Entangled structures in biopolymers

Boštjan Gabrovšek
University of Ljubljana

KOI 2024, sep. 25-27, Brela, Croatia

UNIVERZA V LJUBLJANI

Inštitut za matematiko, fiziko in mehaniko

Sulkowska Group

RUDOLFOVO Znanstveno in tehnološko središče Novo mesto

Plan

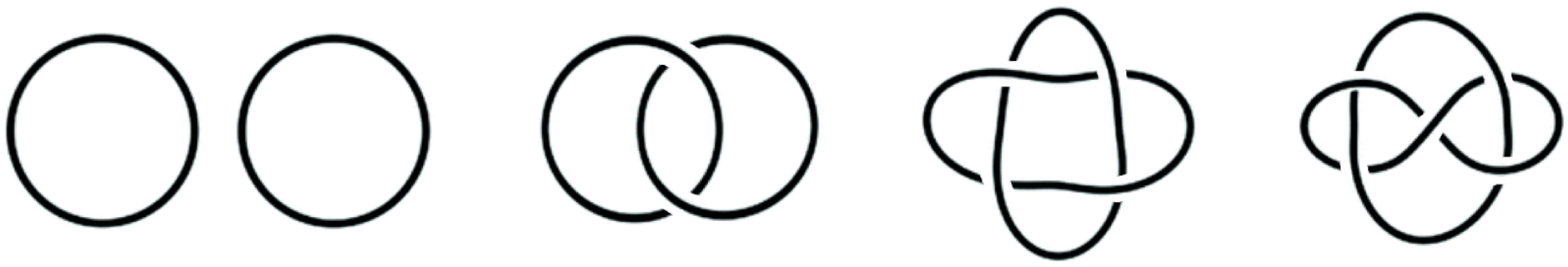
- Knots
- Protein folding
- Knotted proteins
- Mathematical invariants
- Classification of knots in proteins using ML
- Classification of lasso proteins using Topological Data Analysis

What is a knot?

A mathematical knot is an embedding of a circle into $\mathbb{R}^3$, $S^1 \hookrightarrow \mathbb{R}^3$.

0_1 (unknot) | 3_1 (trefoil) | 4_1 (figure eight) | 5_1 (cinquefoil) | 5_2 (3-twist)

A link is an embedding of several circles into $\mathbb{R}^3$, $S^1 \sqcup S^1 \sqcup \ldots \sqcup S^1 \hookrightarrow \mathbb{R}^3$.

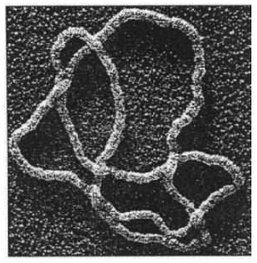


$L0_1$ (unlink) | $L2a1$ (Hopf link) | $L4a1$ (Solomon's knot) | $L5a1$ (Whitehead link)

Where do knots appear?

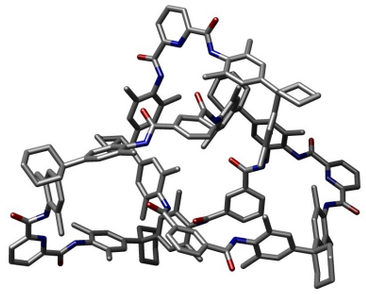
Nature

Industry

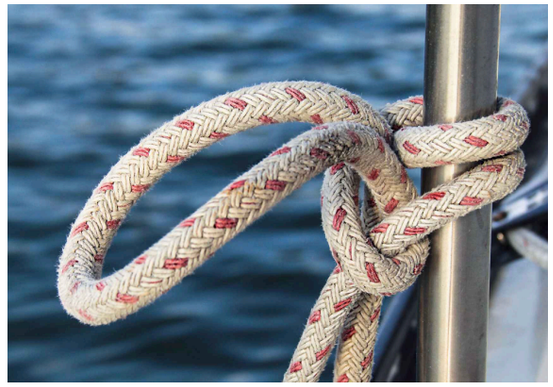


DNA

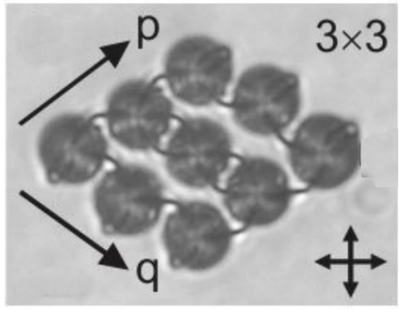
Fluid dynamics



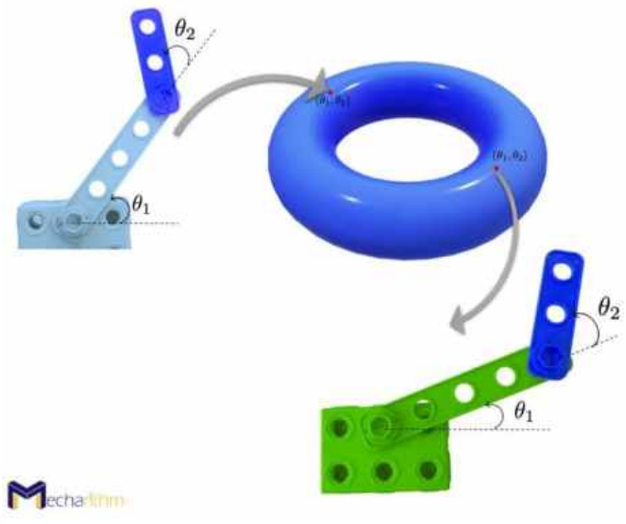
Chemistry



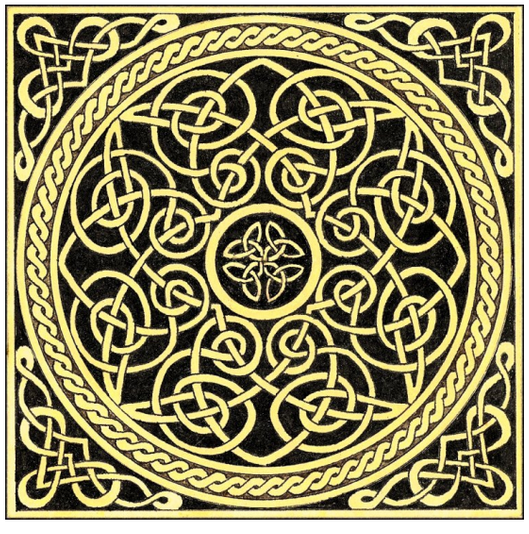
Sailing



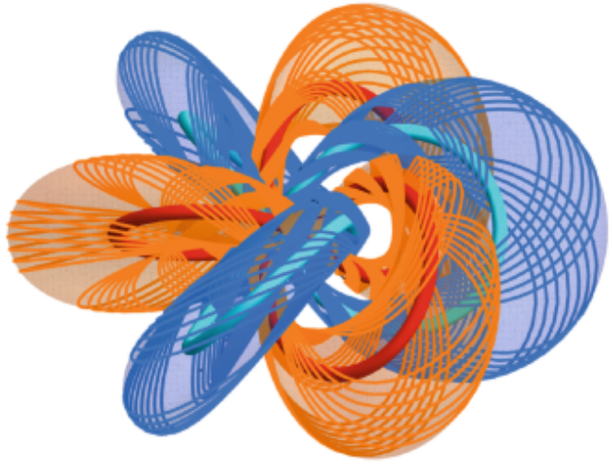
Molecular knots



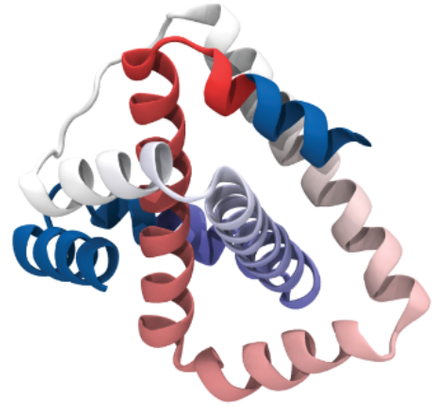
Robotics



Celtic knots (art)



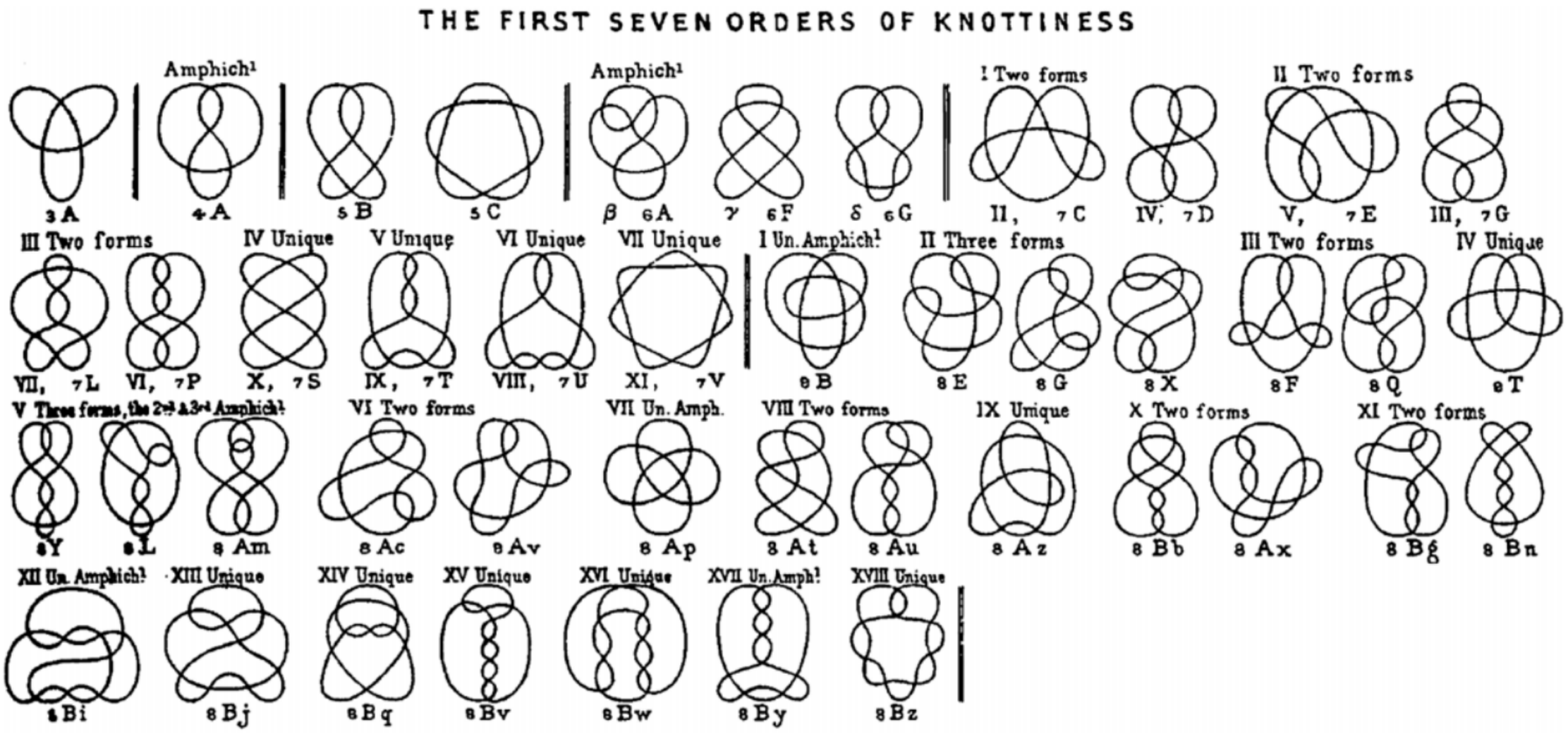
Electromagnetism



Proteins

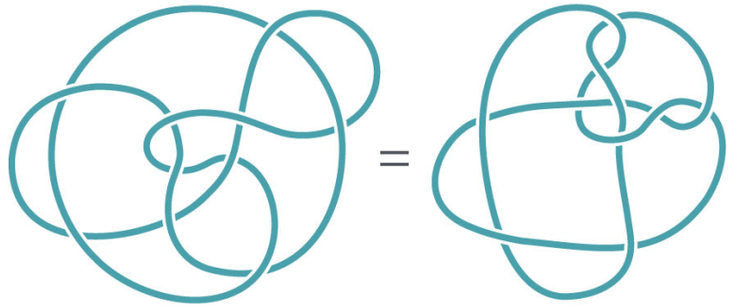
History of knot tabulation

- Tait (1876), a colleague of Kelvin – knots up to 7 crossings (15 knots)



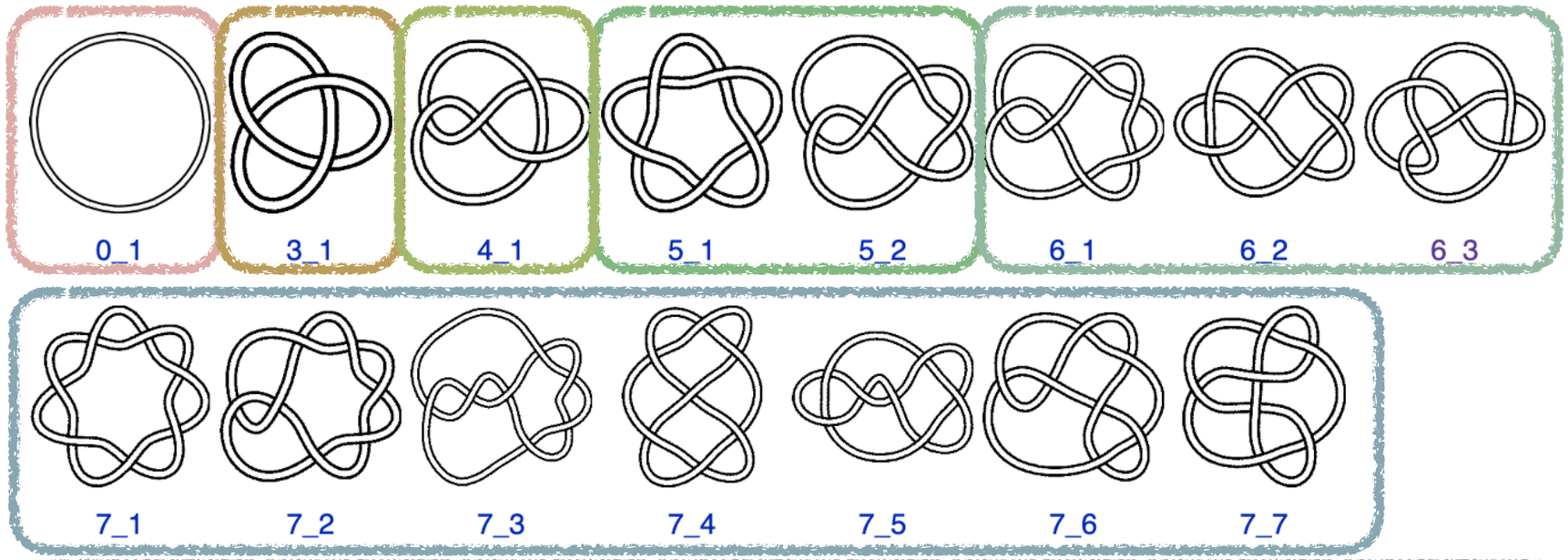
History of knot tabulation

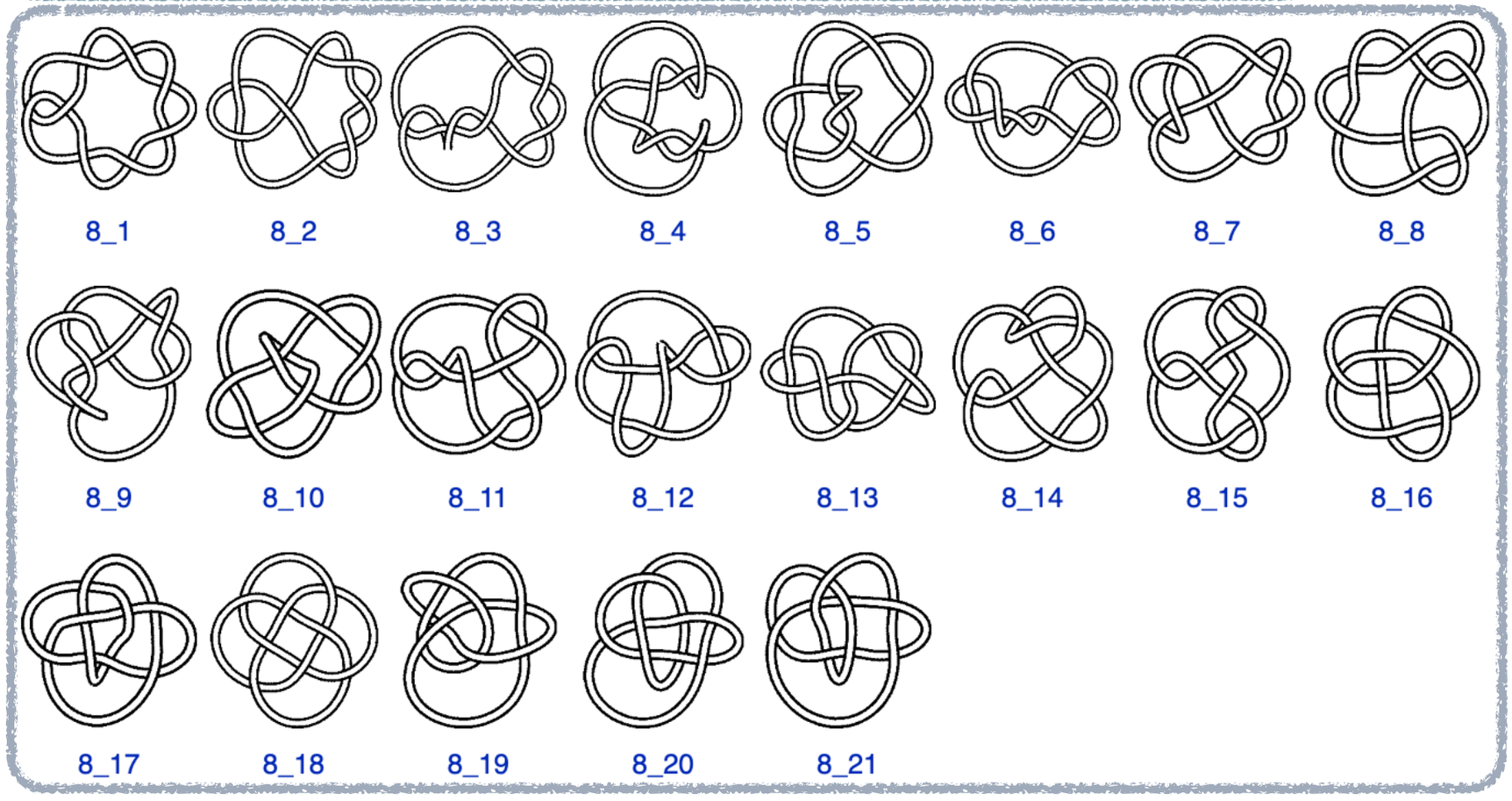
- Little (1885) – knots to 10 crossings (many errors)
- Alexander-Briggs (1927) – all 9 crossing knots (3 errors)
- Conway (1964) - knots to 11 crossings (with errors)
- Rolfsen (1976) - knots to 10 crossings (1 error)



- Caudron (1978) – knots to 11 crossings (correct)
- Hoste/Thistlethwaite/Weeks (1998) - knots to 16 crossings (1,701,936 knots)
  J. Hoste, M. Thistlethwaite, J. Weeks. "The first 1,701,936 knots", The Mathematical Intelligencer, 20:4 (1998).
- Burton (2020) - knots up to 19 crossings (352,152,252 knots)
  B.A. Burton. "The Next 350 Million Knots", 36th International Symposium on Computational Geometry (2020).

Knots up to 8 crossings





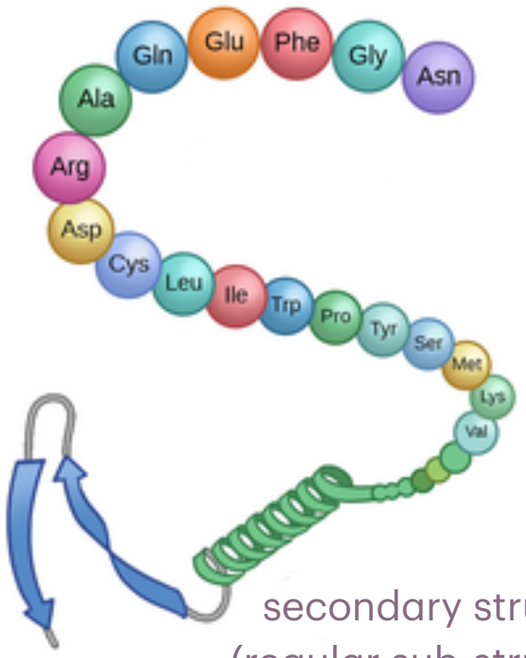
What are proteins?

- Proteins are long molecular chains (biopolymers) composed of dozens, hundreds, or even thousands of amino acids.

- Proteins are the molecular workhorses and building blocks of living organisms as they provide with functions, such as: structural support (kolagen, keratin), enzymatic activity, transport (hemoglobin), storage, signalling, defense (antibodies), movement, regulation (insulin),...

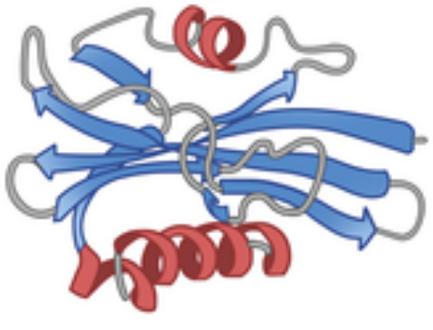
- The native 3D structure is determined by the amino acid sequence.

primary structure
(sequence of amino acids)



secondary structure
(regular sub-structures)

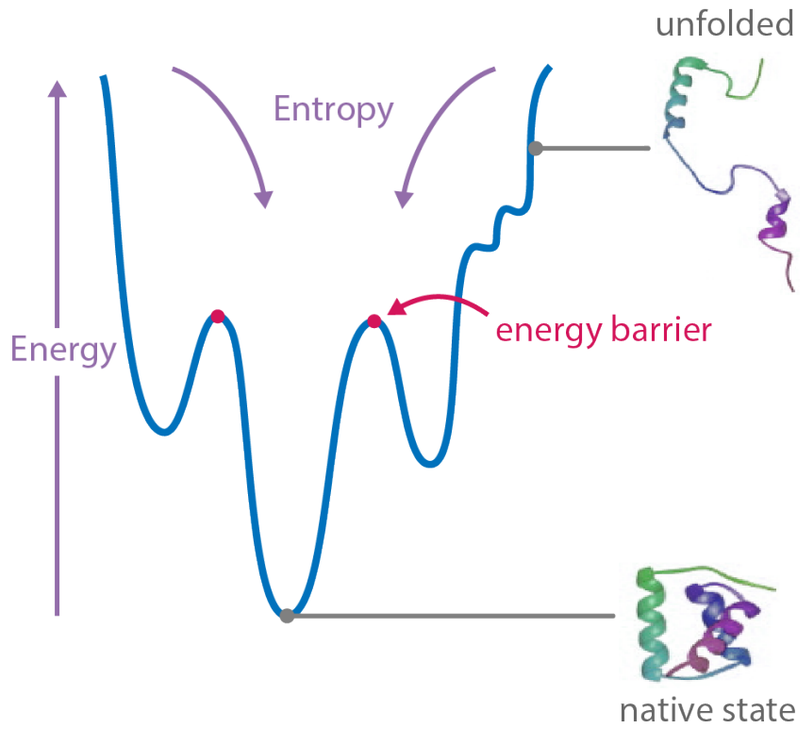
Folding



tertiary structure
(3D structure)

The folding process

- Proteins fold from a linear chain of amino acids into a stable 3D structure.
- Folding is driven by minimizing the protein's free energy to reach its native state.
- Proteins may encounter local energy minima (traps) and need to overcome these barriers to continue folding.



- Misfolding or aggregation occurs when proteins get stuck in non-native states, leading to dysfunction or diseases (Alzheimer's, Parkinson's, Cystic Fibrosis, Prion Diseases,...)

Knots in proteins

- For many years, it was believed that nature could not tie a protein knot, as the kinetic challenges (e.g., high activation barrier) involved in forming such a complex structure would not pass through the evolutional filter.

- Marc L. Mansfield suggested that proteins perhaps can be knotted (Nature, 1994).

"Sir – Most biochemists would probably agree that proteins in the native state are not knotted. The protein folding mechanism is not perceived as including repetitive, snake-like motions of the chain along its own contour."

"In summary, none of the 400 proteins structures analysed were found to have knots. Only one, human carbonic anhydrase B, comes close."

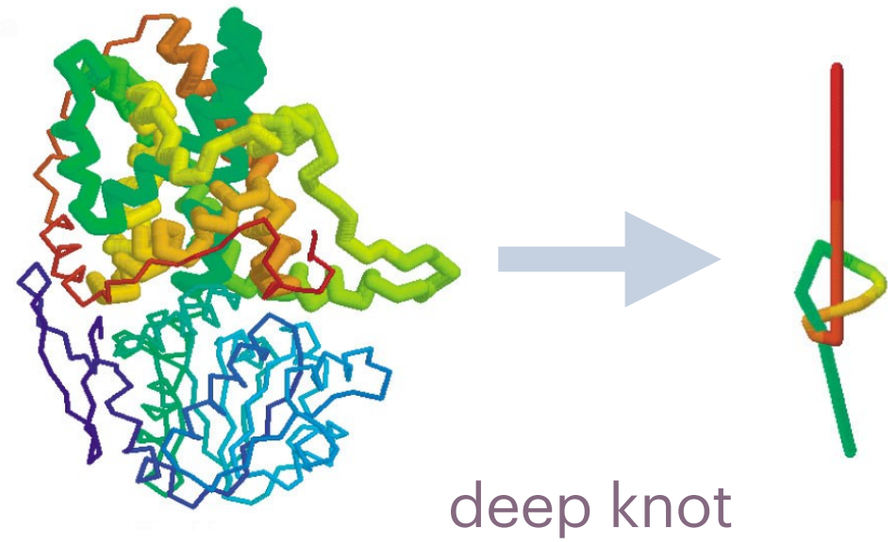
"The absence of knots in proteins would indicate that protein dynamics is non ergodic (all conformation are not accessible)."

"The most reasonable interpretation of these results is that the protein folding mechanism only explores unknotted conformations"

Knots in proteins

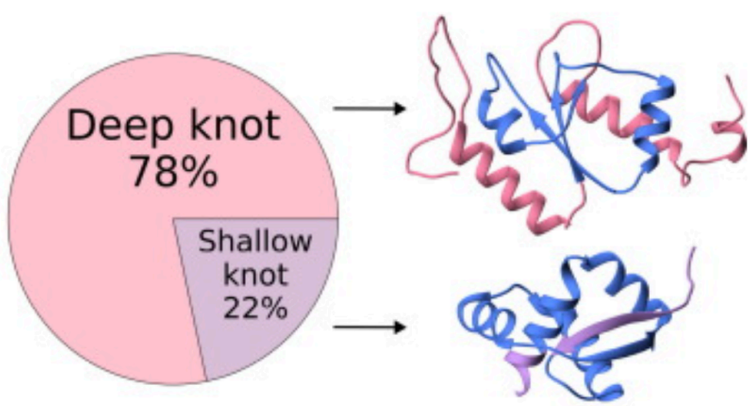
- The First knot 3_1 was confirmed in 3 months later (Liang, Mislow, 1994).
- In 2000 Taylor found nine other knotted proteins and the first deep knot.

shallow knot



deep knot

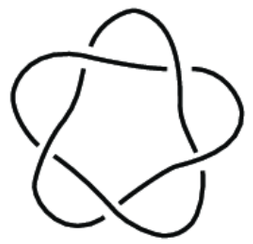
- 78% of knotted proteins are deeply knotted (Sulkowska, 2024)



The case of the human ubiquitin hydrolase

- The ubiquitin hydrolase is an enzyme that cuts ubiquitin chains from proteins, which either:
  rescues the protein from degradation, or recycles ubiquitin molecules for reuse.

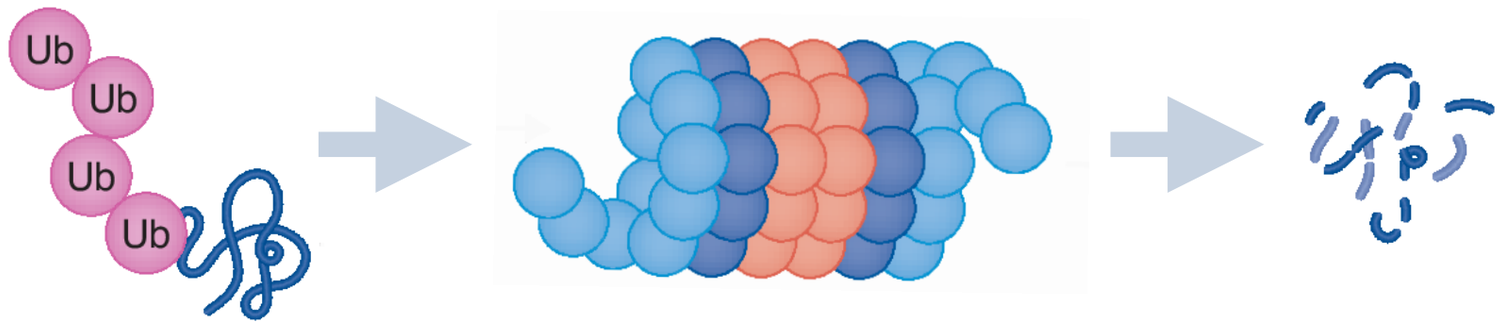
- The enzyme, which is usually in the proximity of the proteasome
  contains the complex 5_1 knot



- It is suggested that the knot prevents the enzyme to be pulled into the proteasome.

- Nobel prize 2004 "for the the discovery of ubiquitin-mediated degradation".

ubiquitin hydrolase
(UCHL)



proteasomal degradation
(the ubiquitin-proteasome system)

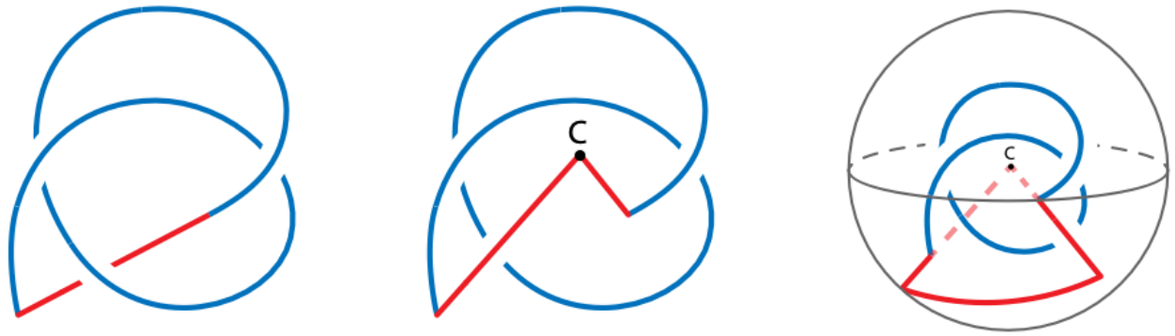
Fun fact: our brain contains 2% of UCHL.

How do we classify a protein as knotted?

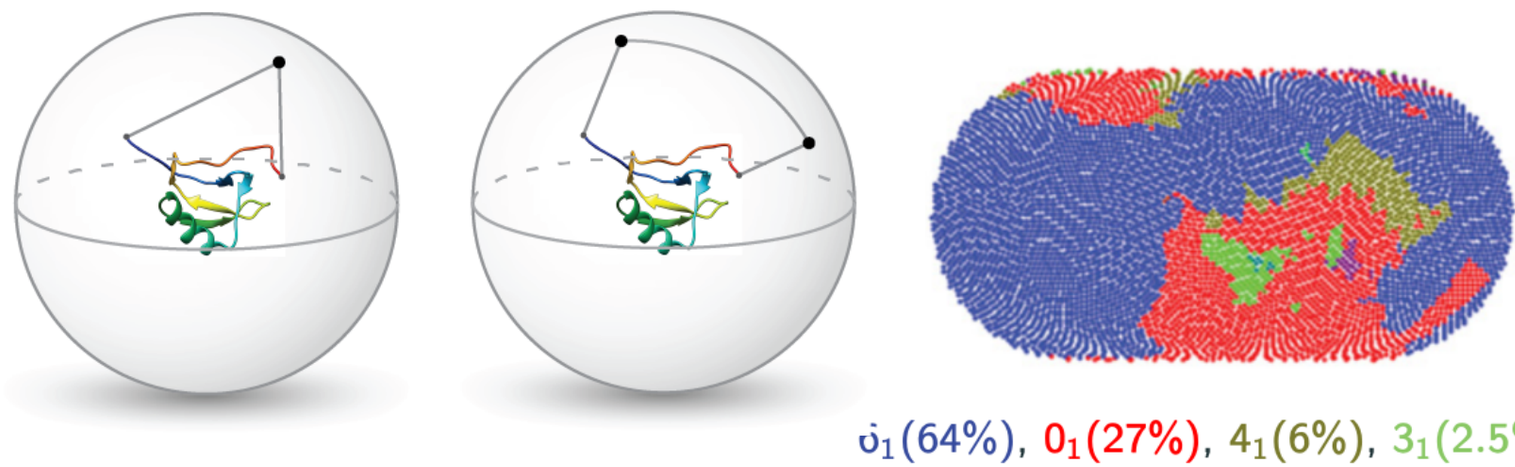
- The protein backbone is an open interval in 3-space. To study knots, we need to close the curve (connect the endpoints) to obtain a closed curve

- A unique method for closure remains an open question. Two approaches are used:

  - direct closures (Virnau, Mirny,...)



  - stohastic closures (Sulkowka, Millet,...)



Big data approach to knots in proteins

- Experimental structures (X-ray Crystallography, NMR, Cyro-EM,...)

RCSB PDB
PROTEIN DATA BANK

contains approx. 200k structures

- AI predicted structures

AlphaFold

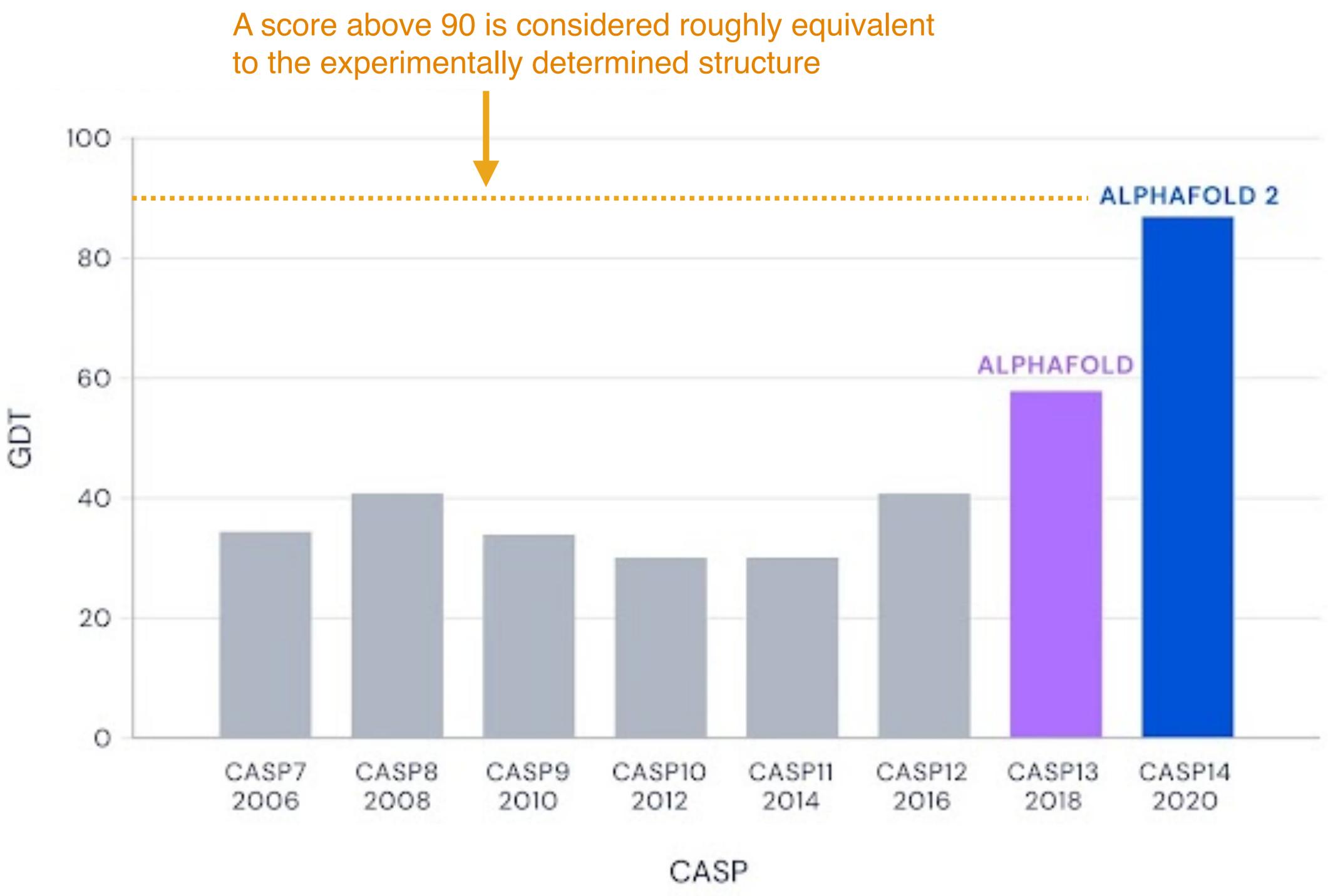
DeepMind
Google

approx. 200M structures

- The protein folding prediction is referred to as the "50 year open research problem".

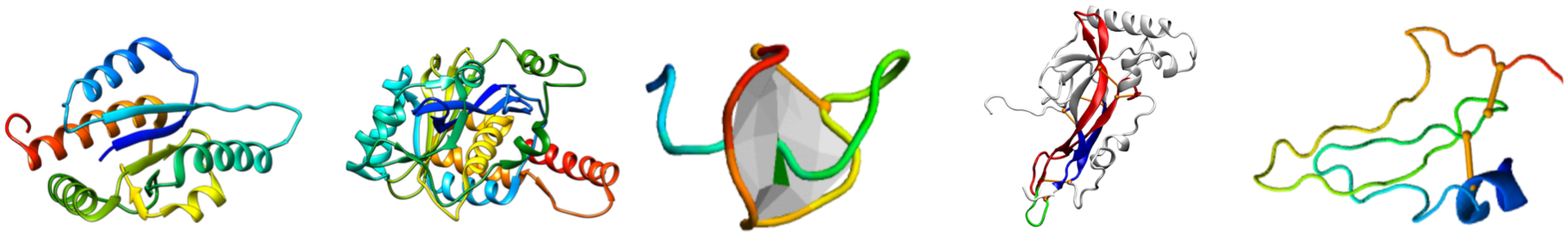
- A typical protein has around 10^{300} different configurations (the the universe is $4 \cdot 10^{17}$ seconds old).

AI prediction accuracy



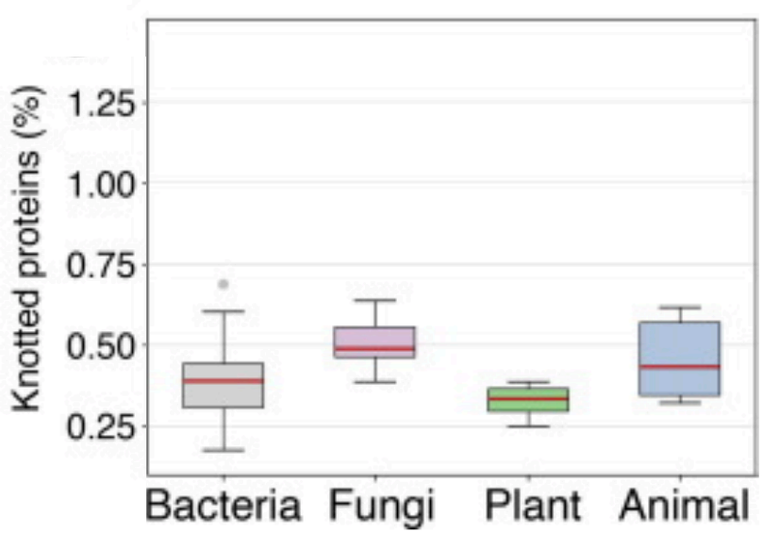
Knotted protein structures

- At least 8% of known proteins are entangled.

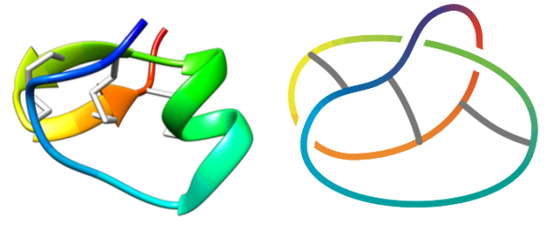


knot slipknot lasso θ-curve deterministic link

- Less than 1% of known proteins are knotted
  KnotProt confirmes 800 - 2100 knots,
  AlphaKnot predicts 341 - 1144494 knots.

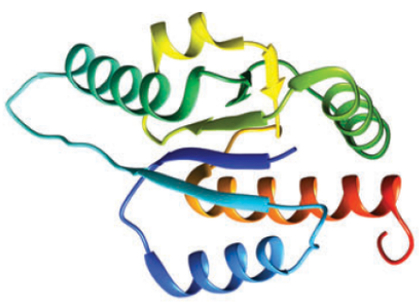


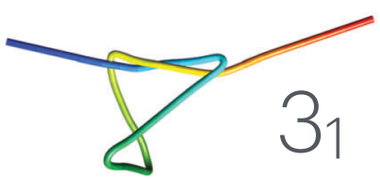
- If we expand the definition of entanglement with bonded knots, more than 20% of proteins are entangled (G.).



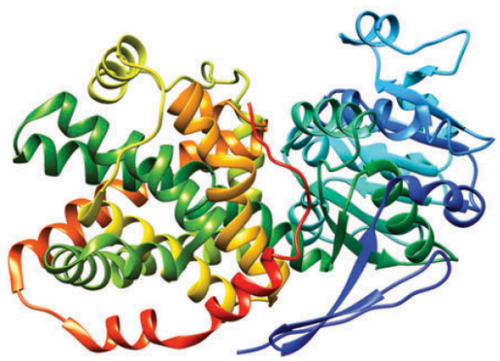
bonded knot

What knots have we found so far?

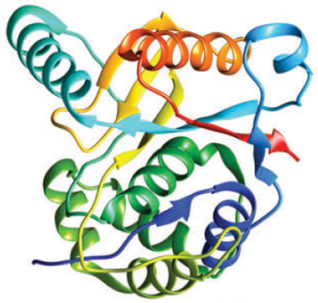




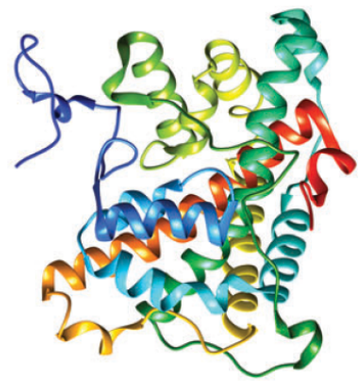
3_1

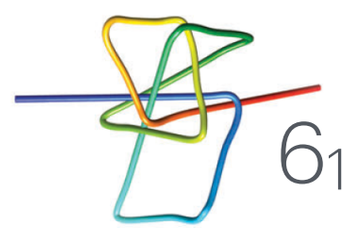


4_1



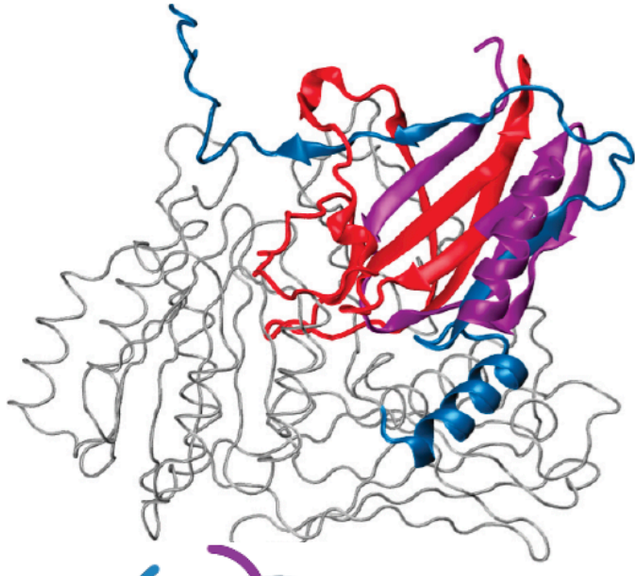
5_2





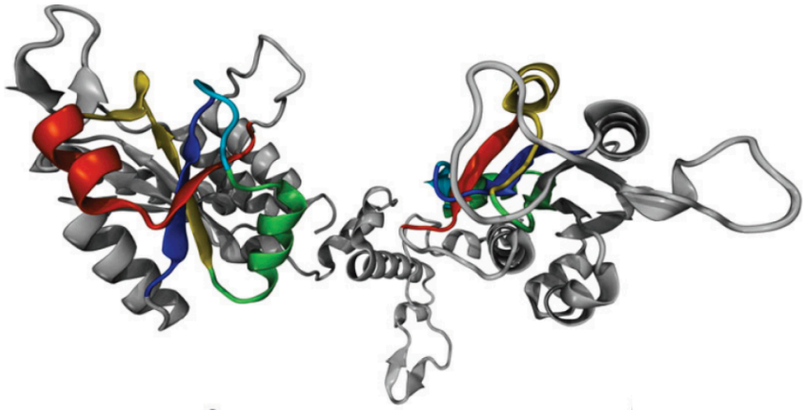
6_1

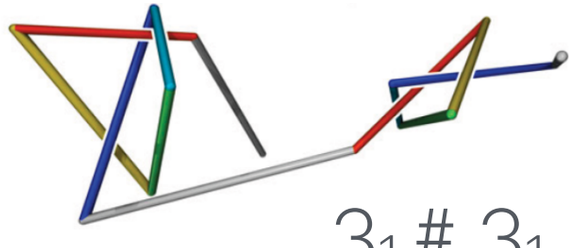
Sulkowska et al. (2022)



6_3

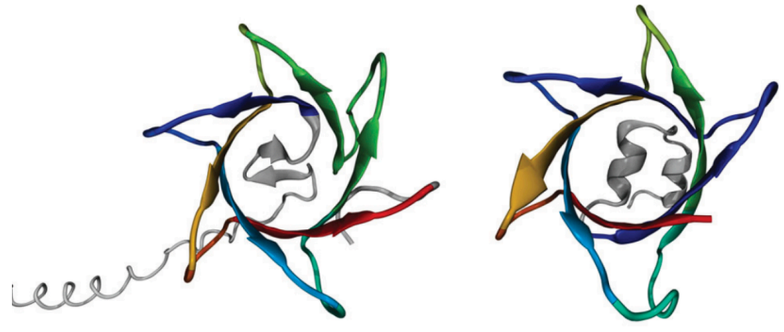
Virnau et al. (2021)

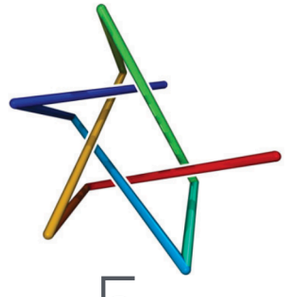




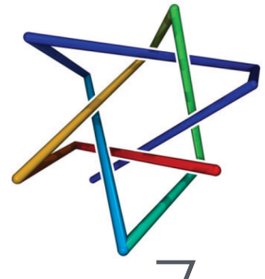
$3_1 \# 3_1$

Virnau et al. (2021)





5_1



7_1

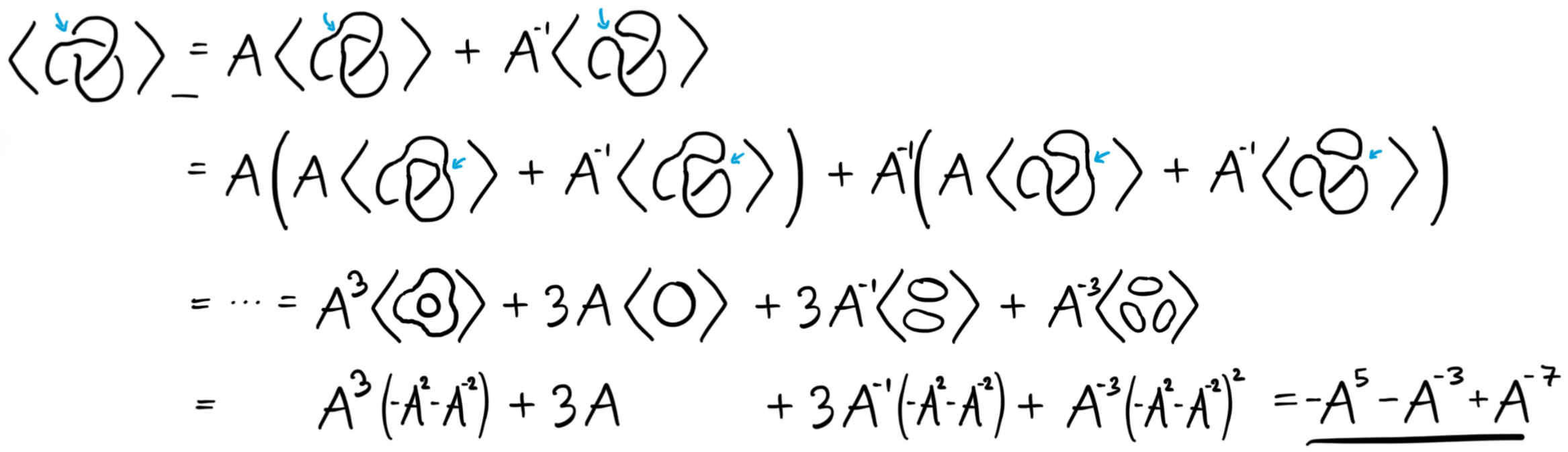
Classification of knotted structures

We focus on the following ways to detect knotted structures in proteins:

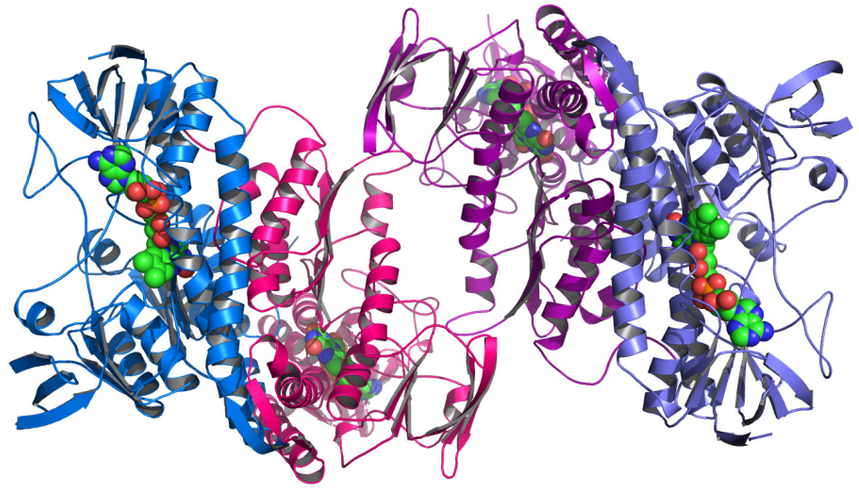
- mathematical invariants
- machine learning
- persistant homology

The Bracket polynomial

- The Bracket polynomial $\langle K \rangle$ of a knot diagram K is a polynomial in variable A obtained by the rules:
  1. $\langle \bigcirc \rangle = 1$
  2. $\langle \times \rangle = A\langle \;) \; (\; \rangle + A^{-1}\langle \;) \; (\; \rangle$
  3. $\langle \bigcirc \cup L \rangle = (-A^2 - A^{-2})\langle L \rangle$



Speeding up the computations



- Time complexity of computing the bracket is $O(n^2)$ where n is the number of crossings.

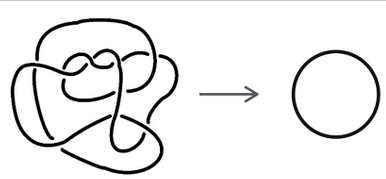
- Proteins can have very complex geometric structures and range in size from tens to several thousand amino acids.

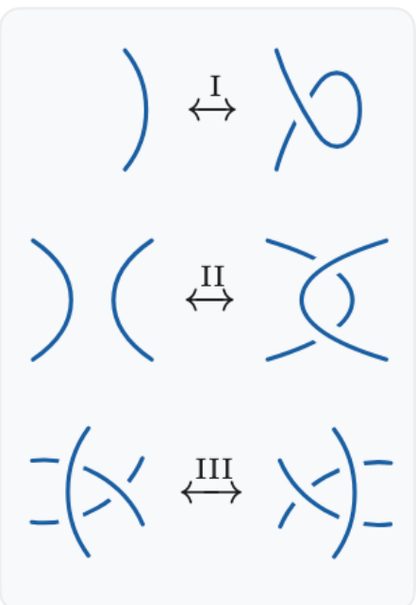
- **Improvements:**

- Put each diagram into canonical form ($O(n)$ with Ewing-Millett notation)

- Use memoization for smaller diagrams

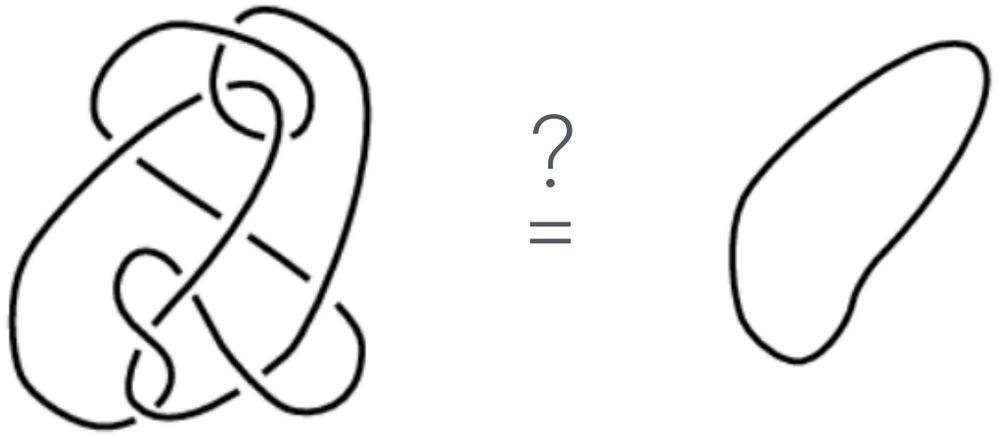
- simplify the diagram at each step (with a smart Reidemeister search)

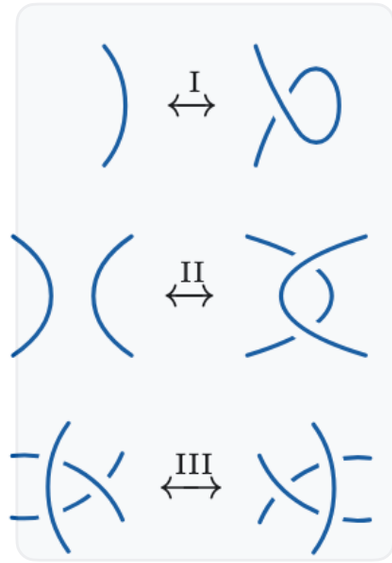




Reidemeister moves

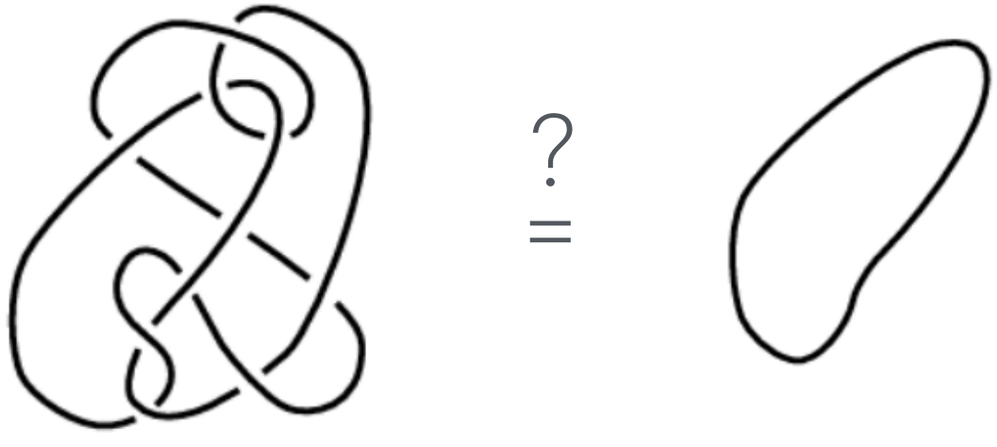
Using Reidemeister moves to simplify a diagram

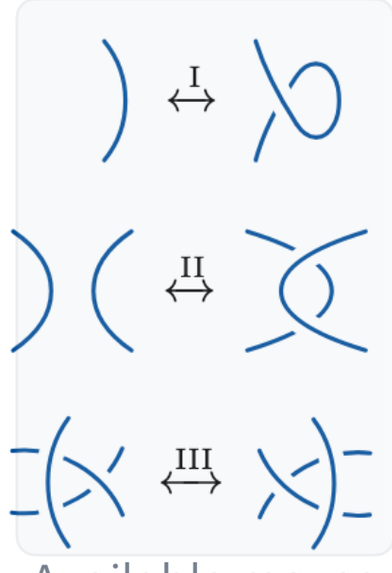




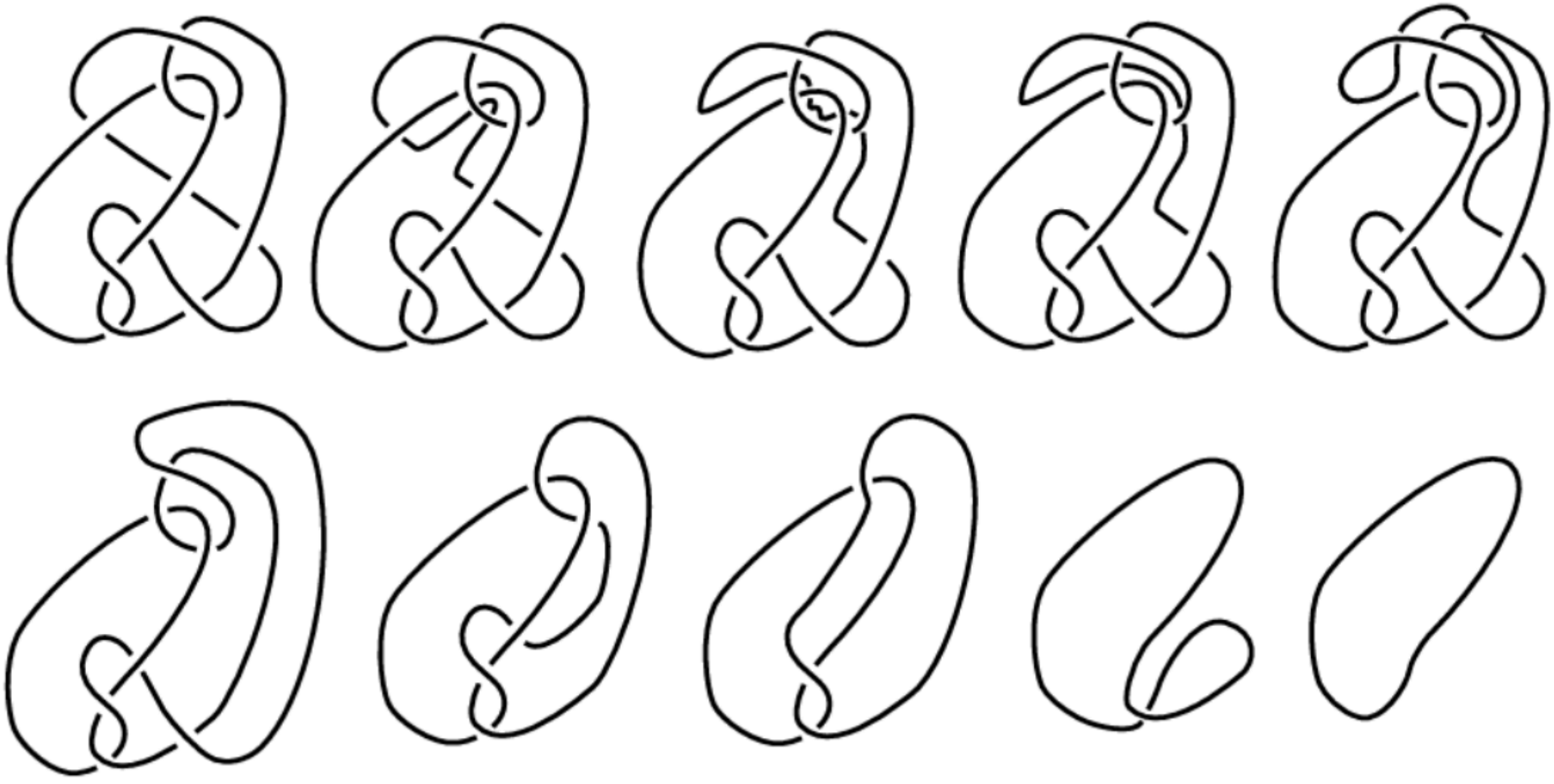
Available moves

Using Reidemeister moves to simplify a diagram





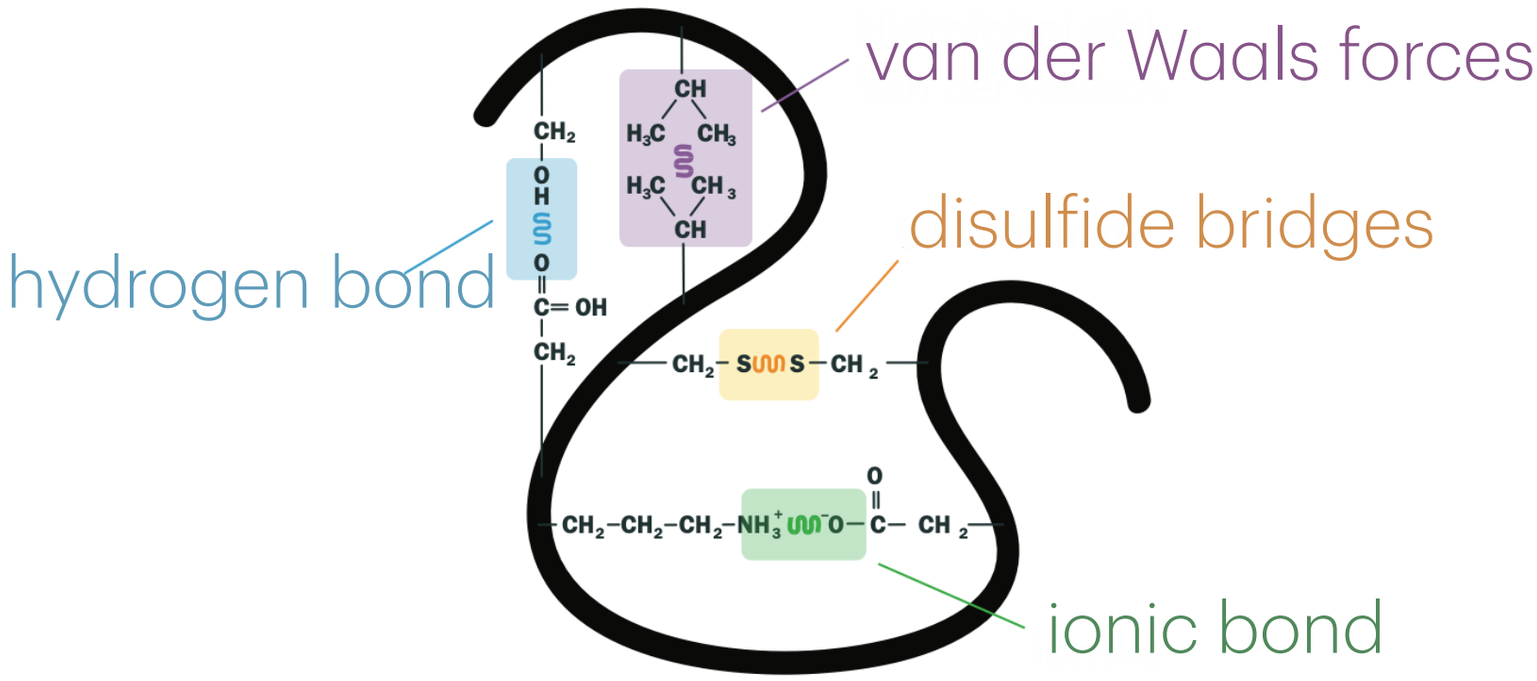
Available moves



Protein bonds

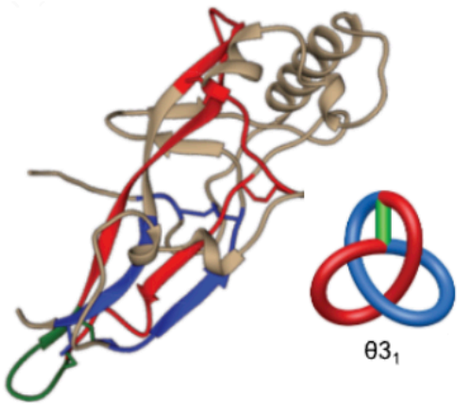
- Peptide Bonds are covalent bonds that link amino acids together to form the protein's backbone

- Non-covalent Bonds (hydrogen bonds, disulfide bridges, ionic bonds, van der Waals forces, and hydrophobic interactions) stabilize the protein's three-dimensional structure, enabling proper folding and function.

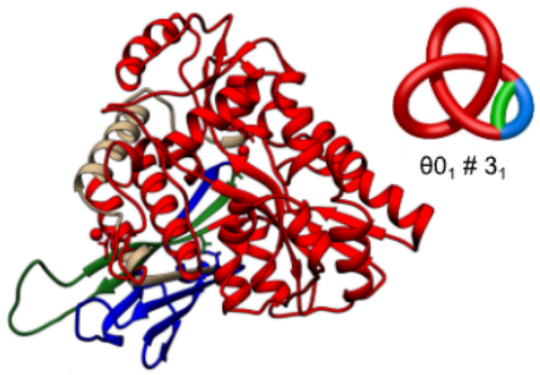


Theta curves and bonded knots

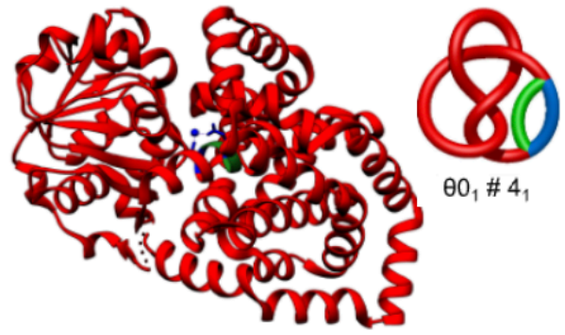
- 7 types of θ-curves have been identified in the Protein Data Bank (PDB) a publicly accessible database of experimentally determined protein, nucleic acid, and complex biomolecular structures (Sulkowska et al., 2024).



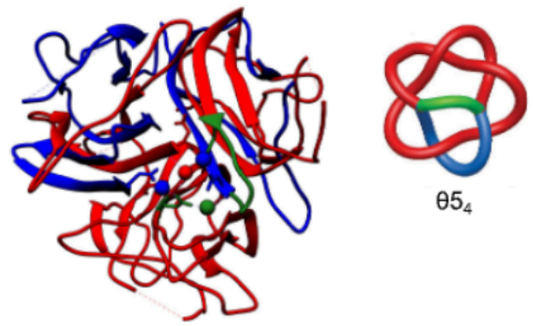
$\theta 3_1$



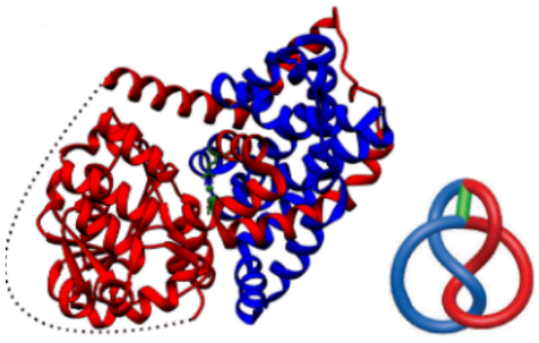
$\theta 0_1 \# 3_1$



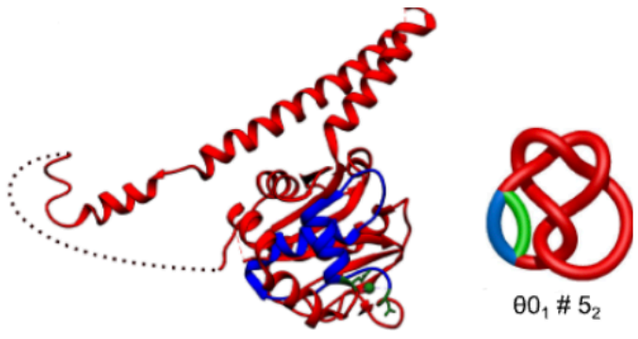
$\theta 0_1 \# 4_1$



$\theta 5_4$



$\theta 4_1$

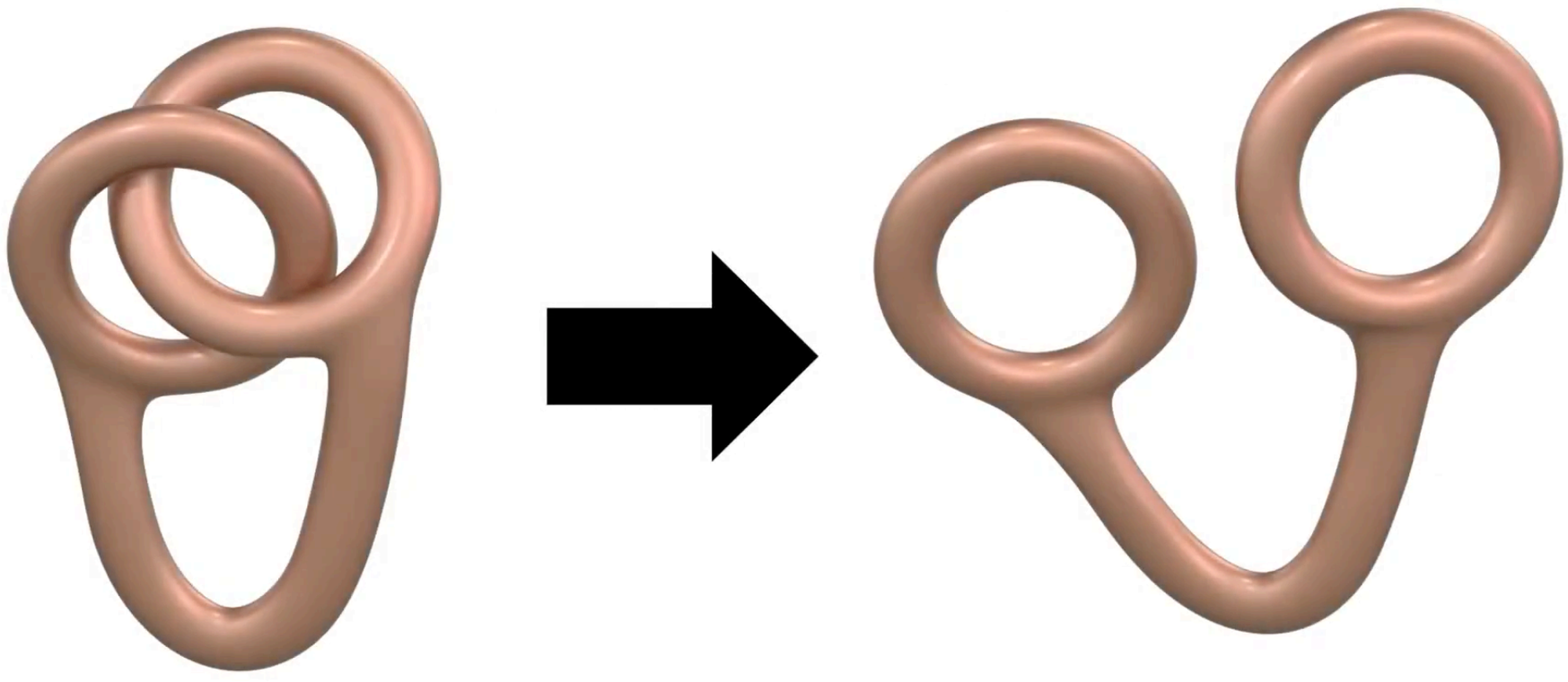


$\theta 0_1 \# 5_2$

$\theta 8_n$

Why topological invariants fall short

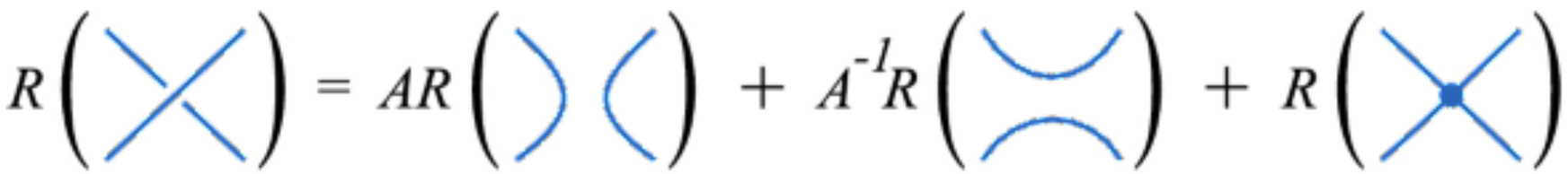
- Most topological invariants fail to detect knottedness of graphs (e.g. θ-curves, handcuffs, bonded knots), the underlying reason being, that they are usually defined via the complement $\mathbb{R}^3 \backslash G$ (G., 2021).

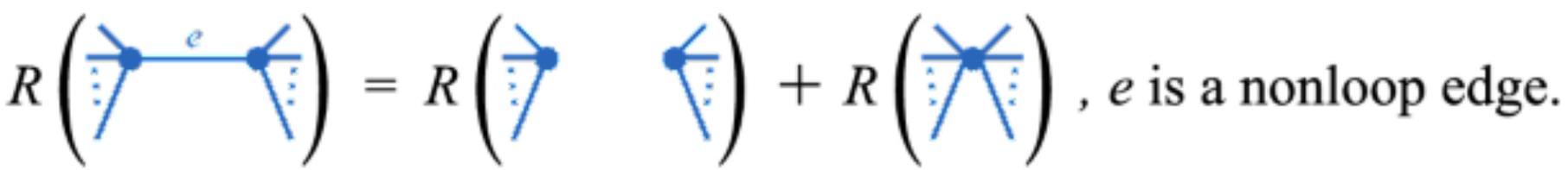


Source: Youtube, Keenan Crane - Computer Science & Robotics, CMU, Pittsburgh, Pennsylvania.

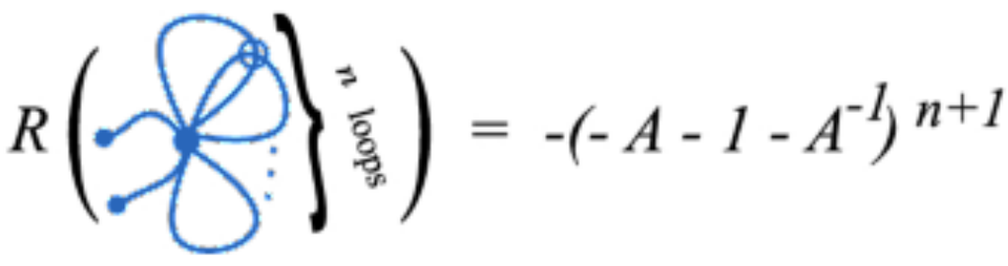
The Yamada polynomial

- The Yamada polynomial is an invariant of (embedded) spatial graphs and is defined by the following rules:

(Y1) 

(Y2) 

(Y3) $R(G \sqcup G') = R(G)R(G')$

(Y4) 

The Yamada polynomial

- The Yamada polynomial is an invariant of (embedded) spatial graphs and is defined by the following rules:

(Y1) $R\left(\times\right) = AR\left)\ (\right) + A^{-1}R\left(\asymp\right) + R\left(\times\right)$

(Y2) $R\left(\overset{e}{\text{—}}\right) = R\left(\ \ \ \right) + R\left(\ \ \ \right)$, e is a nonloop edge.

(Y3) $R(G \sqcup G') = R(G)R(G')$

(Y4) $R\left(\left\{\ \right\}^{n \text{ loops}}\right) = -(-A - 1 - A^{-1})^{n+1}$

Relation from the chromatic polynomial

Relation from the Bracket polynomial

- The time complexity of the Yamada polynomial is $O(3^n)$ and is often not practical to use.

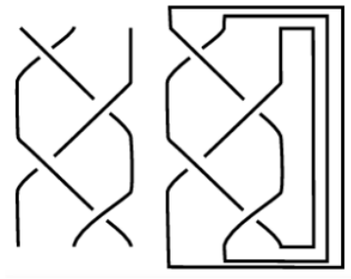
- The Yamada was used to some degree to detect θ-curves and handcuff links in proteins.

- We can speed up Yamada using similar techniques than the Bracket.

Detecting knots using Machine Learning

- L. Dai, O. Vandans, et. al (2020) showed that LSTM-based RNNs are able to predict the knot type for (non-protein like) polymers with 99% accuracy.

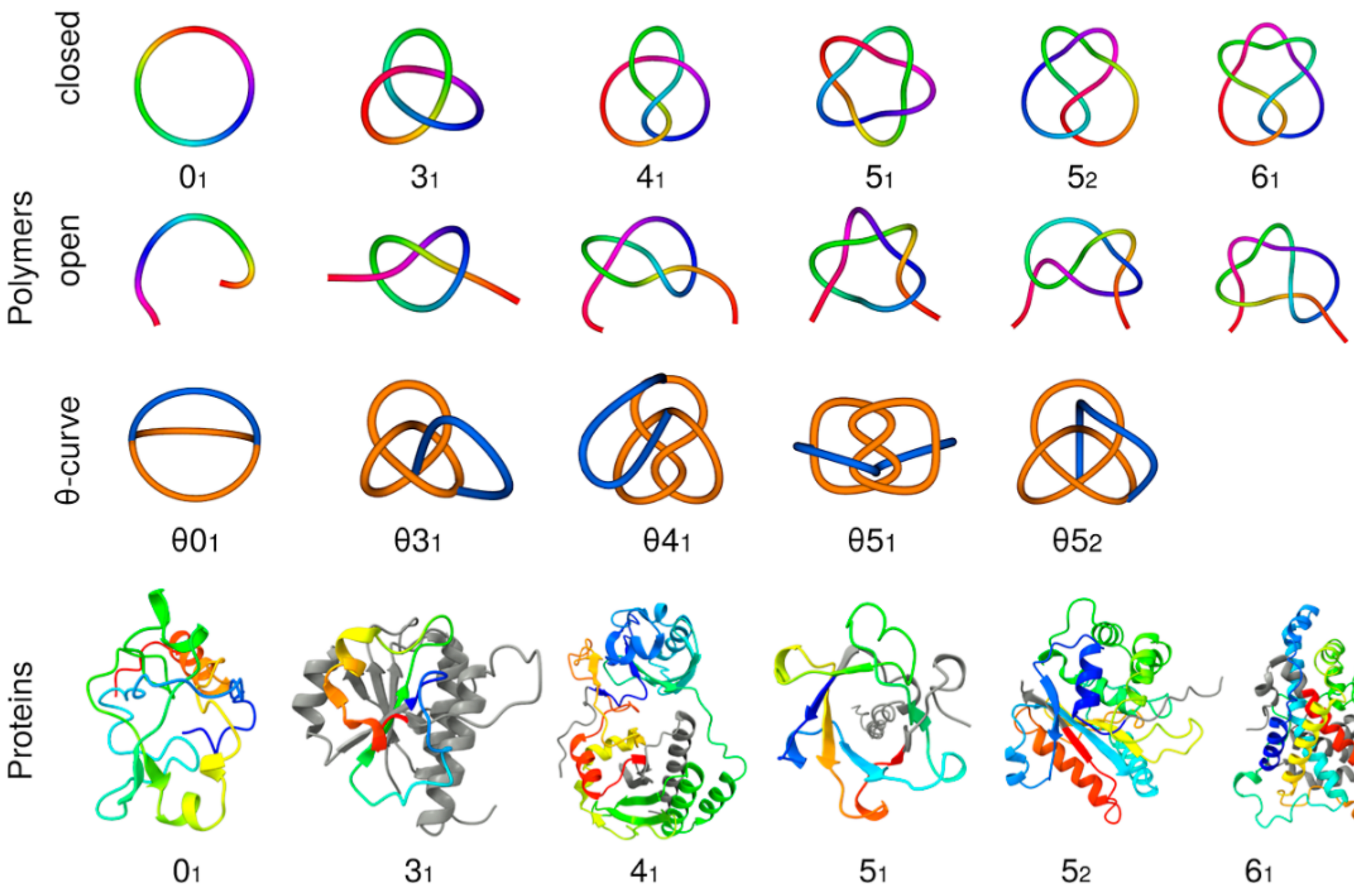
- P. Sułkowski, S. Gukov, et al (2020) used a shared-QK Transformer network architectures to detect the unknot in braided form with 93% - 100% accuracy.



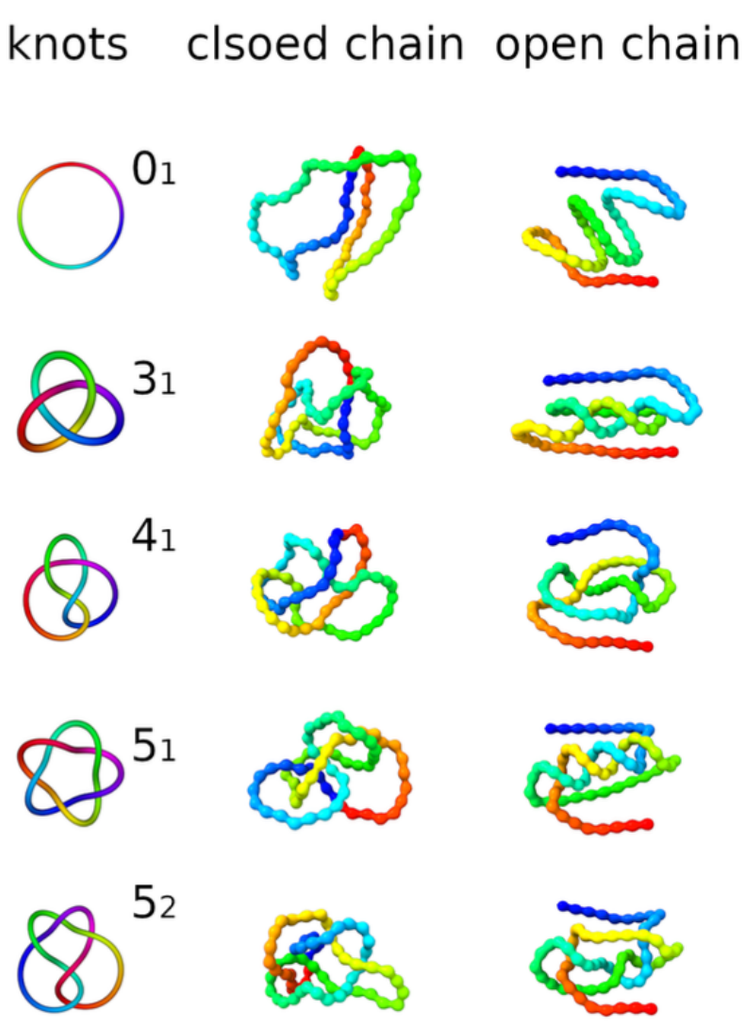
braided form of the figure 8-knot

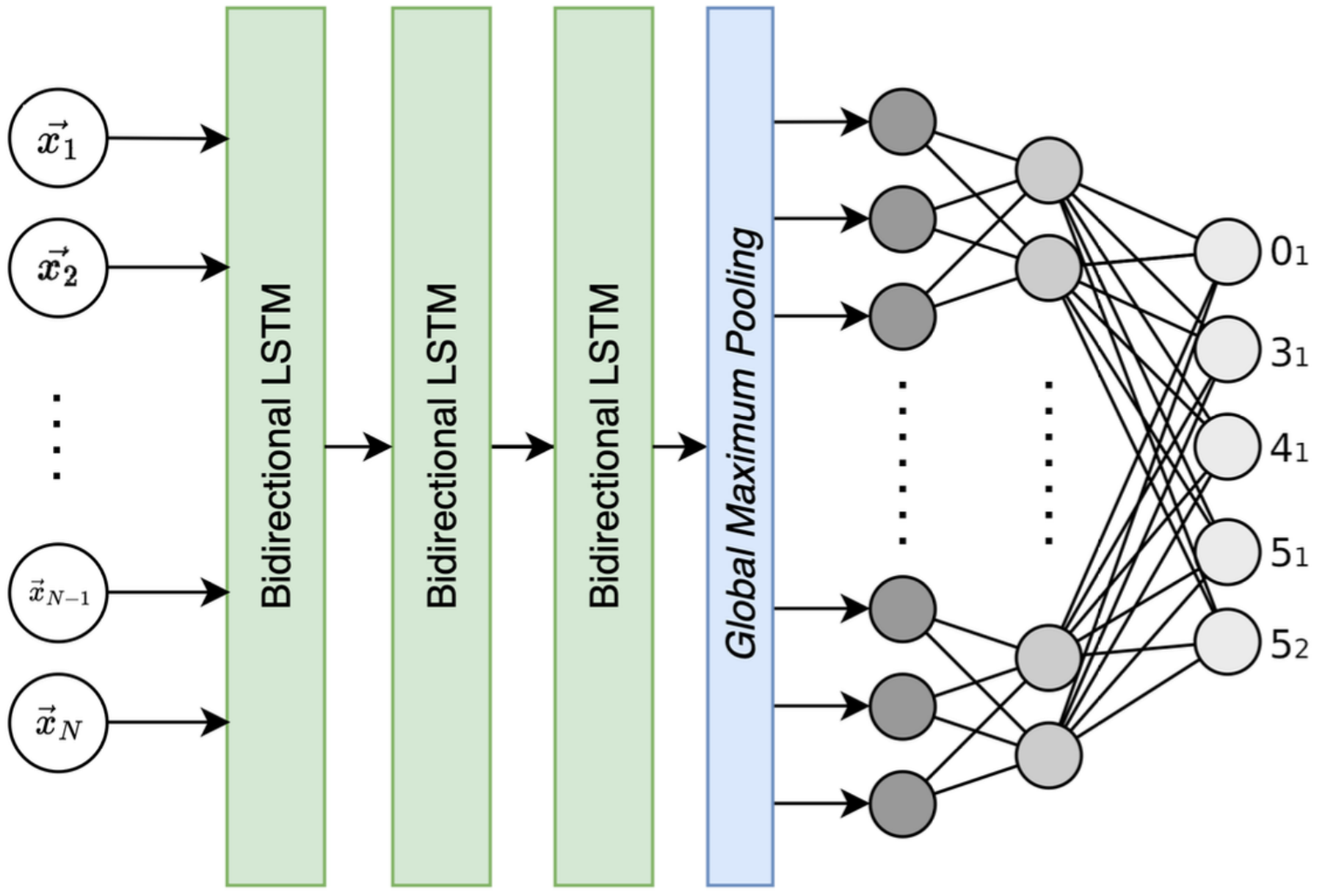
- J. I. Sulkowska, G., et al. (2024) showed that LSTM-based RNNs are able to predict knots, open knots, and θ-curves for (non-protein like) polymers, protein-like polymers, and proteins with 93% - 99% accuracy.

ML knot classification of polymeric and protein structures



The model


knots
clsoed chain
open chain
0_1
3_1
4_1
5_1
5_2


$\vec{x}_1$
$\vec{x}_2$
$\vec{x}_{N-1}$
$\vec{x}_N$
Bidirectional LSTM
Bidirectional LSTM
Bidirectional LSTM
Global Maximum Pooling
0_1
3_1
4_1
5_1
5_2

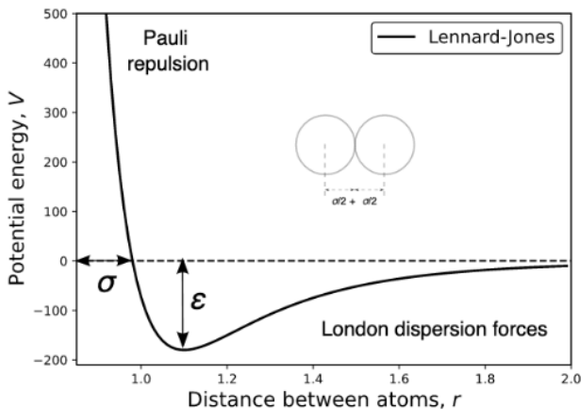
Generating training data

- We generate training, test and validation sets using Molecular Dynamics (MD) simulations.

- Systems of closed knots and open knots consisted of 64, 128, and 256 beads; θ-curves and composite θ-curves were made up of 92, 188, and 286 beads

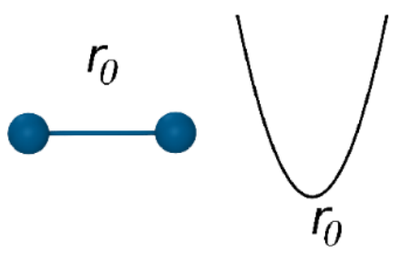
- **Polymer simulations:** we introduce repulsive and binding potentials.

- **Protein-like simulations:** we introduce repulsive, binding, angle, and dihedral angle potentials.



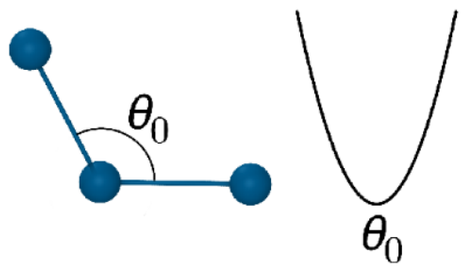
repulsive potential

$V_{Bond} = k_b(r_{ij} - r_0)^2$



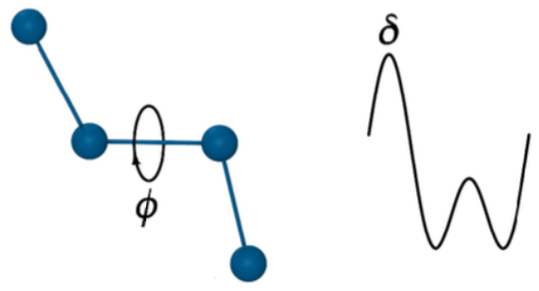
binding potential

$V_{Angle} = k_\theta(\theta_{ijk} - \theta_0)^2$



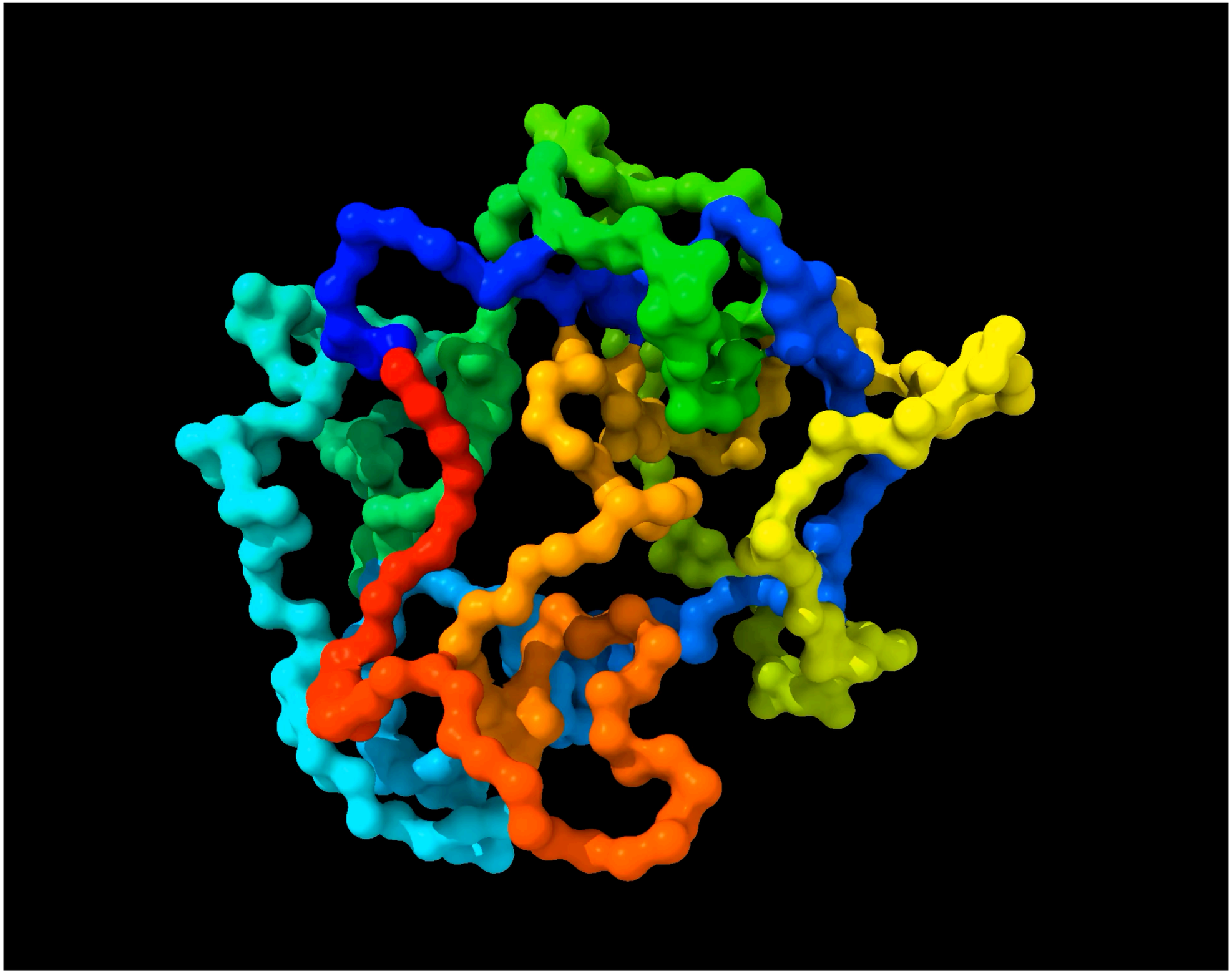
angle potential

$V_{Dihed} = k_\varphi(1 + \cos(n\varphi - \delta)) + \ldots$

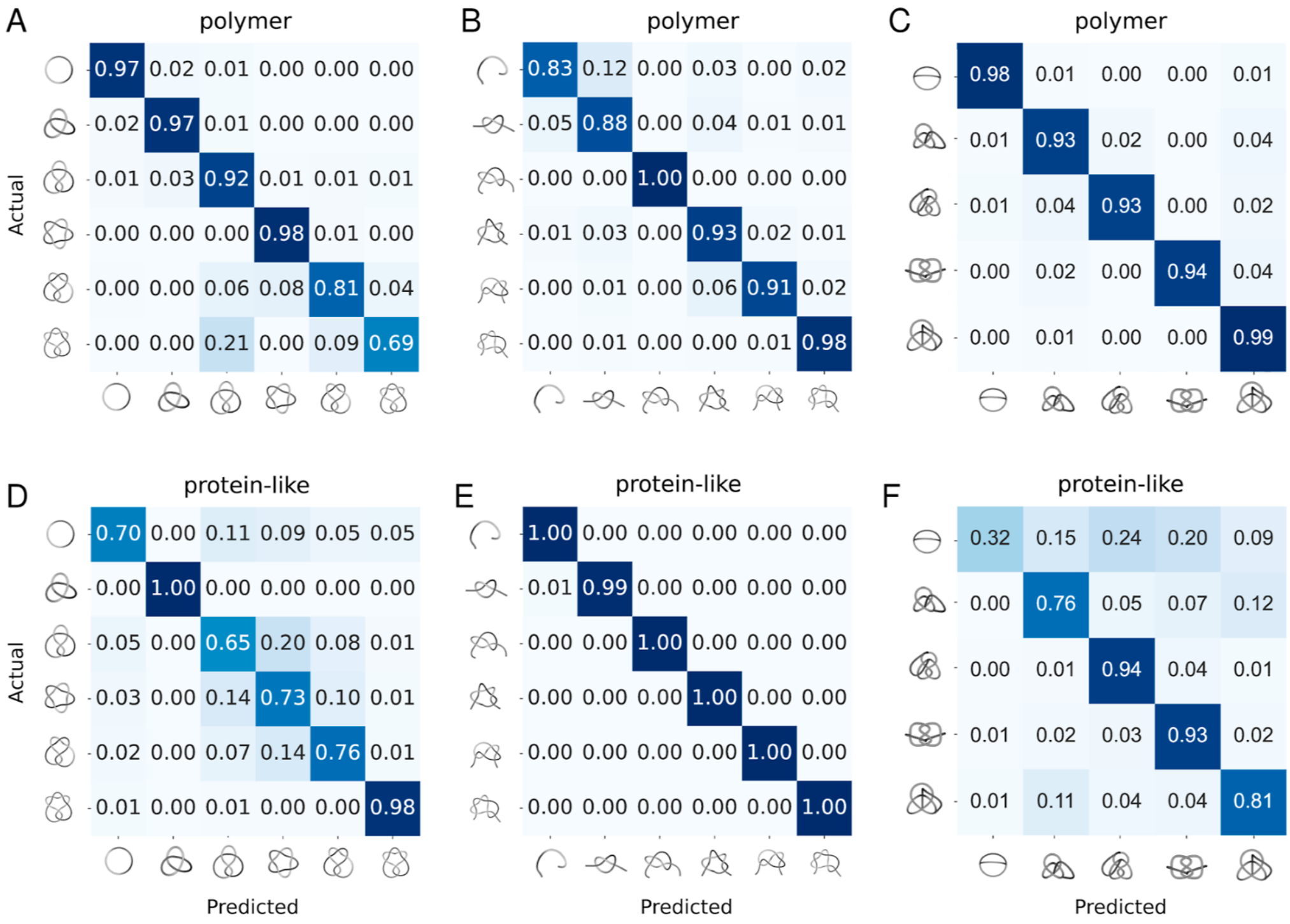


dihedral potential

Molecular dynamics simulation

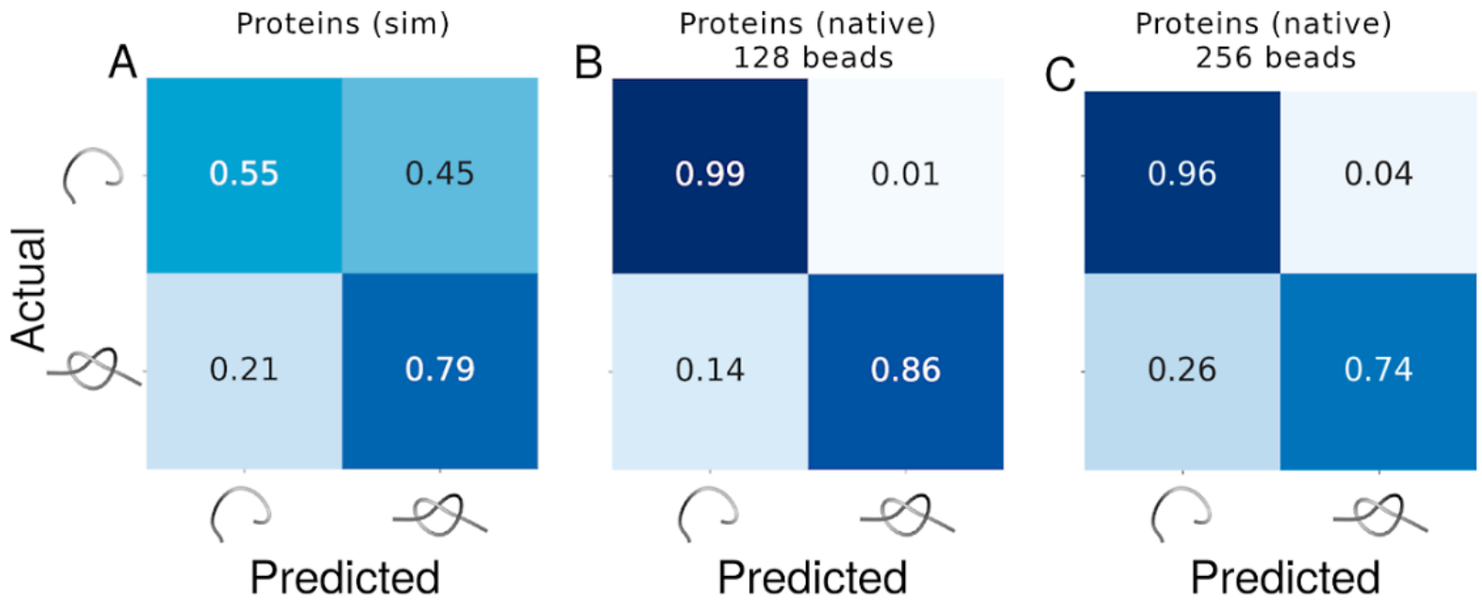


Results for polymer and protein-like simulations



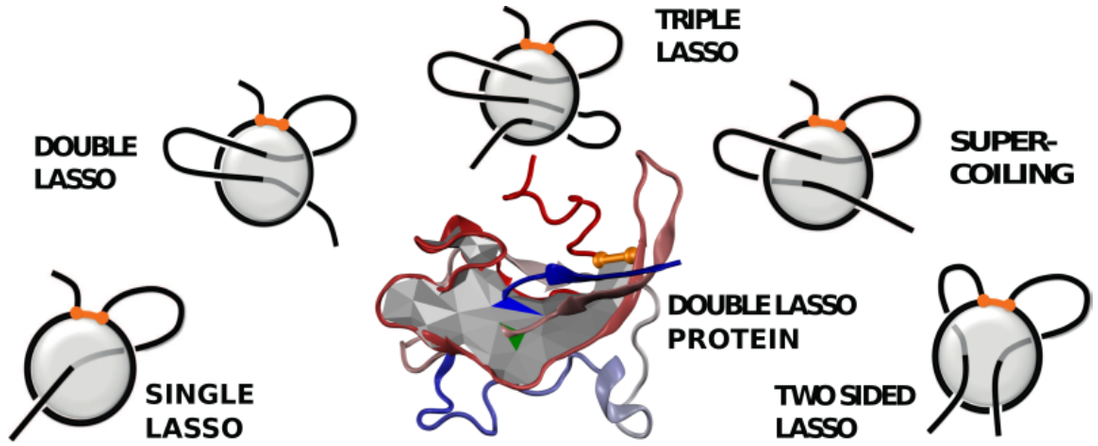
Confusion matrices

Results for proteins

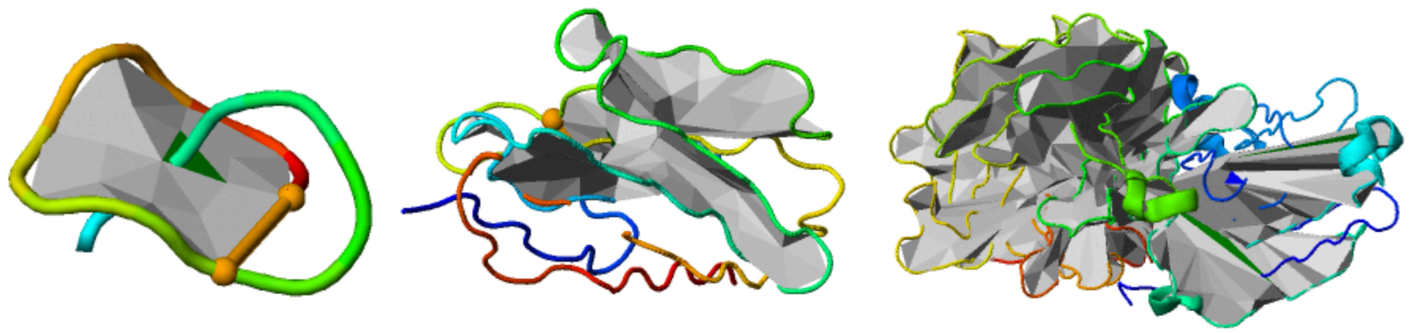


Confusion matrices

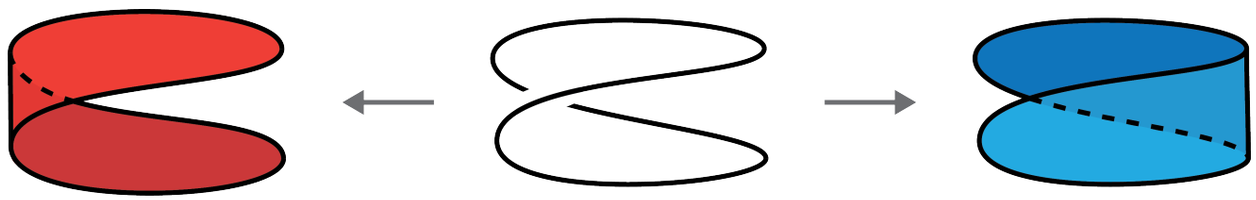
Lassos in proteins



- About 15% of lassos in LassoProt are non-trivially looped
- Traditional approach involves computing the minimal surface and piercings of the tail
- Problems:
  - diverse lasso structures

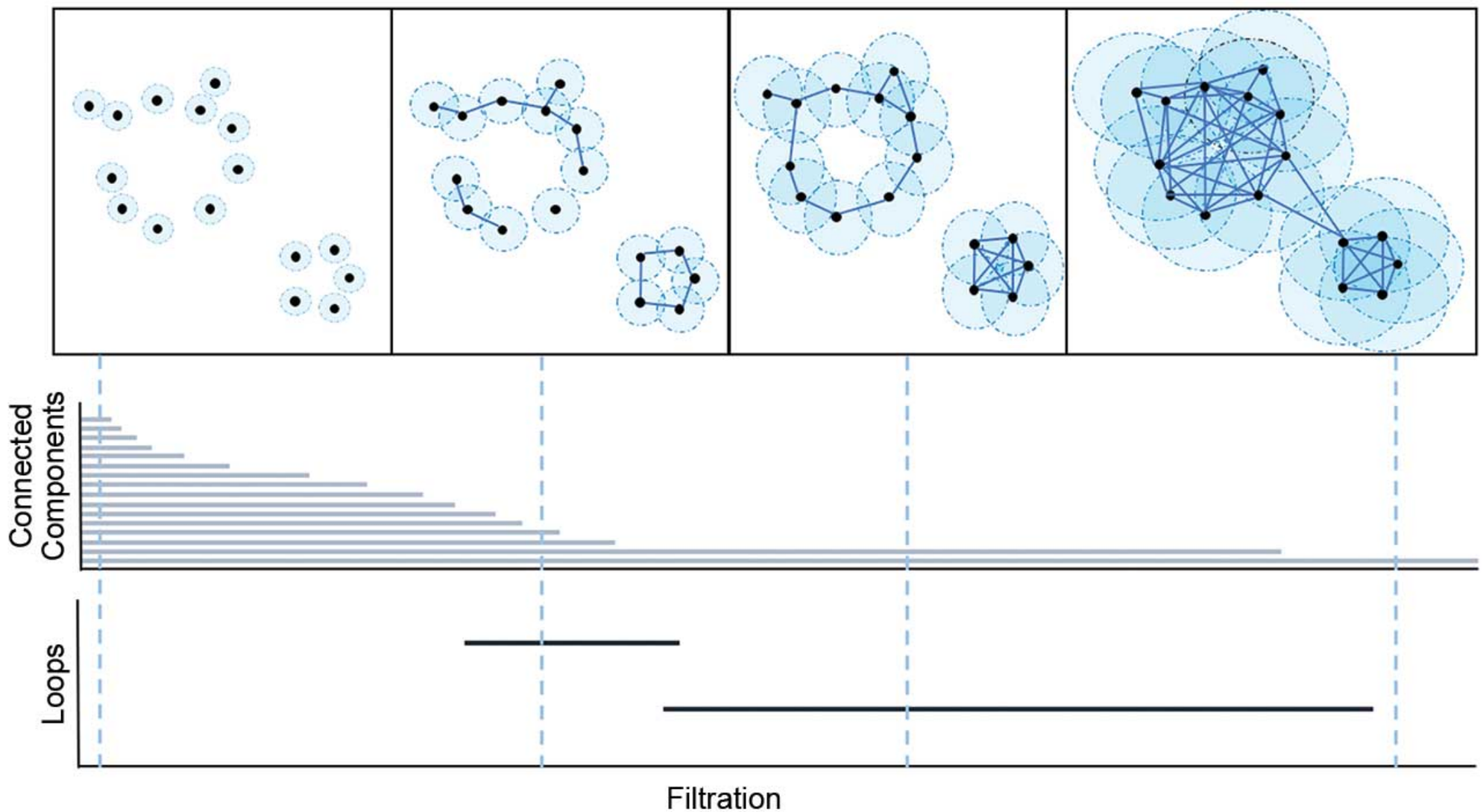


  - computation of the minimal surface can be slow
  - the minimal surface is not stable (and sometimes not well-defined)



Persistent homology (PH)

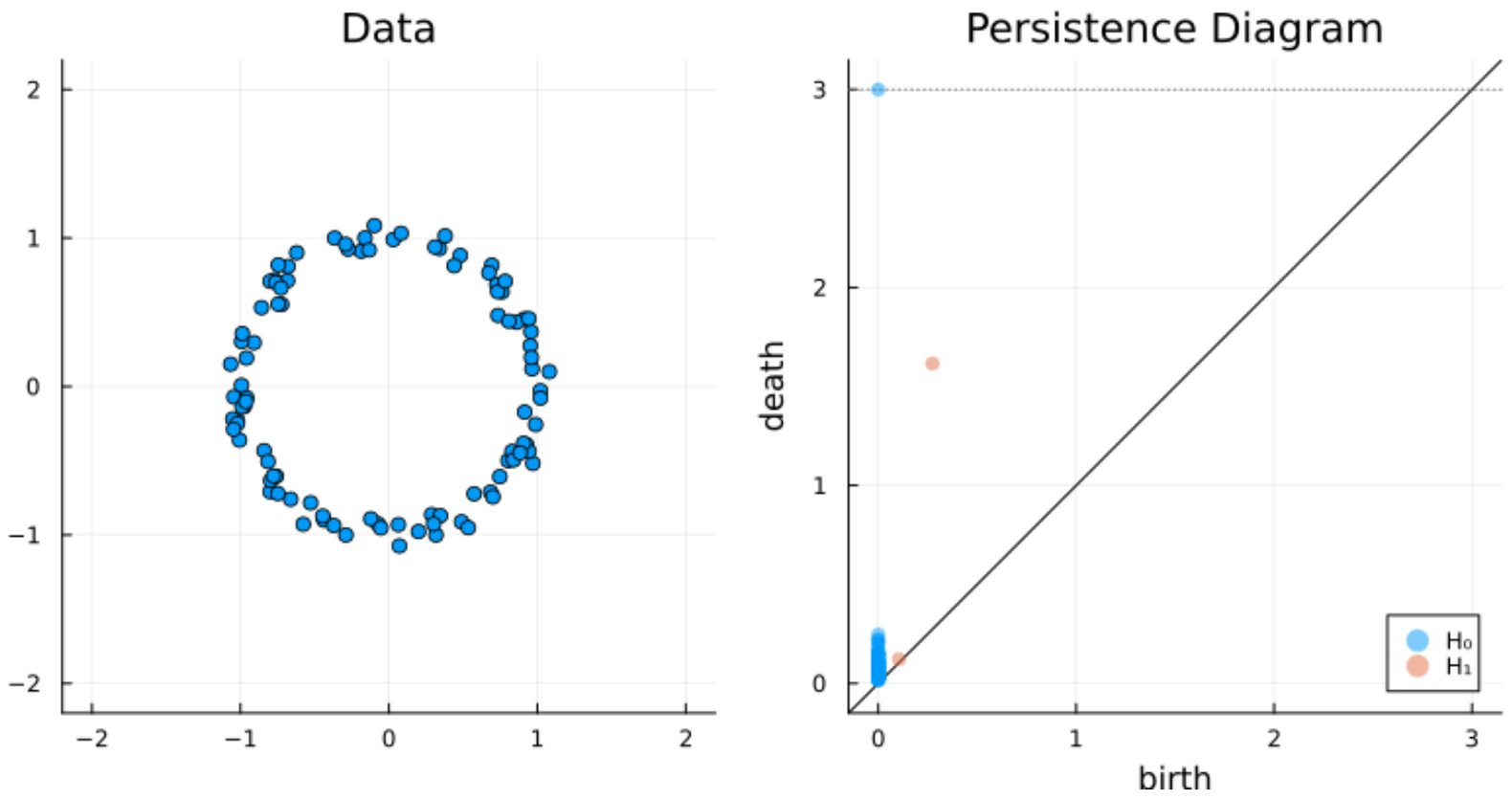
- PH is a method in Topological Data Analysis (TDA) that tracks topological features (e.g. components, holes, voids) across multiple scales by which we can identify and quantify meaning patterns that persist across ranges of scales.



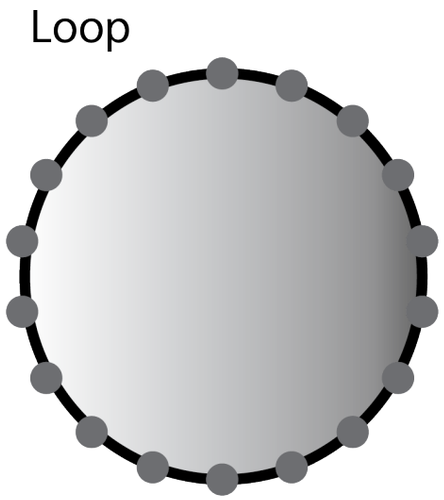
Stability of Persistent Homology

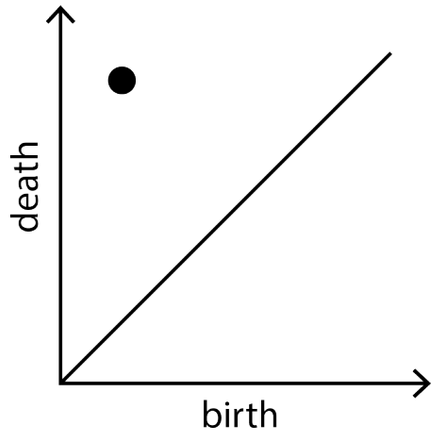
Stability theorem (Cohen-Steiner/Edelsbrunner/Harer 2007). Let X and Y be two finite metric spaces, and let $d_{GH}(X, Y)$ denote their Gromov-Hausdorff distance. If $PH(X)$ and $PH(Y)$ denote the persistence diagrams of the corresponding filtrations, then the bottleneck distance d_B between the persistence diagrams satisfies:

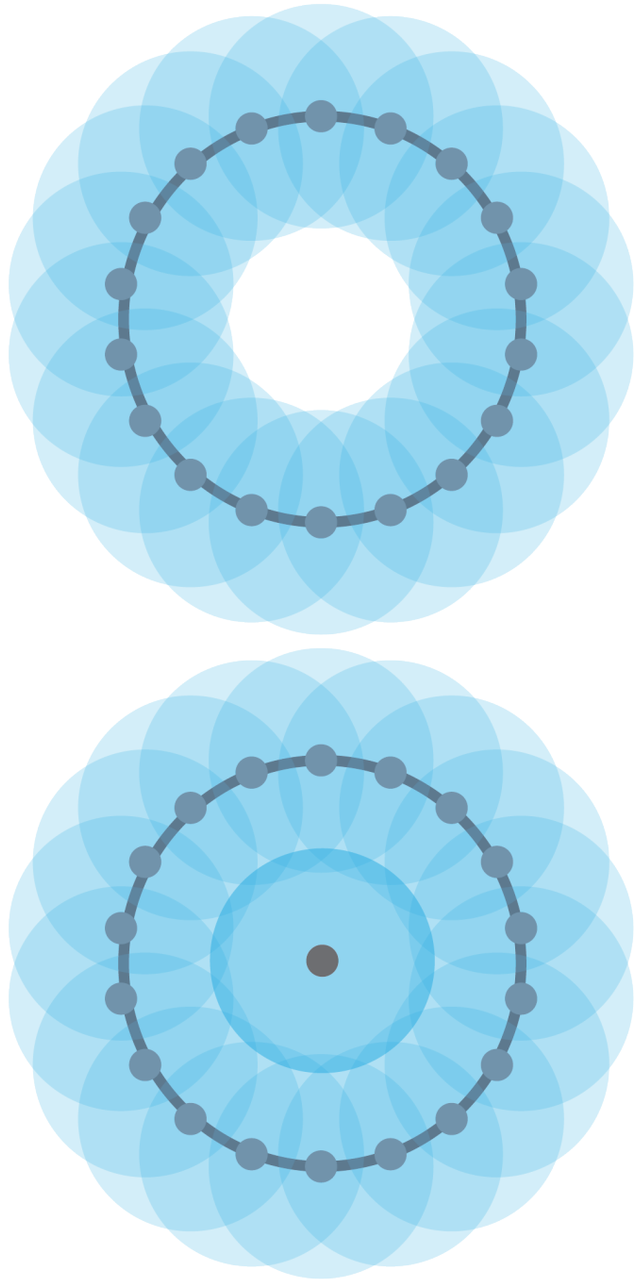
$$d_B\,(\mathrm{PH}(X), \mathrm{PH}(Y)) \leq 2 d_{GH}(X, Y)$$

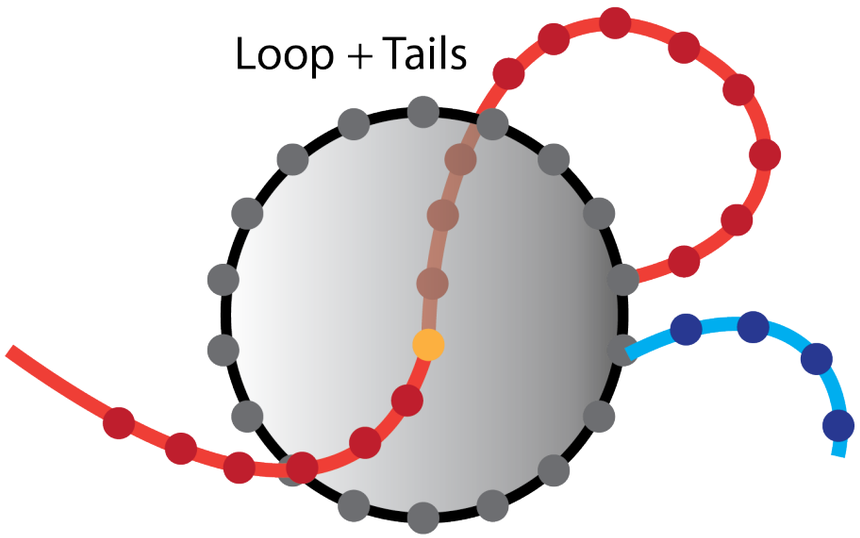


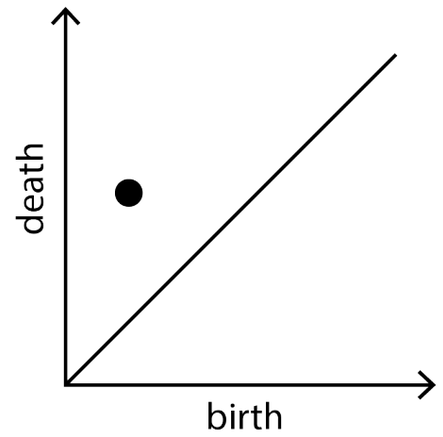
How does PH detect lassos?



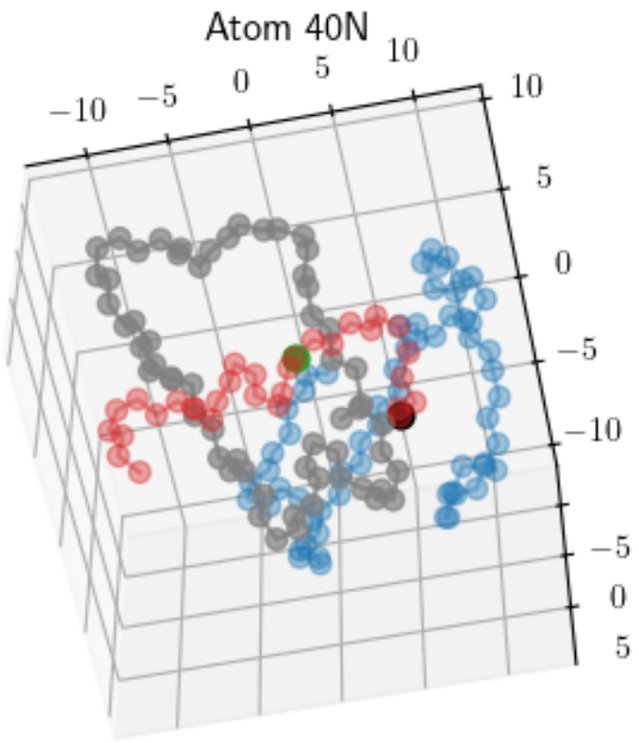


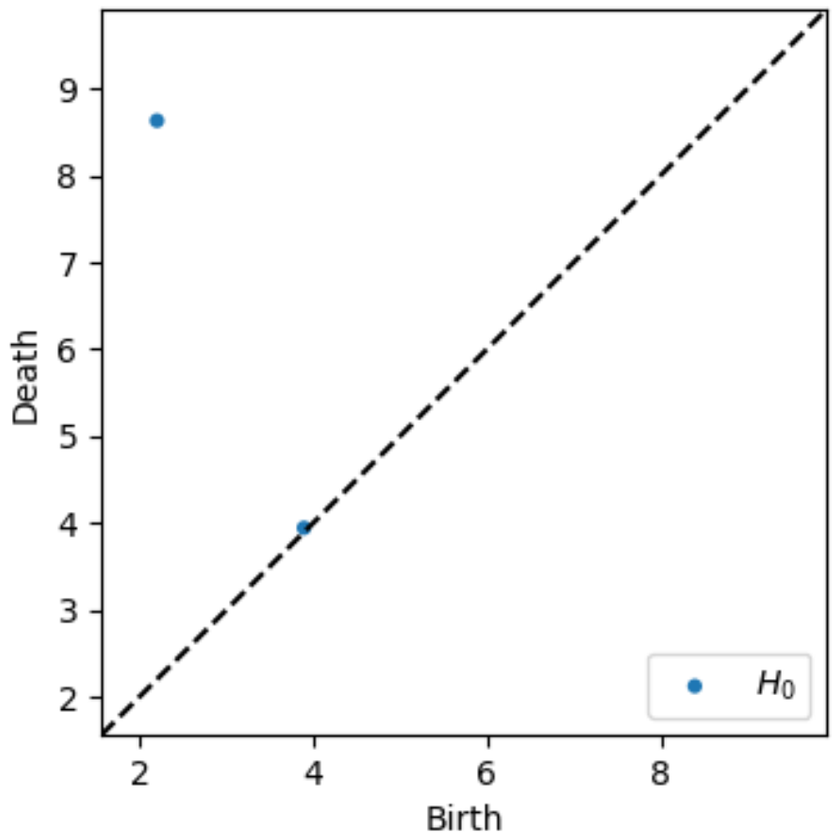


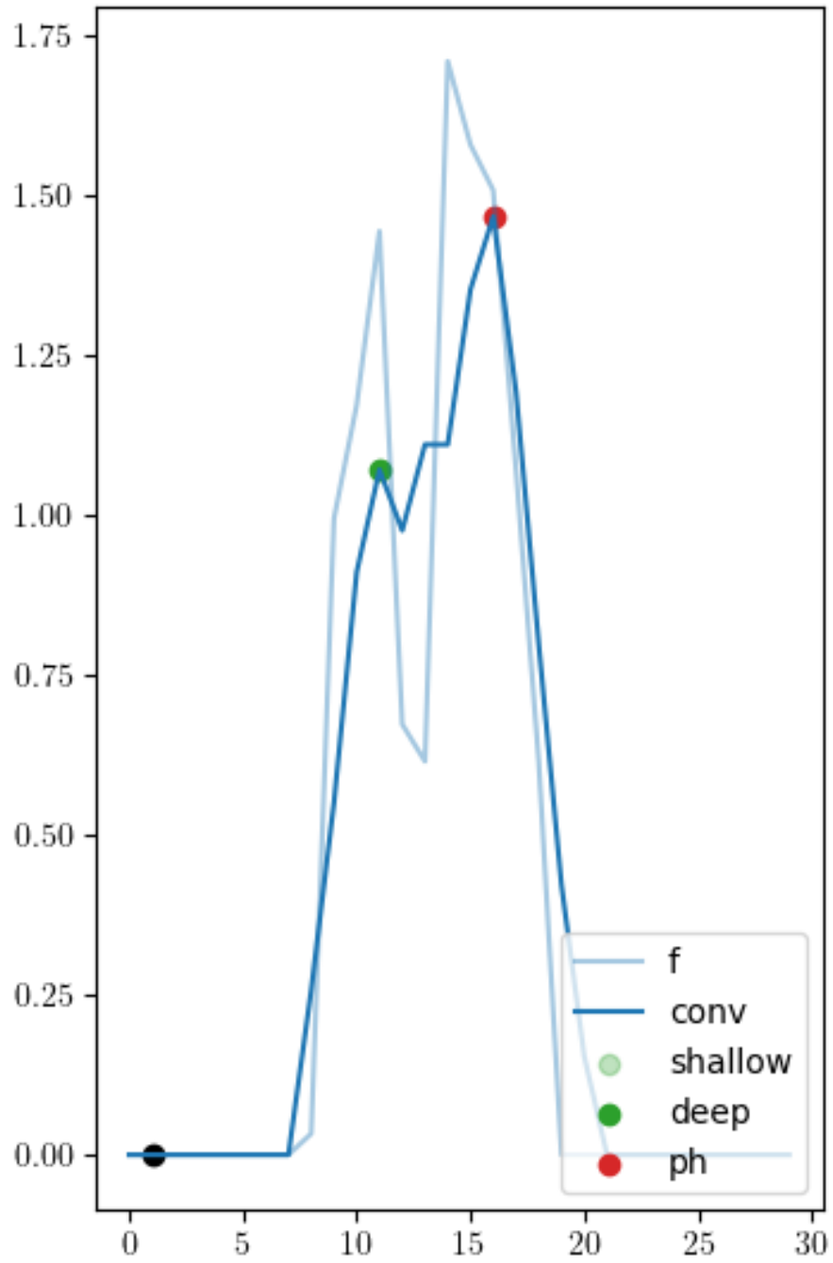




Results







We are 89.9% successful at detecting intersection points in comparison to the minimal surface algorithm (with 180% false positives).

References

- Sulkowka et al., Everything AlphaFold tells us about protein knots, J. Mol. Bio. 436:19, 2024.

- J. I. Sulkowska et al., Theta-curves in proteins, Protein Sci. 33:9, sep. 2024.

- G., An invariant for colored bonded knots, Stud. apply. math, 2021.

- O. Vandans, K. Yang, Z. Wu, L. Dai, L, Identifying knot types of polymer conformations by machine learning. Phys. Rev. E, 2020.

- Sergei Gukov et al, Mach. Learn.: Sci. Technol. 2, 2022.

- F. Bruno da Silva, B. Gabrovšek, M. Korpacz, K. Luczkiewicz, S. Niewieczerzal, M. Sikora, and Joanna I. Sulkowska. Macromolecules 57 (9), 2024.