

Multimodal deep learning for protein engineering

Kevin Kaichuang Yang
Microsoft Research New England
 @KevinKaichuang

Protein engineering requires going from function to sequence

Protein engineering requires going from function to sequence

MGTGDHDD...

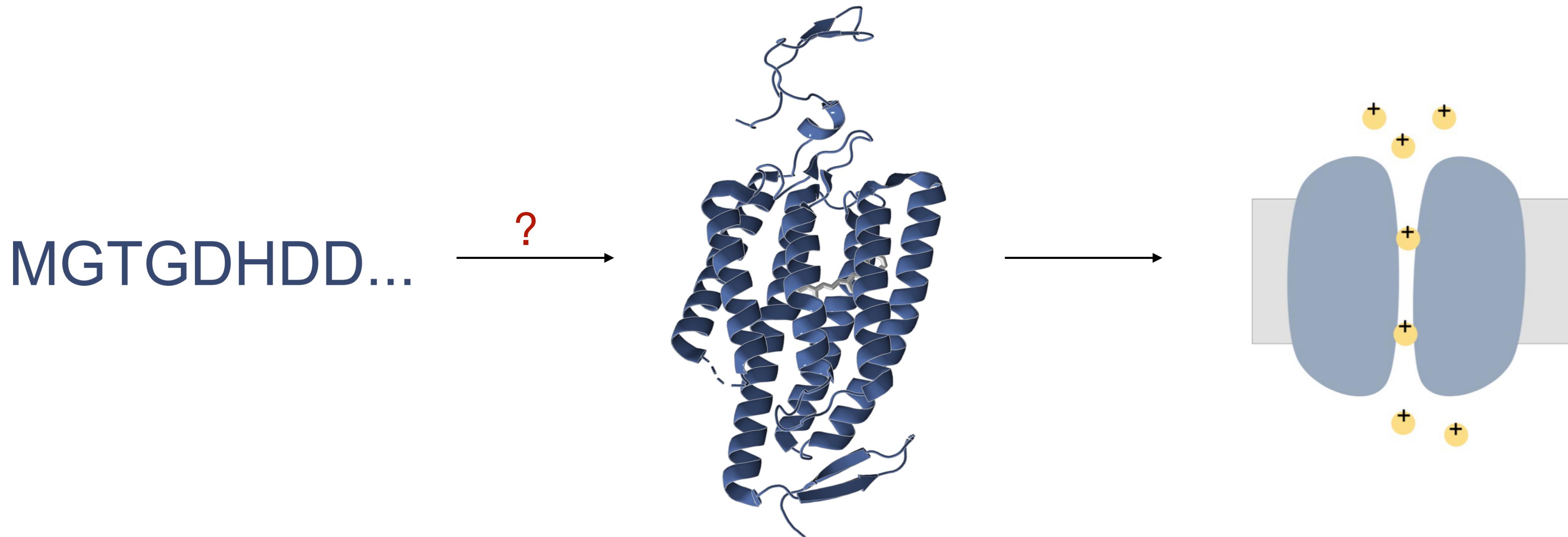
Protein engineering requires going from function to sequence



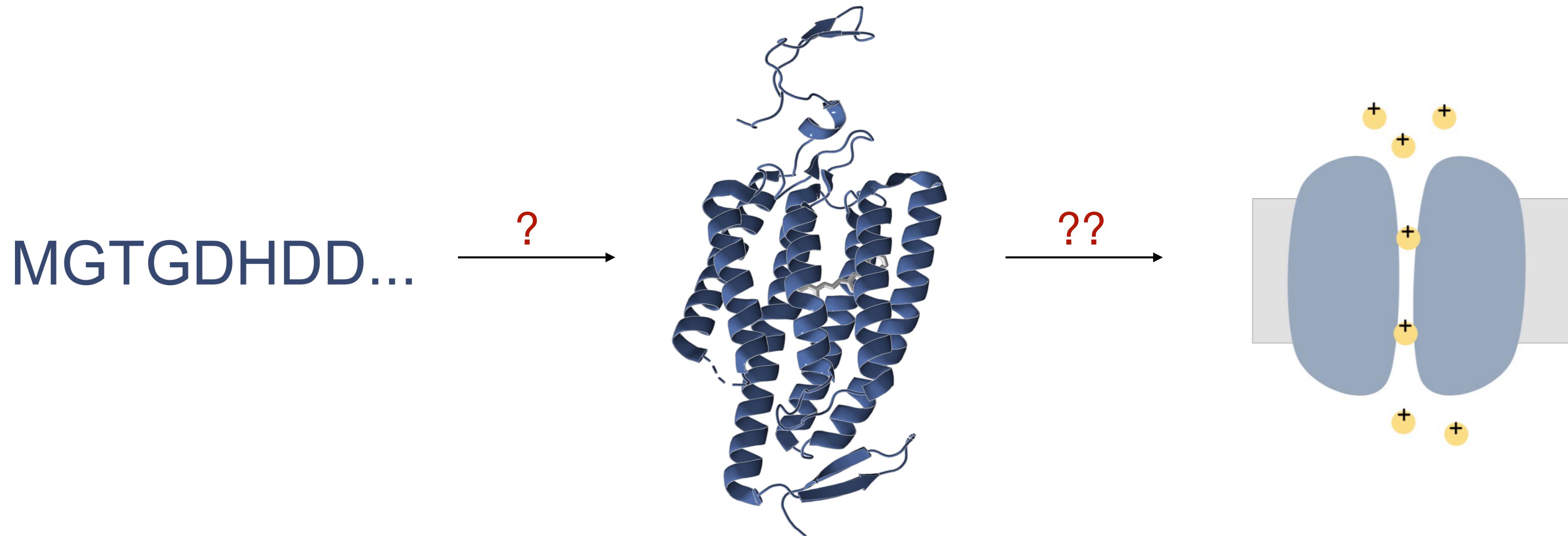
Protein engineering requires going from function to sequence



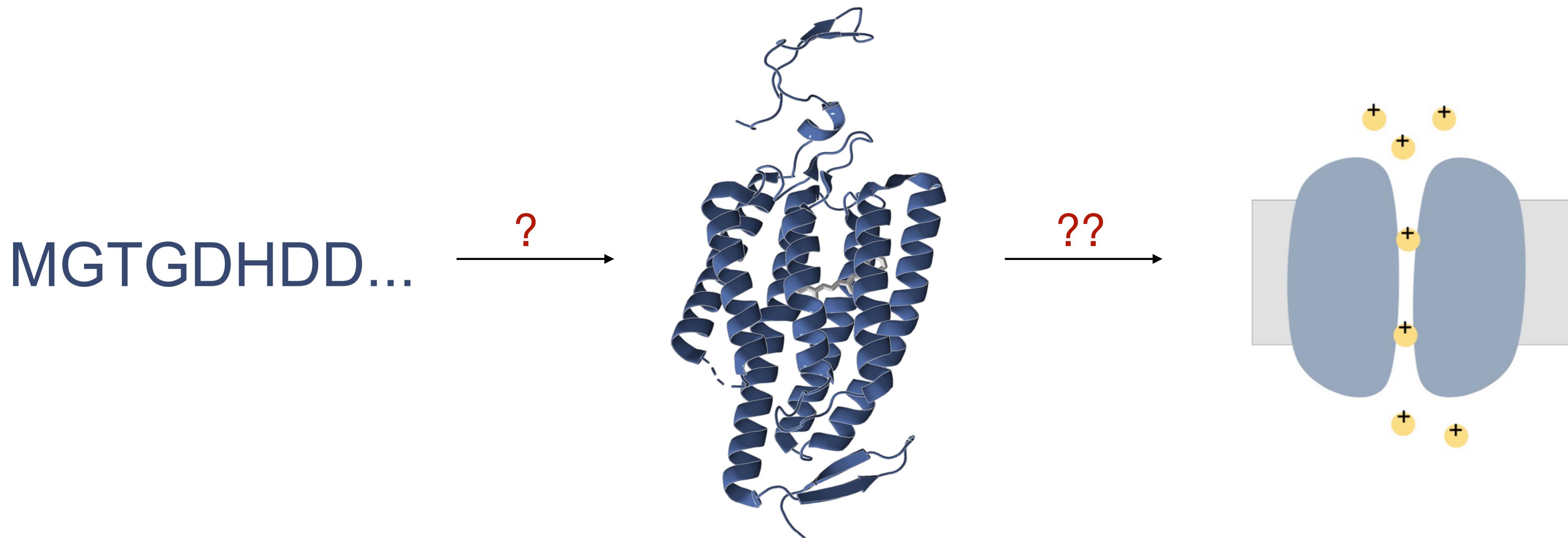
Protein engineering requires going from function to sequence



Protein engineering requires going from function to sequence



Protein engineering requires going from function to sequence



What sequence will give the desired function?

Protein engineering requires going from function to sequence

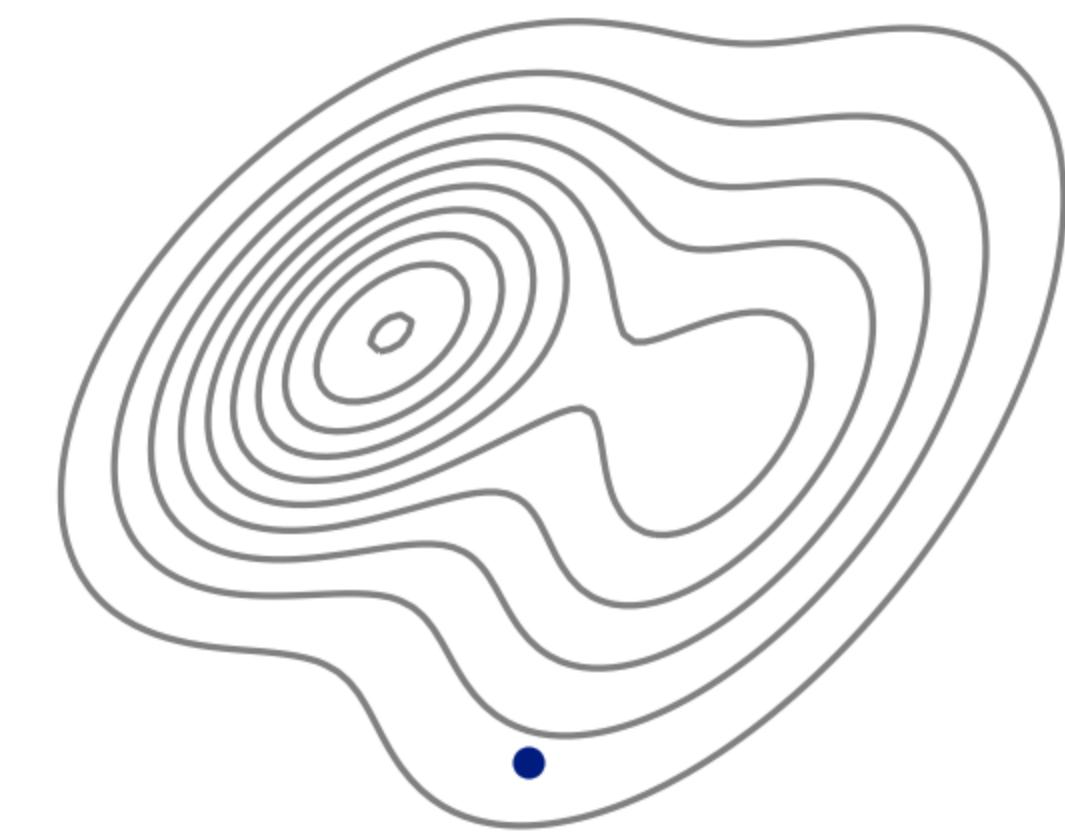
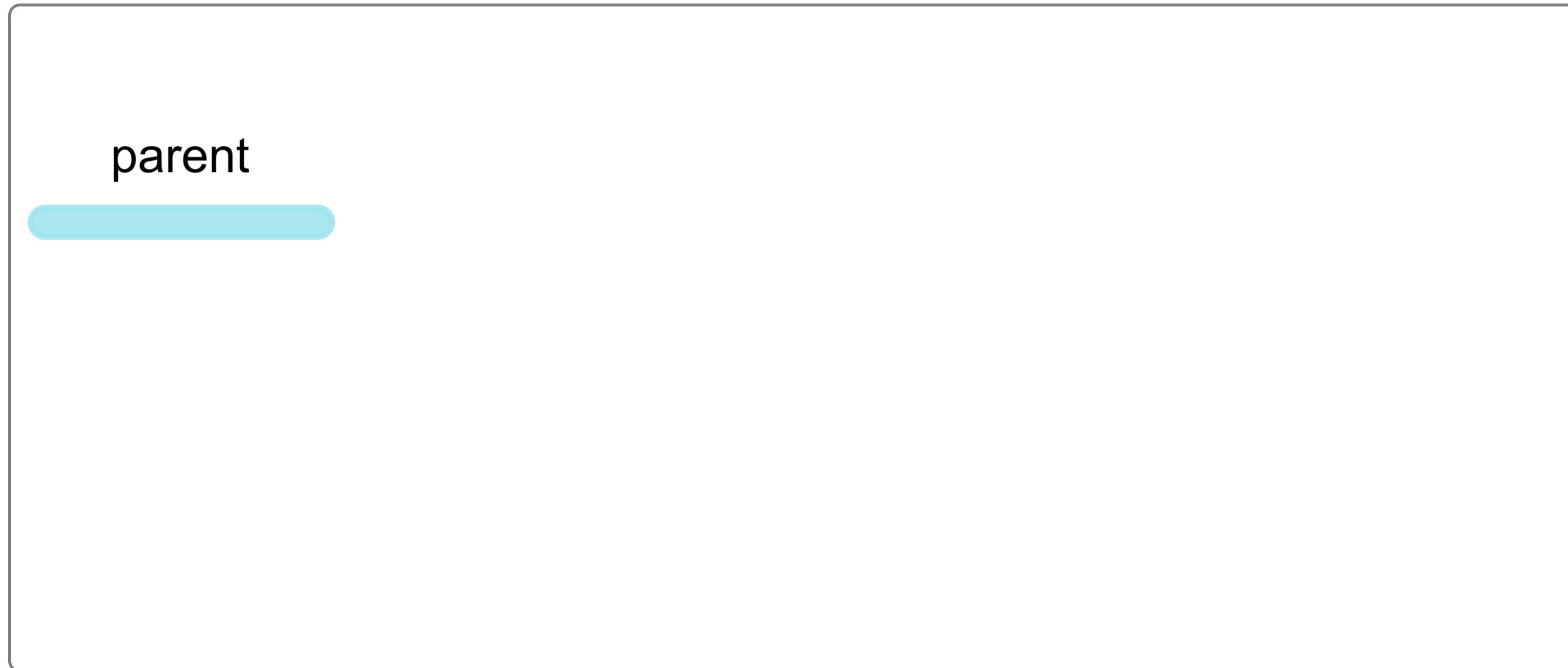


What sequence will give the desired function?

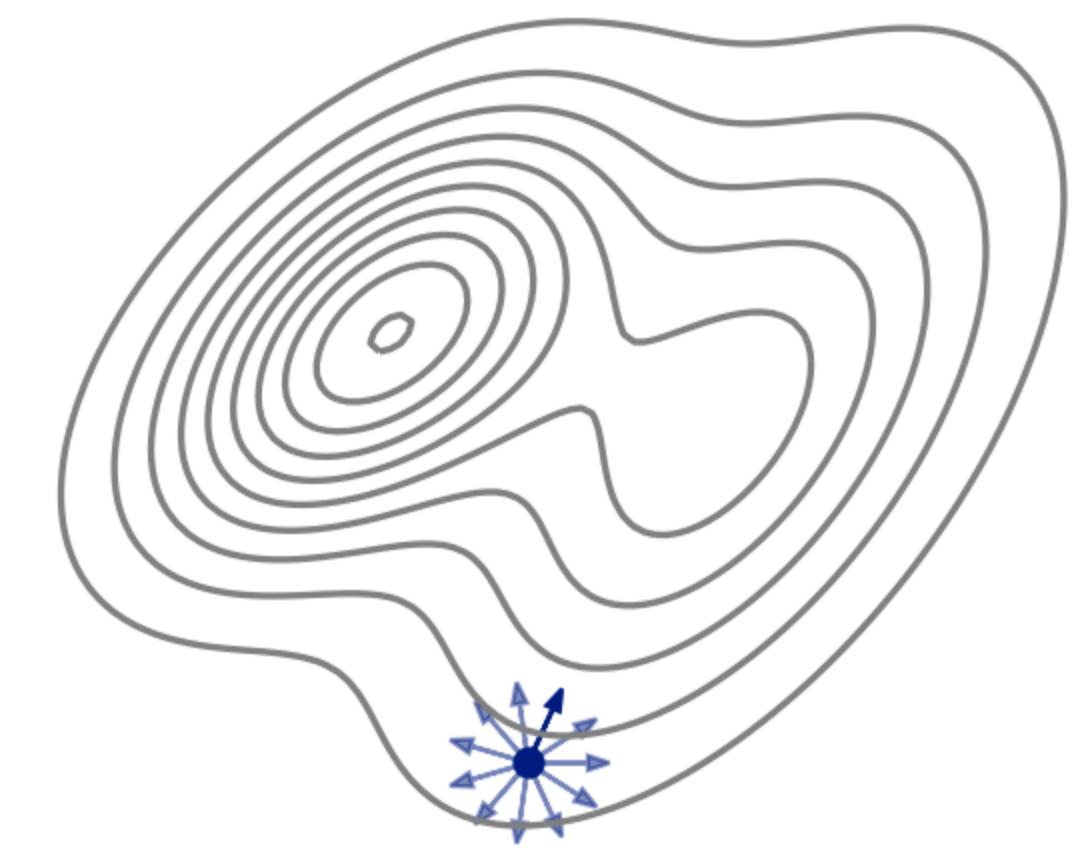
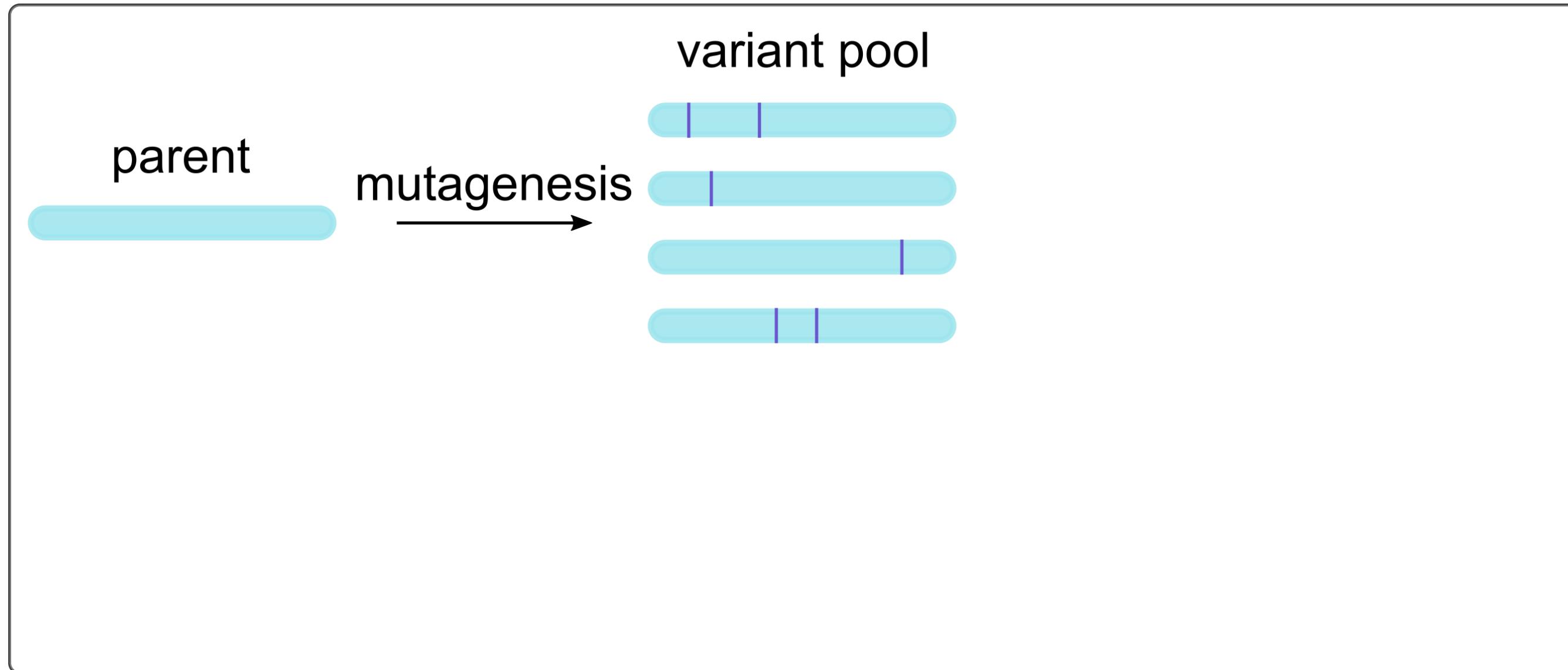
Directed evolution sidesteps the problem



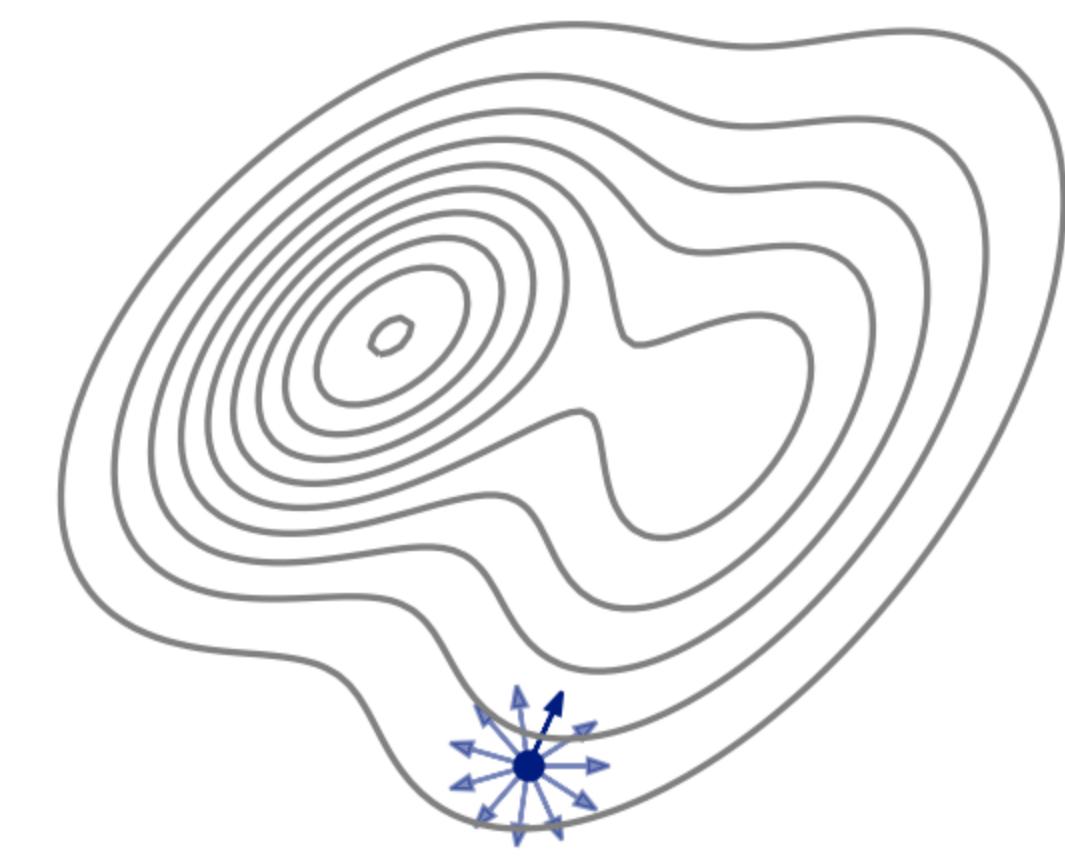
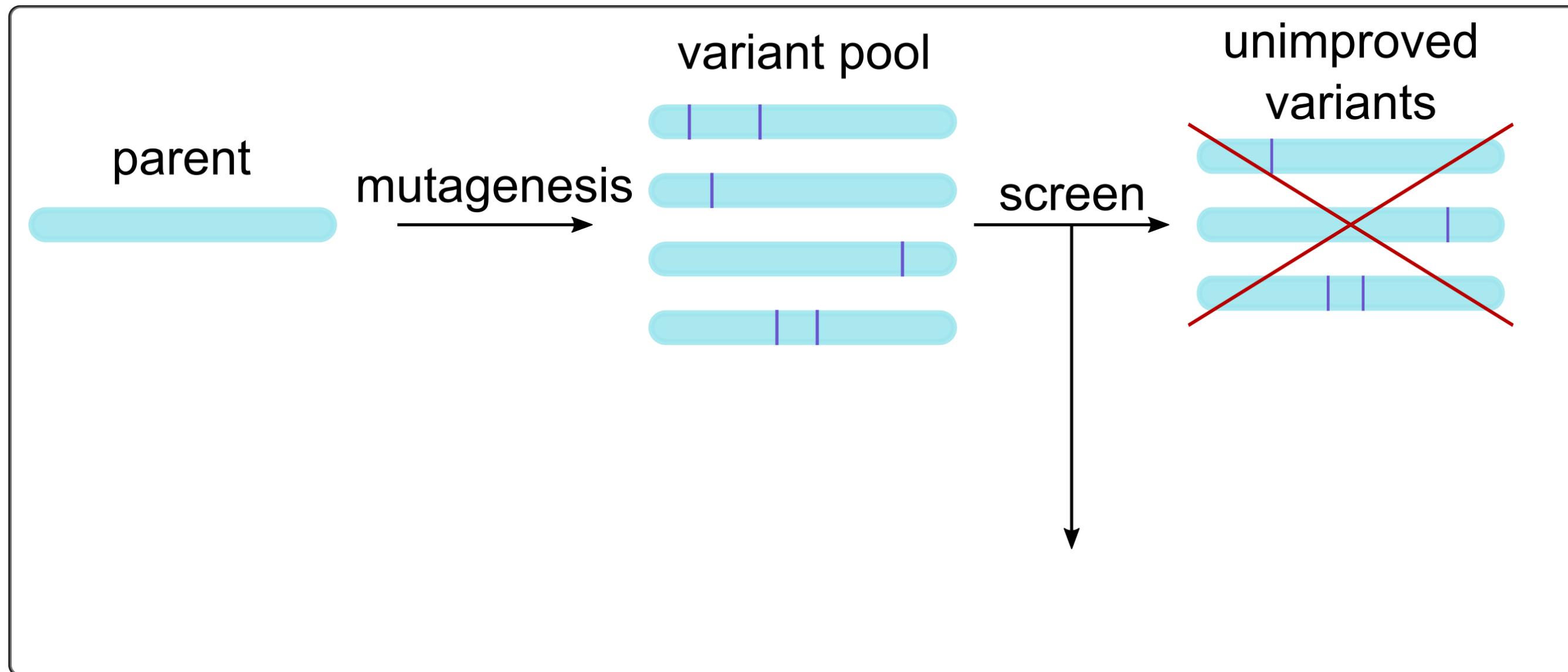
Directed evolution sidesteps the problem



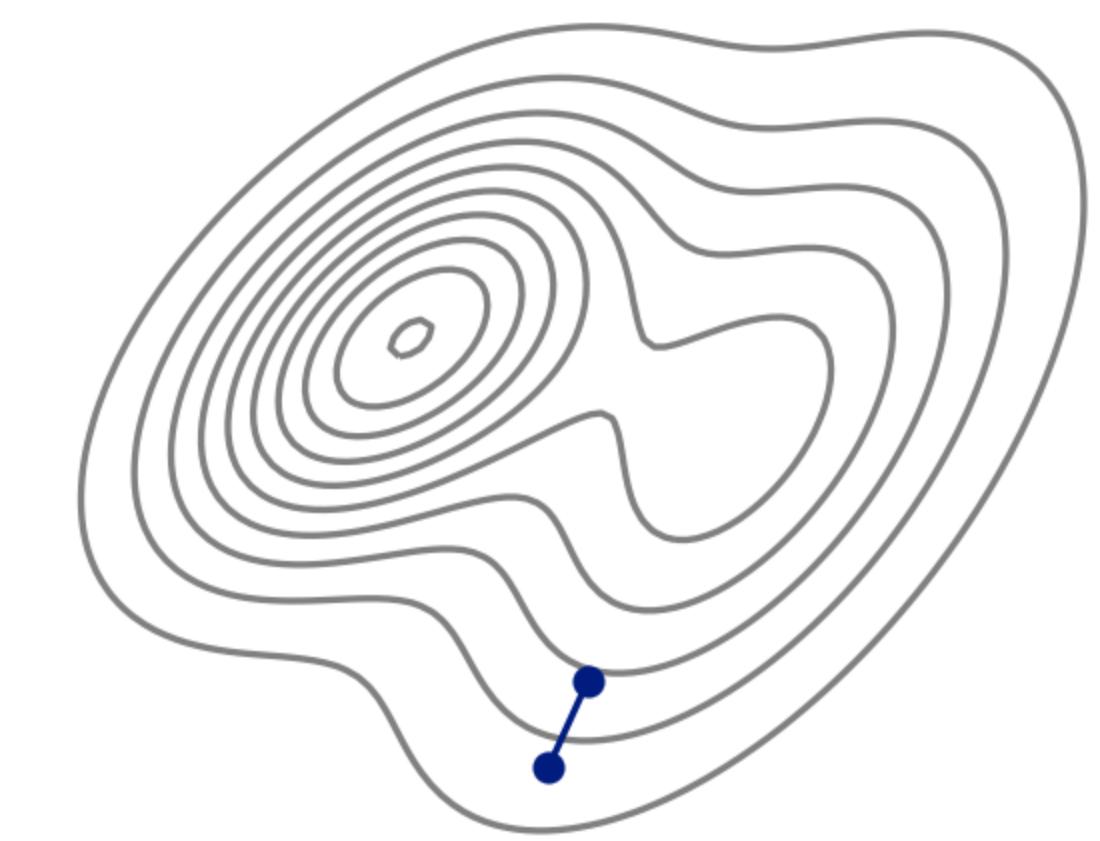
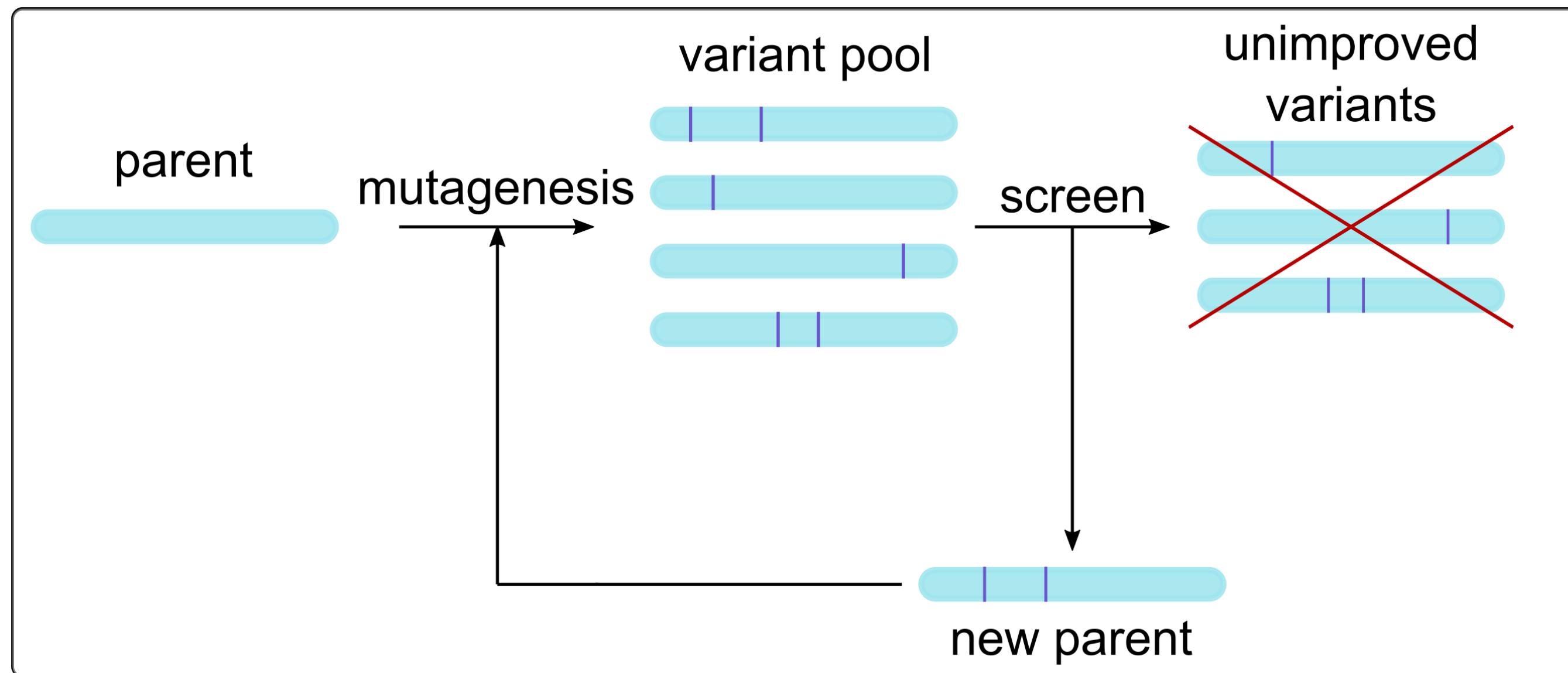
Directed evolution sidesteps the problem



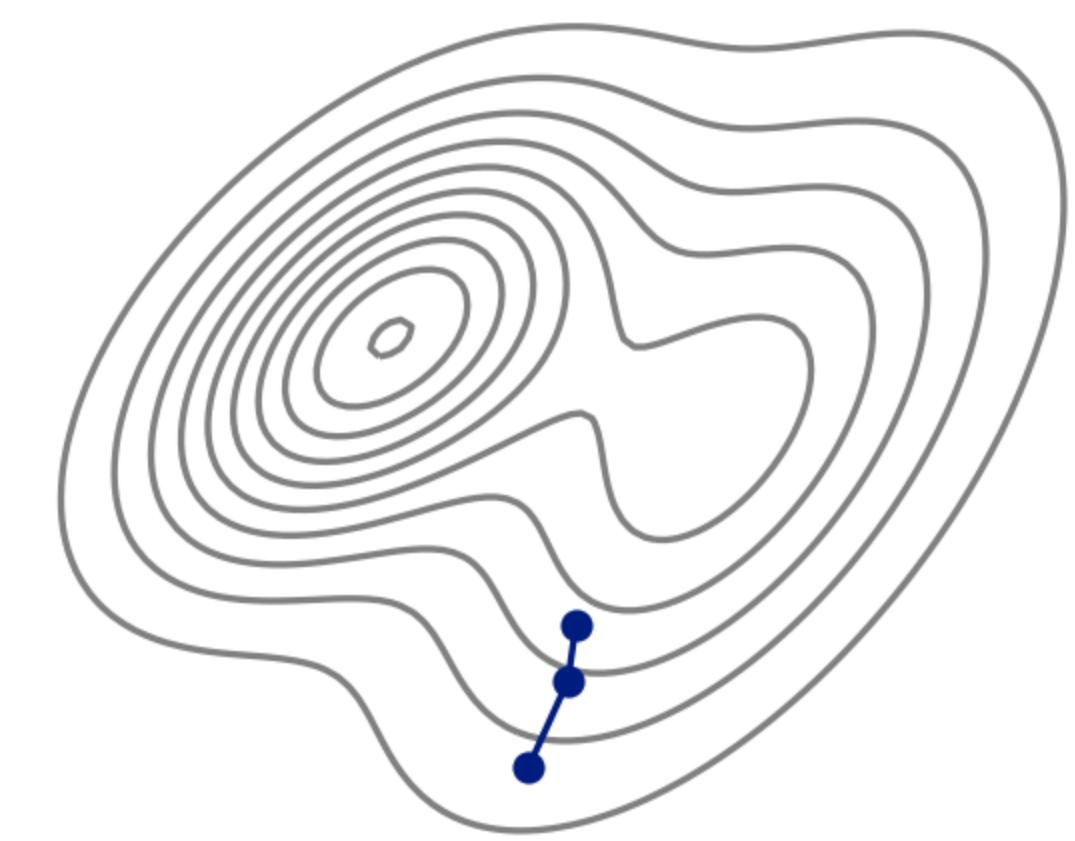
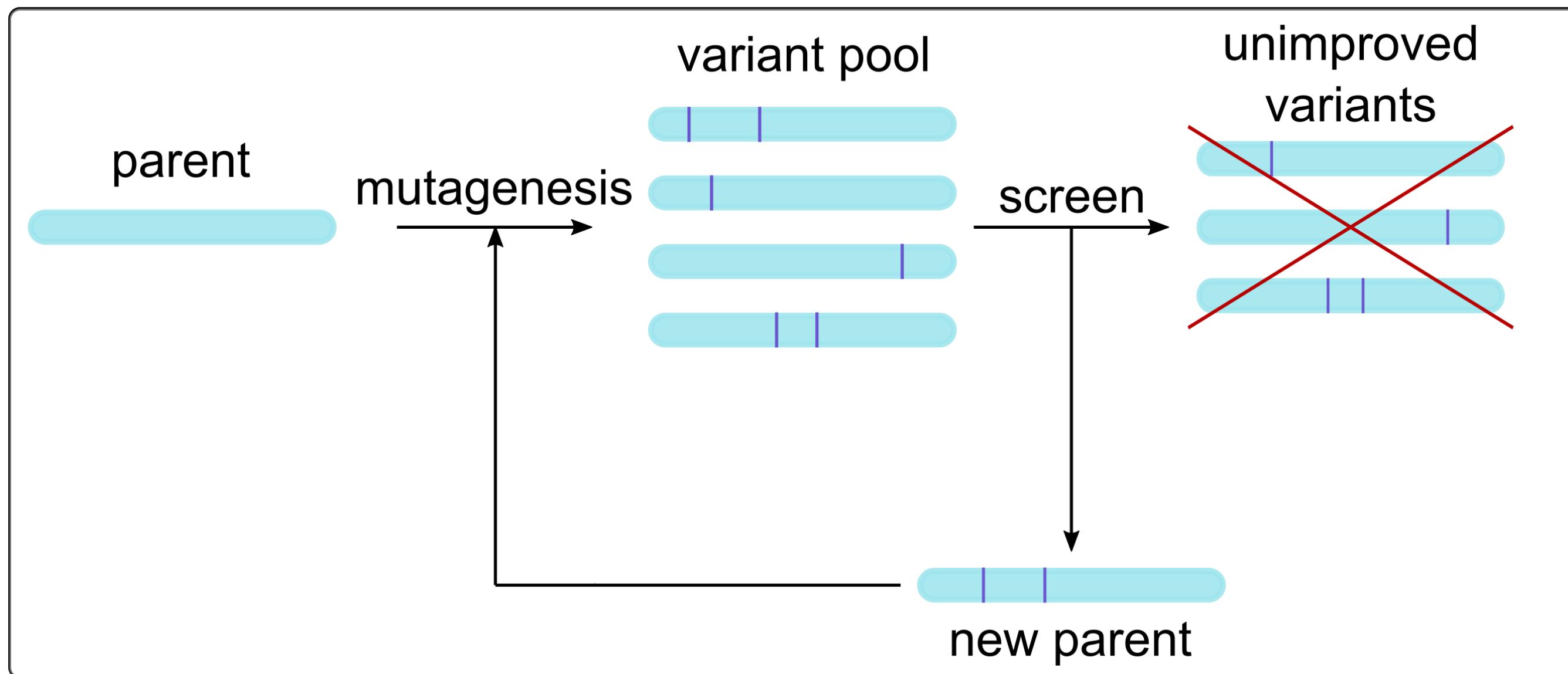
Directed evolution sidesteps the problem



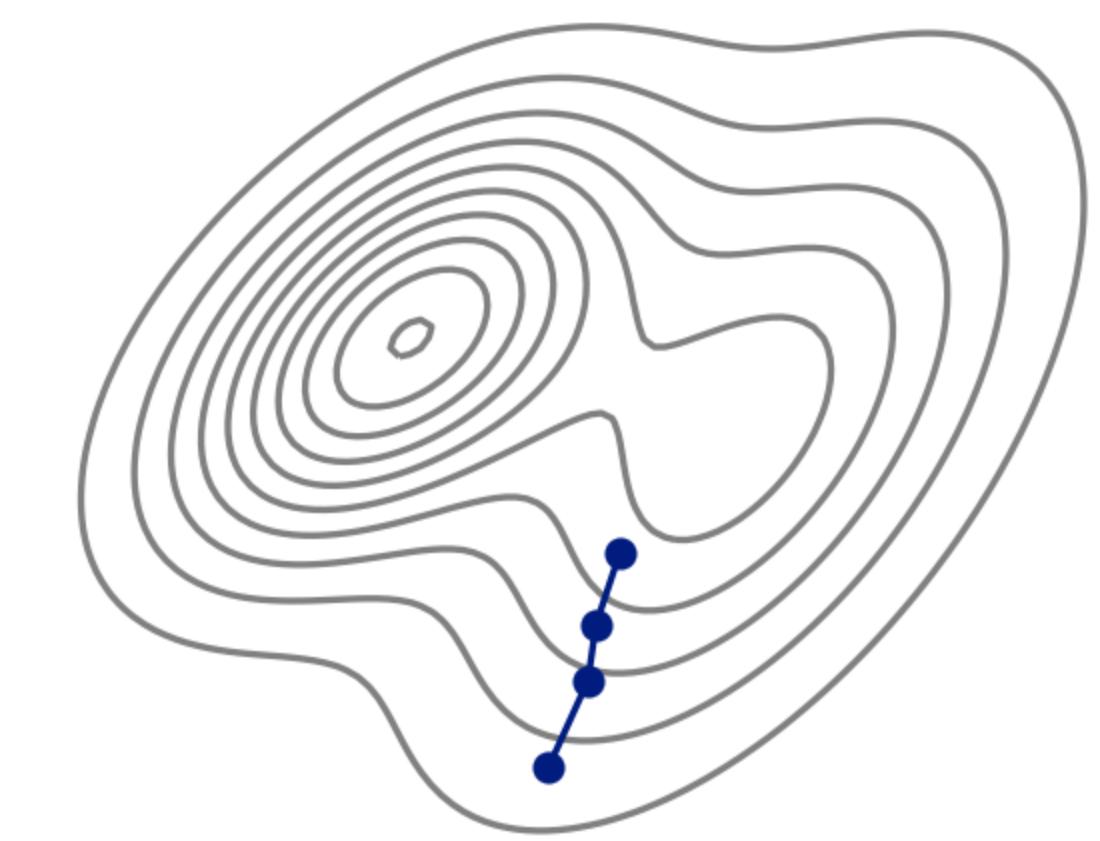
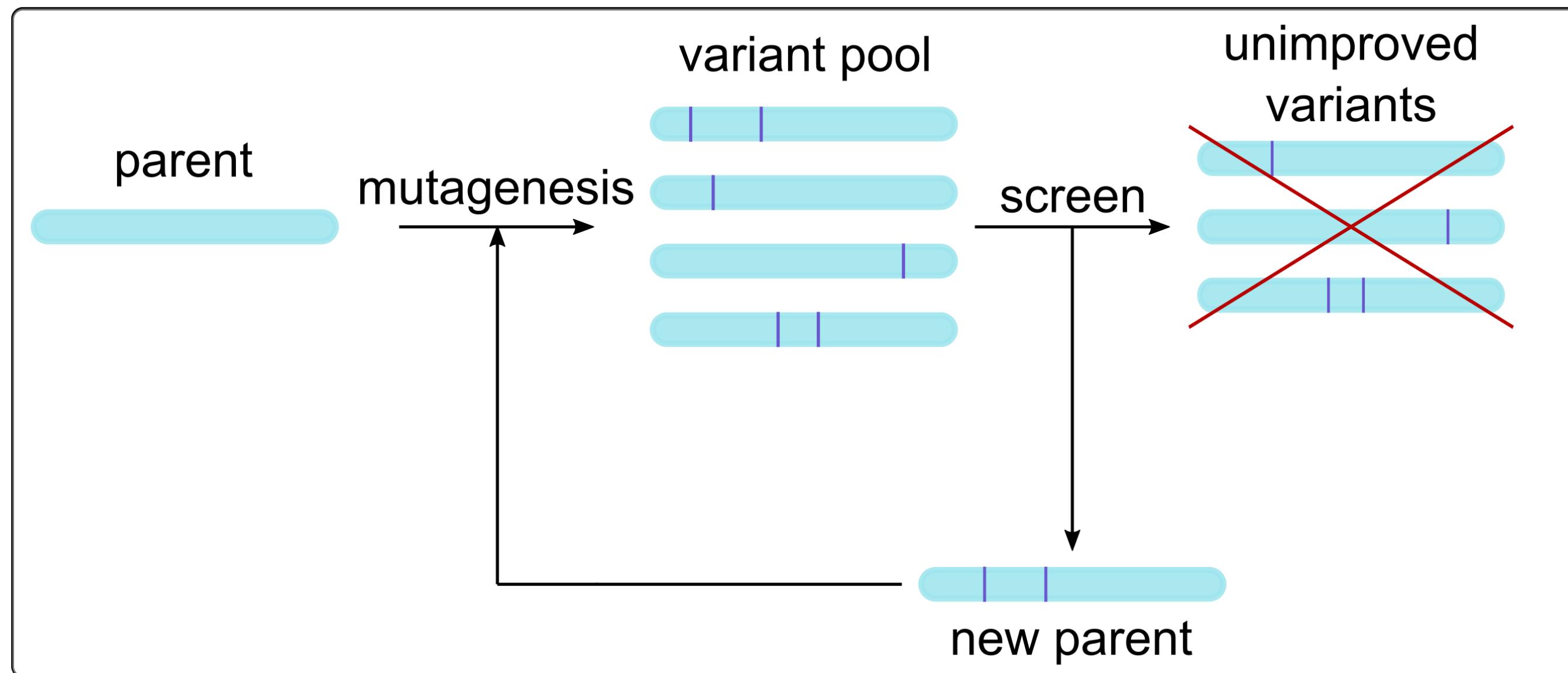
Directed evolution sidesteps the problem



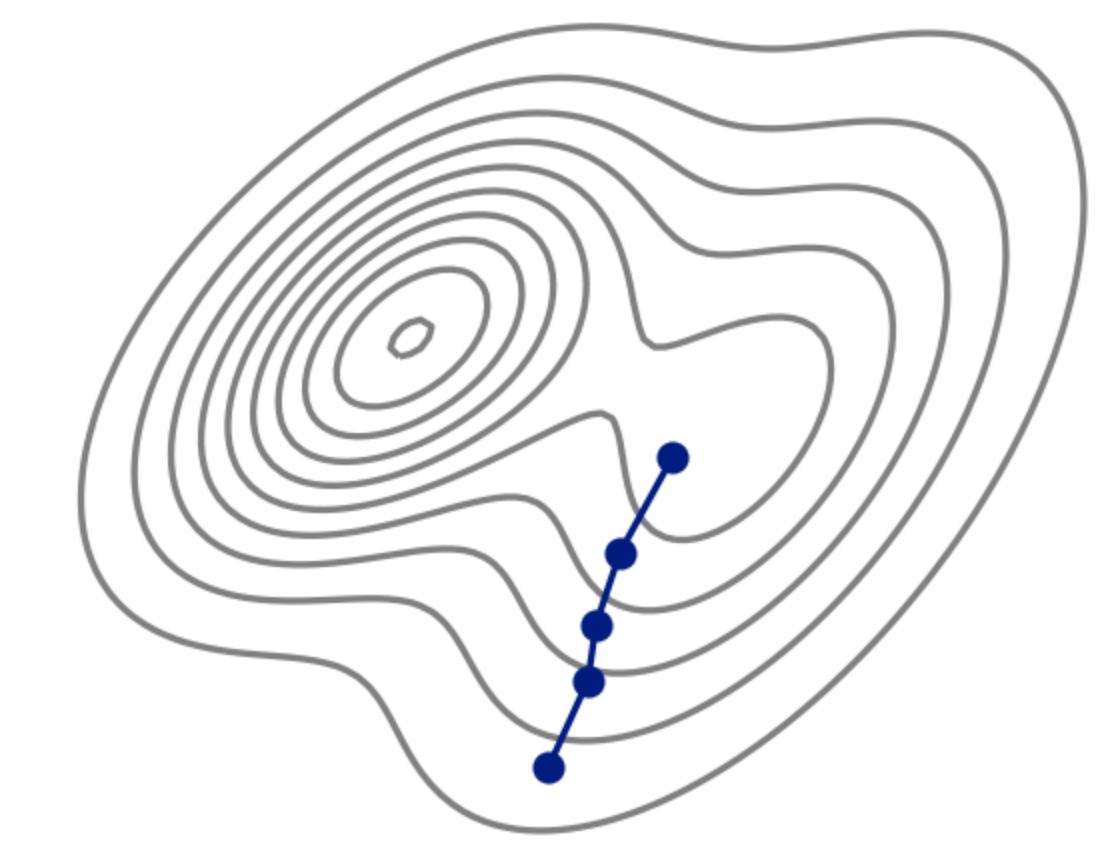
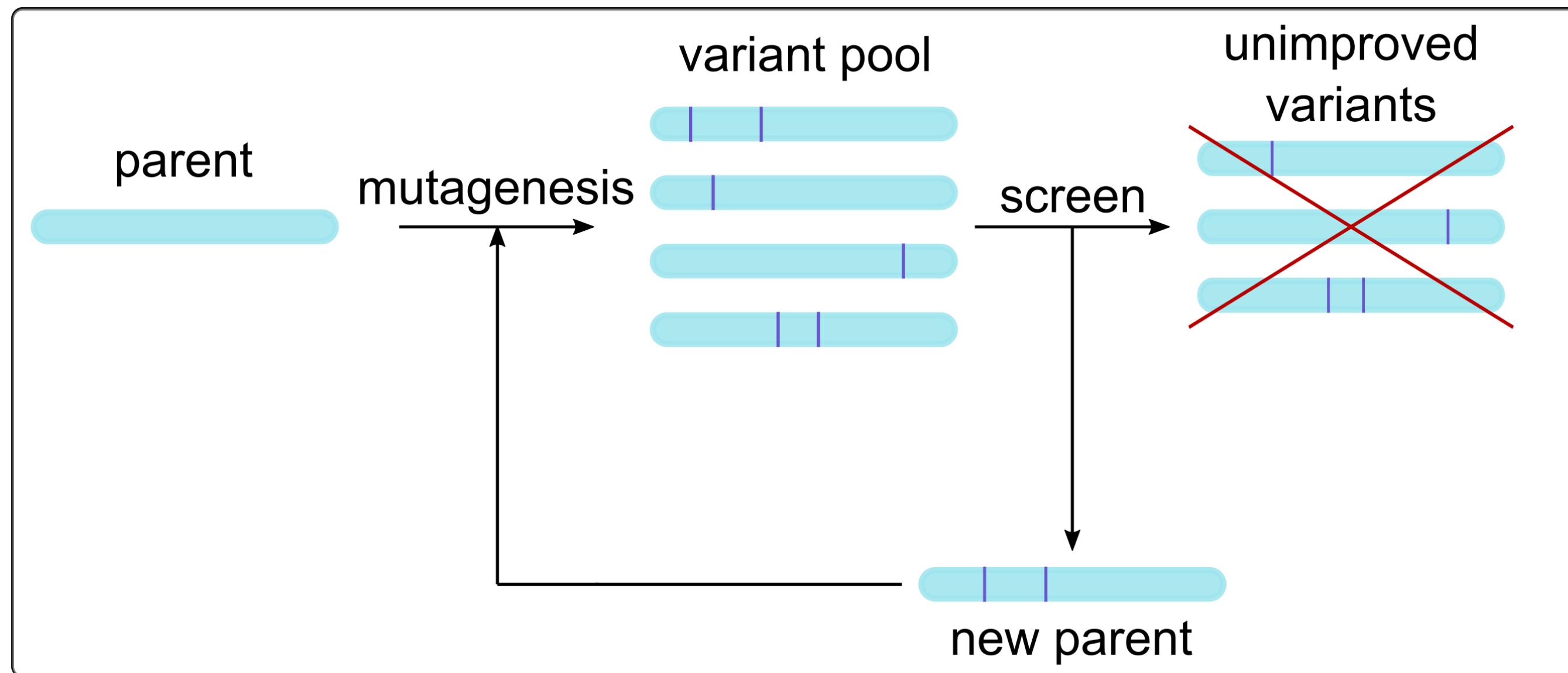
Directed evolution sidesteps the problem



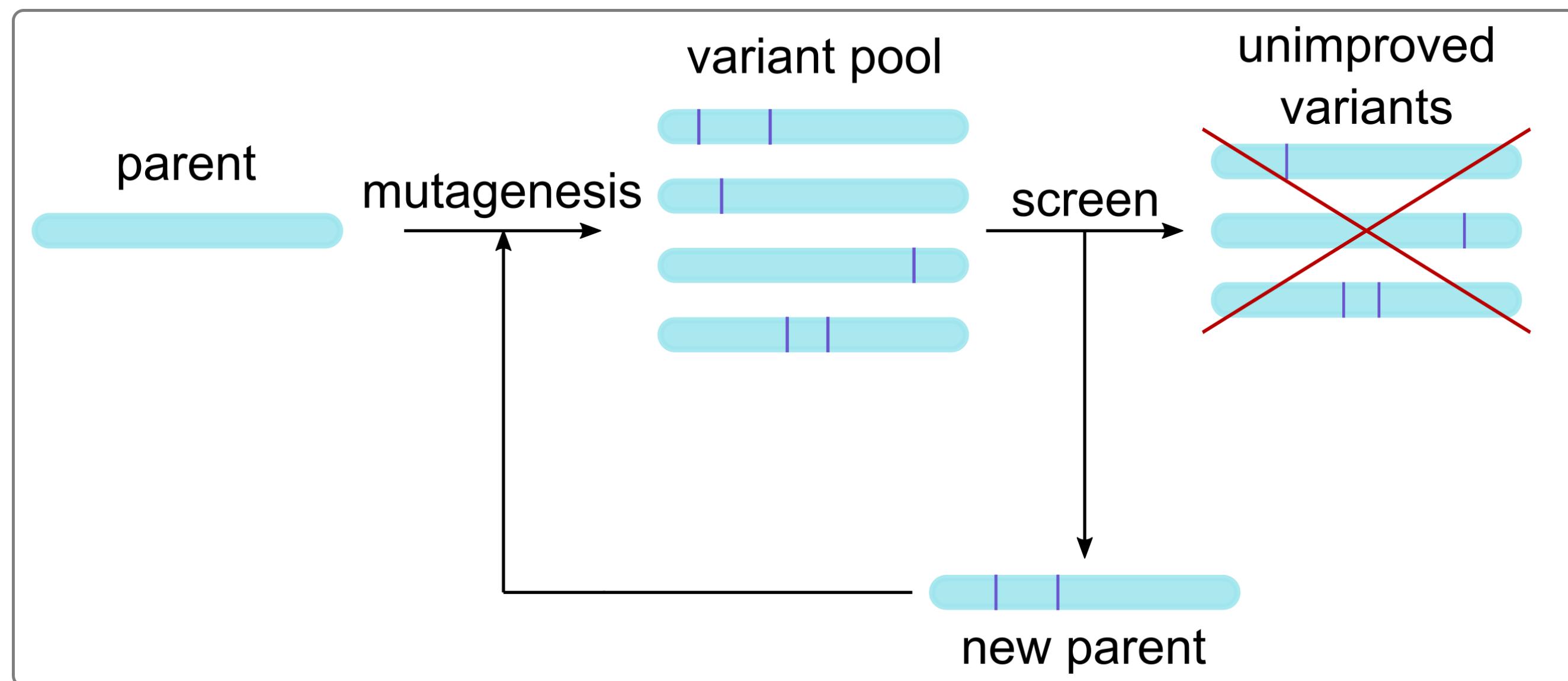
Directed evolution sidesteps the problem



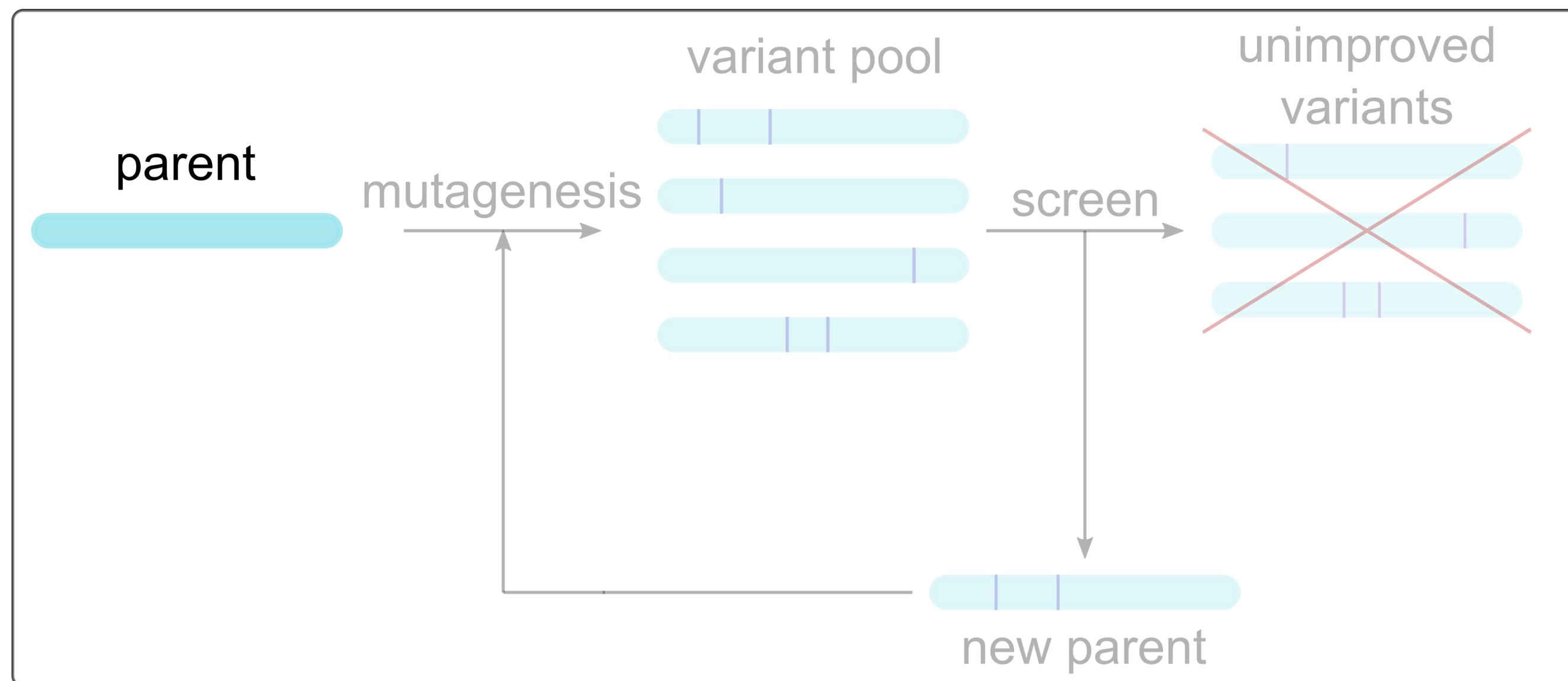
Directed evolution sidesteps the problem



Requirements for directed evolution

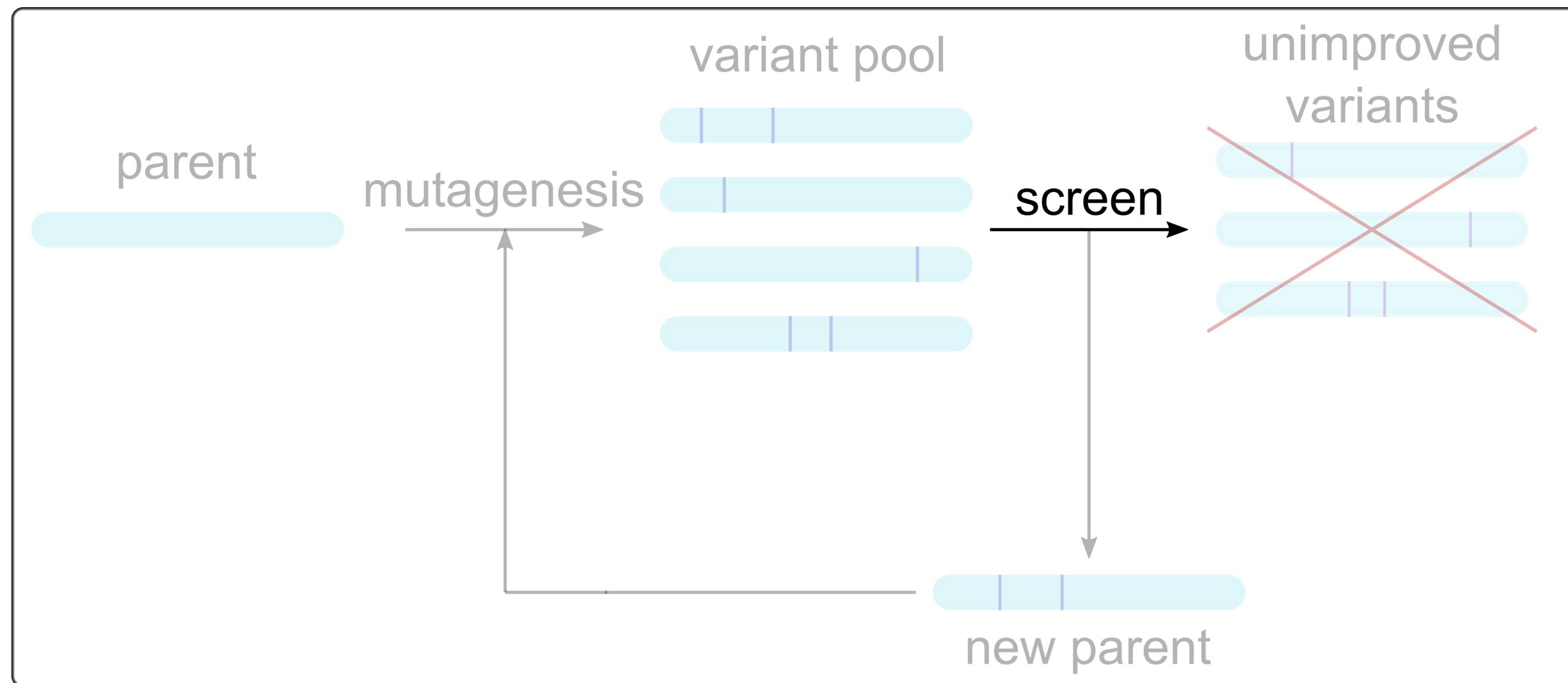


Requirements for directed evolution



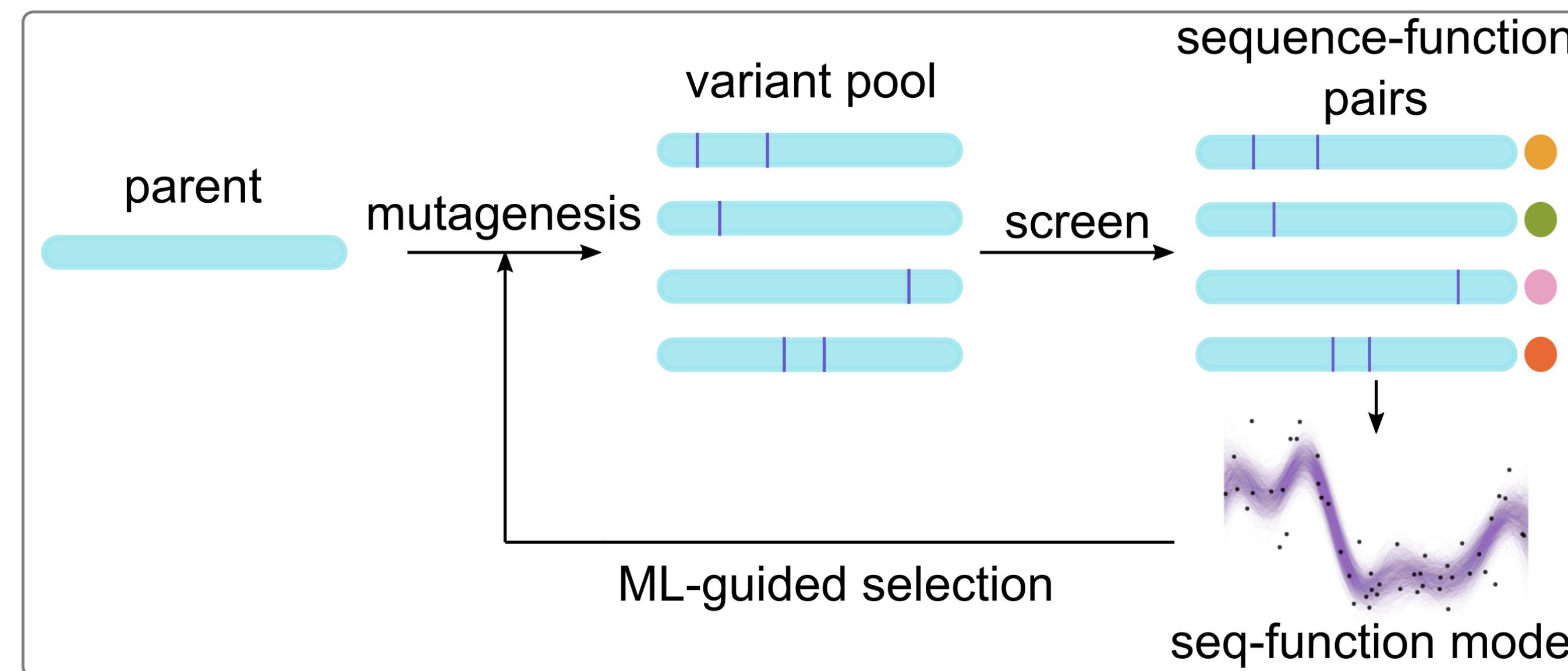
- Parent

Requirements for directed evolution



- Parent
- High-throughput screen (>100 / wk)

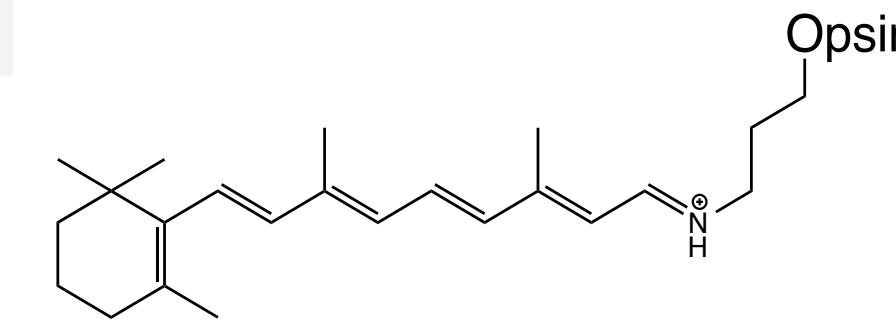
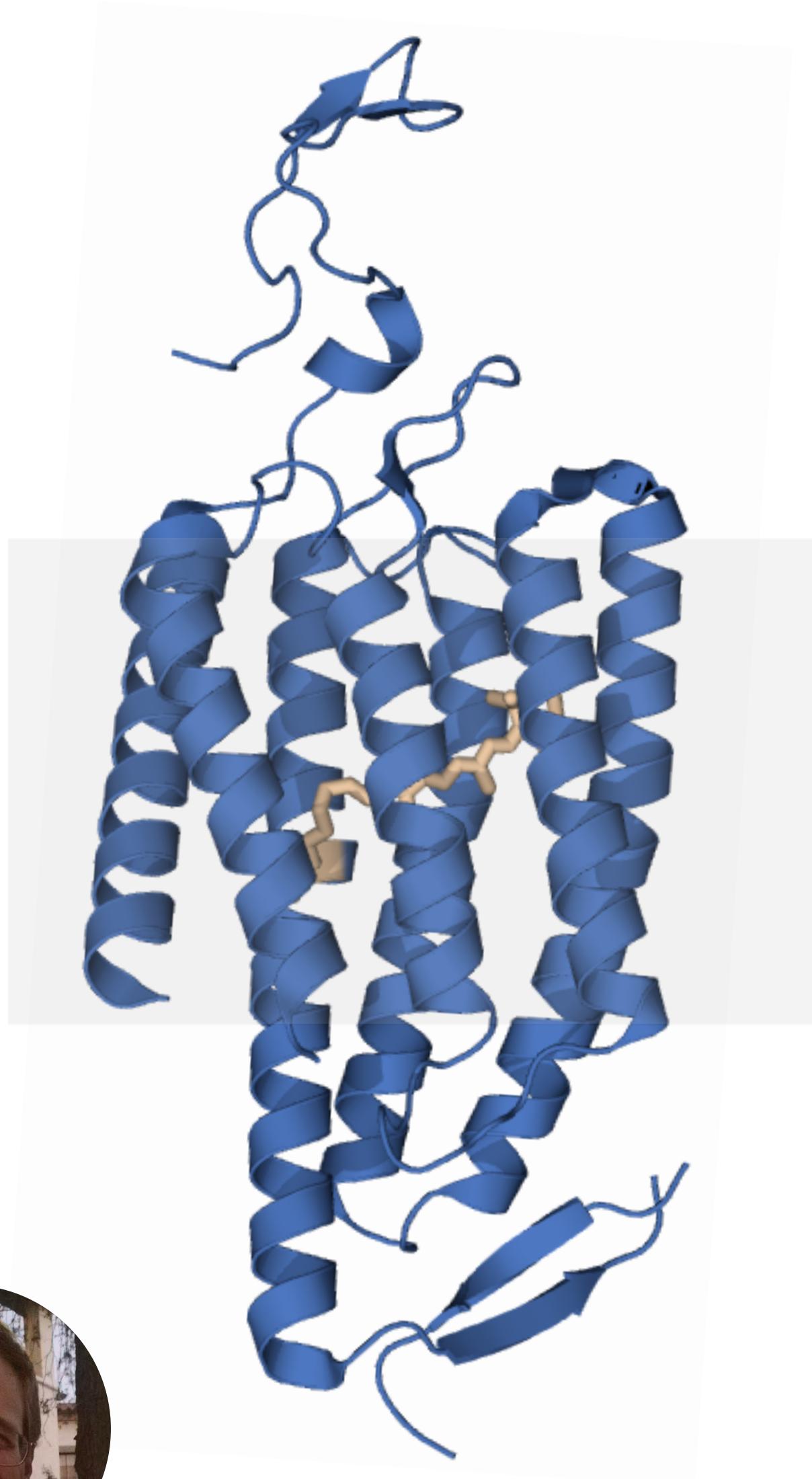
Machine learning enables optimization with fewer measurements



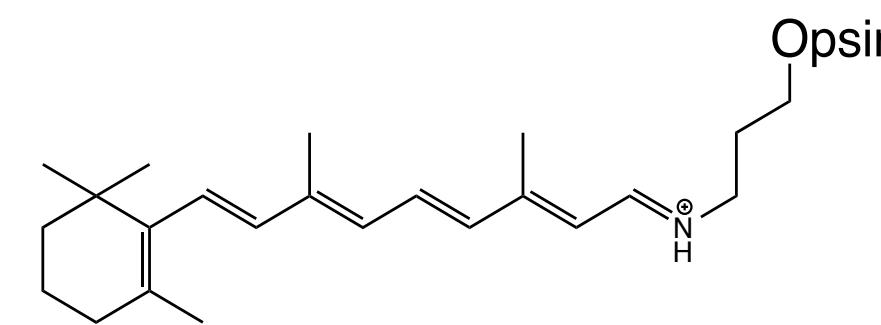
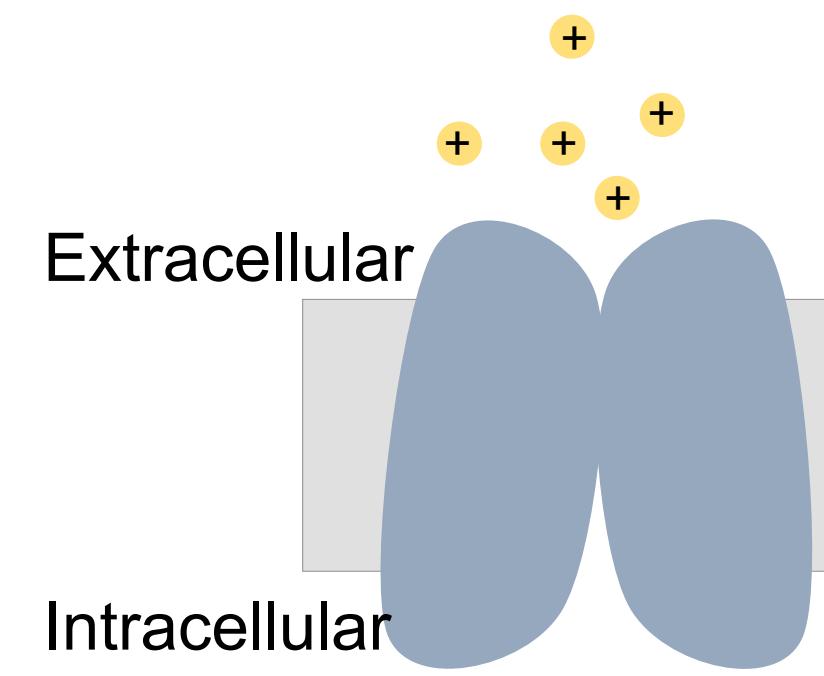
Channelrhodopsins are light-gated ion channels



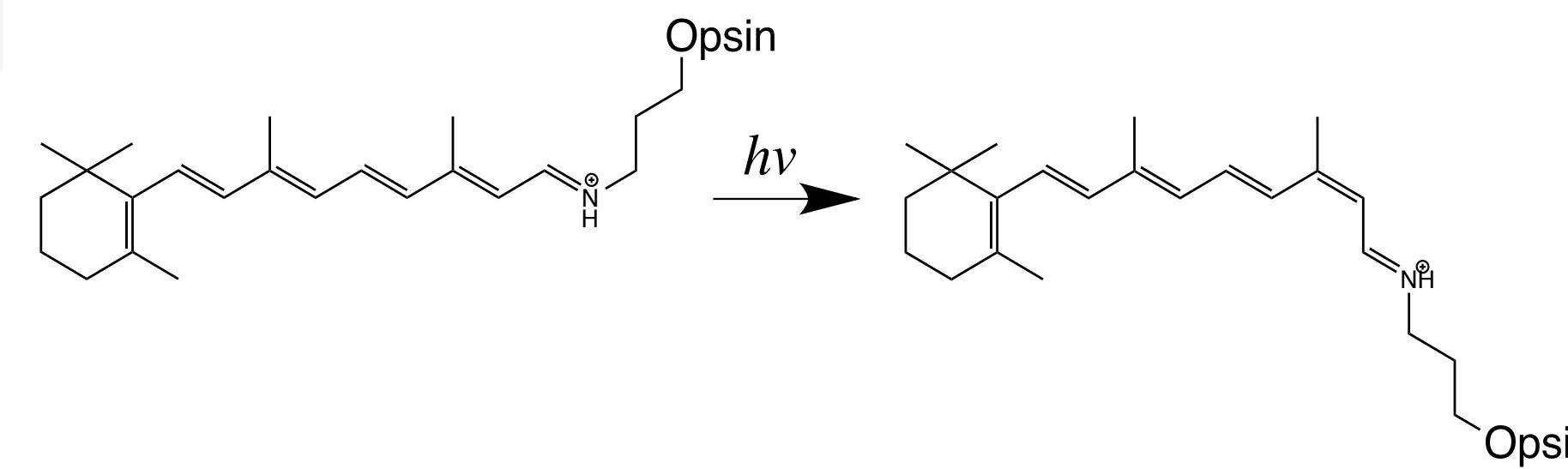
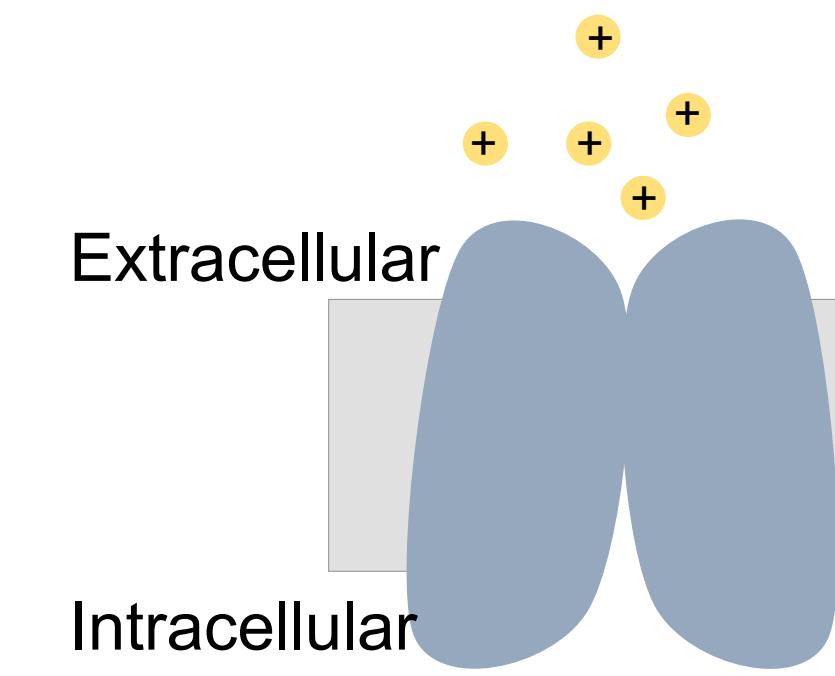
Channelrhodopsins are light-gated ion channels



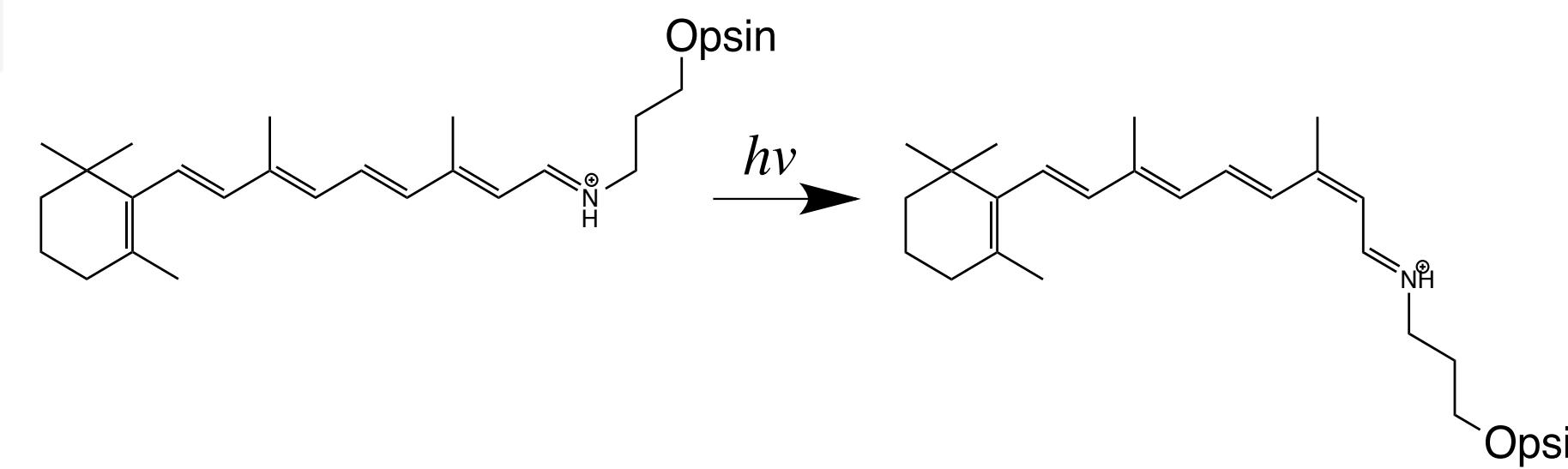
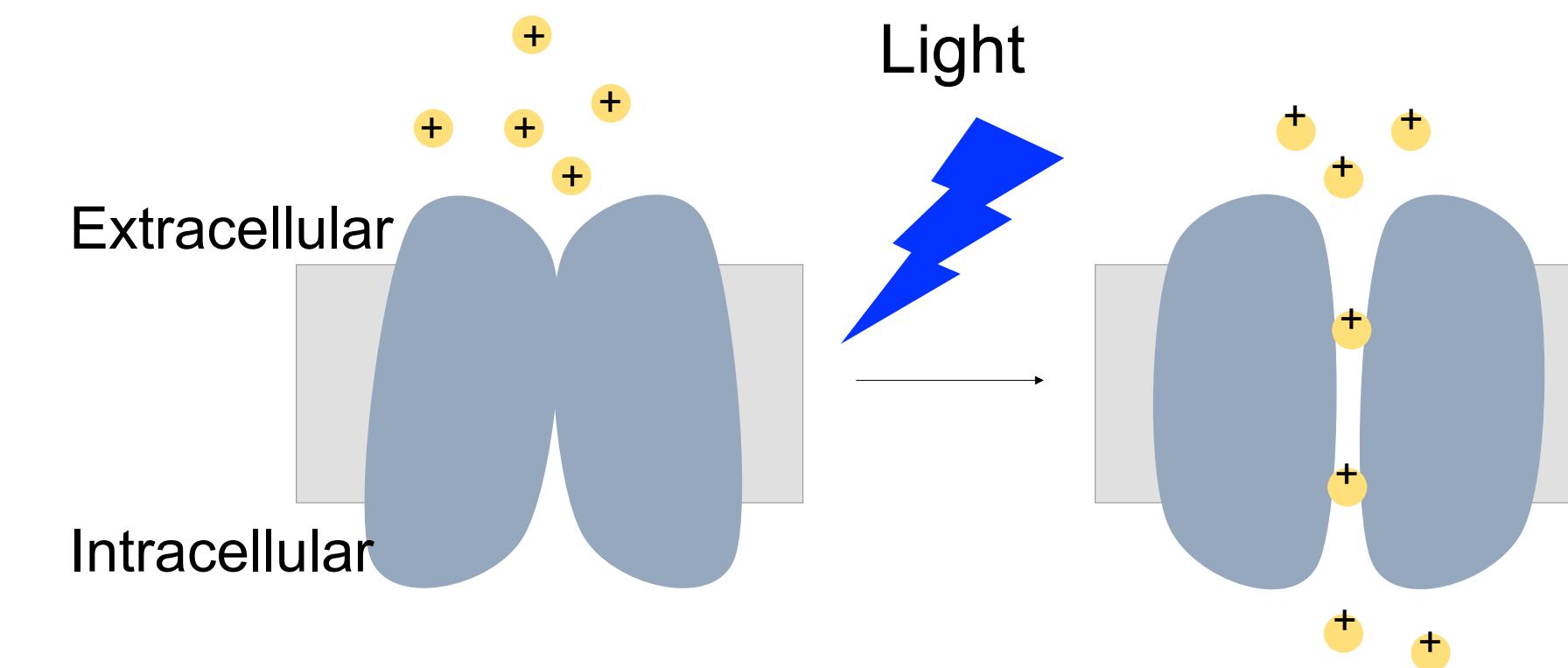
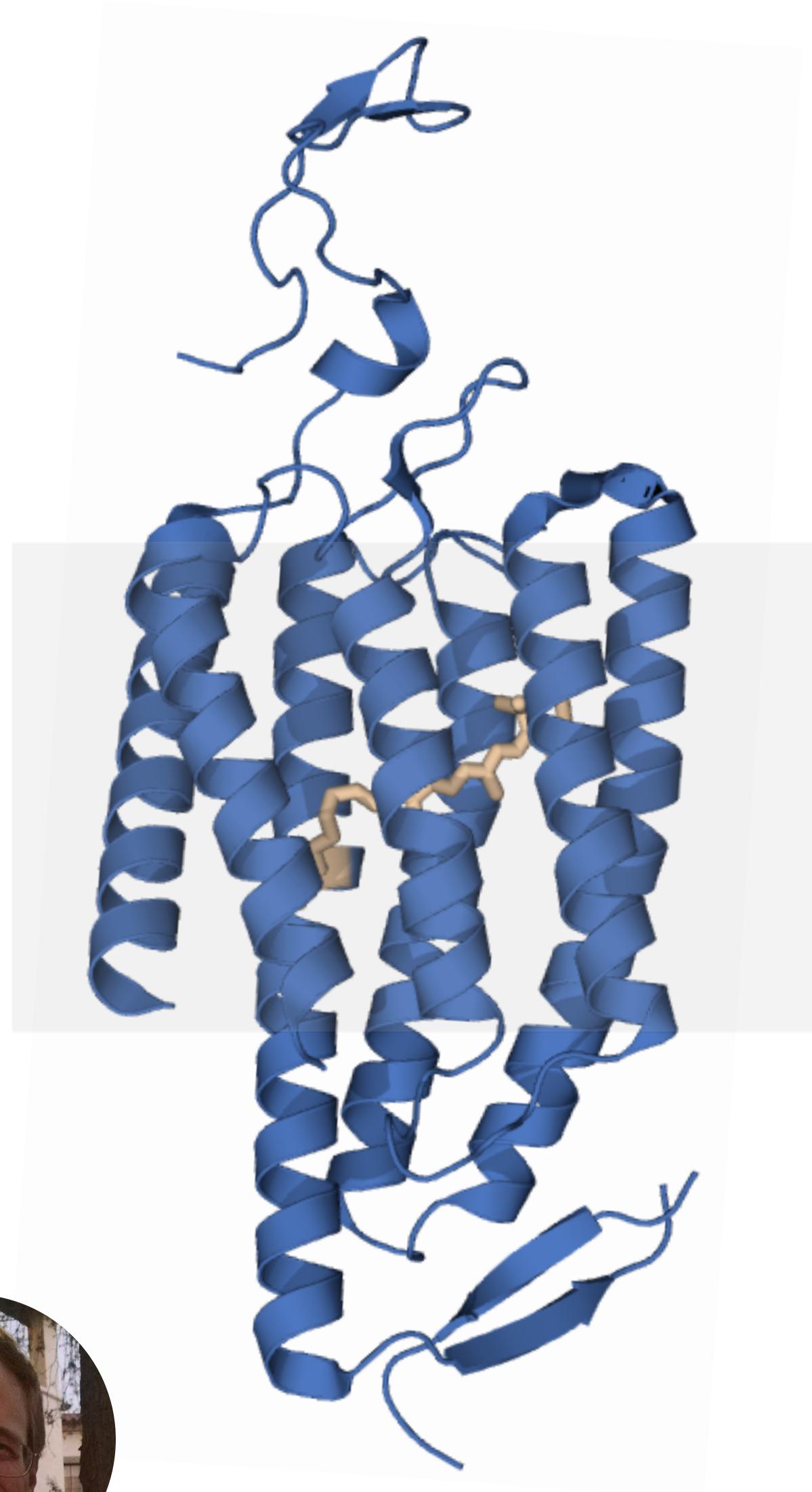
Channelrhodopsins are light-gated ion channels



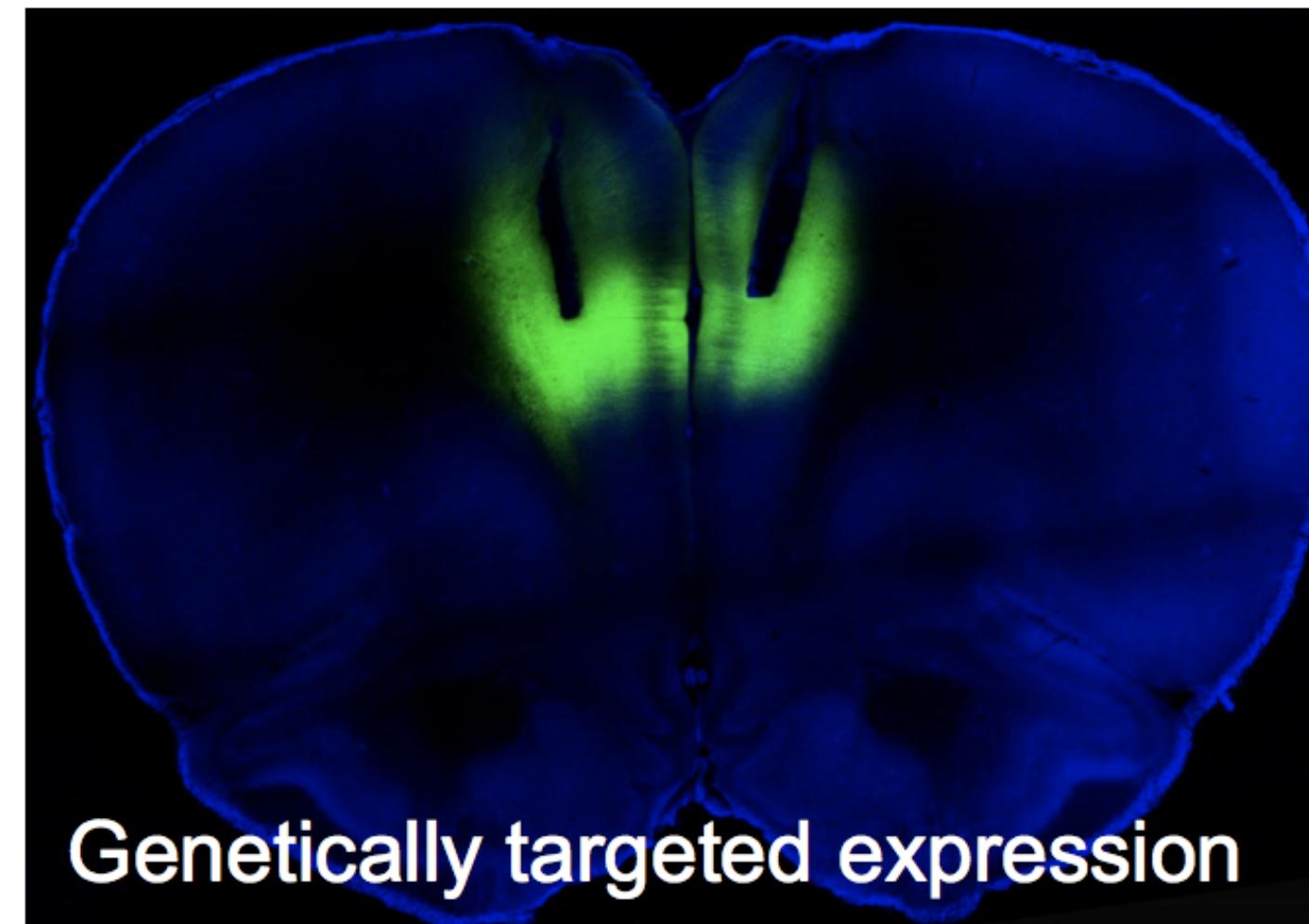
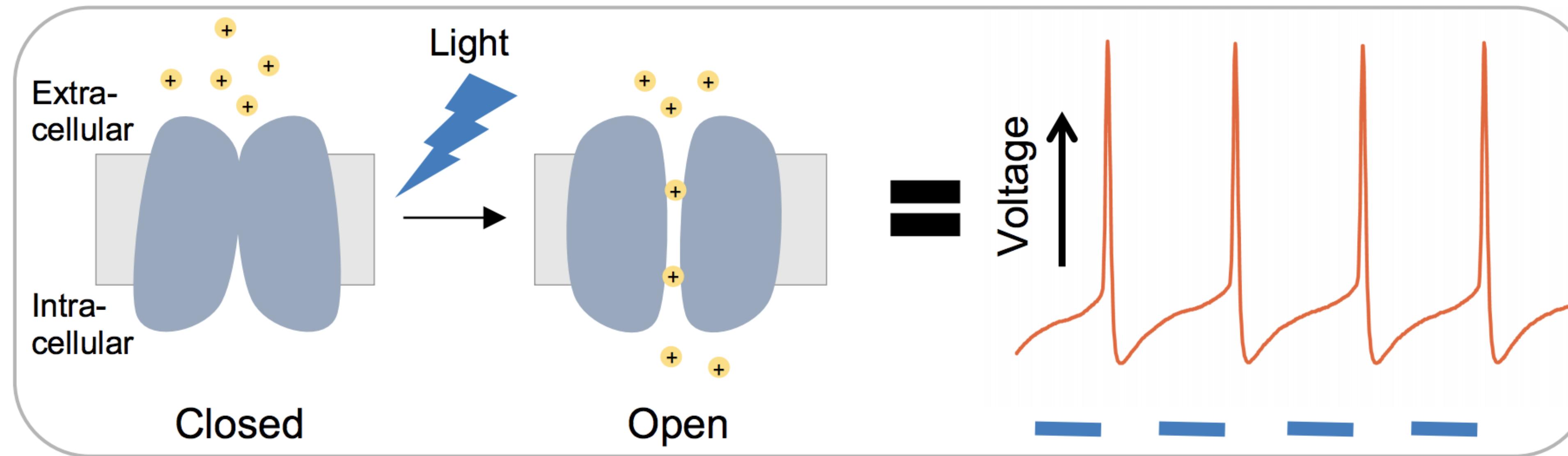
Channelrhodopsins are light-gated ion channels



Channelrhodopsins are light-gated ion channels



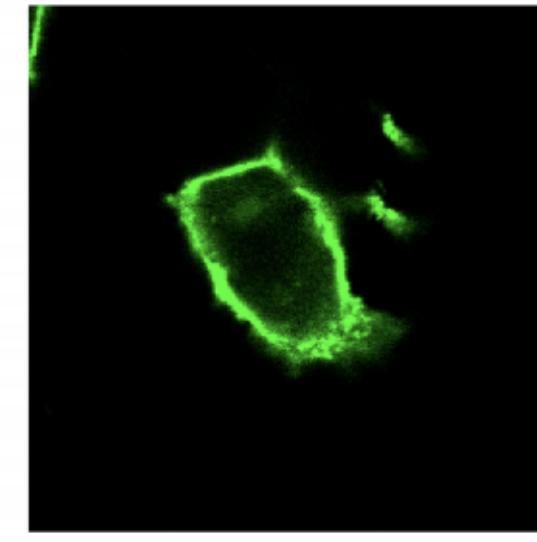
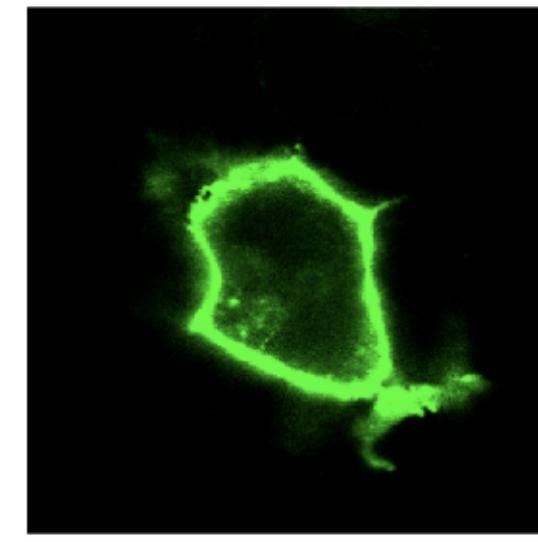
ChRs are optogenetic tools



Multiple engineering goals for optogenetics

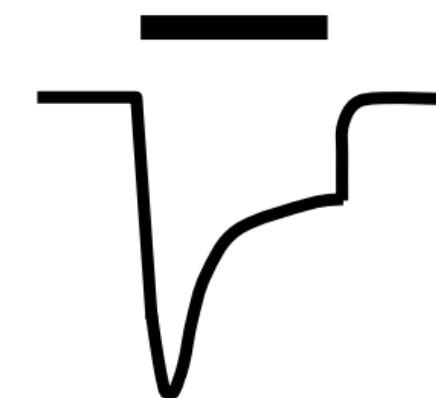
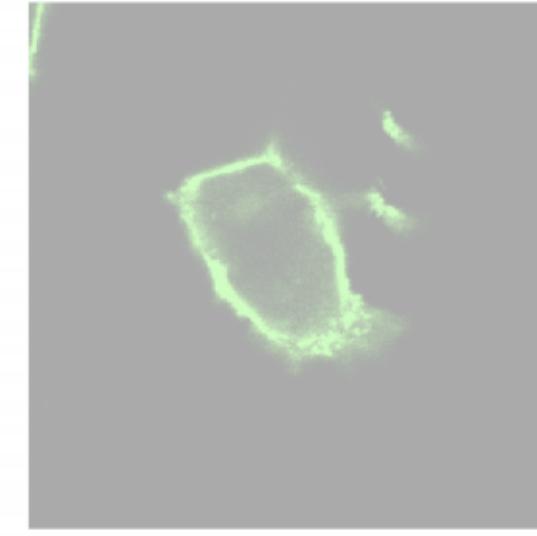
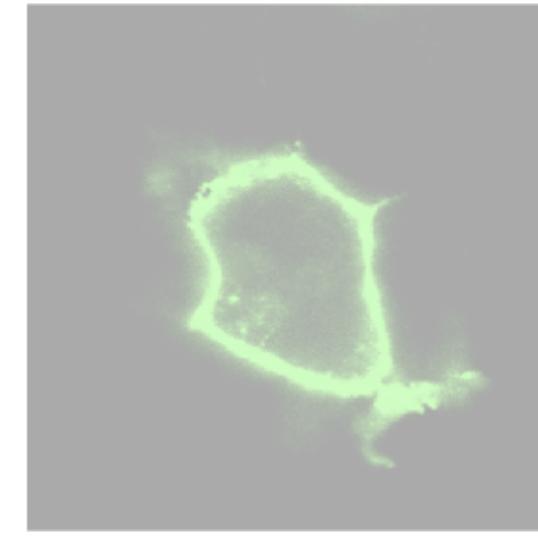
Multiple engineering goals for optogenetics

- Heterologous membrane localization



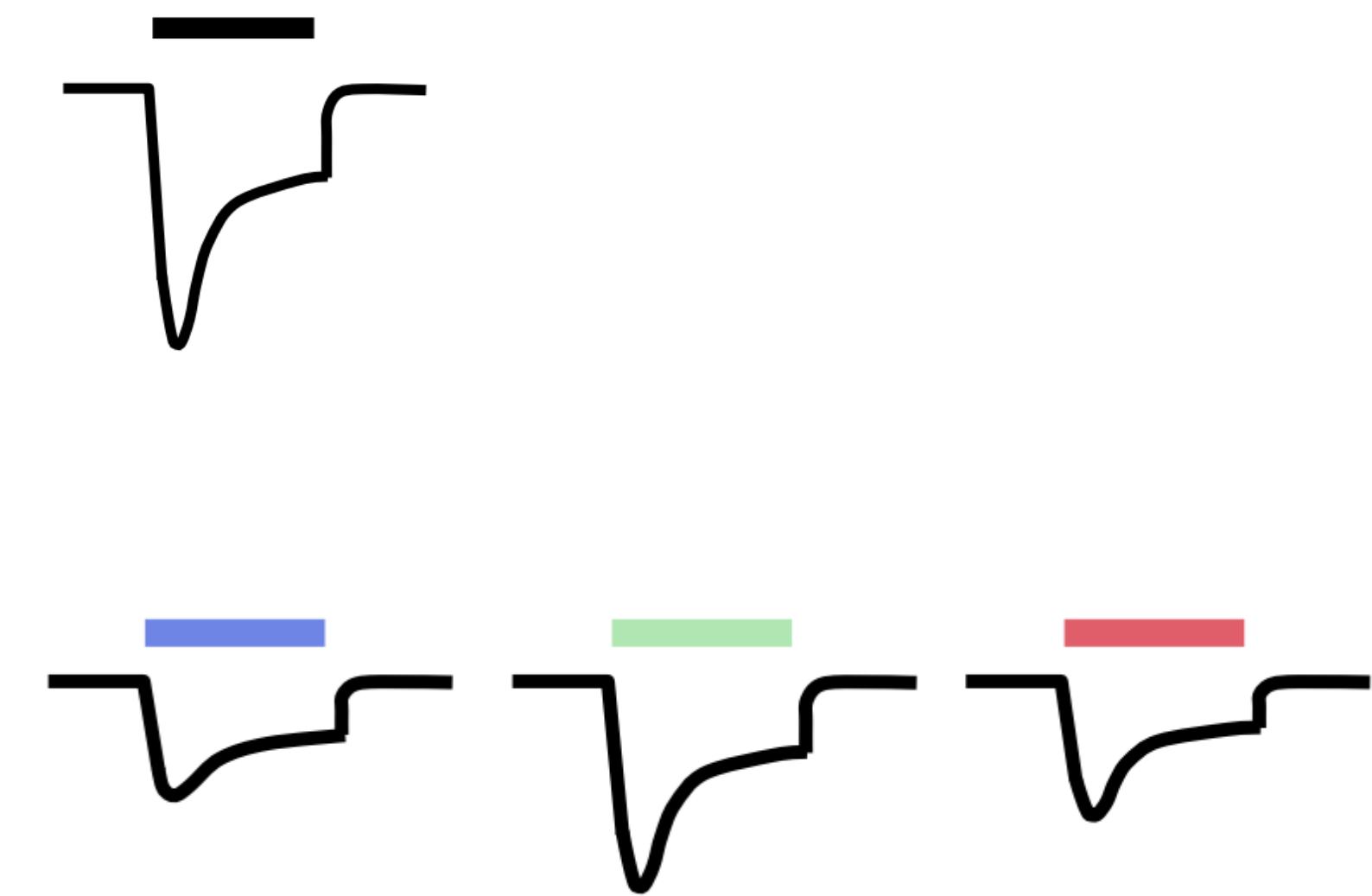
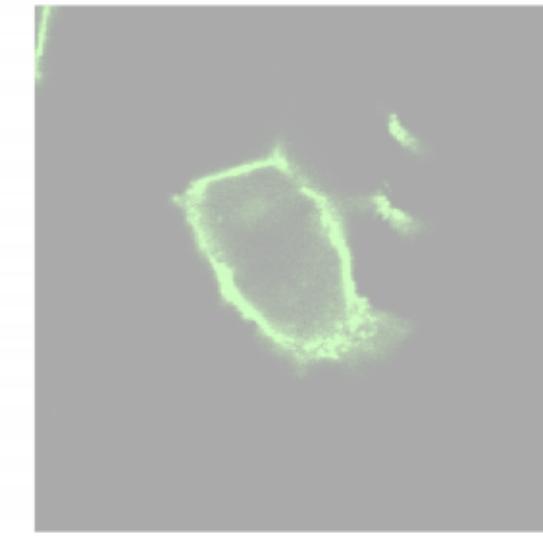
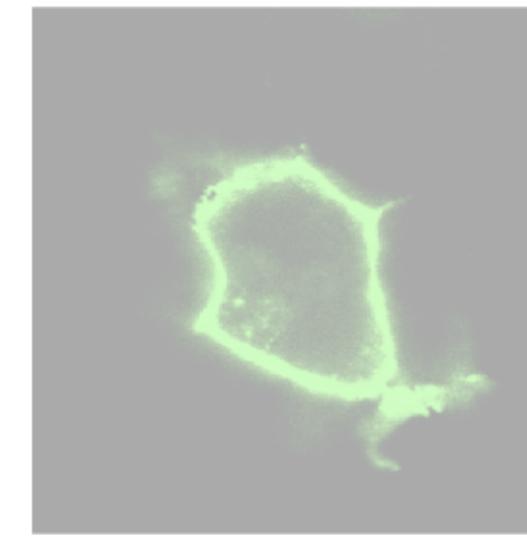
Multiple engineering goals for optogenetics

- Heterologous membrane localization
- Increased sensitivity



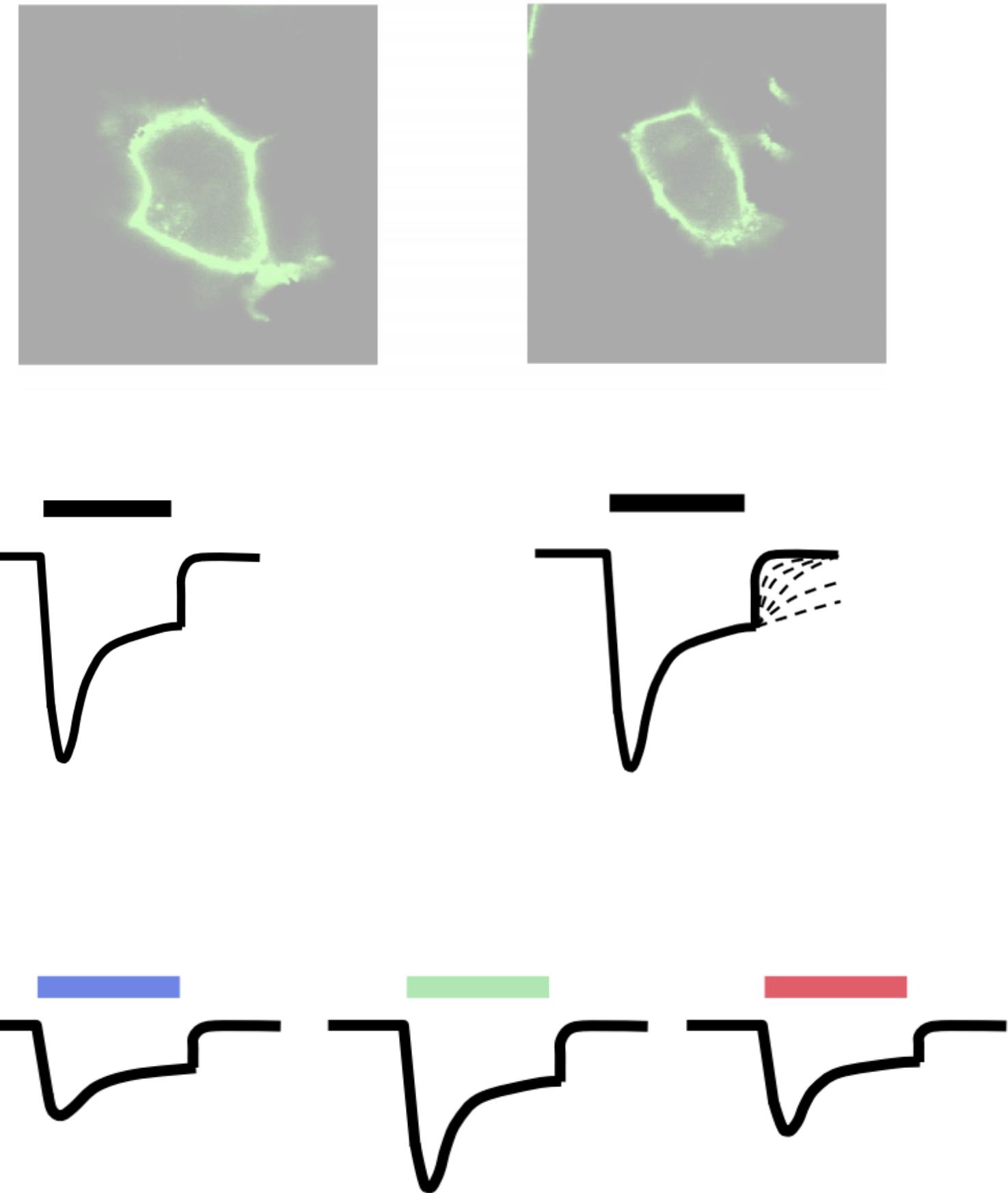
Multiple engineering goals for optogenetics

- Heterologous membrane localization
- Increased sensitivity
- Different activation wavelengths

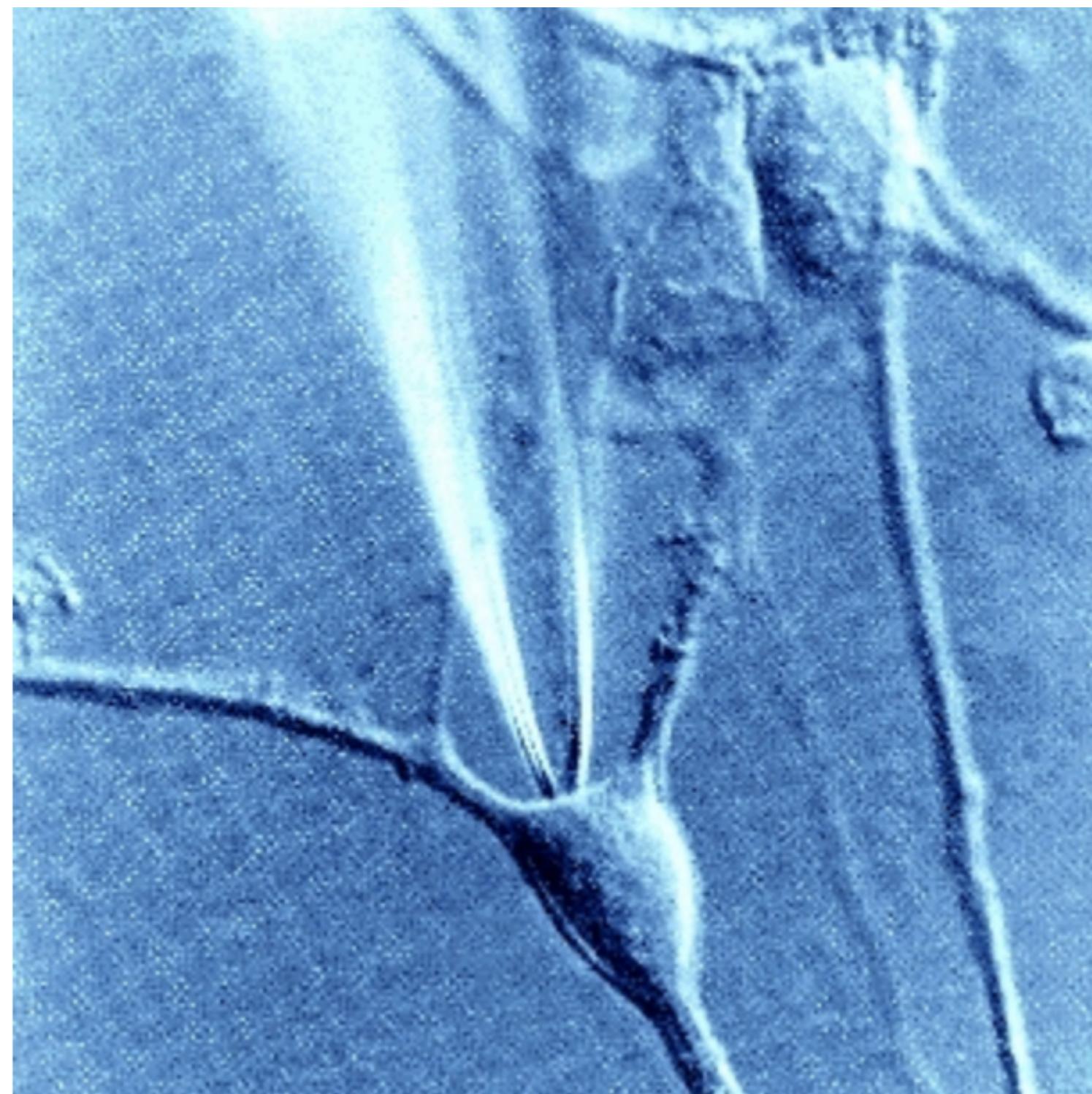


Multiple engineering goals for optogenetics

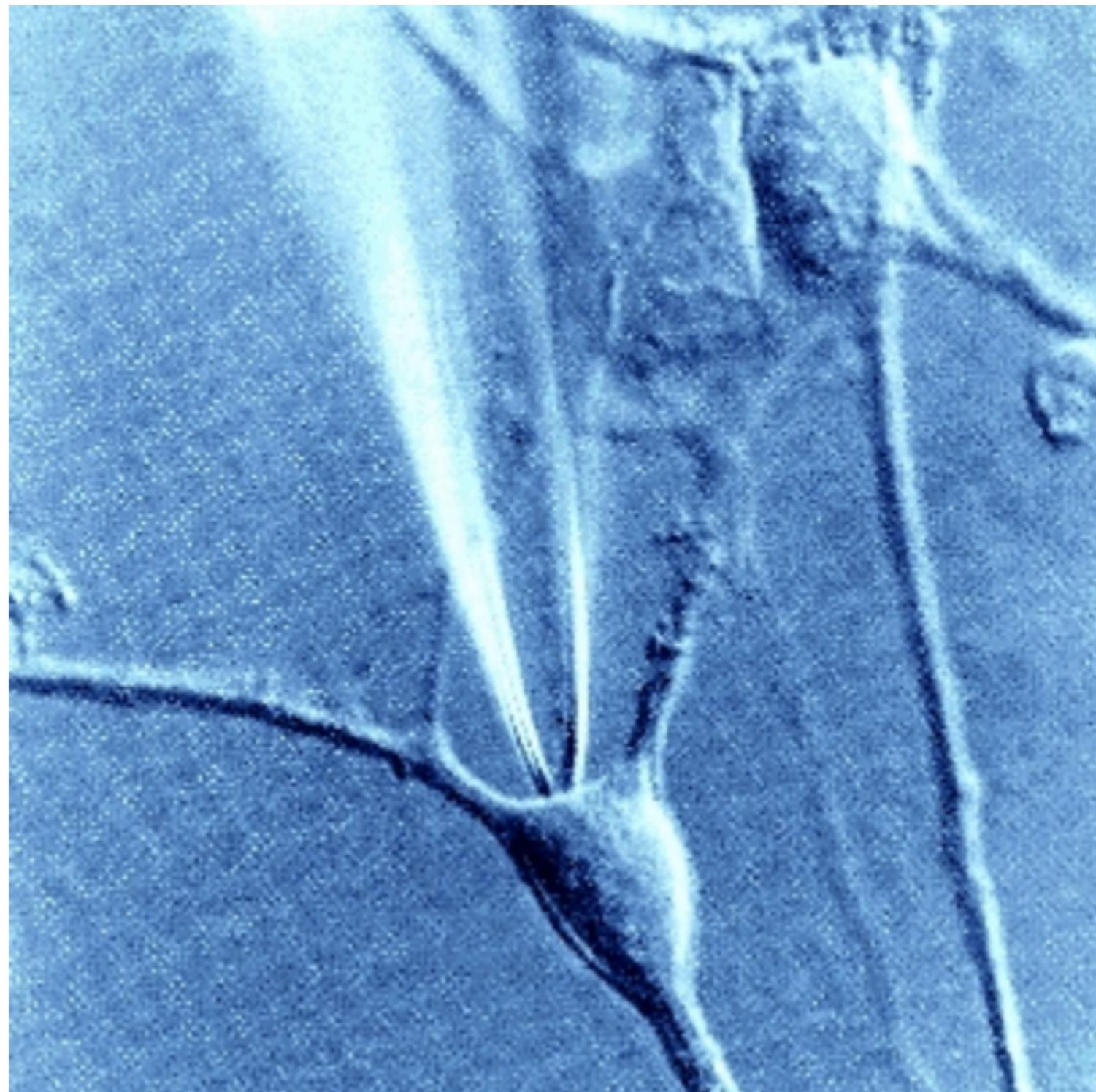
- Heterologous membrane localization
- Increased sensitivity
- Different activation wavelengths
- Different on/off kinetics



No high-throughput screen



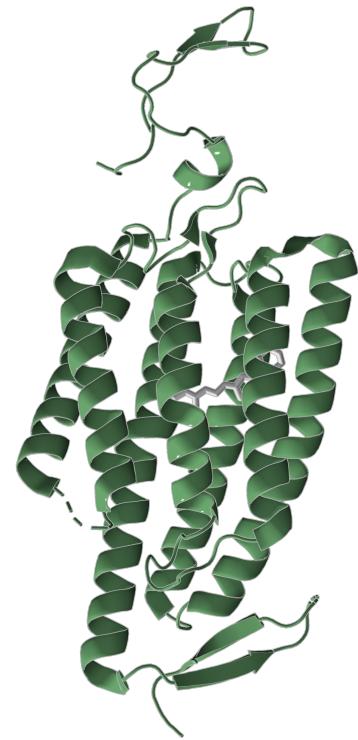
No high-throughput screen



Patch-clamp electrophysiology allows ~2 variants (with replicates) a day

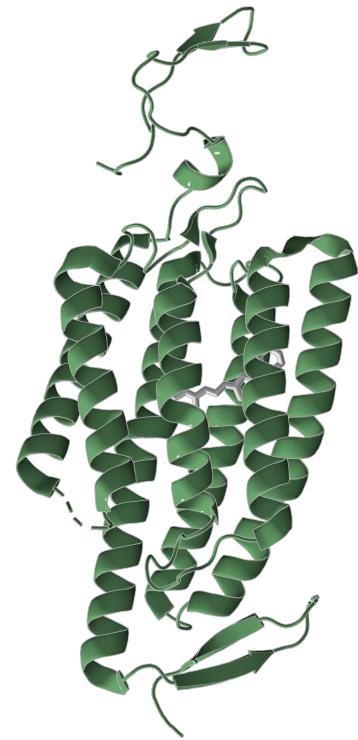
Start from 3 diverse parent ChRs

Start from 3 diverse parent ChRs

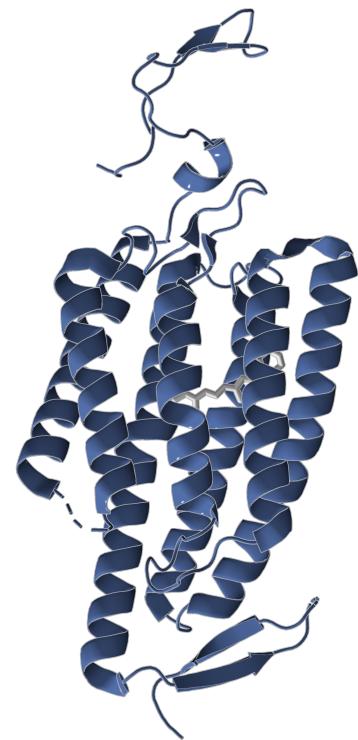


C1C2: has structure

Start from 3 diverse parent ChRs

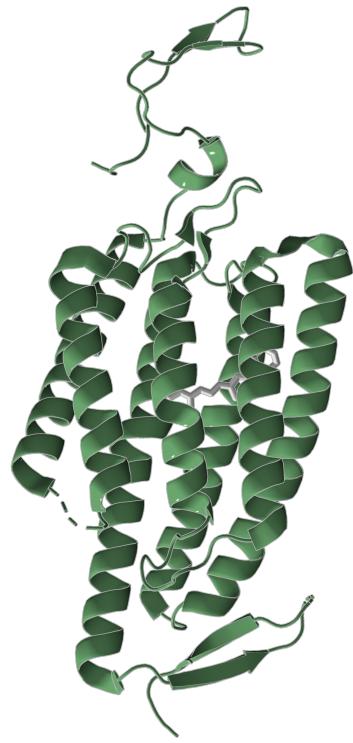


C1C2: has structure

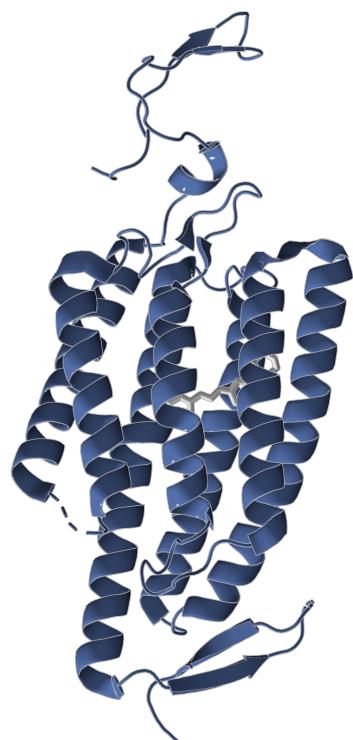


CheRiff: blue-shifted, strong currents

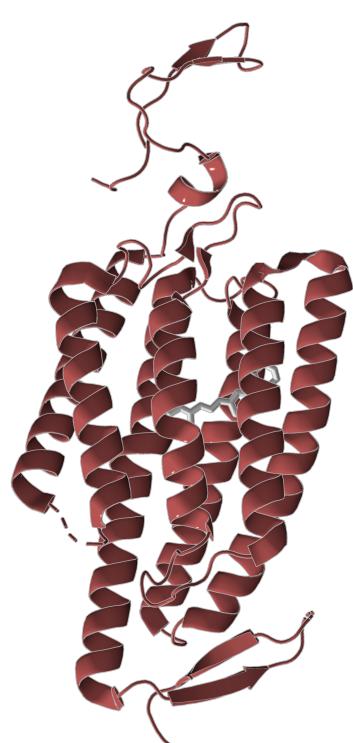
Start from 3 diverse parent ChRs



C1C2: has structure

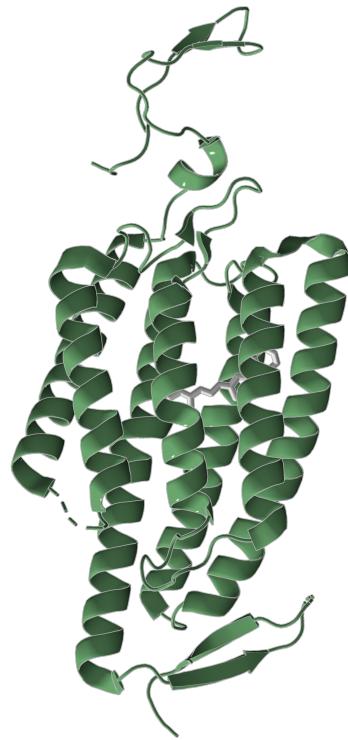


CheRiff: blue-shifted, strong currents

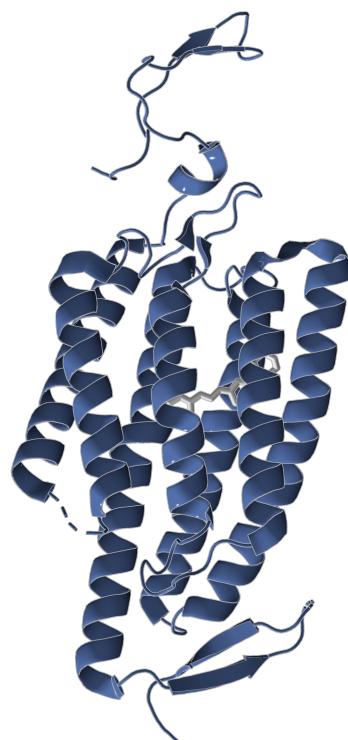


CsChrimsonR: red-shifted

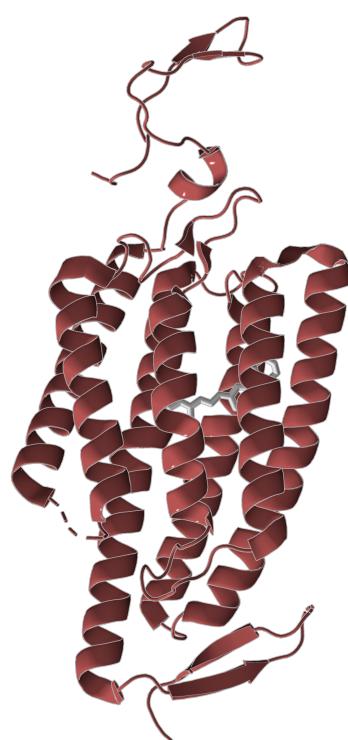
Start from 3 diverse parent ChRs



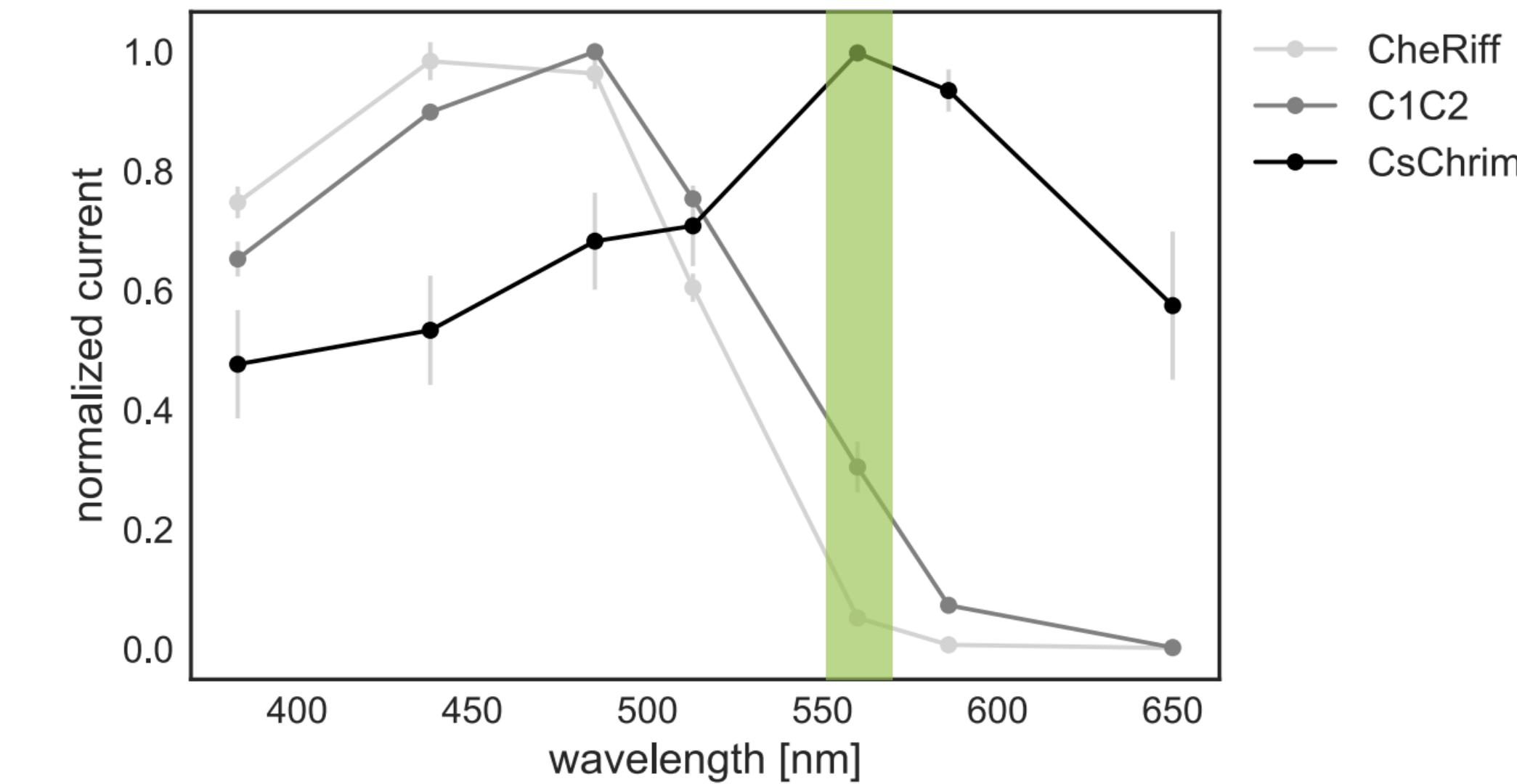
C1C2: has structure



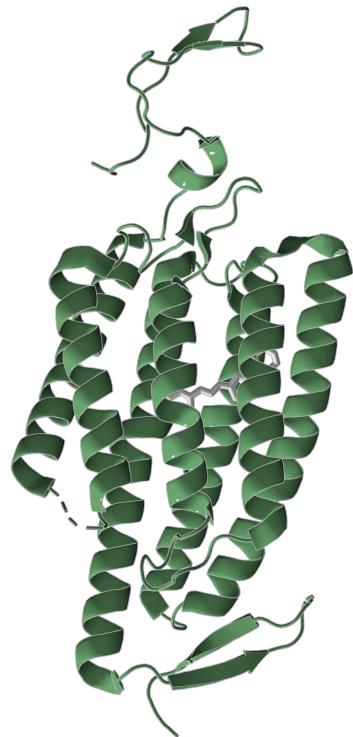
CheRiff: blue-shifted, strong currents



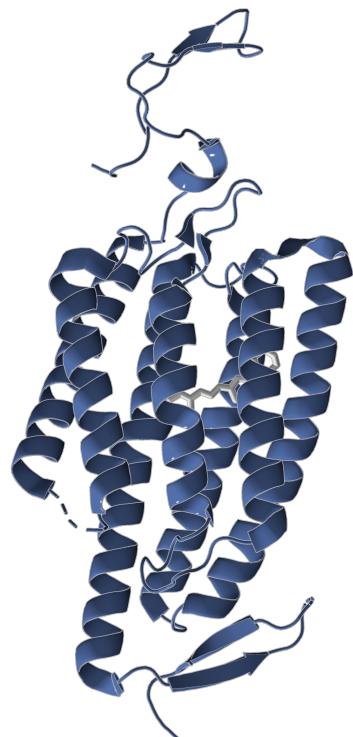
CsChrimsonR: red-shifted



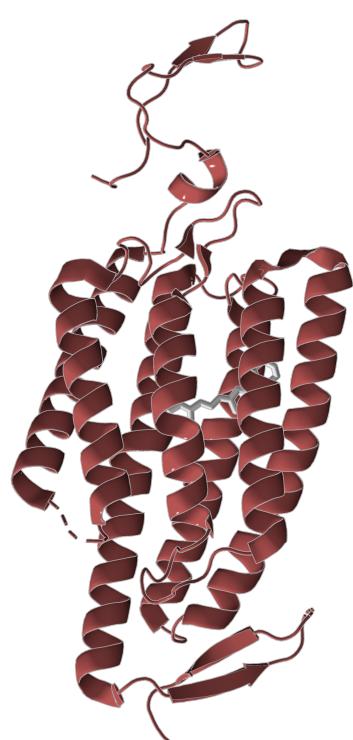
Start from 3 diverse parent ChRs



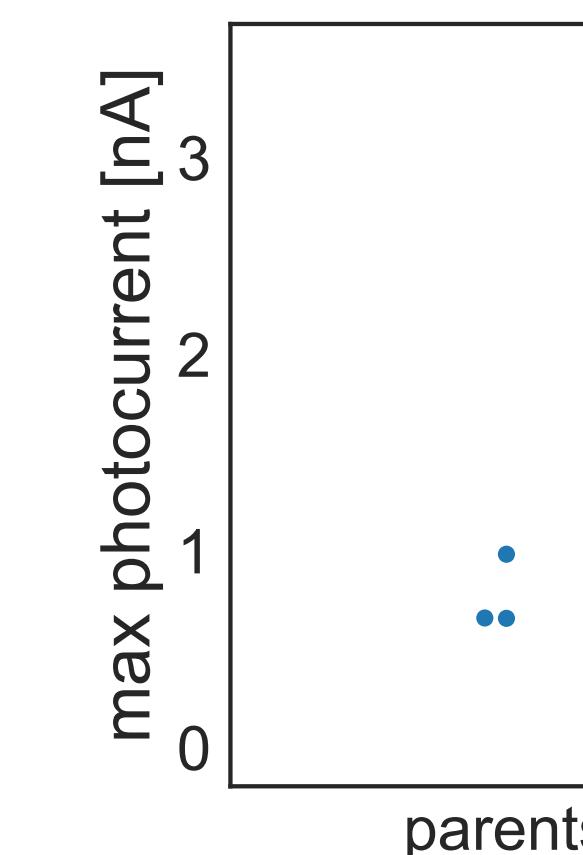
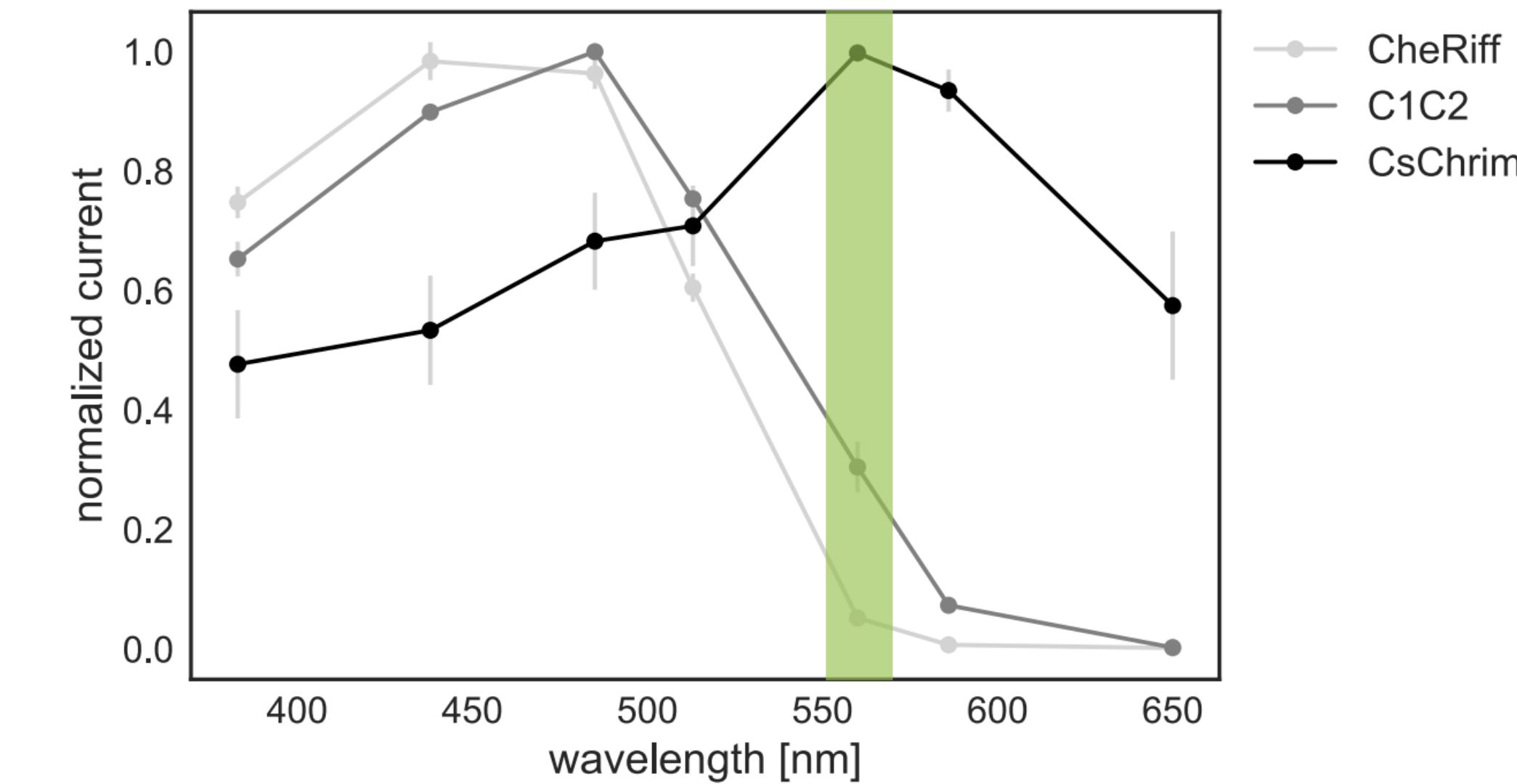
C1C2: has structure



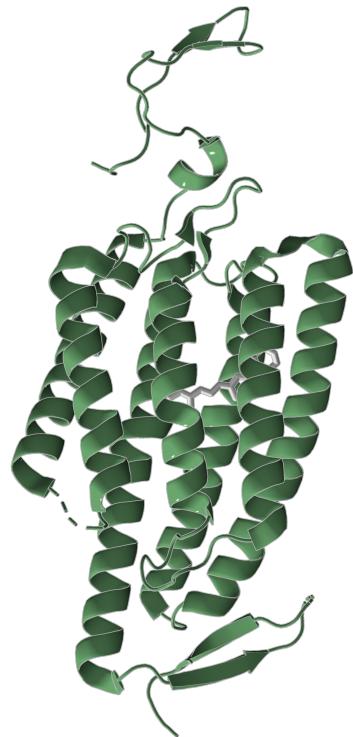
CheRiff: blue-shifted, strong currents



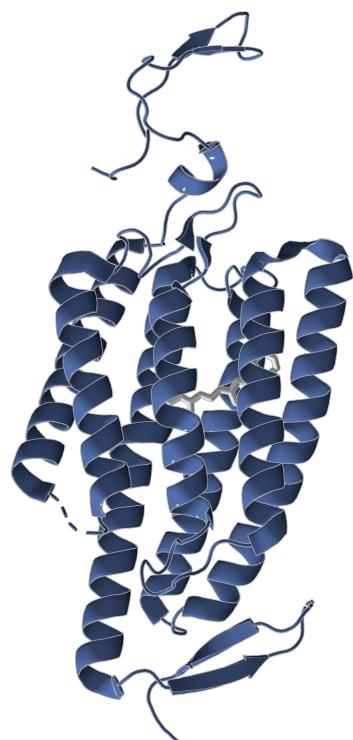
CsChrimsonR: red-shifted



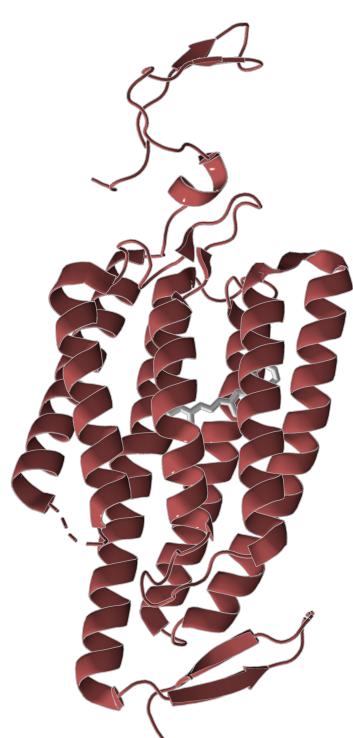
Start from 3 diverse parent ChRs



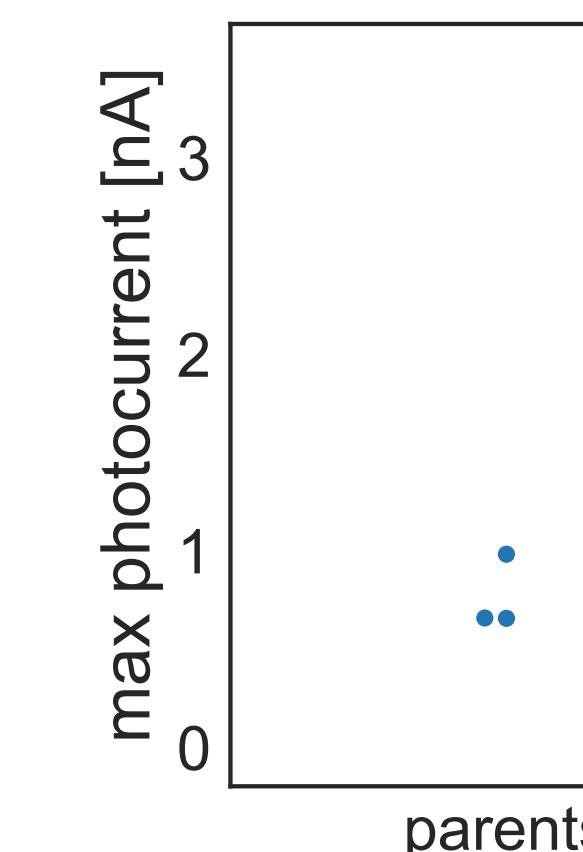
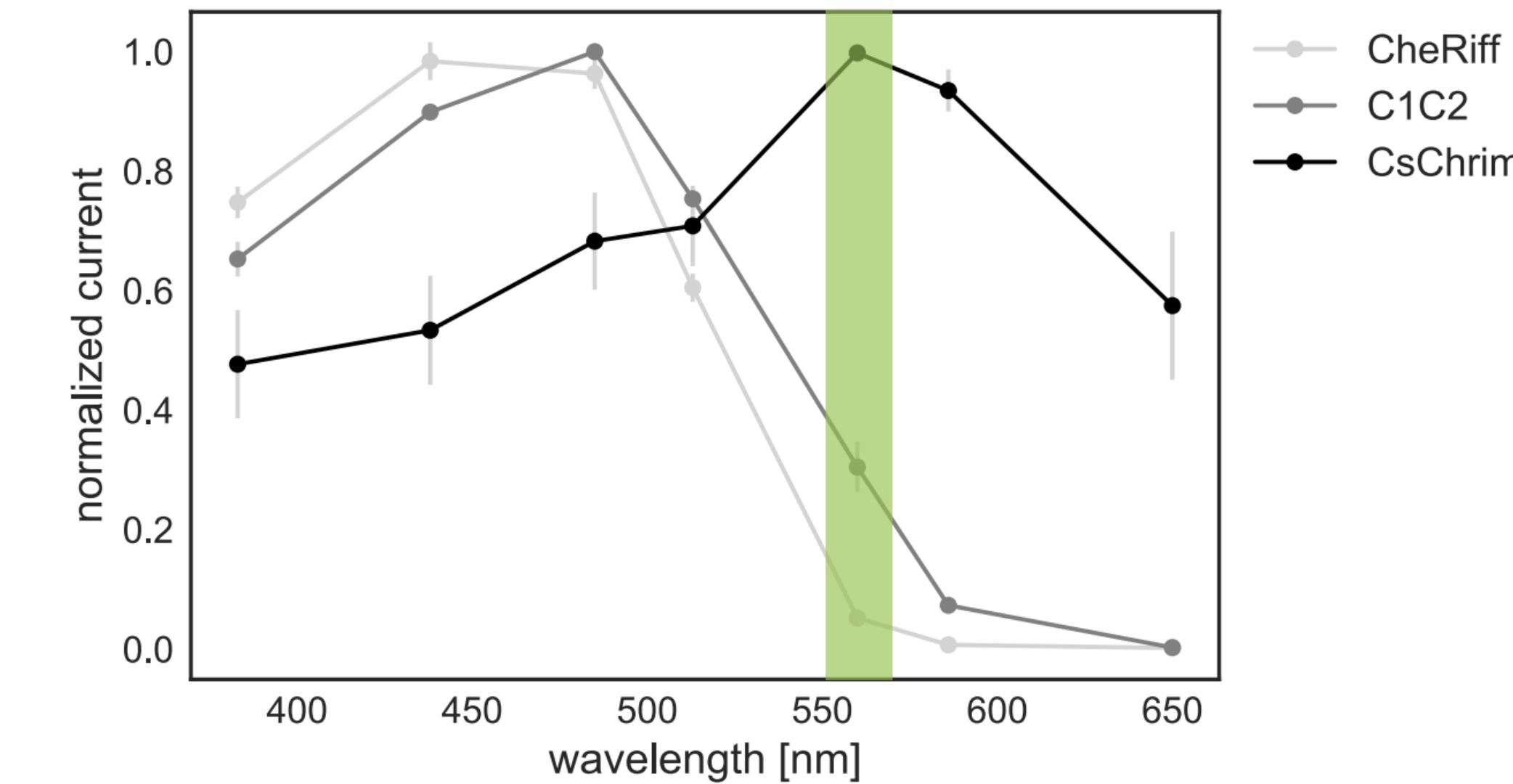
C1C2: has structure



CheRiff: blue-shifted, strong currents



CsChrimsonR: red-shifted

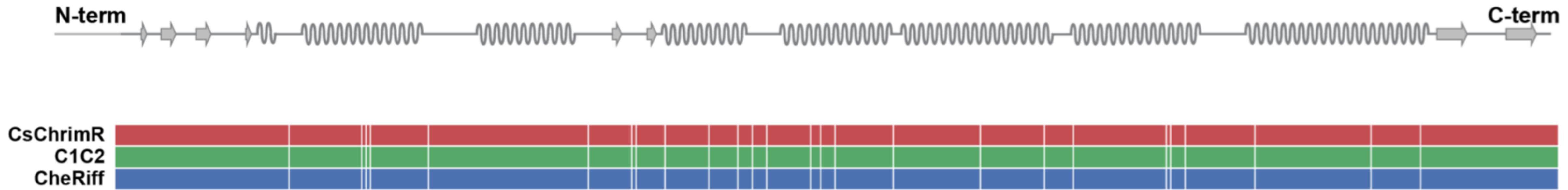


Kato 2012
Hochbaum 2014
Klapoetke 2014

Use recombination to generate diversity

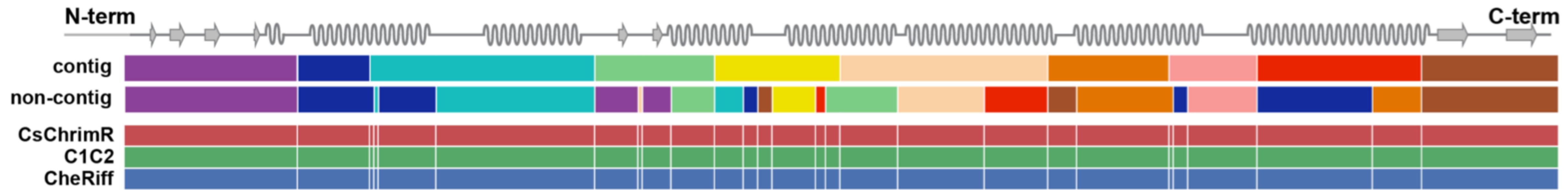
Voigt 2002
Otey 2006
Smith 2013

Use recombination to generate diversity



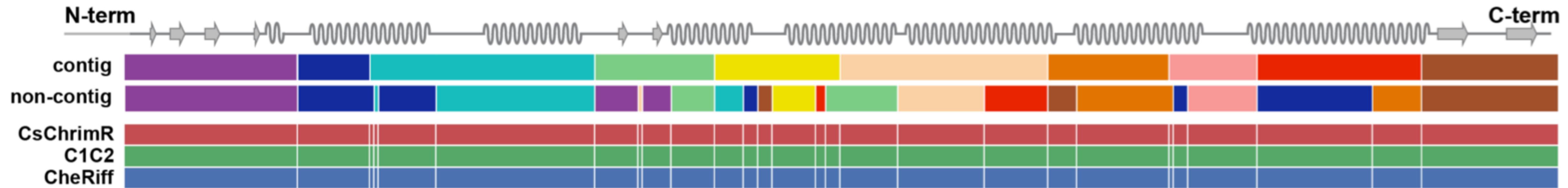
Voigt 2002
Otey 2006
Smith 2013

Use recombination to generate diversity



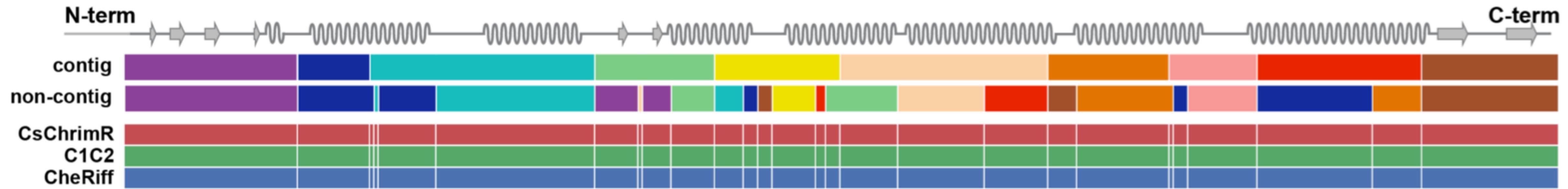
Voigt 2002
Otey 2006
Smith 2013

Use recombination to generate diversity



2 libraries
3 parents
10 blocks

Use recombination to generate diversity

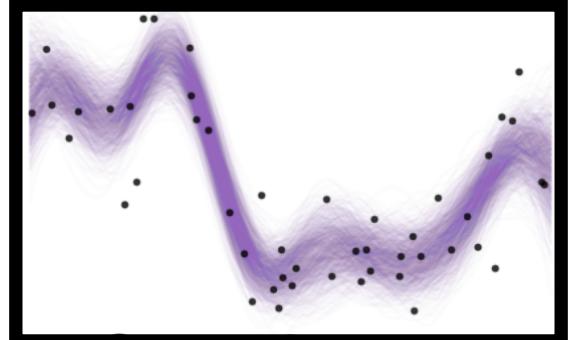


2 libraries

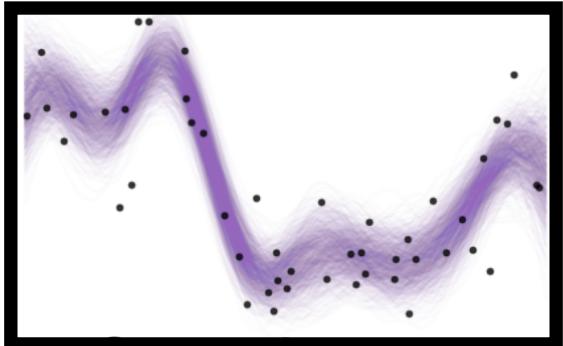
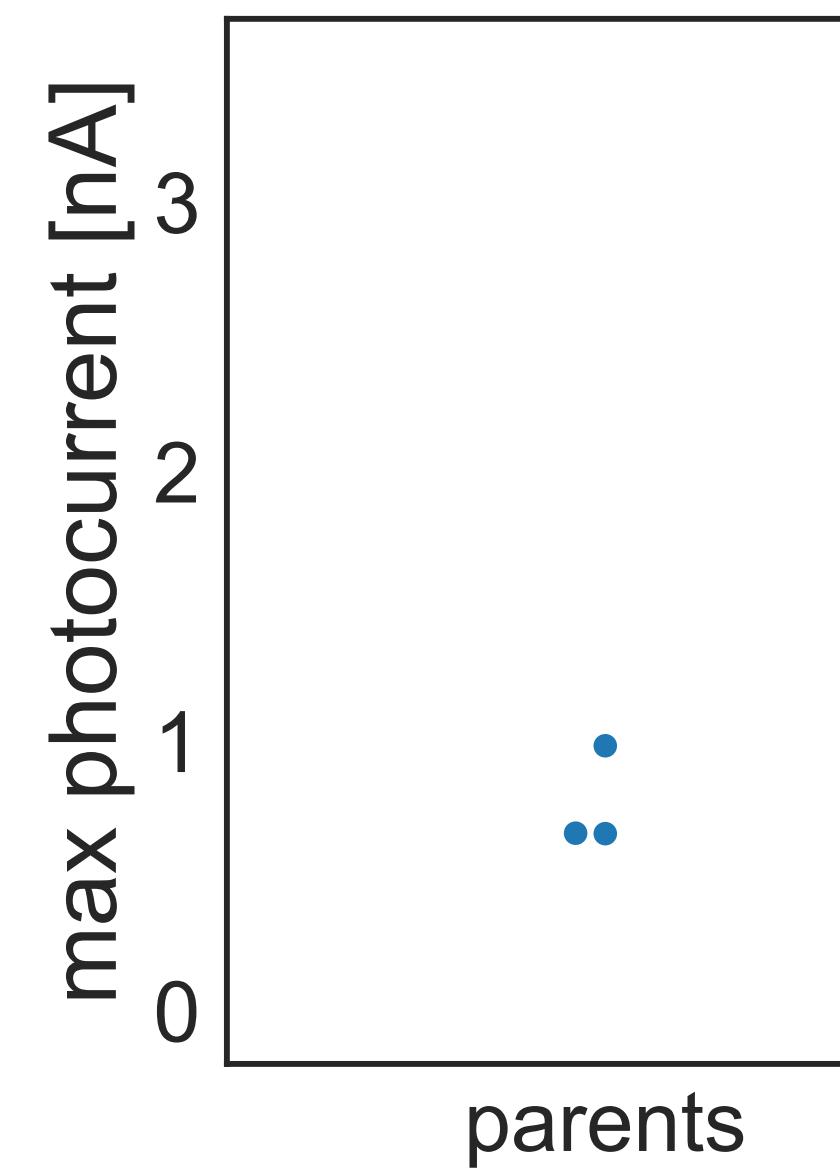
3 parents = 118,098 sequences

10 blocks

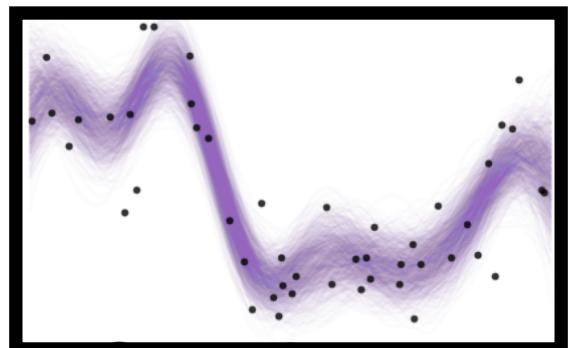
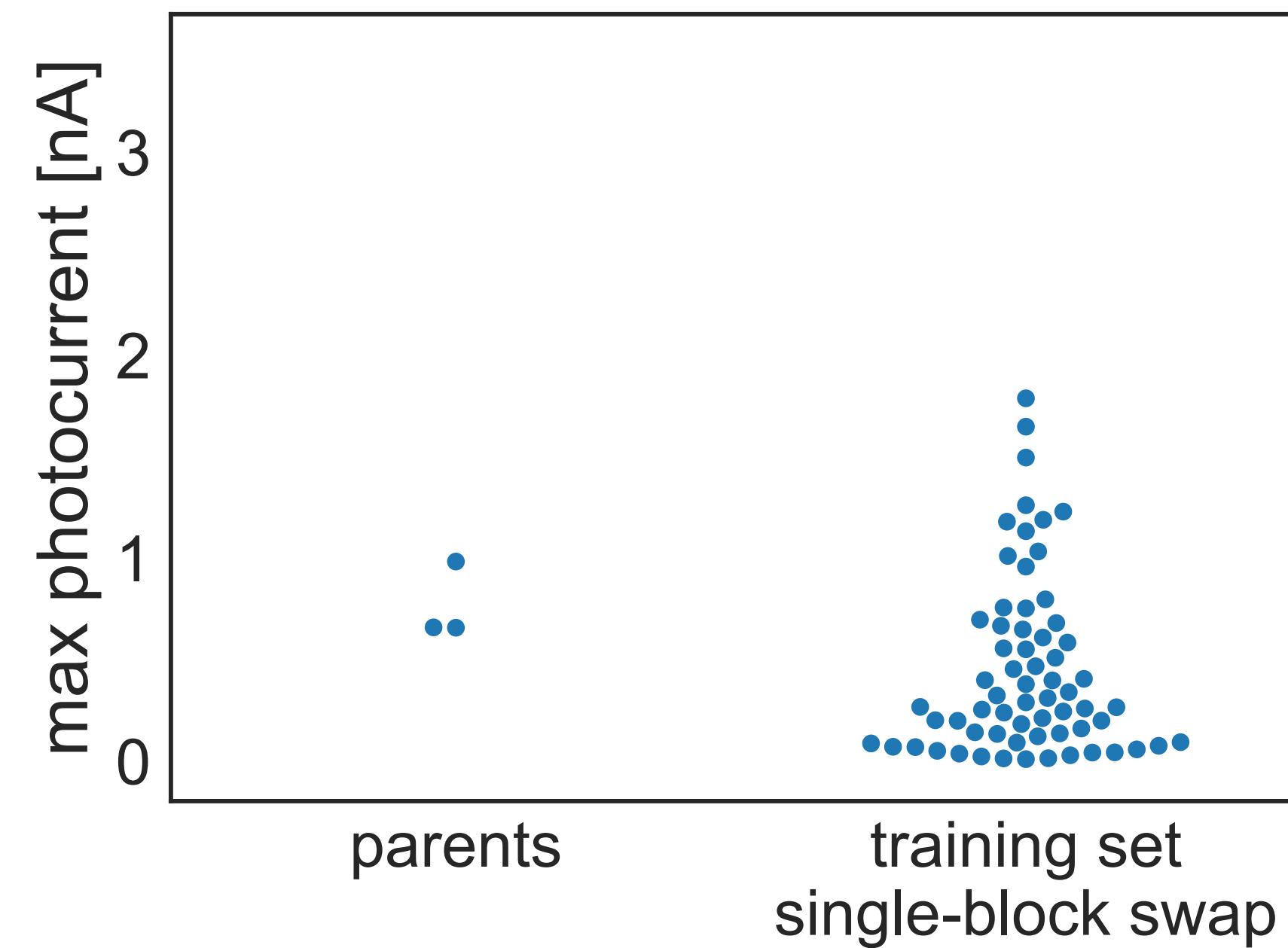
Machine learning enables optimization with fewer measurements



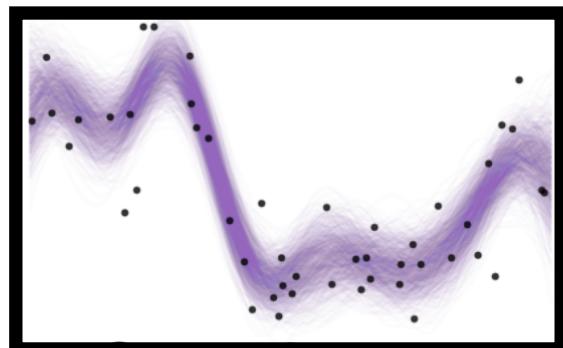
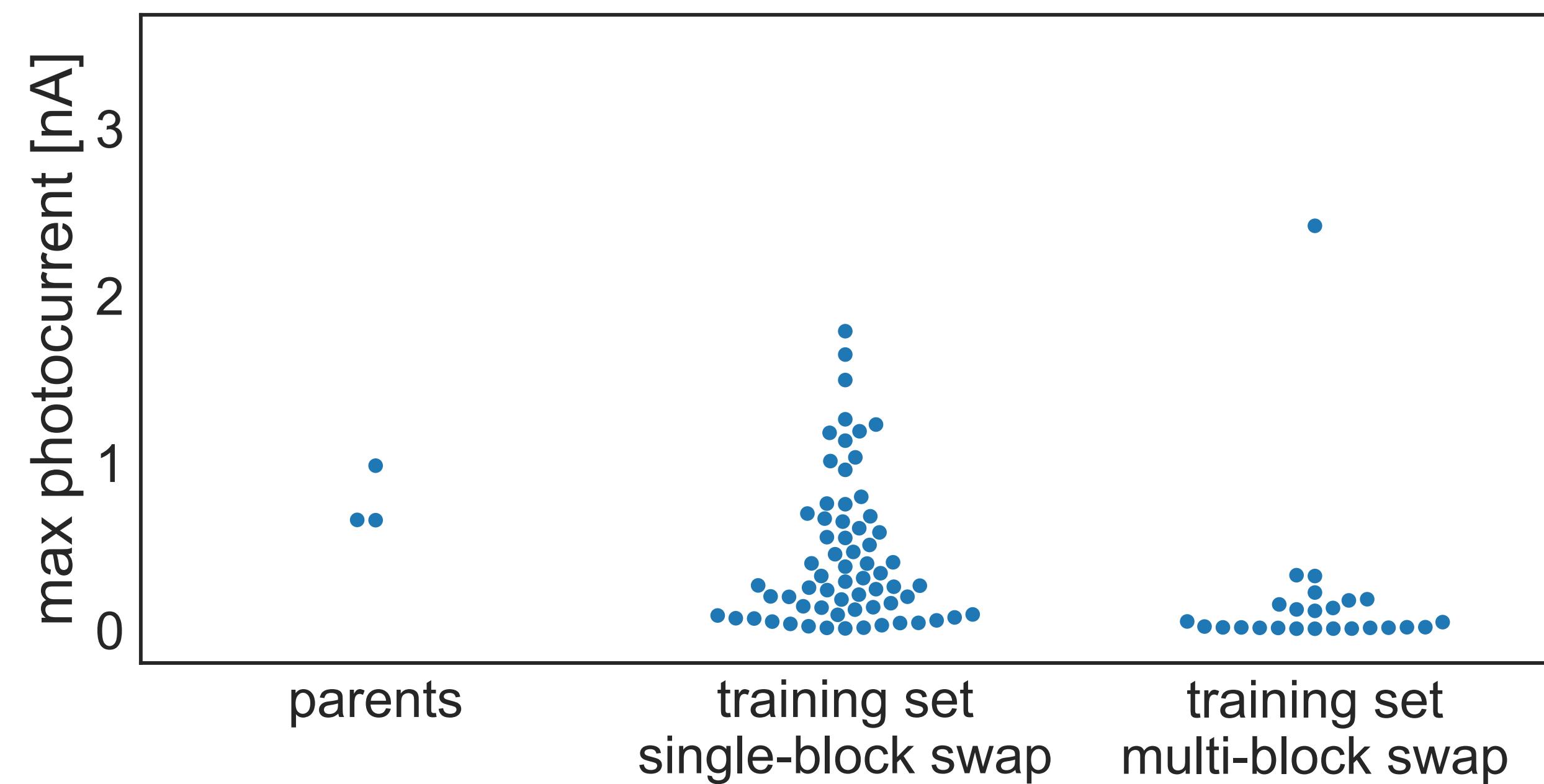
Machine learning enables optimization with fewer measurements



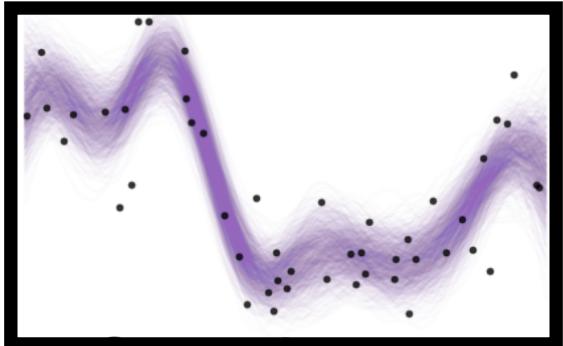
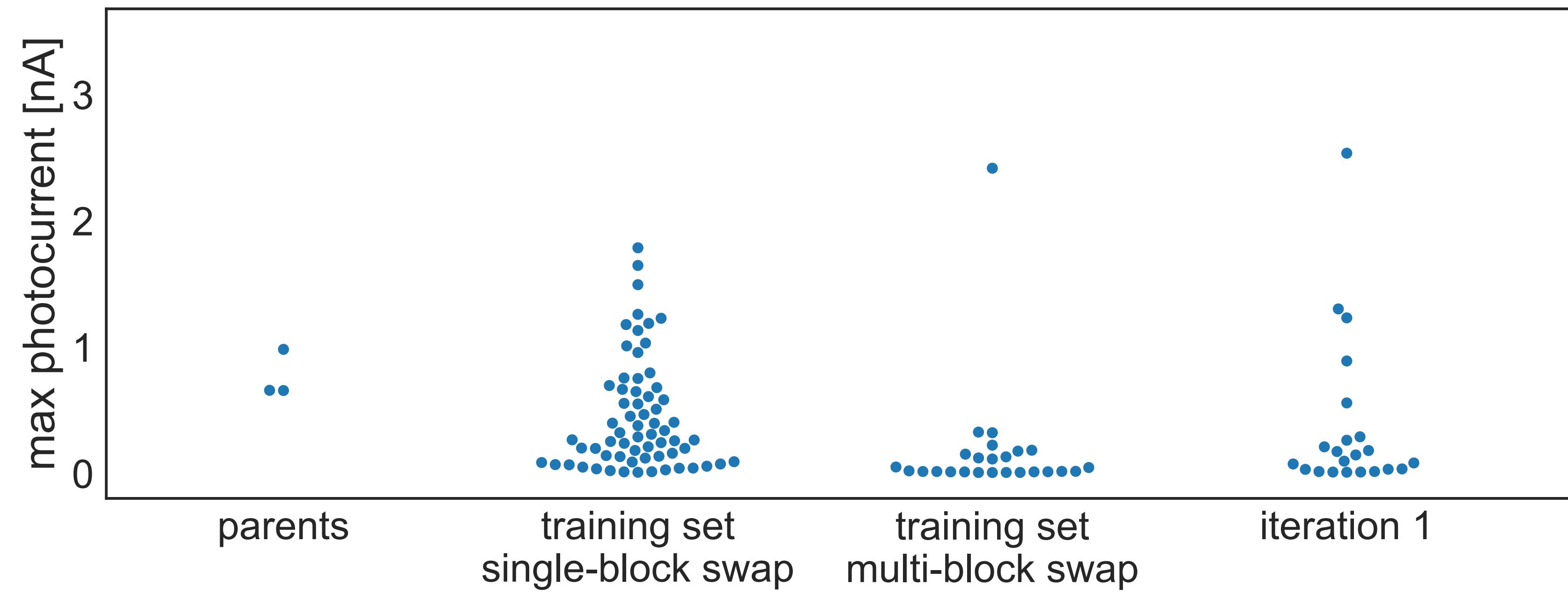
Machine learning enables optimization with fewer measurements



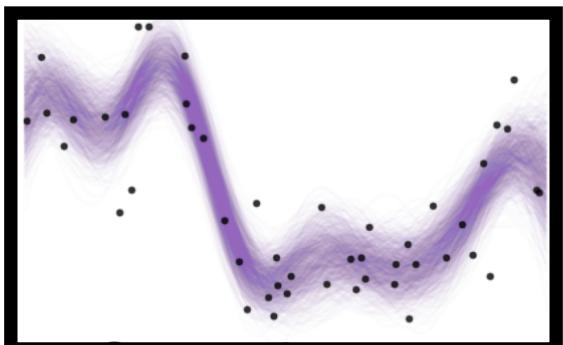
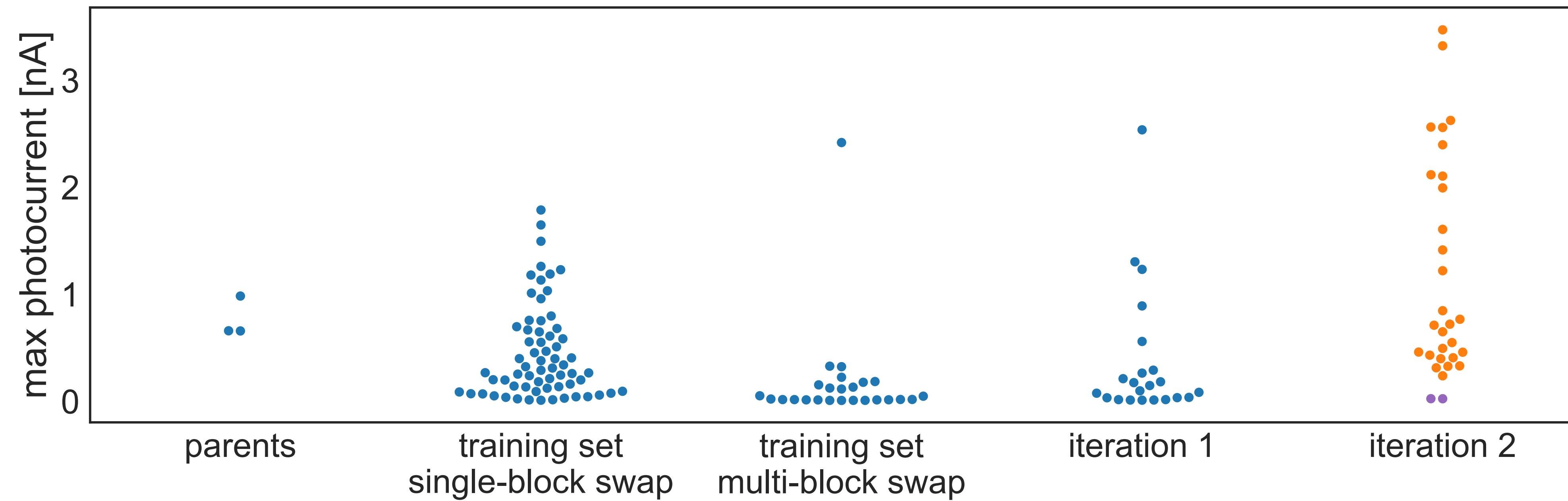
Machine learning enables optimization with fewer measurements



Machine learning enables optimization with fewer measurements



Machine learning enables optimization with fewer measurements

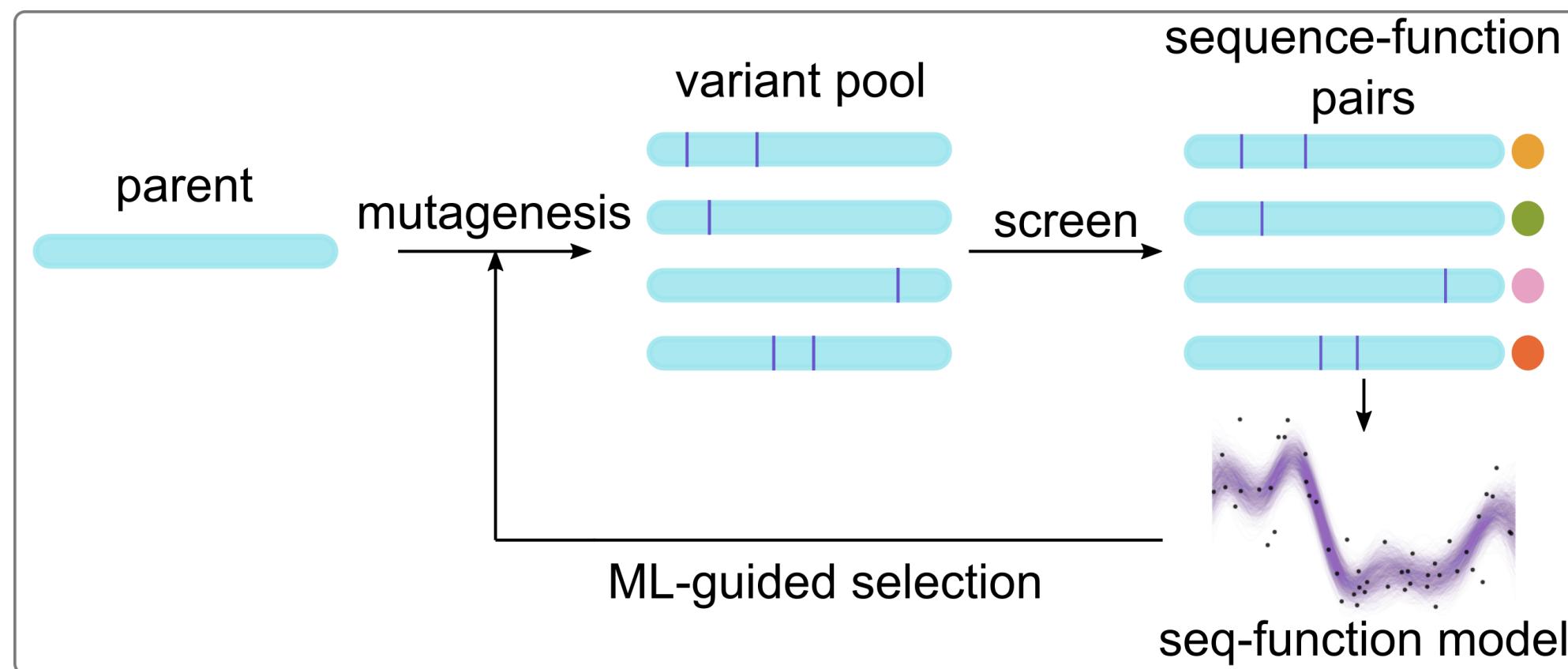


Engineered ChRs control mouse neurons without skull removal

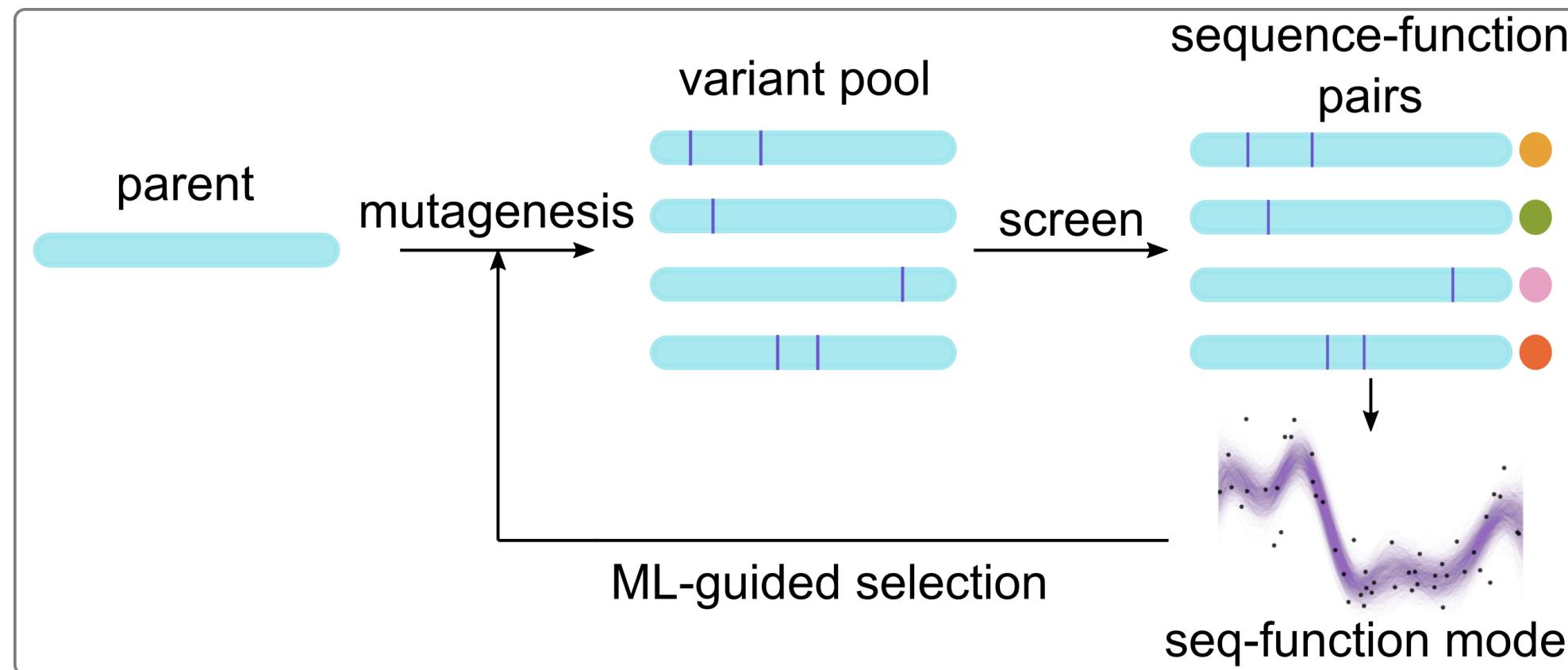
Engineered ChRs control mouse neurons without skull removal



MLDE: successes but limitations

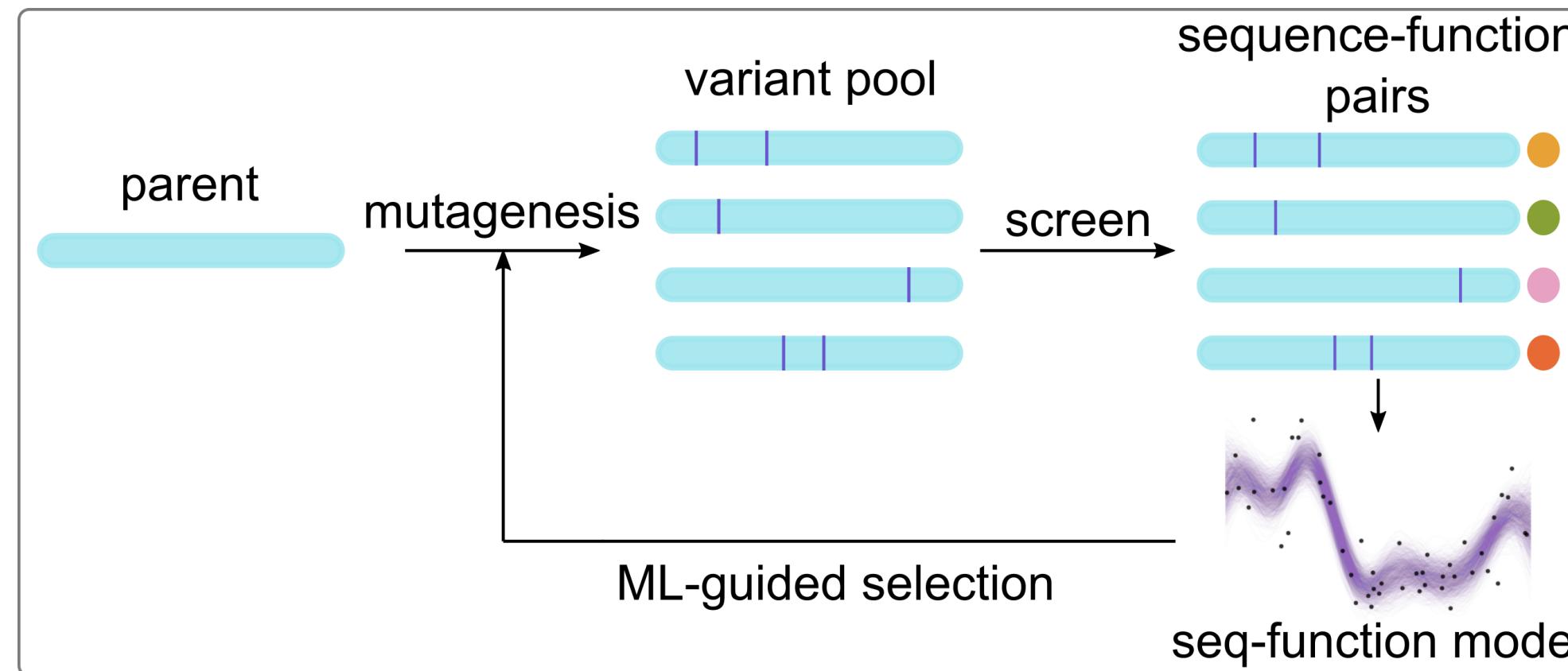


MLDE: successes but limitations



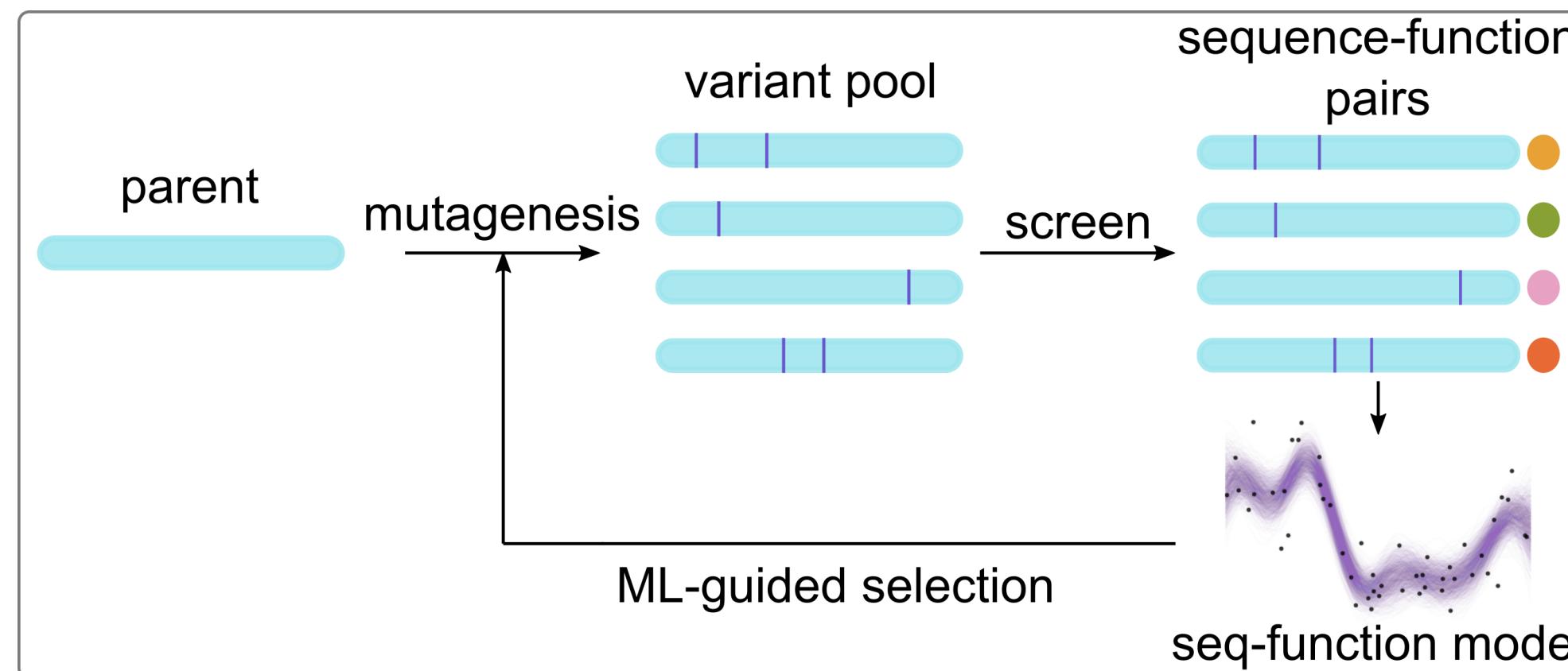
- Accurate models with < 200 measurements

MLDE: successes but limitations



- Accurate models with < 200 measurements
- Find rare, improved sequences with < 200 measurements

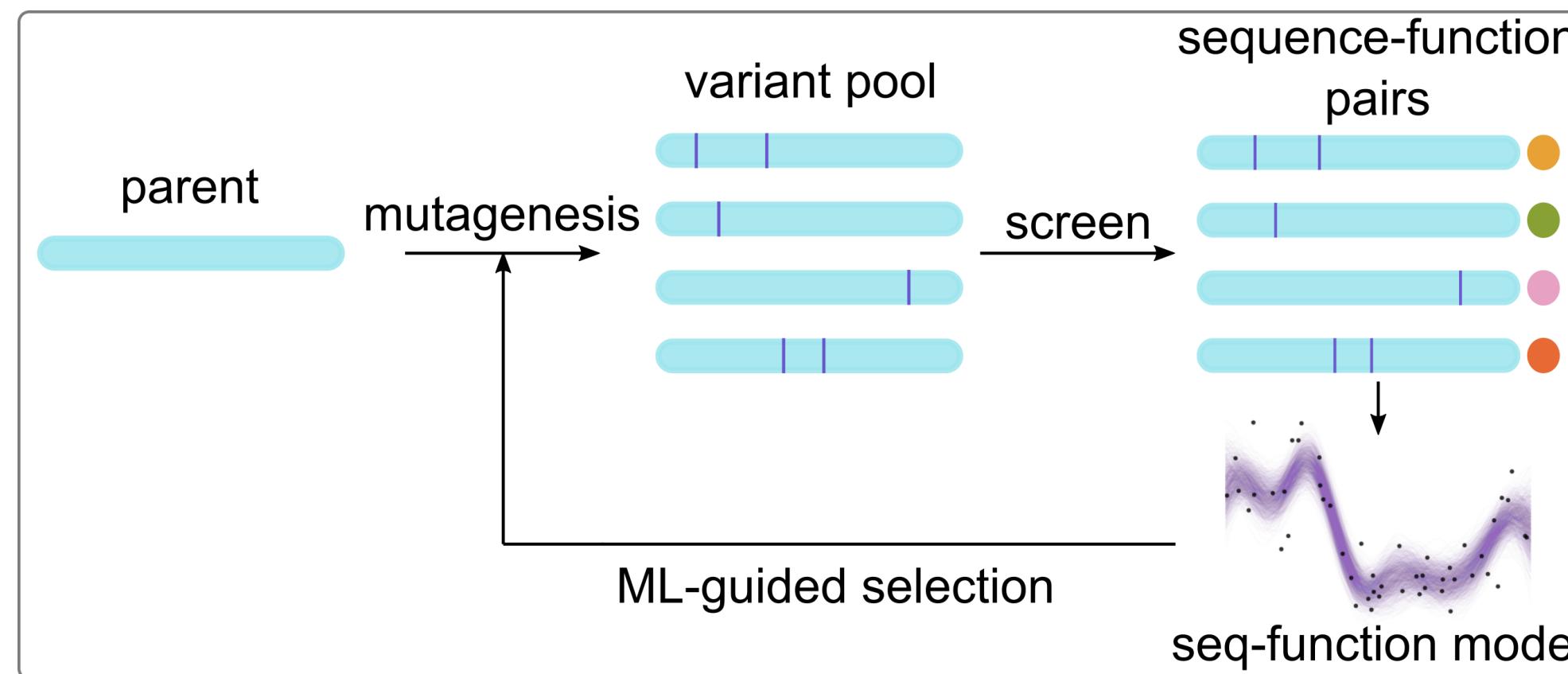
MLDE: successes but limitations



- Accurate models with < 200 measurements
- Find rare, improved sequences with < 200 measurements

Still requires starting point!

MLDE: successes but limitations



- Accurate models with < 200 measurements
- Find rare, improved sequences with < 200 measurements

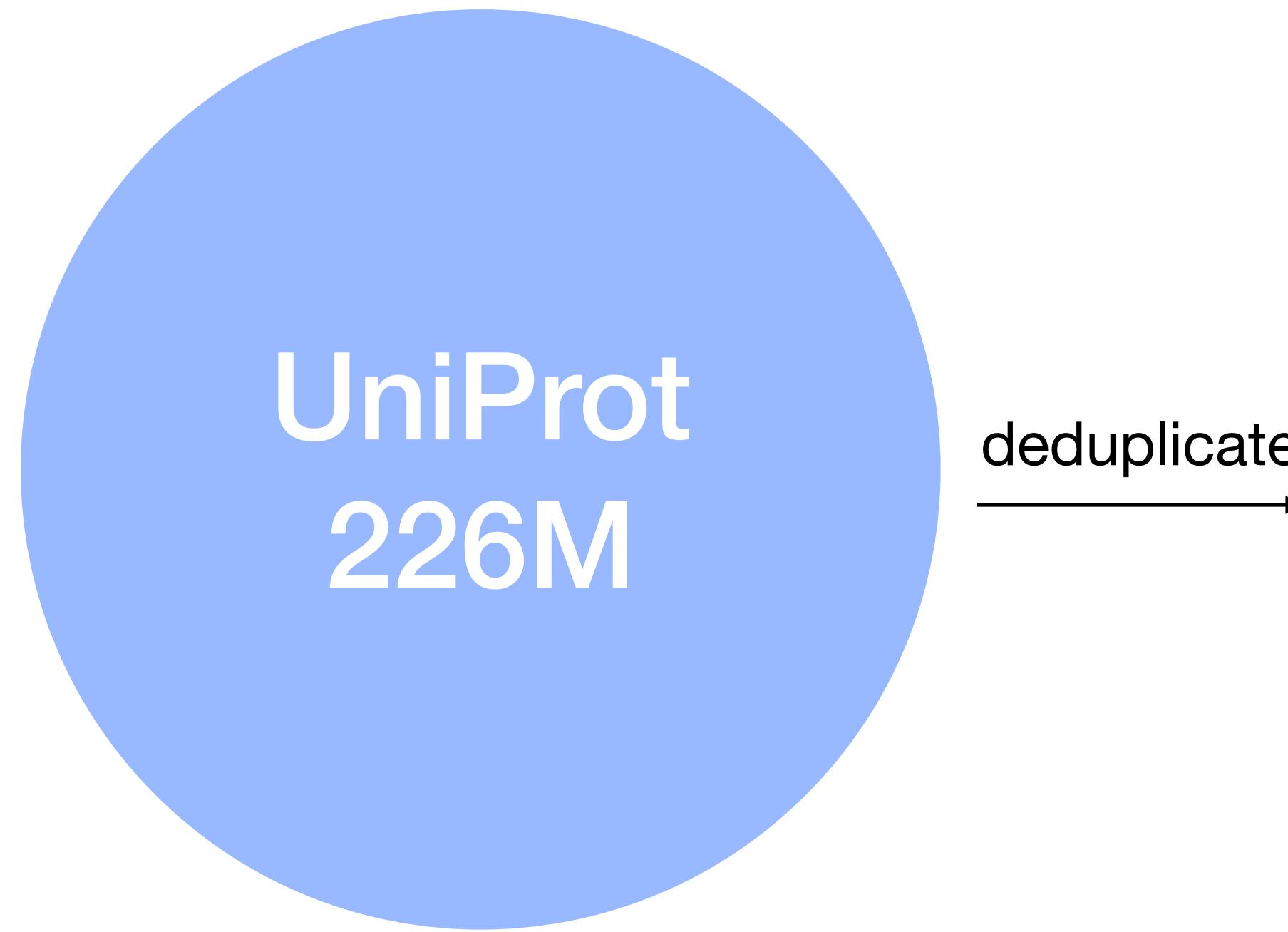
Still requires starting point!
Ignores mountains of protein data

We have access to large protein databases

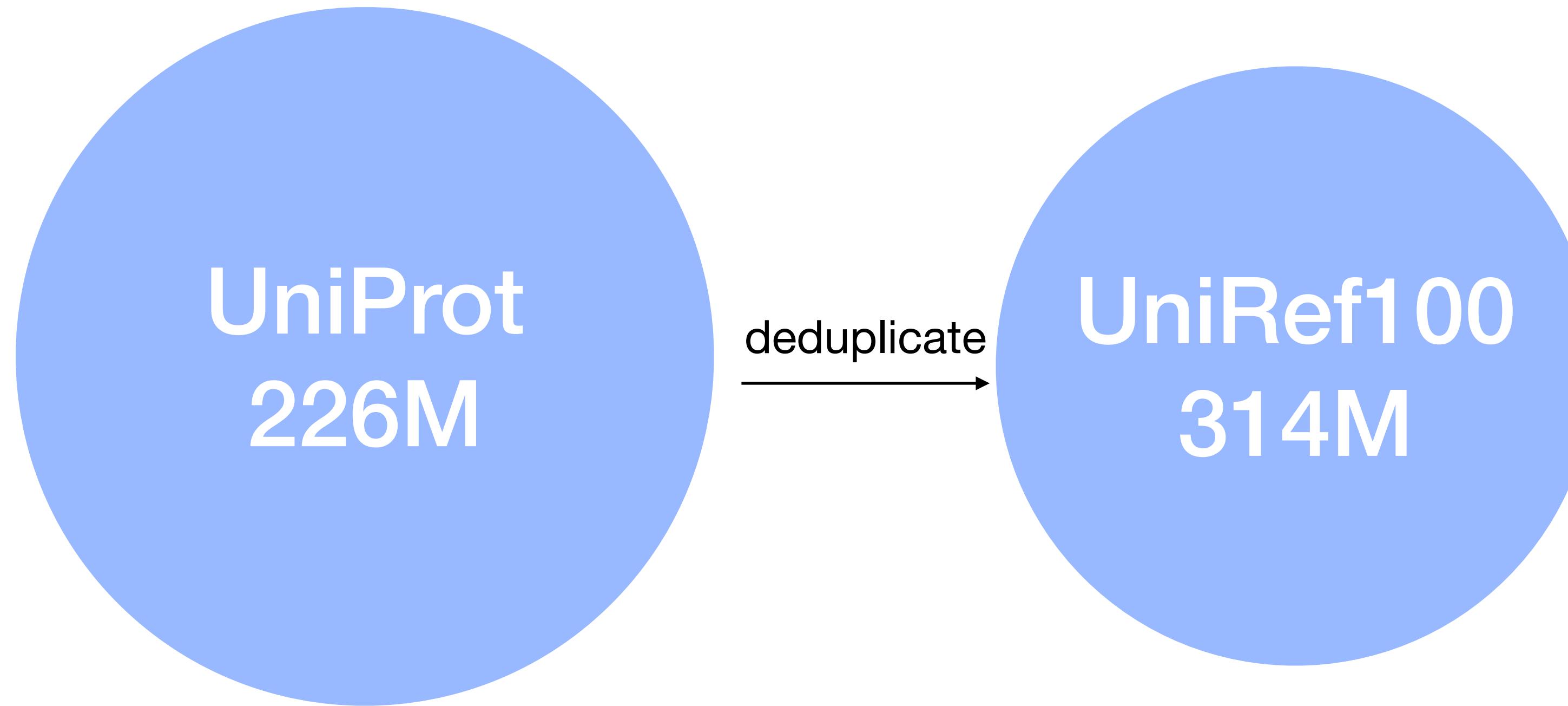
We have access to large protein databases



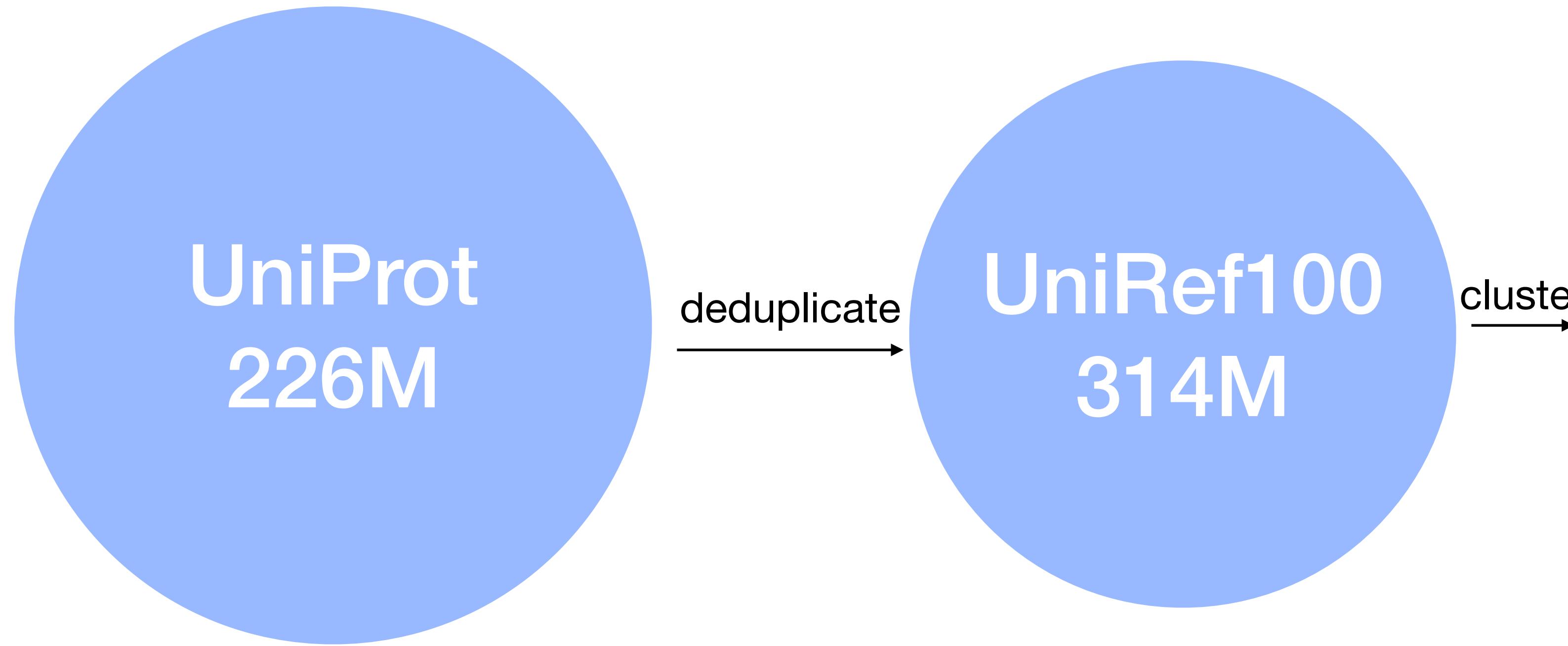
We have access to large protein databases



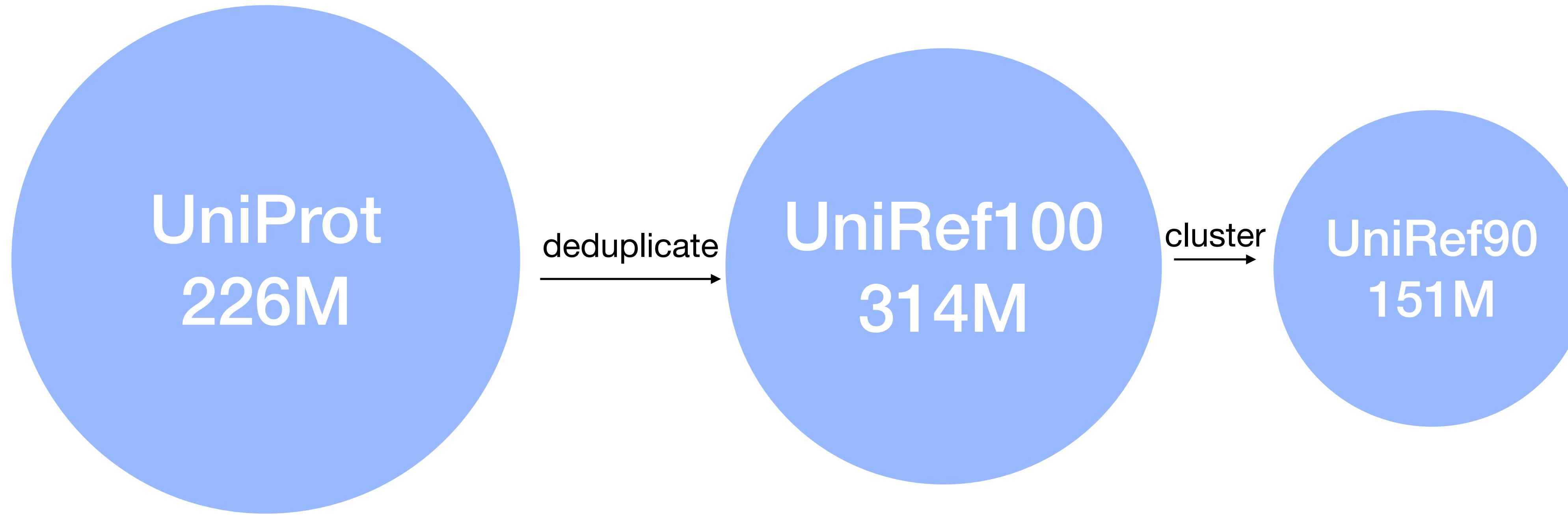
We have access to large protein databases



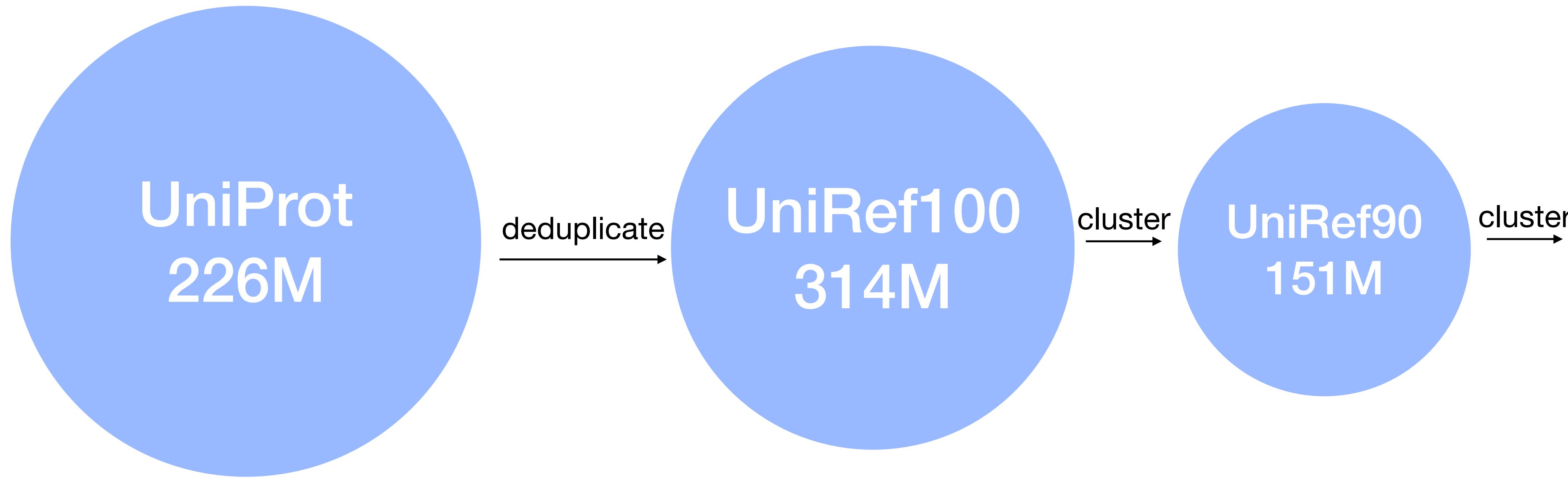
We have access to large protein databases



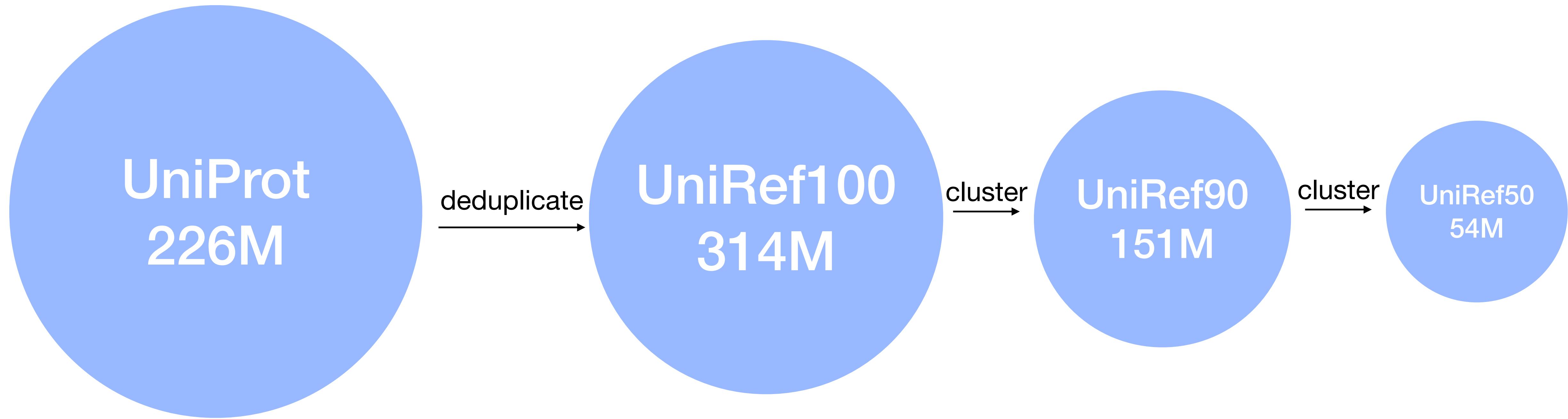
We have access to large protein databases



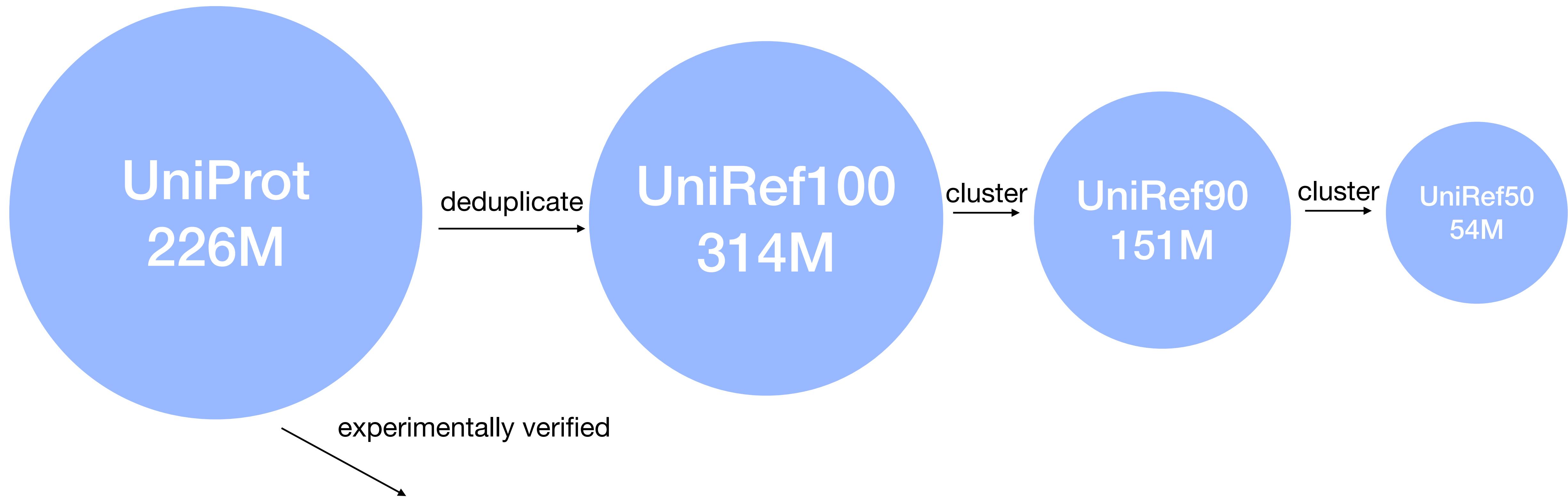
We have access to large protein databases



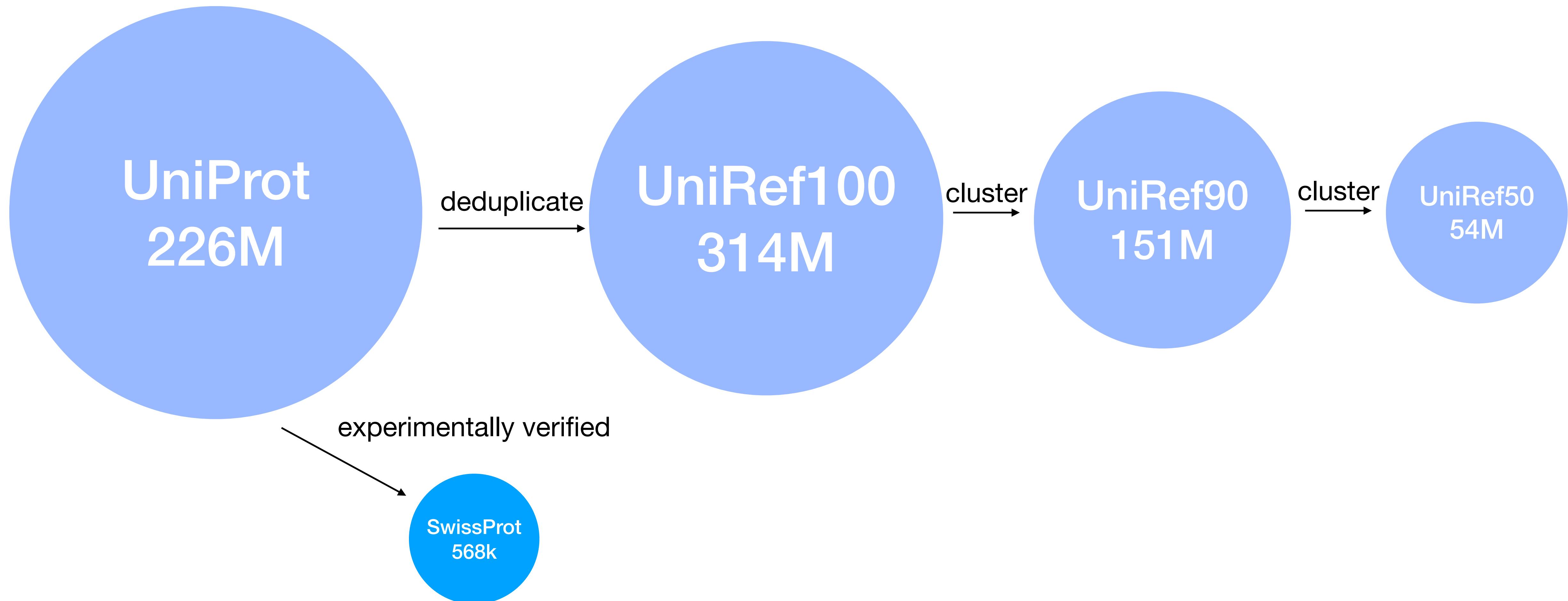
We have access to large protein databases



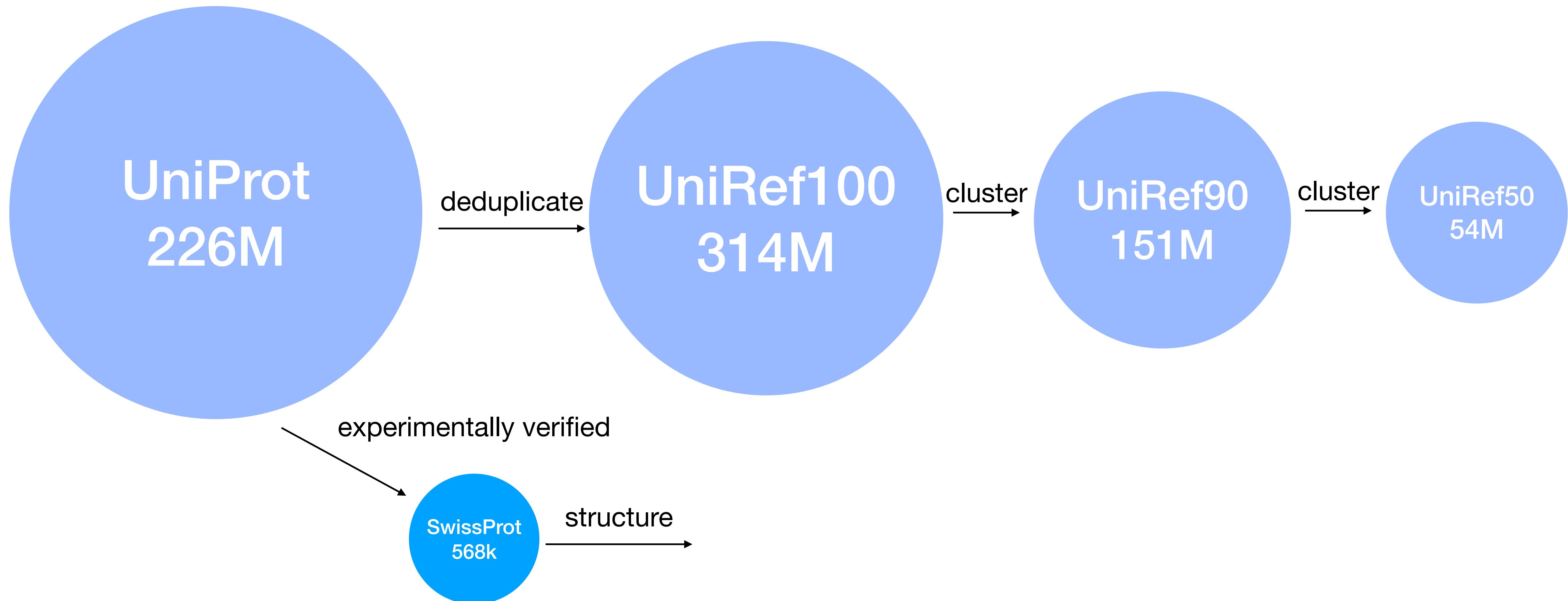
We have access to large protein databases



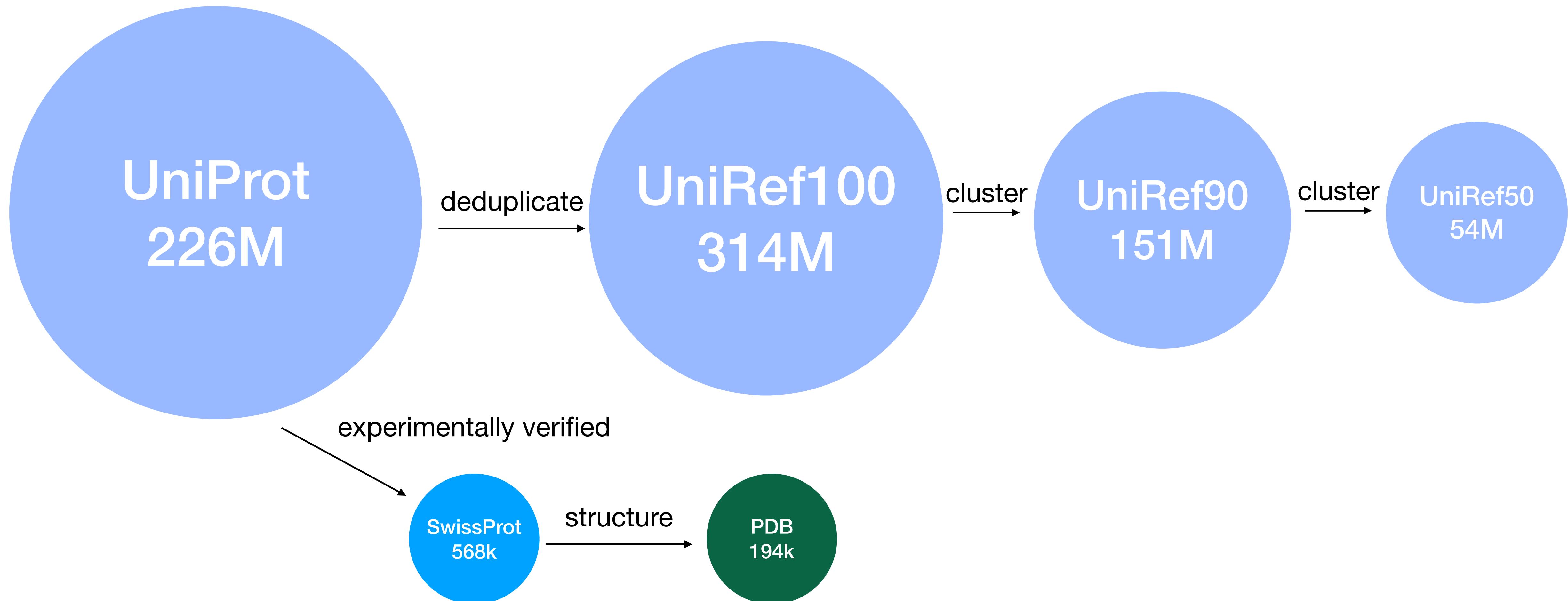
We have access to large protein databases



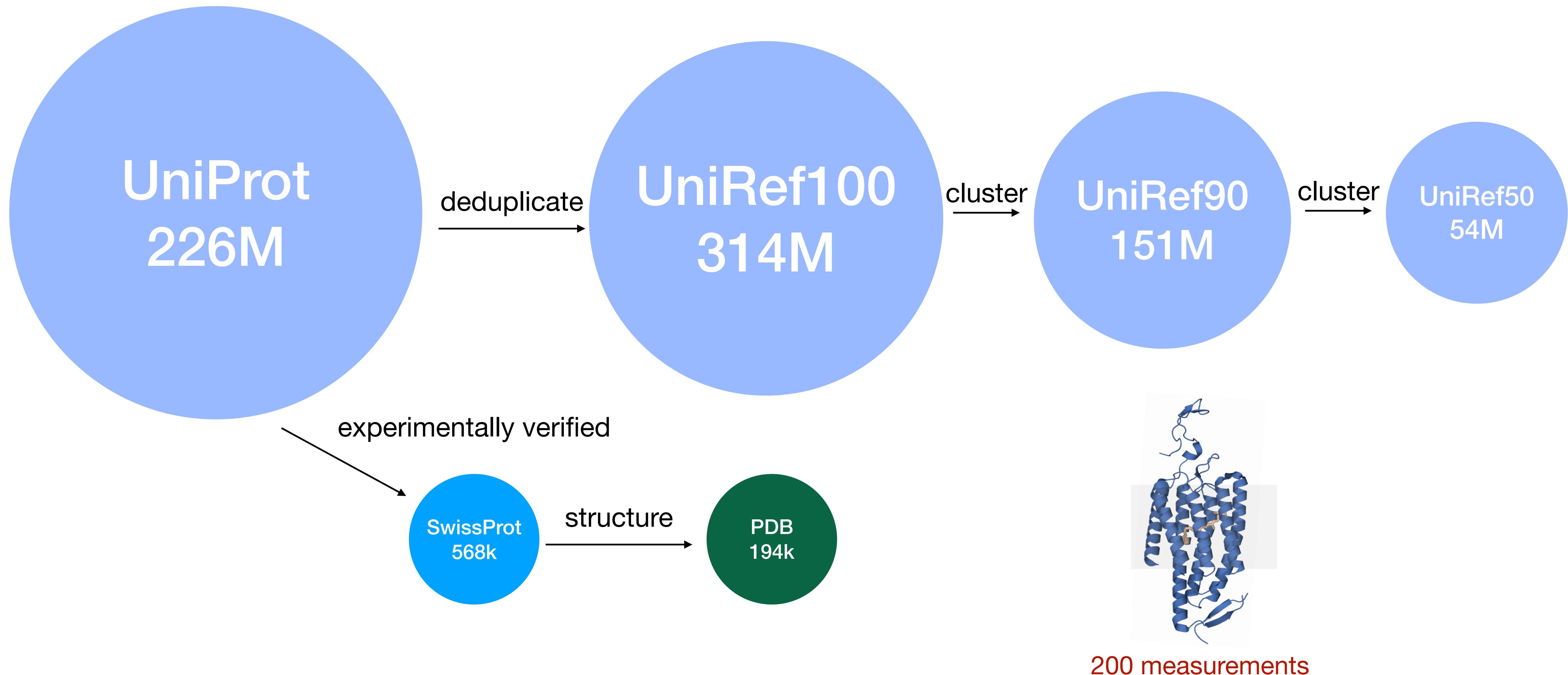
We have access to large protein databases



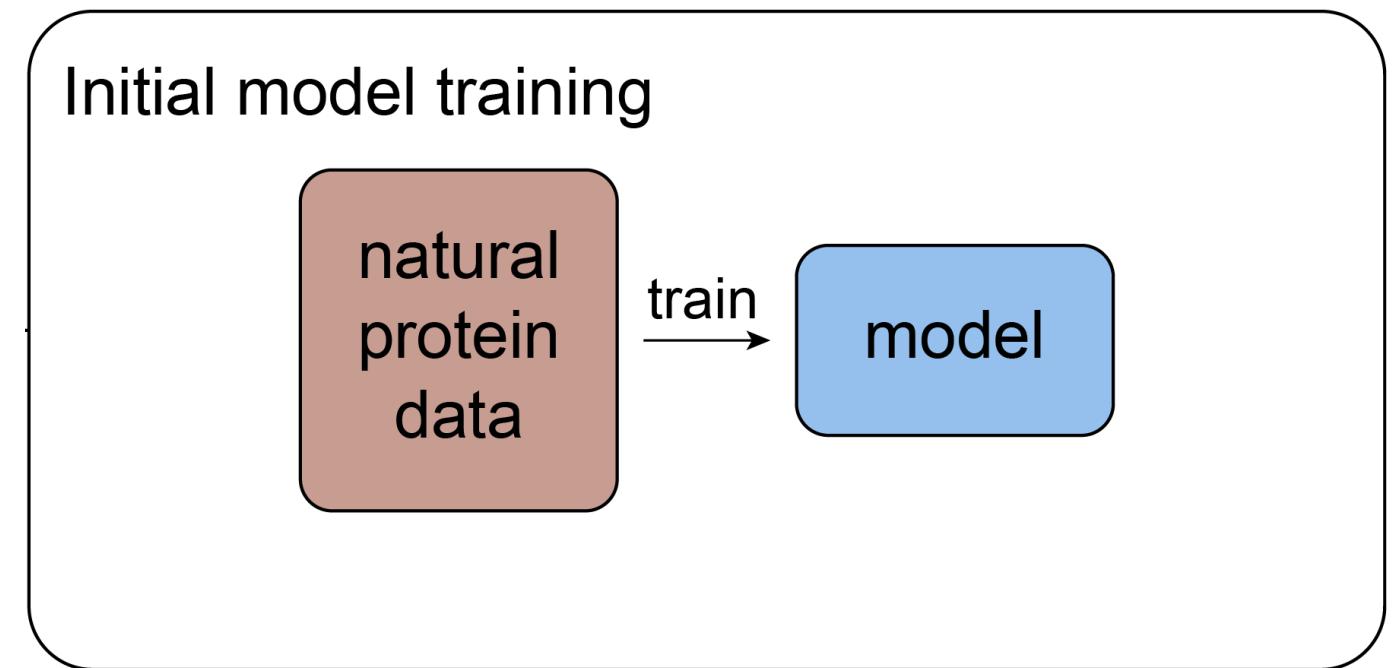
We have access to large protein databases



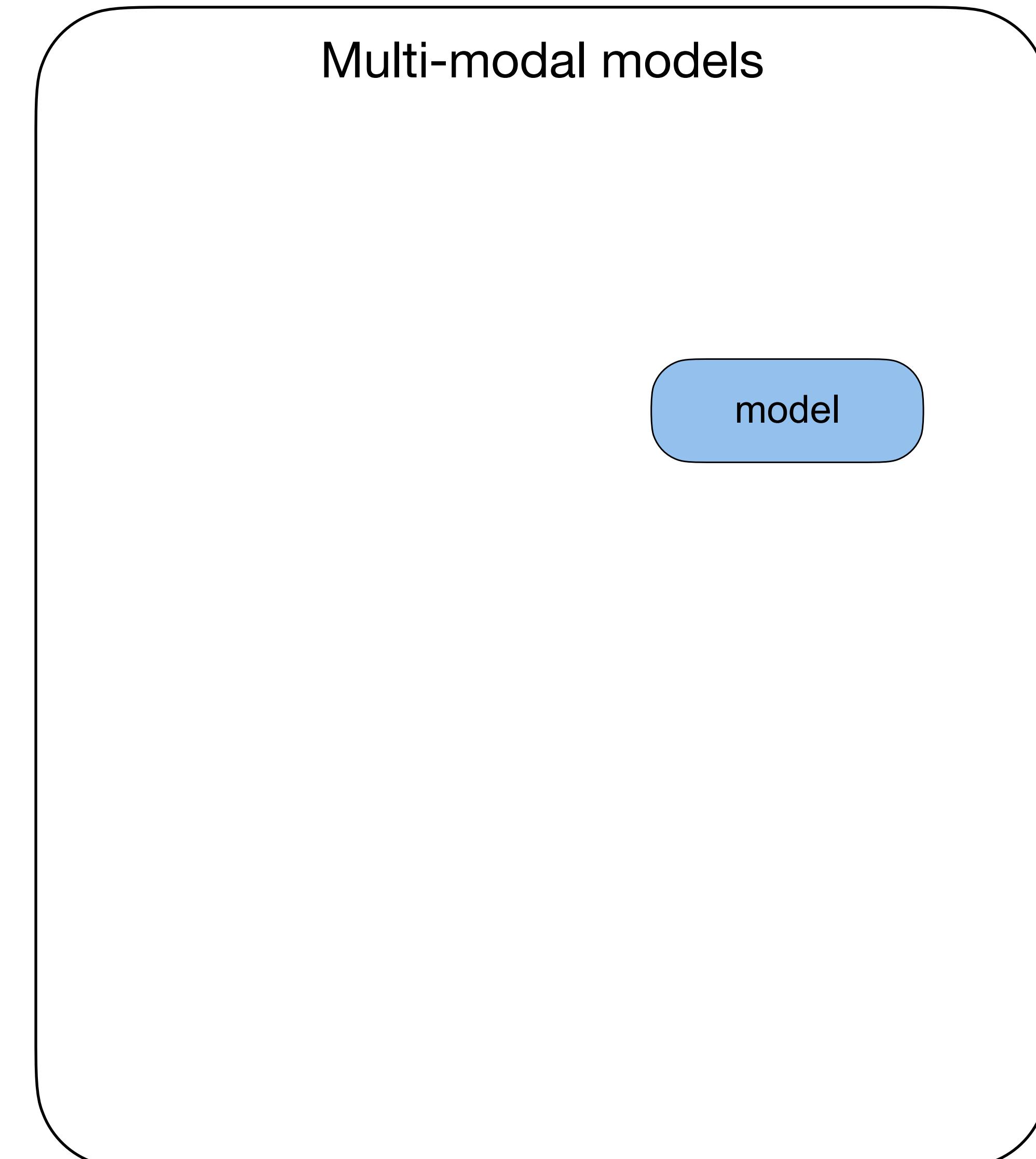
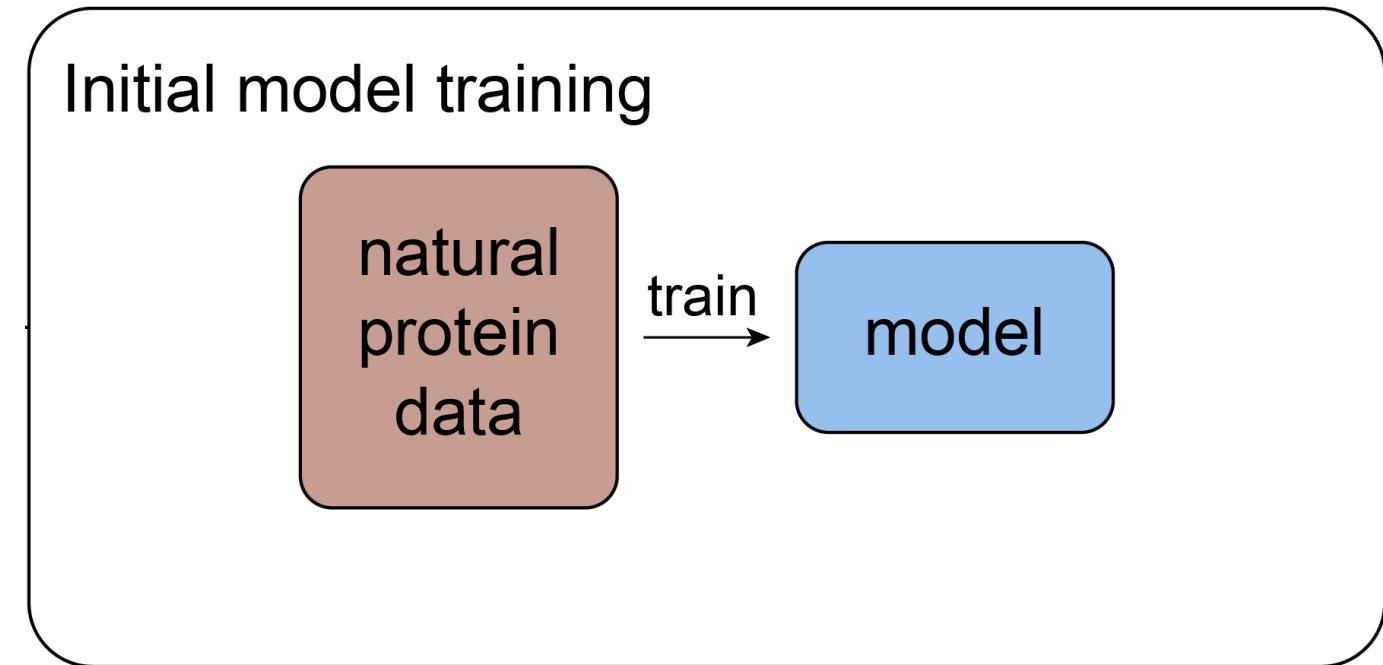
We have access to large protein databases



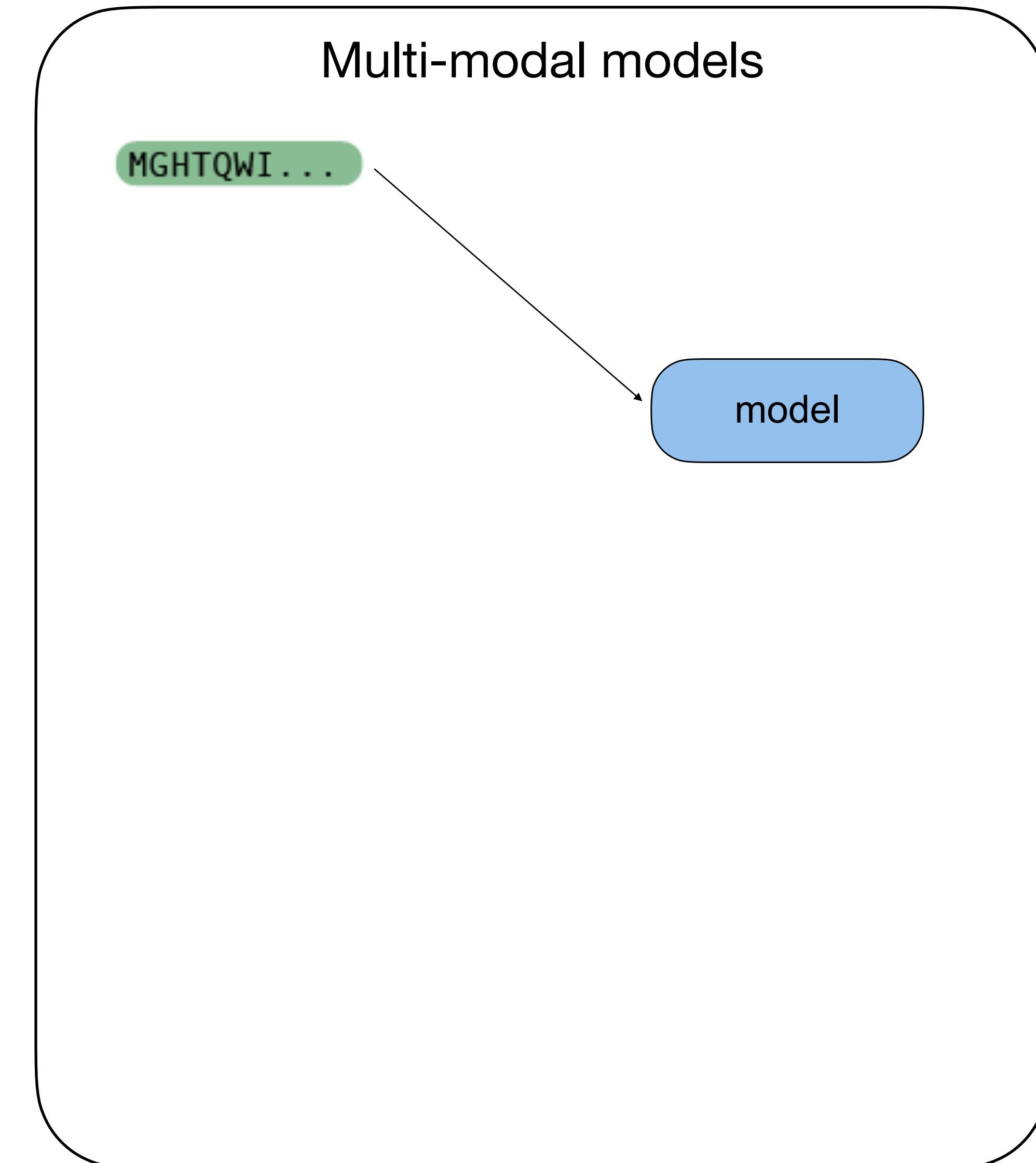
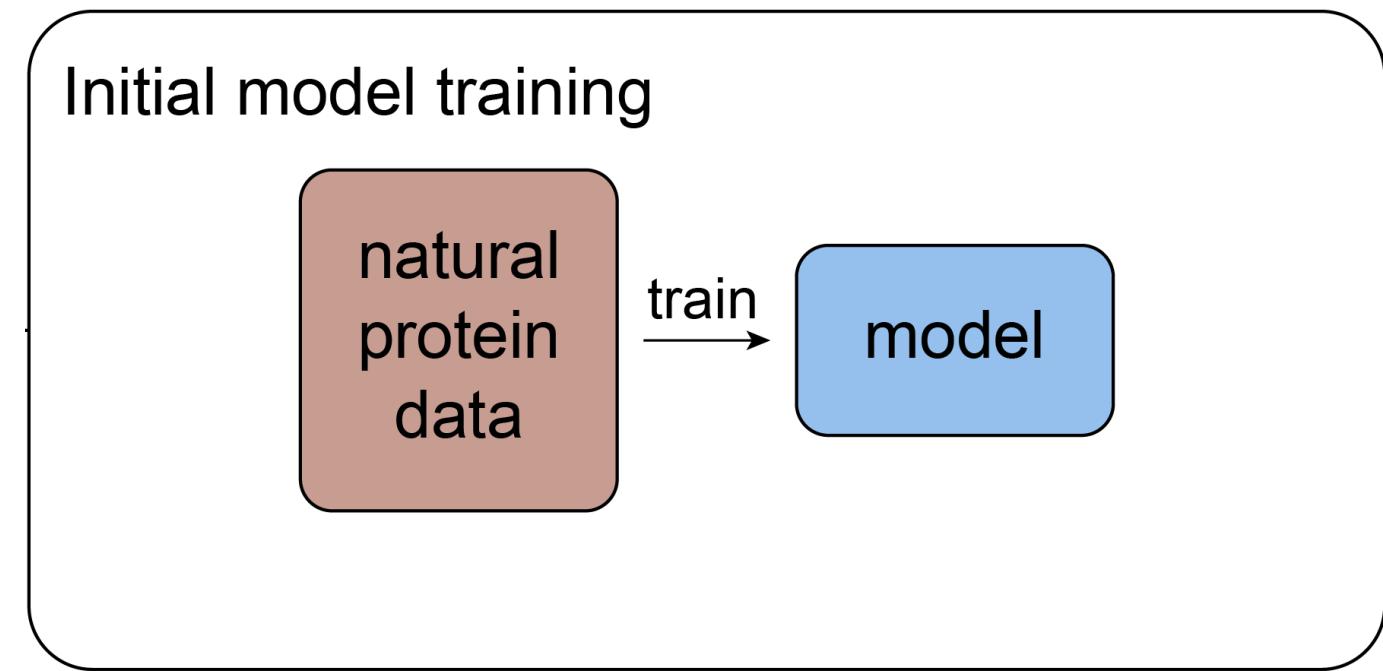
Use multiple data modalities to design proteins



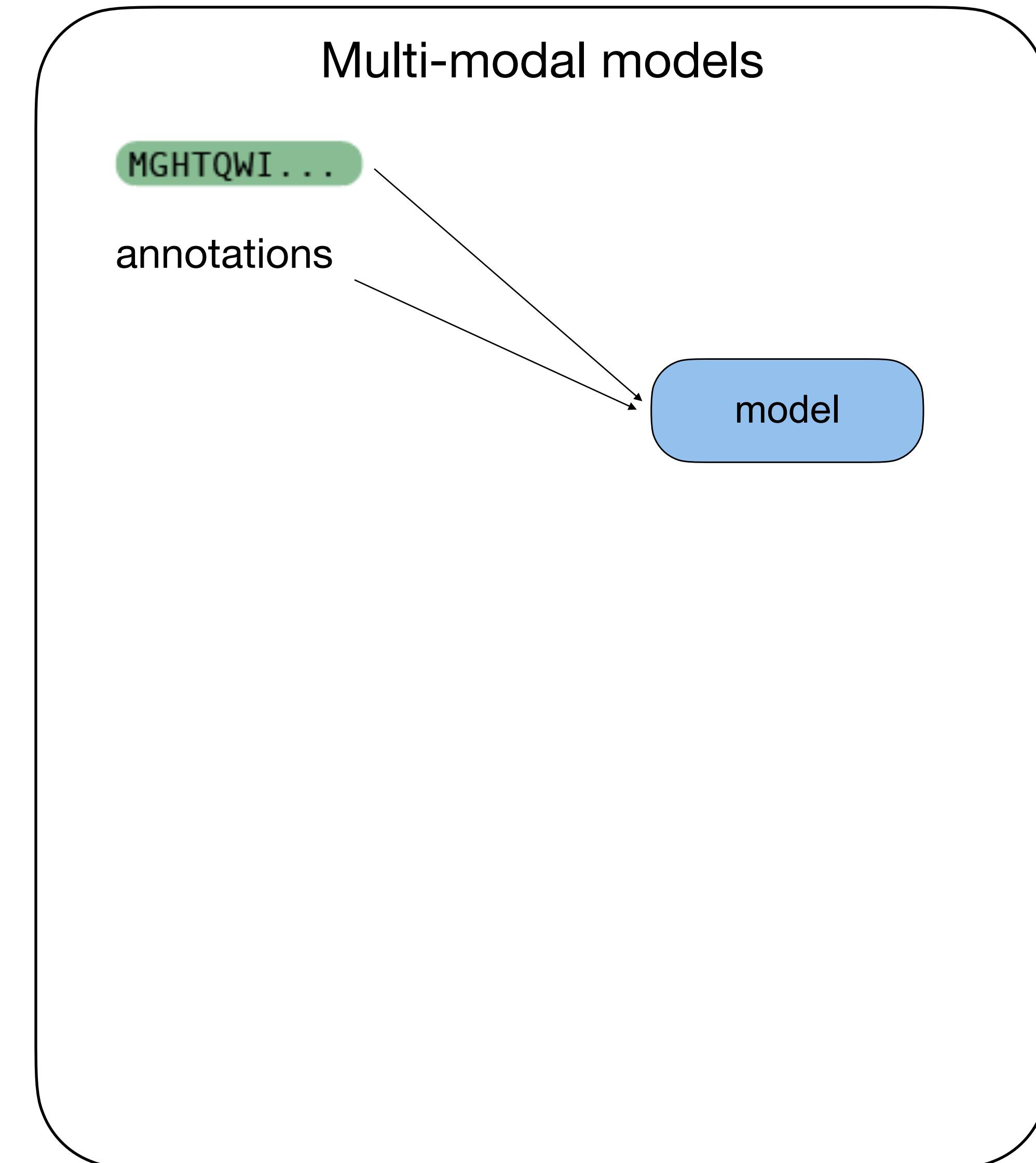
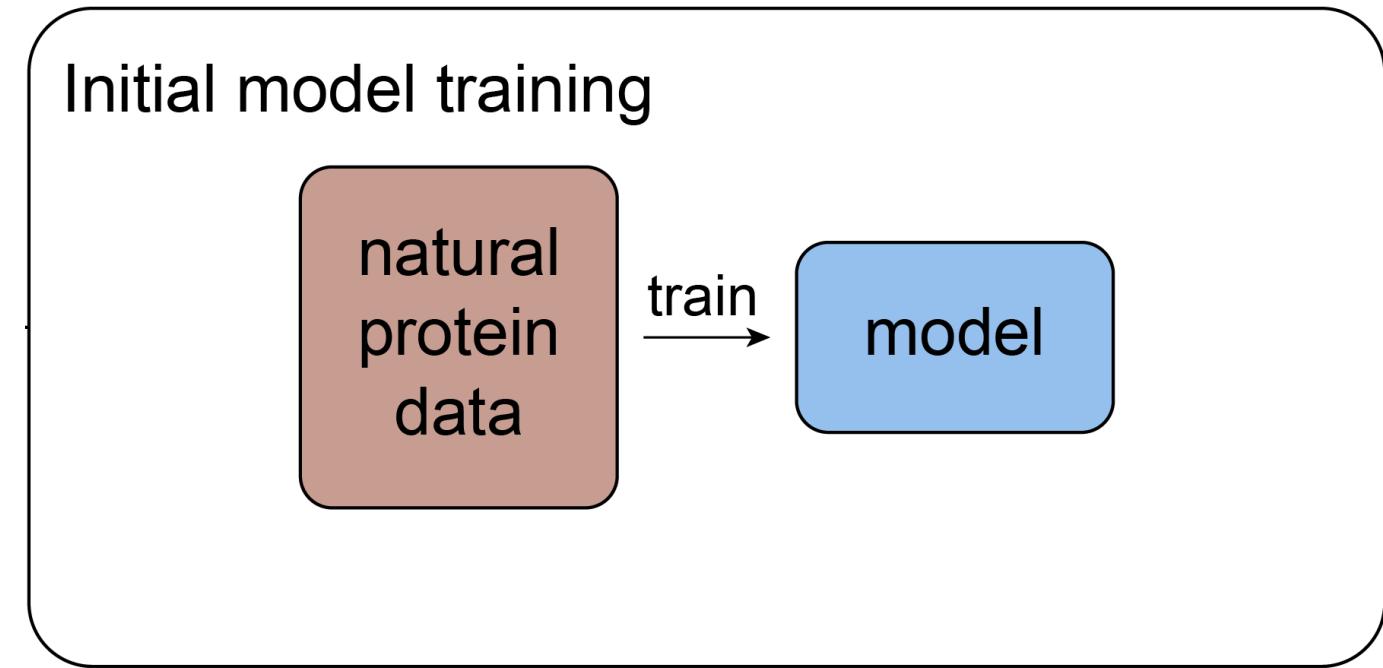
Use multiple data modalities to design proteins



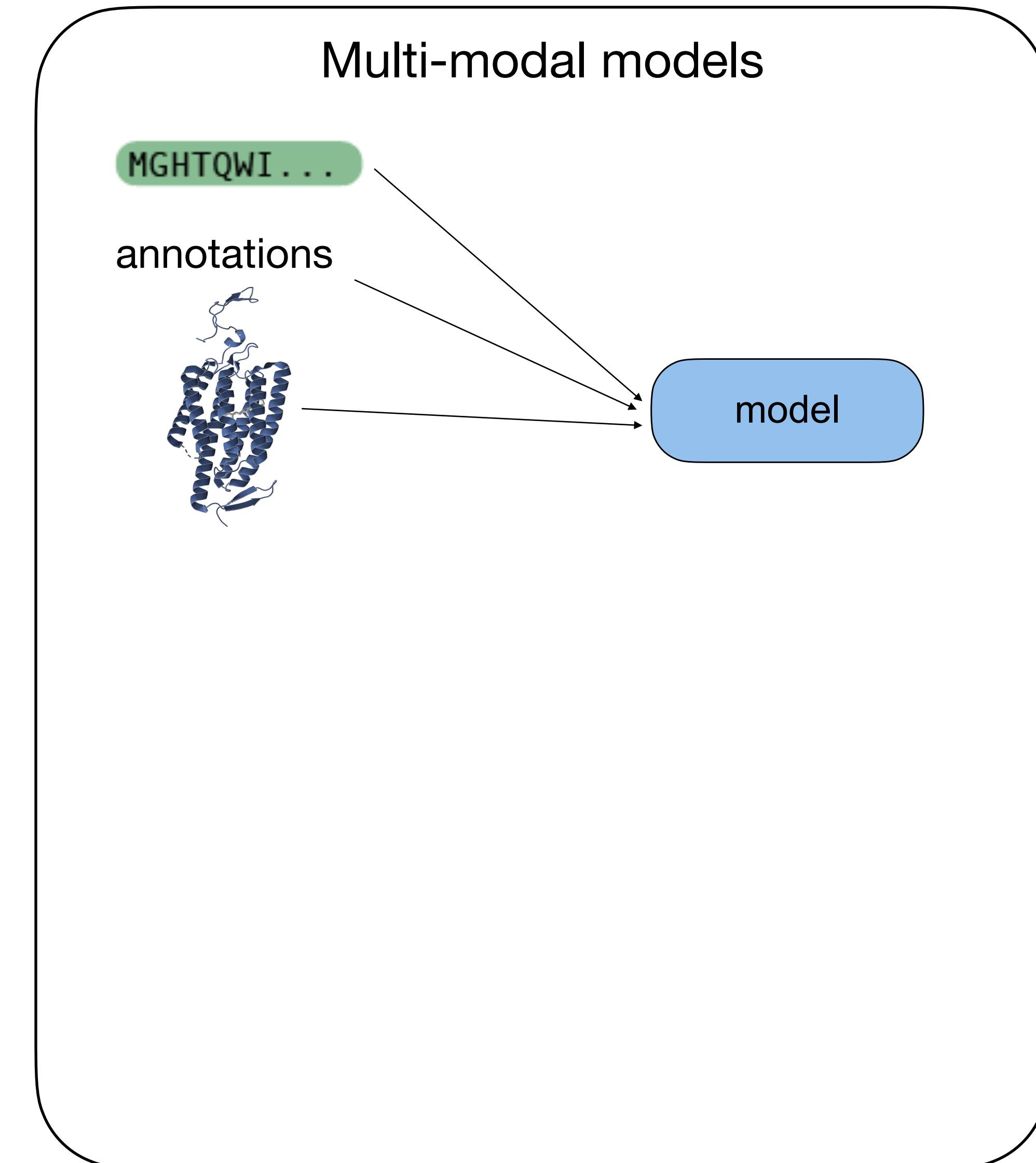
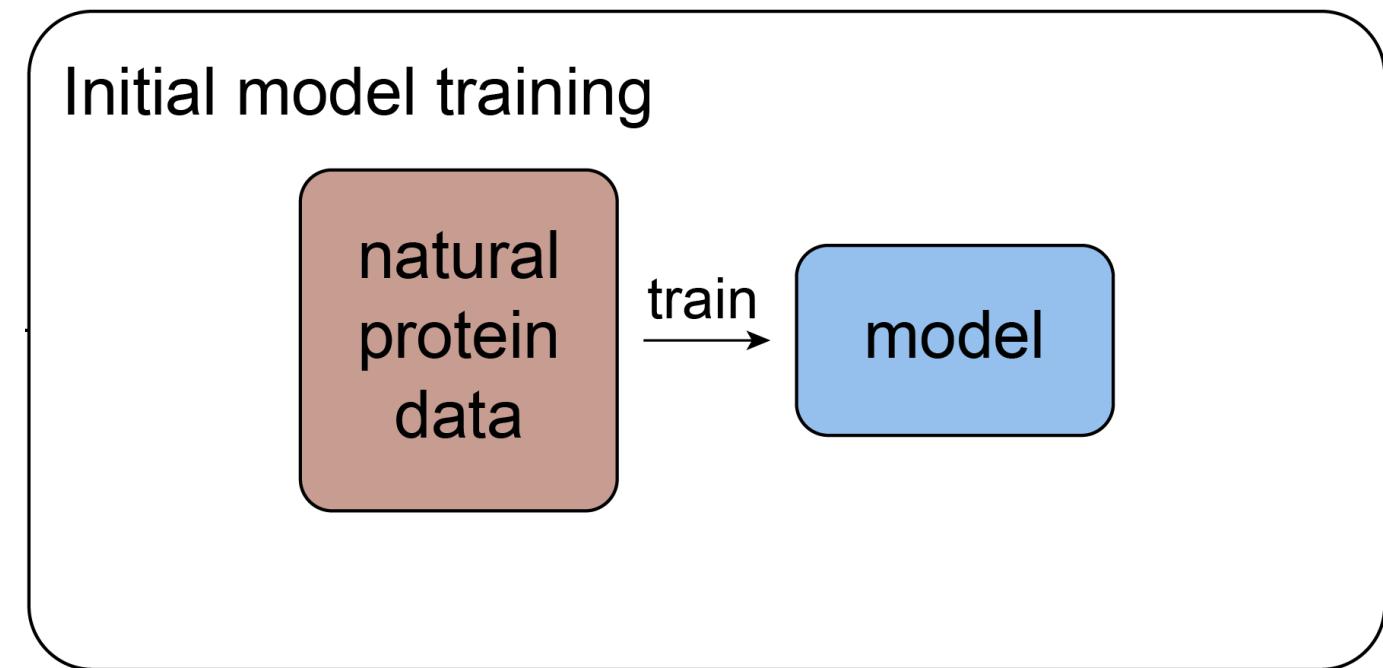
Use multiple data modalities to design proteins



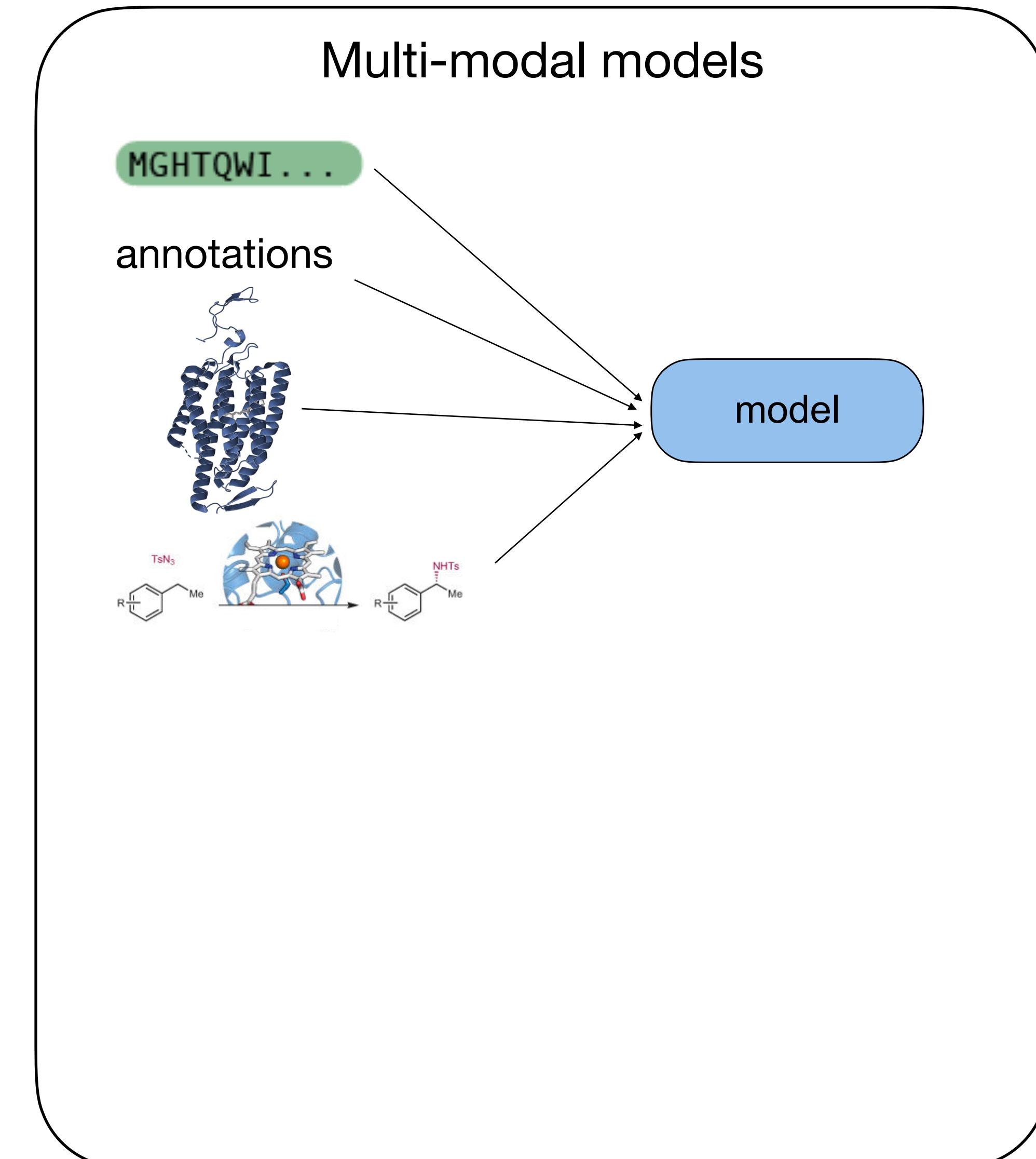
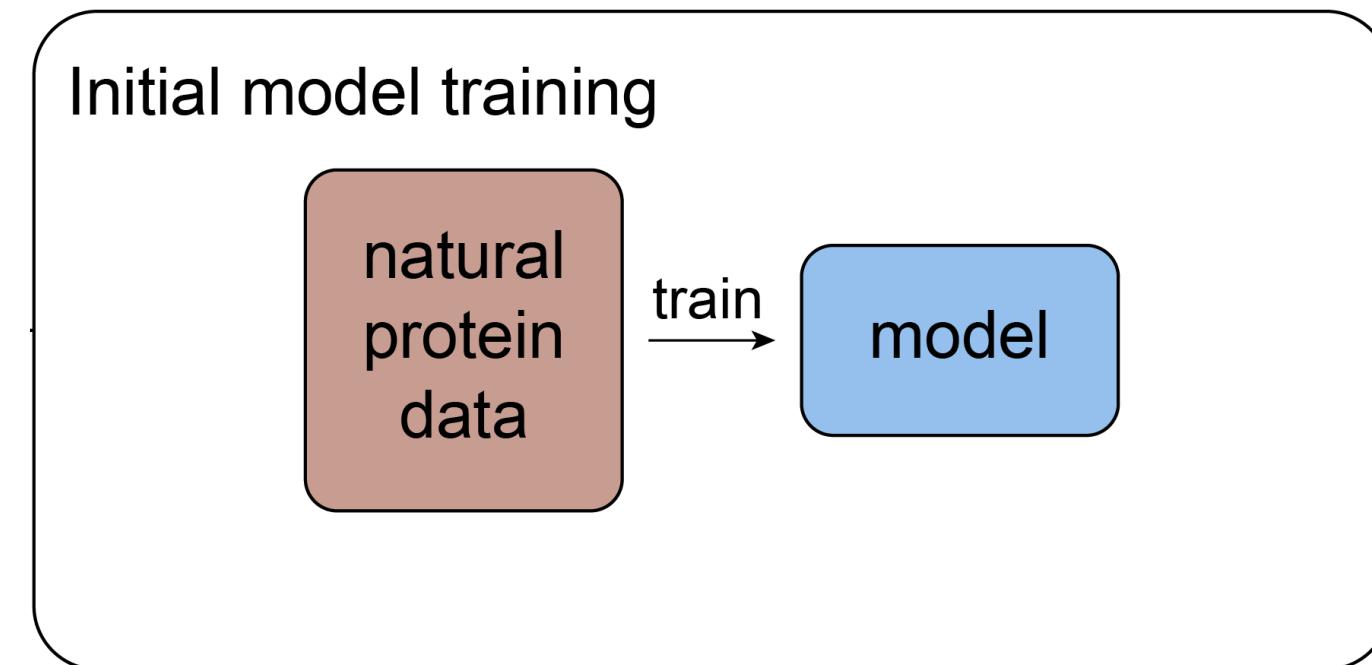
Use multiple data modalities to design proteins



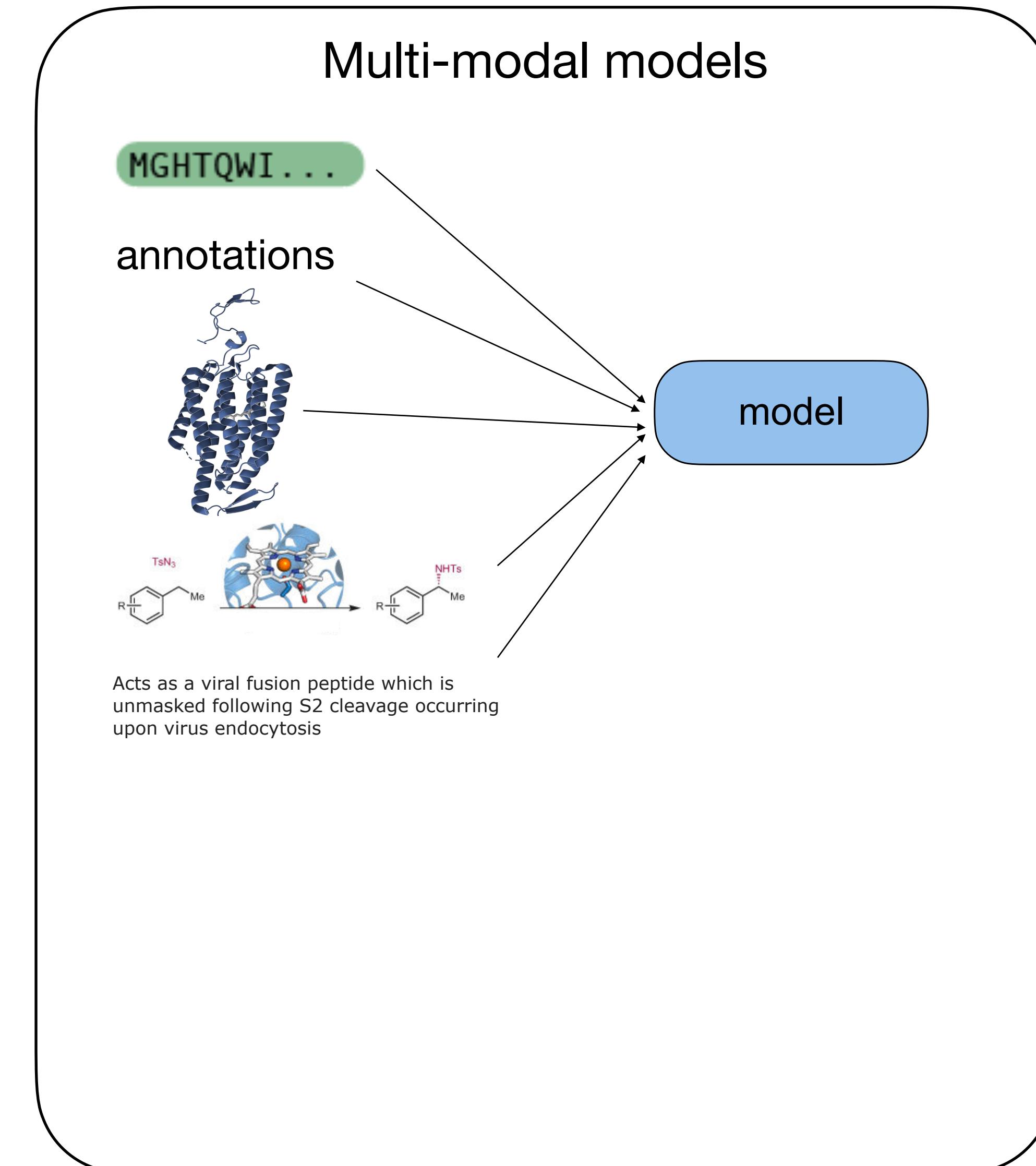
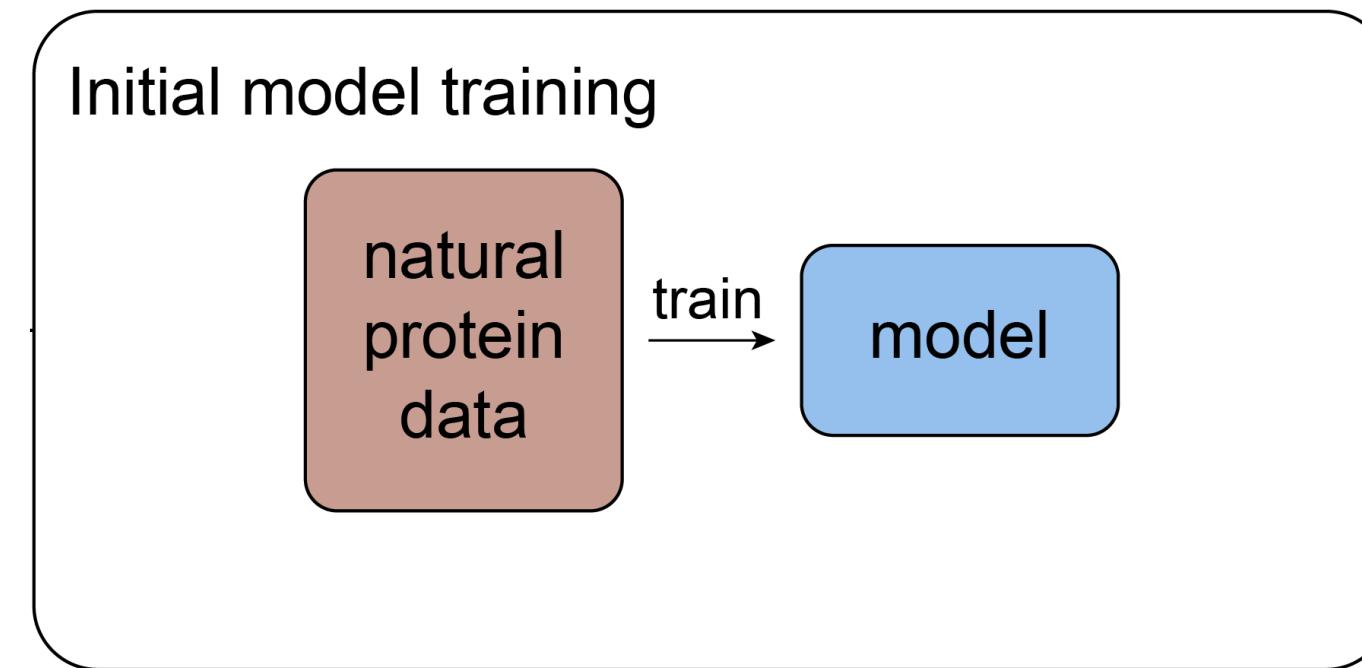
Use multiple data modalities to design proteins



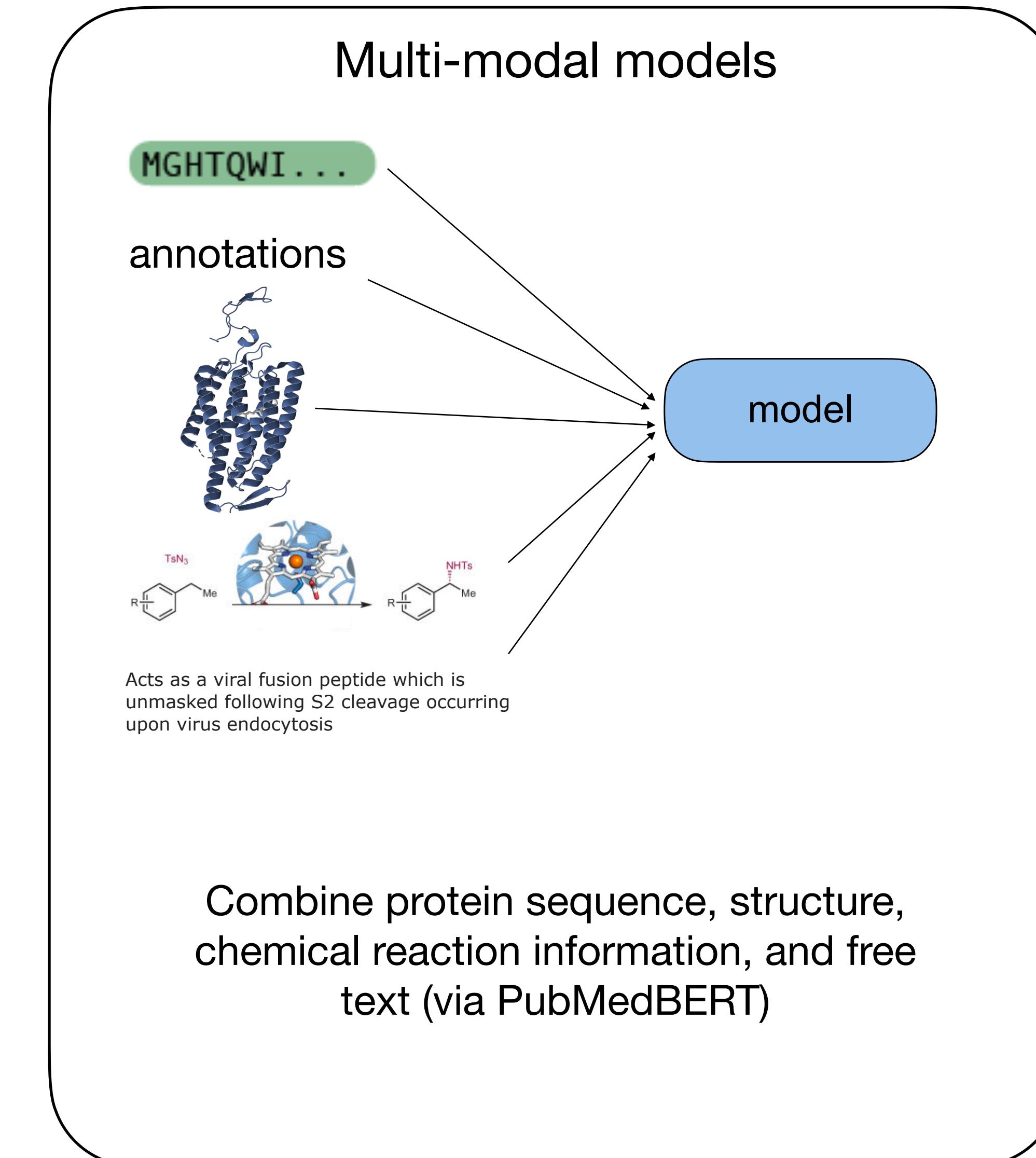
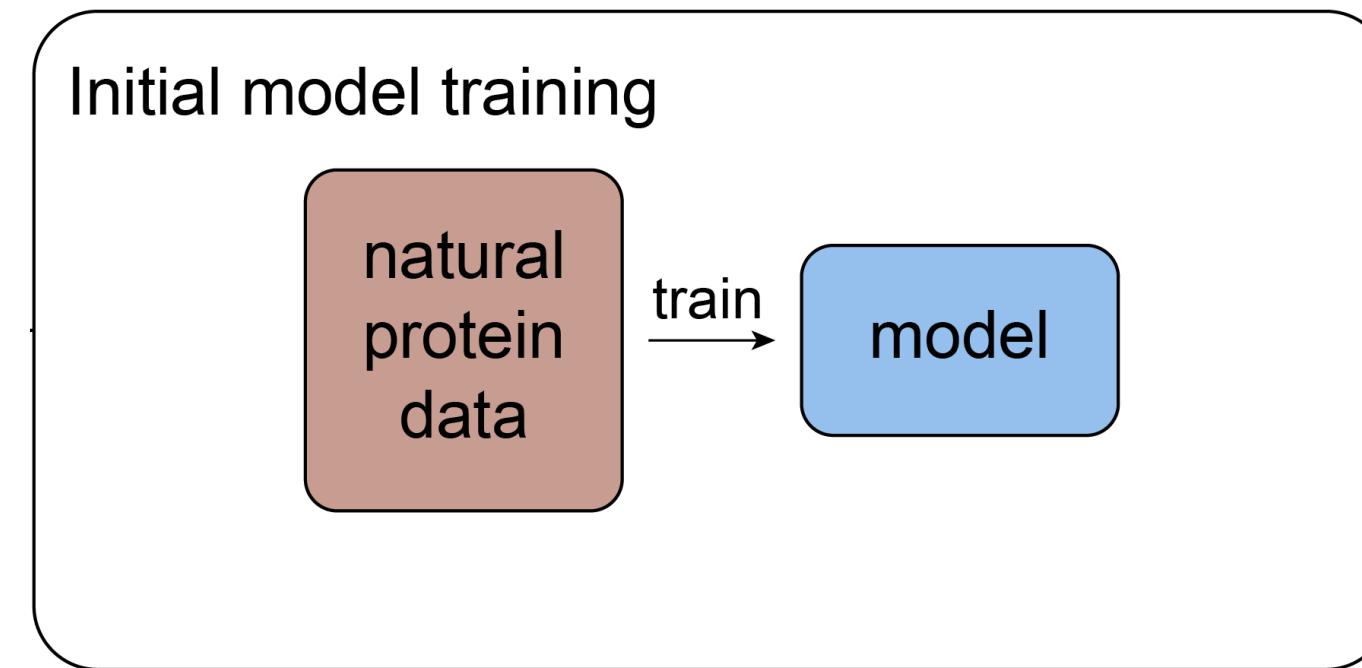
Use multiple data modalities to design proteins



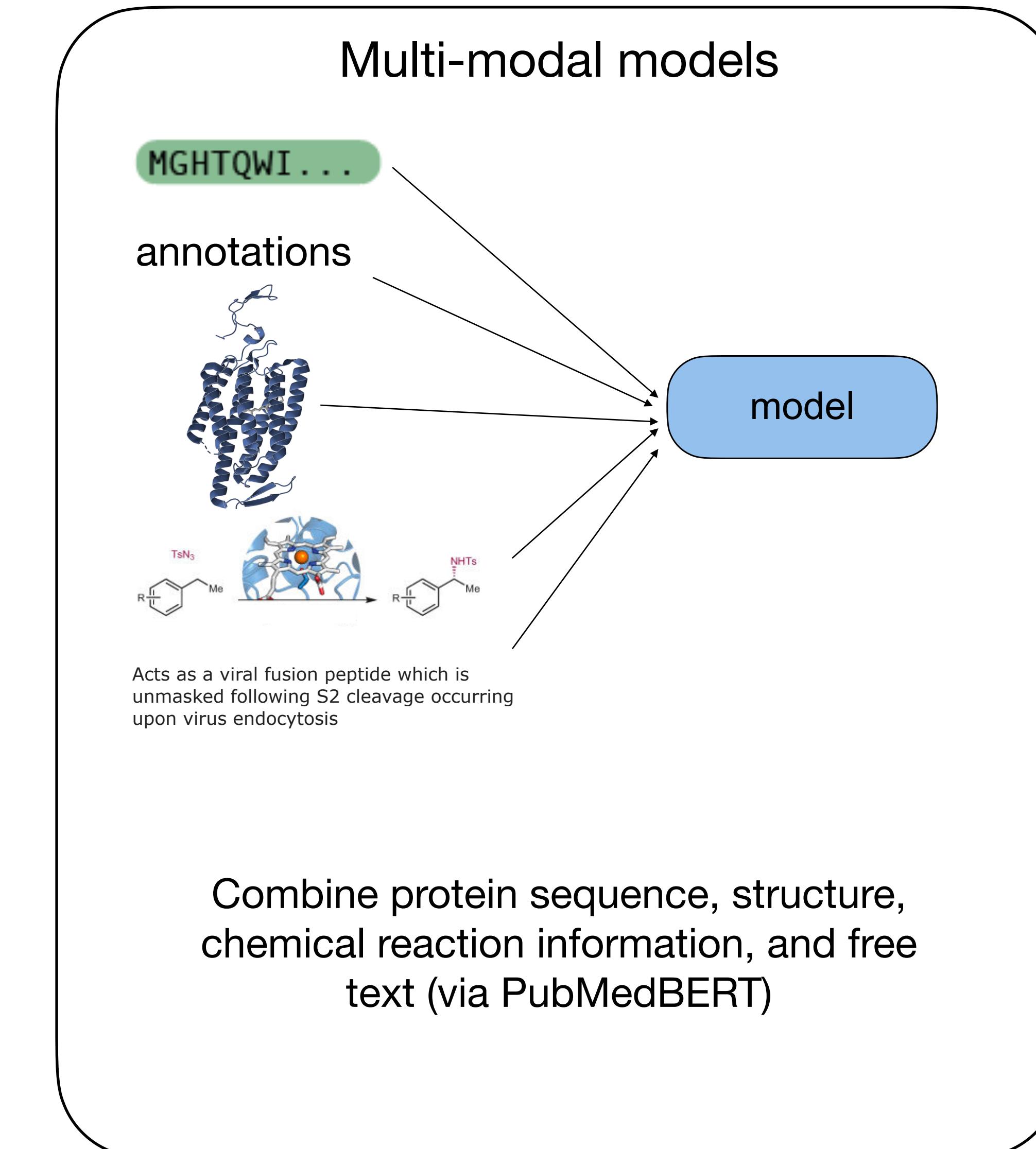
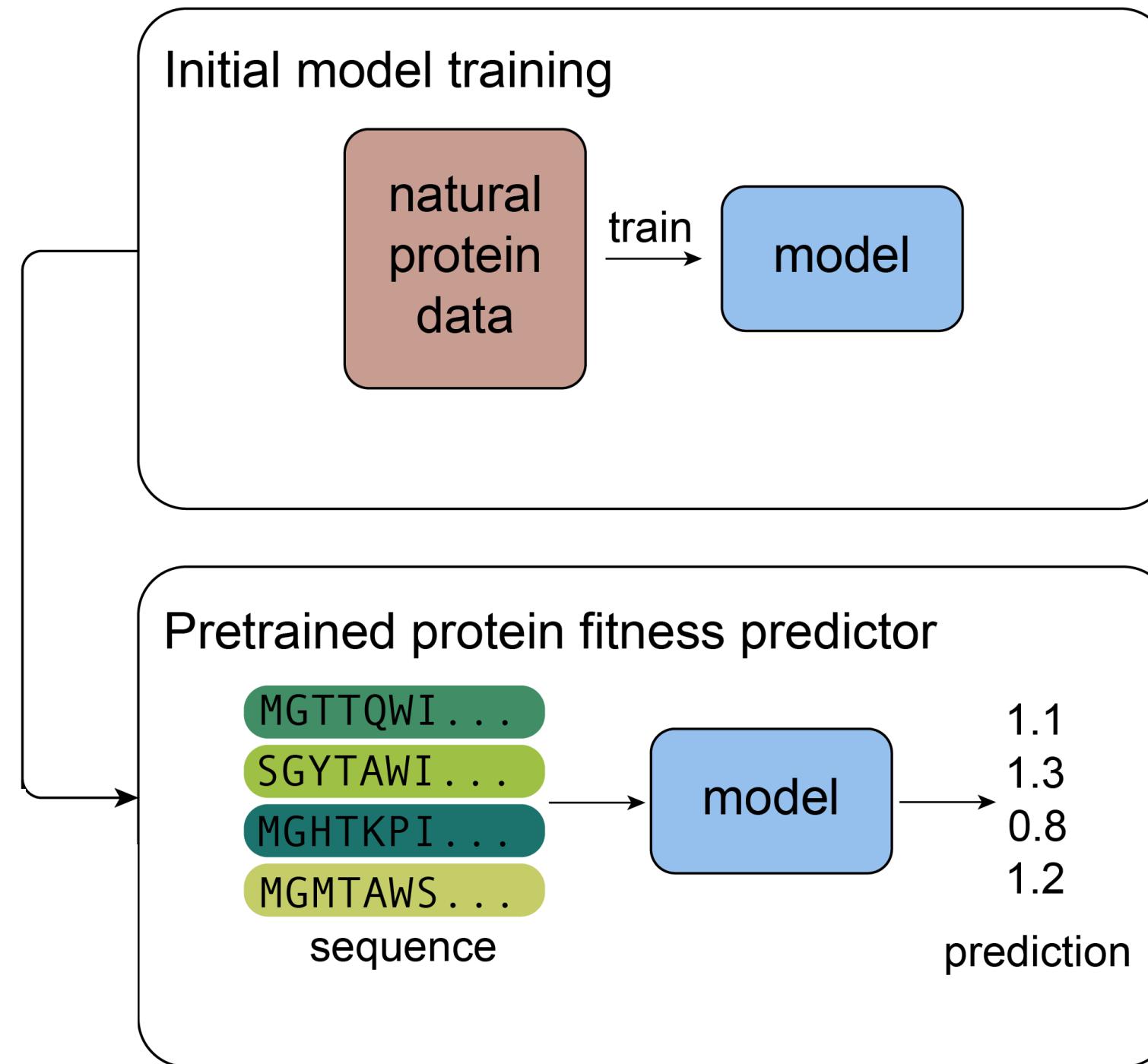
Use multiple data modalities to design proteins



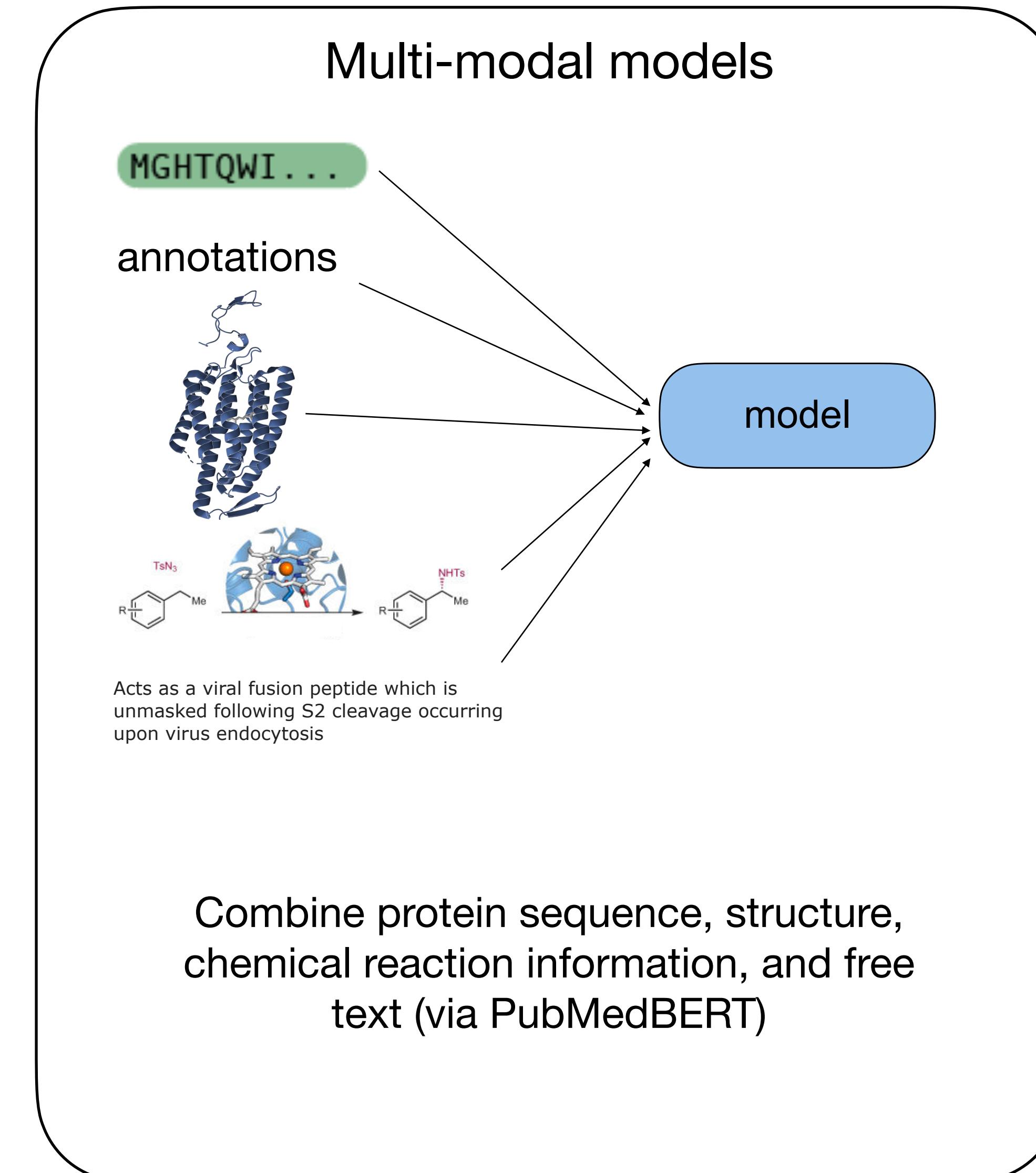
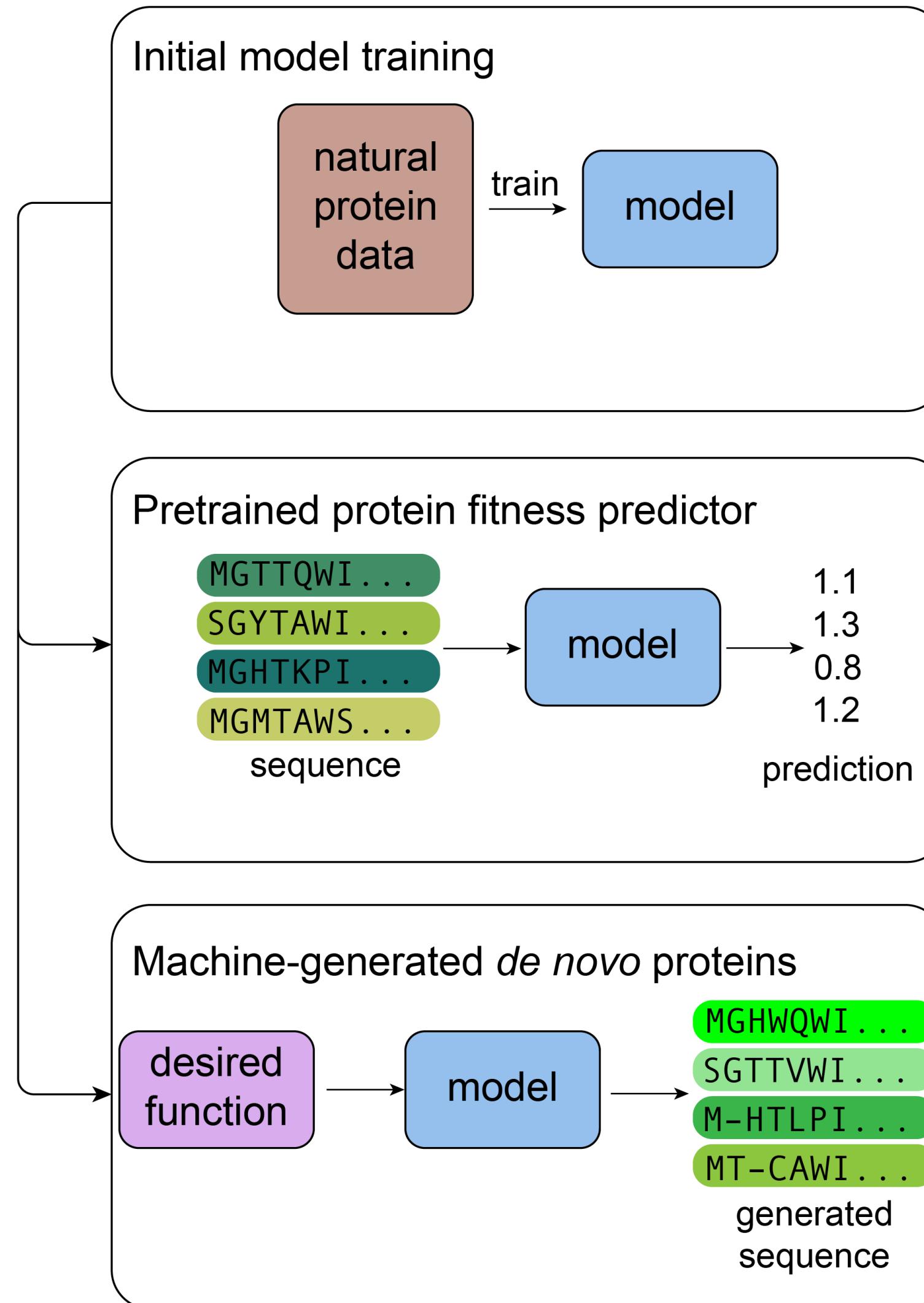
Use multiple data modalities to design proteins



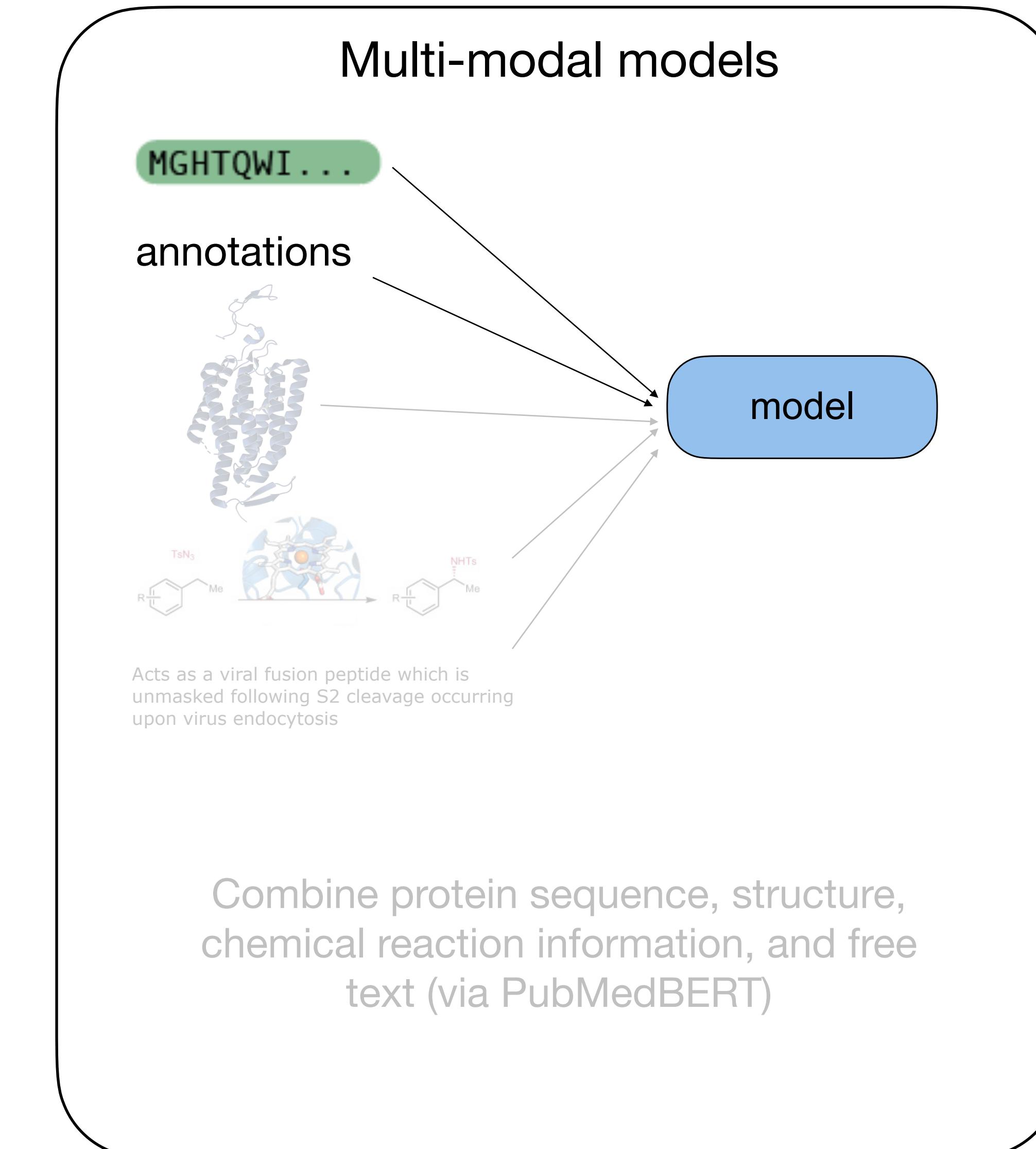
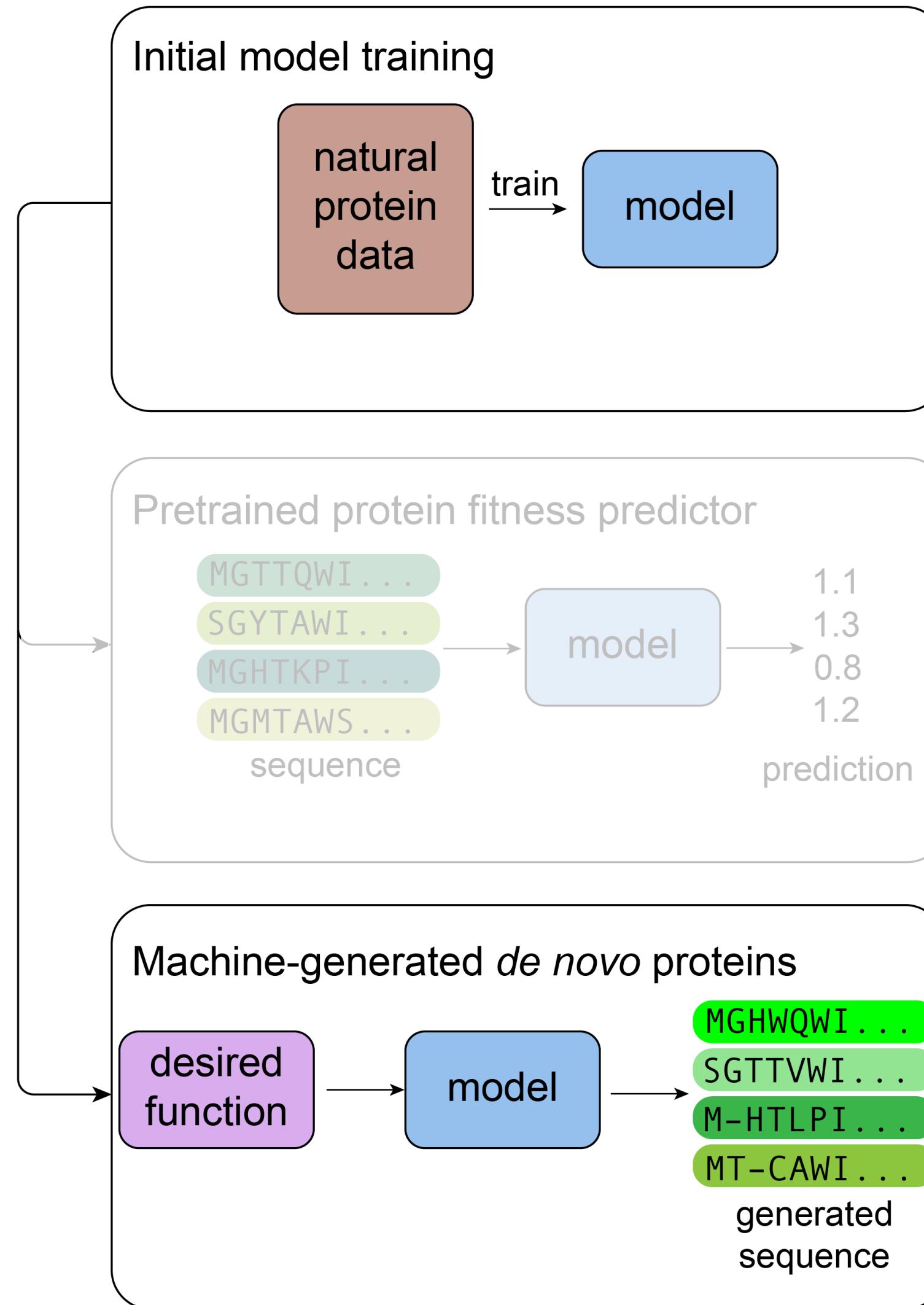
Use multiple data modalities to design proteins



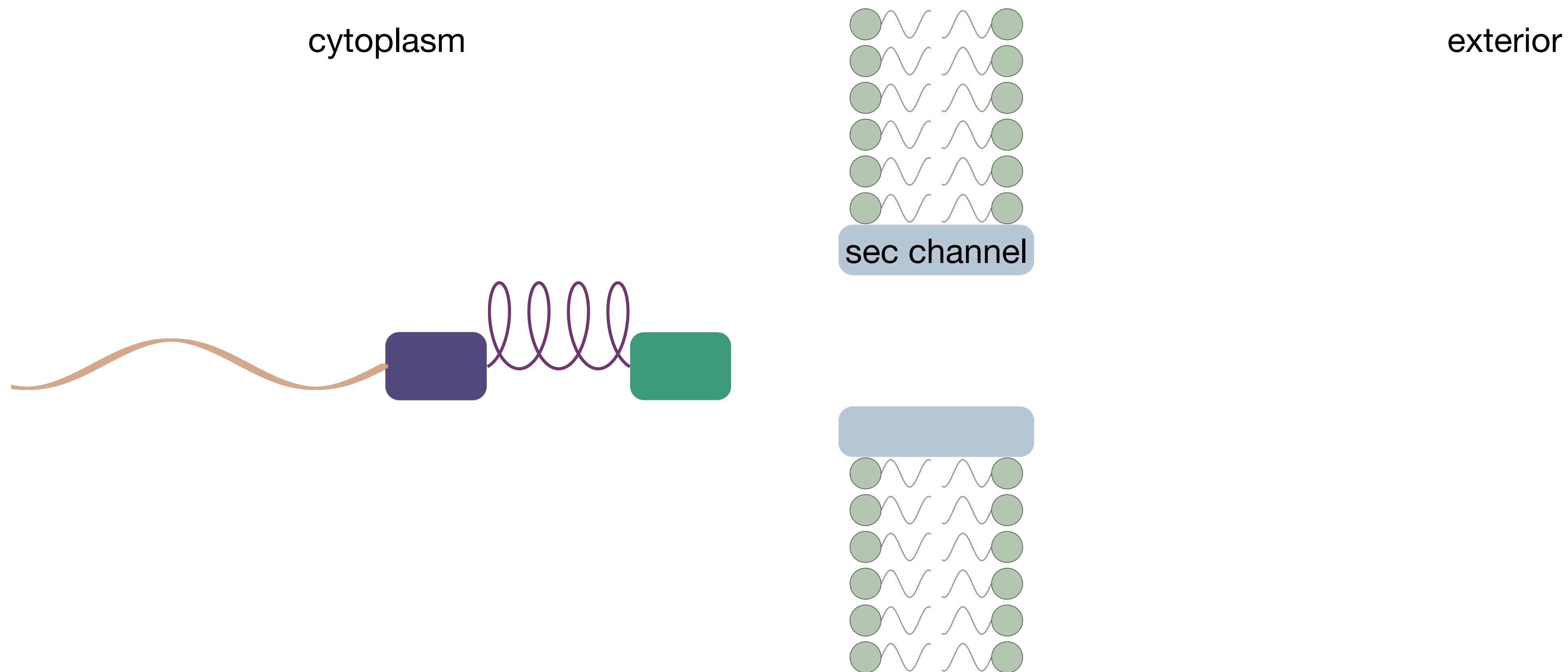
Use multiple data modalities to design proteins



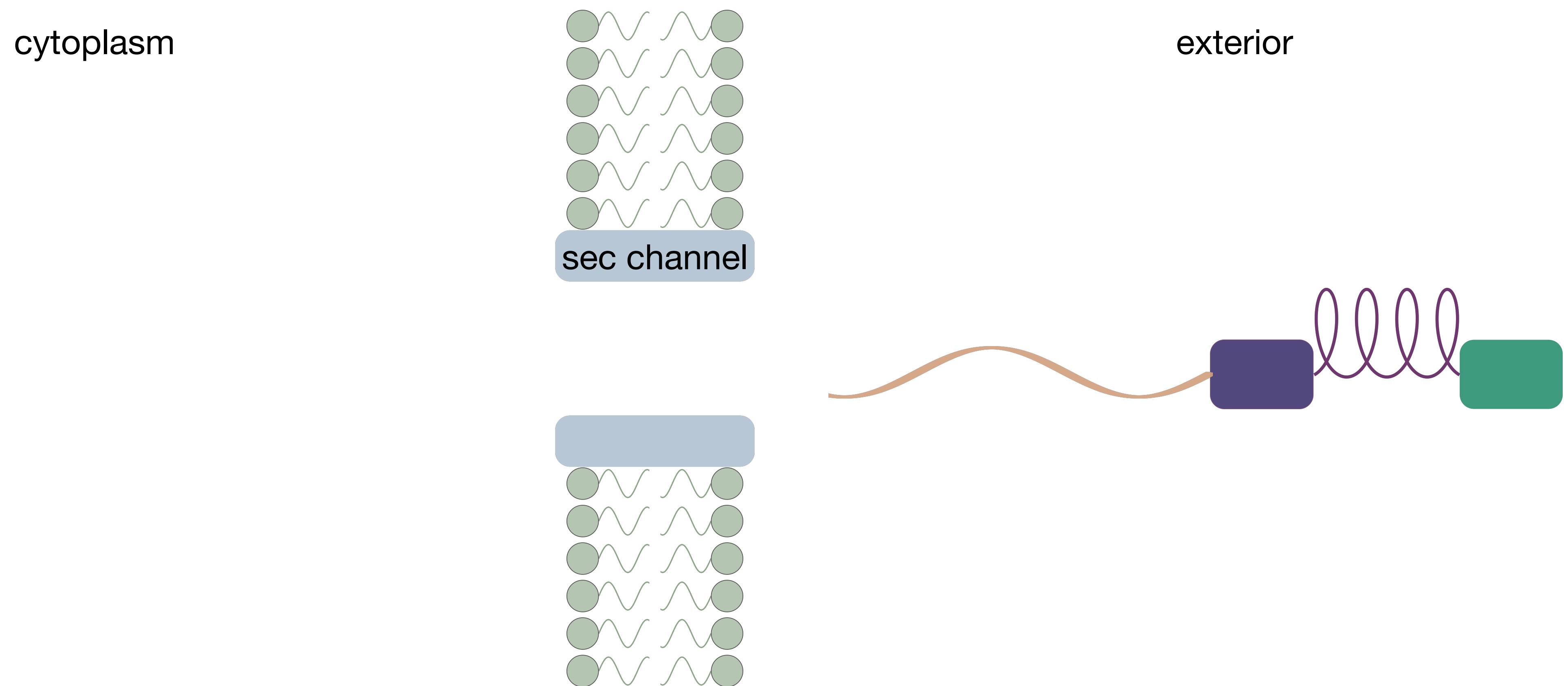
Use multiple data modalities to design proteins



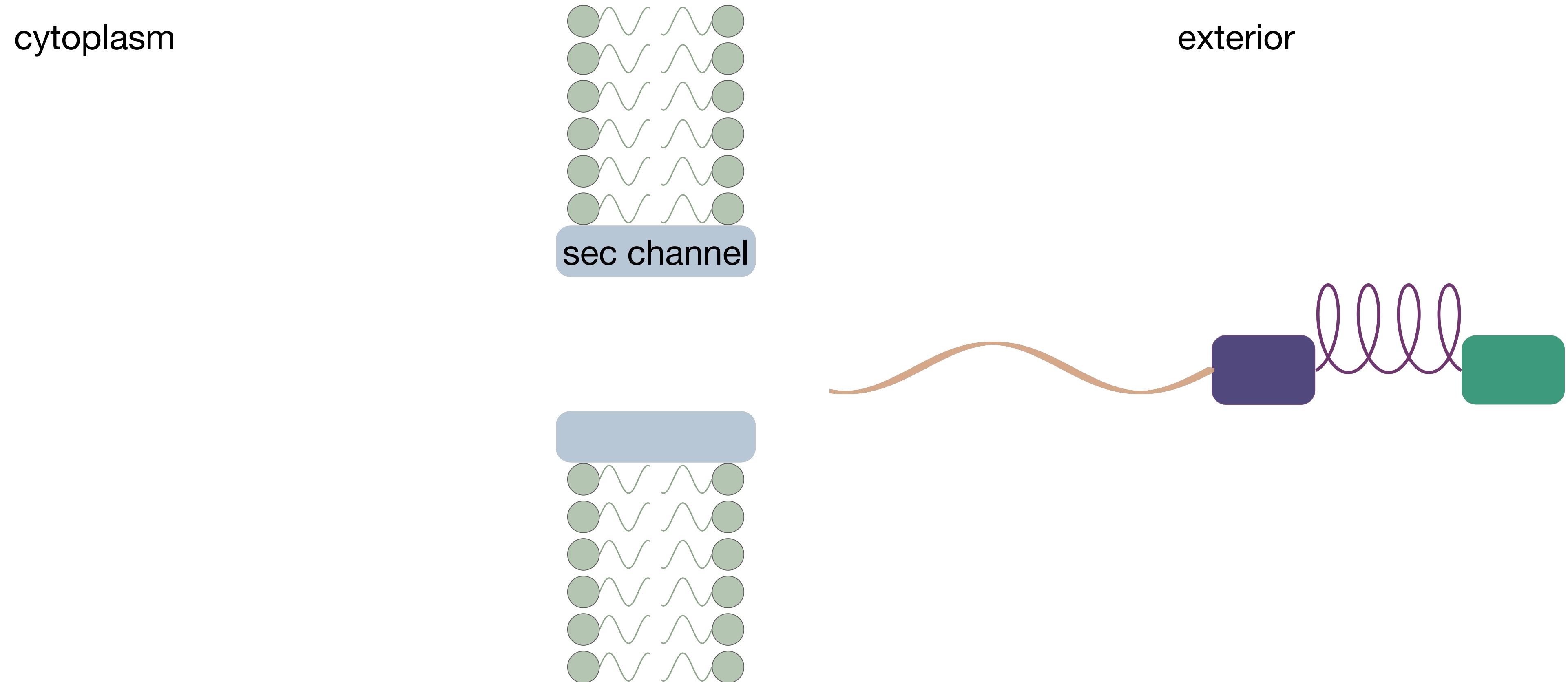
Signal peptides are secretion signals



Signal peptides are secretion signals

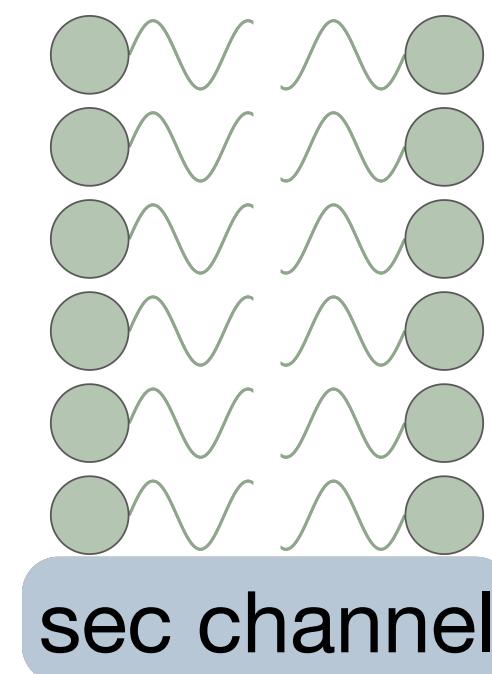


Signal peptides are secretion signals

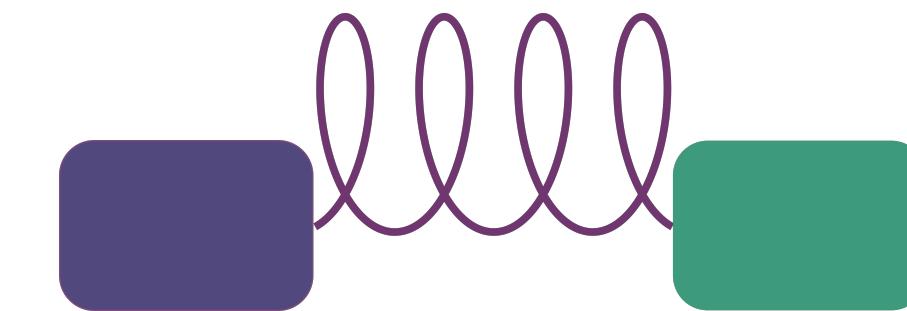
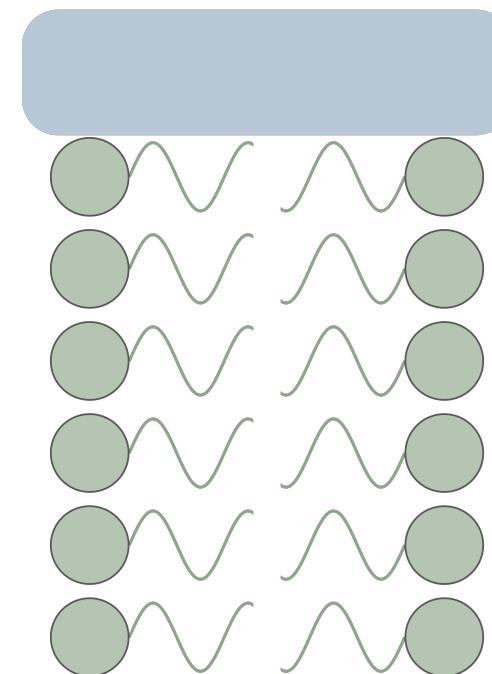


Signal peptides are secretion signals

cytoplasm

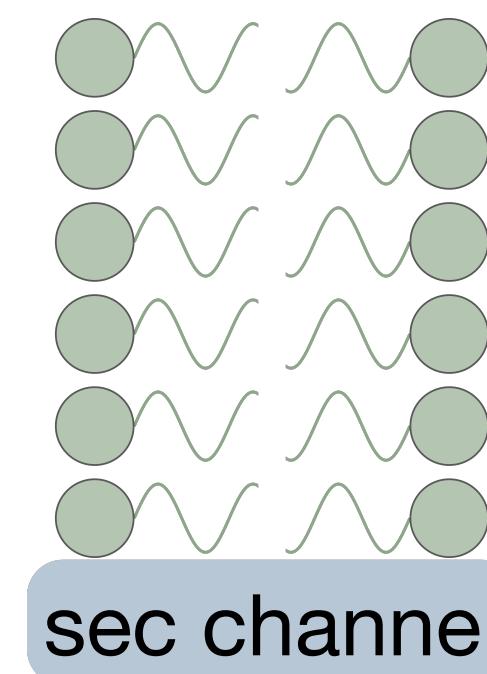


exterior

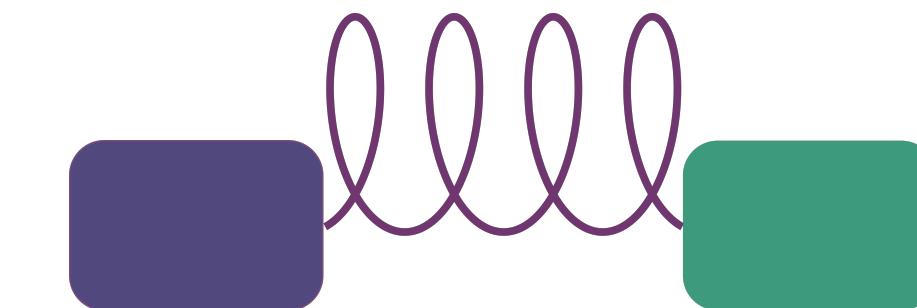
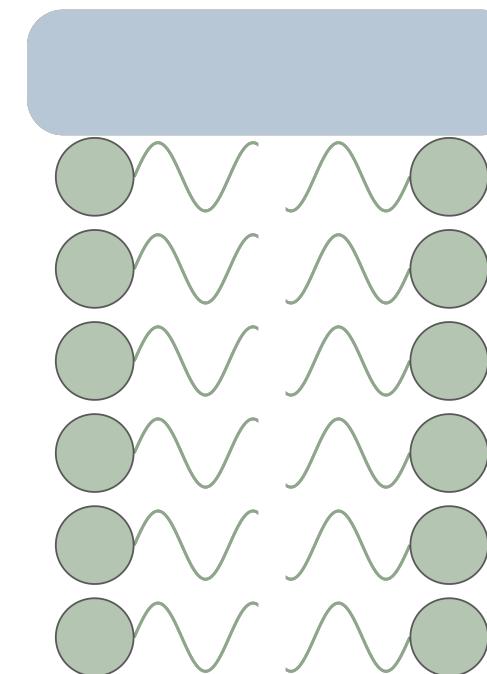
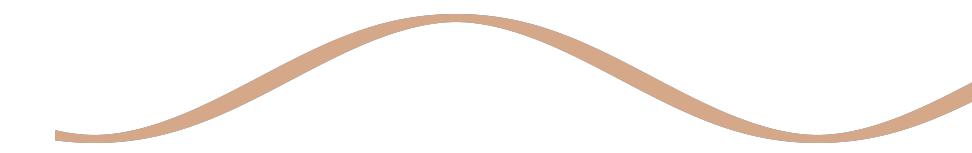


Signal peptides are secretion signals

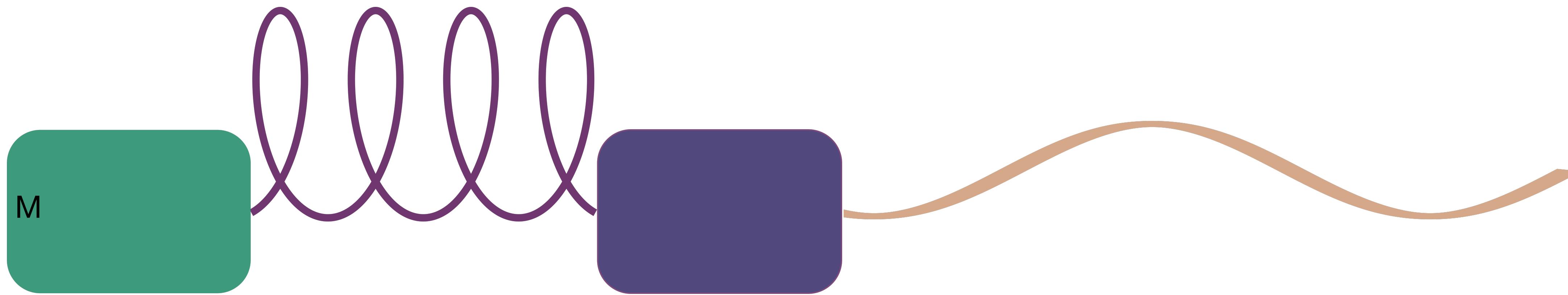
cytoplasm



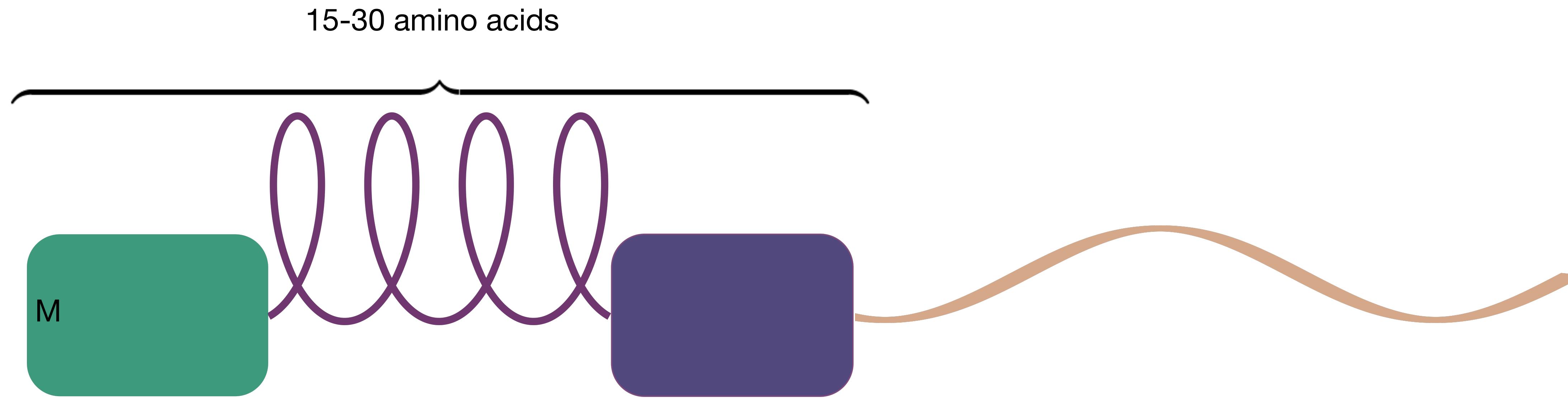
exterior



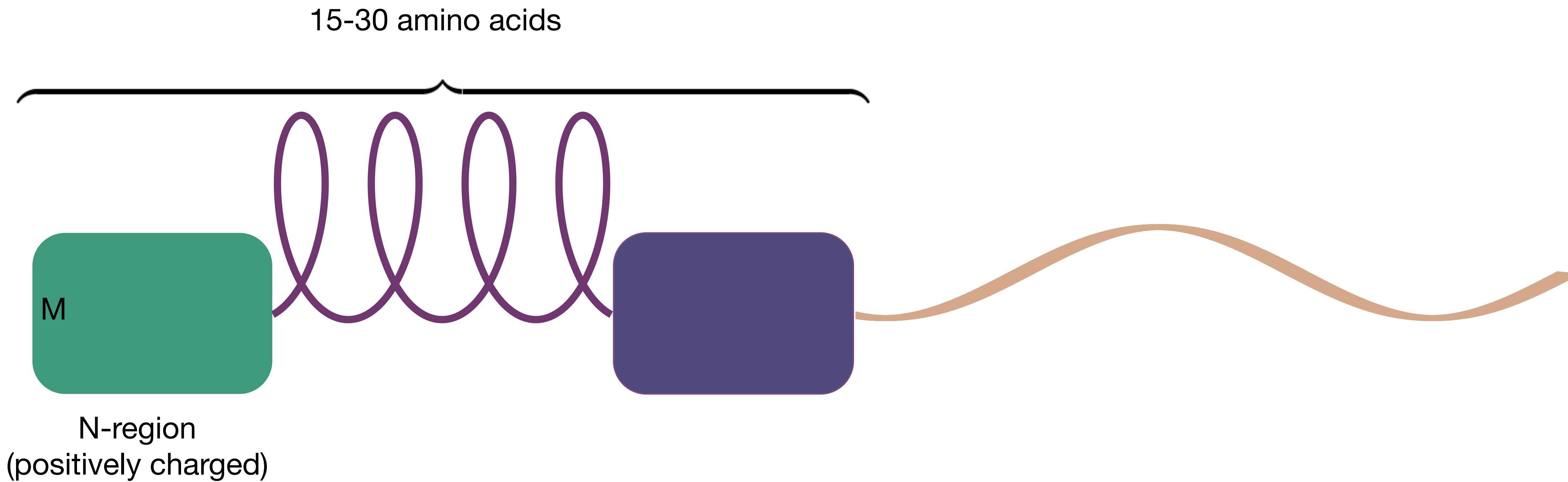
Signal peptides are secretion signals



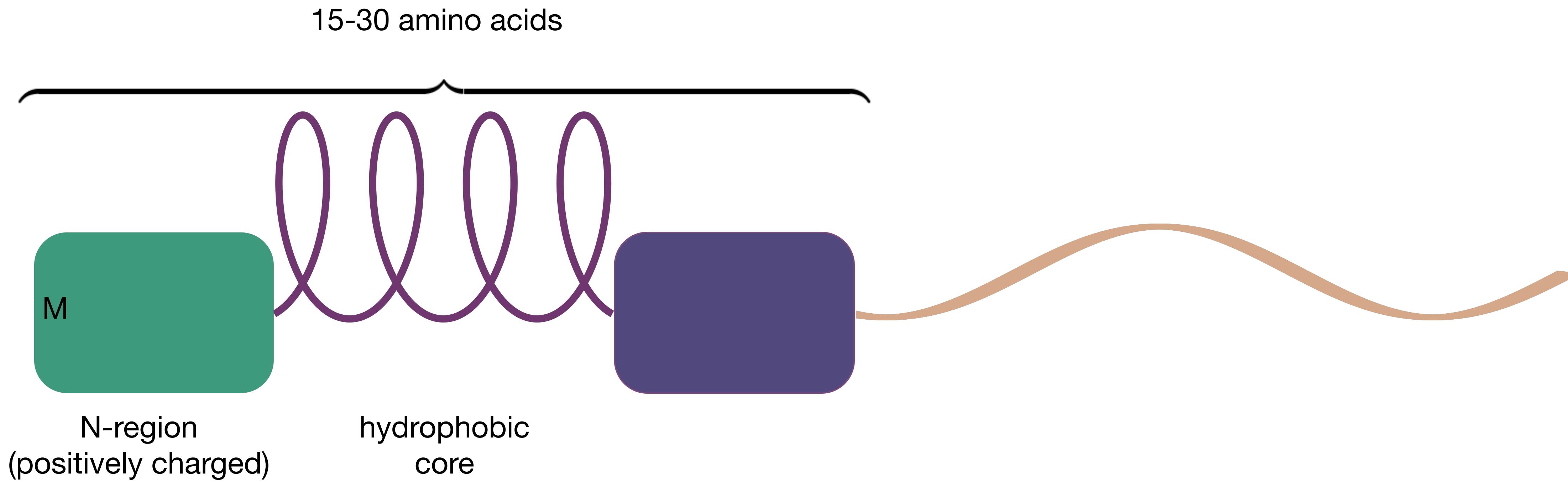
Signal peptides are secretion signals



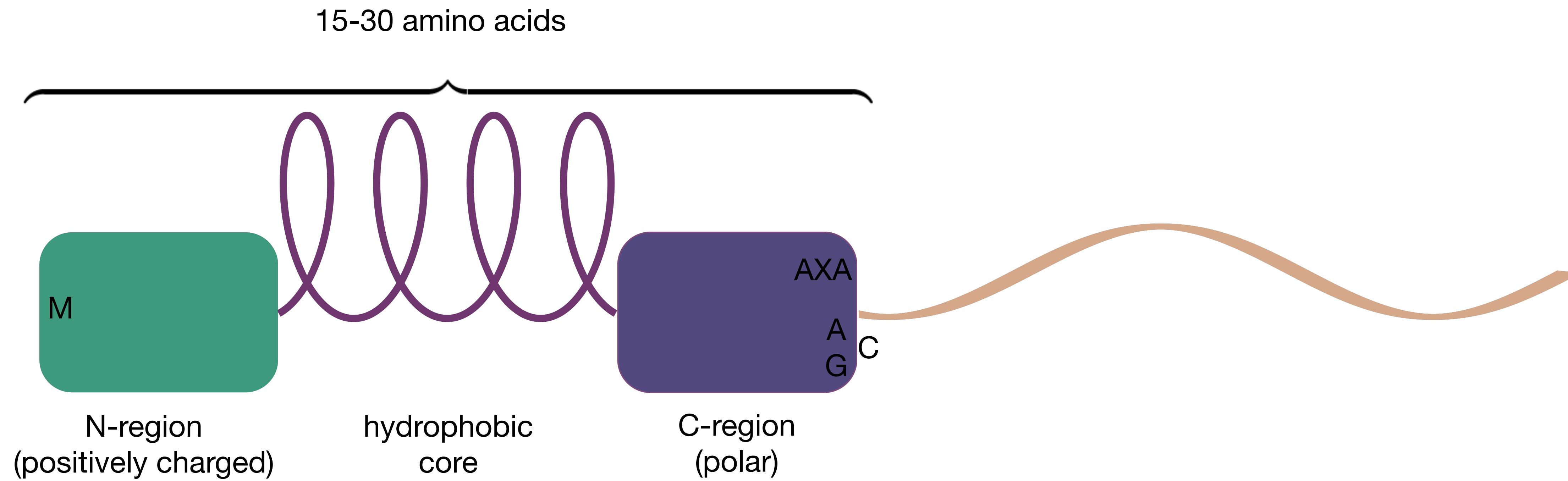
Signal peptides are secretion signals



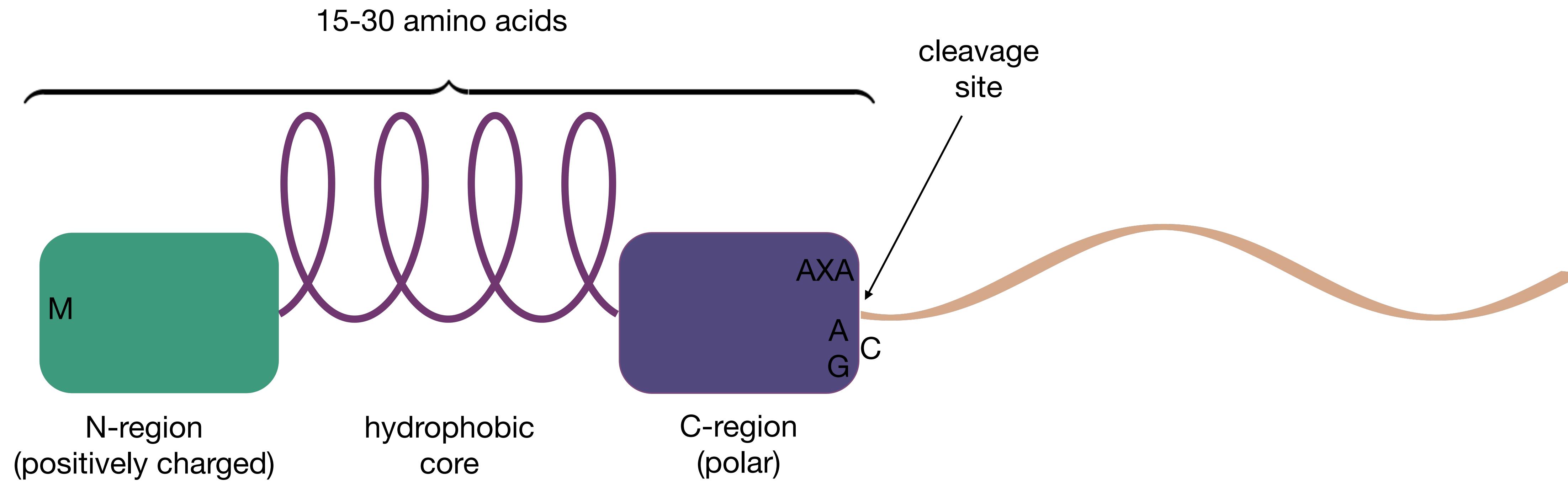
Signal peptides are secretion signals



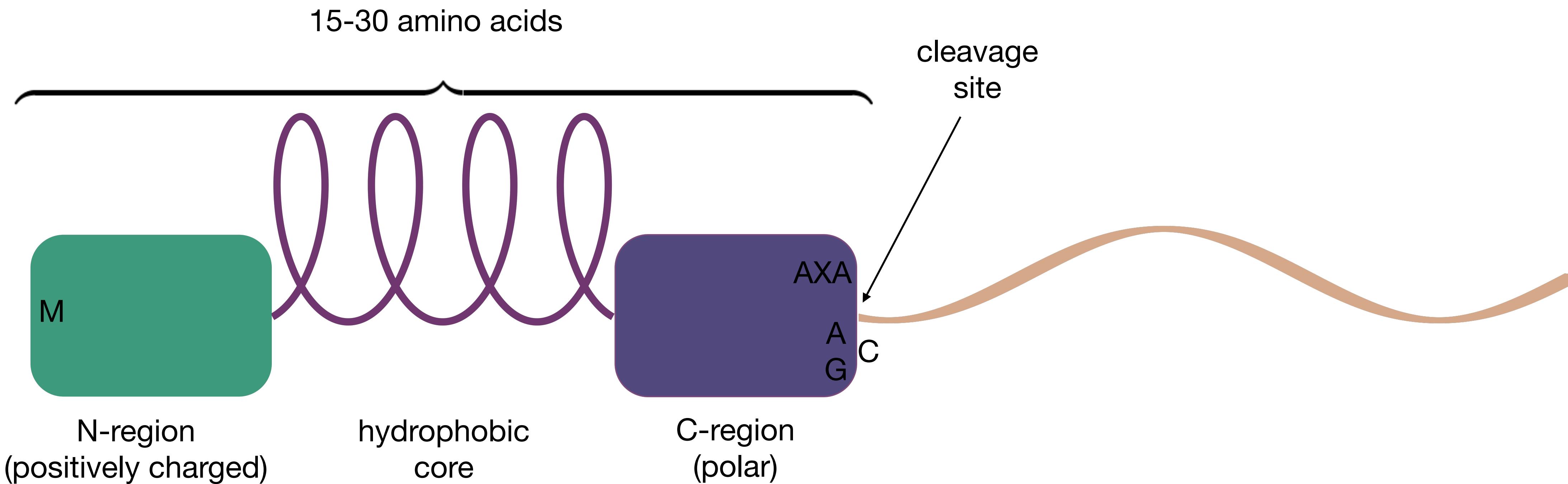
Signal peptides are secretion signals



Signal peptides are secretion signals

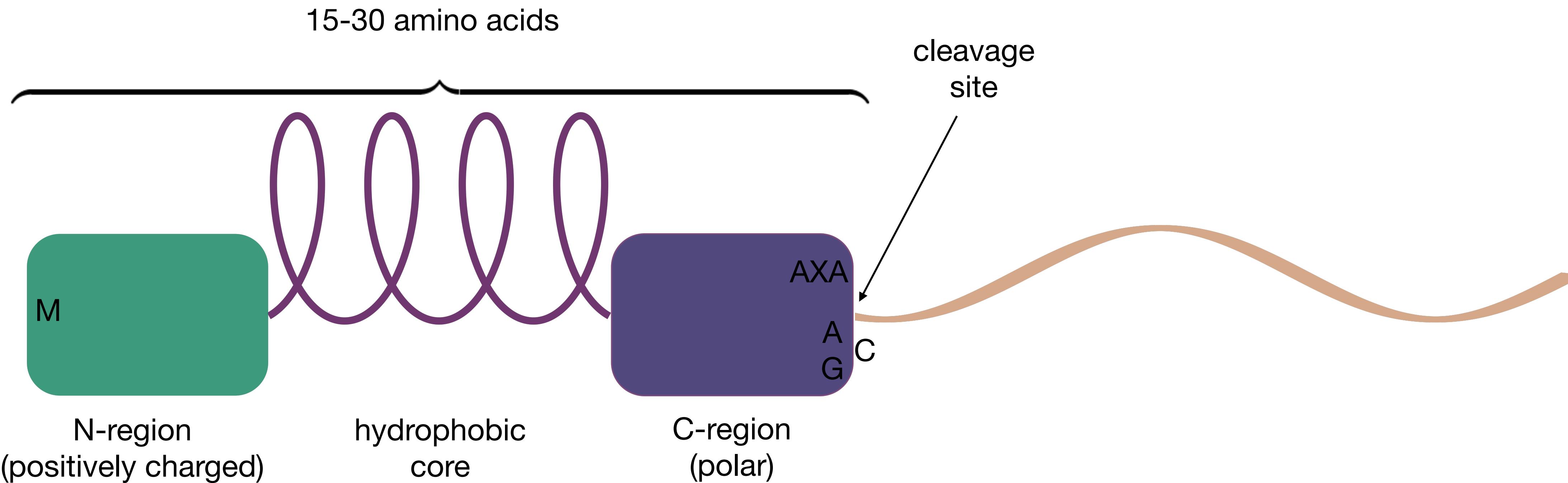


Signal peptides are secretion signals



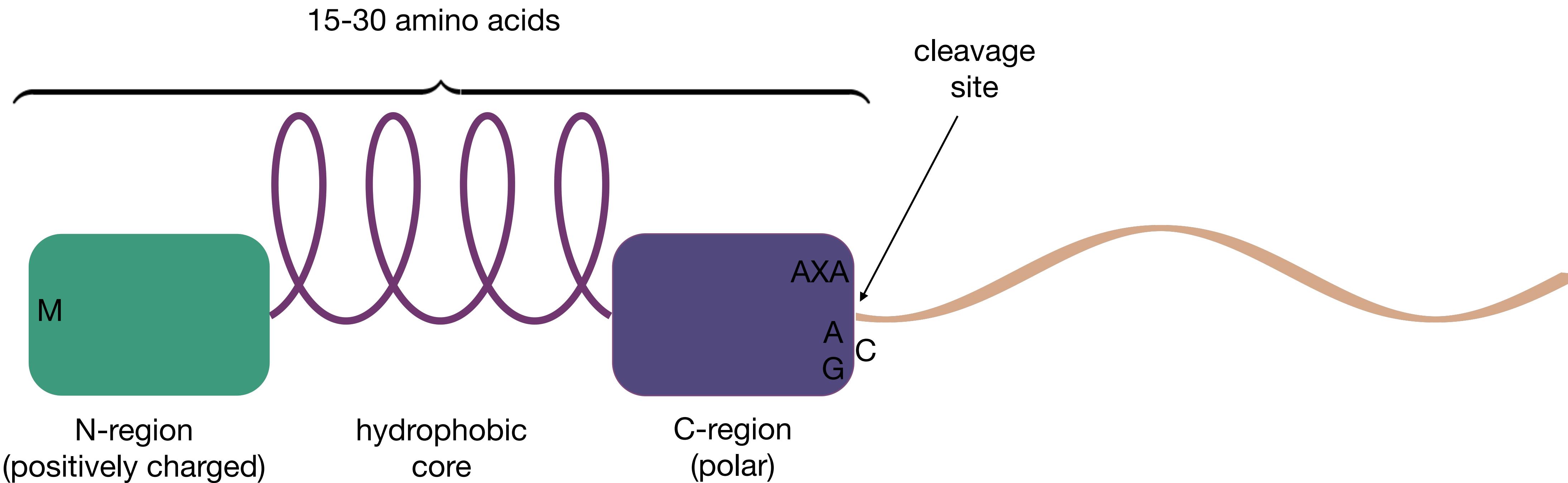
- Extra-cellular secretion in prokaryotes

Signal peptides are secretion signals



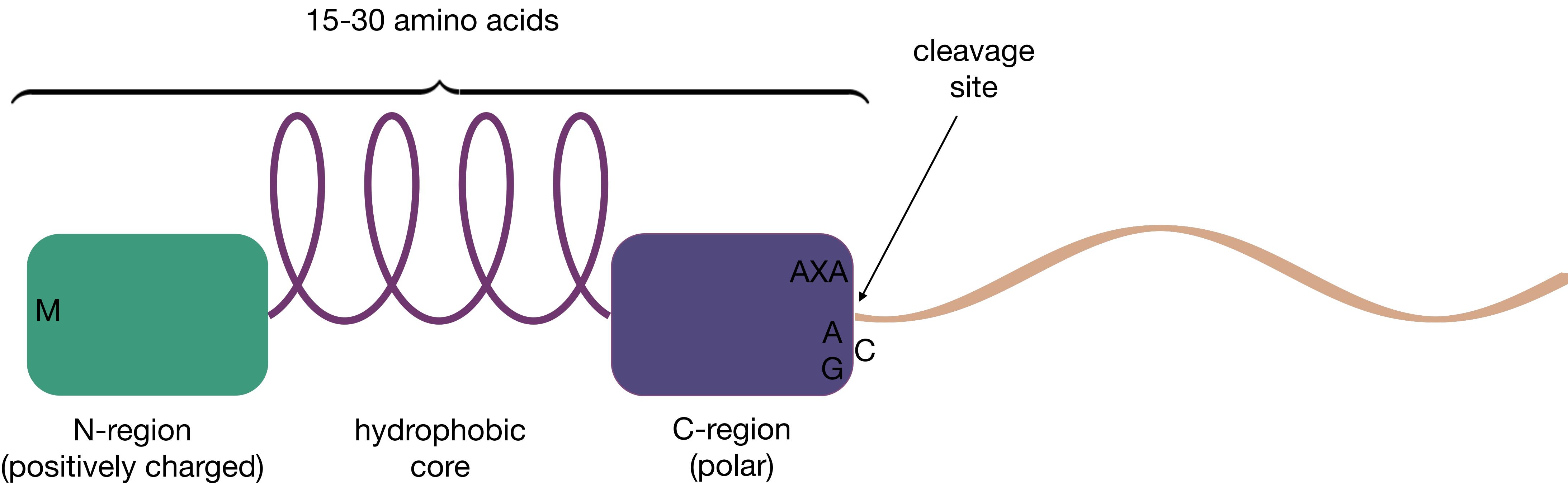
- Extra-cellular secretion in prokaryotes
- Secretion from endoplasmic reticulum in eukaryotes

Signal peptides are secretion signals



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

Signal peptides are secretion signals



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

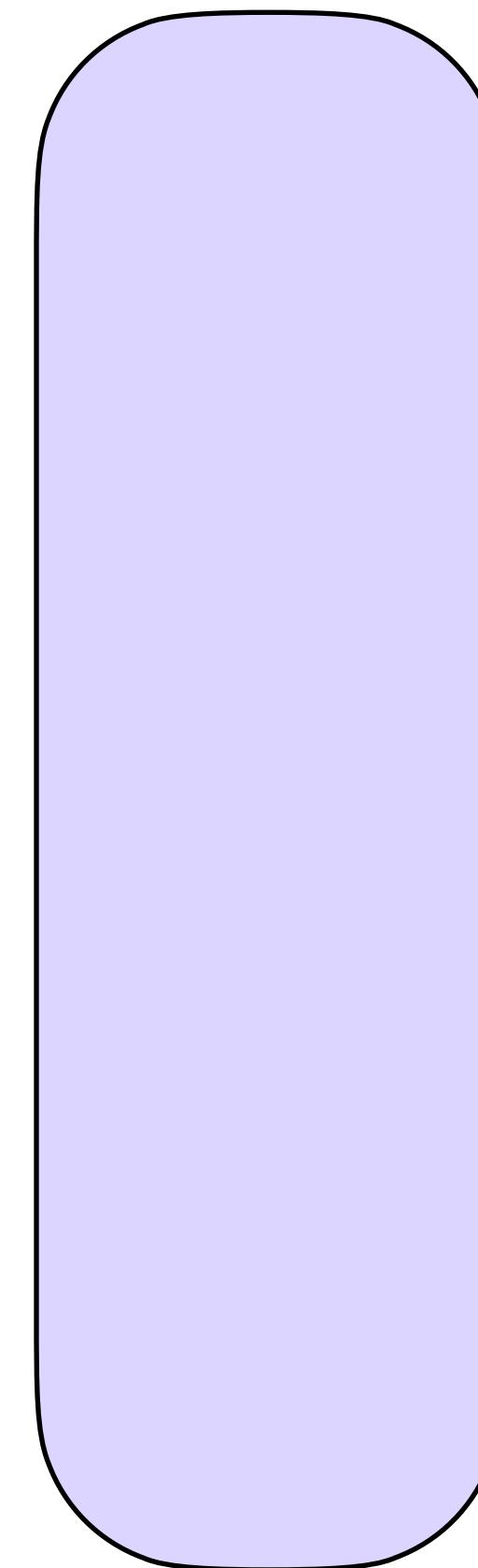
SPs simplify protein production

SPs simplify protein production

- Less burden on cells

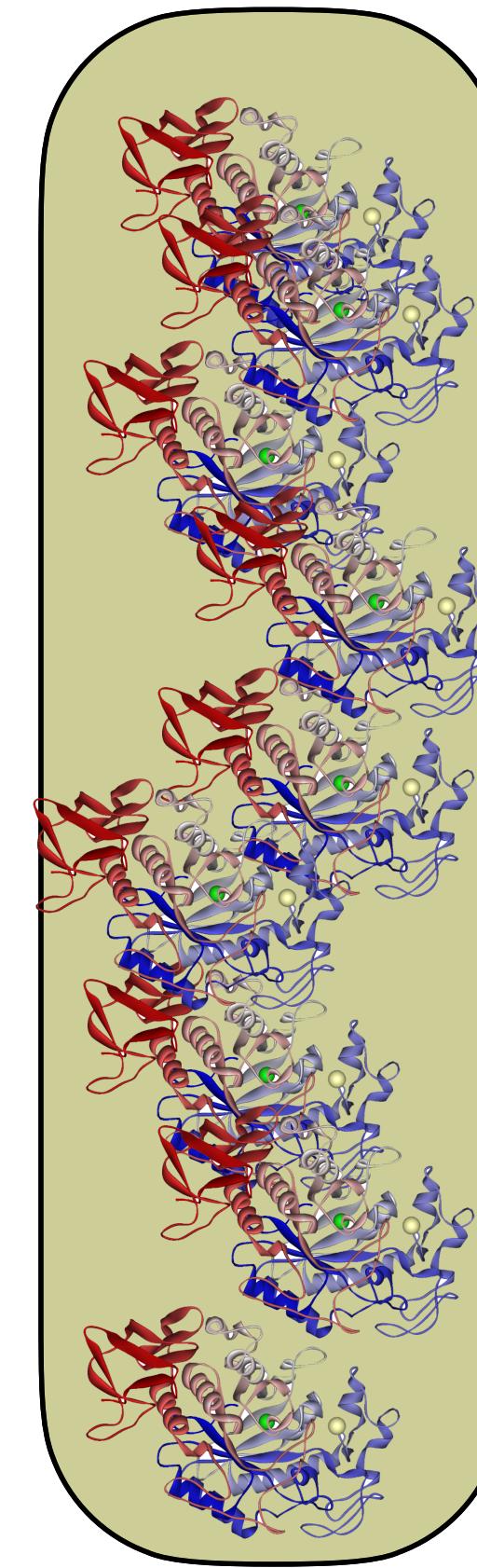
SPs simplify protein production

- Less burden on cells



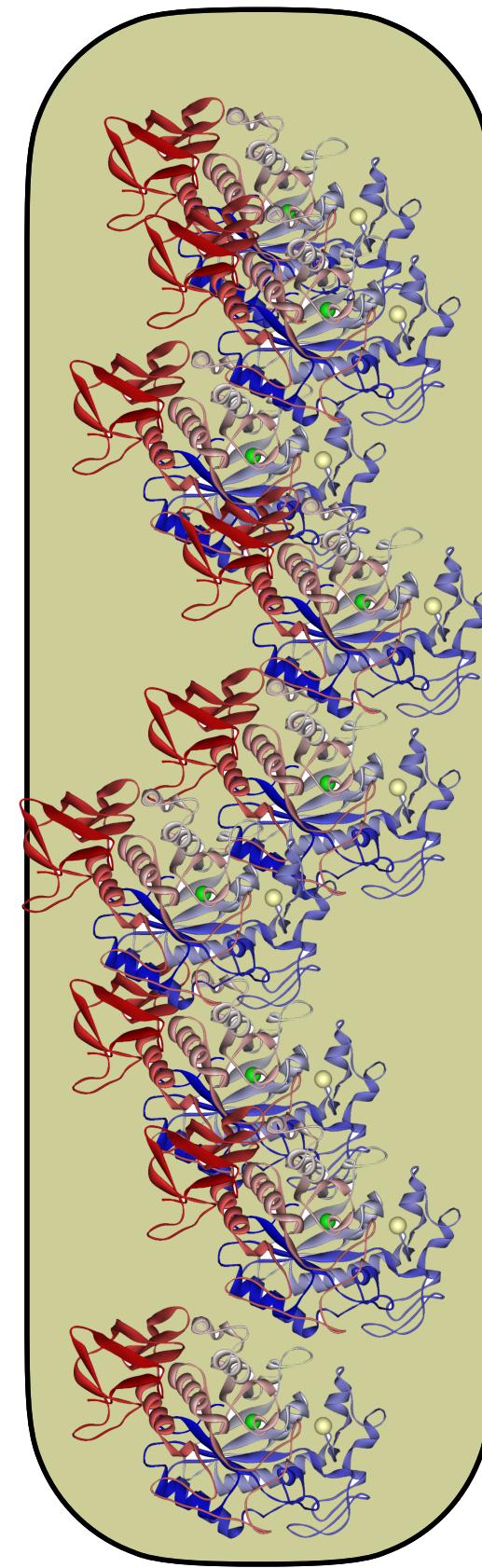
SPs simplify protein production

- Less burden on cells



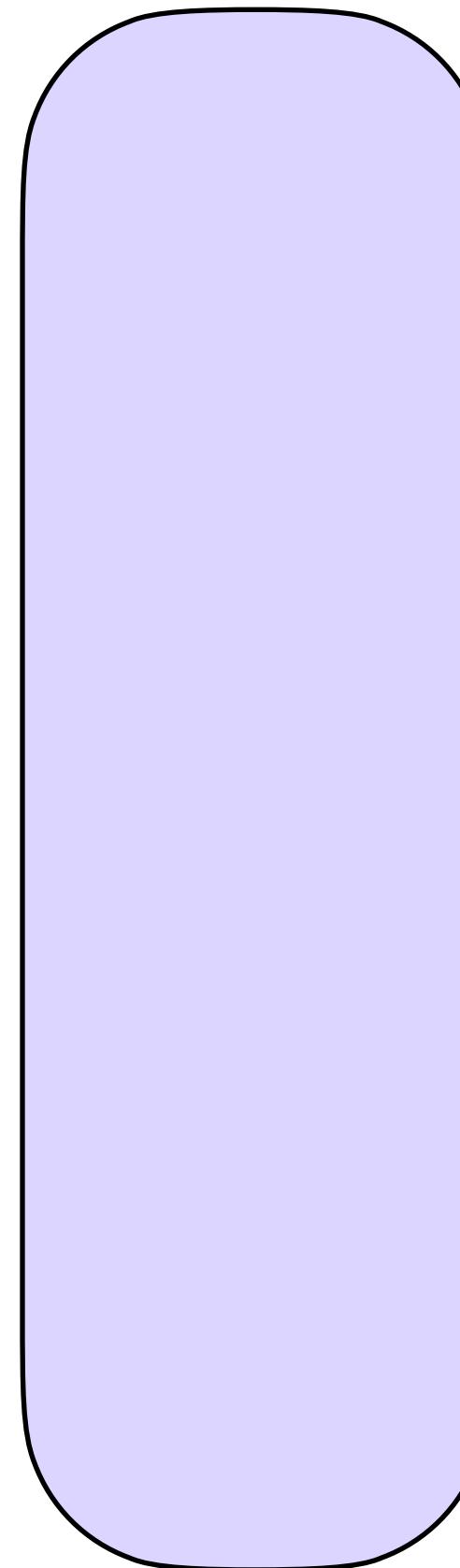
SPs simplify protein production

- Less burden on cells
- Easier separation



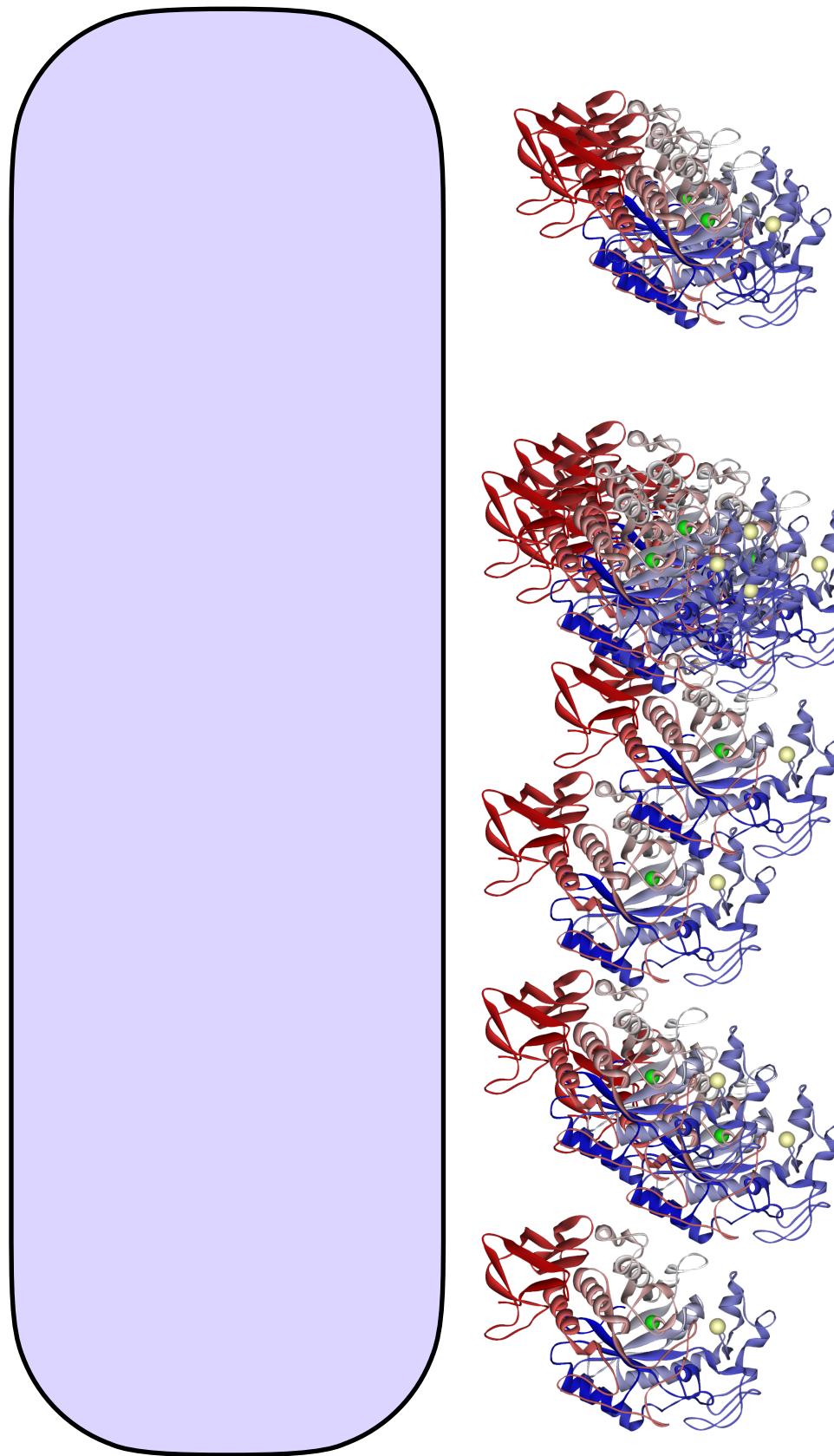
SPs simplify protein production

- Less burden on cells
- Easier separation



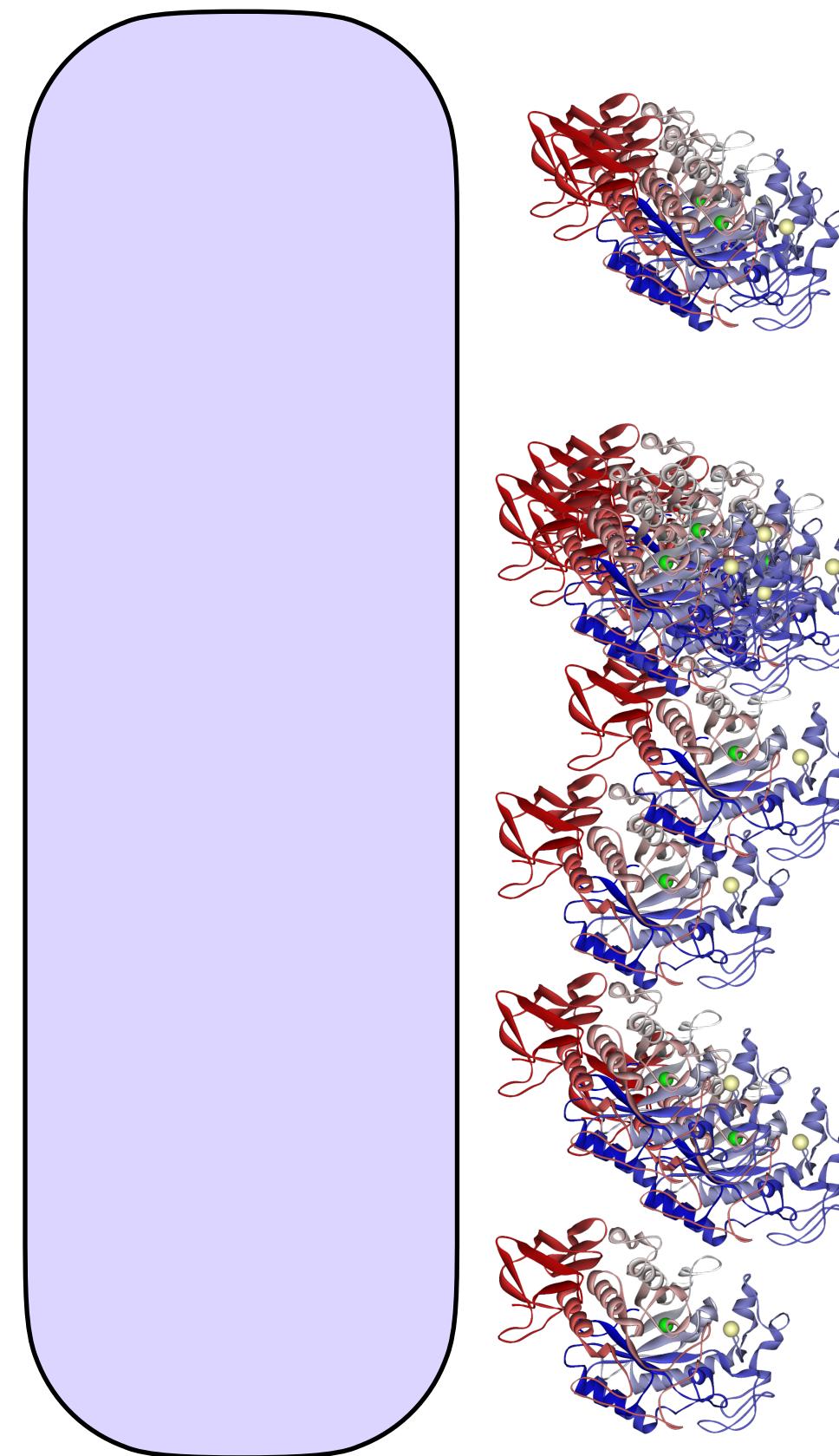
SPs simplify protein production

- Less burden on cells
- Easier separation



SPs simplify protein production

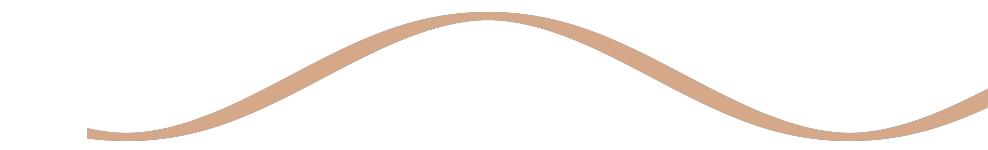
- Less burden on cells
- Easier separation



Need: custom SPs for arbitrary proteins

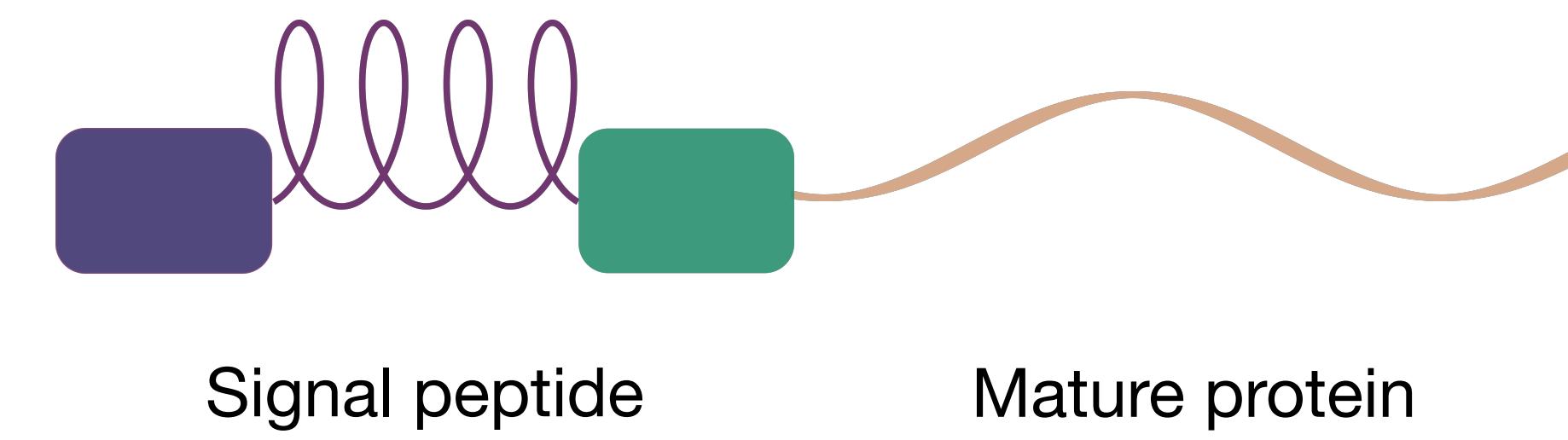
Use machine translation to generate SPs

Use machine translation to generate SPs

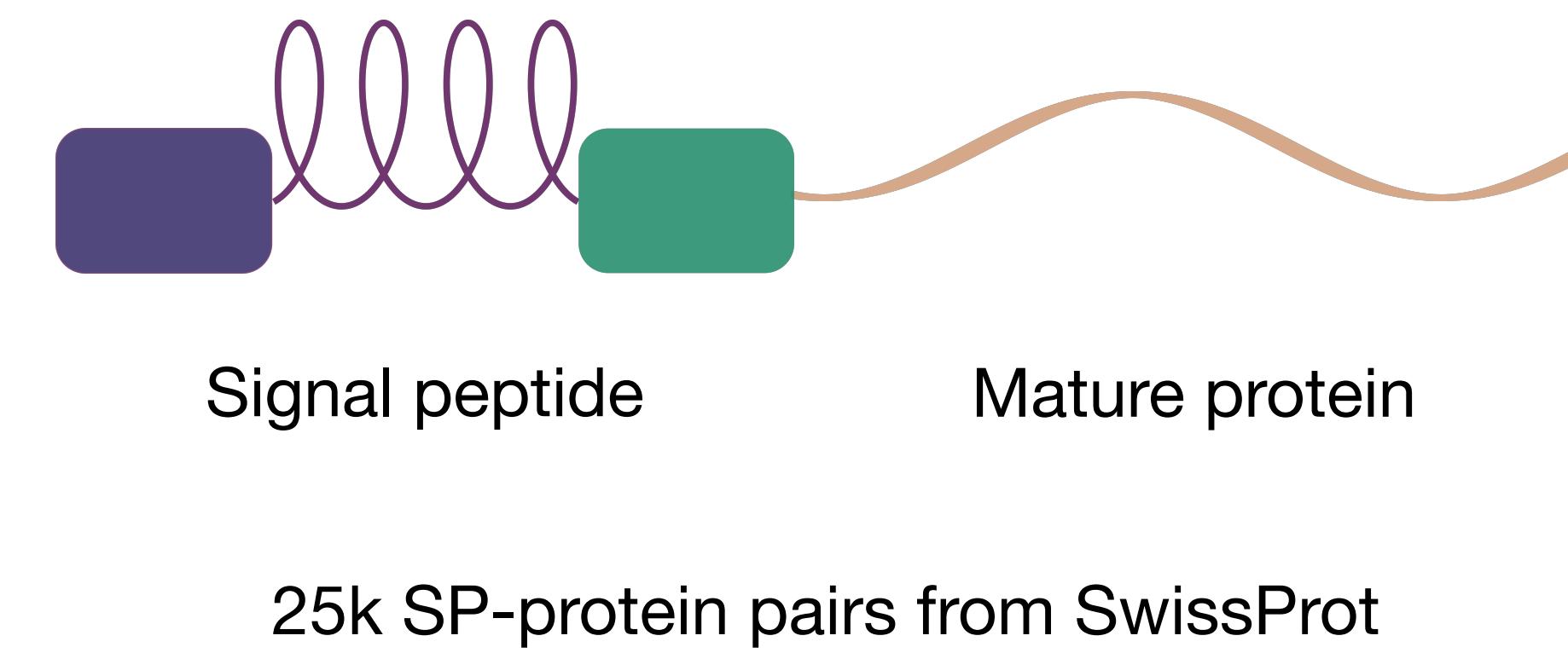


Mature protein

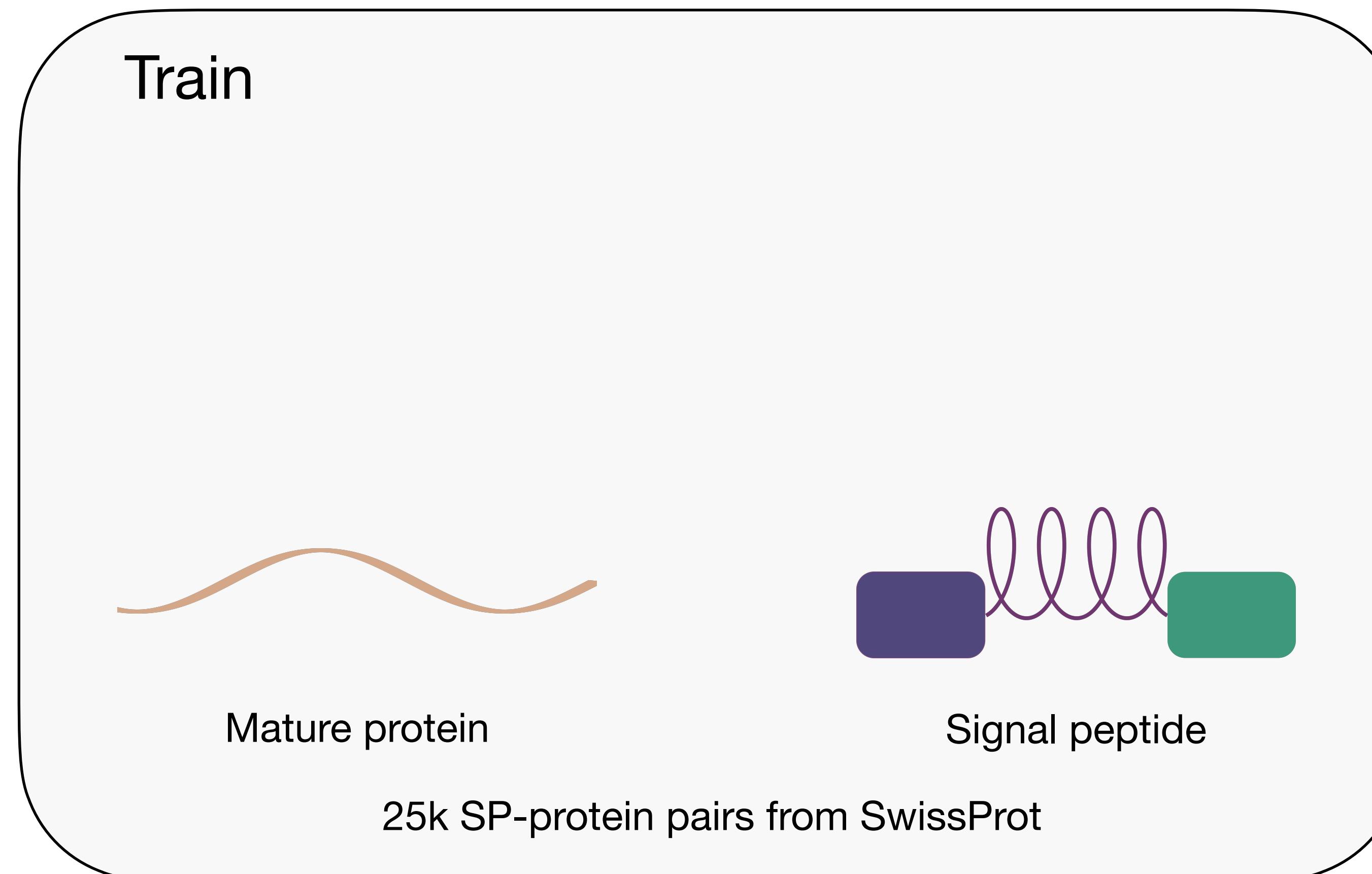
Use machine translation to generate SPs



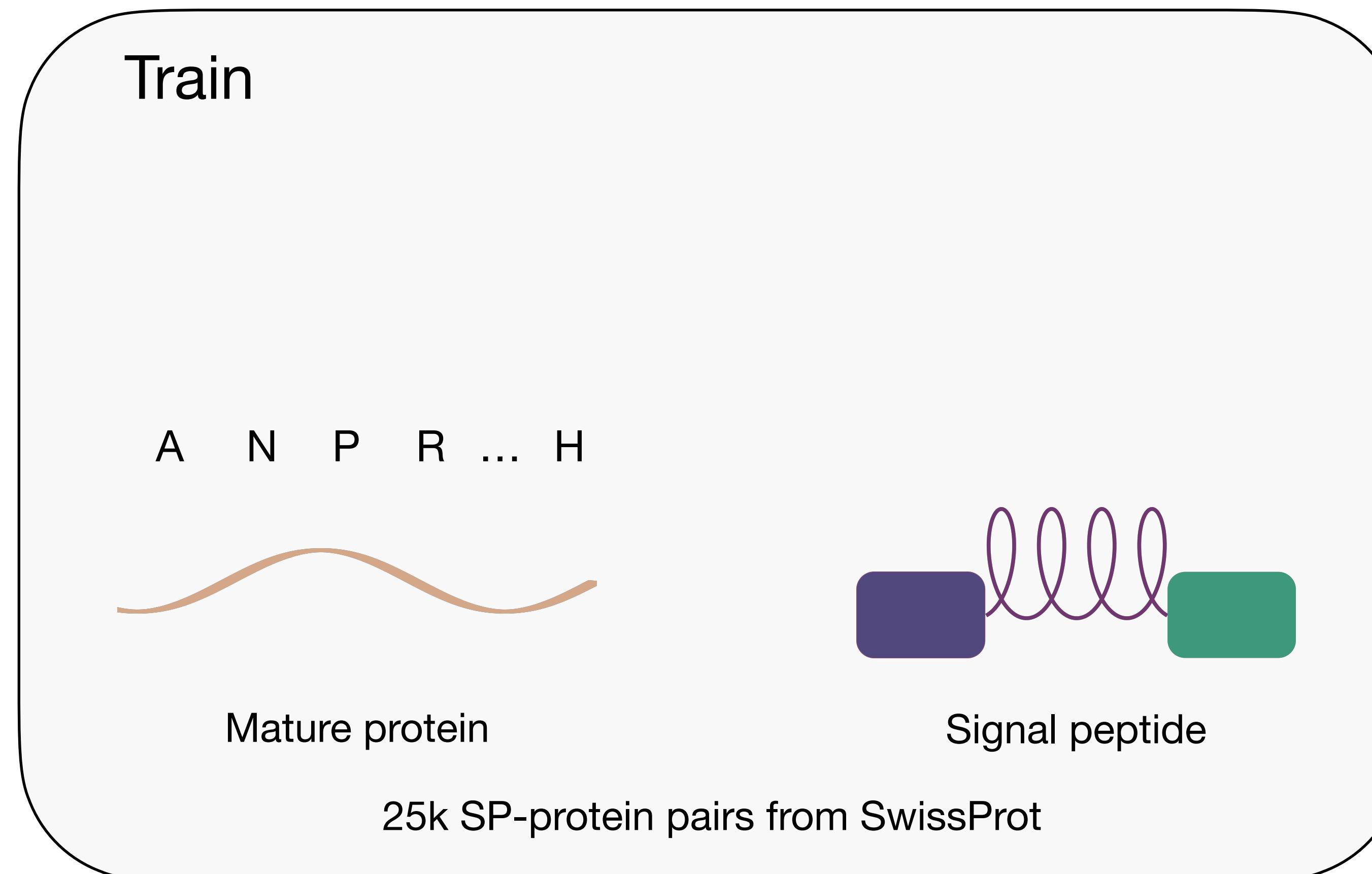
Use machine translation to generate SPs



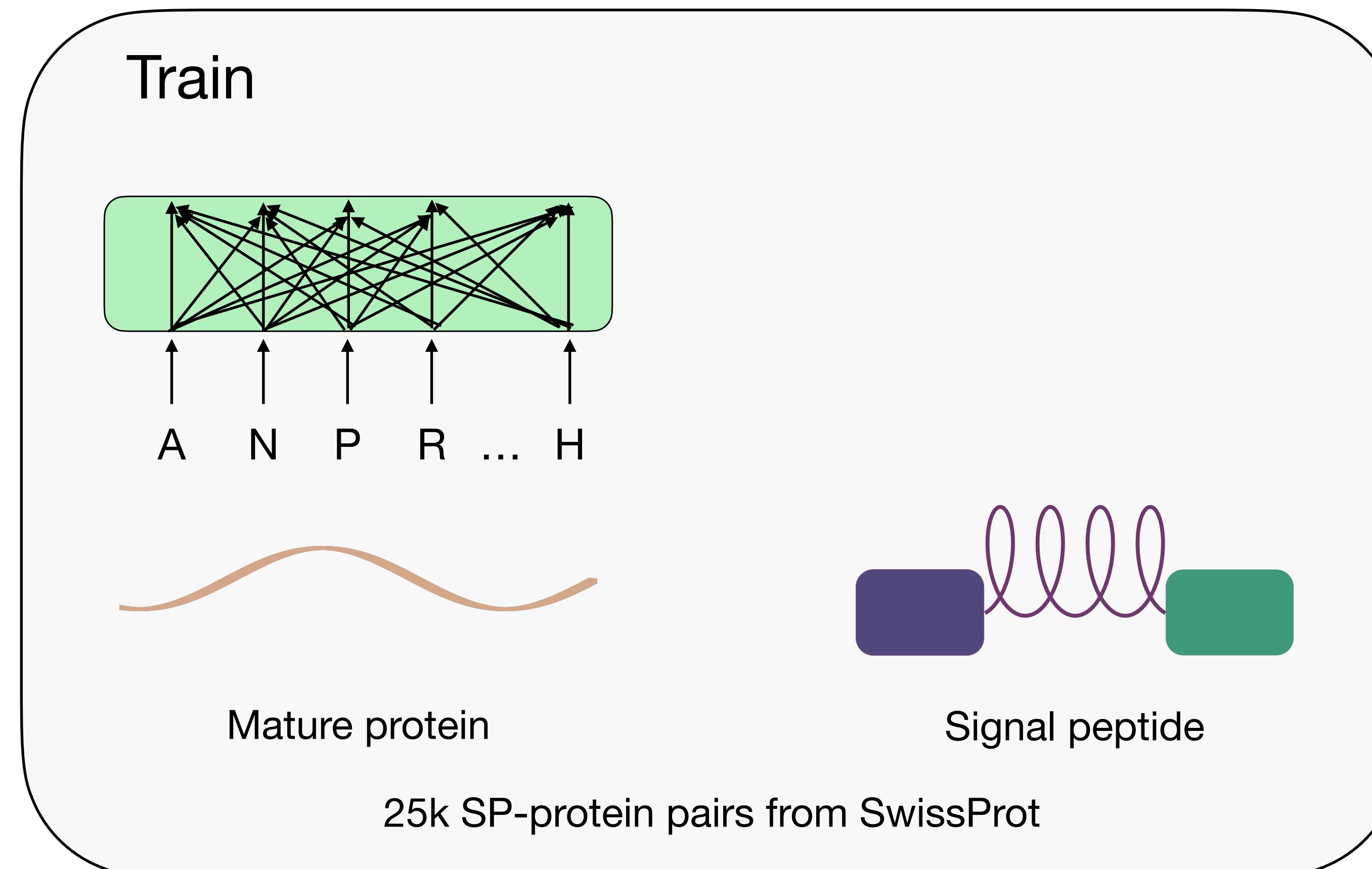
Use machine translation to generate SPs



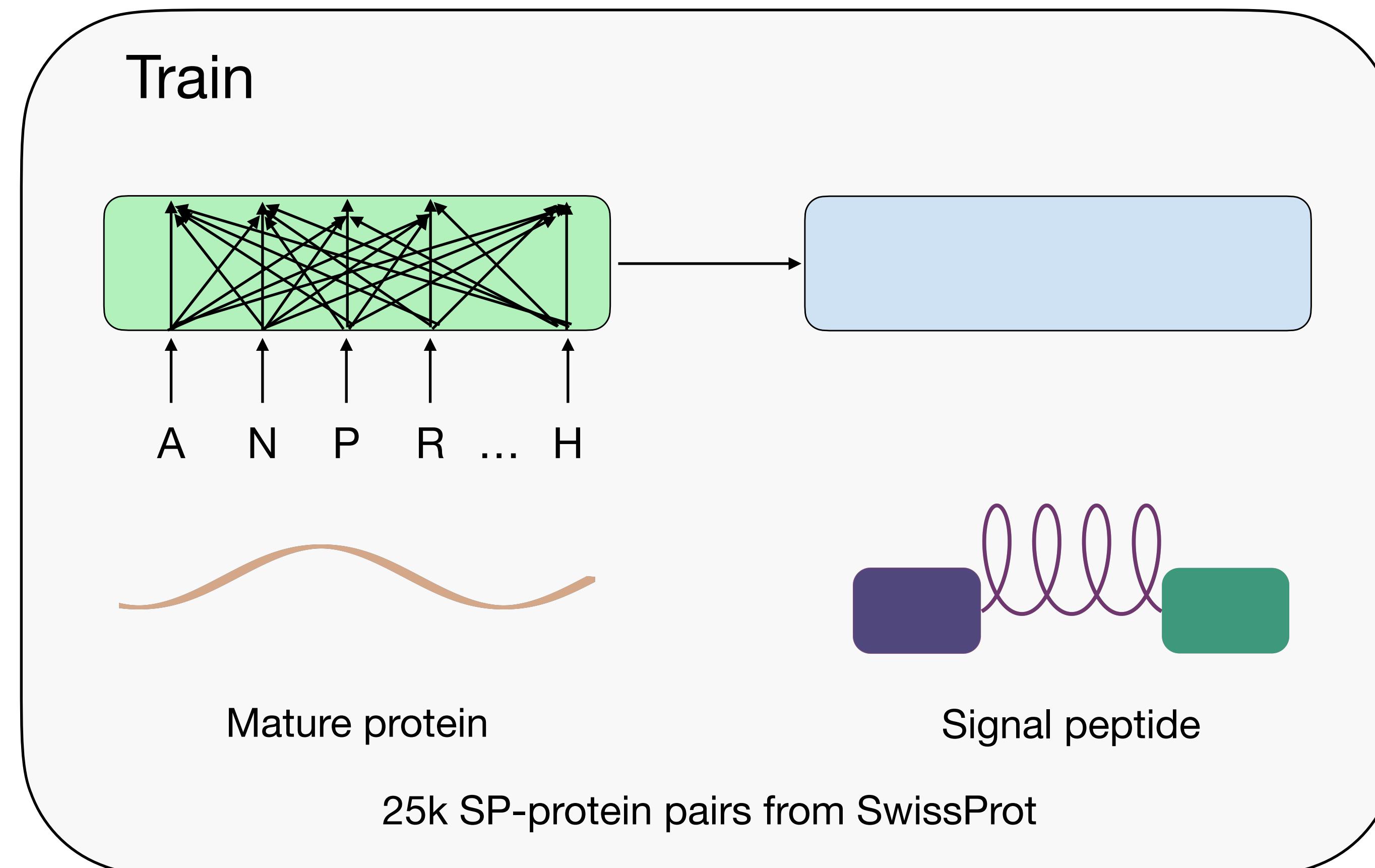
Use machine translation to generate SPs



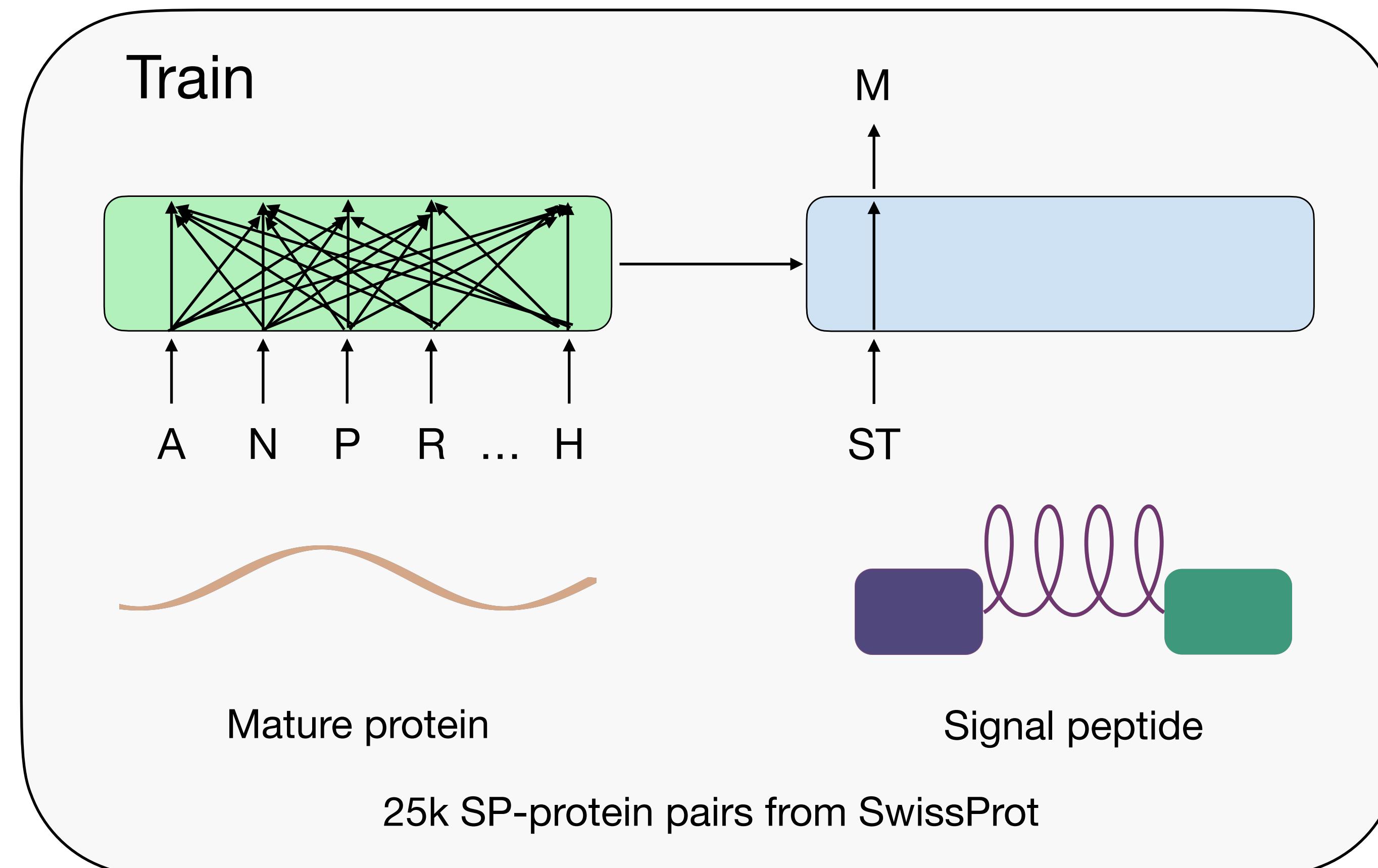
Use machine translation to generate SPs



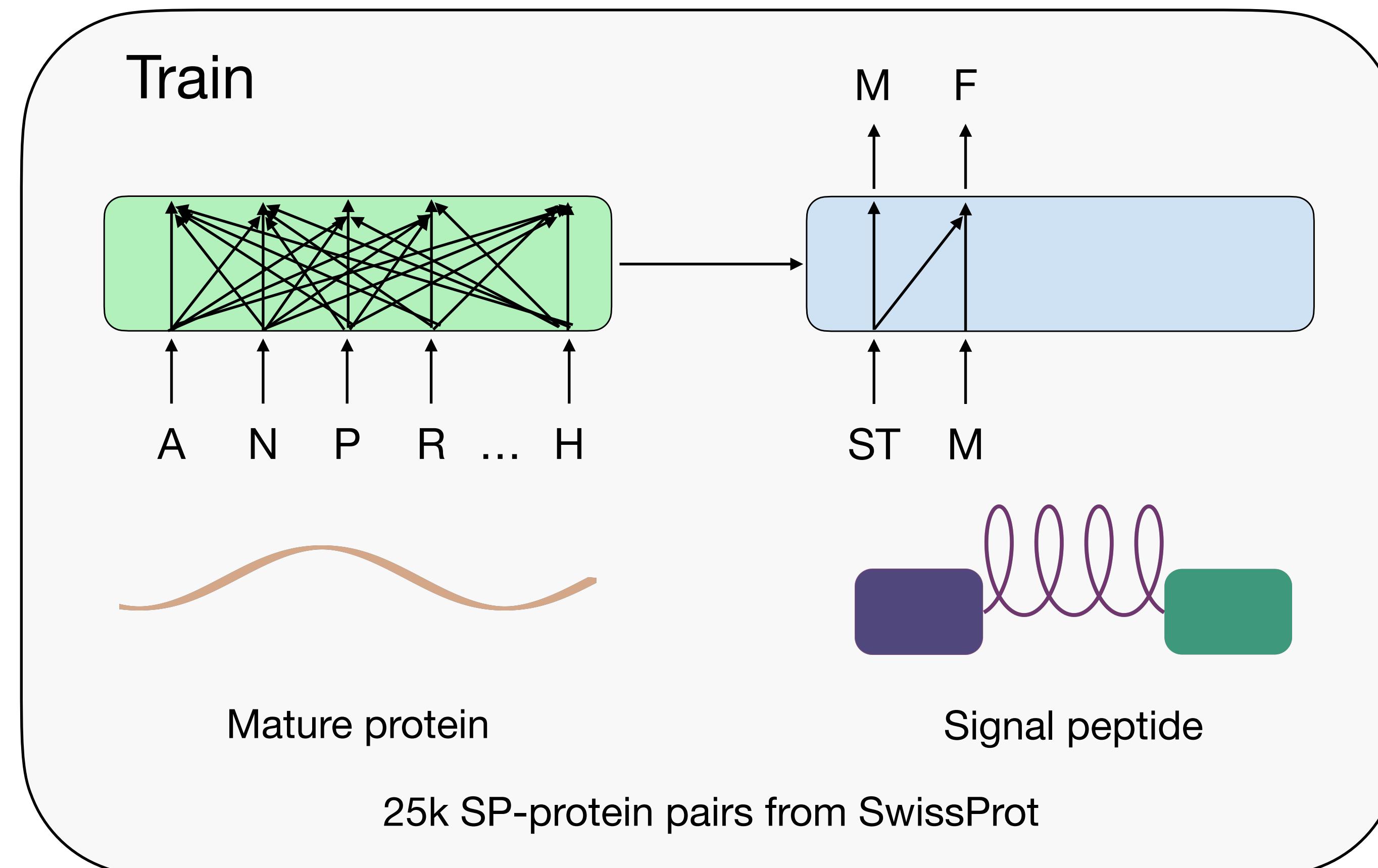
Use machine translation to generate SPs



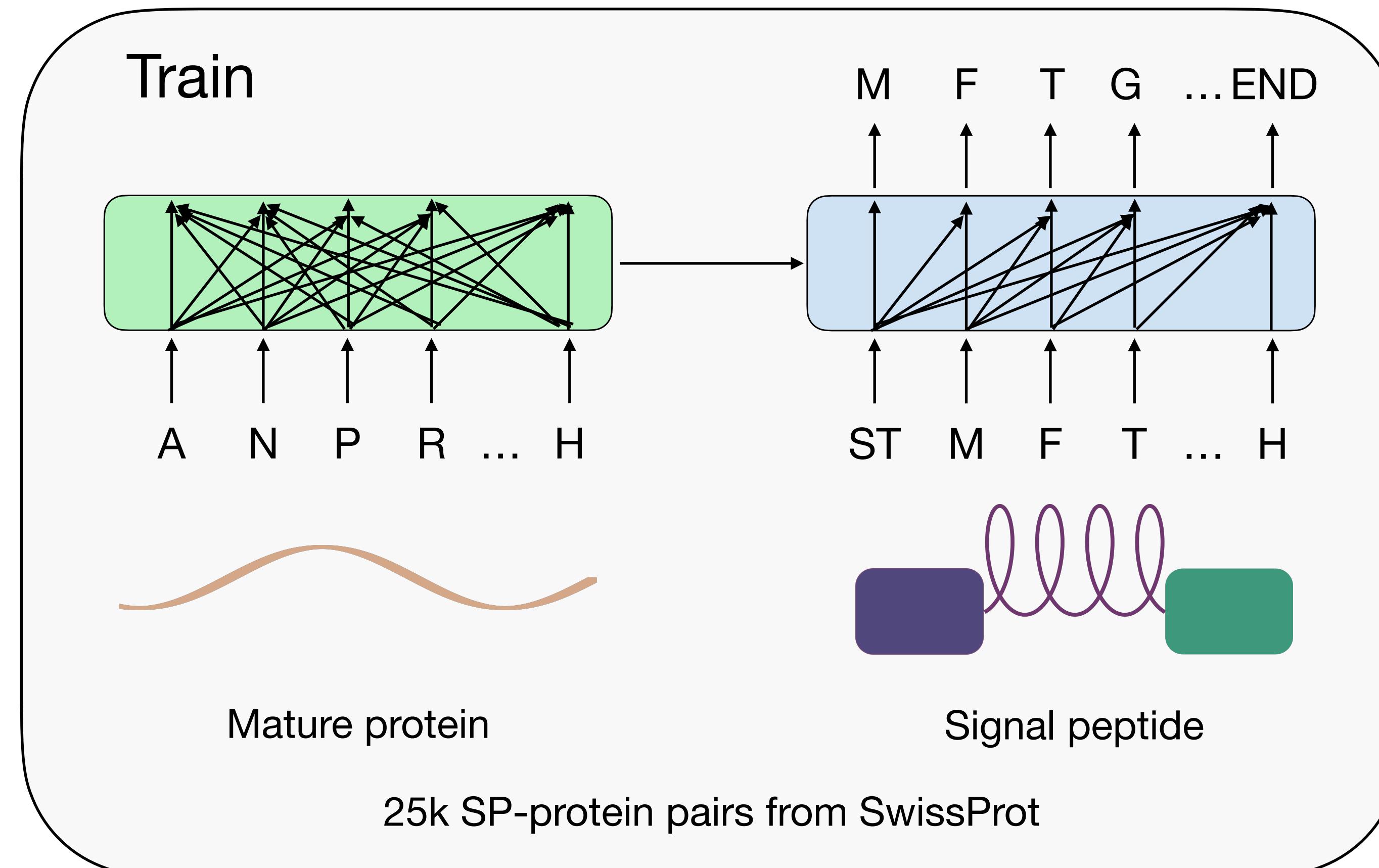
Use machine translation to generate SPs



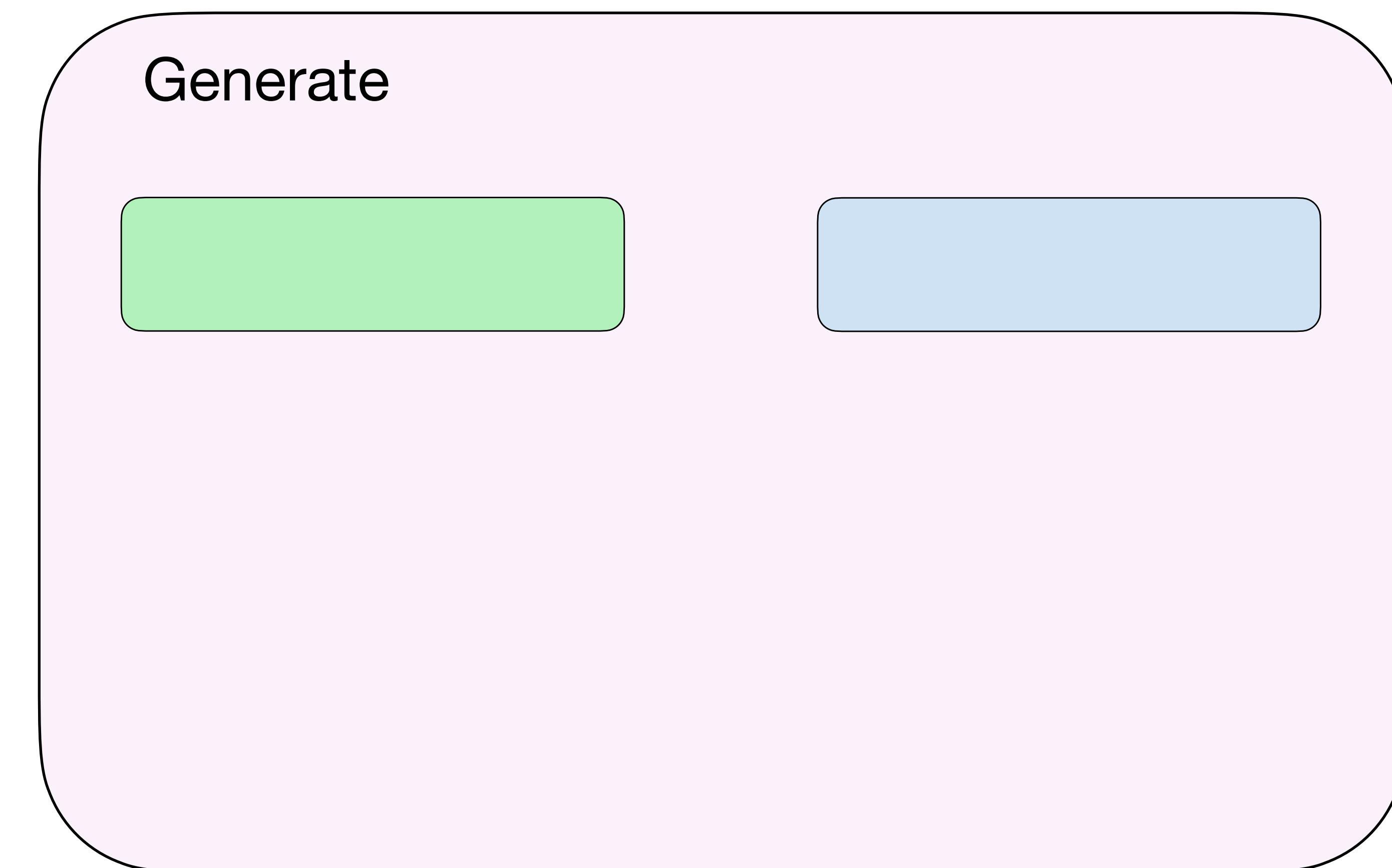
Use machine translation to generate SPs



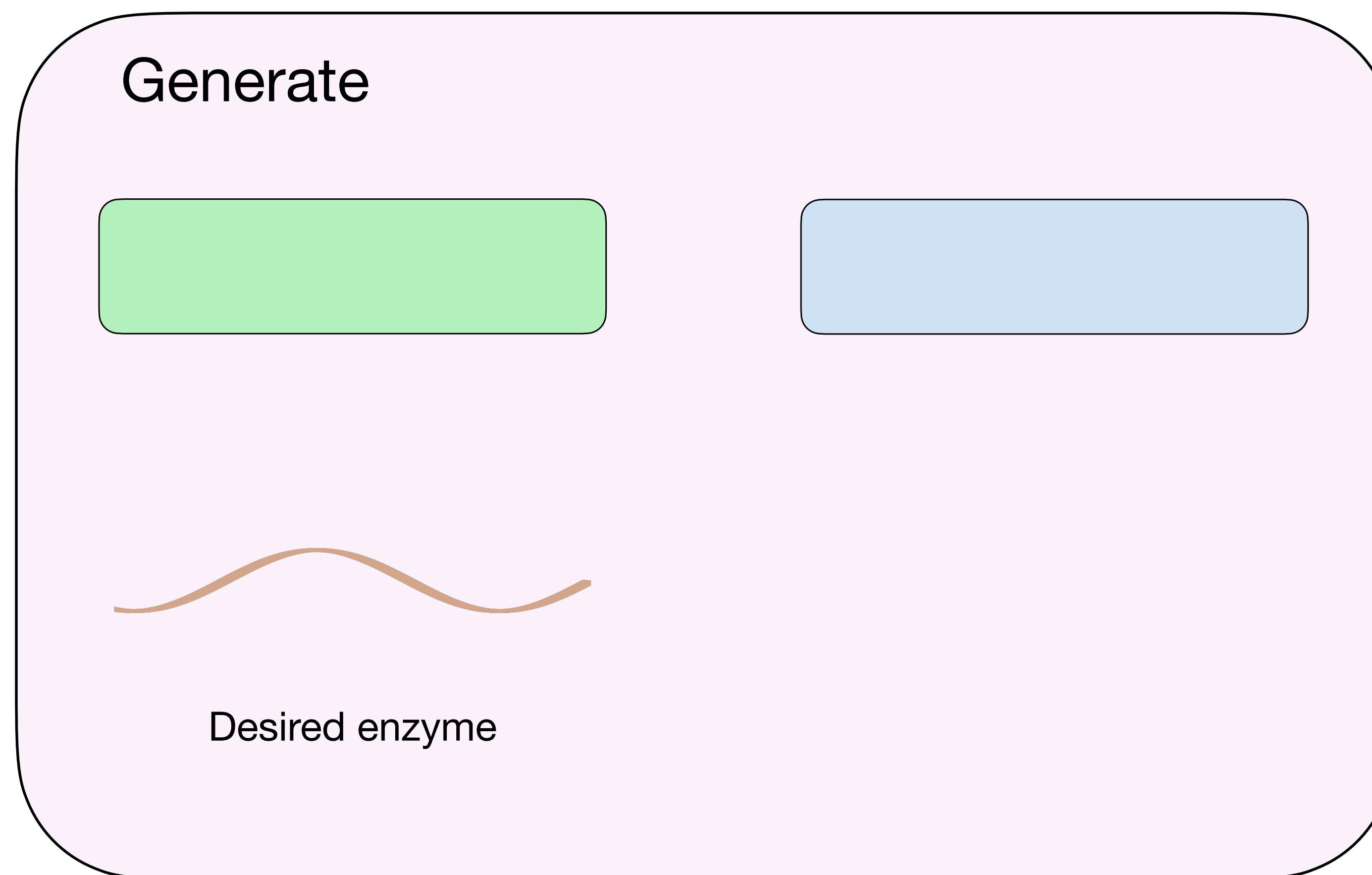
Use machine translation to generate SPs



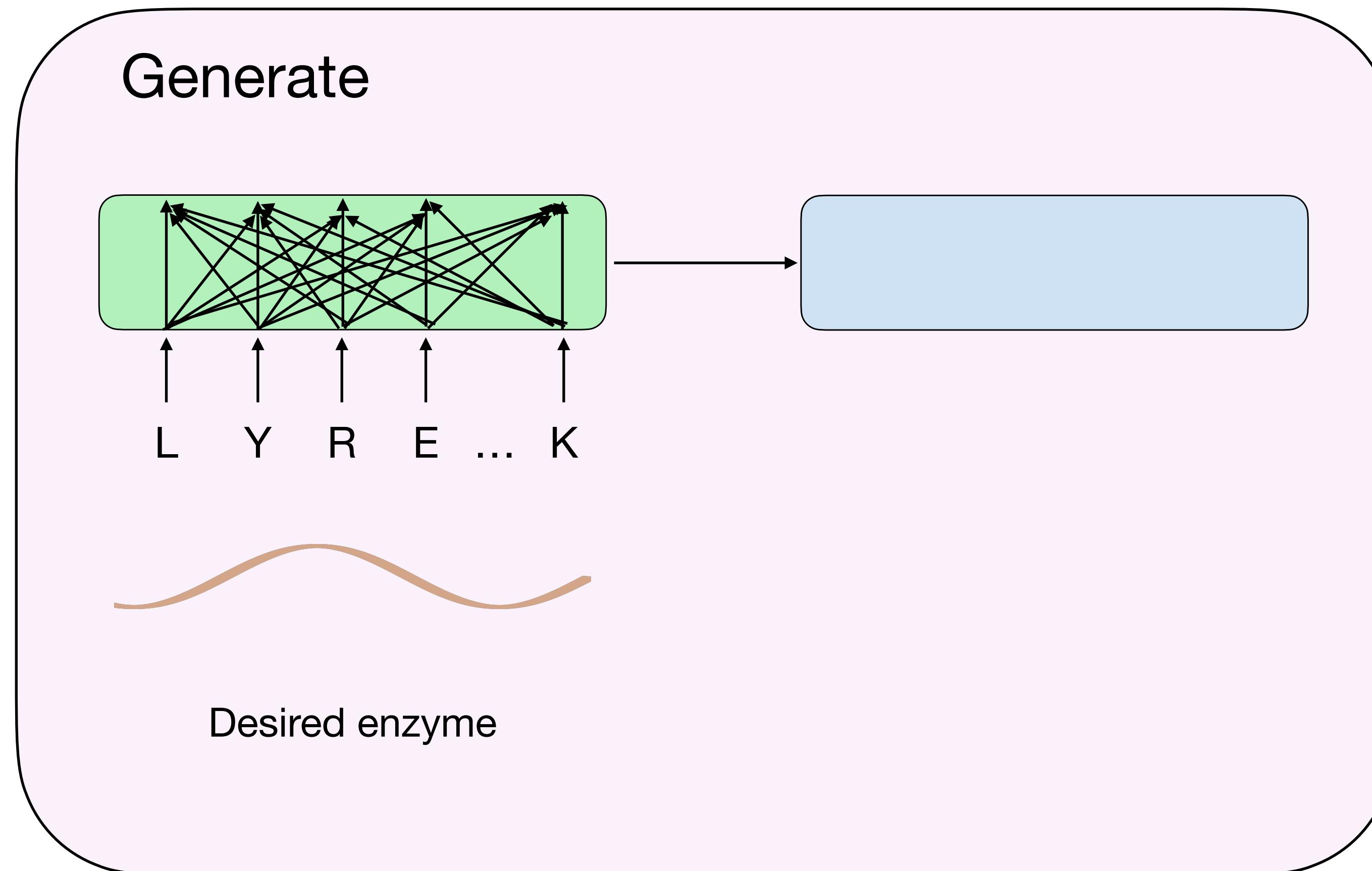
Use machine translation to generate SPs



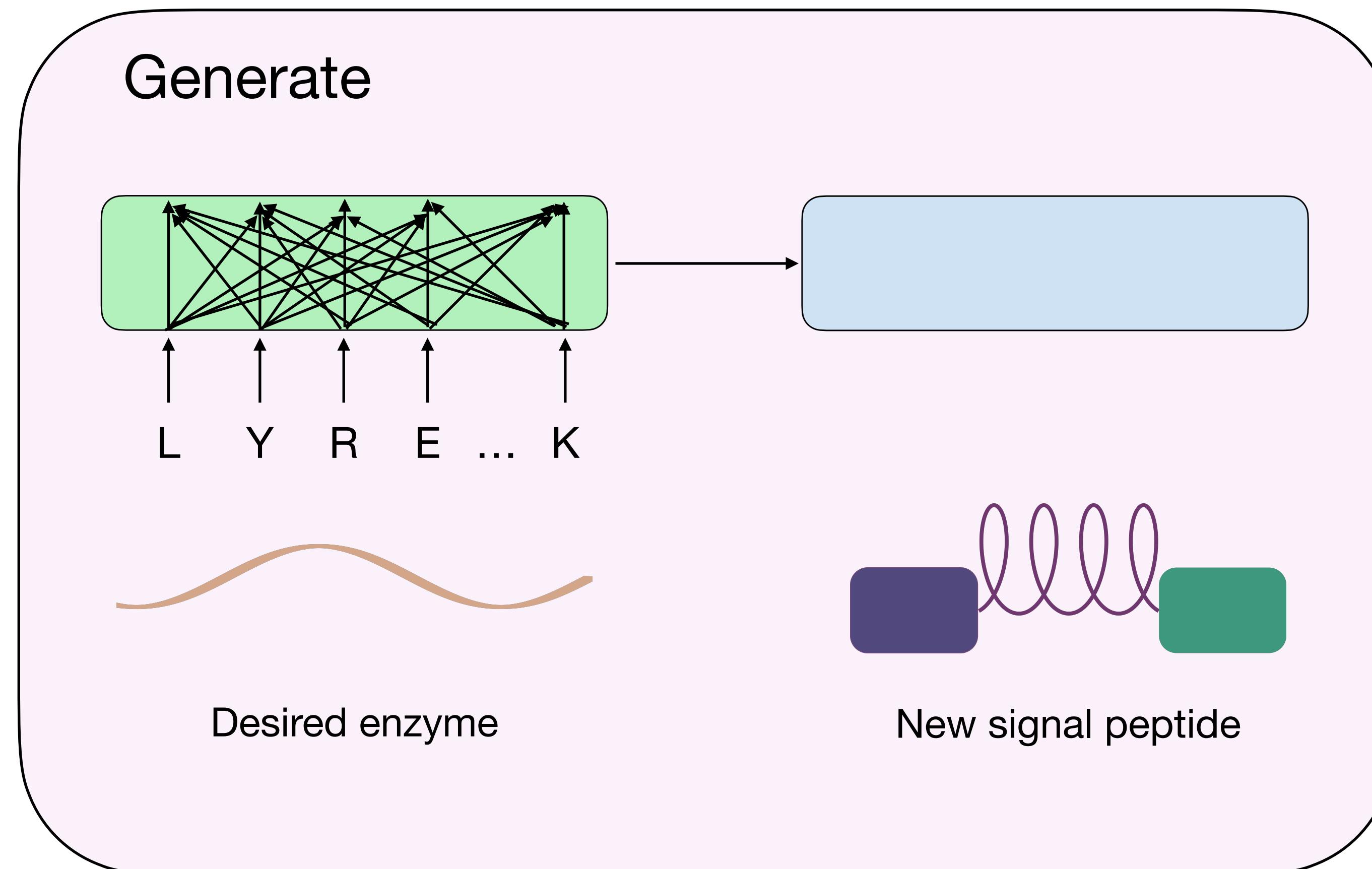
Use machine translation to generate SPs



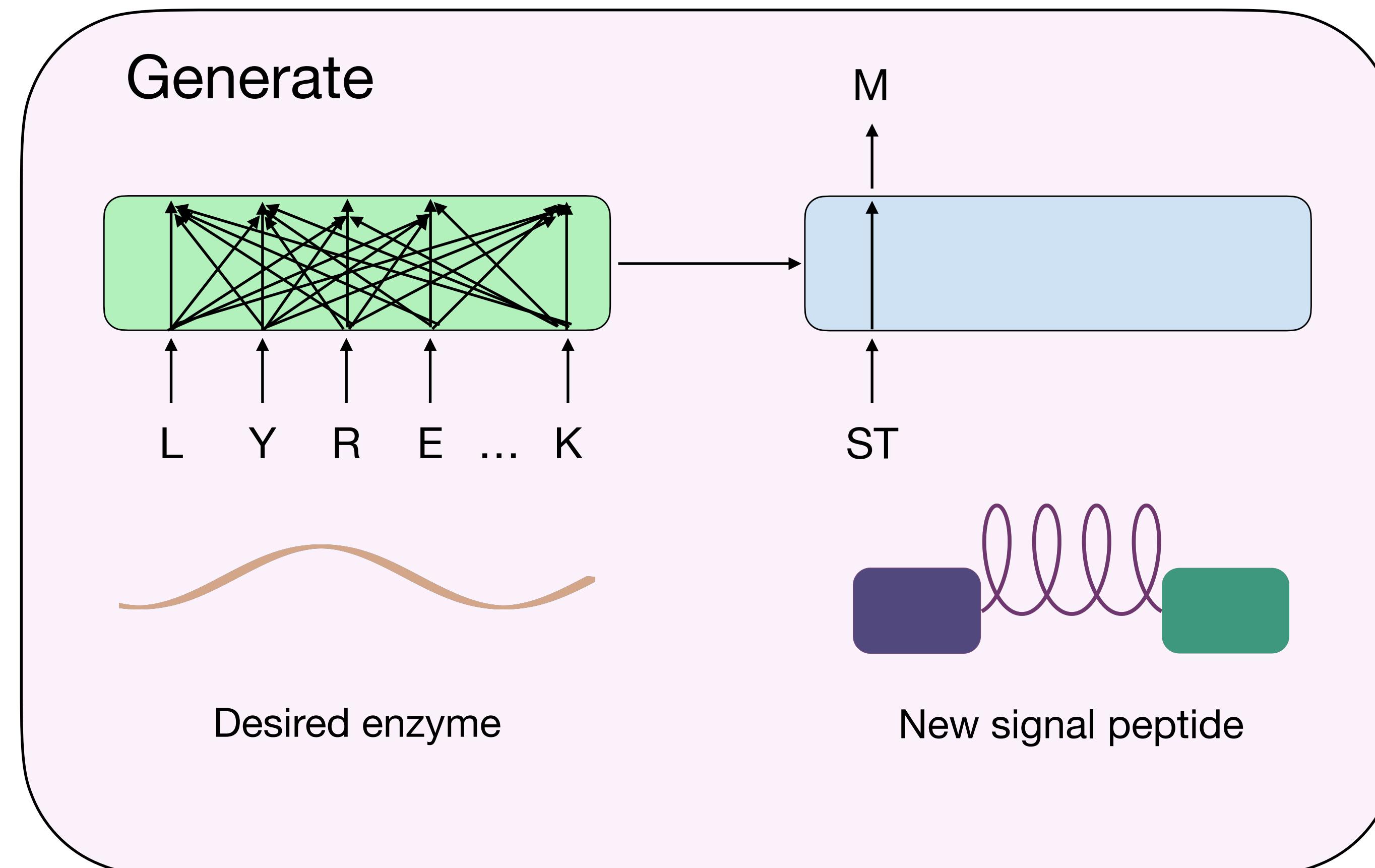
Use machine translation to generate SPs



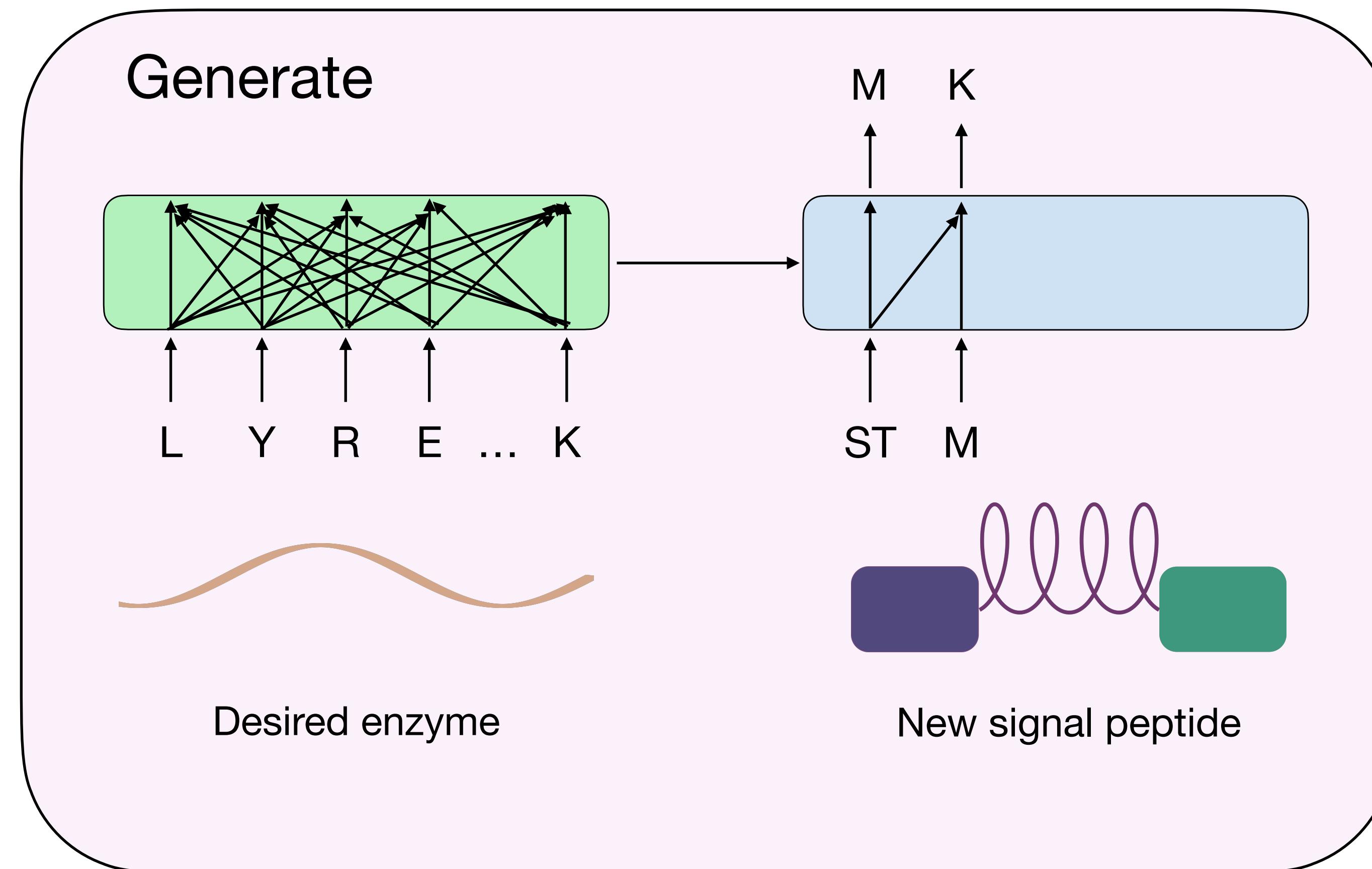
Use machine translation to generate SPs



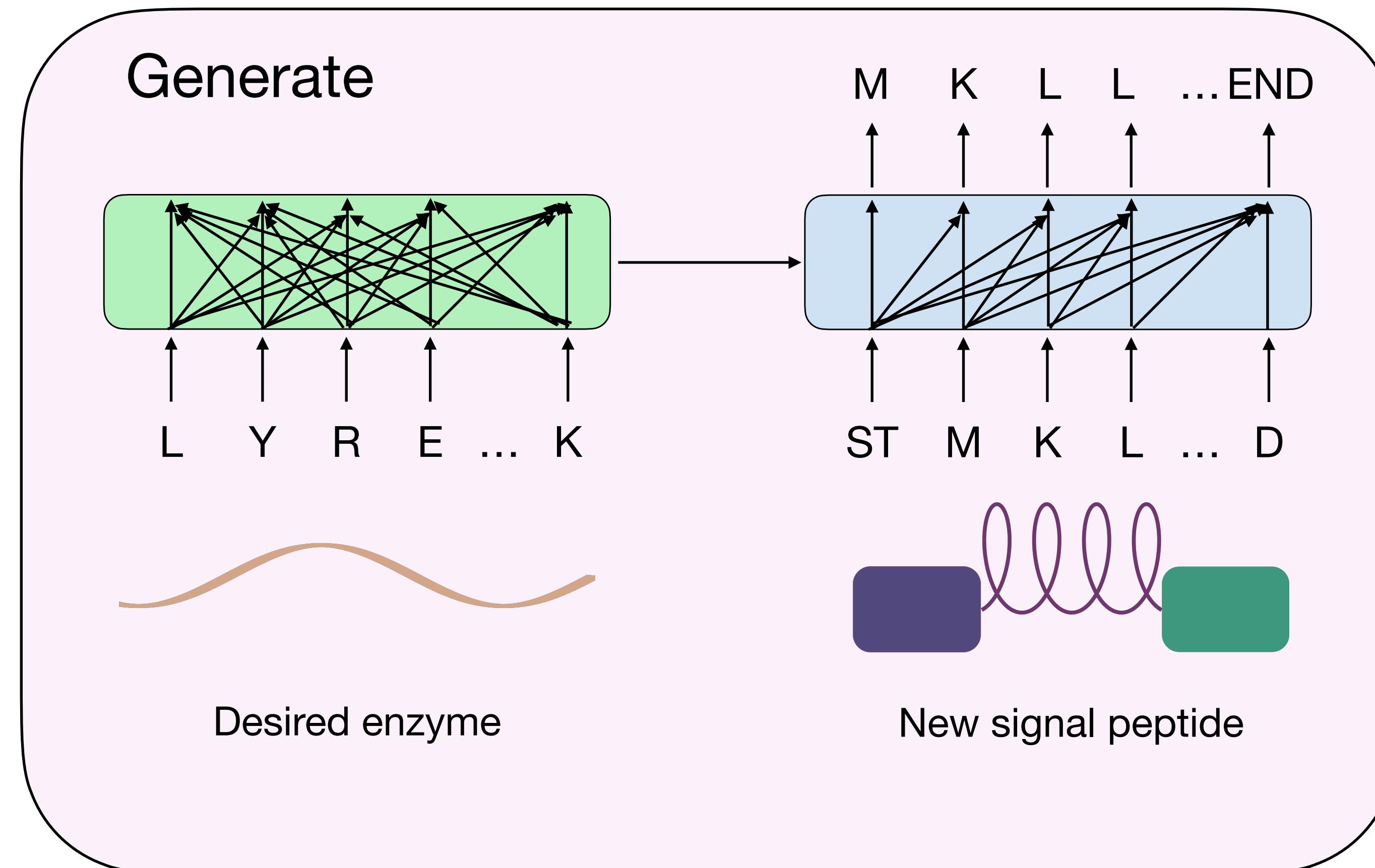
Use machine translation to generate SPs



Use machine translation to generate SPs

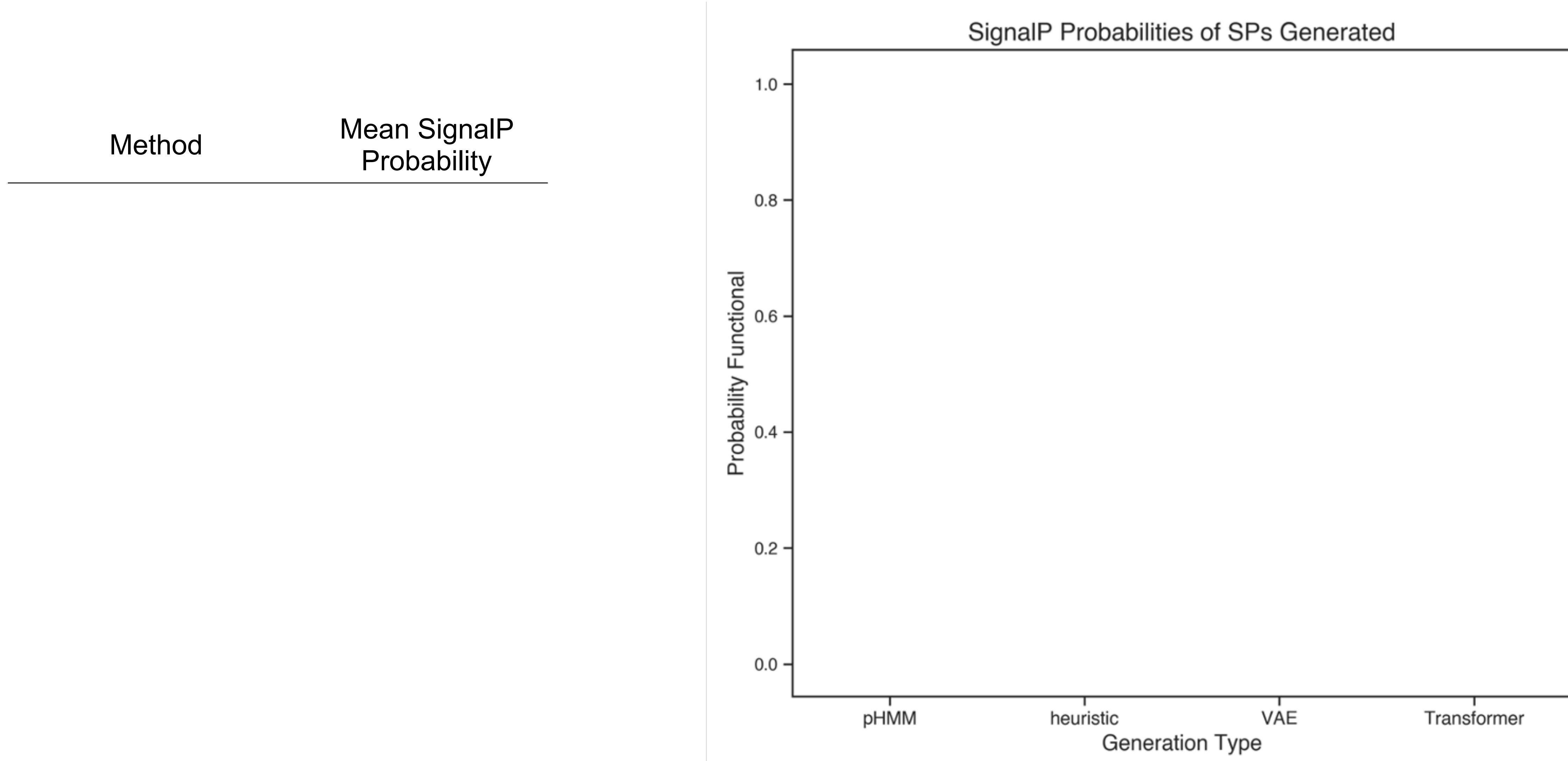


Use machine translation to generate SPs



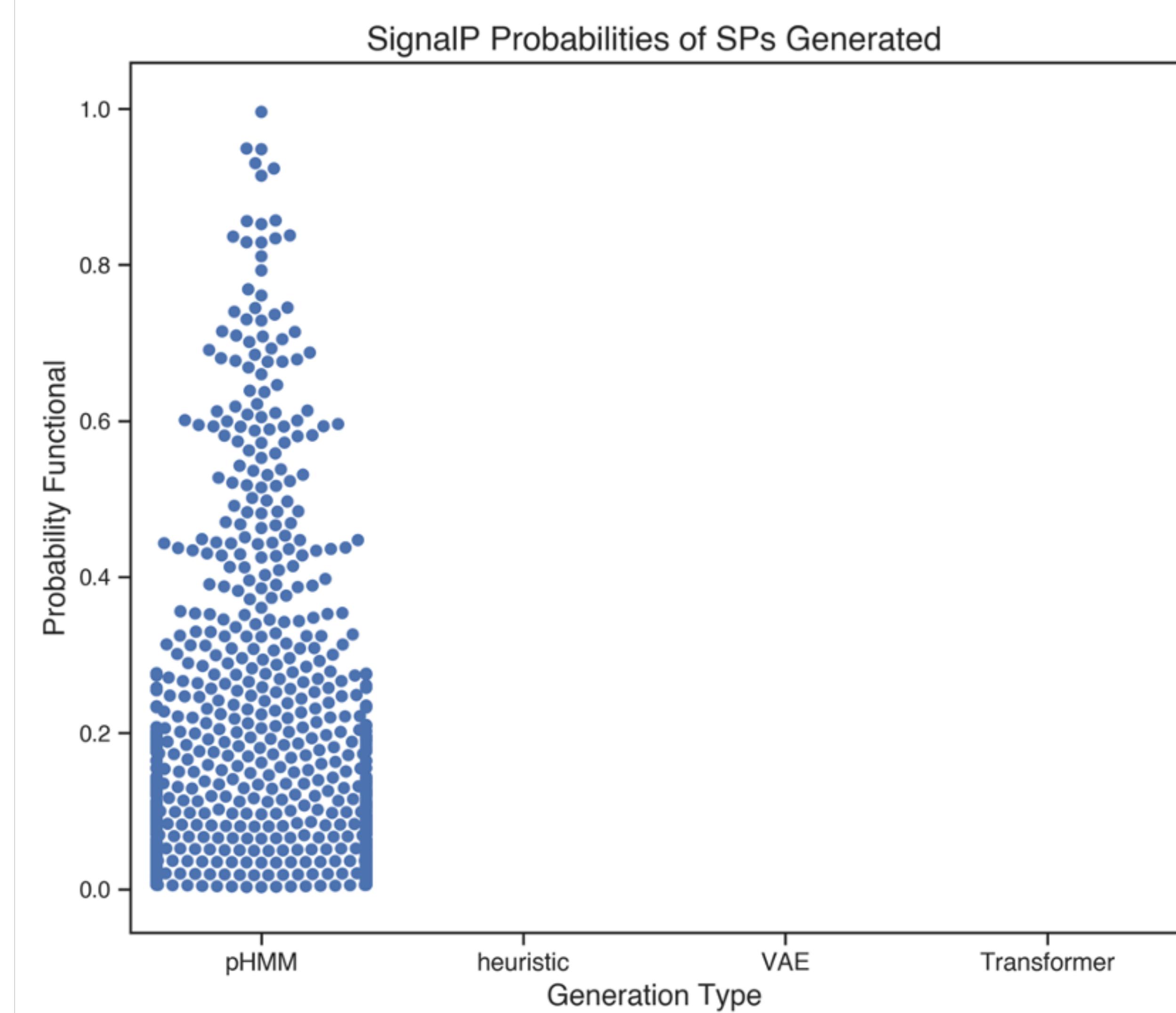
Generated SPs trick SignalP 5.0

Generated SPs trick SignalP 5.0



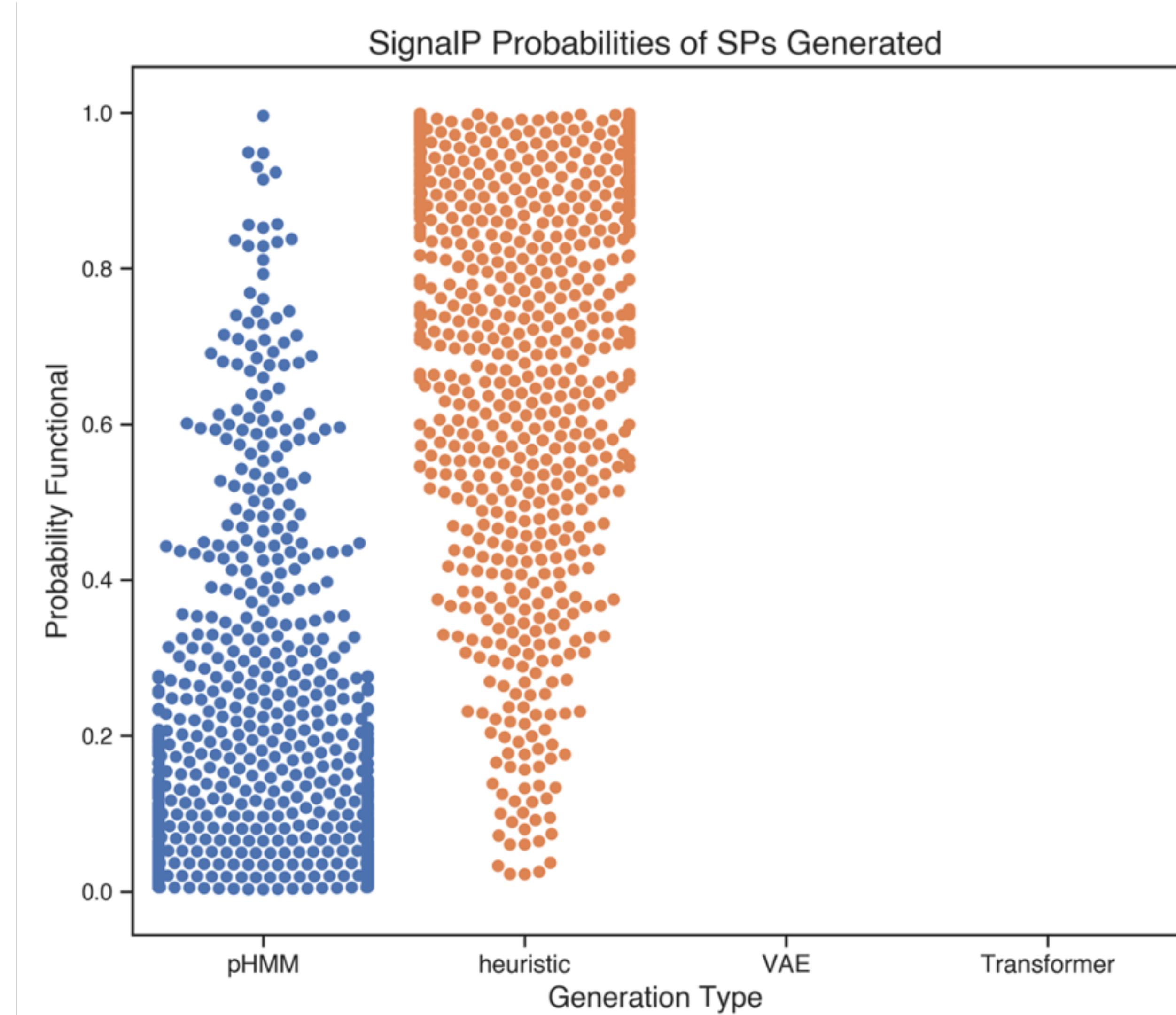
Generated SPs trick SignalP 5.0

Method	Mean SignalP Probability
Hidden Markov Model	16% \pm 19%



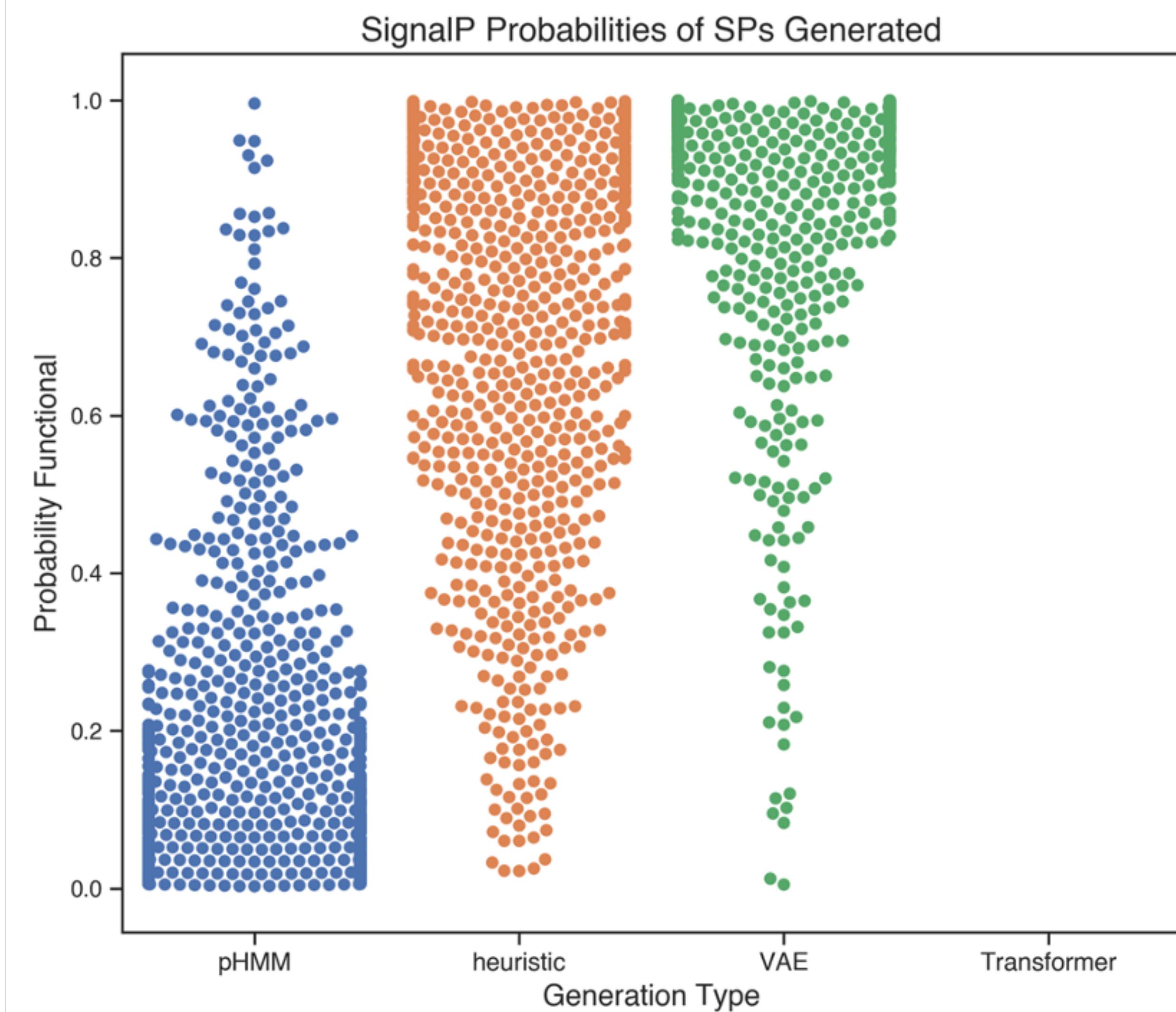
Generated SPs trick SignalP 5.0

Method	Mean SignalP Probability
Hidden Markov Model	16% \pm 19%
heuristic	71% \pm 25%



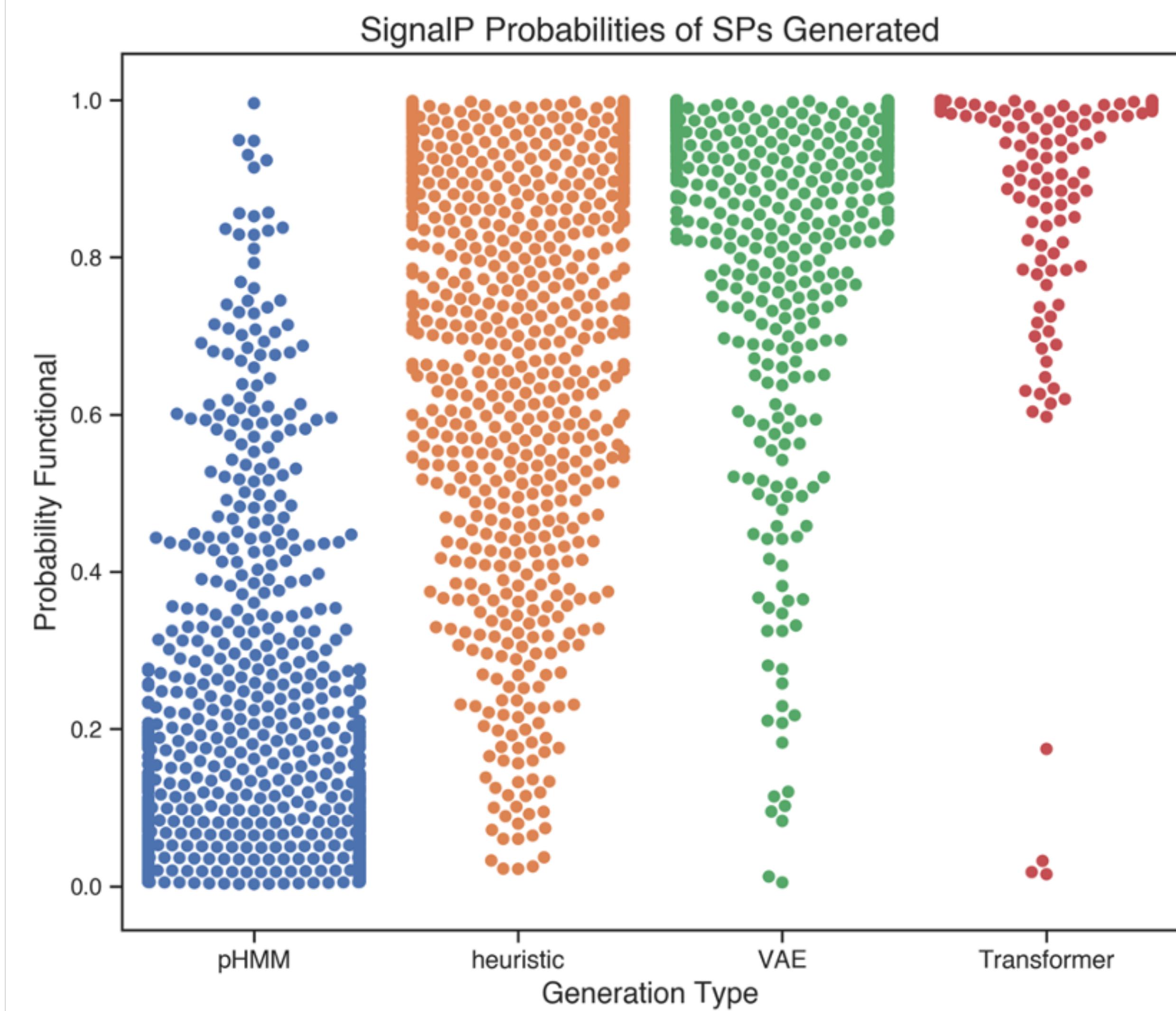
Generated SPs trick SignalP 5.0

Method	Mean SignalP Probability
Hidden Markov Model	16% \pm 19%
heuristic	71% \pm 25%
Variational Autoencoder (VAE)	92% \pm 15%



Generated SPs trick SignalP 5.0

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

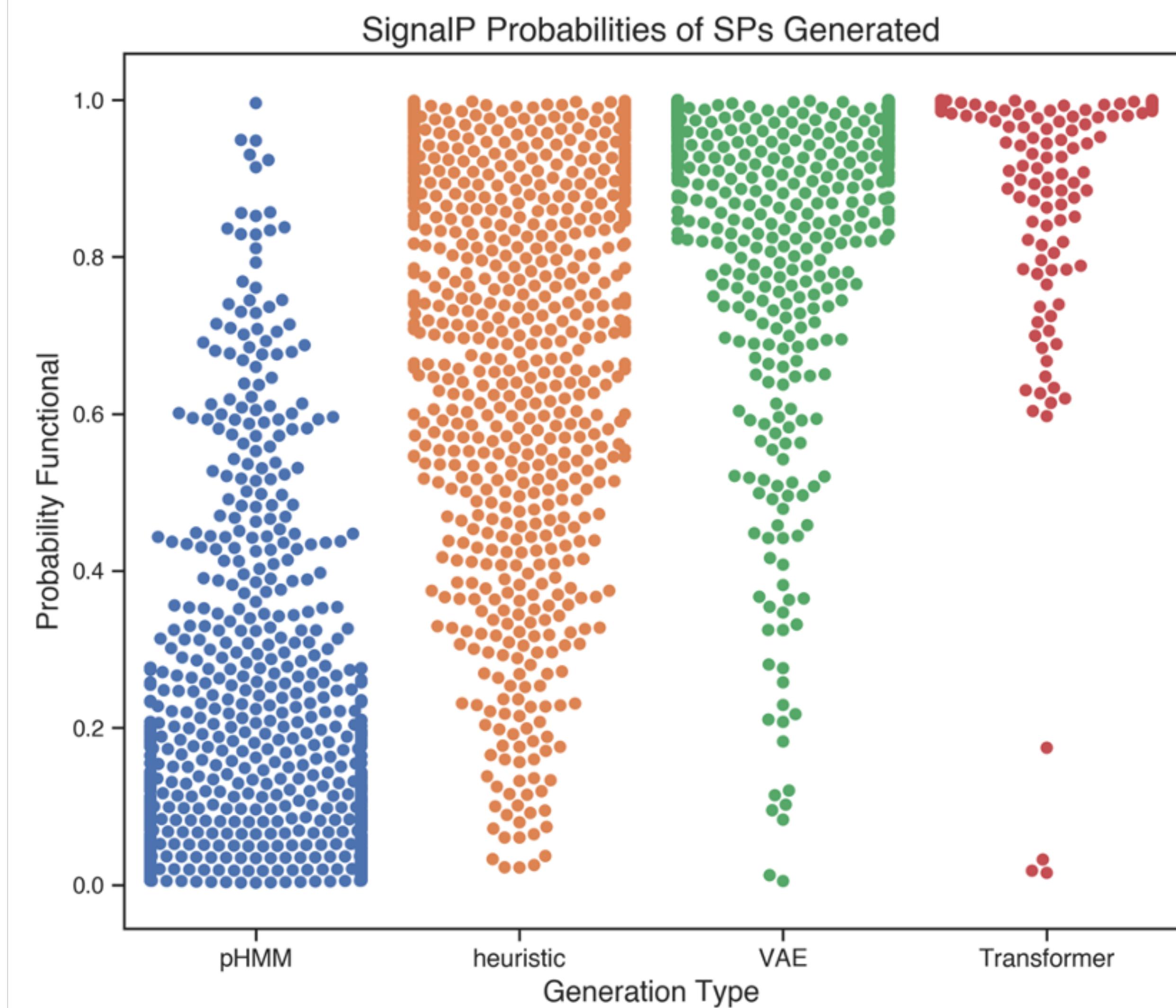


Generated SPs trick SignalP 5.0

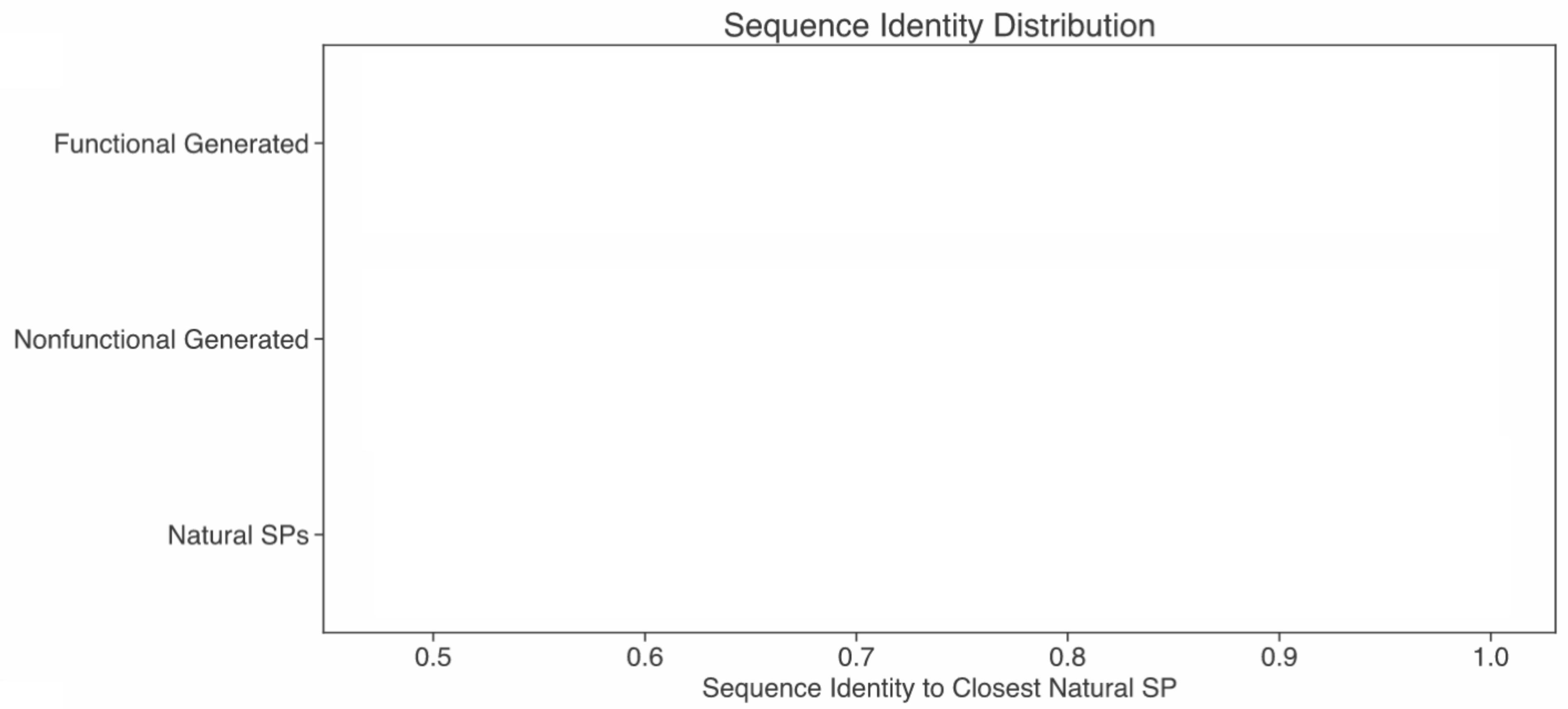
Method	Mean SignalP Probability
Hidden Markov Model	16% \pm 19%
heuristic	71% \pm 25%
Variational Autoencoder (VAE)	92% \pm 15%
Transformer	90% \pm 17%

VAE sequences are repetitive

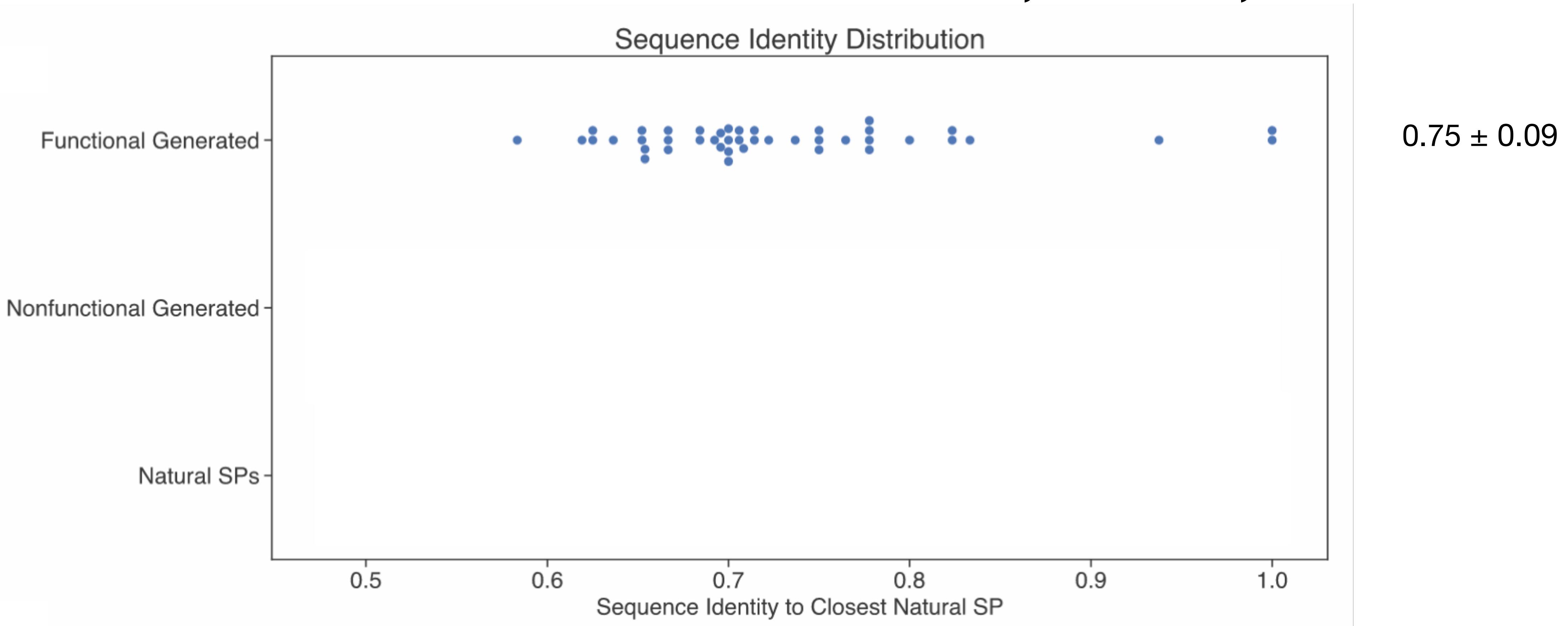
MRKRLLALALALAALSLLLLSFGVKALAGSGA
MRLLLLLLLVLLLAAAPAPPGLS
MKLLLLLVTLTSLVALQAA
MRLLLLALLAAAAVALASA
MASSSSSLFVVVLAVLLLLLTLSSA



Generated SPs are functional, novel, and diverse



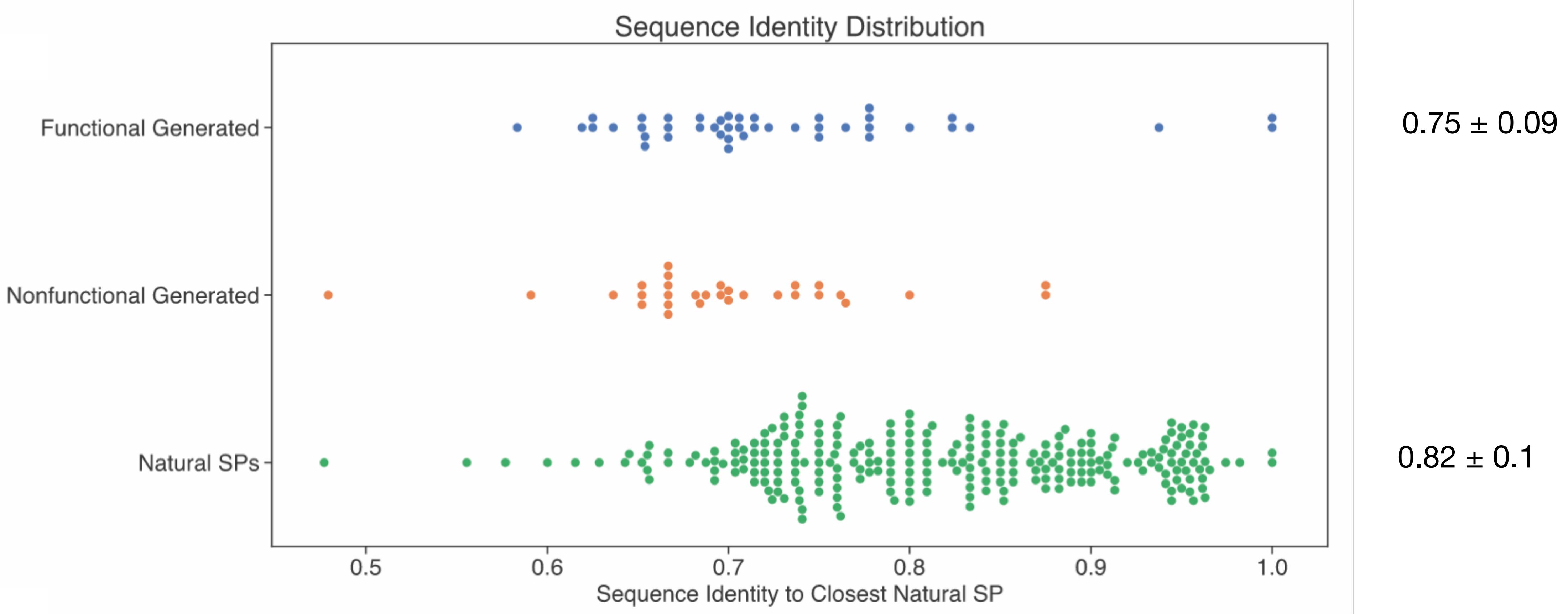
Generated SPs are functional, novel, and diverse



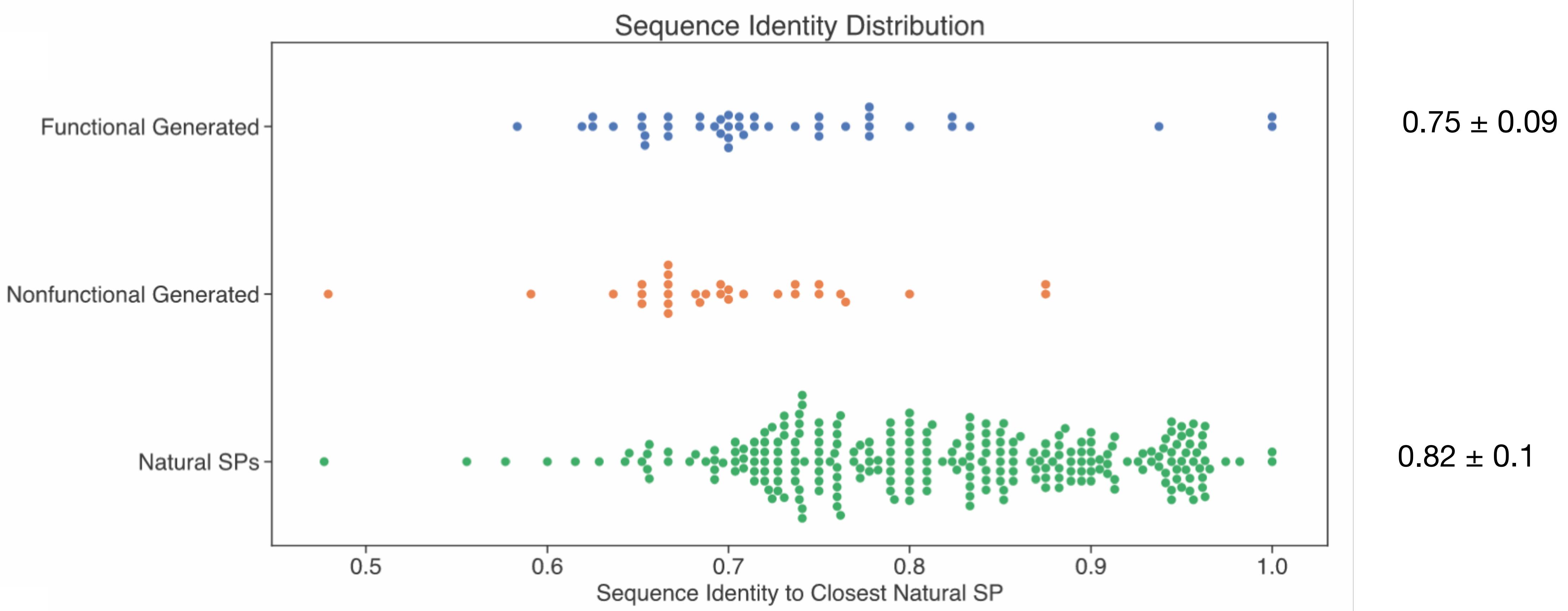
Generated SPs are functional, novel, and diverse



Generated SPs are functional, novel, and diverse

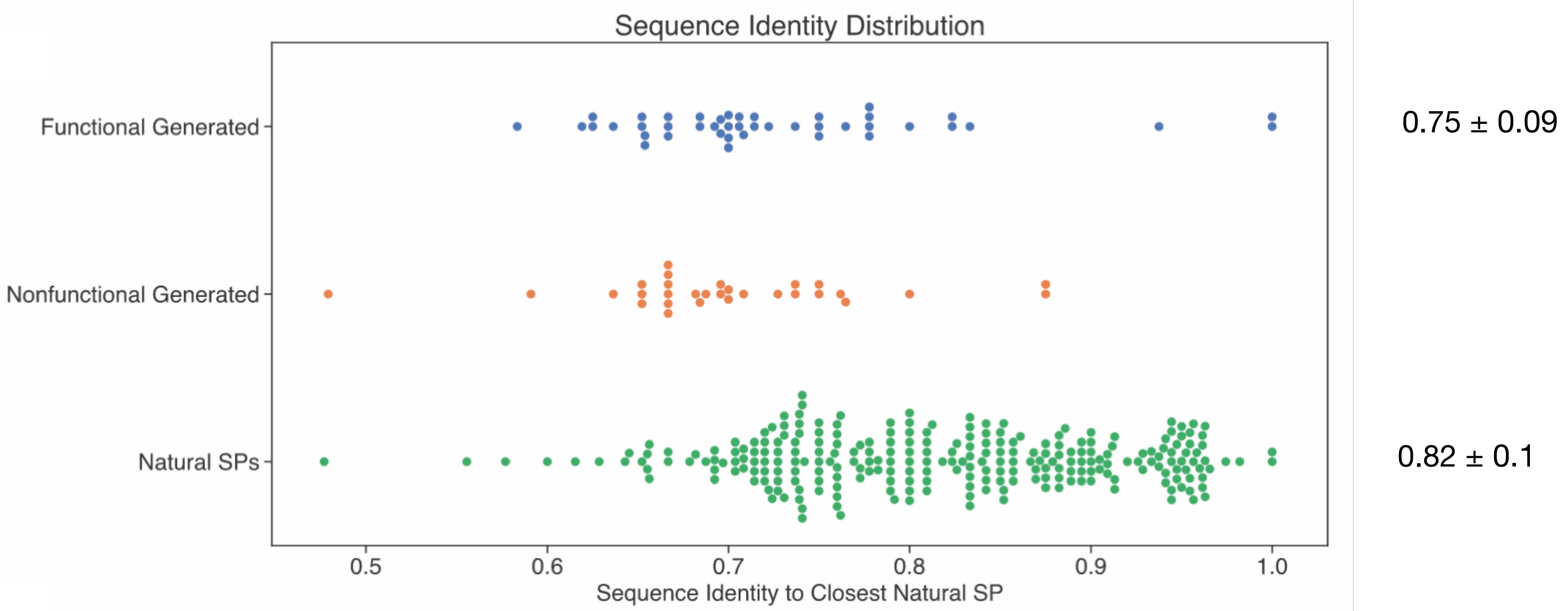


Generated SPs are functional, novel, and diverse



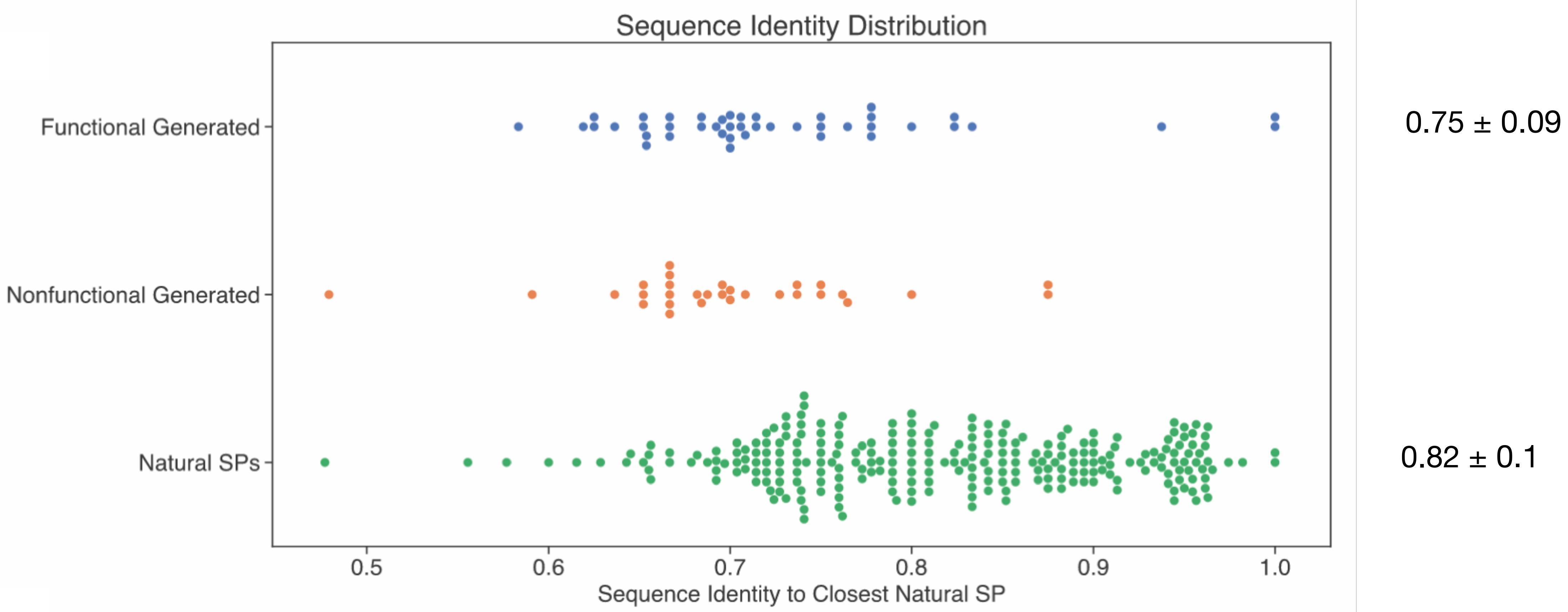
- Can generate *de novo* signal peptides

Generated SPs are functional, novel, and diverse



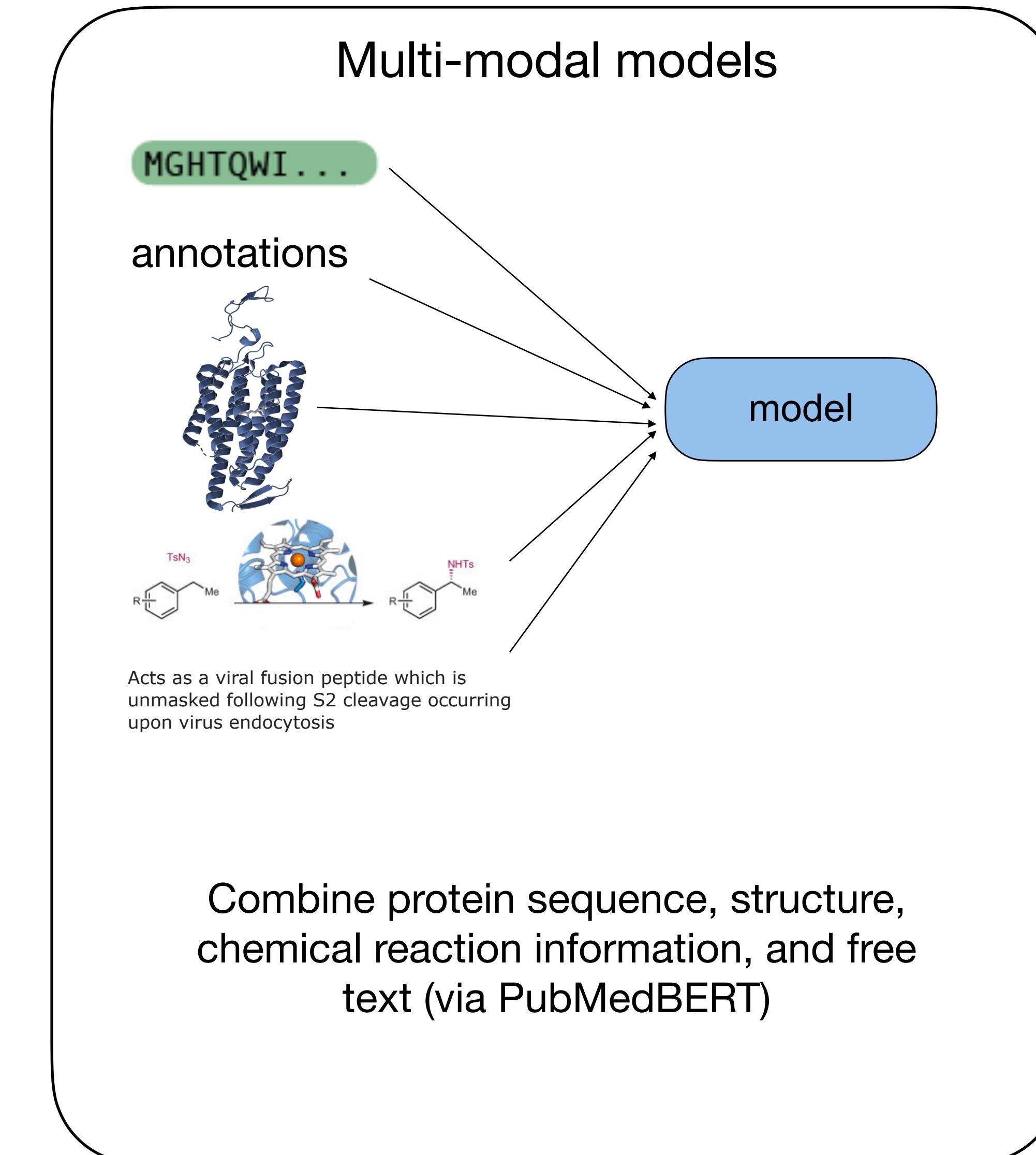
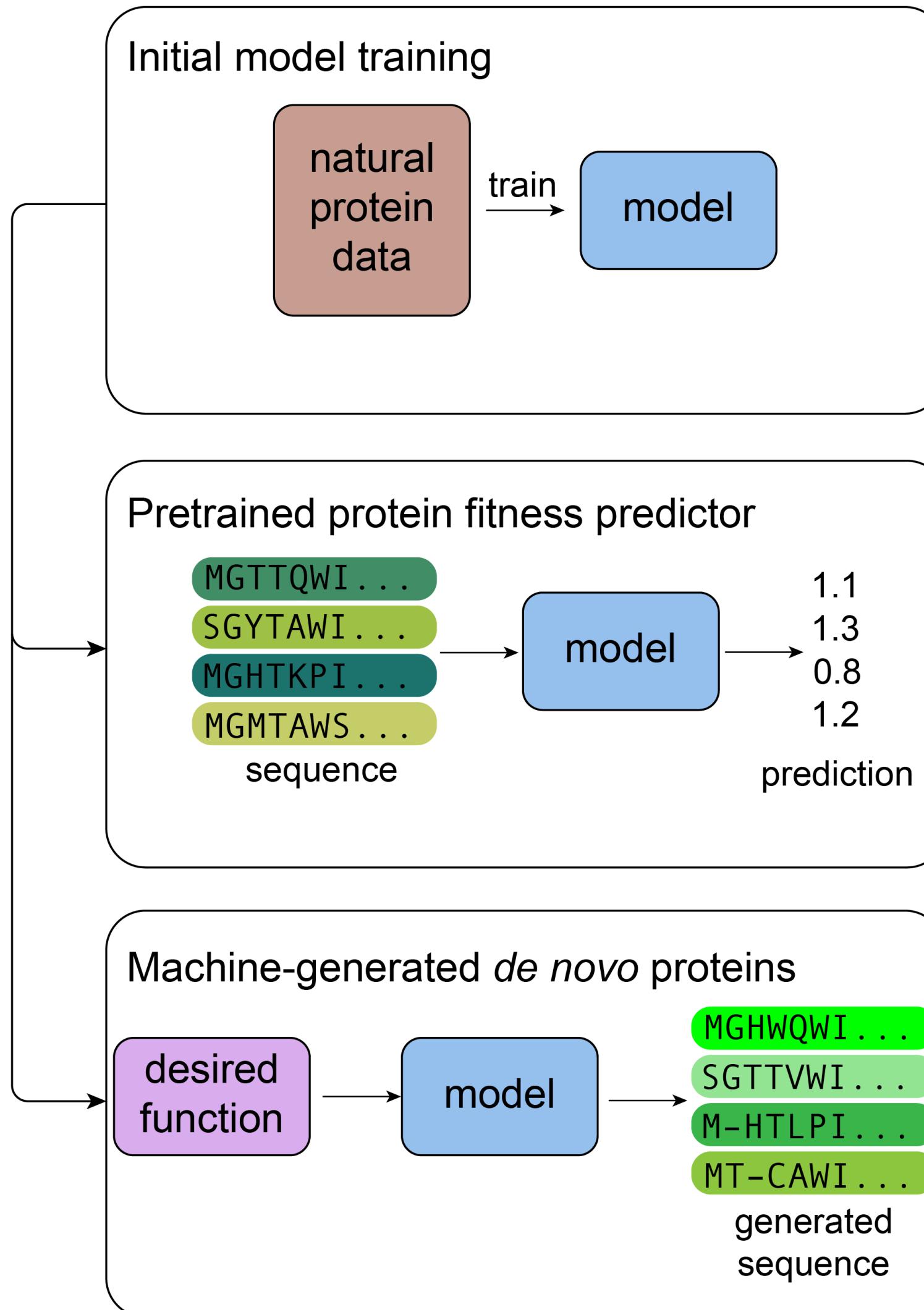
- Can generate *de novo* signal peptides
- That look like real SPs

Generated SPs are functional, novel, and diverse

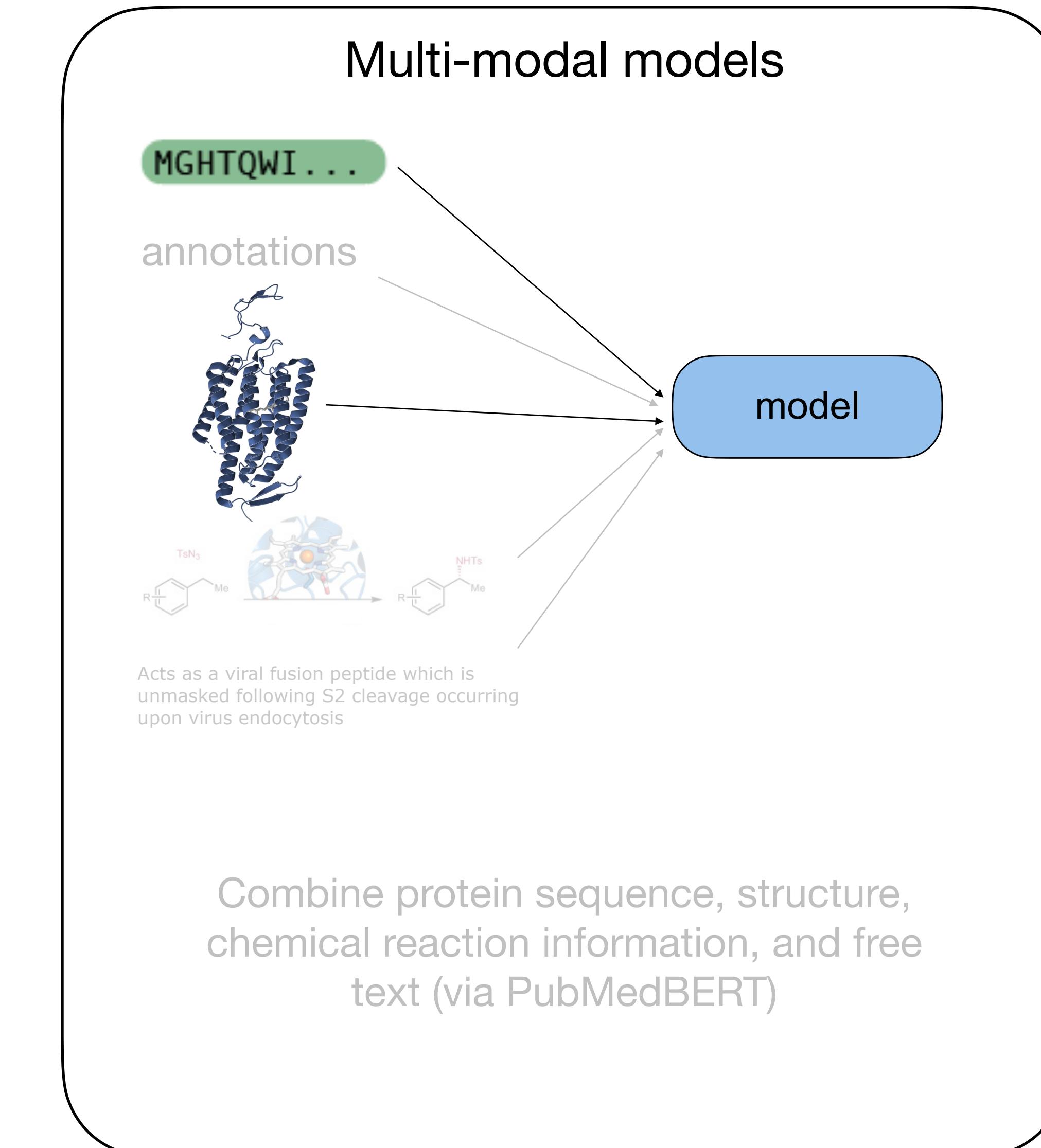
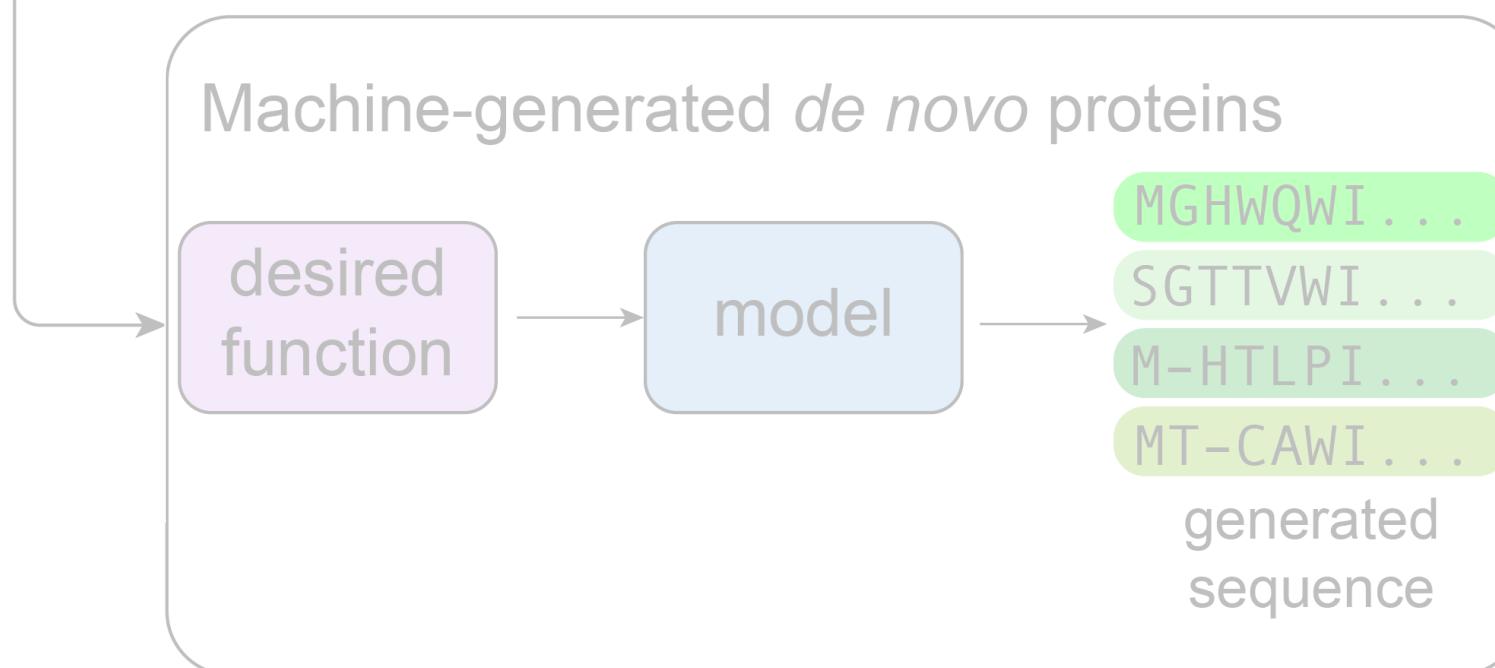
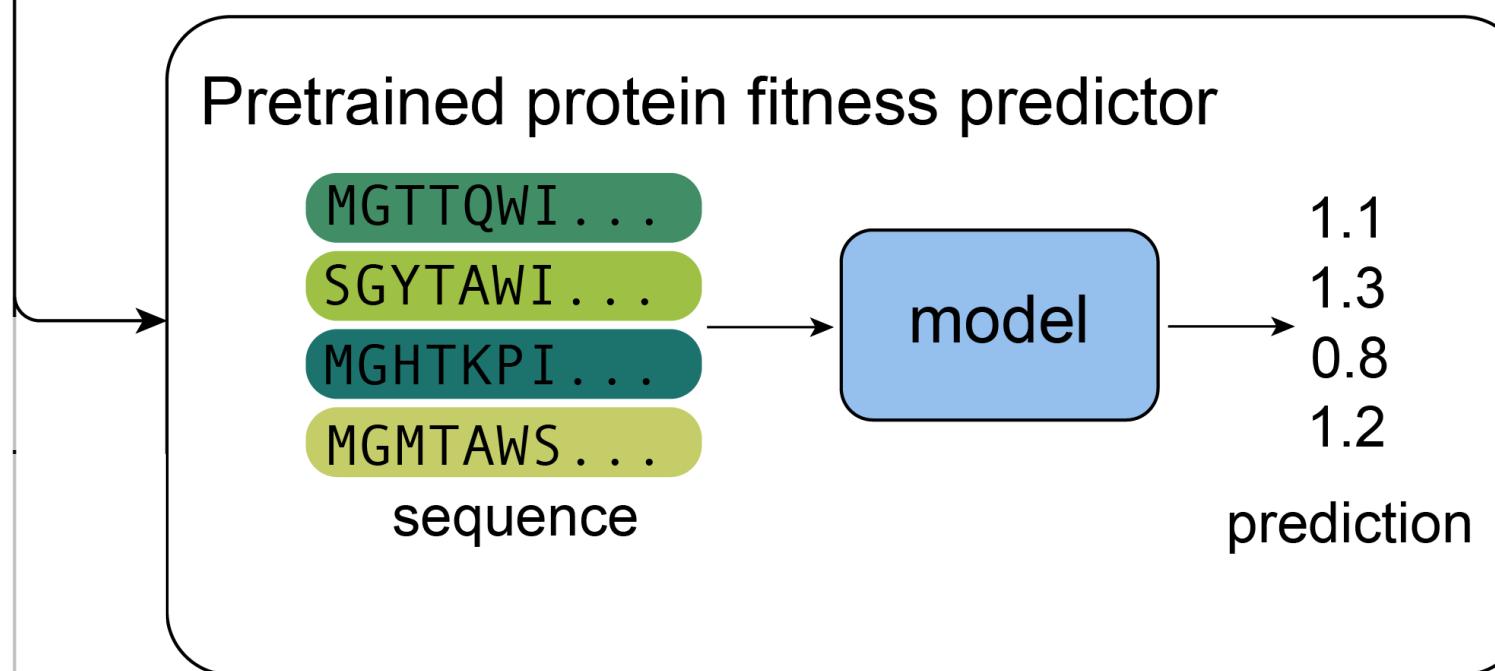
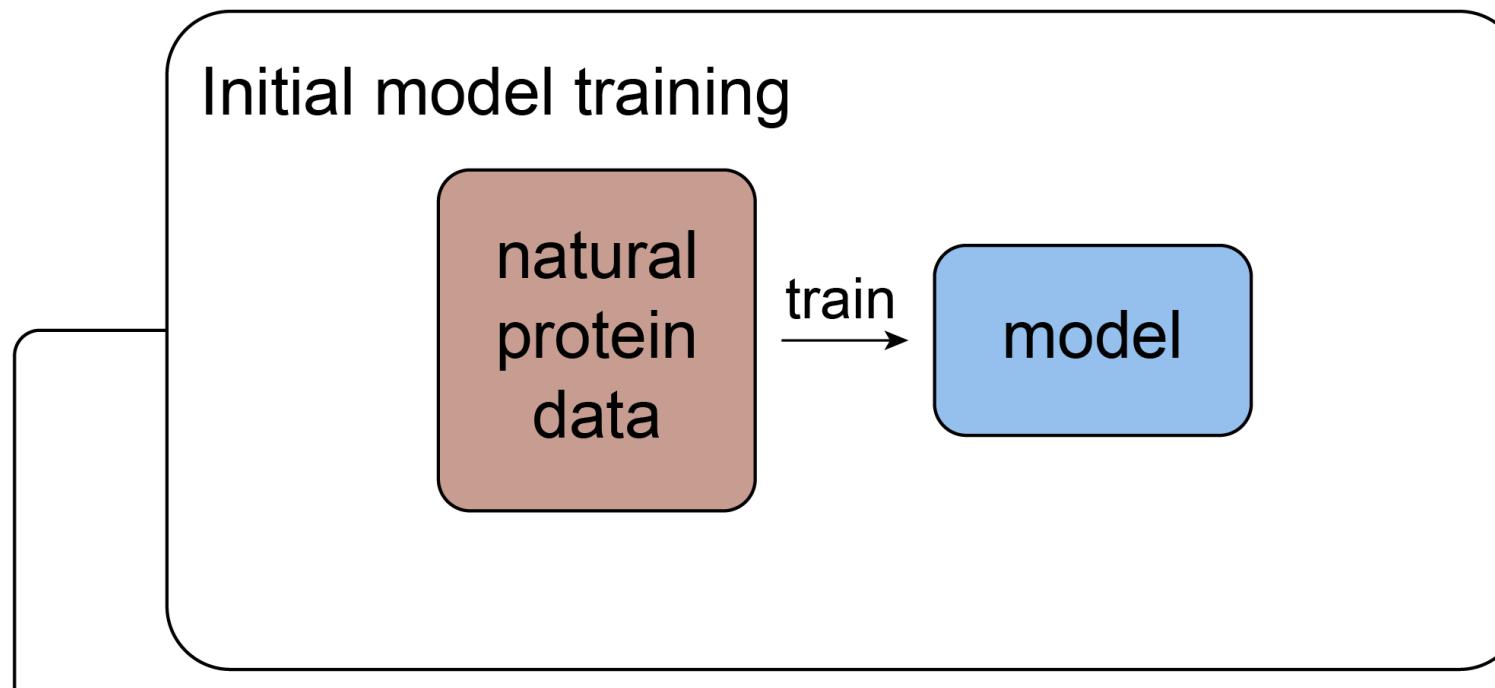


- Can generate *de novo* signal peptides
- That look like real SPs
- And result in functional secreted enzymes

Use multiple data modalities to design proteins



Use multiple data modalities to design proteins



Many methods pretend proteins are language

MFTGNDAGH

Many methods pretend proteins are language

MFTGNDAGH

Treat amino acids like tokens

Many methods pretend proteins are language

MFTGNDAGH

Treat amino acids like tokens

Use tools originally developed for language

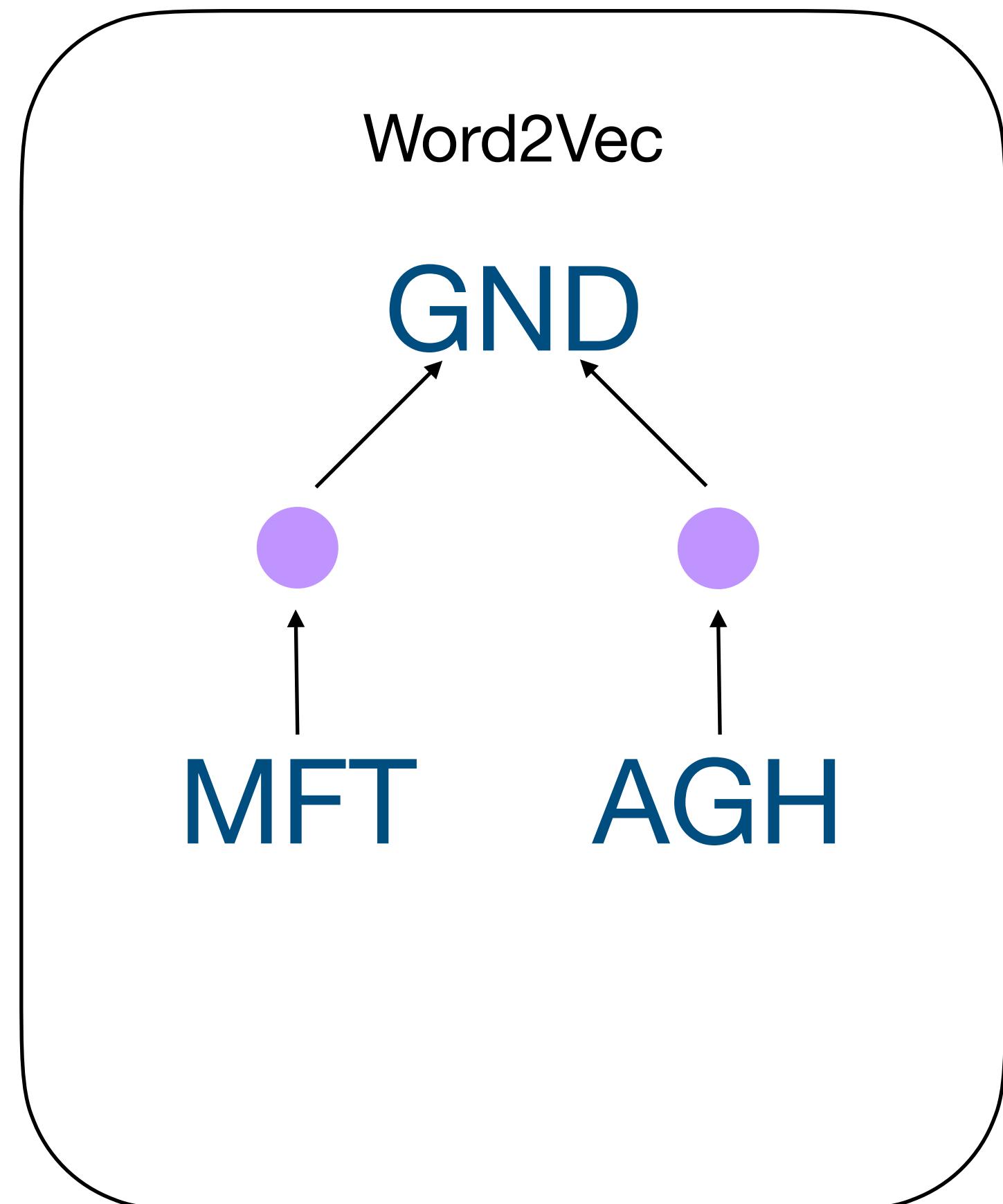
Many methods pretend proteins are language

Word2Vec

MFT AGH

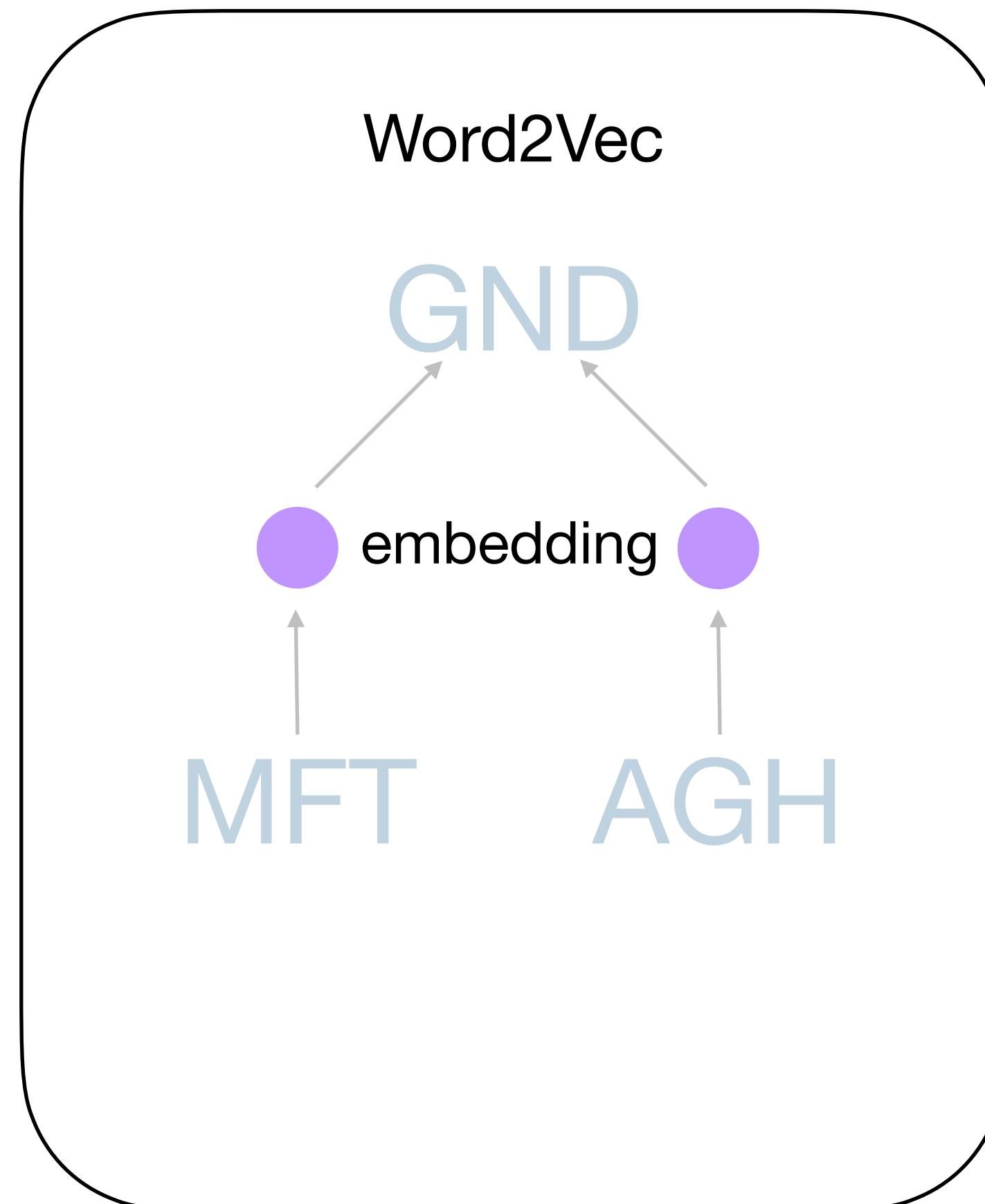
MFTGNDAGH

Many methods pretend proteins are language



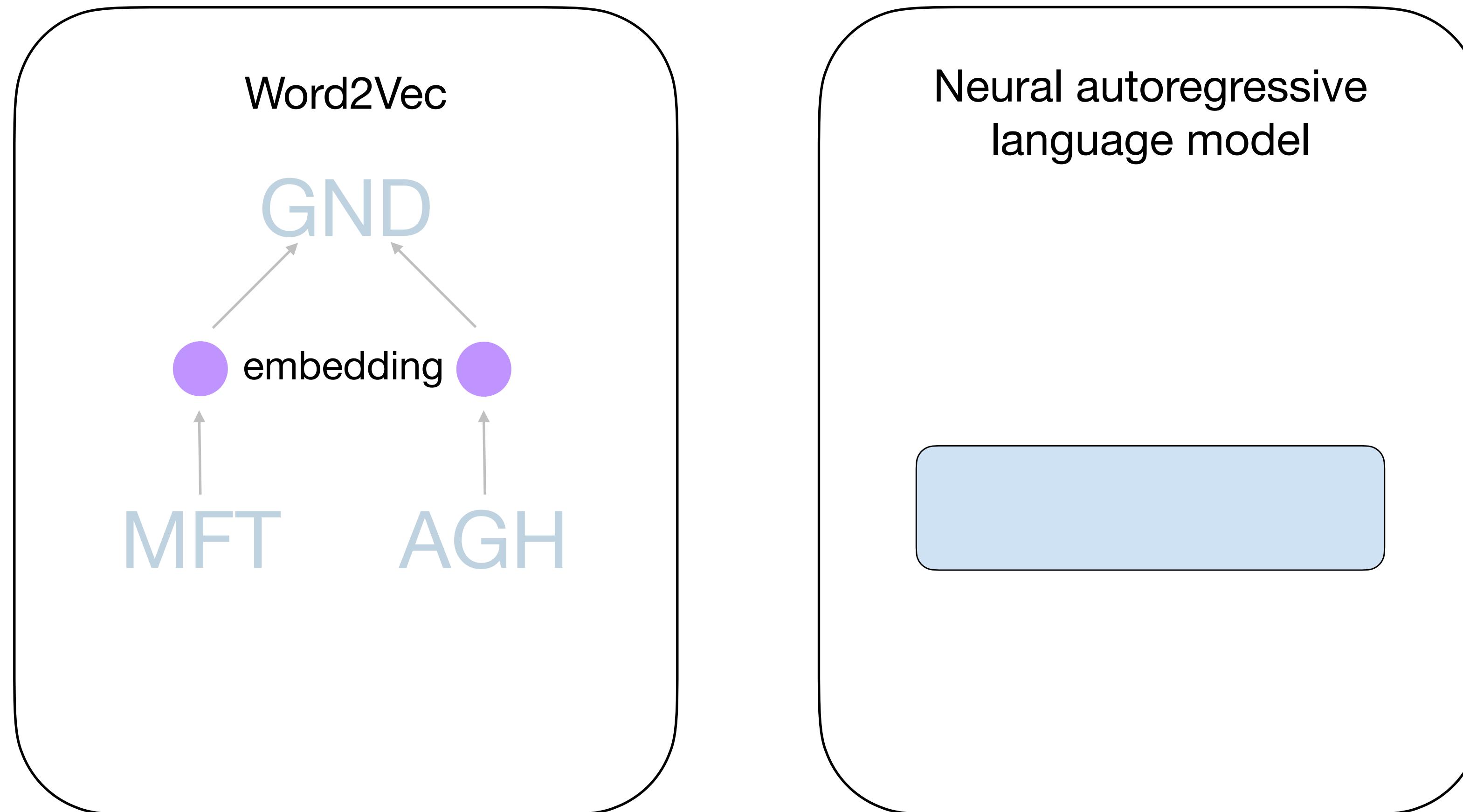
MFTGNDAGH

Many methods pretend proteins are language



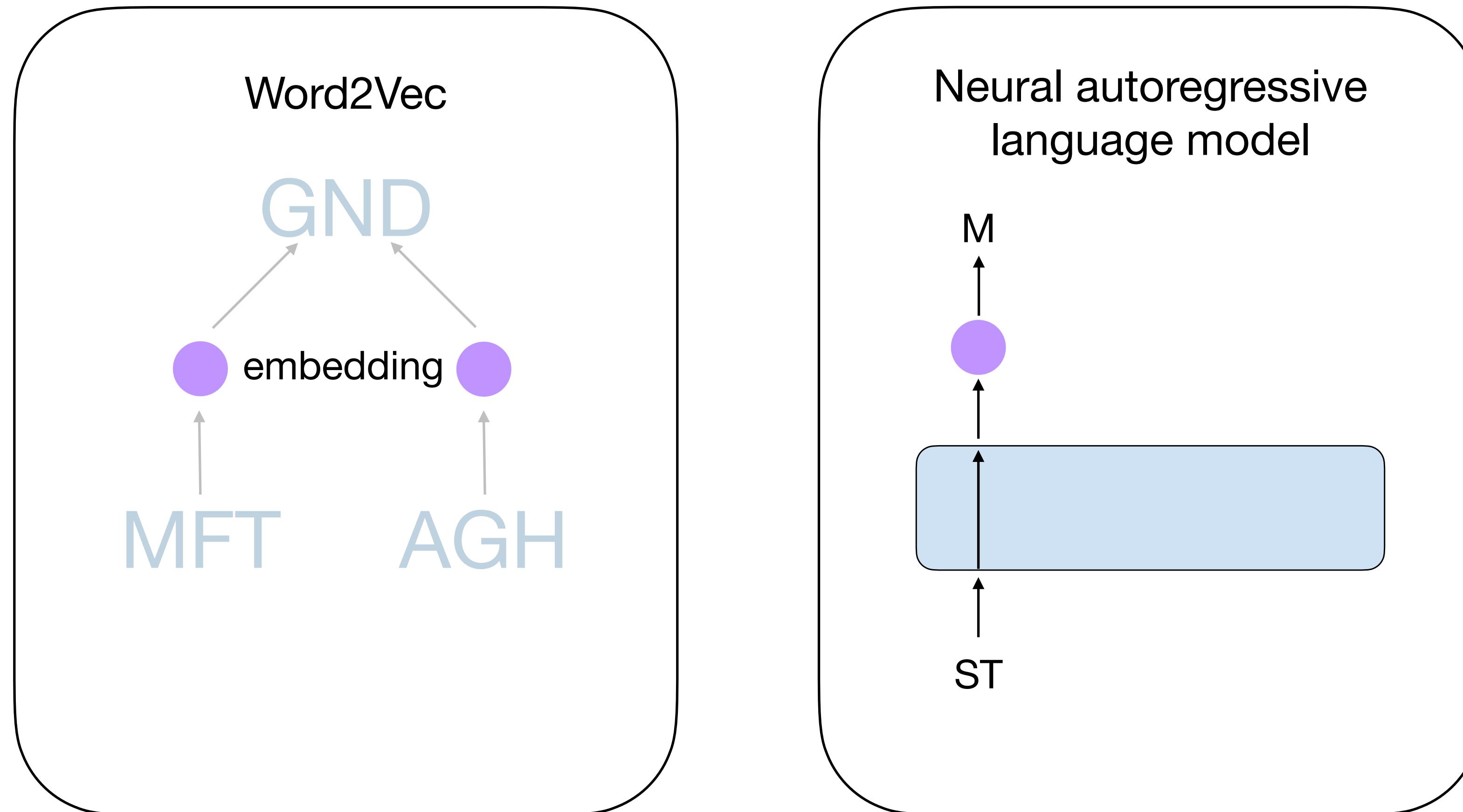
MFTGNDAGH

Many methods pretend proteins are language



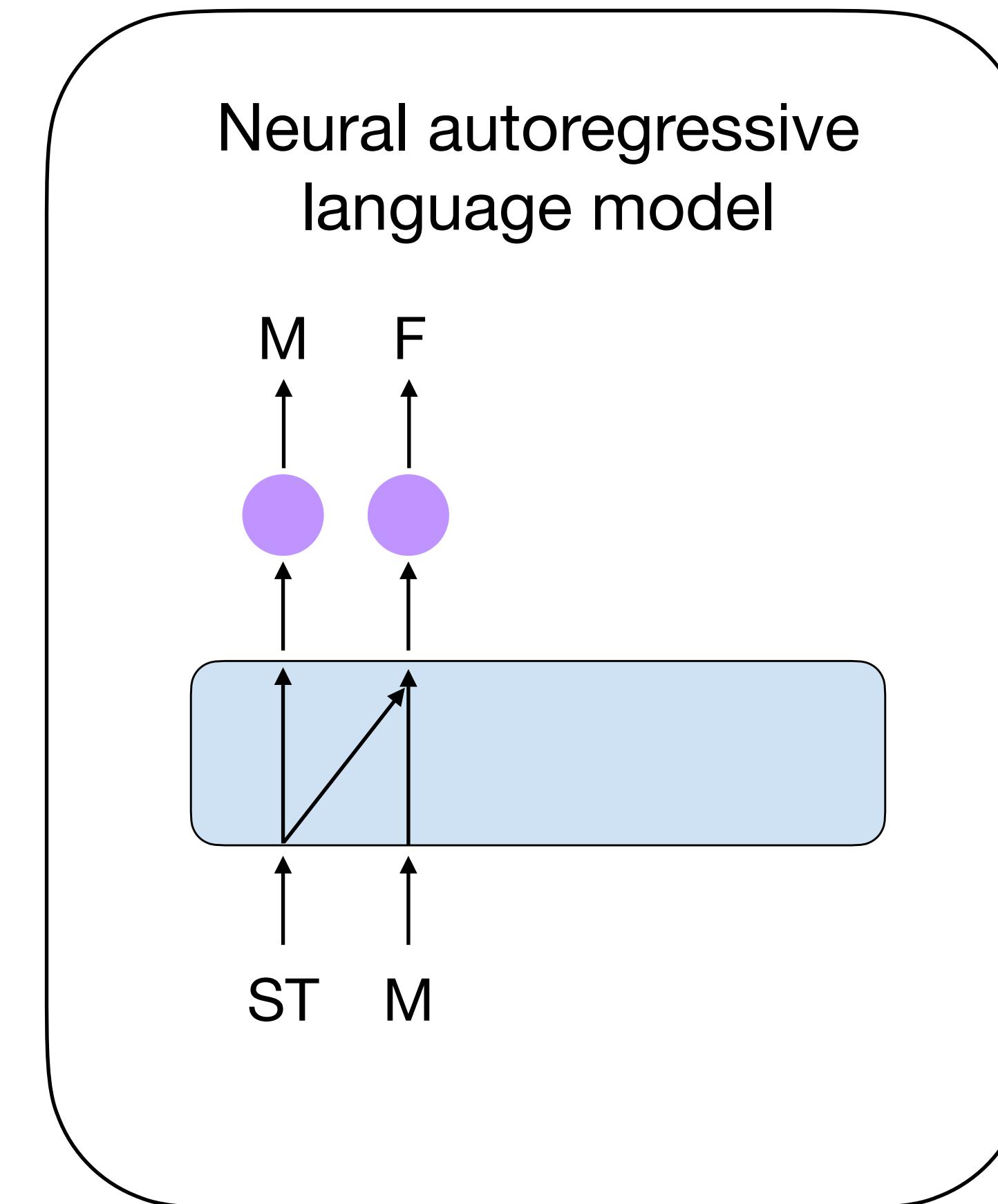
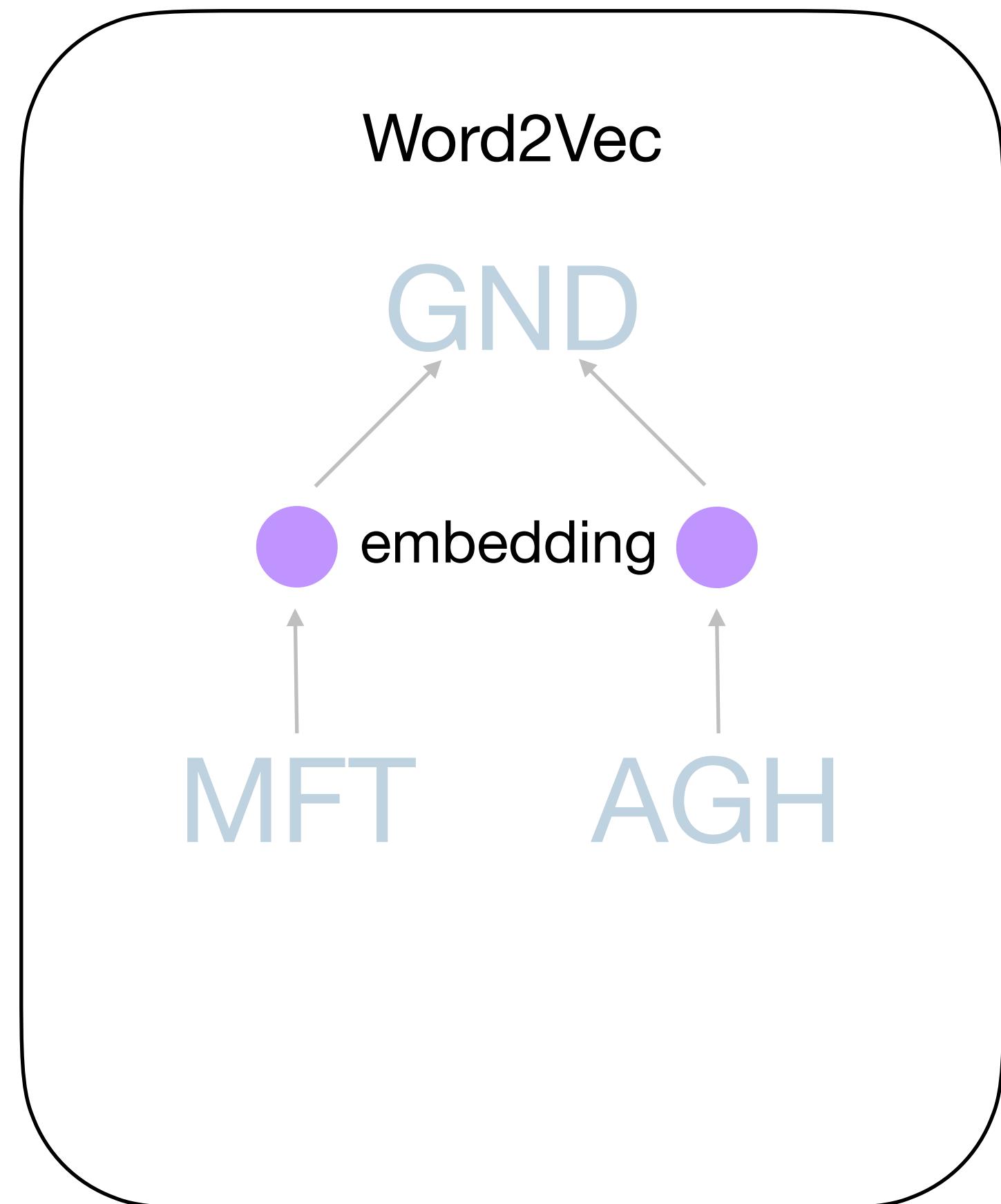
MFTGNDAGH

Many methods pretend proteins are language



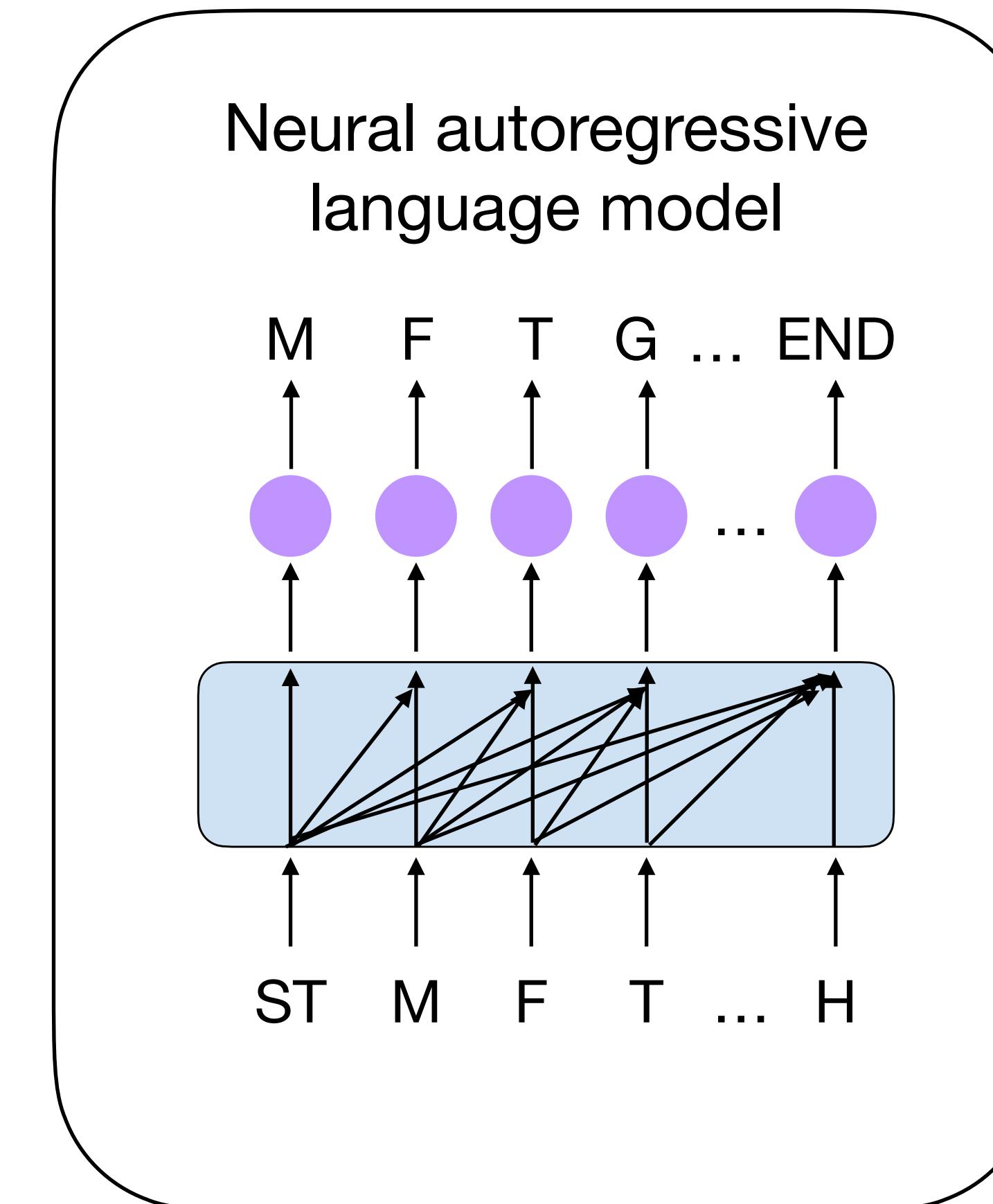
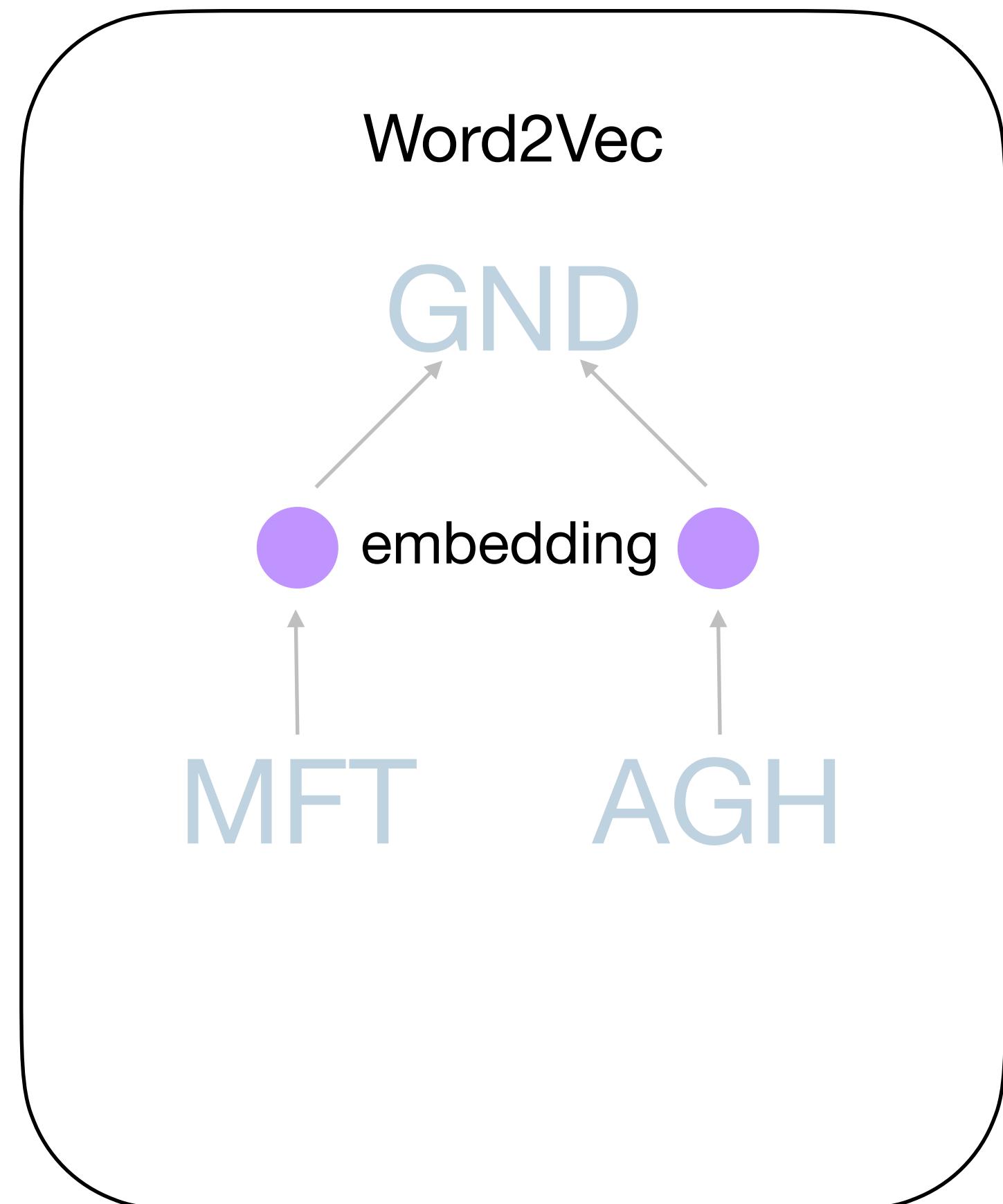
MFTGNDAGH

Many methods pretend proteins are language



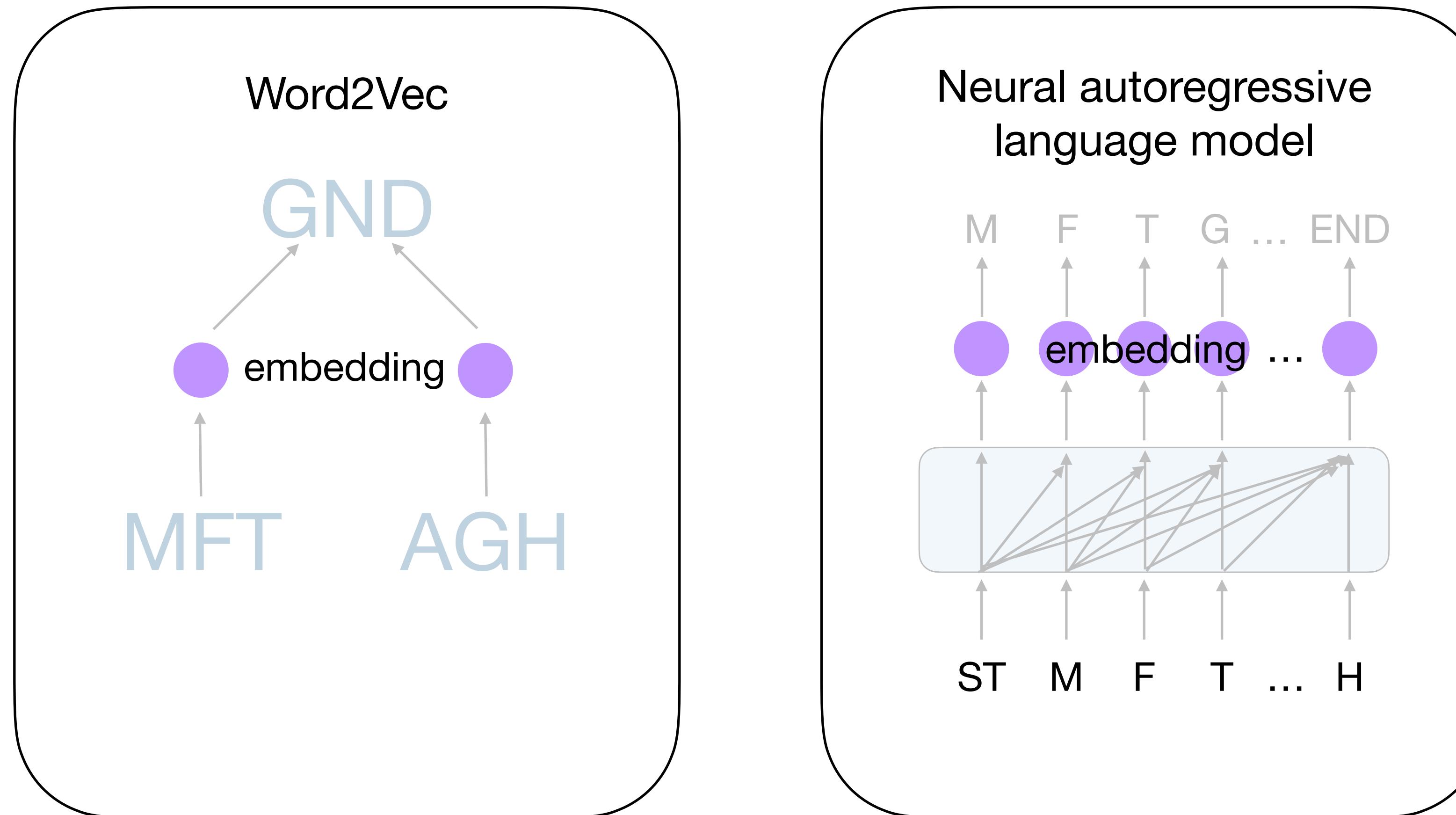
MFTGNDAGH

Many methods pretend proteins are language



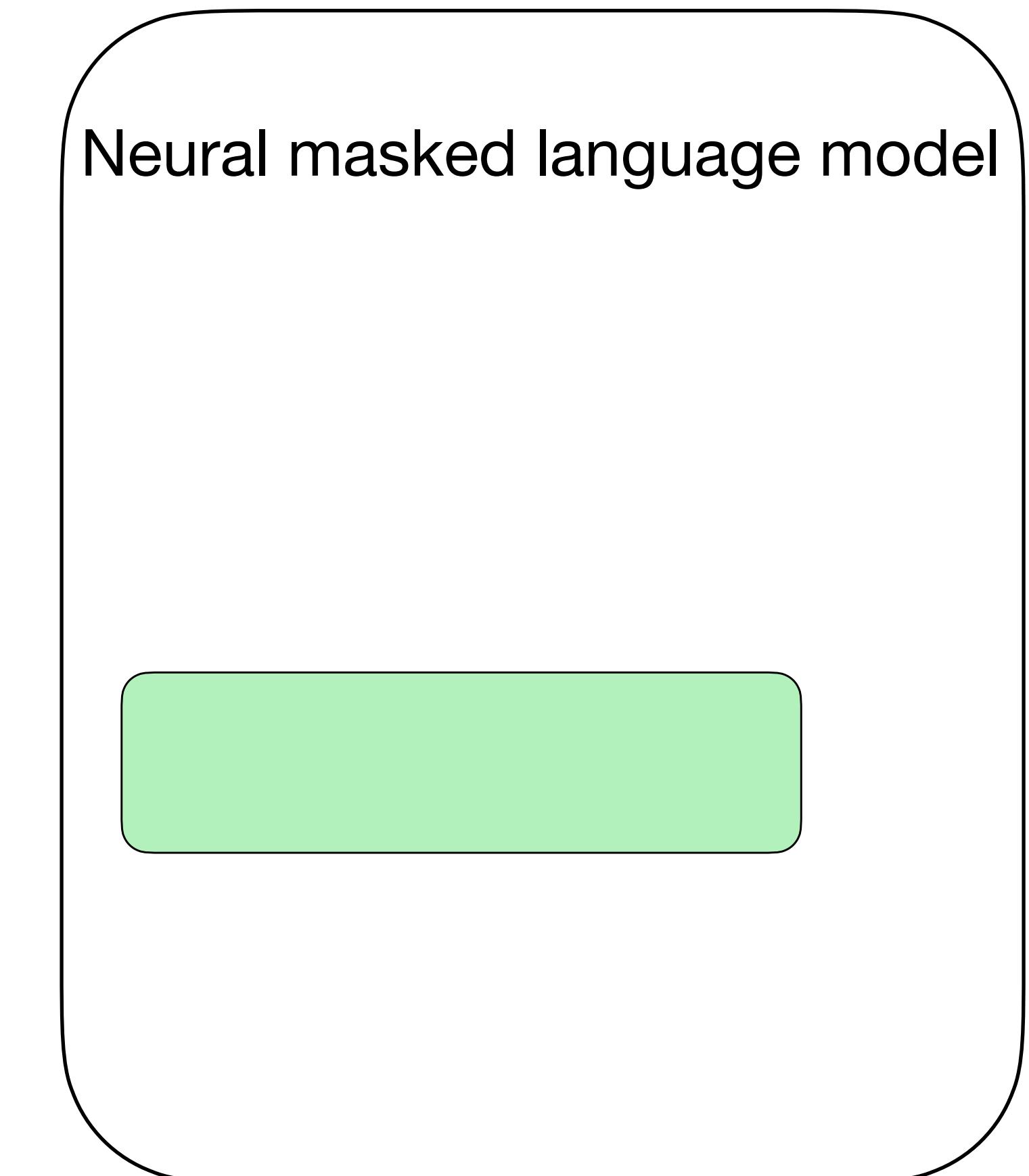
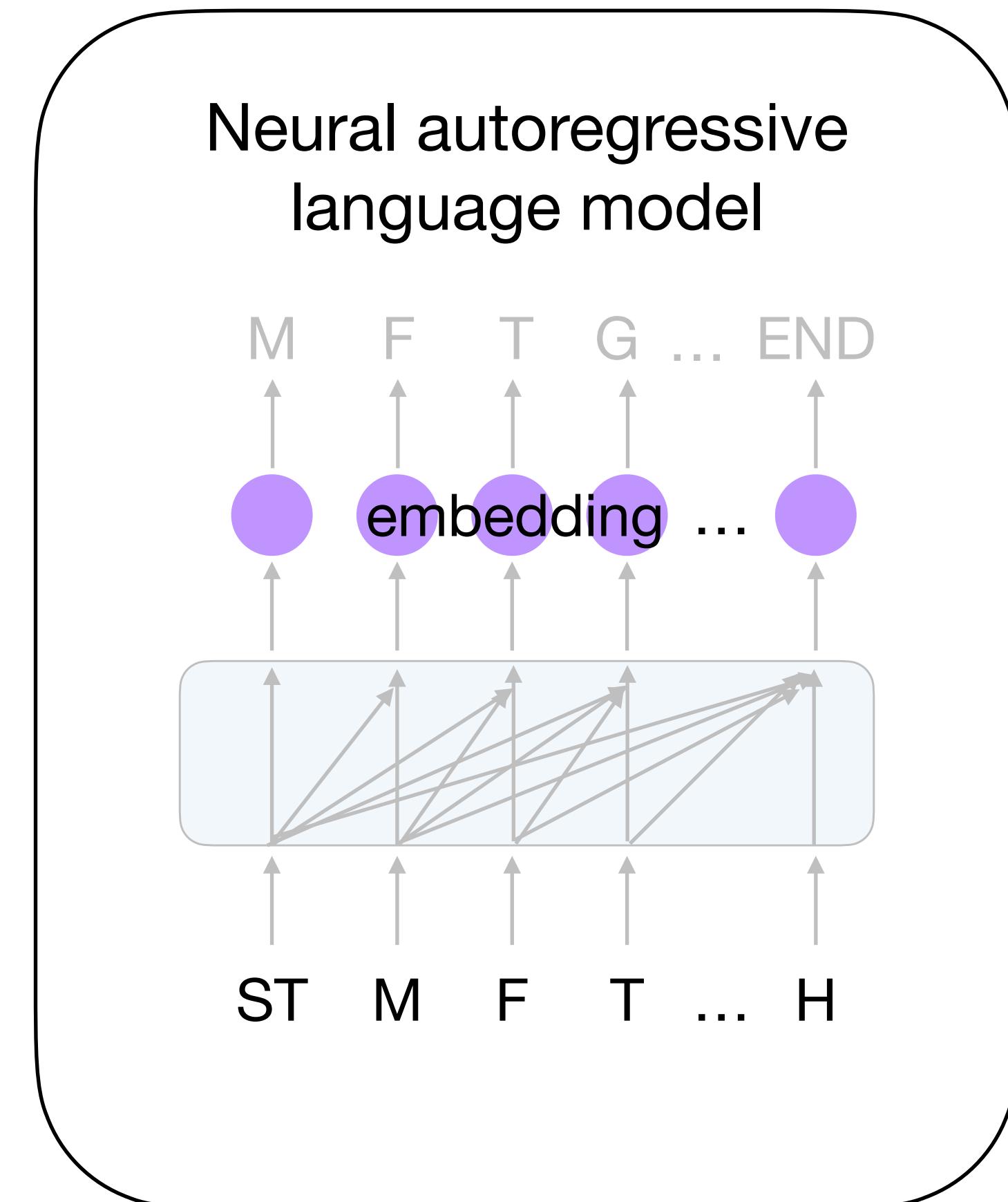
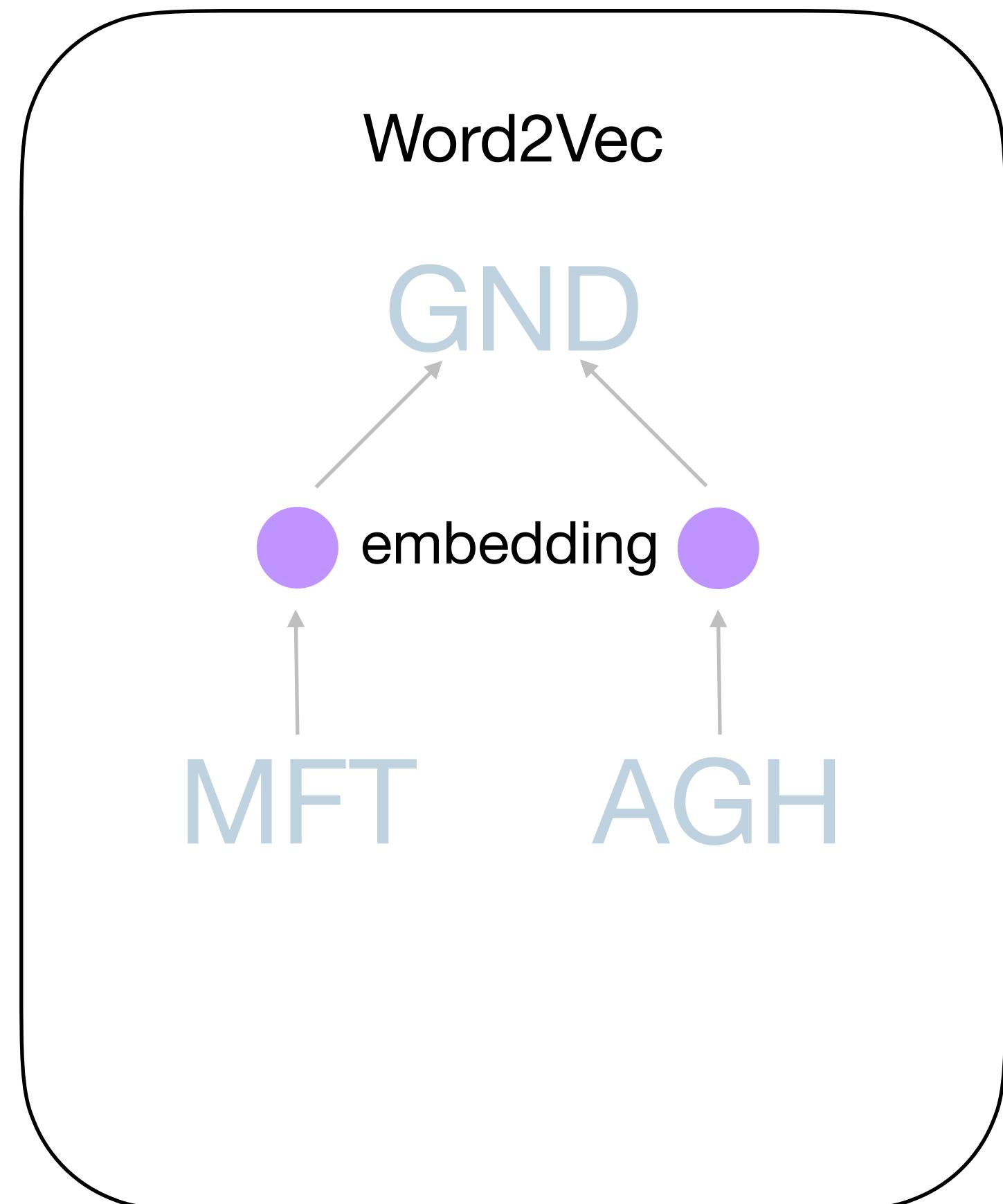
MFTGNDAGH

Many methods pretend proteins are language



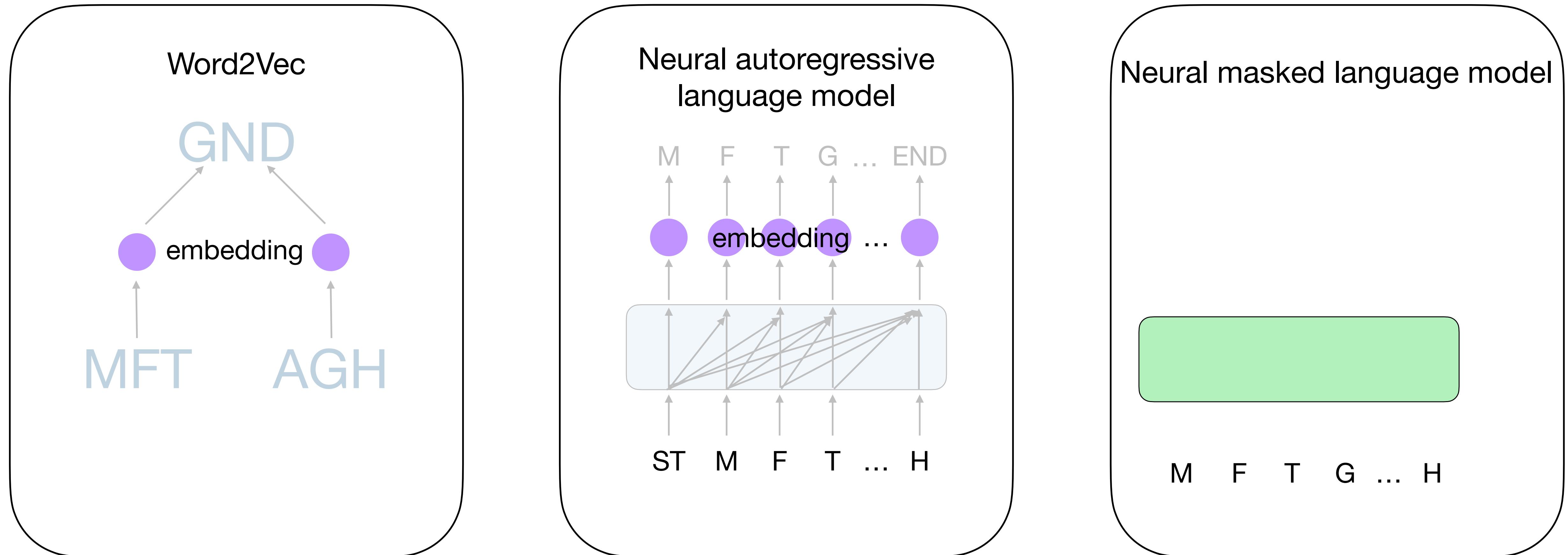
MFTGNDAGH

Many methods pretend proteins are language



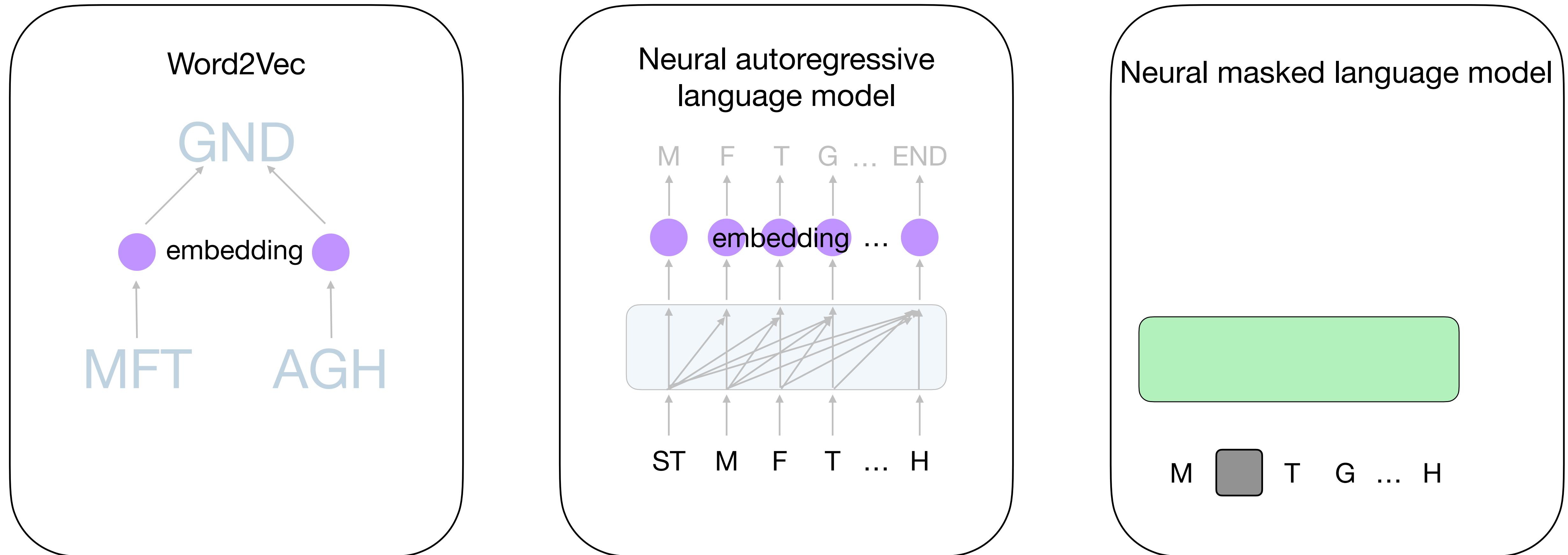
MFTGNDAGH

Many methods pretend proteins are language



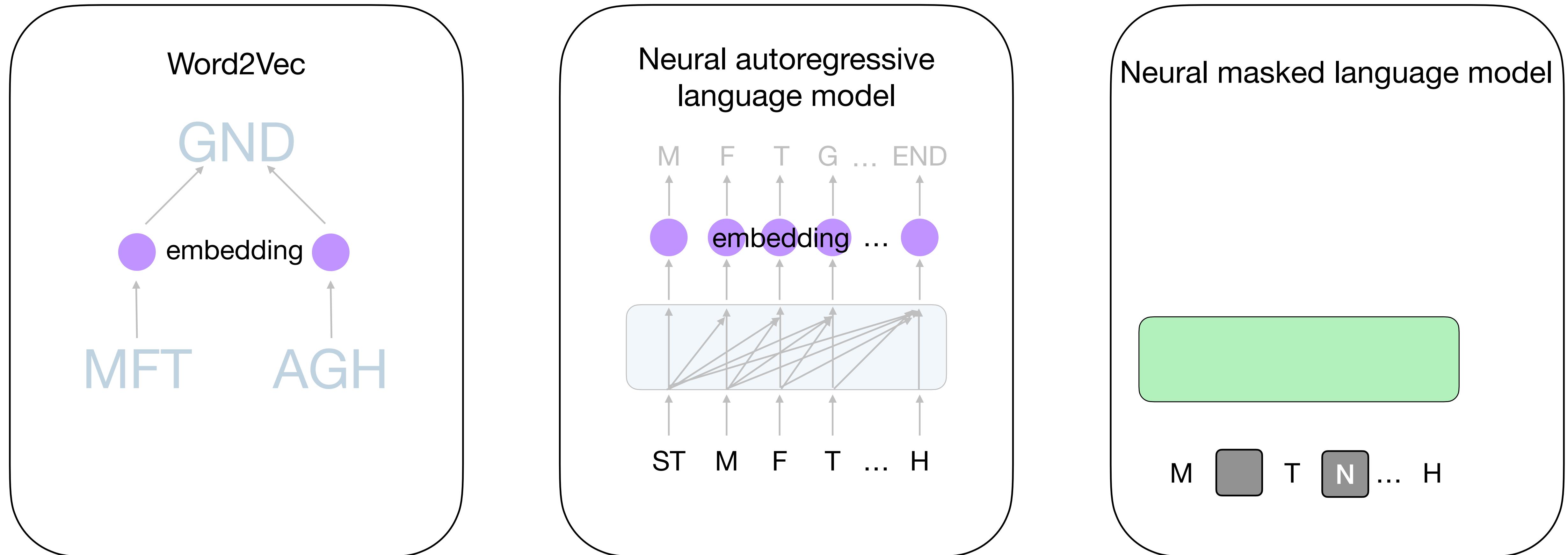
MFTGNDAGH

Many methods pretend proteins are language

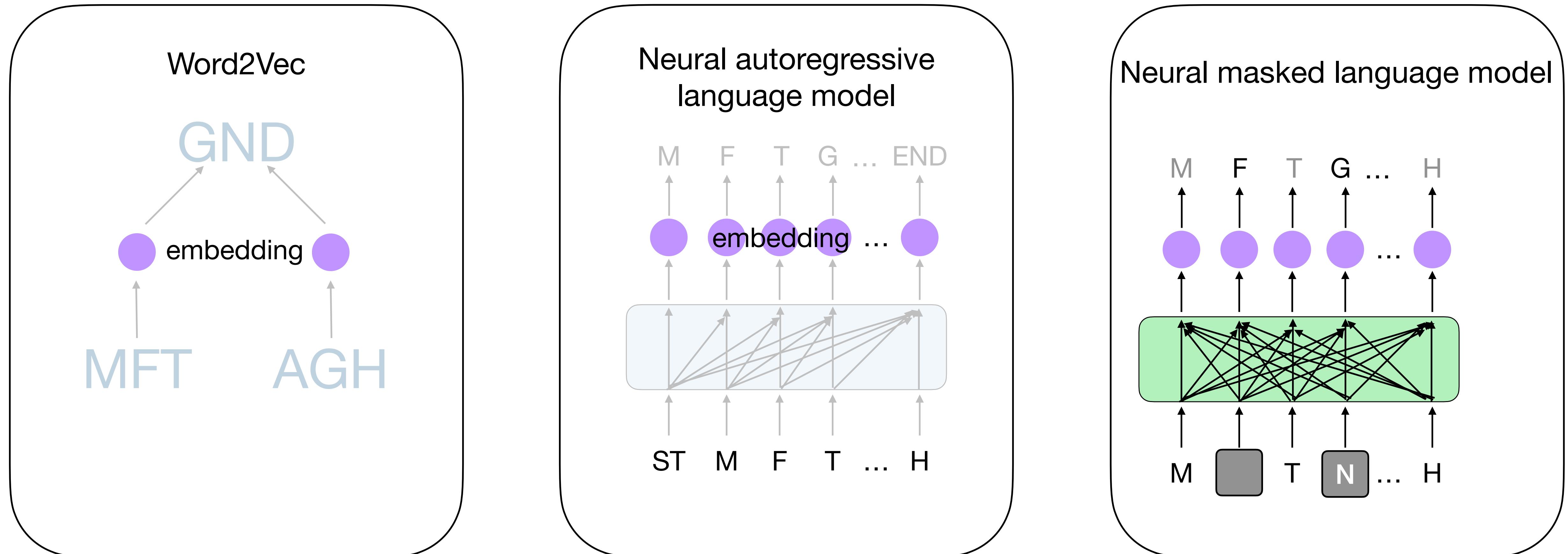


MFTGNDAGH

Many methods pretend proteins are language

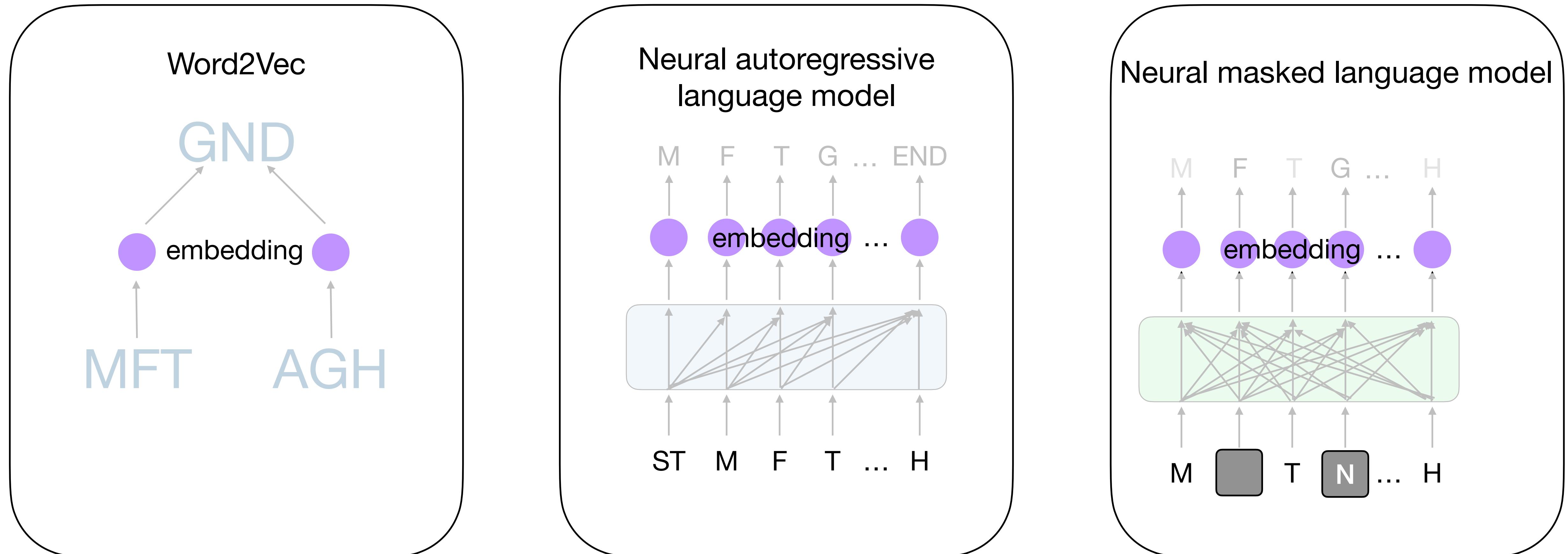


Many methods pretend proteins are language



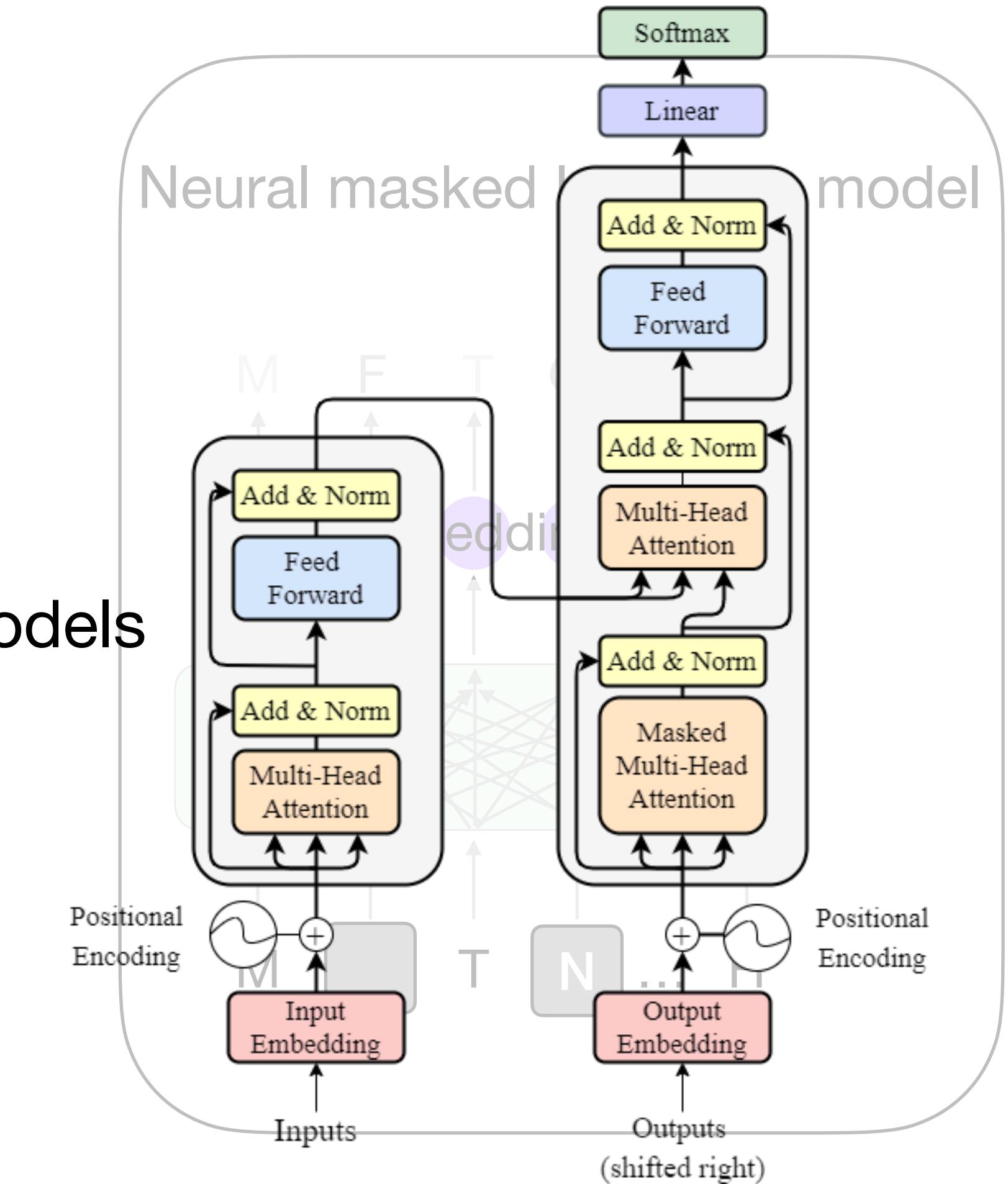
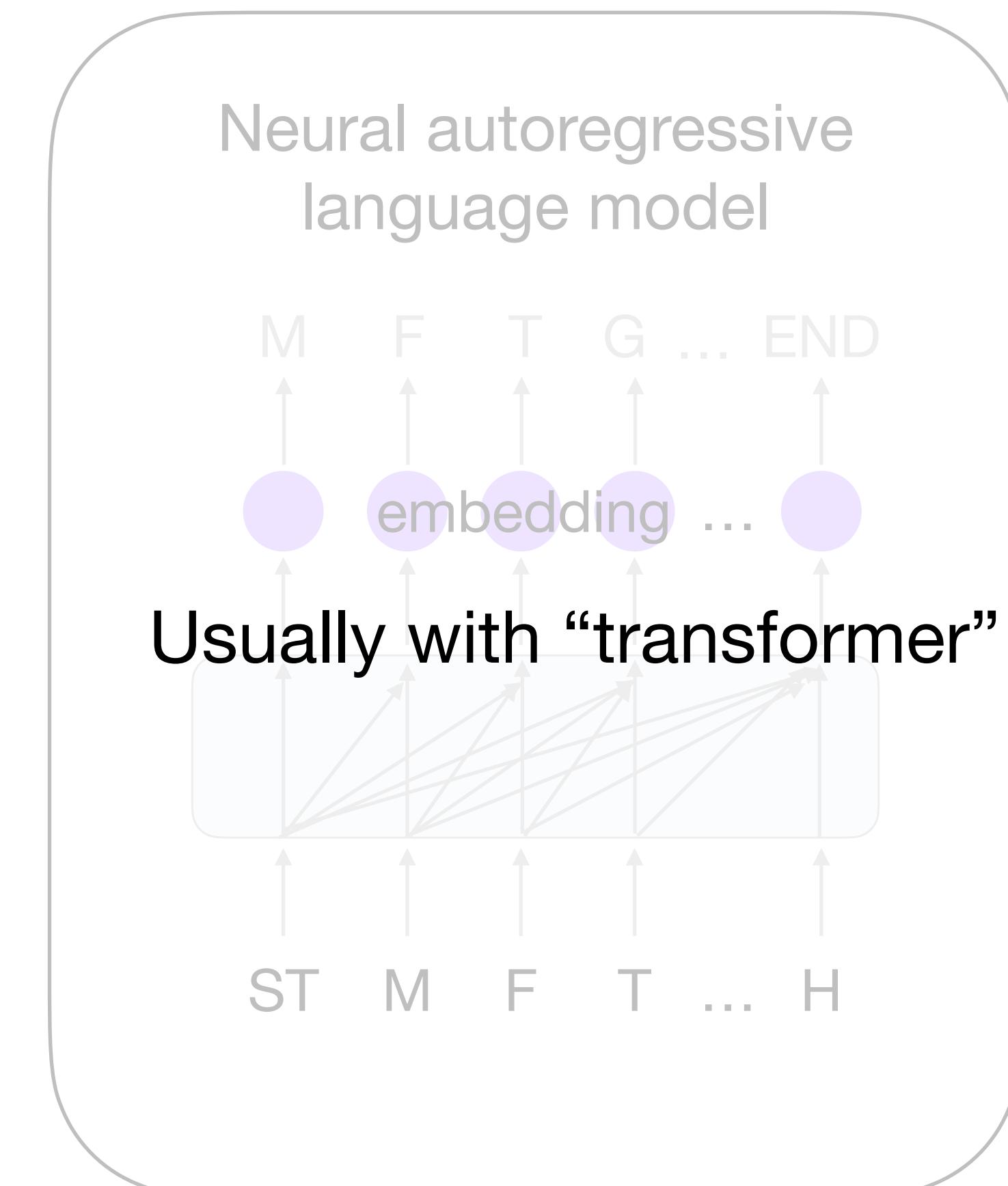
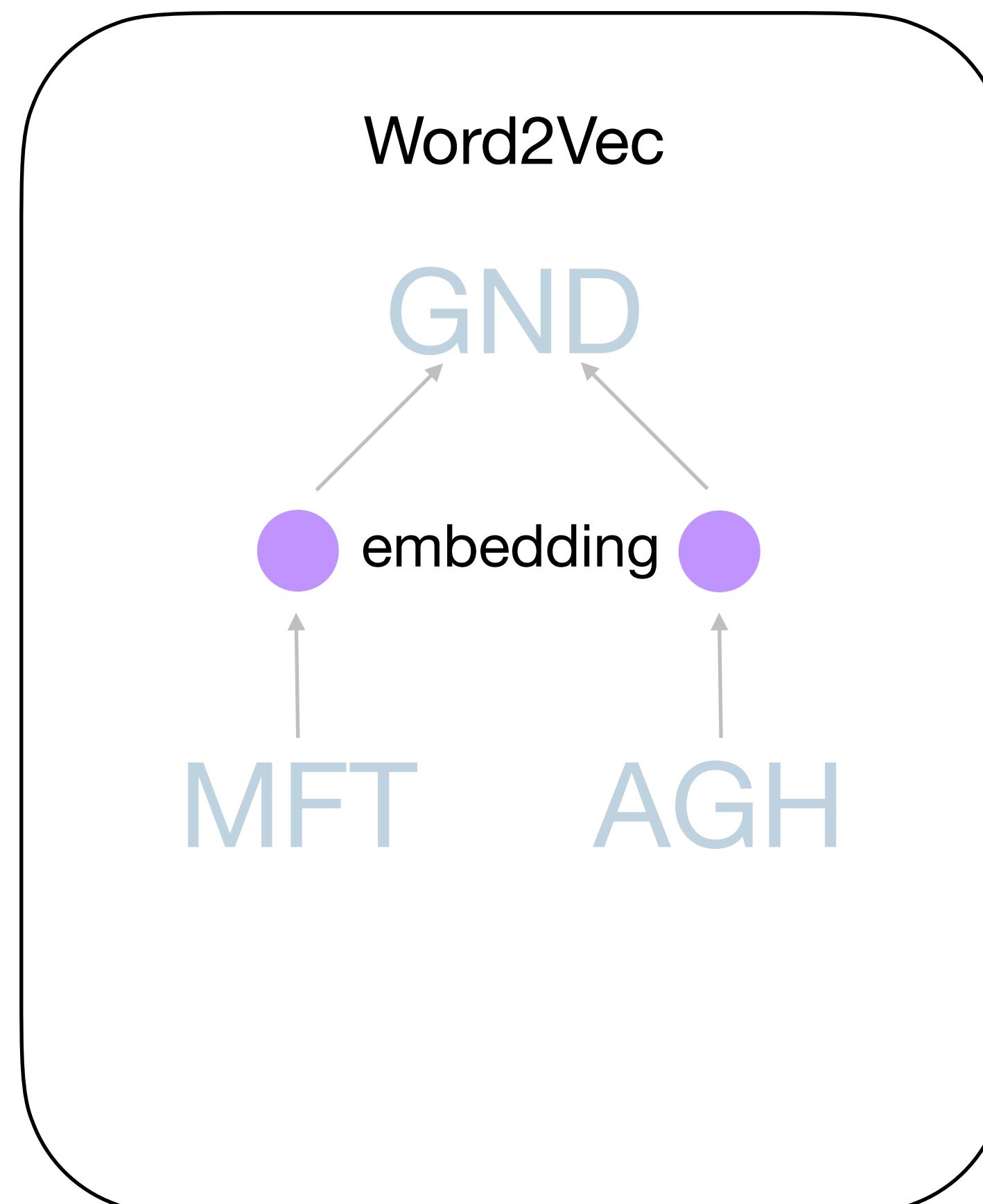
MFTGNDAGH

Many methods pretend proteins are language



MFTGNDAGH

Many methods pretend proteins are language



MFTGNDAGH

Pretrained transformers recapitulate biophysical properties

Pretrained transformers recapitulate biophysical properties

Biological property

- ✖ Negatively charged
- Positively charged

● Hydrophobic

✚ Aromatic

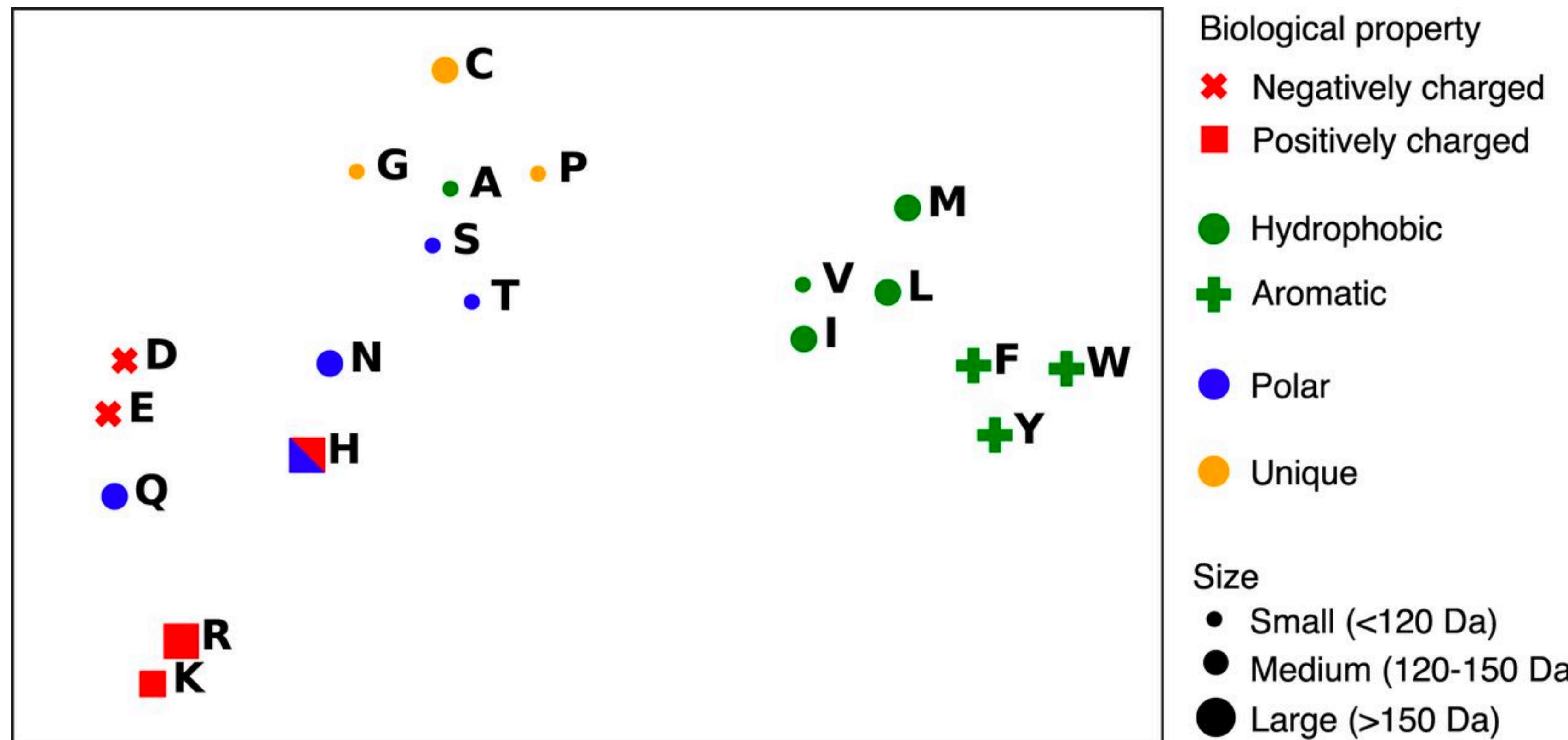
● Polar

● Unique

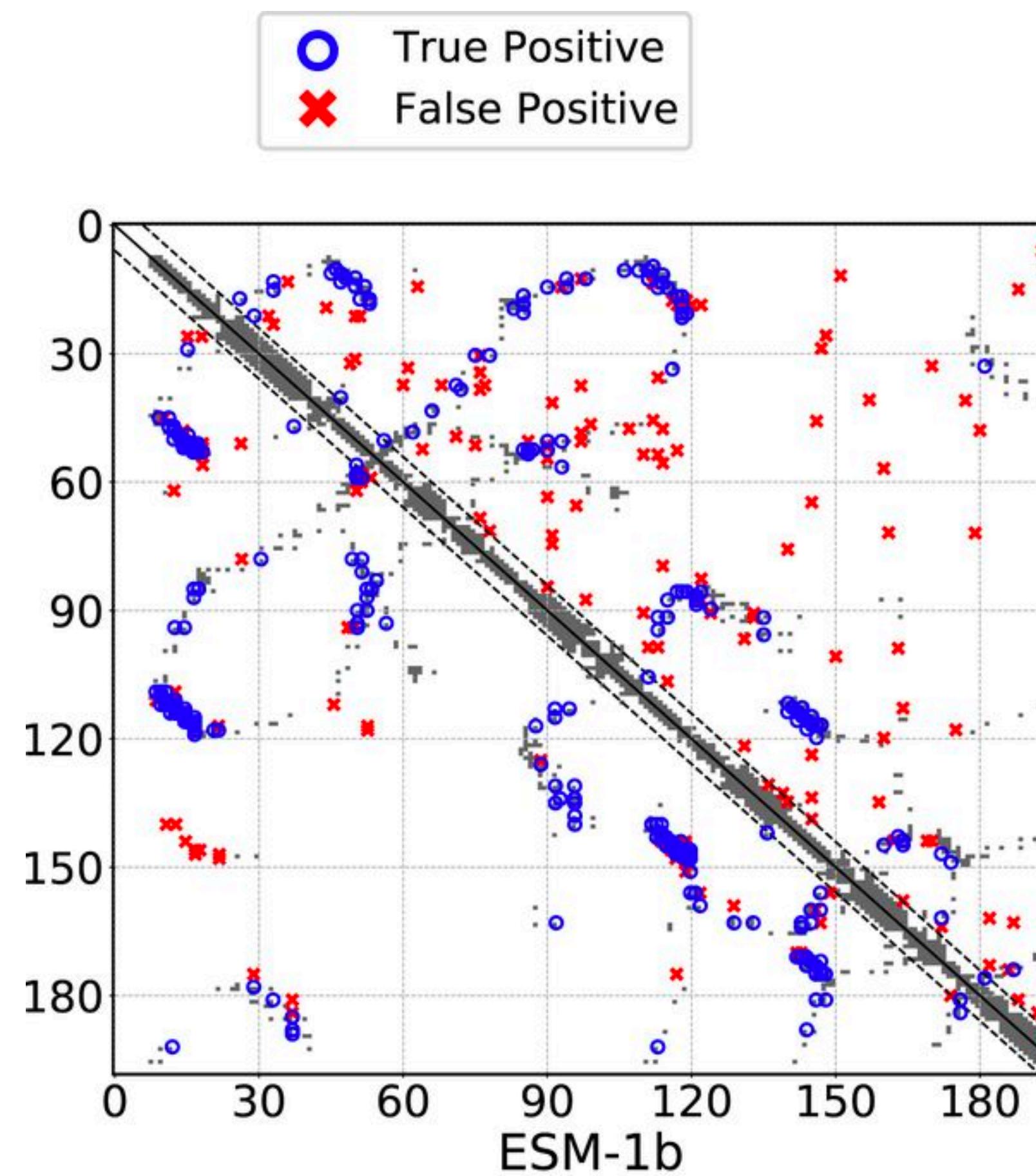
Size

- Small (<120 Da)
- Medium (120-150 Da)
- Large (>150 Da)

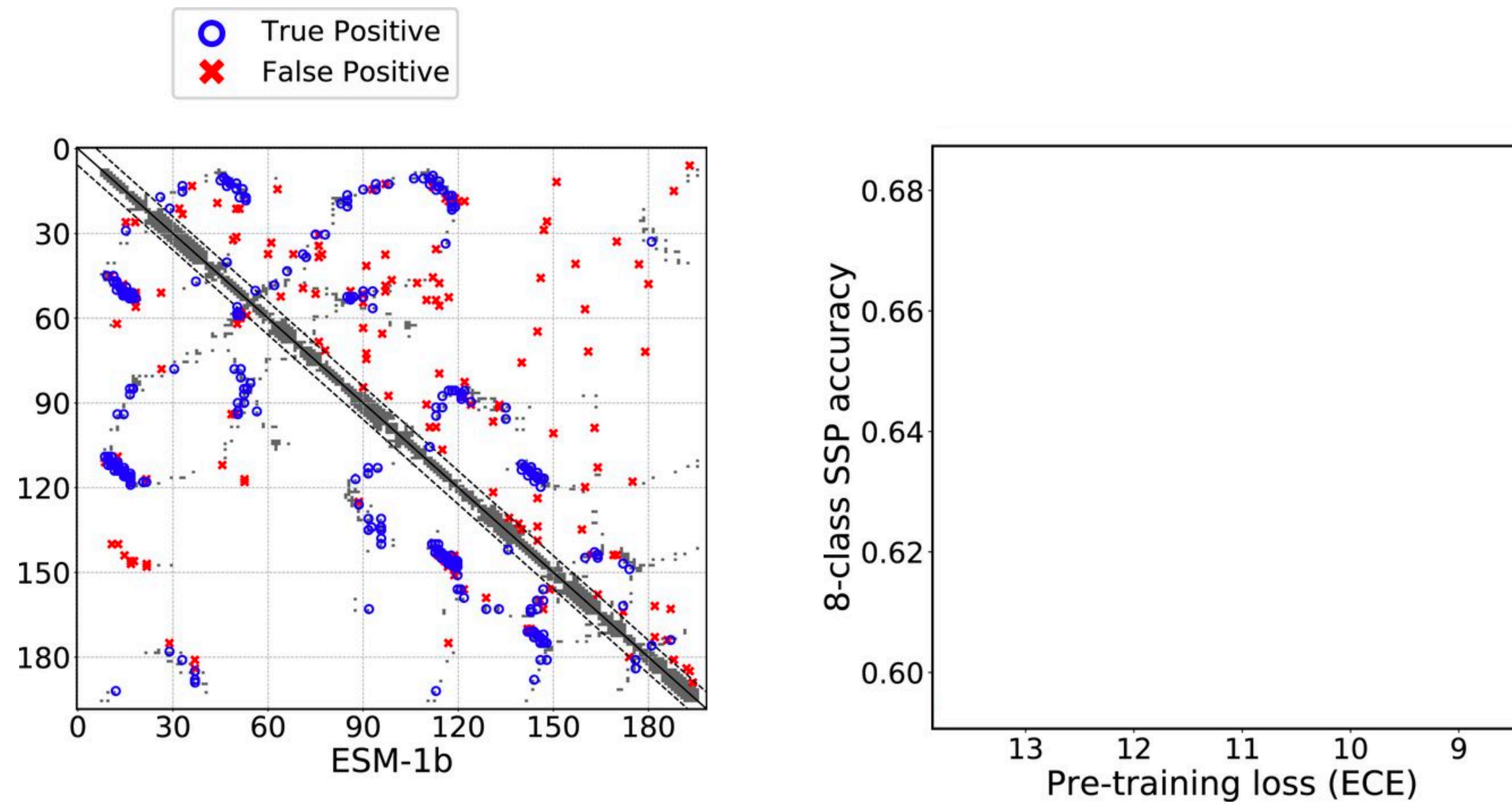
Pretrained transformers recapitulate biophysical properties



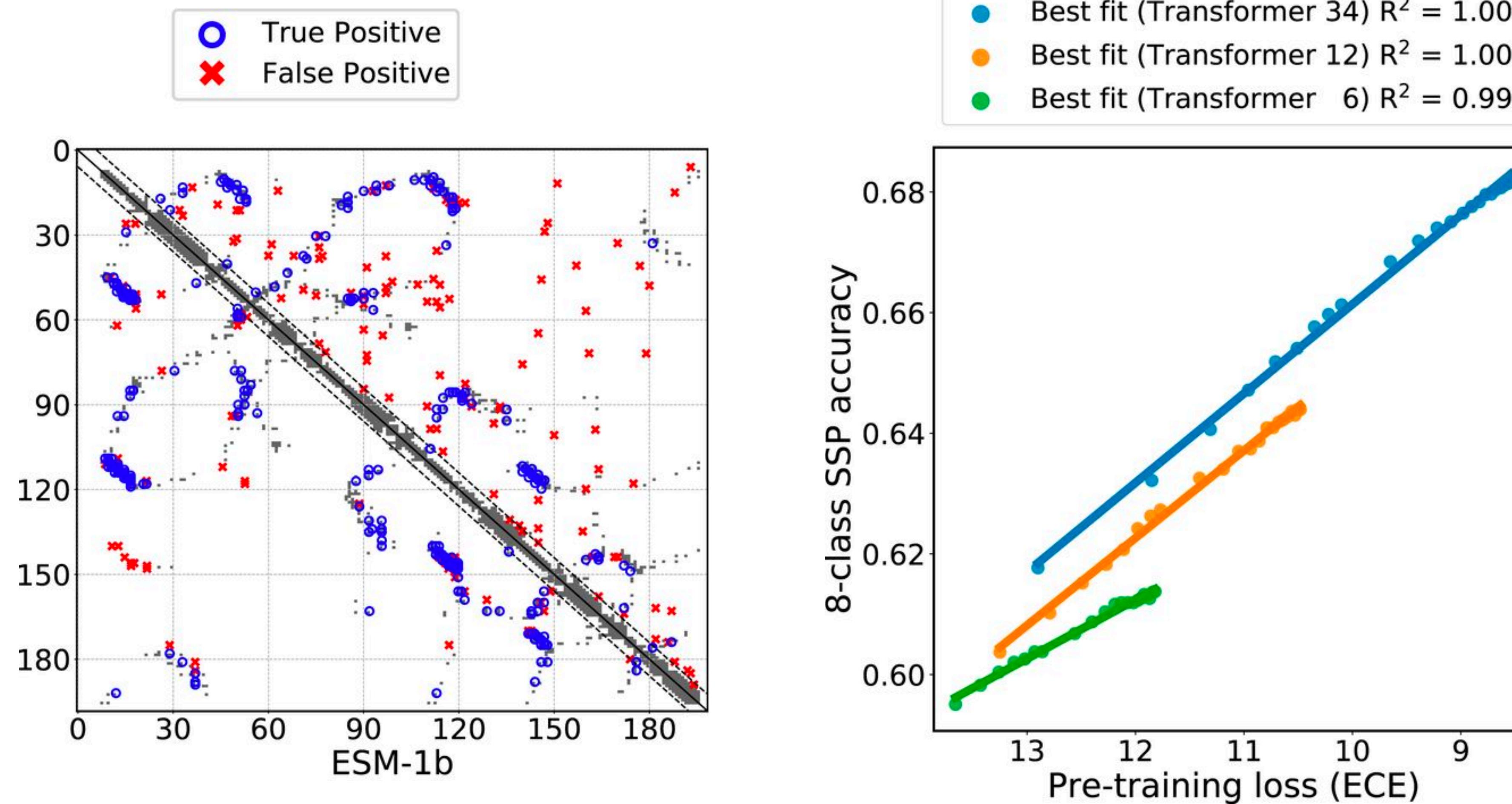
Pretrained transformers contain structural information



Pretrained transformers contain structural information



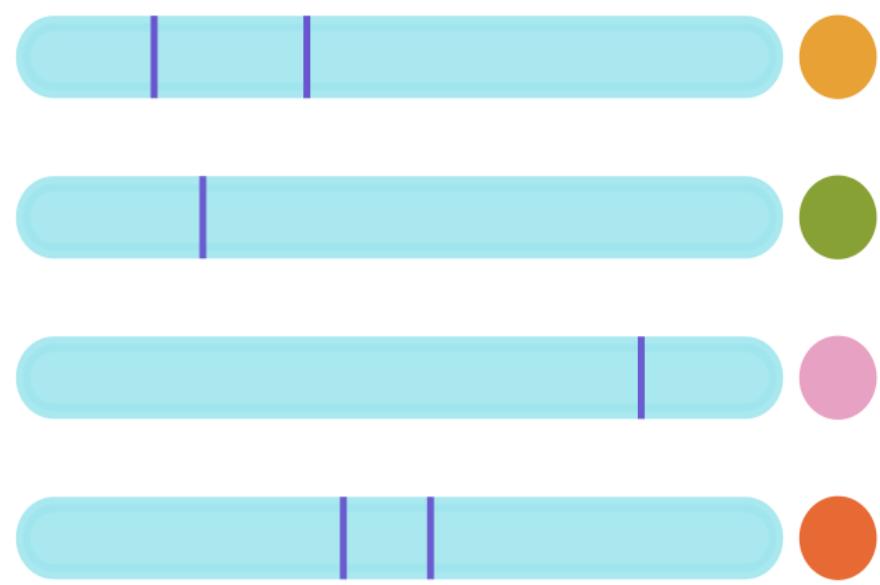
Pretrained transformers contain structural information



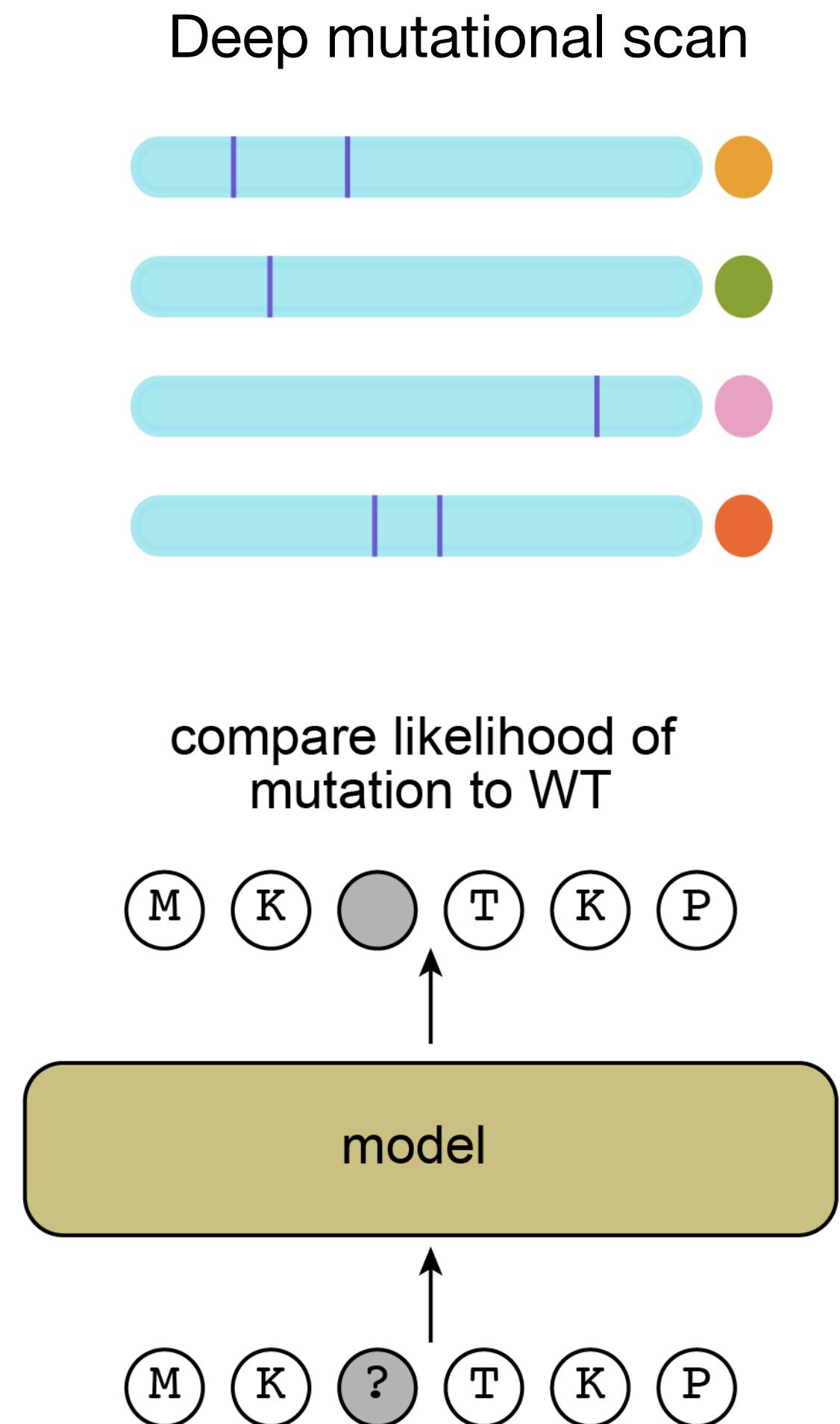
Pretrained transformers are zero-shot fitness predictors

Pretrained transformers are zero-shot fitness predictors

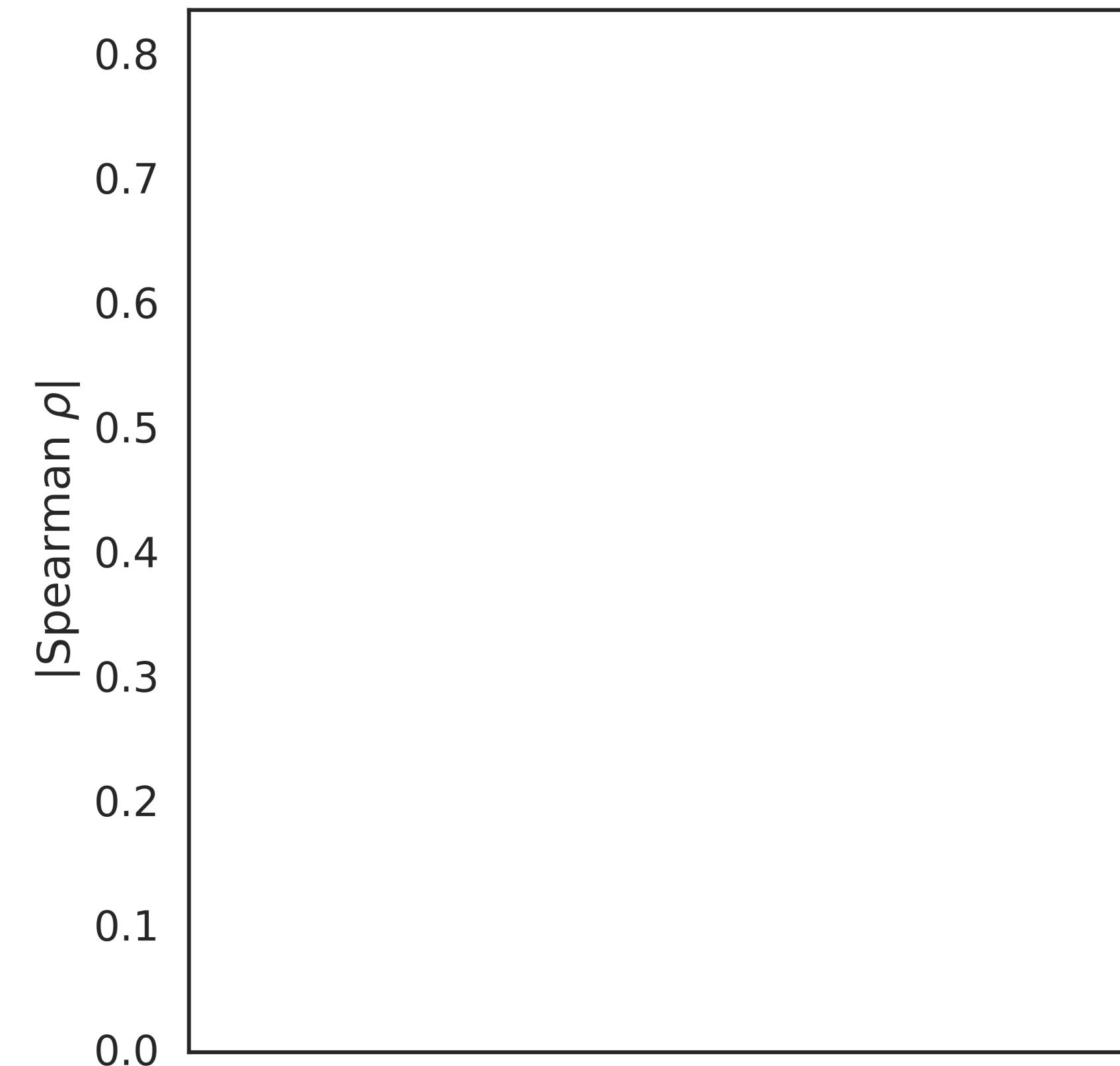
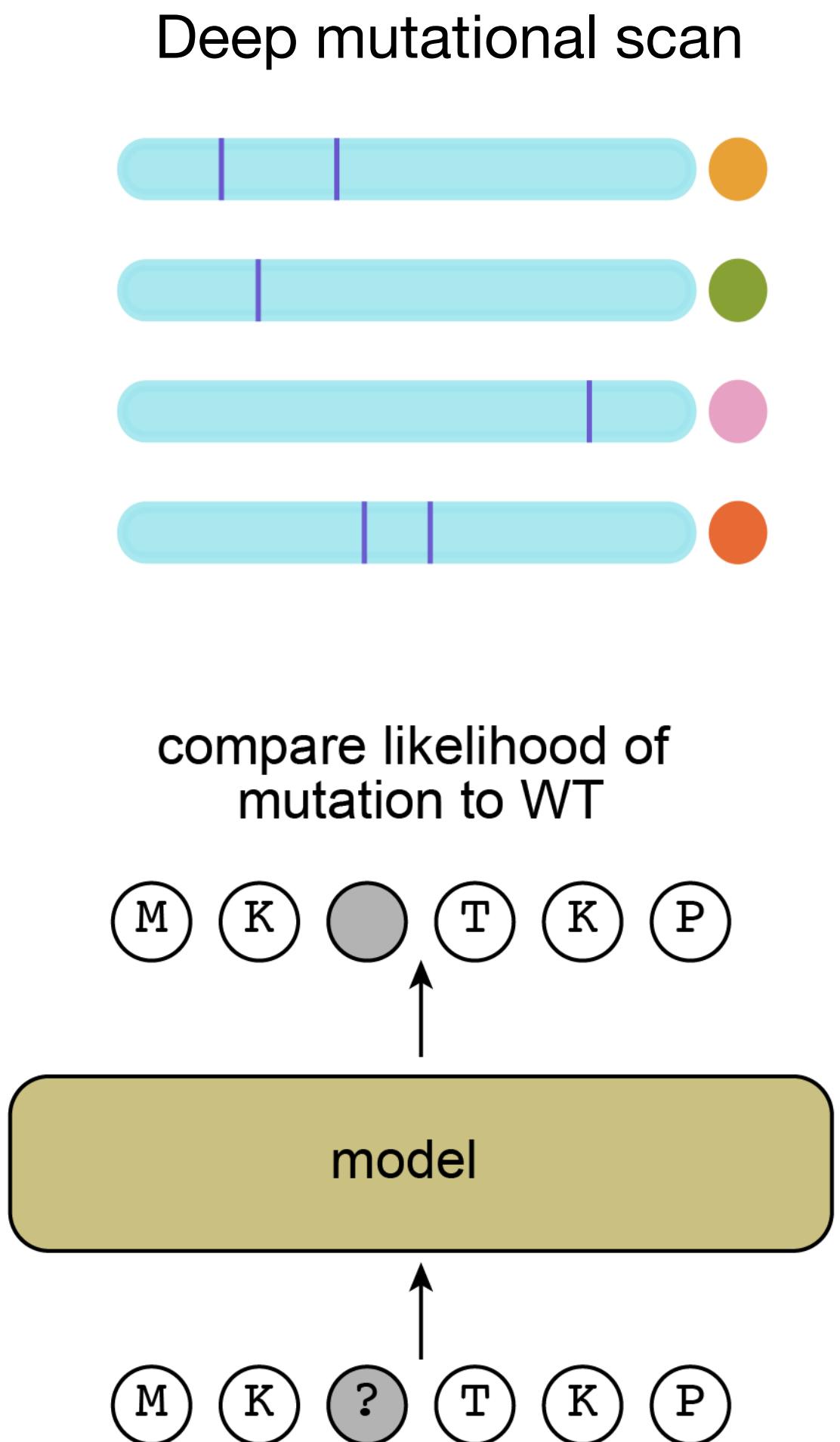
Deep mutational scan



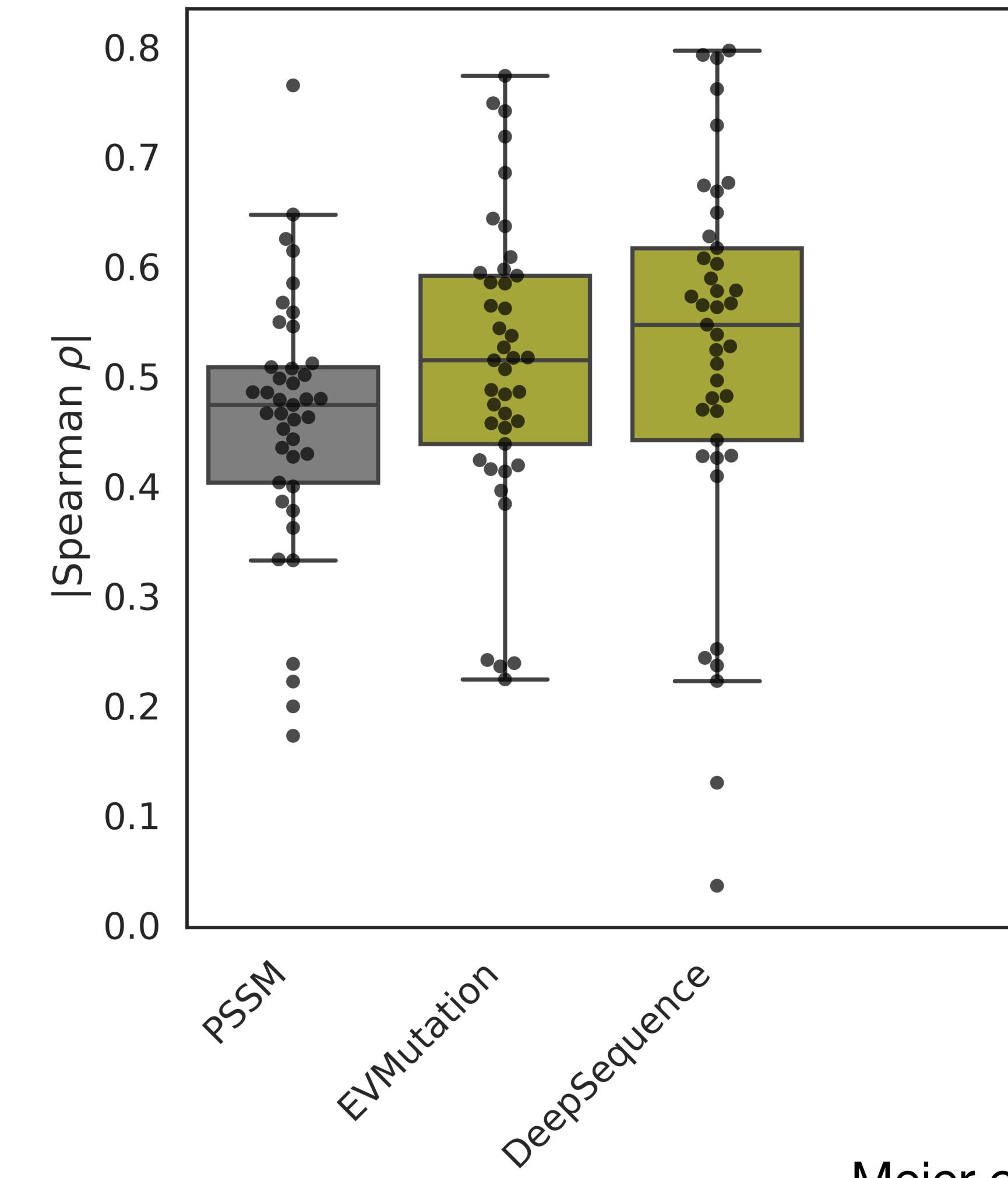
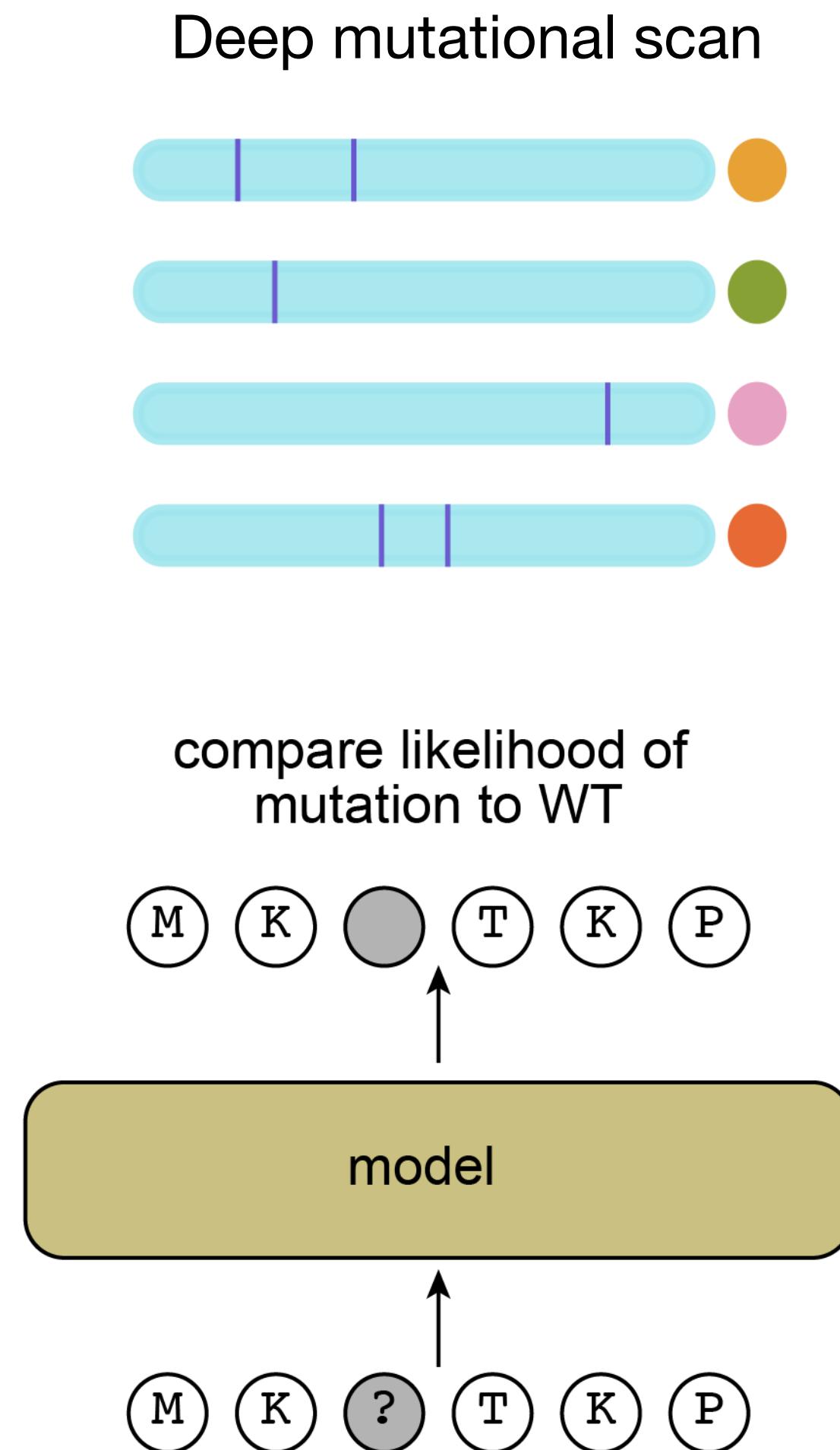
Pretrained transformers are zero-shot fitness predictors



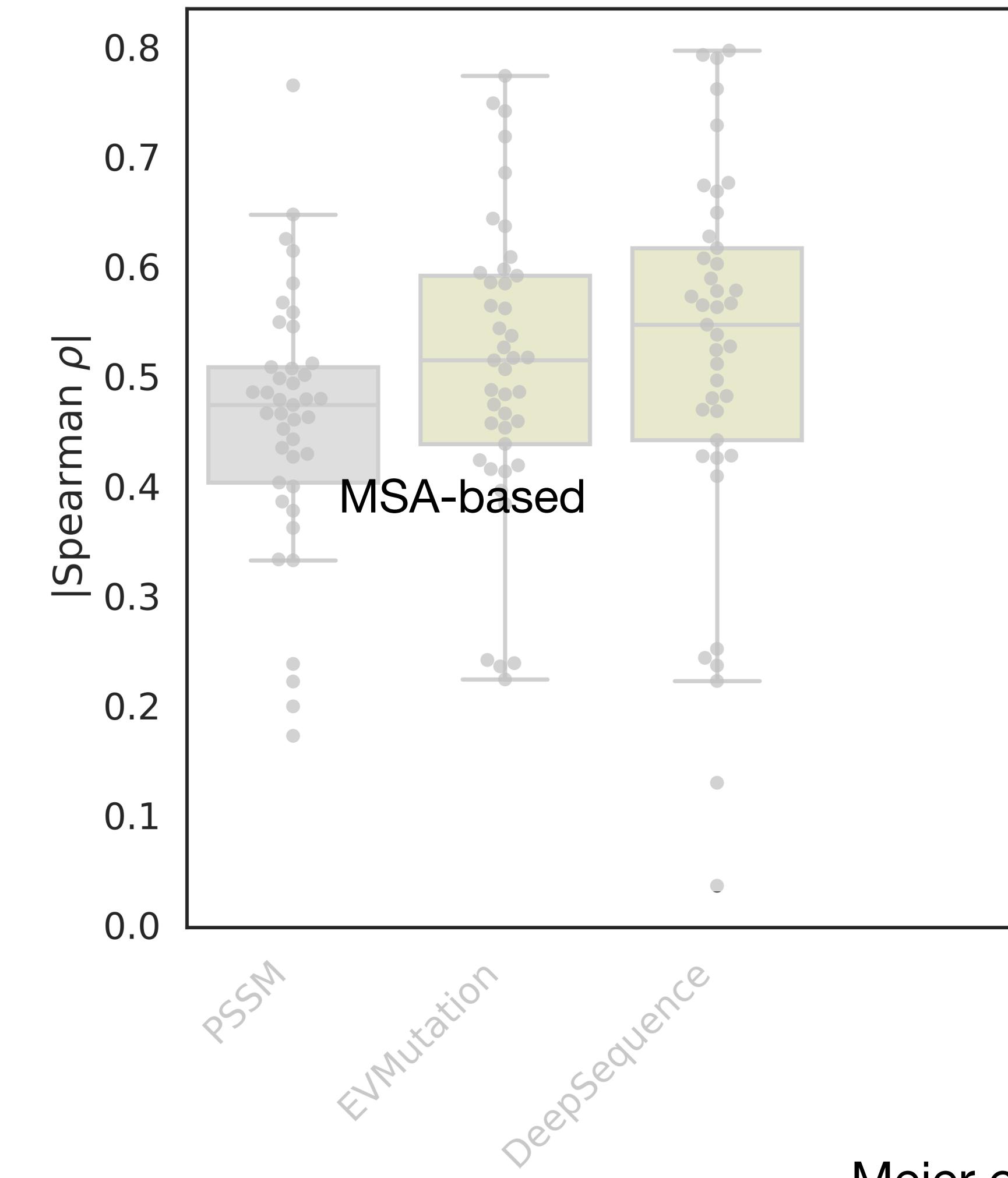
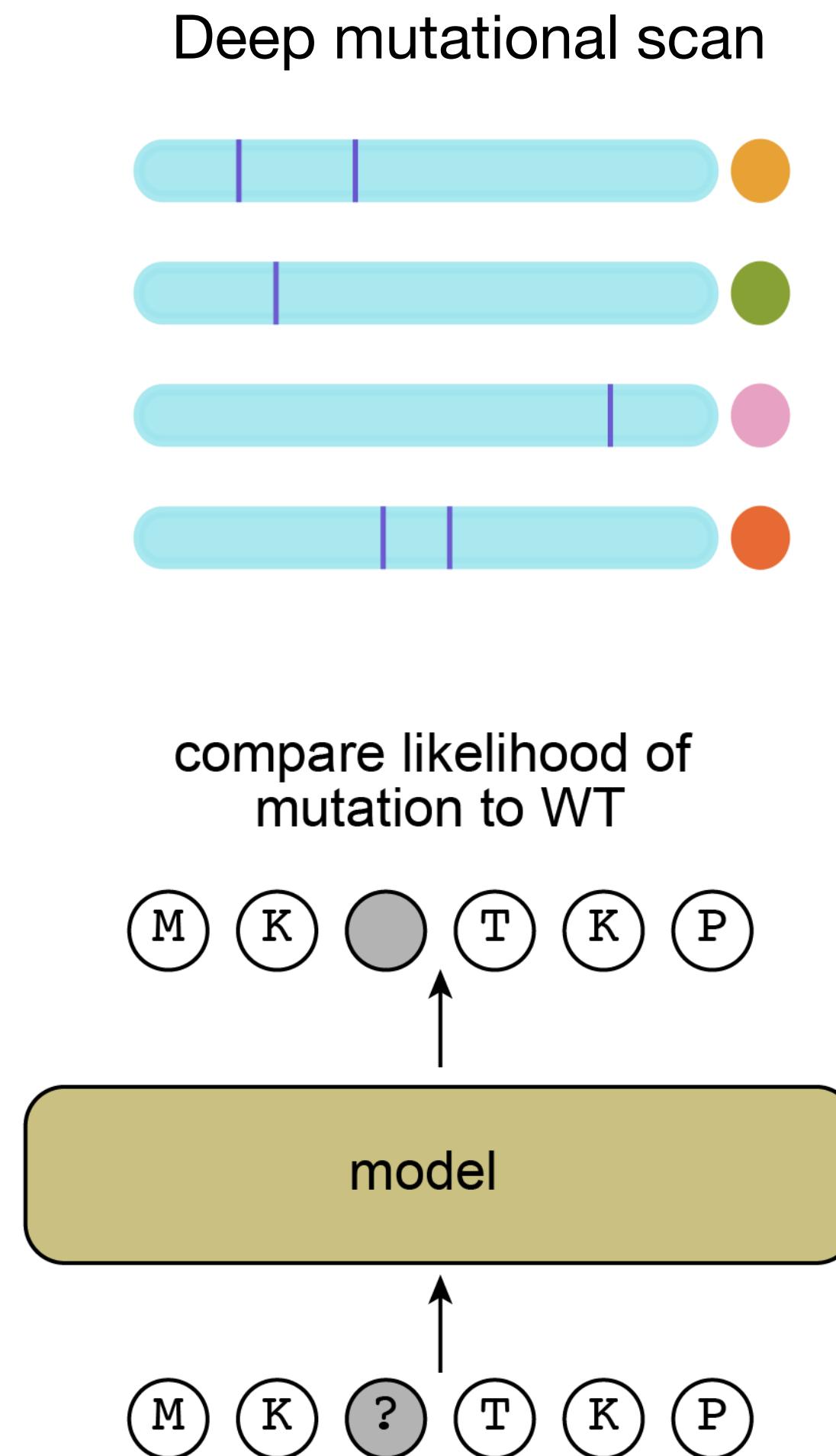
Pretrained transformers are zero-shot fitness predictors



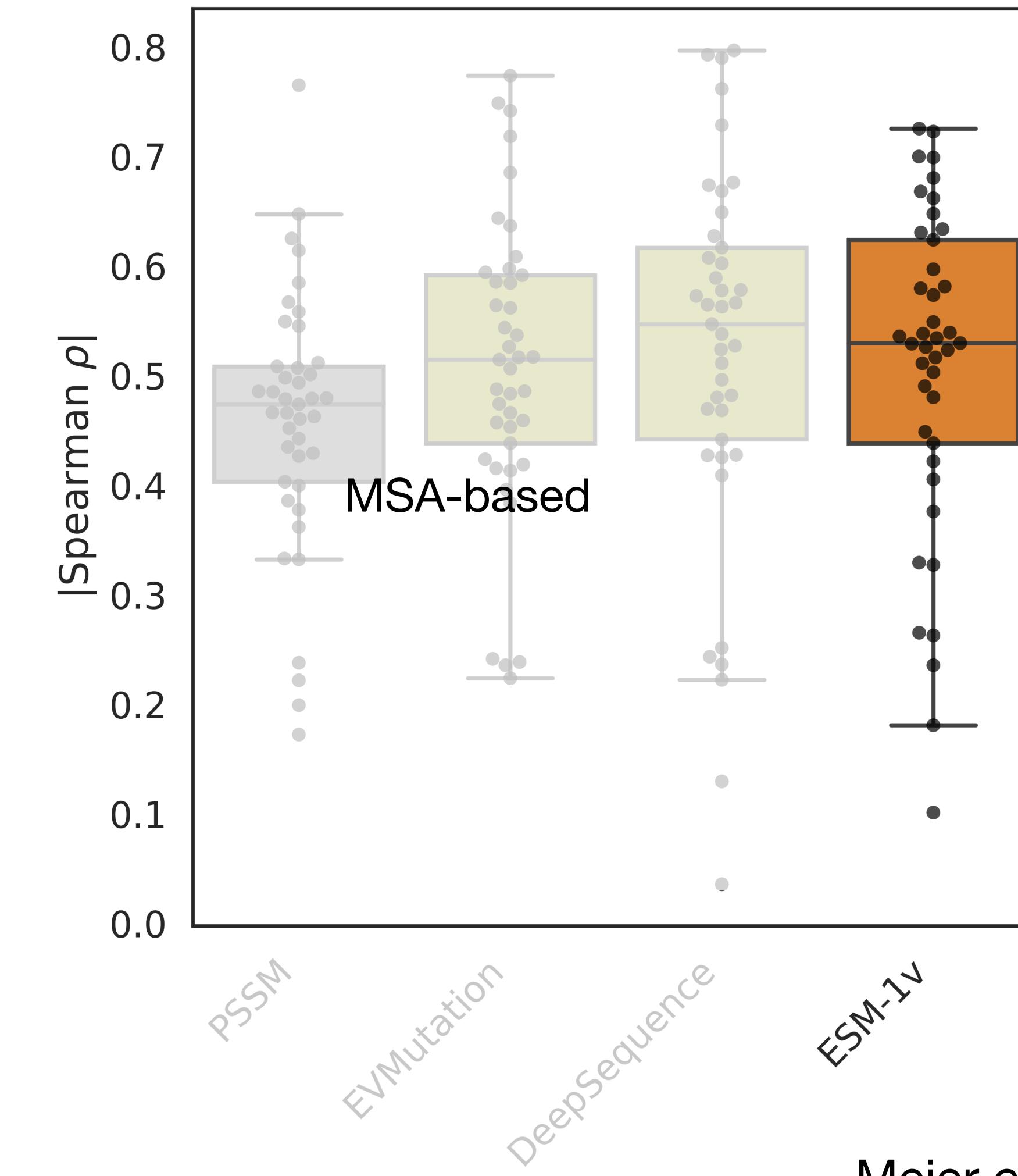
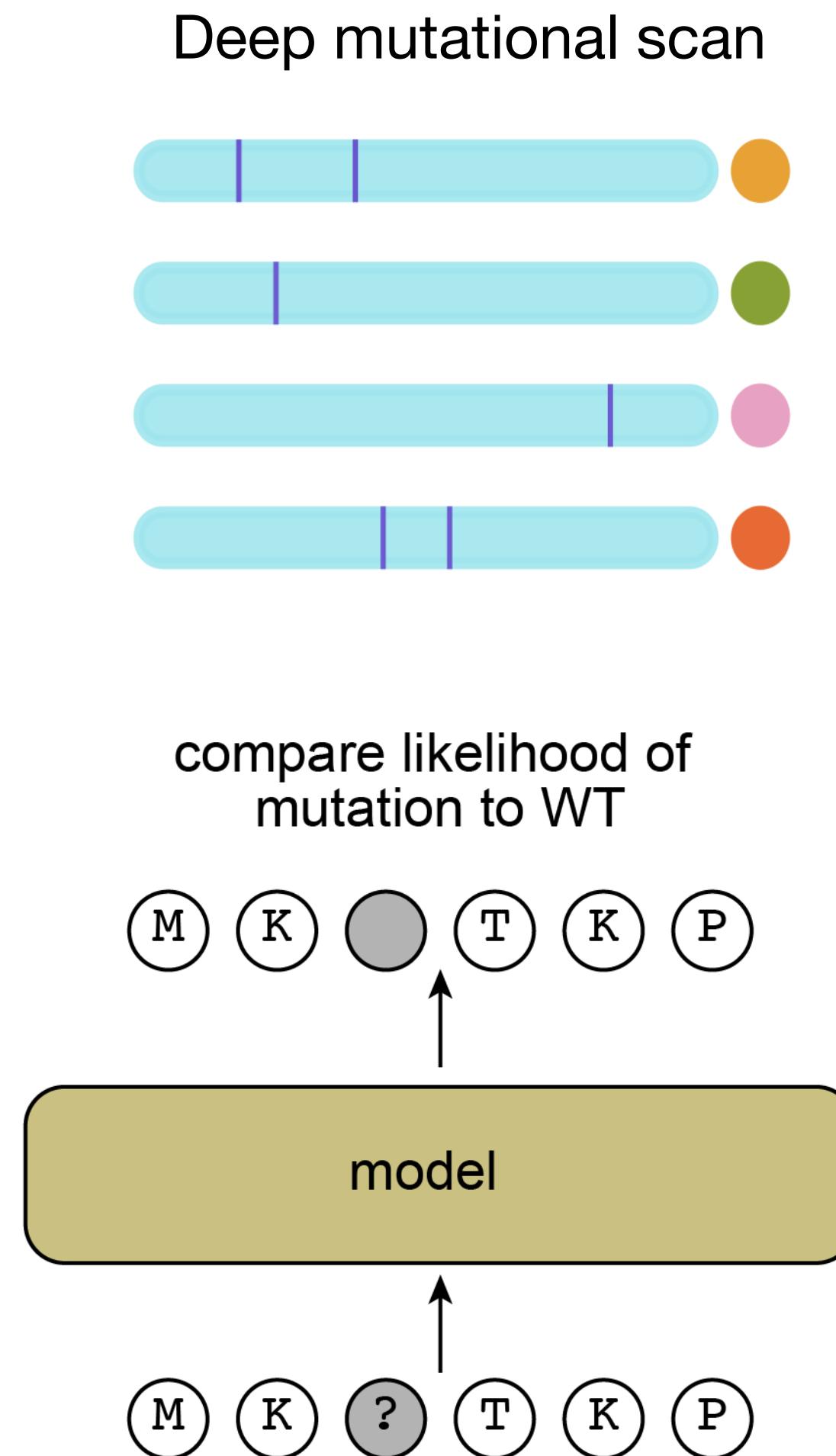
Pretrained transformers are zero-shot fitness predictors



Pretrained transformers are zero-shot fitness predictors

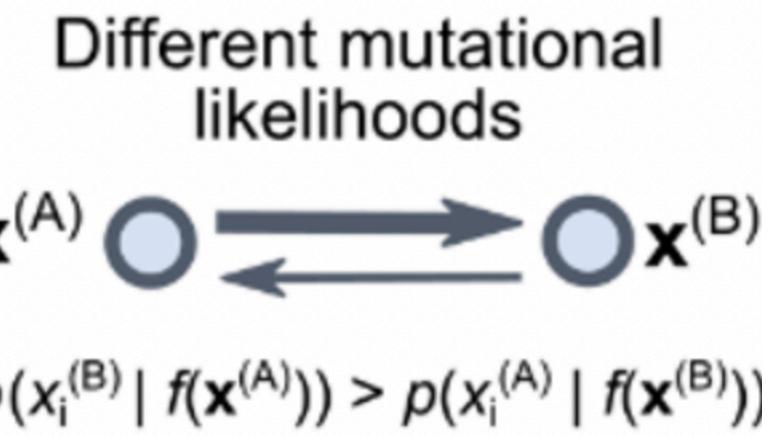


Pretrained transformers are zero-shot fitness predictors

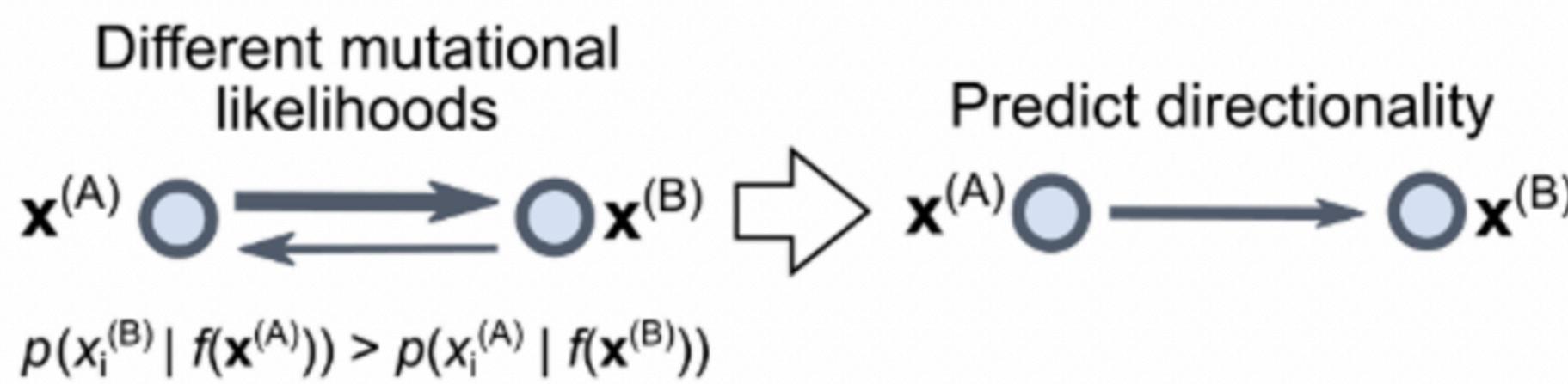


Pretrained transformers reconstruct evolutionary trajectories

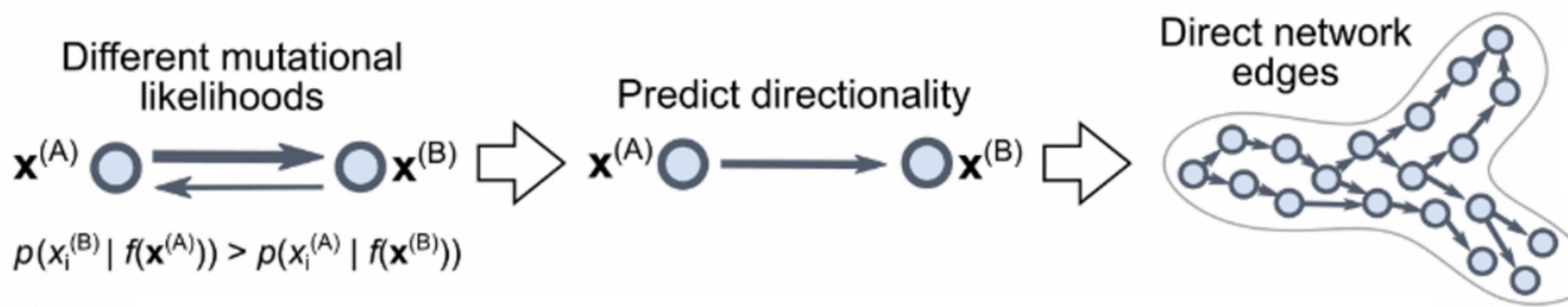
Pretrained transformers reconstruct evolutionary trajectories



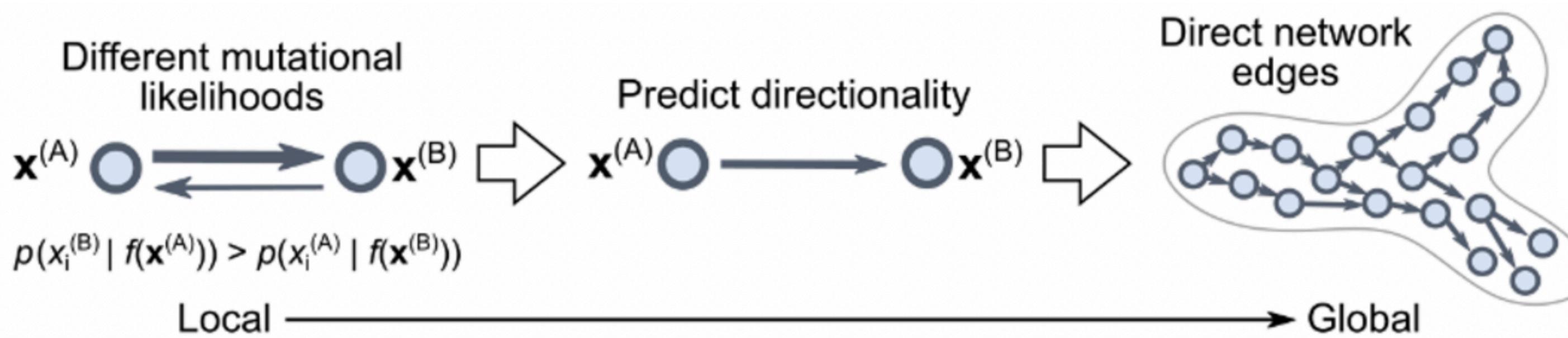
Pretrained transformers reconstruct evolutionary trajectories



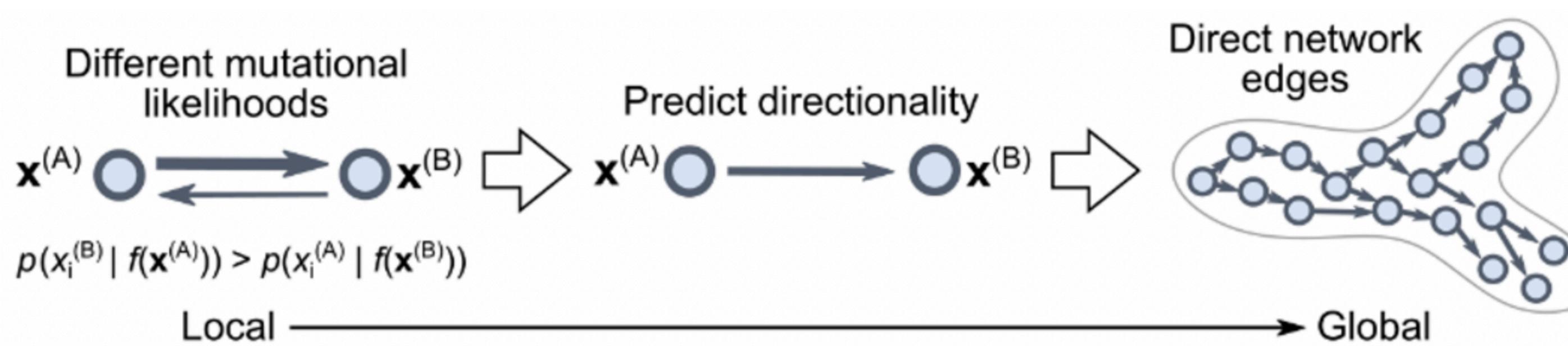
Pretrained transformers reconstruct evolutionary trajectories



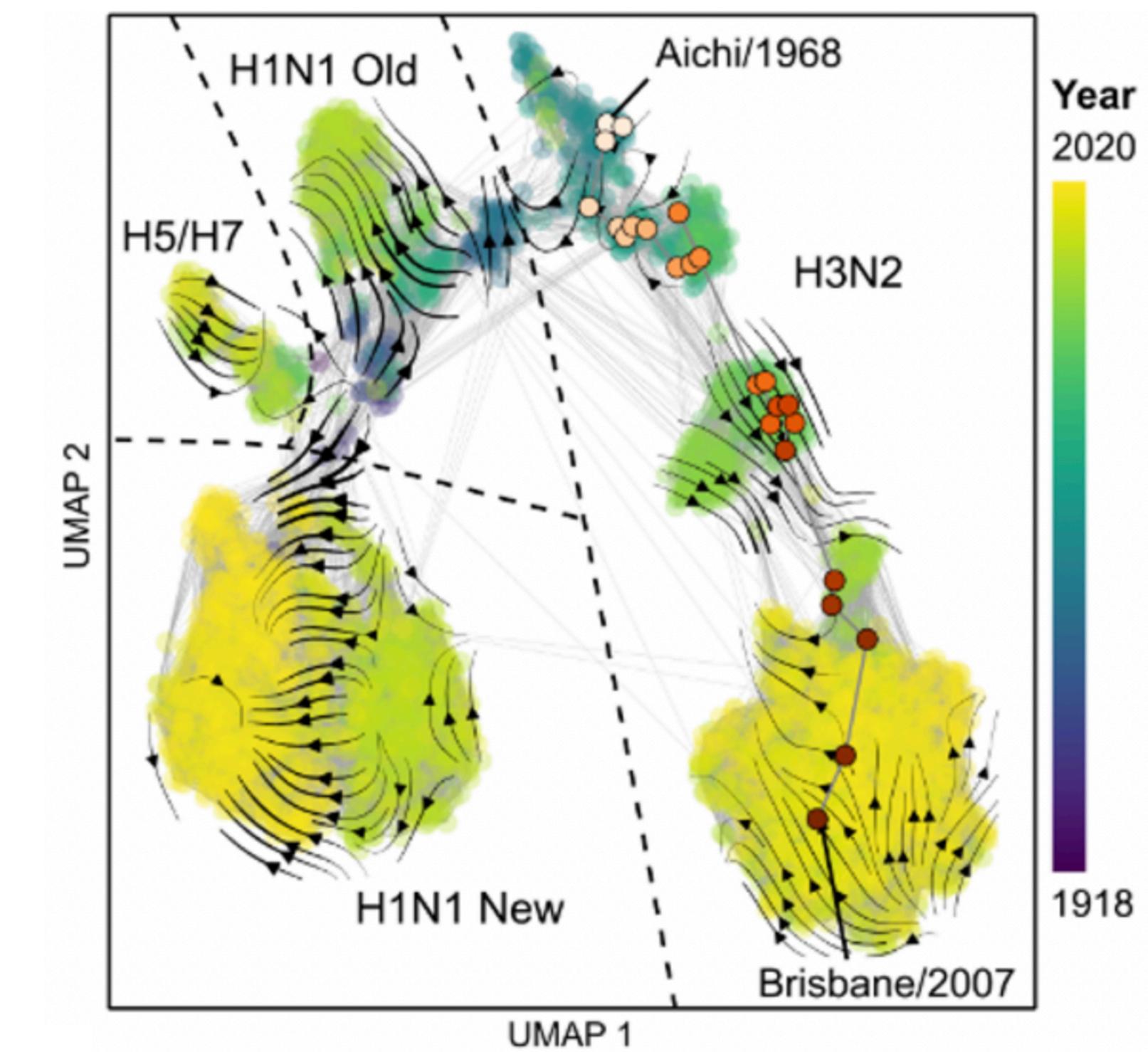
Pretrained transformers reconstruct evolutionary trajectories



Pretrained transformers reconstruct evolutionary trajectories

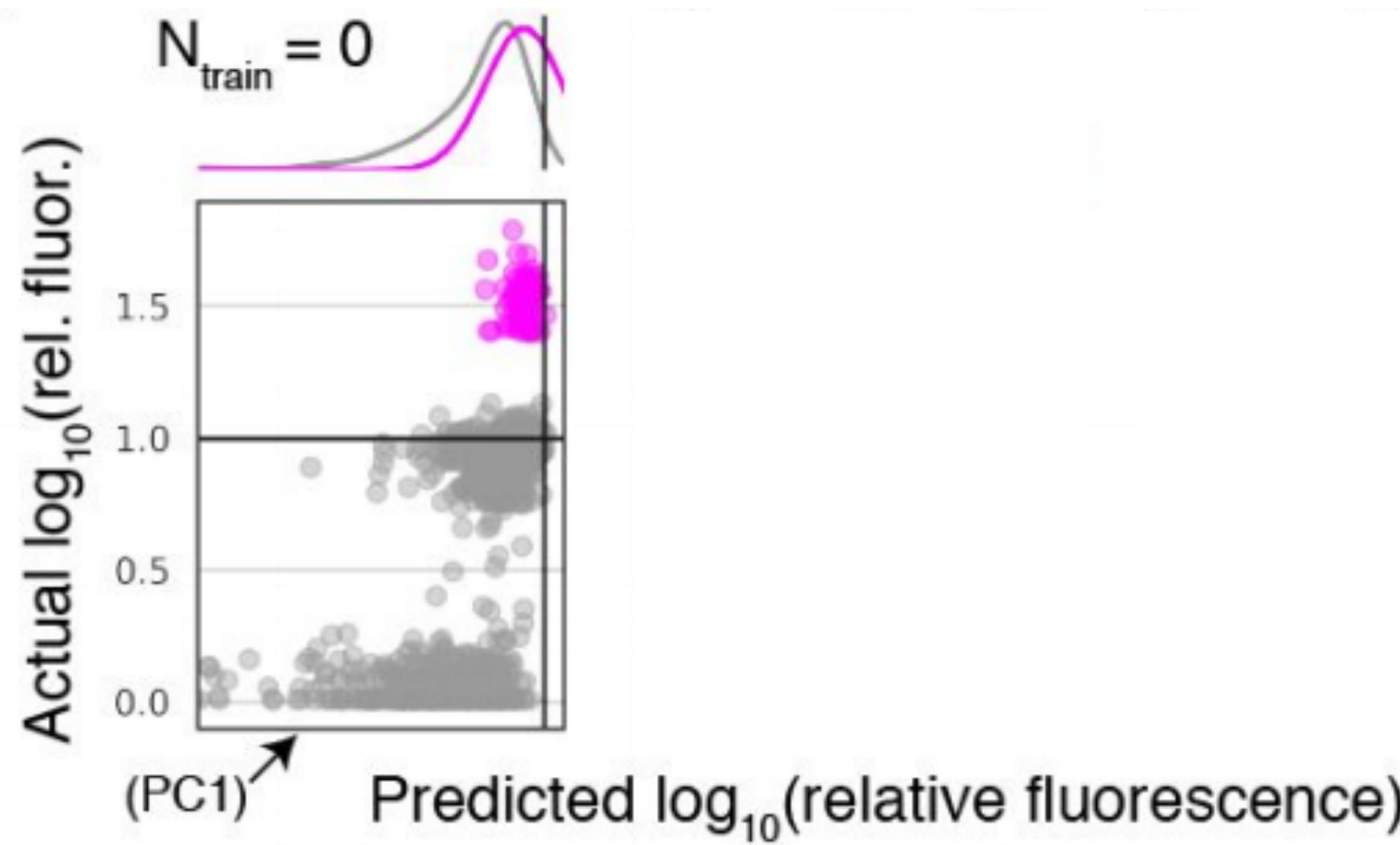


Influenza A nucleoprotein

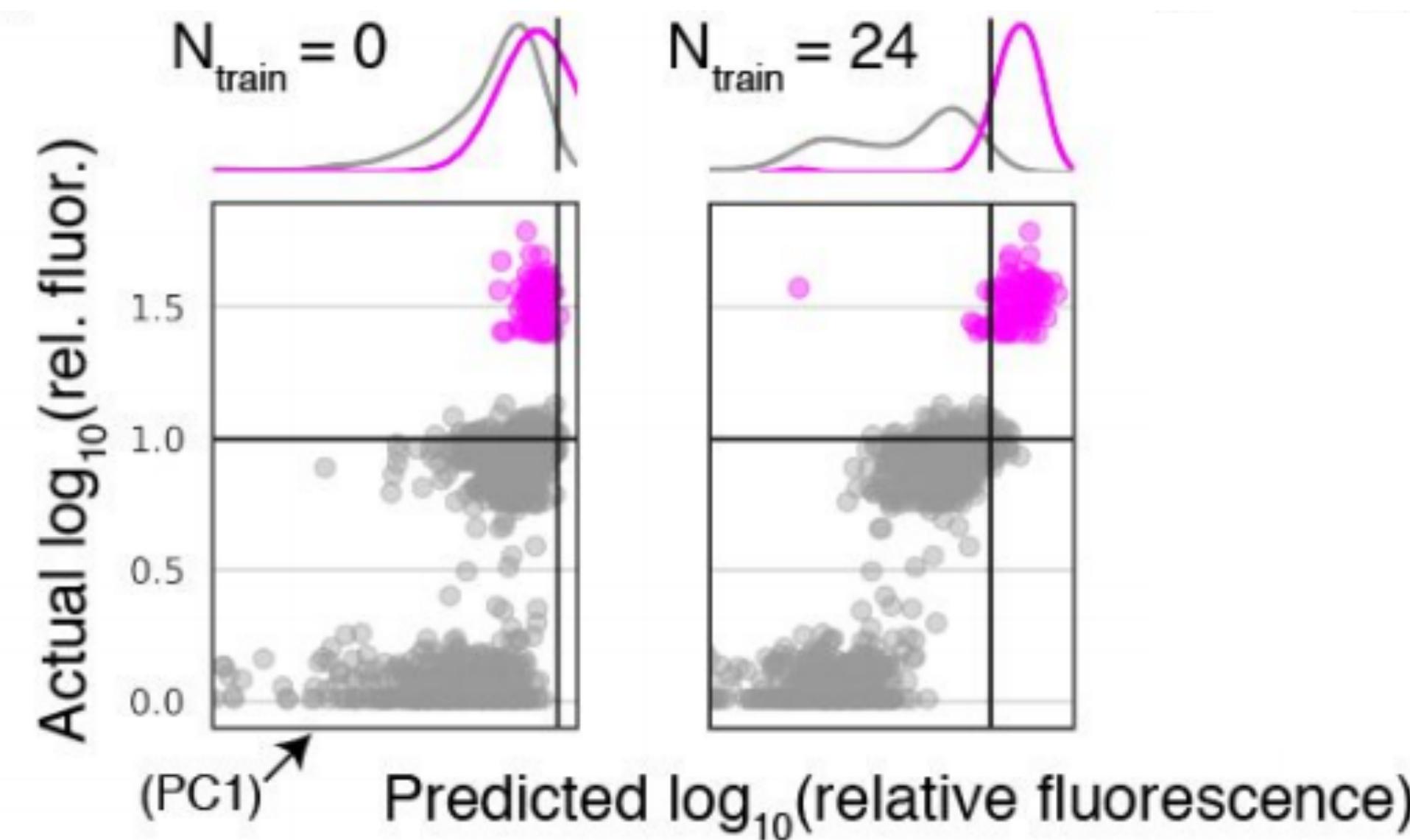


Pretraining guides search away from loss-of-function sequences

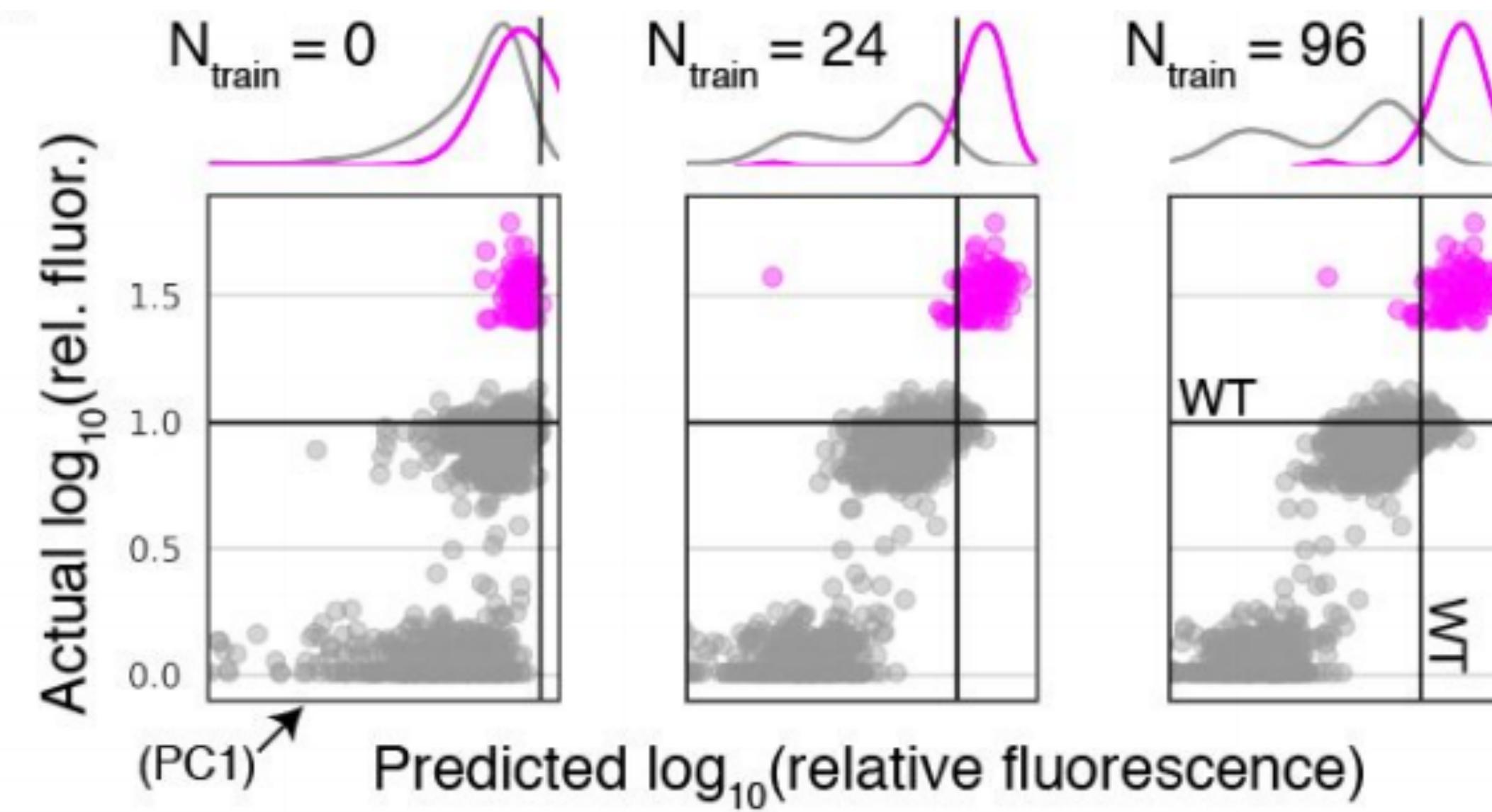
Pretraining guides search away from loss-of-function sequences



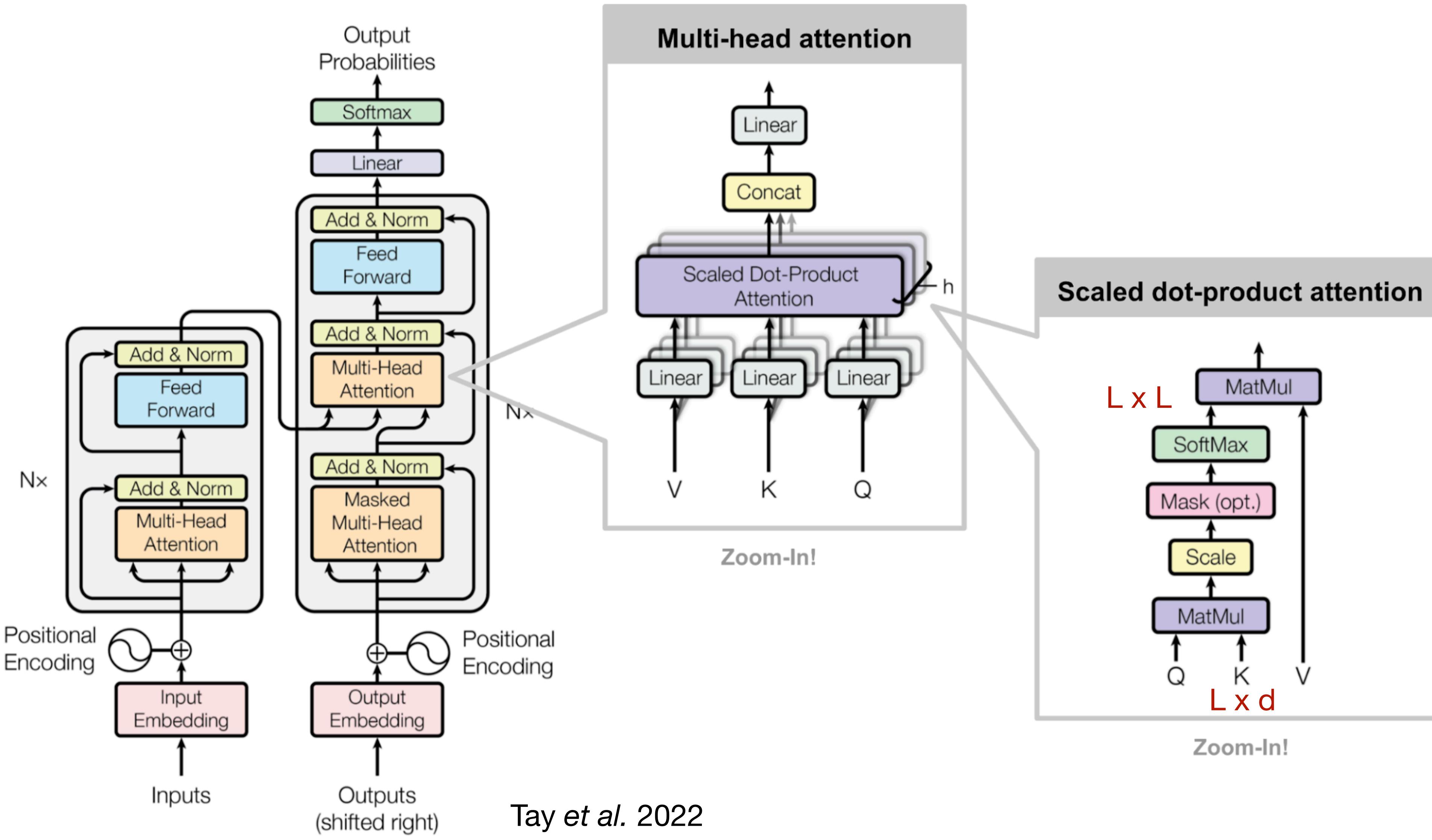
Pretraining guides search away from loss-of-function sequences



Pretraining guides search away from loss-of-function sequences

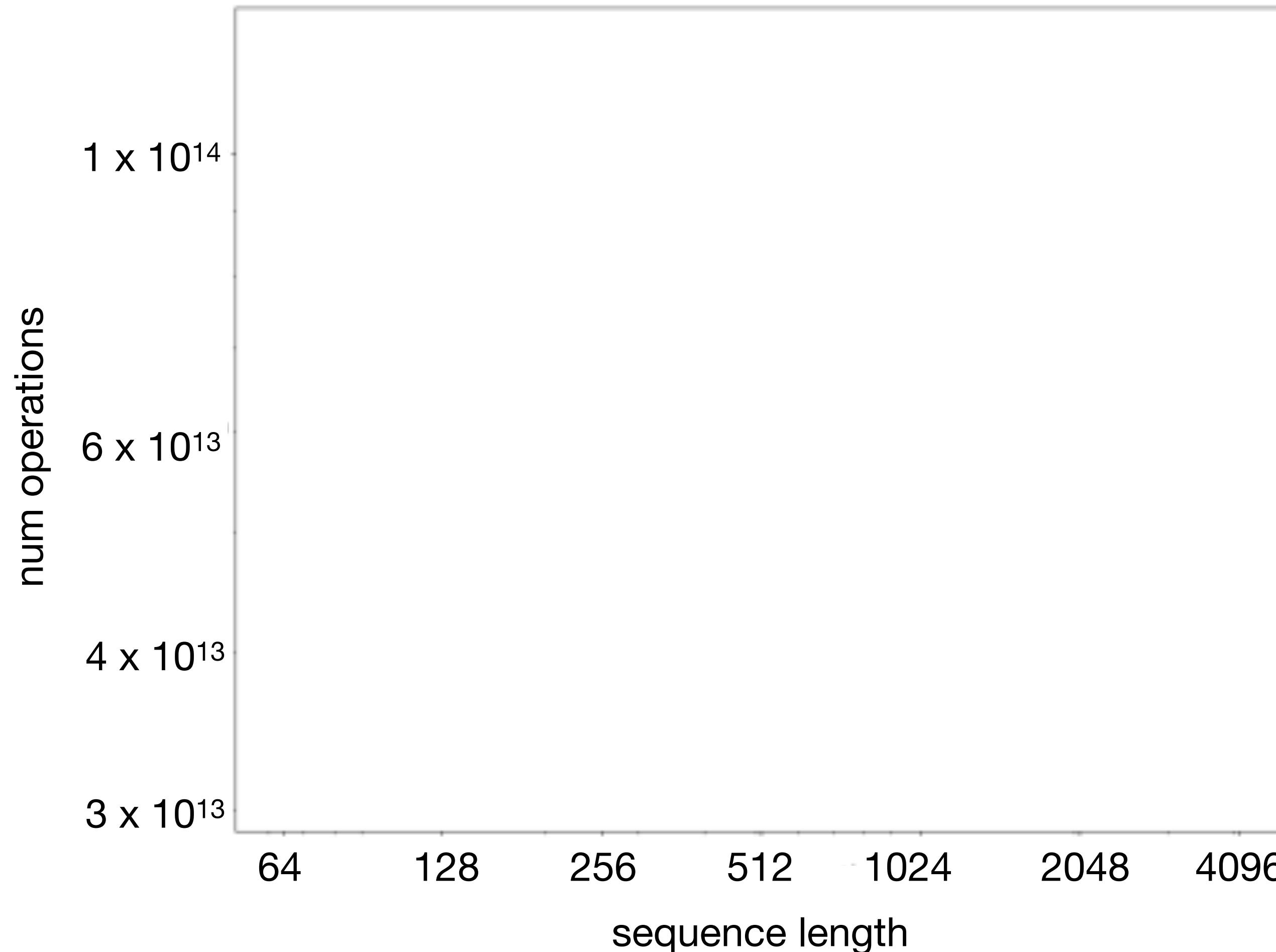


Transformers scale quadratically with length



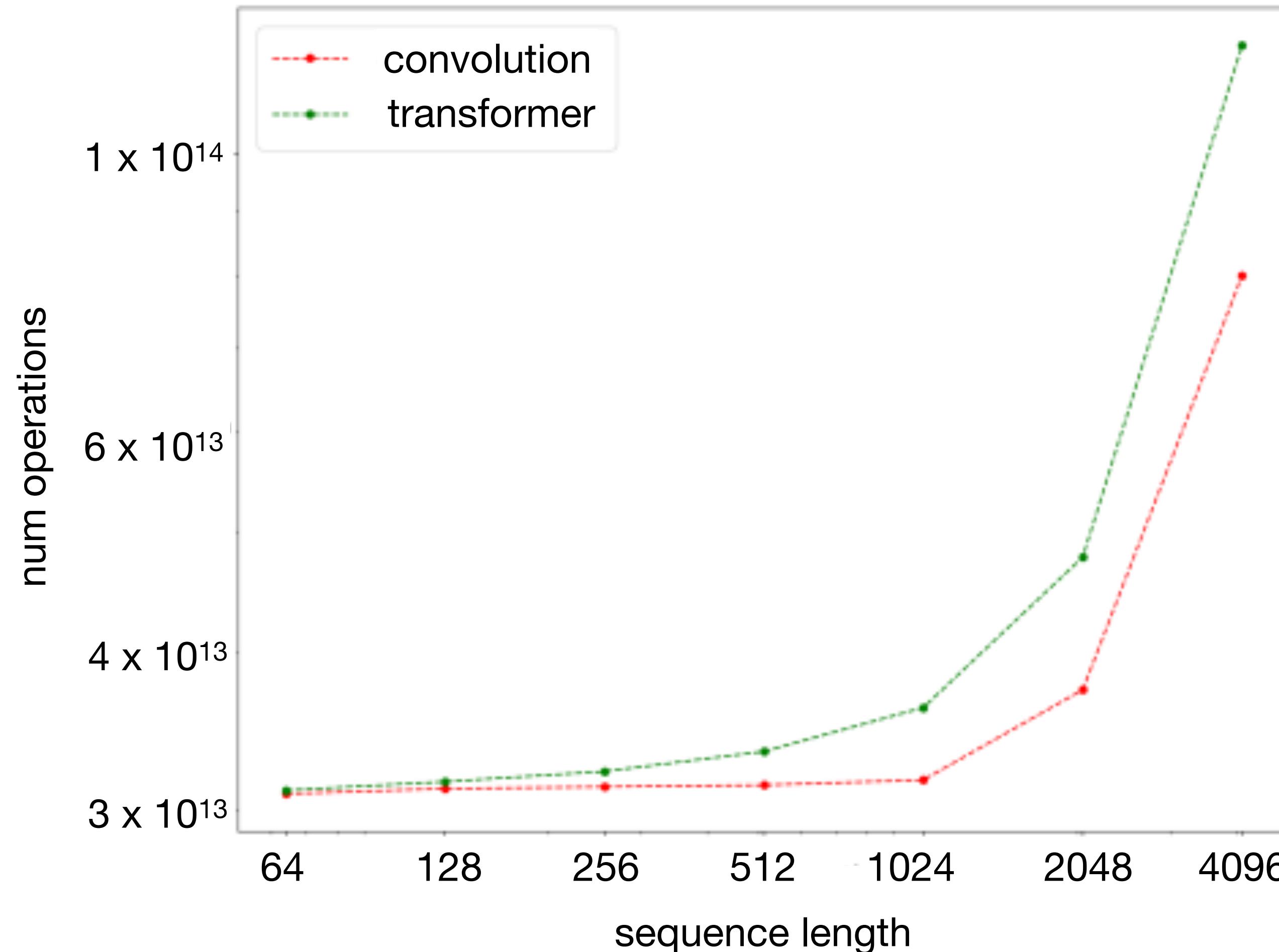
Tay et al. 2022

Convolution scales linearly with length



Tay et al. 2022

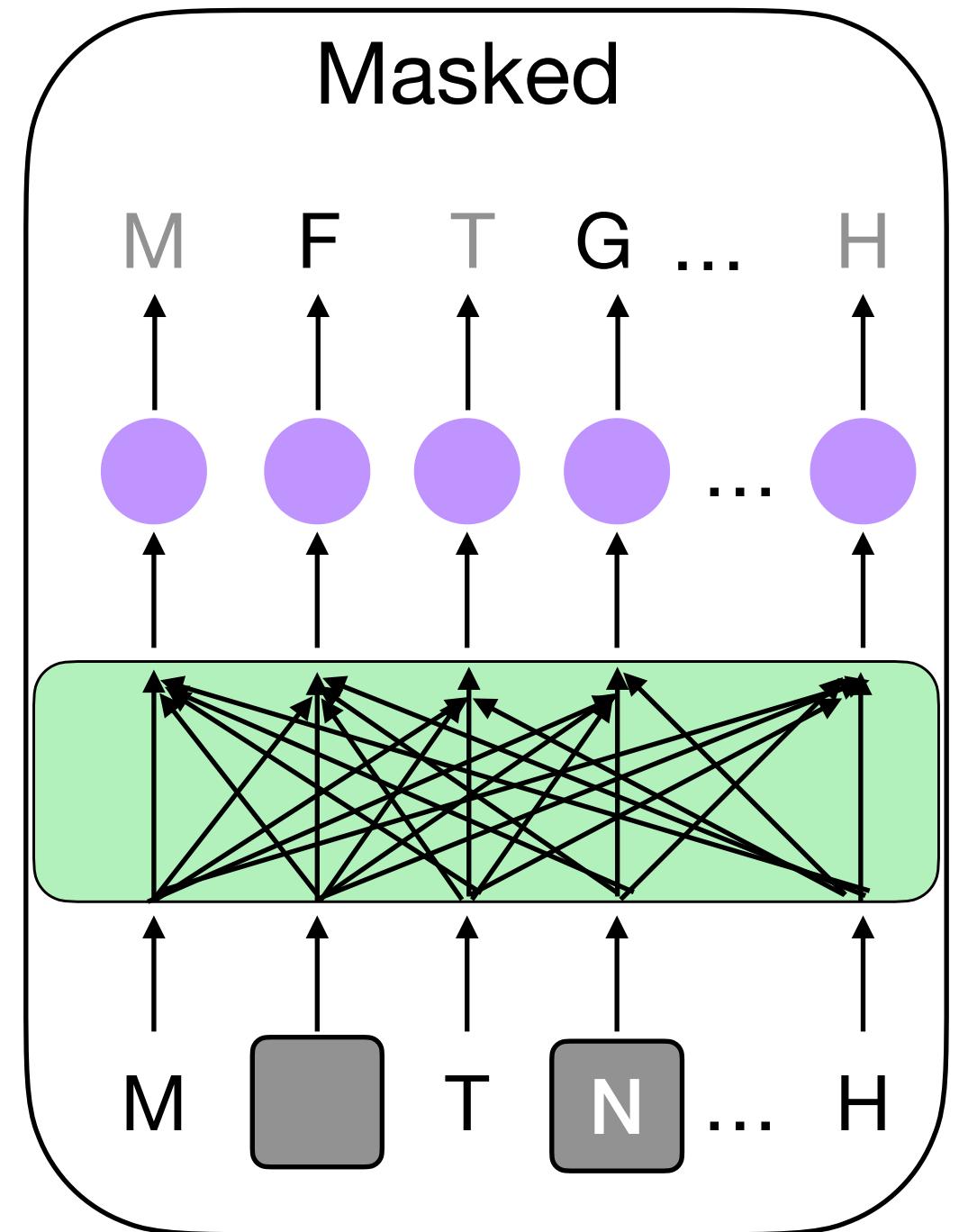
Convolution scales linearly with length



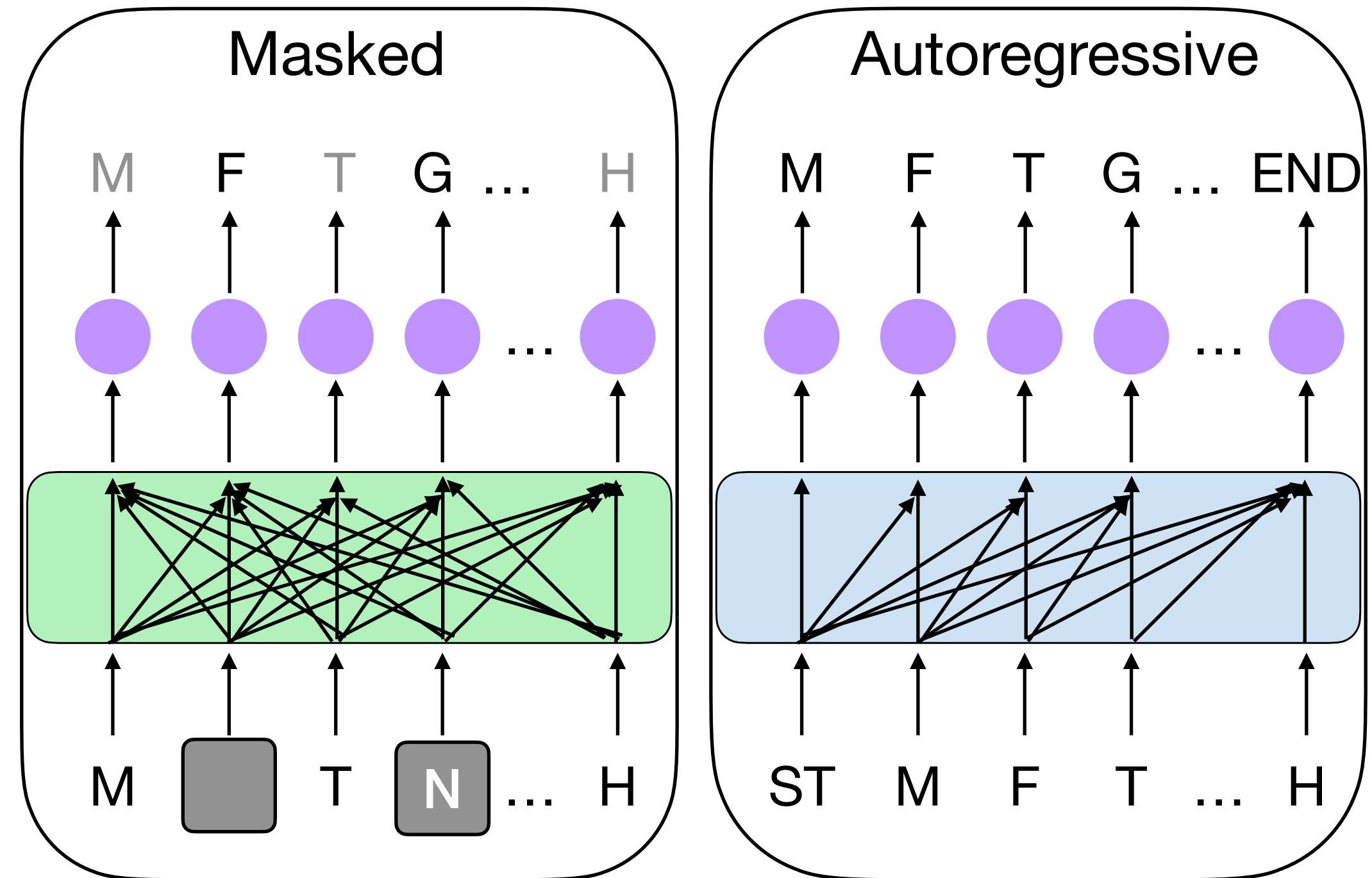
Tay et al. 2022

Separate pretraining task

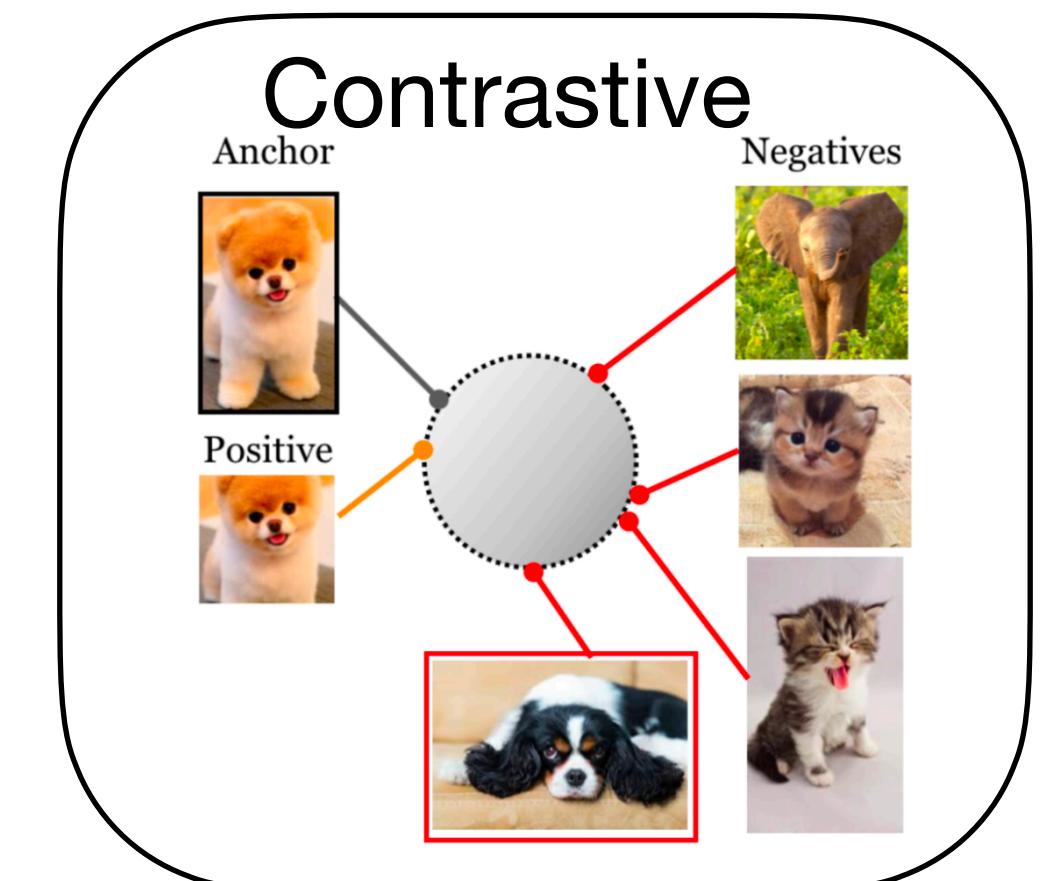
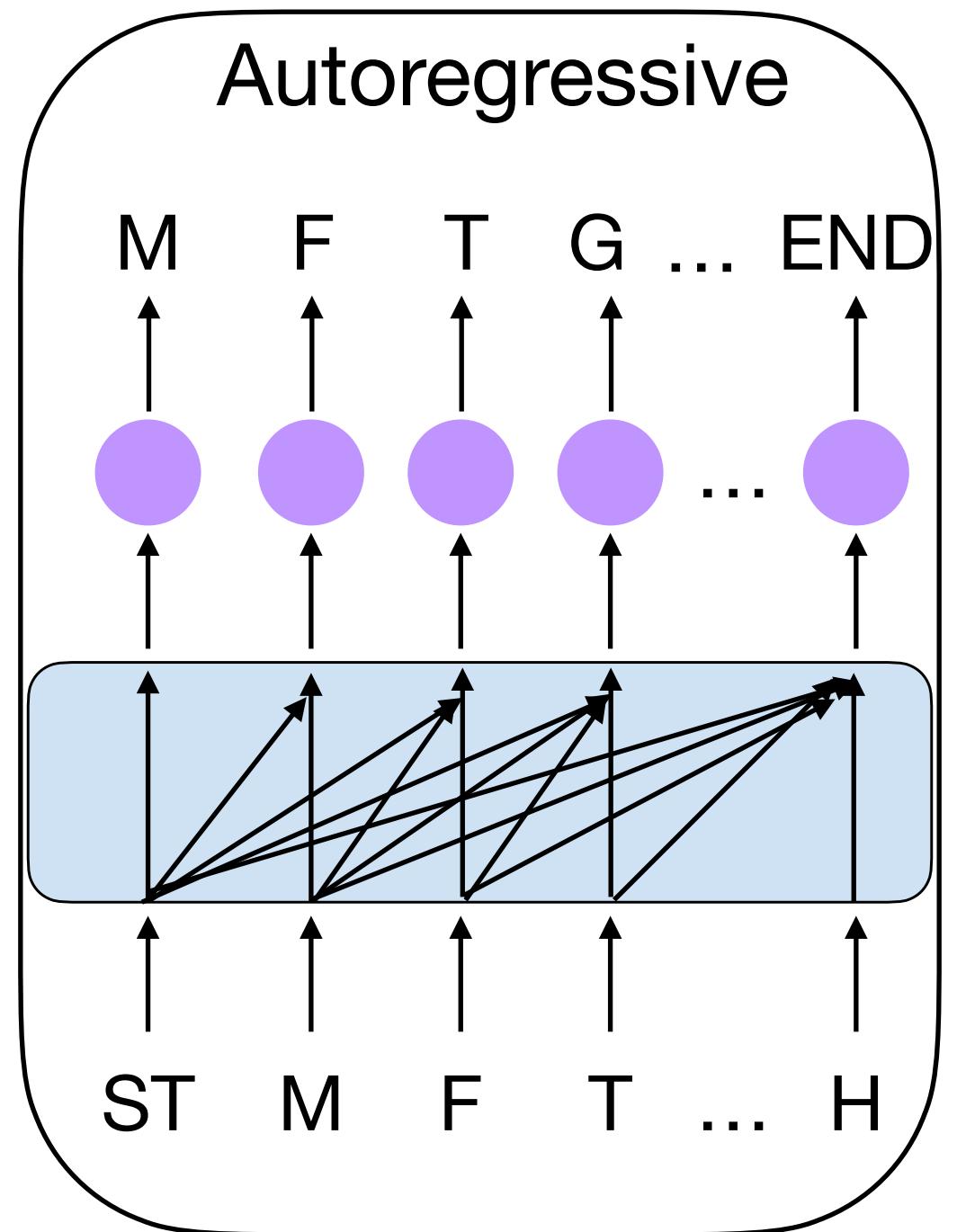
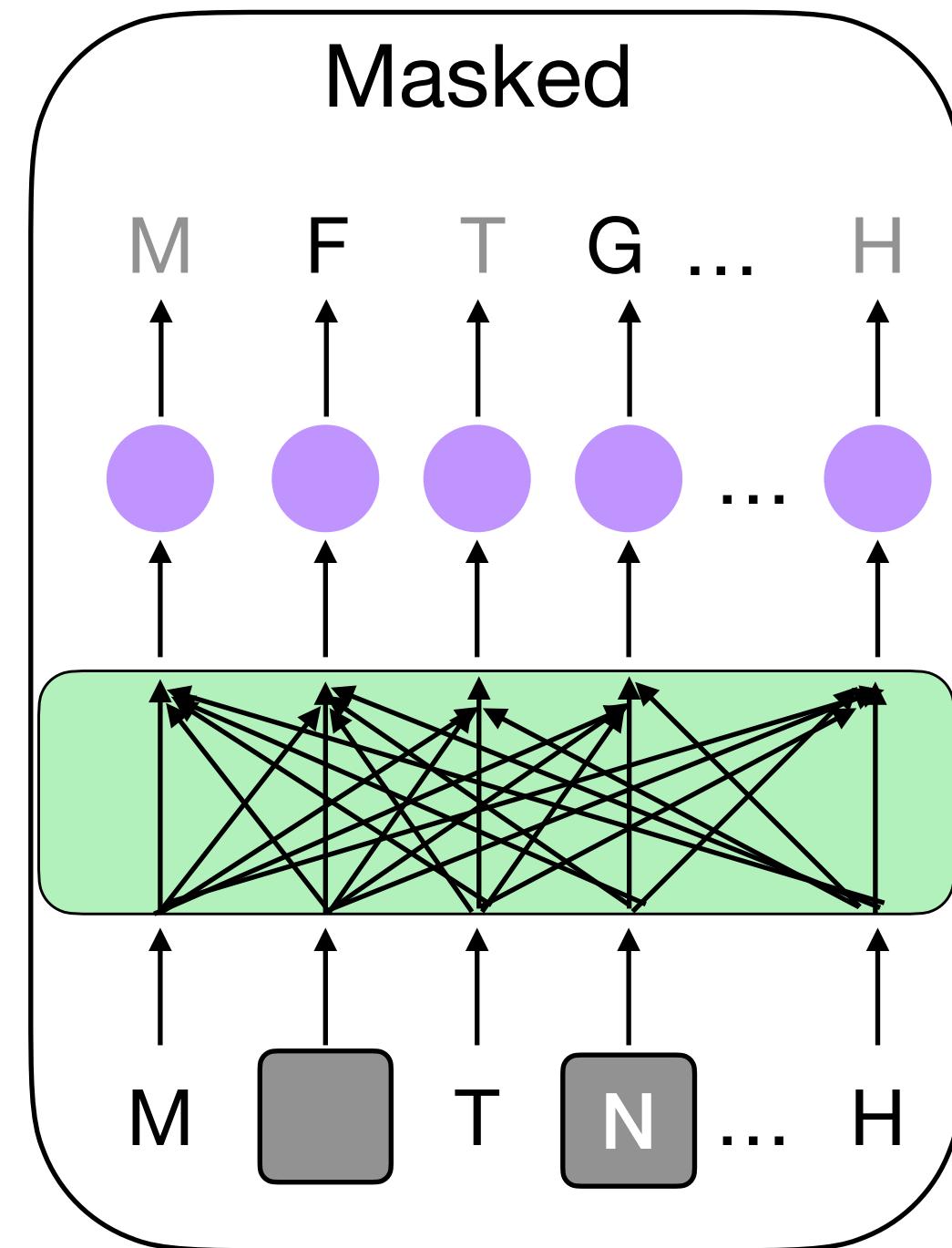
Separate pretraining task



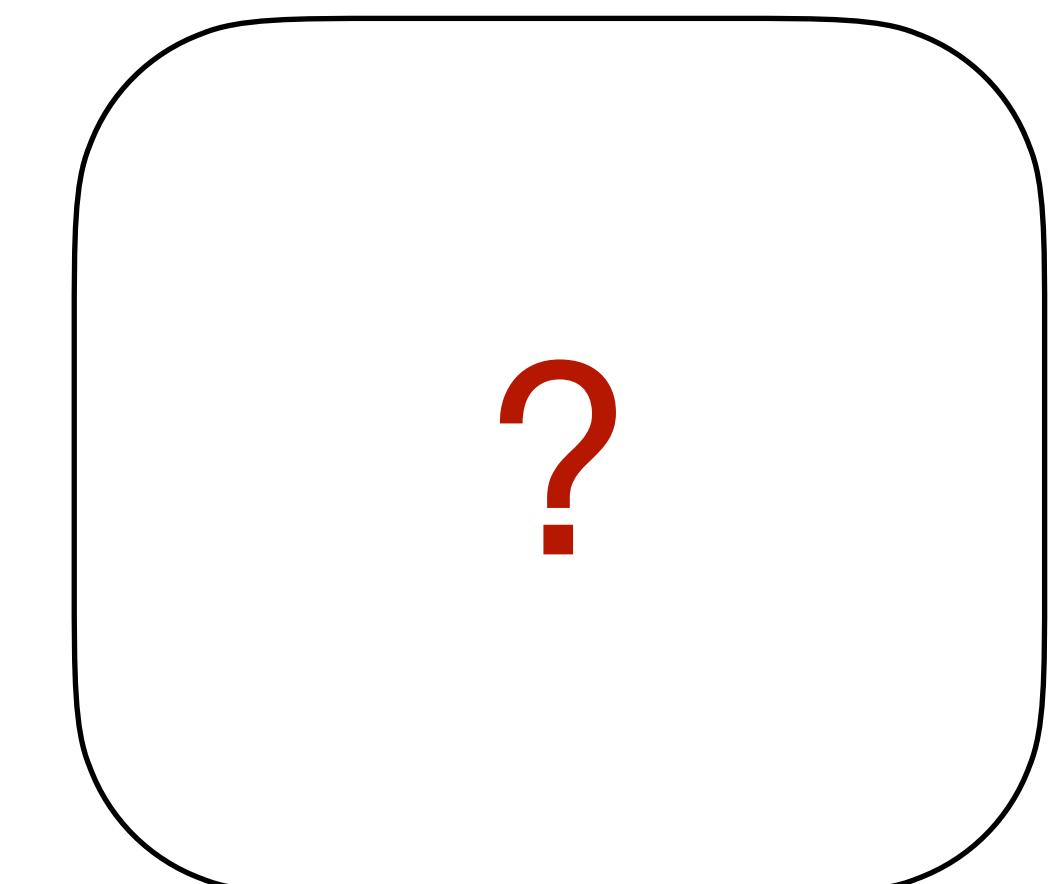
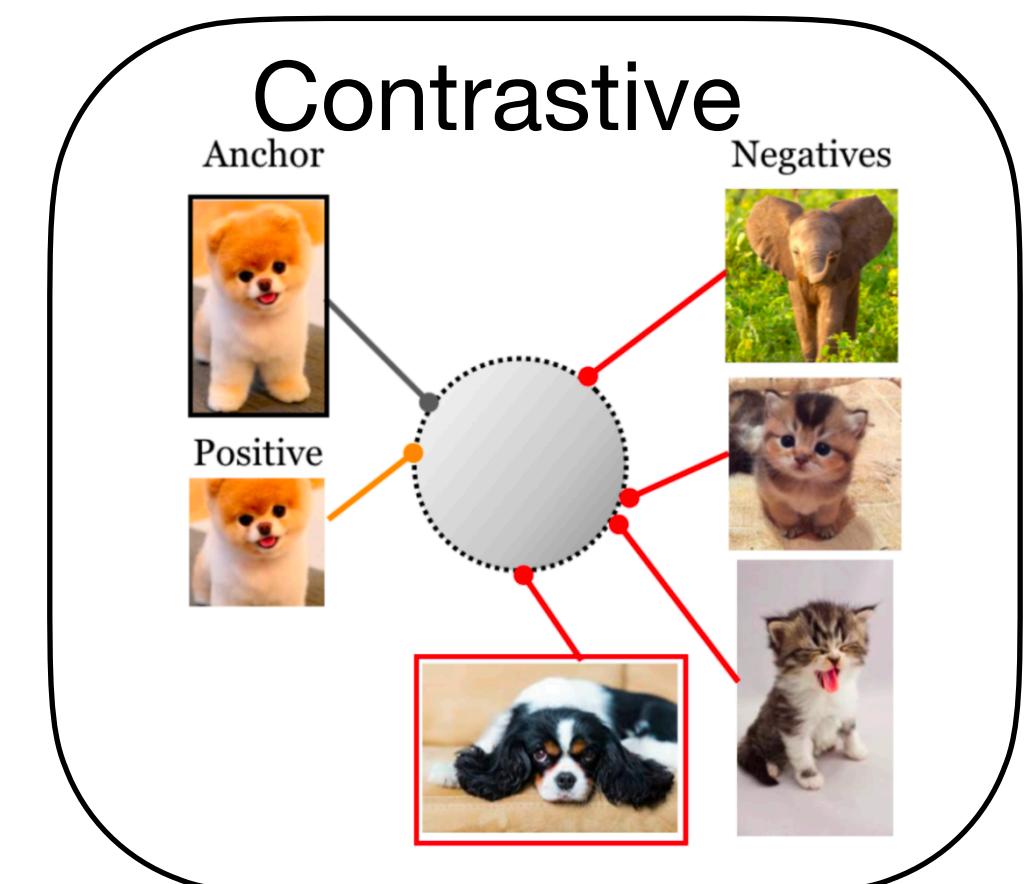
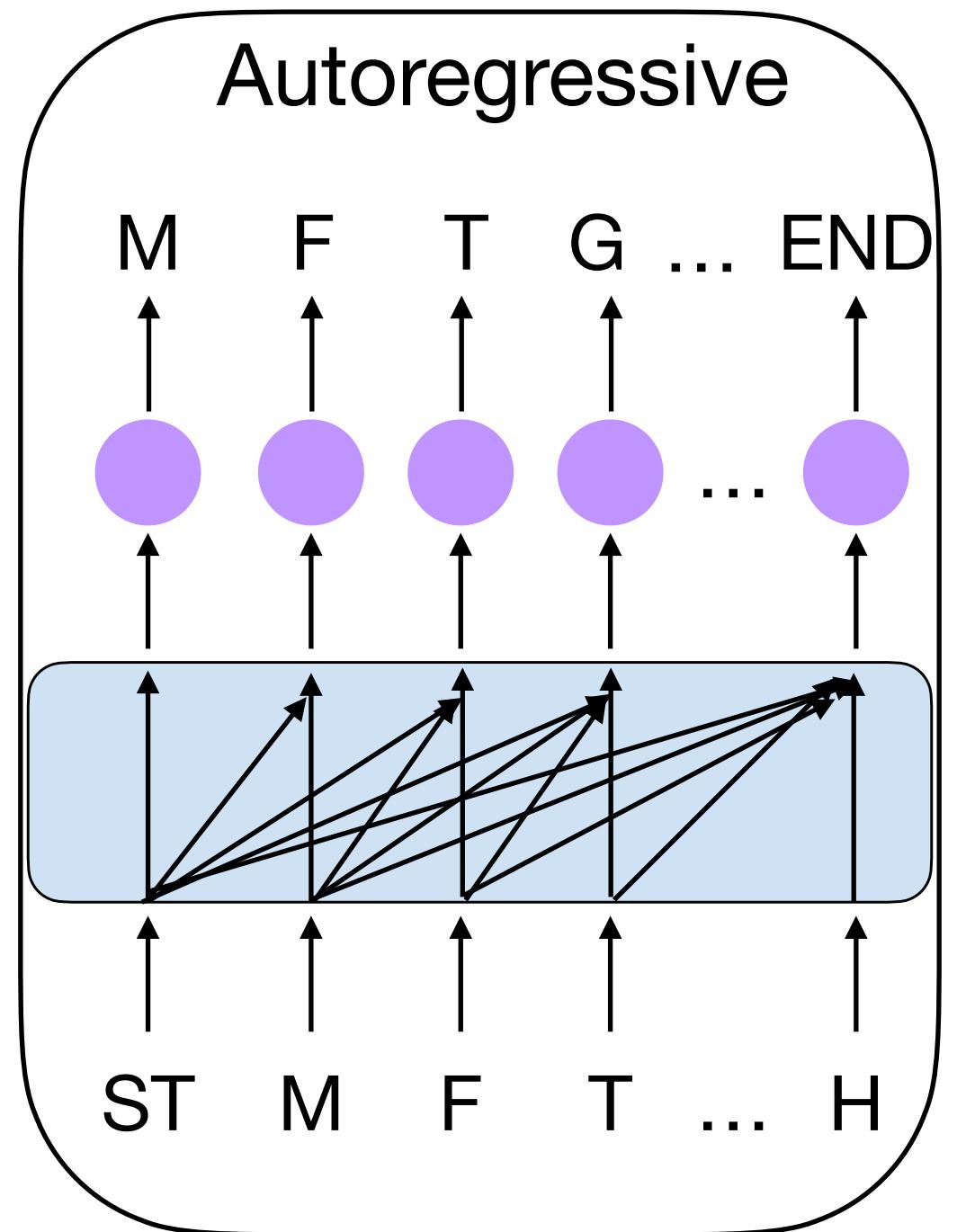
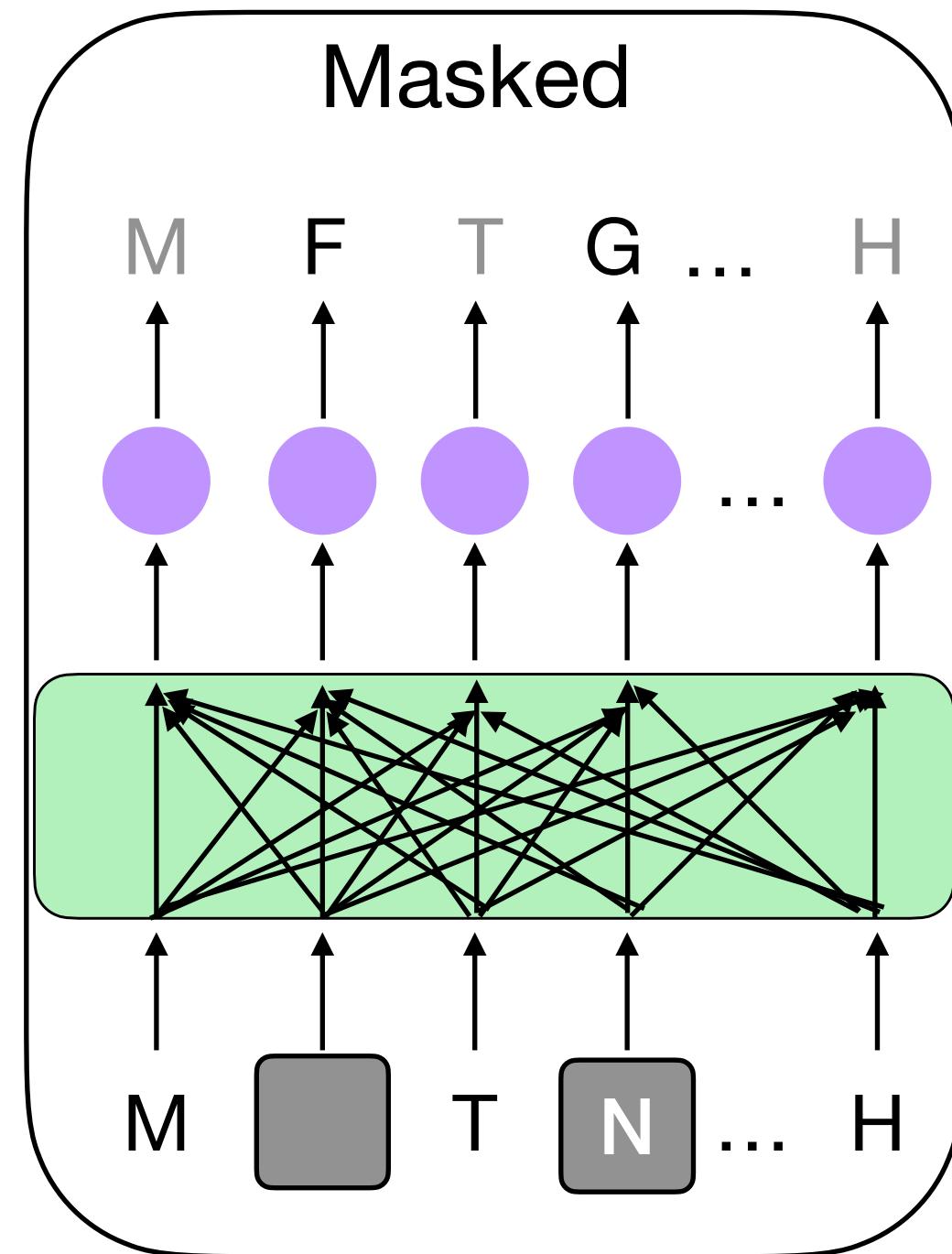
Separate pretraining task



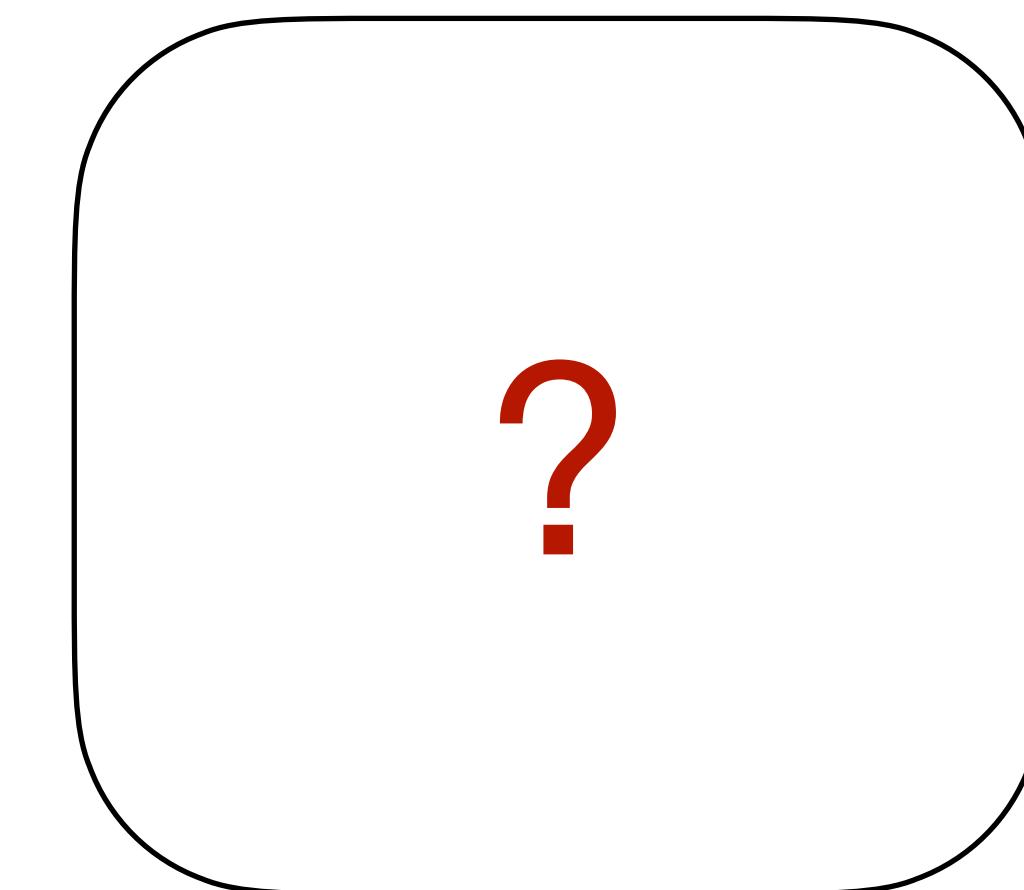
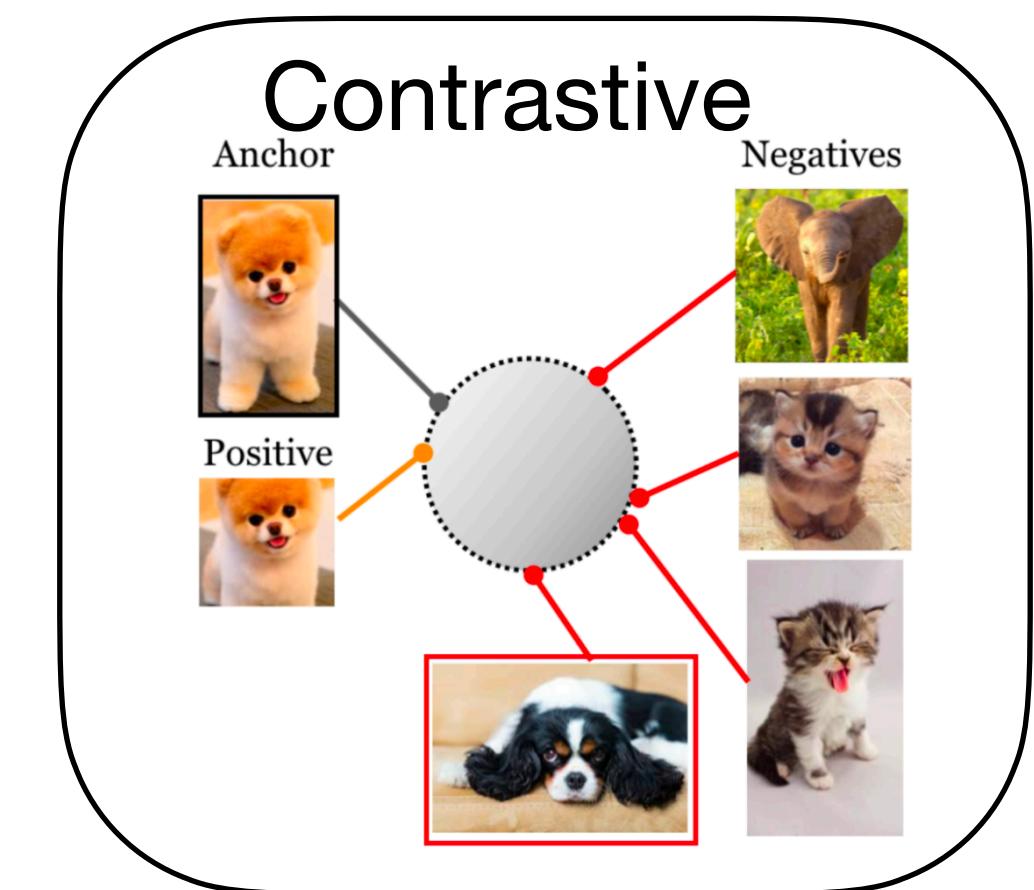
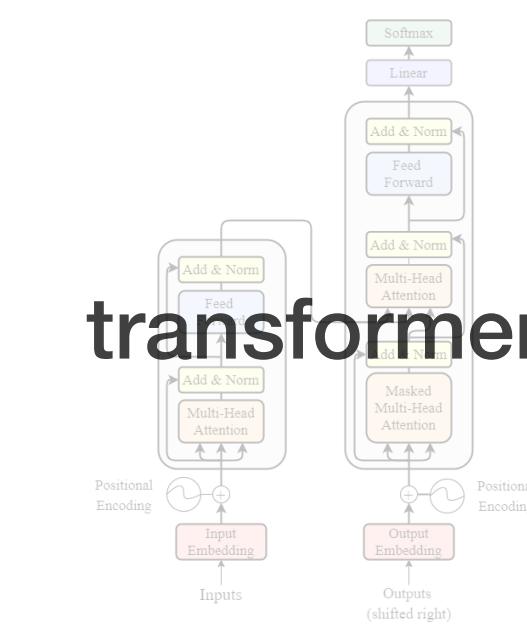
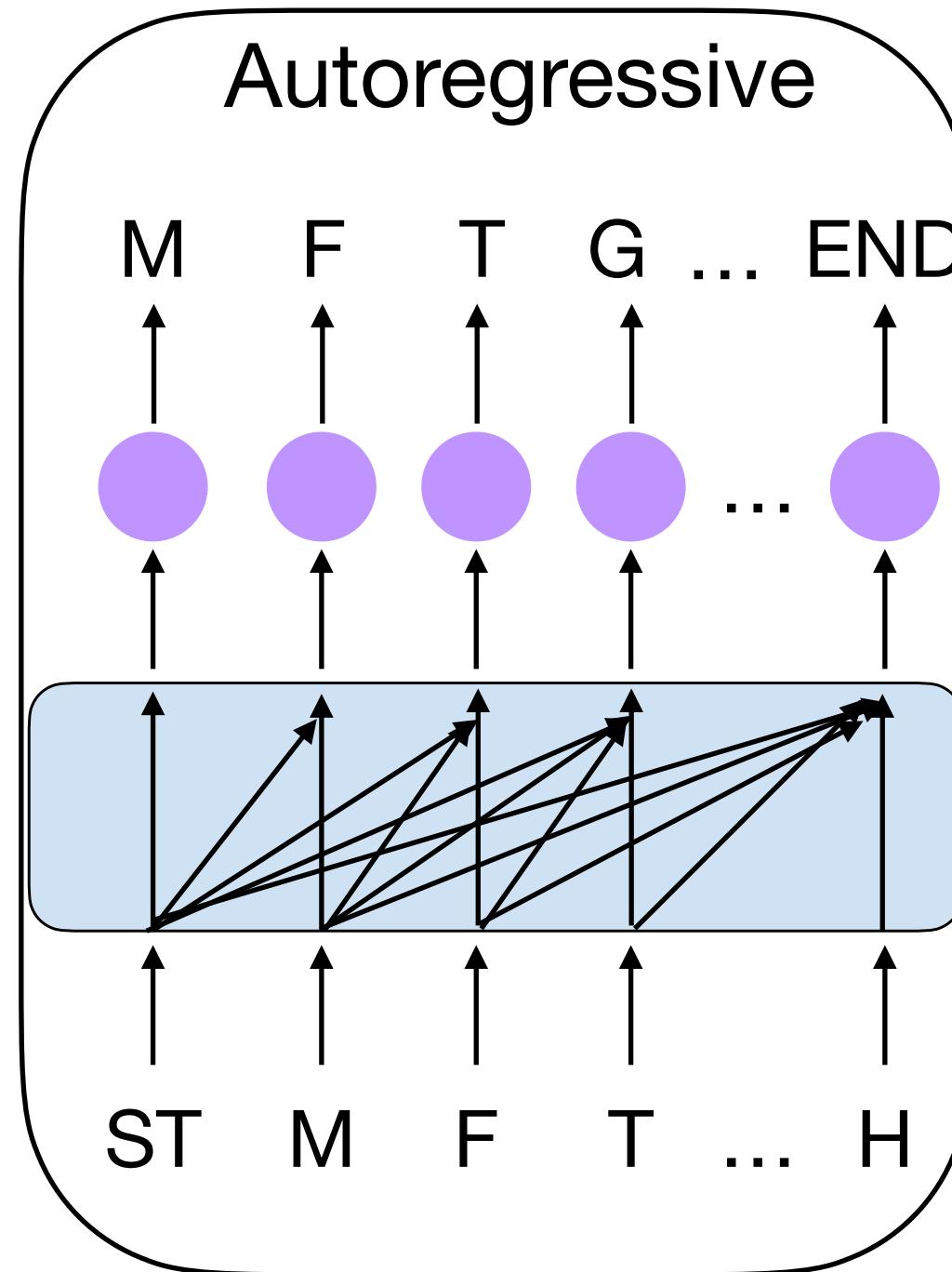
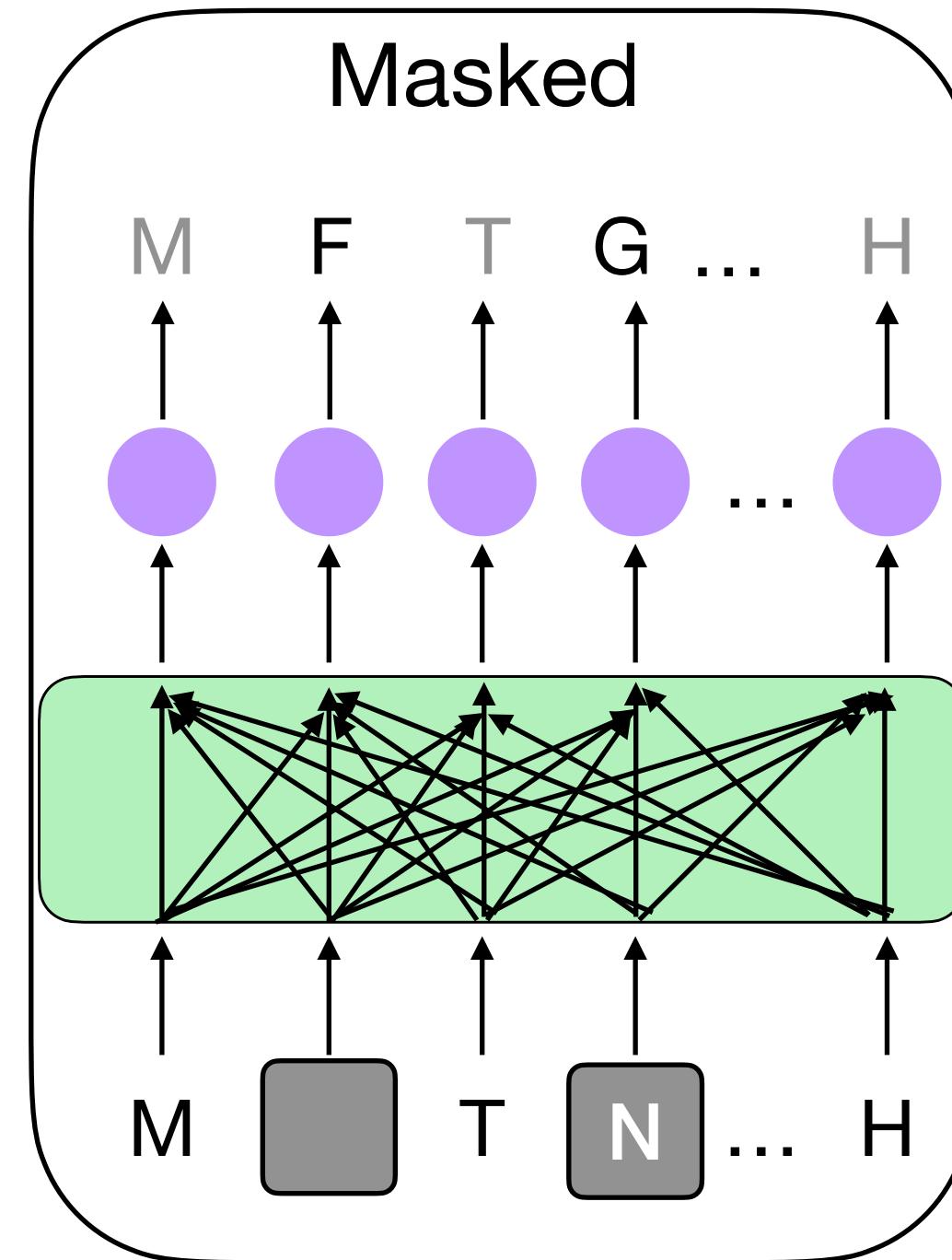
Separate pretraining task



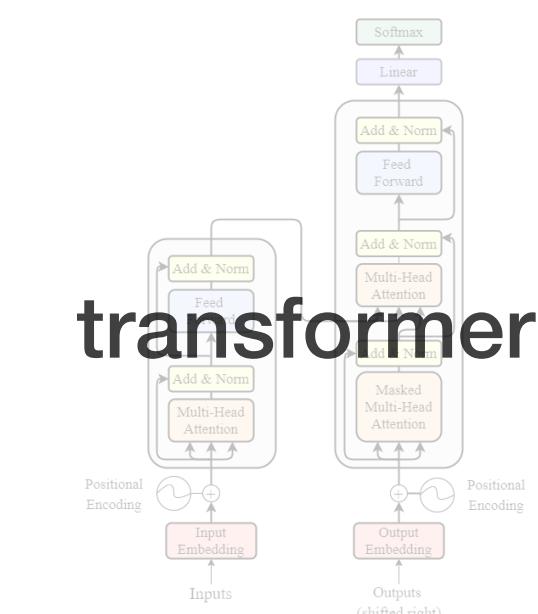
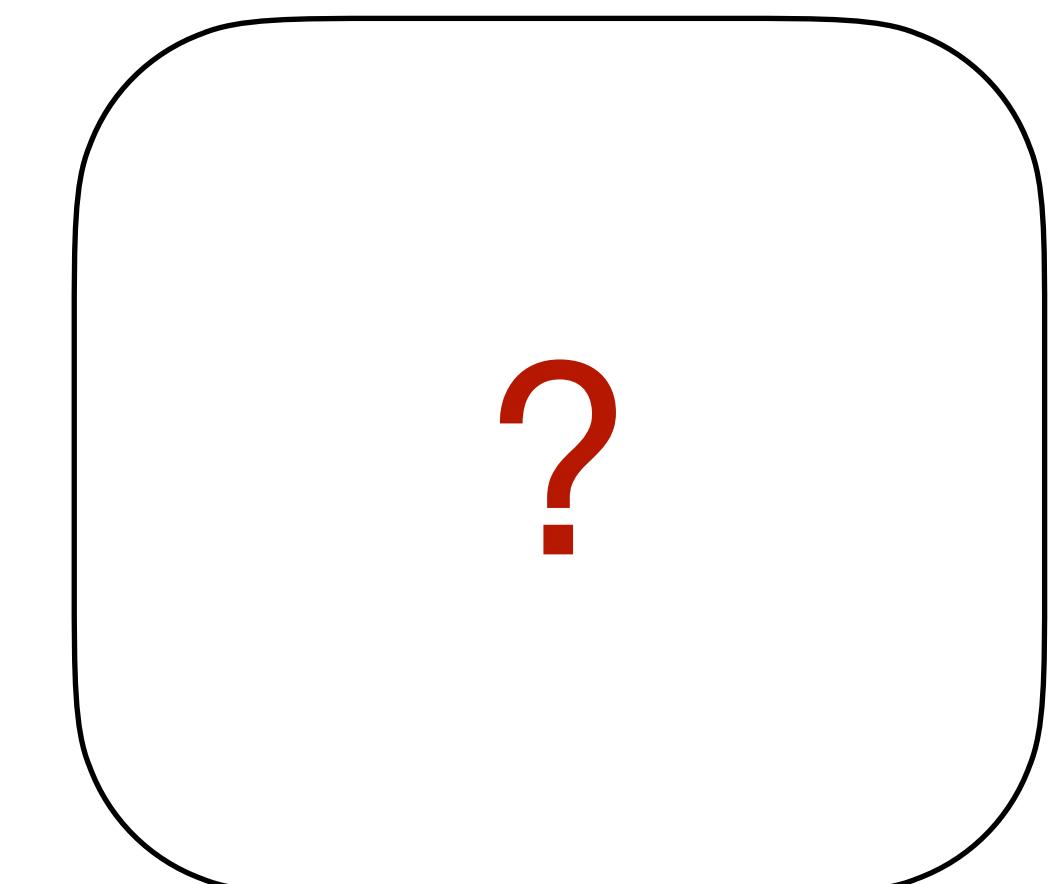
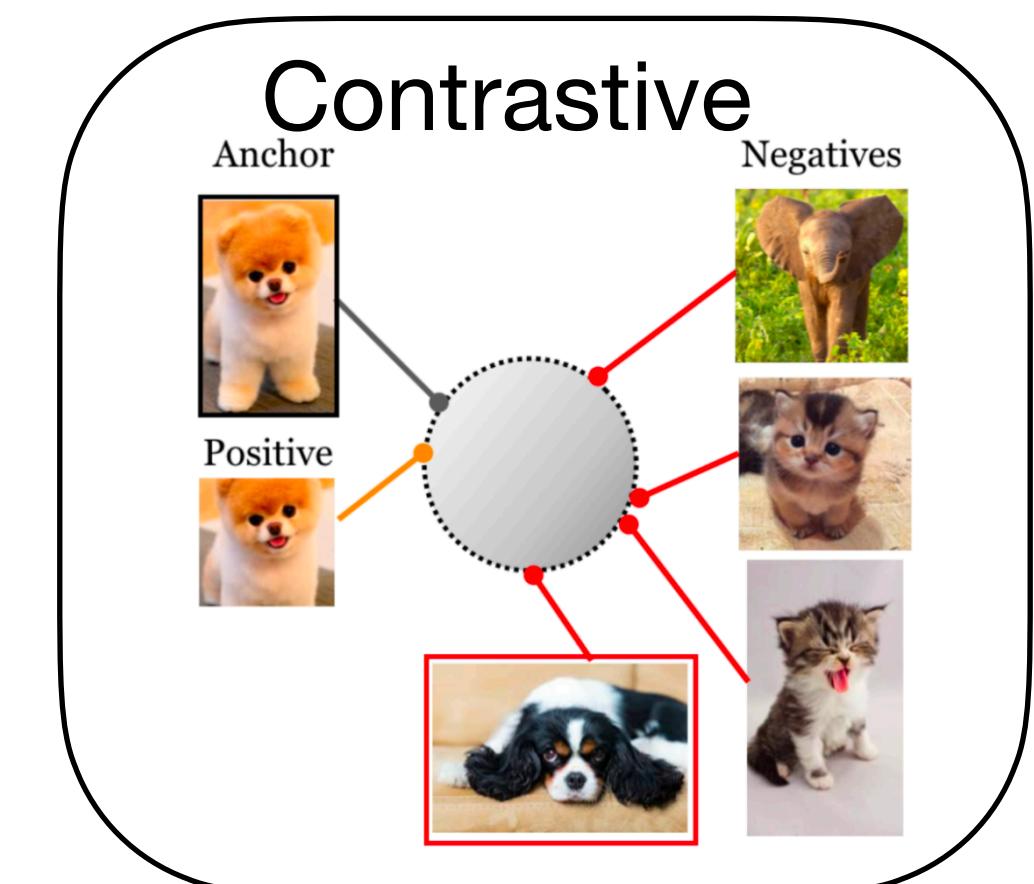
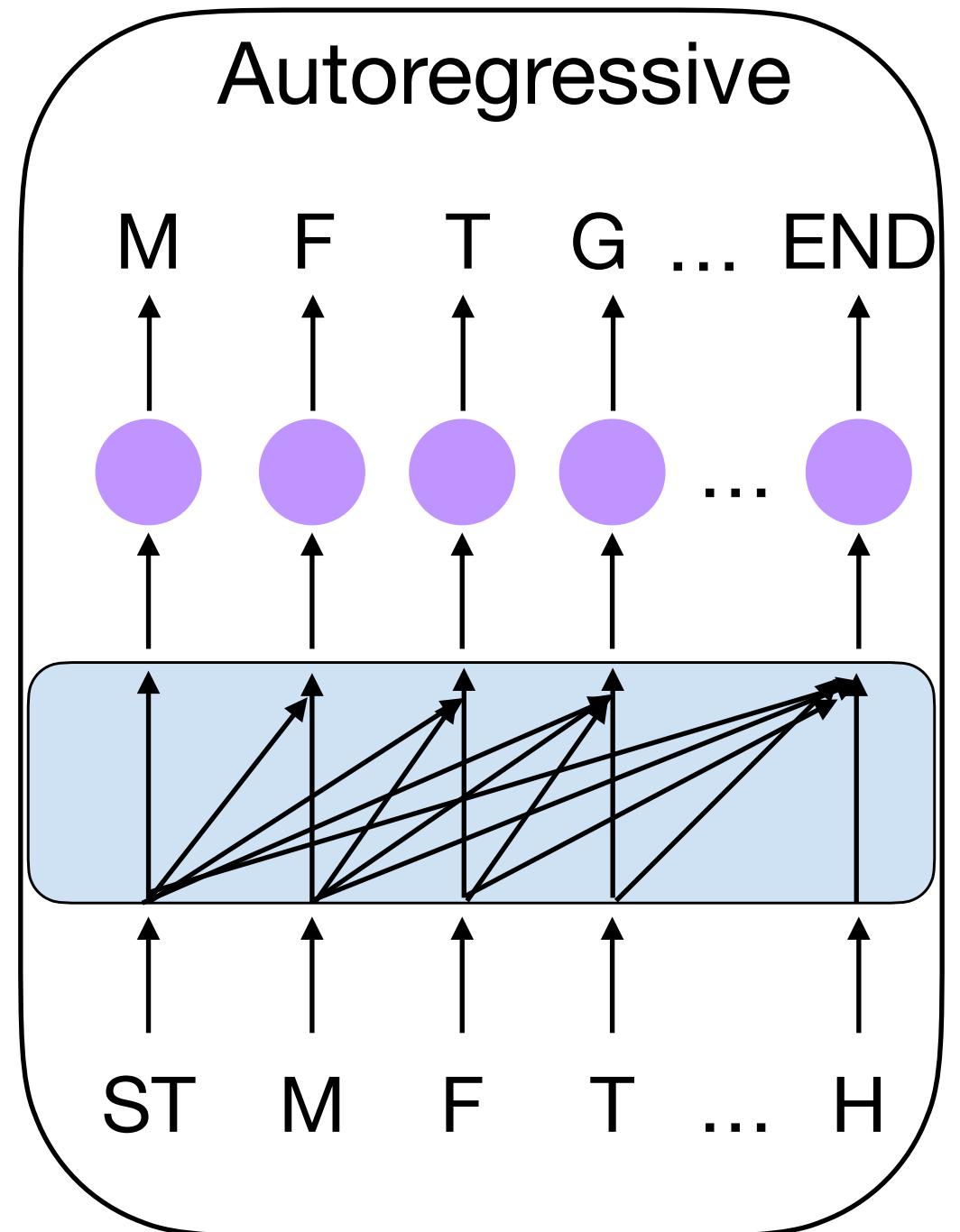
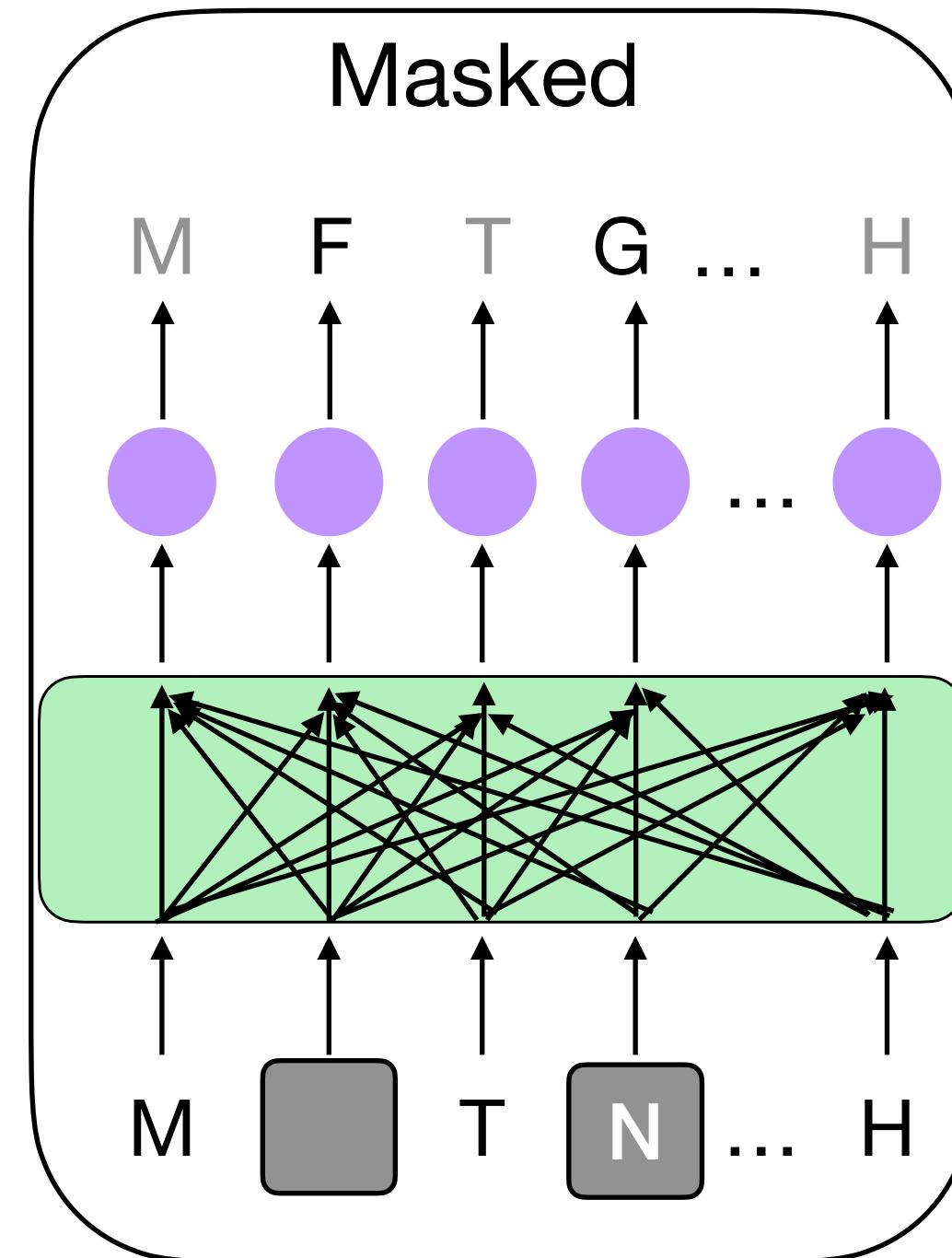
Separate pretraining task



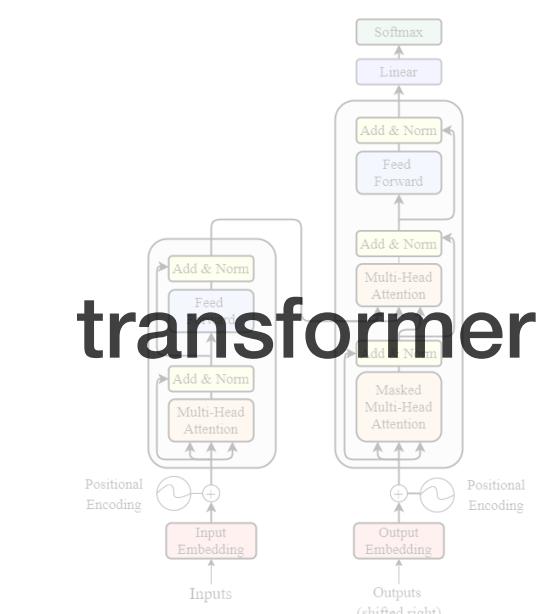
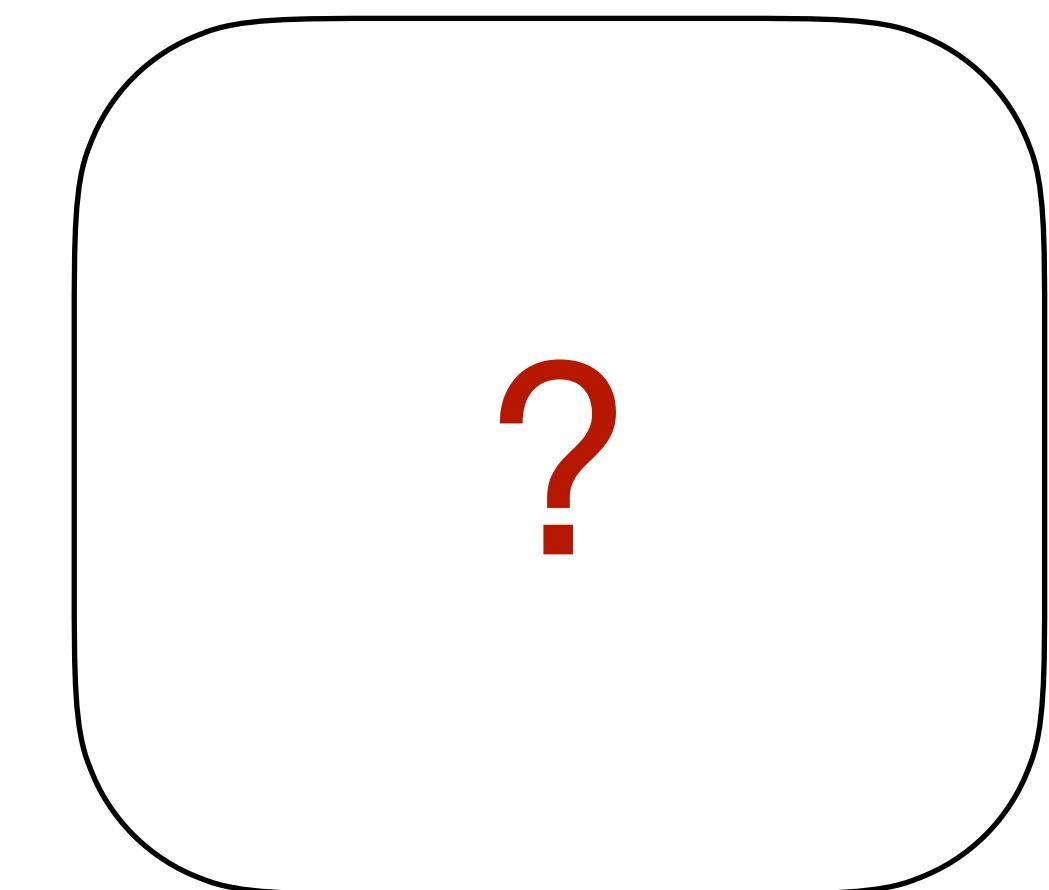
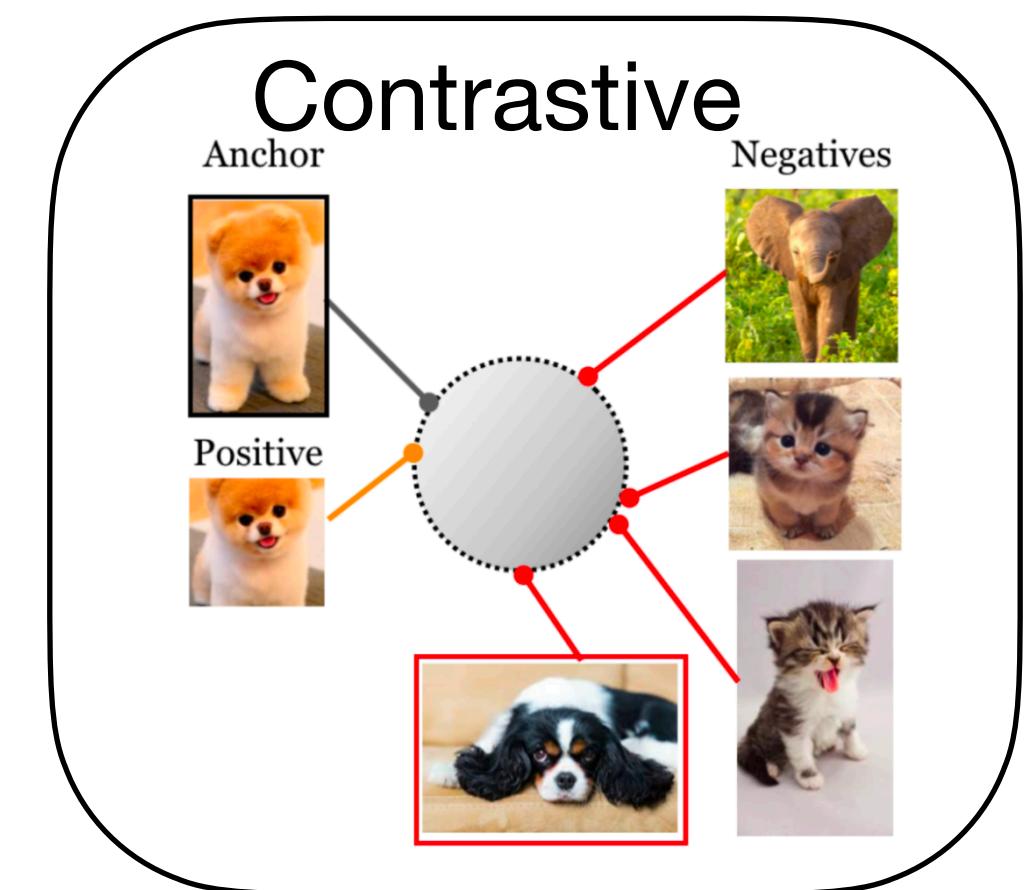
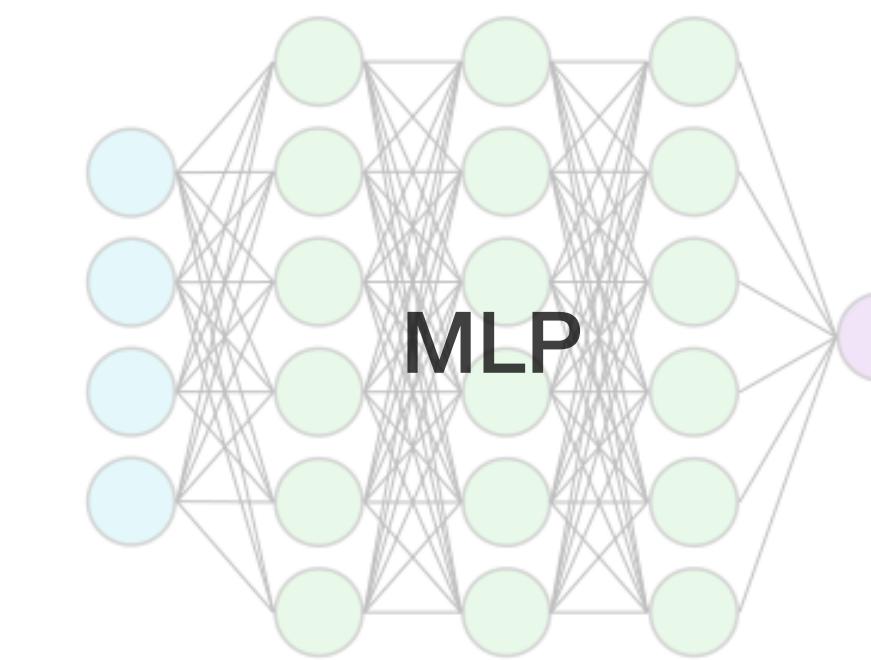
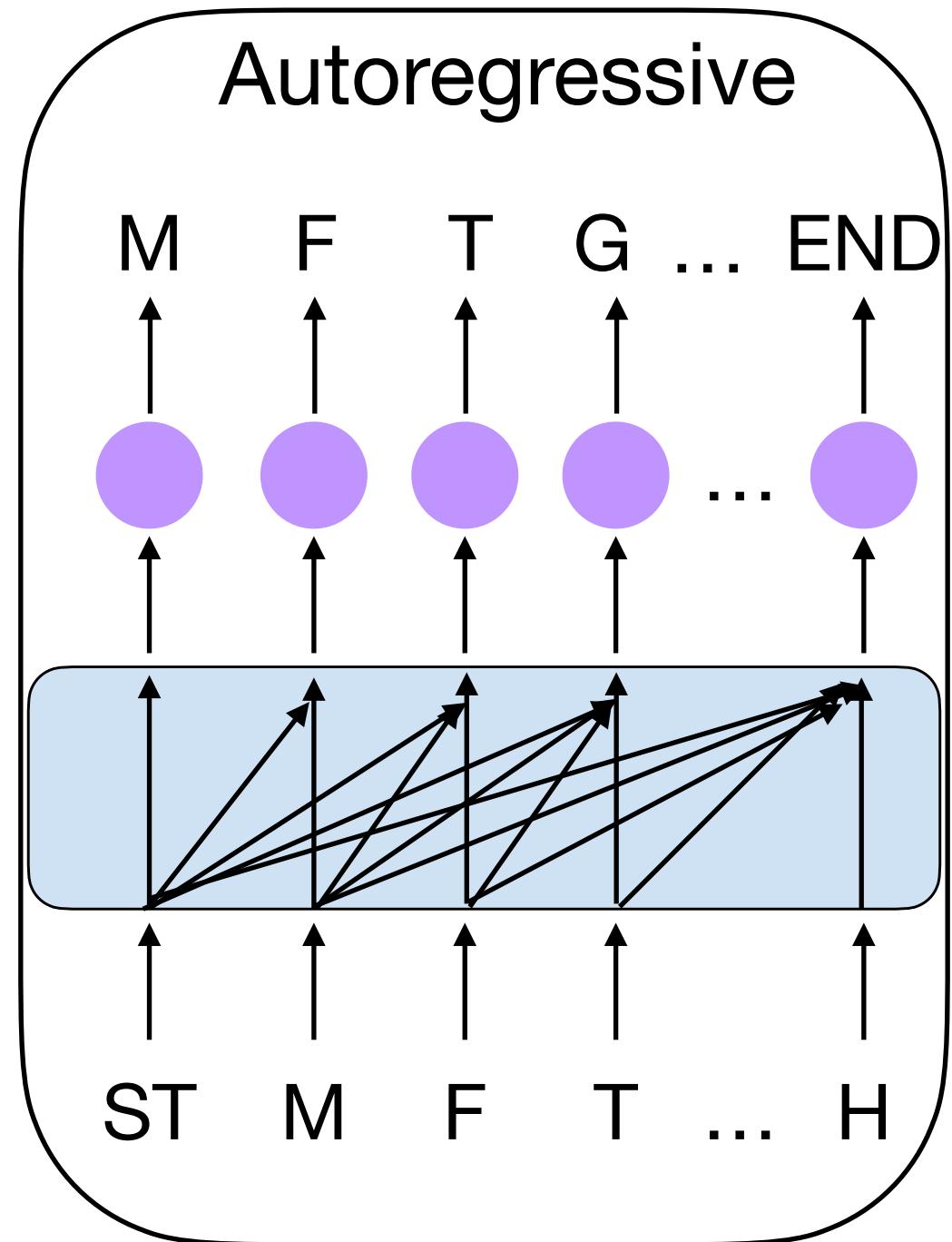
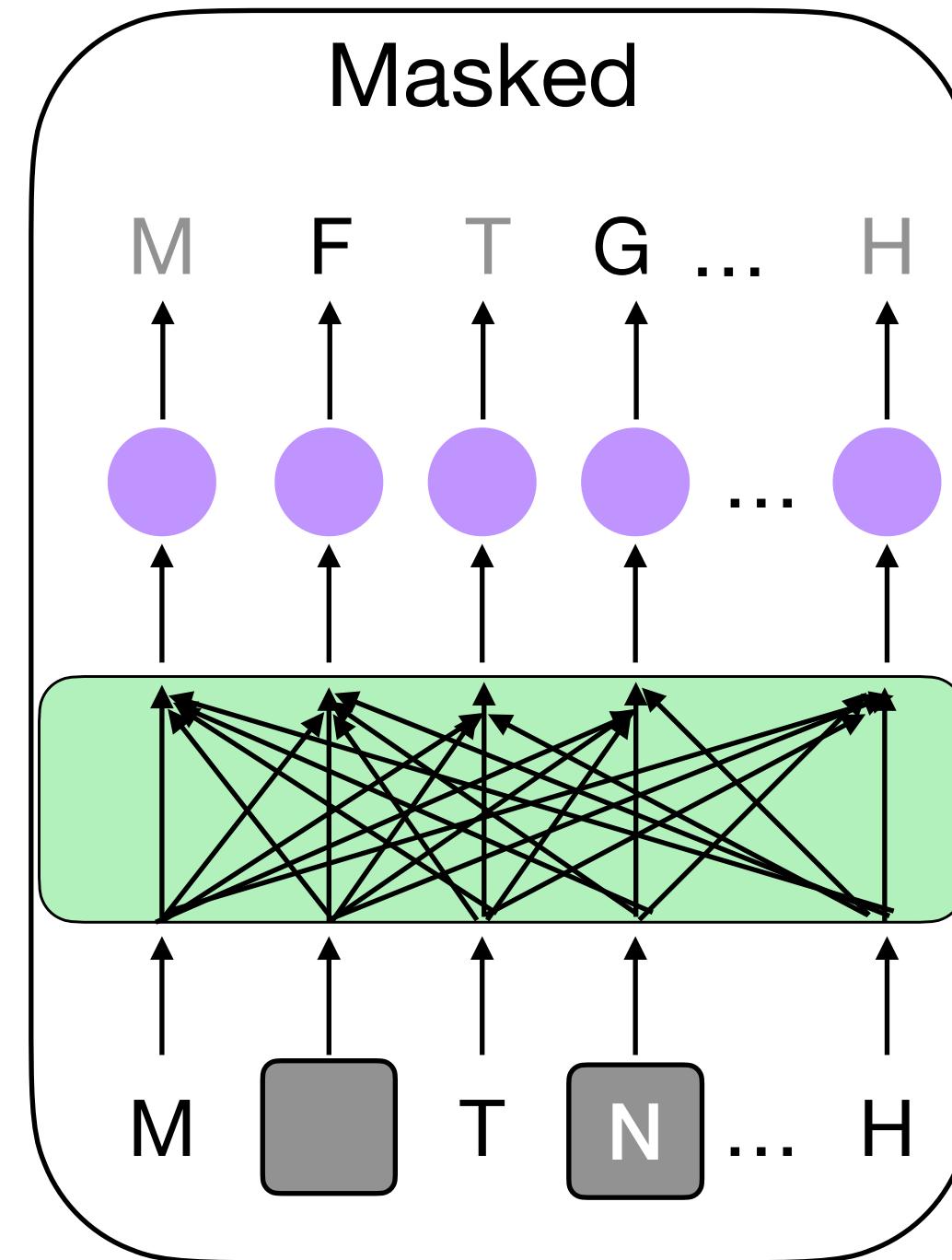
Separate pretraining task and architecture



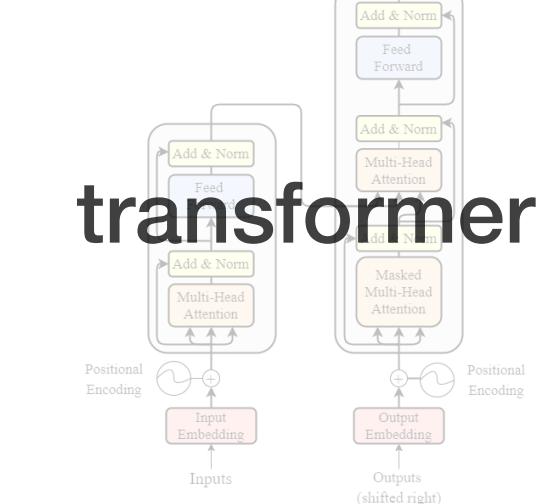
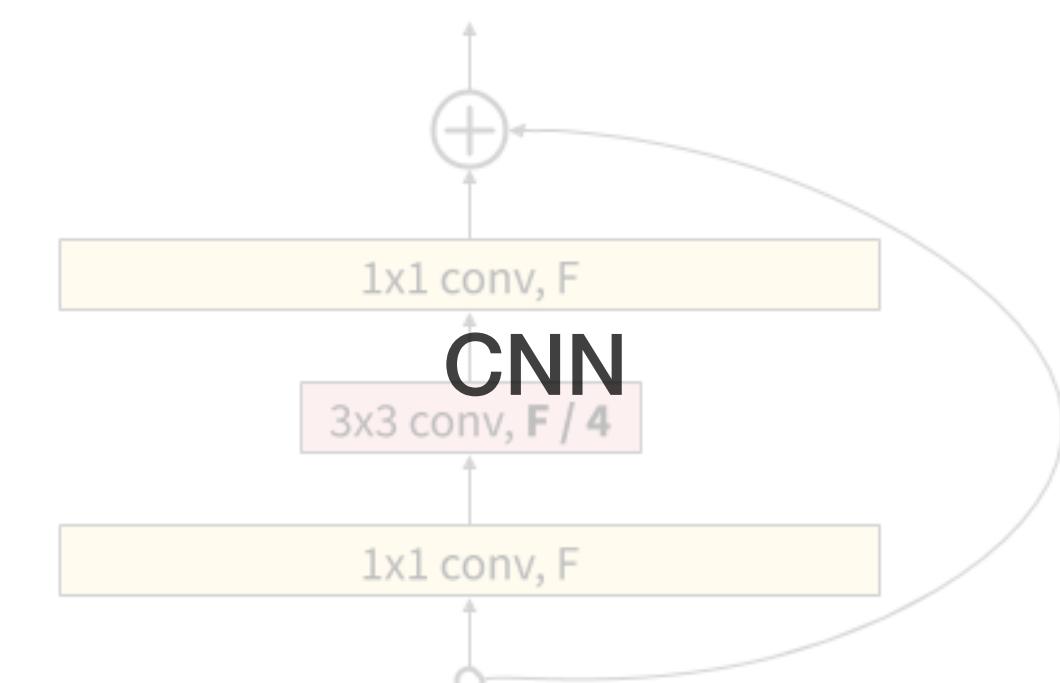
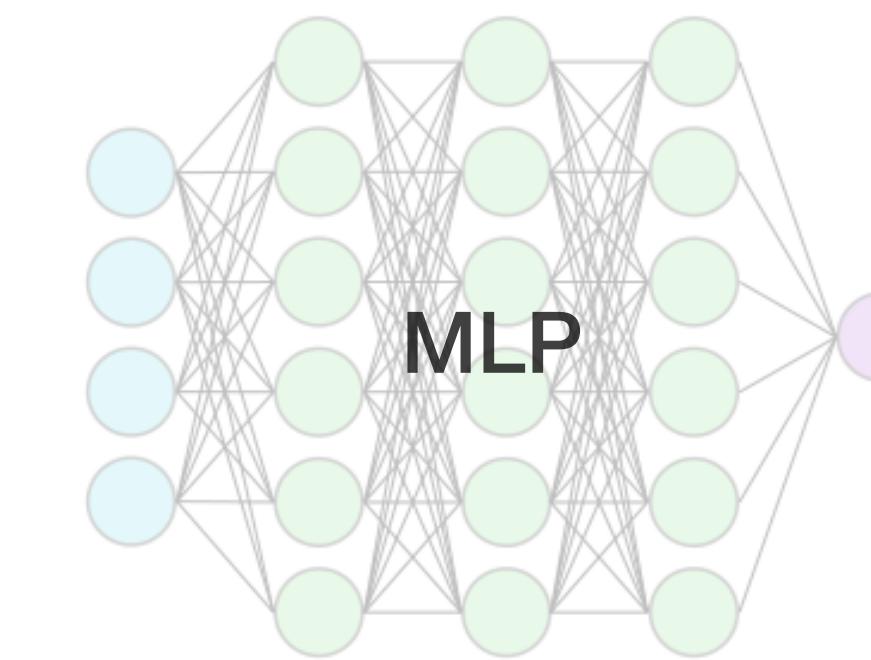
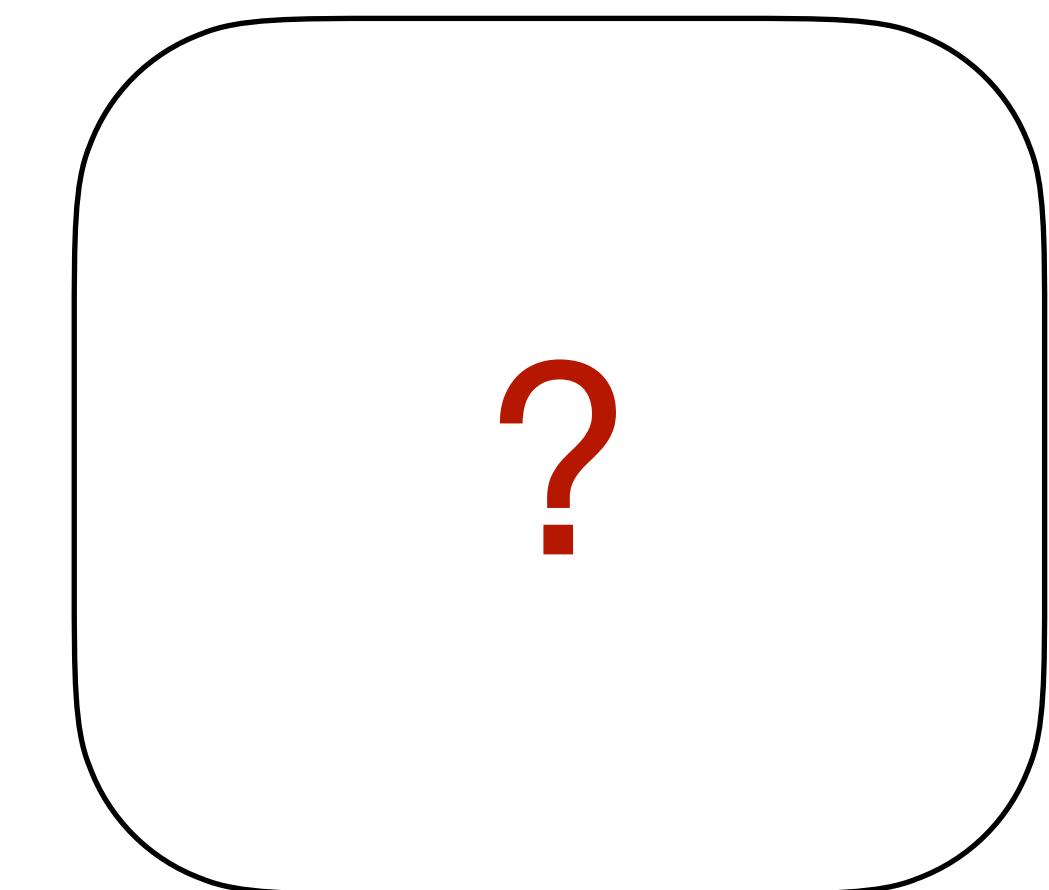
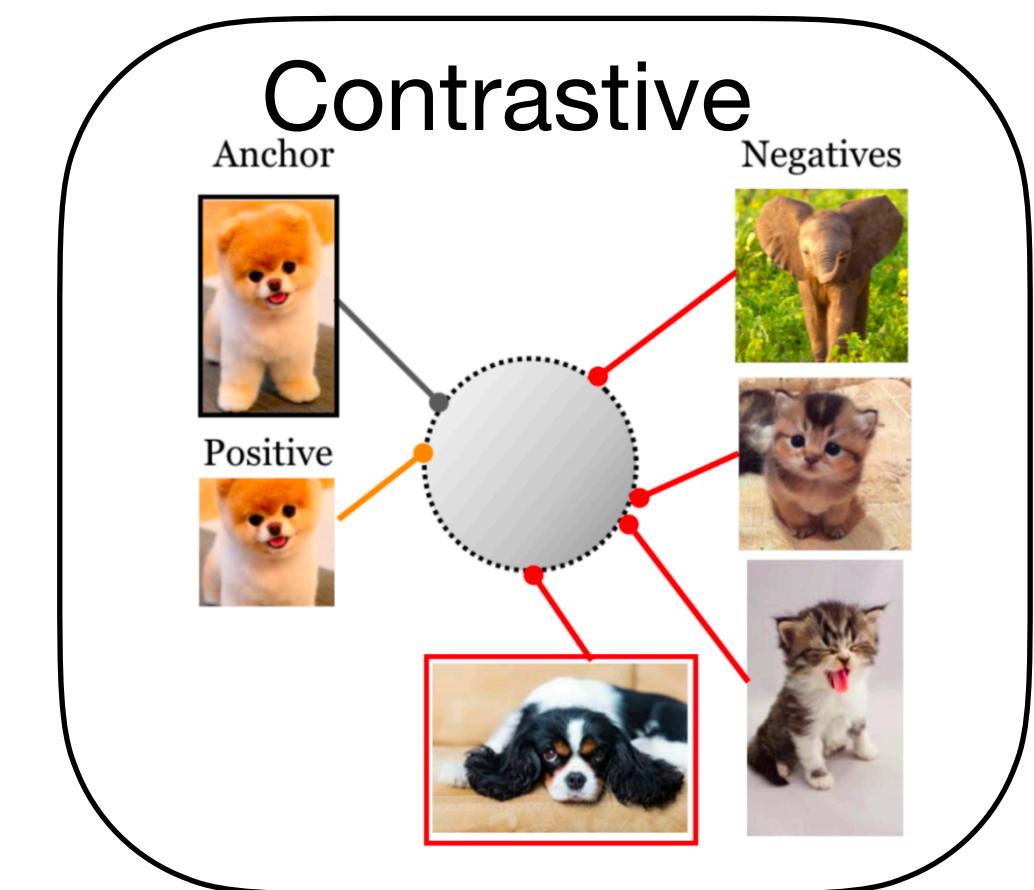
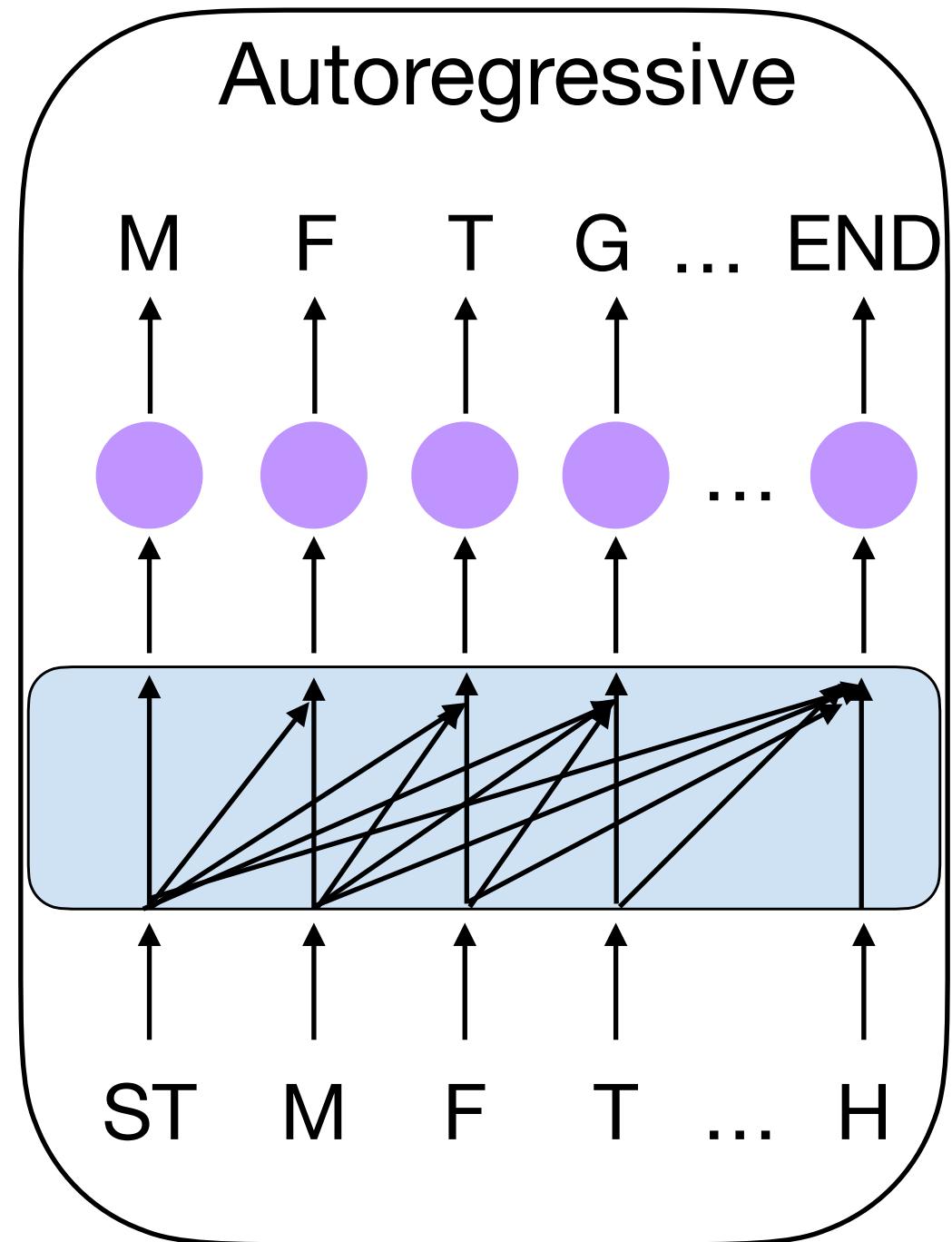
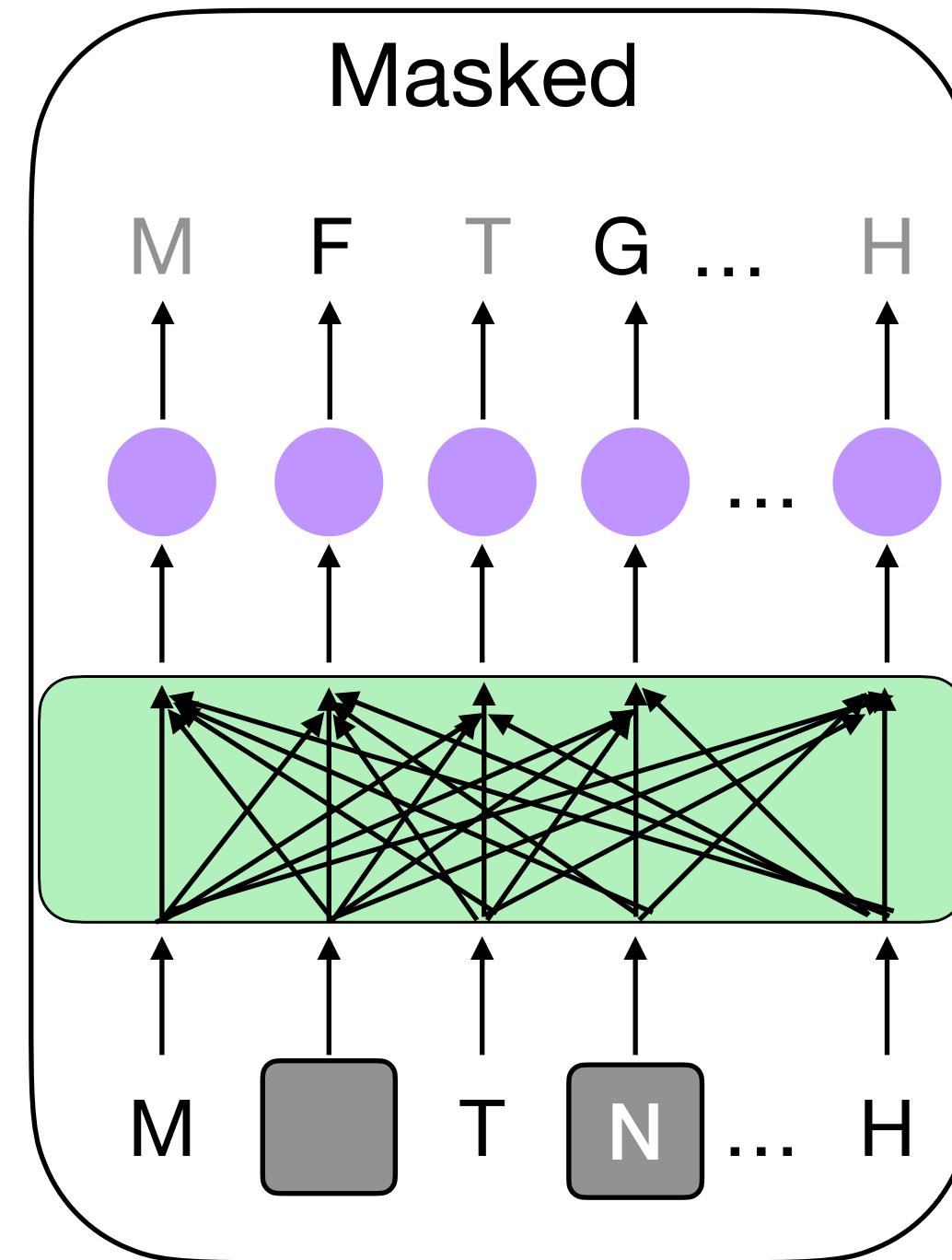
Separate pretraining task and architecture



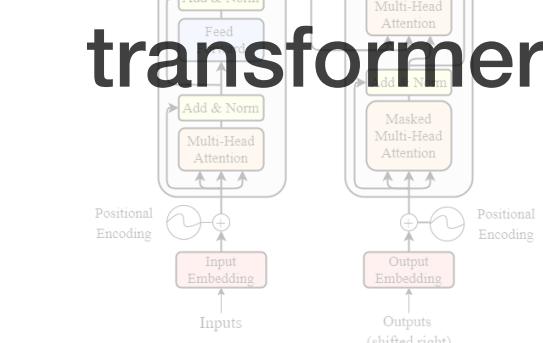
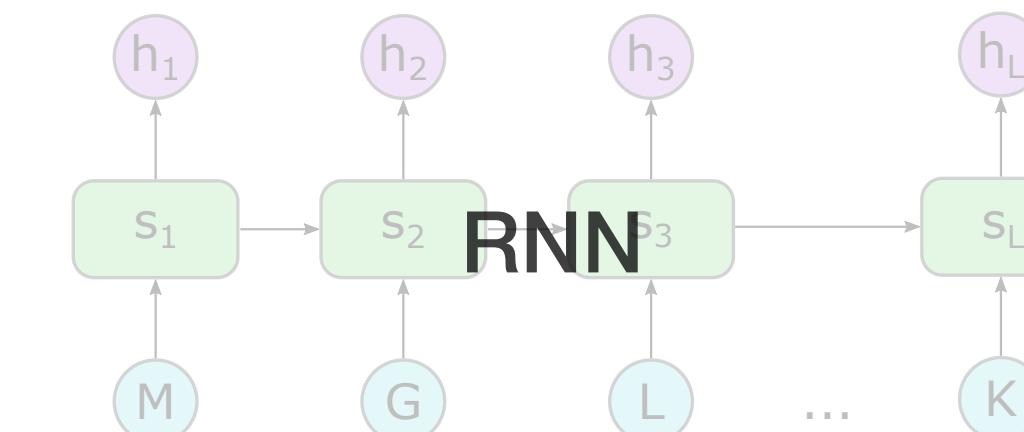
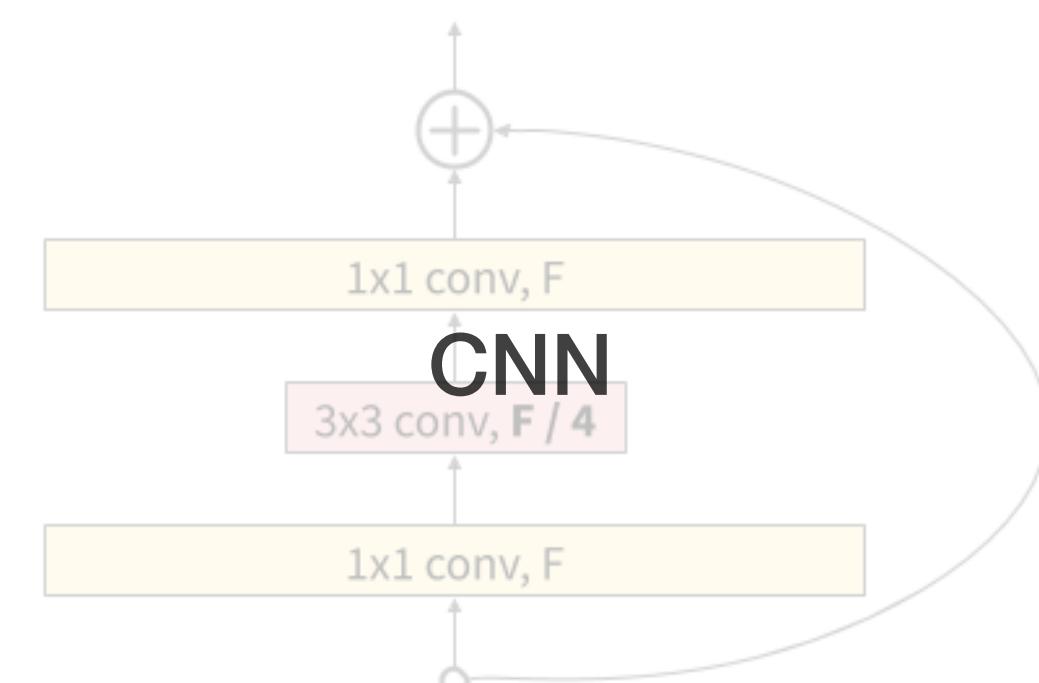
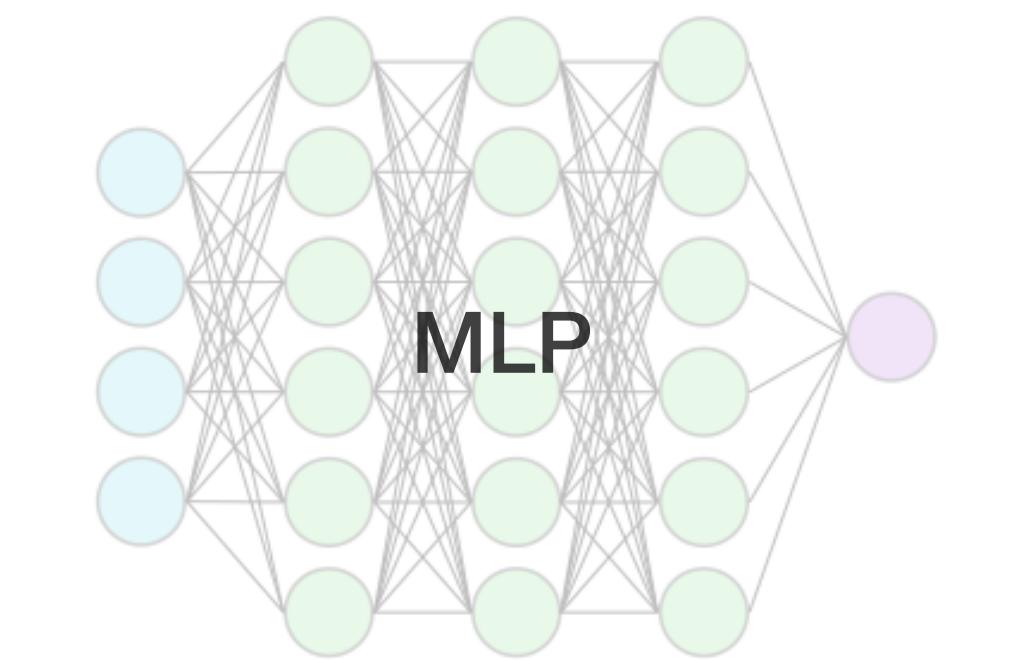
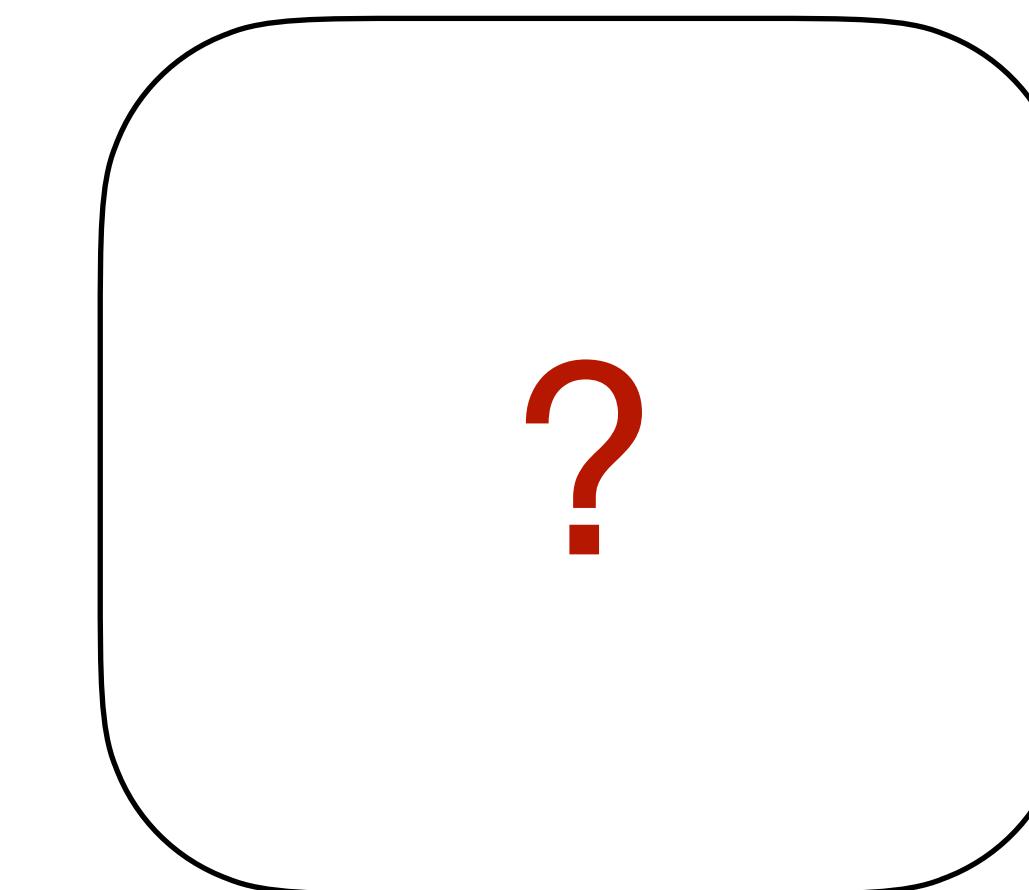
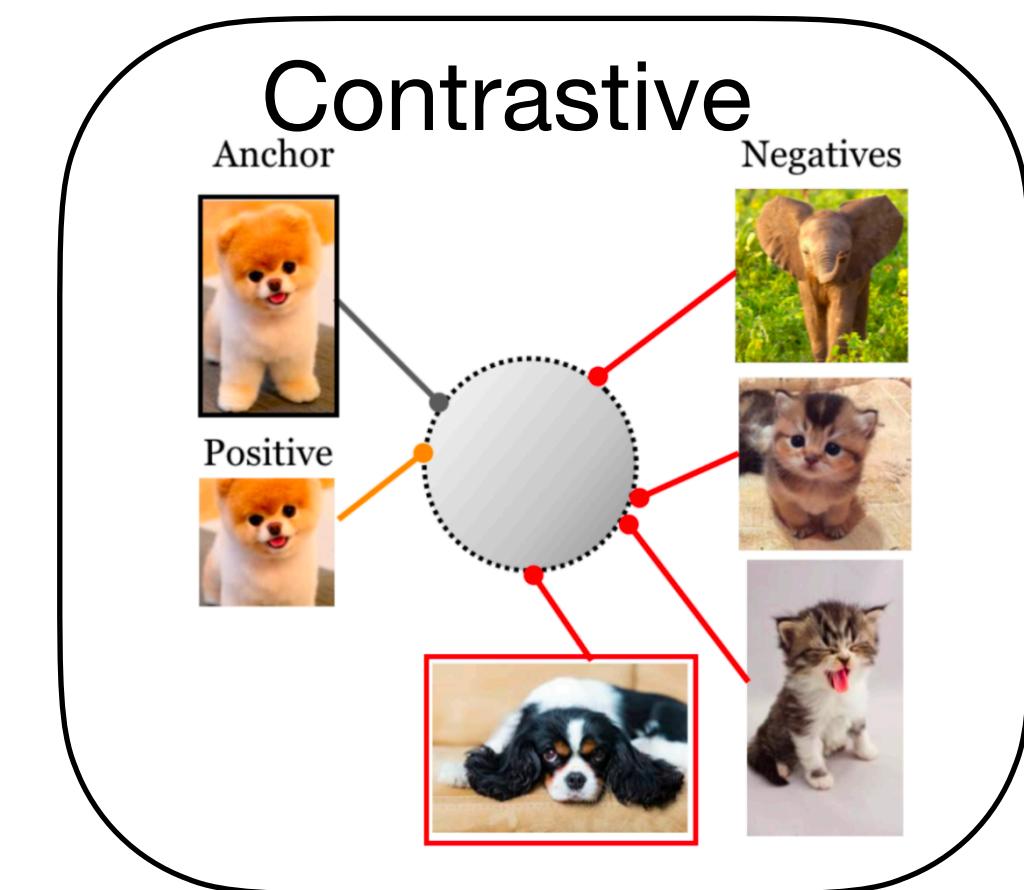
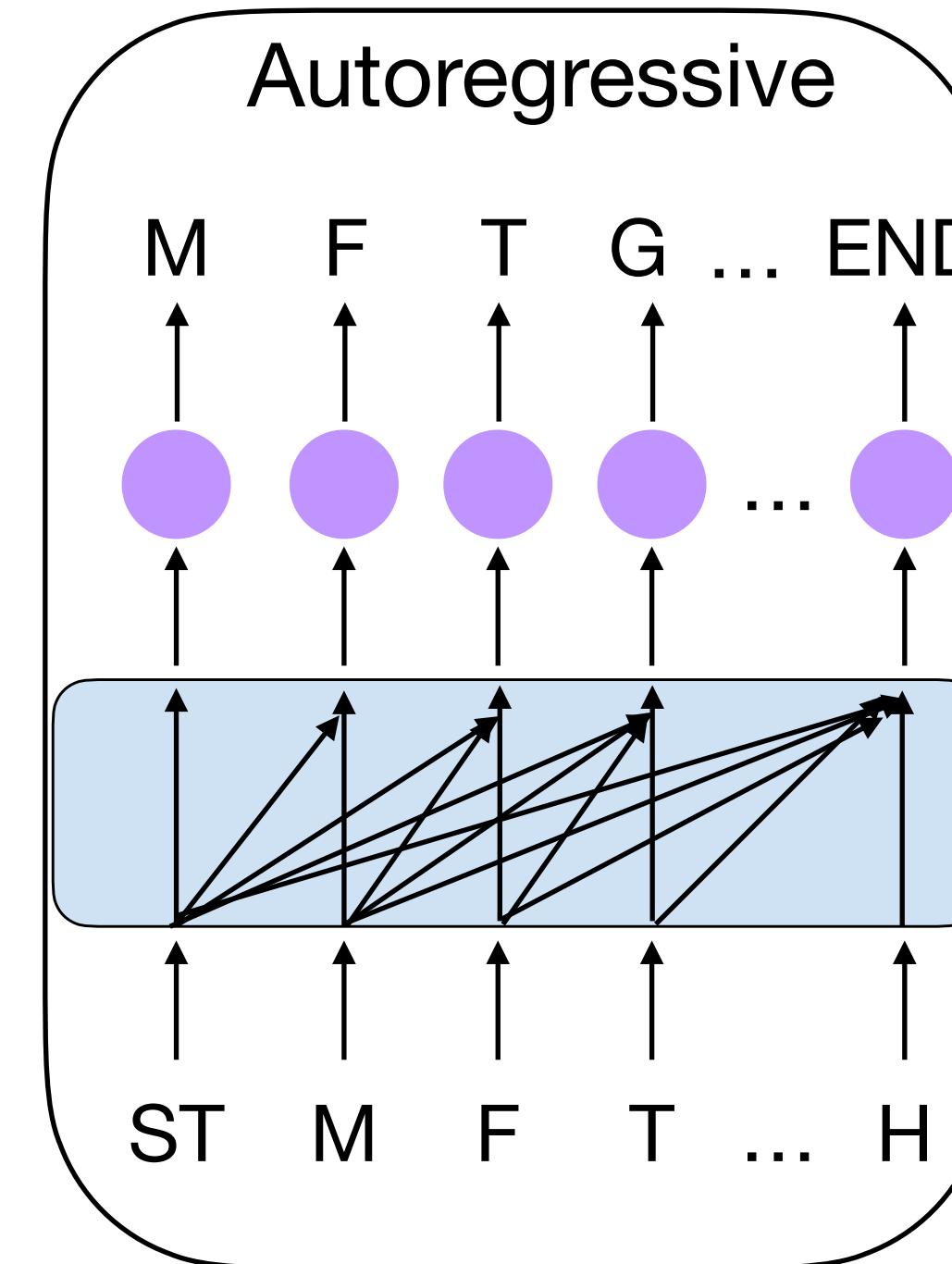
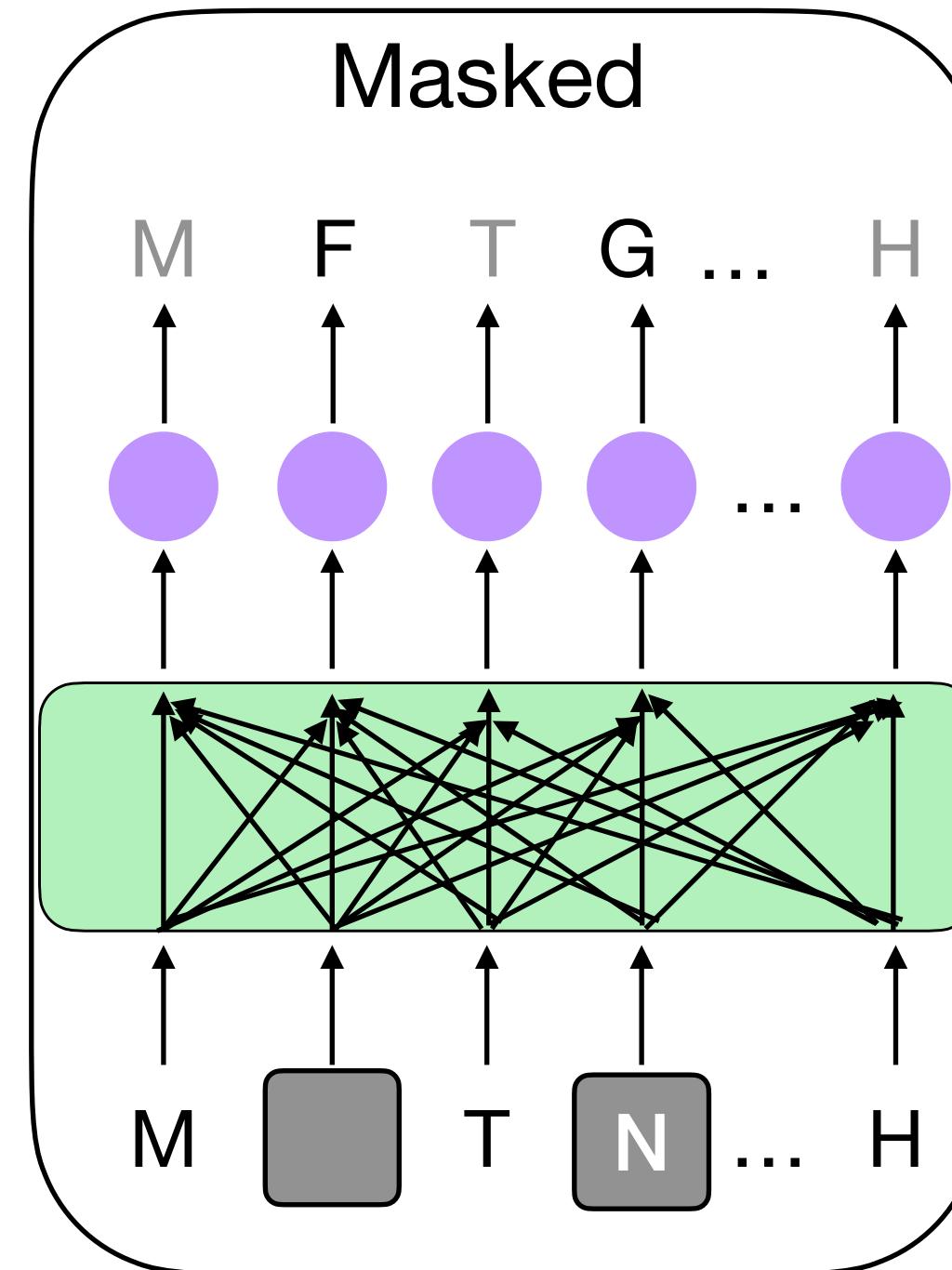
Separate pretraining task and architecture



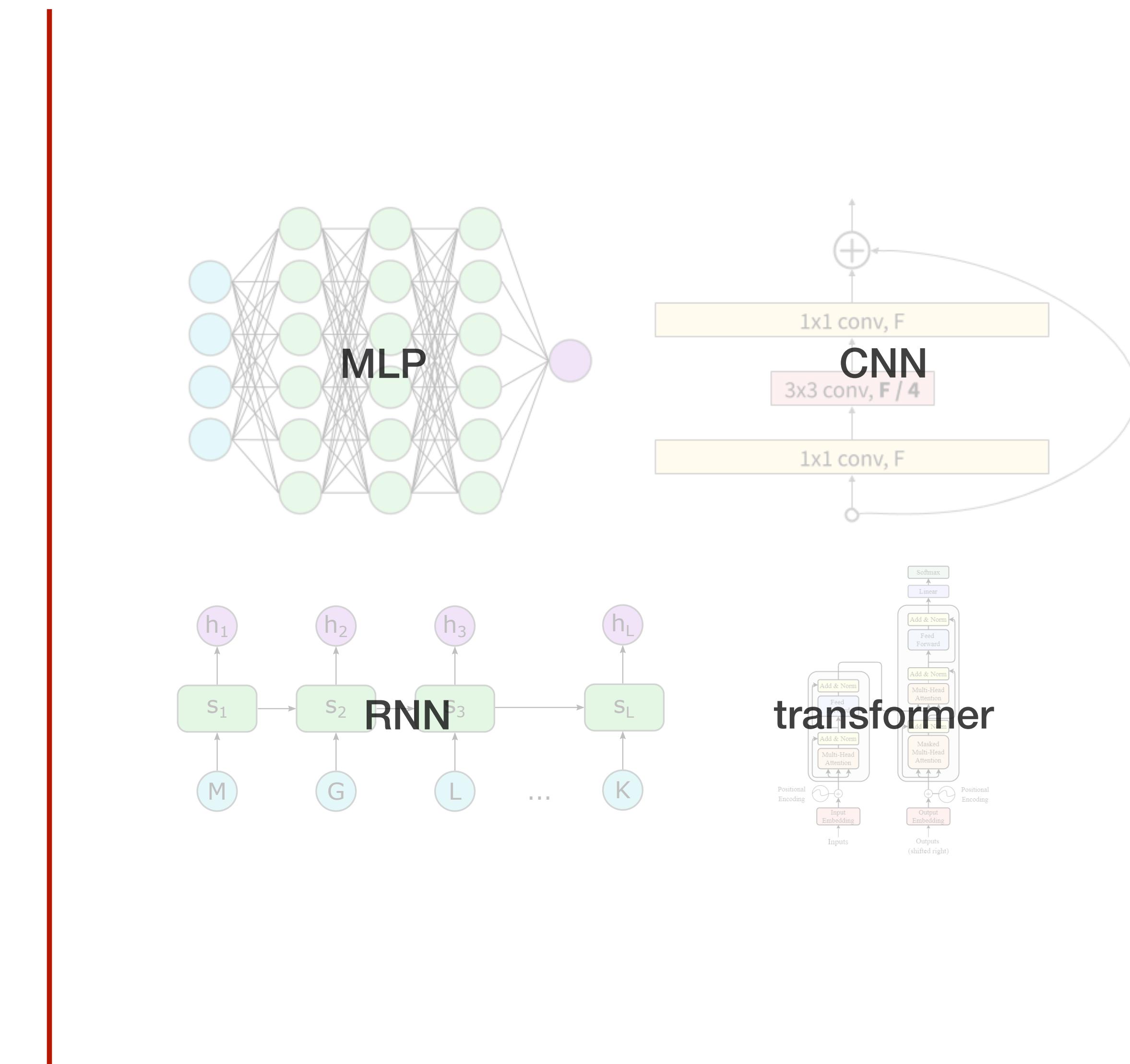
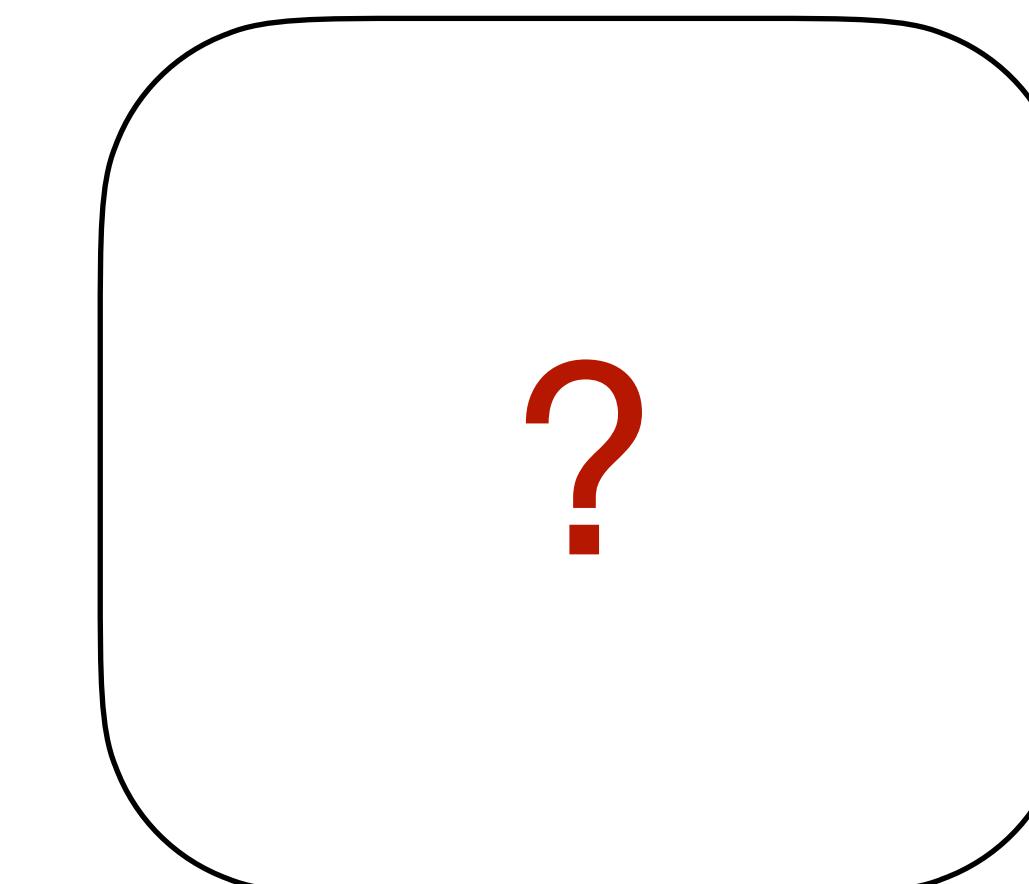
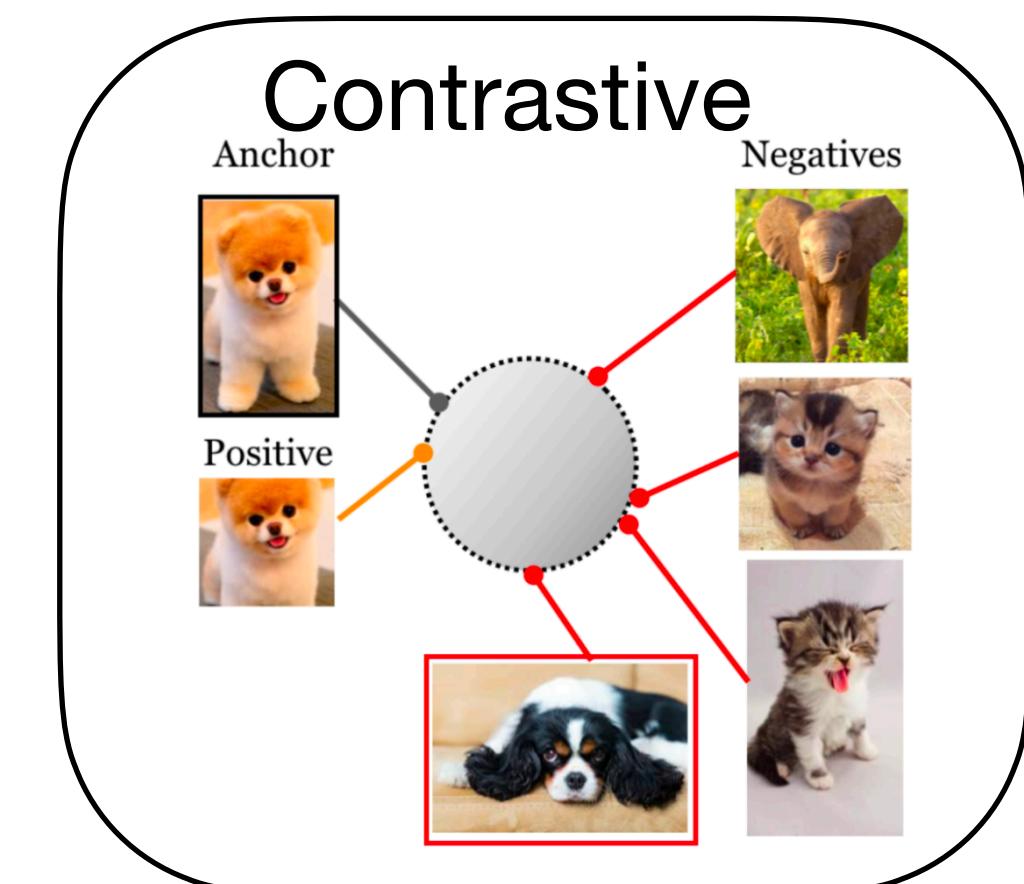
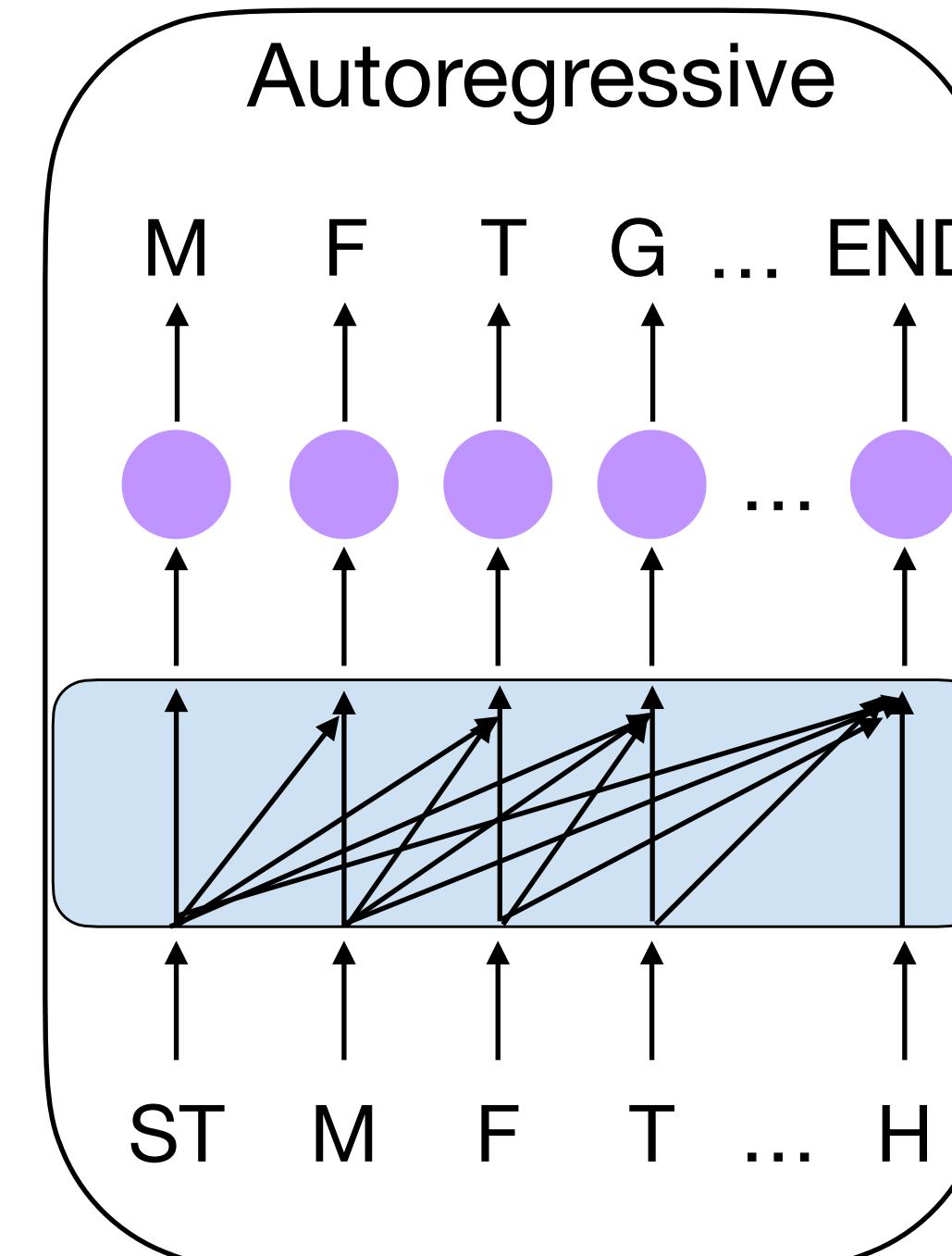
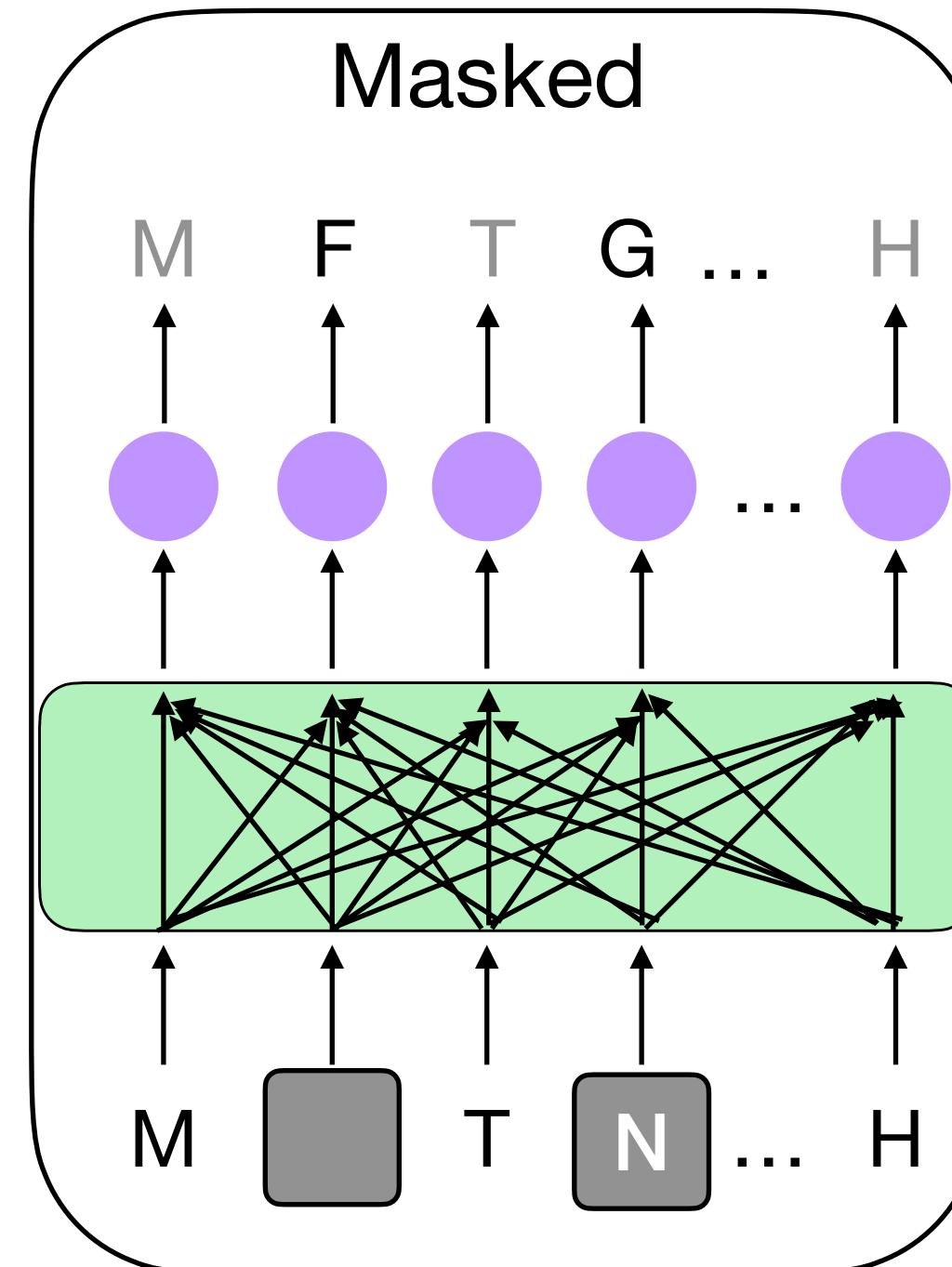
Separate pretraining task and architecture



Separate pretraining task and architecture



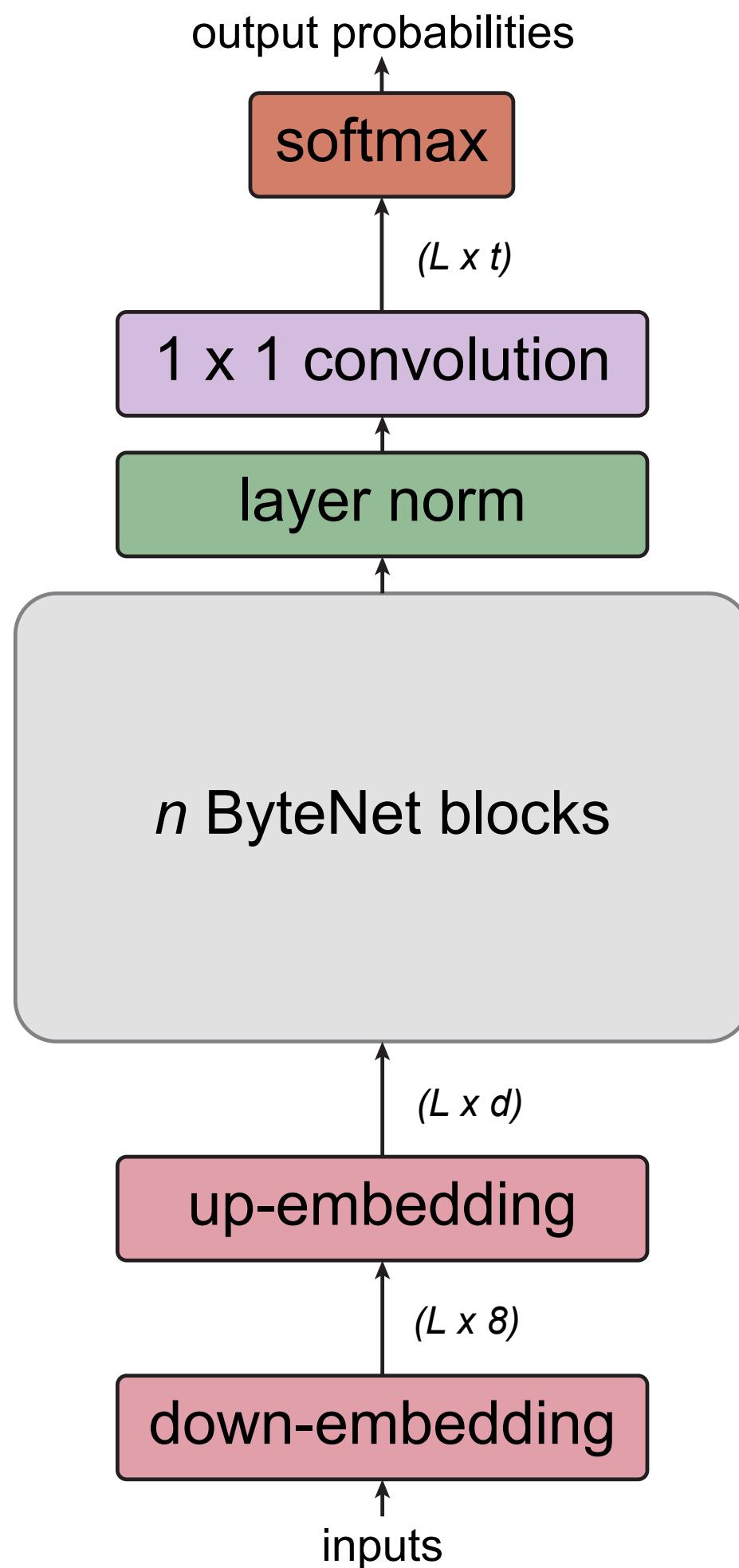
Separate pretraining task and architecture



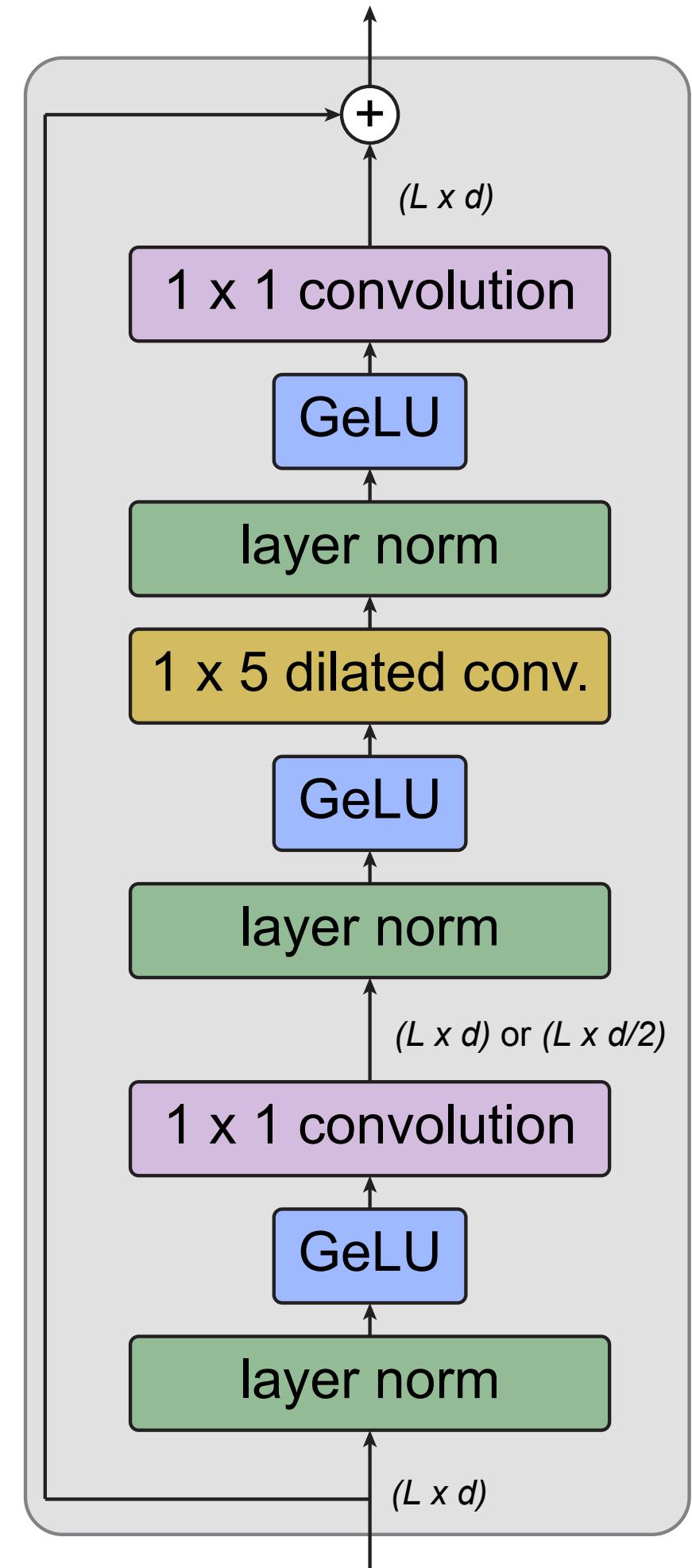
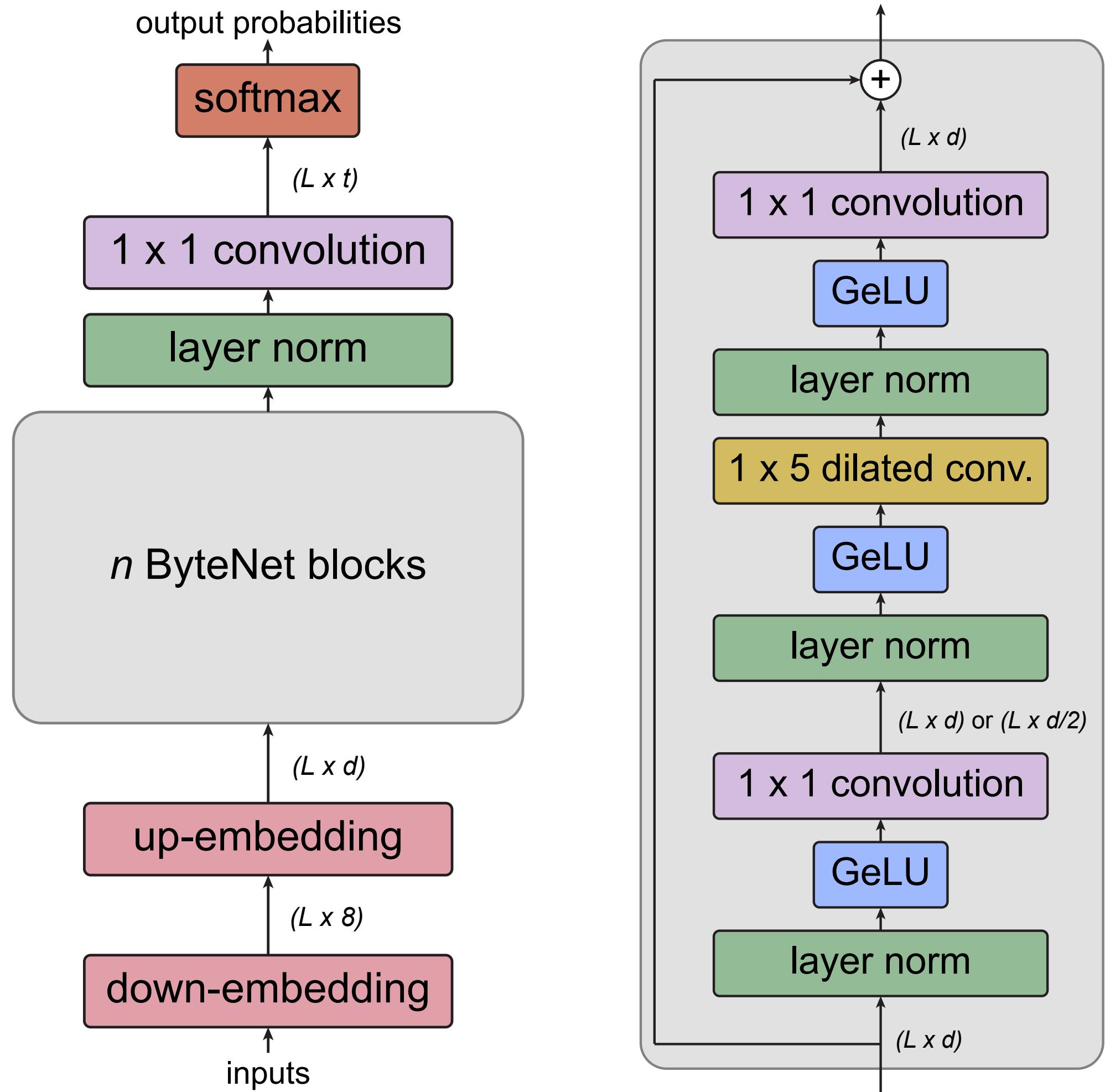
We pretrain CNNs to reconstruct sequences



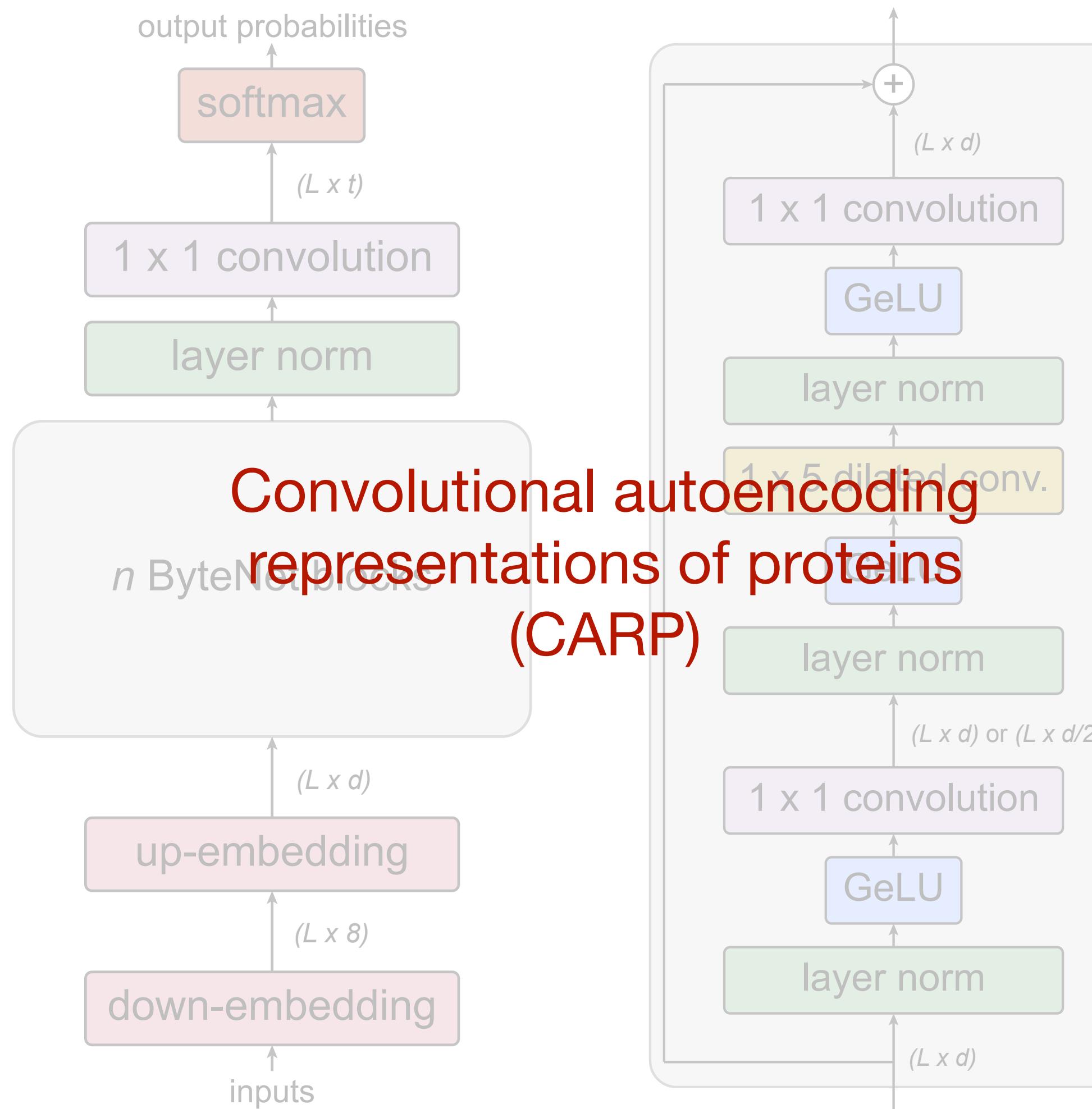
We pretrain CNNs to reconstruct sequences



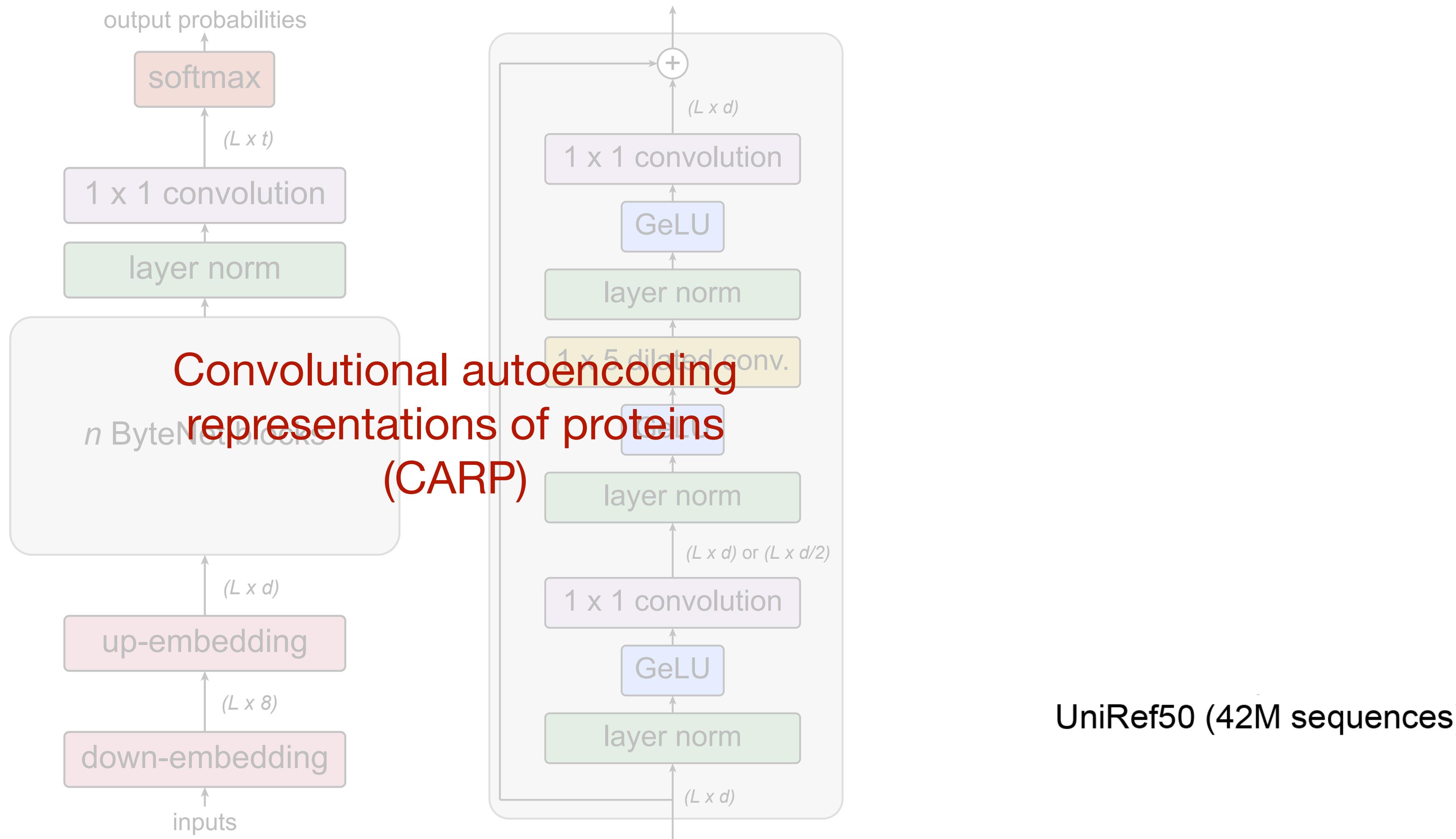
We pretrain CNNs to reconstruct sequences



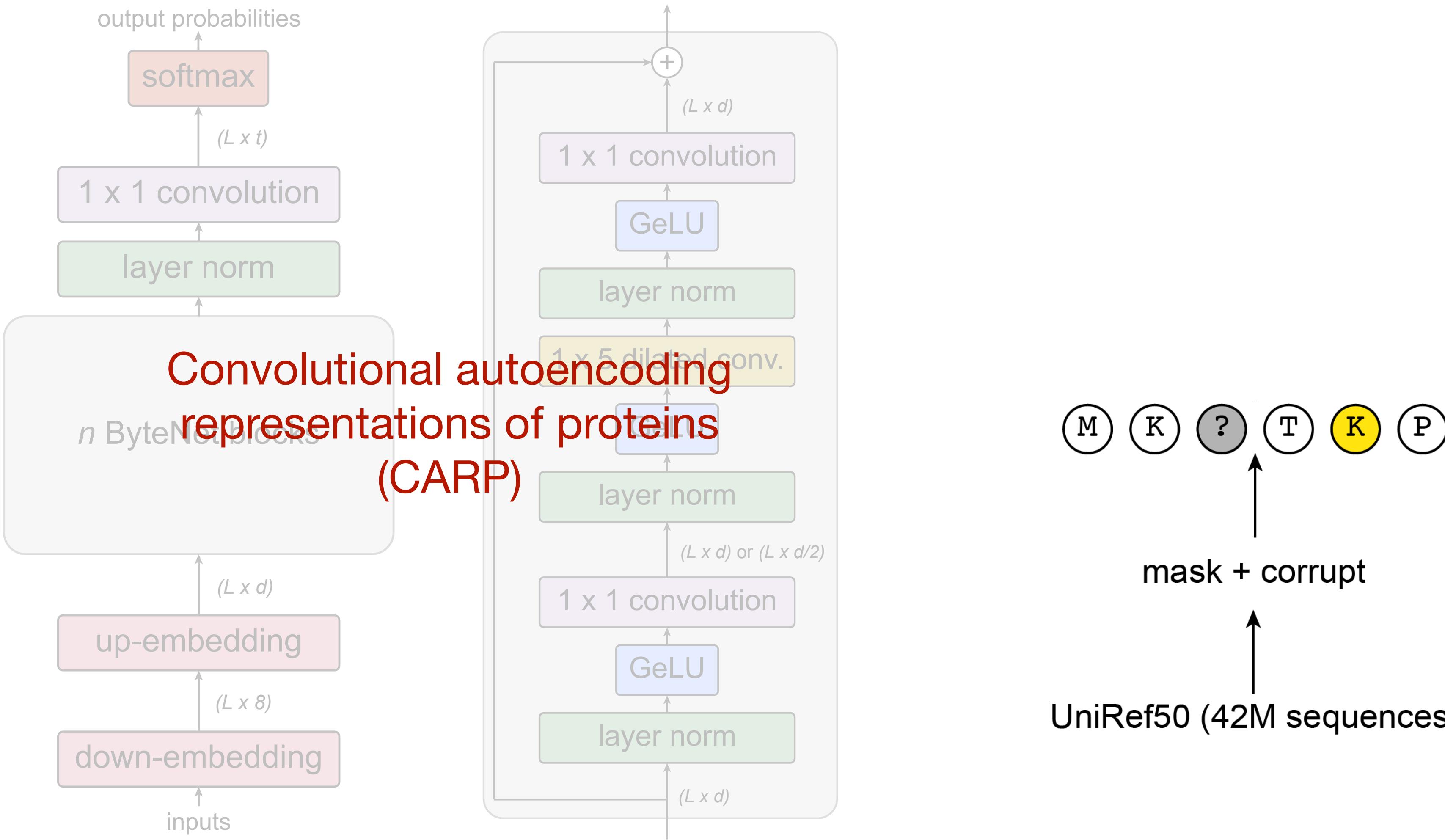
We pretrain CNNs to reconstruct sequences



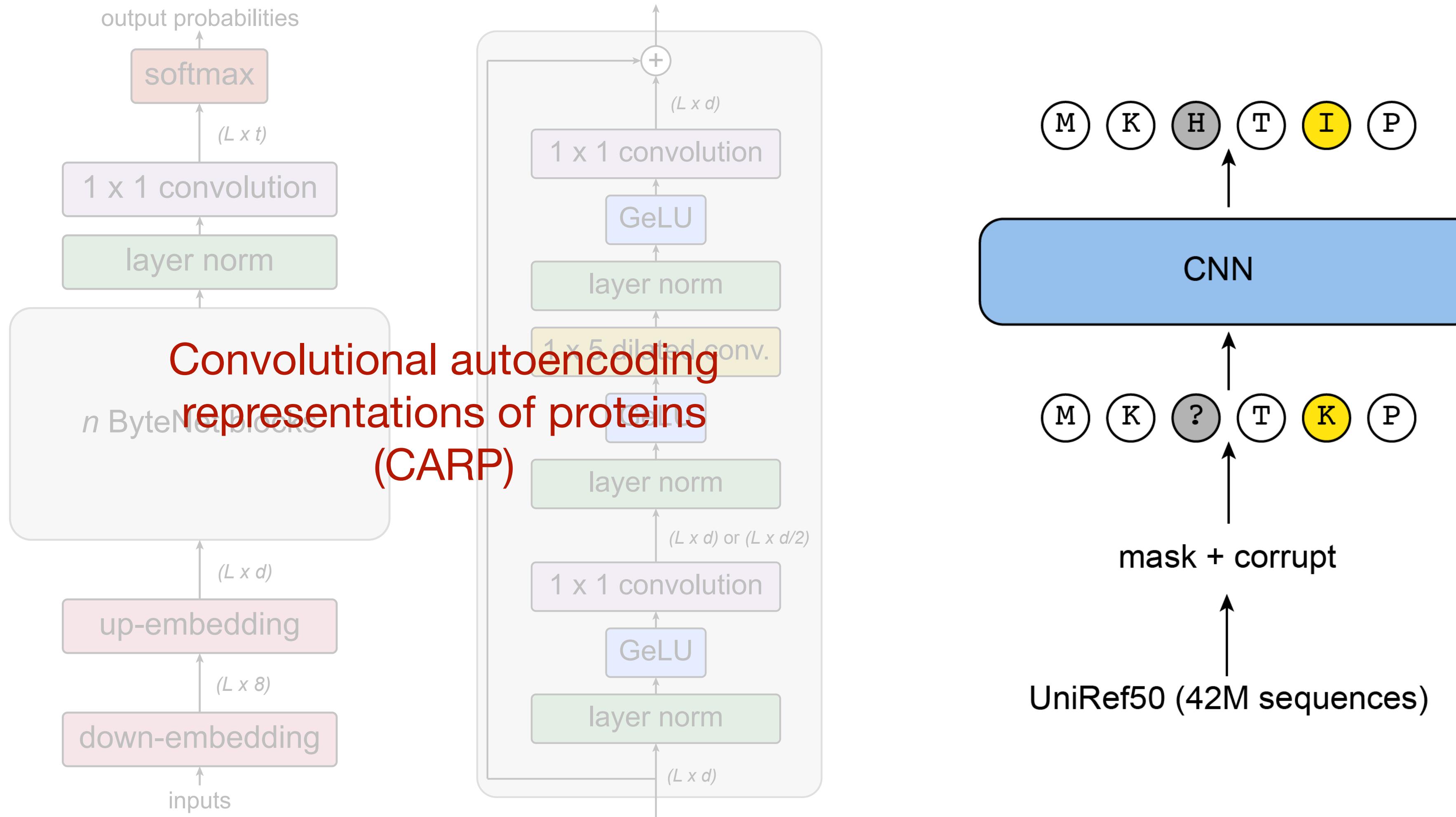
We pretrain CNNs to reconstruct sequences



We pretrain CNNs to reconstruct sequences

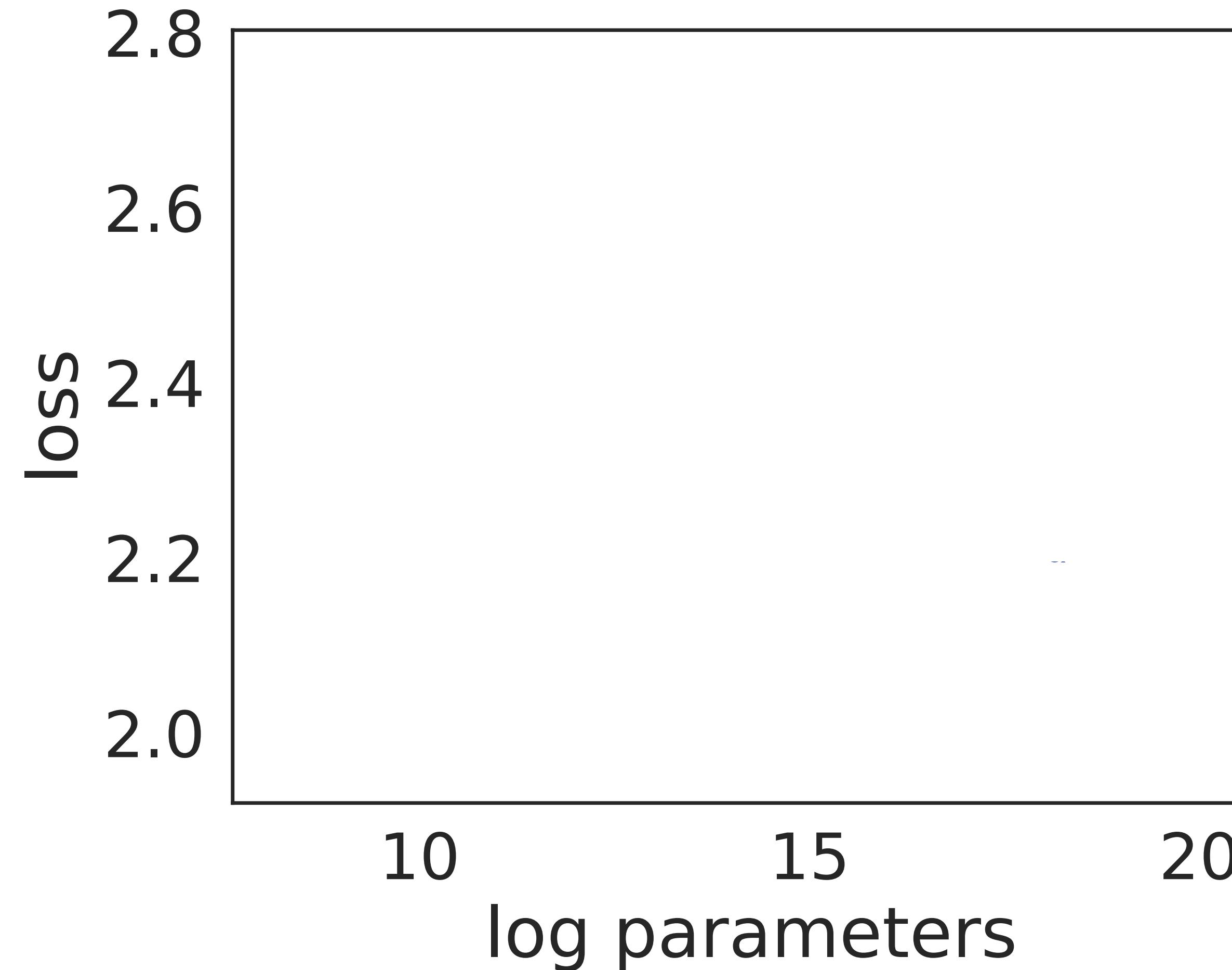


We pretrain CNNs to reconstruct sequences

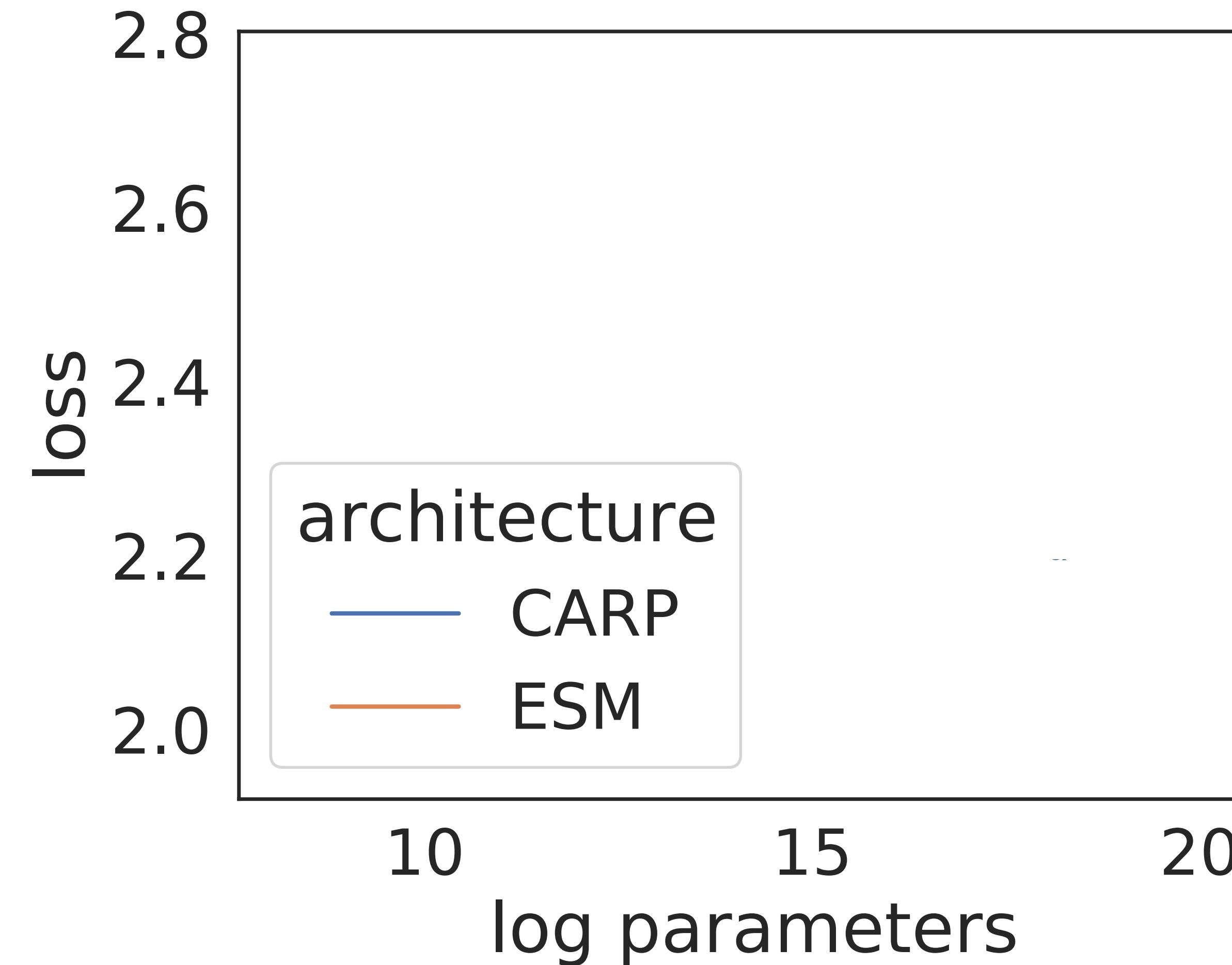


CNNs are competitive with transformers for pretraining

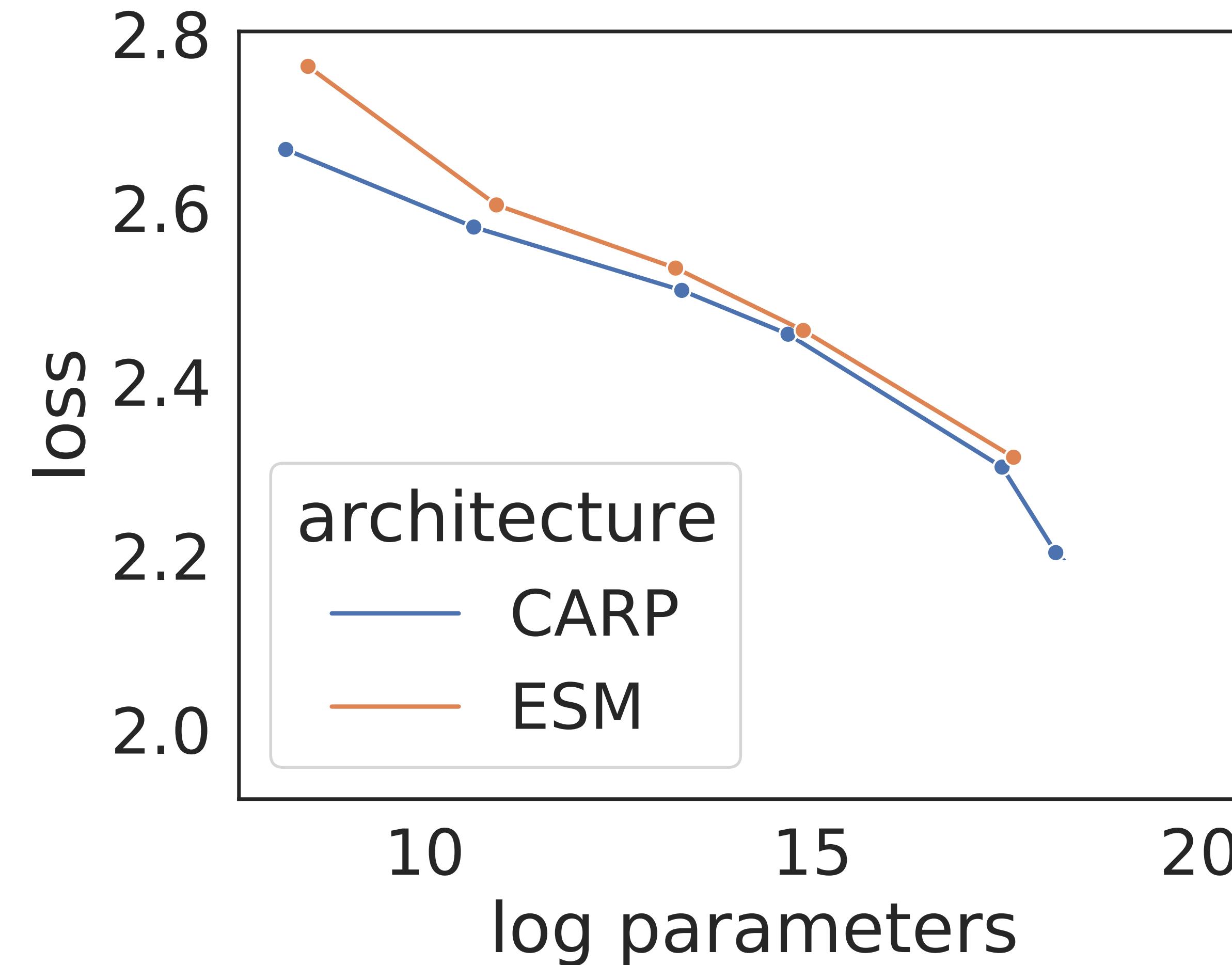
CNNs are competitive with transformers for pretraining



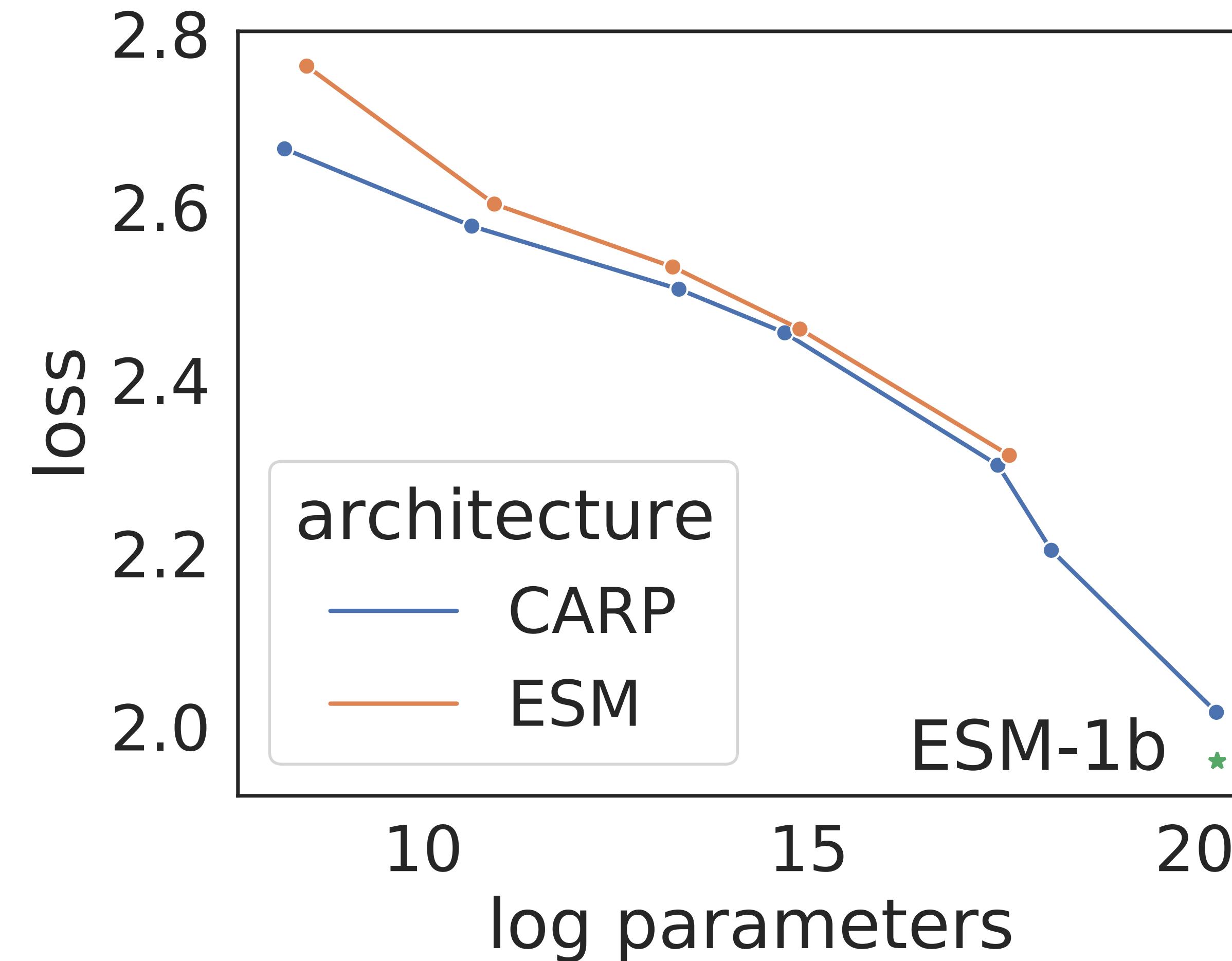
CNNs are competitive with transformers for pretraining



CNNs are competitive with transformers for pretraining

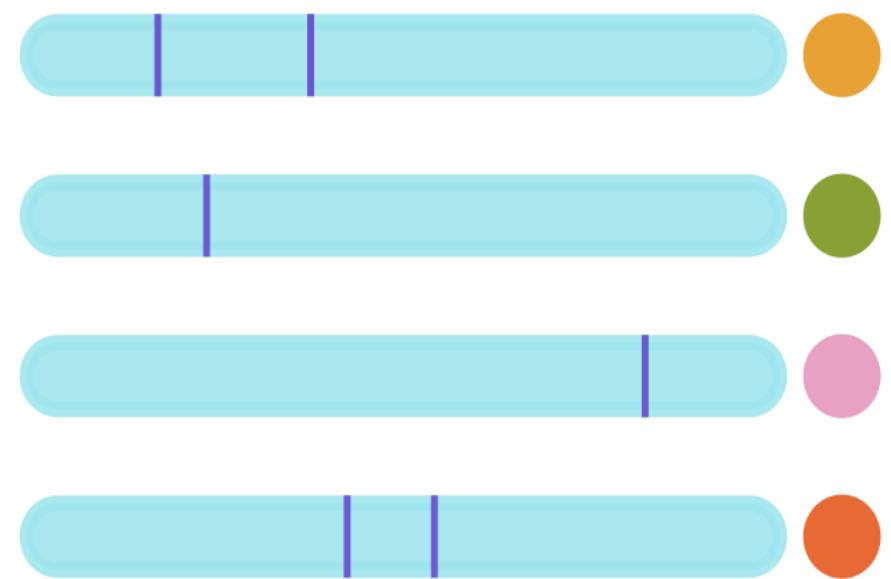


CNNs are competitive with transformers for pretraining

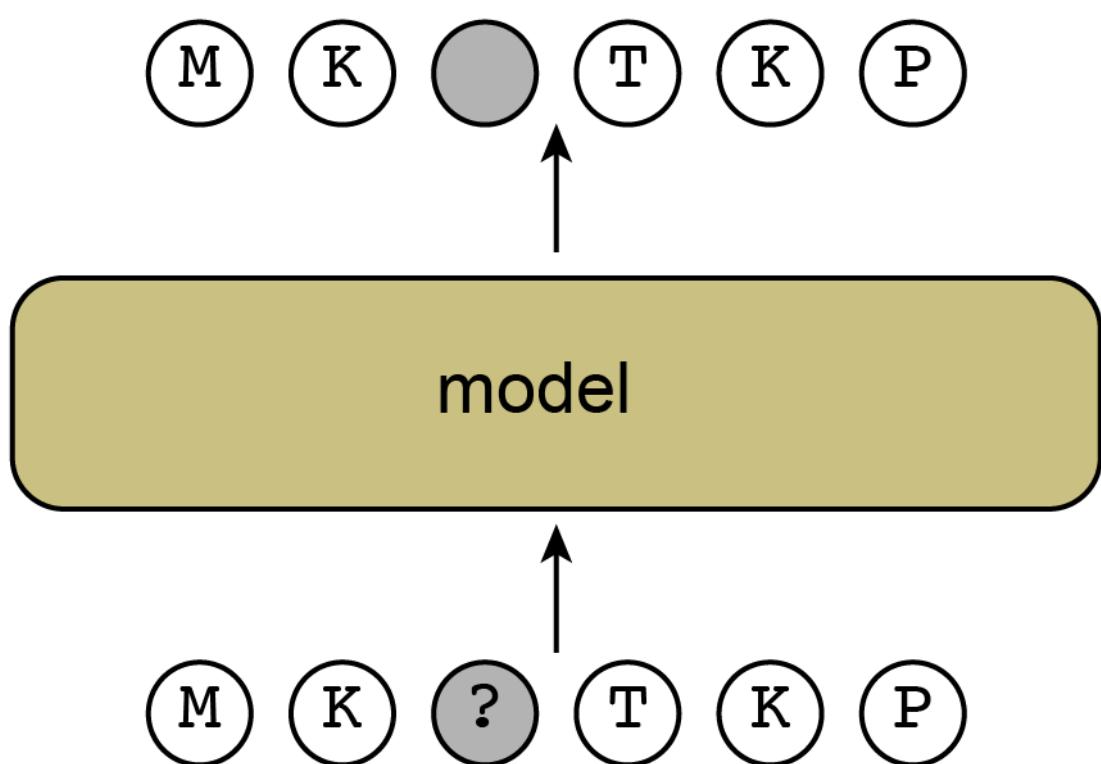


CARP is a zero-shot fitness predictor

Deep mutational scan

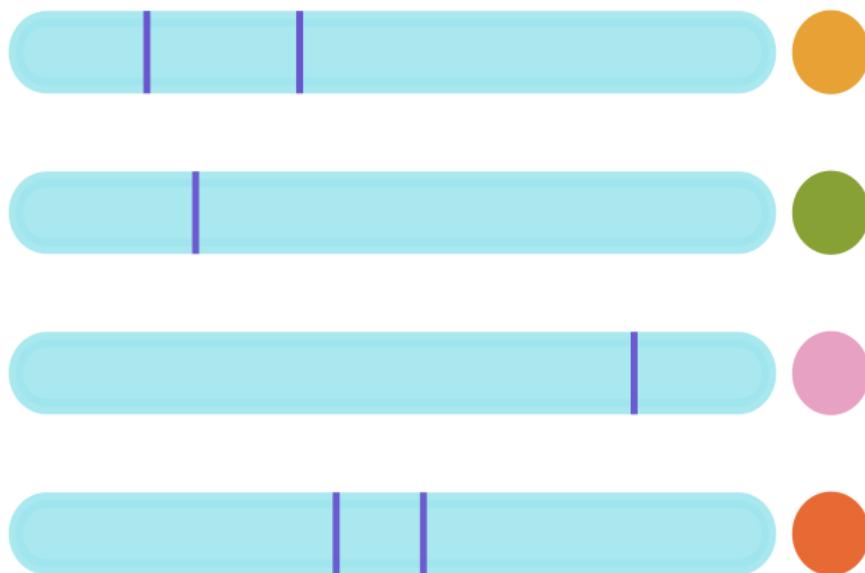


compare likelihood of
mutation to WT



CARP is a zero-shot fitness predictor

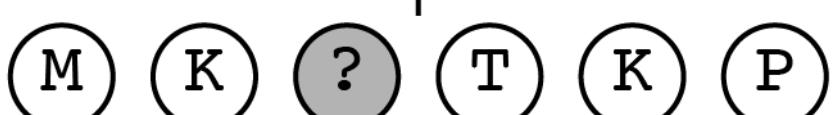
Deep mutational scan



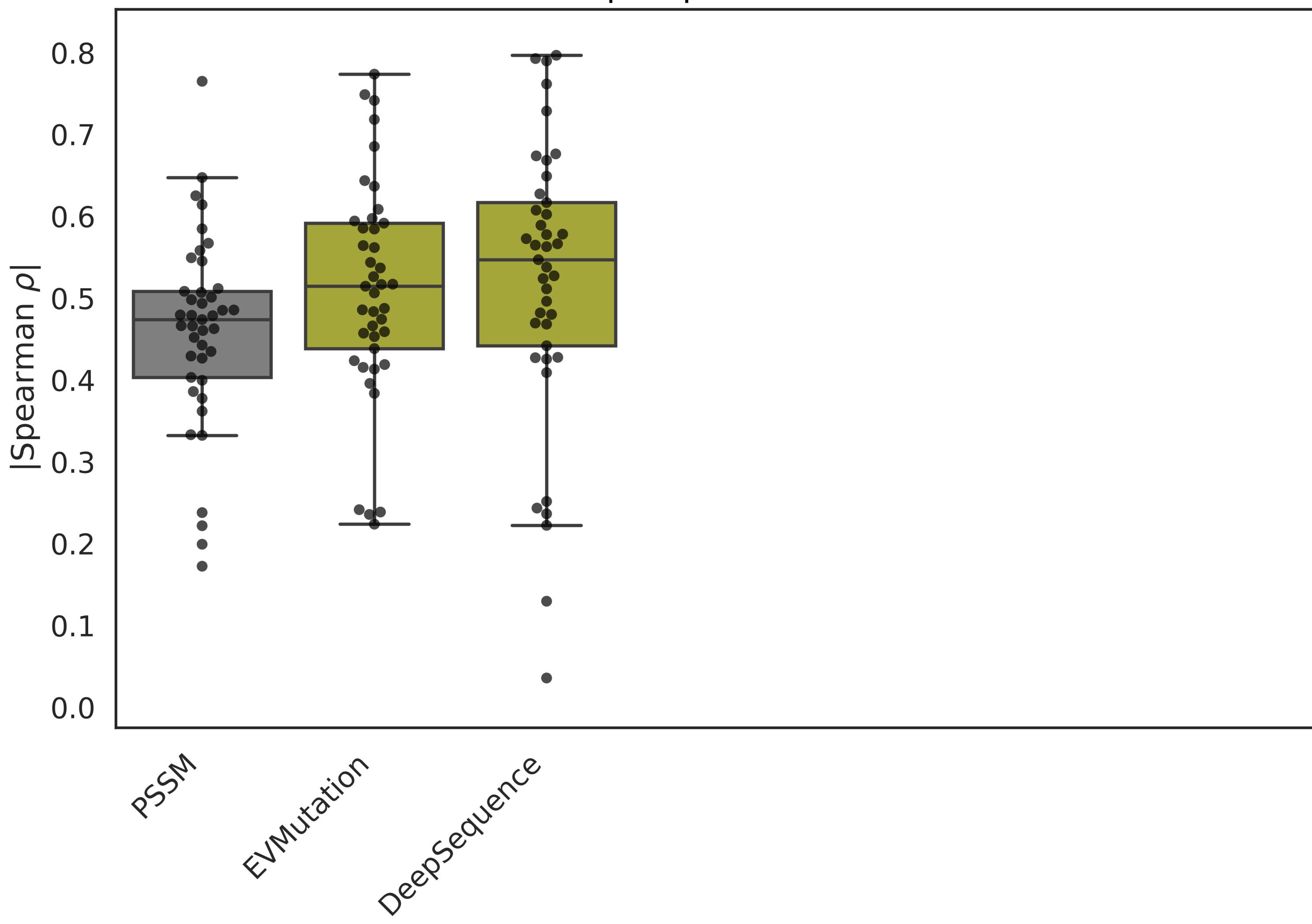
compare likelihood of
mutation to WT



model

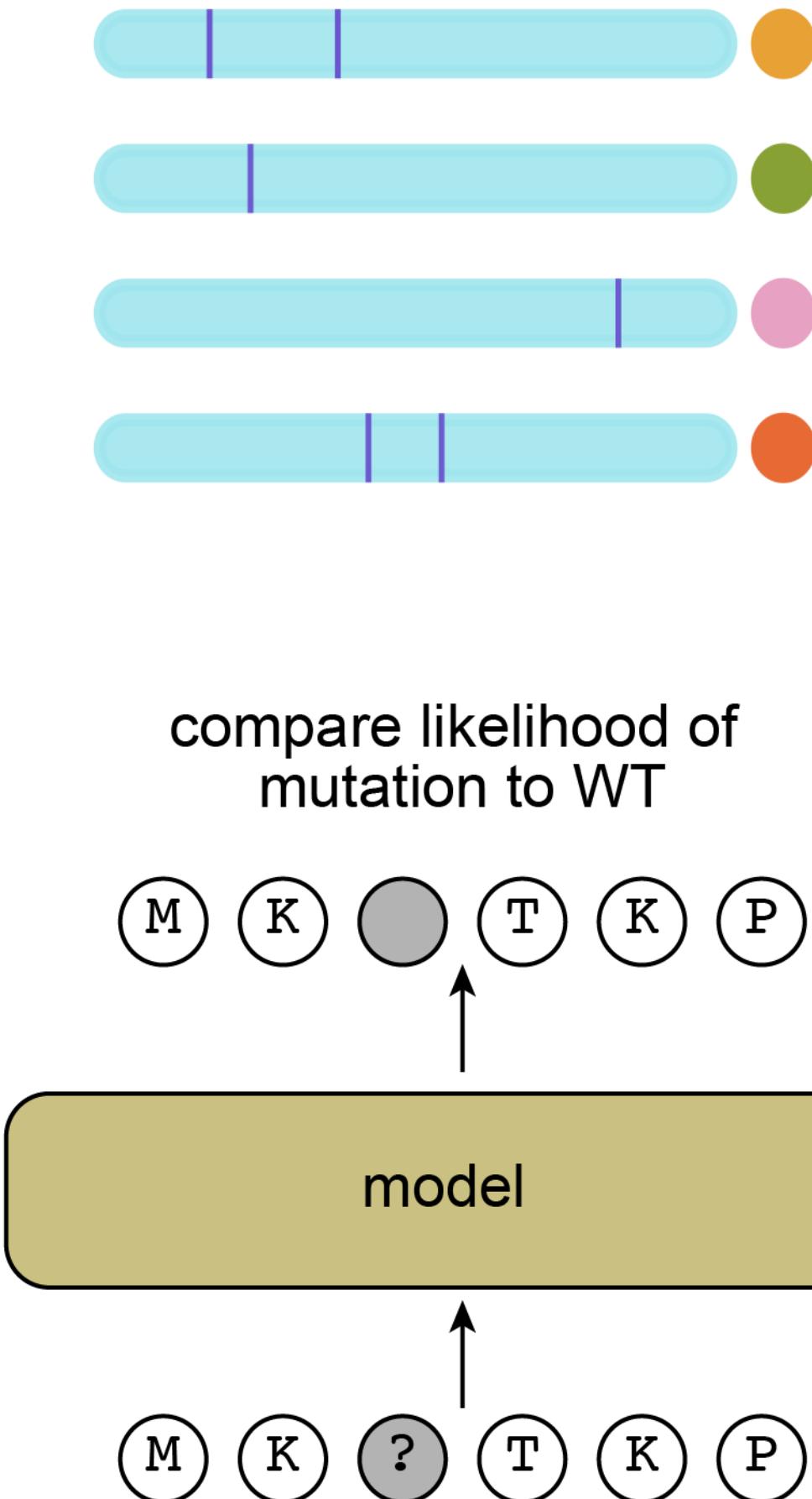


DeepSequence

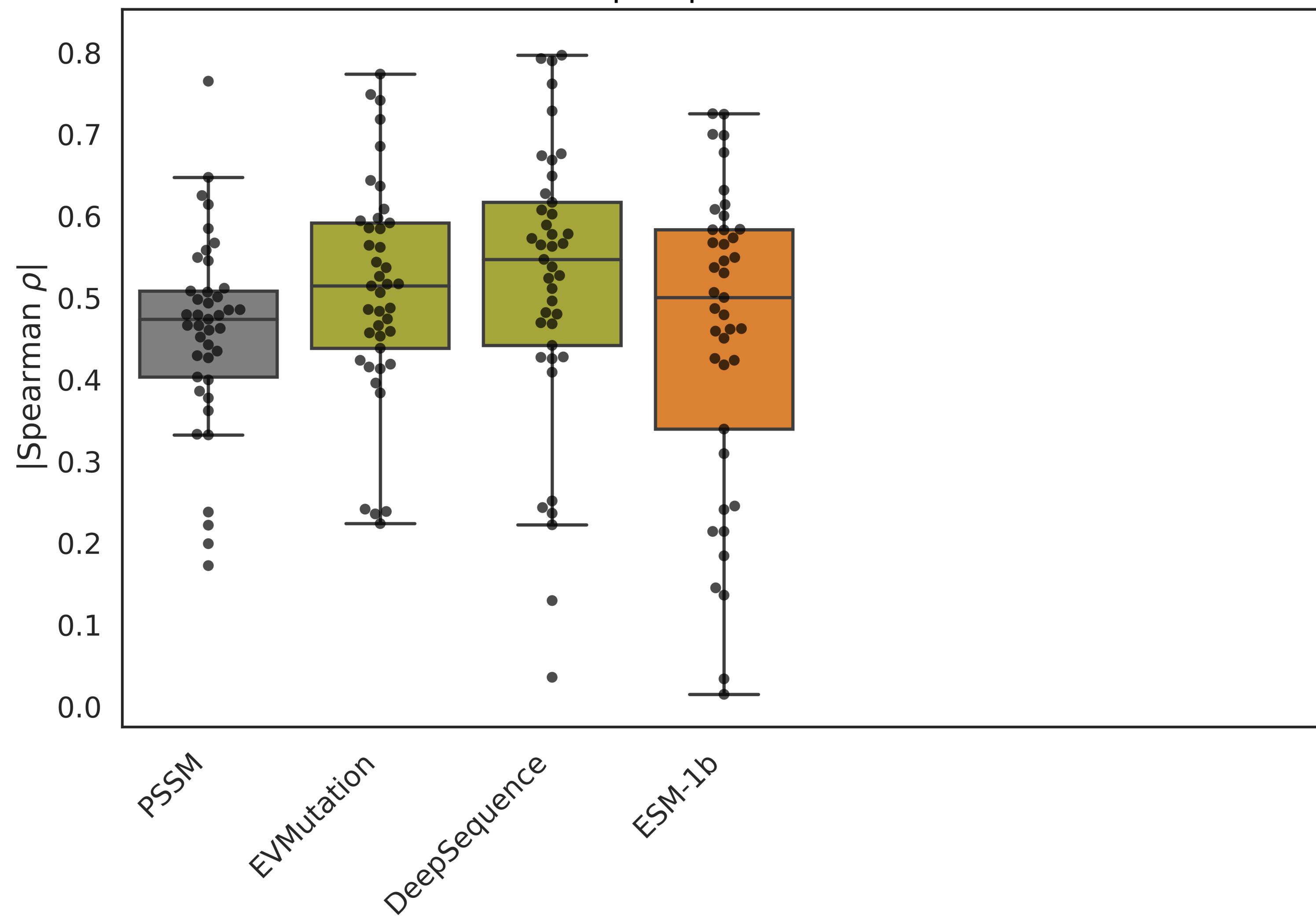


CARP is a zero-shot fitness predictor

Deep mutational scan

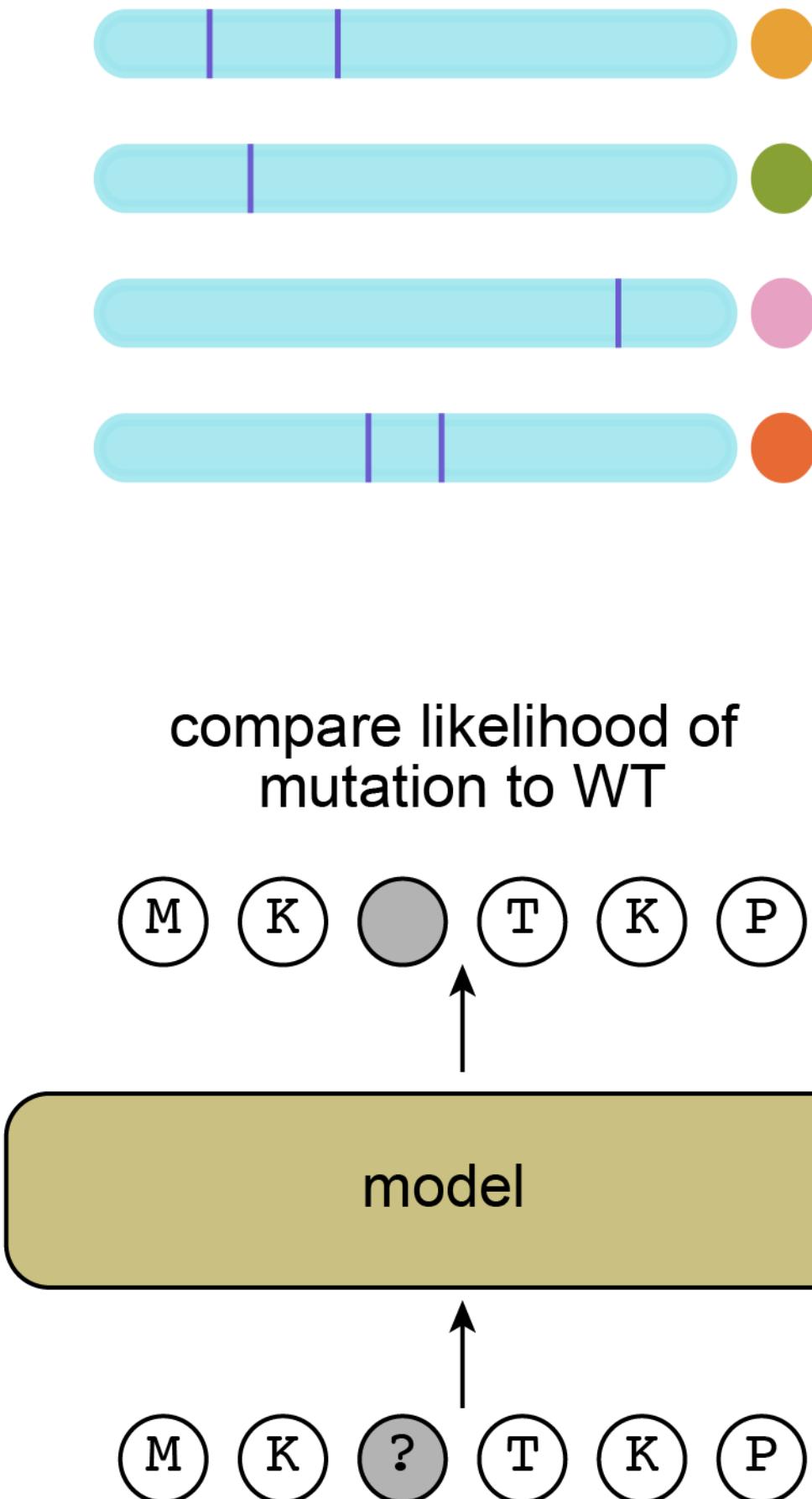


DeepSequence

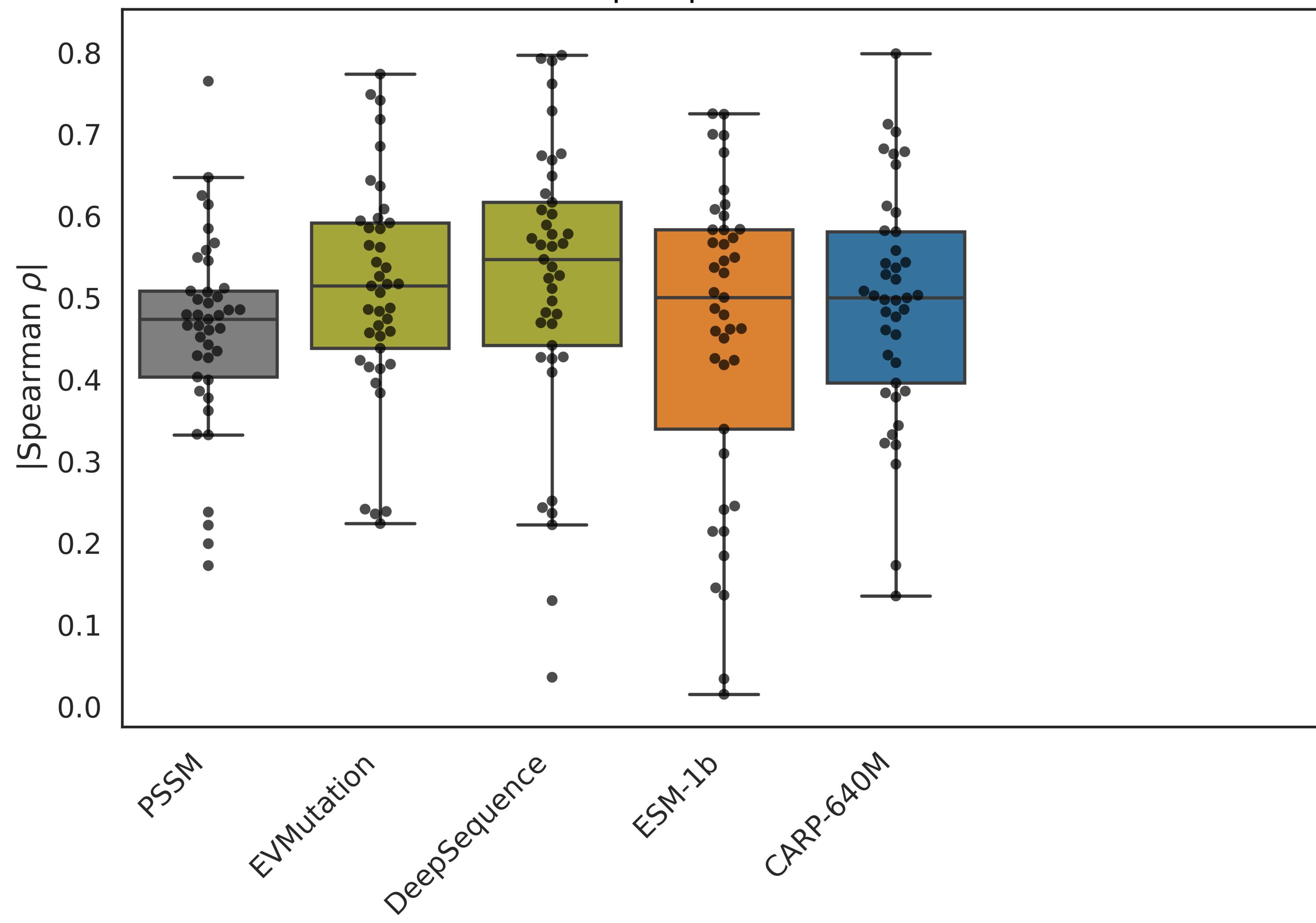


CARP is a zero-shot fitness predictor

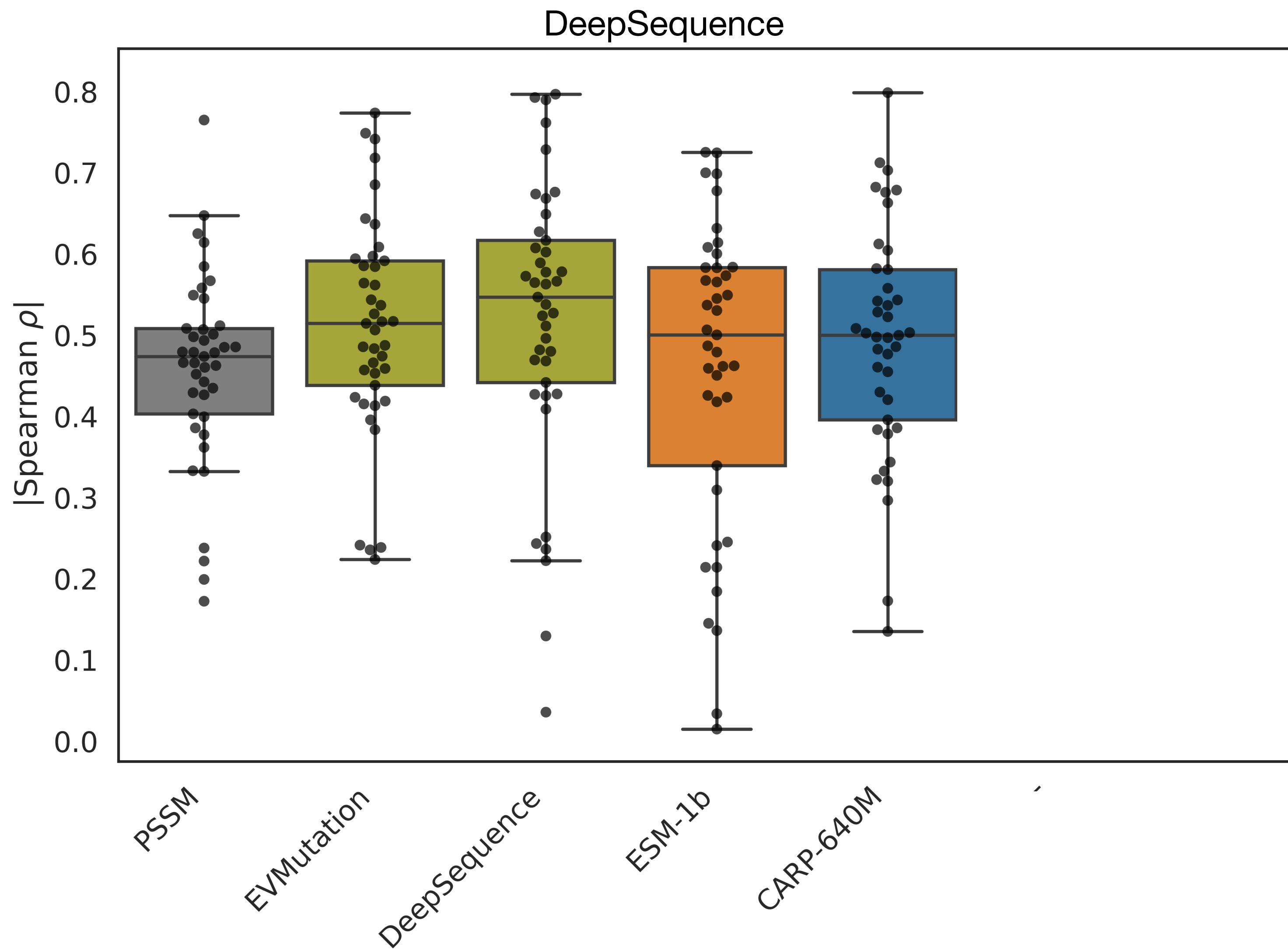
Deep mutational scan



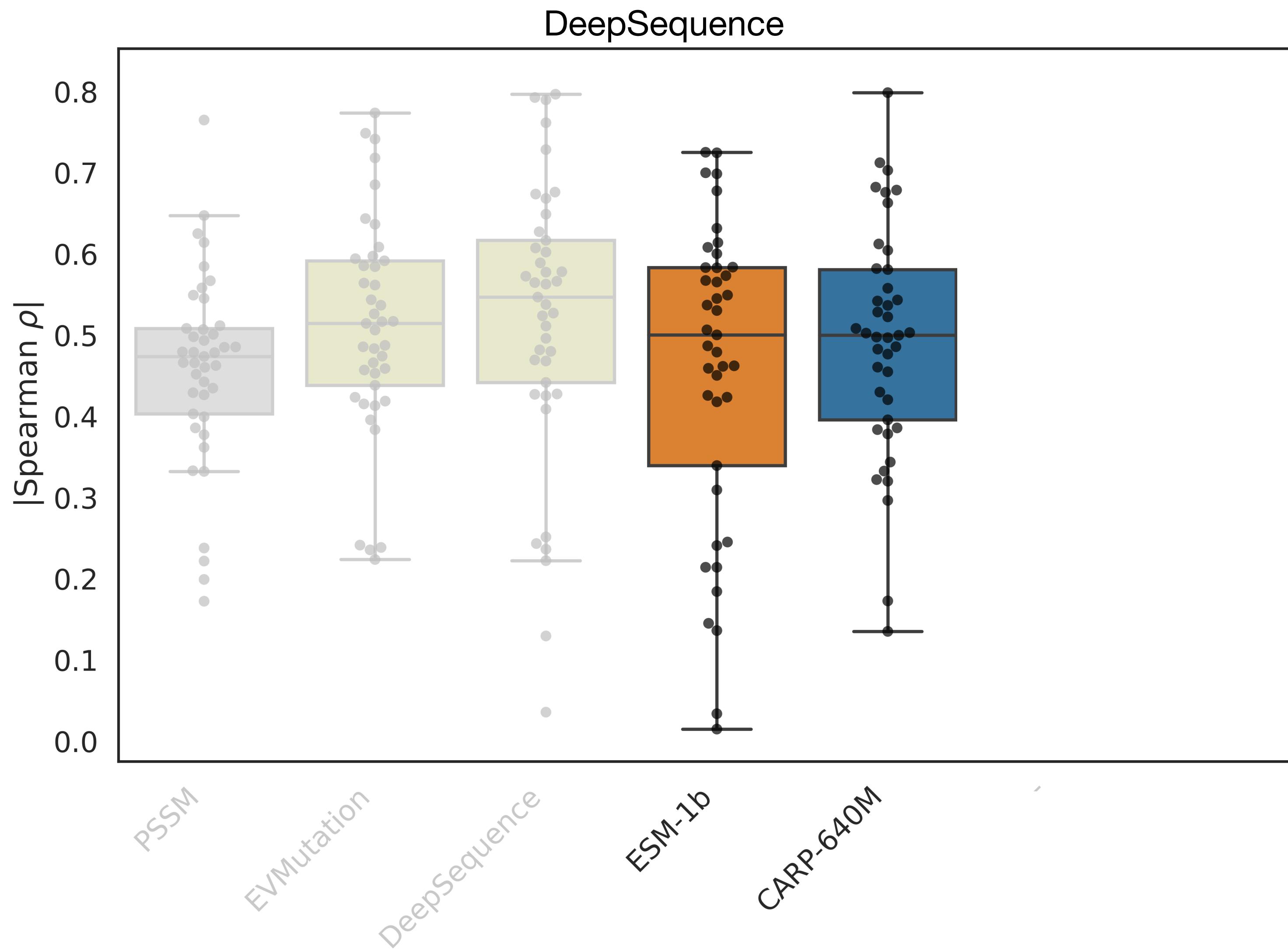
DeepSequence



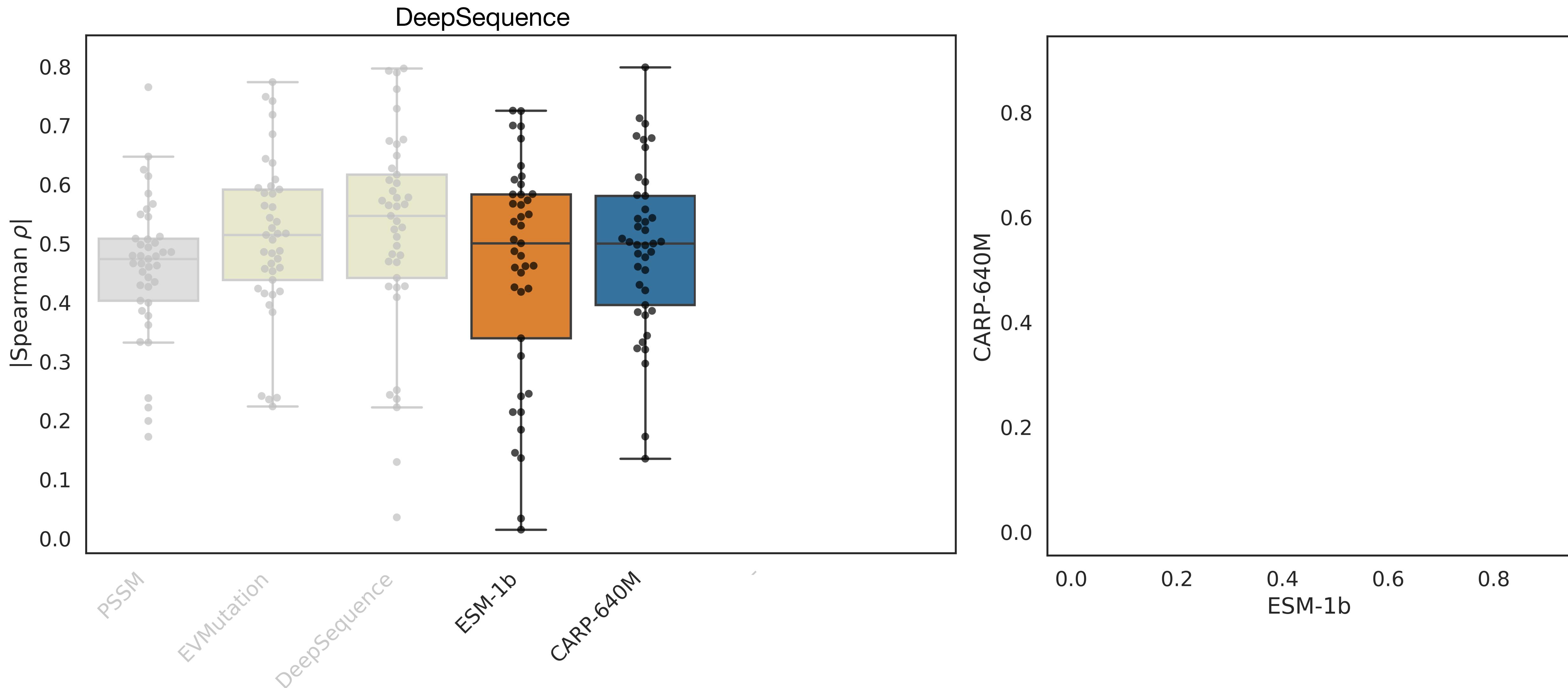
CARP is a zero-shot fitness predictor



CARP is a zero-shot fitness predictor

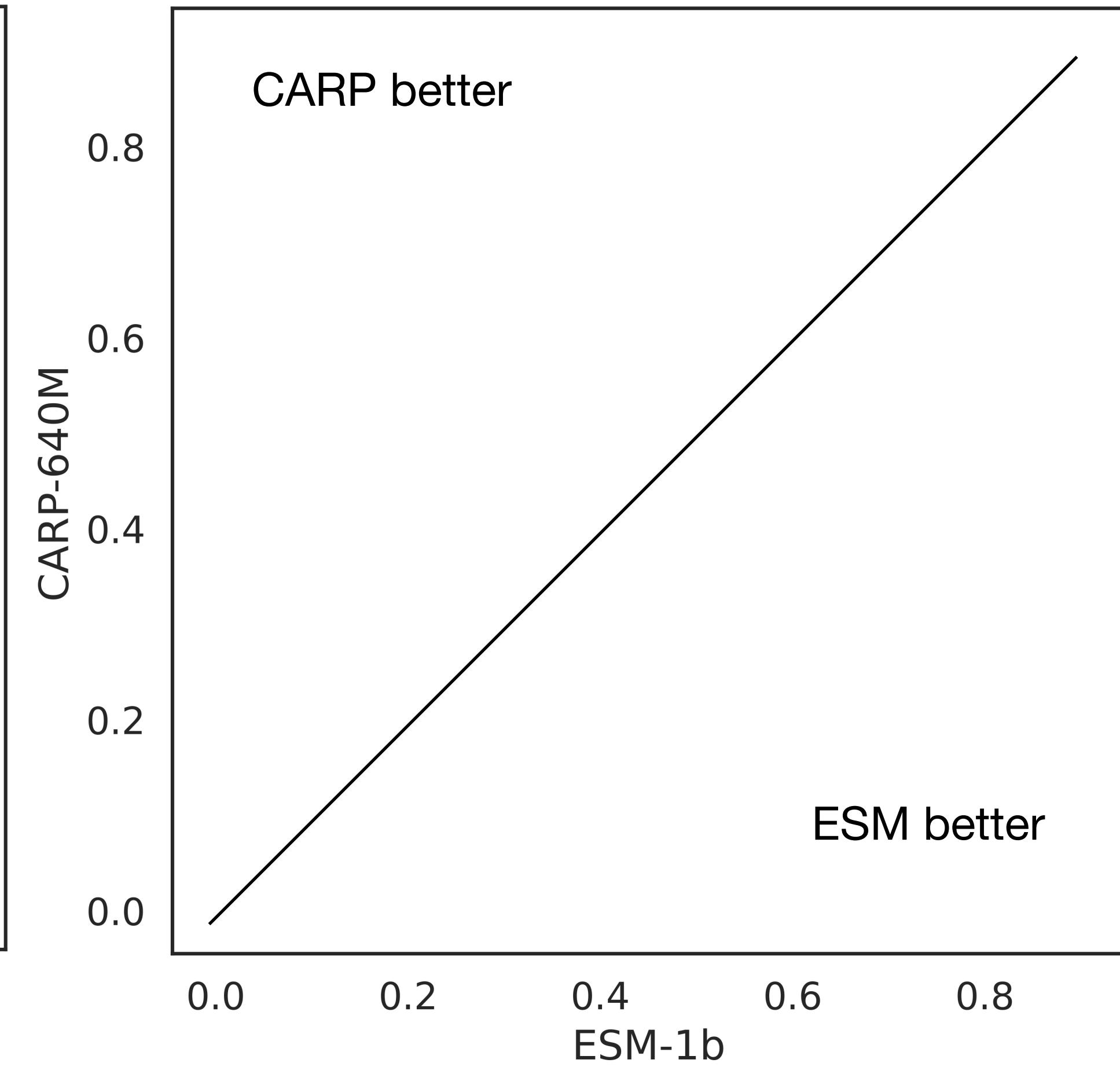
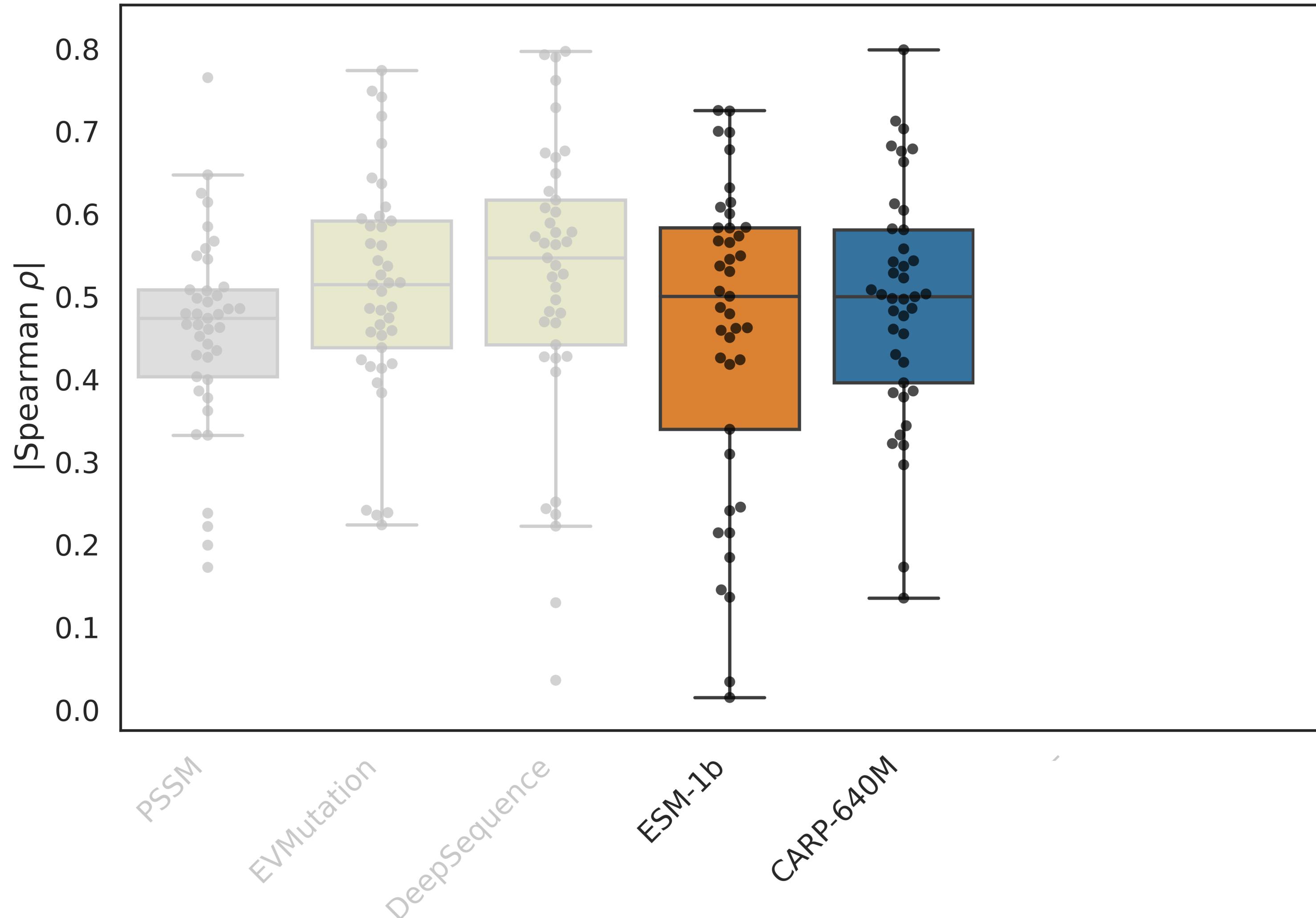


CARP is a zero-shot fitness predictor



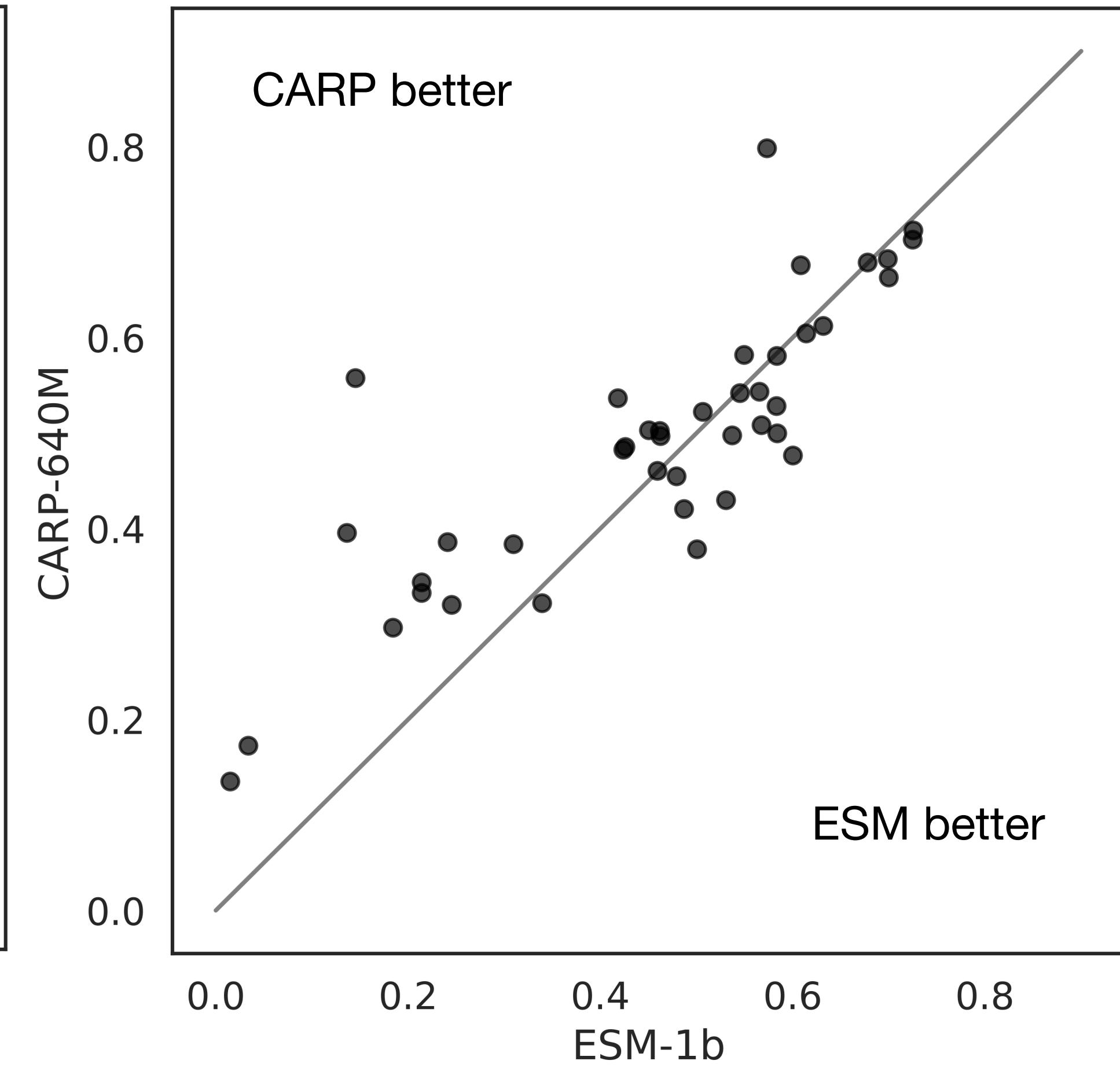
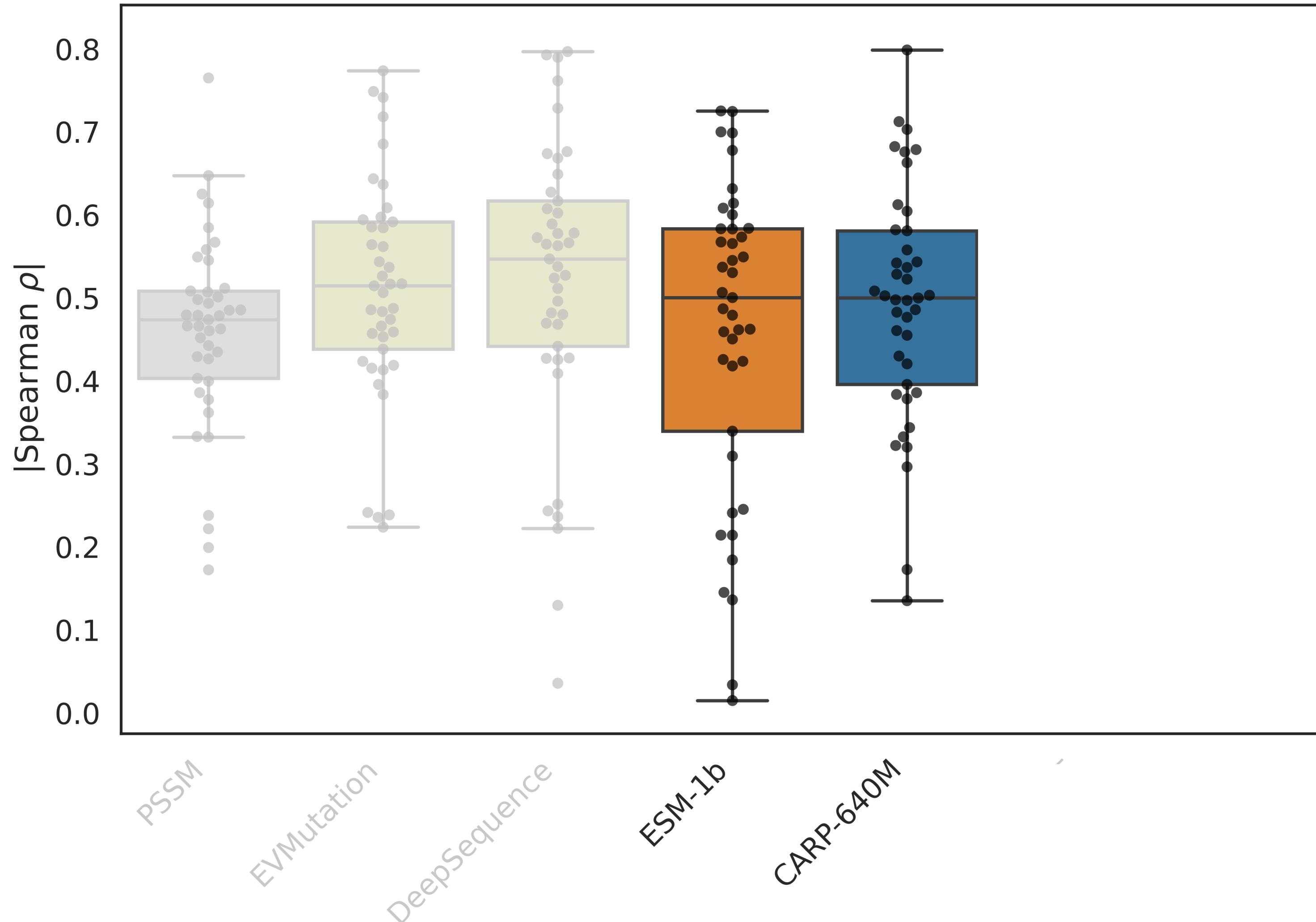
CARP is a zero-shot fitness predictor

DeepSequence



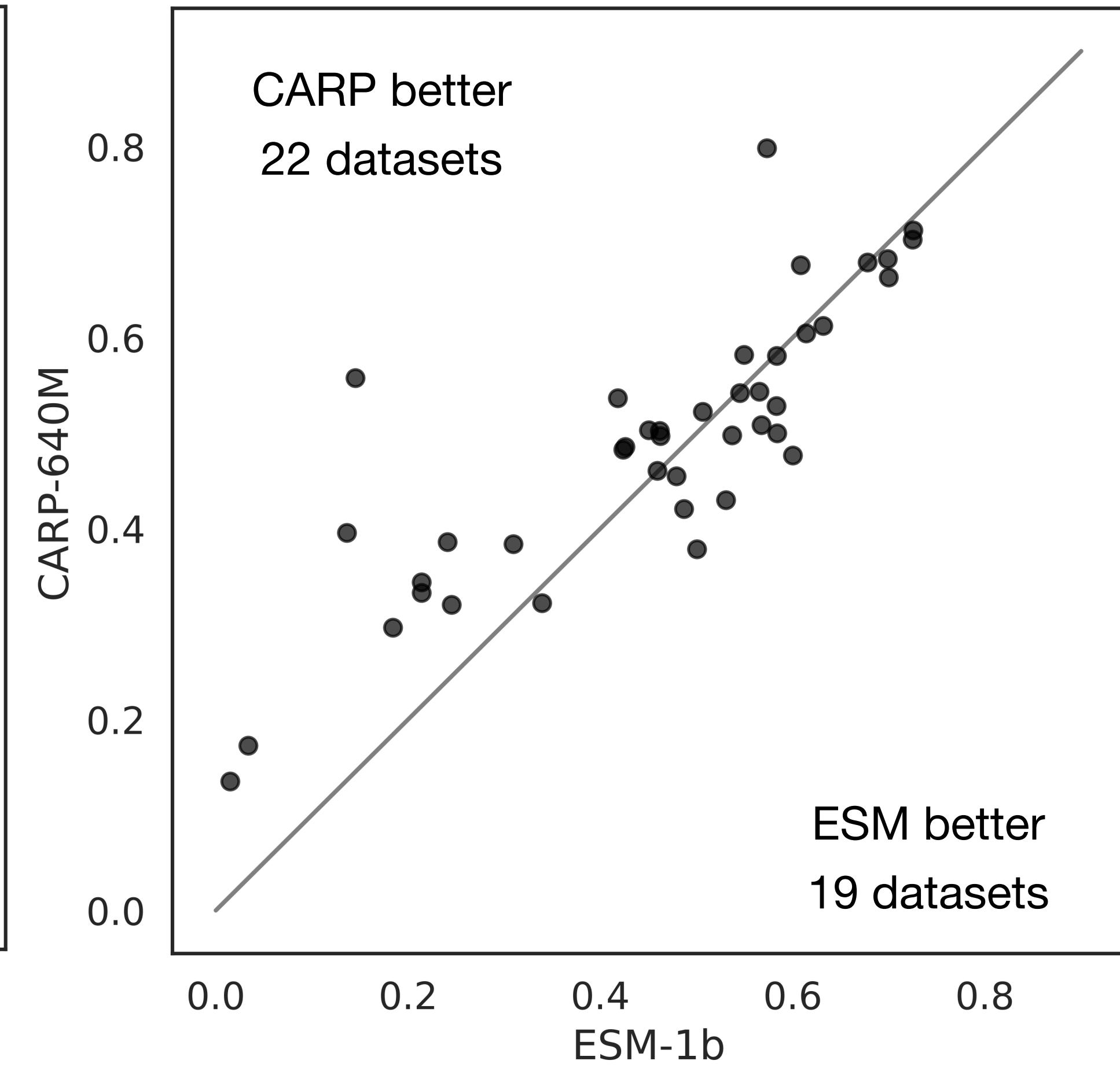
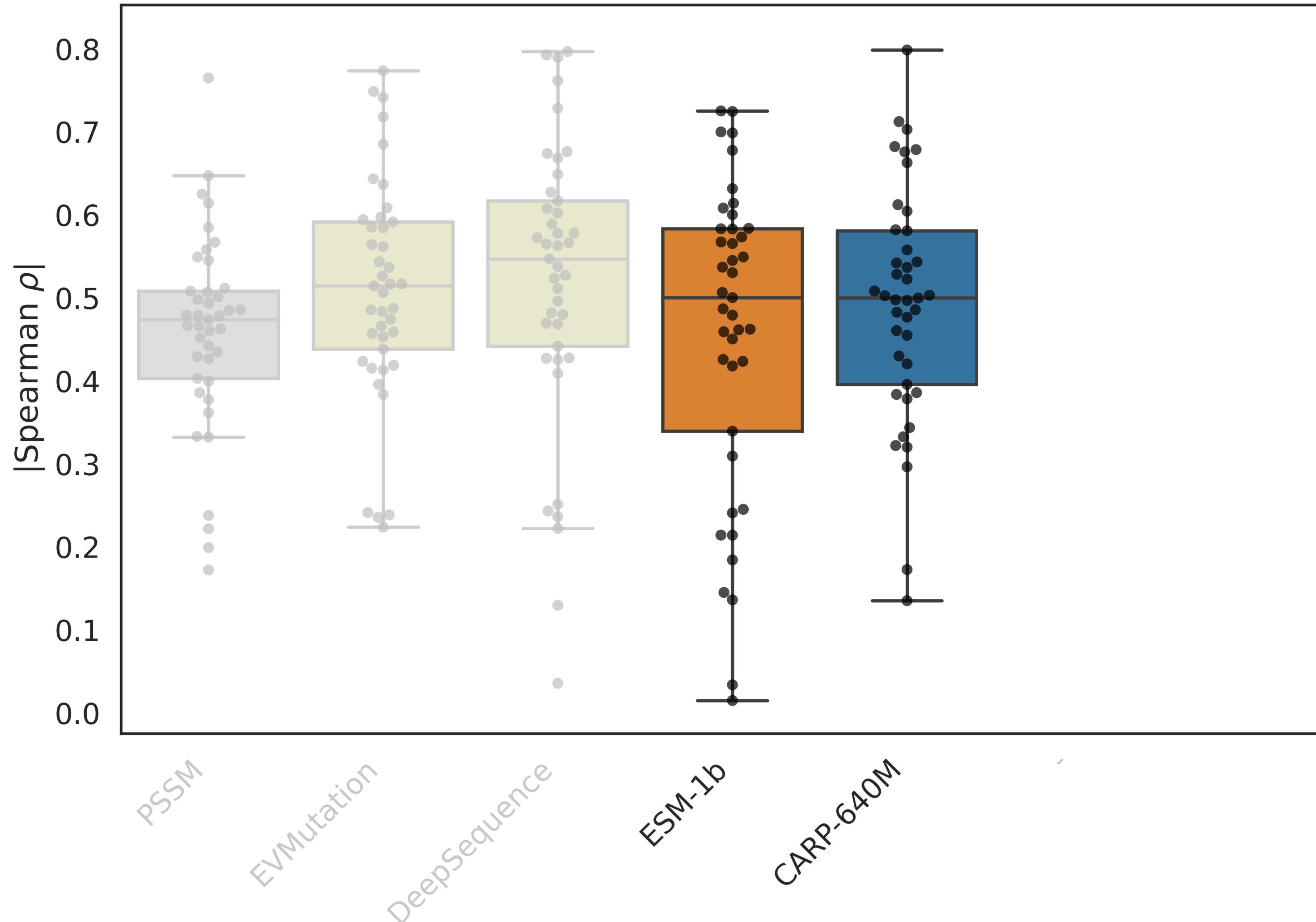
CARP is a zero-shot fitness predictor

DeepSequence



CARP is a zero-shot fitness predictor

DeepSequence



CARP learns structure

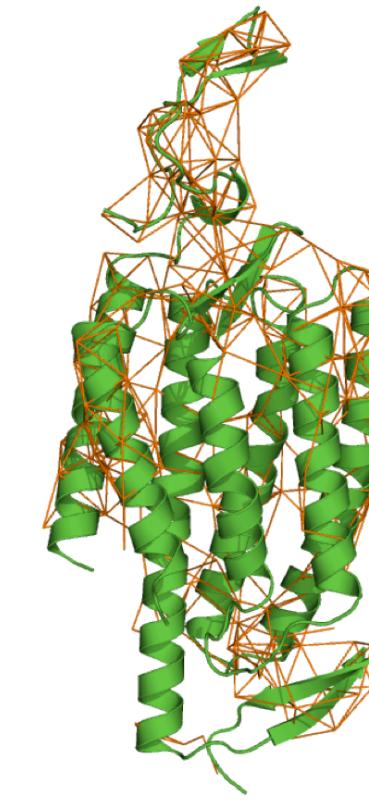
CARP learns structure

Model

ESM-1b (Rives *et al.*)

CARP-640M

CARP learns structure

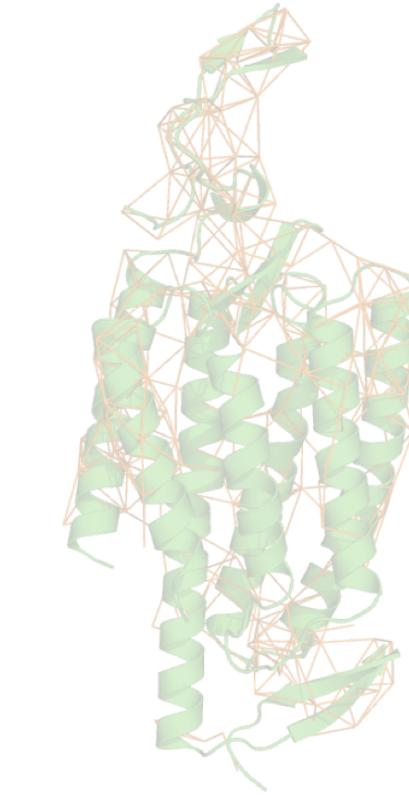


long-range contacts

Model	CAMEO
ESM-1b (Rives <i>et al.</i>)	44.4
CARP-640M	42.0

CARP learns structure

Model	CAMEO	Secondary structure
ESM-1b (Rives <i>et al.</i>)	44.4	0.82
CARP-640M	42.0	0.83



long-range contacts



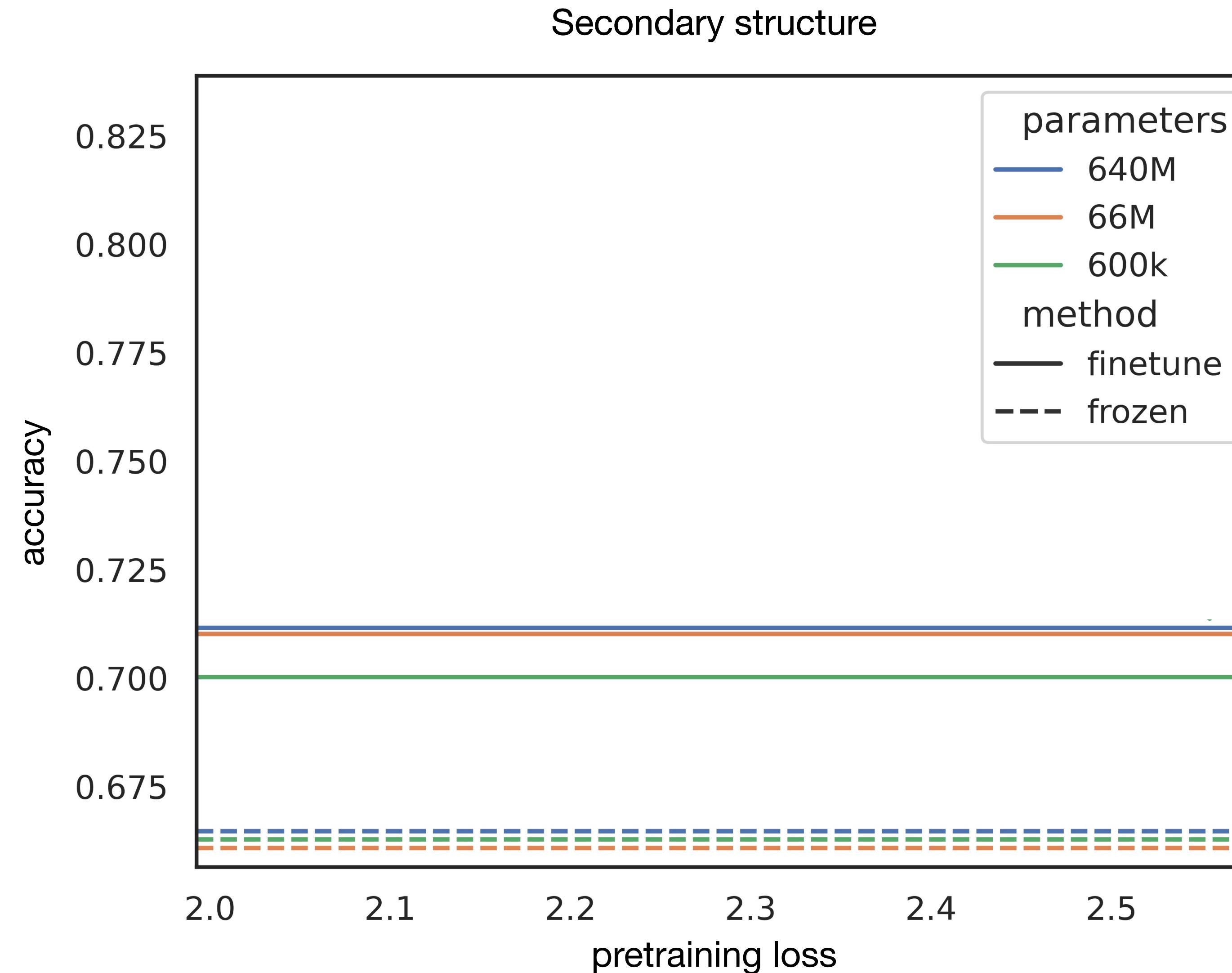
β -Sheet (3 strands)



α -helix

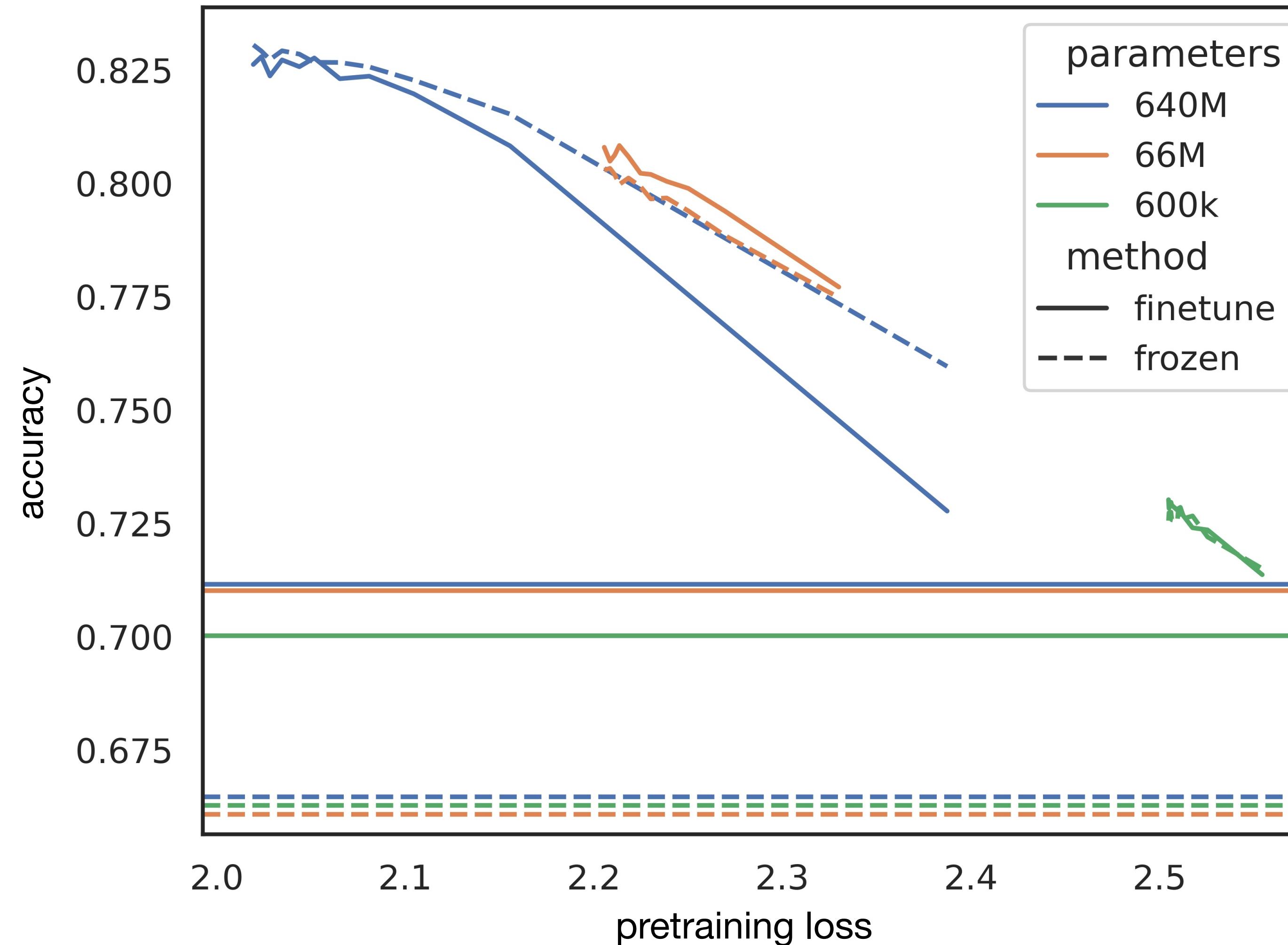
Thomas Shafee

CARP learns structure



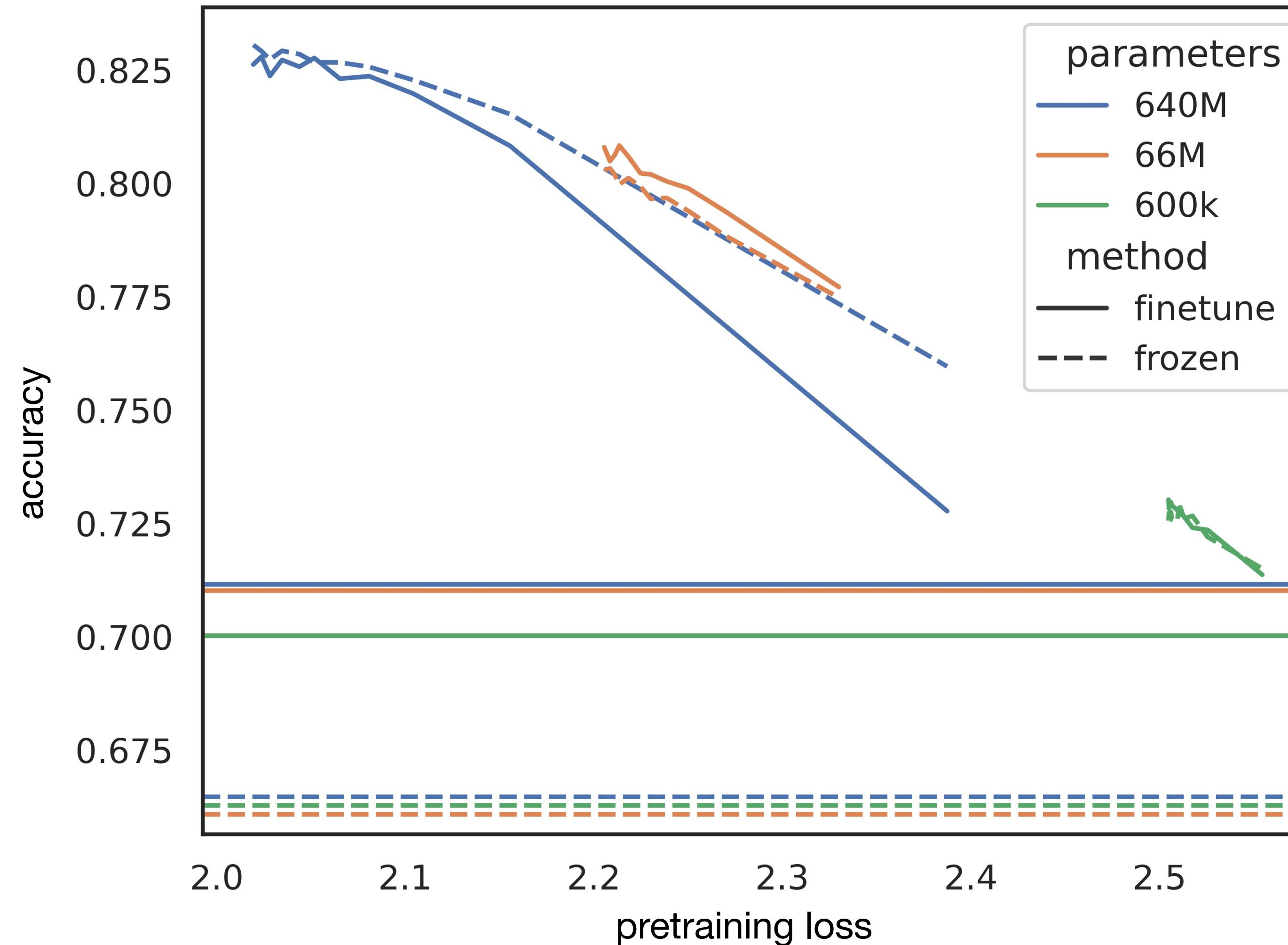
CARP learns structure

Secondary structure

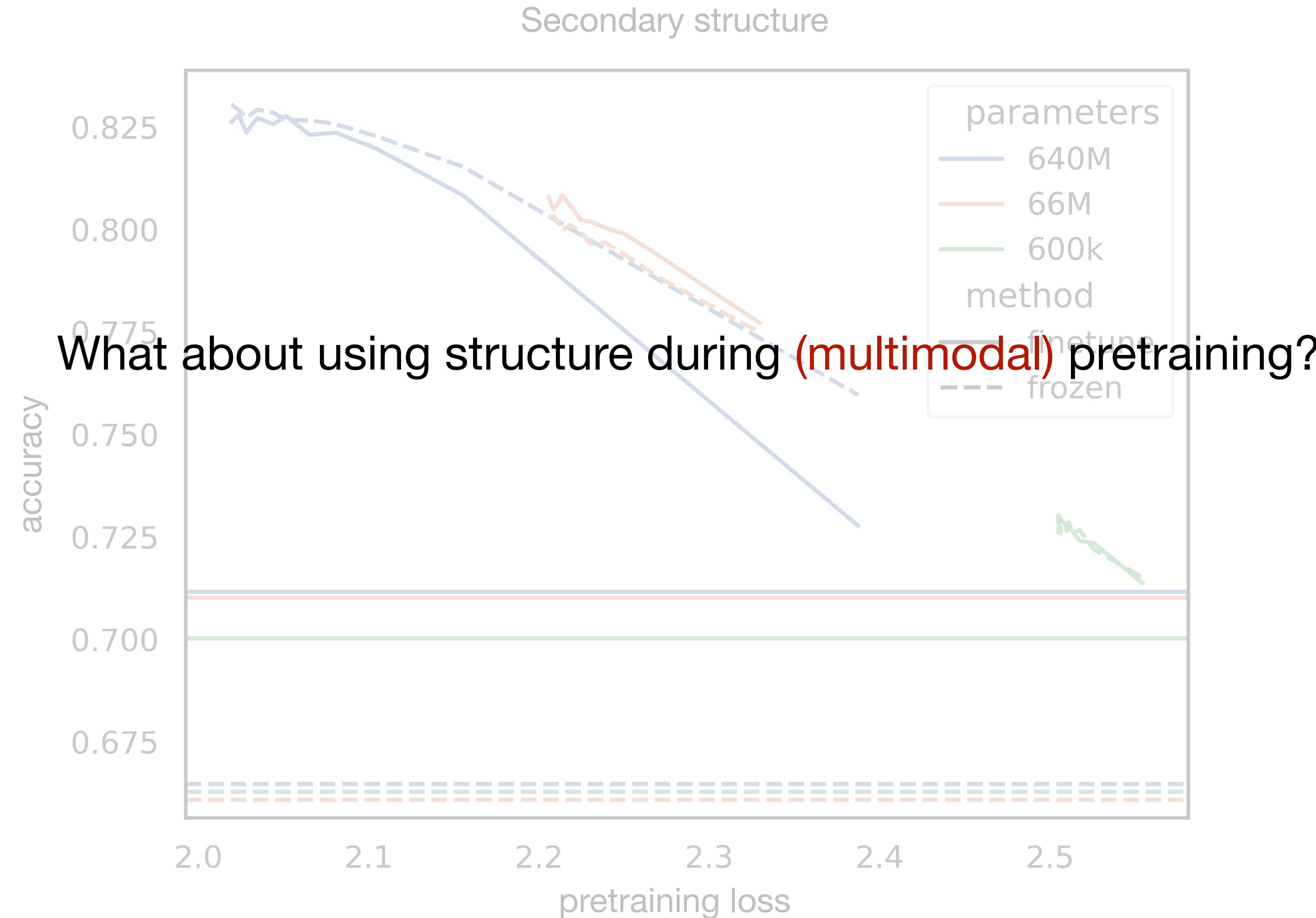


CARP learns structure

Secondary structure

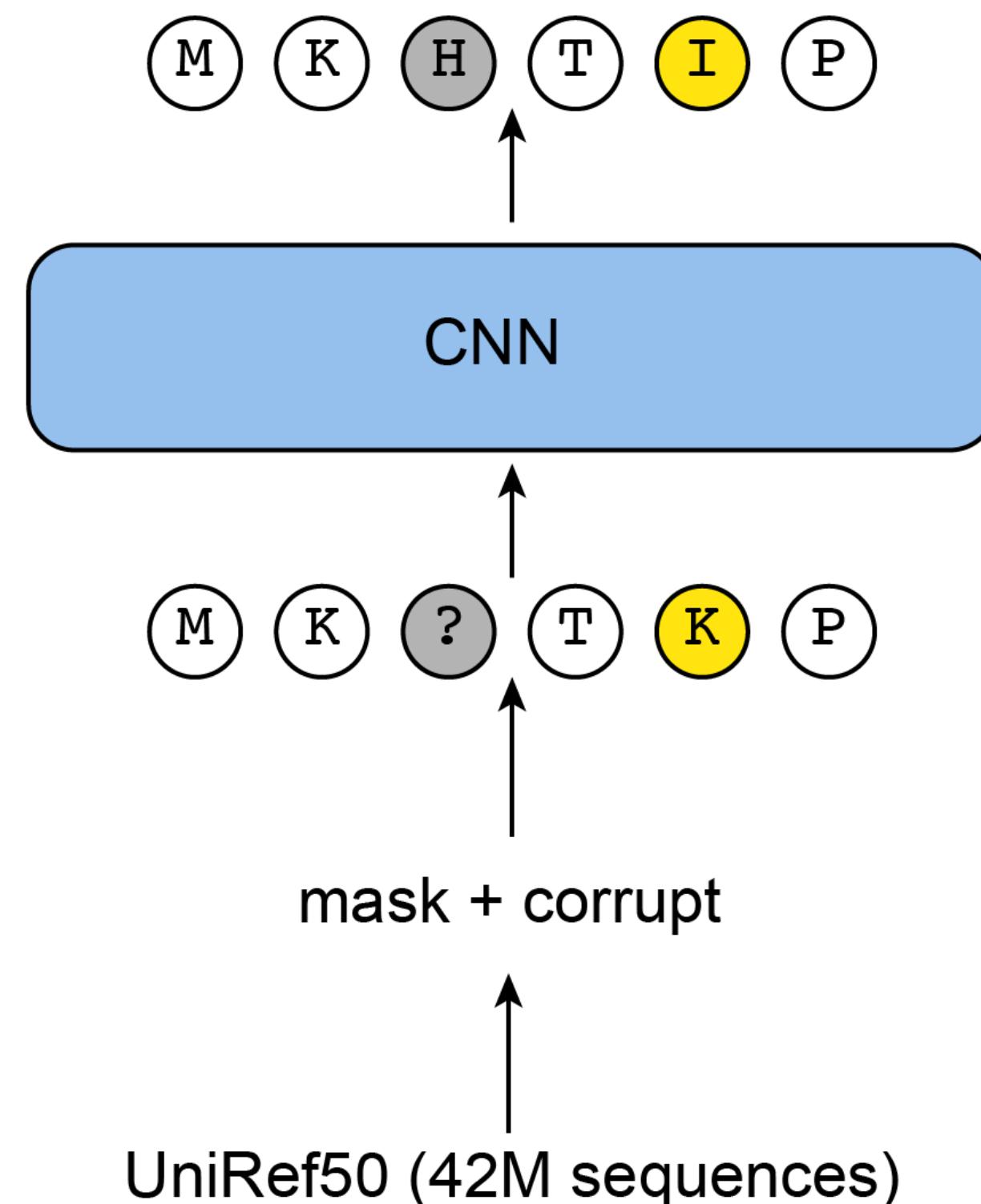


CARP learns structure



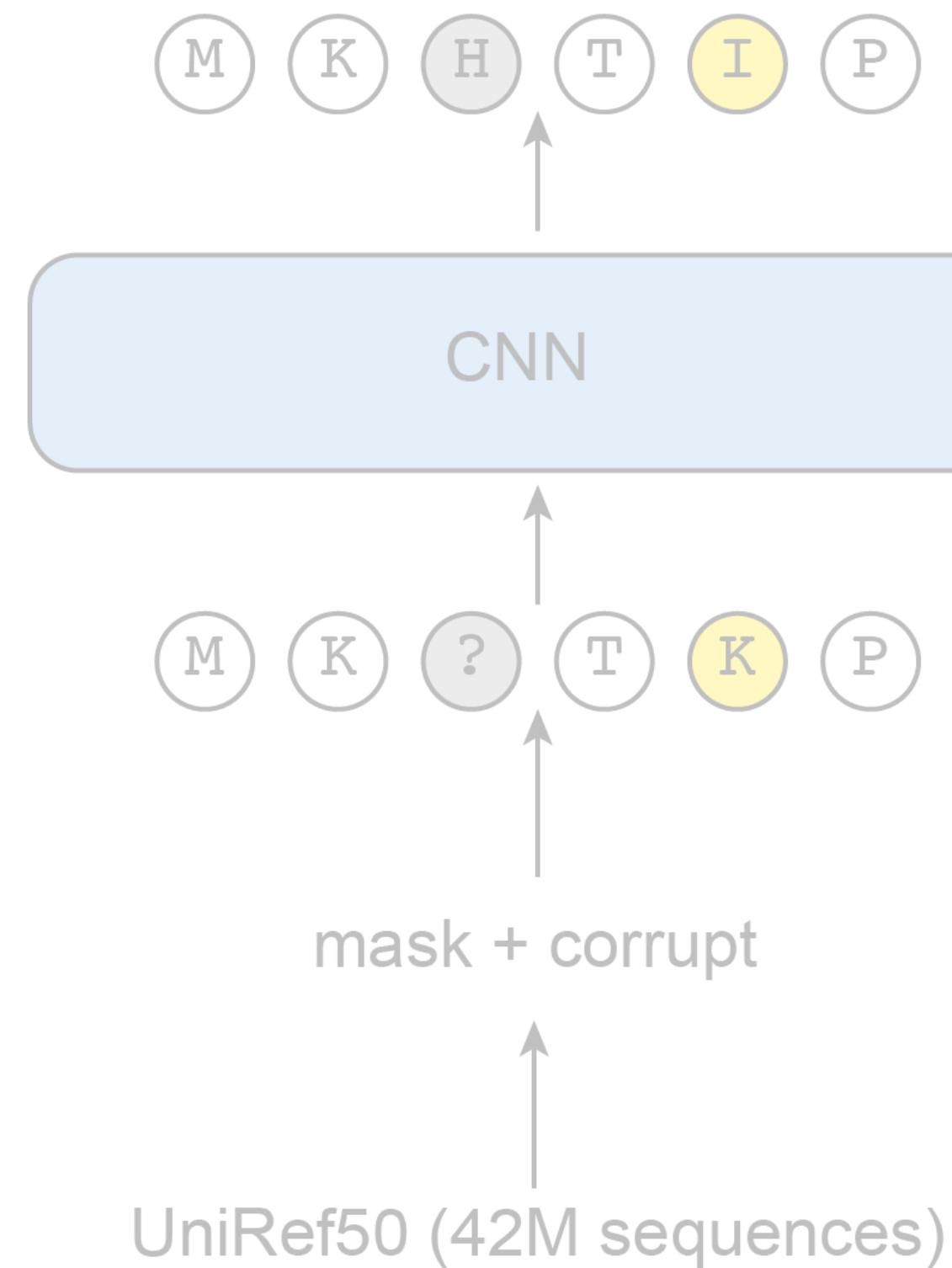
Structural information improves pretraining

Convolutional autoencoding
representations of proteins



Structural information improves pretraining

Convolutional autoencoding
representations of proteins

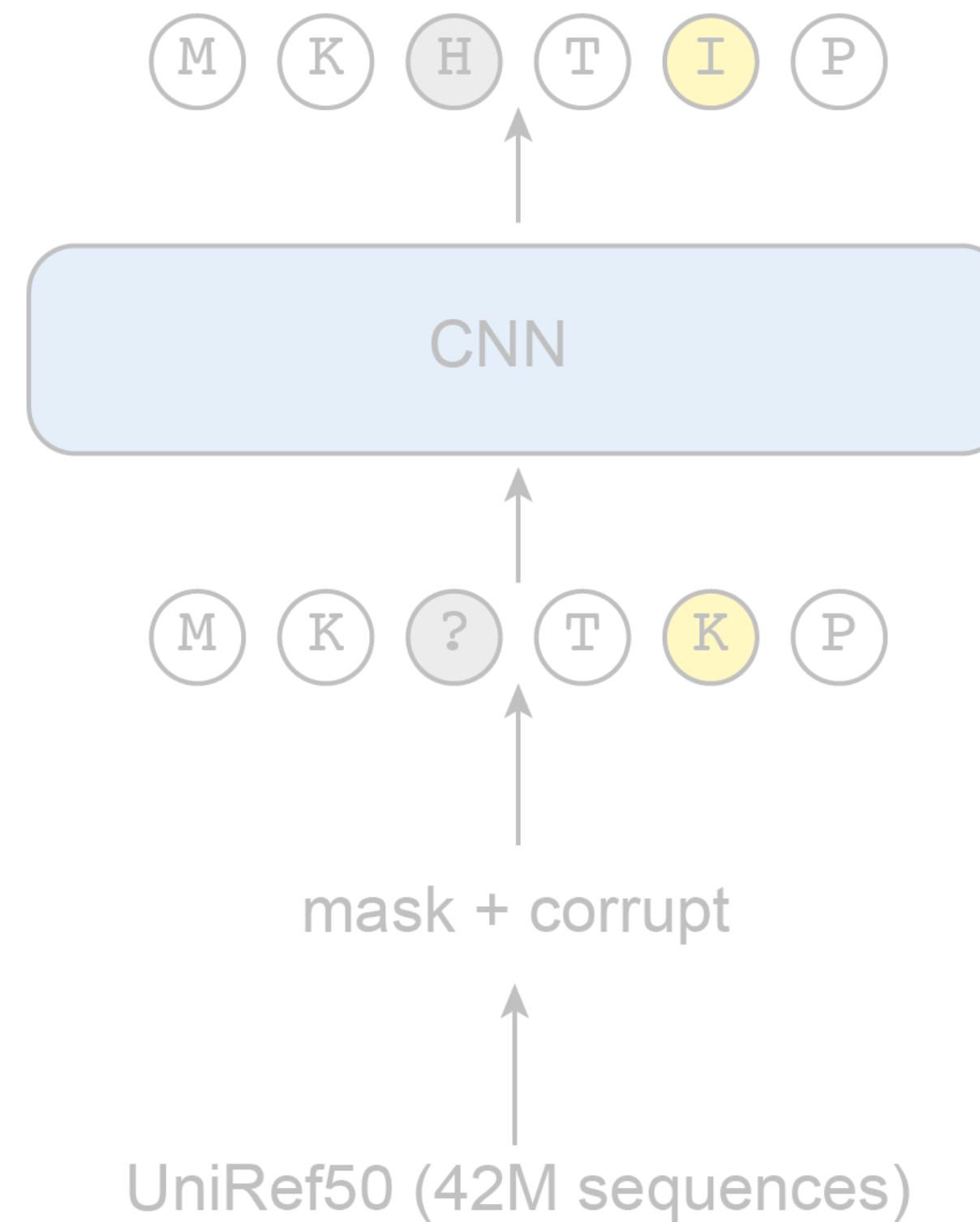


Masked inverse folding



Structural information improves pretraining

Convolutional autoencoding
representations of proteins

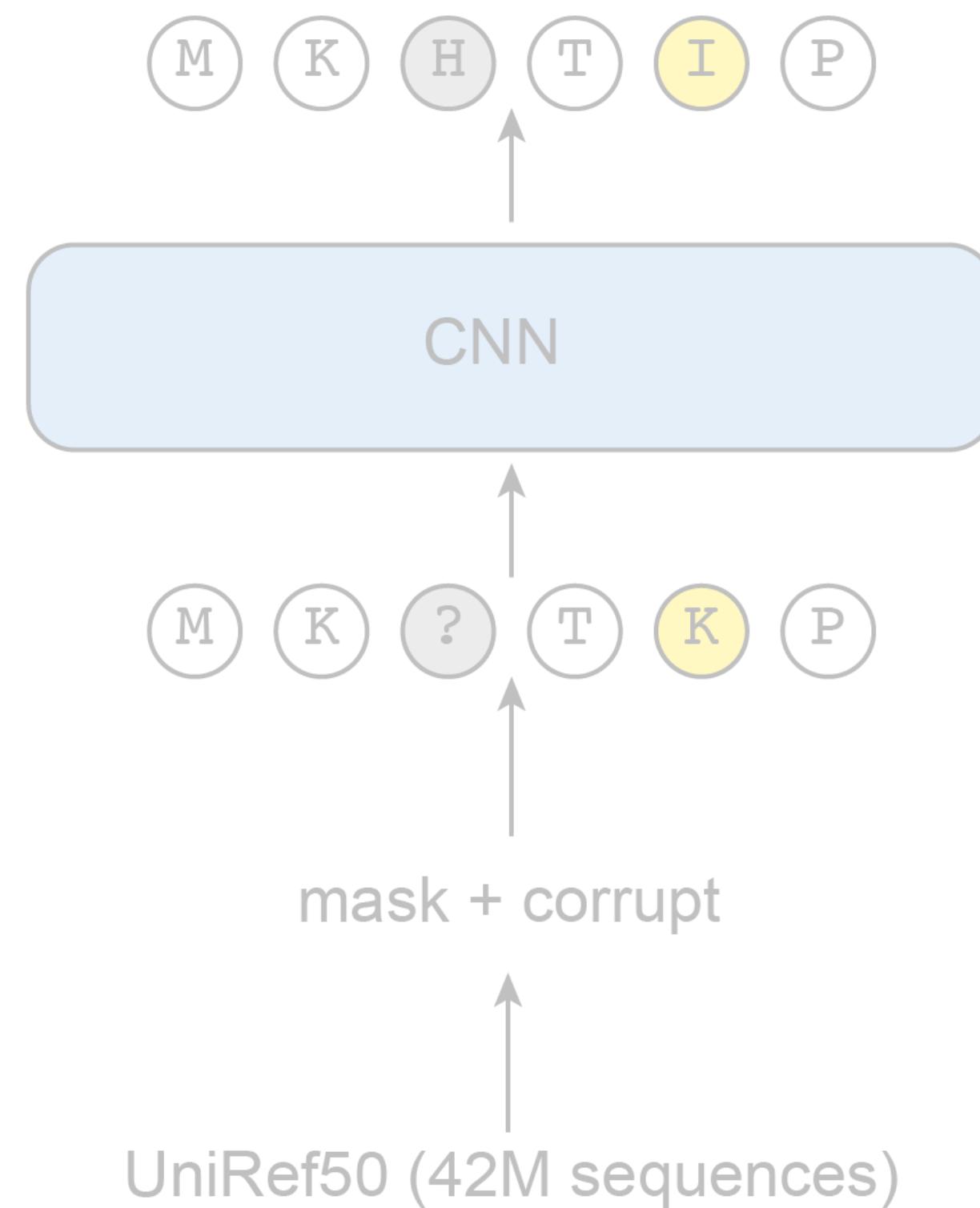


Masked inverse folding

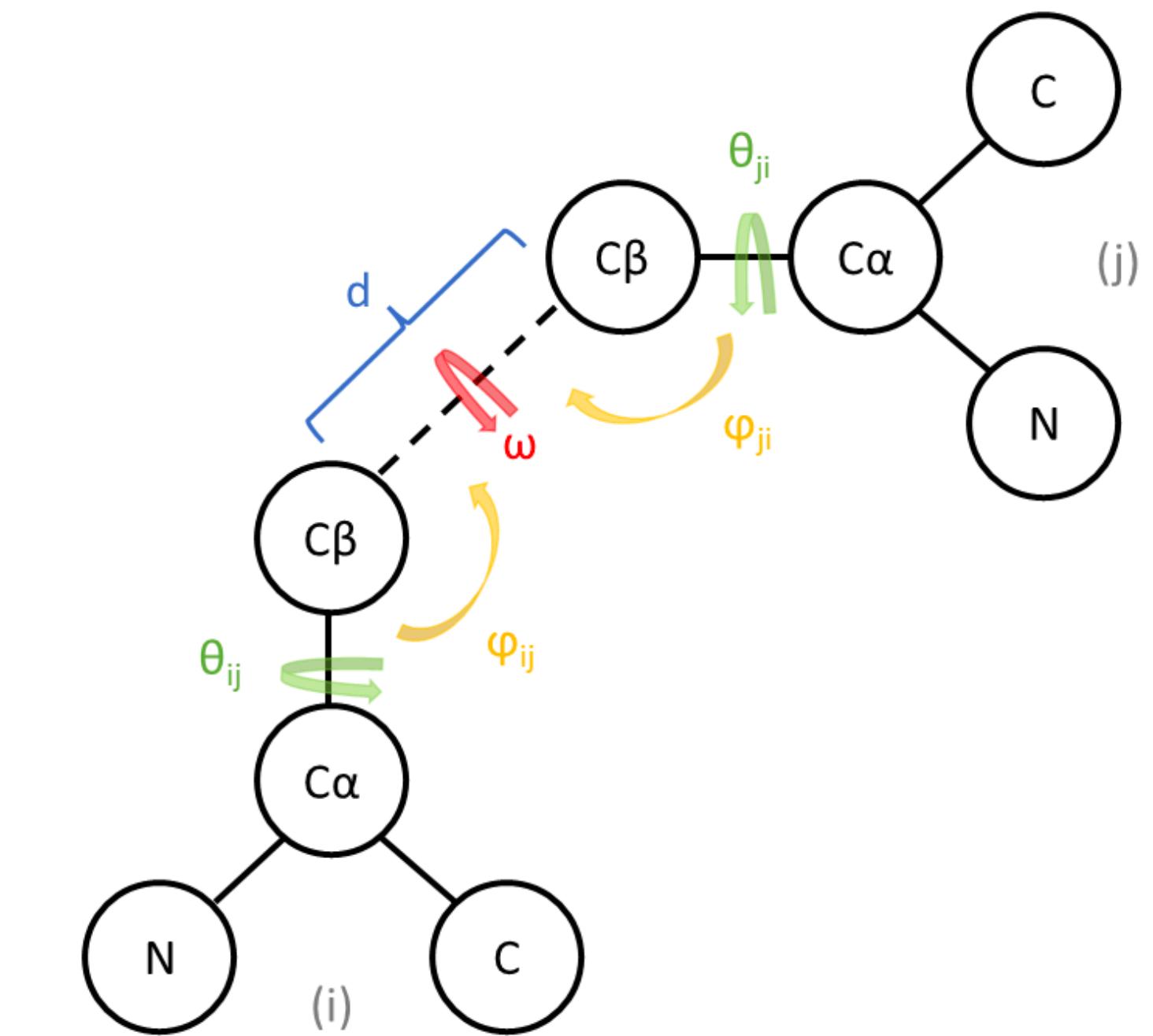


Structural information improves pretraining

Convolutional autoencoding
representations of proteins

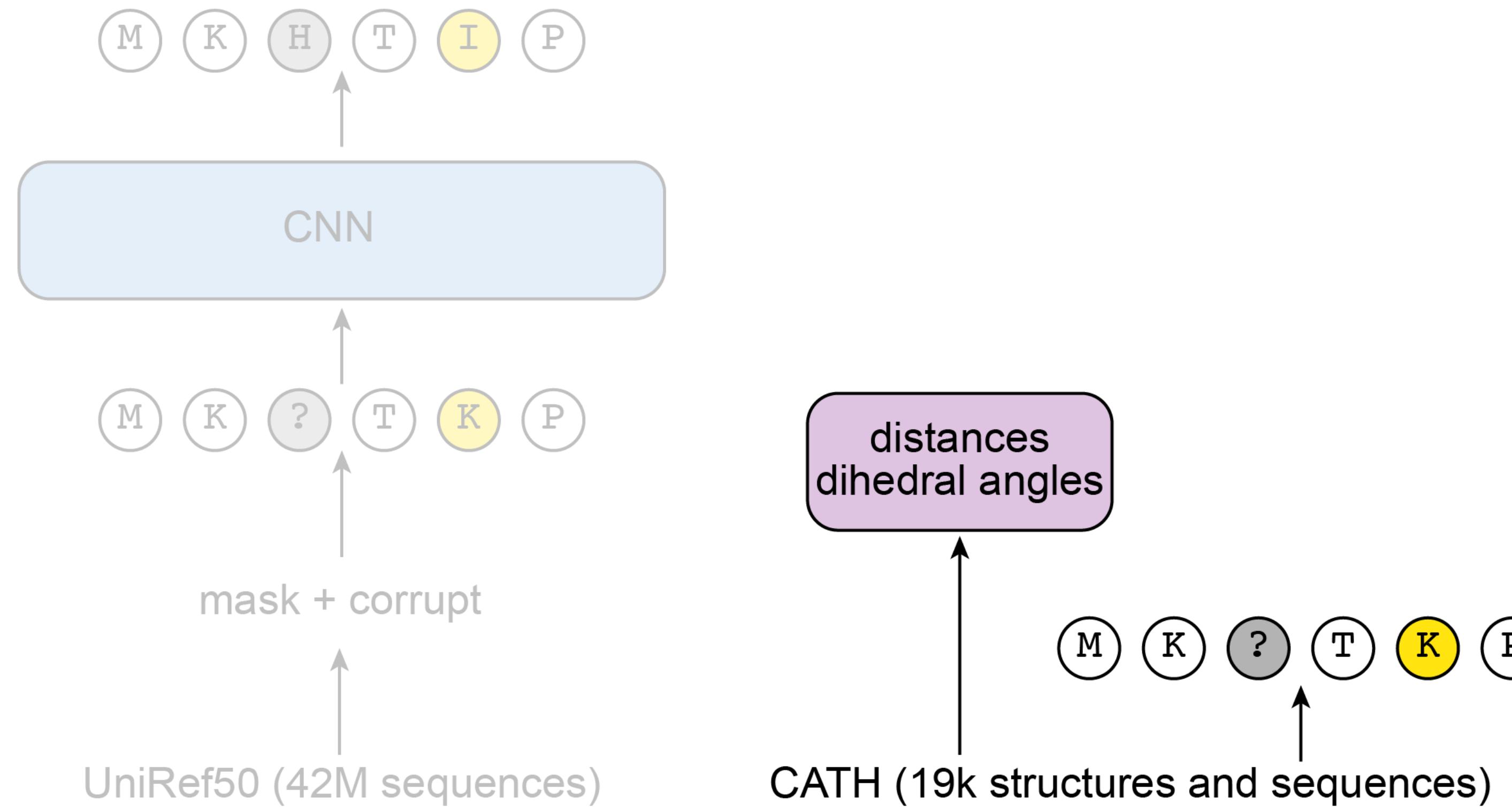


Masked inverse folding

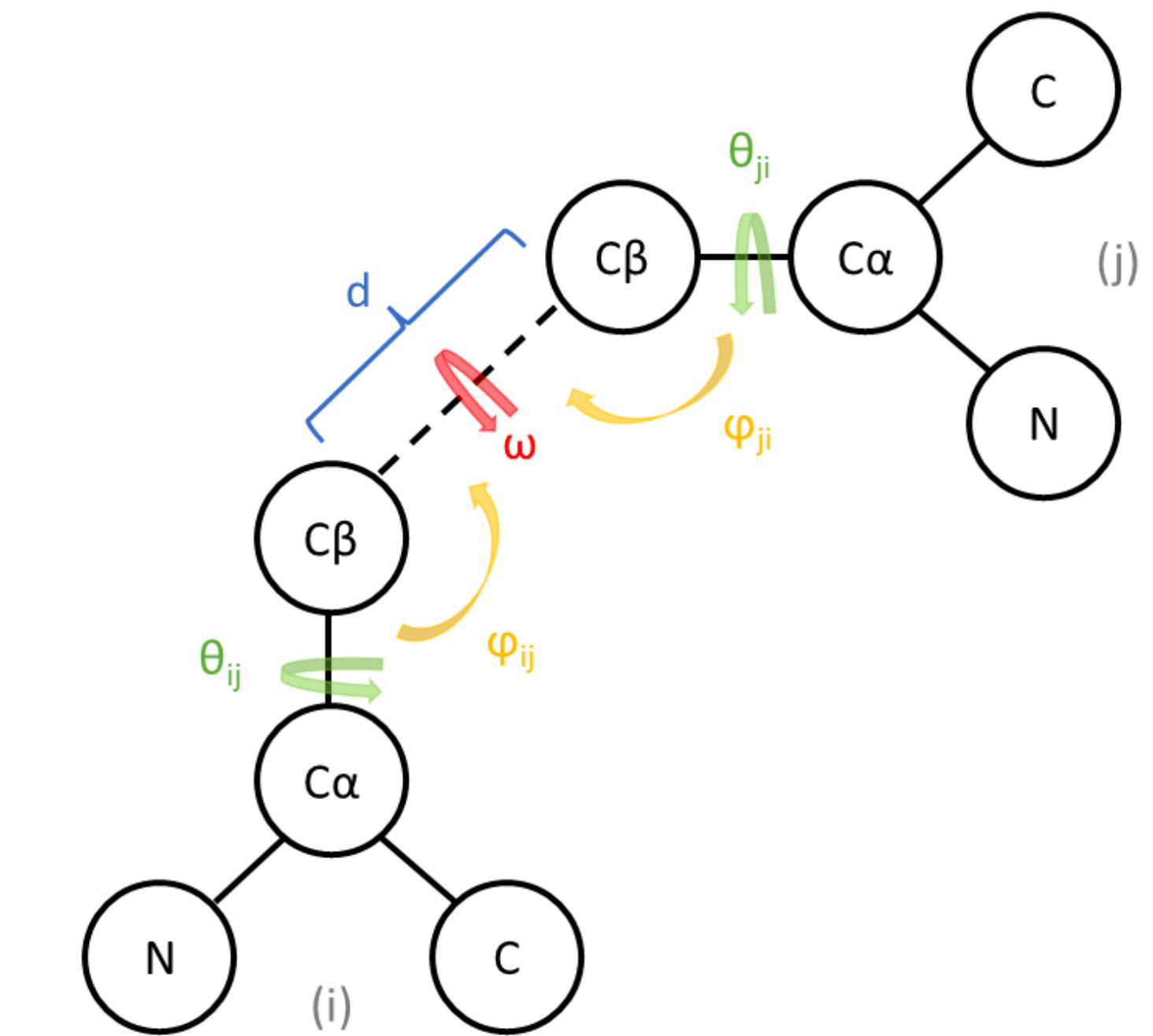


Structural information improves pretraining

Convolutional autoencoding
representations of proteins

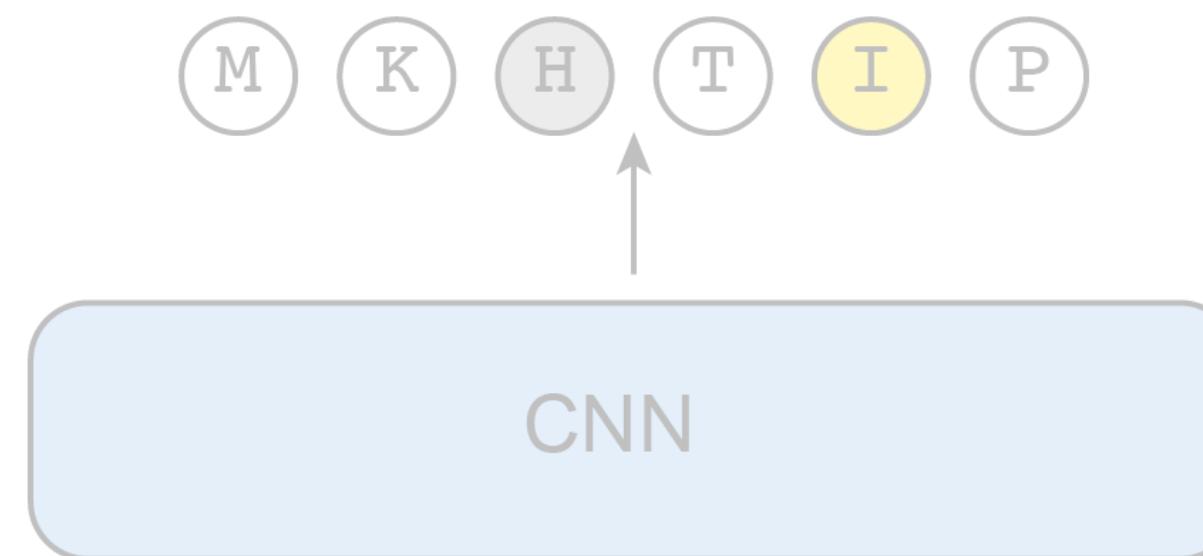


Masked inverse folding

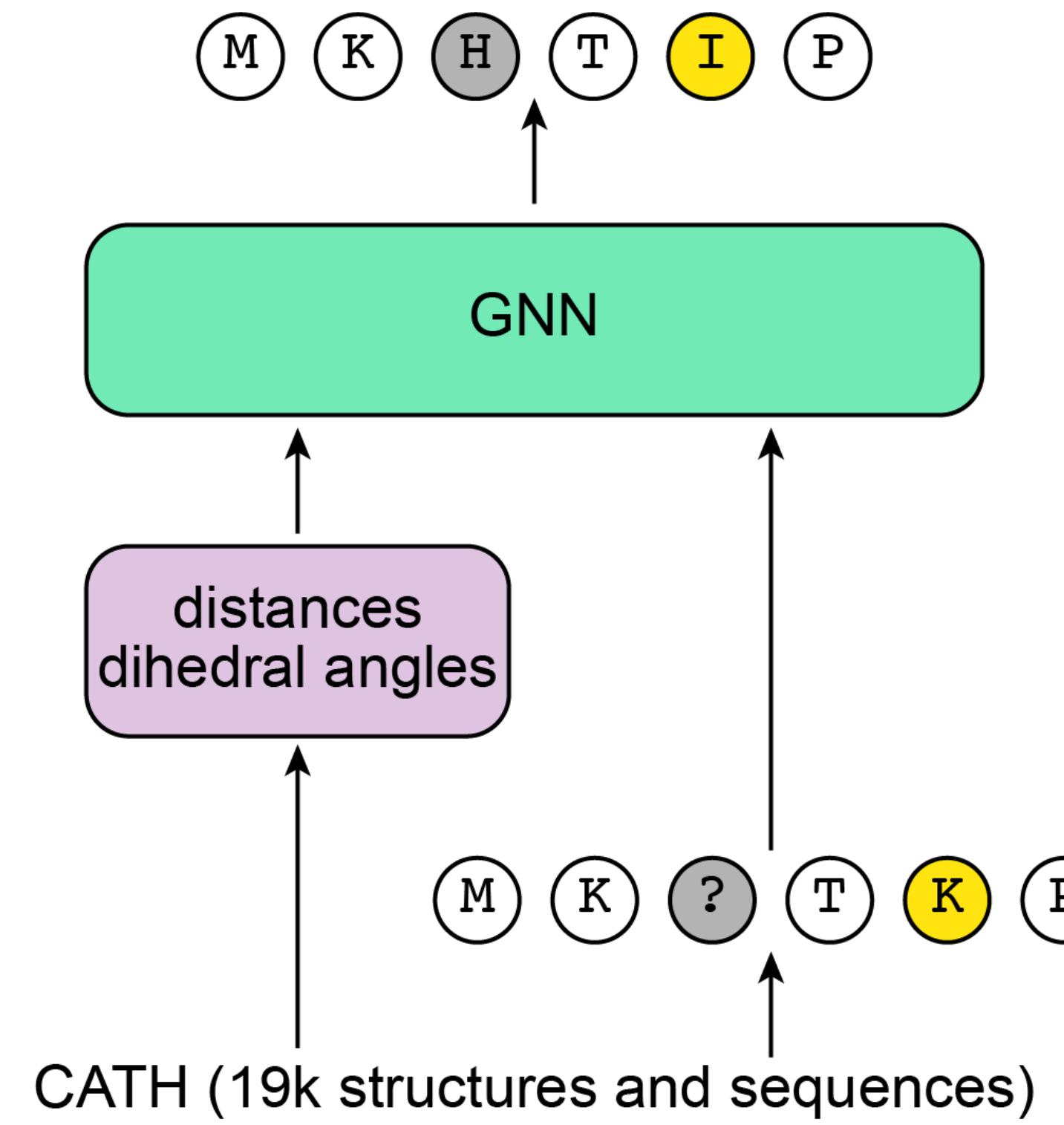


Structural information improves pretraining

Convolutional autoencoding
representations of proteins

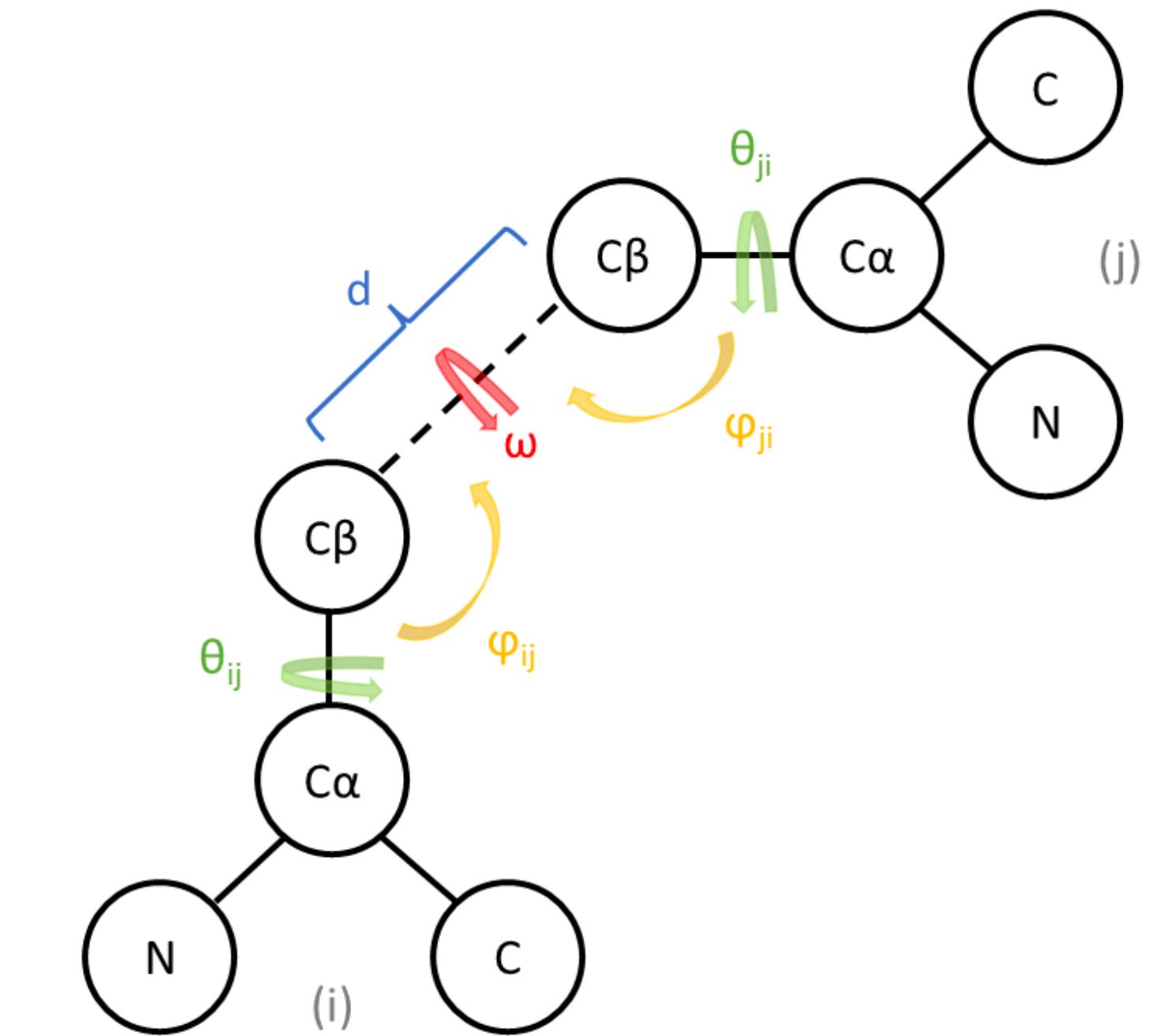


Masked inverse folding



UniRef50 (42M sequences)

CATH (19k structures and sequences)



Structural information improves pretraining

Convolutional autoencoding
representations of proteins

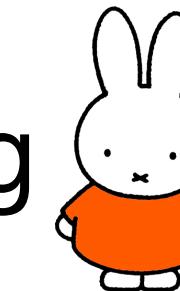


mask + corrupt

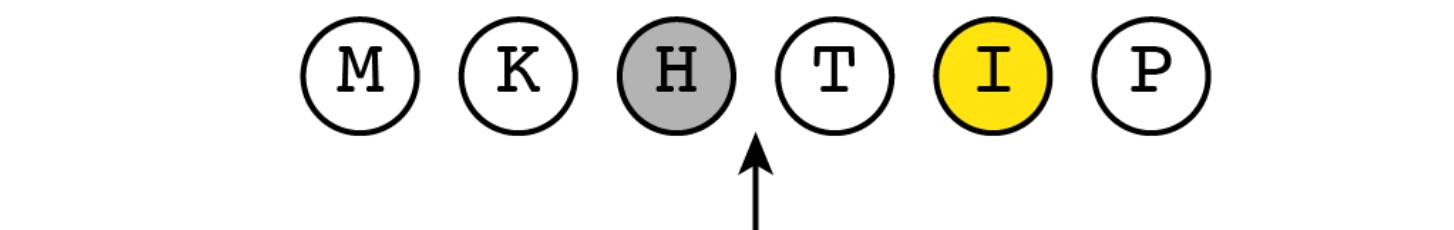
UniRef50 (42M sequences)



Masked inverse folding



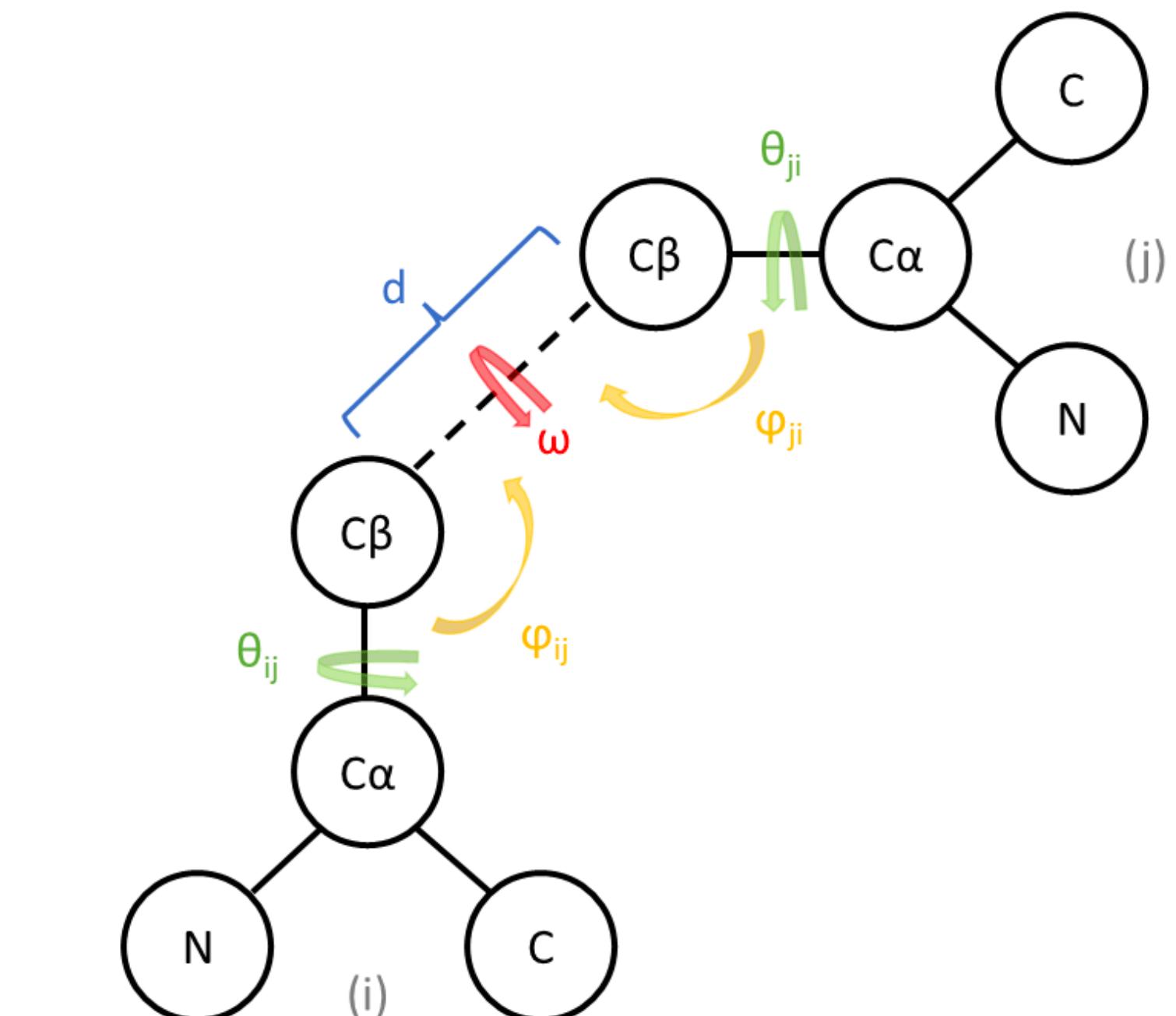
Masked inverse folding



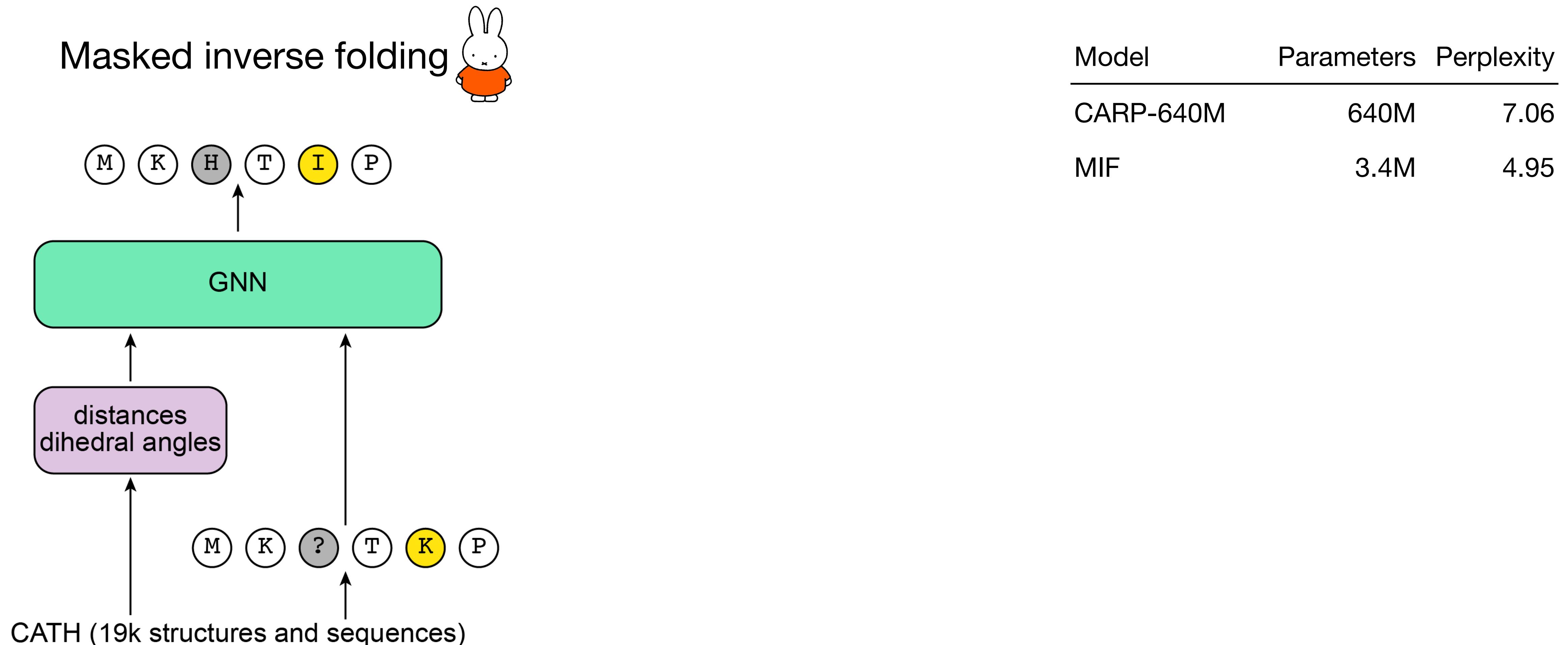
distances
dihedral angles

CATH (19k structures and sequences)

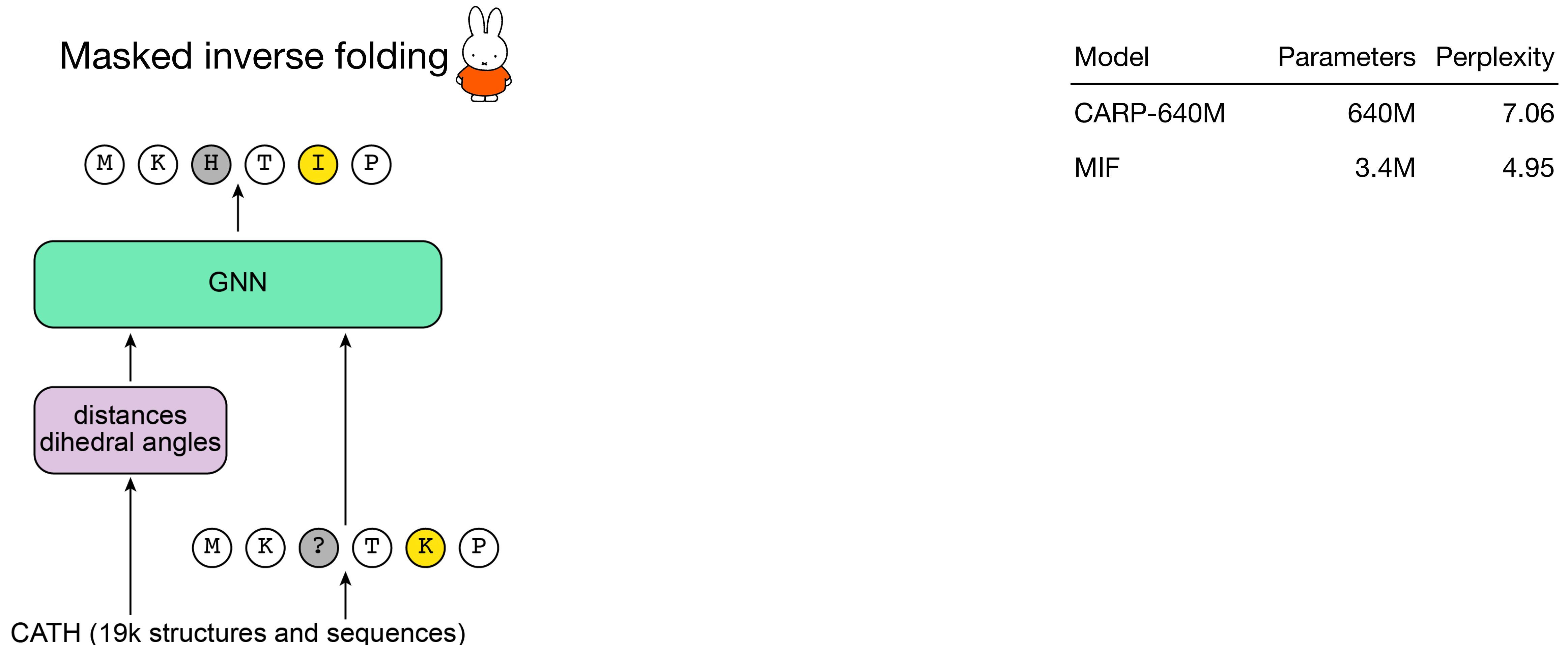
Model	Parameters	Perplexity
CARP-640M	640M	7.06
MIF	3.4M	4.95



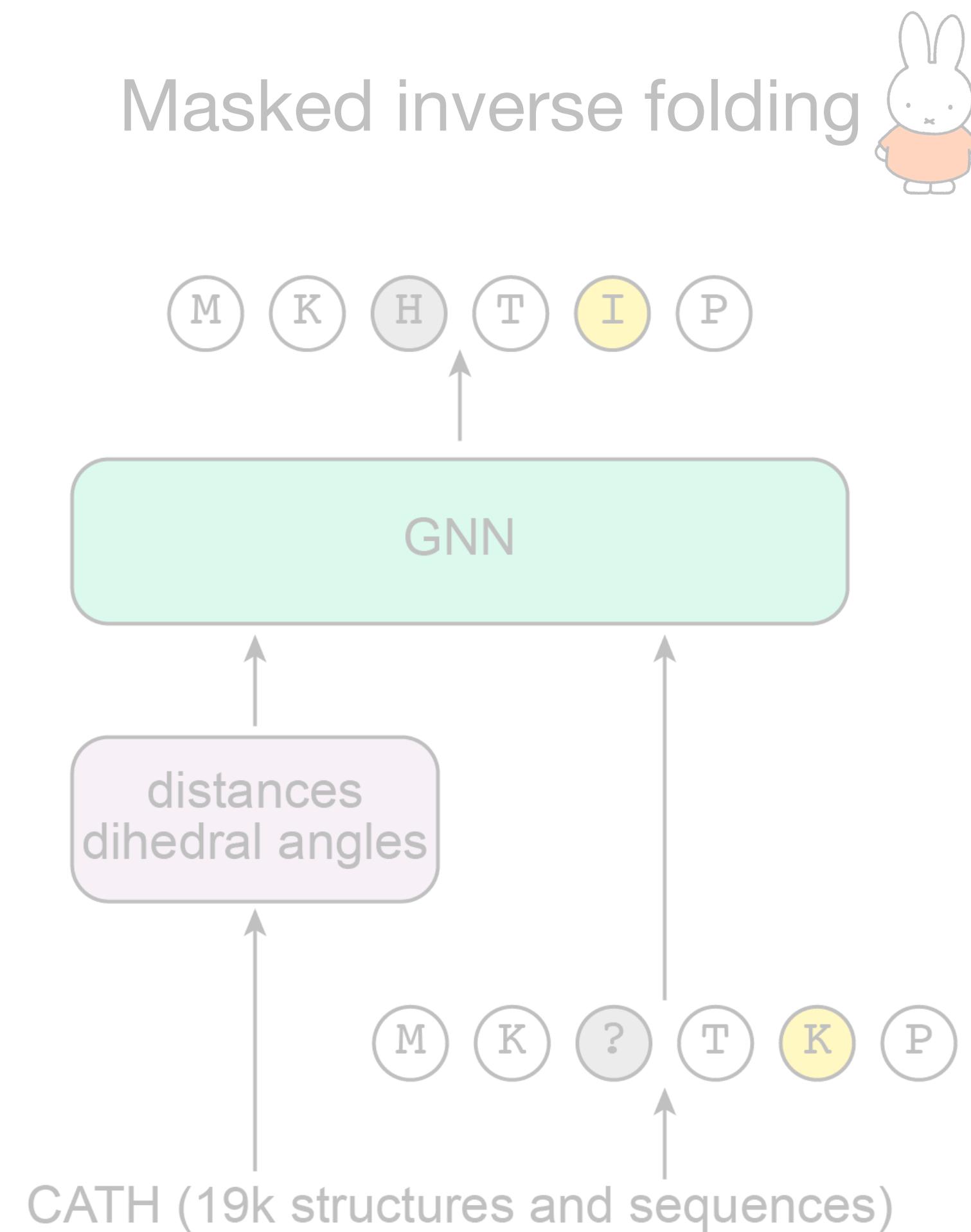
Structural information improves pretraining



Sequence transfer improves pretraining more



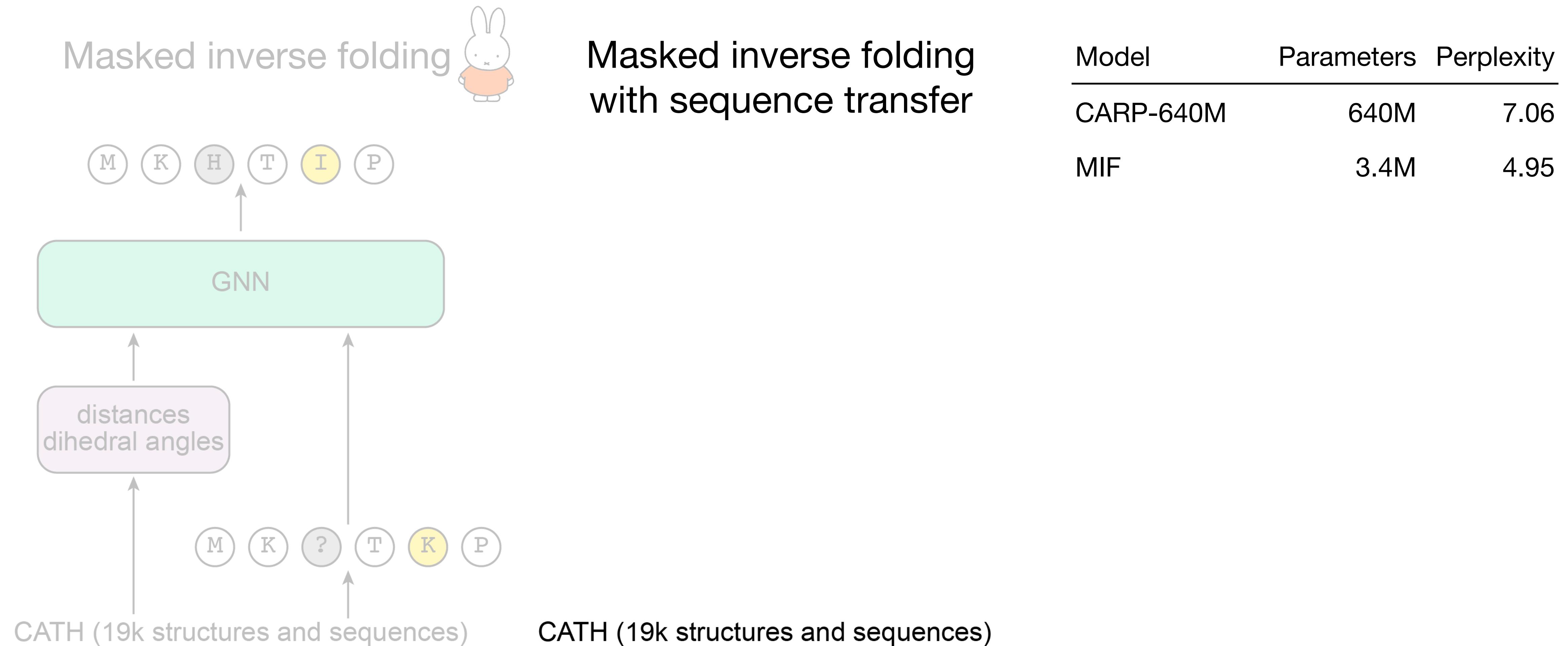
Sequence transfer improves pretraining more



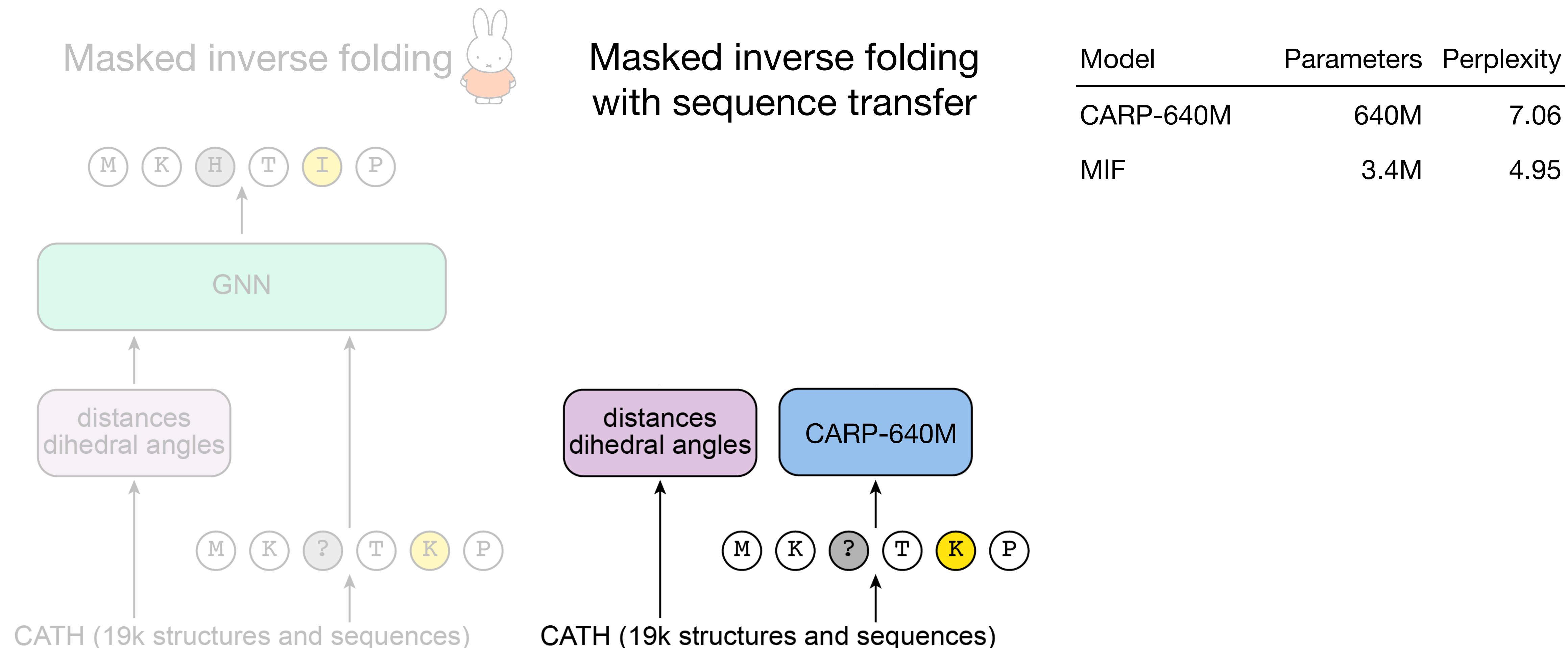
Masked inverse folding
with sequence transfer

Model	Parameters	Perplexity
CARP-640M	640M	7.06
MIF	3.4M	4.95

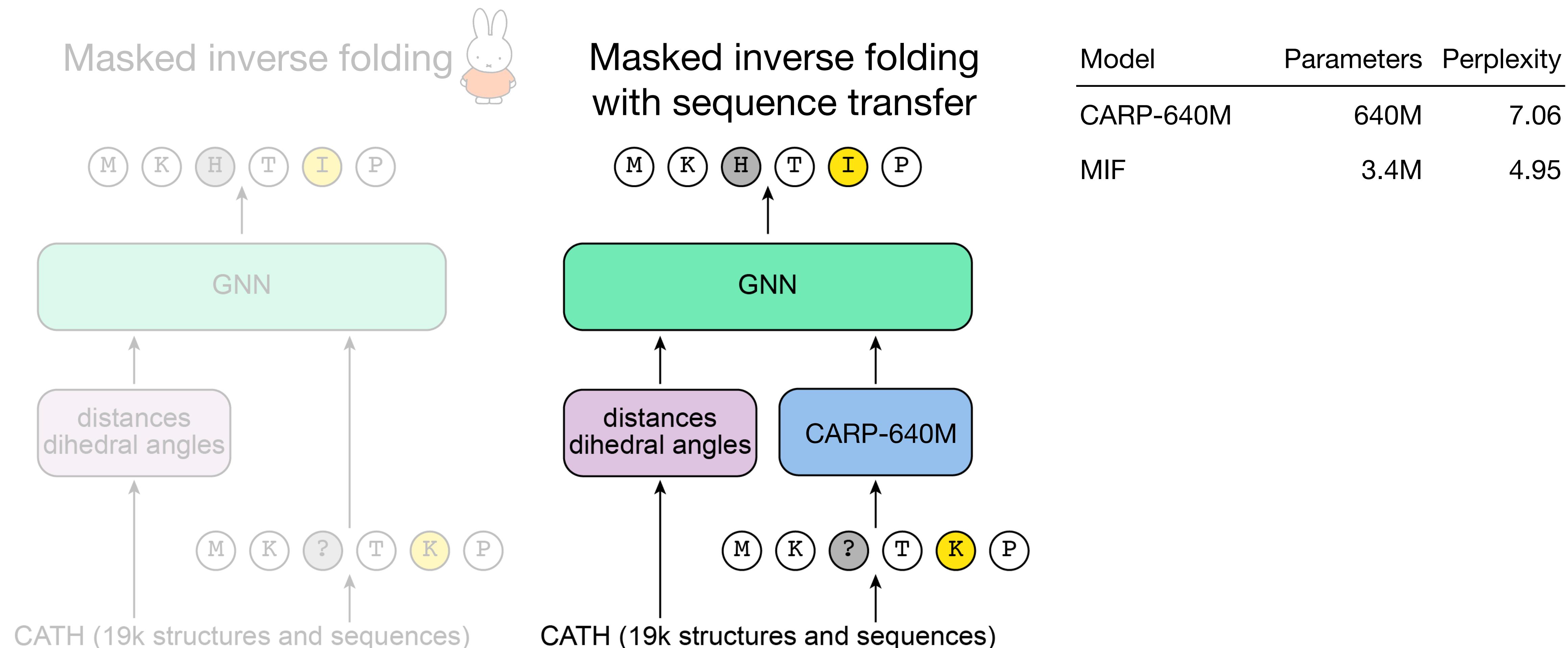
Sequence transfer improves pretraining more



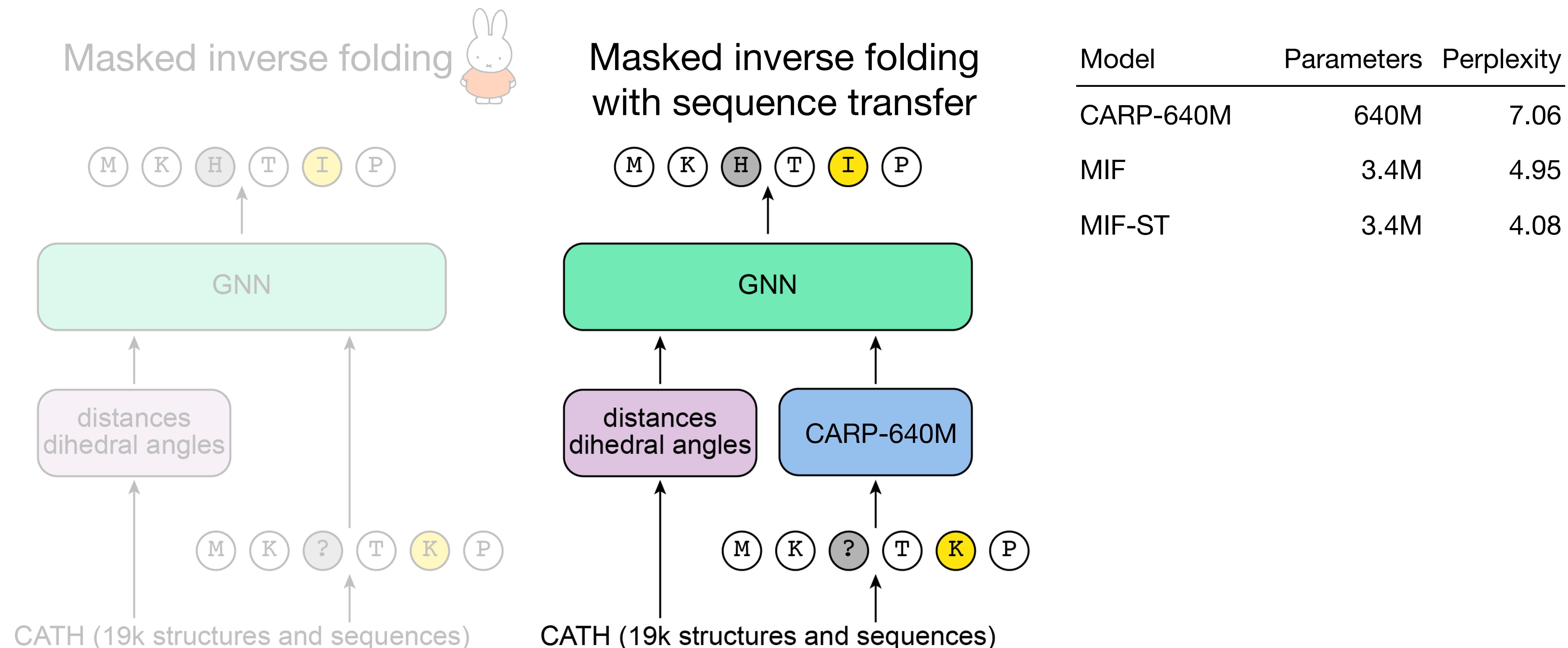
Sequence transfer improves pretraining more



Sequence transfer improves pretraining more

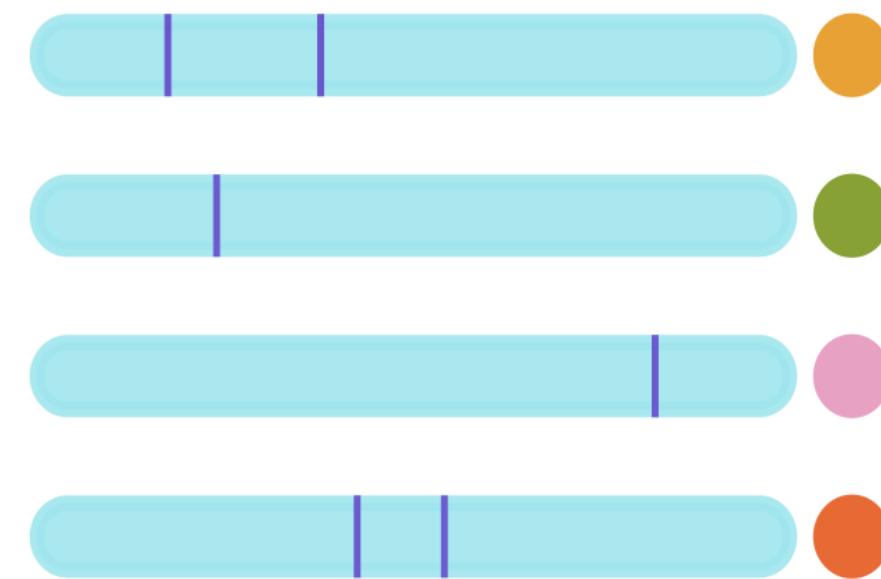


Sequence transfer improves pretraining more

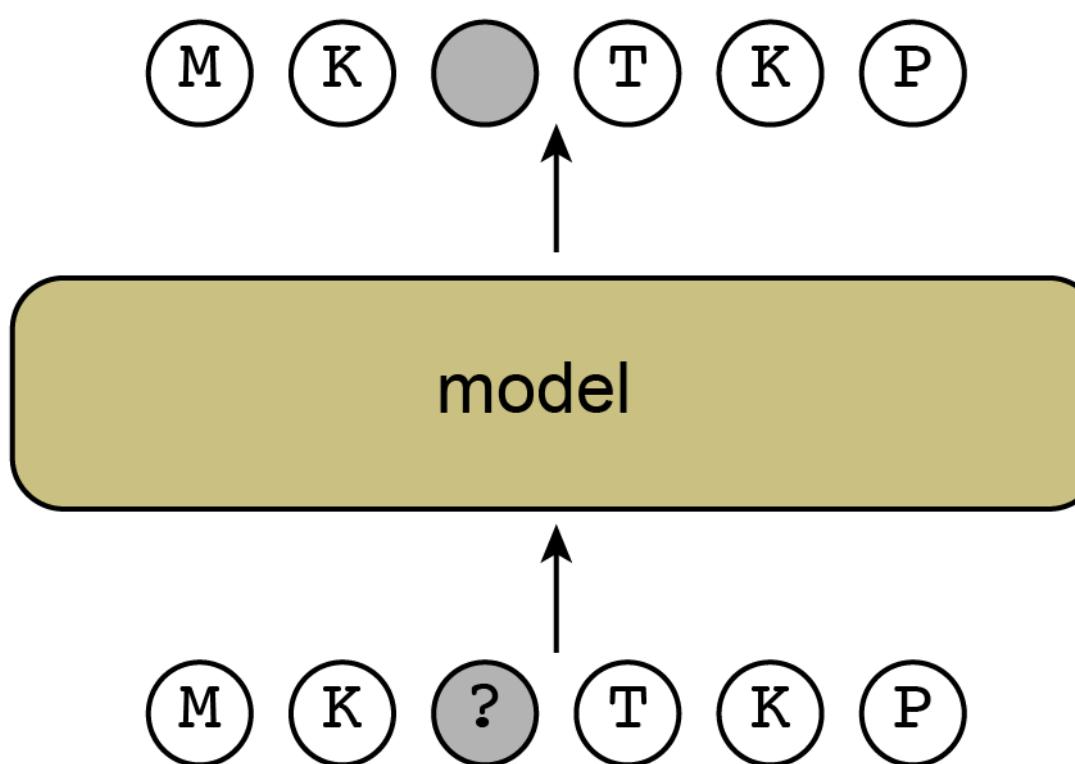


MIF, and MIF-ST are zero-shot fitness predictors

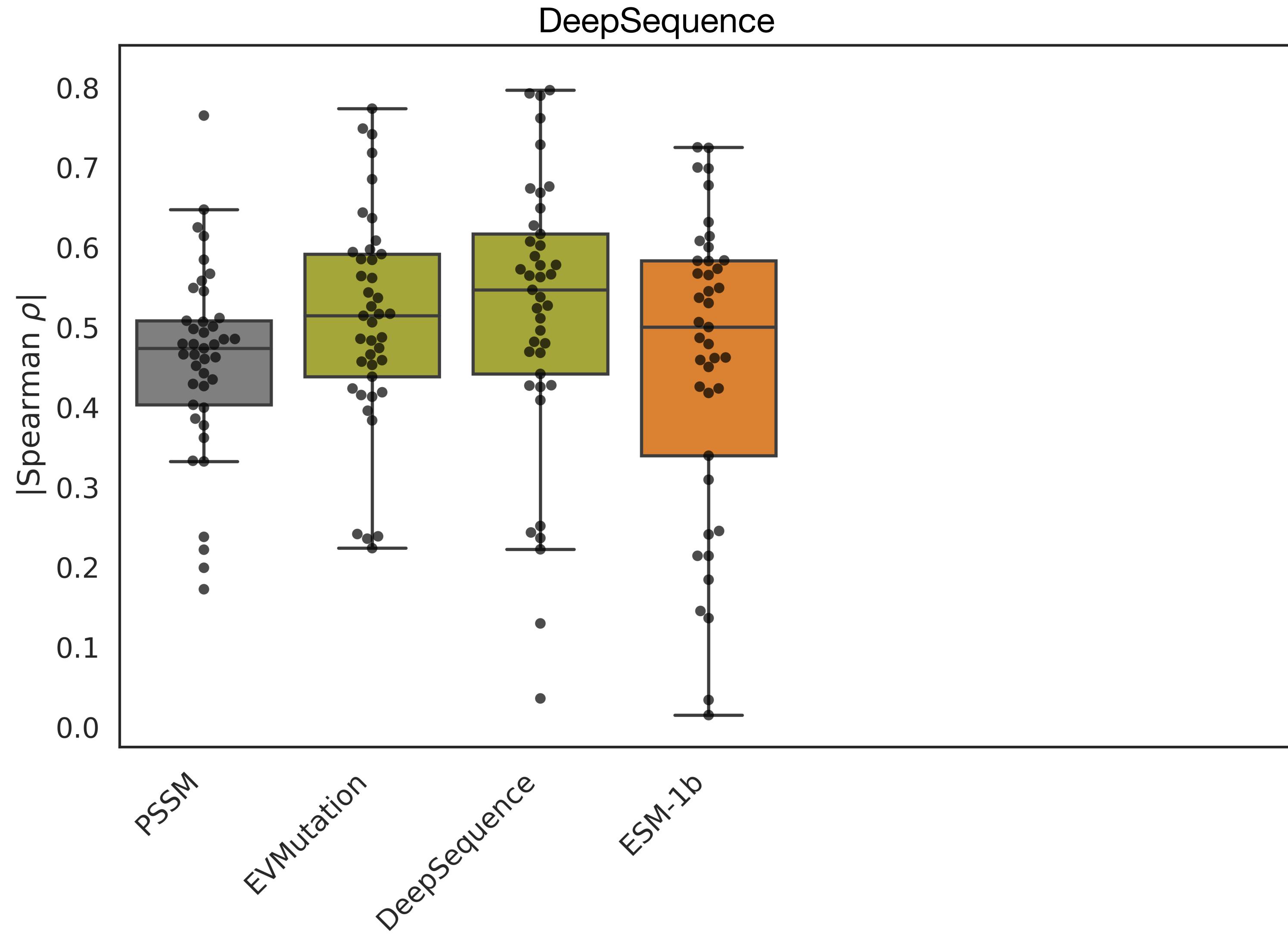
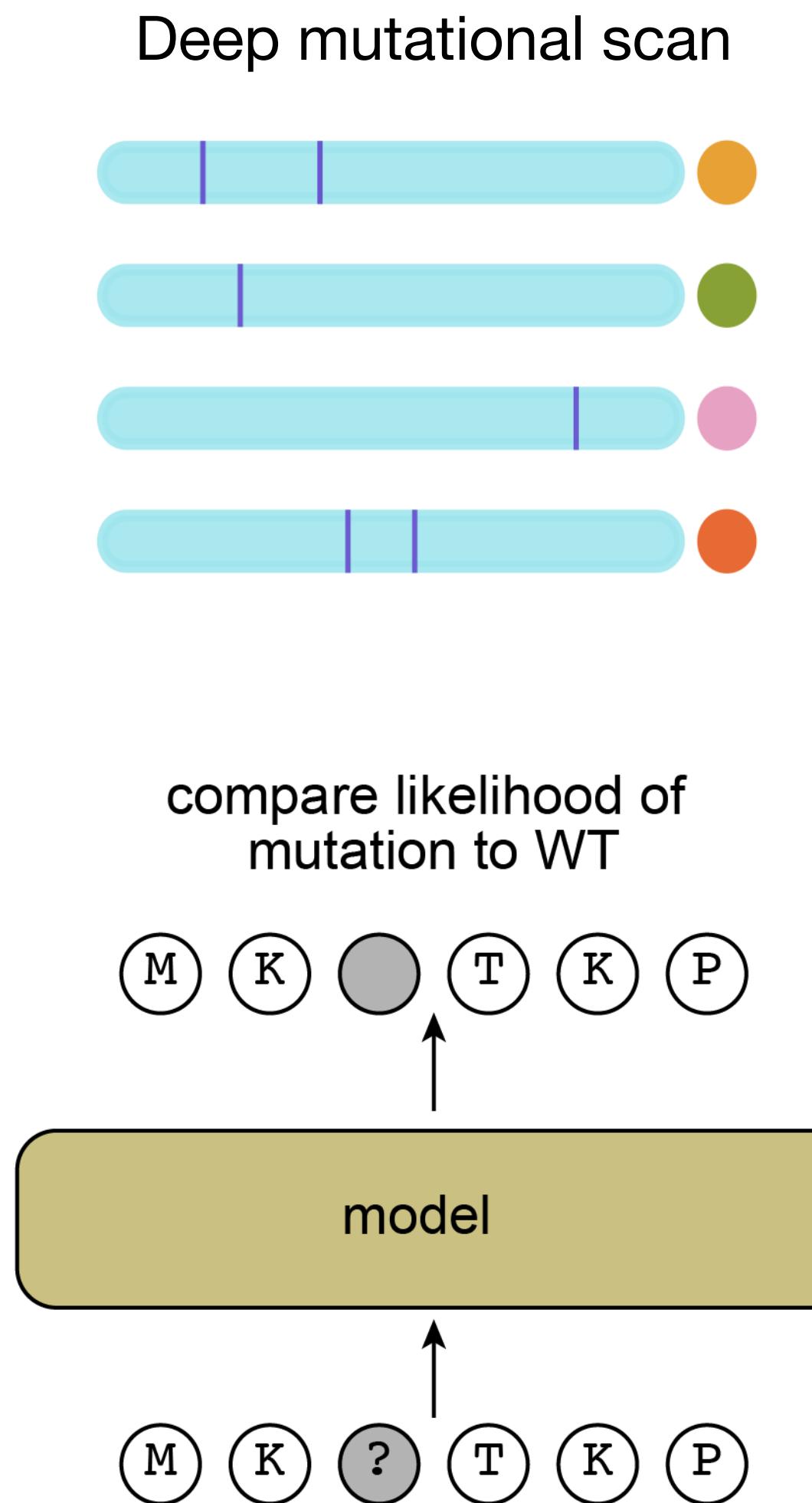
Deep mutational scan



compare likelihood of
mutation to WT



MIF, and MIF-ST are zero-shot fitness predictors



MIF, and MIF-ST are zero-shot fitness predictors

Deep mutational scan



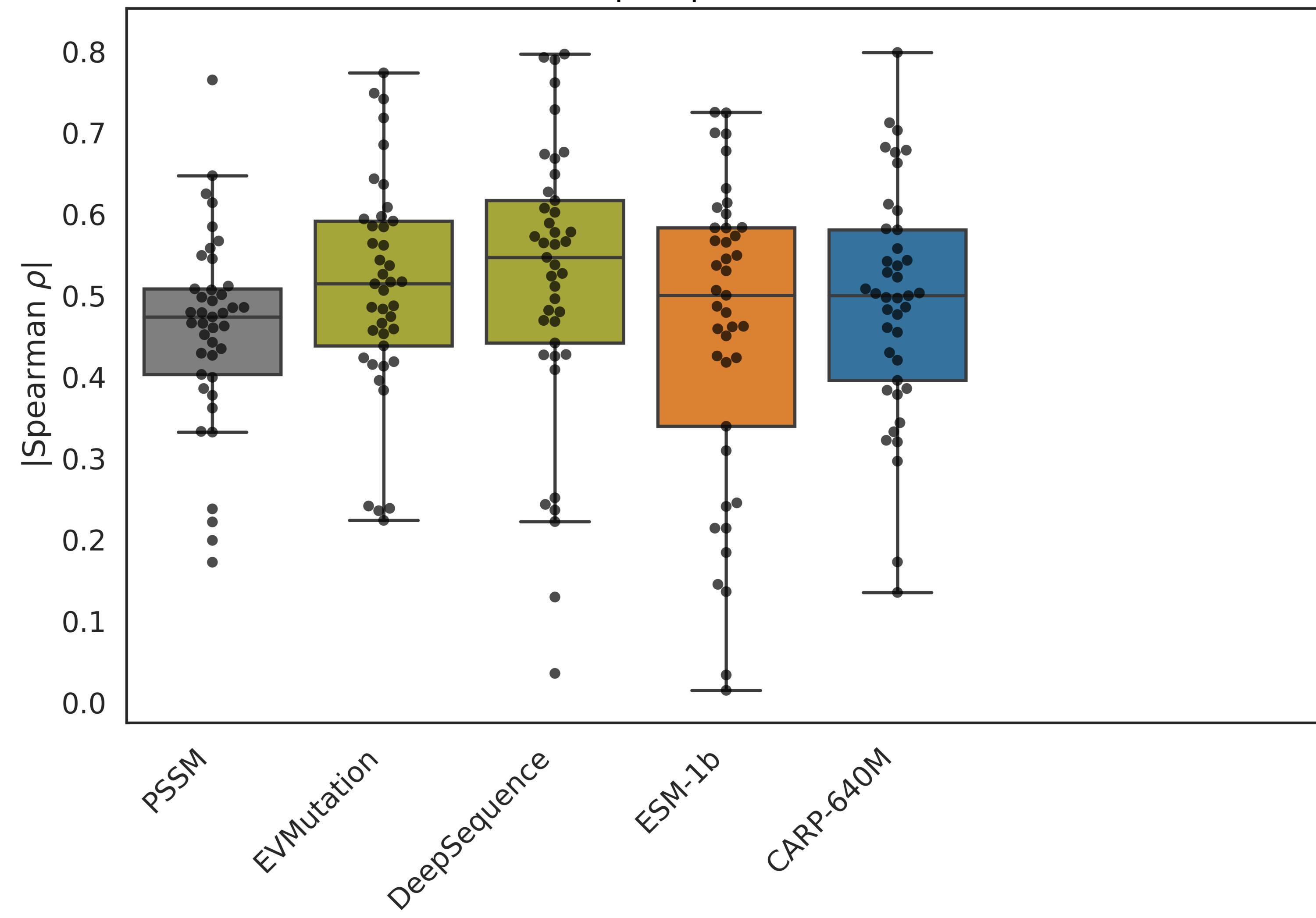
compare likelihood of
mutation to WT



model

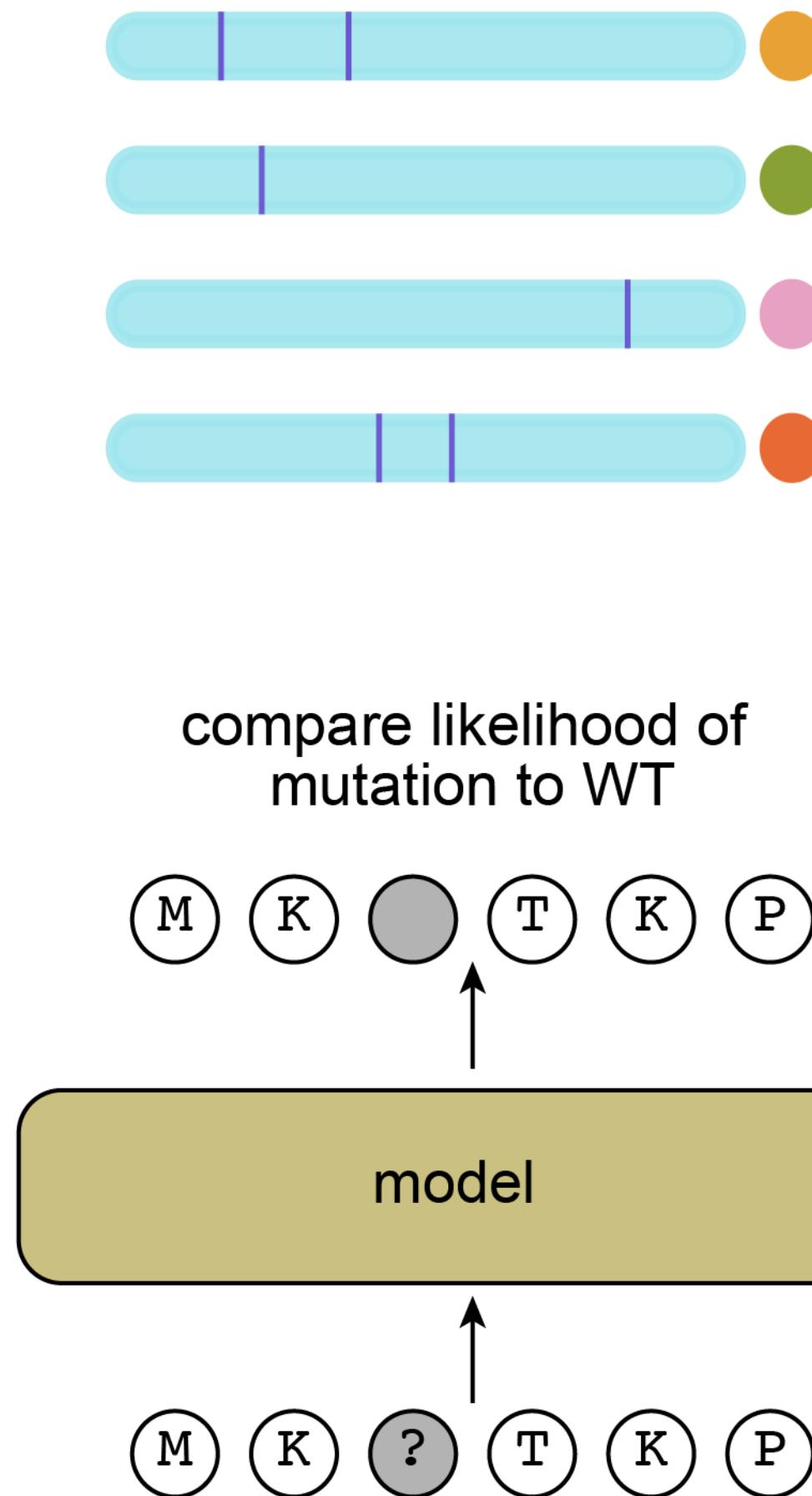


DeepSequence

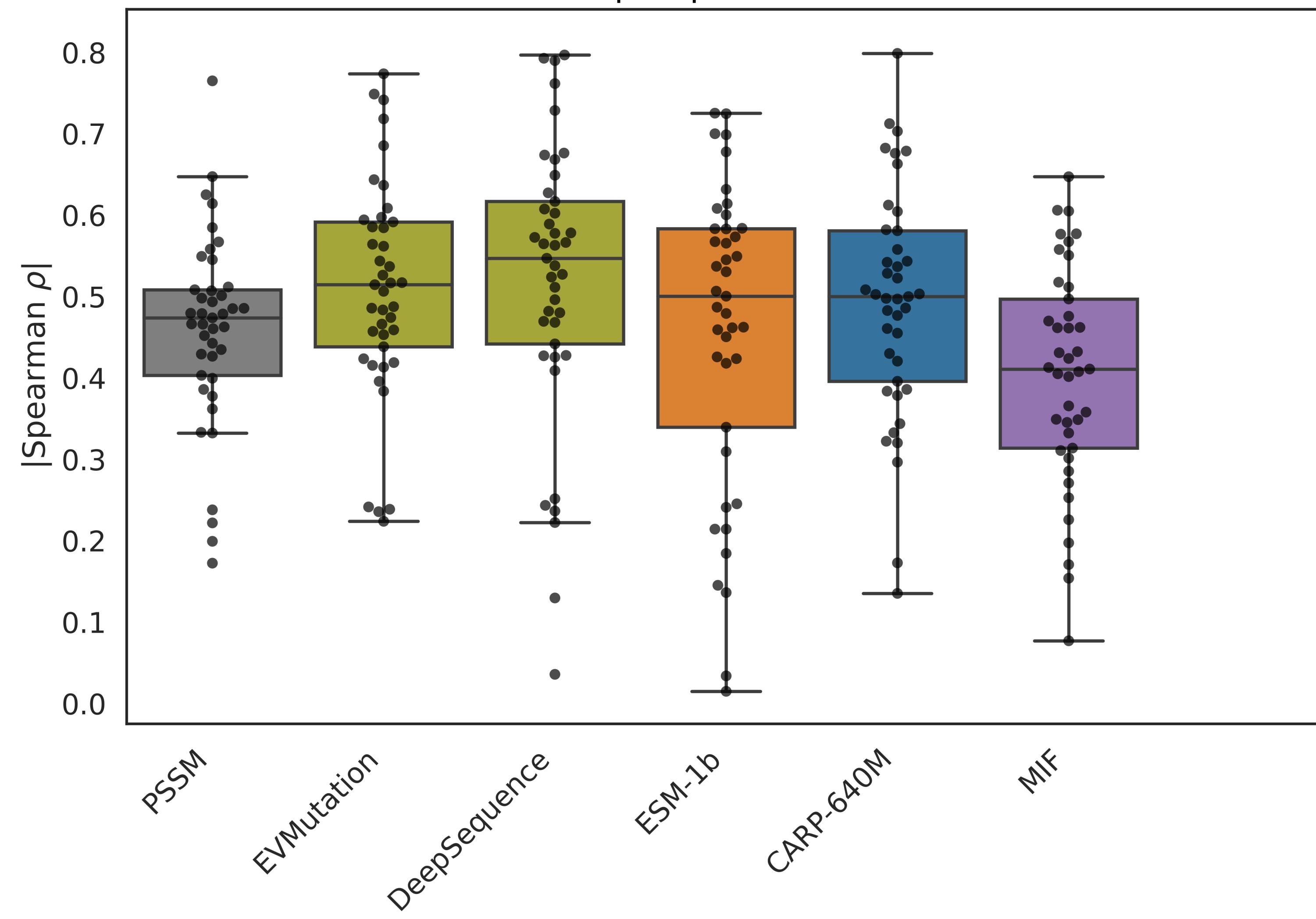


MIF, and MIF-ST are zero-shot fitness predictors

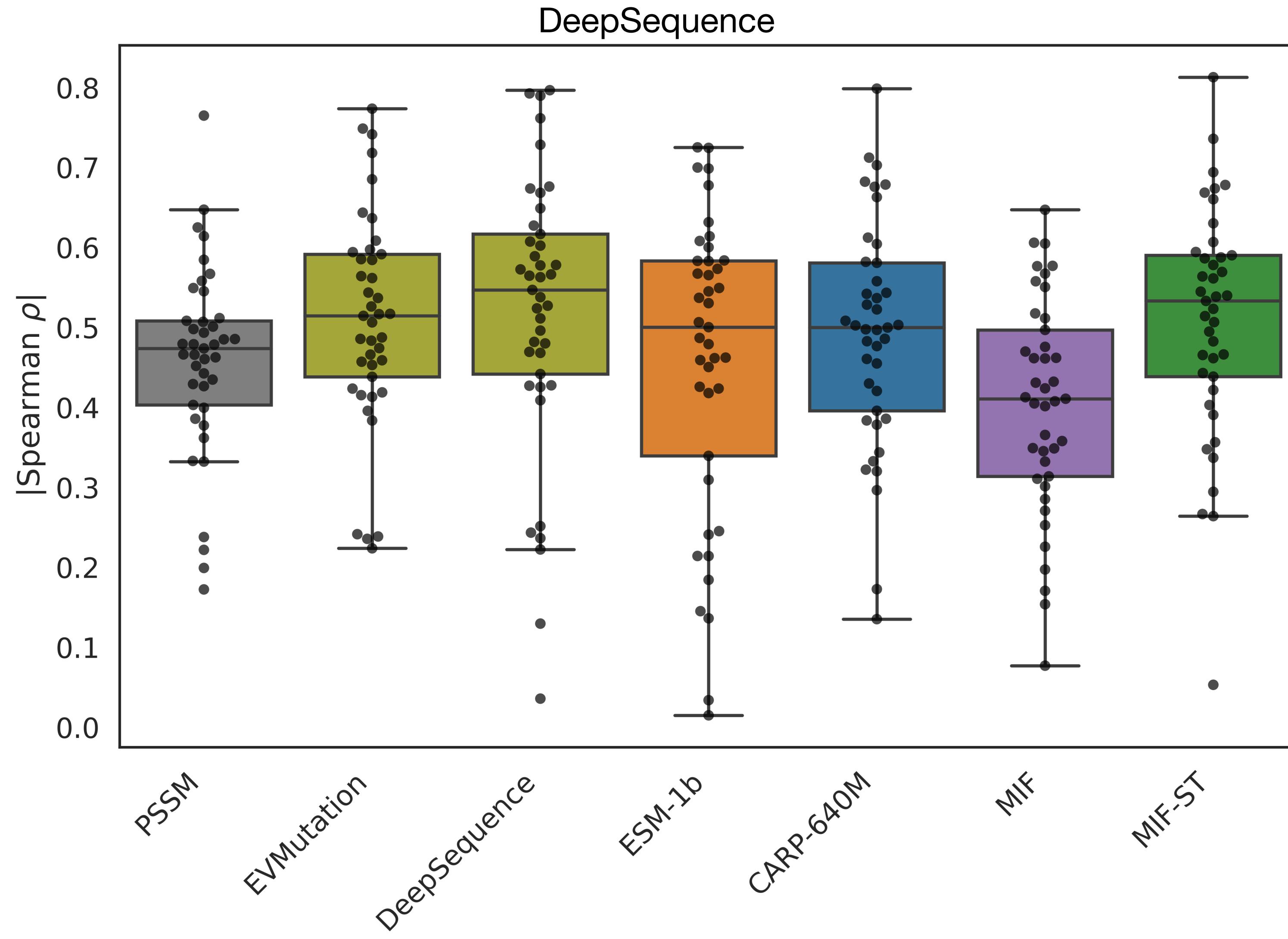
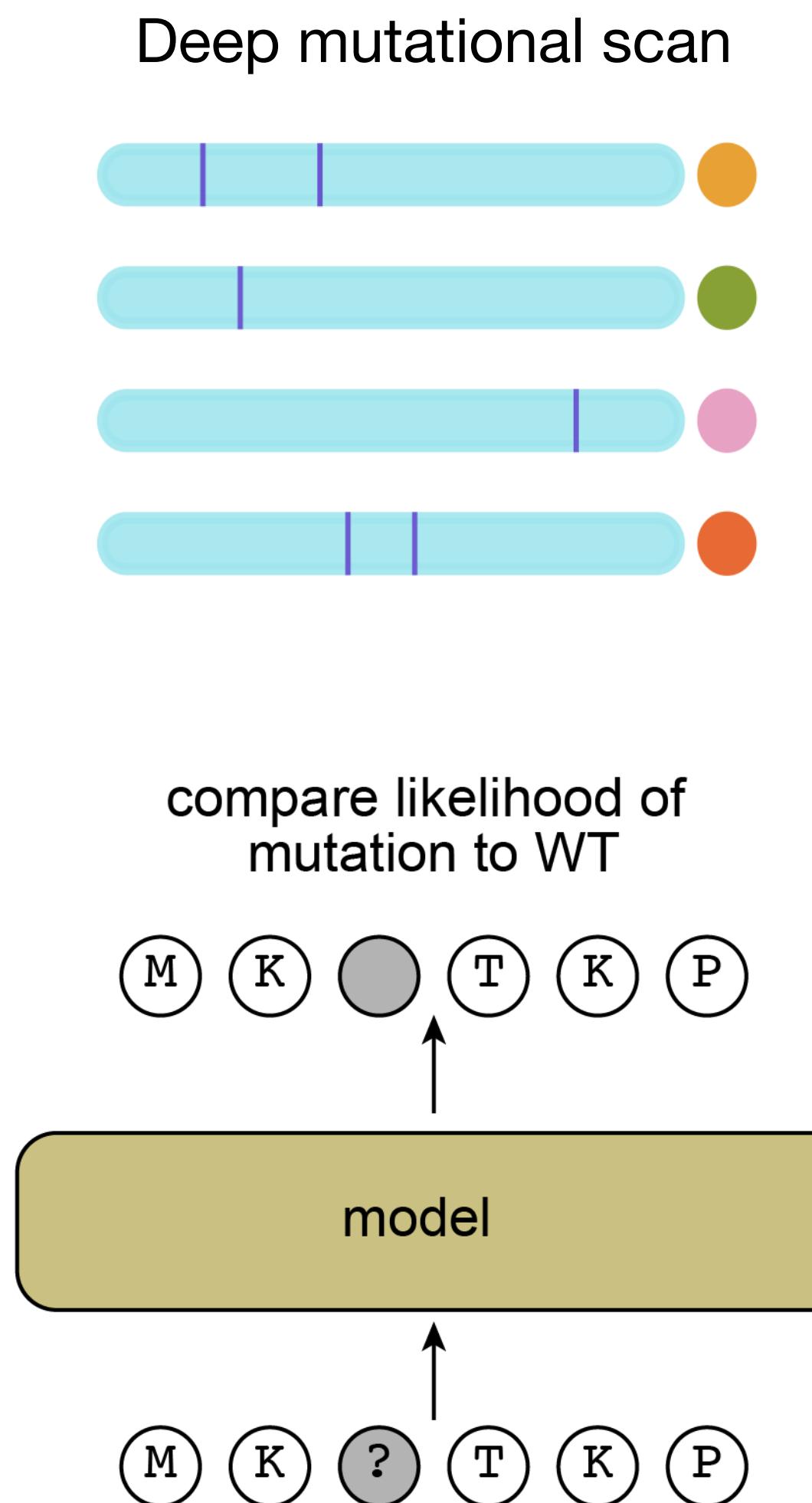
Deep mutational scan



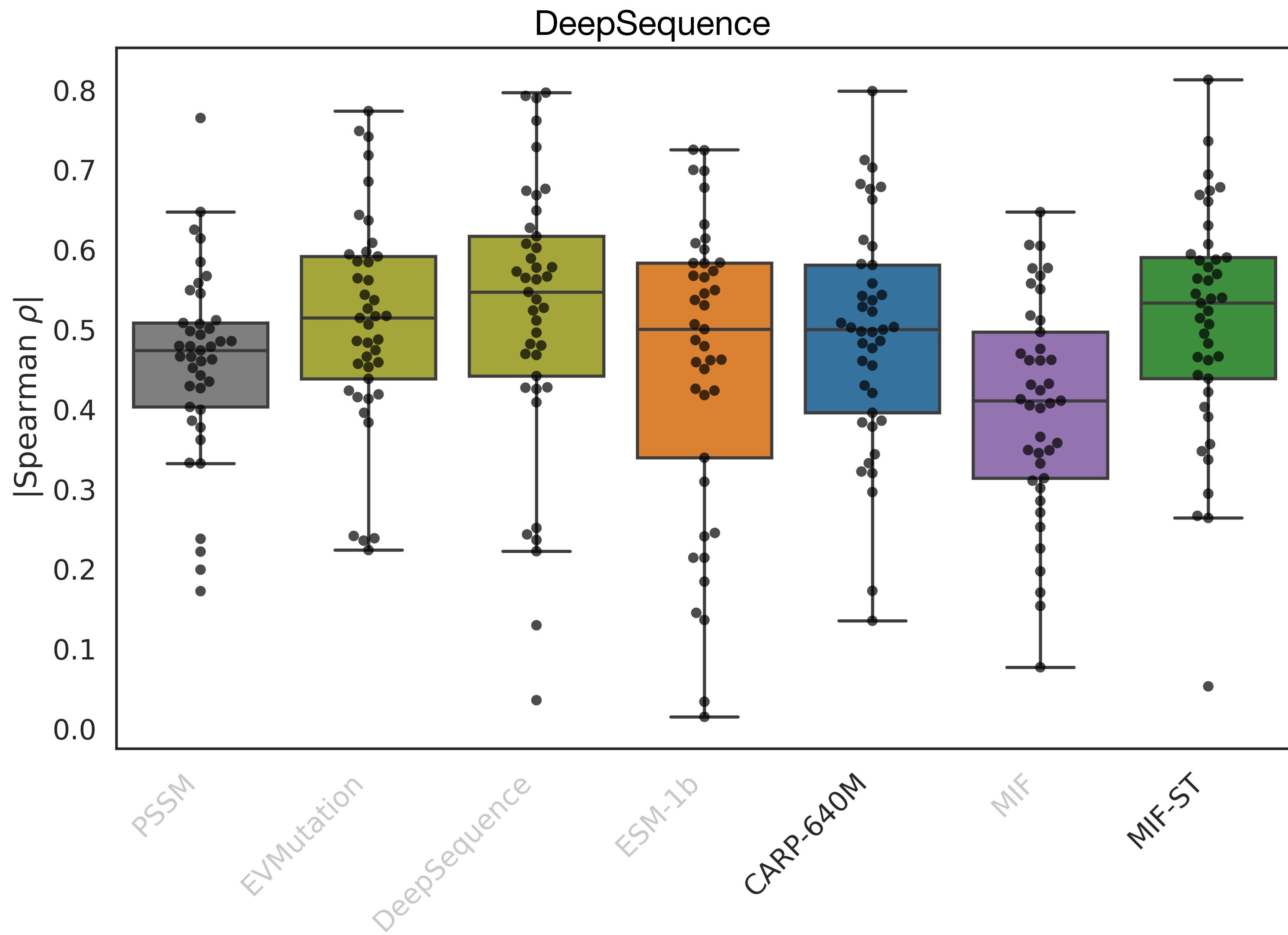
DeepSequence



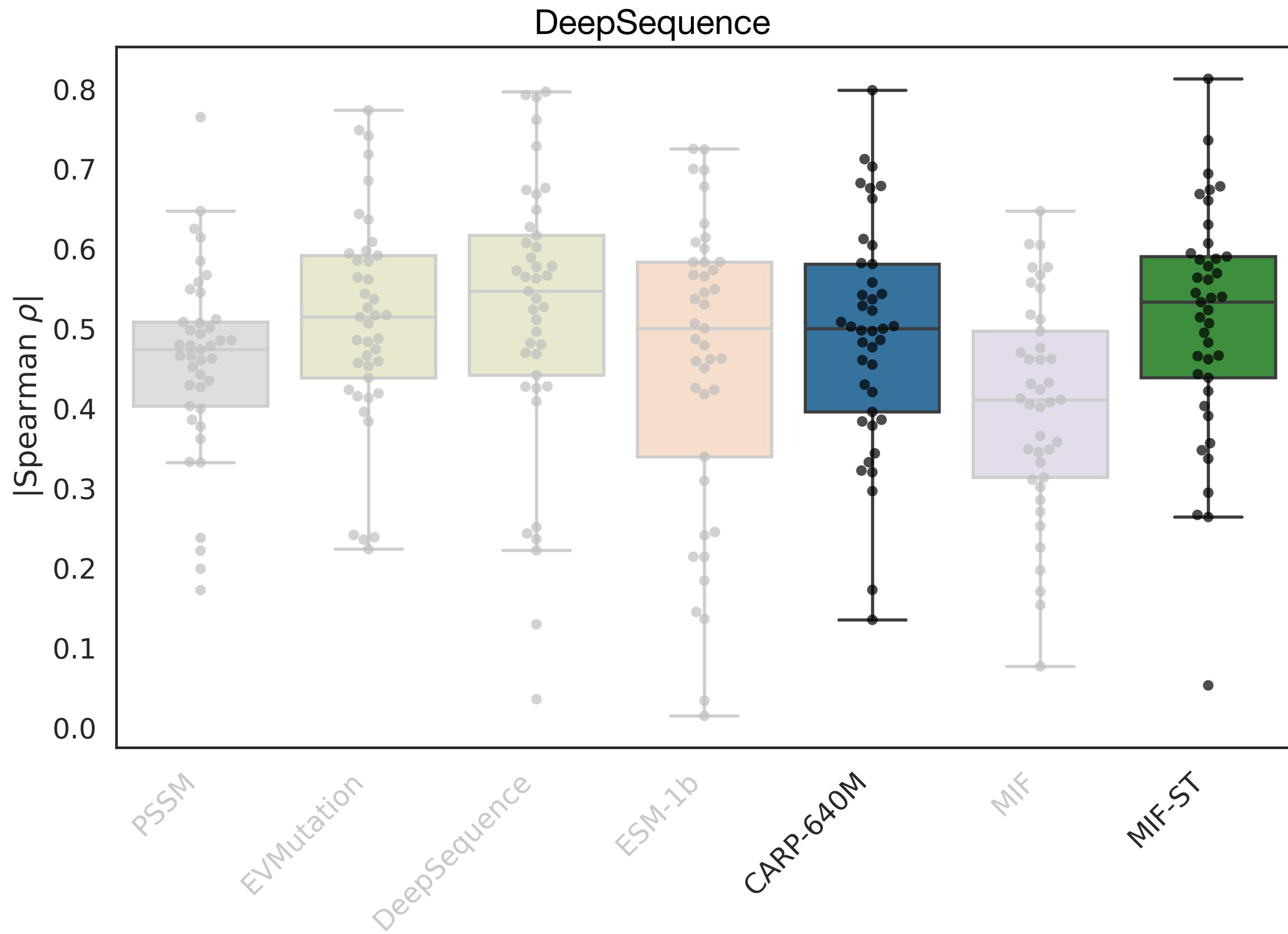
MIF, and MIF-ST are zero-shot fitness predictors



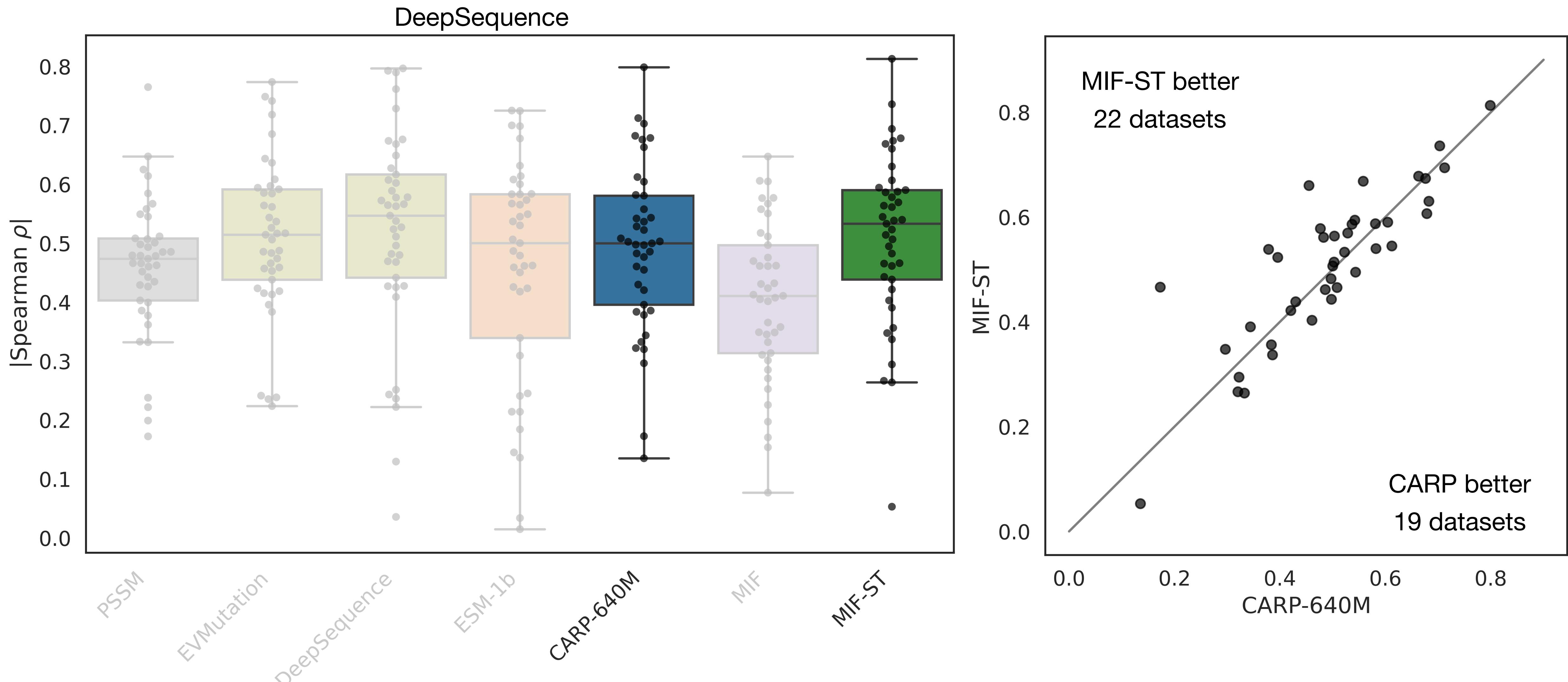
MIF-ST outperforms CARP and MIF on DeepSequence



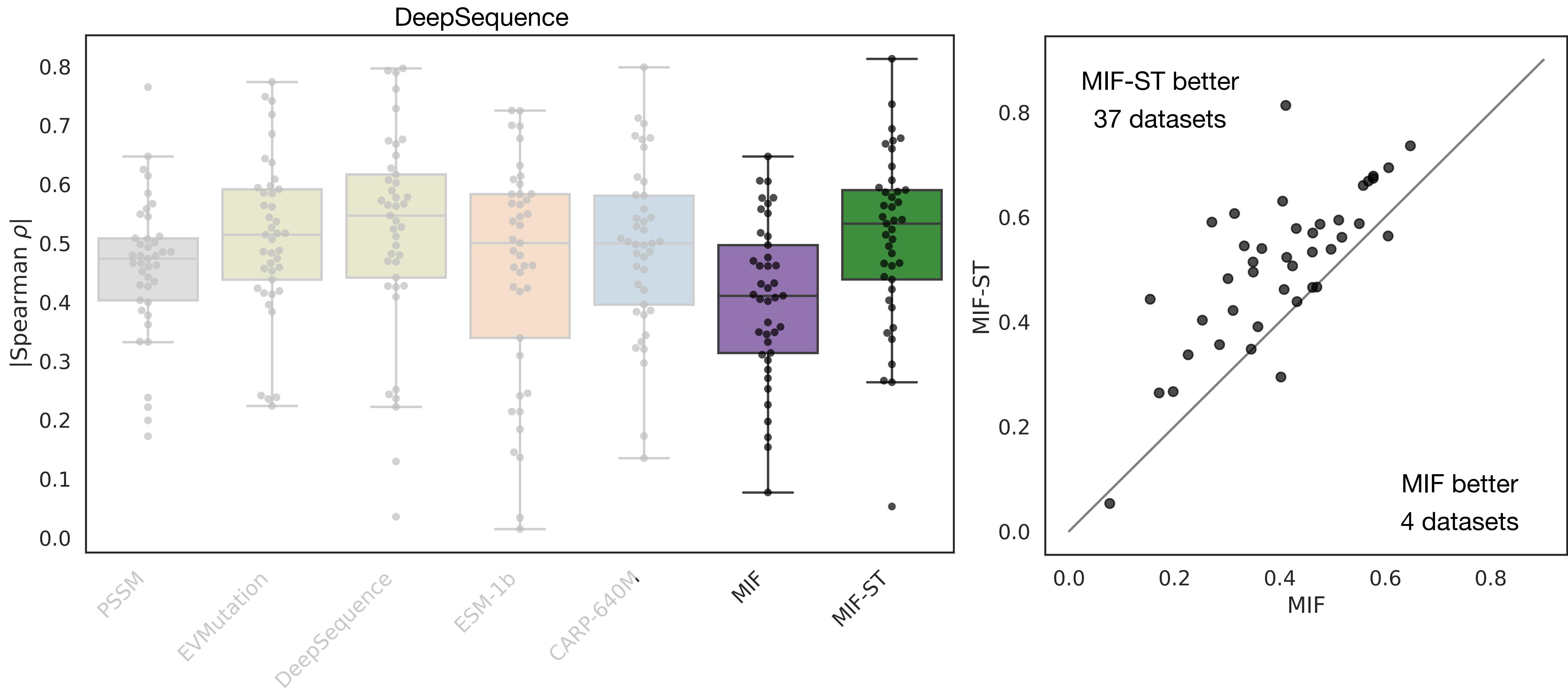
MIF-ST outperforms CARP and MIF on DeepSequence



MIF-ST outperforms CARP and MIF on DeepSequence

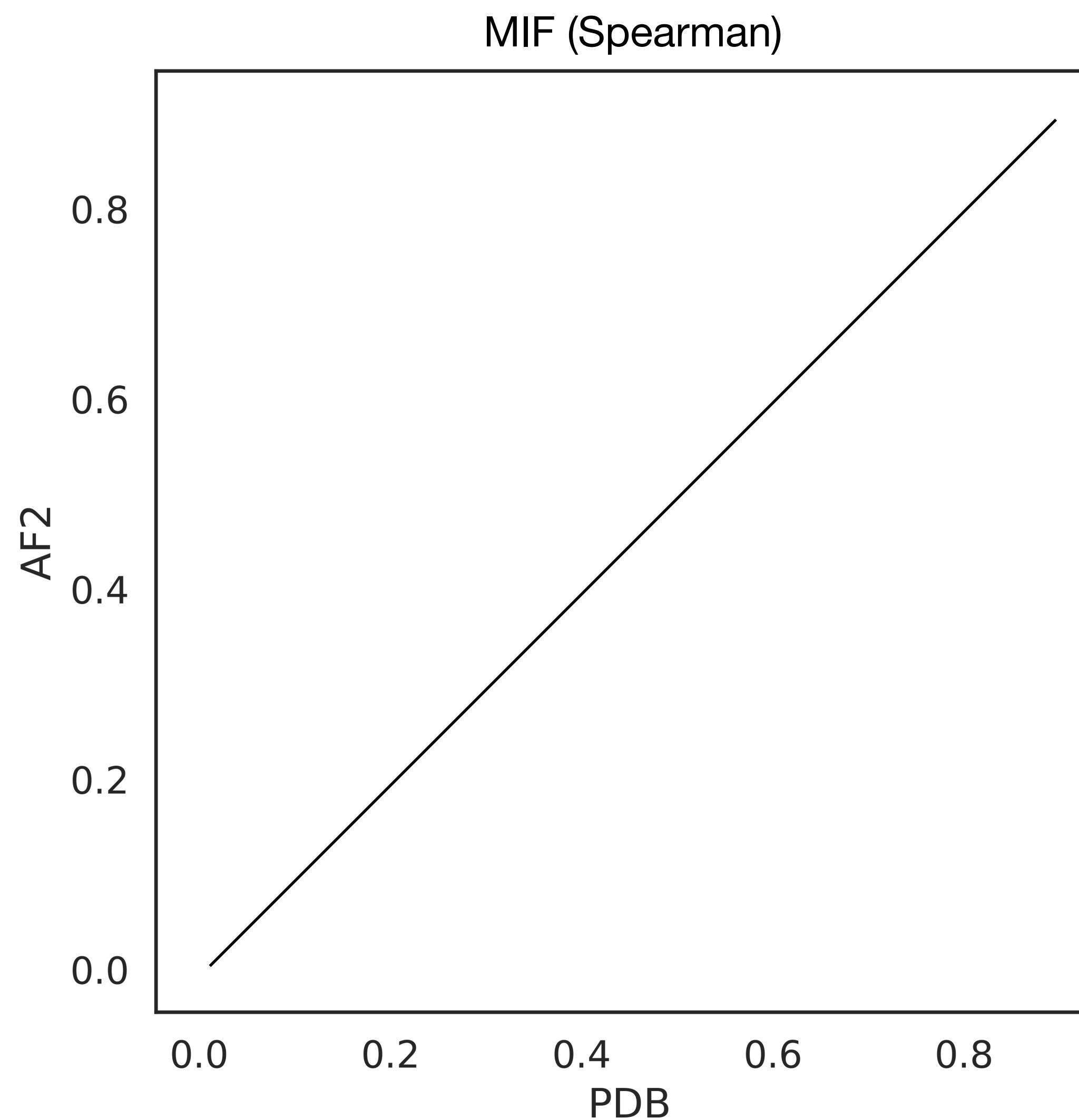


MIF-ST outperforms CARP and MIF on DeepSequence

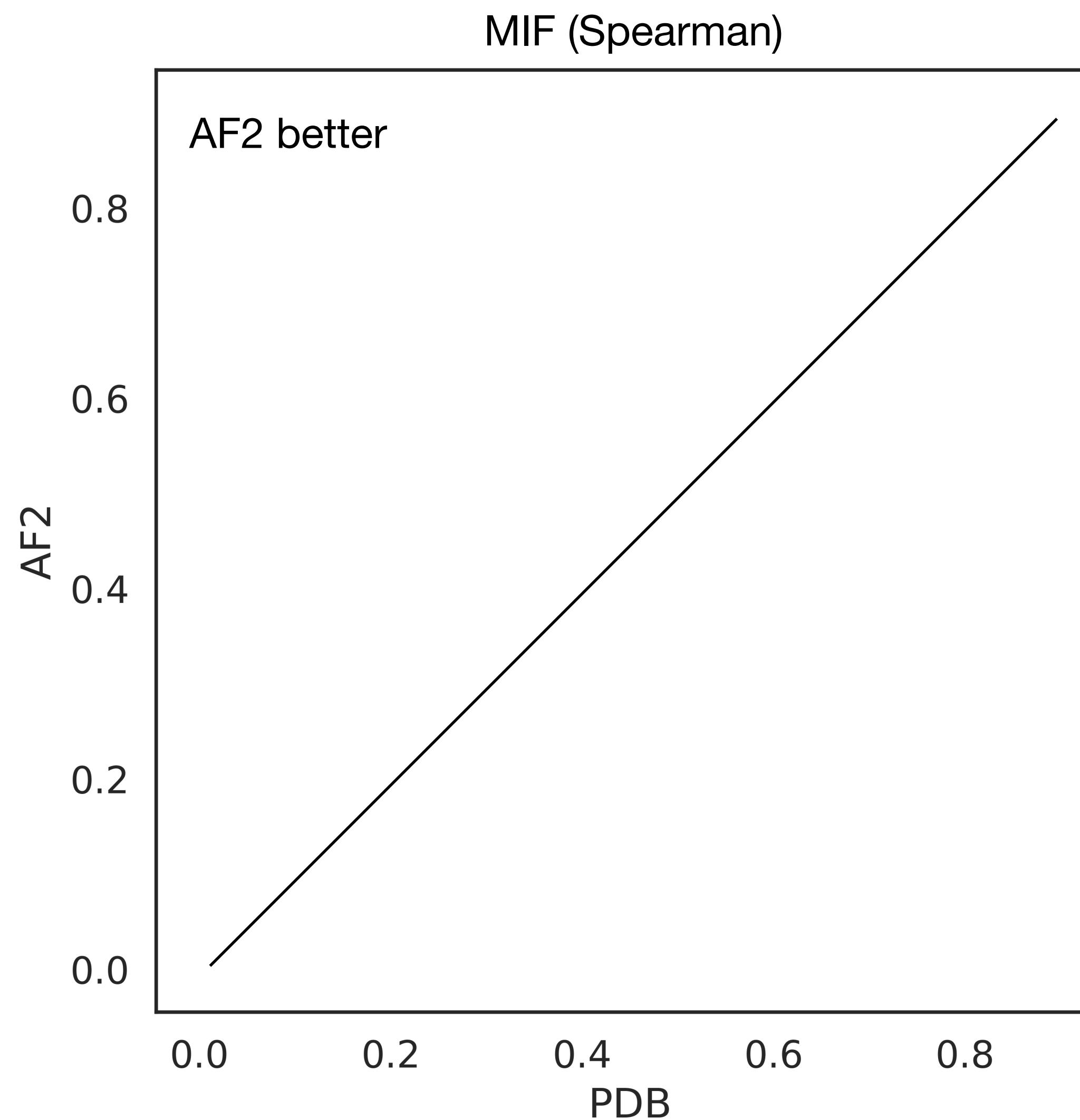


Predictions often better using AF2 structures

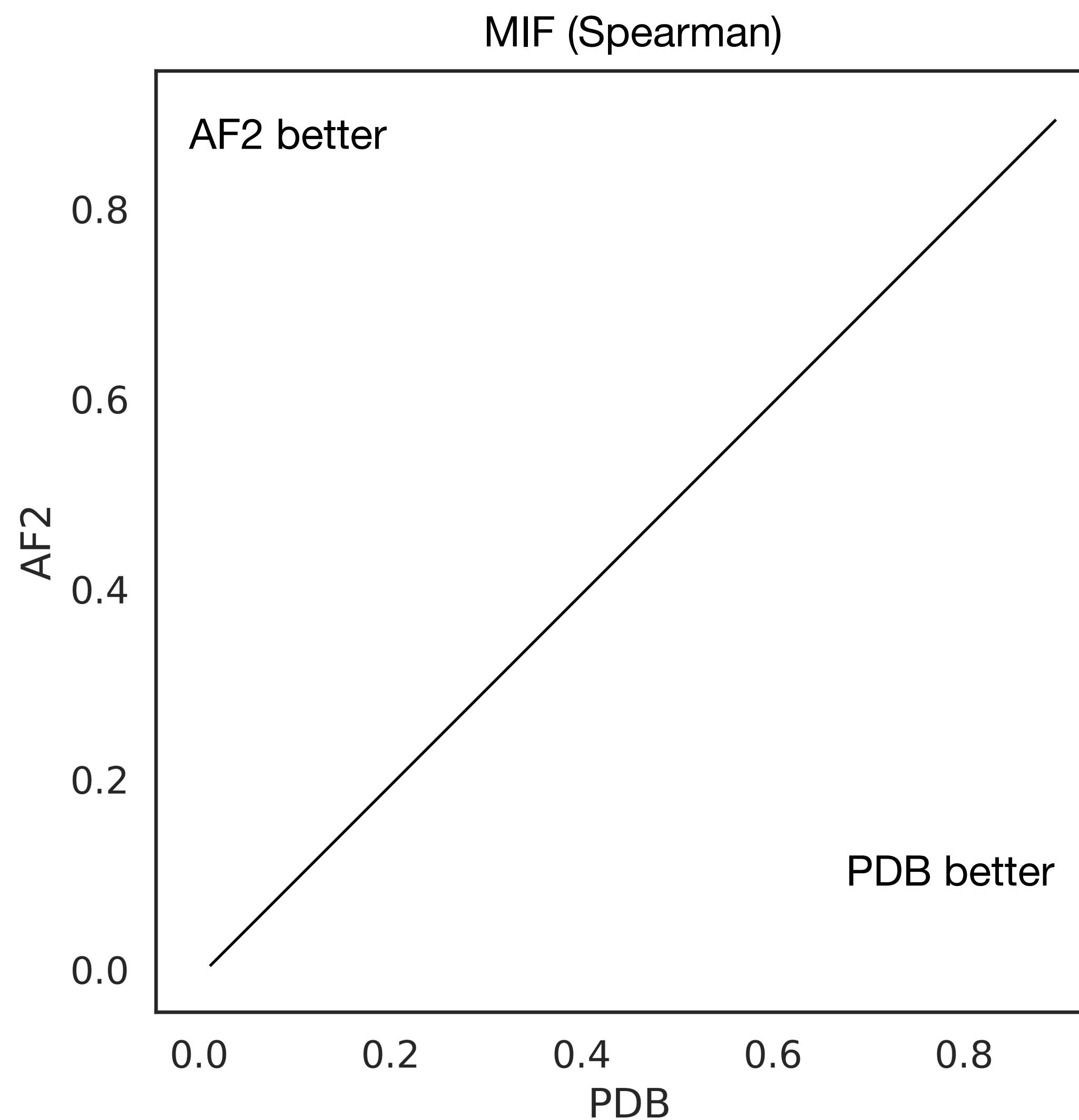
Predictions often better using AF2 structures



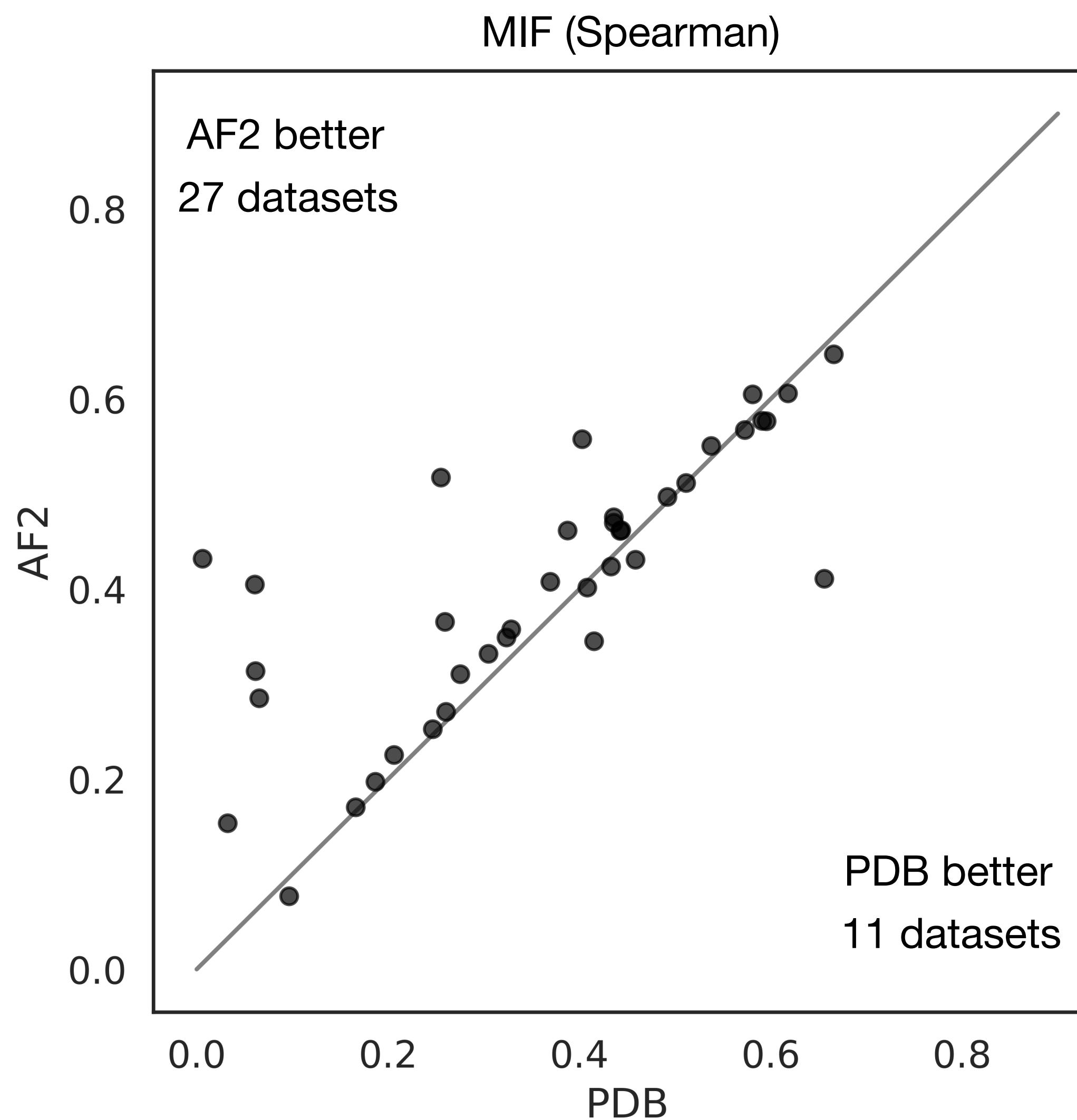
Predictions often better using AF2 structures



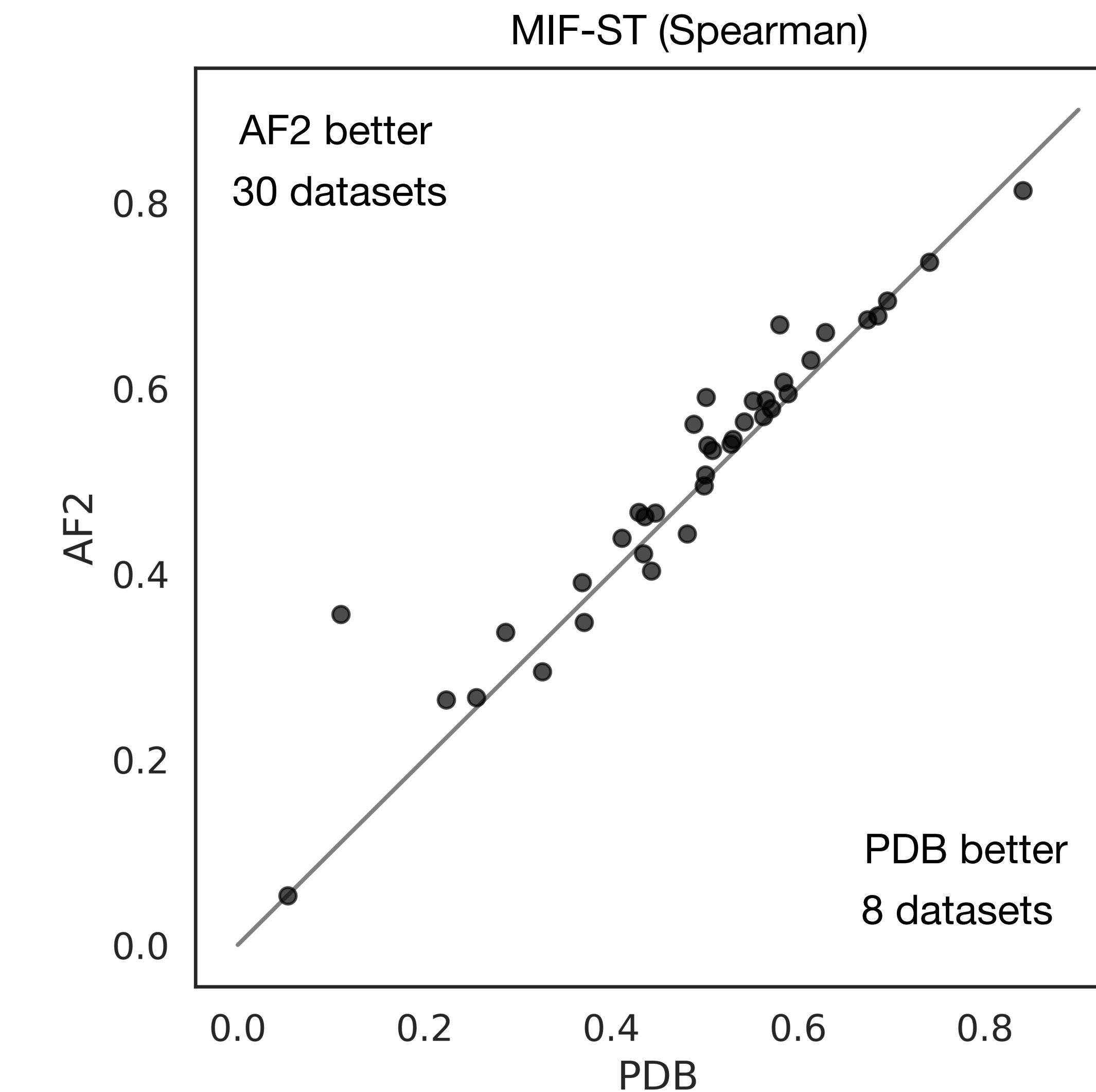
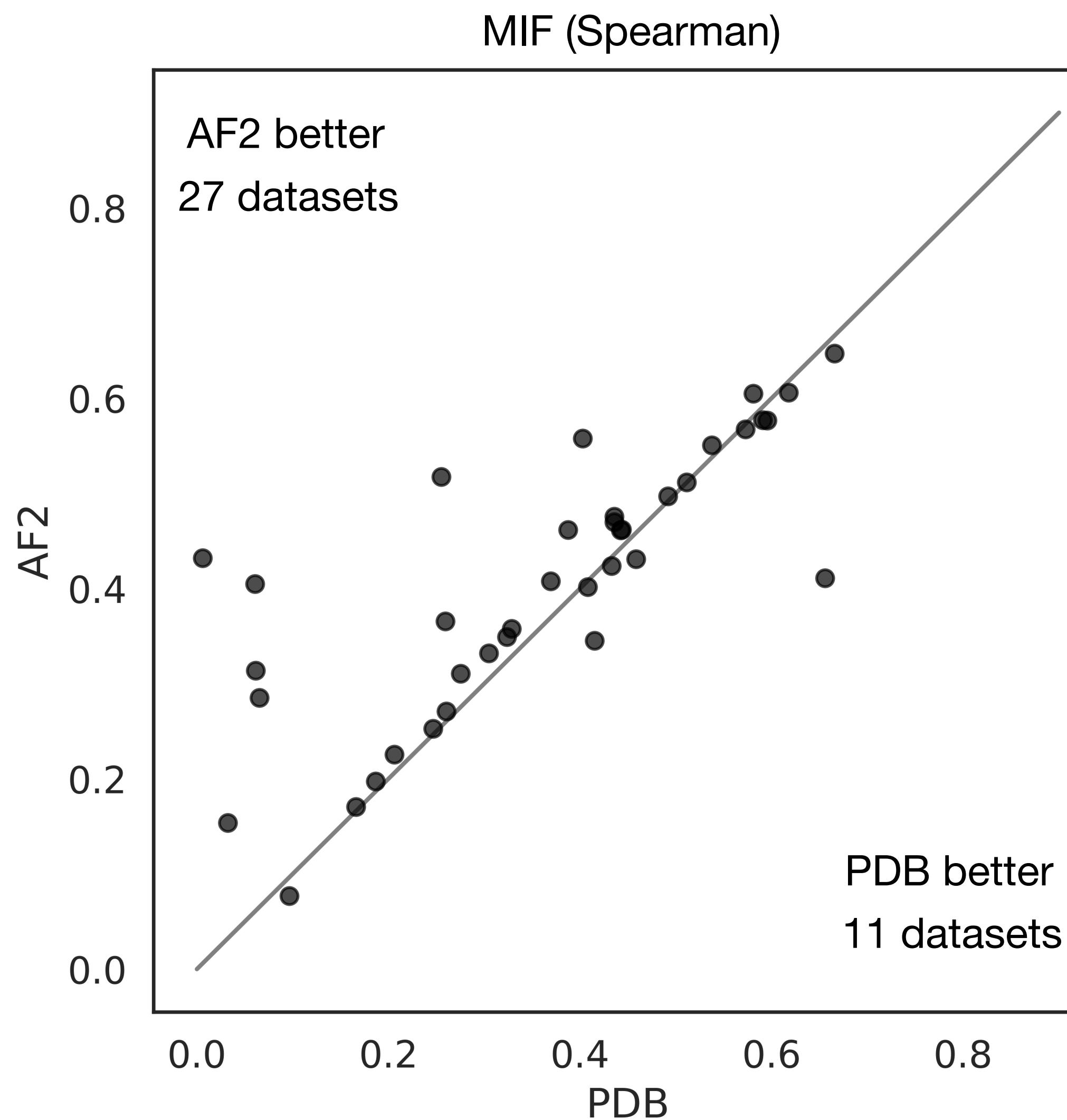
Predictions often better using AF2 structures



Predictions often better using AF2 structures



Predictions often better using AF2 structures



Structure conditioning improves zero-shot predictions on artificial proteins

Structure conditioning improves zero-shot predictions on artificial proteins

Model

Rocklin

Structure conditioning improves zero-shot predictions on artificial proteins

	Model	Rocklin
sequence	CARP-640M	0.28

Structure conditioning improves zero-shot predictions on artificial proteins

	Model	Rocklin
sequence	CARP-640M	0.28
+structure	MIF	0.45

Structure conditioning improves zero-shot predictions on artificial proteins

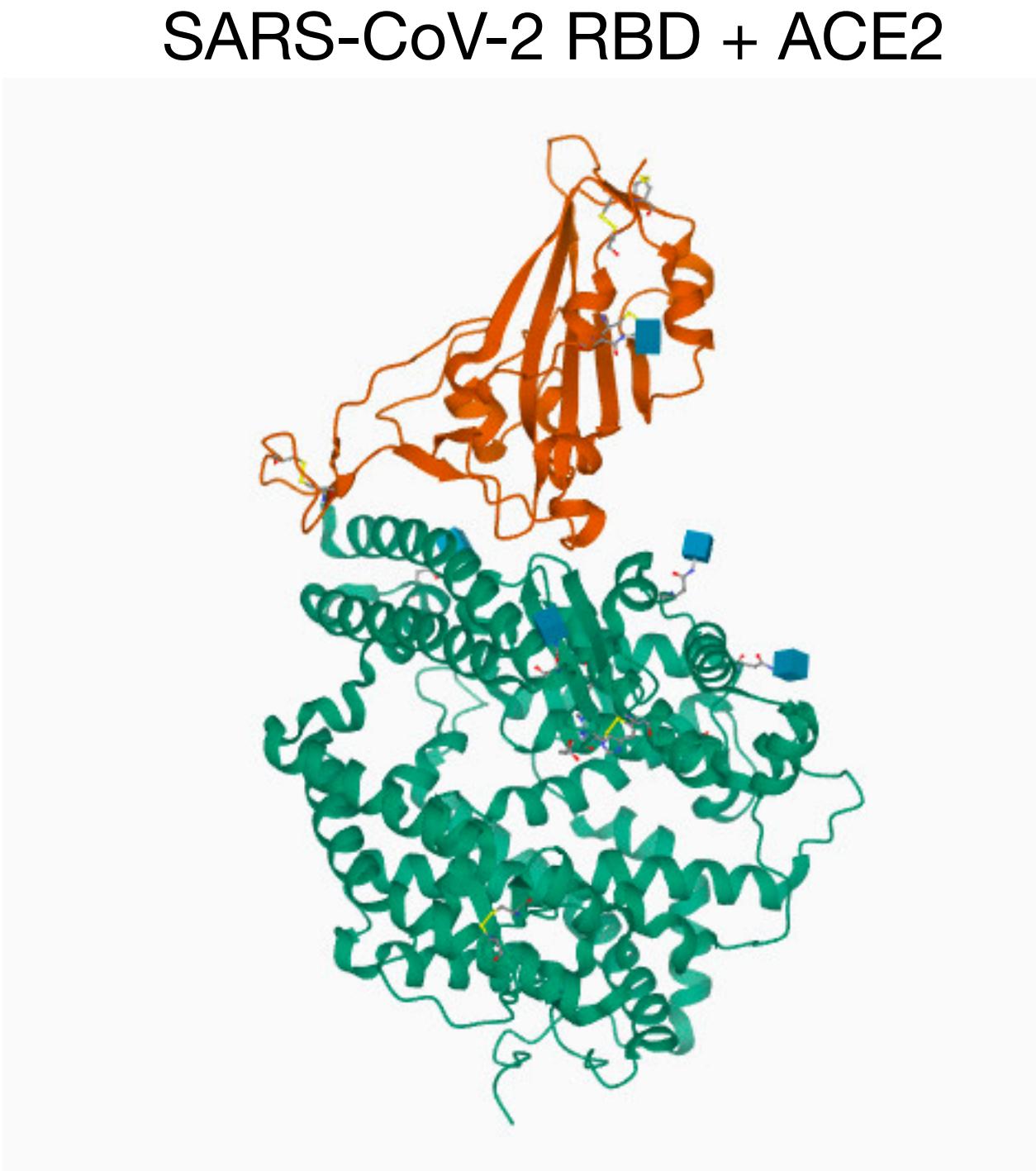
	Model	Rocklin
sequence	CARP-640M	0.28
+structure	MIF	0.45
+sequence transfer	MIF-ST	0.47

Structure conditioning improves zero-shot predictions on artificial proteins

	Model	Rocklin
sequence	CARP-640M	0.28
+structure	MIF	0.45
+sequence transfer	MIF-ST	0.47
+augmentation	GVP+AF2	0.48

Structure conditioning improves zero-shot predictions on RBD binding

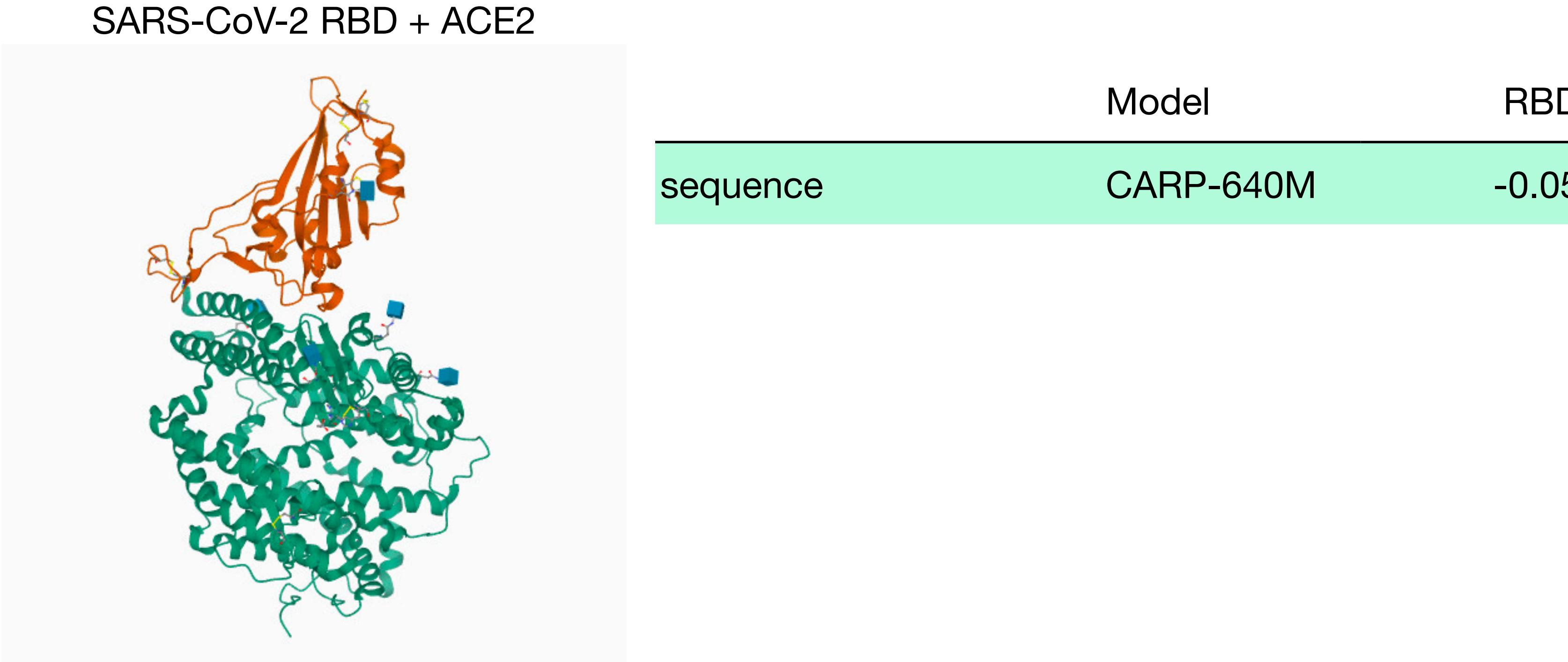
Structure conditioning improves zero-shot predictions on RBD binding



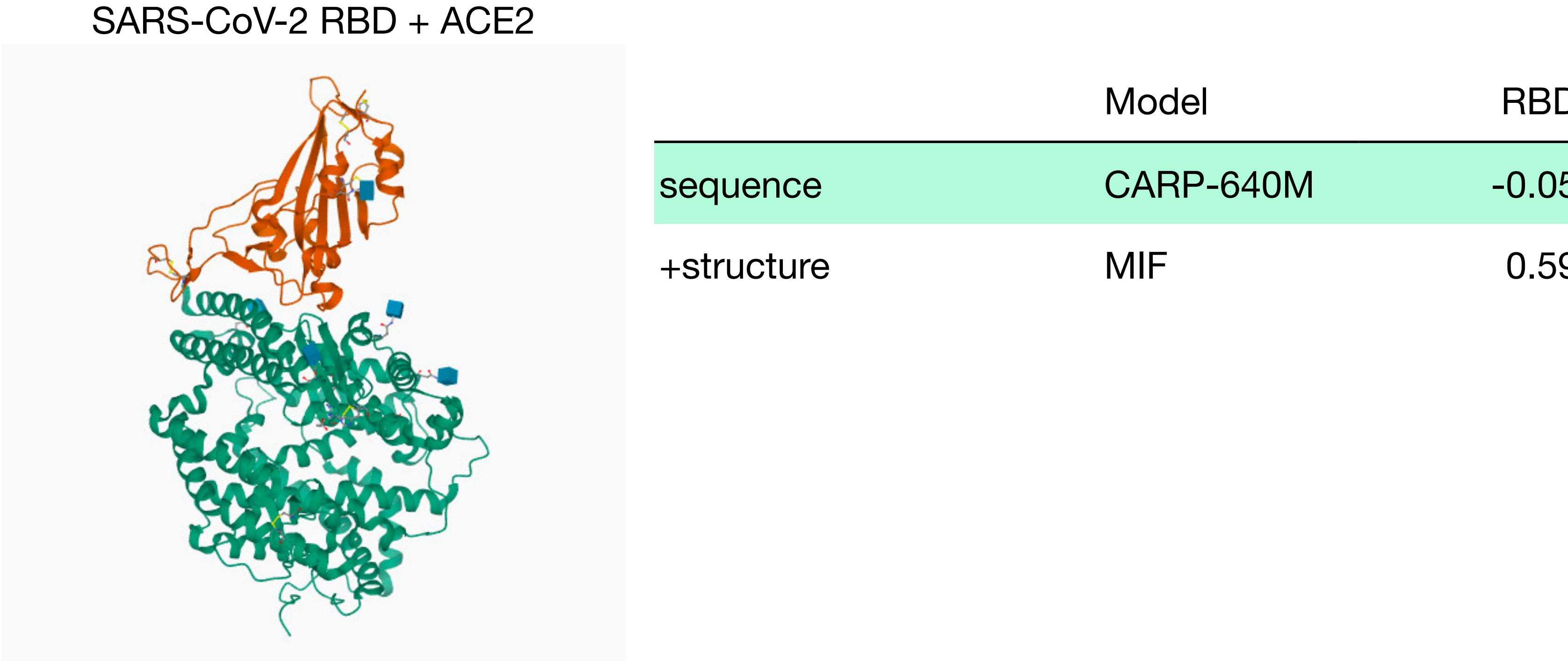
Structure conditioning improves zero-shot predictions on RBD binding



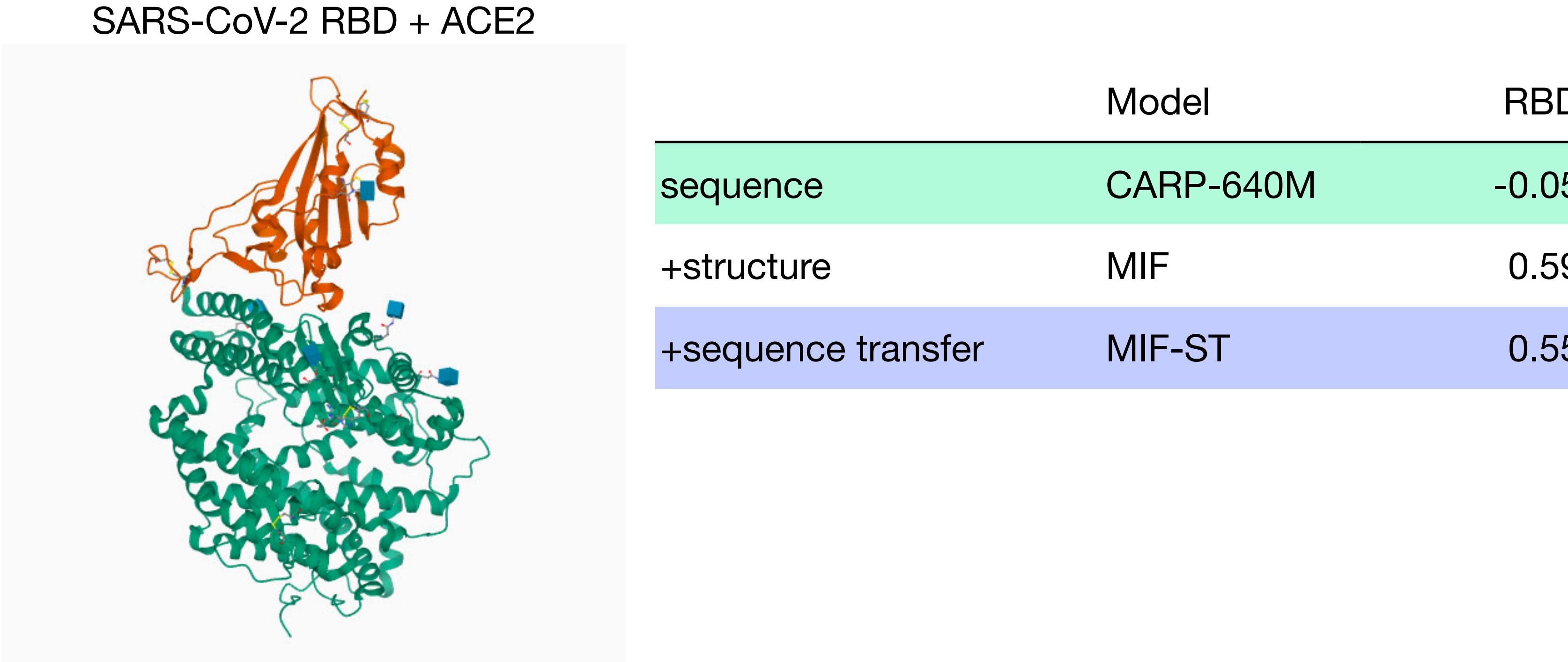
Structure conditioning improves zero-shot predictions on RBD binding



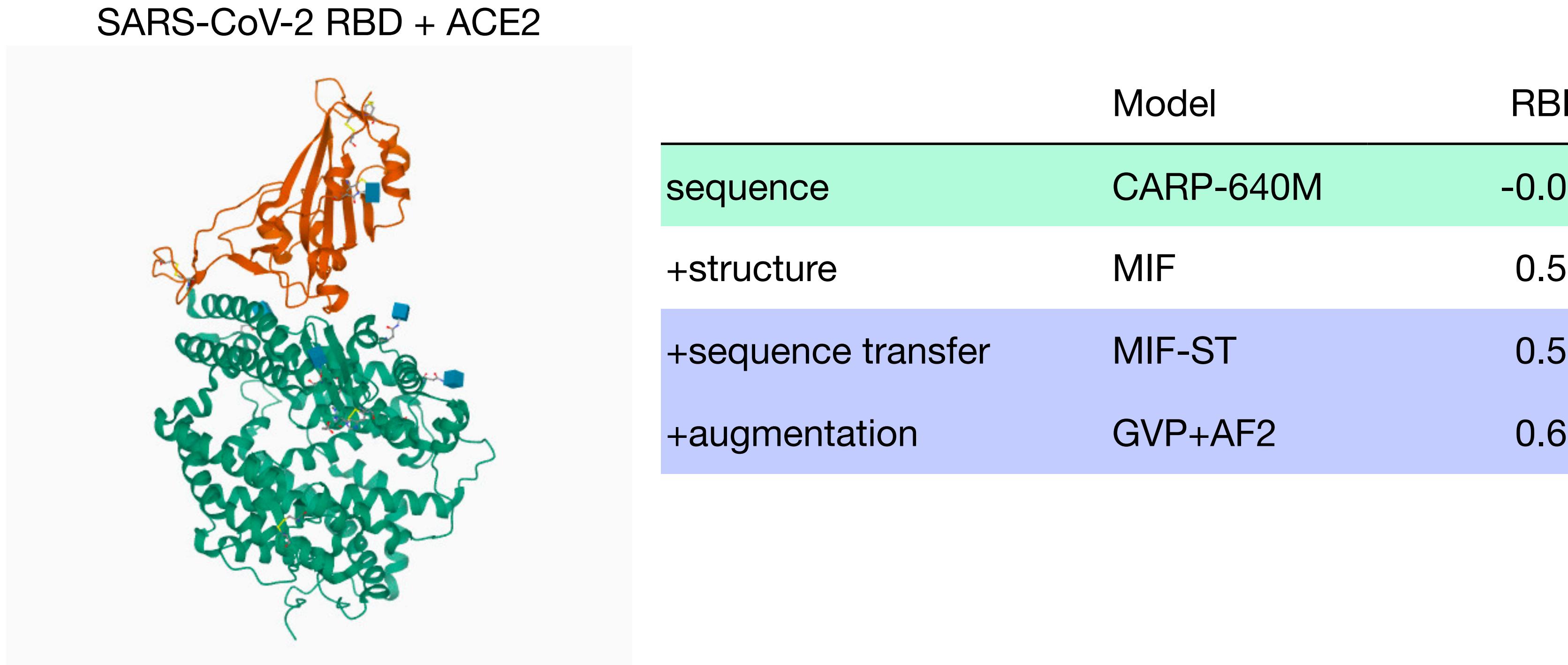
Structure conditioning improves zero-shot predictions on RBD binding



Structure conditioning improves zero-shot predictions on RBD binding

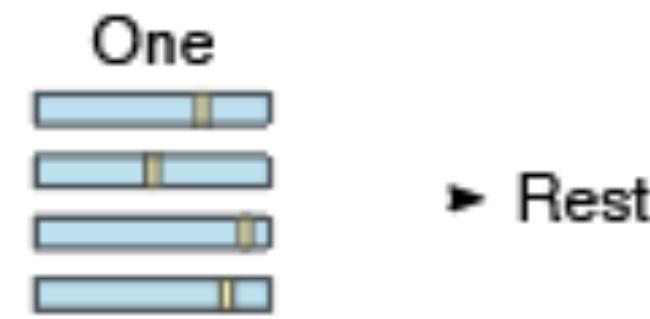


Structure conditioning improves zero-shot predictions on RBD binding

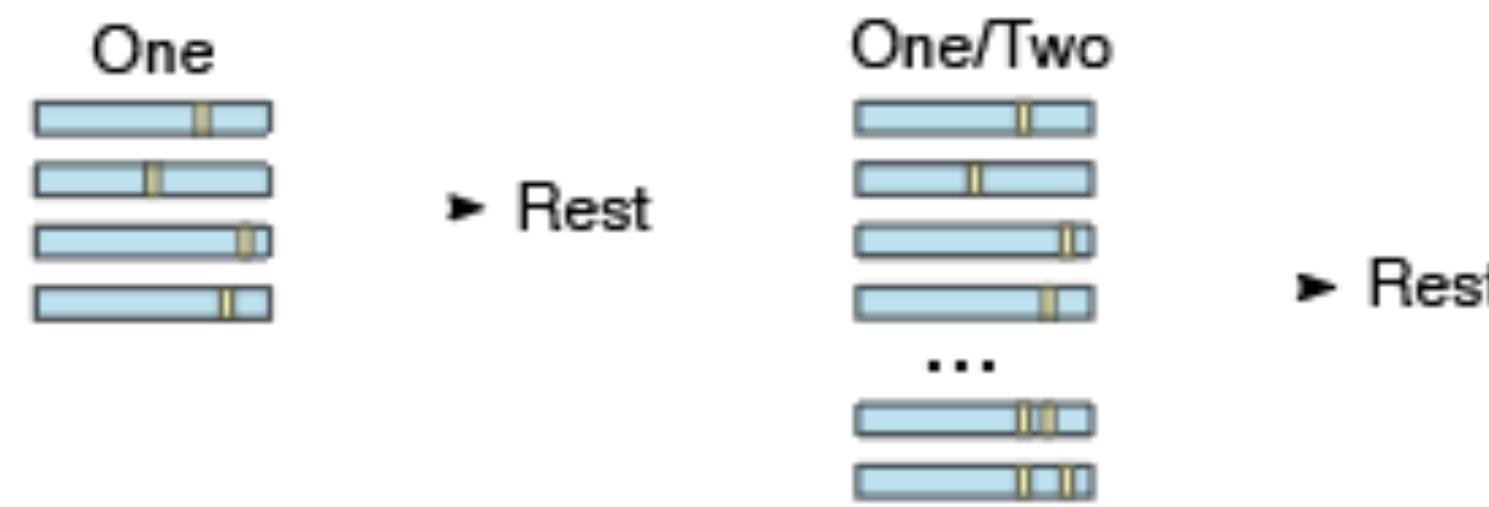


Protein engineering campaigns require out-of-domain generalization

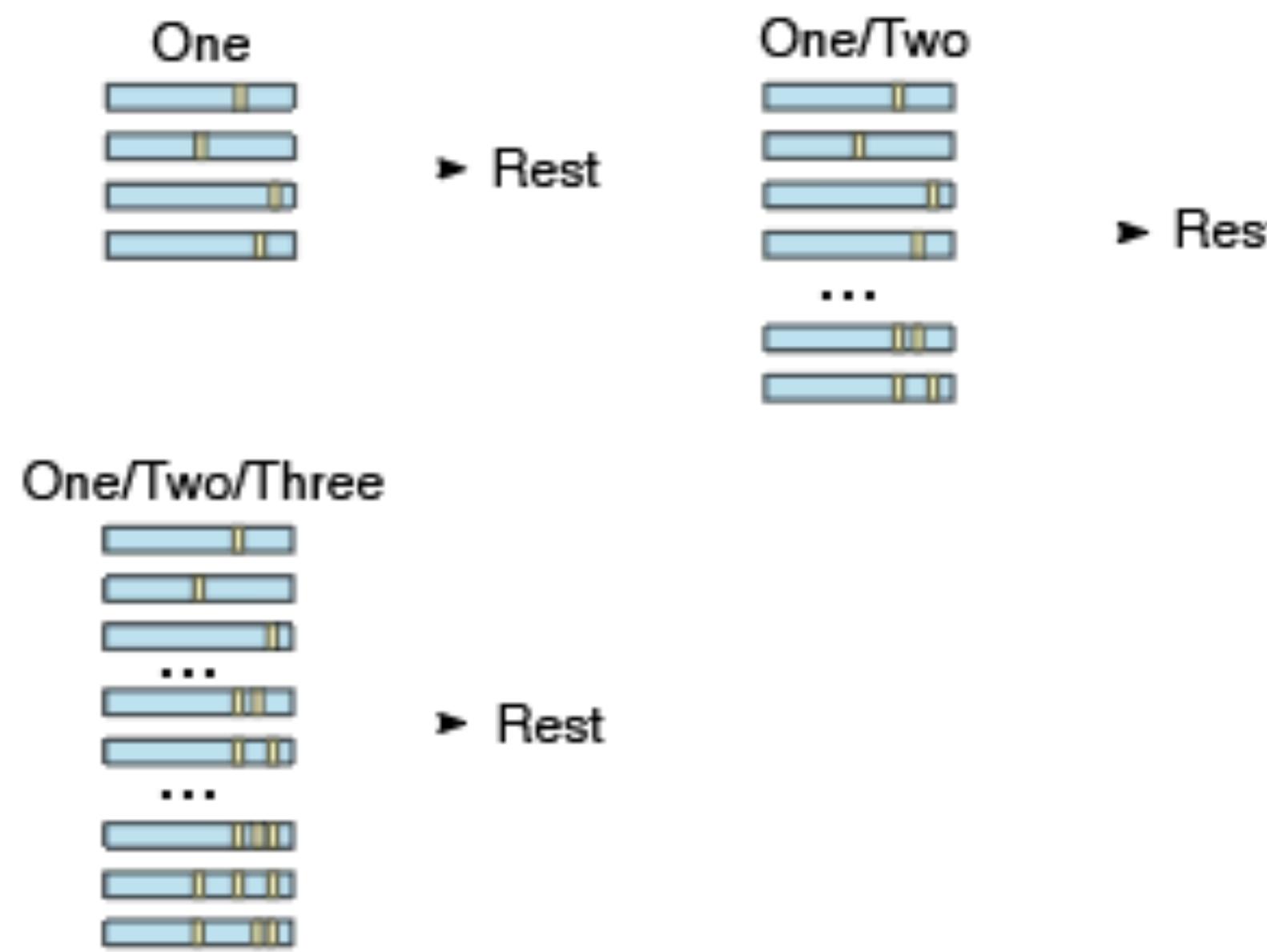
Protein engineering campaigns require out-of-domain generalization



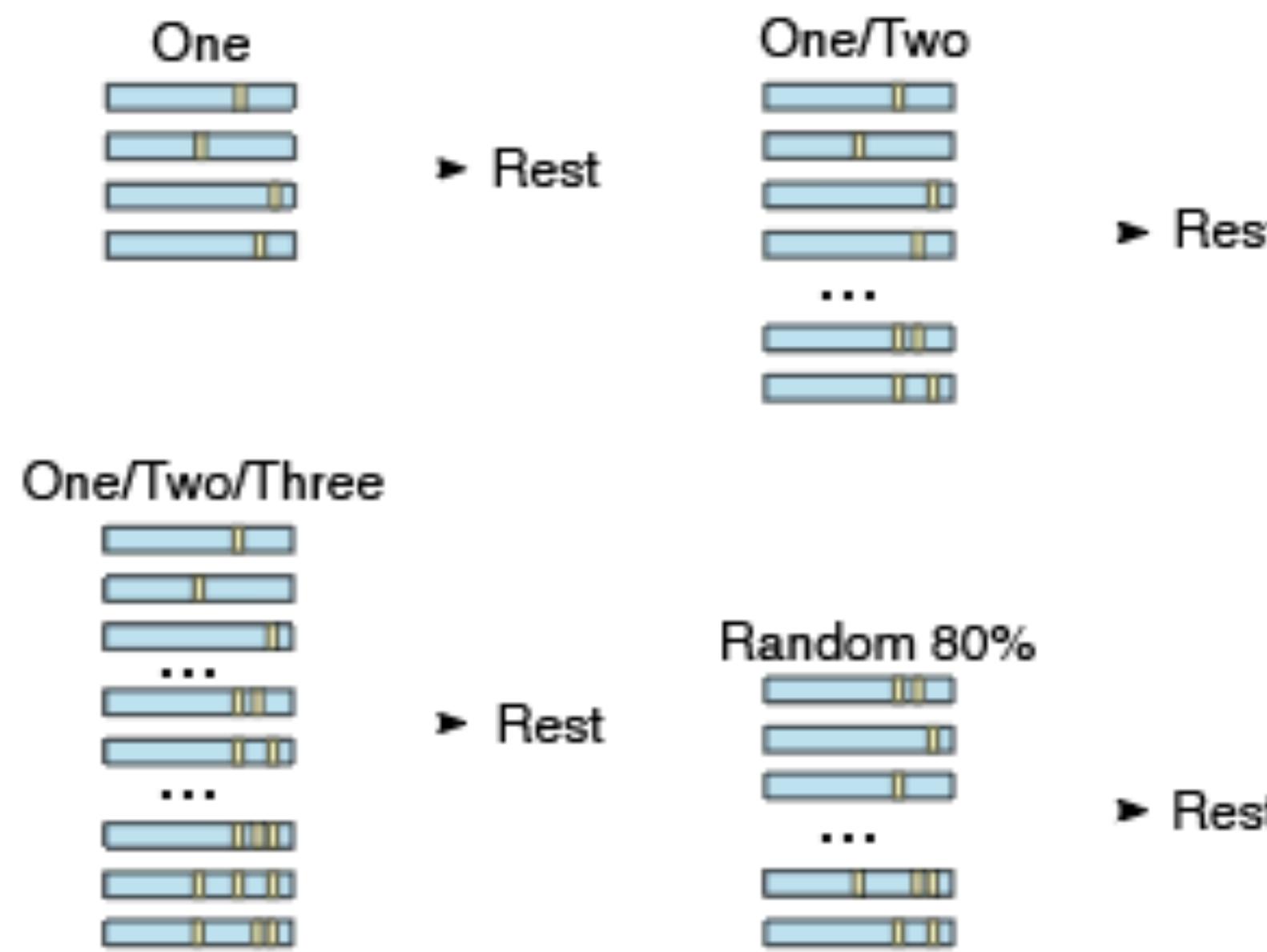
Protein engineering campaigns require out-of-domain generalization



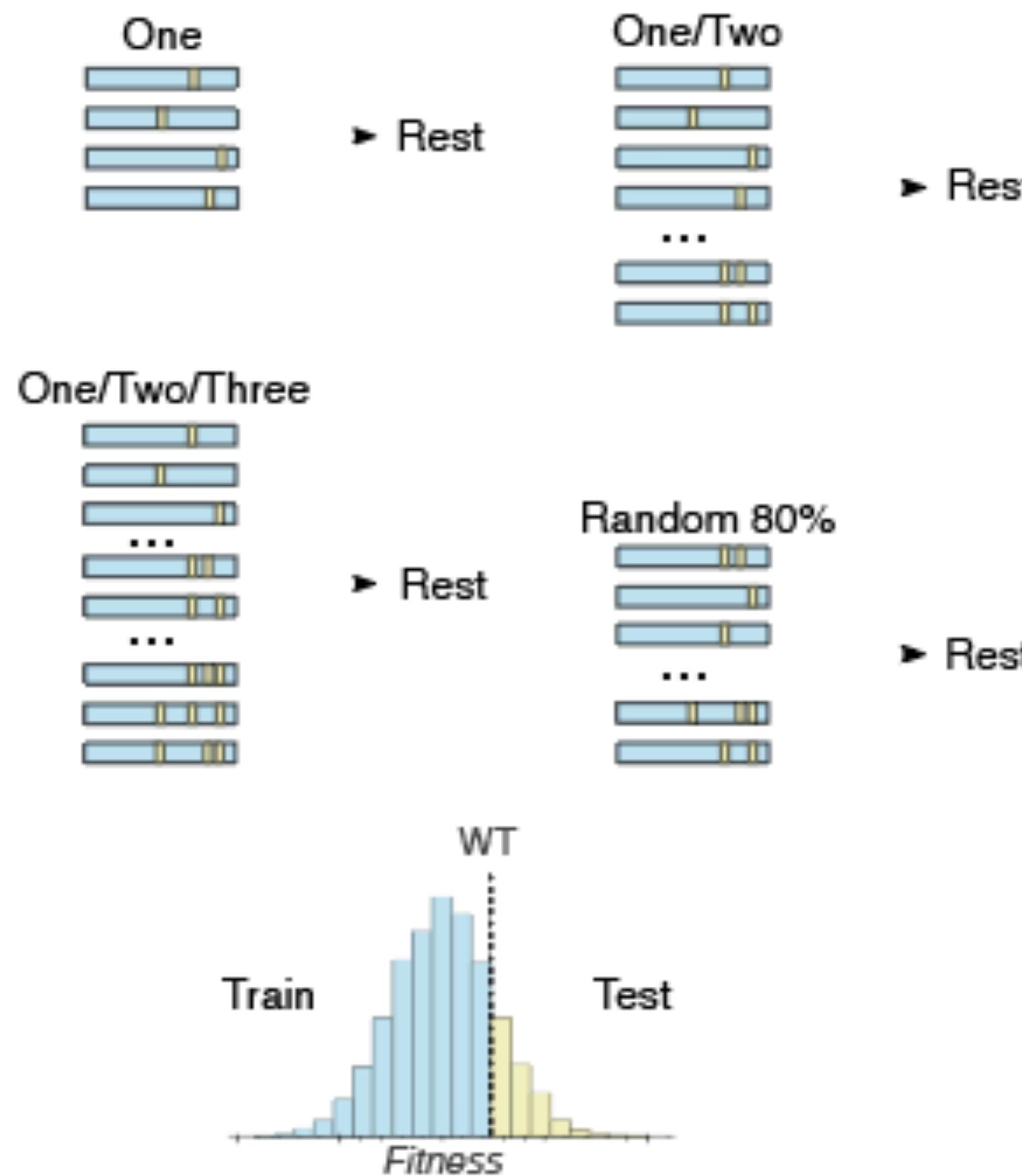
Protein engineering campaigns require out-of-domain generalization



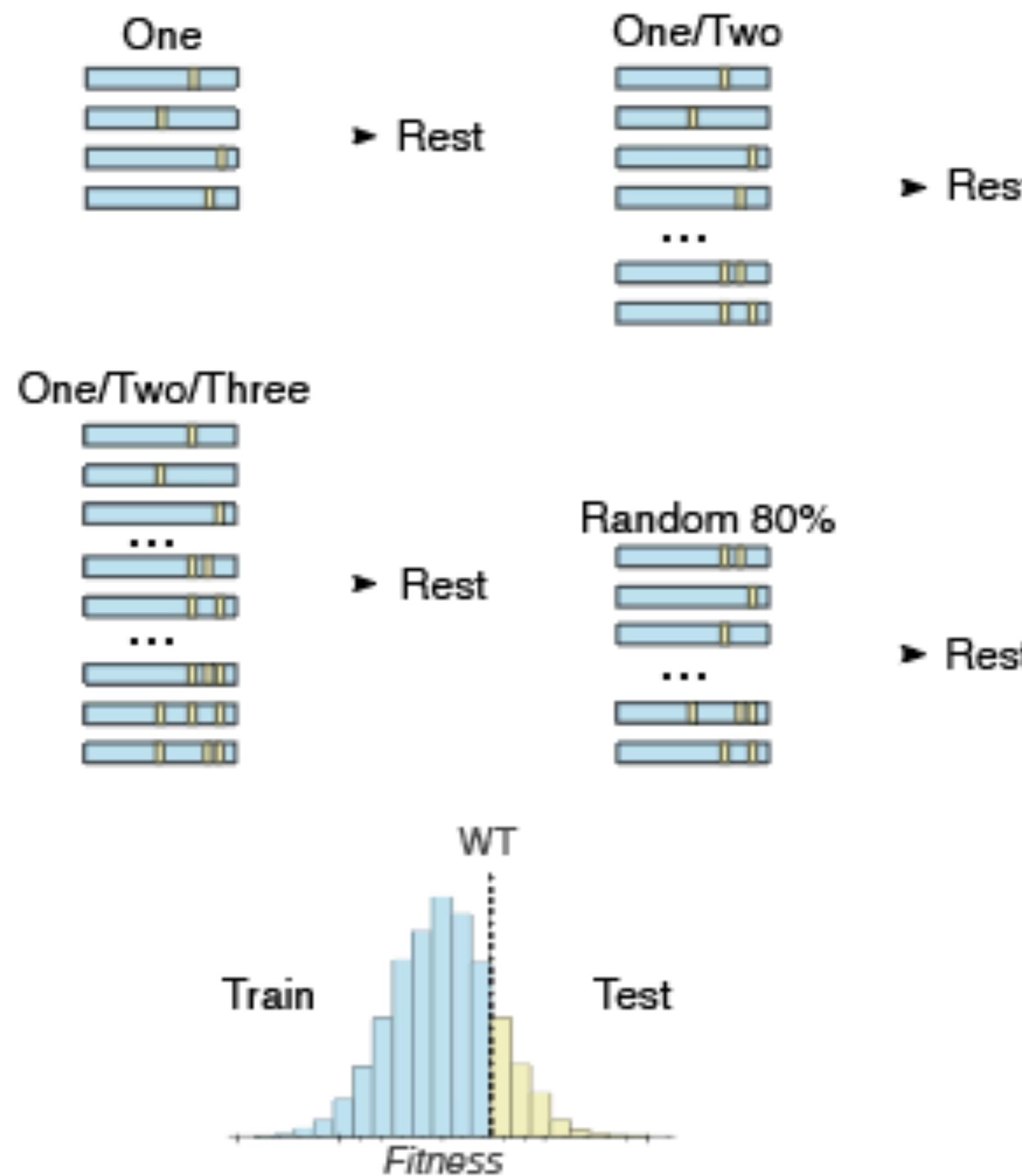
Protein engineering campaigns require out-of-domain generalization



Protein engineering campaigns require out-of-domain generalization

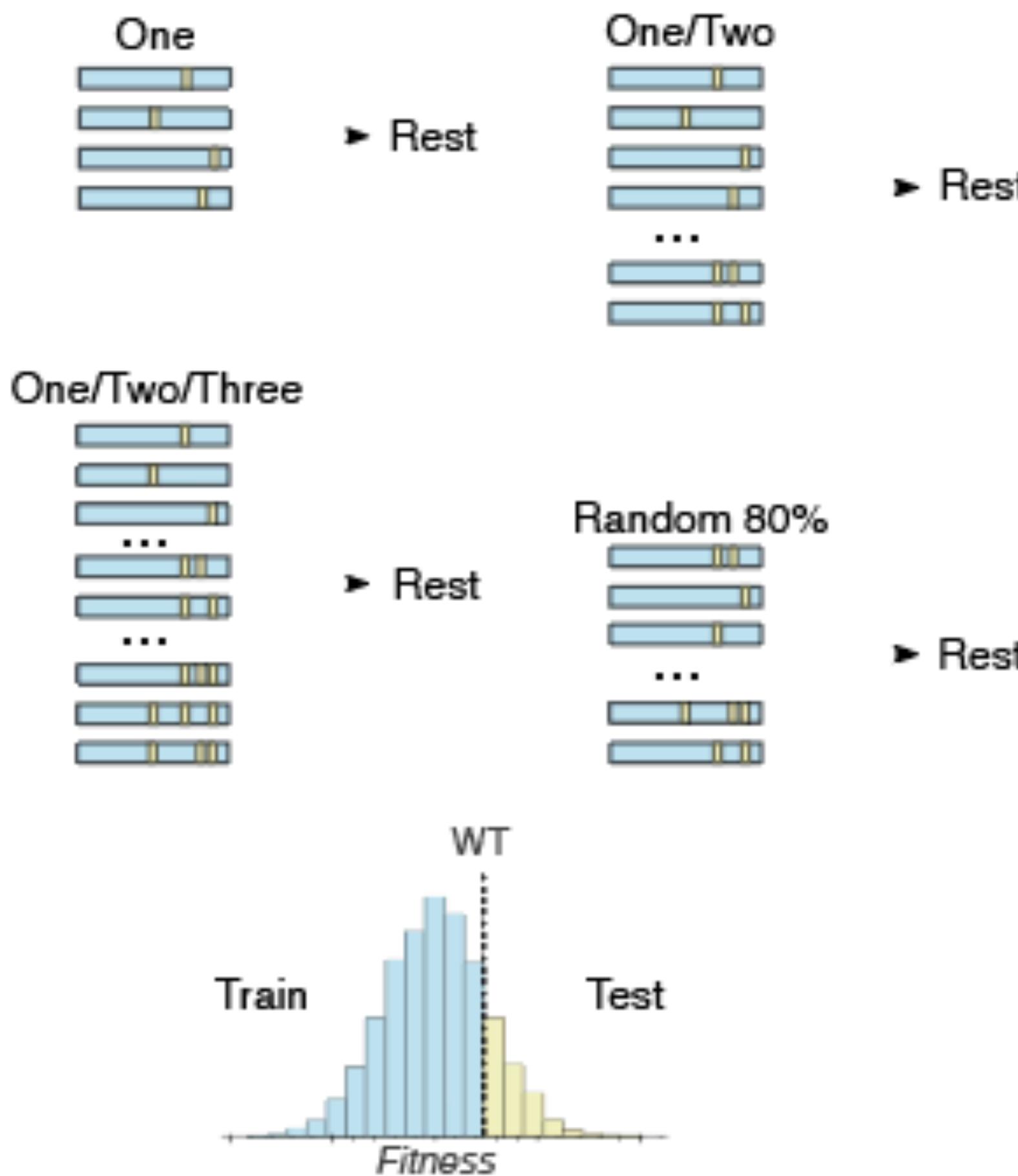


Protein engineering campaigns require out-of-domain generalization



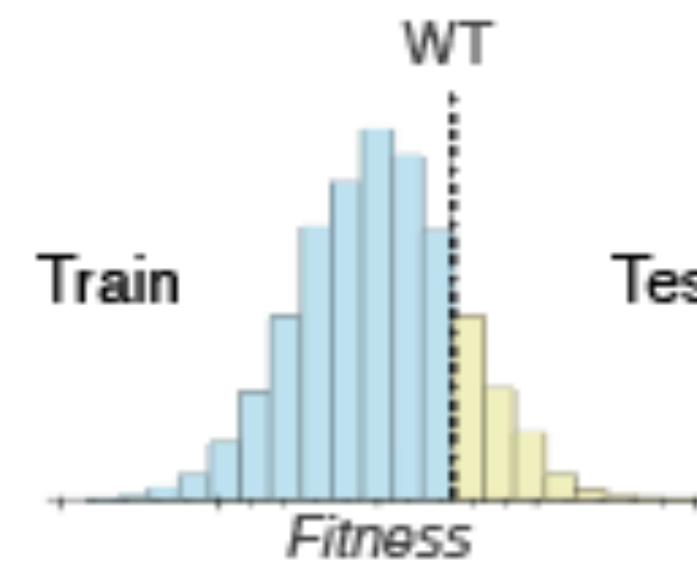
Can usually measure all single mutants

Protein engineering campaigns require out-of-domain generalization



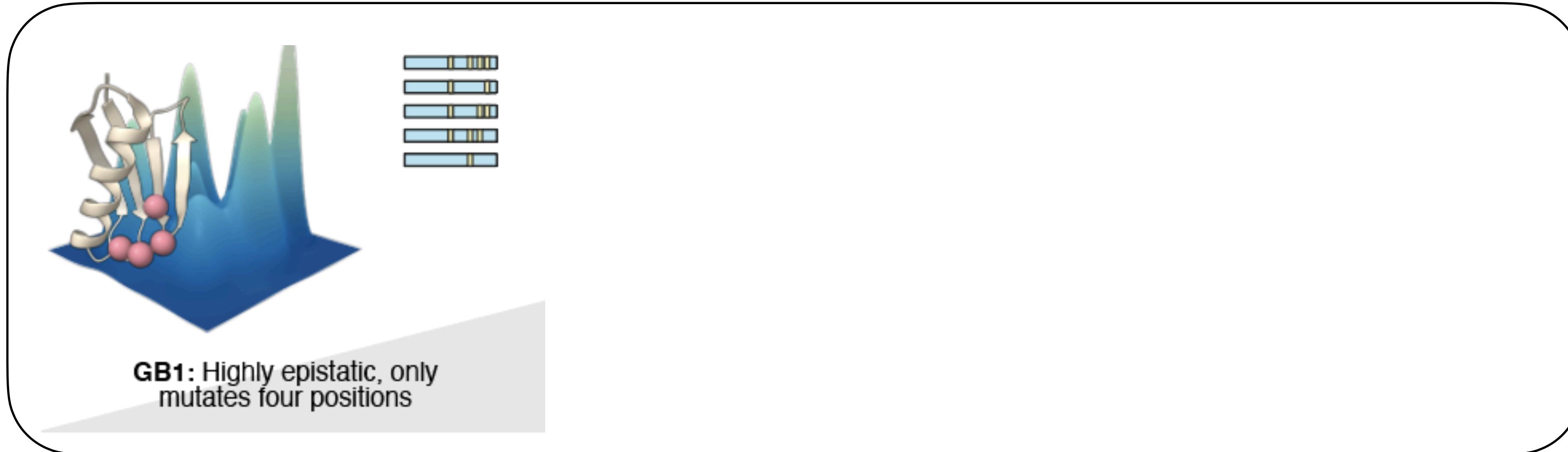
Can usually measure all single mutants

Naively screen bigger spaces
->
mostly non-functional

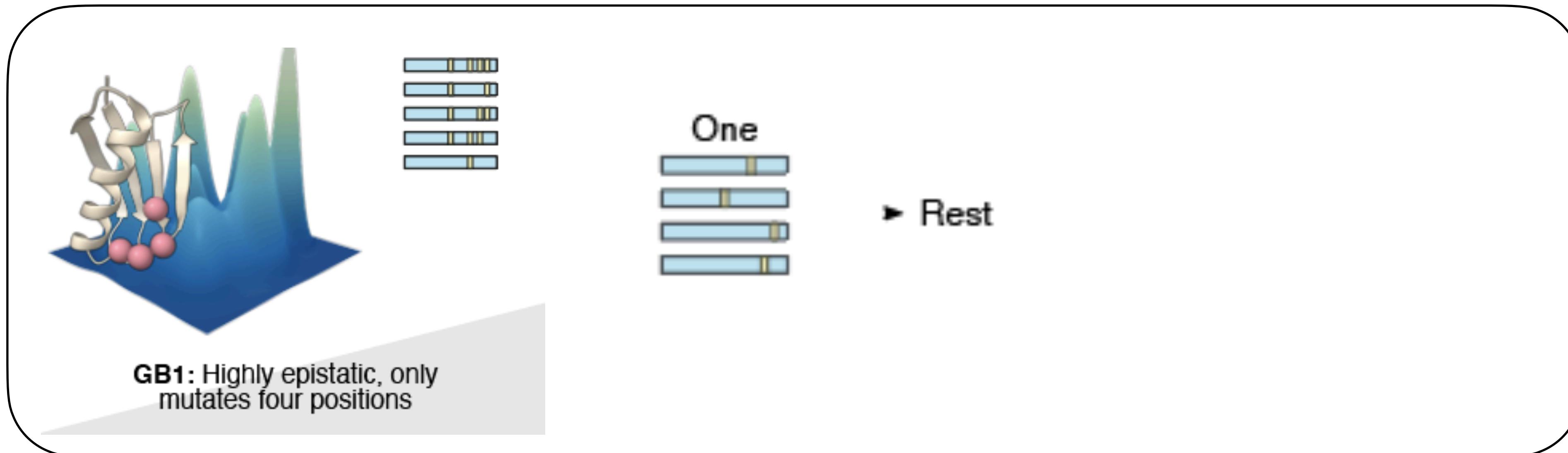


Protein engineering campaigns require out-of-domain generalization

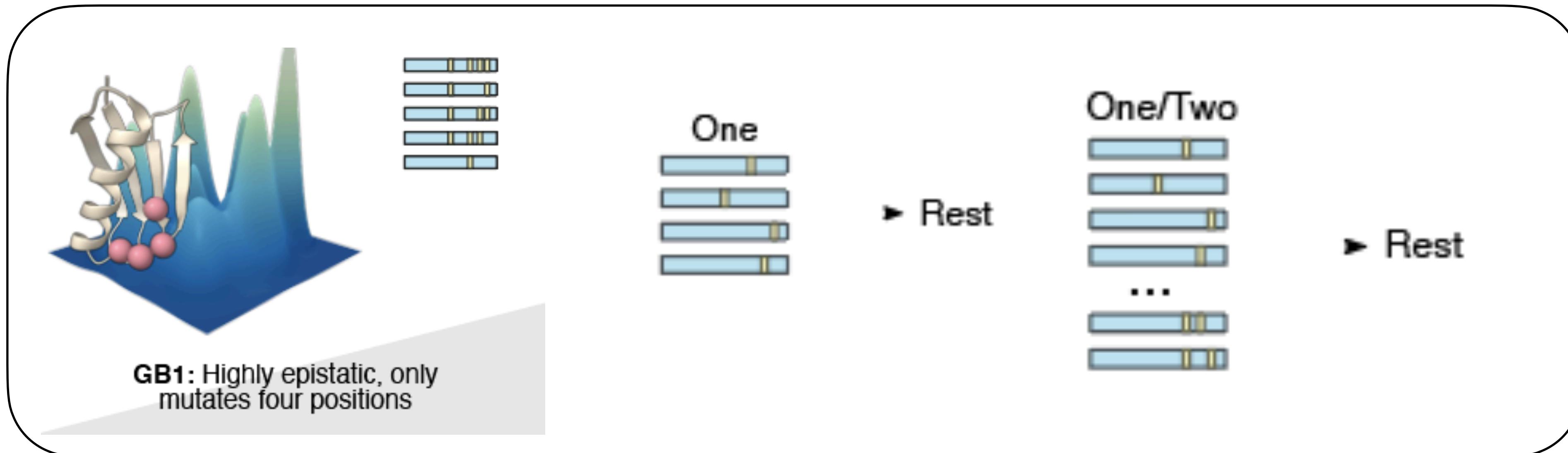
Protein engineering campaigns require out-of-domain generalization



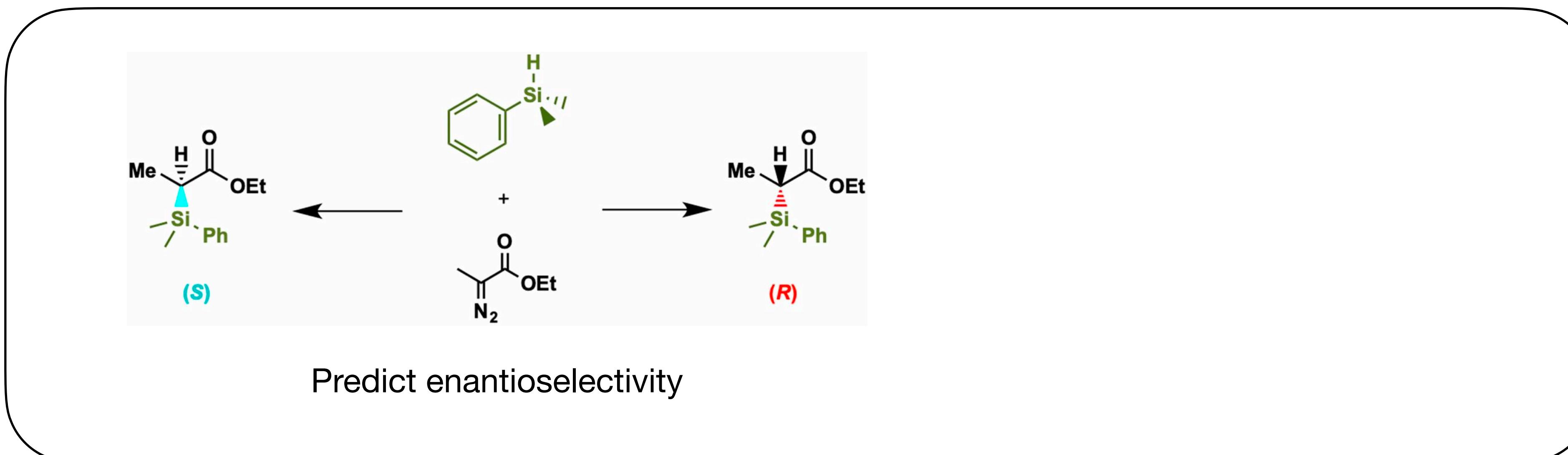
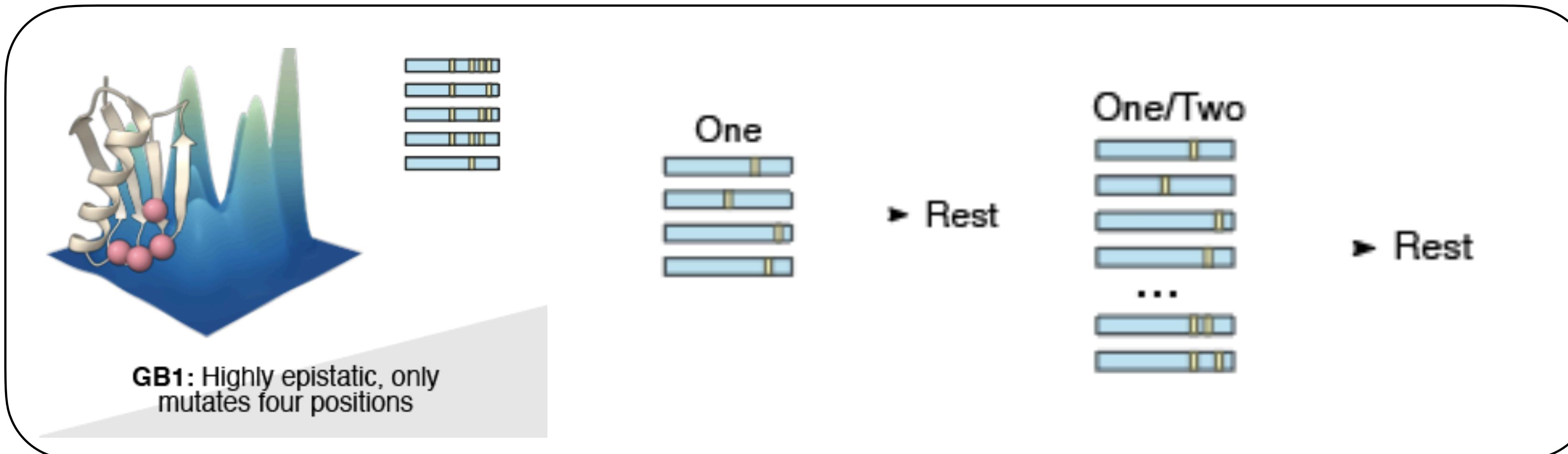
Protein engineering campaigns require out-of-domain generalization



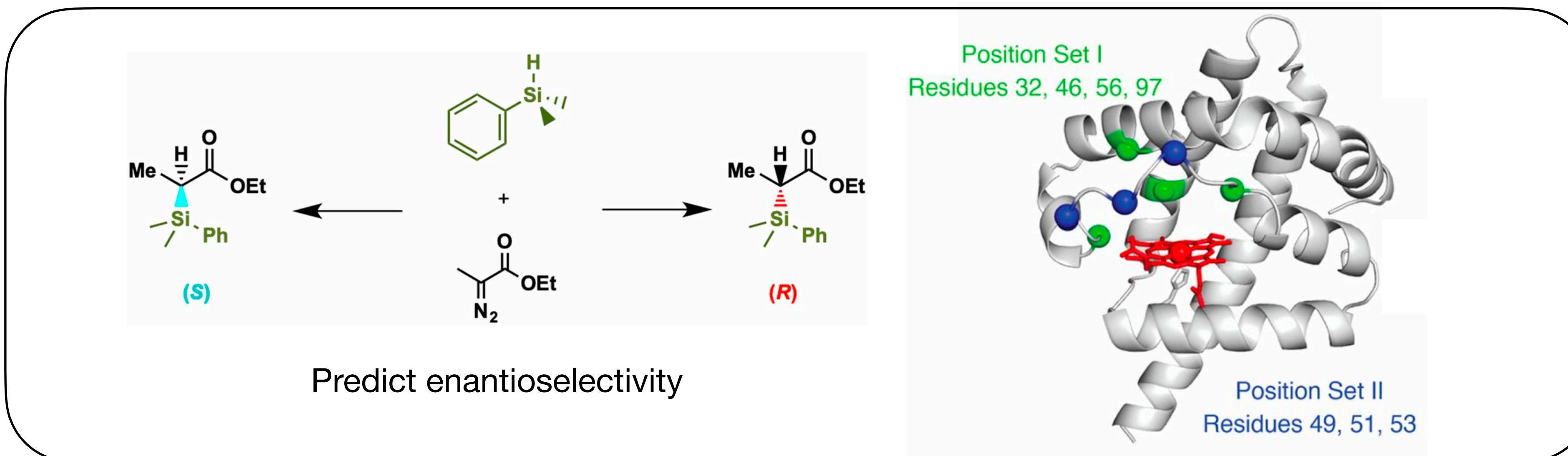
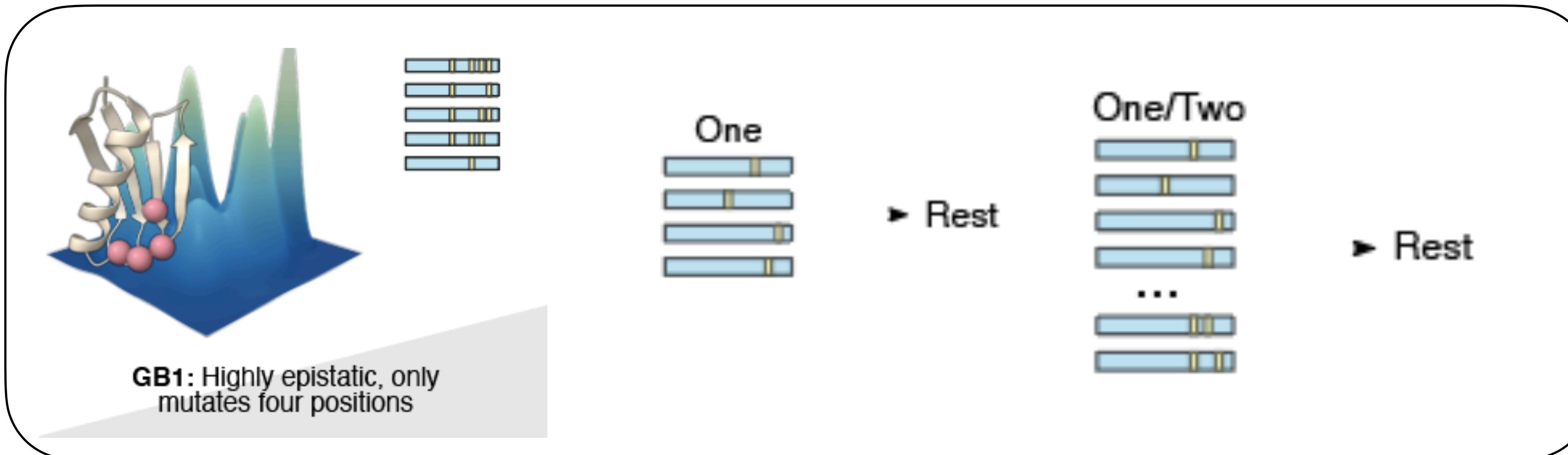
Protein engineering campaigns require out-of-domain generalization



Protein engineering campaigns require out-of-domain generalization



Protein engineering campaigns require out-of-domain generalization



Pretraining improves OOD generalization on GB1

Pretraining improves OOD generalization on GB1

Model	Spearman	
	GB1 1-vs-rest	GB1 2-vs-rest

Pretraining improves OOD generalization on GB1

	Model	Spearman	
		GB1 1-vs-rest	GB1 2-vs-rest
With pretraining	CARP-640M	0.19±0.26	0.73±0.03

Pretraining improves OOD generalization on GB1

	Model	Spearman	
		GB1 1-vs-rest	GB1 2-vs-rest
With pretraining	CARP-640M	0.19±0.26	0.73±0.03
	MIF	0.10±0.20	0.71±0.02

Pretraining improves OOD generalization on GB1

	Model	Spearman	
		GB1 1-vs-rest	GB1 2-vs-rest
With pretraining	CARP-640M	0.19±0.26	0.73±0.03
	MIF	0.10±0.20	0.71±0.02
	MIF-ST	0.22±0.03	0.74±0.03

Pretraining improves OOD generalization on GB1

	Model	Spearman	
		GB1 1-vs-rest	GB1 2-vs-rest
With pretraining	CARP-640M	0.19±0.26	0.73±0.03
	MIF	0.10±0.20	0.71±0.02
	MIF-ST	0.22±0.03	0.74±0.03
No pretraining	CARP-640M	0.11±0.17	0.38±0.26

Pretraining improves OOD generalization on GB1

	Model	Spearman	
		GB1 1-vs-rest	GB1 2-vs-rest
With pretraining	CARP-640M	0.19±0.26	0.73±0.03
	MIF	0.10±0.20	0.71±0.02
	MIF-ST	0.22±0.03	0.74±0.03
No pretraining	CARP-640M	0.11±0.17	0.38±0.26
	MIF	0.03±0.11	0.05±0.12

Pretraining improves OOD generalization on GB1

	Model	Spearman	
		GB1 1-vs-rest	GB1 2-vs-rest
With pretraining	CARP-640M	0.19±0.26	0.73±0.03
	MIF	0.10±0.20	0.71±0.02
	MIF-ST	0.22±0.03	0.74±0.03
No pretraining	CARP-640M	0.11±0.17	0.38±0.26
	MIF	0.03±0.11	0.05±0.12
	MIF-ST	NaN	NaN

Pretraining improves OOD generalization on GB1

	Model	Spearman	
		GB1 1-vs-rest	GB1 2-vs-rest
With pretraining	CARP-640M	0.19±0.26	0.73±0.03
	MIF	0.10±0.20	0.71±0.02
	MIF-ST	0.22±0.03	0.74±0.03
No pretraining	CARP-640M	0.11±0.17	0.38±0.26
	MIF	0.03±0.11	0.05±0.12
	MIF-ST	NaN	NaN
Baseline	CNN	0.15±0.09	0.39±0.04

Pretraining improves OOD generalization on GB1

	Model	Spearman	
		GB1 1-vs-rest	GB1 2-vs-rest
With pretraining	CARP-640M	0.19±0.26	0.73±0.03
	MIF	0.10±0.20	0.71±0.02
	MIF-ST	0.22±0.03	0.74±0.03
No pretraining	CARP-640M	0.11±0.17	0.38±0.26
	MIF	0.03±0.11	0.05±0.12
	MIF-ST	NaN	NaN
Baseline	CNN	0.15±0.09	0.39±0.04

Structure and sequence transfer help a little

Pretraining improves OOD generalization on GB1

	Model	Spearman	
		GB1 1-vs-rest	GB1 2-vs-rest
With pretraining	CARP-640M	0.19±0.26	0.73±0.03
	MIF	0.10±0.20	0.71±0.02
	MIF-ST	0.22±0.03	0.74±0.03
No pretraining	CARP-640M	0.11±0.17	0.38±0.26
	MIF	0.03±0.11	0.05±0.12
	MIF-ST	NaN	NaN
Baseline	CNN	0.15±0.09	0.39±0.04

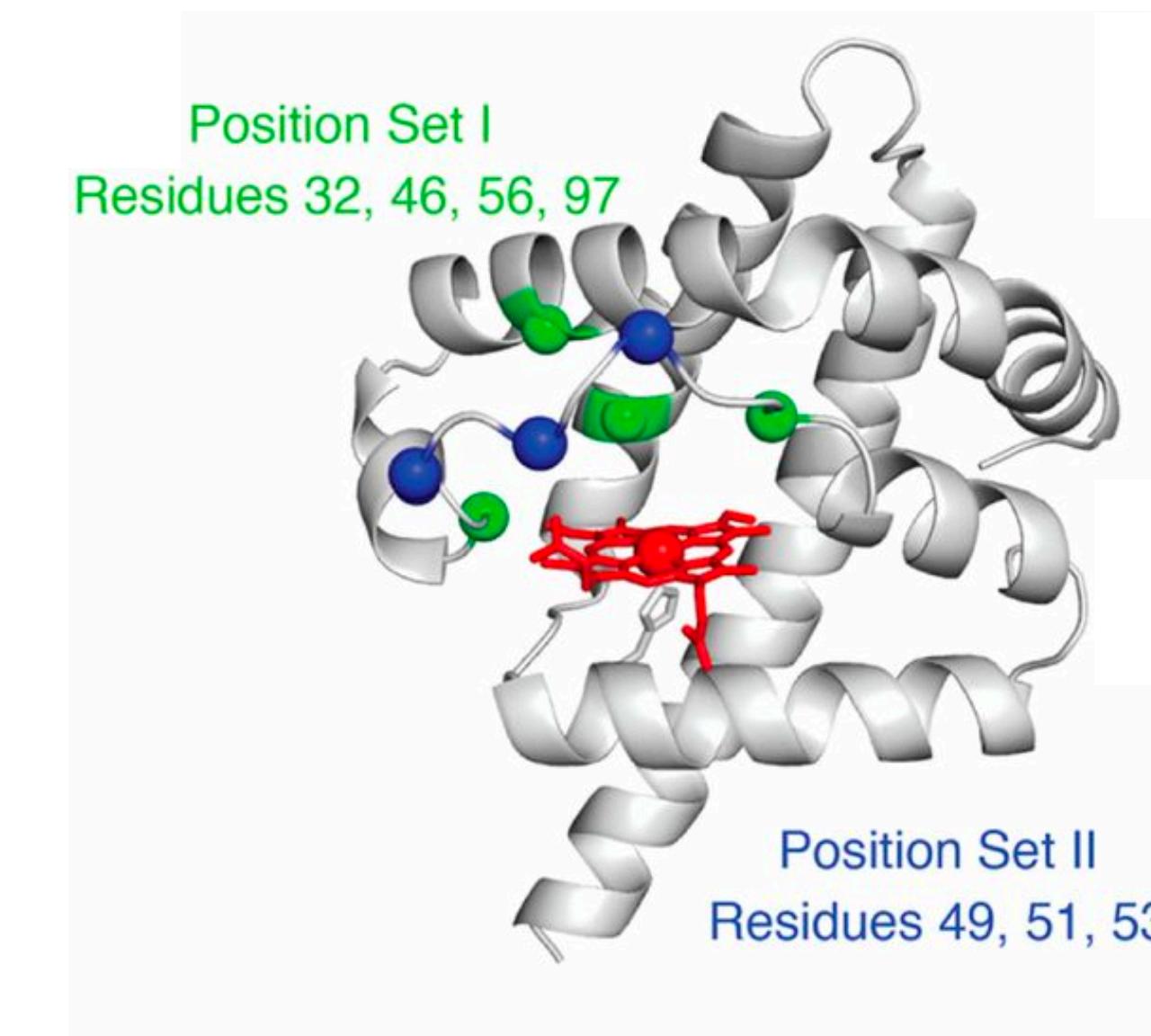
Structure and sequence transfer help a little
Naive MIF-ST predicts all the same values

Sequence and structure both required on *RMA* NOD enantioselectivity

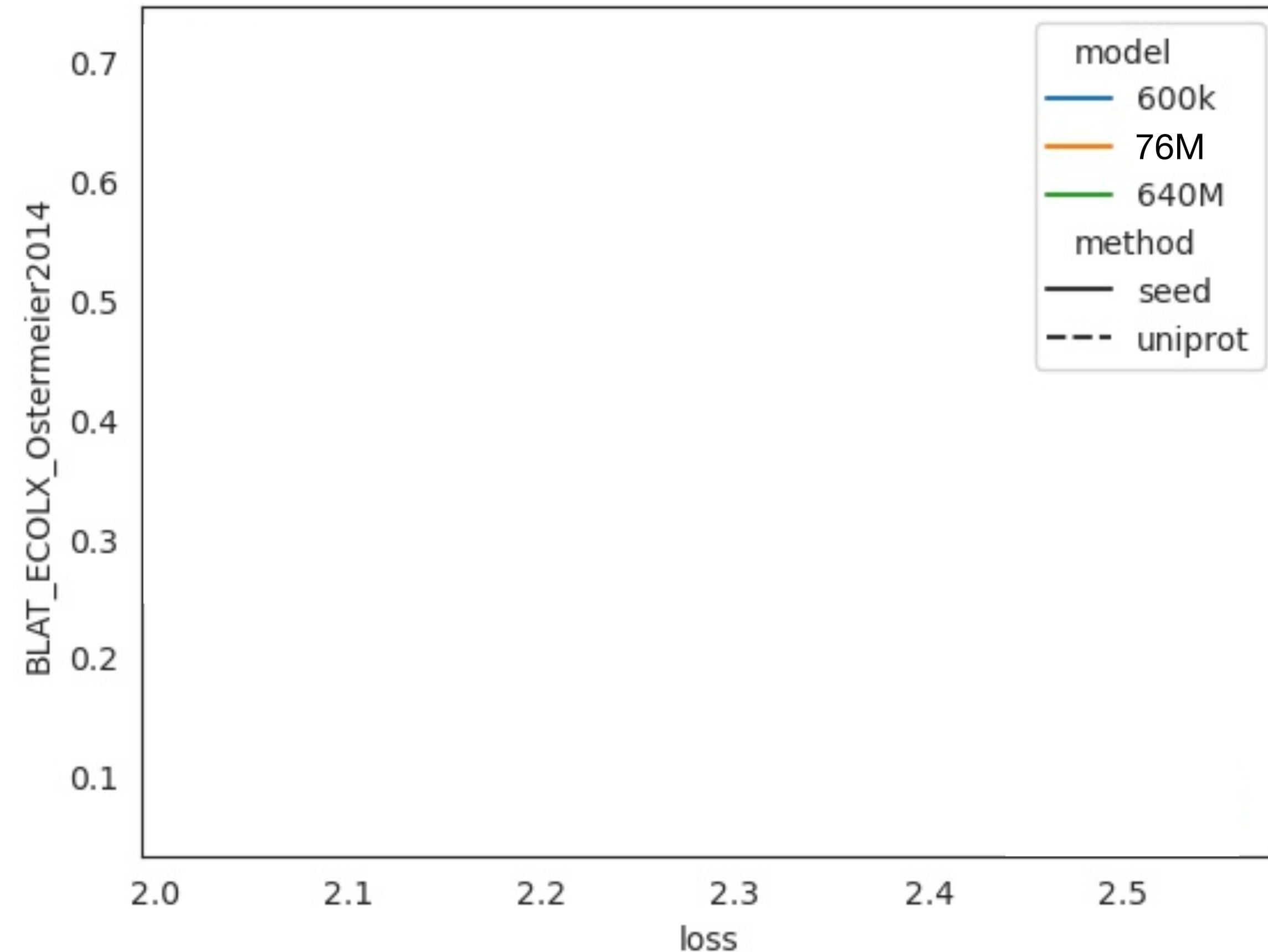
	Model	Spearman
		<i>Rma NOD</i>
With pretraining	CARP-640M	0.69±0.05
	MIF	0.66±0.11
	MIF-ST	0.77±0.03
No pretraining	CARP-640M	0.70±0.03
	MIF	0.66±0.09
	MIF-ST	0.73±0.05

Sequence and structure both required on *RMA* NOD enantioselectivity

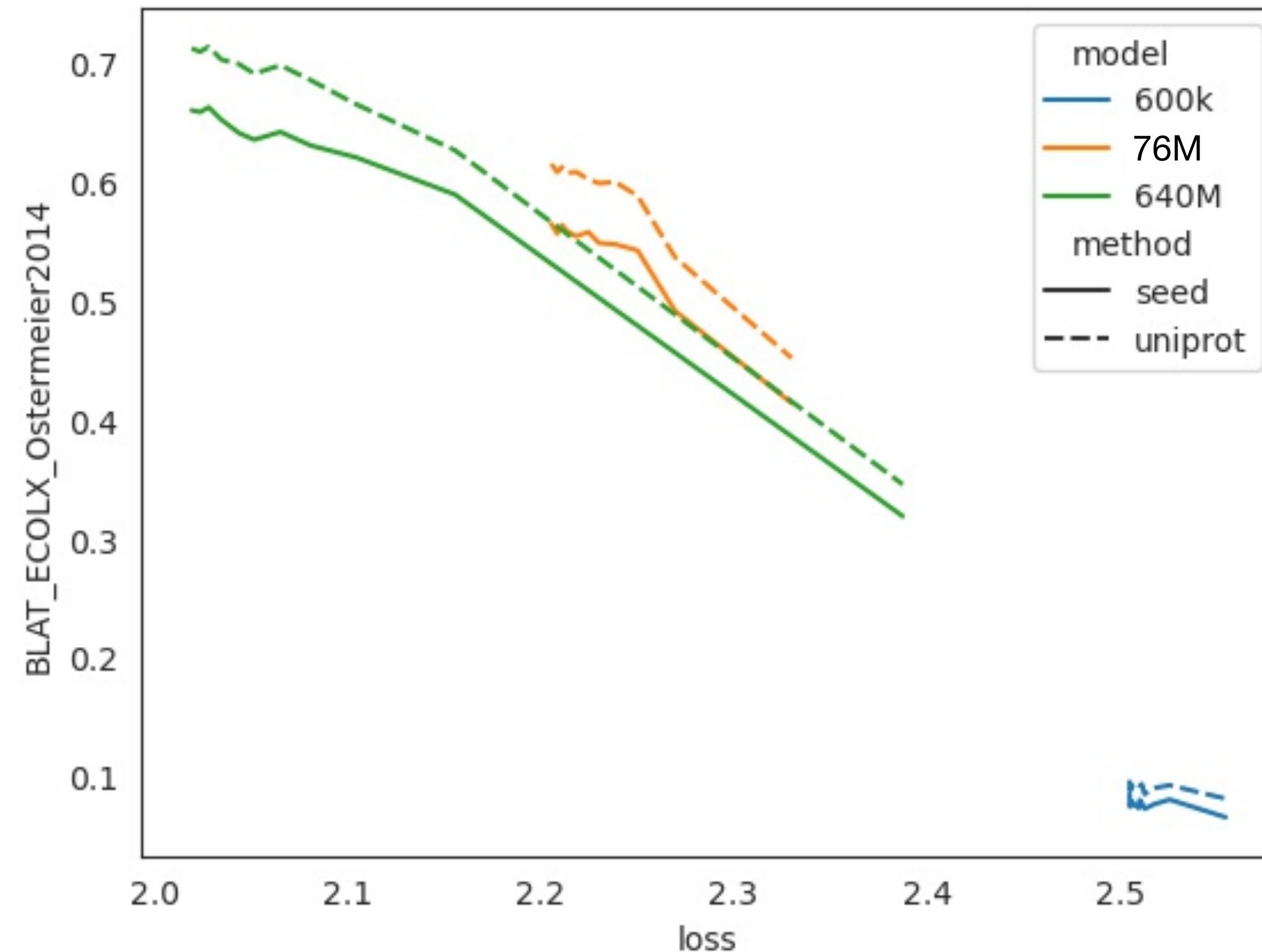
	Model	Spearman <i>Rma NOD</i>
With pretraining	CARP-640M	0.69±0.05
	MIF	0.66±0.11
	MIF-ST	0.77±0.03
No pretraining	CARP-640M	0.70±0.03
	MIF	0.66±0.09
	MIF-ST	0.73±0.05



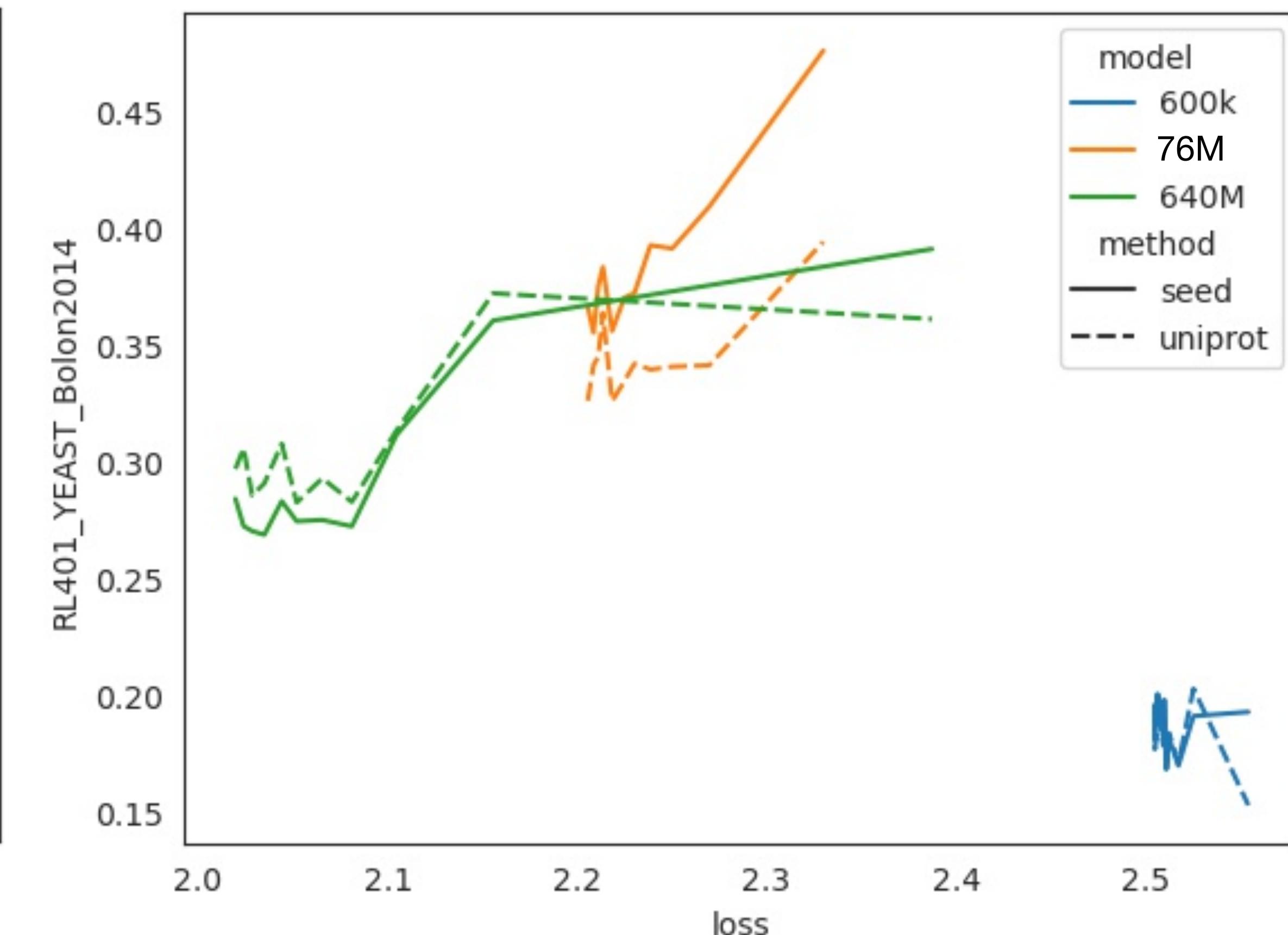
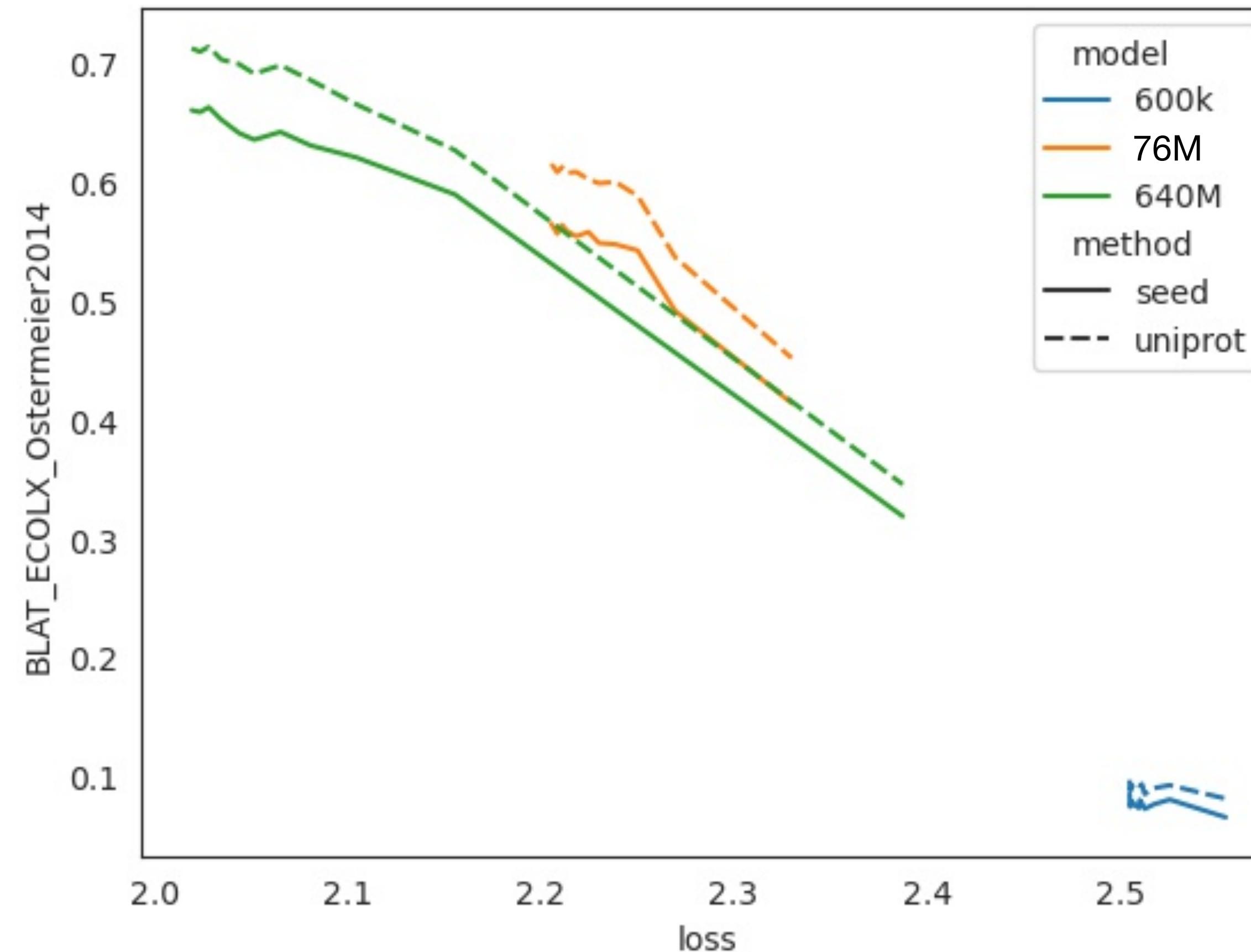
Zero-shot performance mostly improves with more pretraining



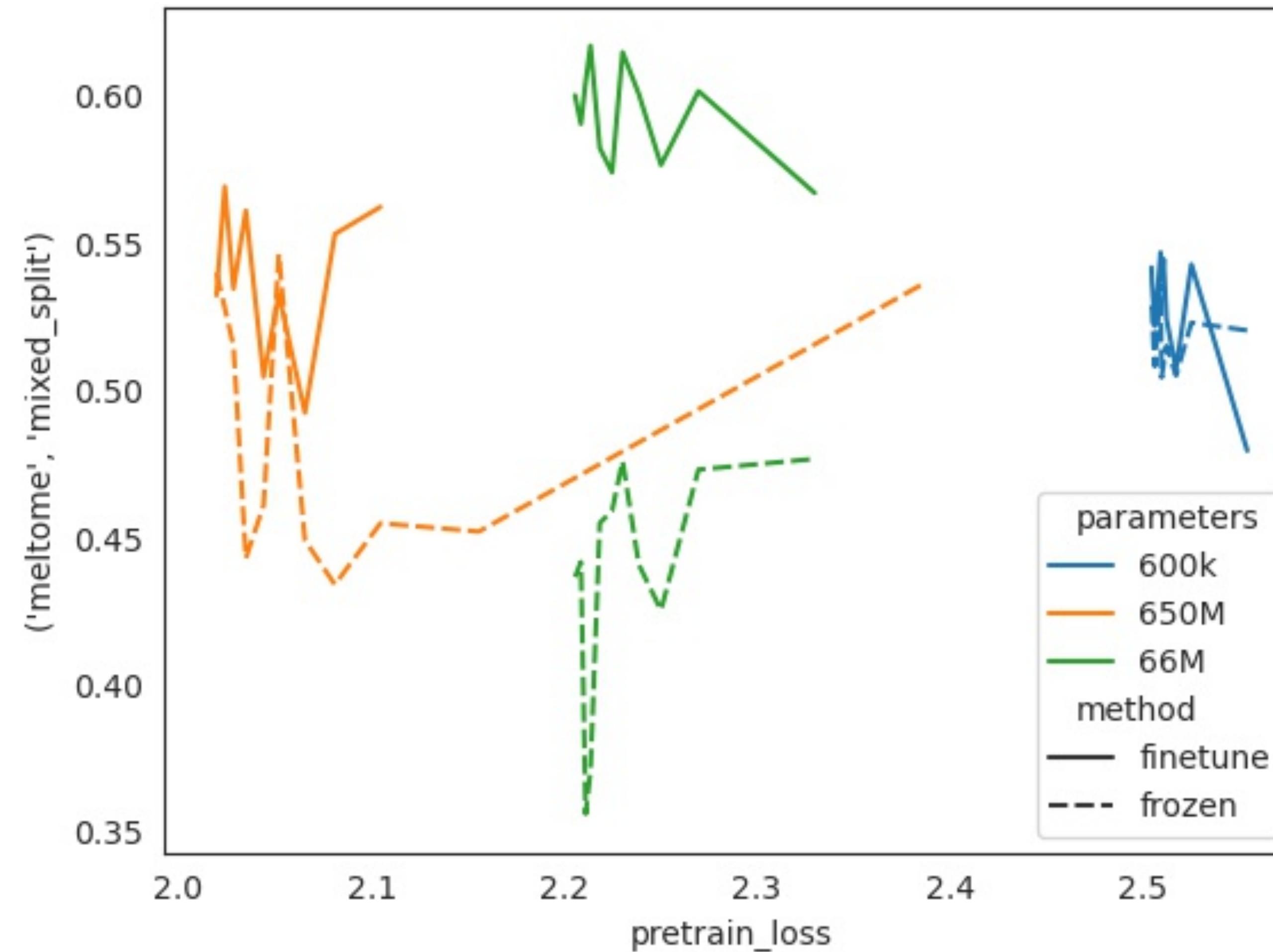
Zero-shot performance mostly improves with more pretraining



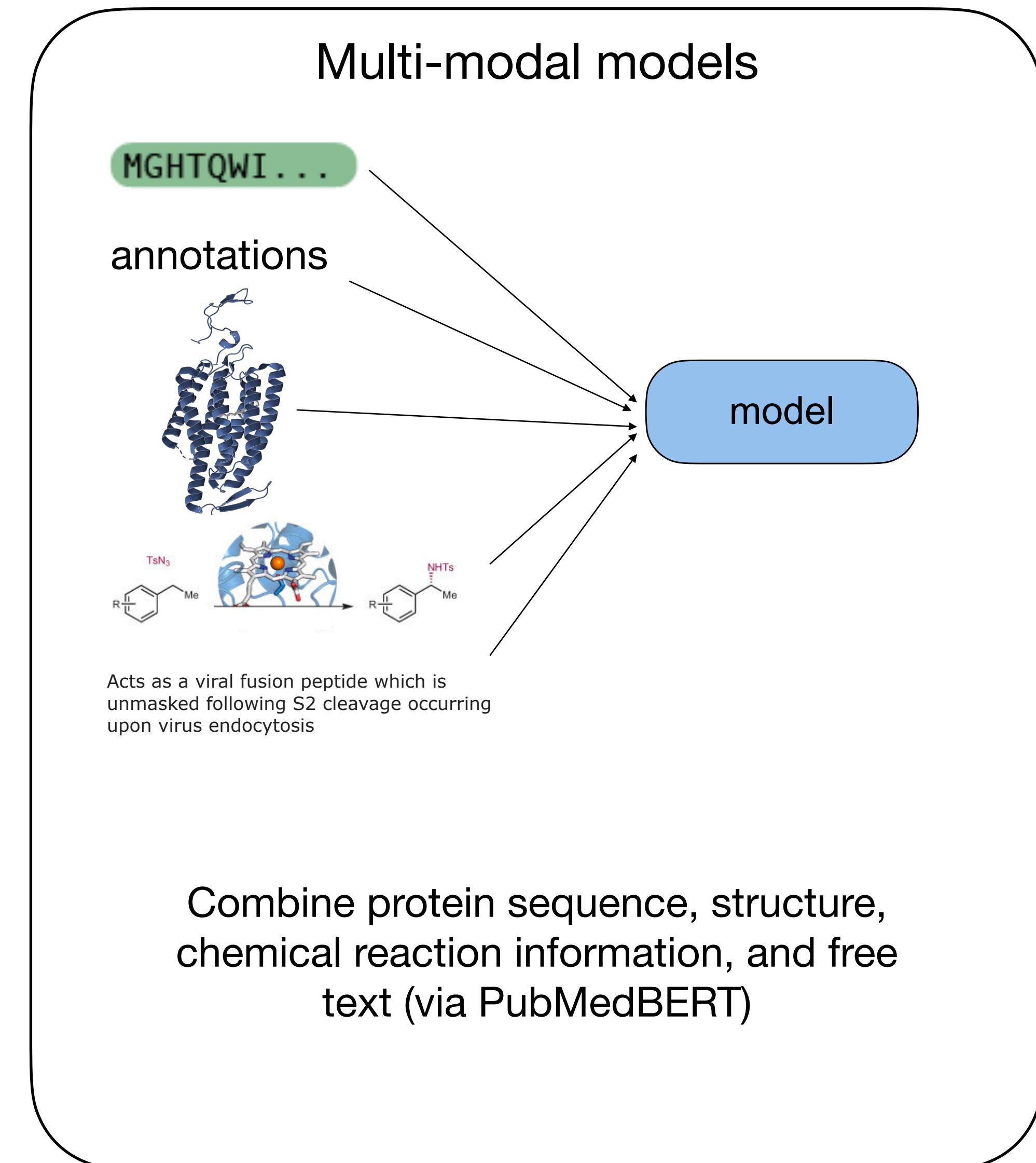
Zero-shot performance mostly improves with more pretraining



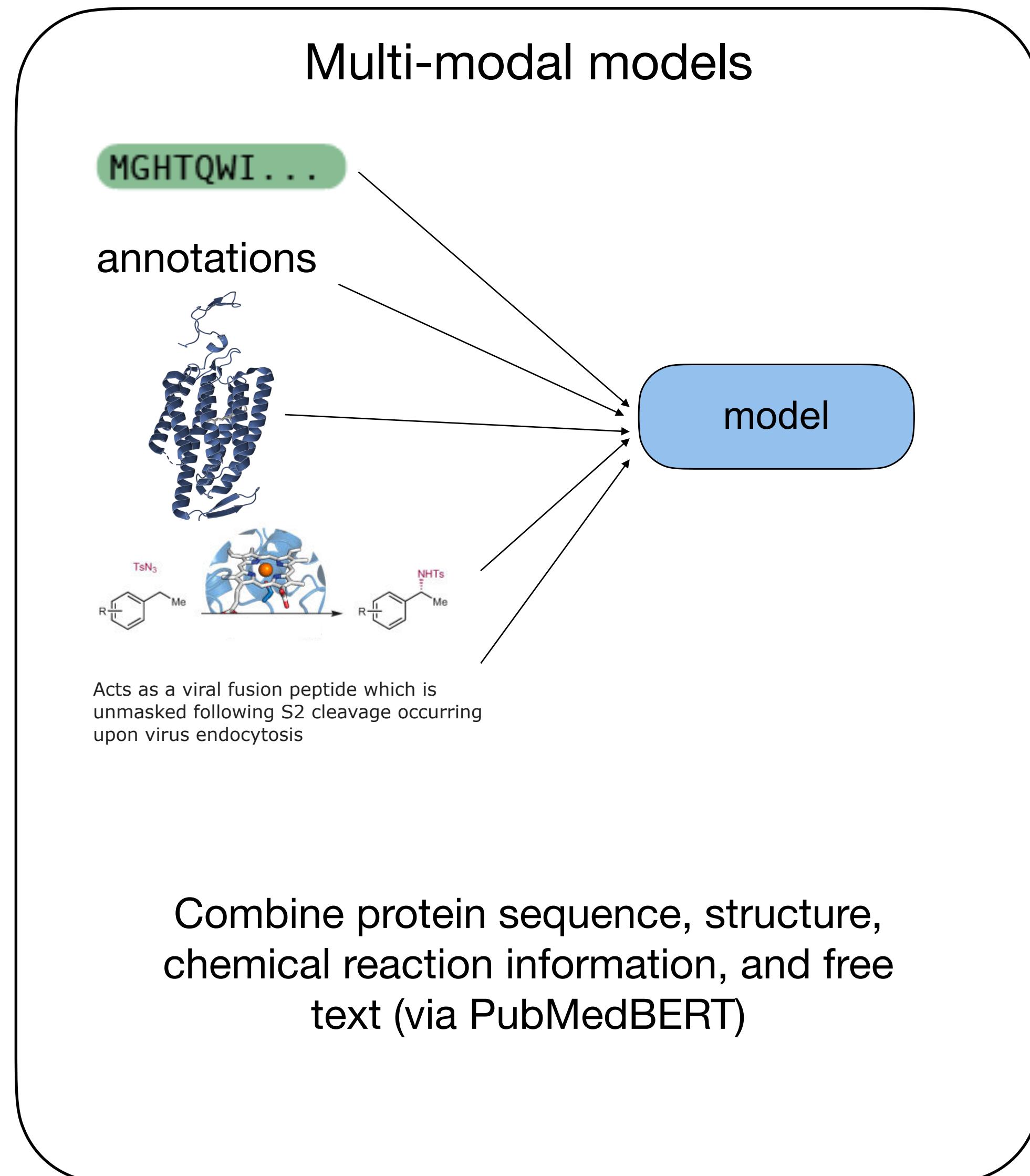
OOD performance mostly does not improve with more pretraining



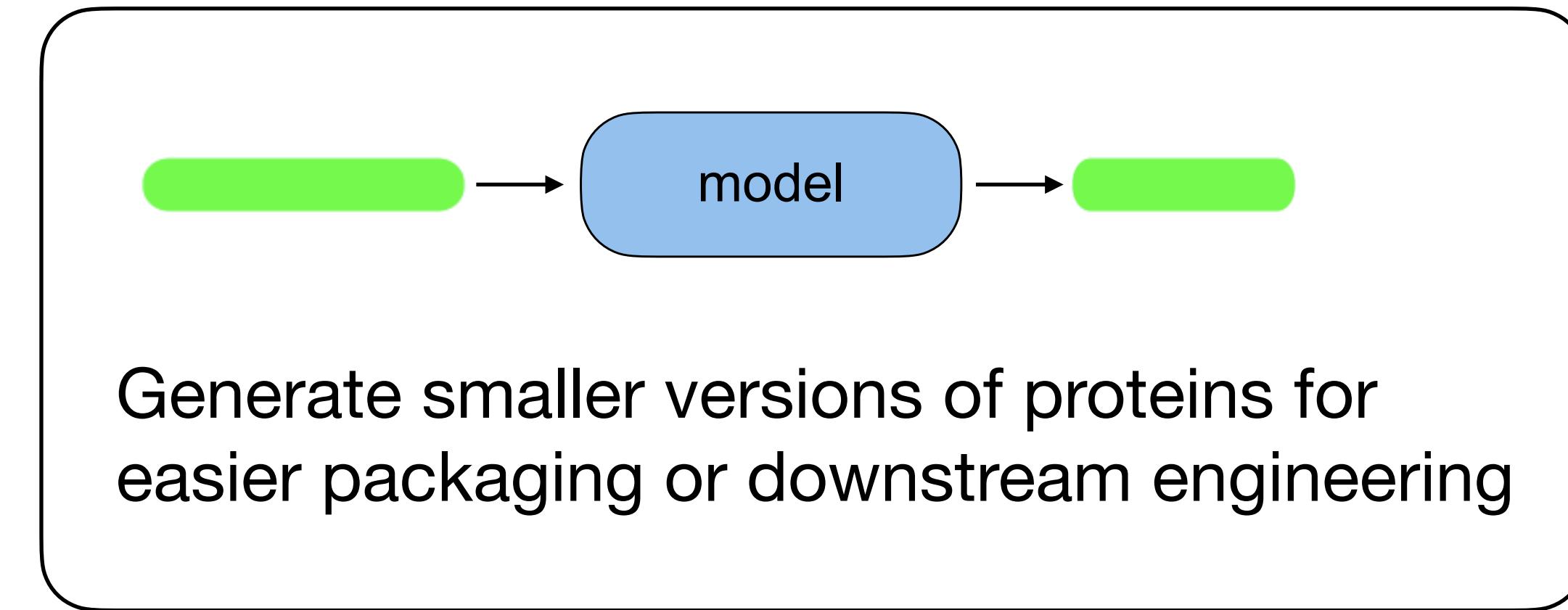
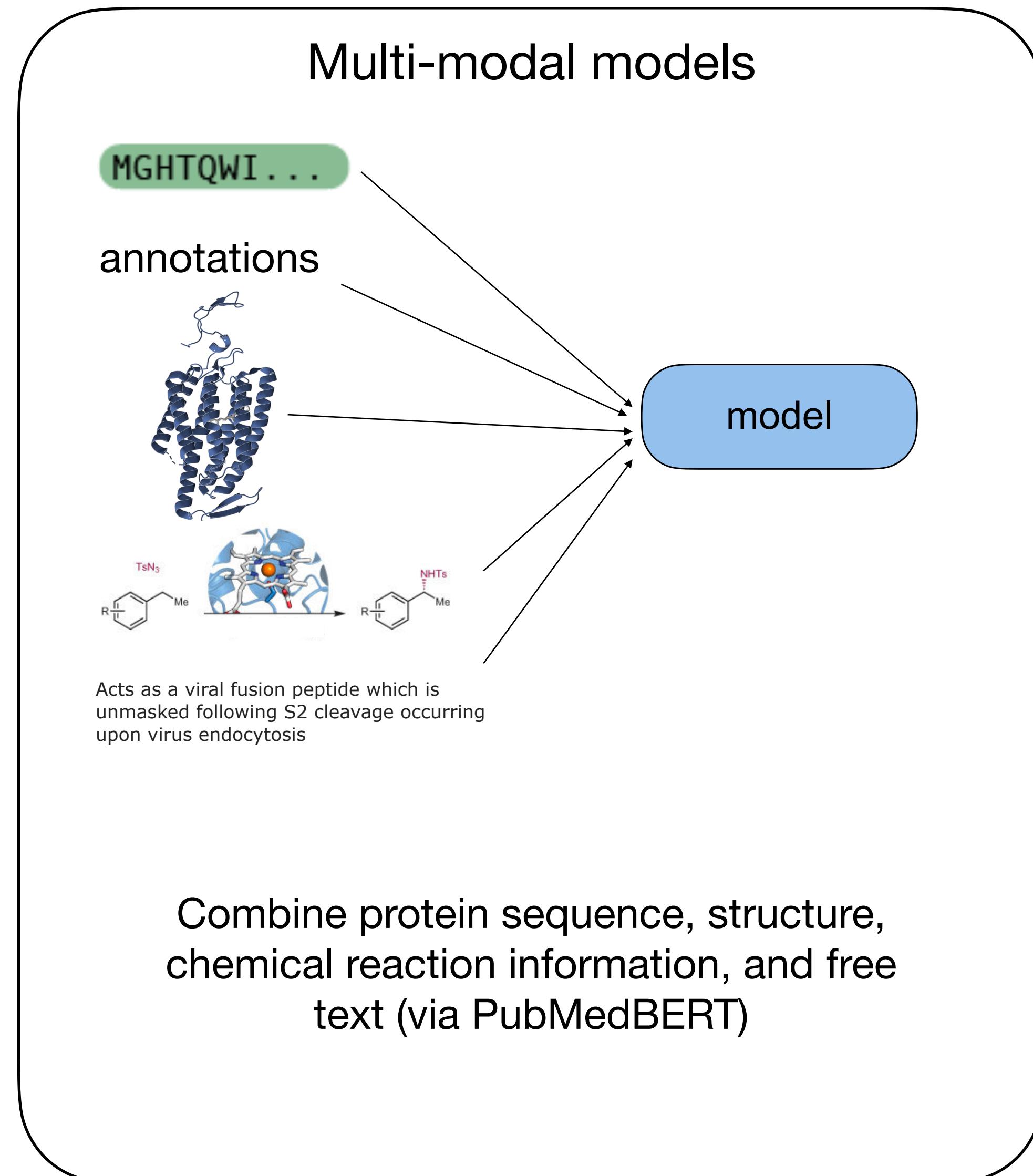
Can different modalities and pretraining tasks do better?



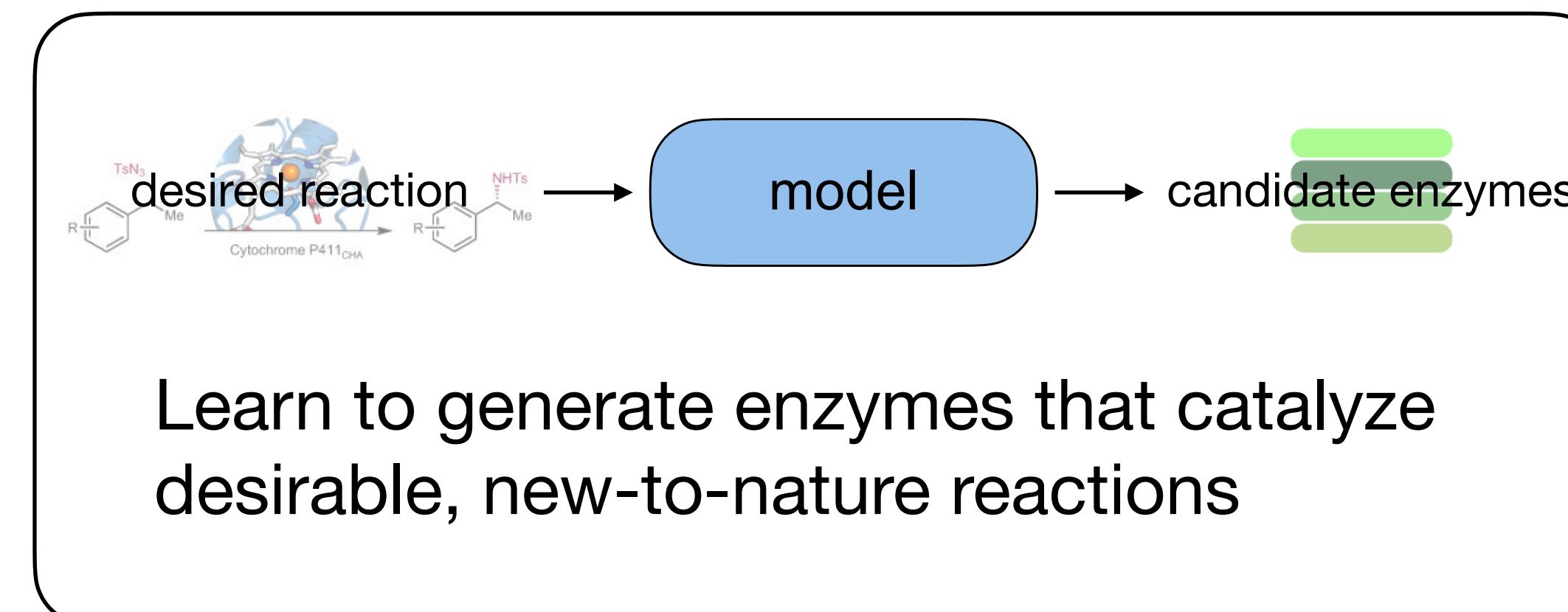
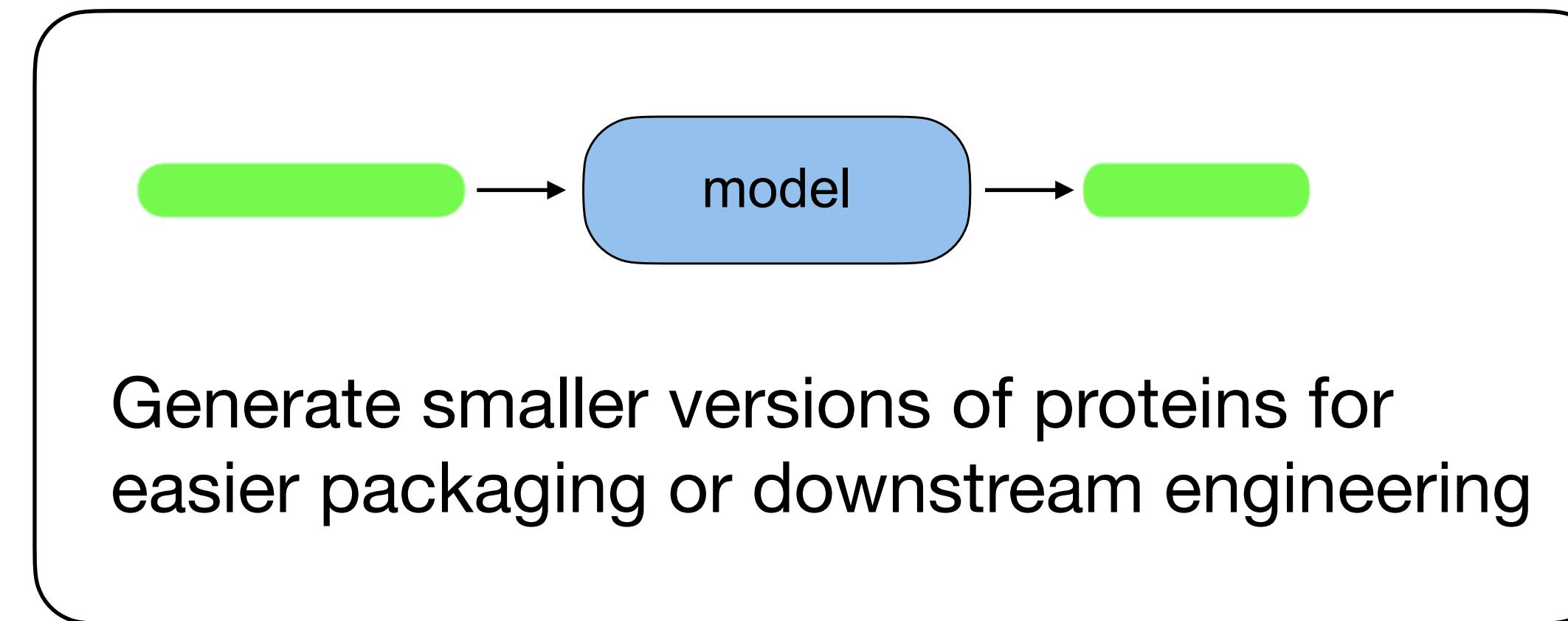
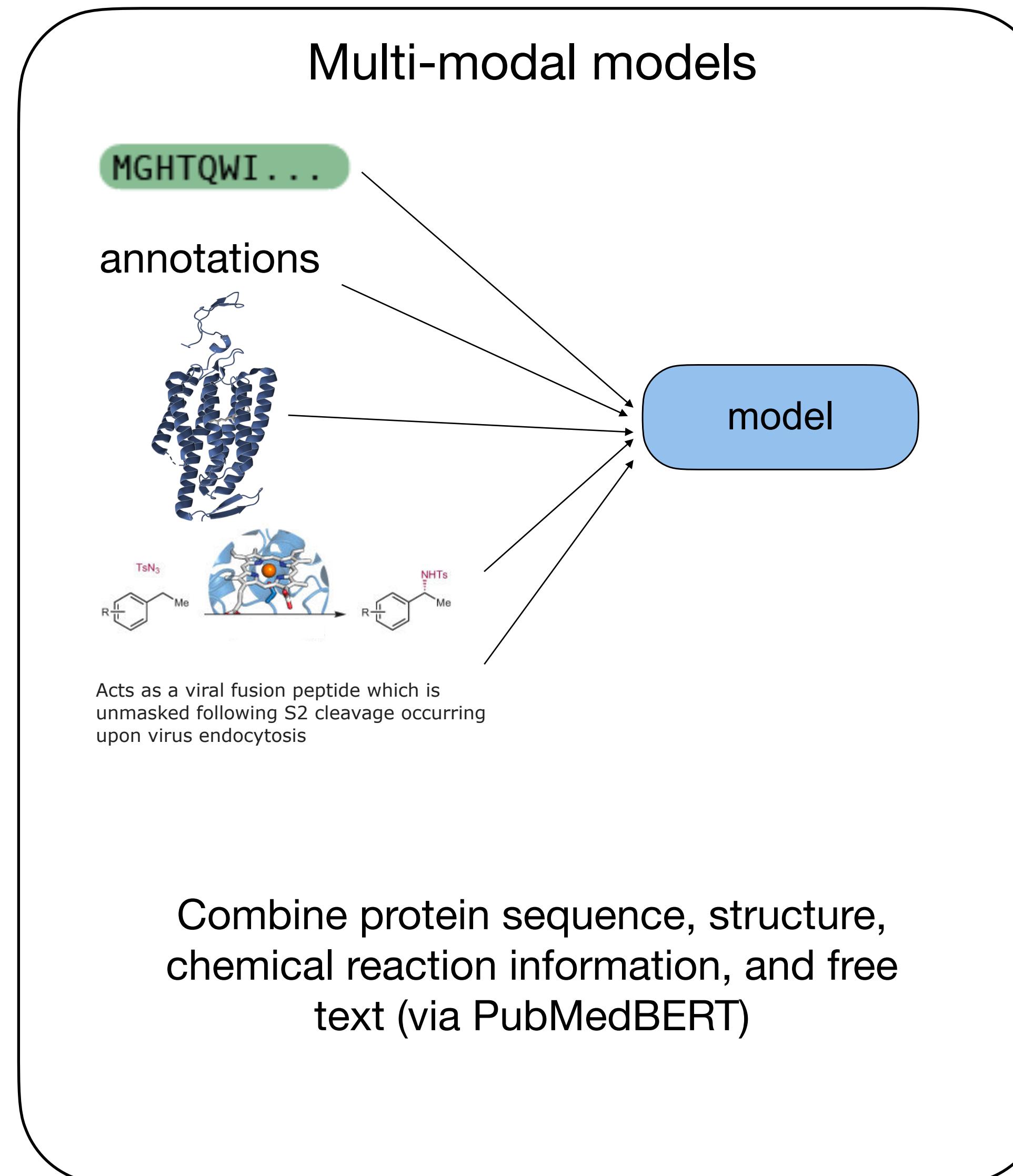
Future work: generating small proteins and enzymes



Future work: generating small proteins and enzymes



Future work: generating small proteins and enzymes



Come to ML for protein engineering!

<https://www.ml4proteinengineering.com/>

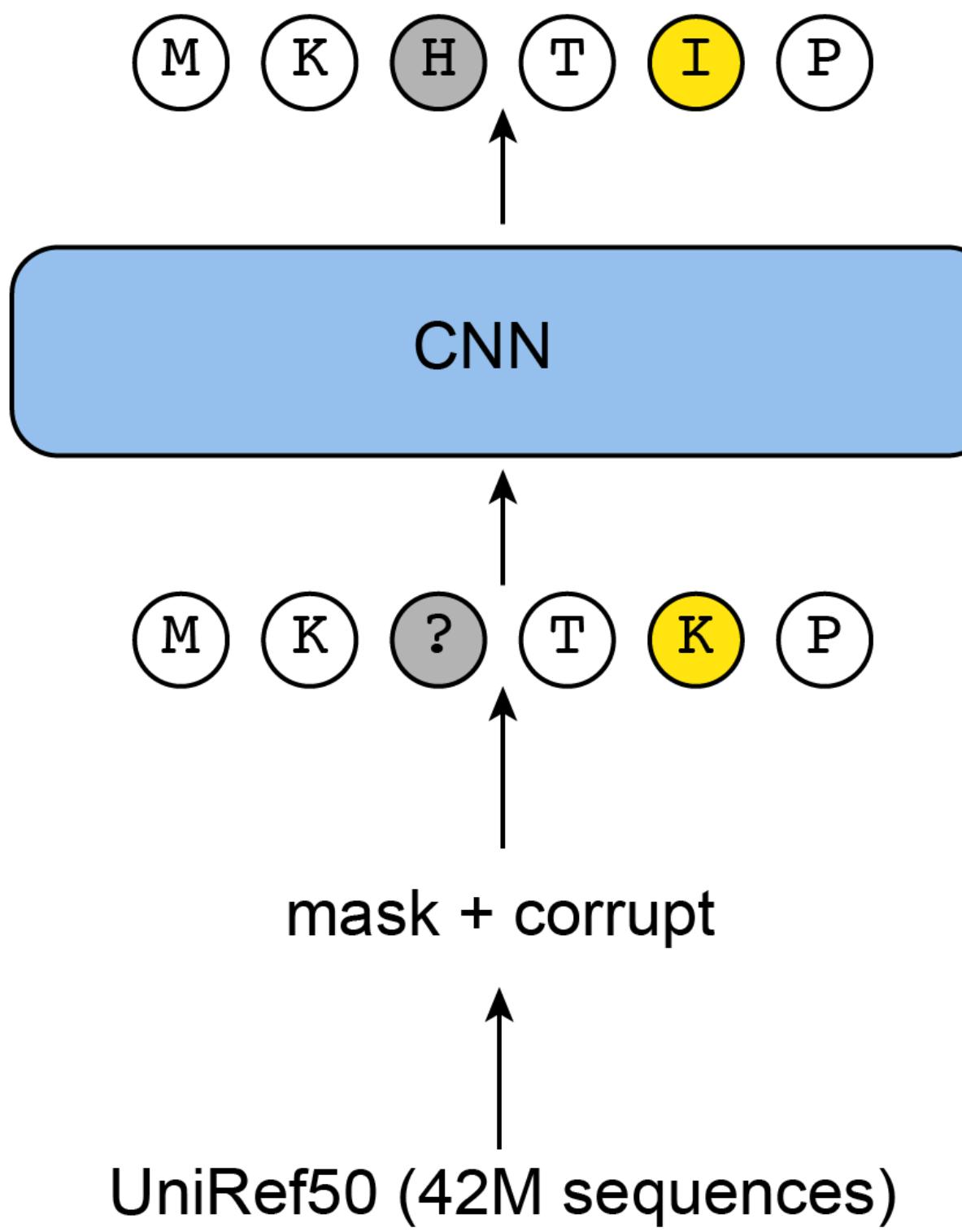
Aug 16th – Justas Dauparas – Postdoctoral Fellow, University of Washington

Sep 6th – Brian Trippe – Postdoctoral Fellow, Columbia University & University of Washington + Jason Yim – PhD Candidate, MIT

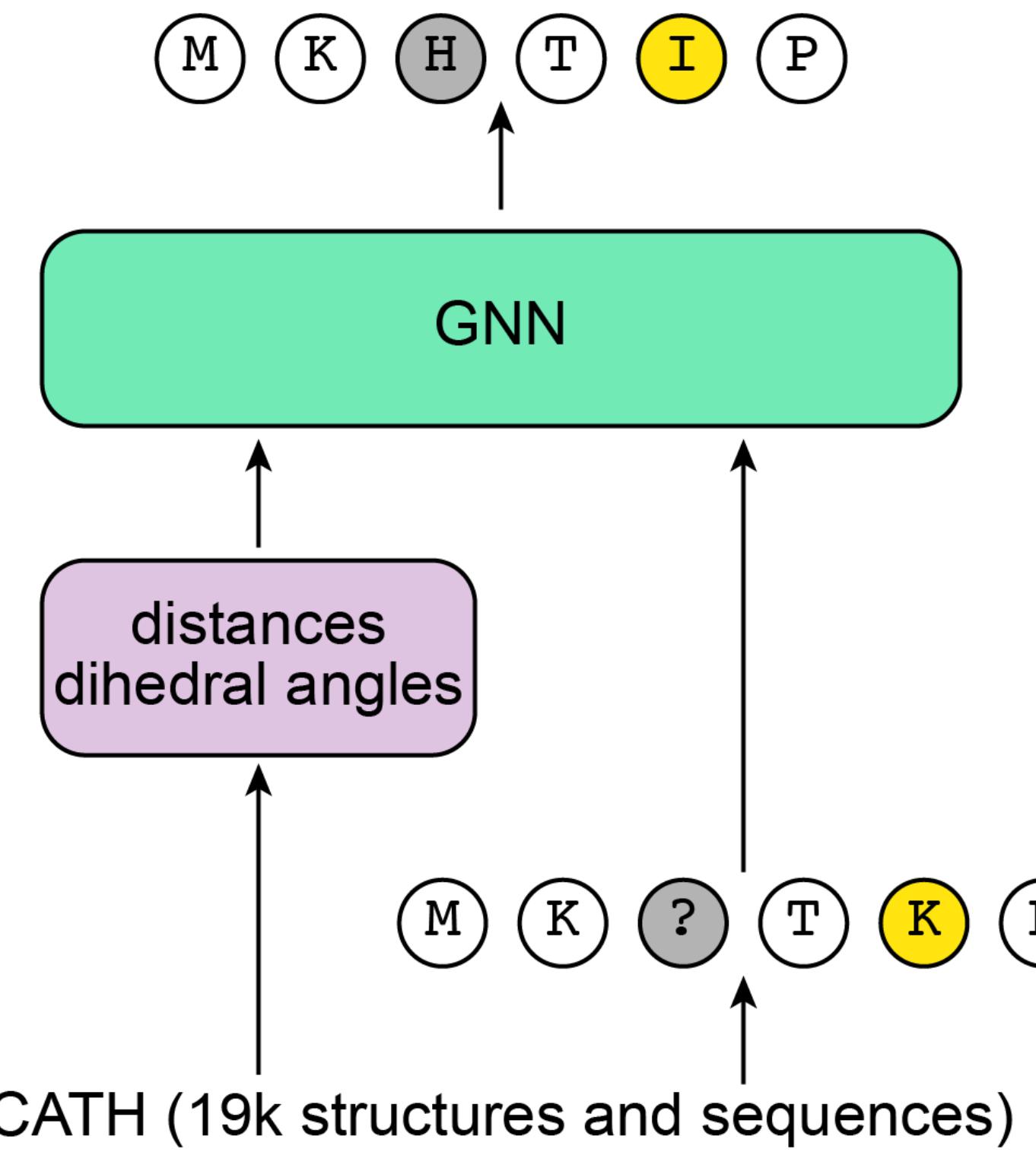
Sep 20th – Moderated discussion panel with all speakers

Try CARP, MIF, and MIF-ST!

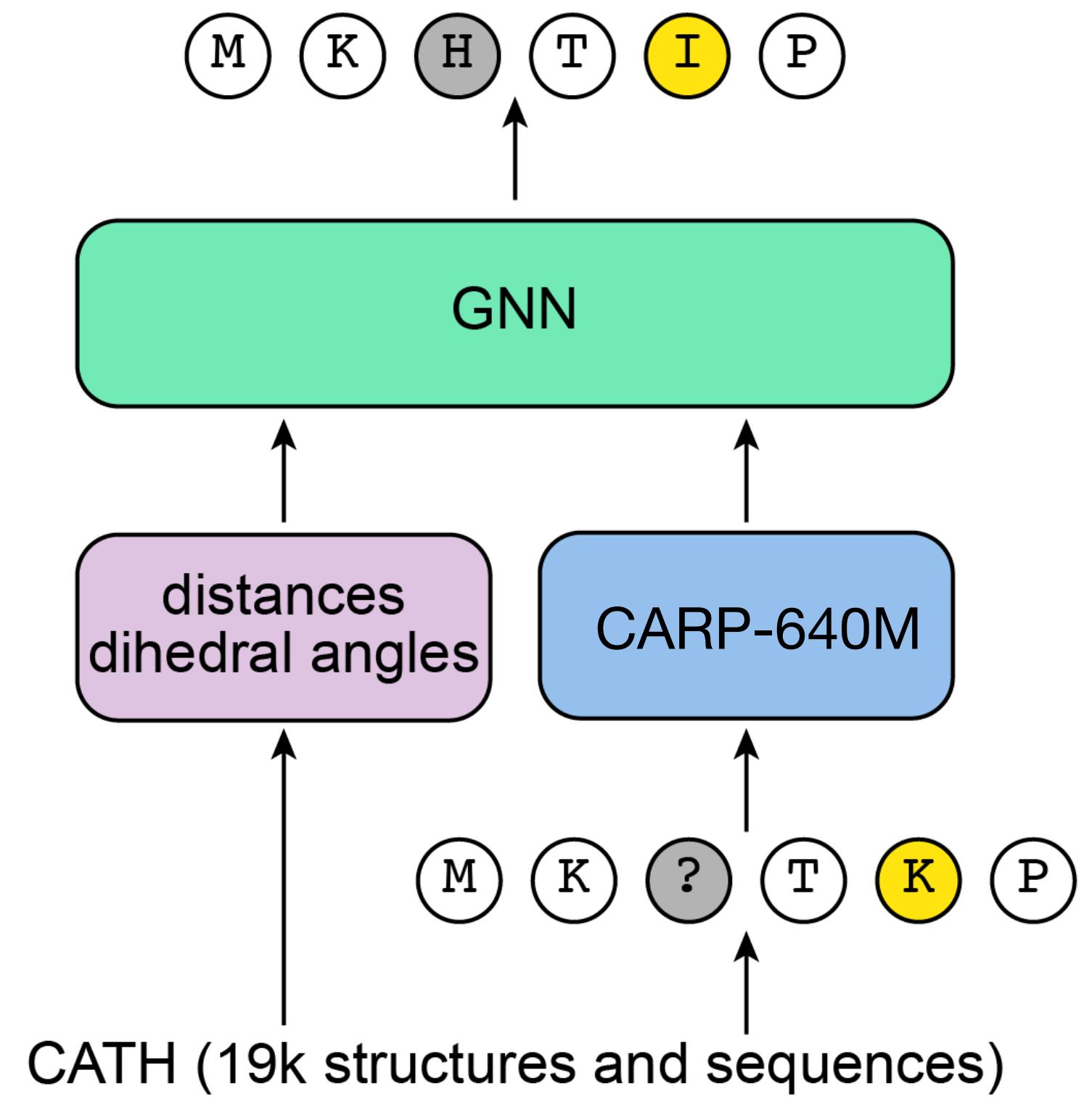
Convolutional autoencoding representations of proteins



Masked inverse folding



Masked inverse folding with sequence transfer



<https://github.com/microsoft/protein-sequence-models>