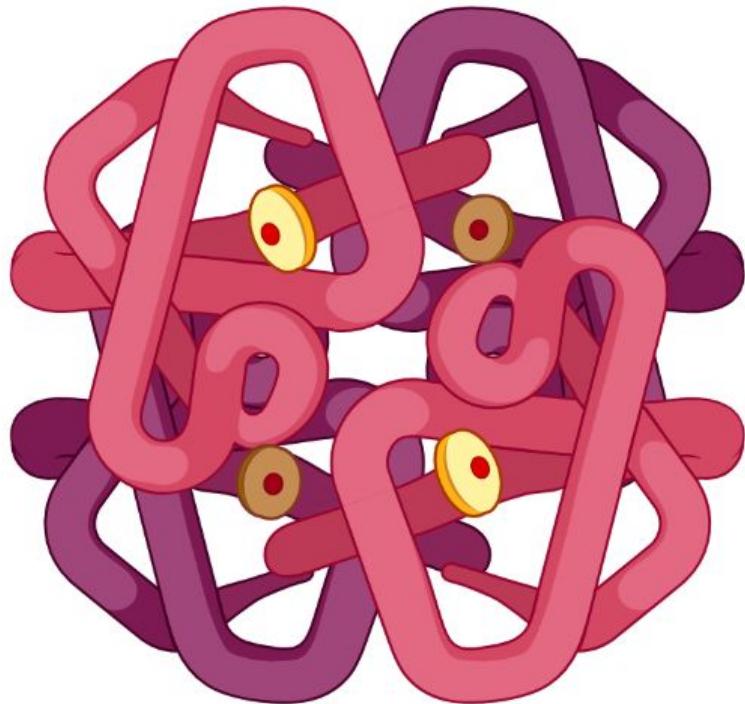


Using NLP to detect knots in protein structures

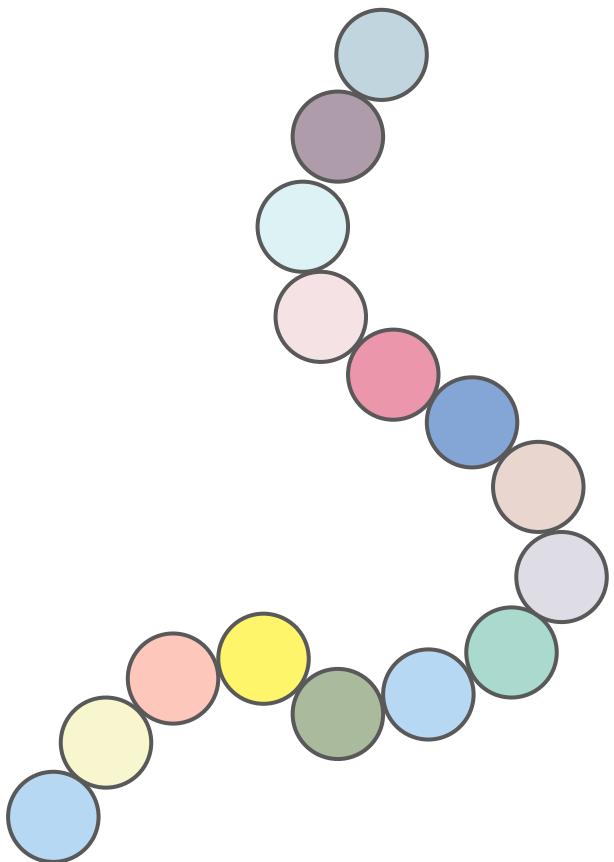
Eva Klimentová

Proteins



hemoglobin

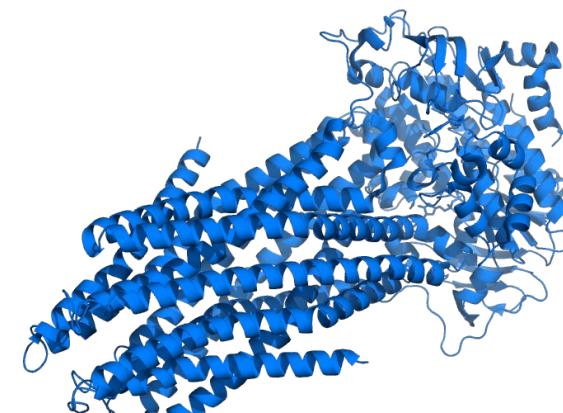
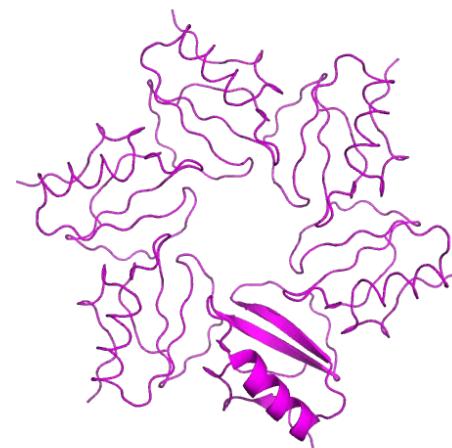
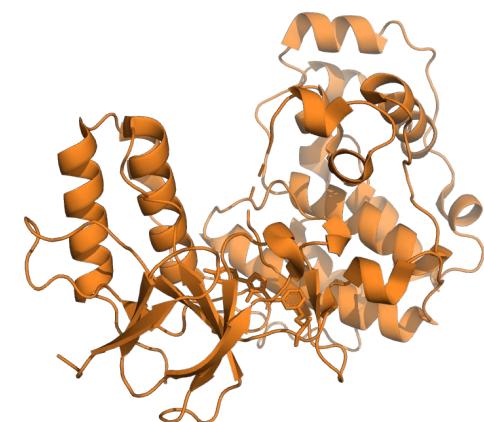
Amino acids = building blocks

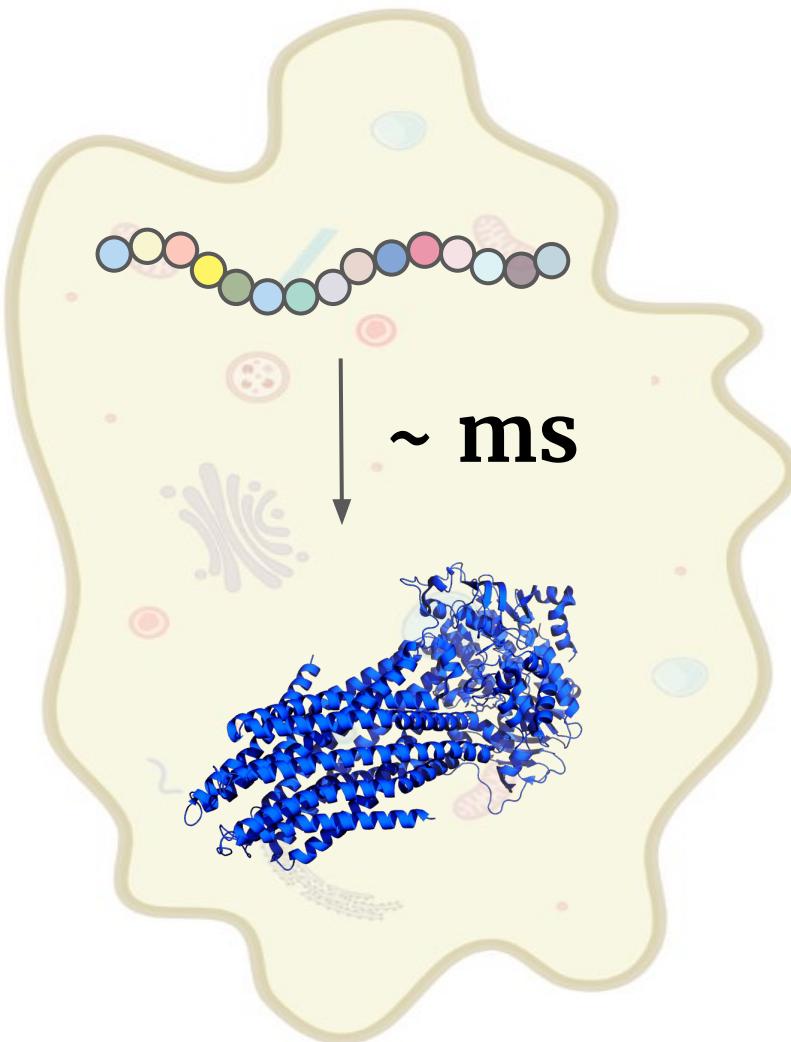


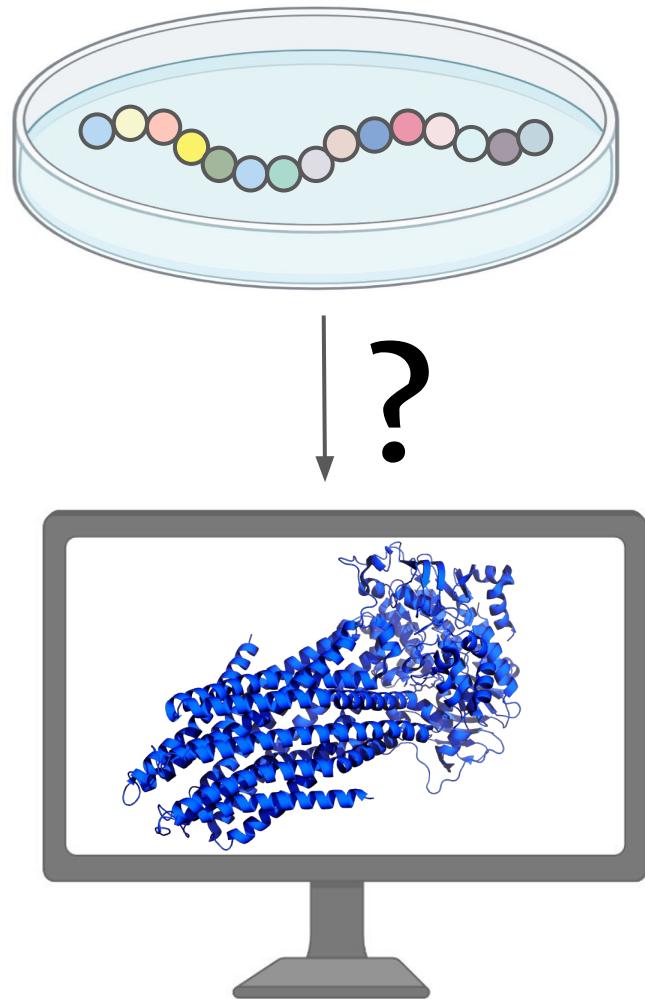
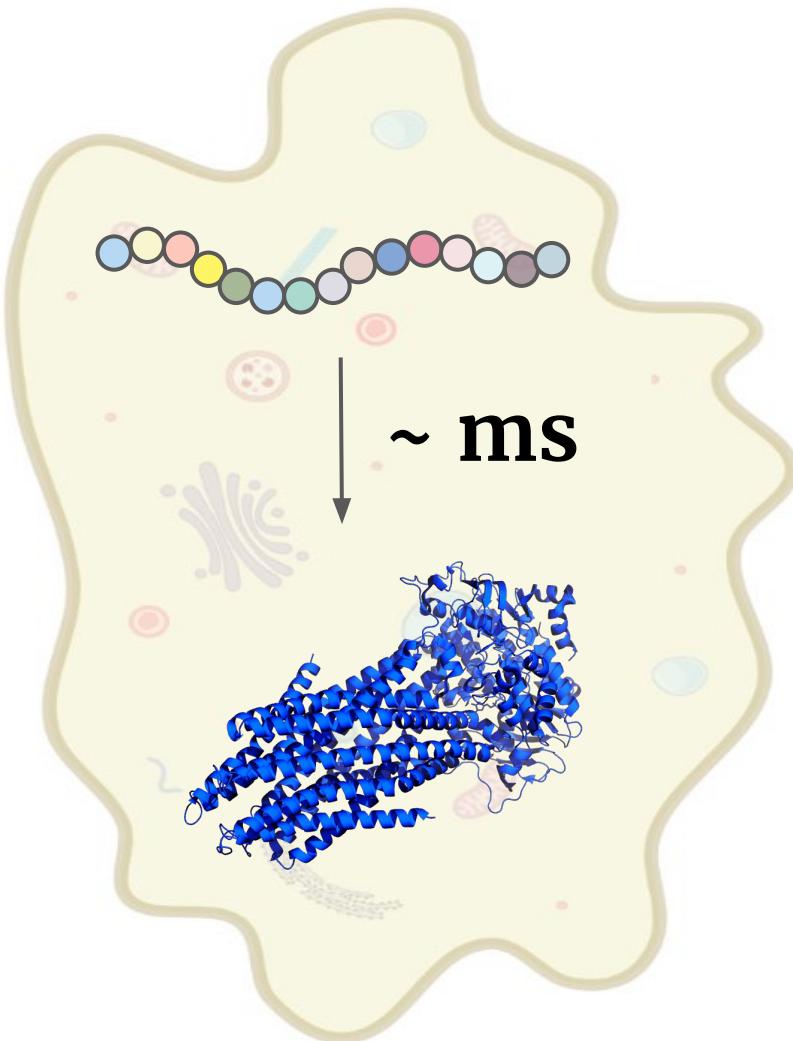
alanine - ala - A	leucine - leu - L
arginine - arg - R	lysine - lys - K
asparagine - asn - N	methionine - met - M
aspartic acid - asp - D	phenylalanine - phe - F
cysteine - cys - C	proline - pro - P
glutamine - gln - Q	serine - ser - S
glutamic acid - glu - E	threonine - thr - T
glycine - gly - G	tryptophan - trp - W
histidine - his - H	tyrosine - tyr - Y
isoleucine - ile - I	valine - val - V

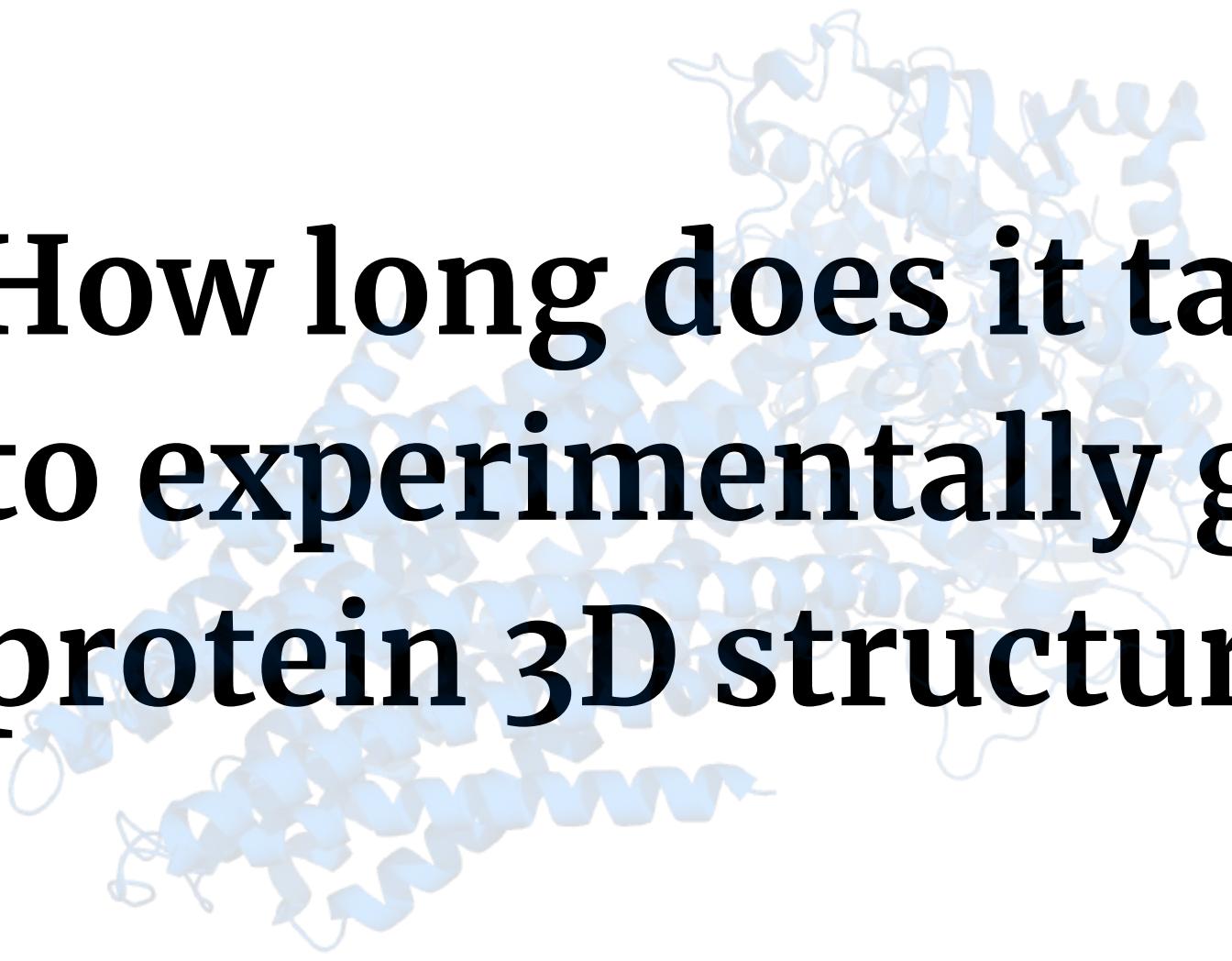
sequence

structure

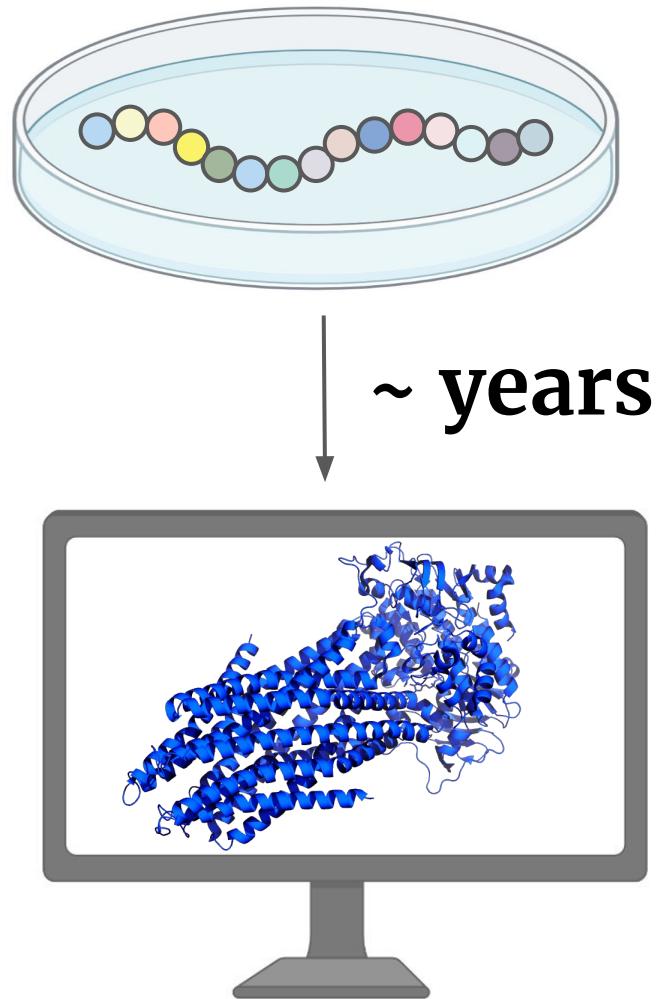
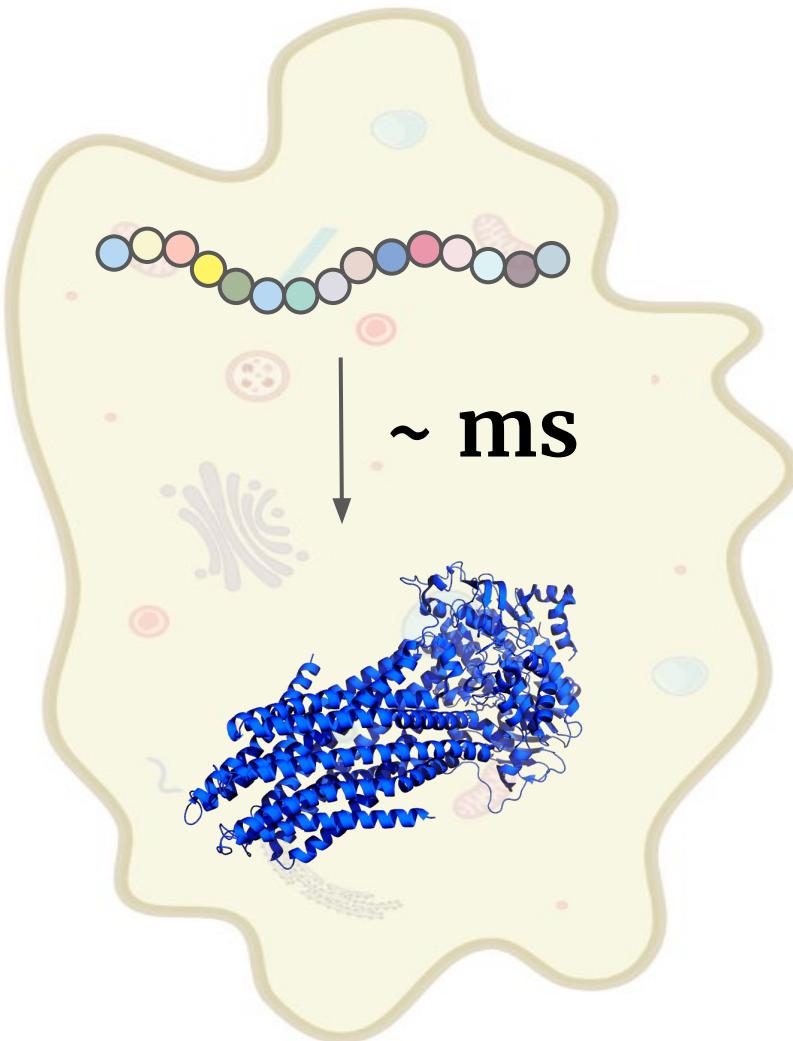


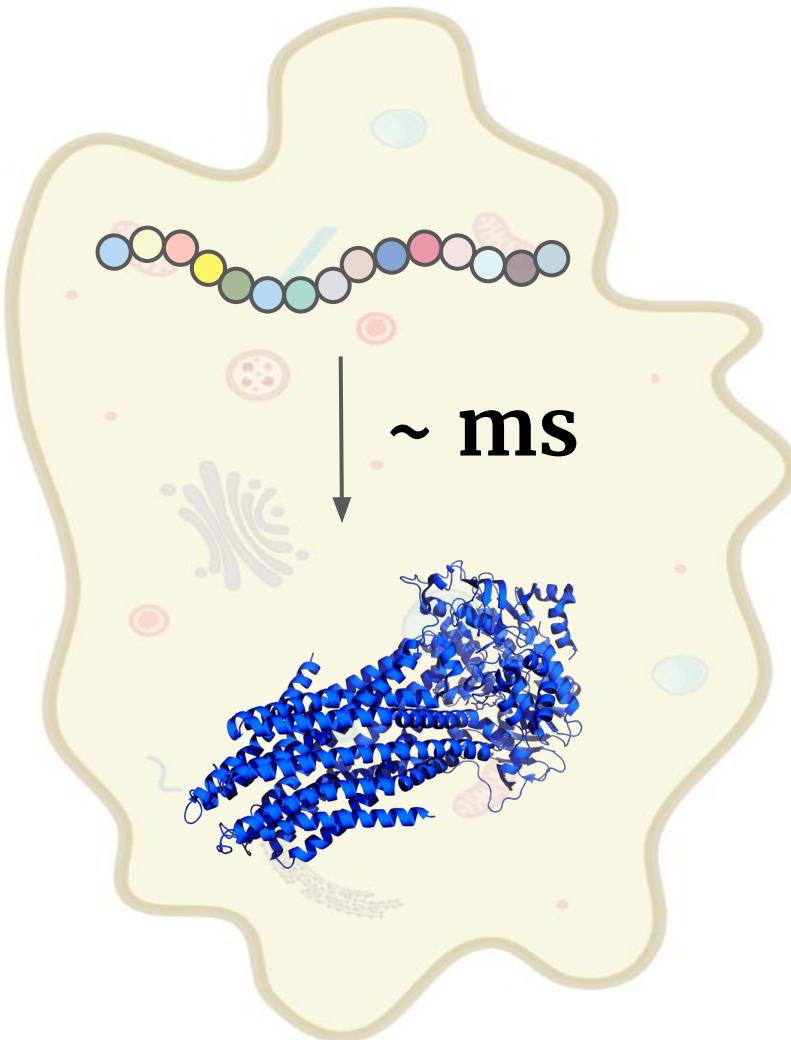




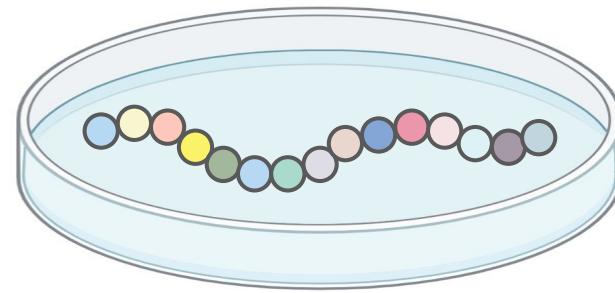


**How long does it take
to experimentally get
protein 3D structure?**

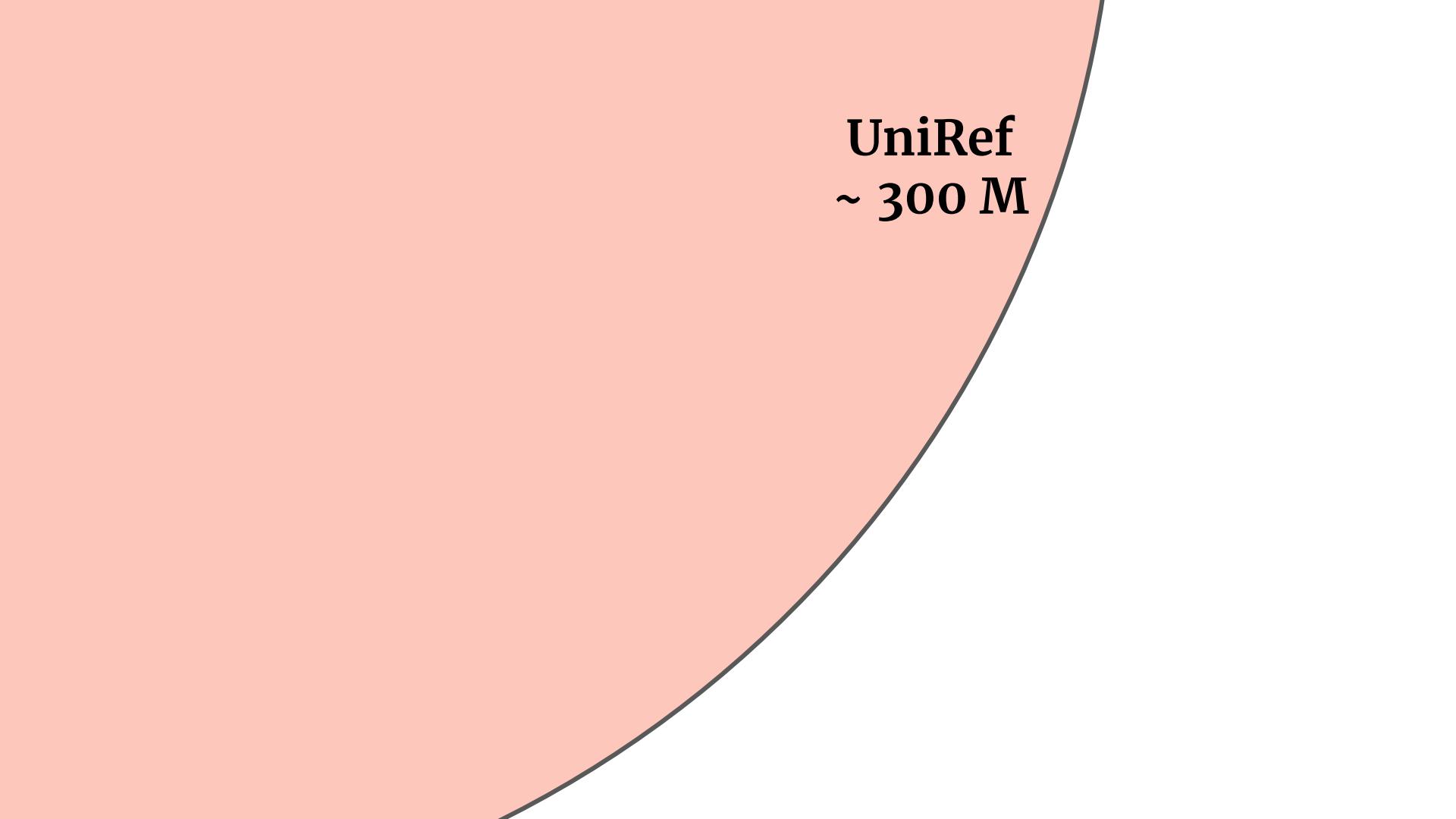




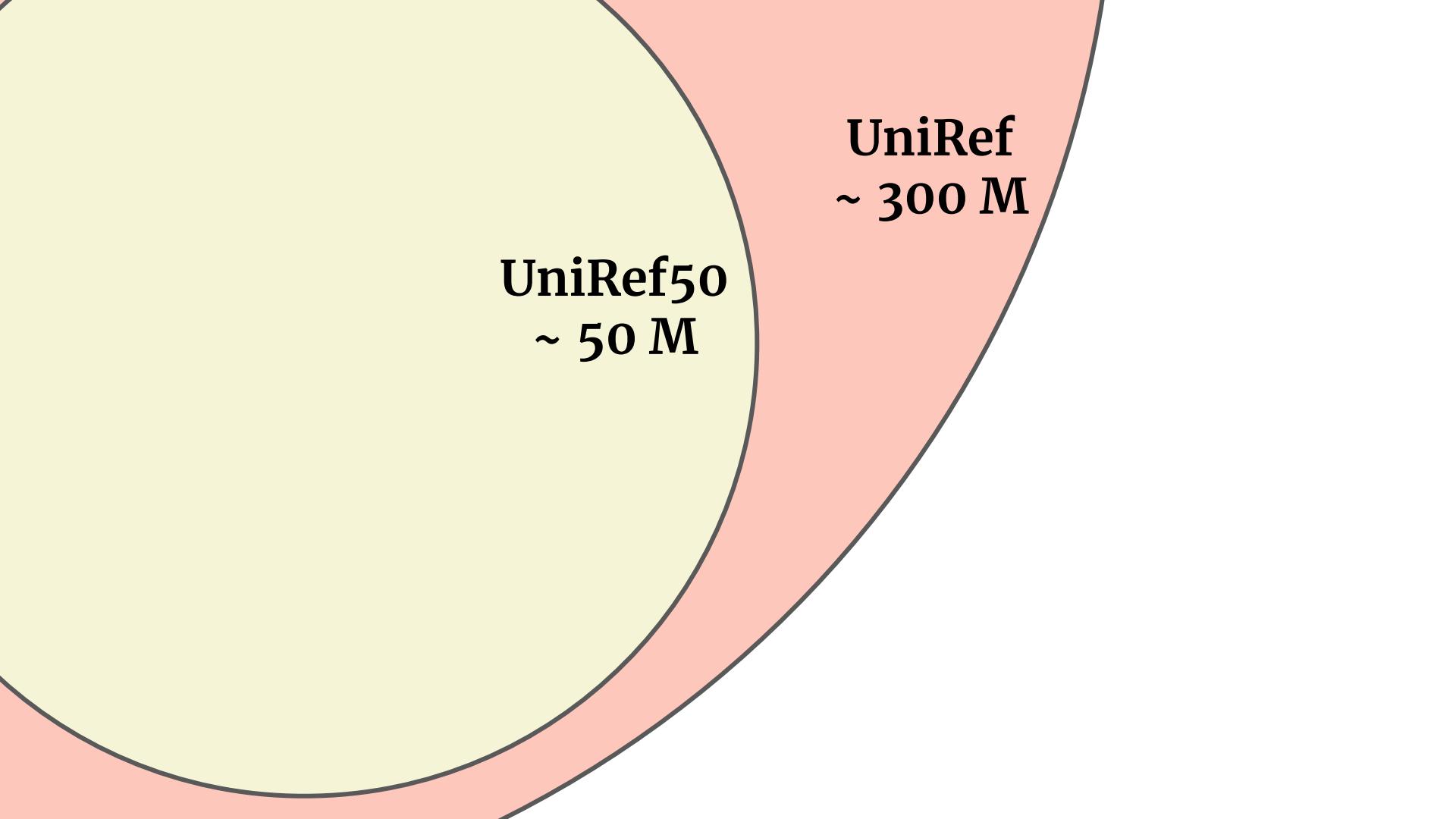
~ ms



PhD life ~ years

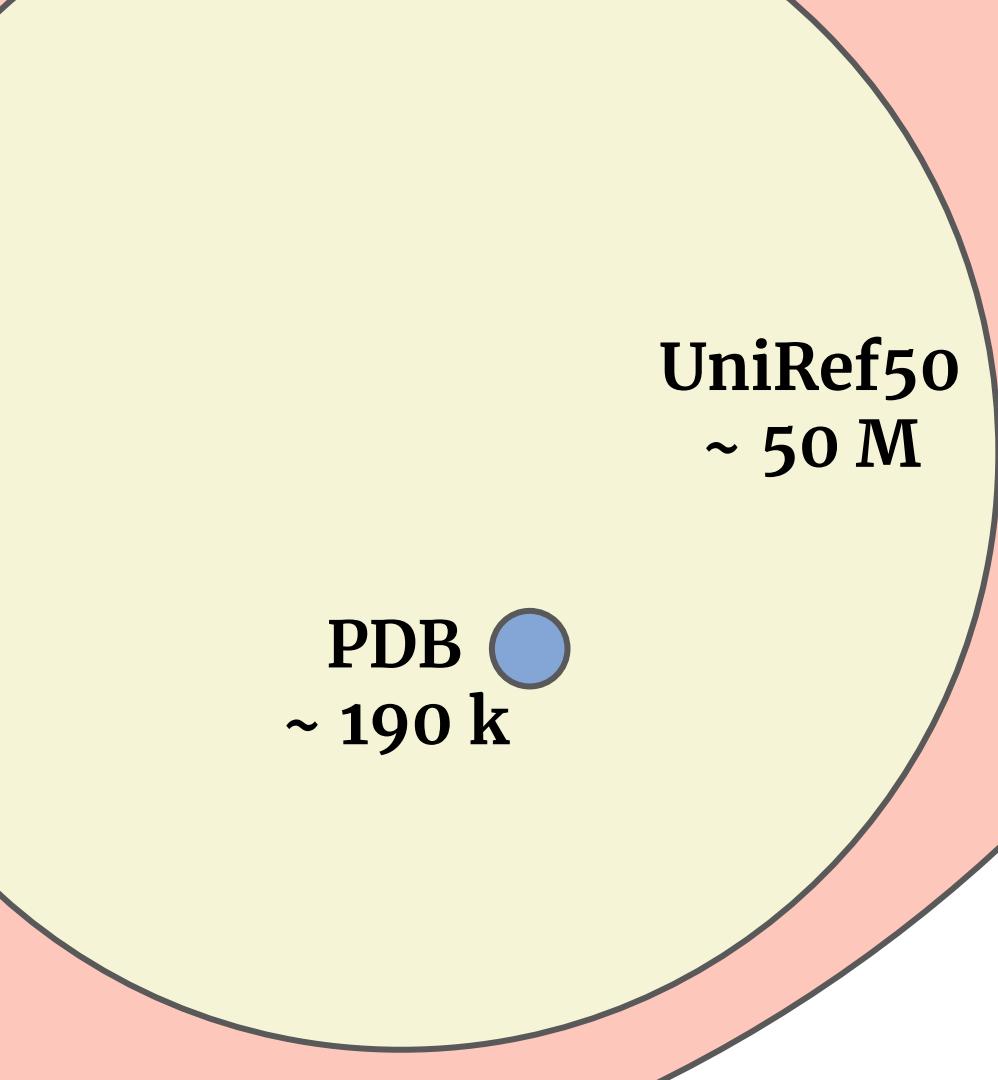
A large, solid orange triangle is positioned in the upper right quadrant of the image. Its base is a thin black diagonal line extending from the bottom center towards the top right corner. The top vertex of the triangle is at the top edge of the frame.

UniRef
~ 300 M

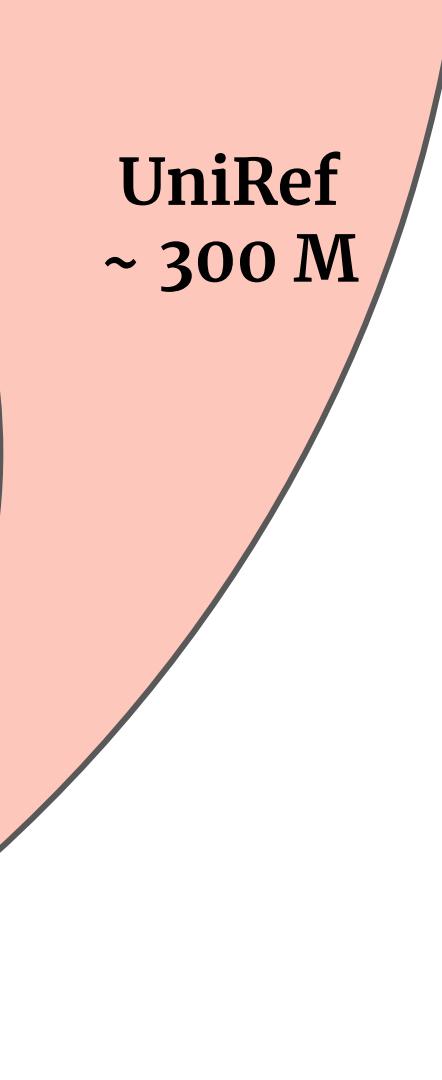


UniRef50
~ 50 M

UniRef
~ 300 M



PDB
~ 190 k

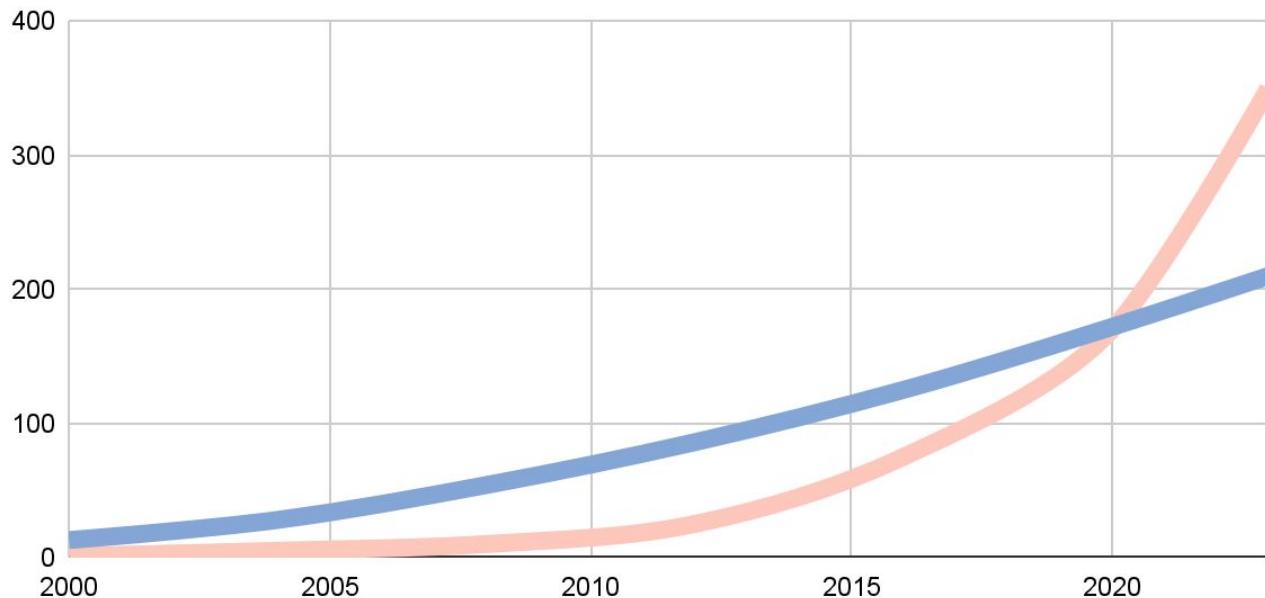


UniRef50
~ 50 M

UniRef
~ 300 M

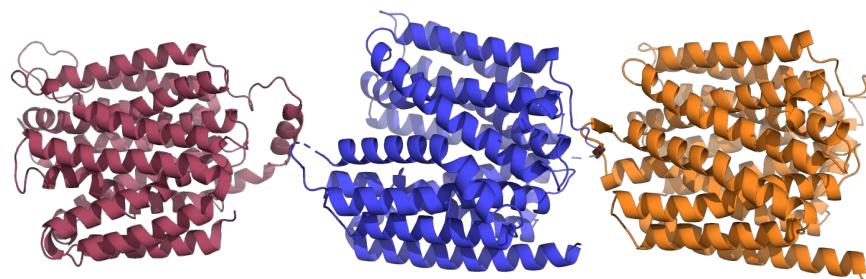
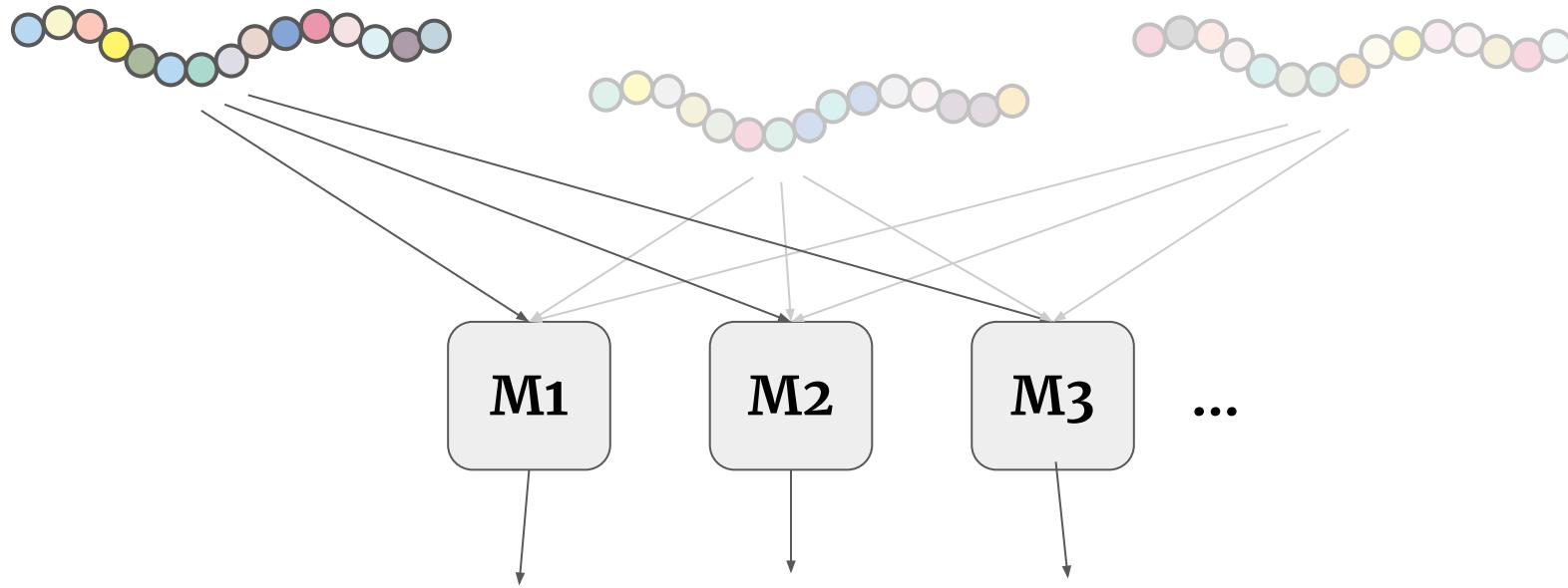
Growth of Protein Databases

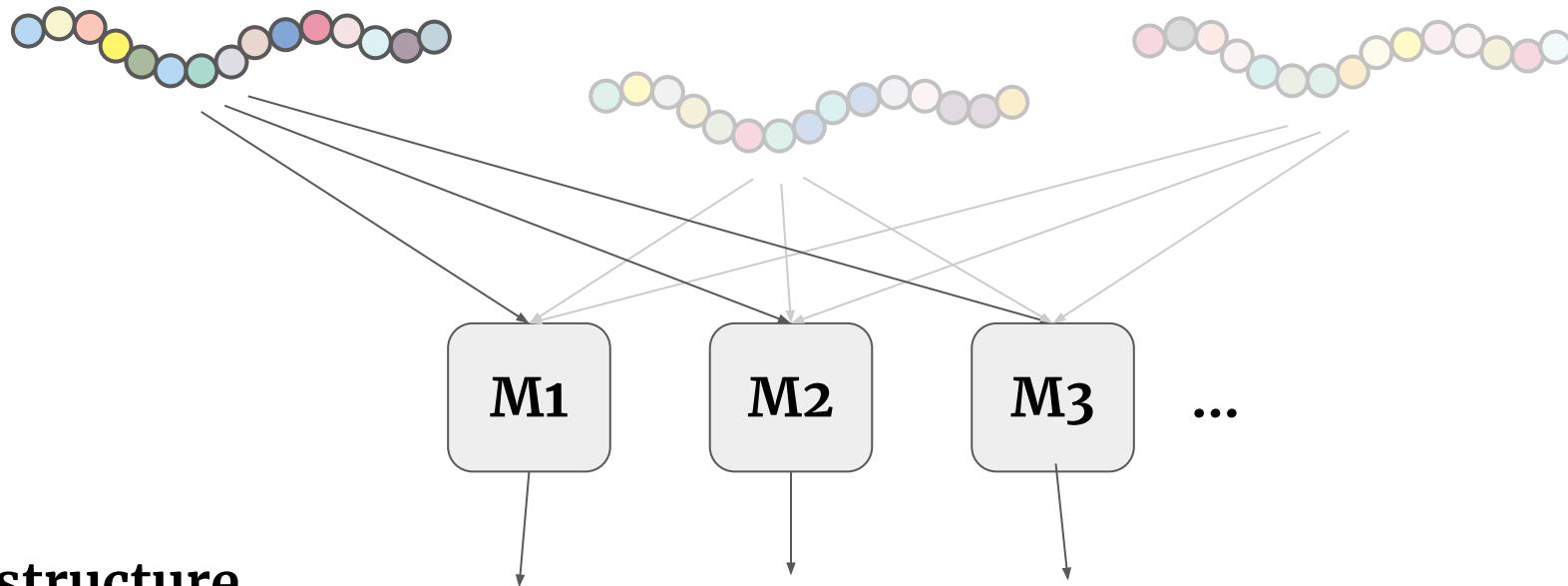
— MILLIONS of protein sequences
— THOUSANDS of protein 3D structures



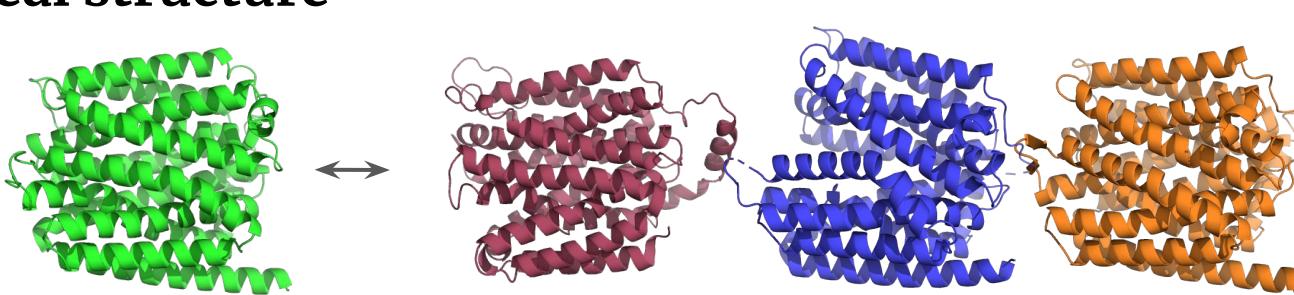
CASP competition



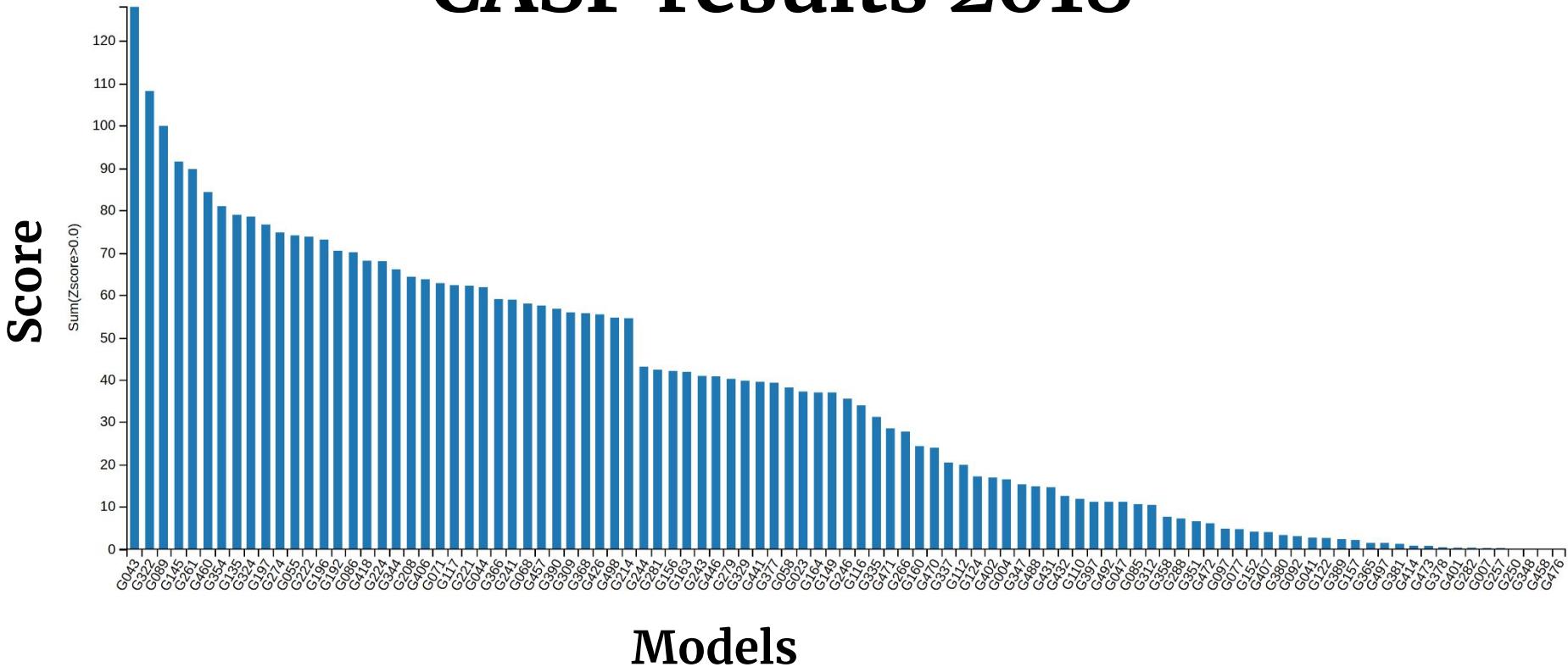




real structure

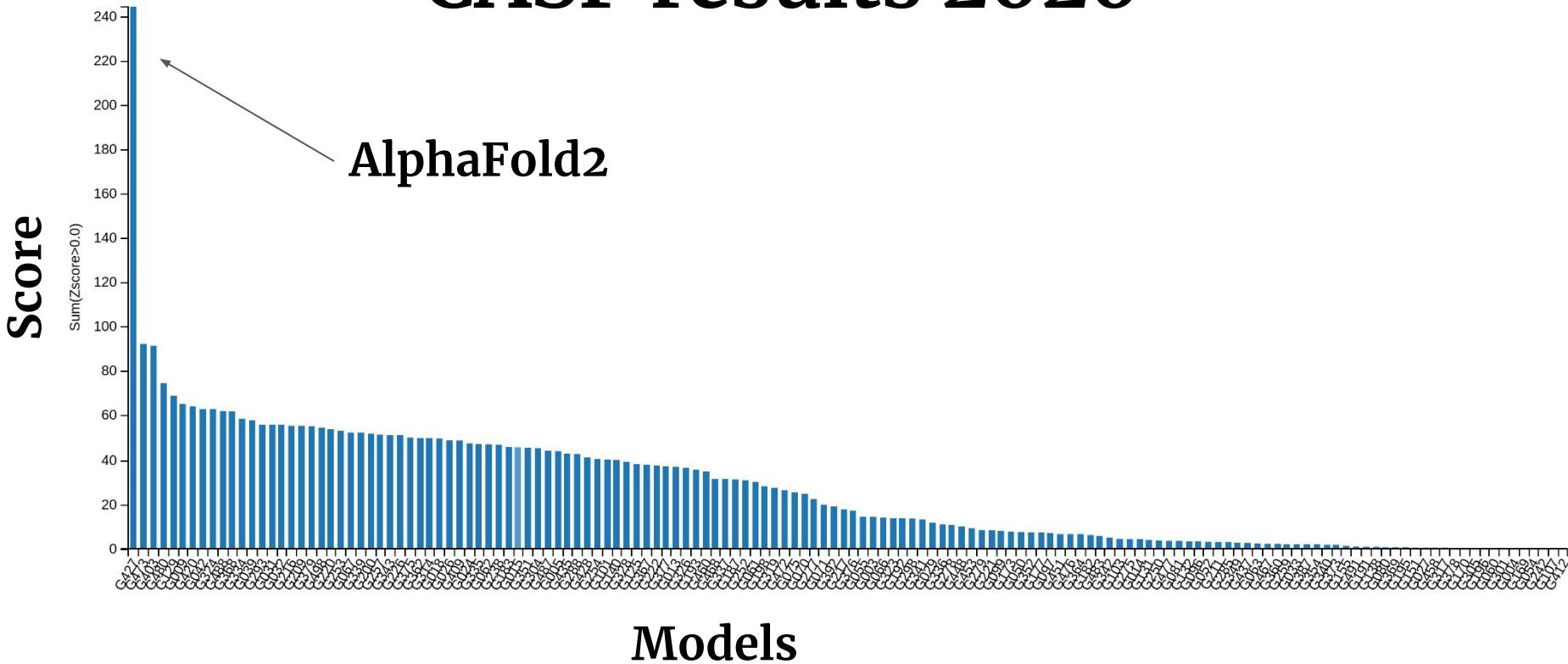


CASP results 2018

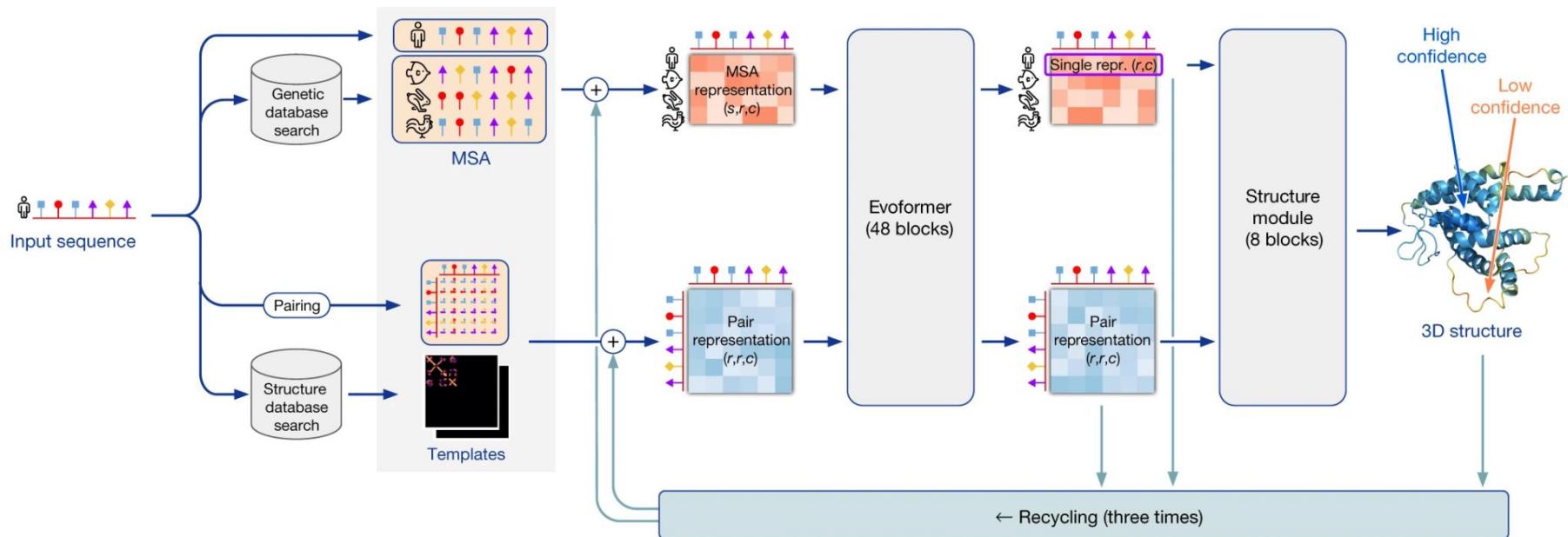


Models

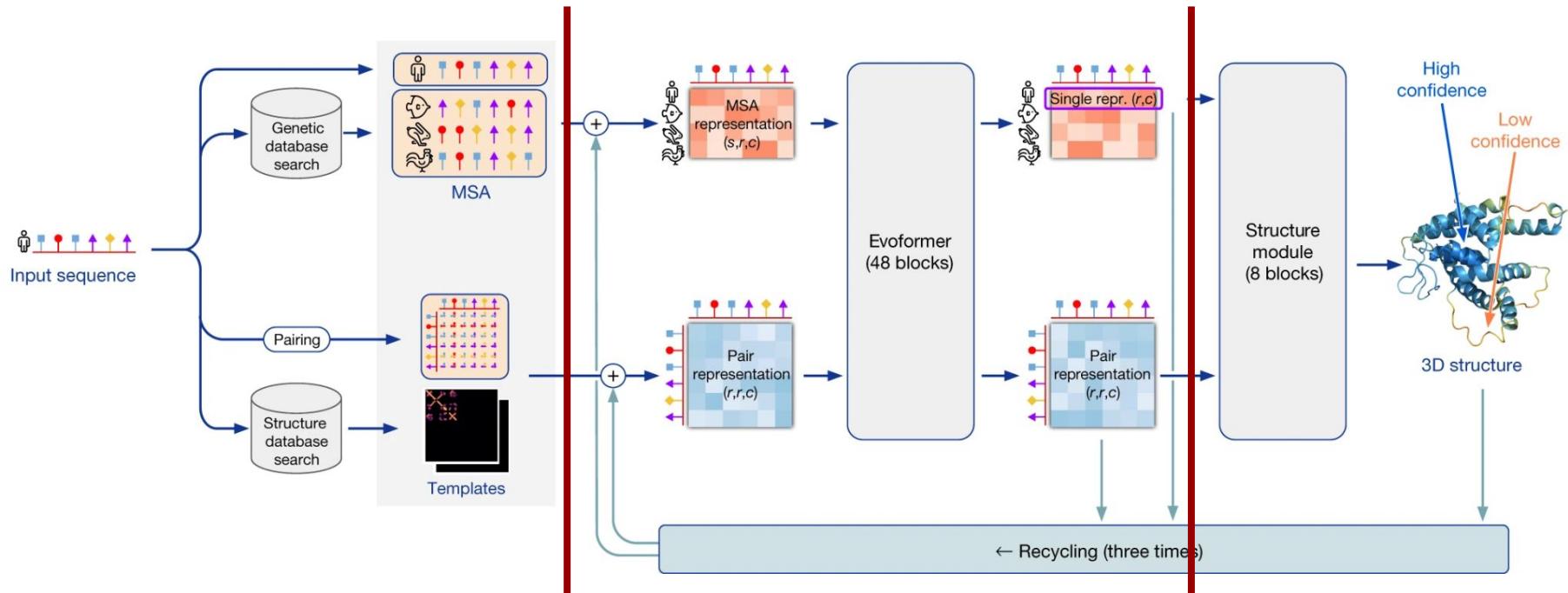
CASP results 2020



AlphaFold2



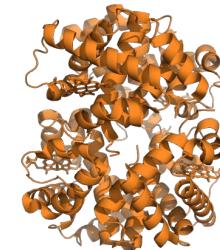
AlphaFold2



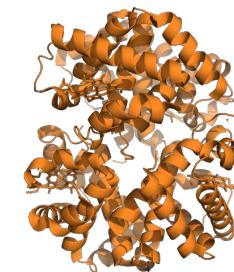
1. encoding



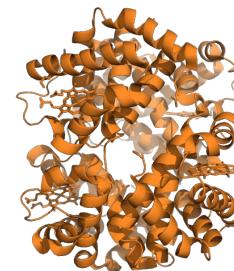
V L S P A D K T N V K A A W G K V G A H ...



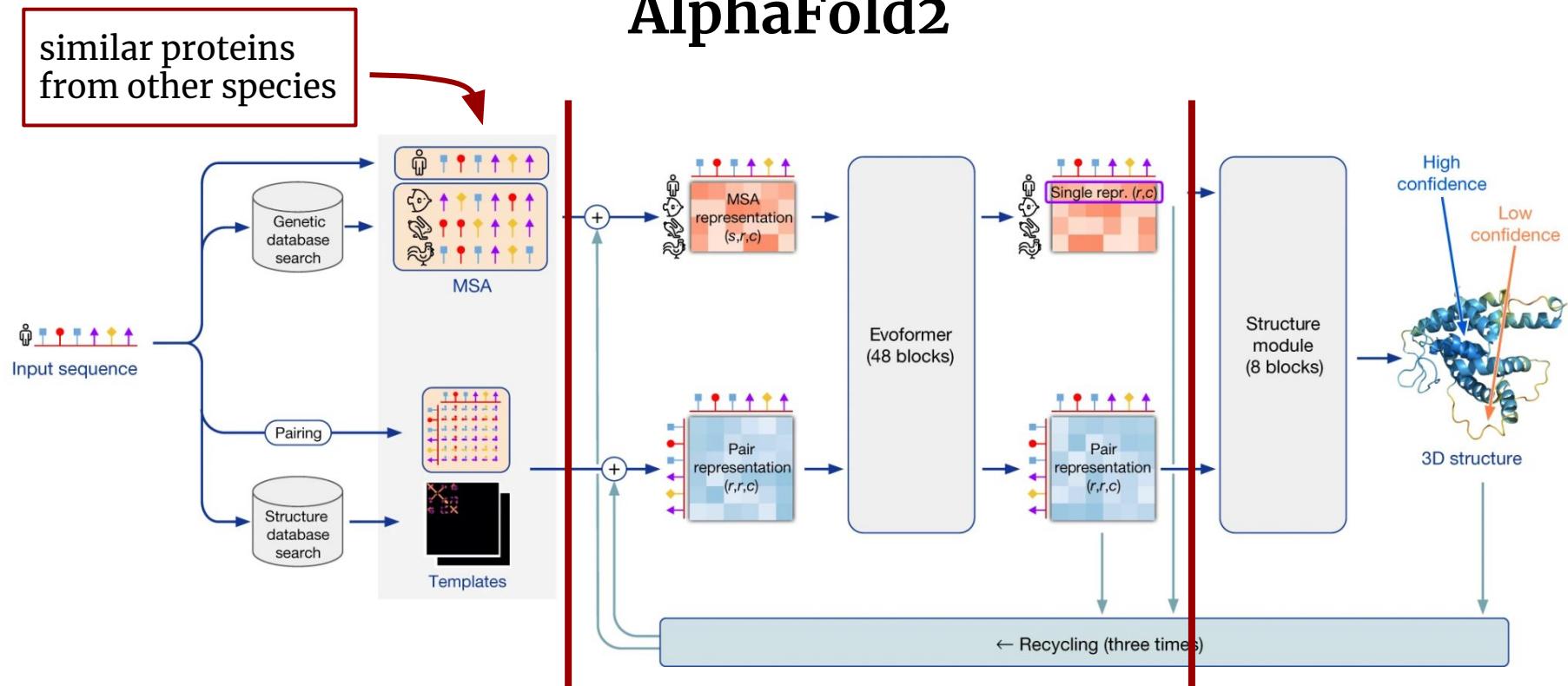
V L S A D D K T N I K N C W G K I G G H ...



X S L S A K D K A N V K A I W G K I L P K ...

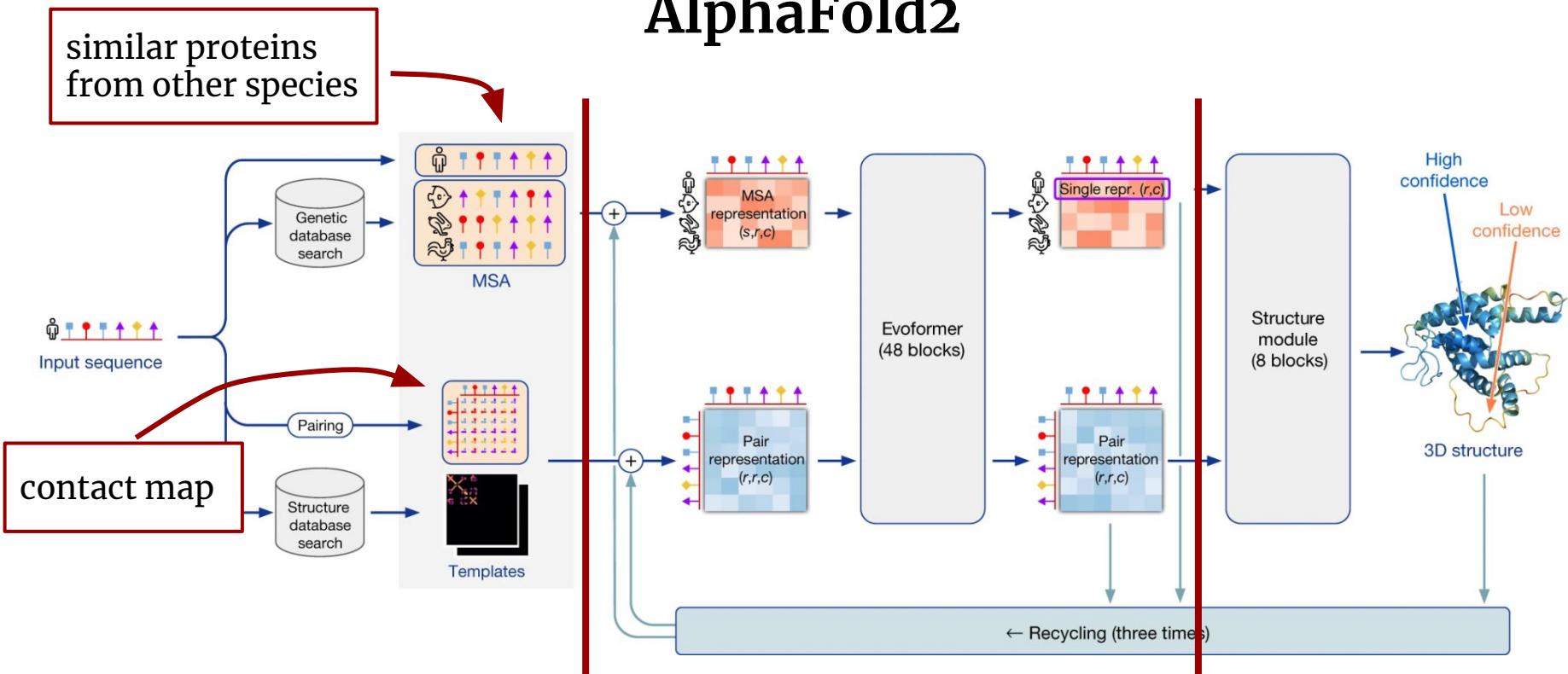


AlphaFold2



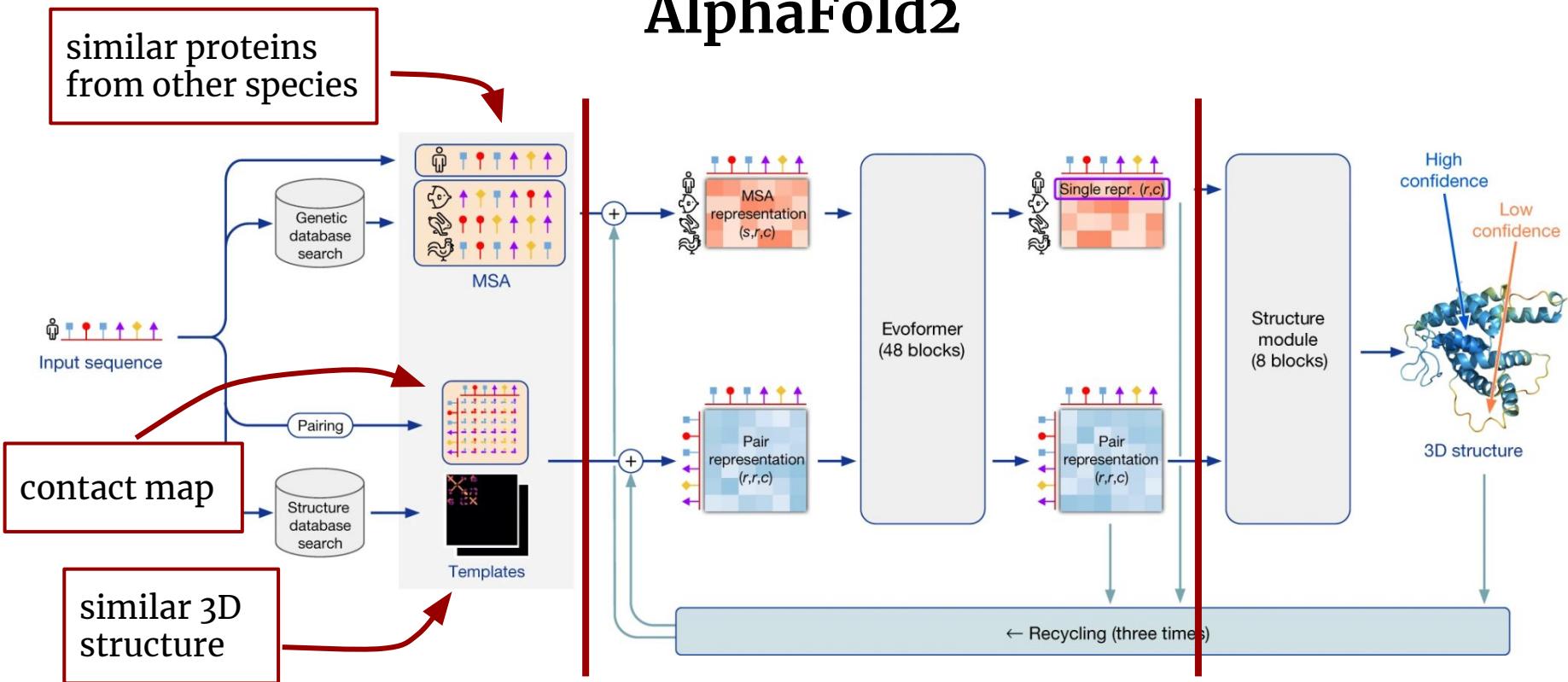
1. encoding

AlphaFold2



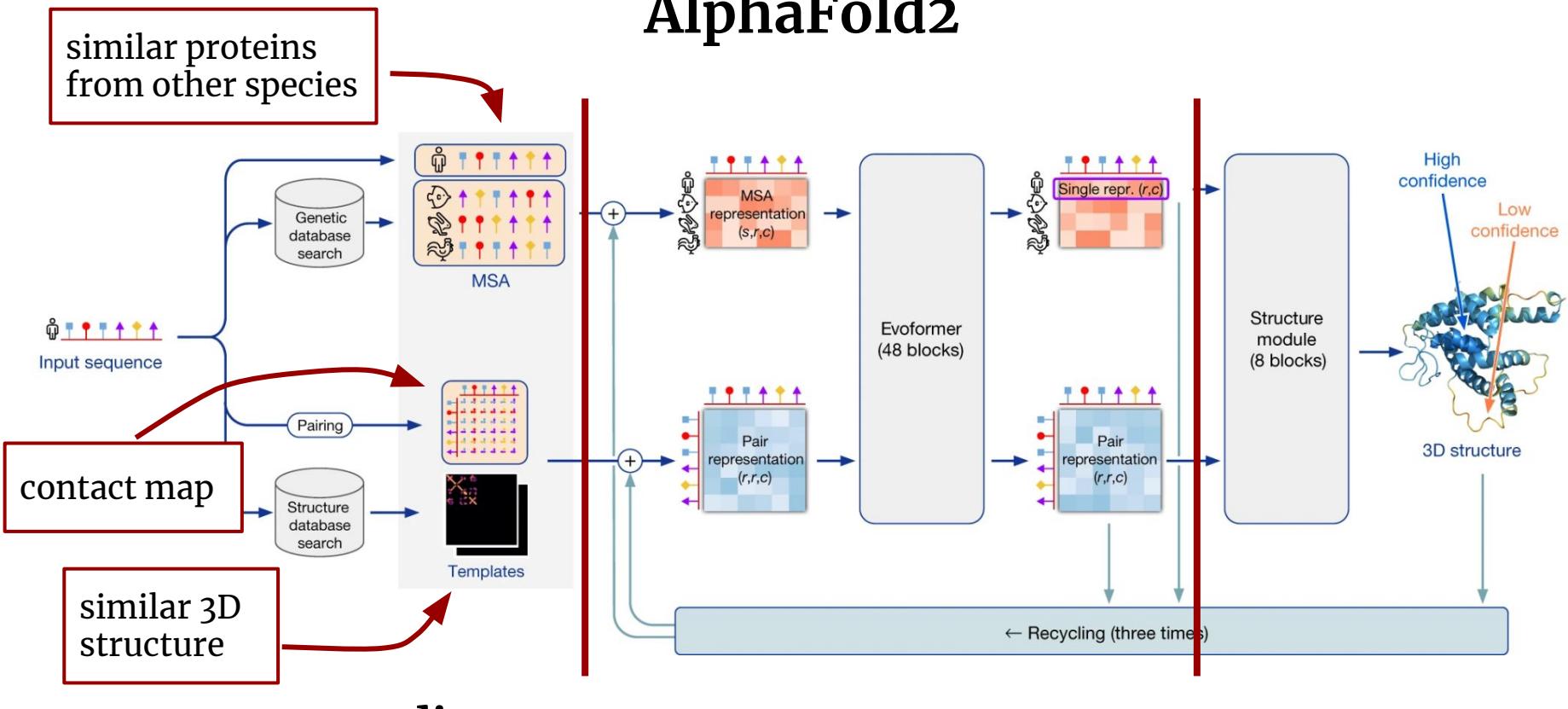
1. encoding

AlphaFold2



1. encoding

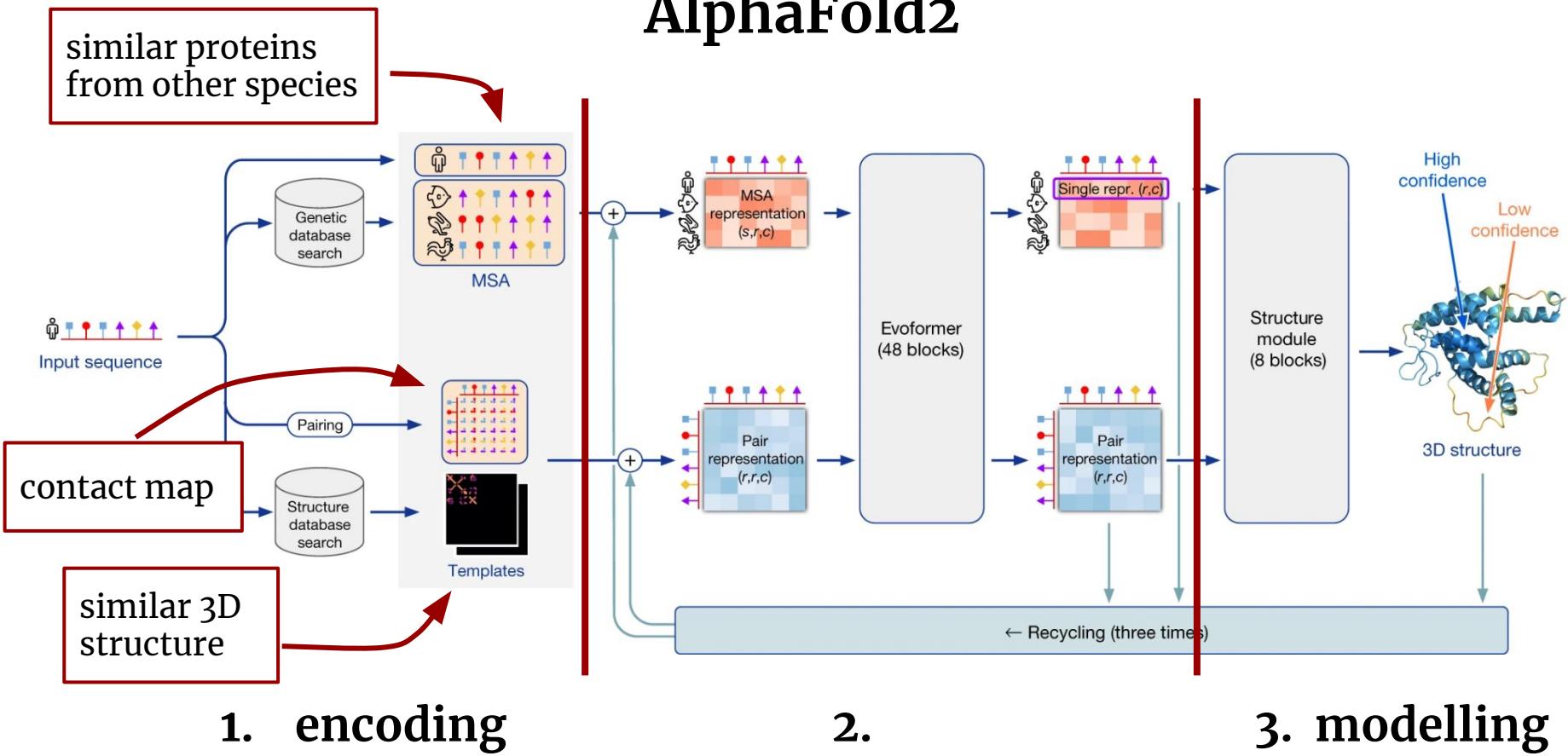
AlphaFold2

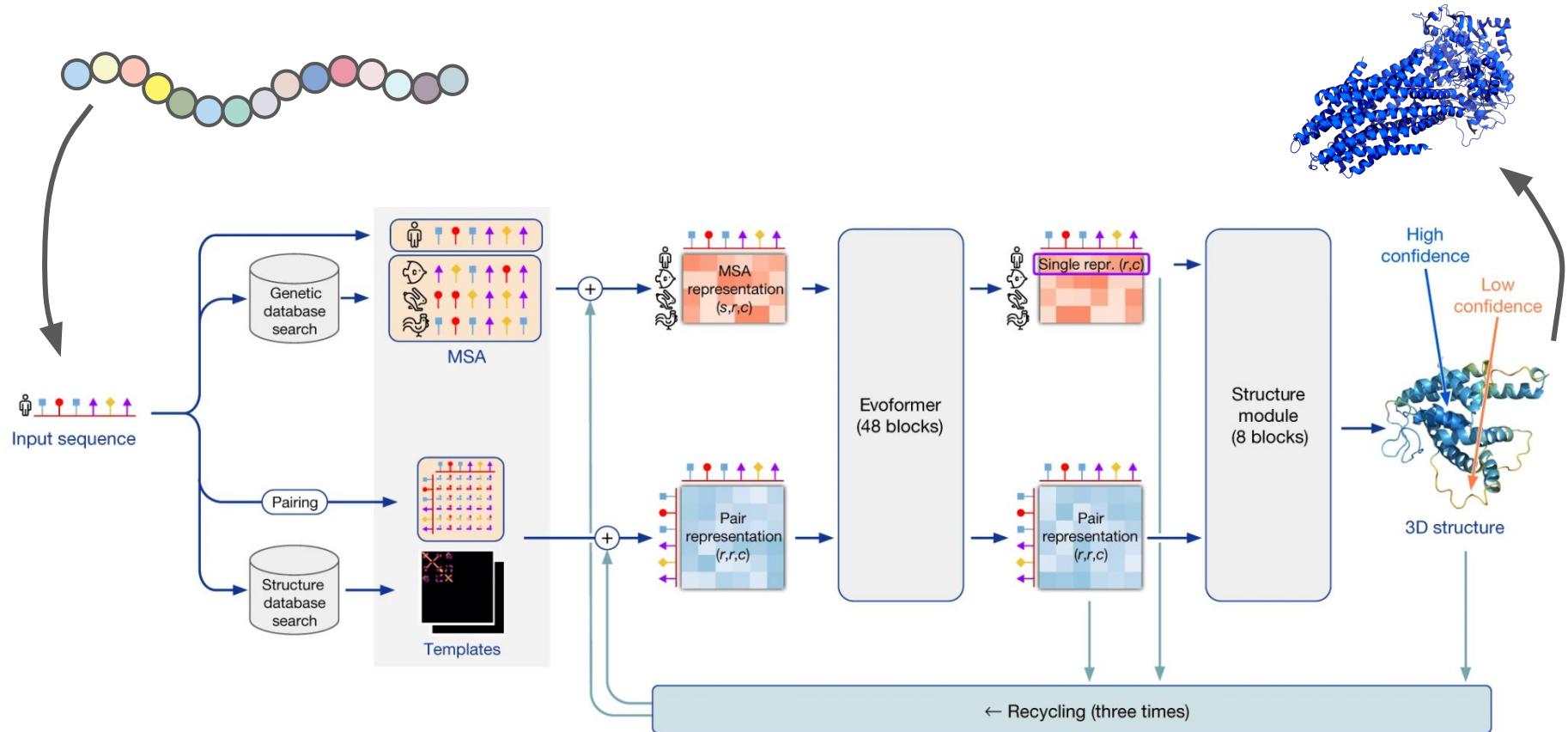


1. encoding

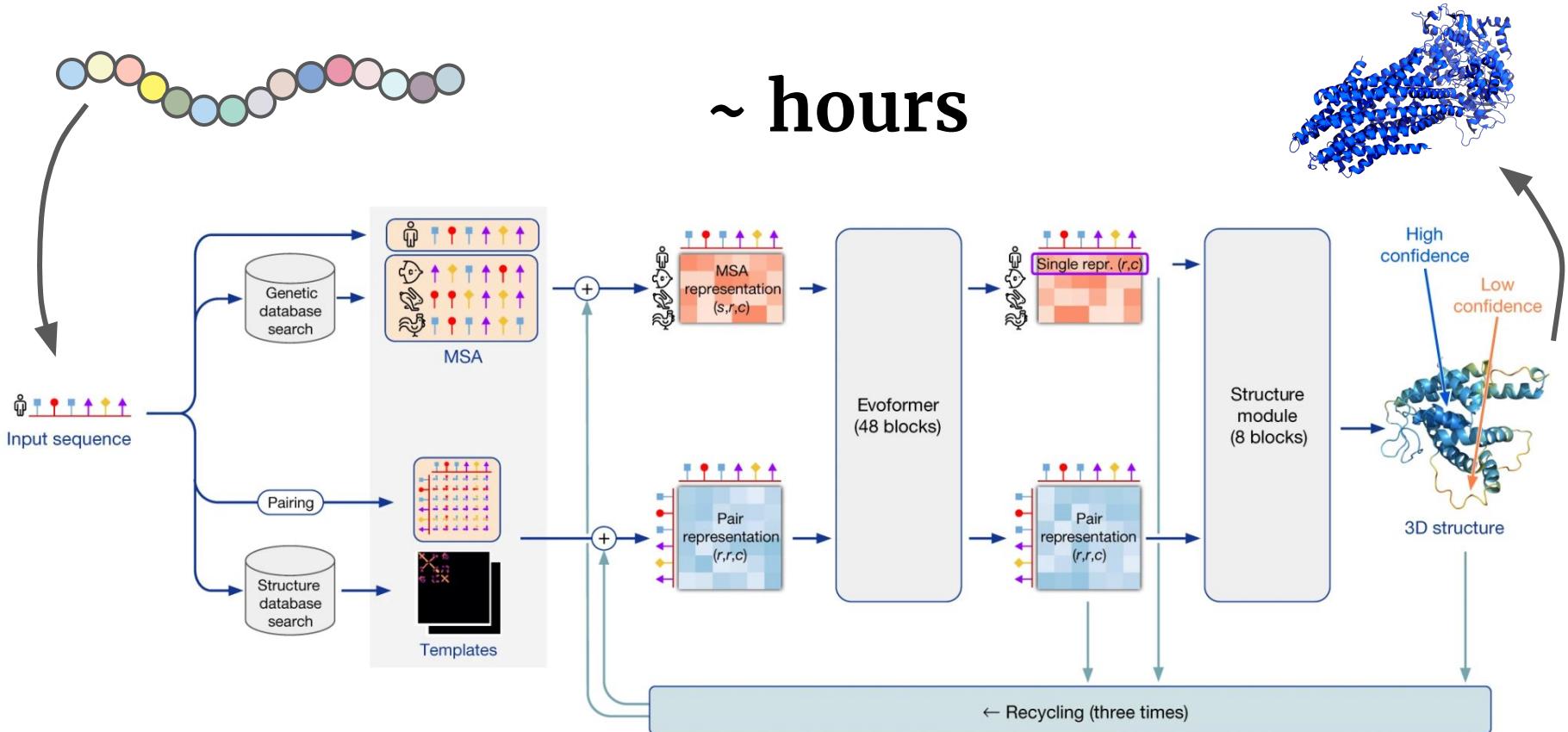
2.

AlphaFold2

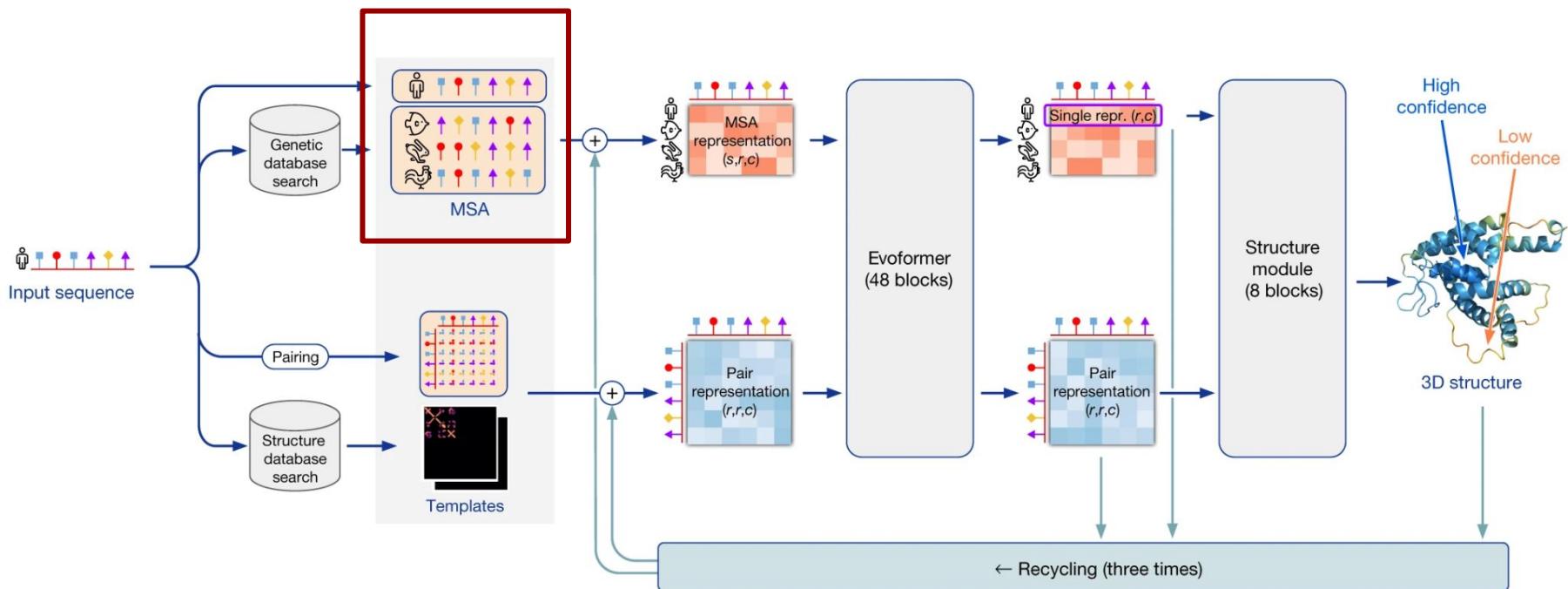




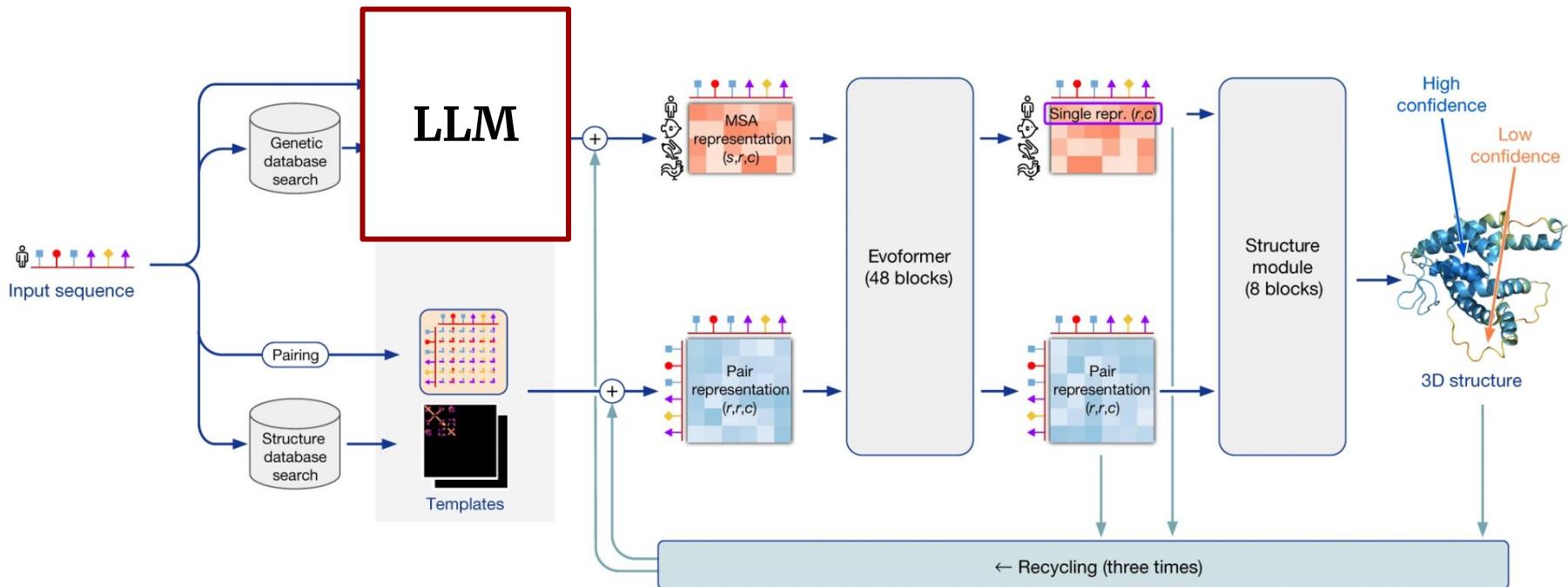
~ hours



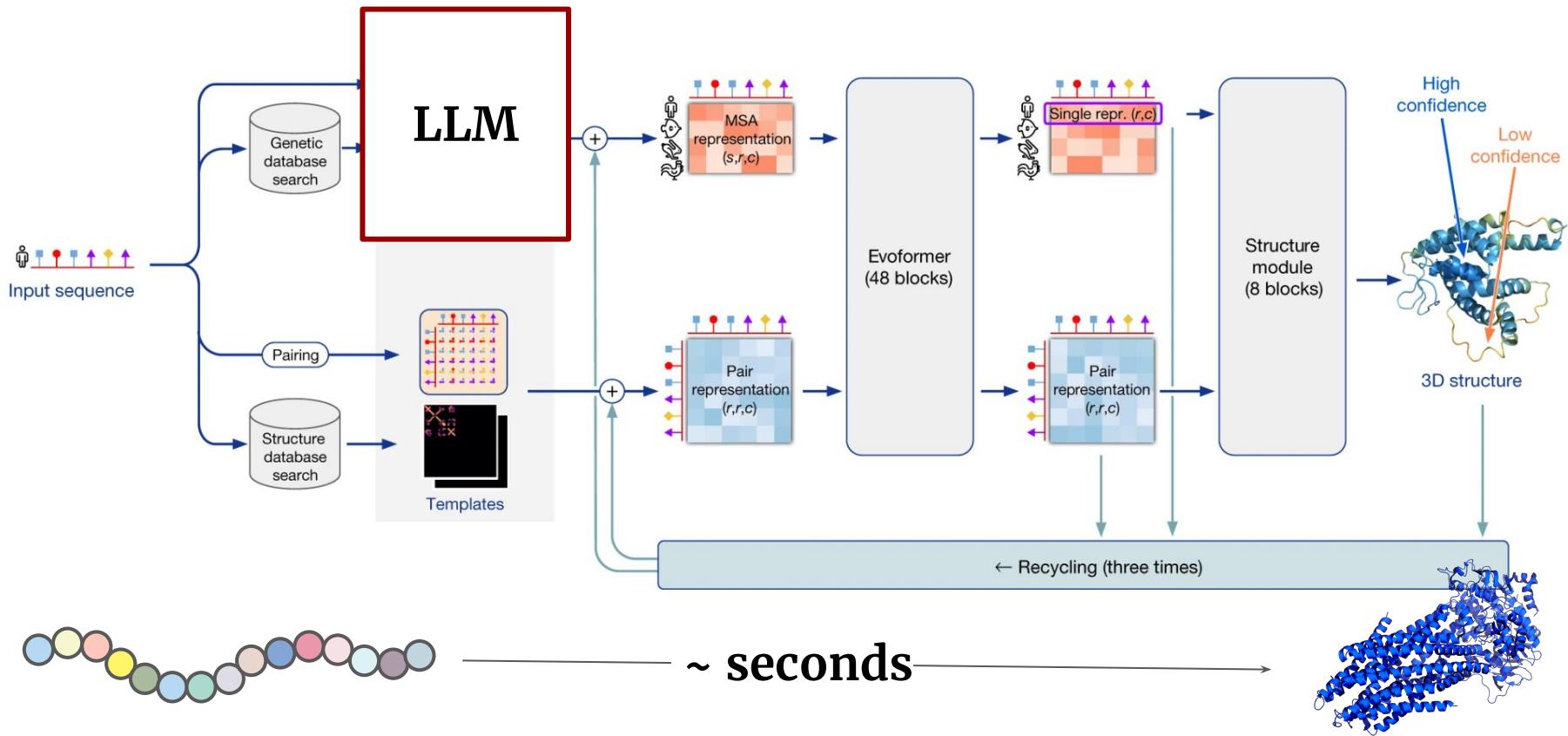
slow



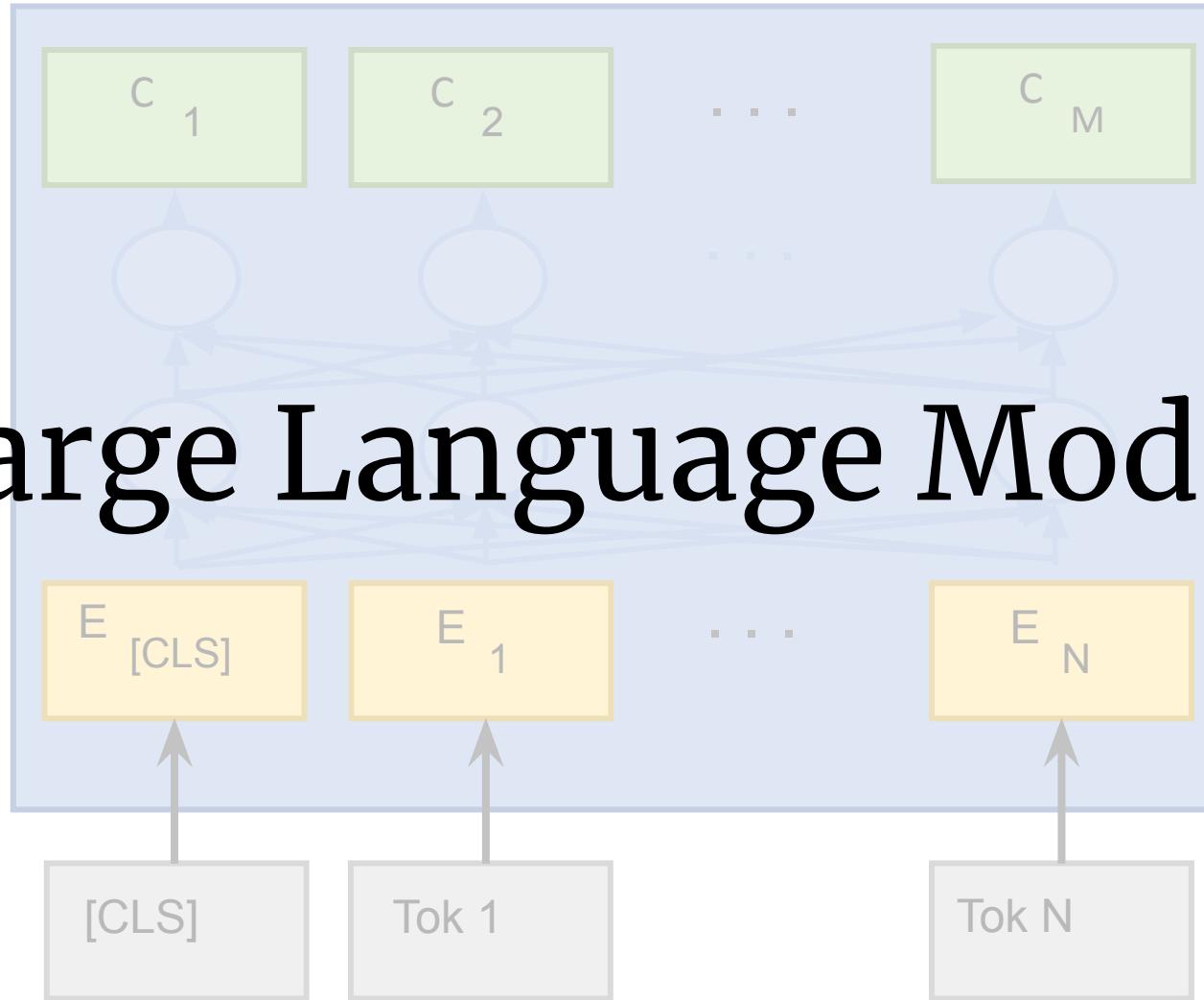
fast ESMfold



fast ESMfold



Large Language Models



big dataset

There

was

a

king

who

had

twelve

beautiful

daughters



M



daughters

children

women

...

0.47

0.28

0.12

This movie is an incredible piece of work. It explores every nook and cranny of the human mind ..

positive

The Best Movie of the 90's"... Aye, right! I went into this movie with pretty high expectations, and it was all downhill from there...

negative

finetune

M
pretrained



Hugging Face

 Hugging Face



Models

Datasets

Spaces

Docs

Solutions

Pricing



Log In

Sign Up

Tasks

Libraries

Datasets

Languages

Licenses

Other

Filter Tasks by name

Multimodal

Feature Extraction

Text-to-Image

Image-to-Text

Text-to-Video

Visual Question Answering

Document Question Answering

Graph Machine Learning

Computer Vision

Depth Estimation

Image Classification

Object Detection

Image Segmentation

Image-to-Image

Unconditional Image Generation

Video Classification

Models 234,899

Filter

Full-text search

Sort: Most Downloads

 jonatasgrosman/wav2vec2-large-xlsr-53-english

Updated Mar 25 · 71.9M · 182

 bert-base-uncased

Updated 26 days ago · 50.5M · 923

 xlm-roberta-large

Updated Apr 6 · 42.6M · 160

 gpt2

Updated Dec 16, 2022 · 17.3M · 1.18k

 openai/clip-vit-large-patch14

Updated Oct 4, 2022 · 16.8M · 460

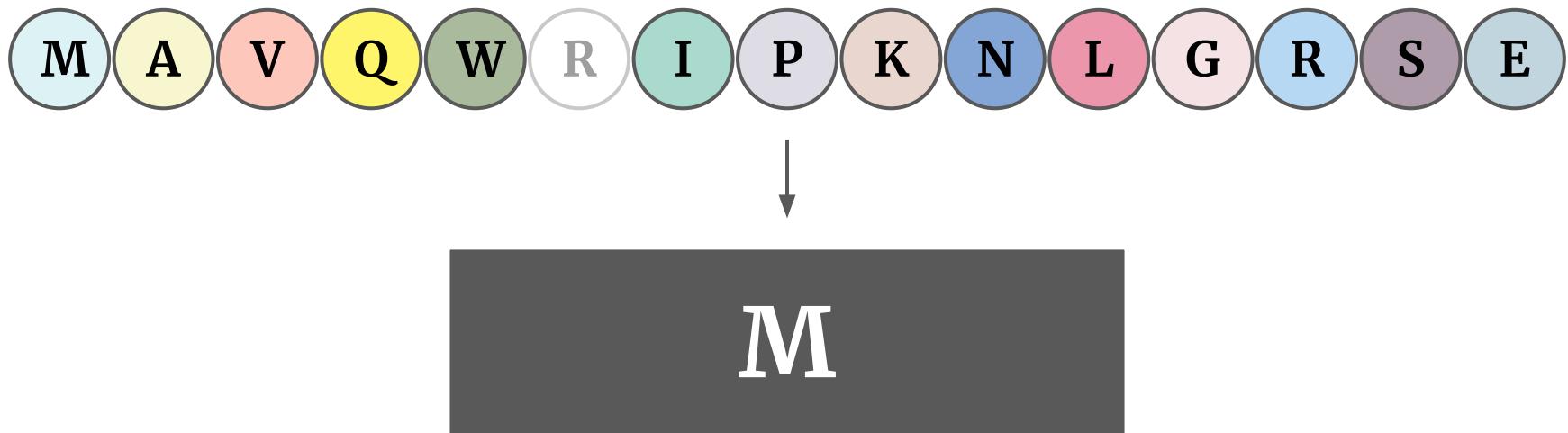
 sociocom/MedNER-CR-JA

Updated Apr 5 · 15.7M · 5

protein ~ language
amino acid ~ word



protein ~ language
amino acid ~ word



 Model card

Files and versions

 Community 3 Edit model card

ProtBert model

Pretrained model on protein sequences using a masked language modeling (MLM) objective. It was introduced in [this paper](#) and first released in [this repository](#). This model is trained on uppercase amino acids: it only works with capital letter amino acids.

Model description

ProtBert is based on Bert model which pretrained on a large corpus of protein sequences in a self-supervised fashion. This means it was pretrained on the raw protein sequences only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those protein sequences.

One important difference between our Bert model and the original Bert version is the way of dealing

Fill-Mask

PyTorch

Transformers

Uniref100

protein

protein language model

AutoTrain Compatible

Model card

[Files and versions](#)

ProtBert model

Pretrained model on protein sequences using a mask introduced in [this paper](#) and first released in [this repository](#). It only works with capital letter amino acids: it only works with capital letter amino acids.

Model description

ProtBert is based on Bert model which pretrained on a supervised fashion. This means it was pretrained on humans labelling them in any way (which is why it can't automatically process to generate inputs and labels from

Fill-Mask

PyTorch

TensorFlow

Transformers

esm

AutoTrain Compatible

Model card

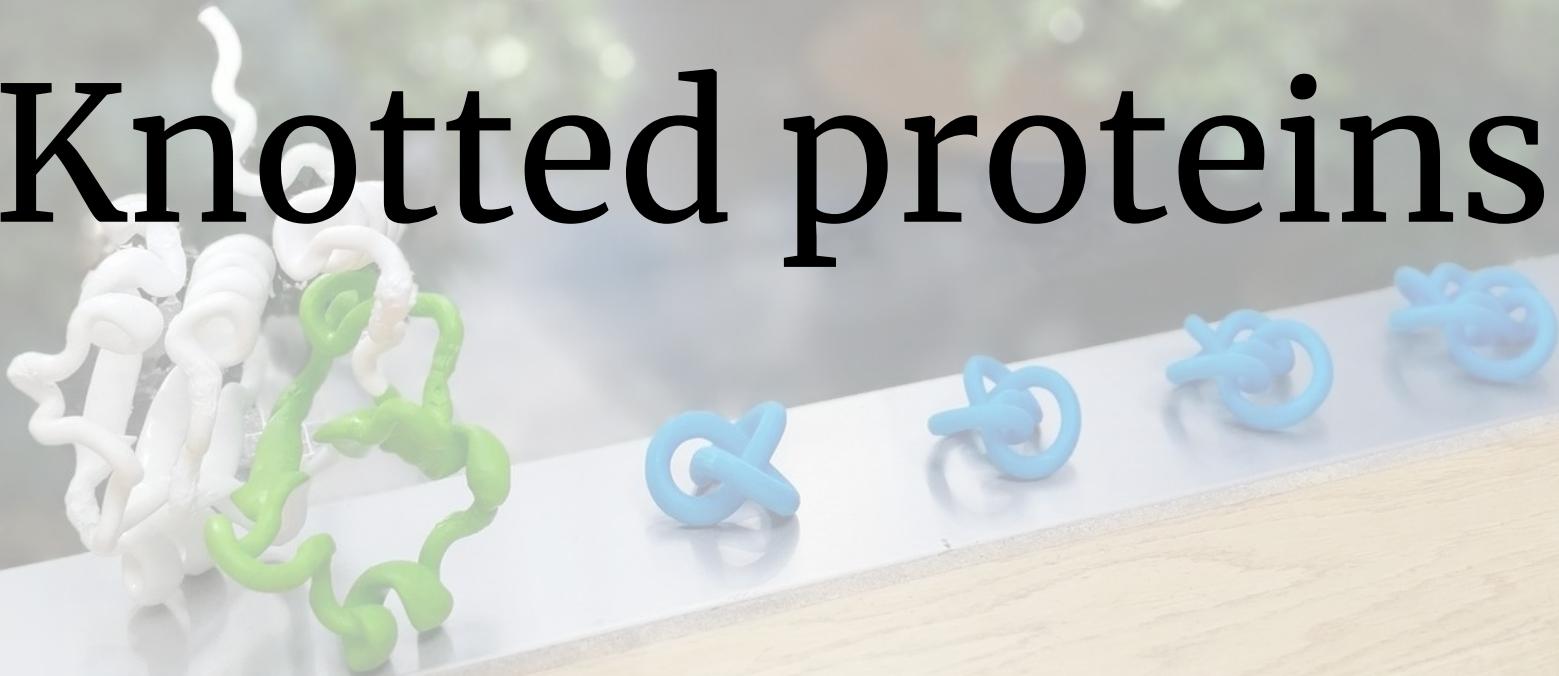
[Files and versions](#)[Community](#) 1

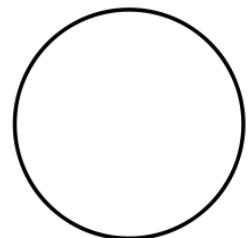
ESM-2

ESM-2 is a state-of-the-art protein model trained on a masked language modelling objective, suitable for fine-tuning on a wide range of tasks that take protein sequences as input. For more information on the model architecture and training data, please refer to the [accompanying paper](#). You may also be interested in some demo notebooks ([PyTorch](#), [TensorFlow](#)) which demonstrate how to fine-tune ESM-2 models on your tasks of interest.

Several ESM-2 checkpoints are available in the Hub with varying sizes. Larger sizes generally provide somewhat better accuracy, but require much more memory and time to train:

Knotted proteins

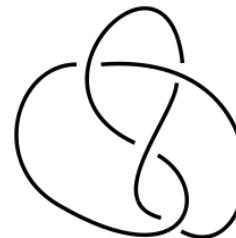




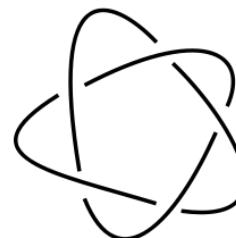
Unknot



3_1



4_1



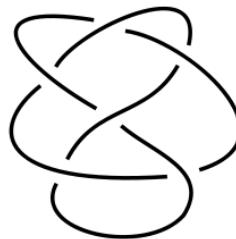
5_1



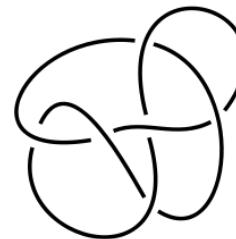
5_2



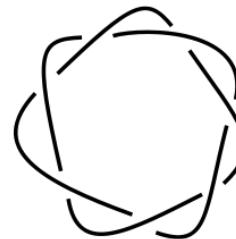
6_1



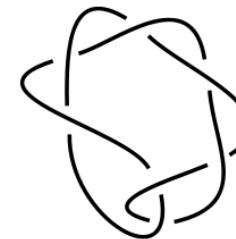
6_2



6_3



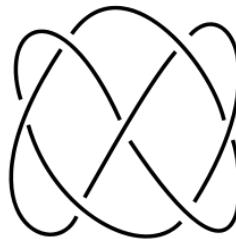
7_1



7_2



7_3



7_4



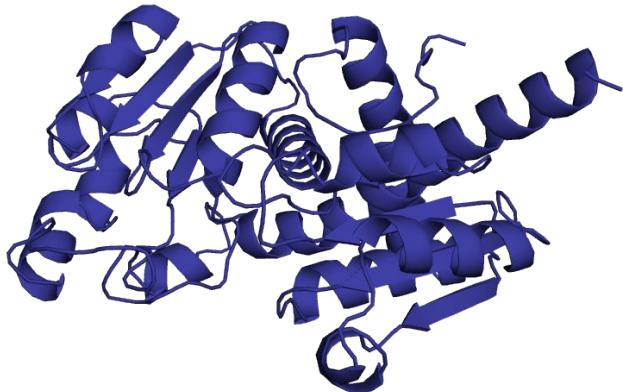
7_5



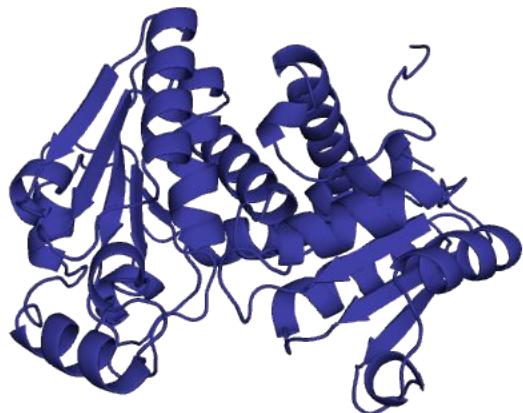
7_6



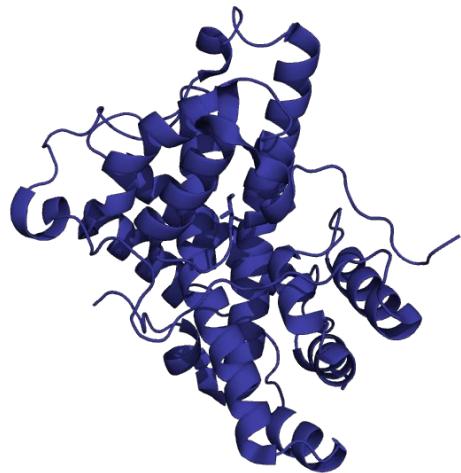
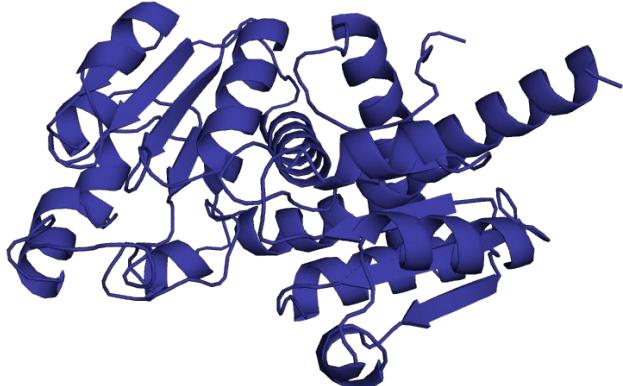
7_7



unknot
99 % proteins



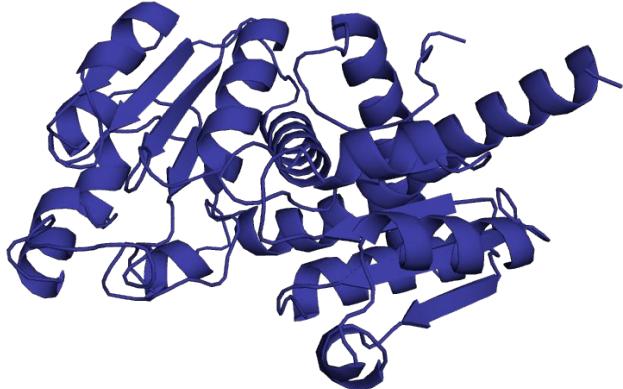
3_1 knot



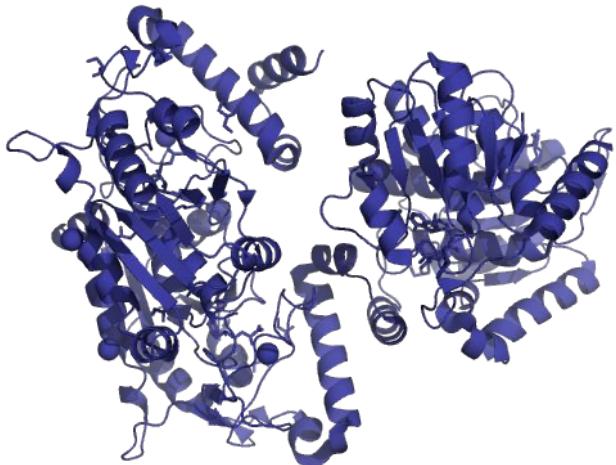
unknot
99 % proteins



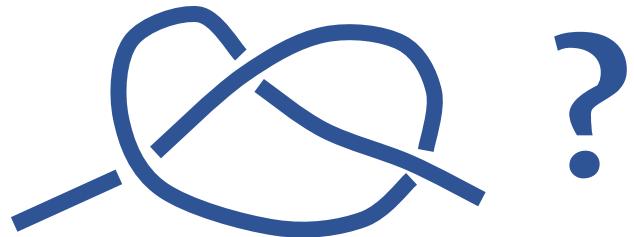
6_1 knot

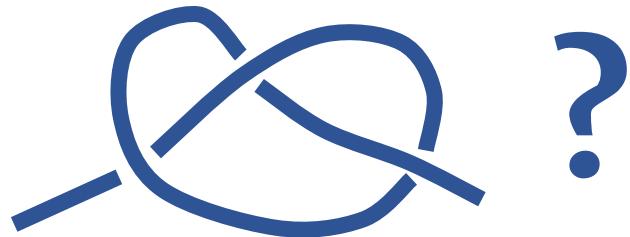


unknot
99 % proteins

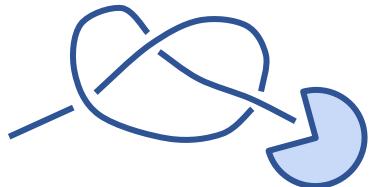


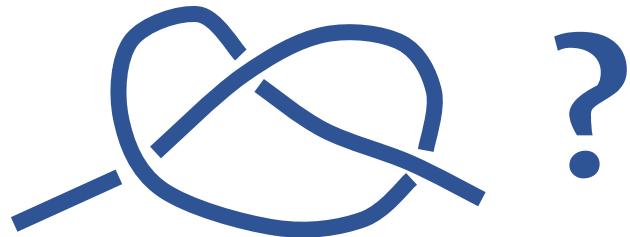
double
3_1 knot



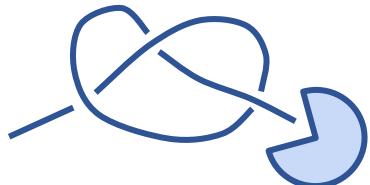


**prevent from
degradation**

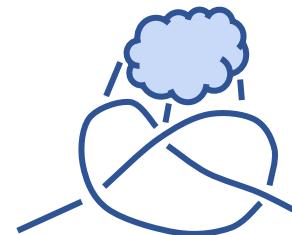


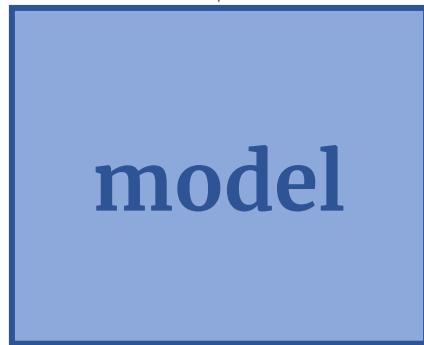


**prevent from
degradation**

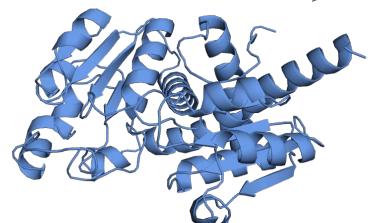


active site

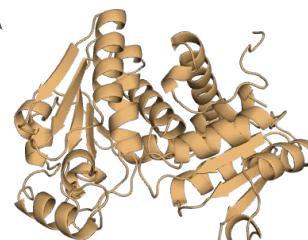




knotting
pattern



unknotted

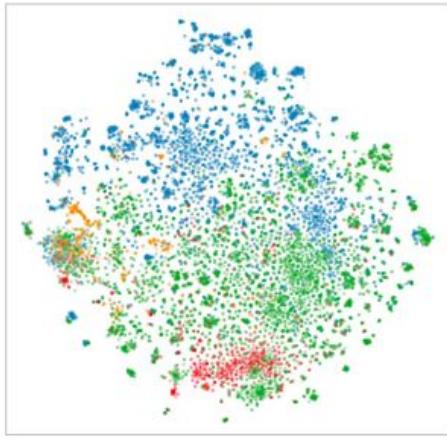


knotted

Dataset

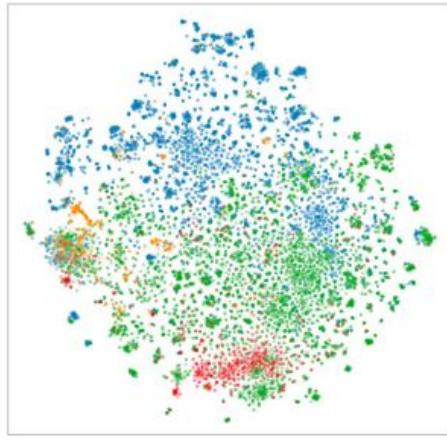
Protein family name	Knotted	Unknotted
ATCase/OTCase	2255	16998
AdoMet synthase	7456	1293
...
SPOUT	34233	2570
Sodium/calcium exchanger	22432	3524
TDD	3005	156
UCH	1785	620
ALL	99 303	103 313

t-SNE projection of embeddings



E Lineage: Kingdoms

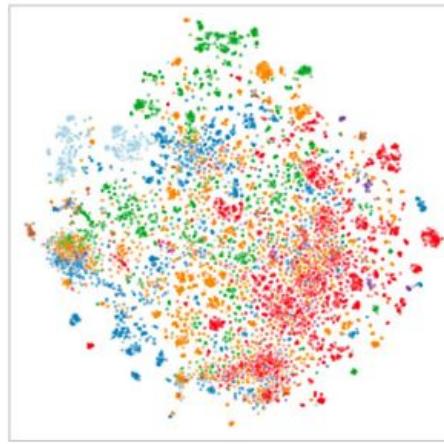
t-SNE projection of embeddings



● Eukaryota ● Archaea

● Bacteria ● Viruses

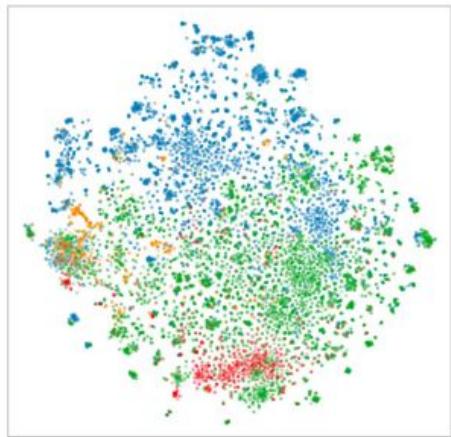
E Lineage: Kingdoms



● All alpha
● All beta
● Alpha & beta (a|b)
● Alpha & beta (a+b)
● Multi-domain
● Membrane, cell surface
● Small proteins

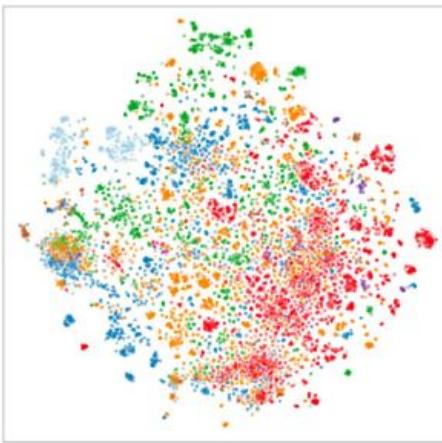
D Structure: SCOPe

t-SNE projection of embeddings

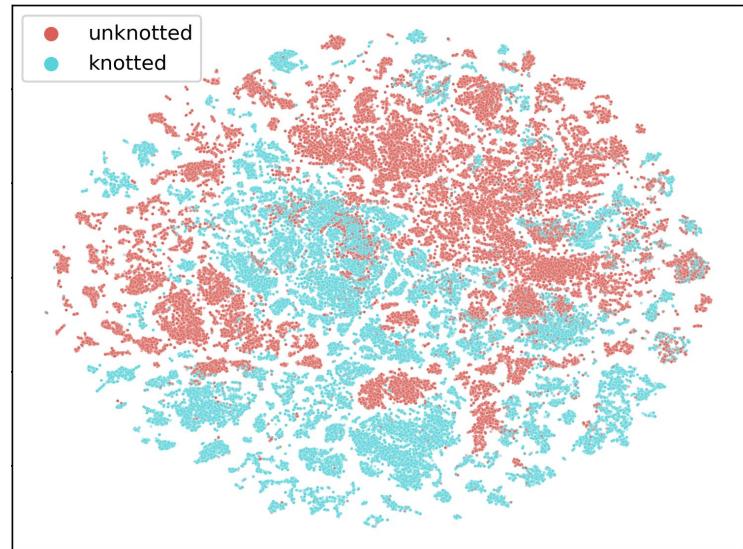


● Eukaryota ● Archaea
● Bacteria ● Viruses

E Lineage: Kingdoms

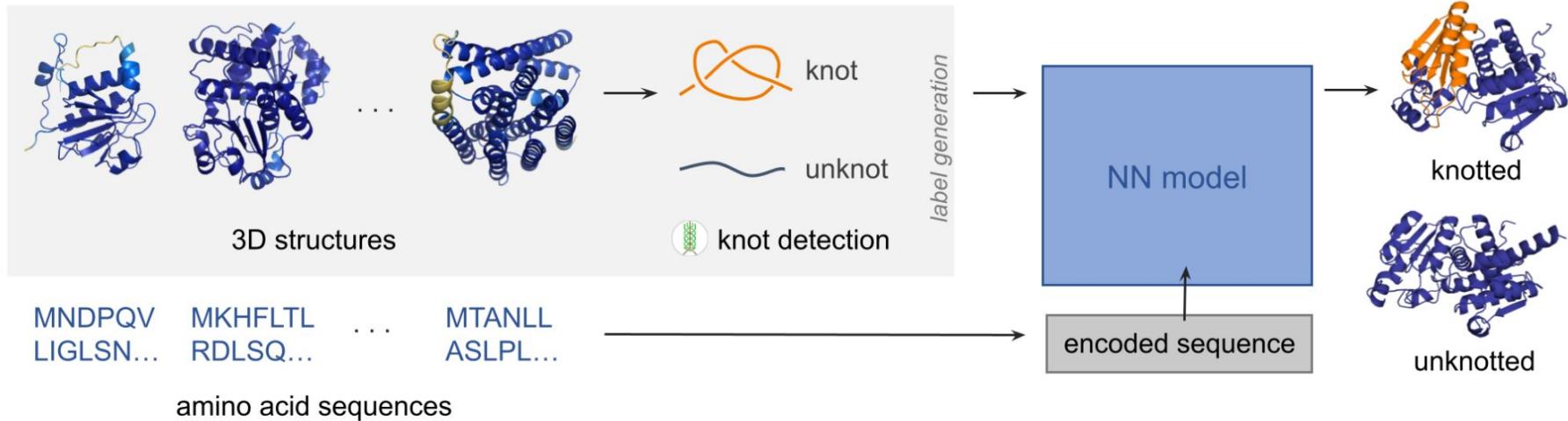


D Structure: SCOPe



● unknotted
● knotted

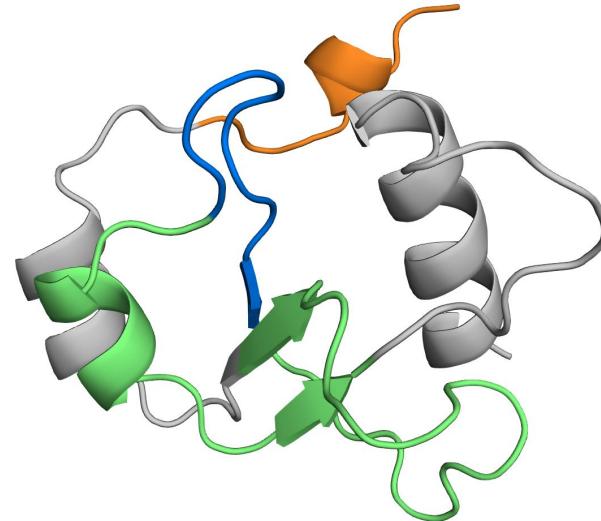
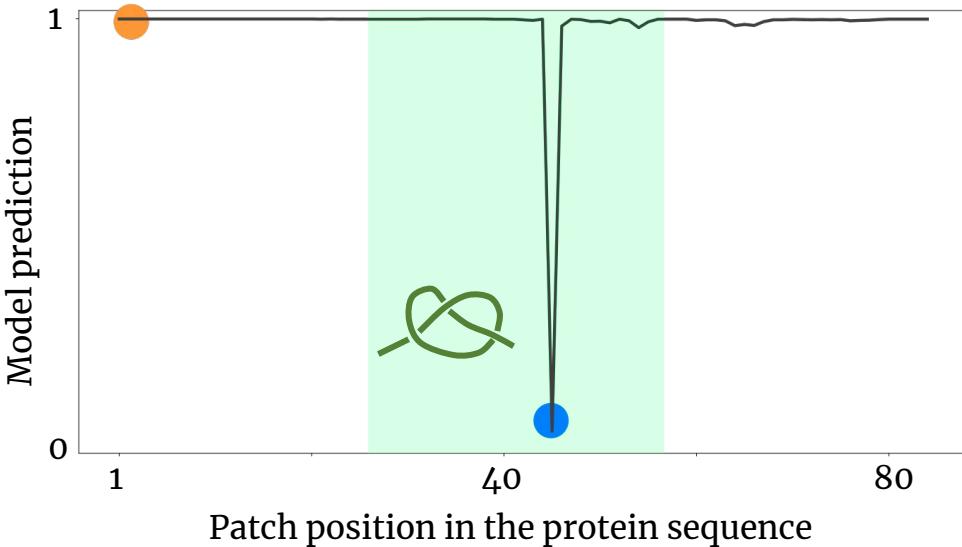
-700K proteins

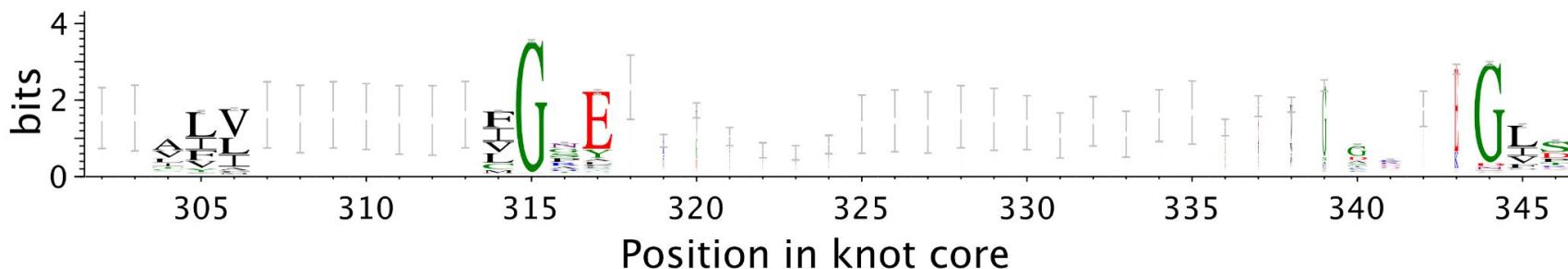
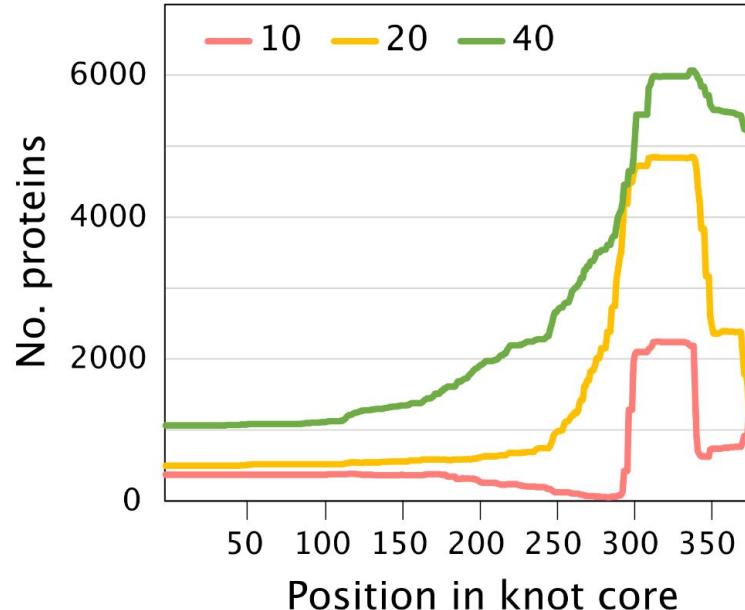


	Dataset size	Accuracy	TPR	TNR
Whole test dataset	39 412	<u>0.9845</u>	0.9865	0.9825
SPOUT	7 371	0.9887	0.9951	0.9090
TDD	612	0.9901	0.9965	0.8333
DUF	716	0.9748	0.9721	0.9766
AdoMet synthase	1 794	0.9899	0.9929	0.9708
Carbonic anhydrase	1531	0.9588	0.9737	0.9313
UCH	477	0.9056	0.9602	0.7520
ATCase/OTCase	3 799	0.9994	0.9977	0.9997
ribosomal-mitochondrial	147	0.8571	1.0000	0.4878
membrane	8 225	0.9811	0.9904	0.9390
VIT	14 262	0.9872	0.9420	0.9933
biosynthesis of lantibiotics	392	0.9642	0.9528	0.9685

Interpretation

XXXXXXXXXXYTDEPYLQSTKLPIYGALLDGENIYELVDKSKGILVIGNESKGIRSTIQNFIQKPIITIPRIGQAESLNAAVATGIIVGQLTL
XXXXXXXXXXXXTDLEPYLQSTKLPIYGALLDGENIYELVDKSKGILVIGNESKGIRSTIQNFIQKPIITIPRIGQAESLNAAVATGIIVGQLTL
...
MGGIFRVNTYYTDLEPYLQSTKLPIYGALLDGENIYELVDKSKGXXXXXXXIRSTIQNFIQKPIITIPRIGQAESLNAAVATGIIVGQLTL
...
MGGIFRVNTYYTDLEPYLQSTKLPIYGALLDGENIYELVDKSKGILVIGNESKGIRSTIQNFIQKPIITIPRIGQAESLNAAVAXXXXXXXX

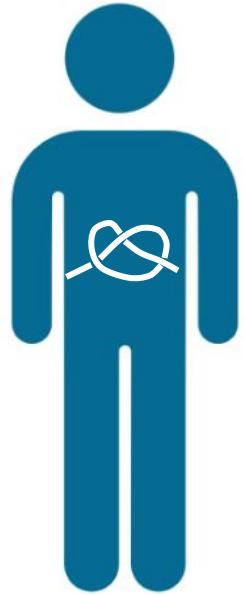




Future ideas

- interpretation for more protein families
- design a new knotted protein

How is it useful?



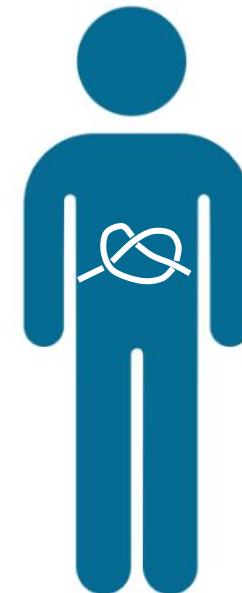
How is it useful?

The Unique Cysteine Knot Regulates the Pleiotropic Hormone Leptin

Ellinor Haglund¹, Joanna I. Sułkowska¹, Zhao He², Gen-Sheng Feng², Patricia A. Jennings^{1*},
José N. Onuchic^{3*}

Abstract

Leptin plays a key role in regulating energy intake/expenditure, metabolism and hypertension. It folds into a four-helix bundle that binds to the extracellular receptor to initiate signaling. Our work on leptin revealed a hidden complexity in the formation of a previously un-described, cysteine-knotted topology in leptin. We hypothesized that this unique topology could offer new mechanisms in regulating the protein activity. A combination of *in silico* simulation and *in vitro* experiments was used to probe the role of the knotted topology introduced by the disulphide-bridge on leptin folding and function. Our results surprisingly show that the free energy landscape is conserved between knotted and unknotted protein, however the additional complexity added by the knot formation is structurally important. Native state analyses led to the discovery that the disulphide-bond plays an important role in receptor binding and thus mediate biological activity by local motions on distal receptor-binding sites, far removed from the disulphide-bridge. Thus, the disulphide-bridge appears to function as a point of tension that allows dissipation of stress at a distance in leptin.



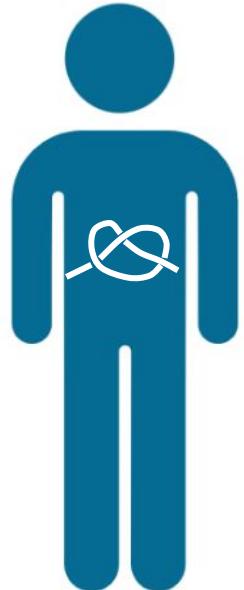
How is it useful?



Tied up in knots: Untangling substrate recognition by the SPOUT methyltransferases

Sarah E. Strassler [†] • Isobel E. Bowles [‡] • Debayan Dey

The SpoU-TrmD (SPOUT) methyltransferase superfamily was designated when structural similarity was identified between the transfer RNA-modifying enzymes TrmH (SpoU) and TrmD. SPOUT methyltransferases are found in all domains of life and predominantly modify transfer RNA or ribosomal RNA substrates, though one instance of an enzyme with a protein substrate has been reported. Modifications placed by SPOUT methyltransferases play diverse roles in regulating cellular processes such as ensuring translational fidelity, altering RNA stability, and conferring bacterial resistance to antibiotics. This large collection of S-adenosyl-L-methionine-dependent methyltransferases is defined by a unique α/β fold with a deep trefoil knot in their catalytic (SPOUT) domain. Herein, we describe current knowledge of SPOUT enzyme structure, domain architecture, and key elements of catalytic function, including S-adenosyl-L-methionine co-substrate binding, beginning with a new sequence alignment that divides the SPOUT methyltransferase superfamily into four major clades. Finally, a major focus of this review will be on our growing understanding of how these diverse enzymes accomplish the molecular feat of specific substrate recognition and modification, as highlighted by recent advances in our knowledge of protein–RNA complex structures and the discovery of the dependence of one SPOUT methyltransferase on metal ion binding for catalysis. Considering the broad biological roles of RNA modifications, developing a deeper understanding of the process of substrate recognition by the SPOUT enzymes will be critical for defining many facets of fundamental RNA biology with implications for human disease.

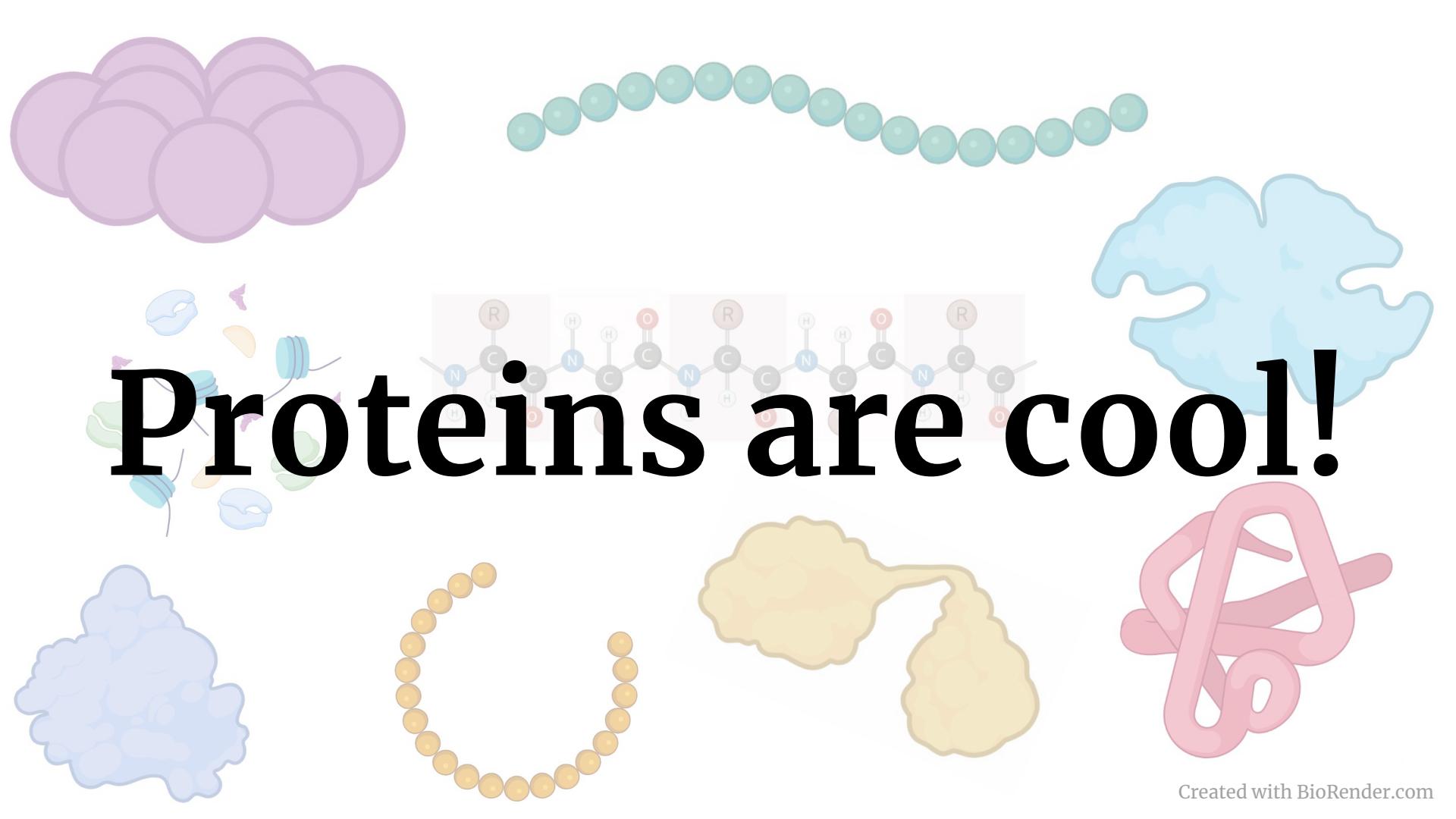


The Unique Cysteine Knot Regulates the Pleotropic Hormone Leptin

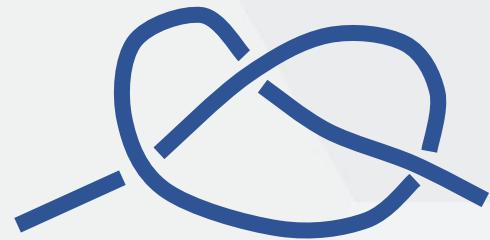
Ellinor Haglund¹, Joanna I. Sułkowska¹, Zhao He², Gen-Sheng Feng², Patricia A. Jennings^{1*},
José N. Onuchic^{3*}

Abstract

Leptin plays a key role in regulating energy intake/expenditure, metabolism and hypertension. It folds into a four-helix bundle that binds to the extracellular receptor to initiate signaling. Our work on leptin revealed a hidden complexity in the formation of a previously un-described, cysteine-knotted topology in leptin. We hypothesized that this unique topology could offer new mechanisms in regulating the protein activity. A combination of *in silico* simulation and *in vitro* experiments was used to probe the role of the knotted topology introduced by the disulphide-bridge on leptin folding and function. Our results surprisingly show that the free energy landscape is conserved between knotted and unknotted protein, however the additional complexity added by the knot formation is structurally important. Native state analyses led to the discovery that the disulphide-bond plays an important role in receptor binding and thus mediate biological activity by local motions on distal receptor-binding sites, far removed from the disulphide-bridge. Thus, the disulphide-bridge appears to function as a point of tension that allows dissipation of stress at a distance in leptin.



Proteins are cool!



Thank you



Petr Šimeček Joanna Sulkowska Denisa Šrámková Marta Korpacz Roksana Malinowska
Mai Lan Nguyen Agata Perlinska Paweł Rubach Maciej Sikora Dawid Uchal



Thank you

GitHub with research on knotted proteins:

https://github.com/ML-Bioinfo-CEITEC/pknots_experiments

slides:

<https://github.com/evaklimentova/presentations>

contact: eva.klimentova@ceitec.muni.cz

Questions?



GitHub with research on knotted proteins:

https://github.com/ML-Bioinfo-CEITEC/pknots_experiments

slides:

<https://github.com/evaklimentova/presentations>

contact: eva.klimentova@ceitec.muni.cz