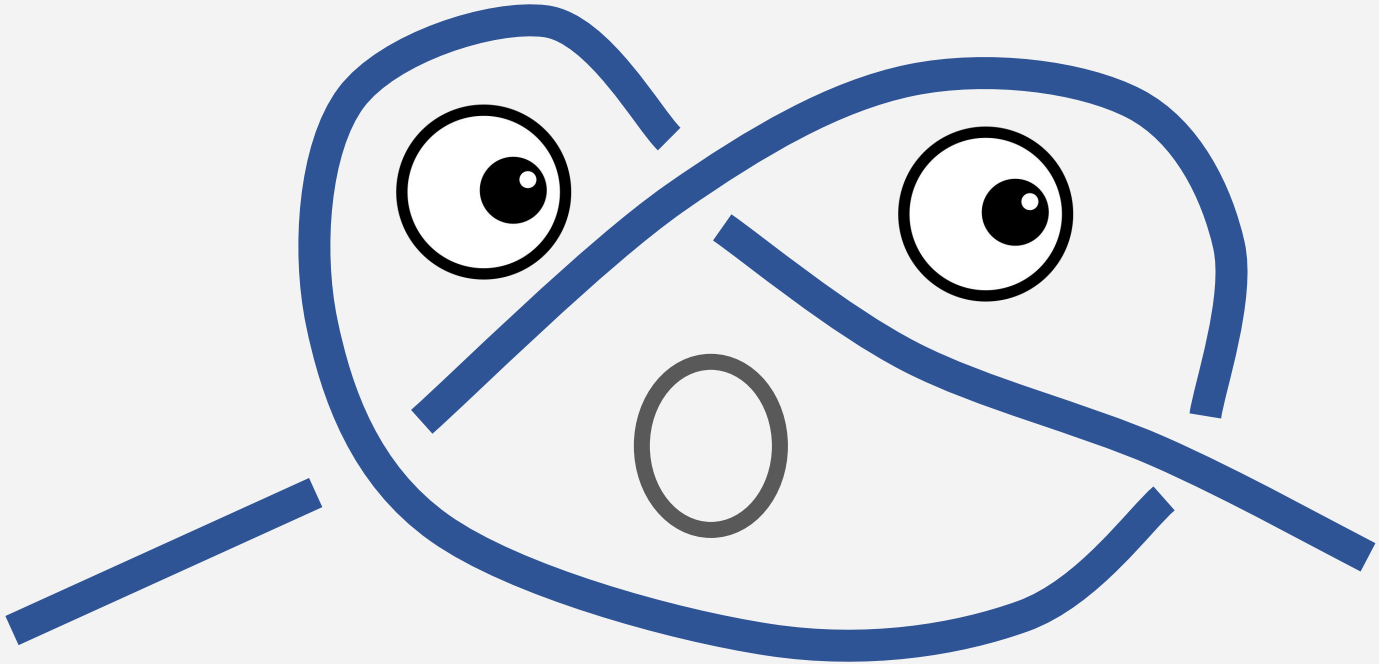
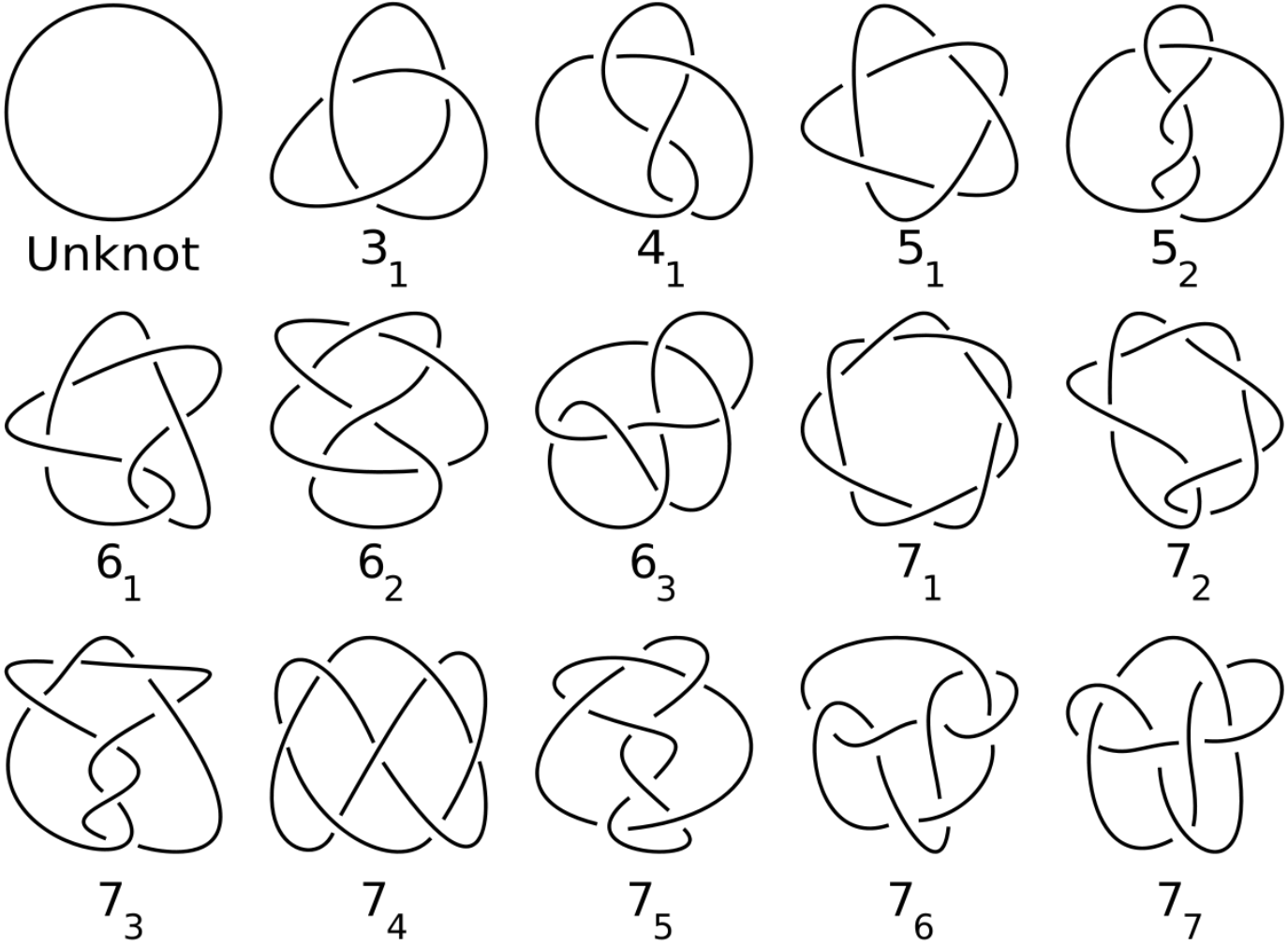
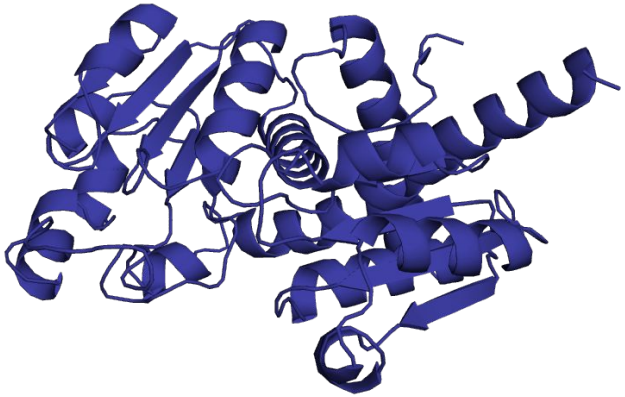
Knot or Not?
Machine learning for knotted proteins

Eva Maršálková
5. 11. 2024

?? knotted protein ??

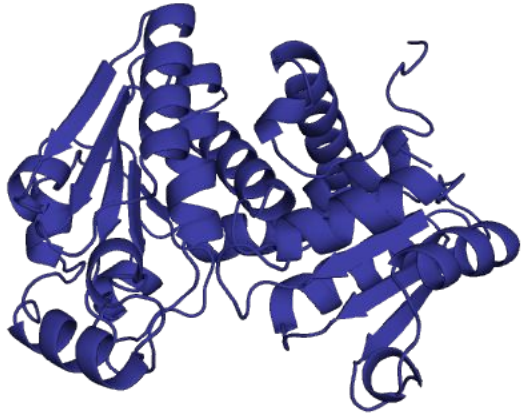


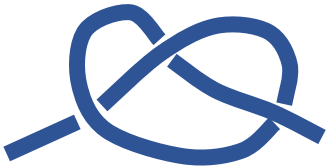

Unknot
3_1
4_1
5_1
5_2
6_1
6_2
6_3
7_1
7_2
7_3
7_4
7_5
7_6
7_7



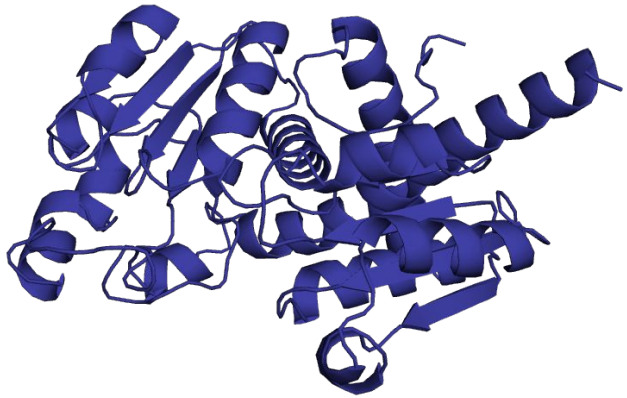
unknot

~ 99 % proteins



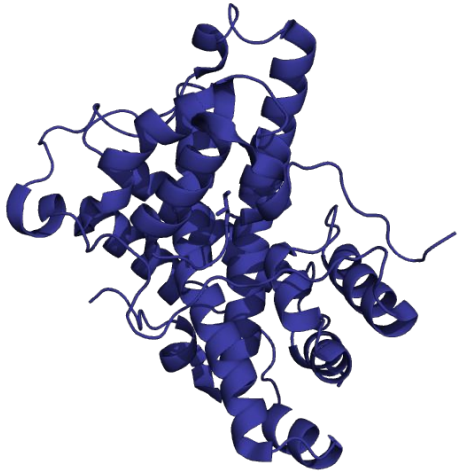


3_1 knot

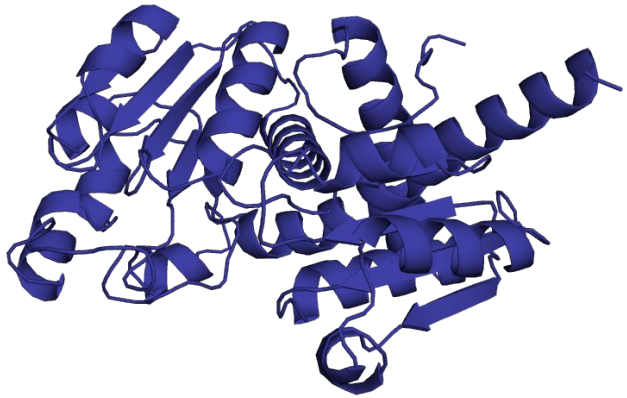


unknot

99 % proteins

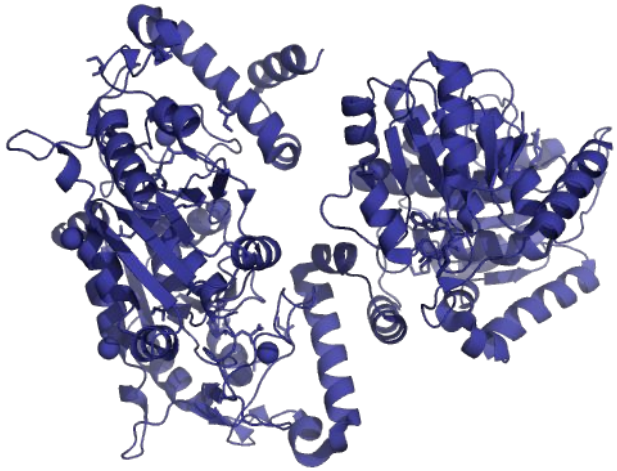


6_1 knot



unknot

99 % proteins

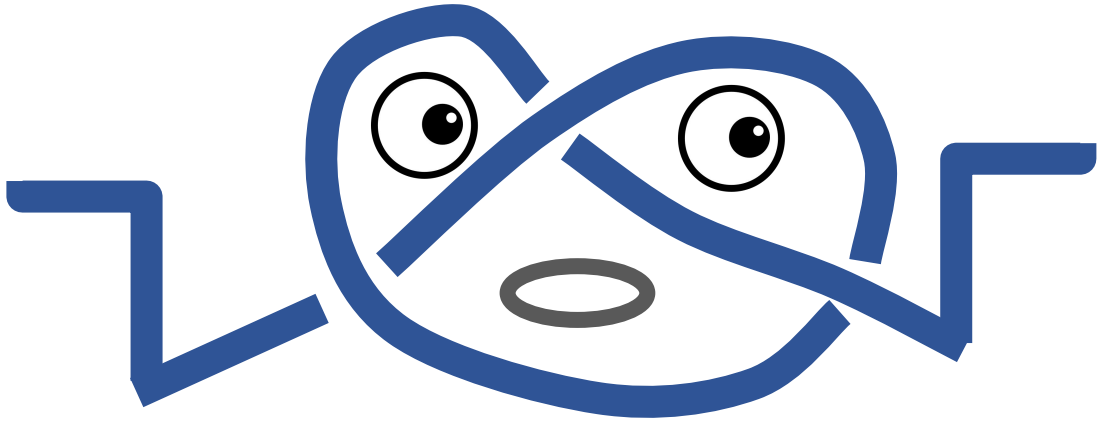


double 3_1 knot

thermal stability?

?

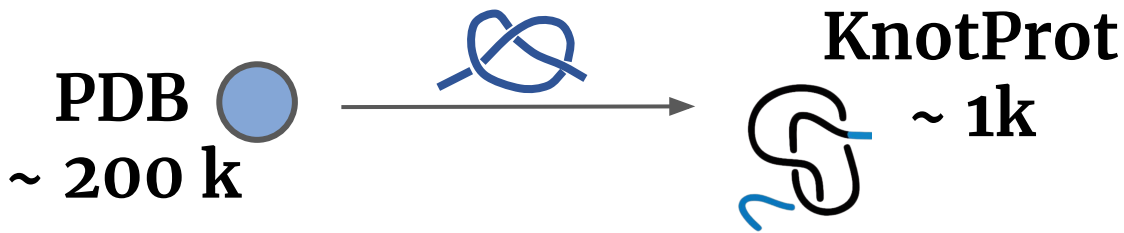
all bacterias have them



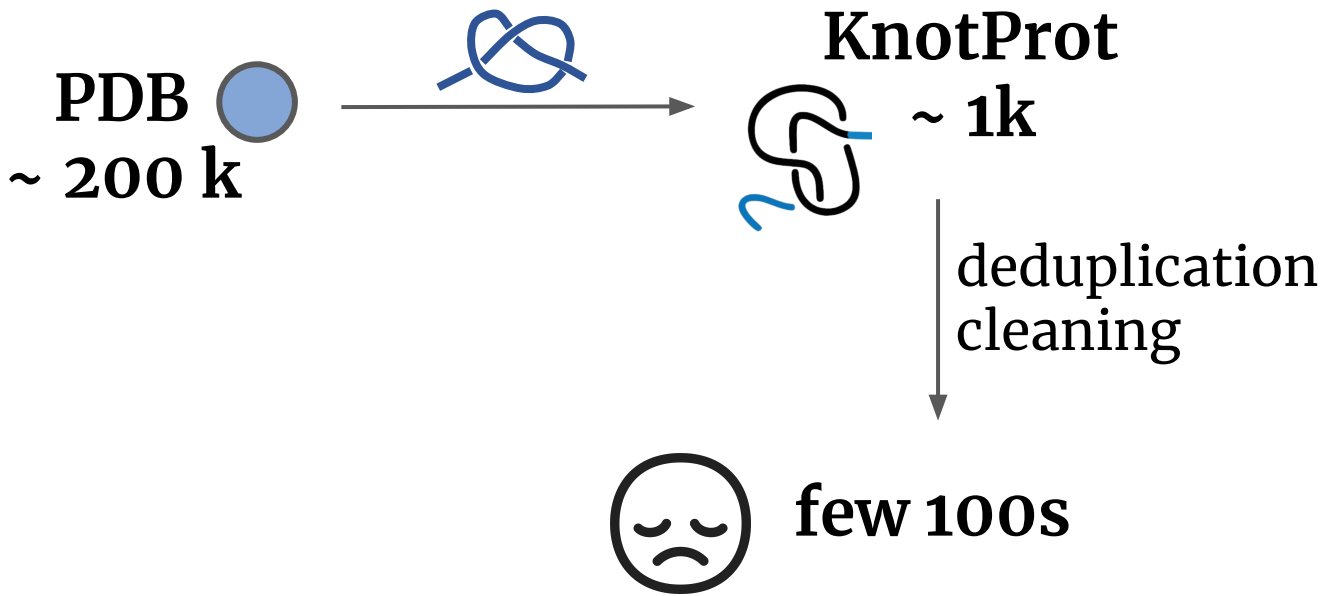
prevent from degradation?

curiosity ♡

of protein structures

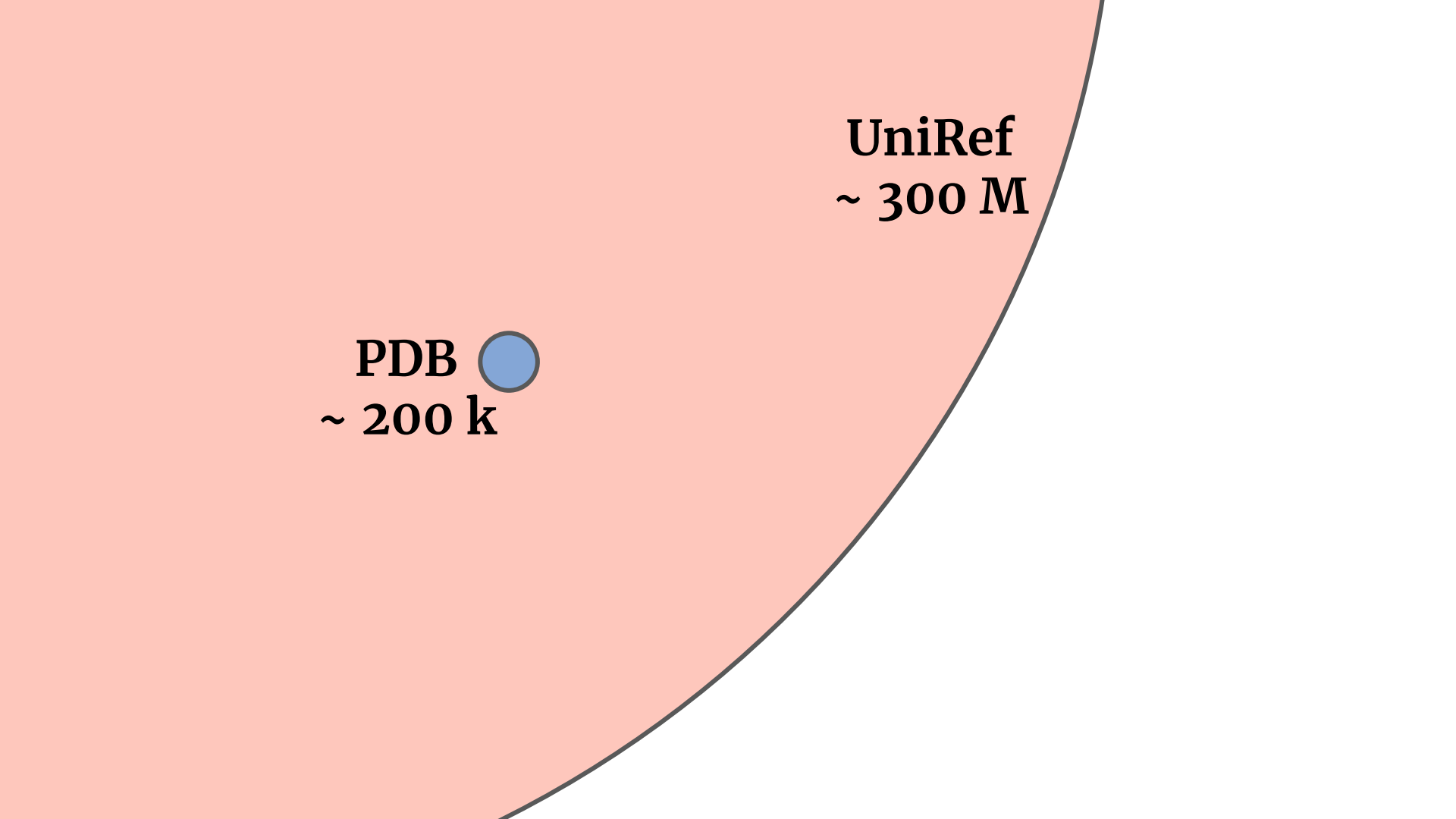


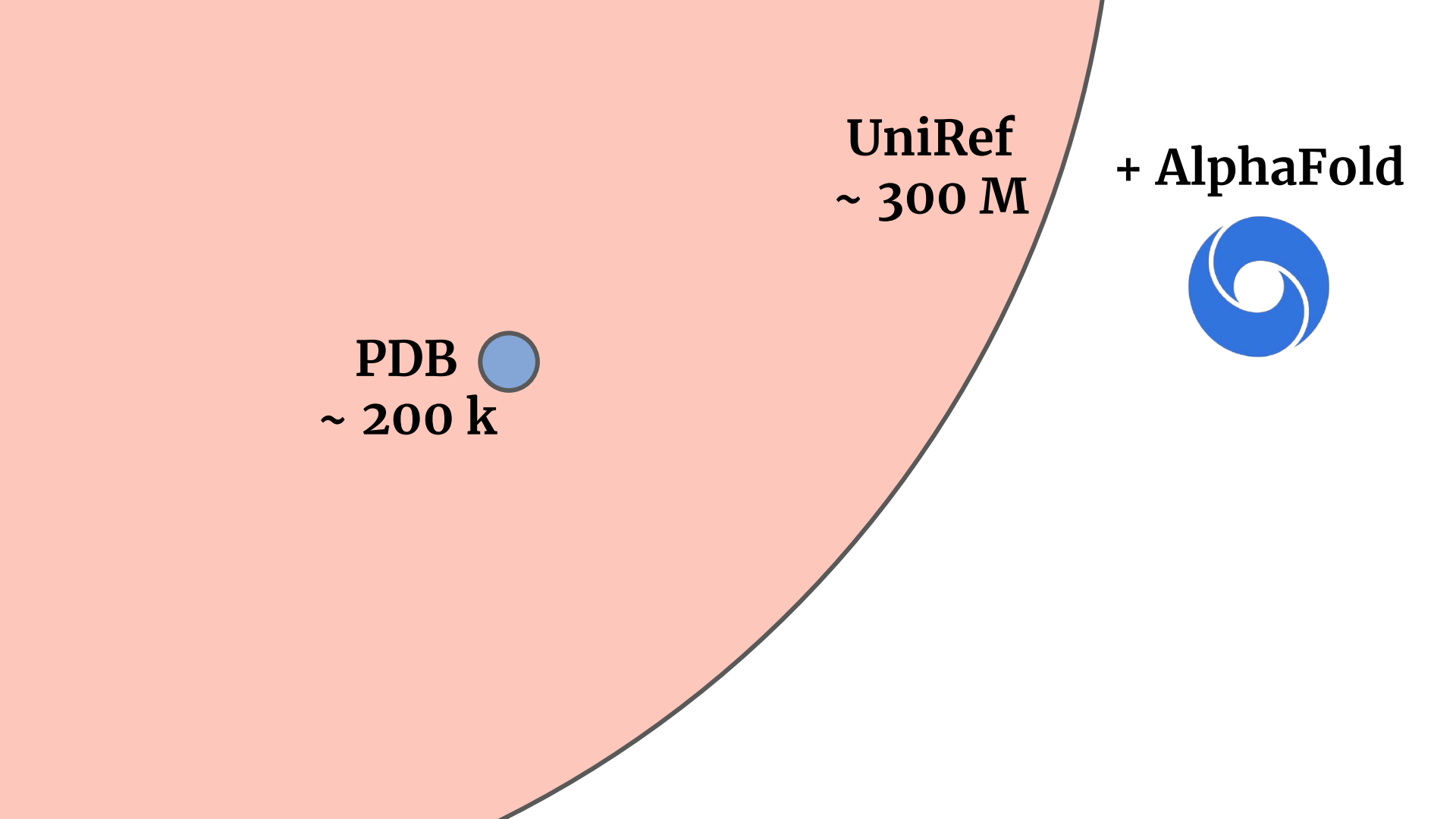
of protein structures



PDB

~ 200 k

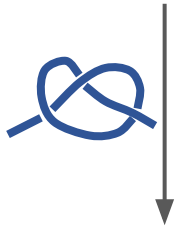

UniRef
~ 300 M
PDB
~ 200 k


UniRef
~ 300 M
+ AlphaFold
PDB
~ 200 k

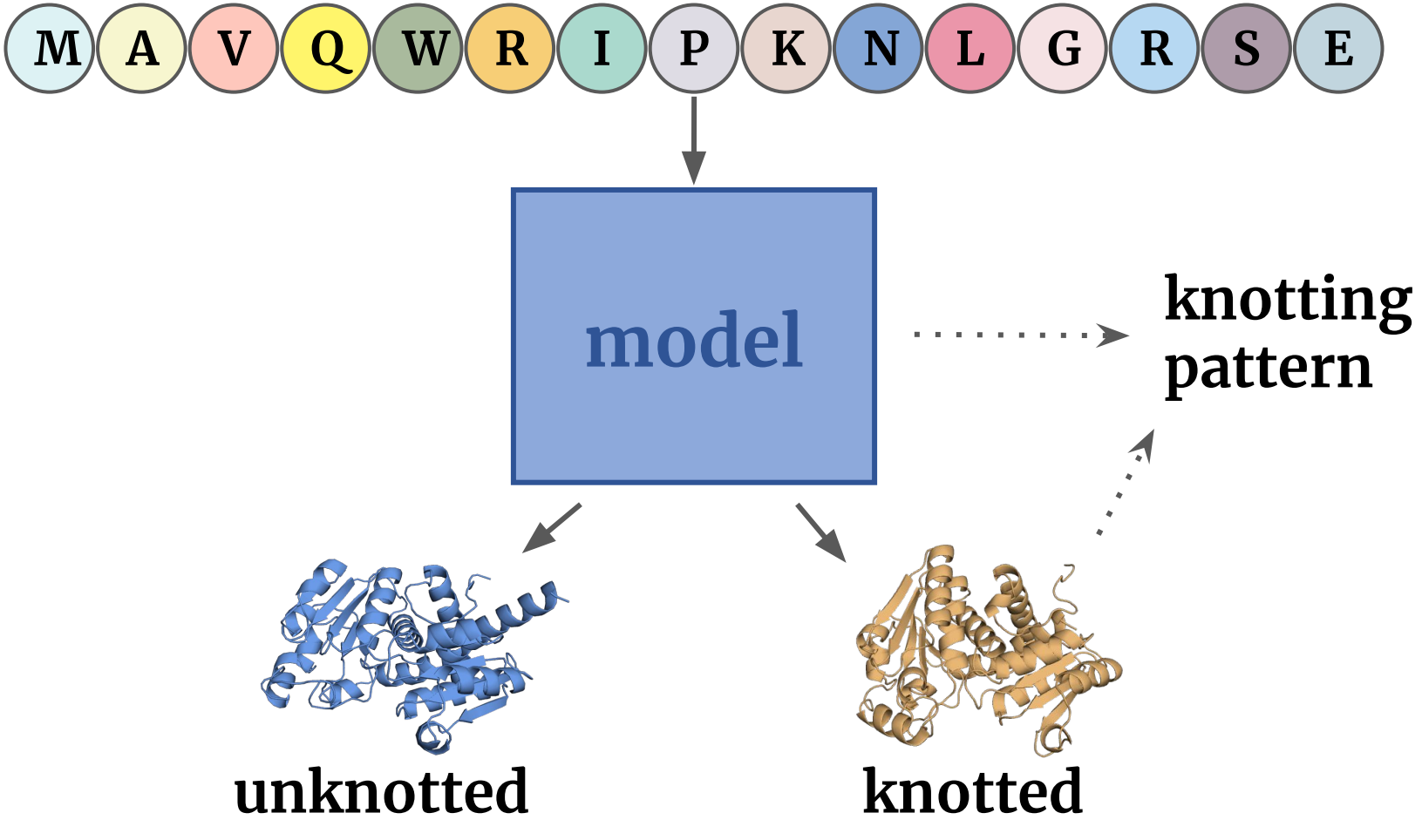
UniRef
~ 300 M

PDB
~ 200 k

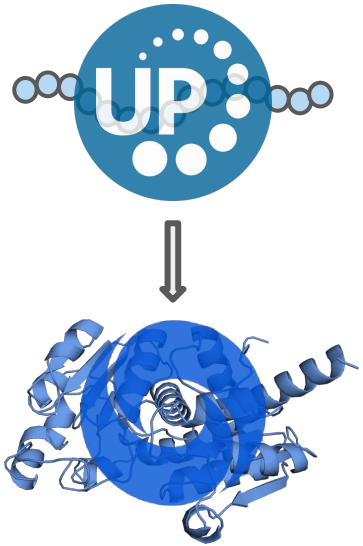
+ AlphaFold



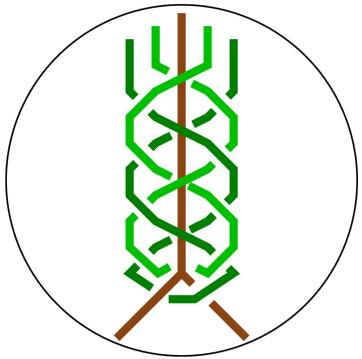
~ 100 000s


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

Getting knotted proteins



all proteins from UniProt predicted with AF



run through Topoly (knot detection)

probability of unknotting < 0.5 & one knot type prob. > 0.4 → knot

probability of unknotting > 0.5 → unknot

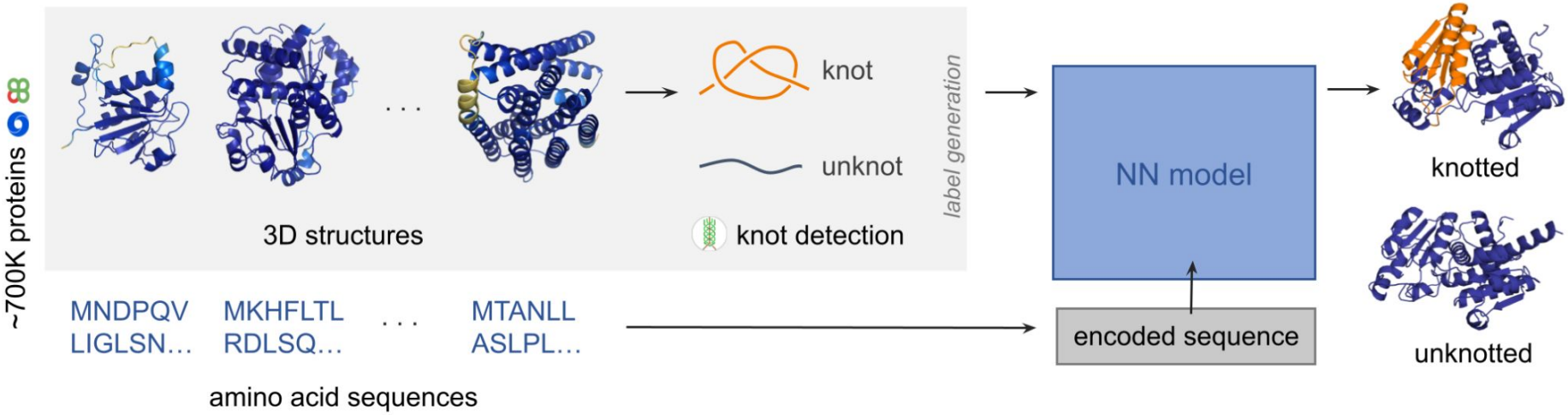
other ⋯→ trash

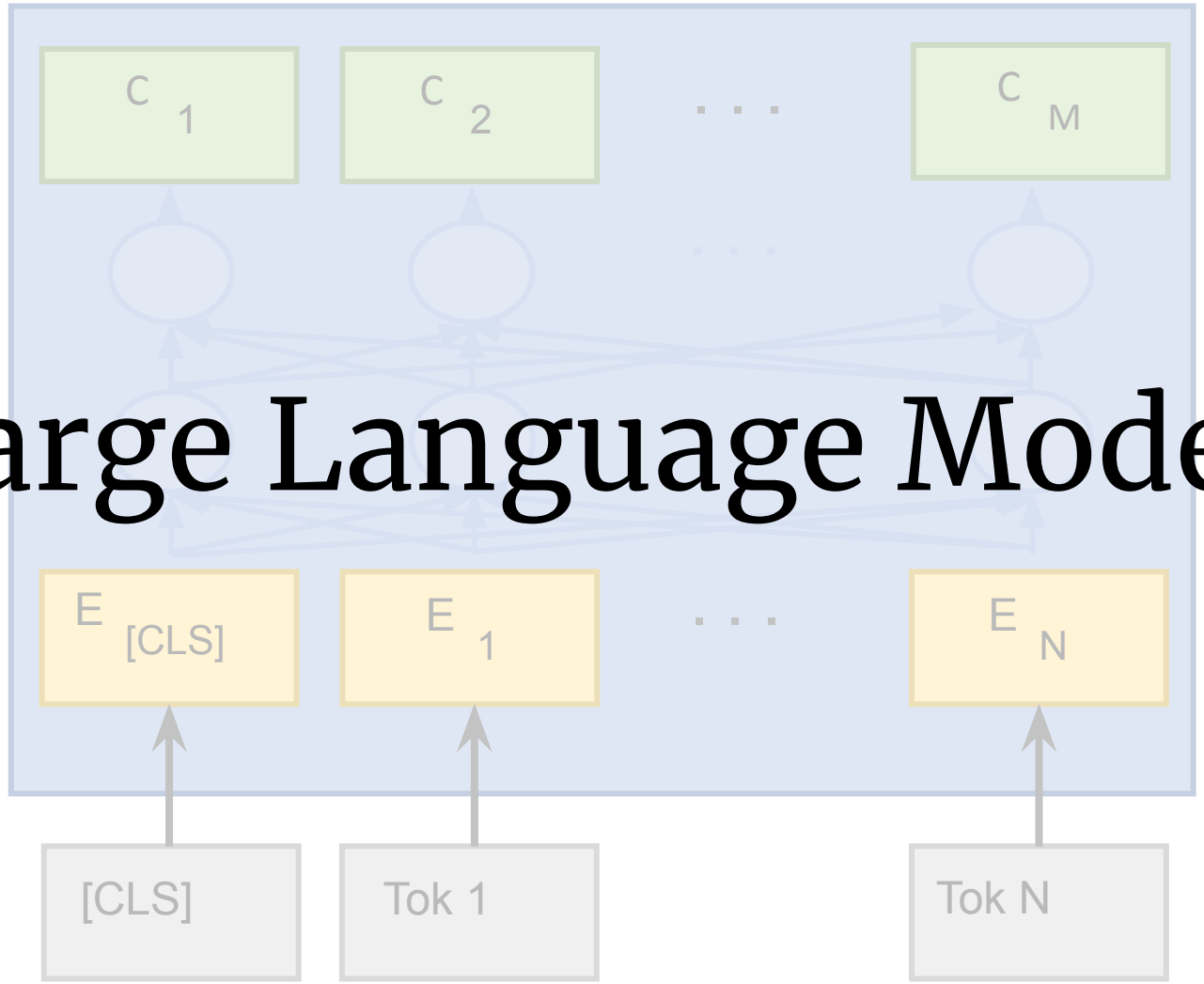
filter for confidence

Dataset

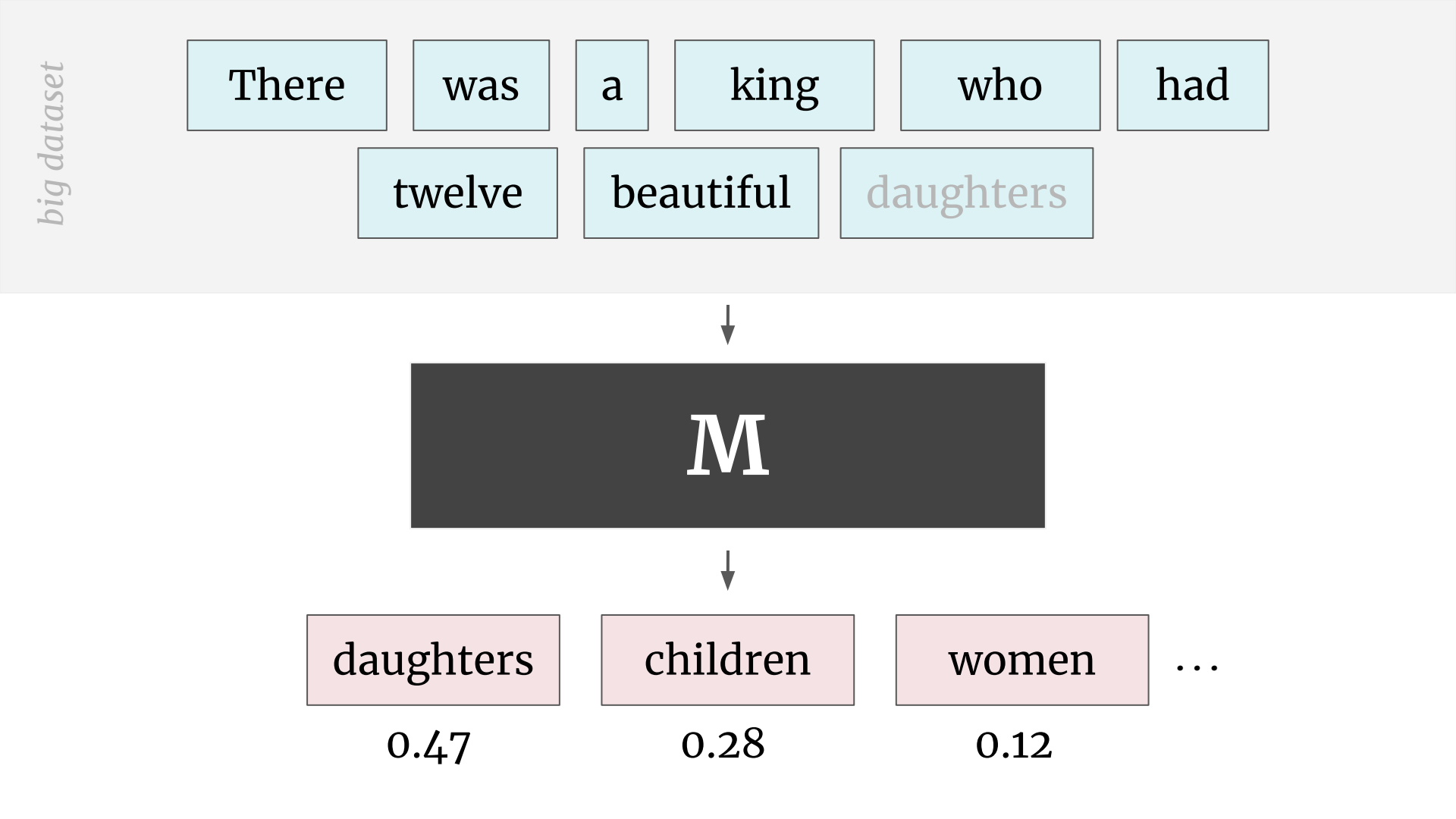
Protein family name	Knotted	Unknotted
ATCase/OTCase	2 255	16 998
AdoMet synthase	7 456	1 293
...	...	...
SPOUT	34 233	2 570
Sodium/calcium exchanger	22 432	3 524
TDD	3 005	156
UCH	1 785	620
ALL	**99 303**	**103 313**

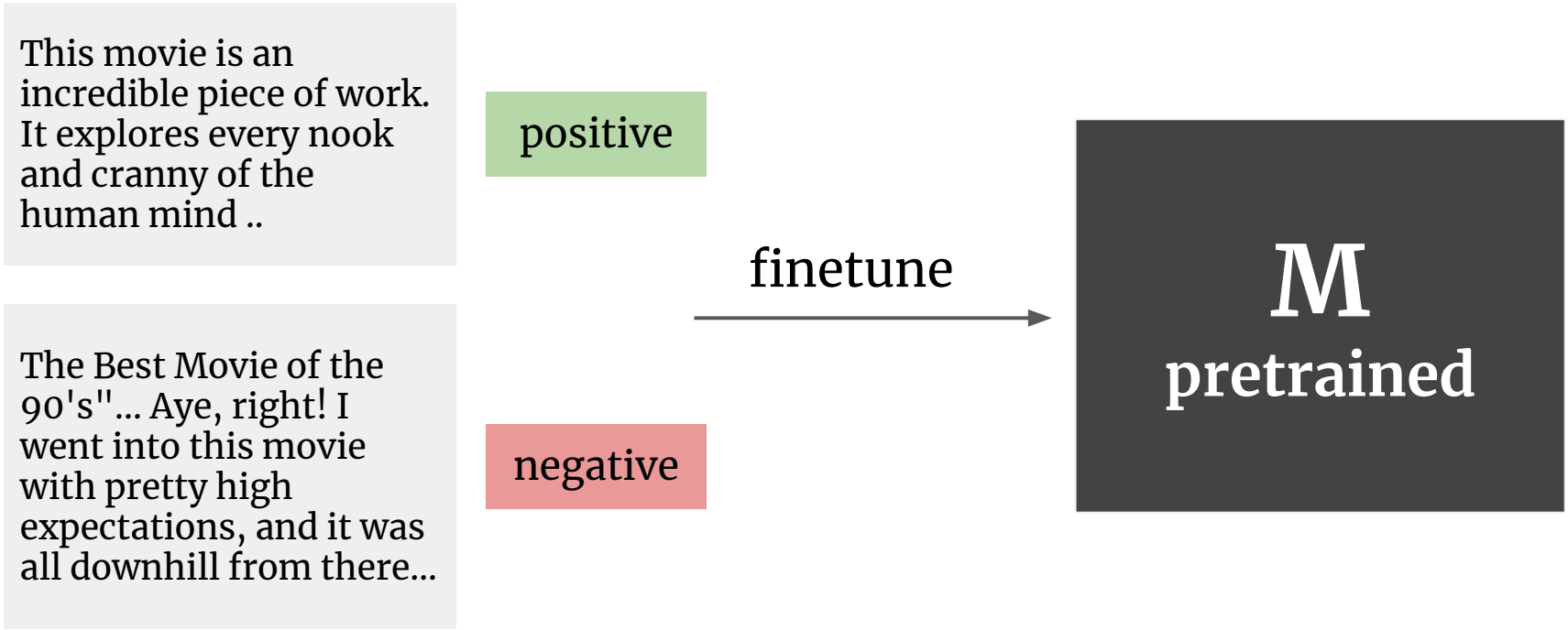
80 / 20 split training / testing


~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



Large Language Models


big dataset
There
was
a
king
who
had
twelve
beautiful
daughters
M
daughters
children
women
...
0.47
0.28
0.12

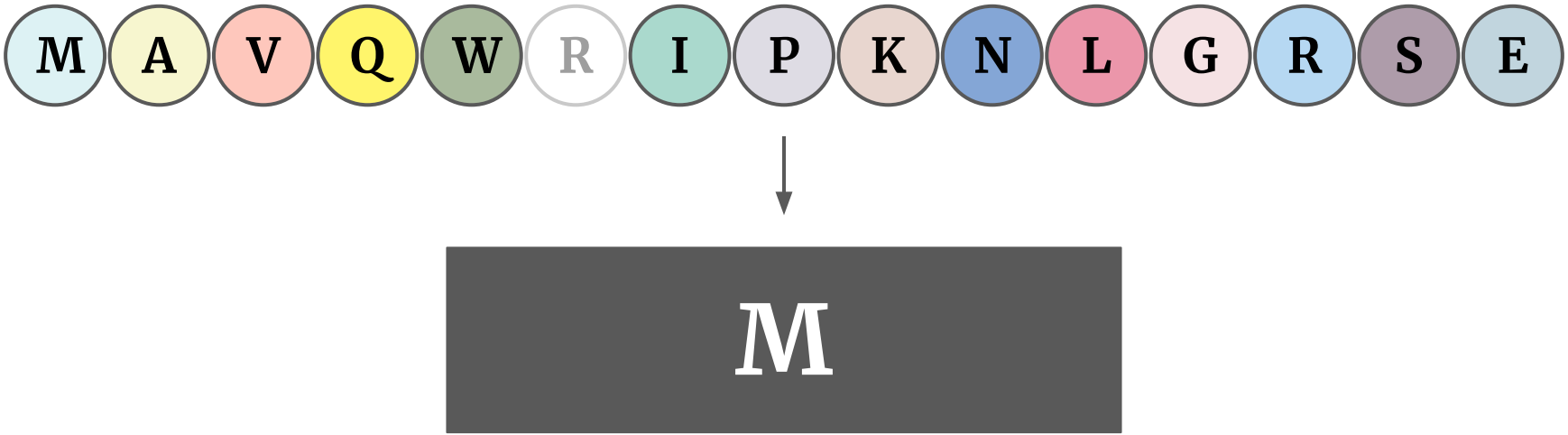

This movie is an incredible piece of work. It explores every nook and cranny of the human mind ..
positive
The Best Movie of the 90's"... Aye, right! I went into this movie with pretty high expectations, and it was all downhill from there...
negative
finetune
M
pretrained

protein ~ language

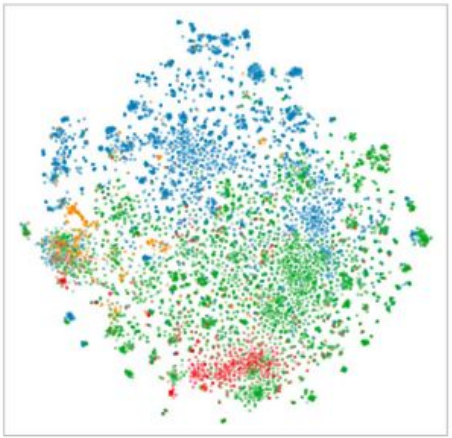
amino acid ~ word

protein ~ language

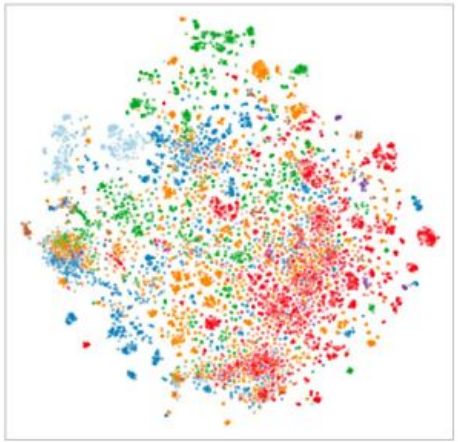
amino acid ~ word



t-SNE projection of ProtBERT embeddings

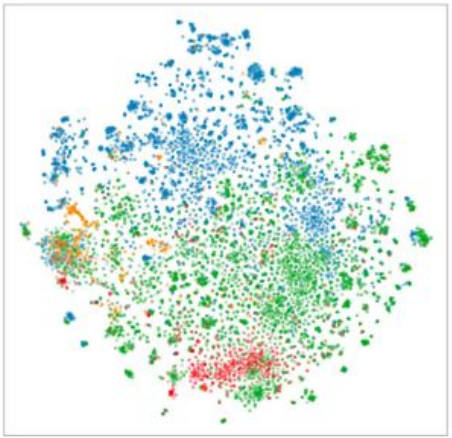


E Lineage: Kingdoms

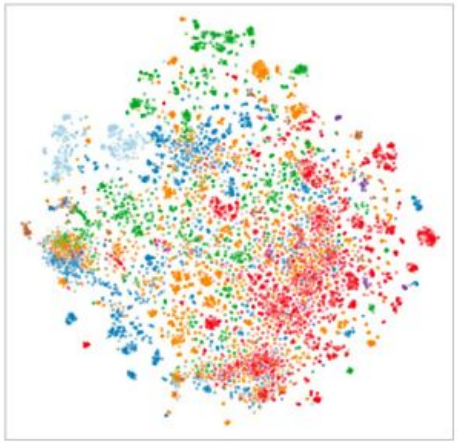


D Structure: SCOPe

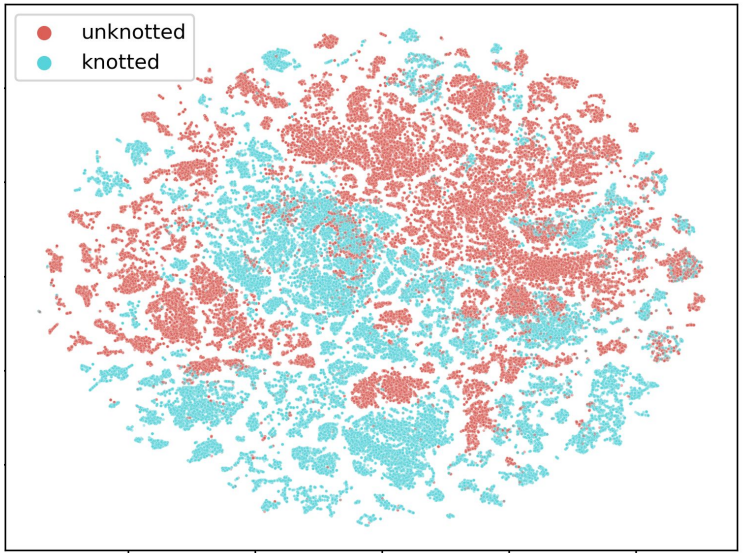
t-SNE projection of ProtBERT embeddings

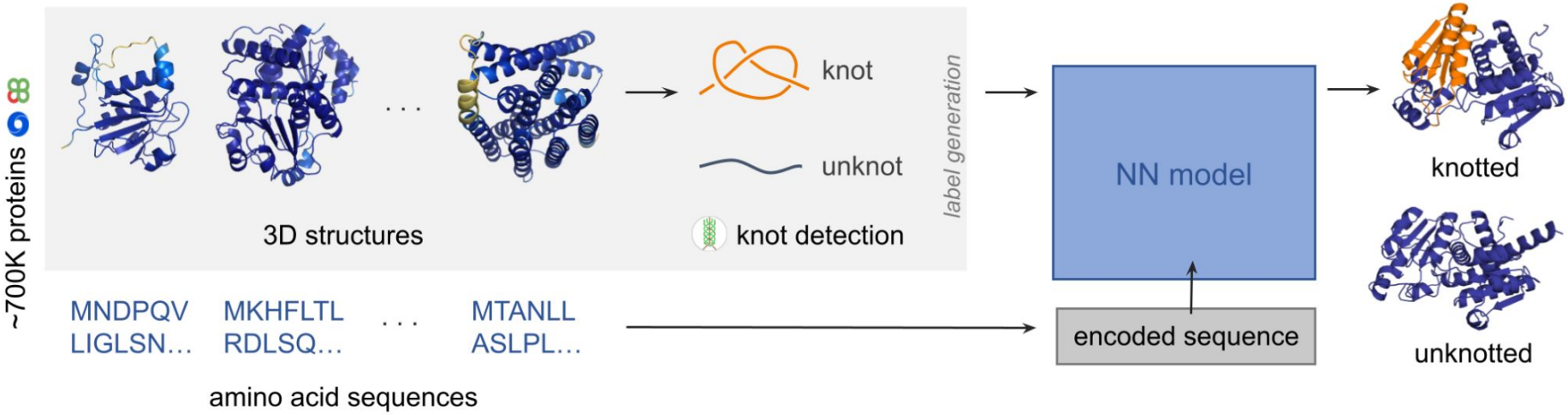


E Lineage: Kingdoms



D Structure: SCOPe

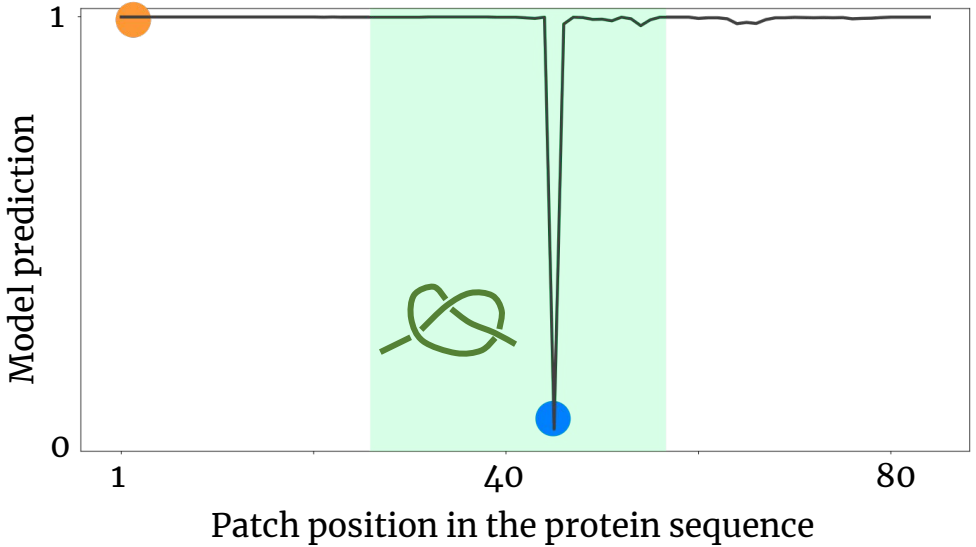


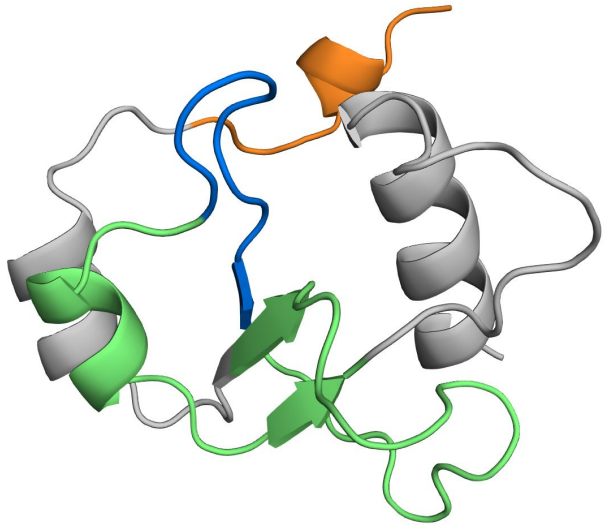

~700K proteins
3D structures
knot
unknot
knot detection
label generation
NN model
knotted
unknotted
MNDPQV LIGLSN…
MKHFLTL RDLSQ…
MTANLL ASLPL…
amino acid sequences
encoded sequence

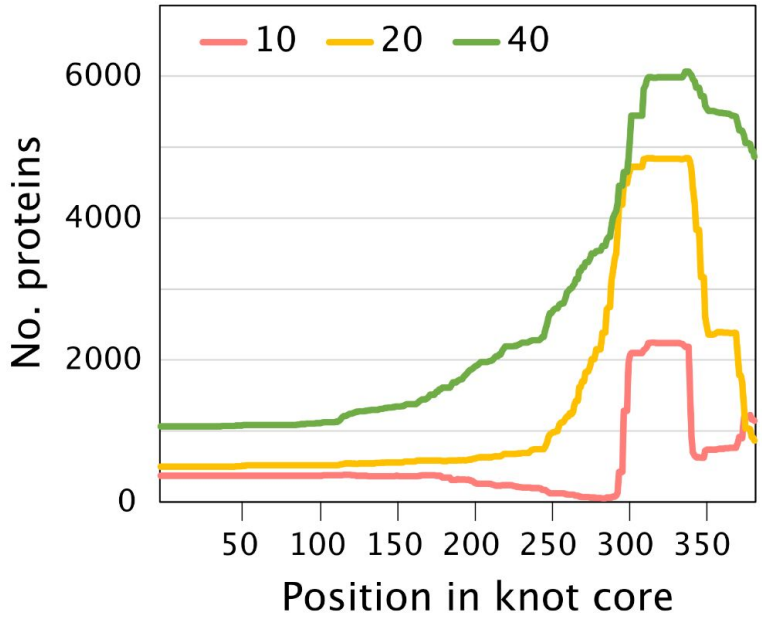
	Dataset size	Accuracy	TPR	TNR
Whole test dataset	39 412	**0.9845**	0.9865	0.9825
SPOUT	7 371	0.9887	0.9951	0.9090
TDD	612	0.9901	0.9965	0.8333
DUF	716	0.9748	0.9721	0.9766
AdoMet synthase	1 794	0.9899	0.9929	0.9708
Carbonic anhydrase	1531	0.9588	0.9737	0.9313
UCH	477	0.9056	0.9602	0.7520
ATCase/OTCase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

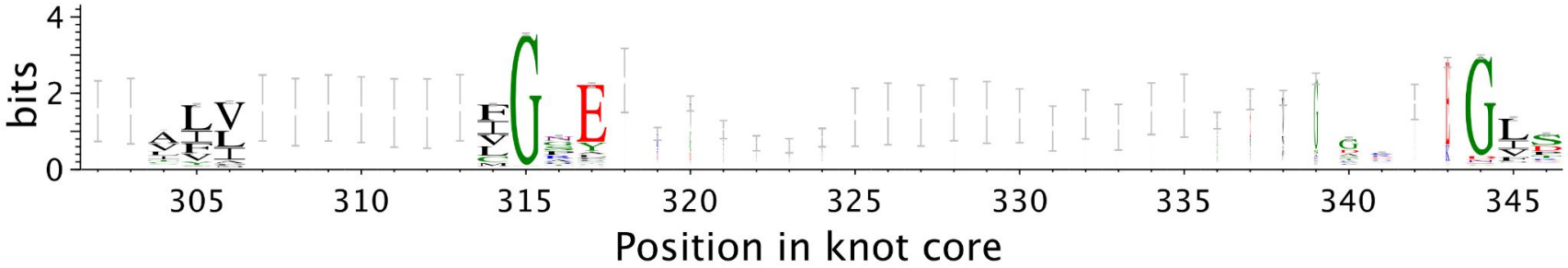
Interpretation

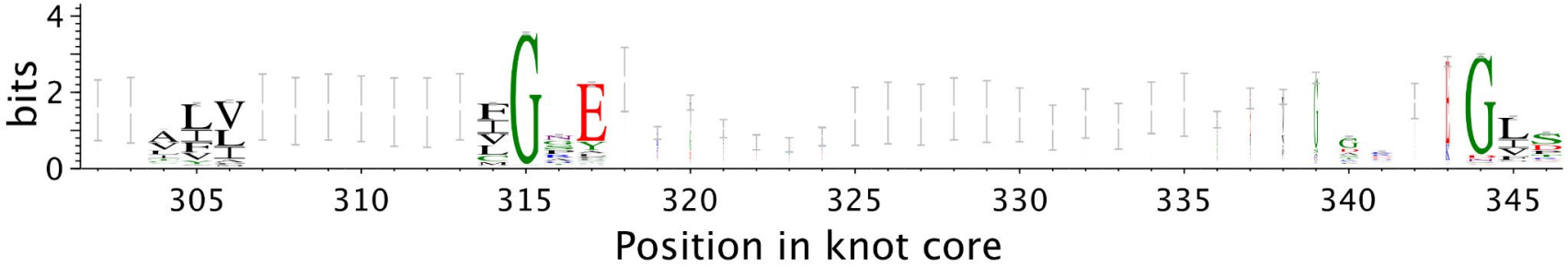
XXXXXXXXXXYTDLEPYLQSTKLPIYGALLDGENIYELVDKSKGILVIGNESKGIRSTIQNFIQKPITIPRIGQAESLNAAVATGIIVGQLTL
MXXXXXXXXXXTDLEPYLQSTKLPIYGALLDGENIYELVDKSKGILVIGNESKGIRSTIQNFIQKPITIPRIGQAESLNAAVATGIIVGQLTL
...
MGGIFRVNTYYTDLEPYLQSTKLPIYGALLDGENIYELVDKSKGXXXXXXXXXXIRSTIQNFIQKPITIPRIGQAESLNAAVATGIIVGQLTL
...
MGGIFRVNTYYTDLEPYLQSTKLPIYGALLDGENIYELVDKSKGILVIGNESKGIRSTIQNFIQKPITIPRIGQAESLNAAVAXXXXXXXXXX






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

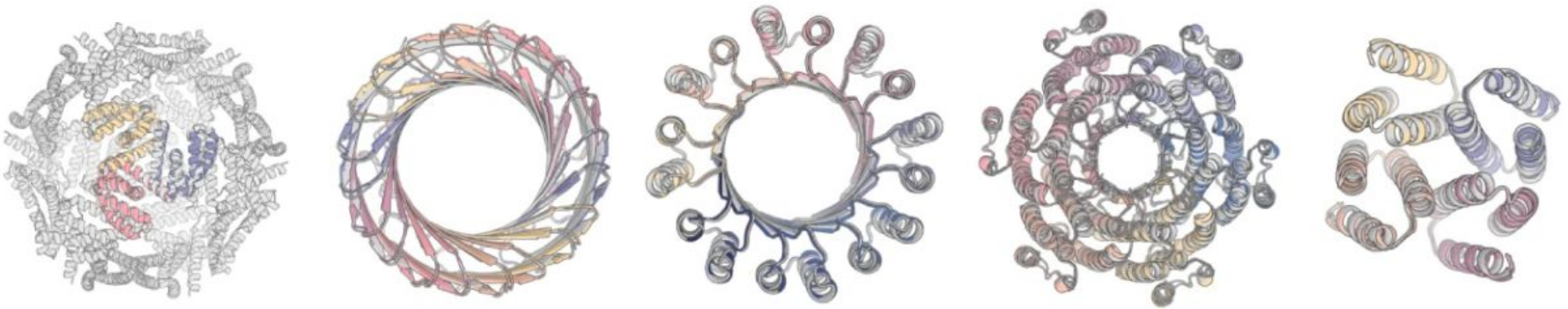

bits
4
2
0
305
310
315
320
325
330
335
340
345
Position in knot core



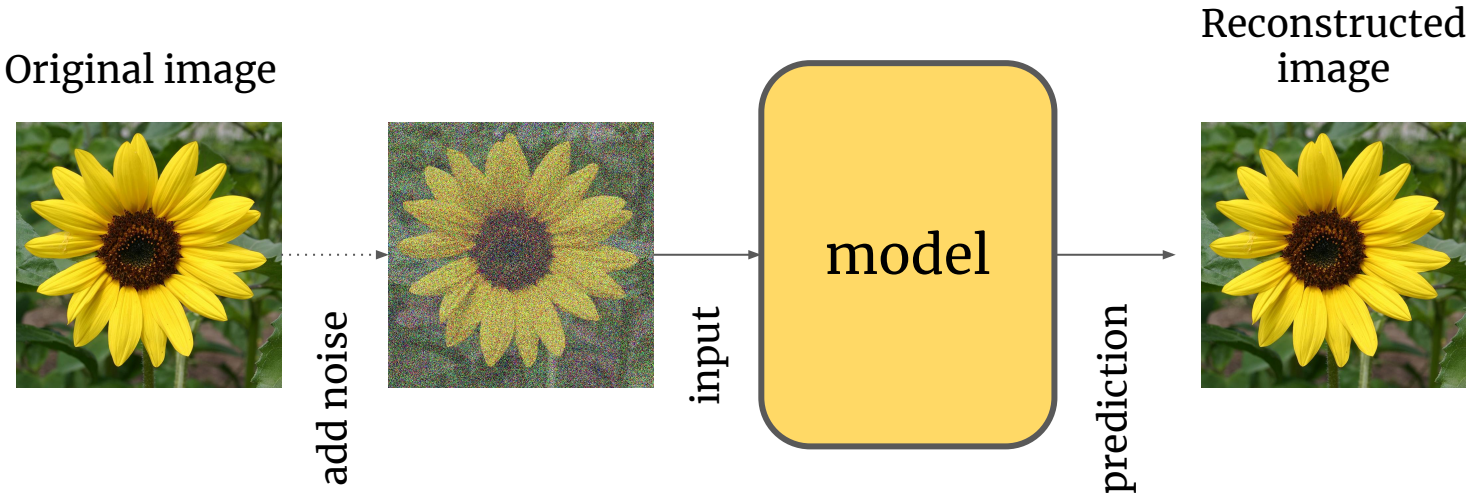
=

family functional site

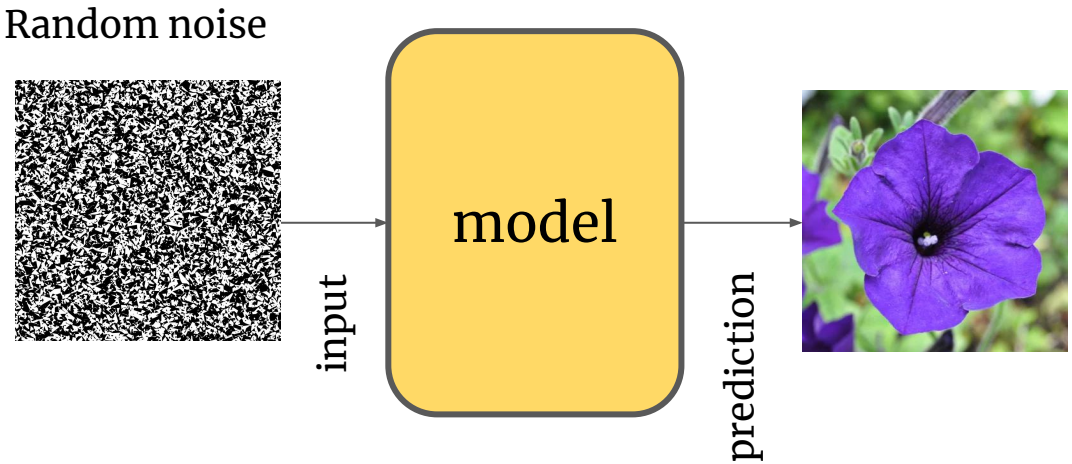
Designing new knots



Diffusion ~ adding and removing noise



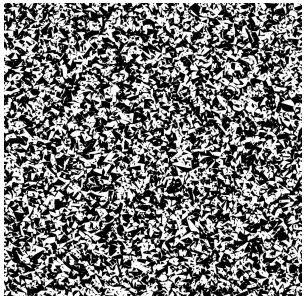
Using diffusion to generate new images



Generate images based on some description

+

curly girl with glasses wearing a sweater cuddling with a rat



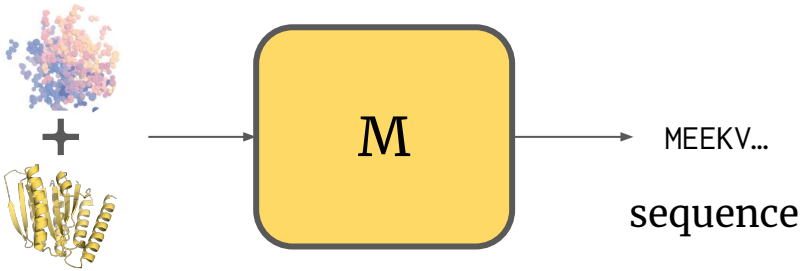
curly girl with glasses wearing a sweater cuddling with a rat

model

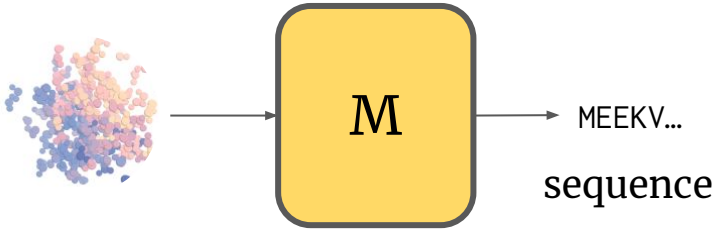
Diffusion for proteins

RFdiffusion

protein MPNN



EvoDiff

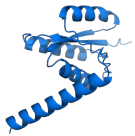


Creating new knots

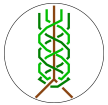
A

RFdiffusion + MPNN

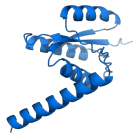
structure generation



knot status



sequence generation



MEEKV...
MKVYI...
MEERL...

structure verification

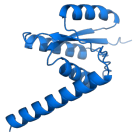
MEEKV...

B

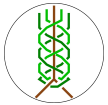
EvoDiff

MEEKVIEIR...

sequence generation



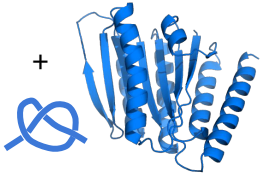
structure prediction



knot status

RFduffusion +MPNN

MEEKV... +



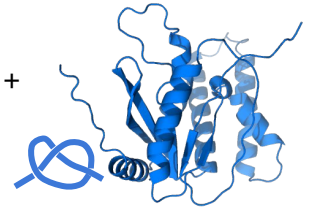
# of designed structures	... processing ...	3D structure with knot
~ 200k		1 037

~ 0.5 %

EvoDiff

MSTSPSGAD...

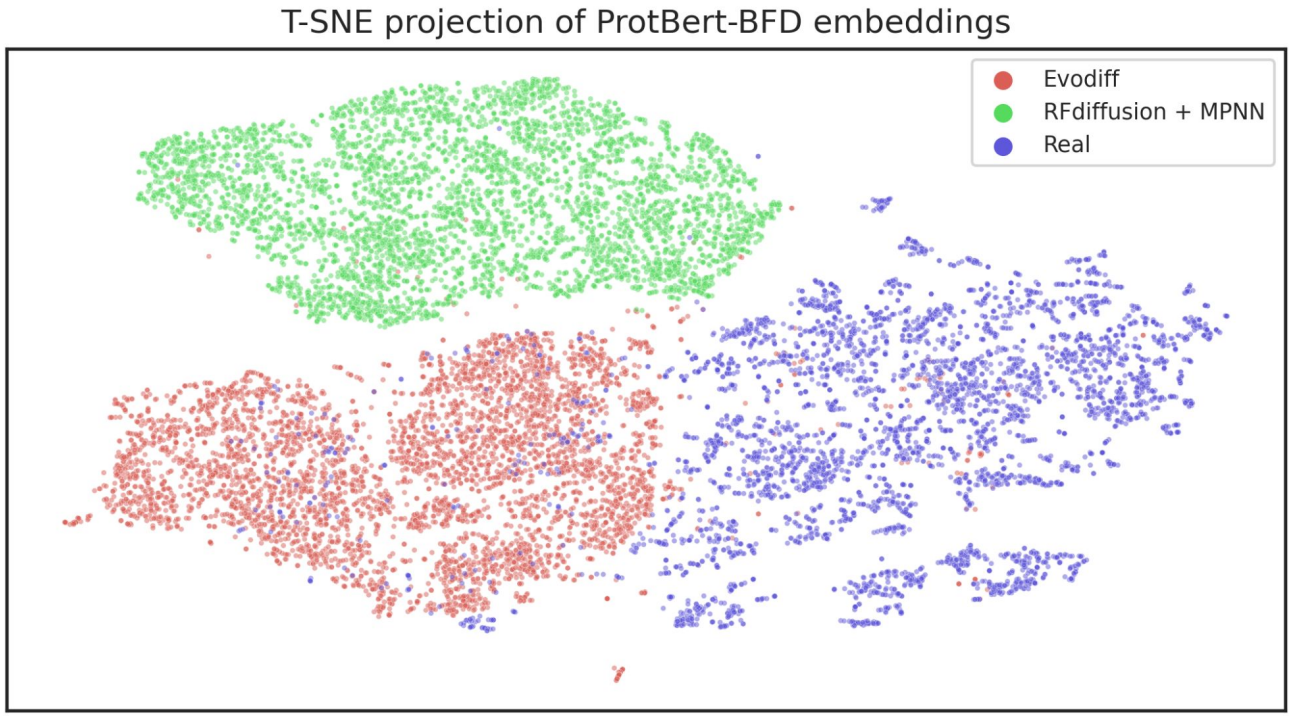
MSTSPS... +



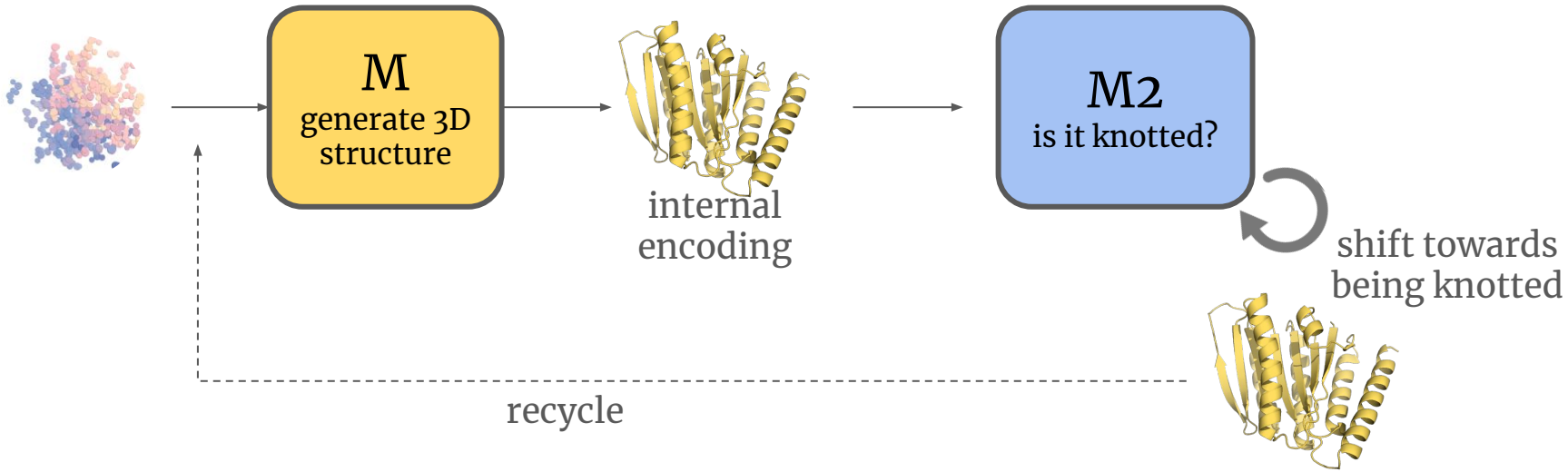
# of designed sequences	... processing ...	3D structure with knot
~ 200k		979

~ 0.5 %

Insights into generated proteins



Another idea how to generate knots



Thank you

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

Supported by the OPUS LAP program of the Grant Agency of Czech Republic
(204/07/1592, “Biological code of knots – identification of knotted patterns in biomolecules via AI approach”)
Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140)