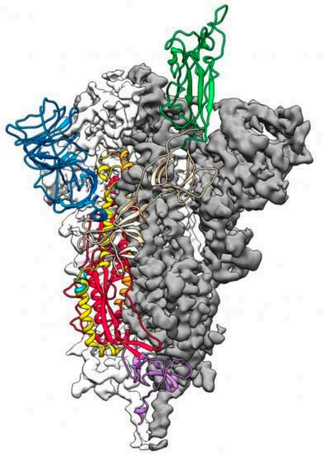
Protein generation with diffusion

Kevin Kaichuang Yang
Microsoft Research New England
@KevinKaichuang

Proteins are biology’s actuators

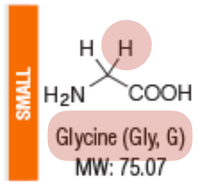
- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling

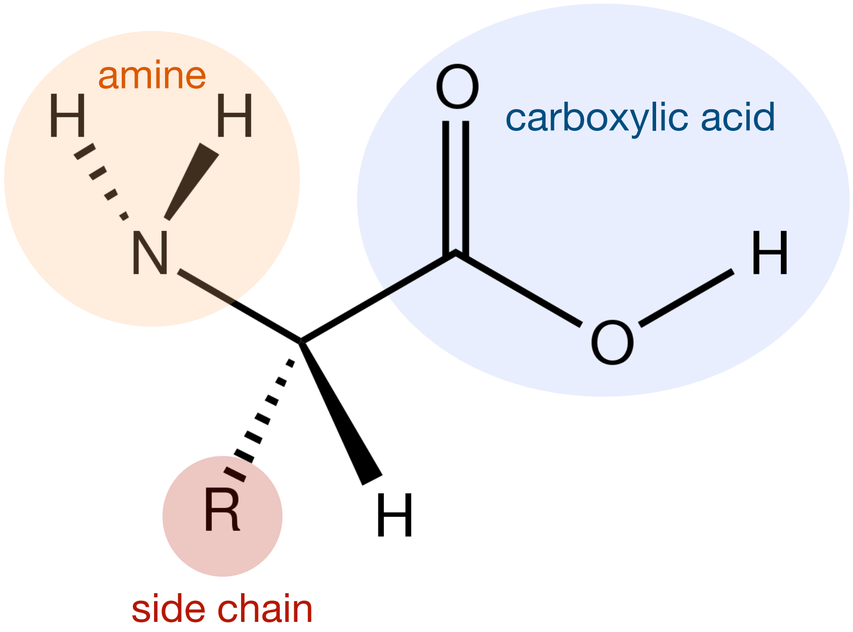


coronavirus spike protein

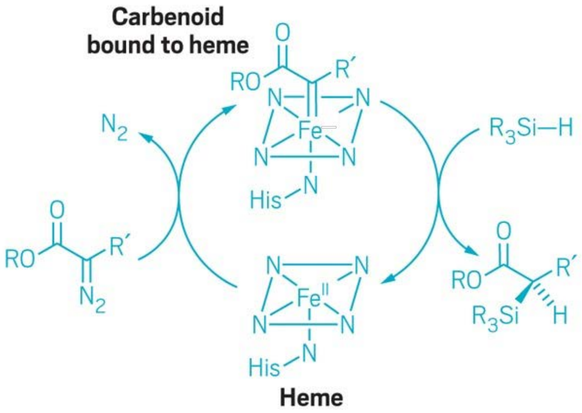
luciferase

Diversity arises from 20 building blocks

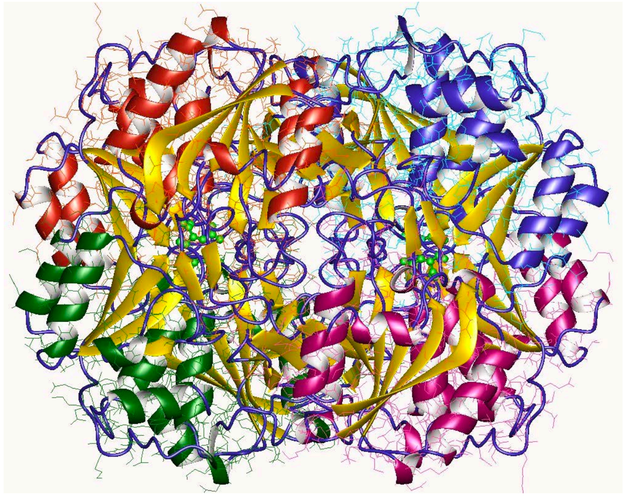




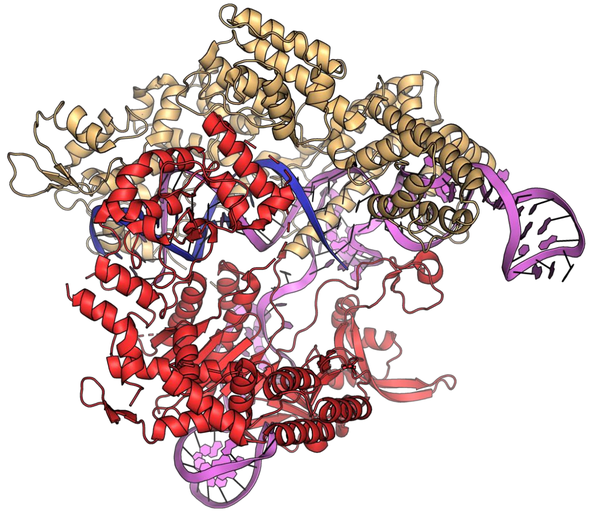
We need proteins with new functions



new chemistry

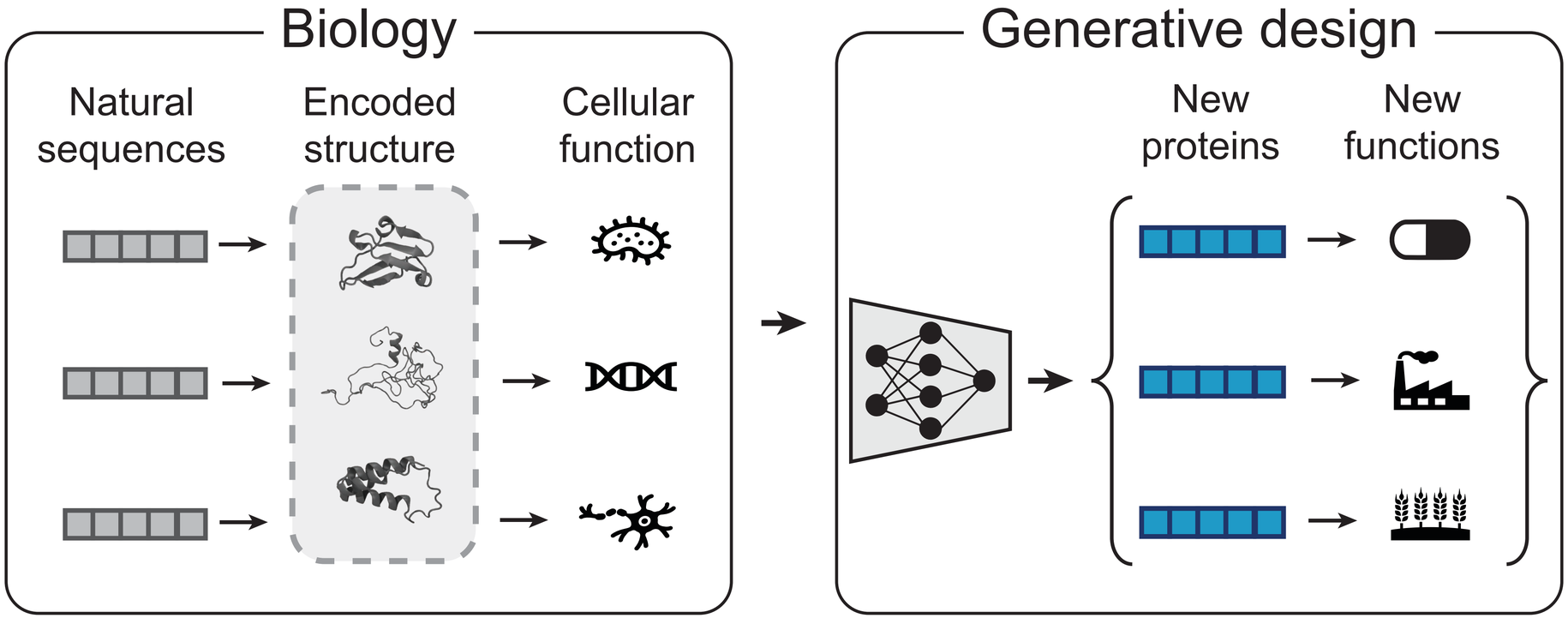


therapeutics



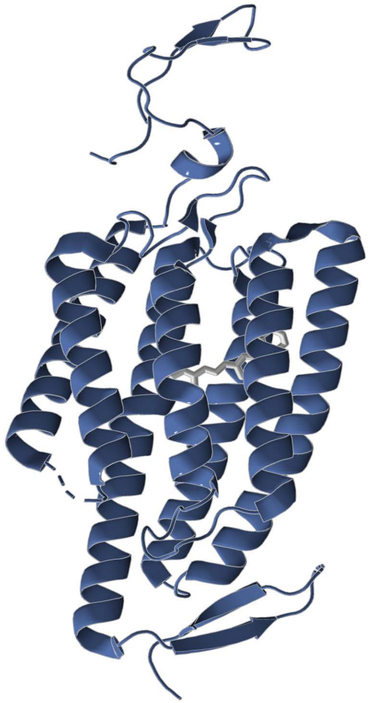
molecular tools

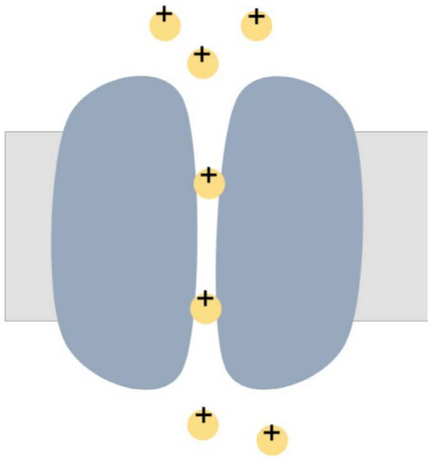
Generate new proteins to expand functional space



Generating new, designable structures expands functional space

MGTGDHDD...





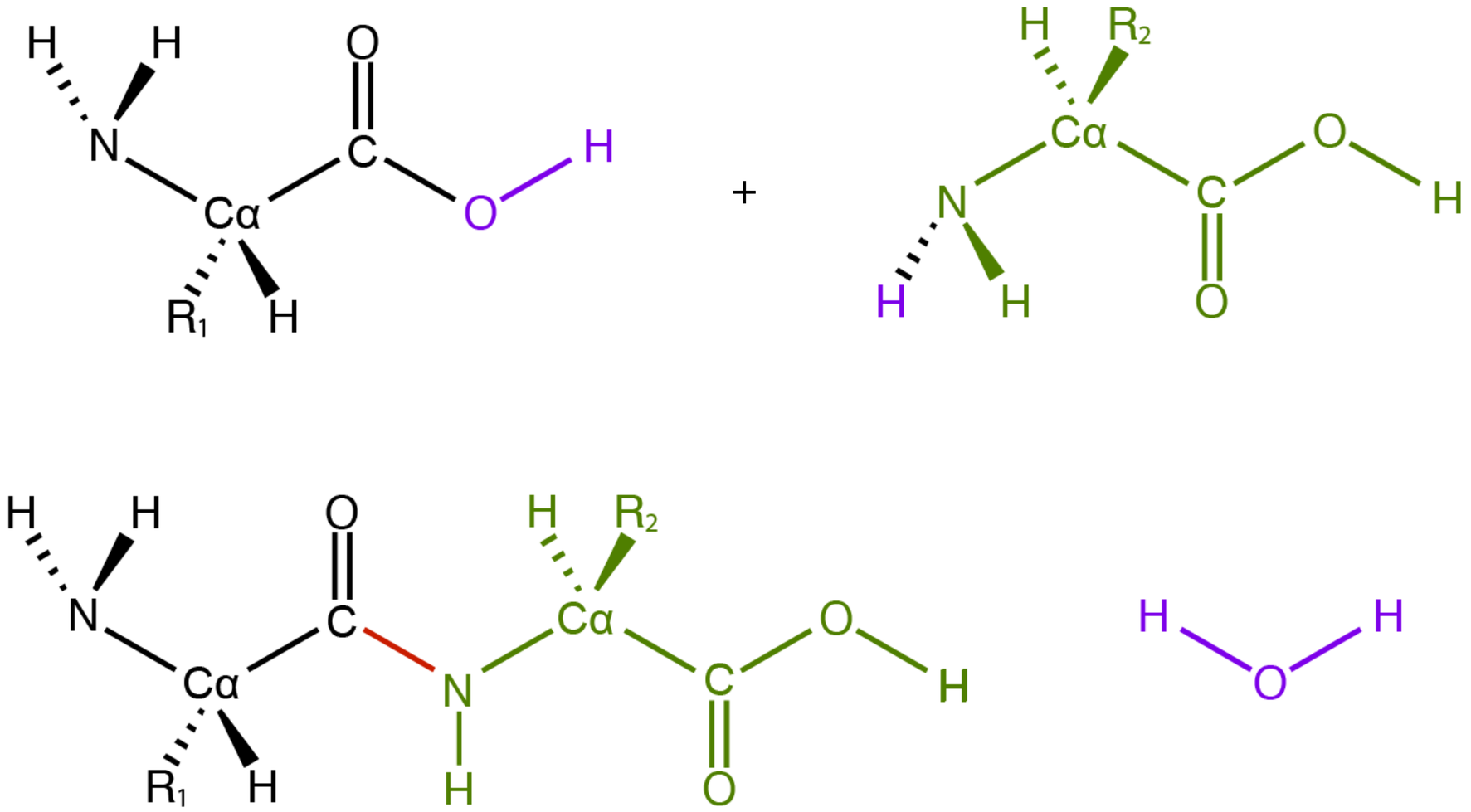
Challenge: Generate diverse and designable structures



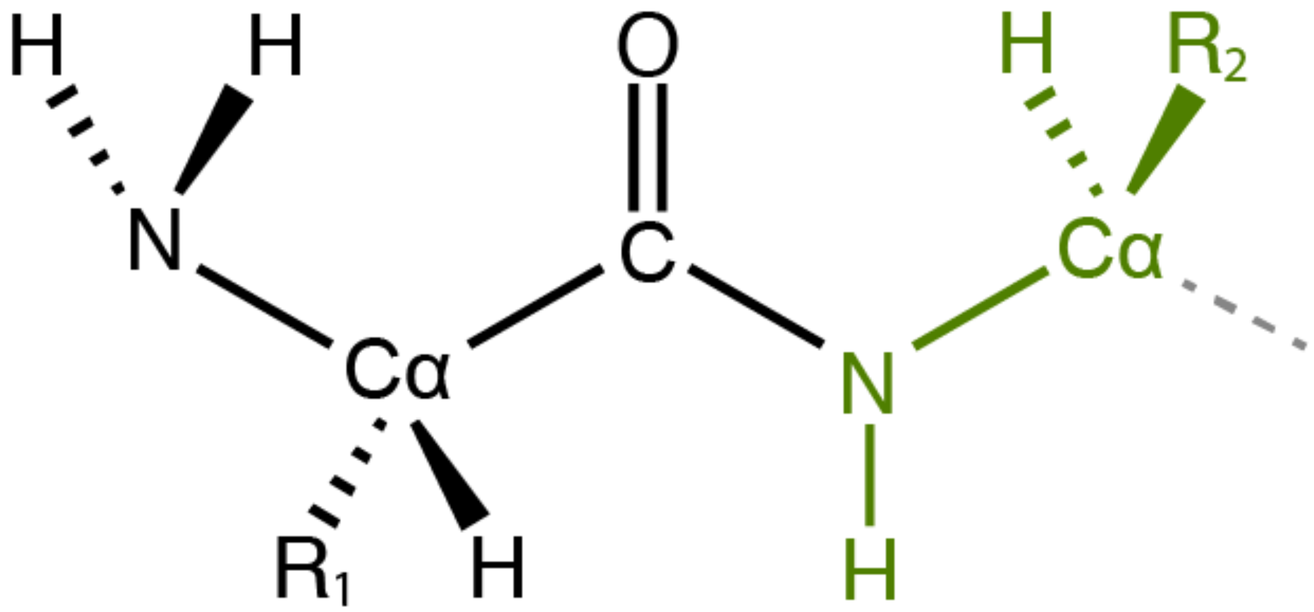


Wu, Yang, van den Berg, Alamdari, Zou, Lu, Amini; *Nature Comm* 2024

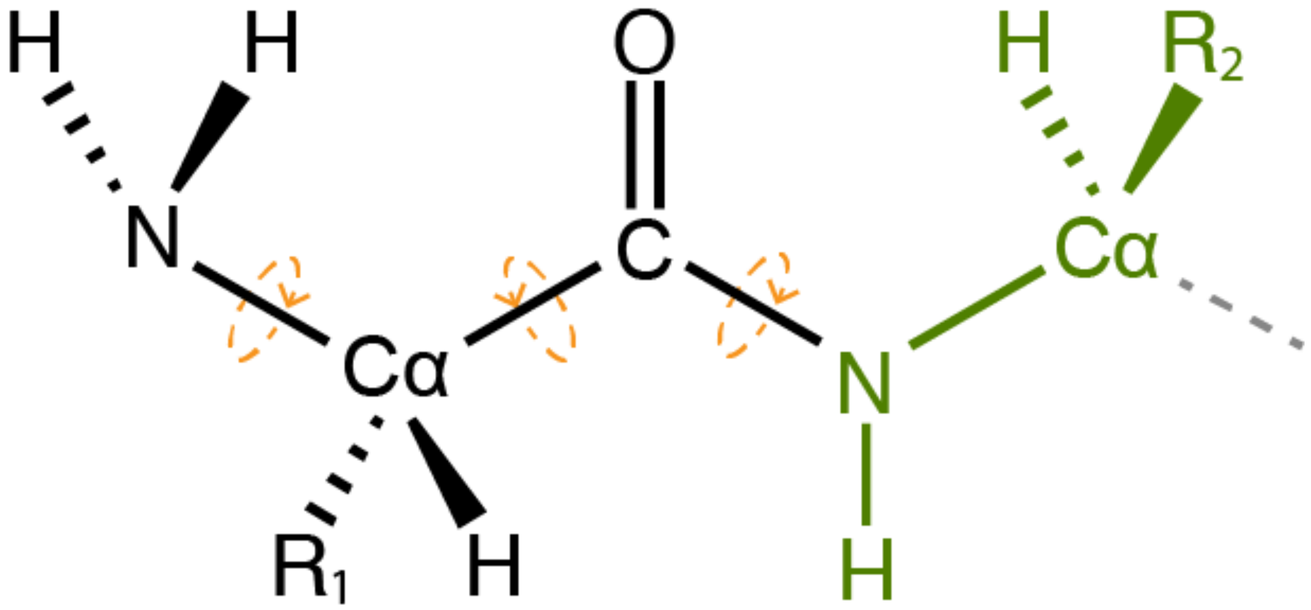
Proteins are polypeptide chains



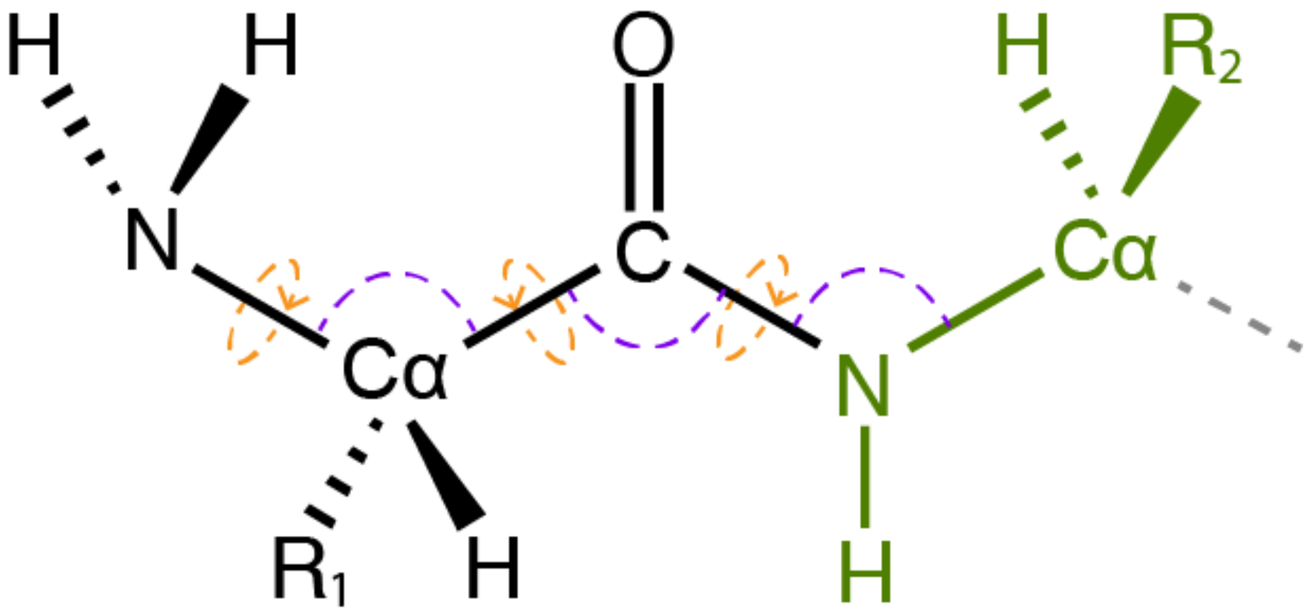
Protein structure is determined by bond angles



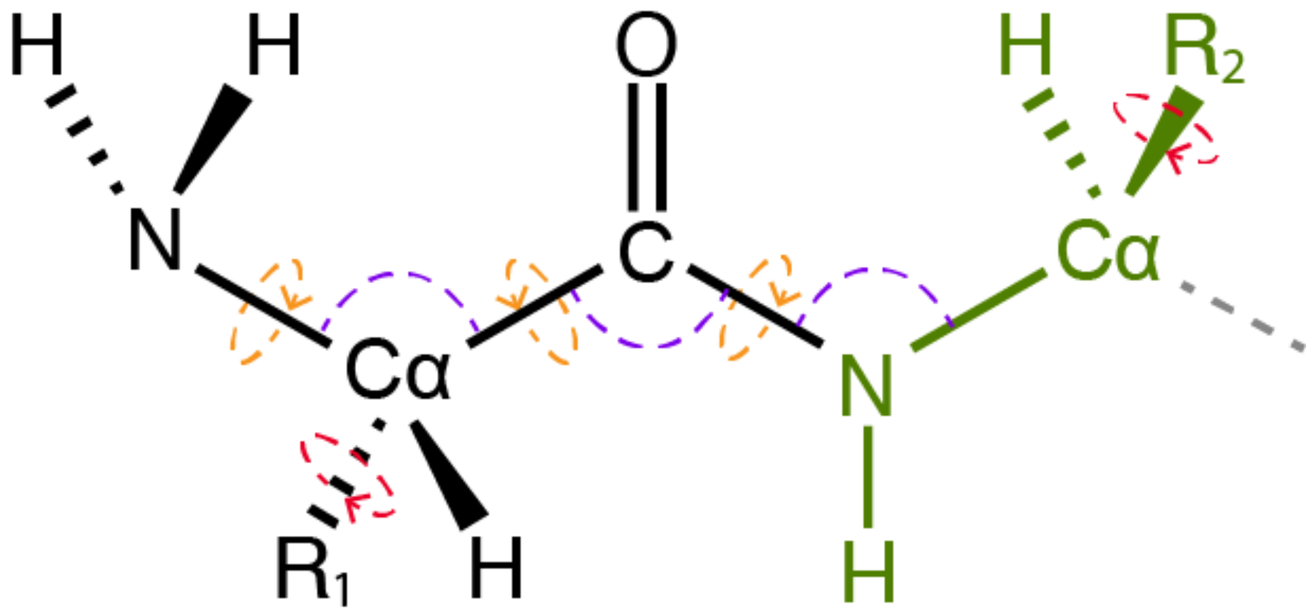
Protein structure is determined by bond angles



Protein structure is determined by bond angles

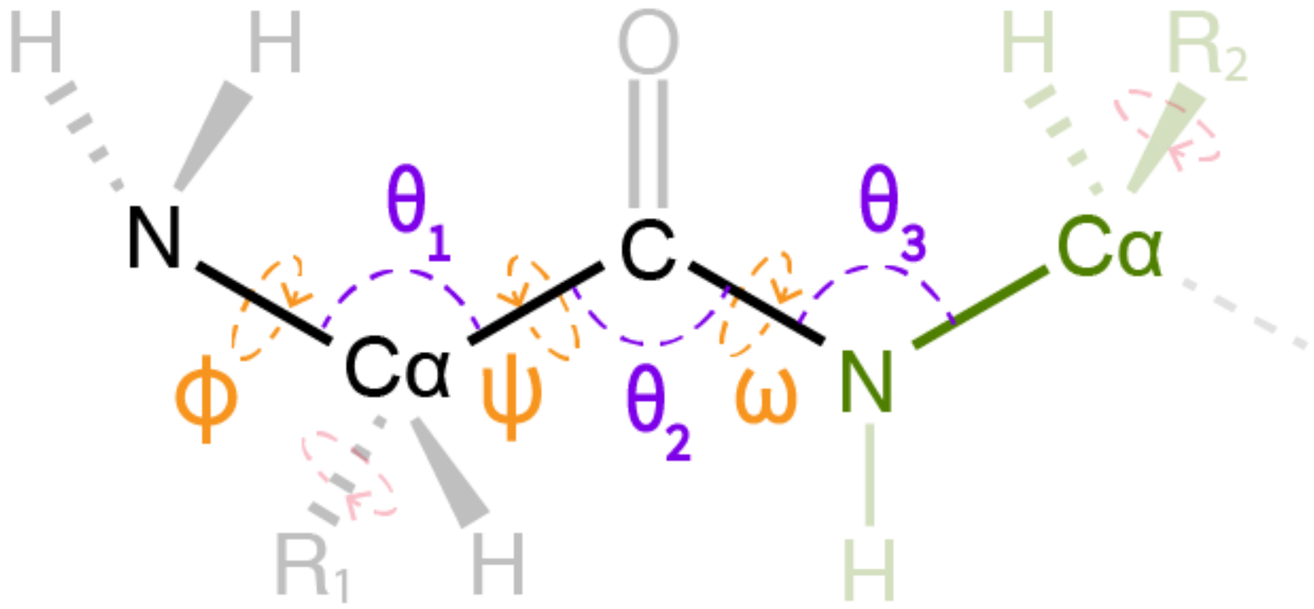


Protein structure is determined by bond angles



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

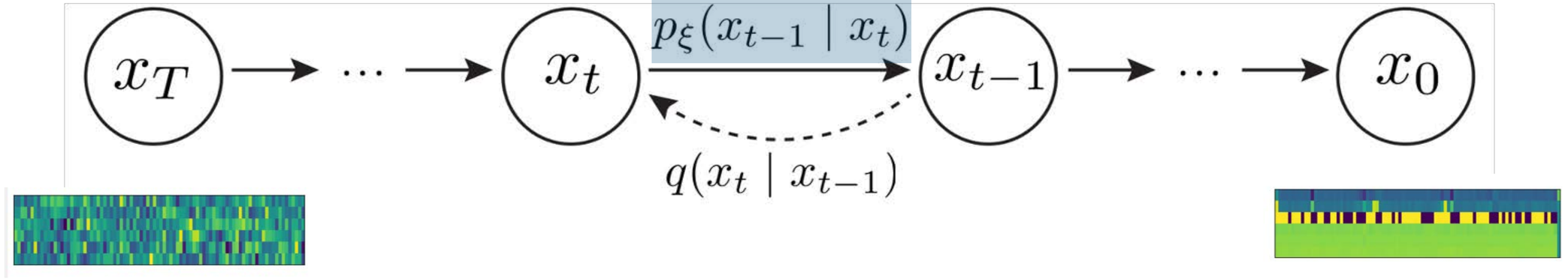
We generate backbone structures represented by dihedral and bond angles



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

FoldingDiff uses diffusion to generate angles

Vanilla transformer

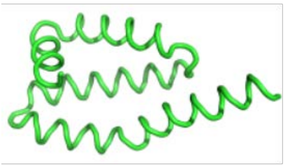


random angles

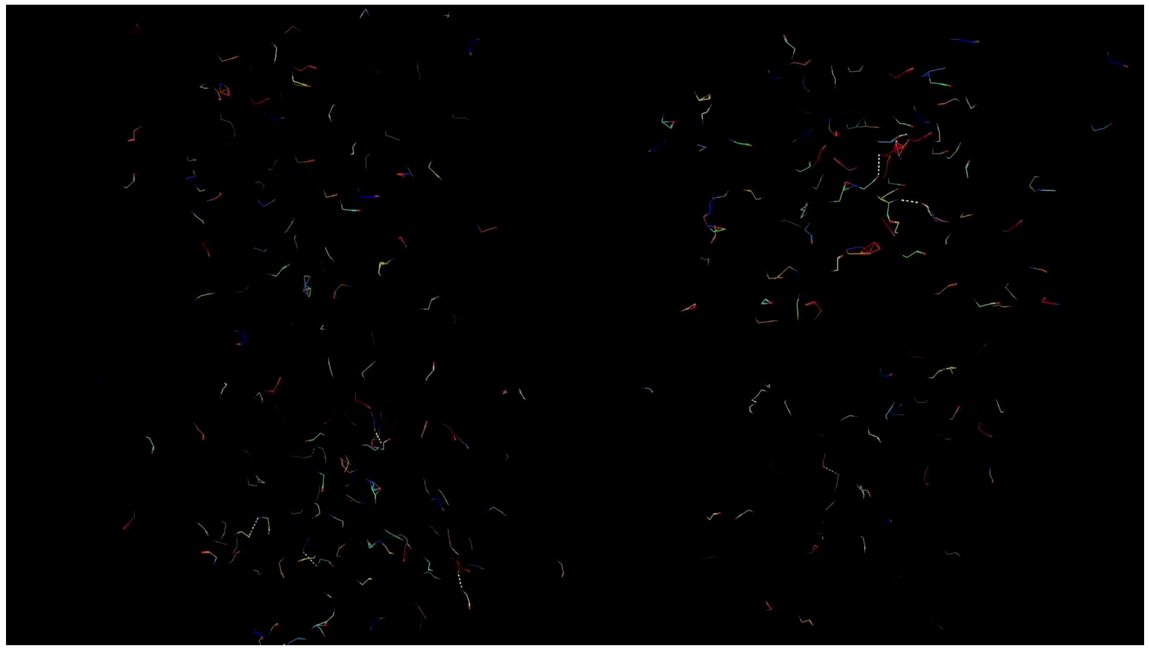
30k structures from CATH

40 - 128 residues

generated angles

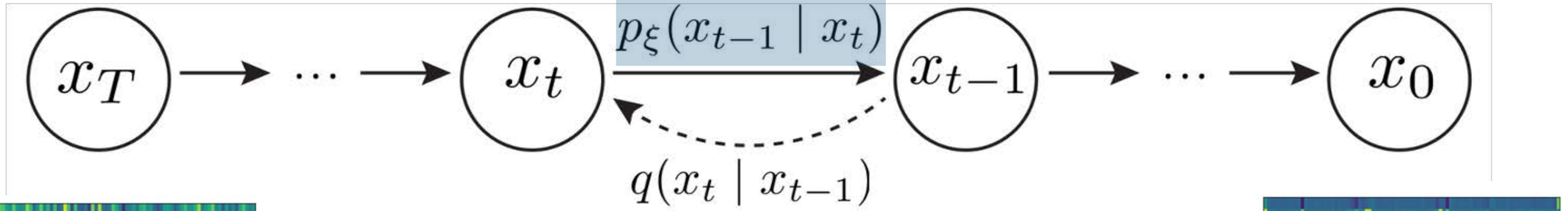


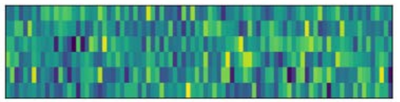
Diffusion on 3D coordinates requires equivariances



Evaluate generations at 3 levels

Vanilla transformer

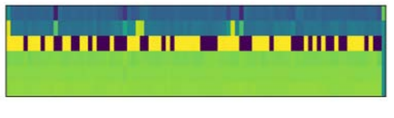




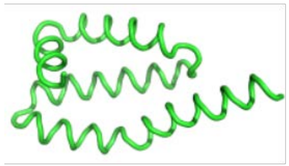
random angles

30k structures from CATH

40 - 128 residues



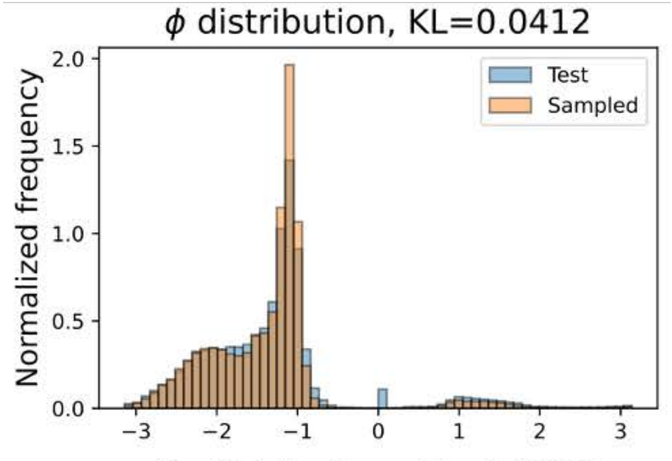
generated angles



generated angles | **structural motifs** | **overall structures**

Generated angles match test distribution

noise → sample → compare to true distribution

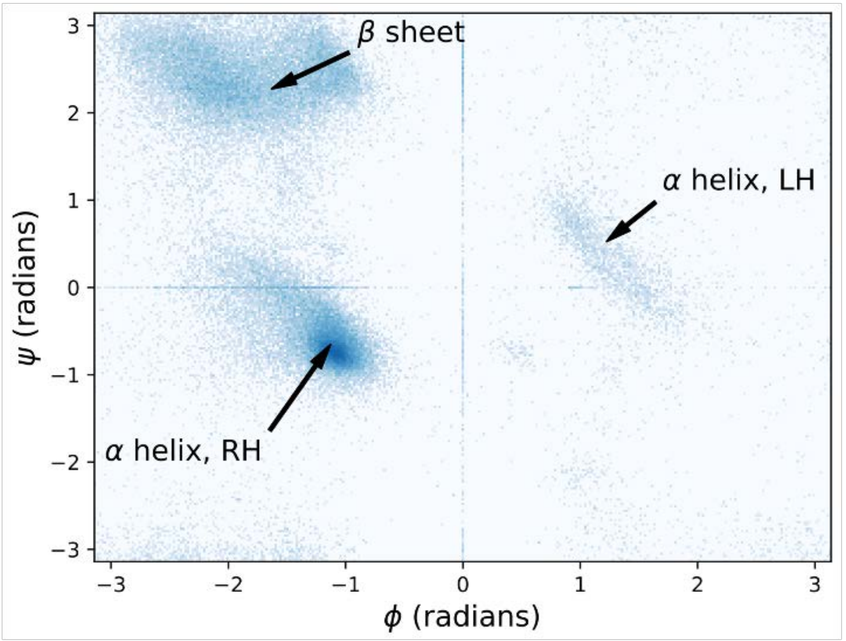


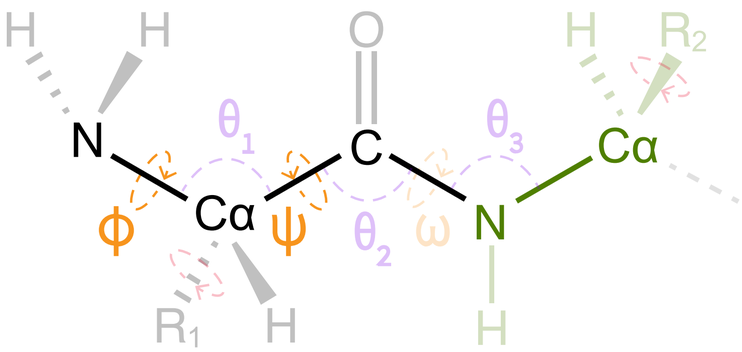
Generated distributions match natural distribution of individual angles

FoldingDiff captures correlations between angles

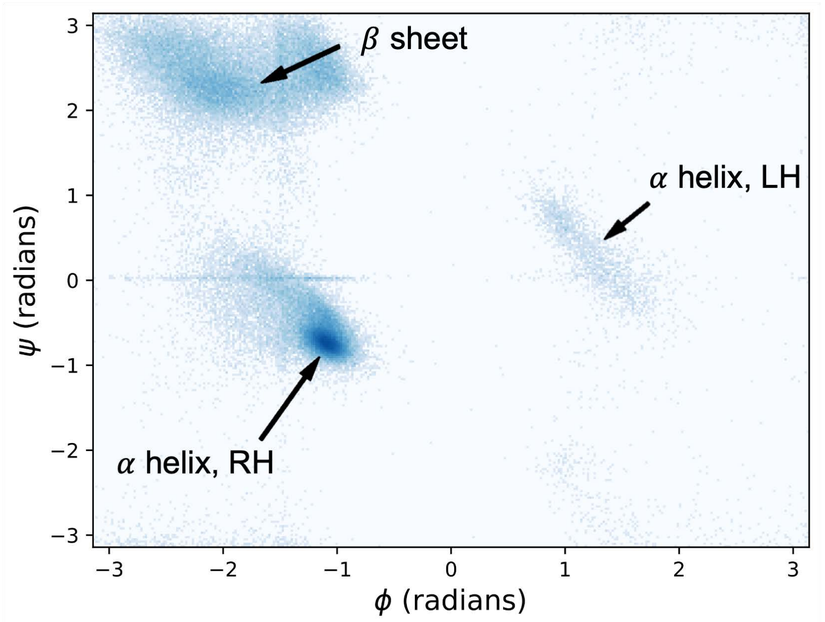
noise ⟶ sample ⟶ compare (ϕ, ψ) co-occurence

True



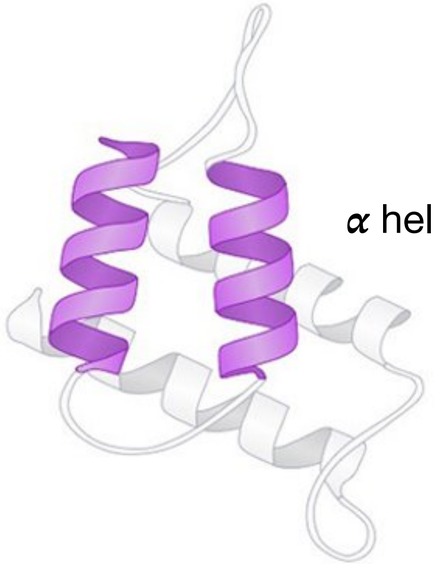


Generated

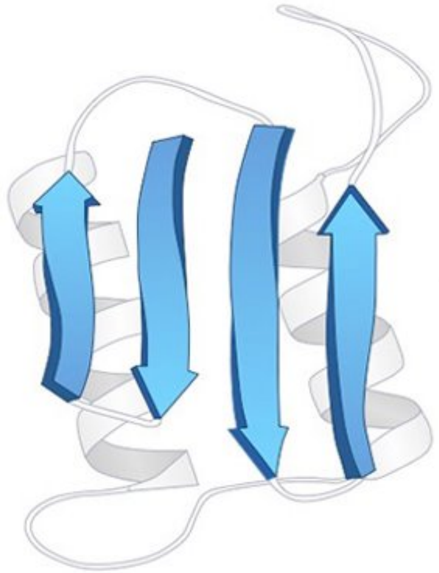


FoldingDiff generates correlations that define common structural motifs

Generated secondary structures match test structures



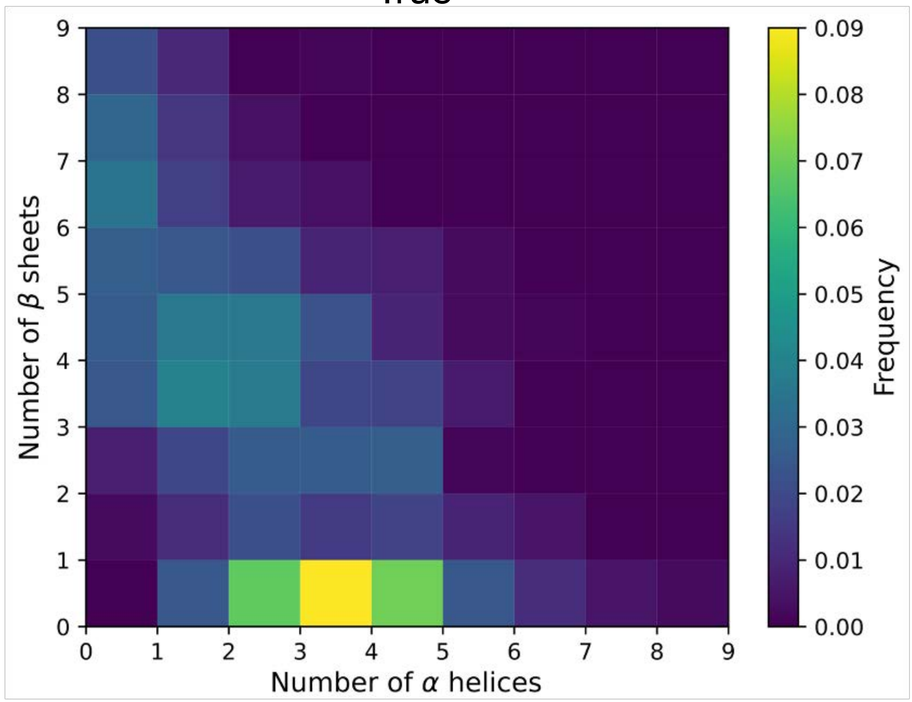
α helix



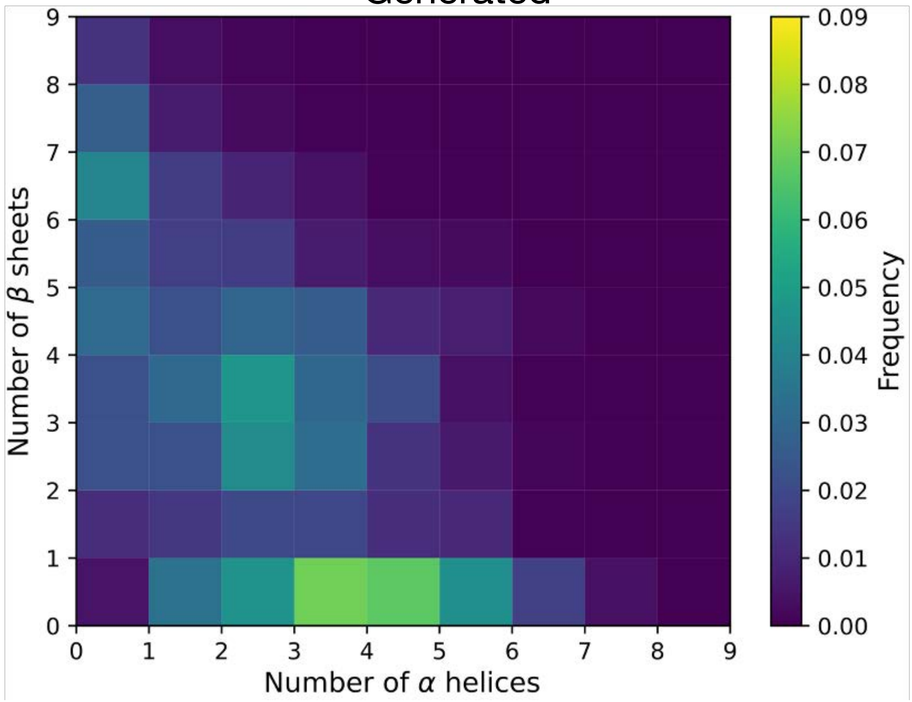
β sheet

noise ⟶ sample ⟶ measure helix, sheet structures

True

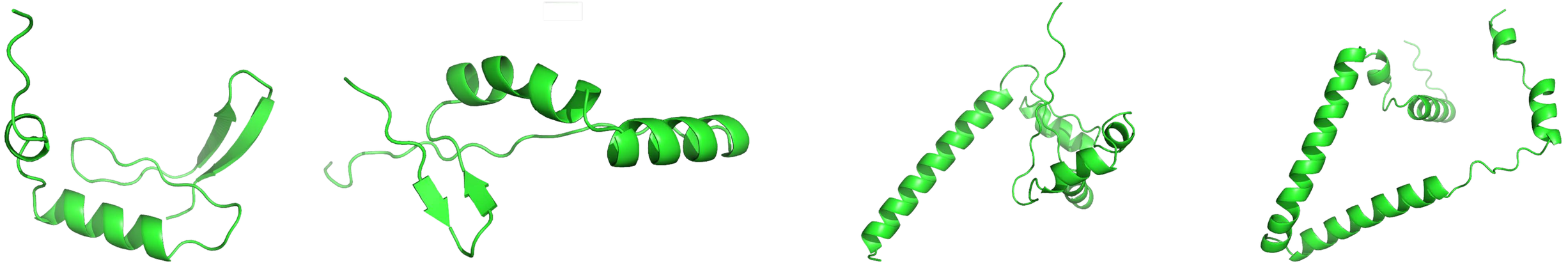


Generated

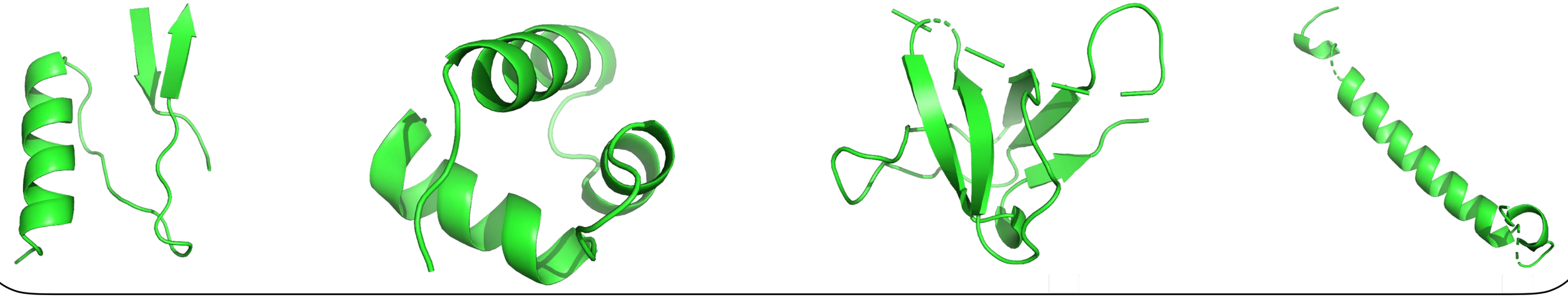


Generated structures look reasonable

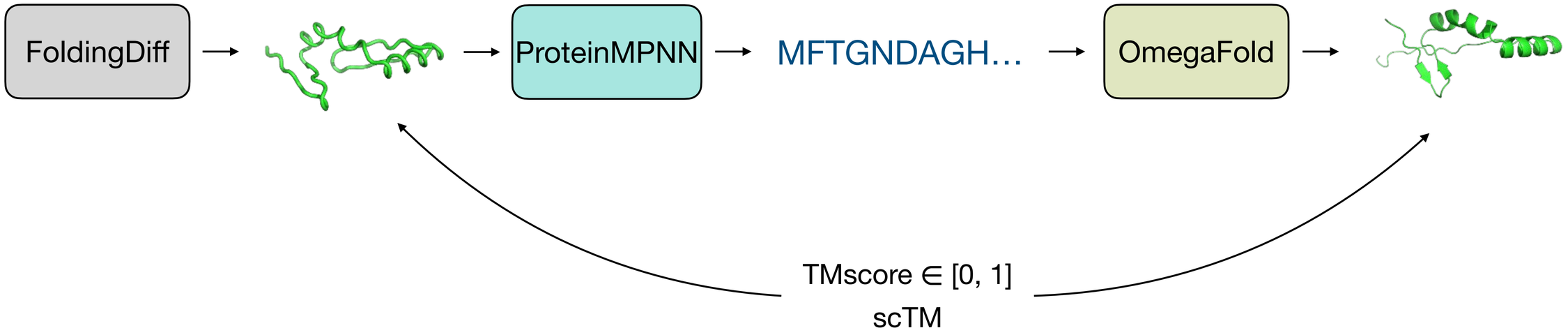
Generated structures



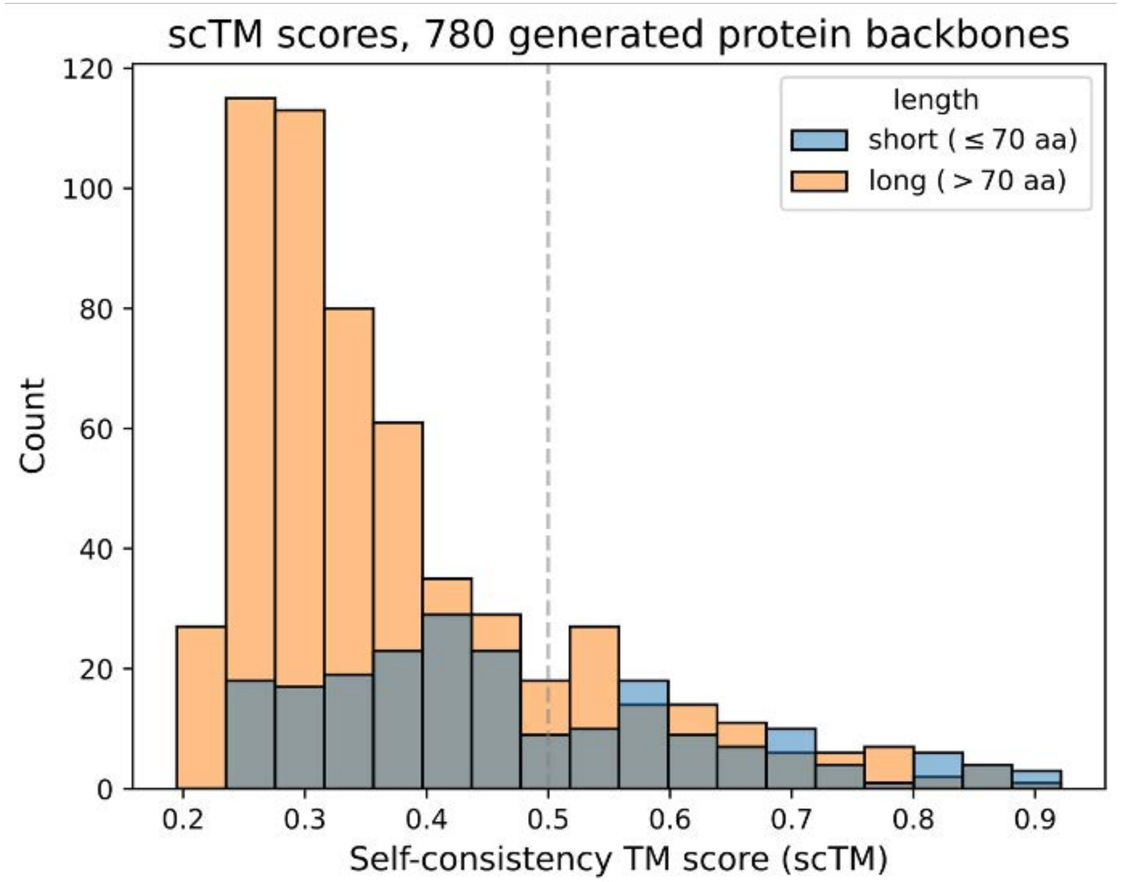
Training set structures



Measure designability of structures with self-consistency TMscore



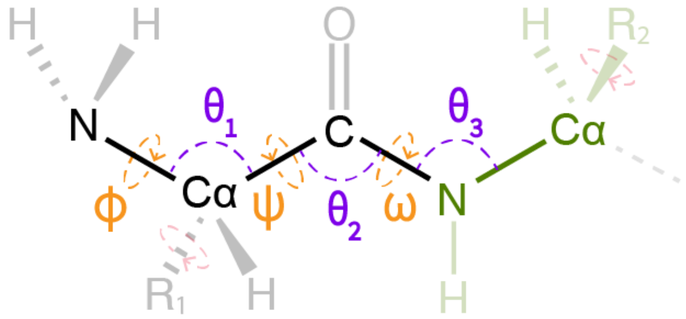
Many generated structures are designable



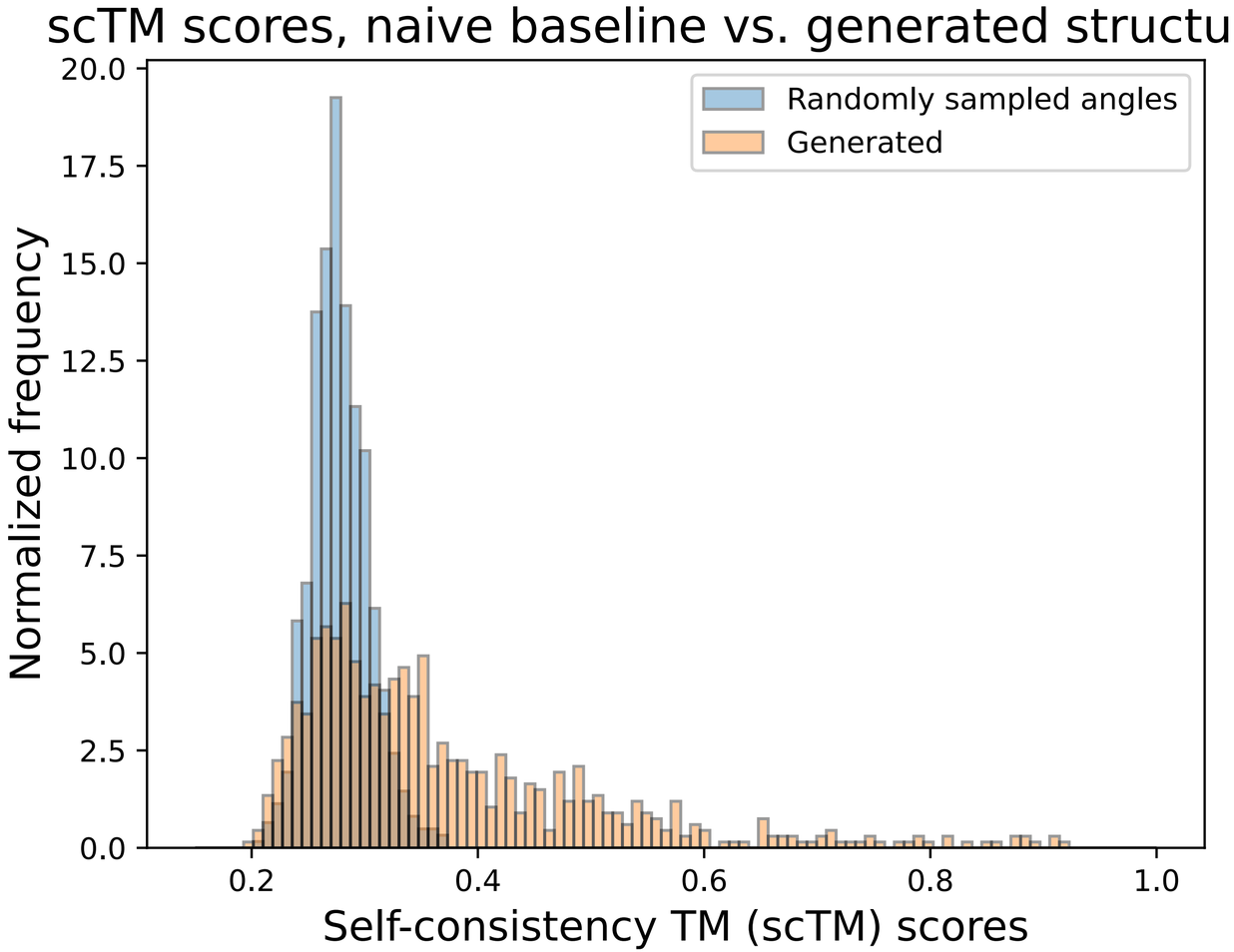
scTM > 0.5	≤ 70 aa	> 70 aa
FoldingDiff	76/210	87/570
ProtDiff (Trippe *et al.*)	36/210	56/570

Significant improvements over point cloud diffusion model

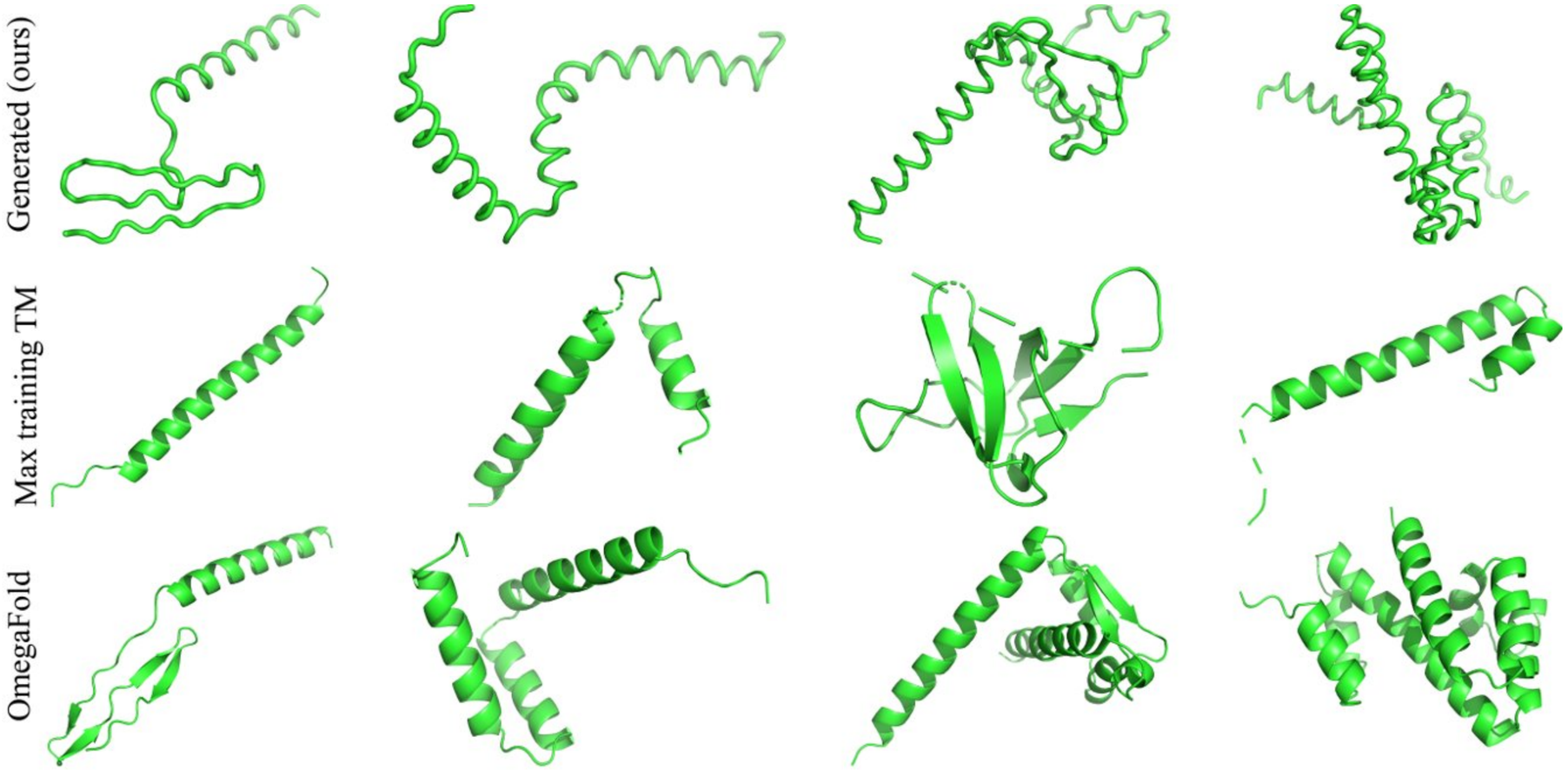
FoldingDiff structures are better than random baseline



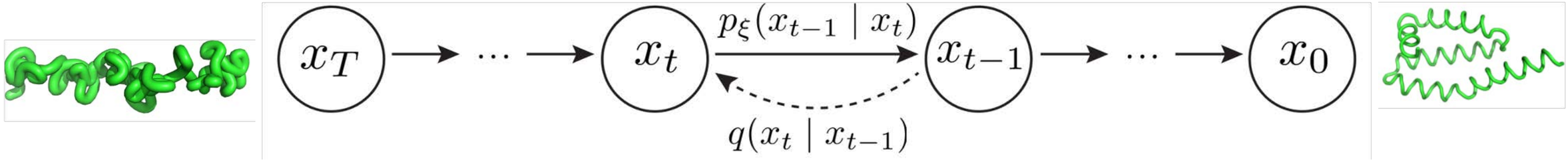
Sample sets of angles
Preserves Ramachandran plot



Generated structures are diverse

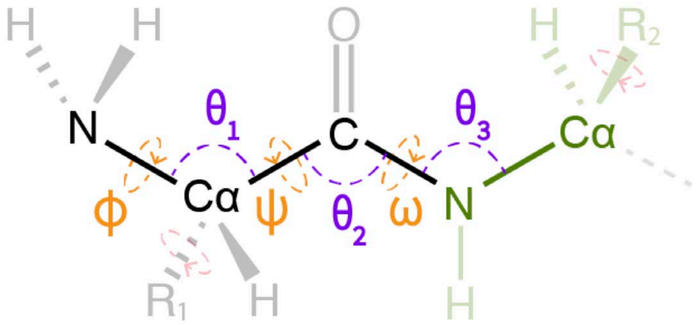


FoldingDiff is first step towards generating new functions

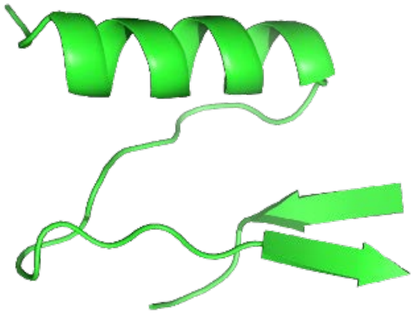


Generate protein backbones by mirroring the folding process

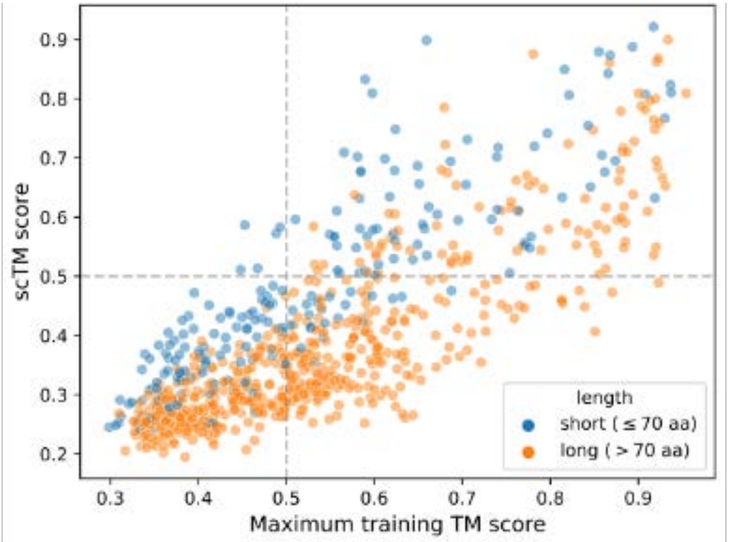
Internal coordinates



Structural motifs



Realistic, diverse samples



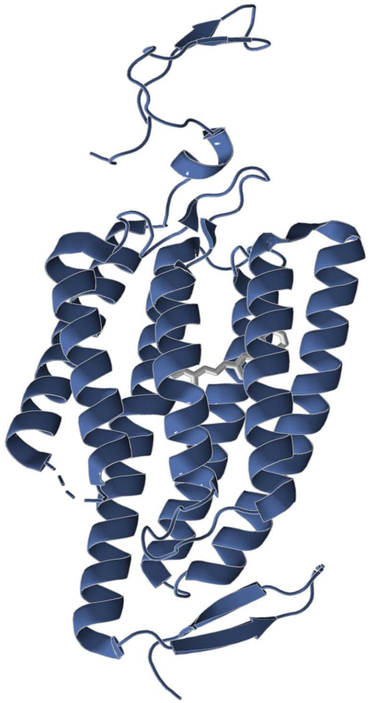
github.com/microsoft/foldingdiff

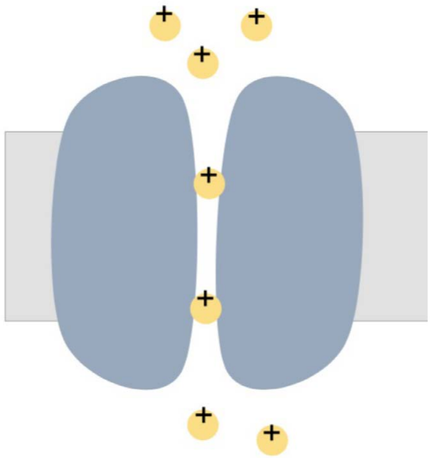
Paper: https://doi.org/10.1038/s41467-024-45051-2

huggingface.co/spaces/wukevin/foldingdiff

Sequence is the universal protein design space

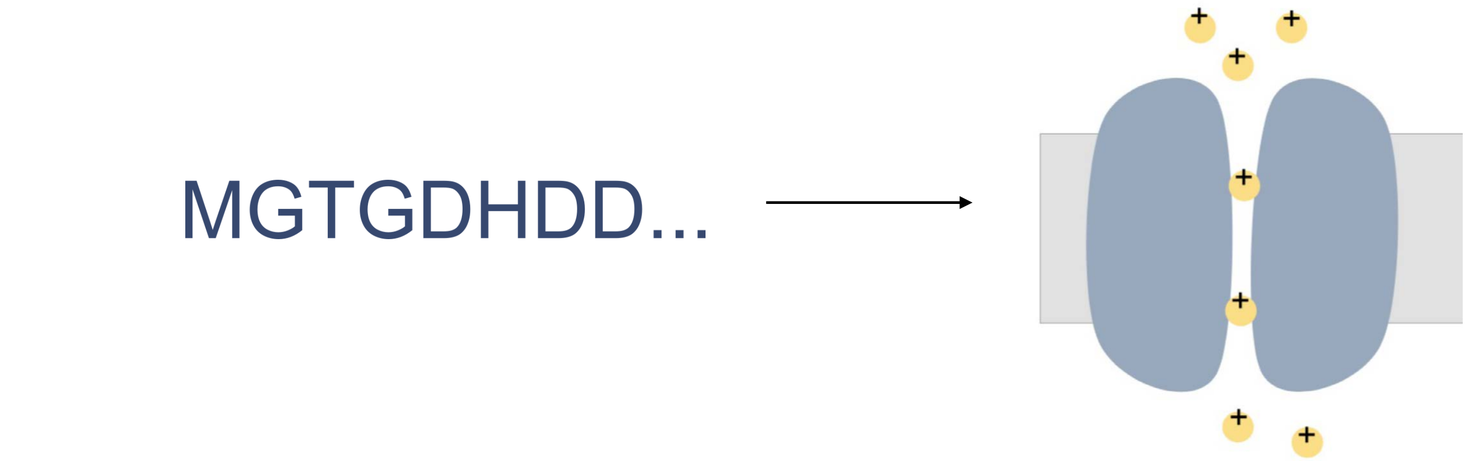
MGTGDHDD...



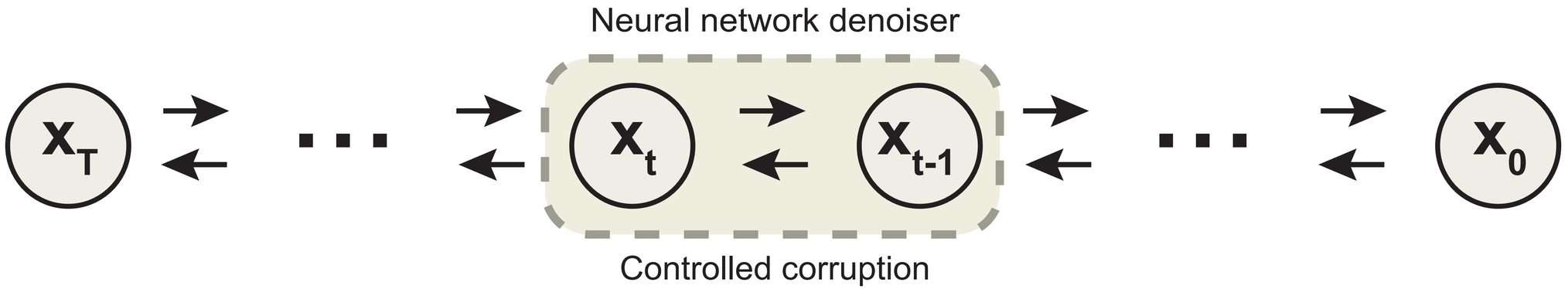


Sequence is the universal protein design space

MGTGDHDD...

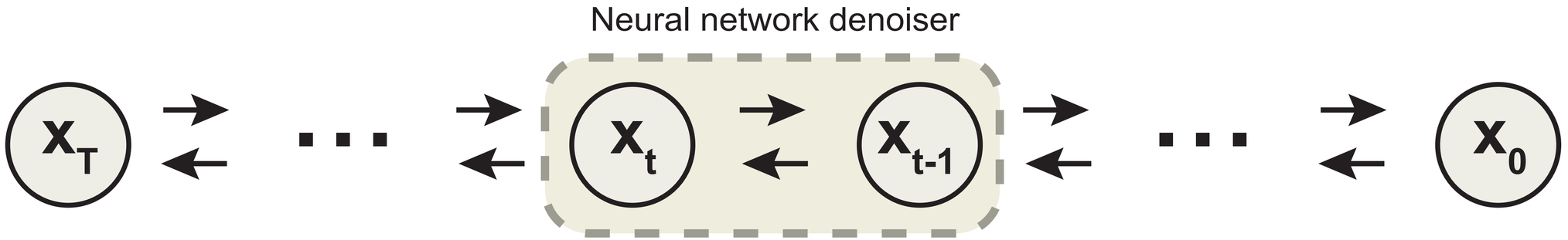


EvoDiff: evolutionary-scale diffusion



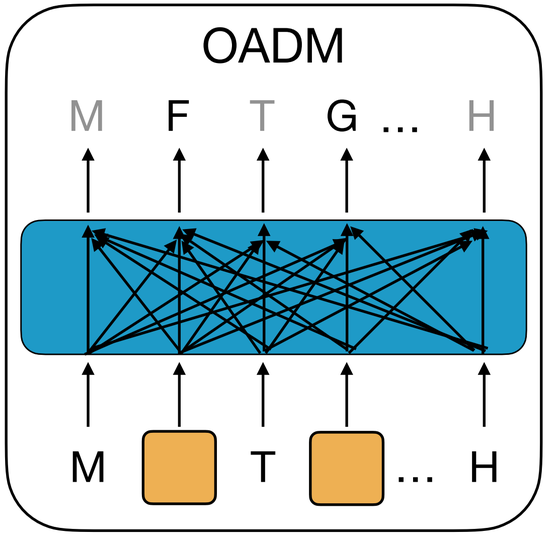
MGTGDHDD…

EvoDiff: evolutionary-scale diffusion

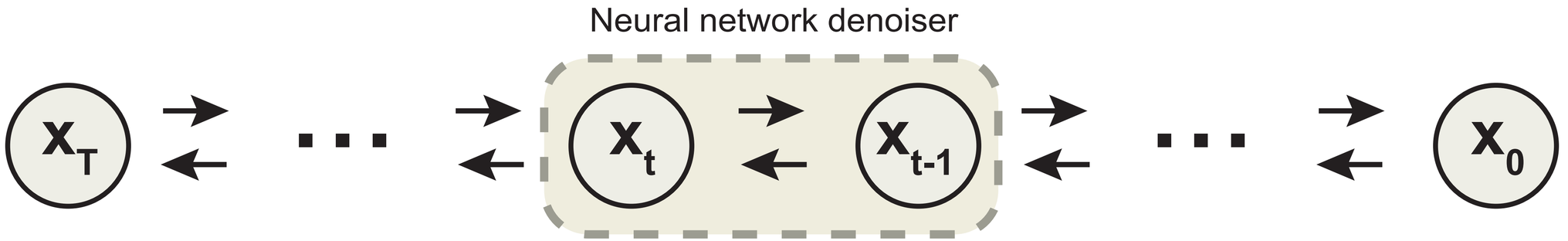


OADM ######## … → MG##D#DD… → MG#GD#DD… → MGTGDHDD…

mask

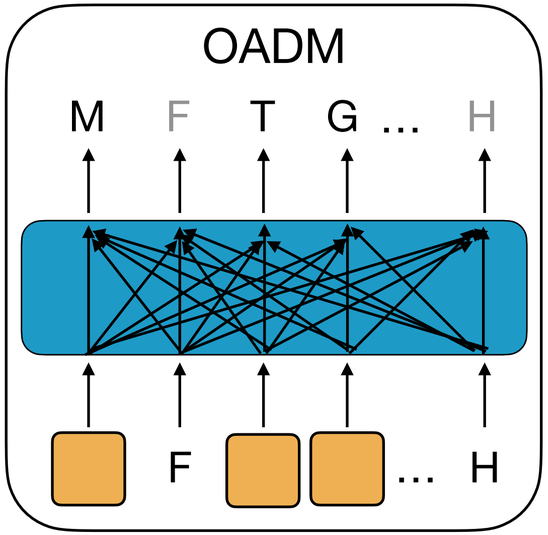


EvoDiff: evolutionary-scale diffusion

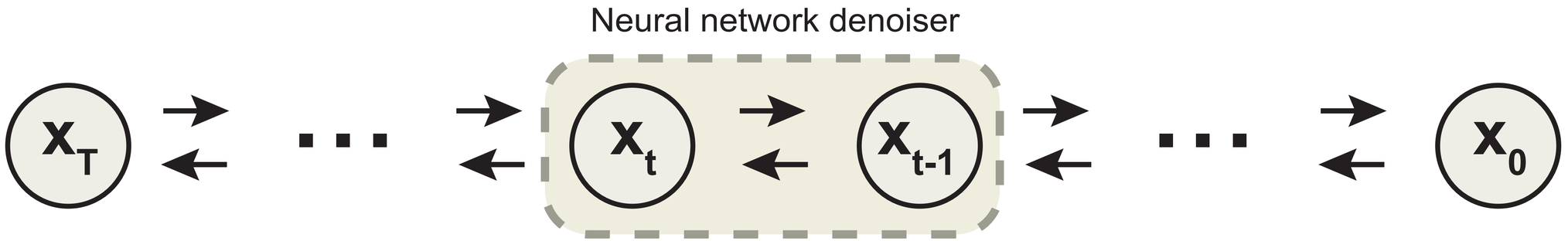


OADM ######## … → MG##D#DD… → MG#GD#DD… → MGTGDHDD…

mask

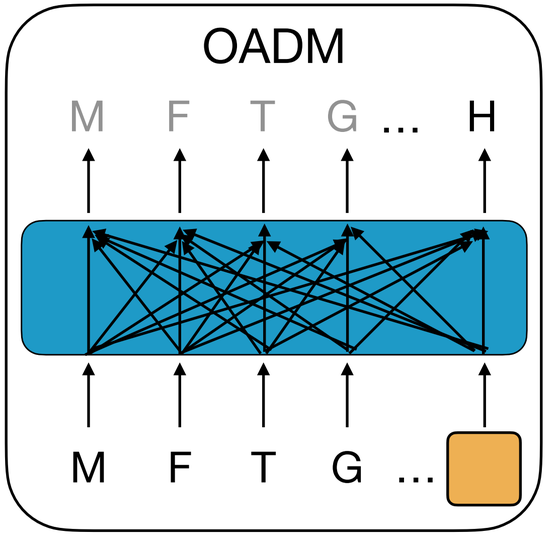


EvoDiff: evolutionary-scale diffusion

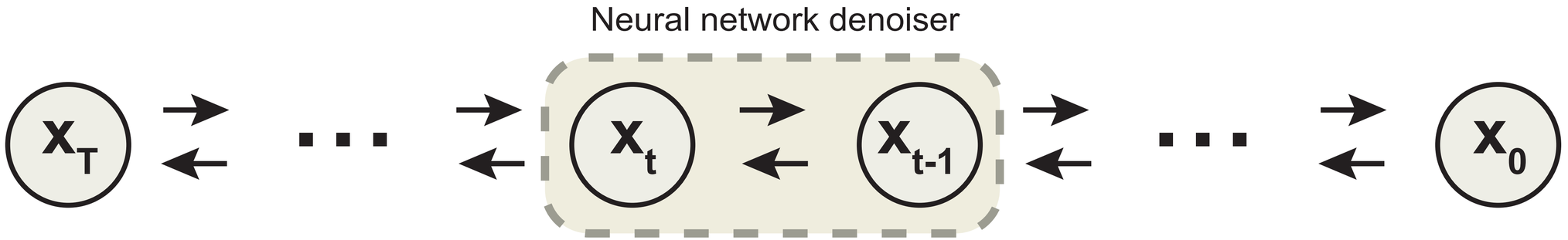


OADM ######## … → MG##D#DD… → MG#GD#DD… → MGTGDHDD…

mask



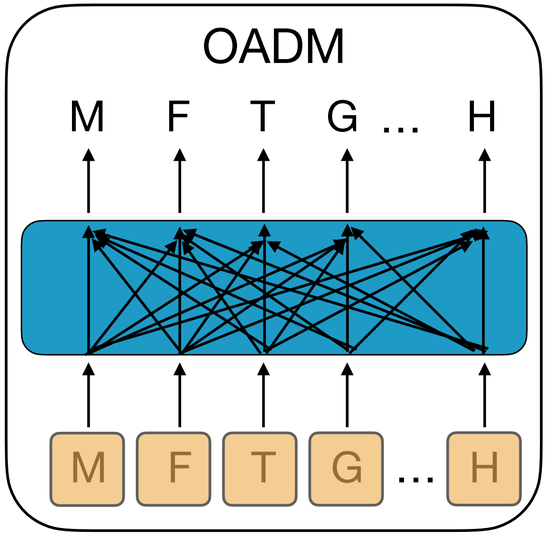
EvoDiff: evolutionary-scale diffusion



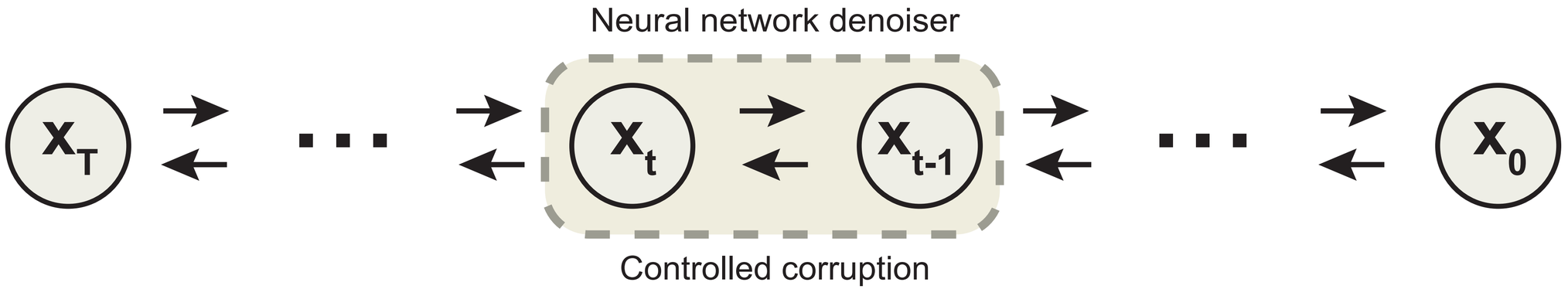
OADM ######## … → MG##D#DD… → MG#GD#DD… → MGTGDHDD…

Hoogeboom *et al., ICLR* 2022

mask



EvoDiff: evolutionary-scale diffusion



OADM ######## … → MG##D#DD… → MG#GD#DD… → MGTGDHDD…

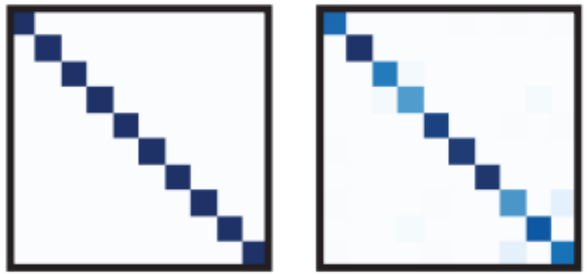
mask

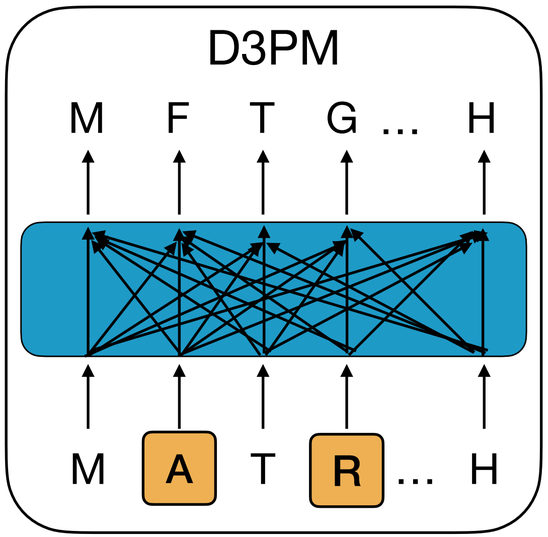
D3PM TYLPLKNE… → MAKTDRDL… → MAGTDRDD… → MGTGDHDD…

mutate

Austin *et al., NeurIPS* 2021

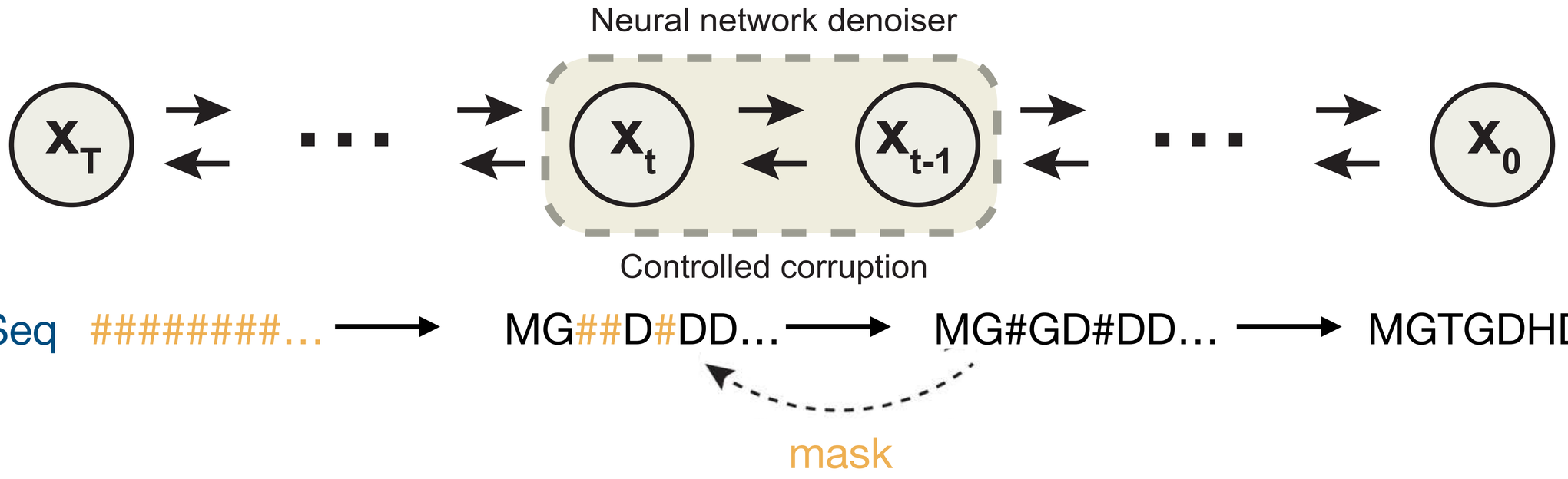
Transition matrix

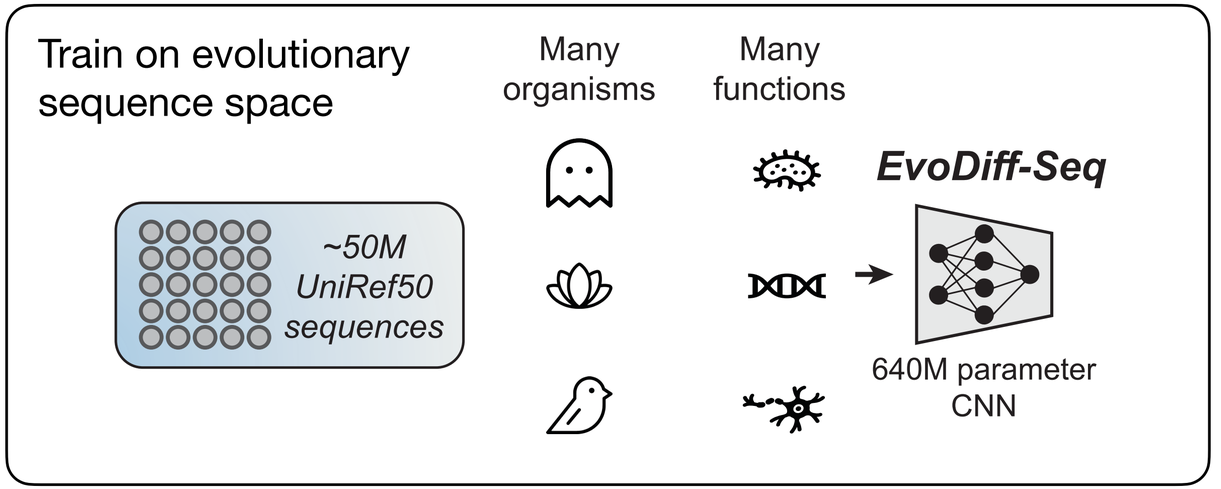




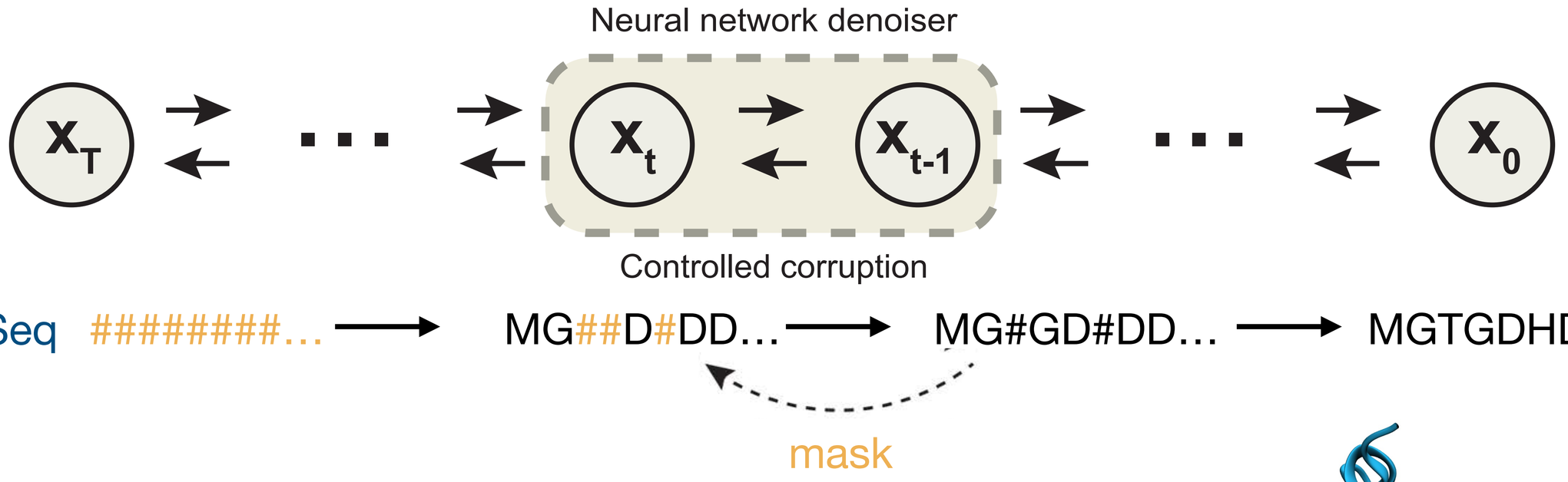
Alamdari+ *bioRxiv* 2023

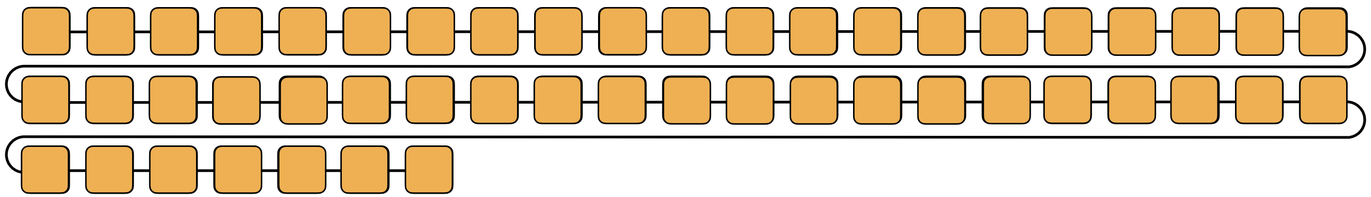
EvoDiff: evolutionary-scale diffusion

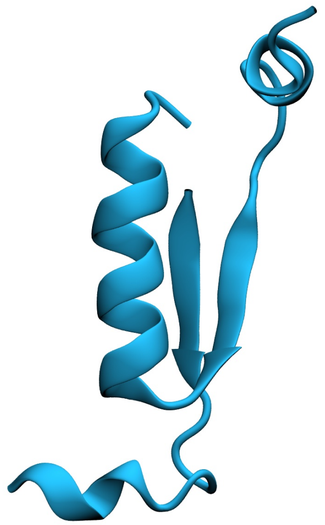




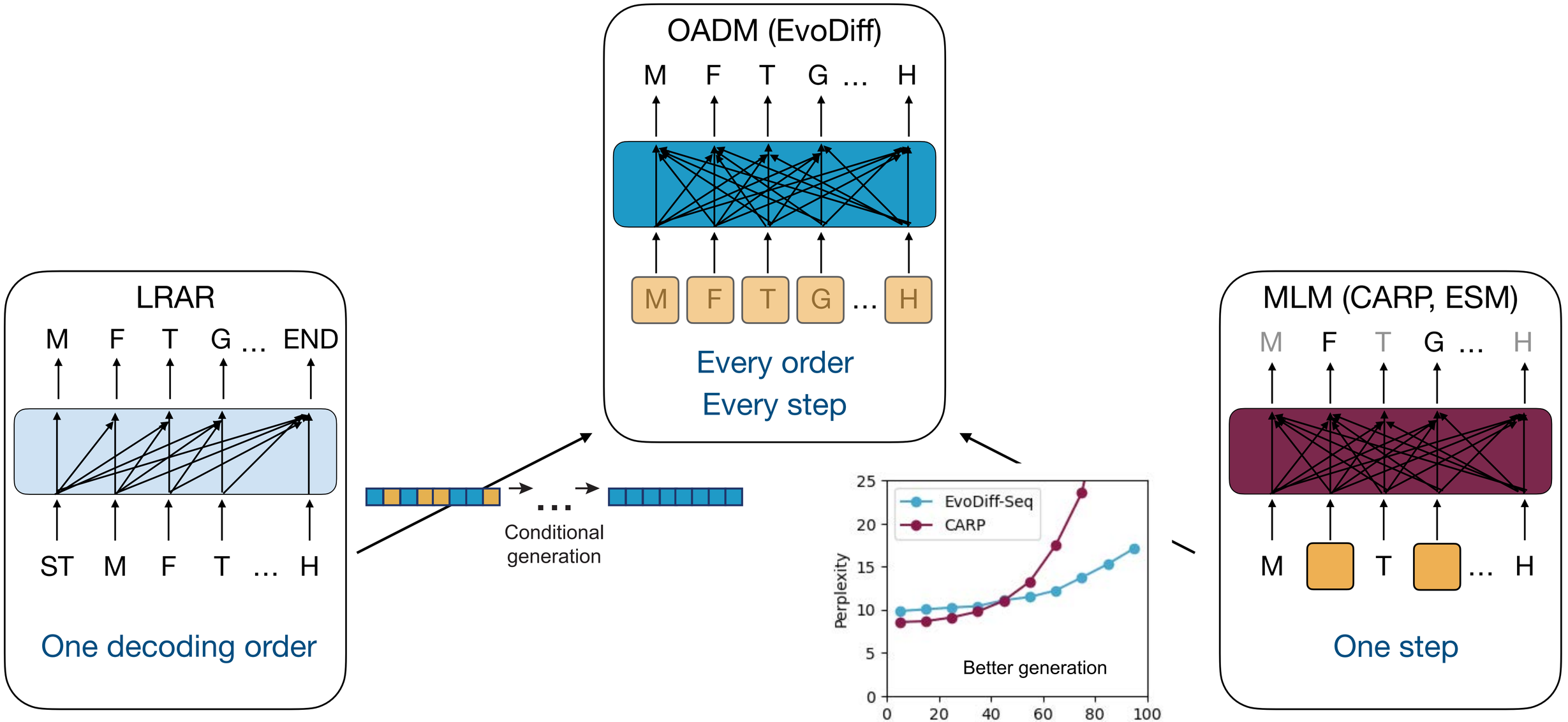
EvoDiff: evolutionary-scale diffusion





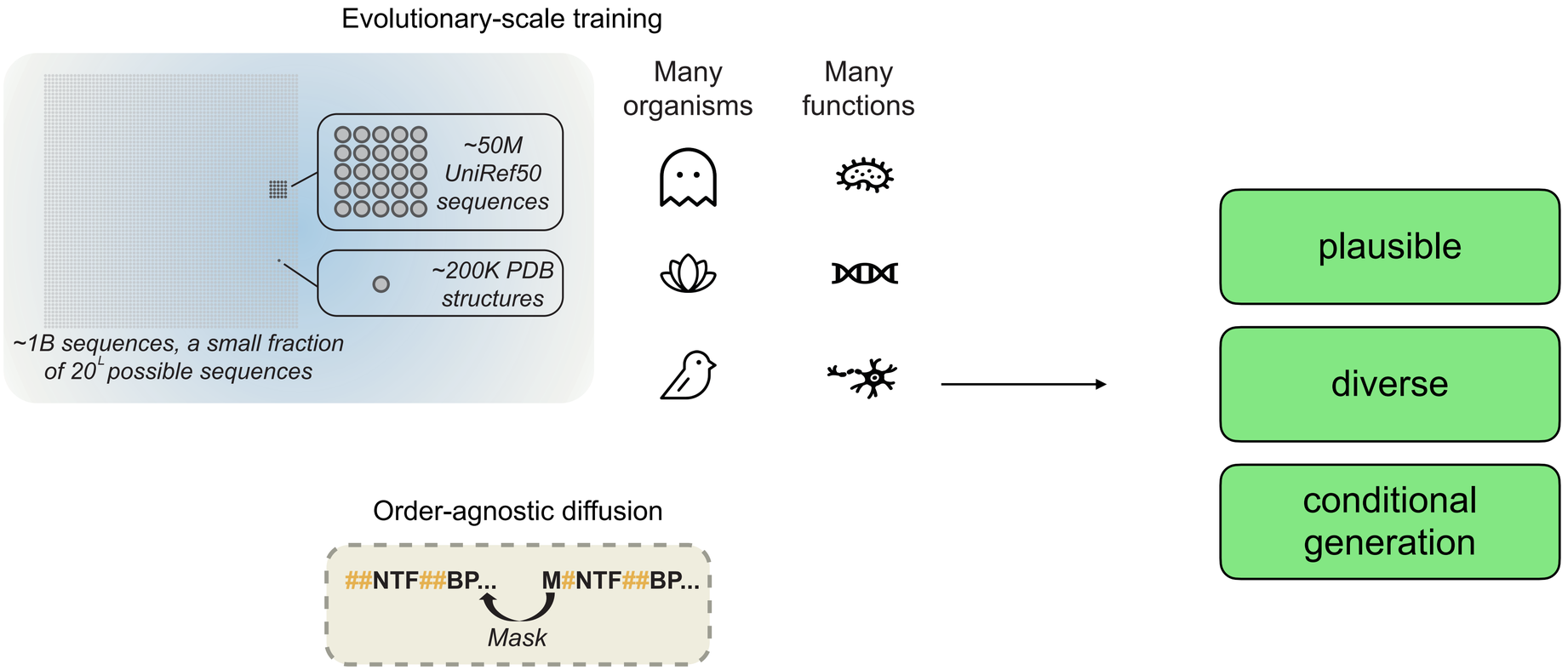


EvoDiff-Seq generalizes masked and autoregressive language models

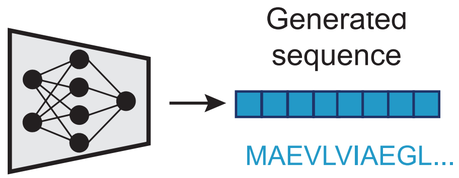


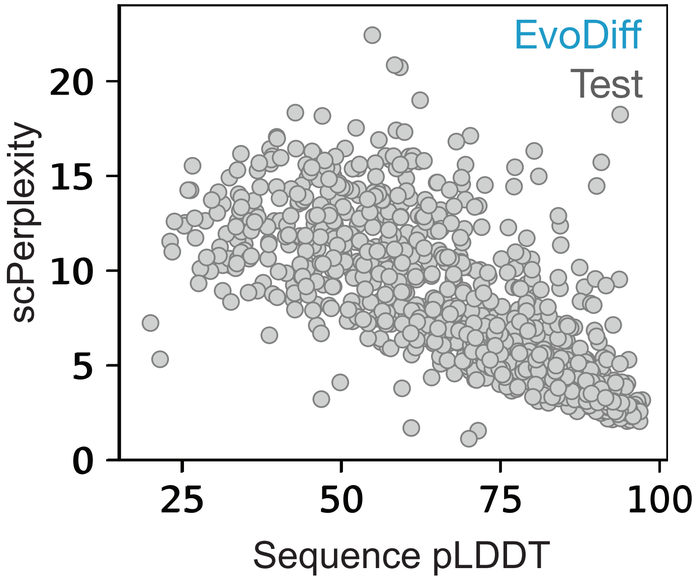
CARP: Yang *et al.*, preprint, 2022
ESM: Rives *et al.*, *PNAS, 2021*

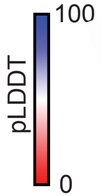
EvoDiff enables controllable generation of plausible, diverse proteins

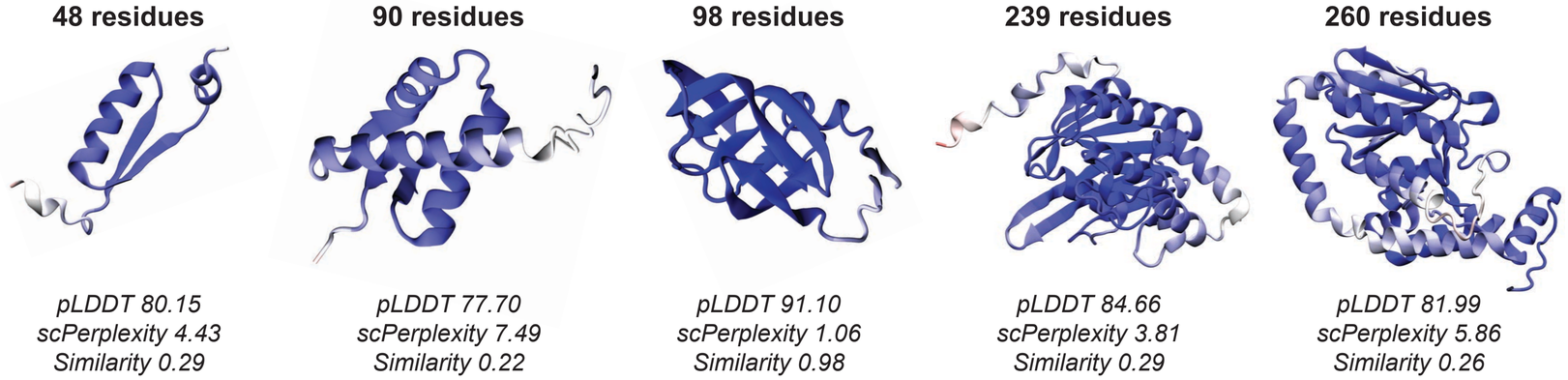


EvoDiff-Seq generates highly-plausible proteins



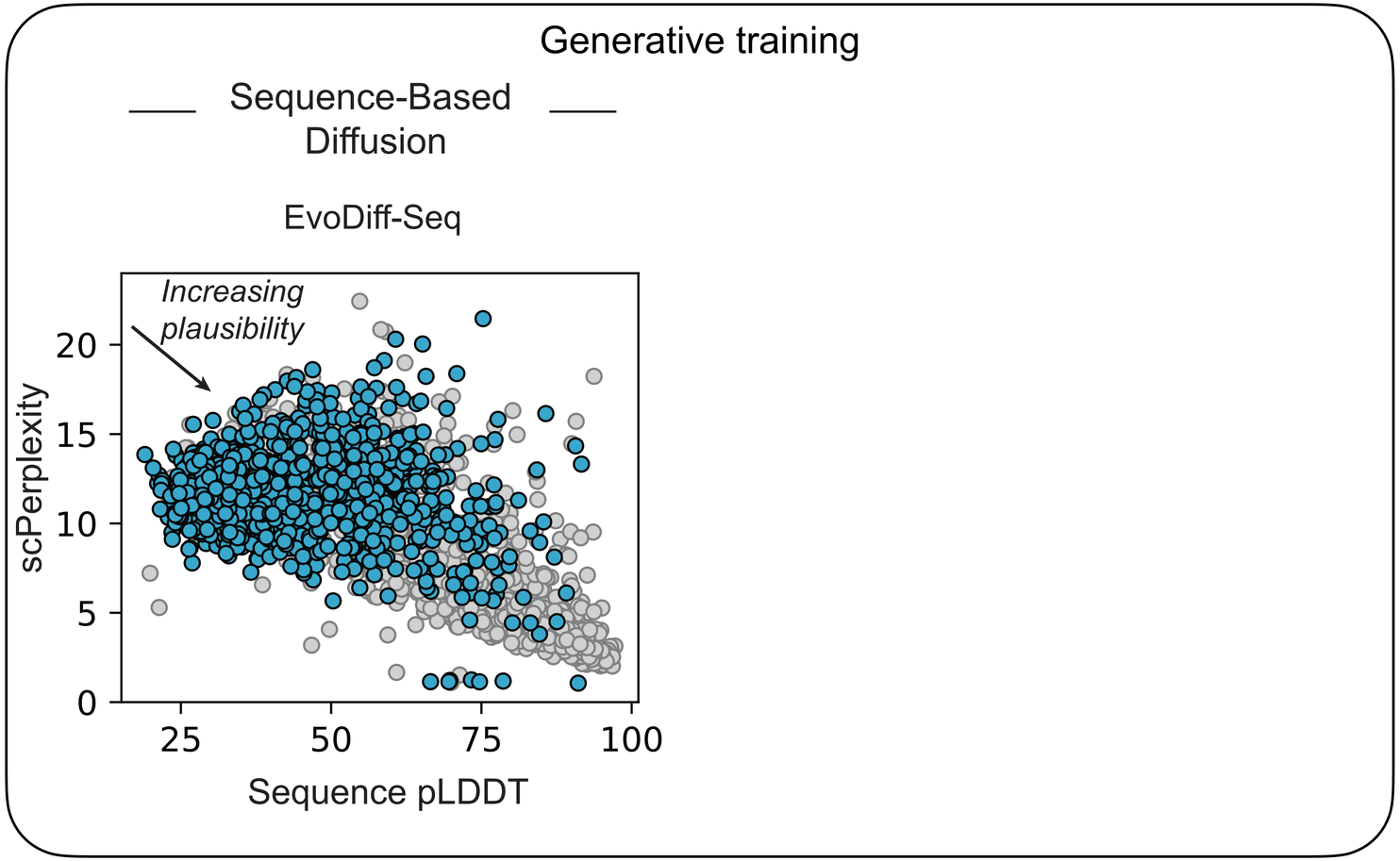






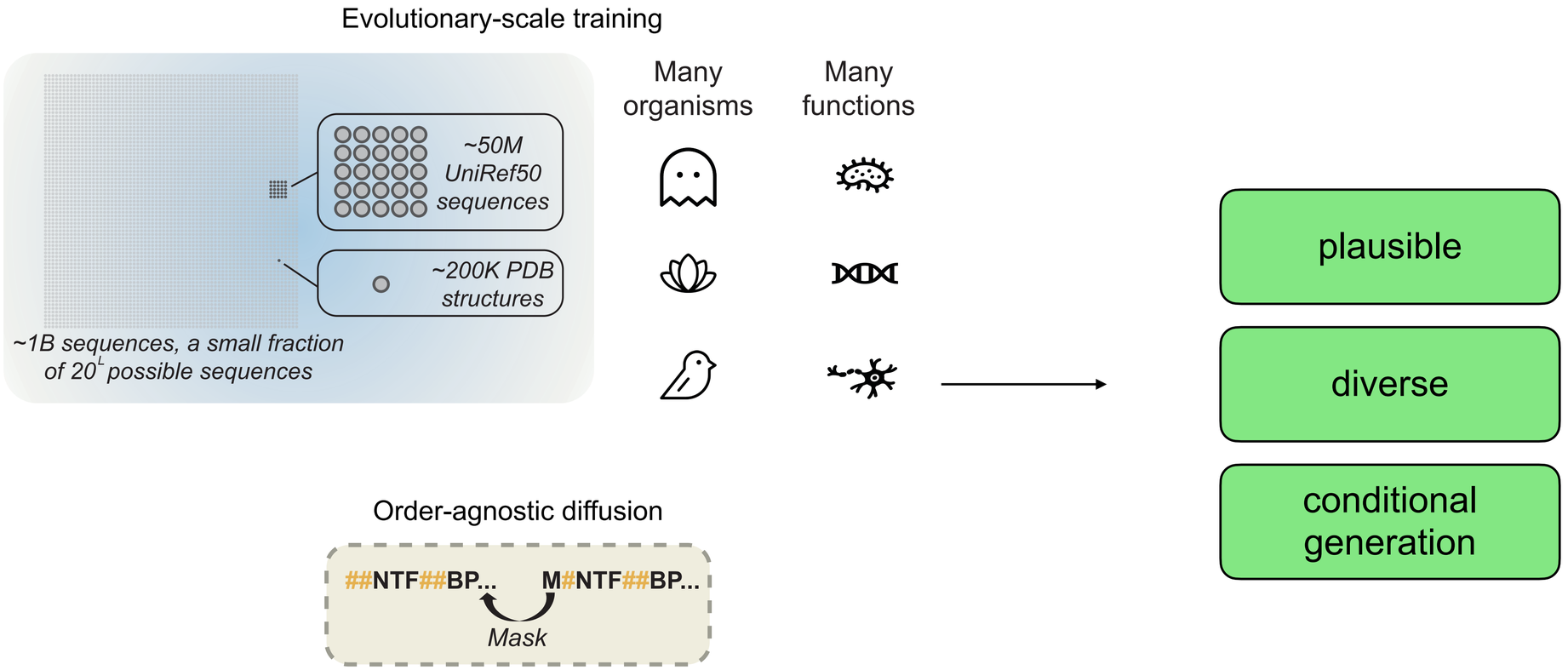
plausible

Generative training results in better sequences

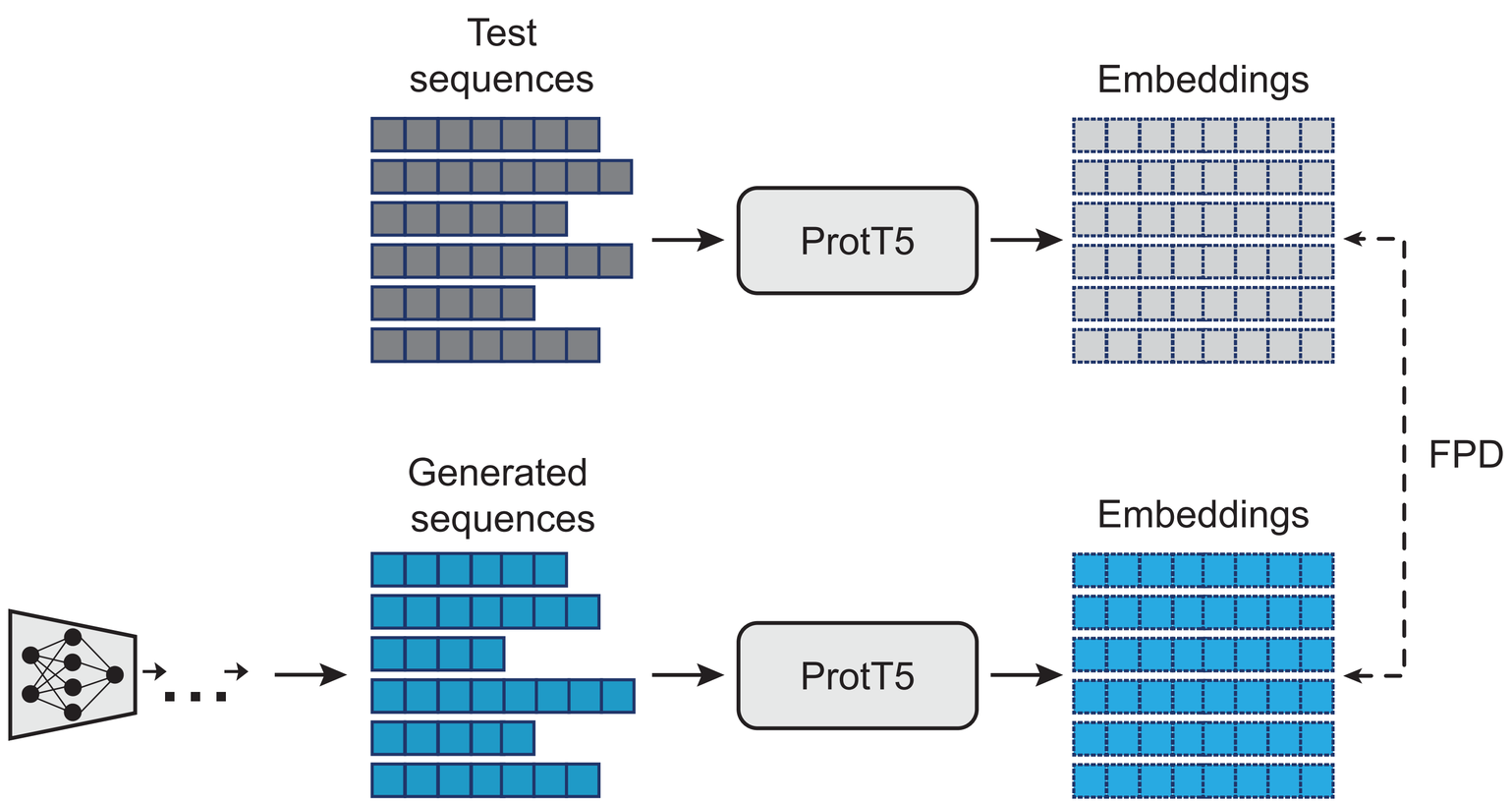


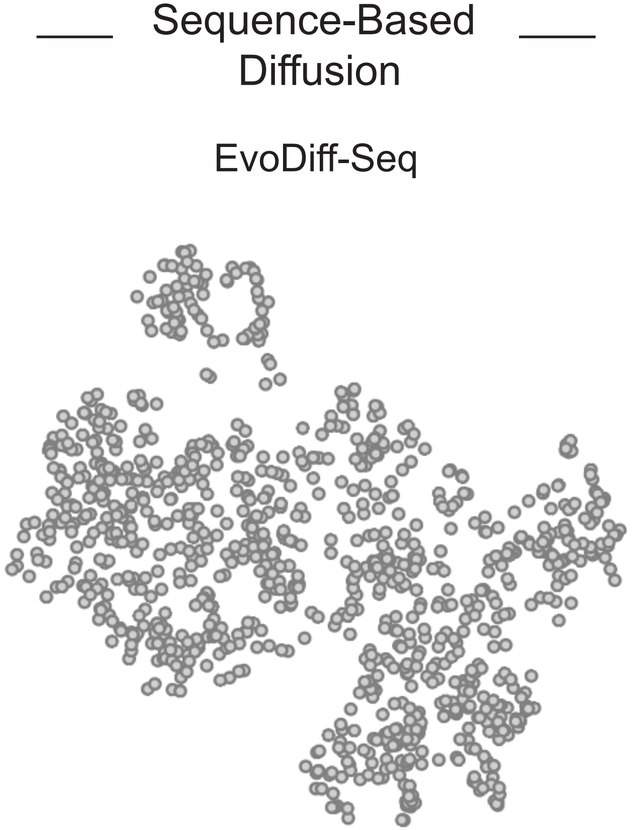
plausible

EvoDiff enables controllable generation of plausible, diverse proteins



EvoDiff-Seq recapitulates natural functional distribution





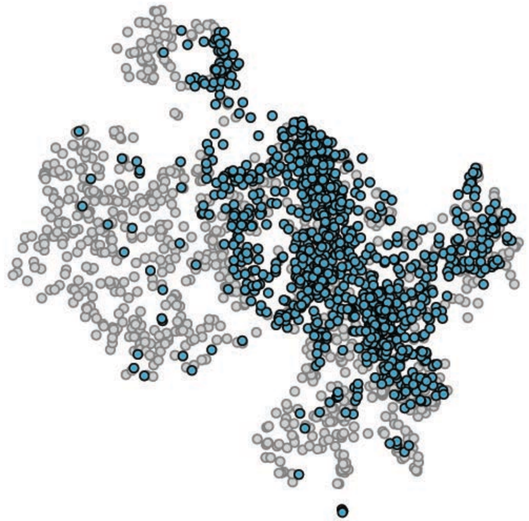
diverse

ProtT5: Elnaggar *et al., IEEE Trans. Pattern Recon.* 2022

Evolutionary-scale diffusion improves FPD

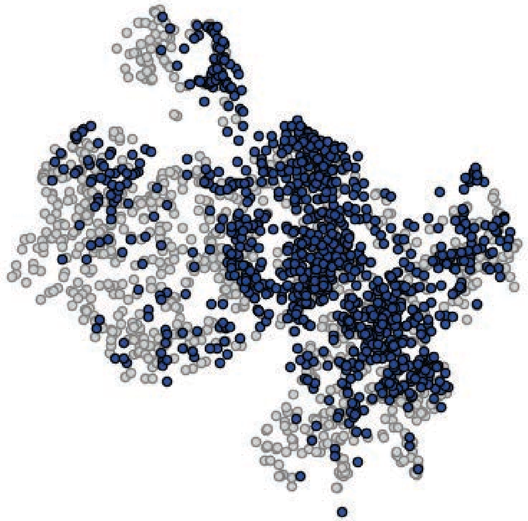
Sequence-Based Diffusion

EvoDiff-Seq
FPD = 0.88



Left-to-right Language Model

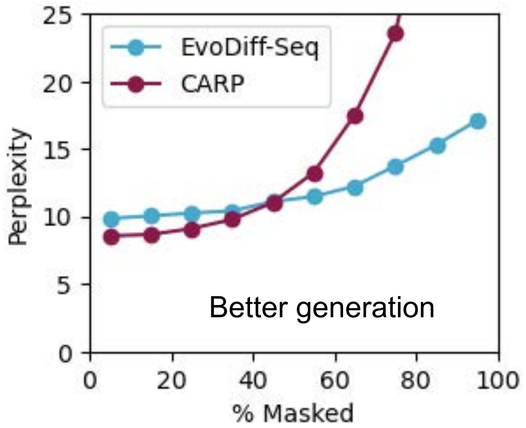
LRAR
FPD = 0.63



Conditional generation

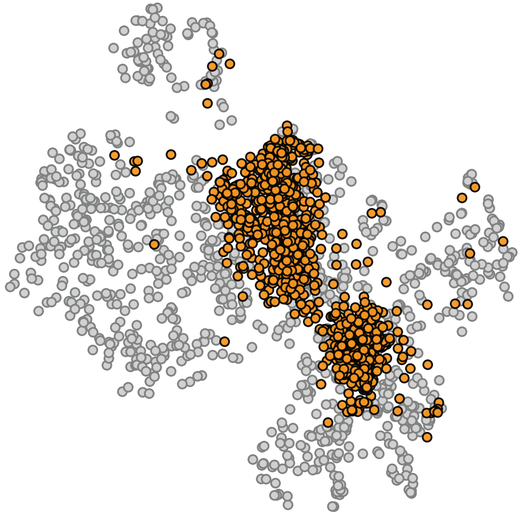
Protein Masked Language Model

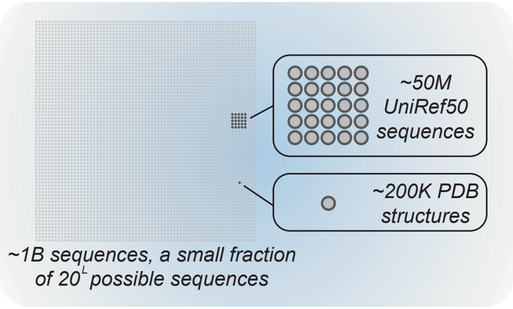
ESM-2
FPD = 2.81



Structure-Based Diffusion

RFdiffusion
FPD = 1.96

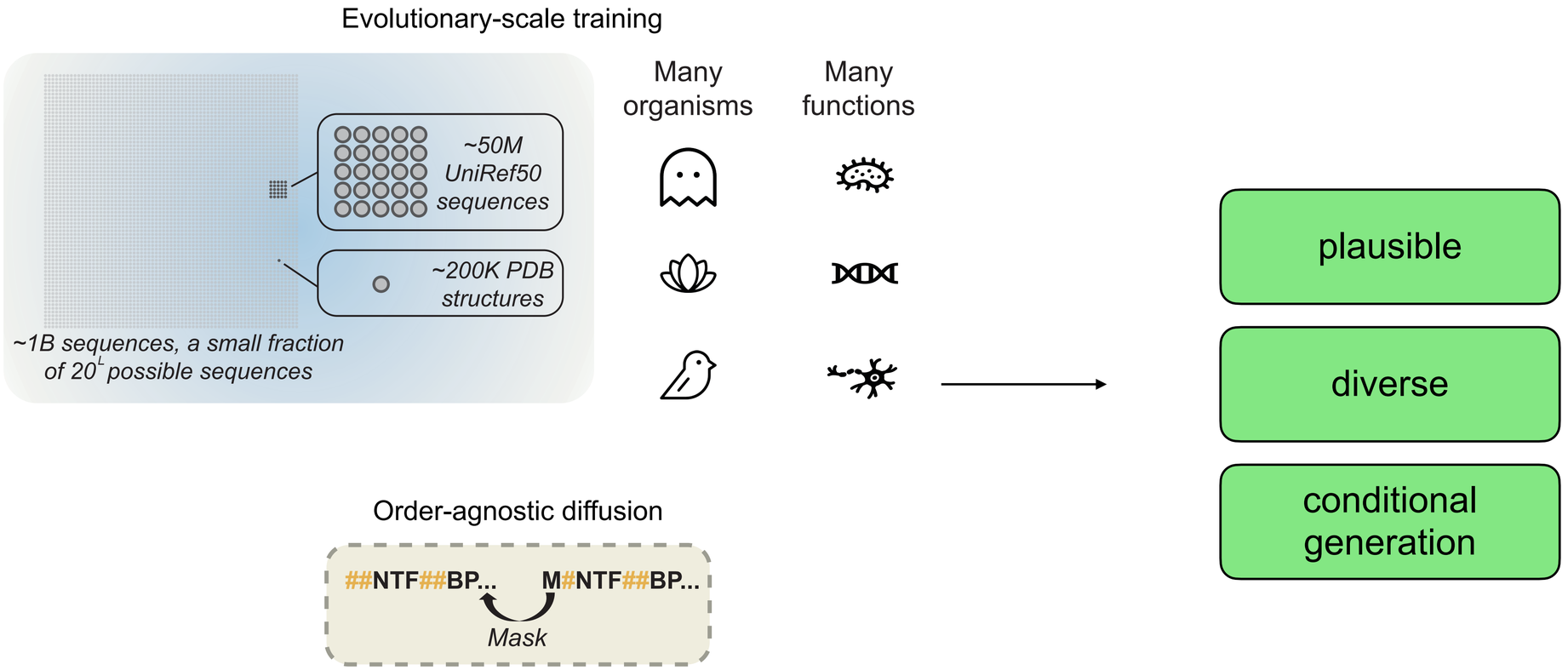




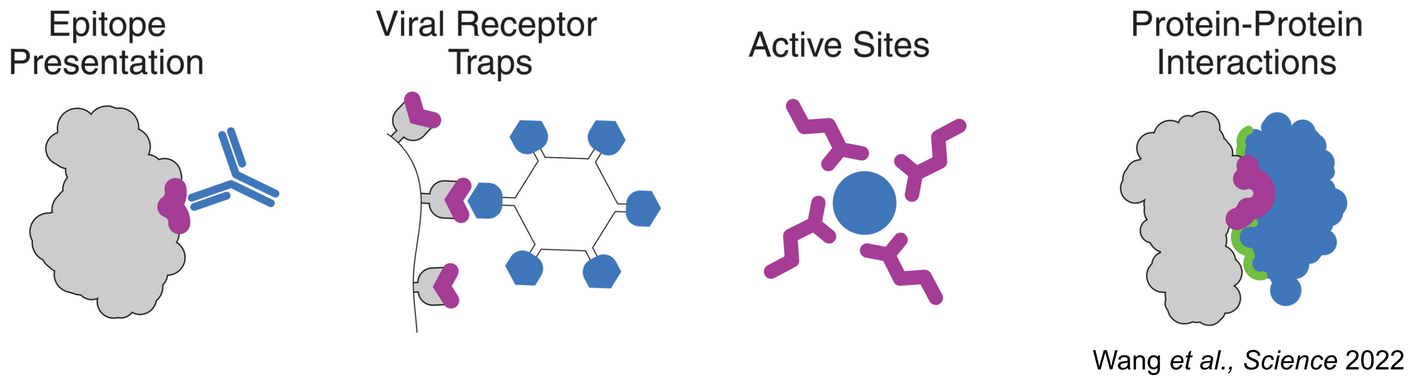
Evolutionary-scale training

diverse

EvoDiff enables controllable generation of plausible, diverse proteins



Many functions are mediated by a motif stabilized by a scaffold



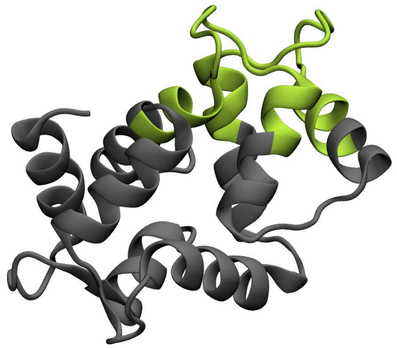
Wang *et al., Science* 2022

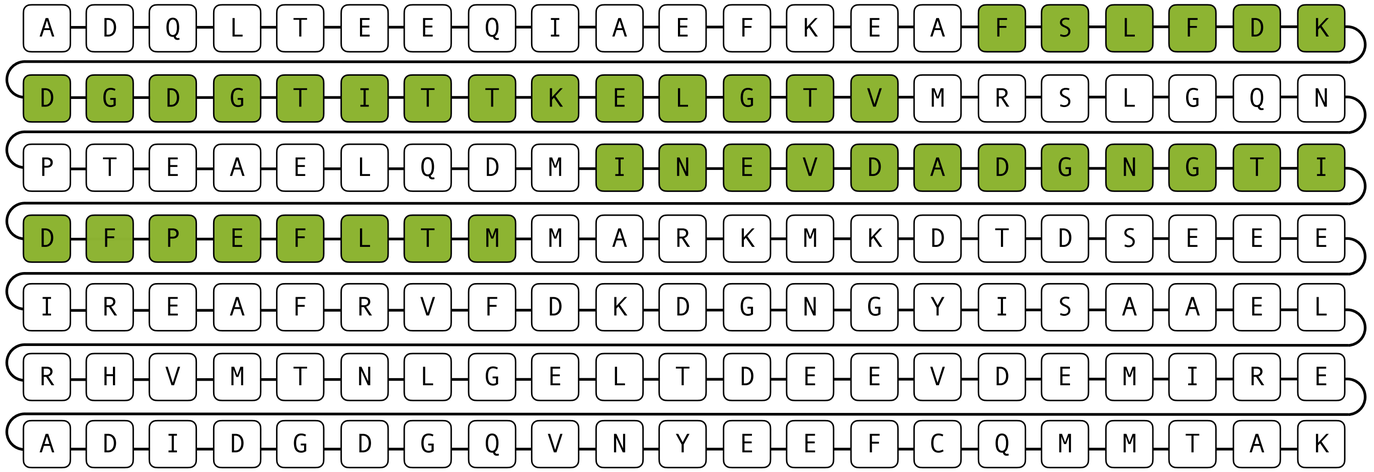
Can we scaffold motifs in sequence space?

conditional generation

EvoDiff can scaffold functional motifs

1PRW: binding site of compact calmodulin

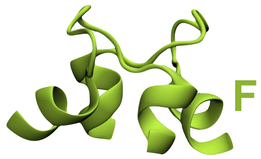




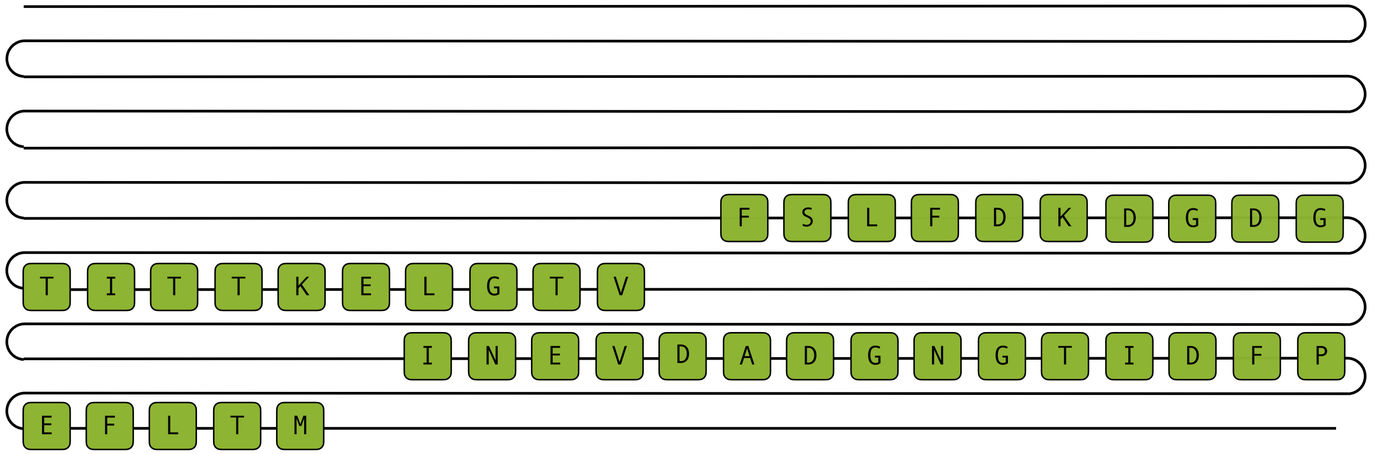
conditional generation

EvoDiff can scaffold functional motifs

1PRW: binding site of compact calmodulin



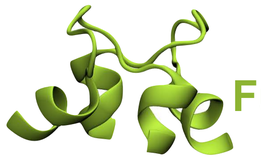
Functional motif



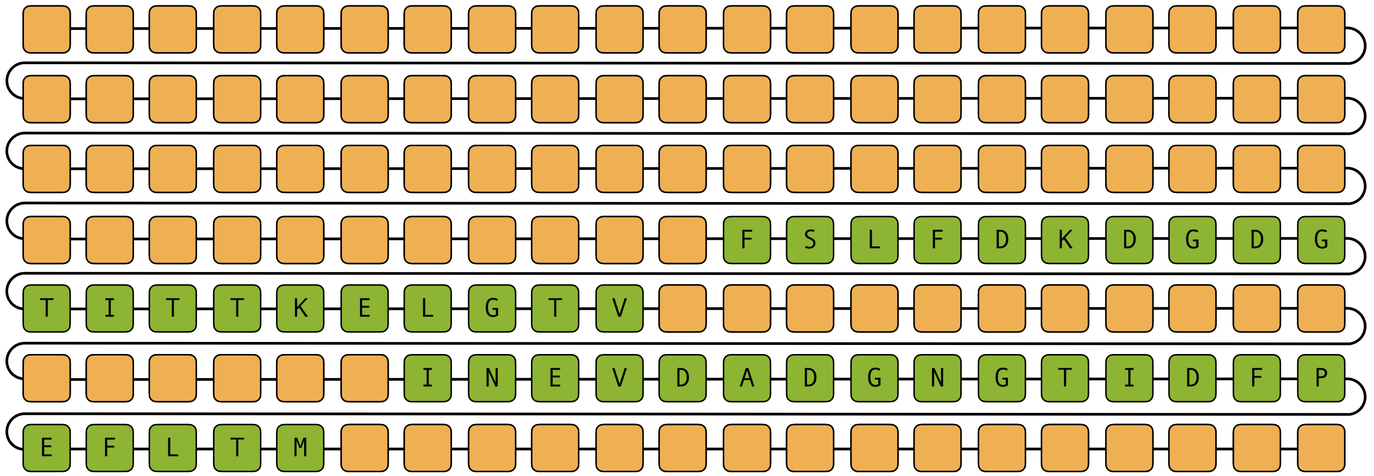
conditional generation

EvoDiff can scaffold functional motifs

1PRW: binding site of compact calmodulin



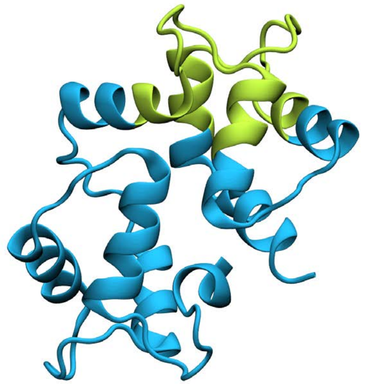
Functional motif

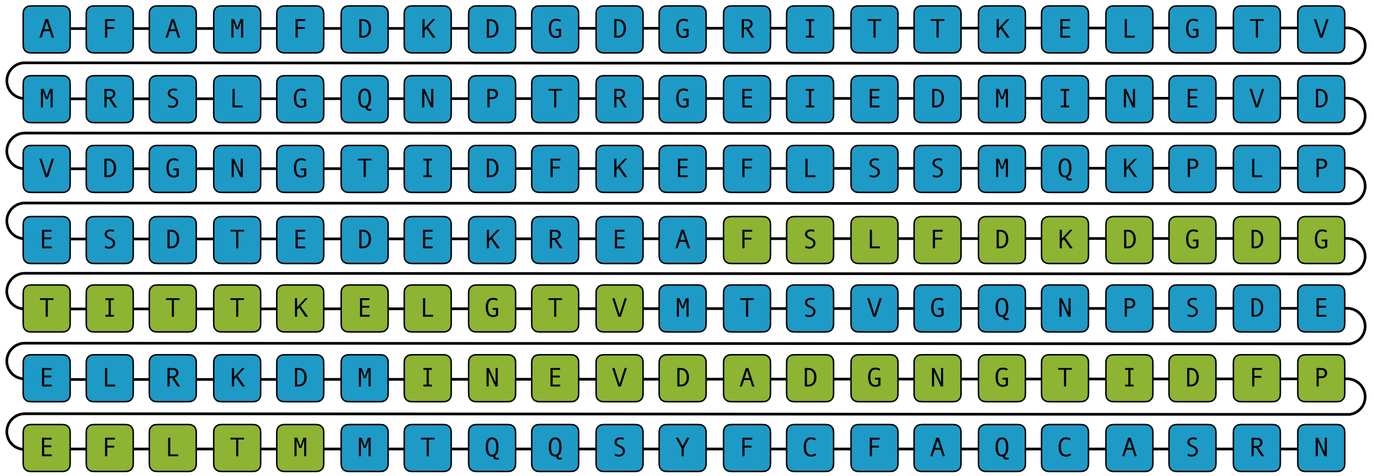


conditional generation

EvoDiff can scaffold functional motifs

1PRW: binding site of compact calmodulin





No structure needed!

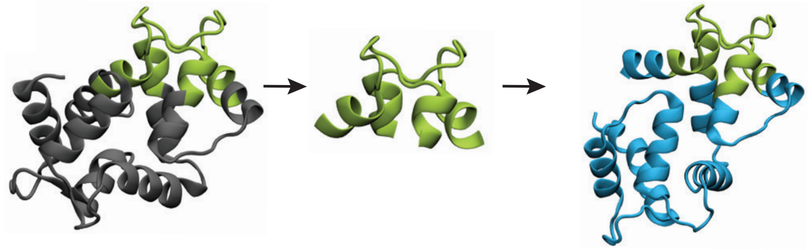
conditional generation

EvoDiff can scaffold functional motifs

EvoDiff-Seq

1PRW: Binding site of compact calmodulin

Native structure | Functional motif | Generated sequence



ADQLTEEQIAEFKEAFSLFDKDGDGTIT
TKELGTVMRSLGQNPTEAELQDMINEV
DADGNGTIDFPEFLTMMARKMKDTDSE
EEIREAFRVFDKDGNGYISAAELRHVMT
NLGELTDEEVDEMIREADIDGDGQVNY
EEFVQMMTAK

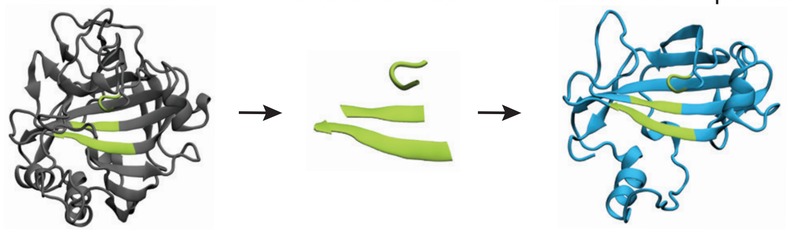
FISDVENAFSLFDKDGDGTITTKELGTVM
RSLGQNPTESELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVF
DRDNNGLISAAELRHVMTNLGEKLTDDE
VDEMIREADVDGDGQVNYEEFVTMMTA
KSLDYNDD

pLDDT 83.3
motifRMSD 0.73Å
TMscore 0.54

EvoDiff-MSA

5YUI: Binding site of carbonic anhydrase metalloenzyme

Native structure | Functional motif | Generated sequence



HWGYGKHNGPEHWHKDFPIAKGERQSP
VDIDTHTAKYDPSLKPLSVSYDQATSLRIL
NNGHAFNVEFDDSQDKAVLKGGPLDGTY
RLIQFHFHWGSLDGQGSEHTVDKKKYAA
ELHLVHWNTKYGDFGKAVQQPDGLAVLG
IFLKVGSAKPGLQKVVDVLDSIKTKGKSAD
FTNFDPRGLLPESLDYWTYPGSLTTPPLL
ECVTWIVLKEPISVSSEQVLKFRKLNFNGE
GEPEELMVDNWRPAQPLKNRQIKASFK

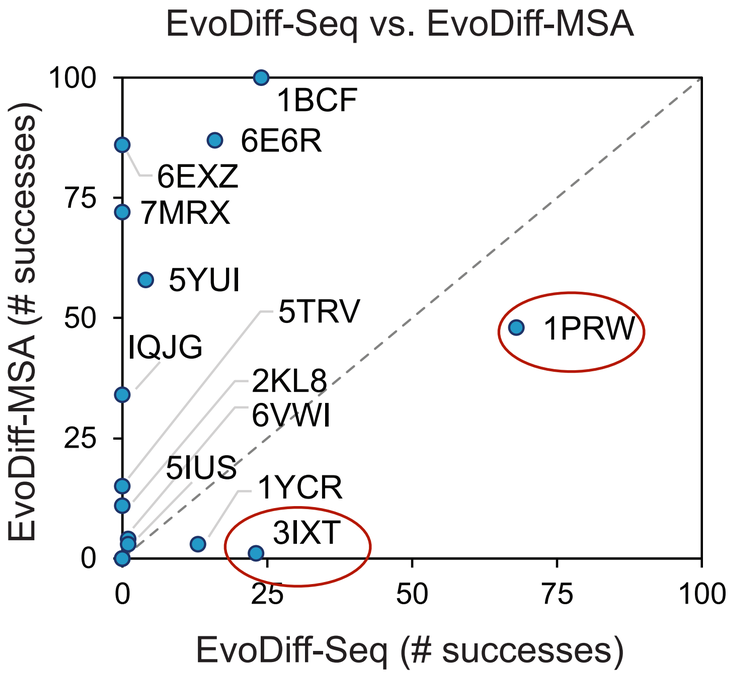
SWAGDAMLSGGGLSGDYSVAEFHFHW
GSTNTAGSEHTINNIRHAAELHLVHVSNR
FGTIEEAARVRNGVAVLGVFFEVGEINAG
LEPITDKLRHLAGRGTHEPVNPLAPHEYM
PSSDDFFTYTGSLTTPPCSTGVLWYVF
DRPTRISVHQ

pLDDT 90.1
motifRMSD 0.47 Å
TMscore 0.88

Model	# Successful (< 1A RMSD)	# Problems solved
RFdiffusion	610	13 / 17
EvoDiff-MSA	522	13 / 17
EvoDiff-Seq	149	8 / 17

conditional generation

EvoDiff can scaffold functional motifs



MSA usually more successful

Orthogonal strengths

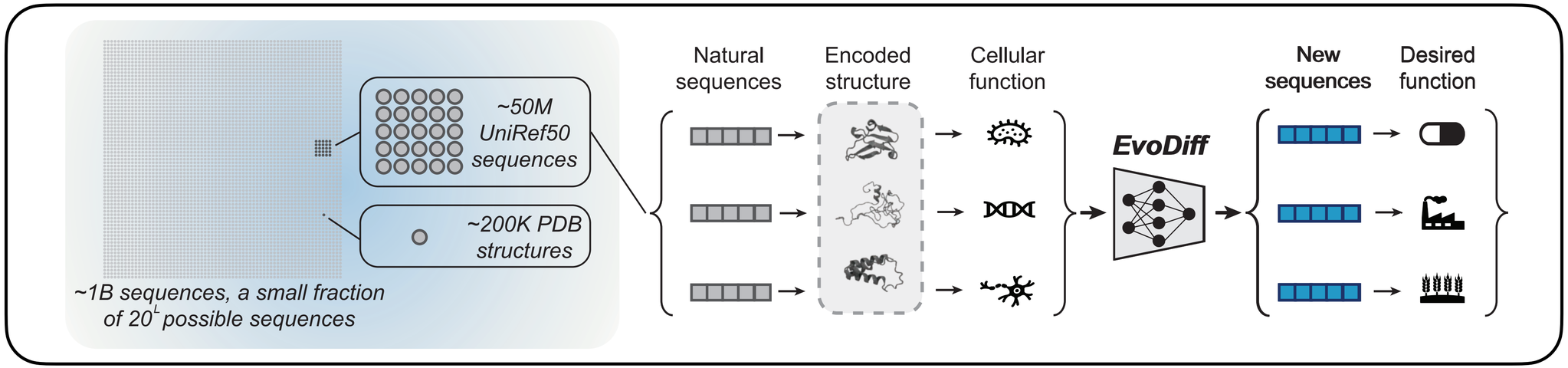
EvoDiff-Seq best

Not many homologs

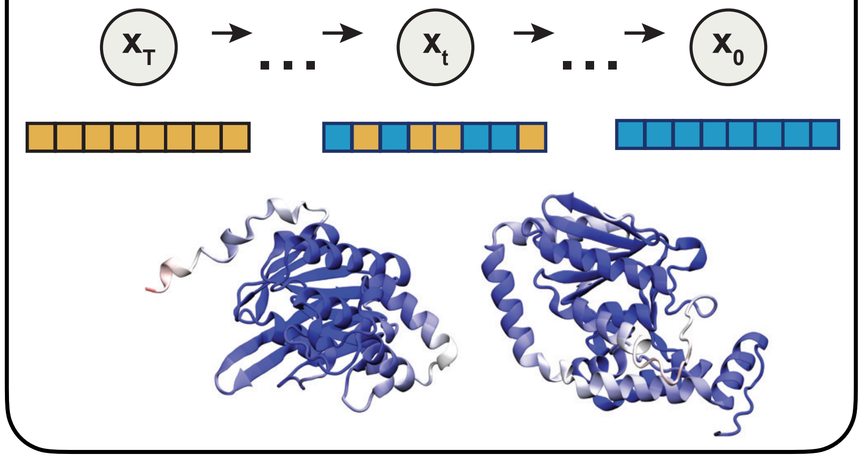
Seq usually more novel

conditional generation

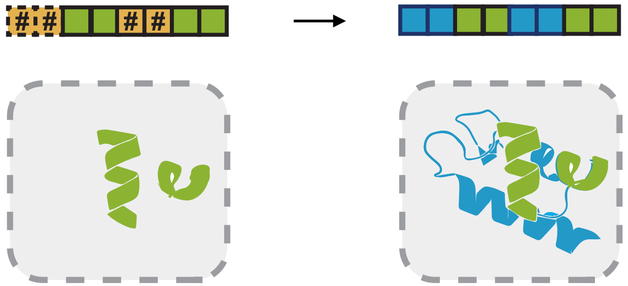
EvoDiff: controllable protein sequence diffusion



Unconditional generation

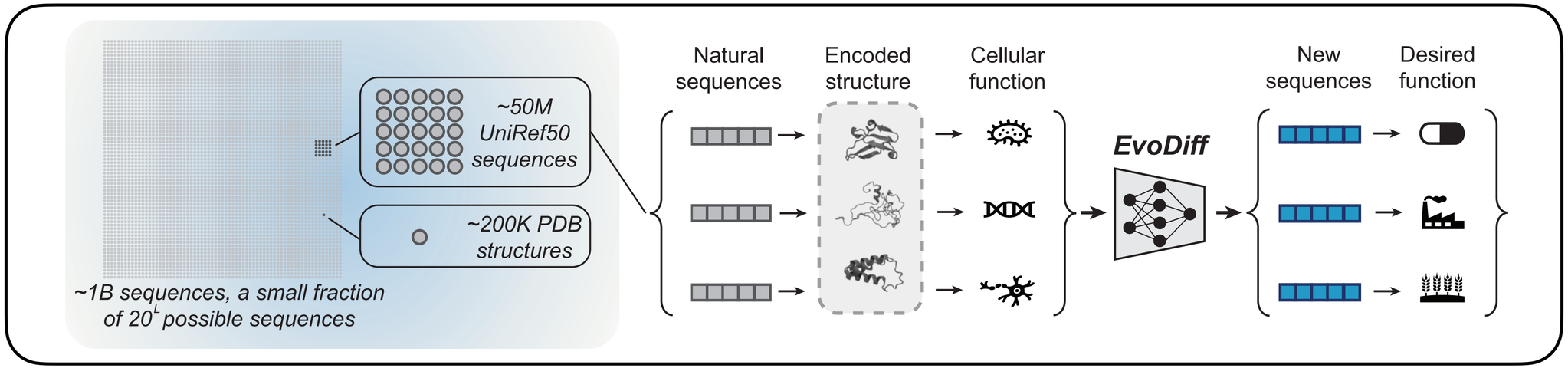


Sequence motif scaffolding

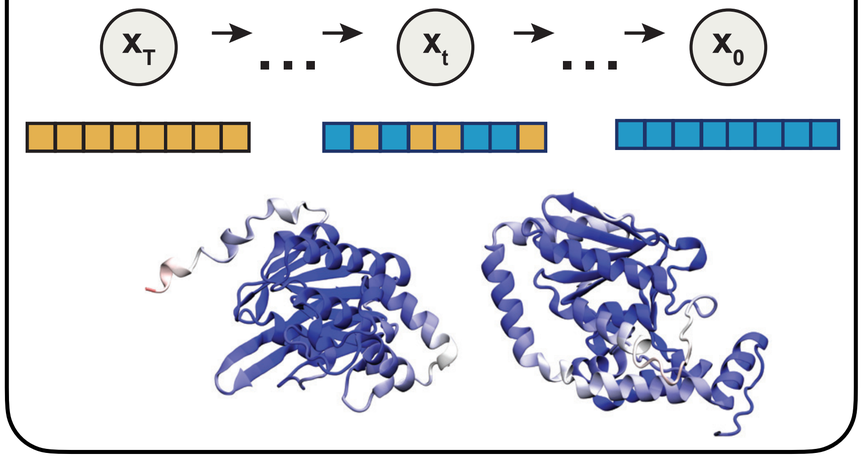


Controllable generation

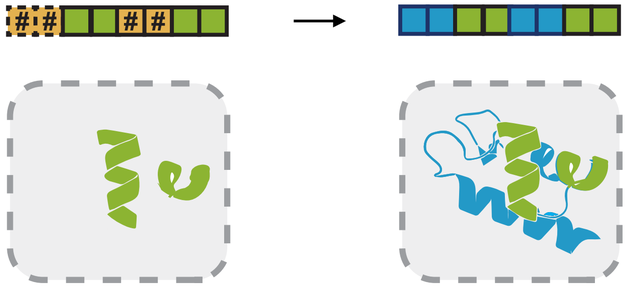
EvoDiff: controllable protein sequence diffusion



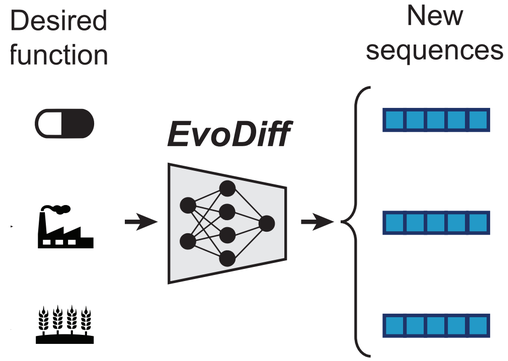
Unconditional generation



Sequence motif scaffolding



Controllable generation



Preprint: https://doi.org/10.1101/2023.09.11.556673

github.com/microsoft/evodiff

Acknowledgments

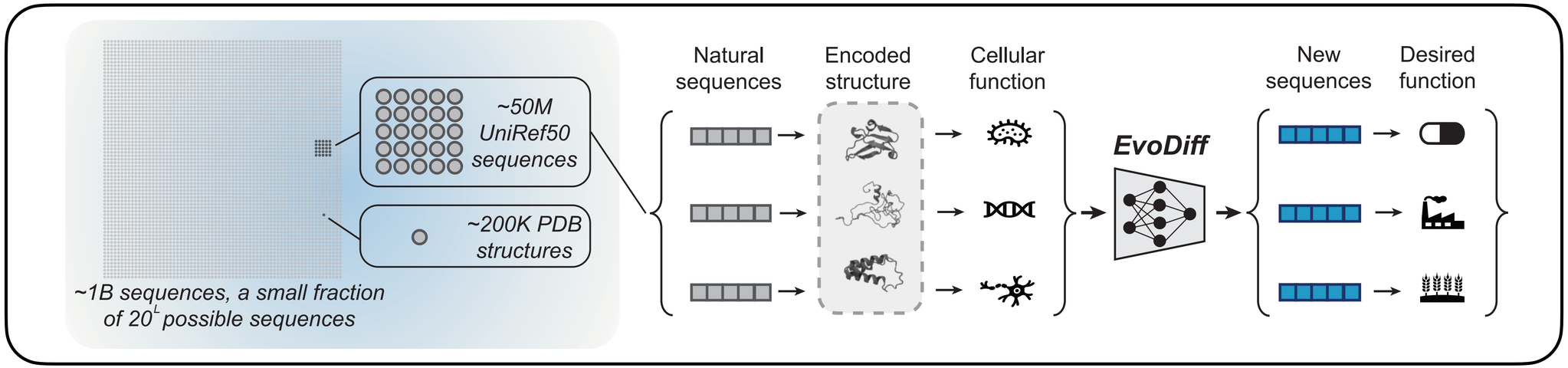




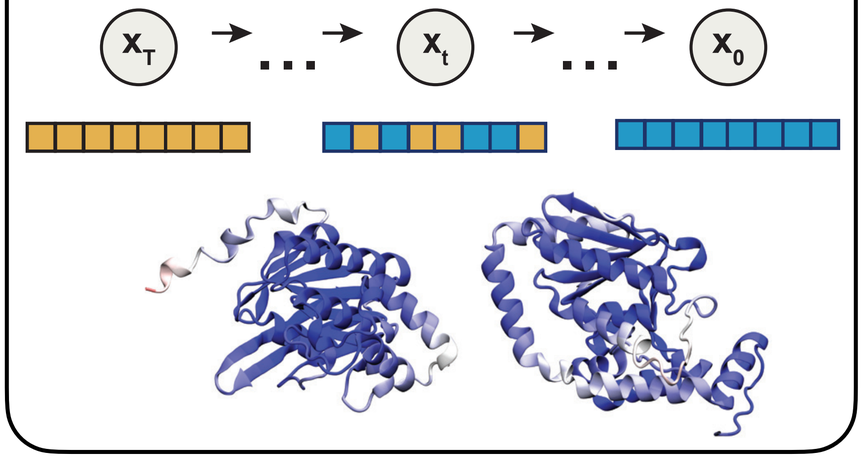



BioML at MSR New England

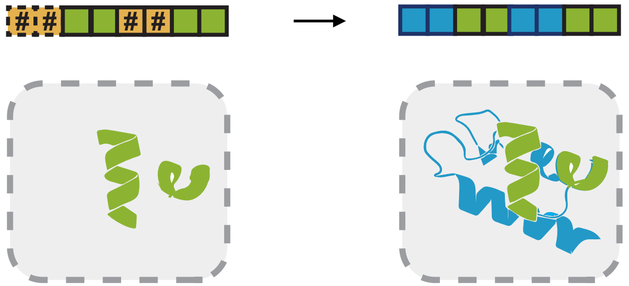
EvoDiff: controllable protein sequence diffusion



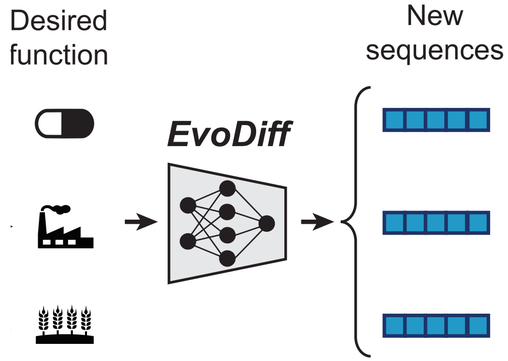
Unconditional generation



Sequence motif scaffolding



Controllable generation



Preprint: https://doi.org/10.1101/2023.09.11.556673

github.com/microsoft/evodiff