1.0−5.0	10^{-9}−10^{-5}	Brownian dynamics?
rigid-body (helix) motions		1.0−5.0	10^{-9}−10^{-6}	enhanced sampling MD methods?
helix−coil transitions		>5.0	10^{-7}−10^{4}	enhanced sampling MD methods?
protein association	≫1.0			Brownian dynamics

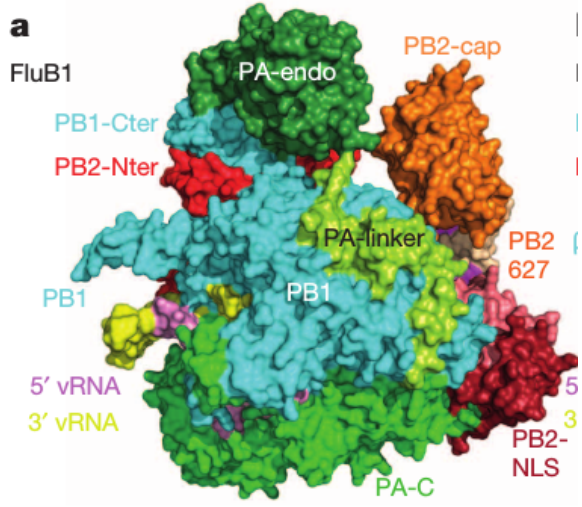
▷Ref: Adcock and McCammon, Chem. Rev., 2006

Challenge *Molecular machines–structure and dynamics*: specification

- ▷ Molecular machines: assemblies with tens / hundreds of subunits
- ▷ Input
  - cryo-electron microscopy (cryo-EM) maps of whole assemblies
  - crystal structures of subunits
  - other data: native mass spectrometry data, . . .
- ▷ Output: structure(s) + mechanism(s)

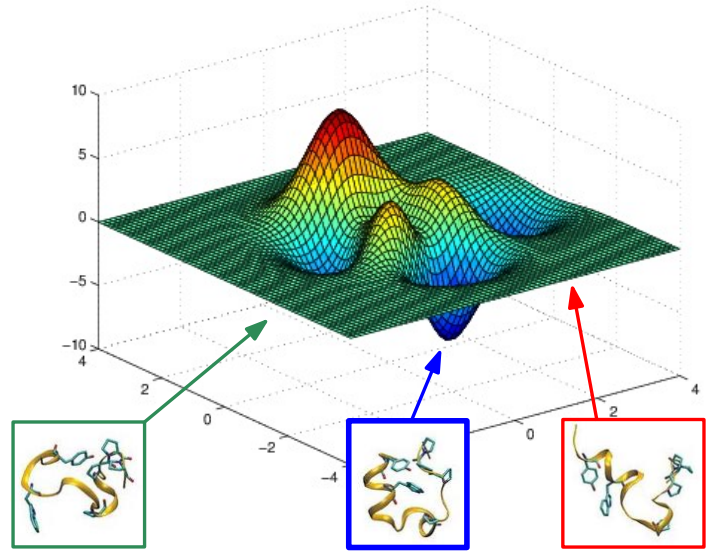
▷ Polymerase of E. coli: structure+dynamics

▷ Polymerase of influenza: structure



▷Ref: Scheres et al, Elife, 2015; ▷Ref: Cusak et al, Nature, 2015

Emergence of macromolecular function(s) from Structure – Thermodynamics – Kinetics

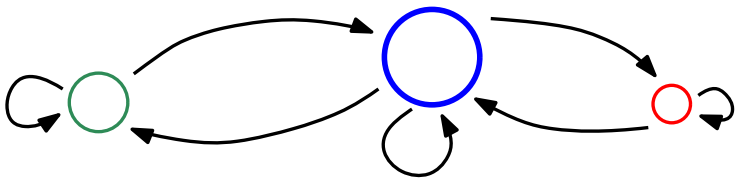


Potential Energy Landscape

- large number of local minima
- enthalpic barriers
- entropic barriers

Structure: stable conformations i.e. local minima of the PEL

Thermodynamics: meta-stable conformations i.e. ensemble of conformations easily inter-convertible into one - another.

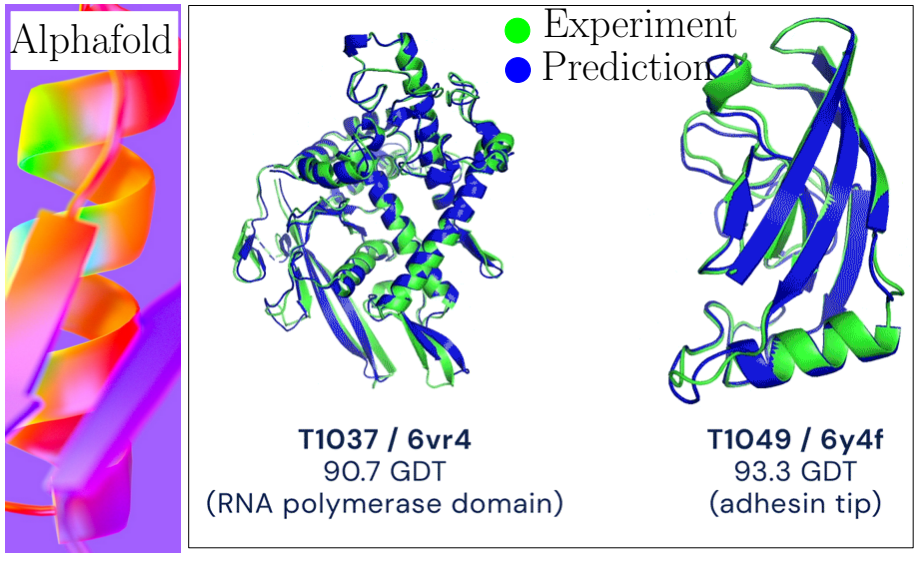


Kinetics: transitions between meta-stable conformations e.g. Markov state model

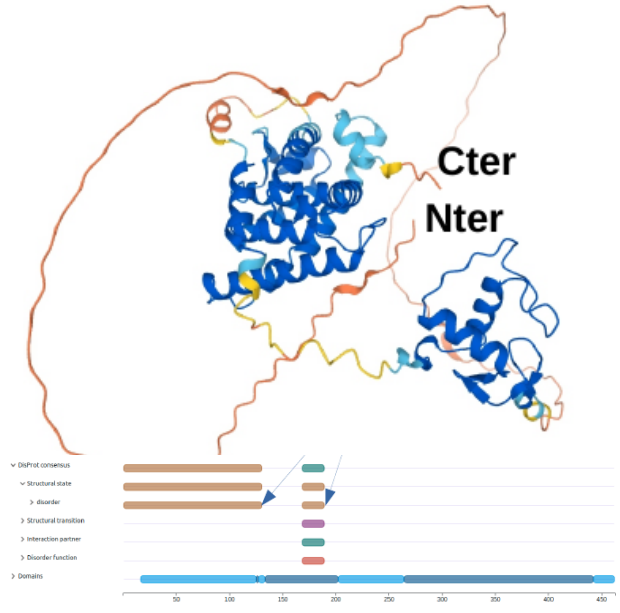
Alphafold by Deepmind

AI: what is being learned?

▷ Successes



▷ . . . and failures



▷ Recent & notable: the Deepmind combined approach (DL, optimization)

- Structure only – neither thermodynamics nor kinetics
- Bias towards well folded structure – no disorder (IDP)
- Predicting is not explaining
- Heavy engineering (team: 34 scientists/engineers)

▷Ref: Jumper et al, Nature, 2021

Mining molecular flexibility: novel tools, novel insights

Computational Structural Biology: what is a protein?

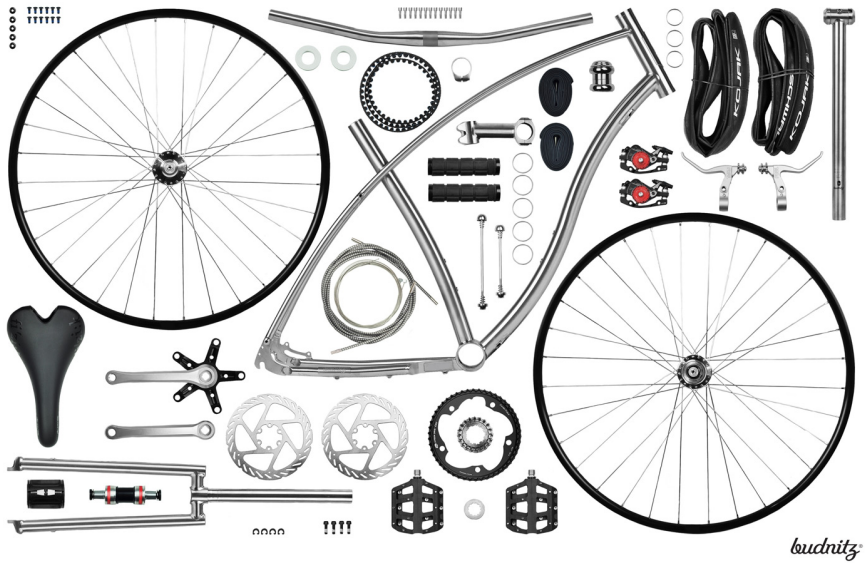
Protein functions: structure and dynamics

Computational Structural Biology: challenges

The importance of dynamics

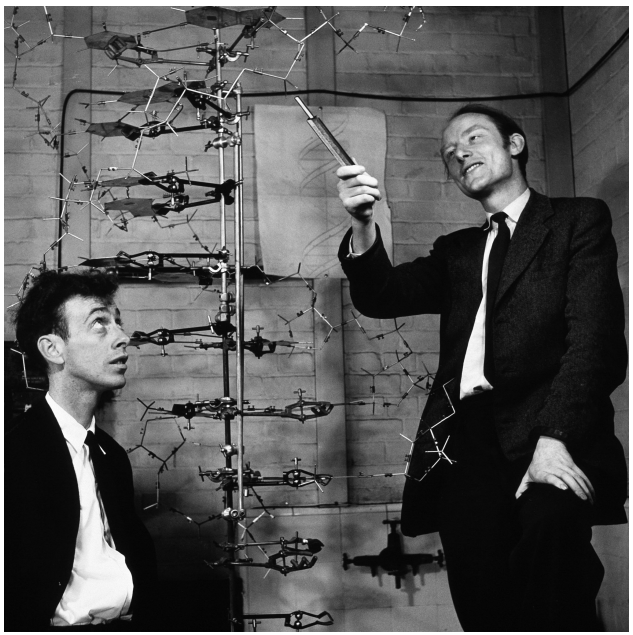
Main difficulties phrased in modern (geometric) terms

Statics vs dynamics



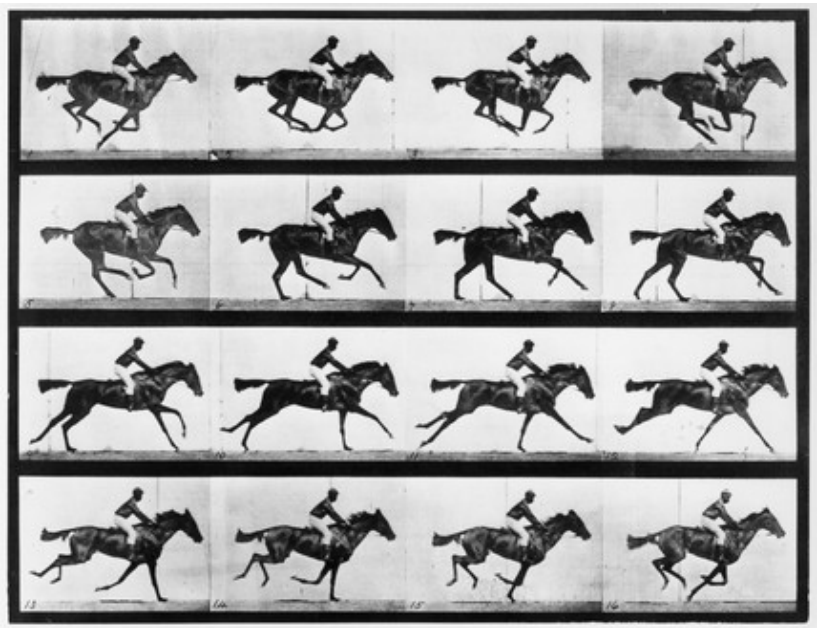
Two schools: static versus dynamic studies

▷ Balls and sticks



(Watson and Crick, DNA model)

▷ The Ballet & time lapse

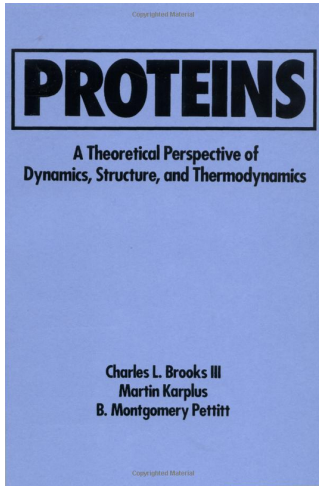


- Static analysis using crystal structure from the Protein Data Bank
  http://rcsb.org
- Dynamical analysis using molecular mechanics

Dynamics of biomolecules: first molecular simulation of a protein

About the simulation duration, quoting M. Levitt "*Cannot remember, but likely less than 100 picoseconds.*
Nb: from the late eighties"

Dynamics: *alea jacta est* in the mid eighties



CONTENTS

▷Ref: Brooks, Karplus, Montgomery Pettitt; Advances in Chemical Physics, Proteins; Wiley, 1988

Mining molecular flexibility: novel tools, novel insights

Computational Structural Biology: what is a protein?

Protein functions: structure and dynamics

Computational Structural Biology: challenges

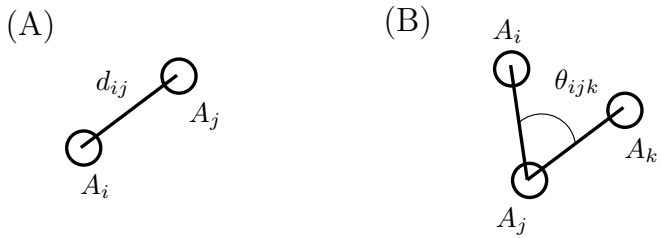
The importance of dynamics

Main difficulties phrased in modern (geometric) terms

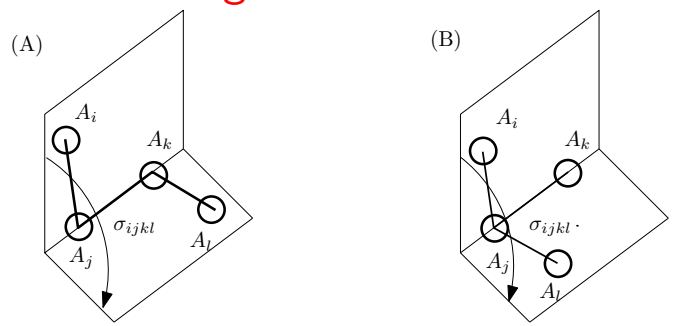
Geometric models: Cartesian and internal coordinates

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

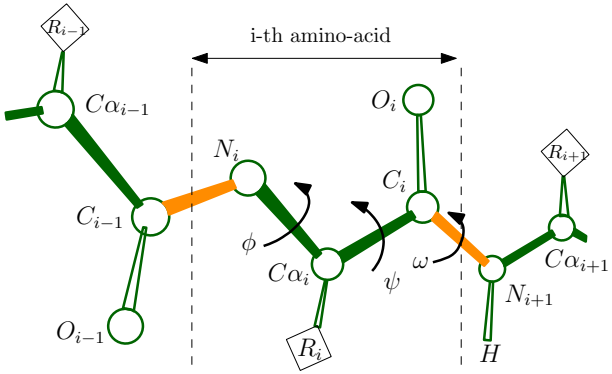
▷ Bond length and valence angle



▷ Dihedral angles



▷ Protein backbone

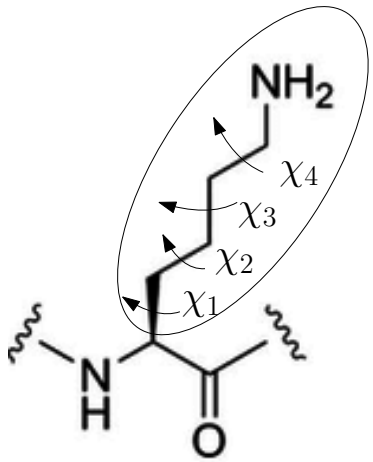


Ramachandran diagram, per a.a. type:

- bivariate distribution for (ϕ, ψ)

▷ Side chain: 20 natural amino acids

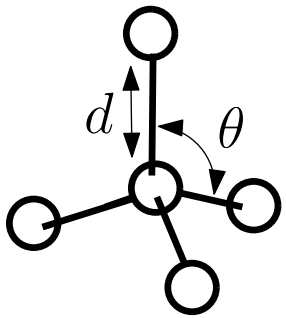
Exple: Lysine, 4 dihedral angles



LYS

The potential energy of (bio-)molecules: force fields

- The $3n-6$ degrees of freedom of a molecule:



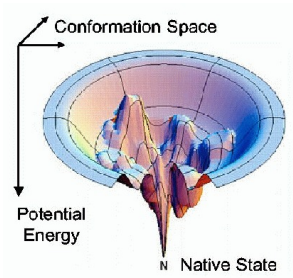
  - types for atoms (element, bonds)
  - covalent: bond lengths, angles
  - non covalent: pairwise distances
  - solvent model

- Potential energy: non linear function

$$V_{\text{total}} = V_{\text{bond}} + V_{\text{angle}} + (V_{\text{proper}} + V_{\text{improper}}) + (V_{\text{vdw}} + V_{\text{electro}}) \qquad (1)$$

V_{bond}: bonds
V_{angle}: covalent angles
V_{proper}: proper dihedrals
V_{improper}: improper dihedrals
V_{vdw}: van der Walls
V_{electro}: electrostatics

- Examples:
  - AMBER: $S_u = (73, 133, 112, 3, 14, 758)$
    1093 unique parameters
  - CHARMM: $S_u = (85, 152, 209, 13, 33, 1)$
    493 unique parameters
  - MARTINI: $S_u = (16, 4, 0, 2, 21, 3)$
    46 unique parameters

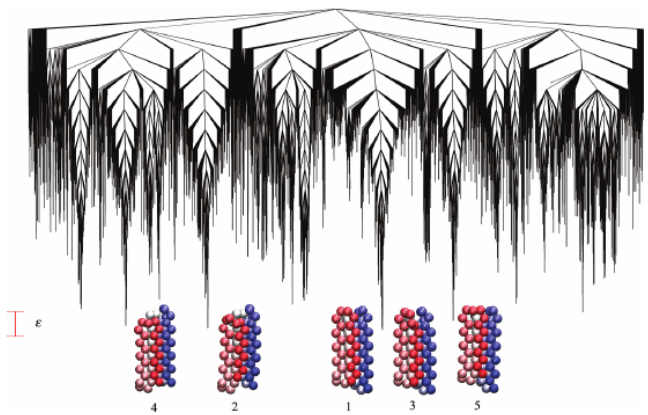


Protein model BLN69: model and force field

- Description:
  - Three types of pseud-atoms a_i: : hydrophobic(B), hydrophylic(L) and neutral(N)
  - Configuration space of intermediate dimension: 207
  - Challenging: frustrated system
  - Exhaustively studied: DB of $\sim 450k$ critical points (Industry)

$$V_{BLN} = \frac{1}{2} \cdot K_r \sum_{i=1}^{N-1} (d_{ij} - R_e)^2 + \frac{1}{2} K_0 \sum_{i=1}^{N-2} (\theta_i - \theta_e)^2 + \epsilon \cdot \sum_{i=1}^{N-3} [A_i(1 + \cos\phi_i) + B_i(1 + 3\cos\phi_i)]$$
$$+4\epsilon \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} \cdot C_{ij}[(\frac{\sigma}{d_{ij}})^{12} - D_{ij}(\frac{\sigma}{d_{ij}})^6]$$

- Disconnectivity graph: describes merge events between basins



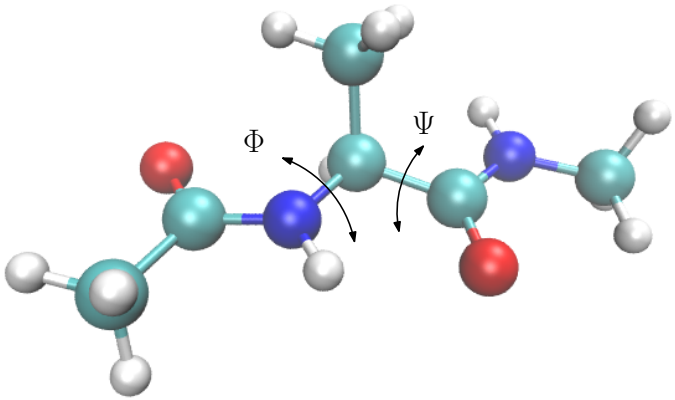
- Ref: Honeycutt, Thirumalai, PNAS, 1990
- Ref: Oakley, Wales, Johnston, J. Phys. Chem., 2011

Thermodynamics

- Quantities defined for a conformation x:
  - potential energy: $V(x)$
  - kinetic energy: $K(x)$
  - total energy: $E(x) = V(x) + K(x)$
  - Boltzmann's distribution: $P^{\text{eq}}(x) = e^{-\beta E(x)}/Z, Z = \sum_{\text{Conformation}x} P^{\text{eq}}(x)$
- Quantities defined for ensembles:
  - Average of observable $\mathcal{O}$ wrt an ensemble: $< \mathcal{O} > \equiv \sum_{\text{Conformation}x} \mathcal{O}(x) P^{\text{eq}}(x)$
  - Exple: average total energy $U =< E >$
  - NVT: Helmholtz free energy $A = U - TS = k_B T \ln Z$
  - NPT: Gibbs free energy $G = U + PV - TS = H - TS$

Density of states and partition functions

Dialanine

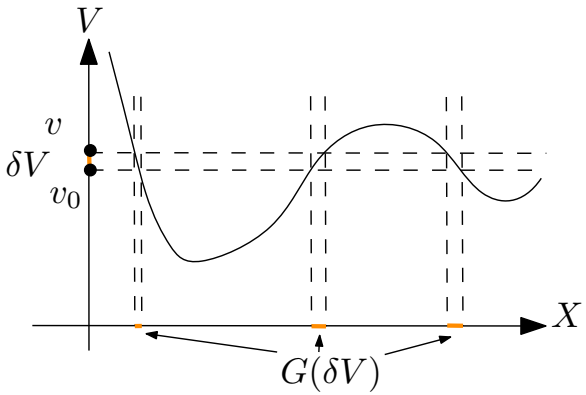


Molecule in water at temperature T

- q: vector of positions of atoms
- Potential energy:

$$V(q)$$

▷ Potential energy landscape:



▷ Density of states (DoS):

- Push forward of the Lebesgue measure by the potential energy V:
- For any $v_0 < v_1$:

$$g([v_0, v_1]) = \int_X 1_{[v_0, v_1]}(V(q))dq$$

▷ Partition function for $A \subset X$: integrate Boltzmann's factor

$$Z_A(T) = \int_{\mathbb{A}} e^{-\beta u} dg(u)$$

▷ NB: n atom: $d = 3n$ Cartesian coordinates. Exple: antibody: $d \sim 42,000$

Key difficulties rephrased

- Key difficulties:
  - Identify low (potential/free) energy conformations
  - Predict observables: thermodynamics / kinetics
- This mini-lecture:
  - Conformational changes in internal coordinates: Tripeptide Loop Closure
  - Thermodynamics and the volume of polytopes

(Open problem) Complexity of Potential Energy Landscape

- Consider a force field of the following type:

$$V_{BLN} = \frac{1}{2} \cdot K_r \sum_{i=1}^{N-1} (R_{i,i+1} - R_e)^2 + \frac{1}{2} K_0 \sum_{i=1}^{N-2} (\theta_i - \theta_e)^2$$
$$+ \epsilon \cdot \sum_{i=1}^{N-3} [A_i(1 + \cos\phi_i) + B_i(1 + 3\cos\phi_i)]$$
$$+ 4\epsilon \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} \cdot C_{ij}[(\frac{\sigma}{R_{i,j}})^{12} - D_{ij}(\frac{\sigma}{R_{i,j}})^6]$$

- Open questions:
  - Number of critical points (local minima, index one saddles)
  - (Topological) Persistence of local minima
  - Geometry of the catchment basins (stable manifolds for $-\nabla V$)
- Rationale:
  - Separation bounds for polynomials
  - Complexity results *à-la* Yomdin-Comte / Tame geometry

Mining molecular flexibility: novel tools, novel insights

PART 1: Introduction to Structural Bioinformatics
PART 2: Protein structure and geometry
PART 3: Thermodynamics and the volume of polytopes
PART 4: Outlook

Mining molecular flexibility: novel tools, novel insights

Tripeptide Loop Closure

TLC: on the quality of solutions

- TLC solutions

Geometric constraints within tripeptides

- TLC steric constraints

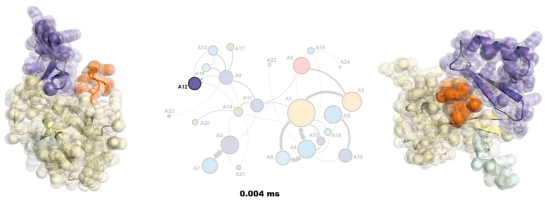
(Uniformly) sampling backbone geometries

- Loop sampling

Challenge *Dynamics of proteins*: specification

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

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

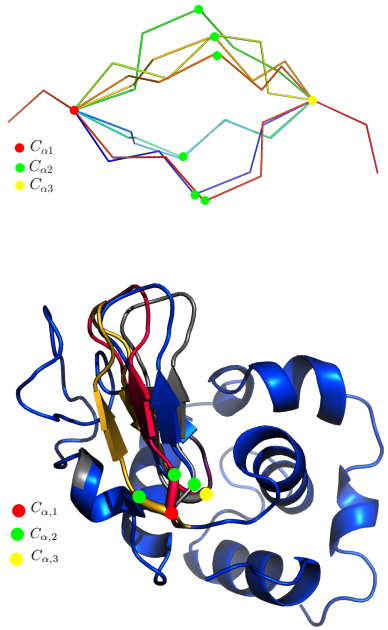


Ref: Chodera et al, eLife, 2019; Youtube link

Overarching goal:

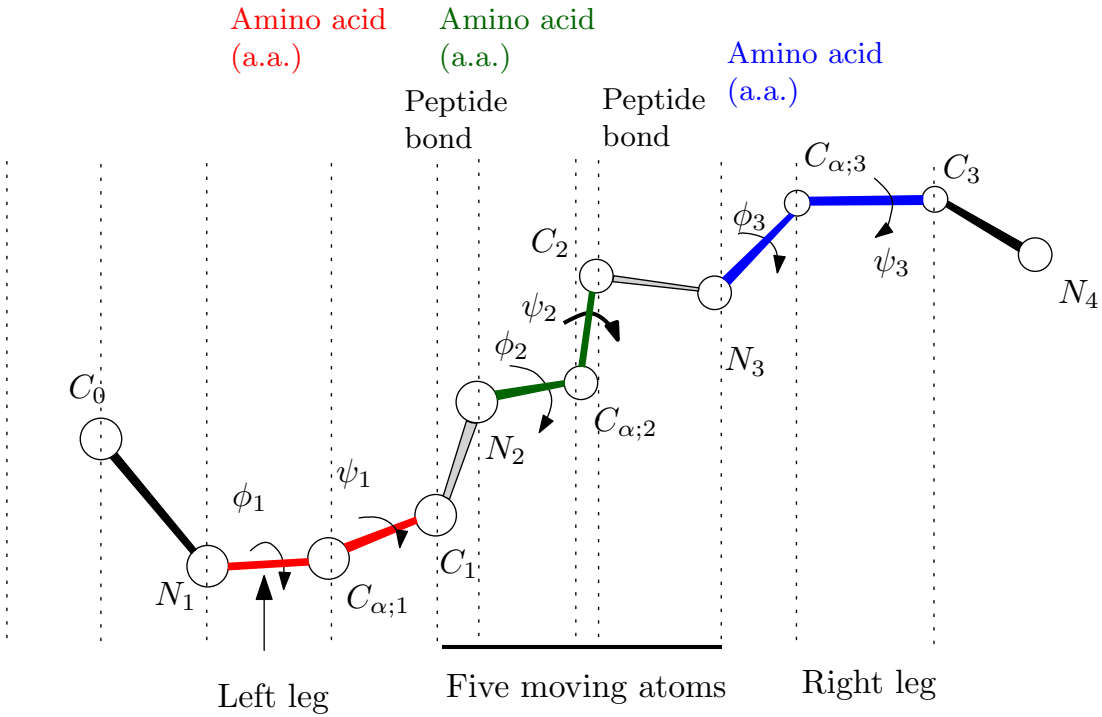
Tired of molecular dynamics? Aim at seven-league boots...

The Tales of Perrault



- Molecular dynamics (Newton's equations of motion): $\Delta t \sim 10^{-15}$ s
- Aim at move sets moving atoms of several angstroms in **one go**, while retaining high quality conformations
- Could result in a gain of several orders of magnitude in simulation time,
- Problem : design the seven-league boots...

Tripeptide: three consecutive amino-acids



▷ Geometry:

- Left+right legs: fixed
- Bond lengths: fixed
- Valence angles: fixed
- ω angles: fixed
- Six $\{\phi, \psi\}$ dihedral angles: free

Overall: 5 moving atoms

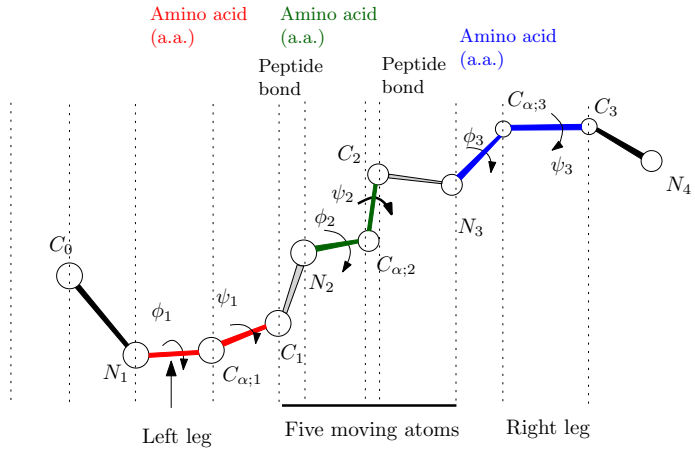
▷ Num. of tripeptide types with the 20 natural amino acids:

- $20 \times 20 \times 20 = 8000$ types

▷ Tripeptides in the Protein Data Bank: $\sim$ 2.6 million tripeptides in high resolution loops, upon filtering out sequence redundancy

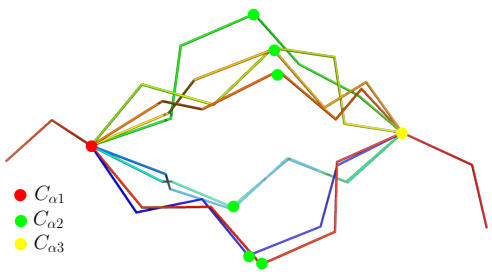
The Tripeptide loop closure – TLC

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

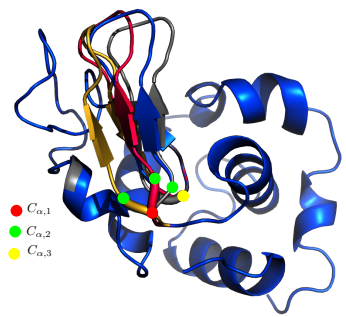


$\Rightarrow$ Find all possible values $(\phi_i, \psi_i)_{i=1,2,3}$ compatible with the fixed internal coordinates

- Theorem: at most 16 solutions



3 consecutive a.a.



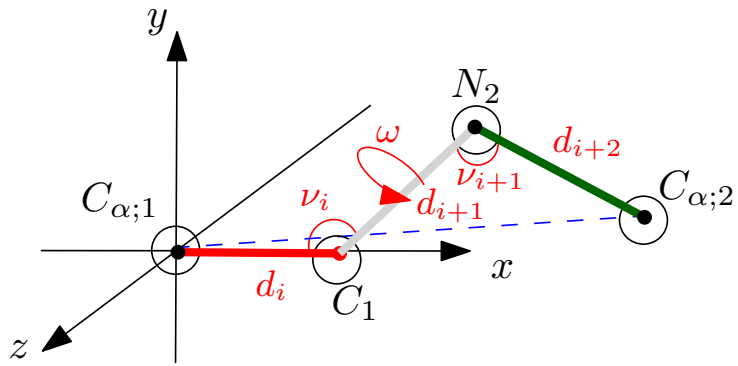
3 a.a. sandwiching SSE–CDRs

▷Ref: Gō and Scheraga, Macromolecules, 1970

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

The peptide bond and peptide rigid bodies

▷ The peptide bond defines a rigid body:

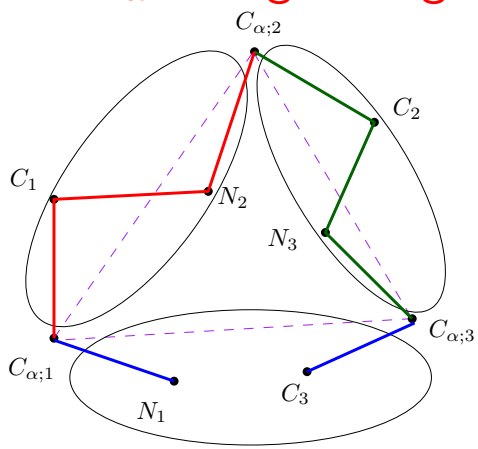


$$||C_{\alpha;2} - C_{\alpha;1}|| = Const$$

Internal coordinates fixed

- Bond lengths
- Valence angles
- ω angle

▷ The C_α triangle is rigid



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

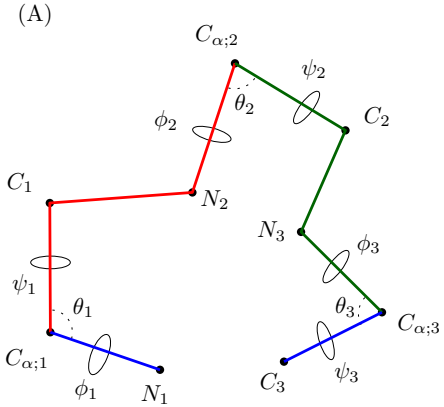
▷ Observations:
- **TLC parameterized in an angular space of dim. 12**
- one solves for six rotation angles $\{(\tau_i, \sigma_i)\}_{i=1,2,3}$

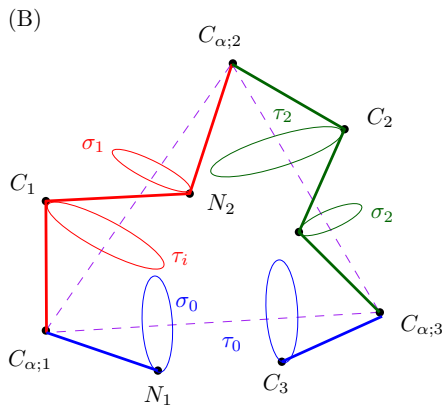
▷Ref: Coutsias et al, J. Comp. Chem., 2004

▷Ref: Cazals et al, Proteins, 2022

TLC model: from six to three angles

▷ Motions of the 3 rigid bodies: 6 angles

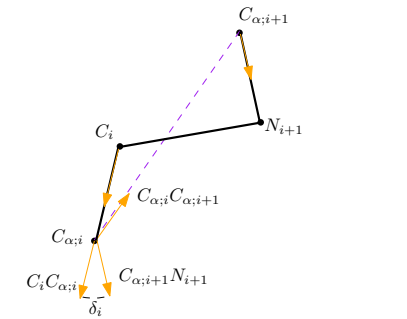




Nb: indices mod(3), e.g., $\sigma_0 = \sigma_3$

▷ . . . which are actually three

$$\sigma_i = \tau_i + \delta_i. \qquad (2)$$



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

▷ Key ingredients of TLC:

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

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

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

▷Ref: Coutsias et al, 2004

The three local frames

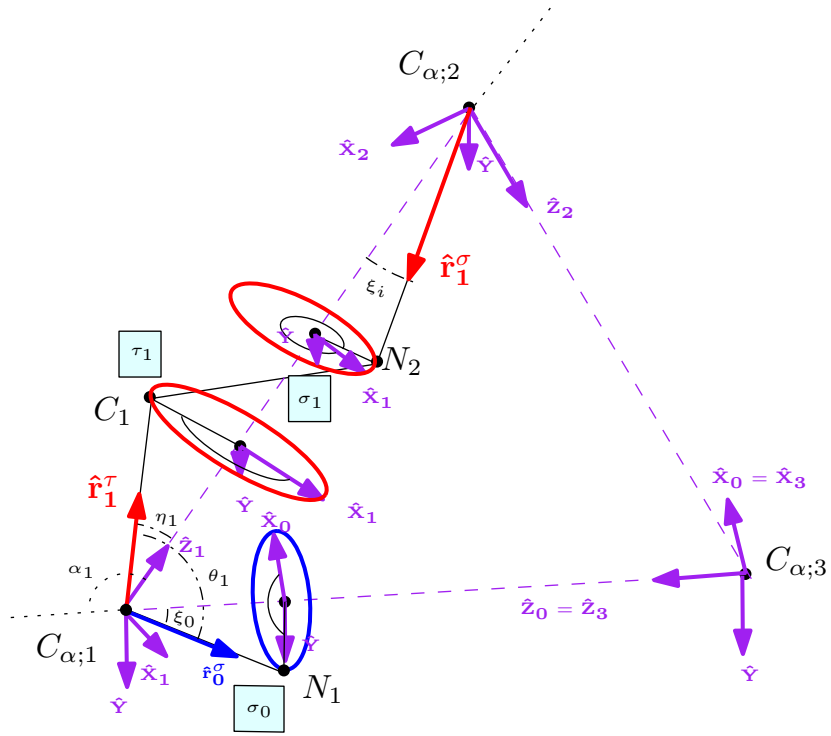
▷ Local frames and individual rotations:

- Orthonormal local frames:

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

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

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



▷ Angular description of the tripeptide:

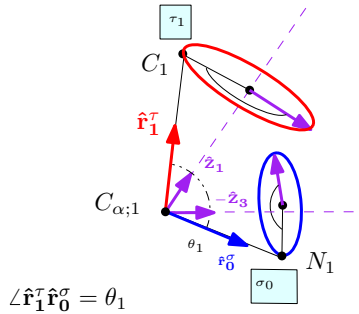
$$\begin{cases} \alpha_i & = \angle \hat{\mathbf{Z}}_\mathbf{i}\hat{\mathbf{Z}}_{\mathbf{i-1}} \\ \xi_i & = \angle - \hat{\mathbf{Z}}_\mathbf{i}\hat{\mathbf{r}}_\mathbf{i}^\sigma \\ \eta_i & = \angle \hat{\mathbf{Z}}_\mathbf{i}\hat{\mathbf{r}}_\mathbf{i}^\tau \\ \delta_i & = \angle \mathsf{Plane}(C_{\alpha;i}C_{\alpha;i+1}C_i), \mathsf{Plane}(C_{\alpha;i}C_{\alpha;i+1}N_{i+1}) \end{cases} \quad (3)$$

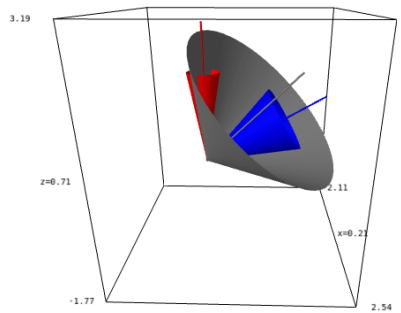
▷ Four tuple of angles for $C_{\alpha;i}$ of tripeptide T_k: $\mathsf{A}_{k,i} = \{\alpha_{i,i}, \eta_{i,i}, \xi_{i,i-1}, \delta_{i,i-1}\}$

Rotations and dot product

▷ Rotations of C_i and N_i: the two cones problem

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





▷ Expressions in local frames:

$$\mathrm{In}(\hat{\mathrm{X}}_{\mathrm{i}-1}, \hat{\mathrm{Y}}, \hat{\mathrm{Z}}_{\mathrm{i}-1}): \quad \hat{\mathrm{r}}^{\sigma}_{\mathrm{i}-1} = -\cos\xi_{i-1}\hat{\mathrm{Z}}_{\mathrm{i}-1} + \sin\xi_{\mathrm{i}-1}(\cos\sigma_{\mathrm{i}-1}\hat{\mathrm{X}}_{\mathrm{i}-1} + \sin\sigma_{\mathrm{i}-1}\hat{\mathrm{Y}}) \quad (4)$$

$$\mathrm{In}(\hat{\mathrm{X}}_{\mathrm{i}}, \hat{\mathrm{Y}}, \hat{\mathrm{Z}}_{\mathrm{i}}): \quad \hat{\mathrm{r}}^{\tau}_{\mathrm{i}} = \cos\eta_i\hat{\mathrm{Z}}_{\mathrm{i}} + \sin\eta_{\mathrm{i}}(\cos\tau_{\mathrm{i}}\hat{\mathrm{X}}_{\mathrm{i}} + \sin\tau_{\mathrm{i}}\hat{\mathrm{Y}}) \quad (5)$$

▷ Valence angle constraint equation: θ_i kept constant

$$\begin{aligned}\langle \hat{\mathrm{r}}^{\sigma}_{\mathrm{i}-1}, \hat{\mathrm{r}}^{\tau}_{\mathrm{i}} \rangle = &-\cos\xi_{i-1}\cos\eta_i\cos\alpha_i \quad (6)\\ &-\cos\xi_{i-1}\sin\eta_i\cos\tau_i\sin\alpha_i\\ &-\cos\eta_i\sin\xi_{i-1}\cos\sigma_{i-1}\sin\alpha_i\\ &+\sin\xi_{i-1}\sin\eta_i(\cos\sigma_{i-1}\cos\tau_i\cos\alpha_i + \sin\sigma_{i-1}\sin\tau_i)\\ &=\cos\theta_i. \quad (7)\end{aligned}$$

Algebra: the degree TLC solutions via the 16 polynomial

▷ Change of variables:

$$u_i = \tan(\tau_i/2), w_i = \tan(\sigma_i/2). \tag{8}$$

▷ Re-write the valence angle constraint – Eq. 19:

$$A_i w_{i-1}^2 u_i^2 + B_i w_{i-1}^2 + C_i w_{i-1} u_i + D_i u_i^2 + E_i = 0, \tag{9}$$

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

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

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

▷ Using resultants: degree 16 polynomial in 1 variable

▷ Nb: the bound it tight.

▷ Robust solutions: requires some care since π is involved

▷Ref: Cox,Little,O'Shea, Using algebraic geometry, 2005

Mining molecular flexibility: novel tools, novel insights

Tripeptide Loop Closure

TLC: on the quality of solutions

- TLC solutions

Geometric constraints within tripeptides

- TLC steric constraints

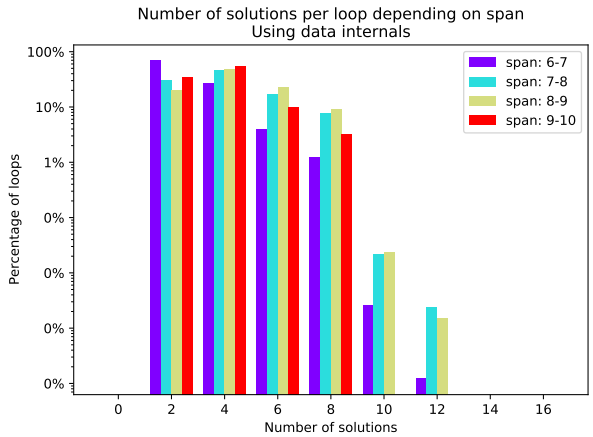
(Uniformly) sampling backbone geometries

- Loop sampling

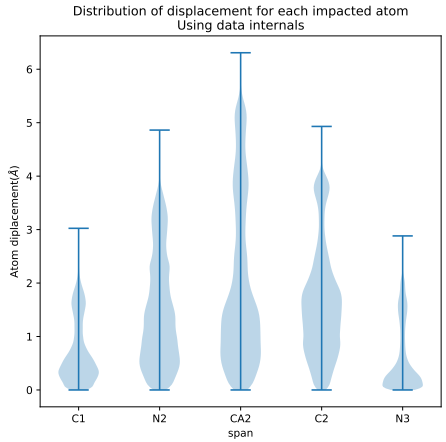
TLC: number of solutions and atomic displacements

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

▷ Number of solutions



▷ Distribution of displacement for the five moving atoms:



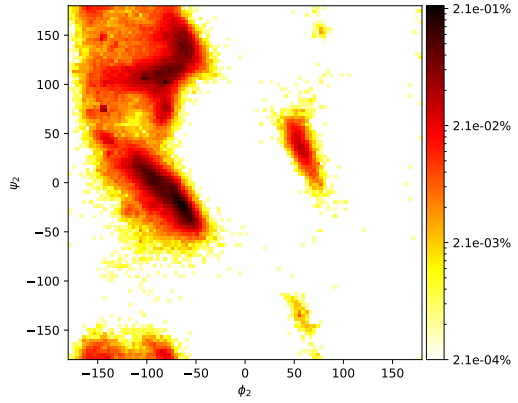
▷ Modeling: **necessary** distance / angular constraints for TLC to admit solutions

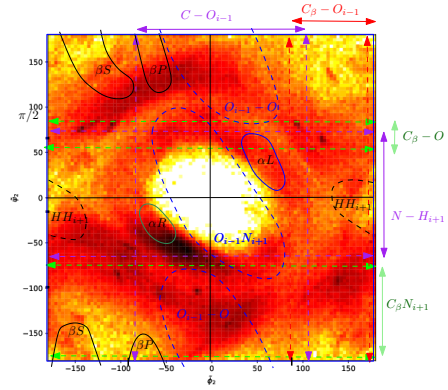
Interpolatory properties of TLC reconstructions

in the 3 Ramachandran domains

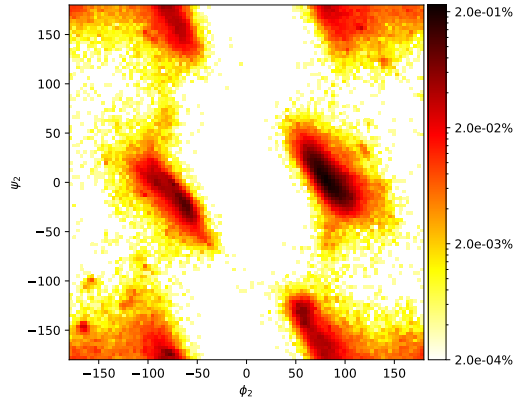
- Method: for the 3 Ramachandran domains (since 3 peptides):
  - compare the distribution of data versus reconstructions
  - distinguish on a per-class amino acid basis
- Ramachandran distributions

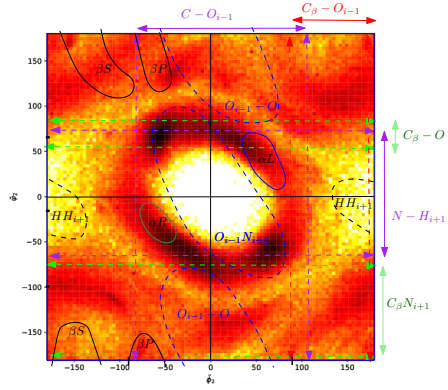
ASP





GLY





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

Mining molecular flexibility: novel tools, novel insights

Tripeptide Loop Closure

TLC: on the quality of solutions

- TLC solutions

Geometric constraints within tripeptides

- TLC steric constraints

(Uniformly) sampling backbone geometries

- Loop sampling

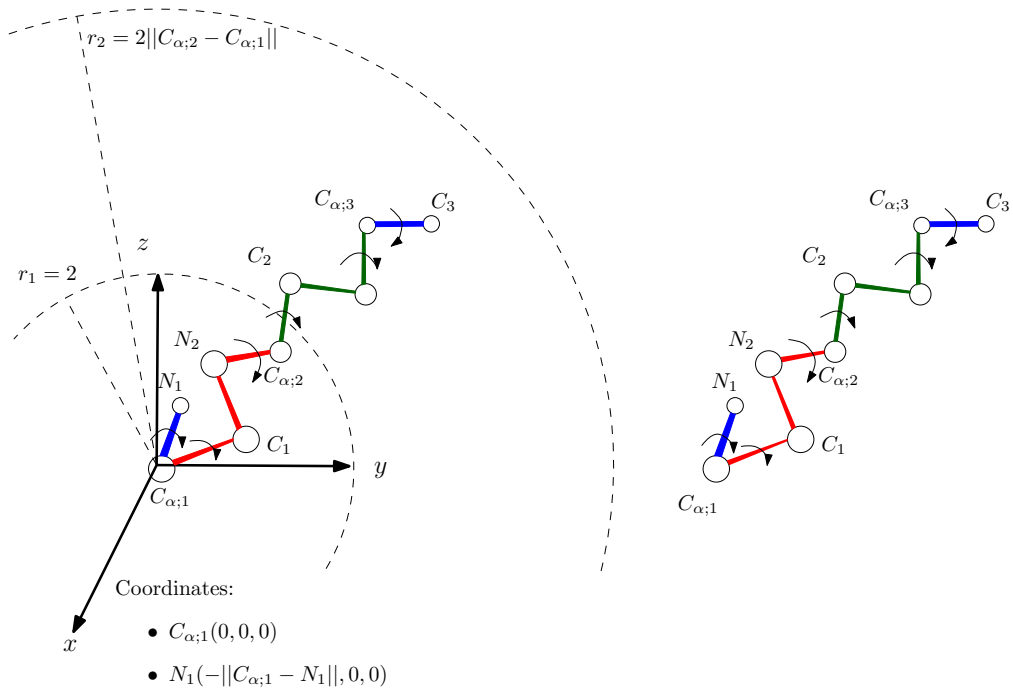
TLC with moving legs and embeddable tripeptides

▷ Geometric model:

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

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

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



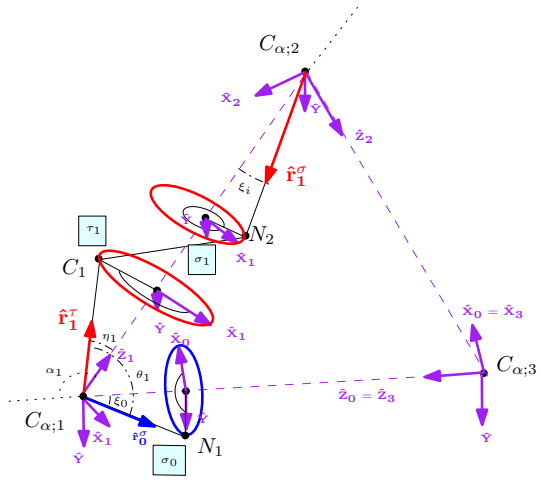
Embedding tripeptides: recap

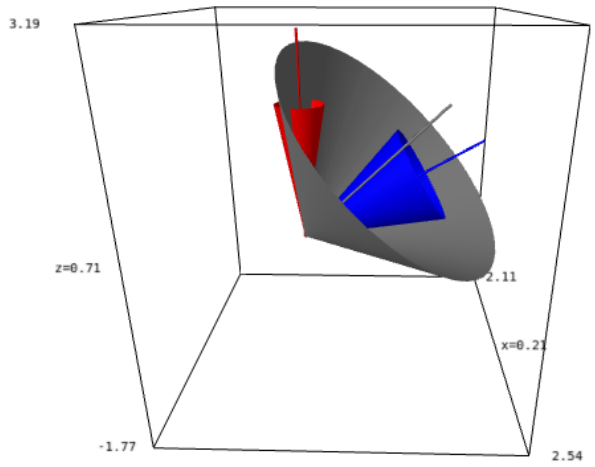
- Orthonormal local frames:

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

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

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





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

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

$$\langle \hat{\mathbf{r}}_{\mathbf{i-1}}^\sigma, \hat{\mathbf{r}}_\mathbf{i}^\tau \rangle = \cos\theta_i. \tag{10}$$

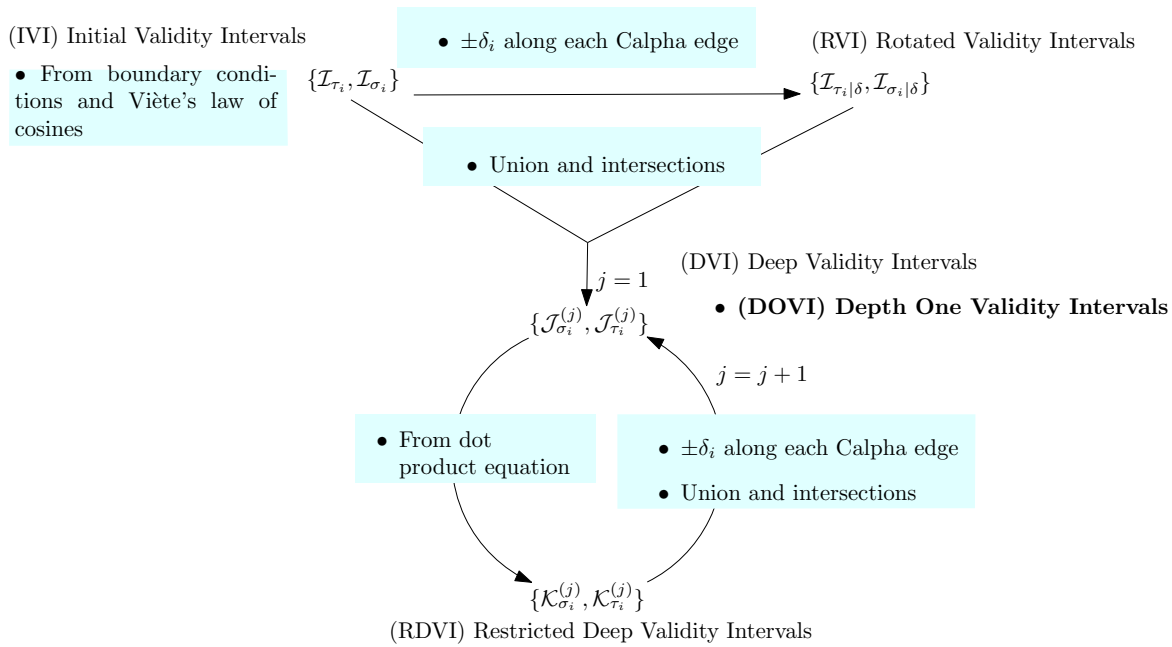
▷ Our goal:

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

Strategy: validity intervals – outline

- Two types of constraints:
  - Coherence along each edge of the C_α triangle – via ω angle
  - Constraint on θ_i at each C_α

- A sequential and iterative construction: interval types used
  - Initial VI
  - Rotated VI
  - Deep VI and Restricted Deep VI



Strategy: validity intervals – details

- We wish to define a *validity intervals*:

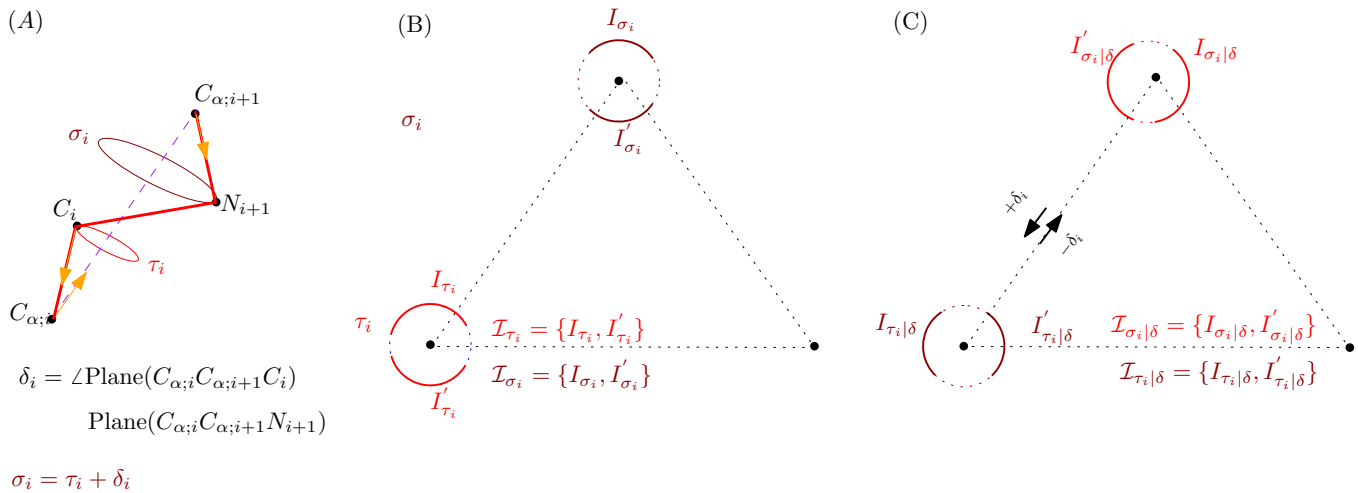
$$\sigma_{i-1} : I_{\sigma_{i-1}} = [\sigma_{i-1;-}, \sigma_{i-1;+}] \subset [0, \pi]$$
$$\tau_i : I_{\tau_i} = [\tau_{i;-}, \tau_{i;+}] \subset [0, \pi].$$

such that $\sigma_{i-1} \in I_{\sigma_{i-1}}$ and $\tau_i \in I_{\tau_i}$ are **tight** necessary conditions

- Taken for granted: Initial Validity Intervals

$$\mathcal{I}_{\tau_i} = \{I_{\tau_i}, I'_{\tau_i}\} \text{ and } \mathcal{I}_{\sigma_{i-1}} = \{I_{\sigma_{i-1}}, I'_{\sigma_{i-1}}\}.$$

- From Initial Validity Intervals to Rotated Validity Intervals:



- Using the coupling $\sigma_i = \tau_i + \delta_i$: Depth One Validity Intervals (DOVI) for σ_{i-1}:

$$\mathcal{J}^{(1)}_{\sigma_{i-1}} = (I_{\sigma_{i-1}} \cap I_{\sigma_{i-1}|\delta}) \cup (I_{\sigma_{i-1}} \cap I'_{\sigma_{i-1}|\delta}) \cup (I'_{\sigma_{i-1}} \cap I_{\sigma_{i-1}|\delta}) \cup (I'_{\sigma_{i-1}} \cap I'_{\sigma_{i-1}|\delta}) \quad (11)$$

For τ_i: mutatis mutandis.

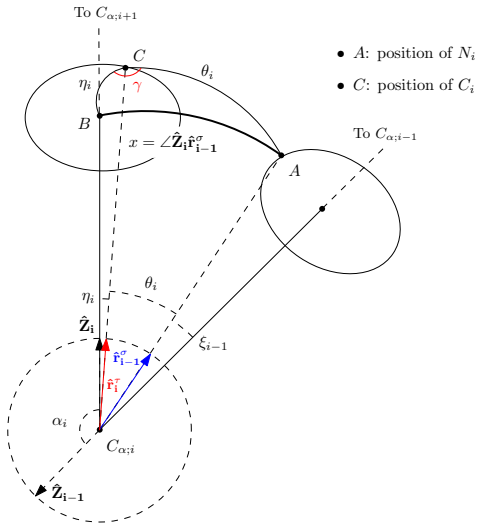
Initial Validity Intervals: bounds

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

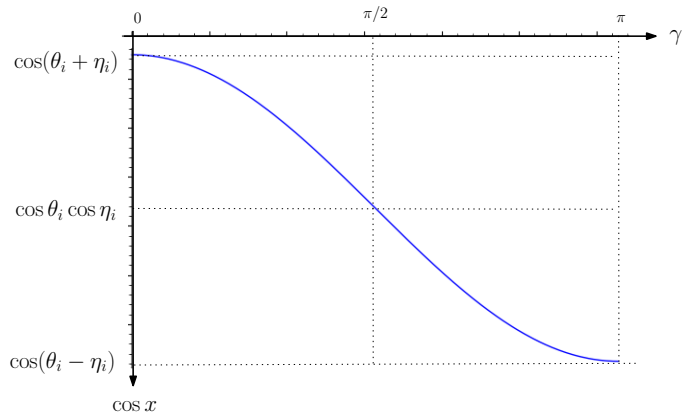
Proof: Viète's law of cosines for the spherical triangle ABC:

$$\cos x = \cos\theta_i \cos\eta_i + \sin\theta_i \sin\eta_i \cos\gamma. \tag{12}$$

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



Viète formula of cosines



$\cos x$

▷ Final step:

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

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

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

$$\langle \hat{\mathbf{r}}^{\sigma}_{\mathbf{i-1}}, \hat{\mathbf{Z}}_{\mathbf{i}} \rangle = -\cos\sigma_{\mathbf{i-1}} \sin\xi_{\mathbf{i-1}} \sin\alpha_{\mathbf{i}} - \cos\xi_{\mathbf{i-1}} \cos\alpha_{\mathbf{i}}. \tag{13}$$

$$\langle \hat{\mathbf{r}}^{\sigma}_{\mathbf{i-1;-}}, \hat{\mathbf{Z}}_{\mathbf{i}} \rangle = \cos(\theta_{\mathbf{i}} + \eta_{\mathbf{i}}) \tag{14}$$

from which we obtain

$$\begin{cases} S^{-} = \frac{+\cos(\theta_i - \eta_i) + \cos\xi_{i-\mathbf{1}} \cos\alpha_i}{\sin\xi_{i-\mathbf{1}} \sin\alpha_i} \\ \sigma_{i-\mathbf{1};-} = \arccos S^{-} \end{cases} \tag{15}$$

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

$$S^{-} > 1, \tag{16}$$

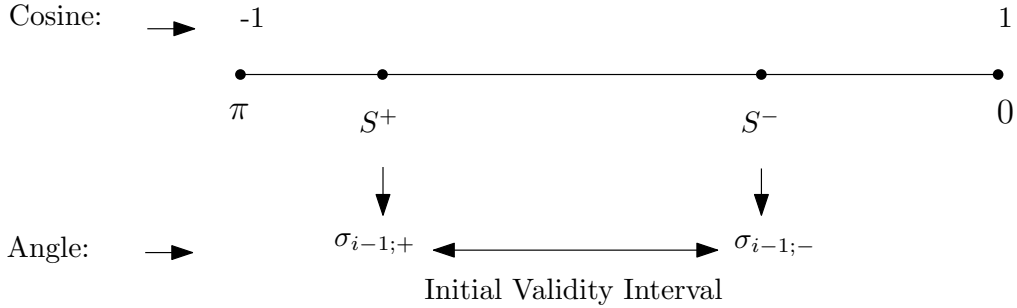
we set $\sigma_{i-\mathbf{1};-} = 0$, so that any value $\sigma_{i-1} \leq \sigma_{i-\mathbf{1};+}$ is valid.

▷ $\sigma_{i-\mathbf{1};+}$: mutatis mutandis

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

C_α valence constraints

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



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

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

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

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

Validity Intervals: Initial and Symmetric

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

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

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

- Angle τ_i: mutatis mutandis

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

$$\mathcal{I}_{\sigma_{i-1}} = I_{\sigma_{i-1}} \cup I'_{\sigma_{i-1}} \tag{17}$$

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

$$\mathcal{I}_{\tau_i} = I_{\tau_{i-1}} \cup I'_{\tau_i}. \tag{18}$$

Extreme angles: illustrations

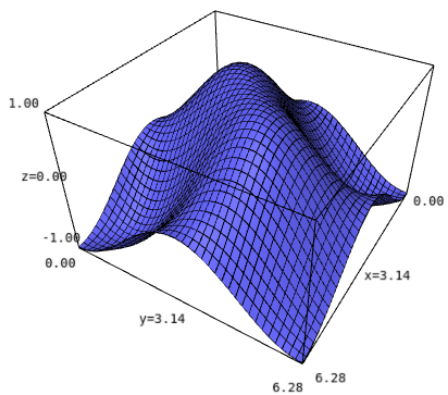
- Dot product surface:

$$f(\sigma_{i-1}, \tau_i) = \langle \hat{\mathbf{r}}_{i-1}^{\sigma}, \hat{\mathbf{r}}_i^{\tau} \rangle \tag{19}$$

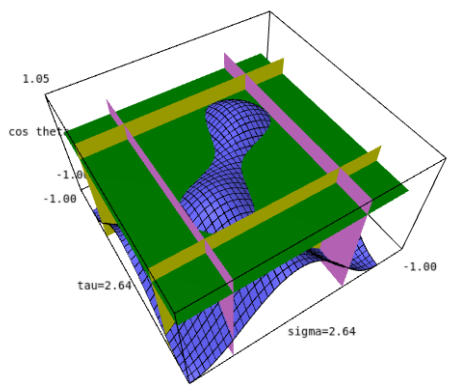
$$= -\cos\xi_{i-1}\cos\eta_i\cos\alpha_i \tag{20}$$
$$-\cos\xi_{i-1}\sin\eta_i\cos\tau_i\sin\alpha_i$$
$$-\cos\eta_i\sin\xi_{i-1}\cos\sigma_{i-1}\sin\alpha_i$$
$$+\sin\xi_{i-1}\sin\eta_i(\cos\sigma_{i-1}\cos\tau_i\cos\alpha_i + \sin\sigma_{i-1}\sin\tau_i)$$

$$= \cos\theta_i \tag{21}$$

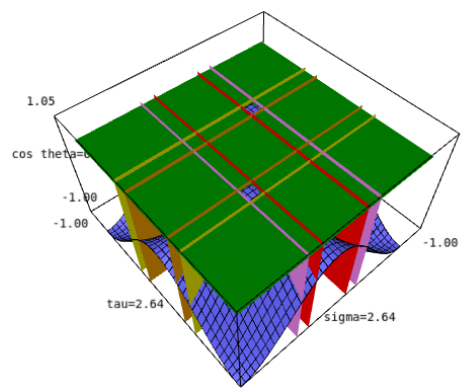
  - angles $\sigma_{i-1;-}$ and $\sigma_{i-1;+}$ correspond to planes orthogonal to the σ_{i-1}; dito for $\tau_{i;-}$ and $\tau_{i;+}$
- Dot product surface and extreme angles $\sigma_{i-1;-}, \sigma_{i-1;+}, \tau_{i-1;-}, \tau_{i-1;+}$



(A)



(B)



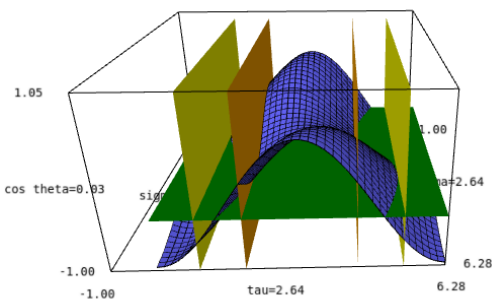
(C)

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

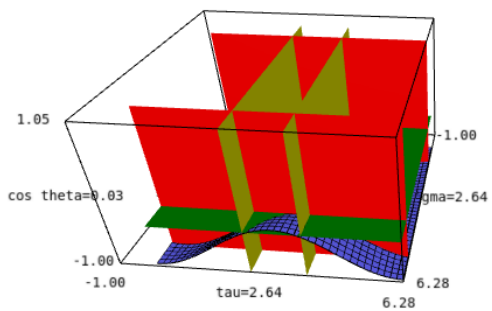
Dot surfaces and their classification

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

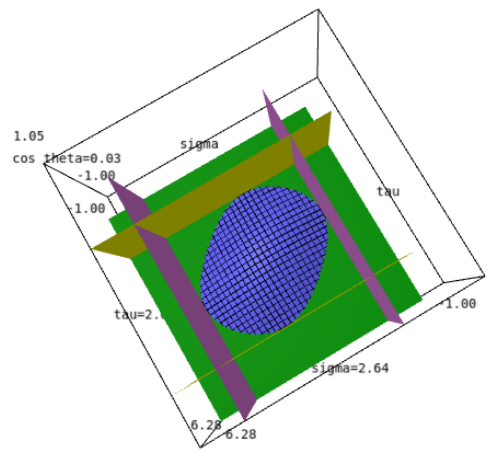
- letter N for $cos(endpoint) < -1$,
- letter P for $cos(endpoint) > 1$,
- letter Z for $-1 < cos(endpoint) < 1$.



PNZZ



PZZN



ZNZN

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

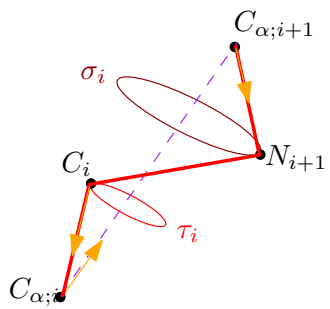
Rotated validity intervals (I)

- Along C_α edge:

$$\sigma_i = \tau_i + \delta_i. \tag{22}$$

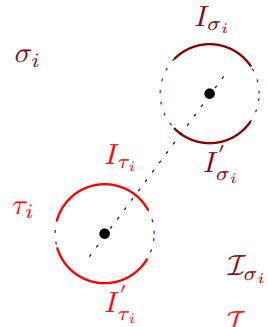
- Rotated interval for an angle: obtained from the value of its twin angle (from τ_i for σ_i, and vice-versa)

(A)



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

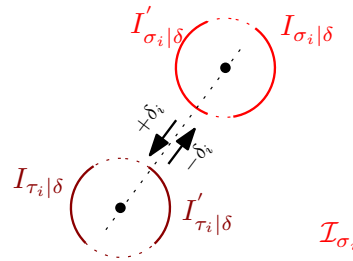
(B)



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

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

(C)



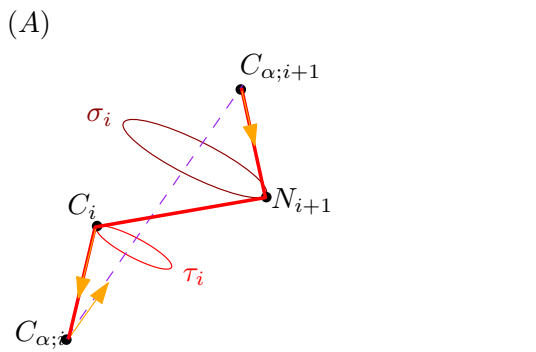
$\mathcal{I}_{\sigma_i|\delta} = \{I_{\sigma_i|\delta}, I'_{\sigma_i|\delta}\}$

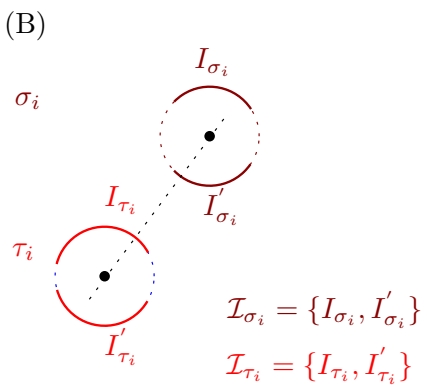
$\mathcal{I}_{\tau_i|\delta} = \{I_{\tau_i|\delta}, I'_{\tau_i|\delta}\}$

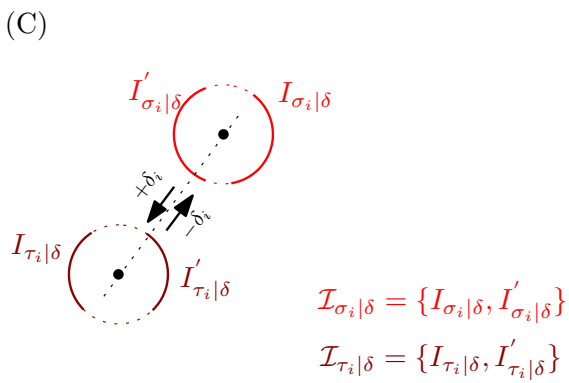
Rotated validity intervals (II)

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

- for σ_{i-1}: $\mathcal{I}_{\sigma_{i-1}|\delta} = I_{\sigma_{i-1}|\delta} \cup I'_{\sigma_{i-1}|\delta}$ with:
  - $I_{\sigma_{i-1}|\delta}$: interval for σ_{i-1} obtained by applying Eq. (22) 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. (22) to $I'_{\tau_{i-1}}$. (Nb: uses the edge $C_{\alpha;i}C_{\alpha;i-1}$ of the C_α triangle.)
- for τ_i: dito







Deep Validity Intervals: depth 1

- ▷ Intervals obtained so far:
  - The conditions on σ_{i-1} and τ_i inherent to the conservation of the valence angles (Eq. (25)).
  - The conditions exploiting rotated validity intervals, stemming from Eq. (22)
- ▷ Combination: intervals combined as follows $(I_{\cdot_{i-1}}, I'_{\cdot_{i-1}}) \times (I_{\cdot_{i-1}|\delta}, I'_{\cdot_{i-1}|\delta})$, which yields *depth one validity intervals*:

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

$$\mathcal{J}^{(1)}_{\sigma_{i-1}} = (I_{\sigma_{i-1}} \cap I_{\sigma_{i-1}|\delta}) \cup (I_{\sigma_{i-1}} \cap I'_{\sigma_{i-1}|\delta}) \cup (I'_{\sigma_{i-1}} \cap I_{\sigma_{i-1}|\delta}) \cup (I'_{\sigma_{i-1}} \cap I'_{\sigma_{i-1}|\delta}) \quad (23)$$

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

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

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

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

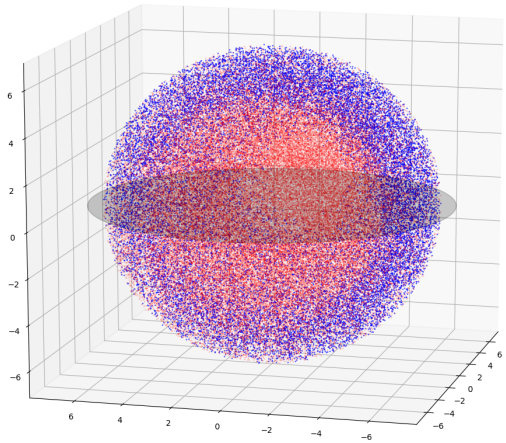
Stringency of necessary conditions: assessment

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

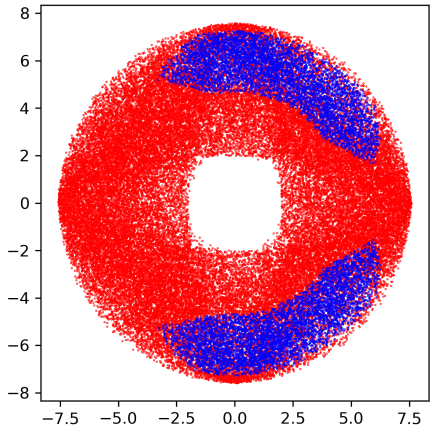
$$\text{Strigency of necessary condition } c : \frac{FP(c)}{N} \tag{24}$$

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

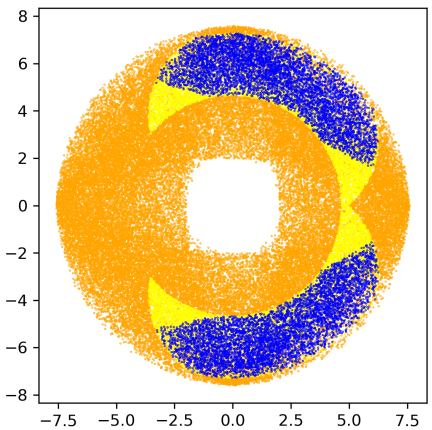
Stringency of necessary conditions: results



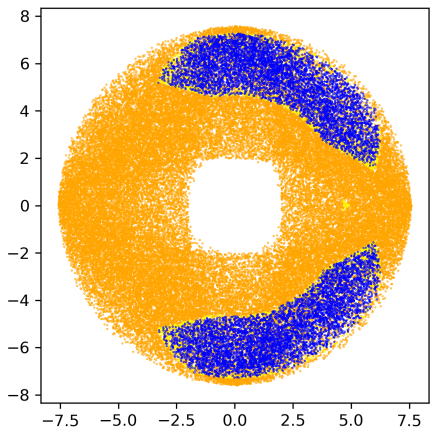
(A)



(B)



(C)



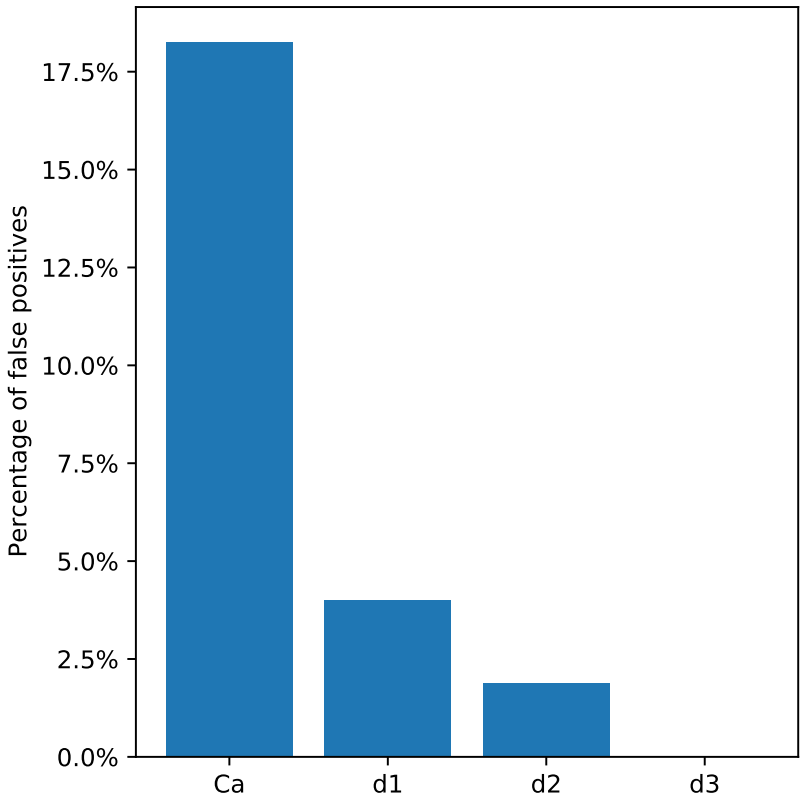
(D)

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

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

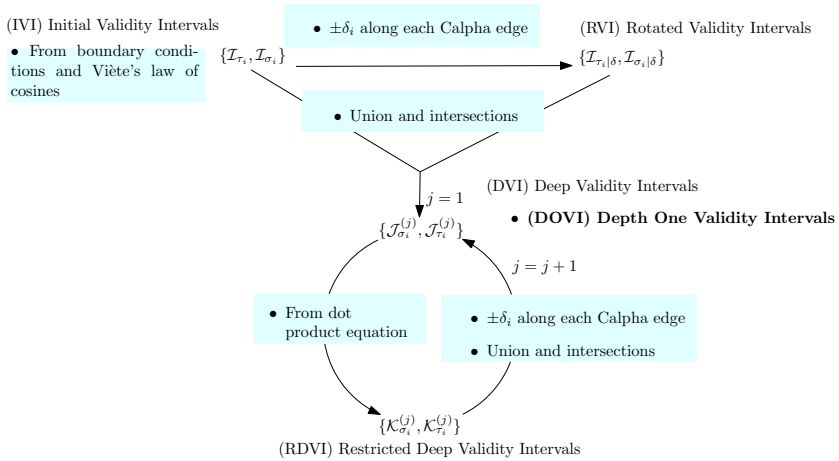
Stringency: statistics

- Stringency: initial validity intervals + deep validity intervals



- Take Home Msg: FP remain

Depth-n validity constraints: outline



▷ Depth 1 validity intervals:

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

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

- Then refinement thanks to intersections with Rotated validity intervals

▷ Depth-n validity intervals:

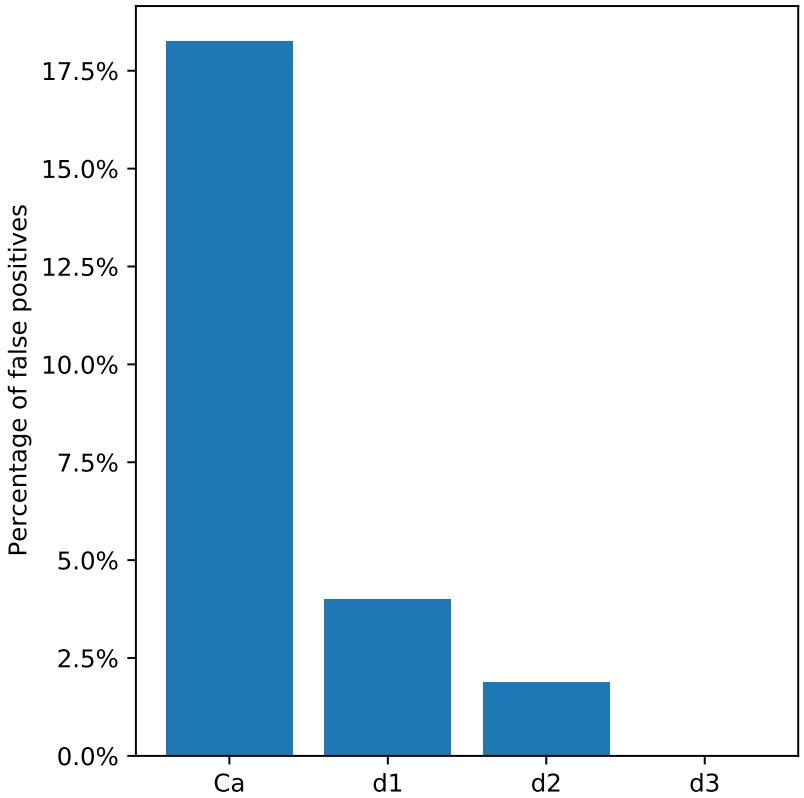
- Given a DVI of depth j (initially, $j = 1$), apply the valence angle constraint to obtain the twin interval on τ_i from σ_{i-1} and vice-versa, using

$$\langle \hat{\mathbf{r}}^{\sigma}_{\mathbf{i}-\mathbf{1}}, \hat{\mathbf{r}}^{\tau}_{\mathbf{i}} \rangle = \cos\theta_i. \tag{25}$$

- Iterate

Stringency: statistics

- Stringency: initial validity intervals + deep validity intervals



- Take Home Msg: (all?) FP filtered out

Mining molecular flexibility: novel tools, novel insights

Tripeptide Loop Closure

TLC: on the quality of solutions

- TLC solutions

Geometric constraints within tripeptides

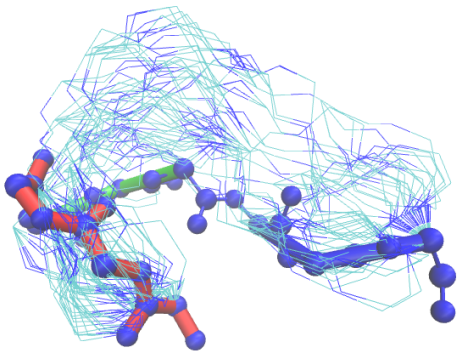
- TLC steric constraints

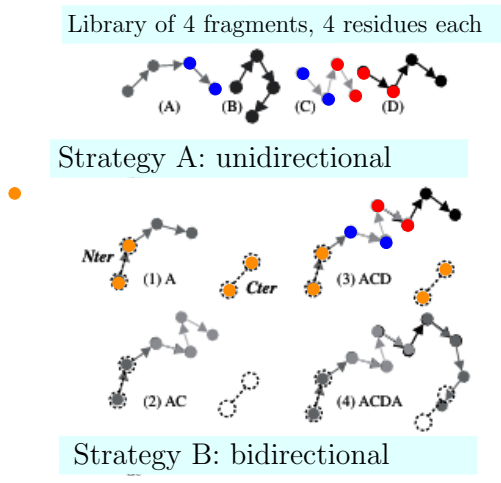
(Uniformly) sampling backbone geometries

- Loop sampling

Protein Loop Sampling: main approaches

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





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

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

▷Ref: Cazals et al; 2022

Loop sampling: difficulties and main approaches

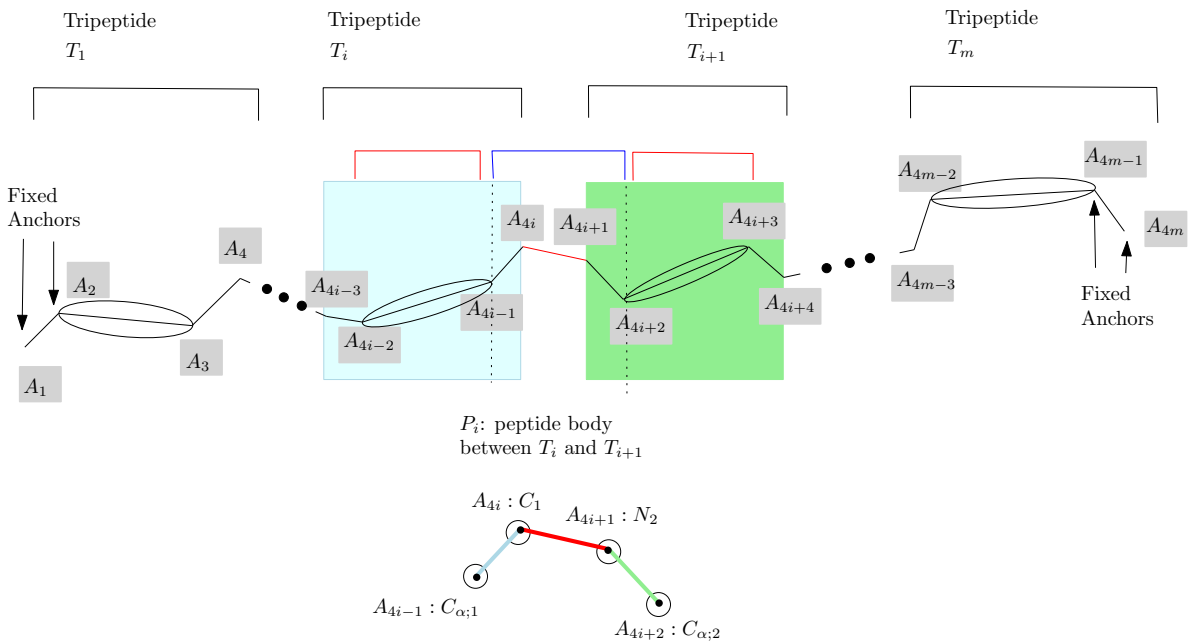
- Main difficulties
  - Space of solutions: a continuous space – if #dihedral angles > 6
  - *Walking* on this constrained *manifold*: geometrically/numerically difficult
  - Incremental construction based on tripeptides: combinatorial explosion
- A mixed discrete - continuous approach
  - `Rosetta`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

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

Global geometric model

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

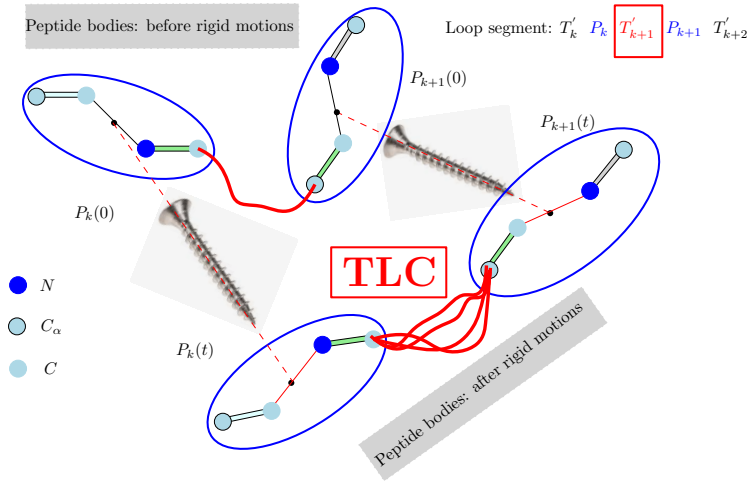
$$L = P_0 \; T_1^{'} \; P_1 \; \ldots \; P_{k-1} \; T_k^{'} \; P_k \; \ldots \; P_{m-1} T_m^{'} P_m. \tag{26}$$



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

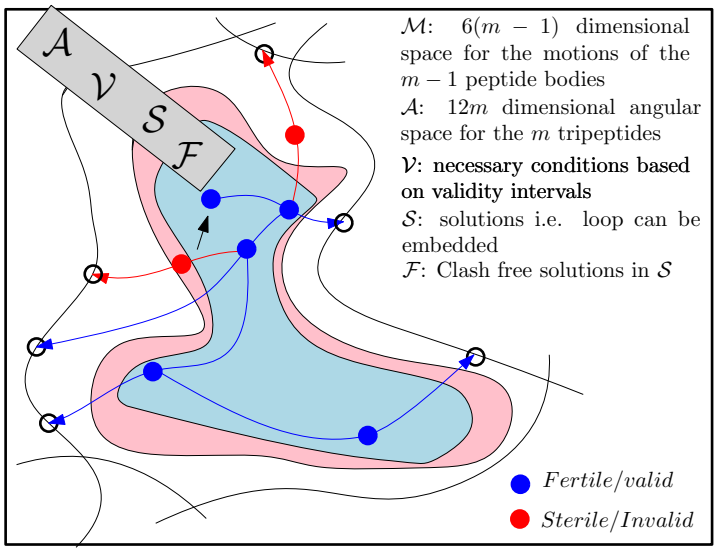
Loop sampling: spaces involved and solution sketch

- Loop decomposition into: rigid peptide bodies and tripeptides cores



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

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



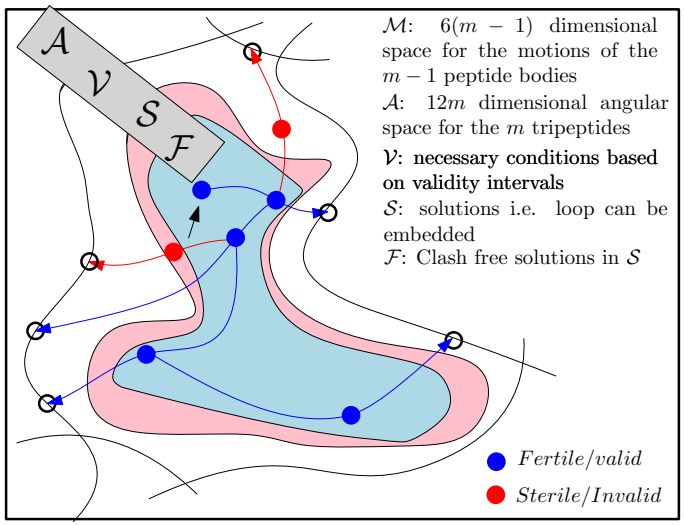
- Aim: perform rejection sampling in a region $\mathcal{V}$ containing all valid loop geometries.
- How: with Hit-and-Run in a domain characterizing necessary conditions – cf validity intervals

Loop sampling: spaces involved and solution sketch

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

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

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



Angular representations: tripeptide and loop

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

Definition 7. Let $\mathrm{A}_{k,i} = \{\alpha_{k,i}, \eta_{k,i}, \xi_{k,i-1}, \delta_{k,i-1}\}$ be the set of angles associated with $C_{\alpha;i}$ in the k-th tripeptide T_k.
The *angular representation* of a tripeptide T_k is the 12-tuple $\mathrm{A}_k = \{\mathrm{A}_{k,1}, \mathrm{A}_{k,2}, \mathrm{A}_{k,3}\}$.
The corresponding 12-dimensional space is denoted $\mathcal{A}_k$.

Definition 8. (Angular conformational space $\mathcal{A}$) The *angular conformational space* of the loop L is the $12m$ dimensional space defined by the product of the m angular space of the individual tripeptides:

$$\mathcal{A} \stackrel{Def}{=} \prod_{k=1}^{m} \mathcal{A}_k. \tag{27}$$

Validity Intervals and Depth One Validity Intervals (DOVI)

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

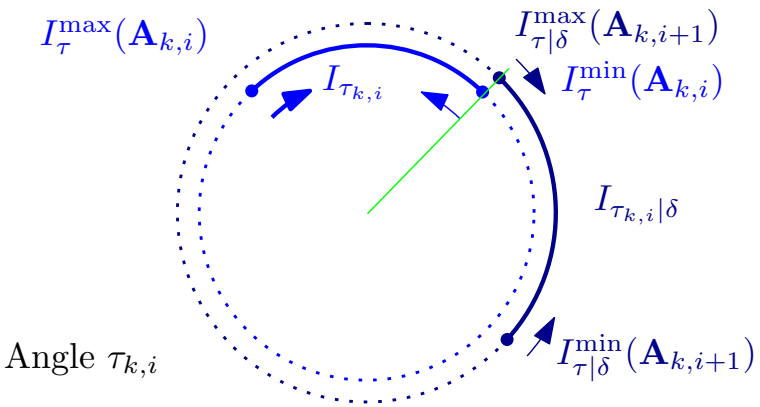
$$\begin{cases} \mathcal{I}_{\tau_{k,i}} = \{I_{\tau_{k,i}}\} \text{ with } I_{\tau_{k,i}} = [I_\tau^{\min}(\mathrm{A}_{k,i}), I_\tau^{\max}(\mathrm{A}_{k,i})] \\ \mathcal{I}_{\tau_{k,i}|\delta} = \{I_{\tau_{k,i}|\delta}\} \text{ with } I_{\tau_{k,i}|\delta} = [I_{\tau|\delta}^{\min}(\mathrm{A}_{k,i+1}), I_{\tau|\delta}^{\max}(\mathrm{A}_{k,i+1})] \end{cases} \tag{28}$$

Indeed:

- $I_{\tau_{k,i}}$: obtained from the conservation of the valence angle at $C_{\alpha;i}$
- $I_{\tau_{k,i}|\delta}$: obtained from $I_{\sigma_{k,i}}$ via the relation $\sigma_i = \tau_i + \delta_i$

▷ Intersection of validity intervals: necessary conditions expressed as intervals

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



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

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

Validity domain for tripeptide T_k and the whole chain L

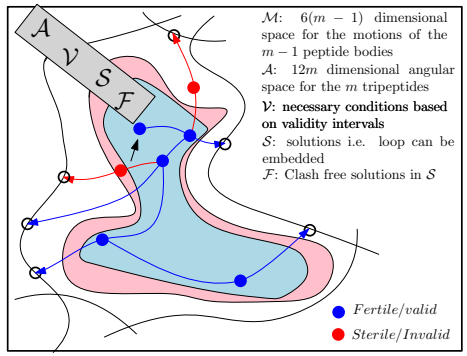
▷ Rigid body I of tripeptide T_k: angles tuples $\rightsquigarrow$ Depth One Validity Intervals

$$\mathrm{DOVI}_{\tau_{k,i}}(\cdot) : \mathcal{A}_k \mapsto \emptyset + (\mathcal{I}_{\tau_{k,i}} \cap \mathcal{I}_{\tau_{k,i}|\delta})^4. \tag{29}$$

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

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

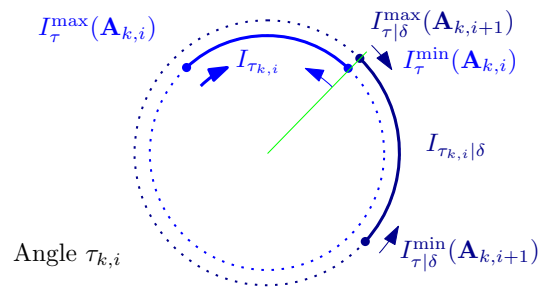
$$\forall k, \forall i, \forall a \in \mathcal{V}_k : \mathrm{DOVI}_{\tau_{k,i}}(a) \neq \emptyset.$$



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

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

$\Rightarrow$ **two implicit equations in $\mathcal{A}_k$: two sub-manifolds (under the appropriate conditions)**



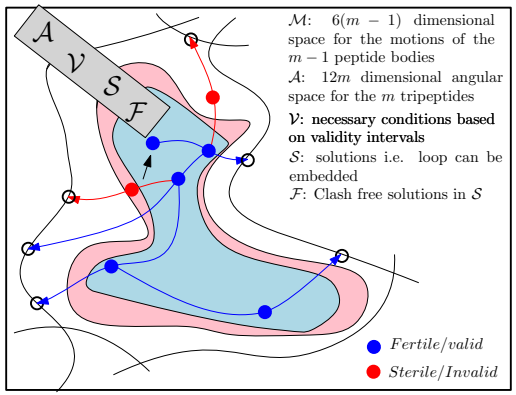
Limit case: $I_\tau^{\max}(\mathbf{A}_{k,i}) = I_{\tau|\delta}^{\min}(\mathbf{A}_{k,i+1})$

Validity domain for the whole chain L with m tripeptides

- ▷ Angles τ: $3m$ angles τ (3 for each tripeptide)
- ▷ Recap per angle τ:
  - For one angle: at most 4 Depth One Validity Intervals (DOVI)
  - For each DOVI: 2 sub-manifolds of $\mathcal{A}_k$ defined by the previous equations; yields (at most) 8 sub-manifolds in $\mathcal{A}_k$.

▷ For one tripeptide: 3 τ angles $\Rightarrow$ 24 constraint surfaces in the 12 dimensional angular space $\mathcal{A}_k$.

▷ For the whole loop: total of $24m$ constraint surfaces.



Motion space for peptide bodies

TLC for peptides with moving legs

▷ Configuration spaces for motions:

- One peptide body: $\mathcal{R} : (S^2 \times [0, A)) \times (S^2 \times [0, 2\pi)) \subset SE(3)$
- The $m-1$ peptide bodies in the loop L: $\mathcal{M} = \mathcal{R}^{m-1}$

▷ Sampling motions: sample the $m-1$ translations and $m-1$ rotations of the peptide bodies independently

▷ Ray in motion space, for $r \in \mathcal{M}$: linear interpolation between the identity and the rigid motion corresponding to r:

$$\text{Ray}(V) = \{\gamma(t) = Id + tV, \text{ with } \gamma(0) = Id\}. \tag{30}$$

▷ Restriction to each peptide body: defines a rigid transformation

$$\gamma_k : [0, 1] \mapsto SE(3), \gamma_k(0) = Id, \tag{31}$$

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

$$P_k(t) = \gamma_k(t) P_k(0). \tag{32}$$

Kinetic angular representation of a tripeptide
Kinetic validity intervals of an angle $\tau_{k,i}$

▷ Functions returning the angles $A_{k,i}$ for $C_{\alpha;i}$ at time t:

$$\begin{cases} f^{(\alpha)}_{(k,i)}(t) & : \text{ function computing the angle } \alpha_{k,i} \text{ at time } t \\ f^{(\xi)}_{(k,i)}(t) & : \text{ function computing the angle } \xi_{k,i} \text{ at time } t \\ f^{(\eta)}_{(k,i)}(t) & : \text{ function computing the angle } \eta_{k,i} \text{ at time } t \\ f^{(\delta)}_{(k,i)}(t) & : \text{ function computing the angle } \delta_{k,i} \text{ at time } t \end{cases} \tag{33}$$

$\Rightarrow$ kinetic vector for the l-th rigid body of the i-th tripeptide:

$$A_{k,i}(t) = (f^{(\alpha)}_{(k,i)}(t), f^{(\xi)}_{(k,i)}(t), f^{(\eta)}_{(k,i)}(t), f^{(\delta)}_{(k,i)}(t)). \tag{34}$$

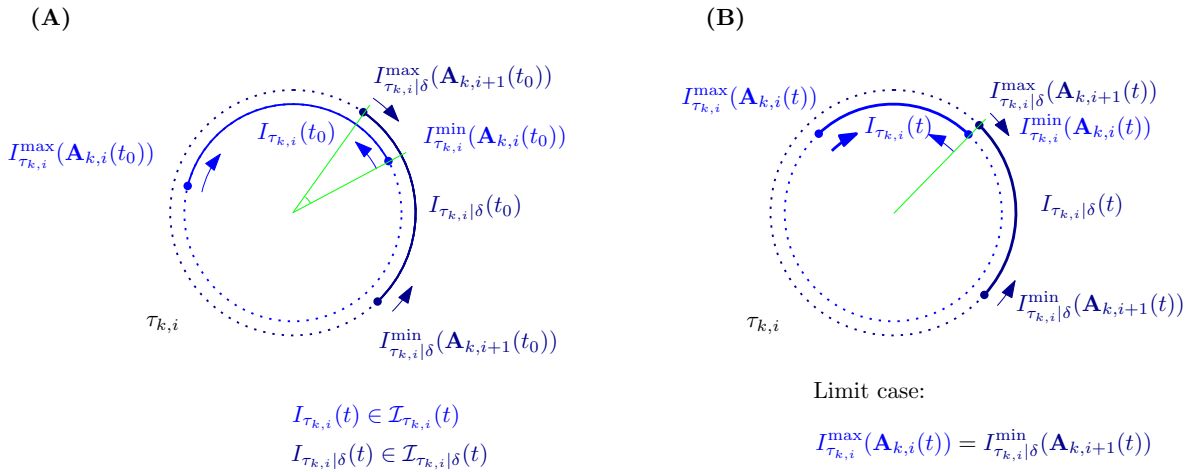
▷ Kinetic validity intervals of angle $\tau_{k,i}$:

$$\begin{cases} I_{\tau_{k,i}}(t) = [I^{\min}_{\tau}(A_{k,i}(t)), I^{\max}_{\tau}(A_{k,i}(t))] \\ I_{\tau_{k,i}|\delta}(t) = [I^{\min}_{\tau|\delta}(A_{k,i+1}(t)), I^{\max}_{\tau|\delta}(A_{k,i+1}(t))] \end{cases} \tag{35}$$

▷ Nb: these are similar to std validity interval: initial VI + transposed VI

Kinetic validity intervals for angle $\tau_{k,i}$

Definition 9. (Kinetic depth 1 validity intervals) Obtained by intersecting the original and transposed VI $I_{\tau_{k,i}}(t) \cap I_{\tau_{k,i}|\delta}(t)$:



- Case study for $\tau_{k,i}$: with $I_{\tau_{k,i}}(t) \in \mathcal{I}_{\tau_{k,i}}(t)$ and $I_{\tau_{k,i}|\delta}(t) \in \mathcal{I}_{\tau_{k,i}|\delta}(t)$
  - **(A)** The interiors of the two intervals intersect.
  - **(B)** The intervals intersect on their boundary. The arrow indicate the derivative of the endpoints of intervals with respect to time t.

- Two conditions for two kinetic intervals $I_{\tau_{k,i}}(t) \in \mathcal{I}_{\tau_{k,i}}(t)$ and $I_{\tau_{k,i}|\delta}(t) \in \mathcal{I}_{\tau_{k,i}|\delta}(t)$:

$$\begin{cases} I^{\max}_{\tau}(\mathrm{A}_{k,i}(t)) & = I^{\min}_{\tau|\delta}(\mathrm{A}_{k,i+1}(t)) \\ \text{or } I^{\min}_{\tau}(\mathrm{A}_{k,i}(t)) & = I^{\max}_{\tau|\delta}(\mathrm{A}_{k,i+1}(t)) \end{cases} \tag{36}$$

Nb: yields 8 equations per angle τ.

Functional form of the time dependent equations

▷ Equations of the form:

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

▷ (Gory) Details:

- Time dependent rigid motions in $SE(3)$ – e.g. angle-vector and Rodrigues' formula for rotations
- Computation of the $\{\alpha, \xi, \eta, \delta\}$ angles – involves $\sqrt{\cdot}$ and inverse trig functions
- Validity intervals via the valence angle constraint

$\Rightarrow$ rather complex univariate (var: t) equations

▷ Solver: Maple

Algorithm overview

▷ For a given angle τ:

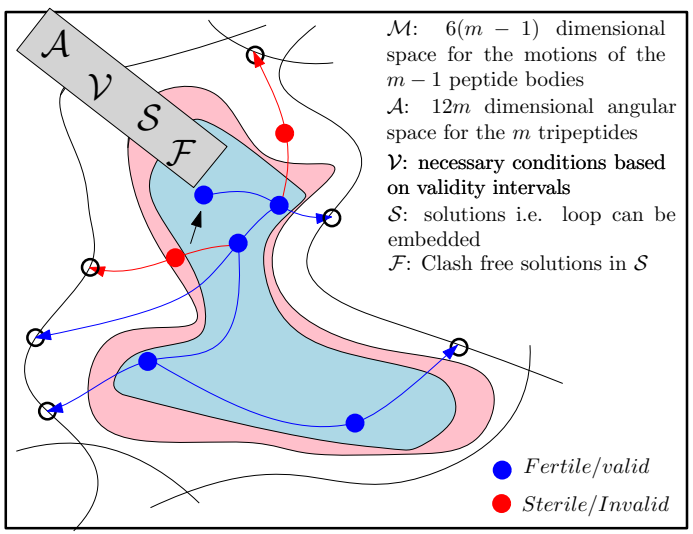
$$t \text{ determines the positions of peptide bodies whence tripeptide legs} \tag{37}$$

$$\rightsquigarrow \text{ kinetic angular representation } \mathsf{A}_i(t) \text{ of } T_k \tag{38}$$

$$\rightsquigarrow \text{ kinetic validity intervals } I_{\tau_{k,i}}(t), I_{\tau_{k,i}|\delta}(t) \tag{39}$$

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

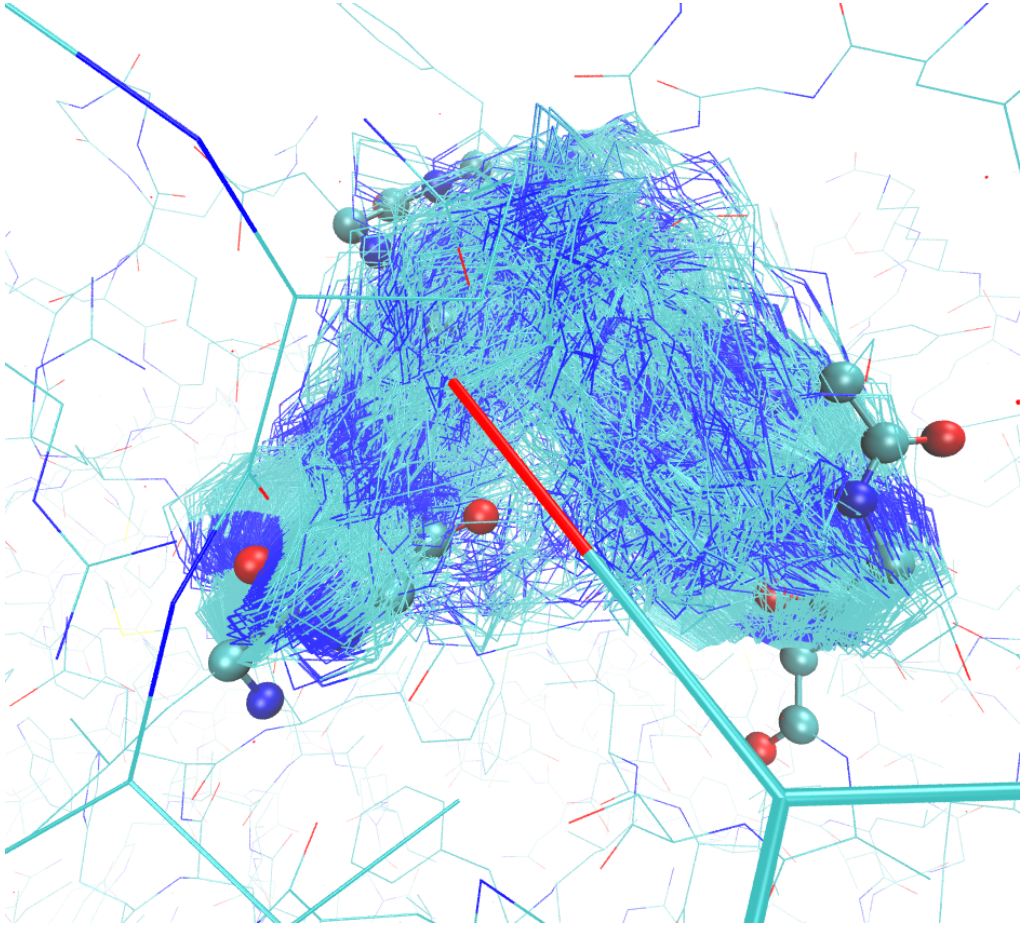
$$I_{\tau}^{\max}(\mathsf{A}_{k,i}(t)) = I_{\tau|\delta}^{\min}(\mathsf{A}_{k,i+1}(t)) \tag{40}$$



▷ Algorithm overview:

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

VMD demo



Loops sampling: ϕ, ψ and ω

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

Loops sampling: ϕ, ψ and ω

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

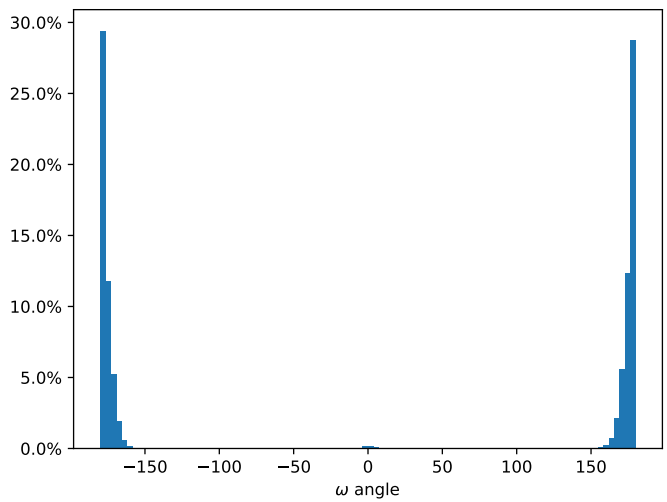
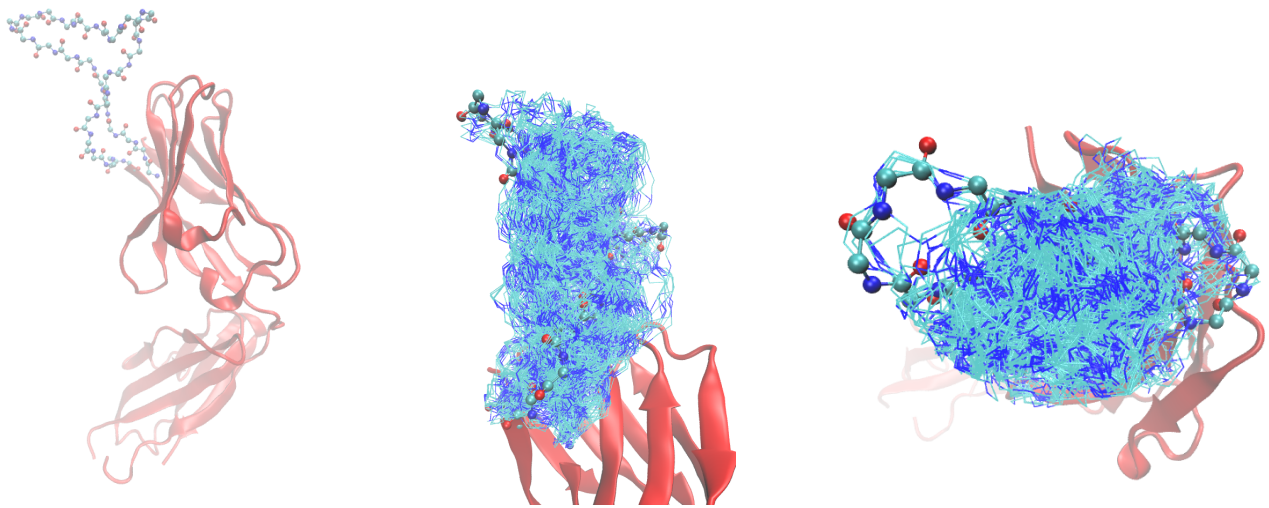


Illustration: CDR-H3-HIV, 30 amino acids

▷ System:

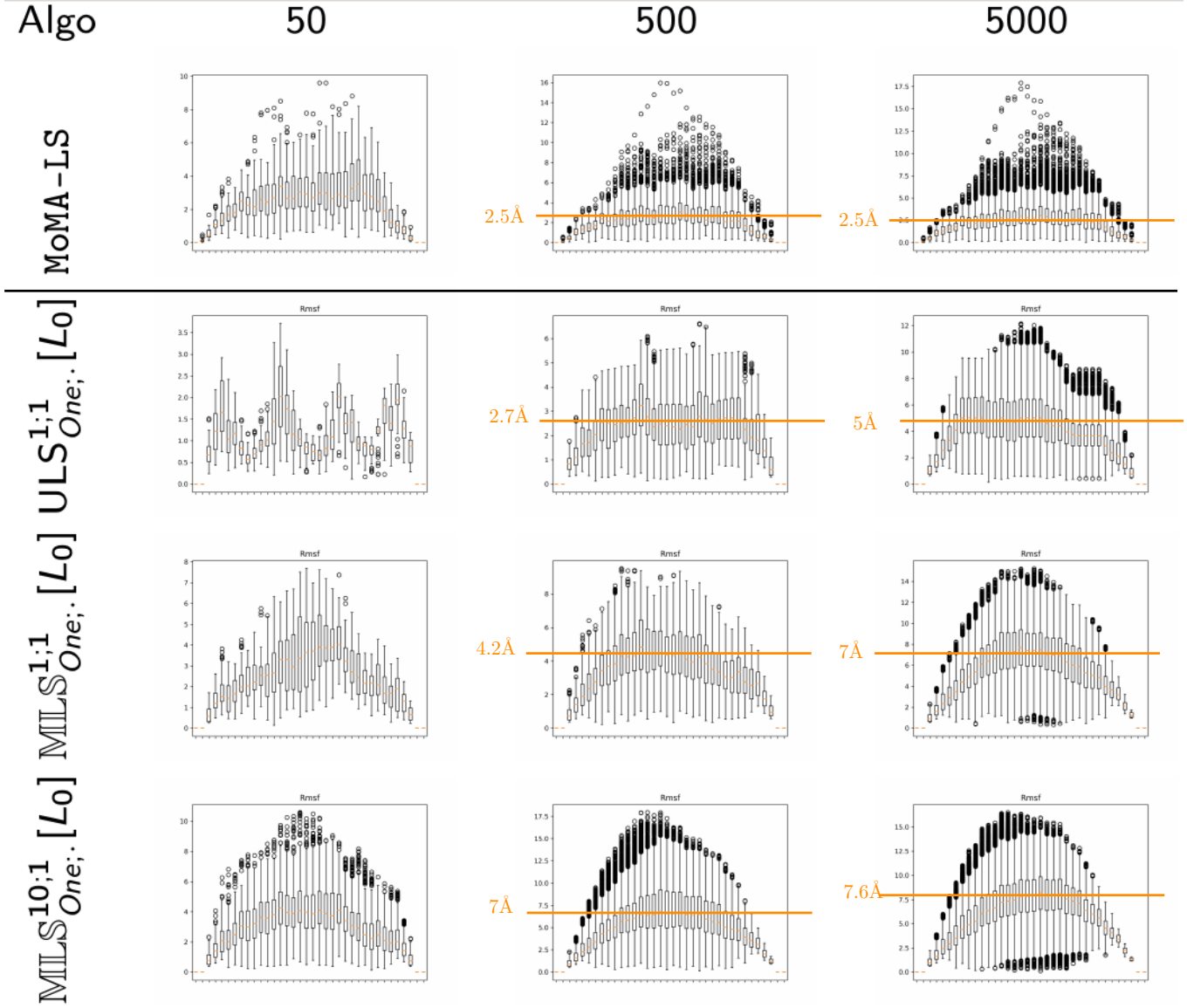
- The loop is a complementarity-determining region (CDR-H3) from PG16, an antibody with neutralization effect on HIV-1.
- pdbid: 3mme, chain A; residues: 93-100, 100A-100T, 101, 102.



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

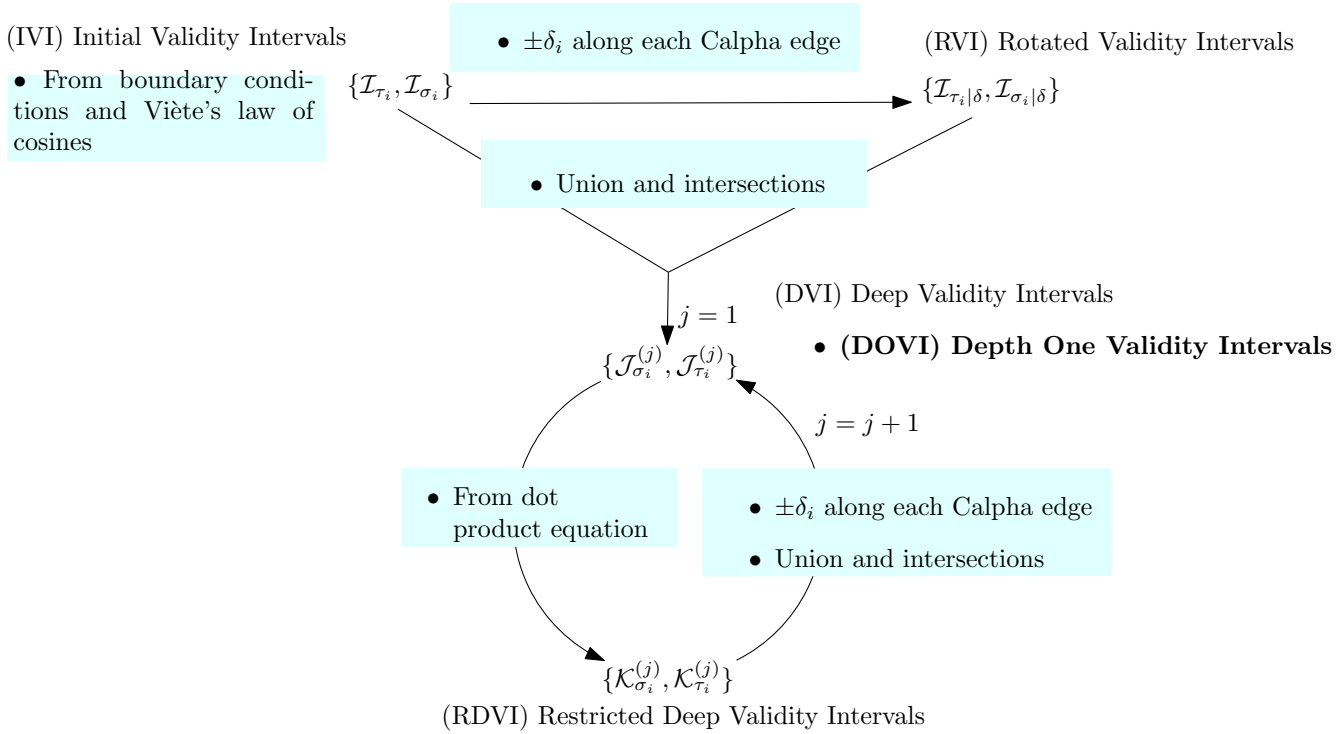
▷ Generation speed: $\sim$ 10 conformations per second

Results: sampling and study of fluctuations



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

(Open problems) Tripeptide Loop Closure



▷ Open problem: do the Deep Validity Intervals converge to the solutions of the degree 16 polynomial?

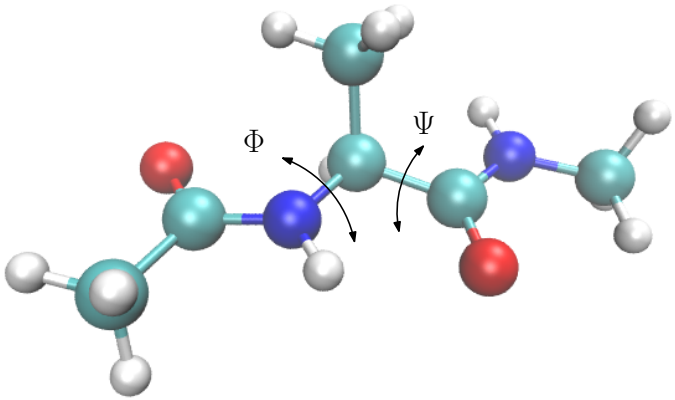
▷ Open problem: Can one leave the realm of zero dimensional problems, and obtain solutions as one (?) dimensional algebraic sets?

Mining molecular flexibility: novel tools, novel insights

PART 1: Introduction to Structural Bioinformatics
PART 2: Protein structure and geometry
PART 3: Thermodynamics and the volume of polytopes
PART 4: Outlook

Density of states and partition functions

Dialanine

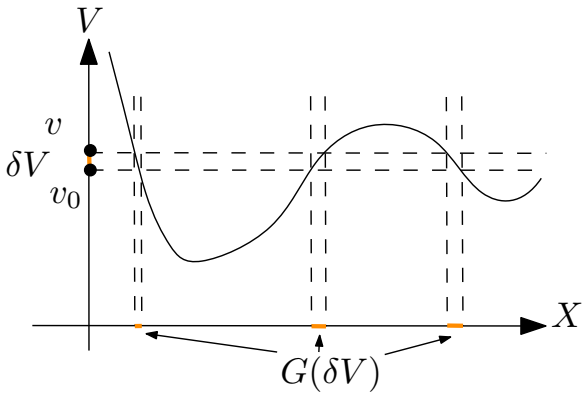


Molecule in water at temperature T

- q: vector of positions of atoms
- Potential energy:

$$V(q)$$

▷ Potential energy landscape:



▷ Density of states (DoS):

- Push forward of the Lebesgue measure by the potential energy V:
- For any $v_0 < v_1$:

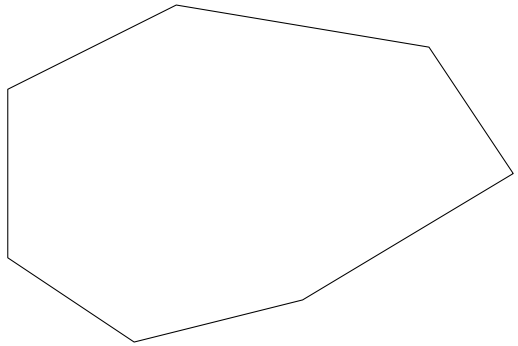
$$g([v_0, v_1]) = \int_X 1_{[v_0, v_1]}(V(q))dq$$

▷ Partition function for $A \subset X$: integrate Boltzmann's factor

$$Z_A(T) = \int_{\mathbb{A}} e^{-\beta u} dg(u)$$

▷ NB: n atom: $d = 3n$ Cartesian coordinates. Exple: antibody: $d \sim 42,000$

Volume of polytopes: hardness, randomized algorithms



▷ Hardness: no polynomial time algorithm with approx factor $(cd/\log d)^d$ – unless P=NP

▷ ε-approximation of the volume: for any parameter $\varepsilon > 0$, a number V

$$(1-\varepsilon)\mathrm{Vol}(K) \leq V \leq (1+\varepsilon)\mathrm{Vol}(K).$$

▷ (ε, δ)-approximation algorithm: algorithm returning an ε-approximation with a probability at least $1-\delta$.

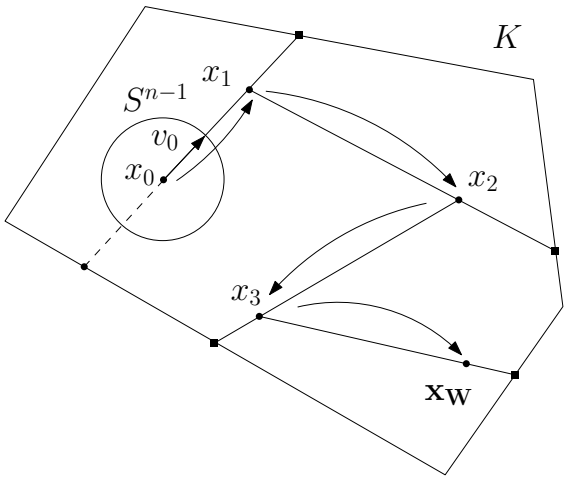
▷ Complexity, the $O^\star(n)$otation:

- $O(d^4)$: upper bound as a function of the dimension d
- $O^\star(d^4)$: term in $\log d, \varepsilon, \delta$ removed; focus on the dimension solely

▷Ref: Cousins, Vempala, SIAM J. Comp., 2018

Random walk: hit-and-run

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



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

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

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

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

▷ NB: precise statement in terms of total variation distance omitted

▷Ref: Berbee et al, Math. Prog., 1987

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

Randomized algorithms: complexity

▷ Thm. For a convex body K given by a membership oracle, and such that $B \subset K \subset RB$, an $(\varepsilon, \delta)-$ approximation can be obtained in time

$$O(\frac{d^4}{\varepsilon^2} \log^9 \frac{d}{\varepsilon\delta} + d^4 \log^8 \frac{d}{\delta} \log R) = O^\star(d^4) \tag{41}$$

▷ Volume estimated using a sequence of isotropic Gaussians:

$$\text{Vol}(K) = \int_K f_0(x)dx \frac{\int_K f_1(x)dx}{\int_K f_0(x)dx} \cdots \frac{\int_K dx}{\int_K f_{m-1}(x)dx} \equiv \int_K f_0(x)dx \prod_{i=1,\ldots,m} R_i \tag{42}$$

▷ Cooling schedule i.e. sequence of Gaussians $f_0, \ldots, f_m$:

- f_0: sharply peaked in K;
- f_m: uniform distribution i.e. $a_m = 0$
-

▷ Complexity, overview: $m = O^\star(\sqrt{d})$ functions used. Each ratio in the telescoping product is estimated (with guarantees) using $O^\star(\sqrt{d})$ samples. The complexity of generating a given sample being $O^\star(d^3)$, the overall algorithm has complexity $O^\star(d^4)$.

▷Ref: Lovász, Vempala, J Comp. Syst. Sciences, 2006
▷Ref: Cousins, Vempala, SIAM J. Comp., 2018

A practical algorithm: outline

▷ Method:

- multi-phase Monte-Carlo using $m = O(\sqrt{d})$ logconcave functions $\{f_0, \ldots, f_{m-1}\}$,
  - $f_i(x) \propto e^{-a_i^T x}$ or $f_i(x) \propto exp(-a_i \|x\|^2)$
- At each step: estimate $r_k \approx \int_K f_k(x)dx / \int_K f_{k-1}(x)dx$

Volume(K, ε): Convex body K, error parameter ε.

- $T =$ **Round**(body: K, steps: $8n^3$), set $K' = T \cdot K$.
- $\{a_0, \ldots, a_m\} =$ **GetAnnealingSchedule**(body: K').
- Set x to be random point from $f_0 \cap K'$, $\varepsilon' = \varepsilon/\sqrt{m}$.
- For $i = 1, \ldots, m$,
  - Set $k = 0, x_0 = x, converged = false, W = 4n^2 + 500$.
  - While $converged = false$,
    - $k = k + 1$.
    - $x_k =$**HitAndRun**(body: K, target distribution: f_{i-1}, current point: x_{k-1}).
    - Set
    $$r_k = \frac{1}{k} \sum_{j=1}^{k} \frac{f_i(x_j)}{f_{i-1}(x_j)}.$$
    - Set $W_{max} = \max\{r_{k-W+1}, \ldots, r_k\}$ and $W_{min} = \min\{r_{k-W+1}, \ldots, r_k\}$.
    - If $W_{max} - W_{min} \leq \varepsilon'/2 \cdot W_{max} \rightarrow converged = true$.
  - Set $R_i = r_k, x = x_k$.
- Return $volume = |T| \cdot (\pi/a_0)^{n/2} \cdot R_1 \ldots R_m$.

▷Ref: Cousins and Vempala, Math. Prog. Comp., 2016

Application to the ratios R_i

- Recall $R_i = \frac{\int_K f_i(x)dx}{\int_K f_{i-1}(x)dx} = \int_K \frac{f_i(x)}{f_{i-1}(x)} \frac{f_{i-1}(x)}{\int_K f_{i-1}(x)dx} dx$
- Define $Y = f_i(X)/f_{i-1}(X)$, and let $X \sim f_{i-1}(X)/\int_K f_{i-1}(x)dx$.
- One has

$$\mathbb{E}[Y] = \int_K \frac{f_i(x)}{f_{i-1}(x)} \frac{f_{i-1}(x)}{\int_K f_{i-1}(x)dx} dx = \frac{\int_K f_i(x)dx}{\int_K f_{i-1}(x)dx}. \tag{43}$$

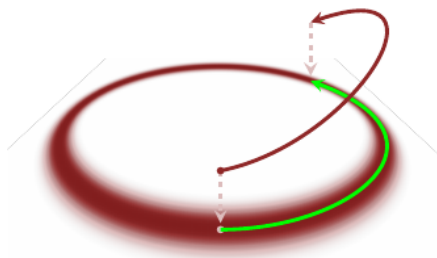
- Associated estimator: with X_i a set of k iid RV $\sim f_{i-1}(x)/\int_K f_{i-1}(y)$:

$$\tilde{R}_i = \frac{1}{k} \sum_j \frac{f_i(X_j)}{f_{i-1}(X_j)}. \tag{44}$$

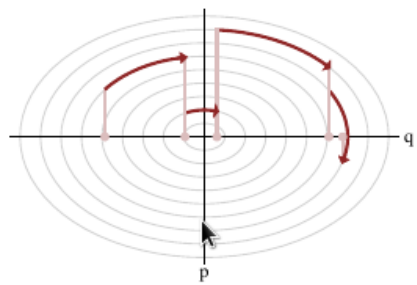
- Importance sampling in disguise

Using HMC to sample a distribution

- Goal: sample a distribution $\pi(q)$
  - Define $U(q) = -log(\pi(q))$ and $K(p) = 1/2\|p\|^2$ (Nb: unit masses)
  - $H(p,q) = U(q) + K(p)$
  - Invariant measure used: $\mu(q,p) = \exp(-H(q,p)) = \pi(q)\exp(-K(p))$, with $\pi(q) = \exp(-U(q))$



- Sampling with HMC: algorithm
  - fix travel time $L > 0$
  - Iterate
    - resample $p \sim \mathcal{N}(0, I_n)$
    - $(q^{(t+1)}, p^{(t+1)}) = \Phi_L(q^{(t)}, p)$



- Rmk: resampling p changes the energy level

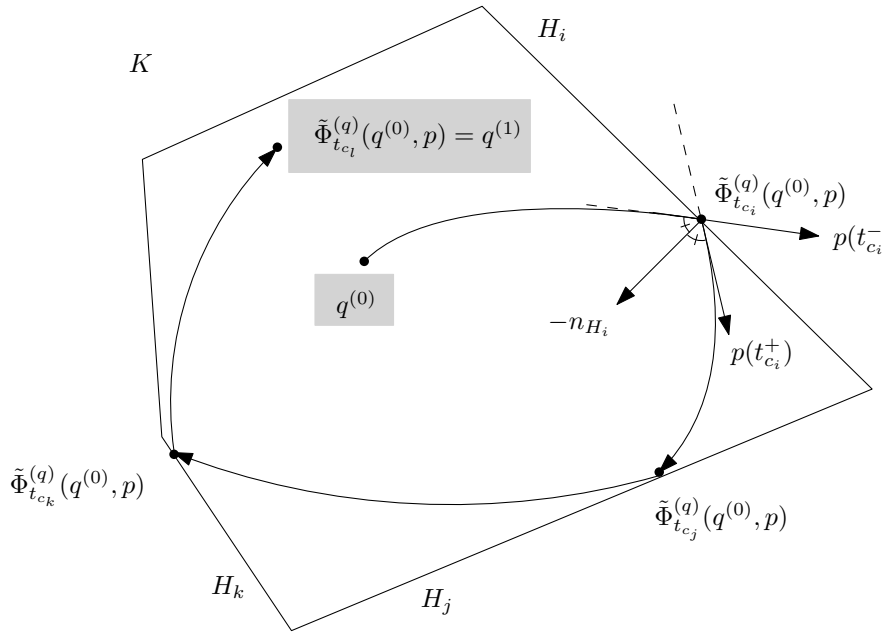
- Nb: HMC and the curse of dimensionality: gliding near the *typical set*

- Ref: Betancourt, ArXiv, 2018

HMC in a polytope: a curved billiard walk

▷ Method:

- HMC with $U(q) = \exp(-a\,\|q\|^2)$
- Reflexions on boundaries of K
- Analytical solutions for trajectories: harmonic oscillator



▷ Parameters:

- Travel time L
- Max number of reflexions Max_{reflex} should be large for the RW to forget its origin and mix

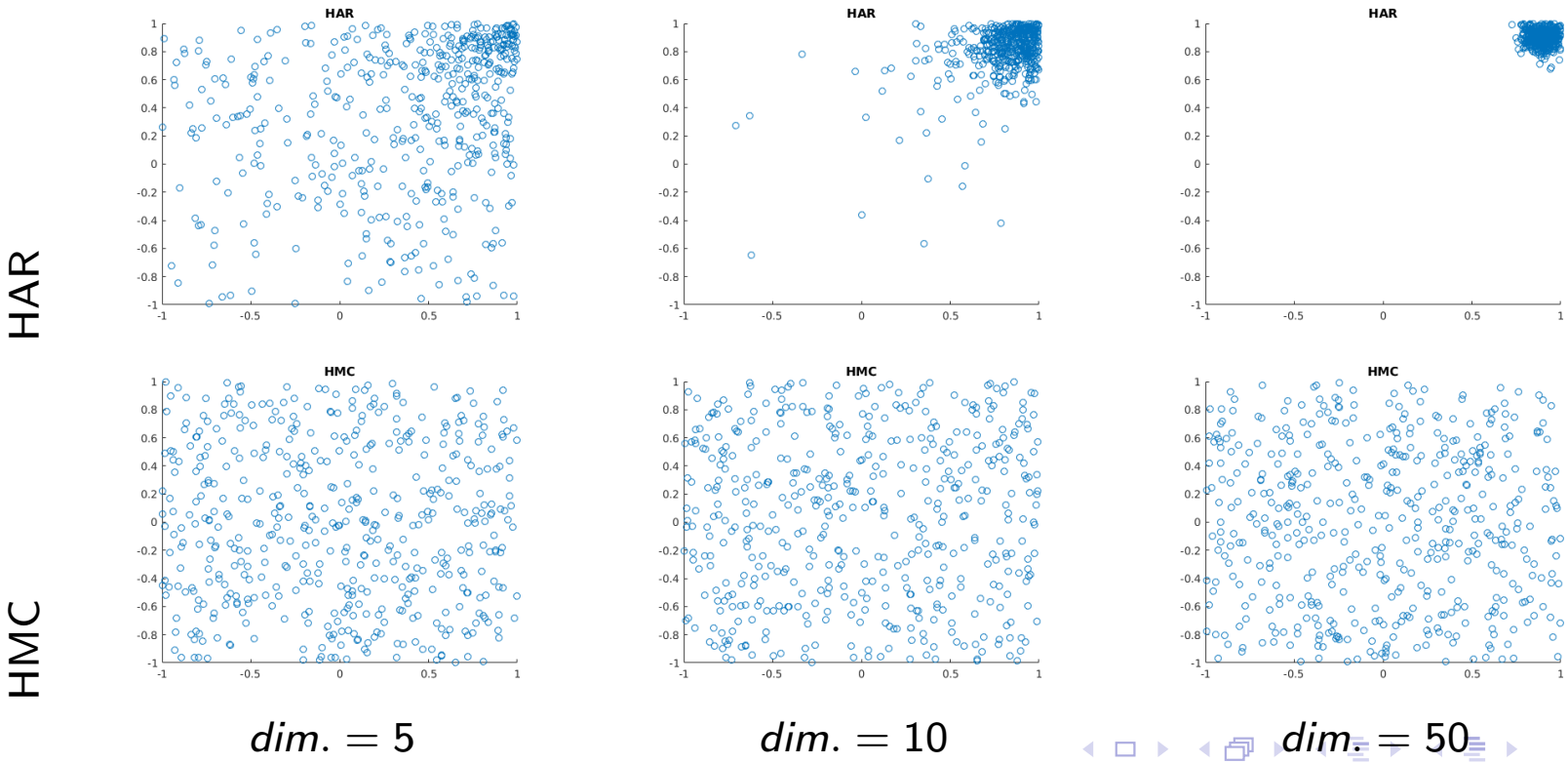
▷ Nb: numerics are **tricky** due to cascaded constructions. It may not be possible to evaluate the *Test_to_zero* predicate ... even if never faced in practice.

Sampling a target distribution with HMC:

illustration of mixing properties

- Setup:
  - Cube $[-1,1]^n$, $n = 5, 10, 50$
  - Target distribution $\pi(q)$: flat isotropic Gaussian ($\sigma_i^2 = 500$)
  - Starting point $q^{(0)}$: $q_i^{(0)} = 0.9, \forall i$, return $q^{(10)}$
  - Repeat 500 times
  - Plots: projection i.e. first 2 coordinates

- HAR vs HMC



Embedding HMC into the volume algorithm

Volume(K, ε): Convex body K, error parameter ε.

- $T =$ **Round**(body: K, steps: $8n^3$), set $K' = T \cdot K$.
- $\{a_0, \ldots, a_m\} =$ **GetAnnealingSchedule**(body: K').
- Set x to be random point from $f_0 \cap K'$, $\varepsilon' = \varepsilon/\sqrt{m}$.
- For $i = 1, \ldots, m$,
  - Set $k = 0, x_0 = x, converged = false, W = 4n^2 + 500$.
  - While $converged = false$,
    - $k = k + 1$.
    - $x_k =$ ~~**HitAndRun**~~(body: K, target distribution: f_{i-1}, current point: x_{k-1}).
    - Set HMC
    $$r_k = \frac{1}{k} \sum_{j=1}^{k} \frac{f_i(x_j)}{f_{i-1}(x_j)}.$$
    - Set $W_{max} = \max\{r_{k-W+1}, \ldots, r_k\}$ and $W_{min} = \min\{r_{k-W+1}, \ldots, r_k\}$.
    - If $W_{max} - W_{min} \leq \varepsilon'/2 \cdot W_{max} \rightarrow converged = true$.
  - Set $R_i = r_k, x = x_k$.
- Return $volume = |T| \cdot (\pi/a_0)^{n/2} \cdot R_1 \ldots R_m$.

Window sizes: $n^0, n^1, n\sqrt{n}, n^2$

▷ Stop condition: the window size W sets the stop criterion

▷ Stats monitored:

- Relative error $| V - \mathrm{Vol}(K) | / \mathrm{Vol}(K)$
- # calls to the oracle

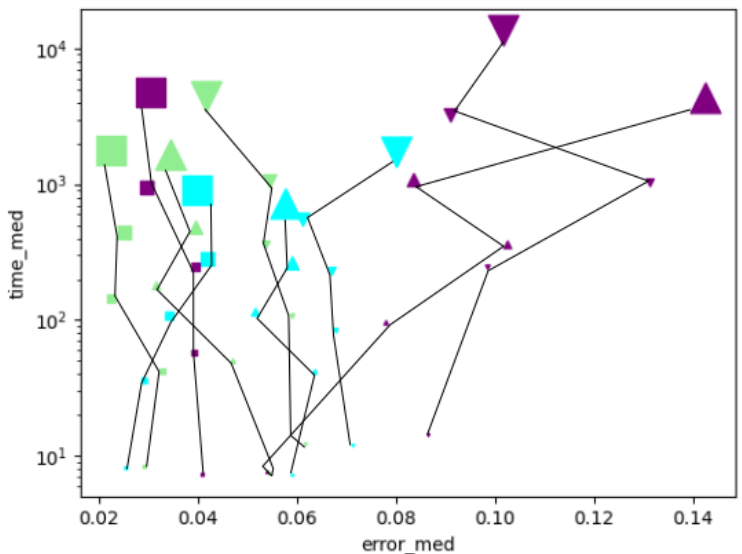
▷ Polytopes tested in $\mathbb{R}^n$, for $n = 10, \ldots, 50$:

- Cube: a must
- Simplex: standard simplex,isotropic simplex

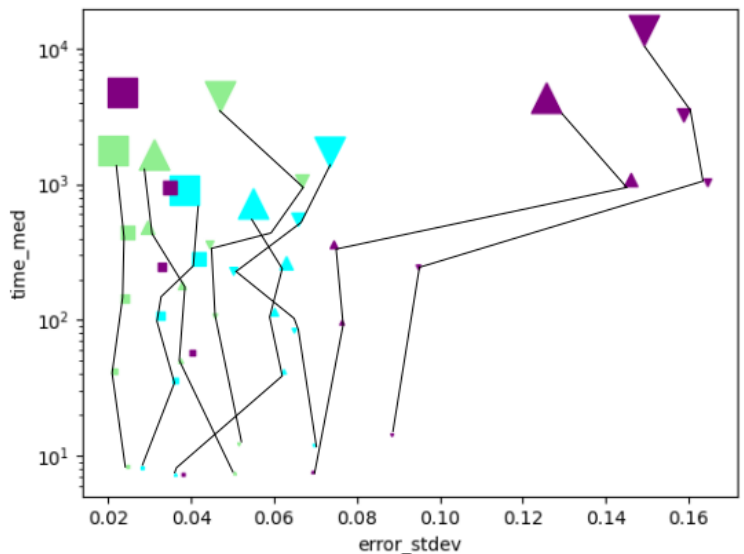
▷Ref: Cousins and Vempala, Math. Prog. Comp., 2016

Volume calculation: relative error and running times

- **Relative errors on volume and running times**
  - Three polytopes: cube (square), isotropic simplex (triangle up), standard simplex (triangle down), in dimension $n = 10, 30, 50, 70, 100$.
  - Statistics reported over 50 runs.
  - The error of an estimate $\tilde{V}$ is defined as $\mathrm{err}_r = |\tilde{V} - V|/V$.
  - Three algorithms: purple/light blue/light green HAR/HMC0/HMC1.5



median error $\times$ median time



stdev(error) $\times$ median time

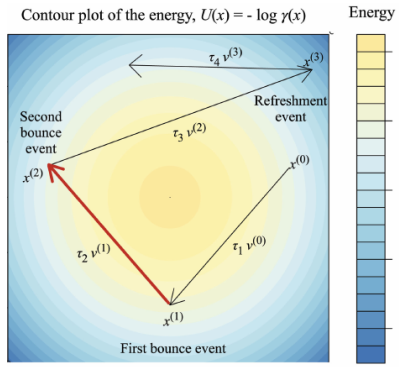
Piecewise deterministic Markov processes (PDMP)

the non-reversible Bouncy Particle Sampler (BPS)

▷ Notations: state space (position, velocity): $z = (x, v) \in E = \mathbb{R}^d \times \mathbb{R}^d$.

▷ PDMP z_t: a continuous time Markov process defined by:

1. a deterministic flow $\phi_t(z)$,
2. function determining the length of steps: jump kernel $\lambda(z)$
3. a jump kernel in phase (x, v) space: $q(\cdot|z)$



▷ BPS: PDMP to sample a distribution $\pi(x)$ in $\mathbb{R}^d$ using piecewise linear trajectories bouncing on high energy level set surfaces

1. Linear trajectories: $\phi_t(x, v) = (x + tv, v)$,
2. Arrival time of 1D inhomogeneous Poisson process of intensity $\lambda(x, v) = \max(0, -\langle \nabla_x(\log \pi)(x), v \rangle)$,
3. $q(\cdot|z)$: reflection w.r.t. the gradient of the potential:

$$(x, v') = \left(x, v - 2 \frac{\langle v, \nabla_x(\log \pi)(x) \rangle}{\|\nabla_x(\log \pi)(x)\|^2} \nabla_x(\log \pi)(x) \right) \tag{45}$$

4. +Refresh of velocity to ensure ergodicity

▷Ref: Doucet et al, Stats. and probability letters, 136, 2018

Extension: BPS on a bounded domain – a polytope

▷ Three types of events:

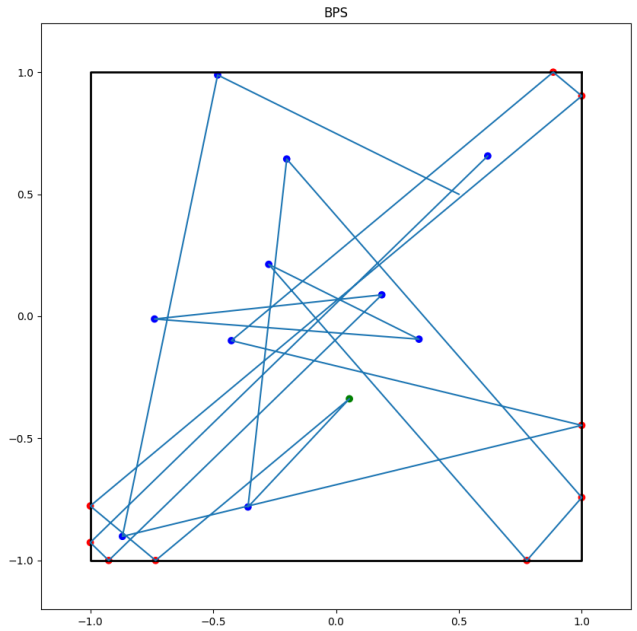
- PDMP events: as usual
- Reflexions on the boundary

$$v' = v - 2\frac{\langle n, v\rangle}{\|n\|^2} n, \qquad (46)$$

- Refresh events: velocity resampled from isotropic normal distribution

▷ Numerics: lazy update of linear algebra operations

▷ Example BPS trajectory in the 2d cube $[-1, 1]^2$:



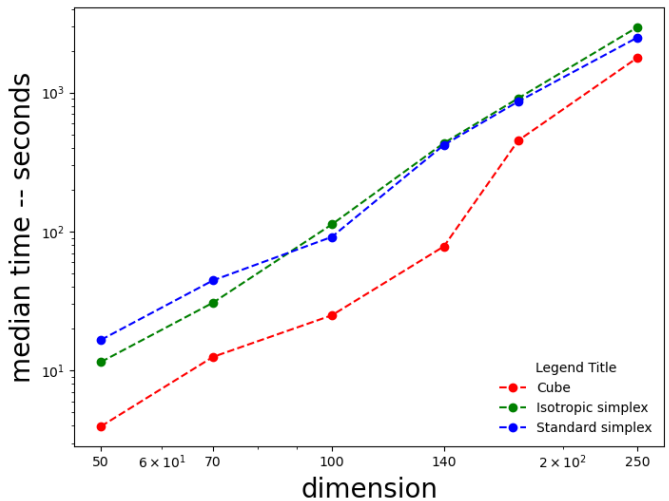
Blue: PDMP jump events, Red: reflections on the boundary, Green: refresh events
Nb: $\pi(x)$: Gaussian of variance $\sigma = 1$.

PDMP to compute volumes of polytopes: experiments

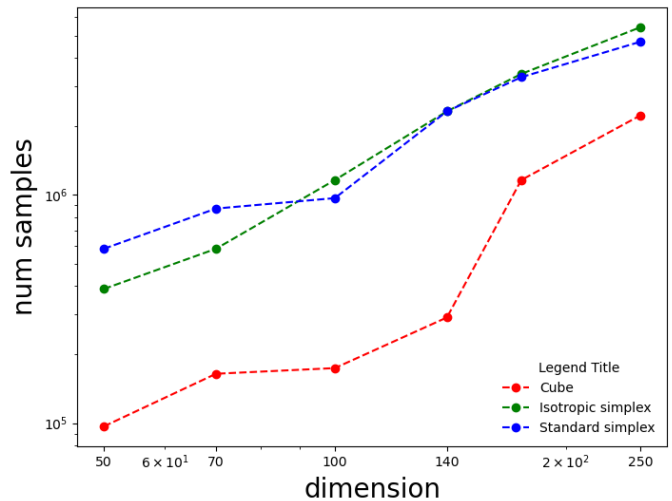
▷ Complexity: $C = O(d^c)$, dimension up to $d = 250$

▷ Protocol: find the smallest number of samples so that the estimated volume is within $err\%$ from the exact value

log(median time) versus log(dim)



log(num samples) versus log(dim)



▷ Linear regression in log log scale for the three polytopes:

	Time		Num. samples	
model	slope	R^2	slope	R^2
cube	3.77	0.96	1.94	0.88
Δ_{iso}	3.52	1.00	1.72	0.99
Δ_{std}	3.18	0.99	1.37	0.96

Mining molecular flexibility: novel tools, novel insights

PART 1: Introduction to Structural Bioinformatics
PART 2: Protein structure and geometry
PART 3: Thermodynamics and the volume of polytopes
PART 4: Outlook

Computational Structural Biology/Theoretical Biophysics

- Physical model (at least for non covalent interactions) are mature
- Outstanding high dimensional (geometric) problems. . .
- With potentially critical applications in biology and (molecular) medicine

$\Rightarrow$ Joint the effort !

▷ Software: the Structural Bioinformatics Library, see `http://sbl.inria.fr`

Bibliography

- Protein geometry
  - **Tripeptide loop closure: a detailed study of reconstructions based on Ramachandran distributions**; T. O'Donnell, and C.H. Robert, and F. Cazals; Proteins: structure, function, and bioinformatics, 90 (3), 2022.
  - **Geometric constraints within tripeptides and the existence of tripeptide reconstructions**; `https://www.biorxiv.org/content/10.1101/2022.06.21.497005v1`; T. O'Donnell and F. Cazals; Submitted.
  - **Enhanced conformational exploration of protein loops using a global parameterization of the backbone geometry**; `https://www.biorxiv.org/content/10.1101/2022.06.21.497022v1`; T. O'Donnell and F. Cazals; Submitted.
- Thermodynamics and the volume of polytopes
  - **Efficient computation of the volume of a polytope in high-dimensions using Piecewise Deterministic Markov Processes**; A. Chevallier, and F. Cazals, and P. Fearnhead; AISTATS, 2022.
  - **Improved polytope volume calculations based on Hamiltonian Monte Carlo with boundary reflections and sweet arithmetics**; A. Chevallier, and S. Pion, and F. Cazals; J. of Computational Geometry, NA (NA), 2022
  - **Wang-Landau algorithm: an adapted random walk to boost convergence**; A. Chevallier, and F. Cazals; J. of Computational Physics, 410 (1), 2020

`http://www-sop.inria.fr/teams/abs/publications/frederic-cazals.html`

Acknowledgments

- Collaborators
  - Timothee O'Donnell (Inria ABS)
  - Augustin Chevallier (Univ. Lancaster)
  - Dorian Mazauric (Inria ABS)
  - Charles Robert (CNRS IPBC)
  - David Wales (Cambridge University)
  - Juan Cortés (CNRS LAAS)
- Inria
- 3IA Côte d'Azur
- Université Côte d'Azur