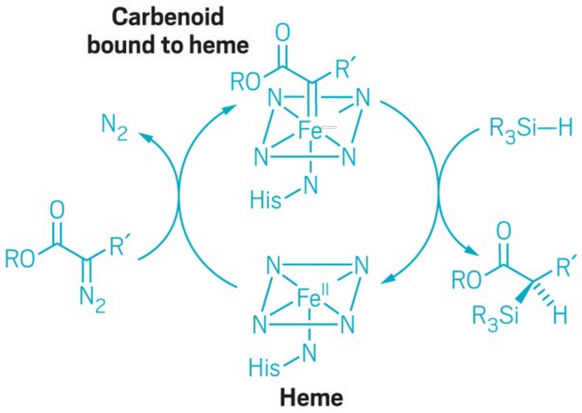
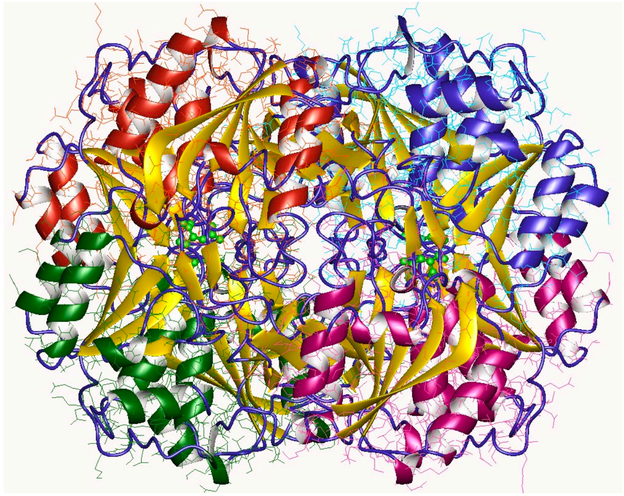
Protein sequence generation with evolutionary diffusion

Kevin Kaichuang Yang
Microsoft Research New England
@KevinKaichuang

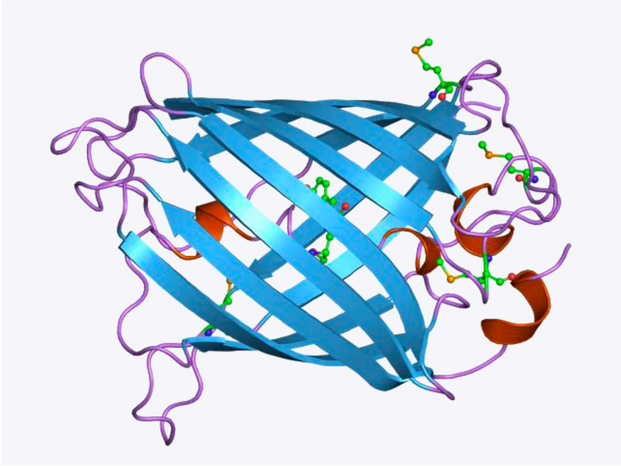
We need proteins with new functions



new chemistry

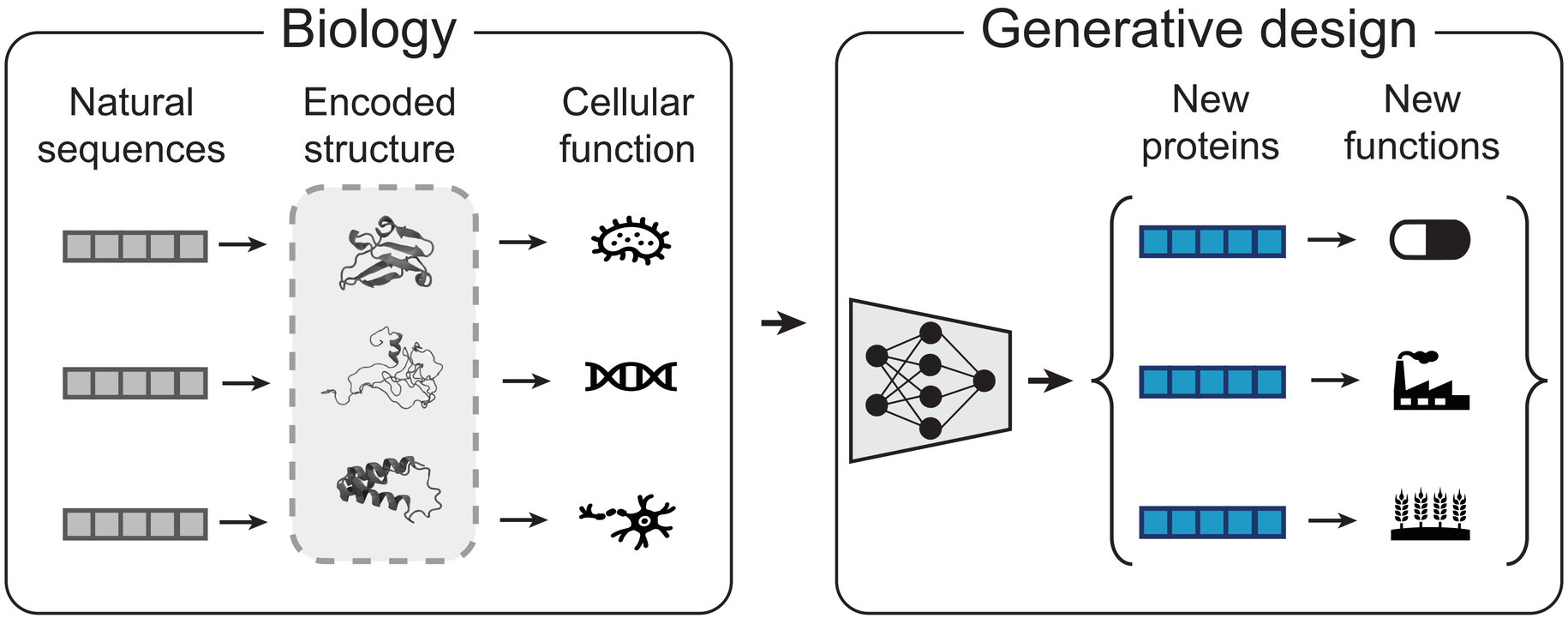


therapeutics

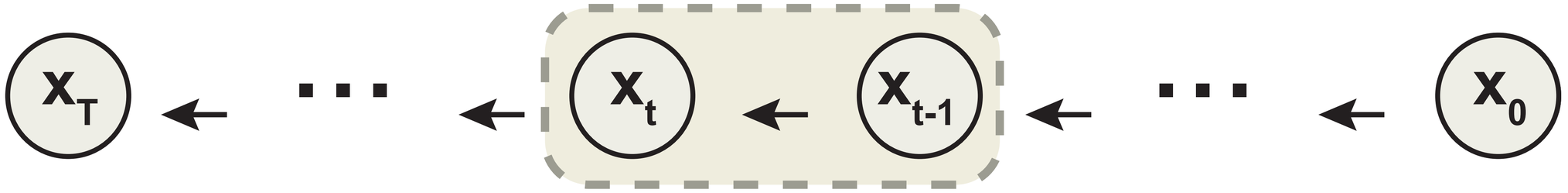


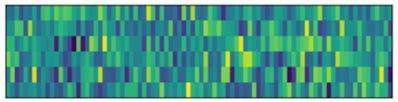
molecular tools

Generate new proteins to expand functional space

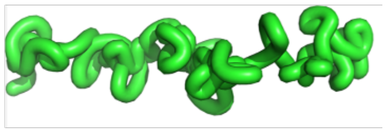


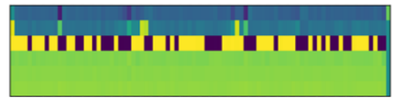
Diffusion models generate realistic structures



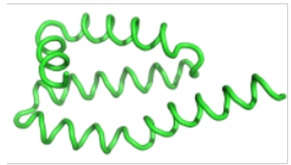


random structures

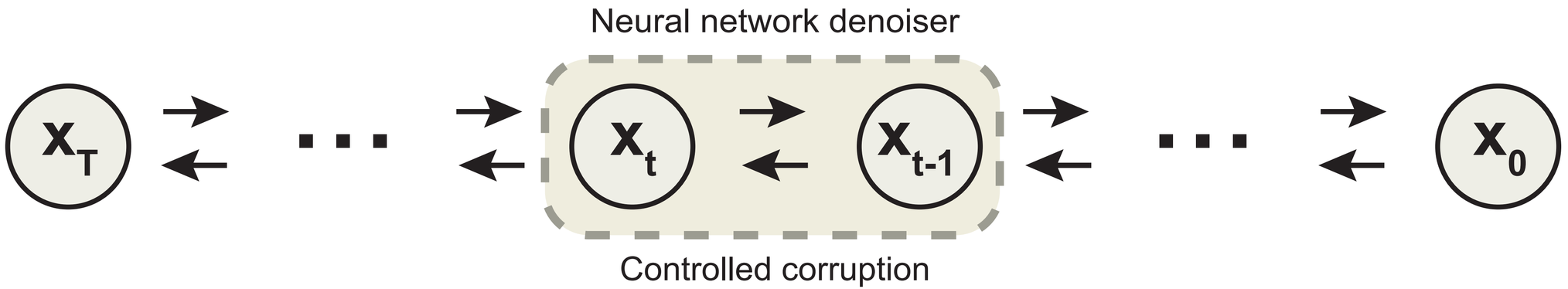


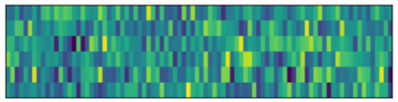


real structures

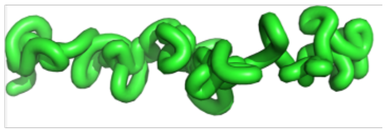


Diffusion models generate realistic structures

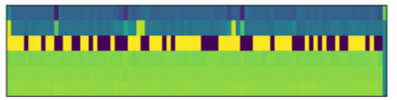




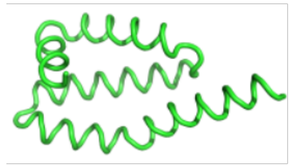
random structures



Wu *et al., preprint 2022*



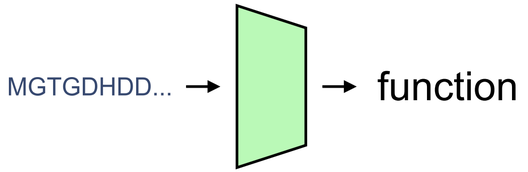
generated structures



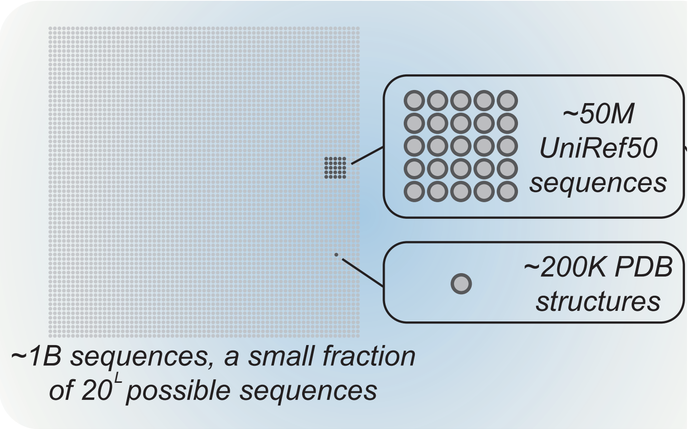
Sequence is the universal protein design space

No sequence design

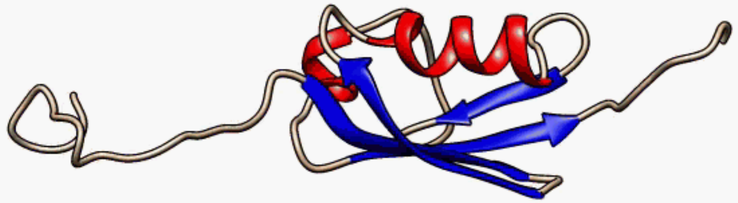
Many function predictors



Many more sequences than structures



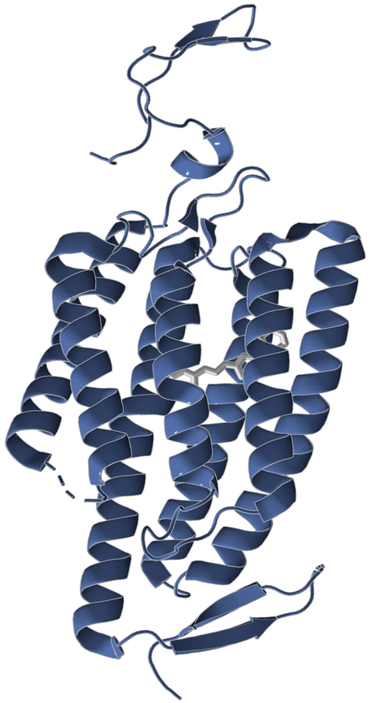
Handle disordered regions

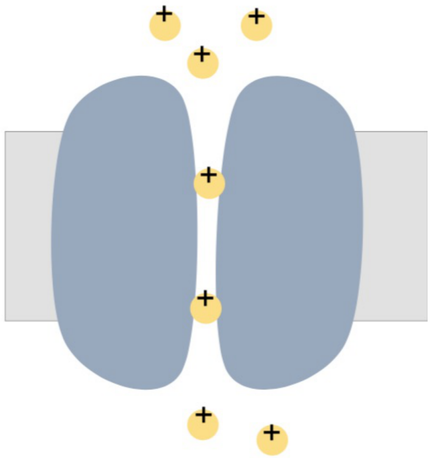


By Lukasz Kozlowski - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=36298697

Sequence is the universal protein design space

MGTGDHDD...

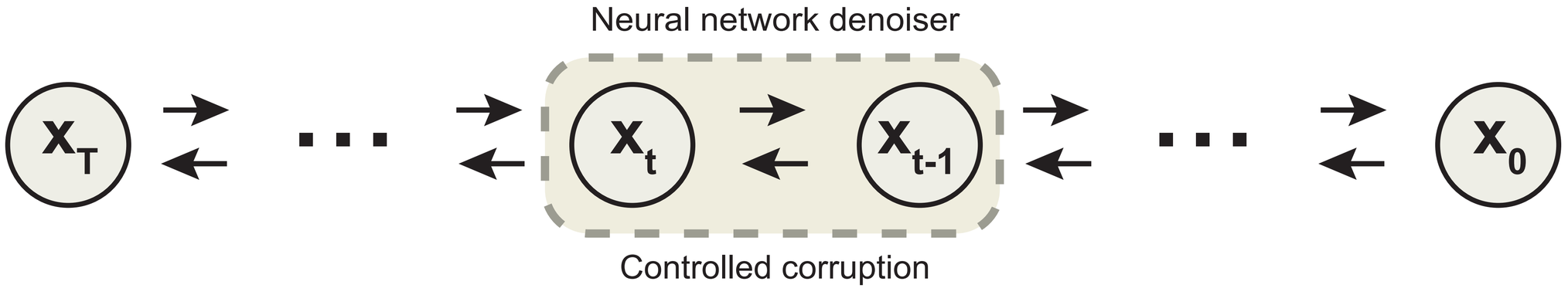




Sequence is the universal protein design space

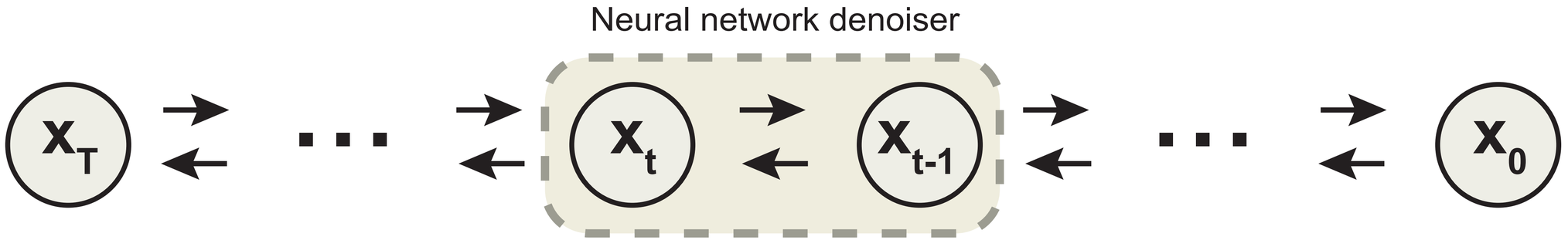
MGTGDHDD...

EvoDiff: evolutionary-scale diffusion



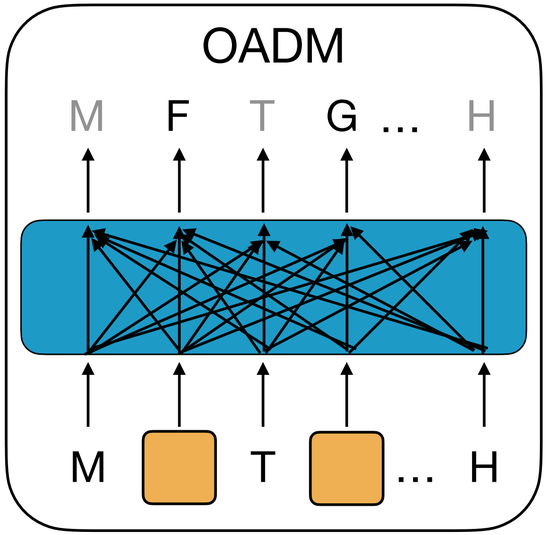
MGTGDHDD…

EvoDiff: evolutionary-scale diffusion

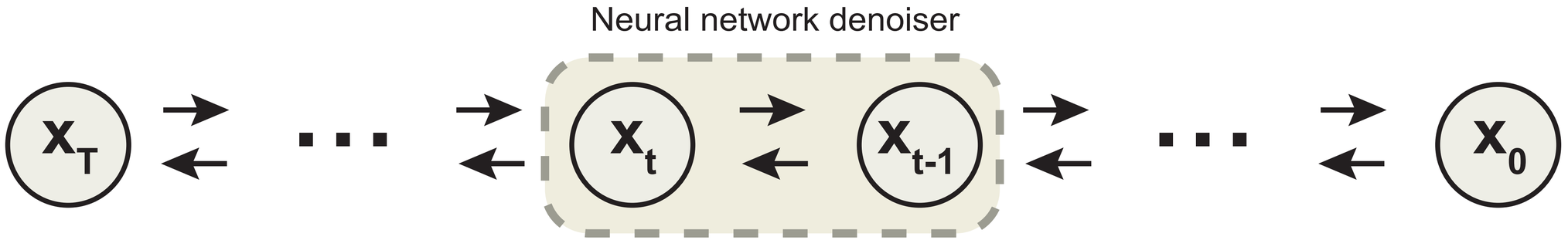


OADM ######## … → MG##D#DD… → MG#GD#DD… → MGTGDHDD…

mask

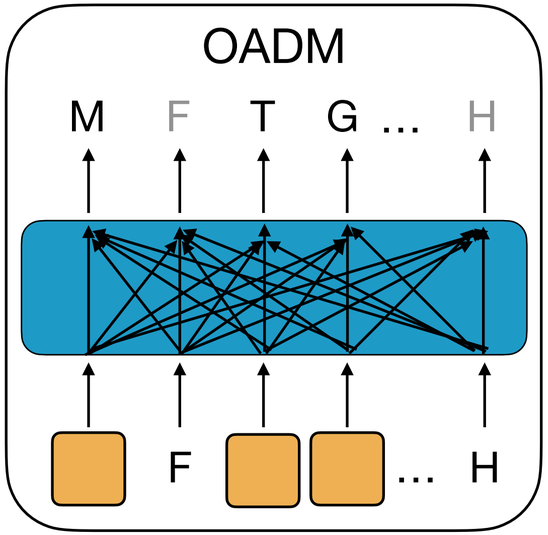


EvoDiff: evolutionary-scale diffusion

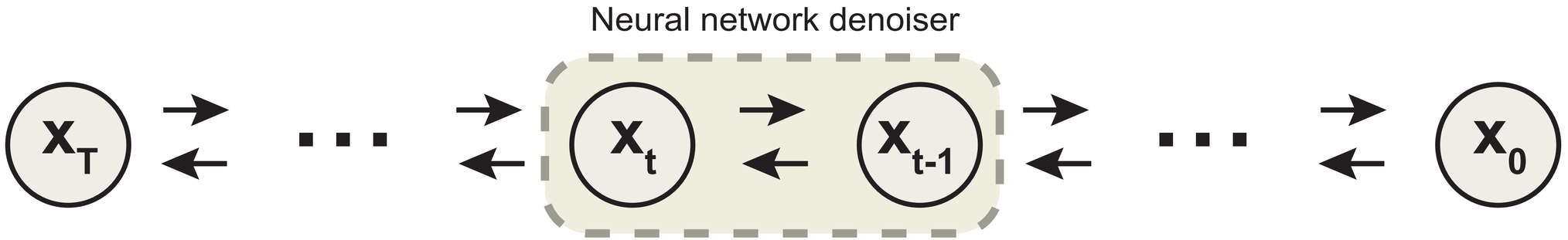


OADM ######## … → MG##D#DD… → MG#GD#DD… → MGTGDHDD…

mask

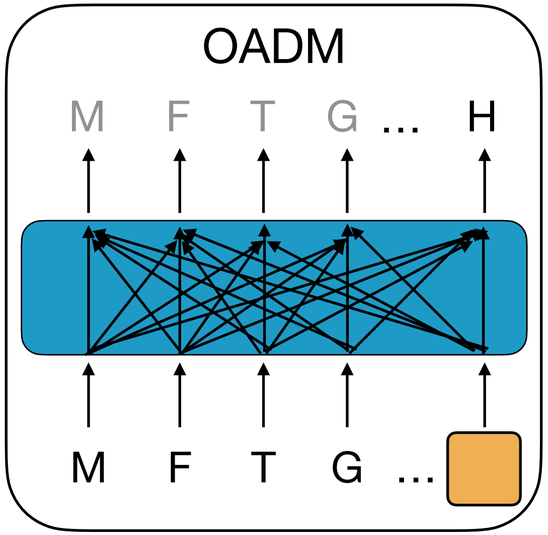


EvoDiff: evolutionary-scale diffusion

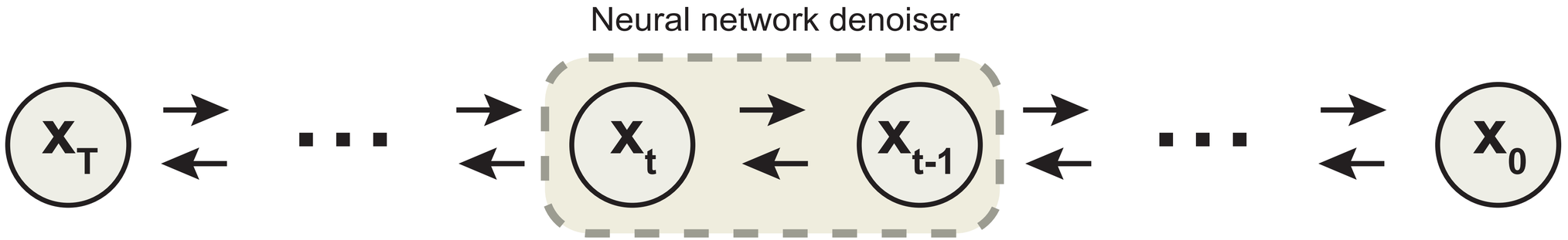


OADM ######## … → MG##D#DD… → MG#GD#DD… → MGTGDHDD…

mask



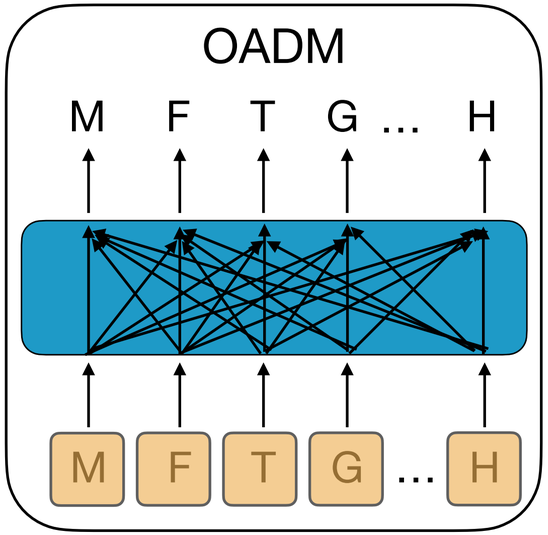
EvoDiff: evolutionary-scale diffusion



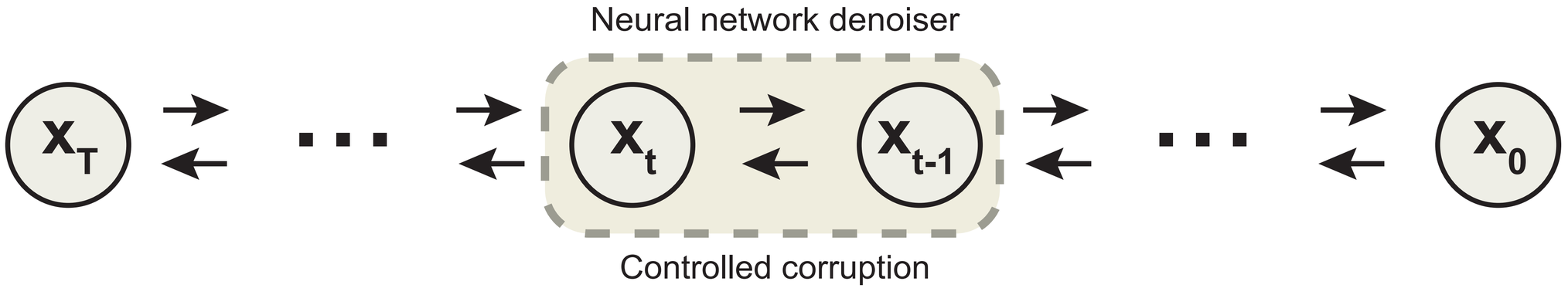
OADM ########… → MG##D#DD… → MG#GD#DD… → MGTGDHDD…

Hoogeboom *et al., ICLR* 2022

mask



EvoDiff: evolutionary-scale diffusion



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

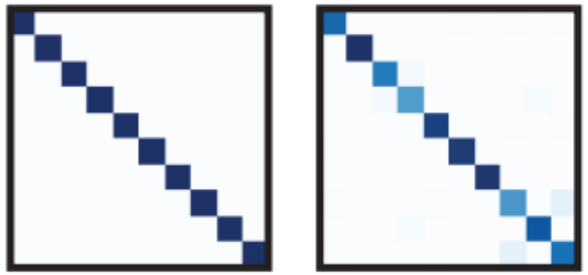
mask

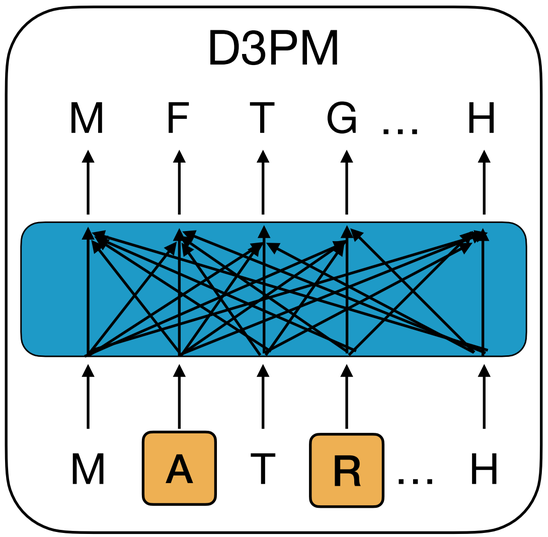
D3PM TYLPLKNE… → MAKTDRDL… → MAGTDRDD… → MGTGDHDD…

mutate

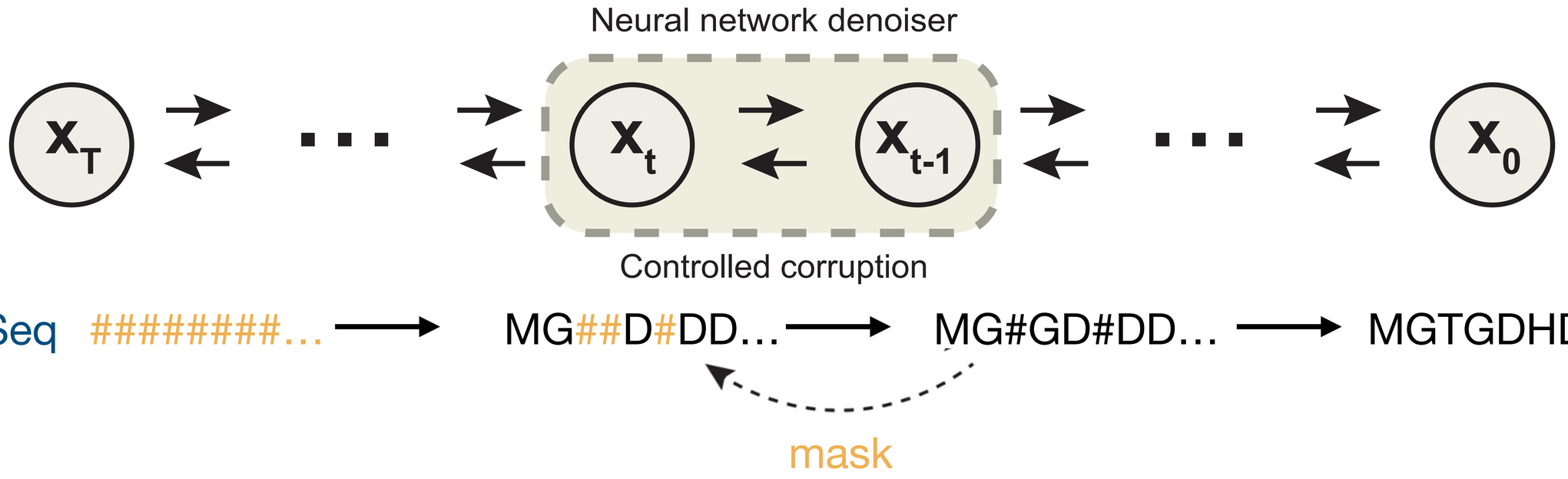
Austin *et al., NeurIPS* 2021

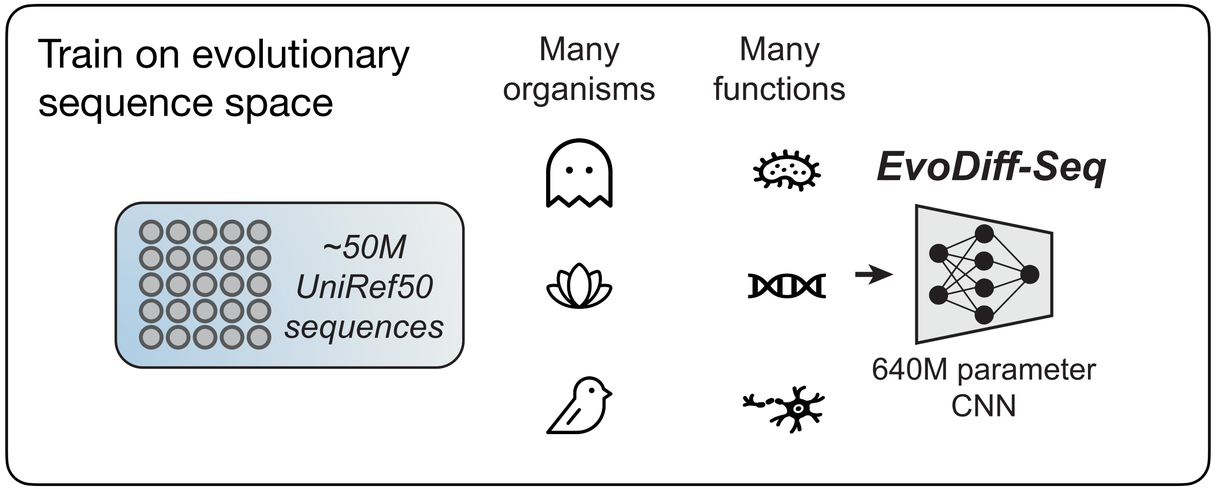
Transition matrix



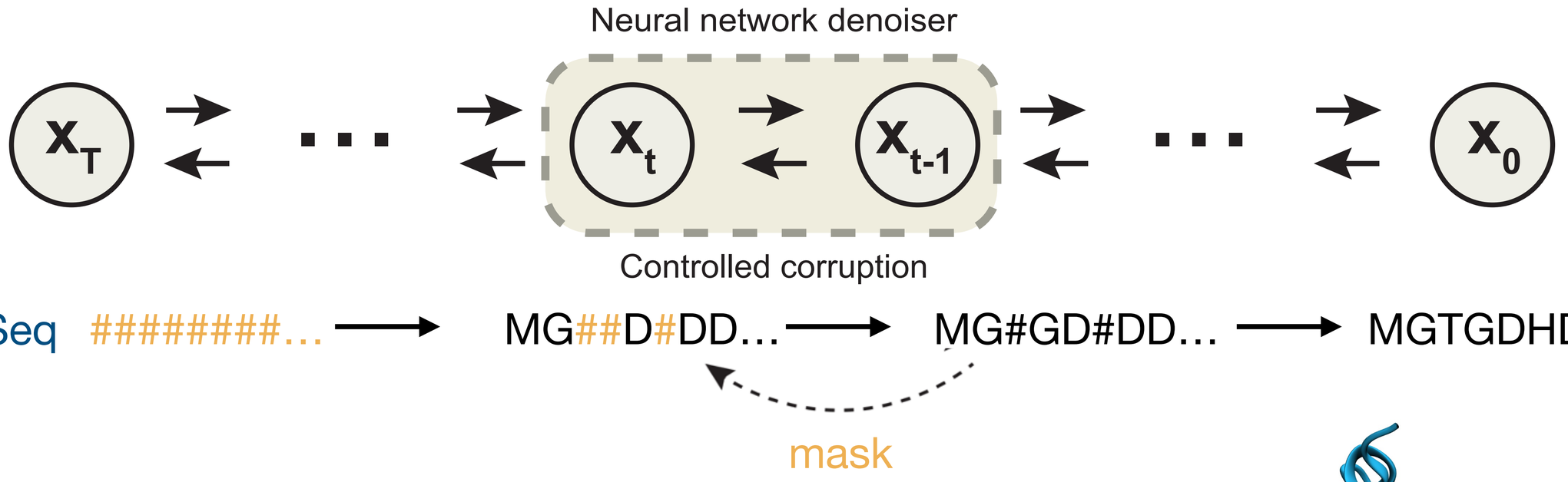


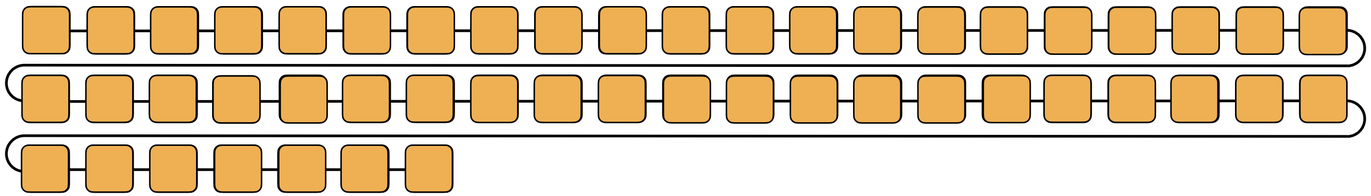
EvoDiff: evolutionary-scale diffusion

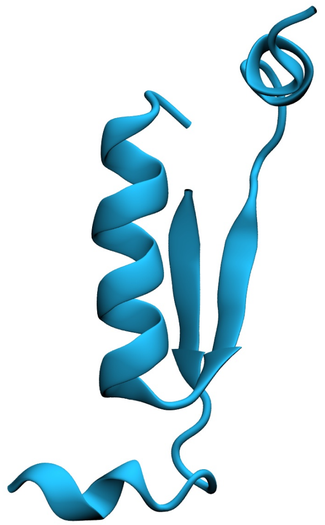




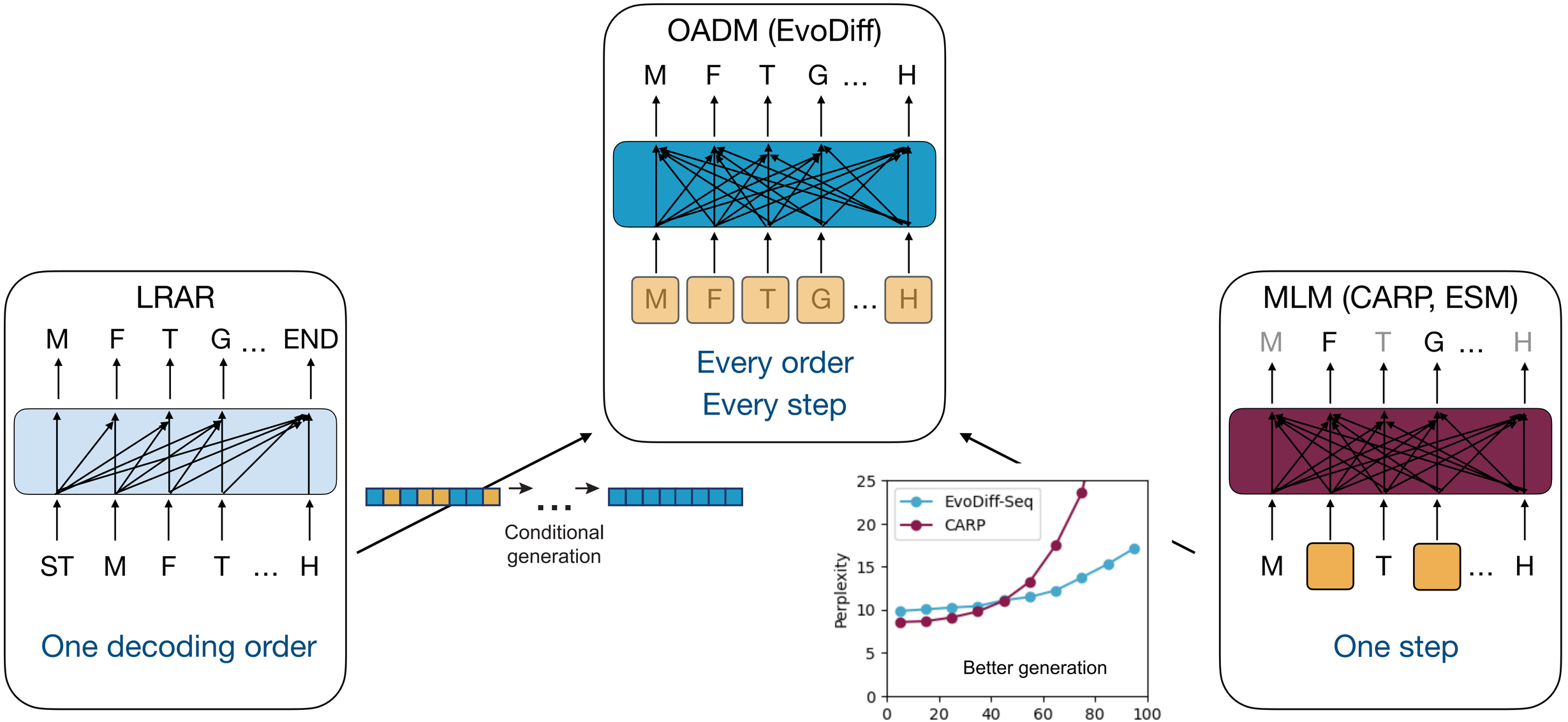
EvoDiff: evolutionary-scale diffusion





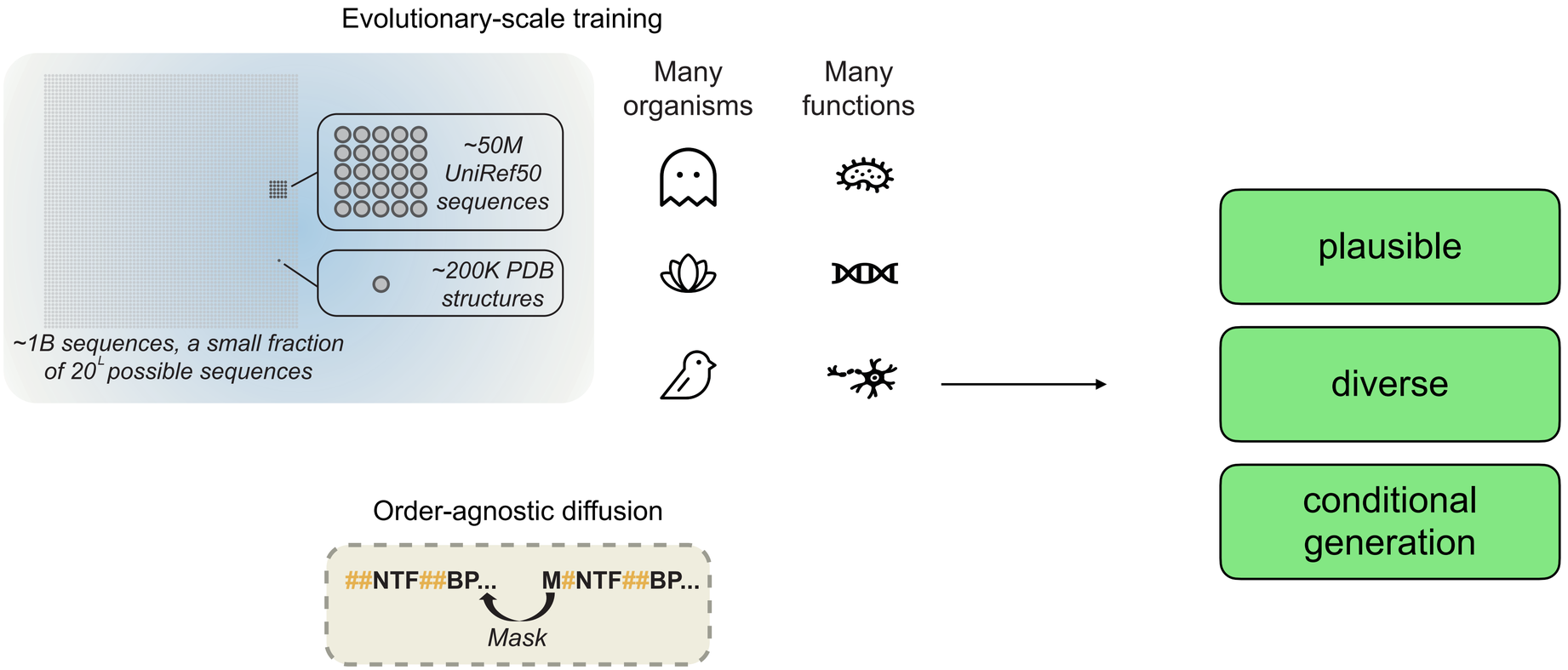


EvoDiff-Seq generalizes masked and autoregressive language models

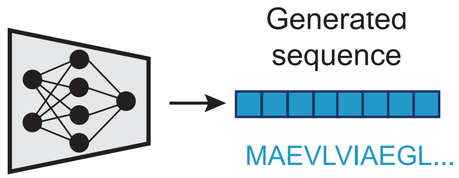


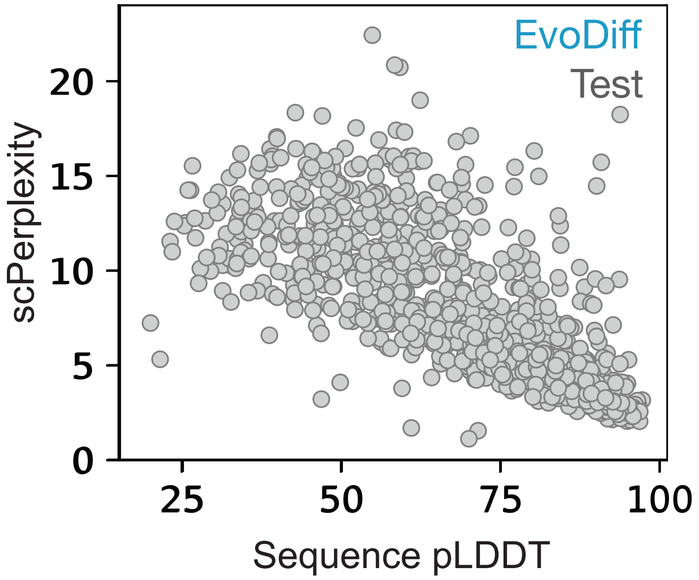
CARP: Yang *et al.,* preprint, 2022
ESM: Rives *et al., PNAS, 2021*

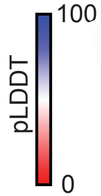
EvoDiff enables controllable generation of plausible, diverse proteins

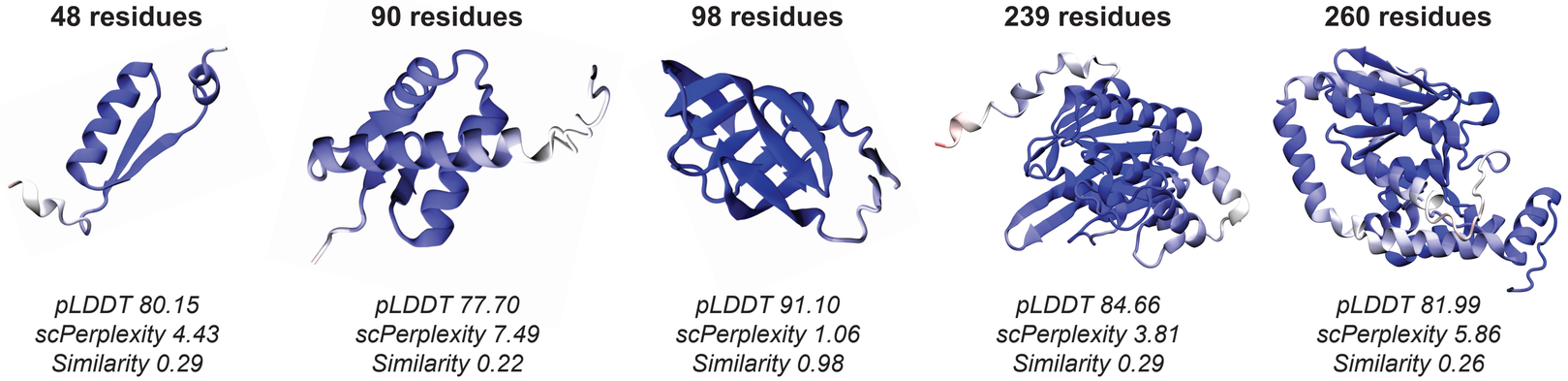


EvoDiff-Seq generates highly-plausible proteins



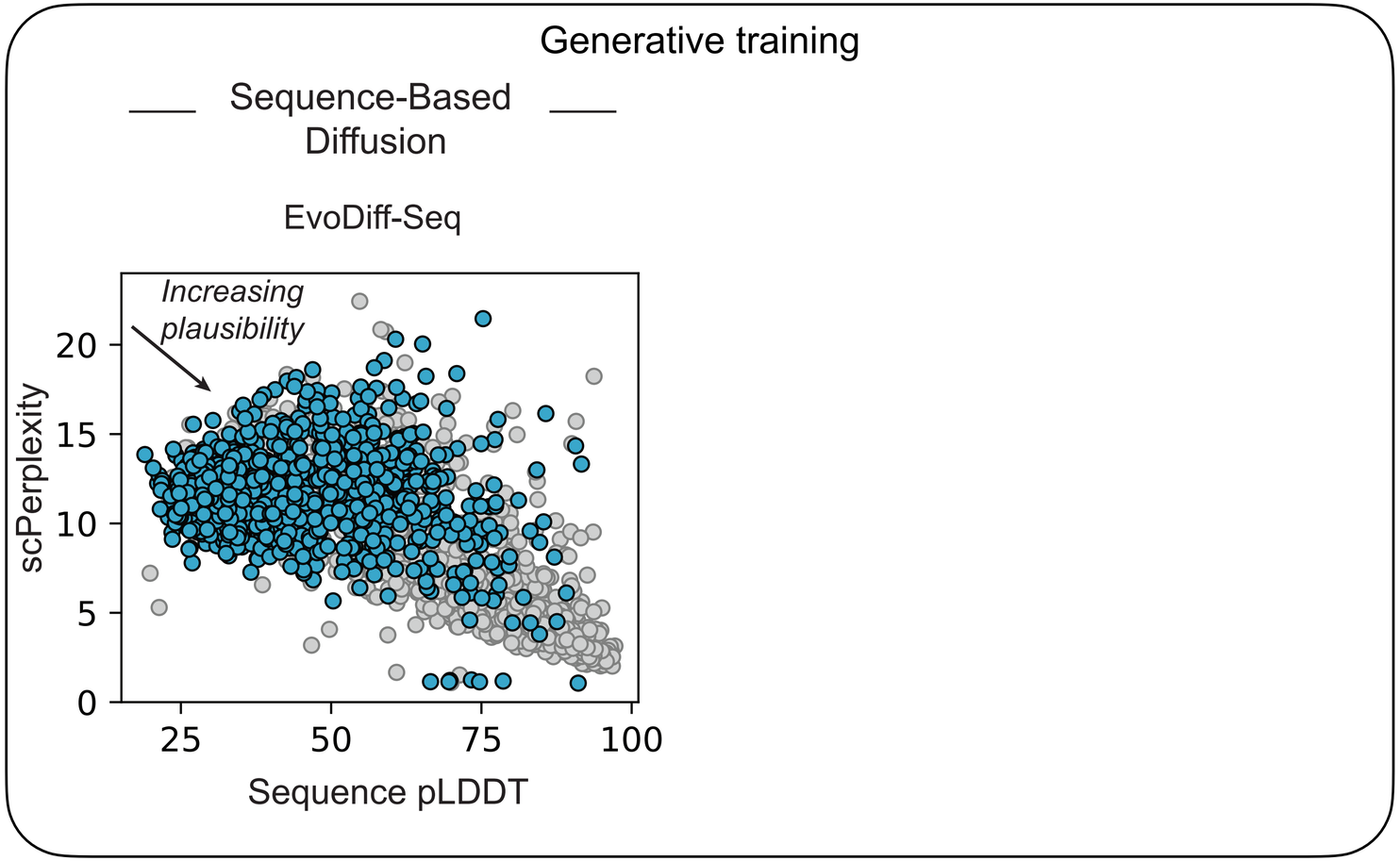






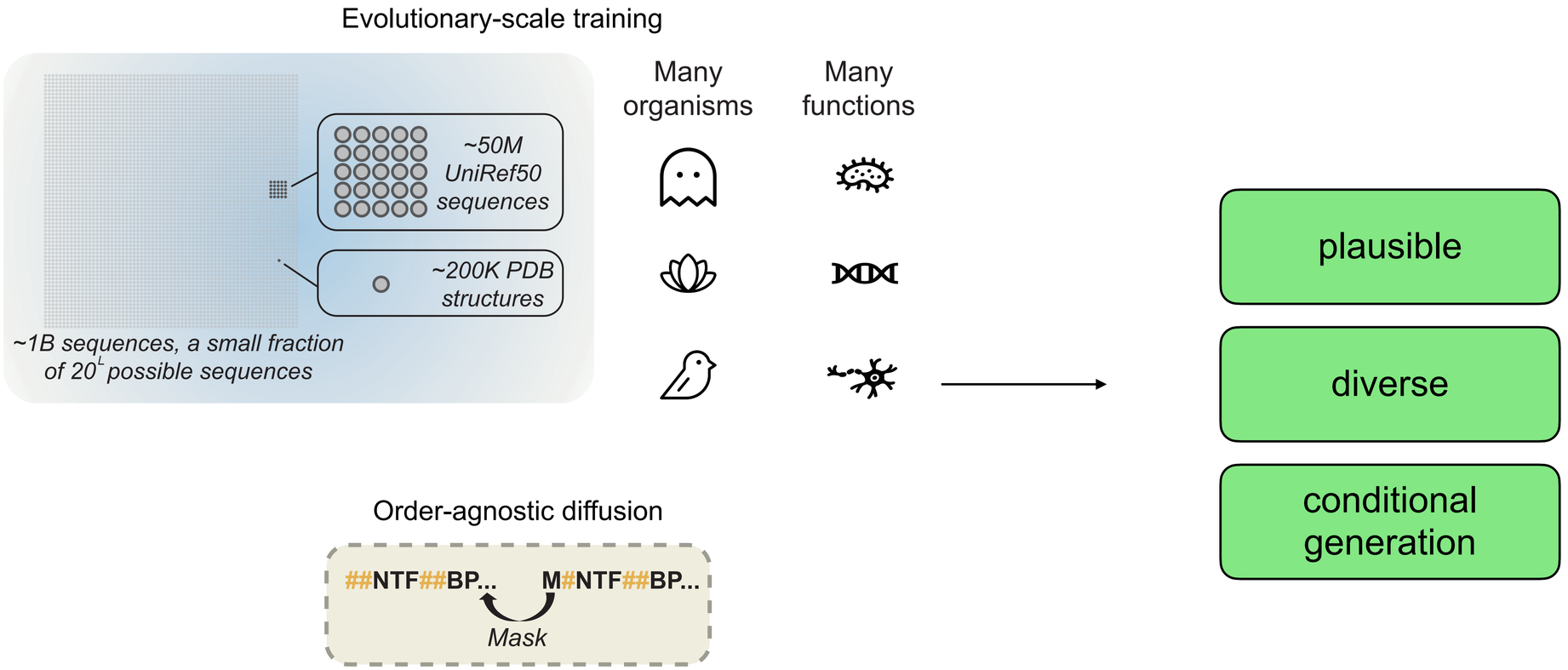
plausible

Generative training results in better sequences

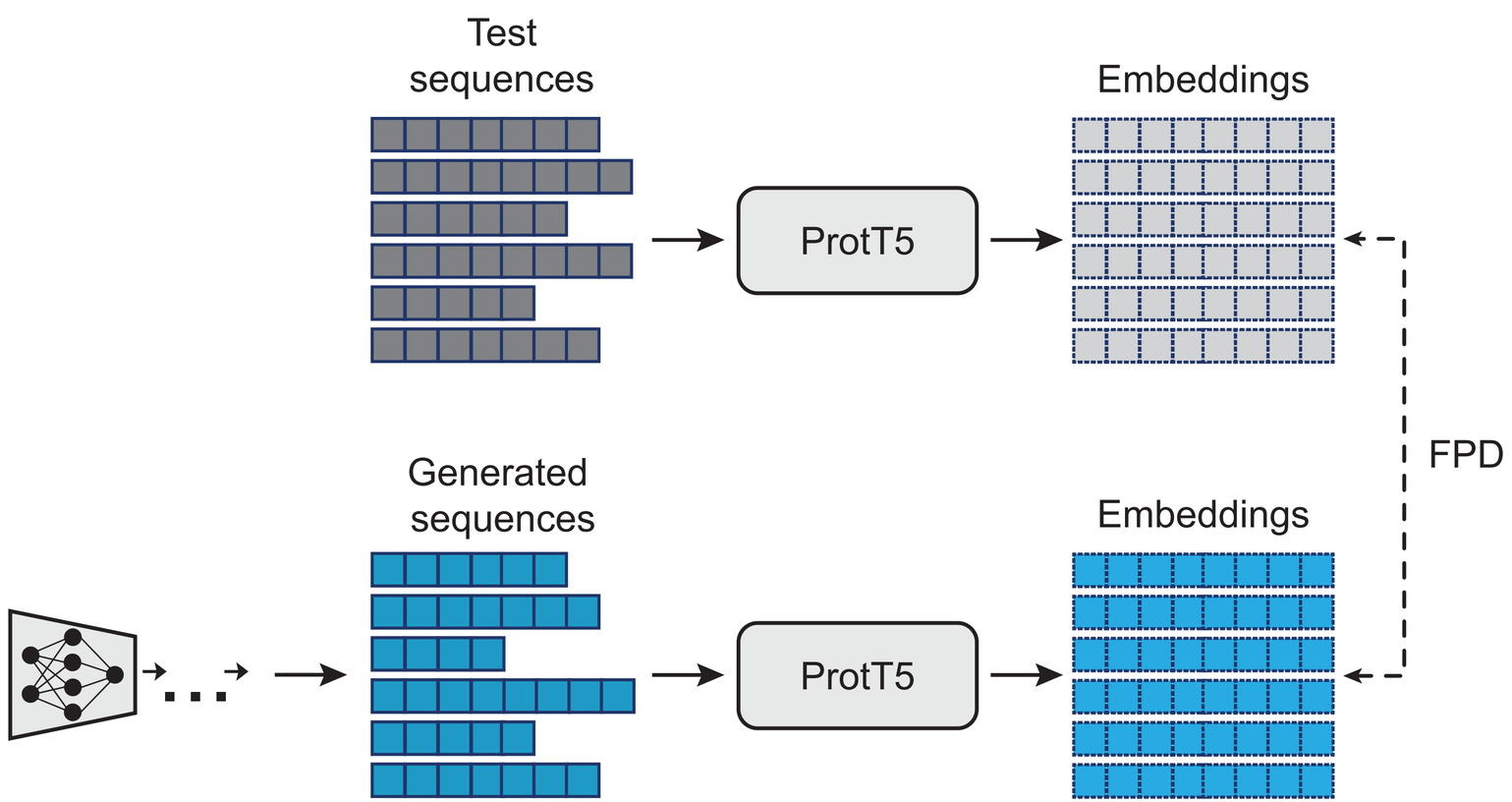


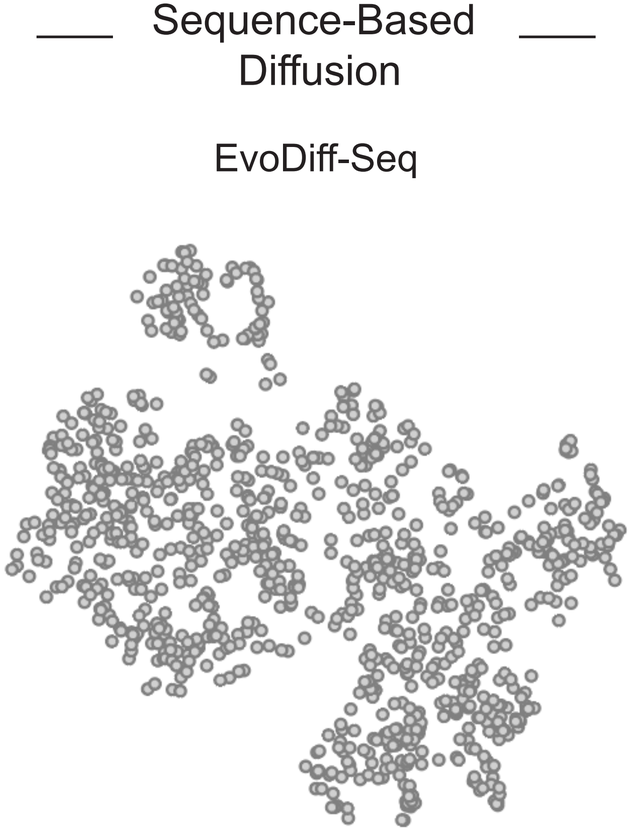
plausible

EvoDiff enables controllable generation of plausible, diverse proteins



EvoDiff-Seq recapitulates natural functional distribution





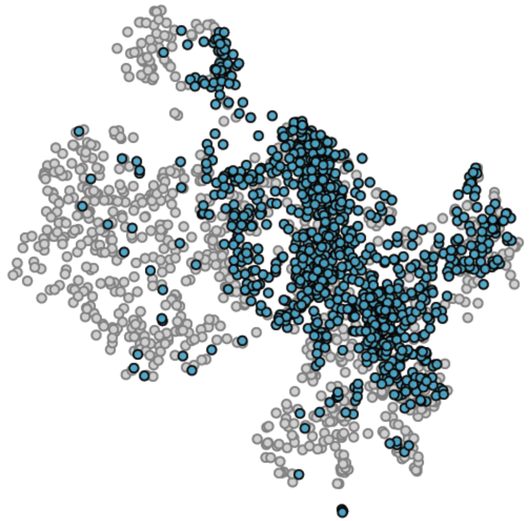
diverse

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

Evolutionary-scale diffusion improves FPD

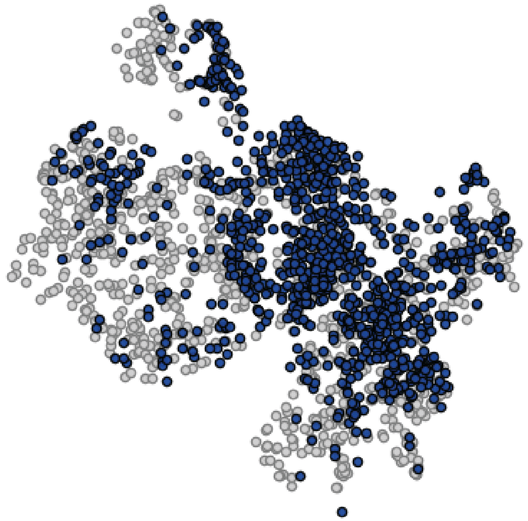
Sequence-Based Diffusion

EvoDiff-Seq
FPD = 0.88



Left-to-right Language Model

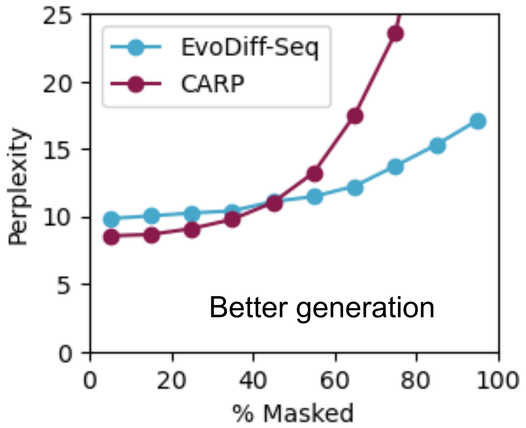
LRAR
FPD = 0.63



Conditional generation

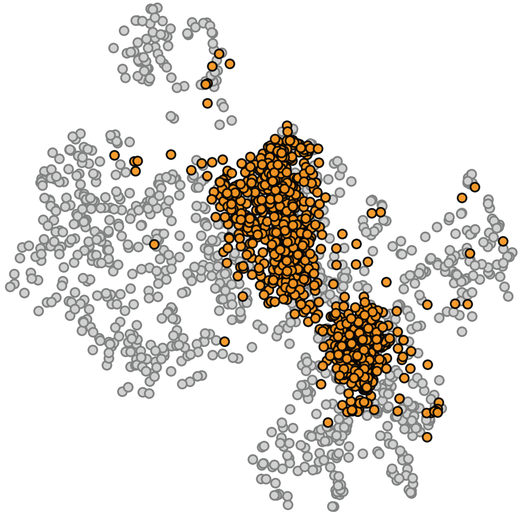
Protein Masked Language Model

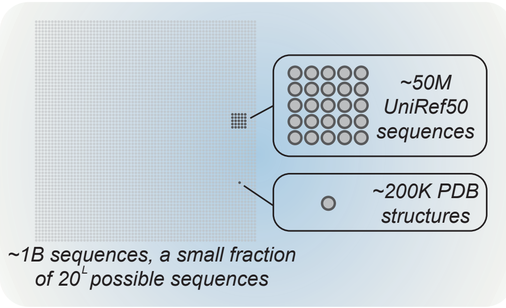
ESM-2
FPD = 2.81



Structure-Based Diffusion

RFdiffusion
FPD = 1.96

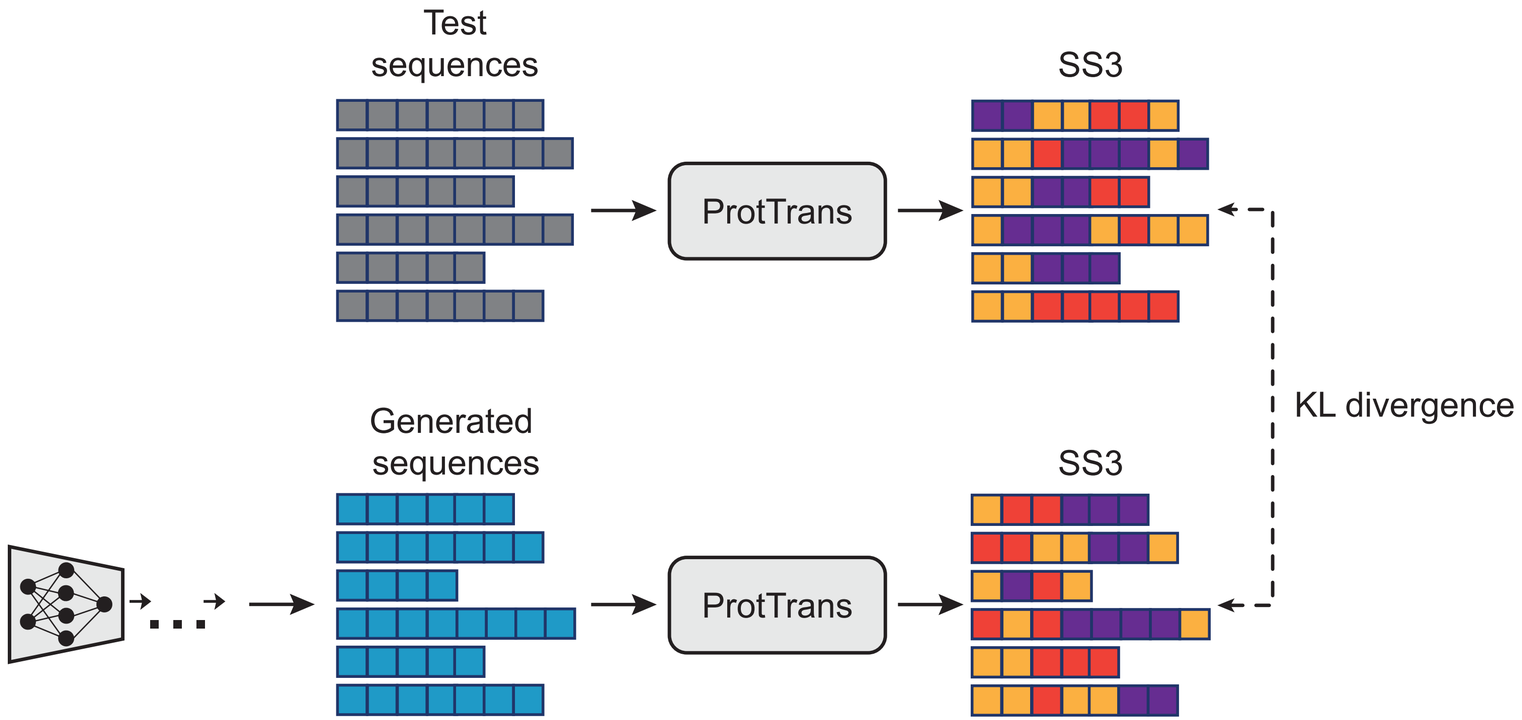


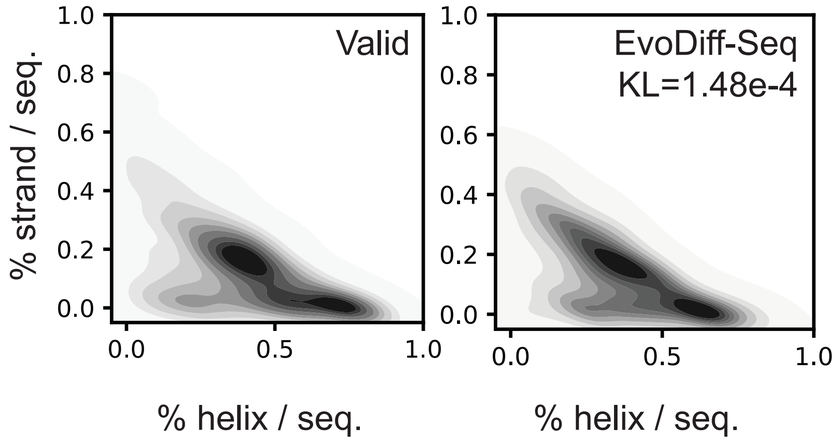


Evolutionary-scale training

diverse

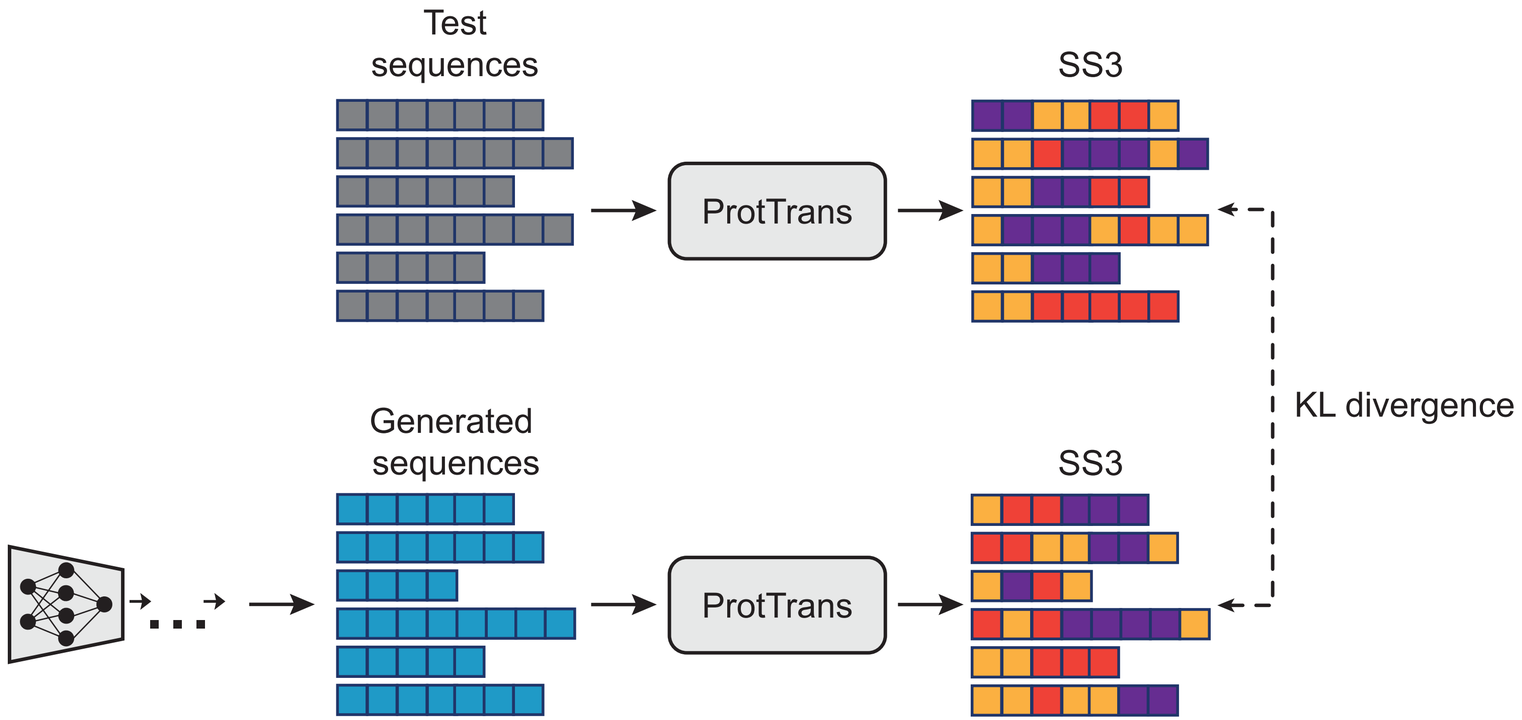
EvoDiff-Seq recapitulates natural secondary structure distribution

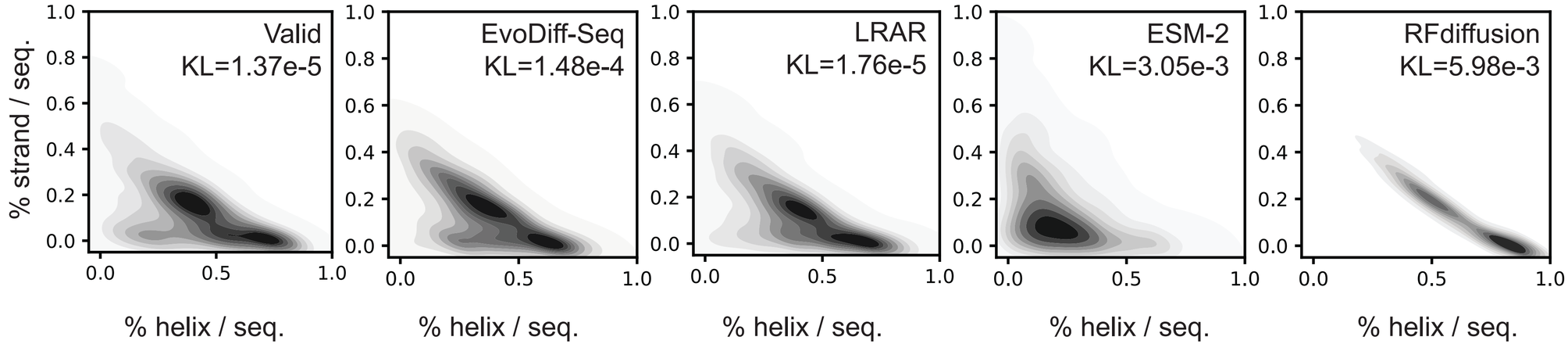




diverse

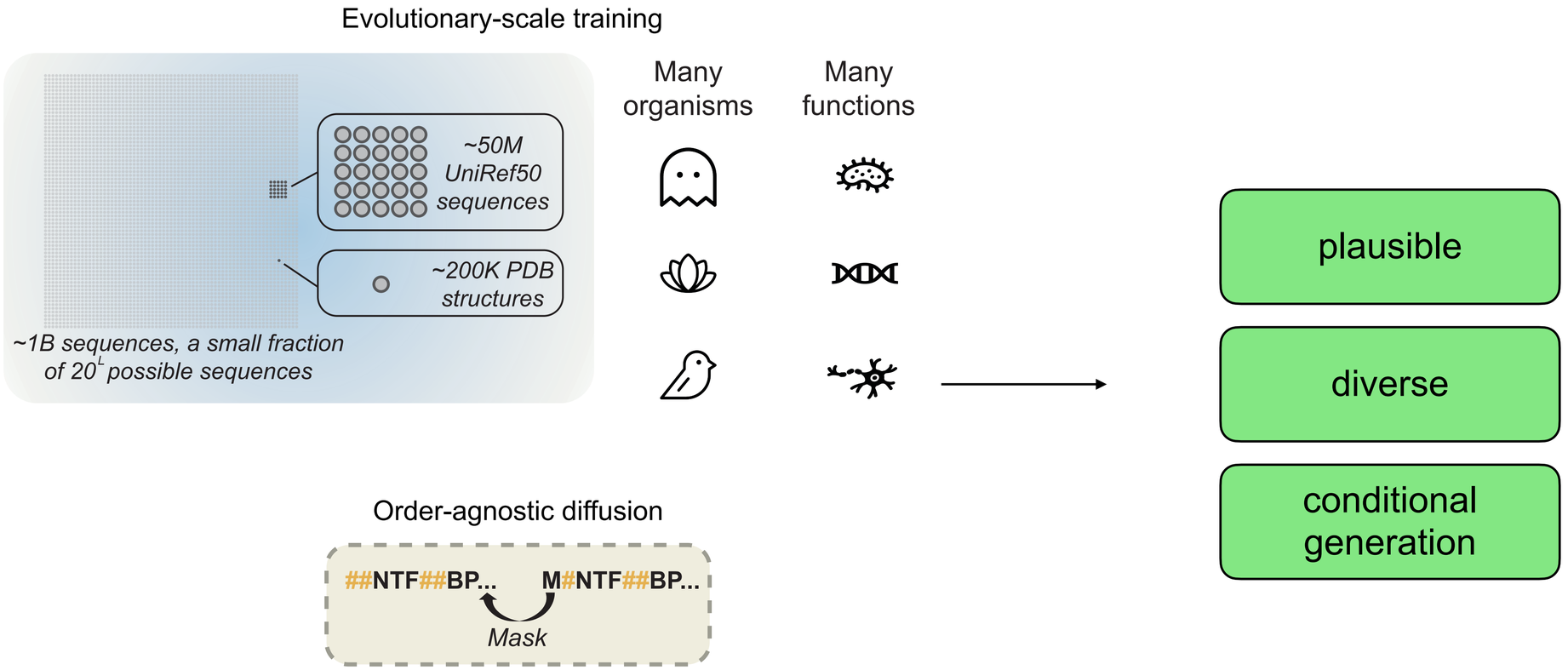
Evolutionary-scale diffusion improves secondary structures





diverse

EvoDiff enables controllable generation of plausible, diverse proteins



EvoDiff-MSA leverages evolutionary information

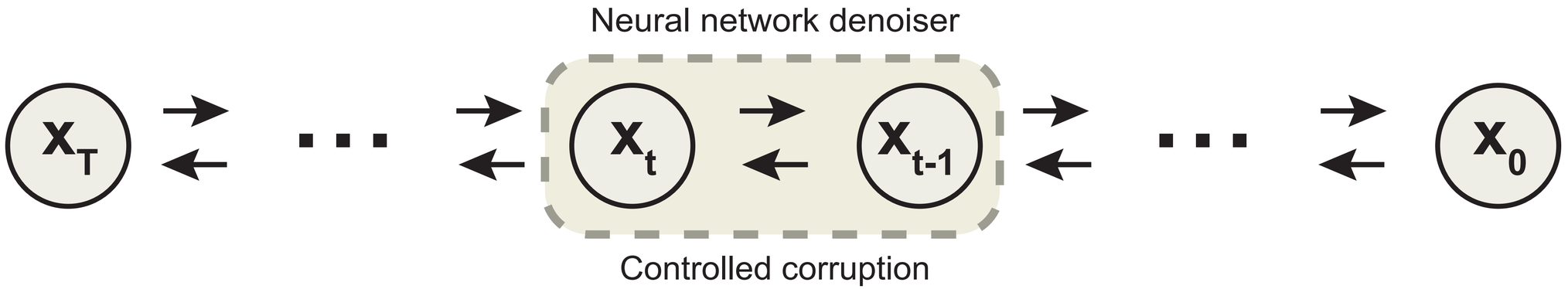
query MFTGNDAGH…

homology search

subsample to 64

conditional generation

EvoDiff-MSA leverages evolutionary information



EvoDiff-MSA

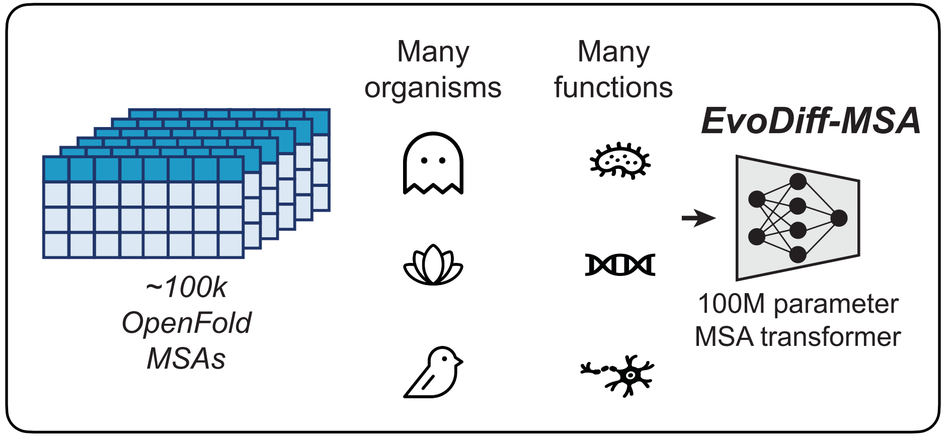
```
########…
########…
########…
########…
…
```

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

mask

```
MFTGN#AGH…
MH#G-DAGL…
MHTGN--GL…
MH#NN#AGL…
…
```

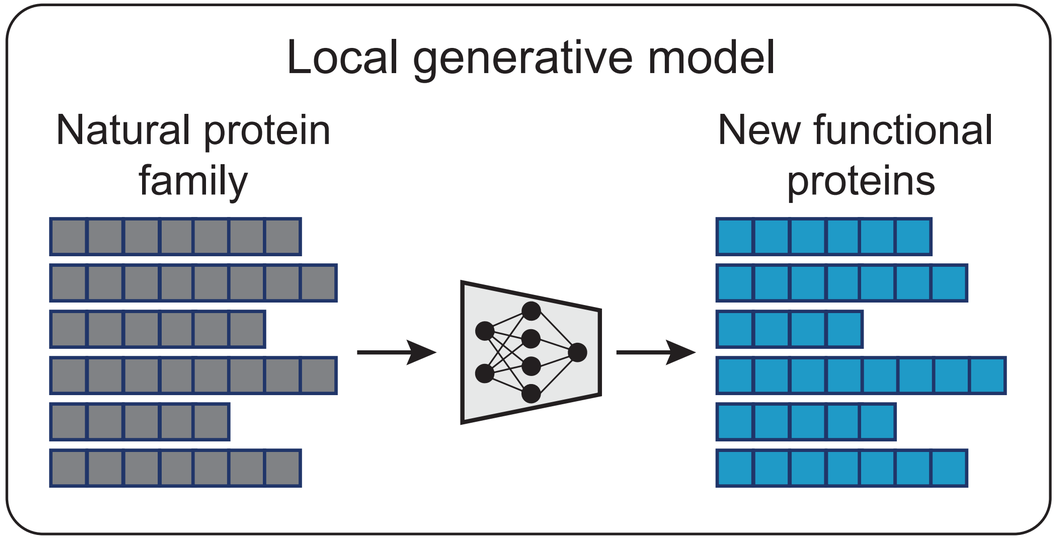
```
MFTGNDAGH…
MHTG-DAGL…
MHTGN--GL…
MHYNNDAGL…
…
```



conditional generation

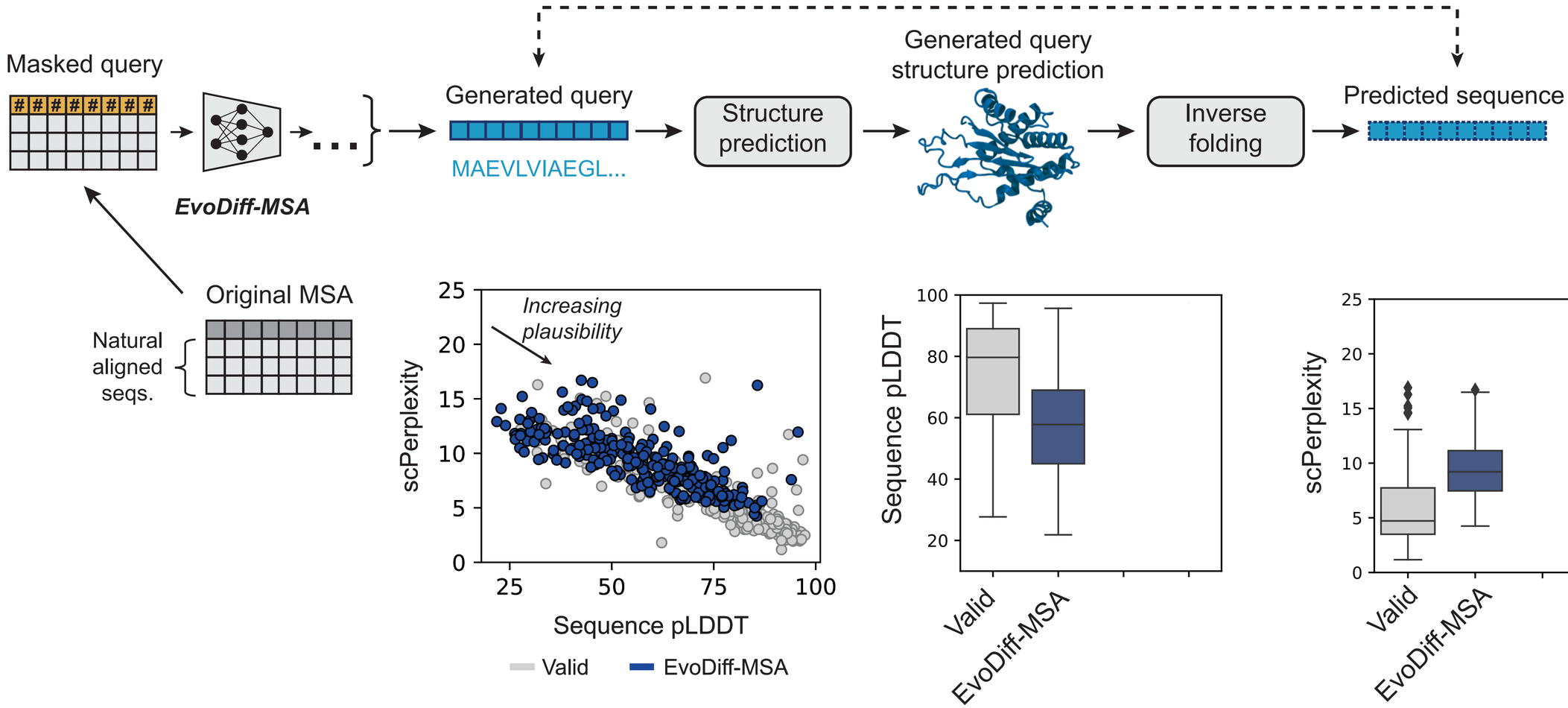
OpenFold: Ahdritz *et al., Preprint* 2022

EvoDiff-MSA is an amortized local generative model



conditional generation

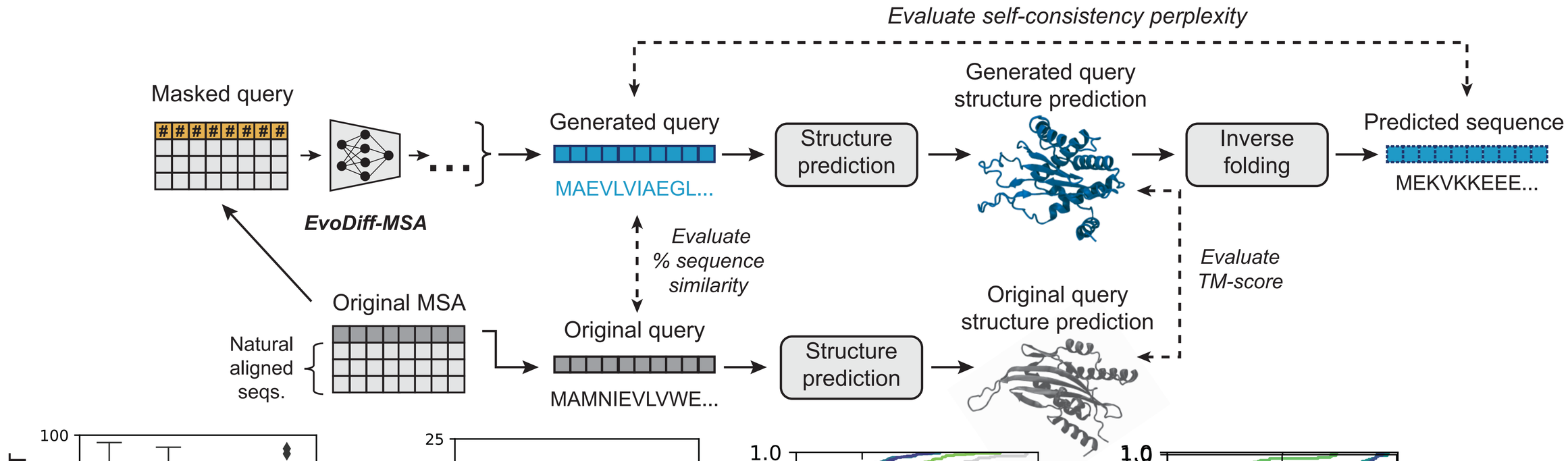
Evolutionary diffusion improves local generation

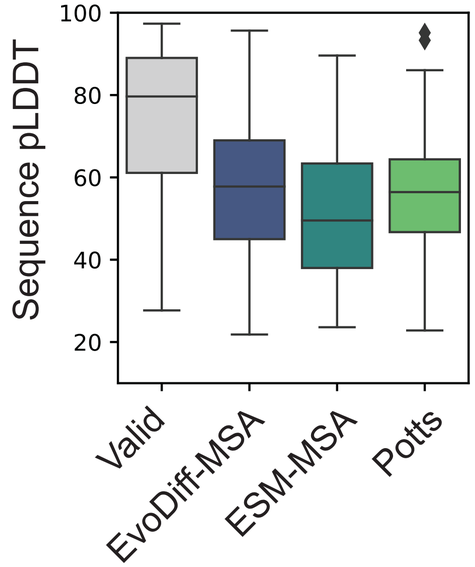


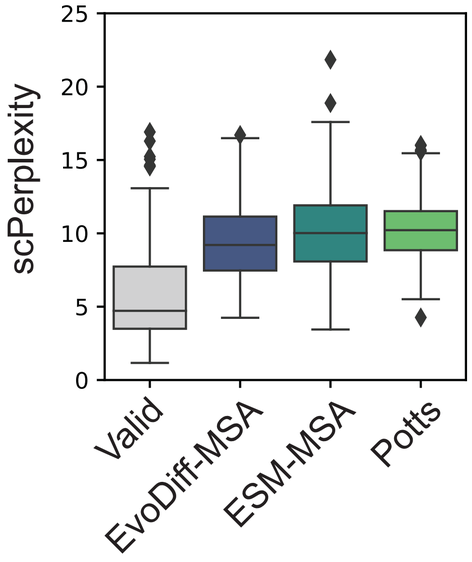
ESM-MSA: Rao *et al., ICML* 2021
CCMgen: Vorberg *et al., PLoS Comp Bio 2018*

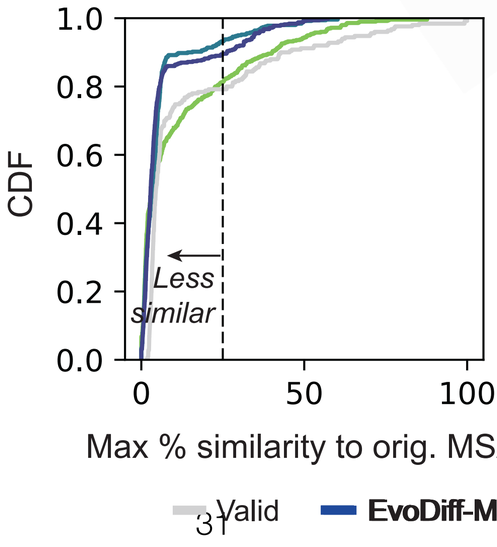
conditional generation

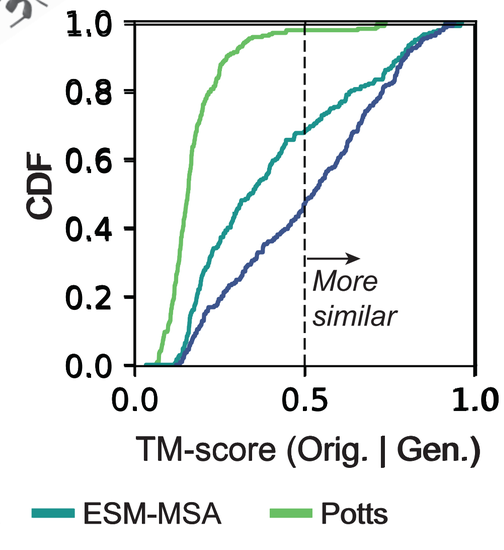
Evolutionary diffusion improves local generation





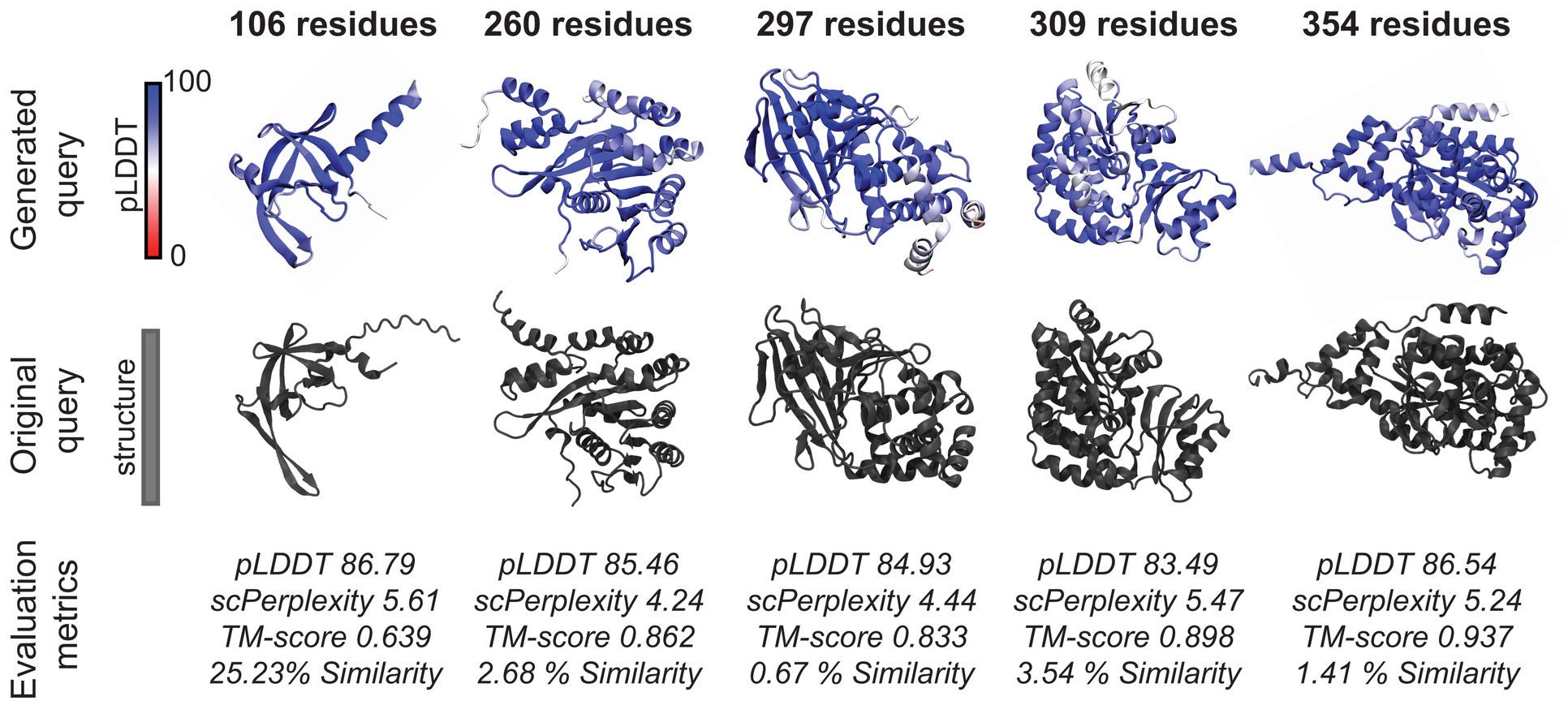






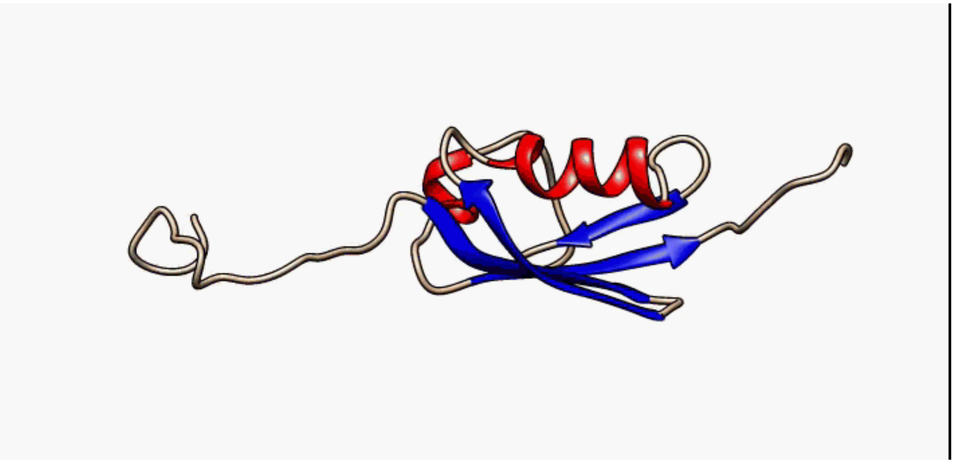
conditional generation

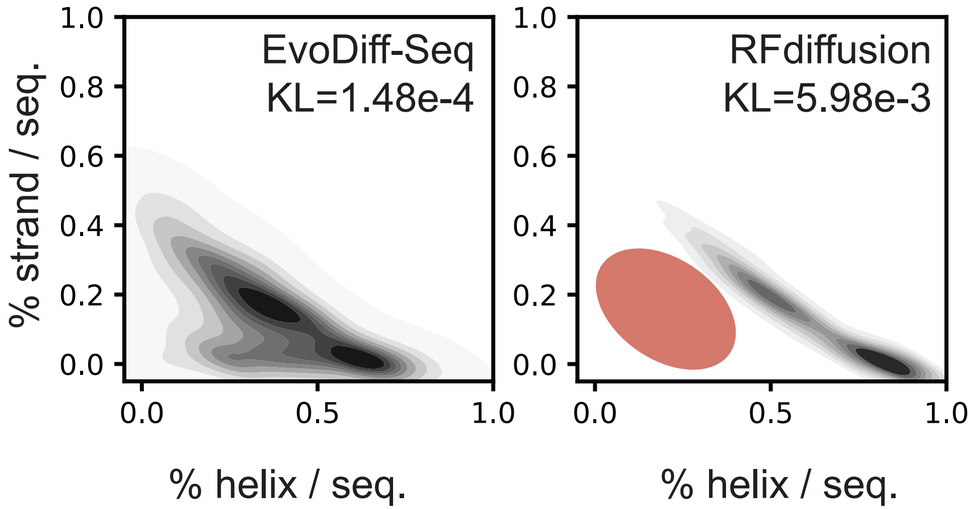
Evolutionary diffusion improves local generation



conditional generation

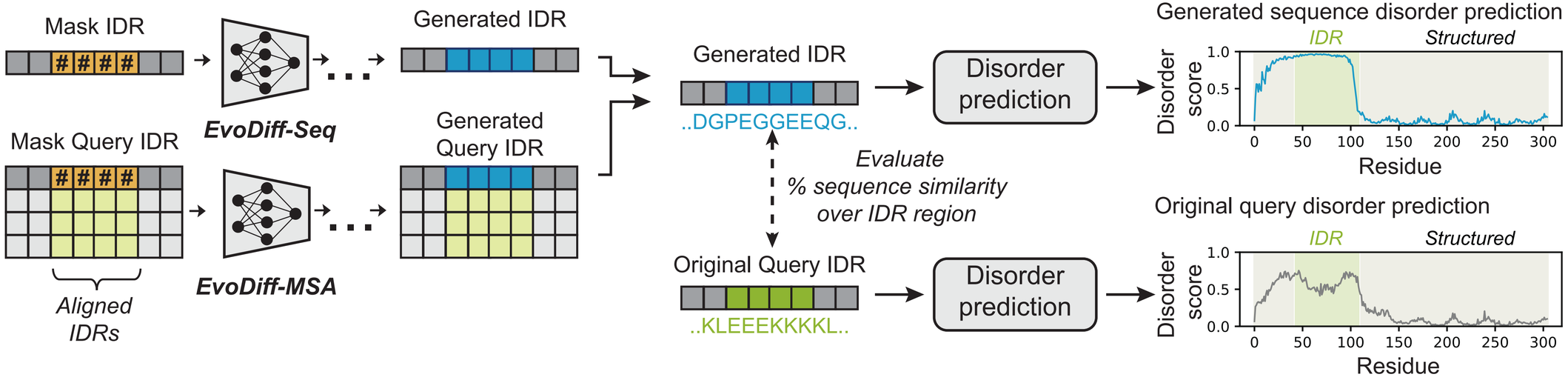
Intrinsically disordered regions perform many cellular functions





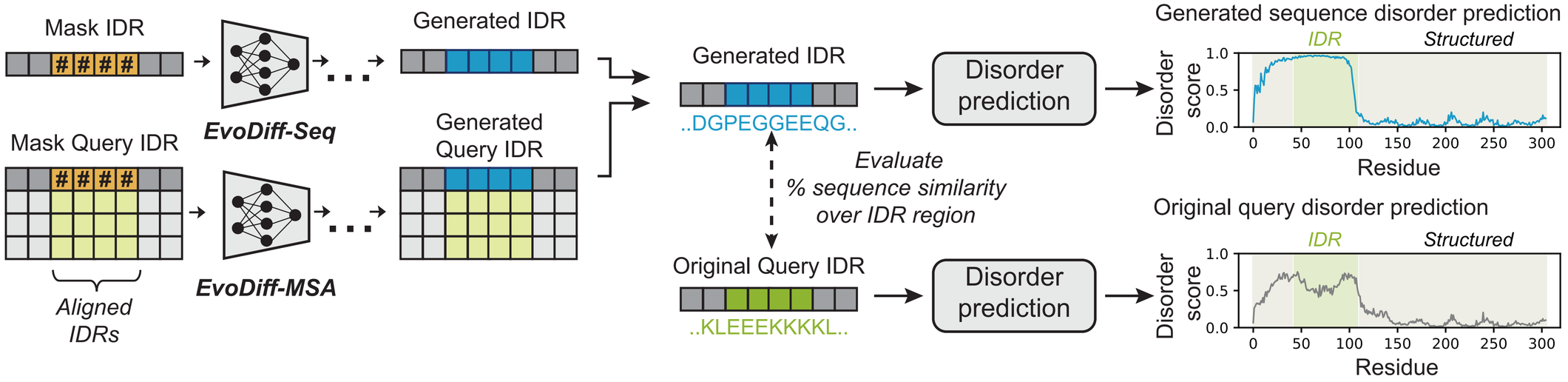
conditional generation

EvoDiff can generate disordered regions



conditional generation

EvoDiff can generate disordered regions



EvoDiff-Seq IDR generation

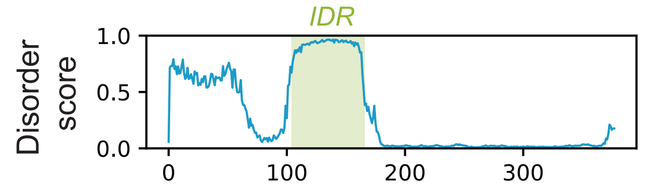
EvoDiff-MSA IDR generation

UniProtID Q96Q15: Serine/threonine protein kinase

Sequence disorder prediction

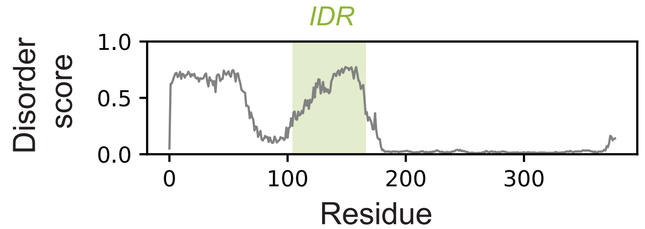
IDR sequence

Gen. IDR



PEEETTIAGVYTTSDSKVQDSS
STGPTIGEQIQKANLESSVFGS
SAPAAGDADQDGSQAQPV

True IDR



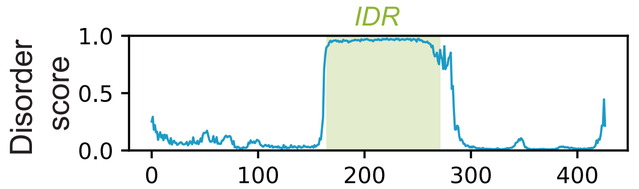
PSQNQNHLKDWFLENKSEVCE
CRNNEDGPGLIMEEQHKCSSK
SLEHKTQTPPVEENVTQKIS

17.7% Similarity

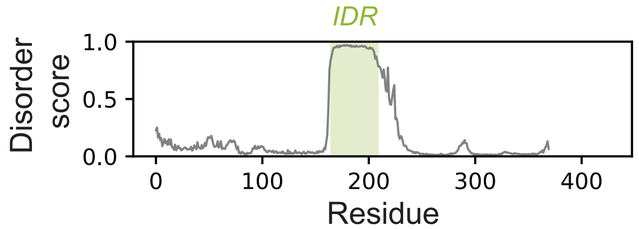
UniProtID F5H008: Endosome/lysosome vesicle associated protein

Sequence disorder prediction

IDR sequence



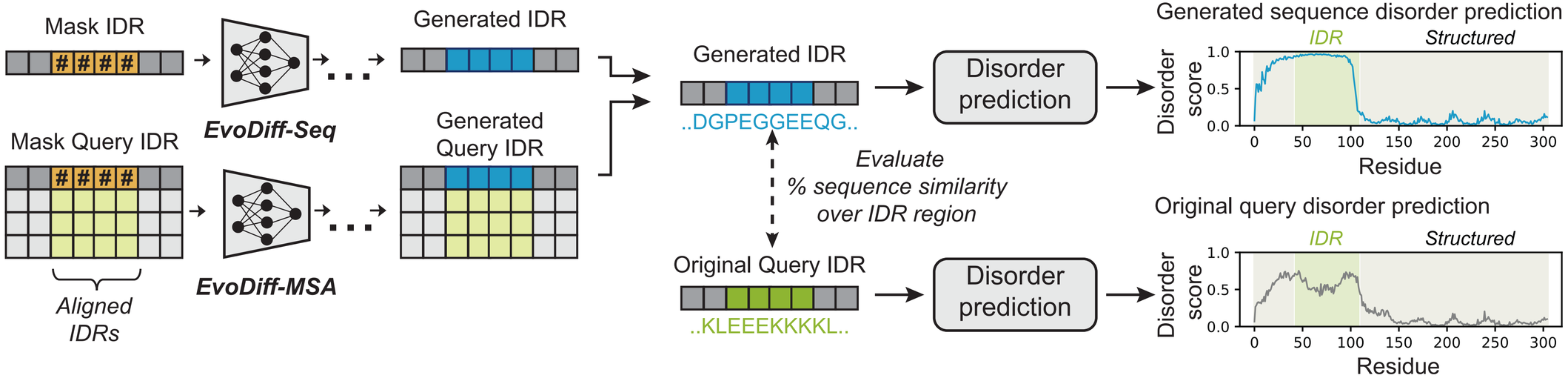
NNDEINGKQSIEKKTEKLGFKE
EEDAVHAIEKNKISGNTKAKKK
KEEDGKNKGELLAKAMDKPH-
L--EEKEKEKKDKEEE------QEQ
EKPRIE--DIGSDEENE-K----GK
DKKKKTKK

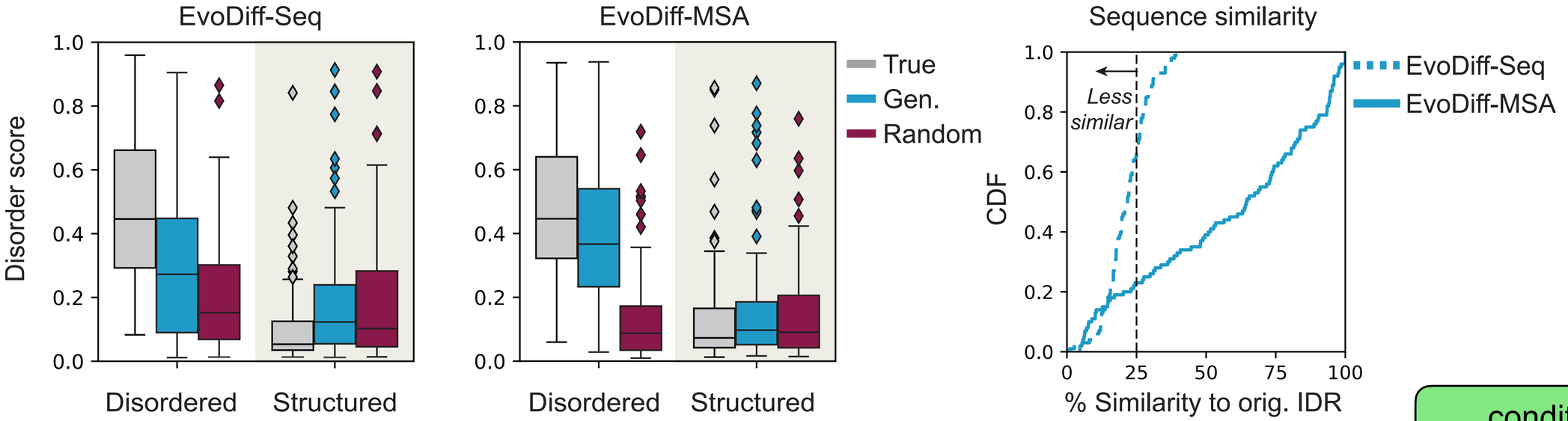


EKEISDDEA---E----E----------
----------E-K
GE--K-EEE--DK--DDEEKPKIE--
DVGSDEEDD-S----GKDKKKK
TKK

37.0% Similarity

EvoDiff can generate disordered regions

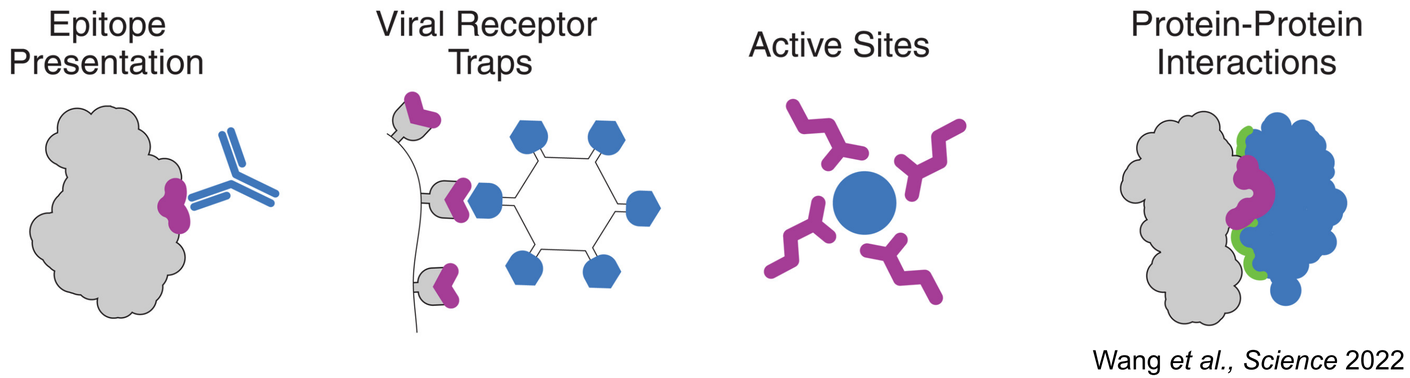




conditional generation

DR-BERT: Nambiar *et al., Preprint* 2023

Many functions are mediated by a motif stabilized by a scaffold



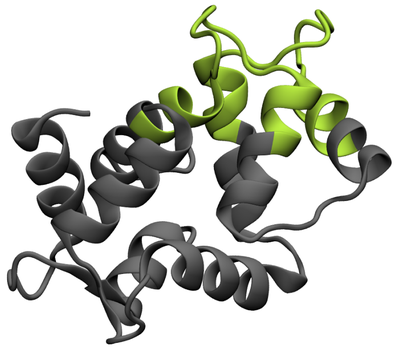
Wang *et al., Science* 2022

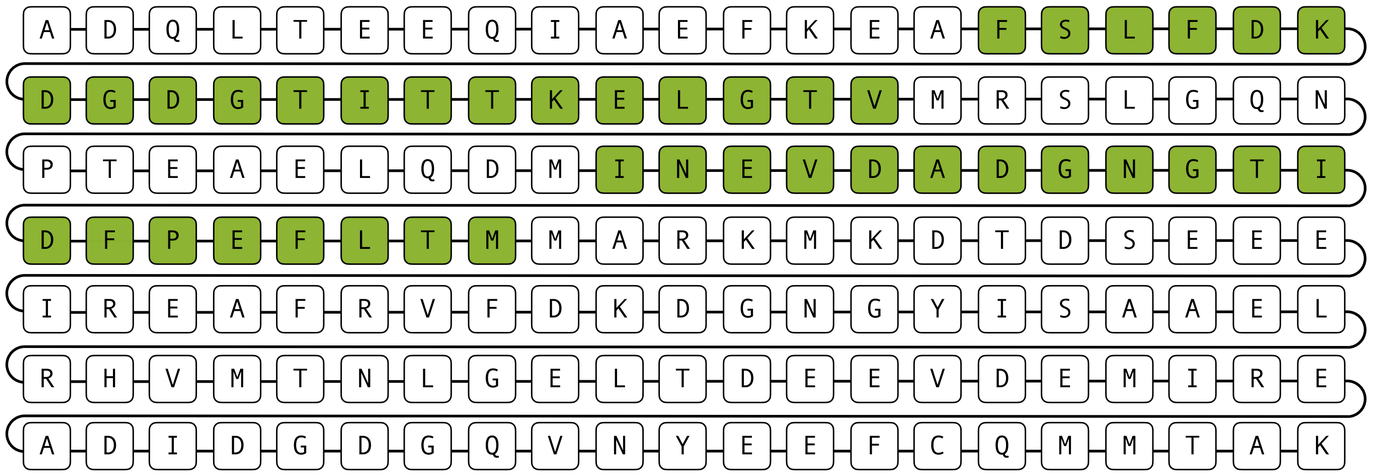
Can we scaffold motifs in sequence space?

conditional generation

EvoDiff can scaffold functional motifs

1PRW: binding site of compact calmodulin

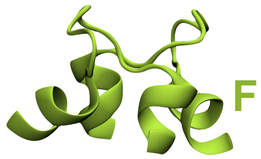




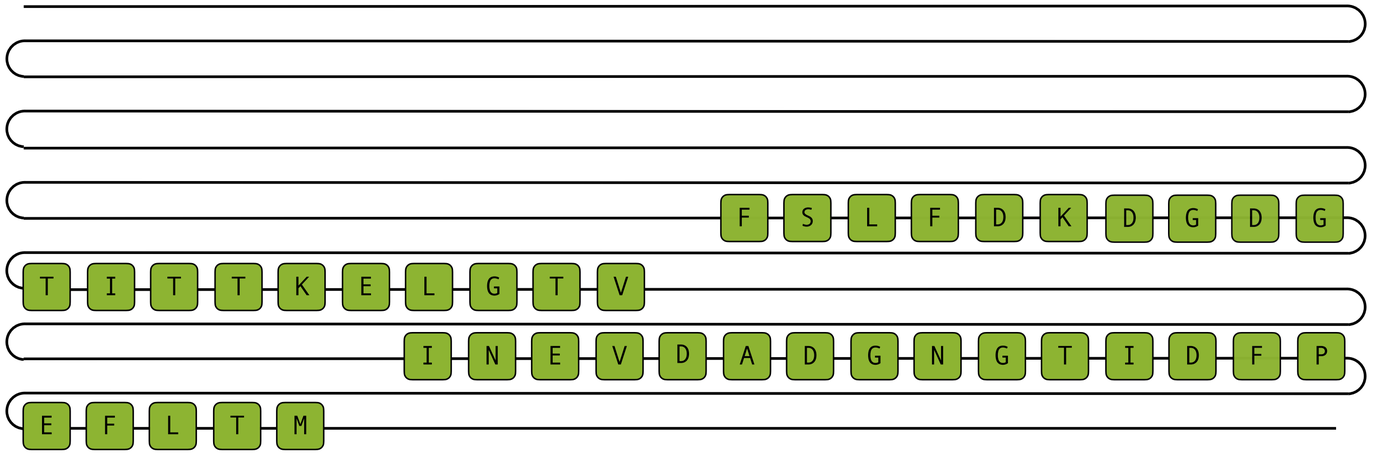
conditional generation

EvoDiff can scaffold functional motifs

1PRW: binding site of compact calmodulin



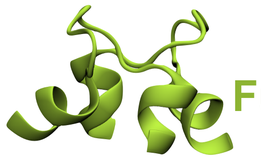
Functional motif



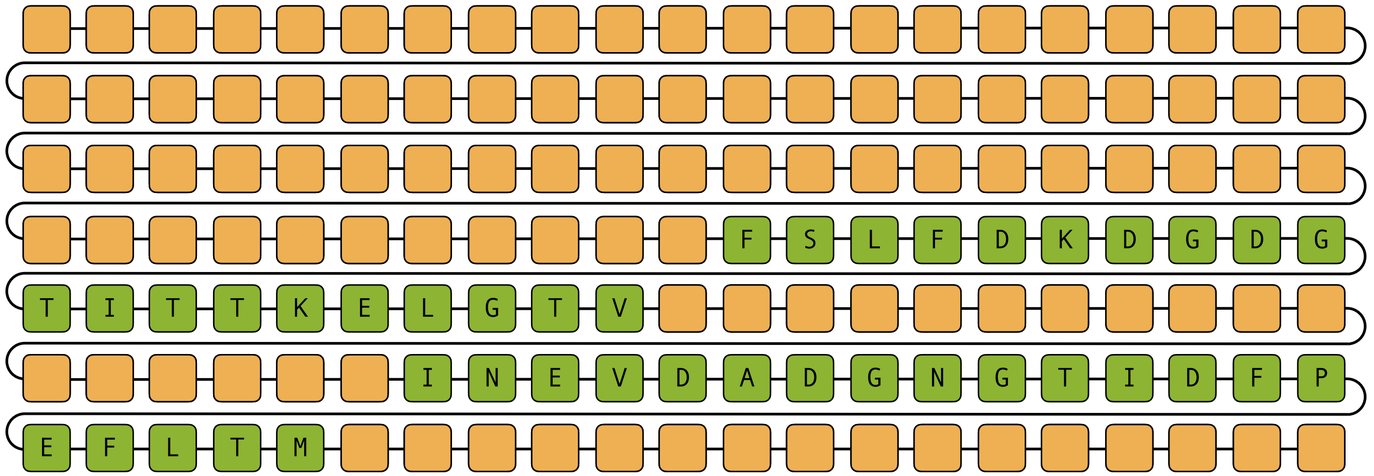
conditional generation

EvoDiff can scaffold functional motifs

1PRW: binding site of compact calmodulin



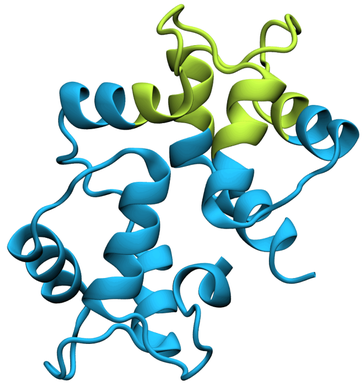
Functional motif

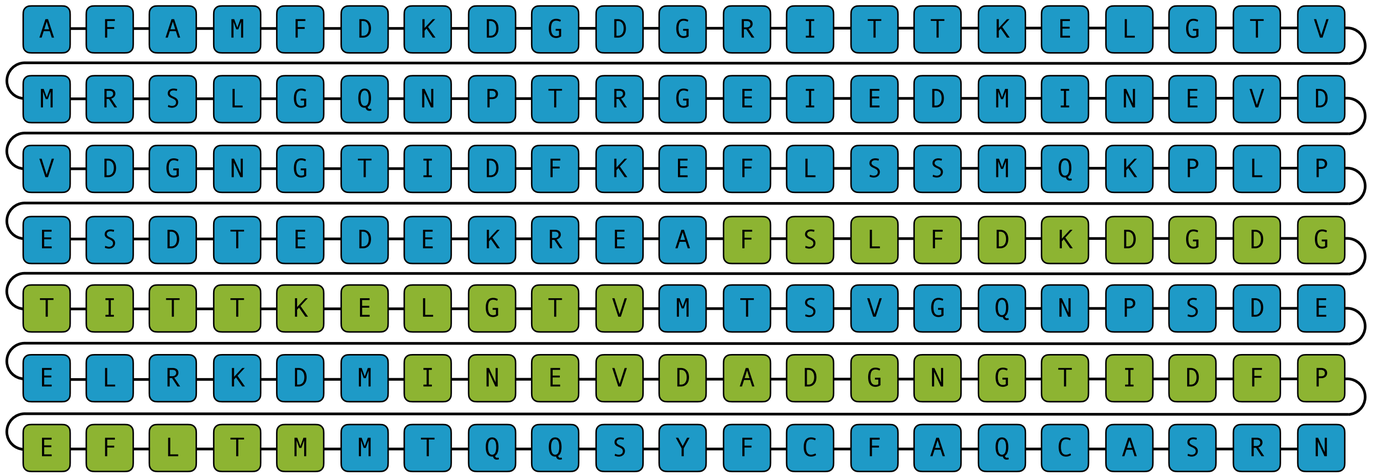


conditional generation

EvoDiff can scaffold functional motifs

1PRW: binding site of compact calmodulin





No structure needed!

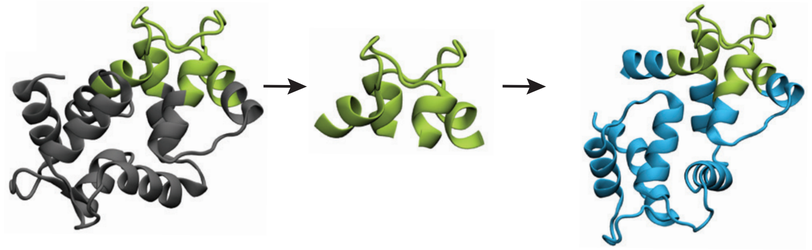
conditional generation

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

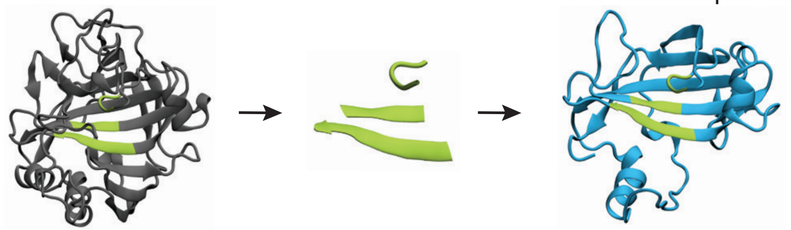
FISDVENAFSLFDKDGDGTITTKELGTVM
RSLGQNPTESELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVF
DRDNNGLISAAELRHVMTNLGEKLTDDE
VDEMIREADVDGDGQVNYEEFVTMMTA
KSLDYNDD

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

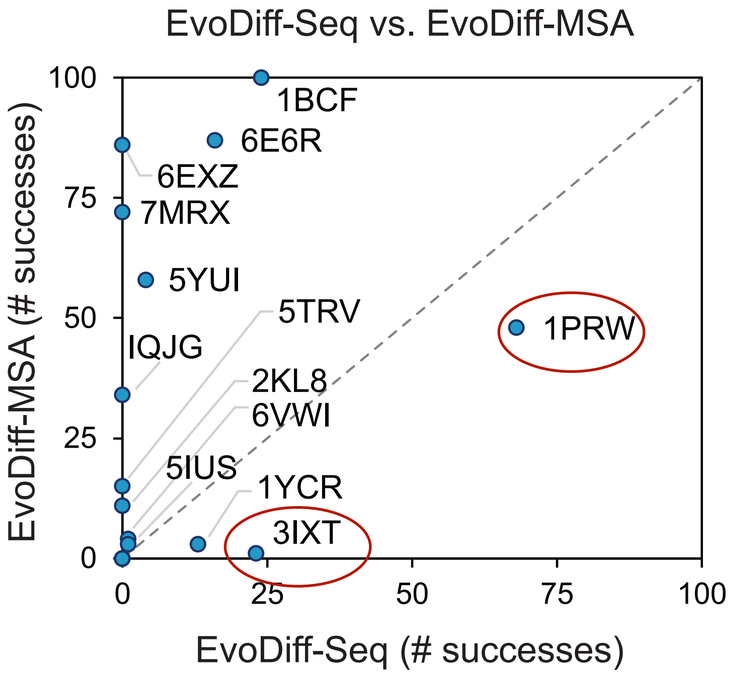
SWAGDAMLSGGGLSGDYSVAEFHFHW
GSTNTAGSEHTINNIRHAAELHLVHVSNR
FGTIEEAARVRNGVAVLGVFFEVGEINAG
LEPITDKLRHLAGRGTHEPVNPLAPHEYM
PSSDDFFTYTGSLTTPPCSTGVLWYVF
DRPTRISVHQ

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

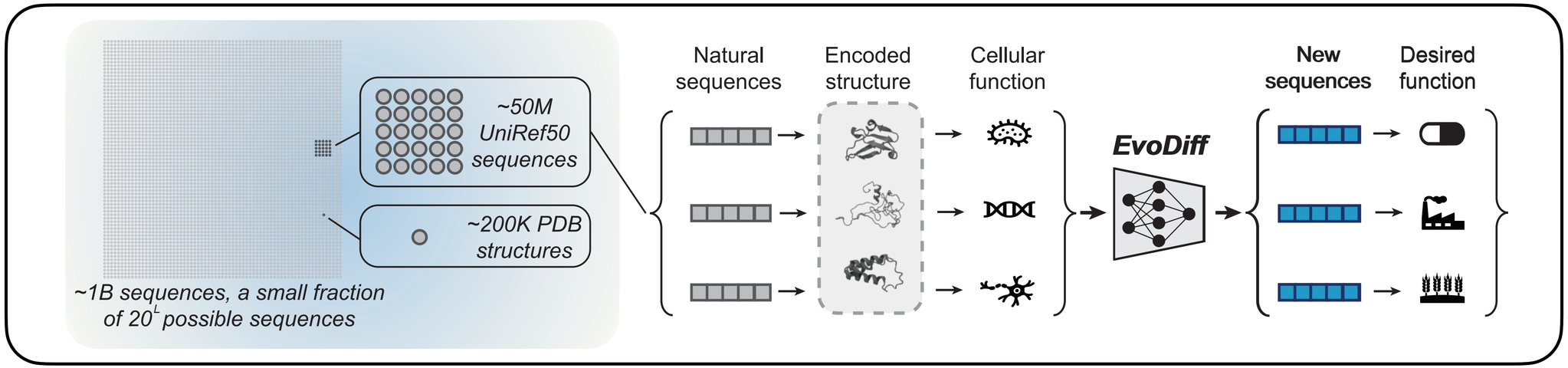
EvoDiff-Seq best

Not many homologs

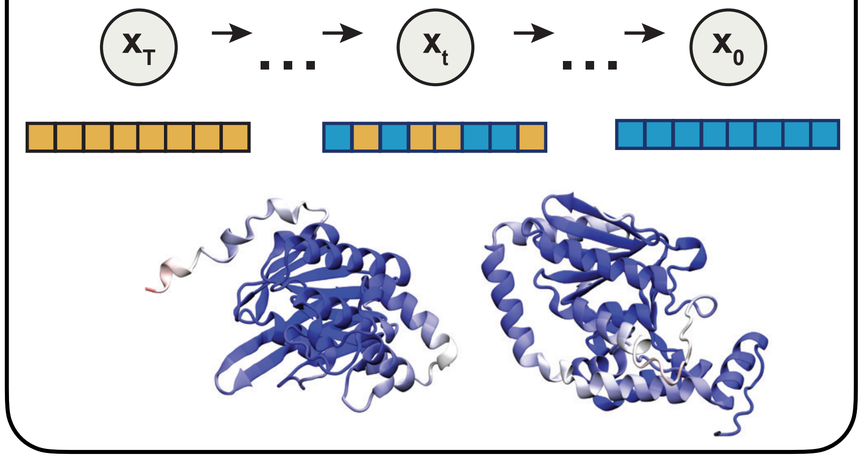
Seq usually more novel

conditional generation

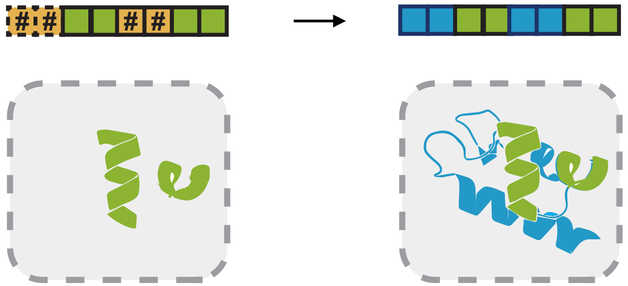
EvoDiff: controllable protein sequence diffusion



Unconditional generation

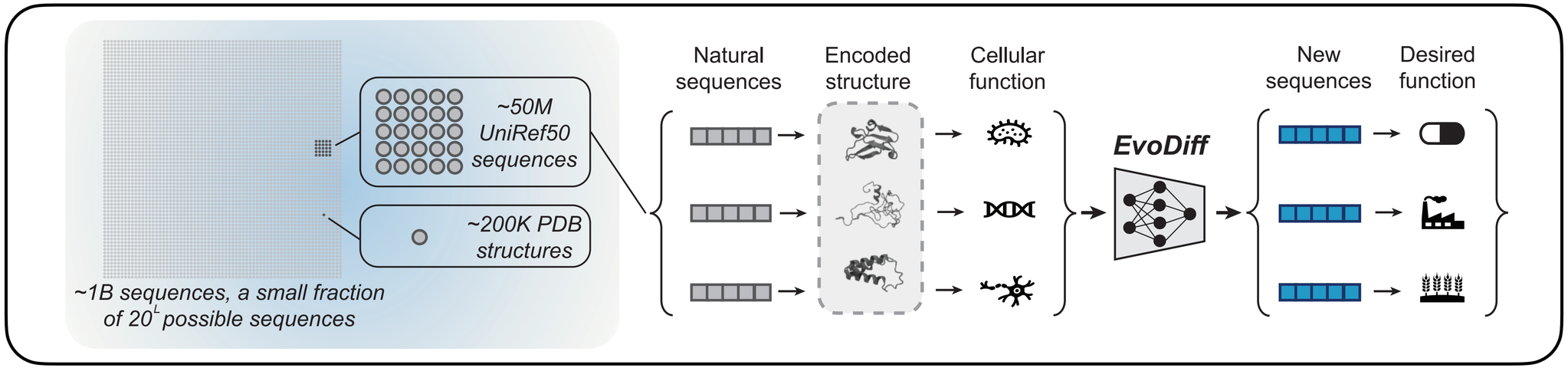


Sequence motif scaffolding

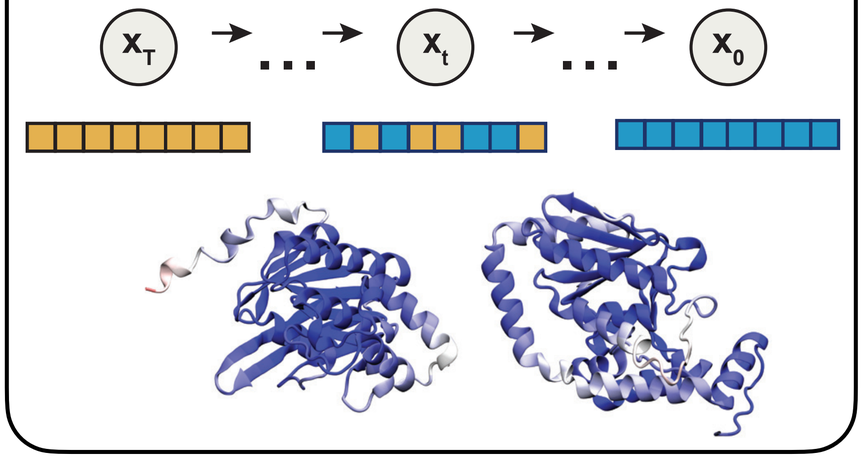


Controllable generation

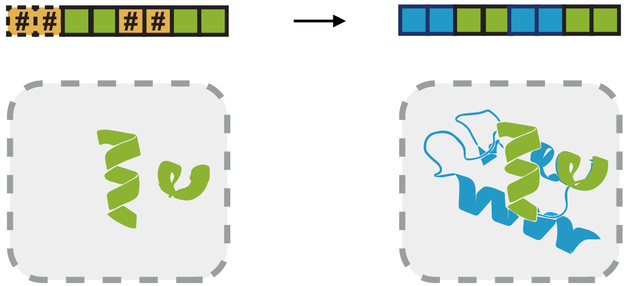
EvoDiff: controllable protein sequence diffusion



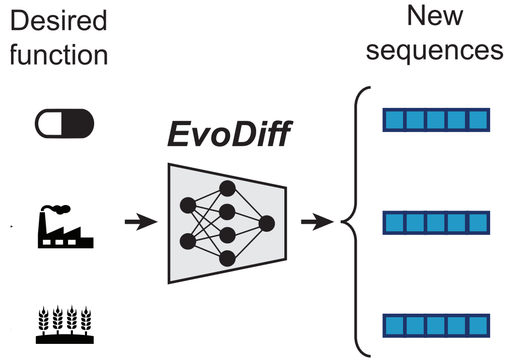
Unconditional generation



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