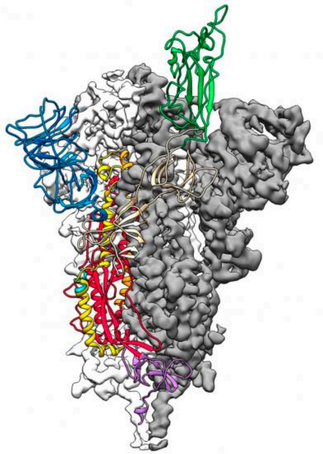
Multimodal deep learning for protein engineering

Kevin Kaichuang Yang
Microsoft Research New England
@KevinKaichuang

Proteins are biology’s actuators

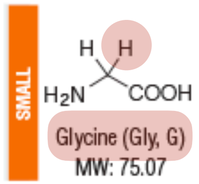
- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling

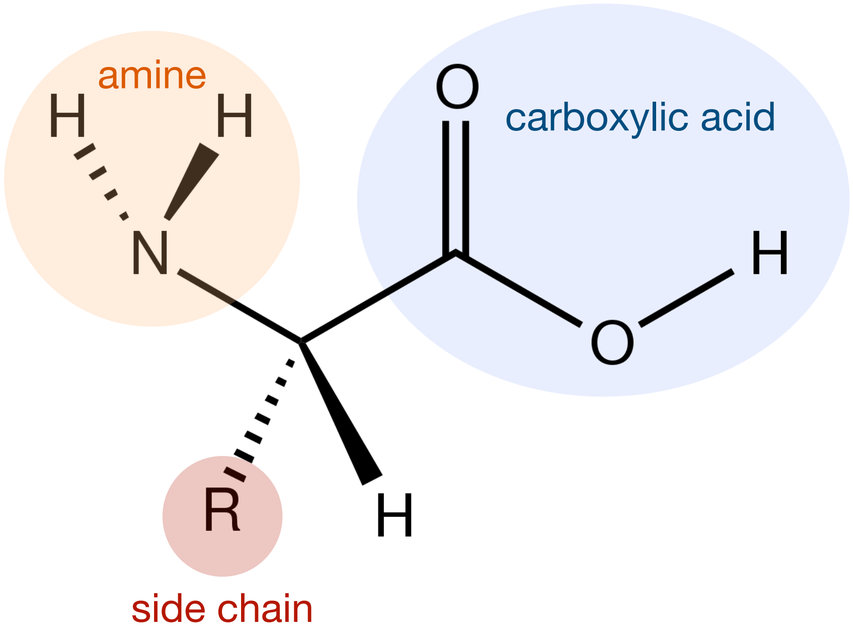


coronavirus spike protein

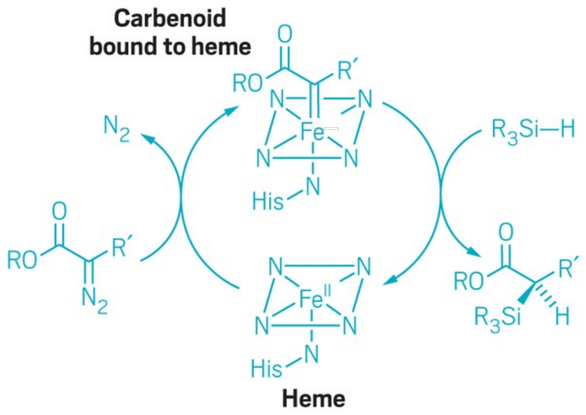
luciferase

Diversity arises from 20 building blocks

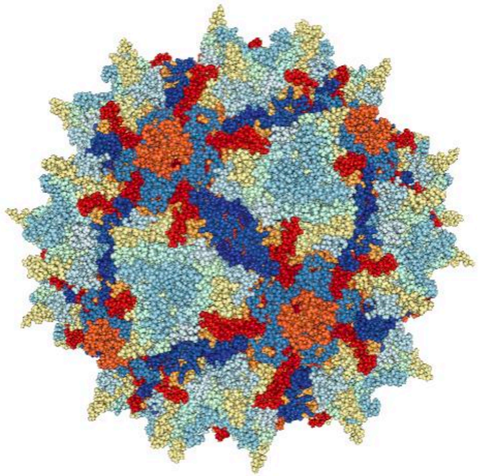




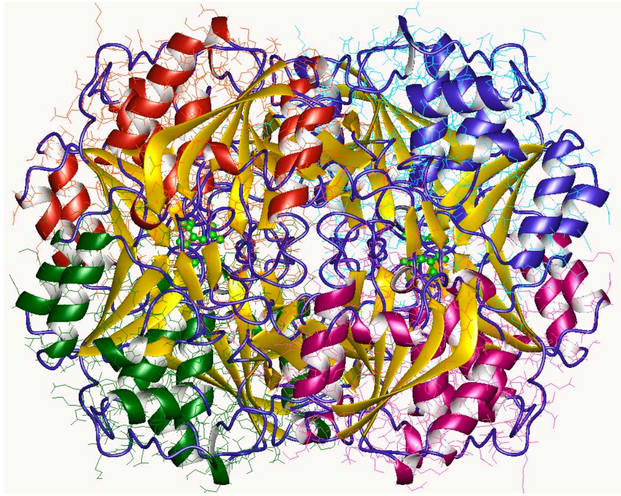
Why design proteins?



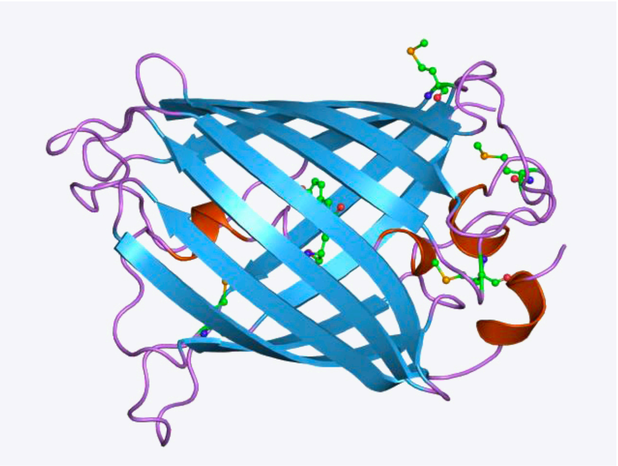
new chemistry



understand function

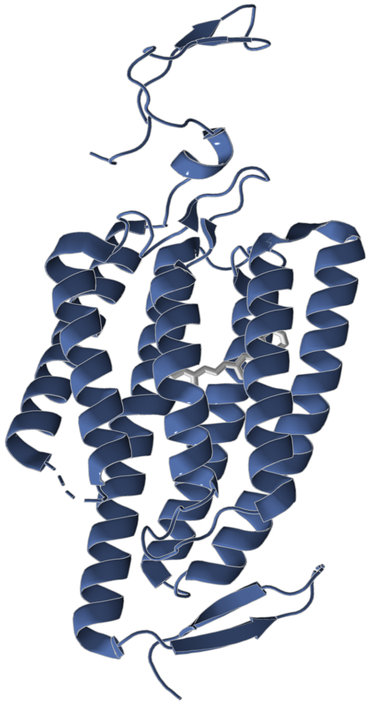


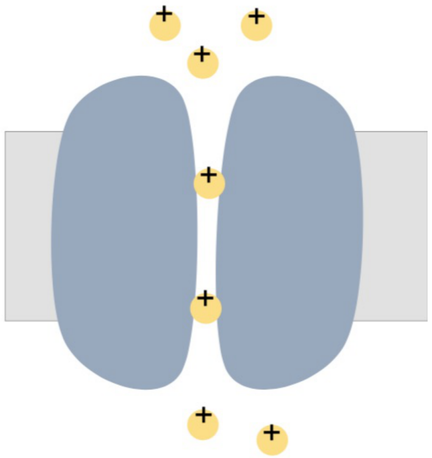
therapeutics



molecular tools

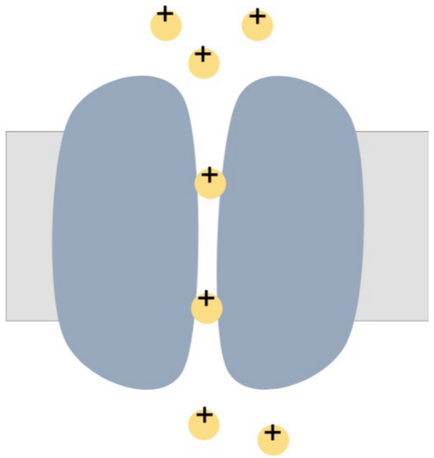
The protein engineering problem

MGTGDHDD... ?  ??



What sequence will give the desired function?

The protein engineering problem

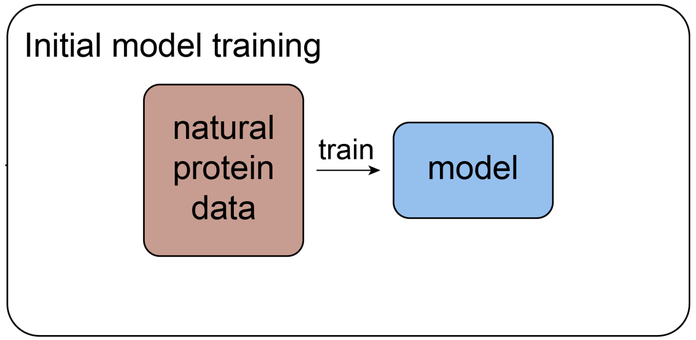


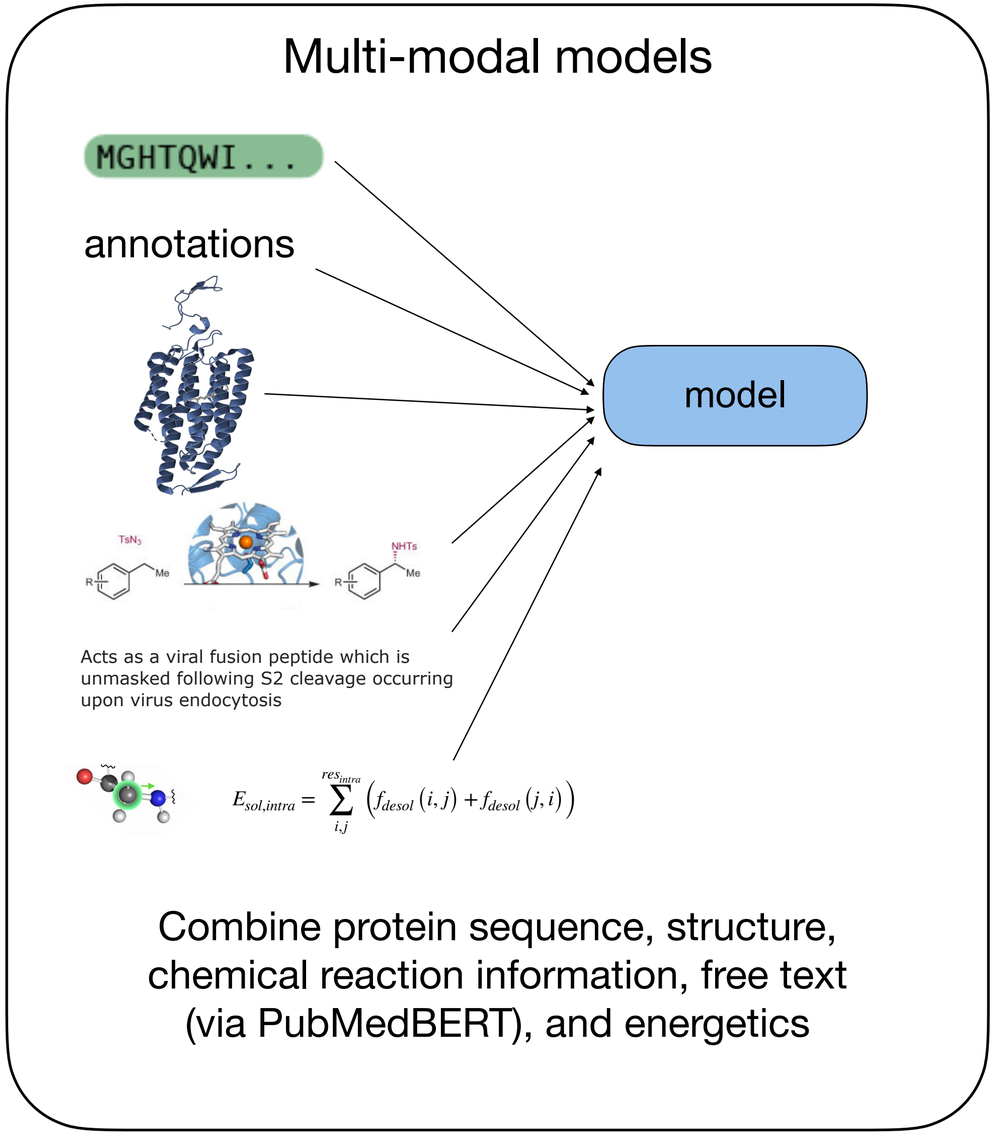
???

MGTGDHDD...

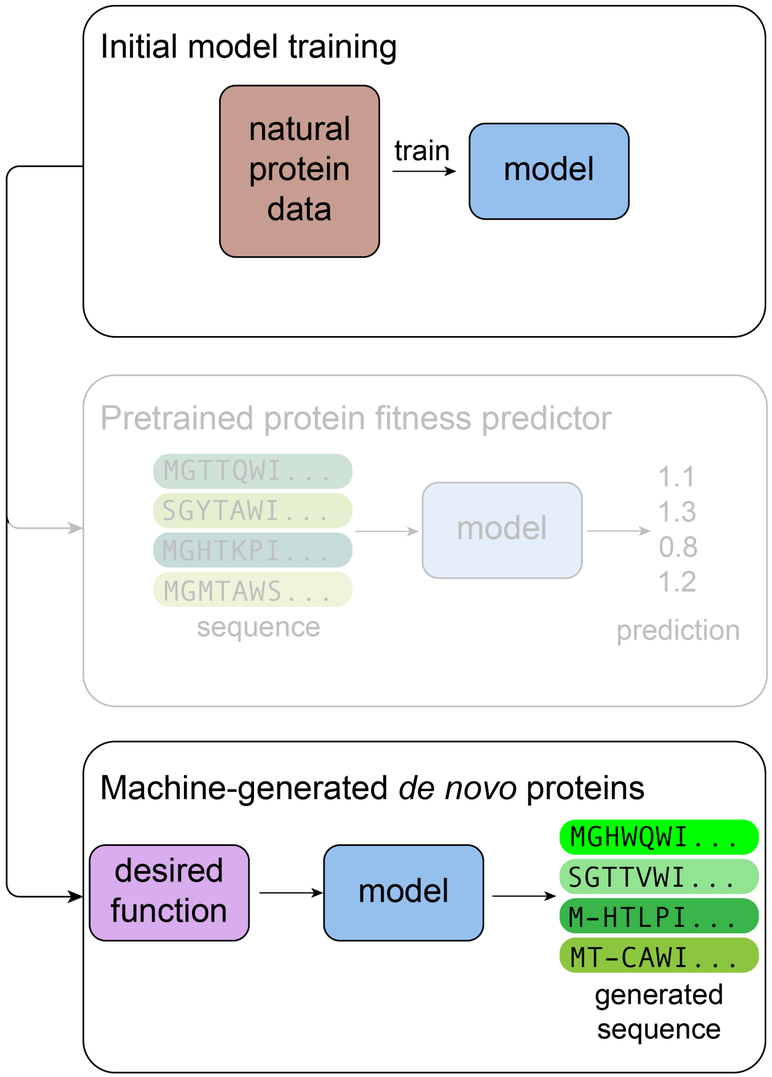
What sequence will give the desired function?

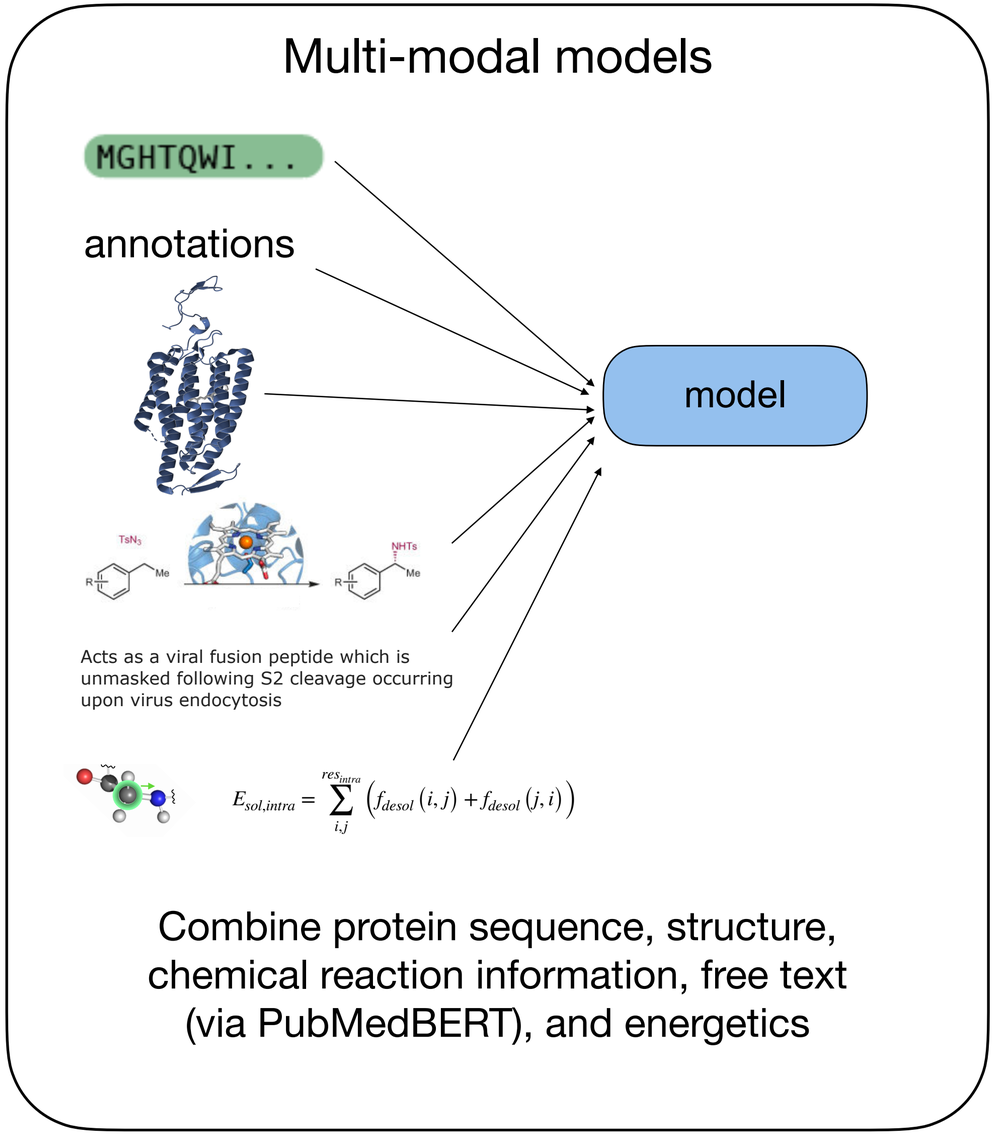
Use multiple data modalities to discover and design proteins



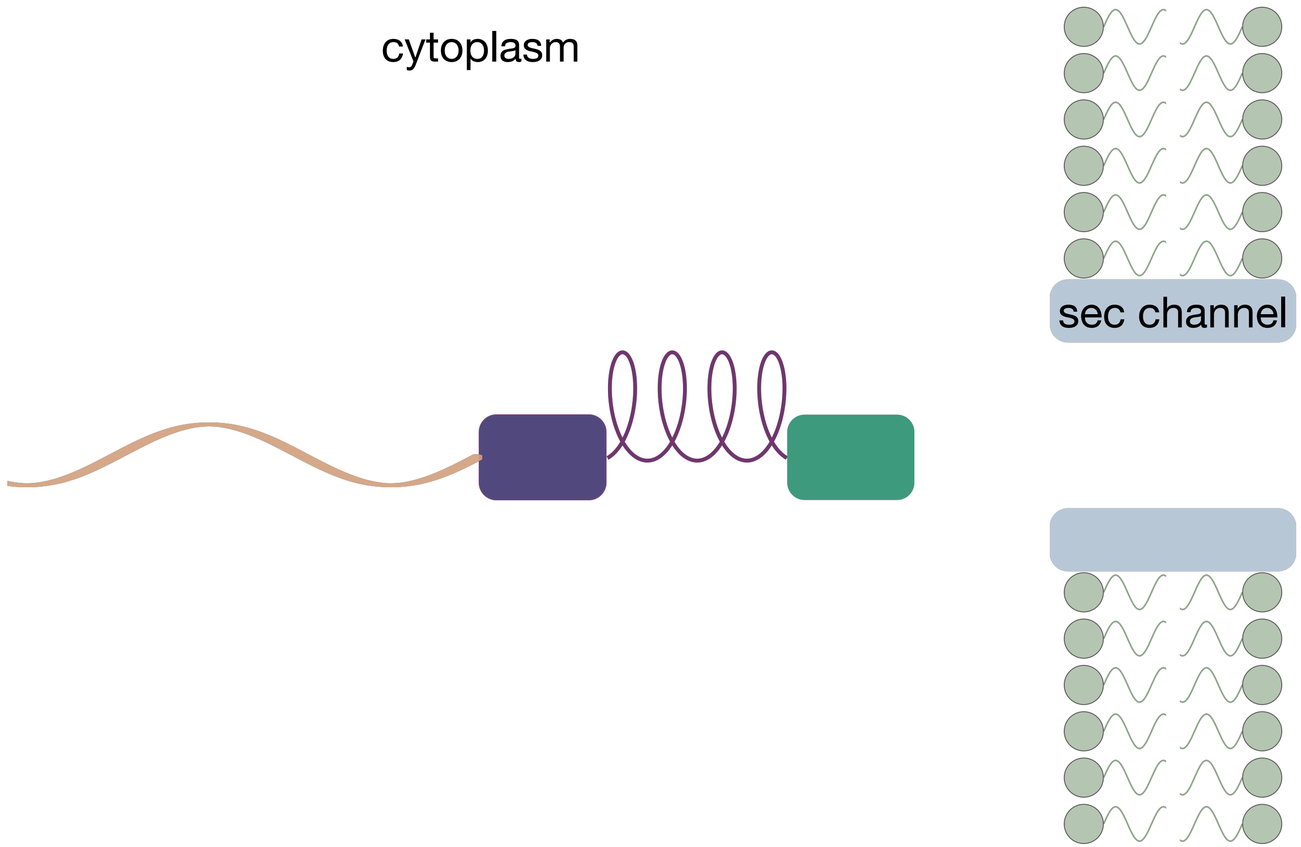


Use multiple data modalities to discover proteins

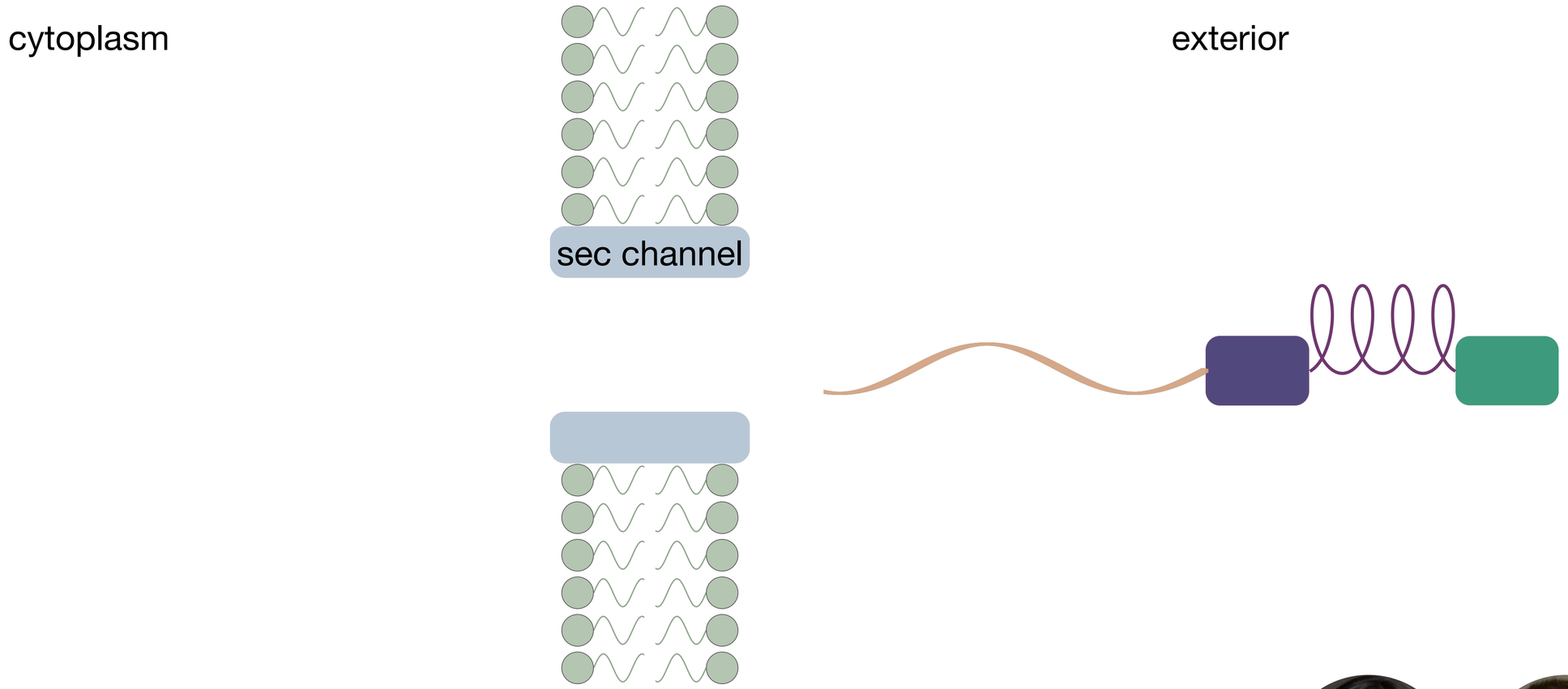




Signal peptides are secretion signals



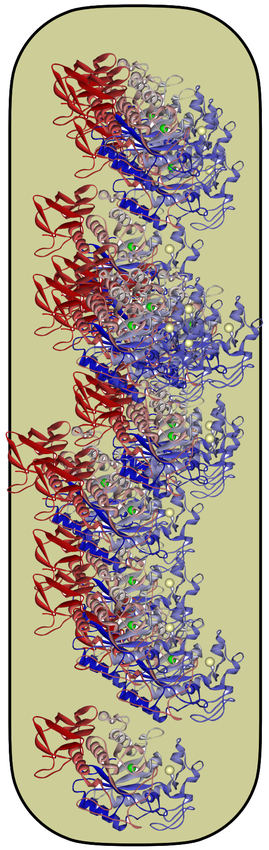
Signal peptides are secretion signals



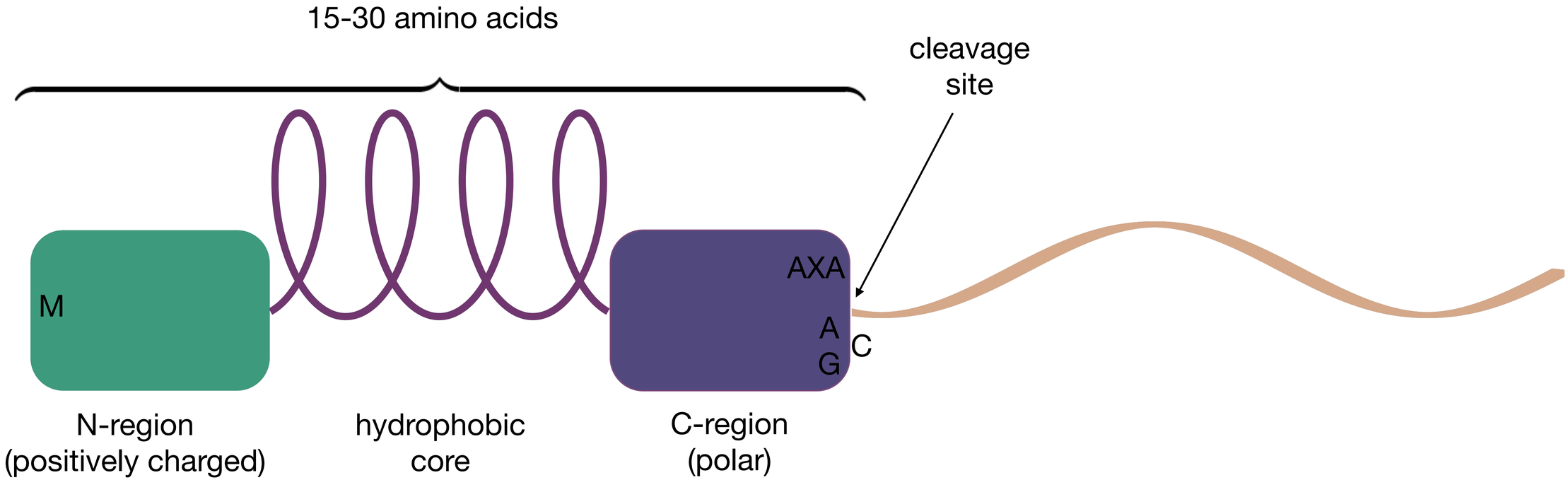
Need: custom SPs for arbitrary proteins


Signal peptides simplify protein production

- Less burden on cells
- Easier separation

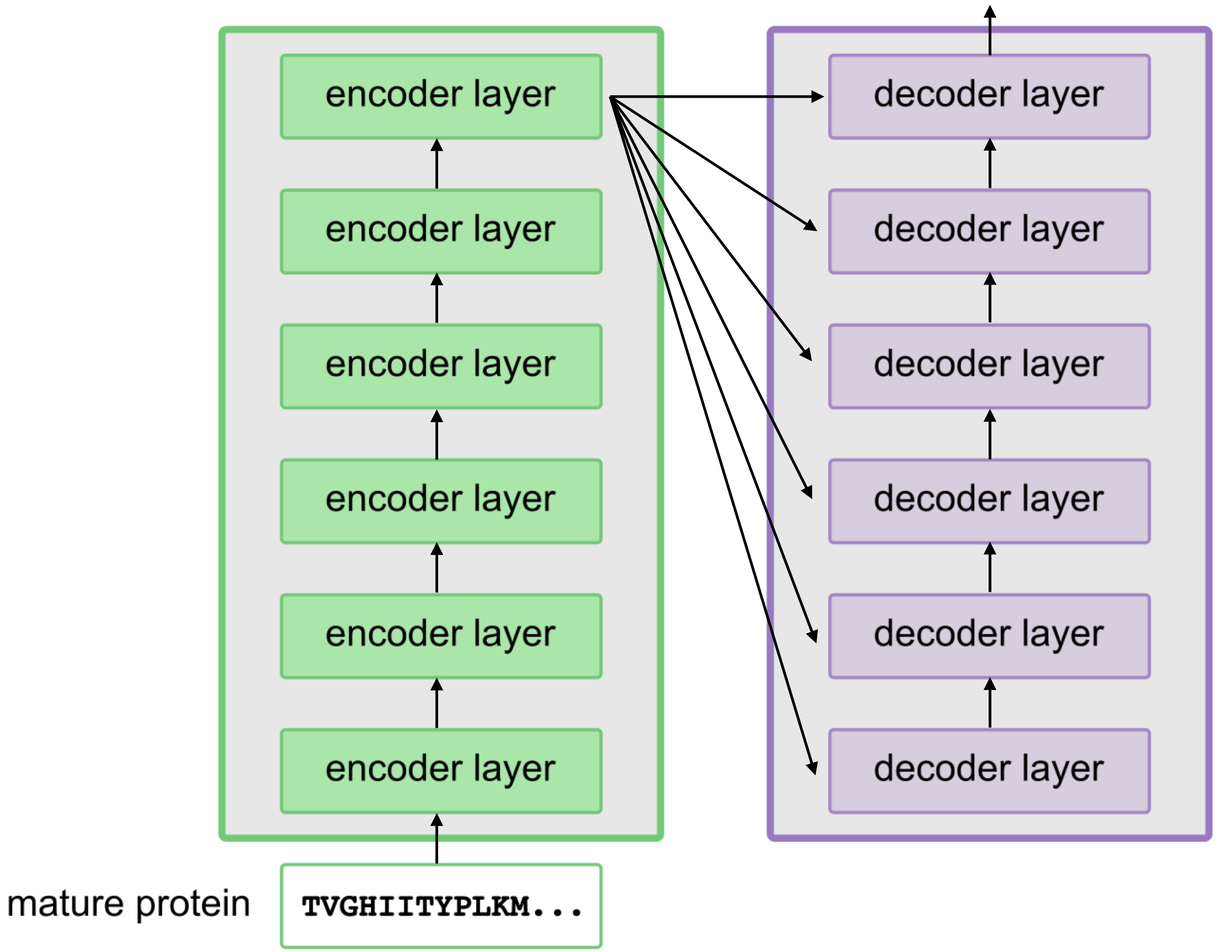


Signal peptides are structured



- Extra-cellular secretion in prokaryotes
- Secretion from endoplasmic reticulum in eukaryotes
- SP dependent on species and protein
- Rules insufficient for generation

Machine translation for SP generation



- Transformer architecture
- 40k examples from UniProt
- 40% test accuracy

Actual `MNFSKIFIFVVLAVLLLCSQTEA`
Predicted `MNFKLIFLVALVLMAAFLGQTEG`

Actual `MLRKSAVLAGVVLLGASAQA`
Predicted `MRLSTSALVLGAASSAV`

Actual `MASRVECLVVCLLVLRVAQSSG`
Predicted `MNLASVLLLLAACHLSVSVNG`

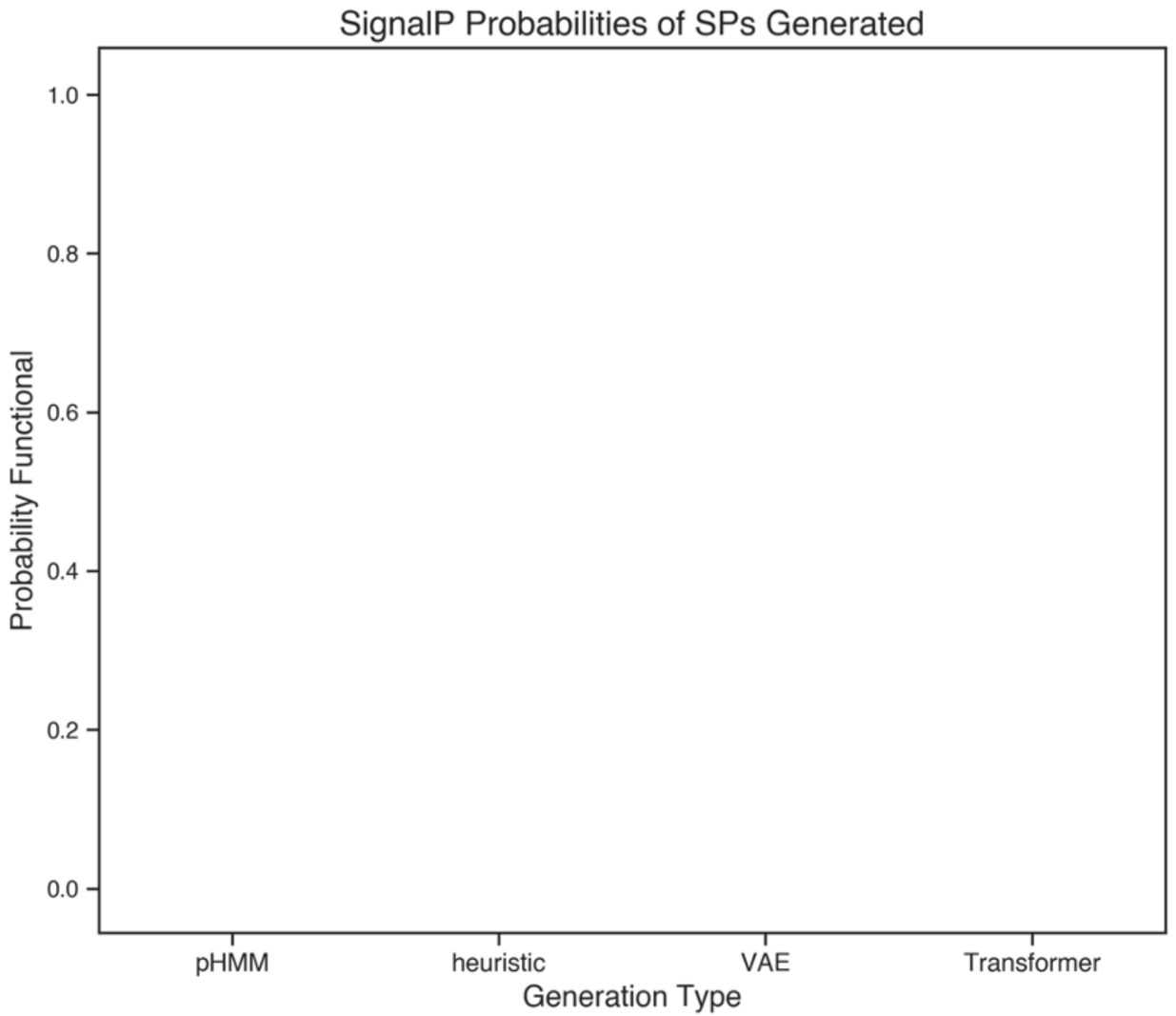
Actual `MLWIRAVLPLVAGSDT`
Predicted `MLGIWTLLPLVLTYVVRLLSKCVNA`

Generated SPs trick SignalP 5.0

Method	Mean SignalP Probability
Hidden Markov Model	16% ± 19%
heuristic	71% ± 25%
Variational Autoencoder (VAE)	**92% ± 15%**
Transformer	90% ± 17%

VAE sequences are repetitive

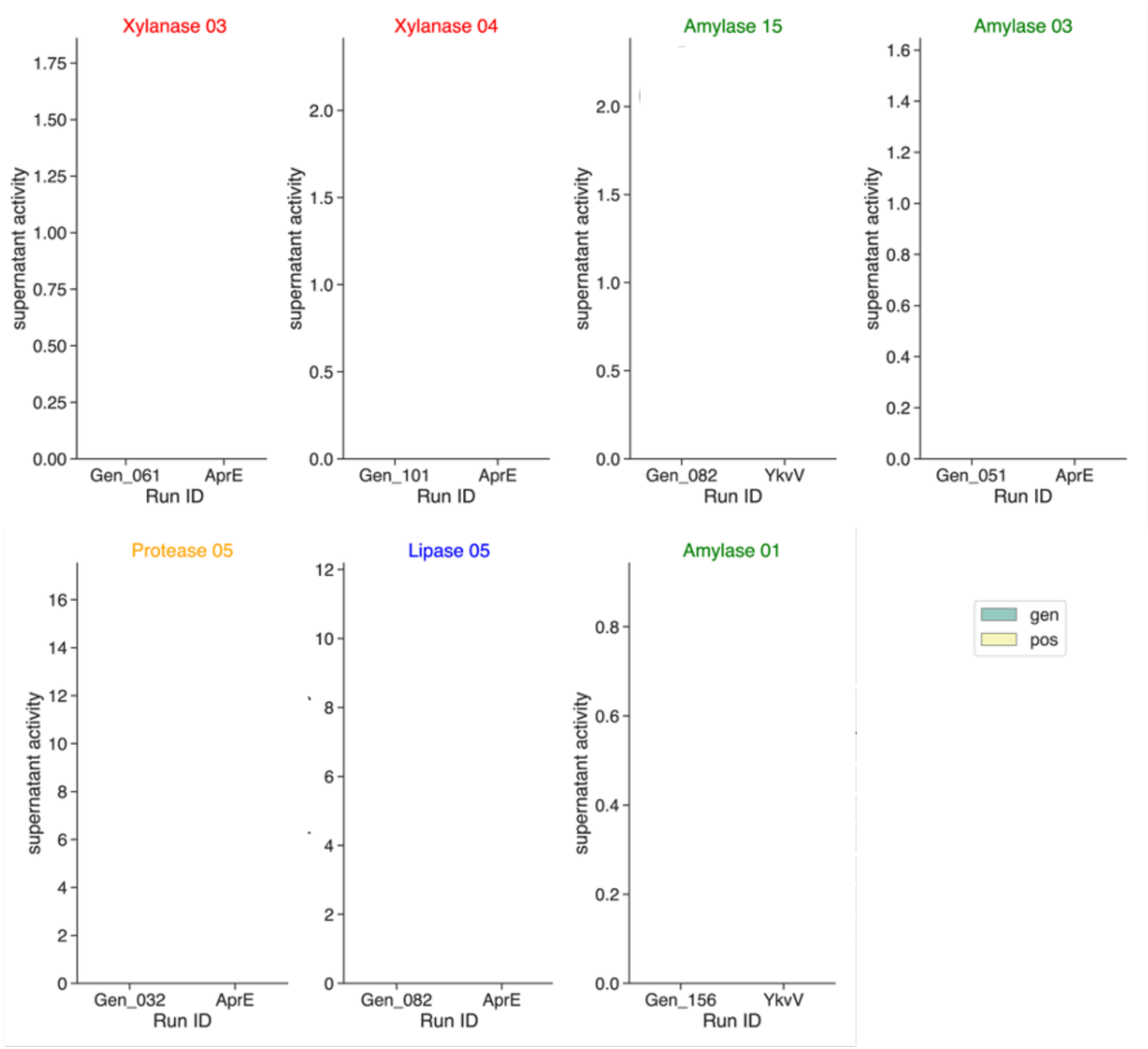
```
MRKRLLALALAALSLLLLLSFGVKALAGSGA
MRLLLLLLLVLLLAAPAPPGLS
MKLLLLLVTLLTSVLALQA
MRLLLLALLAAAAVALASA
MASSSSSLFVVLAVLLLLLLLTLSSA
```



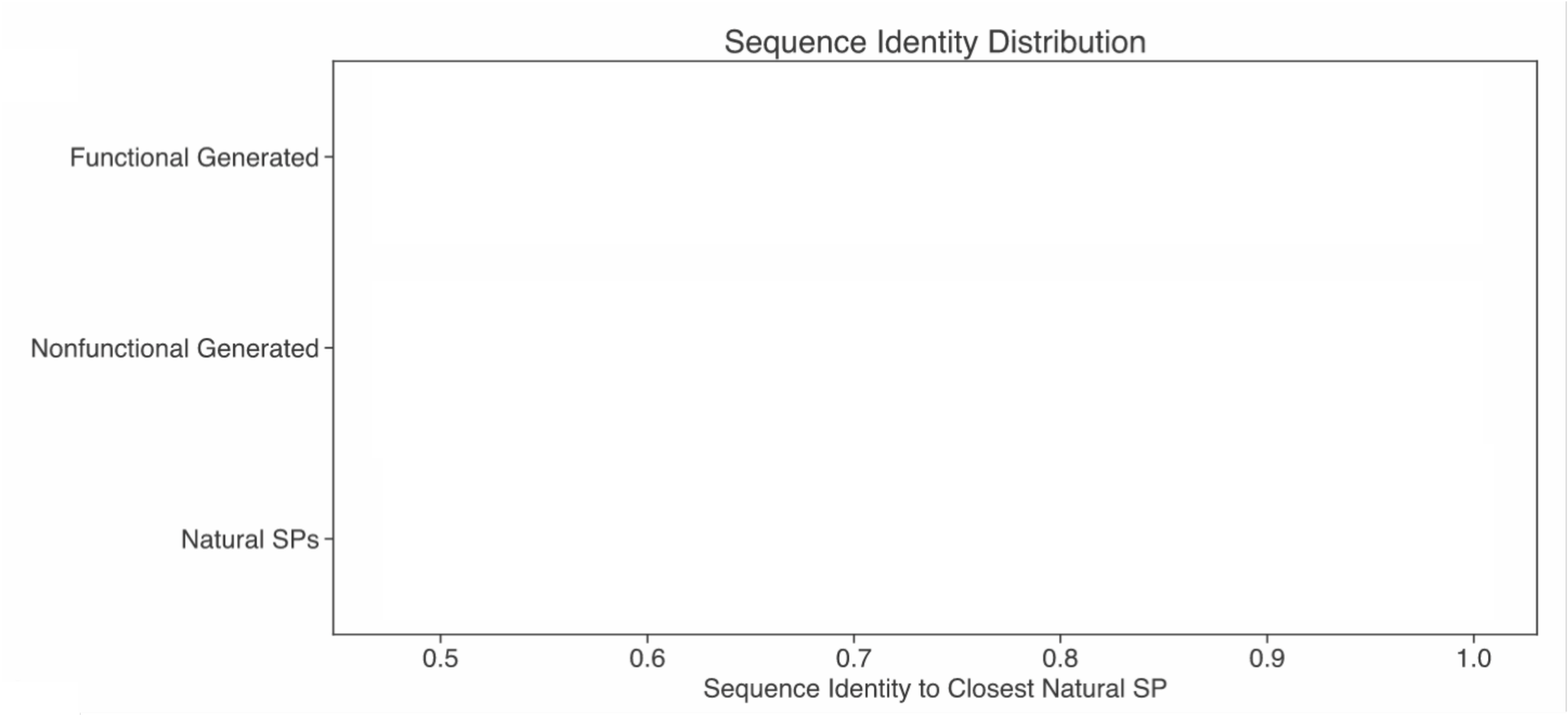
Nielsen *et al.,* 2019

Generated SPs are functional in *Bacillus subtilis*

- 79% of natural SPs are functional
- 48% of generated SPs are functional
- Enzymatic activity is comparable



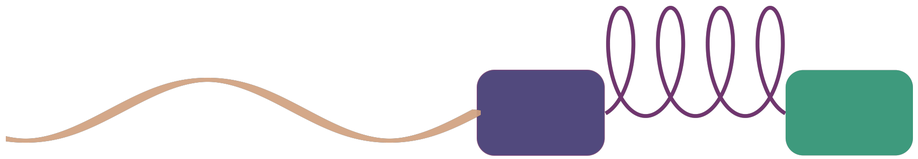
Generated functional SPs are novel and diverse



0.75 ± 0.09

0.82 ± 0.1

SP generation conclusions

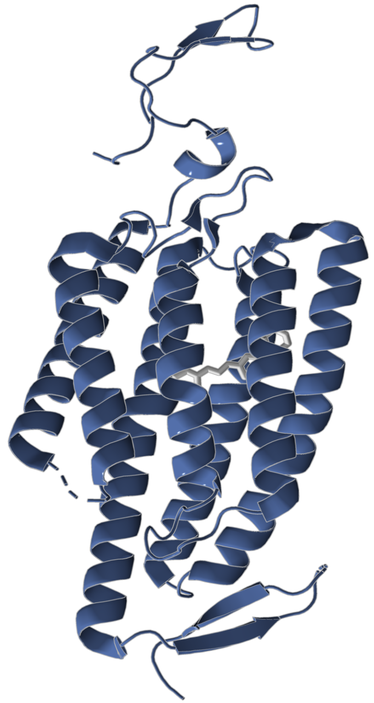


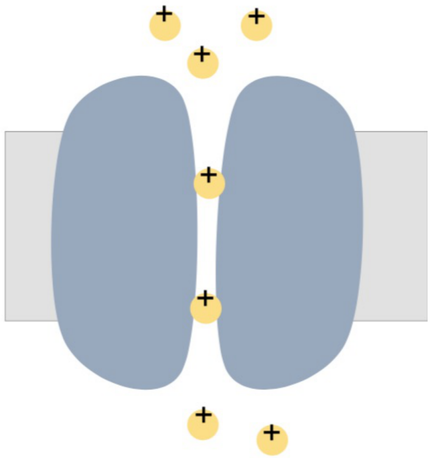
- Can generate *de novo* signal peptides
- That look like real SPs
- And result in functional secreted enzymes

How about entire proteins?

Generating new, designable structures expands functional space

MGTGDHDD...



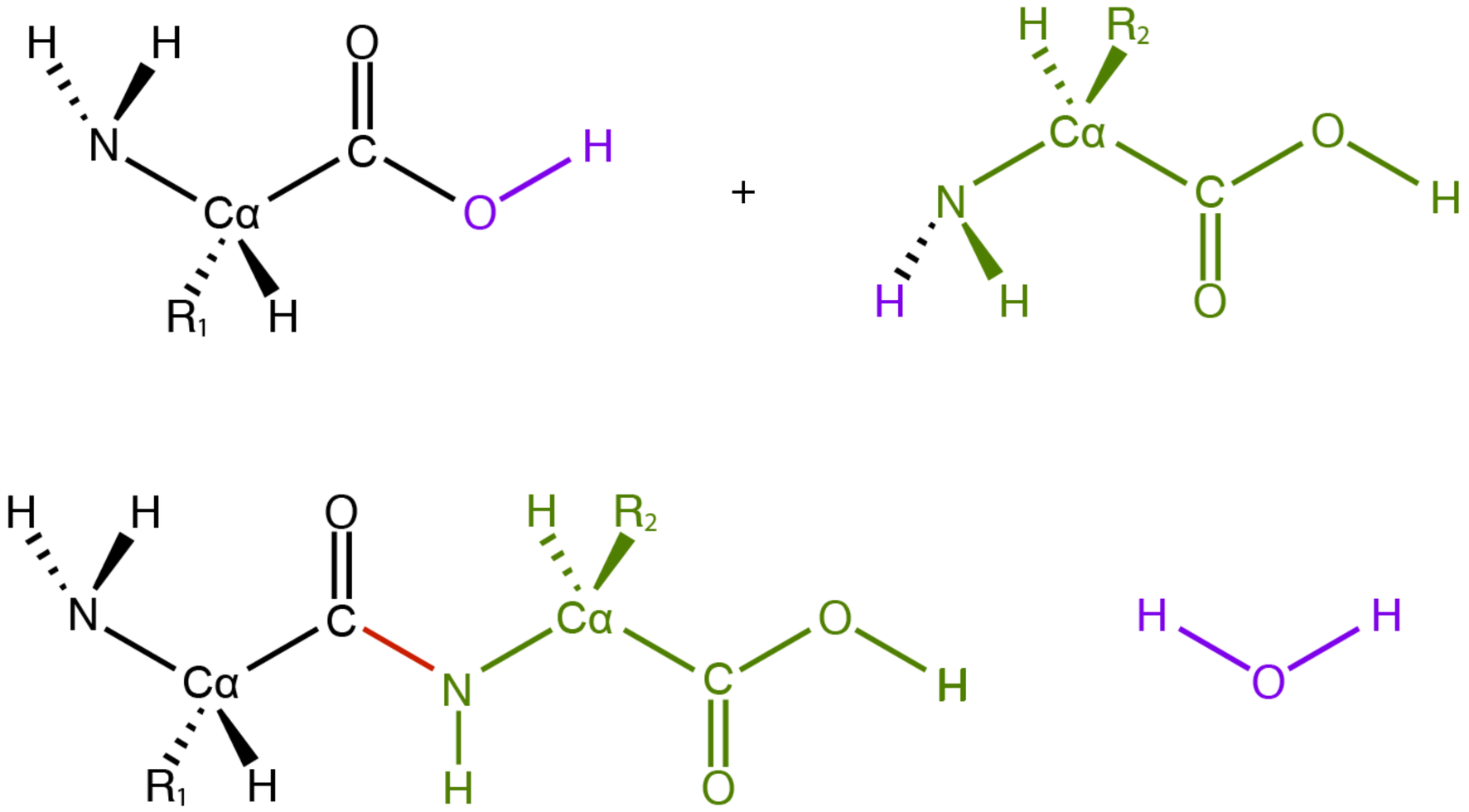


Challenge: Generate diverse and designable structures

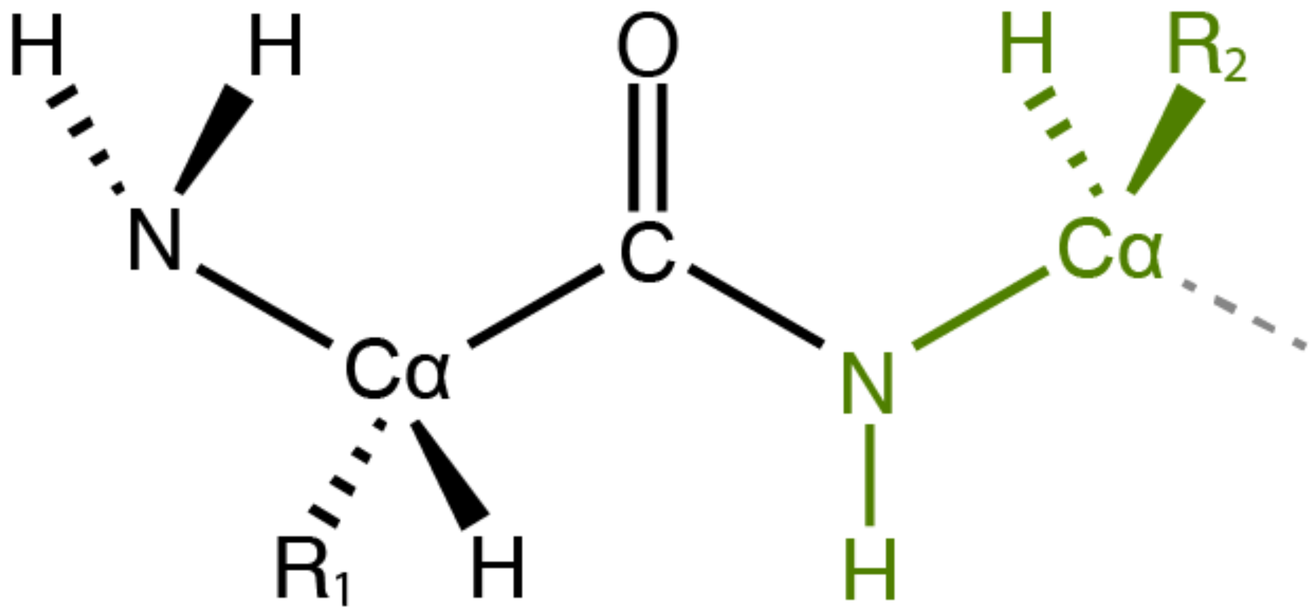




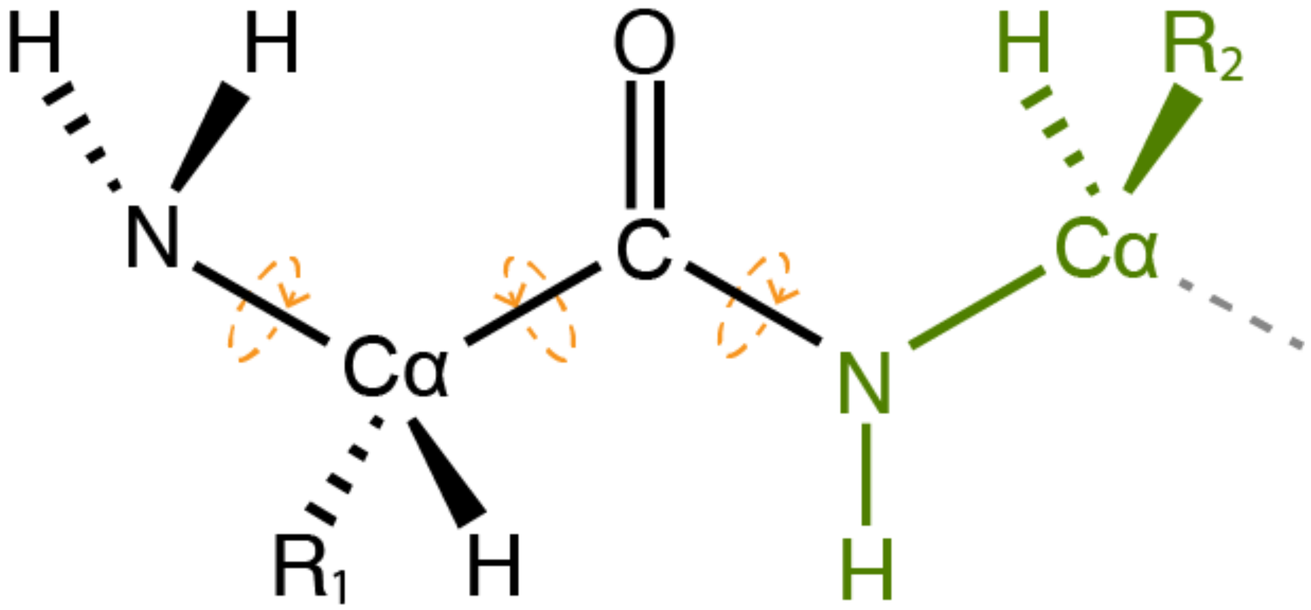
Proteins are polypeptide chains



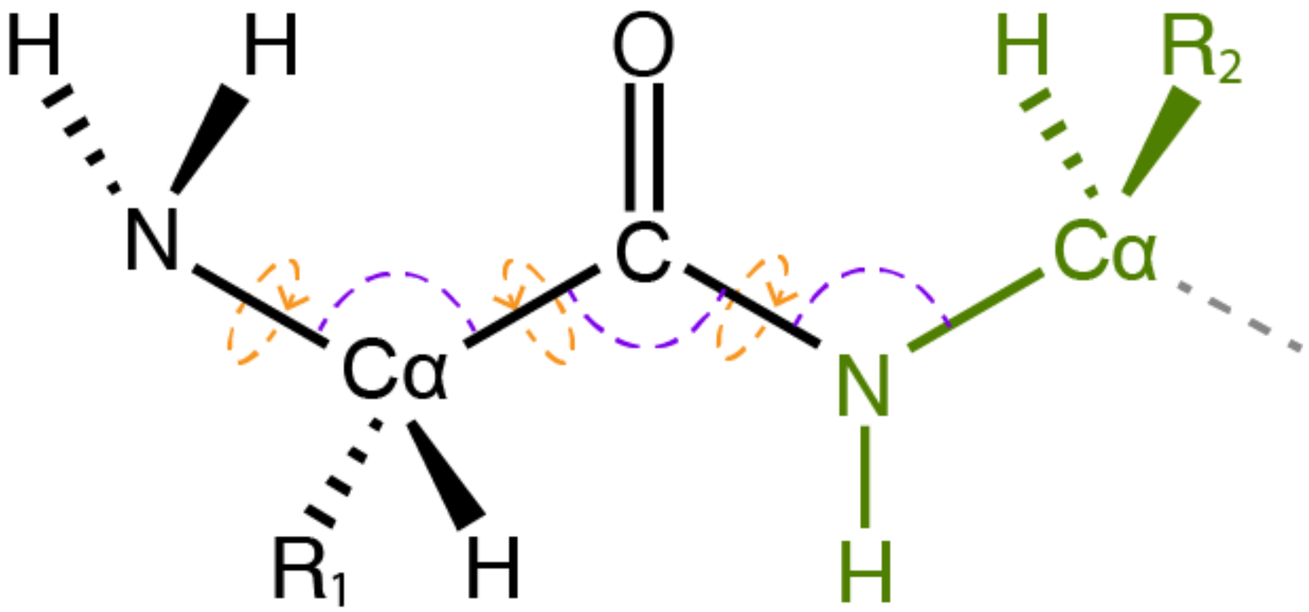
Protein structure is determined by bond angles



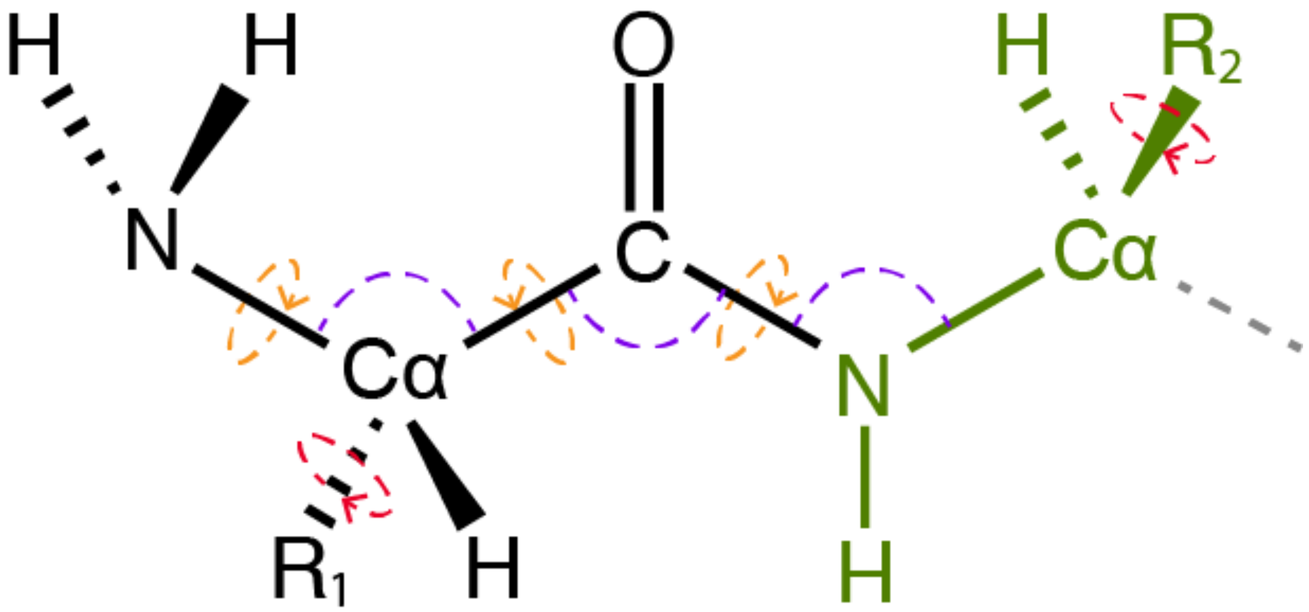
Protein structure is determined by bond angles



Protein structure is determined by bond angles

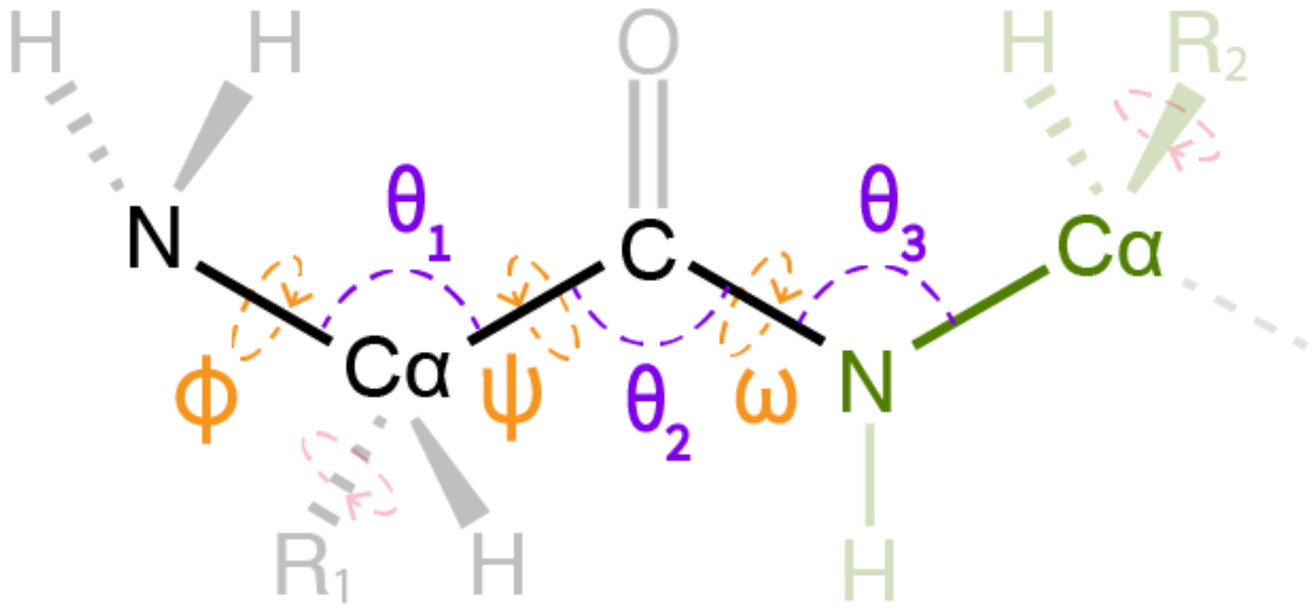


Protein structure is determined by bond angles



These six angles (for every consecutive pair of amino acids) fully determine the structure

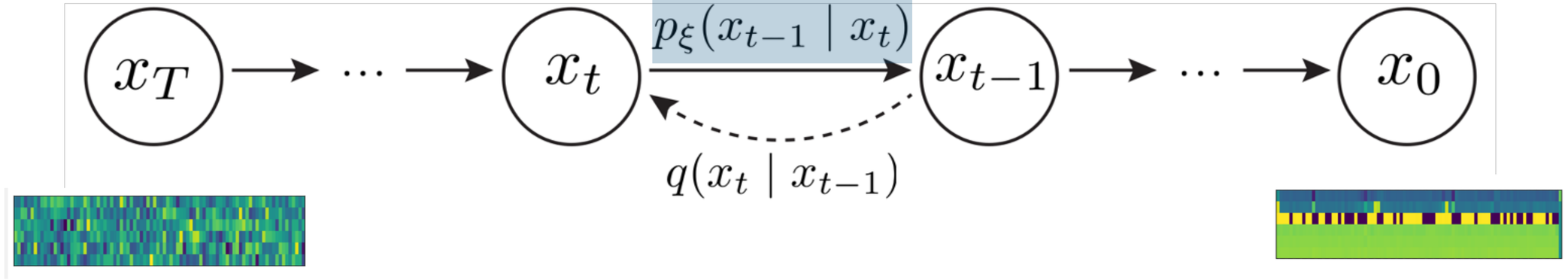
We generate backbone structures represented by dihedral and bond angles



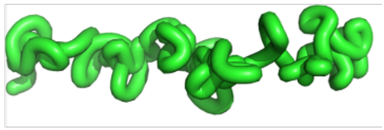
These six angles (for every consecutive pair of amino acids) fully determine the backbone structure

FoldingDiff uses diffusion to generate angles

Vanilla transformer



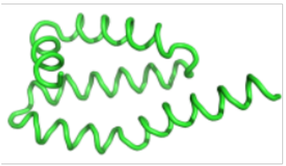
random angles



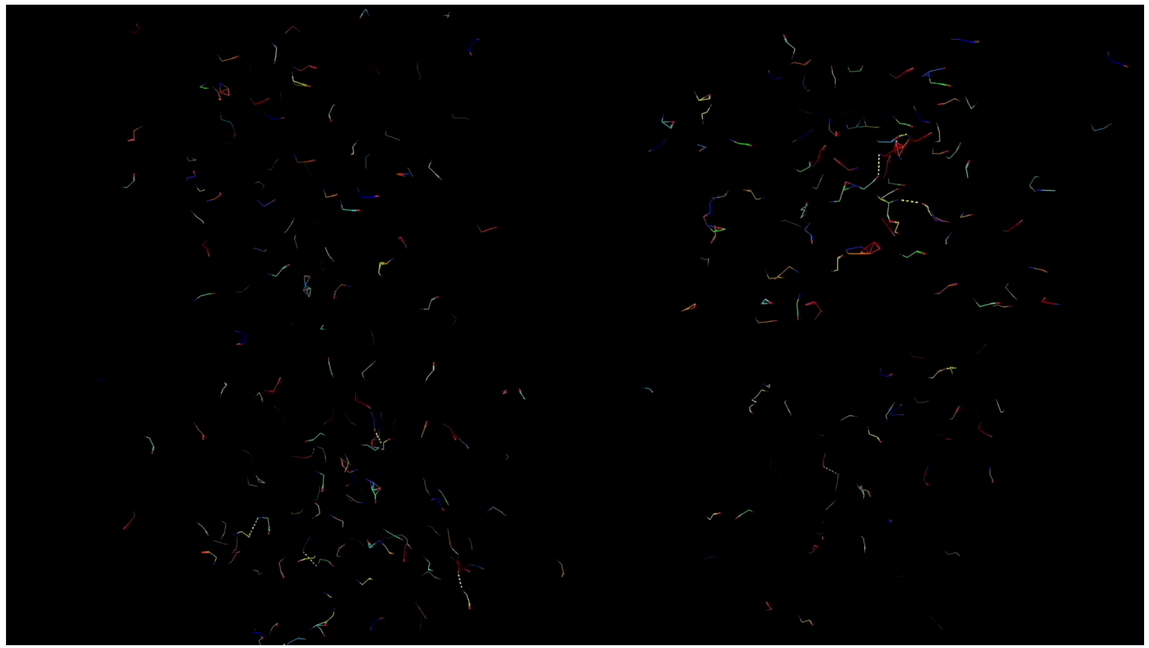
30k structures from CATH

40 - 128 residues

generated angles

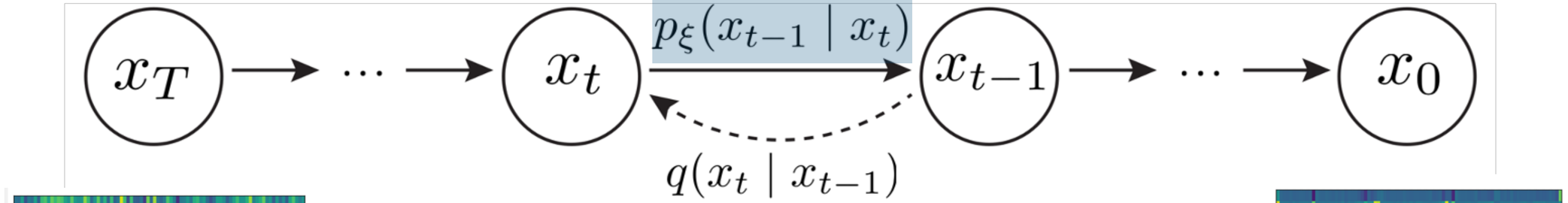


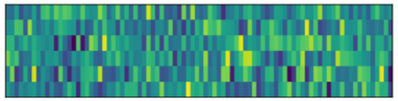
Diffusion on 3D coordinates requires equivariances



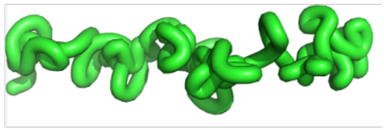
Evaluate generations at 3 levels

Vanilla transformer



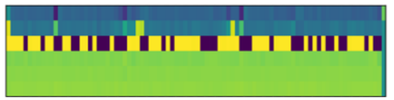


random angles

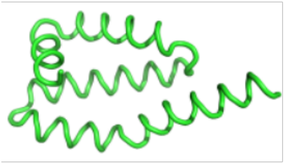


30k structures from CATH

40 - 128 residues



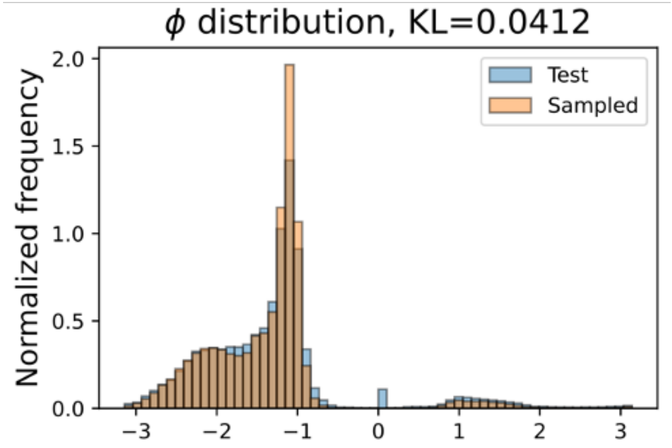
generated angles



generated angles | structural motifs | overall structures

Generated angles match test distribution

noise → sample → compare to true distribution

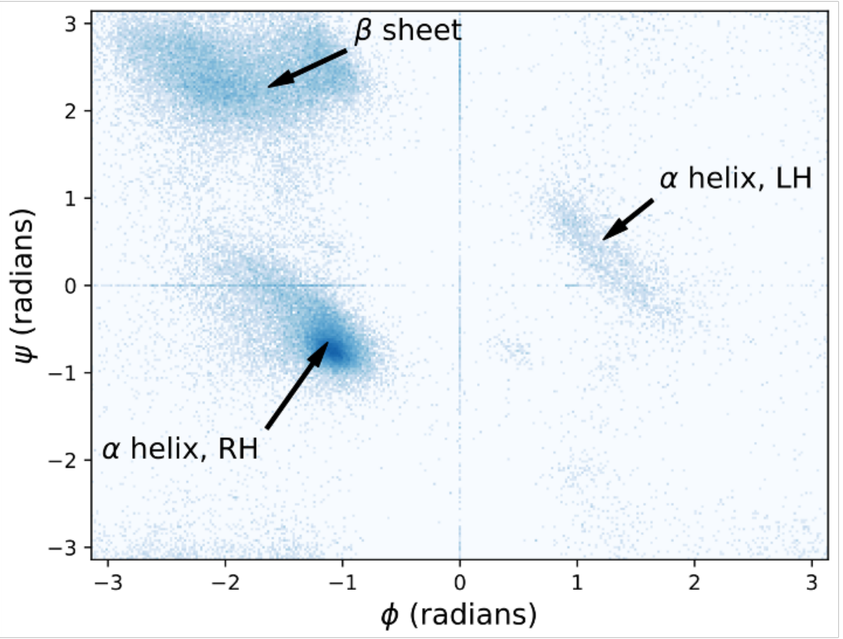


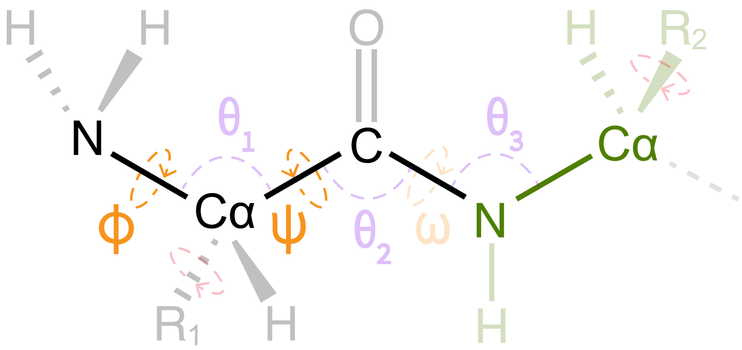
Generated distributions match natural distribution of individual angles

FoldingDiff captures correlations between angles

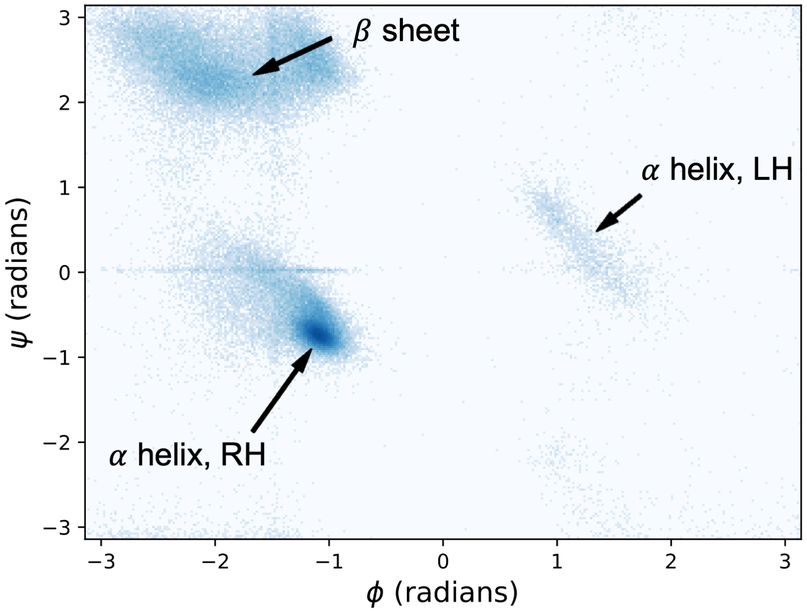
noise ⟶ sample ⟶ compare (ϕ, ψ) co-occurence

True



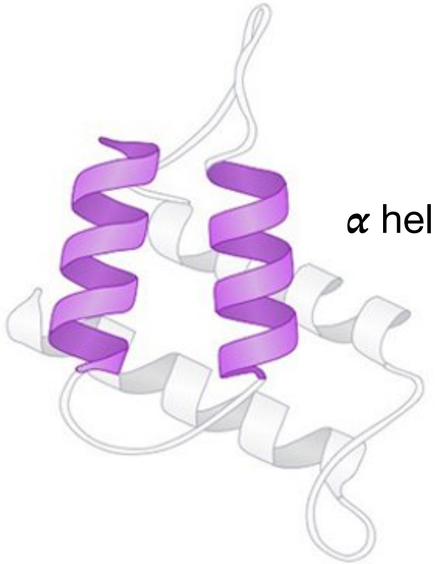


Generated

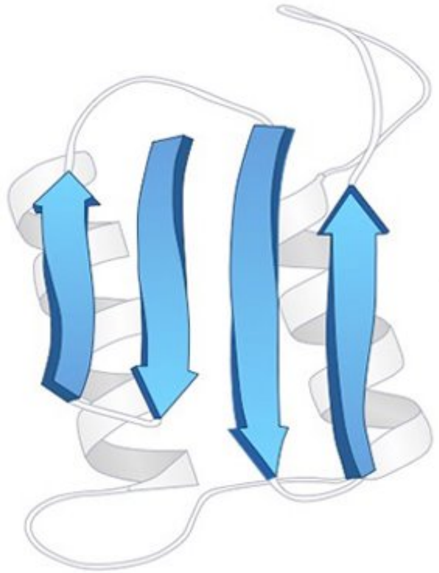


FoldingDiff generates correlations that define common structural motifs

Generated secondary structures match test structures



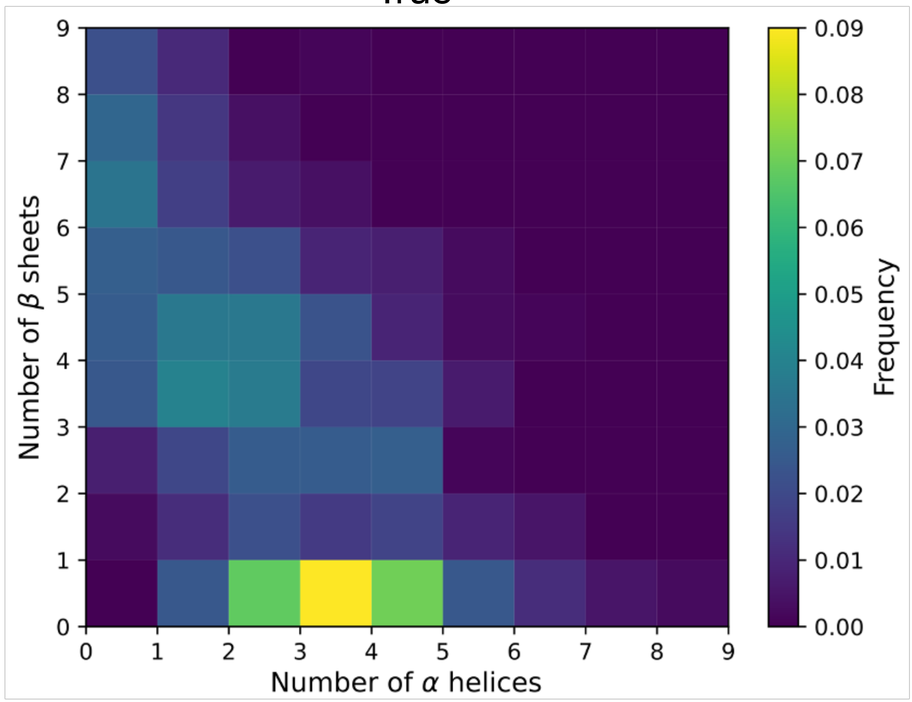
α helix



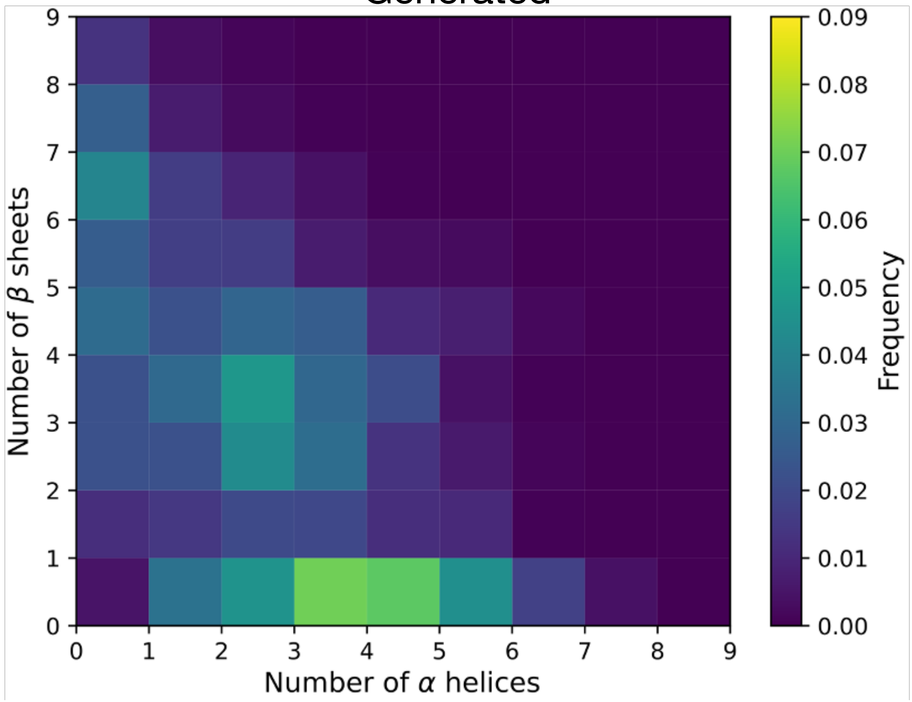
β sheet

noise ⟶ sample ⟶ measure helix, sheet structures

True

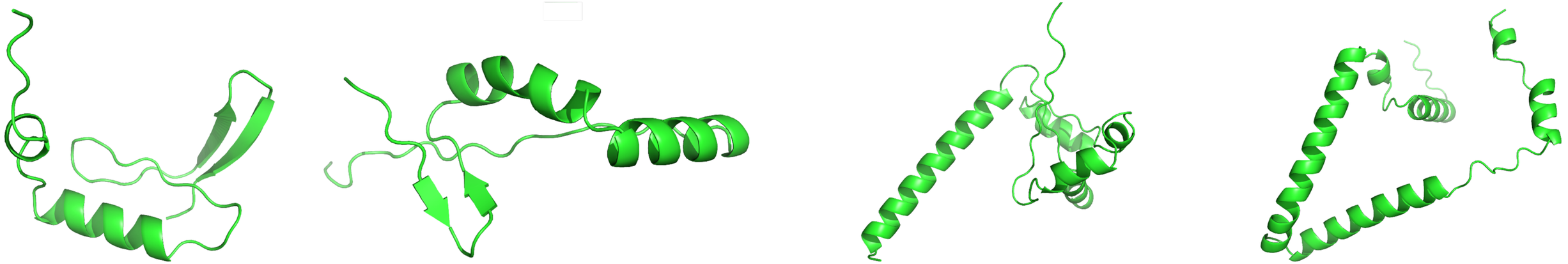


Generated

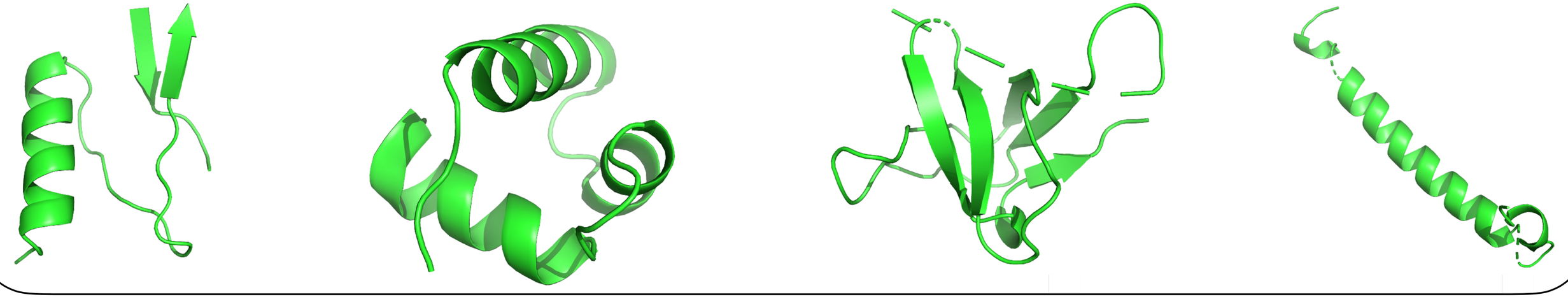


Generated structures look reasonable

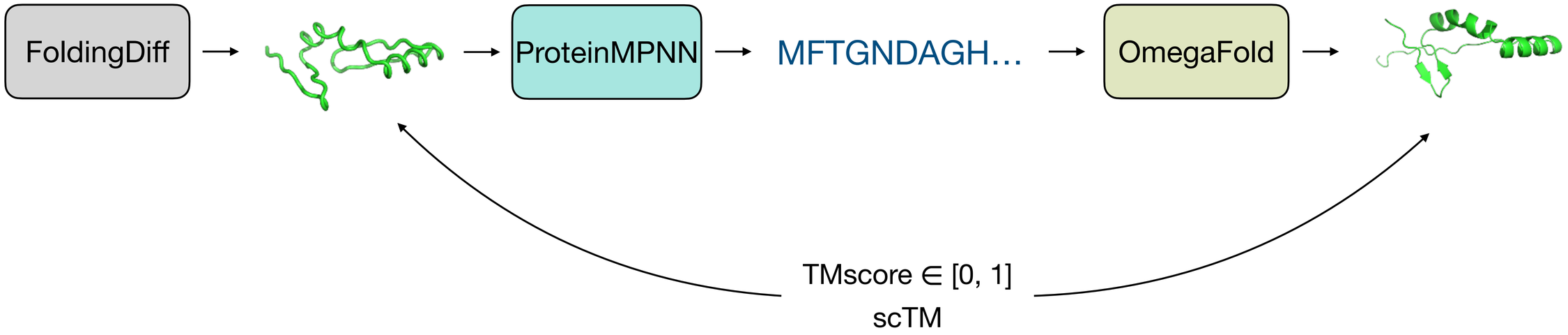
Generated structures



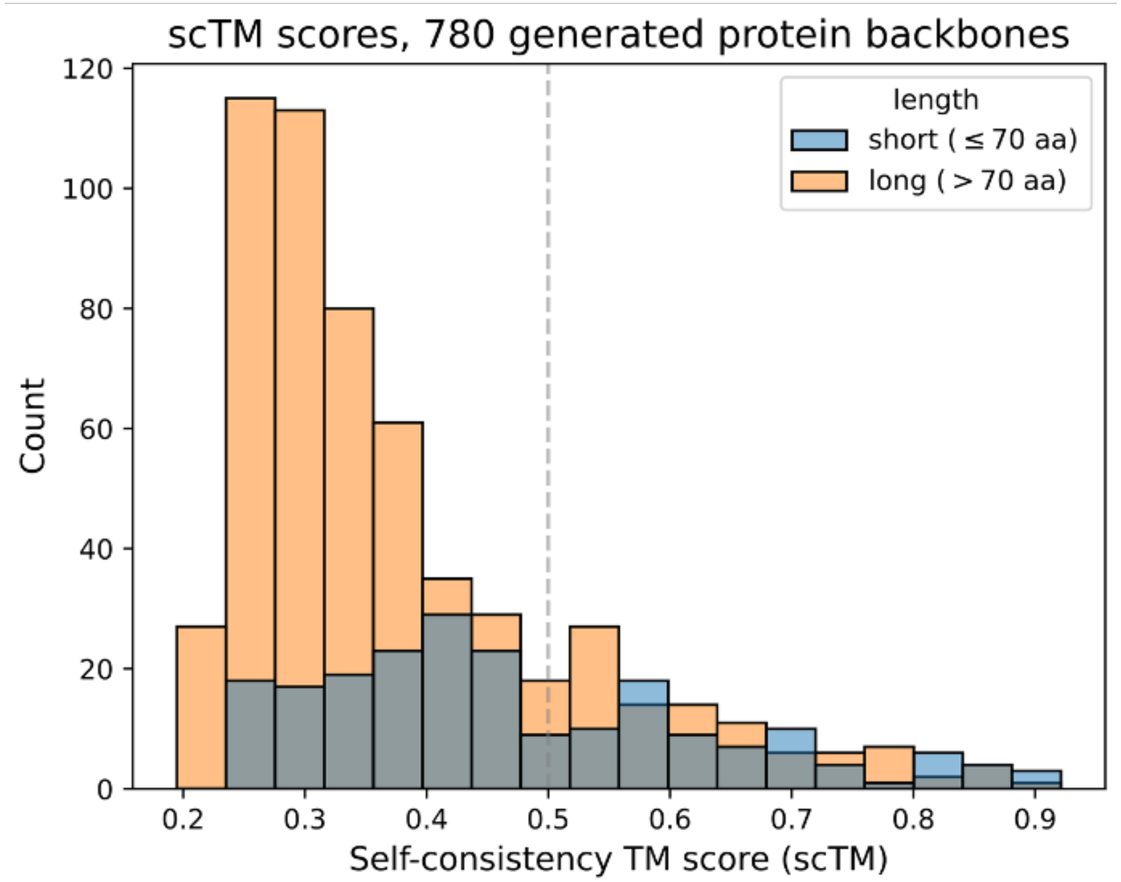
Training set structures



Measure designability of structures with self-consistency TMscore



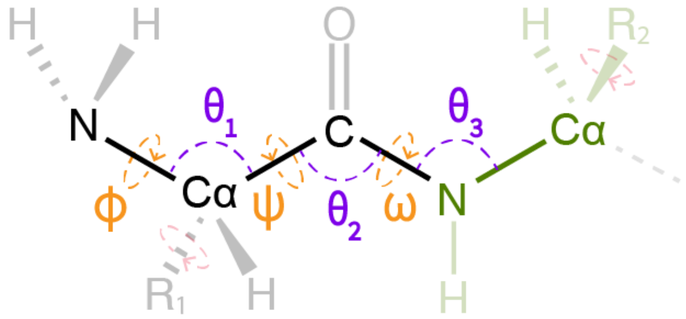
Many generated structures are designable



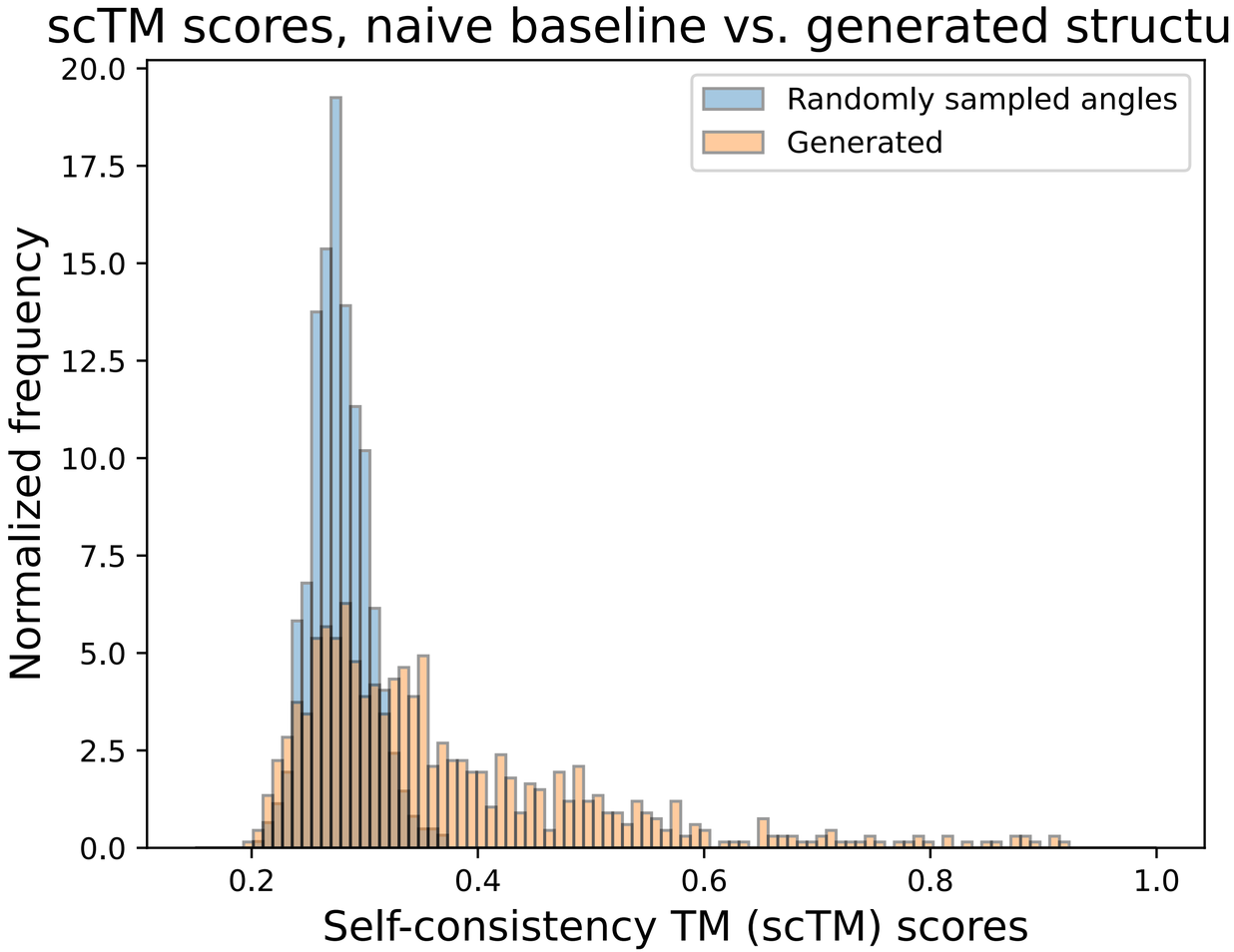
scTM > 0.5	≤ 70 aa	> 70 aa
FoldingDiff	76/210	87/570
ProtDiff (Trippe *et al.*)	36/210	56/570

Significant improvements over point cloud diffusion model

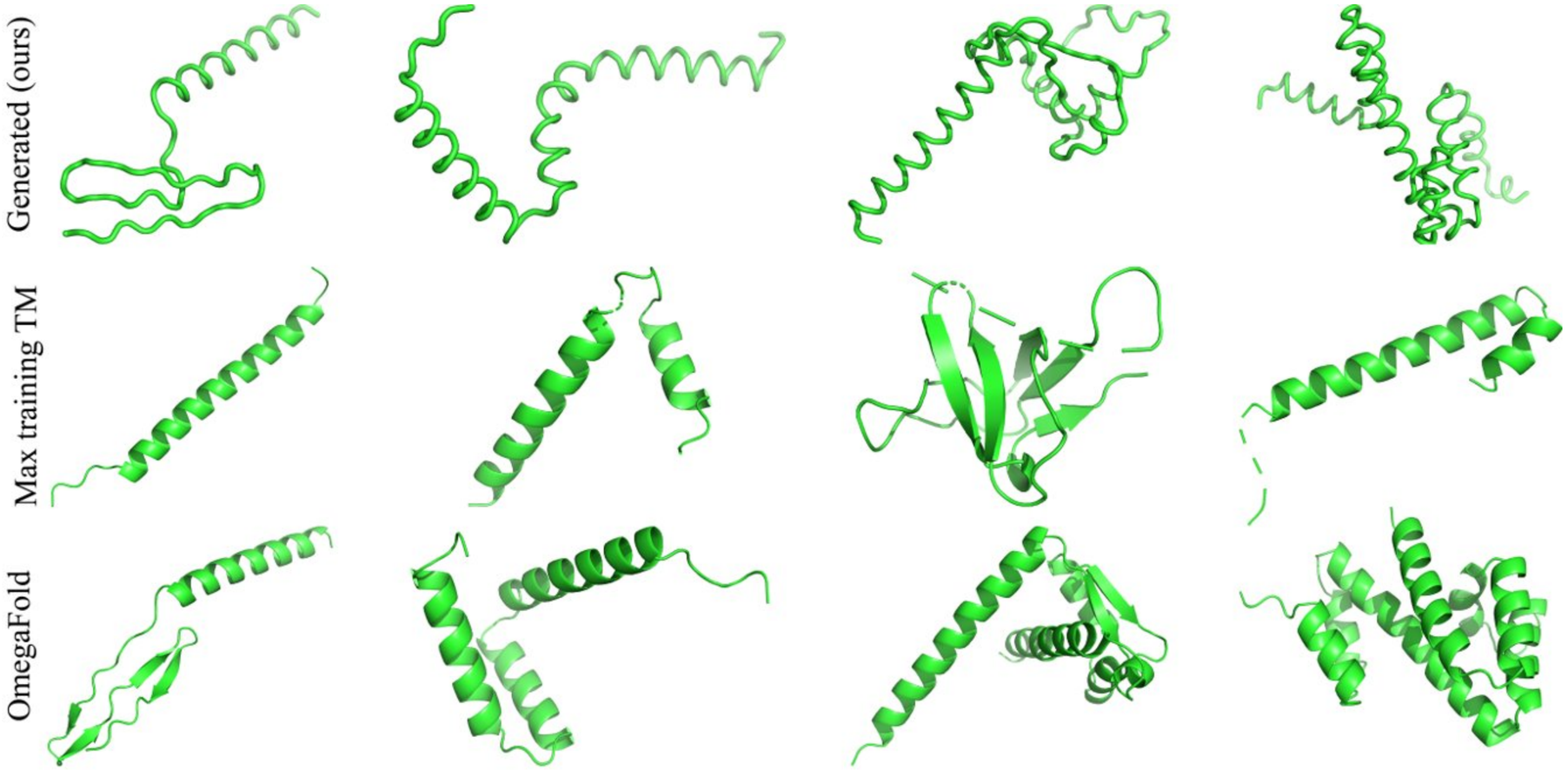
FoldingDiff structures are better than random baseline



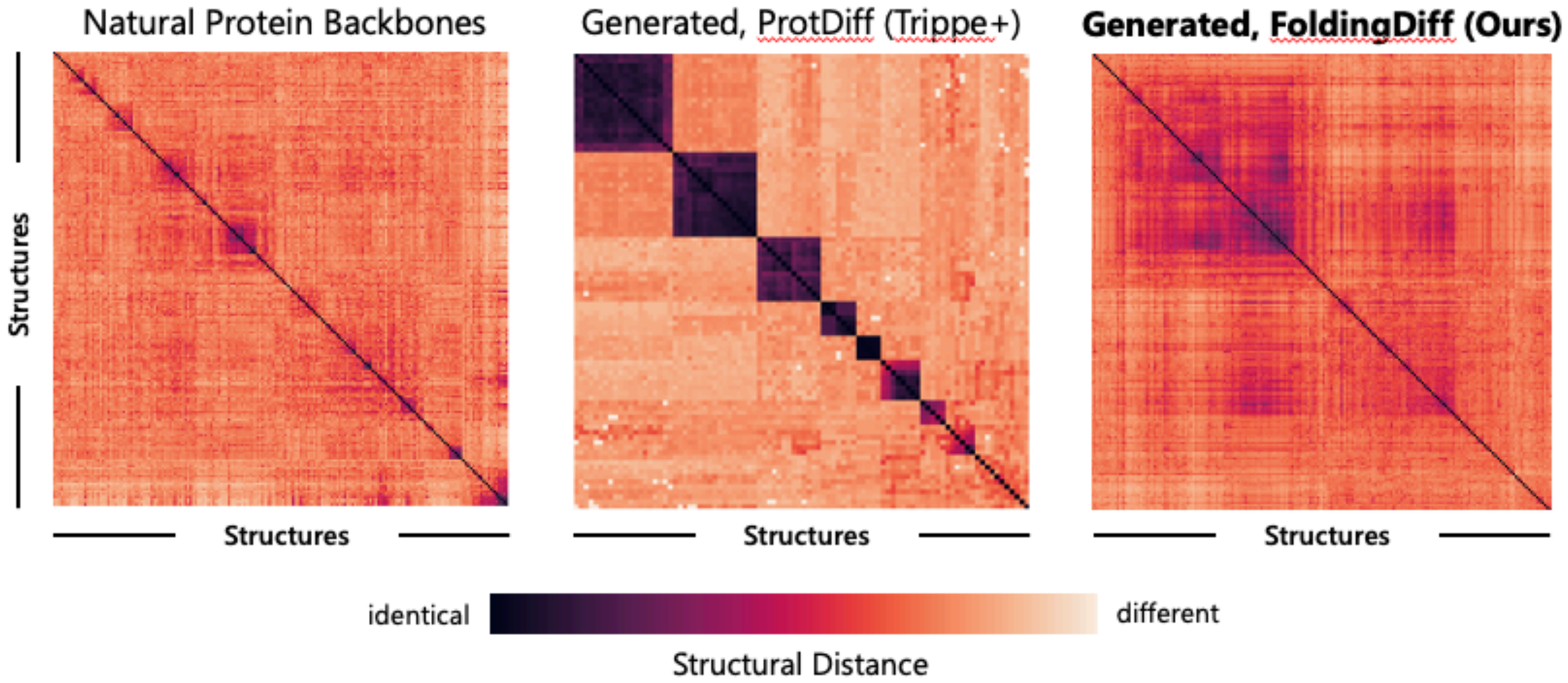
Sample sets of angles
Preserves Ramachandran plot



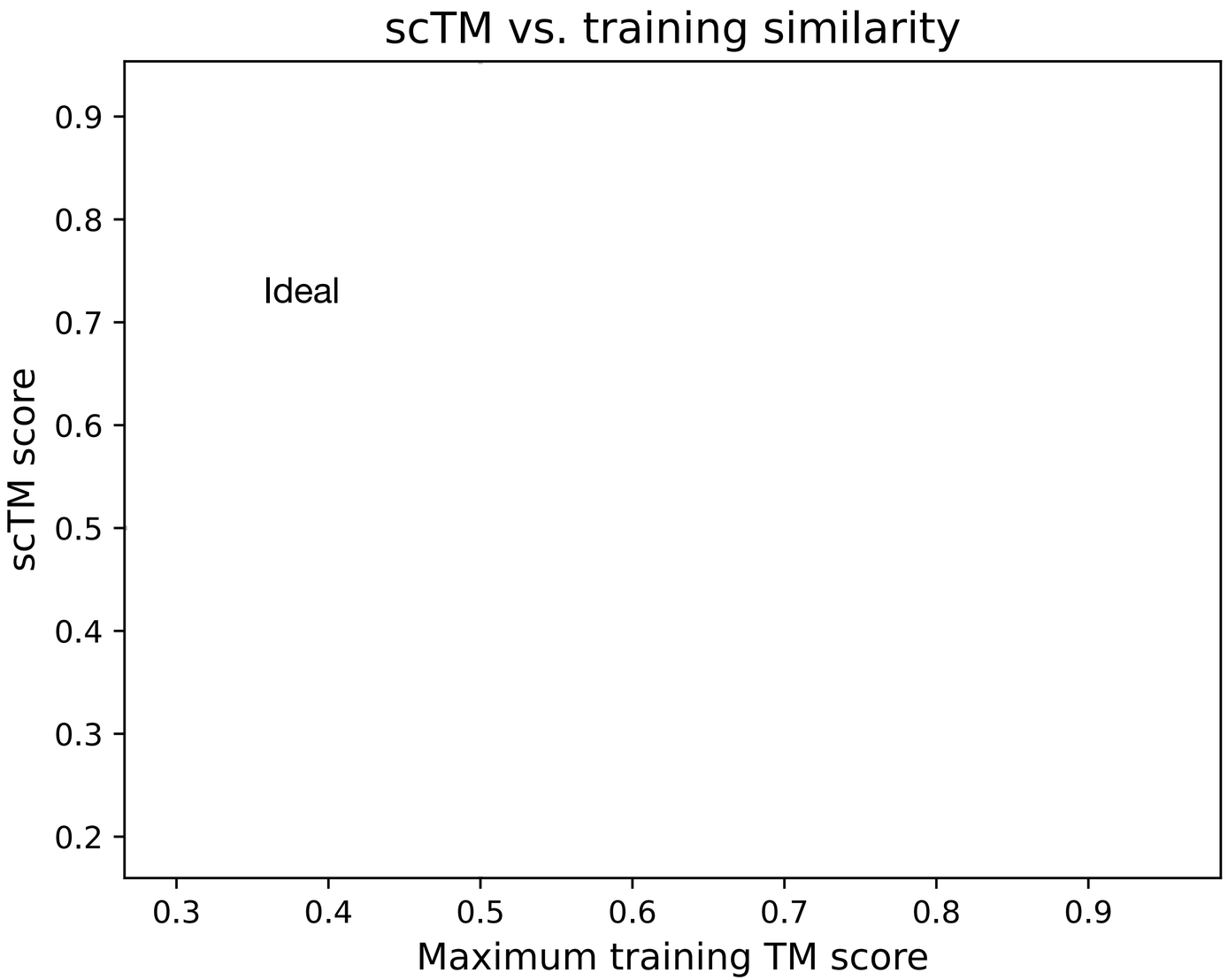
Generated structures are diverse



Generated structures are diverse

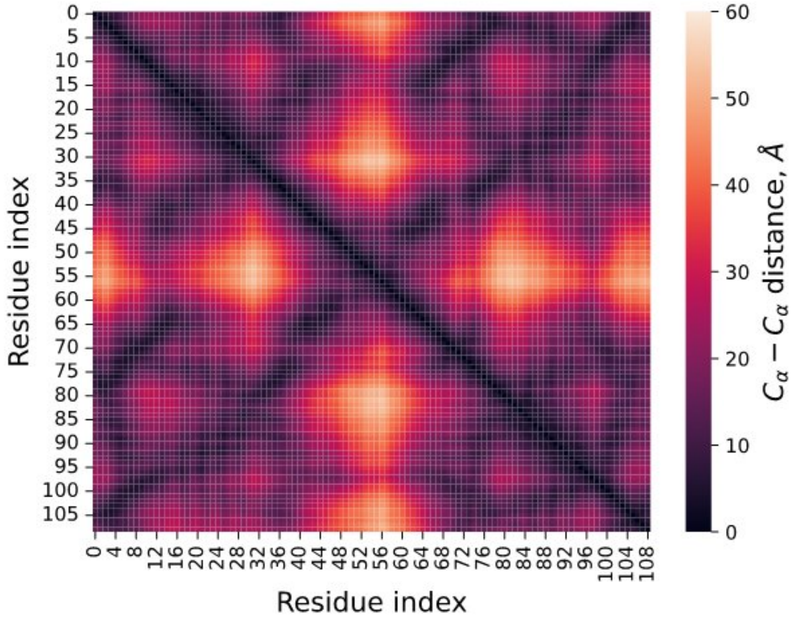


Generated structures are diverse

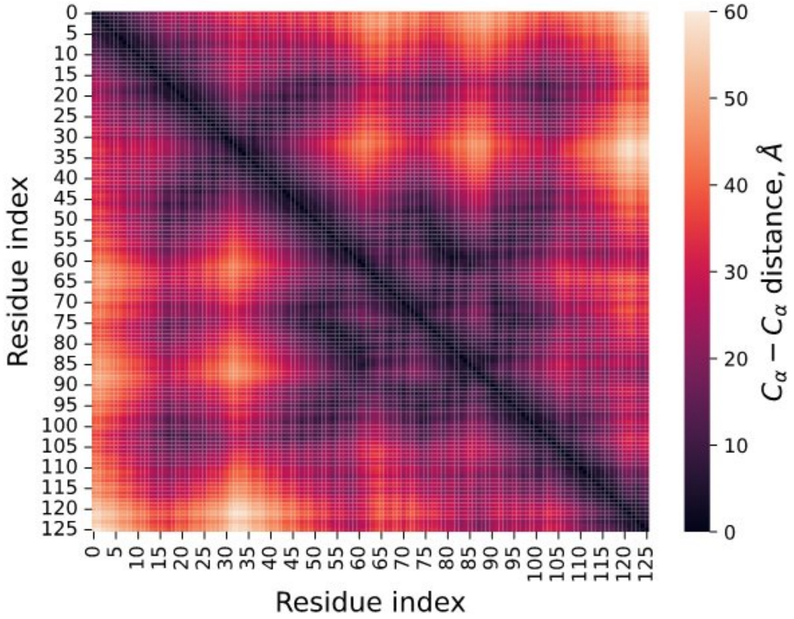


Angle formulation enables generation from vanilla transformers

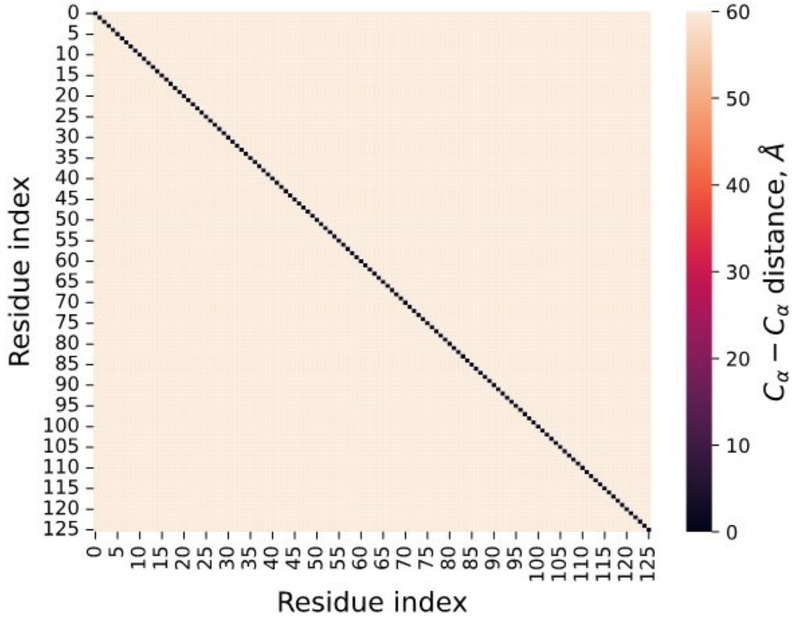
FoldingDiff



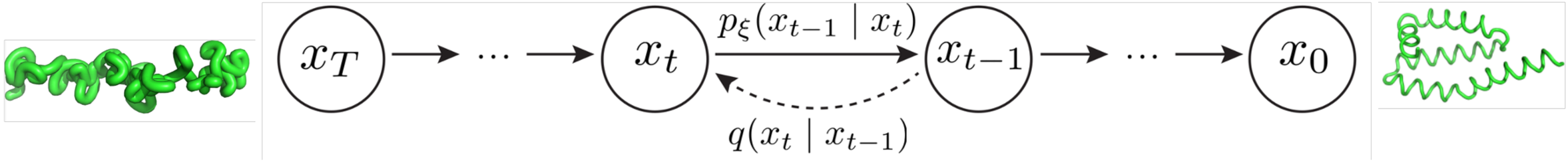
Real protein



Transformer + point cloud

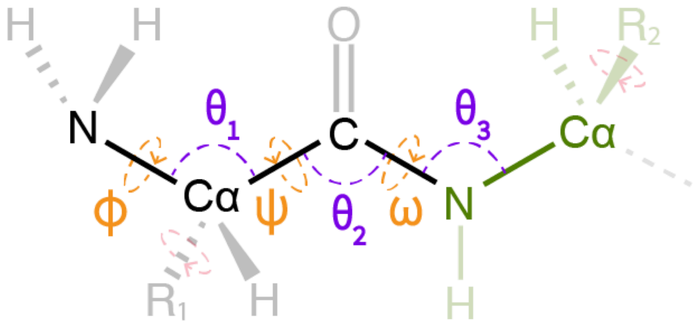


FoldingDiff is first step towards generating new functions

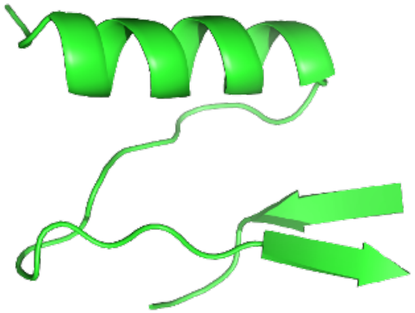


Generate protein backbones by mirroring the folding process

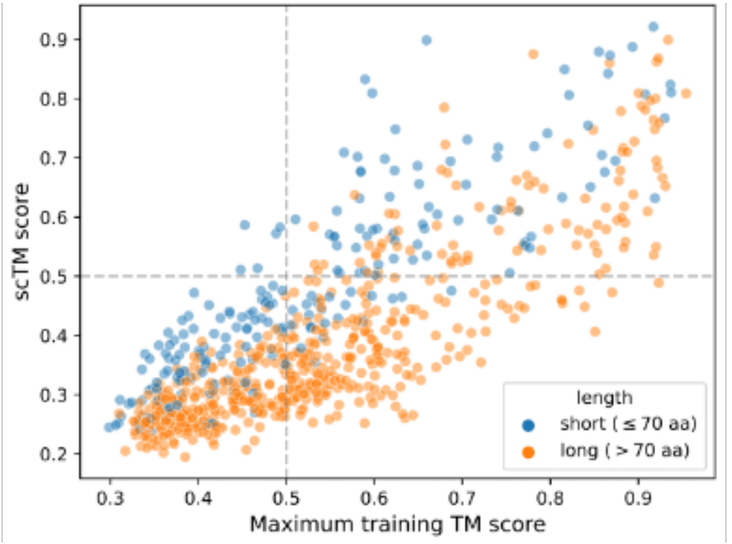
Internal coordinates



Structural motifs



Realistic, diverse samples



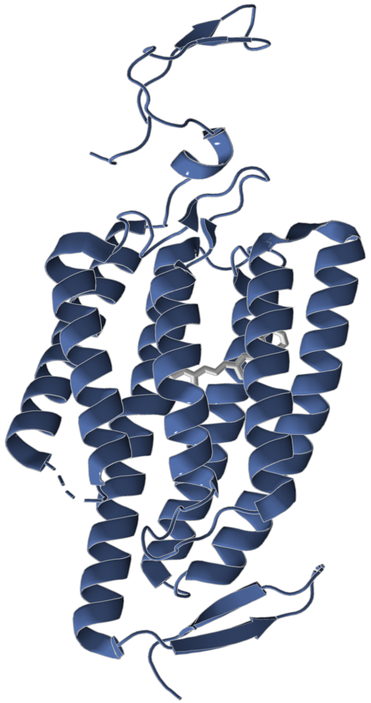
github.com/microsoft/foldingdiff

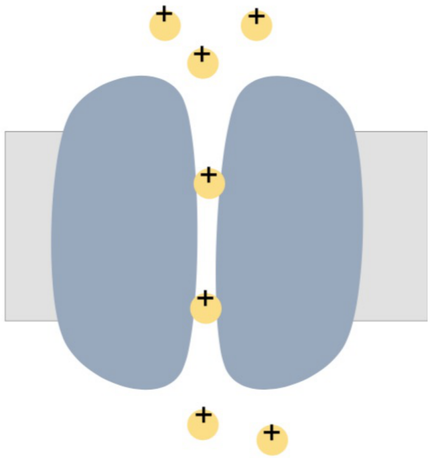
huggingface.co/spaces/wukevin/foldingdiff

BioNeMo?

Generating new, valid sequences expands functional space

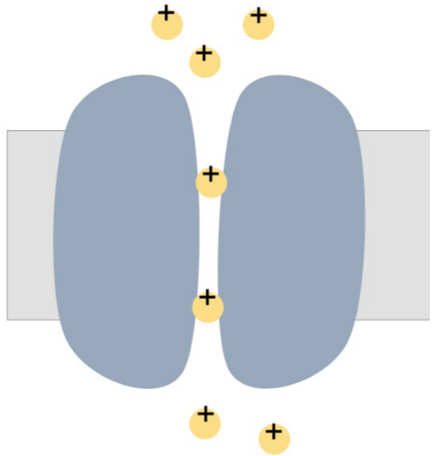
MGTGDHDD...





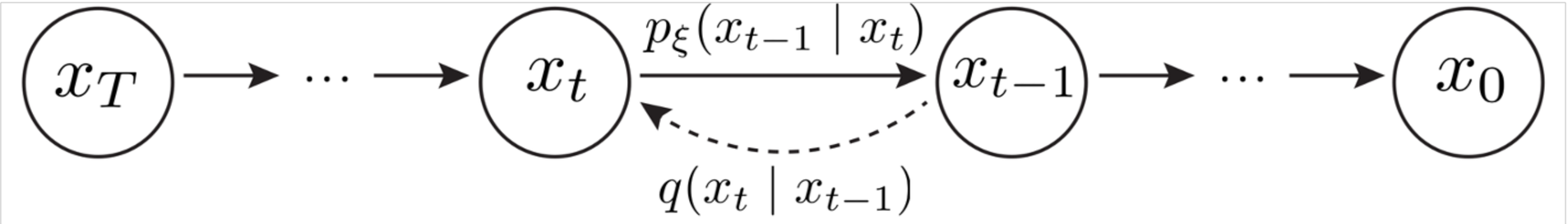
Generating new, valid sequences expands functional space

MGTGDHDD...



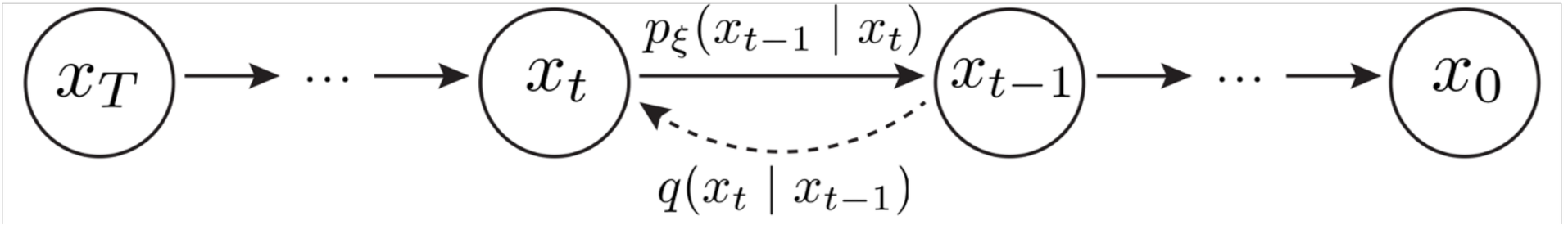
Challenge: Generate valid and diverse sequences

Use discrete diffusion to generate sequences



MGTGDHDD…

Order-agnostic autoregressive diffusion links diffusion, masked language models, and autoregressive language models



OA-ARDM ########… ← MG##D#DD… ← MG#GD#DD… ← MGTGDHDD…

mask

MLM

MG#GD#DD…

↓

MGTGDHDD…

learn single diffusion step

LR-AR

MGTGD#

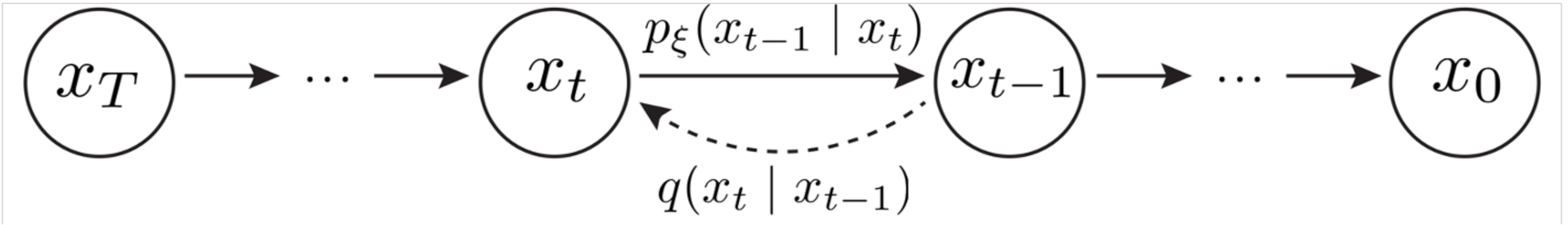
↓

MGTGDH

learn single decoding order

OA-ARDM generalizes masked and left-right autoregressive language models

Discrete diffusion mutates towards a random string



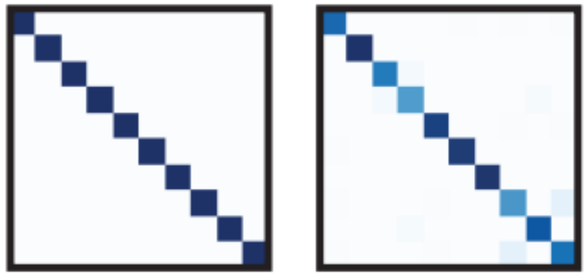
OA-ARDM ########… ← MG##D#DD… ← MG#GD#DD… ← MGTGDHDD…

mask

D3PM TYLPLKNE… ← MAKTDRDL… ← MAGTDRDD… ← MGTGDHDD…

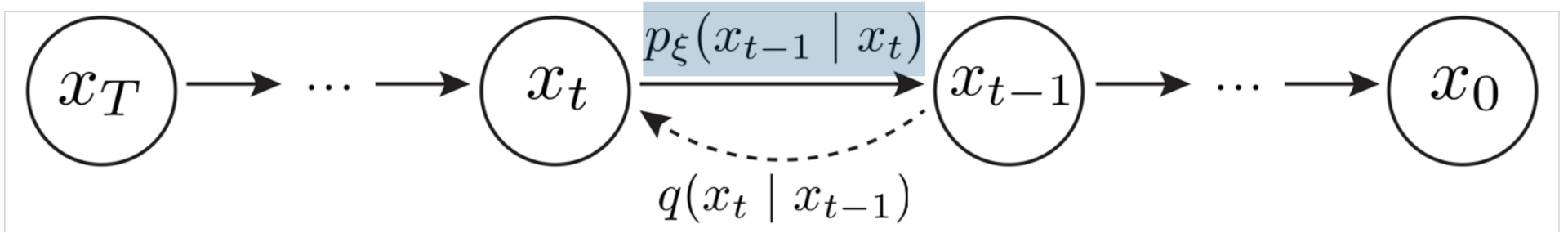
mutate

Transition matrix



Train with same dataset and architecture for fair comparisons

640M parameter CNN 42M sequences from UniRef50



OA-ARDM ########… ← MG##D#DD… ← MG#GD#DD… ← MGTGDHDD…

mask

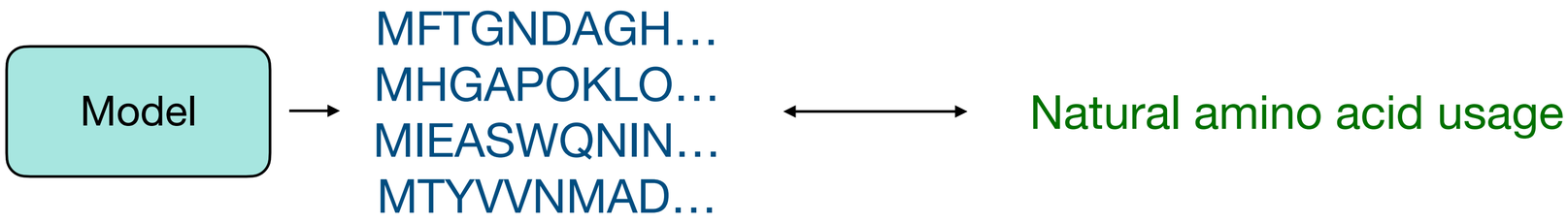
D3PM TYLPLKNE… ← MAKTDRDL… ← MAGTDRDD… ← MGTGDHDD…

mutate

---

generated residues | structural motifs | overall sequence and structure | distribution match

Scale and diffusion improve generated amino acid distribution



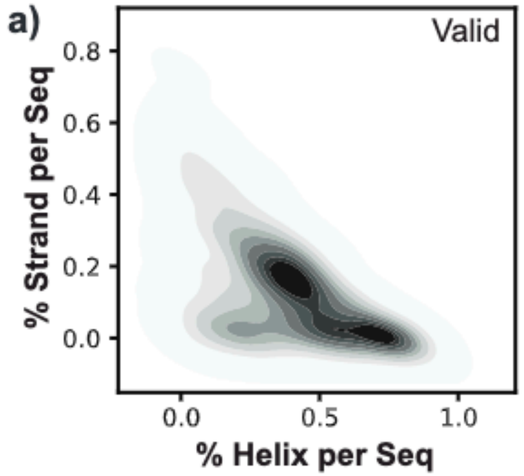
Model	Params	KL
Valid	-	0.00059
Random	-	0.165

Unlike NLP, OA-ARDM is better than left-right

As expected, masked language models are not good at unconditional generation

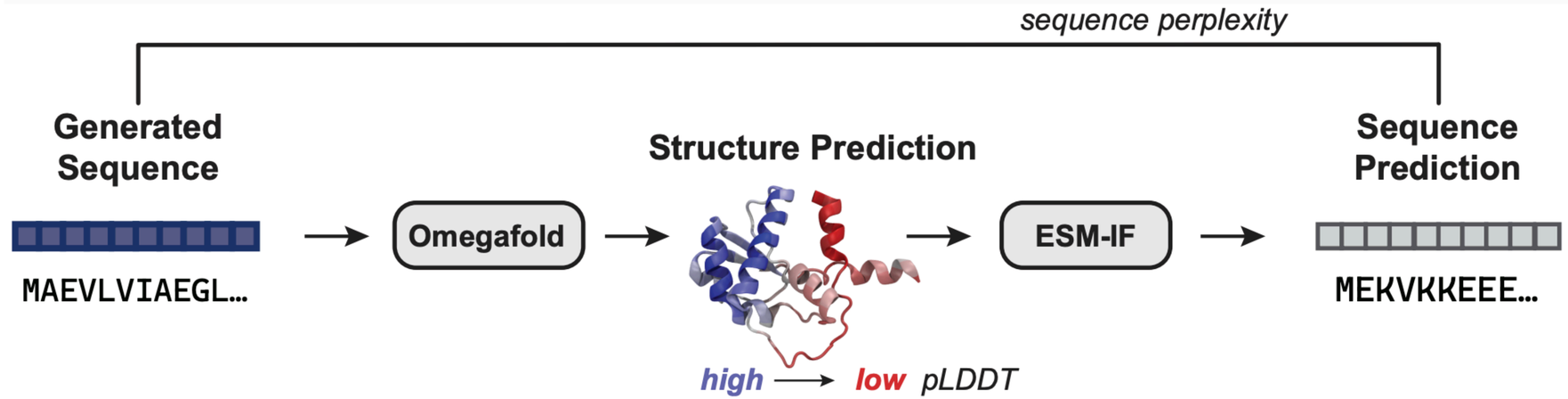
generated residues

Diffusion models recapitulate secondary structure distributions



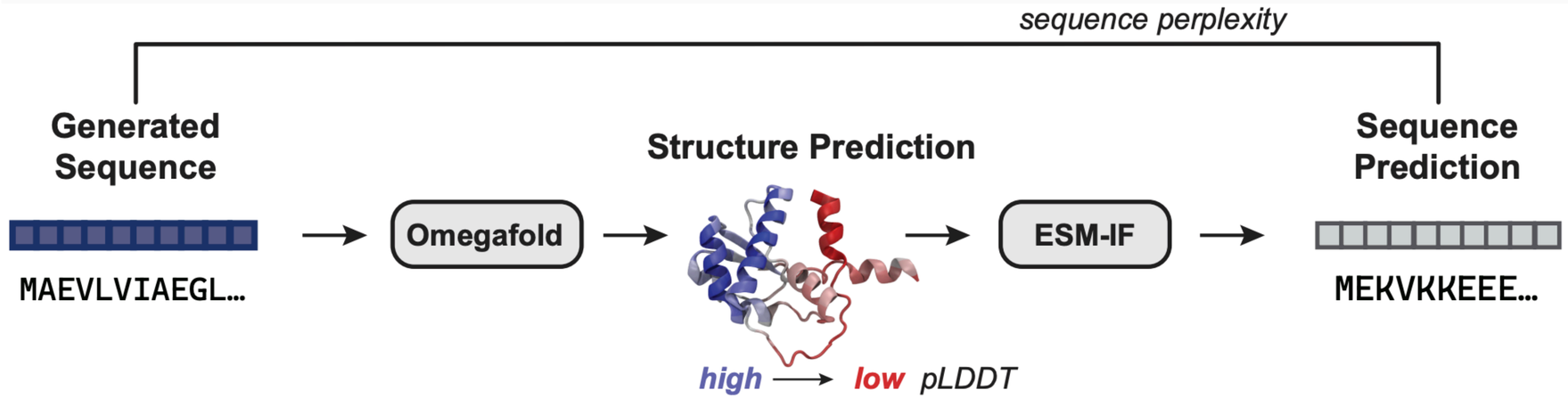
structural motifs

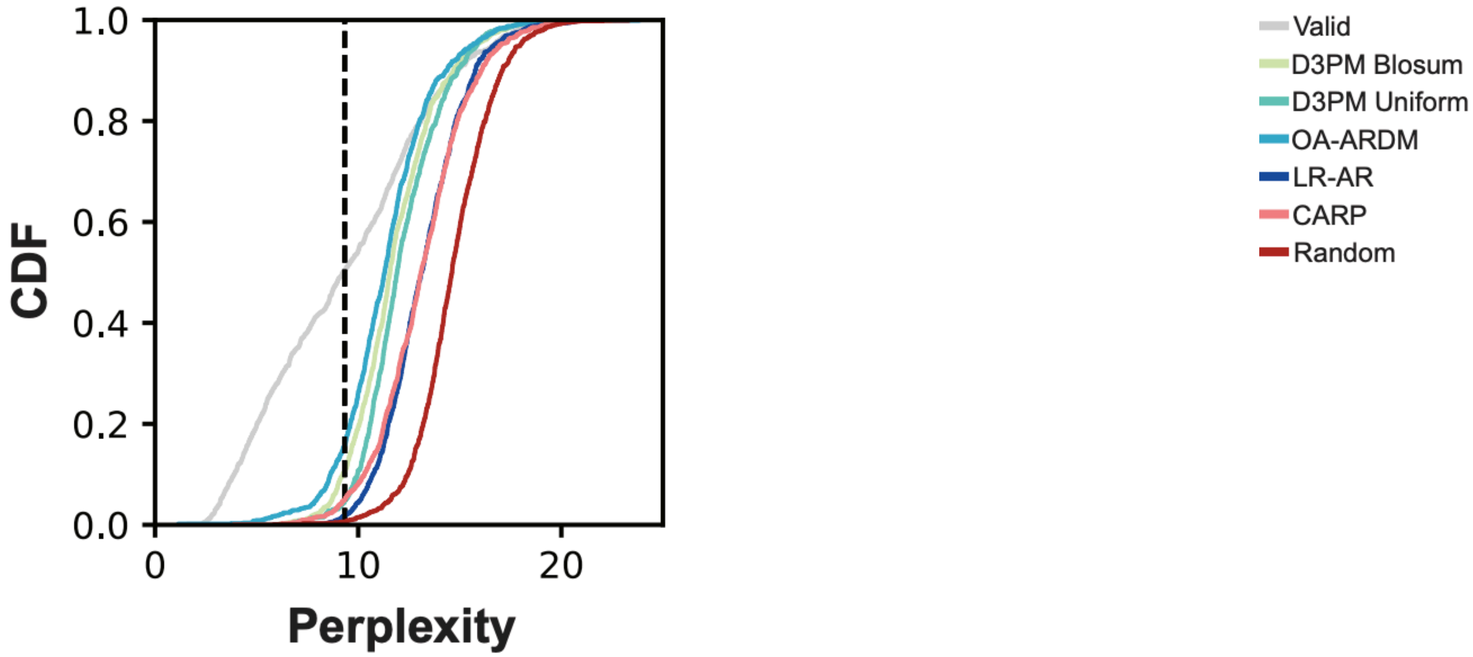
Evaluate structural confidence and self-consistency



overall sequence and structure

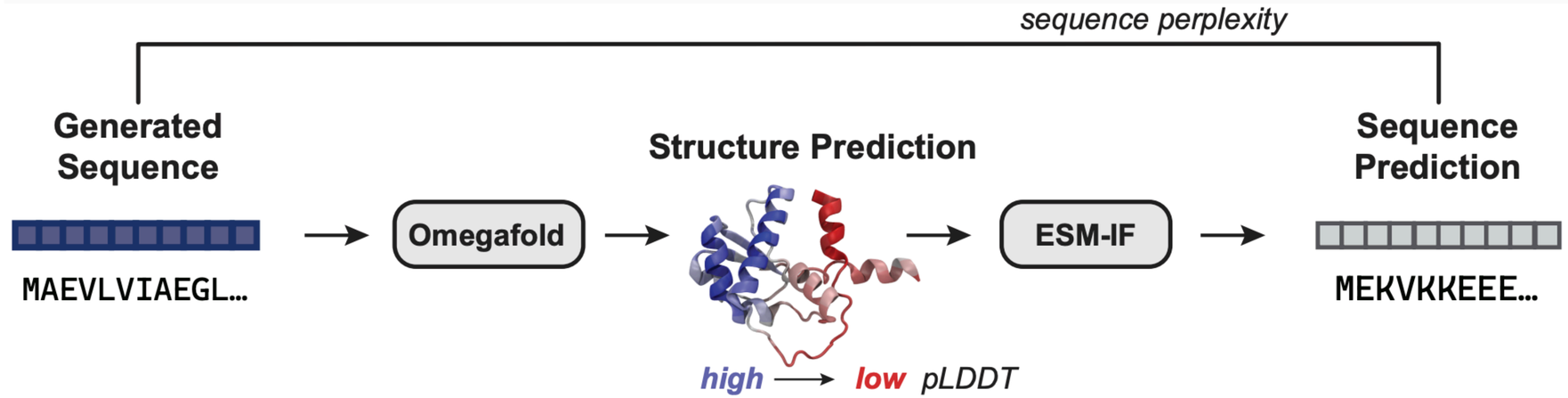
OA-ARDM produces more structured and self-consistent sequences





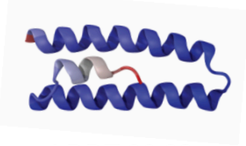
overall sequence and structure

OA-ARDM produces more structured and self-consistent sequences



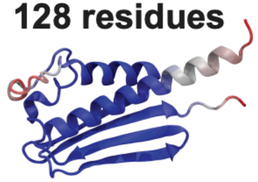
Model	Foldable Structs	Low-Perplexity Seqs	High-fidelity Samples	Minimum Hamming
Valid	506	505	267	0.016
Random	54	6	4	0.73

64 residues



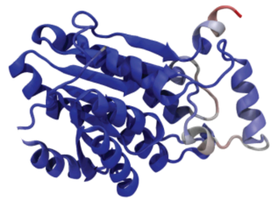
pLDDT 90.28
perplexity 6.18
Hamming 0.73

128 residues



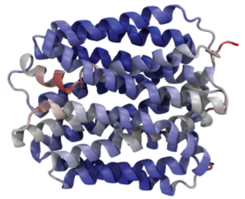
pLDDT 85.06
perplexity 6.09
Hamming 0.83

256 residues



pLDDT 89.84
perplexity 3.76
Hamming 0.71

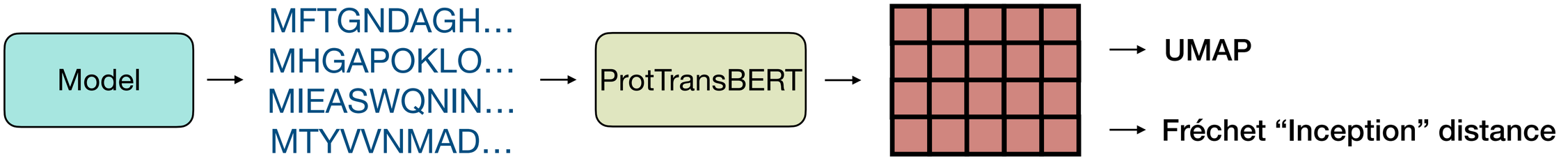
384 residues

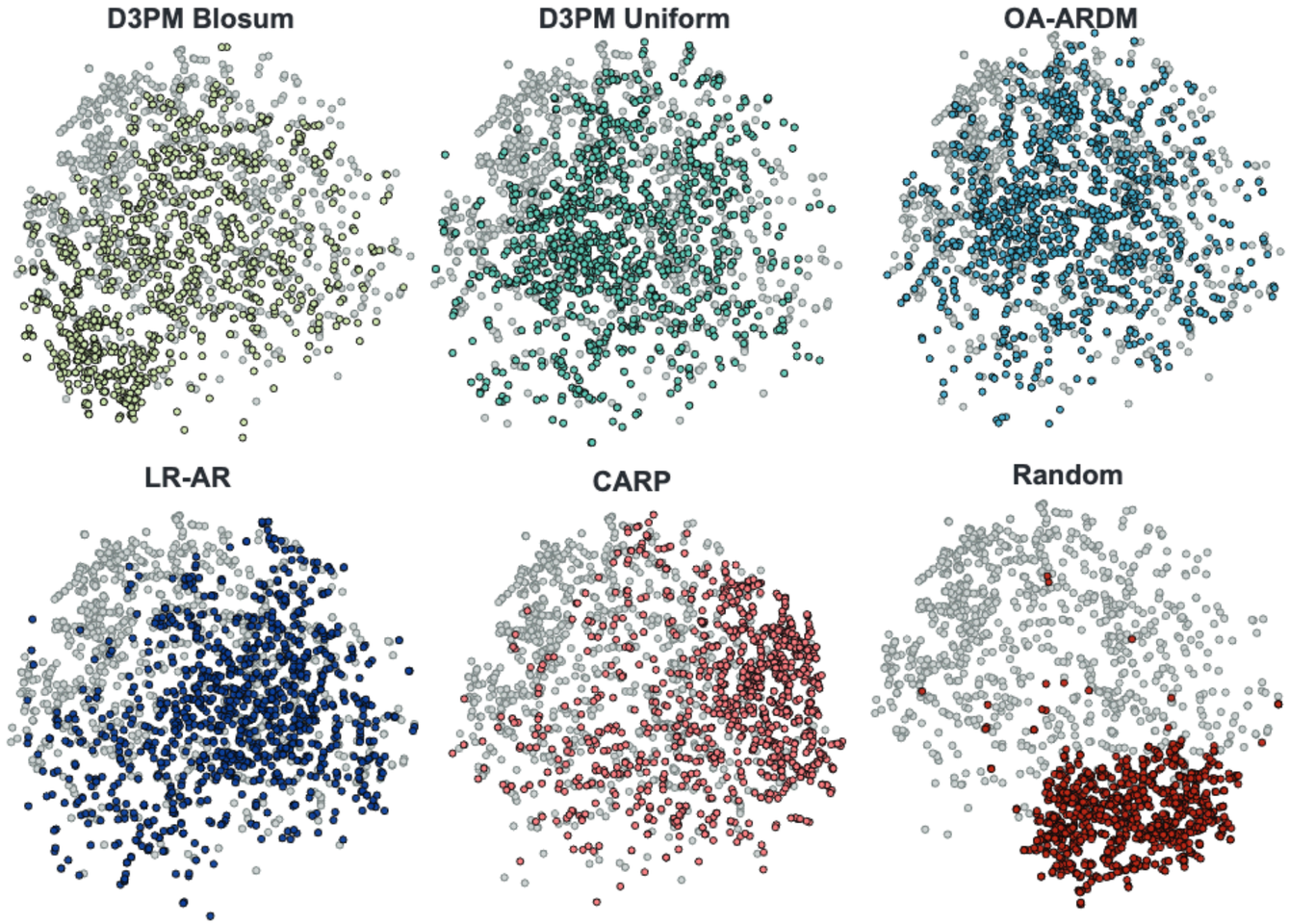


pLDDT 75.56
perplexity 6.30
Hamming 0.85

overall sequence and structure

OA-ARDM better recapitulates the natural sequence distribution

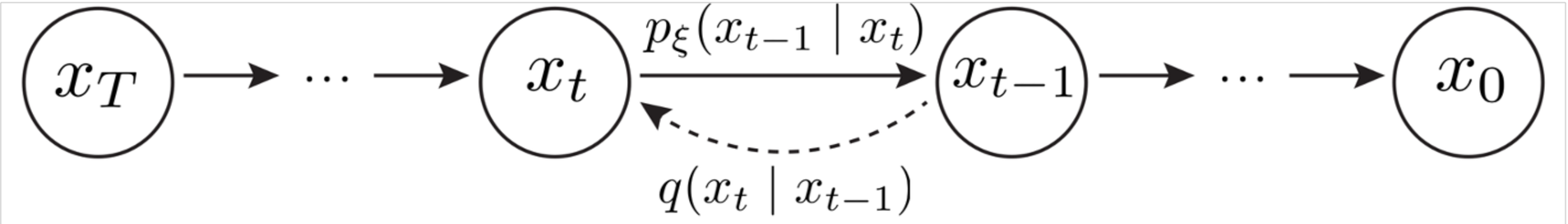




Model	FID
D3PM Blosum	1.50
D3PM Uniform	1.36
OA-ARDM	**0.88**
LR-AR	1.57
CARP	1.82
Random	3.16

distribution match

Generate multiple sequence alignments to leverage evolutionary information



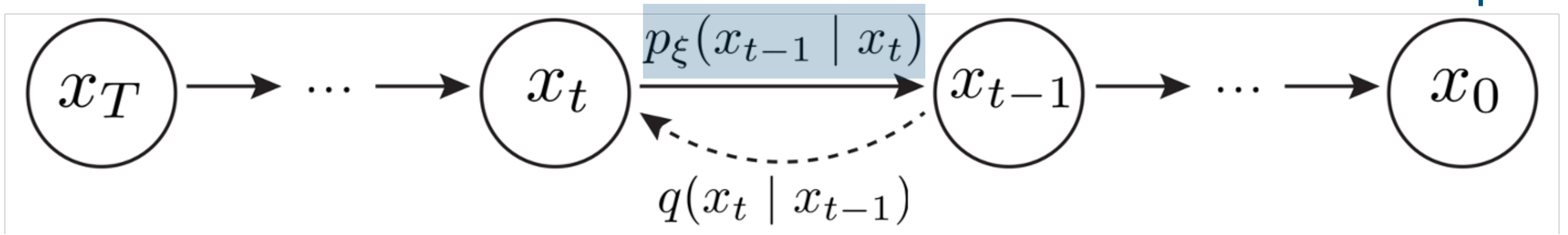
query `MFTGNDAGH…`

homology search

Generate multiple sequence alignments to leverage evolutionary information

100M parameter MSA transformer

100k MSAs from OpenFold



OA-ARDM

```
########…   MFTGN#AGH…   MFTGN#AGH…   MFTGNDAGH…
########…   MH#G-DAGL…   MH#G-DAGL…   MHTG-DAGL…
########…   MHTG#—GL…    MHTGN--GL…   MHTGN--GL…
########…   MH#NN#AGL…   MH#NN#AGL…   MHYNNDAGL…
…           …            …            …
```

mask

D3PM

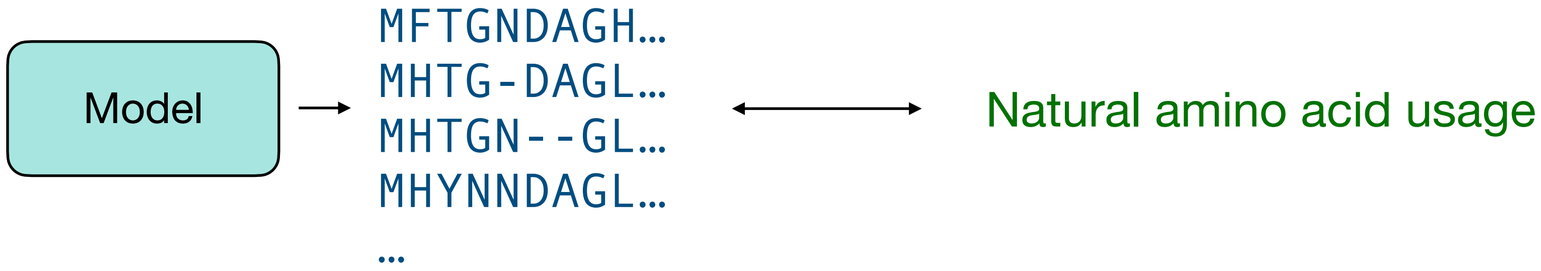
```
HQWTYIPAQ…   MFTGNDAGH…   MFTGNDAGH…   MFTGNDAGH…
PLKMNVDFW…   MHTL-DAGL…   MHTH-DAGL…   MHTG-DAGL…
MNDFHIGSD…   M-TKN—GL…    MHTKN--GL…   MHTGN--GL…
ALKENIRTA…   MHYVNDLGL…   MHYNNDLGL…   MHYNNDAGL…
…            …            …            …
```

mutate

generated residues

overall sequence and structure

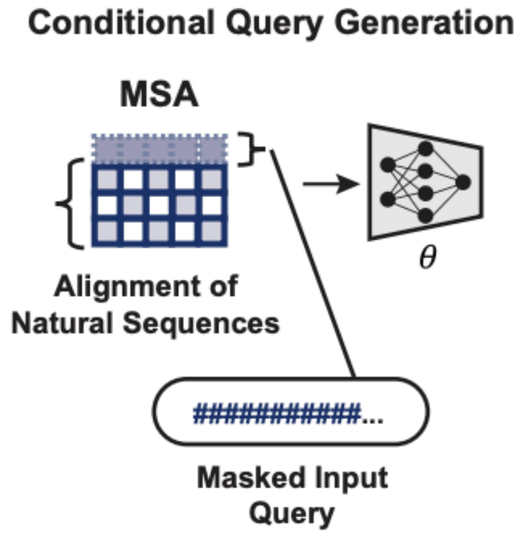
OA-ARDM better recapitulates the natural amino acid distribution

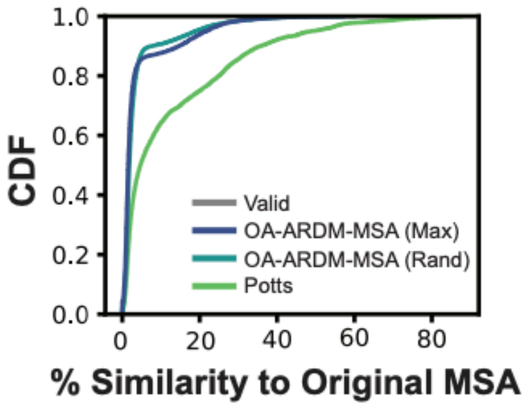


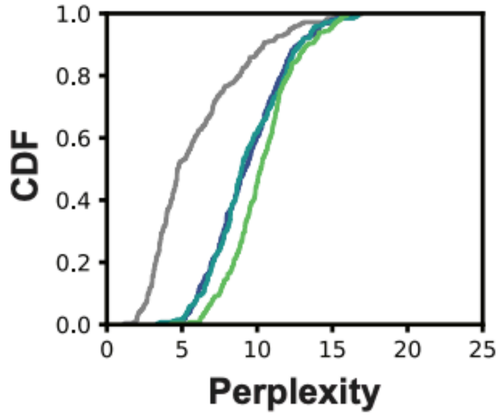
Model	Params	KL (AA+GAP)	KL (AA)
Valid	-	0.056	0.00093
ESM-MSA-1b (Rao et al., 2021)	100M	>1	> 1

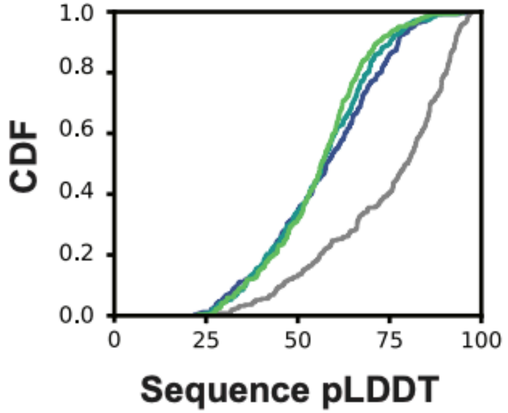
generated residues

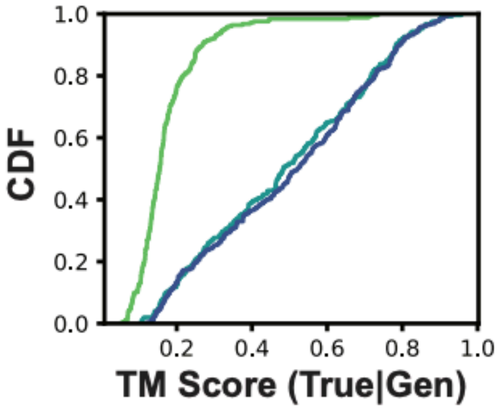
OA-ARDM can generate realistic queries conditioned on the MSA





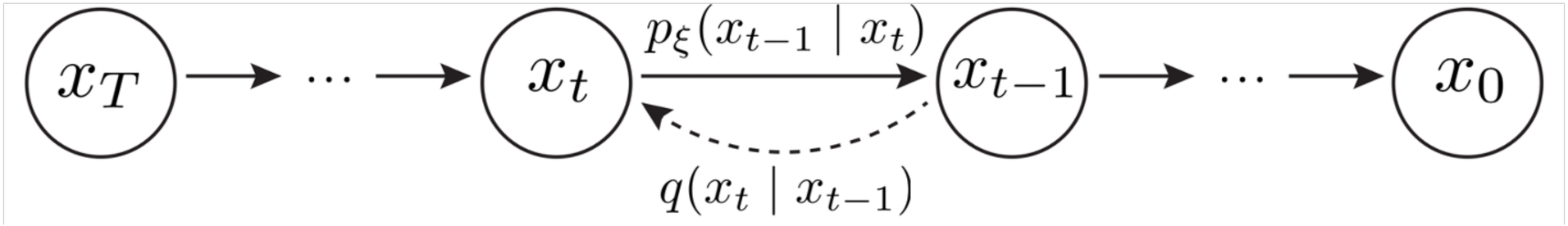


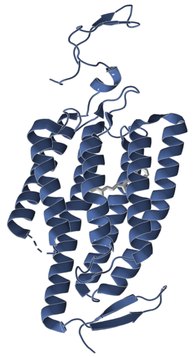




overall sequence and structure

We use diffusion models to generate structure, sequence, and MSA



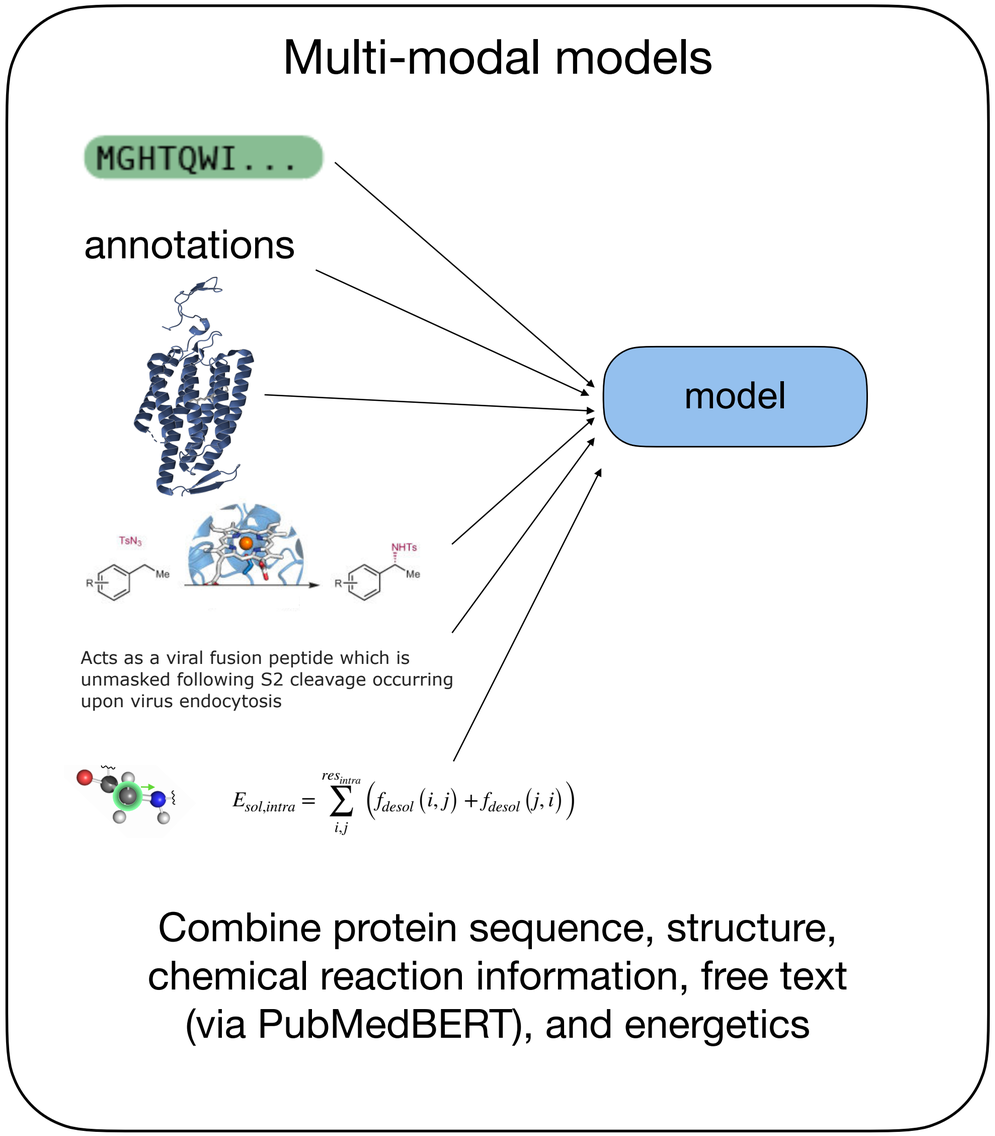


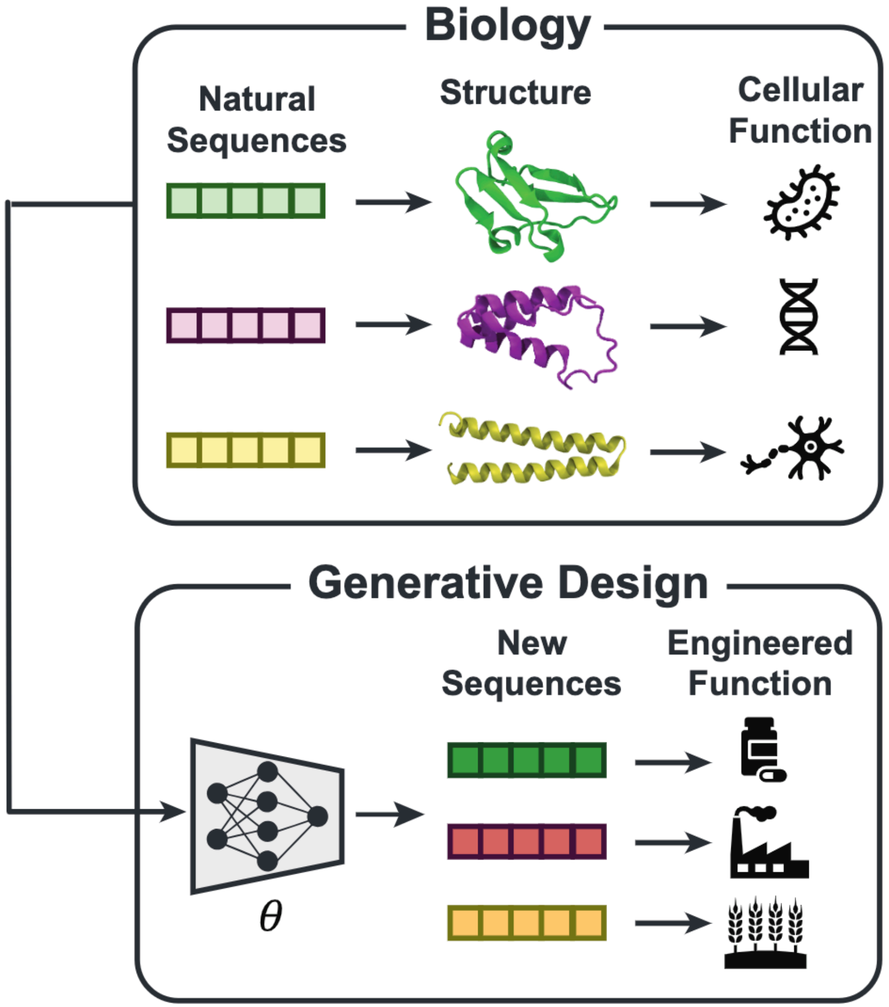
MFTGNDAGH…
MHGAPOKLO…
MIEASWQNIN…
MTYVVNMAD…

MFTGNDAGH…
MHTG-DAGL…
MHTGN--GL…
MHYNNDAGL…
…

Eventually always need sequences!

Can we generate proteins with new functions?





Acknowledgments

BioML at MSR New England