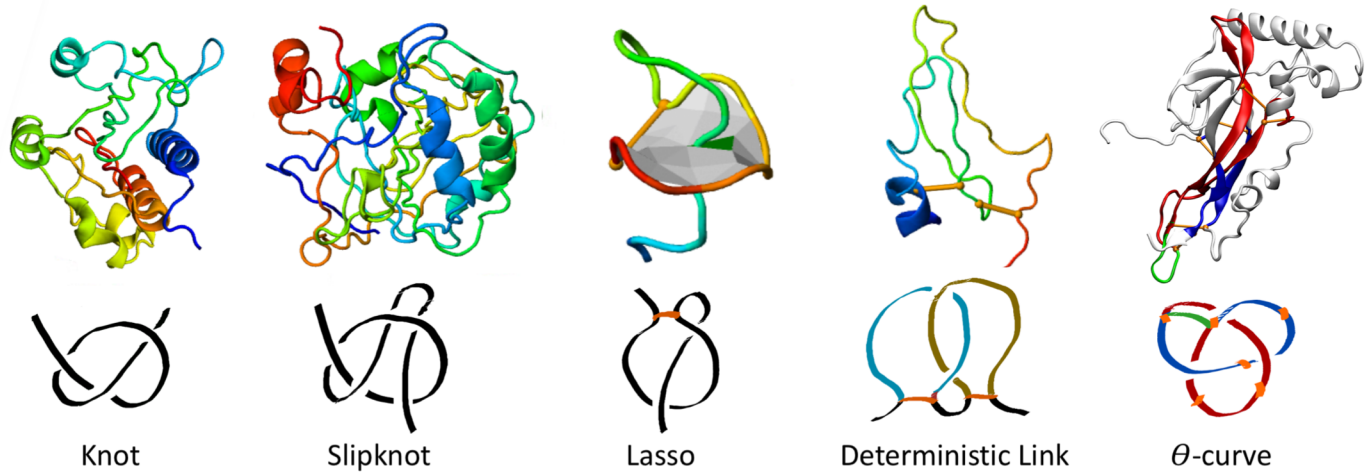
Using LSTM-type models to classify knots/entanglement in polymers and proteins

Joanna Sulkowska

Centre of New Technologies, University of Warsaw

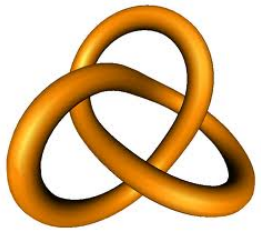


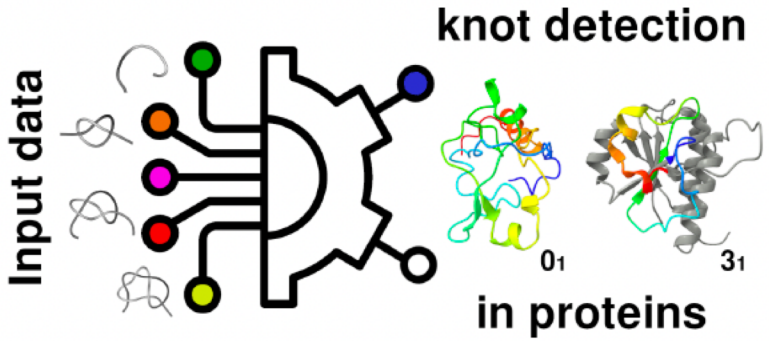
What is machine learning?

"Machine learning is concerned with computer programs that automatically improve their performance through experience"

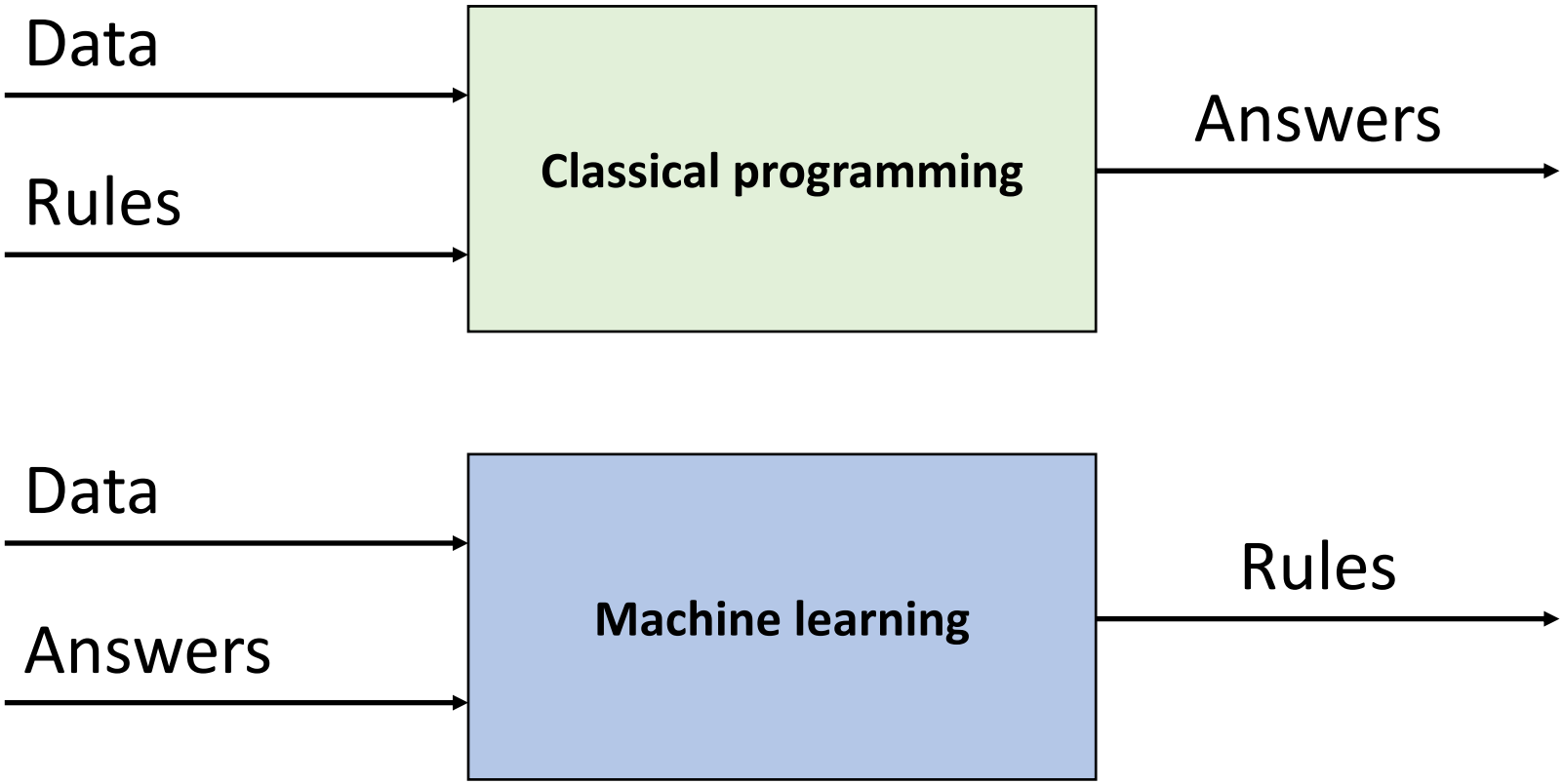
Herbert Alexander Simon

Learning is when **improving** with **experience** at some **task**

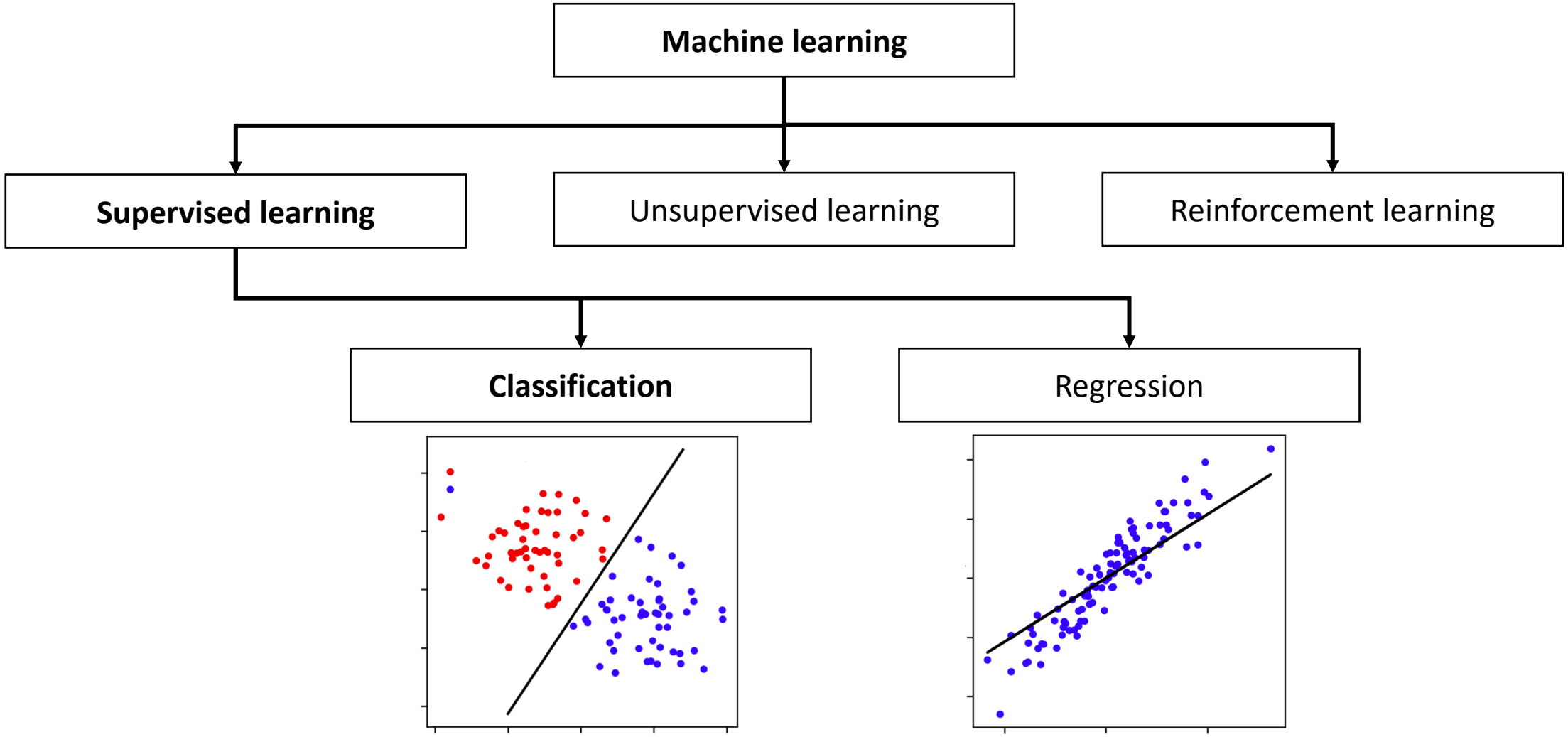




What is machine learning?



Machine learning types



Important elements of supervised ML classification

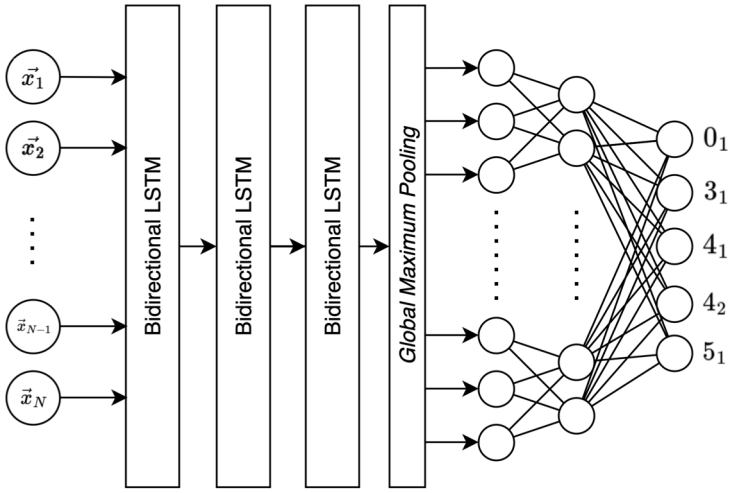
Data (features)

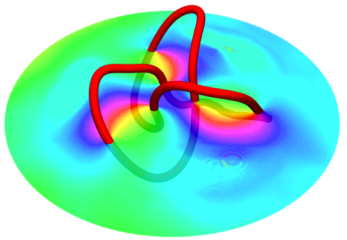
	x	y	z
0	18.48	6.75	-18.54
1	16.56	5.63	-15.42
2	12.97	6.65	-14.46
3	11.06	6.62	-11.18
4	8.20	4.21	-12.00
...	...	...	...

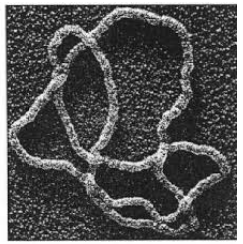
Labels (target)

	id	label
0	AF-A0A1B4BU79-F1	0_1
1	AF-A0A0M3YTH5-F1	3_1
2	AF-P10289-F1	0_1
3	AF-A0A1S6R5S1-F1	0_1
4	AF-A0A0U1I0E7-F1	3_1
...	...	...

Model (classifier)



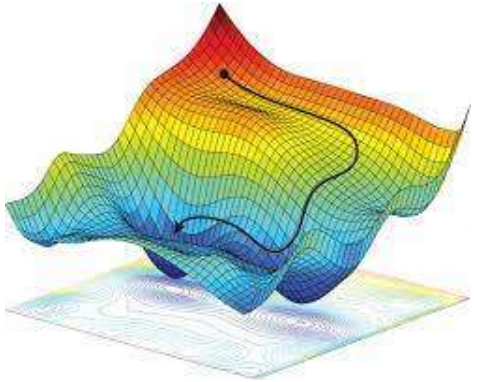




Loss function

$$L_{CE} = -\sum_{i=1}^{N} t_i \log(p_i)$$

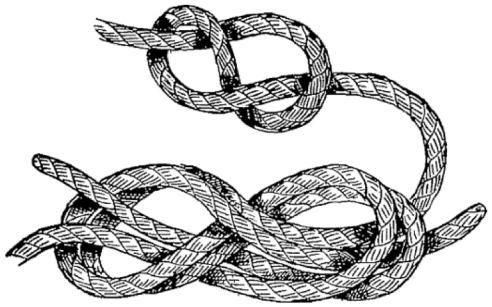
Optimizer



Source: https://krishnaik.in/2022/03/28/understanding-all-optimizers-in-deep-learning/

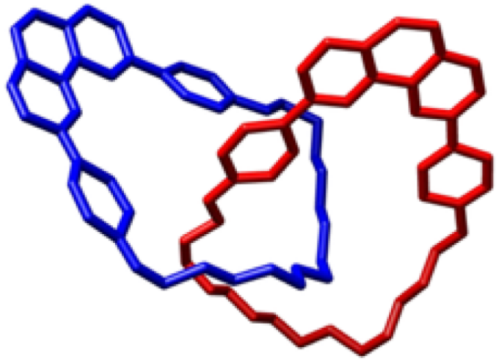
Non-trivial topology

INDUSTRY

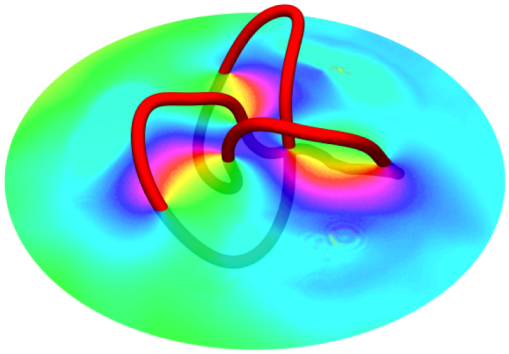


NATURE

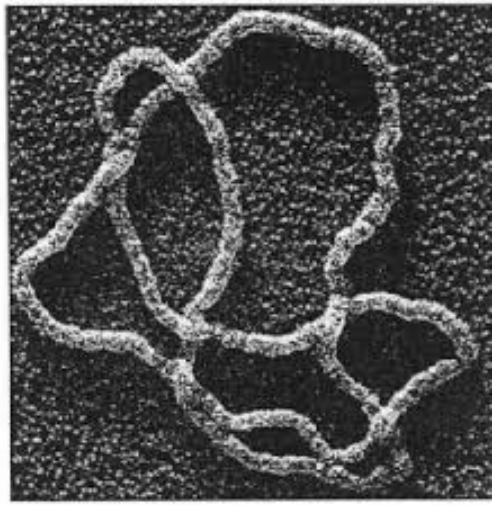
ART



CHEMISTRY



PHYSICS

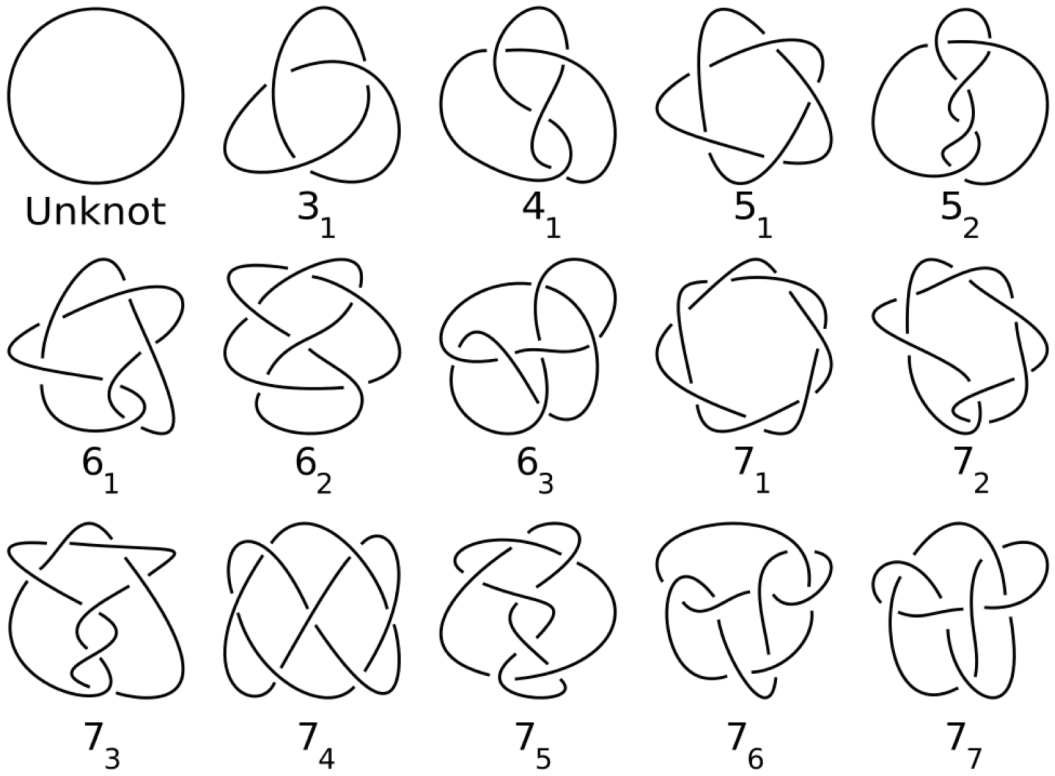


BIOLOGY

Knots

A **knot** is an **embedding** of the **circle** into **three-dimensional Euclidean space (R3)**

The **Jones**, **Alexander**, and **HOMFLYPT polynomials** are commonly used **invariants** used for **knot classification** [1]



Source*: https://en.wikipedia.org/wiki/Knot_(mathematics)*

[1] Eric J Rawdon, Kenneth C Millett, Joanna I. Sulkowska and Andrzej Stasiak. Knot localization in proteins. Biochemical Society Transactions, 41:538–41, 2013 Apr 2013. ISSN 1470-8752. doi: 10.1042/BST20120329

Are there knots in proteins?
Marc L. Mansfield (1994) *Nature Structure Biology*

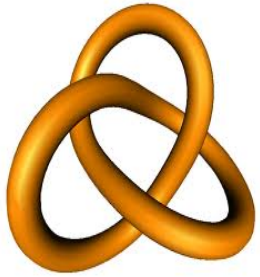
"Sir - Most biochemists would probably agree that proteins in the native state are not knotted. The protein folding mechanism is not perceived as including repetitive, snake-like motions of the chain along its own contour."

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

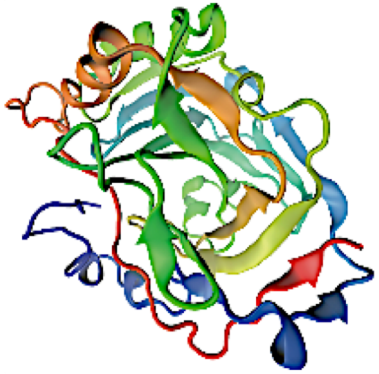
"The absence of knots in proteins would indicate that protein dynamics is non ergodic (all conformation are not accessible)."

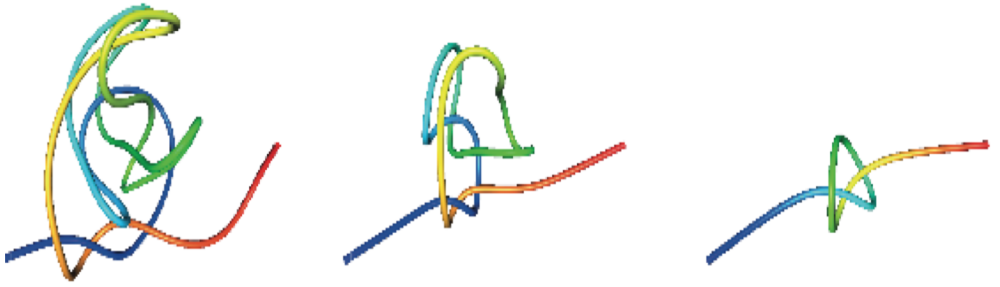
"The most reasonable interpretation of these results is that the protein folding mechanism only explores unknotted conformations"

RCSB PDB PROTEIN DATA BANK



KNOT – proteins



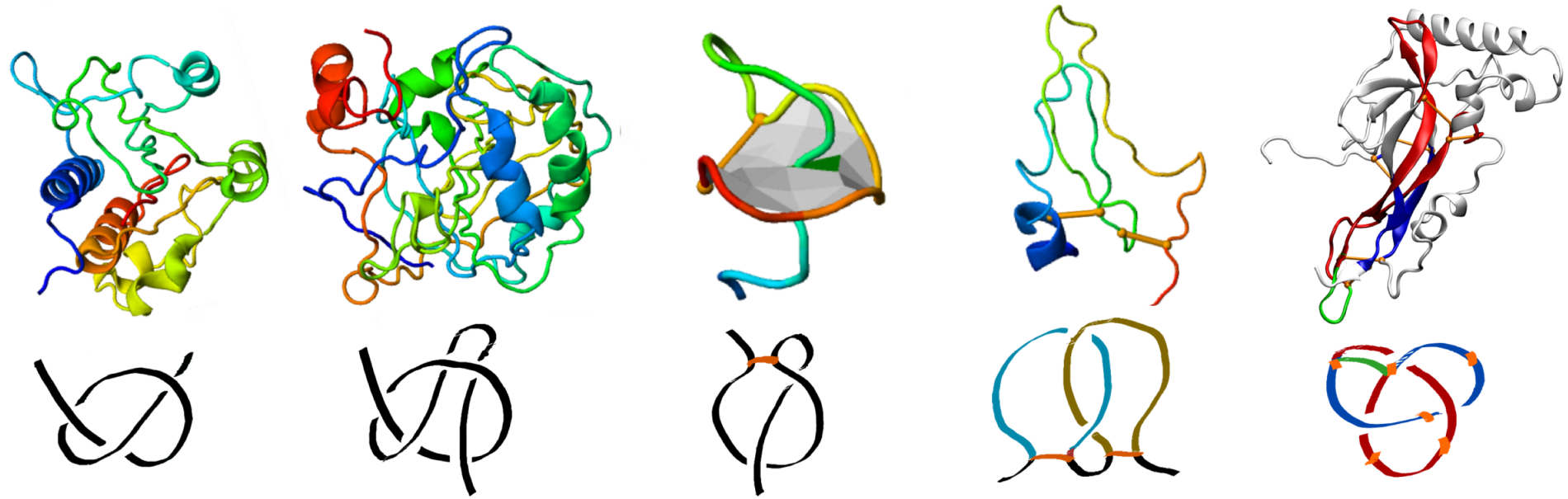


Probabilistic knot

Knot theorists would say NO –
since without cutting we can transform it to the straight line.

However, sailors and climbers can recognize a lot of “linear” knots.

AT LEAST 8% OF KNOWN PROTEINS ARE ENTANGLED



Knot

Sułkowska, et al. PNAS 2012

Slipknot

Lasso

Niemyska, et al. SR 2016

Deterministic Link

Dabrowski-Tumanski & Sułkowska PNAS 2017

θ-curve

Dabrowski-Tumanski et al.,

RCSB PDB PROTEIN DATA BANK

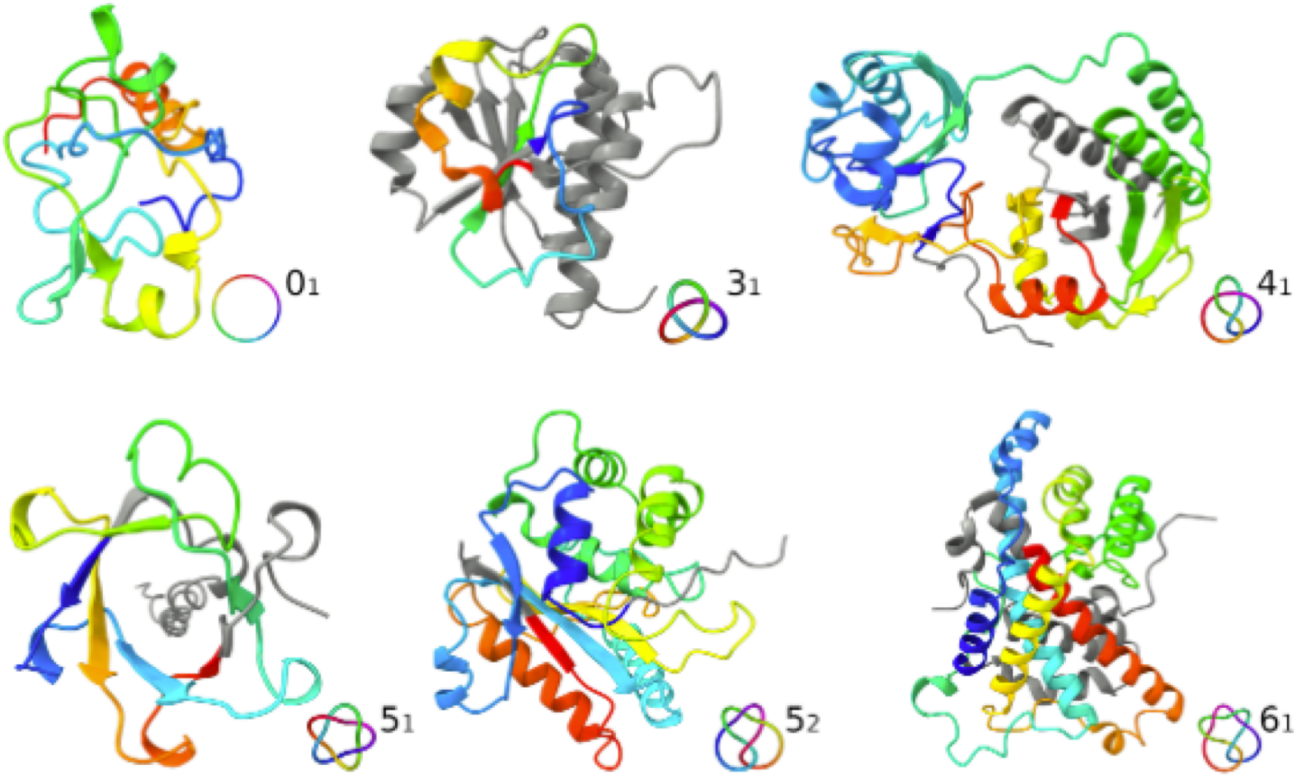
200 K 3D structures

What is a biological role of entanglement?
How knotted proteins fold?
How did knotted structures appear in evolution? How many knotted proteins are actually there?
What types of knots exist in proteins?

Knots in proteins Big data approach to knotted proteins!

Evidence suggests that **knots** serve essential **biological functions**, contributing to the **protein's structural stability**, **enzymatic activity** and **molecular interactions** [2, 3, 4, 5]

Complex protein structure makes **traditional methods** of distinguishing knots **ineffective**



[2] Tobias C. Sayre, Toni M. Lee, Neil P. King, and Todd O. Yeates. Protein stabilization in a highly knotted protein polymer. Protein Engineering, Design and Selection, 24(8):627–630, 06 2011

[3] Szymon Niewieczerzał and Joanna I. Sulkowska. Supercoiling in a protein increases its stability. Phys. Rev. Lett.,123:138102, Sep 2019

[4] Joanna Sulkowska, Piotr Sulkowski, Piotr Szymczak, and Marek Cieplak. Stabilizing effect of knots on proteins. Proceedings of the National Academy of Sciences of the United States of America, 105:19714–9, 01 2009

[5] Thomas Christian, Reiko Sakaguchi, Agata Perlinska, Georges Lahoud, Takuhiro Ito, Erika Taylor, Shigeyuki Yokoyama, Joanna Sulkowska, and Ya-Ming Hou. Methyl transfer by substrate signaling from a knotted protein fold. Nature Structural Molecular Biology, 23, 08 2016

Big data approach to knotted proteins!

- Experimental structures
- X-ray, NMR, Cryo-EM
- Not always possible to determine structure

RCSB PDB PROTEIN DATA BANK

200 K 3D structures

1 %

Knotted proteins

- Predictions
- Source: UniProt
- Well-annotated API

237M sequences

AlphaFold Protein Structure Database

214 M

- Predictions
- Source: MGnify
- Sparse annotations, non-reliable API

772M sequences

ESM Metagenomic Atlas

600 M

ML application !!!

How many knotted proteins exist?
What are the possible topologies?

Big data approach to knotted proteins!

- Experimental structures
- X-ray, NMR, Cryo-EM
- Not always possible to determine structure

RCSB PDB PROTEIN DATA BANK

⬇

200 K 3D structures

⬇

1 %

Knotted proteins

- Predictions
- Source: UniProt
- Well-annotated API

237M sequences

⬇

AlphaFold Protein Structure Database

⬇

214 M

0.4 %

Knotted proteins

- Predictions
- Source: MGnify
- Sparse annotations, non-reliable API

772M sequences

⬇

ESM Metagenomic Atlas

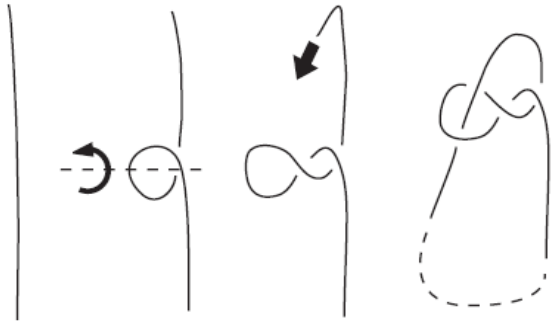
⬇

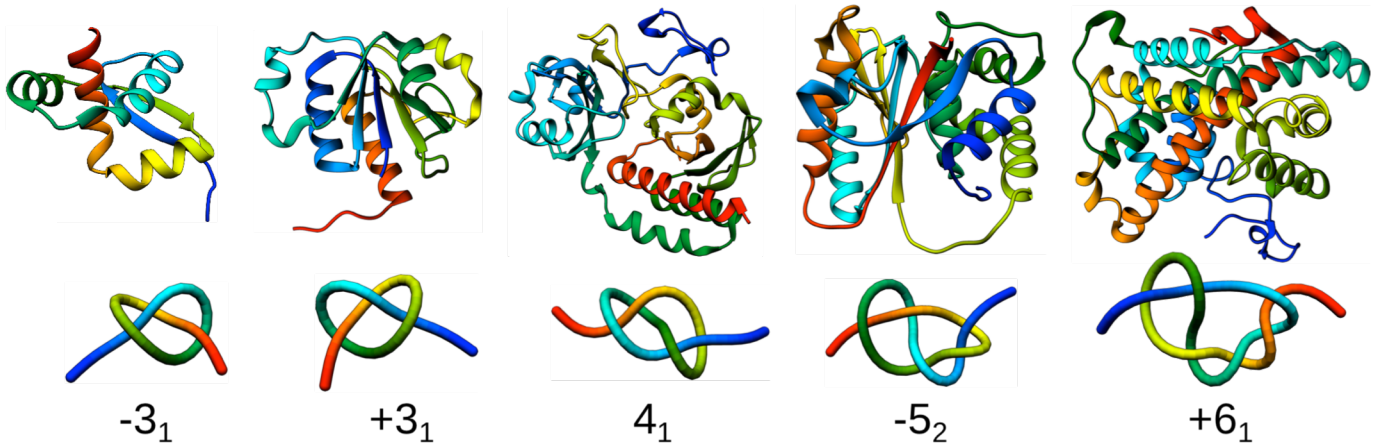
600 M

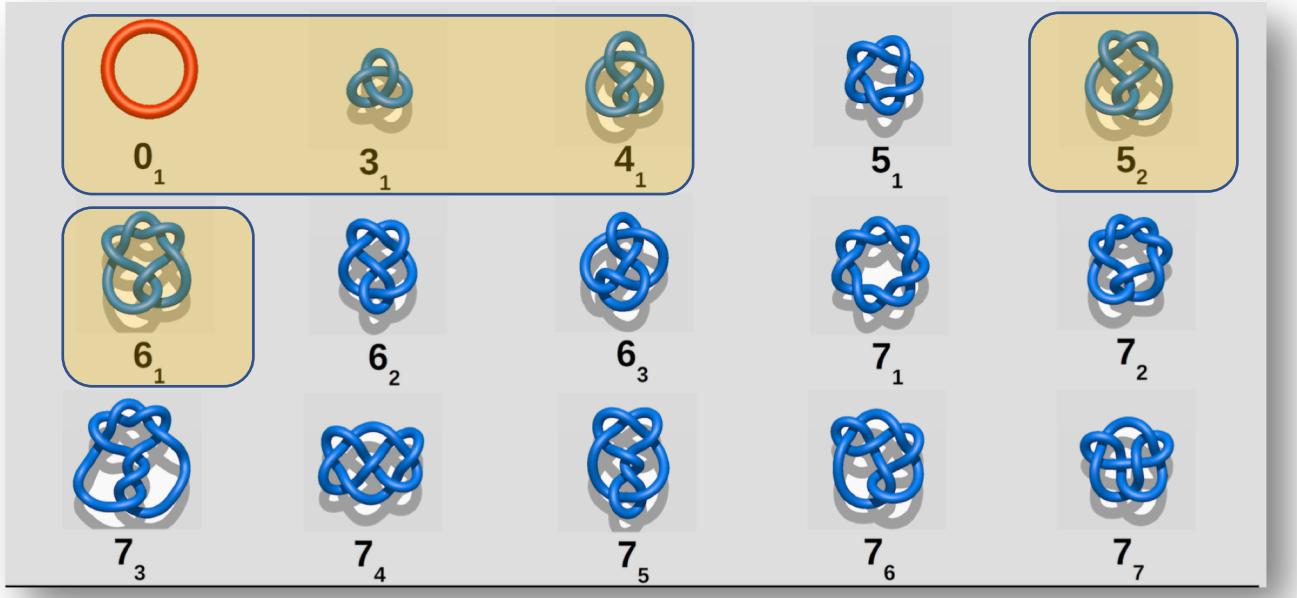
0.5 % (?)

Knotted proteins

Are there more complex non-trivial topologies?







Dodging the crisis of folding proteins with knots
JI Sulkowska, et. al.., Proc. Natl. Acad. Sci. USA (2009)

Are there more complex non-trivial topologies?

- **NEW KNOTTED FAMILYS**
- **POTENTIALLY NEW TYPE OF KNOTS**

DeepMind

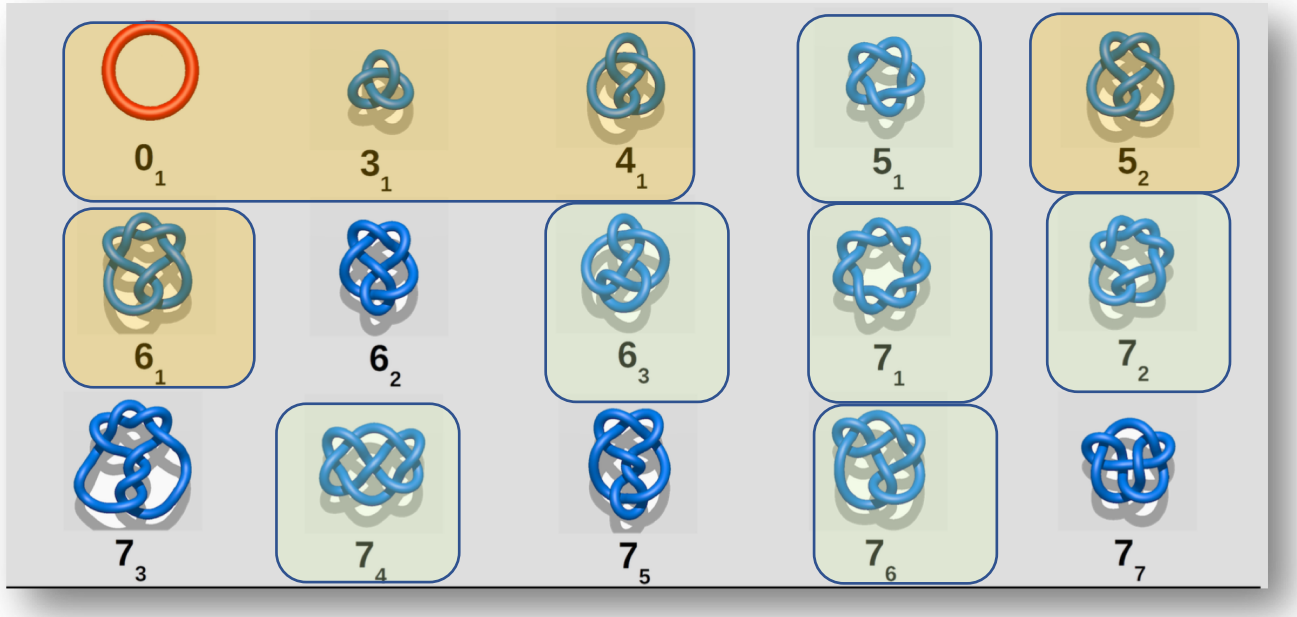
CONSERVATION OF KNOT TYPE in family

RoseTTaFold

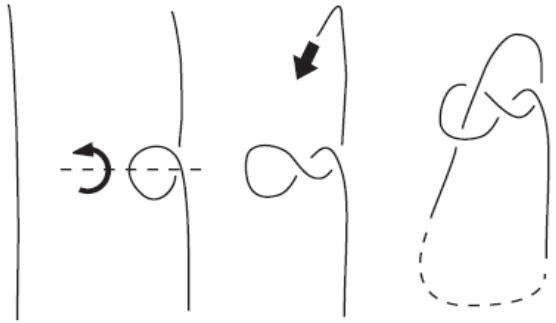
(cross-validation)

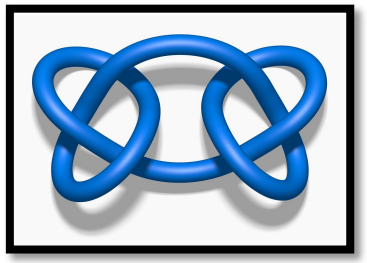
EMSFold

(cross-validation)



NO FOLDING THEORY !

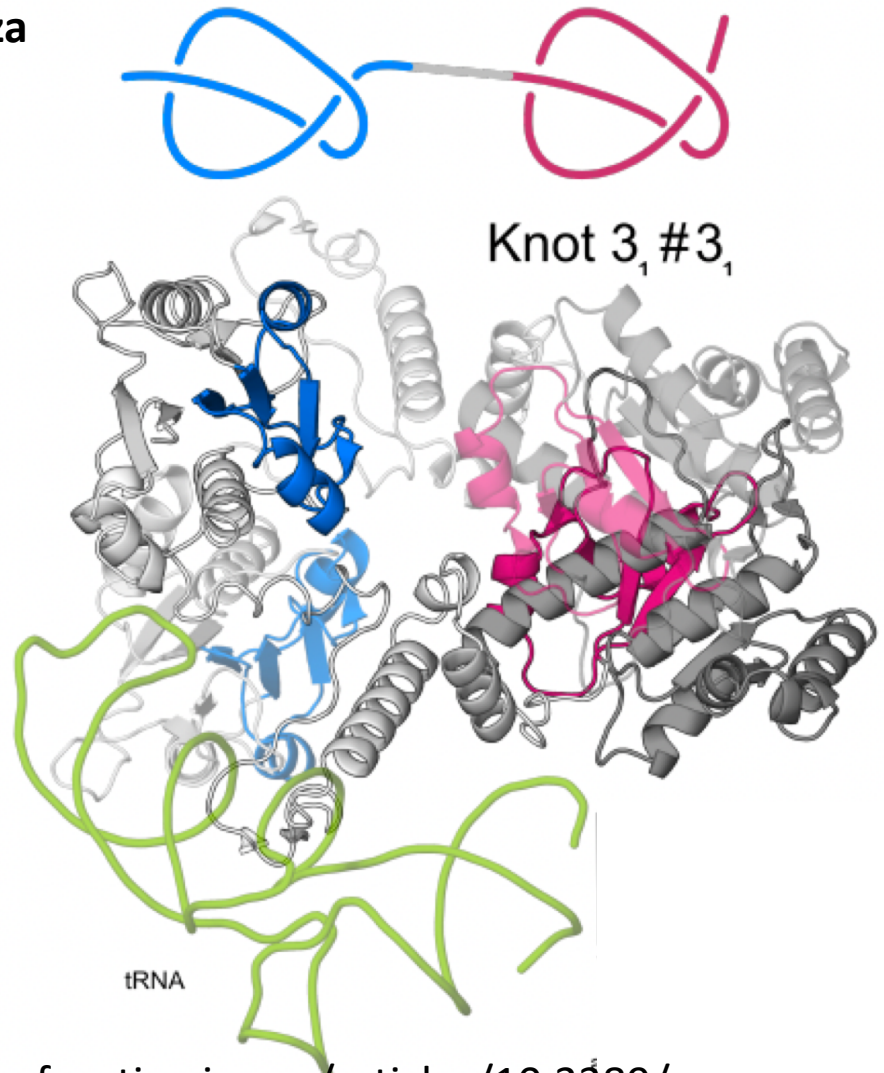


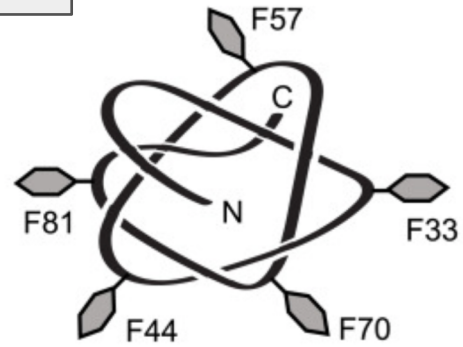


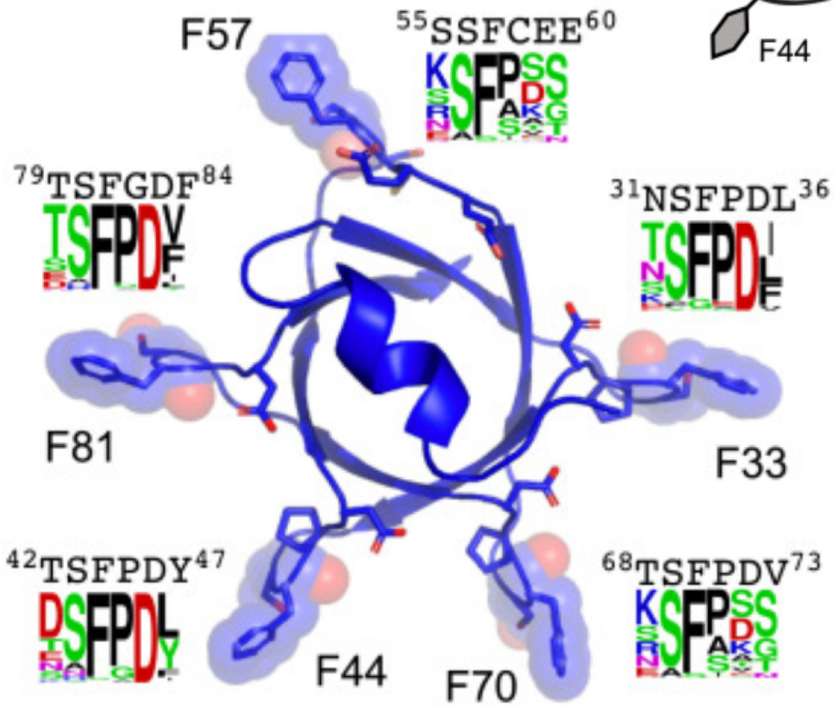
$3_1 \# 3_1$

New types of topology predicted by AF

X-ray 2.1 A
Anna Kluza







https://www.frontiersin.org/articles/10.3389/fmolb.2023.1223830/abstract

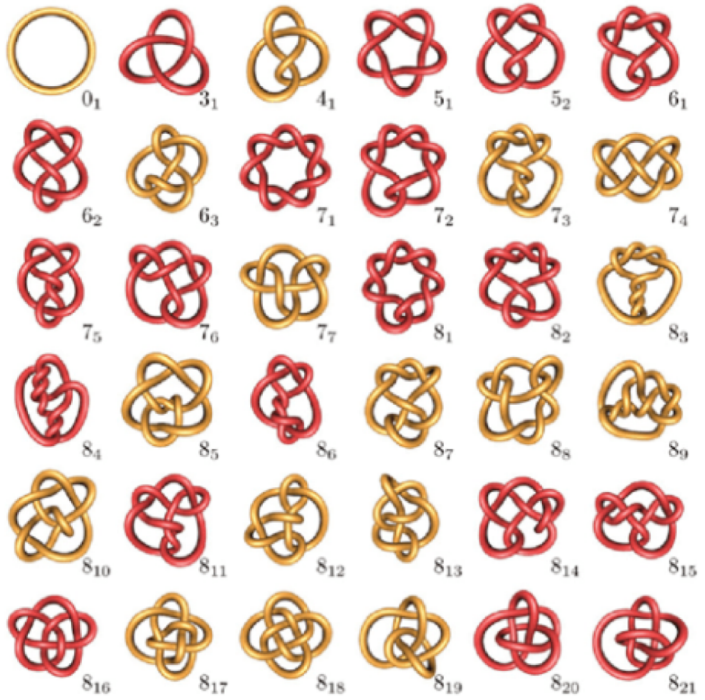
https://www.sciencedirect.com/science/article/pii/S0021925823025814

Machine Learning Understands Knotted Polymers
A. Braghetto et al. *Macromolecules* 2023, 56, 7, 2899–2909

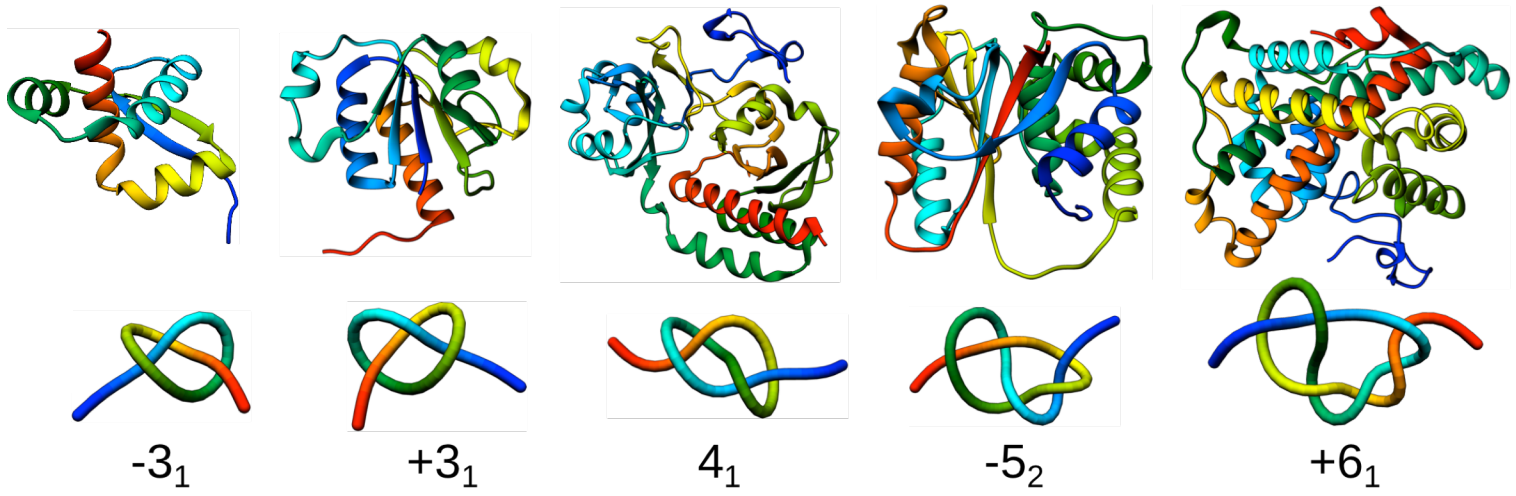
Non-trivial topology

A main issue to face in studying knotting and knot complexity in polymeric systems is the identification and classification of the knotted states in a very large set of ring/open knot configurations (embeddings).

This has been mostly achieved by first projecting a given configuration on a plane, then identifying and coding the resulting self-crossings and finally feeding this information to build a topological invariant such as the Alexander polynomial, if the invariant differs for two different configurations, then they have different knotted states (or do belong to two different knot types).



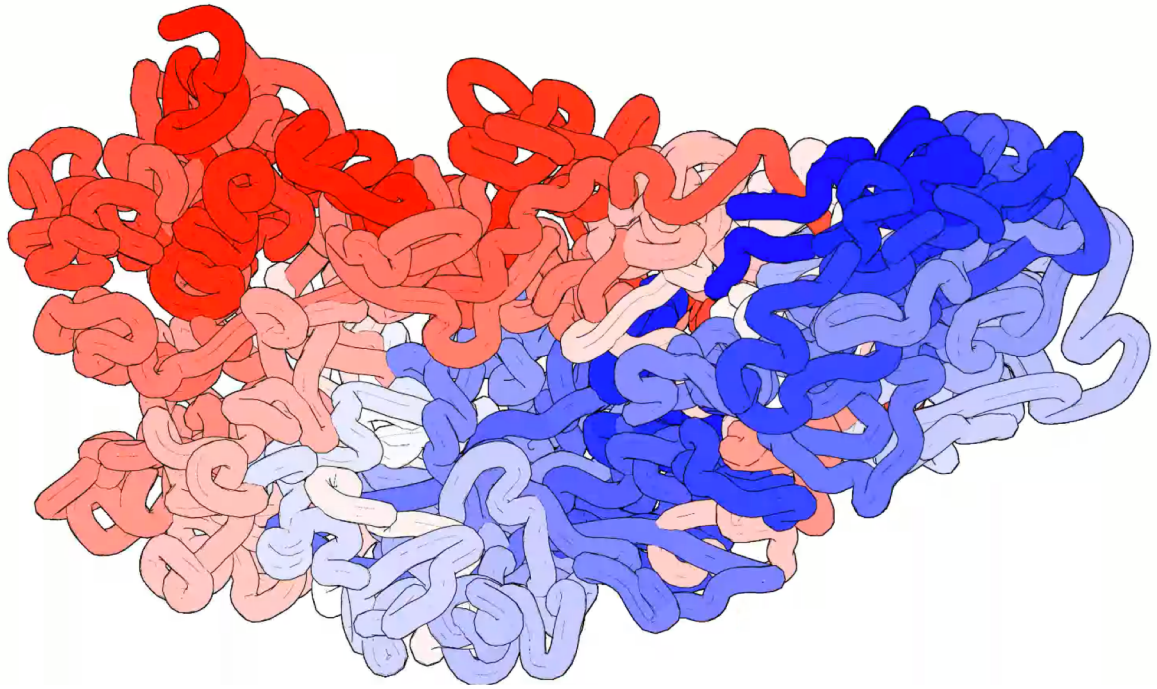
- With this method, one can distinguish prime knots up to the minimal crossing number 7

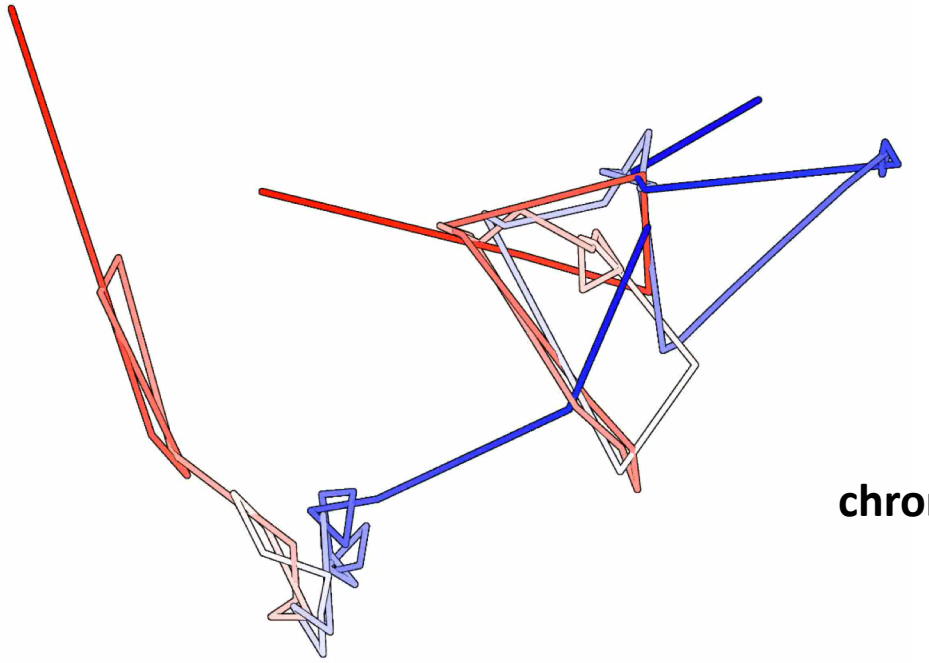


Non-trivial topology – problems

The computation time of the whole procedure is known to increase with the number of crossings n_c formed after the projection and, if more sophisticated topological invariants as the Jones and the HOMFLY polynomials are considered, the growth of time with n_c is dramatic.

This problem is particularly important when either the length of the polymer N is very large or when the 3D configurations are geometrically badly organized in space (compact or isotropically confined rings, proteins, DNA, RNA).





Knotted
chromosome

KnotGenom: server to detect entanglement in and between chromosomes,
Sulkowska et al. Nucleic Acids Res. (2018)

DATABASE AND SERVER – AlphaKnot 2.0

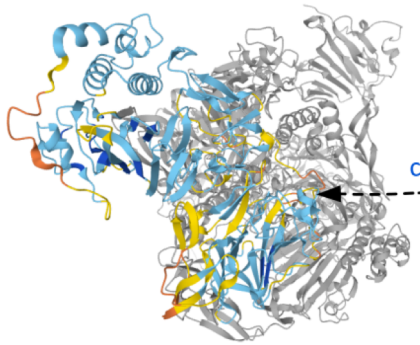
A knot or not a knot?

AlphaFold proteome predictions

Knot types	Category	Uniprot ID	Organism	Knot pLDDT
6_3	Knot	O00534-F1	SOYBN	89
6_3	Unsure	Q75WE7-F1	RAT	88
5_1	Artifact	Q4D805-F1	TRYCC	75

OR

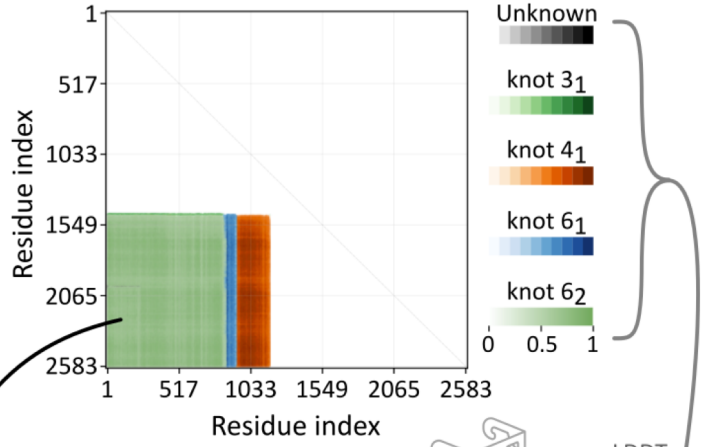
Your own protein or prediction



colored knotted region (6_2)

AlphaKnot

Knot map



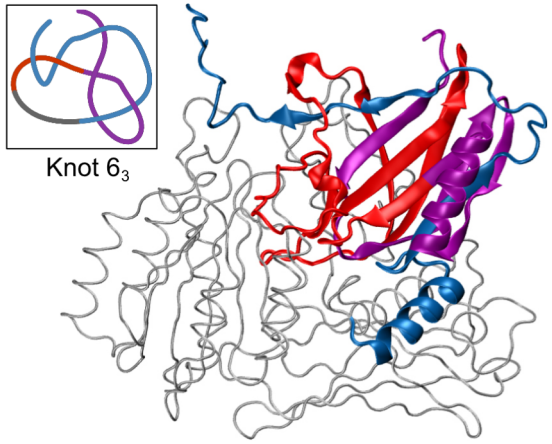
pLDDT
C-alpha clashes
Topological complexity

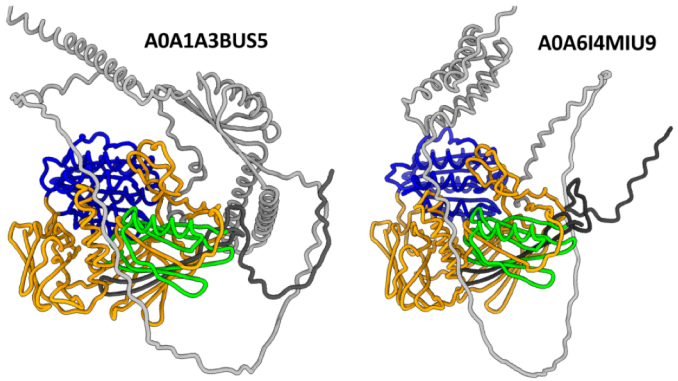
Knot types & Model sequence

	Knot pLDDT	Knot core range
6_2	75	839-1477
3_1	75	831-1471
Unknown	76	242-1989
6_1	75	927-1477
4_1	72	1169-1479

Category

Artifact
OR
Unsure
OR
Knot

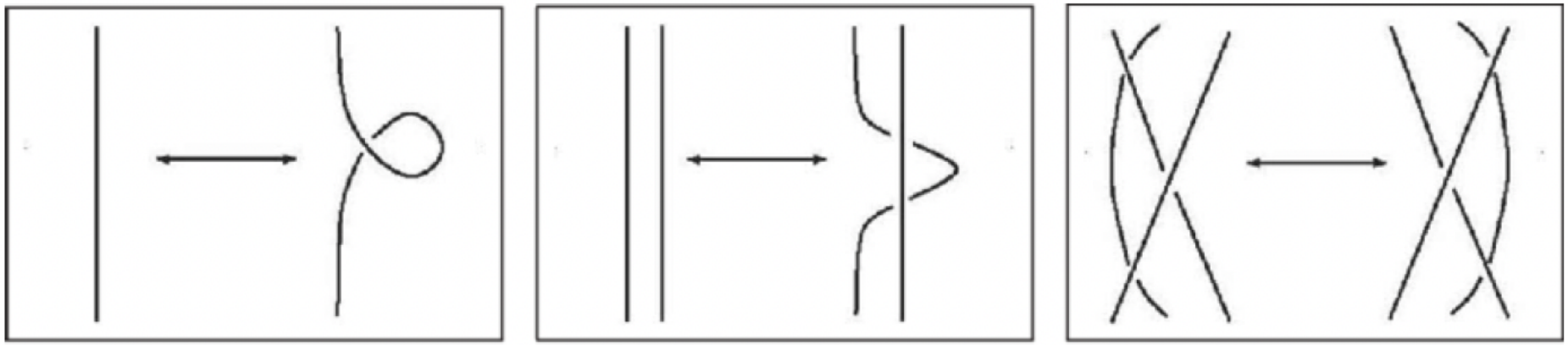




AlphaKnot: server to analyze entanglement in structures predicted by AlphaFold methods
W Niemyska et. al., NAR 2022

Non-trivial topology – problems

To solved the problem, we can conduct local deformations that reduce as much as possible the geometrical entanglement of the curve in 3D space (and thus the number of crossings after projection) without changing its topological state.

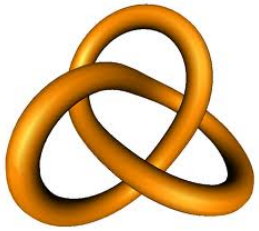


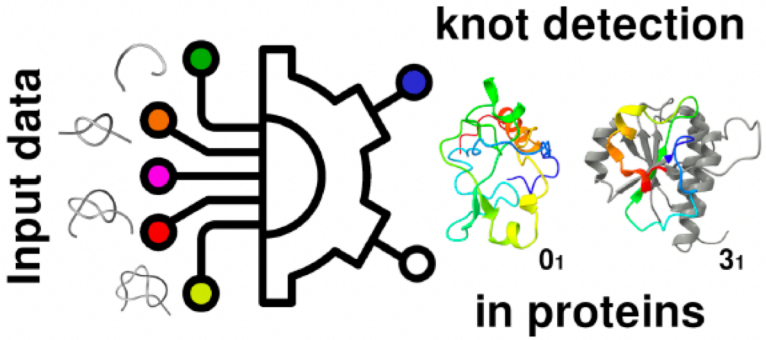
- However, these procedures are time consuming and there is no guarantee that they can always decrease the number of crossings to the point at which the knot identification based on an algebraic invariant is actually doable.

- but this can be done with ML approach (discover hidden patterns within a large set of complex data).

Machine learning understands knotted polymers
A. Braghetto et. al., *Macromolecules* 2023
Knots and θ-Curves Identification in Polymeric Chains and
Native Proteins Using Neural Networks, F. Bruno da Silva et. al., *Macromolecules* 2024

ML supervised approaches based on neural networks (NNs) a tool in providing new insights into the mathematical problem of knot identification





Questions:

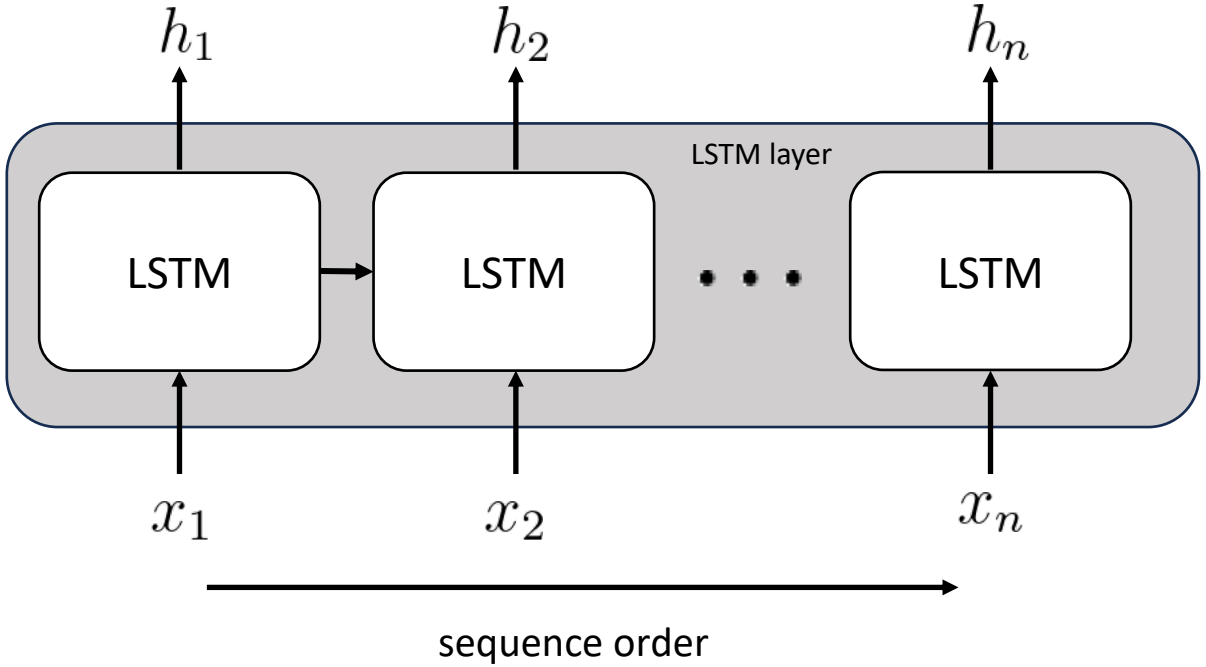
- Which configurational features or NN architectures are more efective in learning global properties that best identify knotted states in polymer rings under diferent physical conditions?

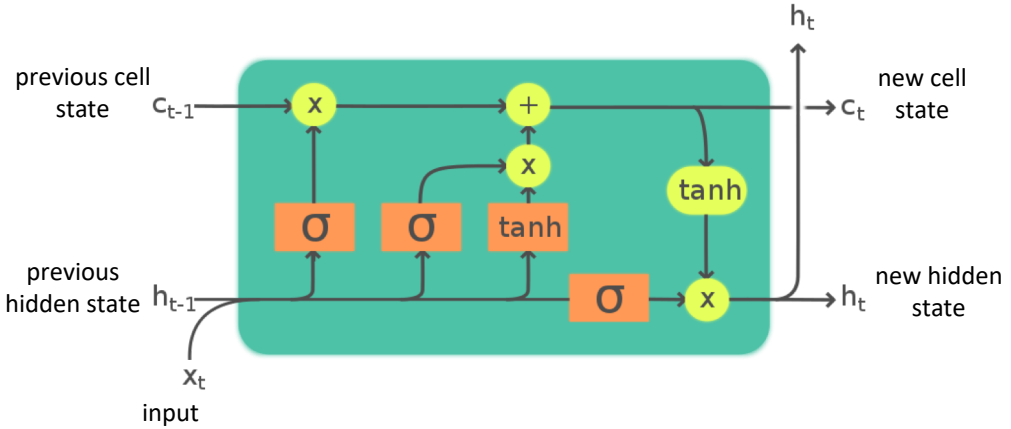
- Unlike algebraic methods, the ML approach cannot identify with certainty a knotted state but it can furnish a probability that a configuration has a given knot type.

A. Braghetto et. al. *Macromolecules* 2023
F. Bruno da Silva et. al.., *Macromolecules* 2024

LSTM

Long Short-Term Memory (LSTM) is an advanced type of **recurrent neural network (RNN)** designed for handling **sequential data** by capturing, retaining or removing **long-term dependencies**, thanks to its **memory cells** and **gating mechanisms.**



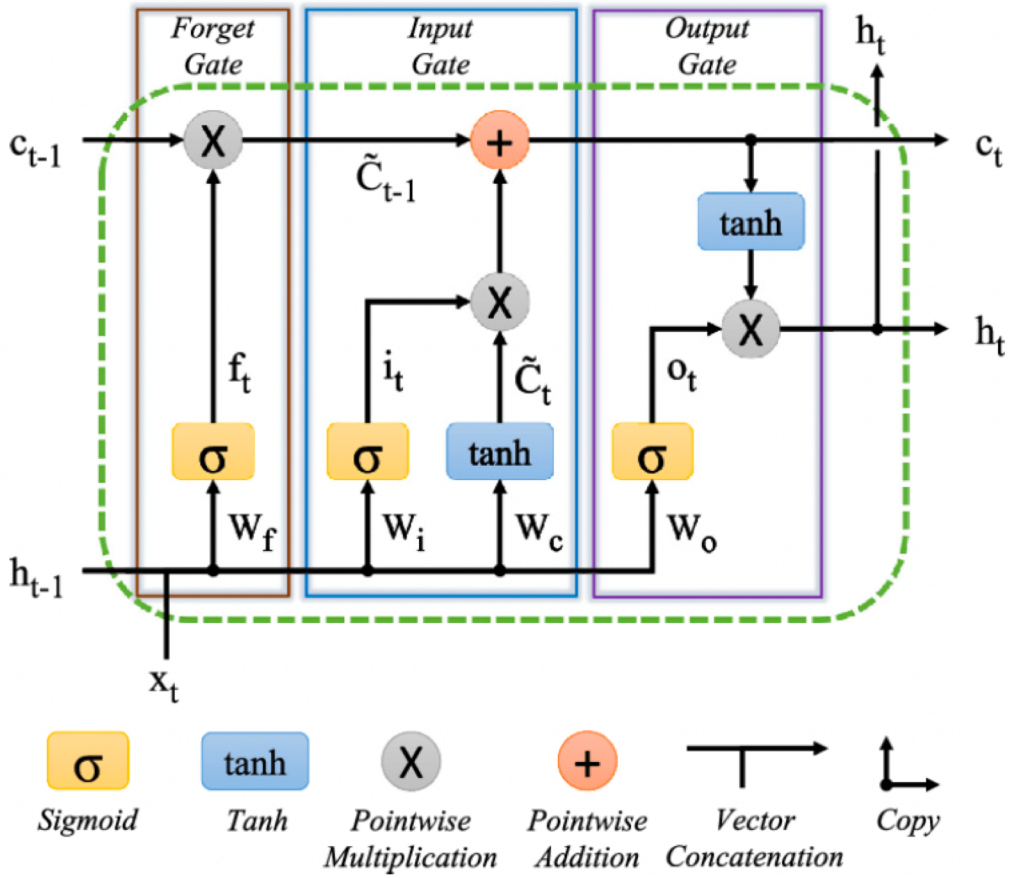


RNN differs from convolutional neural network (CNN), which is the second main type of NN, in the way it processes information: in CNN, inputs and outputs are processed independently, in the case of RNN, the output of the previous cell is the input to the current one.

LSTM

Long Short-Term Memory (LSTM) is an advanced type of **recurrent neural network (RNN)** designed for handling **sequential data** by capturing, retaining or removing **long-term dependencies**, thanks to its **memory cells** and **gating mechanisms.**

- In forget gate, an input (x_t) and the hidden state (ht–1) are passed by the sigmoid function. The result takes values from 0 to 1 depending on their importance and is multiplied pointwise with the previous state of the cell (c_{t-1}).
- The input gate processes then x_t and h_{t-1} through a sigmoid function and a tanh function that returns a vector of values between –1 and 1. The results are pointwise multiplicated and then pointwise added to the state of the cell (C_{t-1}).
- The output gate processes the updated state of the cell with the tanh function and the previous hidden state with the sigmoid function. The results obtained are multiplied pointwise and returned as the next hidden state (h_t).

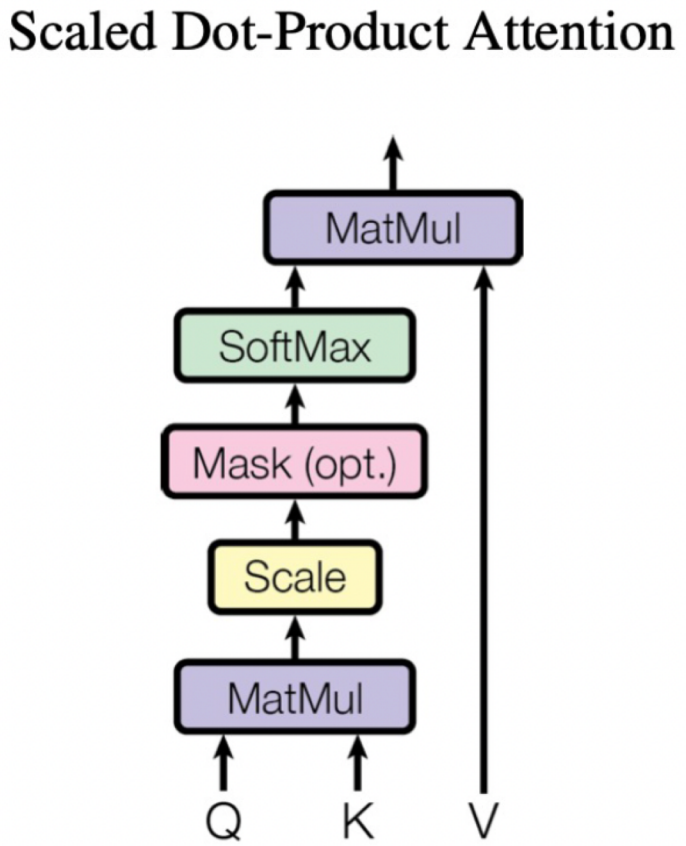


Attention is all you need

The **attention mechanism**, which was first presented in the "*Attention is All You Need*" paper by Vaswani et al. [1], enables models to dynamically determine the relative importance of various words in a sequence, improving the ability to capture long-range dependencies

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

where Q, K, V are the query, key, and value matrices, d_k is the dimension of the keys.



An Encoder-Decoder kind of neural network architecture that allows the model to focus on specific sections of the input while executing a task. It dynamically assigns weights to different elements in the input, indicating their relative importance or relevance.

[1] Vaswani, A., Shazeer, N.,Parmar, N., et. al., *Attention is All you Need*, 2017

Multi-Head Attention

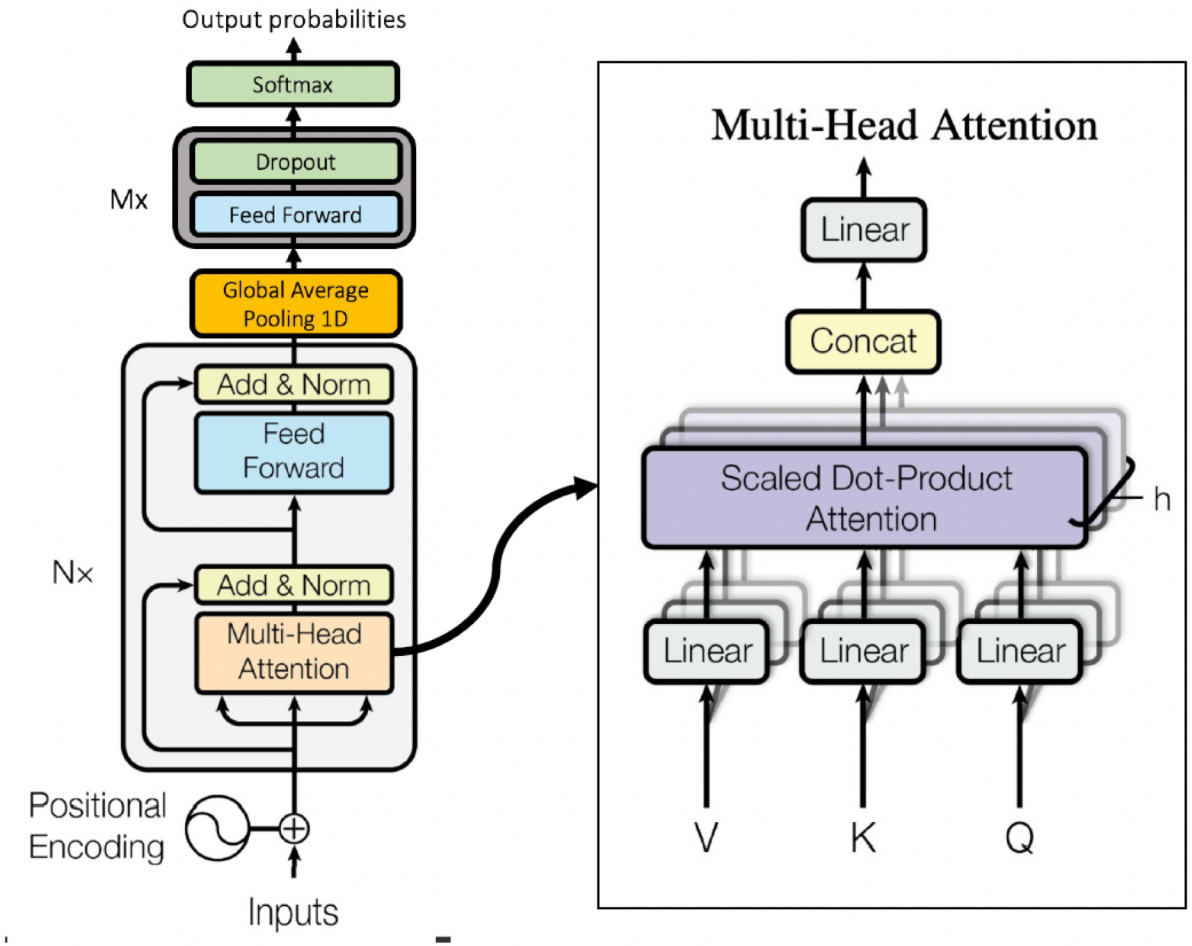
Multi-head Attention is a module for attention mechanisms which runs through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

where each head is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

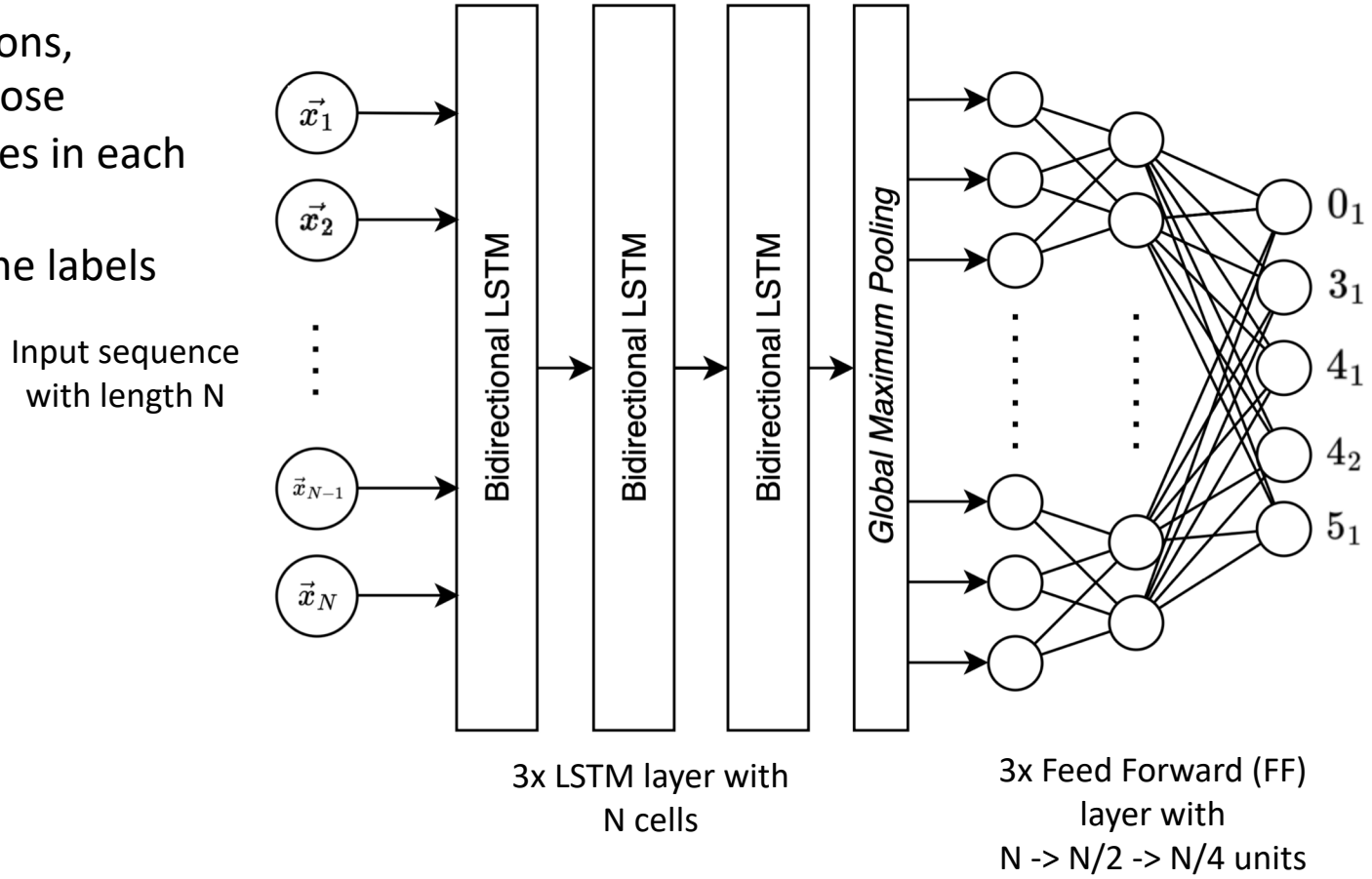
and W_i^Q, W_i^K, W_i^V, and W^O are parameter matrices.



LSTM-based model to classify topologies

Model architecture proposed by Olafs Vandans, et al. [6]

- Time series are replaced by polymer configurations,
- the signal is represented by a set of features whose dimension corresponds to the number of features in each set (i.e., 3)
- the instants of the time series are replaced by the labels of the bead along the rings/protein chain



Model main parameters

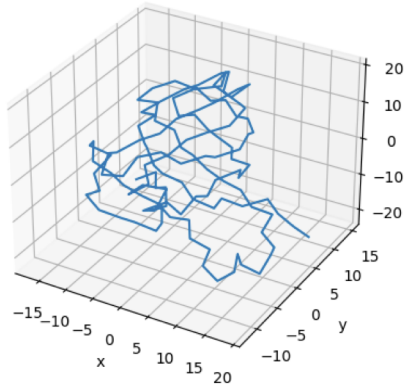
Loss function	Crossentropy
Optimizer	ADAM (RMSprop)
Metric	Accuracy
Batch size	32

[6] Olafs Vandans, Kaiyuan Yang, Zhongtao Wu, and Liang Dai. Identifying knot types of polimer conformations by machine learning. Physical Review E, 101(2), feb 2020

Data, features and preprocessing

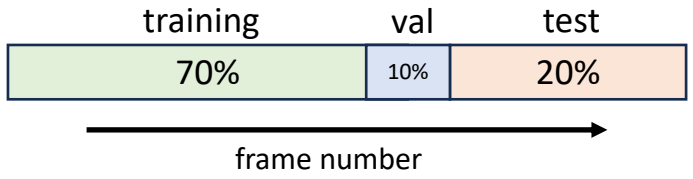
Data

dynamics simulation sequentially divided



	x	y	z
0	18.48	6.75	-18.54
1	16.56	5.63	-15.42
2	12.97	6.65	-14.46
3	11.06	6.62	-11.18
4	8.20	4.21	-12.00
...	...	...	...
123	-6.84	5.37	-19.98
124	-8.93	8.00	-18.20
125	-11.71	9.00	-20.63
126	-14.01	12.02	-20.02
127	-16.92	10.80	-22.22

150,000 frames



Features

1. $[x_i, y_i, z_i]$

2. $[d_i(s), \theta_i(s), \varphi_s(s)] \quad s = 1$

$$d_i(s) = |\mathbf{x}_{i+s} - \mathbf{x}_i| \tag{1.1}$$

is a distance between beads $\mathbf{x}_{i+s}$ and $\mathbf{x}_i$,

$$\theta_i(s) = \arccos[(\mathbf{x}_{i+s} - \mathbf{x}_i) \cdot (\mathbf{x}_{i+2s} - \mathbf{x}_{i+s})] \tag{1.2}$$

we consider as a angle between three monomers and

$$\begin{aligned} \varphi_s(s) &= \arctan\left[\frac{\mathbf{u} \cdot \mathbf{w}}{\mathbf{v} \cdot \mathbf{w}}\right] \\ \mathbf{v} &= \frac{(\mathbf{r}_2 - \mathbf{r}_1) \times (\mathbf{r}_3 - \mathbf{r}_2)}{||(\mathbf{r}_2 - \mathbf{r}_1) \times (\mathbf{r}_3 - \mathbf{r}_2||} \\ \mathbf{w} &= \frac{(\mathbf{r}_3 - \mathbf{r}_2) \times (\mathbf{r}_4 - \mathbf{r}_3)}{||(\mathbf{r}_3 - \mathbf{r}_2) \times (\mathbf{r}_4 - \mathbf{r}_3||} \\ \mathbf{w} &= \frac{(\mathbf{r}_3 - \mathbf{r}_2) \times \mathbf{v}}{||(\mathbf{r}_3 - \mathbf{r}_2) \times \mathbf{v}||} \end{aligned} \tag{1.3}$$

is the angle between the plane with points (r_1, r_2, r_3) and the plane with (r_2, r_3, r_4), where we used notation: $r_1 = x_i$, $r_2 = x_{i+s}$, $r_3 = x_{i+2s}$ and $r_4 = x_{i+3s}$.

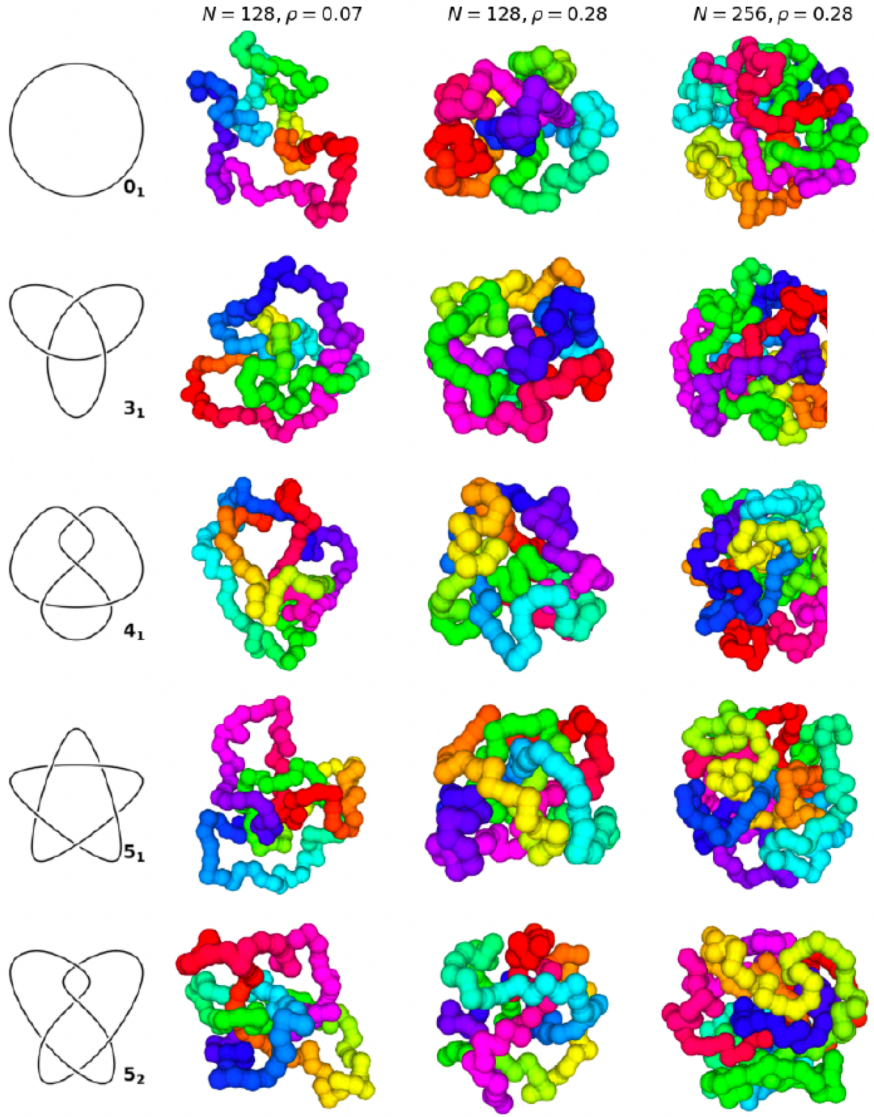
Preprocessing

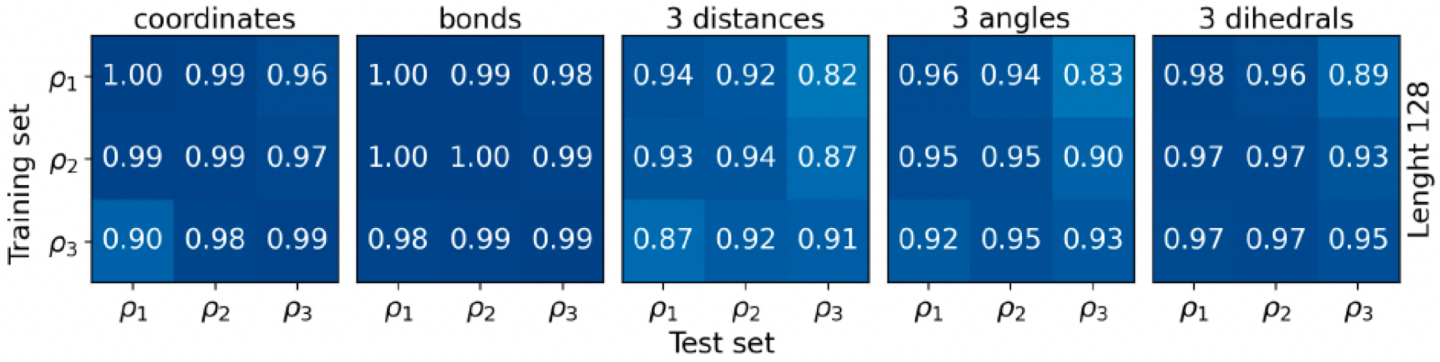
Normalization

StandardScaler()

$$X_i' = \frac{X_i - \mu}{\sigma} = \frac{X_i - X_{\text{mean}}}{X_{std}}$$

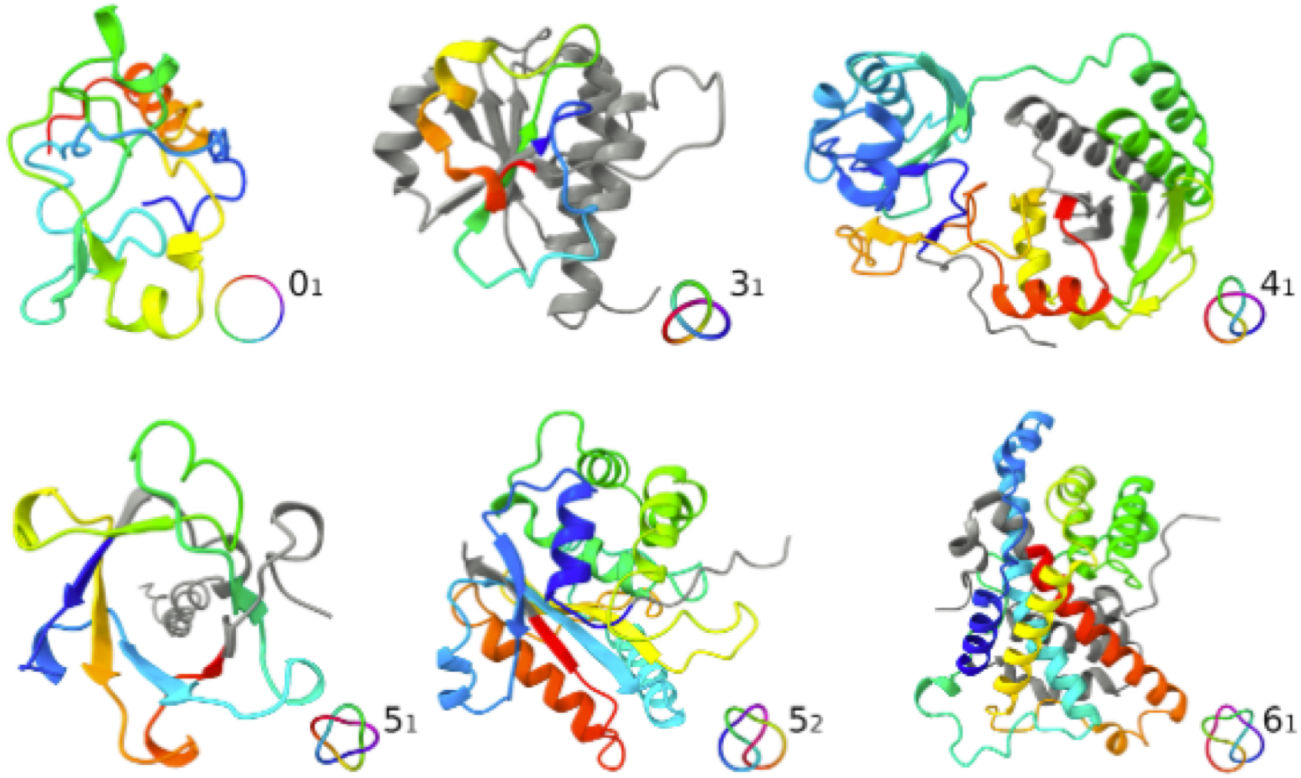
Results for polymers

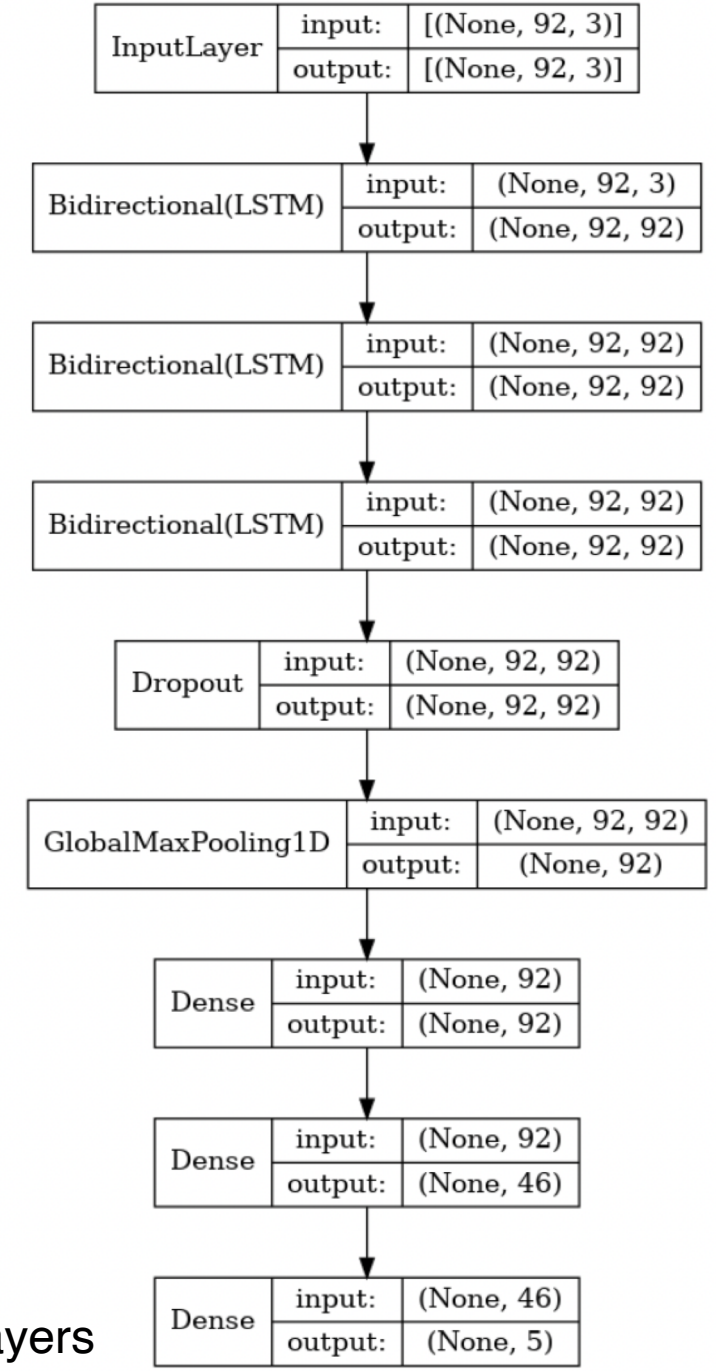




Machine learning understands knotted polymers
A. Braghetto et. al. *Macromolecules* 2023

Knots and θ-Curves Identification in Polymeric Chains and Native Proteins Using Neural Networks





The diagram shows the sequence of data processing layers

Results for polymers and protein-like

Closed knots
(w/o interactions)

Best accuracy: 99%

$[x_i, y_i, z_i]$

Closed knots

Best accuracy: 84%

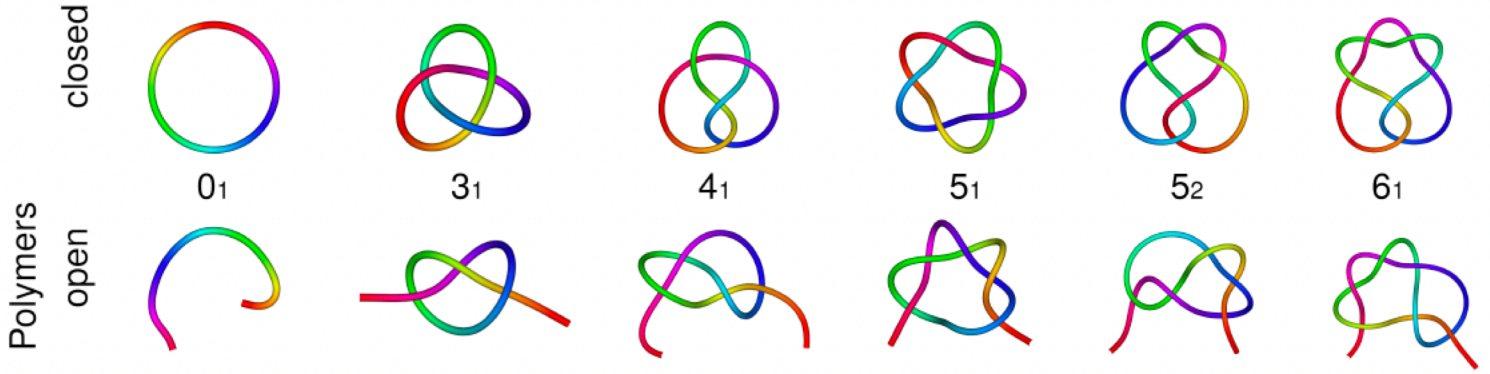
$[d_i(s), \theta_i(s), \varphi_s(s)]$

Open knots

Best accuracy: 99%

$[d_i(s), \theta_i(s), \varphi_s(s)]$

Optimizer: RMSprop



F. Bruno da Silva et. Al.., *Macromolecules* 2024

Results for proteins

Different approach

Simulations generated for selected proteins

Analogous training as before

Testing on a set of 1106 native conformations of other proteins

Only **unknot** (823) and **trefoil** (283)

Large class imbalance in testing set

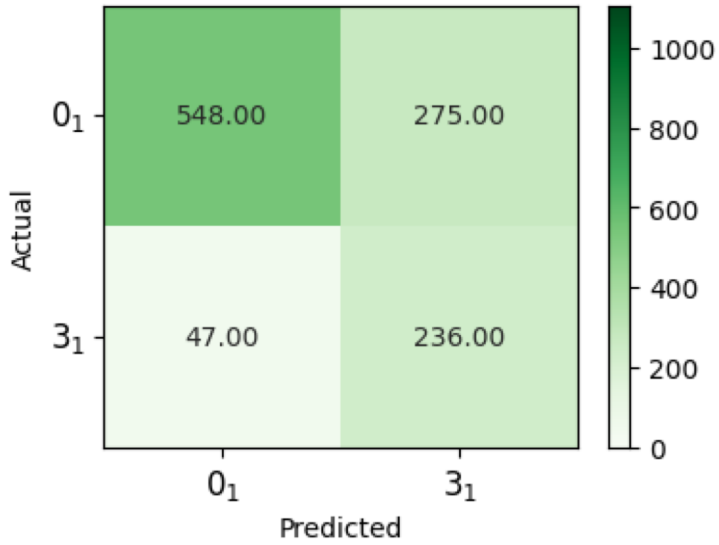
Dynamics for 6 proteins were generated:

0_1: *A0A1H4VHL9* (1), *A0A1M4N8C8* (2), *P85286* (3)
3_1: *A0A2D6CS53* (4), *A0A7L1ILP5* (5), *AF-A0A2E8PTH8* (6)

at 7 different temperatures *T* = 20, 30, 40, 50, 75, 90, 100

Best results for model trained on (2) and (5) for *T*=20

Accuracy: 71%



F. Bruno da Silva et. Al.., *Macromolecules* 2024

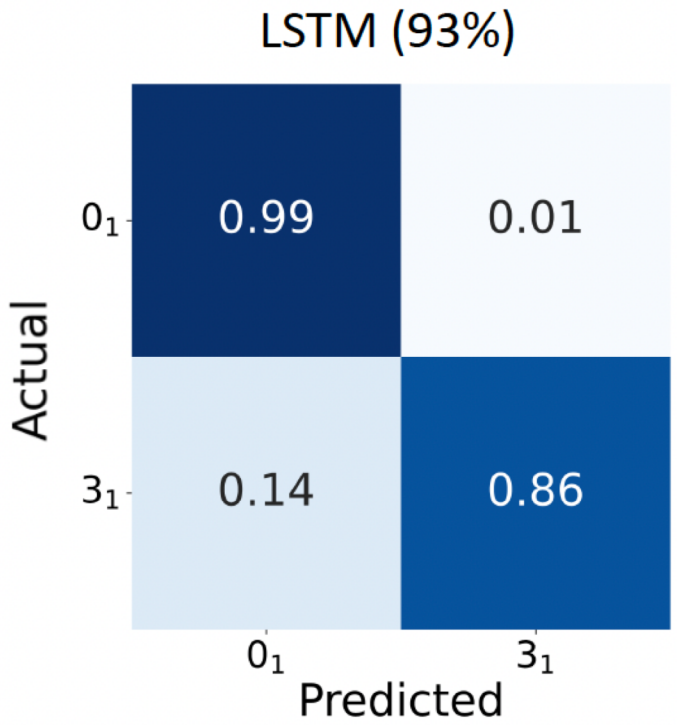
Proteins classification

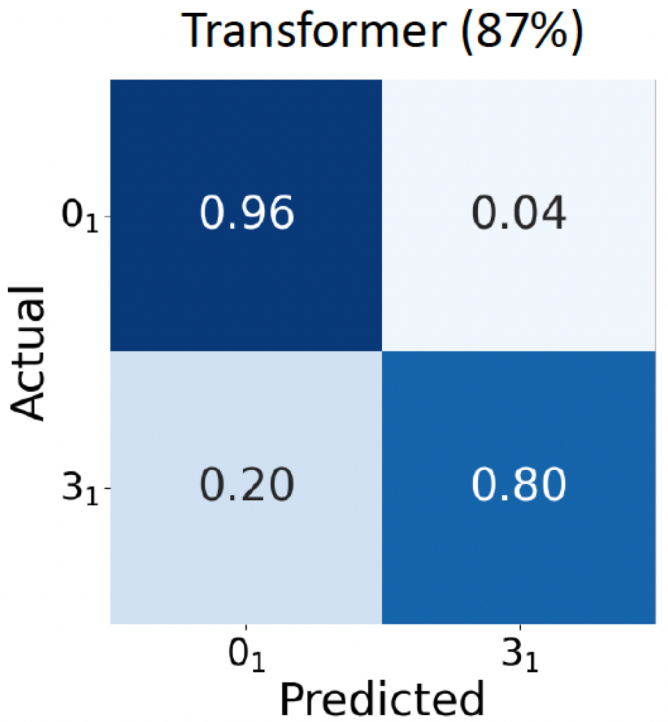
Training set assembled from real proteins set

Proteins data set prepared by **Paweł Rubach**, **Maciej Sikora** and **Fernando Bruno**

55025 unknots and **47862 trefoils** proteins in a training set (102887 in total)

Testing set assembled from **23571 unknotted** (***0_1***) and **20513 trefoil** (***3_1***) real proteins structures

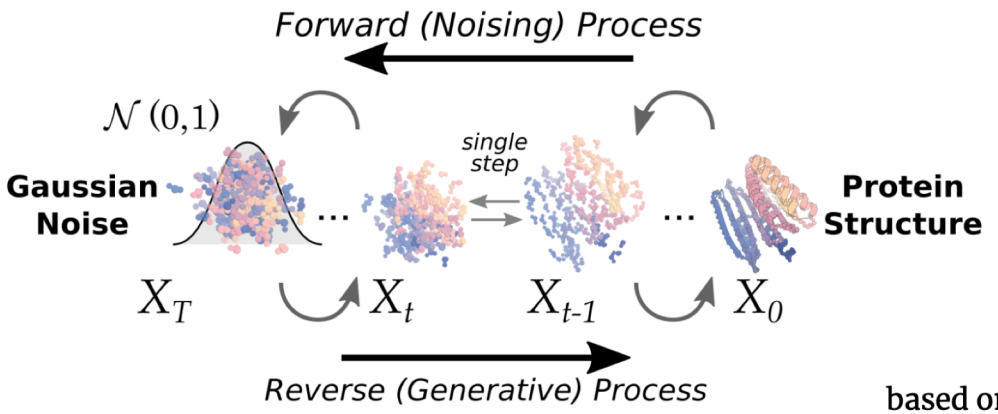




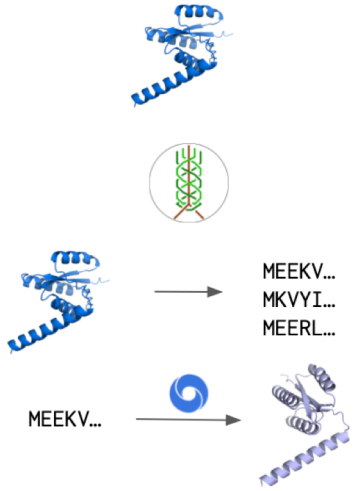
F. Bruno da Silva et. Al.., *Macromolecules* 2024

TOWARD ARITIFICALLY KNOTTED PROTEIN and RNA

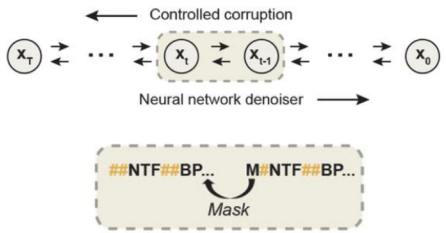
Diffusion for protein backbone: **RFdiffusion**

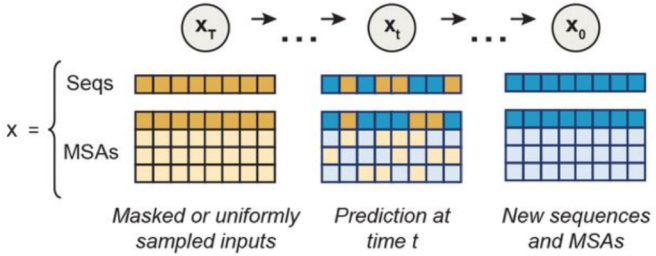


based on RoseTTAFold model

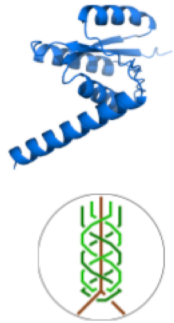


EvoDiff - directly generate sequence





MEEKVIEIR...



Eva Klimentová

References

[1] Eric J Rawdon, Kenneth C Millett, Joanna I. Sulkowska and Andrzej Stasiak. Knot localization in proteins. Biochemical Society Transactions, 41:538–41, 2013 Apr 2013. ISSN 1470-8752. doi: 10.1042/BST20120329

[2] Tobias C. Sayre, Toni M. Lee, Neil P. King, and Todd O. Yeates. Protein stabilization in a highly knotted protein polymer. Protein Engineering, Design and Selection, 24(8):627–630, 06 2011

[3] Szymon Niewieczerzał and Joanna I. Sulkowska. Supercoiling in a protein increases its stability. Phys. Rev. Lett.,123:138102, Sep 2019

[4] Joanna Sulkowska, Piotr Sulkowski, Piotr Szymczak, and Marek Cieplak. Stabilizing effect of knots on proteins. Proceedings of the National Academy of Sciences of the United States of America, 105:19714–9, 01 2009

[5] Thomas Christian, Reiko Sakaguchi, Agata Perlinska, Georges Lahoud, Takuhiro Ito, Erika Taylor, Shigeyuki Yokoyama, Joanna Sulkowska, and Ya-Ming Hou. Methyl transfer by substrate signaling from a knotted protein fold. Nature Structural Molecular Biology, 23, 08 2016

[6] Olafs Vandans, Kaiyuan Yang, Zhongtao Wu, and Liang Dai. Identifying knot types of polimer conformations by machine learning. Physical Review E, 101(2), feb 2020

[7] Machine Learning Understands Knotted Polymers, A. Braghetto et al. *Macromolecules* 2023, 56, 7, 2899–2909

[8] Knots and θ-Curves Identification in Polymeric Chains and Native Proteins Using Neural Networks
Fernando Bruno da Silva, *Macromolecules* 2024