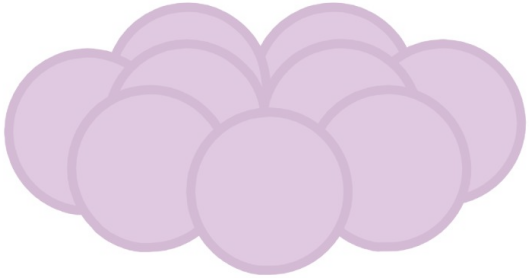
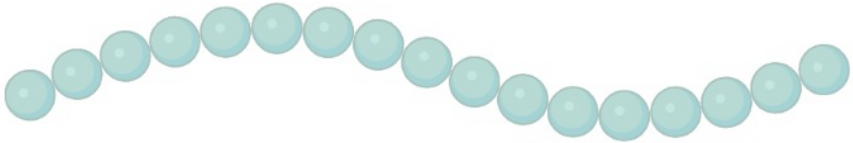
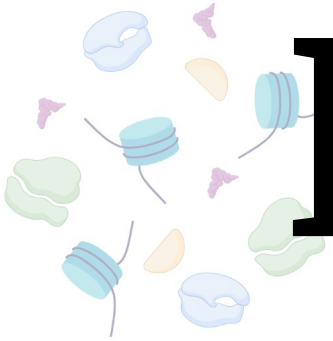
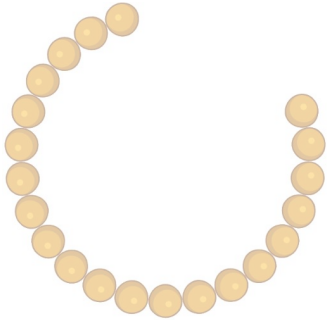
Using NLP to detect knots in protein structures

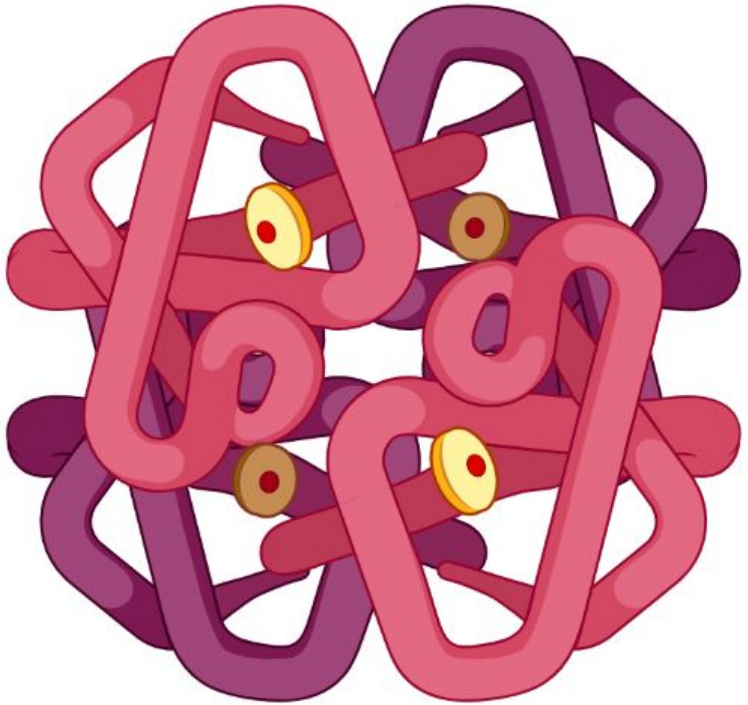
Eva Klimentová

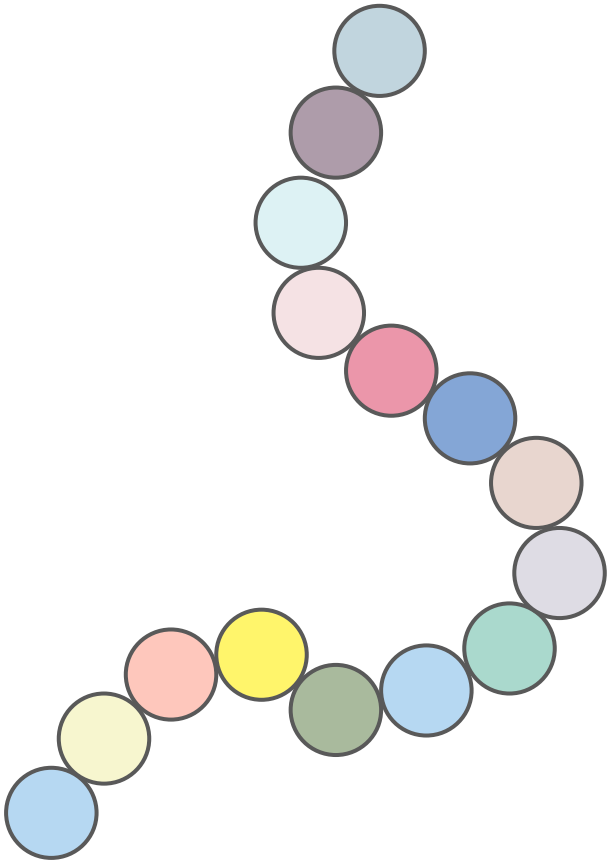


Proteins





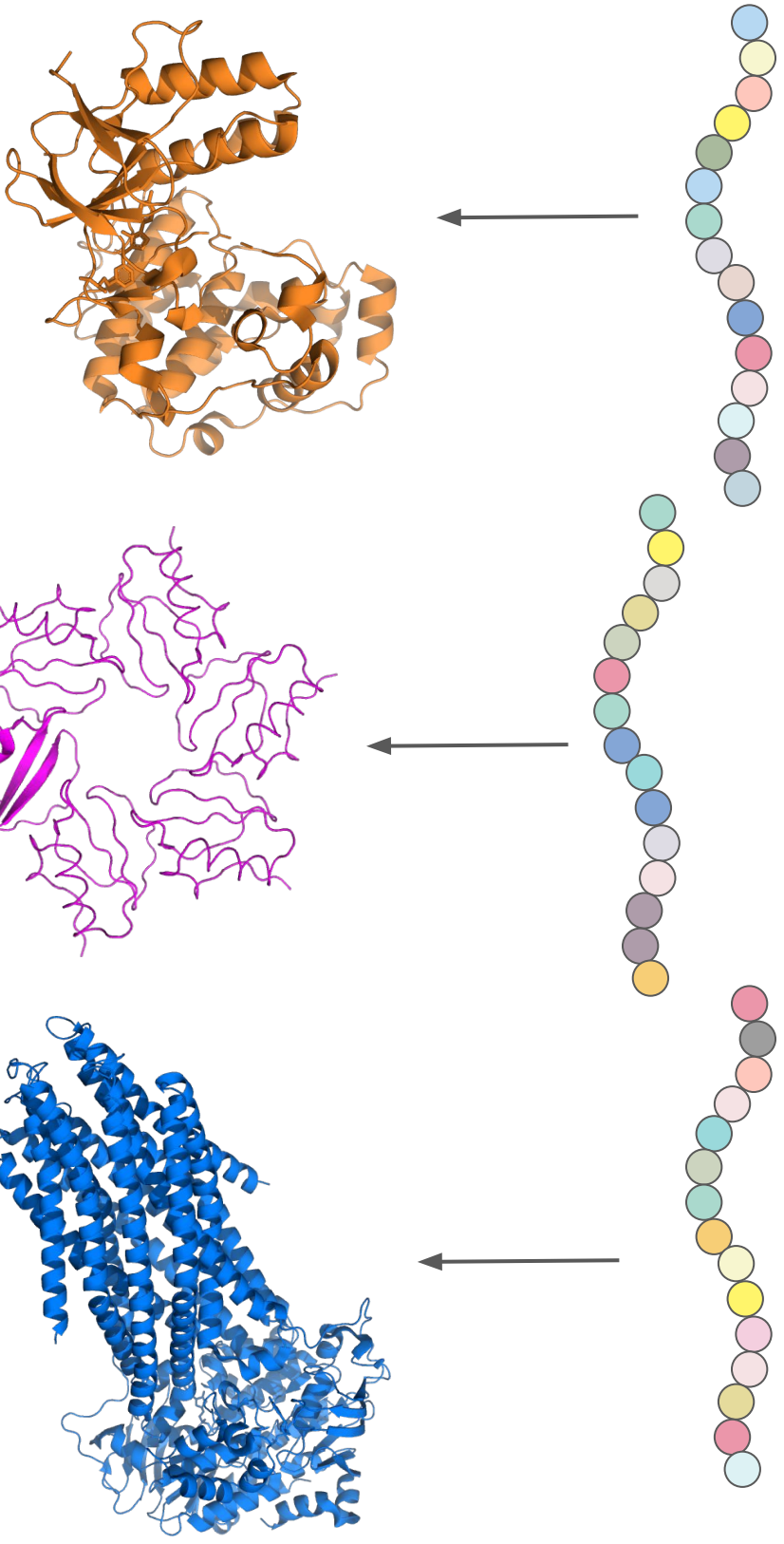
hemoglobin

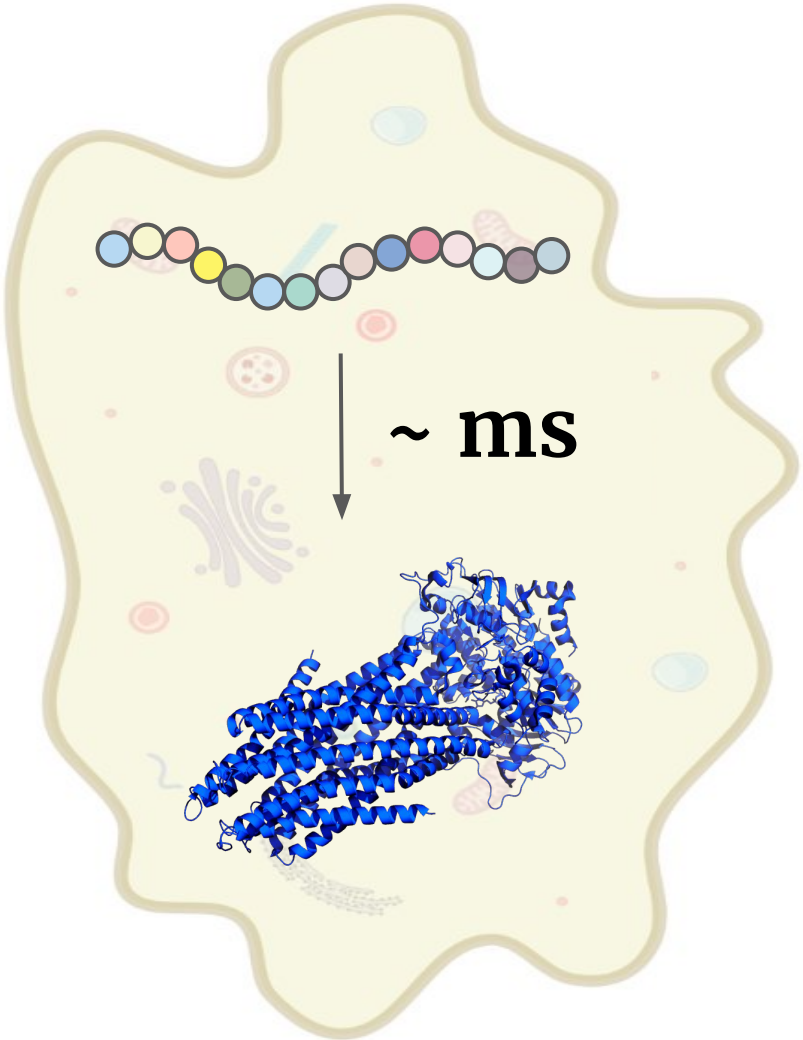


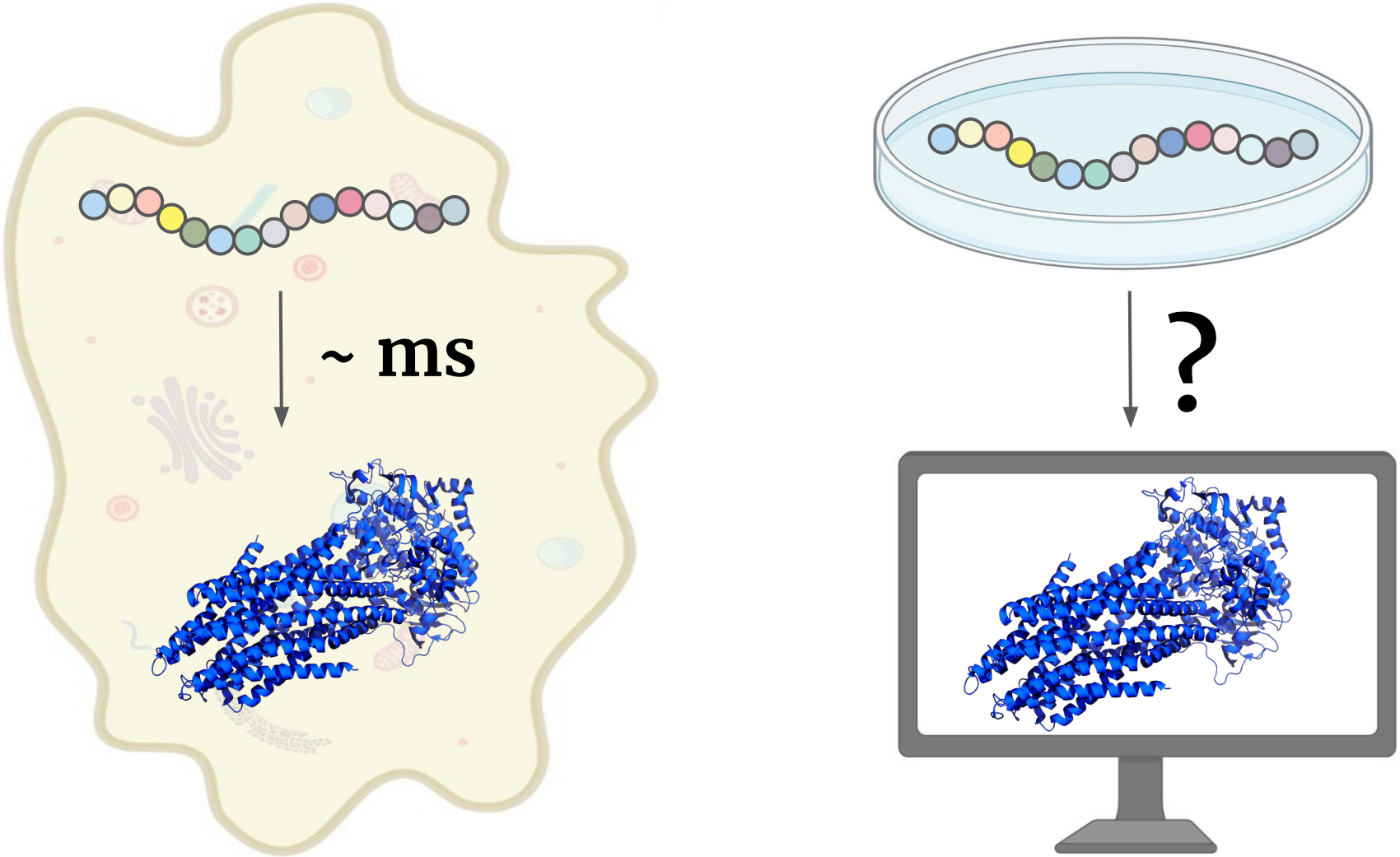
Amino acids = building blocks

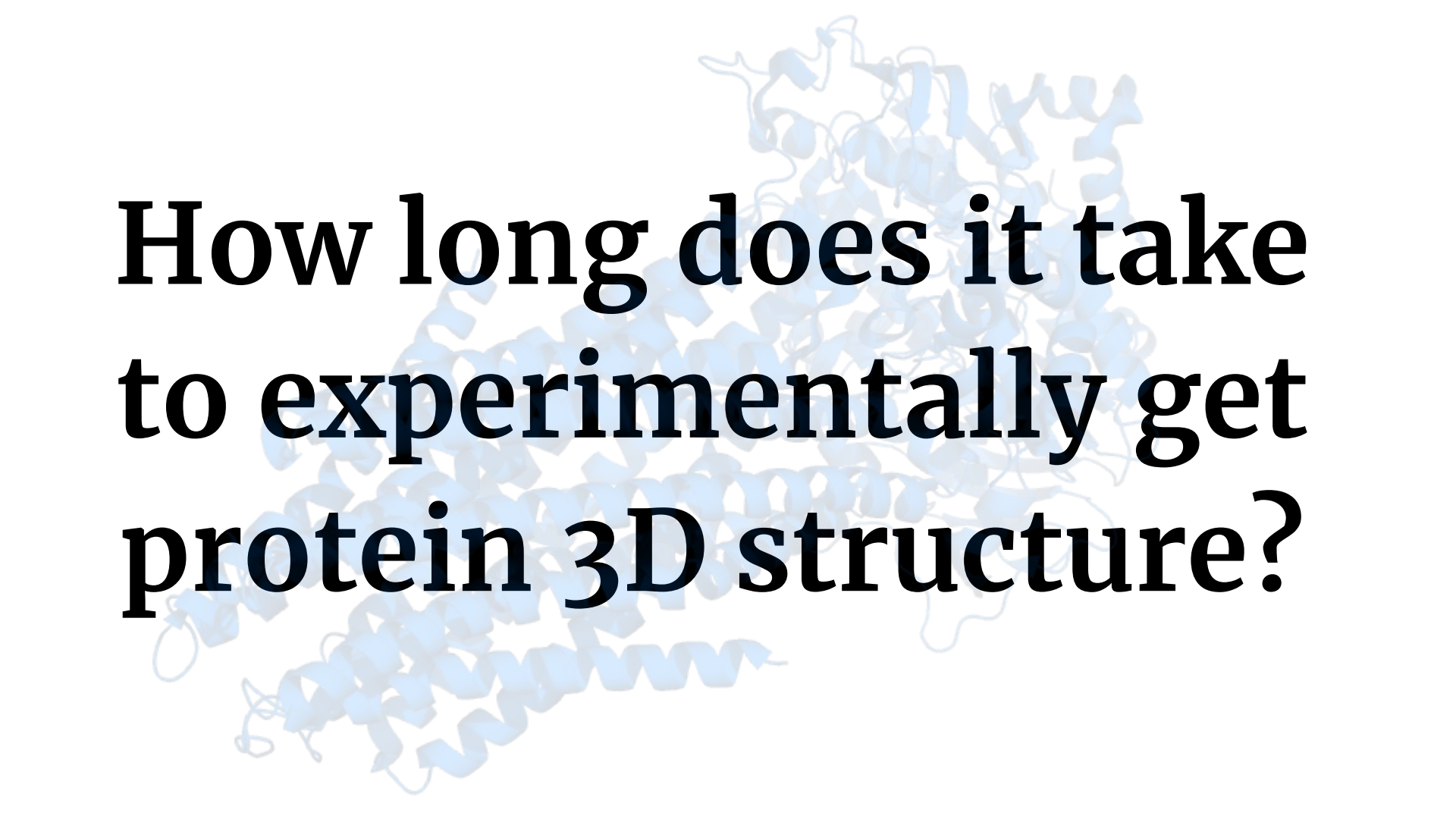
alanine - ala - A	leucine - leu - L
arginine - arg - R	lysine - lys - K
asparagine - asn - N	methionine - met - M
aspartic acid - asp - D	phenylalanine - phe - F
cysteine - cys - C	proline - pro - P
glutamine - gln - Q	serine - ser - S
glutamic acid - glu - E	threonine - thr - T
glycine - gly - G	tryptophan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

structure sequence

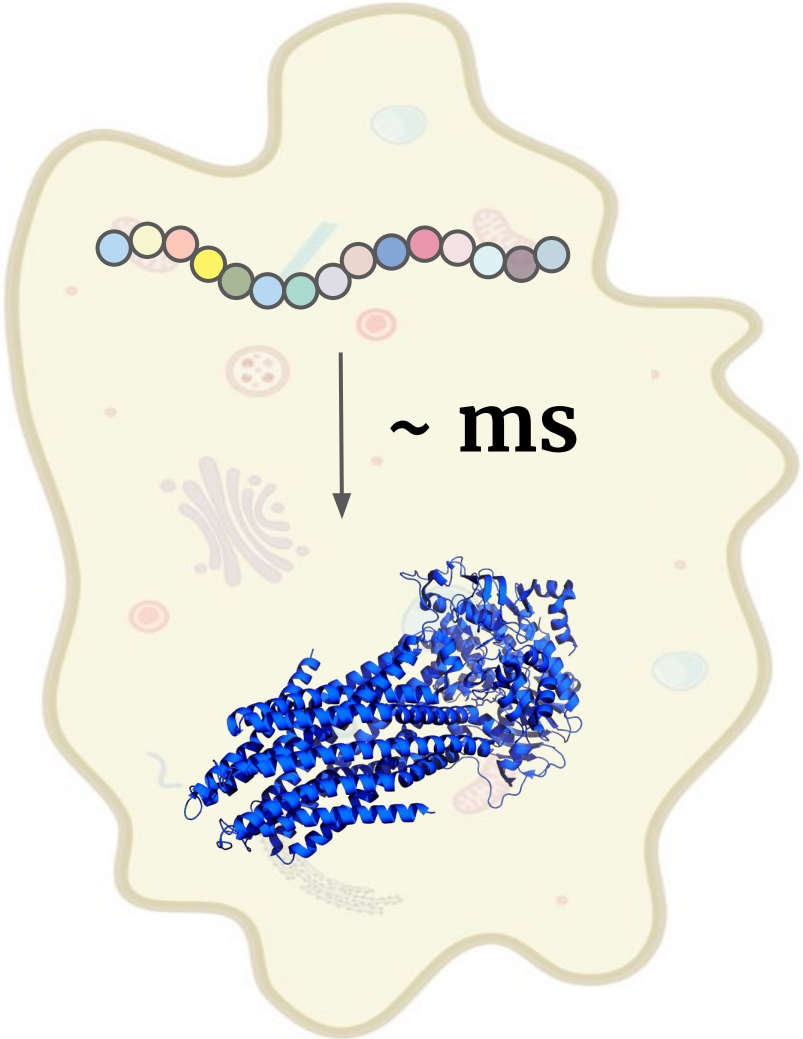


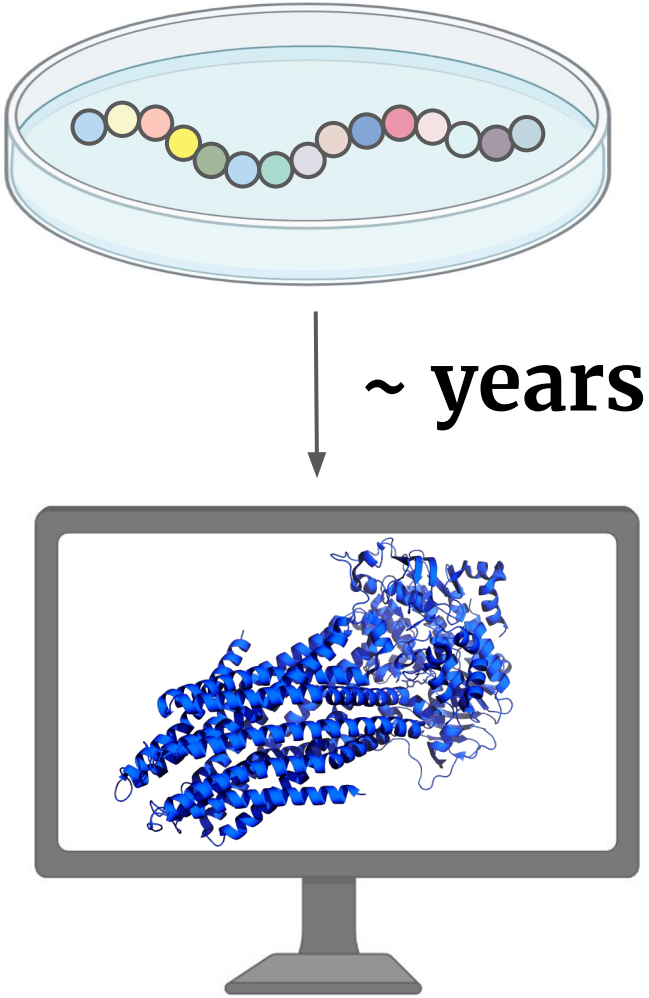

~ ms

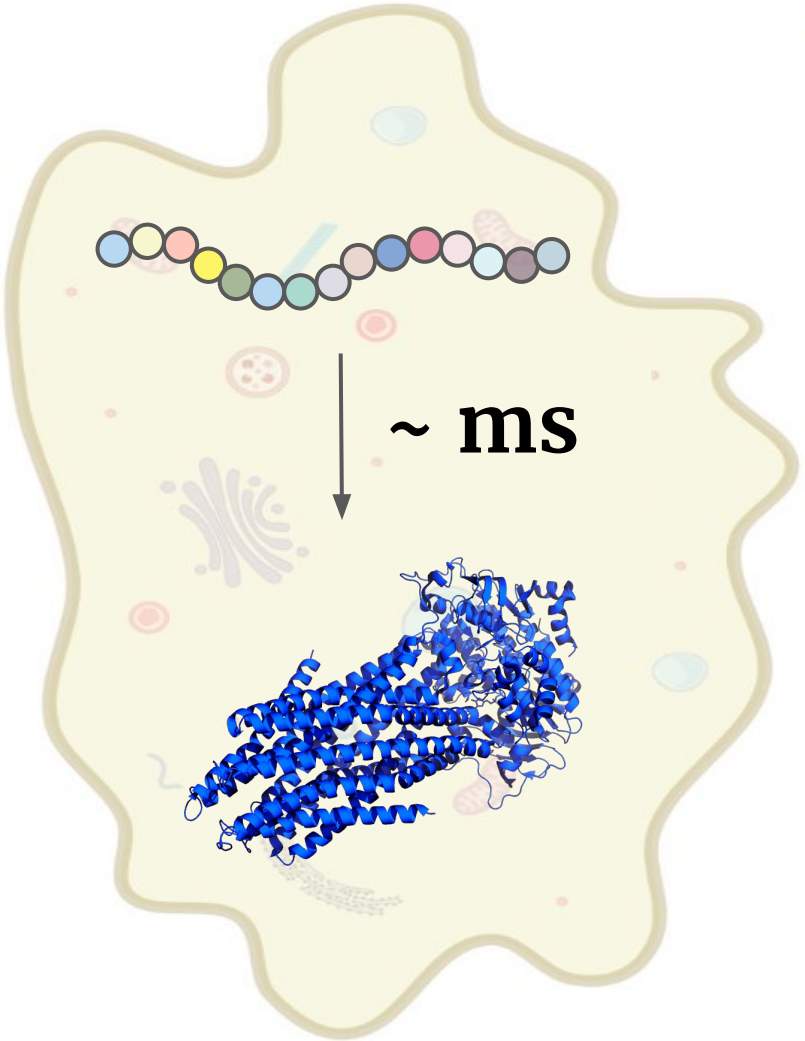

~ ms
?



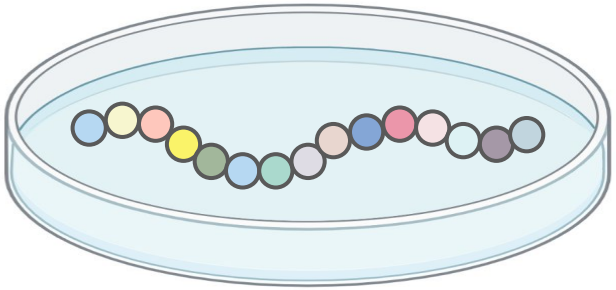
How long does it take to experimentally get protein 3D structure?


~ ms


~ years

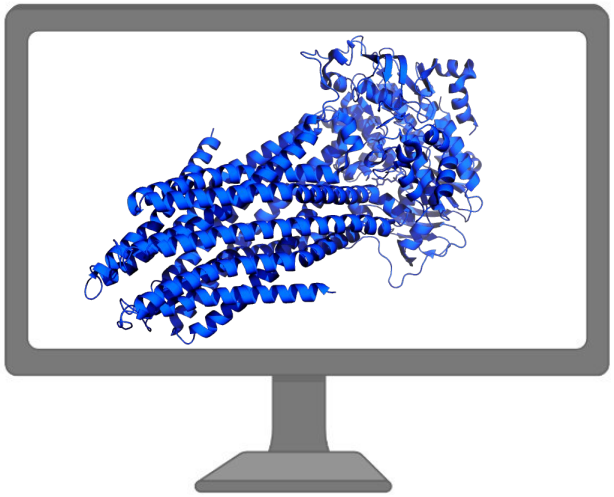


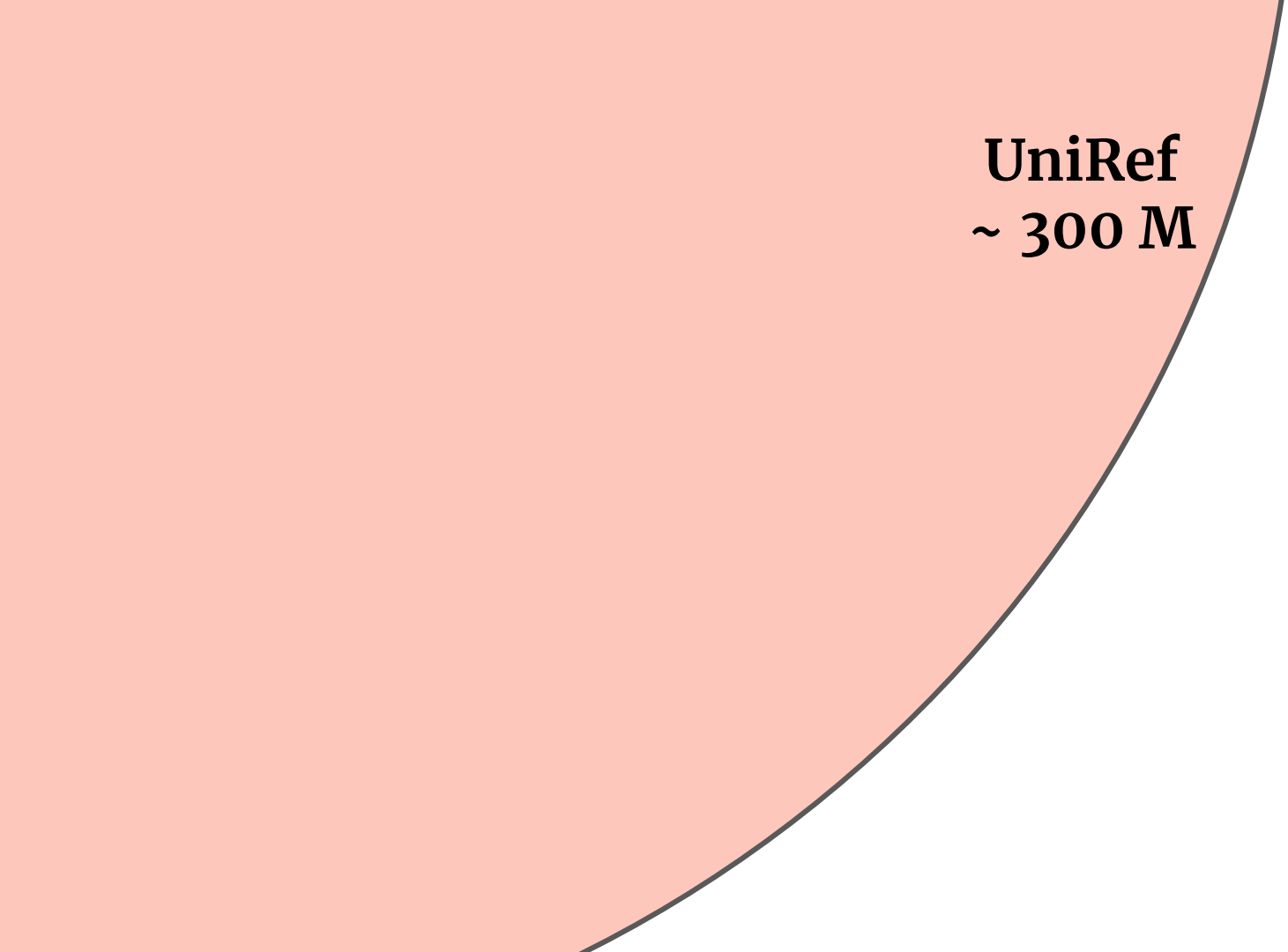
~ ms

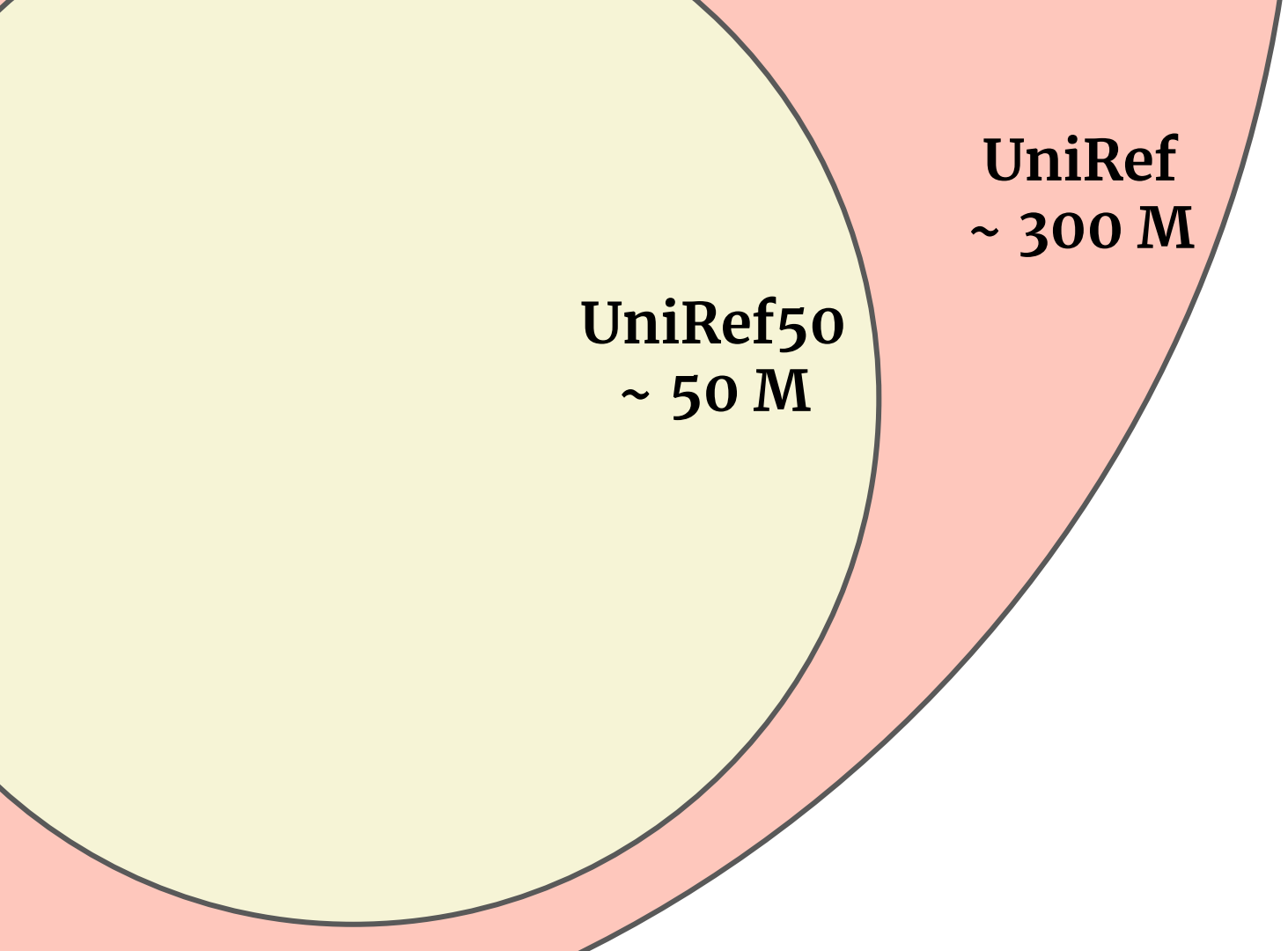


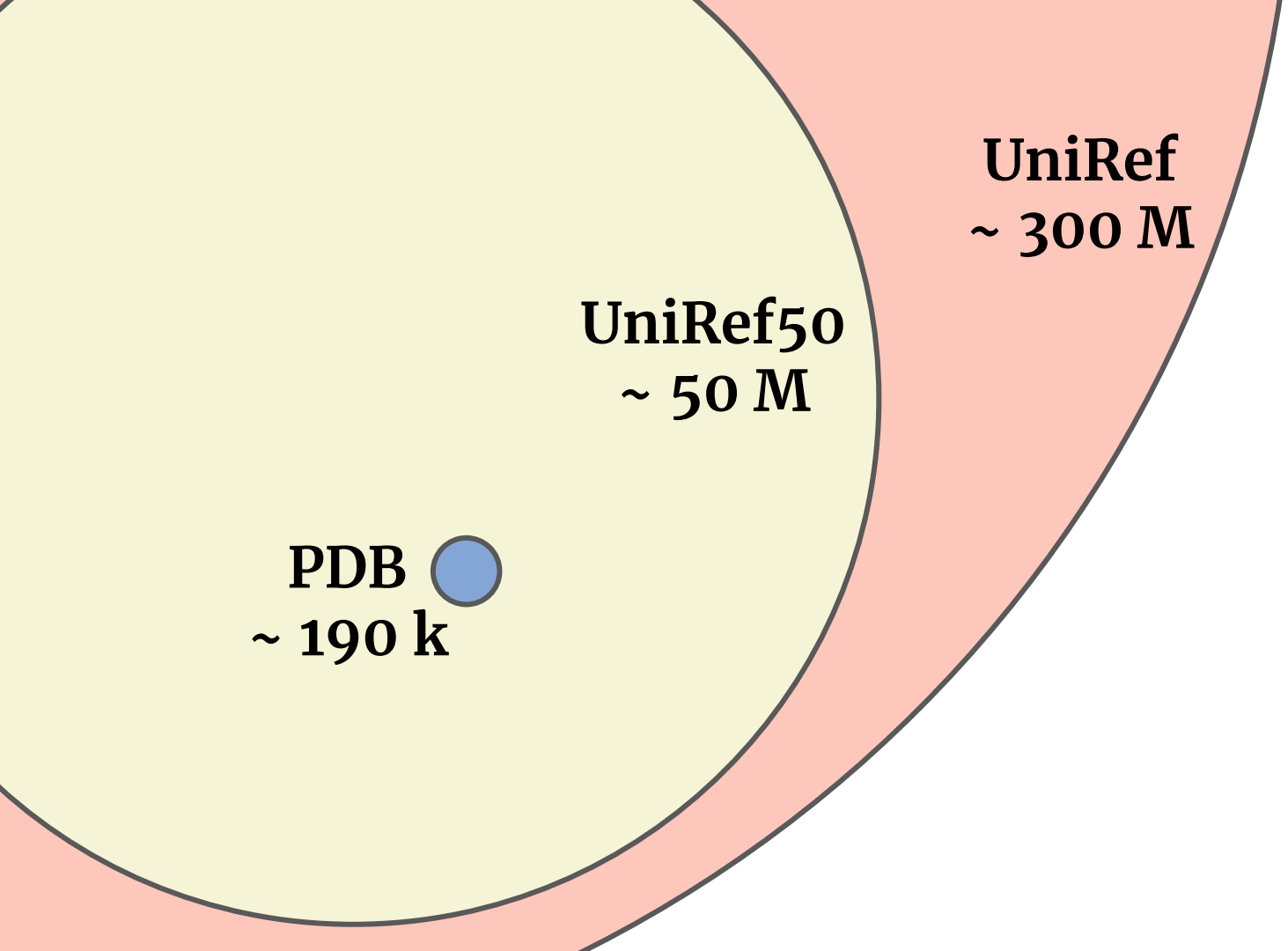
PhD life

~ years

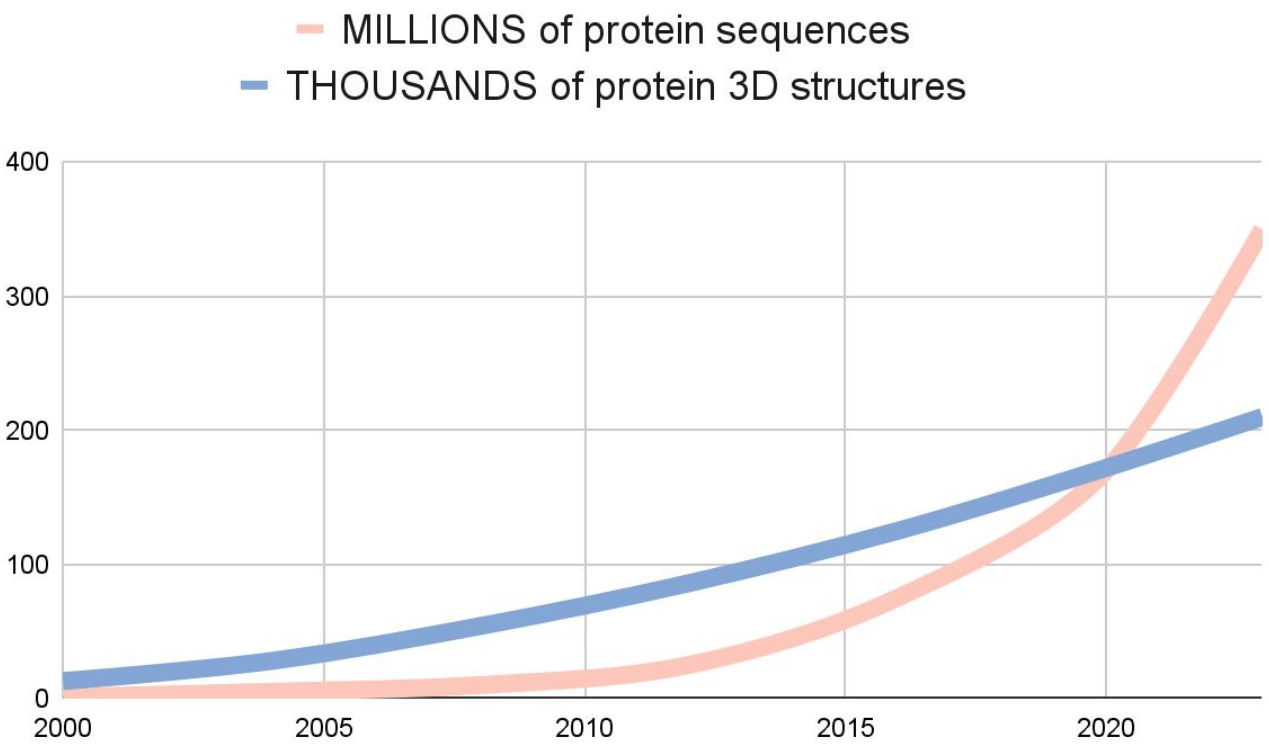



UniRef
~ 300 M

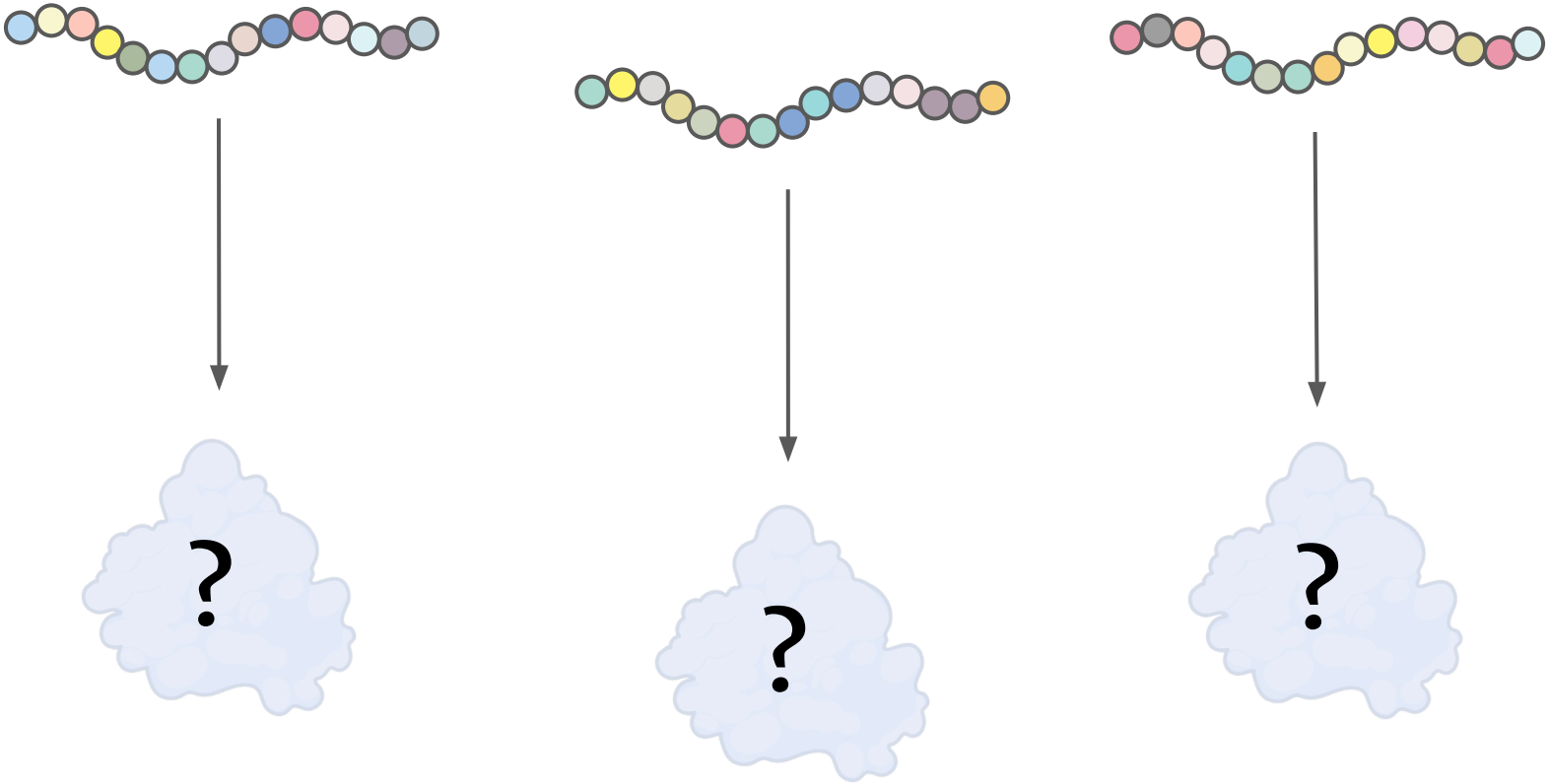

UniRef
~ 300 M
UniRef50
~ 50 M

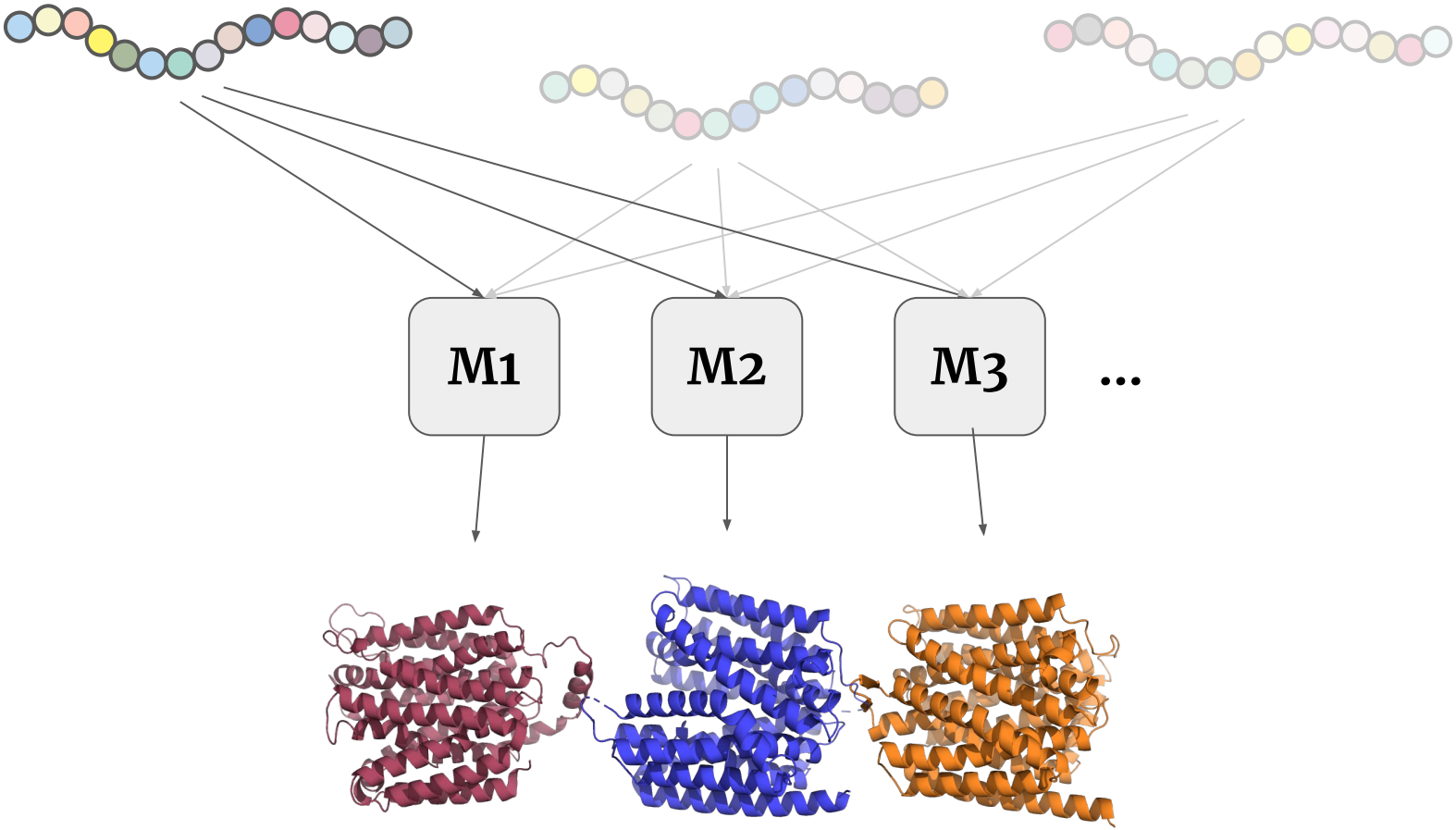

UniRef
~ 300 M
UniRef50
~ 50 M
PDB
~ 190 k

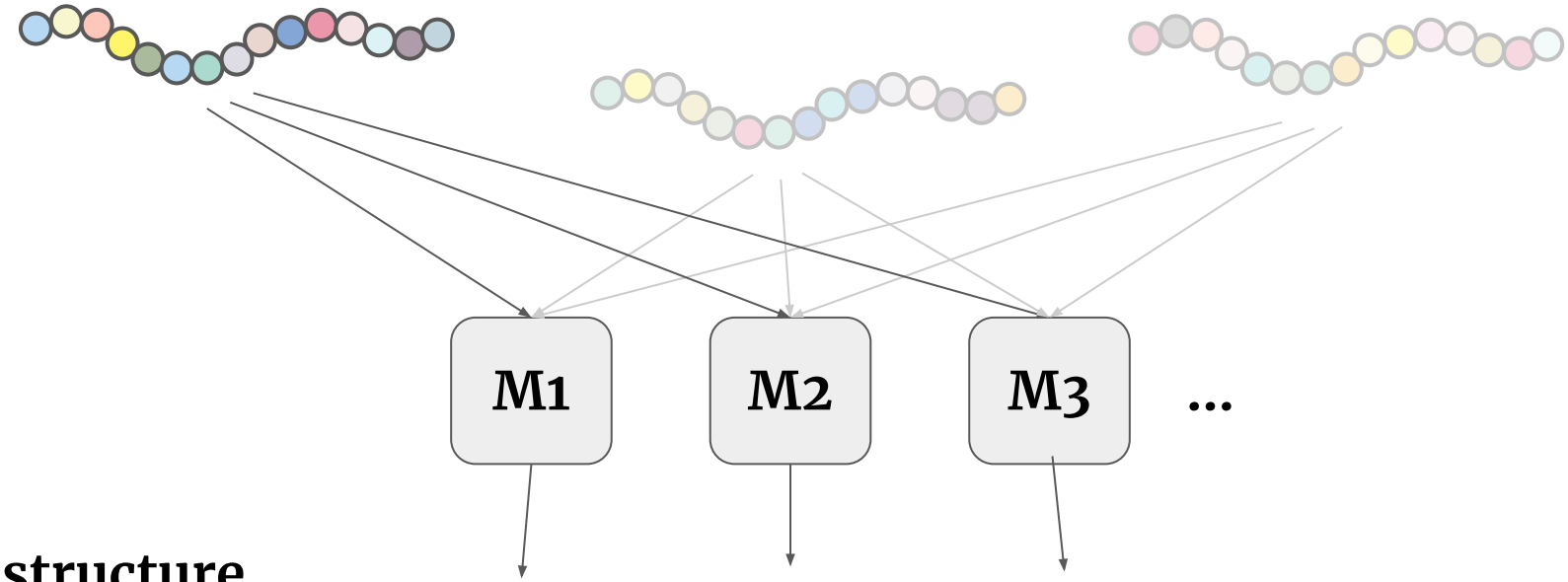
Growth of Protein Databases



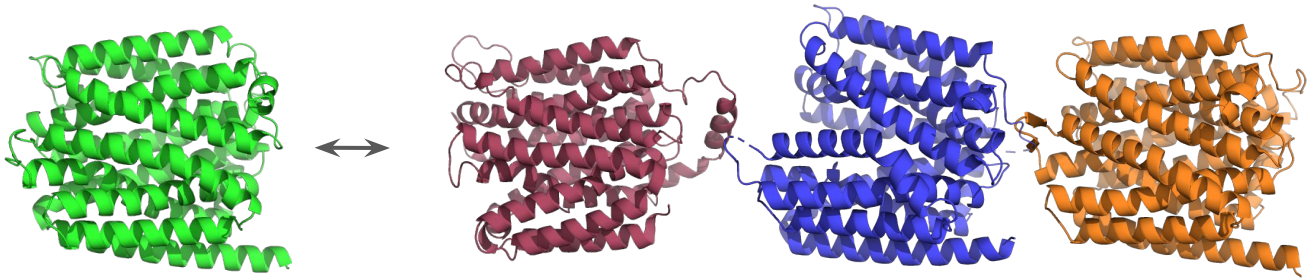
CASP competition



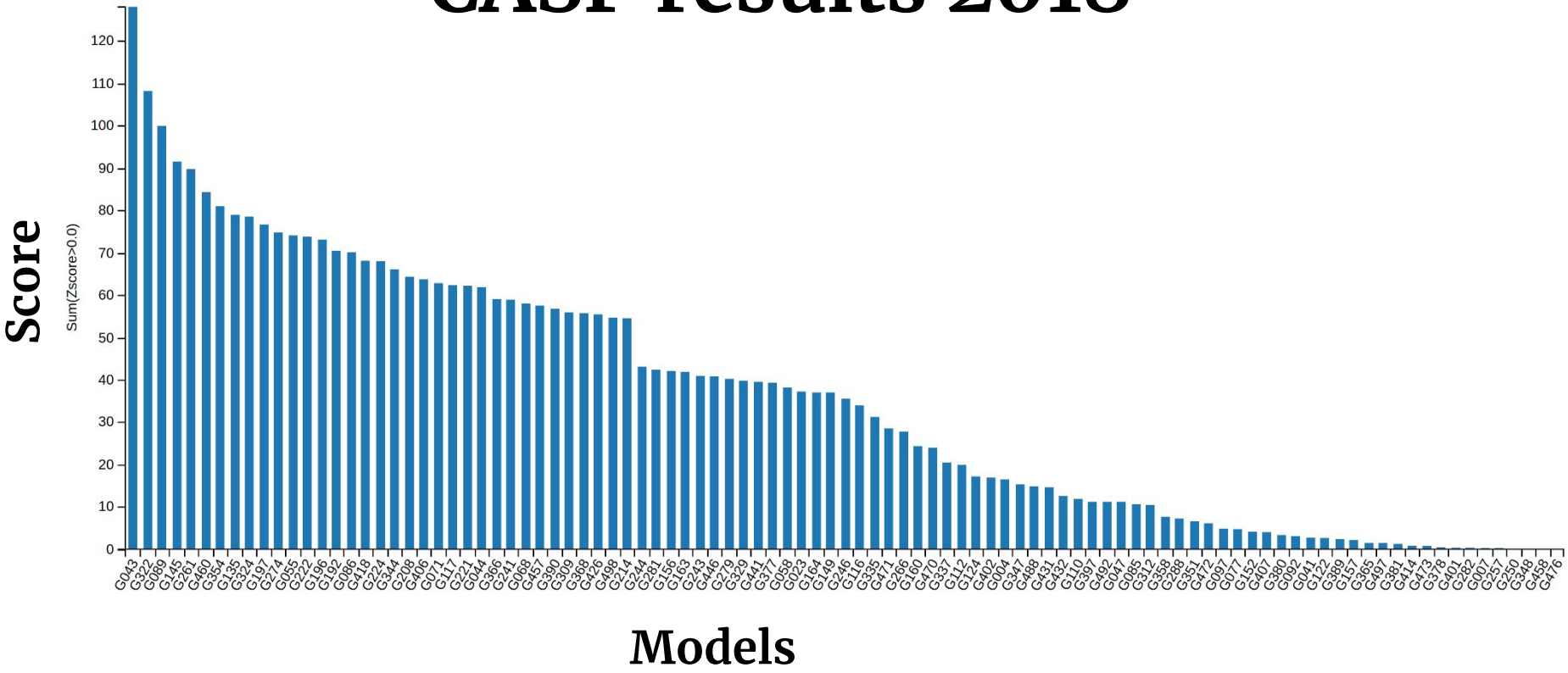

M1
M2
M3
...



real structure

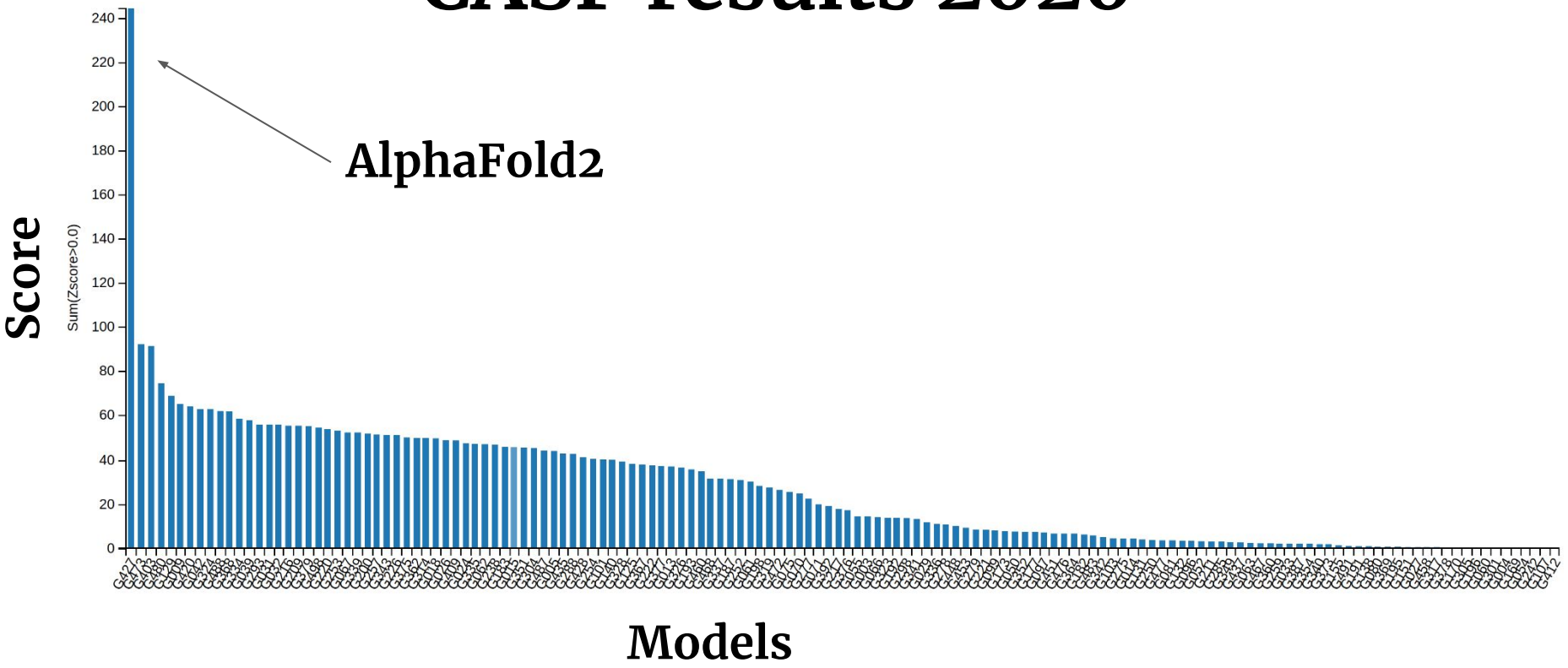


CASP results 2018



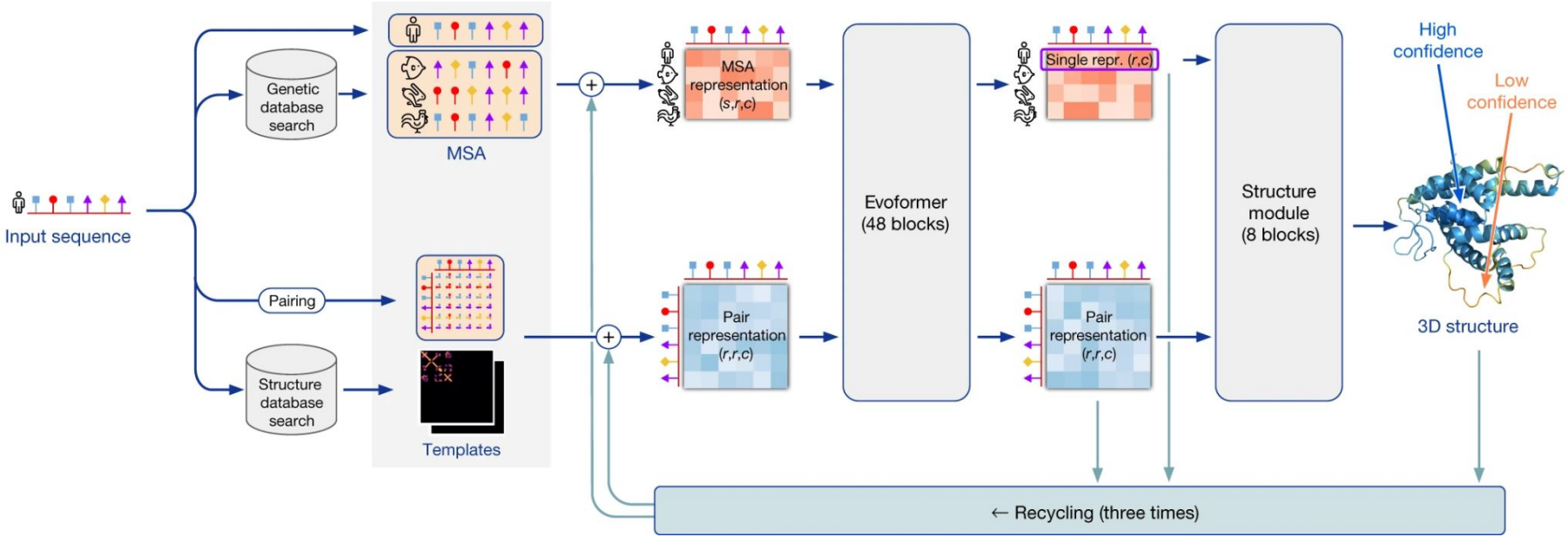
https://predictioncenter.org/index.cgi

CASP results 2020



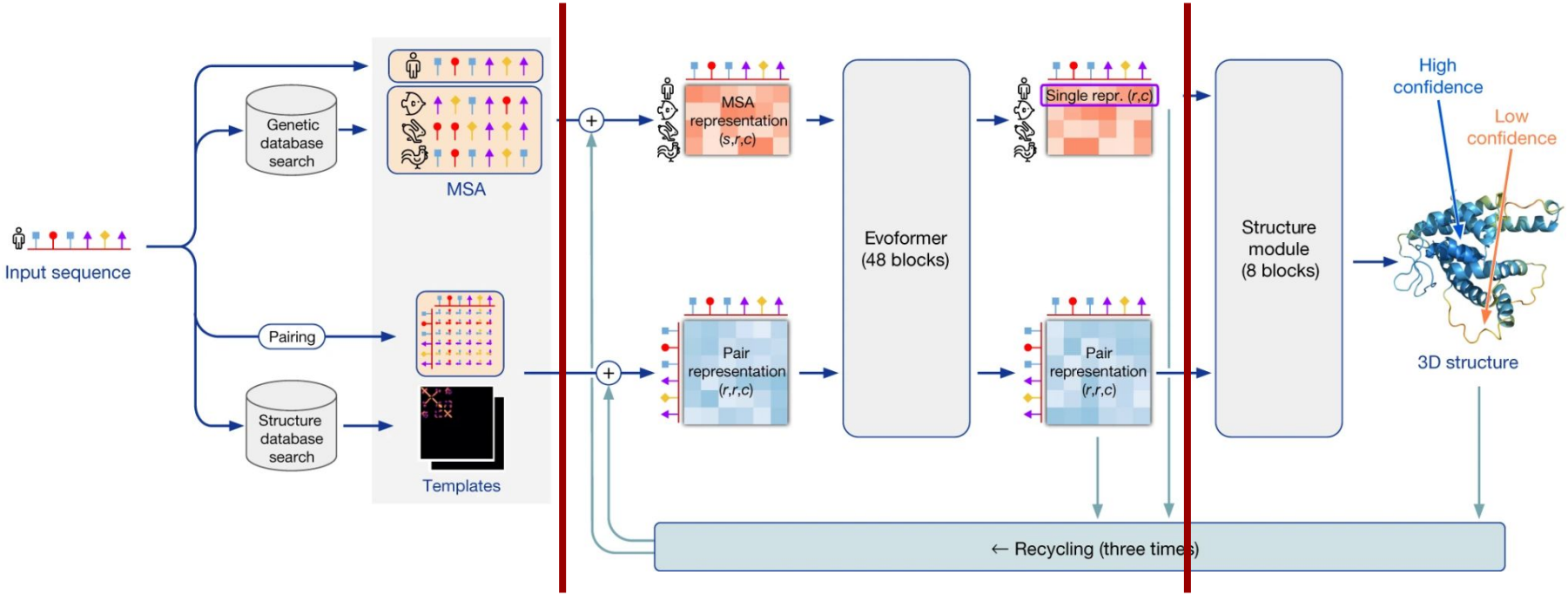
https://predictioncenter.org/index.cgi

AlphaFold2



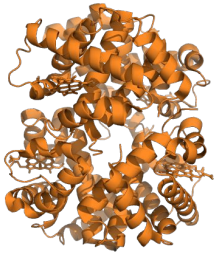
Highly accurate protein structure prediction with AlphaFold

AlphaFold2



1. encoding

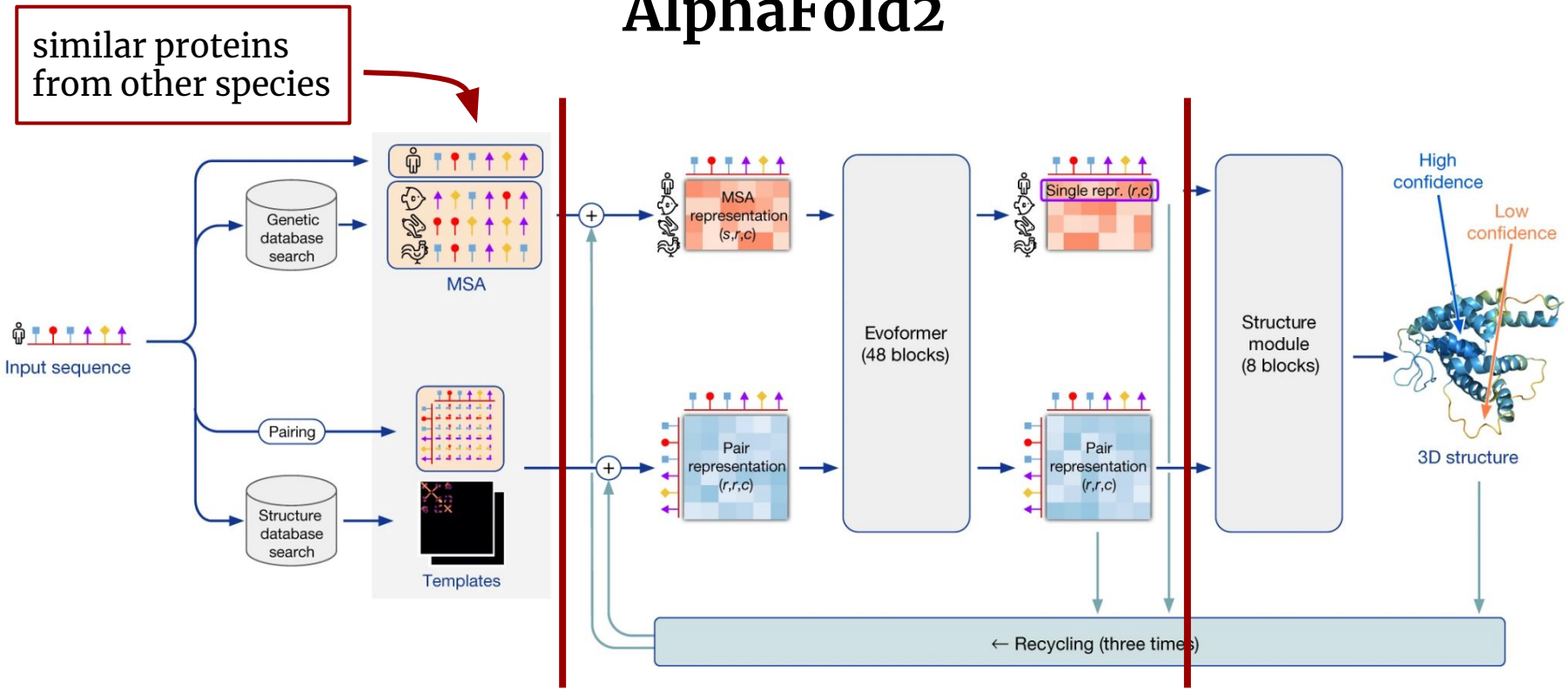
VLSPADKTNVKAAWGKVGAH...



VLSADDKTNIKNCWGKIGGH...

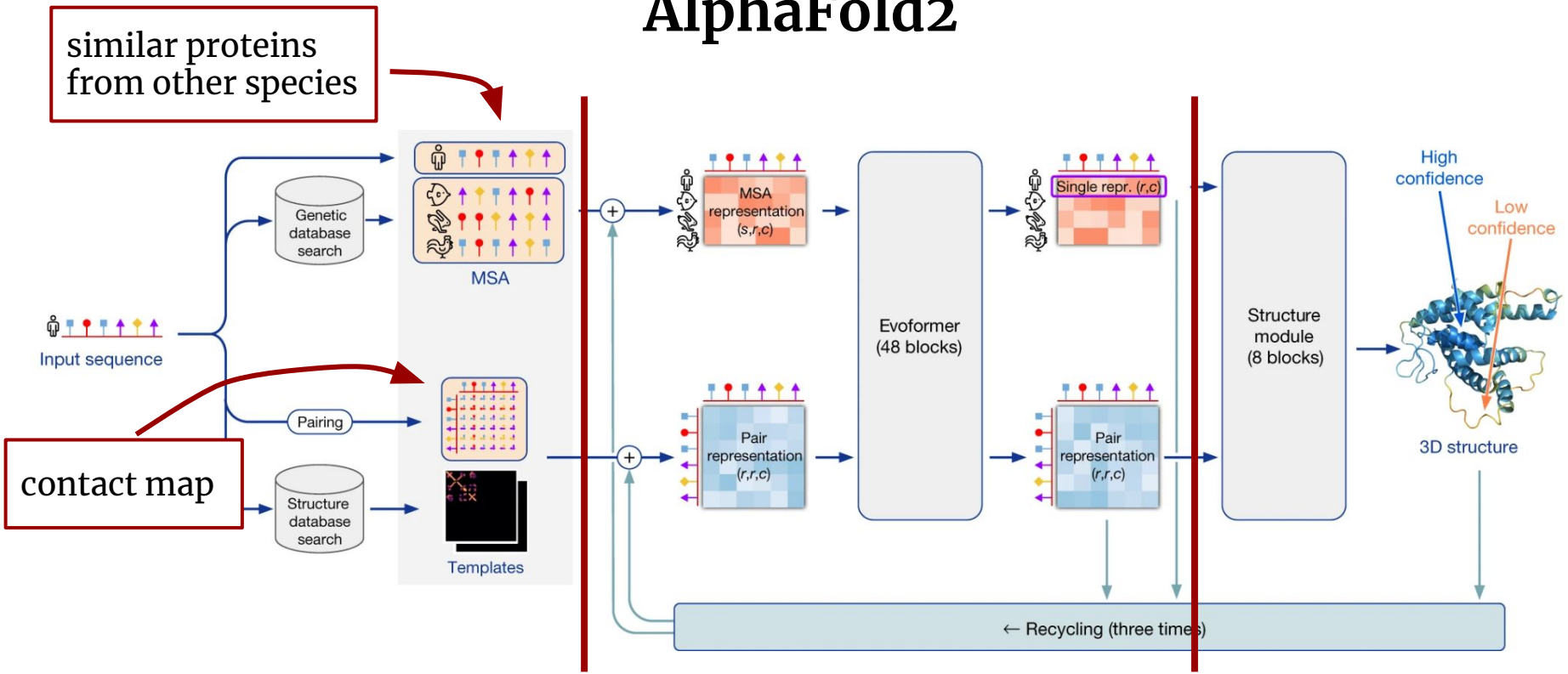
XSLSAKDKANVKAIWGKILPK...

AlphaFold2



1. encoding

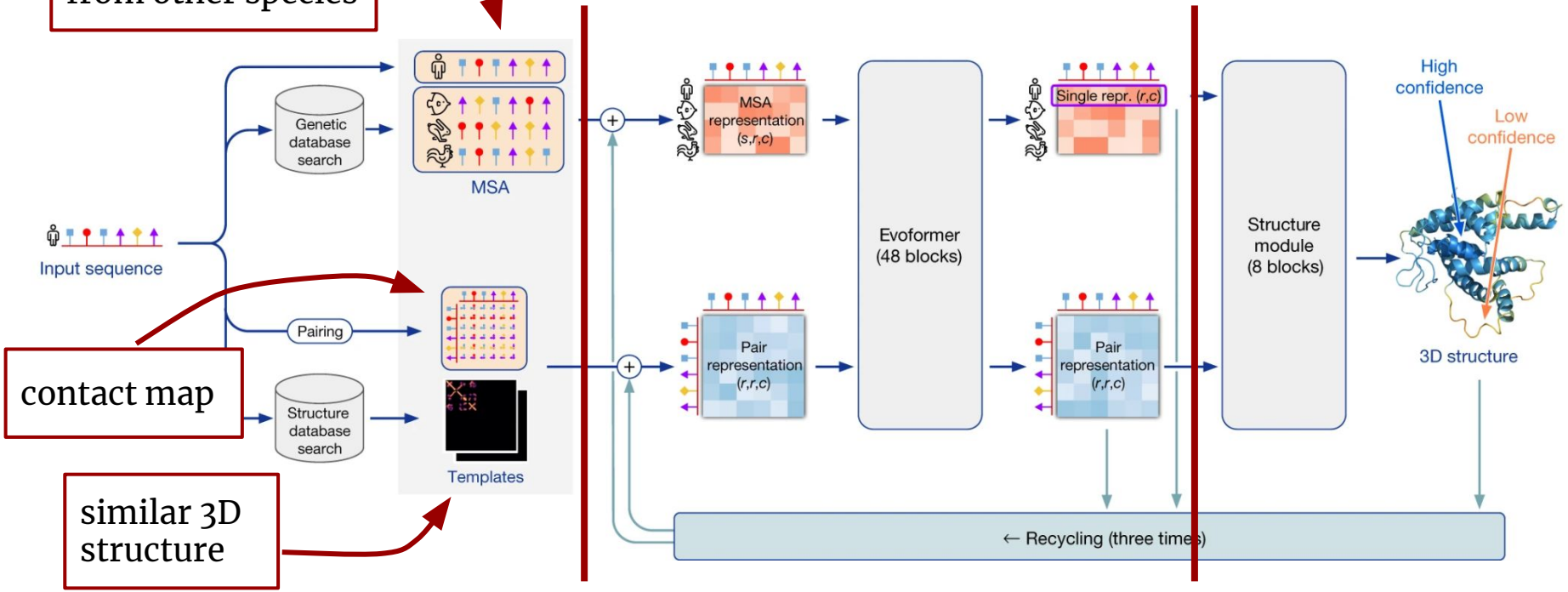
AlphaFold2



1. **encoding**

AlphaFold2

similar proteins from other species



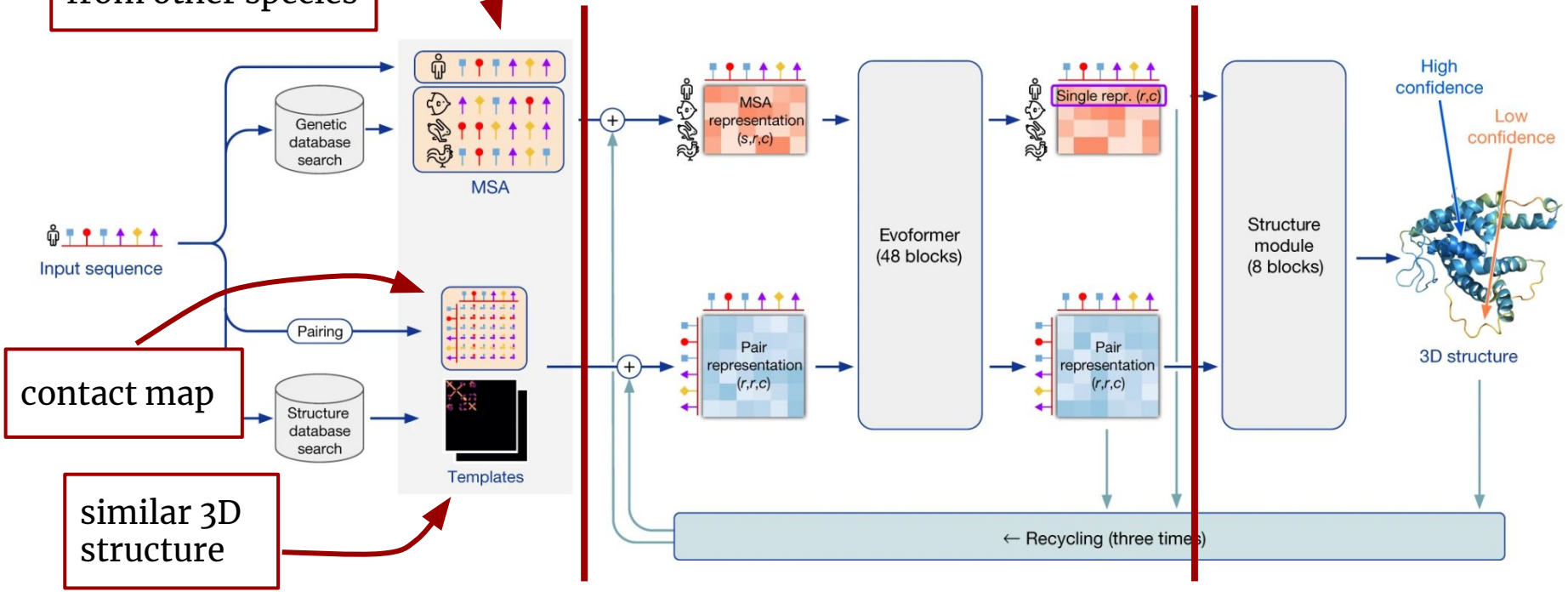
contact map

similar 3D structure

1. encoding

AlphaFold2

similar proteins from other species



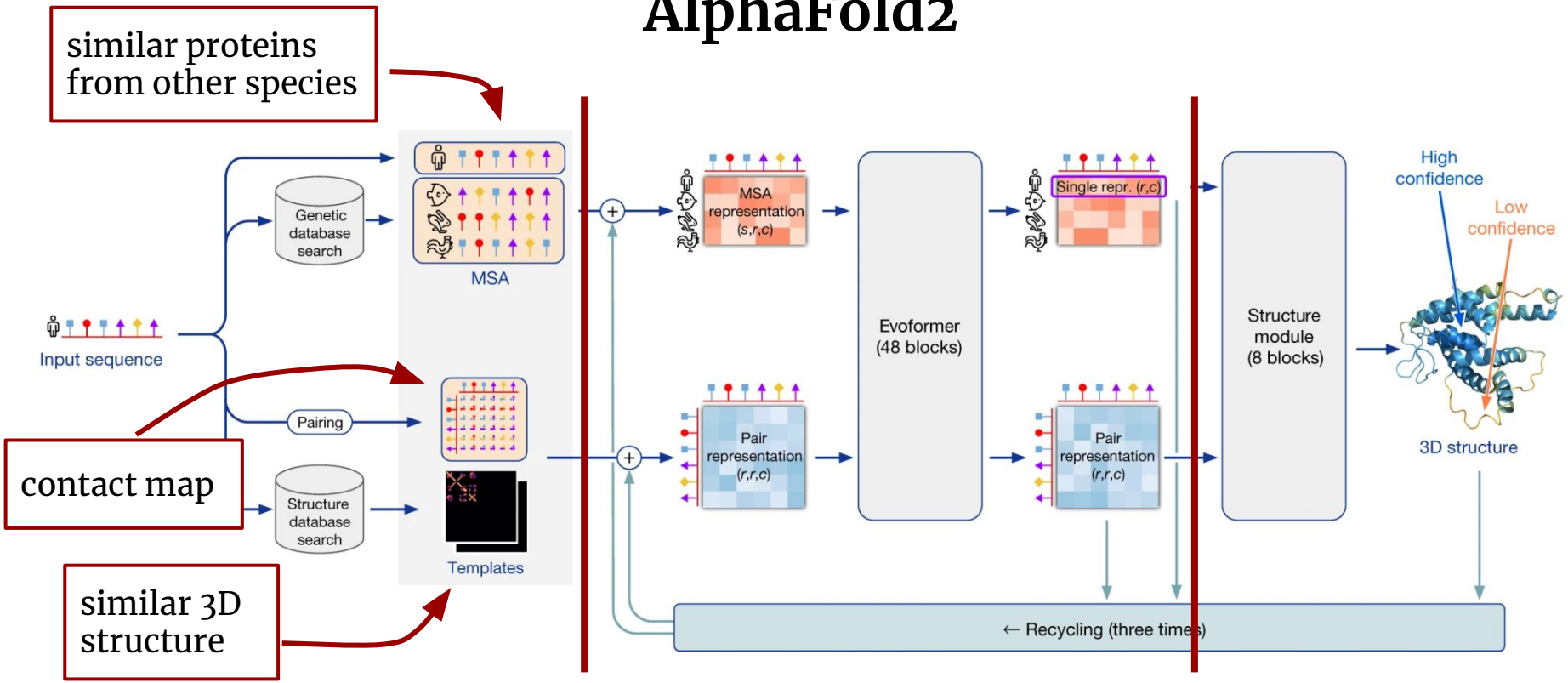
contact map

similar 3D structure

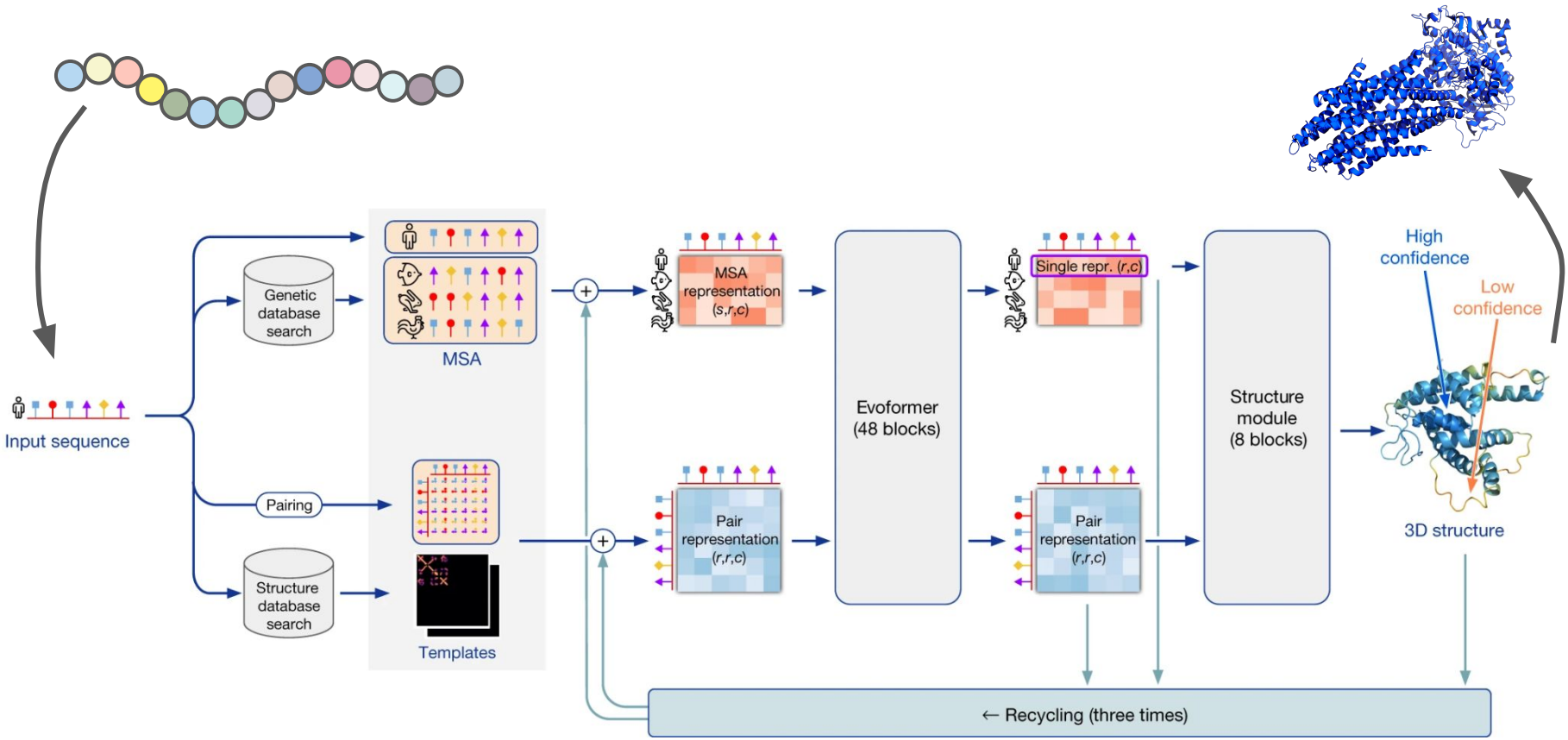
1. encoding

2.

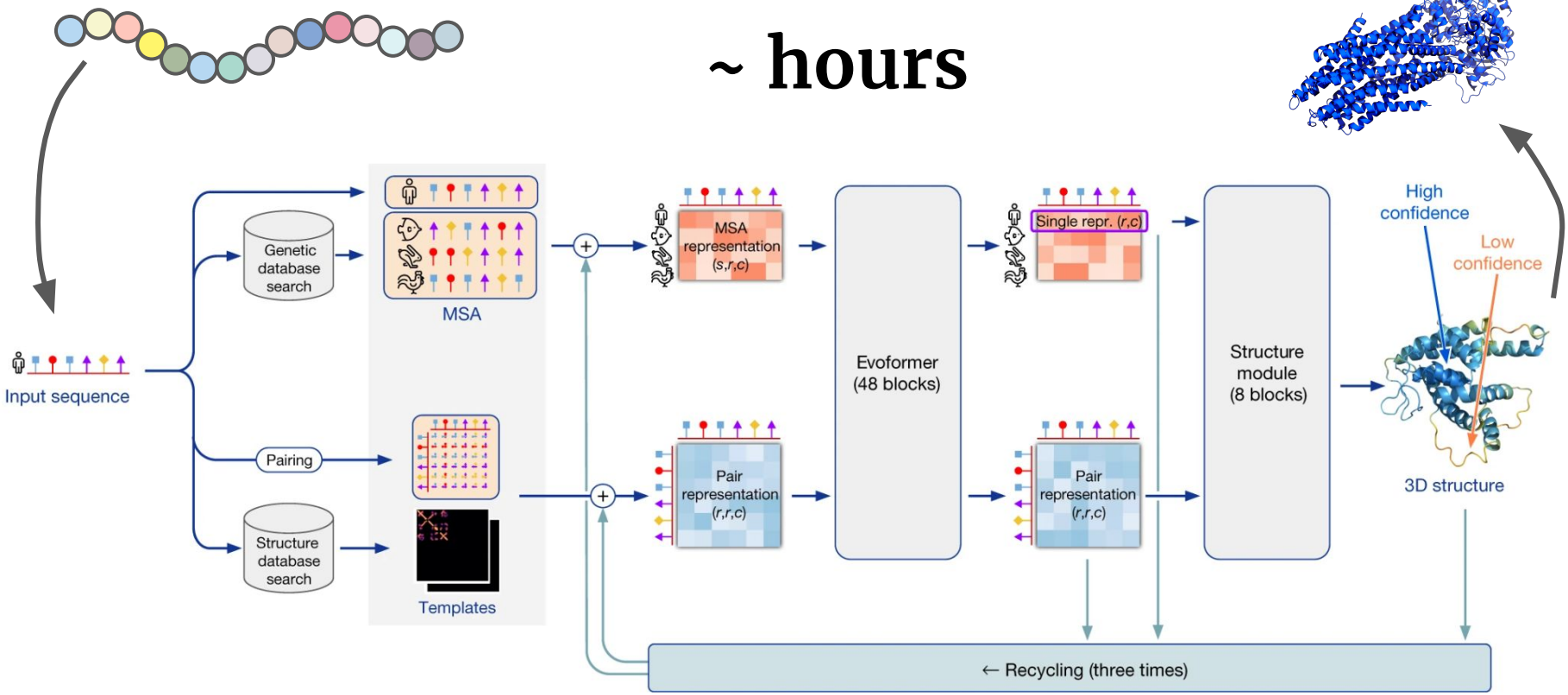
AlphaFold2



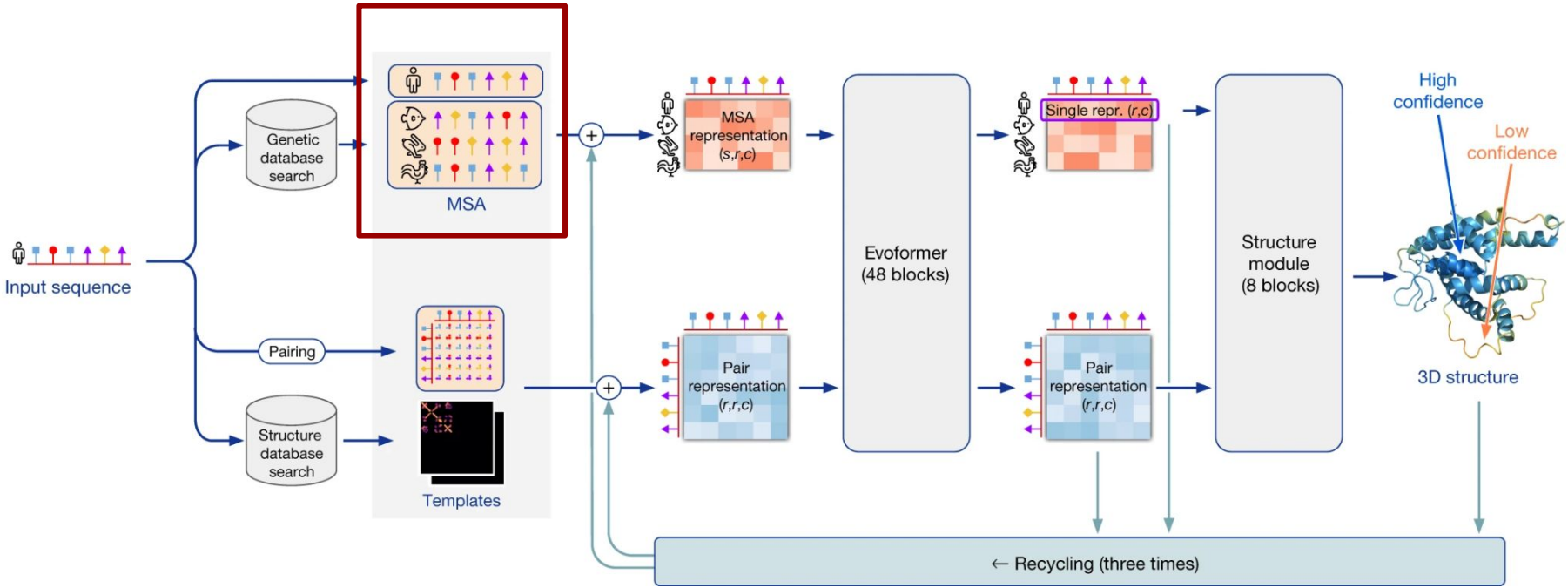
1. encoding
2.
3. modelling


Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

~ hours

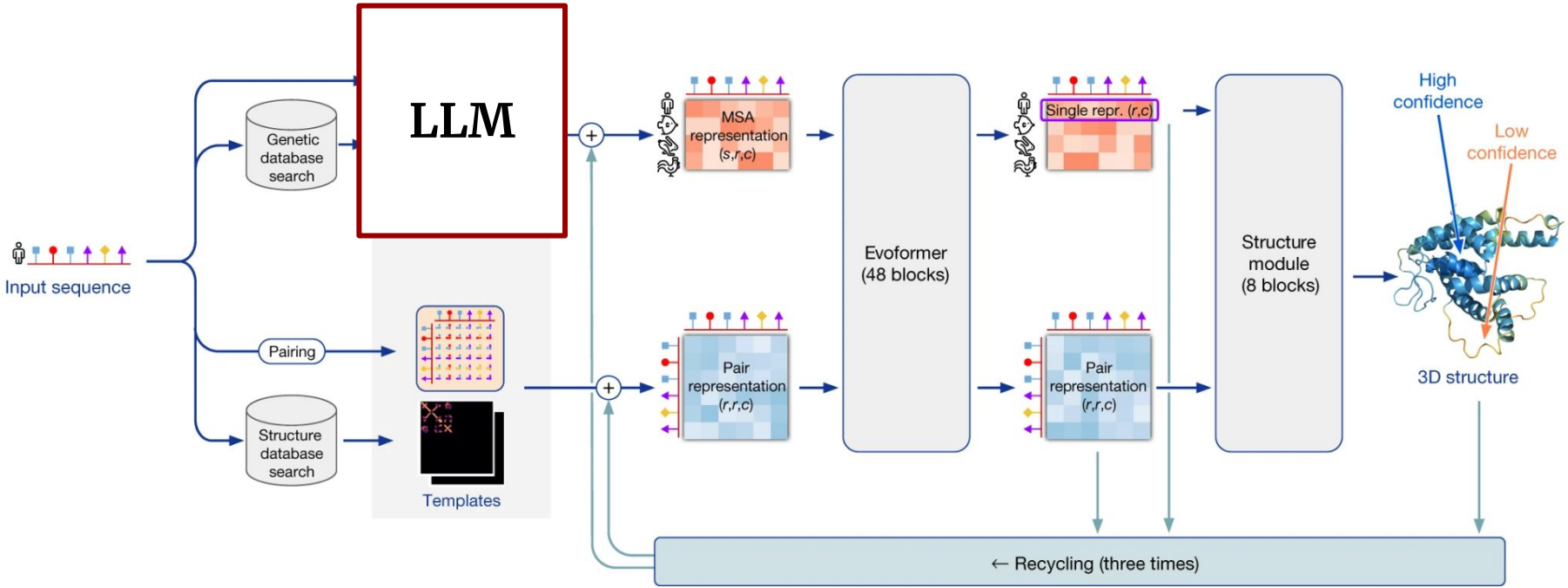

Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

slow


Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

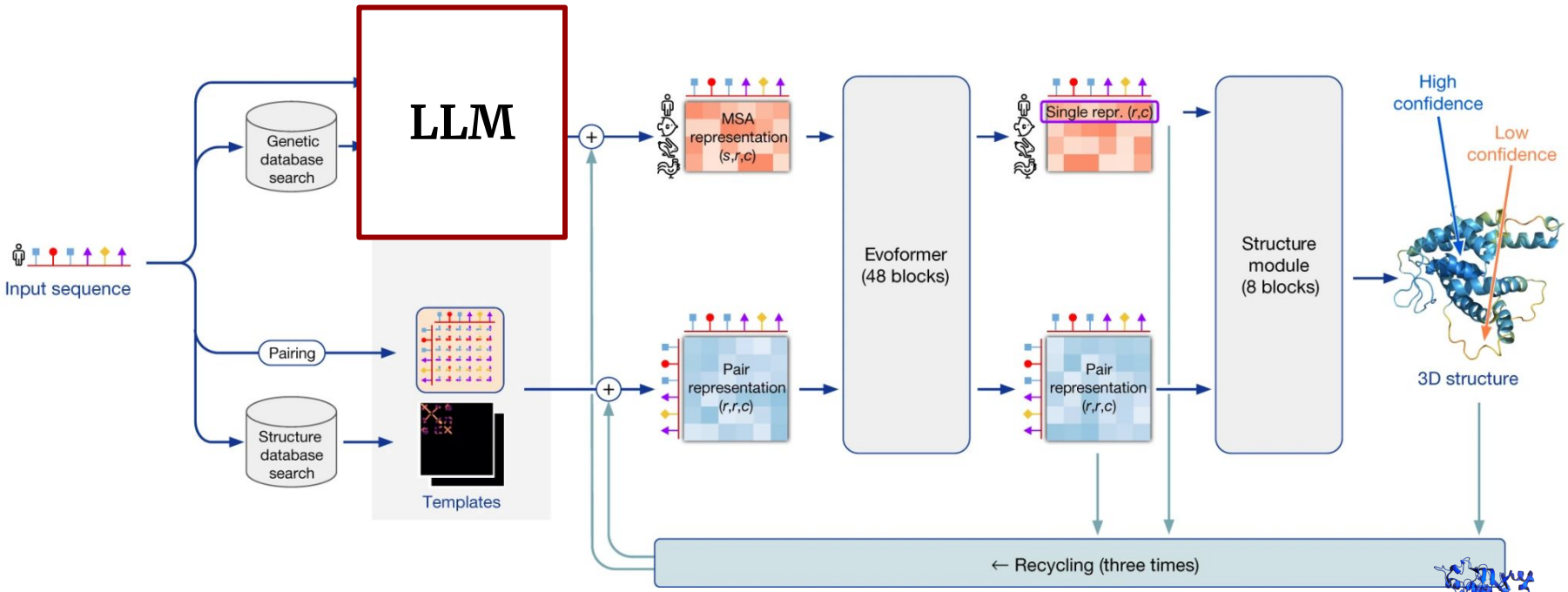
ESMfold

fast

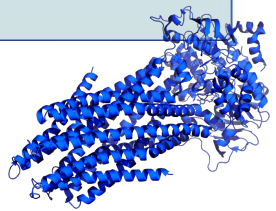


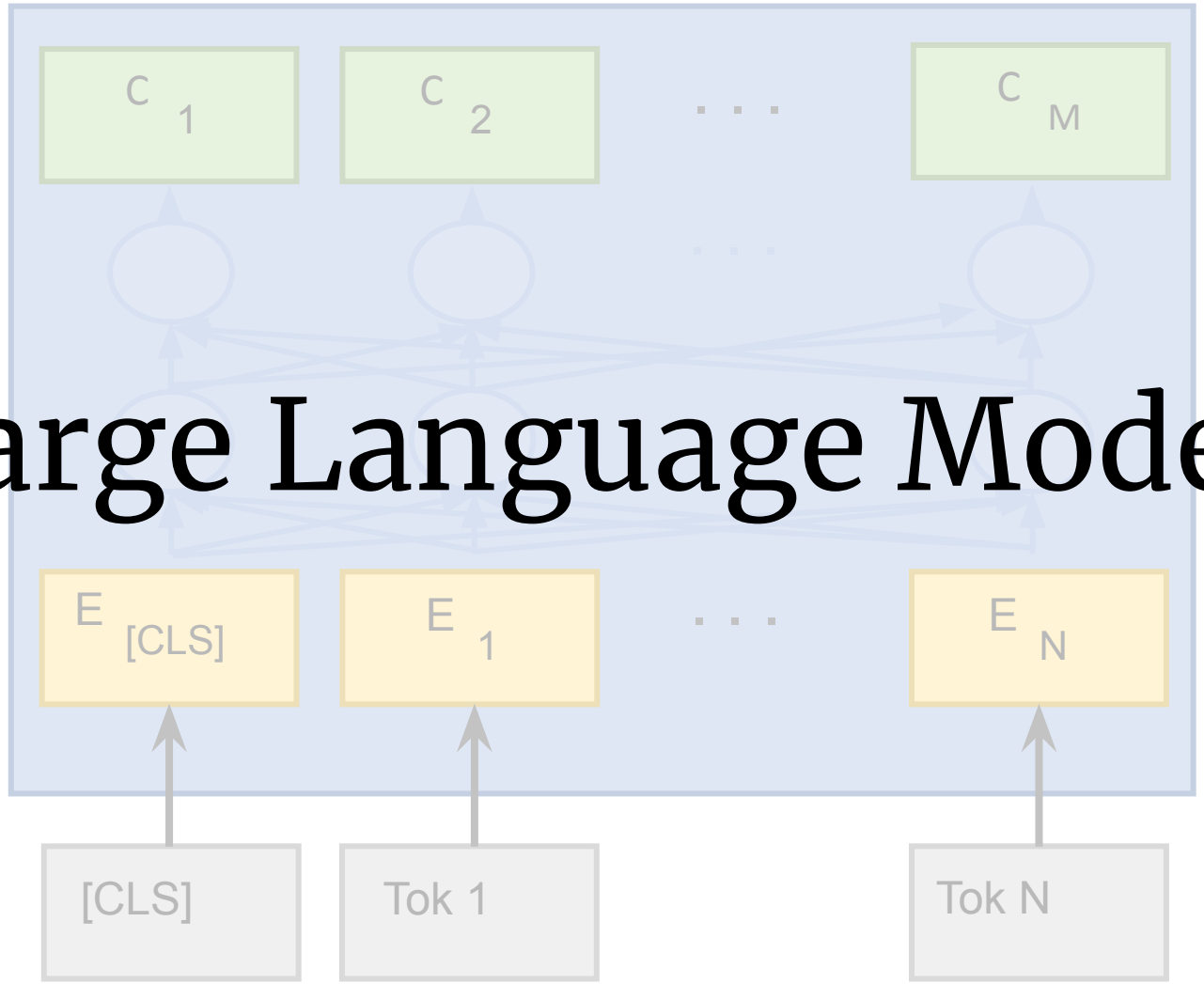
ESMfold

fast

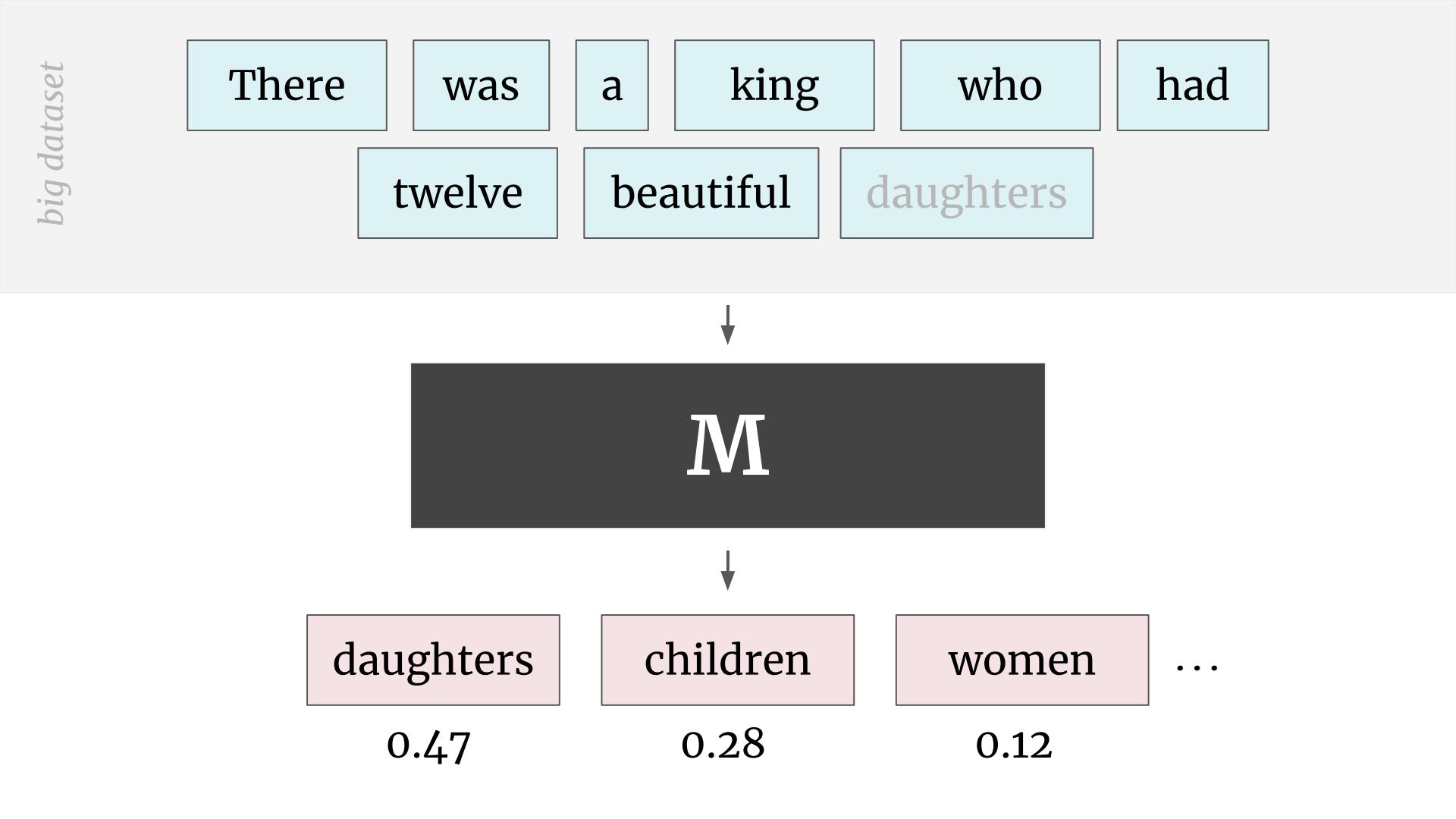


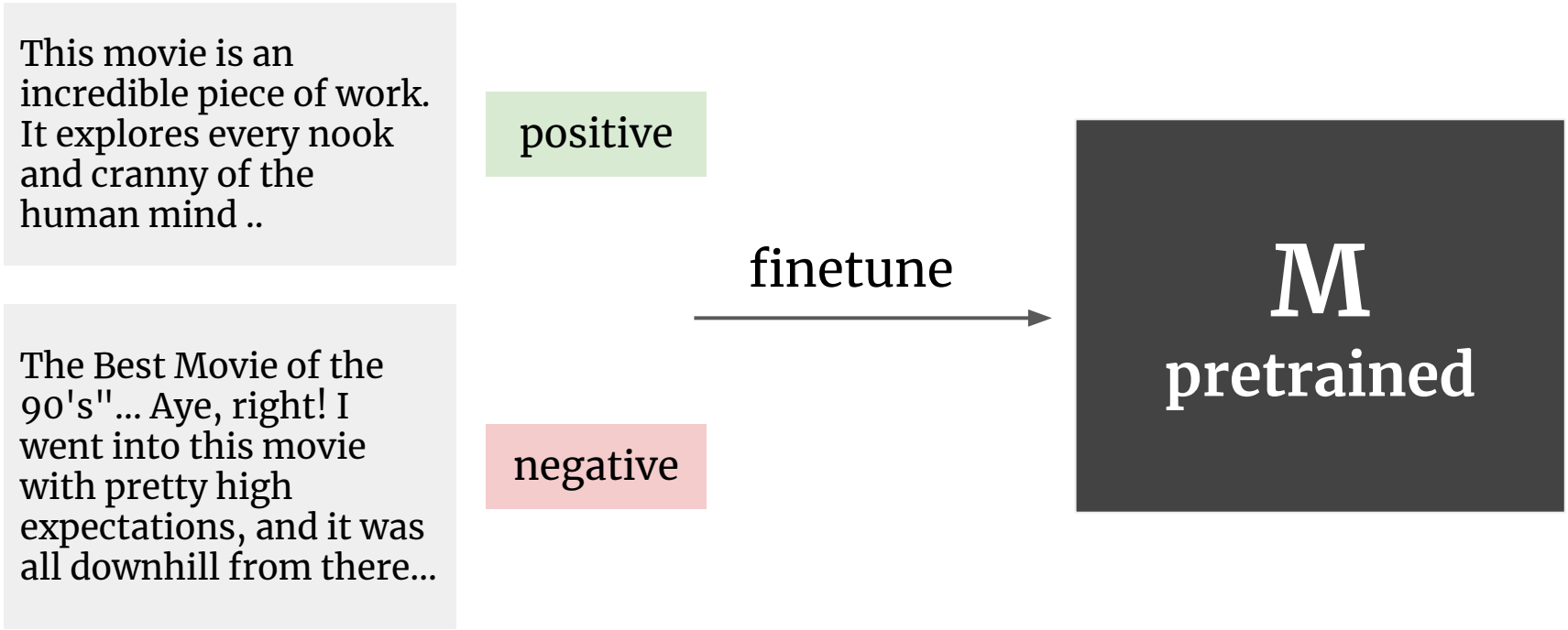
~ seconds





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 t | new 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

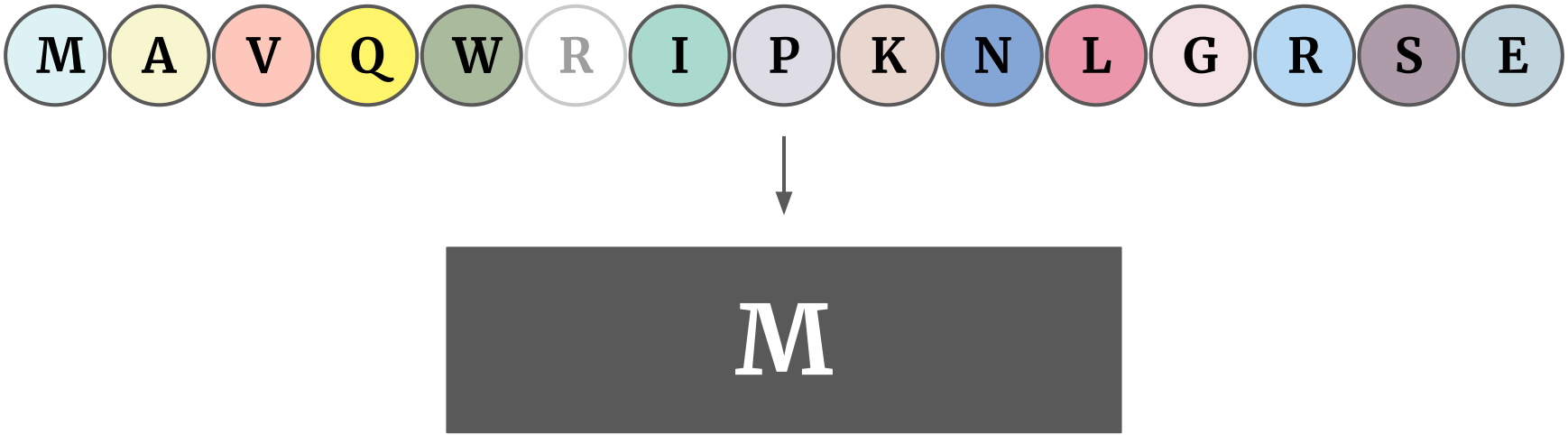
```
from transformers import AutoTokenizer, AutoModelForSequenceClassification
from transformers import TrainingArguments, Trainer


dataset = load_my_dataset()

model = AutoModelForSequenceClassification.from_pretrained("distilbert-base-uncased")
tokenizer = AutoTokenizer.from_pretrained("distilbert-base-uncased")

tokenized_dataset = dataset.map(tokenizer, batched=True)

training_args = TrainingArguments(learning_rate=2e-5, num_train_epochs=5)

trainer = Trainer(model=model,
    args=training_args,
    train_dataset=tokenized_dataset["train"],
    eval_dataset=tokenized_dataset["test"],
    tokenizer=tokenizer)

trainer.train()
```

protein ~ language

amino acid ~ word

protein ~ language

amino acid ~ word



Rostlab / **prot_bert** like 44

Fill-Mask PyTorch Transformers Uniref100 protein protein language model AutoTrain Compatible

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

Rostlab / **prot_bert** like 44

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 mas
introduced in this paper and first released in this rep
amino acids: it only works with capital letter amino a

Model description

ProtBert is based on Bert model which pretrained or
supervised fashion. This means it was pretrained on
humans labelling them in any way (which is why it ca
automatic process to generate inputs and labels fro

facebook / **esm2_t36_3B_UR50D** like 2

Fill-Mask PyTorch TensorFlow Transformers esm AutoTrain Compati

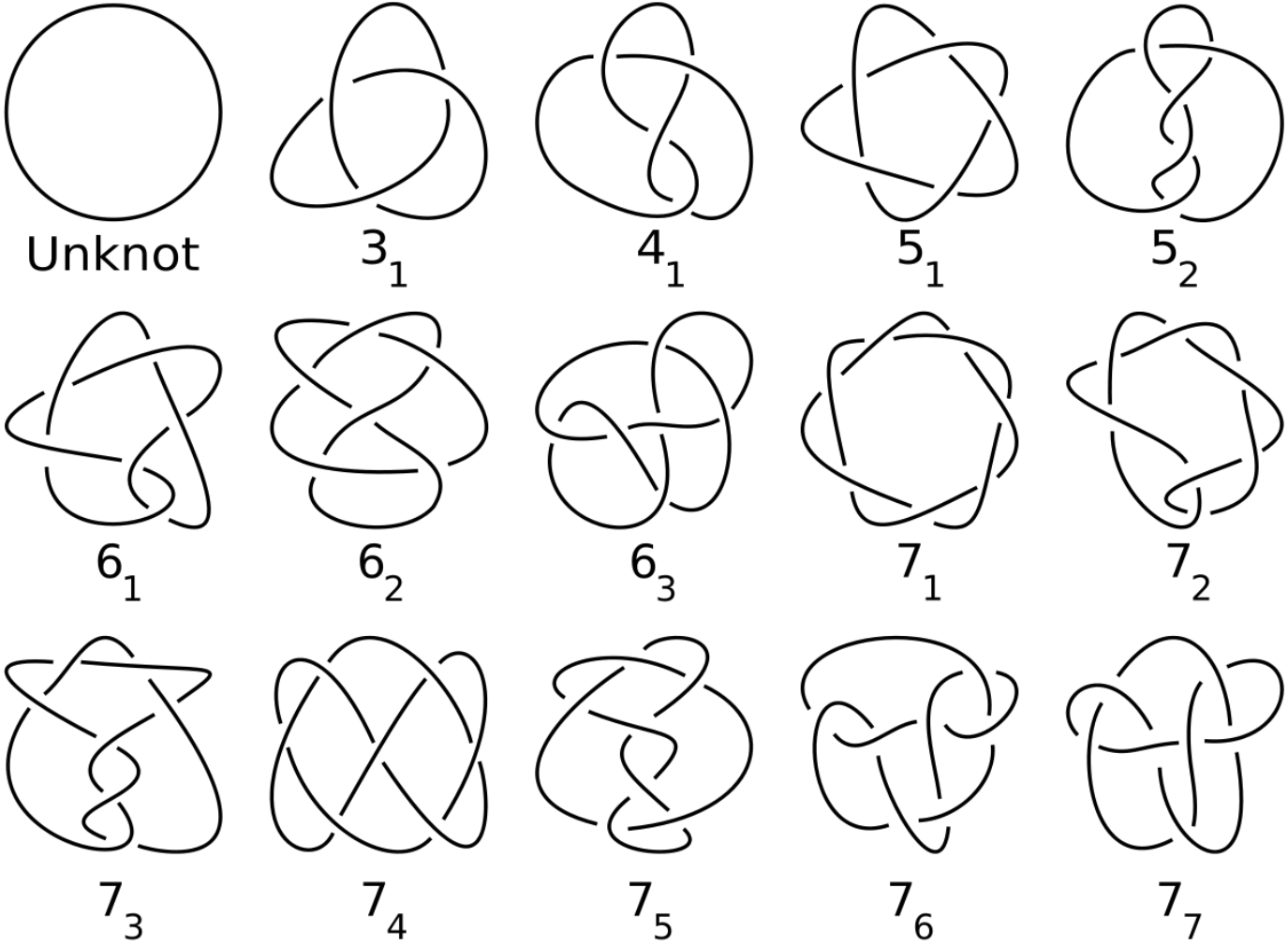
Model card Files and versions Community 1

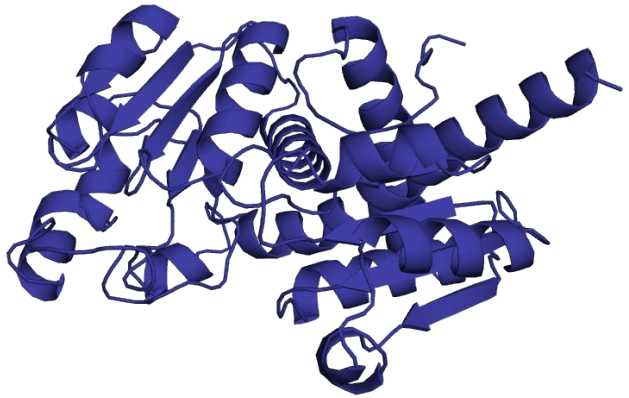
ESM-2

ESM-2 is a state-of-the-art protein model trained on a masked language modelling objecti
suitable for fine-tuning on a wide range of tasks that take protein sequences as input. For
information on the model architecture and training data, please refer to the accompanyin
You may also be interested in some demo notebooks (PyTorch, TensorFlow) which demon
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 genera
somewhat better accuracy, but require much more memory and time to train:

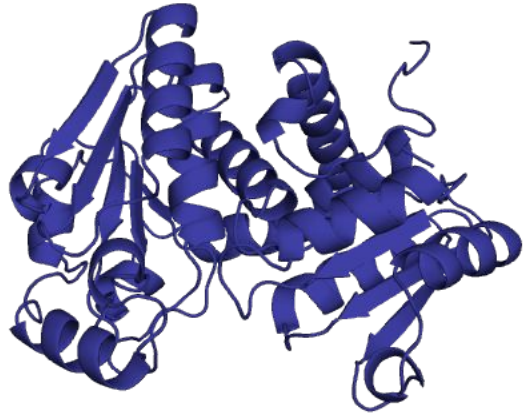
Knotted proteins

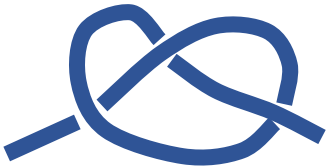

Unknot
3_1
4_1
5_1
5_2
6_1
6_2
6_3
7_1
7_2
7_3
7_4
7_5
7_6
7_7



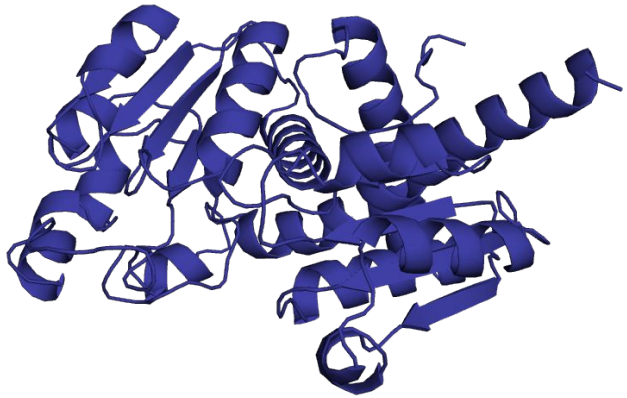
unknot

99 % proteins



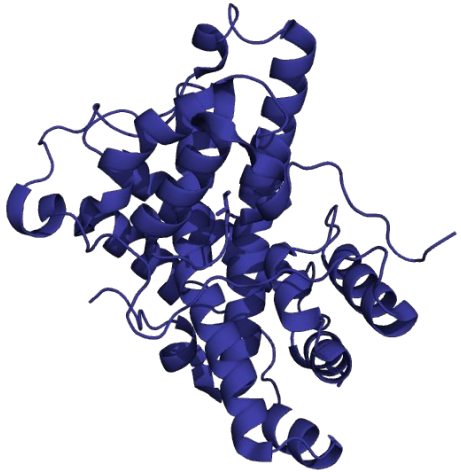


3_1 knot

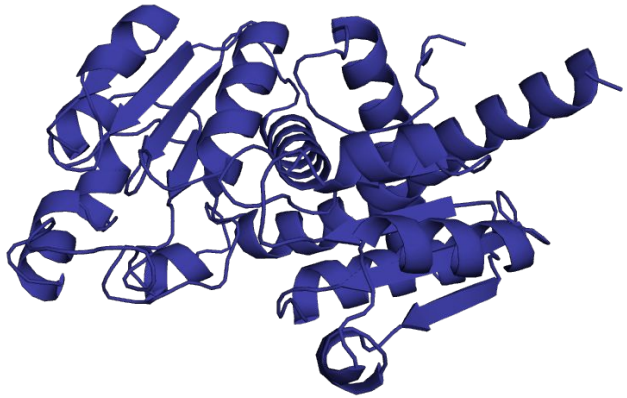


unknot

99 % proteins

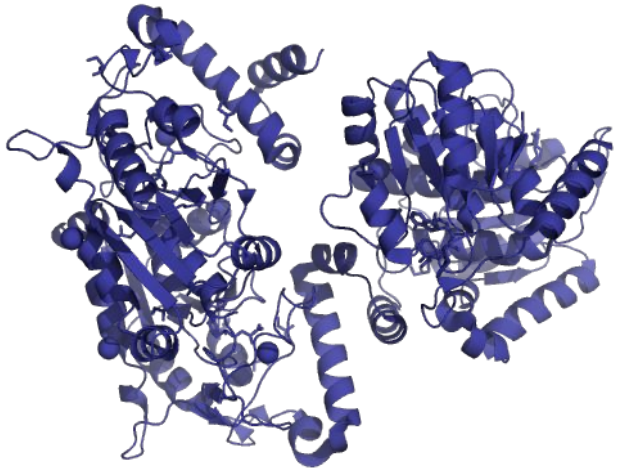


6_1 knot

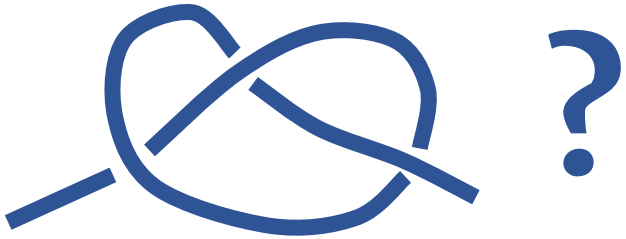


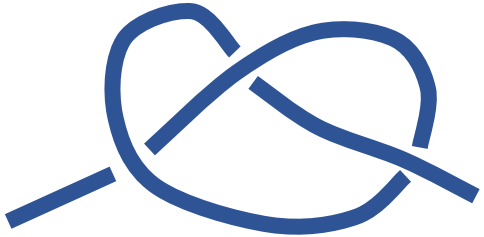
unknot

99 % proteins



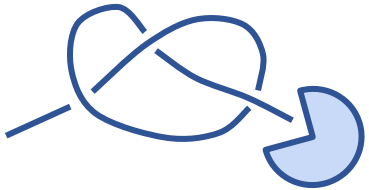
double 3_1 knot

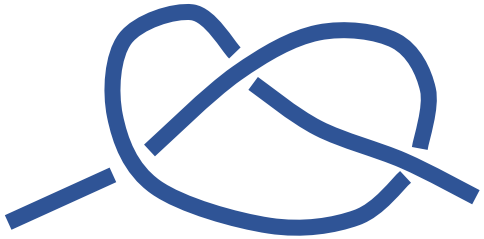




?

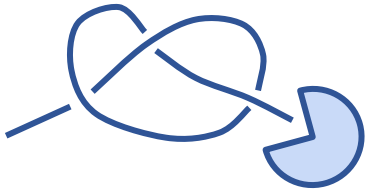
prevent from degradation



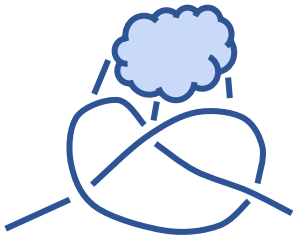


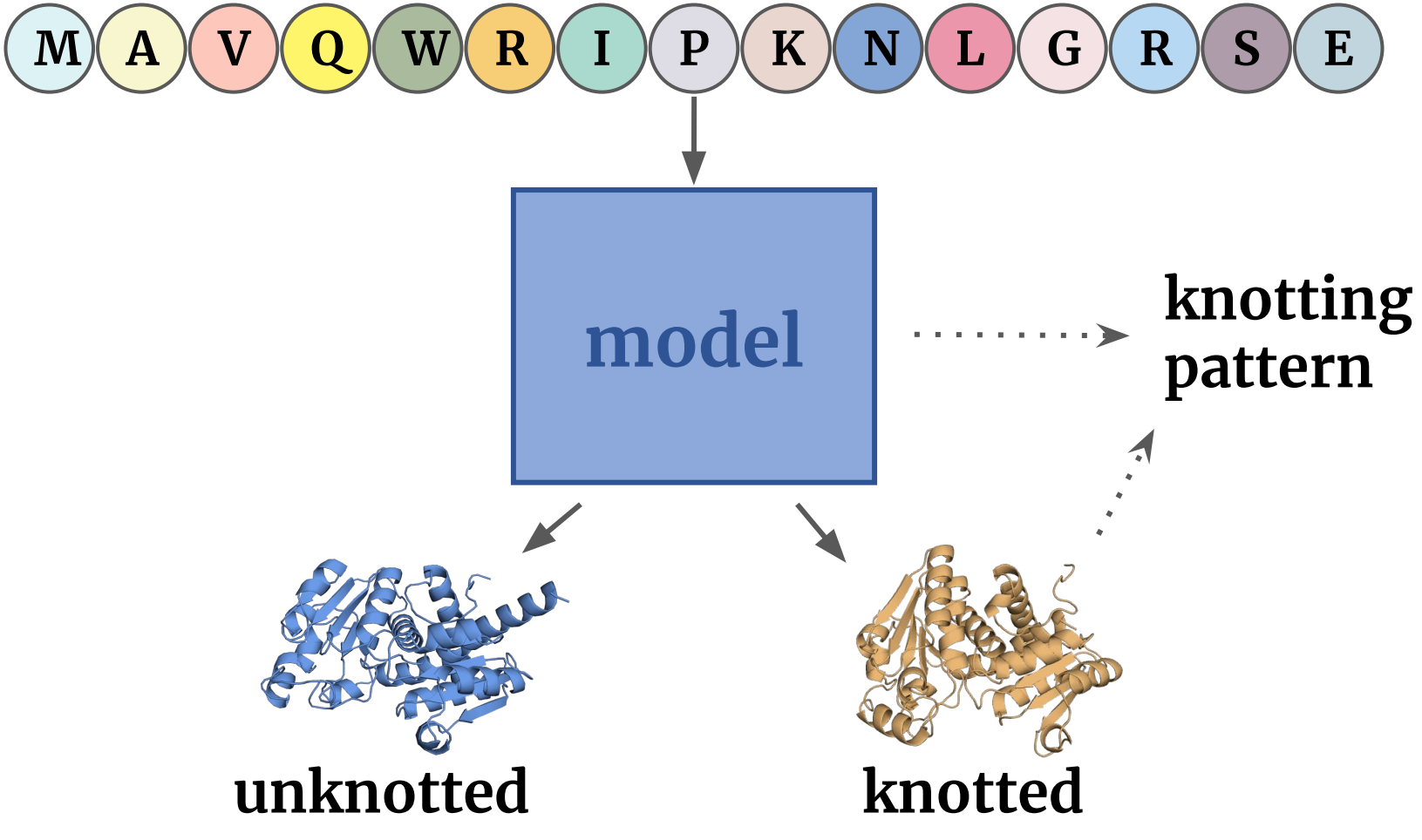
?

prevent from degradation



active site

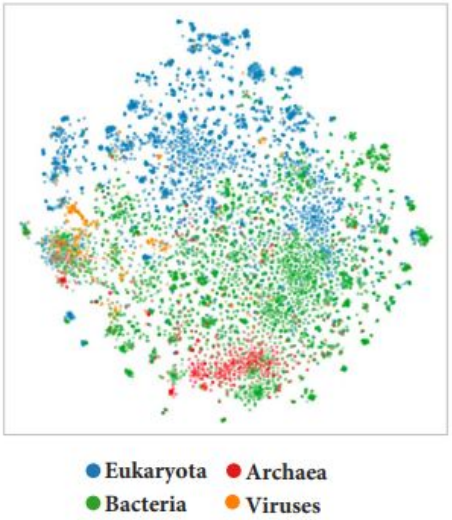



M A V Q W R I P K N L G R S E
model
knotting pattern
unknotted
knotted

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**

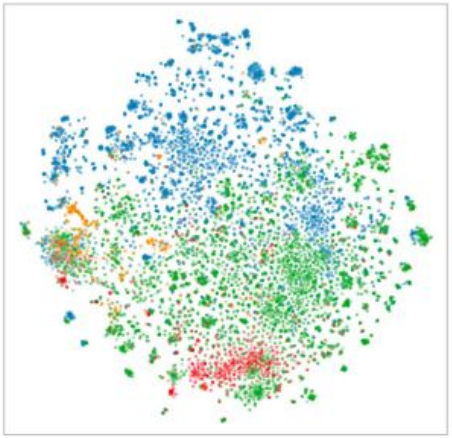
t-SNE projection of embeddings



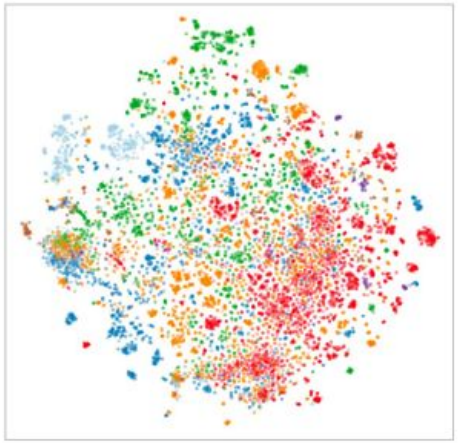
E Lineage: Kingdoms

from ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning

t-SNE projection of embeddings



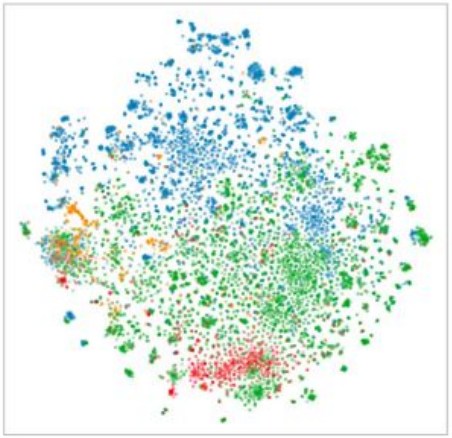
E Lineage: Kingdoms



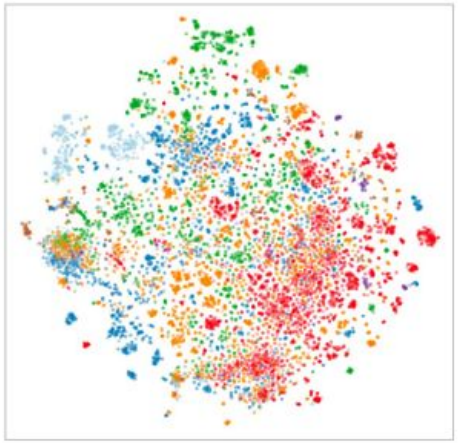
D Structure: SCOPe

from ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning

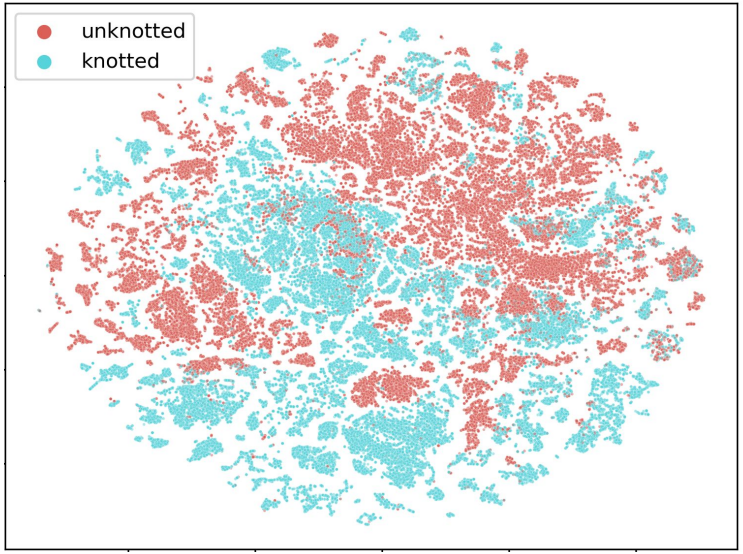
t-SNE projection of embeddings



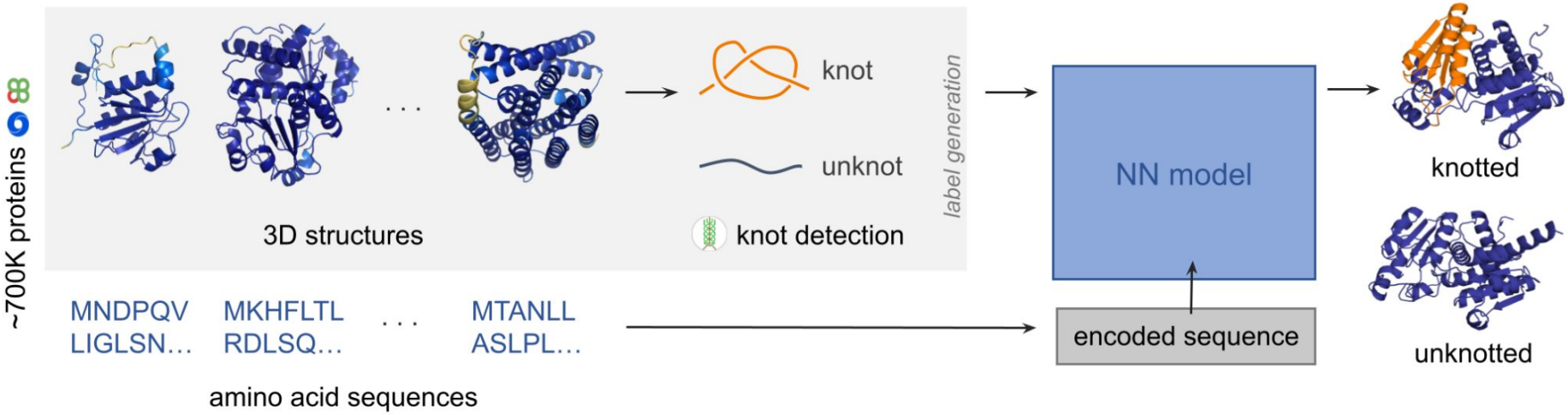
E Lineage: Kingdoms



D Structure: SCOPe



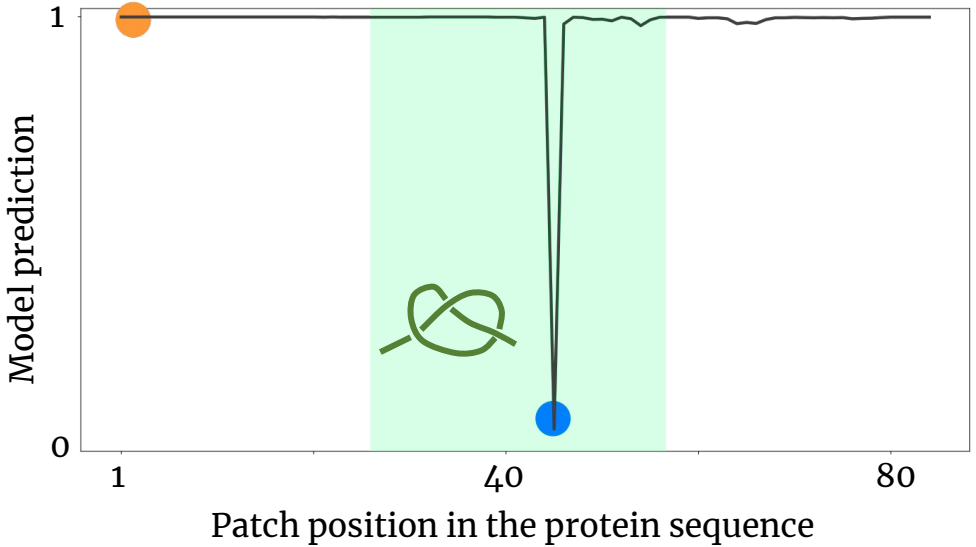
from ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning

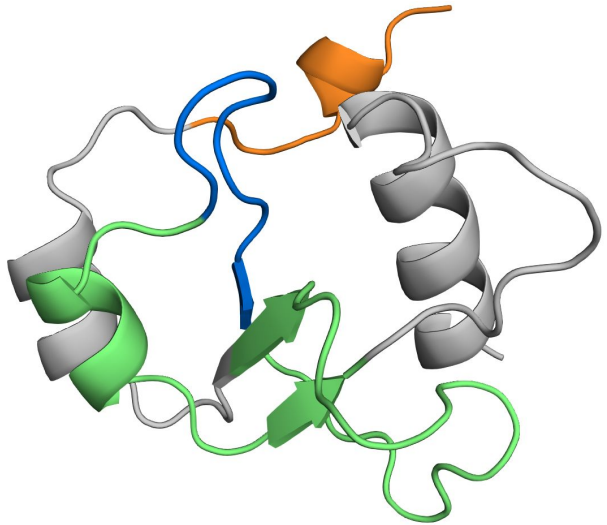

~700K proteins
3D structures
knot
unknot
knot detection
label generation
NN model
knotted
unknotted
MNDPQV LIGLSN…
MKHFLTL RDLSQ…
MTANLL ASLPL…
amino acid sequences
encoded sequence

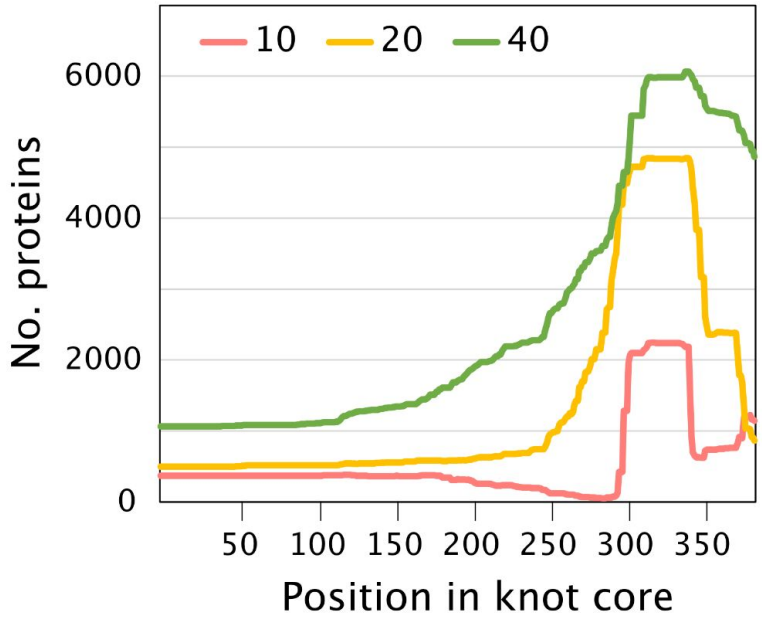
	Dataset size	Accuracy	TPR	TNR
Whole test dataset	39 412	**0.9845**	0.9865	0.9825
SPOUT	7 371	0.9887	0.9951	0.9090
TDD	612	0.9901	0.9965	0.8333
DUF	716	0.9748	0.9721	0.9766
AdoMet synthase	1 794	0.9899	0.9929	0.9708
Carbonic anhydrase	1531	0.9588	0.9737	0.9313
UCH	477	0.9056	0.9602	0.7520
ATCase/OTCase	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

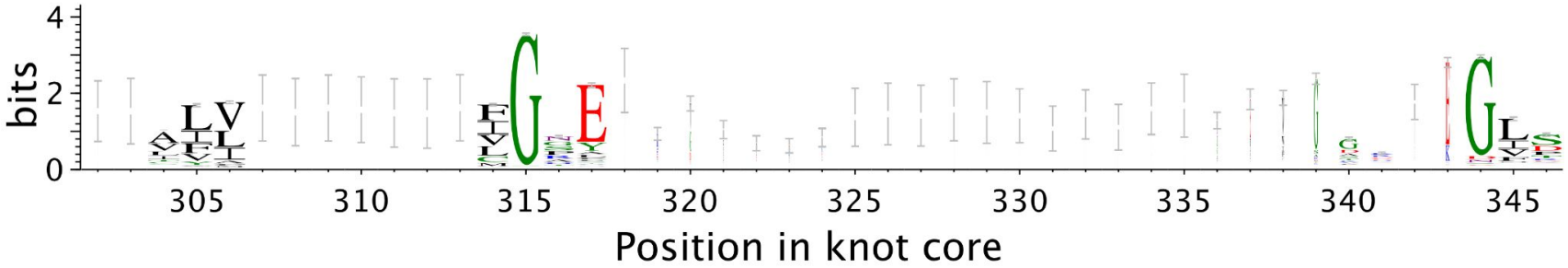
Interpretation

XXXXXXXXXXYTDLEPYLQSTKLPIYGALLDGENIYELVDKSKGILVIGNESKGIRSTIQNFIQKPITIPRIGQAESLNAAVATGIIVGQLTL
MXXXXXXXXXXTDLEPYLQSTKLPIYGALLDGENIYELVDKSKGILVIGNESKGIRSTIQNFIQKPITIPRIGQAESLNAAVATGIIVGQLTL
...
MGGIFRVNTYYTDLEPYLQSTKLPIYGALLDGENIYELVDKSKGXXXXXXXXXXIRSTIQNFIQKPITIPRIGQAESLNAAVATGIIVGQLTL
...
MGGIFRVNTYYTDLEPYLQSTKLPIYGALLDGENIYELVDKSKGILVIGNESKGIRSTIQNFIQKPITIPRIGQAESLNAAVAXXXXXXXXXX





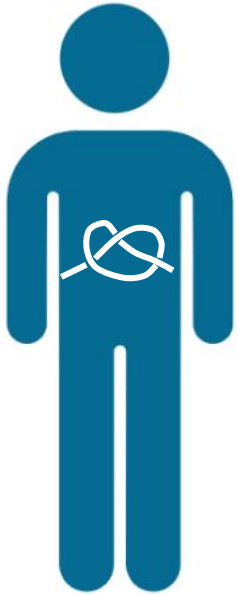

10
20
40
6000
4000
2000
0
No. proteins
50
100
150
200
250
300
350
Position in knot core


bits
4
2
0
305
310
315
320
325
330
335
340
345
Position in knot core

Future ideas

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

How is it useful?



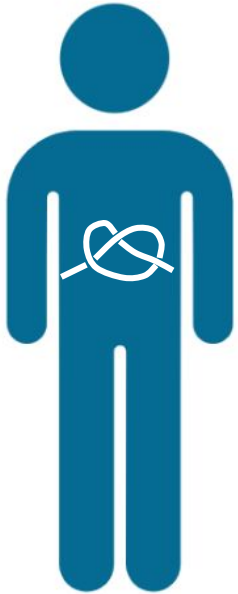
How is it useful?

The Unique Cysteine Knot Regulates the Pleotropic Hormone Leptin

Ellinor Haglund[1], Joanna I. Sułkowska[1], Zhao He[2], Gen-Sheng Feng[2], 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

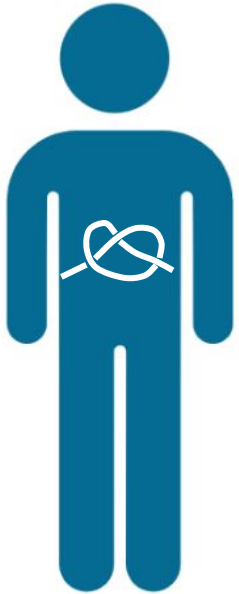
The SpoU-TrmD (SPOUT) methyltransferase superfamily was designated when structural similarity was identified between the transfer RNA–modifying enzymes TrmH (SpoU) and TrmD. SPOUT methyltransferases are found in all domains of life and predominantly modify transfer RNA or ribosomal RNA substrates, though 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.

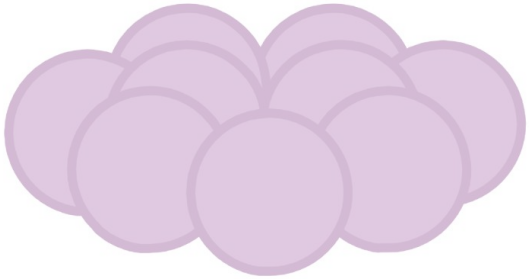
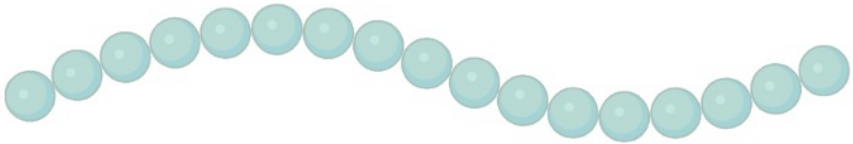
The Unique Cysteine Knot Regulates the Pleotropic Hormone Leptin

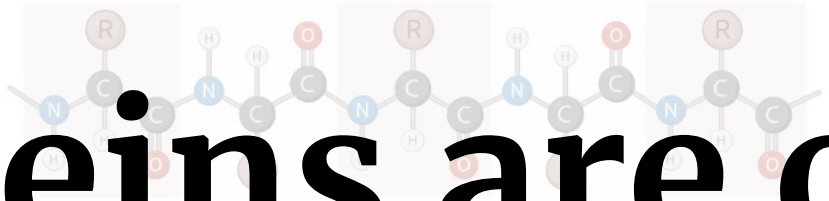
Ellinor Haglund[1], Joanna I. Sułkowska[1], Zhao He[2], Gen-Sheng Feng[2], Patricia A. Jennings[1]*, José N. Onuchic[3]*

Abstract

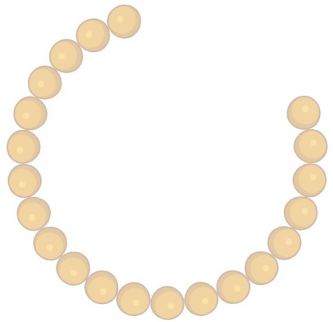
Leptin plays a key role in regulating energy intake/expenditure, metabolism and hypertension. It folds into a four-helix bundle that binds to the extracellular receptor to initiate signaling. Our work on leptin revealed a hidden complexity in the formation of a previously un-described, cysteine-knotted 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 Pawel Rubach Maciej Sikora Dawid Uchal

Thank you

GitHub with research on knotted proteins:

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

slides:

contact: eva.klimentova@ceitec.muni.cz

Questions?

GitHub with research on knotted proteins:

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

slides:

contact: eva.klimentova@ceitec.muni.cz