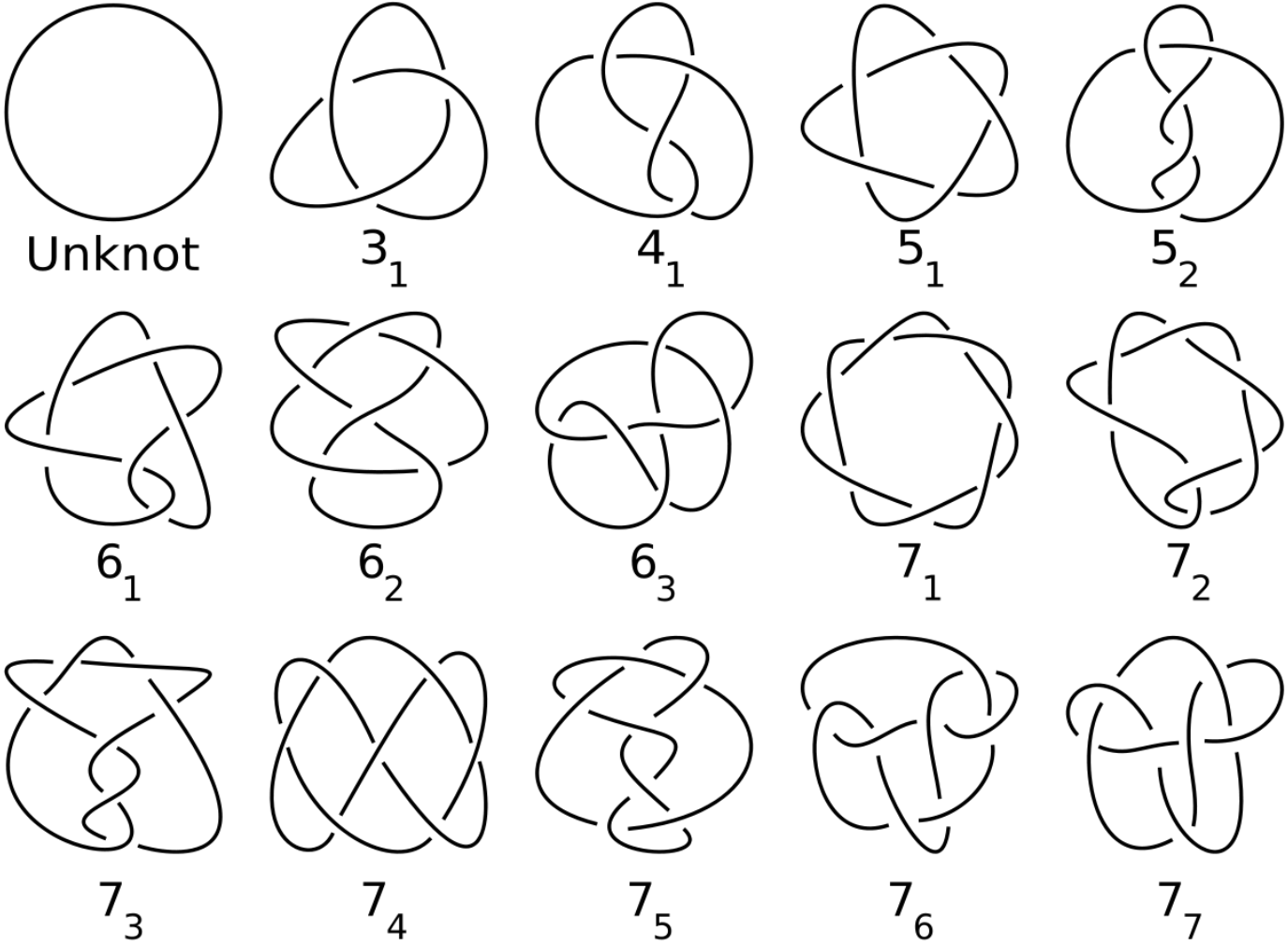
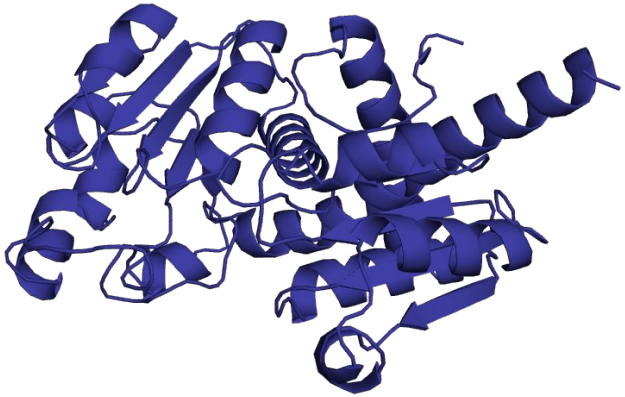
Artificial Knotted Proteins

Eva Klimentová
5. 4. 2024

...work in progress...

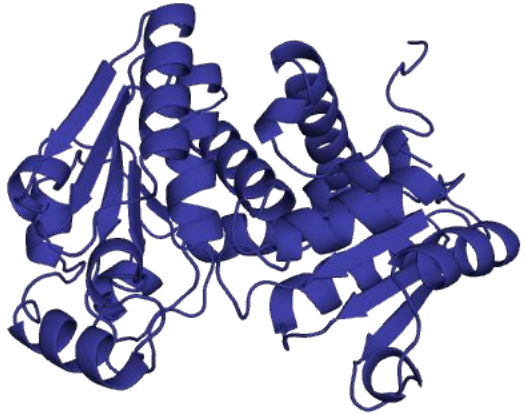
Knotted proteins

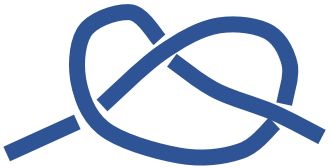

Unknot
3_1
4_1
5_1
5_2
6_1
6_2
6_3
7_1
7_2
7_3
7_4
7_5
7_6
7_7



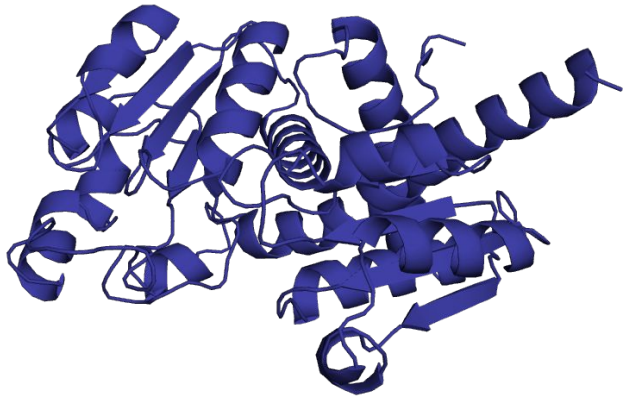
unknot

~ 99 % proteins



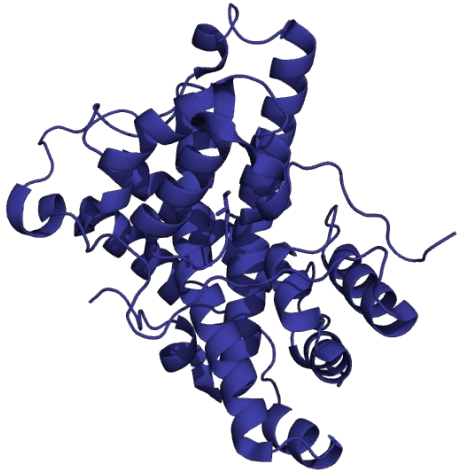


3_1 knot

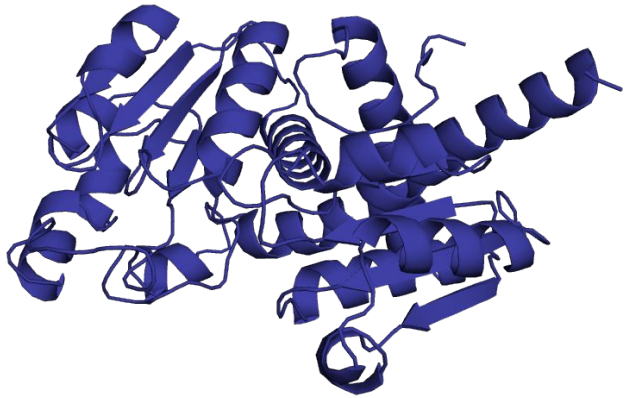


unknot

99 % proteins

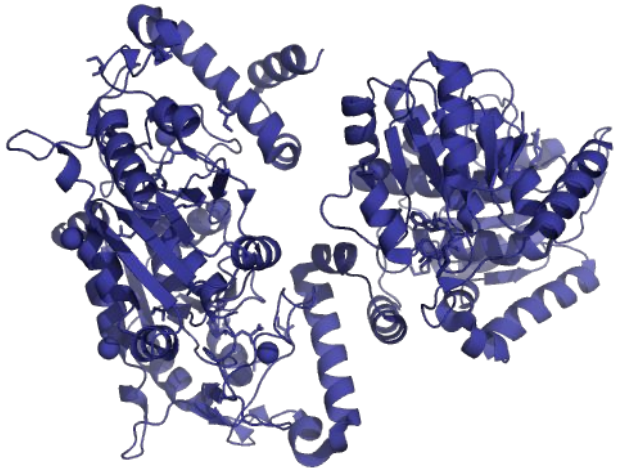


6_1 knot

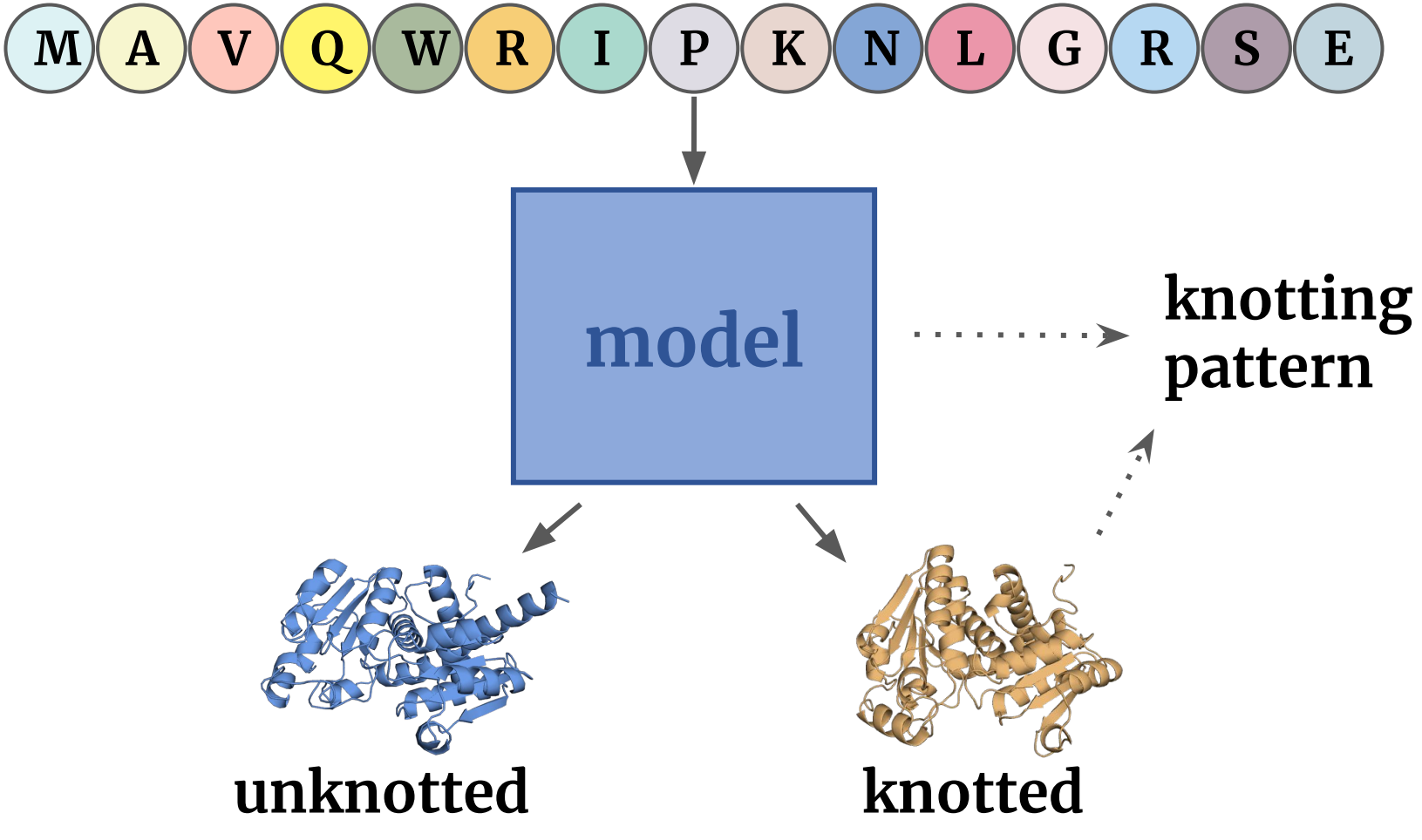


unknot

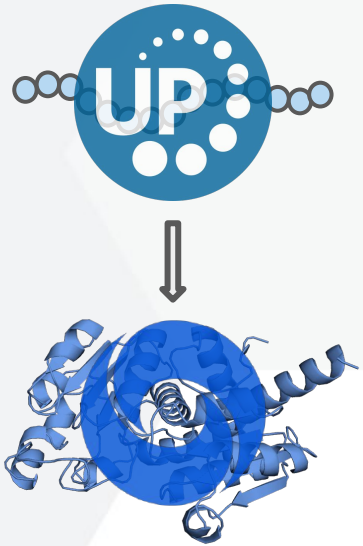
99 % proteins



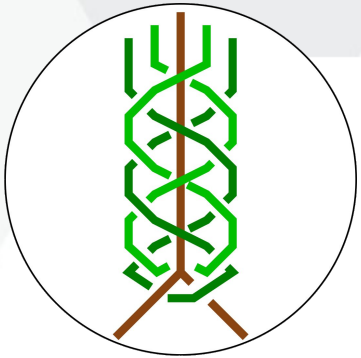
double 3_1 knot


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

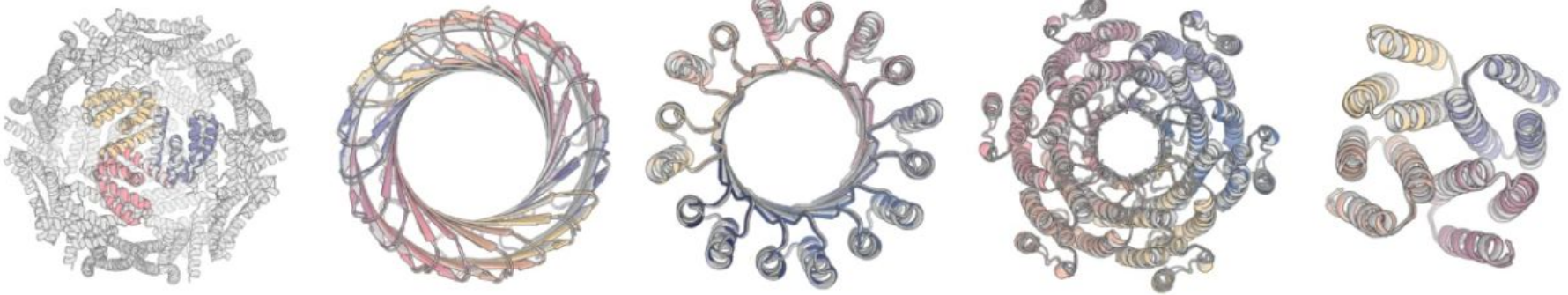
other ⋯→ trash

filter for confidence

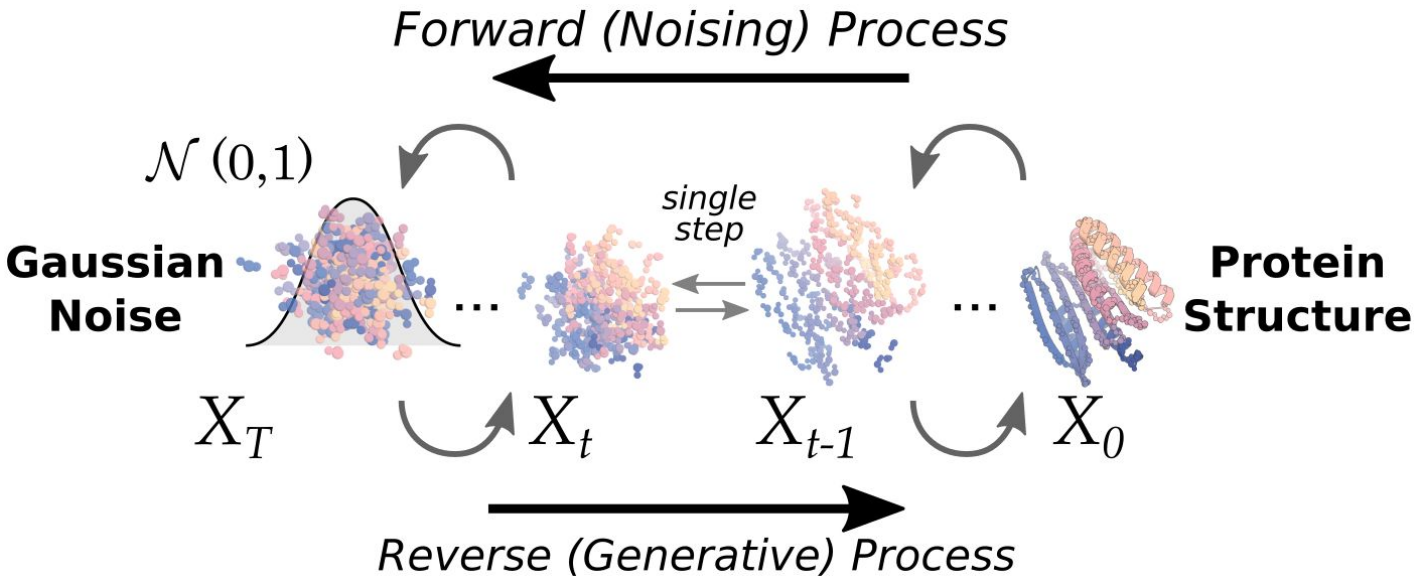
Dataset

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

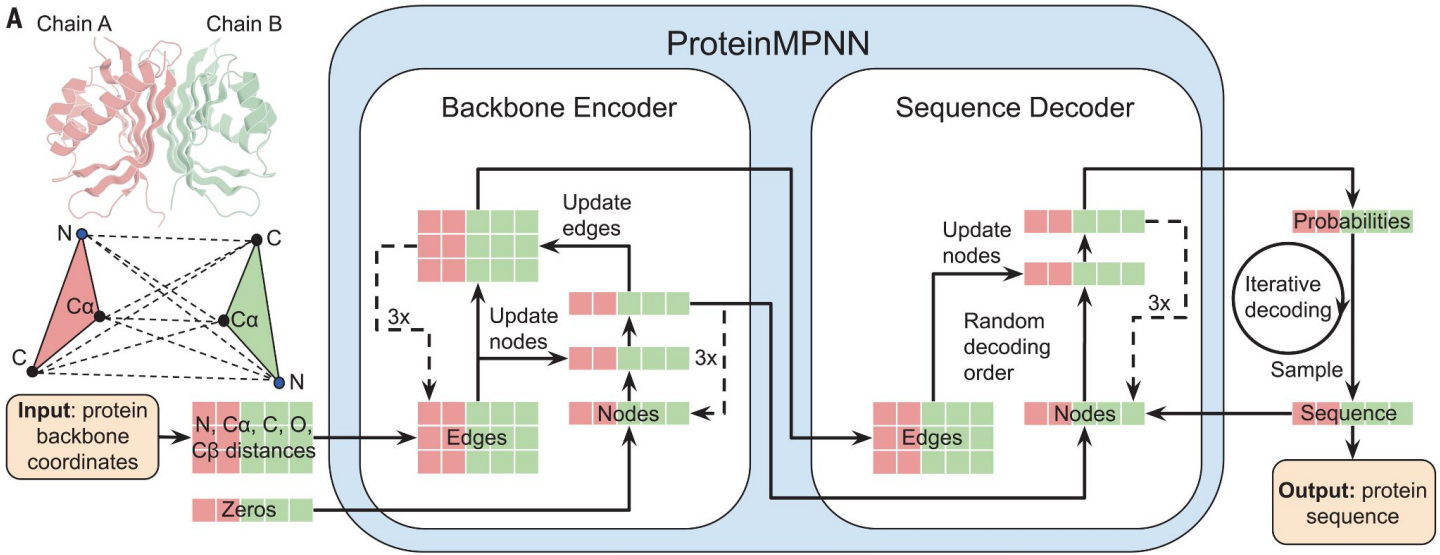


Diffusion for protein backbone: **RFdiffusion**



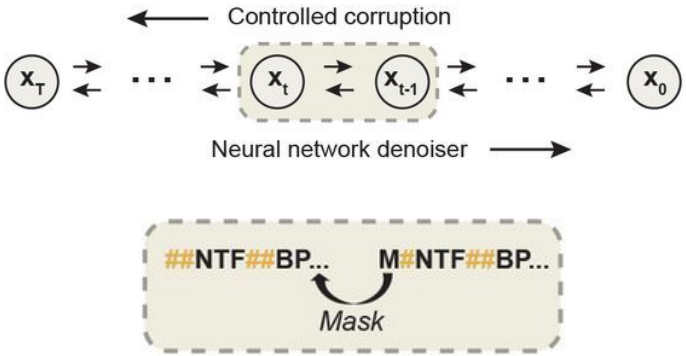
https://github.com/RosettaCommons/RFdiffusion
https://www.nature.com/articles/s41586-023-06415-8

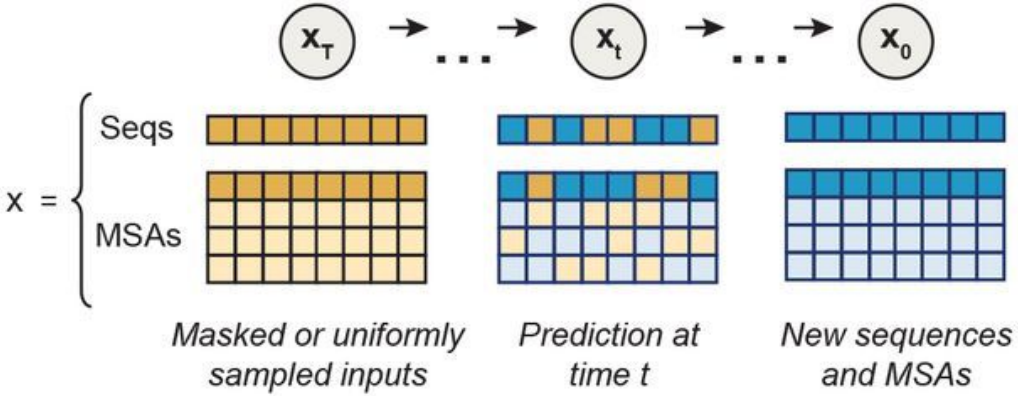
Diffusion for protein sequence based on structure: **ProteinMPNN**



https://github.com/dauparas/ProteinMPNN/tree/main
https://www.science.org/doi/10.1126/science.add2187
https://huggingface.co/spaces/simonduerr/ProteinMPNN

EvoDiff – directly generate sequence





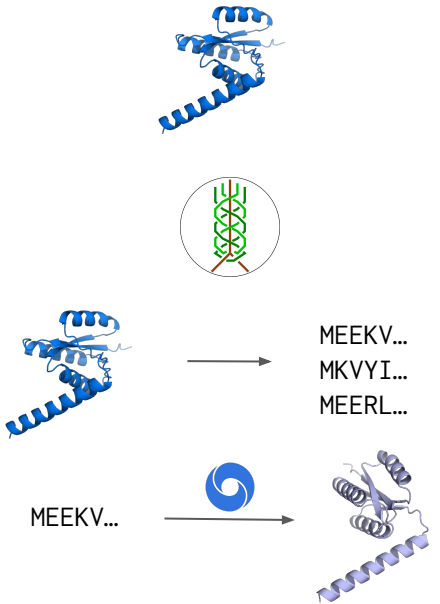
https://github.com/microsoft/evodiff/tree/main

Kevin Kaichuang Yang

Generating random proteins with RFdiffusion + MPNN

Workflow:

1. design random structure with RFdiffusion
2. check it’s topology
3. design multiple sequences with MPNN based on structure
4. predict structure from sequences
    - check pLDDT
    - check topology again

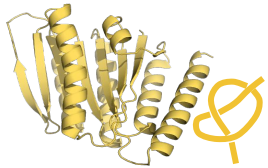


Generating random proteins with EvoDiff

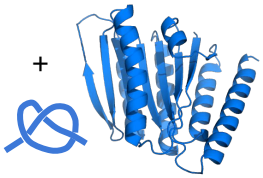
Workflow:

1. design random protein sequence with EvoDiff
2. predict 3D structure with OmegaFold
3. check topology of predicted structure

RFduffusion +MPNN



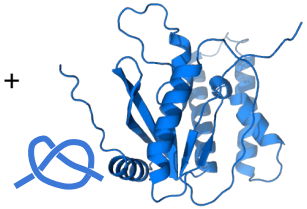
MEEKV... +



1. # of designed structures	2. Non-trivial topology in designed structure	4. Successful design of sequence with non-trivial topology
212 681	2 814	1 037

~ 0.5 %

MSTSPS... +



EvoDiff

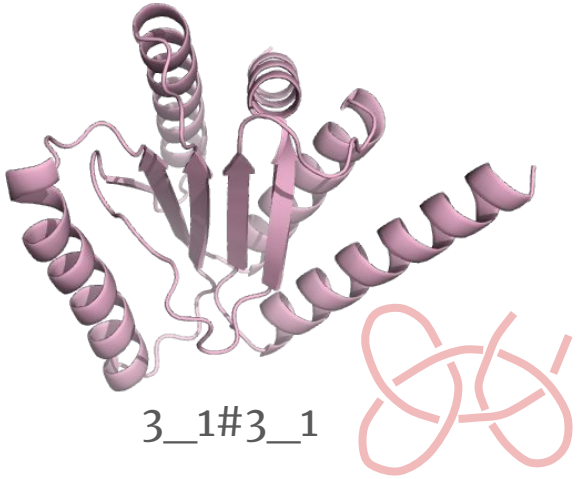
MSTSPSGAD...

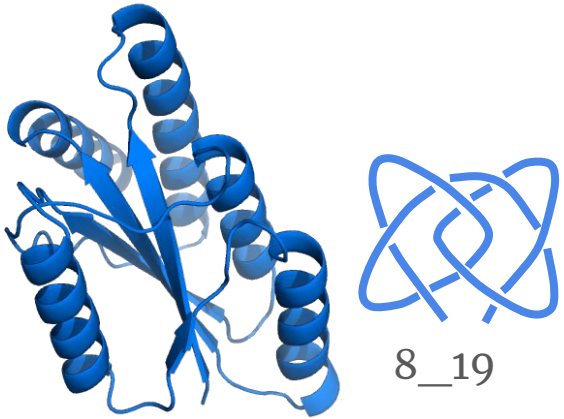
1. # of designed sequences		3. 3D structure of sequence with non-trivial topology
212 681		979

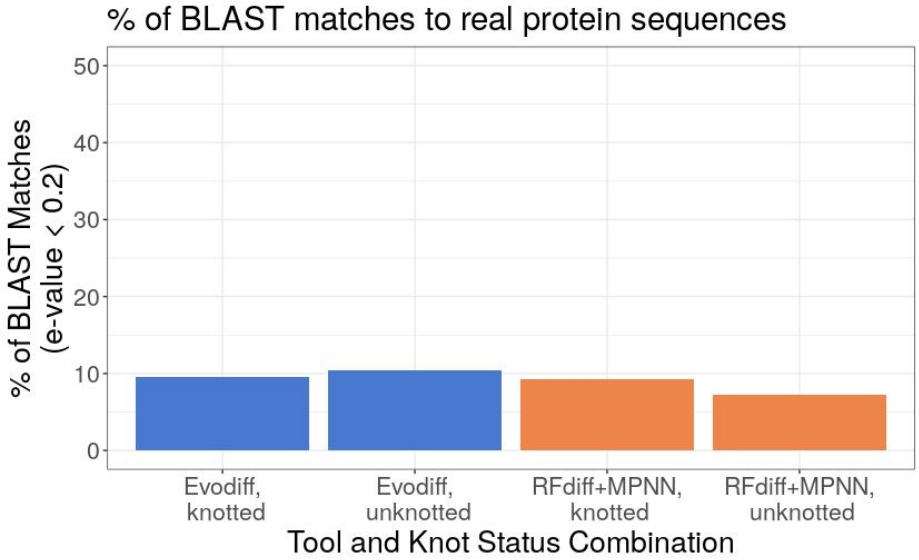
~ 0.5 %

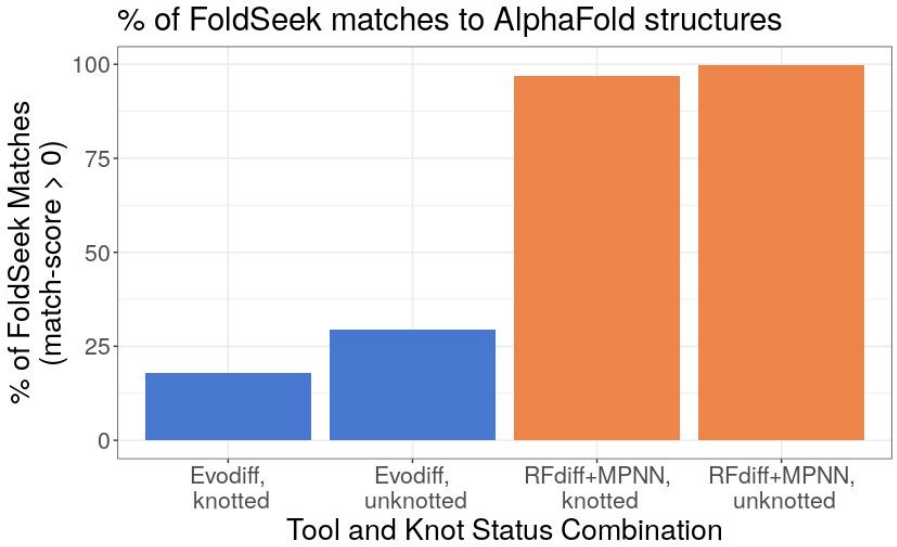
Generated knotted proteins

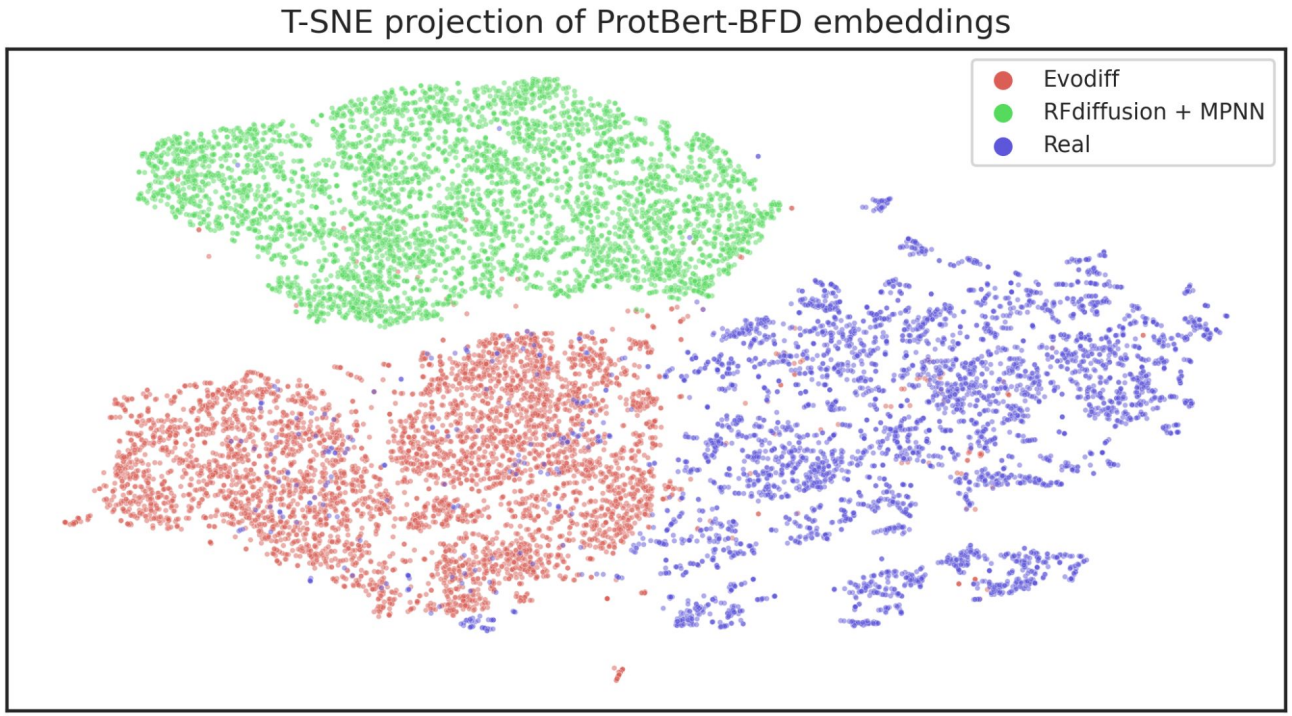
Major knot type	# of RF+MPNN designs	# of EvoDiff designs
3_1	950	866
3_1#3_1	5	2
4_1	7	44
5_1	26	15
5_2	1	11
8_19	3	0
and others		

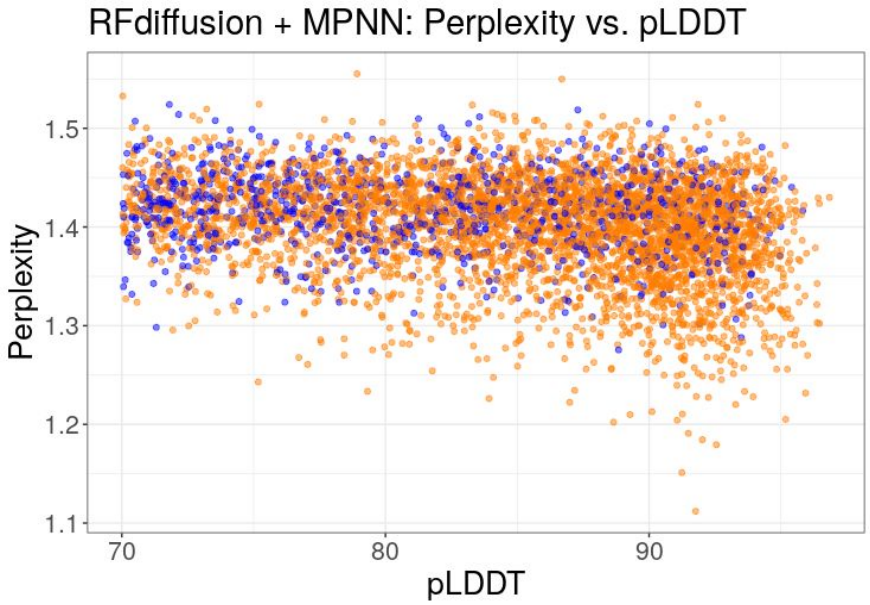


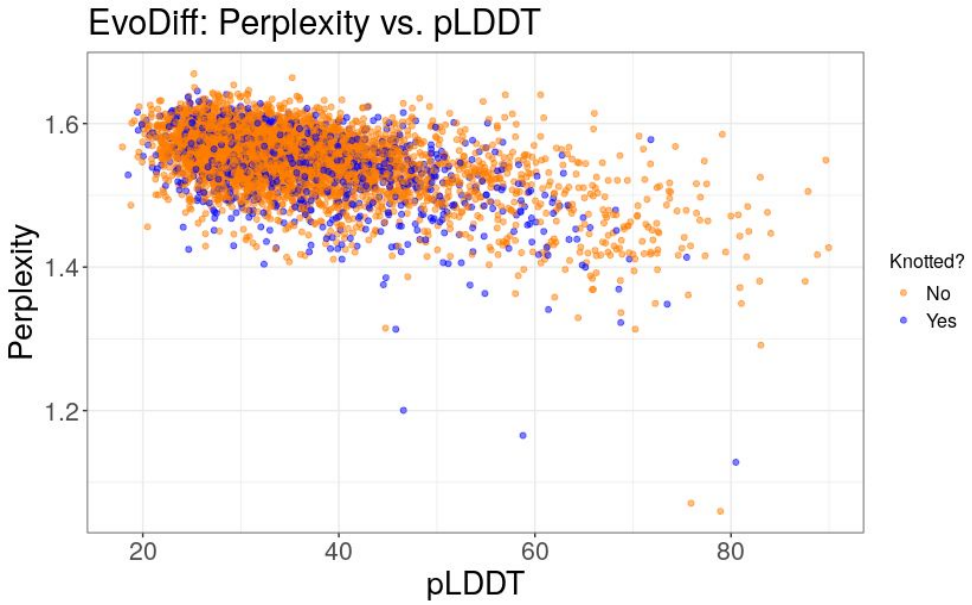



% of BLAST matches to real protein sequences
% of BLAST Matches
(e-value < 0.2)
50
40
30
20
10
0
Evodiff,
knotted
Evodiff,
unknotted
RFdiff+MPNN,
knotted
RFdiff+MPNN,
unknotted
Tool and Knot Status Combination


% of FoldSeek matches to AlphaFold structures
% of FoldSeek Matches
(match-score > 0)
100
75
50
25
0
Evodiff,
knotted
Evodiff,
unknotted
RFdiff+MPNN,
knotted
RFdiff+MPNN,
unknotted
Tool and Knot Status Combination


T-SNE projection of ProtBert-BFD embeddings
Evodiff
RFdiffusion + MPNN
Real


RFdiffusion + MPNN: Perplexity vs. pLDDT
Perplexity
1.5
1.4
1.3
1.2
1.1
70
80
90
pLDDT


EvoDiff: Perplexity vs. pLDDT
Perplexity
1.6
1.4
1.2
20
40
60
80
pLDDT
Knotted?
No
Yes

RFdiffusion + pMPNN knotted proteins

ID	length	pLDDT	pMPNN likelihood	knot size	solubility	visual inspection


Questions?
Ideas?