

Multimodal deep learning for protein engineering

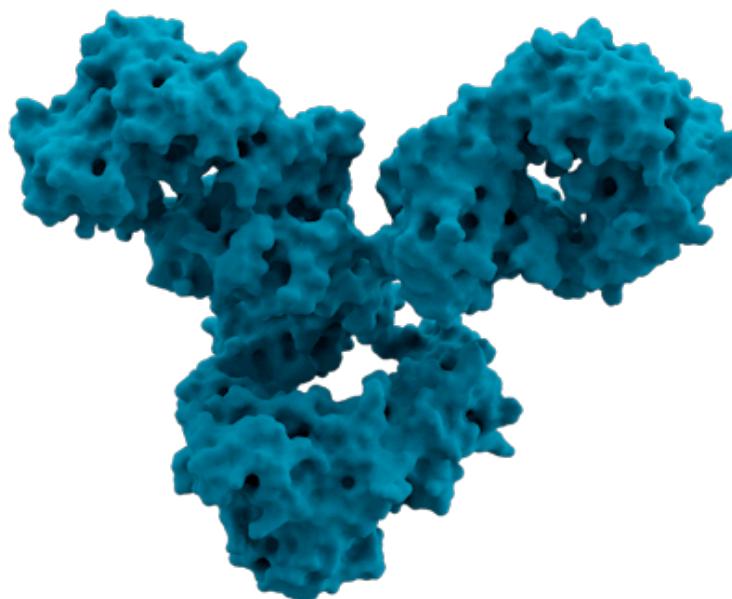
Kevin Kaichuang Yang
Microsoft Research New England
 @KevinKaichuang

Proteins are biology's actuators

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

Proteins are biology's actuators

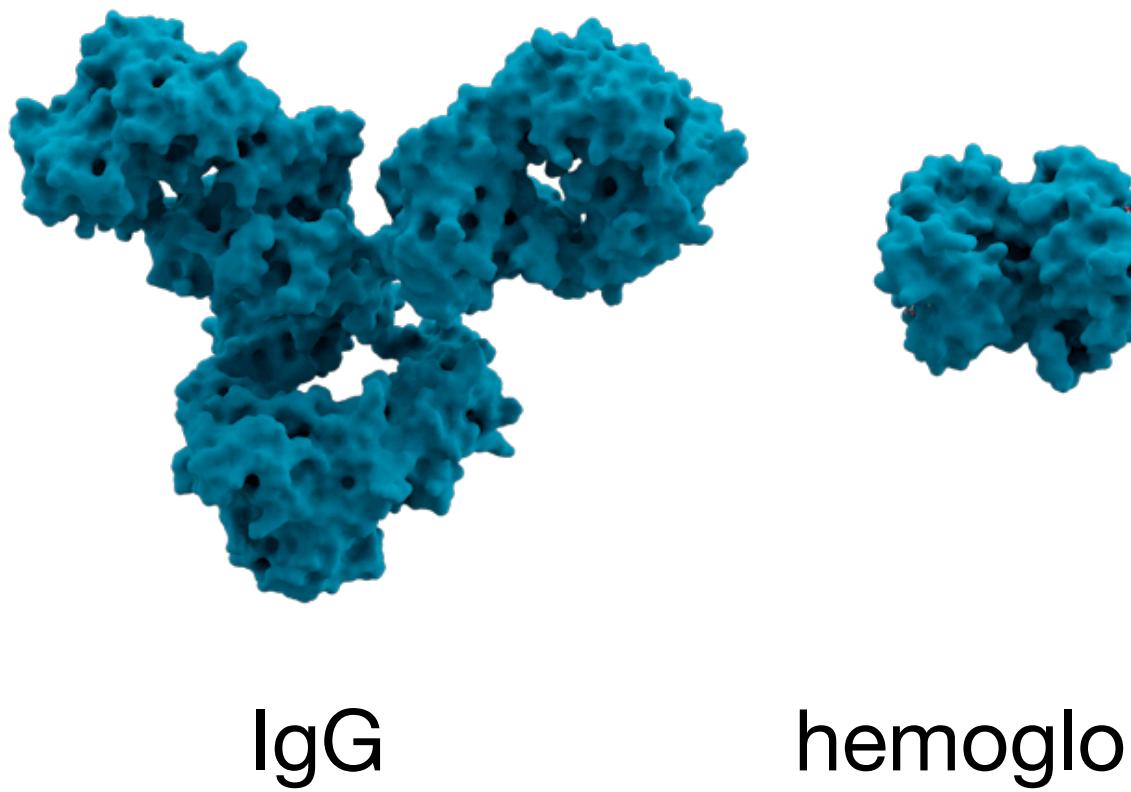
- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling



IgG

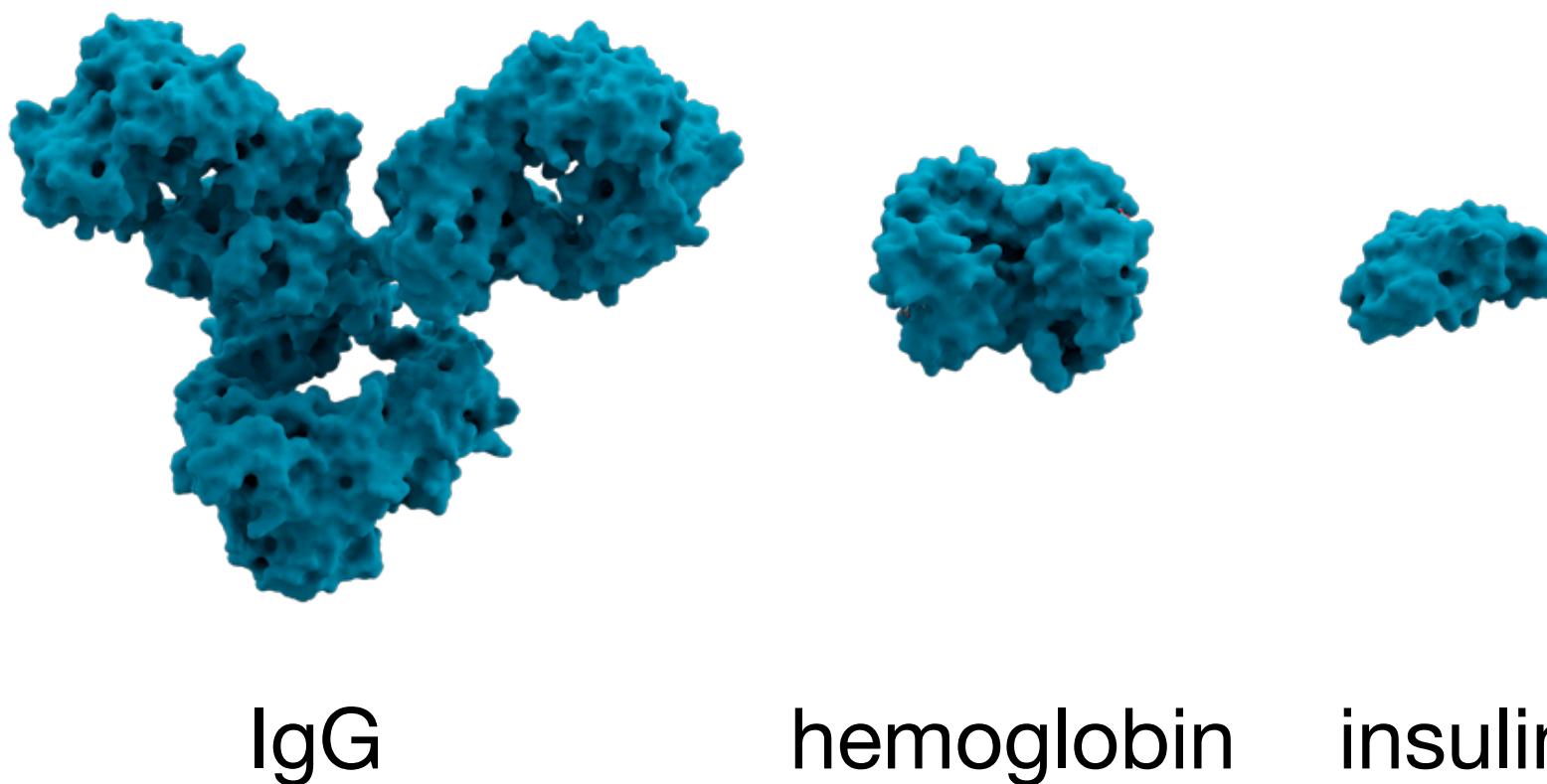
Proteins are biology's actuators

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



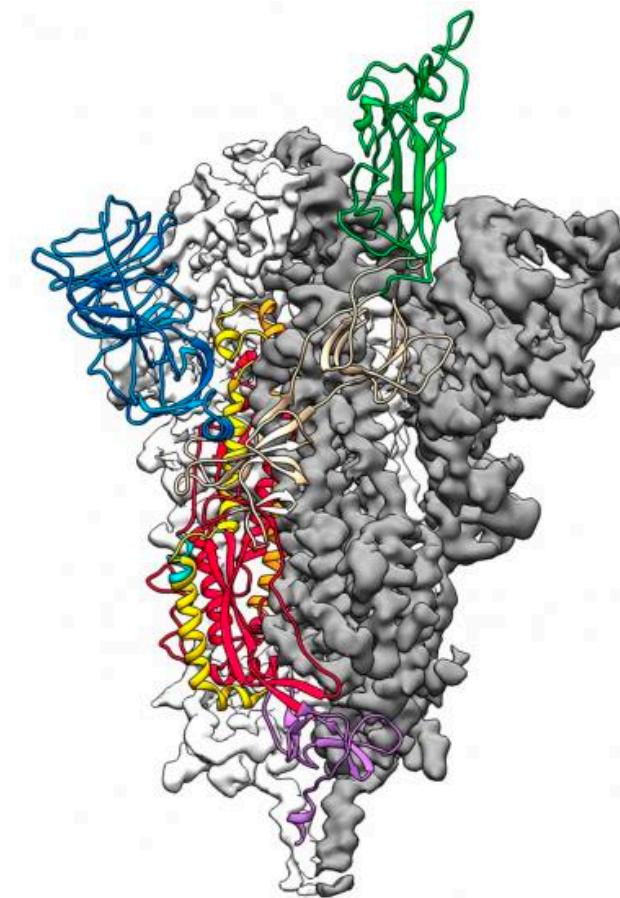
Proteins are biology's actuators

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

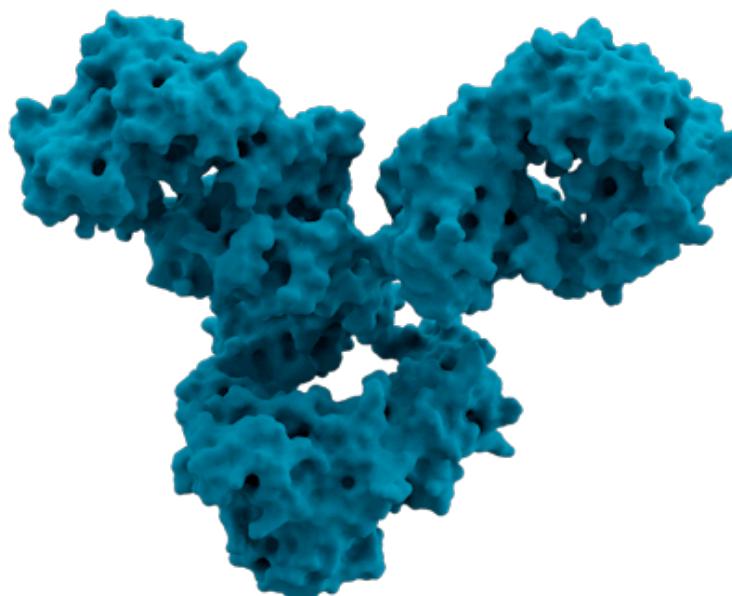


Proteins are biology's actuators

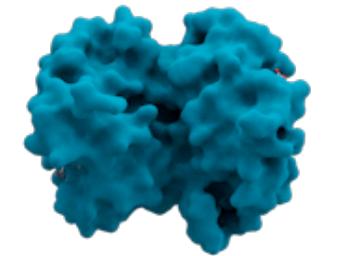
- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling



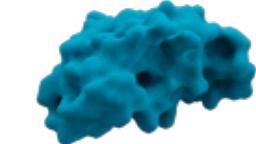
coronavirus spike protein



IgG



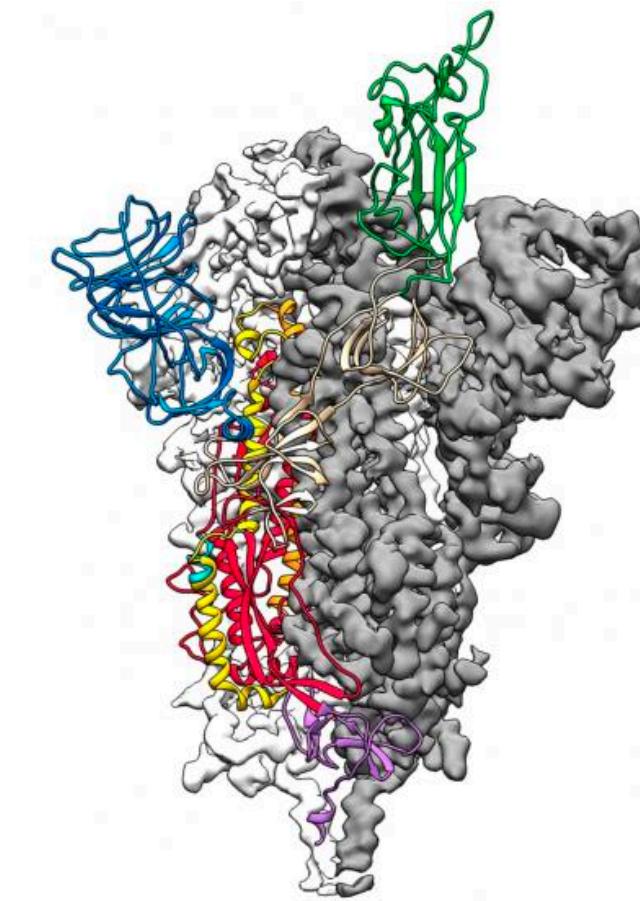
hemoglobin



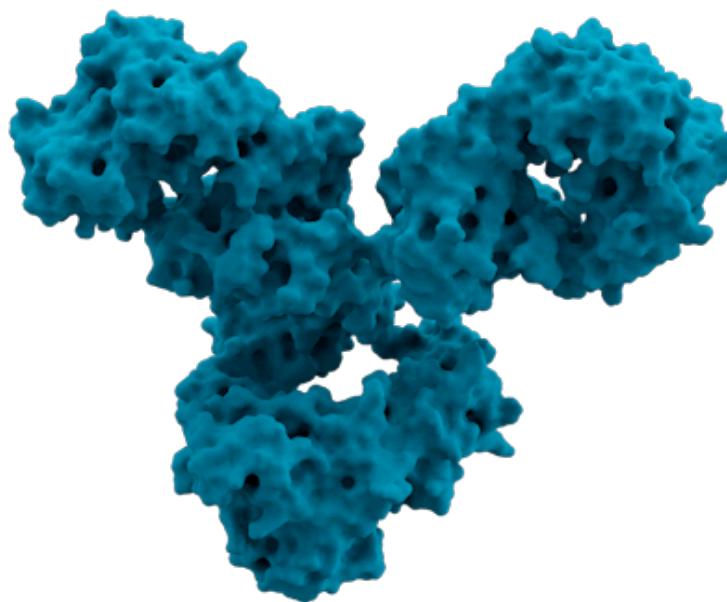
insulin

Proteins are biology's actuators

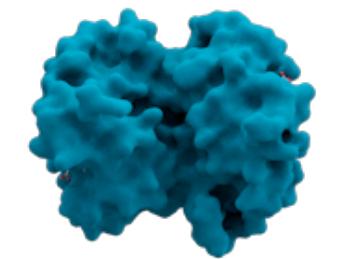
- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling



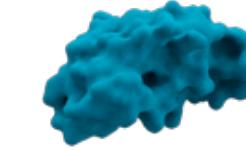
coronavirus spike protein



IgG



hemoglobin

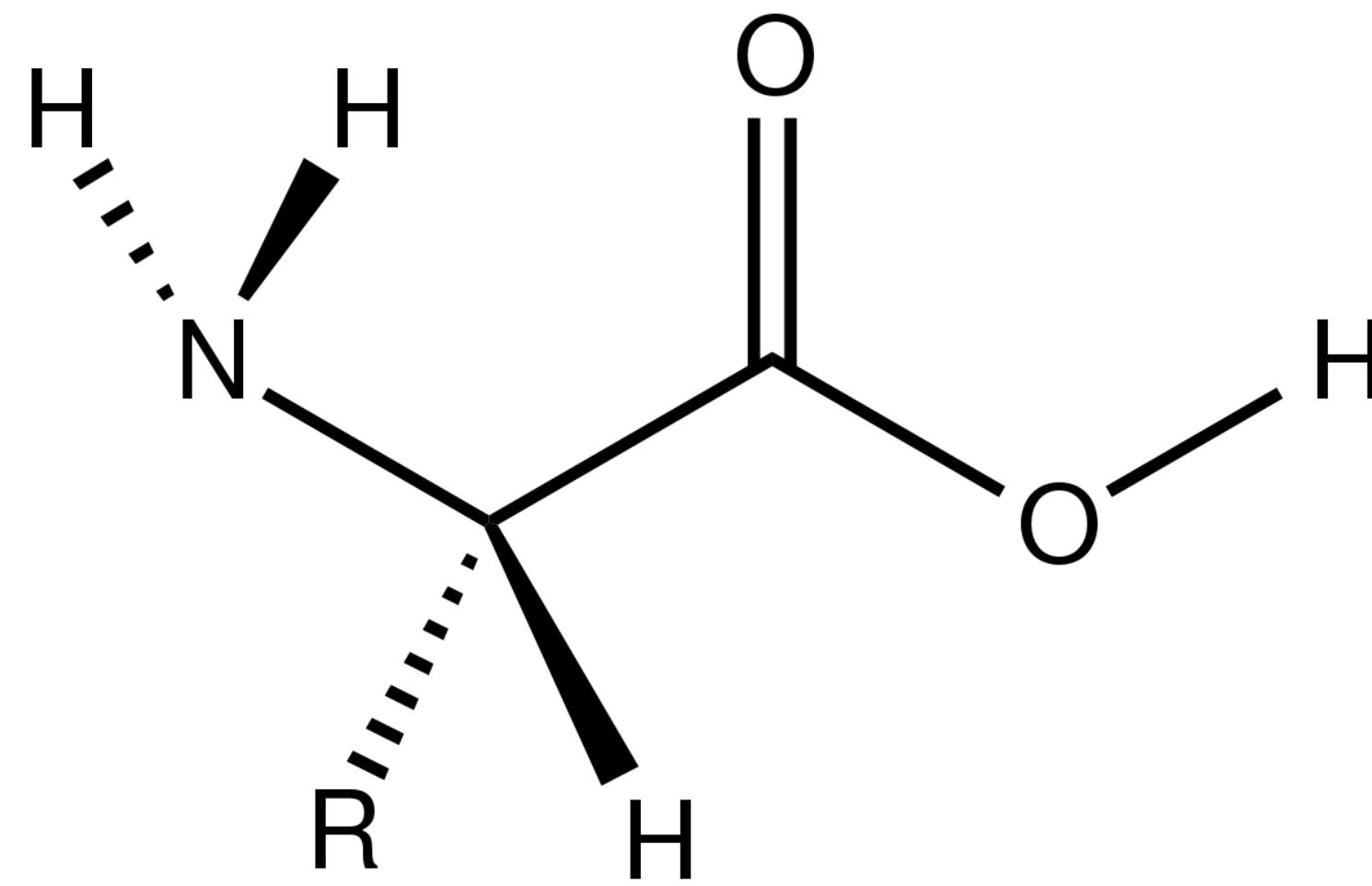


insulin

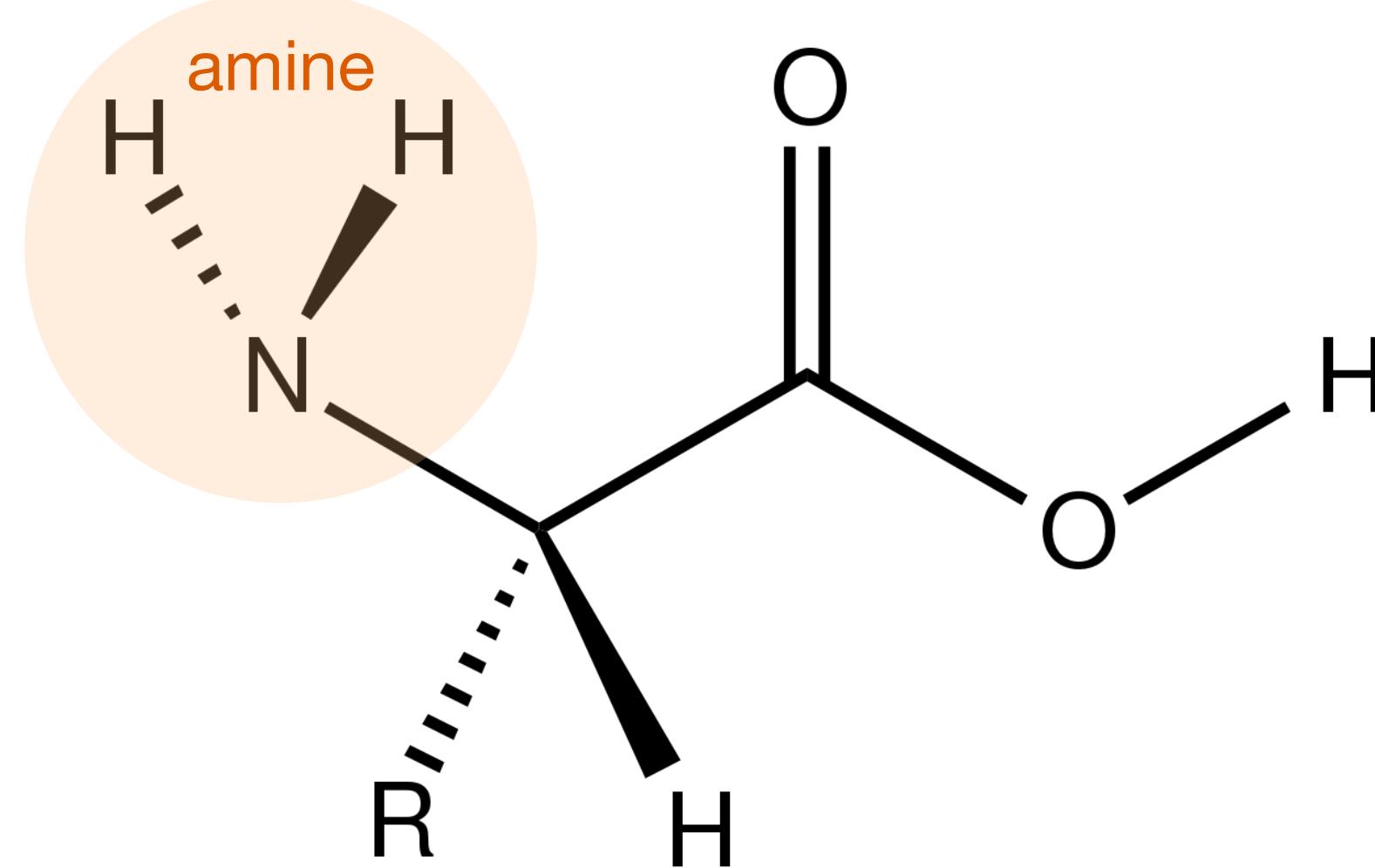


luciferase

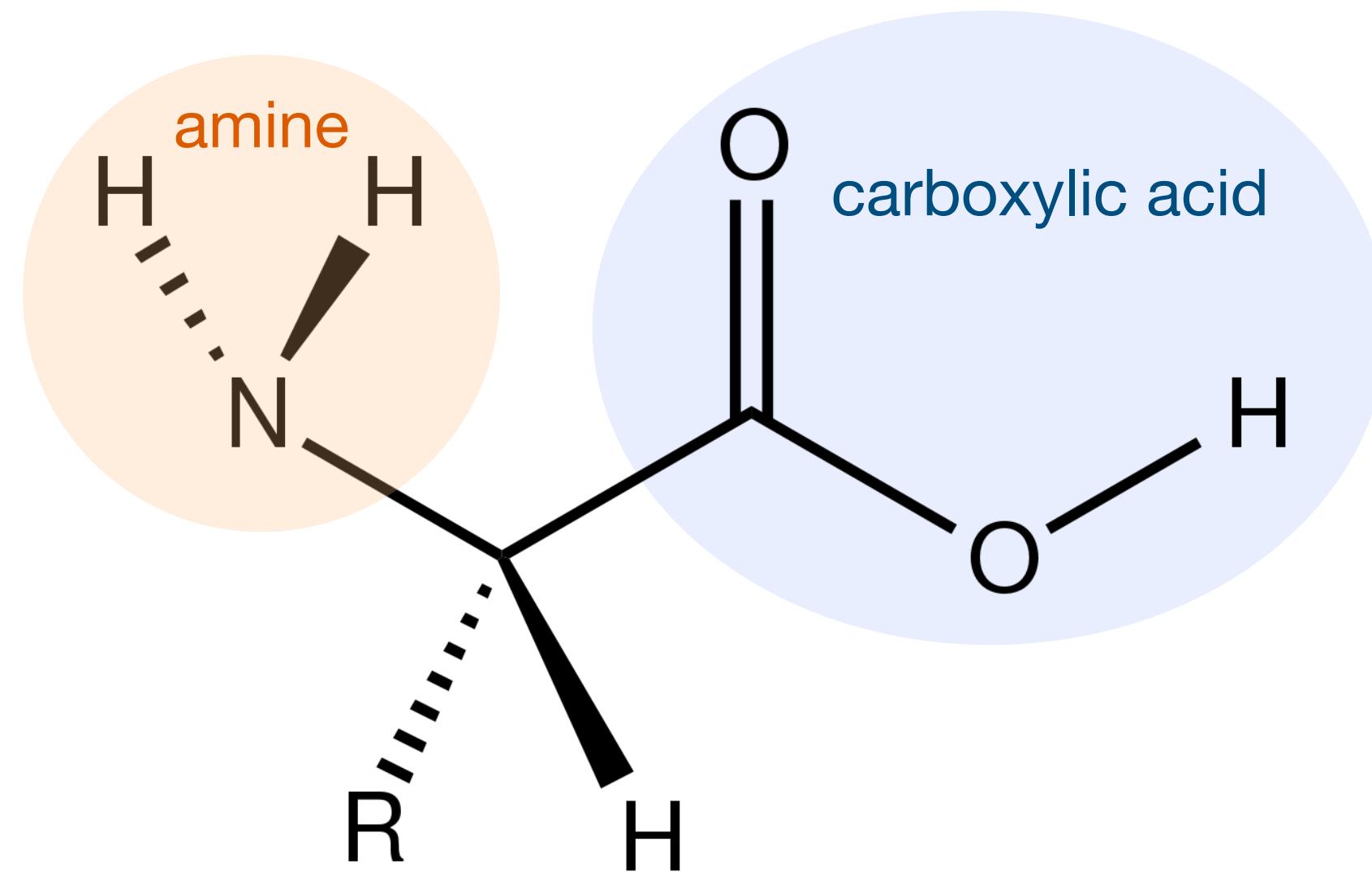
Diversity arises from 20 building blocks



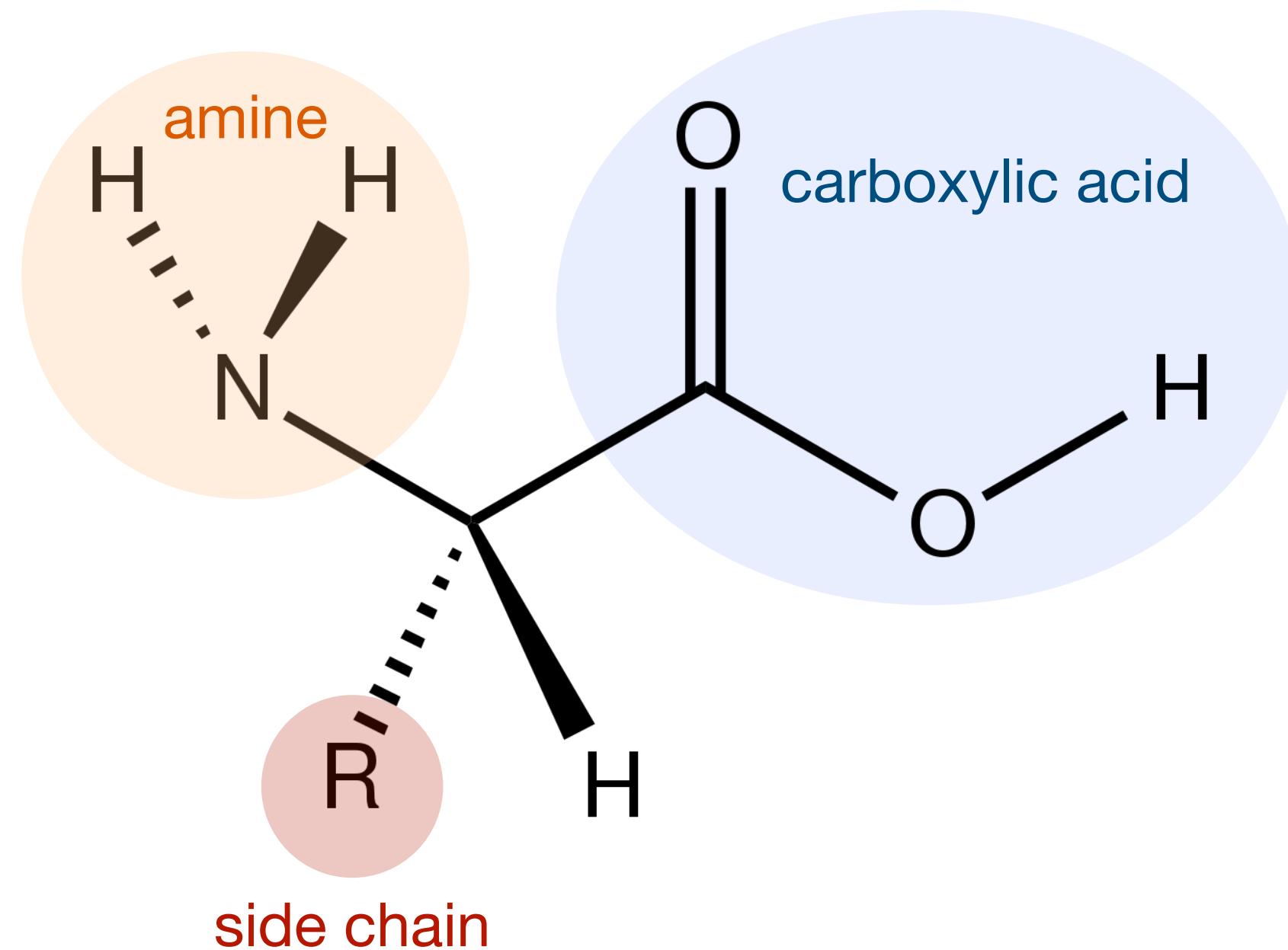
Diversity arises from 20 building blocks



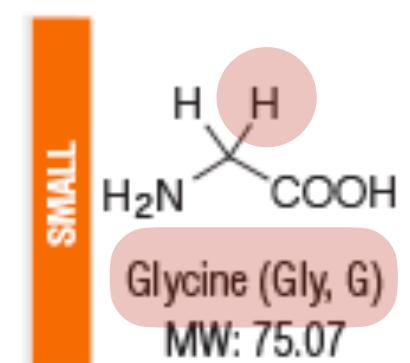
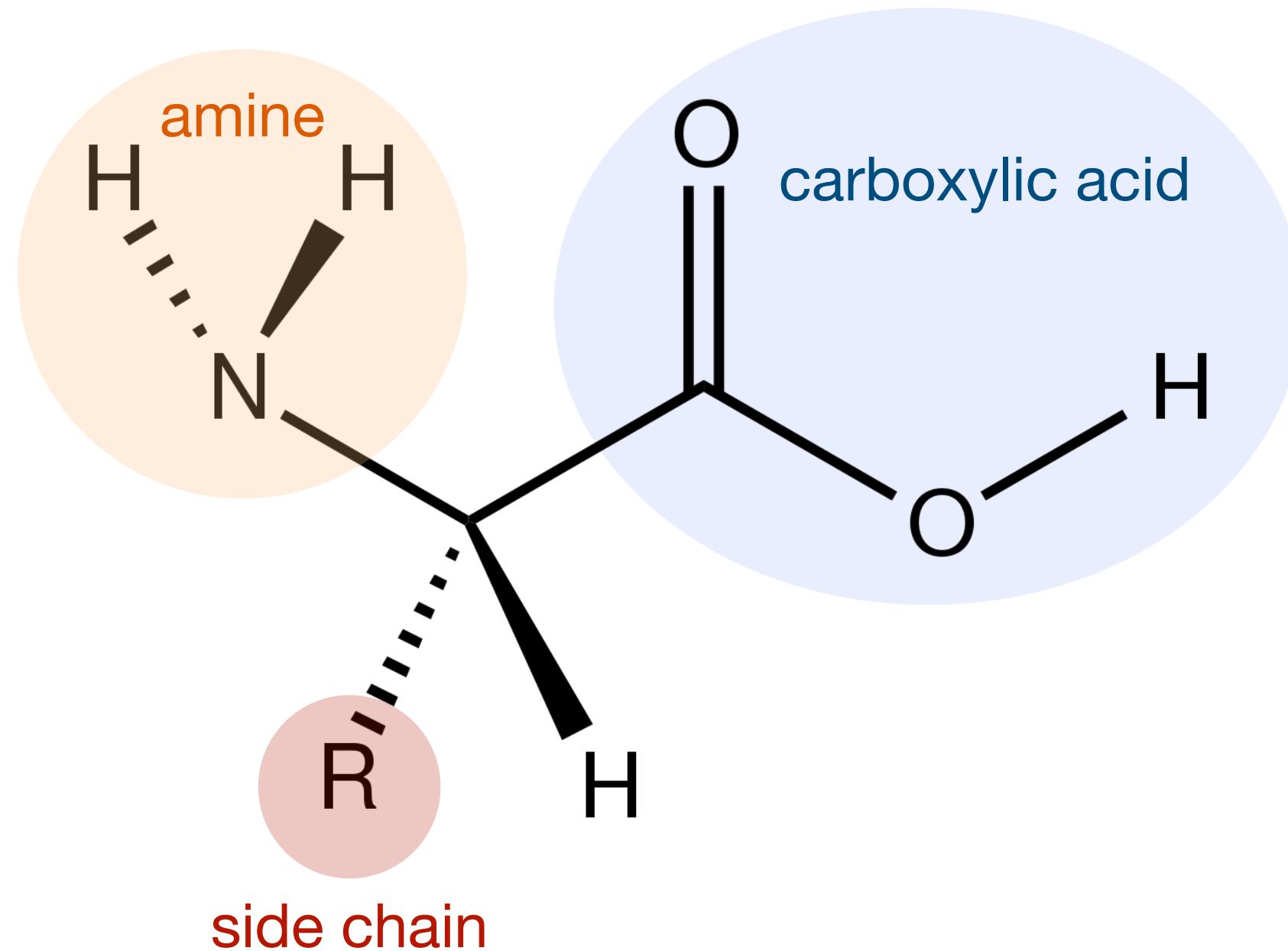
Diversity arises from 20 building blocks



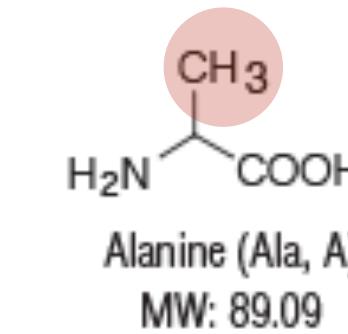
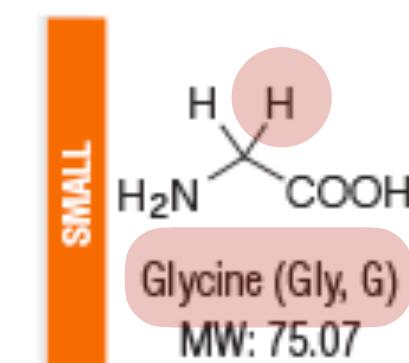
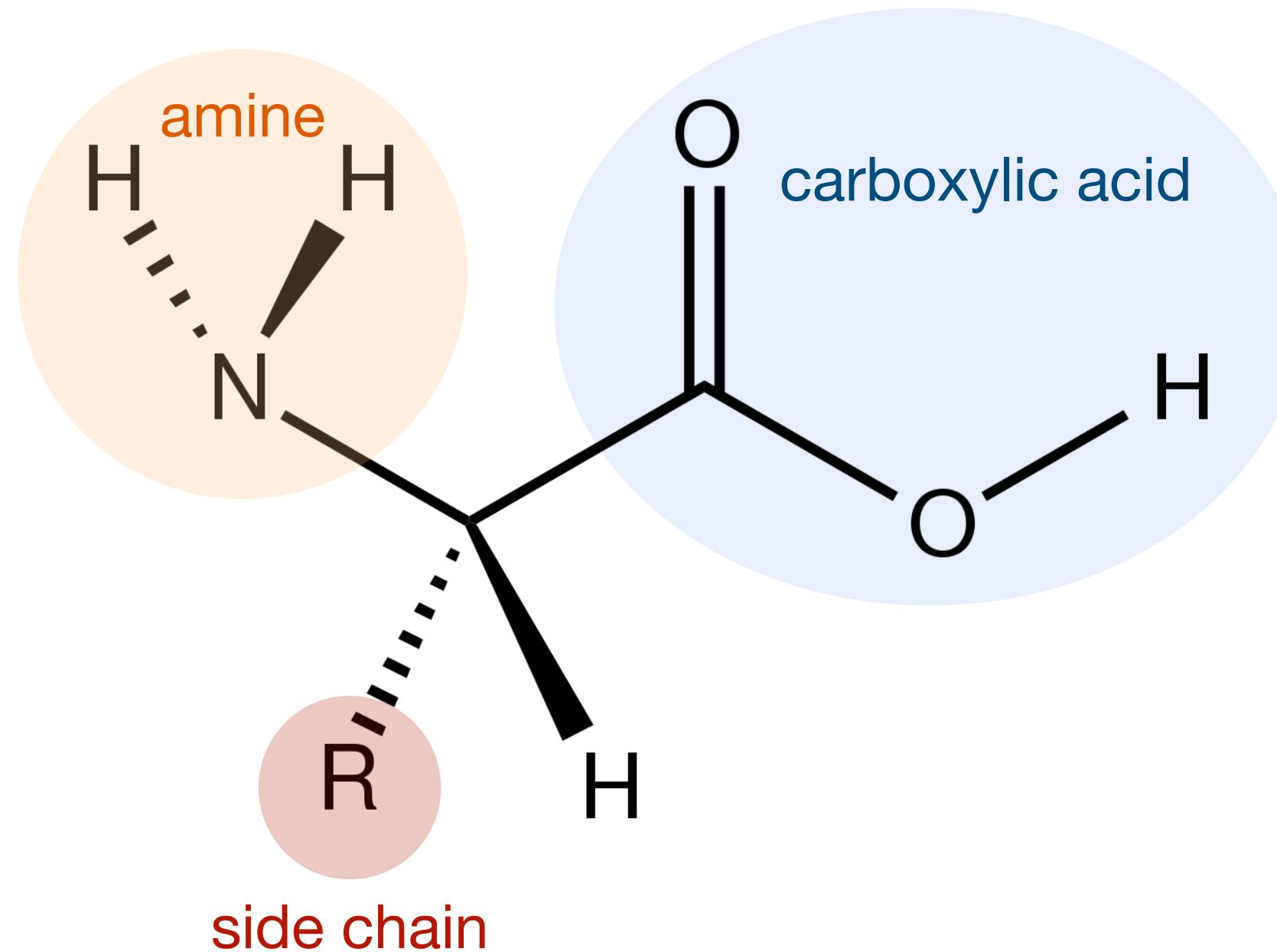
Diversity arises from 20 building blocks



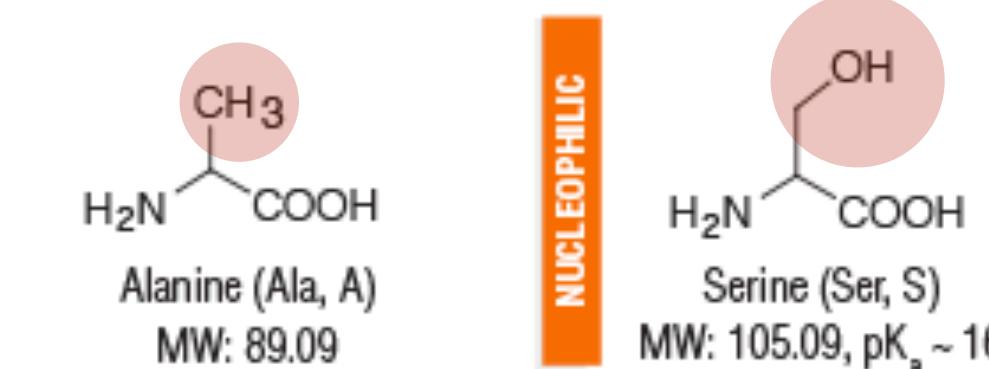
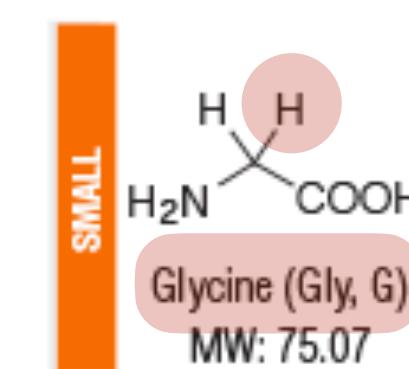
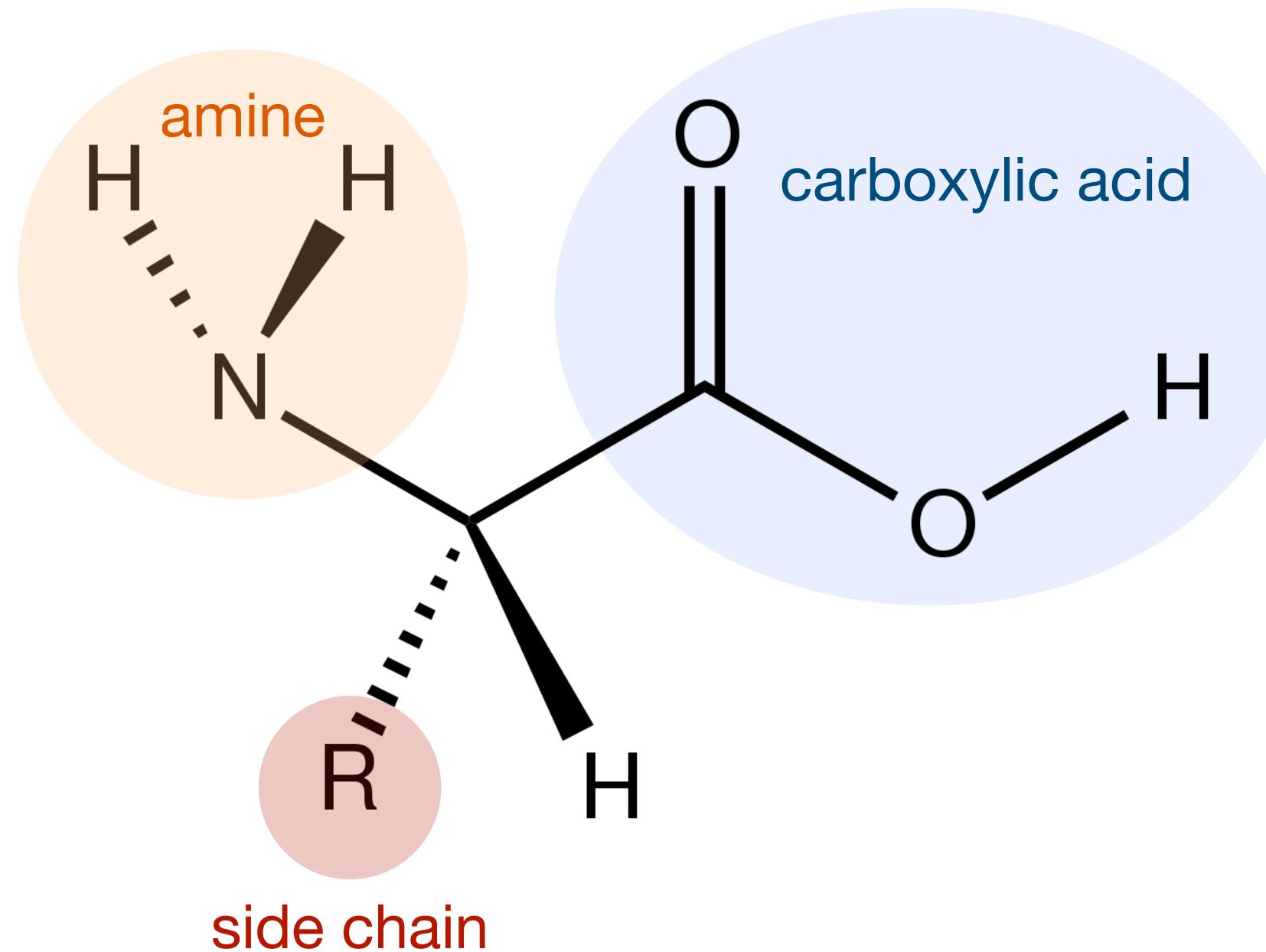
Diversity arises from 20 building blocks



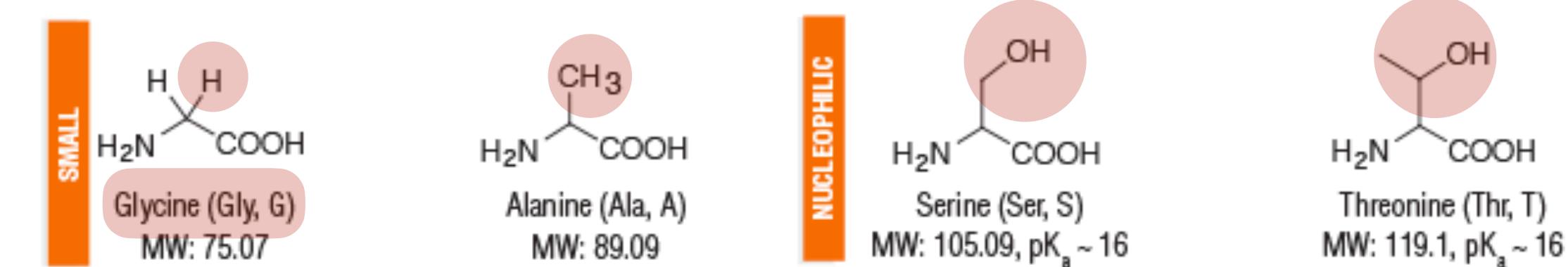
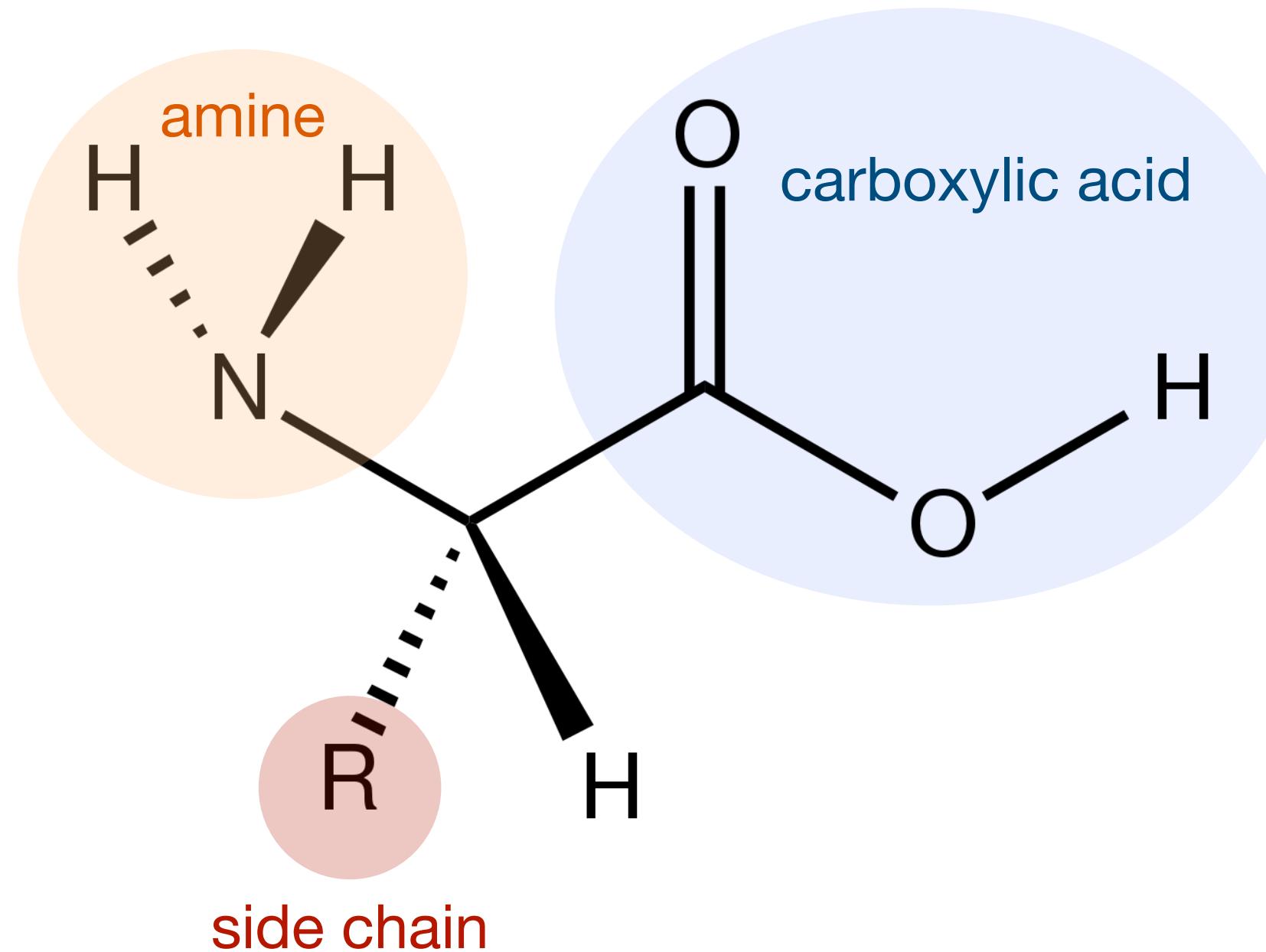
Diversity arises from 20 building blocks



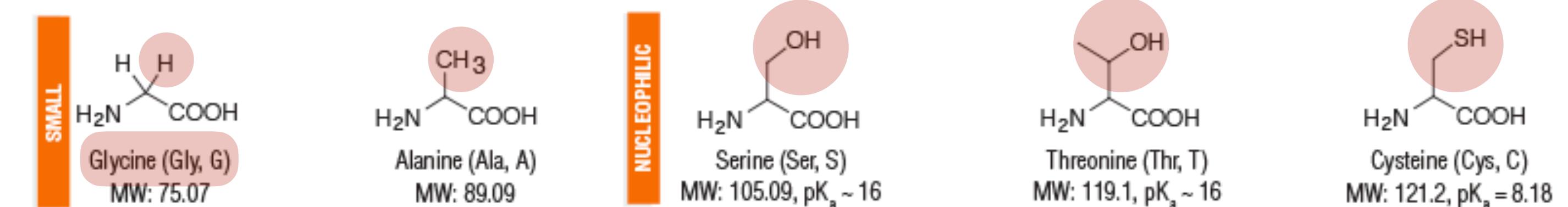
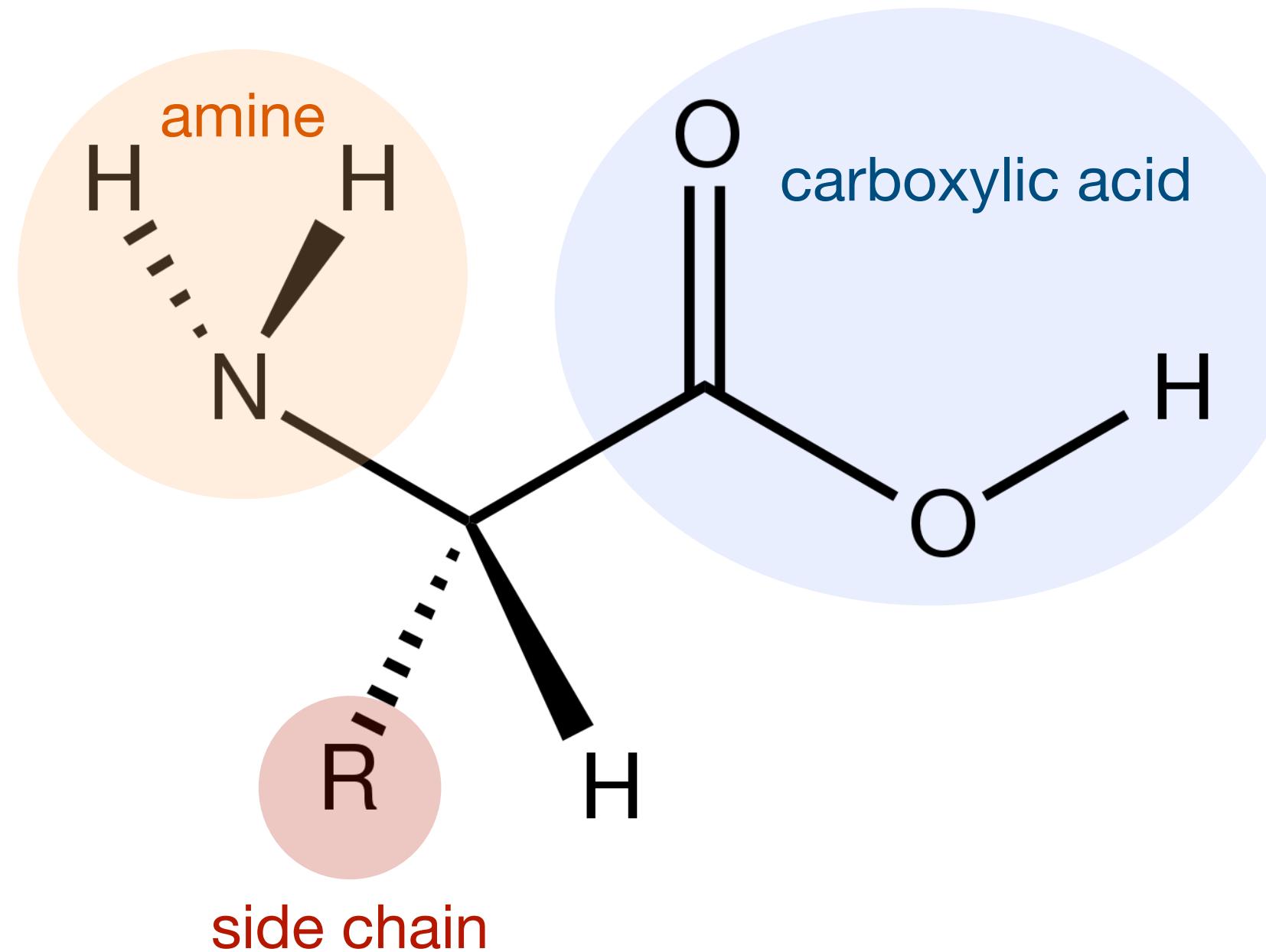
Diversity arises from 20 building blocks



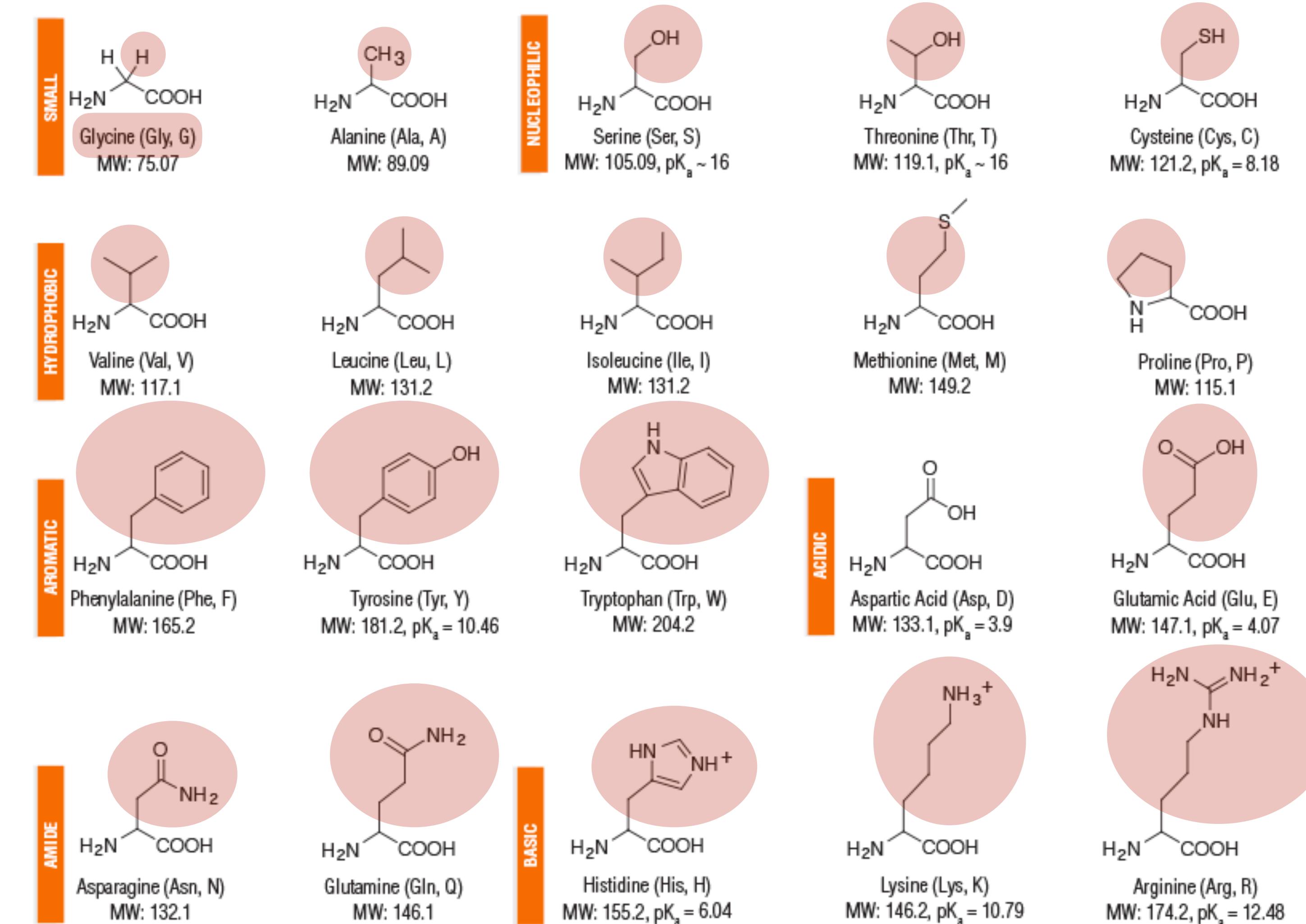
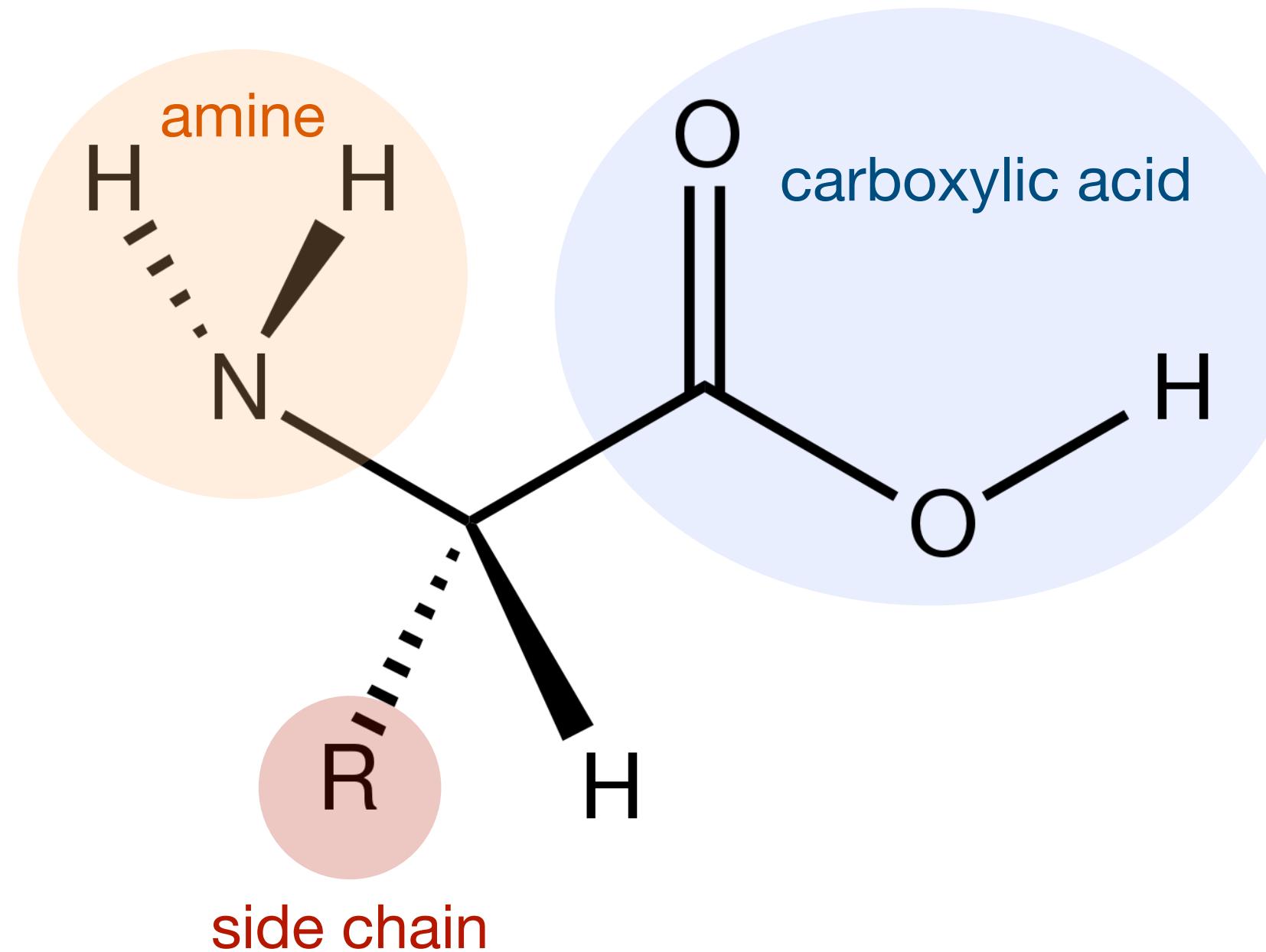
Diversity arises from 20 building blocks



Diversity arises from 20 building blocks

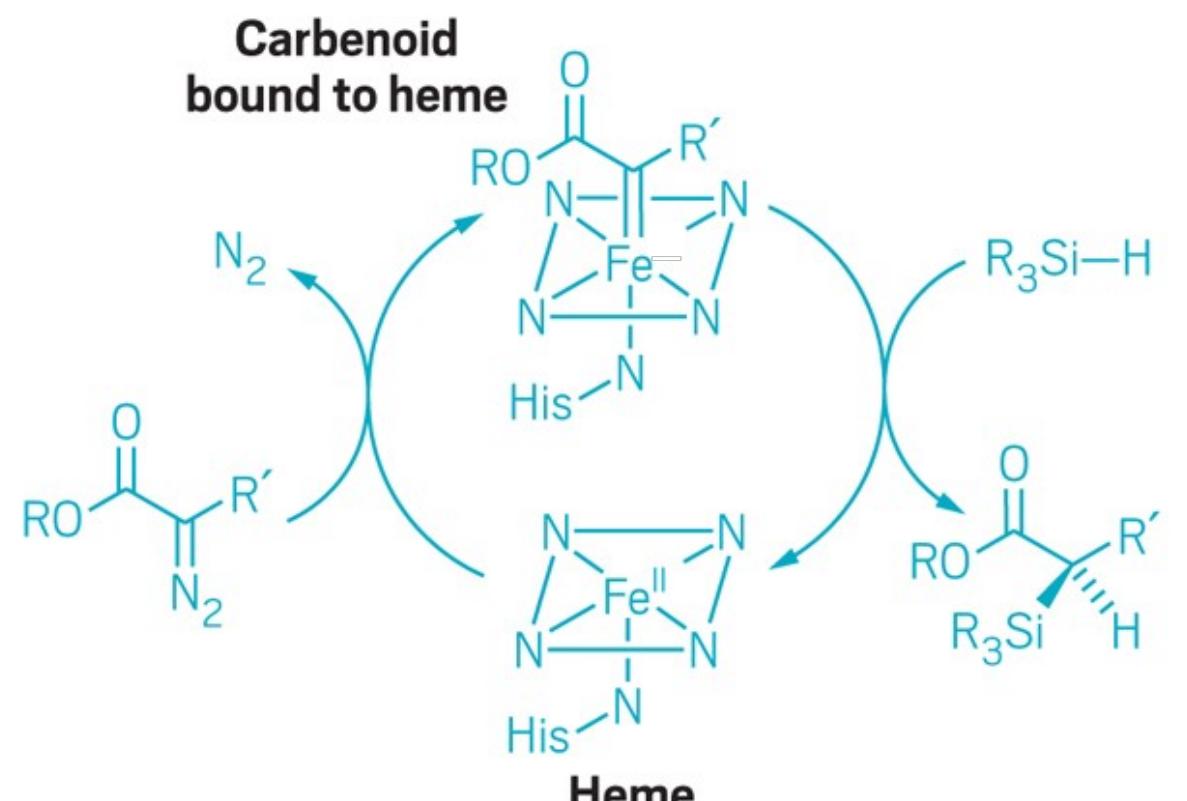


Diversity arises from 20 building blocks



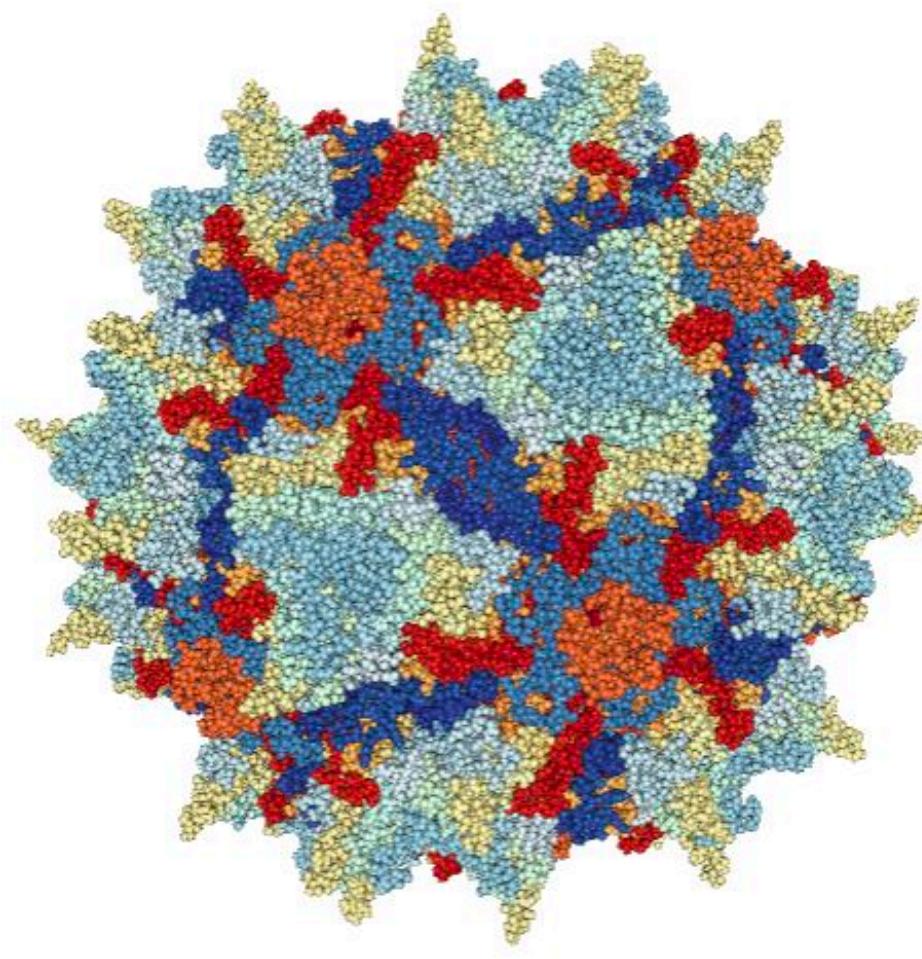
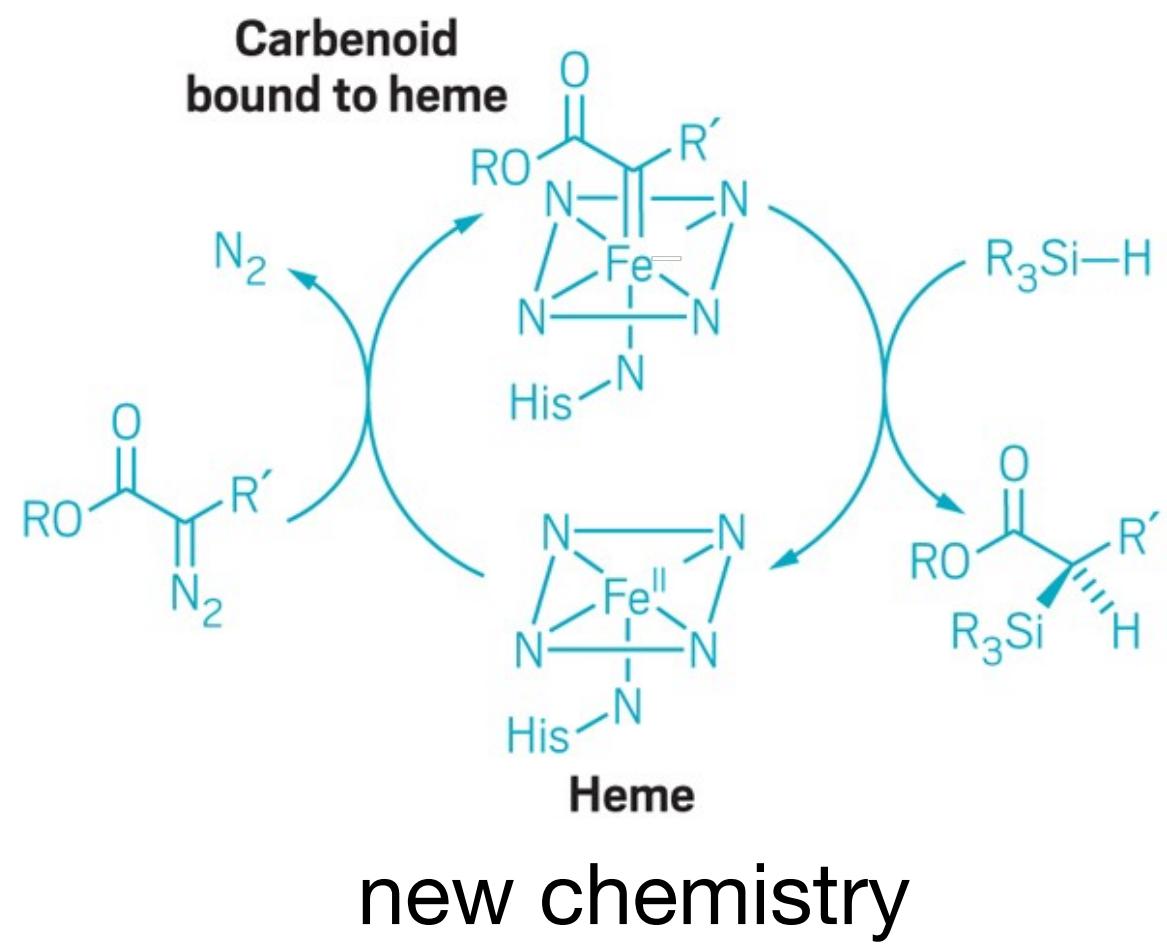
Why design proteins?

Why design proteins?

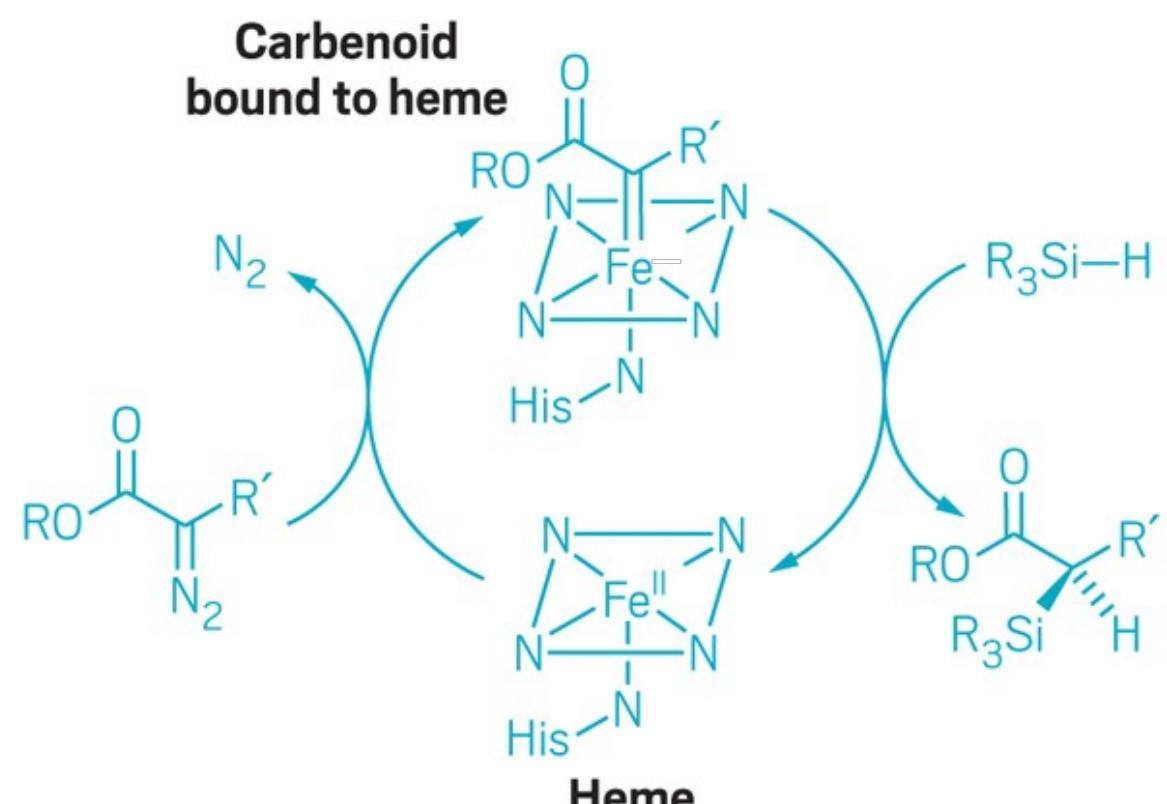


new chemistry

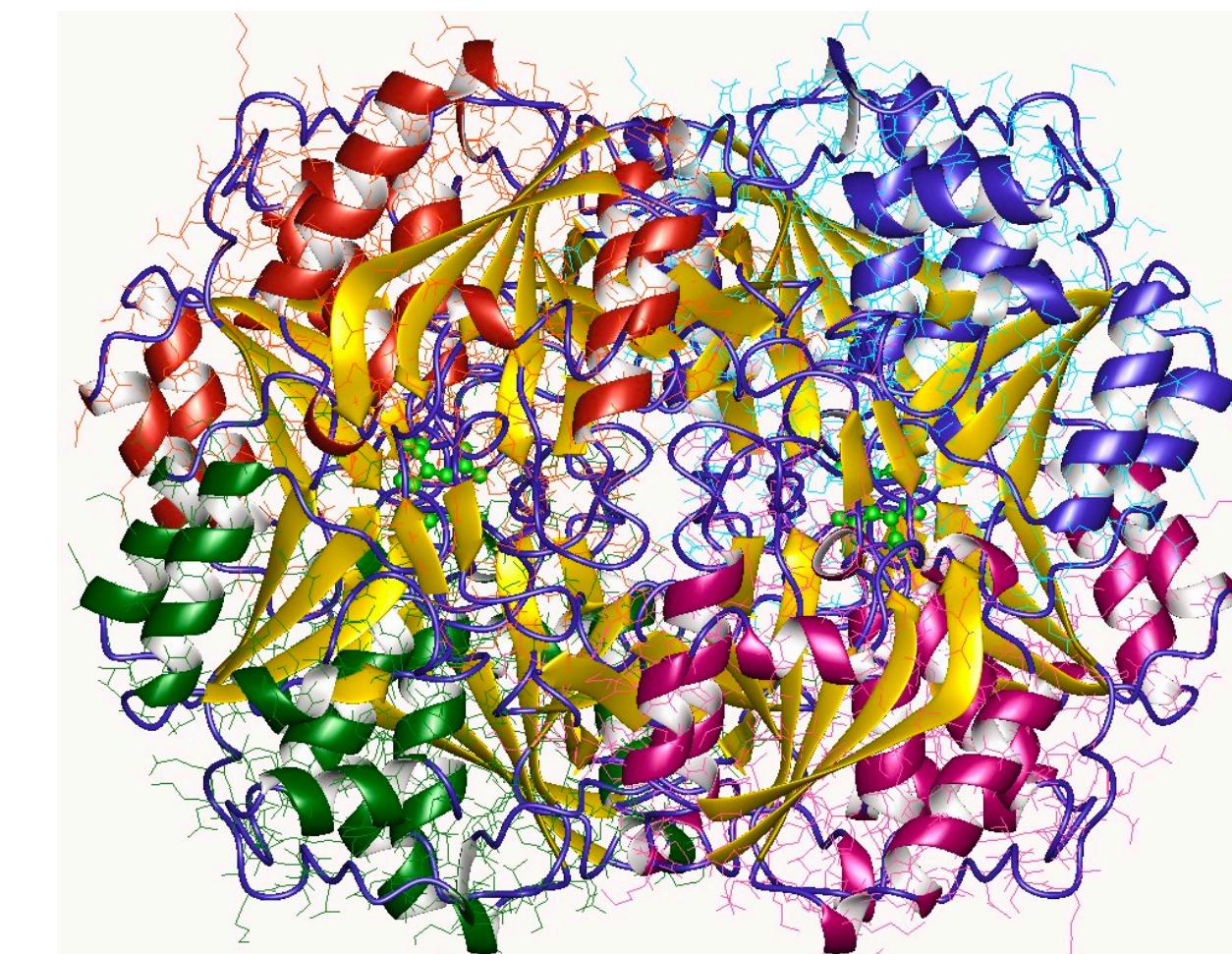
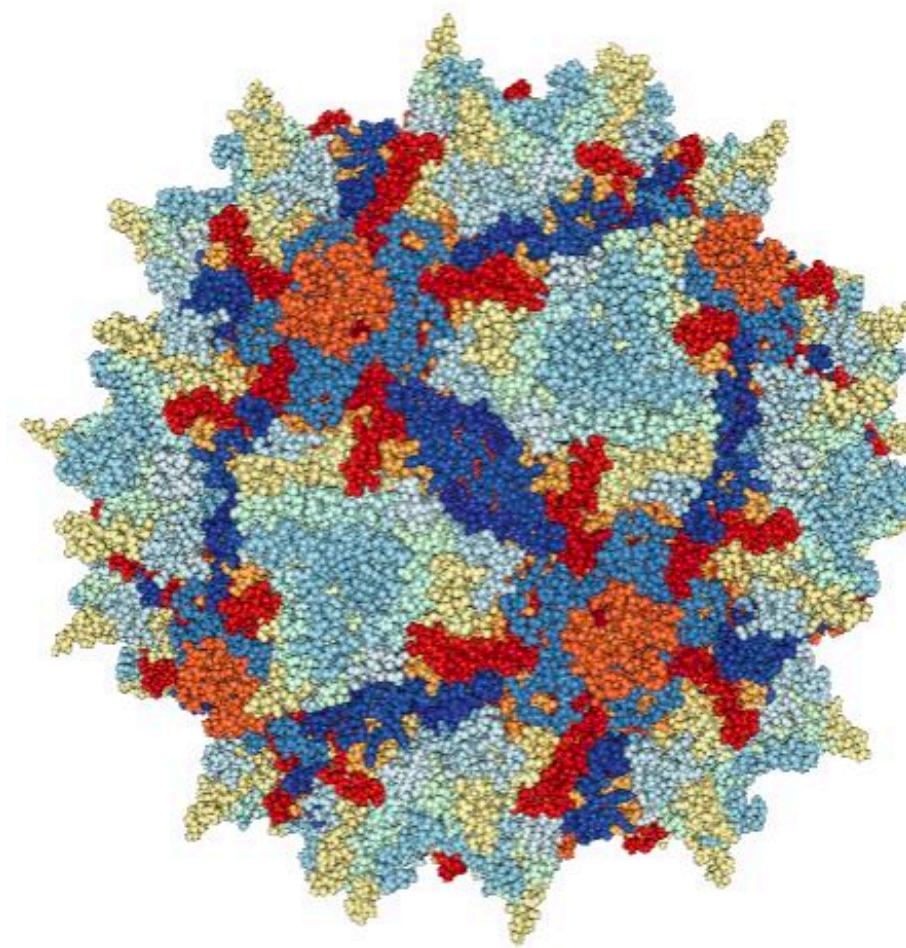
Why design proteins?



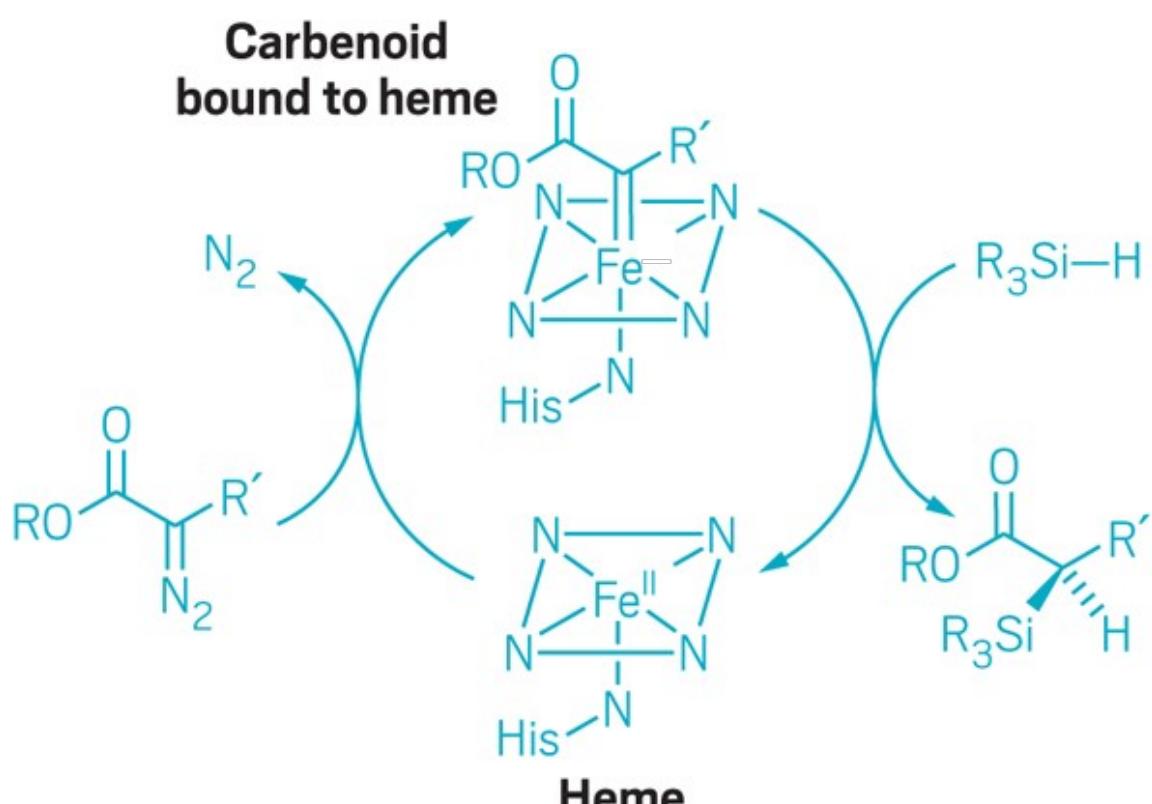
Why design proteins?



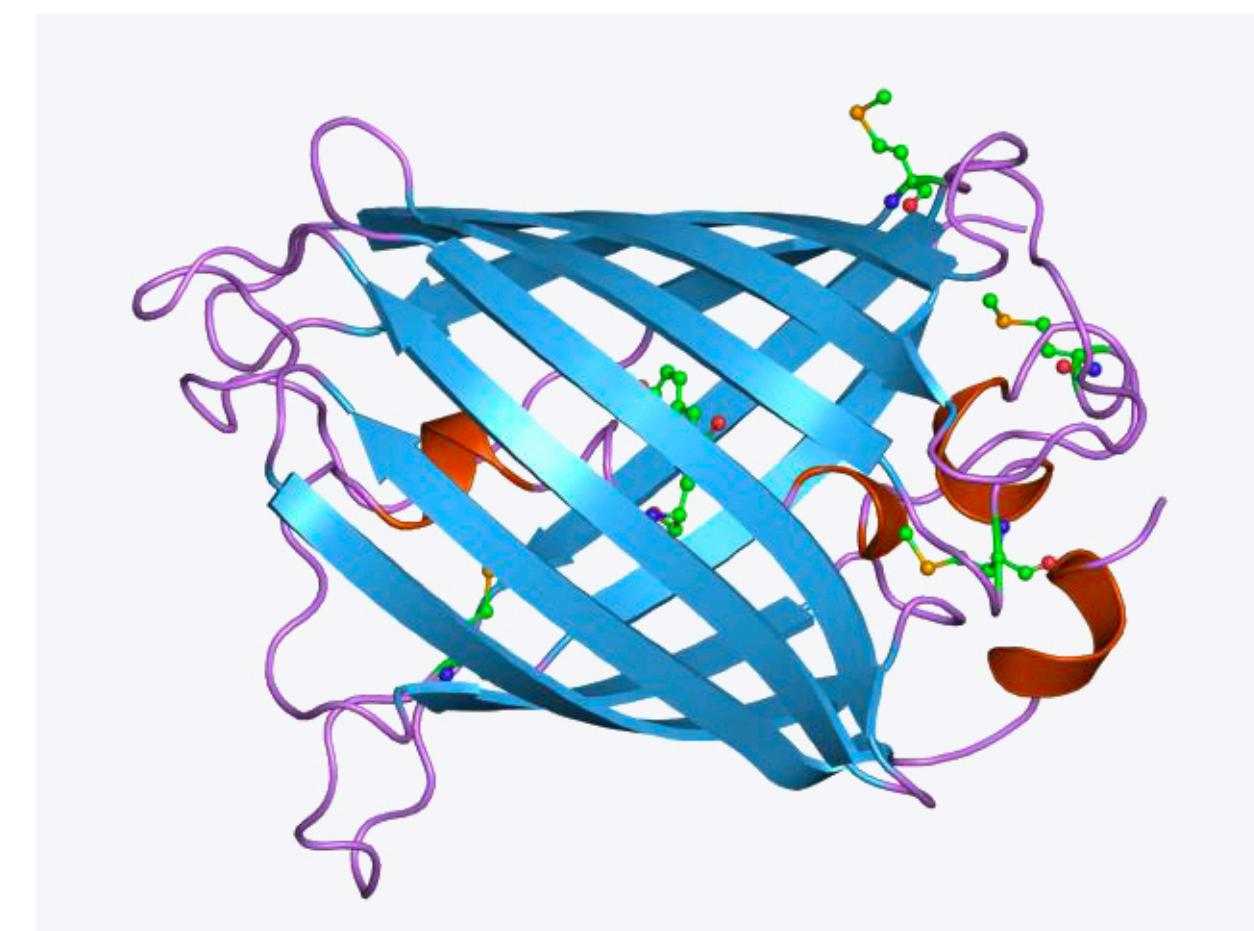
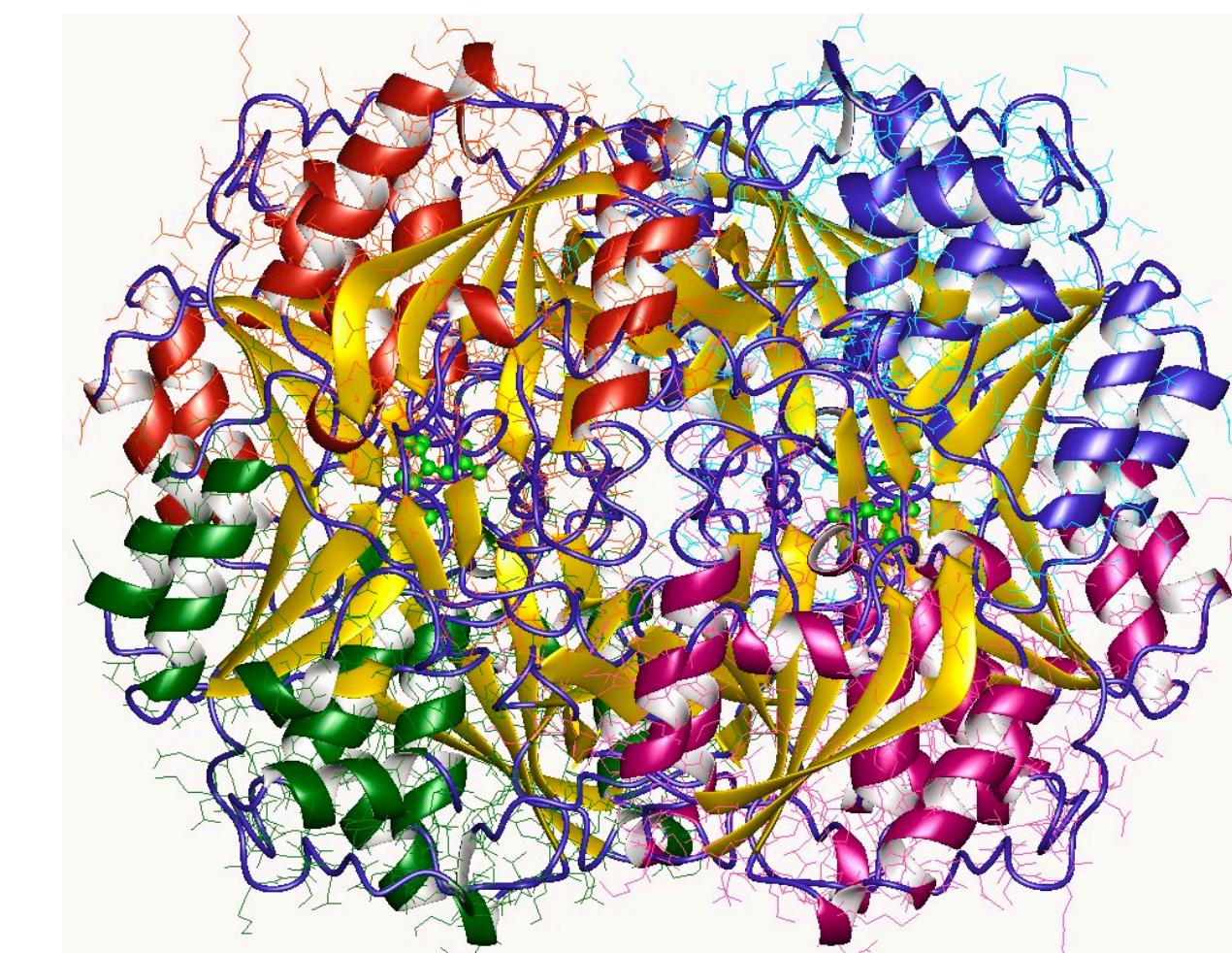
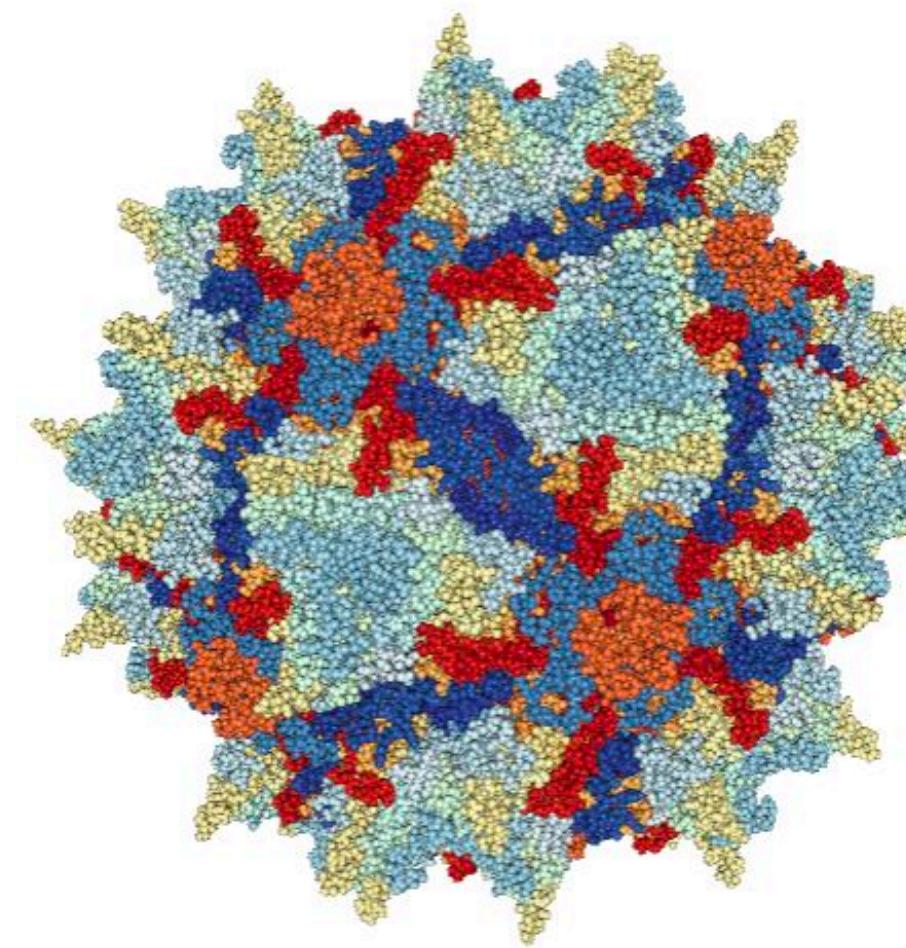
new chemistry



Why design proteins?



new chemistry

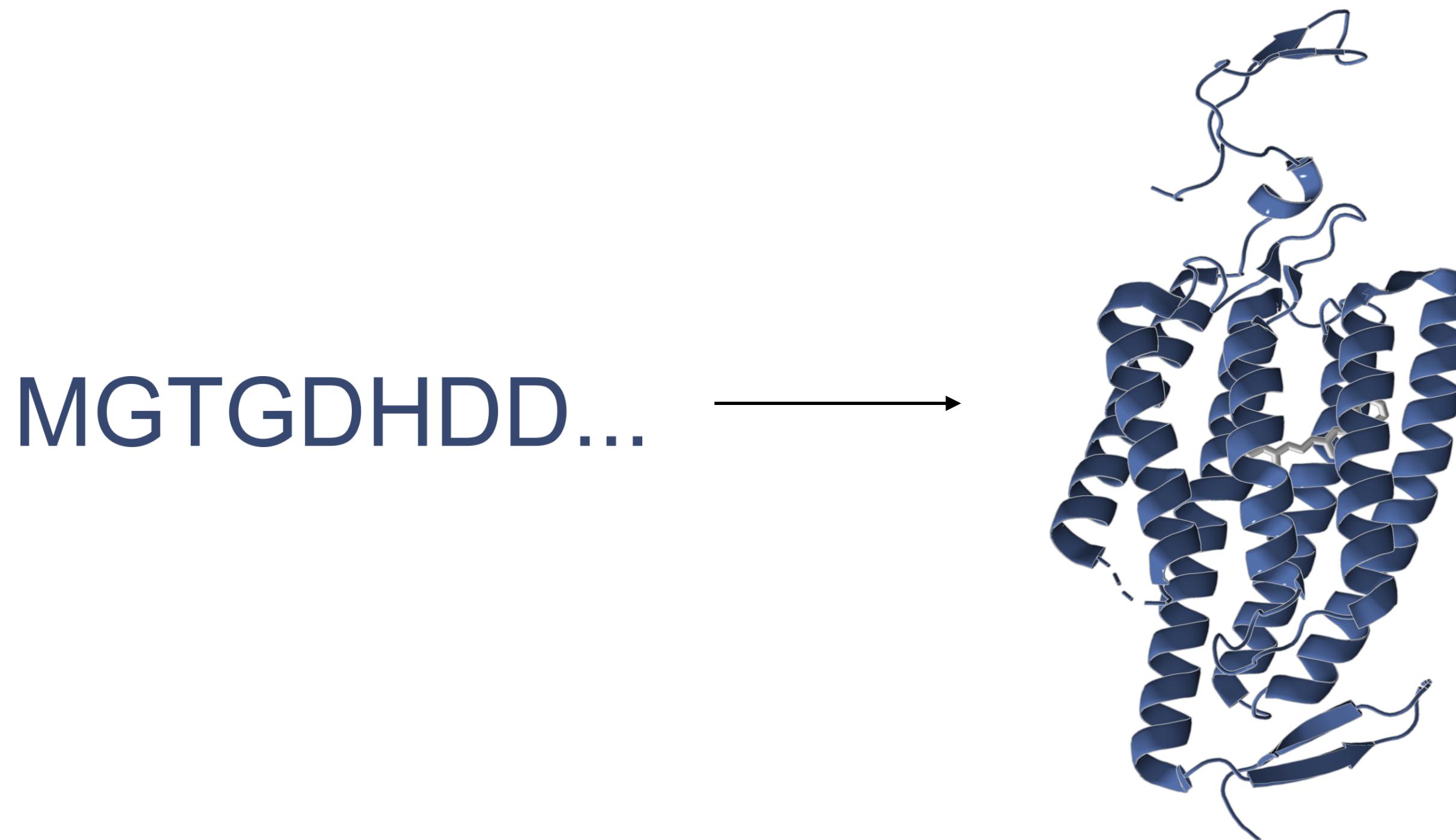


Protein engineering requires going from function to sequence

Protein engineering requires going from function to sequence

MGTGDHDD...

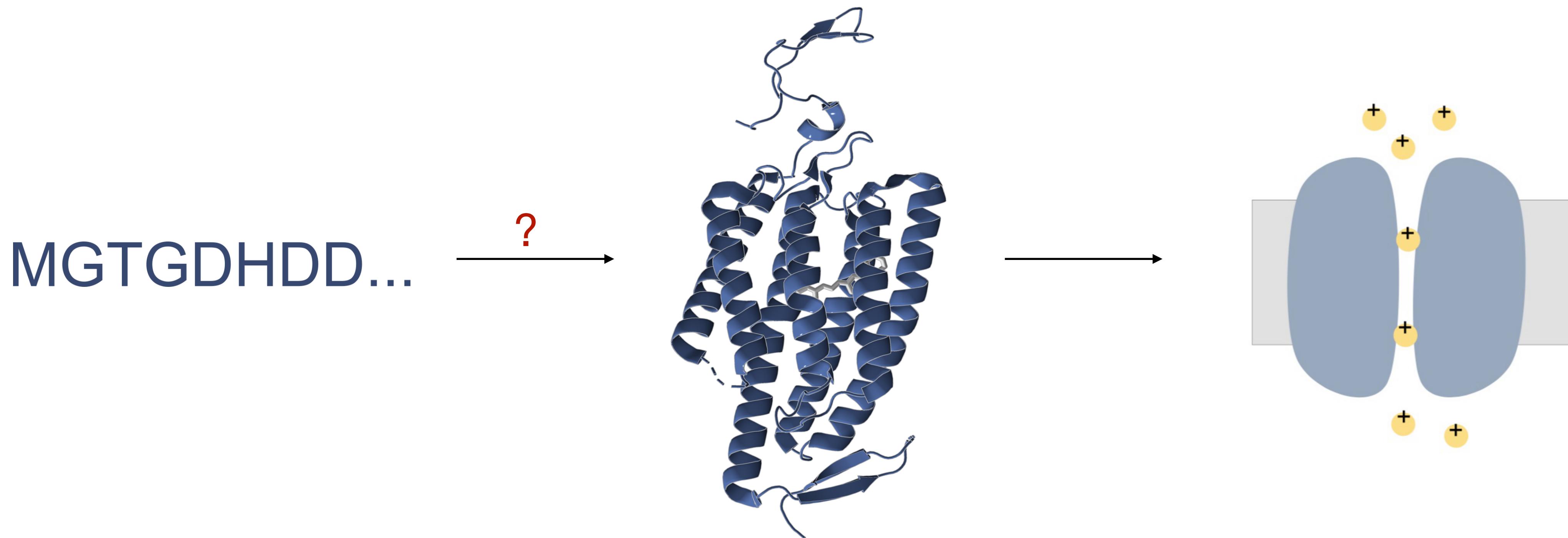
Protein engineering requires going from function to sequence



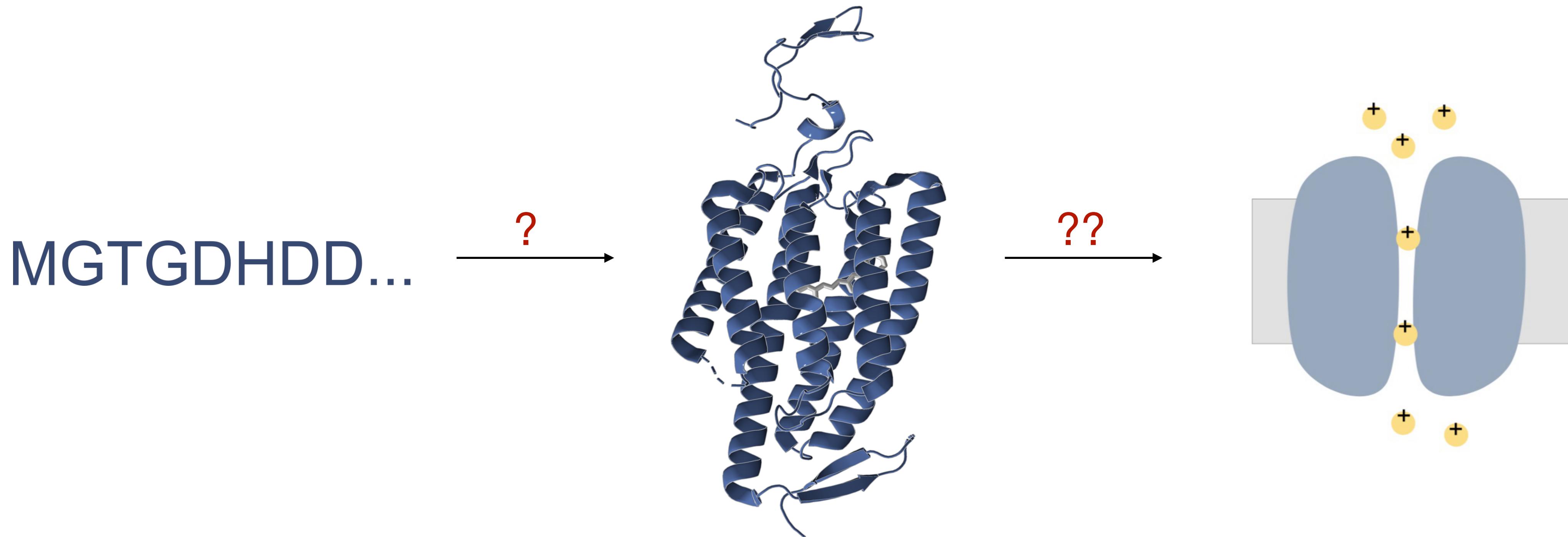
Protein engineering requires going from function to sequence



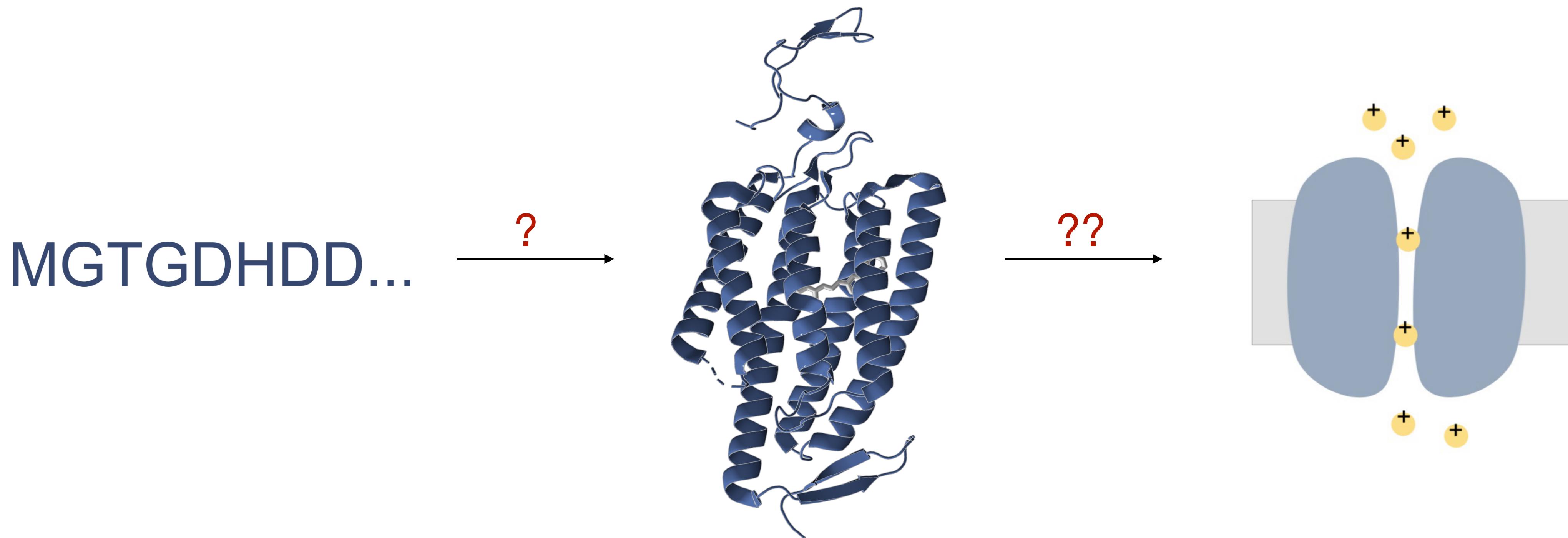
Protein engineering requires going from function to sequence



Protein engineering requires going from function to sequence

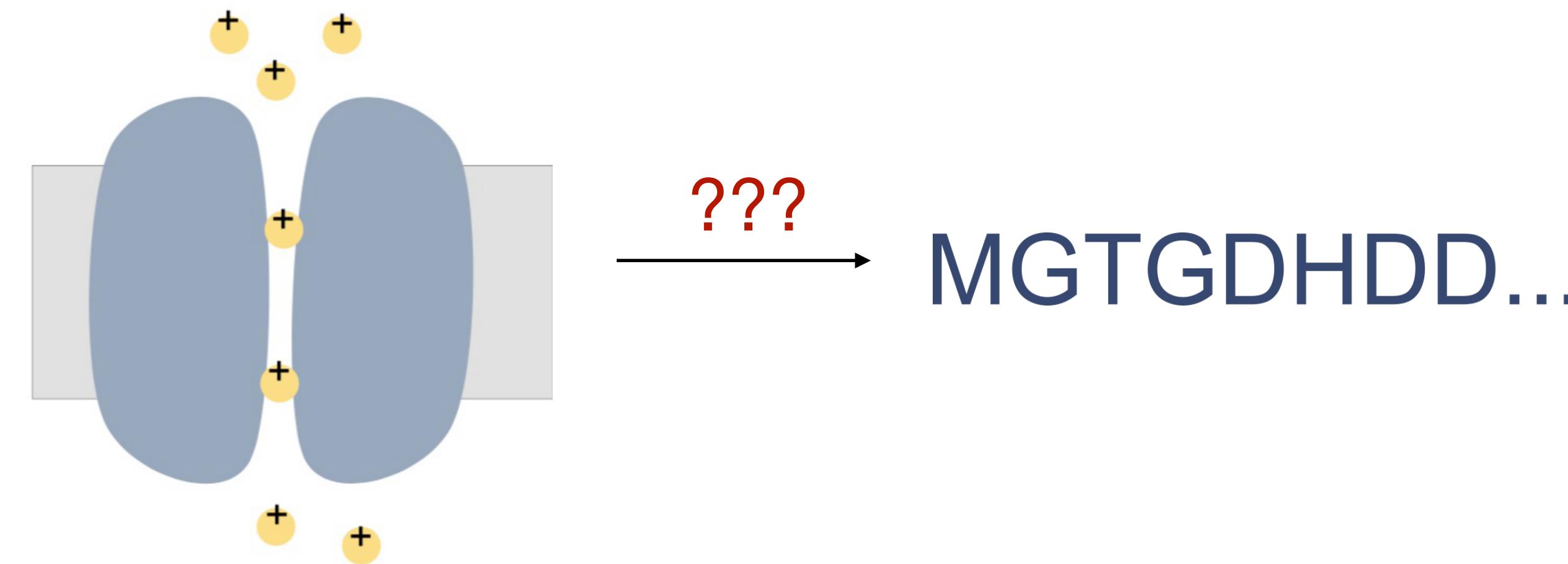


Protein engineering requires going from function to sequence



What sequence will give the desired function?

Protein engineering requires going from function to sequence

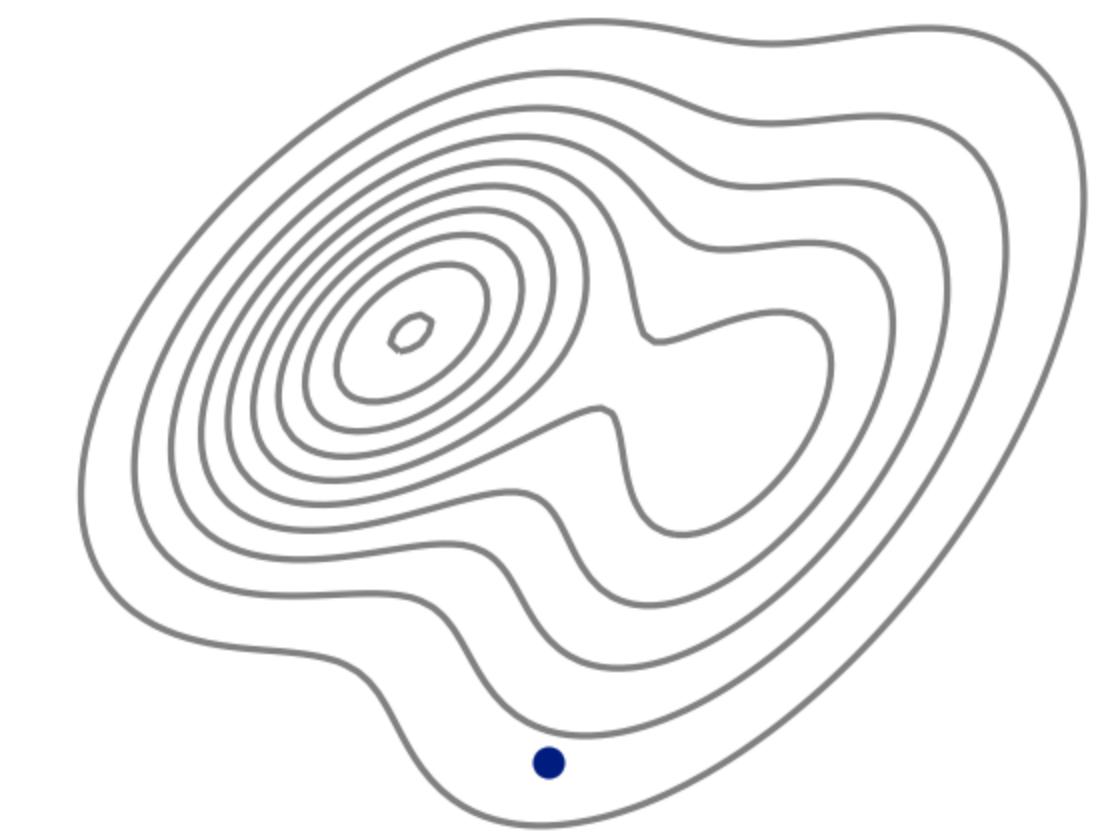
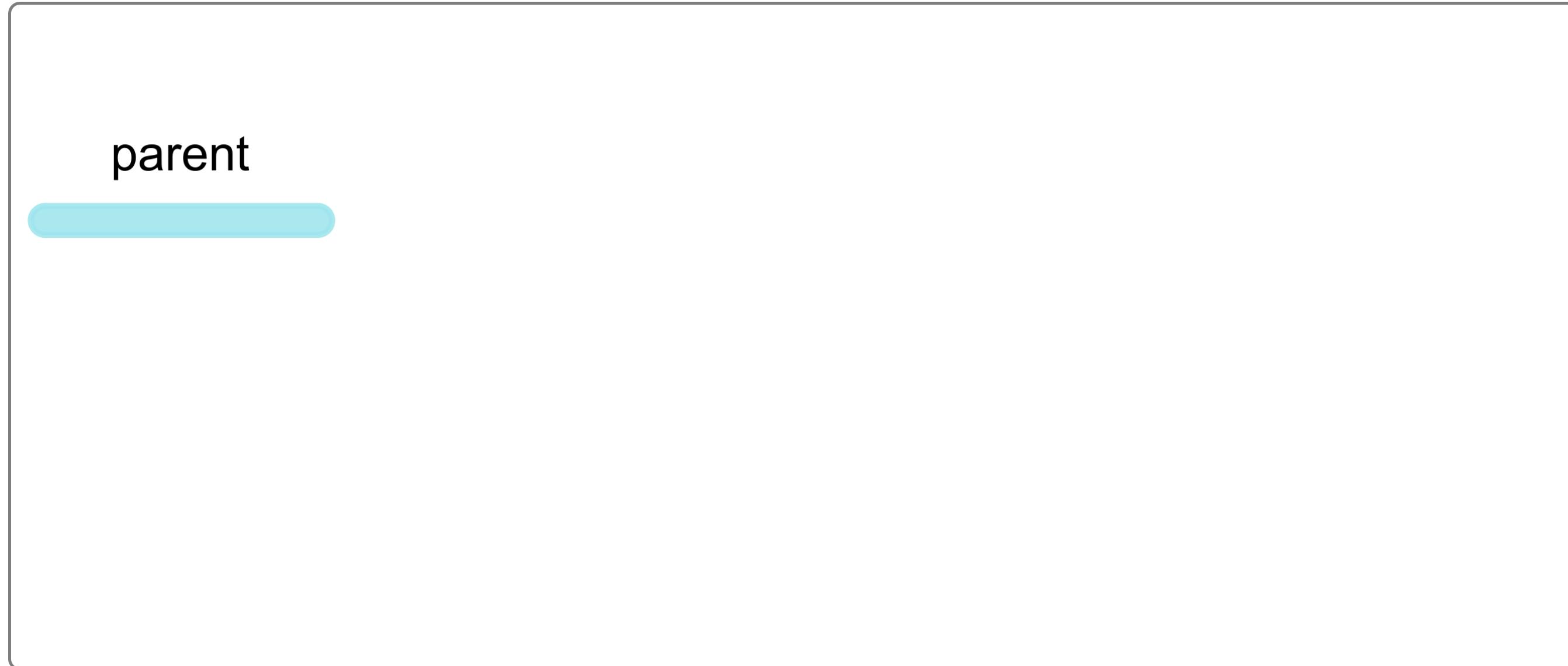


What sequence will give the desired function?

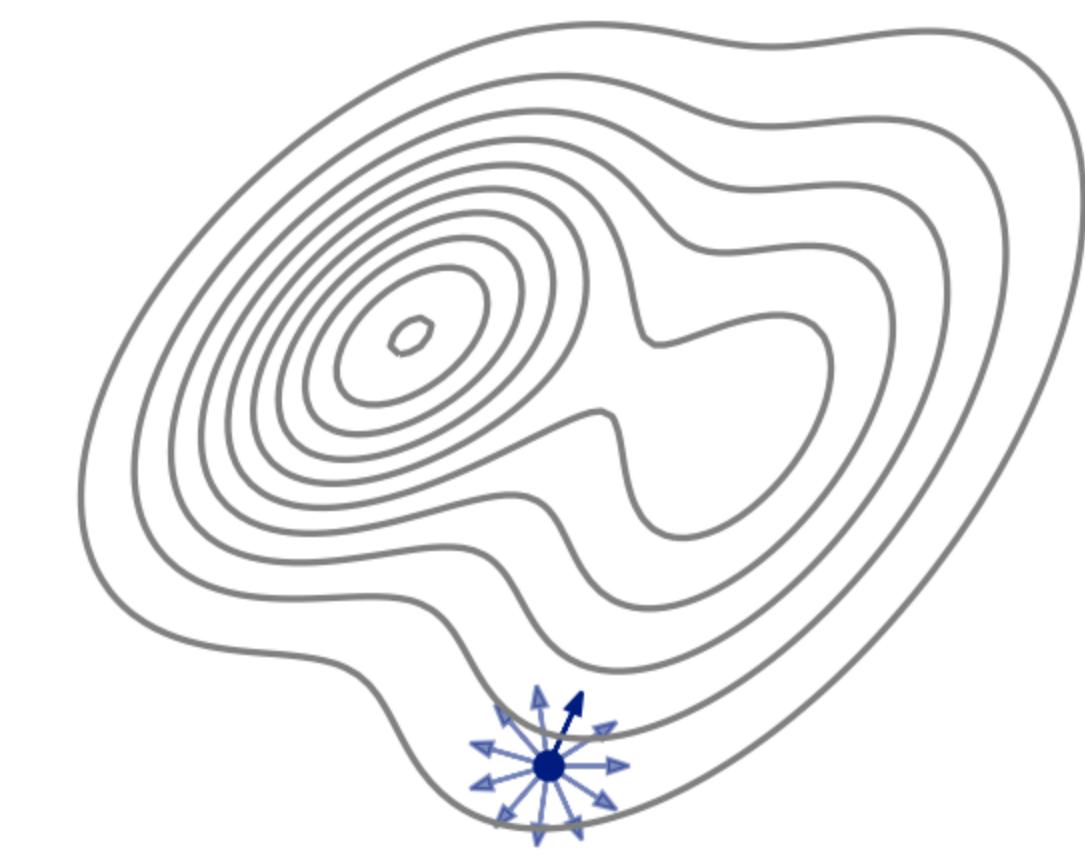
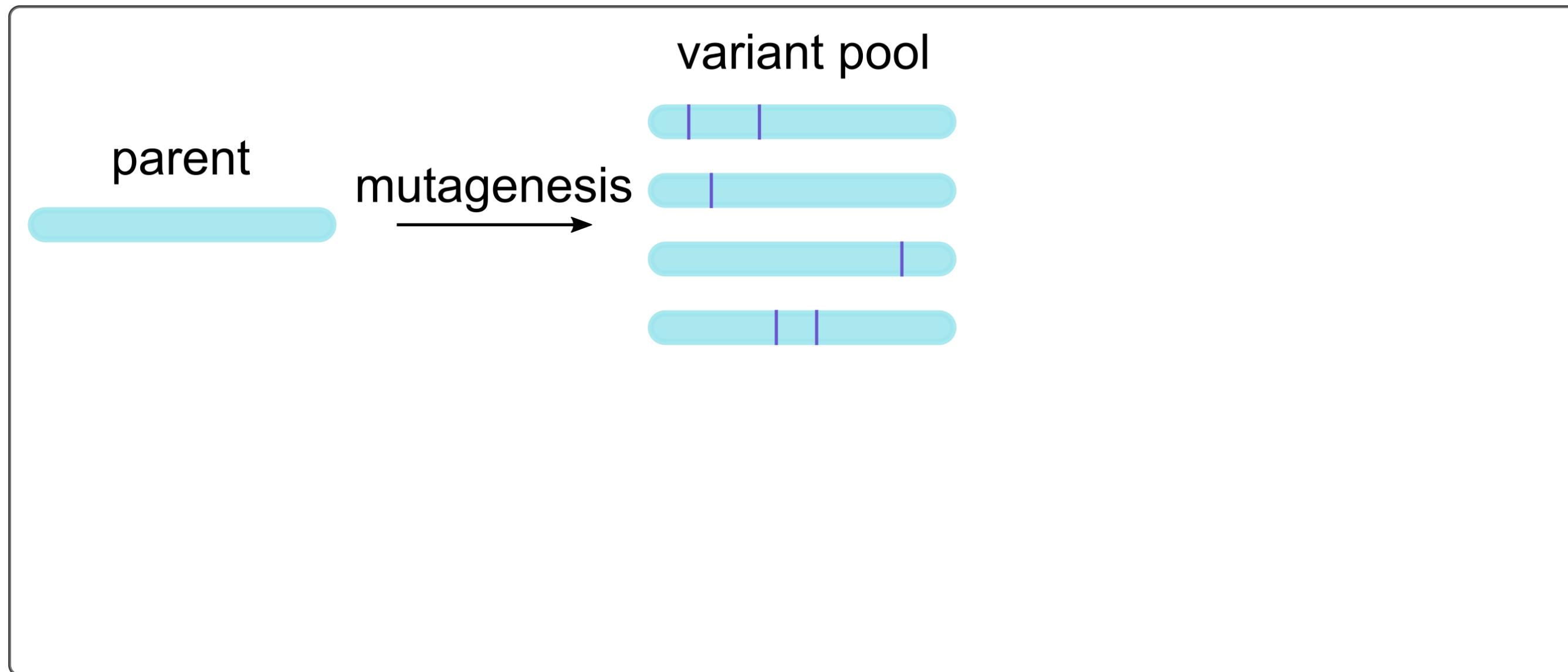
Directed evolution sidesteps the problem



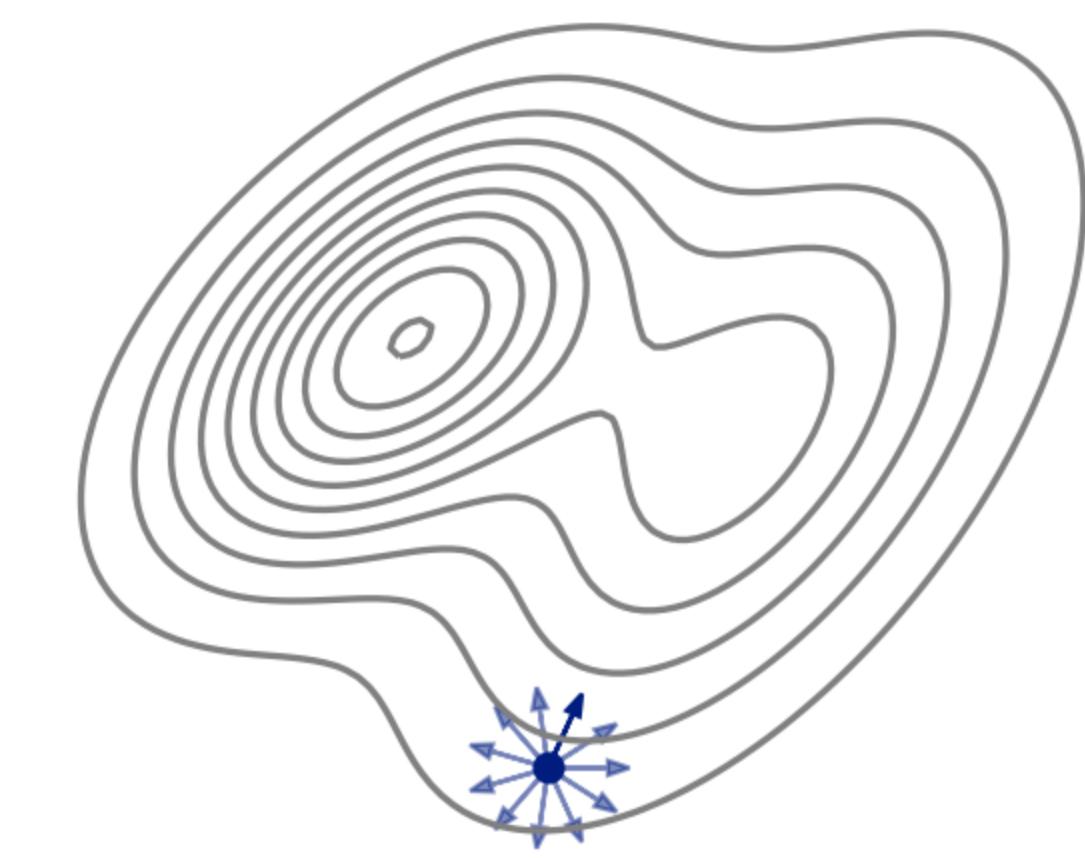
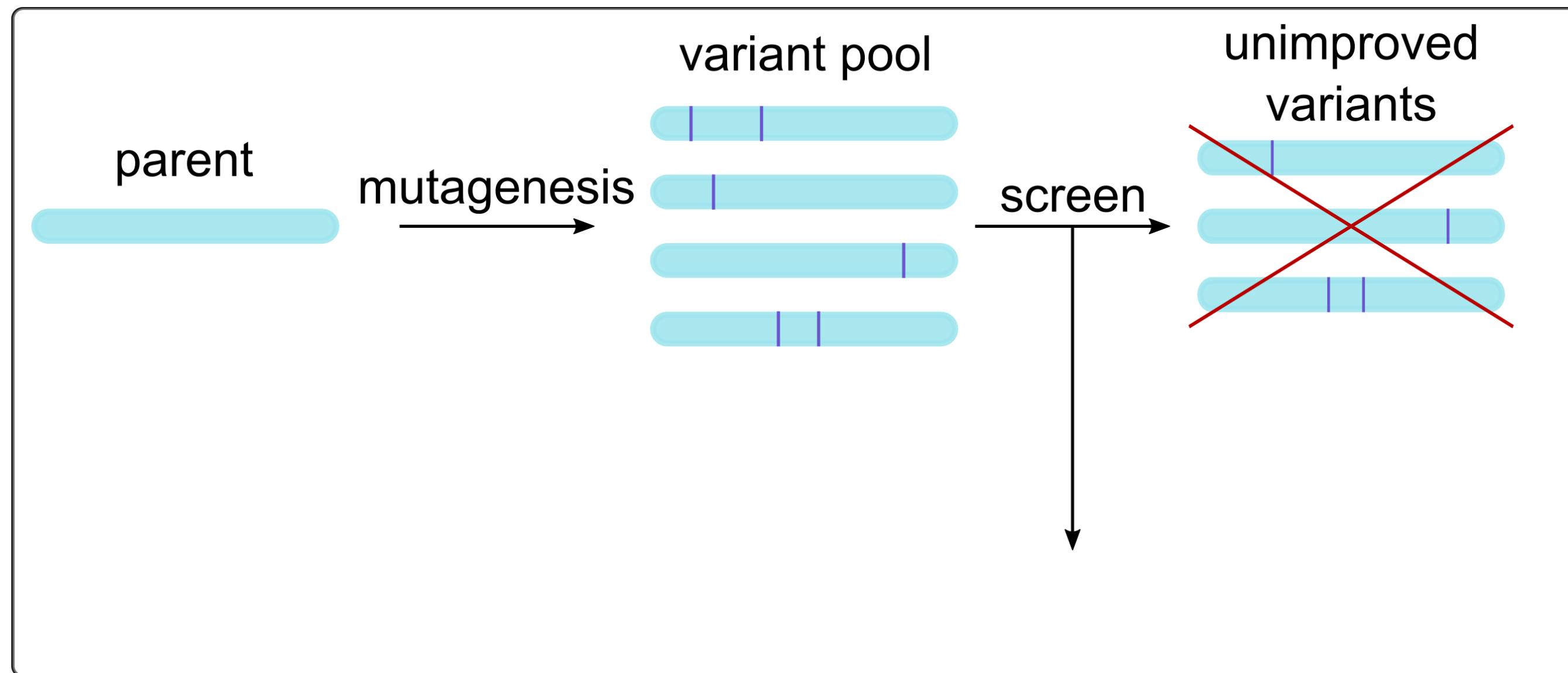
Directed evolution sidesteps the problem



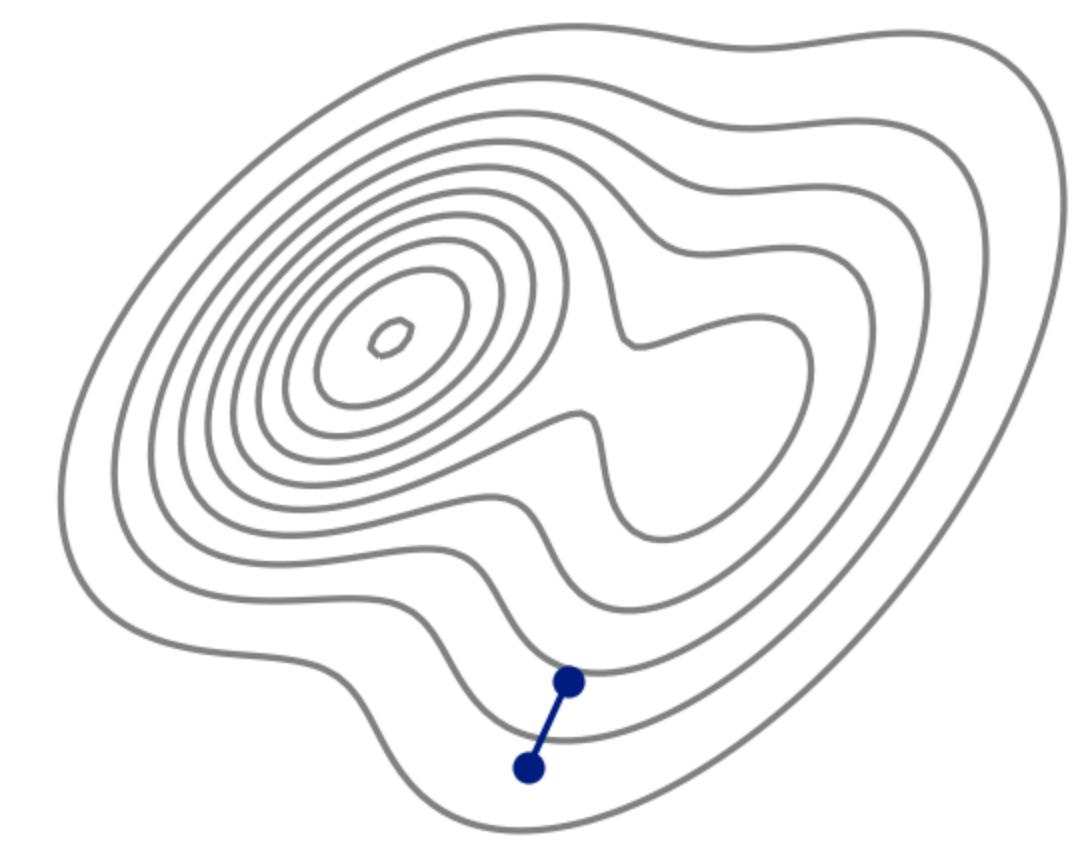
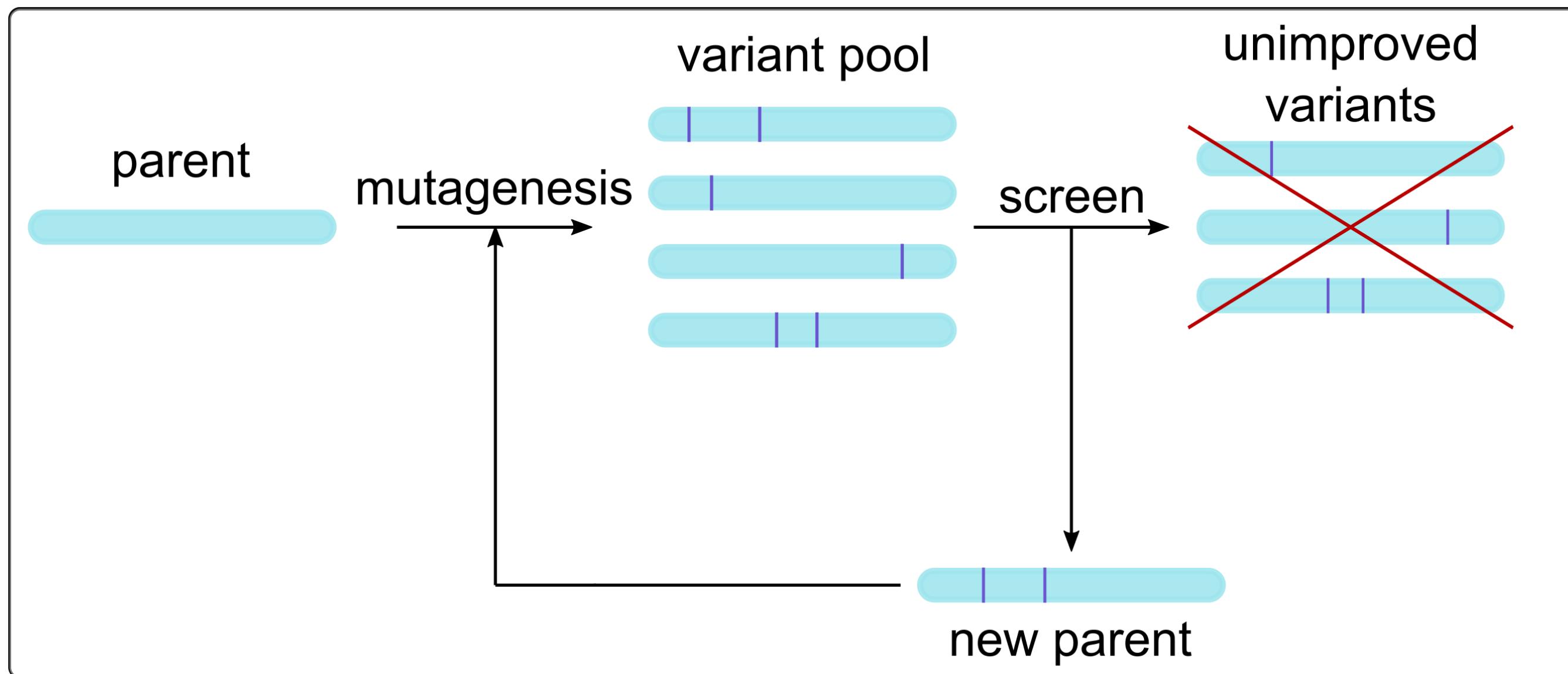
Directed evolution sidesteps the problem



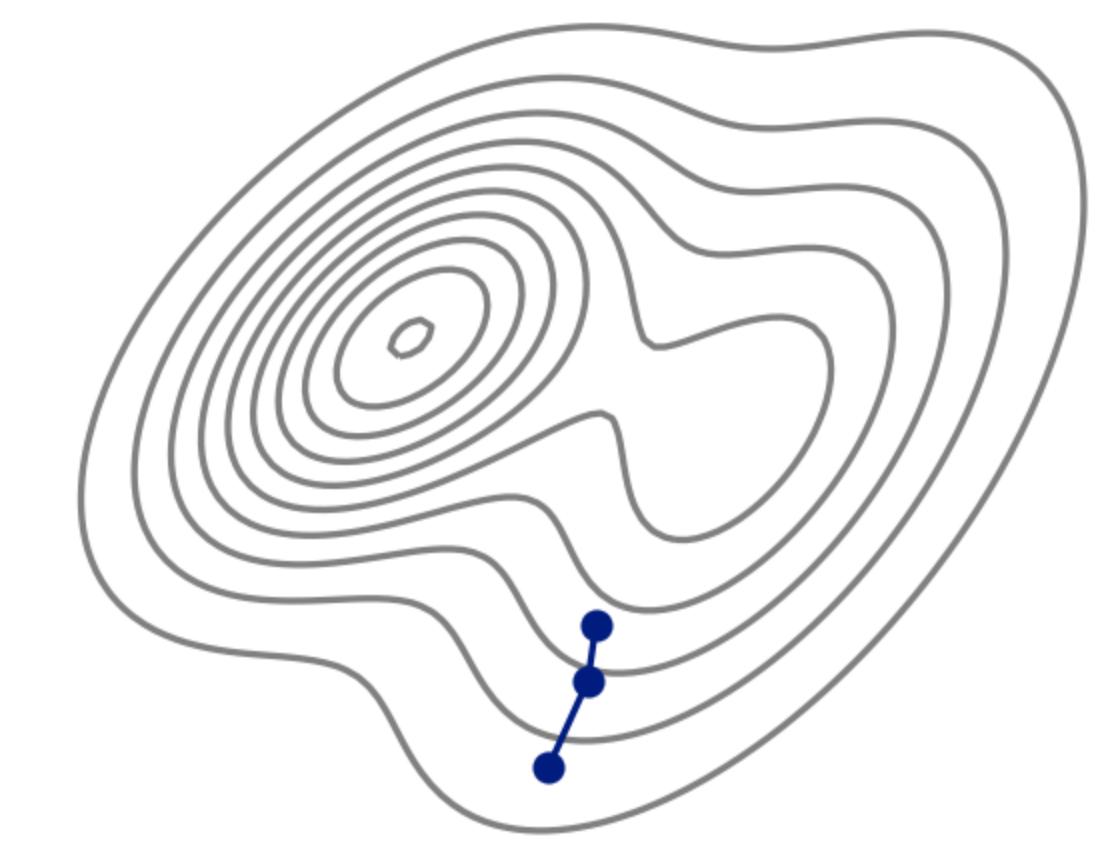
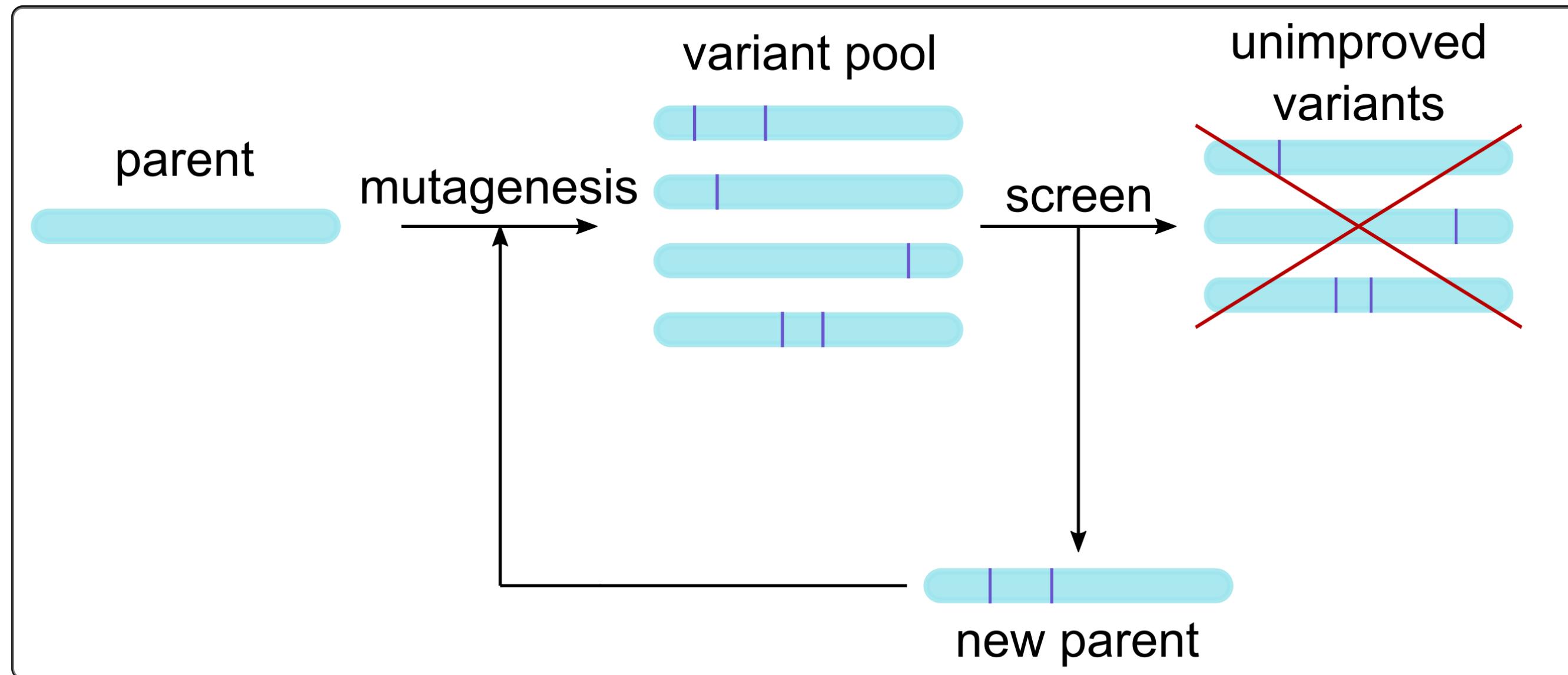
Directed evolution sidesteps the problem



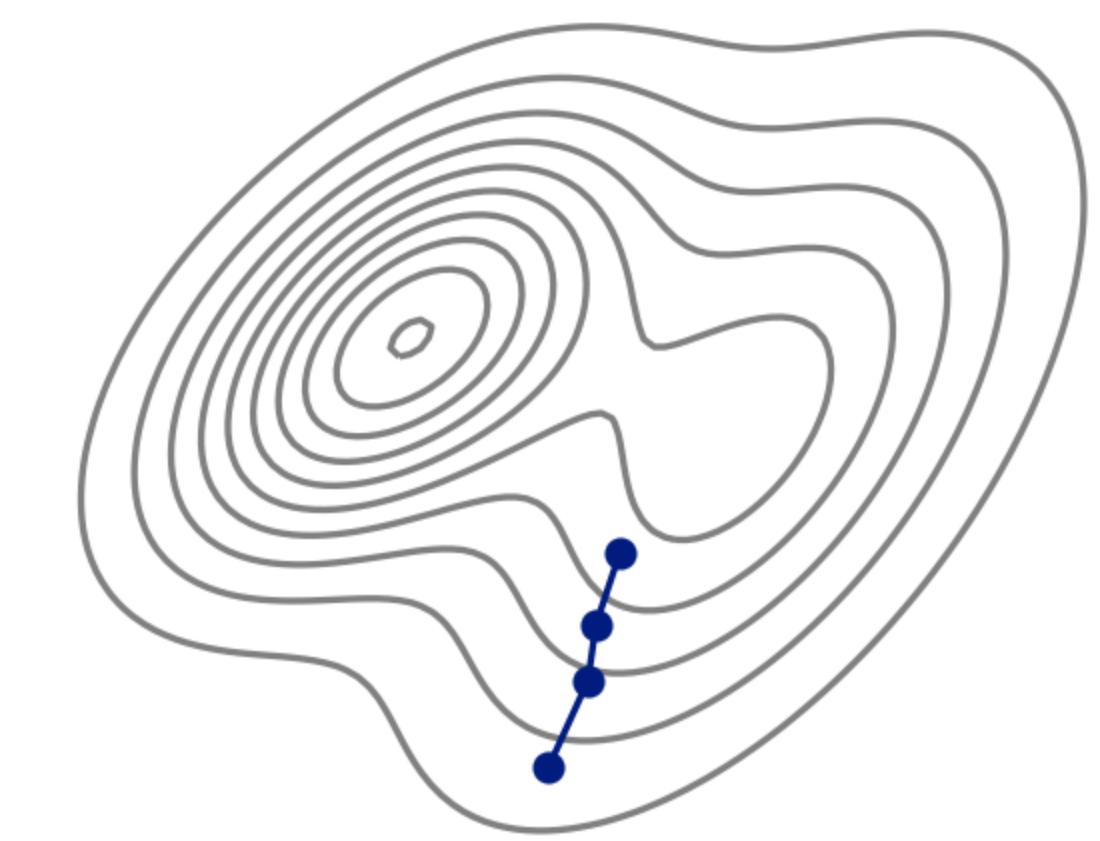
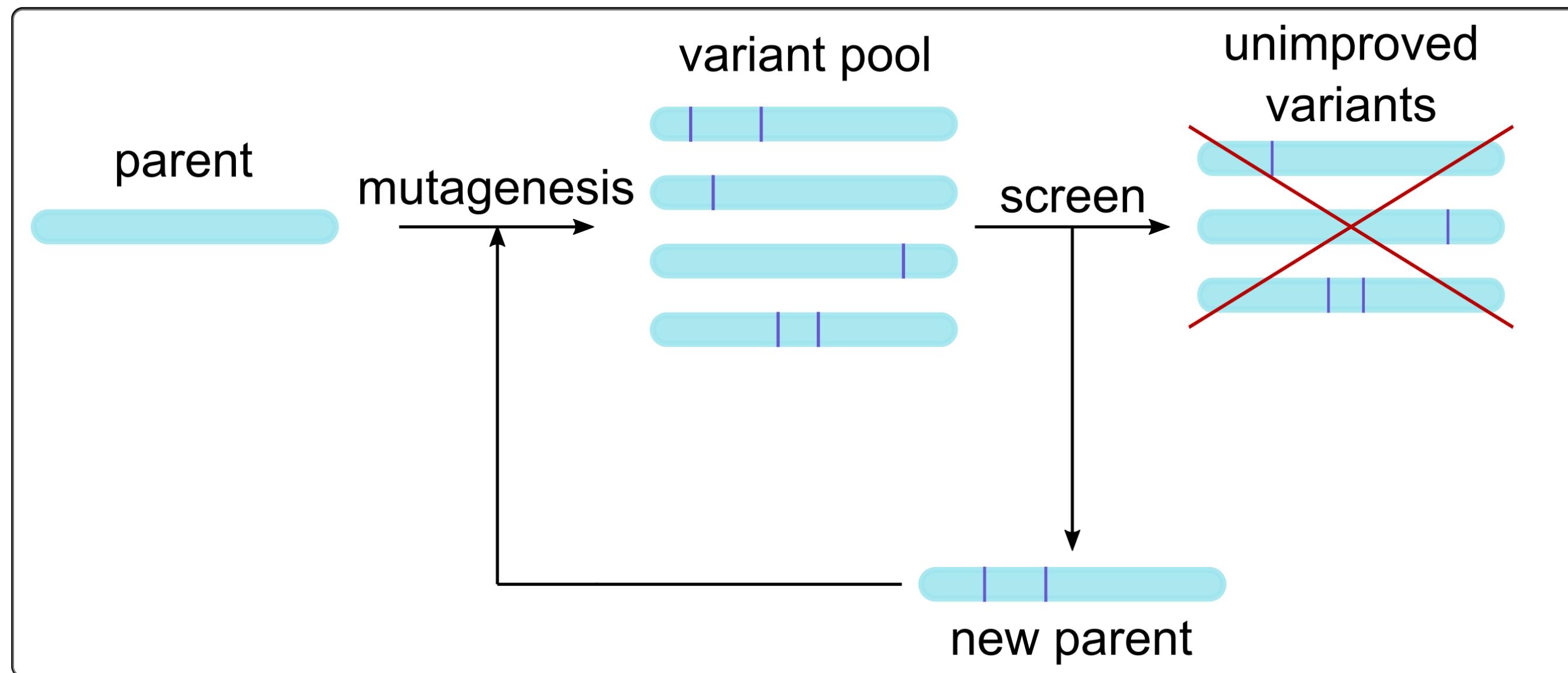
Directed evolution sidesteps the problem



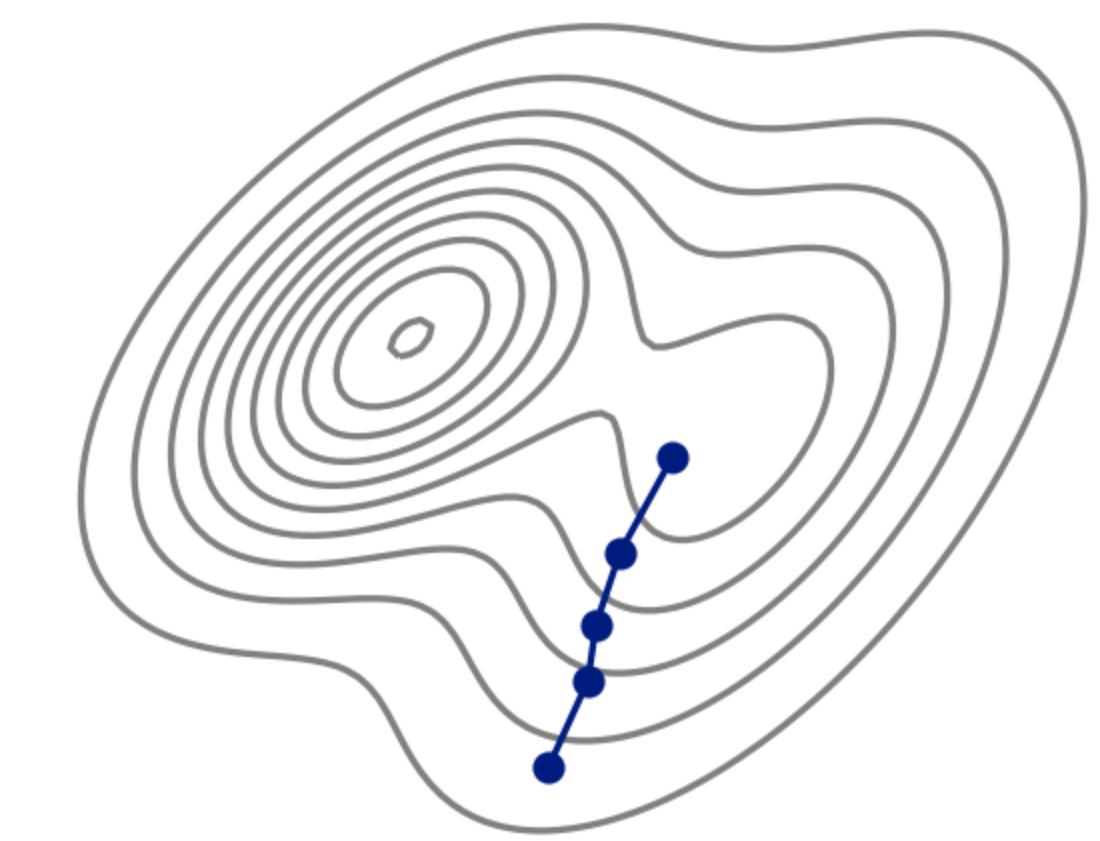
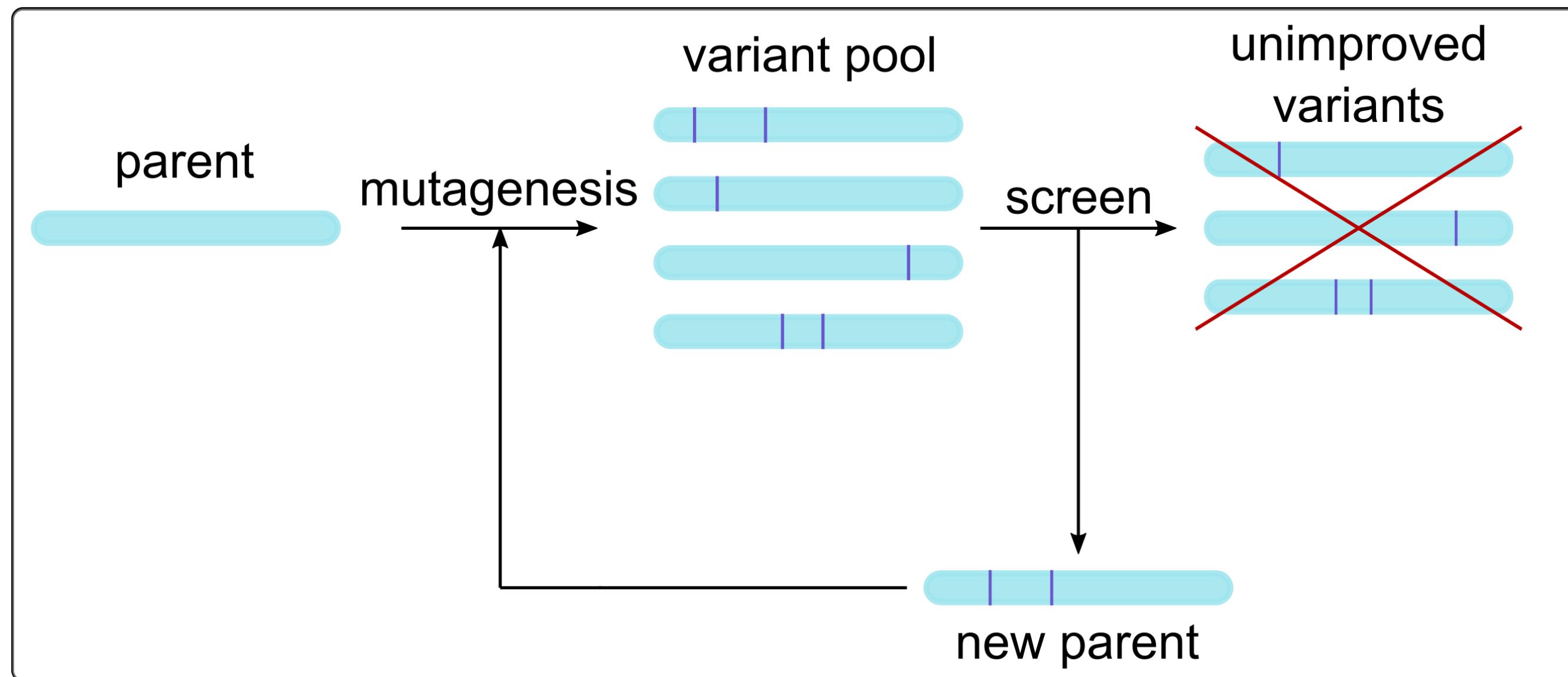
Directed evolution sidesteps the problem



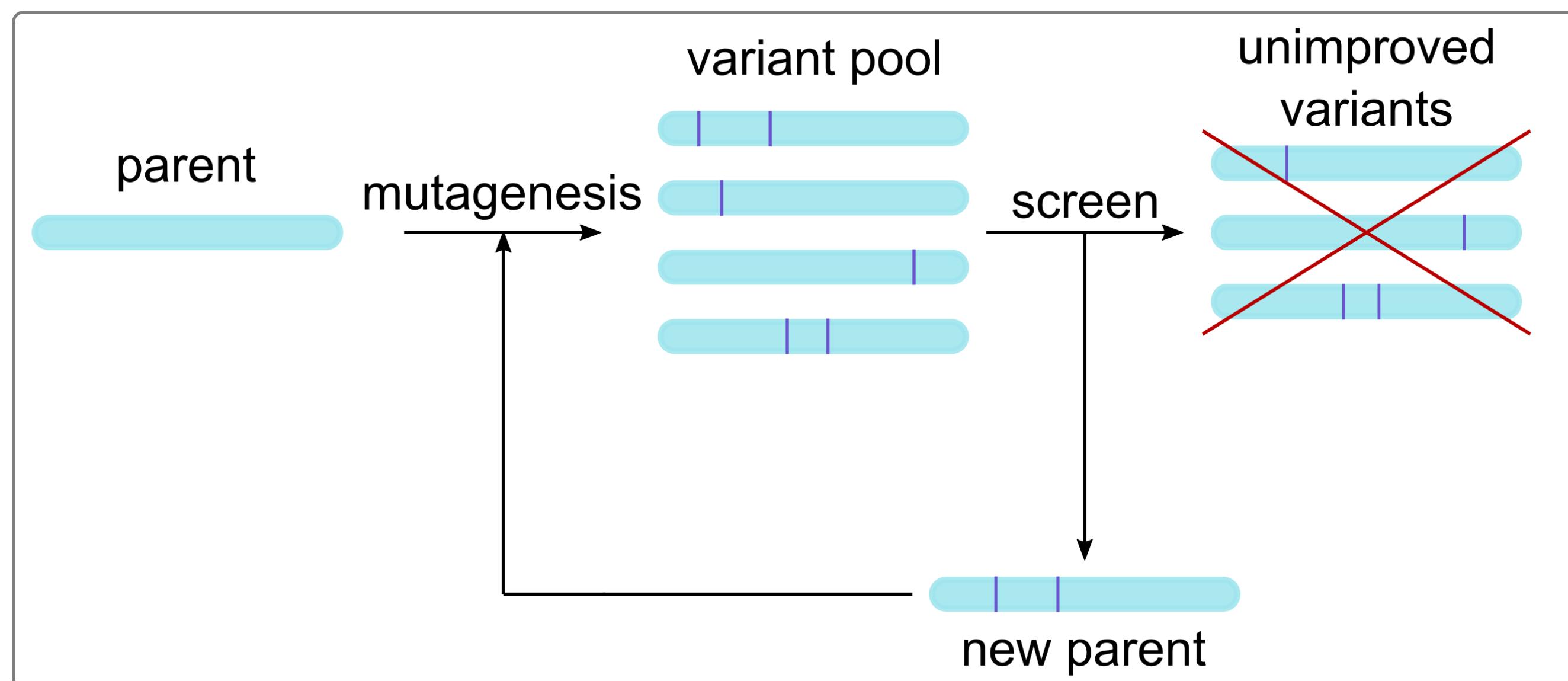
Directed evolution sidesteps the problem



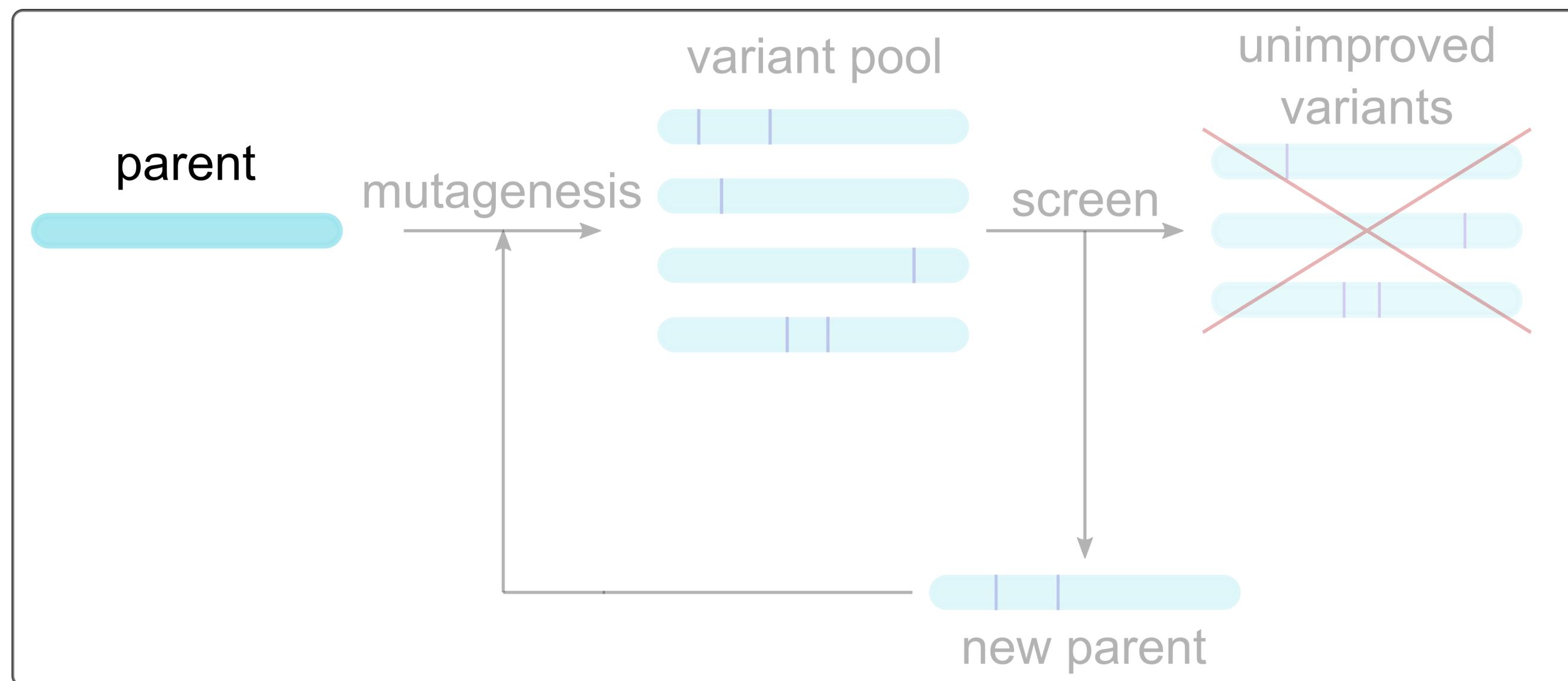
Directed evolution sidesteps the problem



Requirements for directed evolution

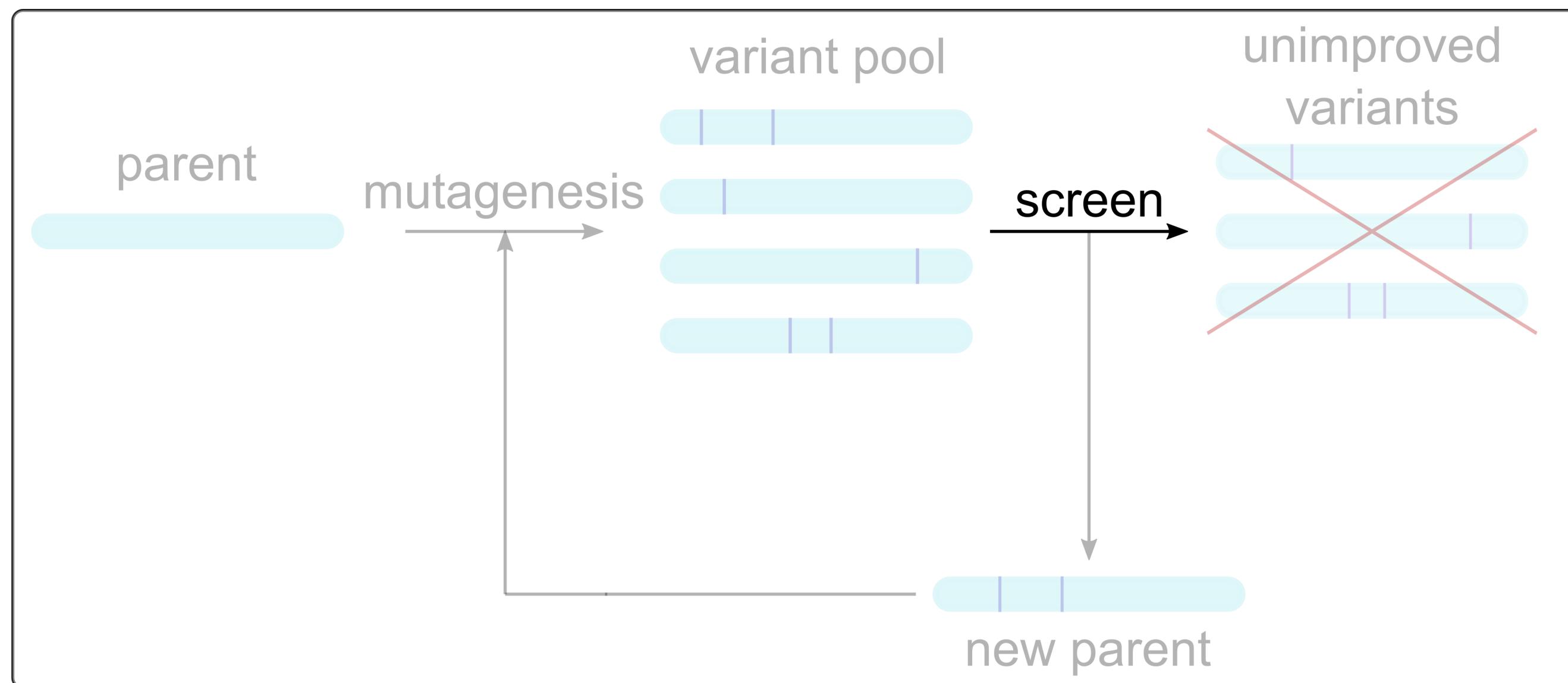


Requirements for directed evolution



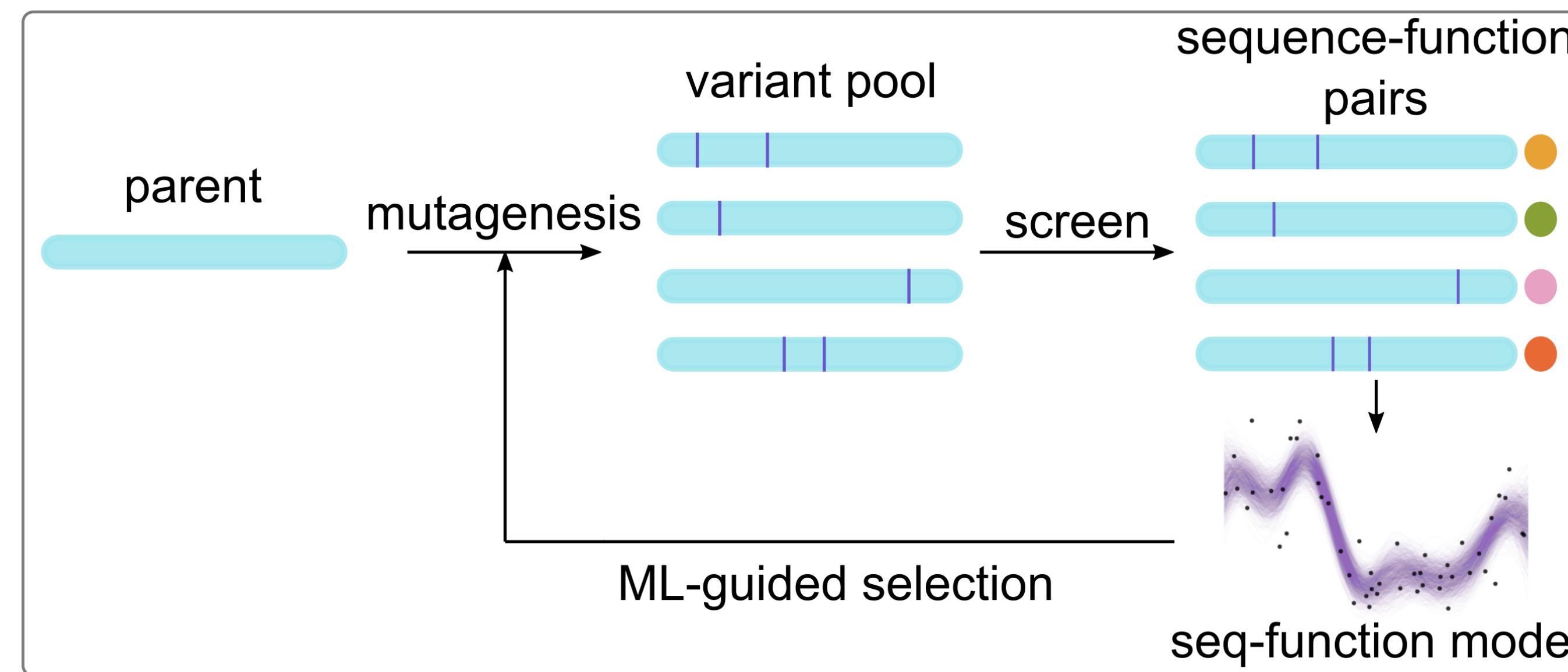
- Parent

Requirements for directed evolution

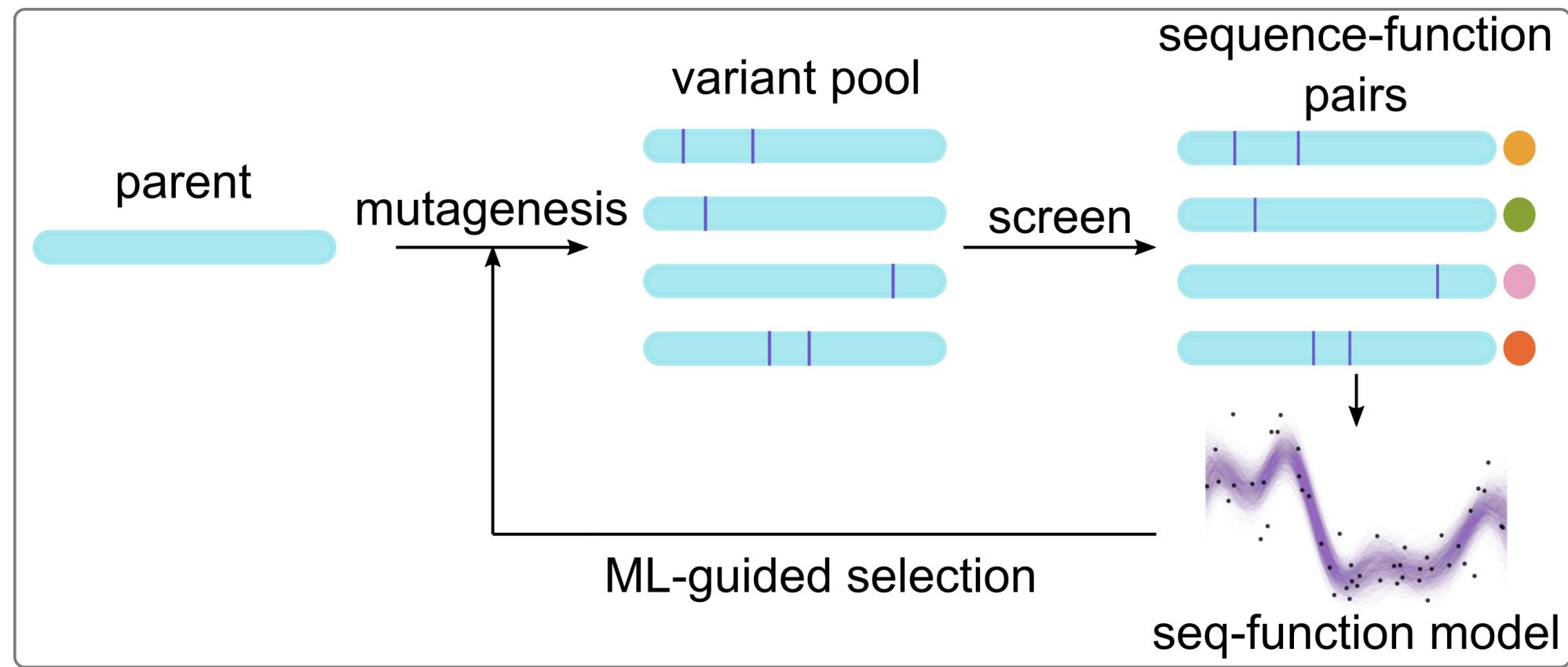


- Parent
- High-throughput screen (>100 / wk)

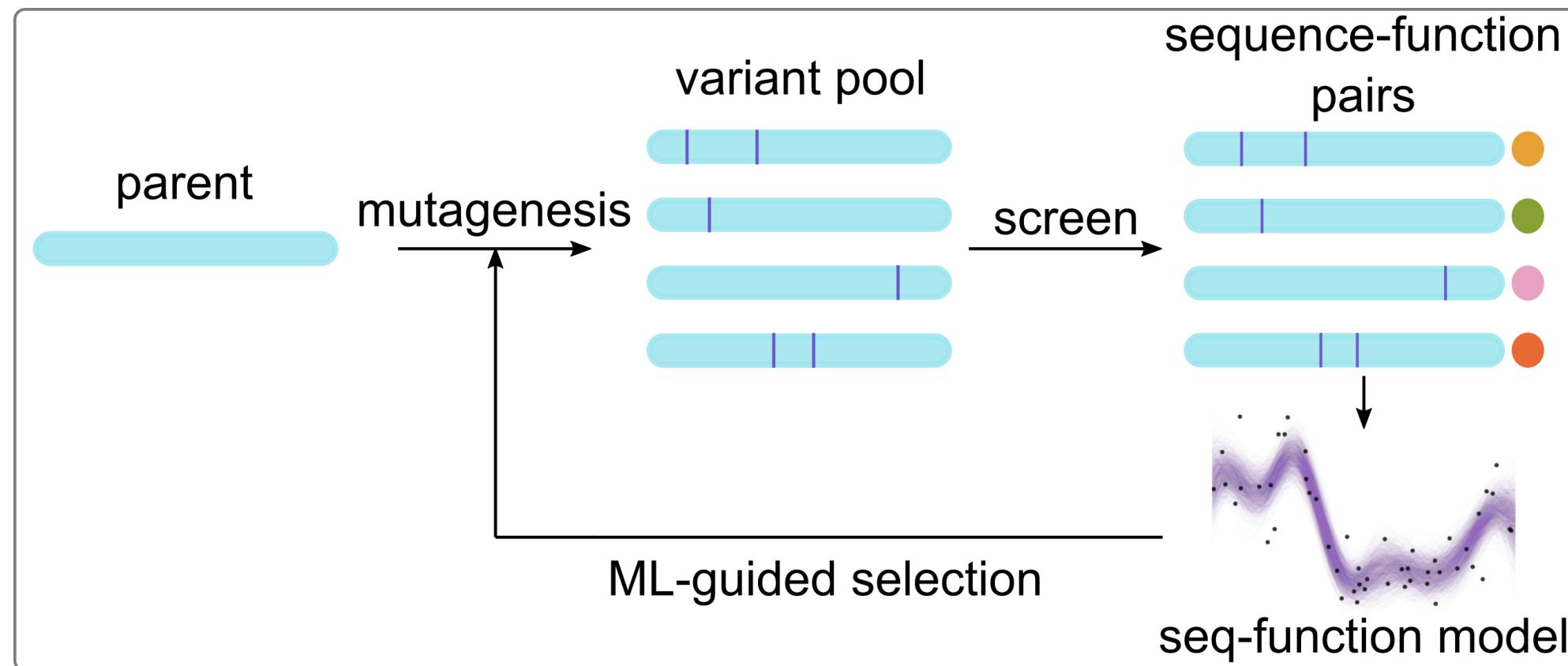
Machine learning enables optimization with fewer measurements



But still requires starting points

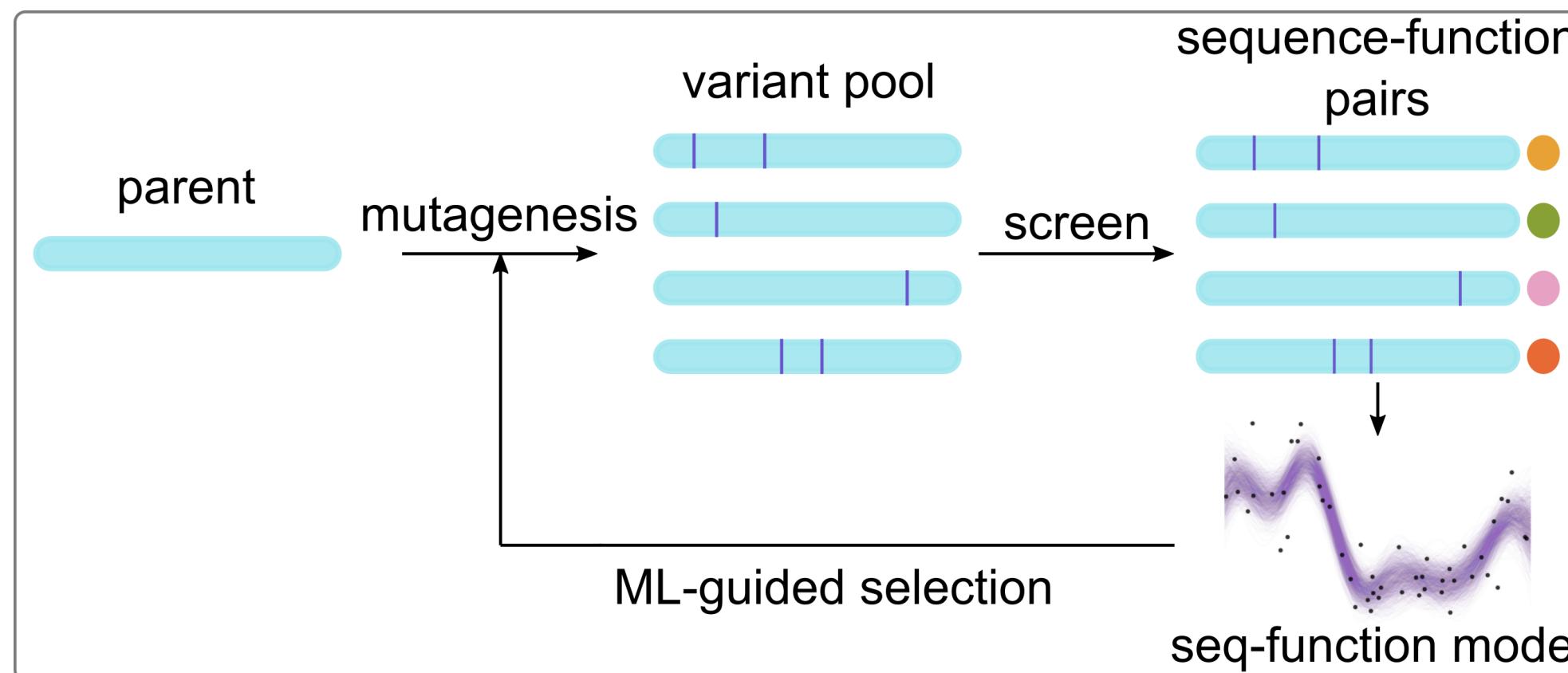


But still requires starting points



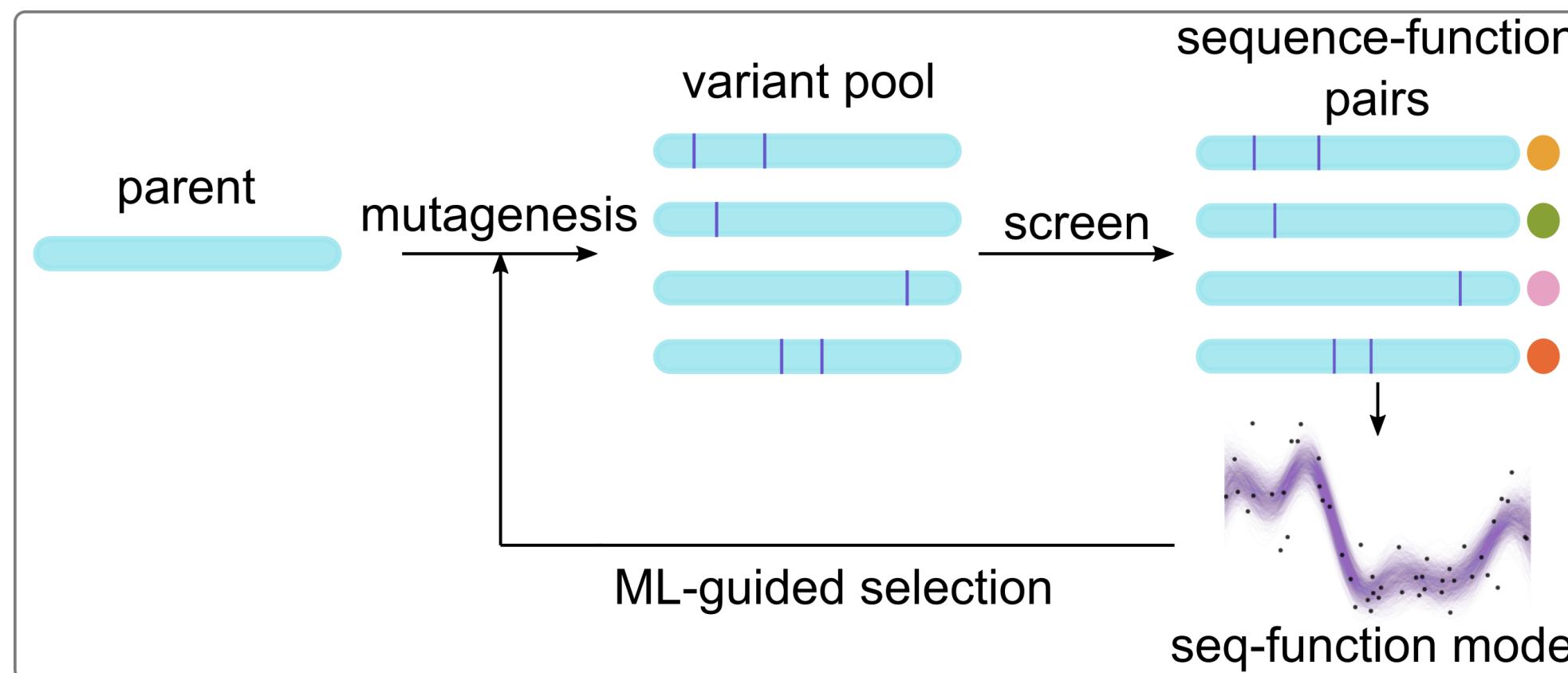
- Accurate models with < 200 measurements

But still requires starting points



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

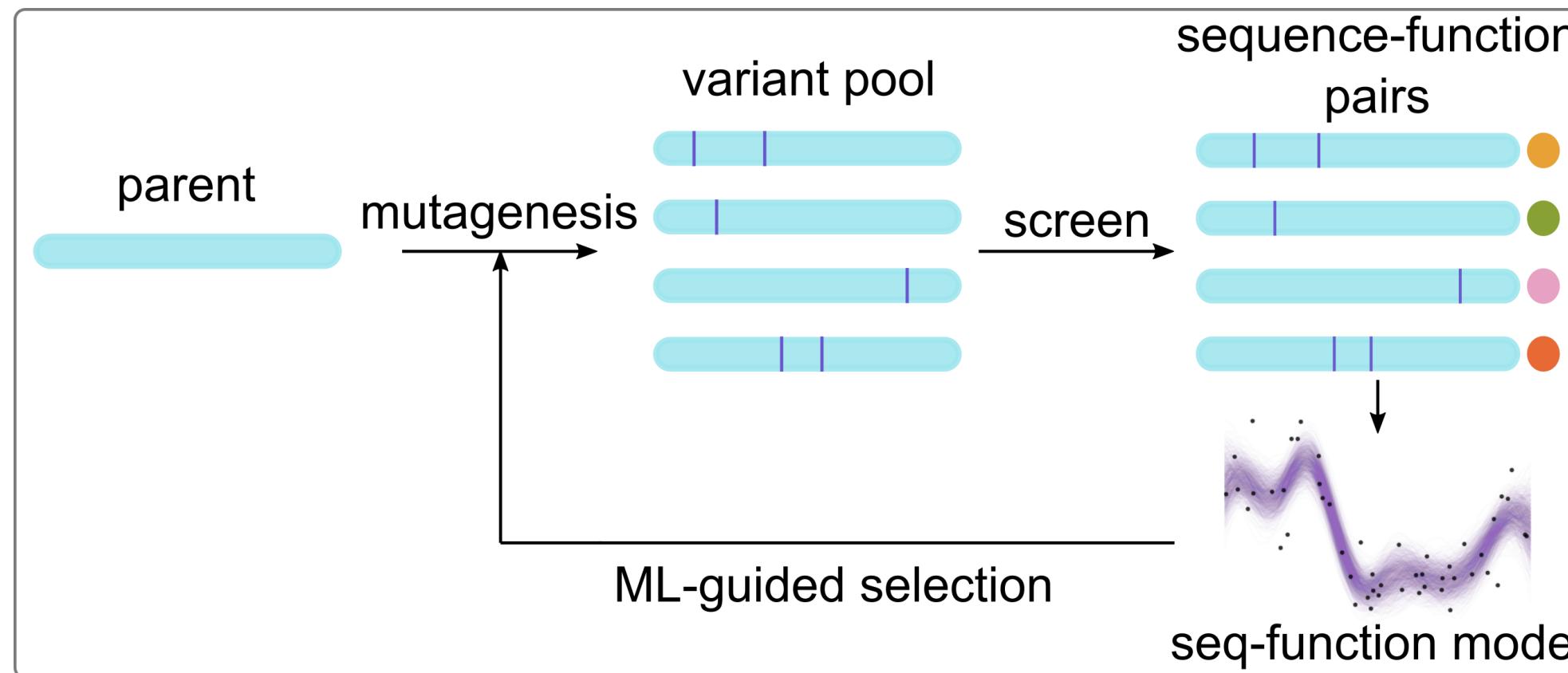
But still requires starting points



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

Still requires starting point!

But still requires starting points



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

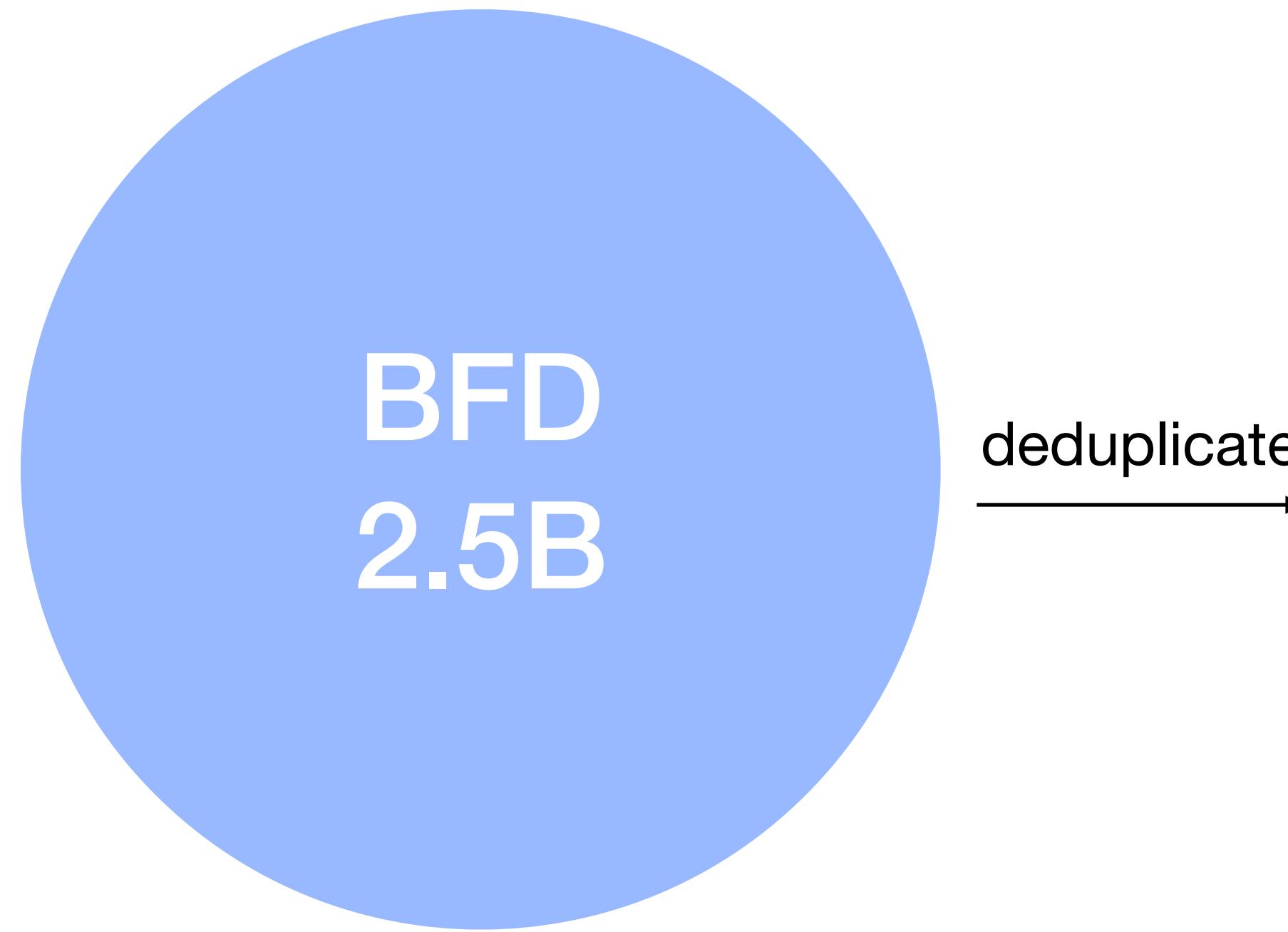
Still requires starting point!
Ignores mountains of protein data

We have access to large protein databases

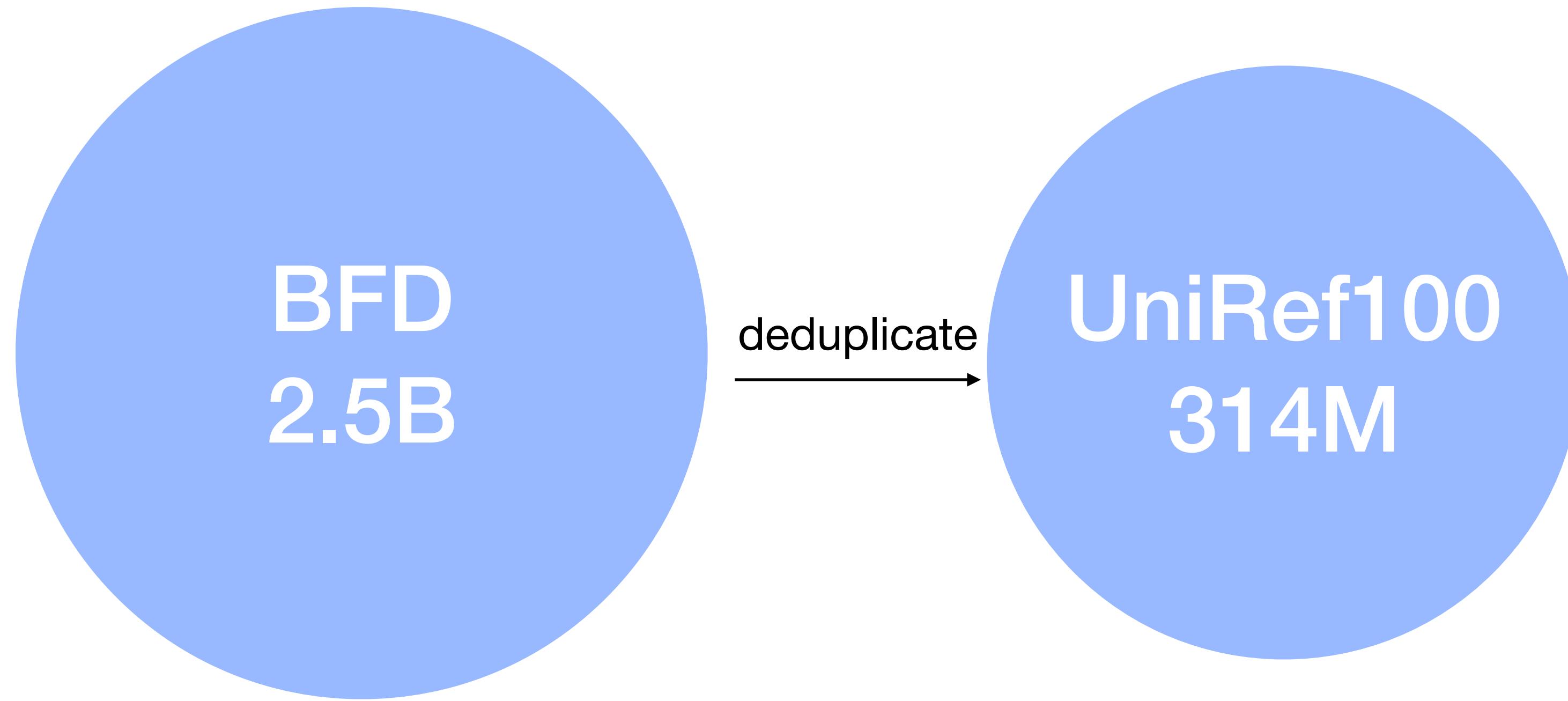
We have access to large protein databases



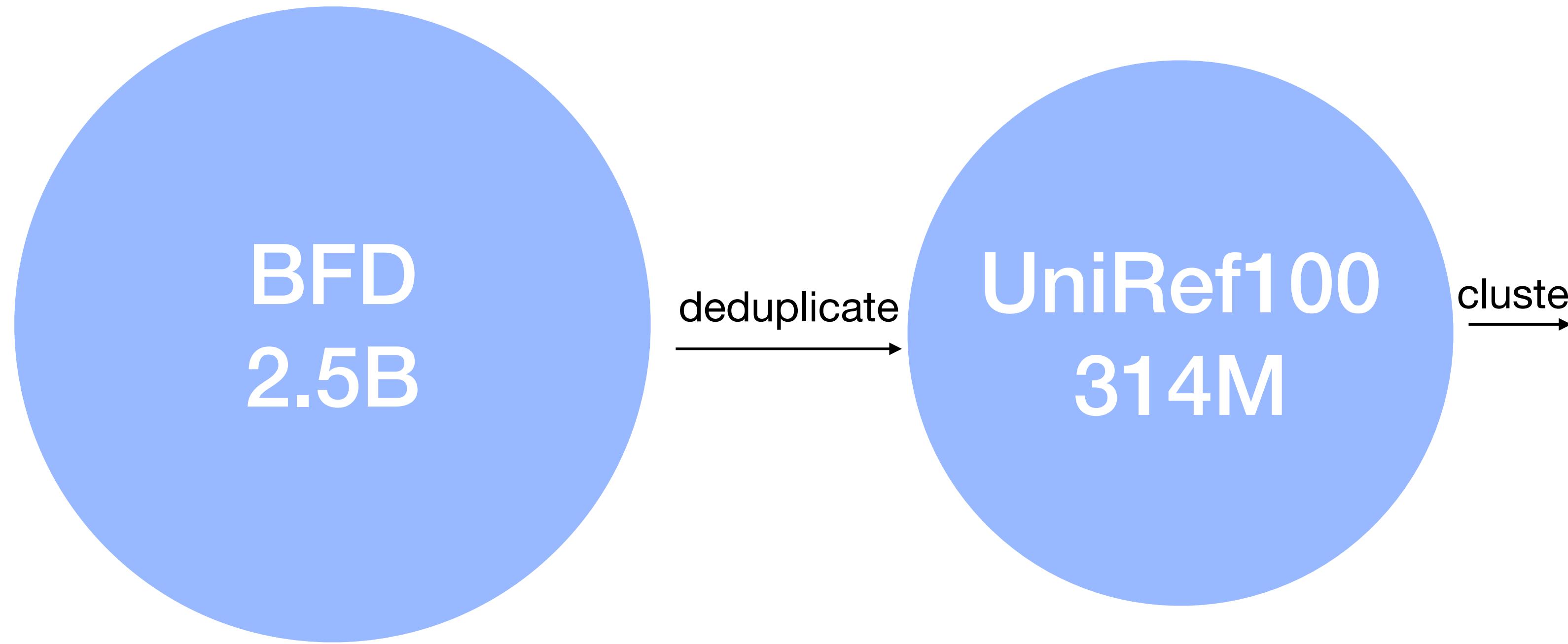
We have access to large protein databases



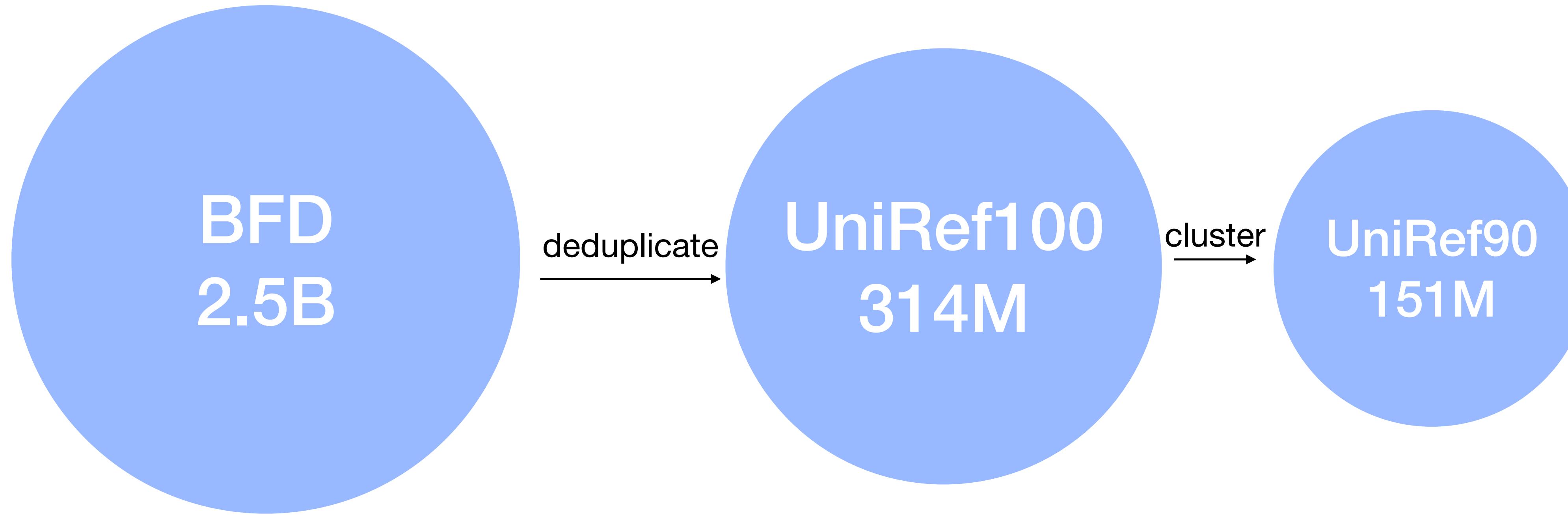
We have access to large protein databases



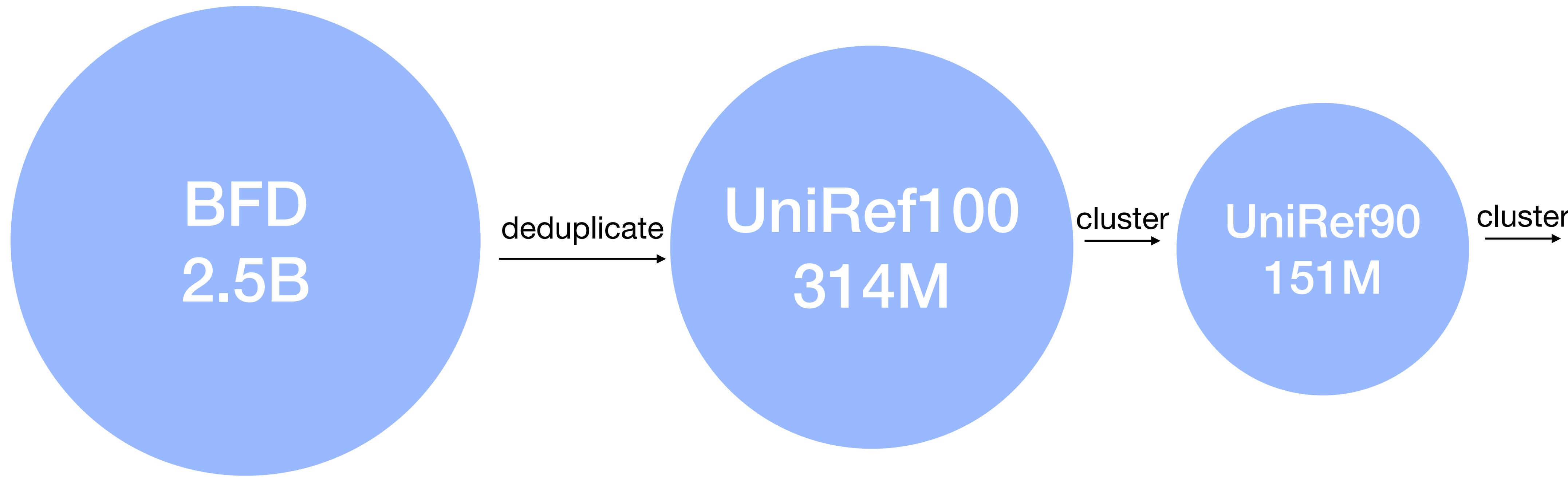
We have access to large protein databases



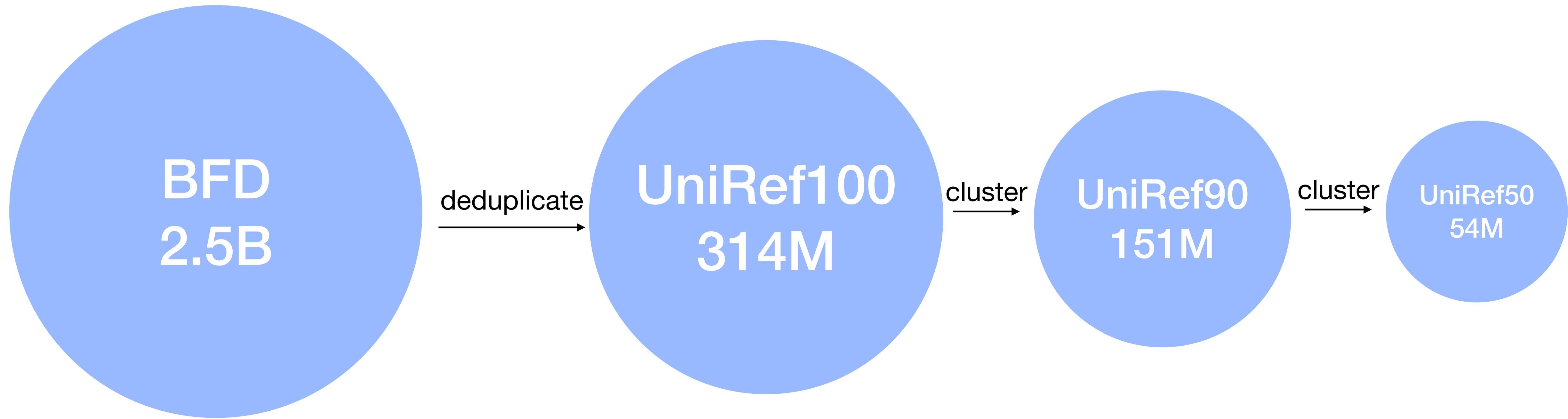
We have access to large protein databases



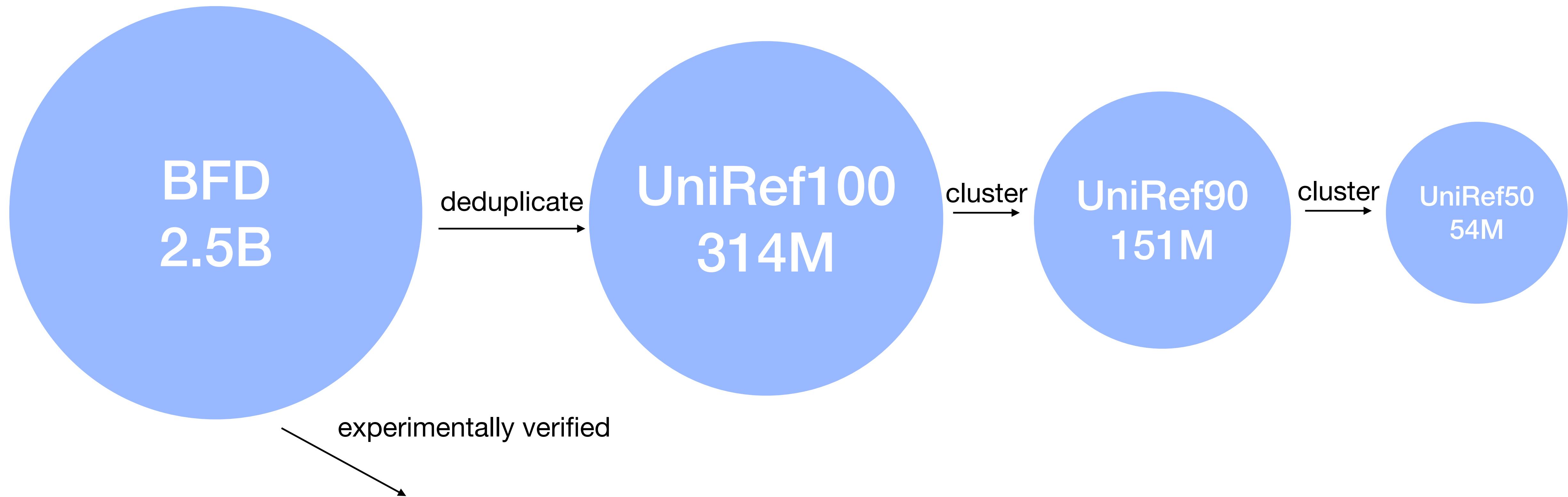
We have access to large protein databases



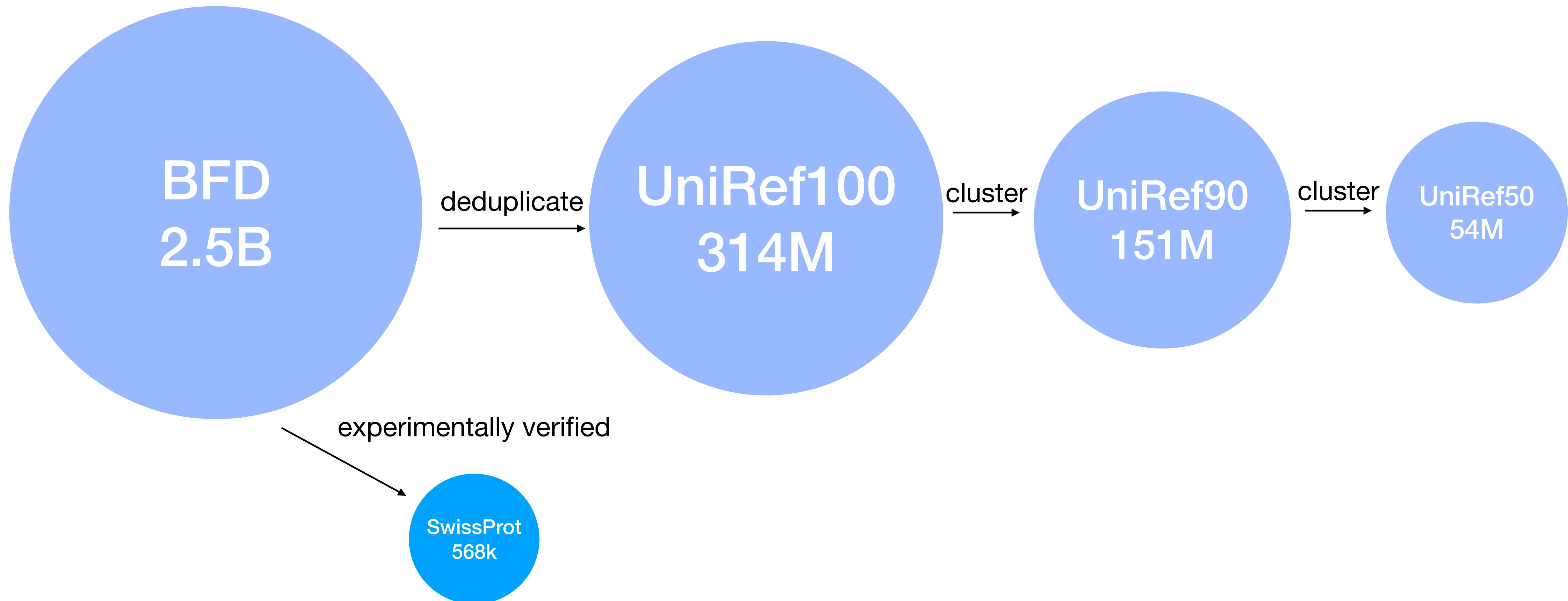
We have access to large protein databases



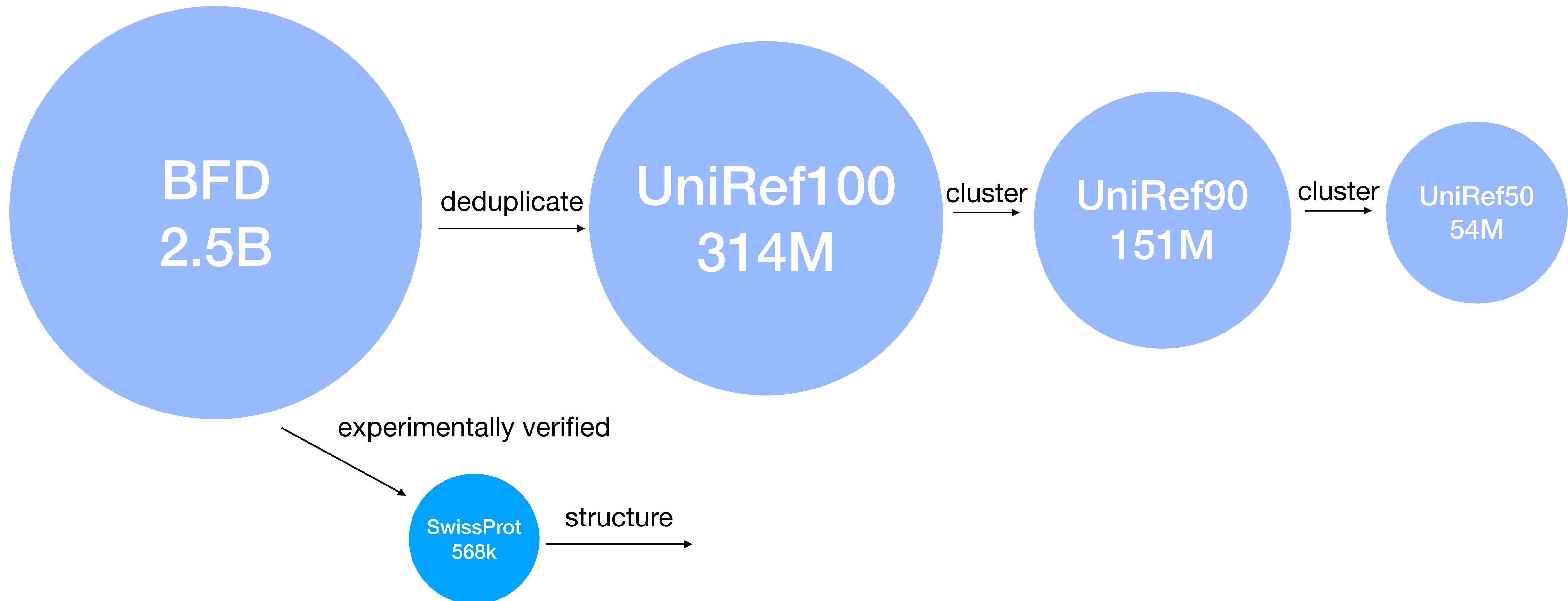
We have access to large protein databases



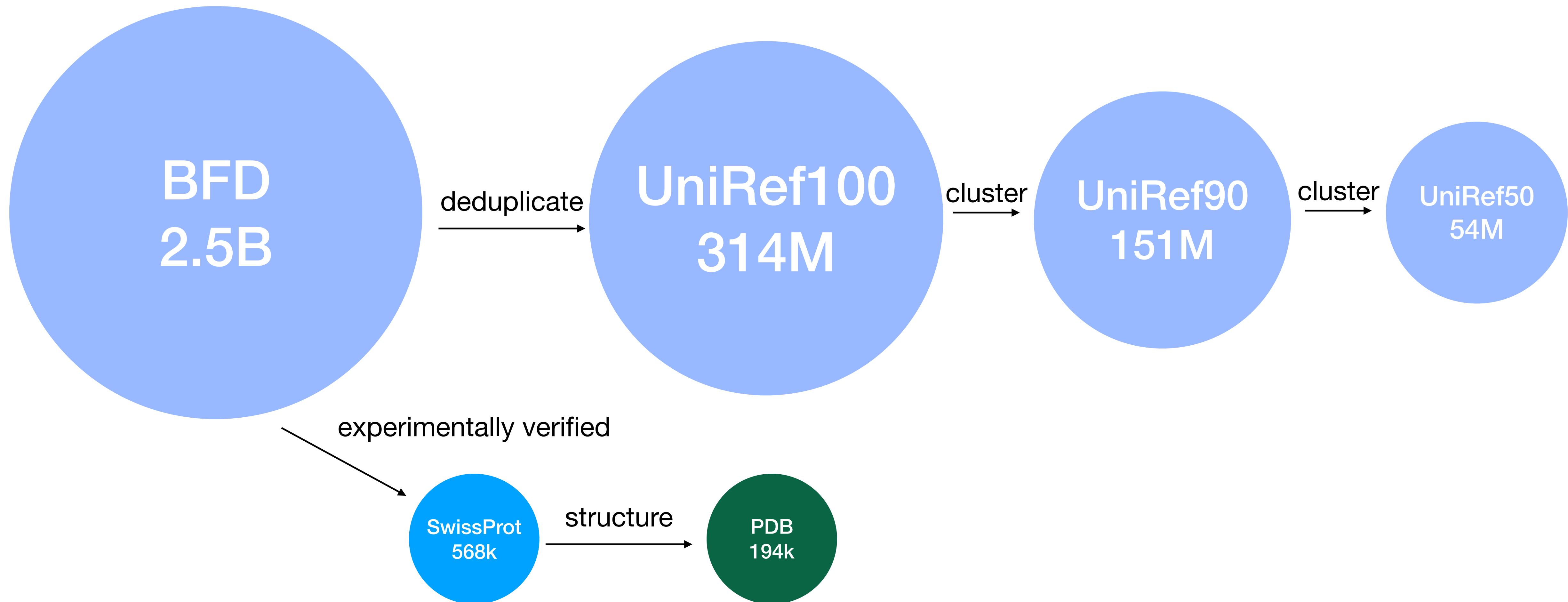
We have access to large protein databases



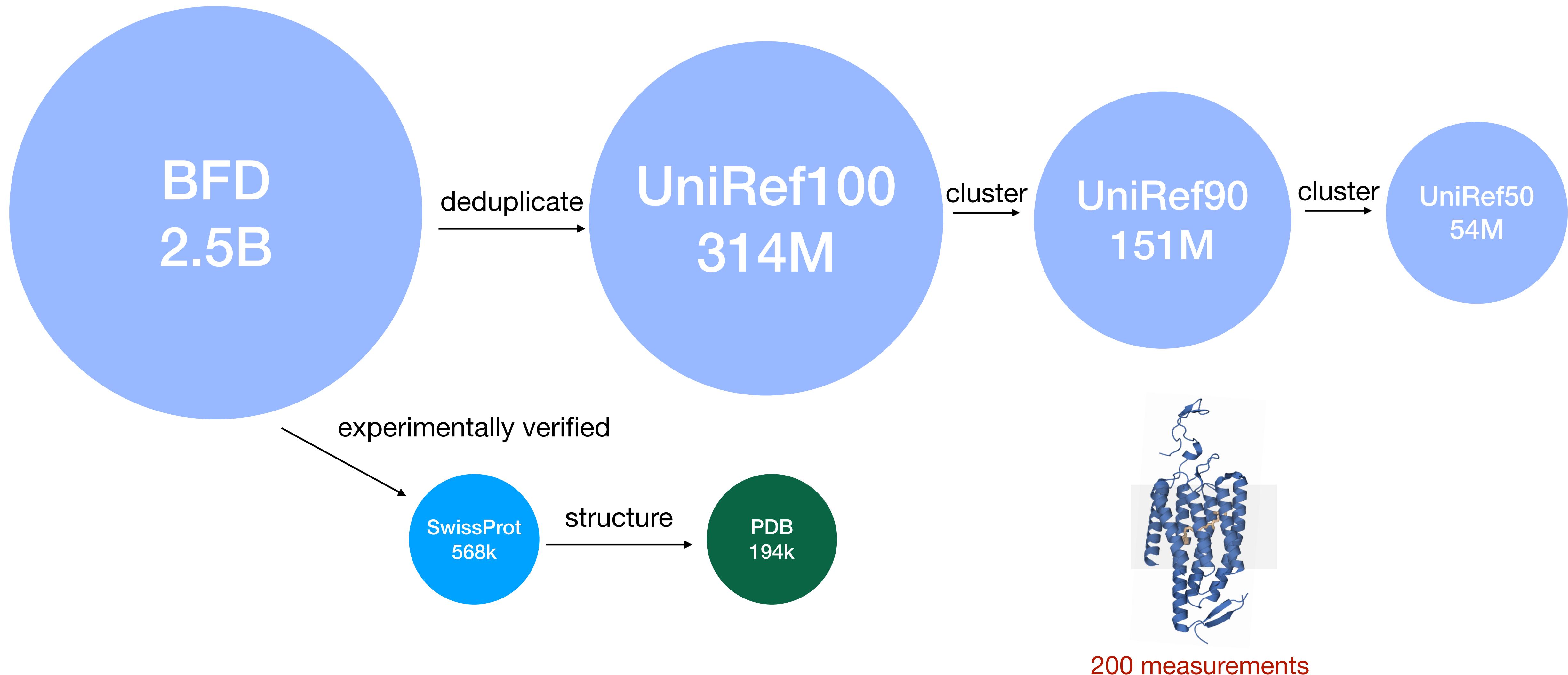
We have access to large protein databases



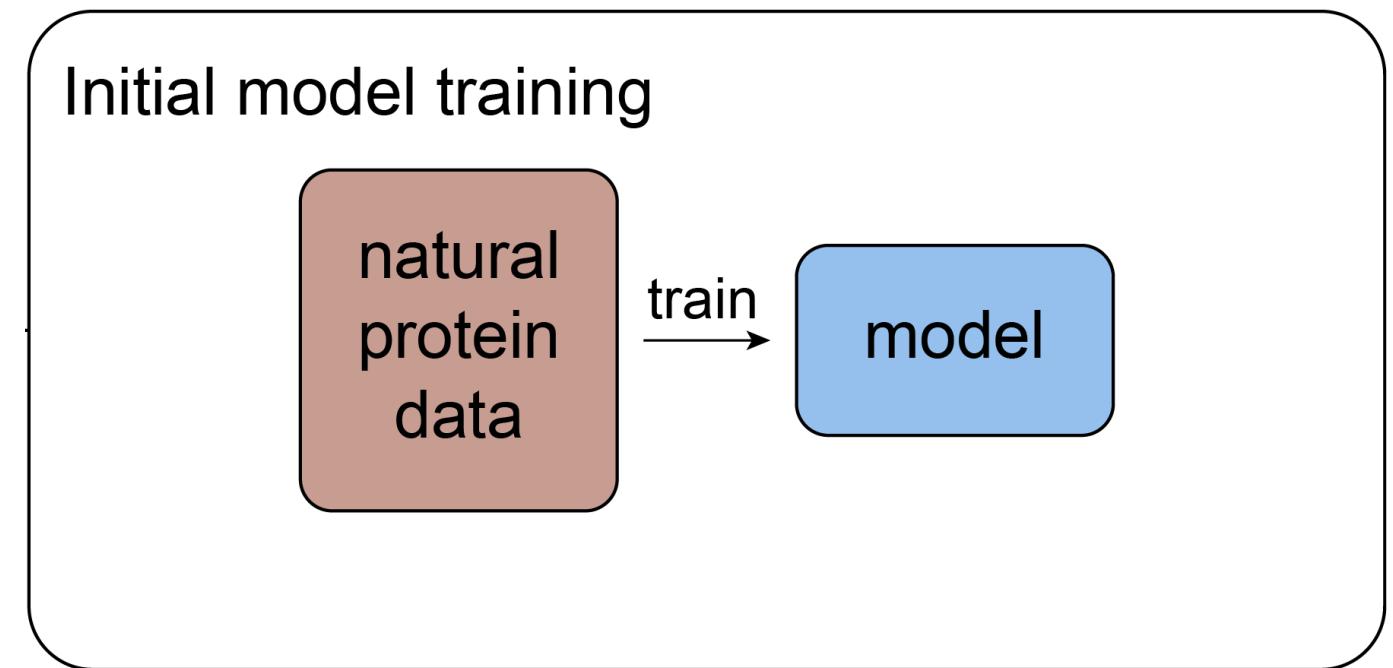
We have access to large protein databases



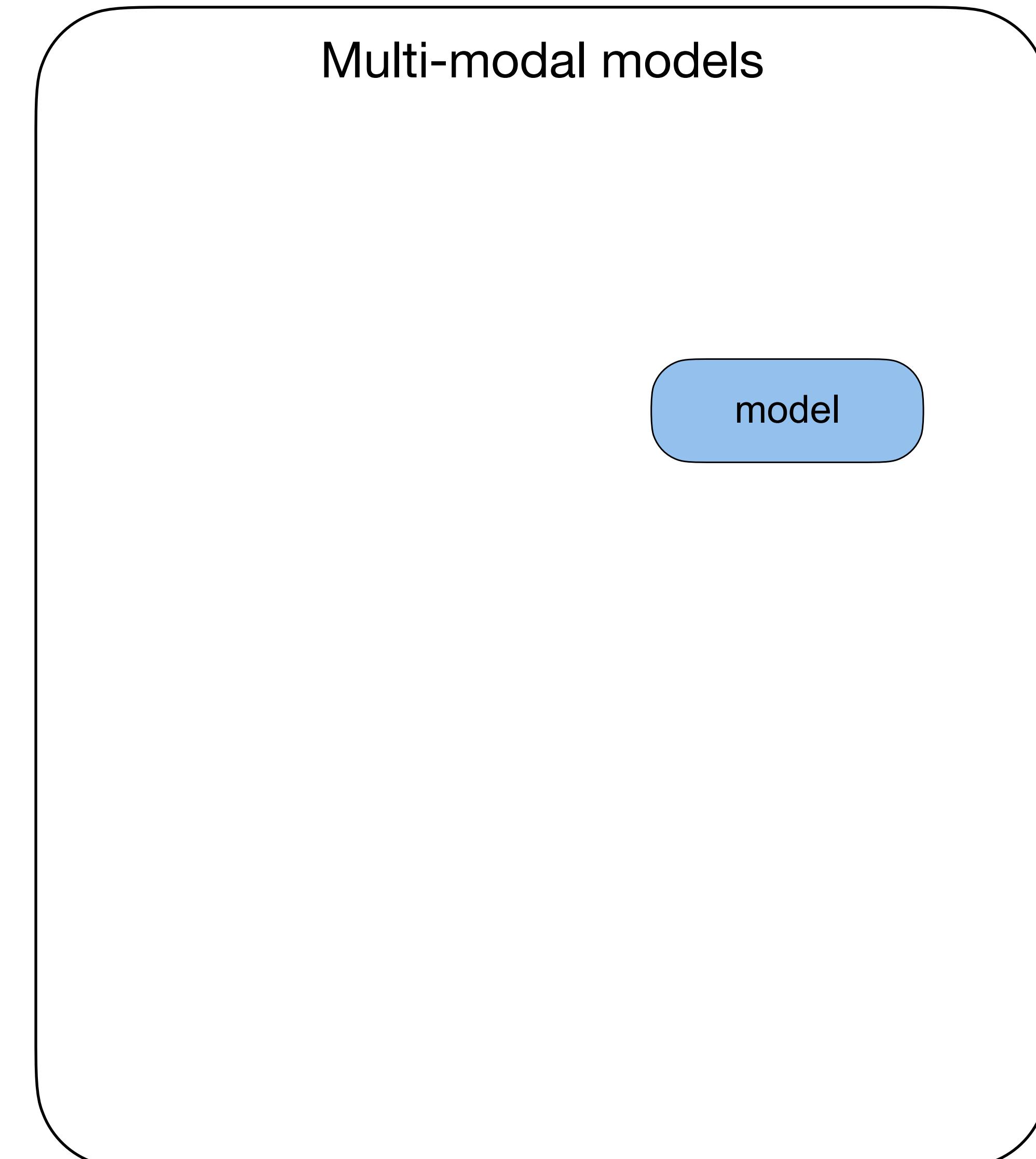
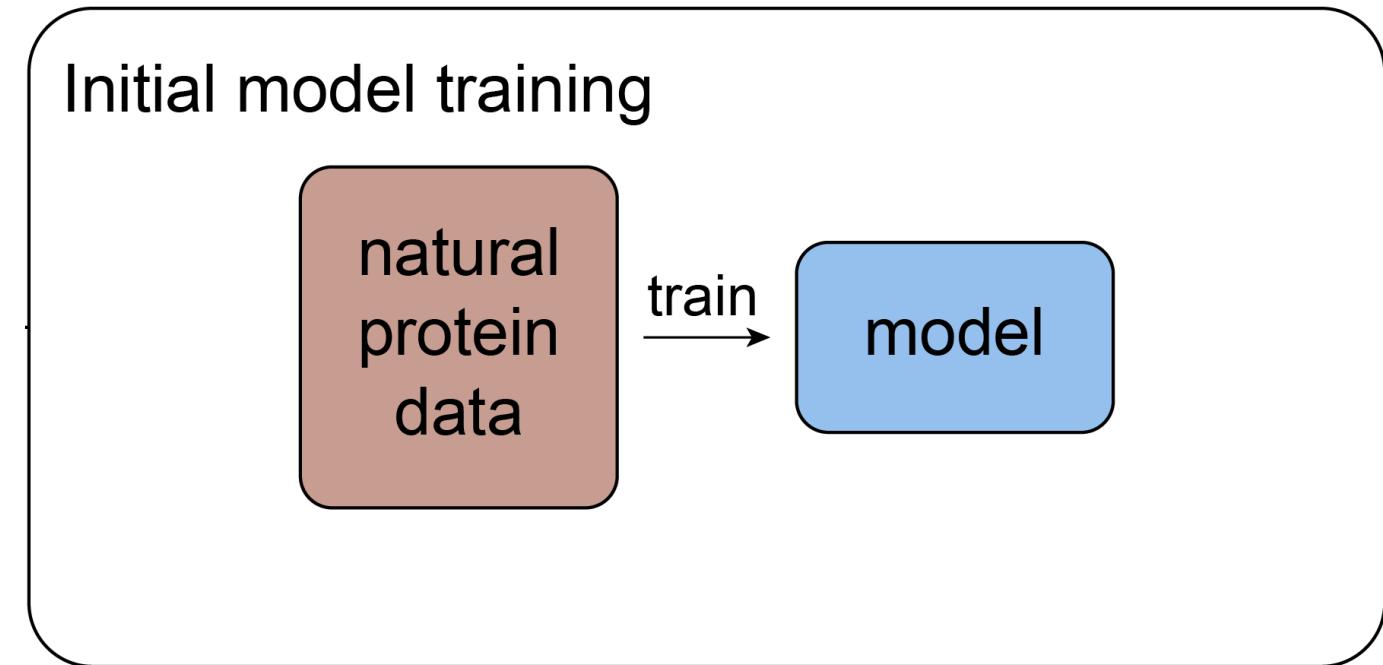
We have access to large protein databases



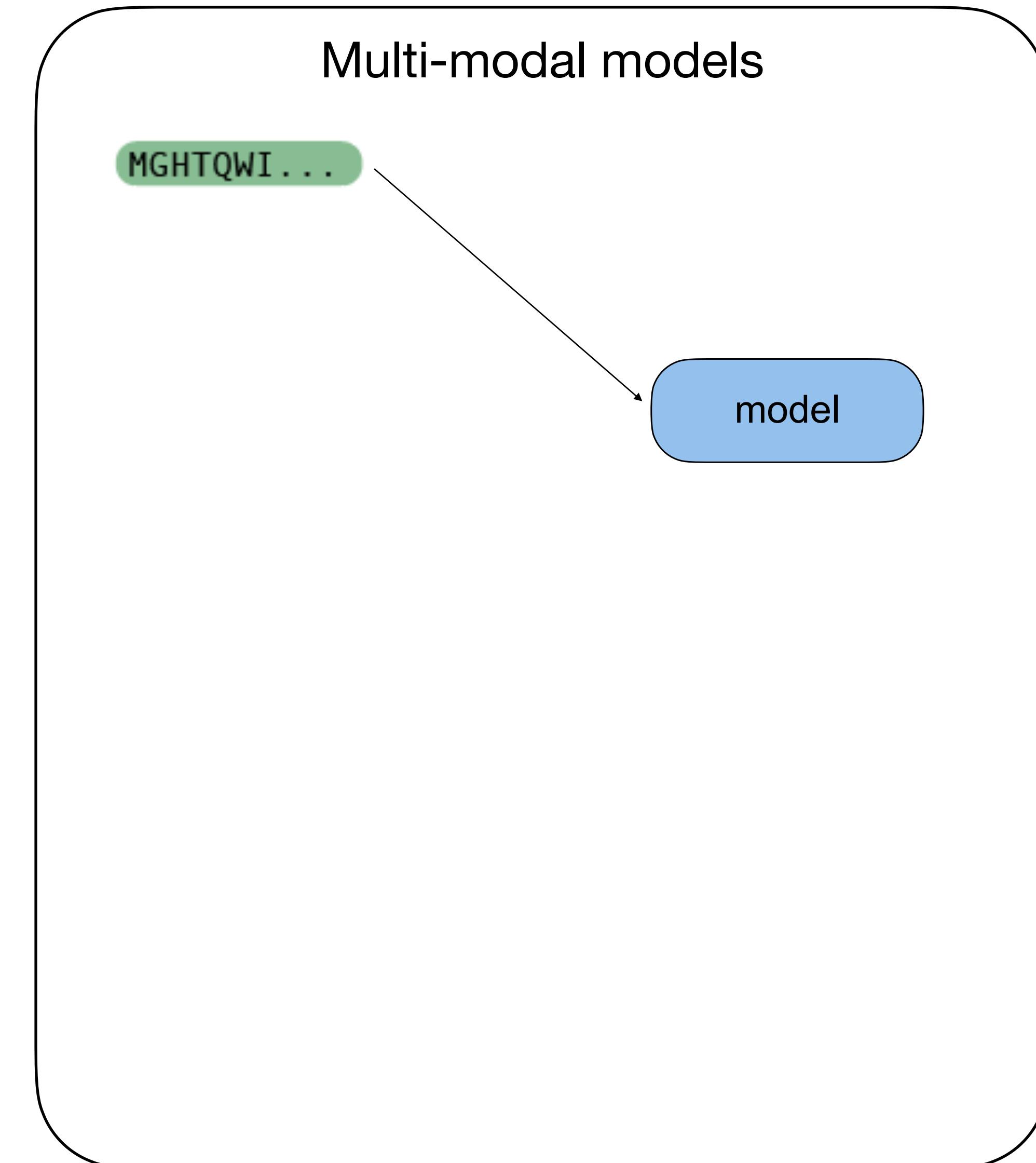
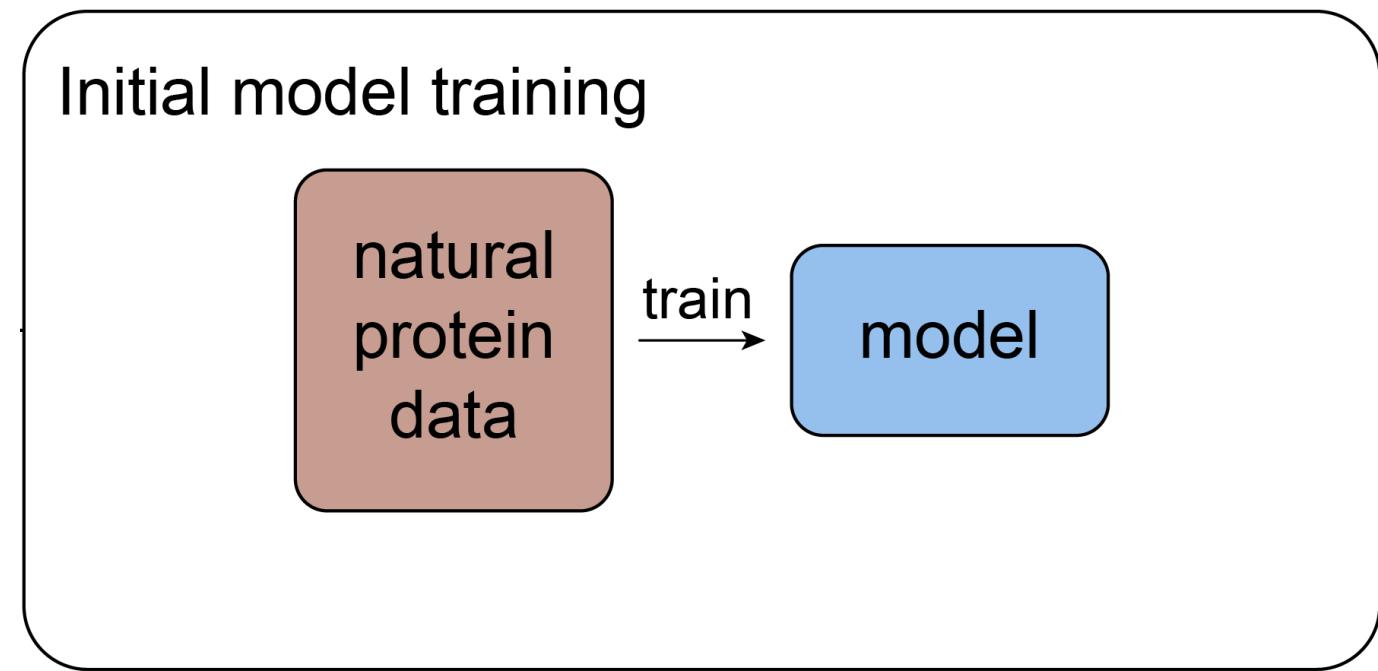
Use multiple data modalities to design proteins



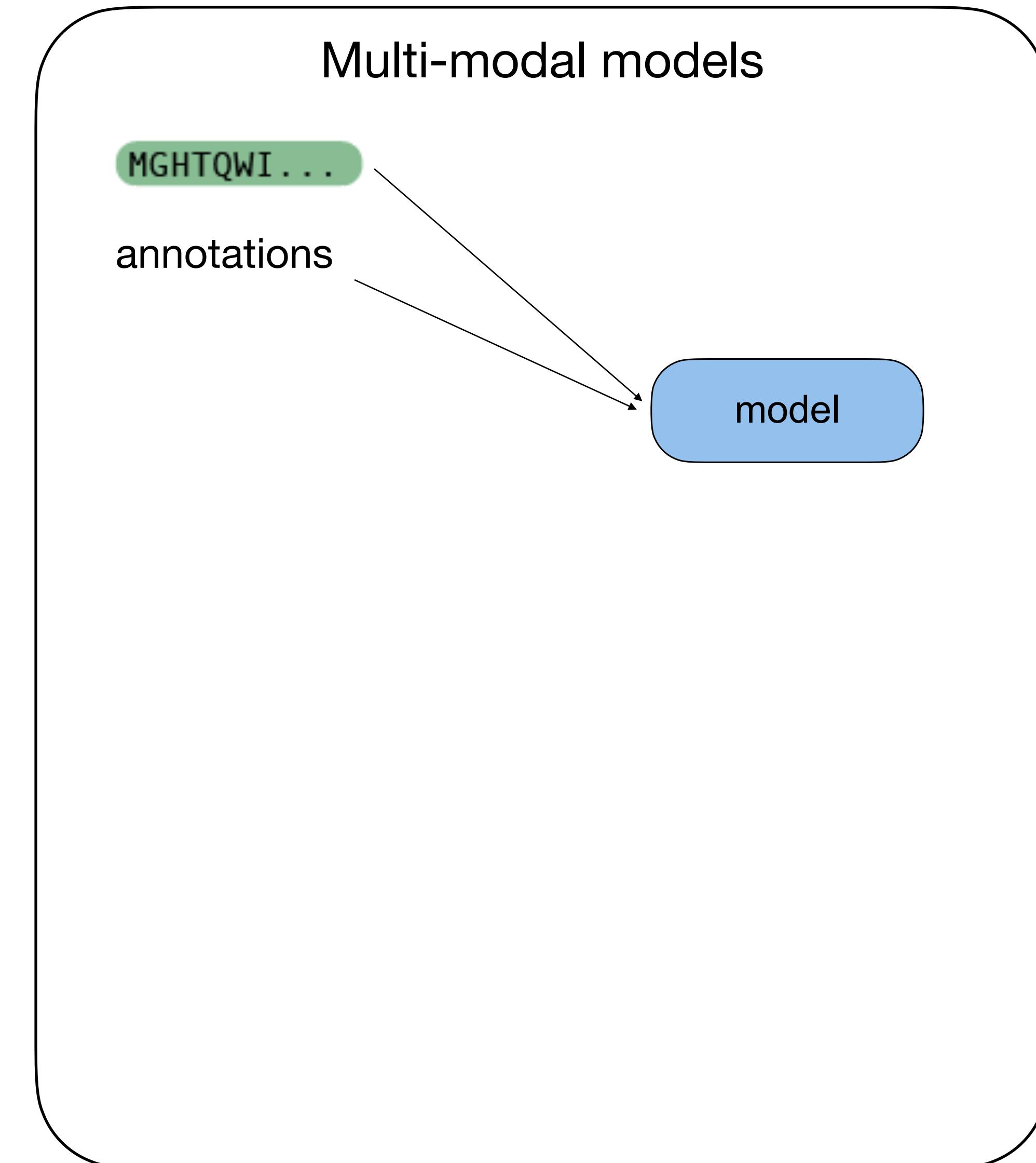
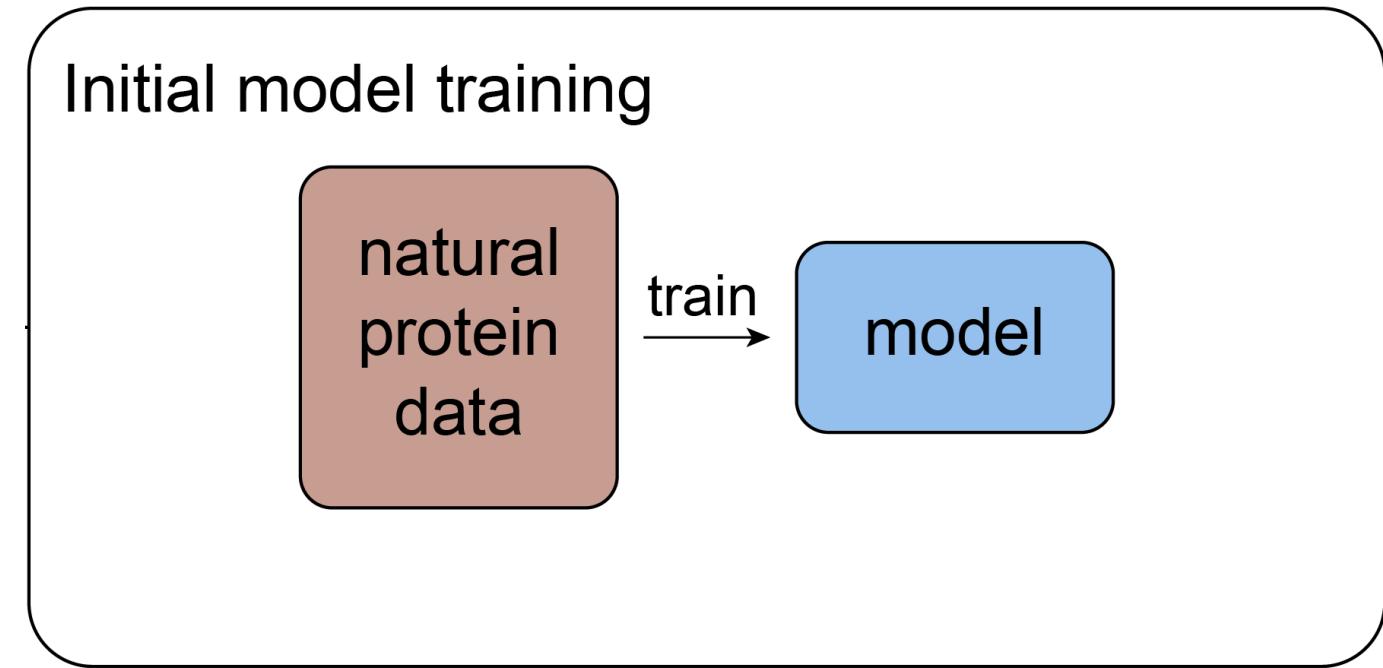
Use multiple data modalities to design proteins



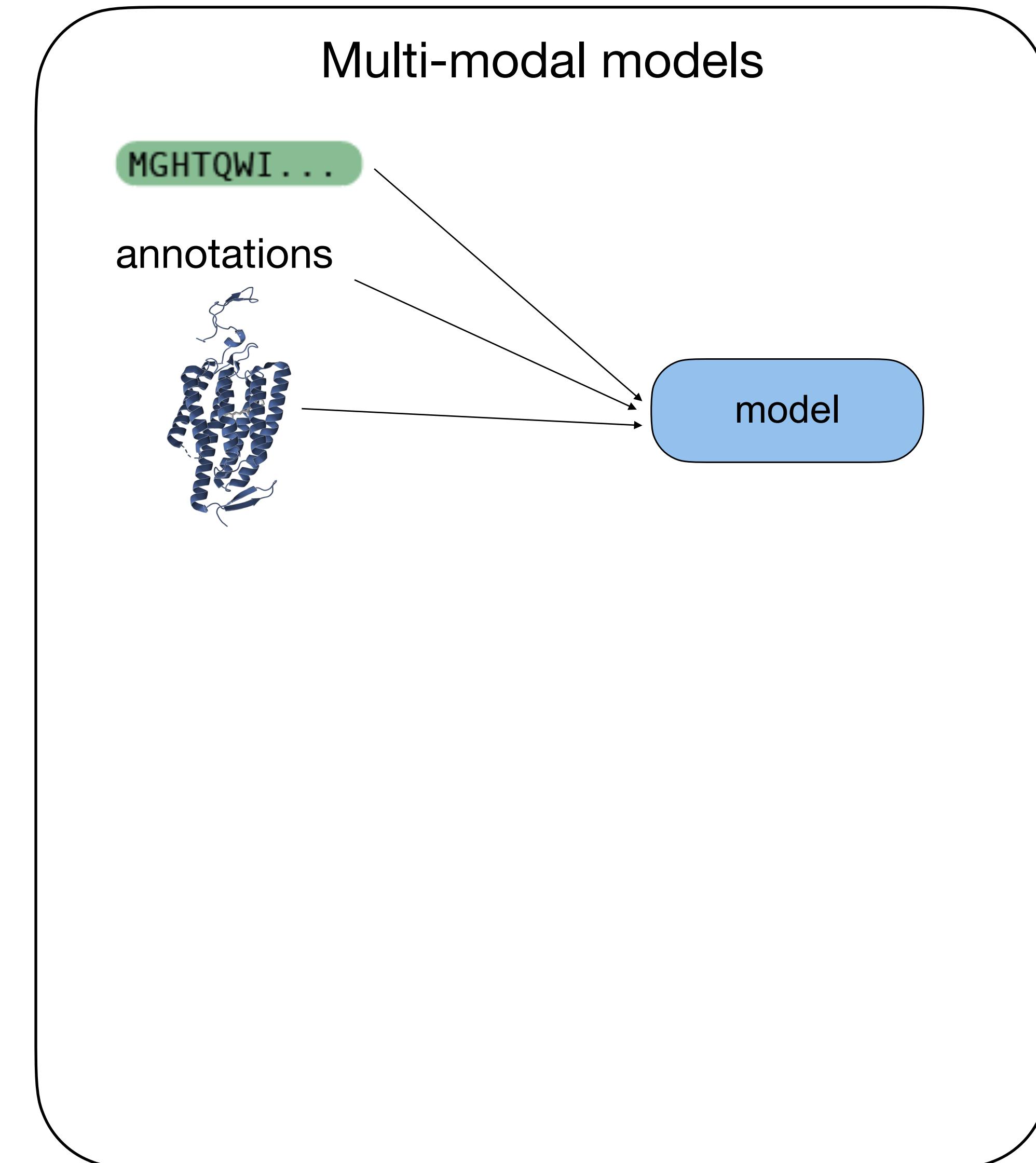
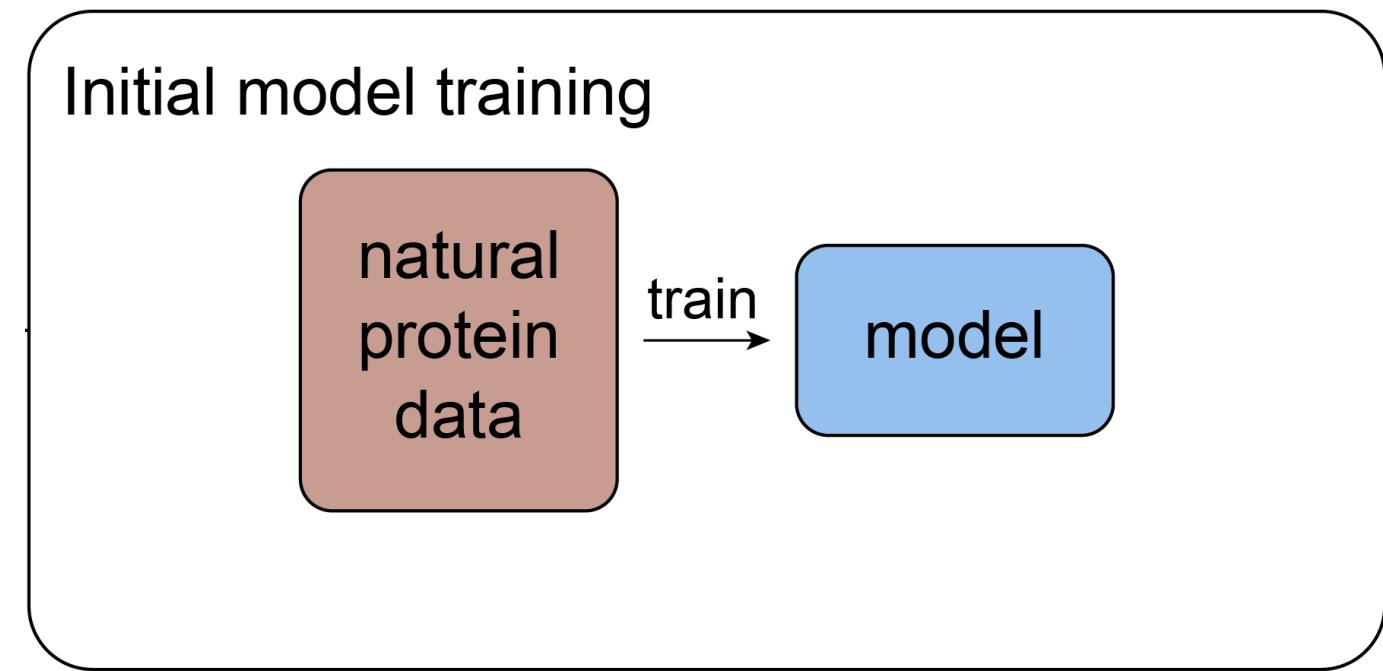
Use multiple data modalities to design proteins



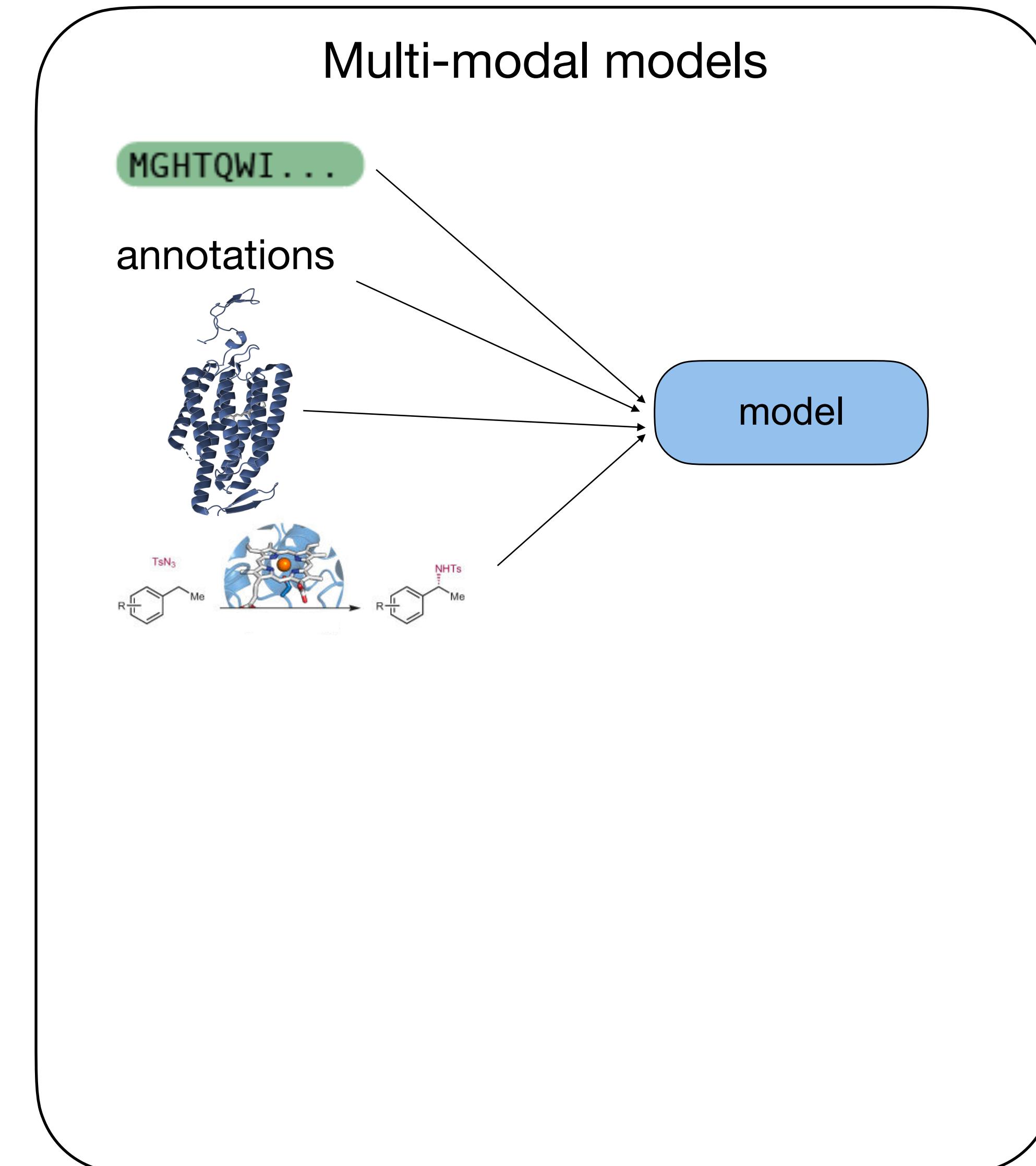
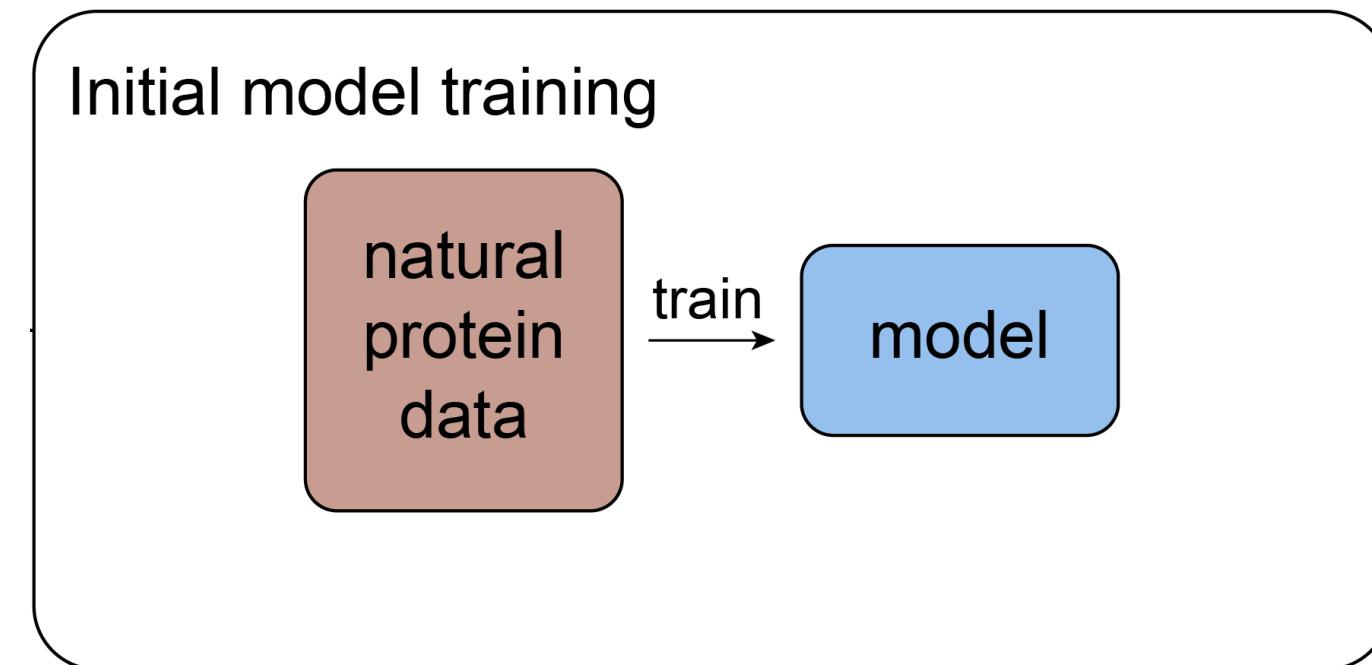
Use multiple data modalities to design proteins



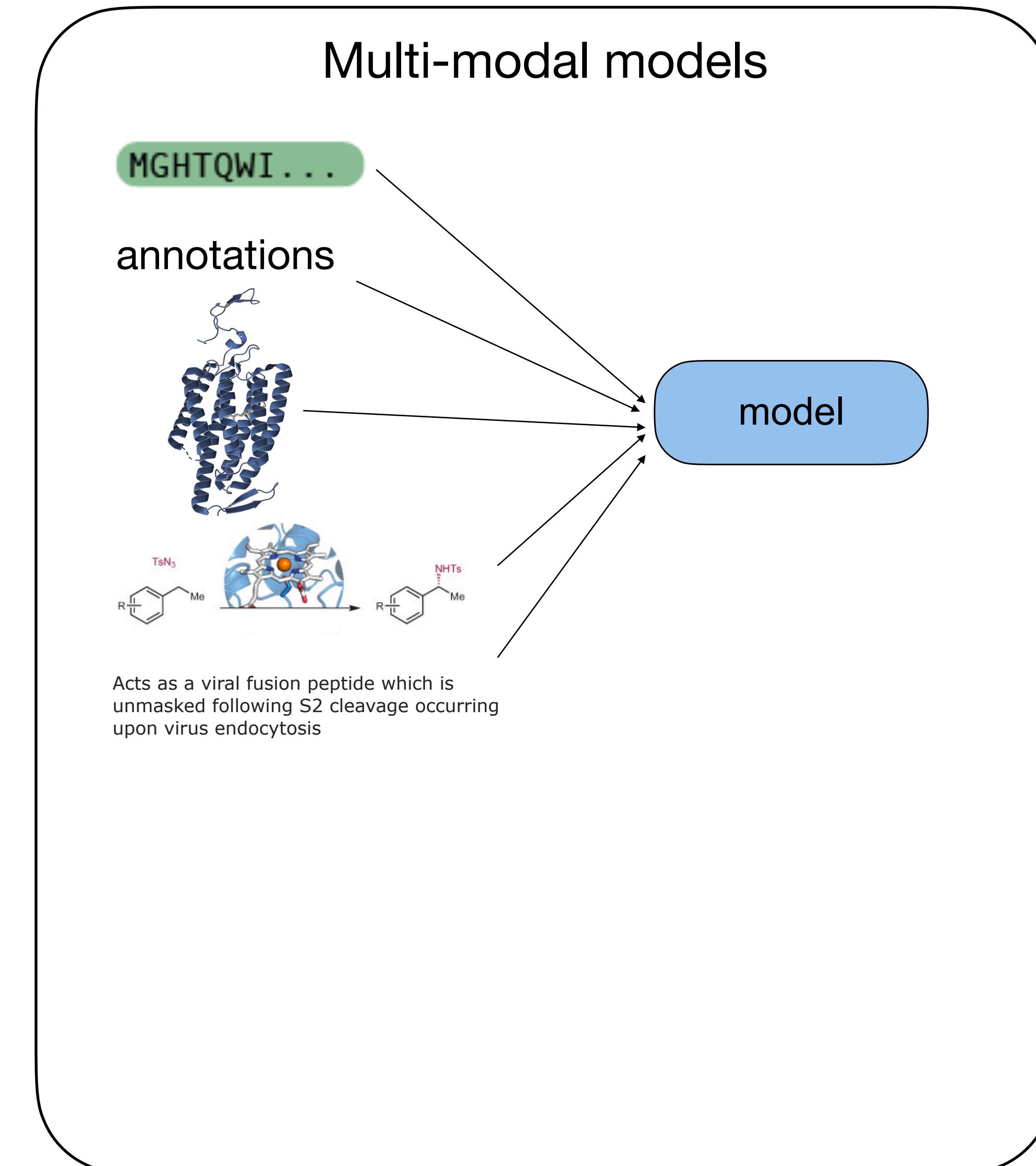
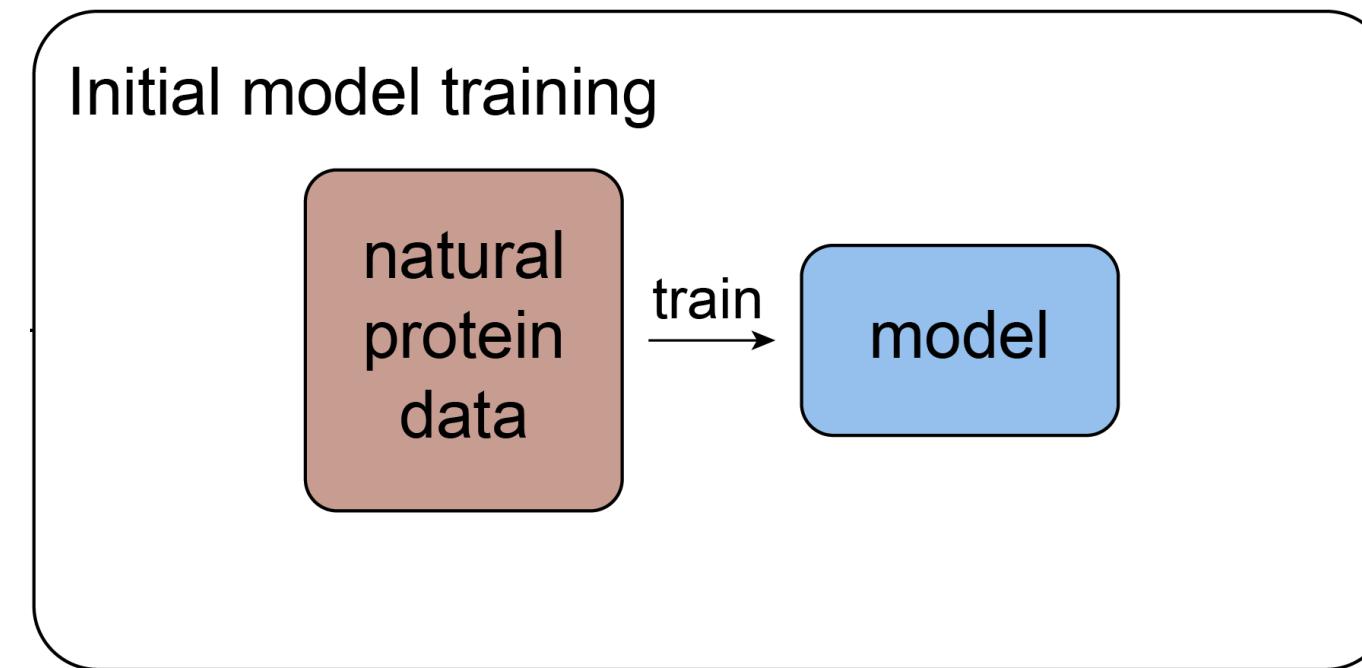
Use multiple data modalities to design proteins



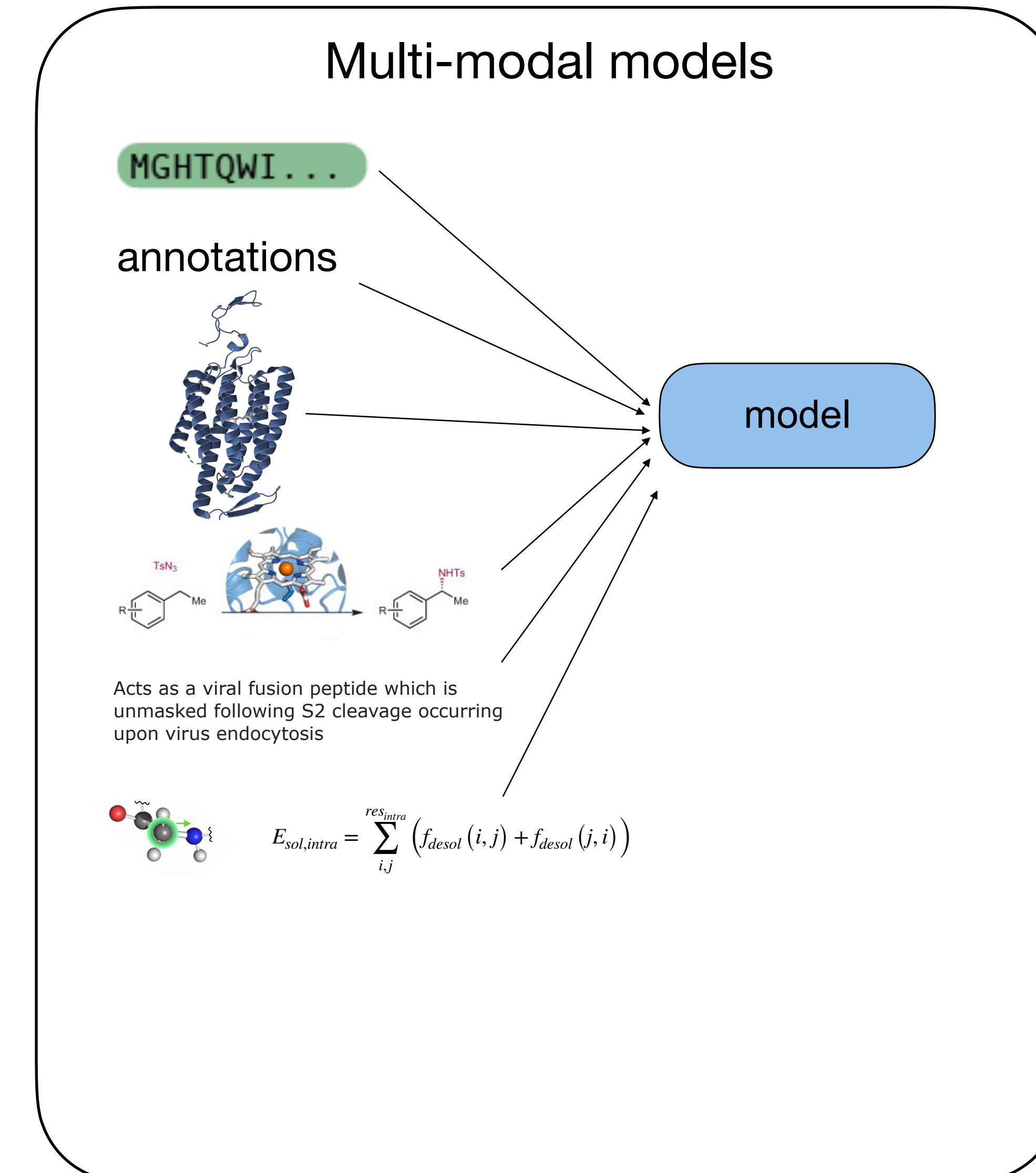
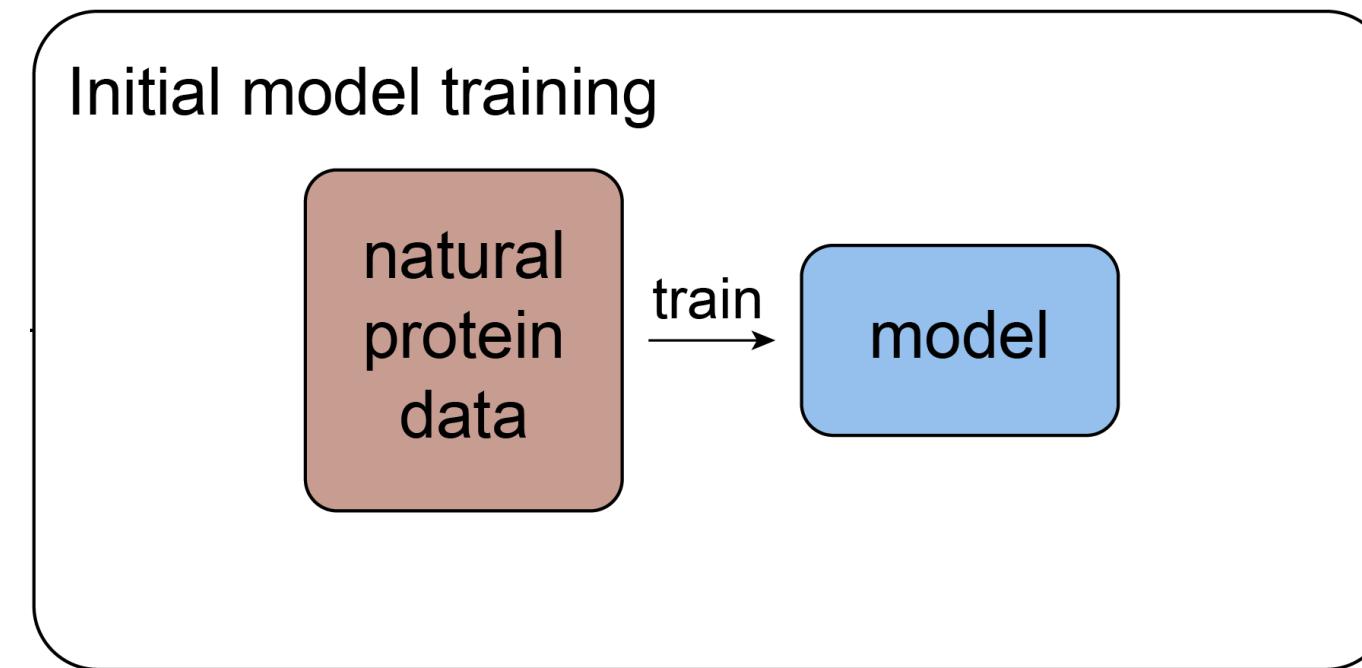
Use multiple data modalities to design proteins



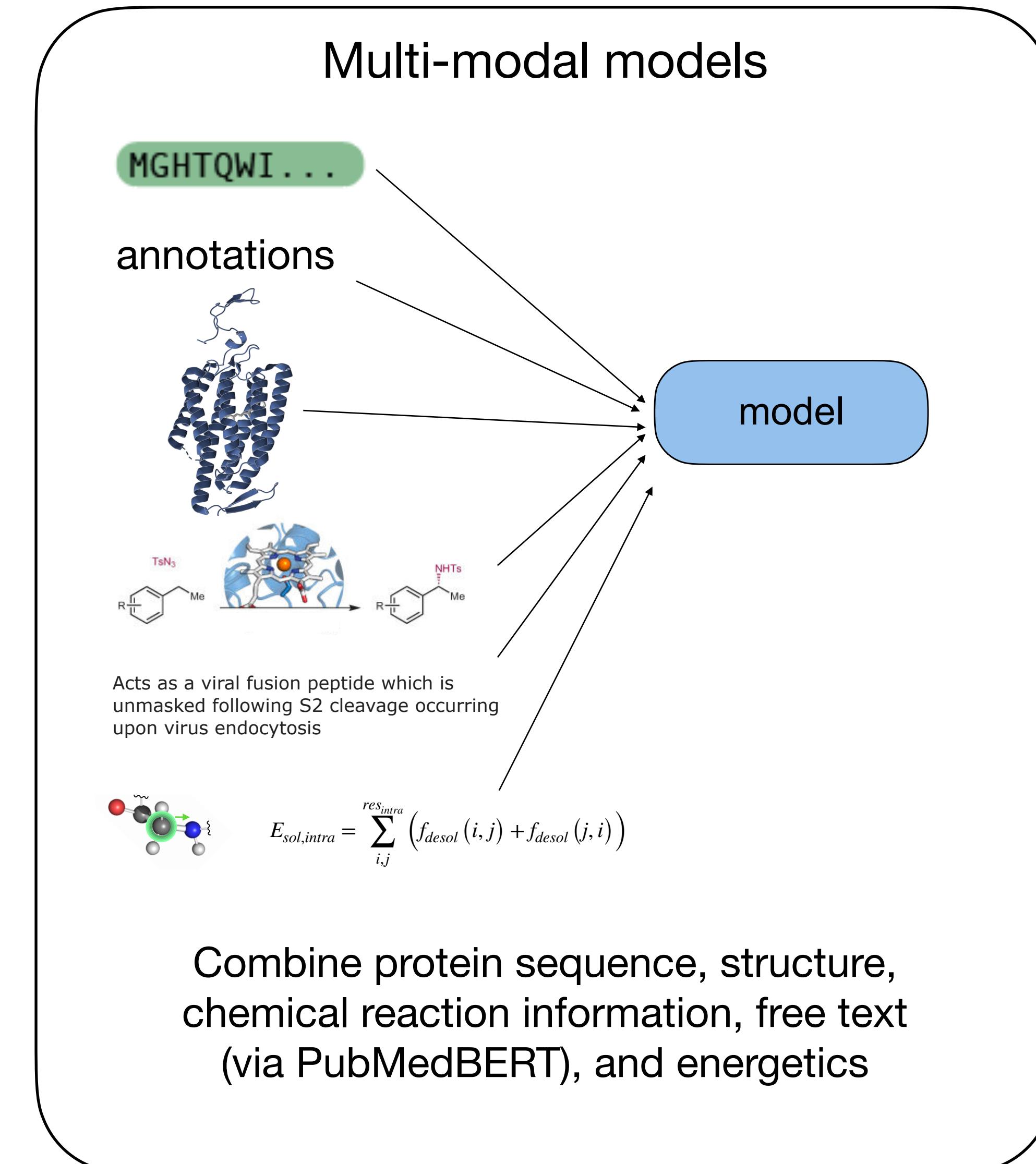
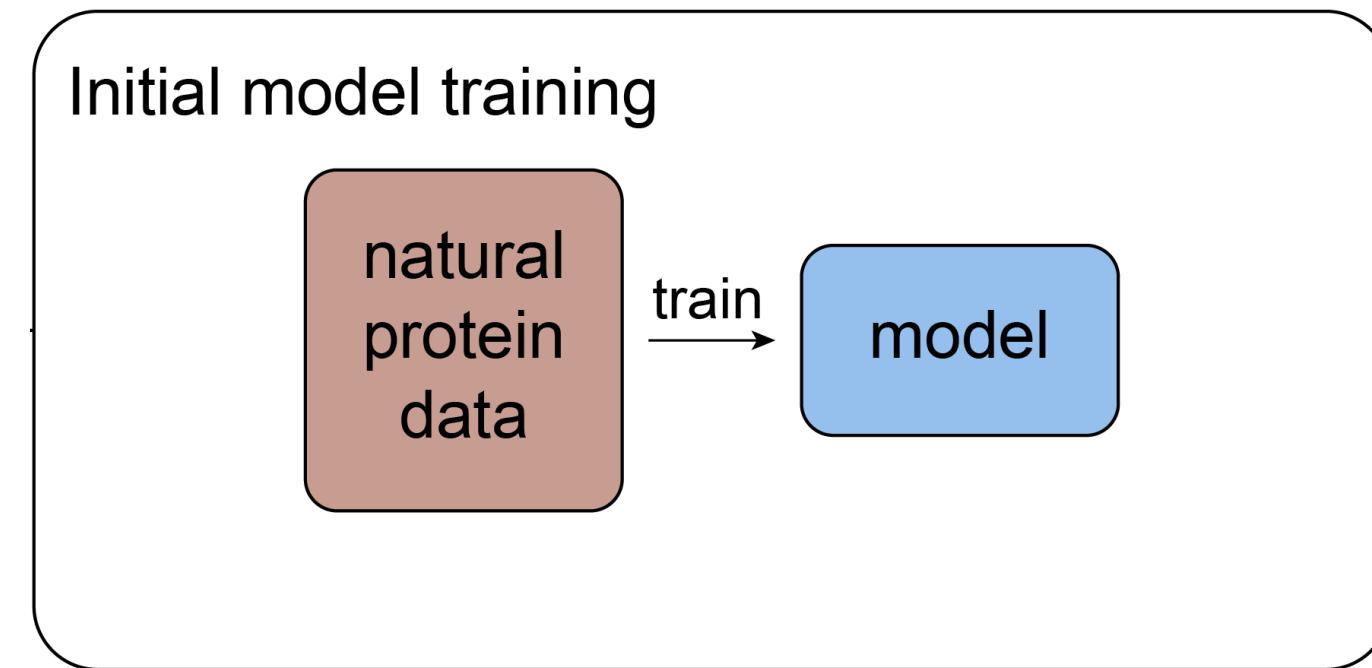
Use multiple data modalities to design proteins



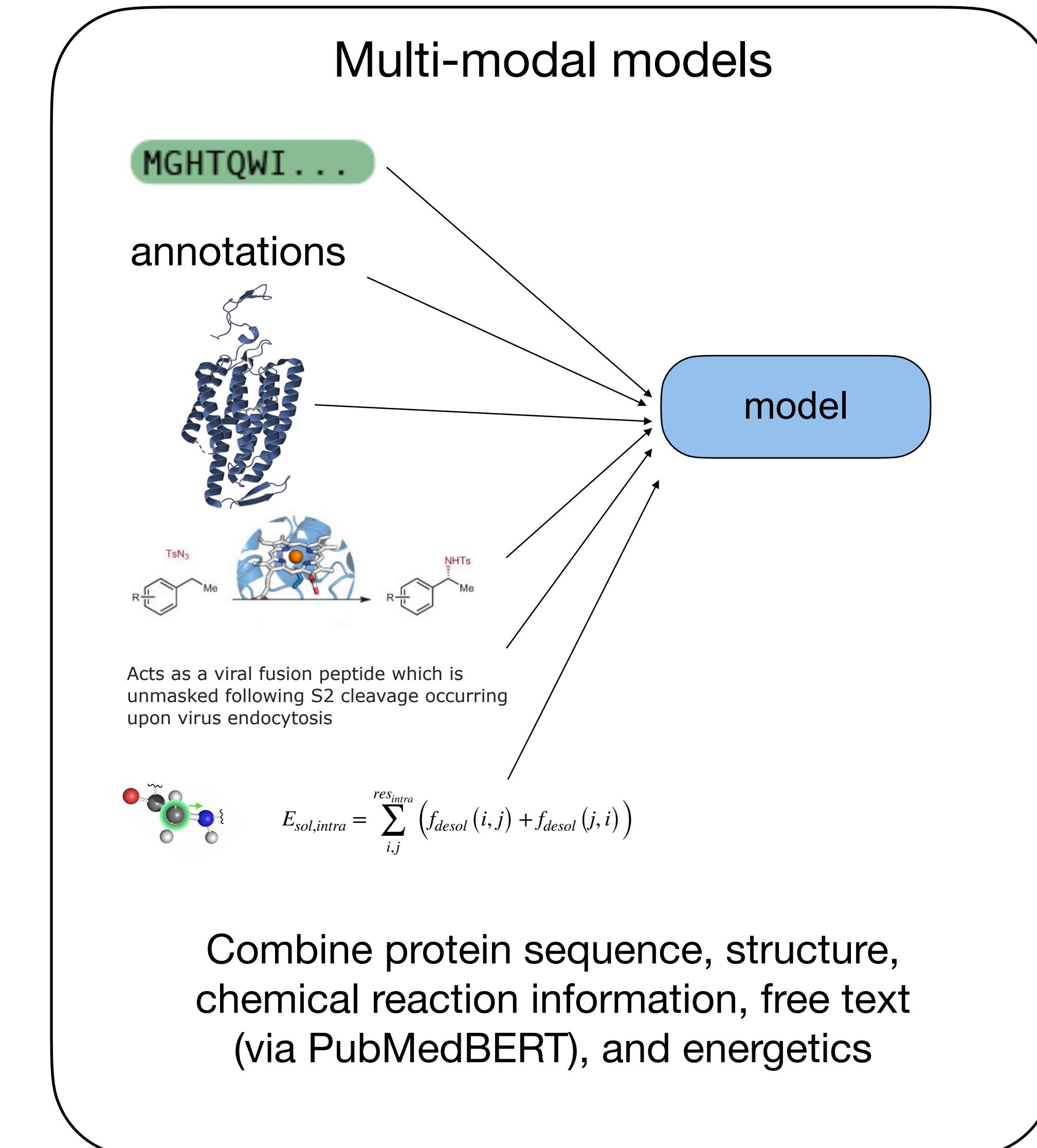
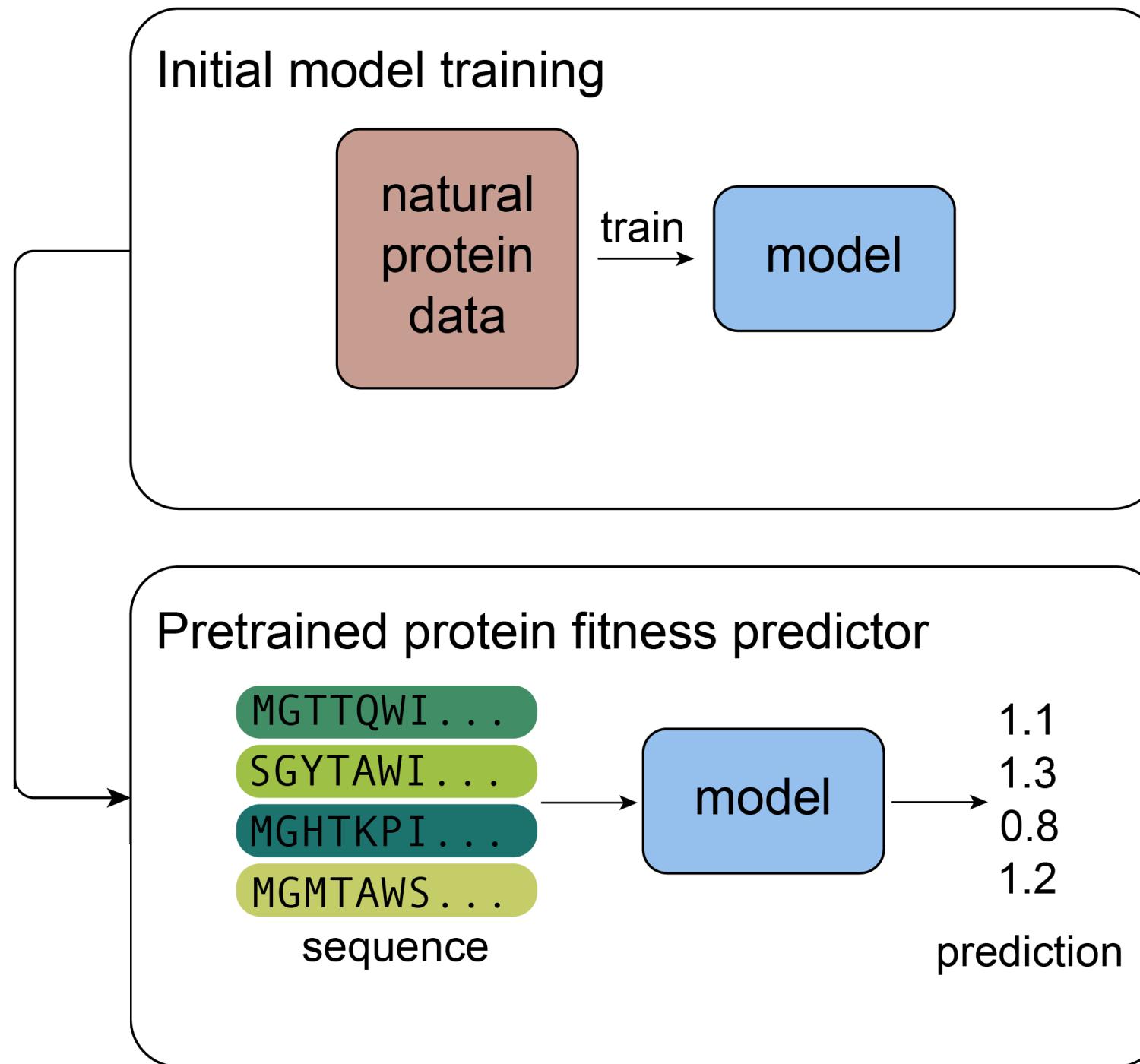
Use multiple data modalities to design proteins



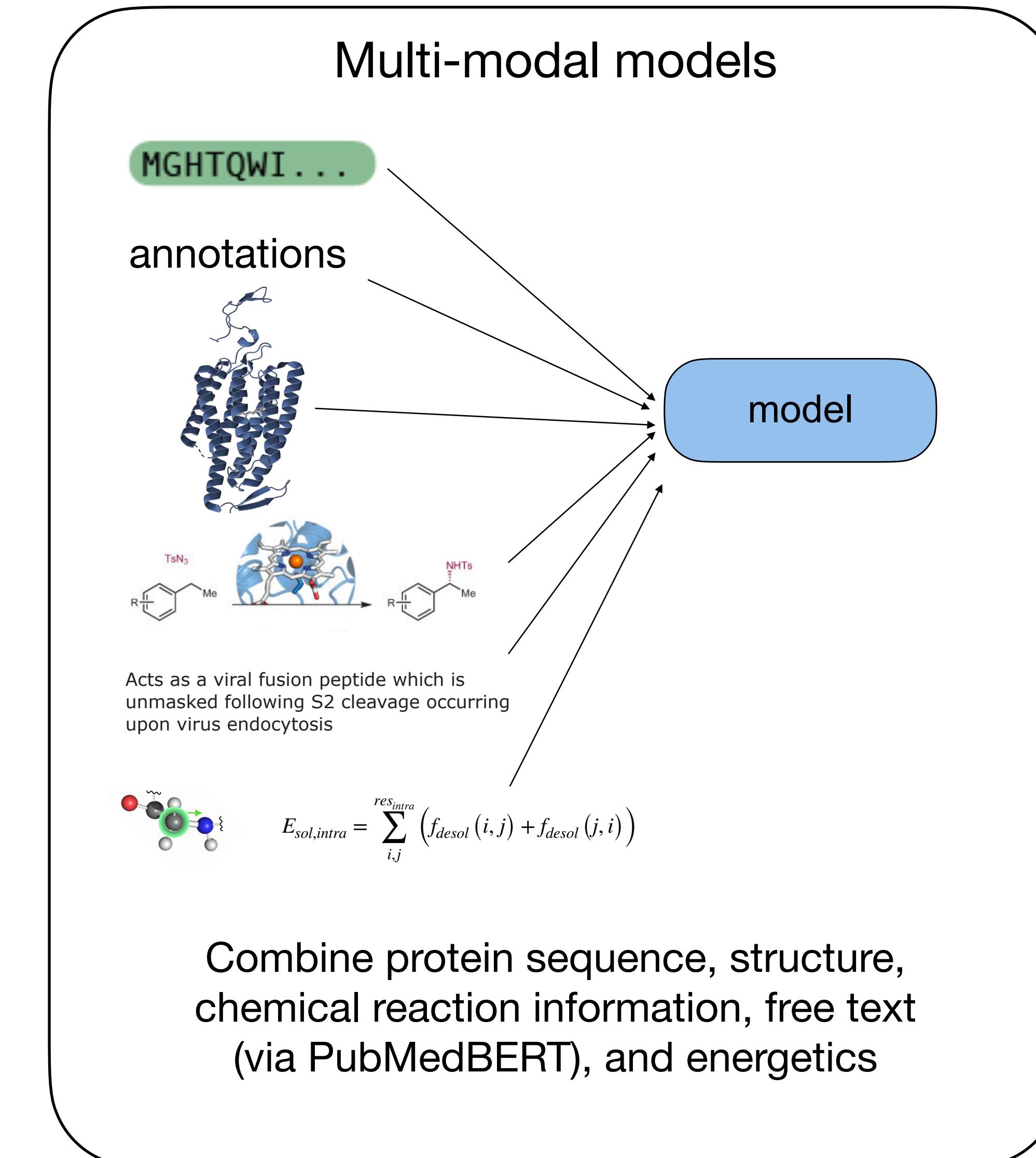
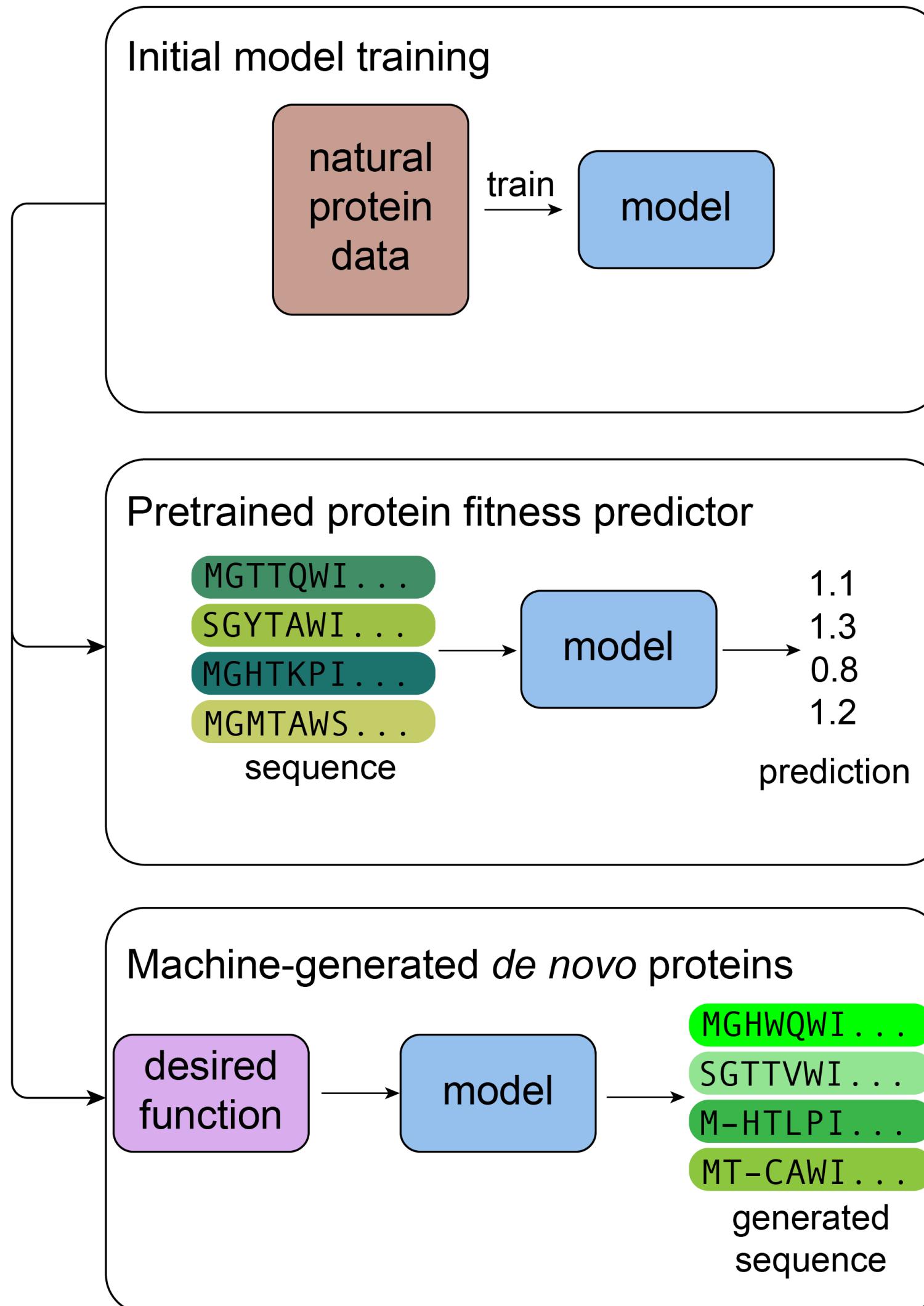
Use multiple data modalities to design proteins



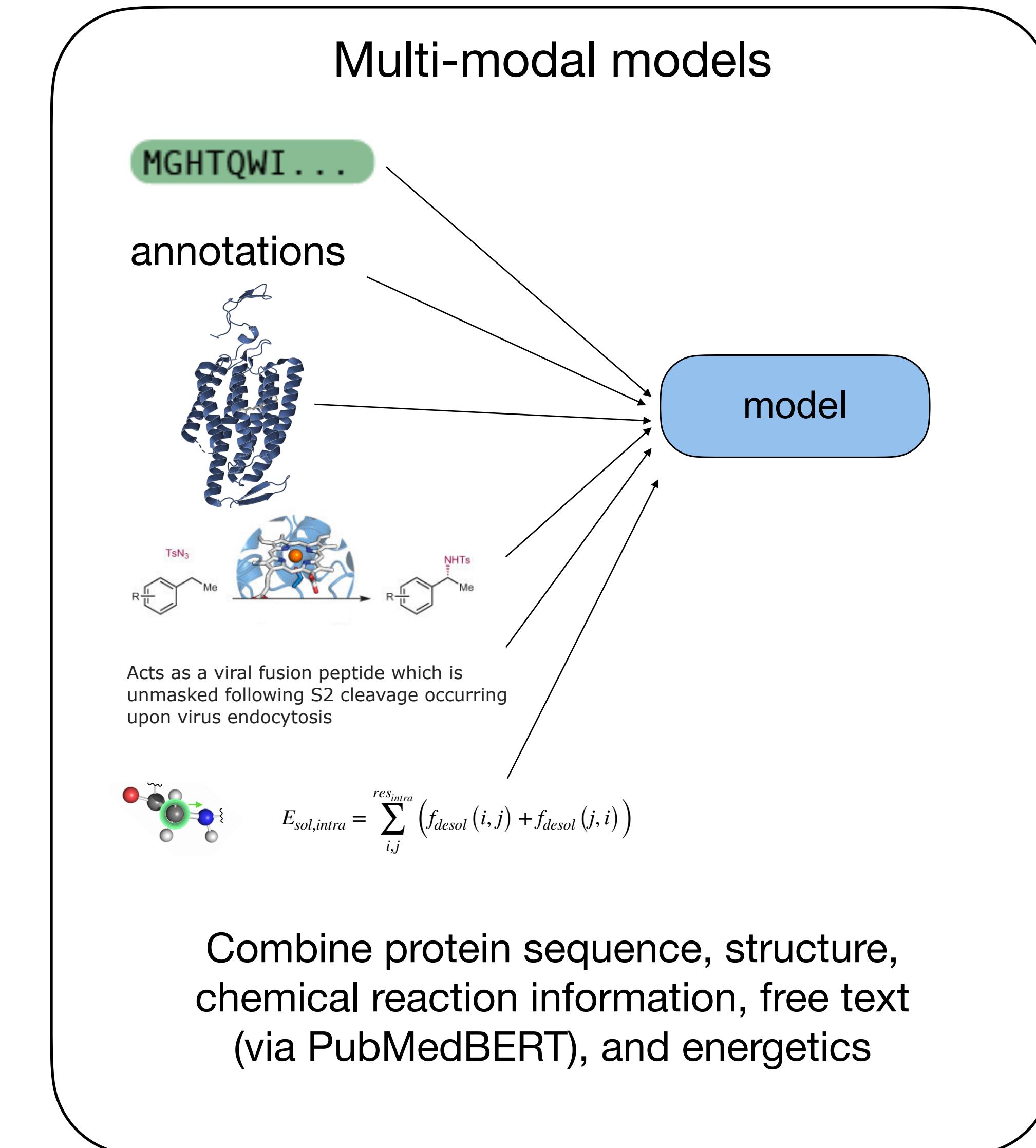
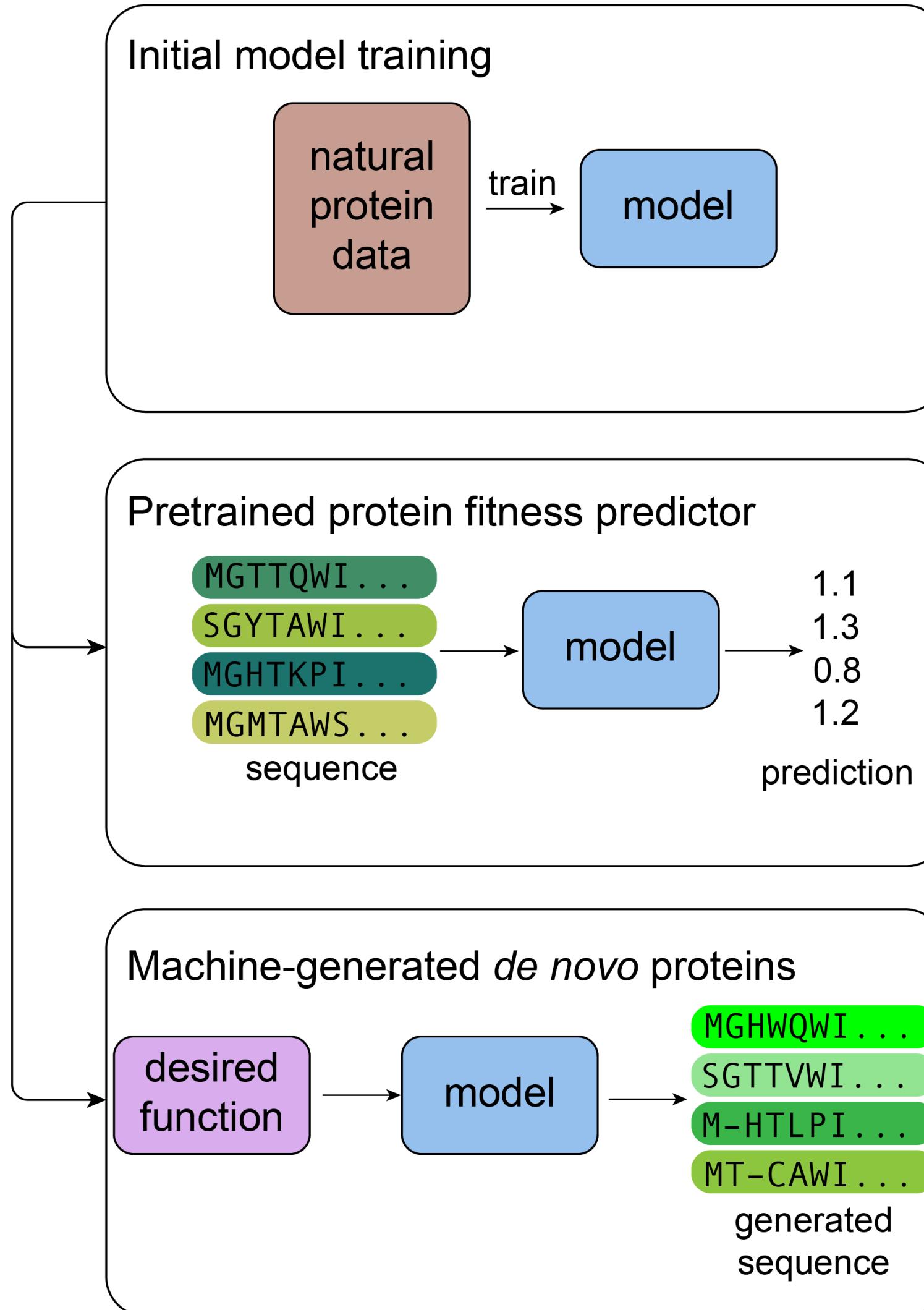
Use multiple data modalities to design proteins



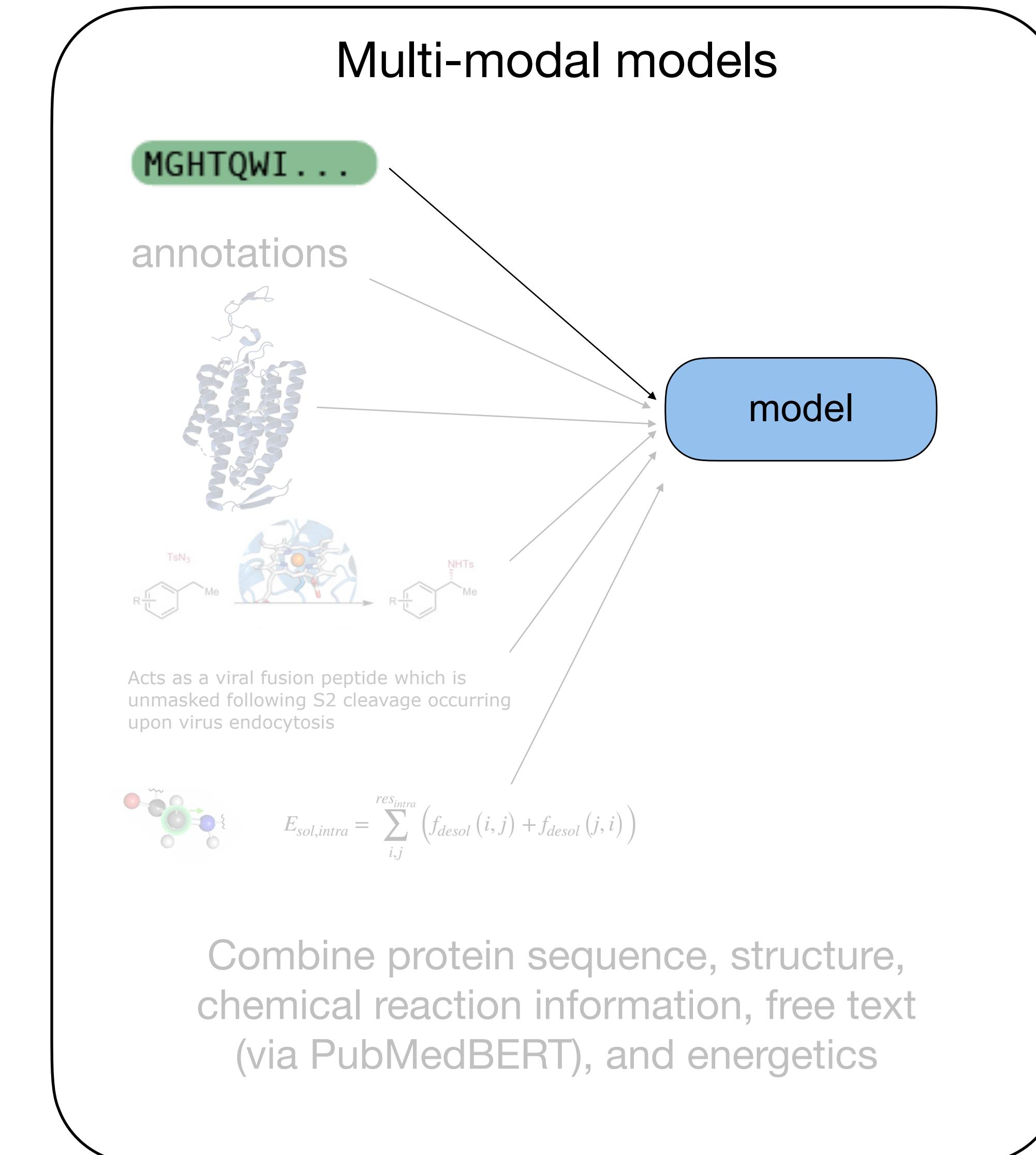
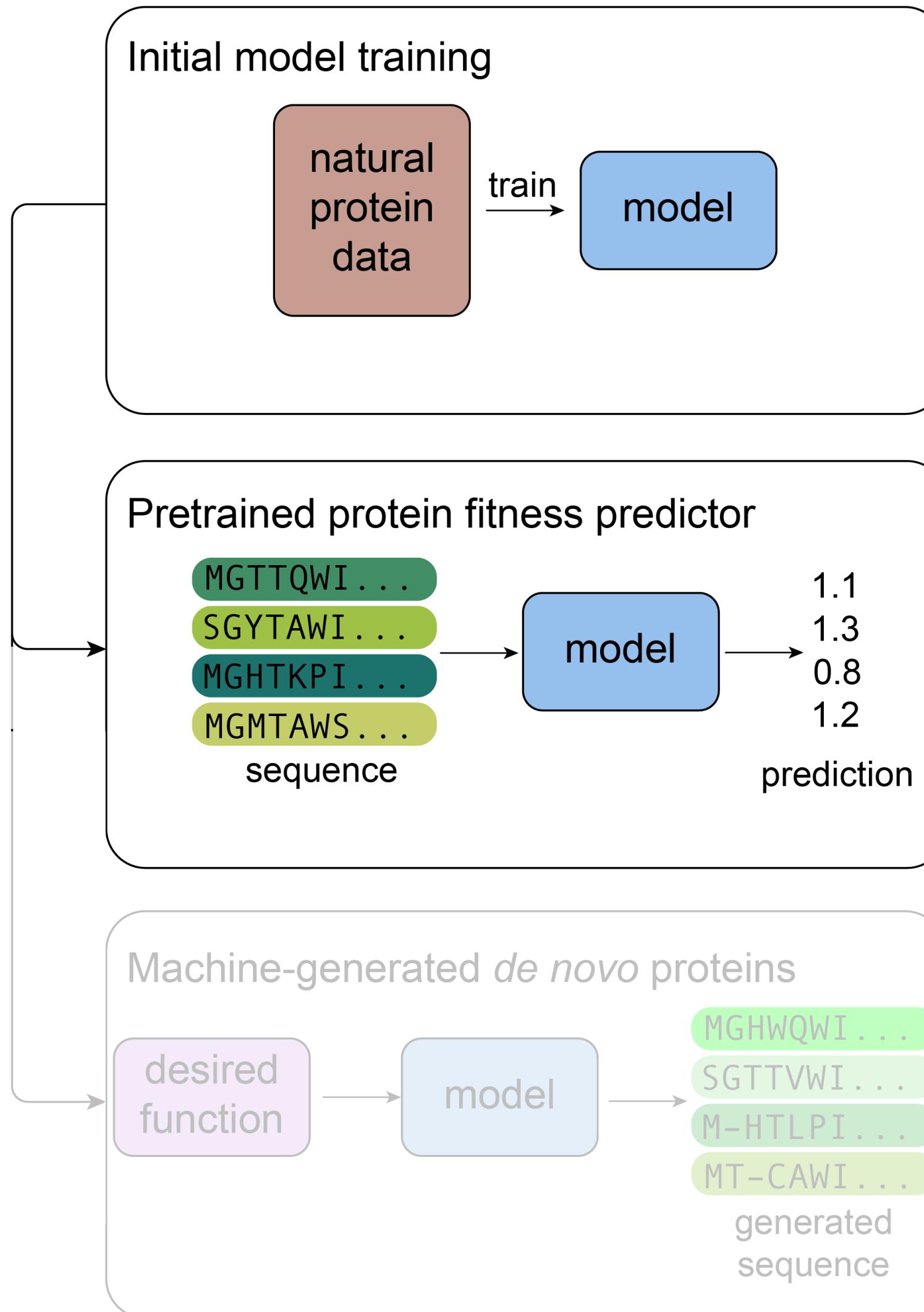
Use multiple data modalities to design proteins



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

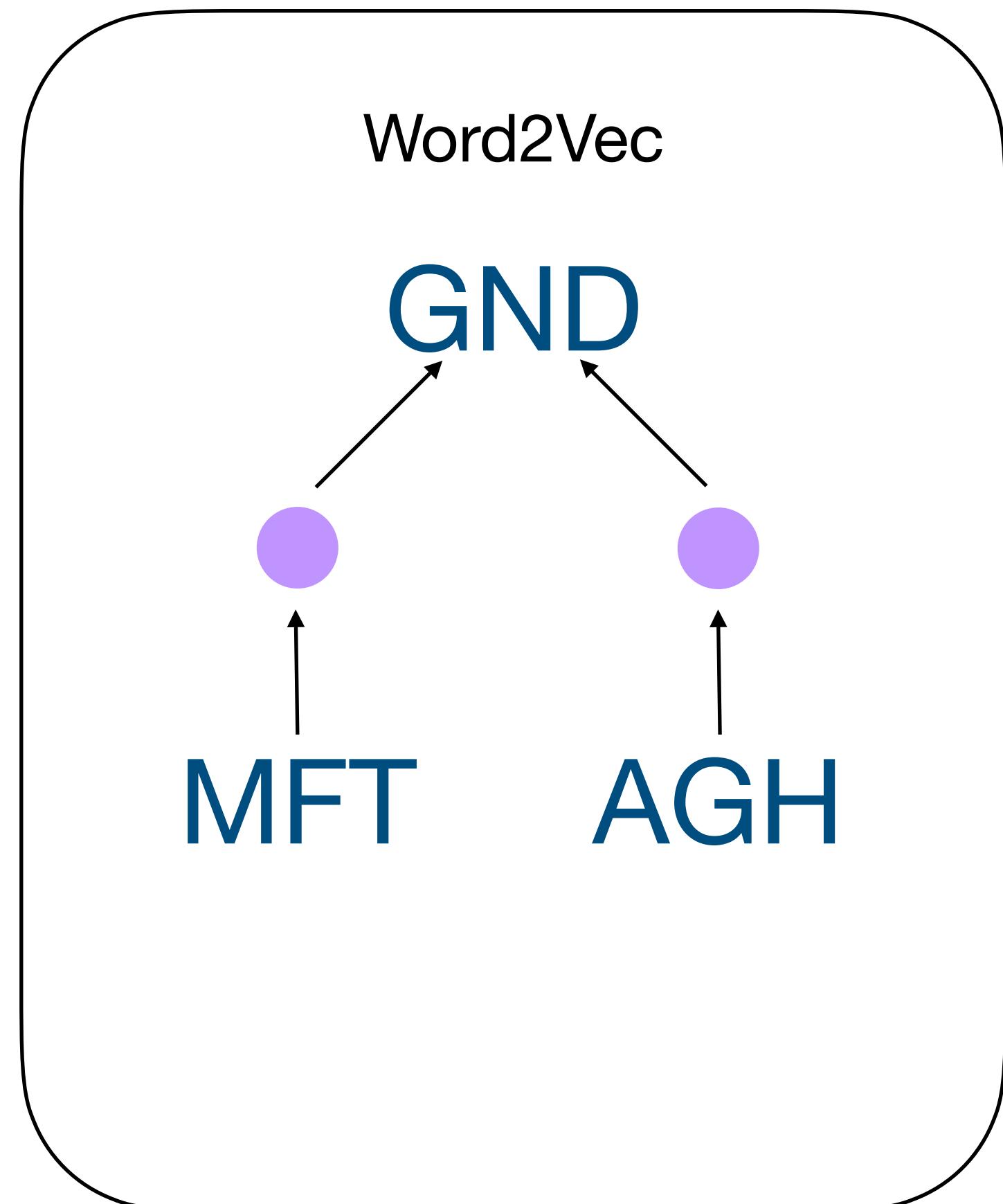
Many methods pretend proteins are language

Word2Vec

MFT AGH

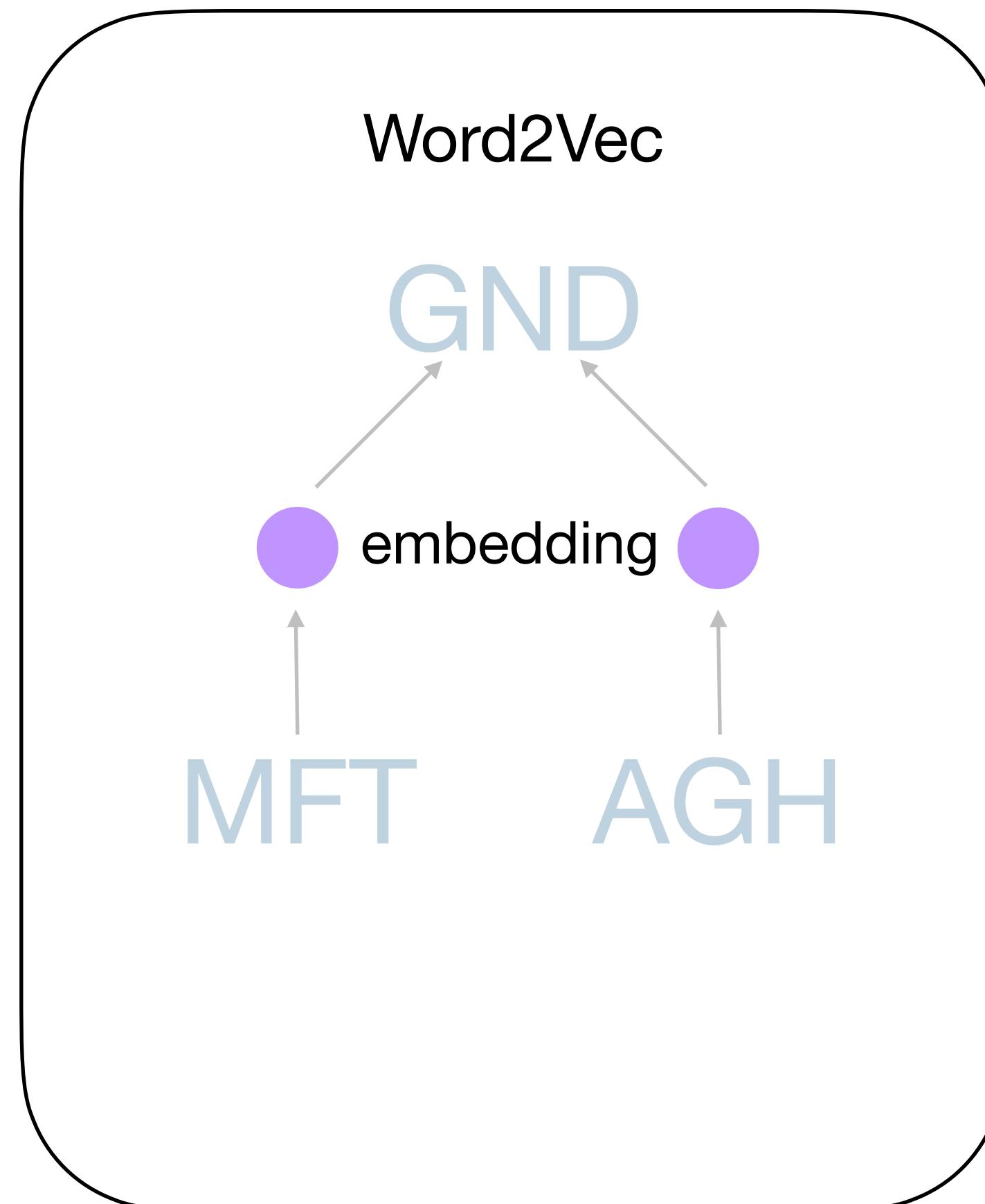
MFTGNDAGH

Many methods pretend proteins are language



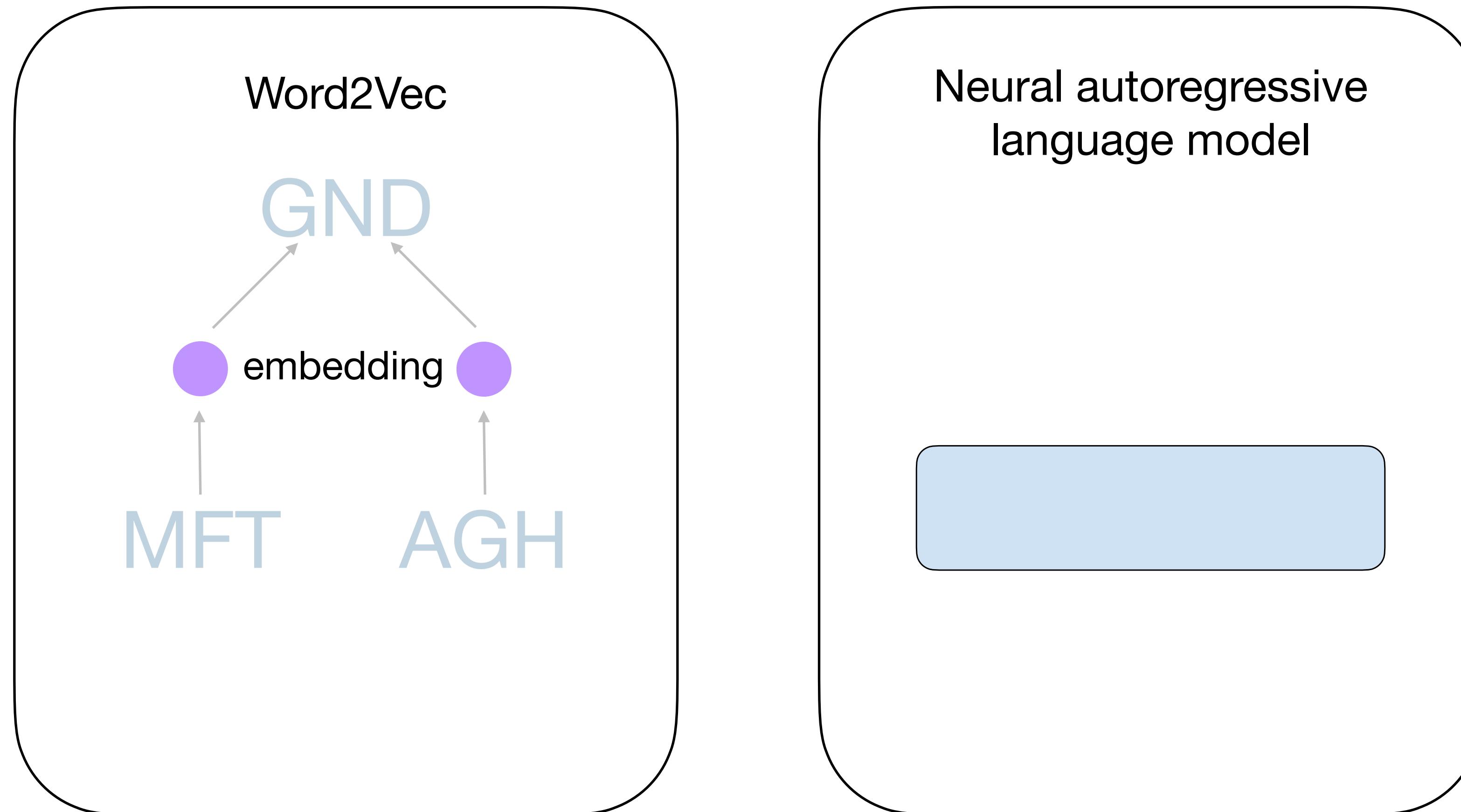
MFTGNDAGH

Many methods pretend proteins are language



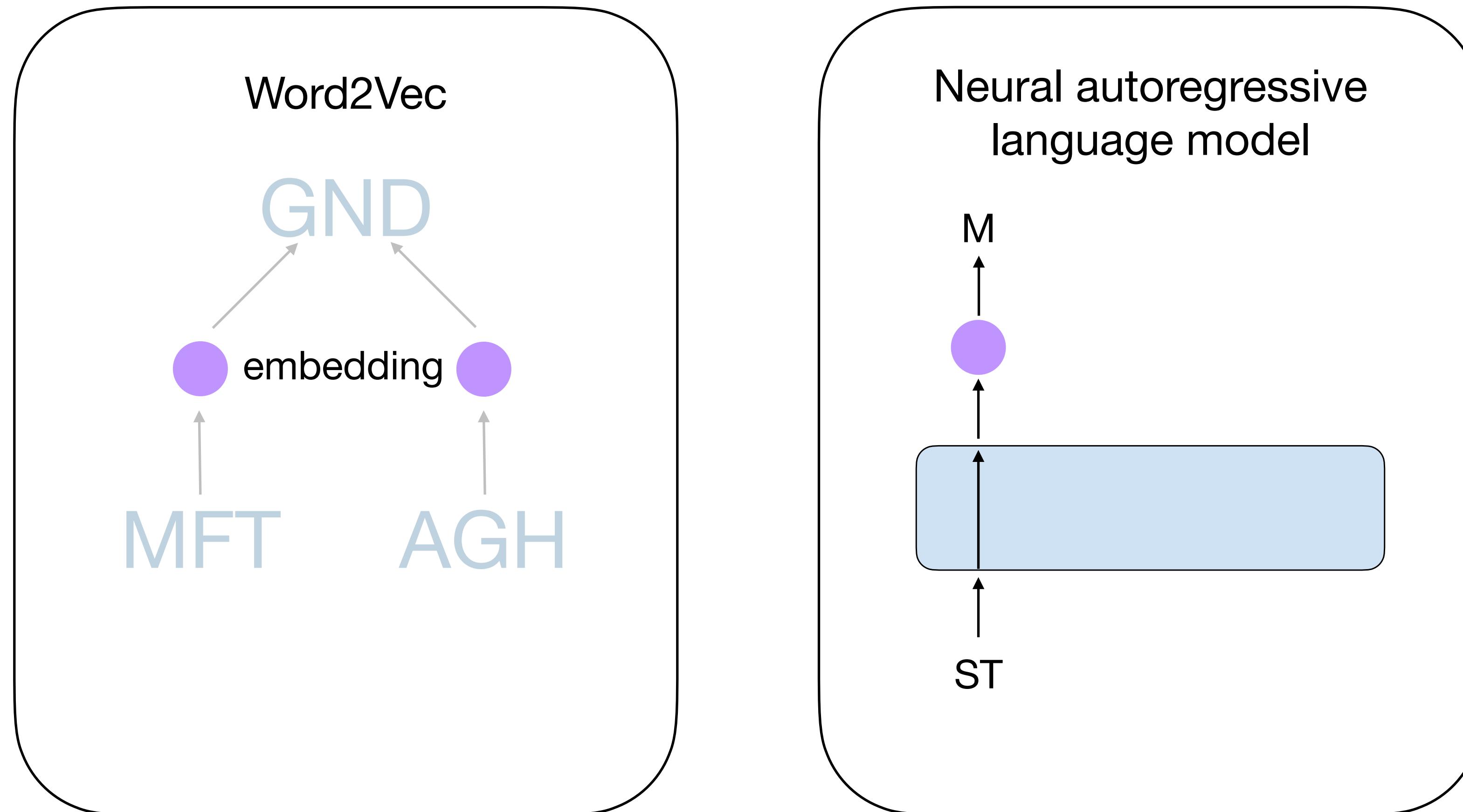
MFTGNDAGH

Many methods pretend proteins are language



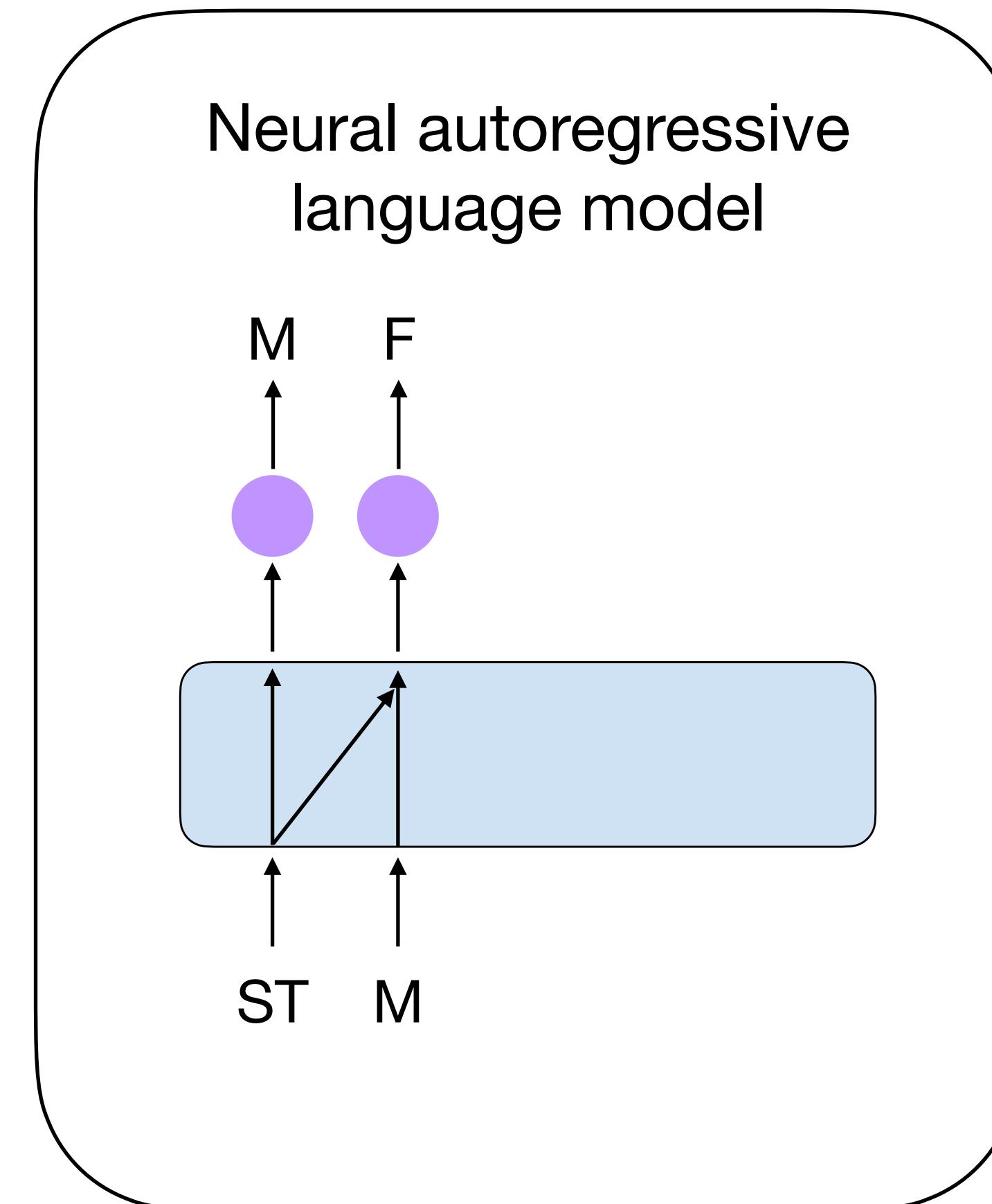
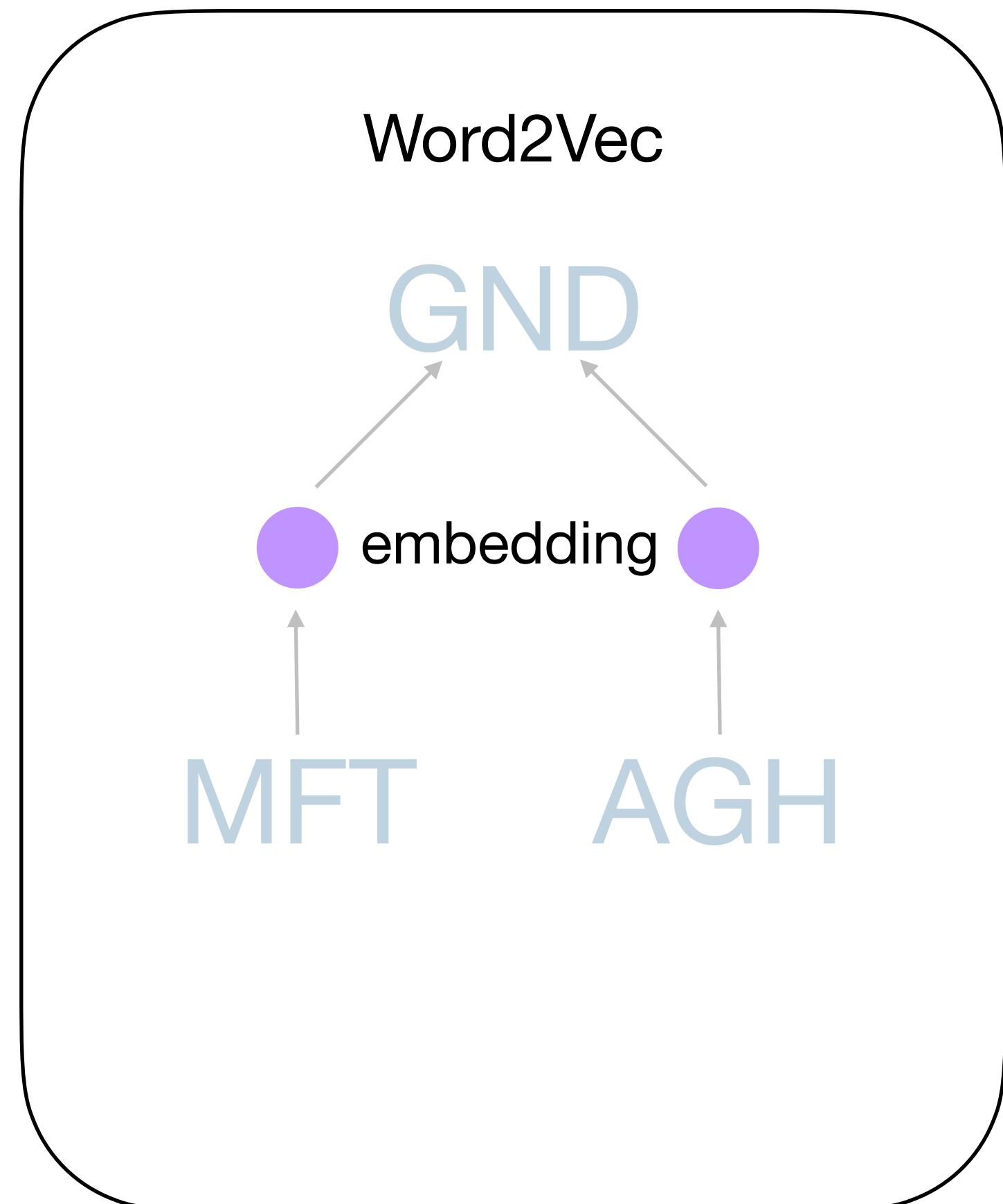
MFTGNDAGH

Many methods pretend proteins are language



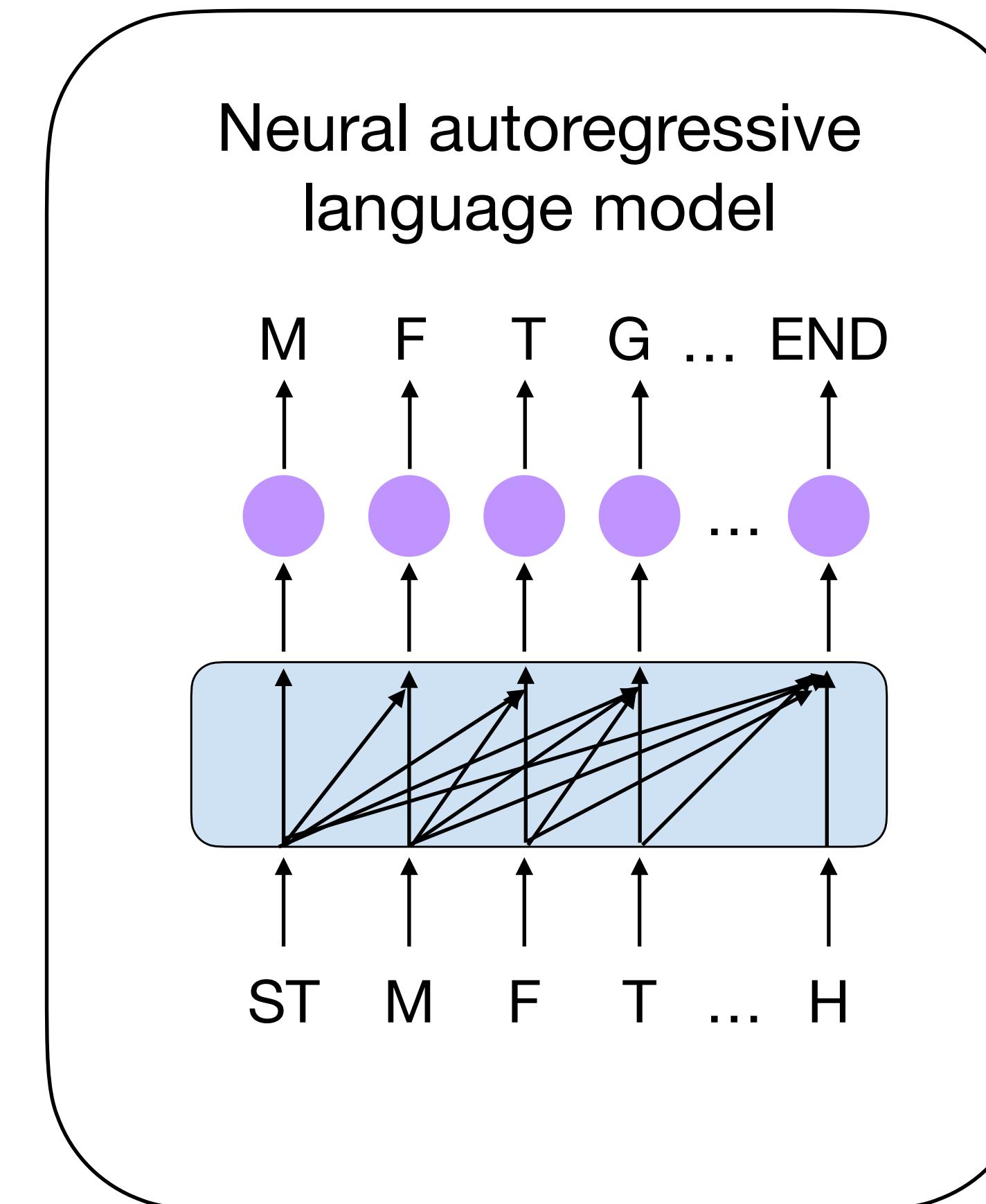
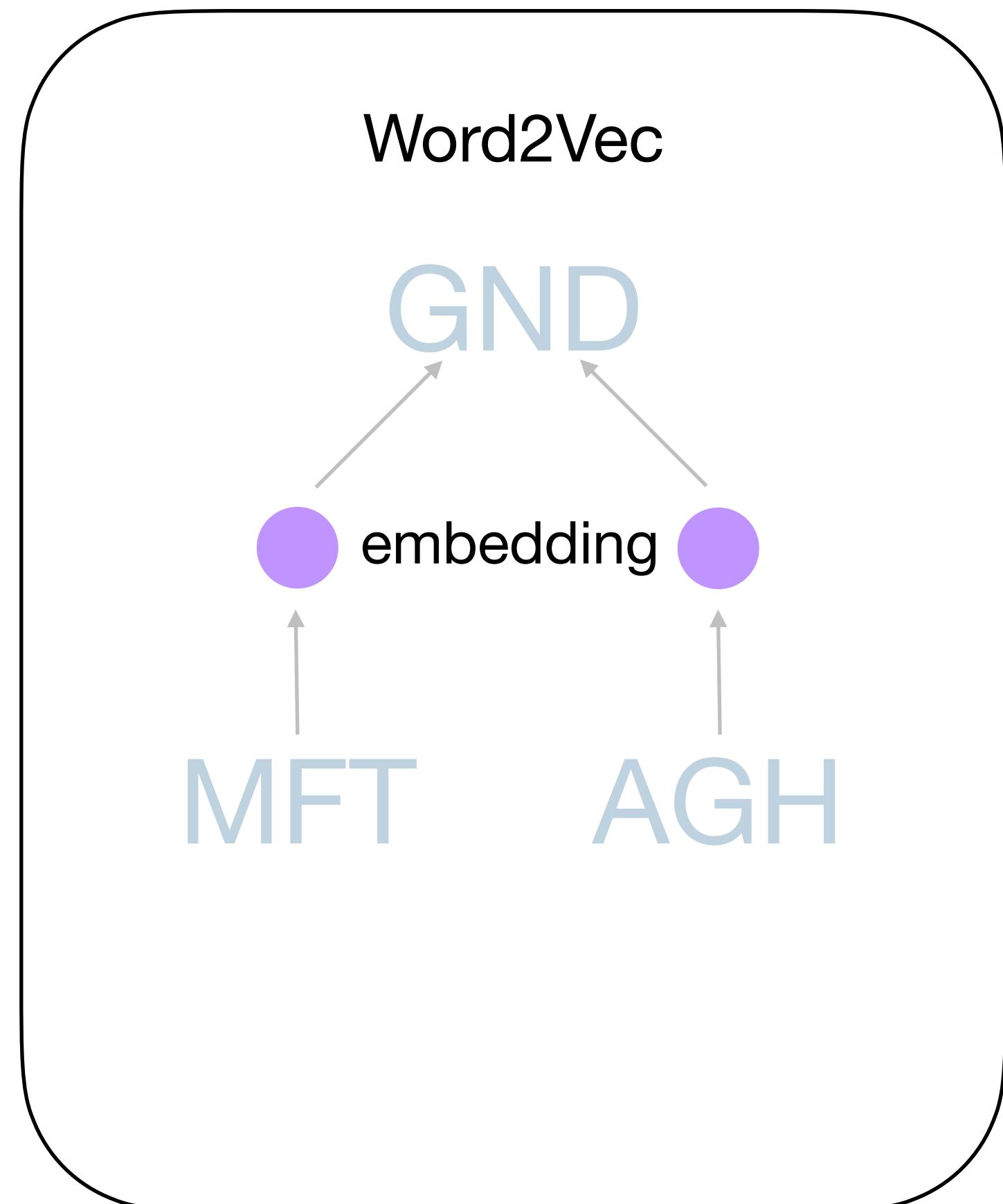
MFTGNDAGH

Many methods pretend proteins are language



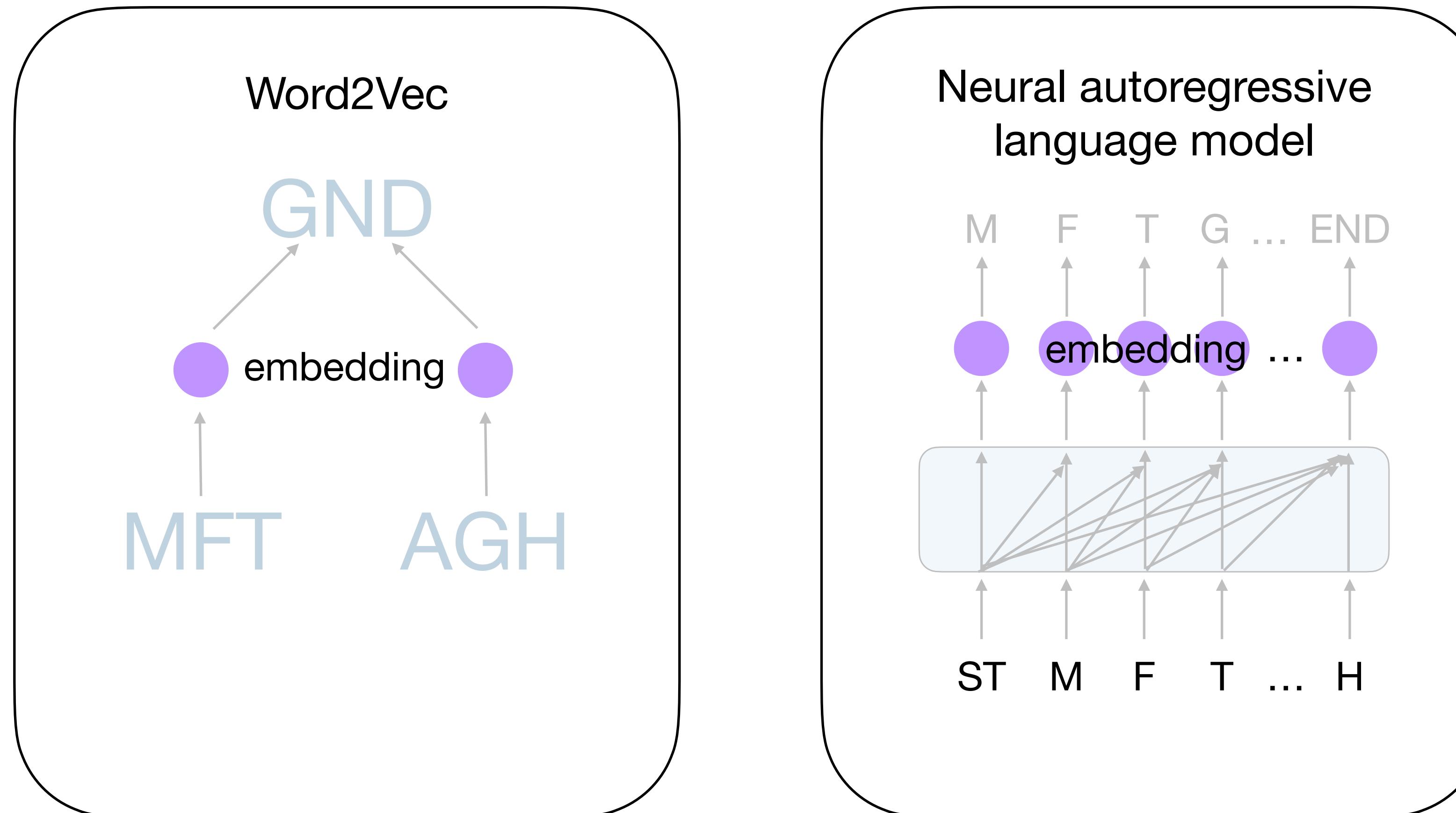
MFTGNDAGH

Many methods pretend proteins are language



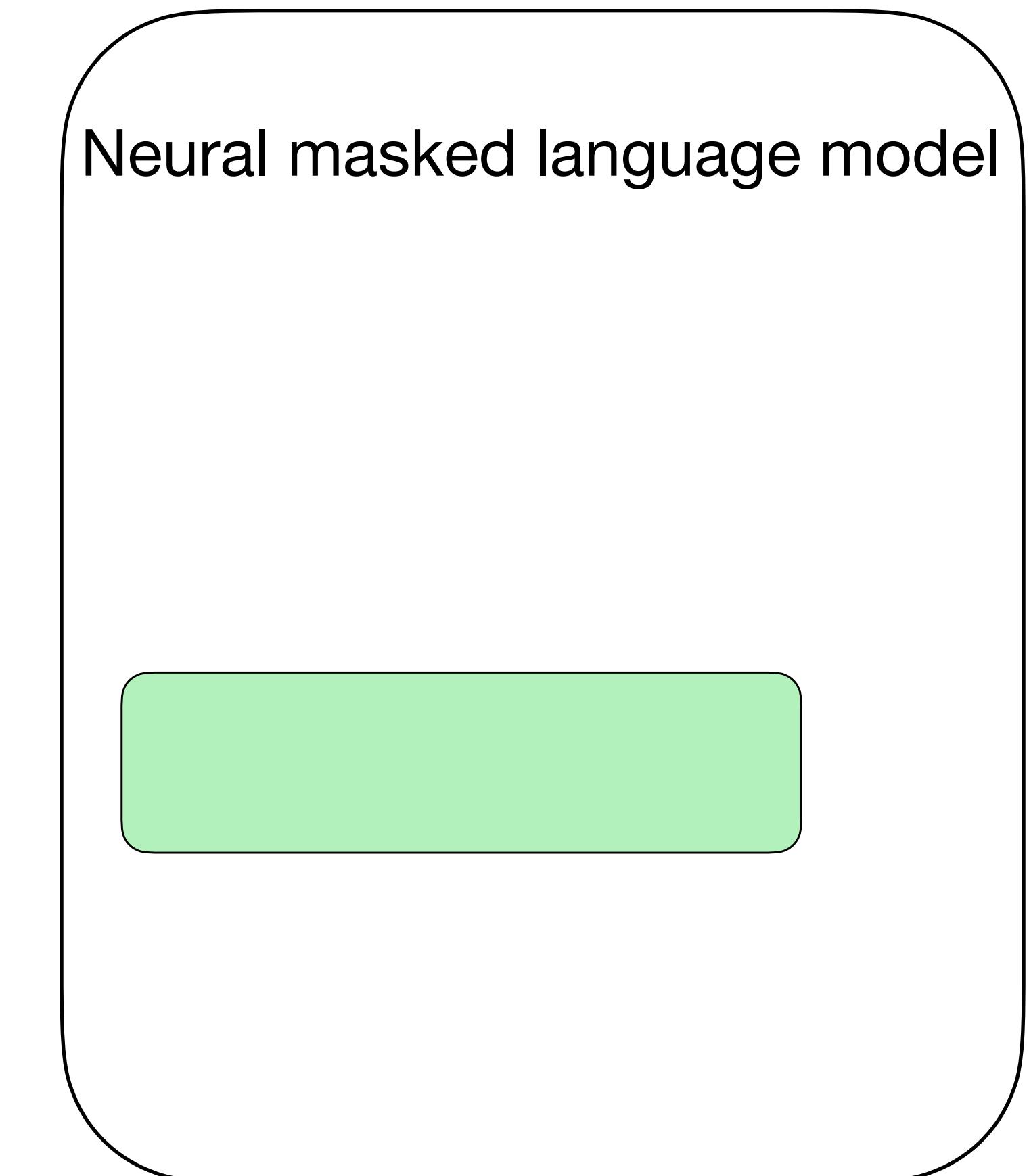
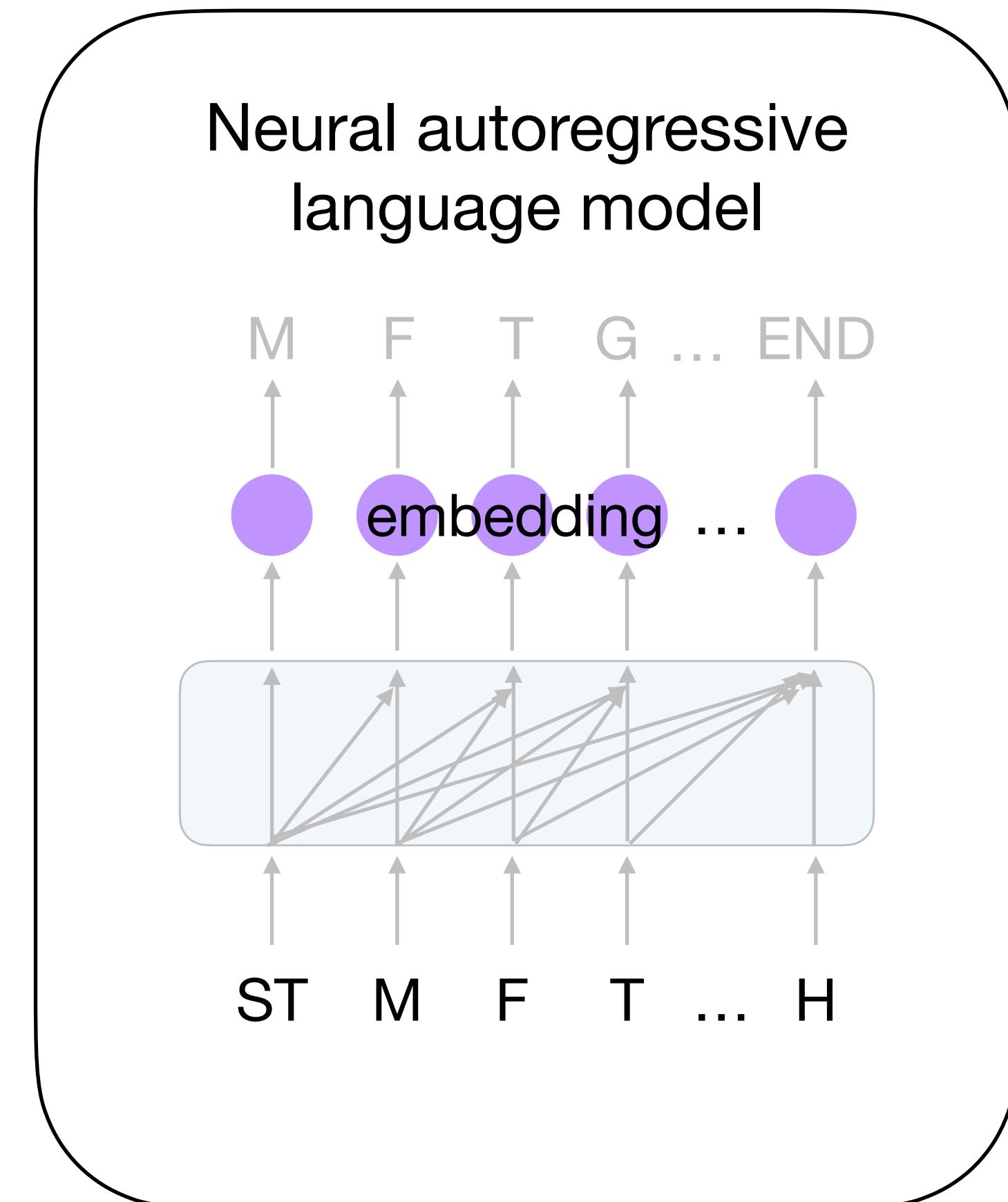
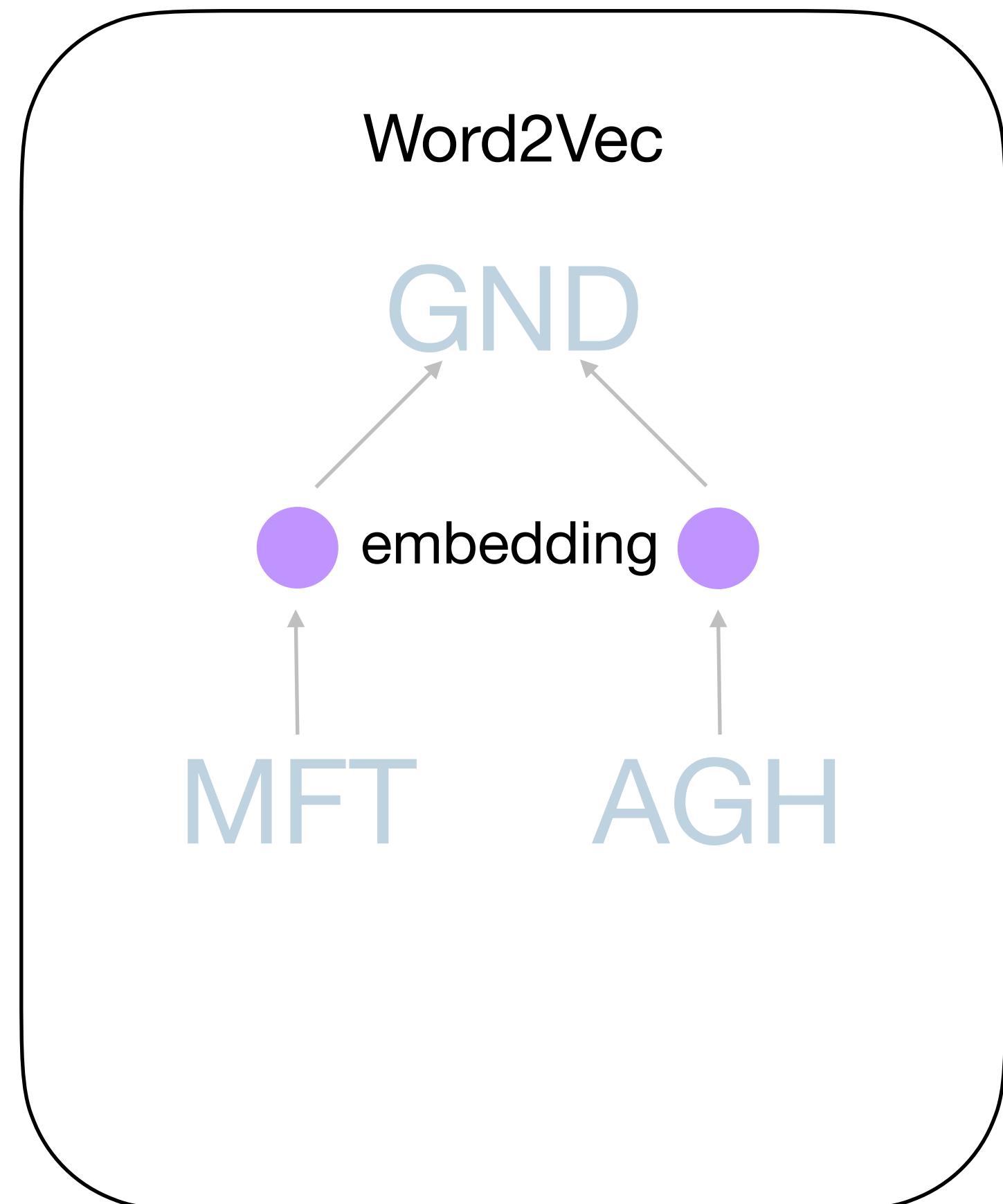
MFTGNDAGH

Many methods pretend proteins are language



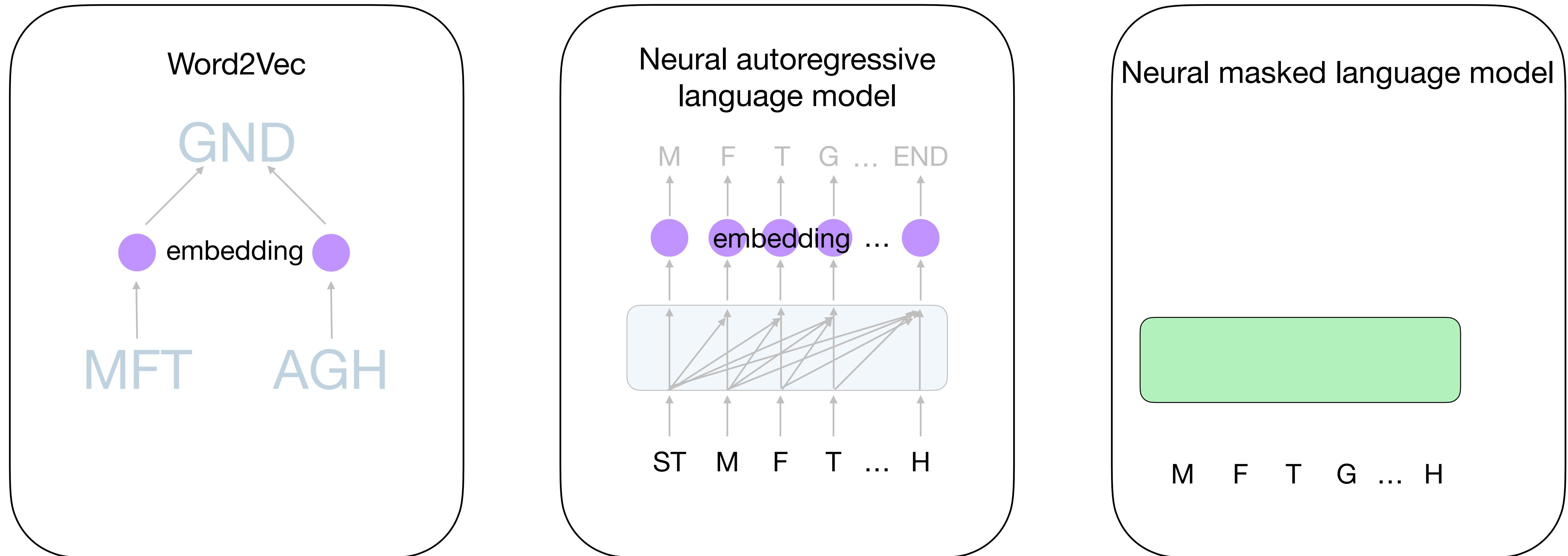
MFTGNDAGH

Many methods pretend proteins are language



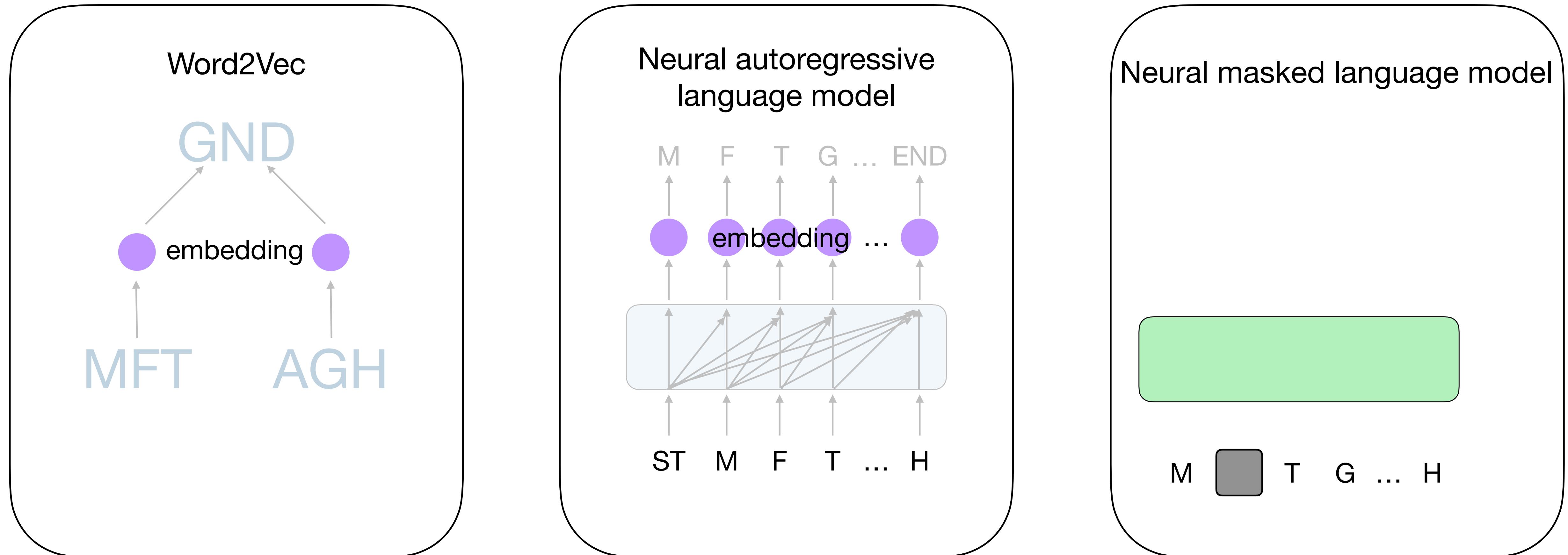
MFTGNDAGH

Many methods pretend proteins are language



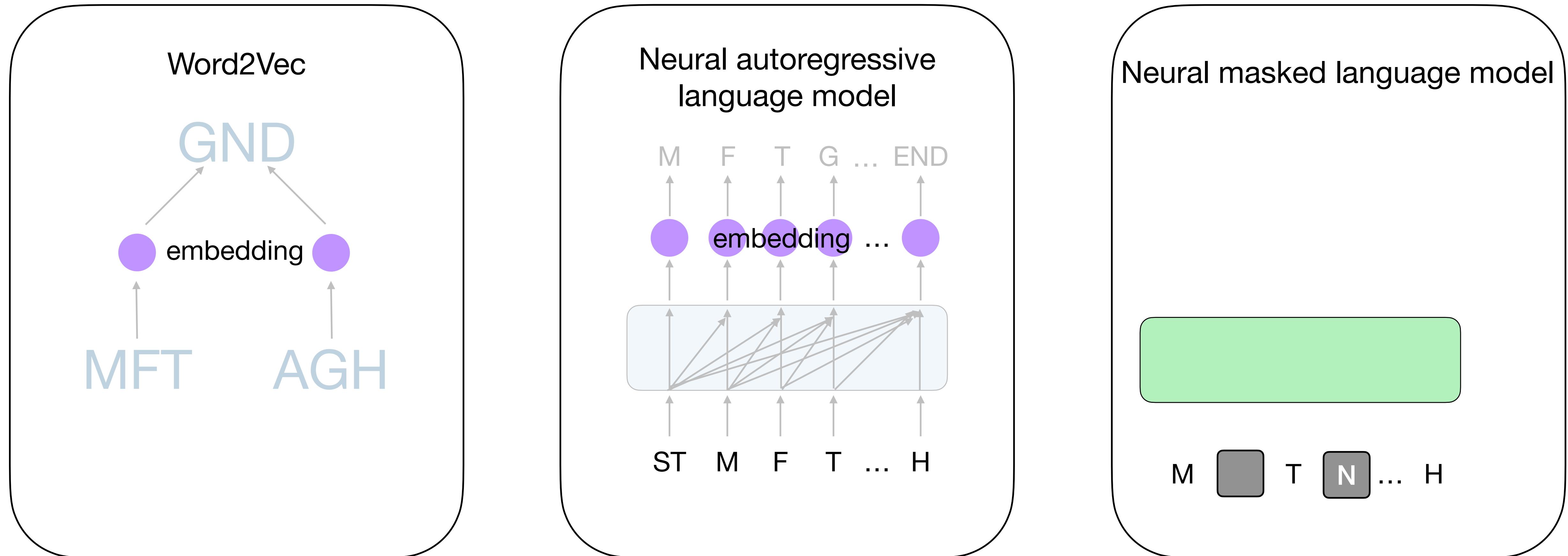
MFTGNDAGH

Many methods pretend proteins are language

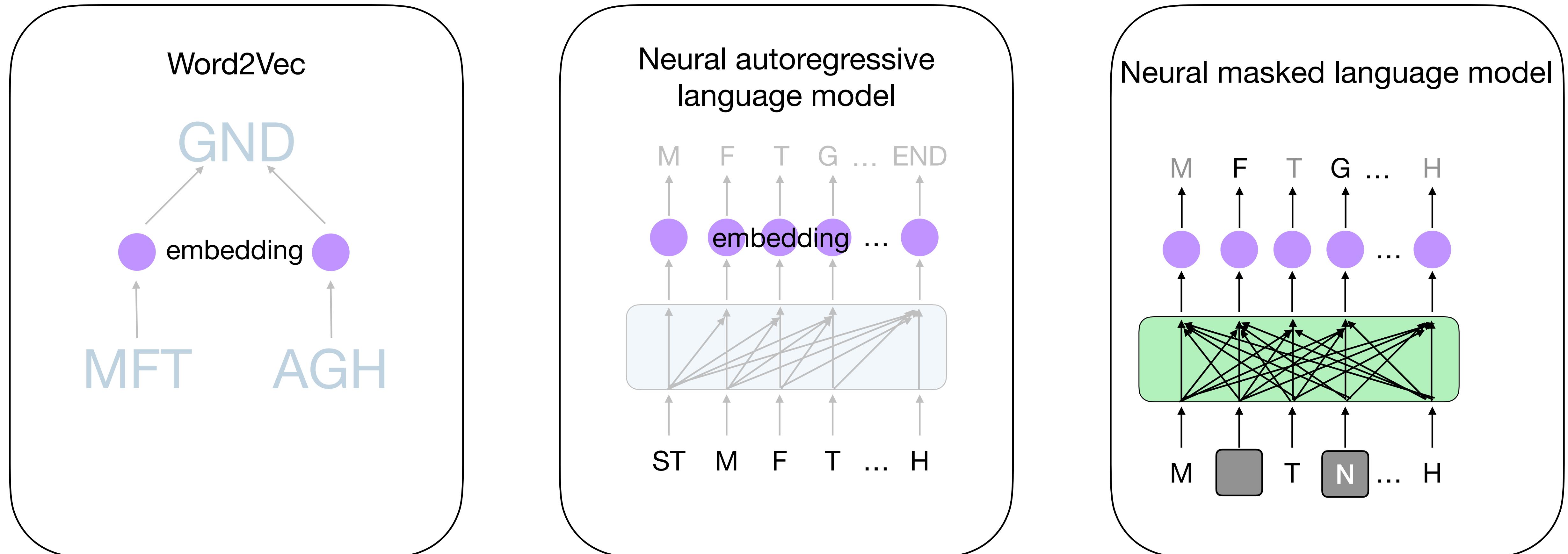


MFTGNDAGH

Many methods pretend proteins are language

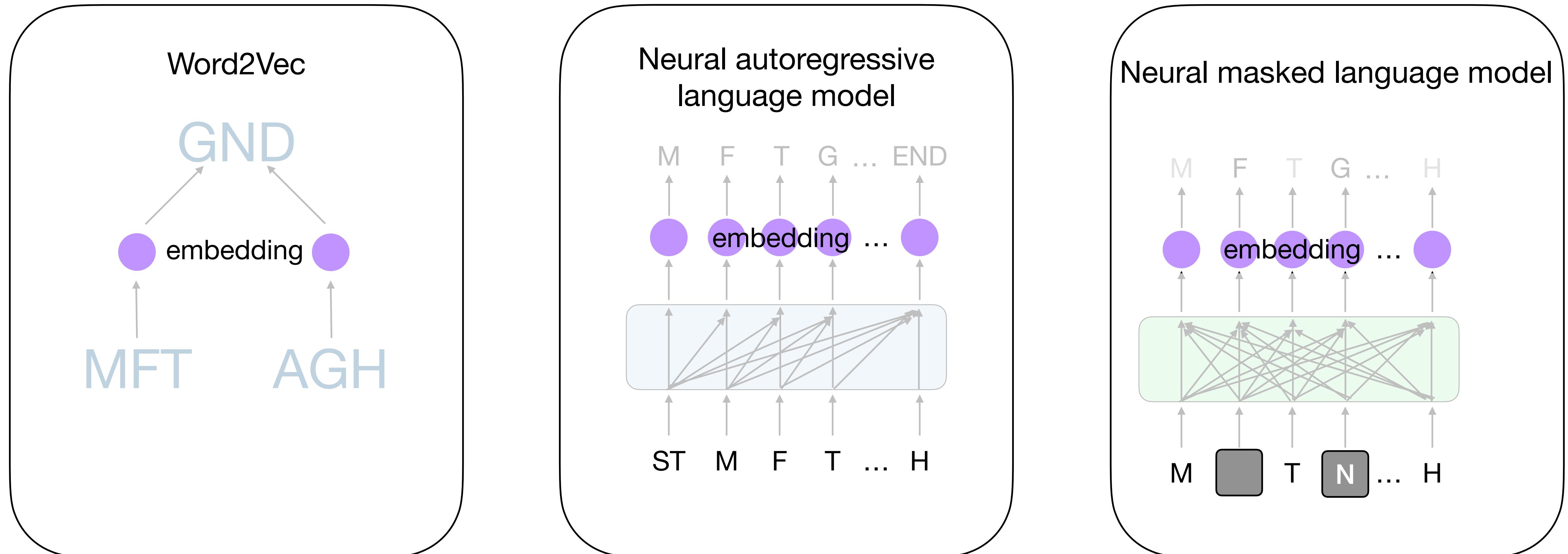


Many methods pretend proteins are language



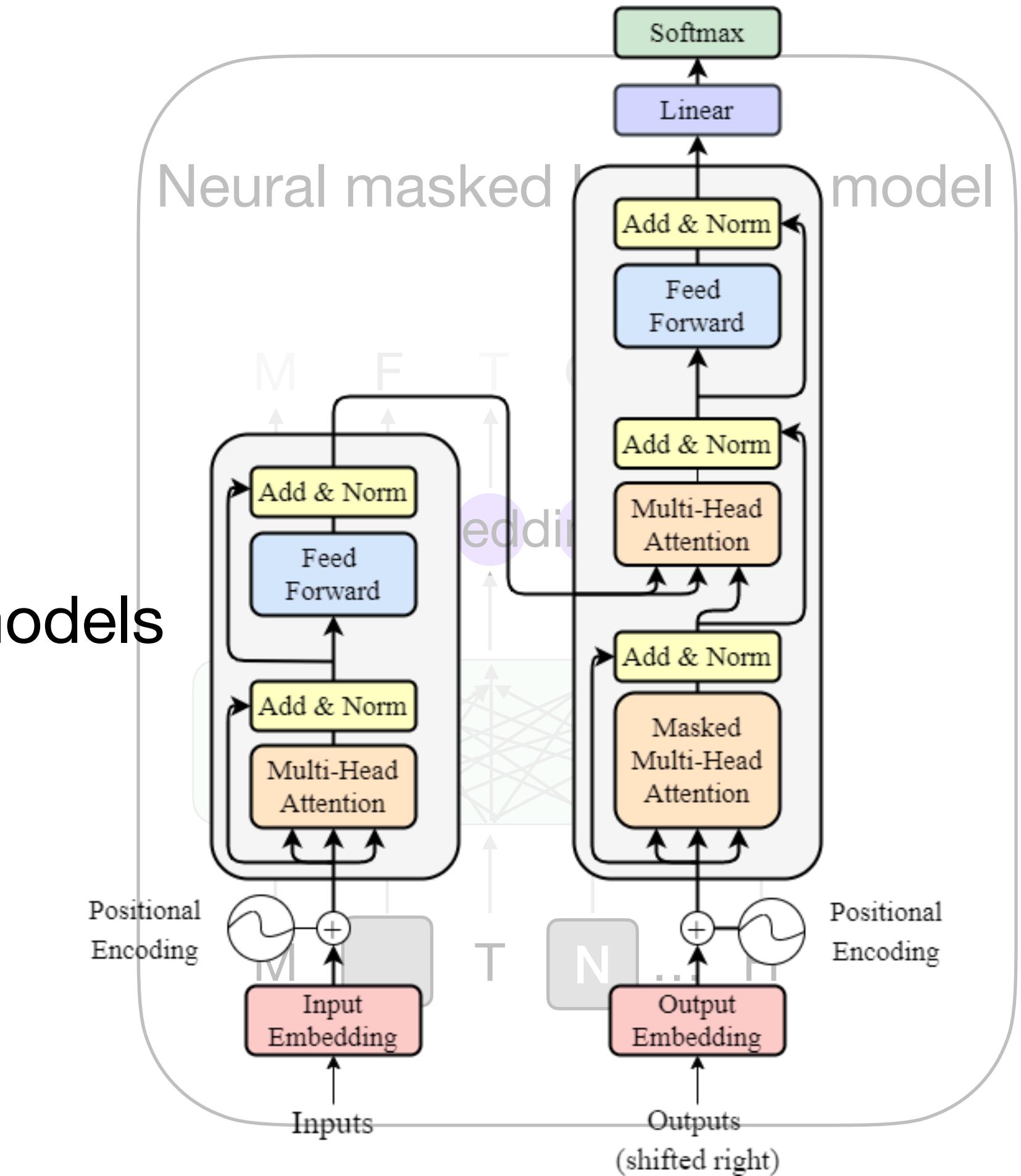
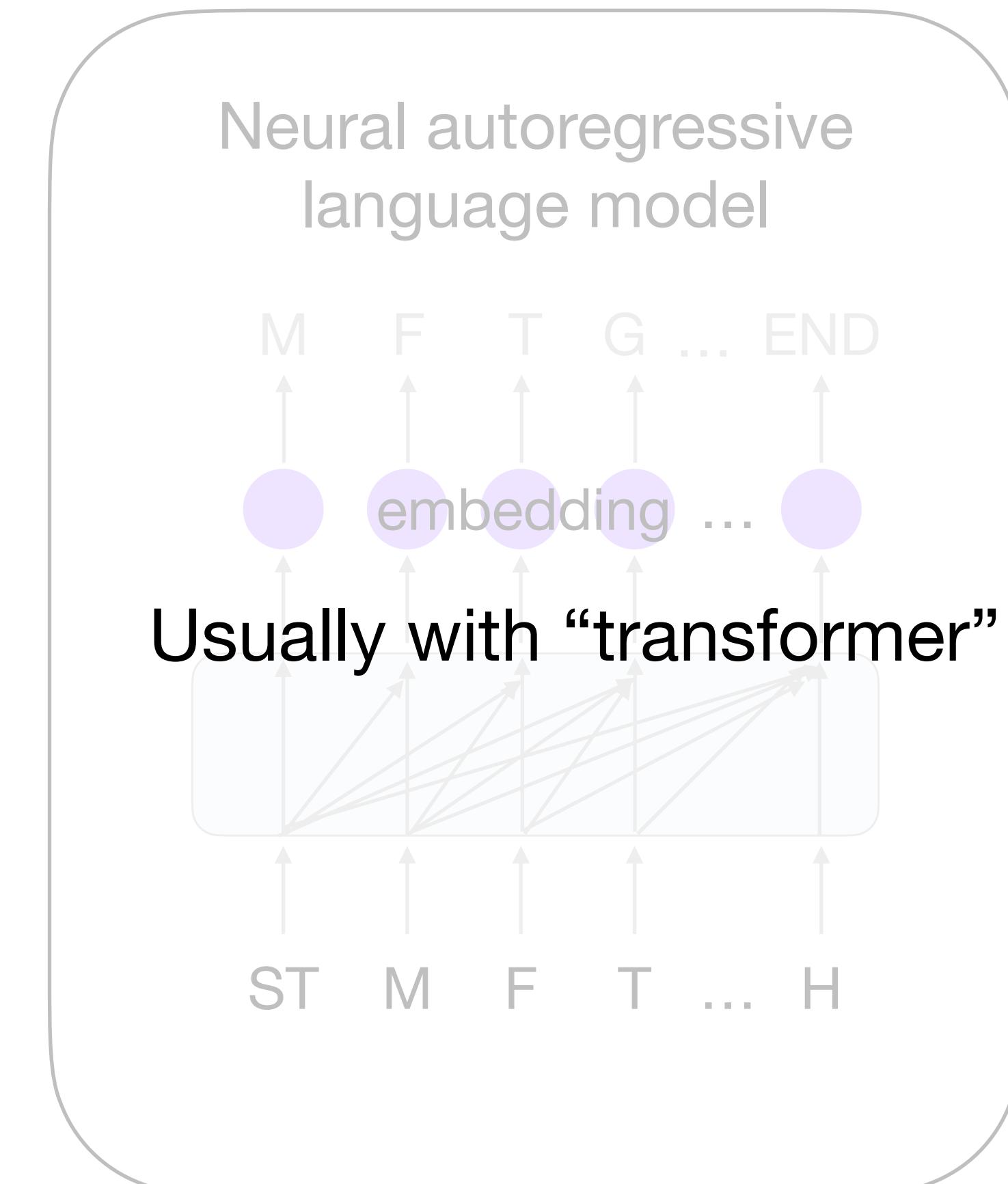
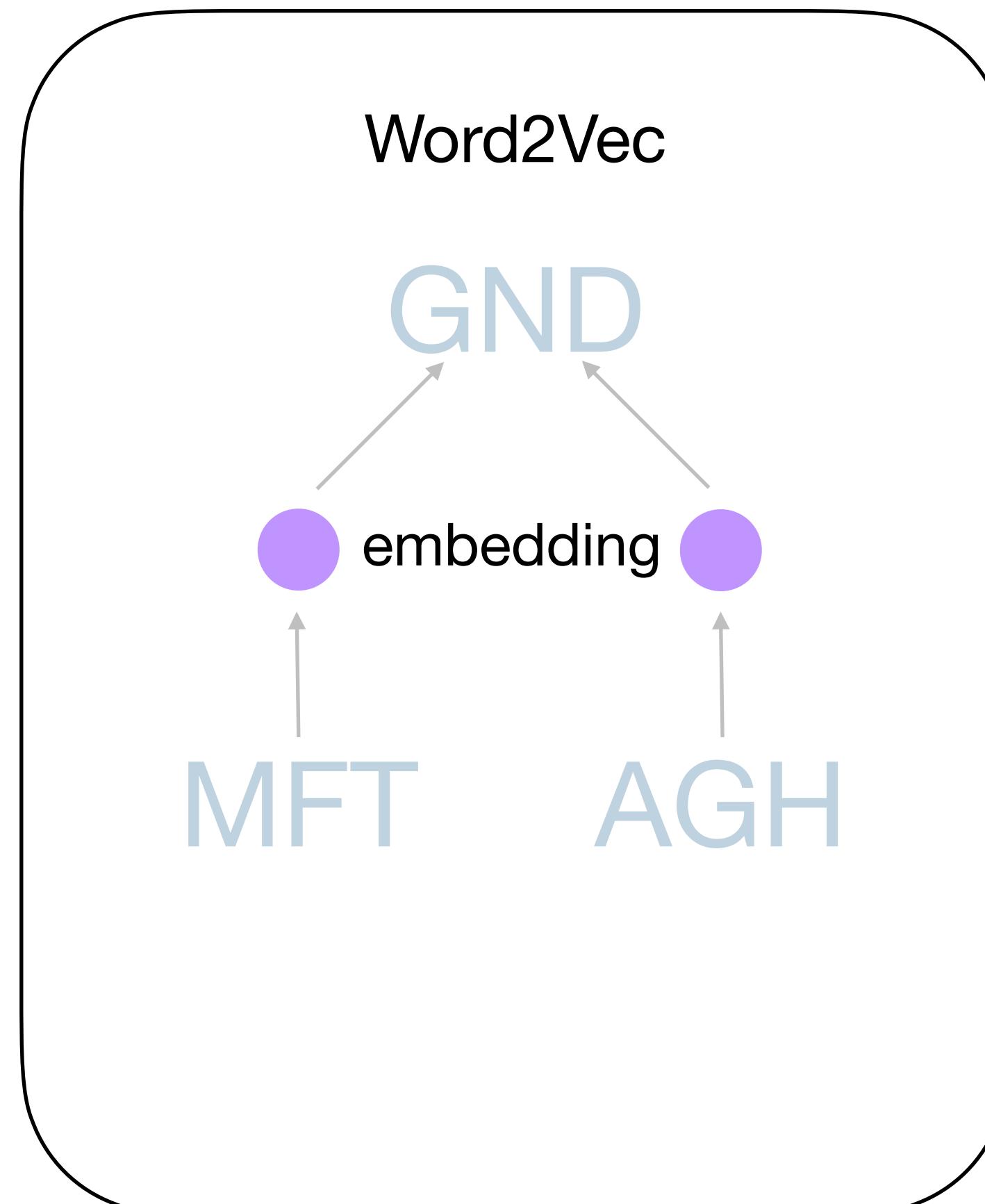
MFTGNDAGH

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

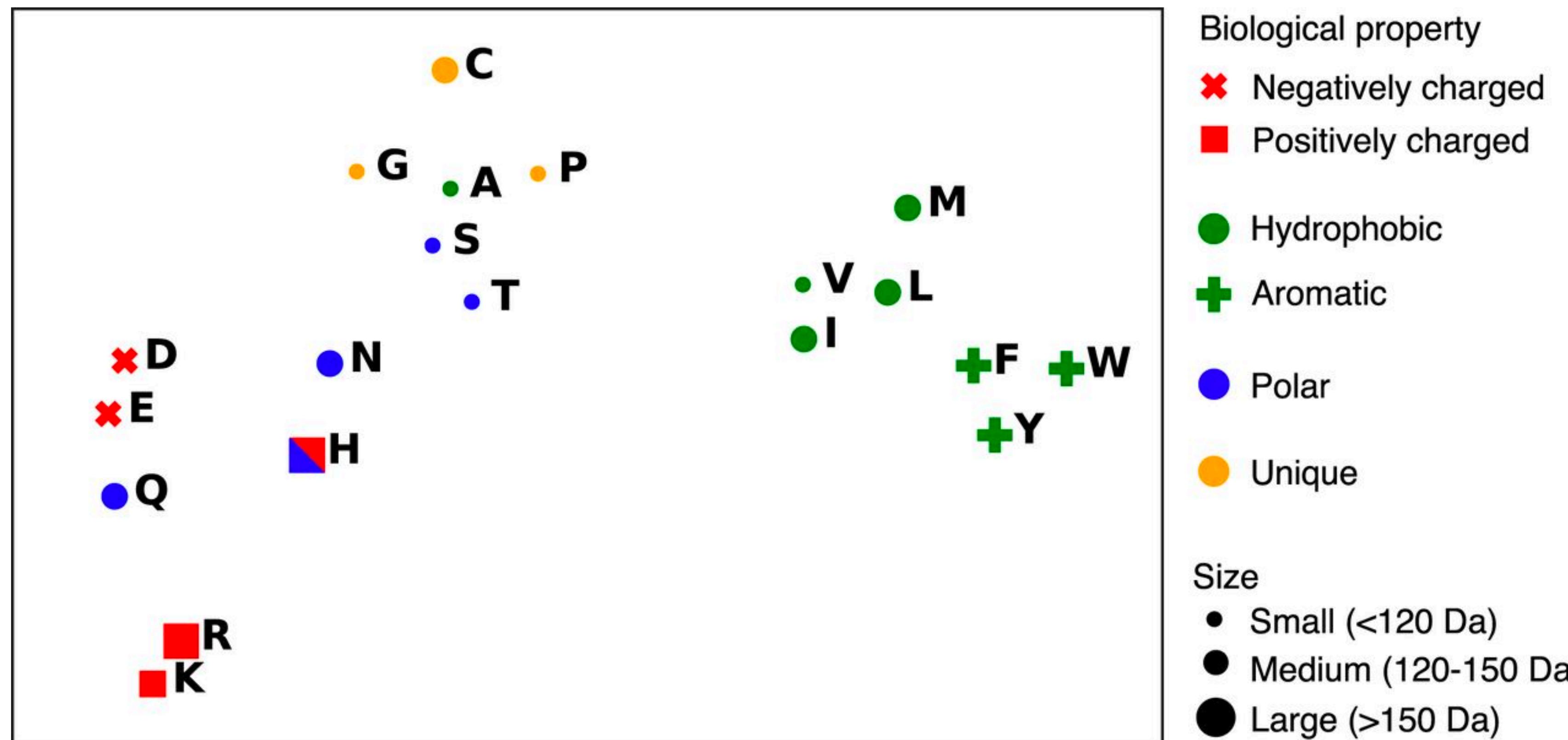
● Polar

● Unique

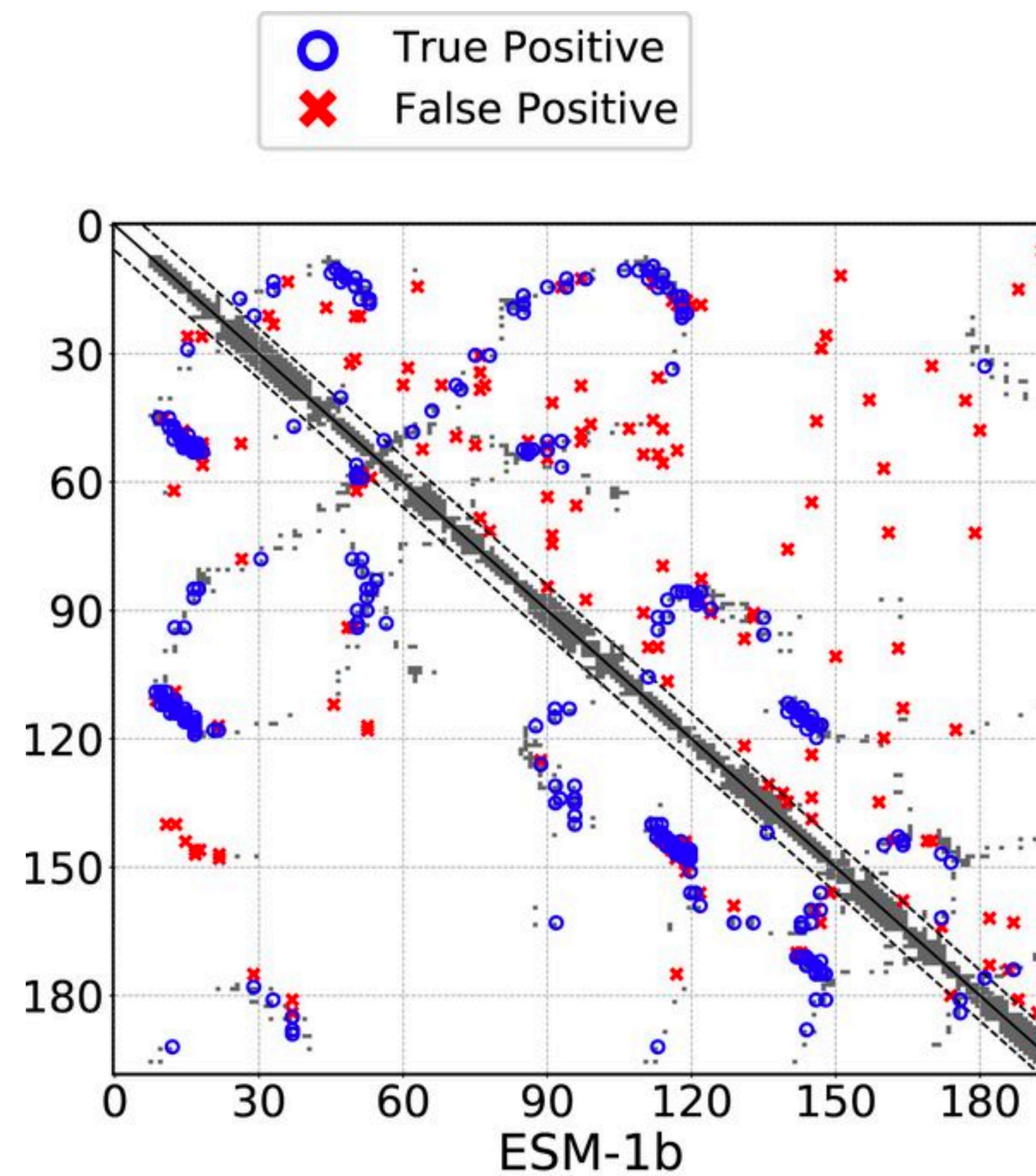
Size

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

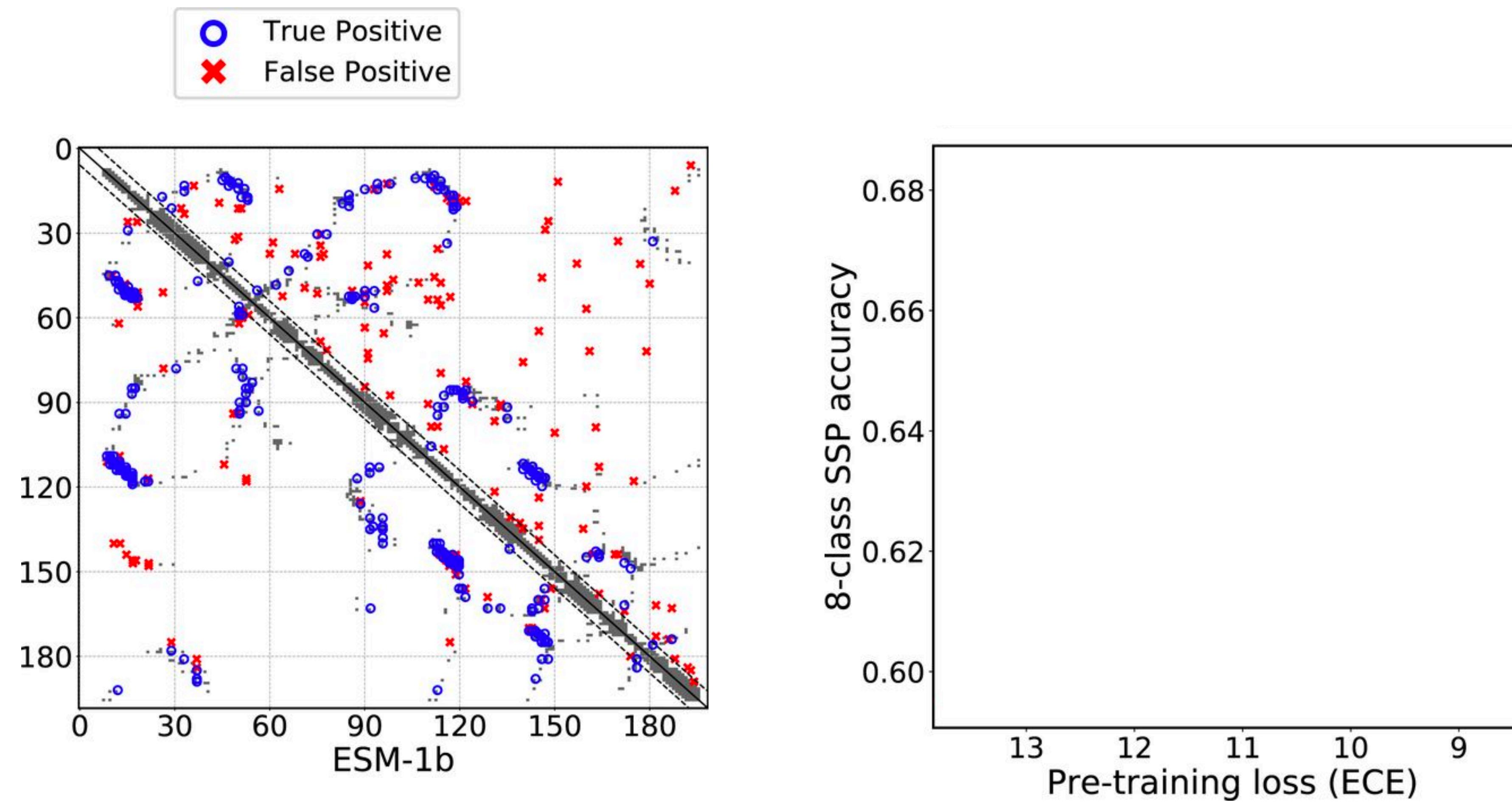
Pretrained transformers recapitulate biophysical properties



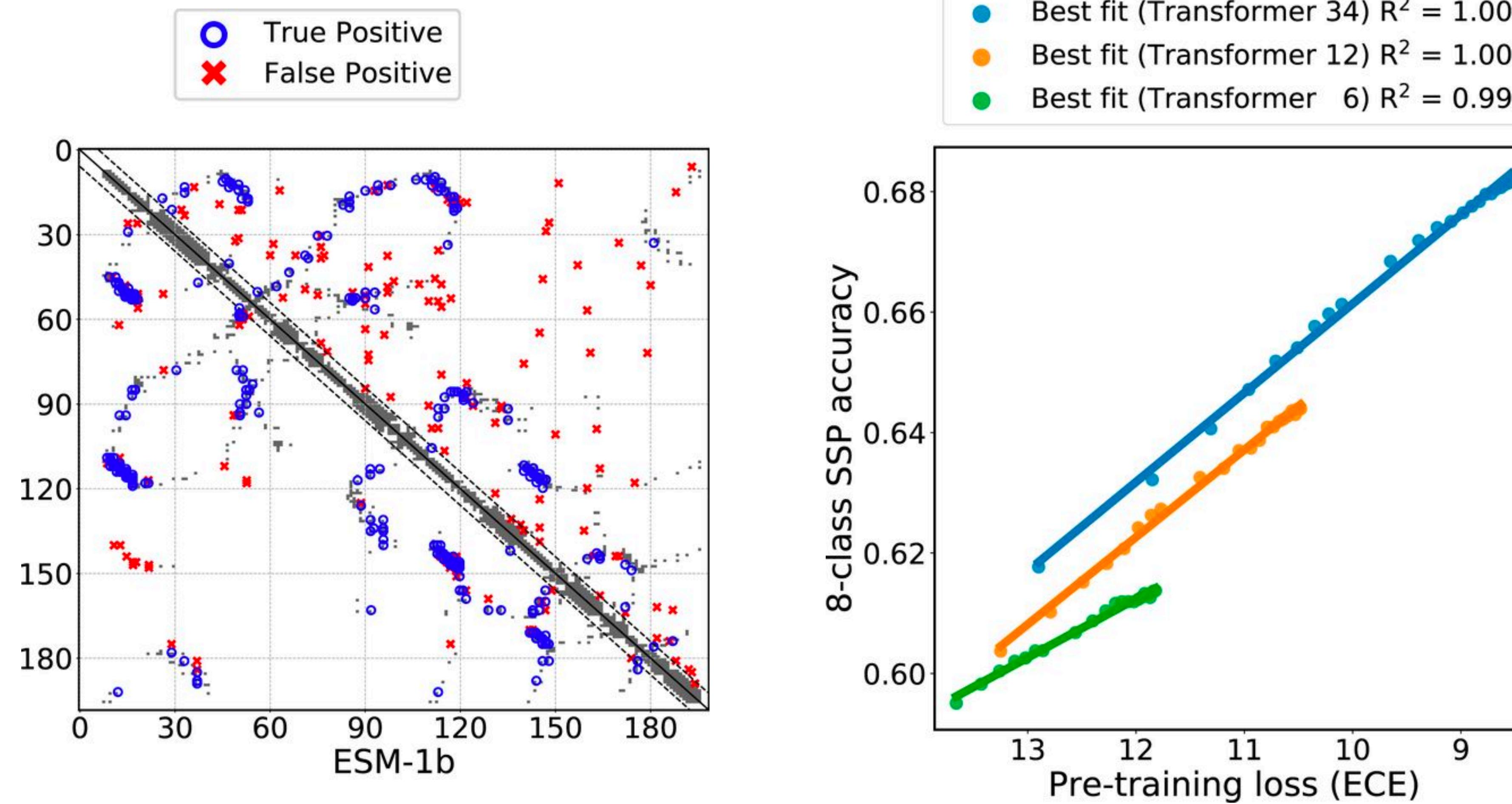
Pretrained transformers contain structural information



Pretrained transformers contain structural information



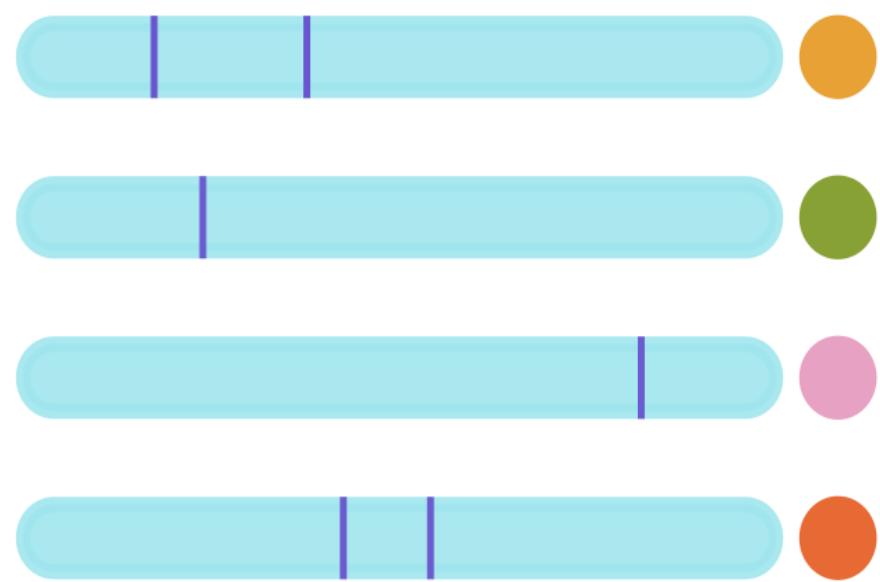
Pretrained transformers contain structural information



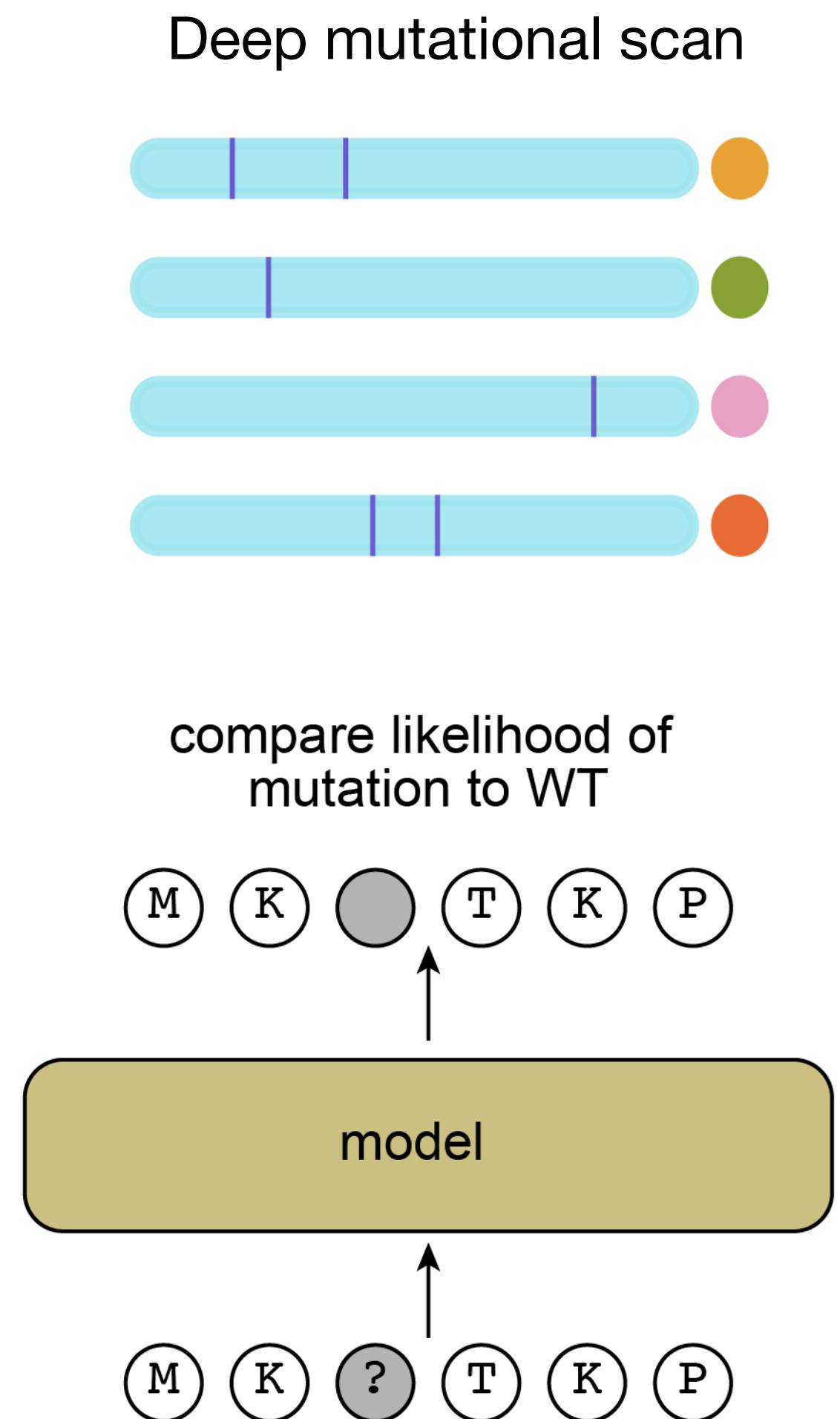
Pretrained transformers are zero-shot fitness predictors

Pretrained transformers are zero-shot fitness predictors

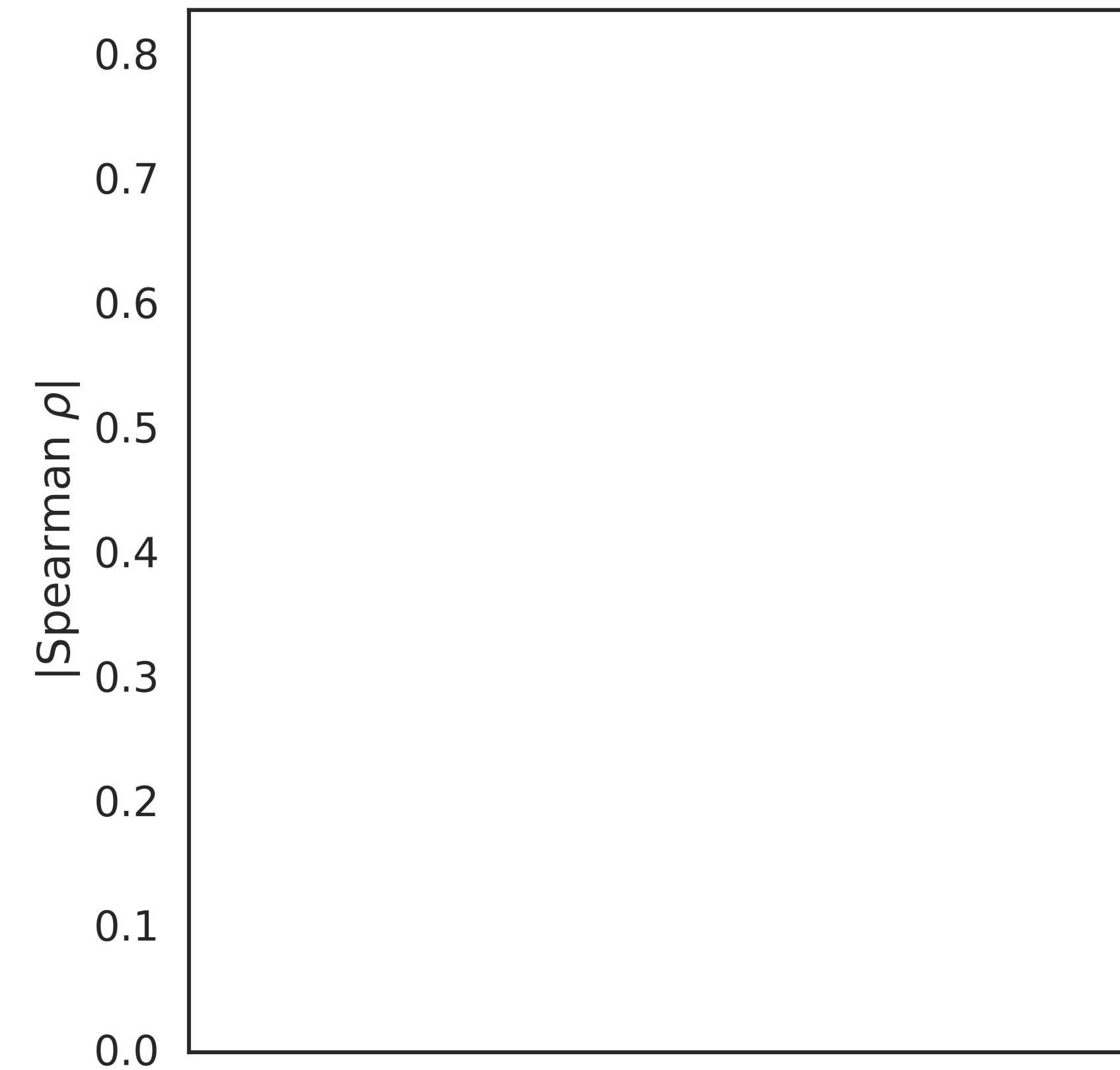
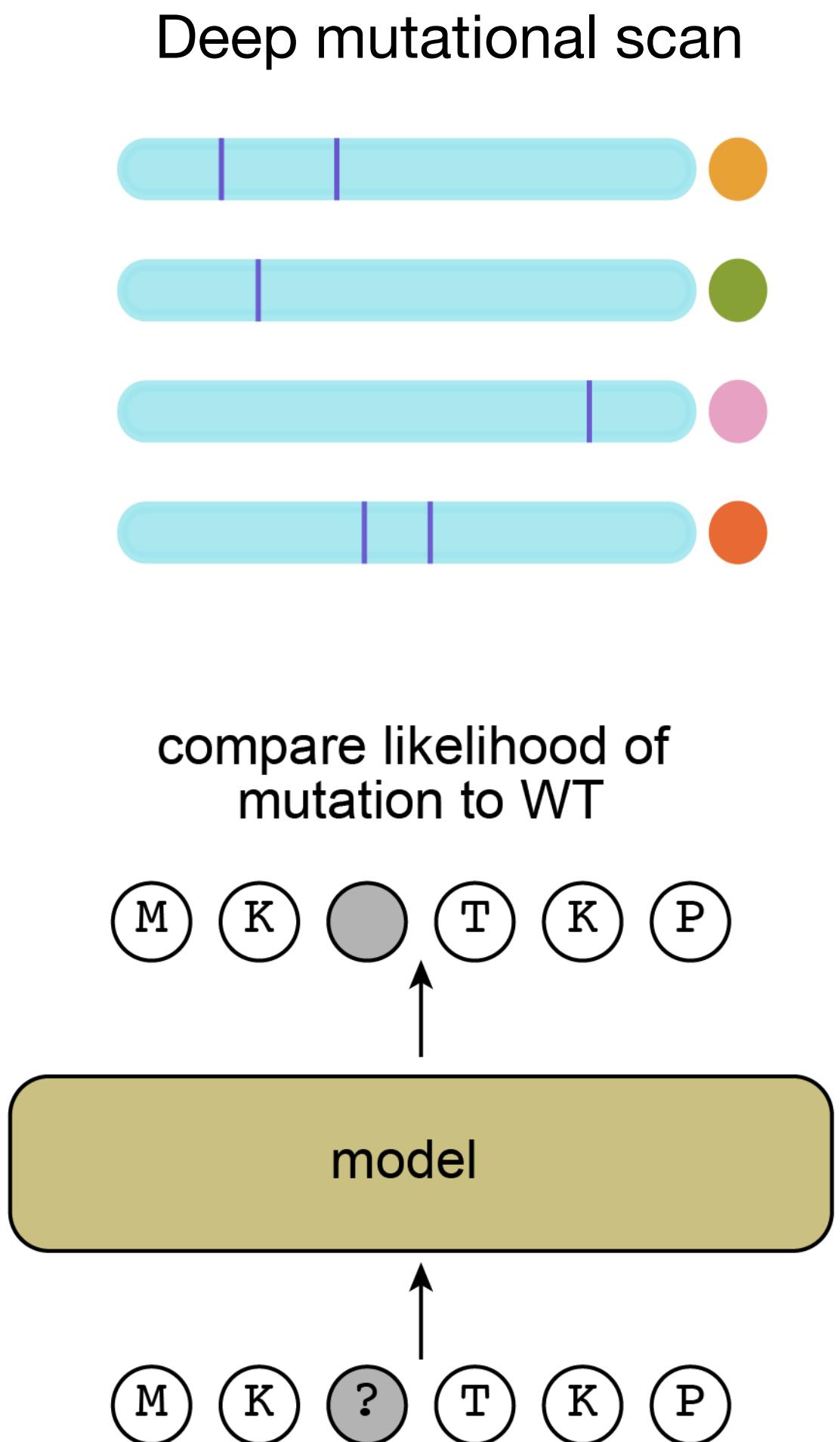
Deep mutational scan



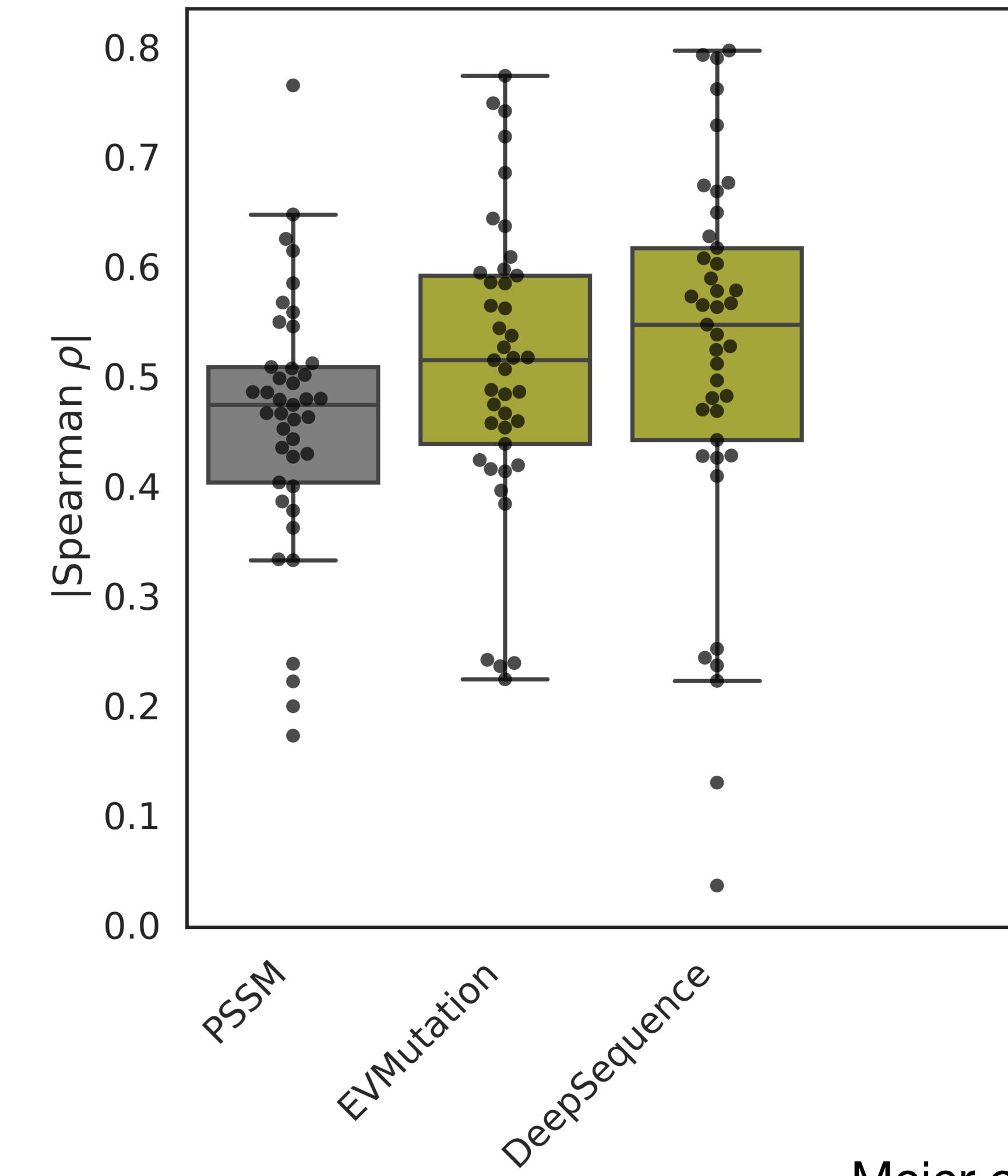
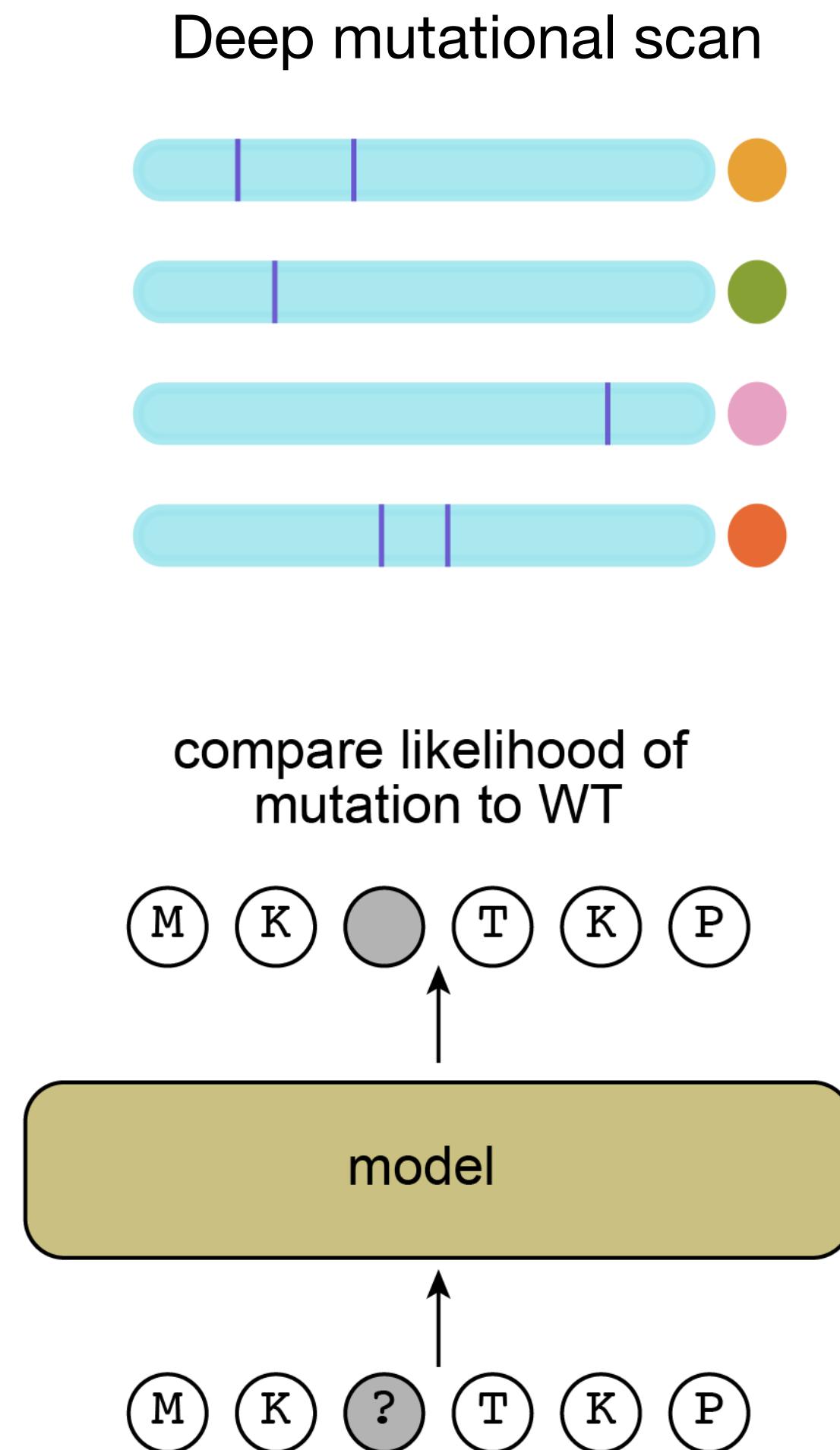
Pretrained transformers are zero-shot fitness predictors



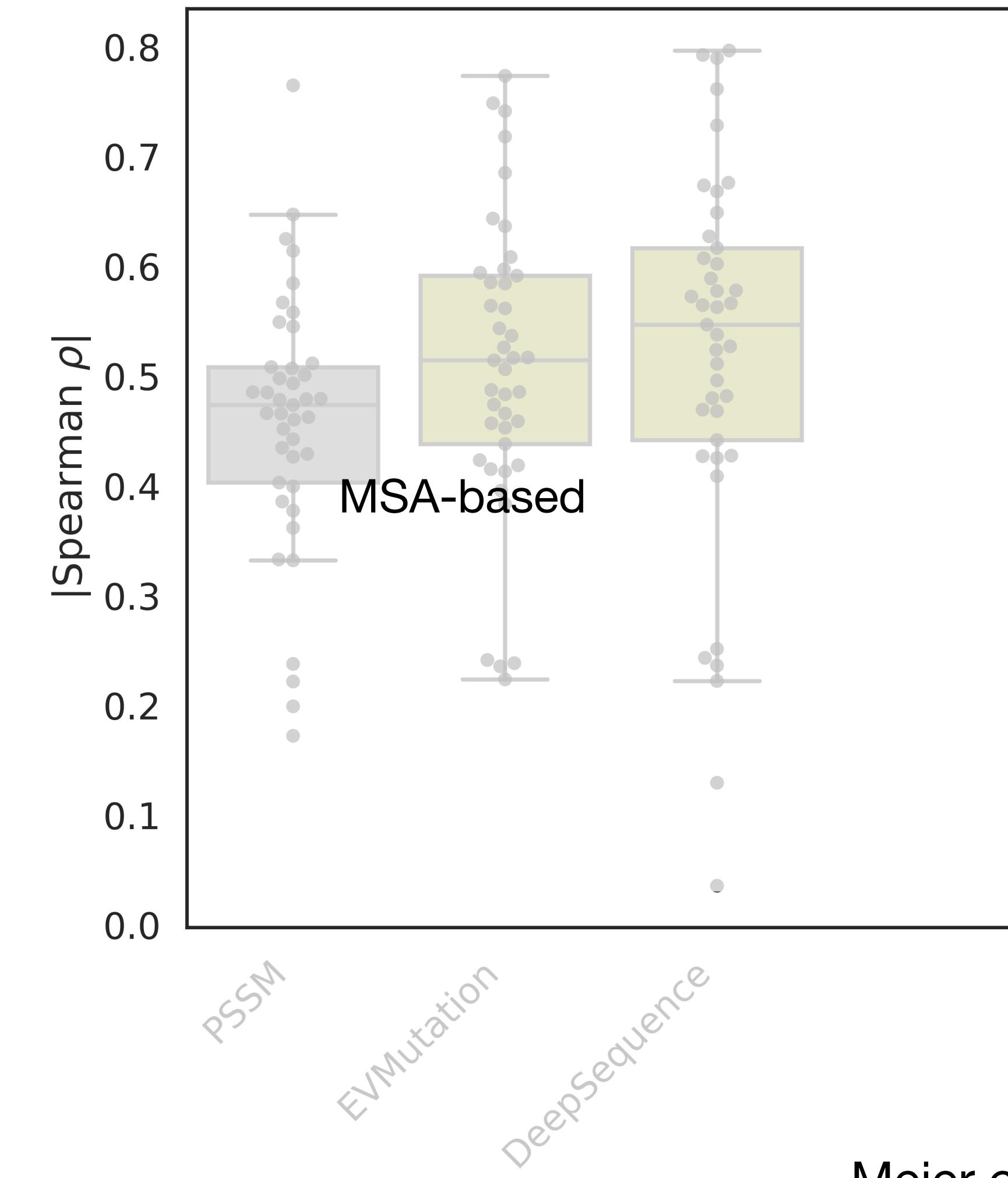
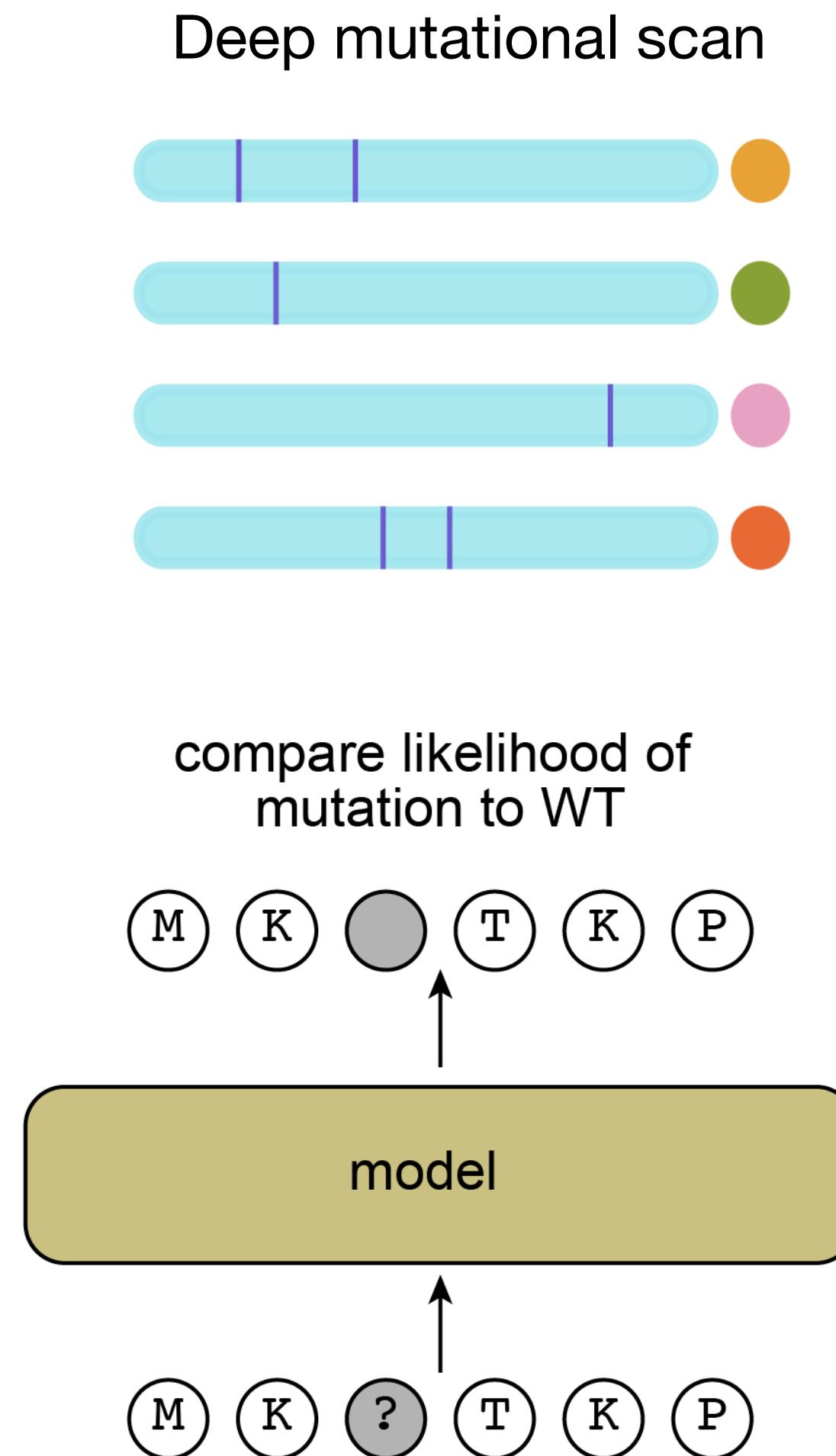
Pretrained transformers are zero-shot fitness predictors



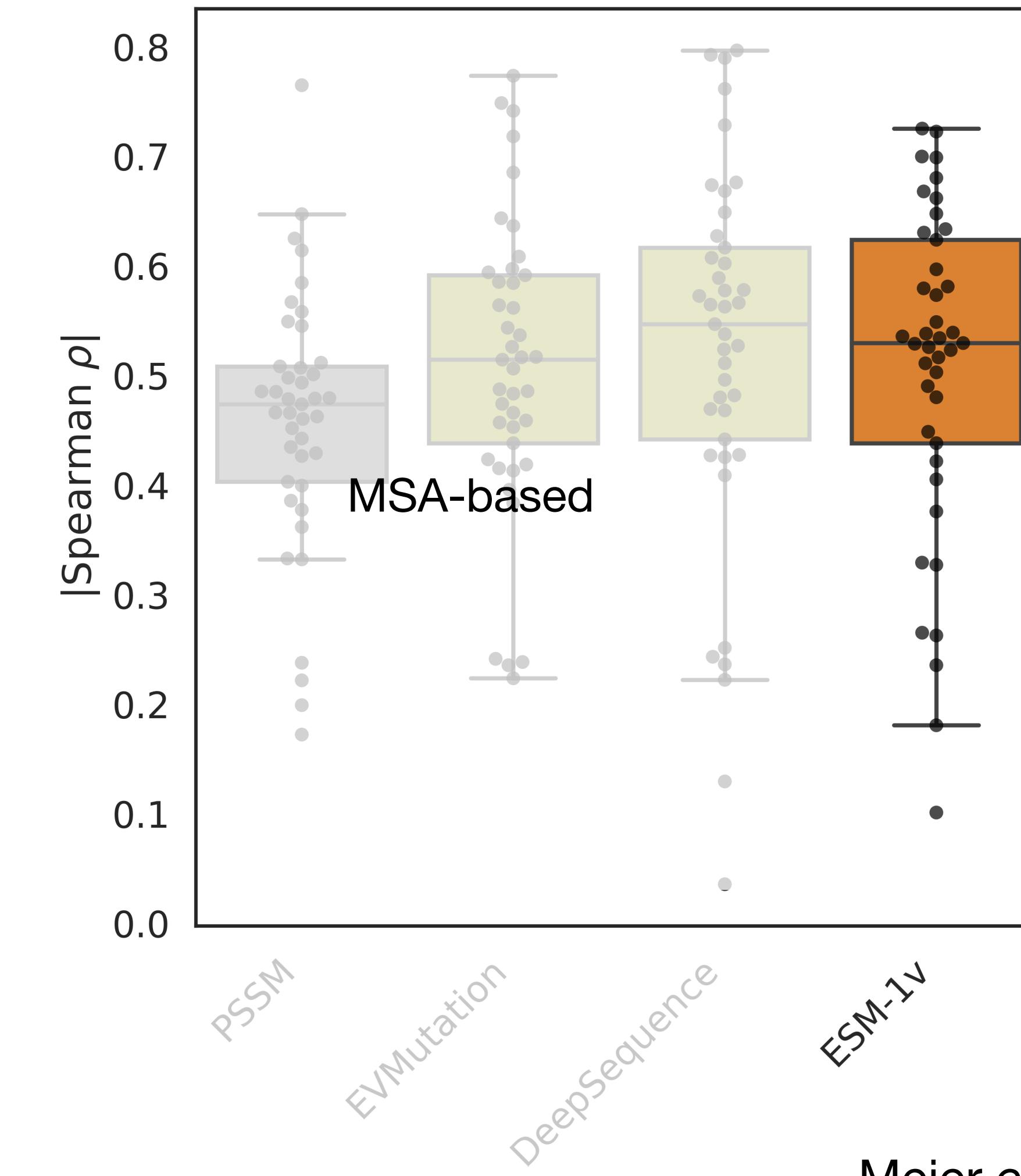
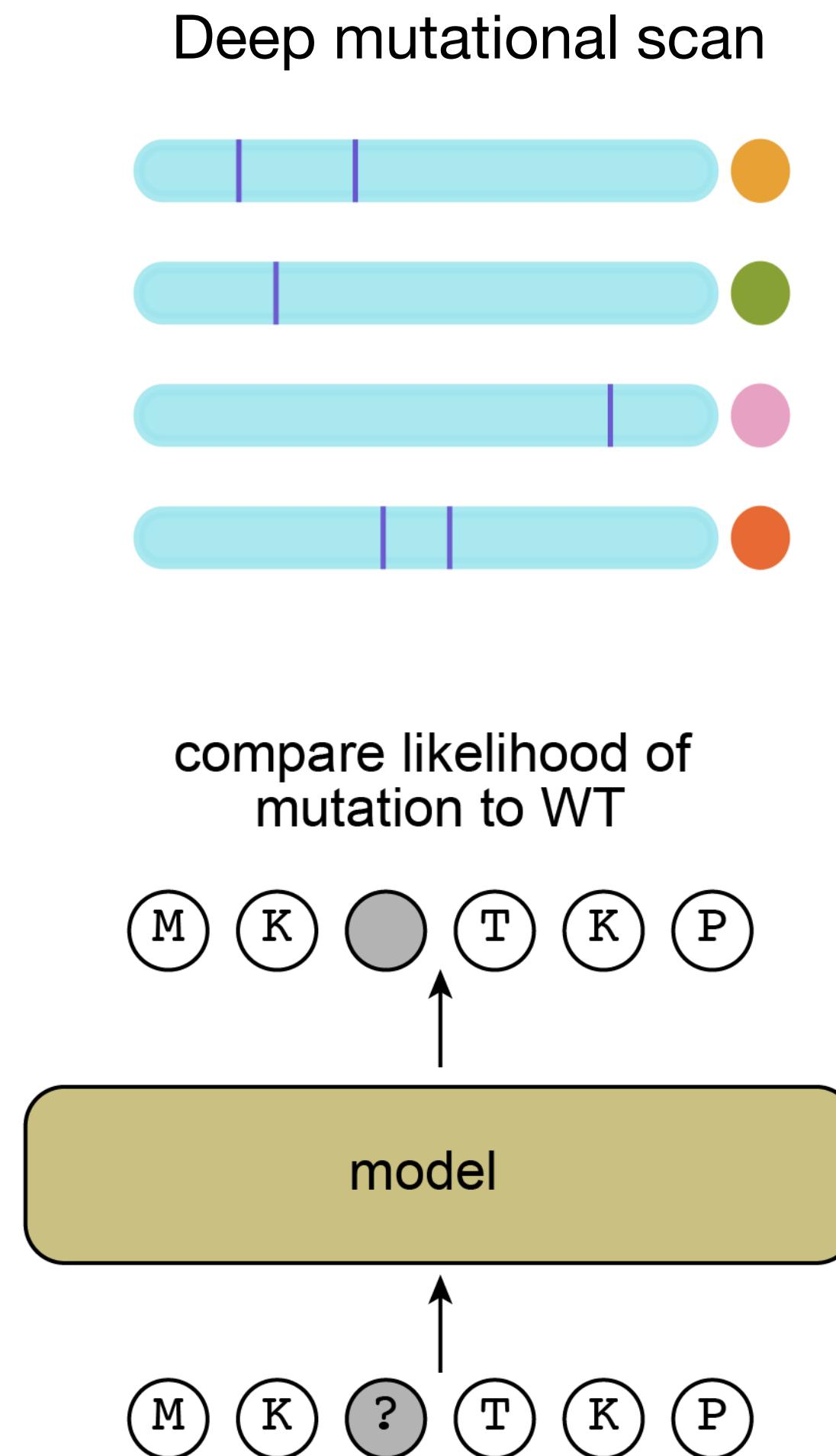
Pretrained transformers are zero-shot fitness predictors



Pretrained transformers are zero-shot fitness predictors

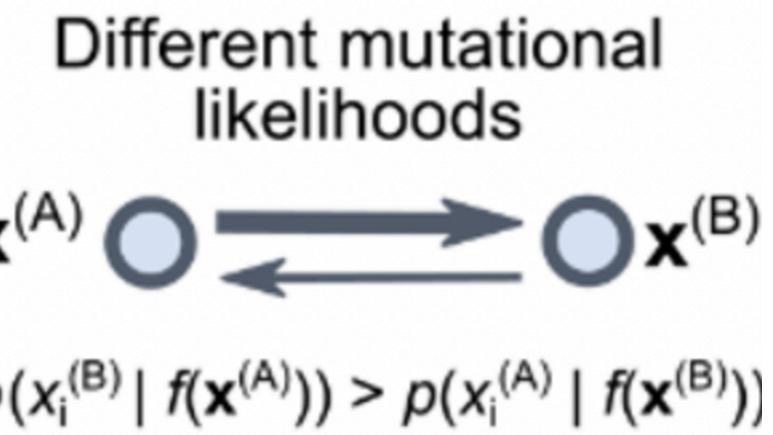


Pretrained transformers are zero-shot fitness predictors

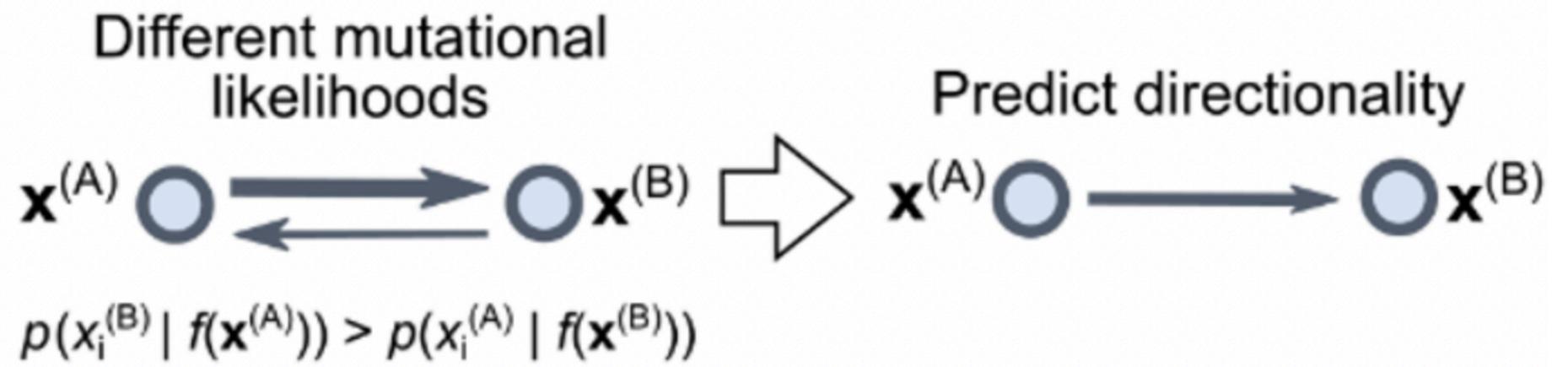


Pretrained transformers reconstruct evolutionary trajectories

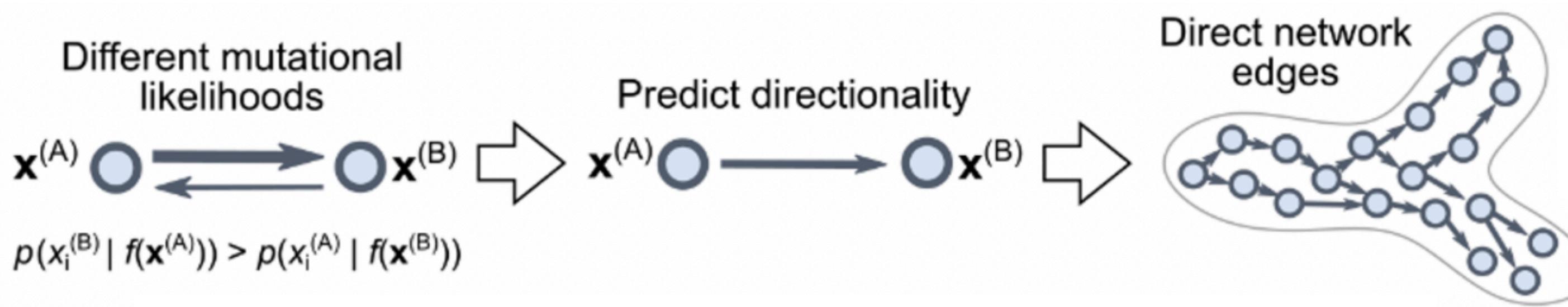
Pretrained transformers reconstruct evolutionary trajectories



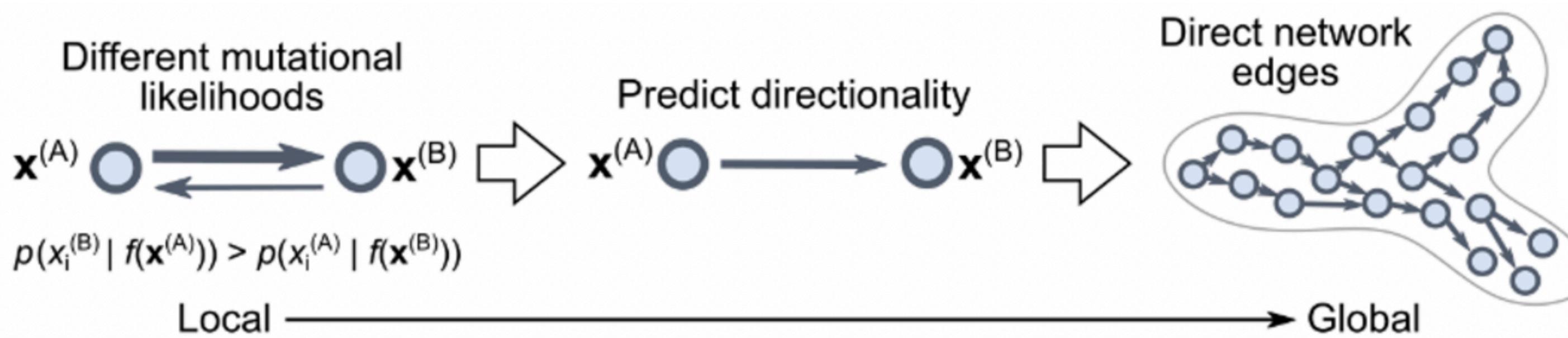
Pretrained transformers reconstruct evolutionary trajectories



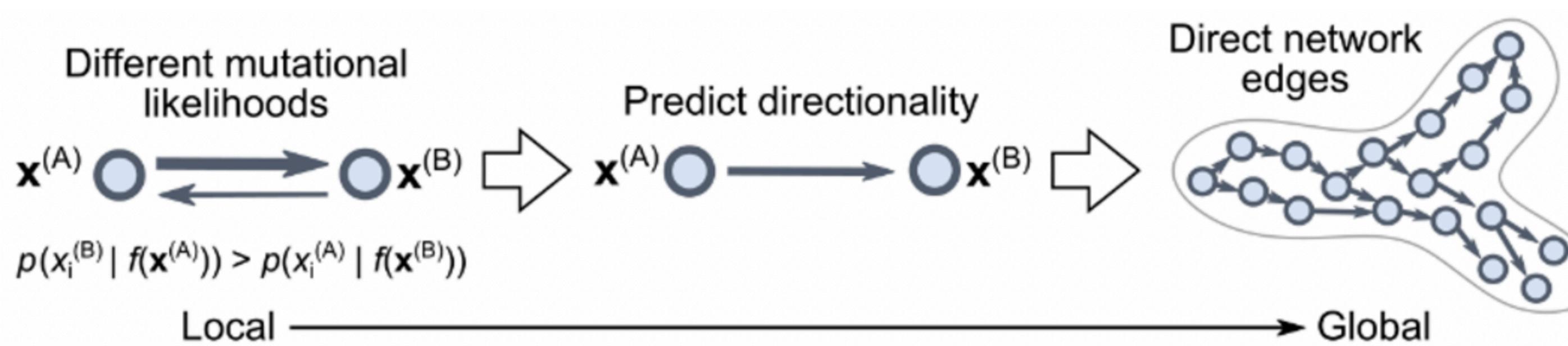
Pretrained transformers reconstruct evolutionary trajectories



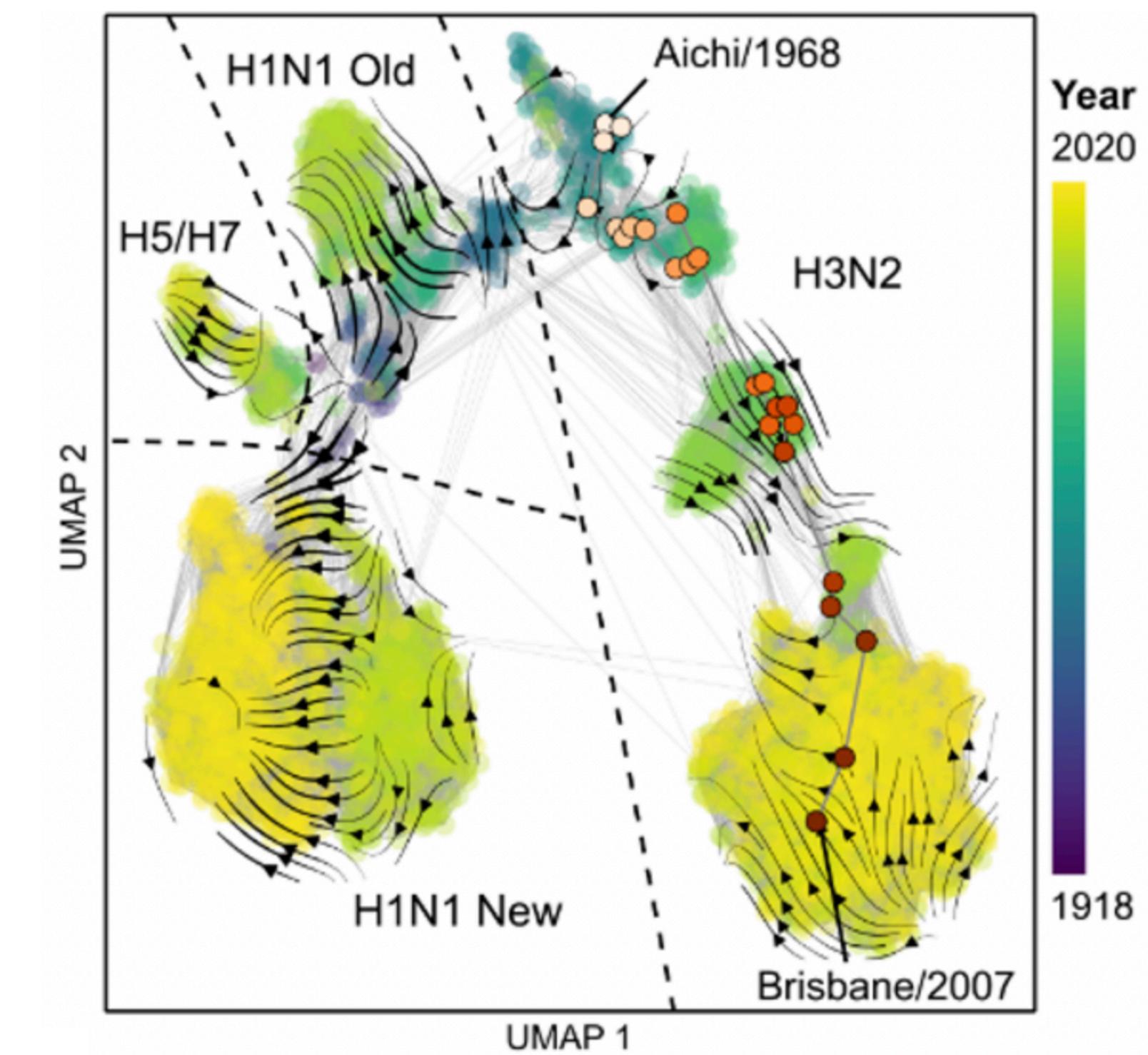
Pretrained transformers reconstruct evolutionary trajectories



Pretrained transformers reconstruct evolutionary trajectories

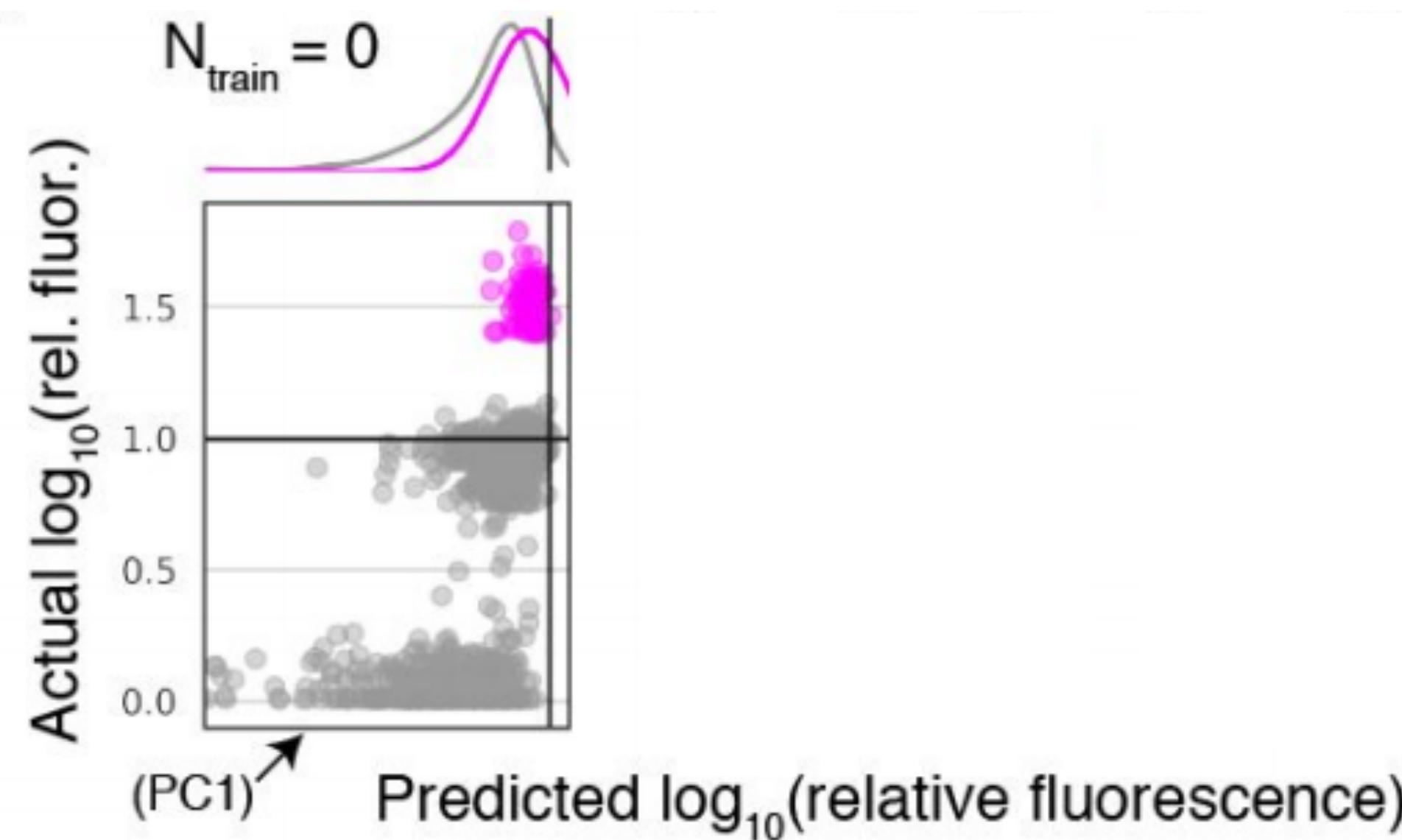


Influenza A nucleoprotein

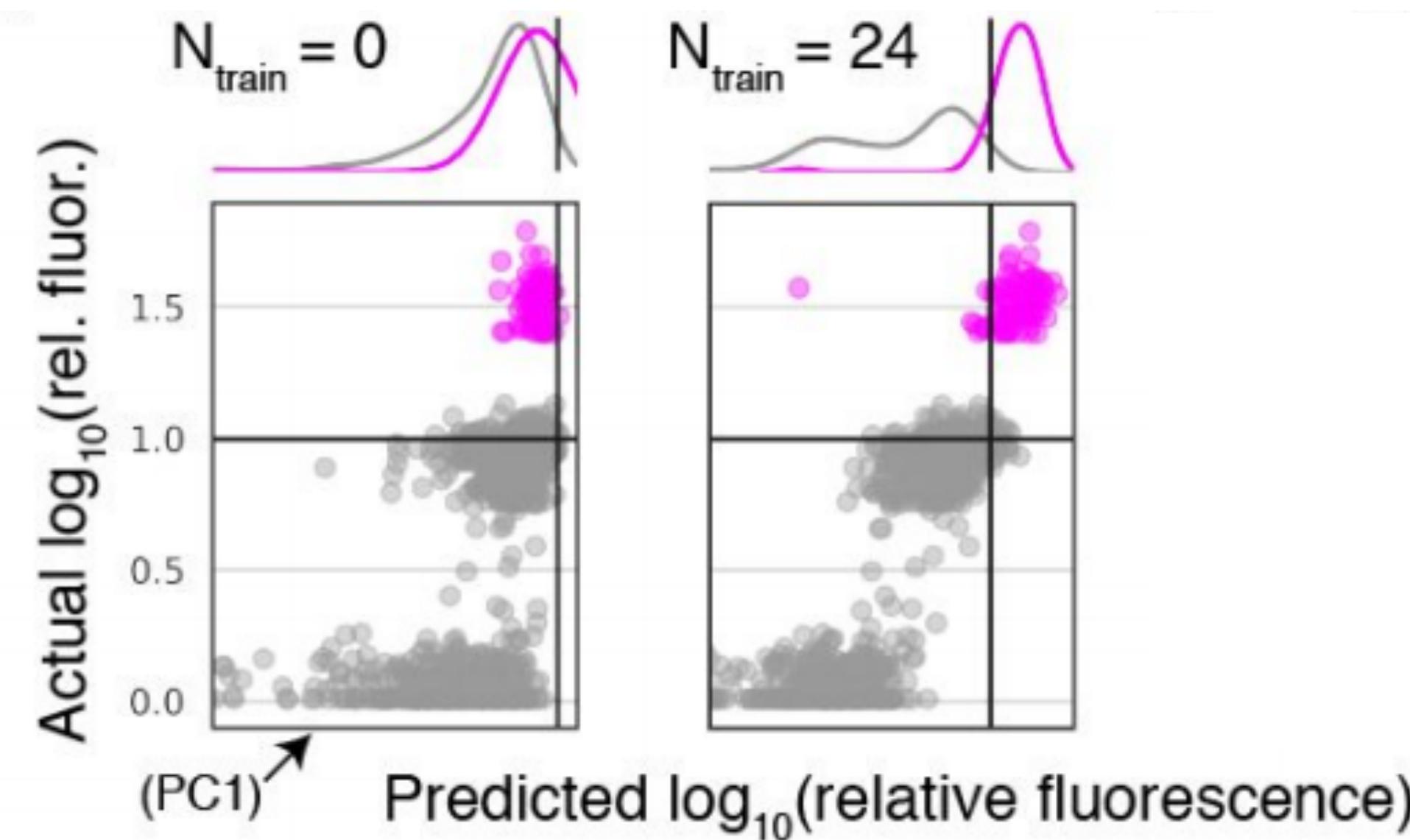


Pretraining guides search away from loss-of-function sequences

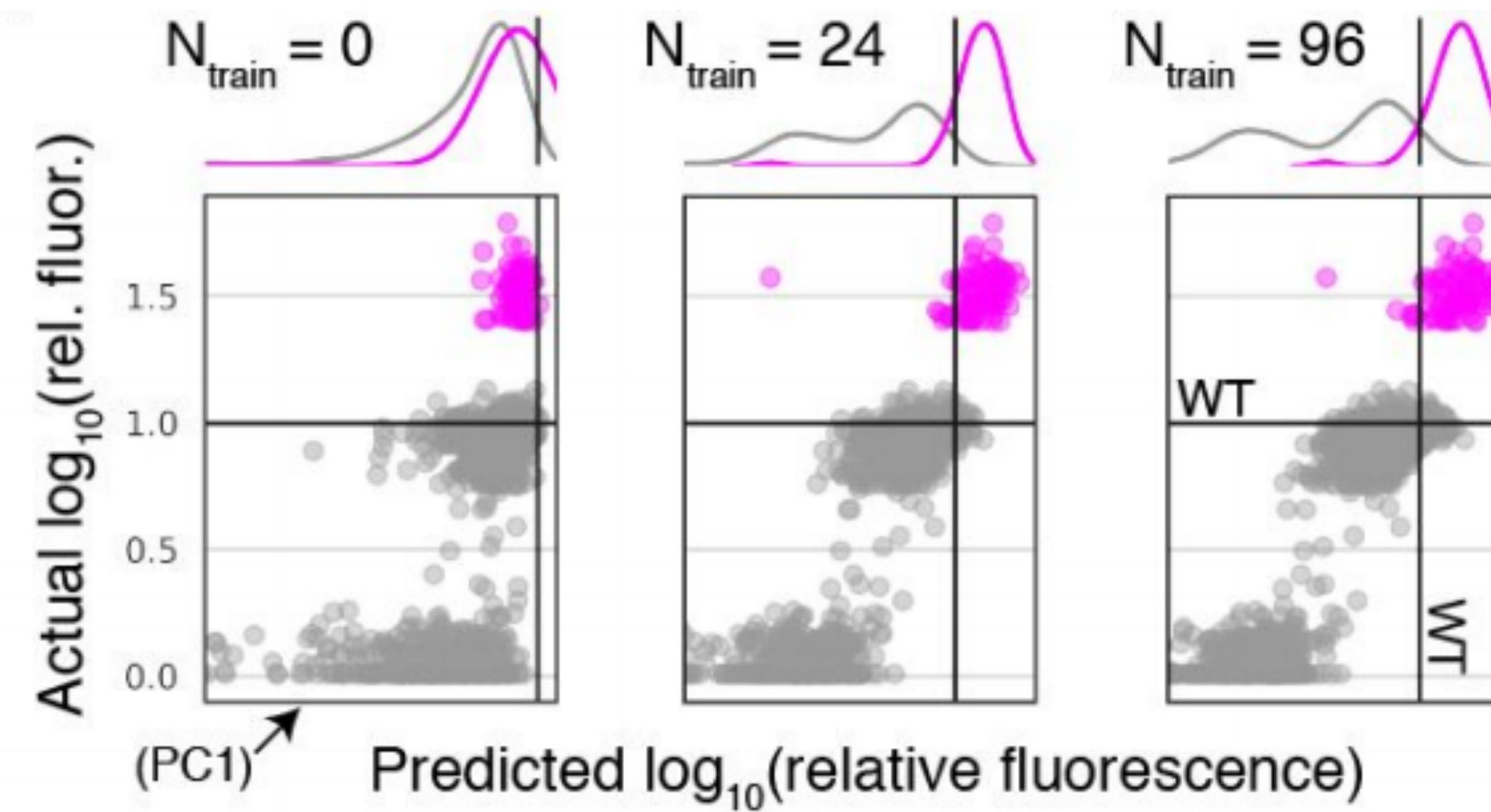
Pretraining guides search away from loss-of-function sequences



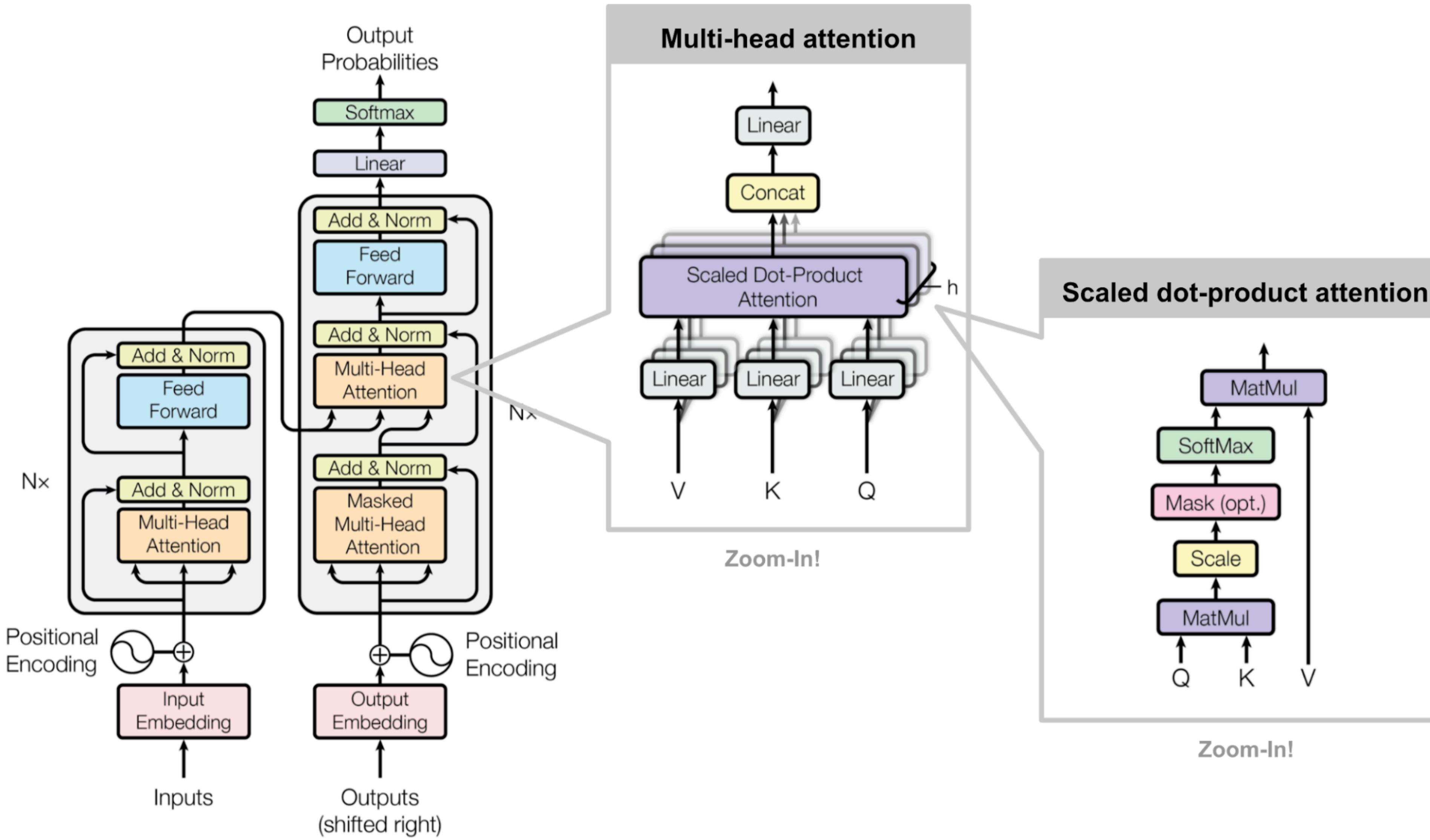
Pretraining guides search away from loss-of-function sequences



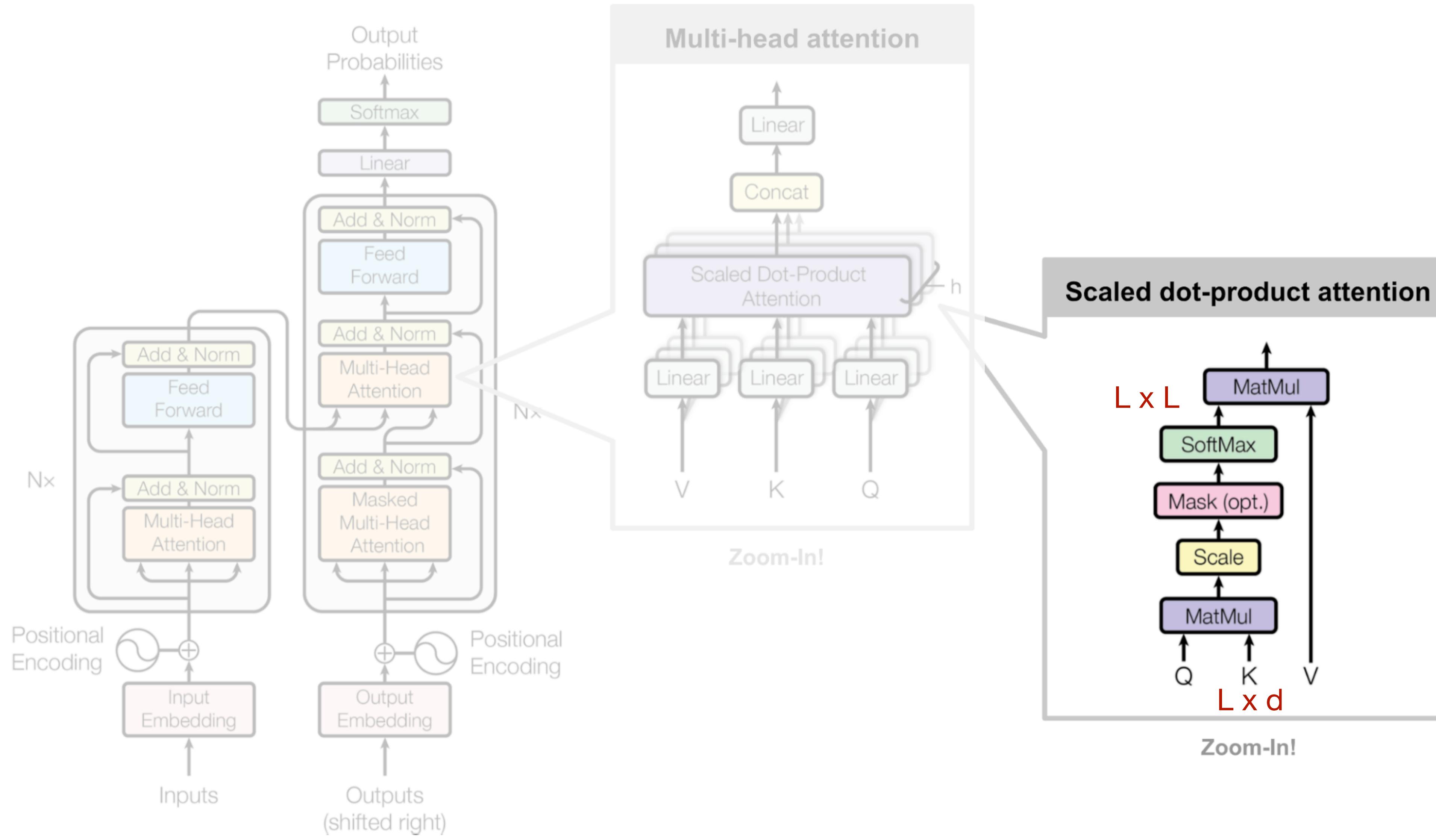
Pretraining guides search away from loss-of-function sequences



Transformers scale quadratically with length

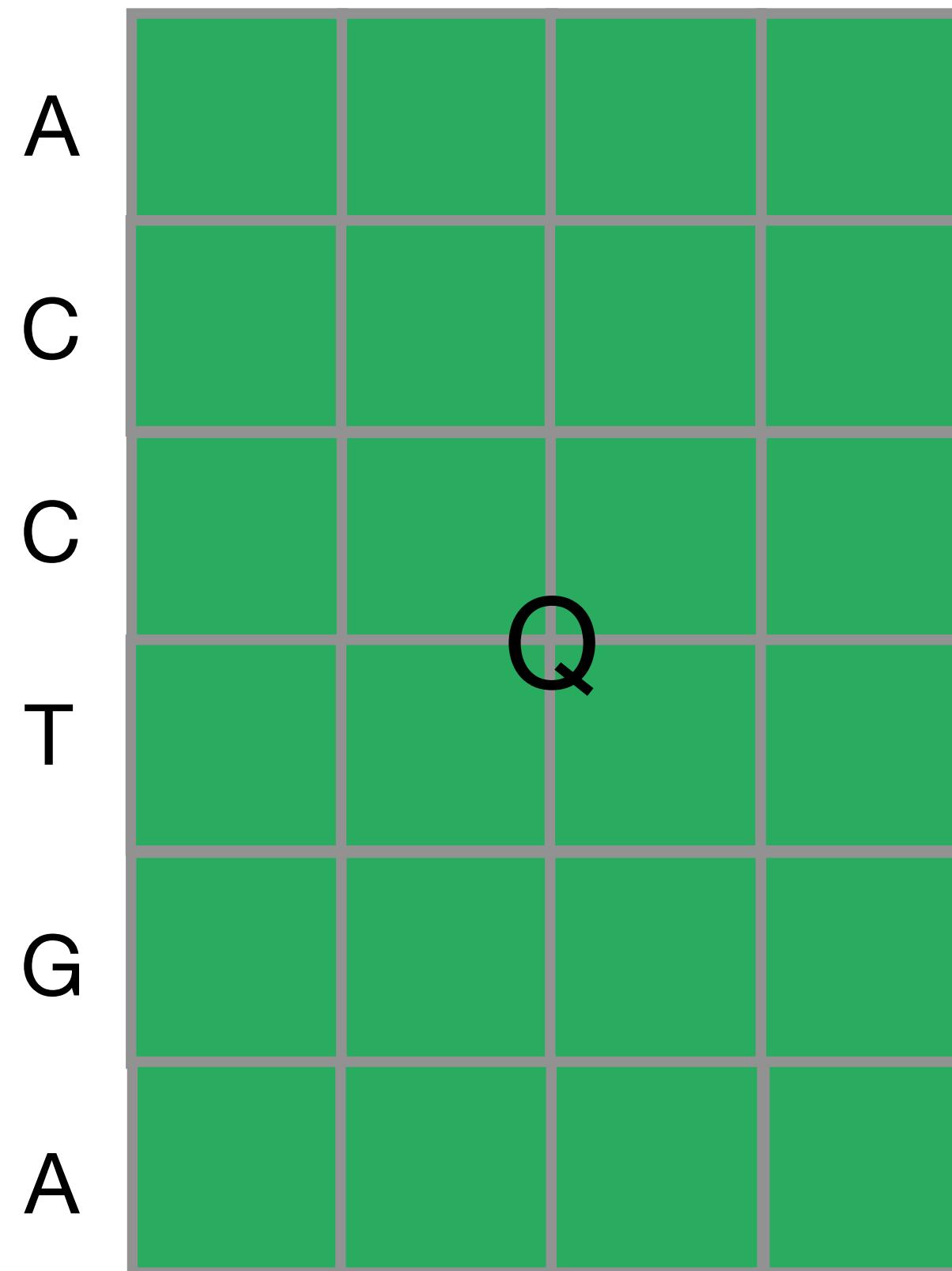


Transformers scale quadratically with length

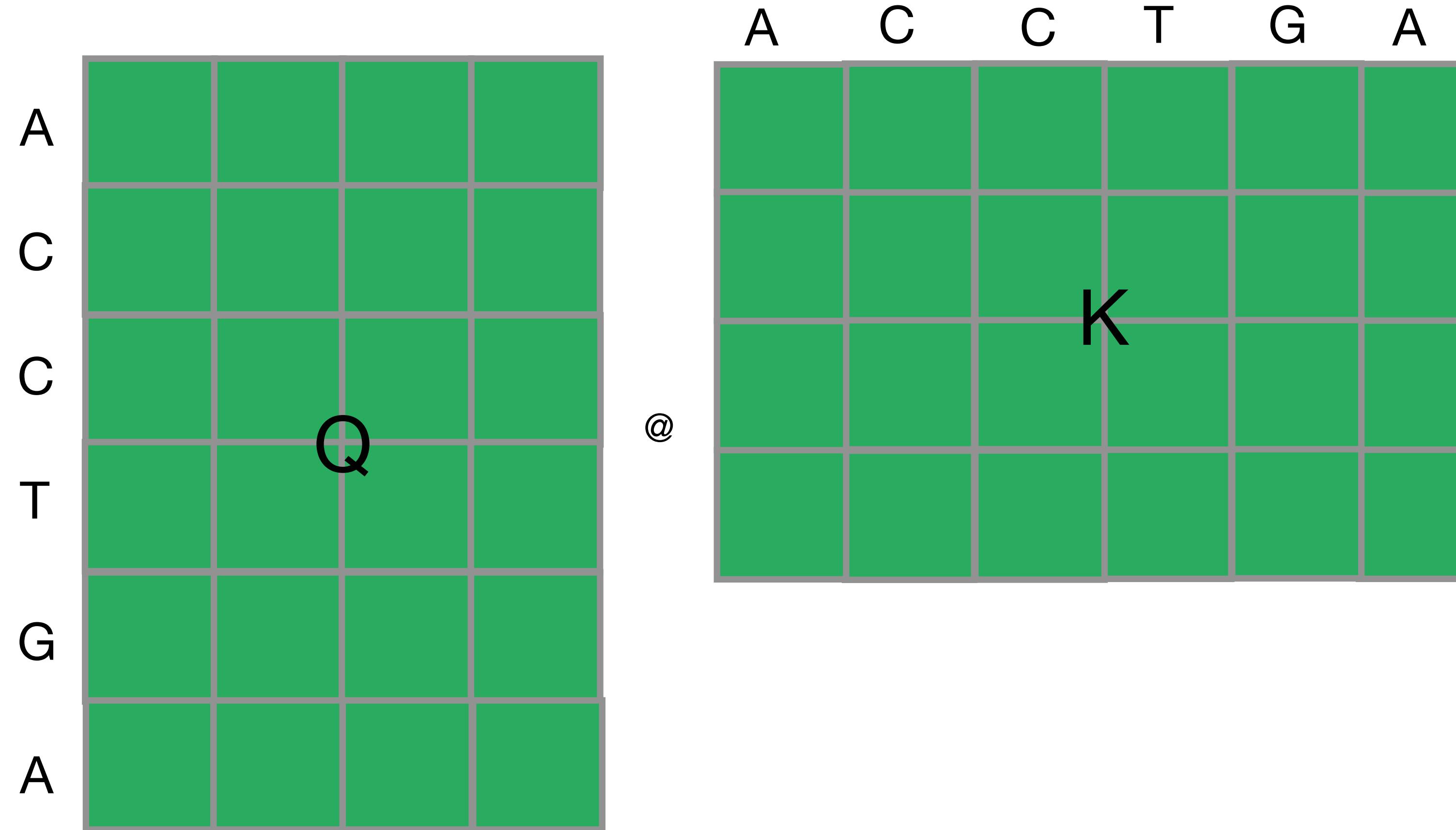


Transformers scale quadratically with length

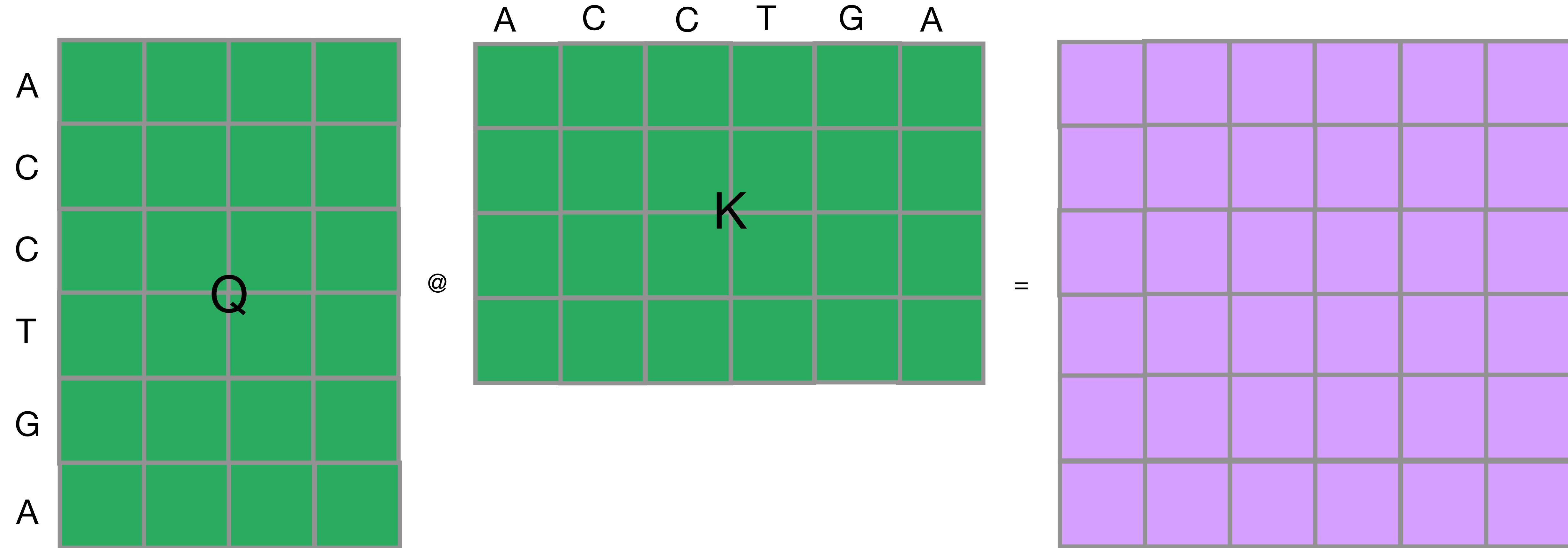
Transformers scale quadratically with length



Transformers scale quadratically with length



Transformers scale quadratically with length



Convolution scales linearly with length

A	1	0	0	0
C	0	1	0	0
C	0	1	0	0
T	0	0	0	1
G	0	0	1	0
A	1	0	0	0

0.2	0.4	0.1	0.1
0.1	0.7	0.1	0.1
0.1	0.1	0.1	0.7

Convolution scales linearly with length

A	0.2	0.4	0.1	0.1
C	0.1	0.7	0.1	0.1
C	0.1	0.1	0.1	0.7
T	0	0	0	1
G	0	0	1	0
A	1	0	0	0

0.2	0.4	0.1	0.1
0.1	0.7	0.1	0.1
0.1	0.1	0.1	0.7

Convolution scales linearly with length

A	0.2	0.4	0.1	0.1
C	0.1	0.7	0.1	0.1
C	0.1	0.1	0.1	0.7
T	0	0	0	1
G	0	0	1	0
A	1	0	0	0

0.2	0.4	0.1	0.1
0.1	0.7	0.1	0.1
0.1	0.1	0.1	0.7

1.0

$$0.2 + 0.7 + 0.1$$

Convolution scales linearly with length

A	1	0	0	0
C	0.2	0.4	0.1	0.1
C	0.1	0.7	0.1	0.1
T	0.1	0.1	0.1	0.7
G	0	0	1	0
A	1	0	0	0

0.2	0.4	0.1	0.1
0.1	0.7	0.1	0.1
0.1	0.1	0.1	0.7

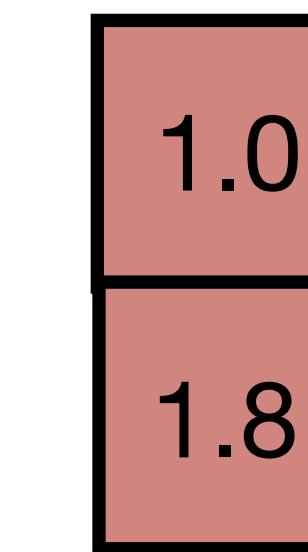
1.0

$$0.2 + 0.7 + 0.1$$

Convolution scales linearly with length

A	1	0	0	0
C	0.2	0.4	0.1	0.1
C	0.1	0.7	0.1	0.1
T	0.1	0.1	0.1	0.7
G	0	0	1	0
A	1	0	0	0

0.2	0.4	0.1	0.1
0.1	0.7	0.1	0.1
0.1	0.1	0.1	0.7



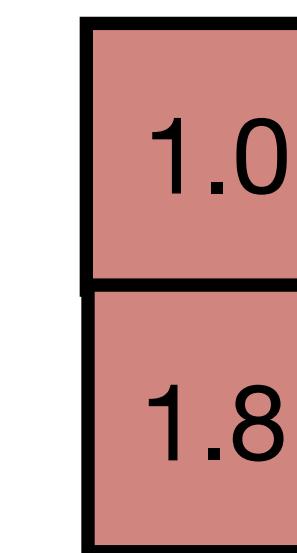
$$0.2 + 0.7 + 0.1$$

$$0.4 + 0.7 + 0.7$$

Convolution scales linearly with length

A	1	0	0	0
C	0	1	0	0
C	0.2	0.4	0.1	0.1
T	0.1	0.7	0.1	0.1
G	0.1	0.1	0.1	0.7
A	1	0	0	0

0.2	0.4	0.1	0.1
0.1	0.7	0.1	0.1
0.1	0.1	0.1	0.7



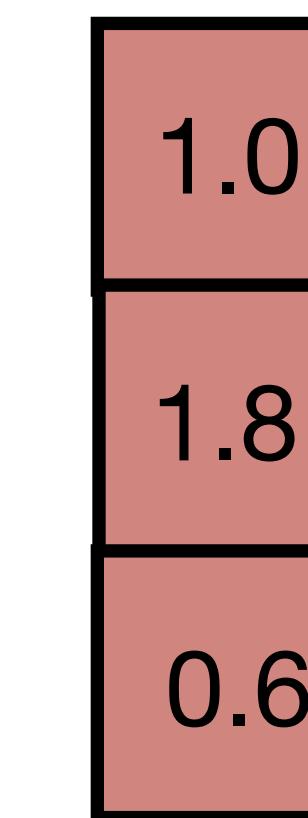
$$0.2 + 0.7 + 0.1$$

$$0.4 + 0.7 + 0.7$$

Convolution scales linearly with length

A	1	0	0	0
C	0	1	0	0
C	0.2	0.4	0.1	0.1
T	0.1	0.7	0.1	0.1
G	0.1	0.1	0.1	0.7
A	1	0	0	0

0.2	0.4	0.1	0.1
0.1	0.7	0.1	0.1
0.1	0.1	0.1	0.7



$$0.2 + 0.7 + 0.1$$

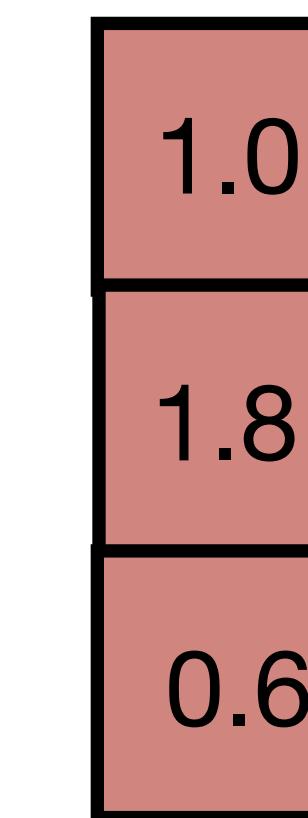
$$0.4 + 0.7 + 0.7$$

$$0.4 + 0.1 + 0.1$$

Convolution scales linearly with length

A	1	0	0	0
C	0	1	0	0
C	0	1	0	0
T	0.2	0.4	0.1	0.1
G	0.1	0.7	0.1	0.1
A	0.1	0.1	0.1	0.7

0.2	0.4	0.1	0.1
0.1	0.7	0.1	0.1
0.1	0.1	0.1	0.7



$$0.2 + 0.7 + 0.1$$

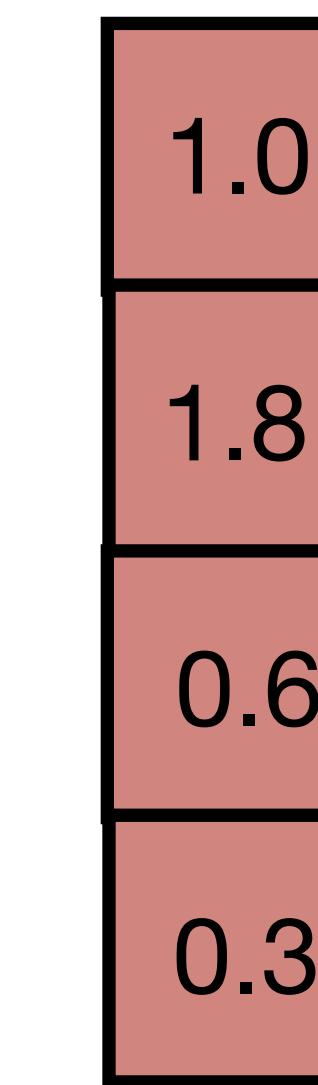
$$0.4 + 0.7 + 0.7$$

$$0.4 + 0.1 + 0.1$$

Convolution scales linearly with length

A	1	0	0	0
C	0	1	0	0
C	0	1	0	0
T	0.2	0.4	0.1	0.1
G	0.1	0.7	0.1	0.1
A	0.1	0.1	0.1	0.7

0.2	0.4	0.1	0.1
0.1	0.7	0.1	0.1
0.1	0.1	0.1	0.7



$$0.2 + 0.7 + 0.1$$

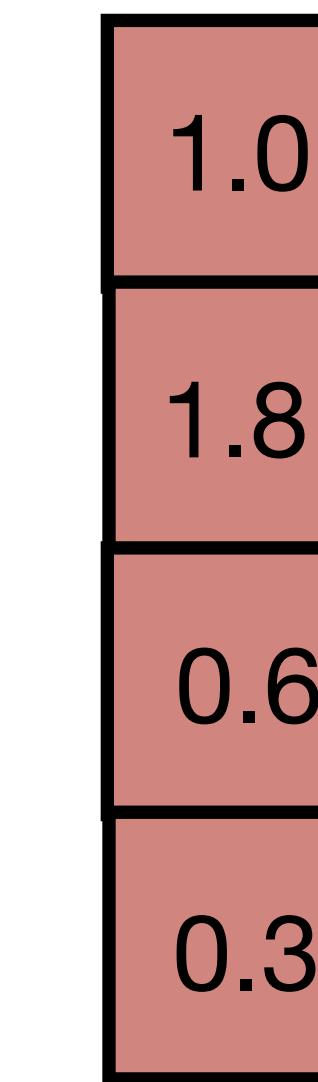
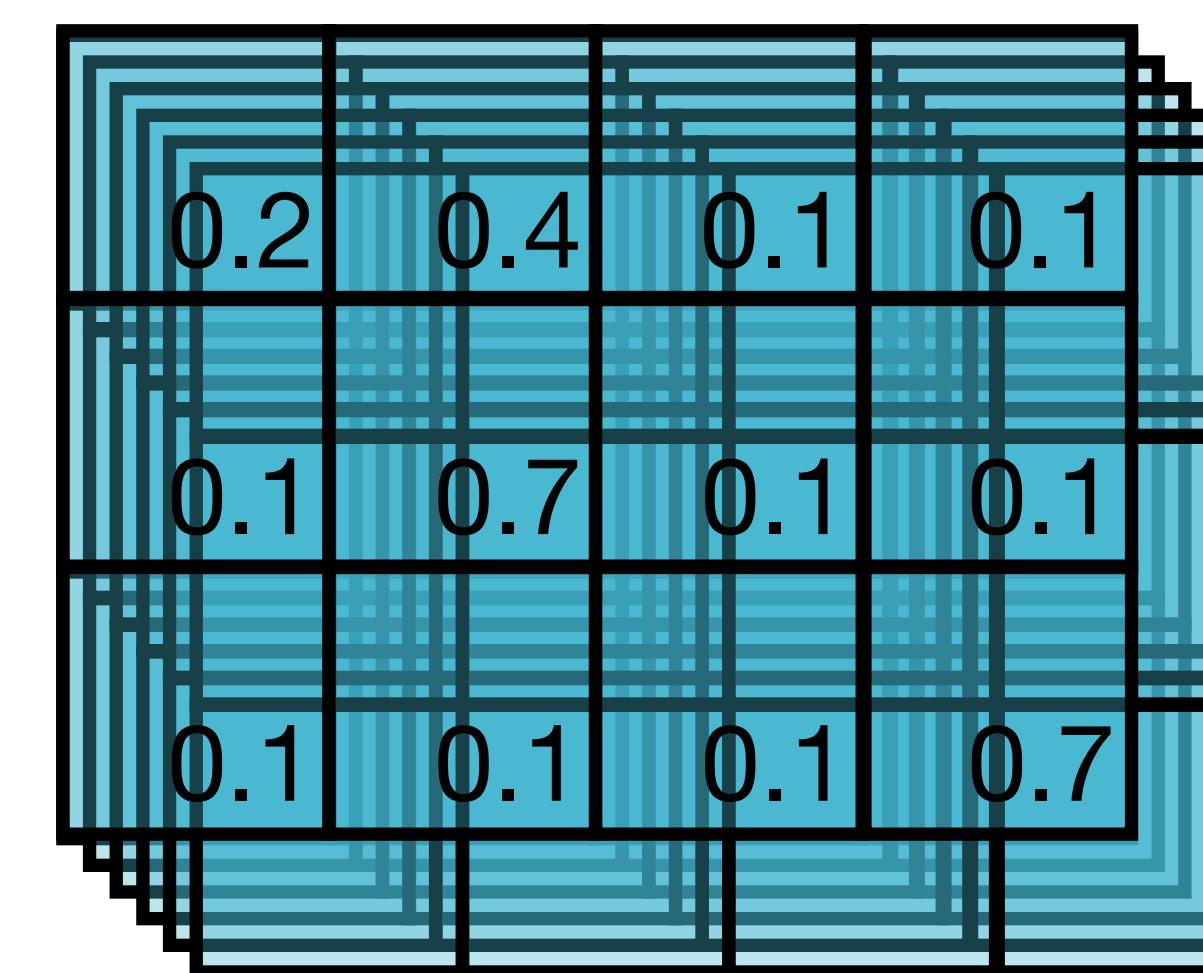
$$0.4 + 0.7 + 0.7$$

$$0.4 + 0.1 + 0.1$$

$$0.1 + 0.1 + 0.1$$

Convolution scales linearly with length

A	1	0	0	0
C	0	1	0	0
C	0	1	0	0
T	0.2	0.4	0.1	0.1
G	0.1	0.7	0.1	0.1
A	0.1	0.1	0.1	0.7



$$0.2 + 0.7 + 0.1$$

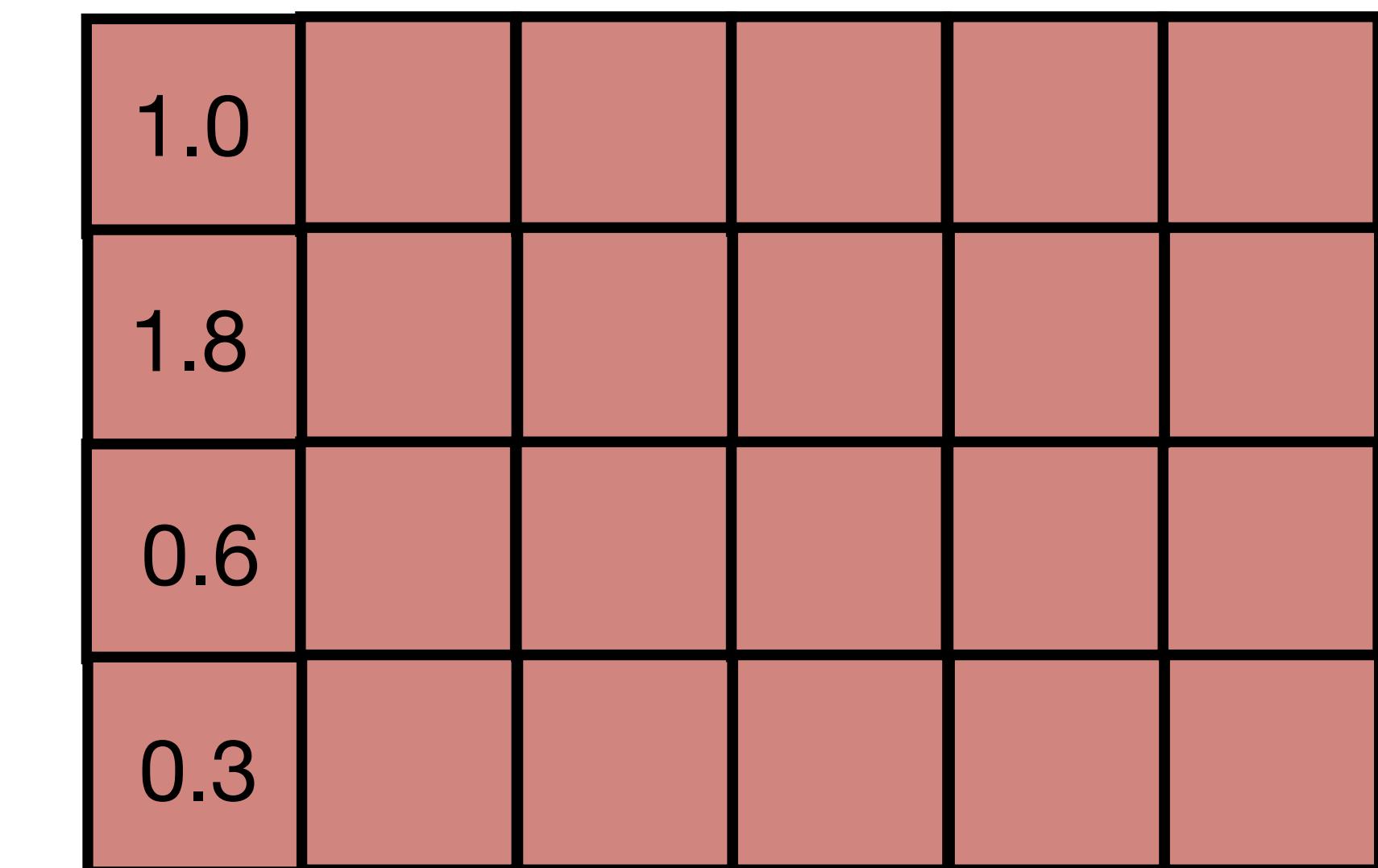
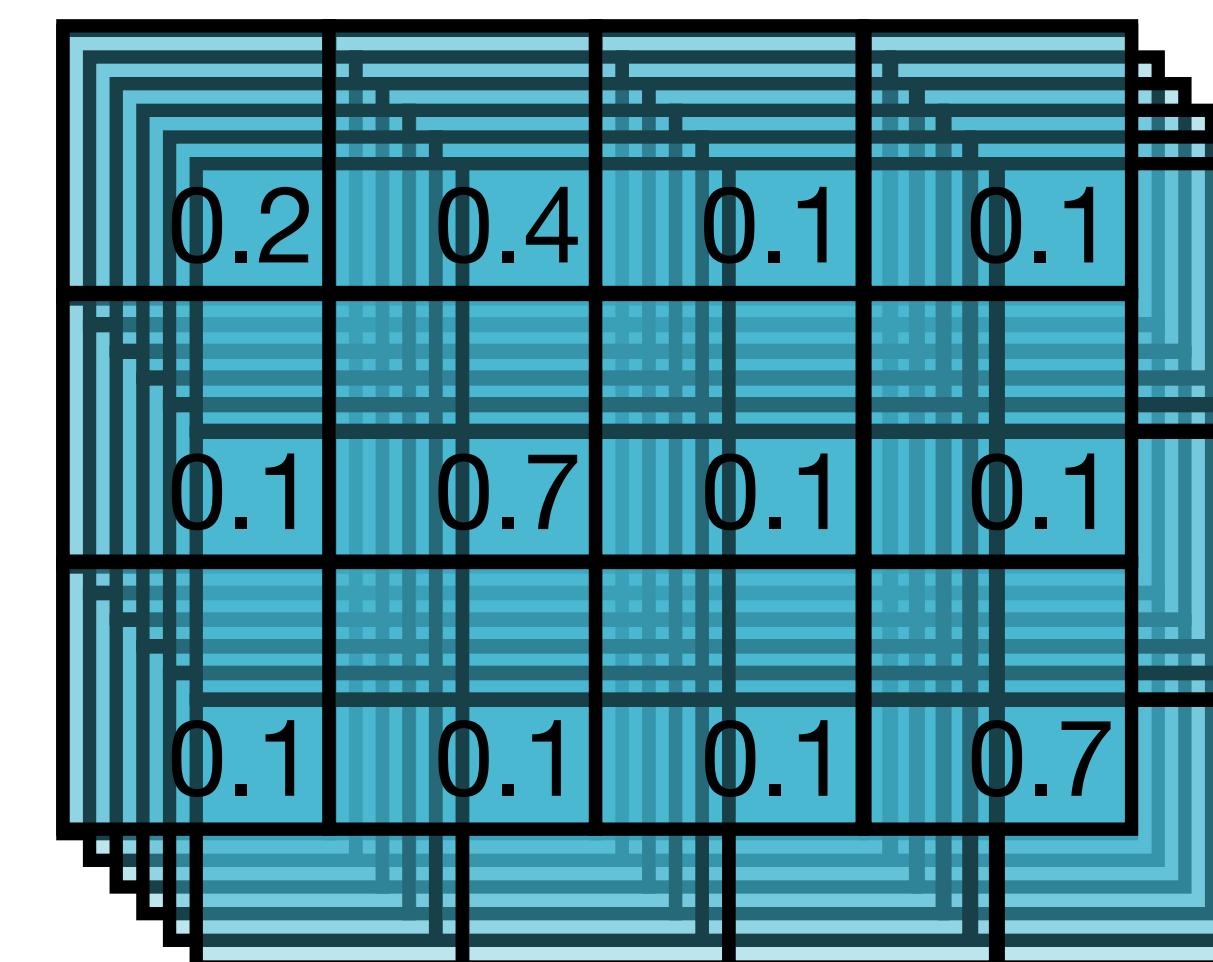
$$0.4 + 0.7 + 0.7$$

$$0.4 + 0.1 + 0.1$$

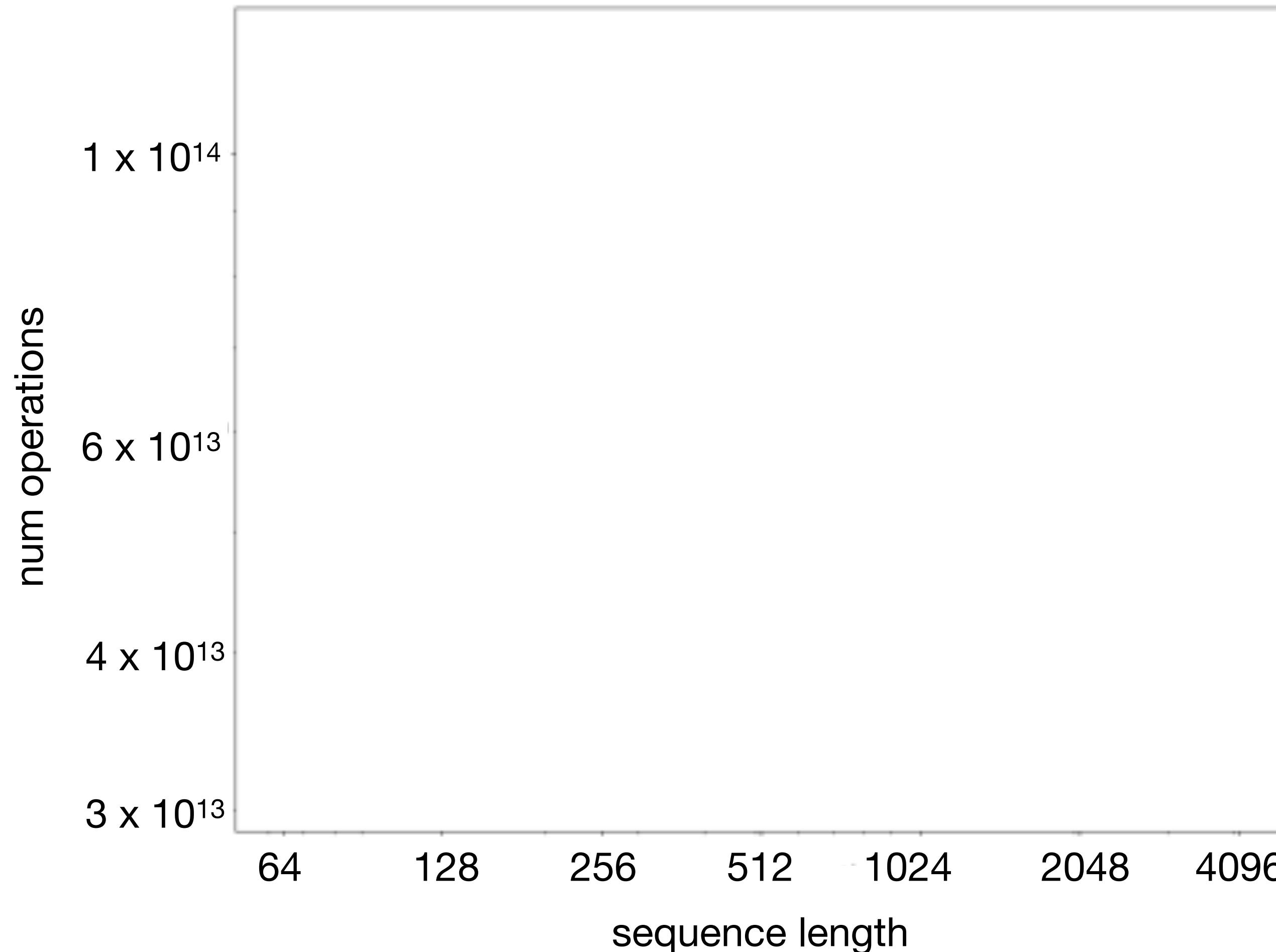
$$0.1 + 0.1 + 0.1$$

Convolution scales linearly with length

A	1	0	0	0
C	0	1	0	0
C	0	1	0	0
T	0.2	0.4	0.1	0.1
G	0.1	0.7	0.1	0.1
A	0.1	0.1	0.1	0.7

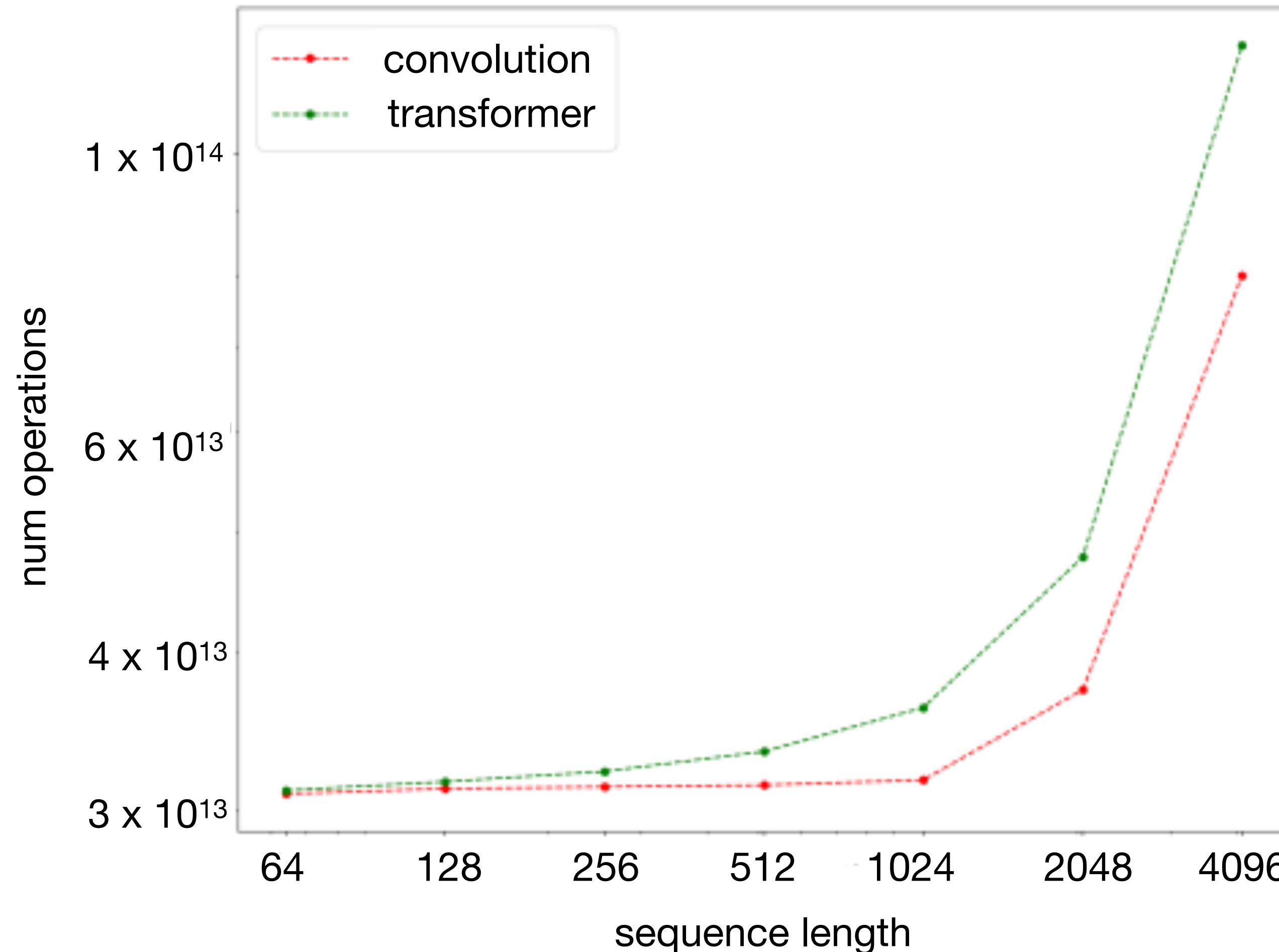


Convolution scales linearly with length



Tay et al. 2022

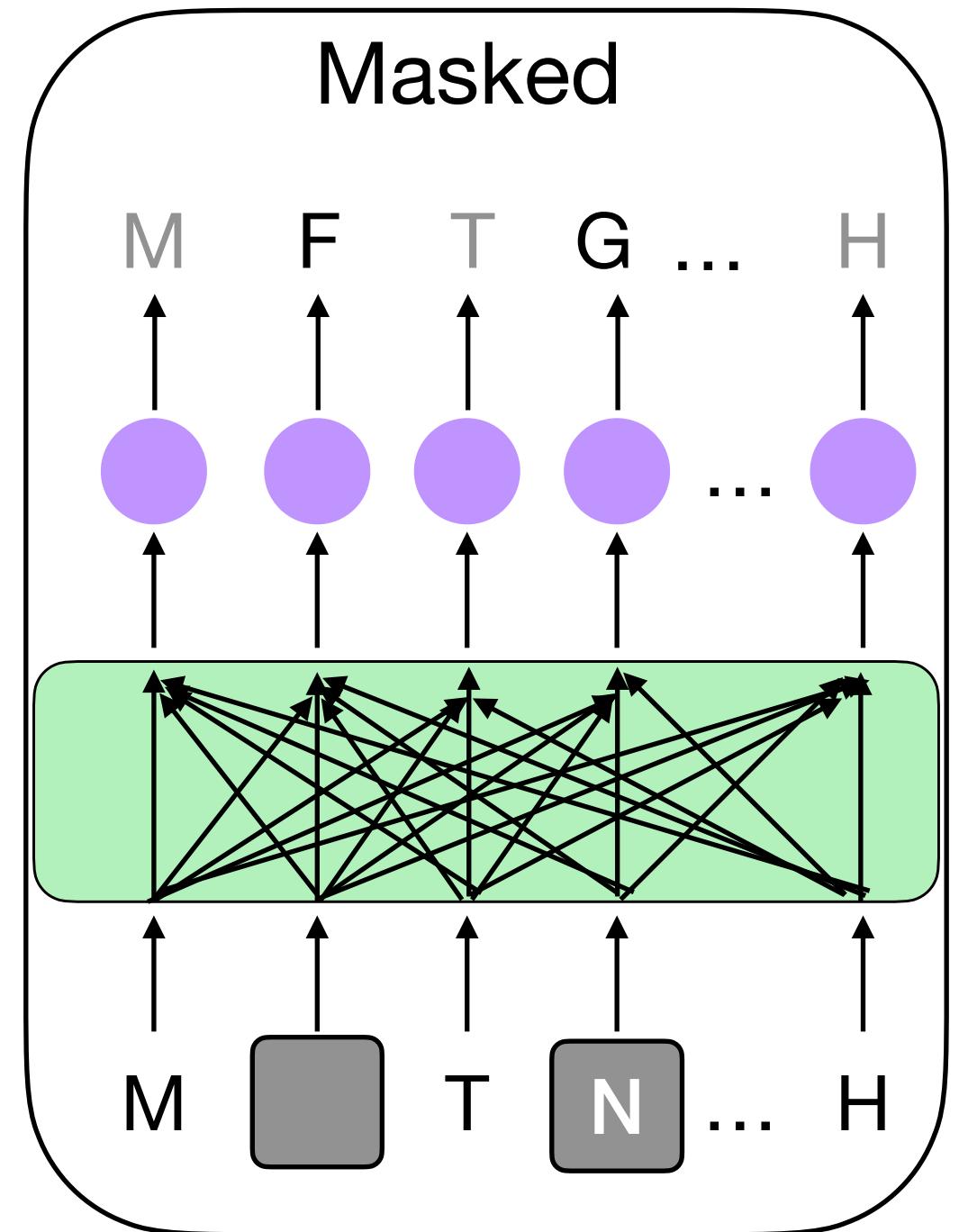
Convolution scales linearly with length



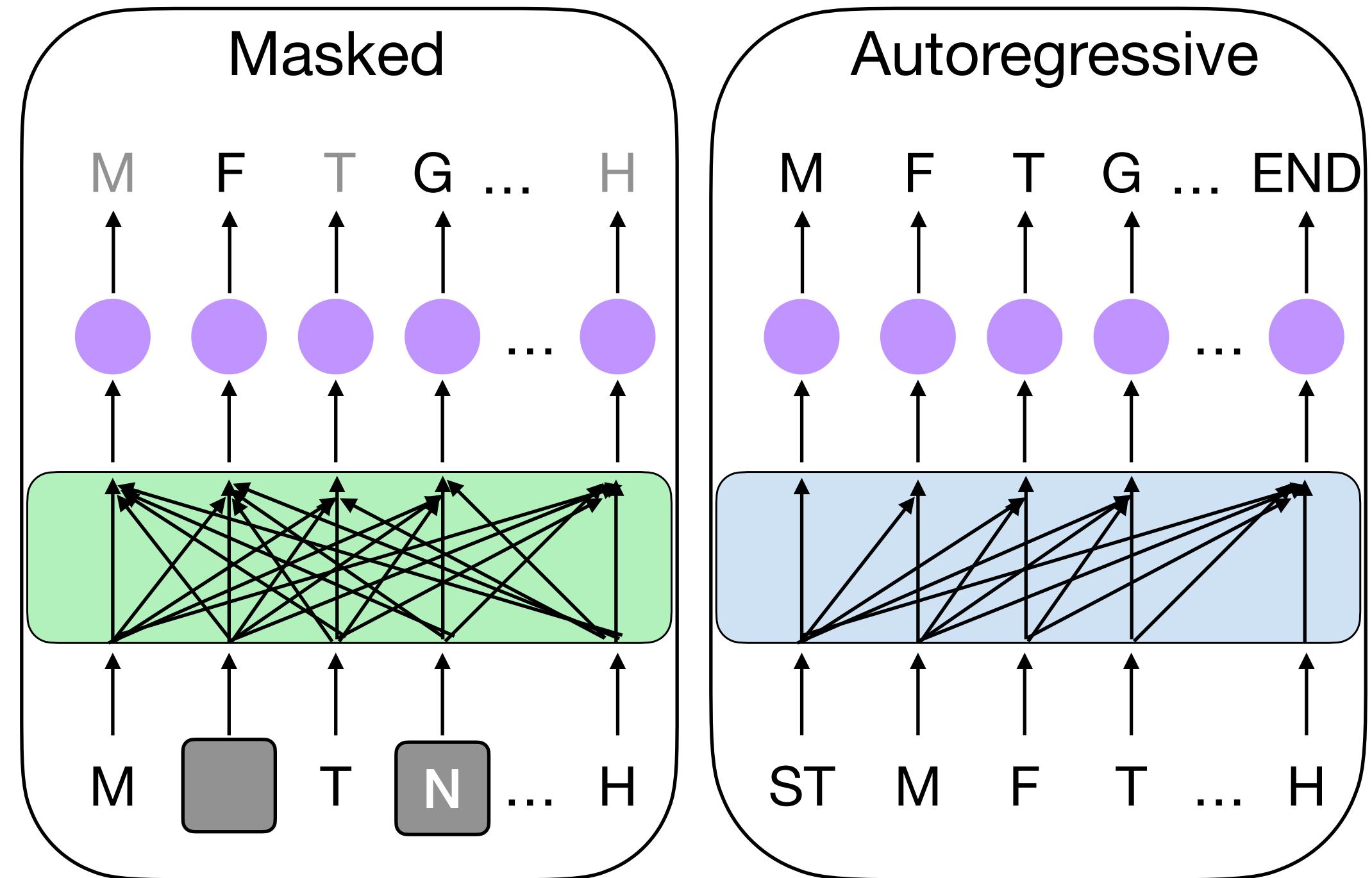
Tay et al. 2022

Separate pretraining task

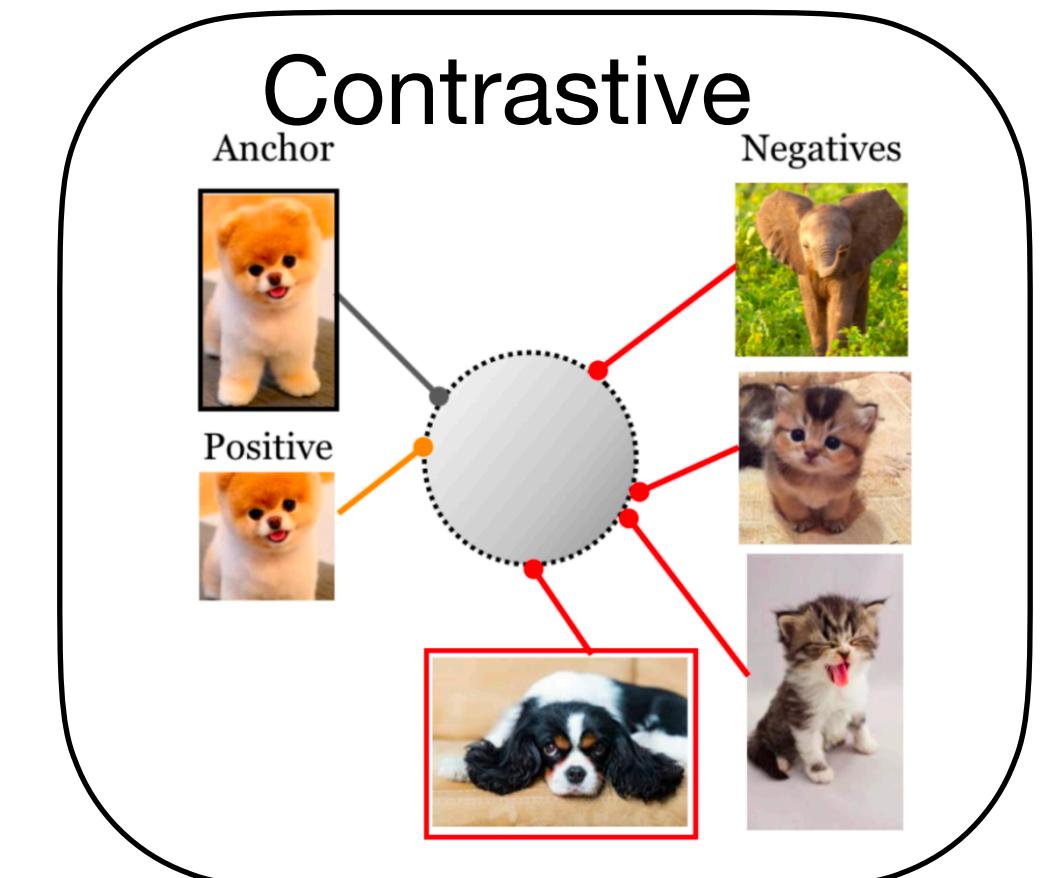
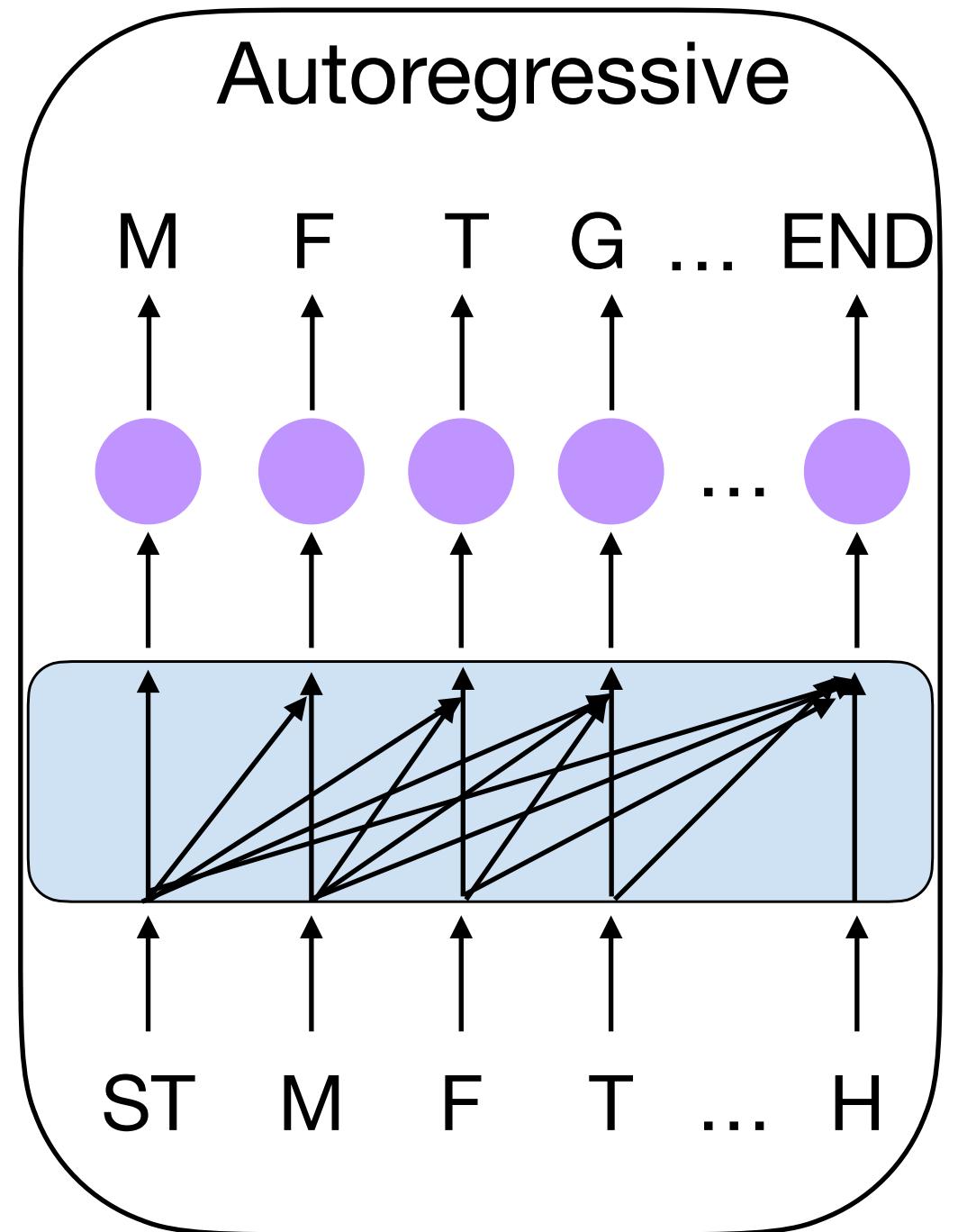
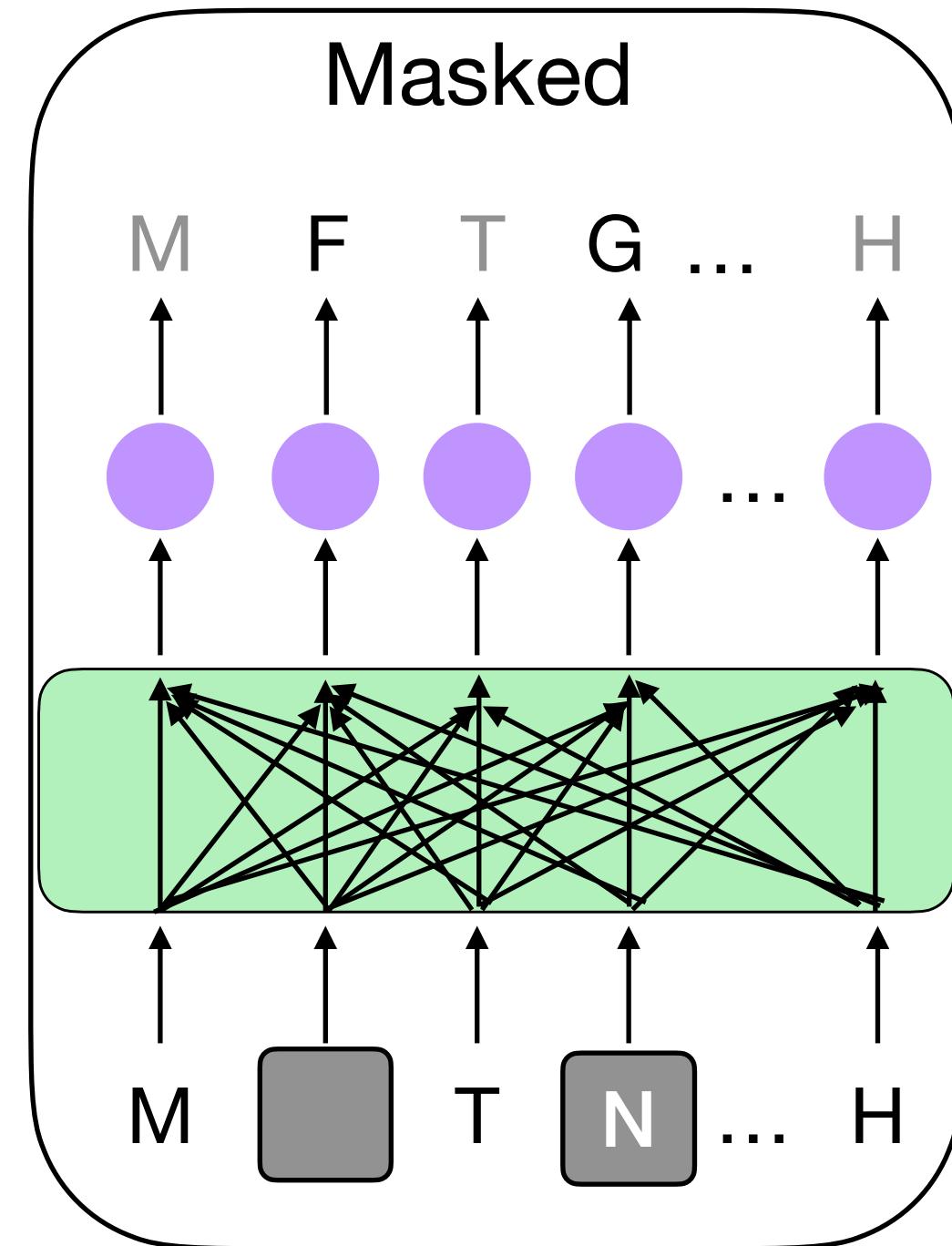
Separate pretraining task



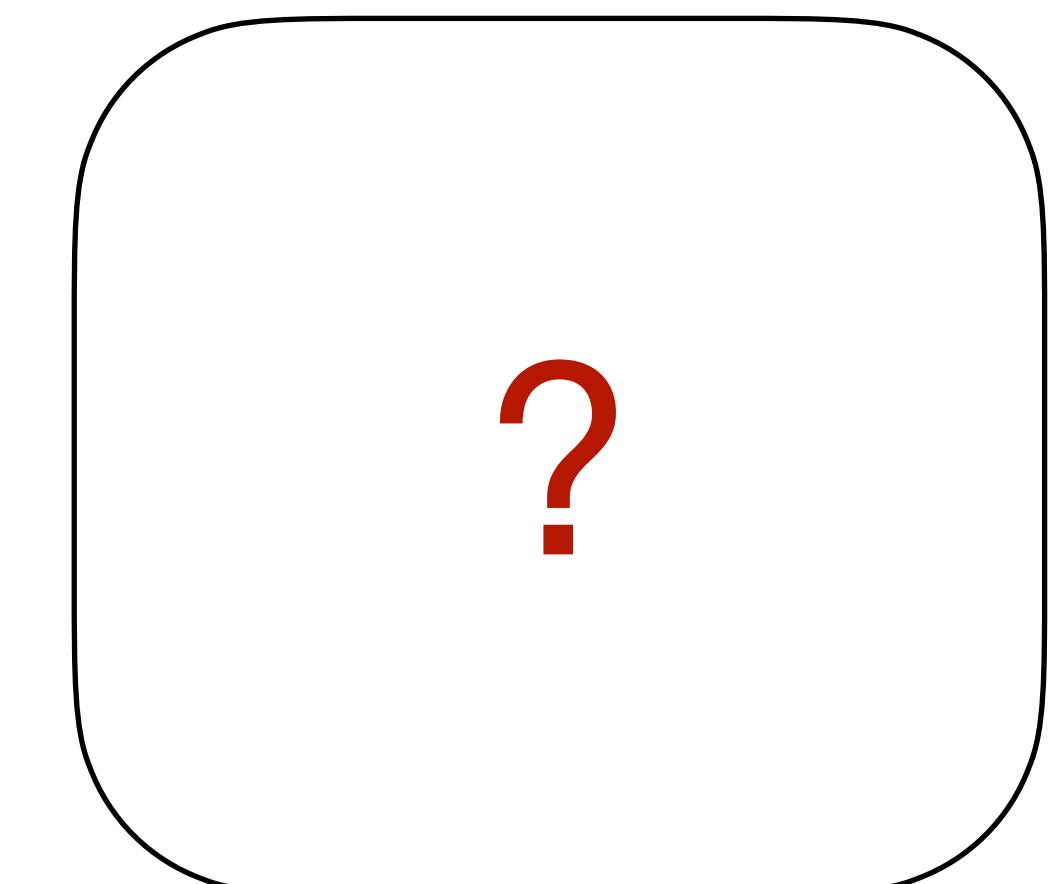
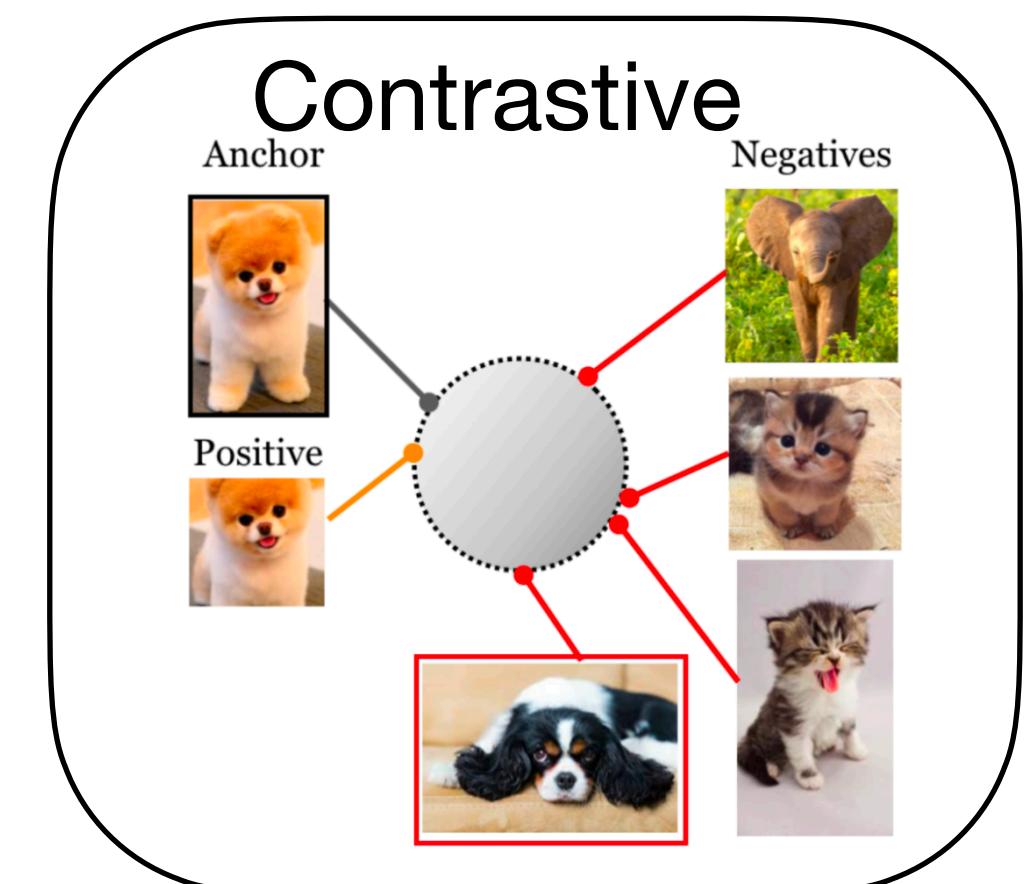
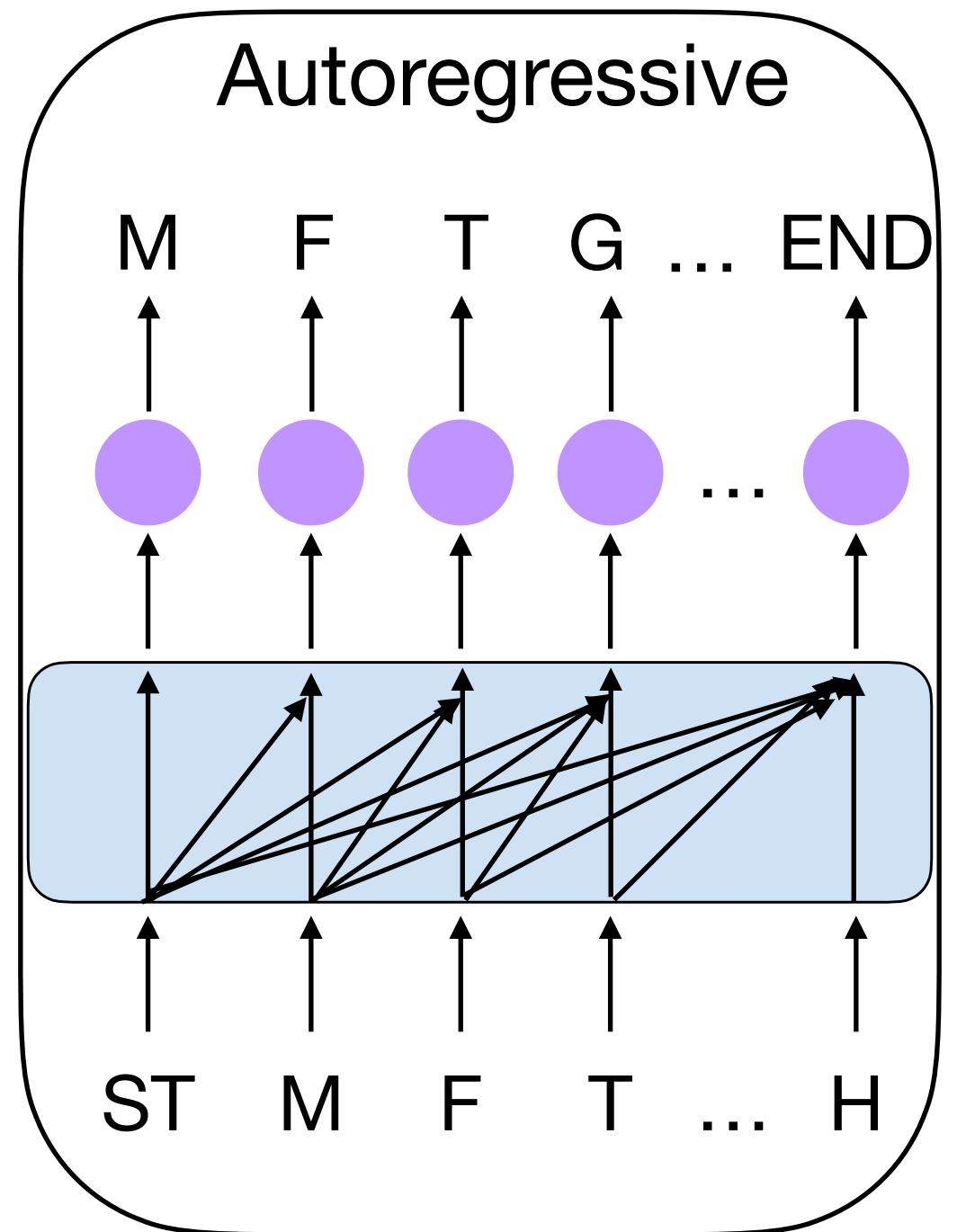
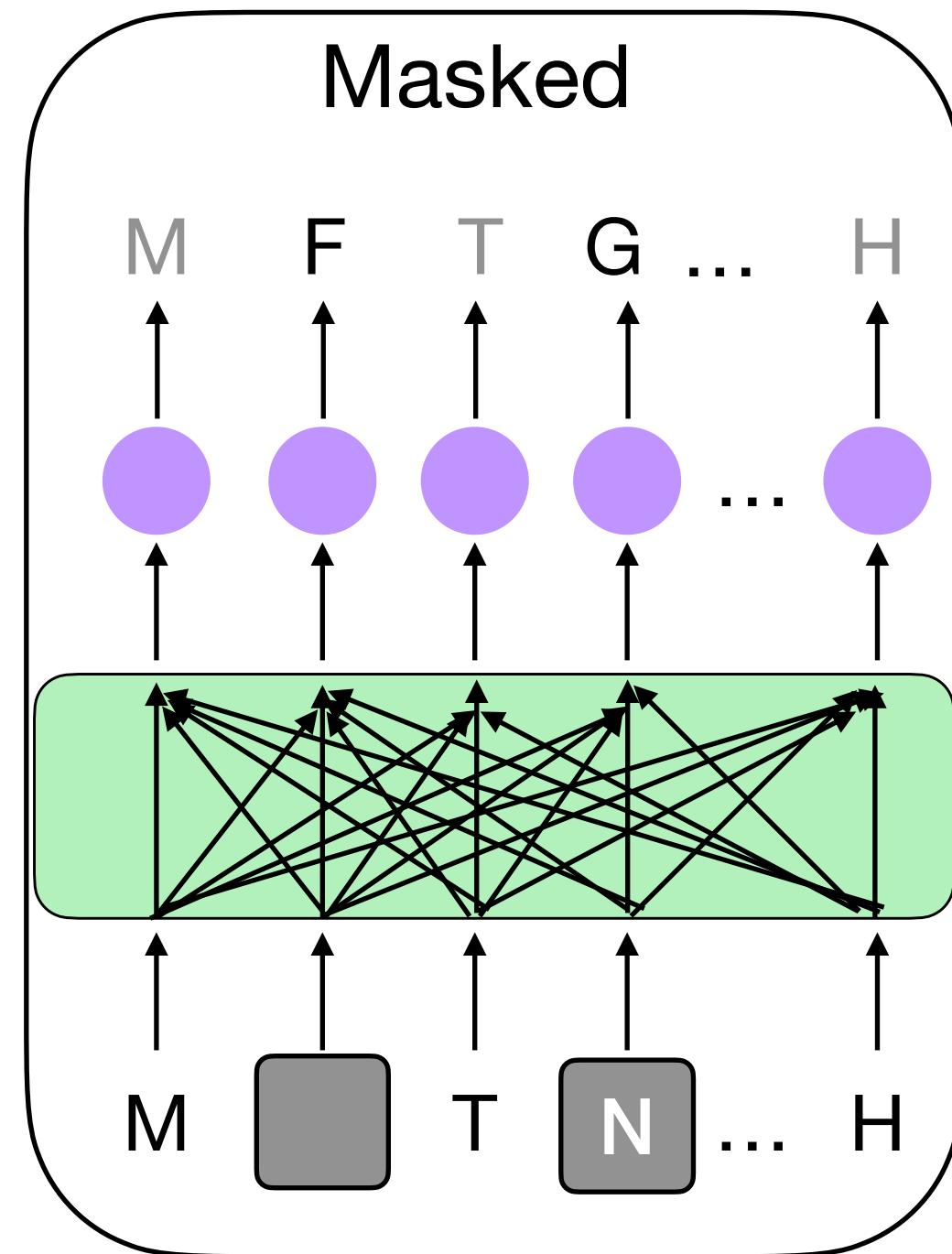
Separate pretraining task



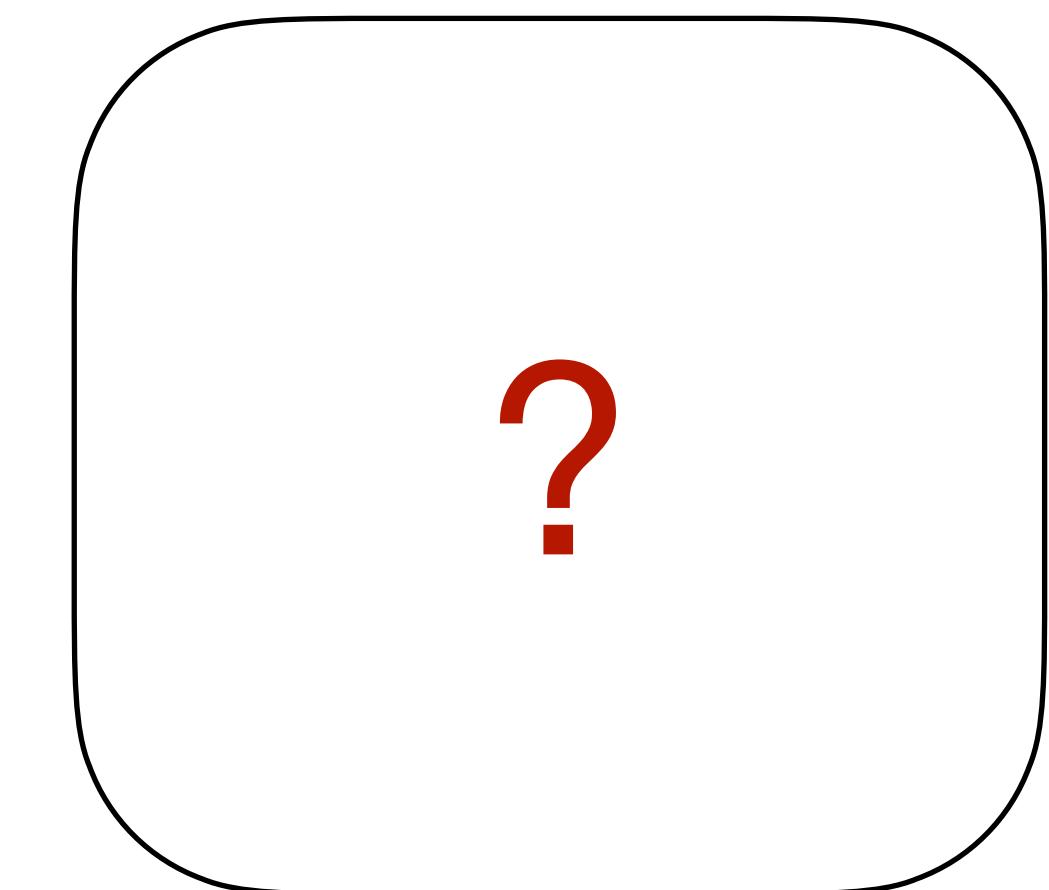
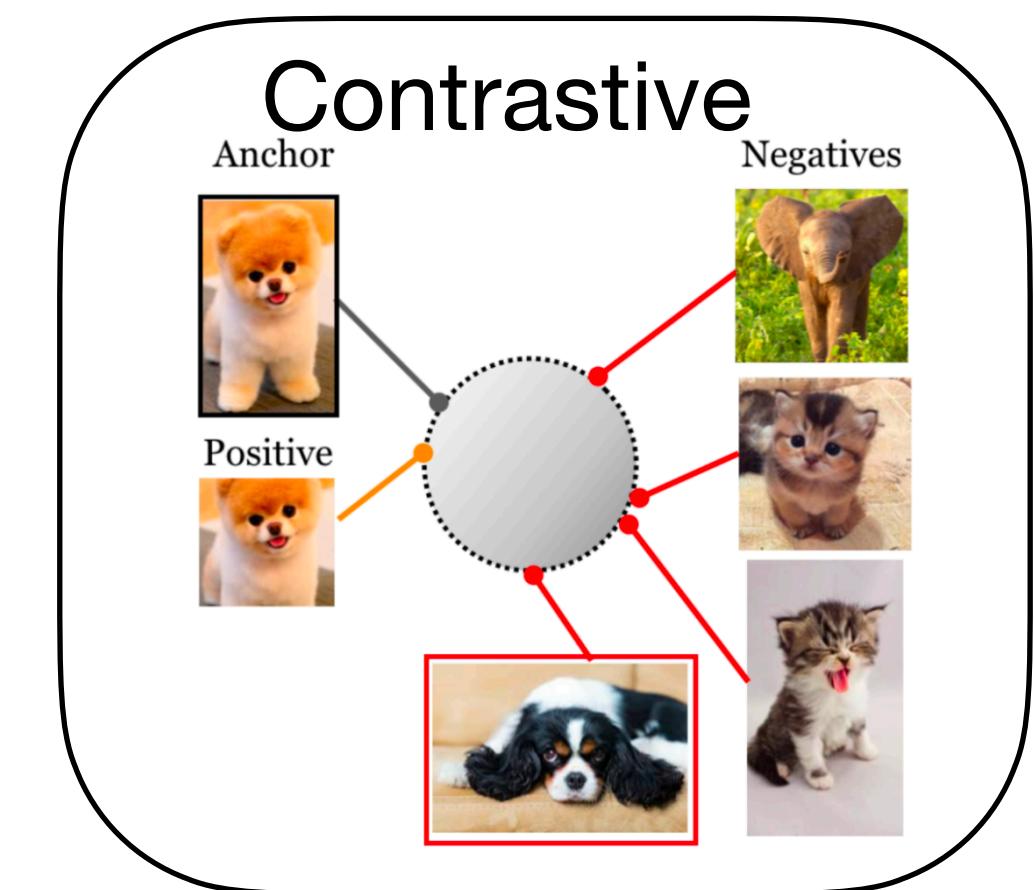
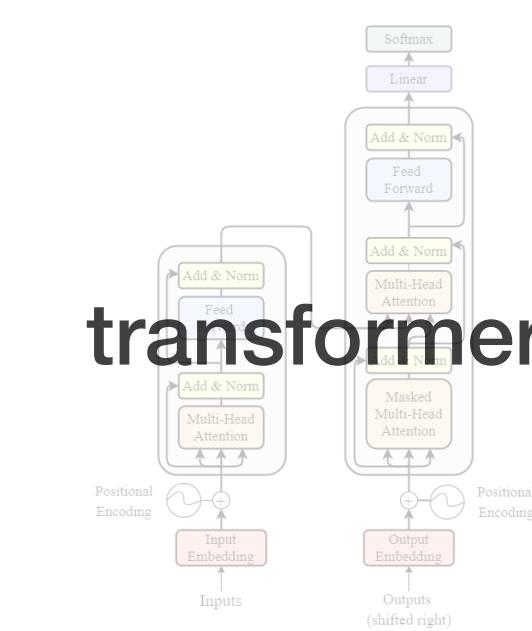
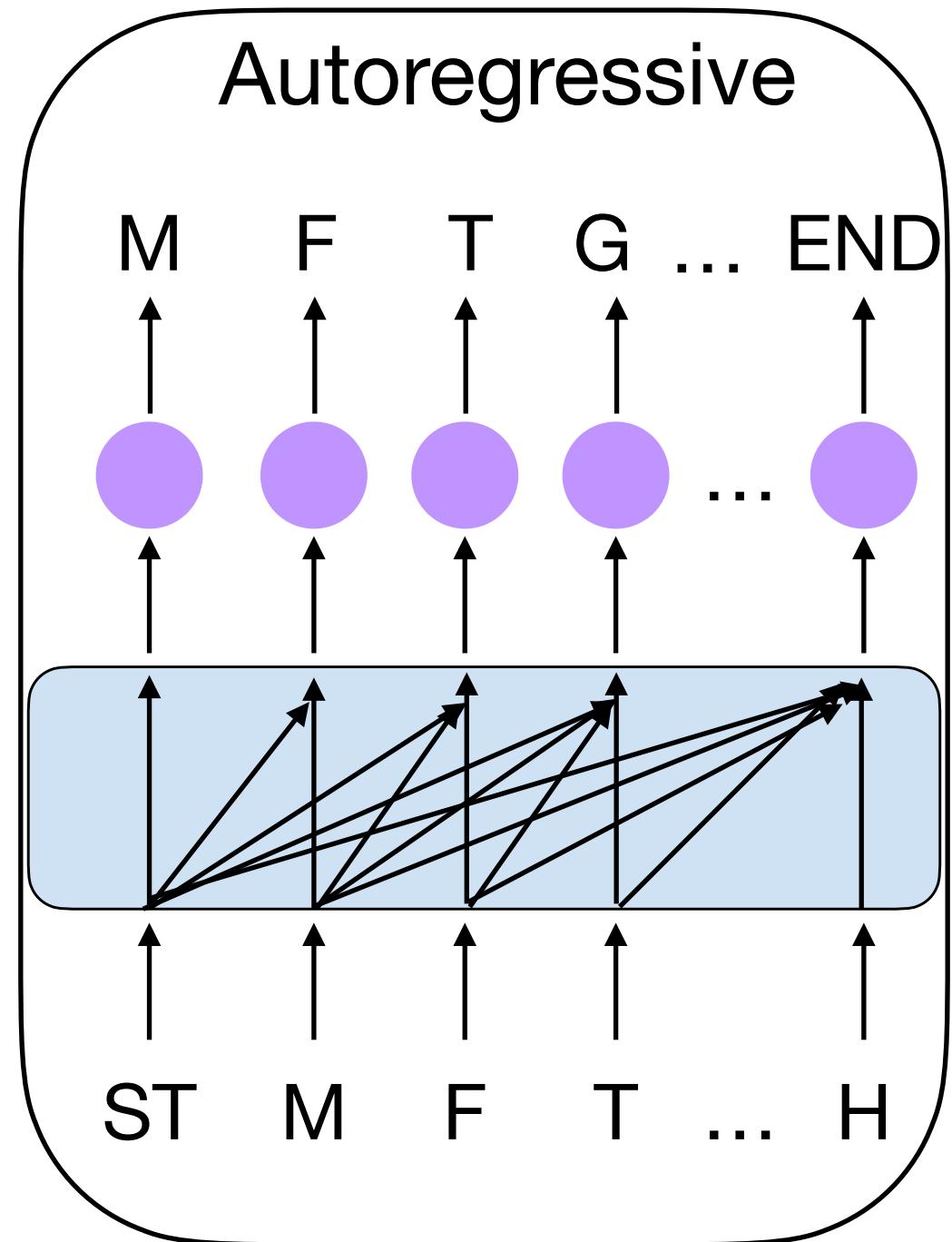
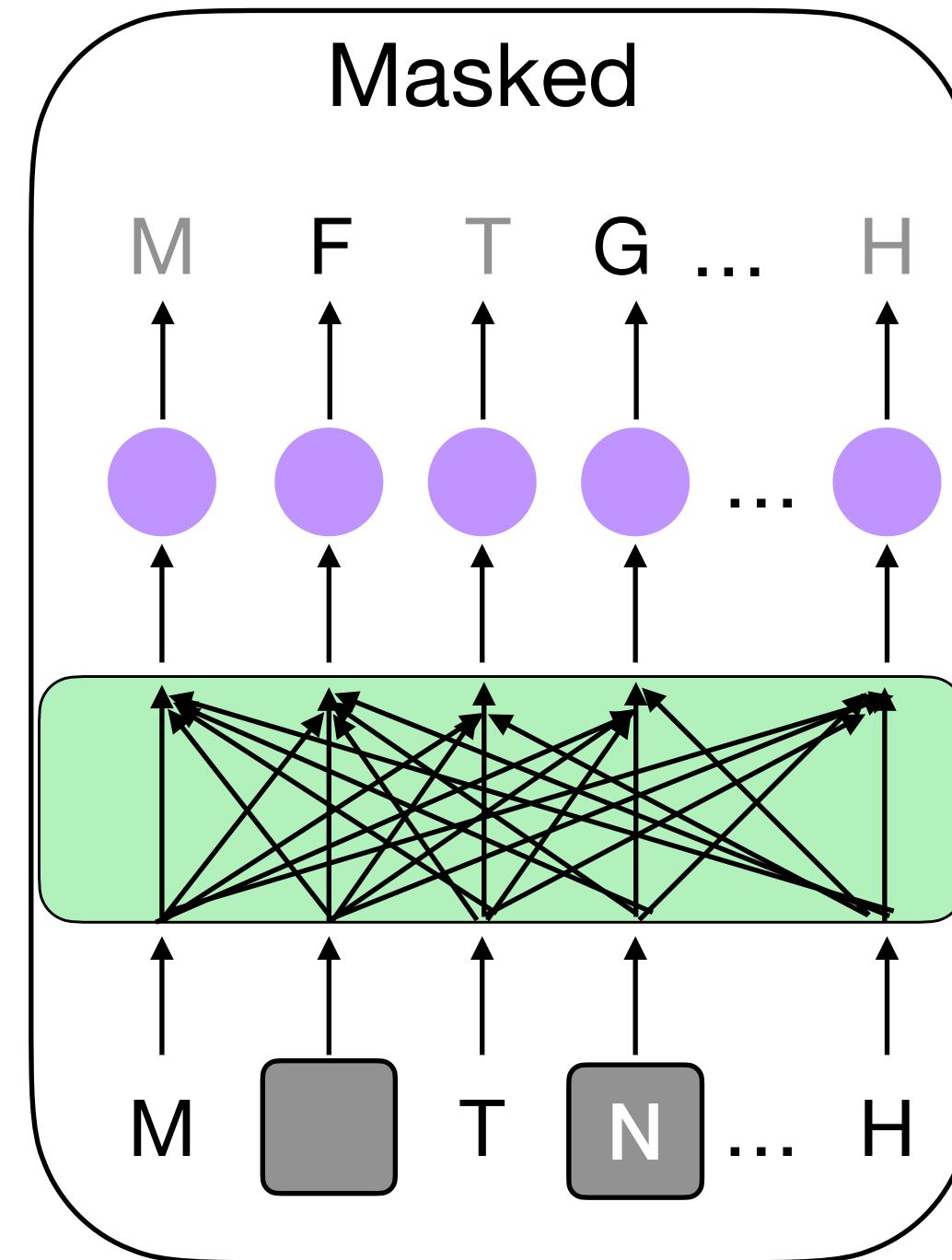
Separate pretraining task



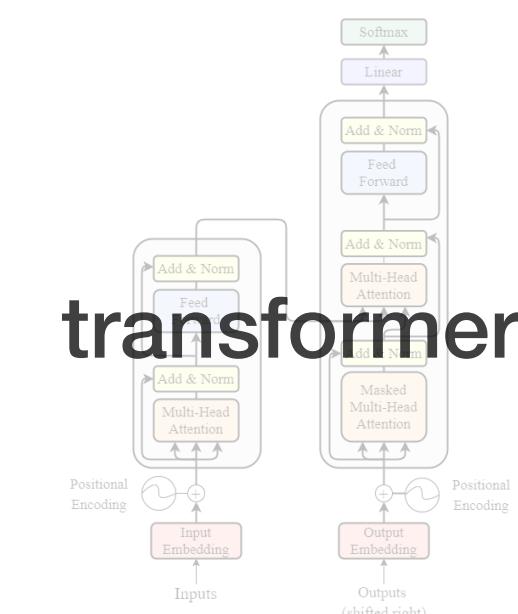
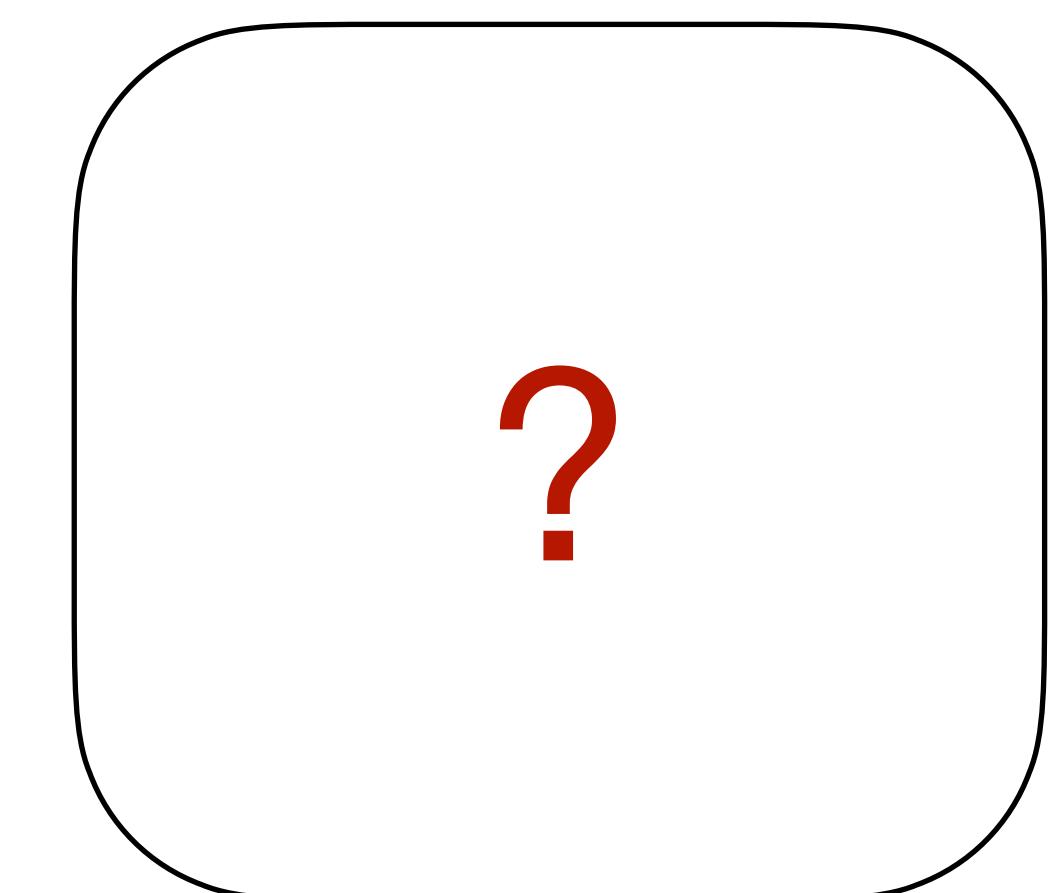
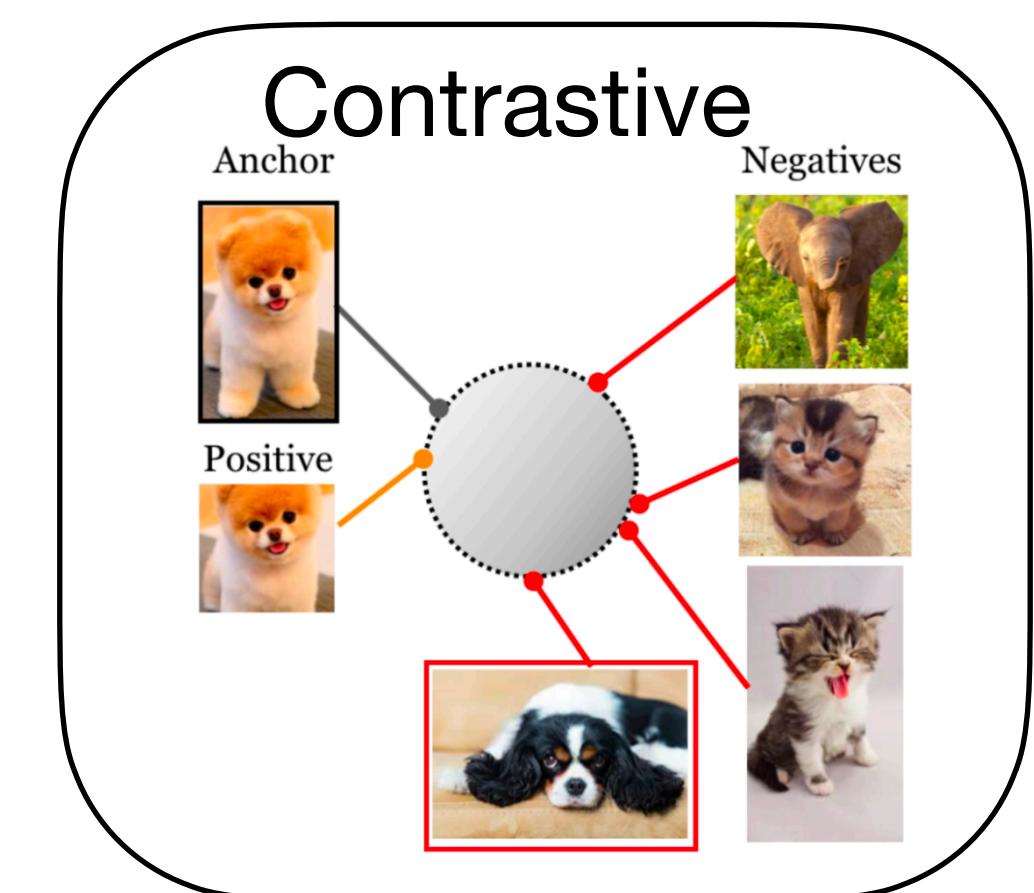
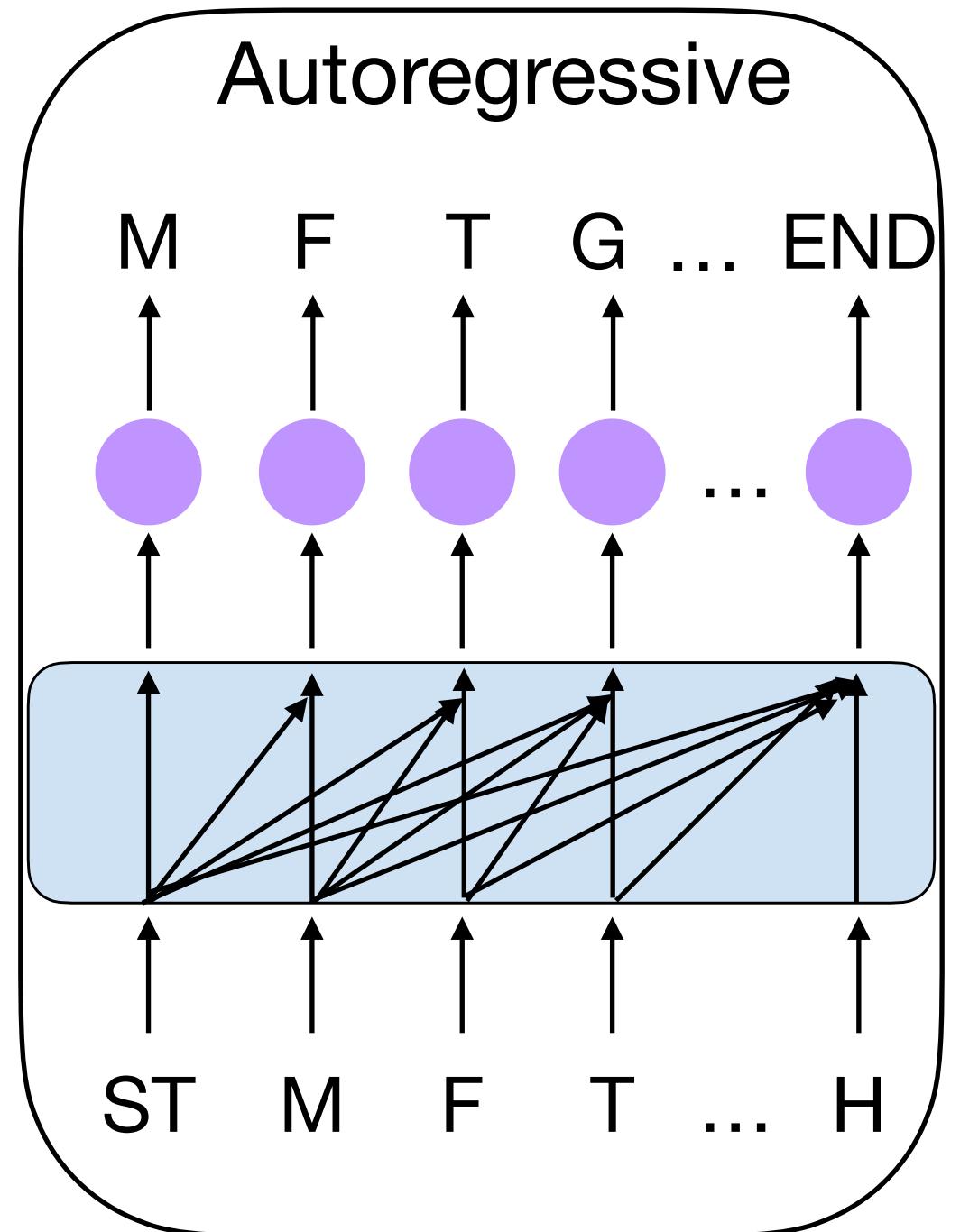
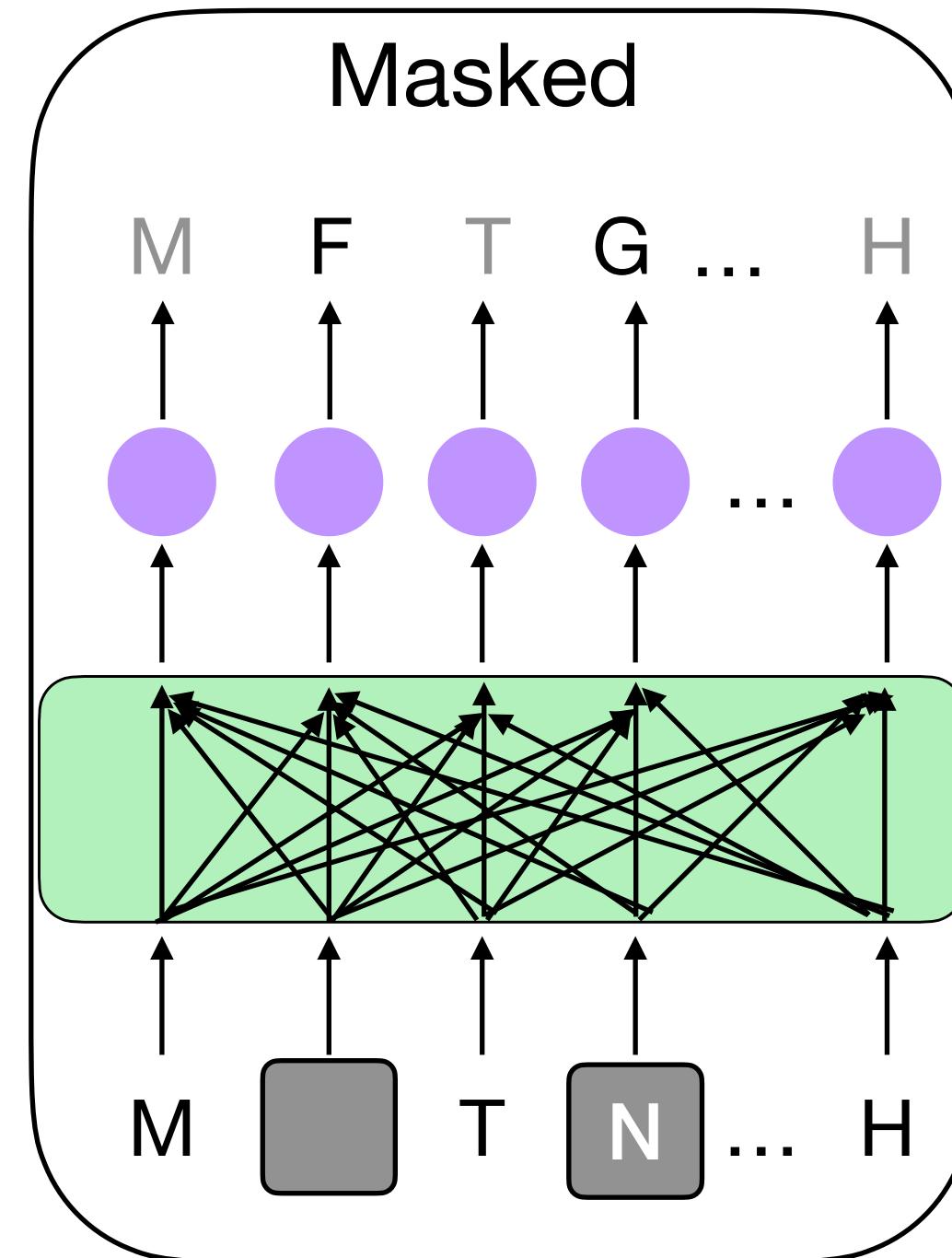
Separate pretraining task



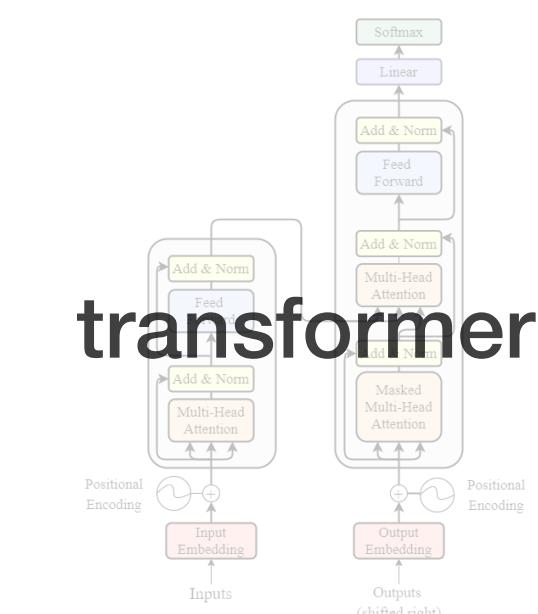
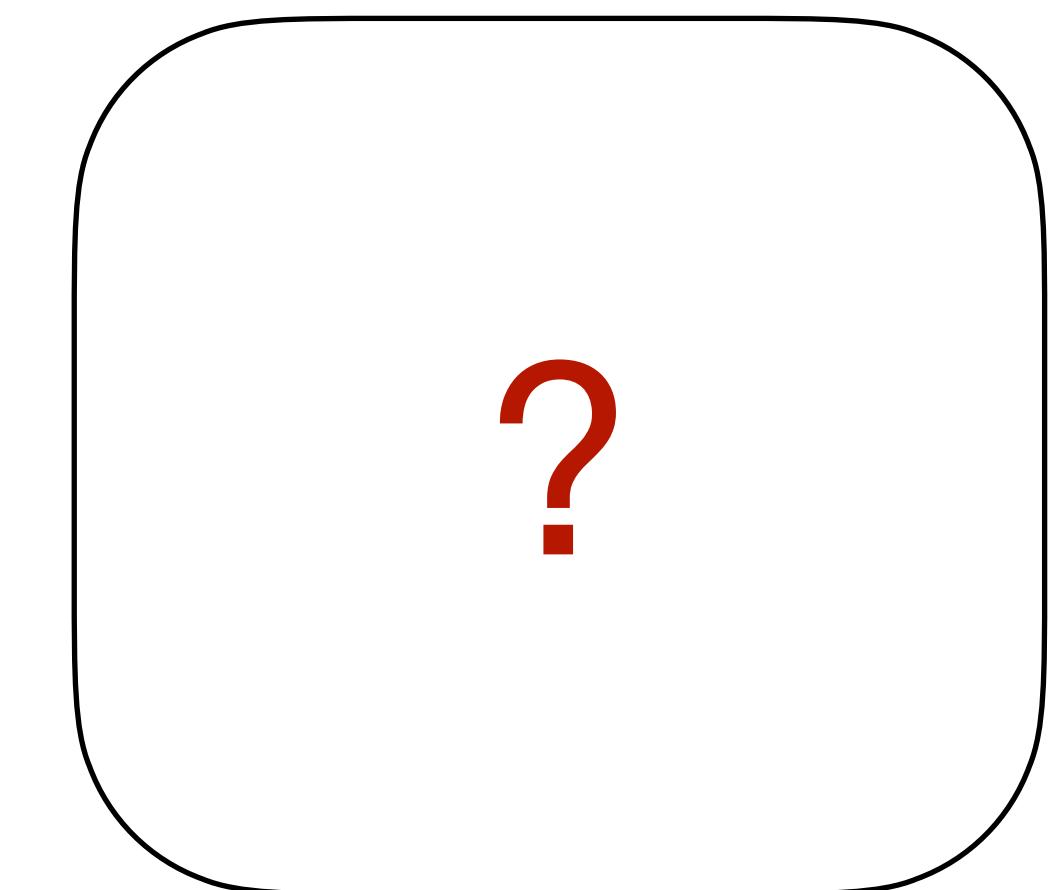
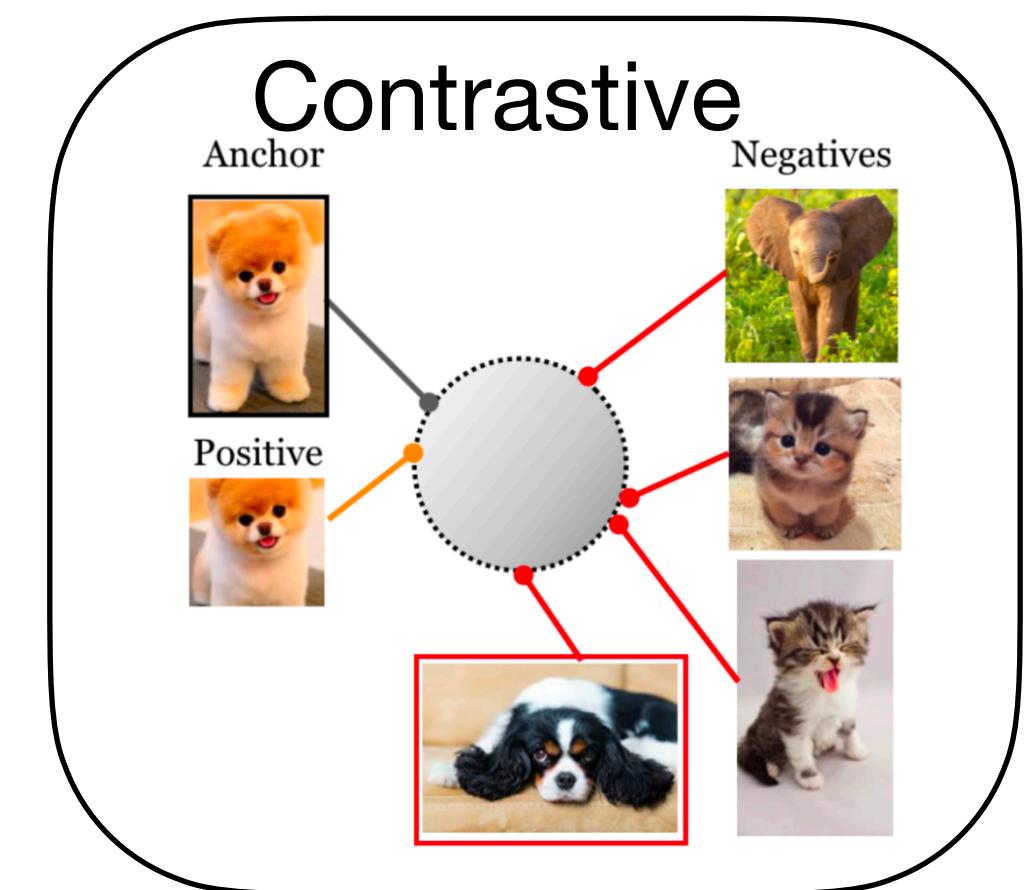
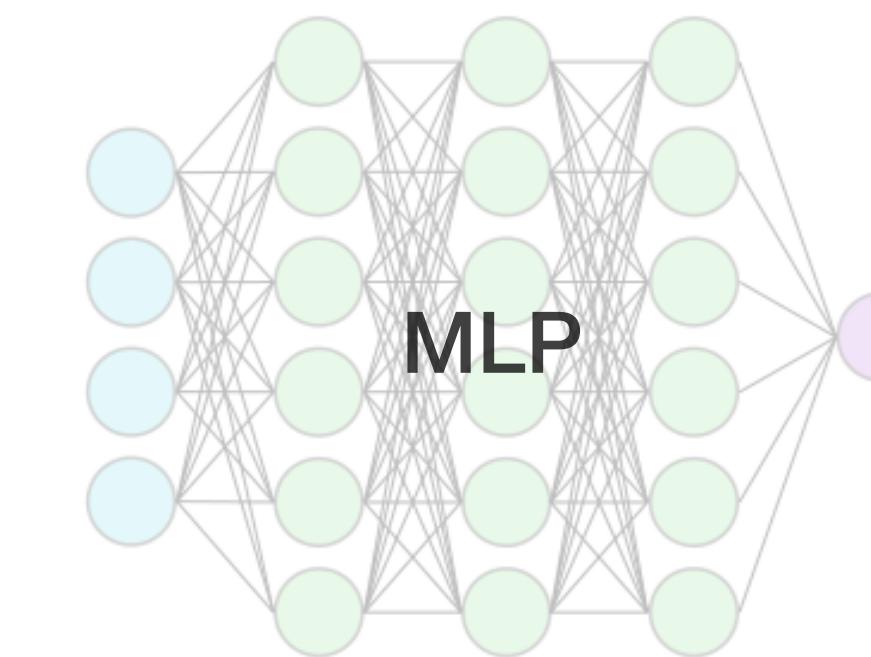
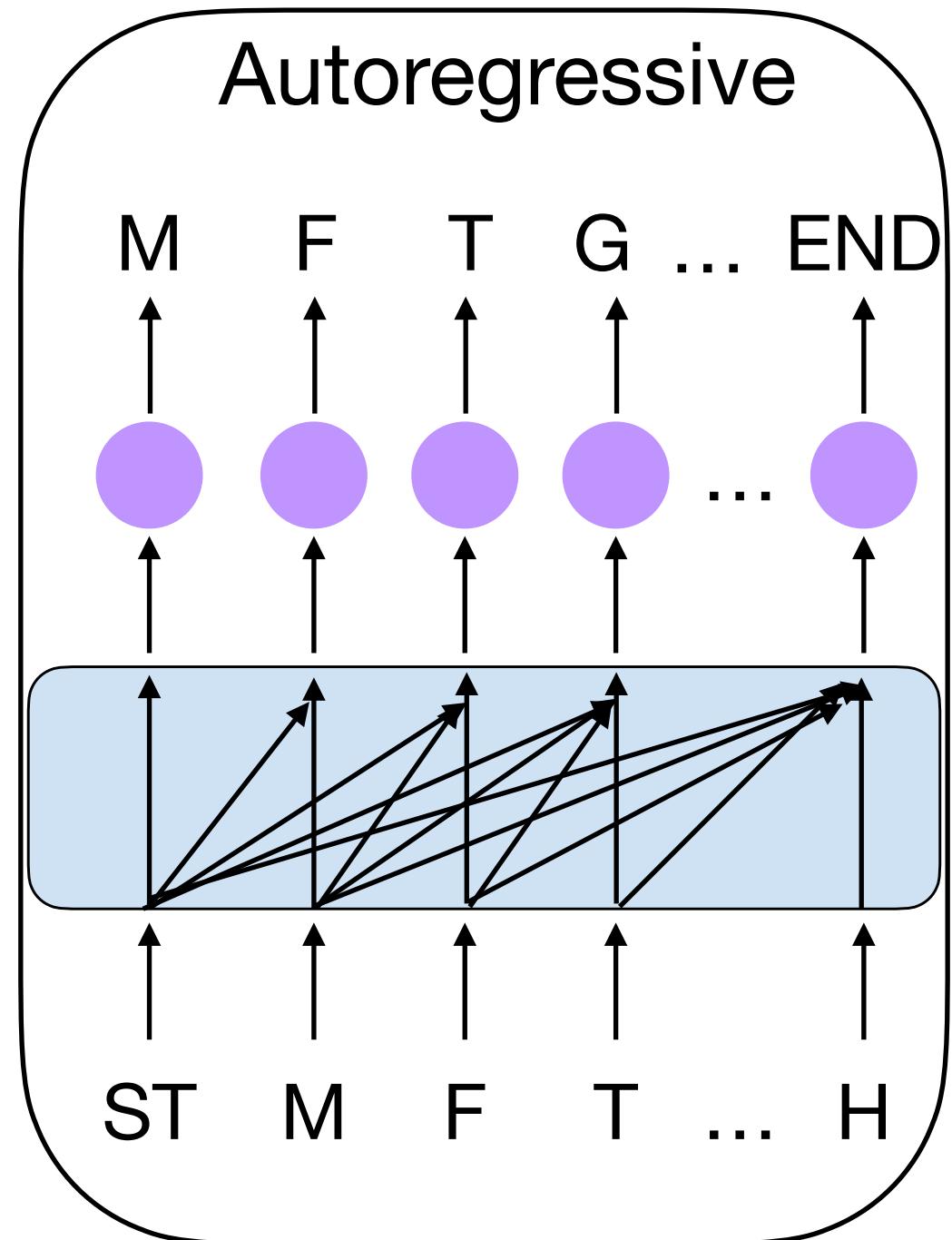
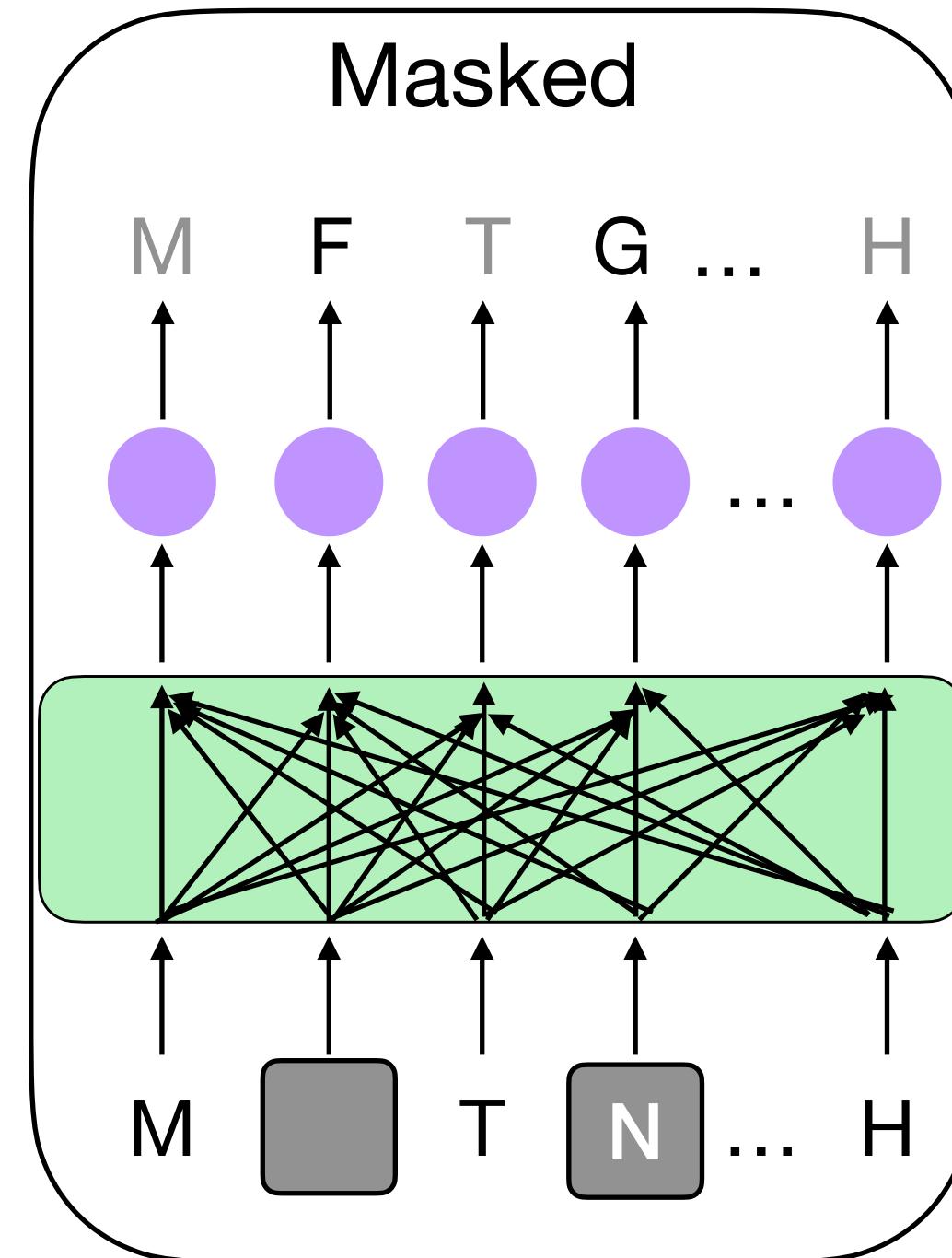
Separate pretraining task and architecture



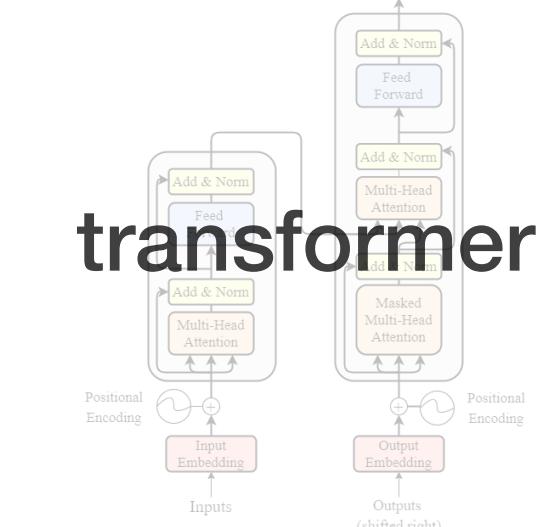
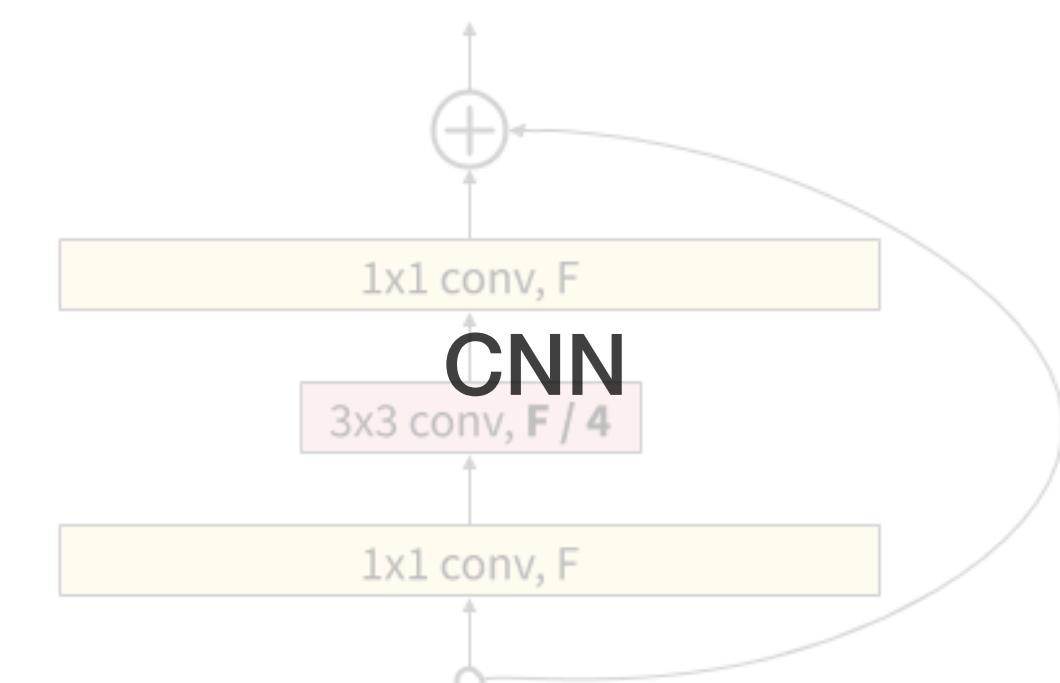
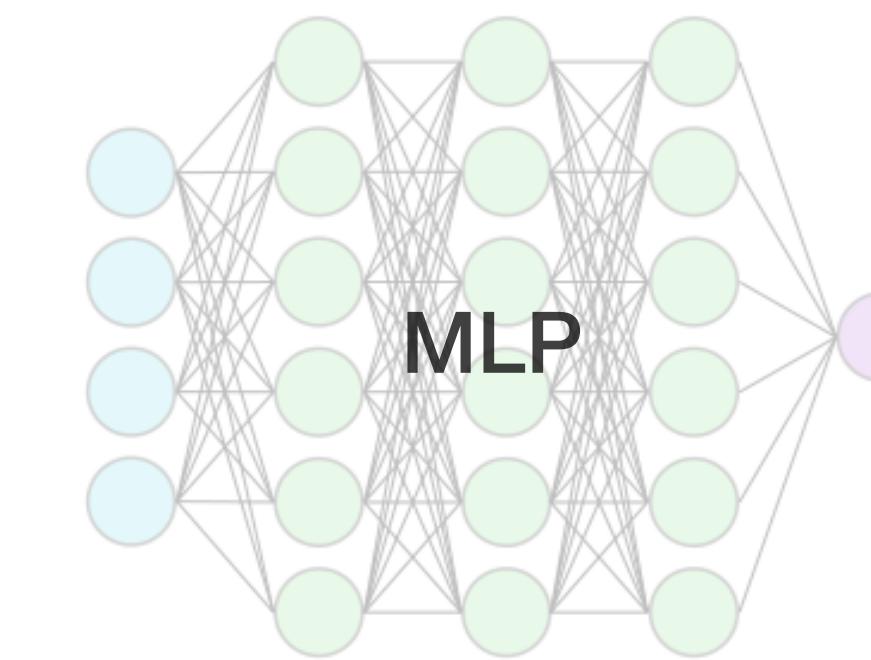
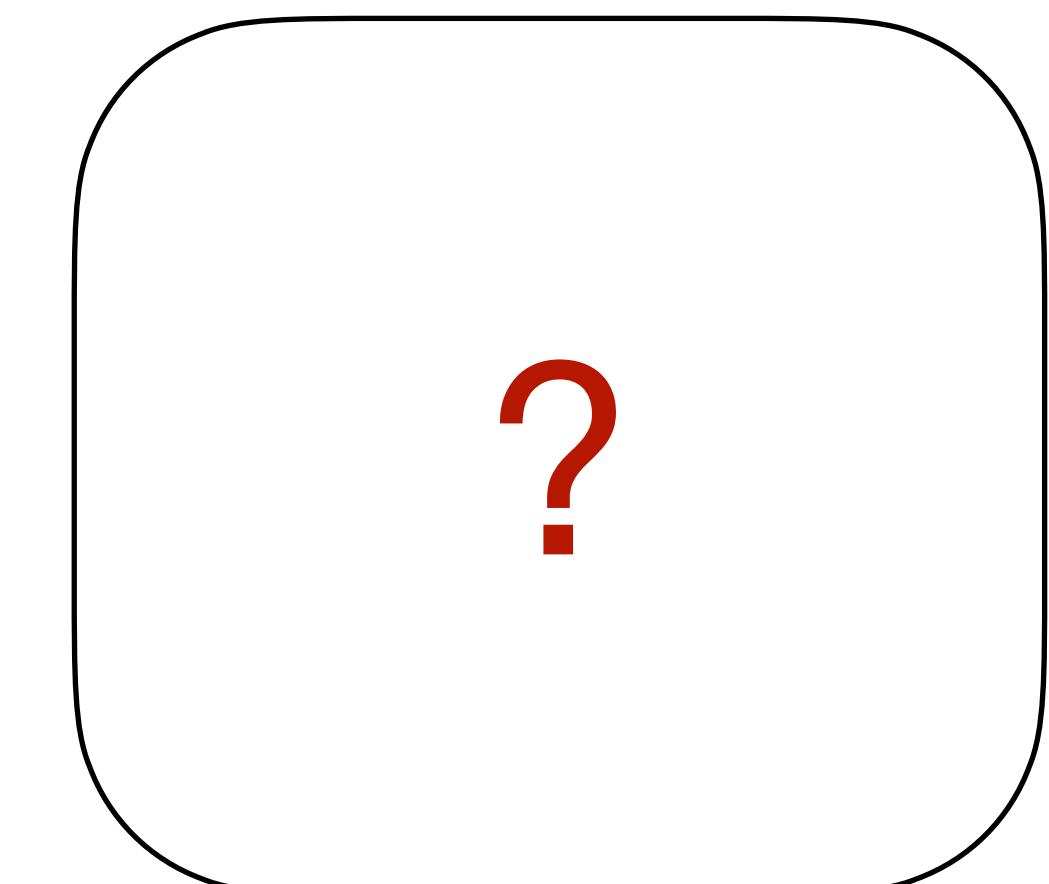
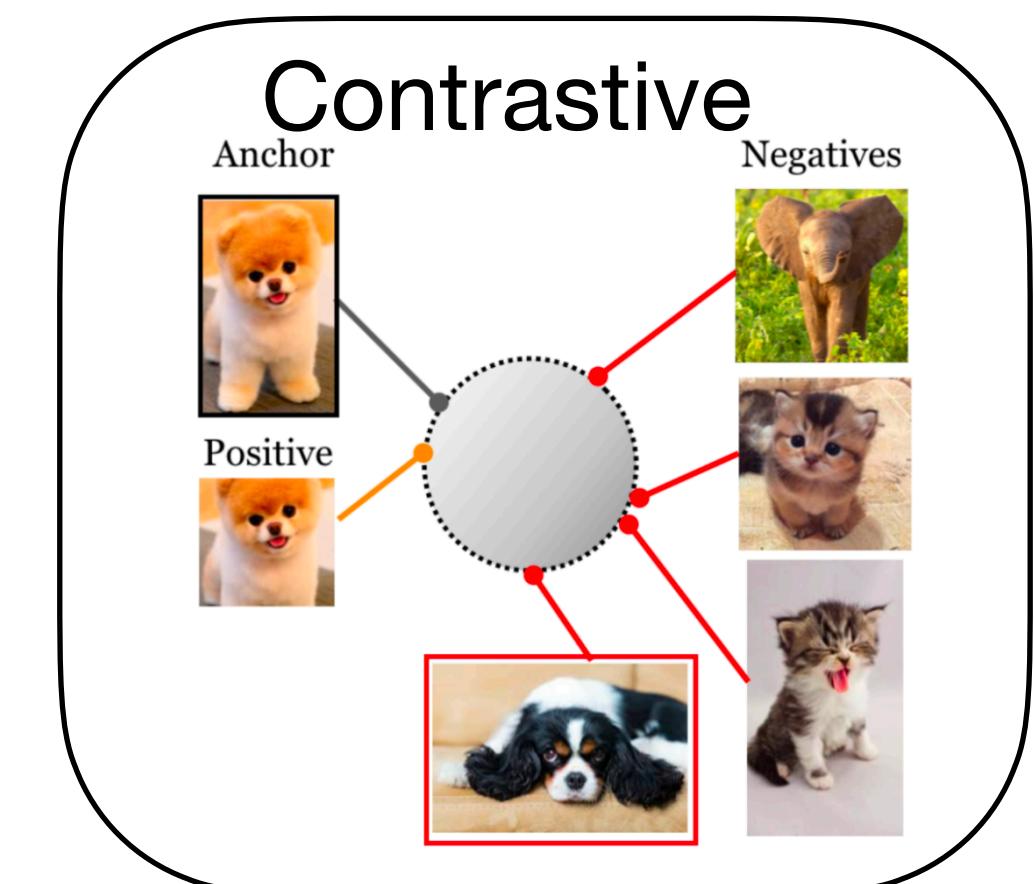
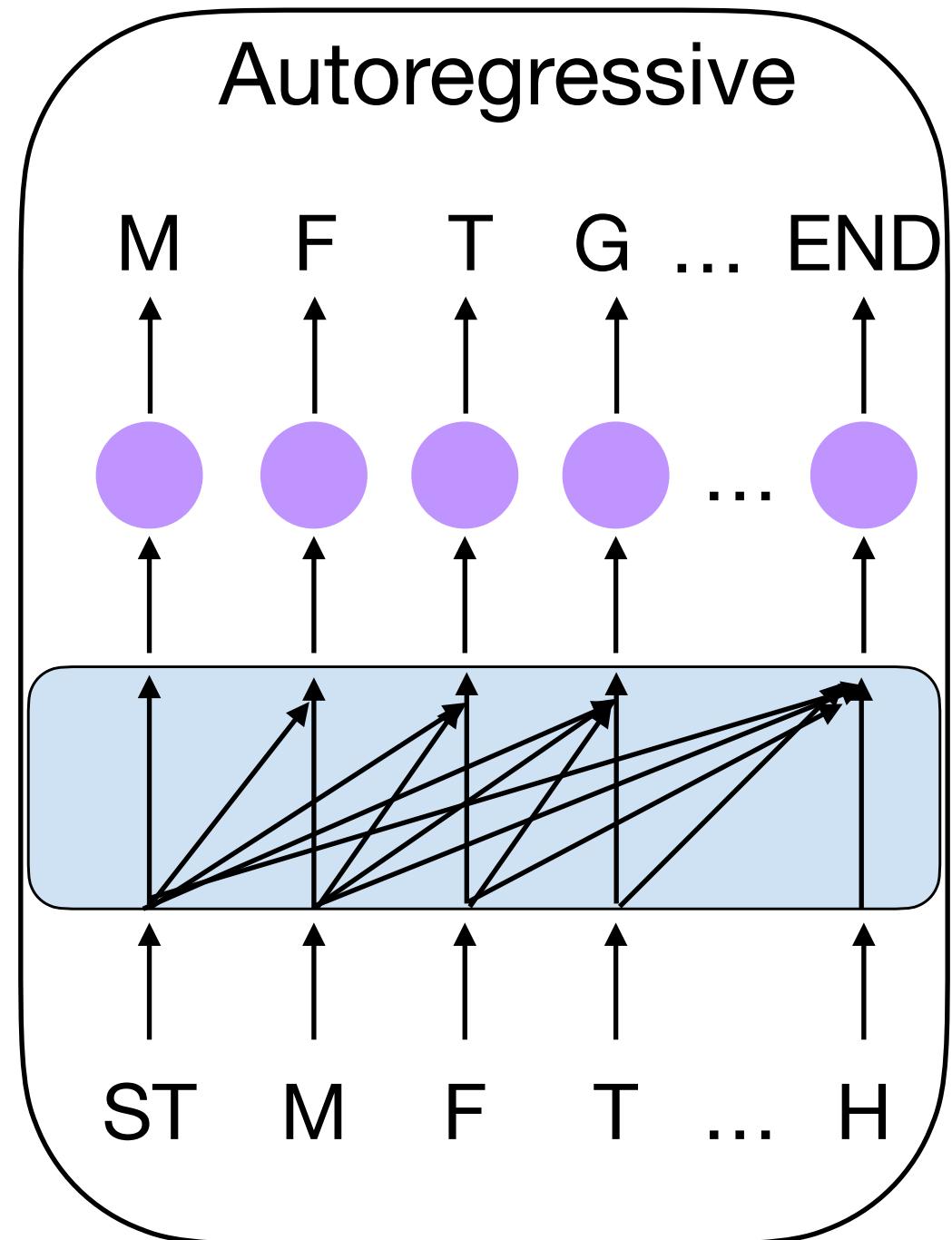
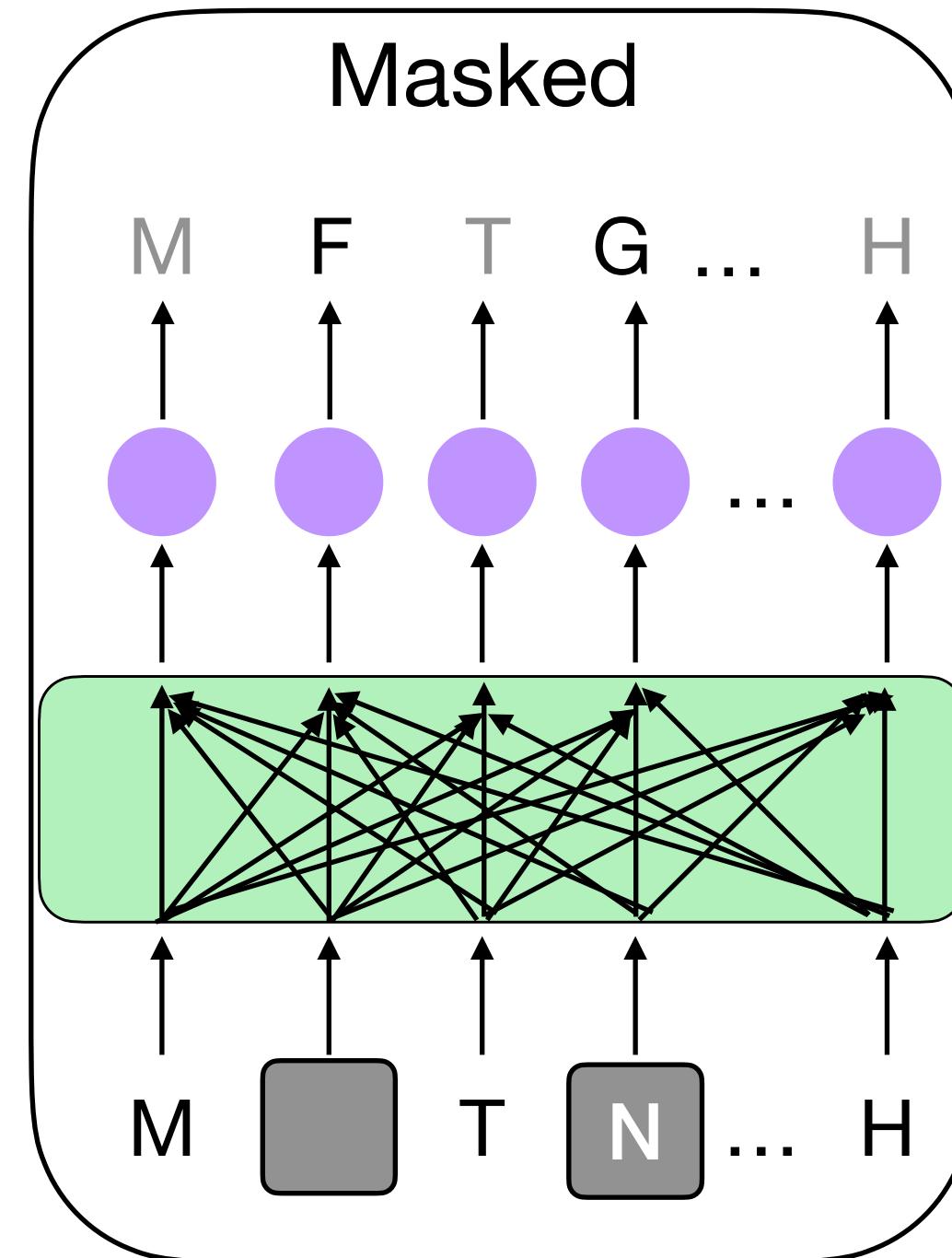
Separate pretraining task and architecture



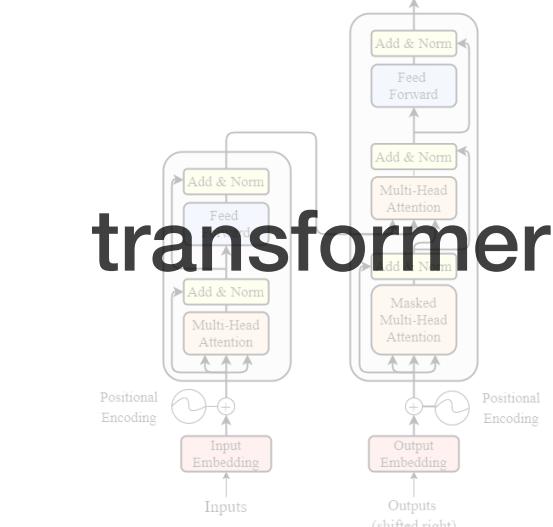
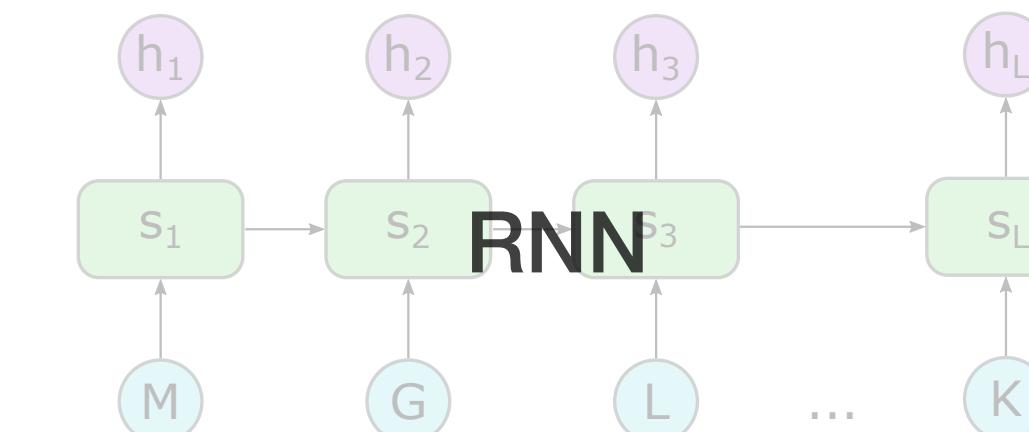
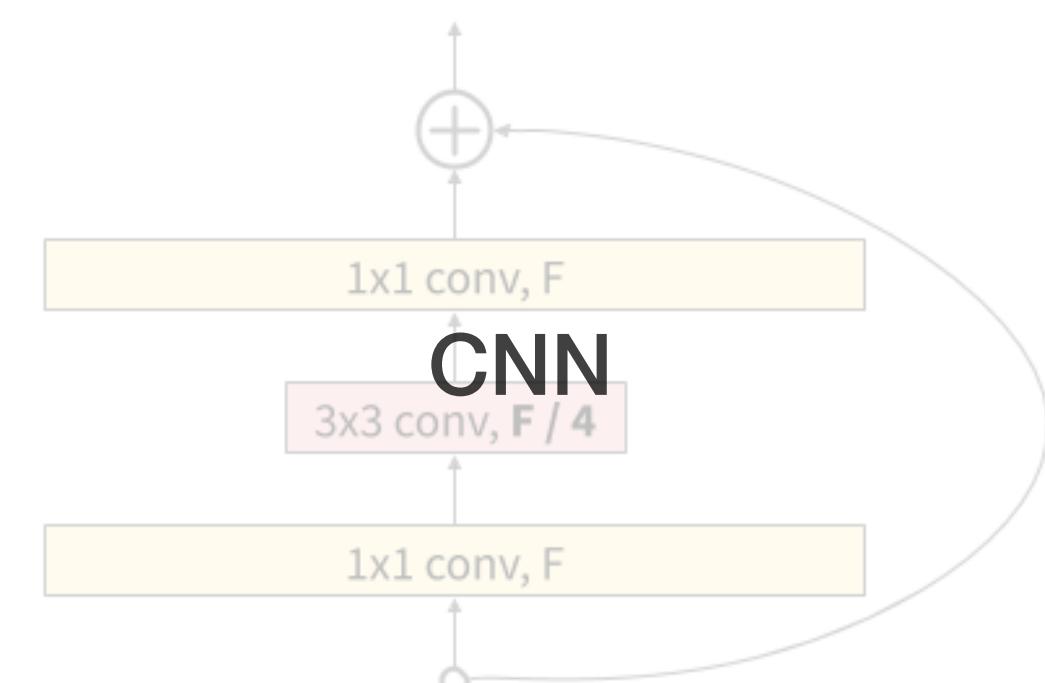
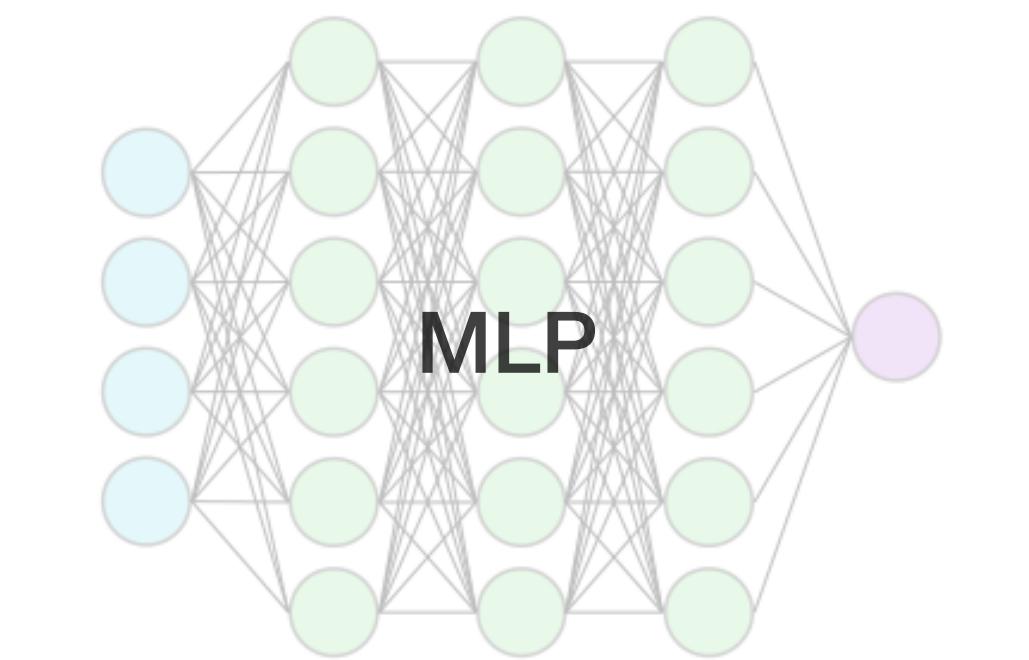
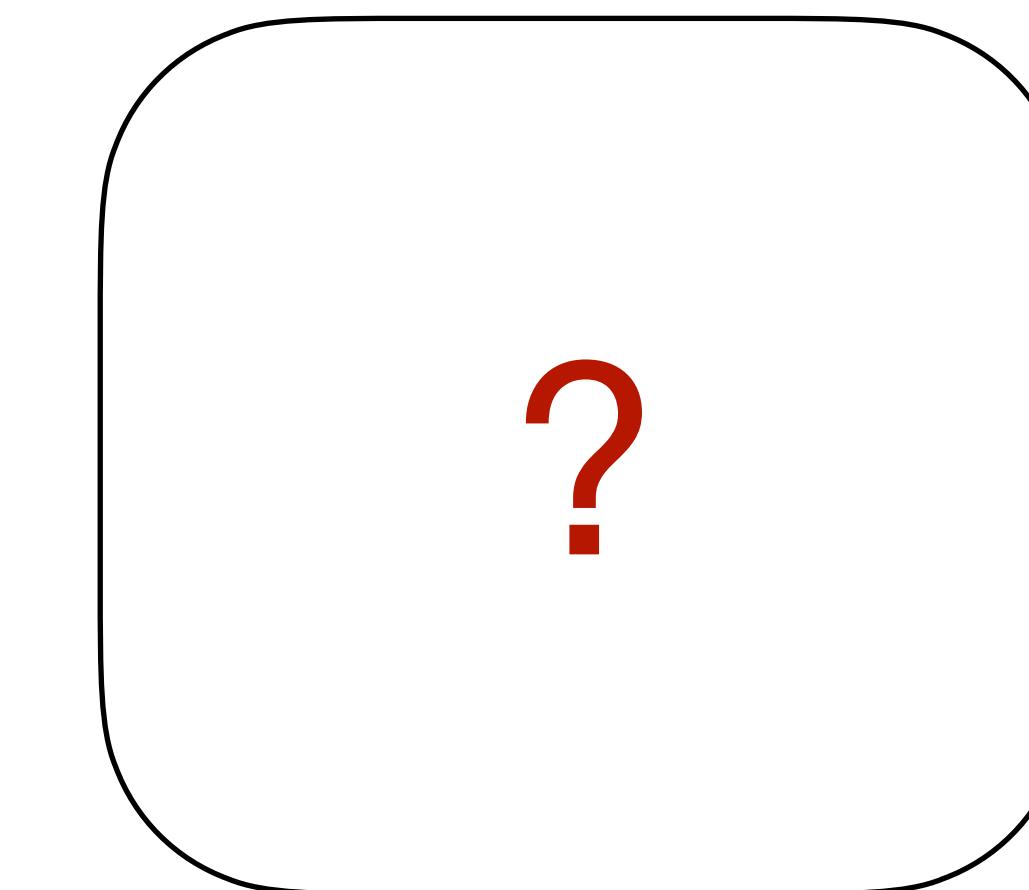
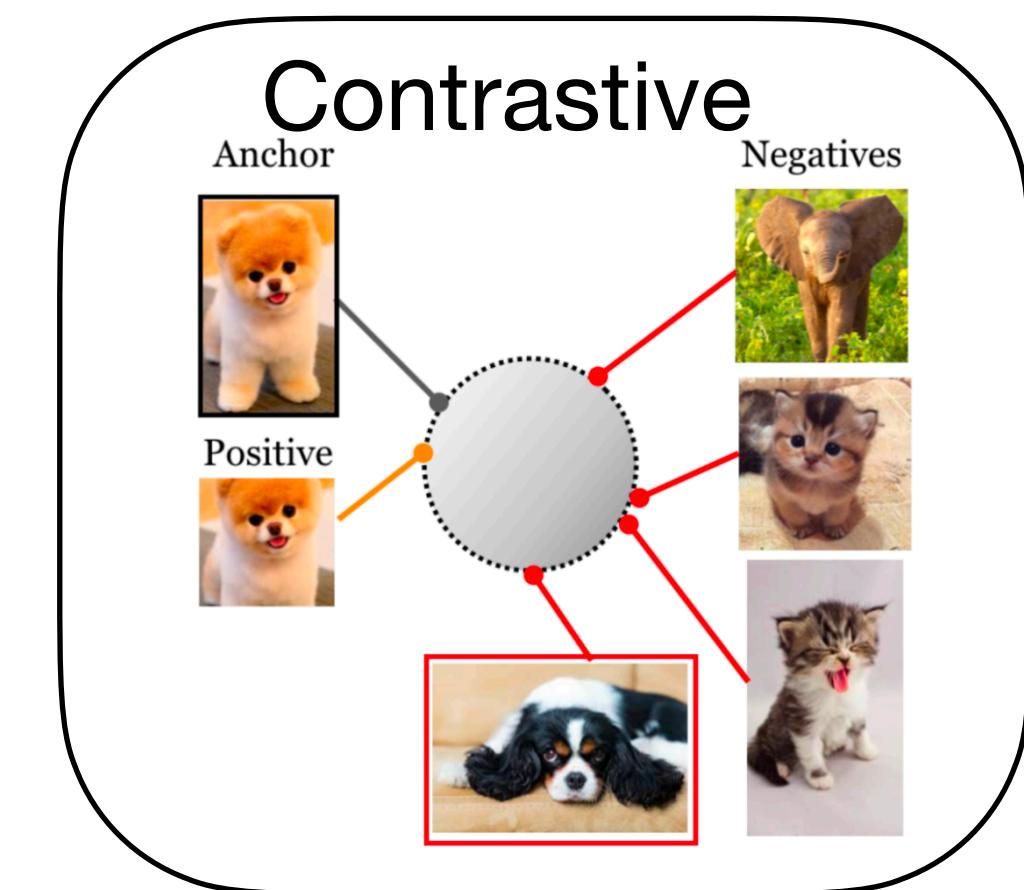
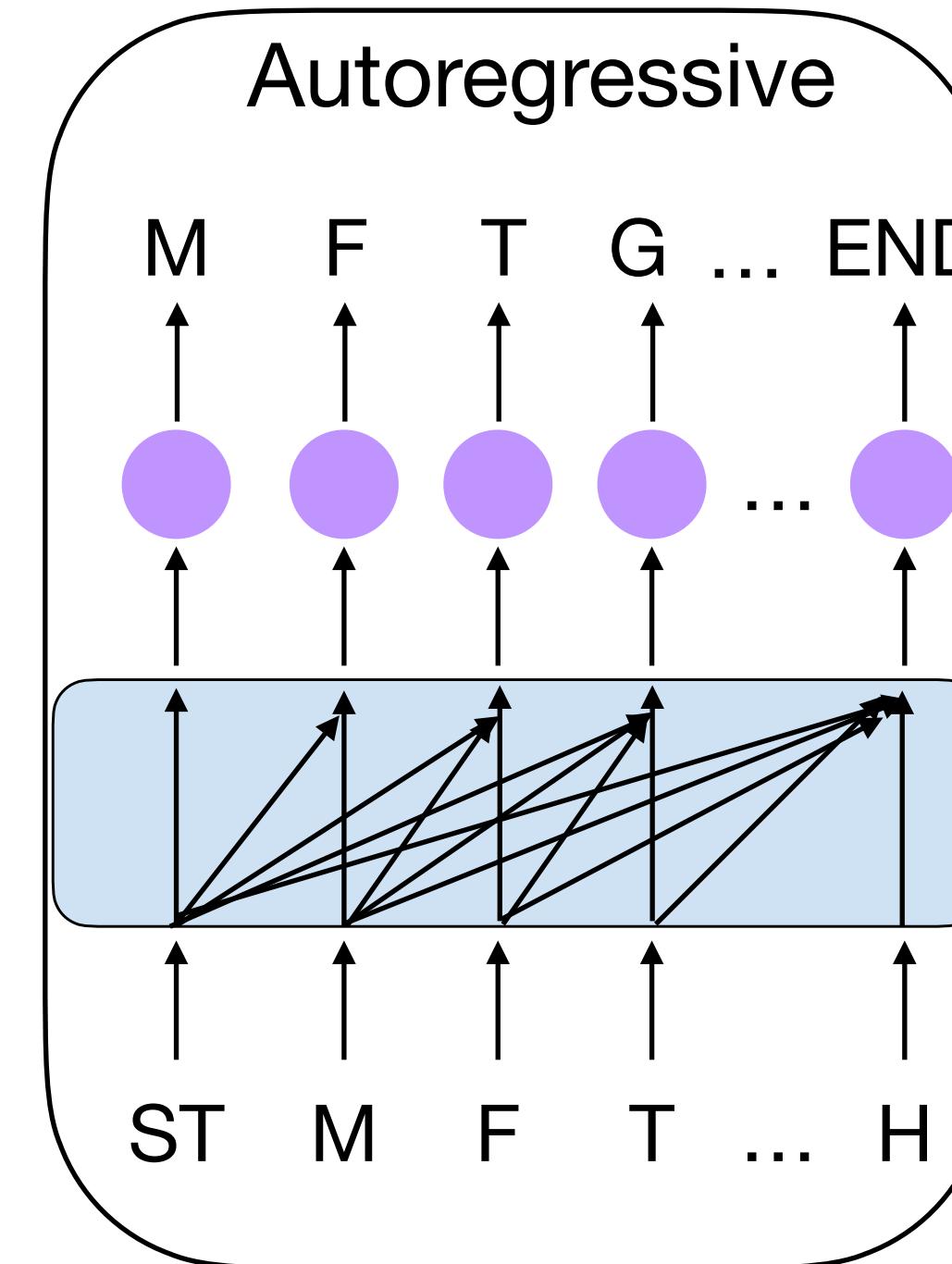
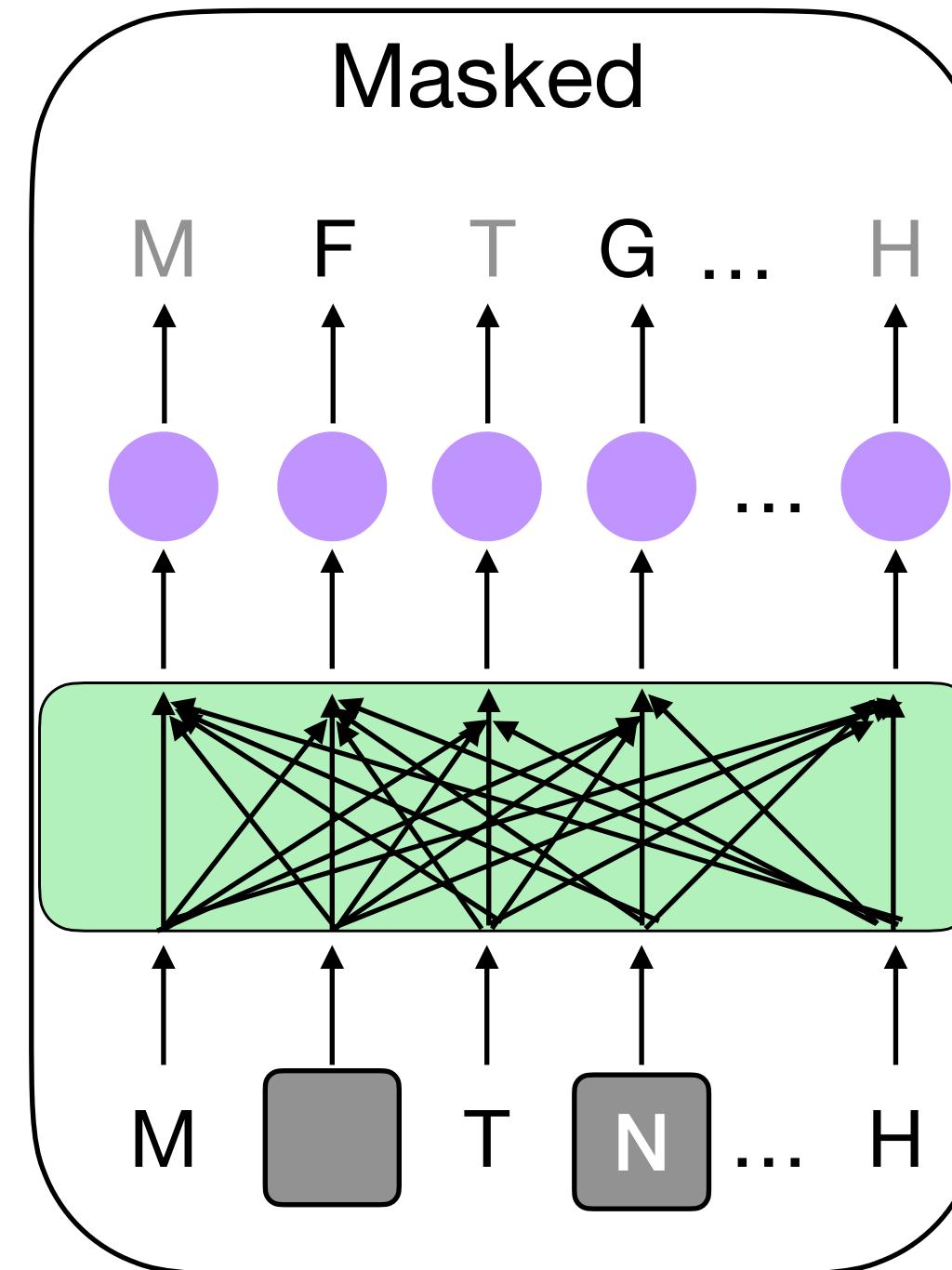
Separate pretraining task and architecture



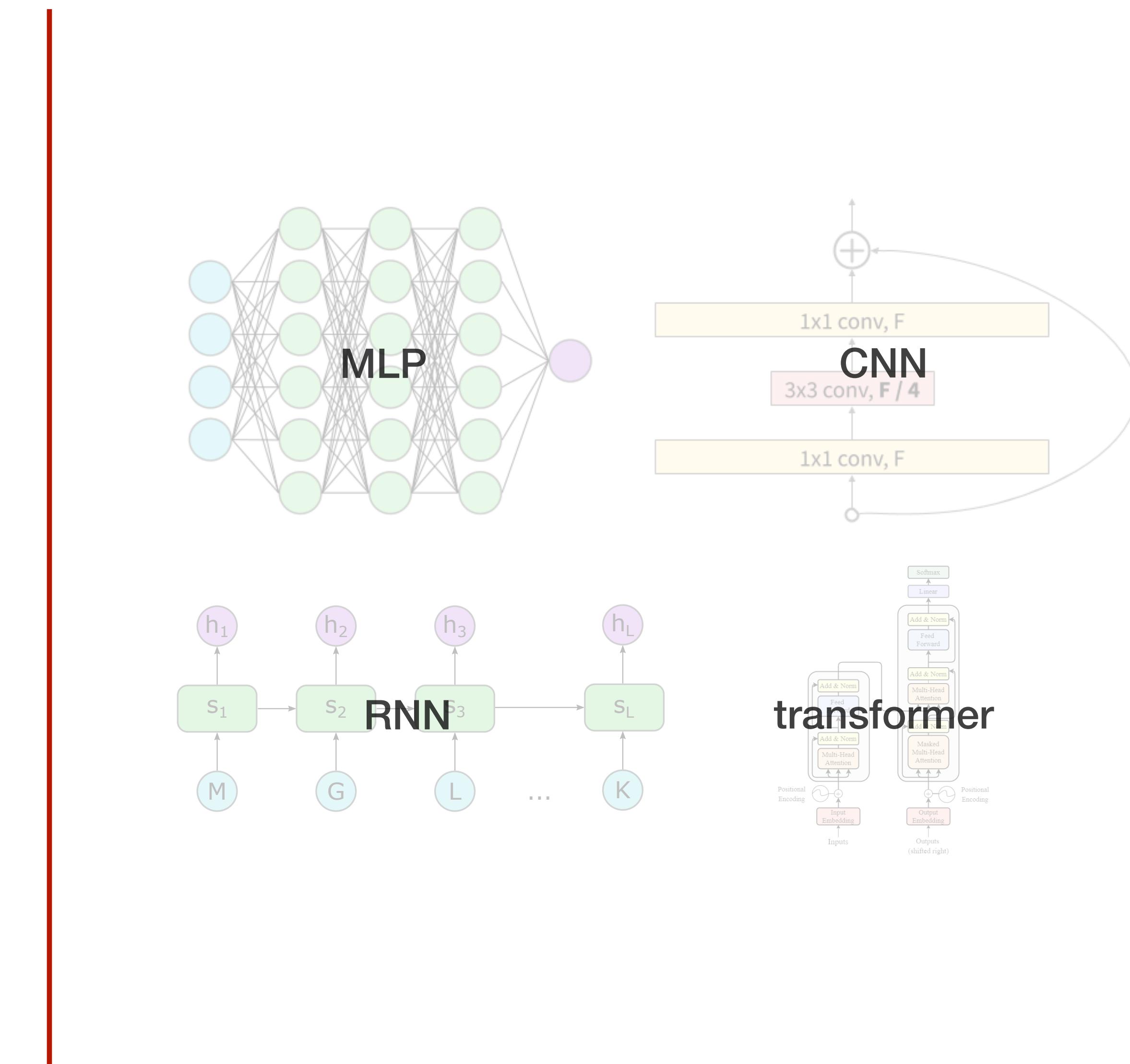
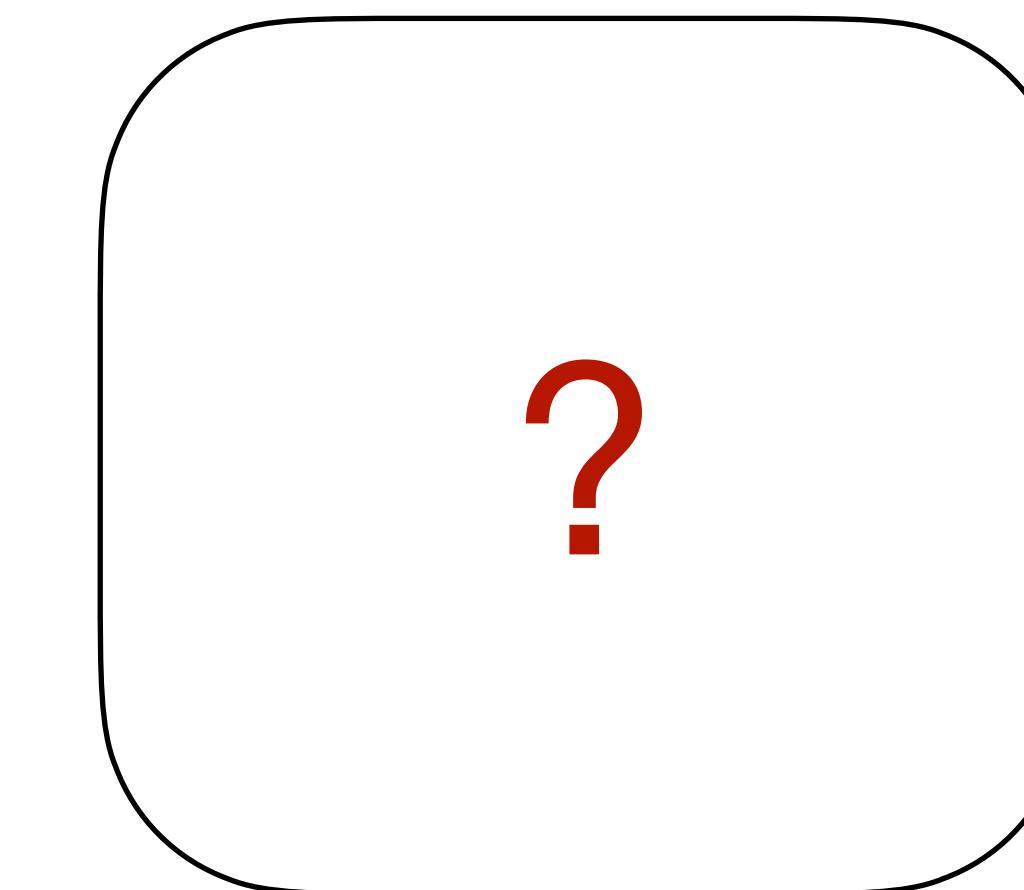
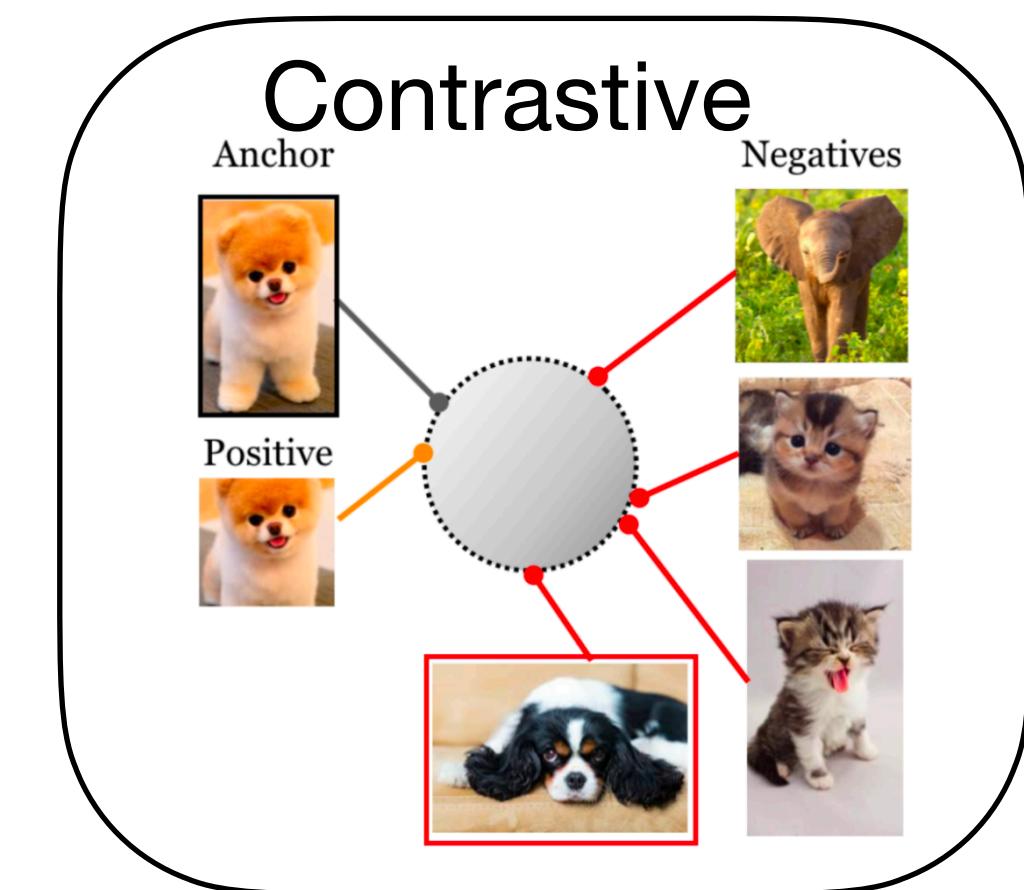
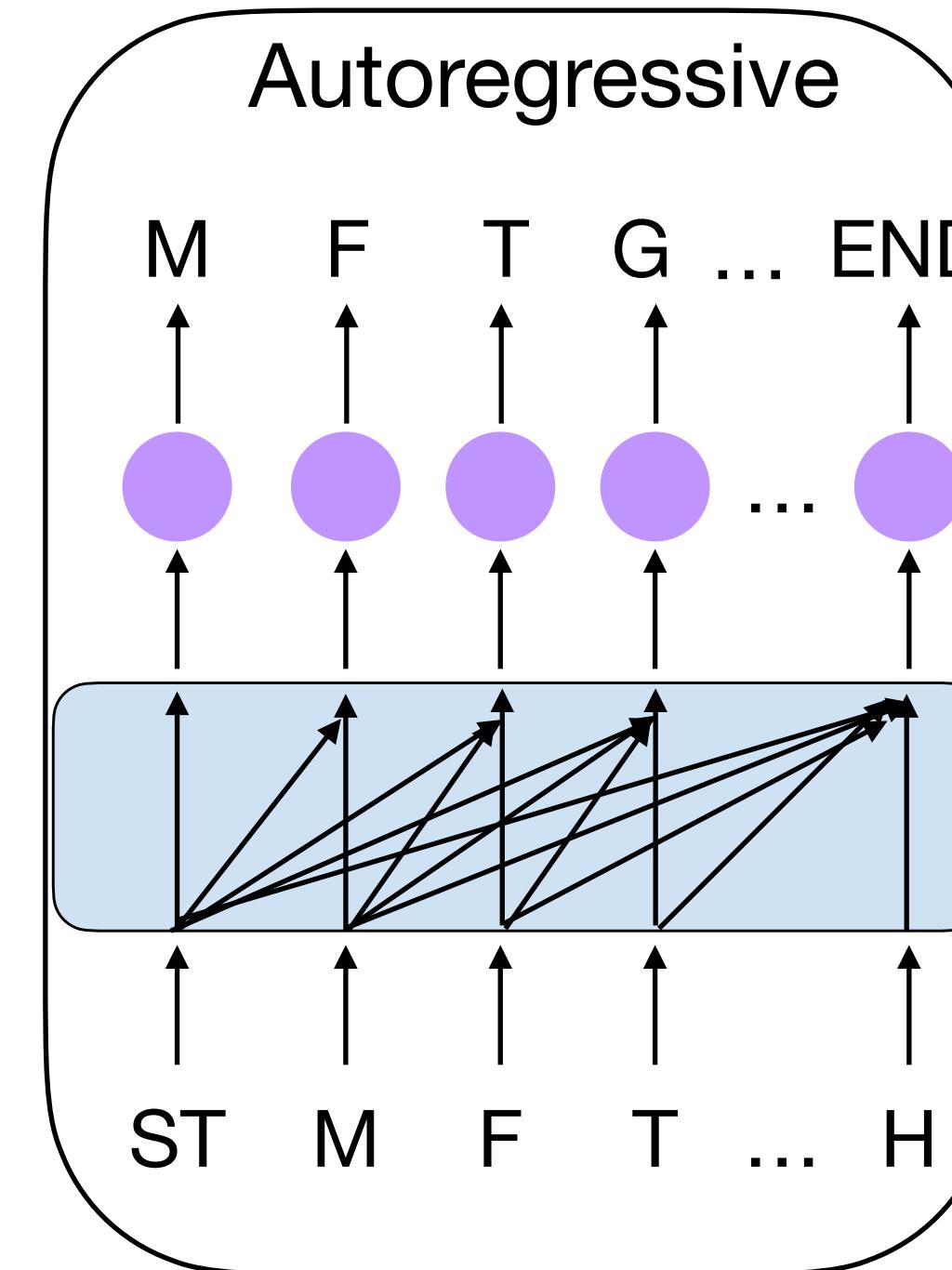
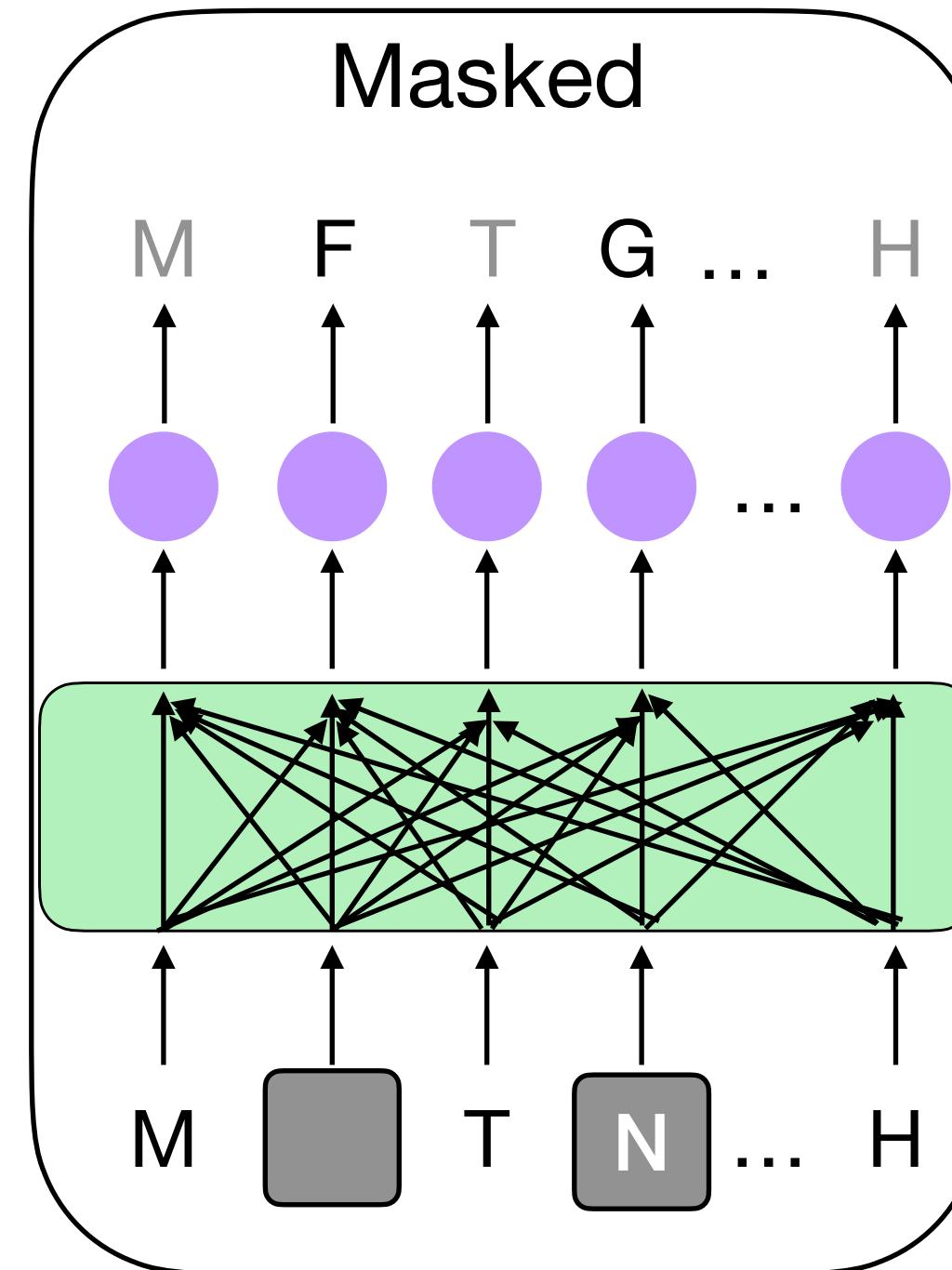
Separate pretraining task and architecture



Separate pretraining task and architecture



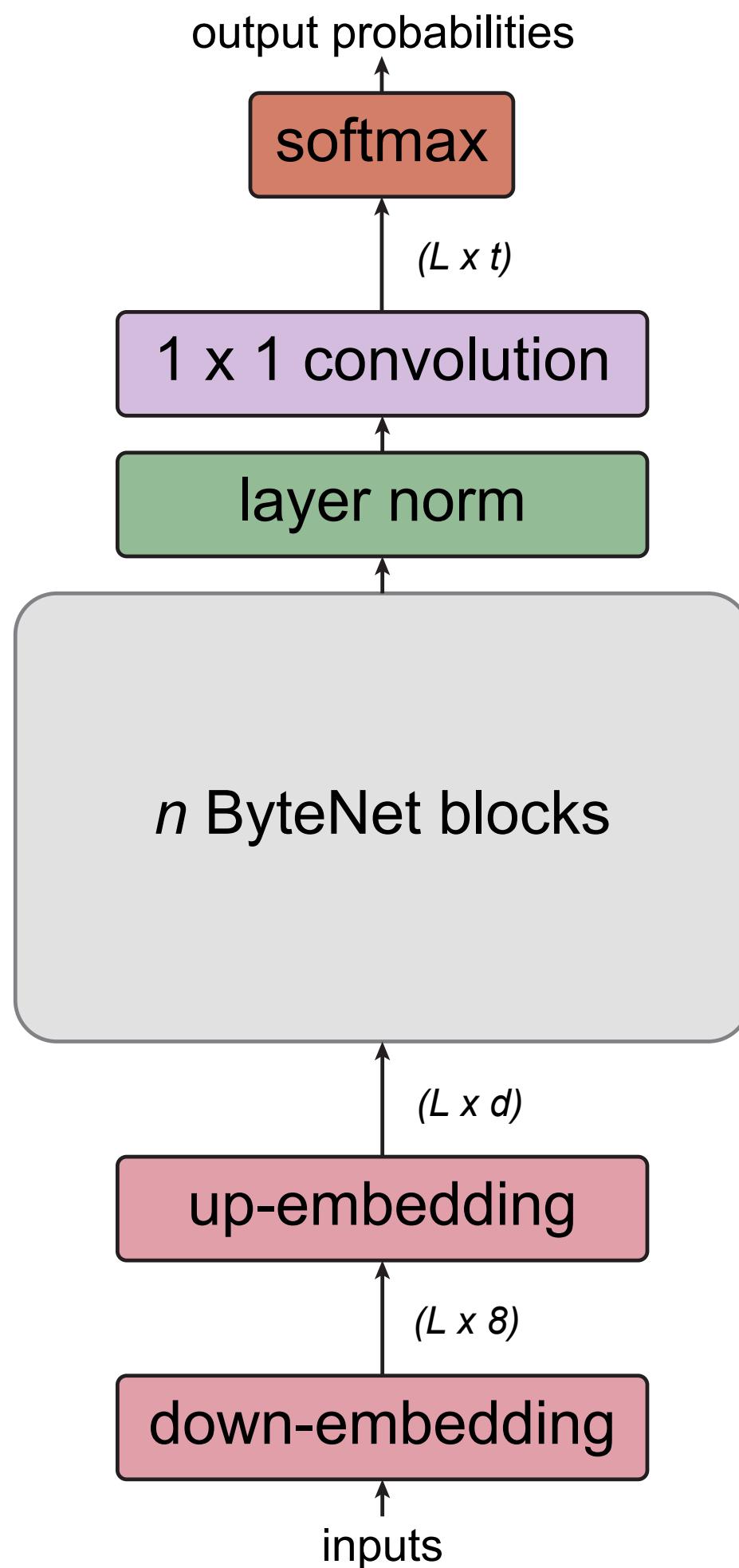
Separate pretraining task and architecture



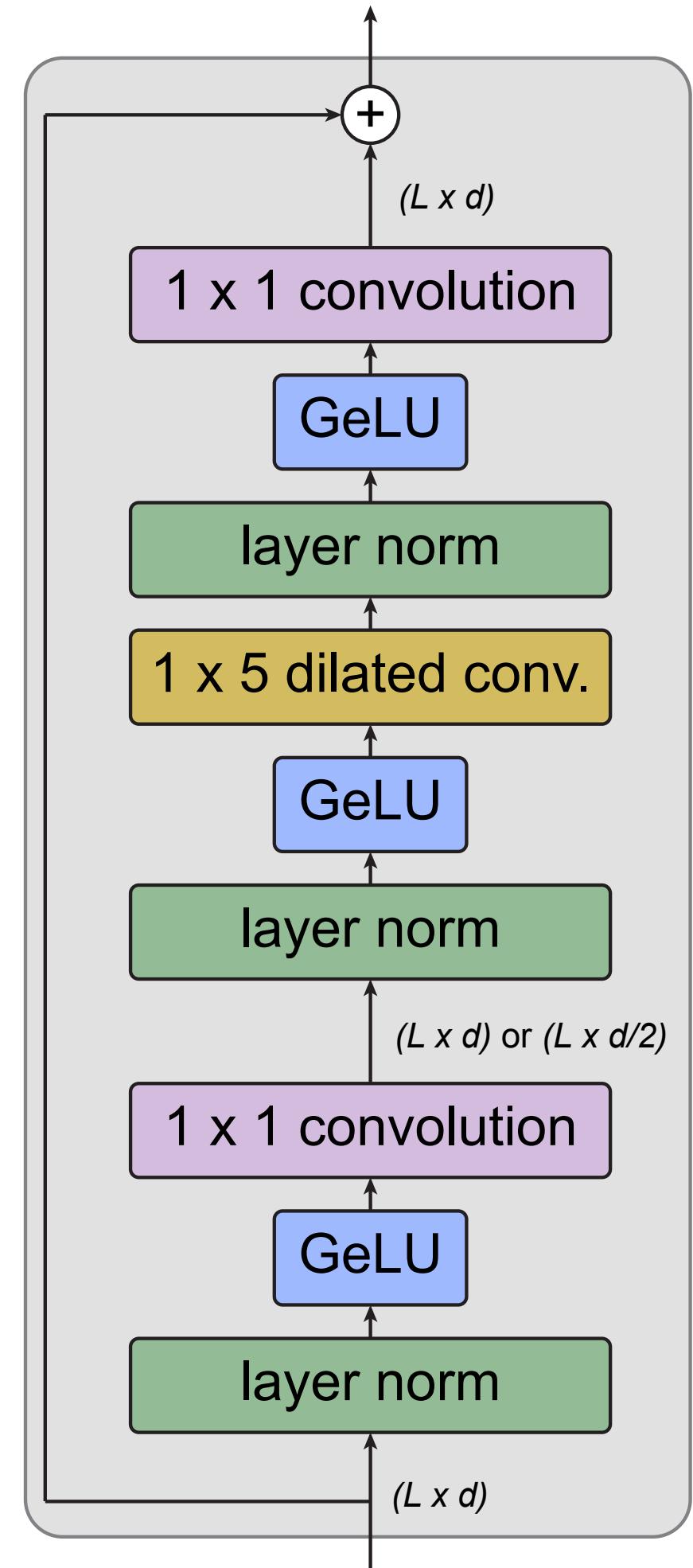
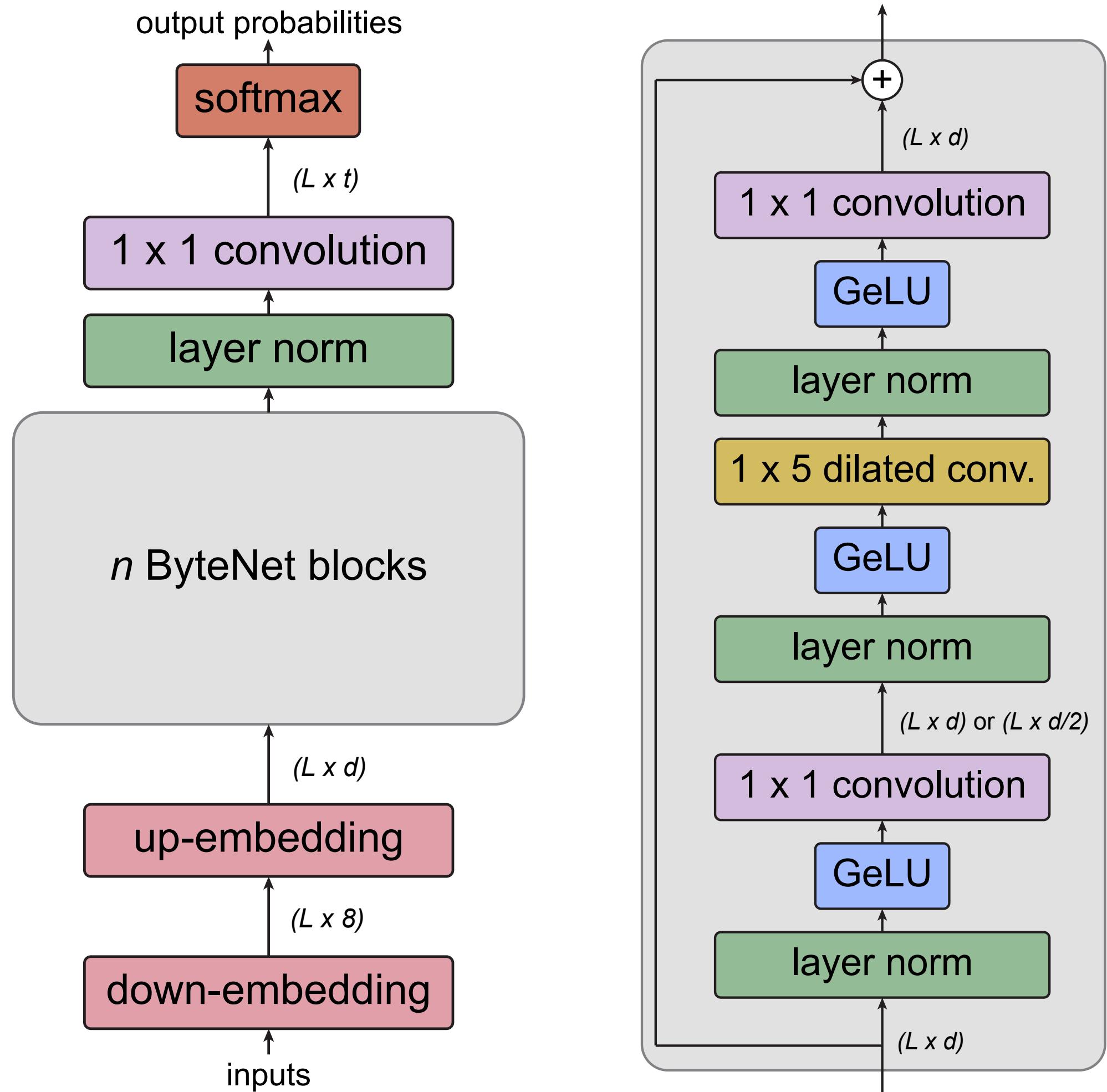
We pretrain CNNs to reconstruct sequences



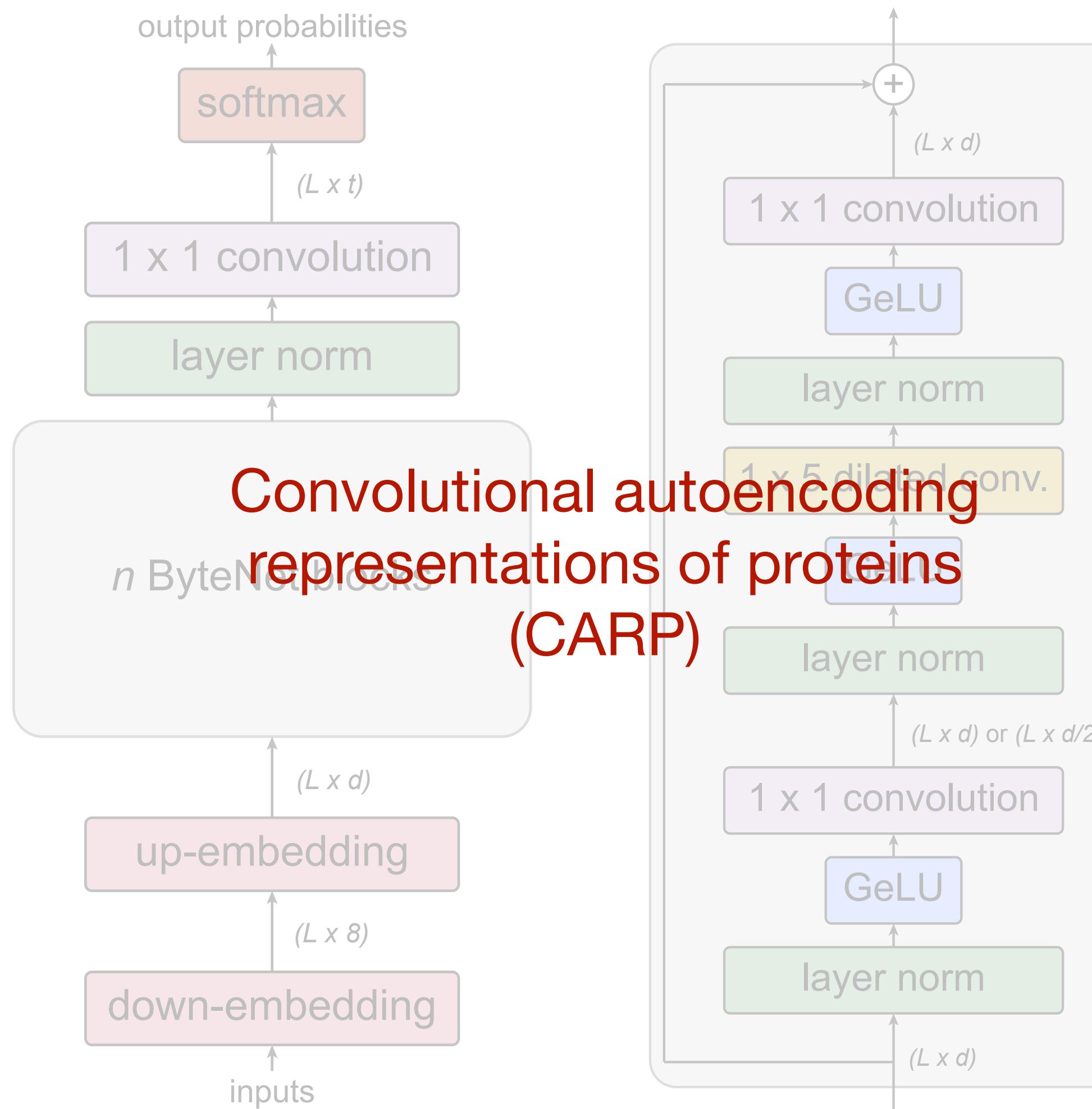
We pretrain CNNs to reconstruct sequences



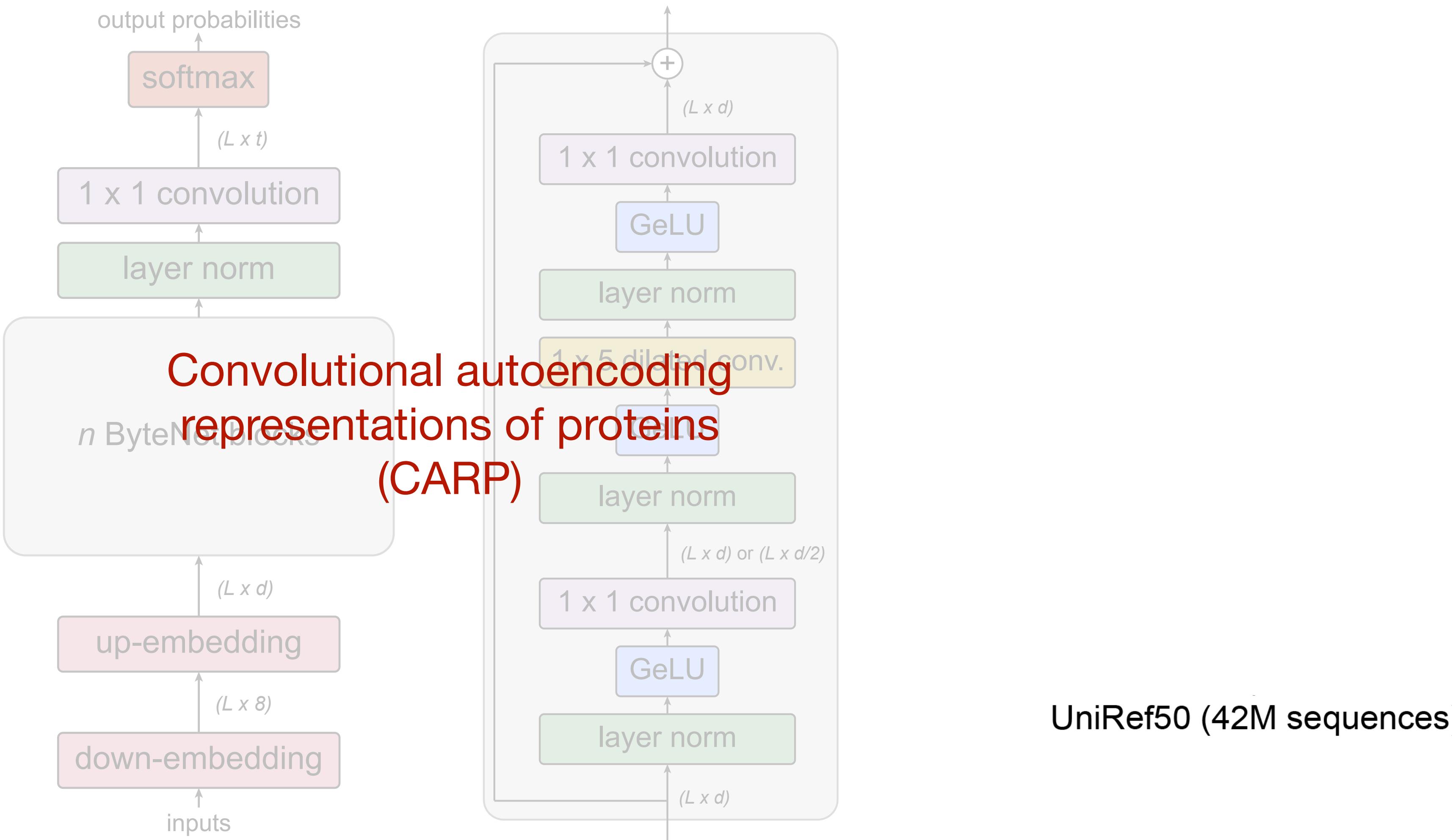
We pretrain CNNs to reconstruct sequences



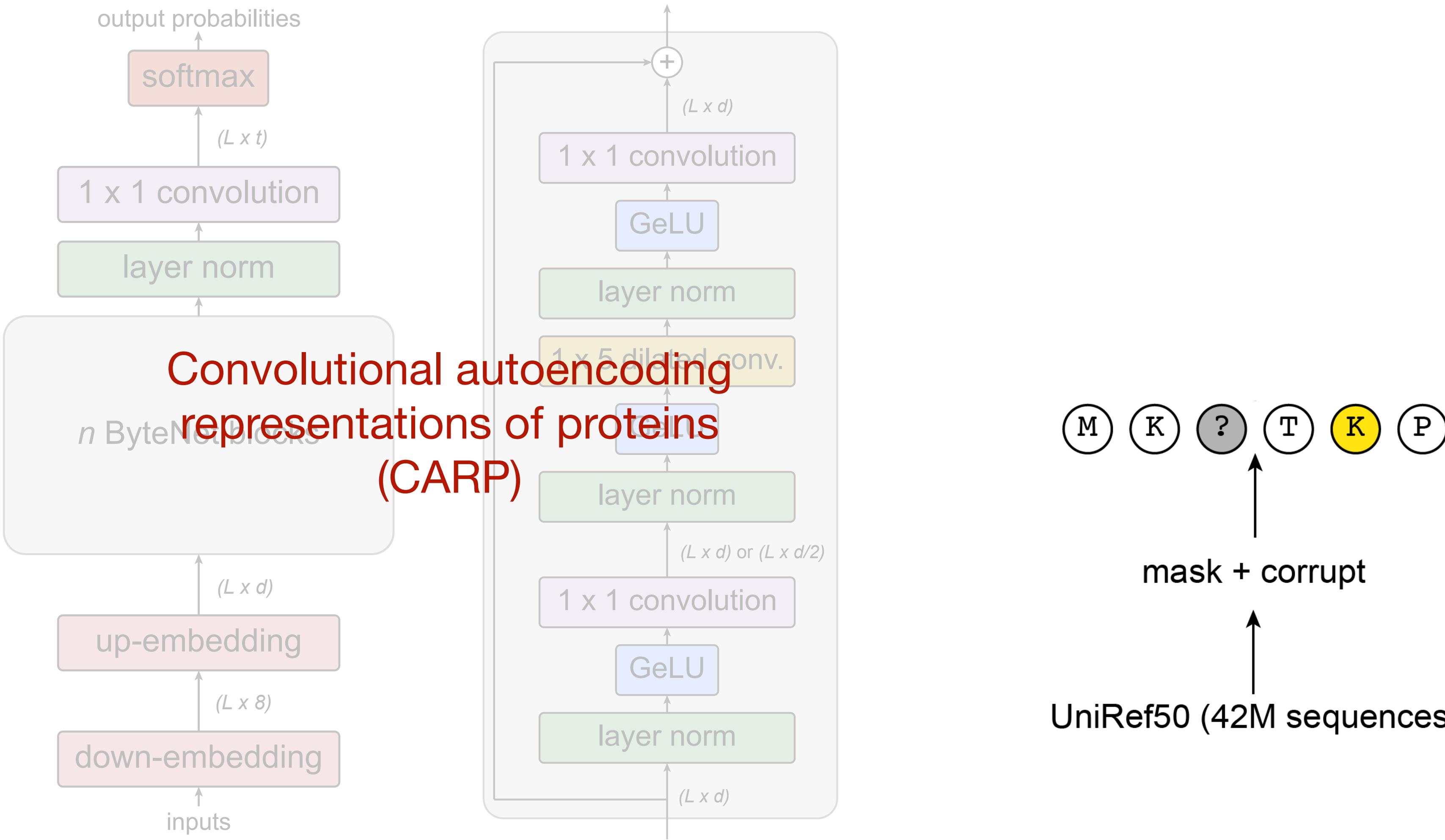
We pretrain CNNs to reconstruct sequences



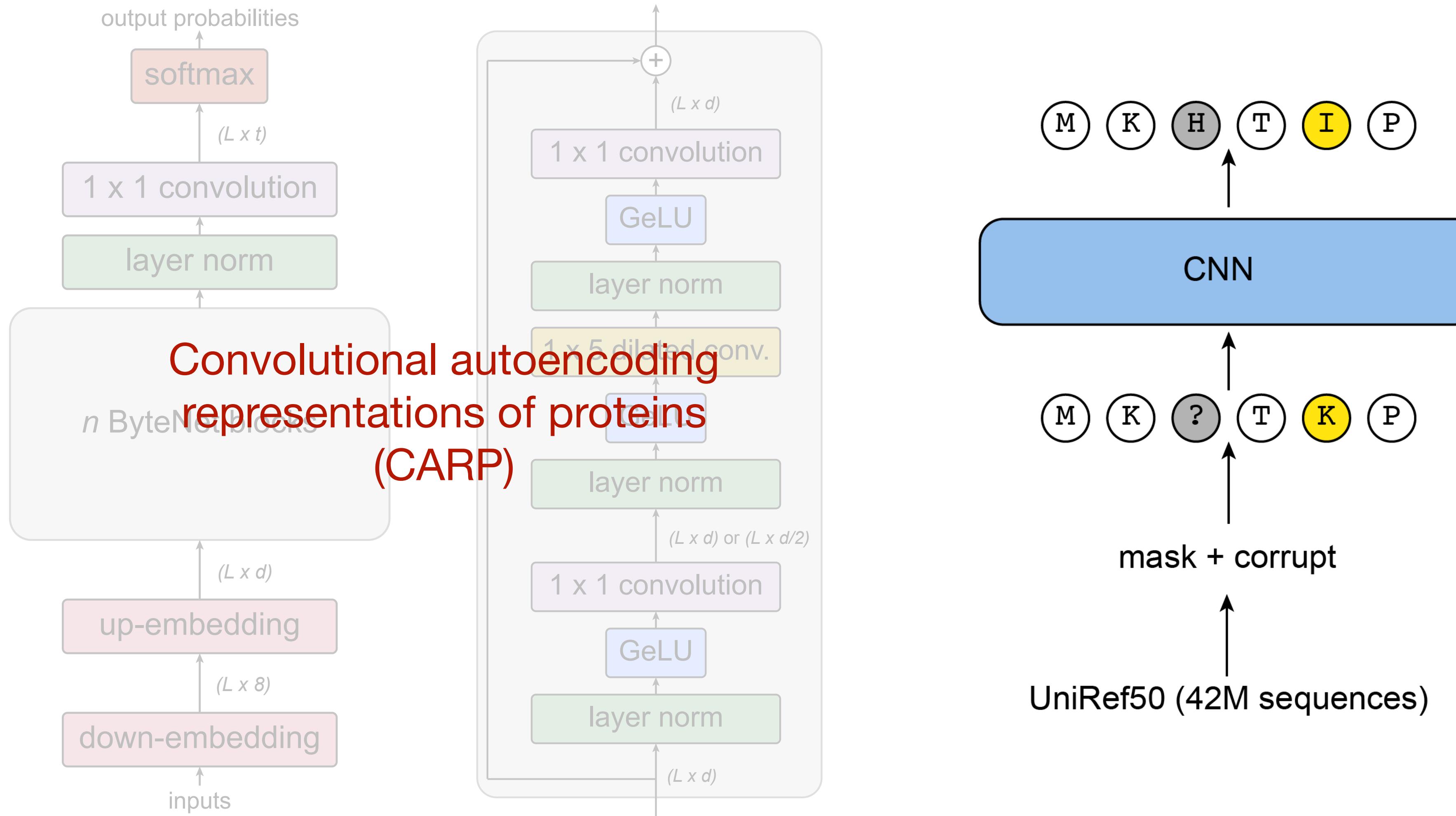
We pretrain CNNs to reconstruct sequences



We pretrain CNNs to reconstruct sequences

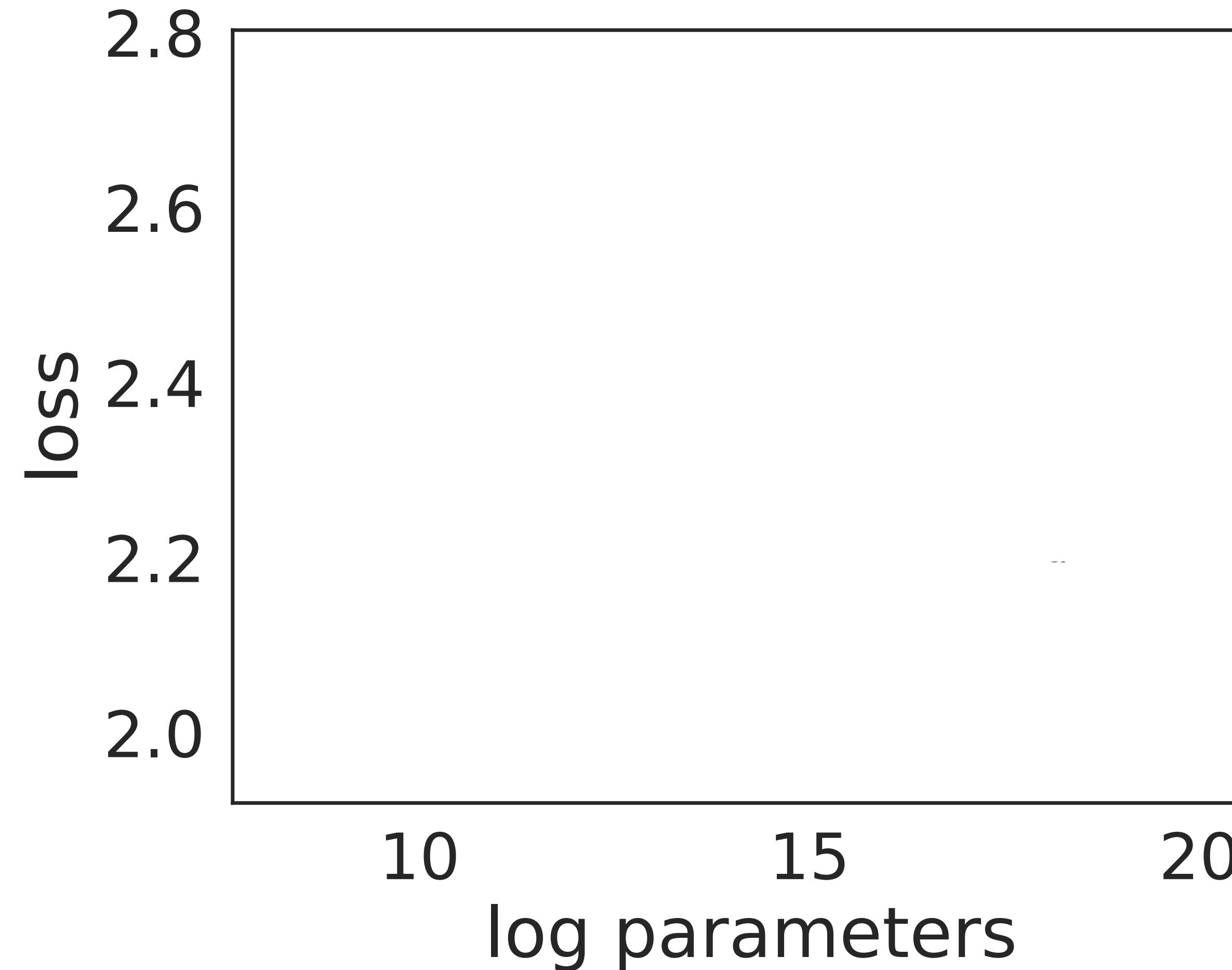


We pretrain CNNs to reconstruct sequences

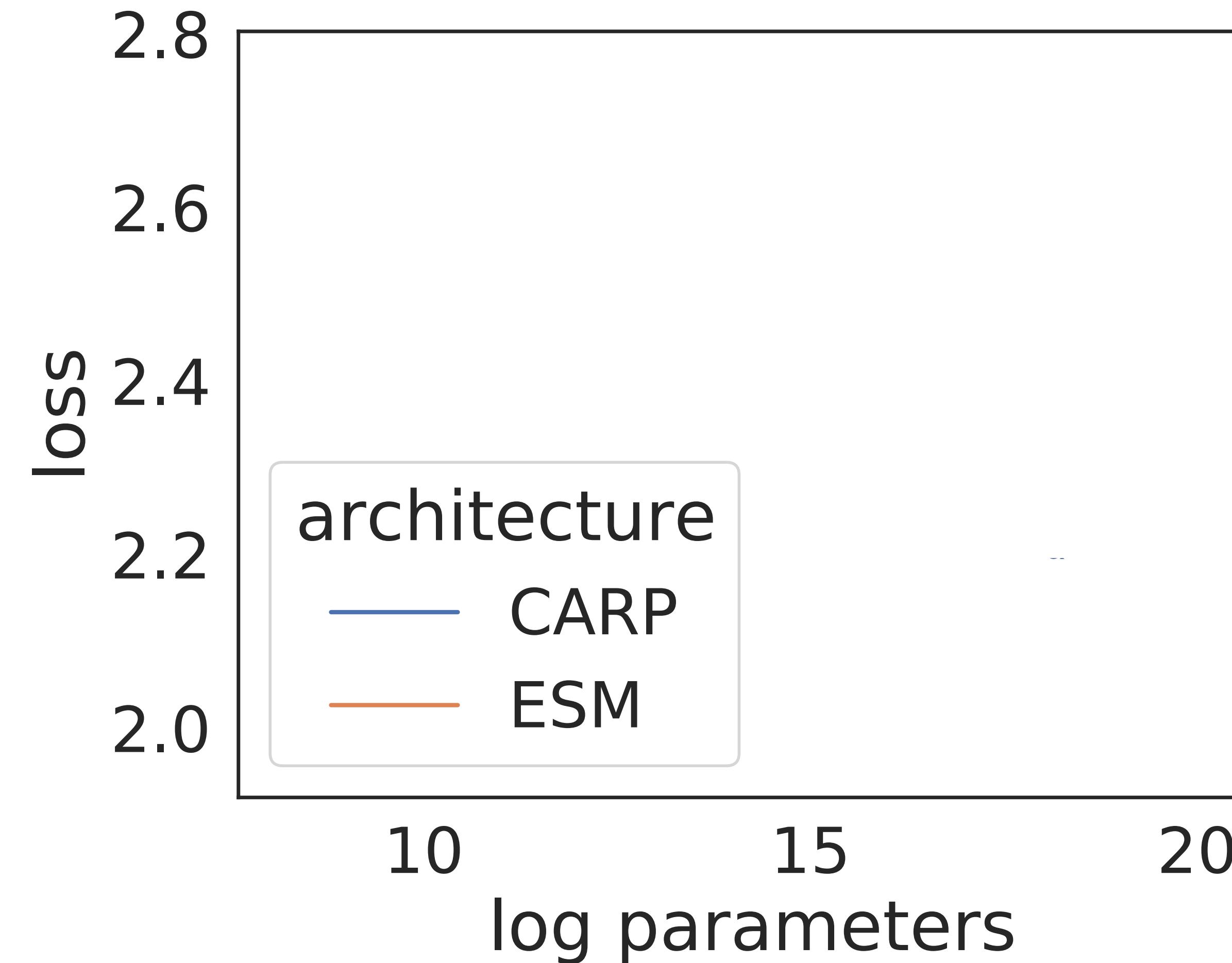


CNNs are competitive with transformers for pretraining

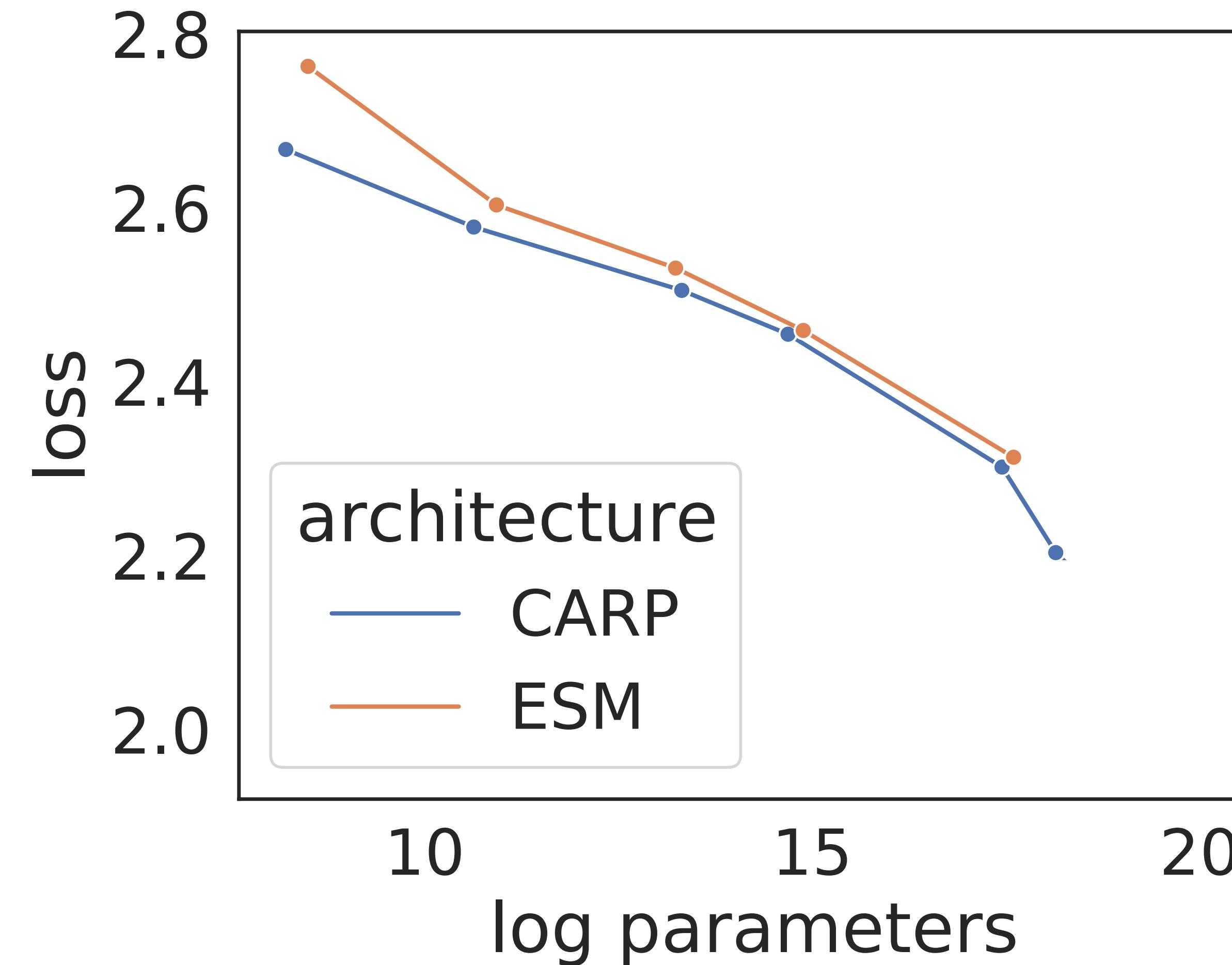
CNNs are competitive with transformers for pretraining



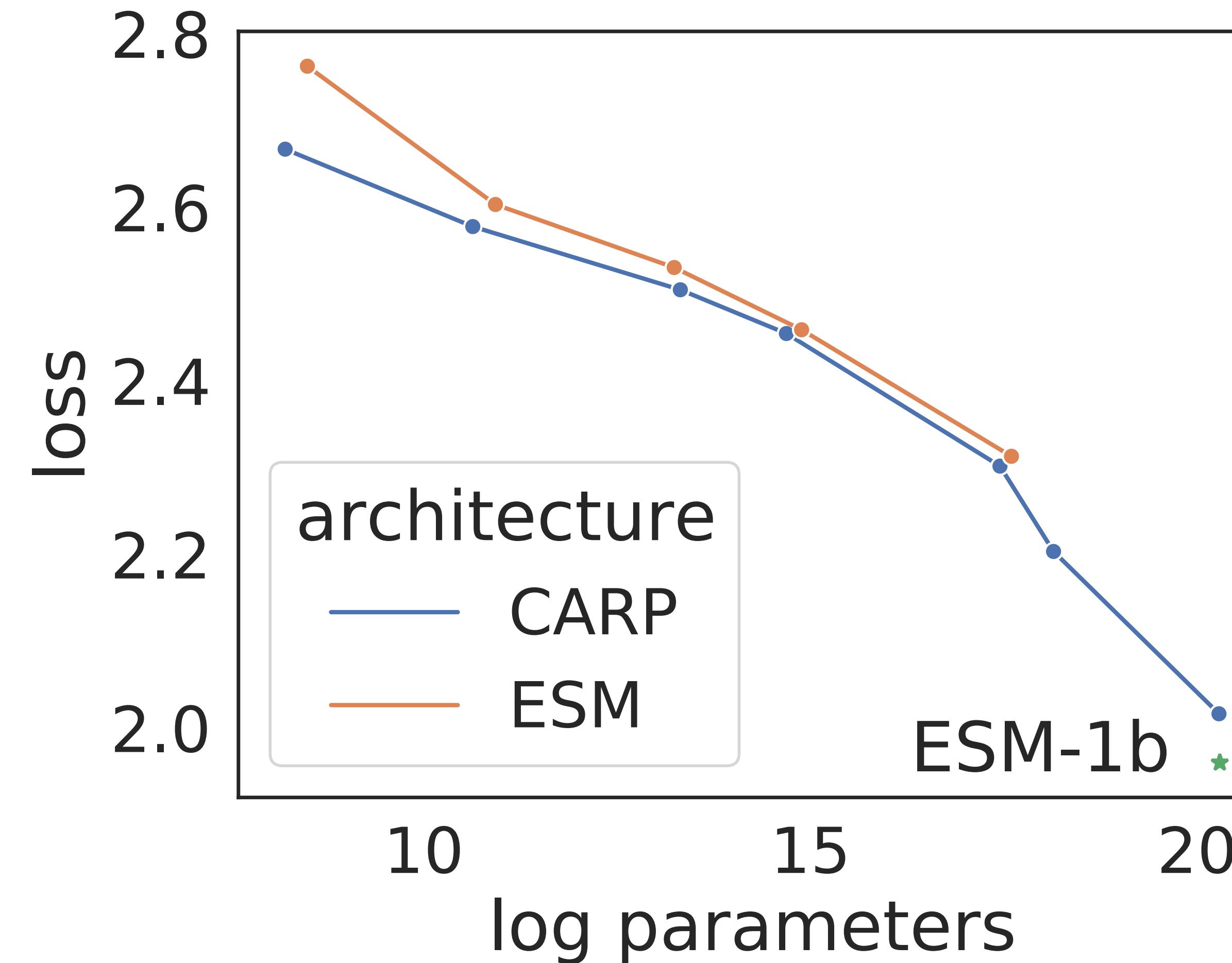
CNNs are competitive with transformers for pretraining



CNNs are competitive with transformers for pretraining

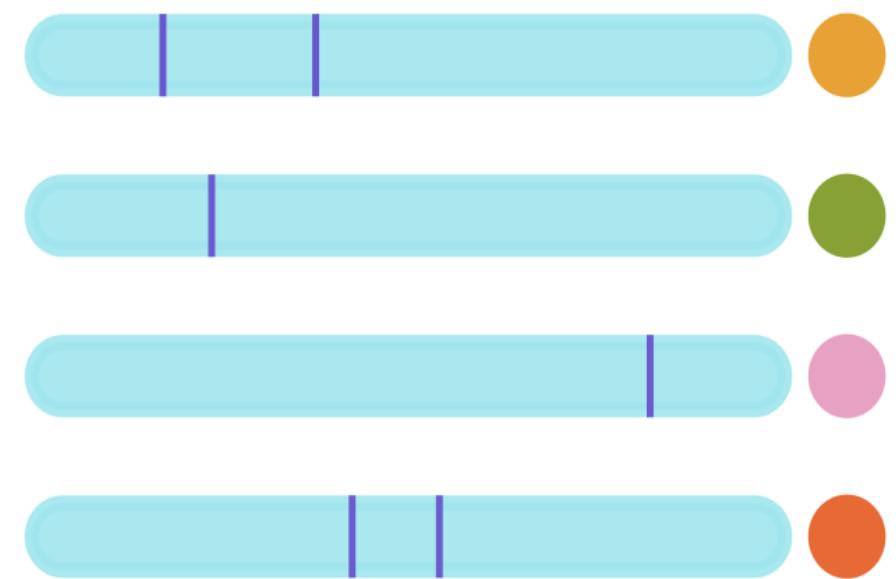


CNNs are competitive with transformers for pretraining

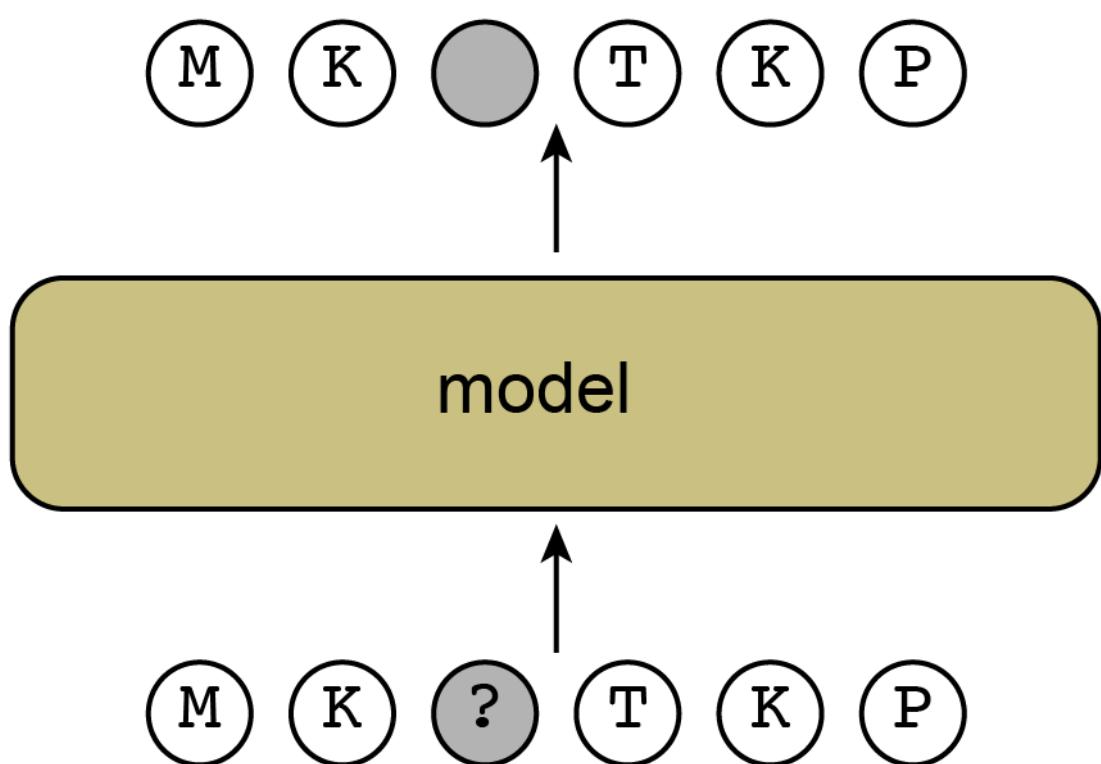


CARP is a zero-shot fitness predictor

Deep mutational scan

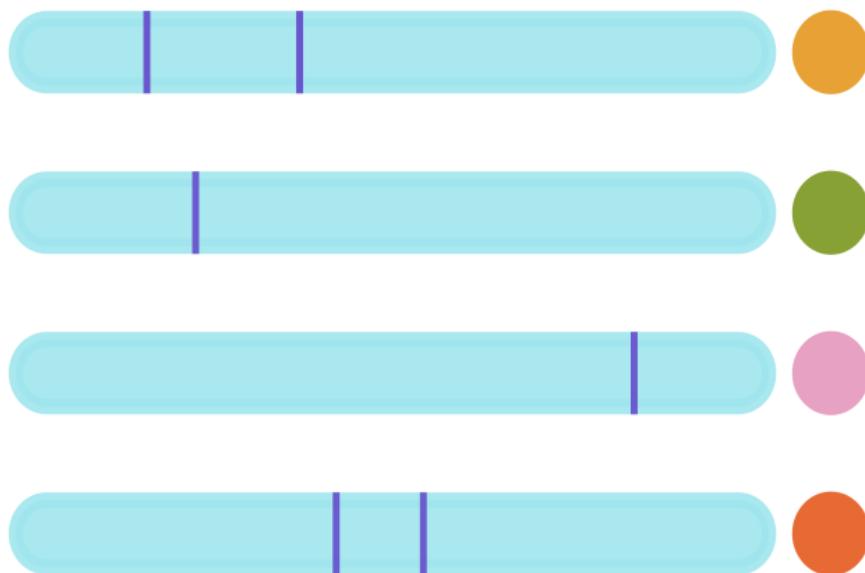


compare likelihood of
mutation to WT



CARP is a zero-shot fitness predictor

Deep mutational scan



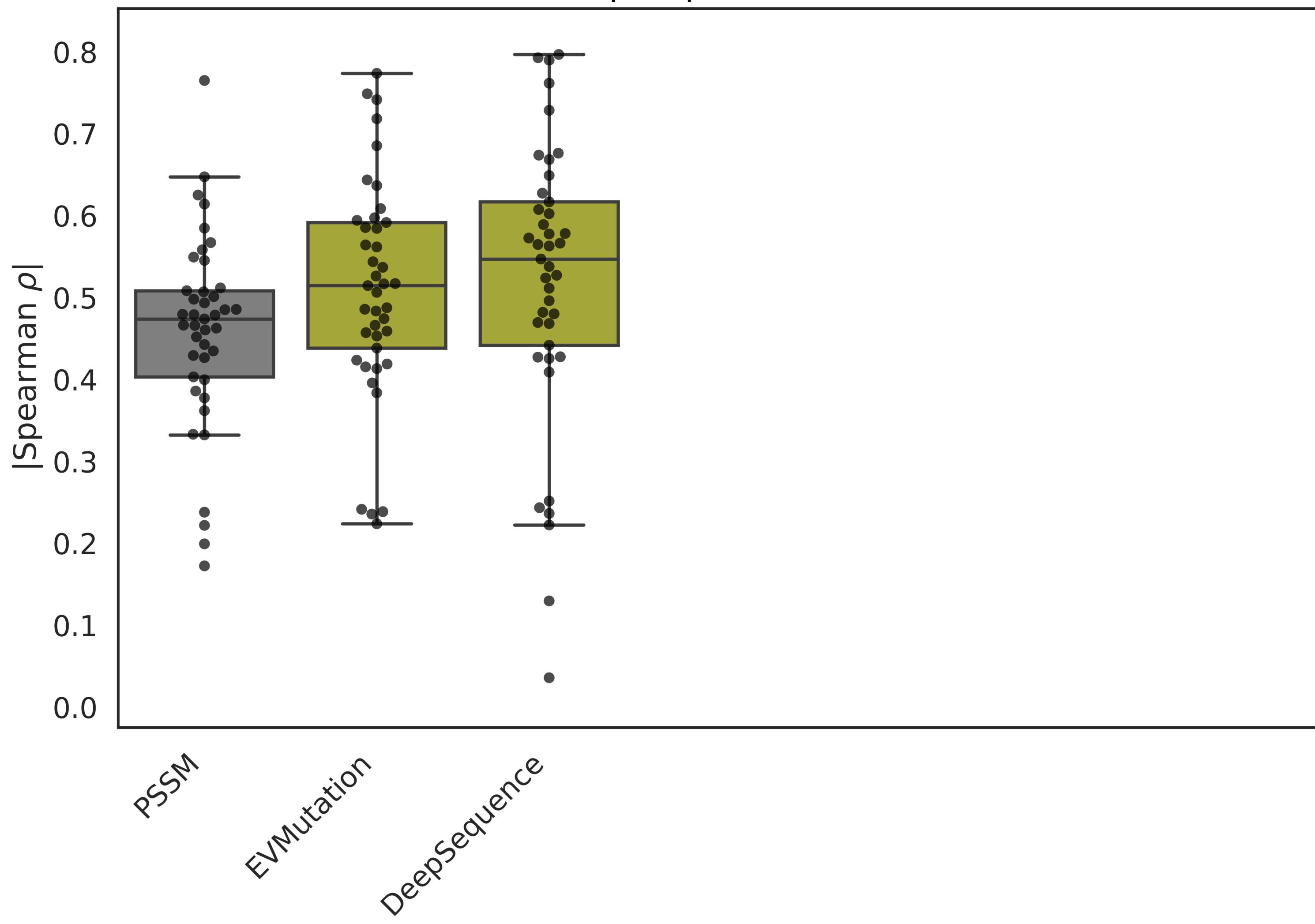
compare likelihood of
mutation to WT



model



DeepSequence



CARP is a zero-shot fitness predictor

Deep mutational scan



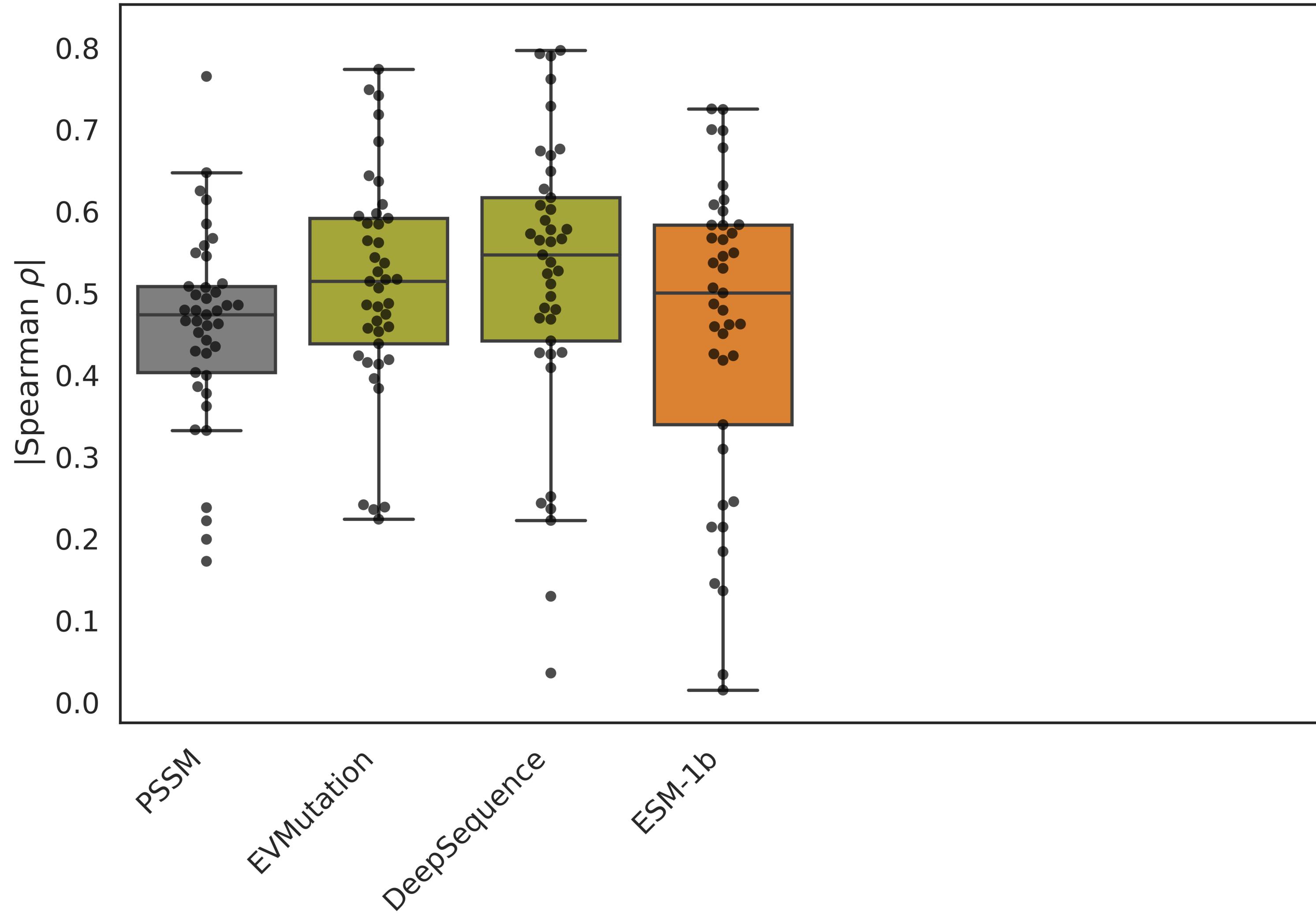
compare likelihood of
mutation to WT



model

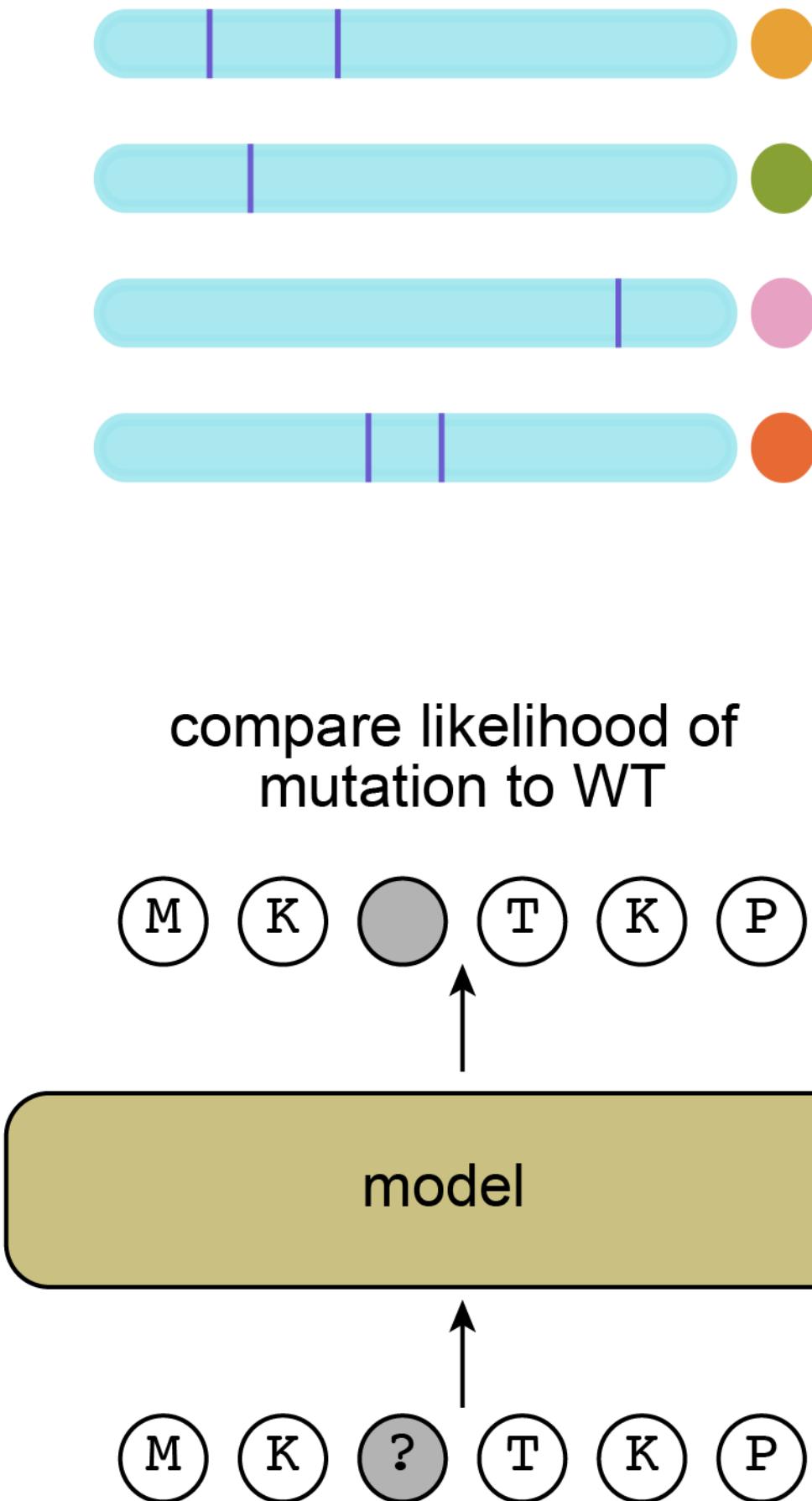


DeepSequence

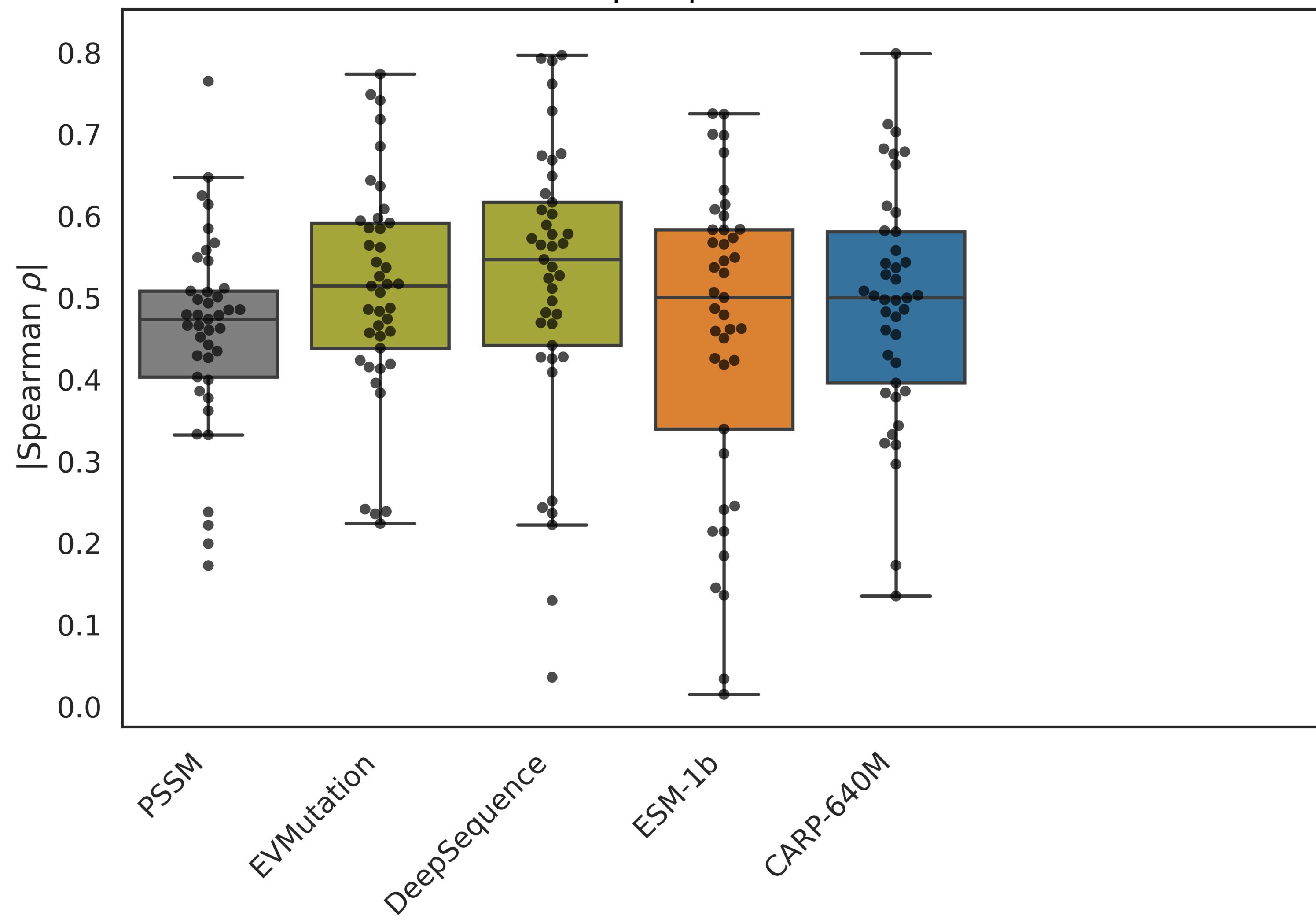


CARP is a zero-shot fitness predictor

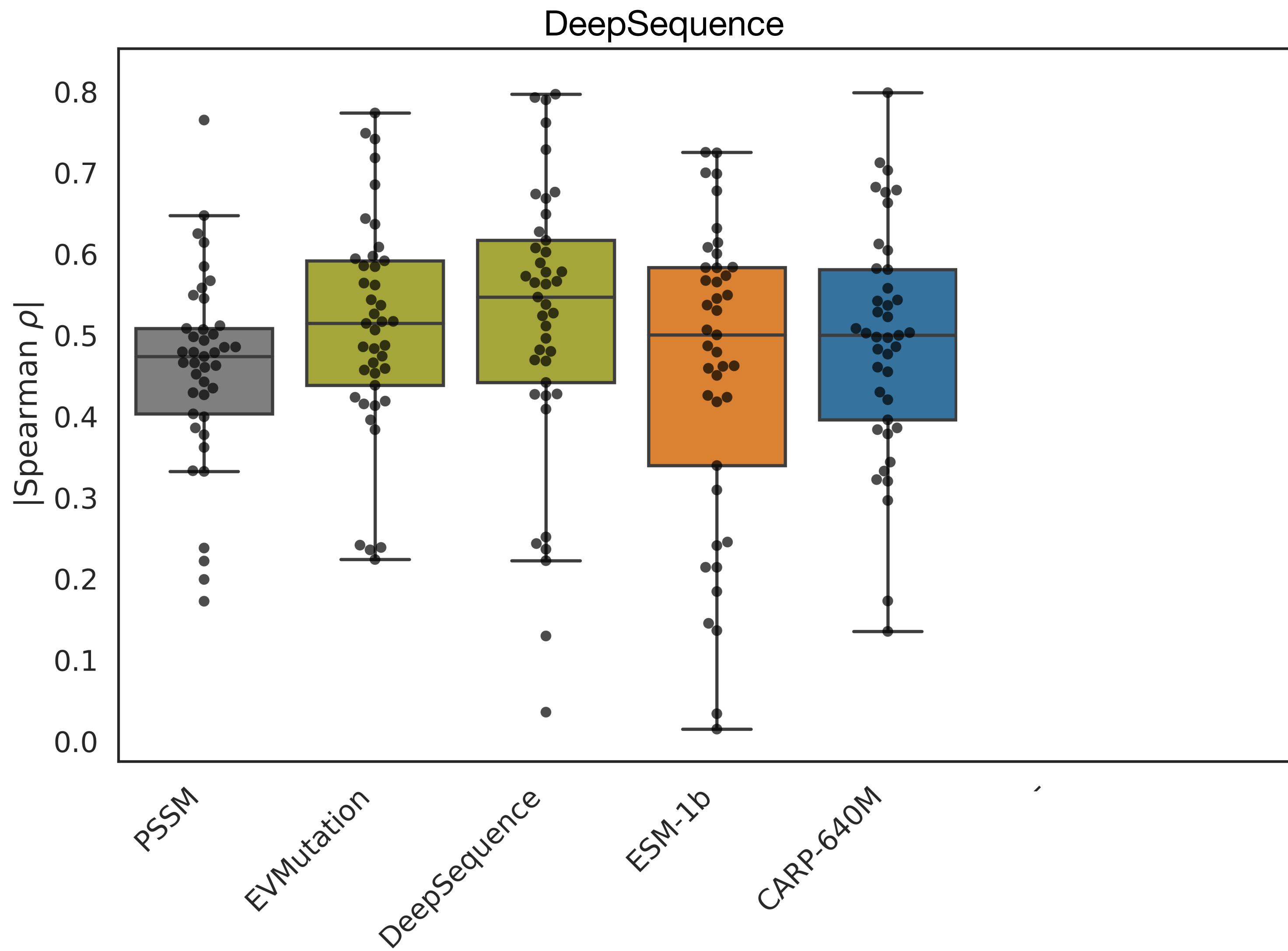
Deep mutational scan



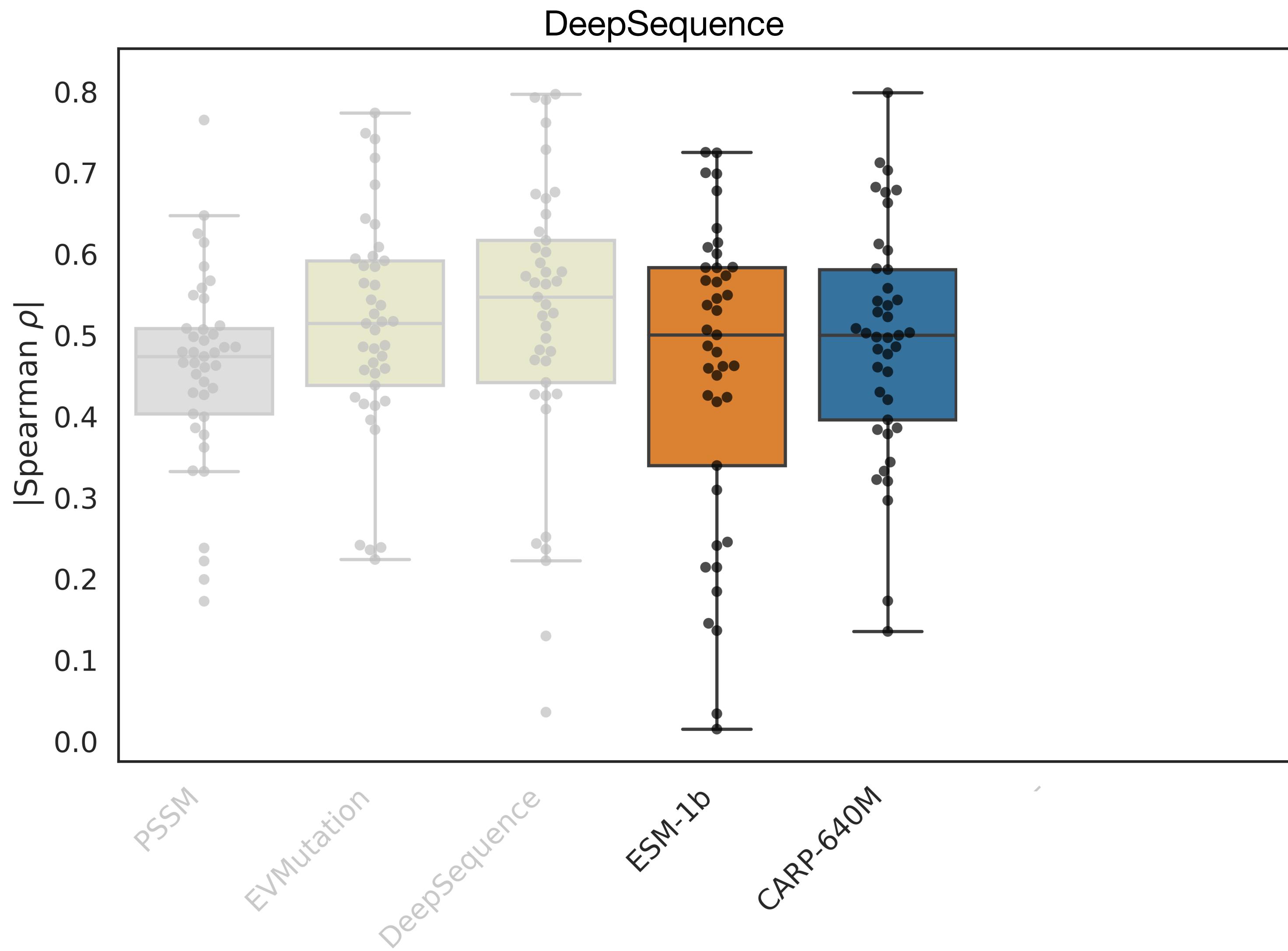
DeepSequence



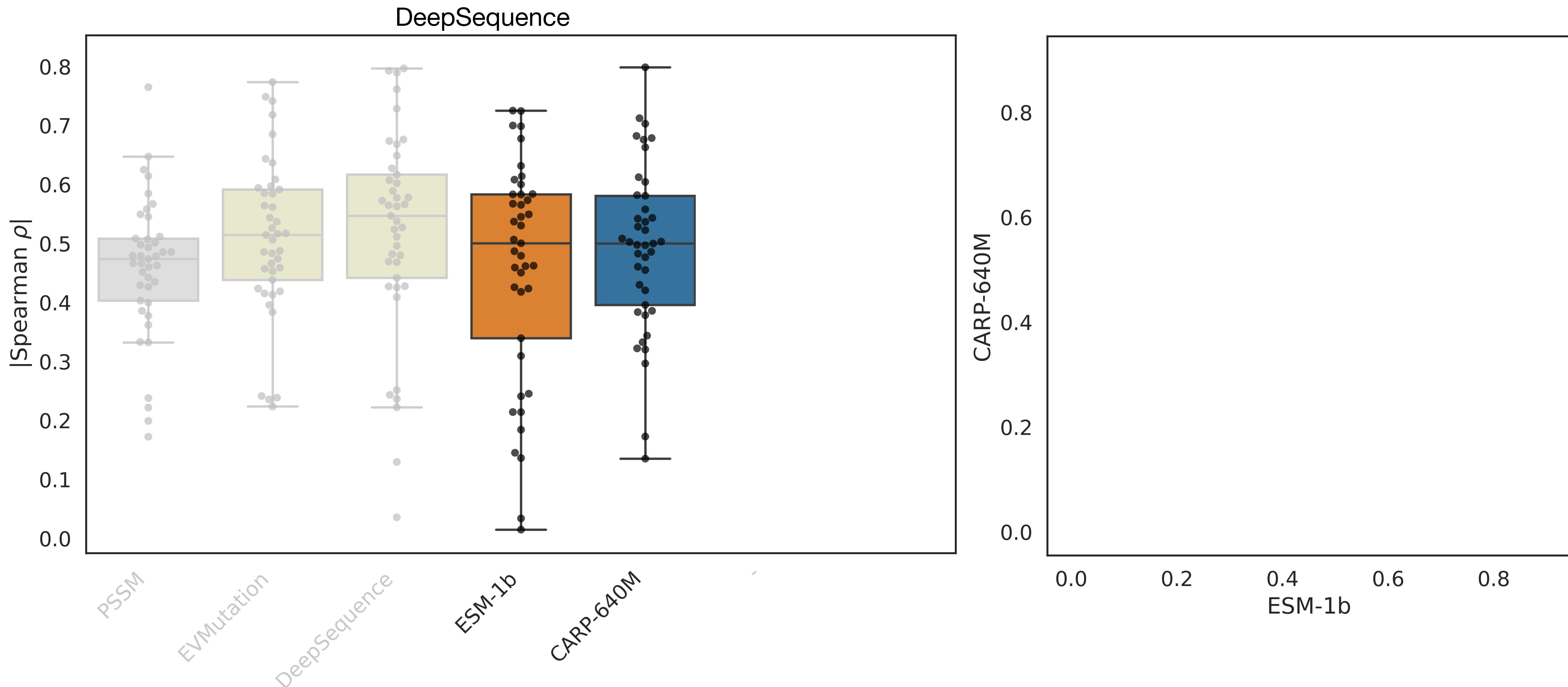
CARP is a zero-shot fitness predictor



CARP is a zero-shot fitness predictor

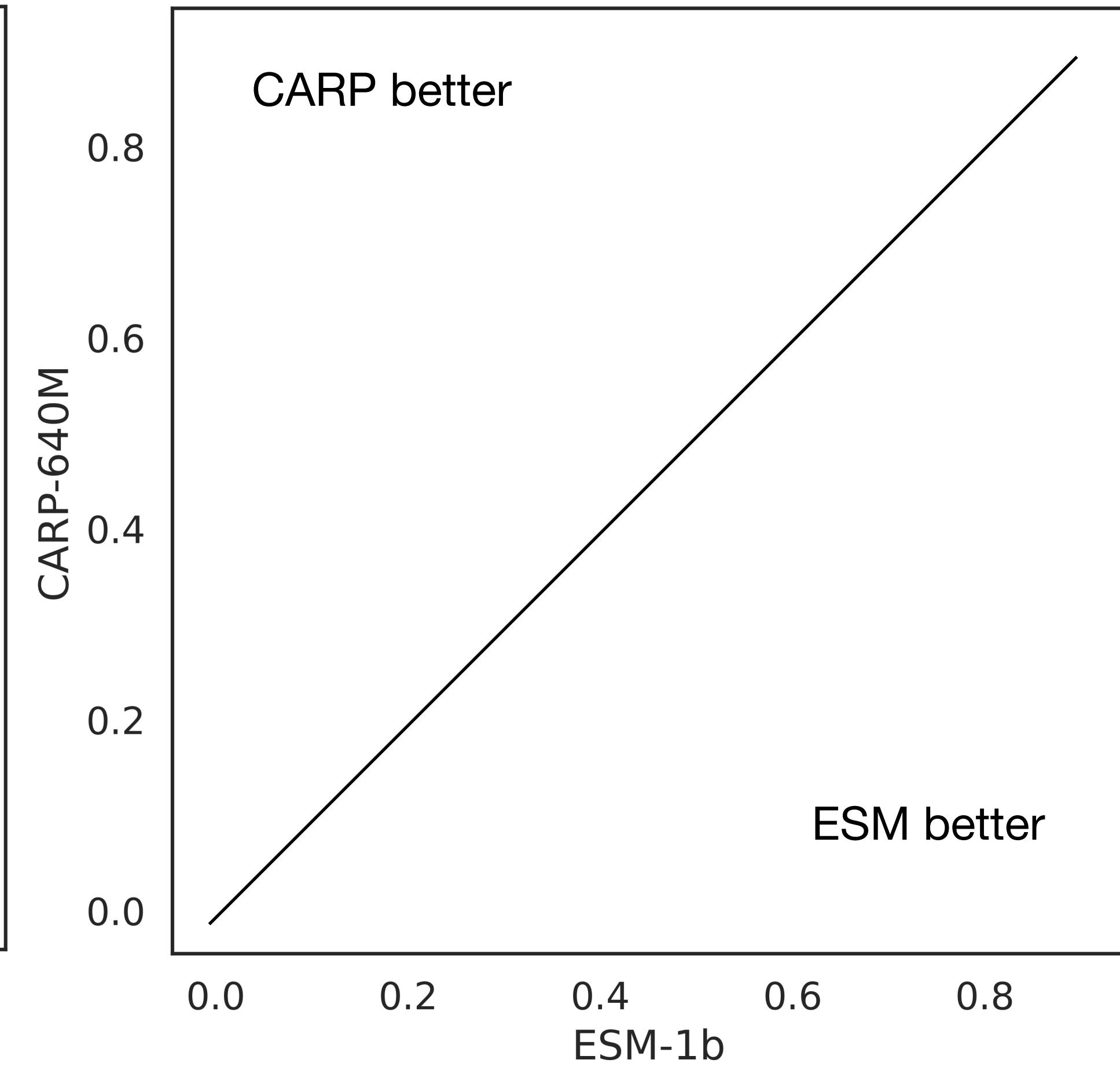
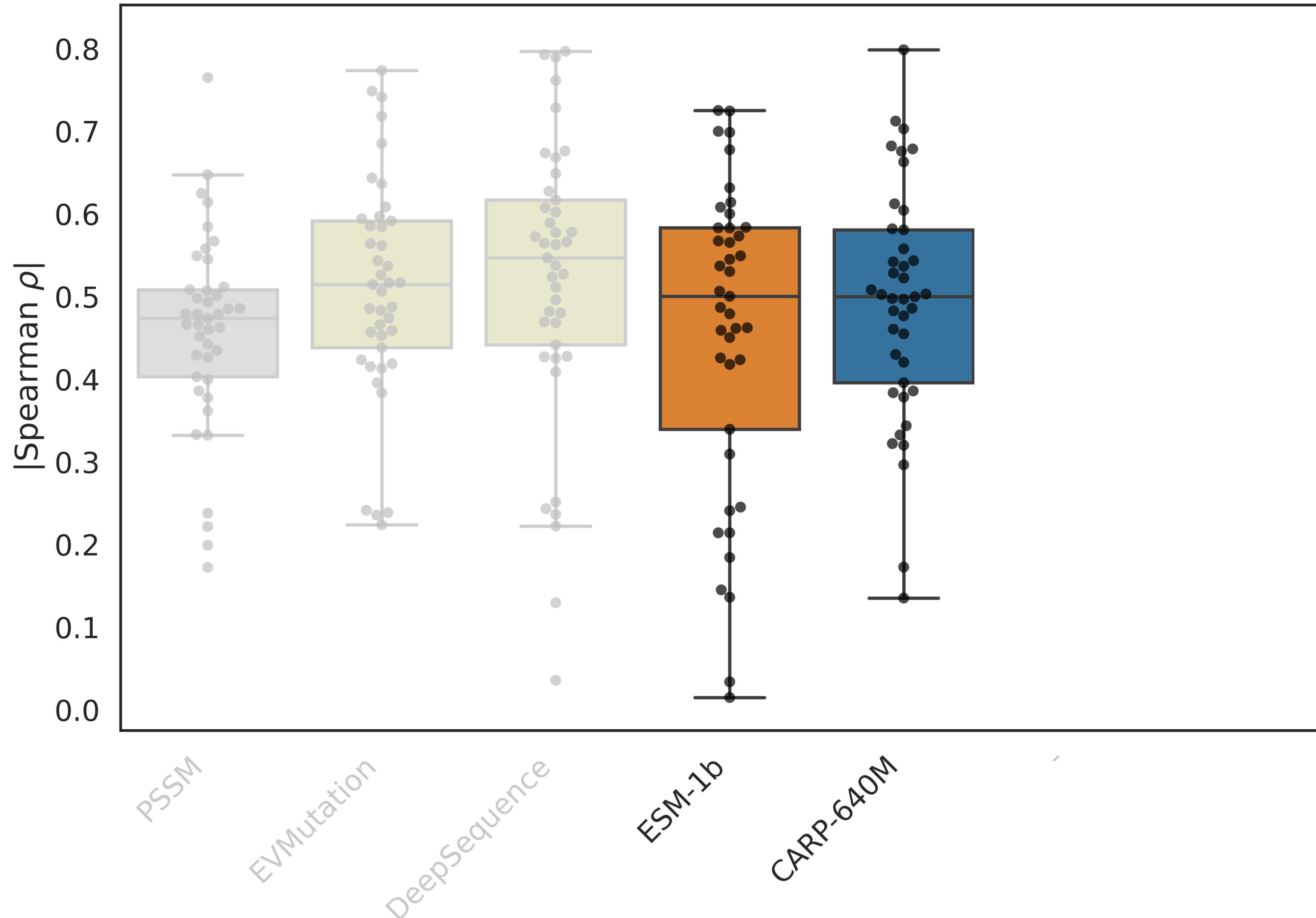


CARP is a zero-shot fitness predictor



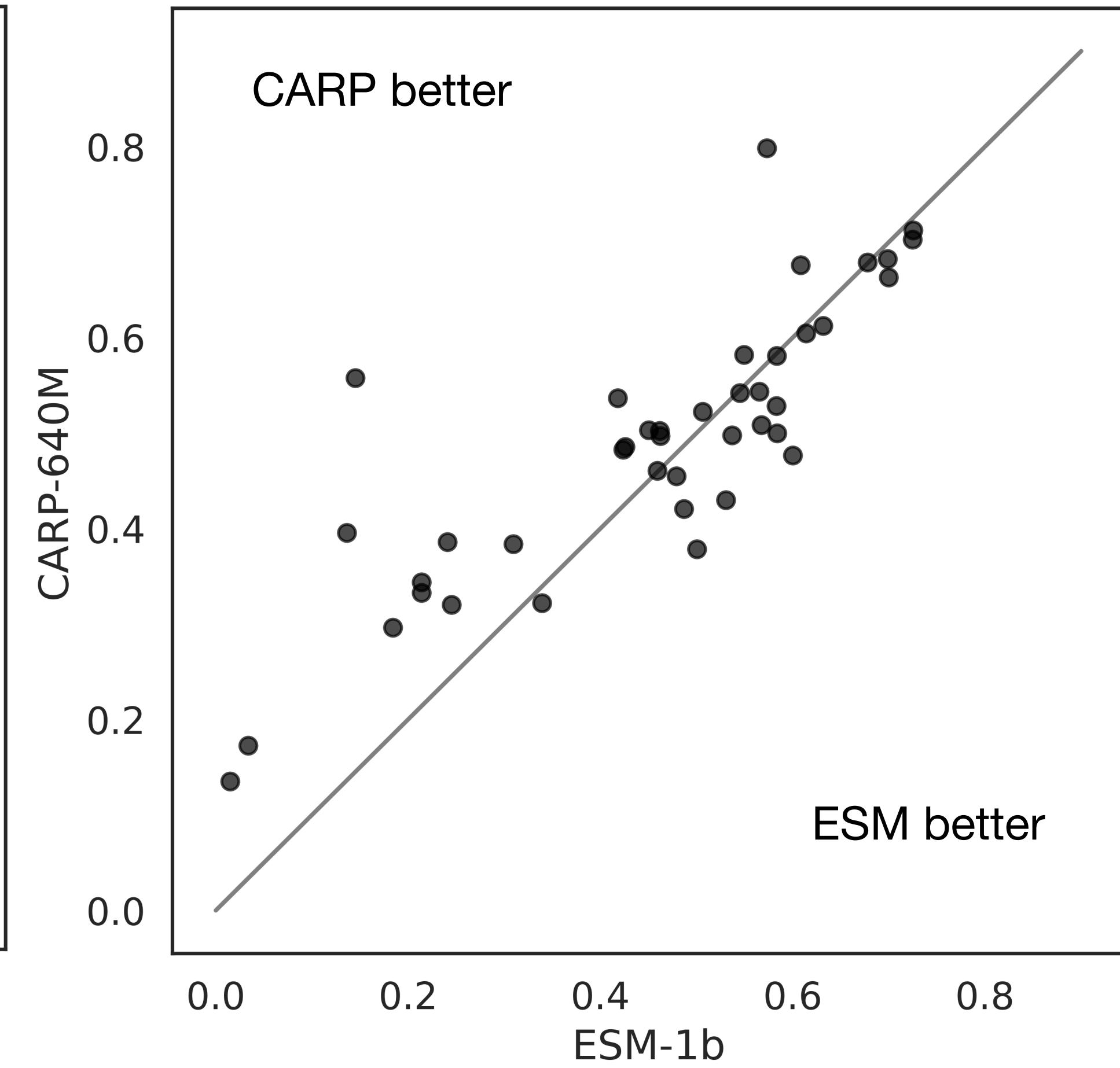
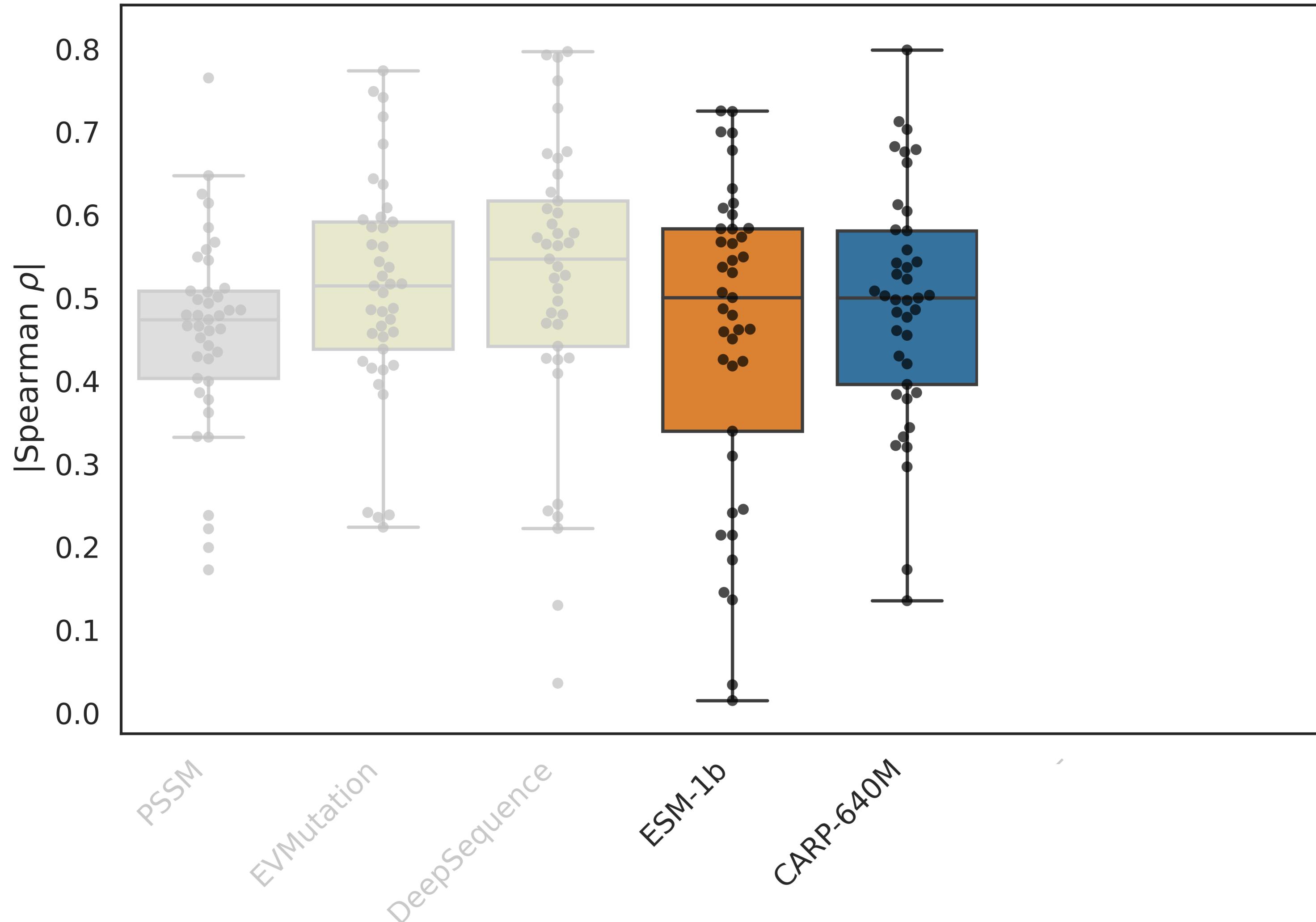
CARP is a zero-shot fitness predictor

DeepSequence



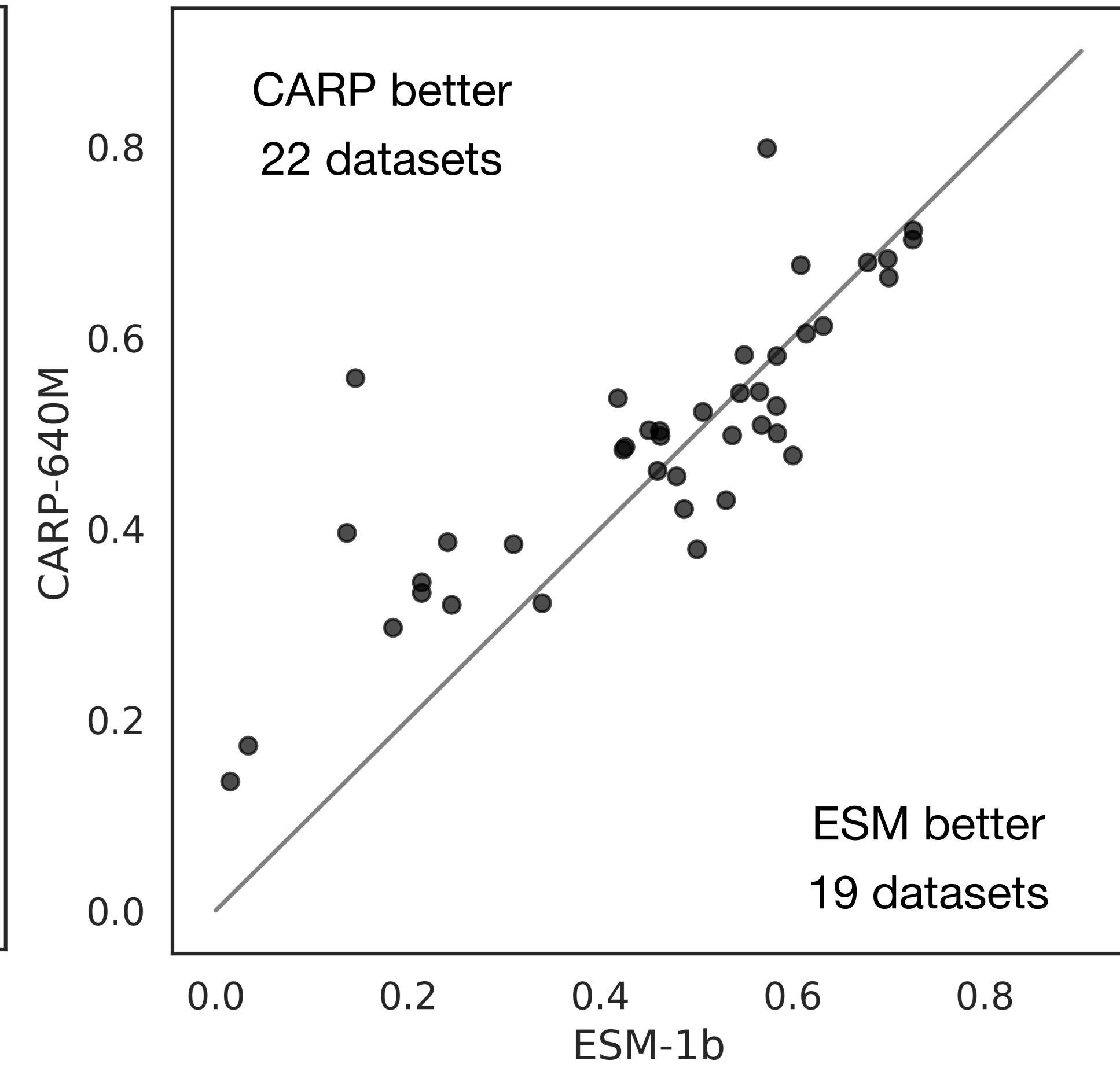
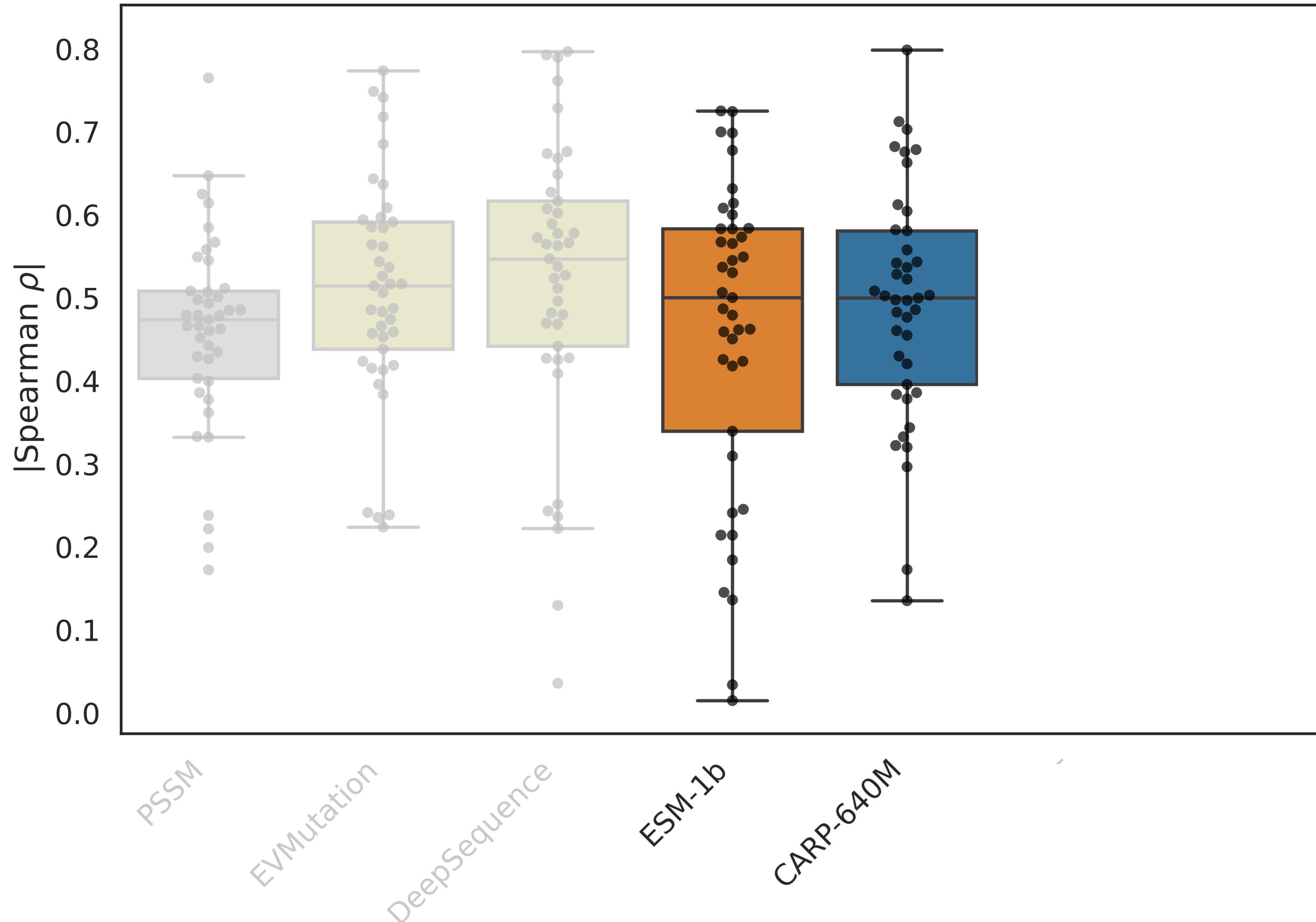
CARP is a zero-shot fitness predictor

DeepSequence



CARP is a zero-shot fitness predictor

DeepSequence



CARP learns structure

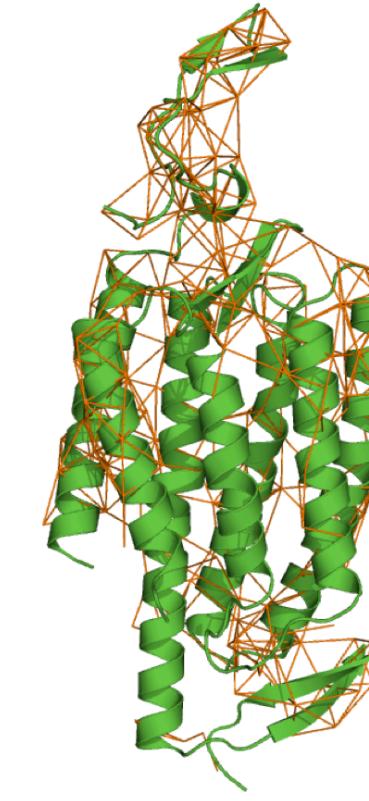
CARP learns structure

Model

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

CARP-640M

CARP learns structure

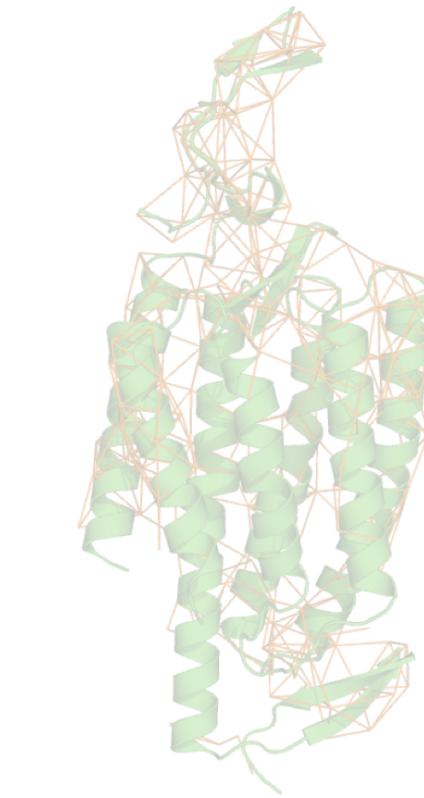


long-range contacts

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)

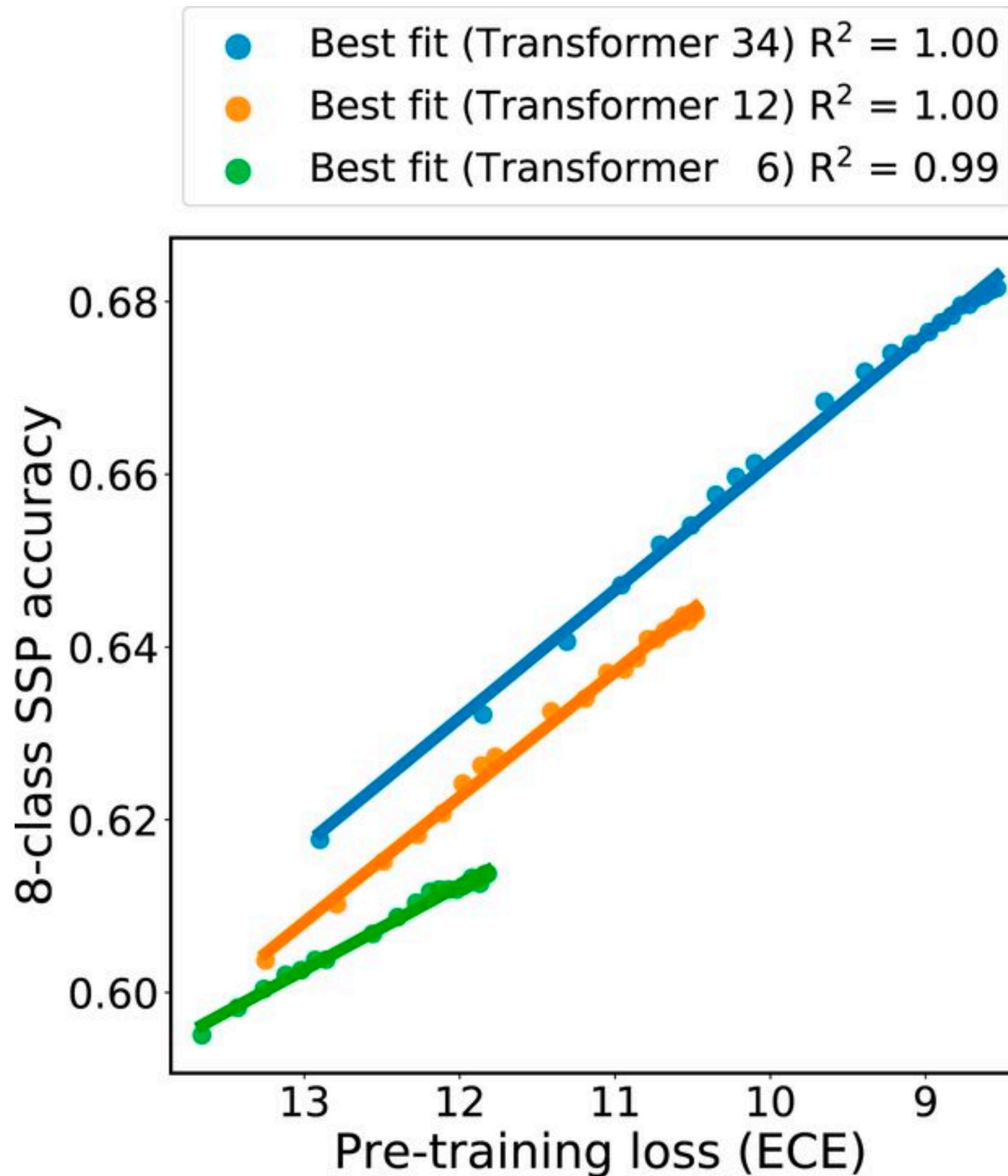


α -helix

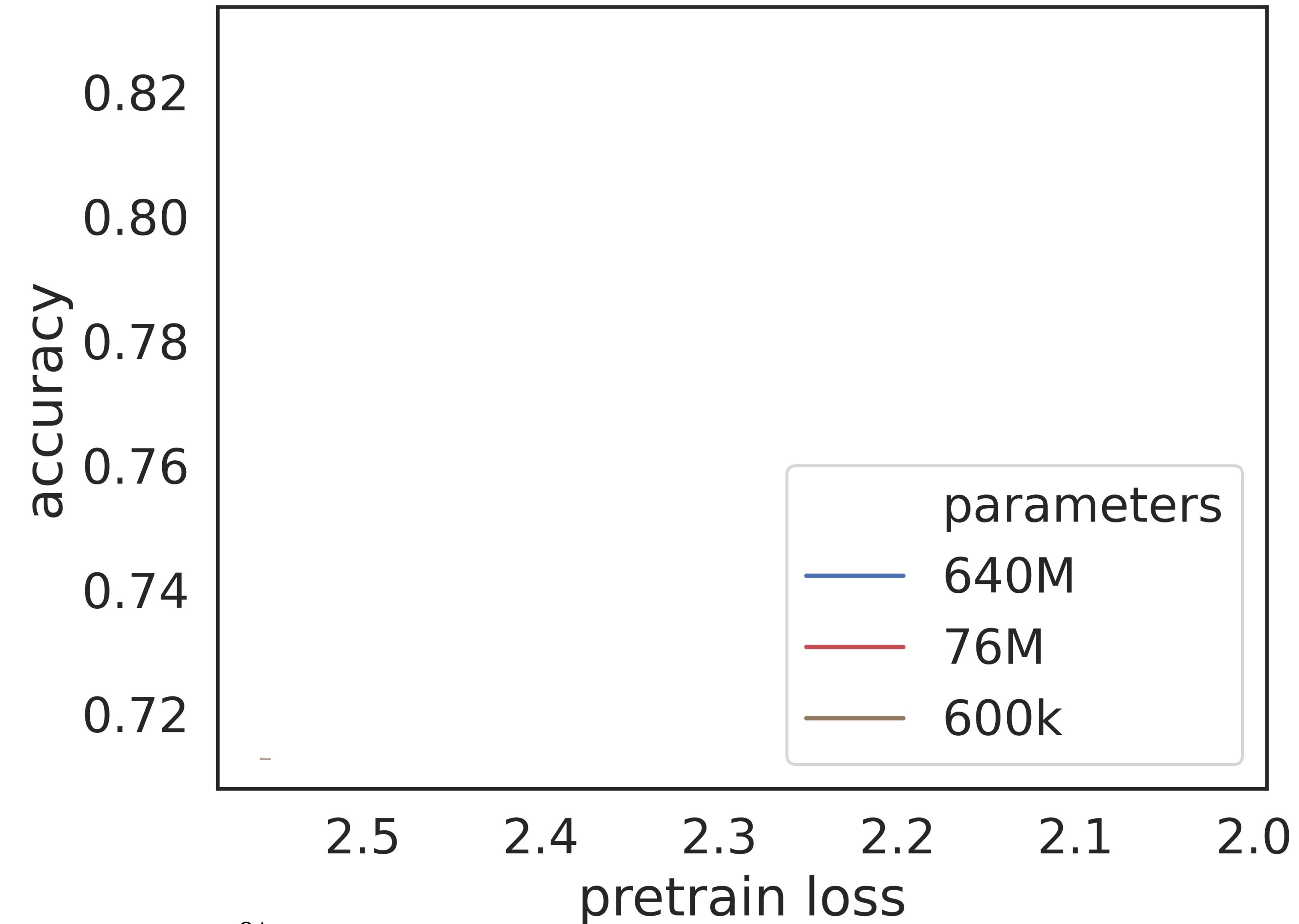
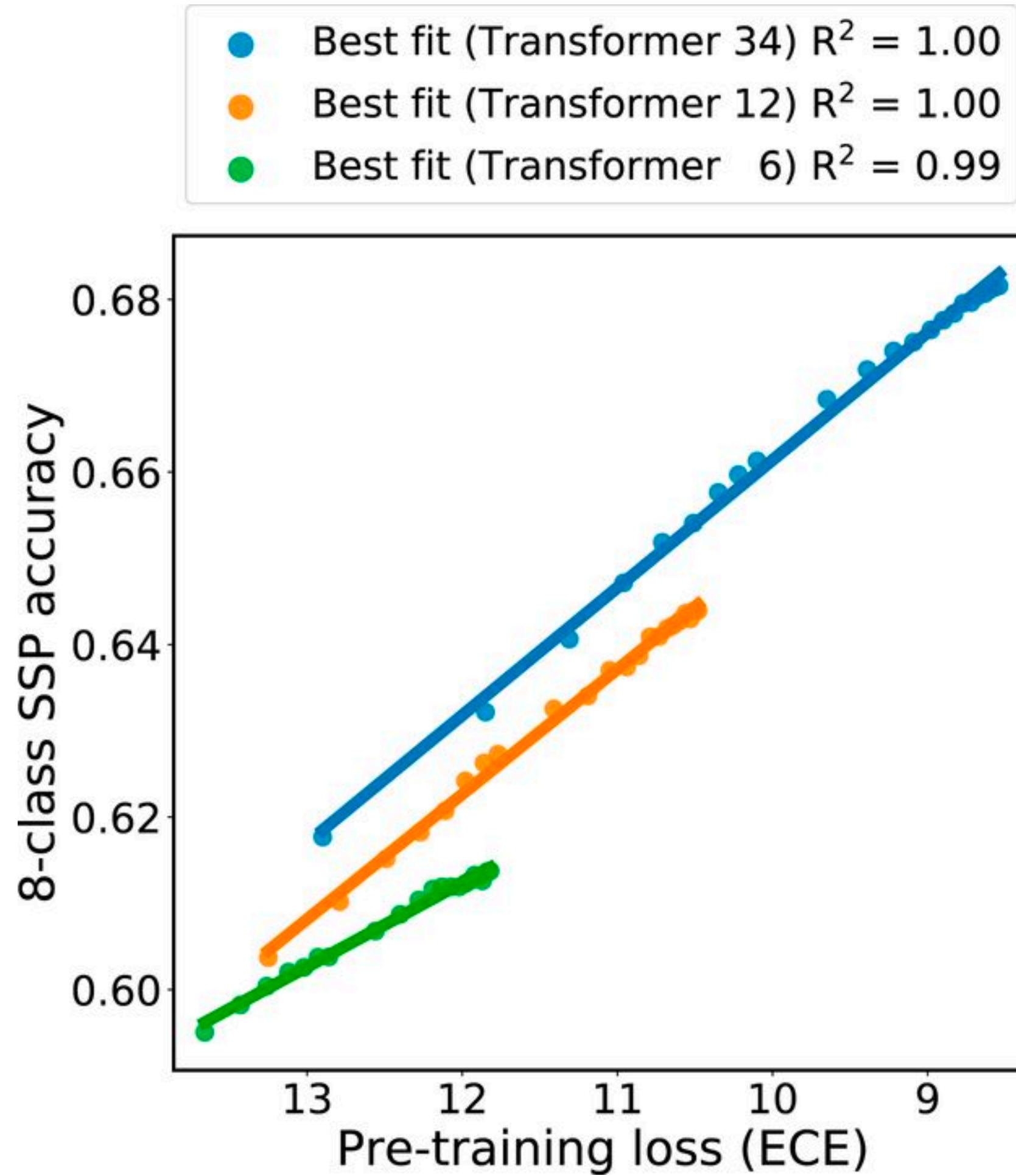
Thomas Shafee

Structure predictions improve smoothly with pretraining

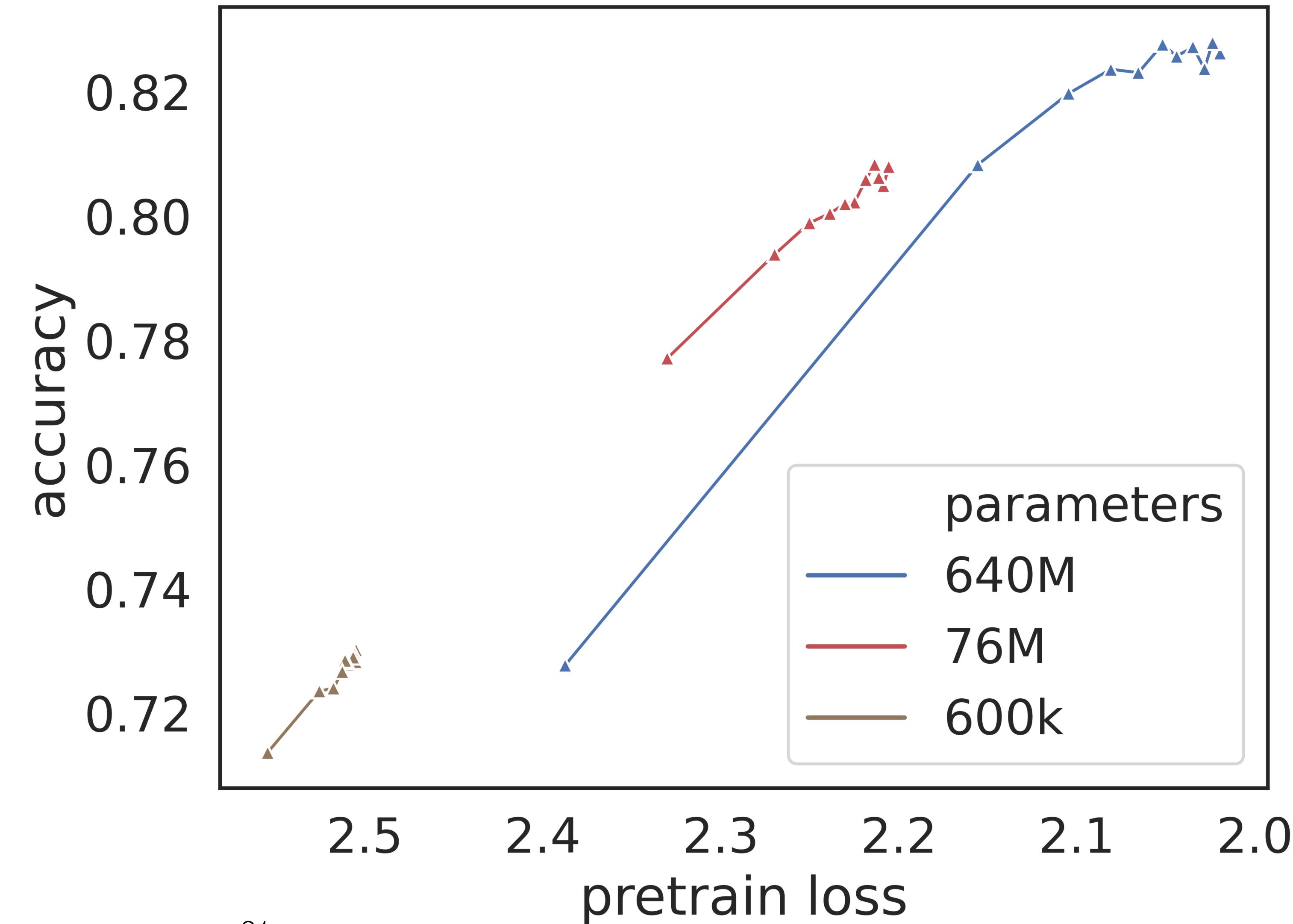
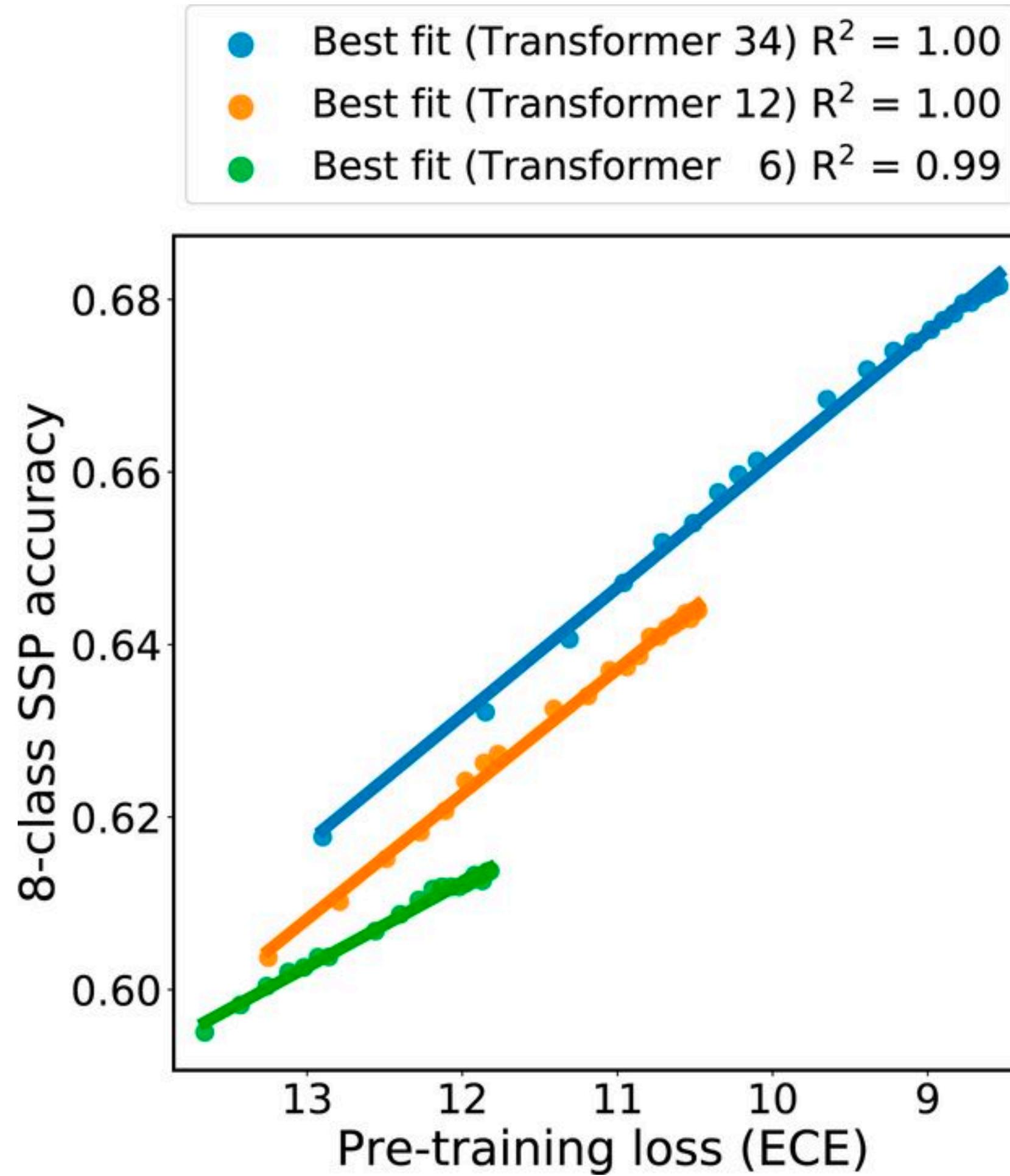
Structure predictions improve smoothly with pretraining



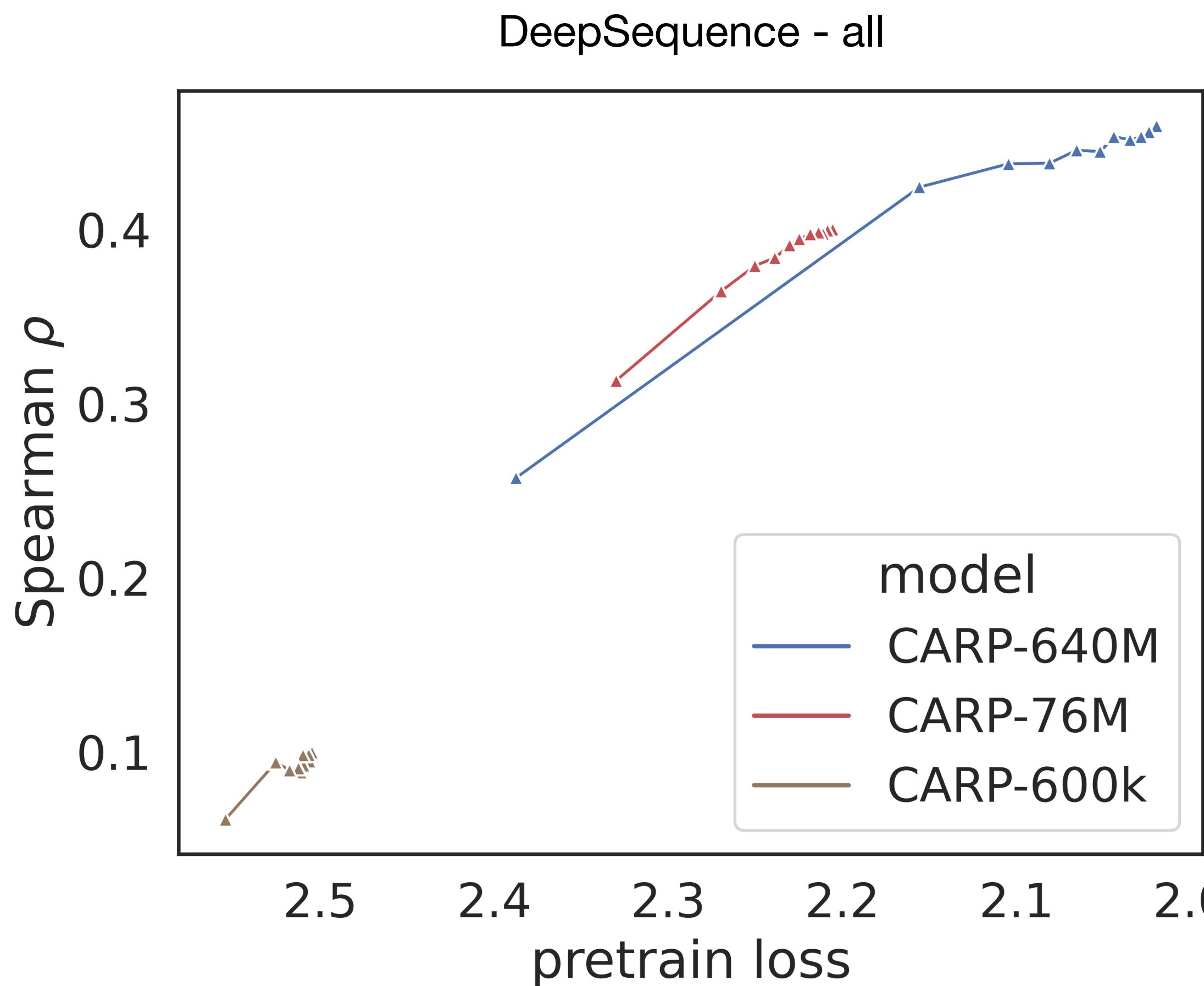
Structure predictions improve smoothly with pretraining



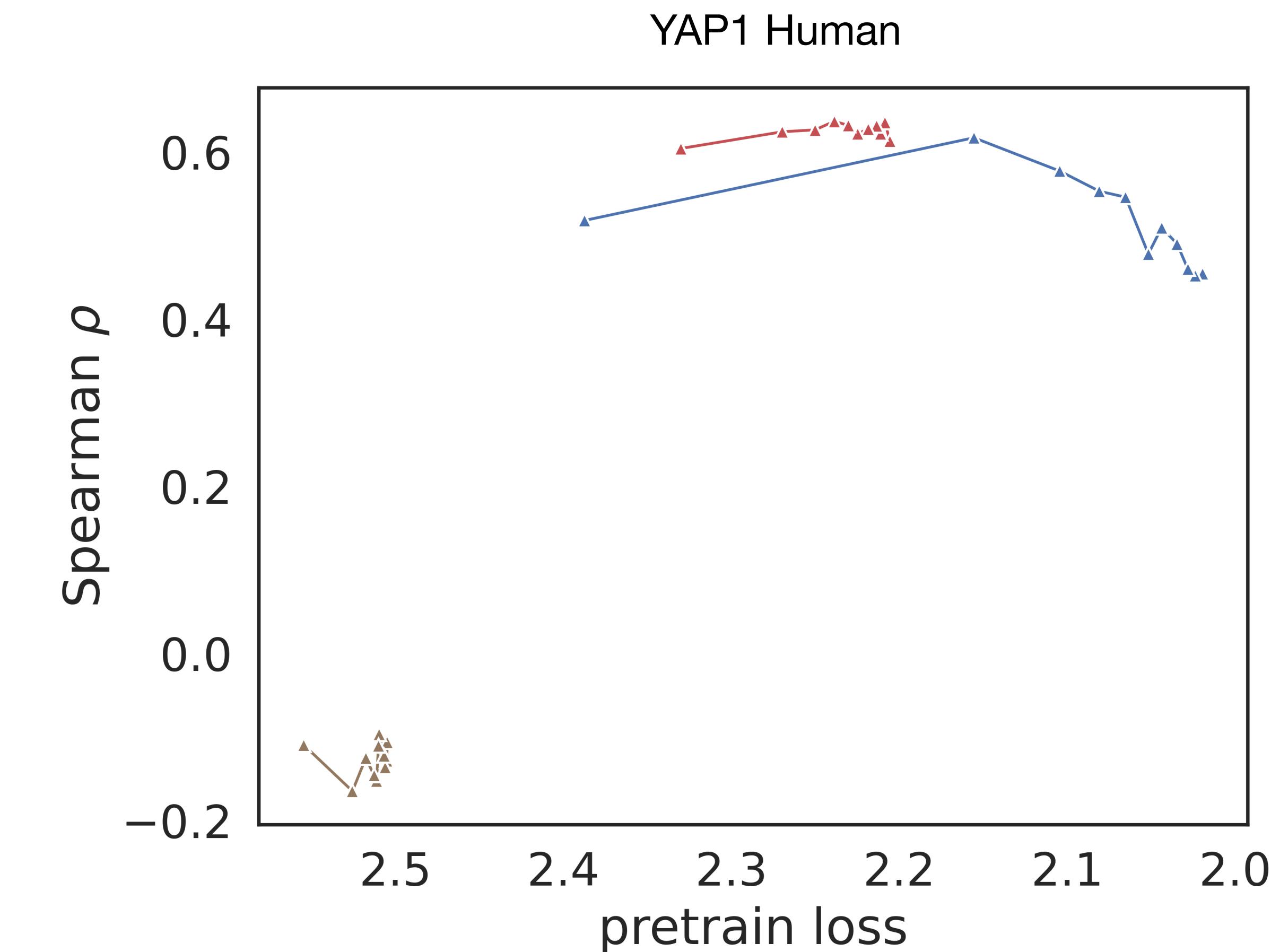
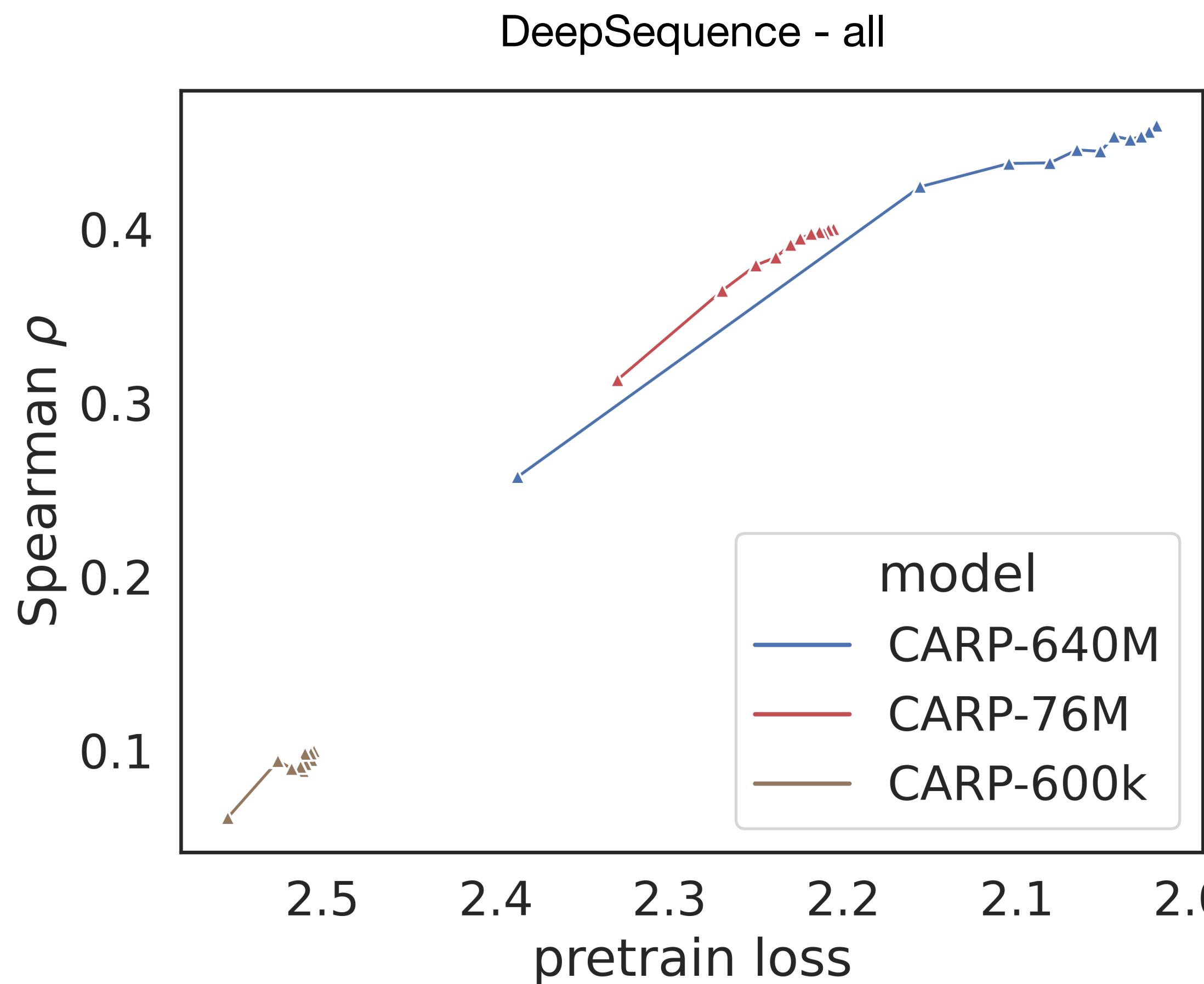
Structure predictions improve smoothly with pretraining



CARP zero-shot performance mostly improves with pretraining

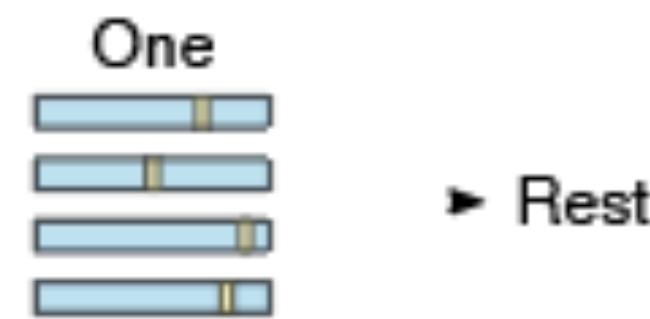


CARP zero-shot performance mostly improves with pretraining

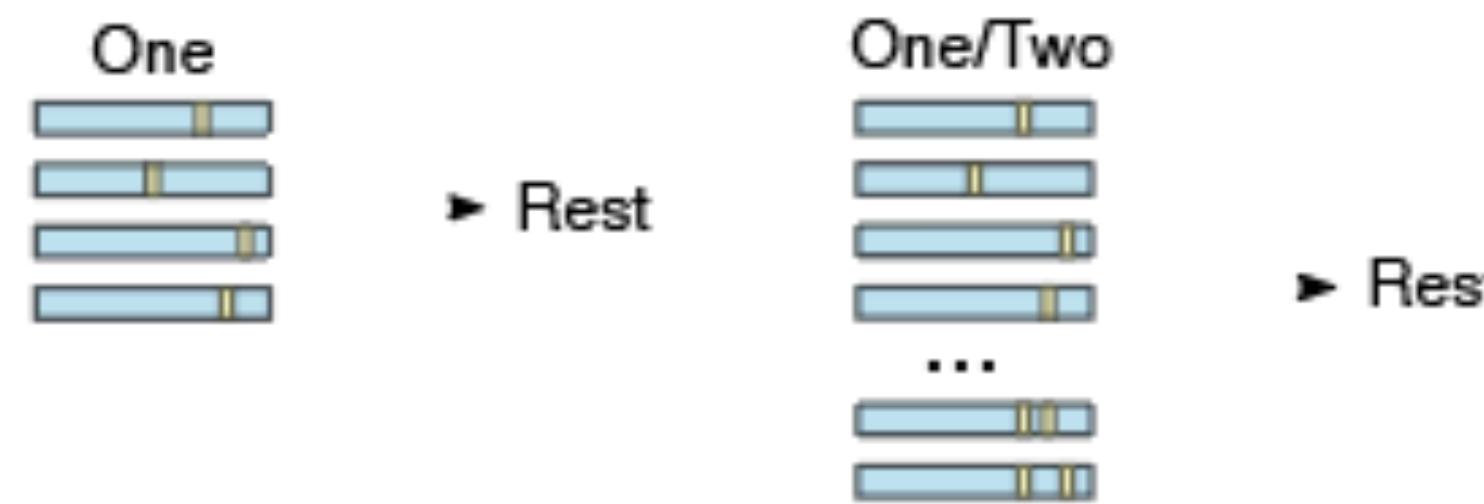


Protein engineering campaigns require out-of-domain generalization

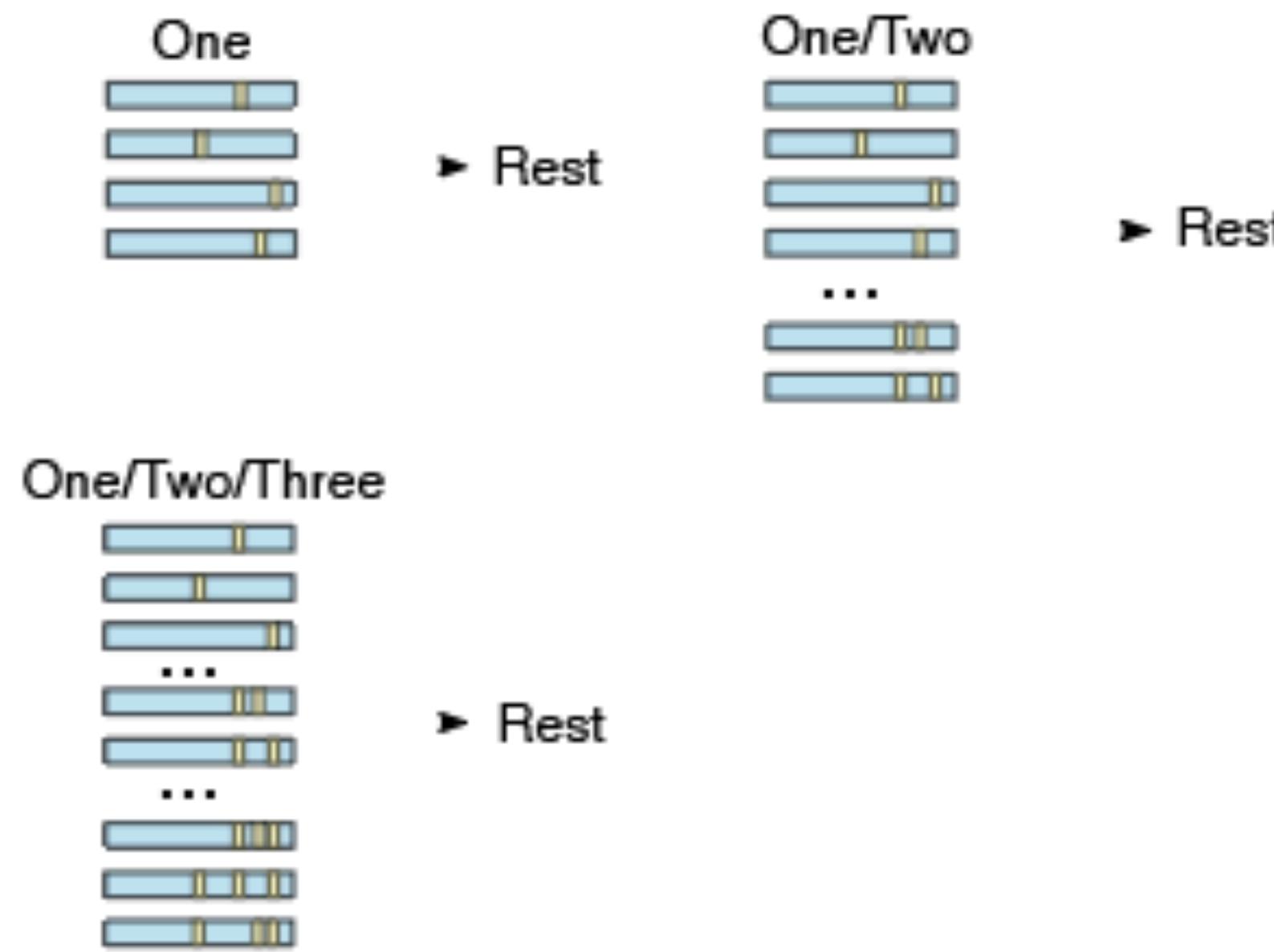
Protein engineering campaigns require out-of-domain generalization



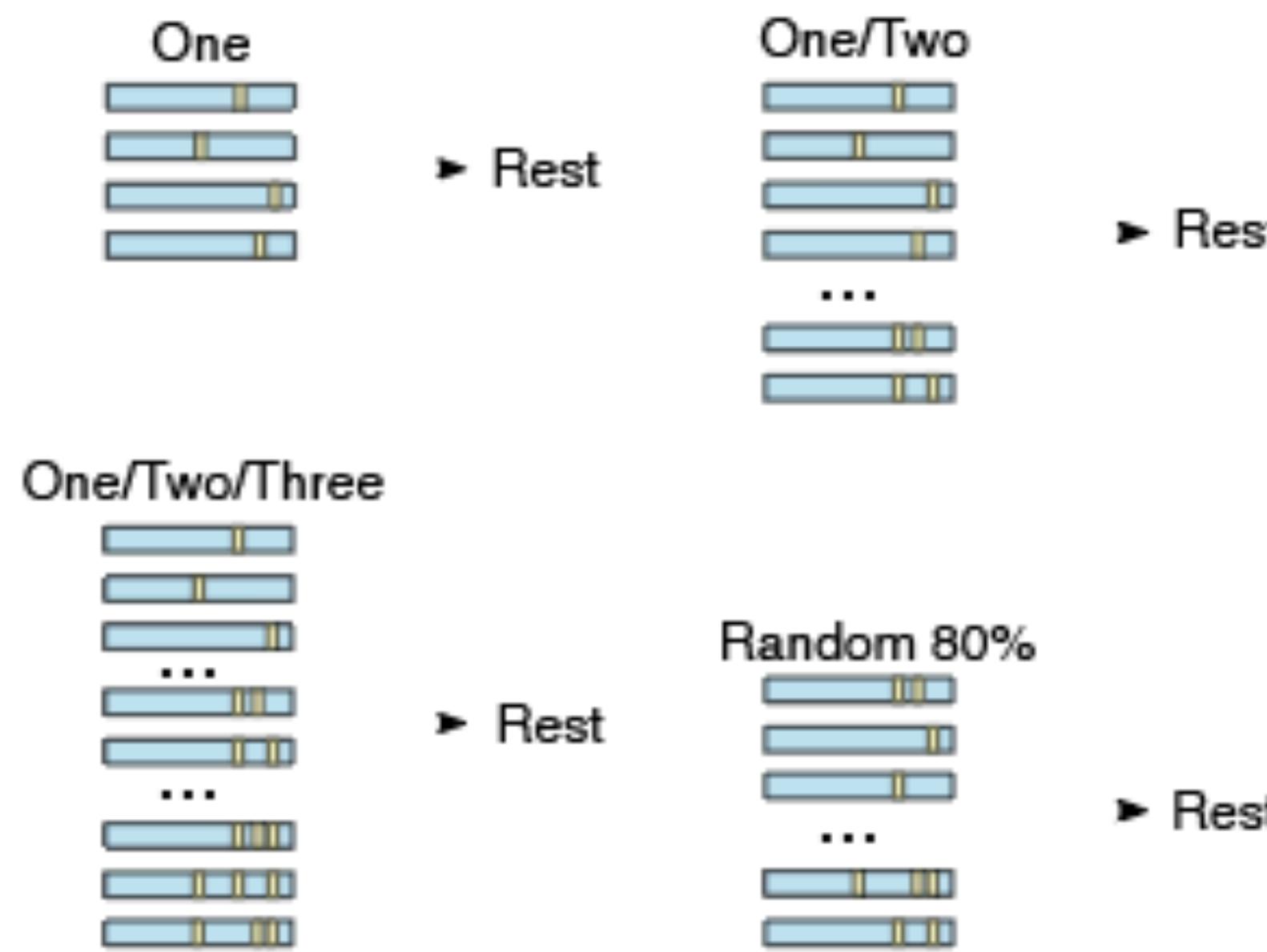
Protein engineering campaigns require out-of-domain generalization



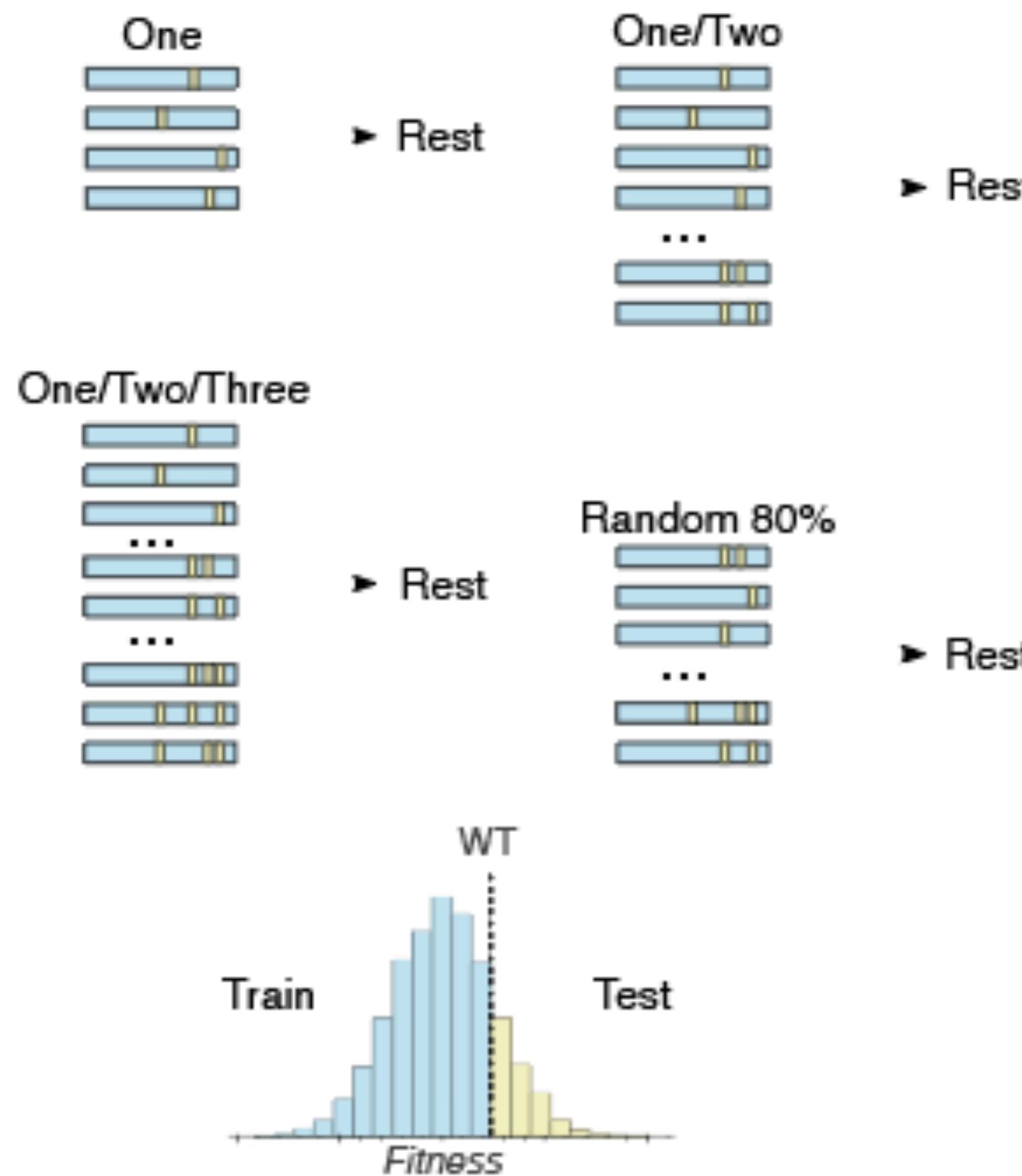
Protein engineering campaigns require out-of-domain generalization



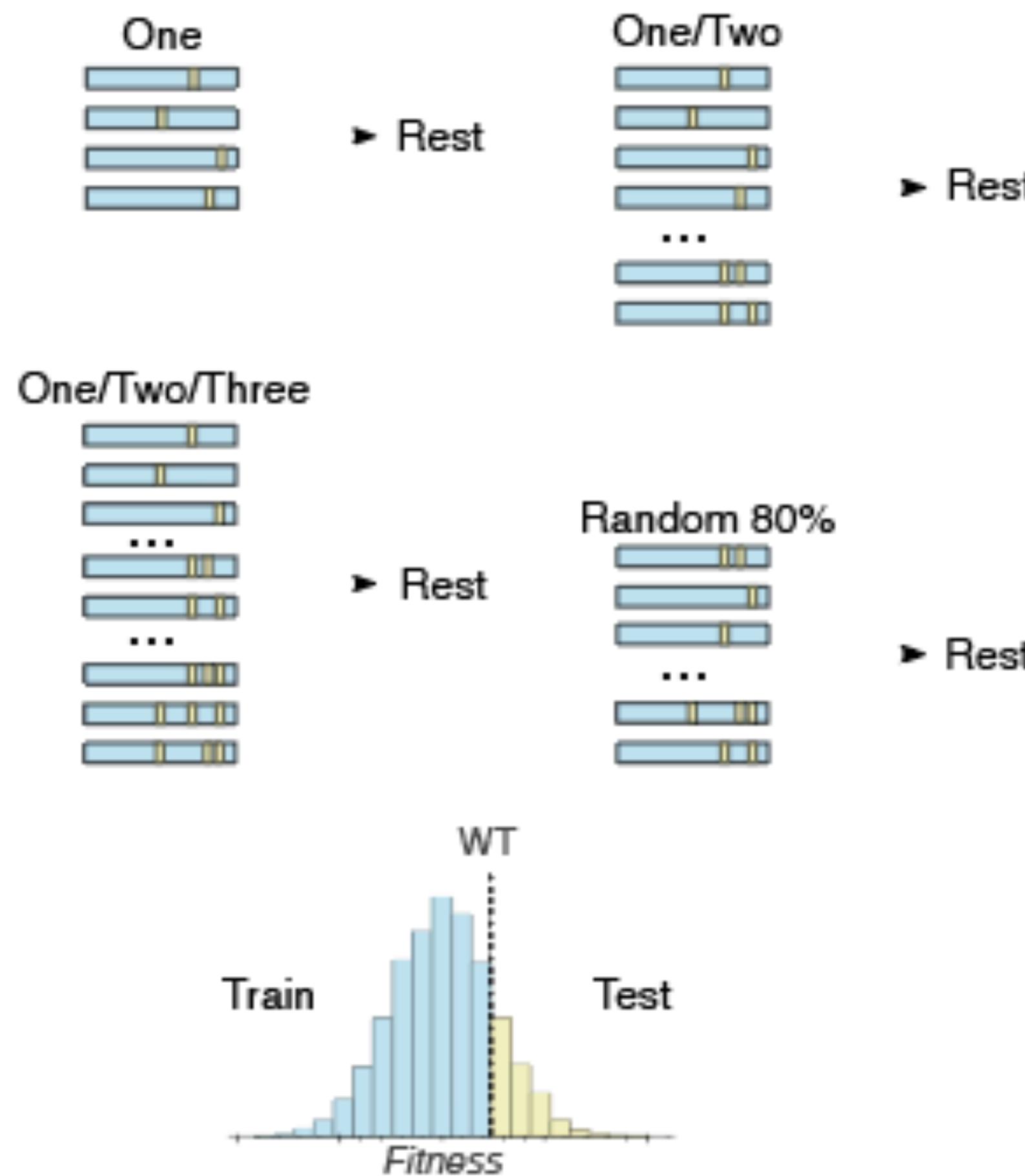
Protein engineering campaigns require out-of-domain generalization



Protein engineering campaigns require out-of-domain generalization

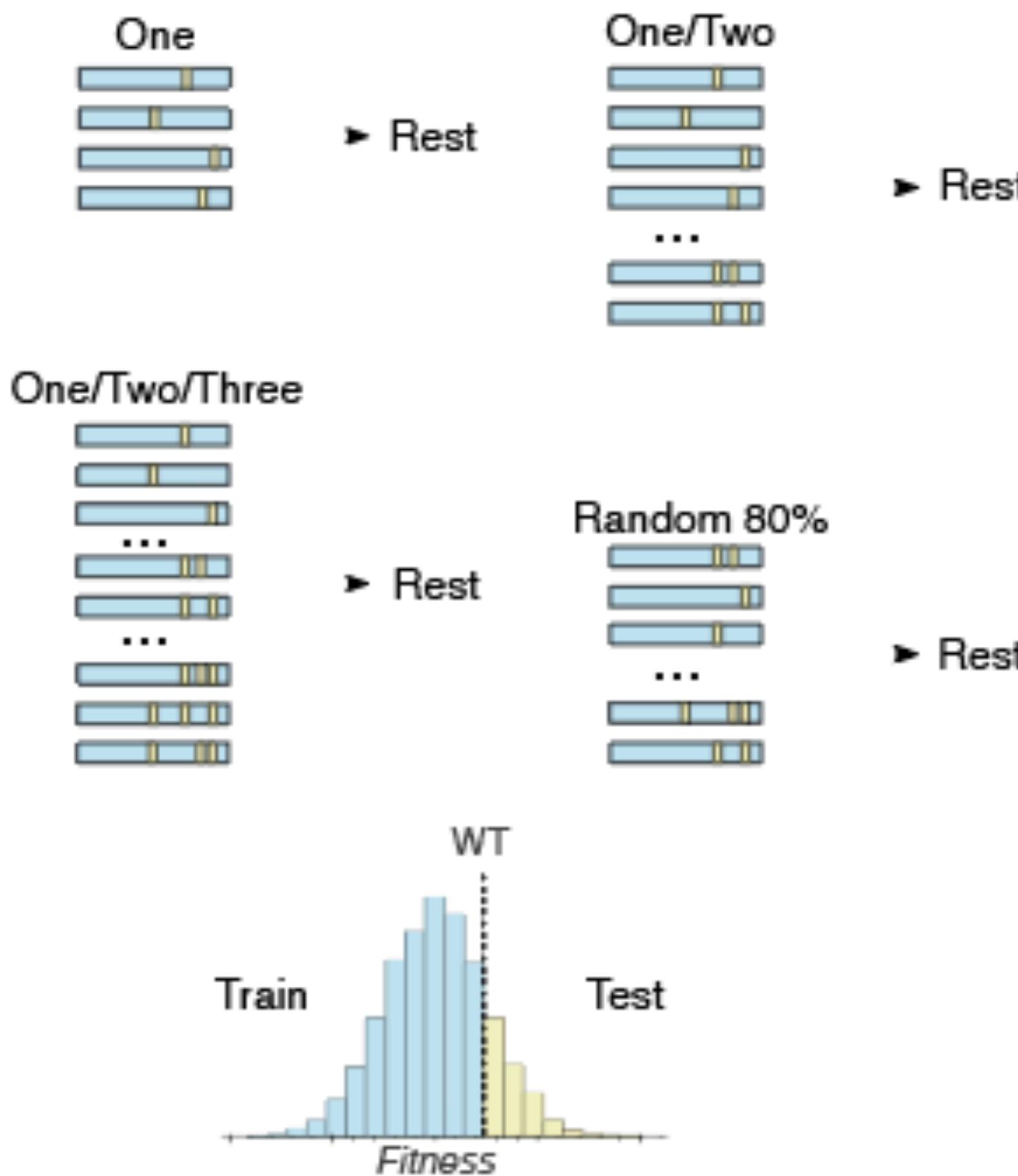


Protein engineering campaigns require out-of-domain generalization



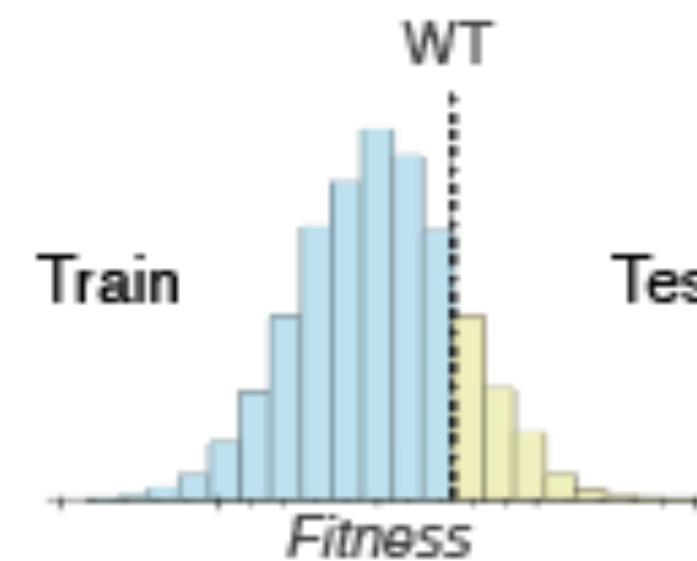
Can usually measure all single mutants

Protein engineering campaigns require out-of-domain generalization



Can usually measure all single mutants

Naively screen bigger spaces
->
mostly non-functional

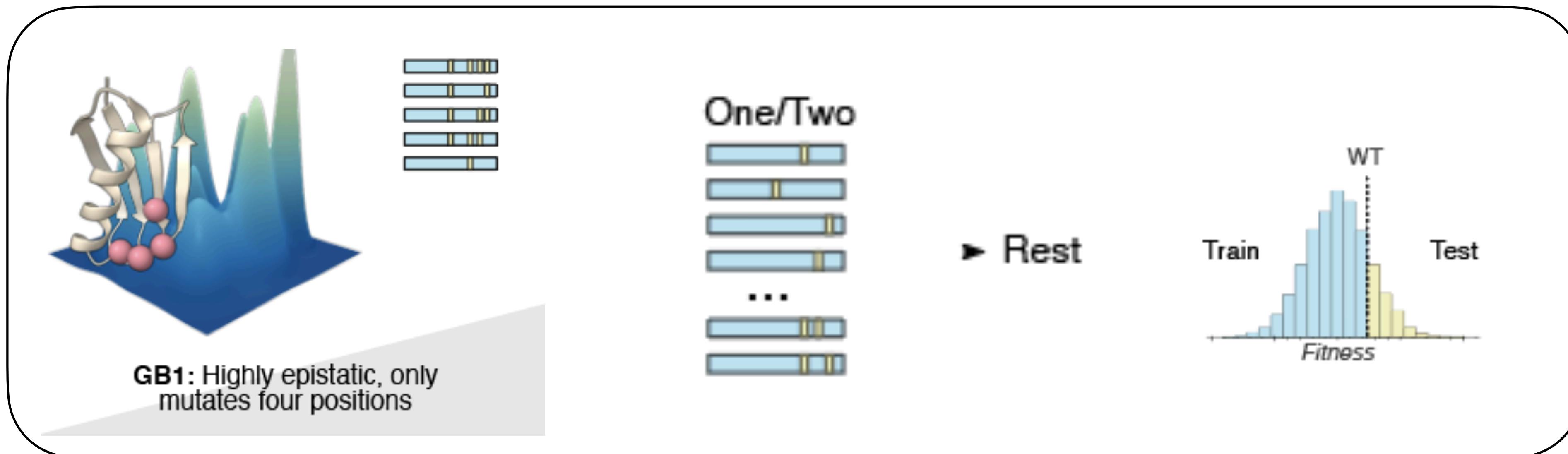


Protein engineering campaigns require out-of-domain generalization

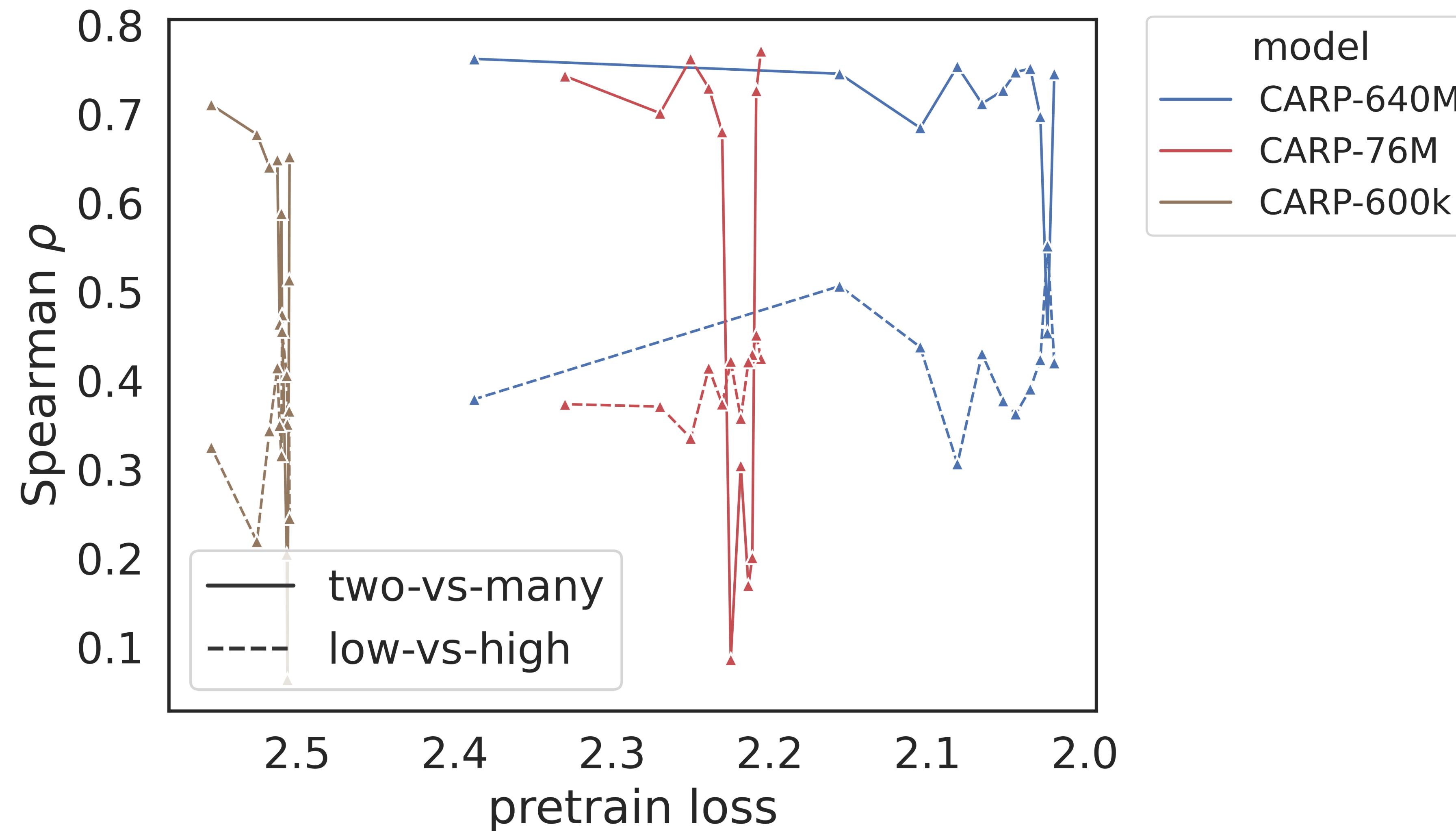
Protein engineering campaigns require out-of-domain generalization



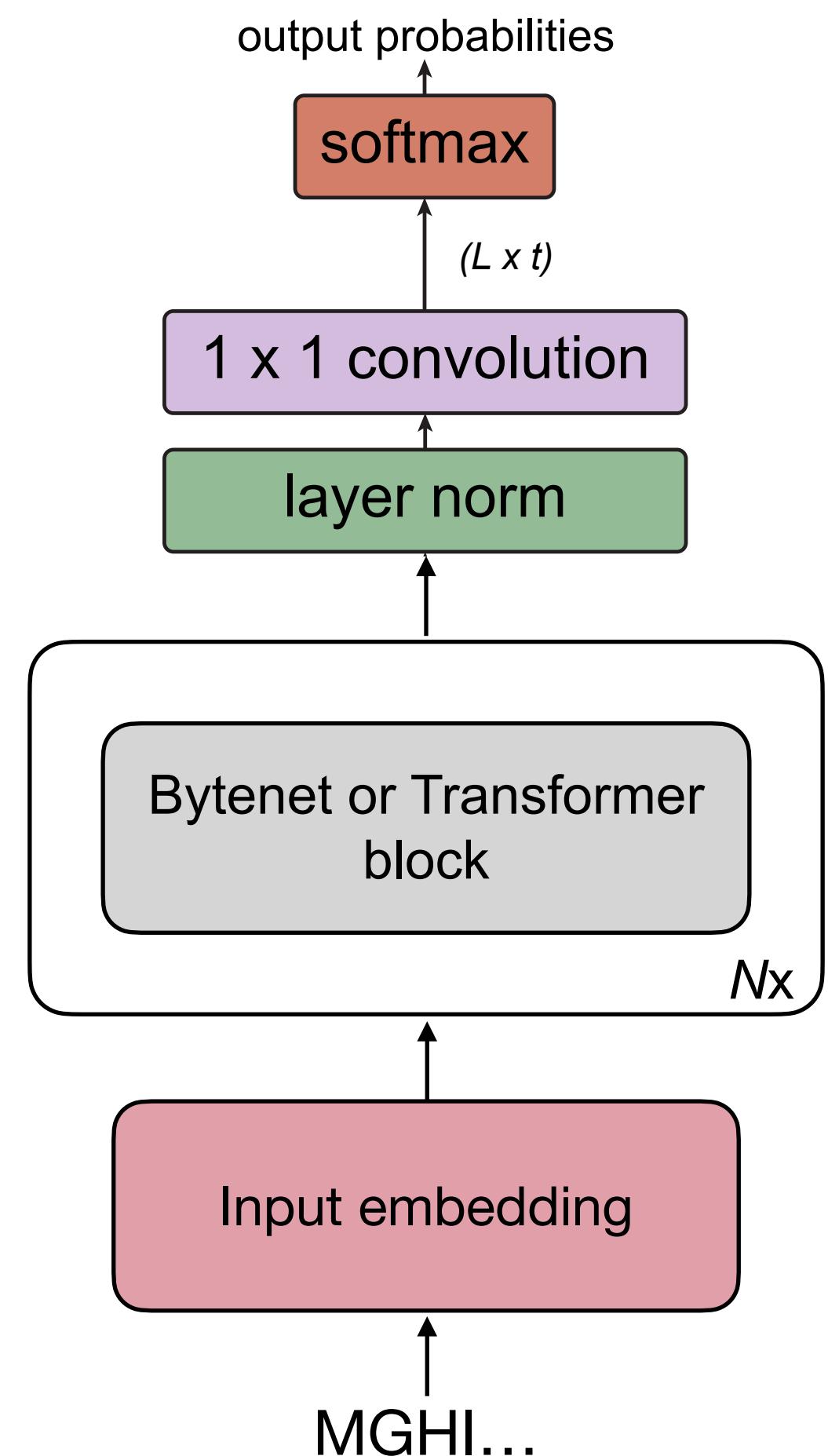
Protein engineering campaigns require out-of-domain generalization



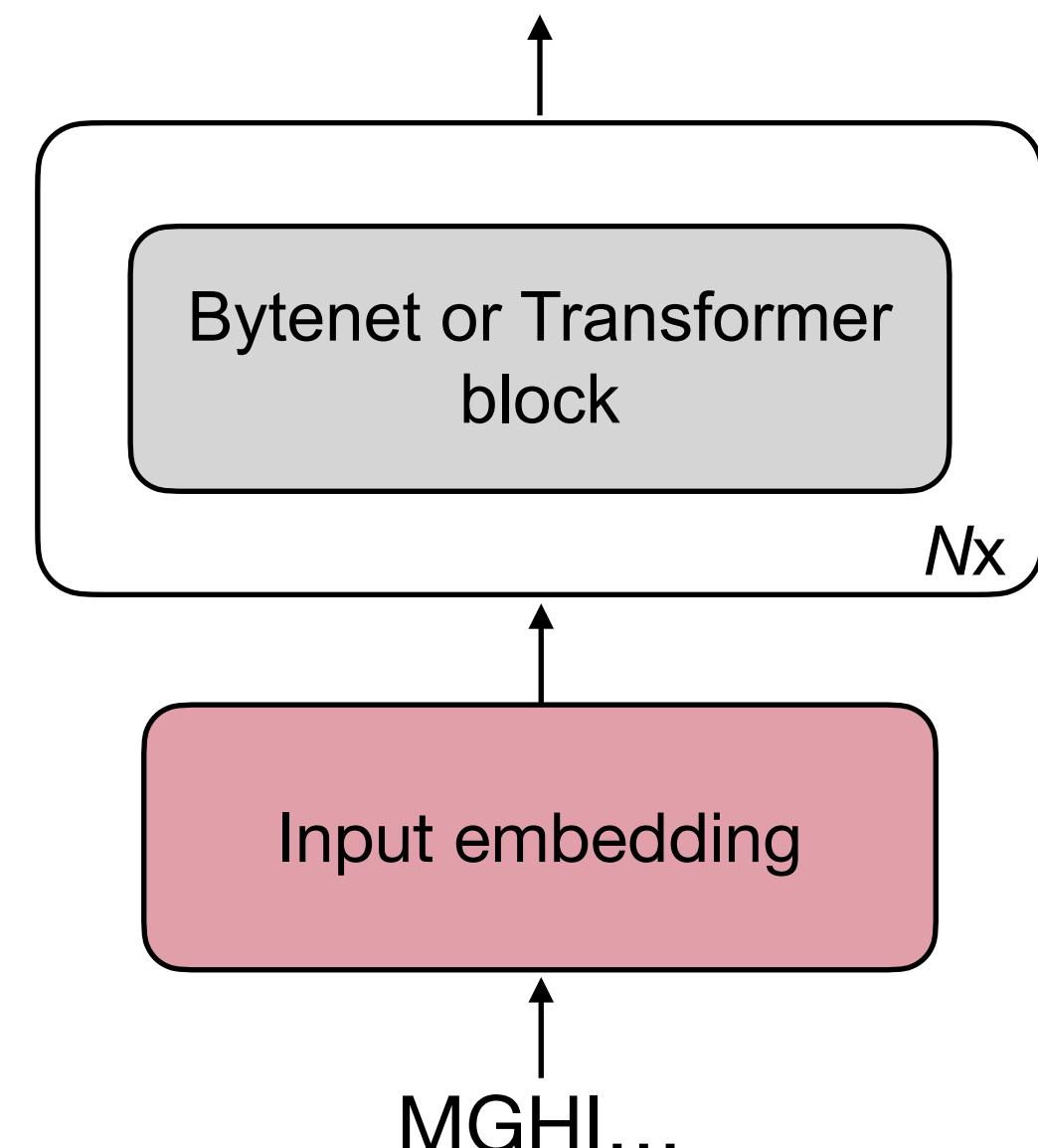
OOD performance does not improve with more pretraining



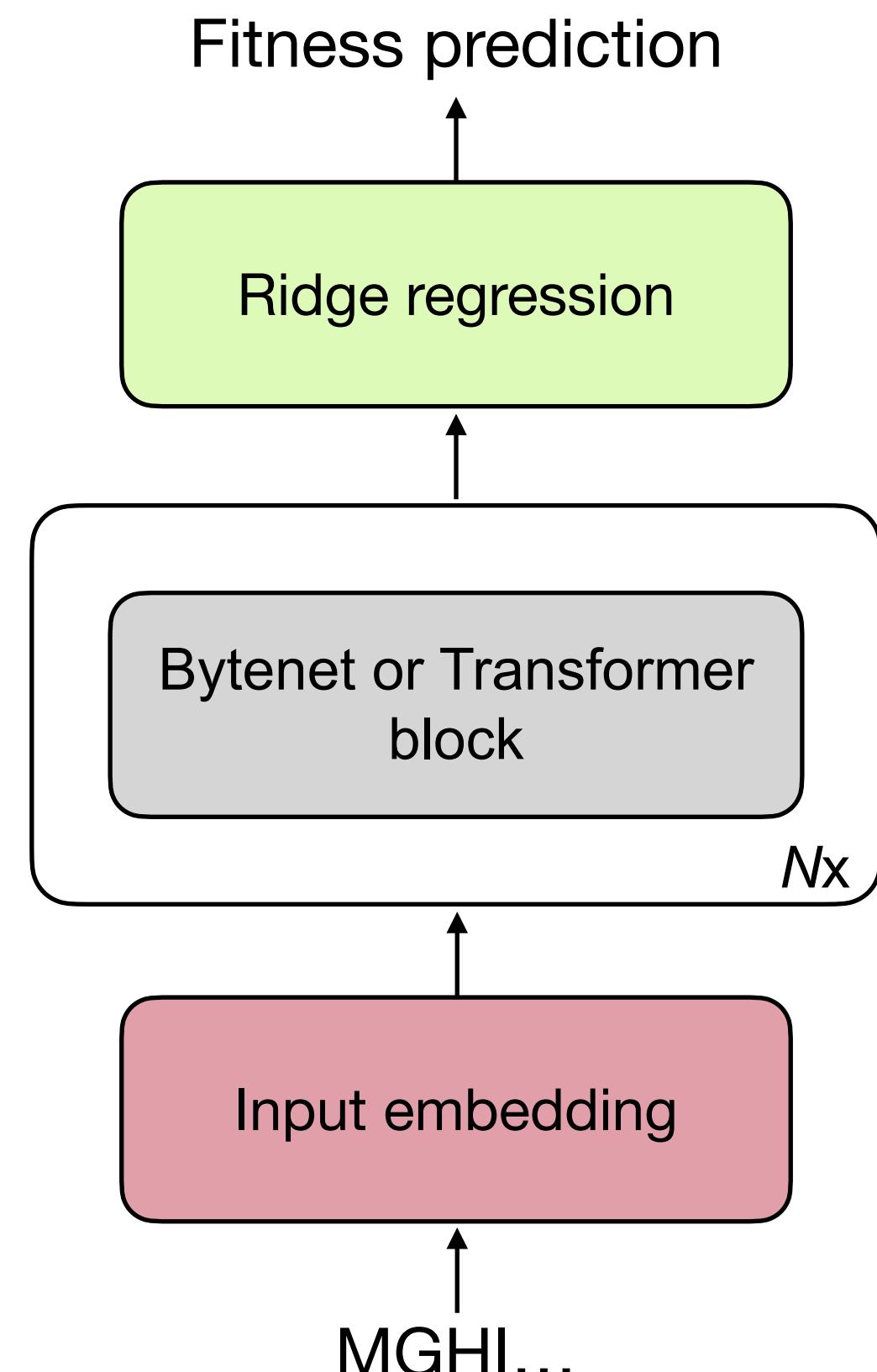
Not all layers are necessary for transfer



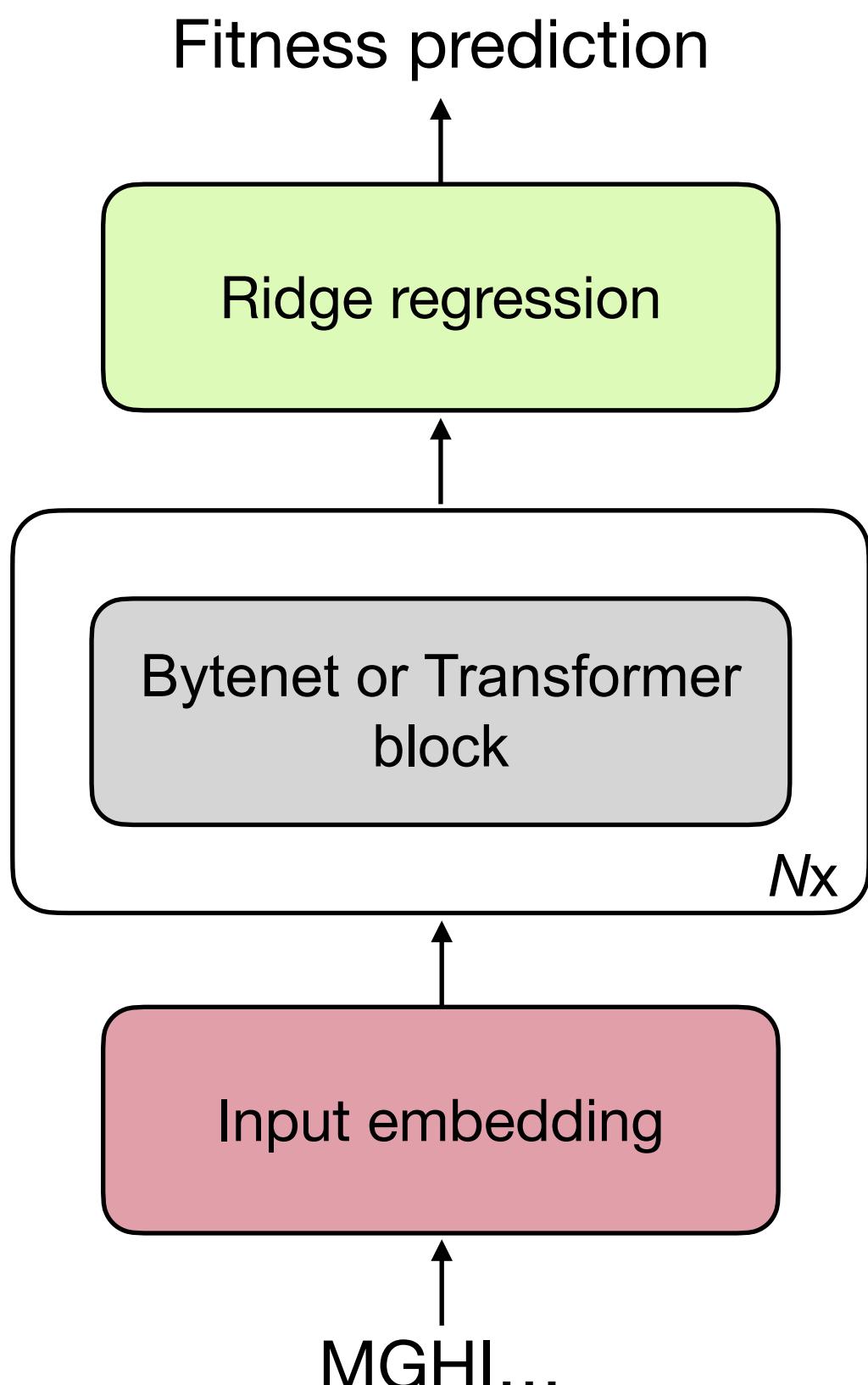
Not all layers are necessary for transfer



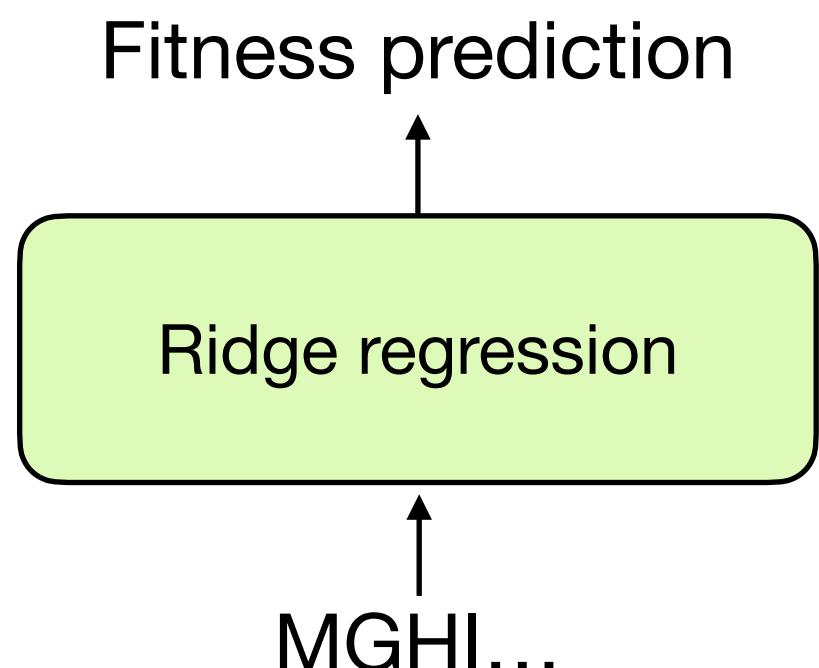
Not all layers are necessary for transfer



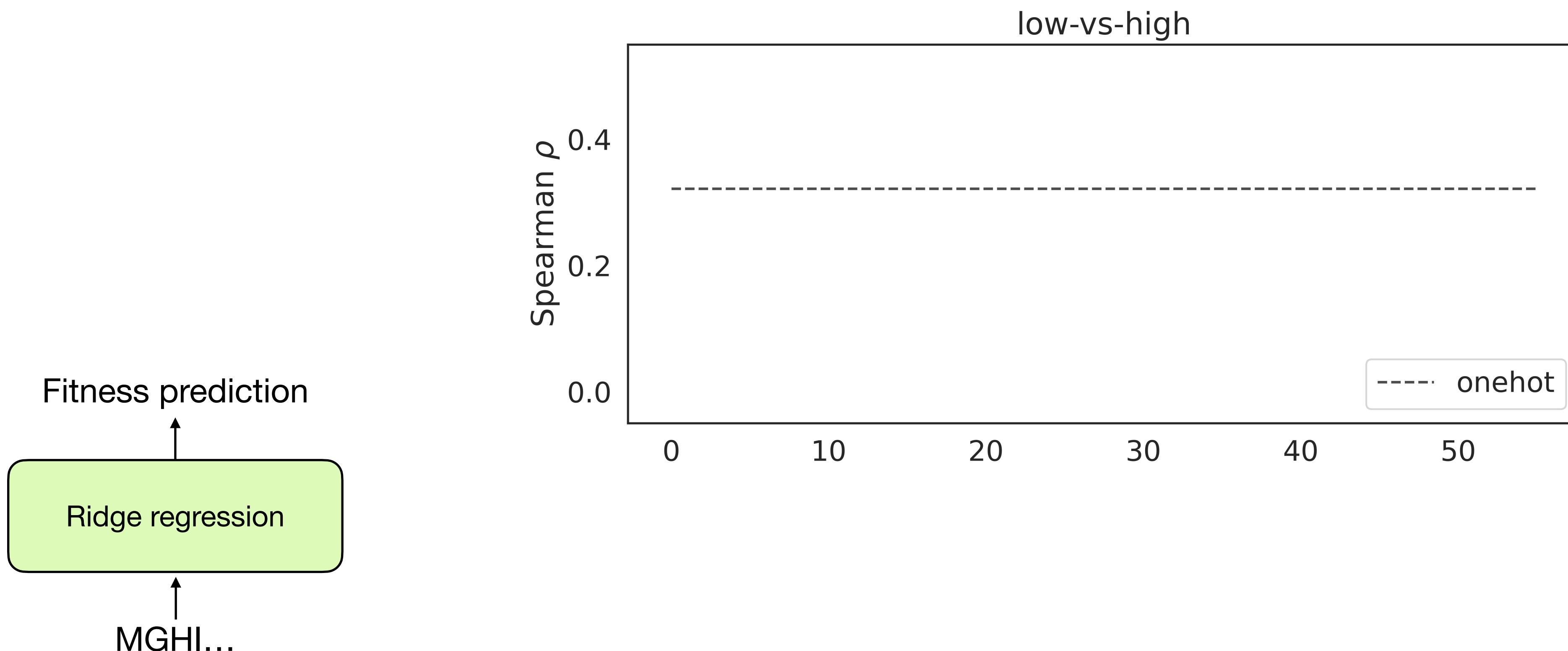
Not all layers are necessary for transfer



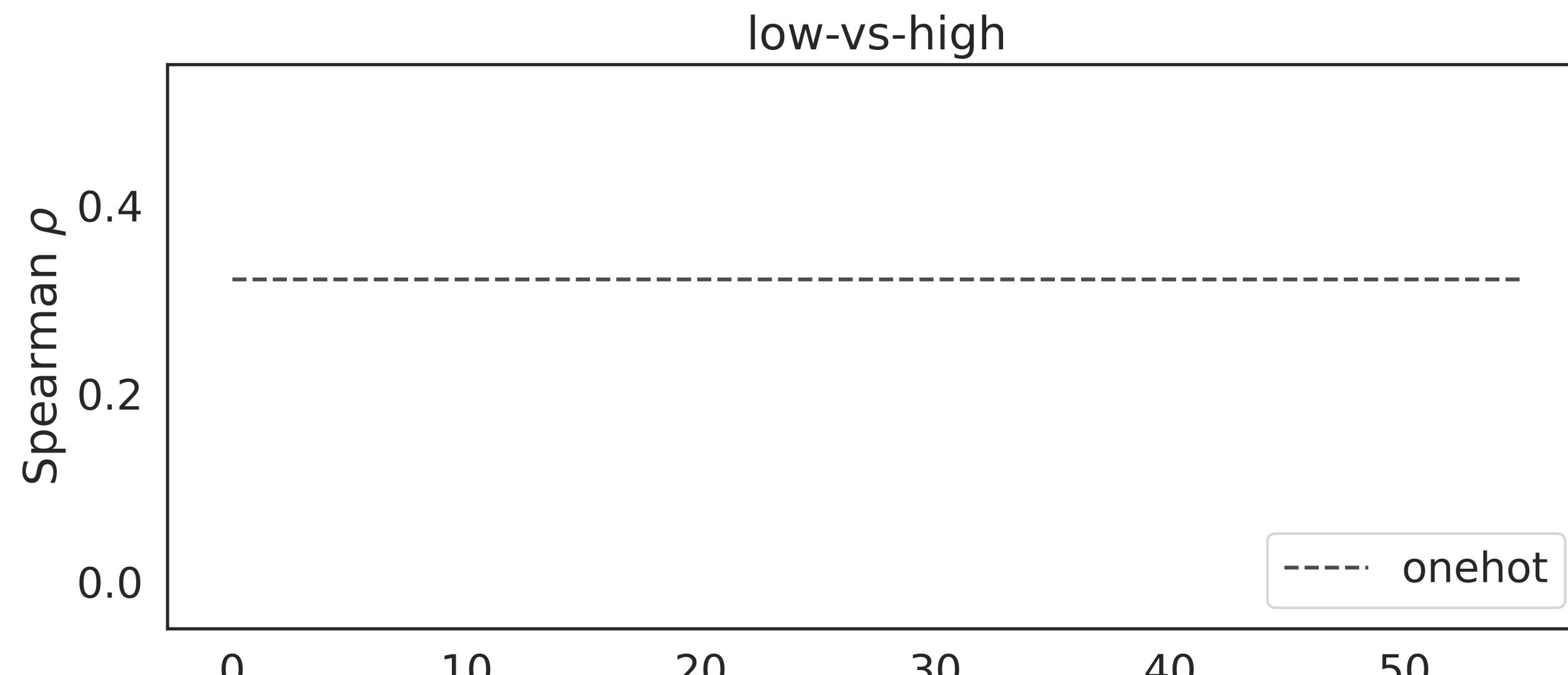
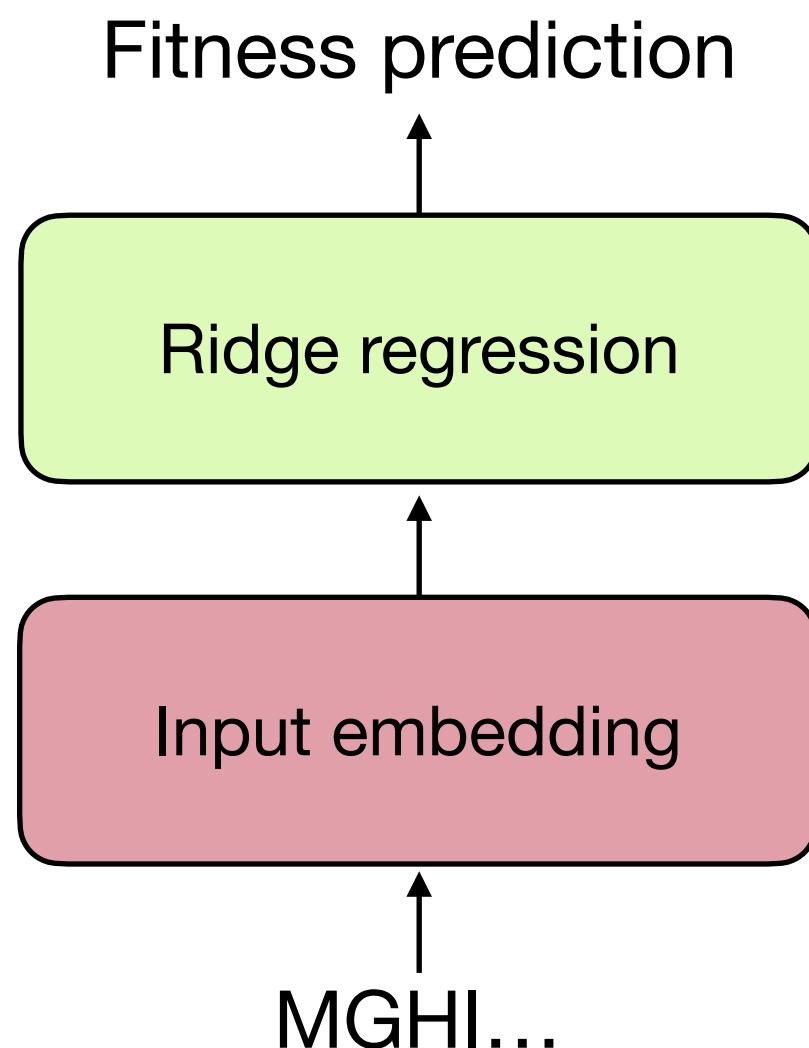
Not all layers are necessary for transfer



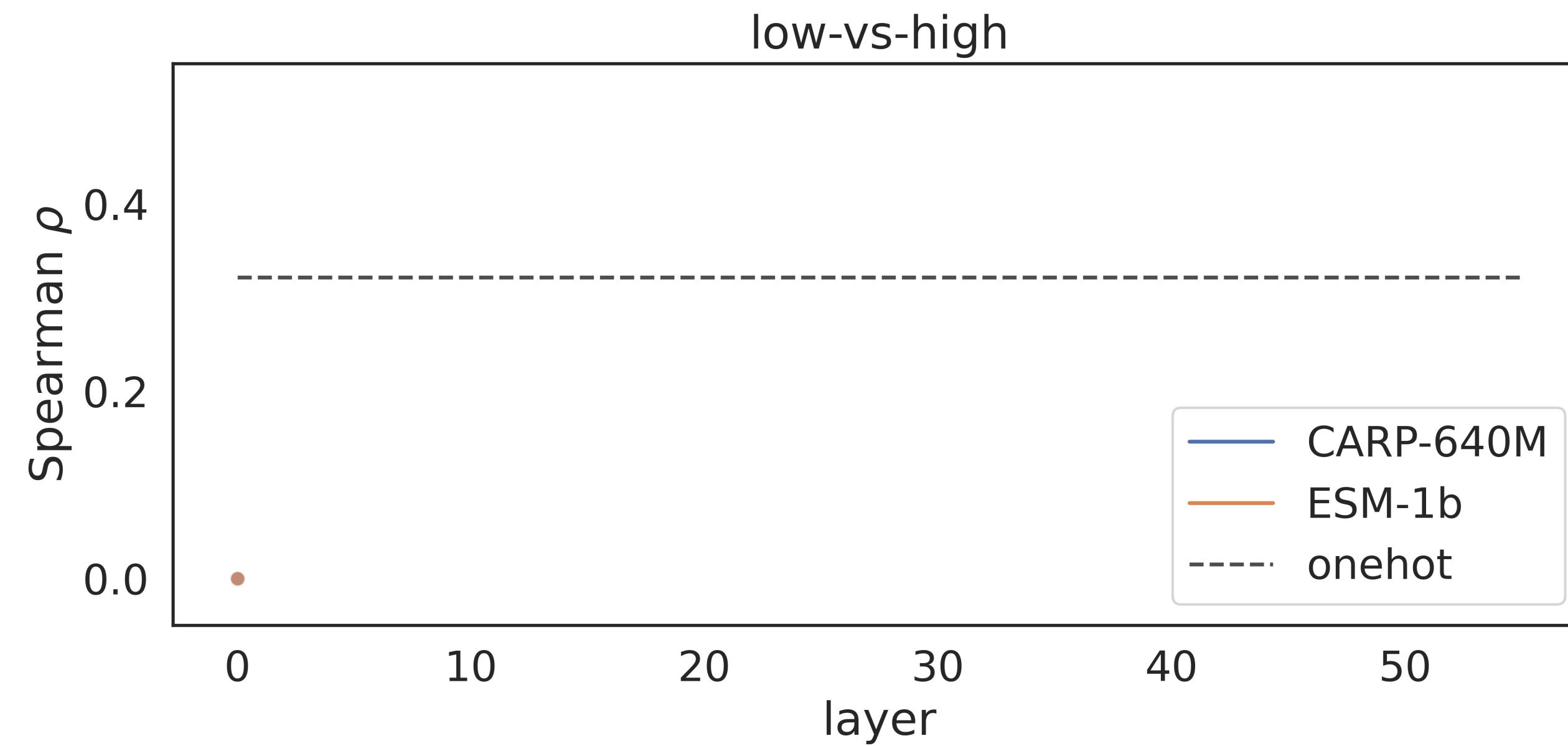
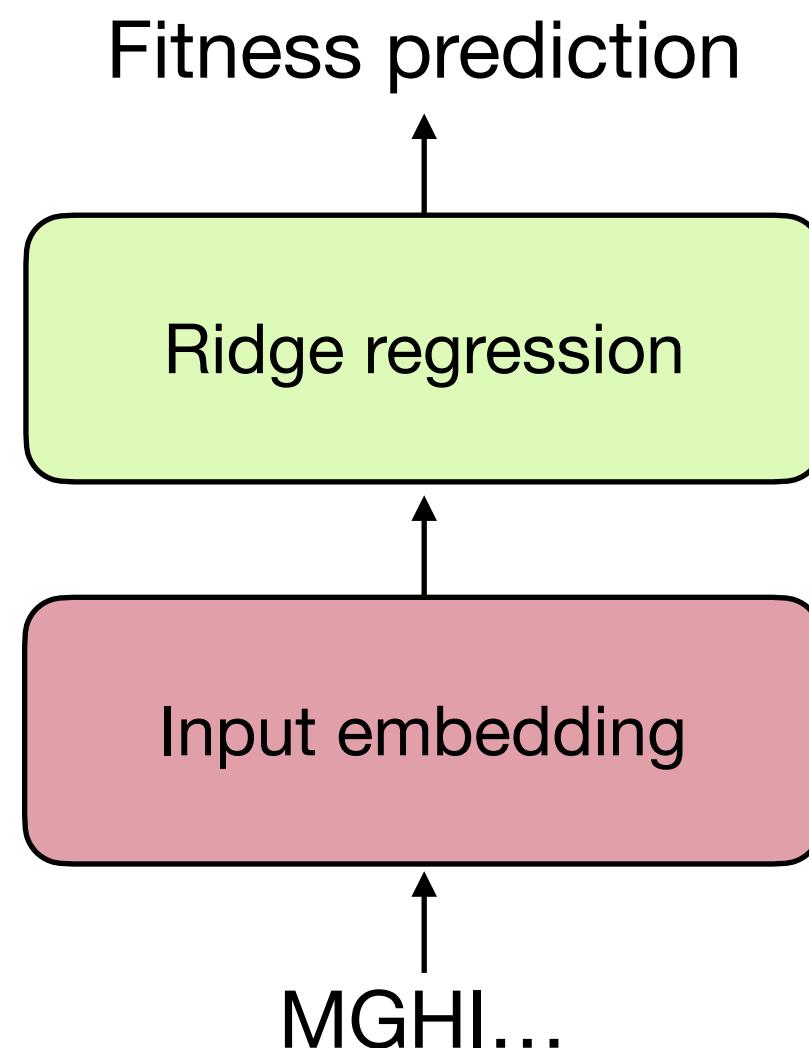
Not all layers are necessary for transfer



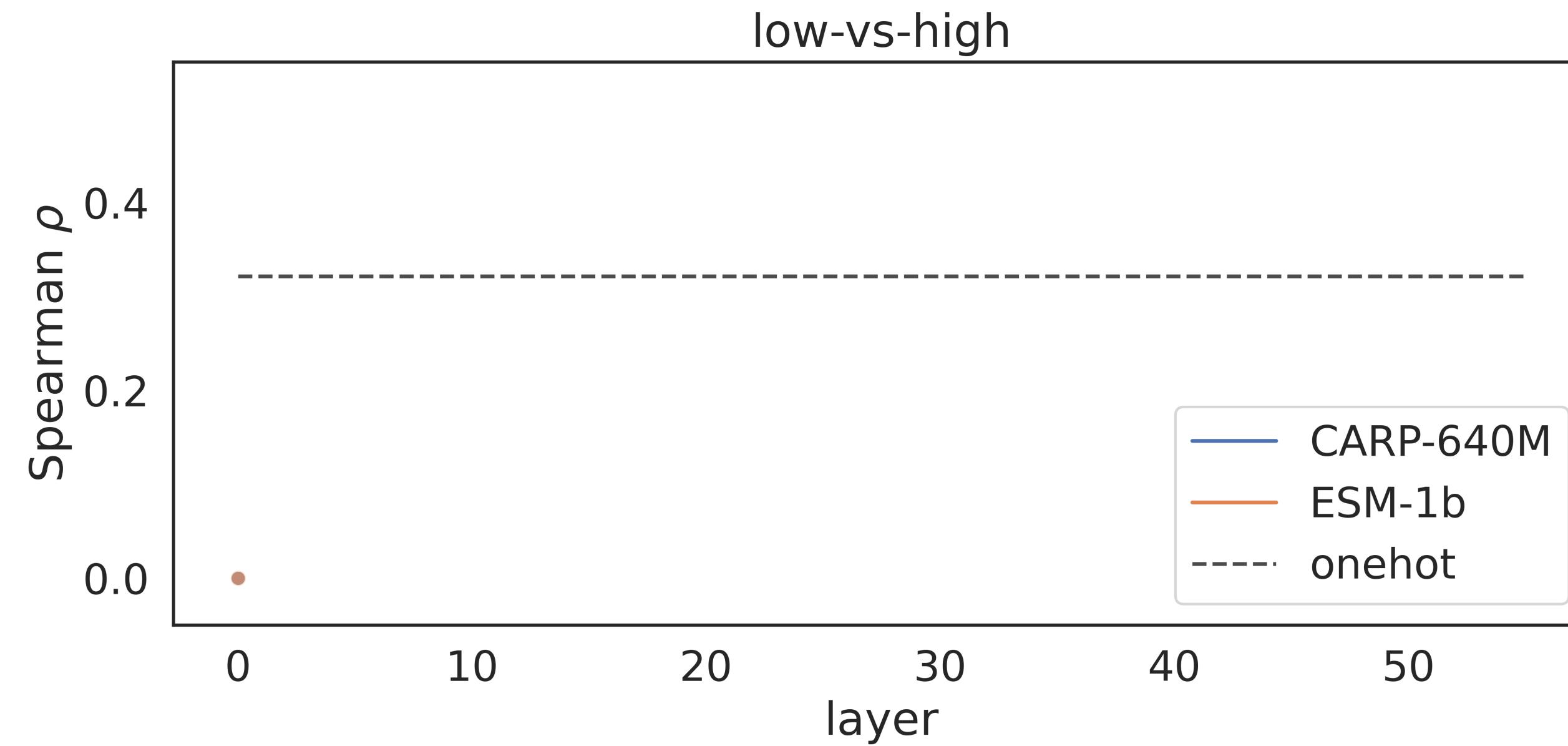
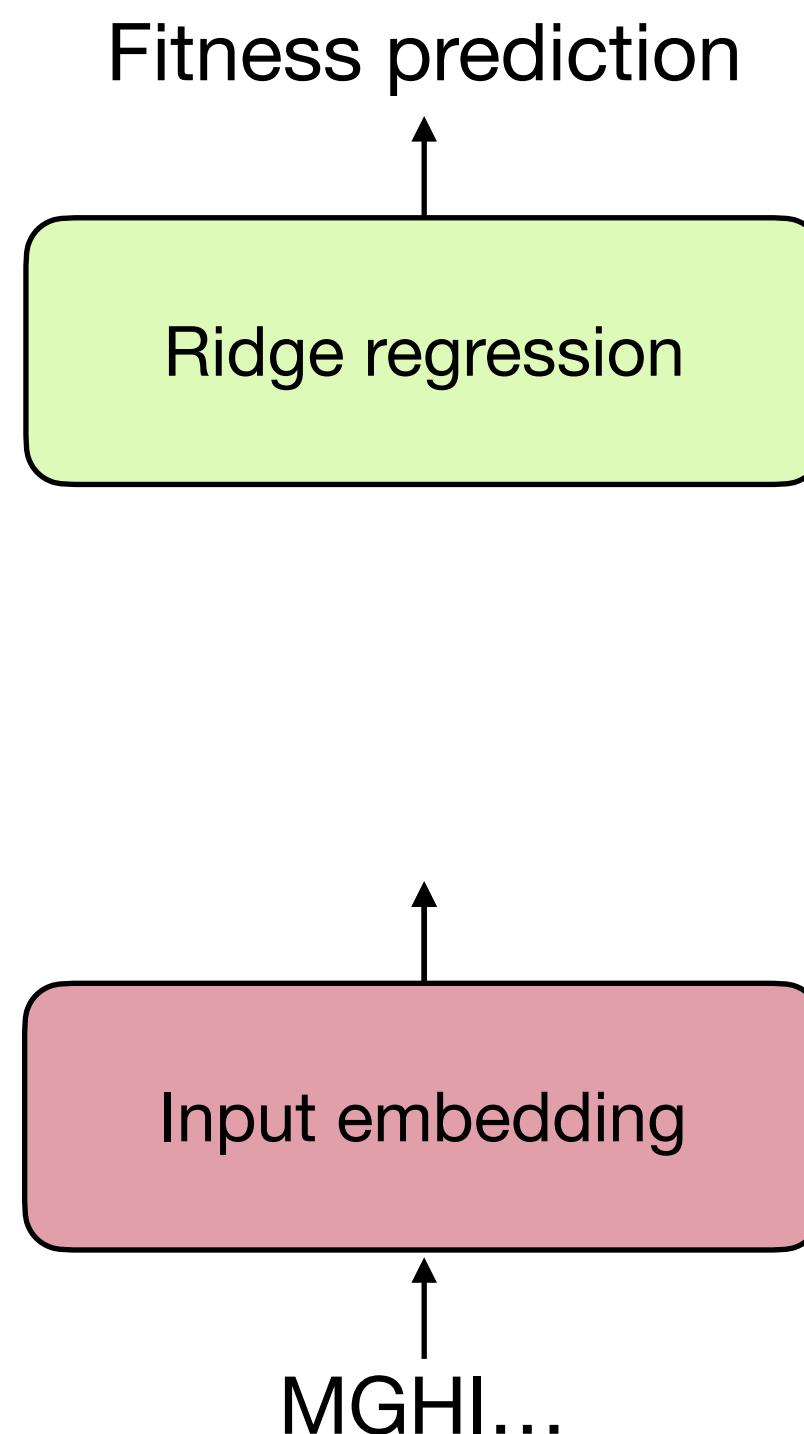
Not all layers are necessary for transfer



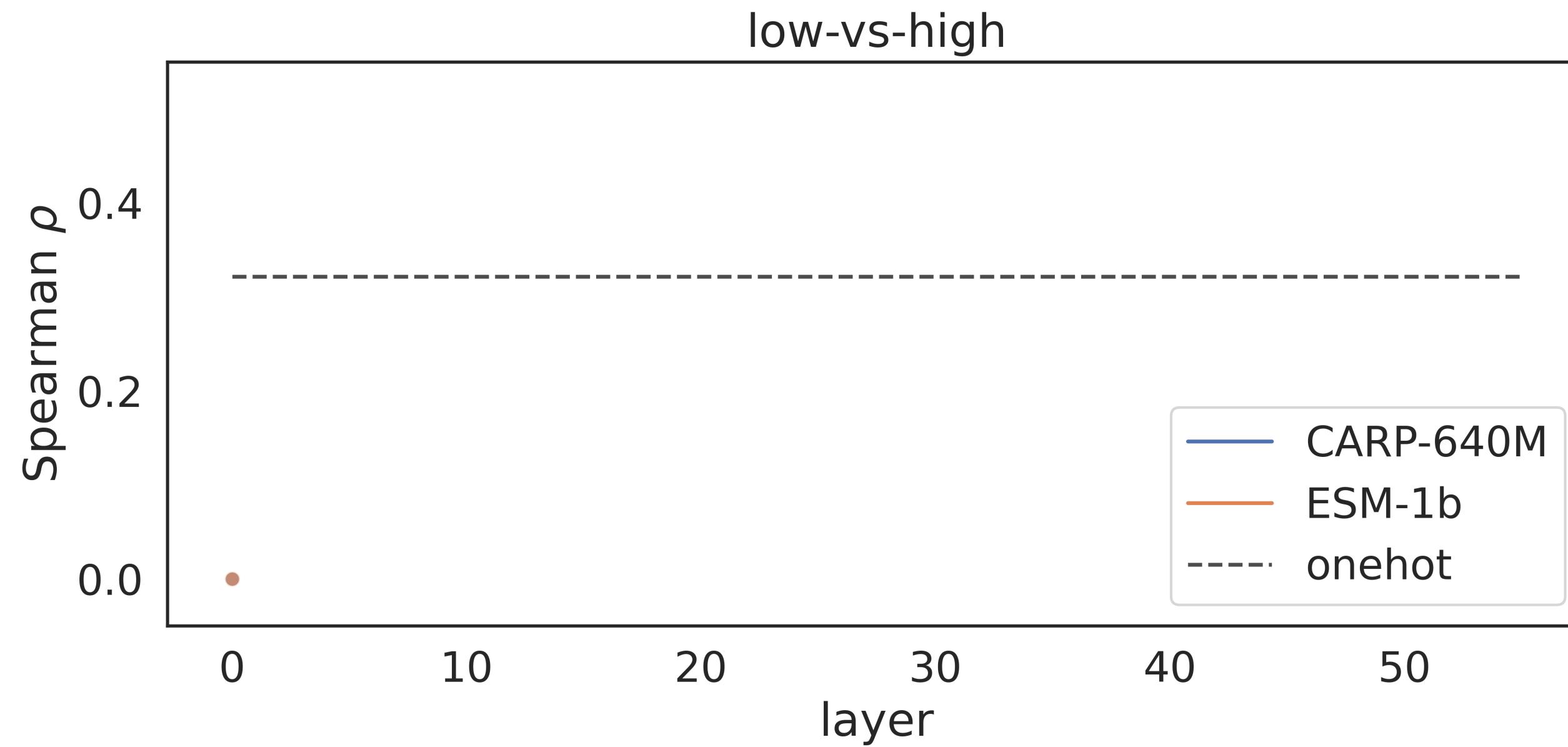
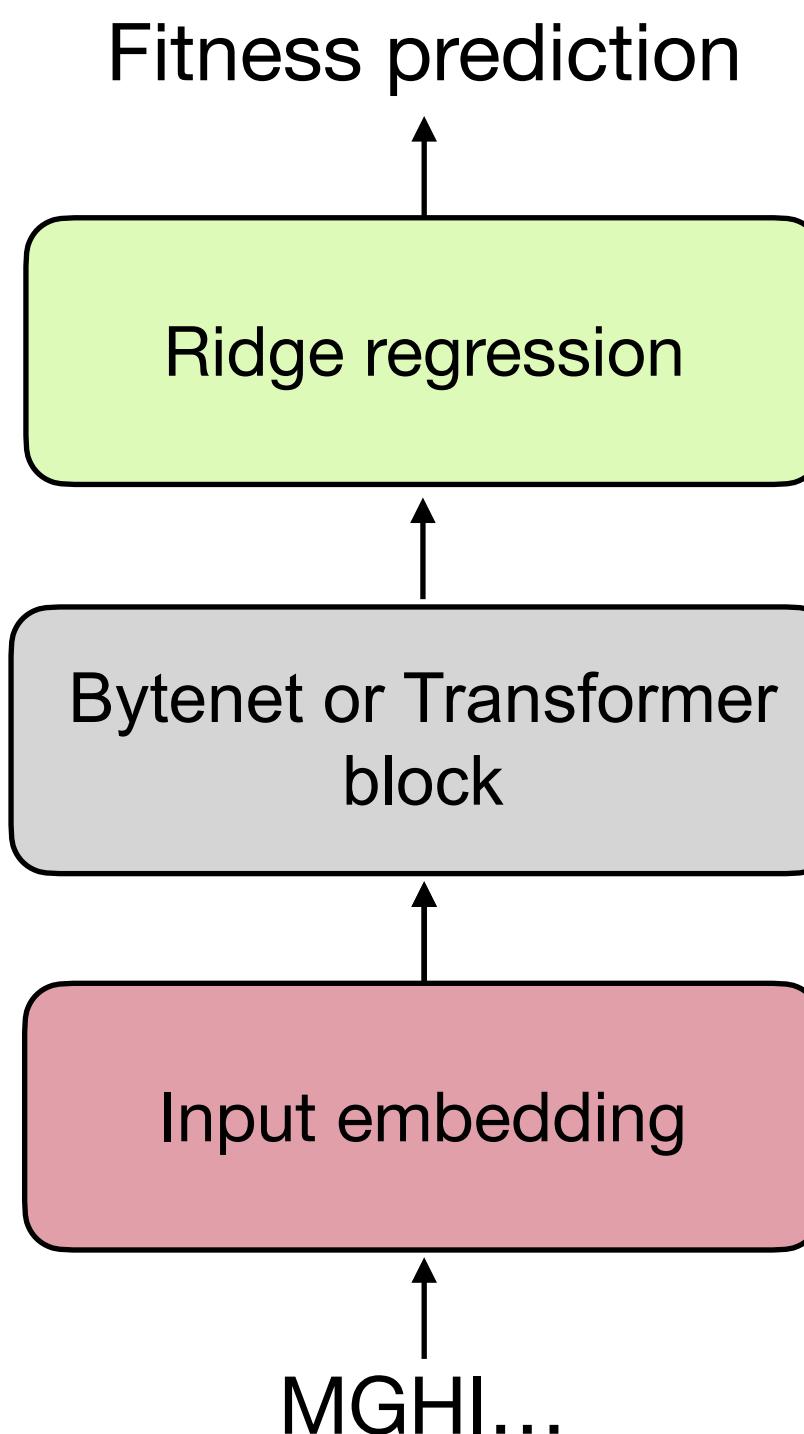
Not all layers are necessary for transfer



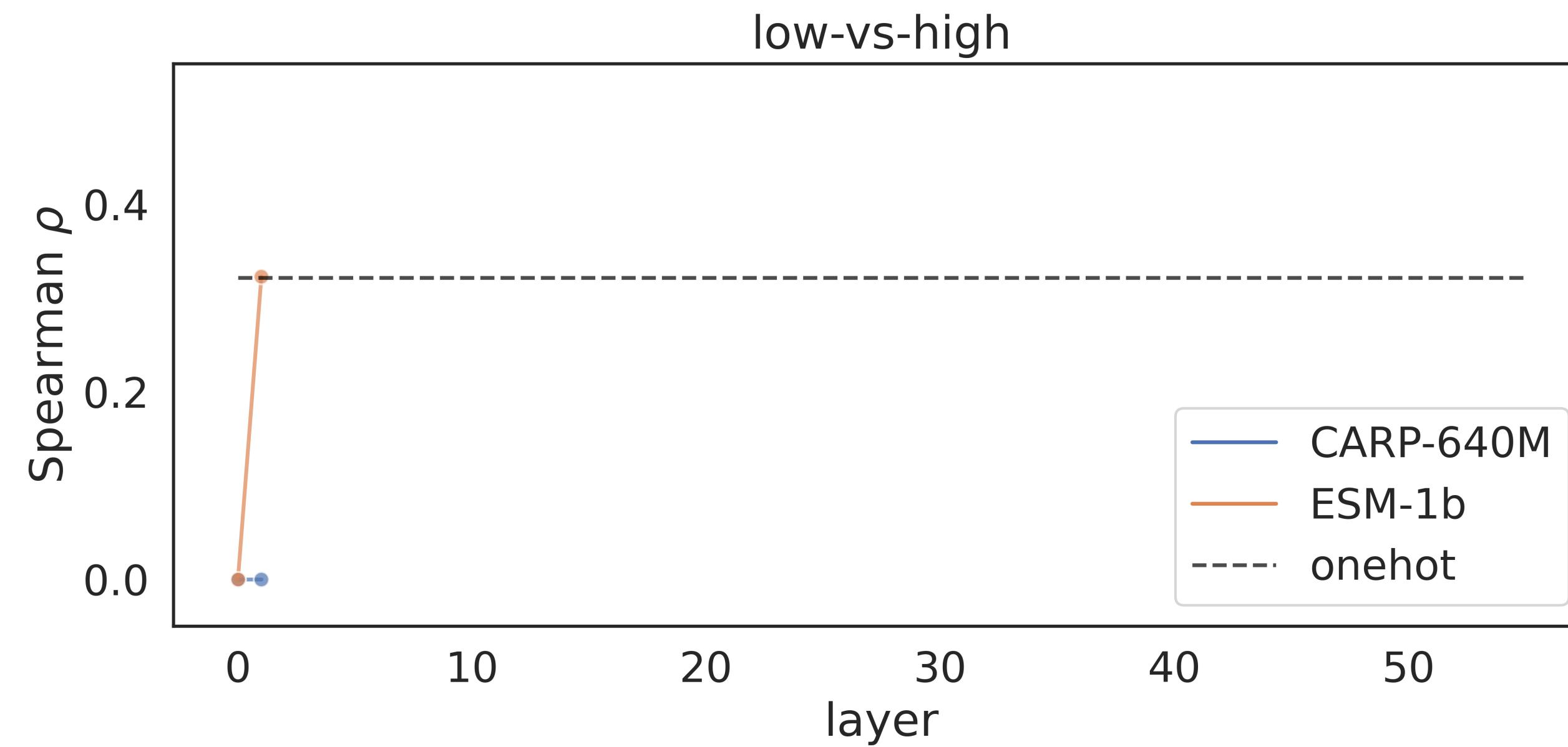
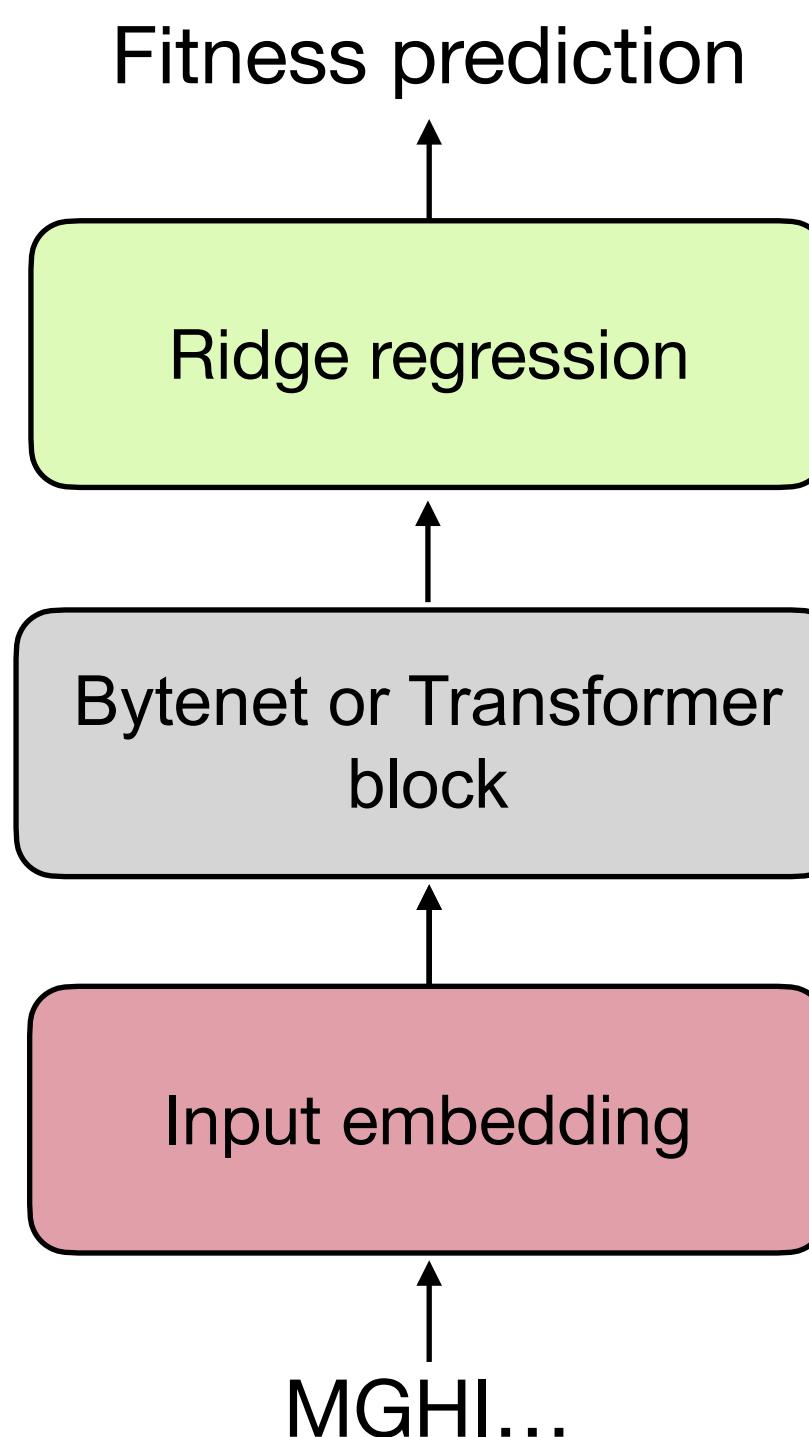
Not all layers are necessary for transfer



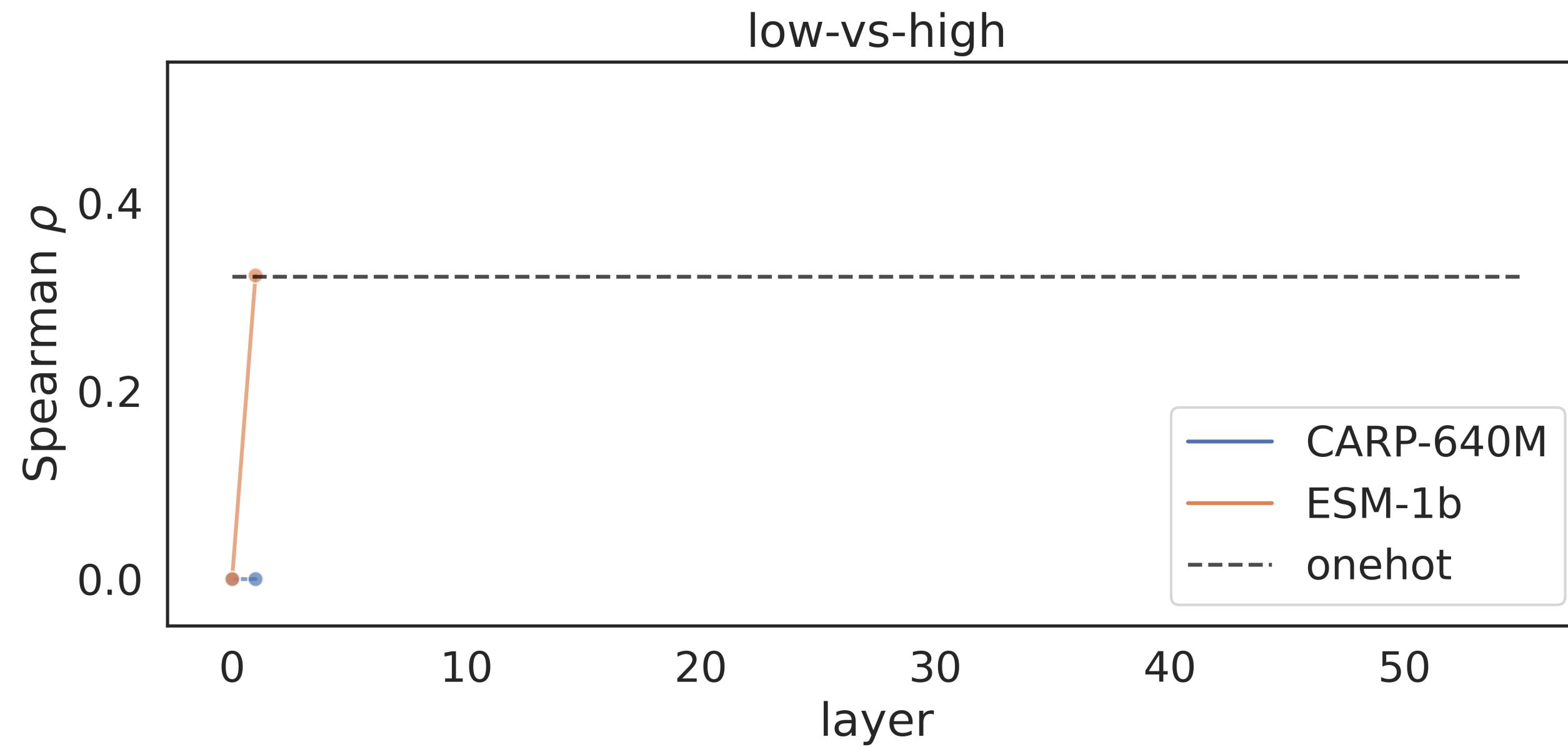
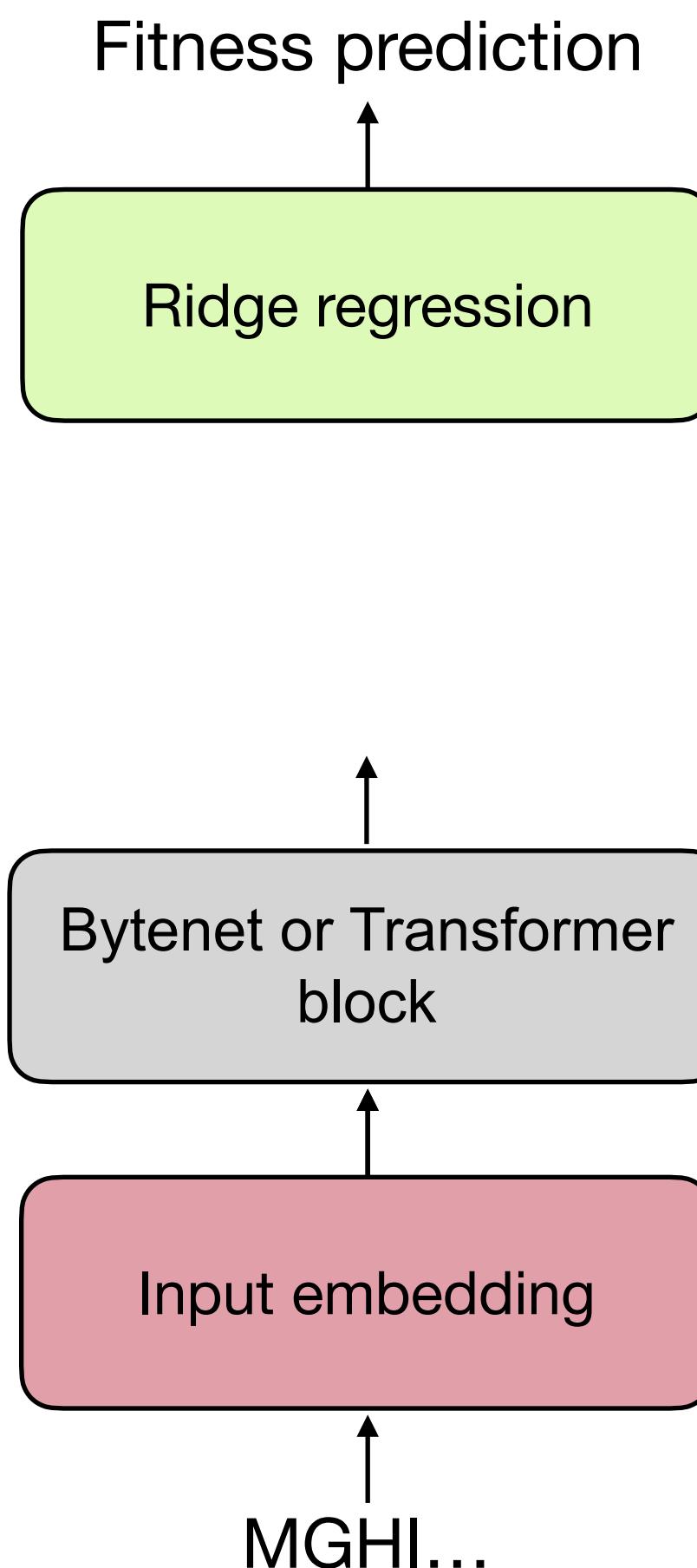
Not all layers are necessary for transfer



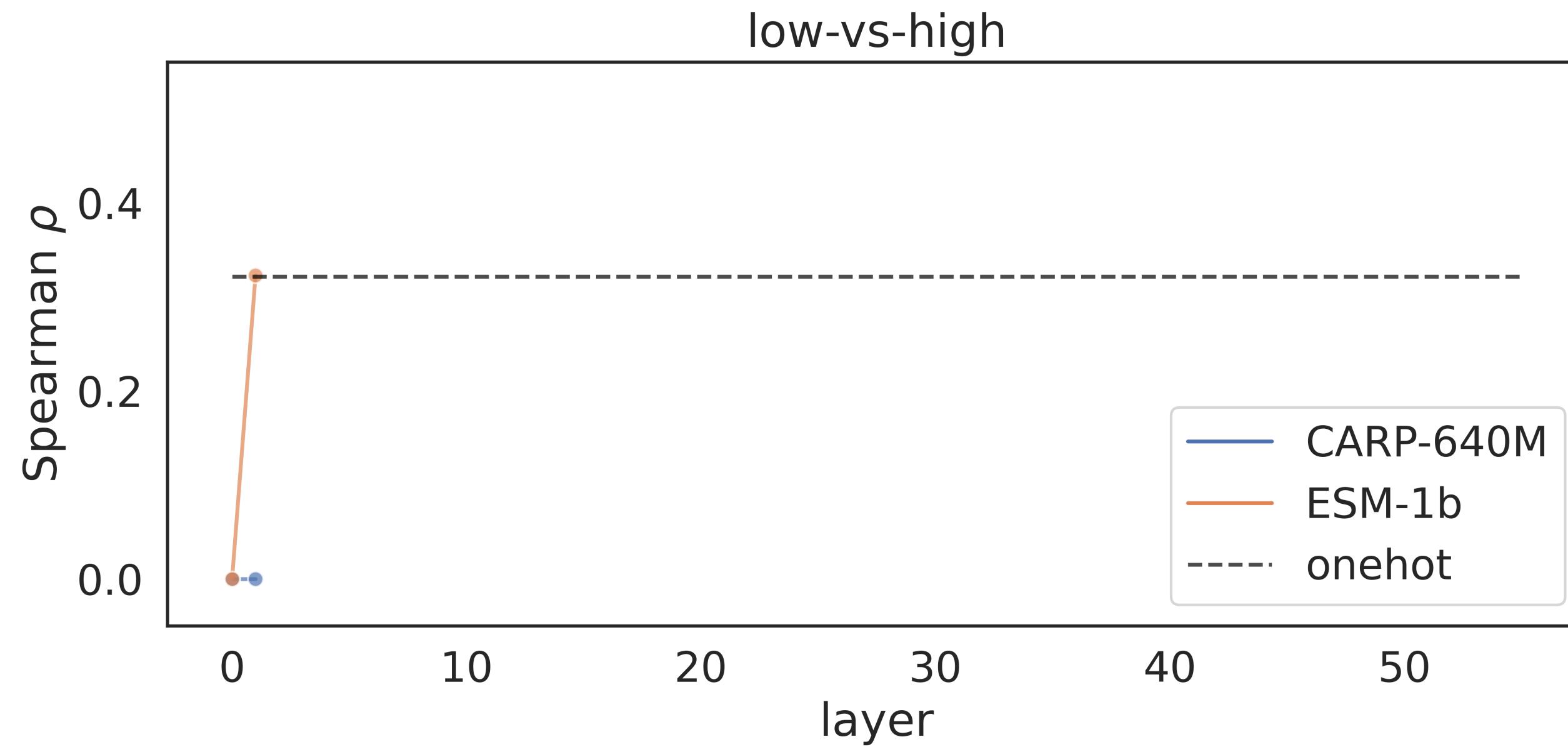
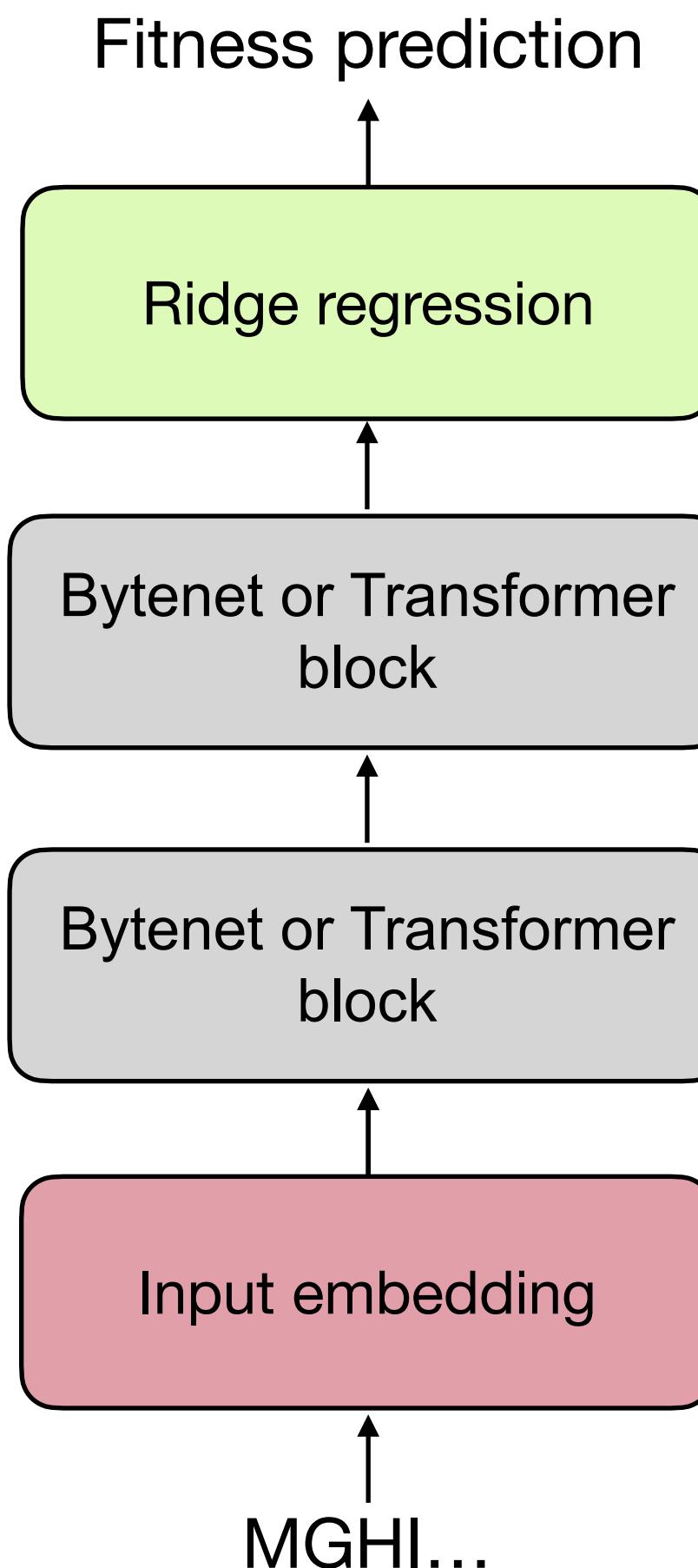
Not all layers are necessary for transfer



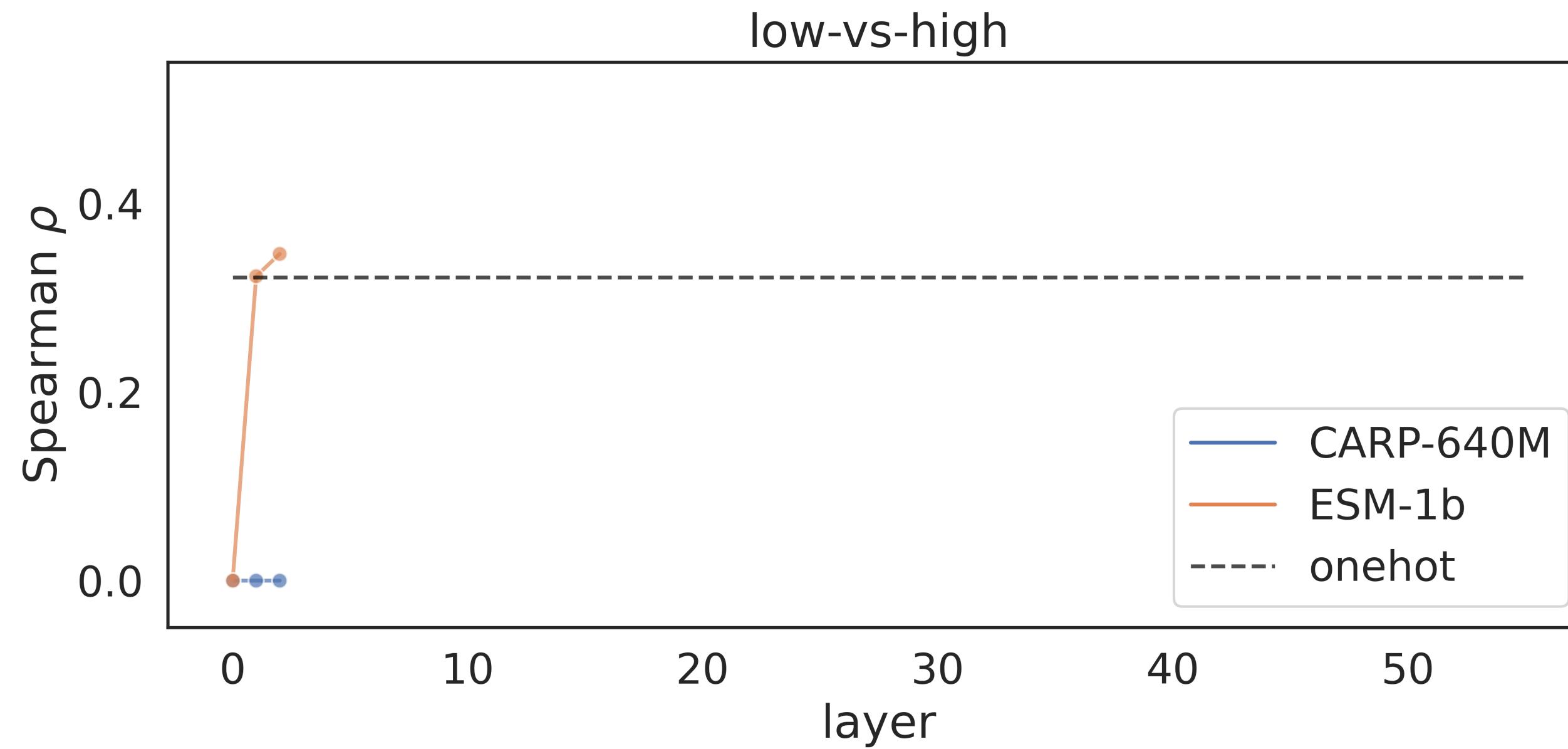
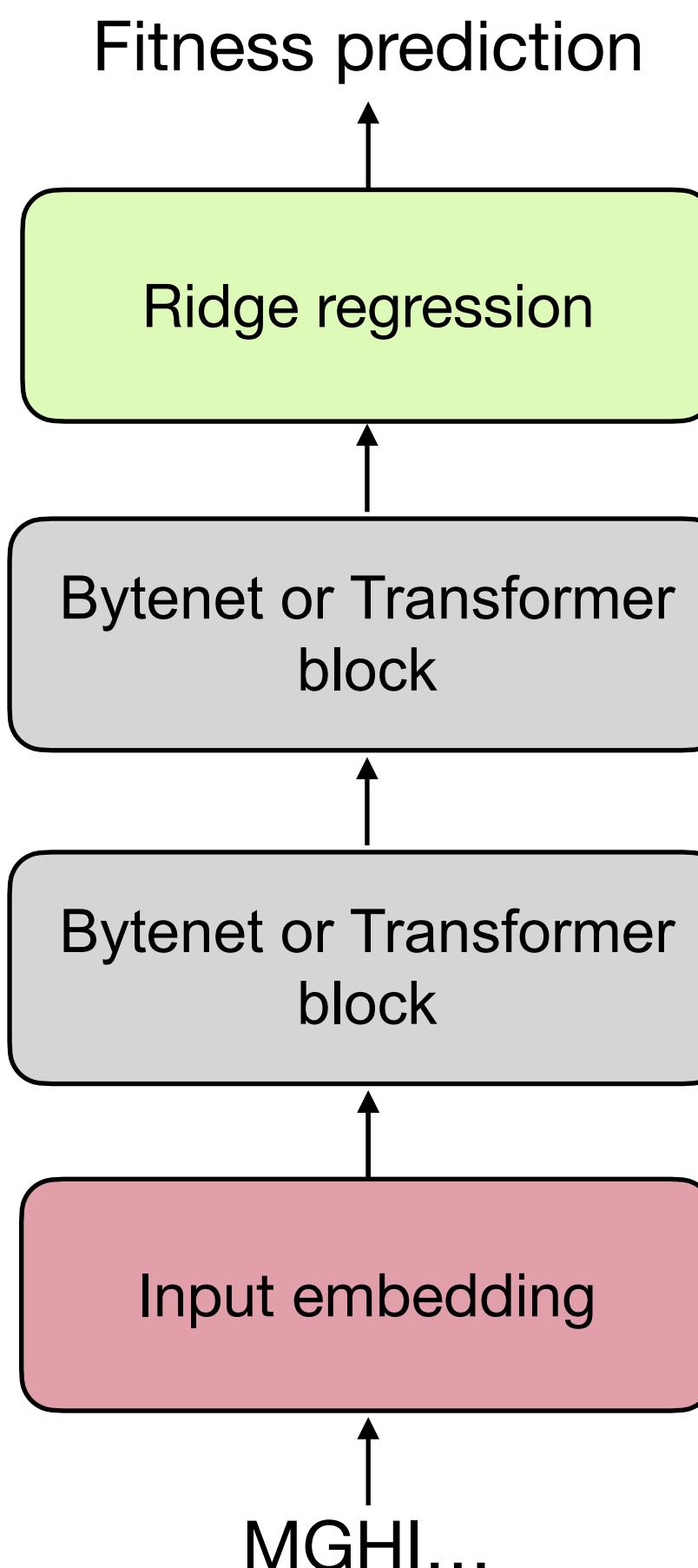
Not all layers are necessary for transfer



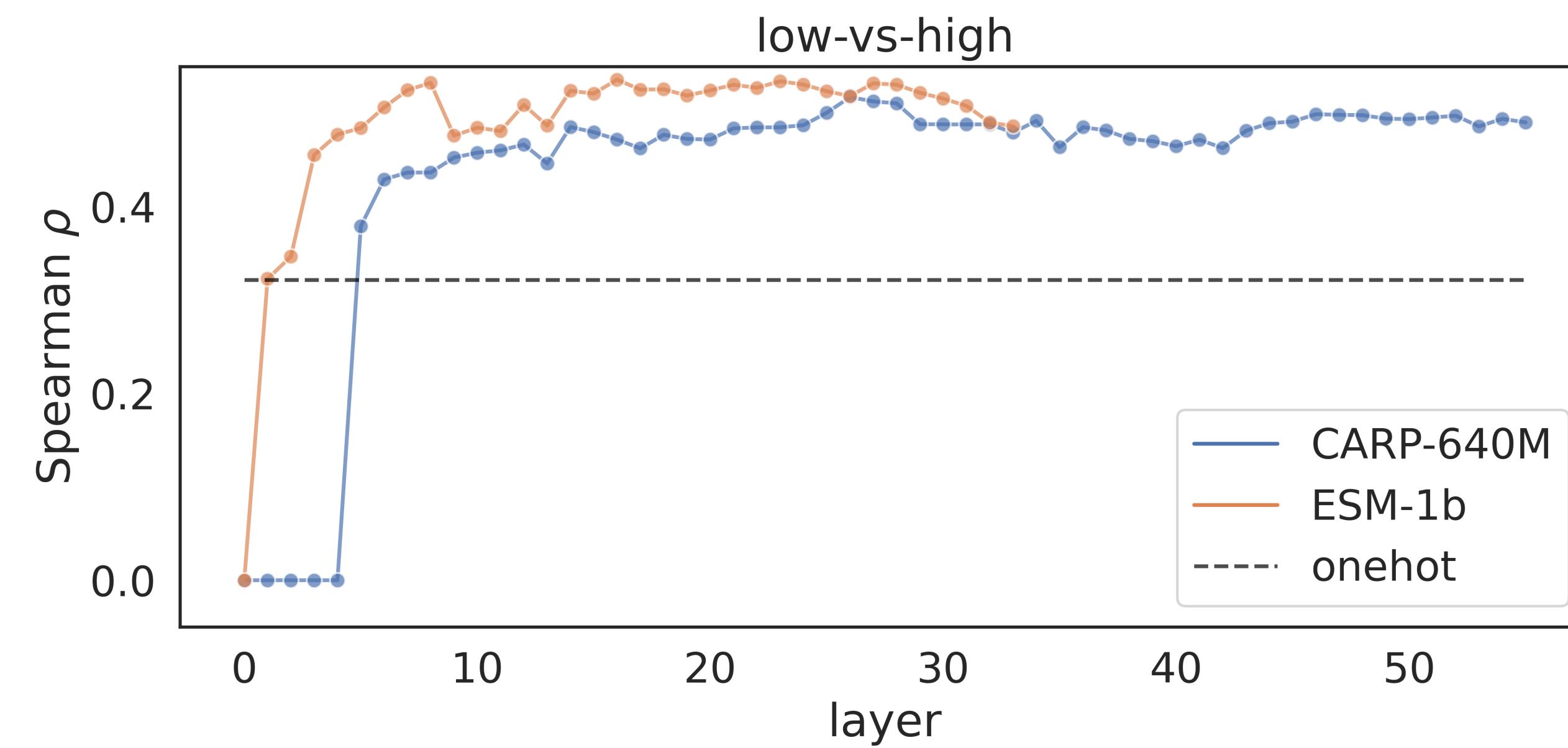
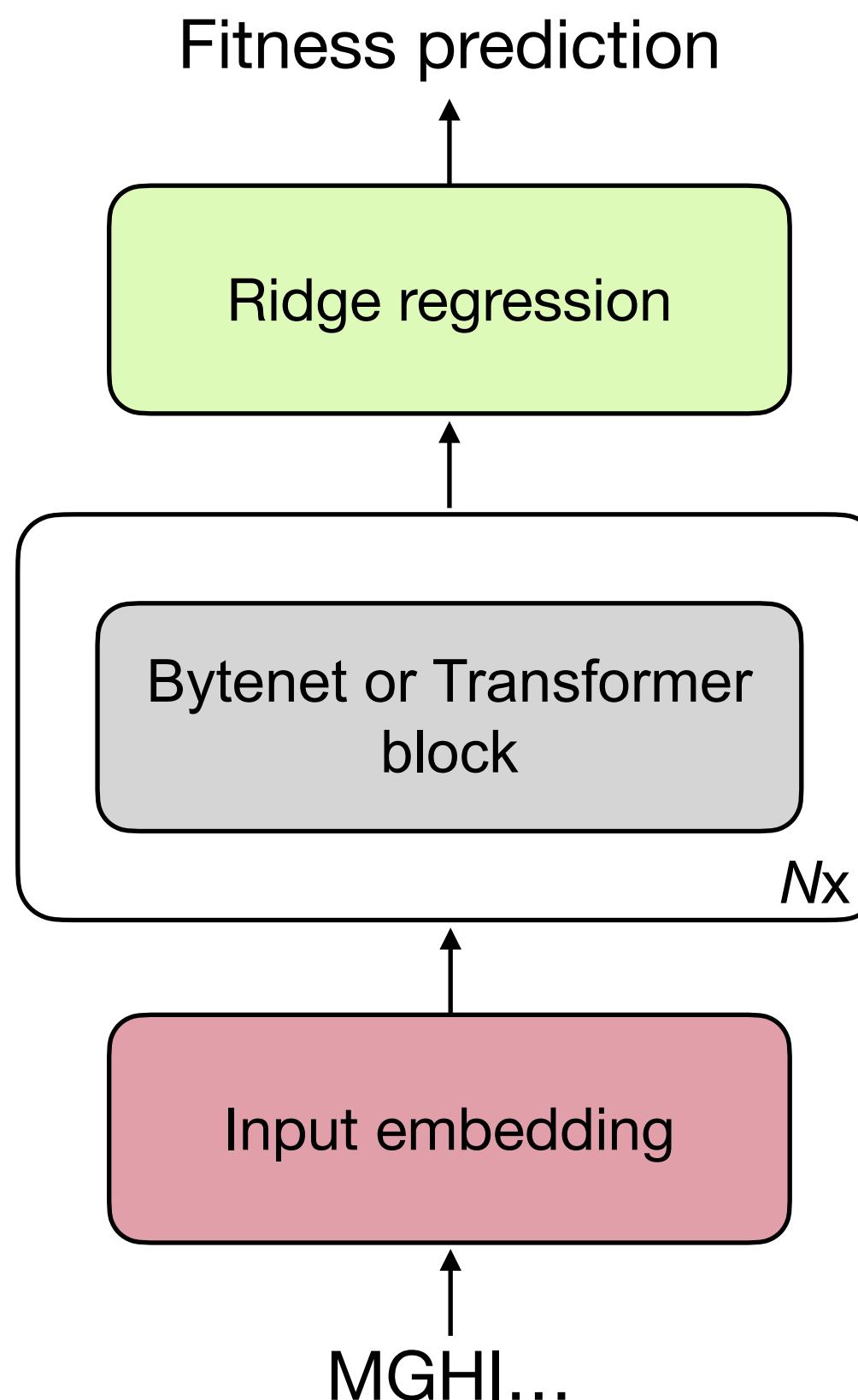
Not all layers are necessary for transfer



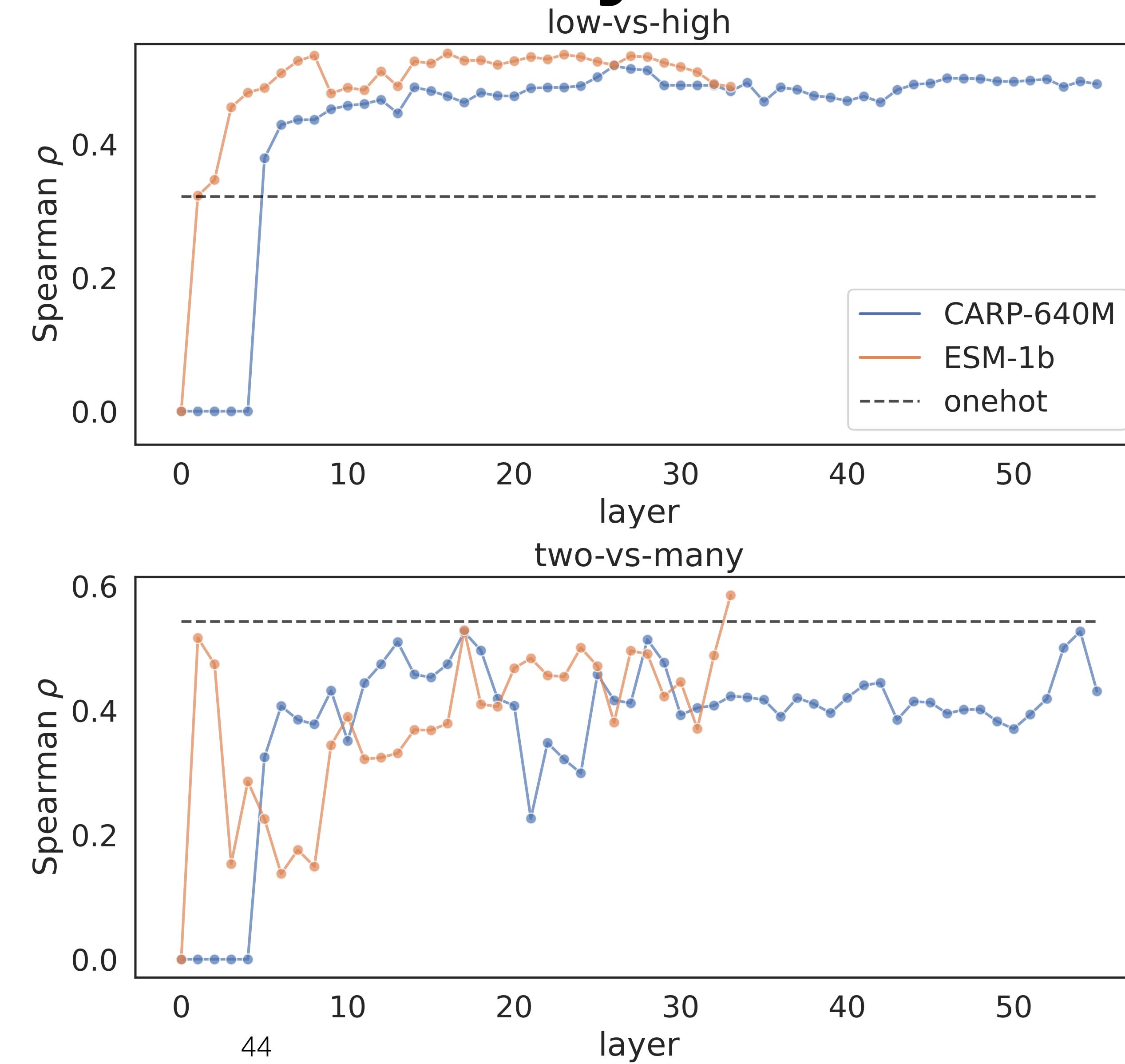
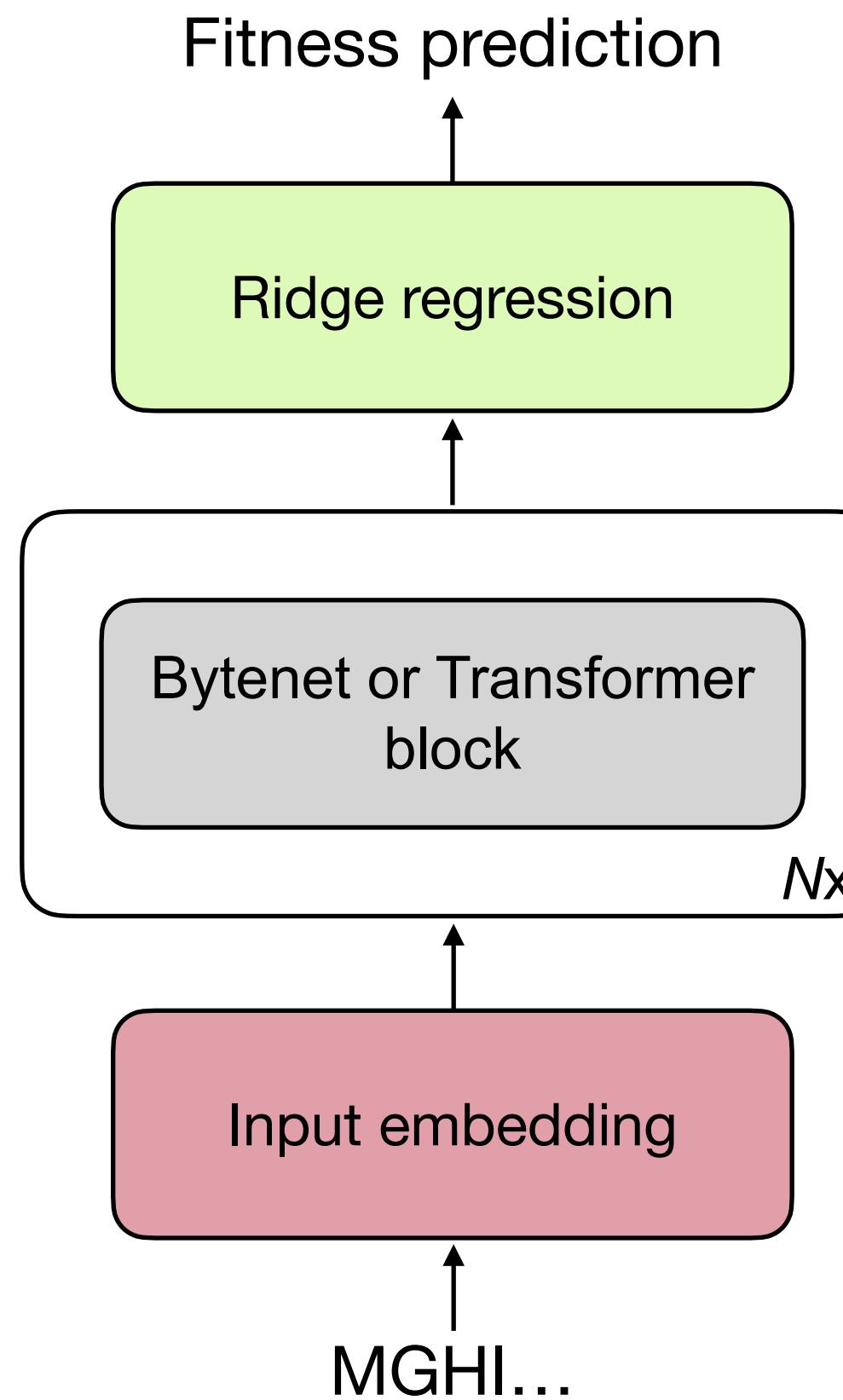
Not all layers are necessary for transfer



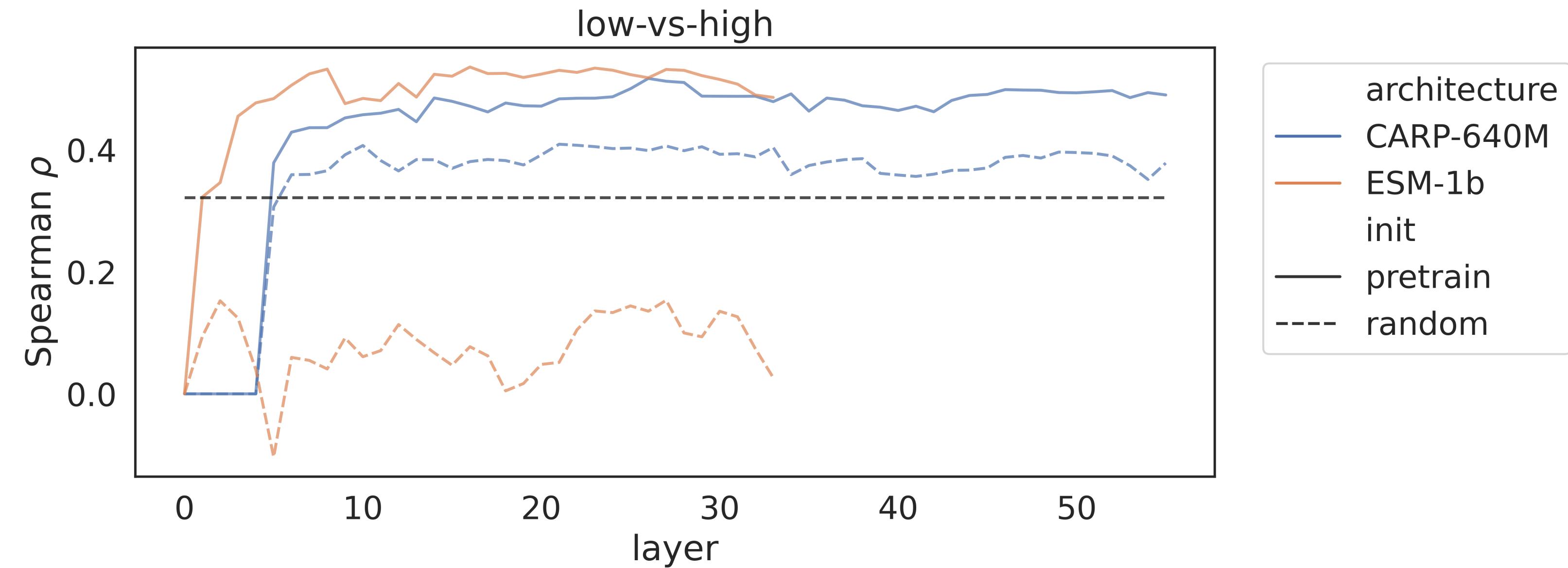
Not all layers are necessary for transfer



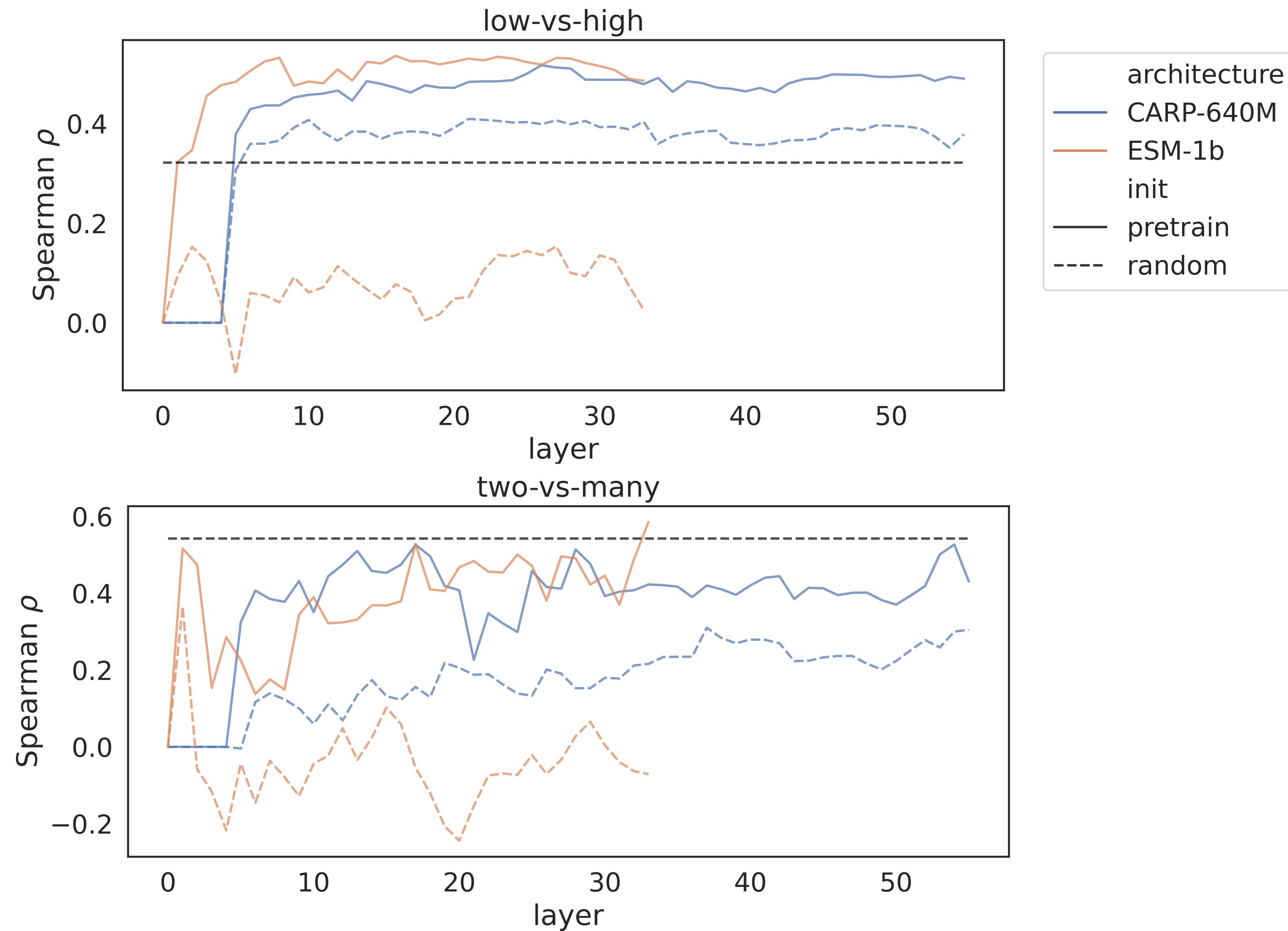
Not all layers are necessary for transfer



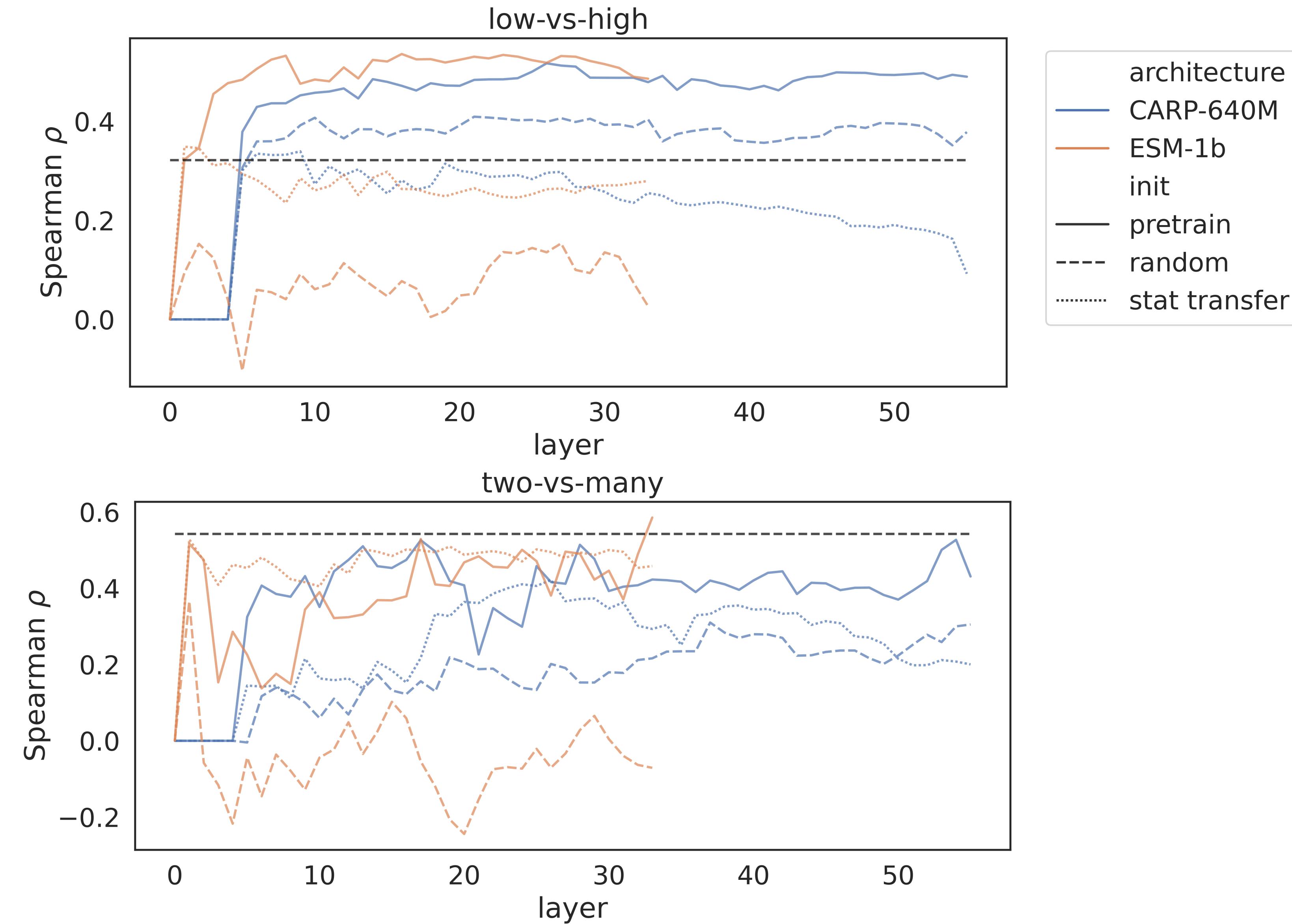
Pretraining helps beyond the architecture



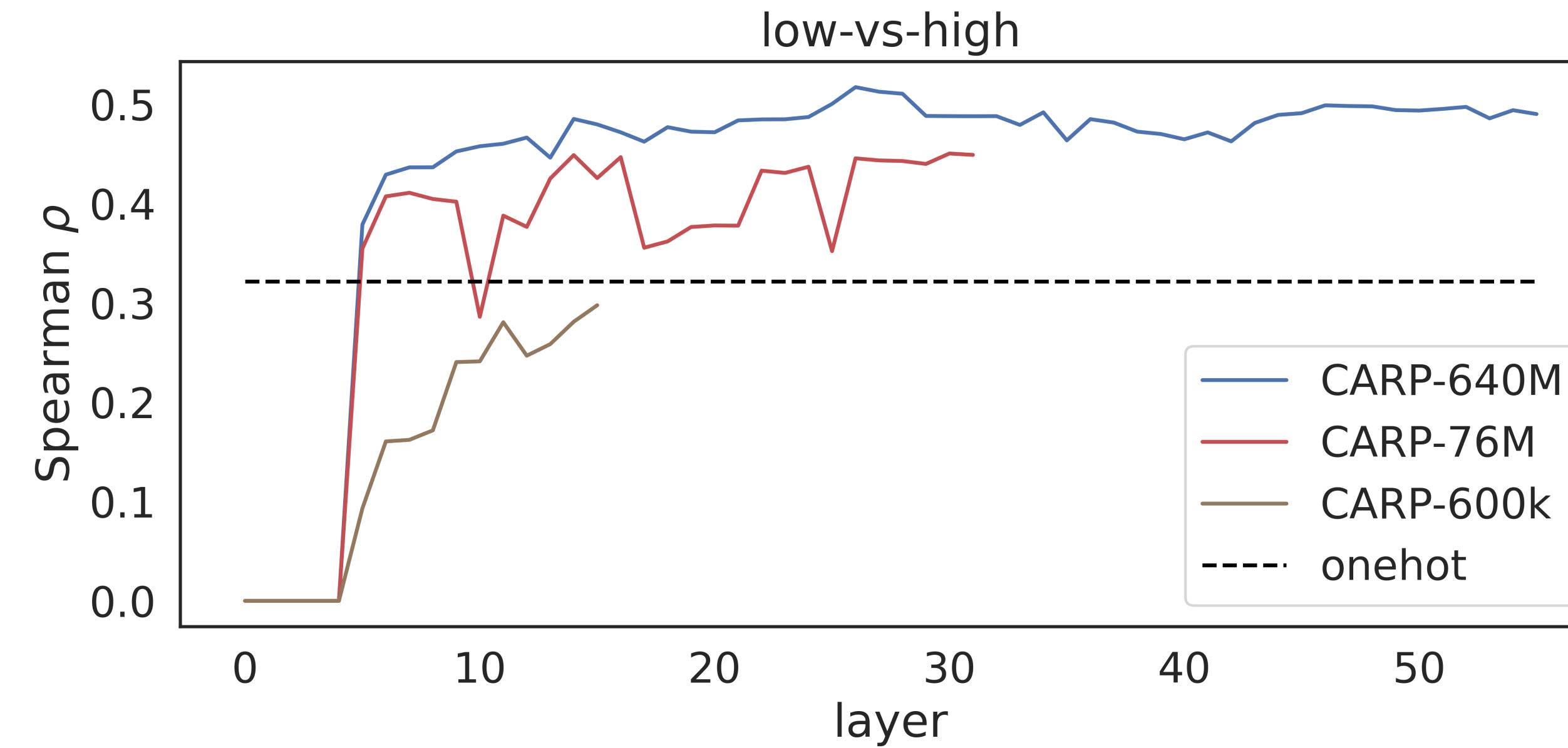
Pretraining helps beyond the architecture



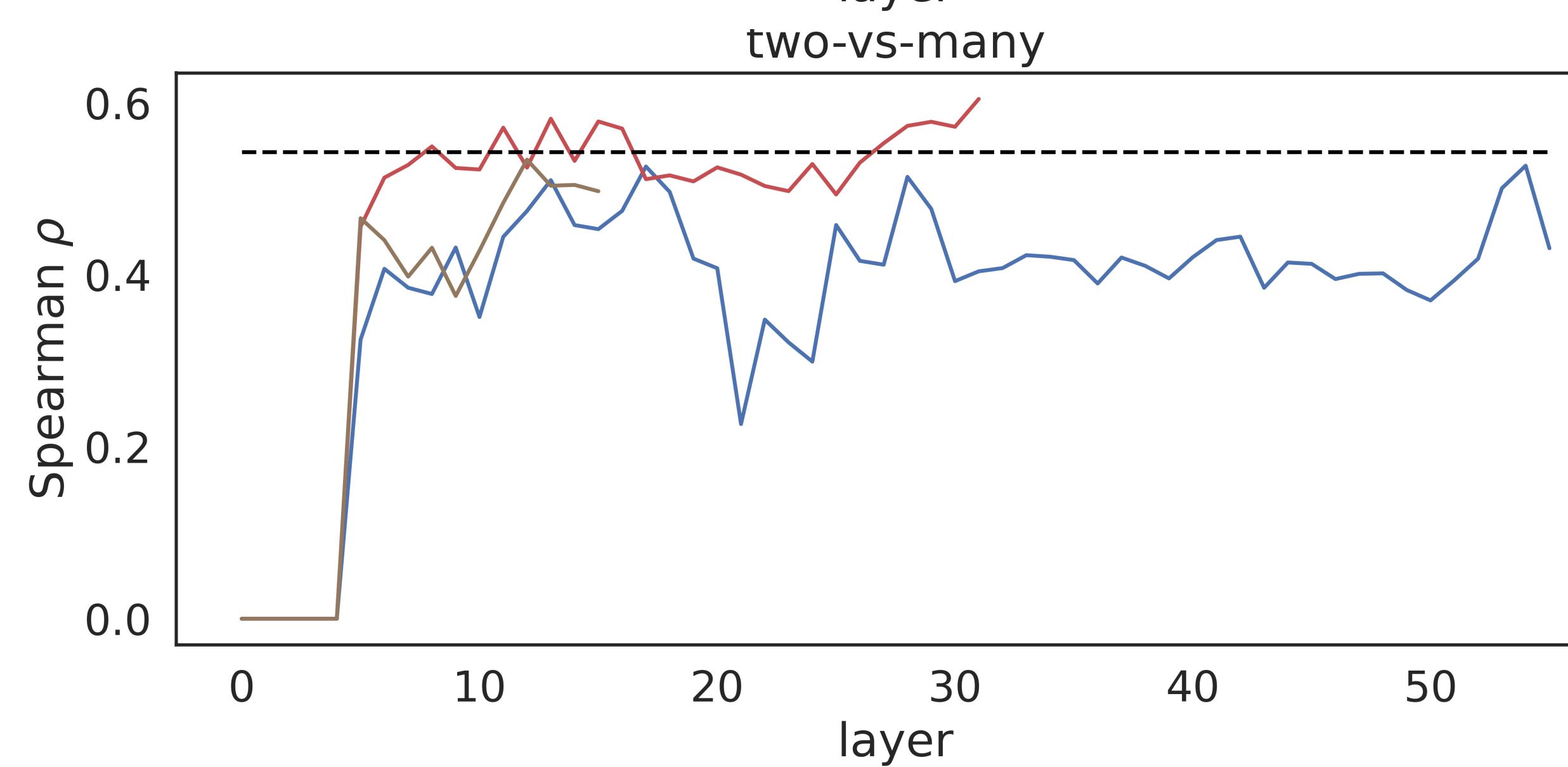
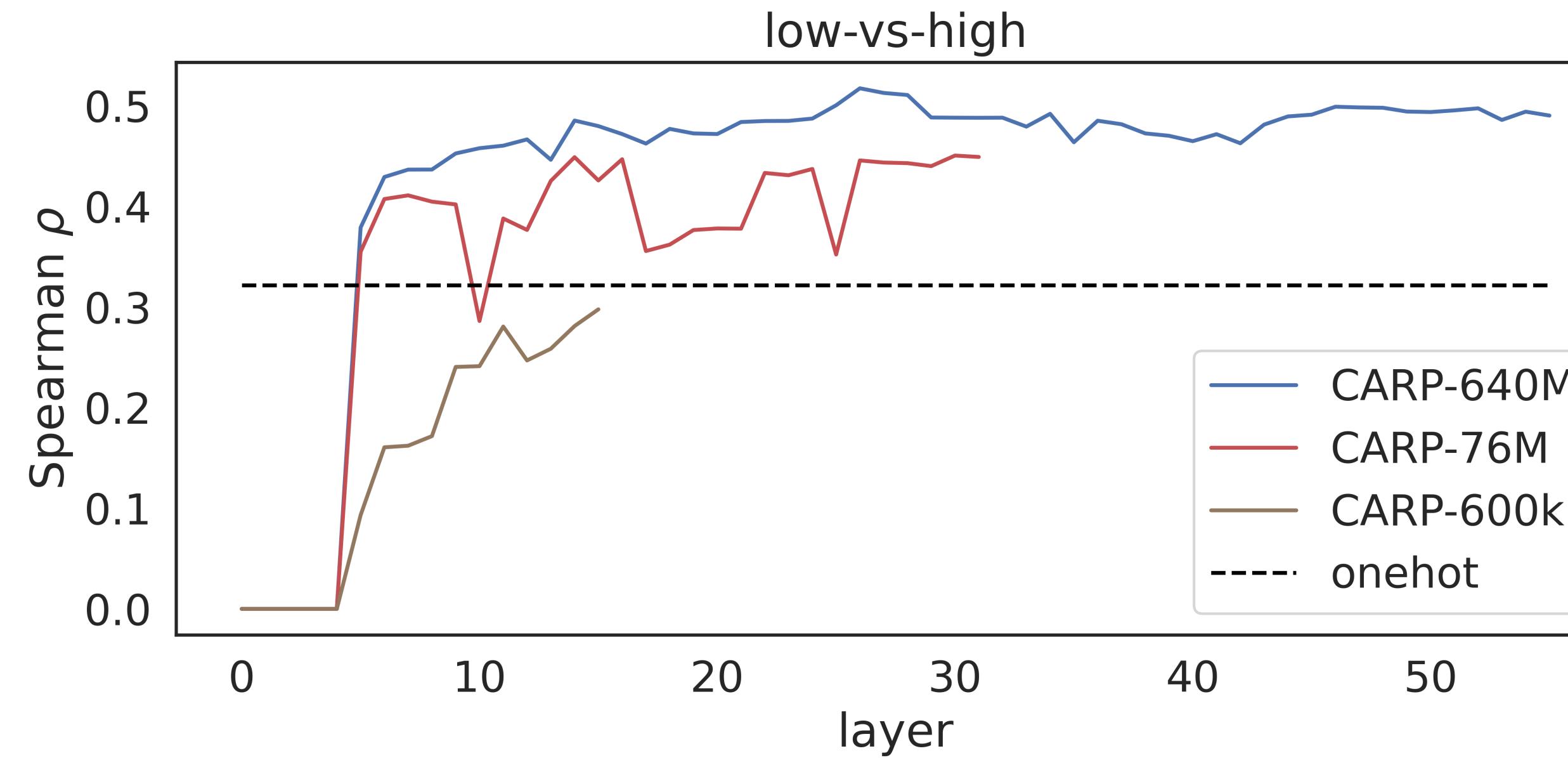
Sometimes, shuffled weights work too



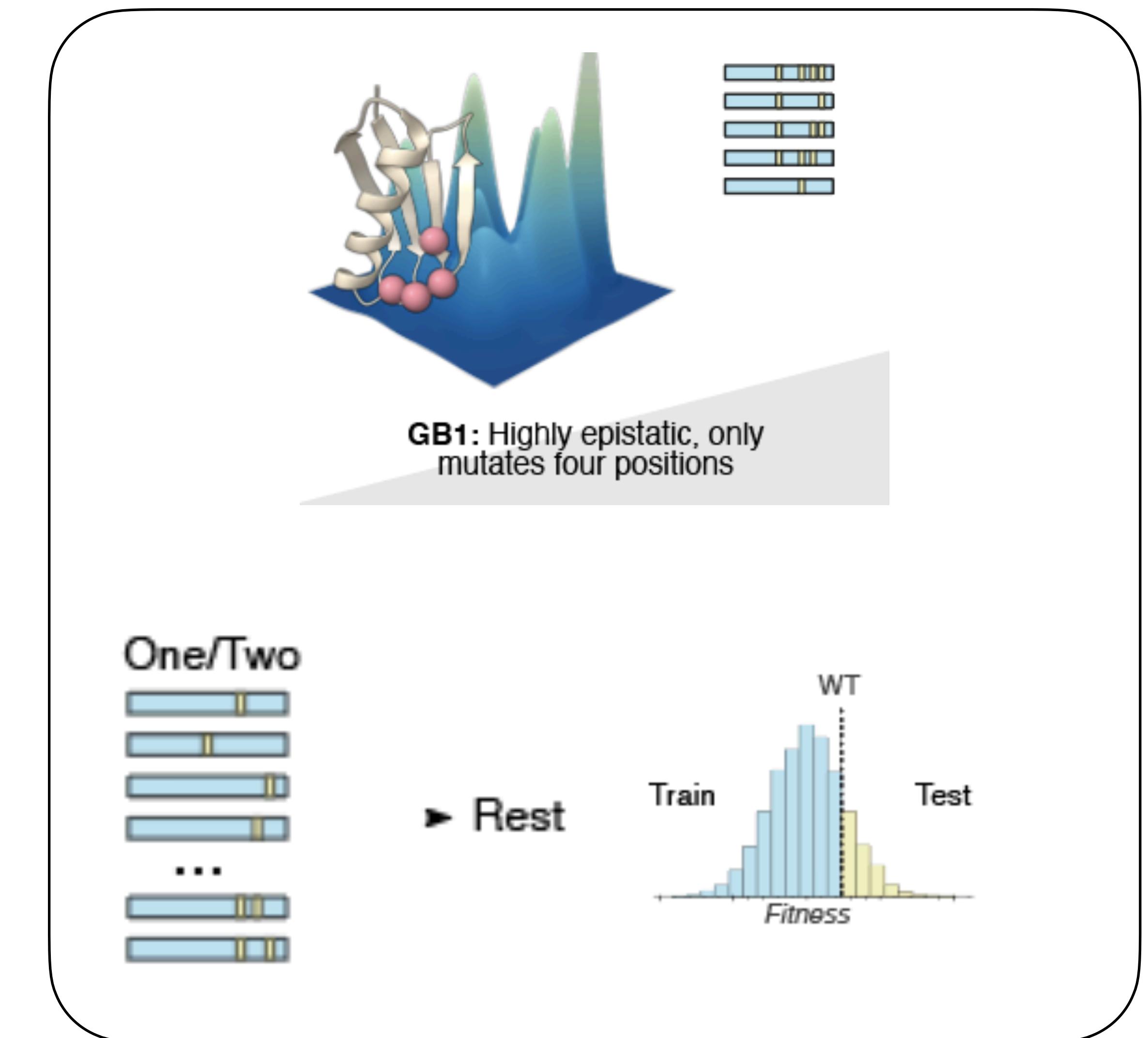
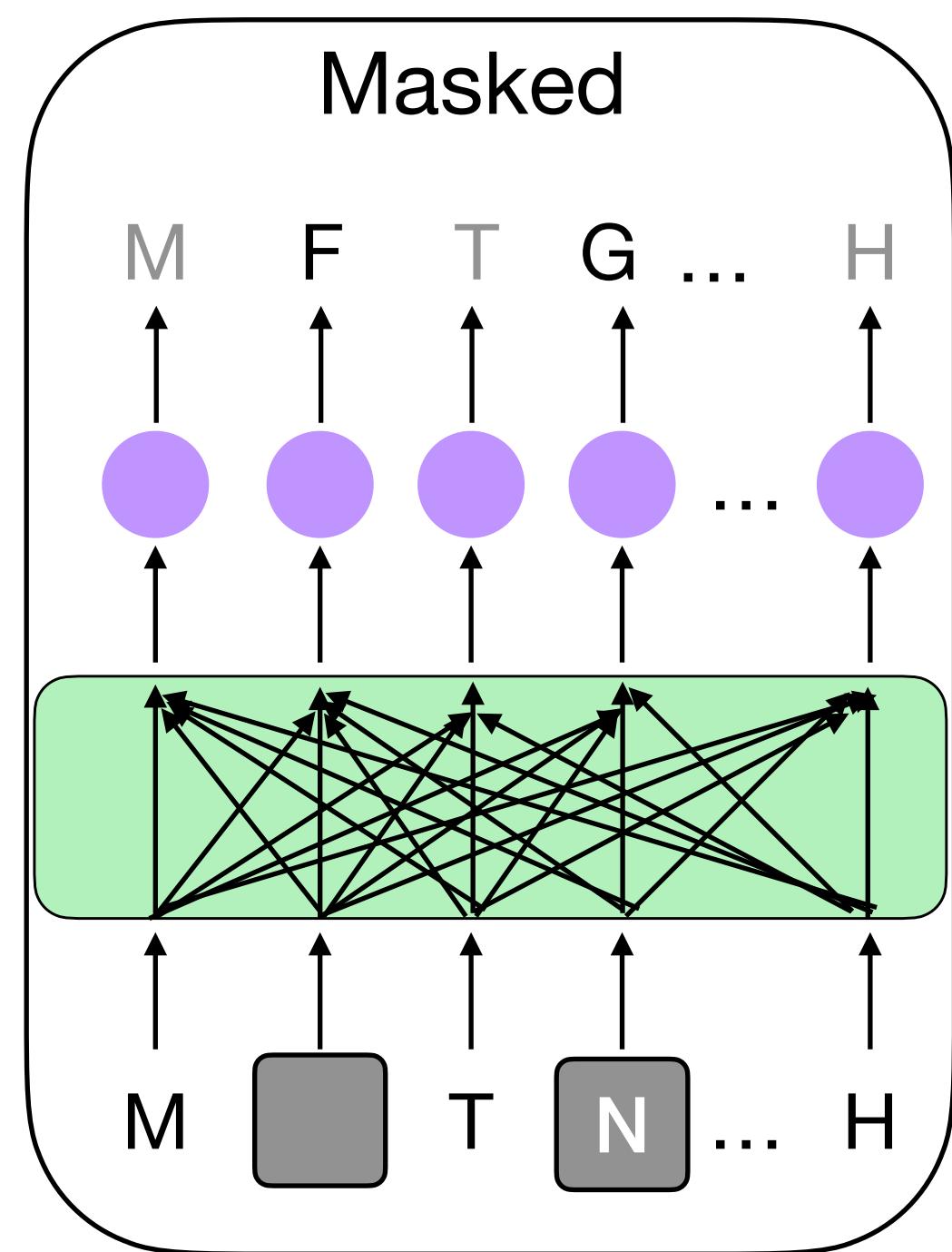
Bigger models do not always help



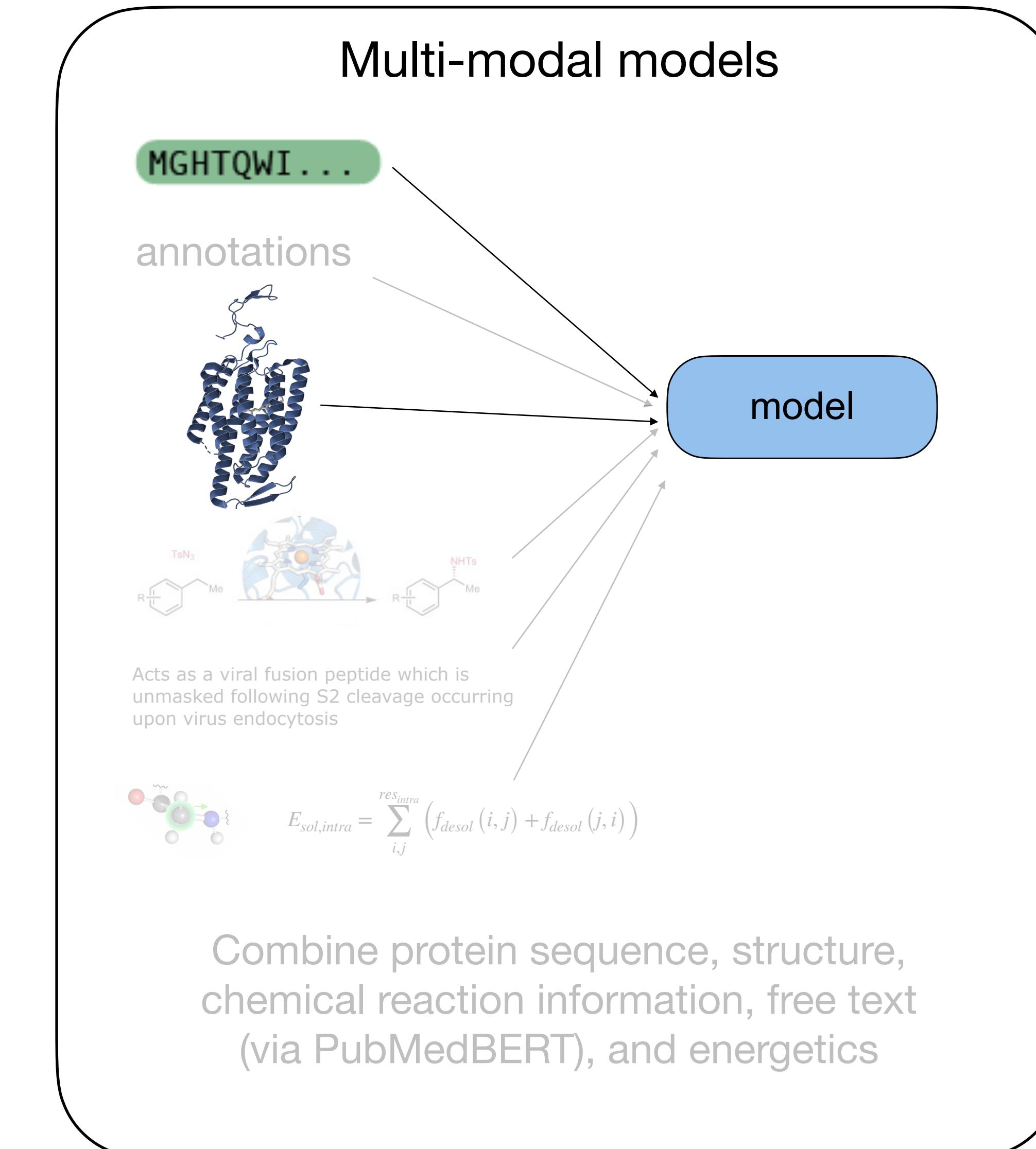
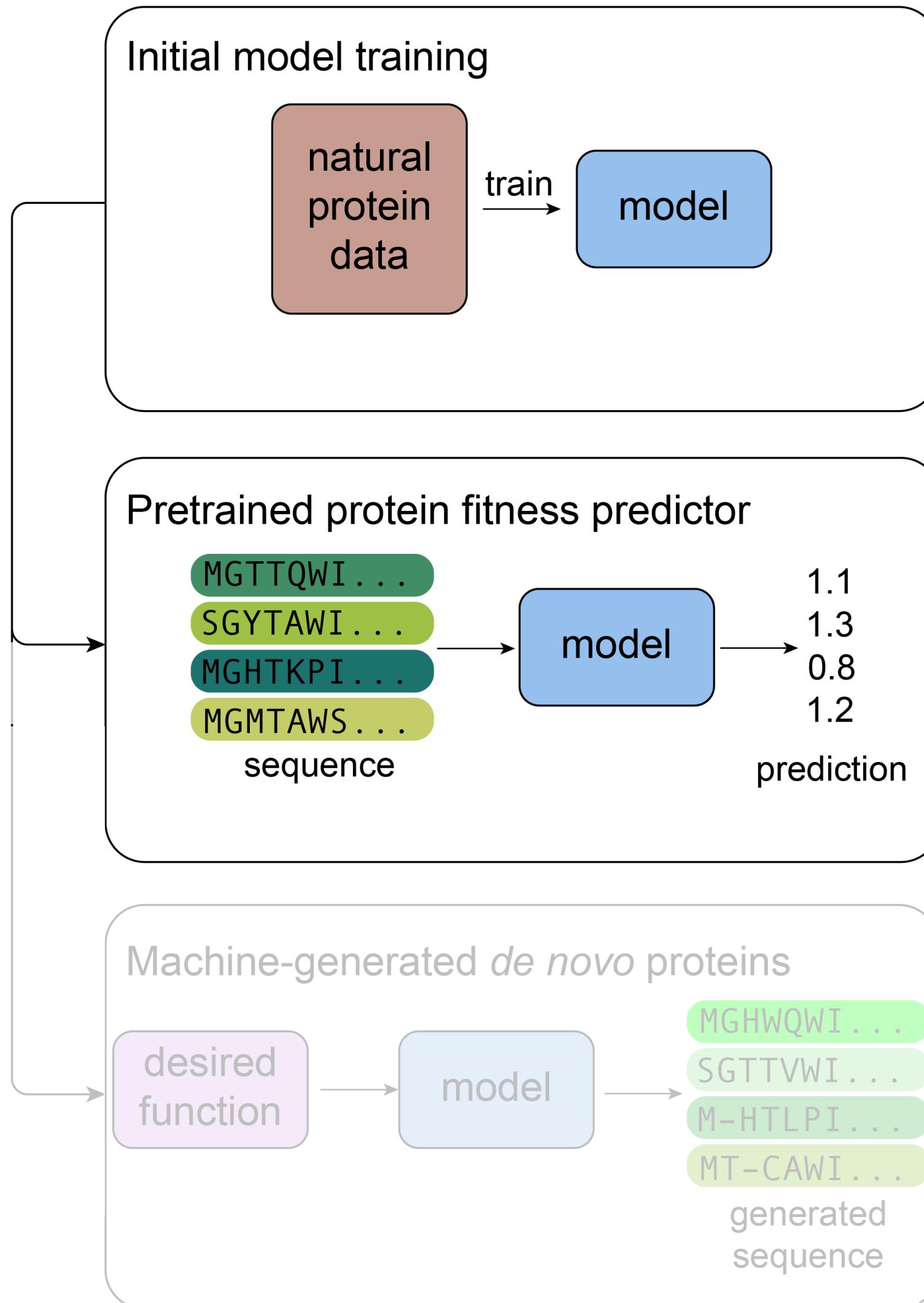
Bigger models do not always help



Pretrain and downstream tasks are mismatched

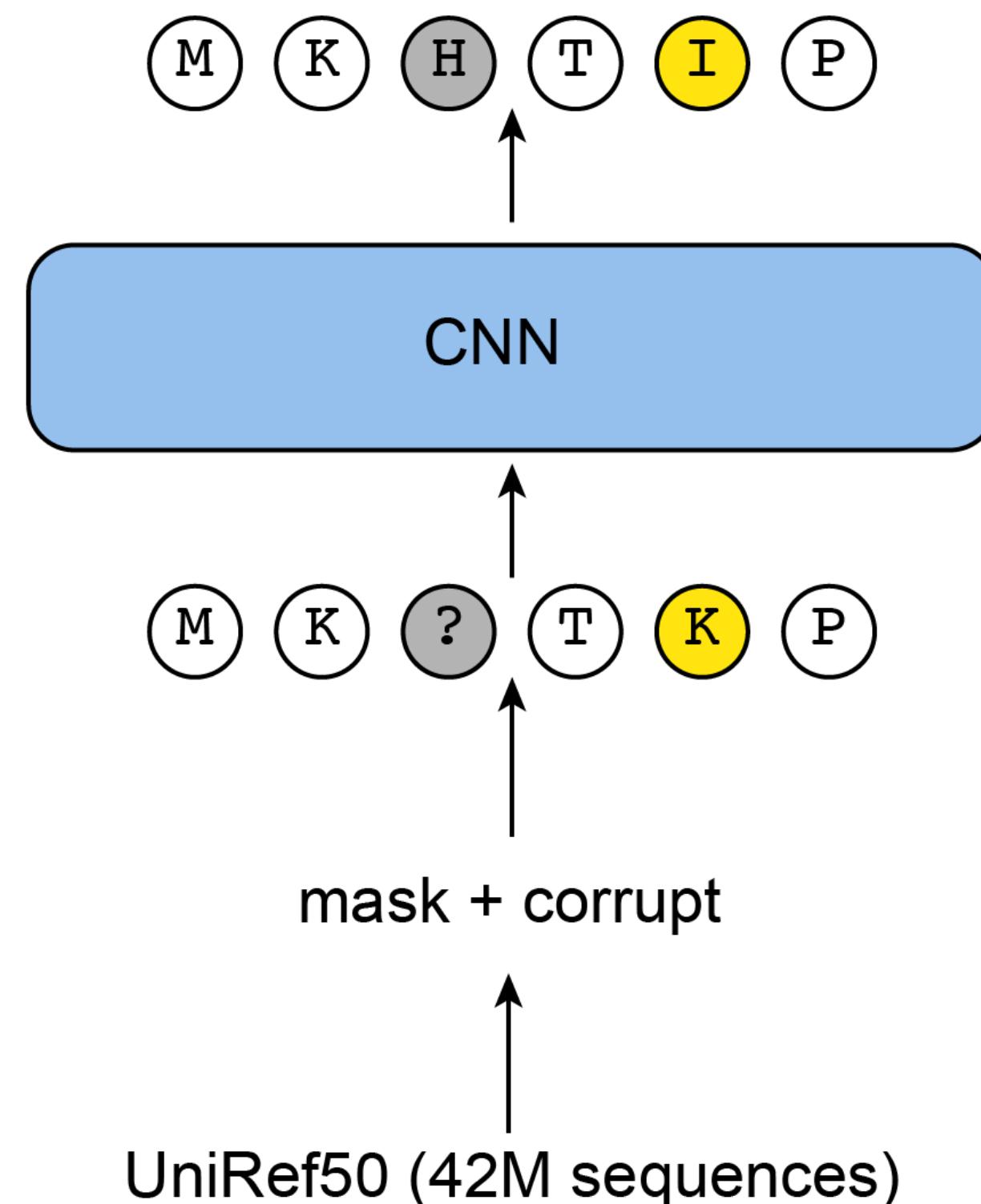


Pretrain with structure and sequence



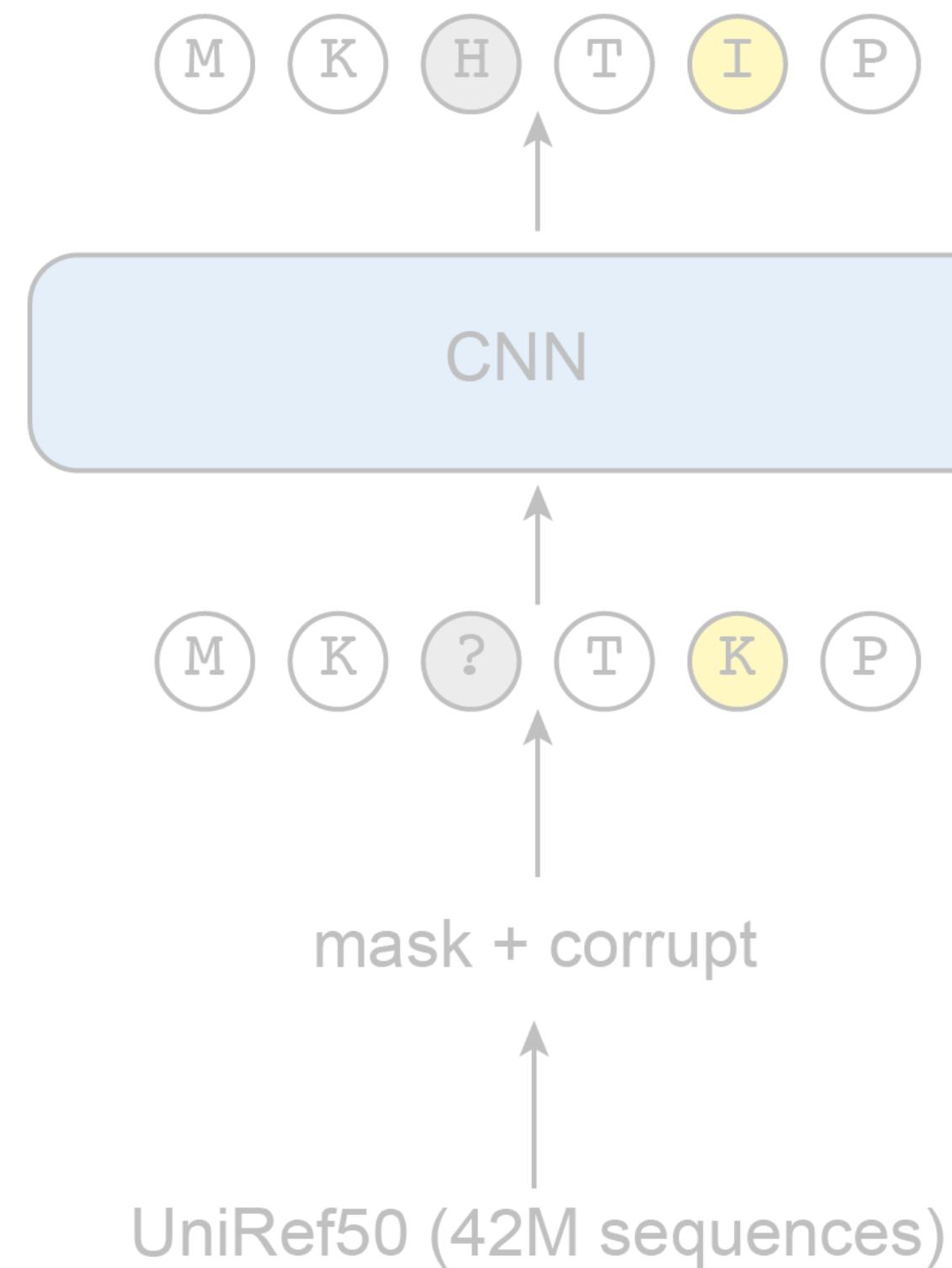
Structural information improves pretraining

Convolutional autoencoding
representations of proteins

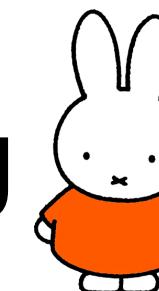


Structural information improves pretraining

Convolutional autoencoding
representations of proteins

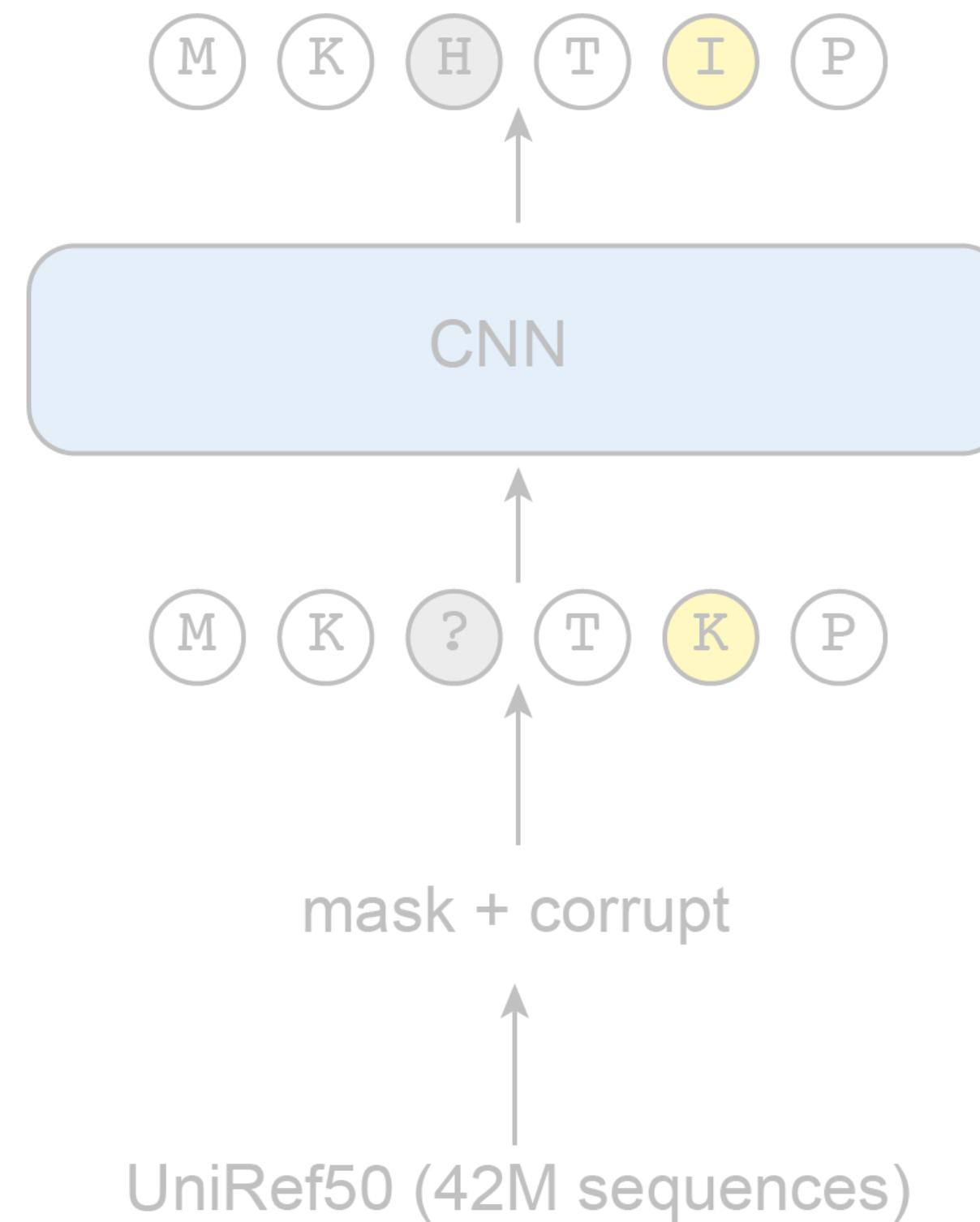


Masked inverse folding

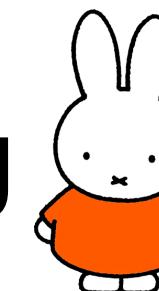


Structural information improves pretraining

Convolutional autoencoding
representations of proteins

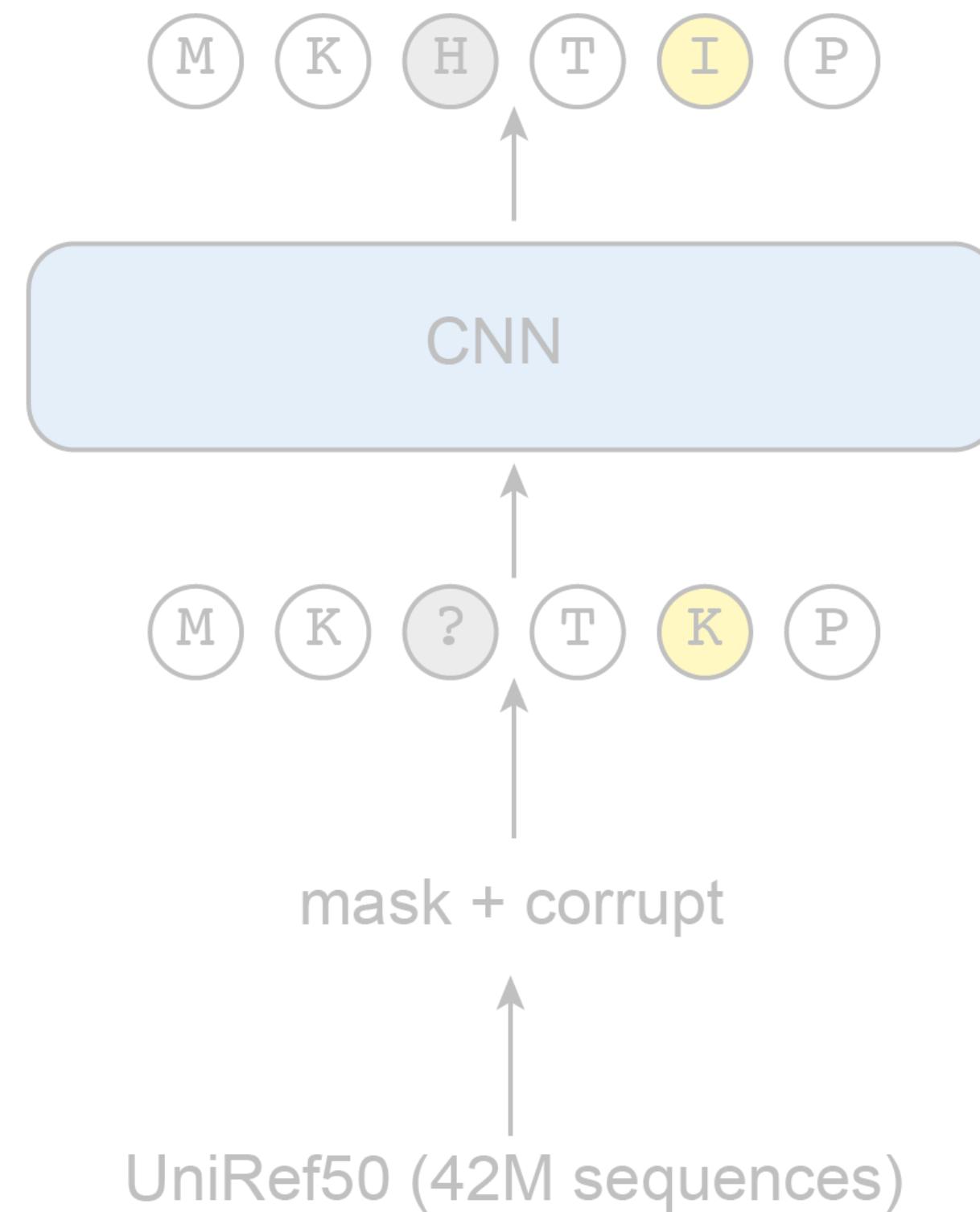


Masked inverse folding

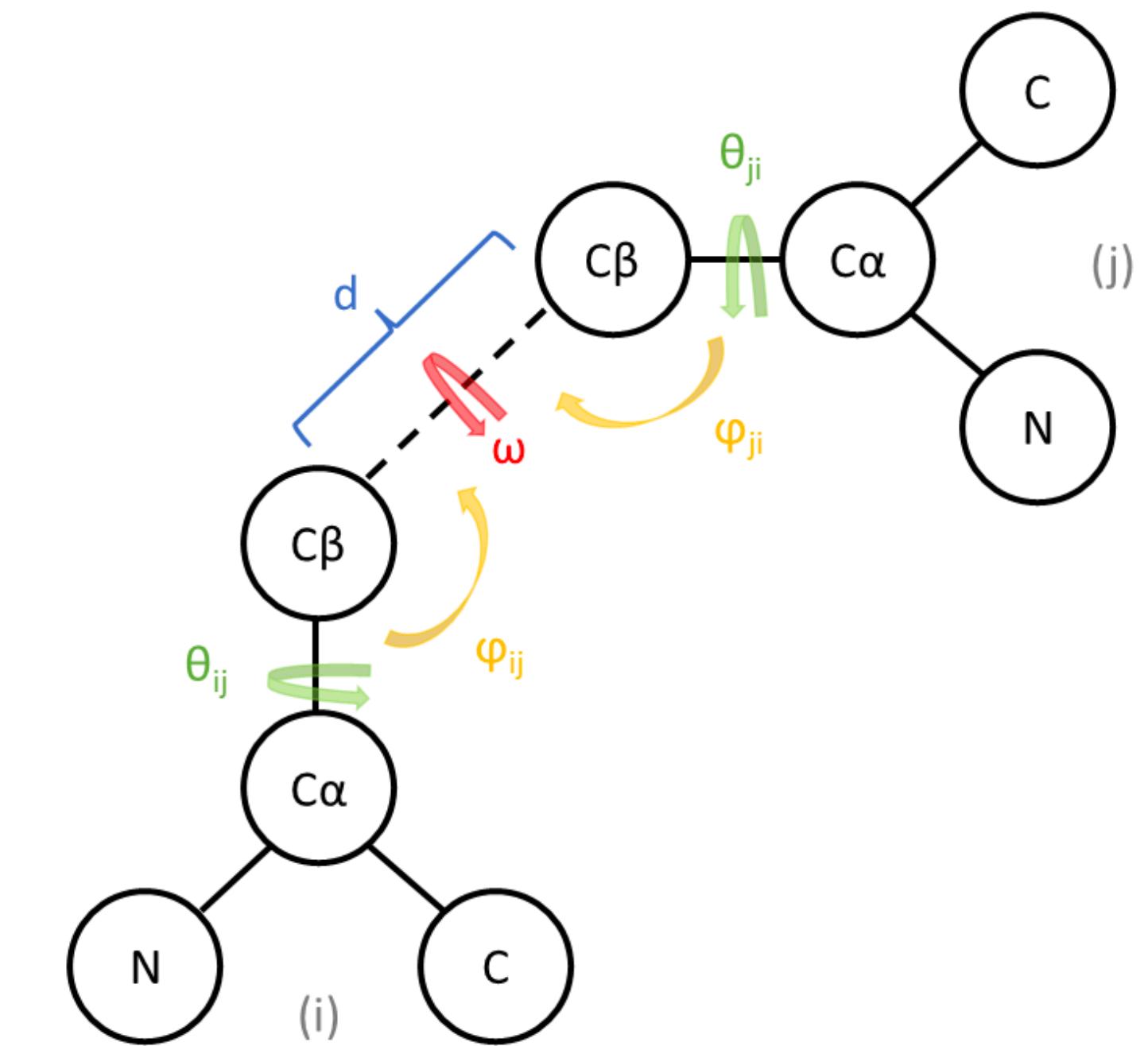
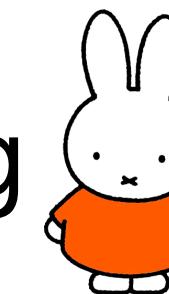


Structural information improves pretraining

Convolutional autoencoding
representations of proteins

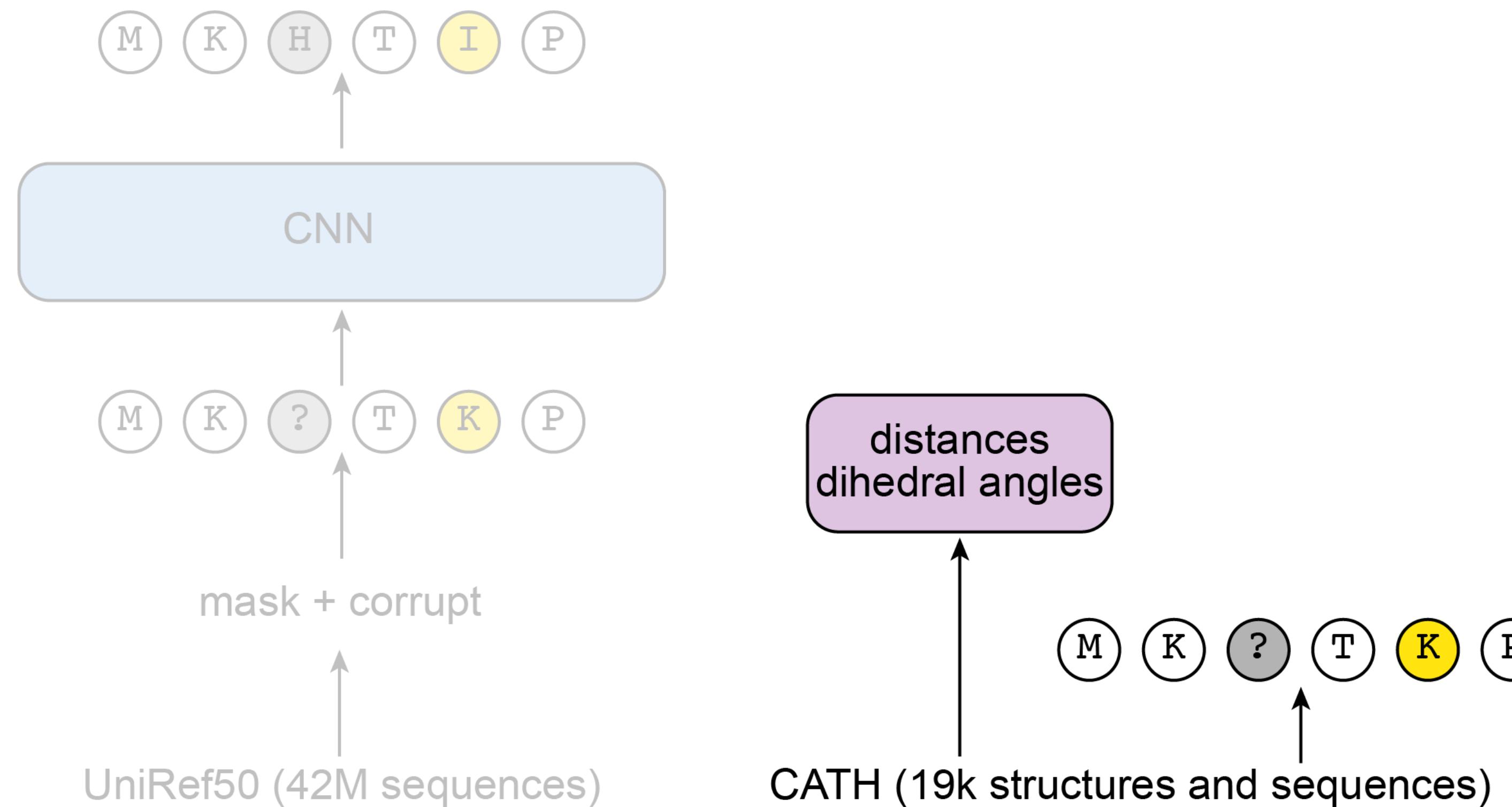


Masked inverse folding

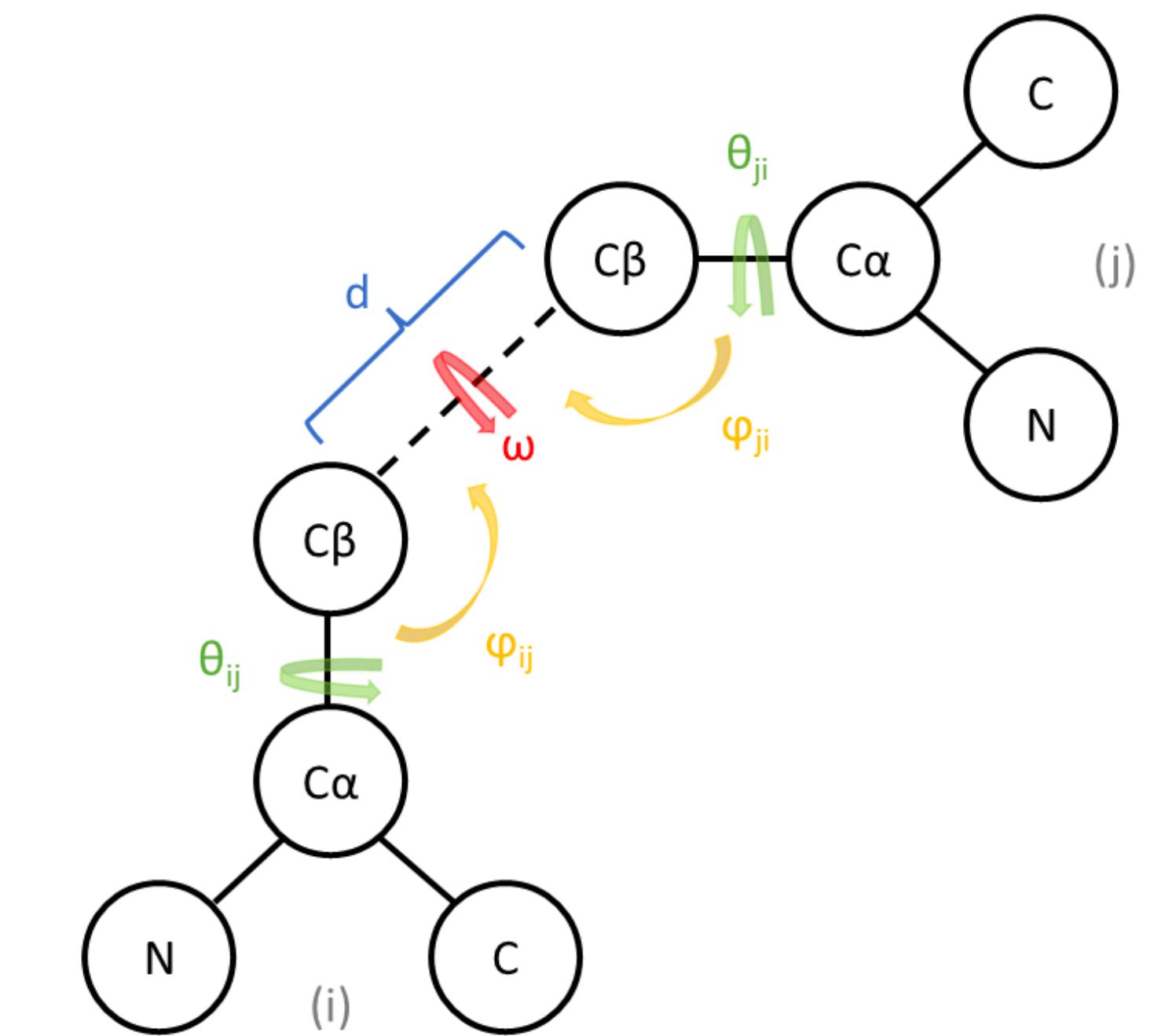
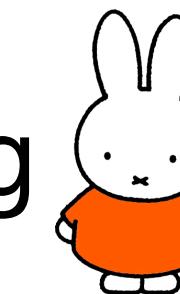


Structural information improves pretraining

Convolutional autoencoding
representations of proteins

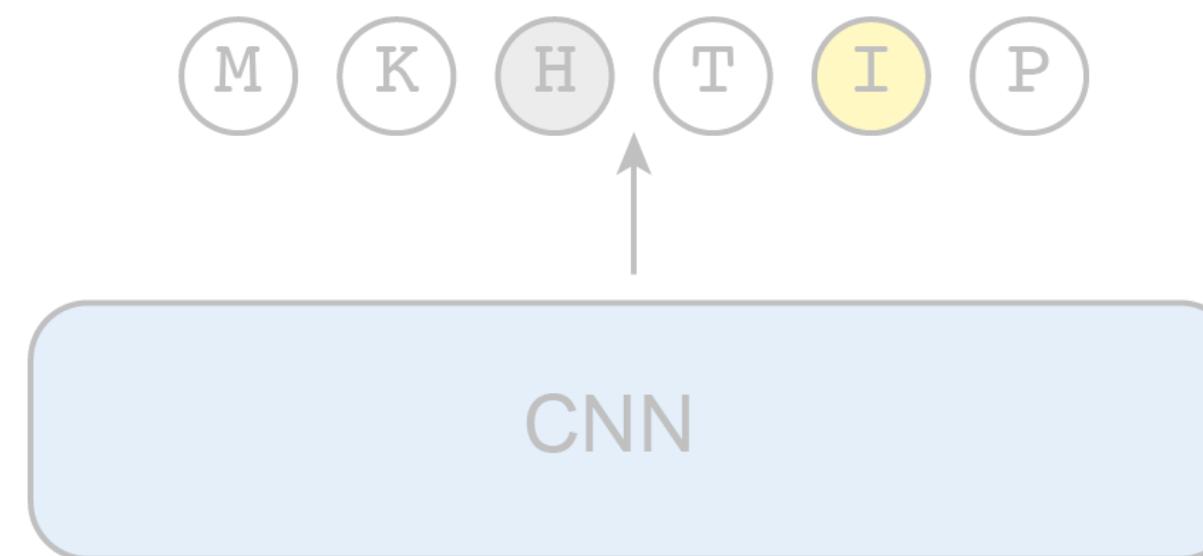


Masked inverse folding

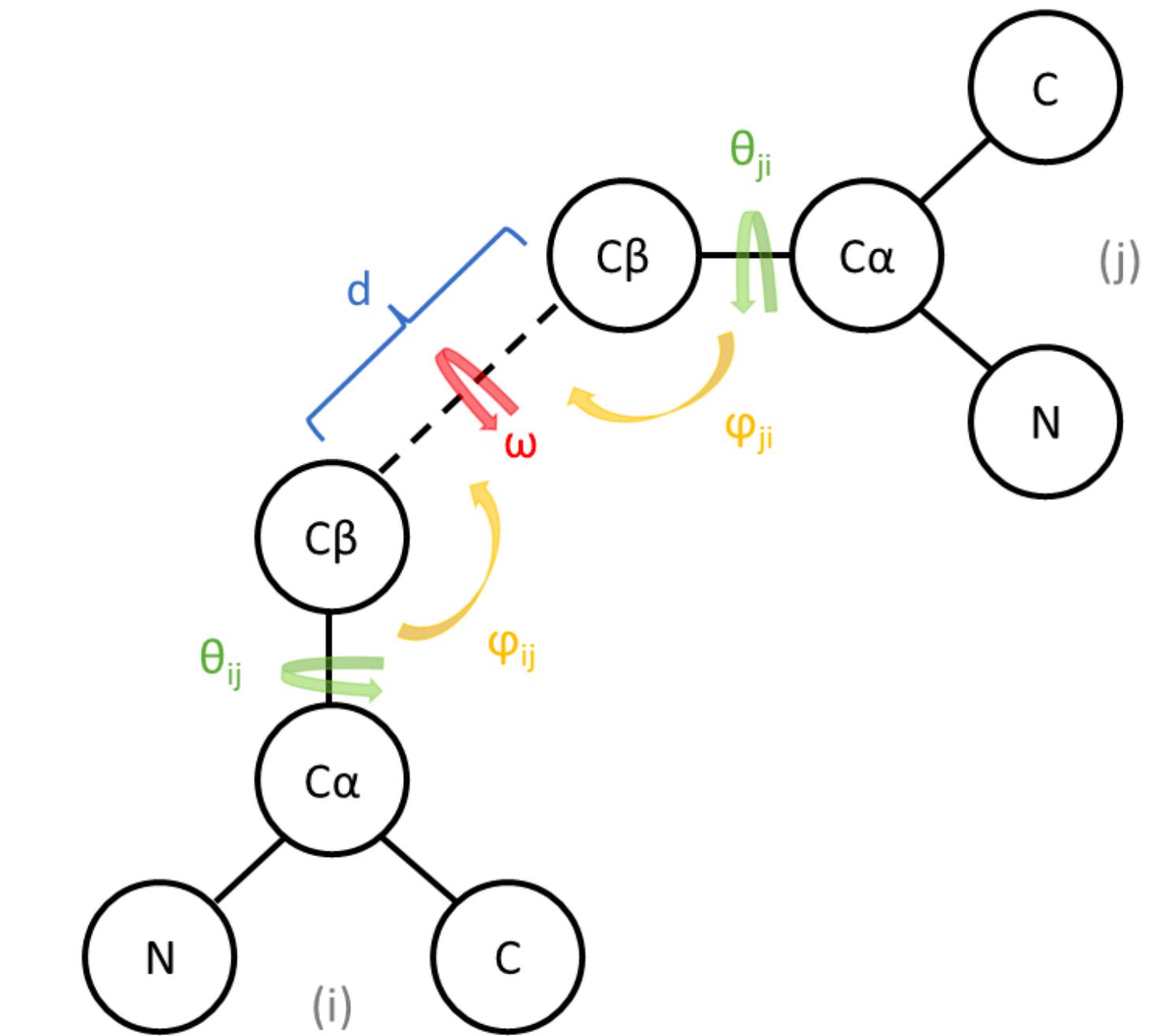
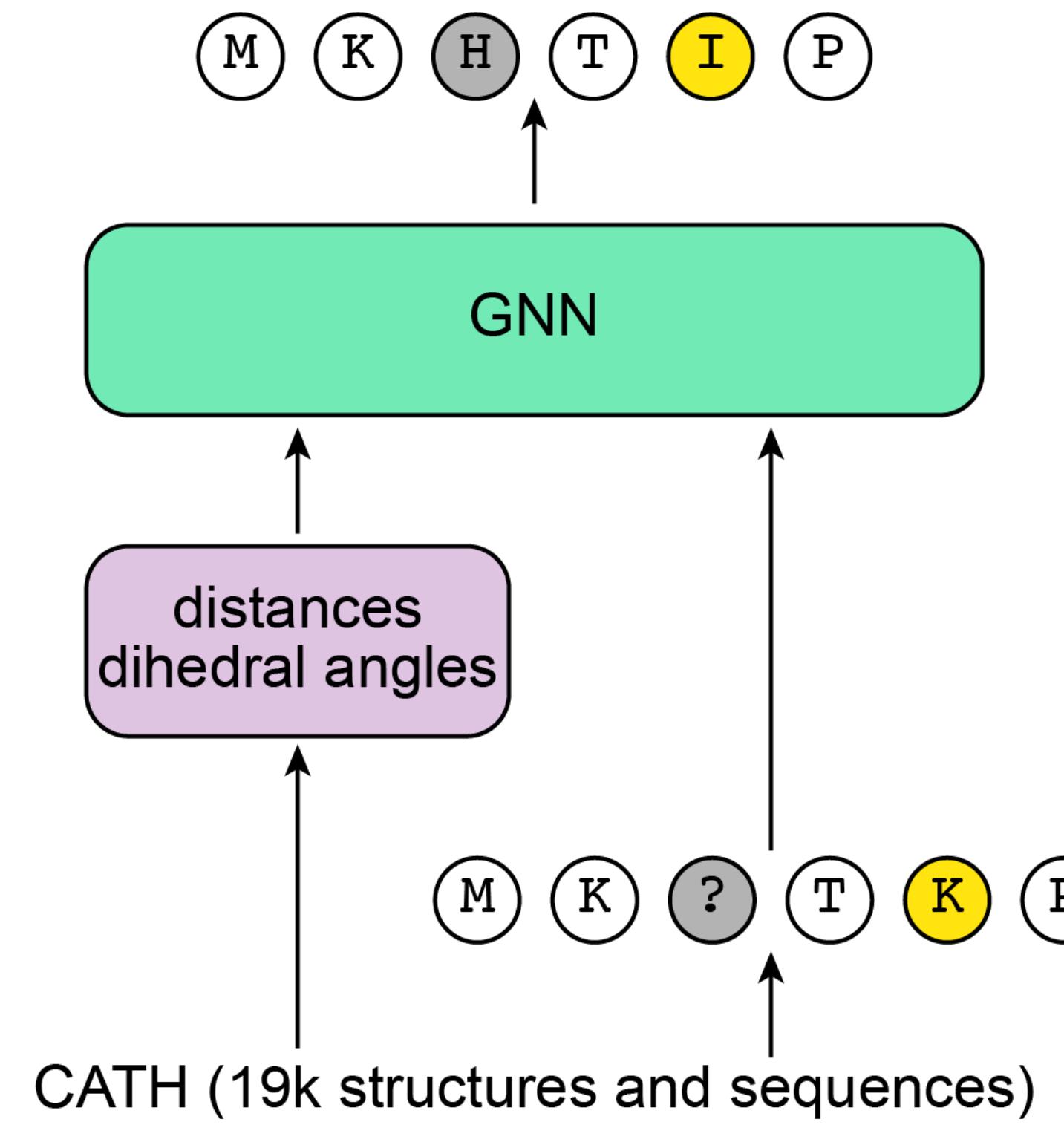


Structural information improves pretraining

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

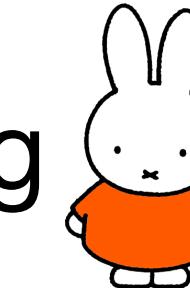


mask + corrupt

UniRef50 (42M sequences)



Masked inverse folding

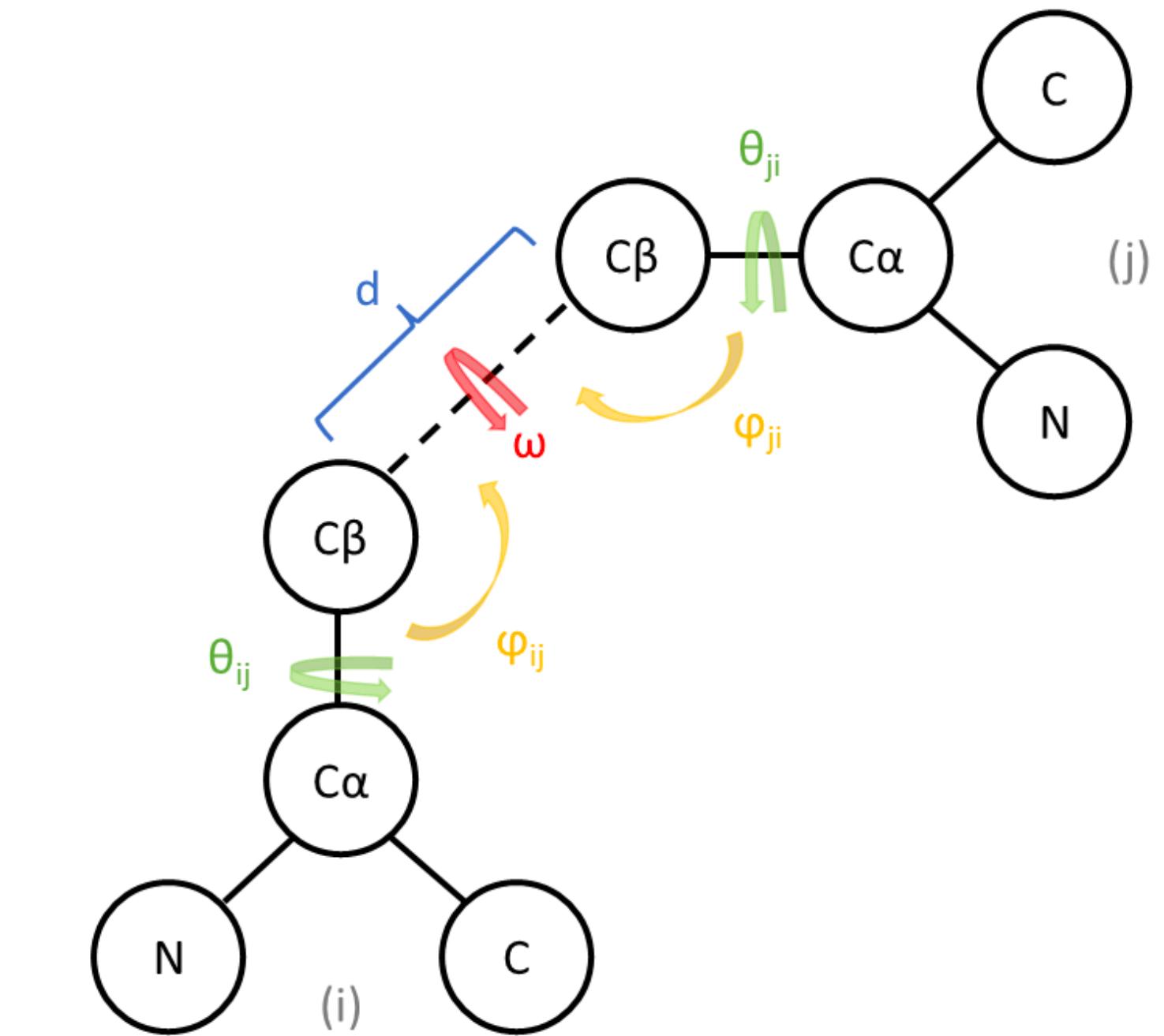


distances
dihedral angles

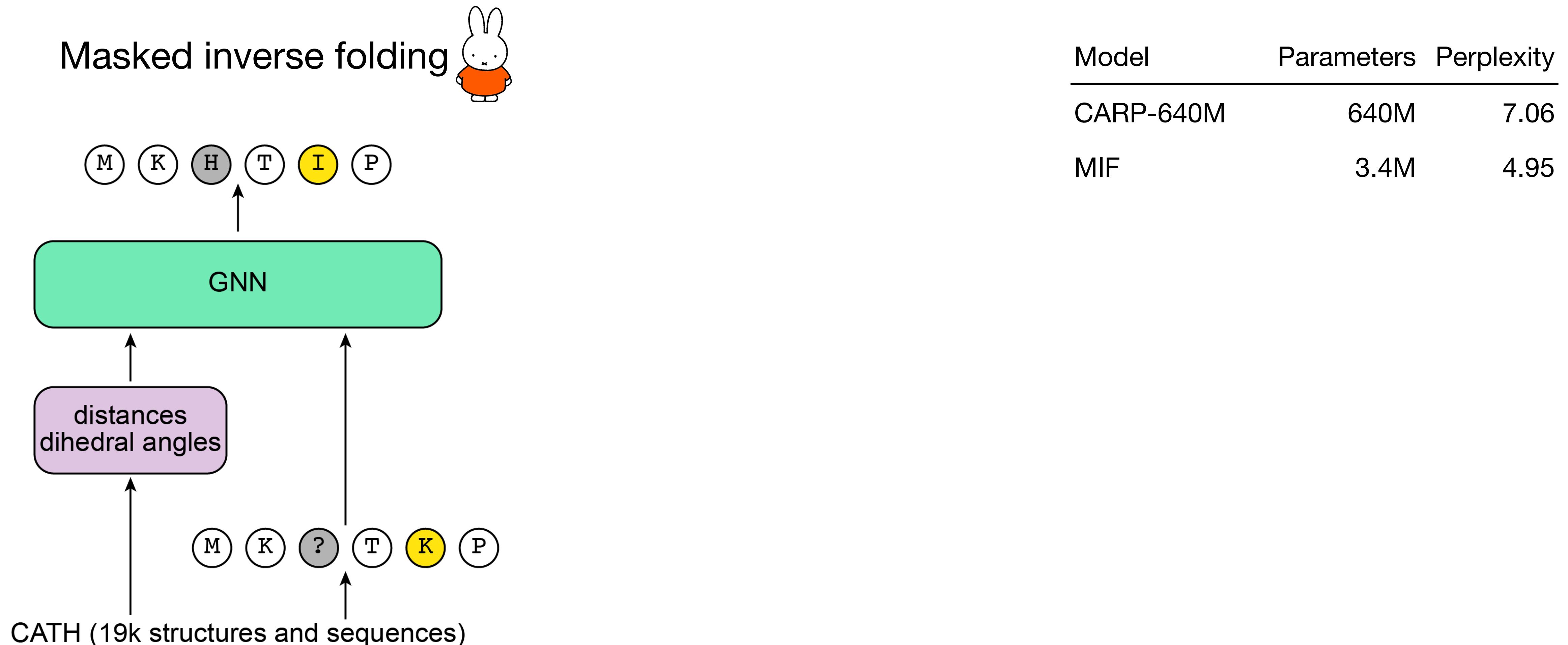
M K ? T K P

CATH (19k structures and sequences)

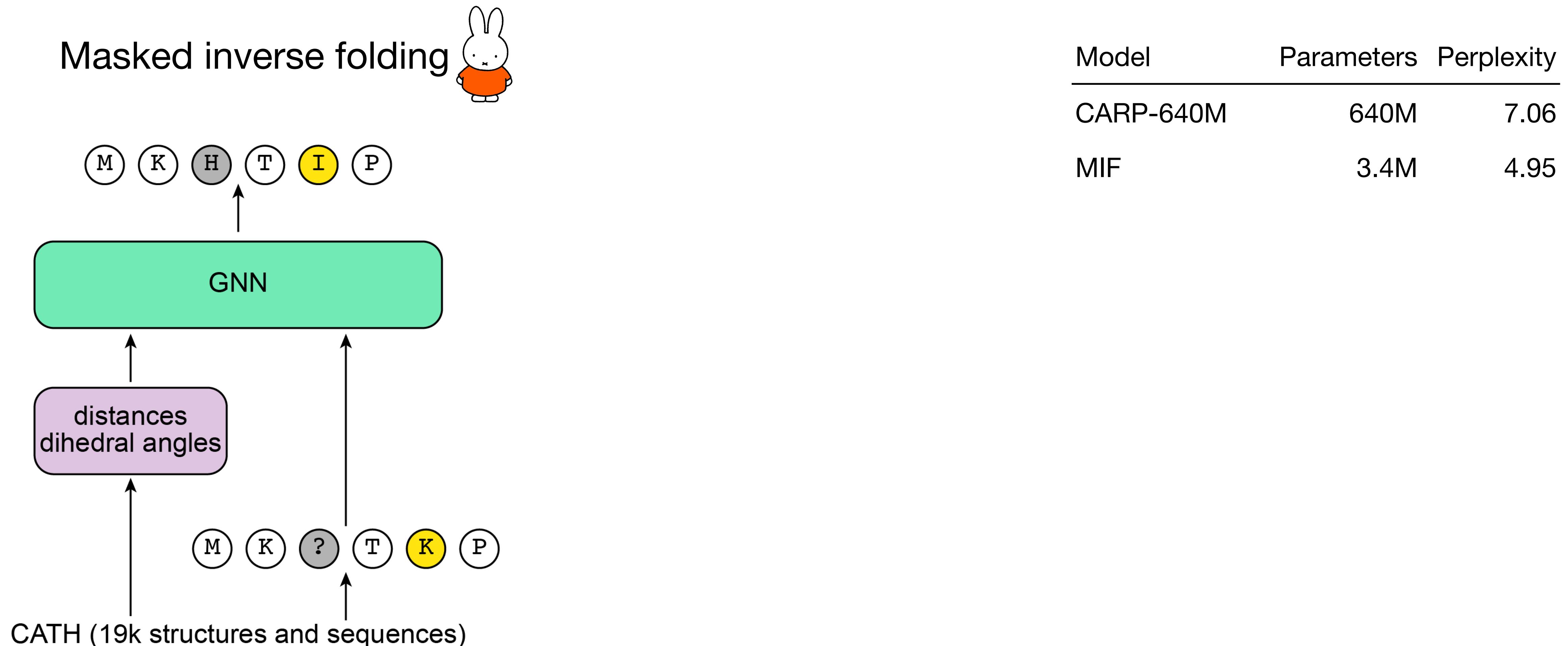
Model	Parameters	Perplexity
CARP-640M	640M	7.06
MIF	3.4M	4.95



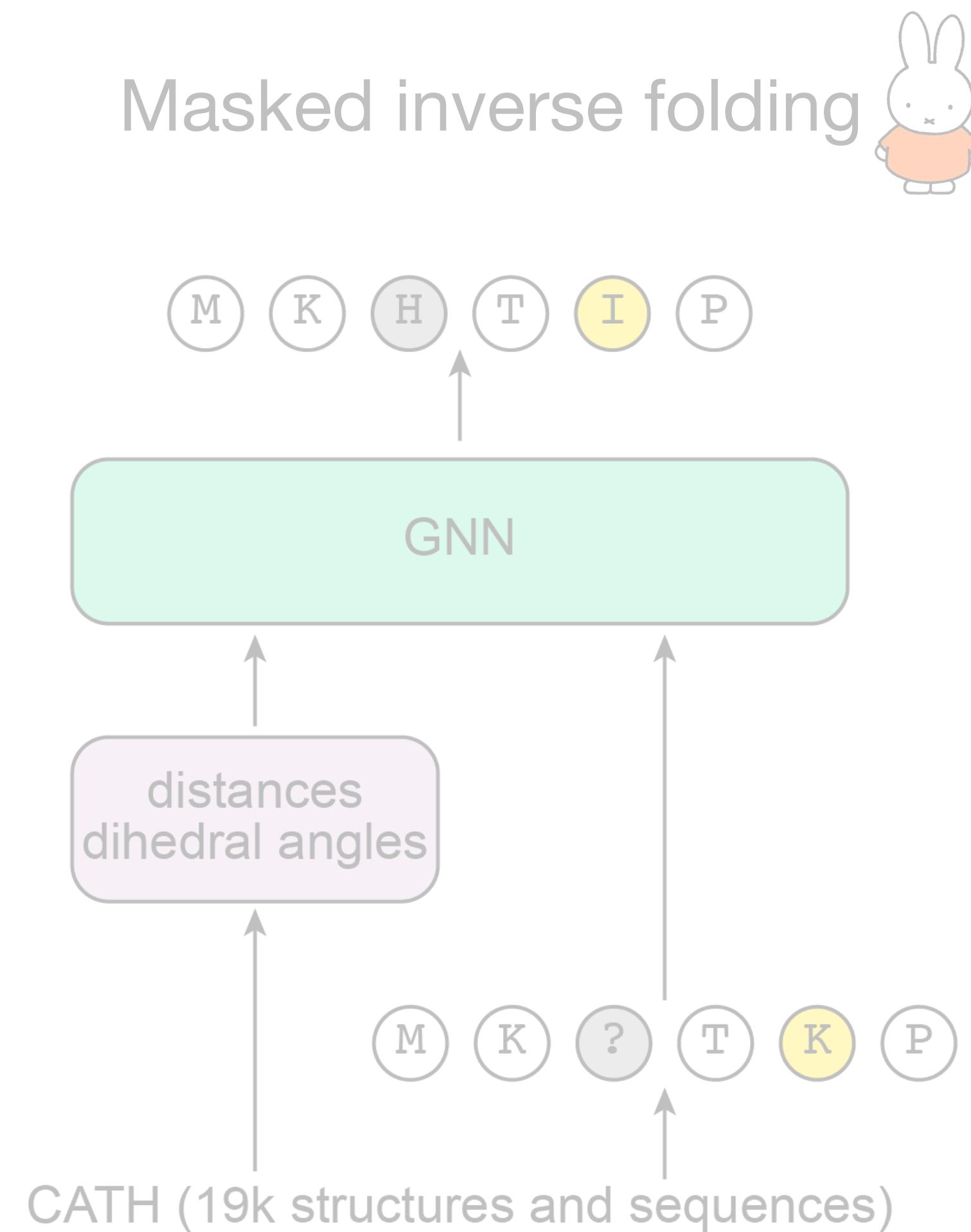
Structural information improves pretraining



Sequence transfer improves pretraining more



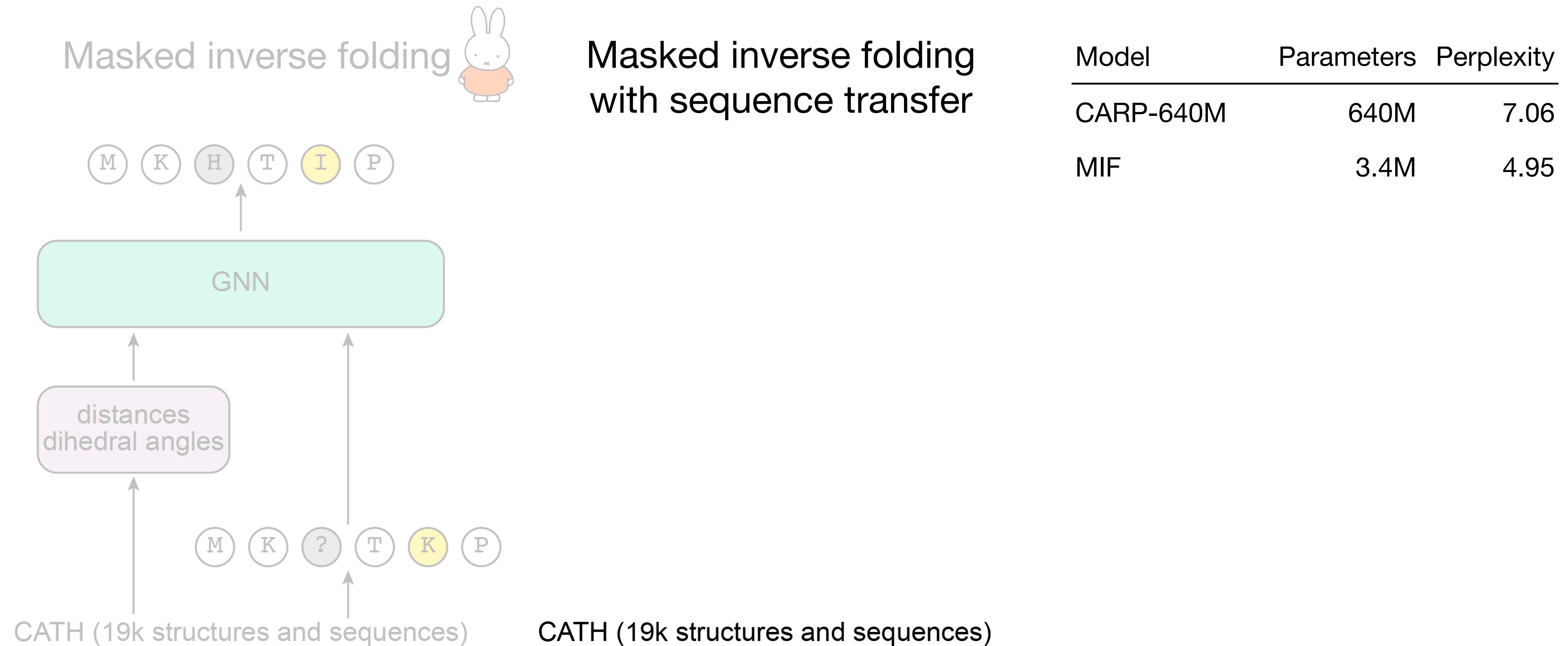
Sequence transfer improves pretraining more



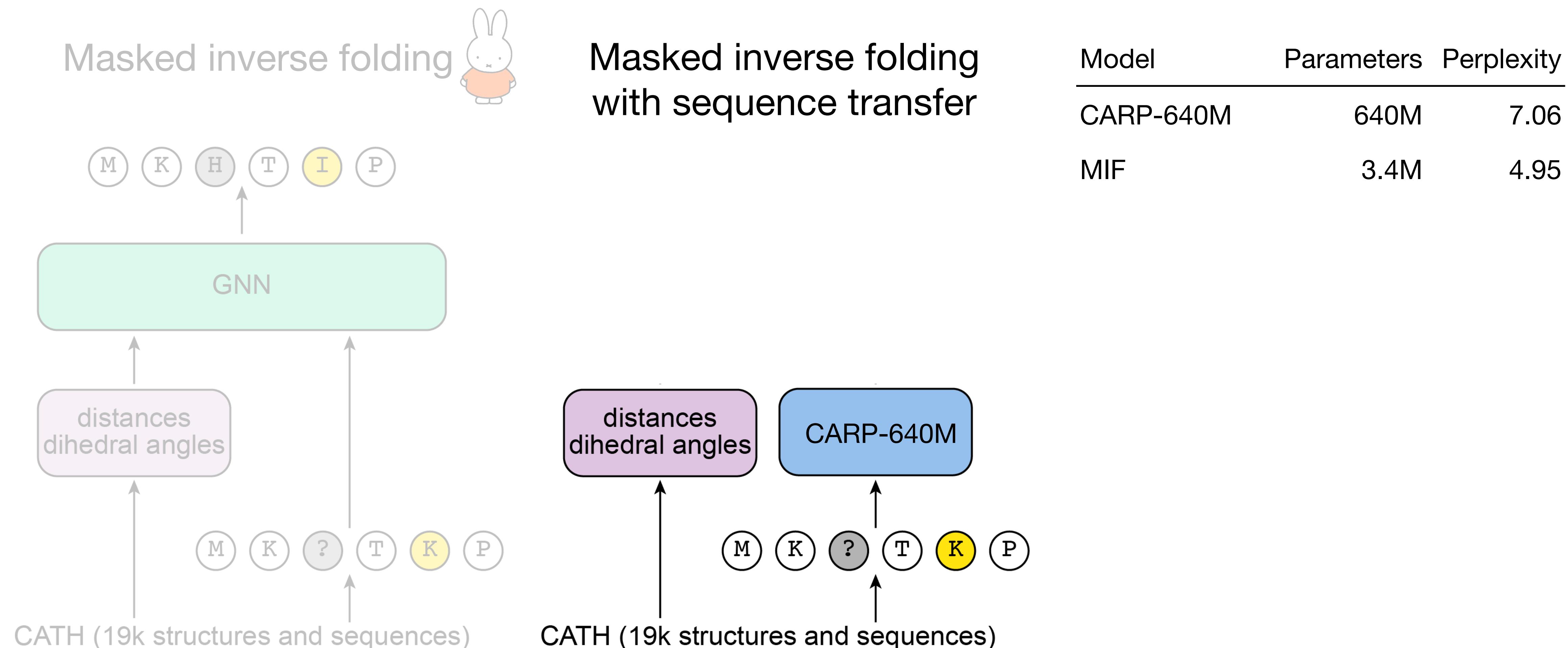
Masked inverse folding
with sequence transfer

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

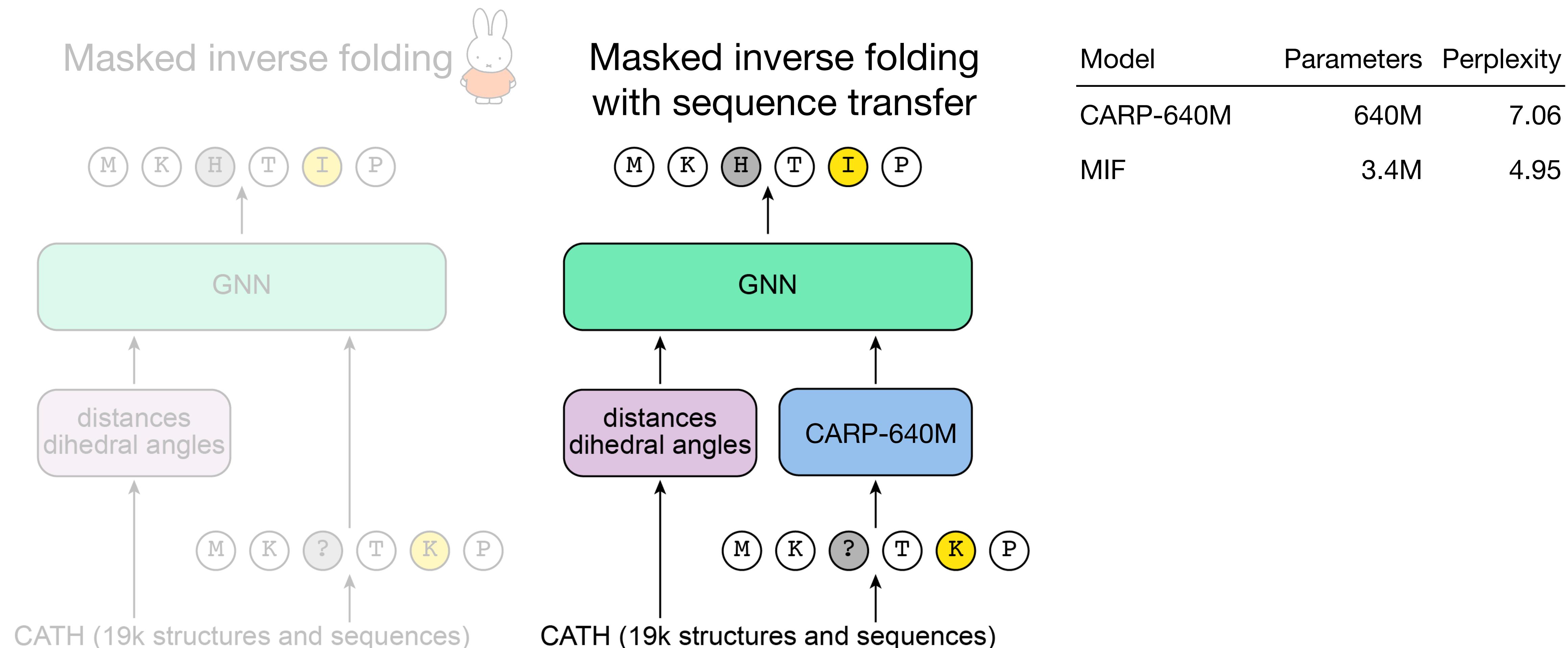
Sequence transfer improves pretraining more



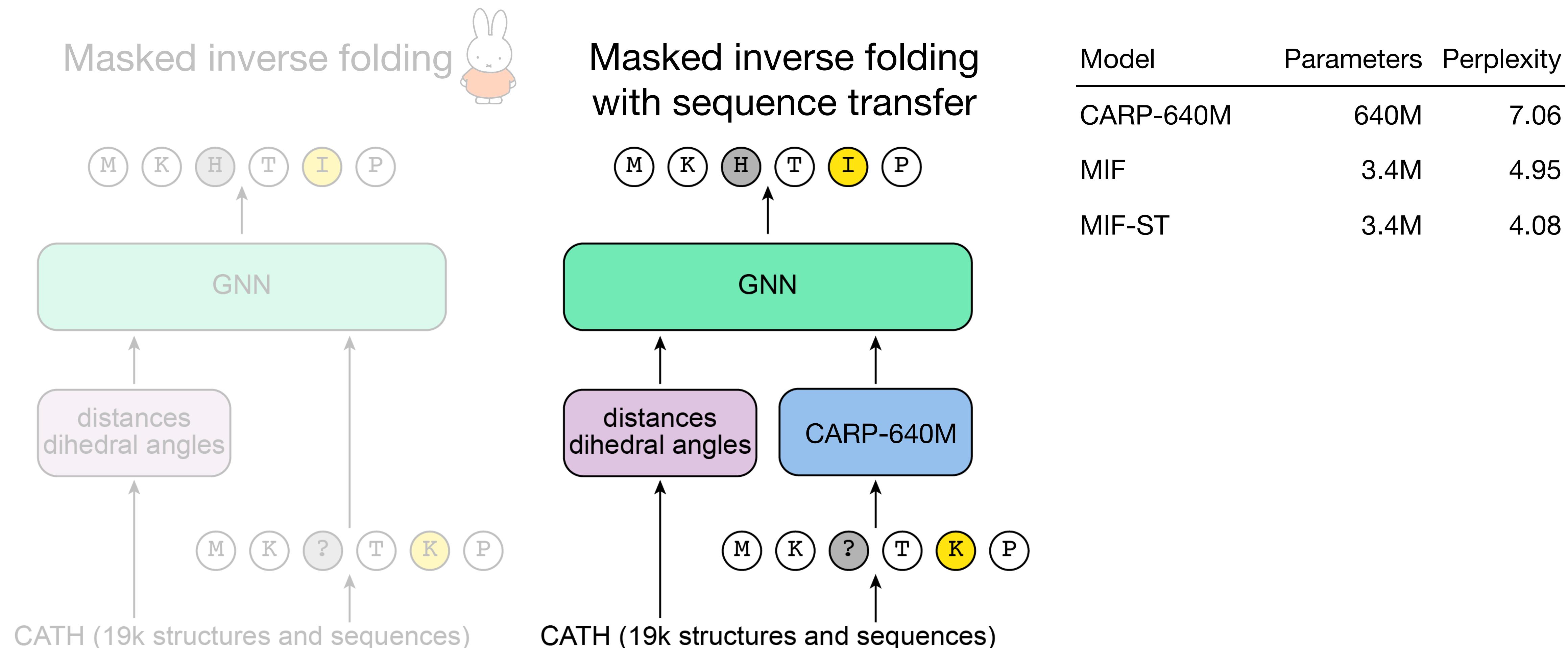
Sequence transfer improves pretraining more



Sequence transfer improves pretraining more

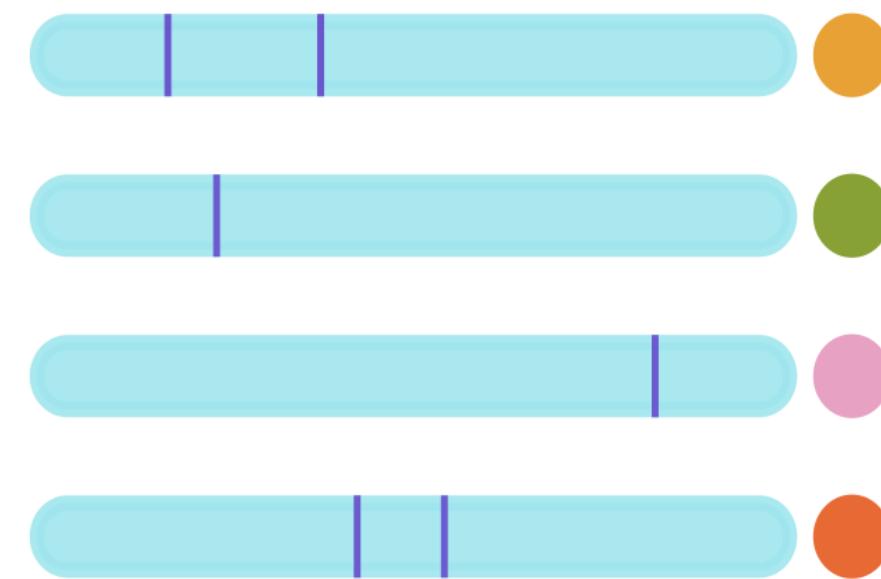


Sequence transfer improves pretraining more

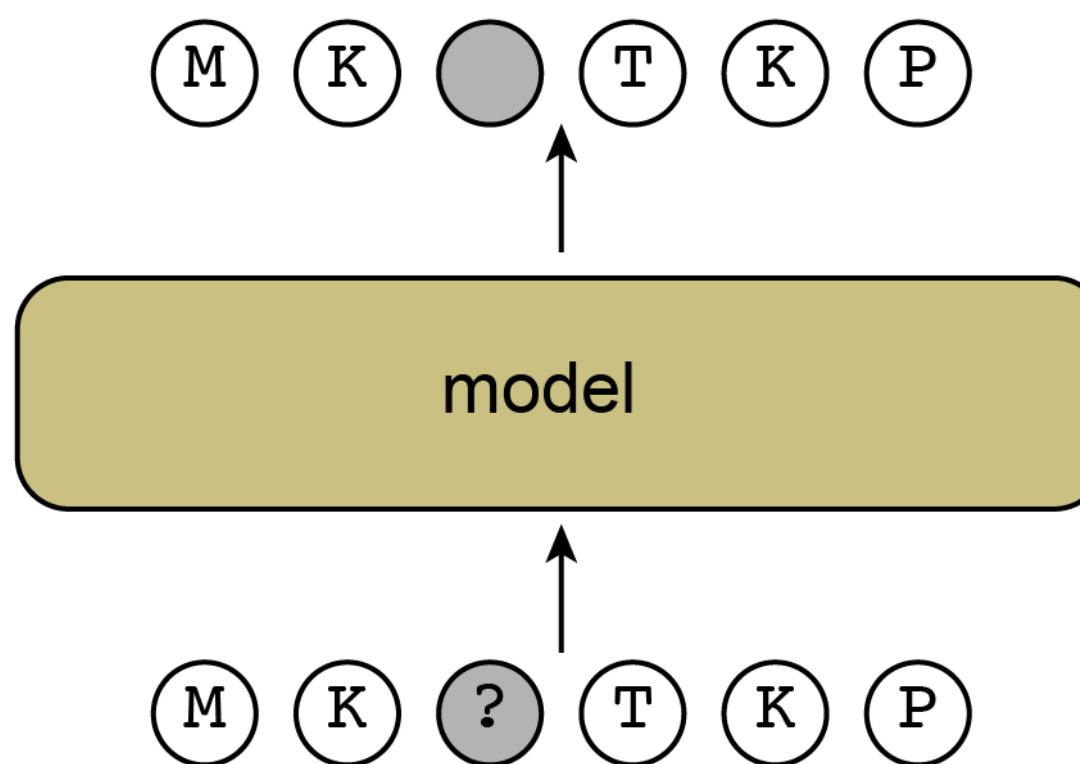


MIF, and MIF-ST are zero-shot fitness predictors

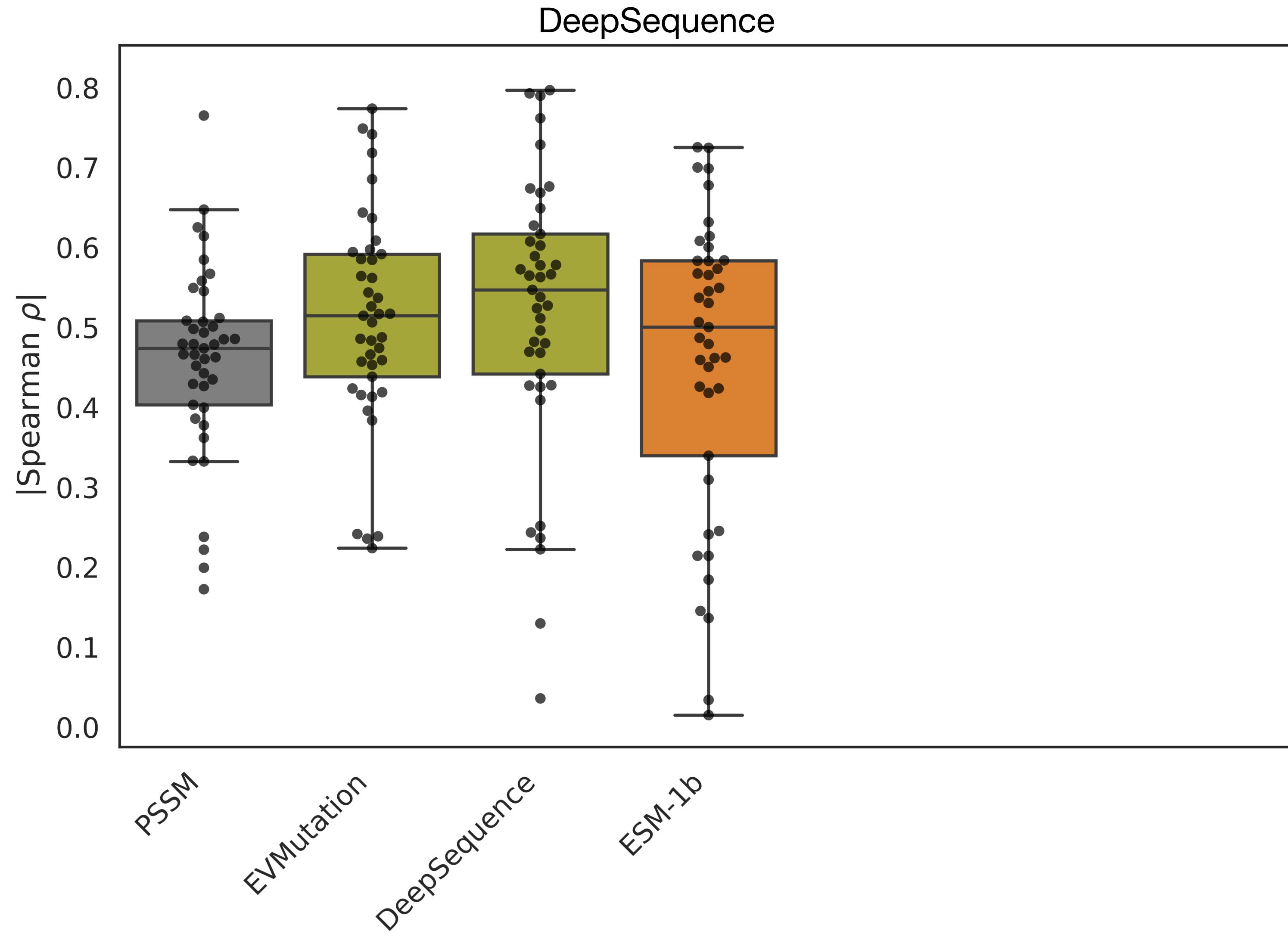
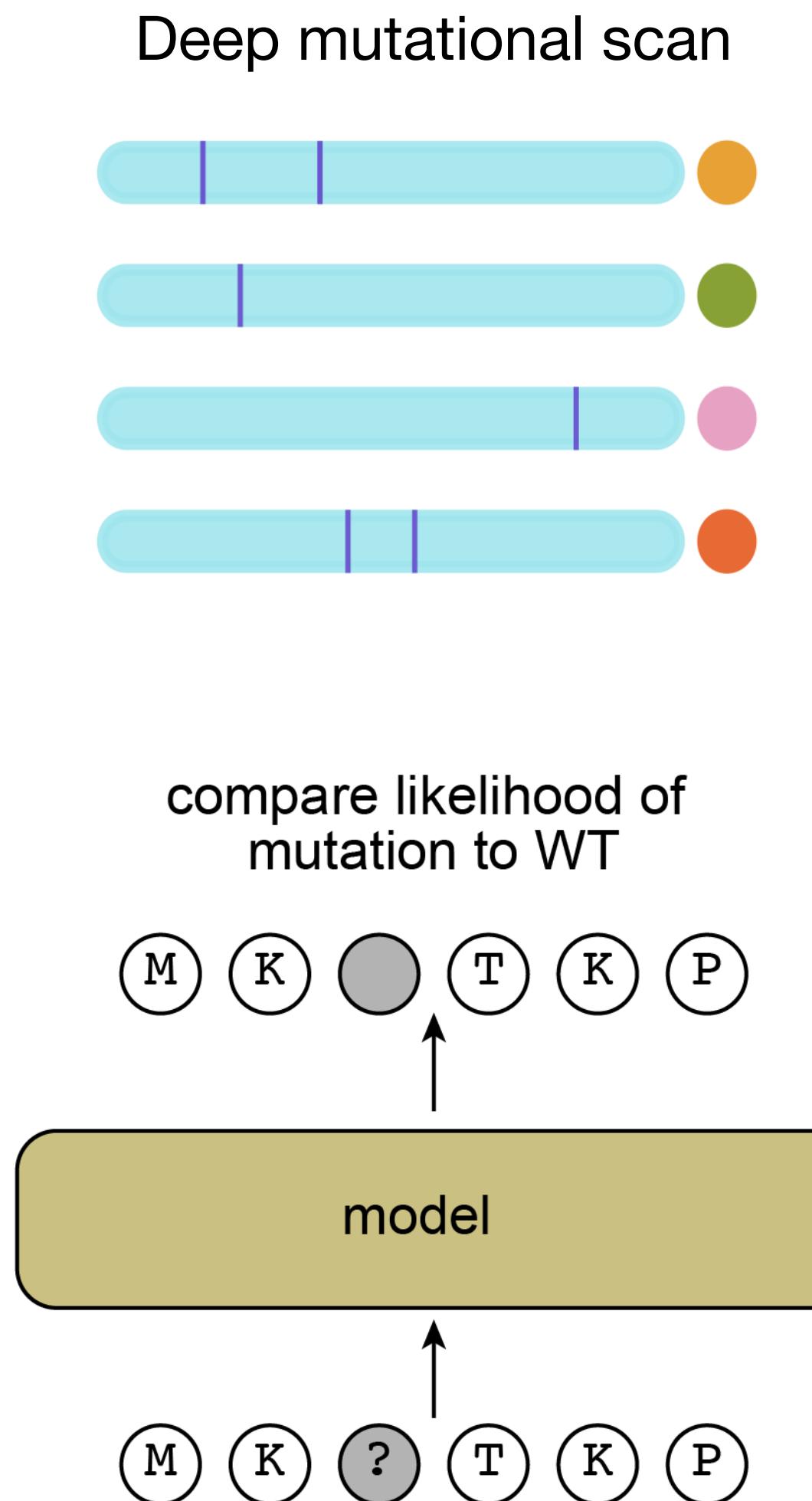
Deep mutational scan



compare likelihood of
mutation to WT

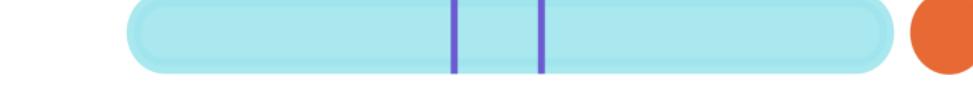
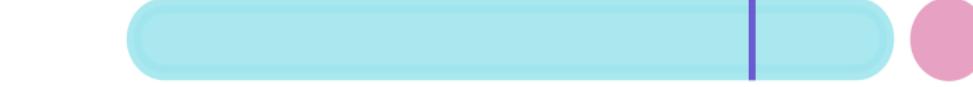


MIF, and MIF-ST are zero-shot fitness predictors



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

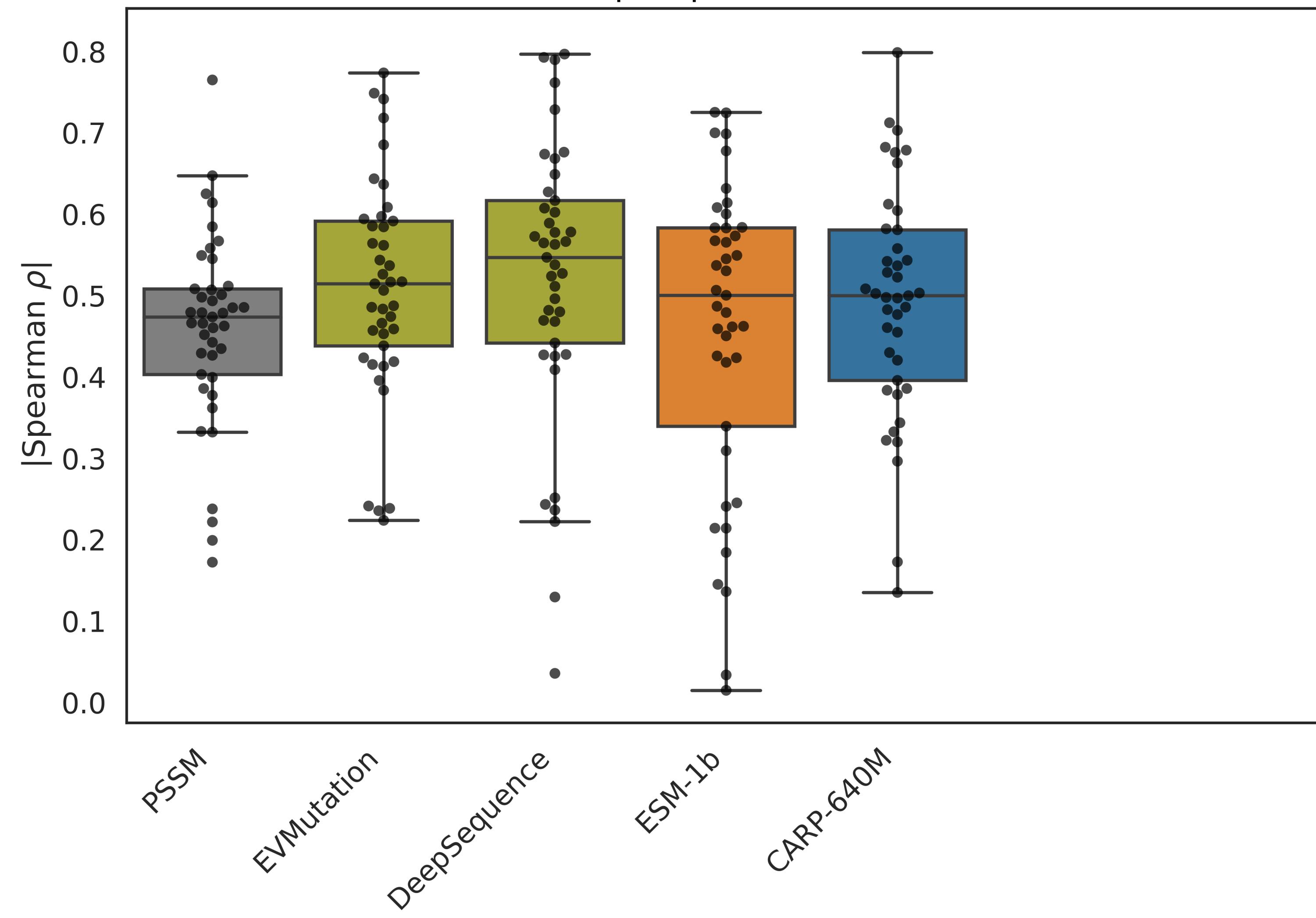
Deep mutational scan



compare likelihood of
mutation to WT

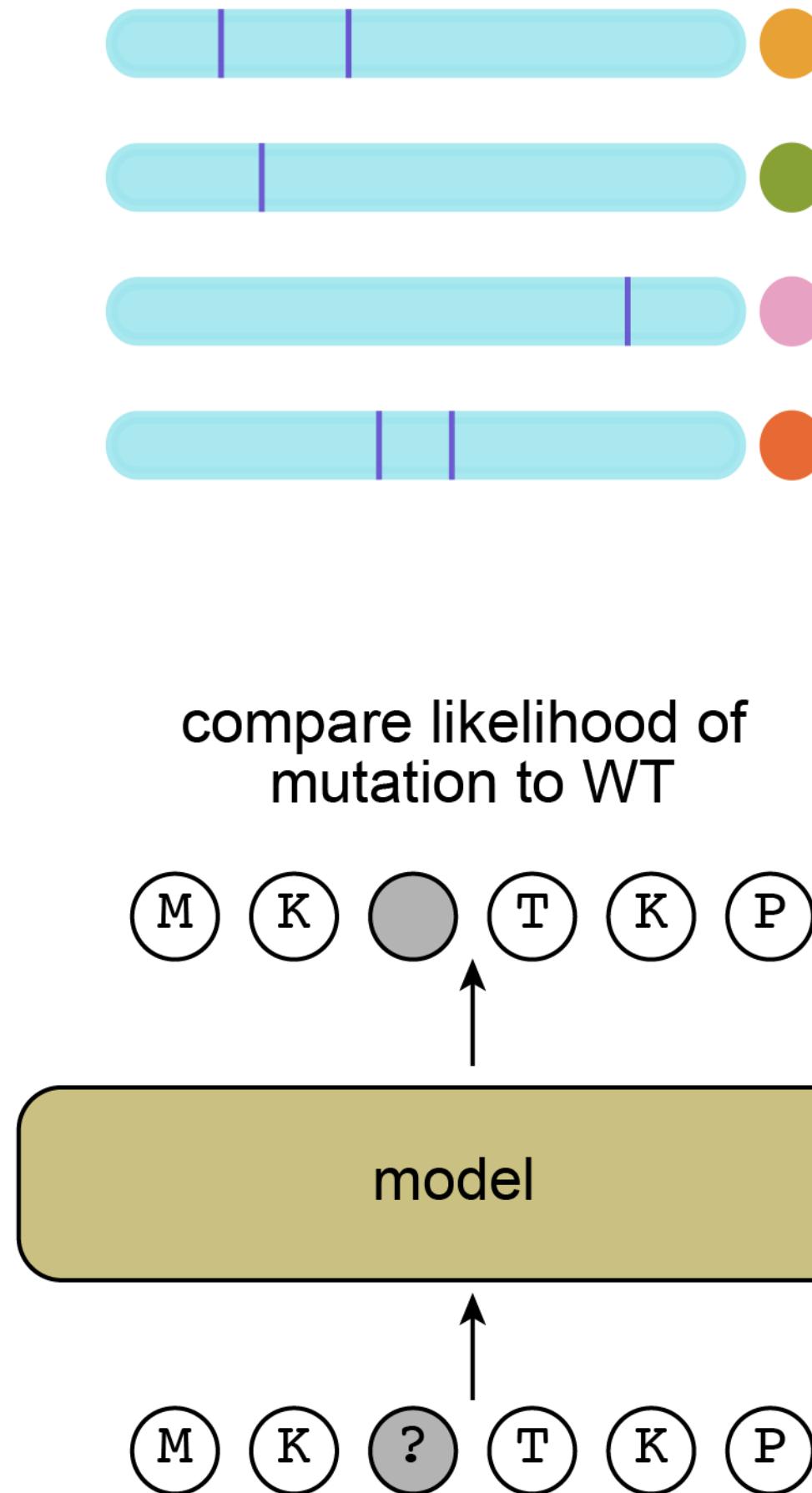


DeepSequence

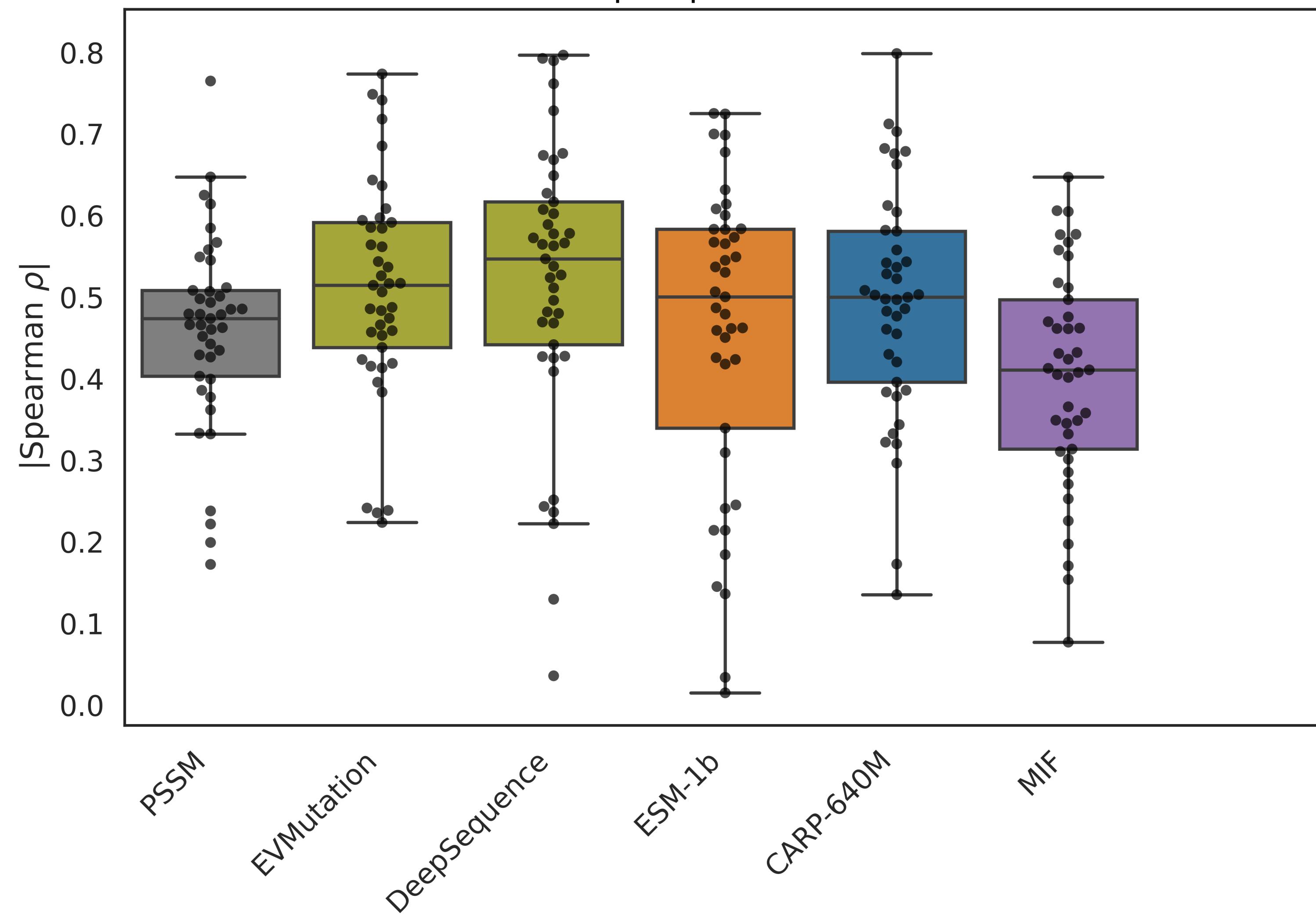


MIF, and MIF-ST are zero-shot fitness predictors

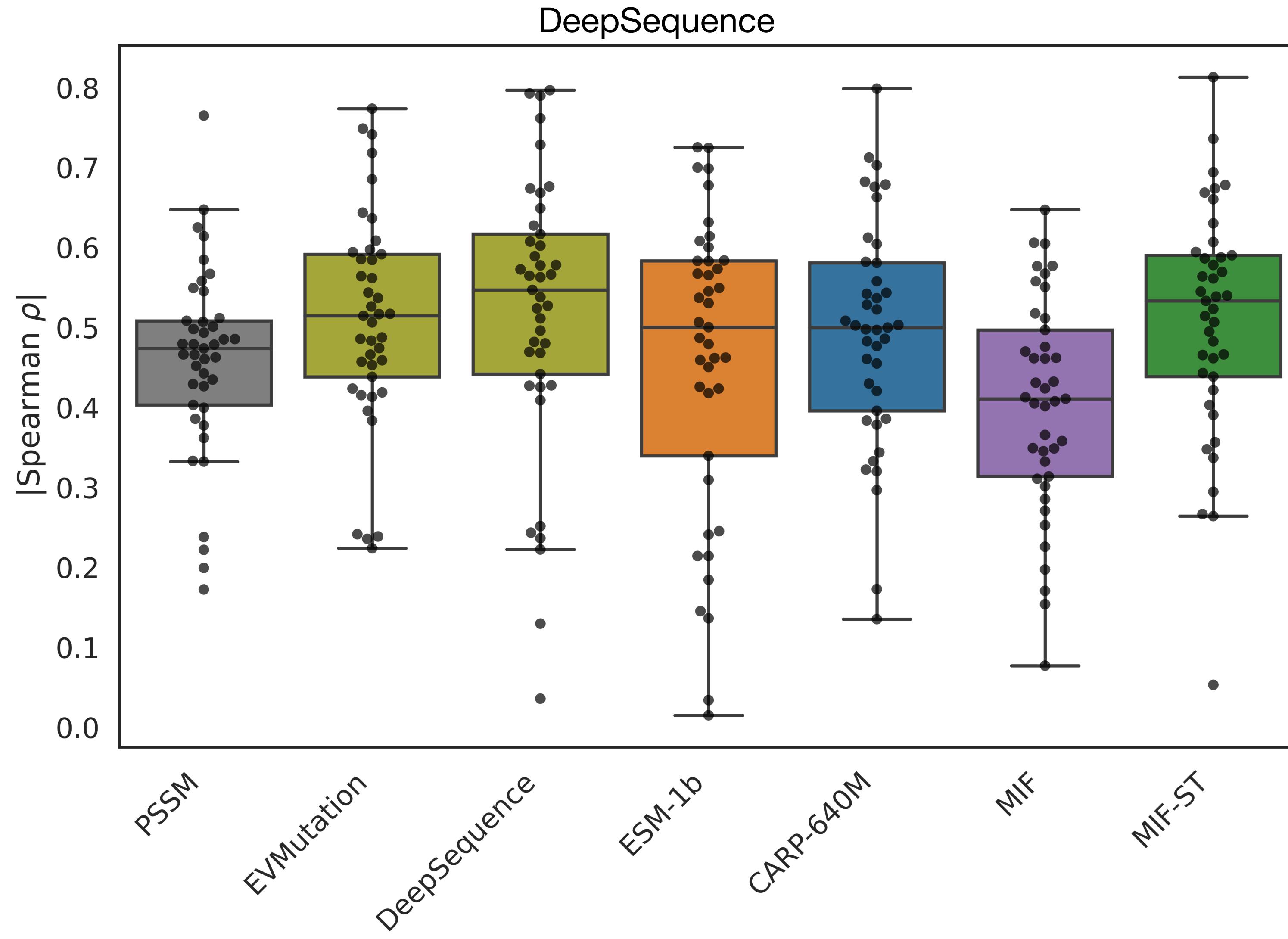
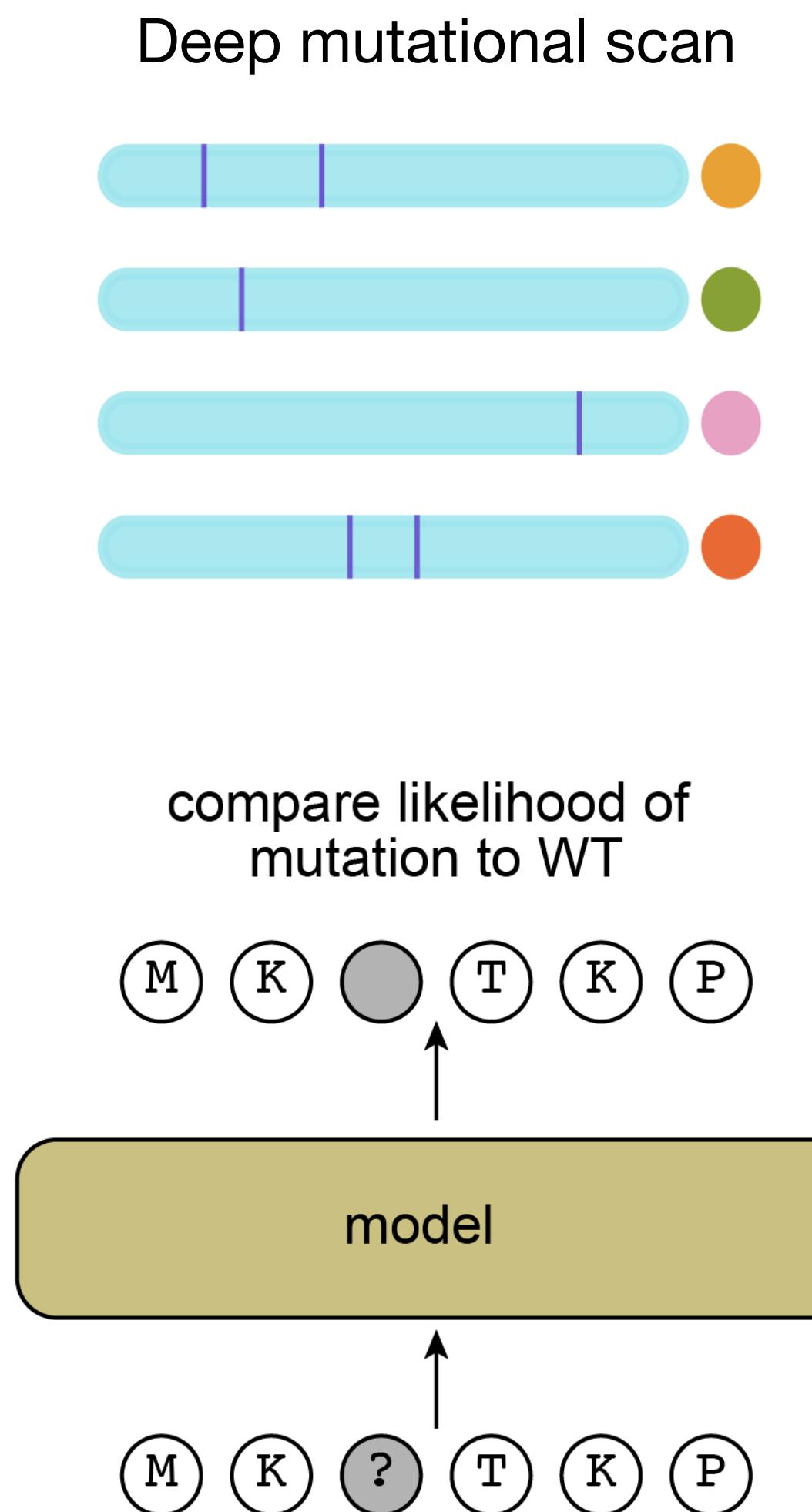
Deep mutational scan



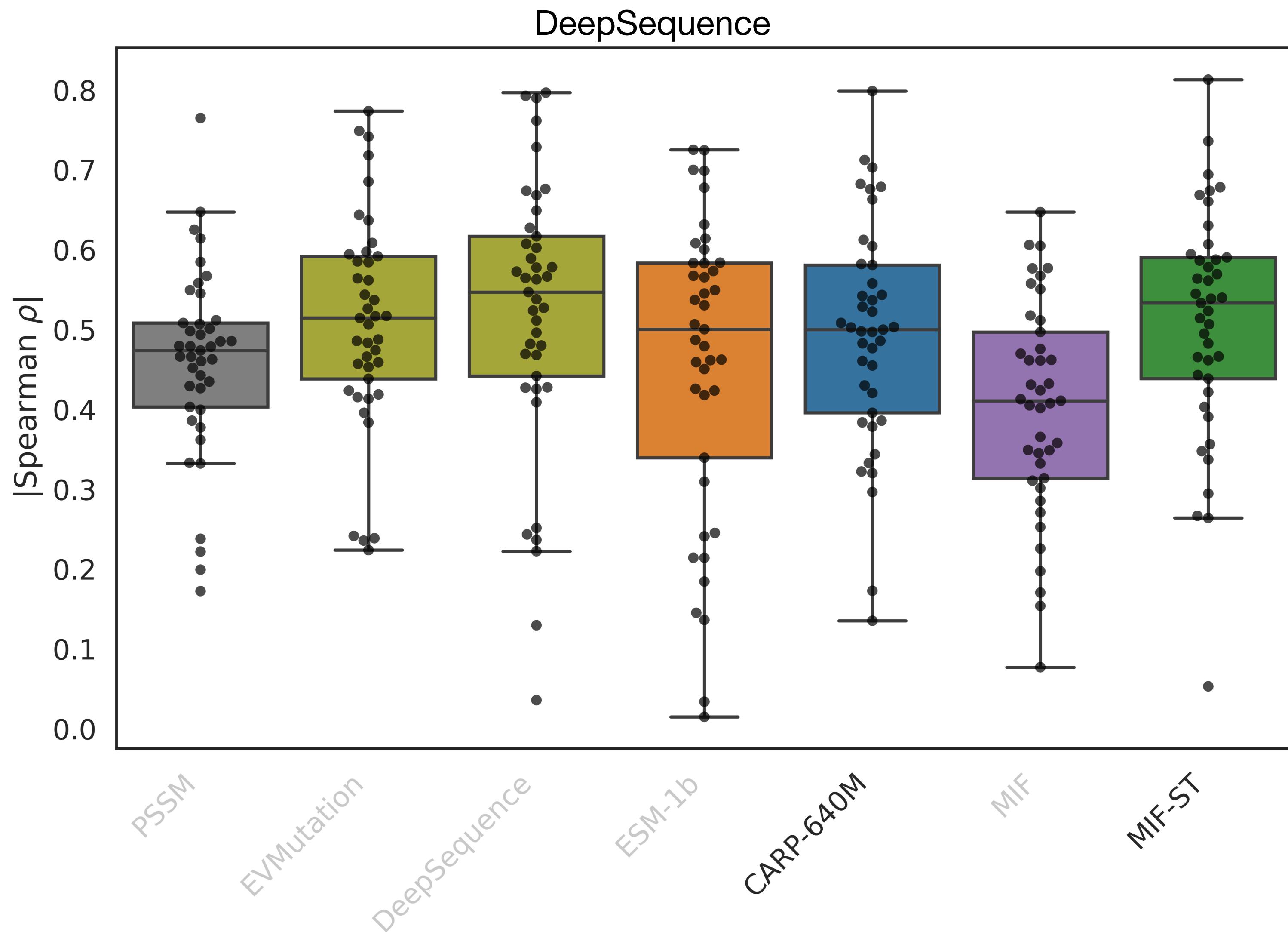
DeepSequence



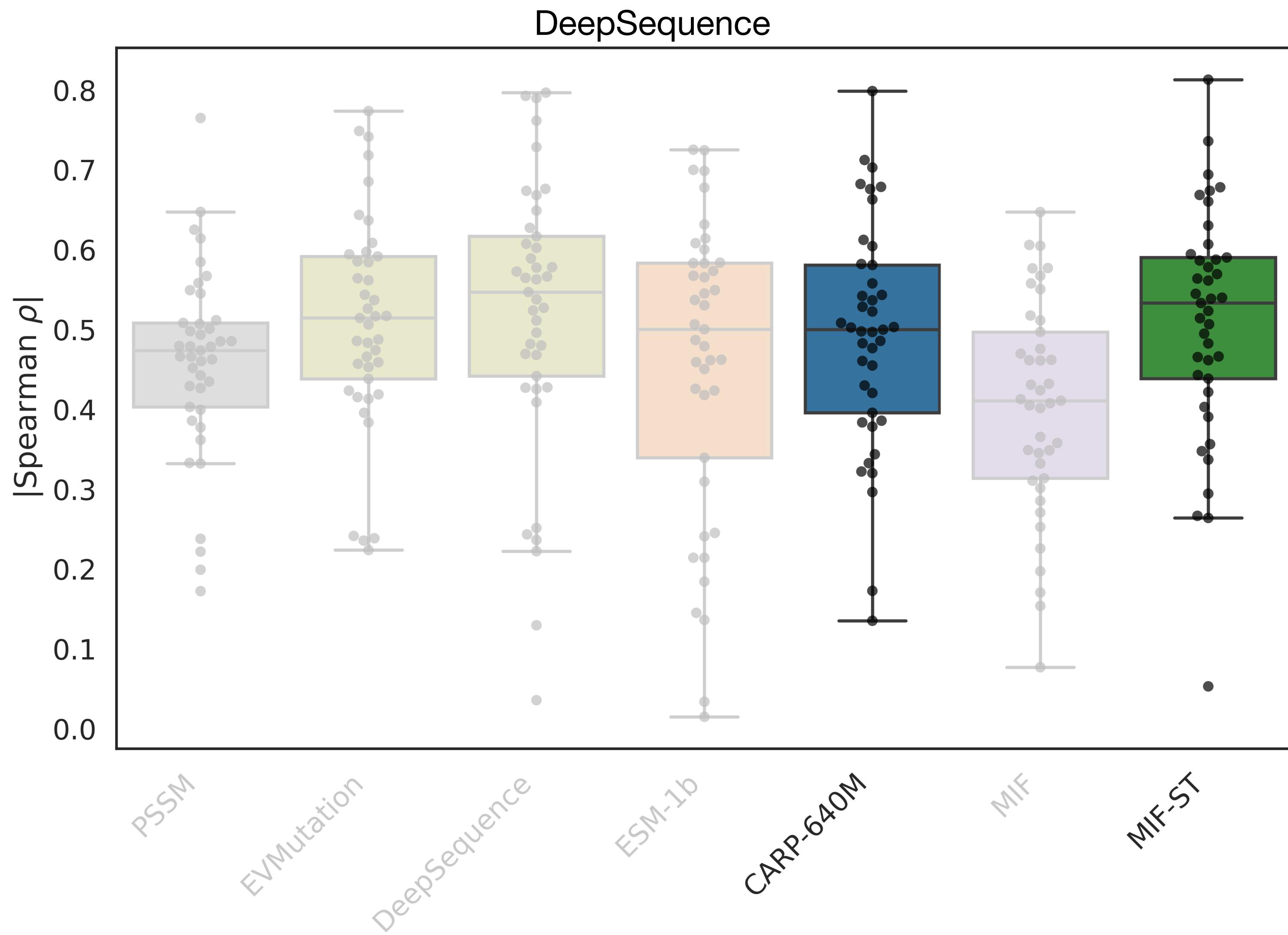
MIF, and MIF-ST are zero-shot fitness predictors



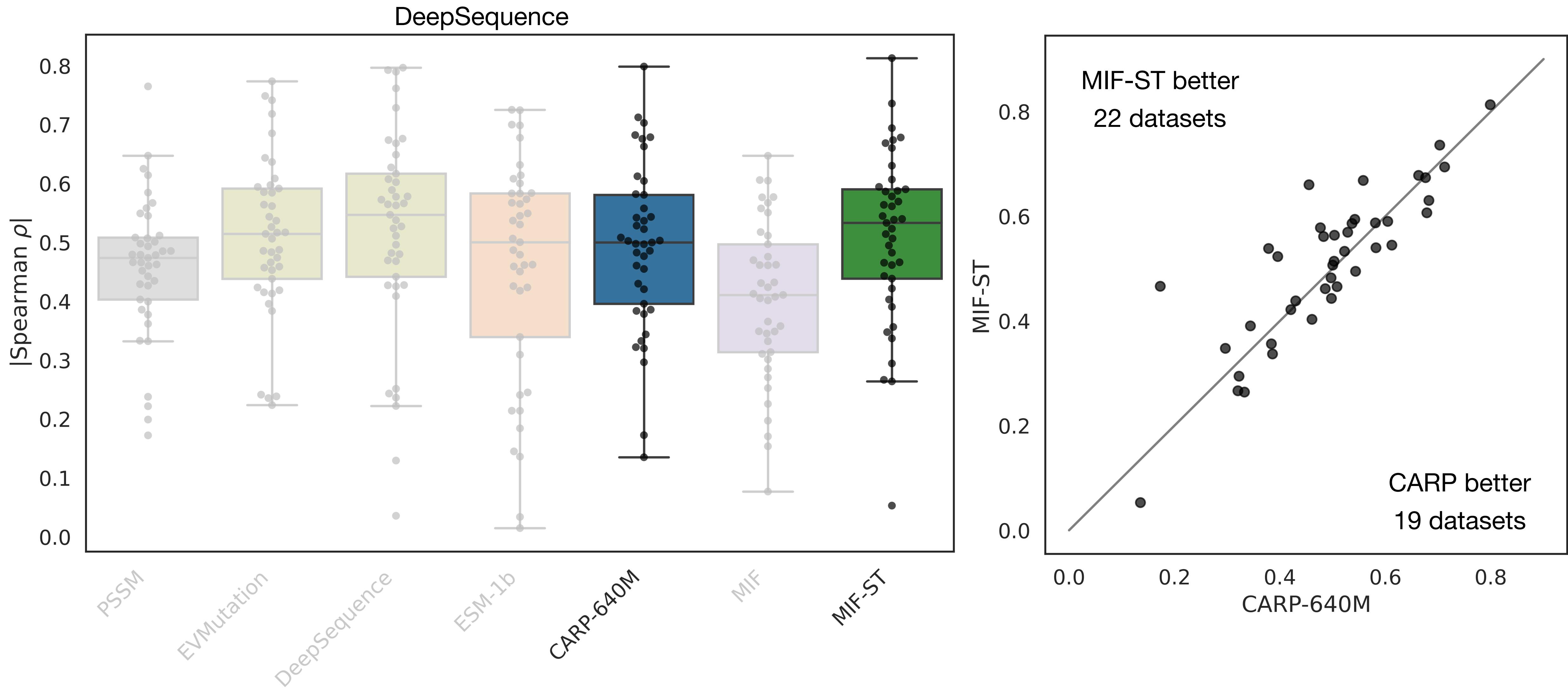
MIF-ST outperforms CARP and MIF on DeepSequence



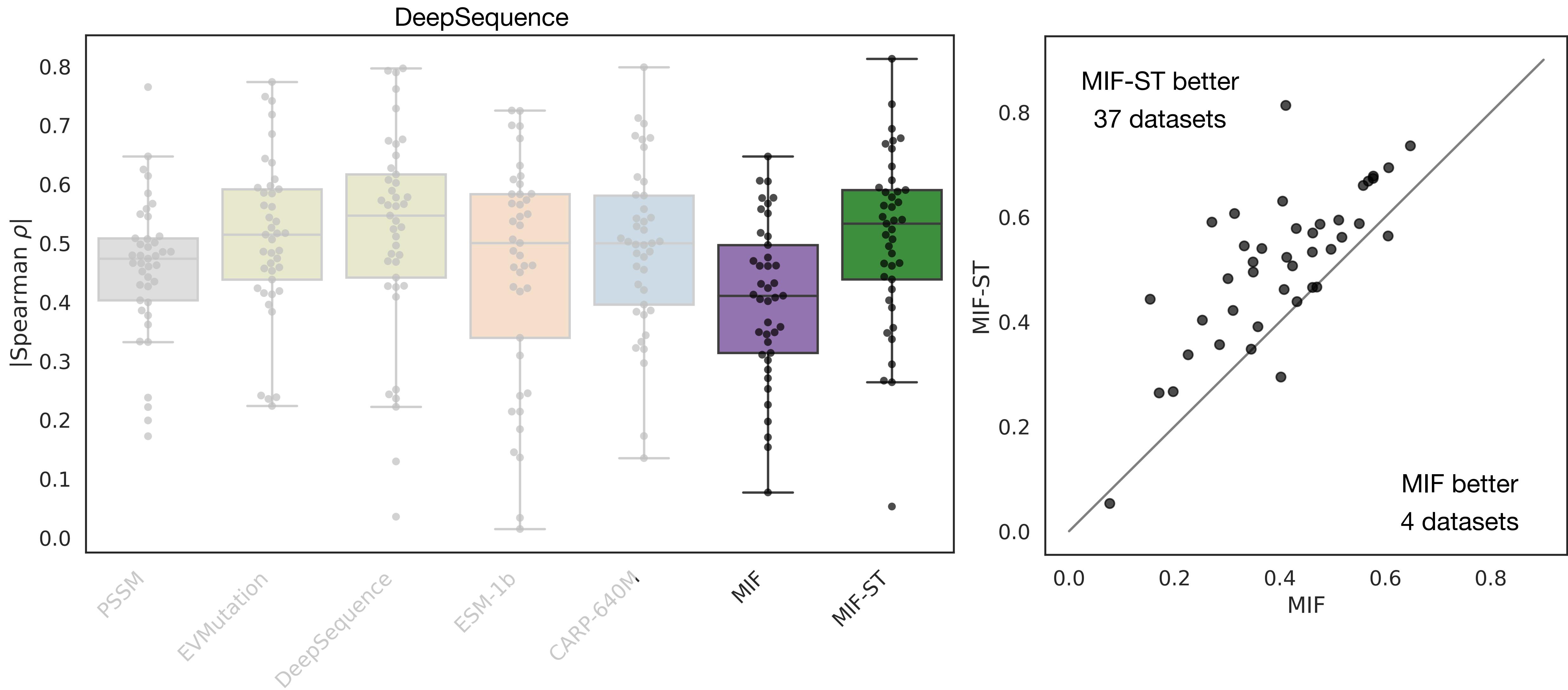
MIF-ST outperforms CARP and MIF on DeepSequence



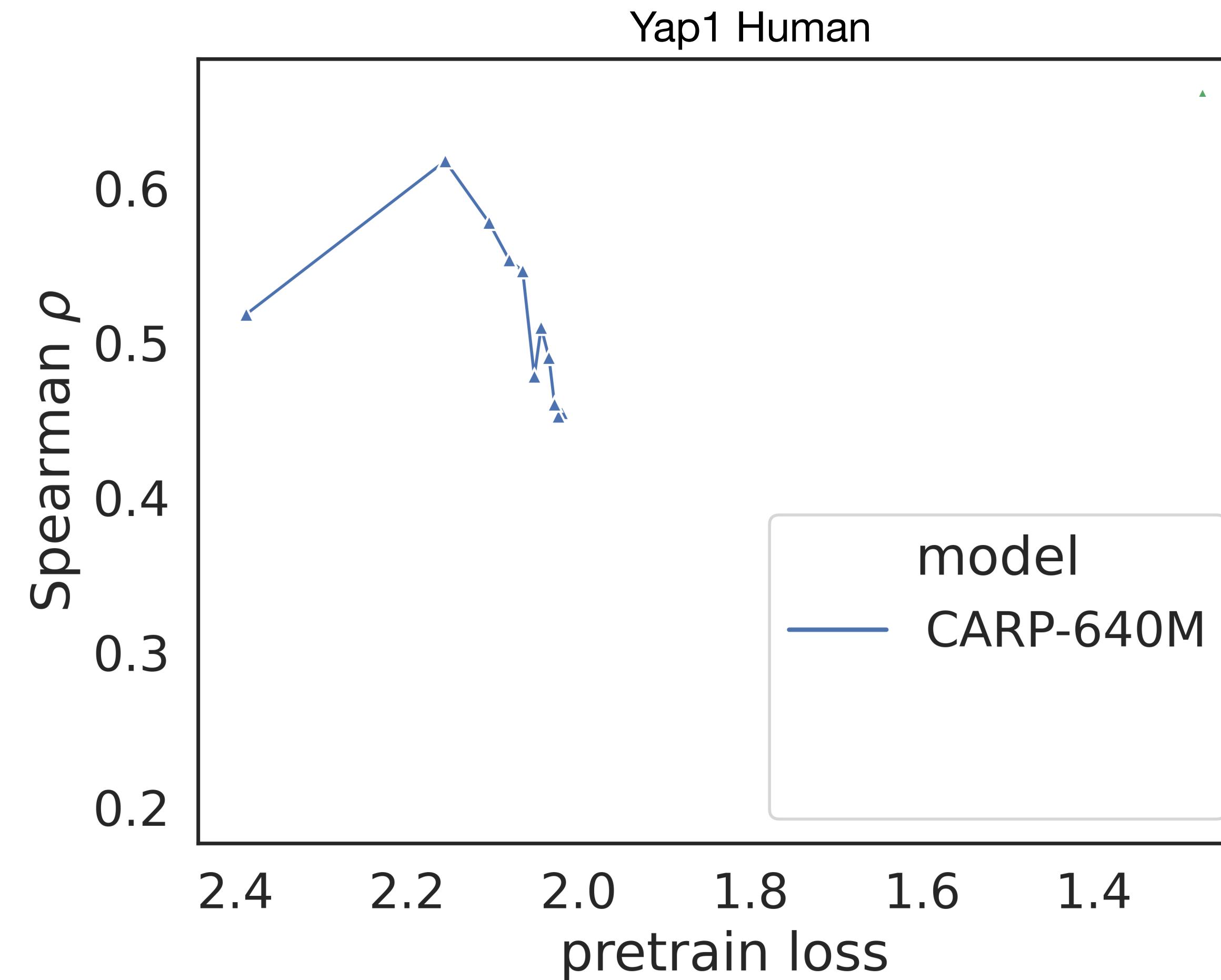
MIF-ST outperforms CARP and MIF on DeepSequence



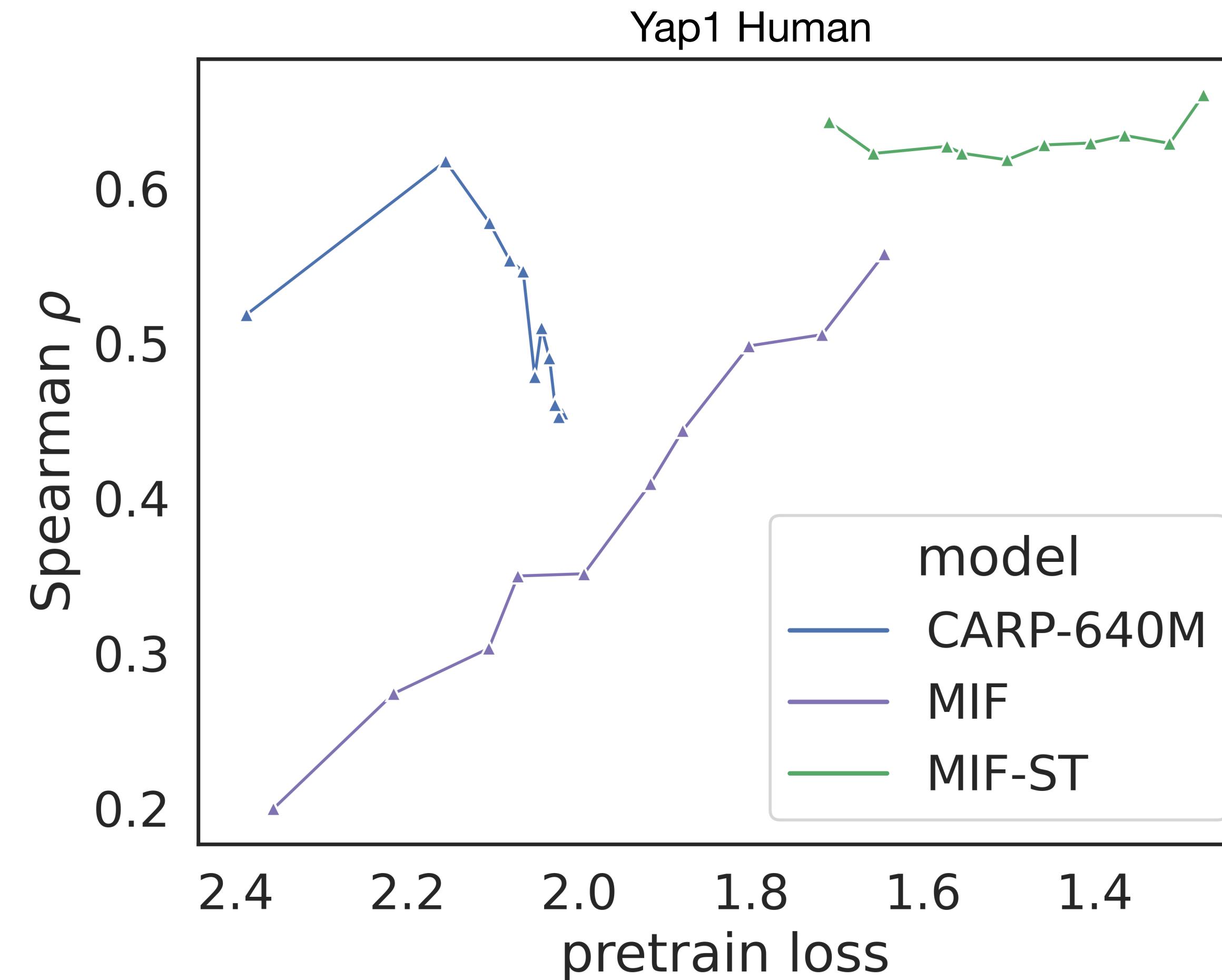
MIF-ST outperforms CARP and MIF on DeepSequence



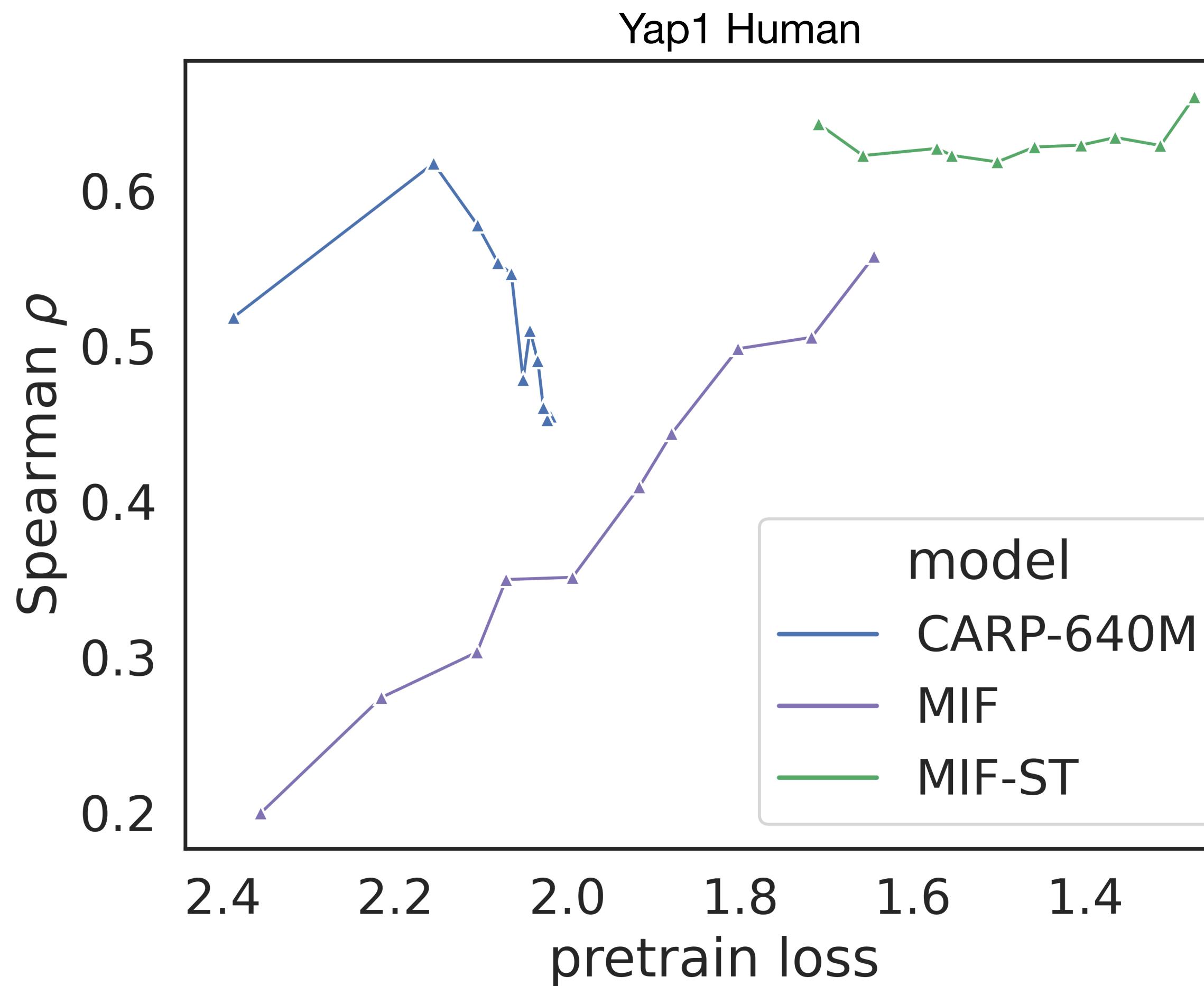
Structure-conditioned zero-shot improves more consistently with pretraining



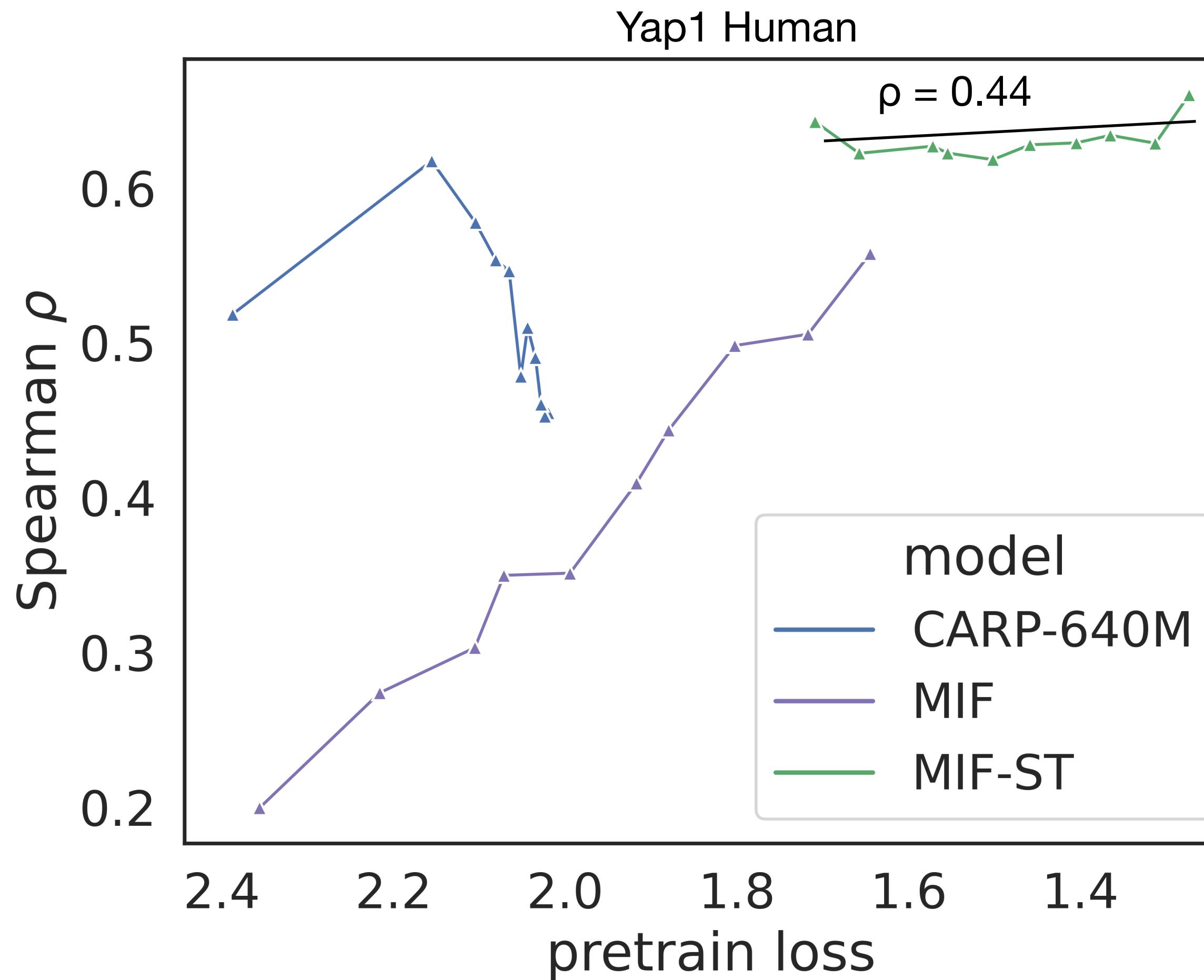
Structure-conditioned zero-shot improves more consistently with pretraining



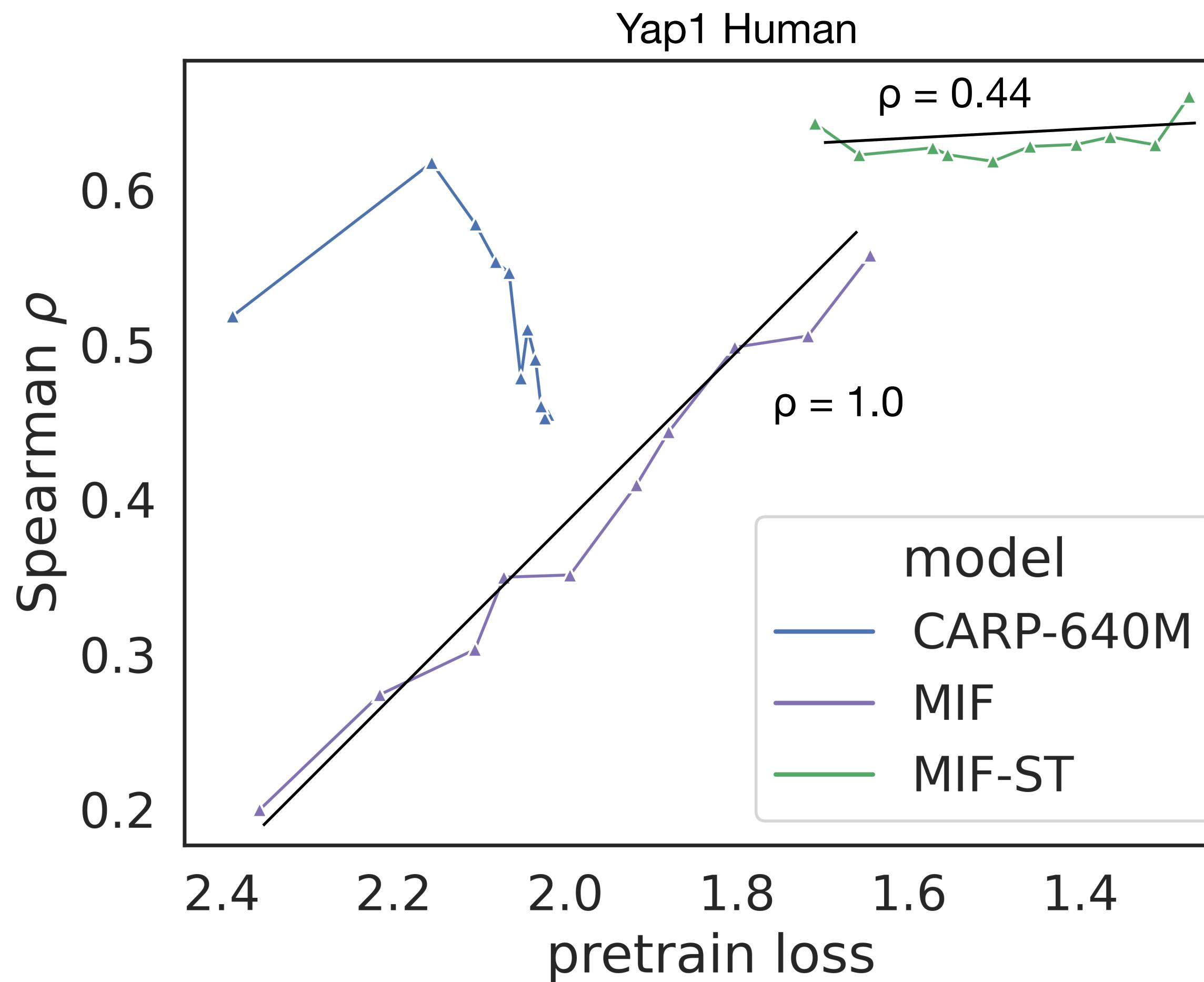
Structure-conditioned zero-shot improves more consistently with pretraining



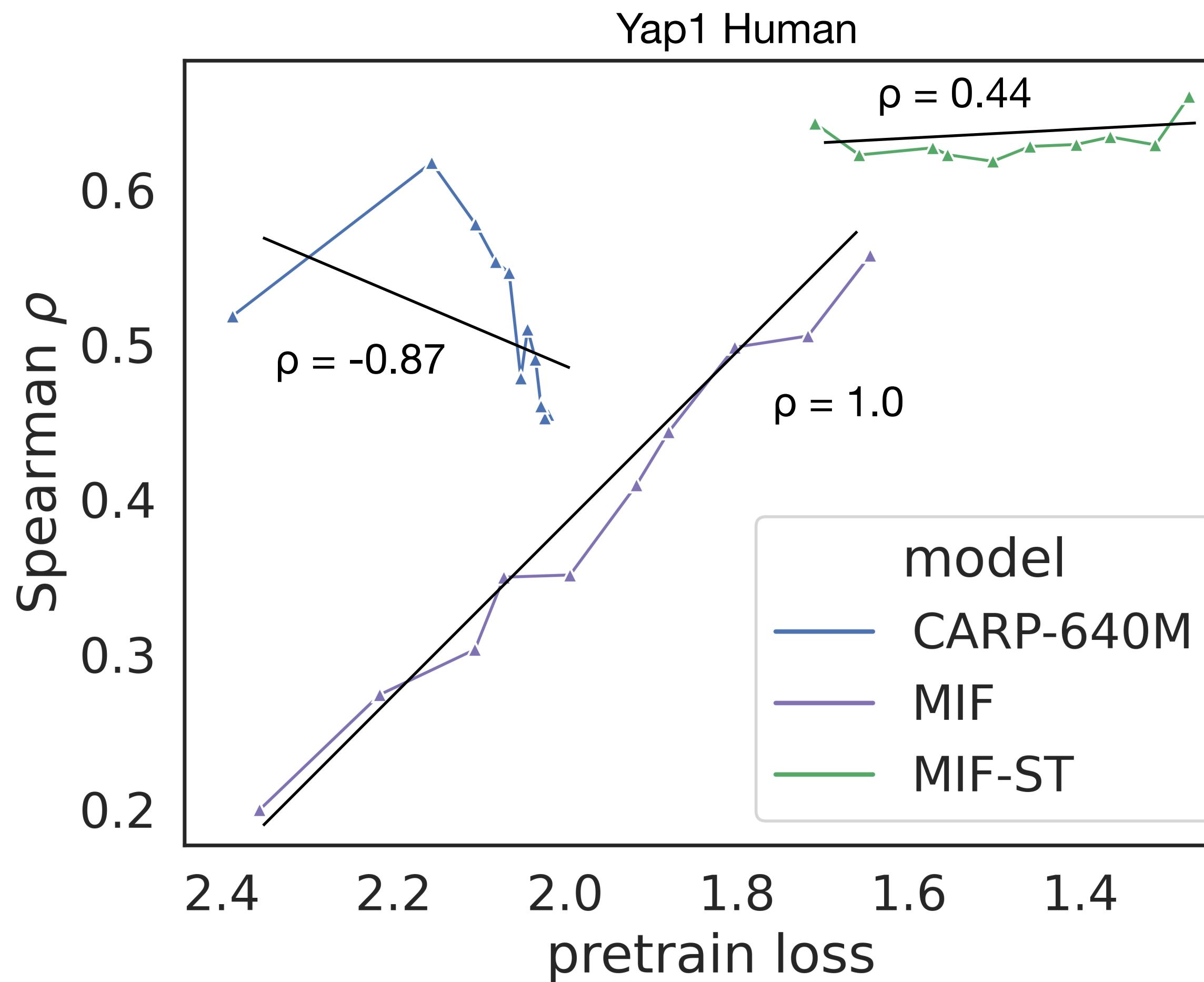
Structure-conditioned zero-shot improves more consistently with pretraining



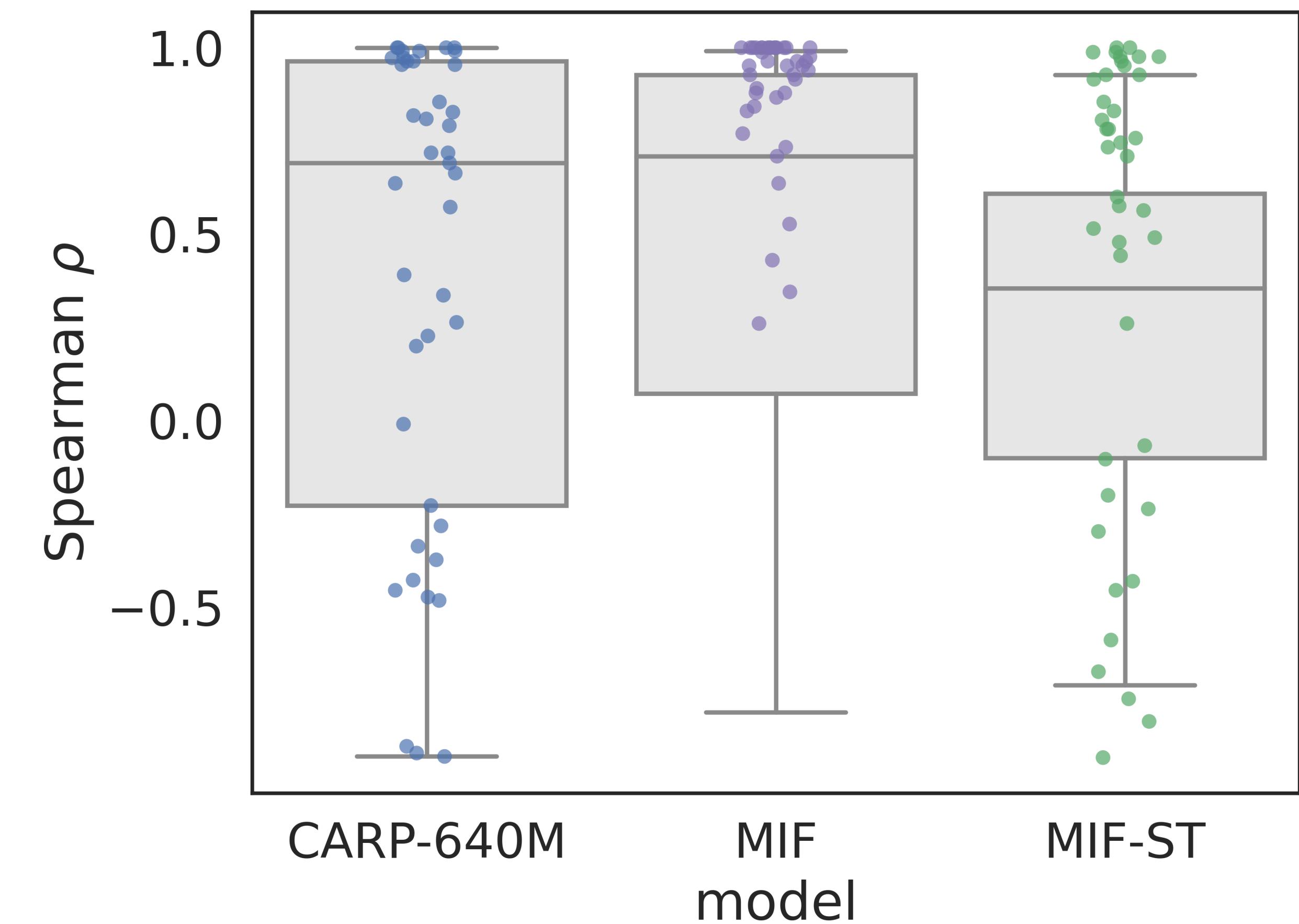
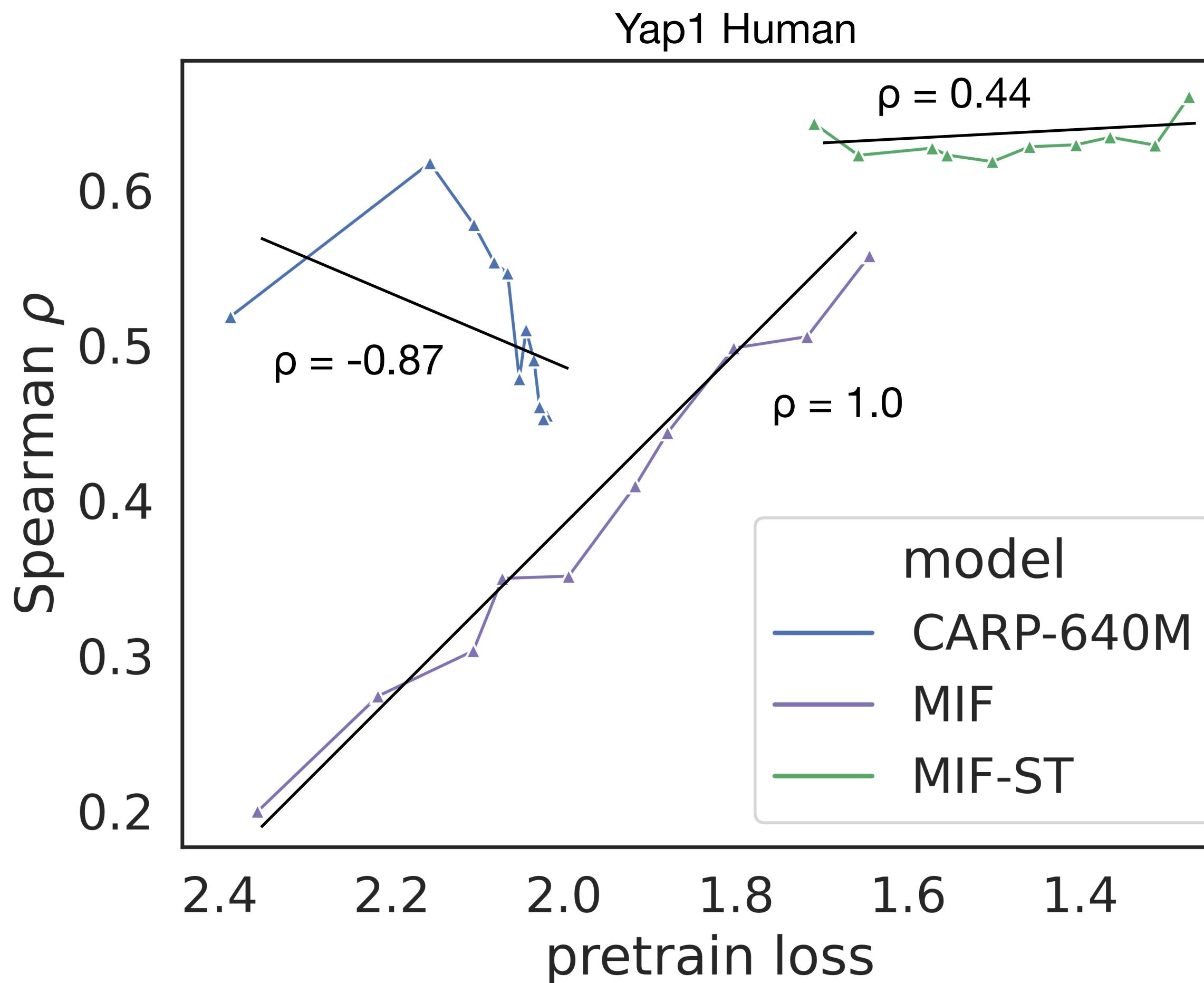
Structure-conditioned zero-shot improves more consistently with pretraining



Structure-conditioned zero-shot improves more consistently with pretraining

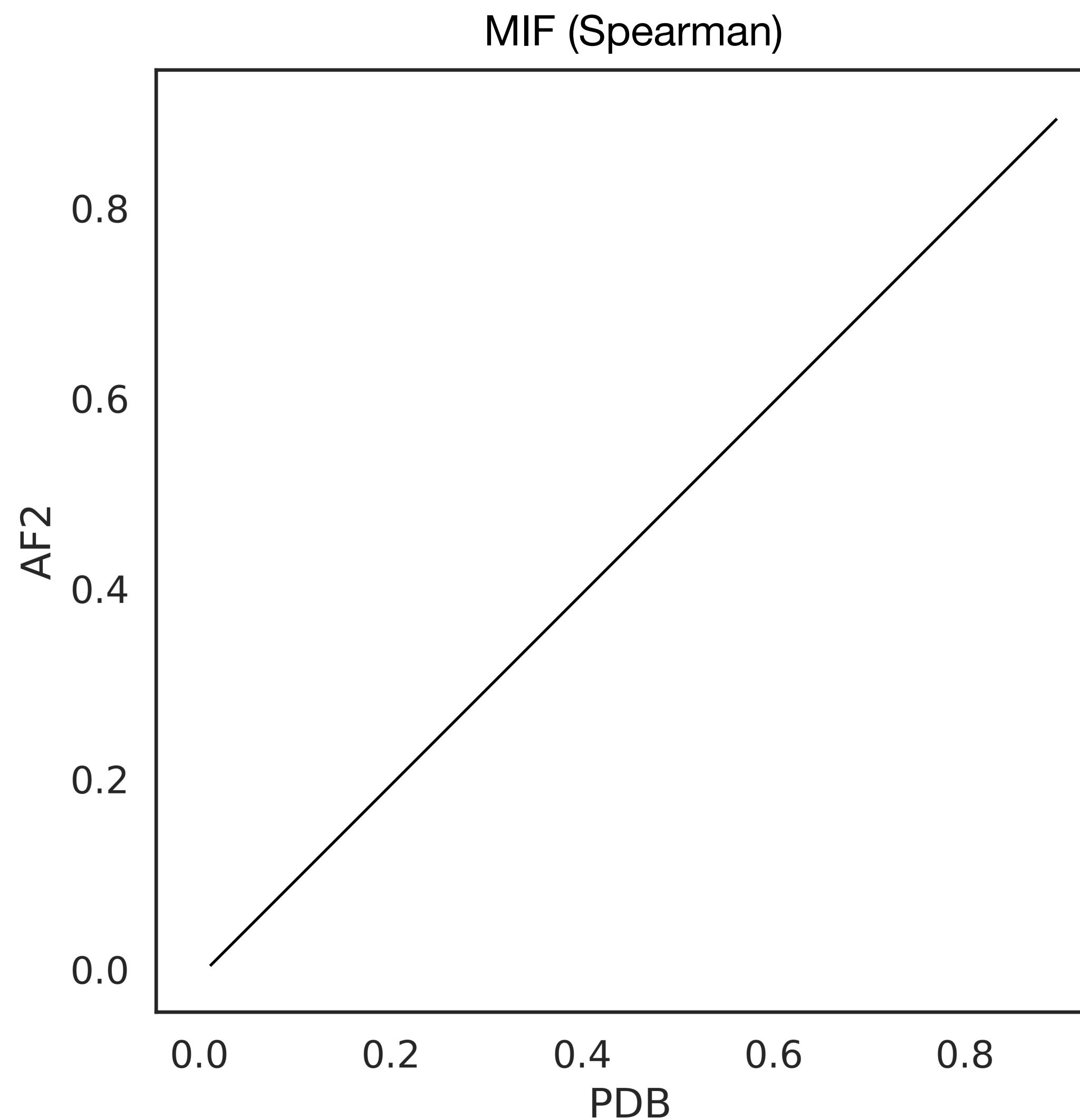


Structure-conditioned zero-shot improves more consistently with pretraining

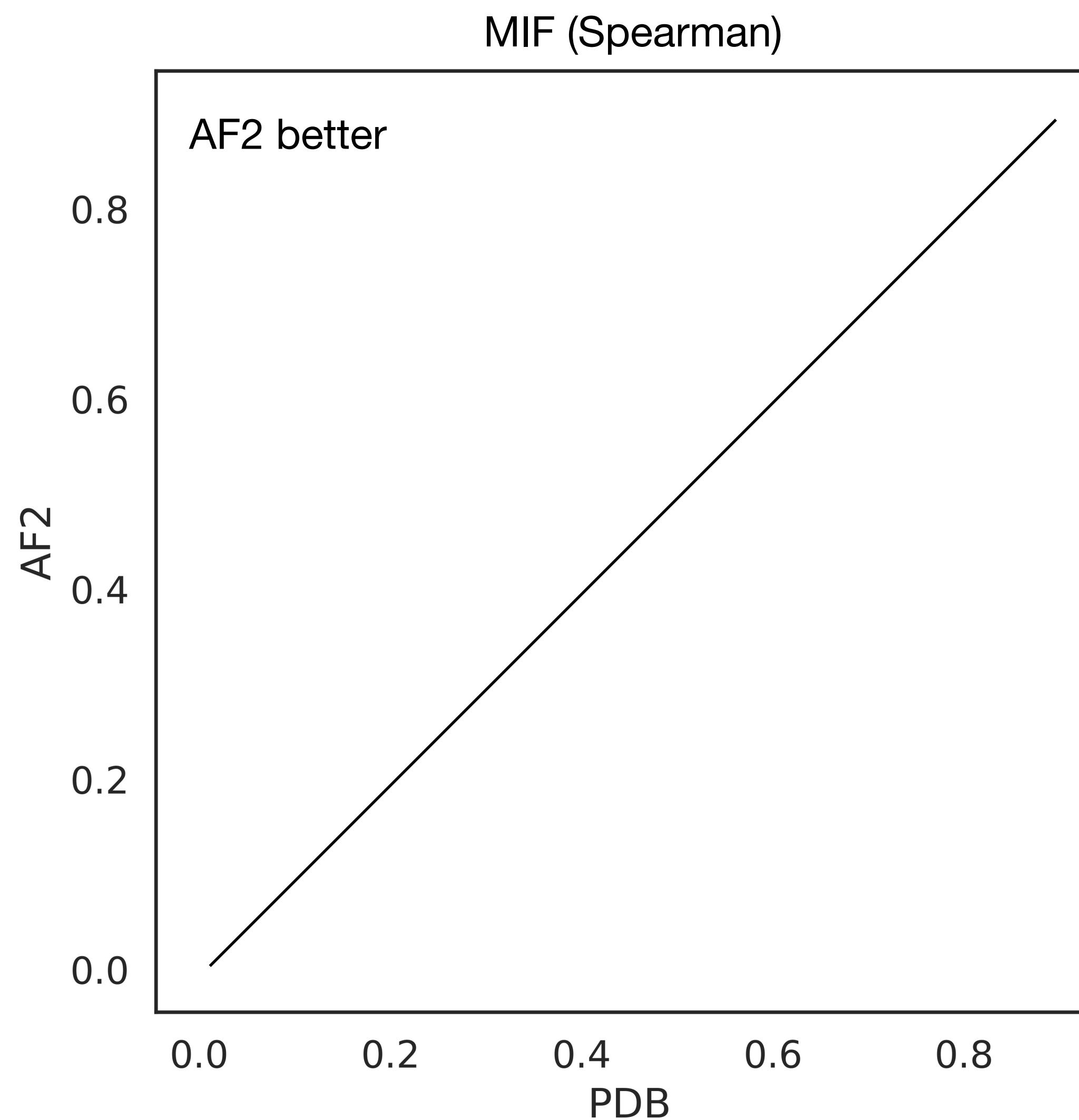


Predictions often better using AF2 structures

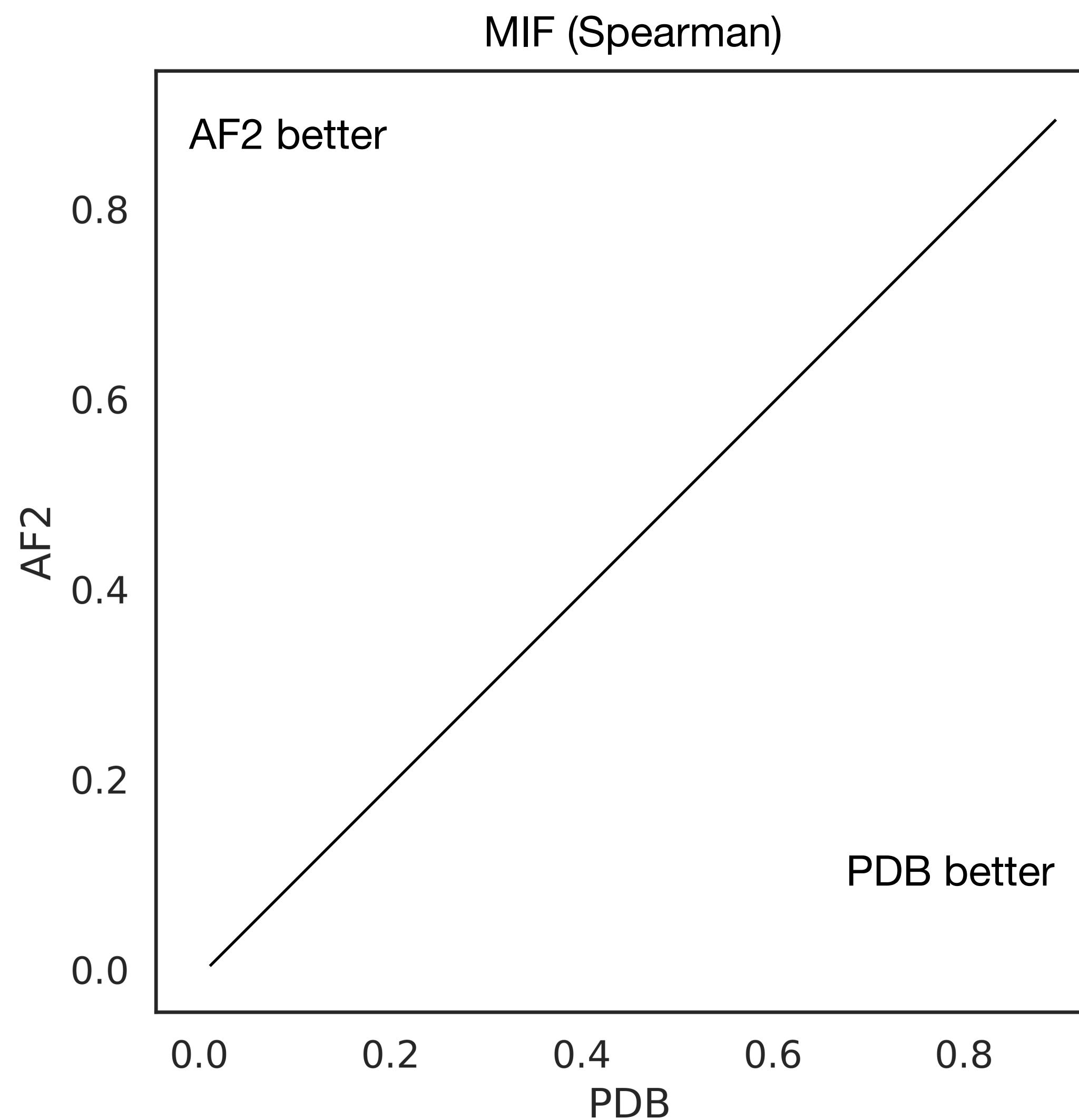
Predictions often better using AF2 structures



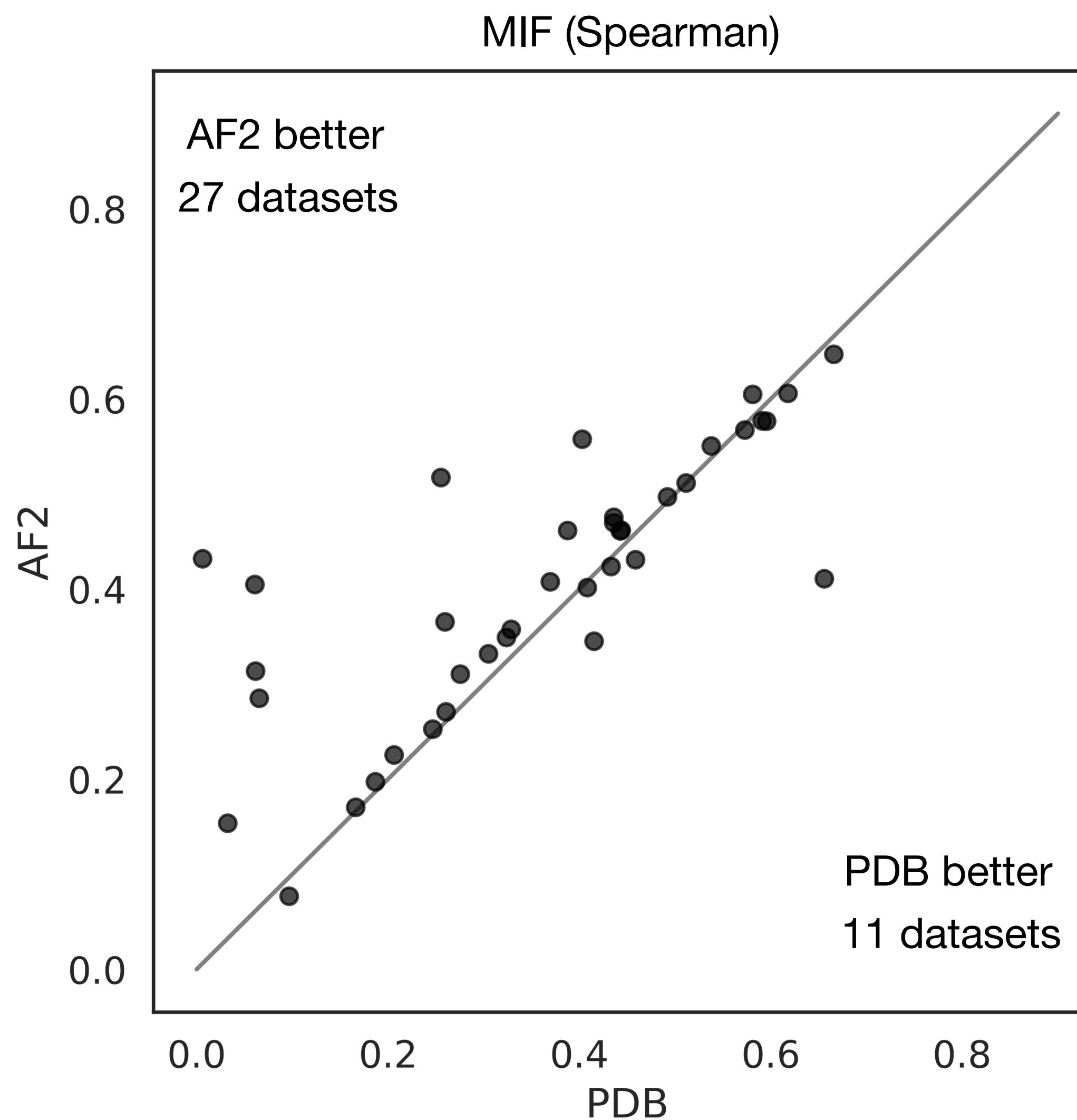
Predictions often better using AF2 structures



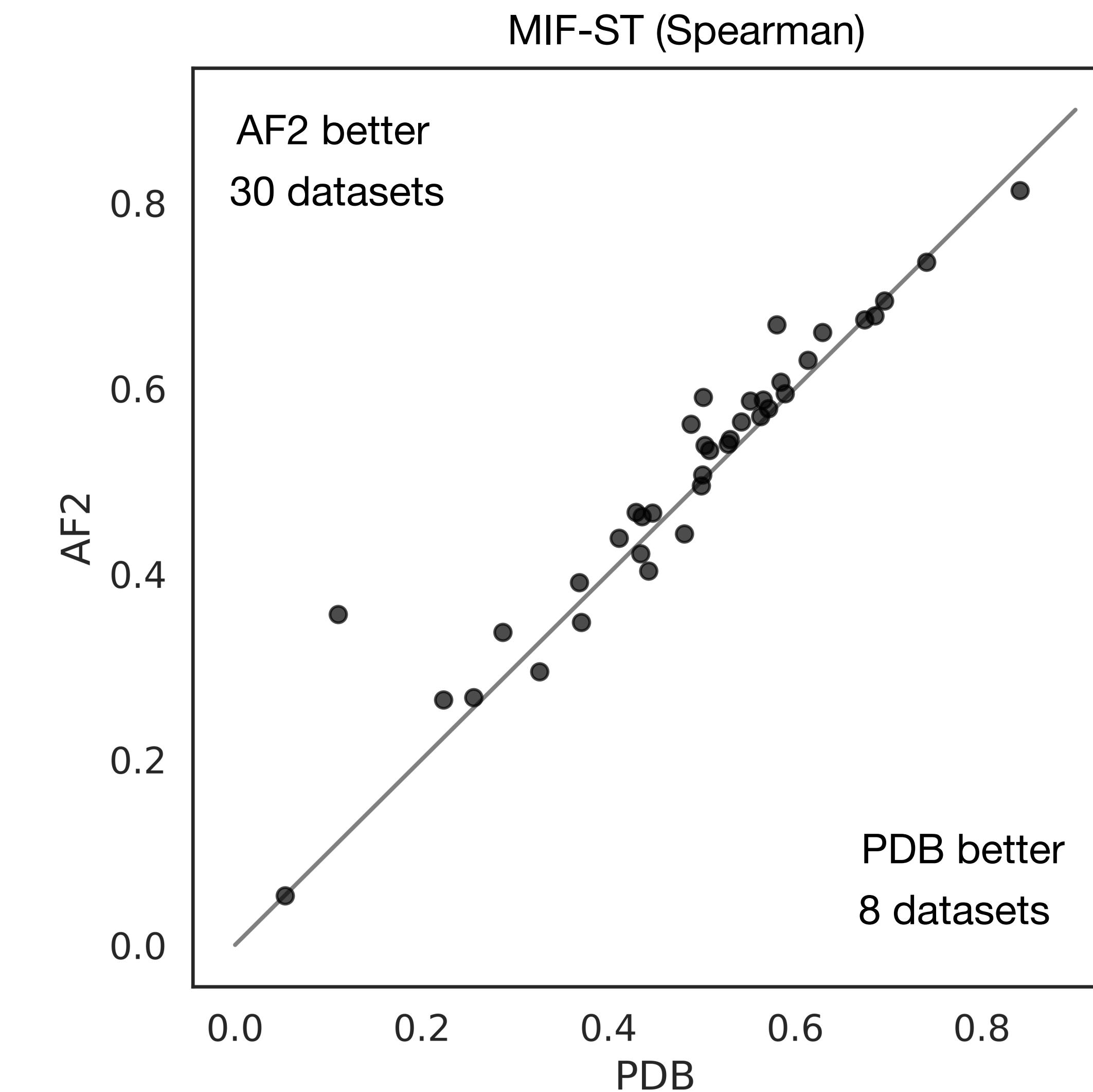
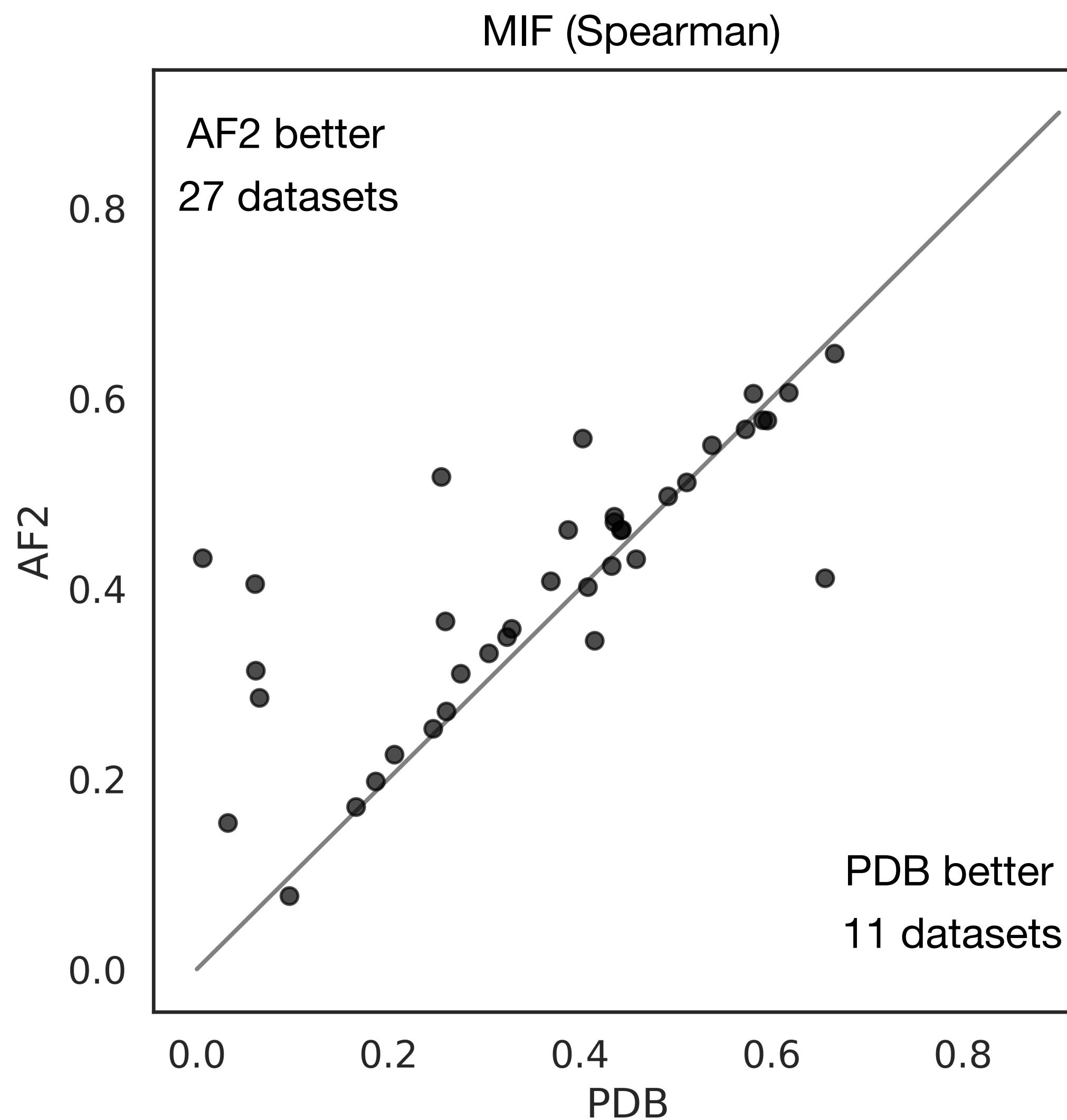
Predictions often better using AF2 structures



Predictions often better using AF2 structures



Predictions often better using AF2 structures



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

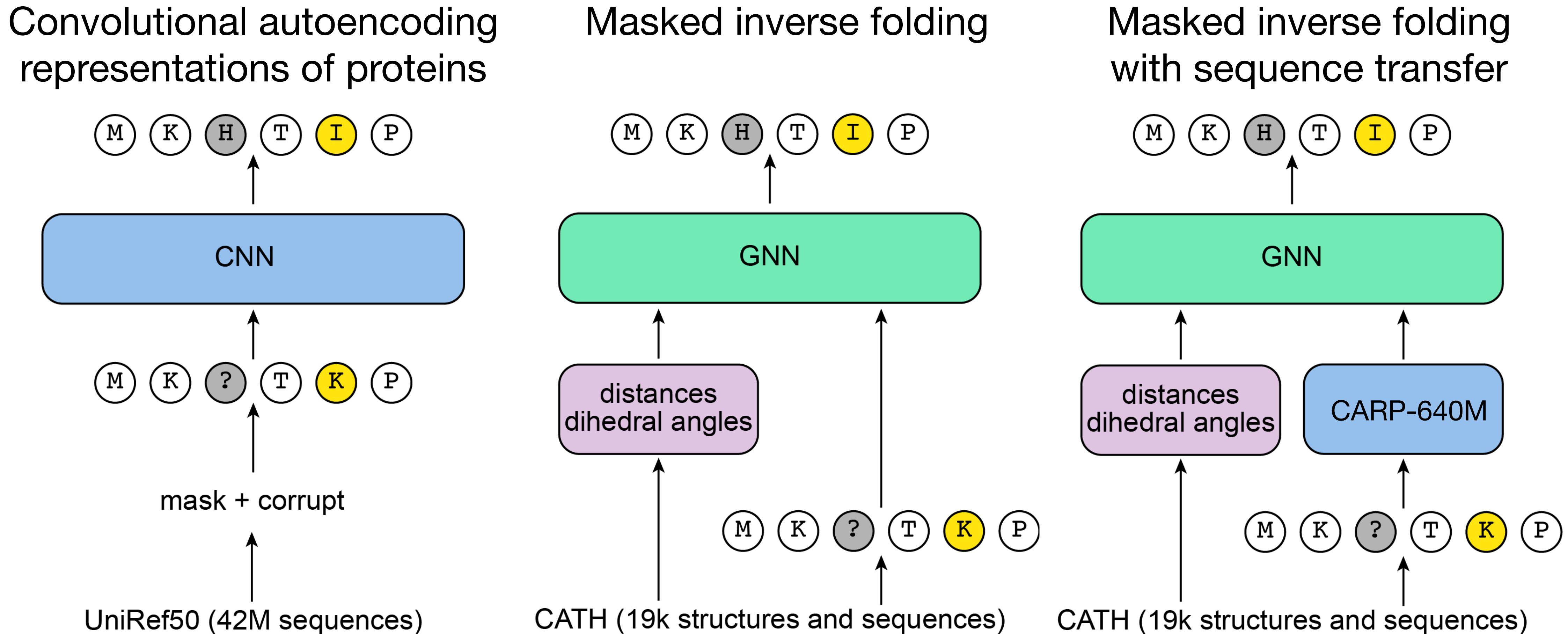
Structure and sequence transfer help a little

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

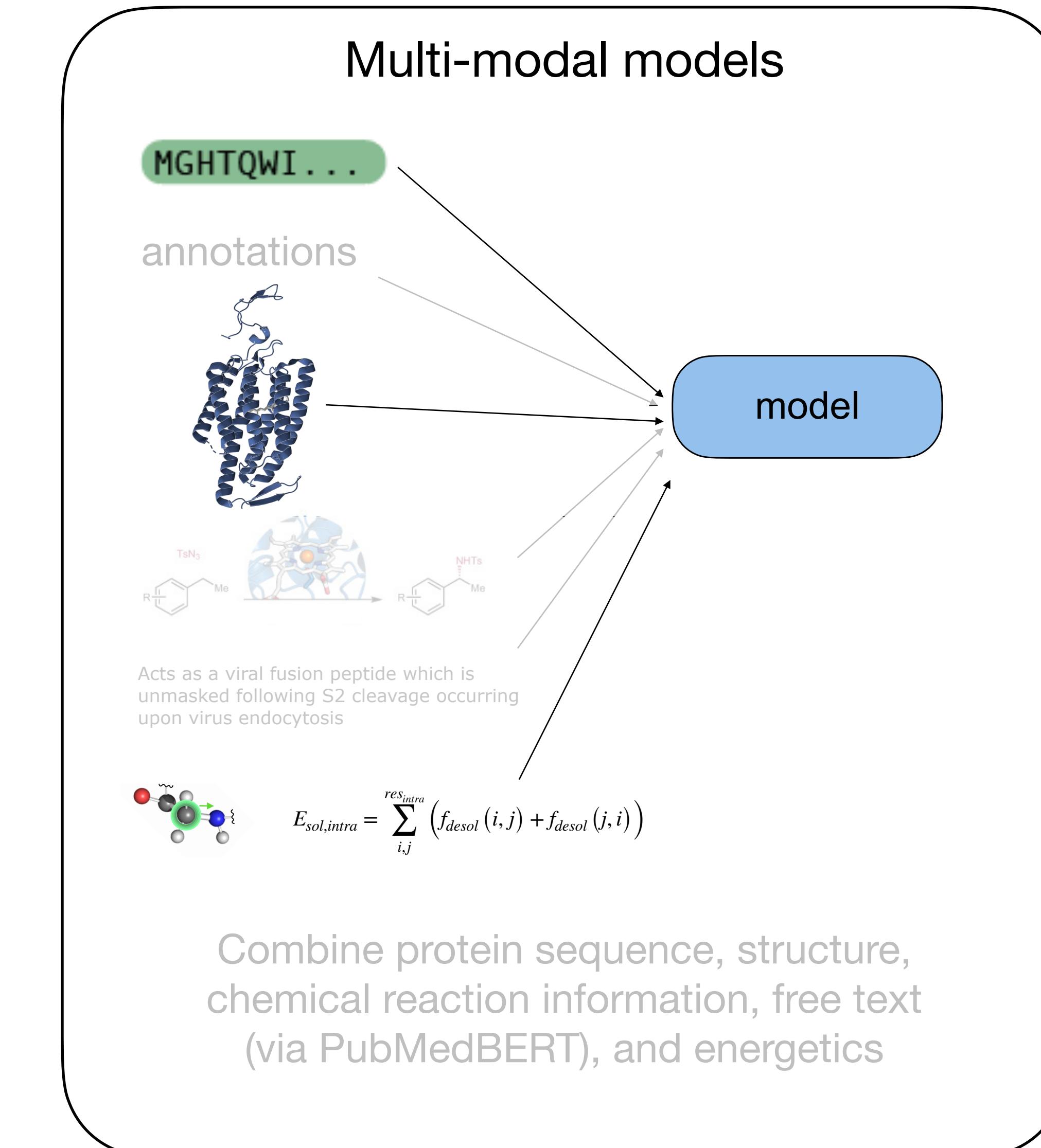
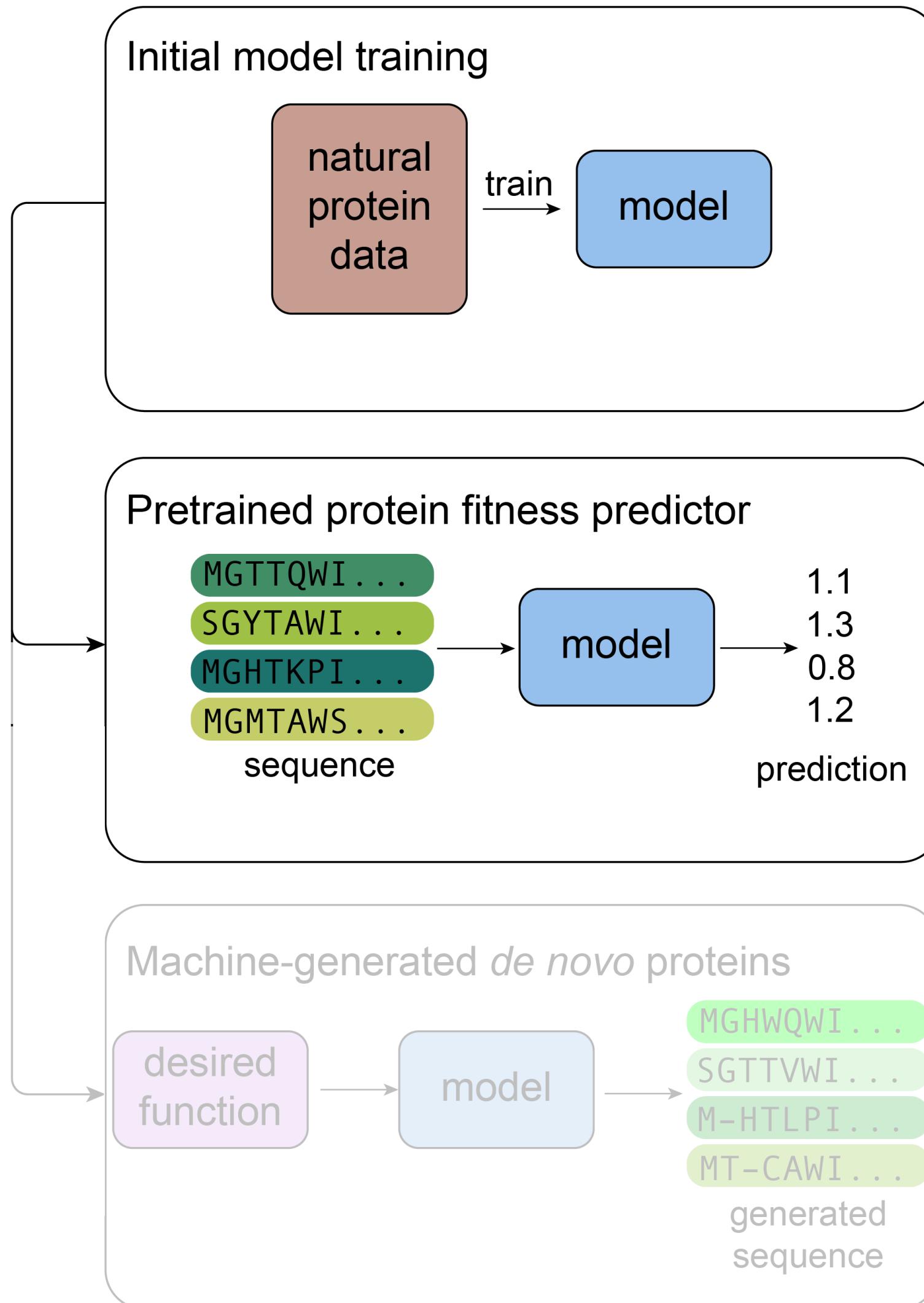
Structure and sequence transfer help a little
Naive MIF-ST predicts all the same values

Try CARP, MIF, and MIF-ST!

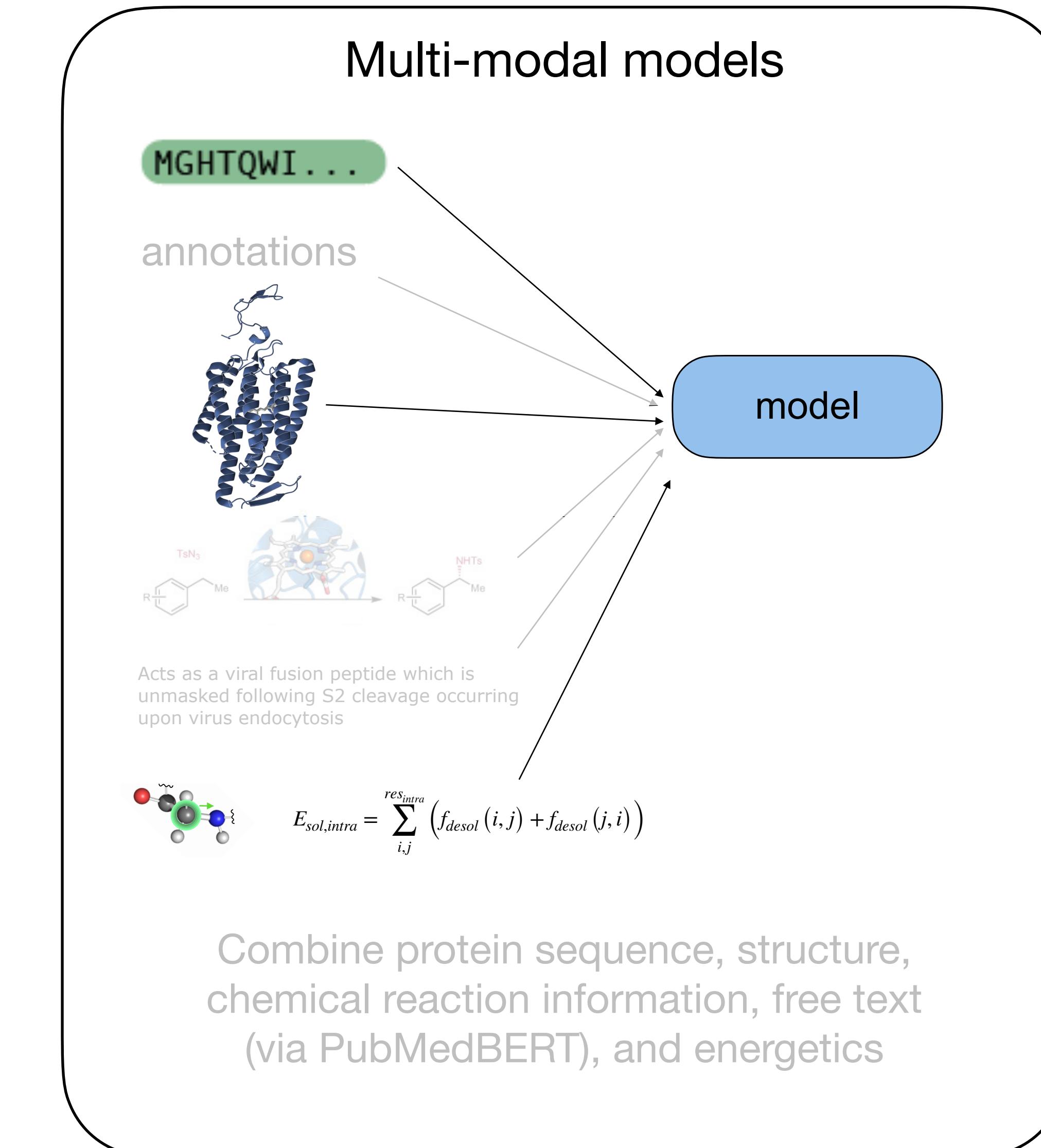
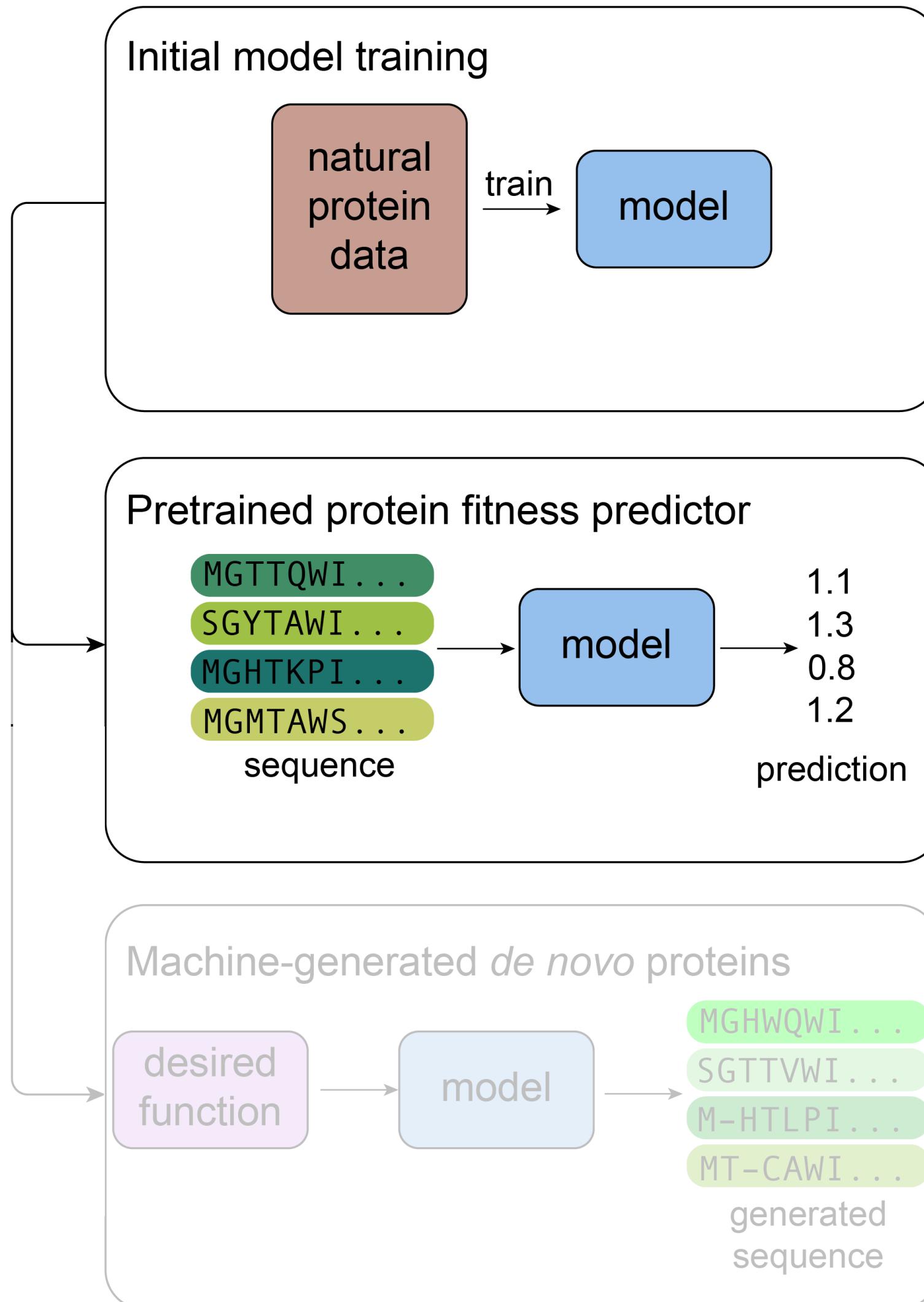


<https://github.com/microsoft/protein-sequence-models>

Use energetic features as inputs



Use energetic features as inputs

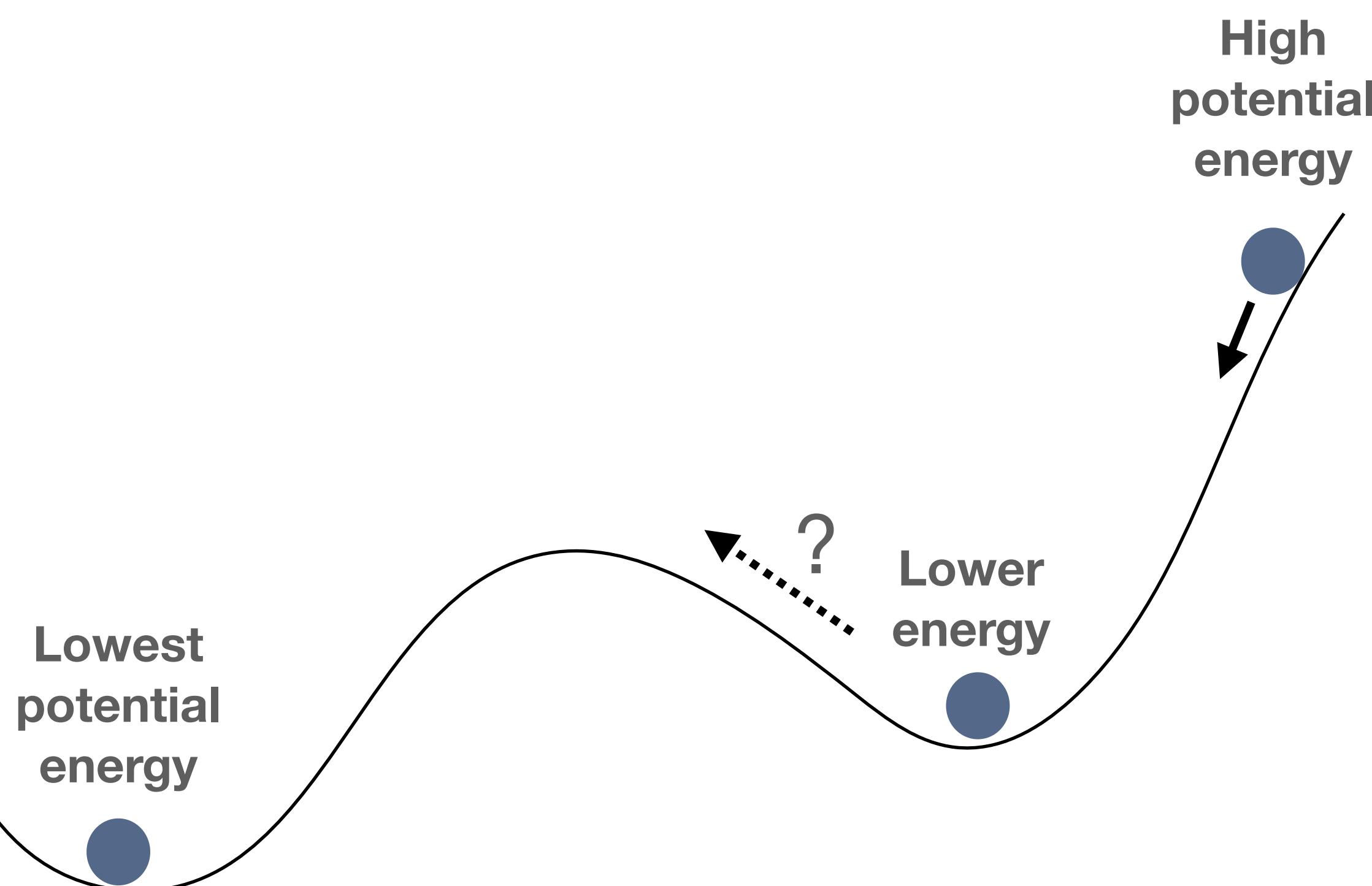


Energy is a physics concept that describes natural processes

Energy is a physics concept that describes natural processes

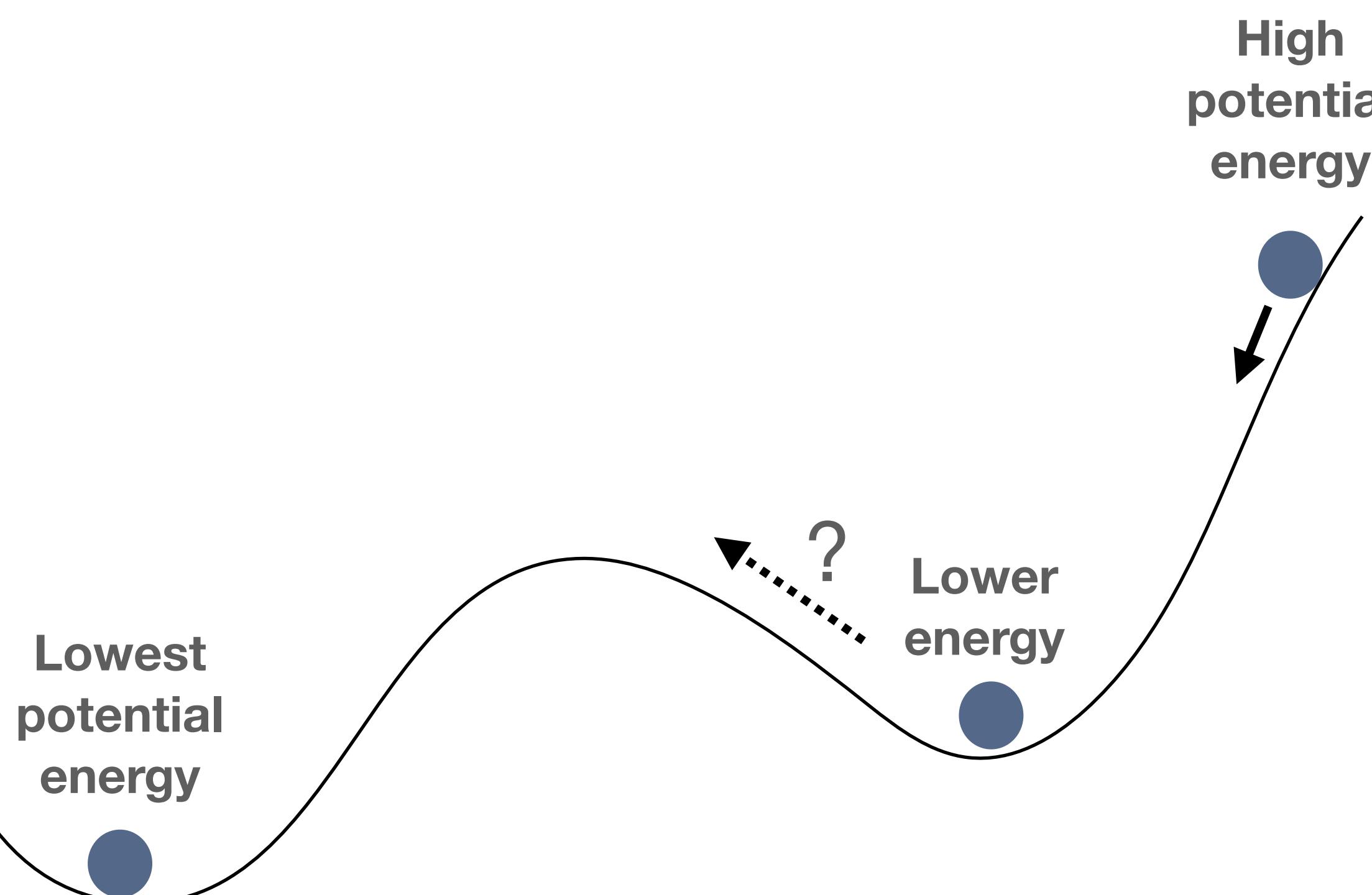
Principle of minimum energy

Energy is a physics concept that describes natural processes



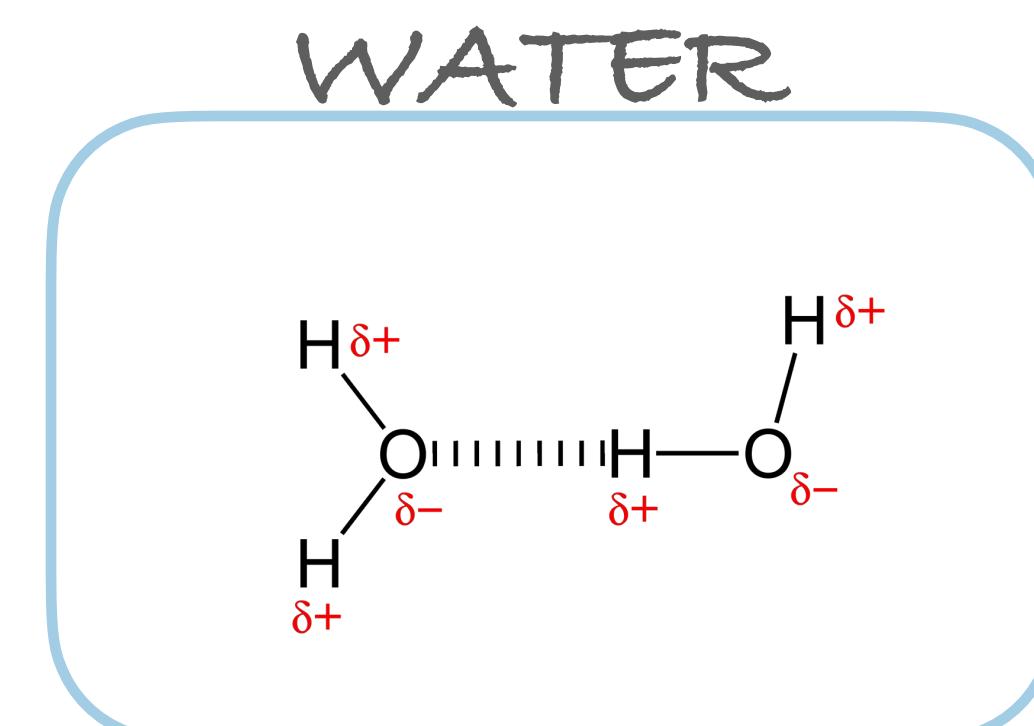
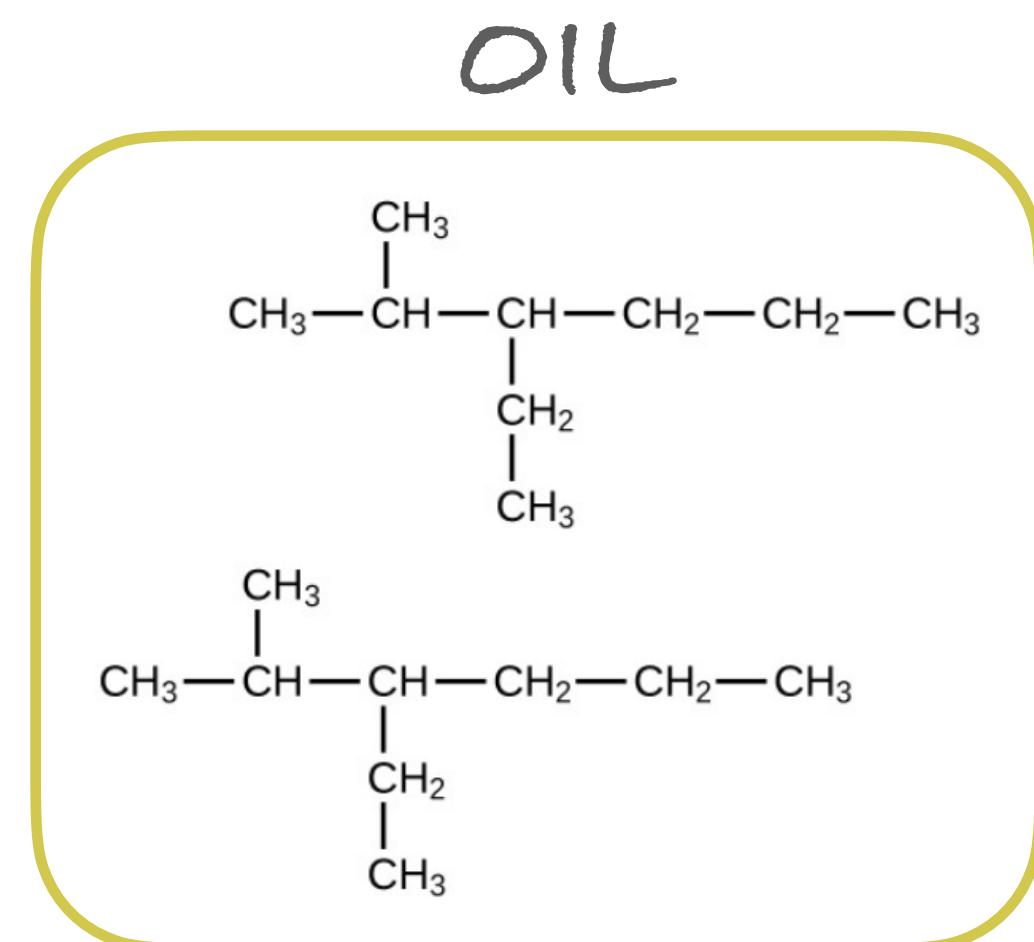
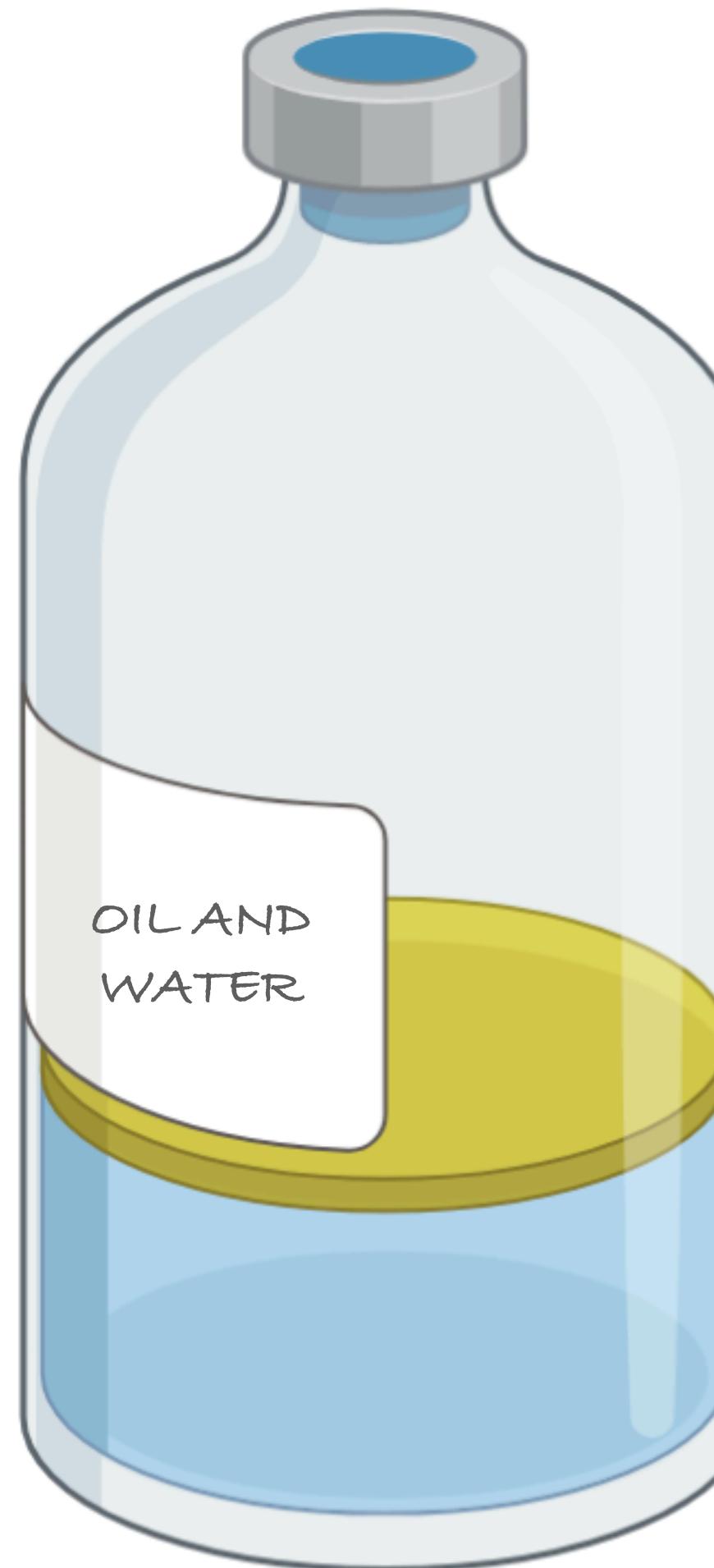
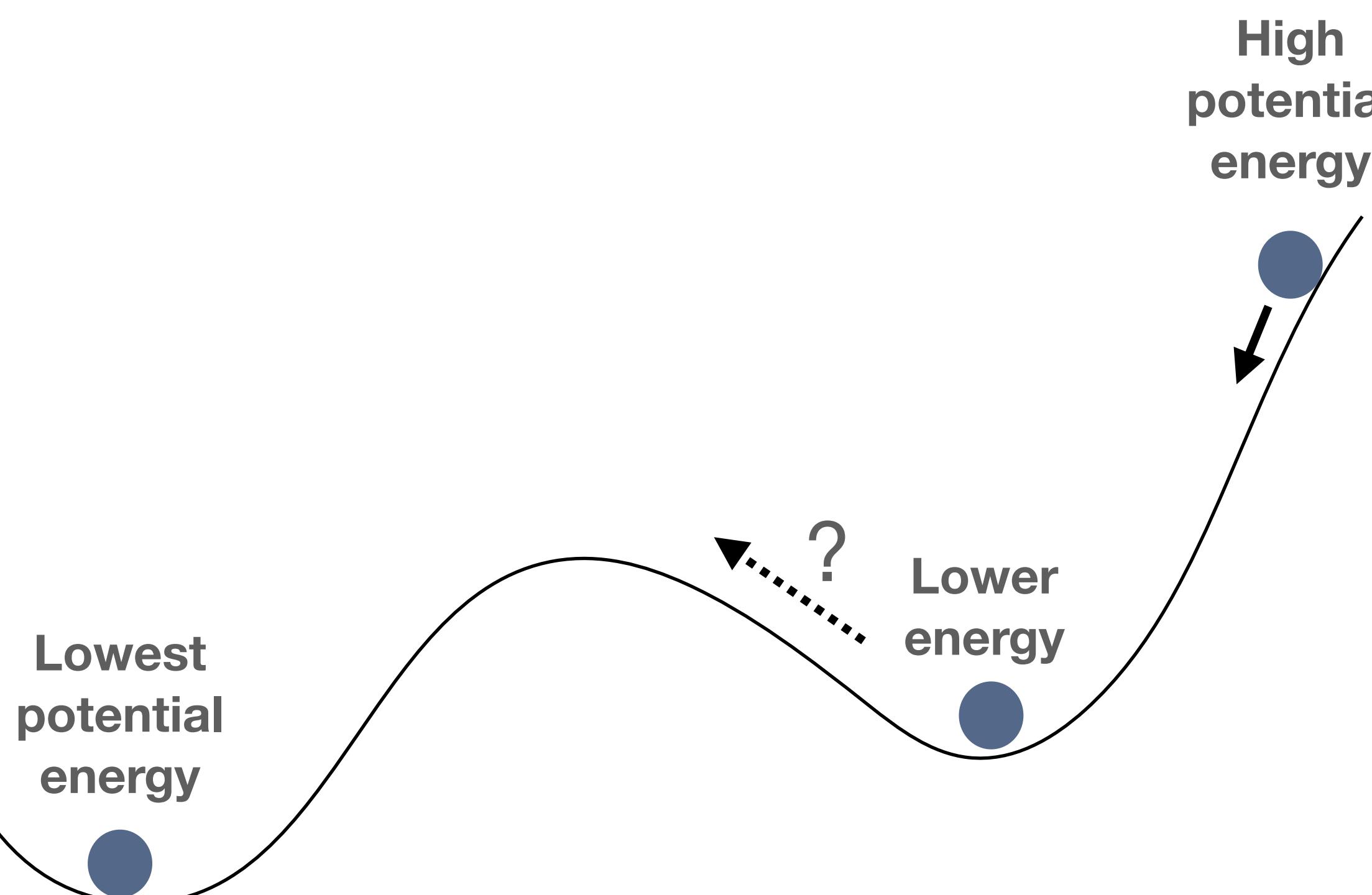
Principle of minimum energy

Energy is a physics concept that describes natural processes



Principle of minimum energy

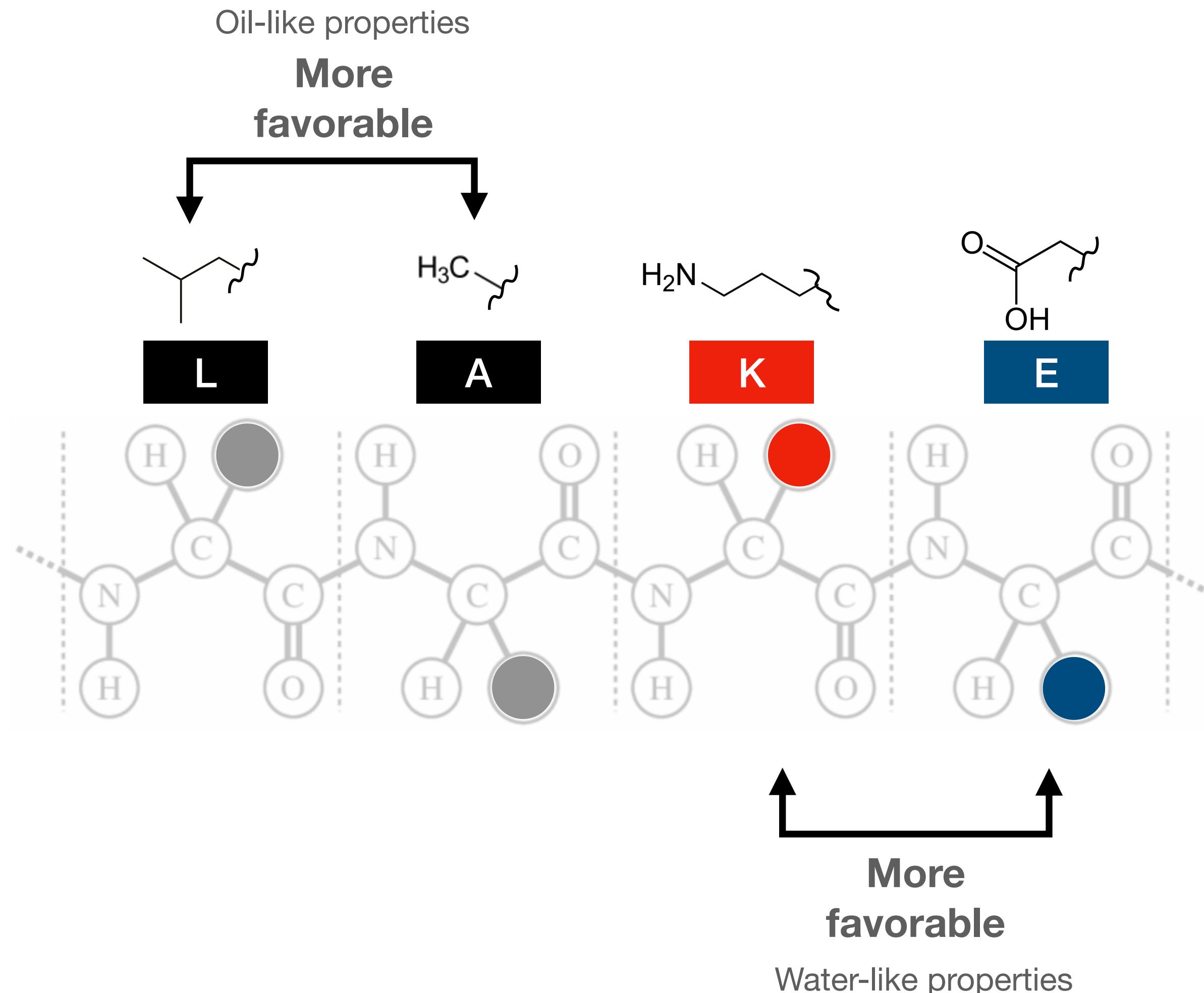
Energy is a physics concept that describes natural processes



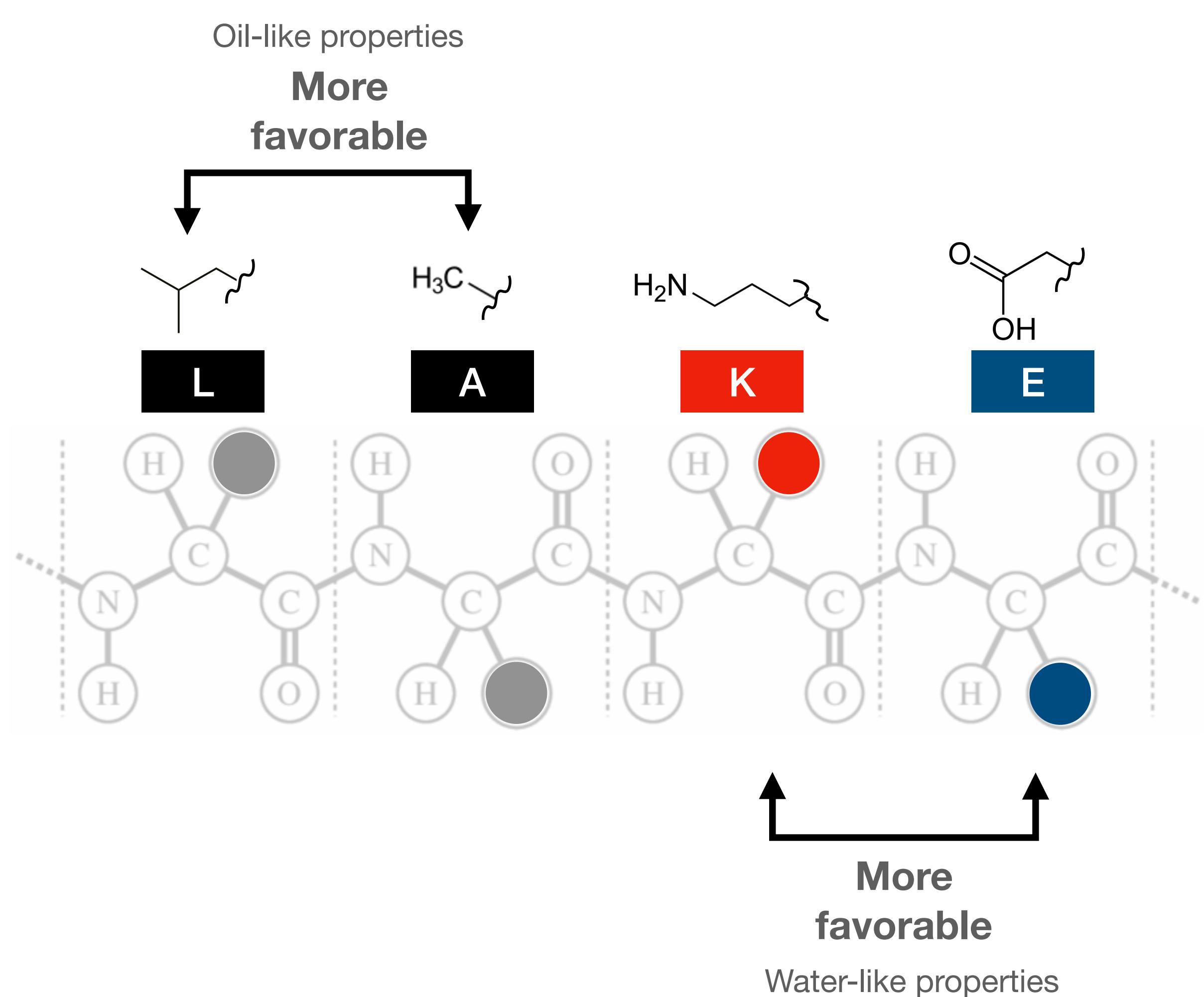
Principle of minimum energy

Protein properties are bound by physical laws

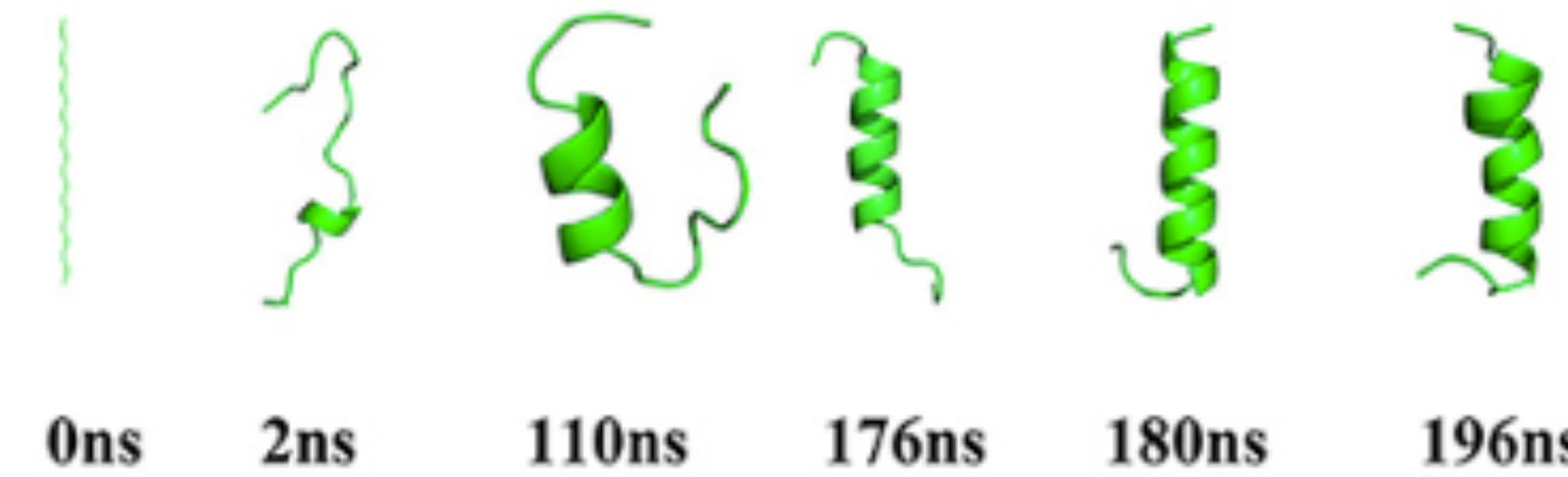
Protein properties are bound by physical laws



Protein properties are bound by physical laws

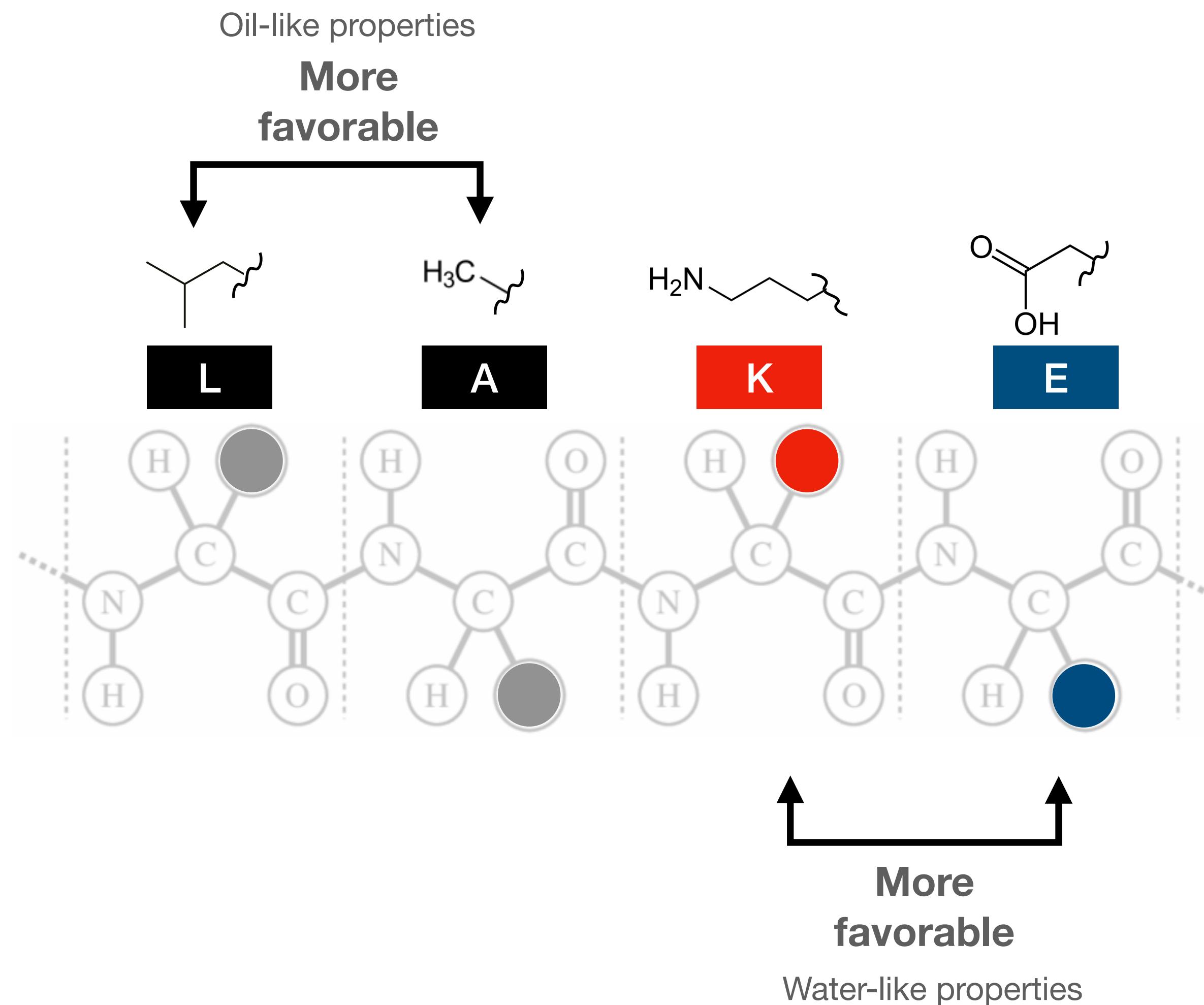


All atom physics-based simulation

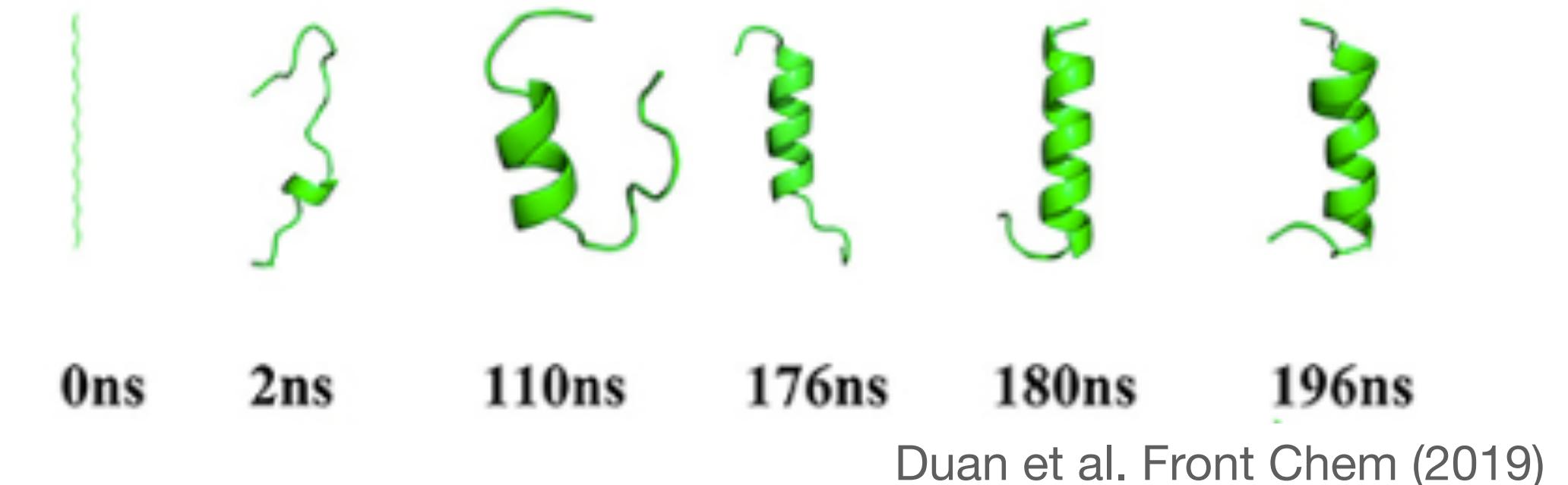


Duan et al. Front Chem (2019)

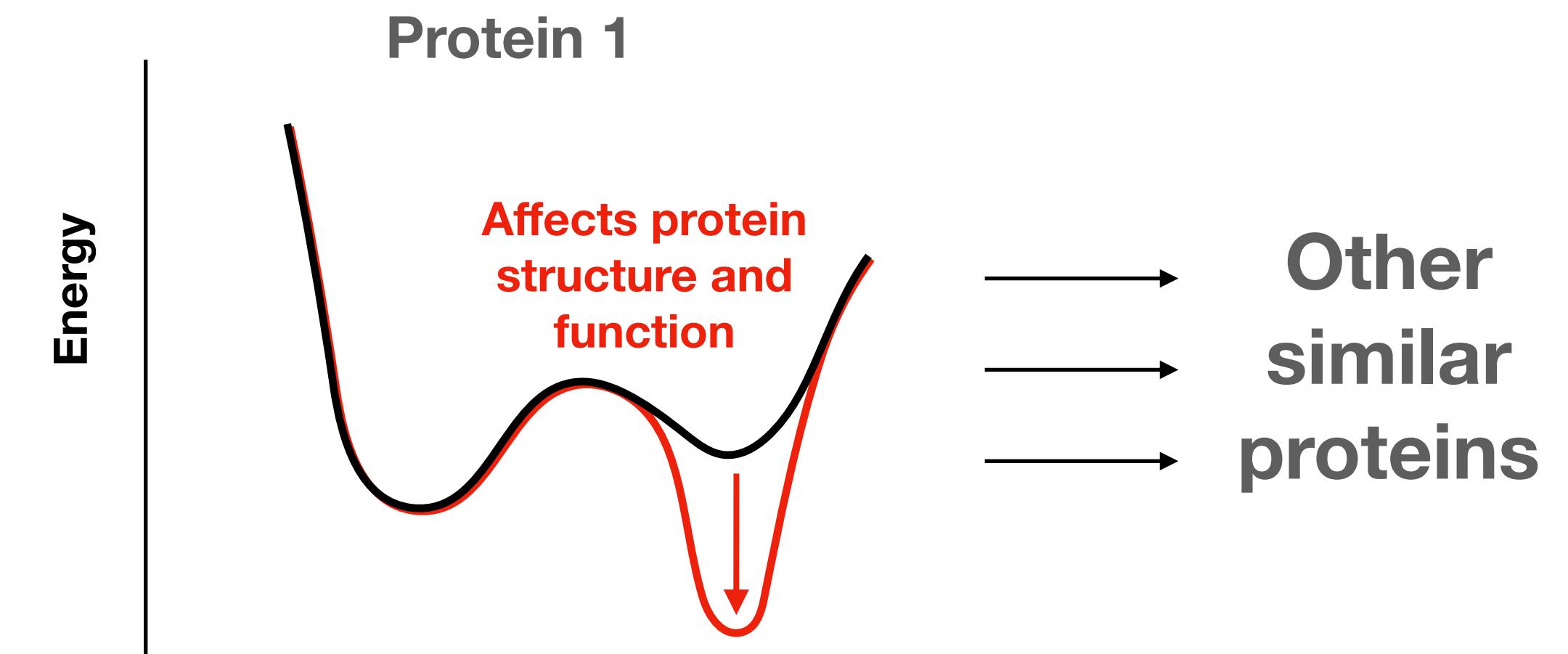
Protein properties are bound by physical laws



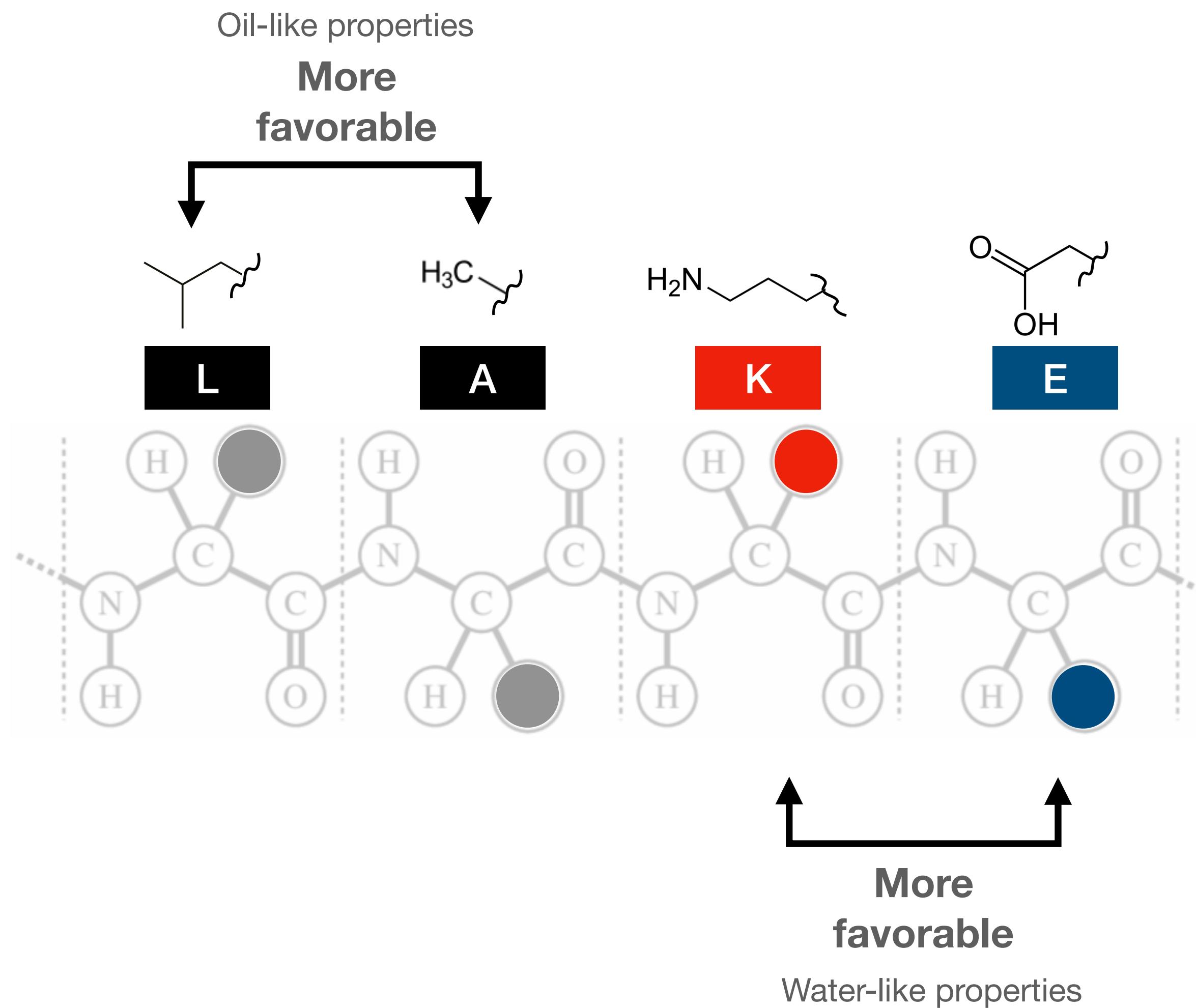
All atom physics-based simulation



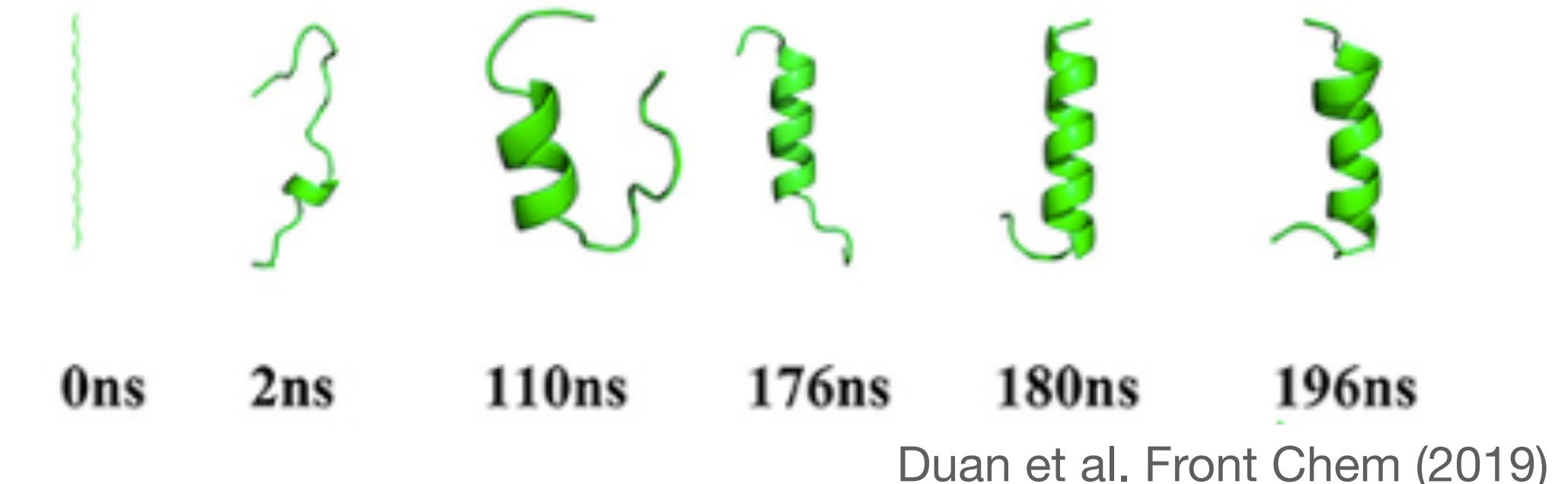
Protein functional mechanism



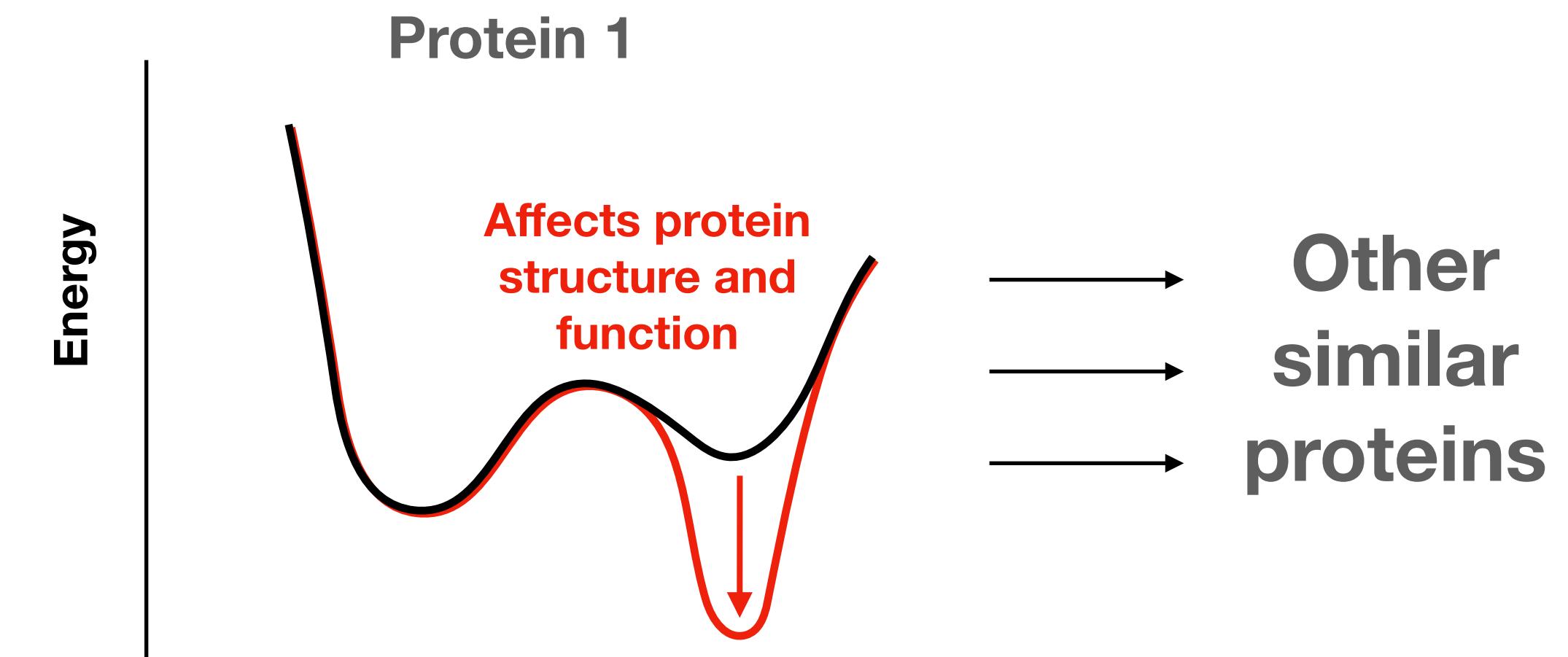
Protein properties are bound by physical laws



All atom physics-based simulation

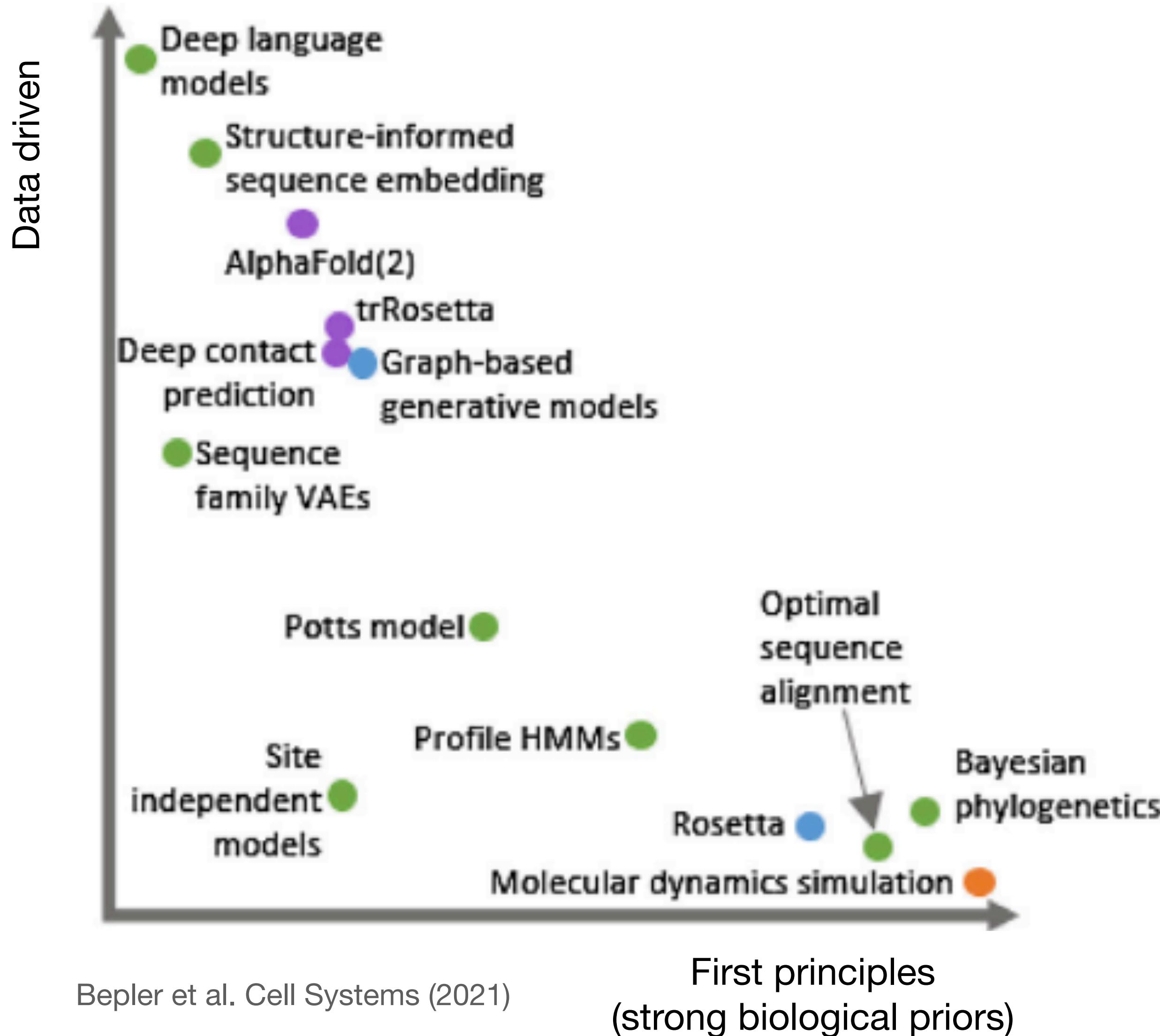


Protein functional mechanism

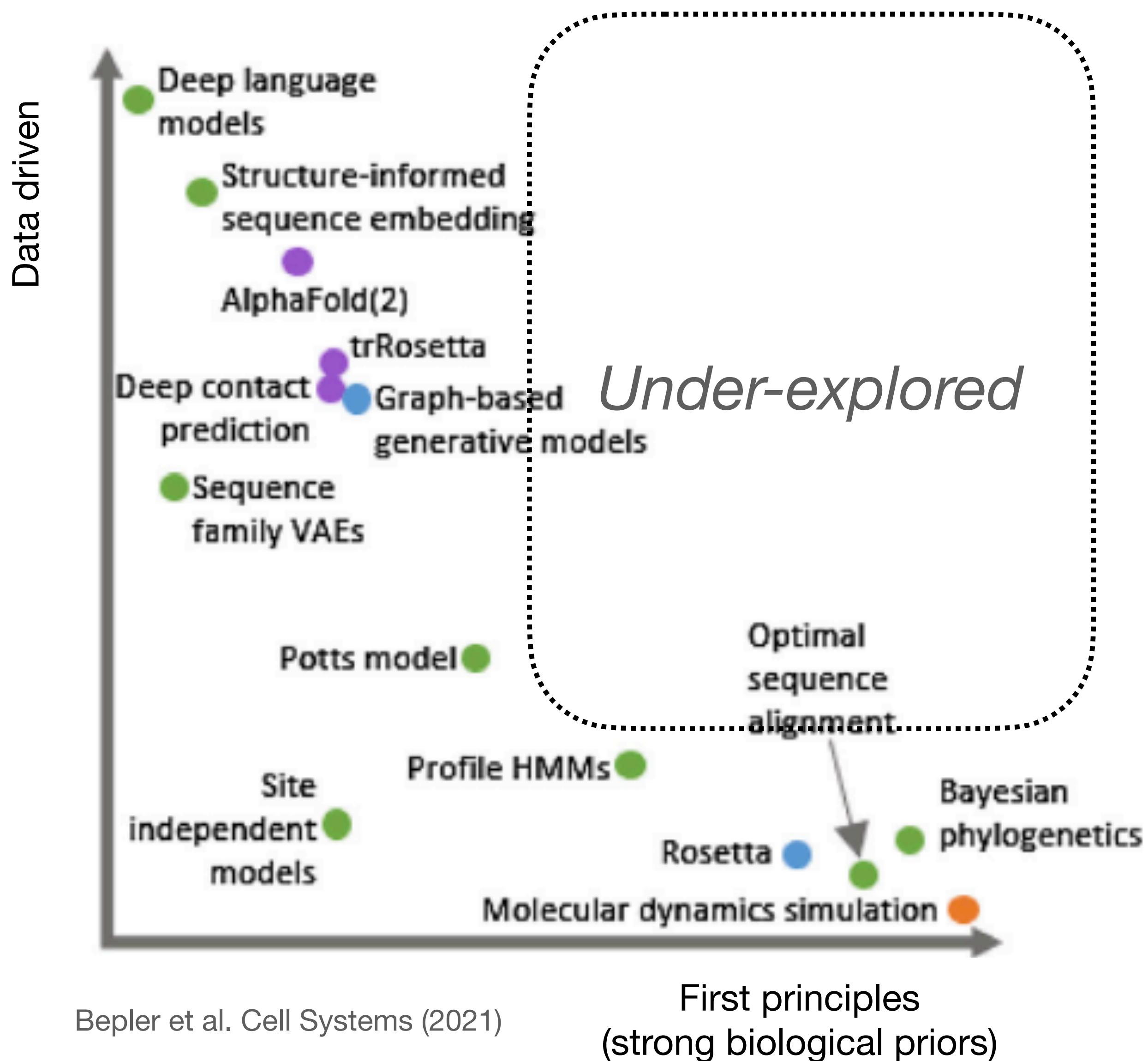


Predictive power limited by computational efficiency and accuracy

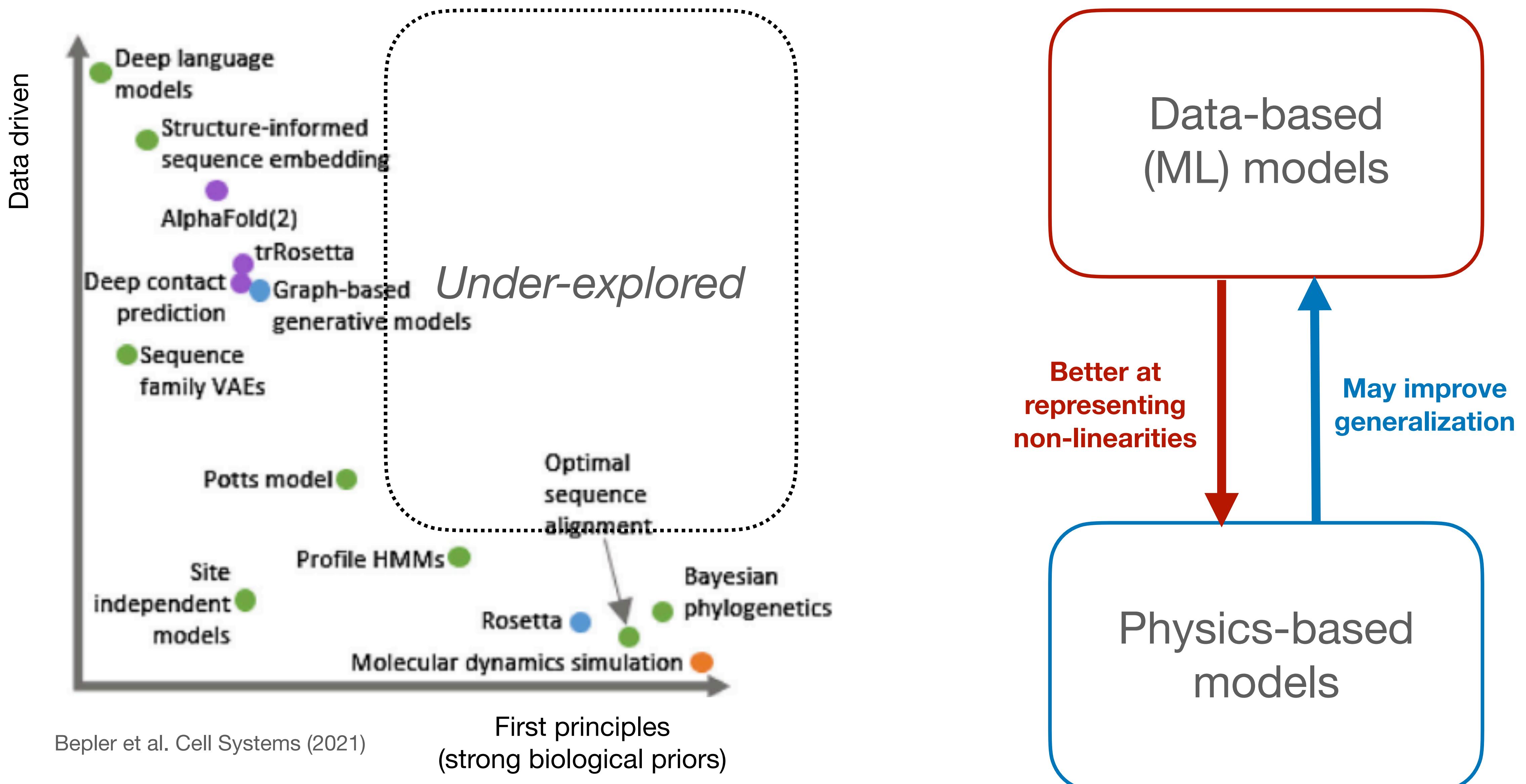
Physics and ML may be complementary



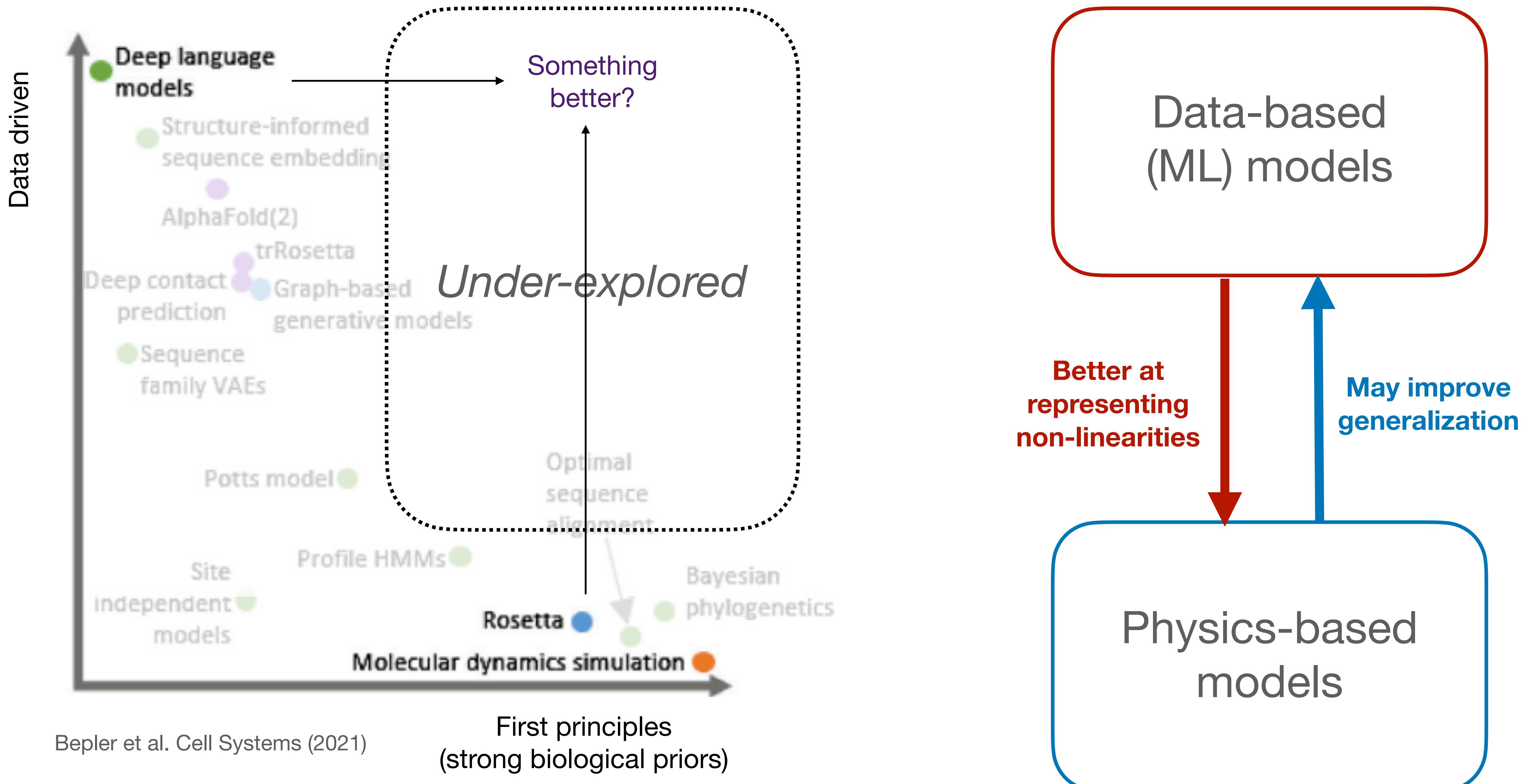
Physics and ML may be complementary



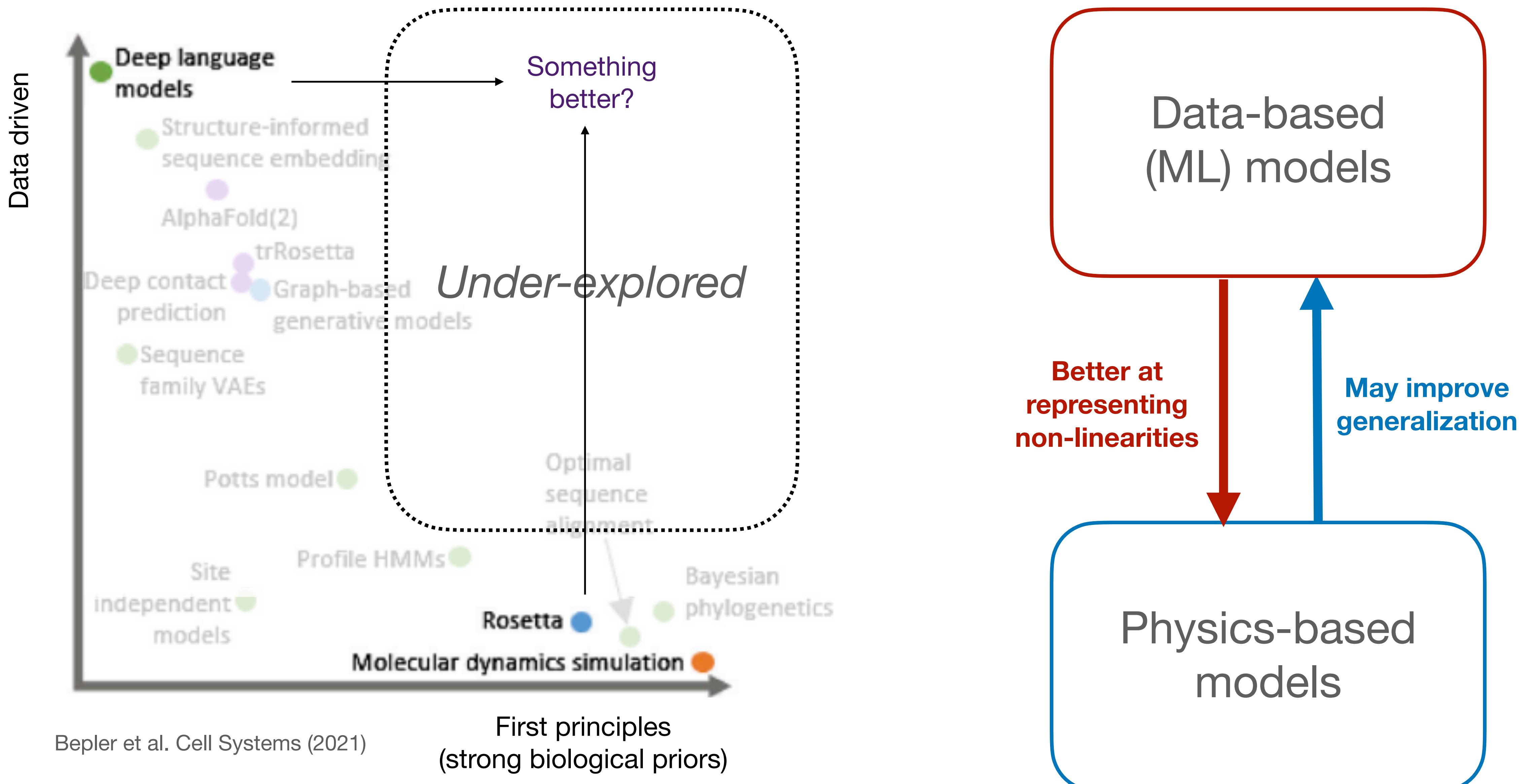
Physics and ML may be complementary



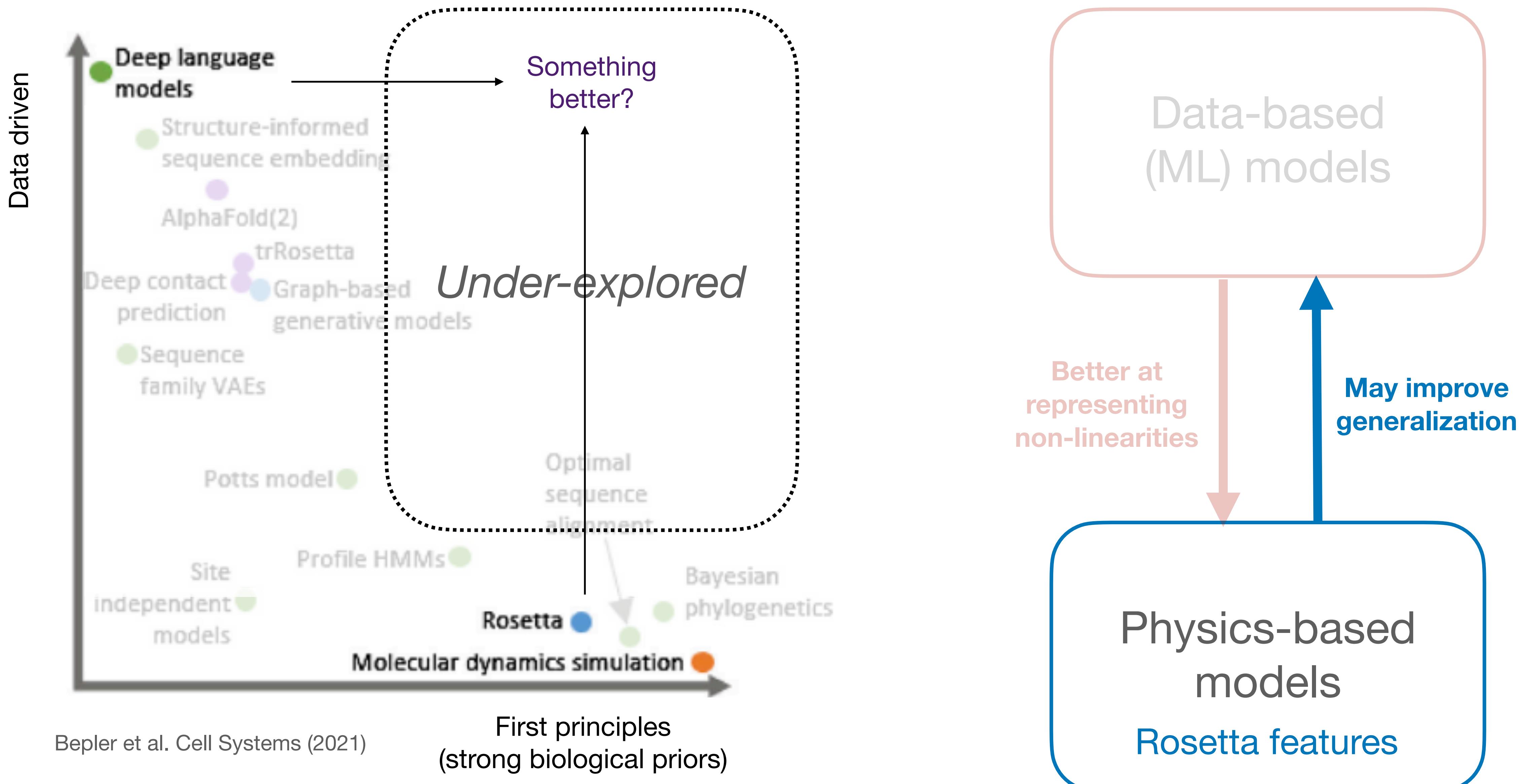
Physics and ML may be complementary



Physics-based features may improve ML models

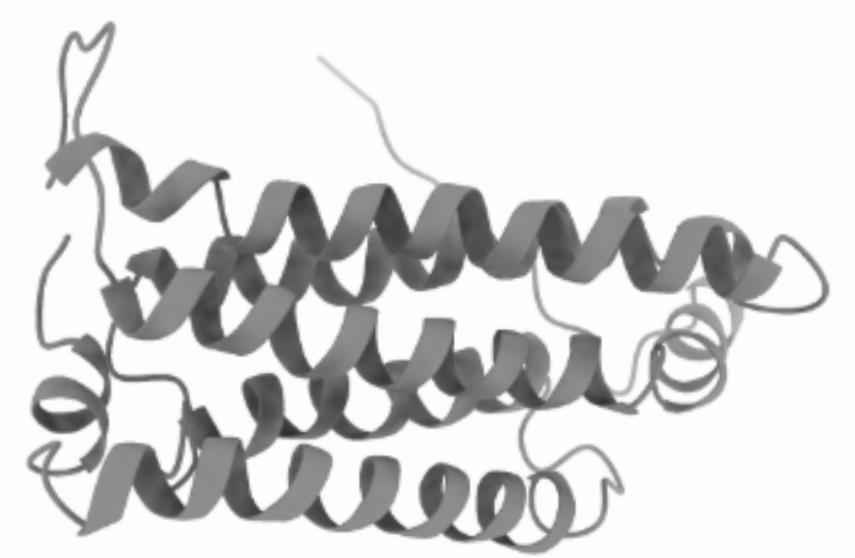


Physics-based features may improve ML models



Use precomputed physics-based energies as input features

Use precomputed physics-based energies as input features

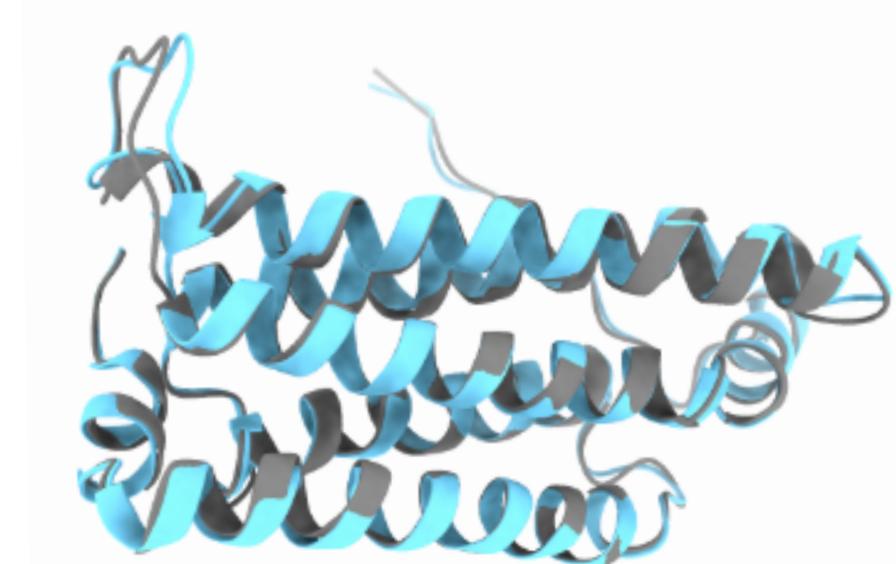


Reference protein

Use precomputed physics-based energies as input features

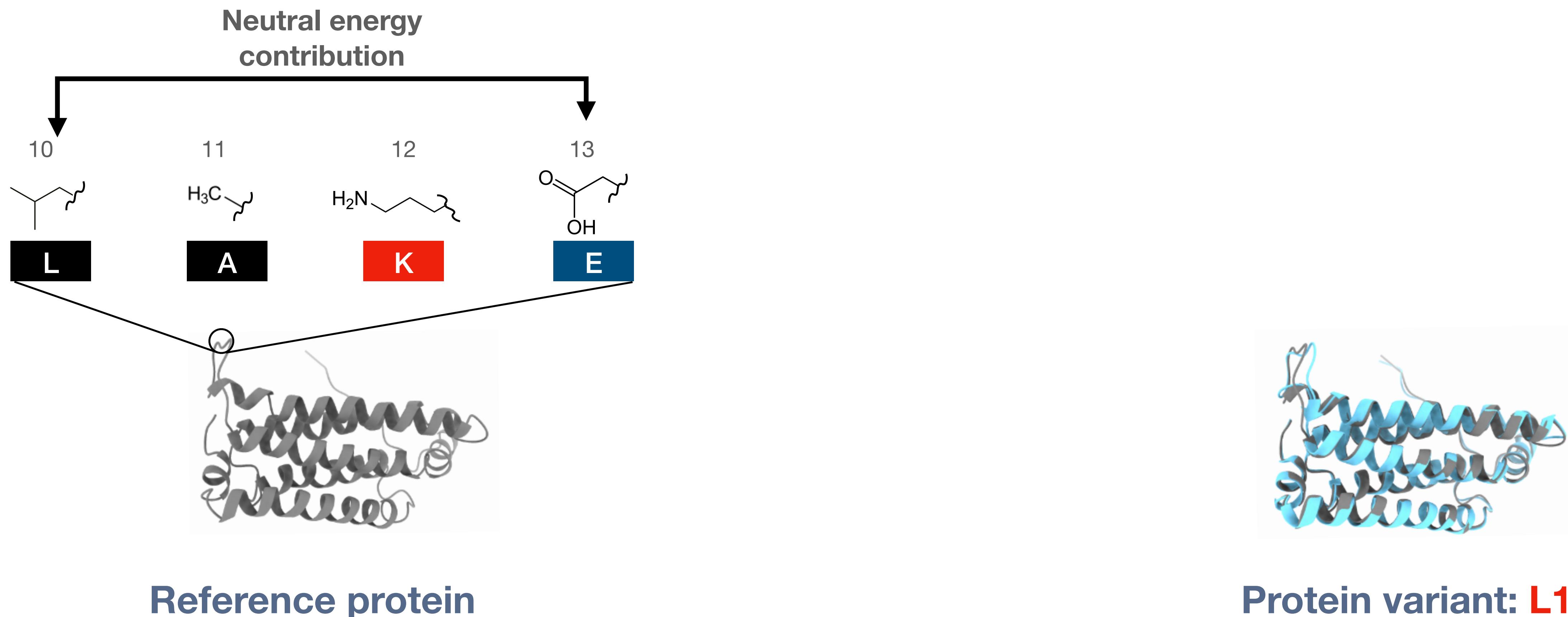


Reference protein

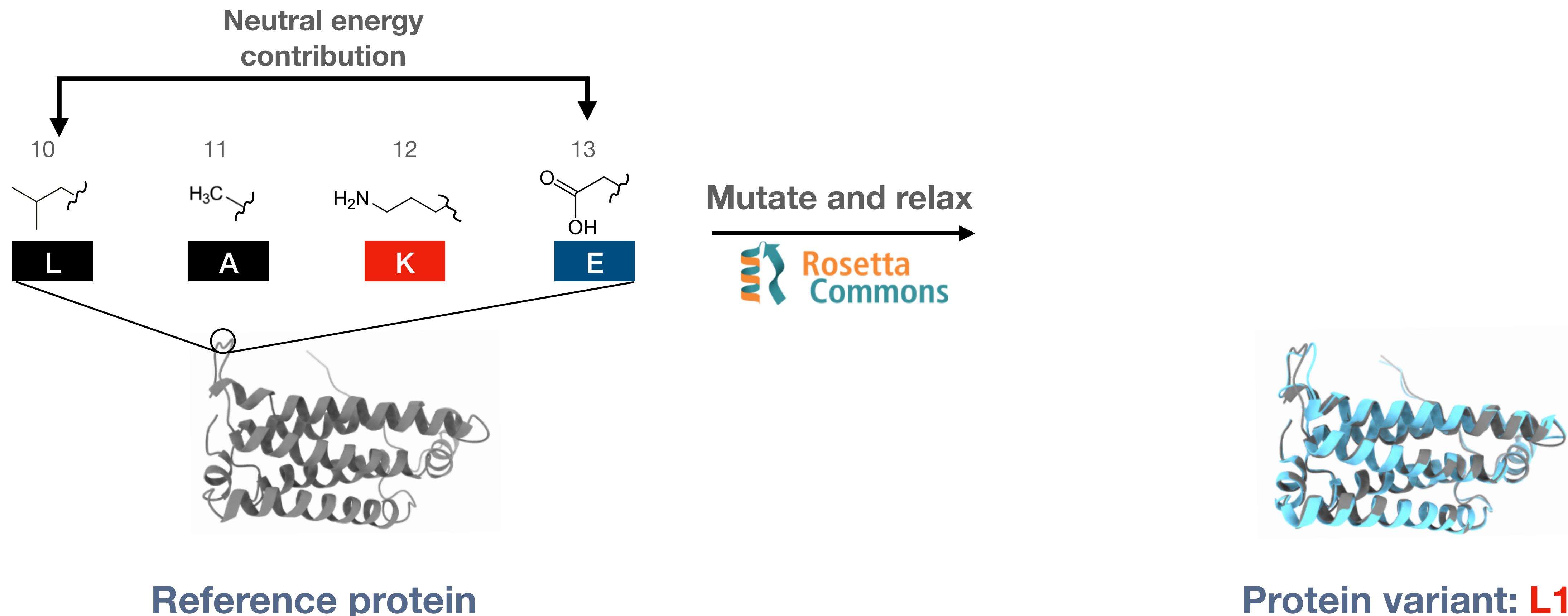


Protein variant: **L10E**

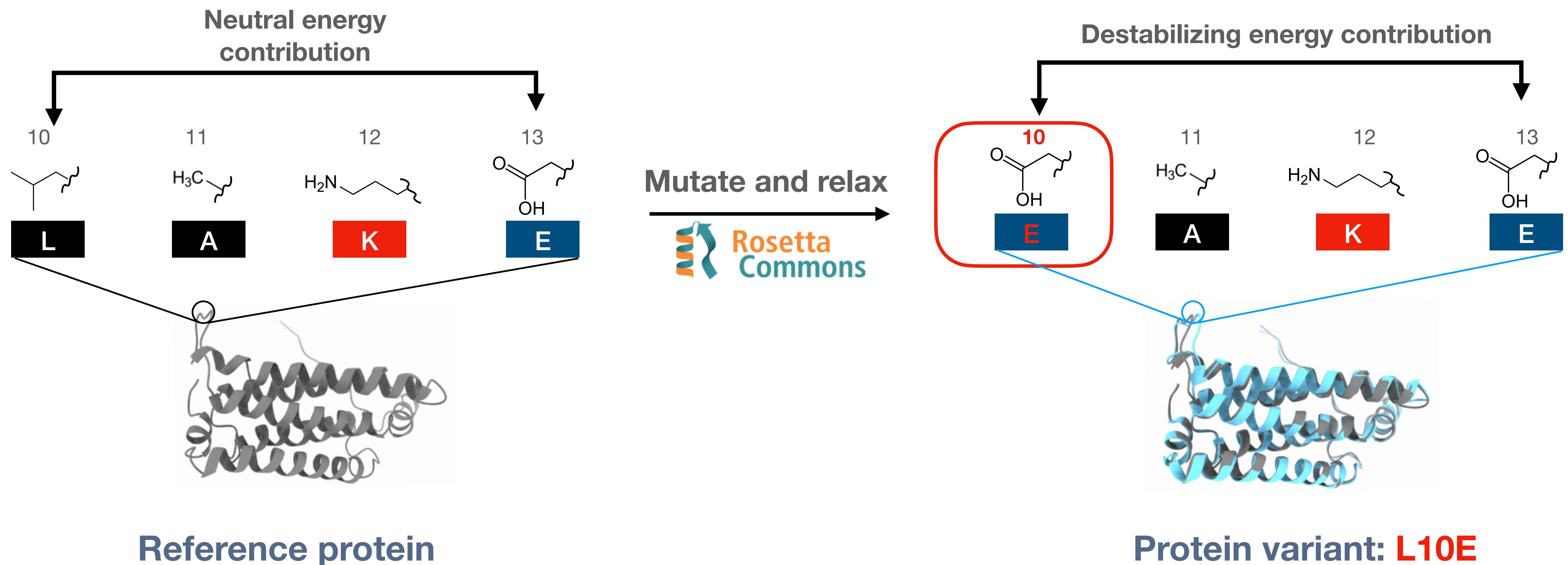
Use precomputed physics-based energies as input features



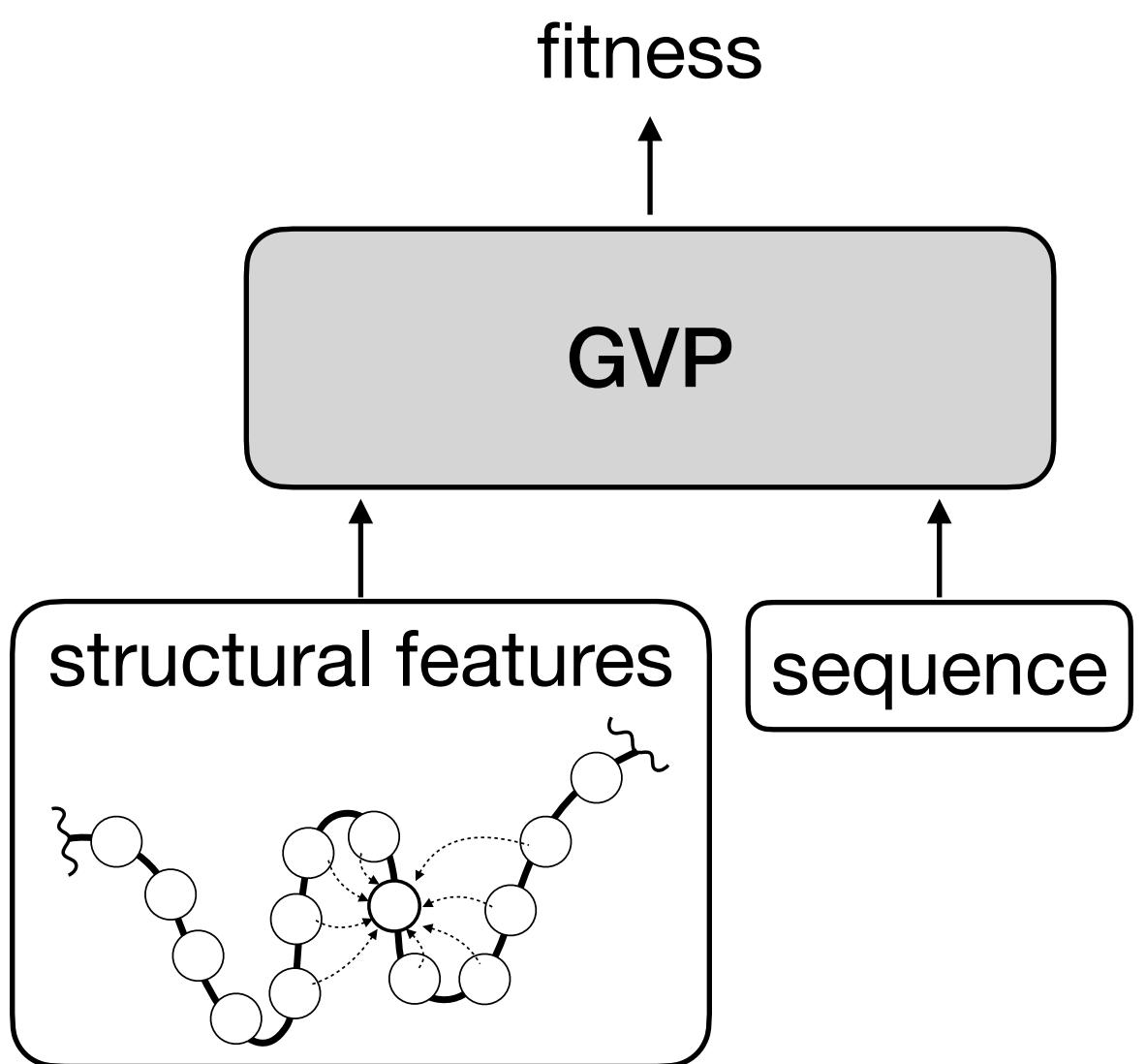
Use precomputed physics-based energies as input features



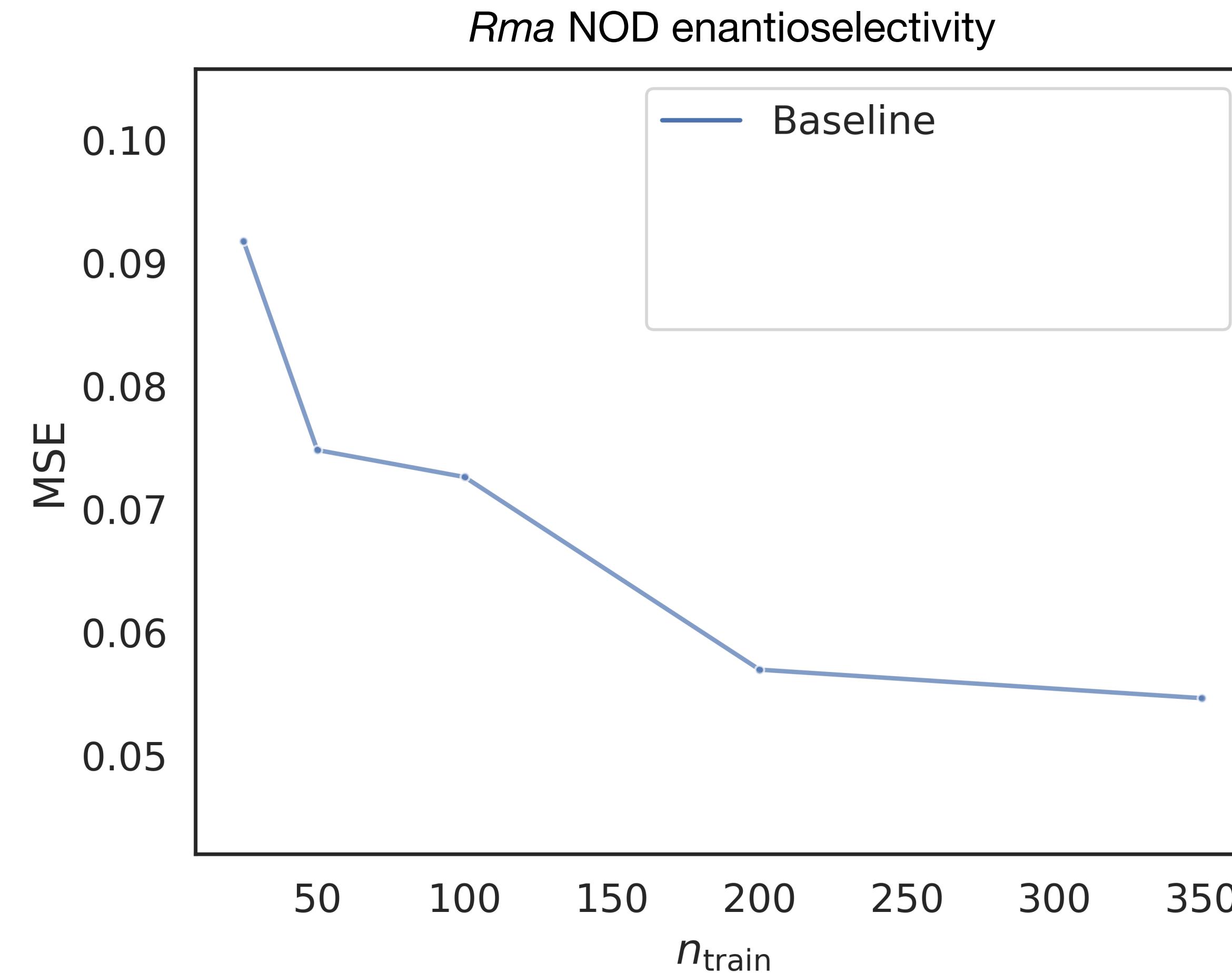
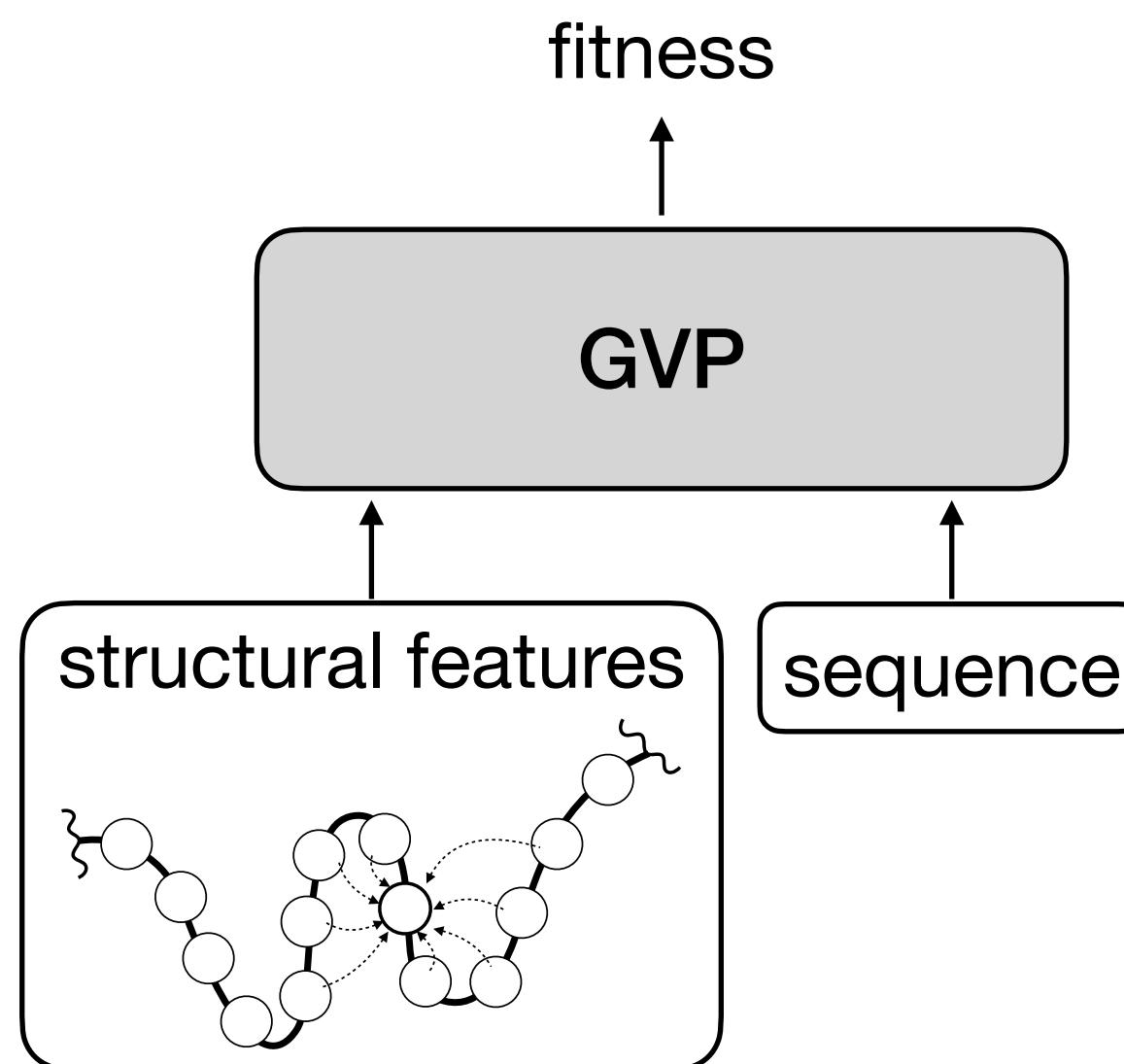
Use precomputed physics-based energies as input features



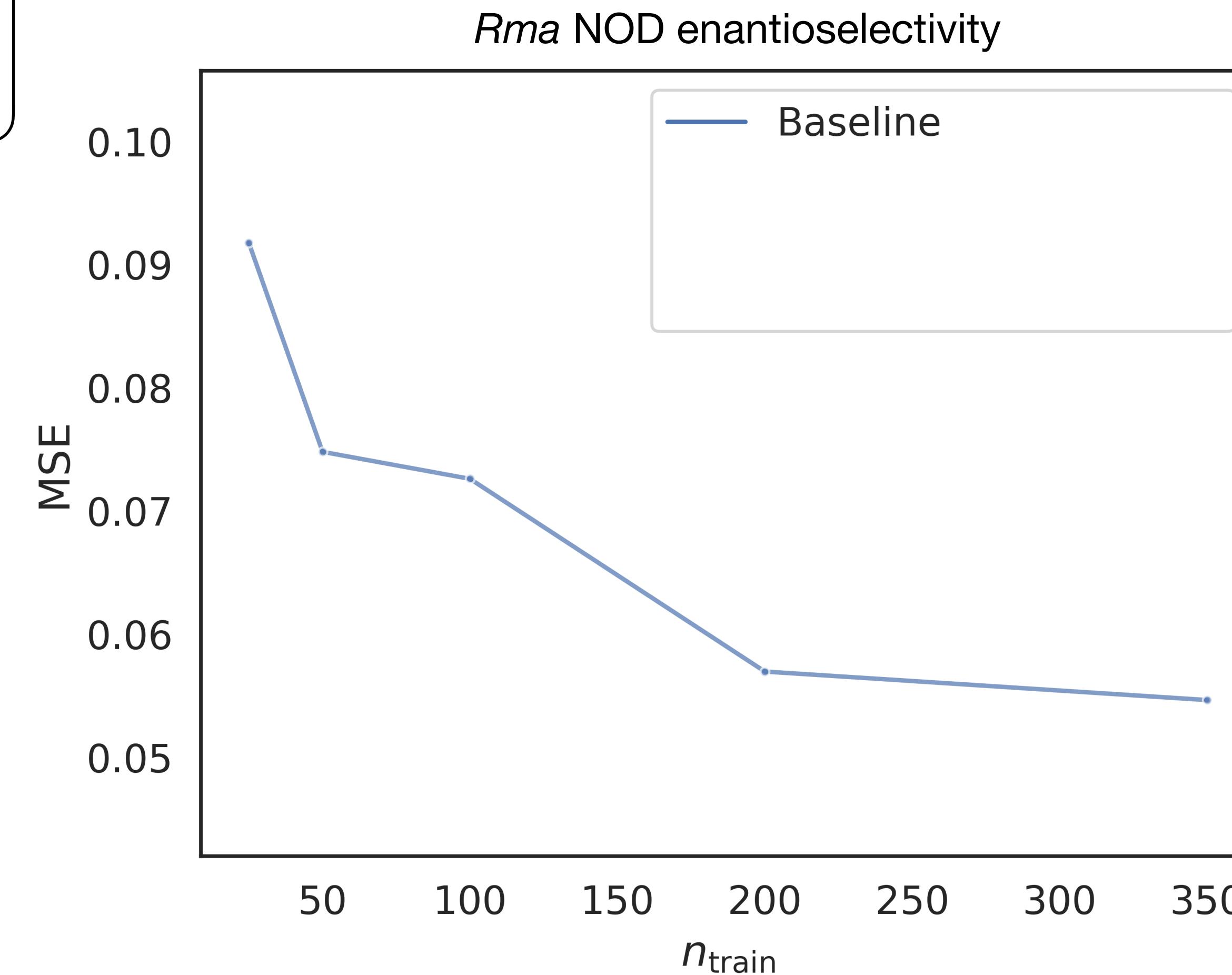
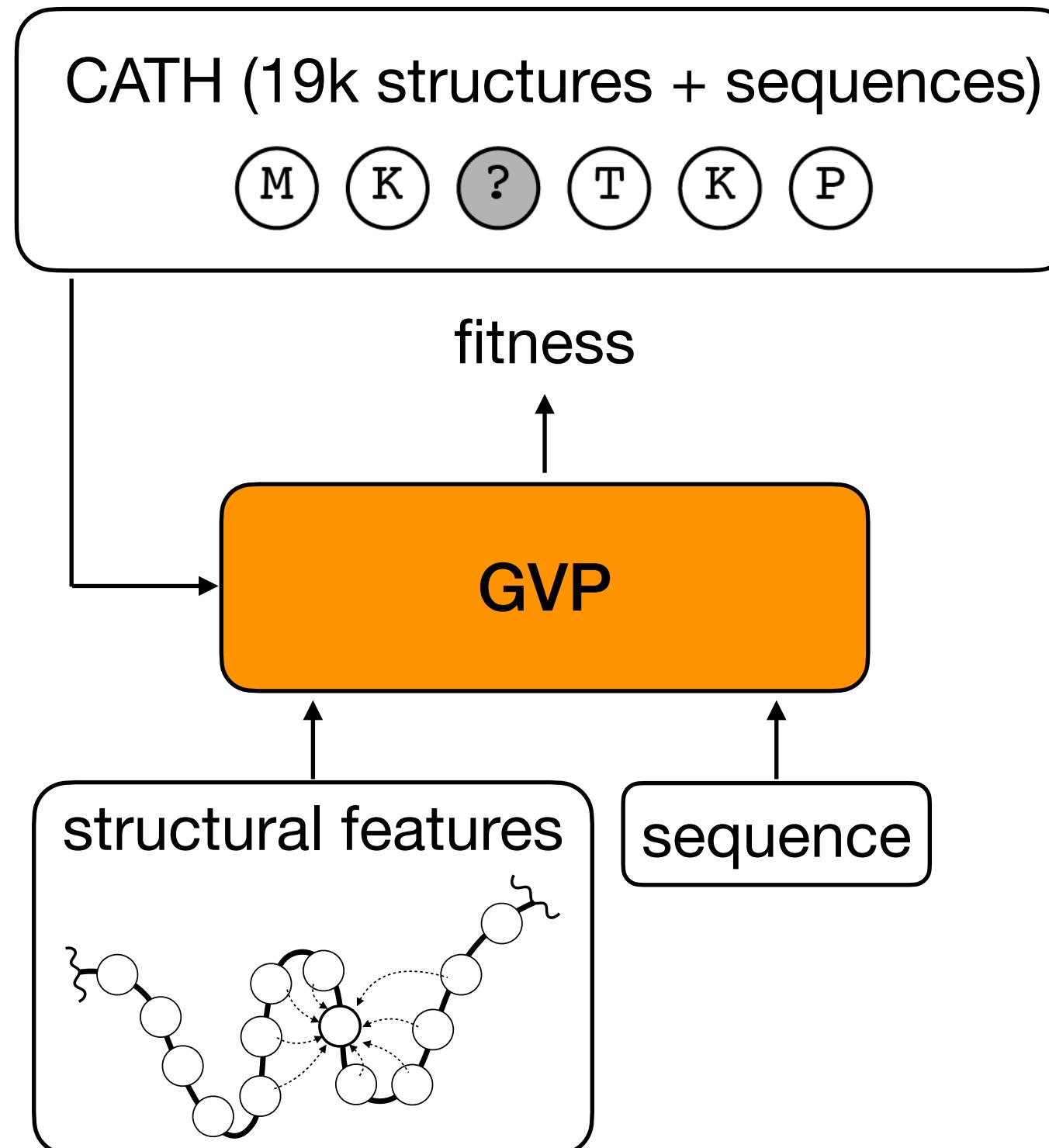
Energetics features improve performance in low-data regimes



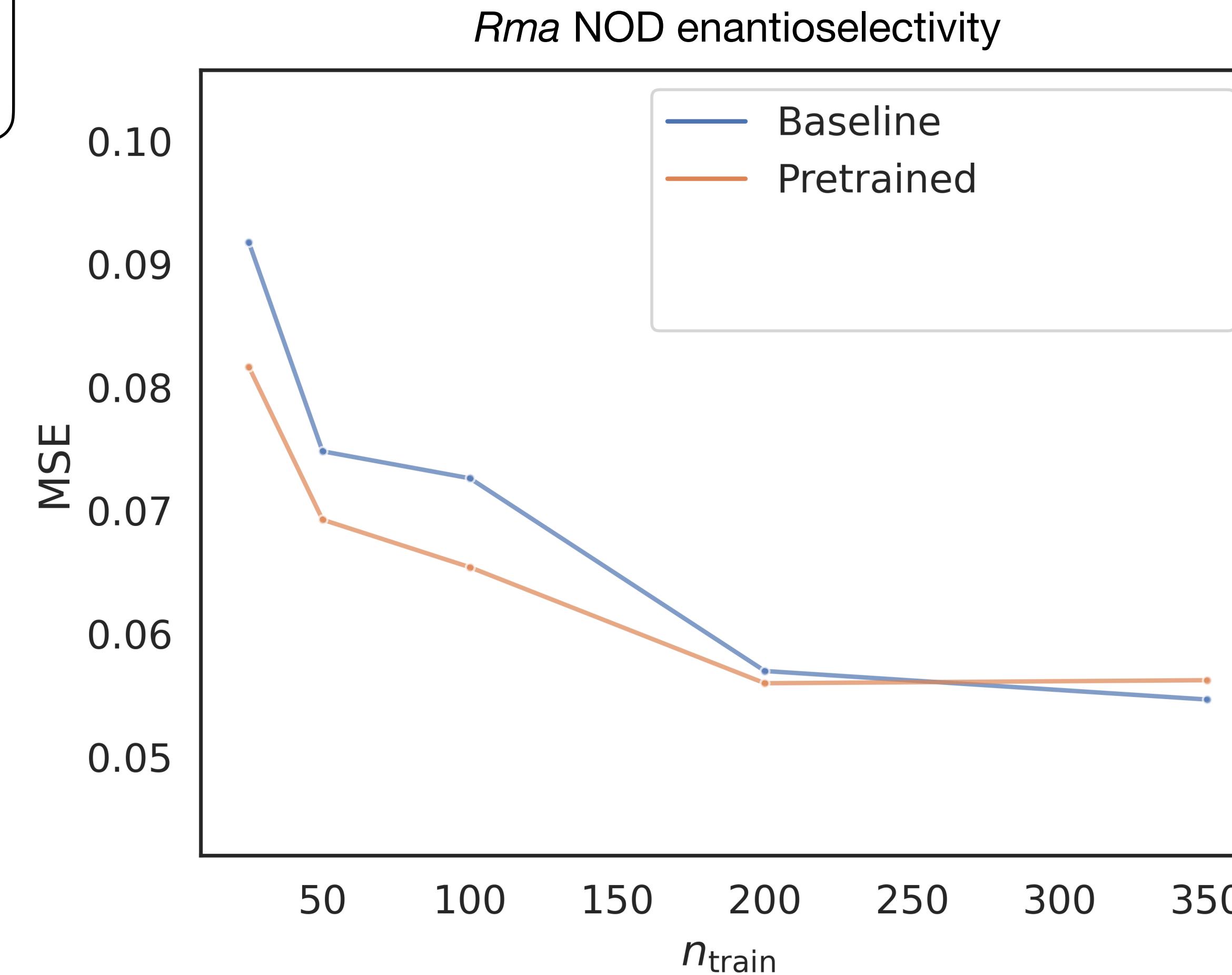
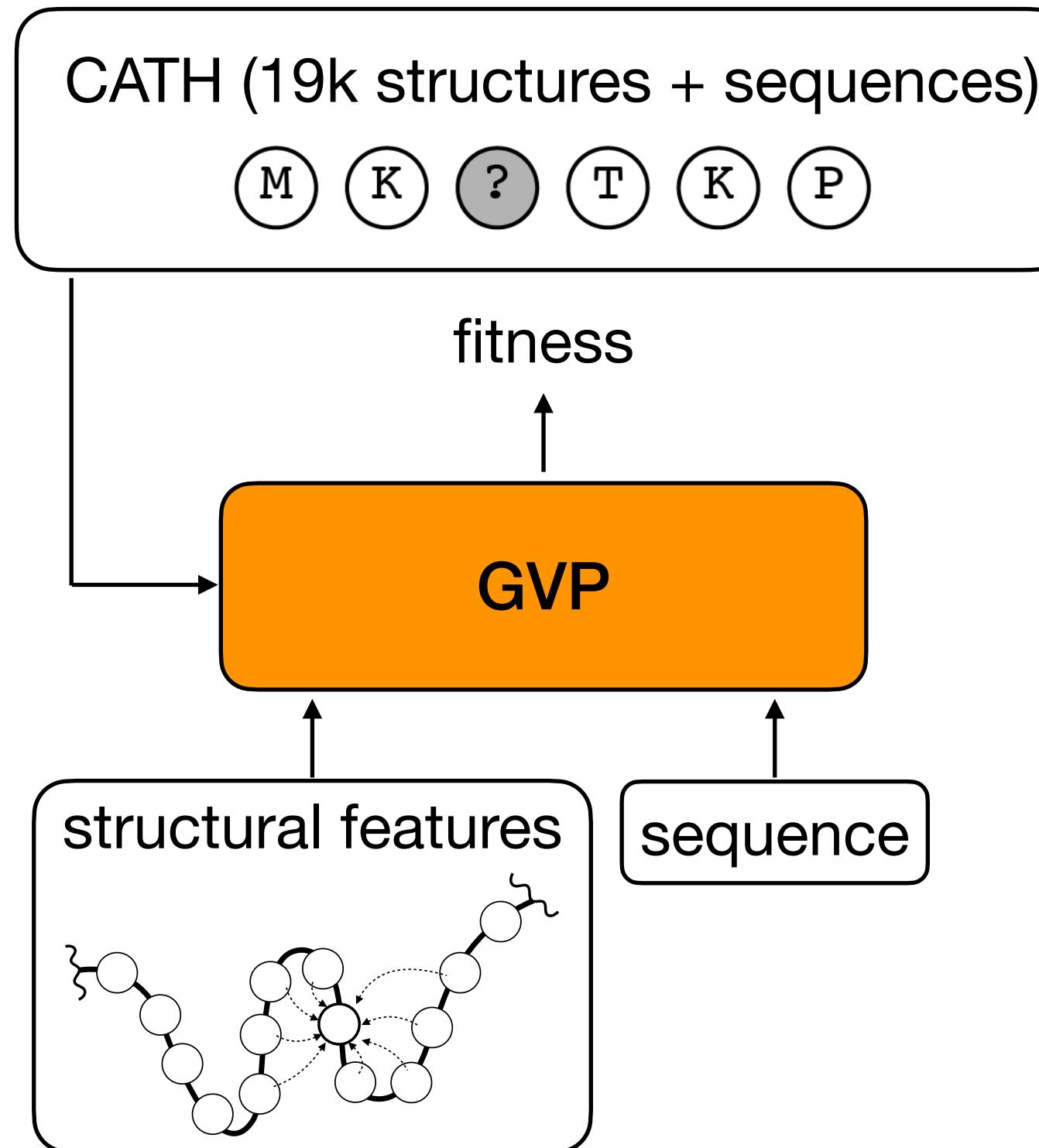
Energetics features improve performance in low-data regimes



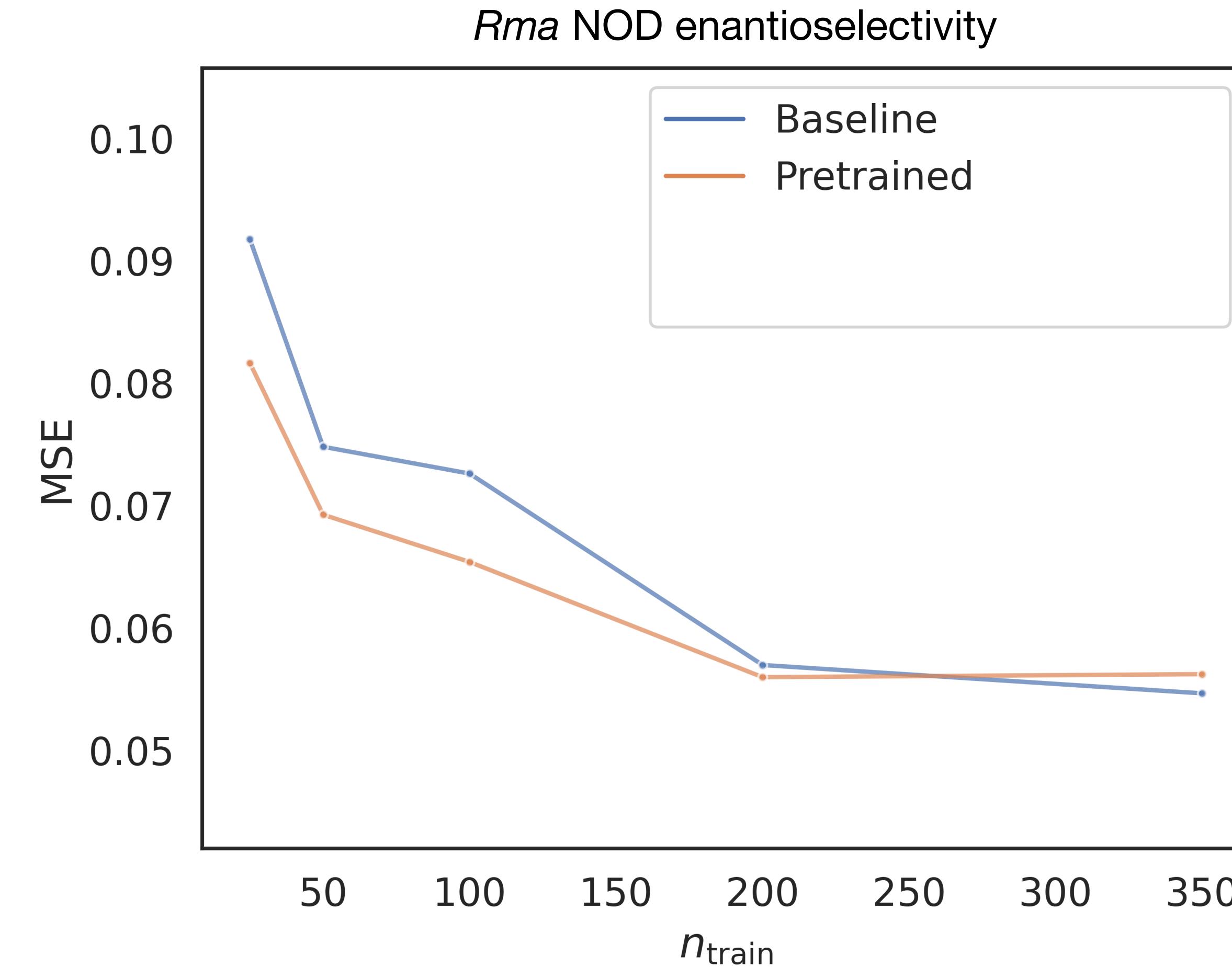
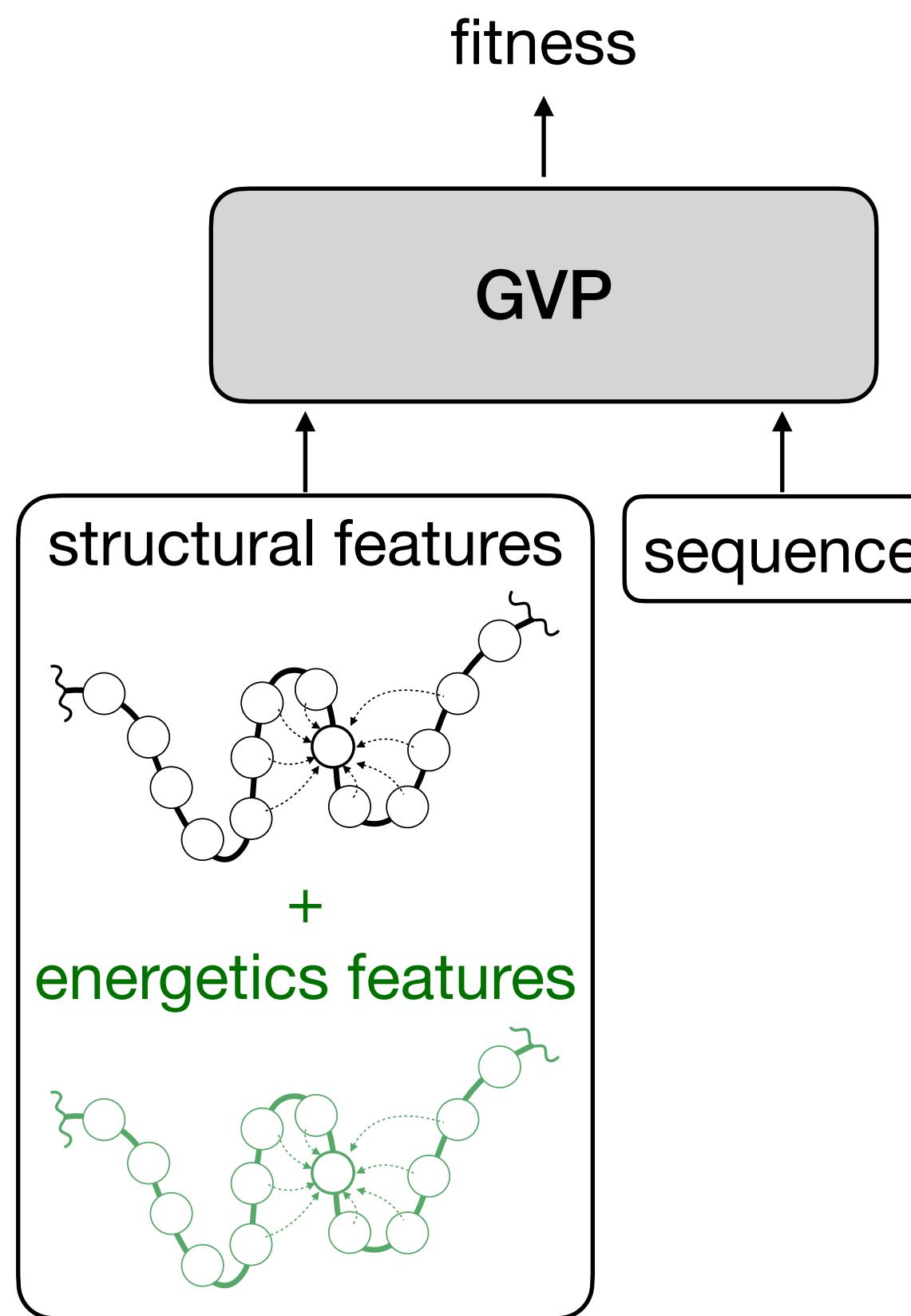
Energetics features improve performance in low-data regimes



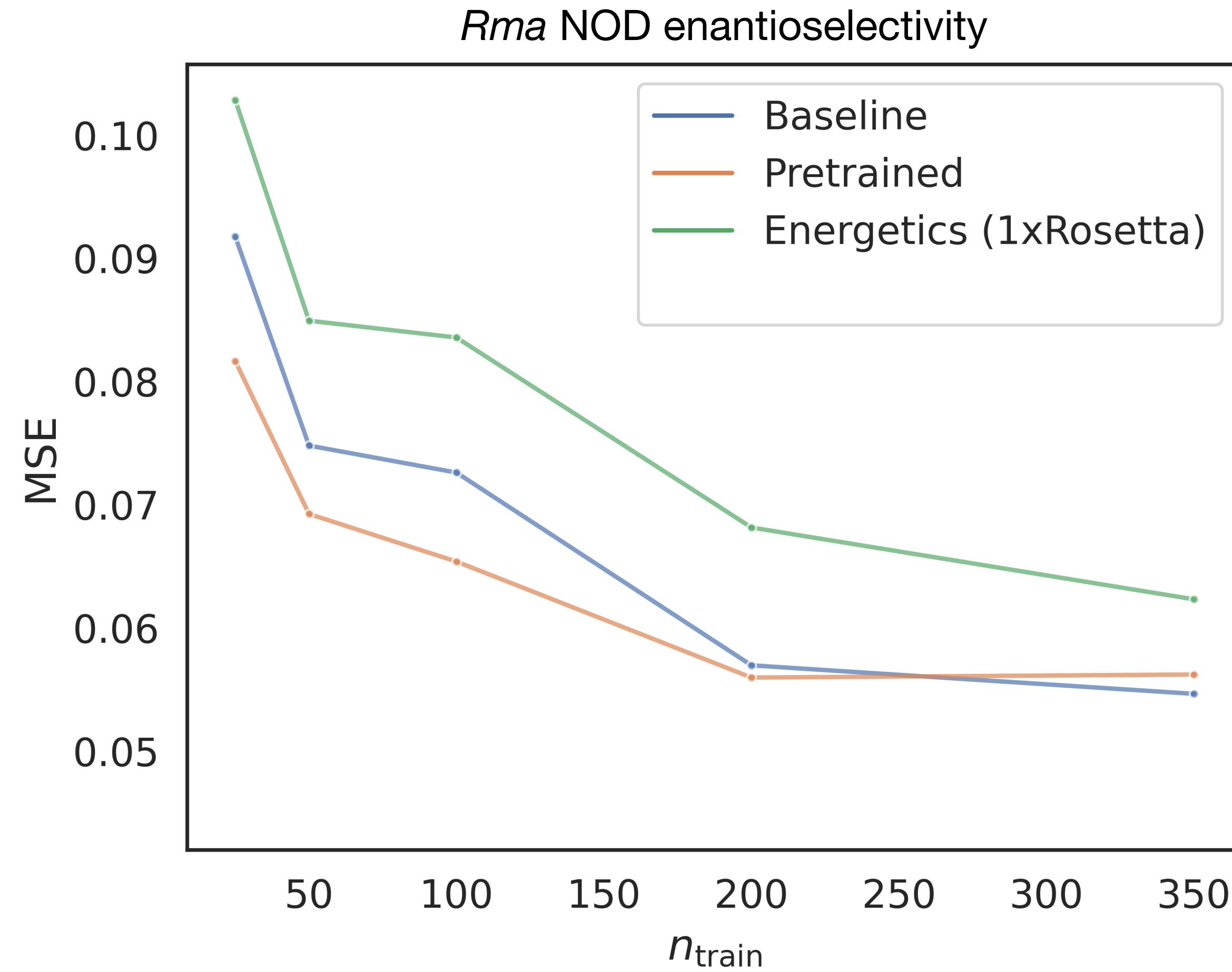
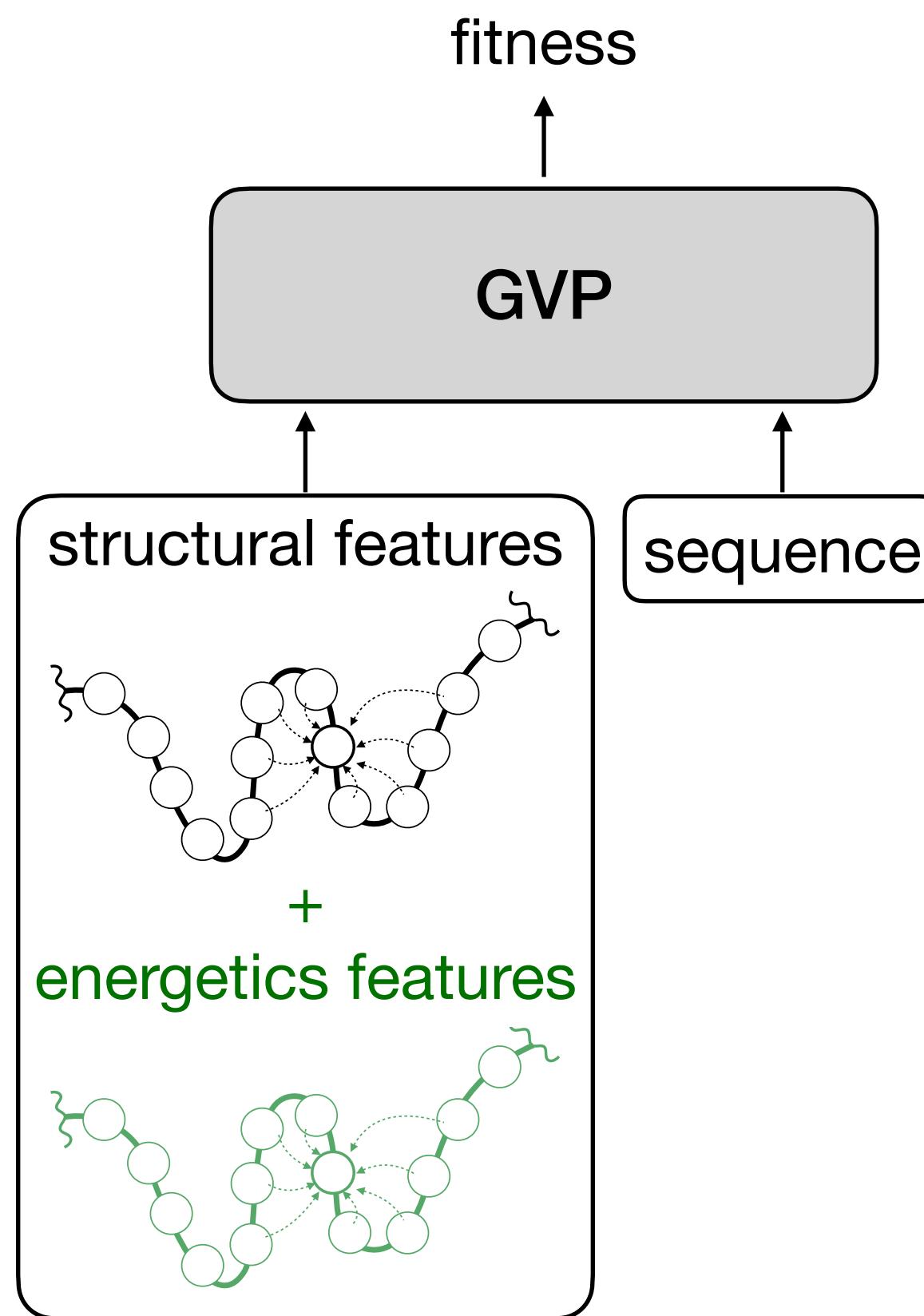
Energetics features improve performance in low-data regimes



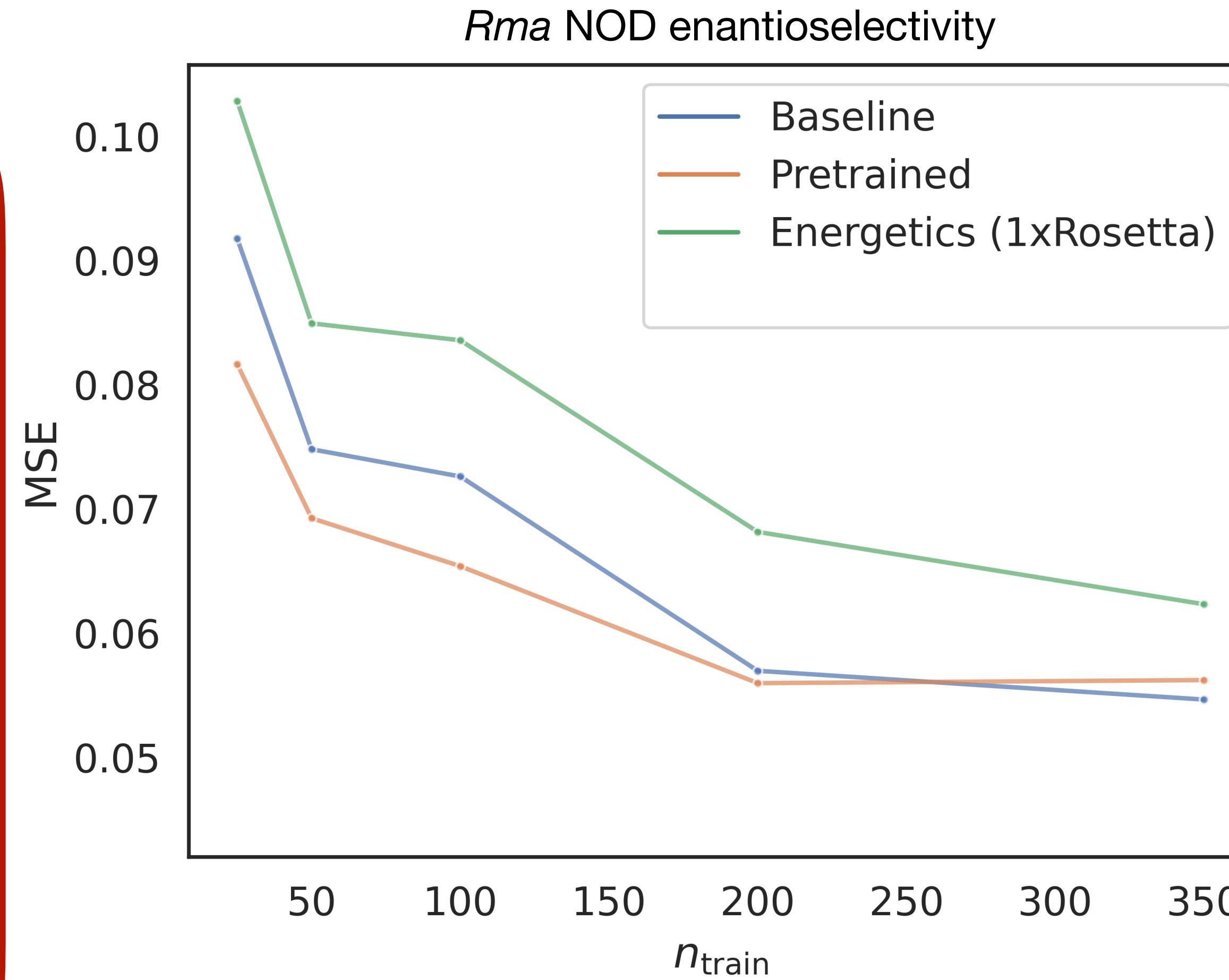
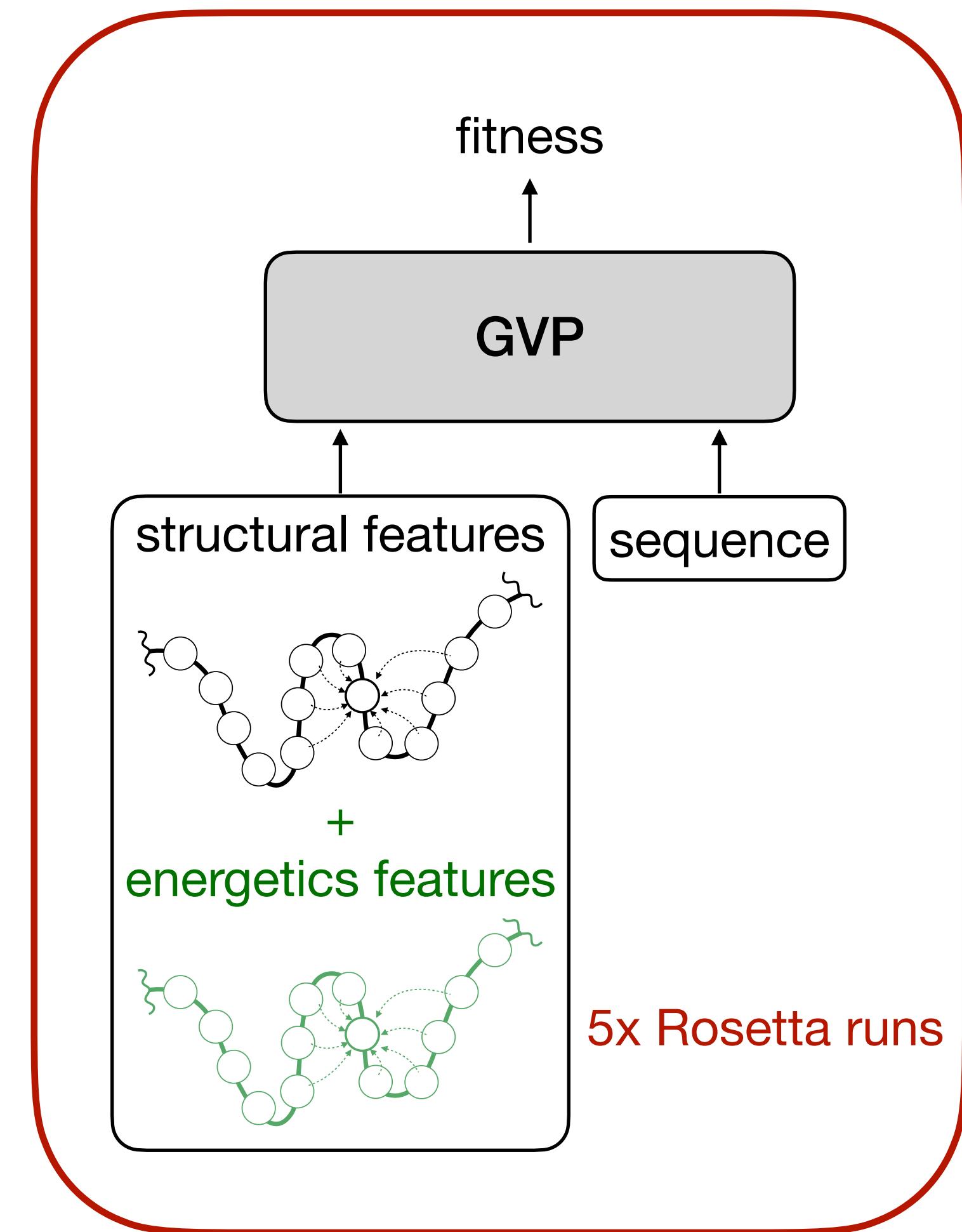
Energetics features improve performance in low-data regimes



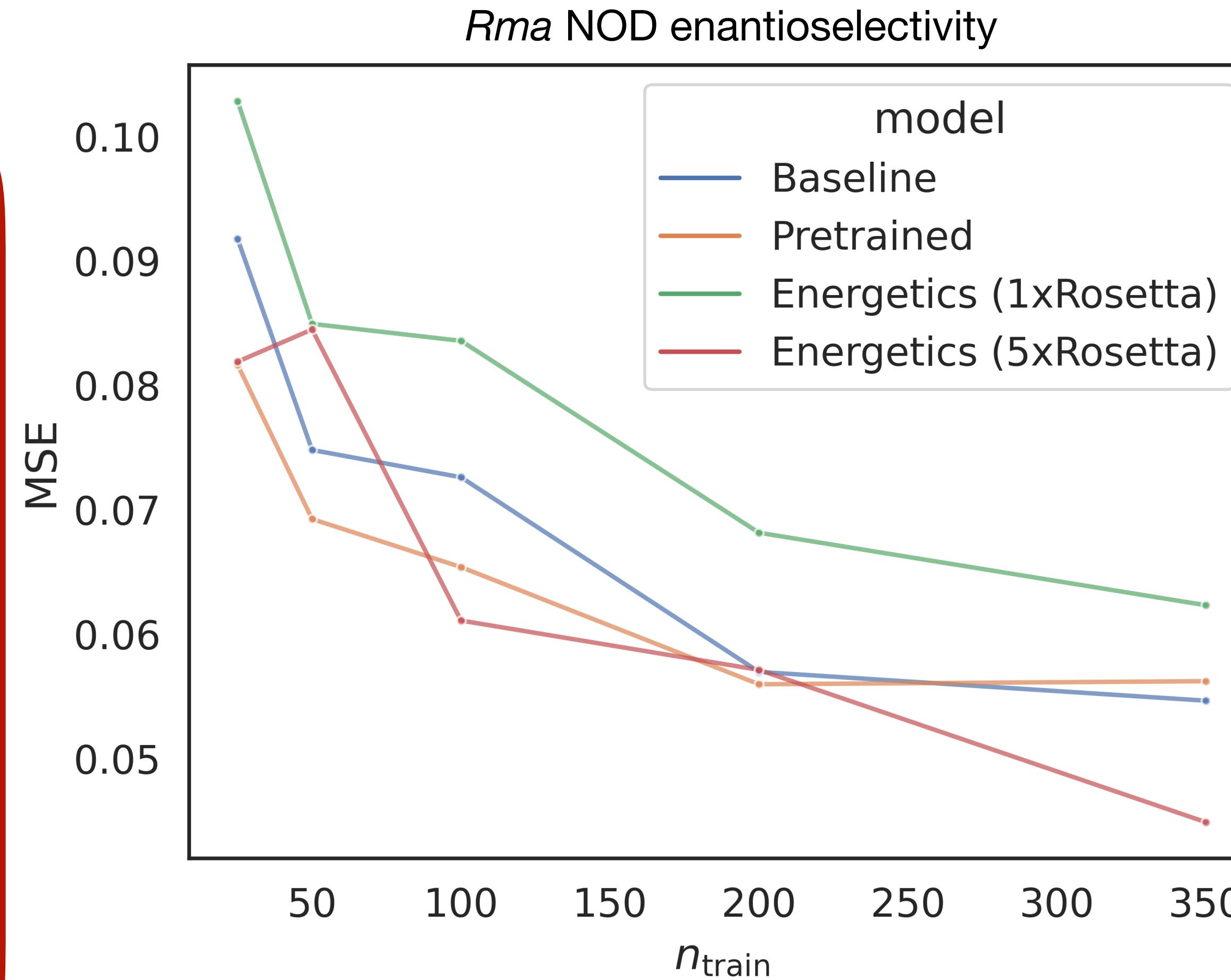
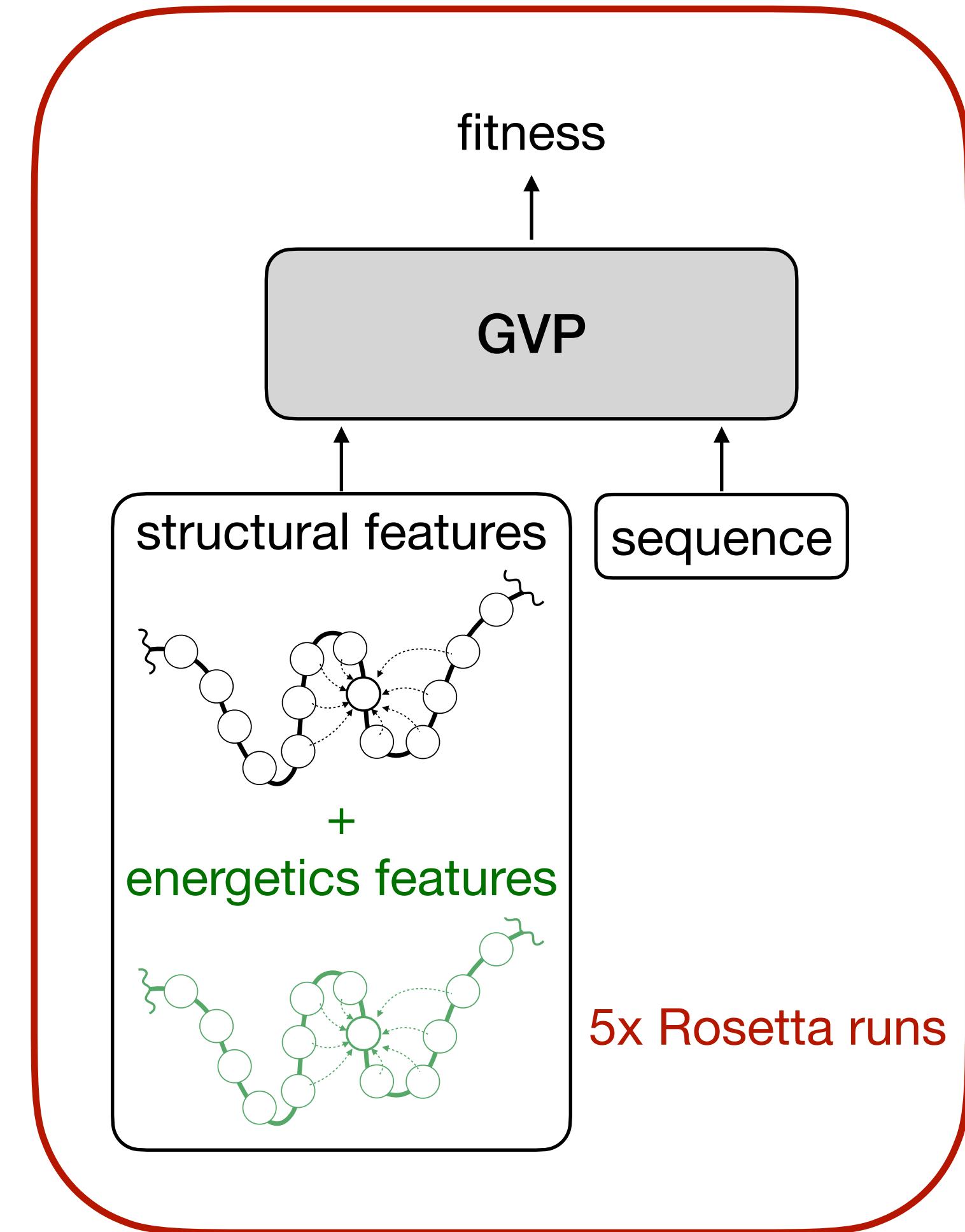
Energetics features improve performance in low-data regimes



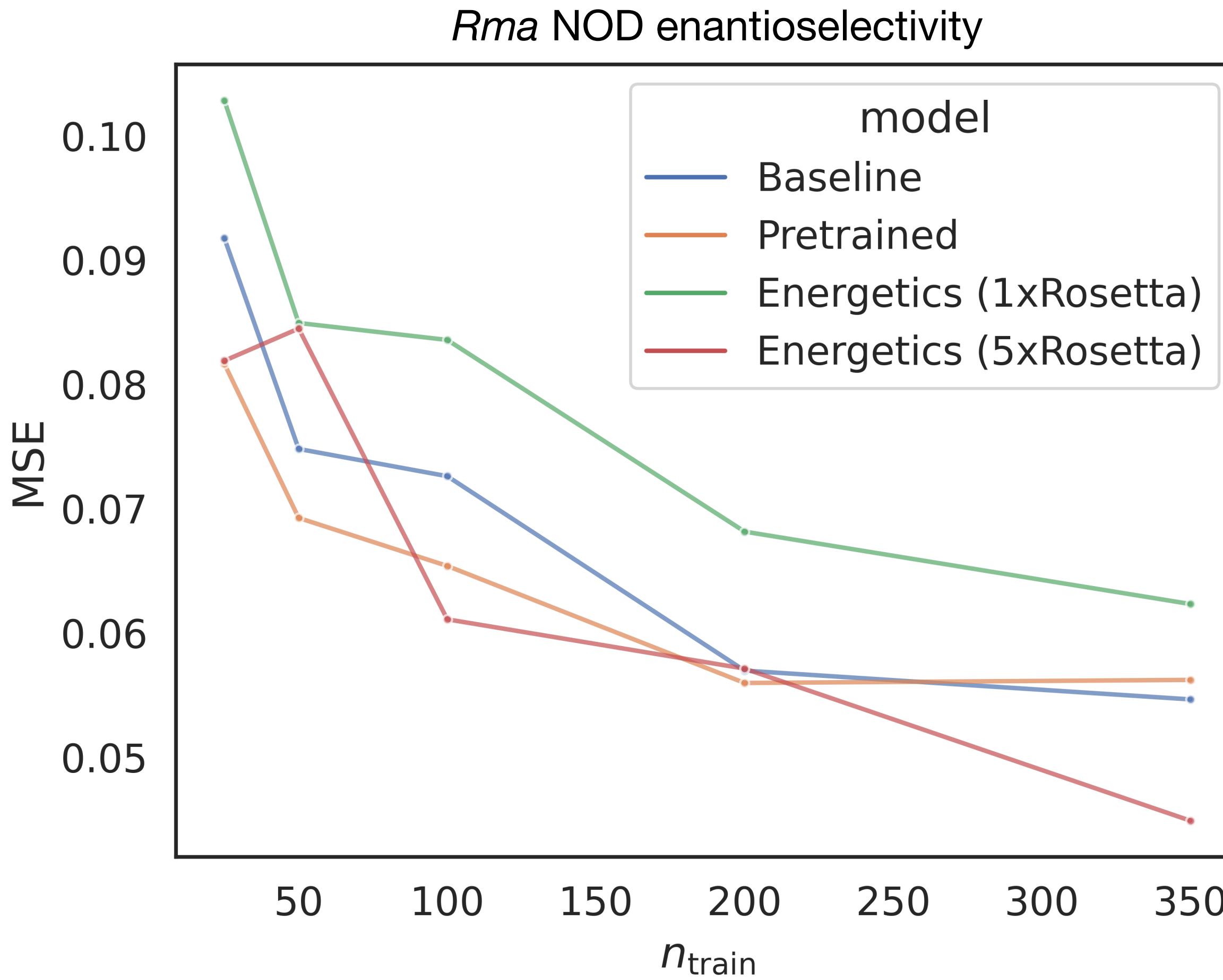
Energetics features improve performance in low-data regimes



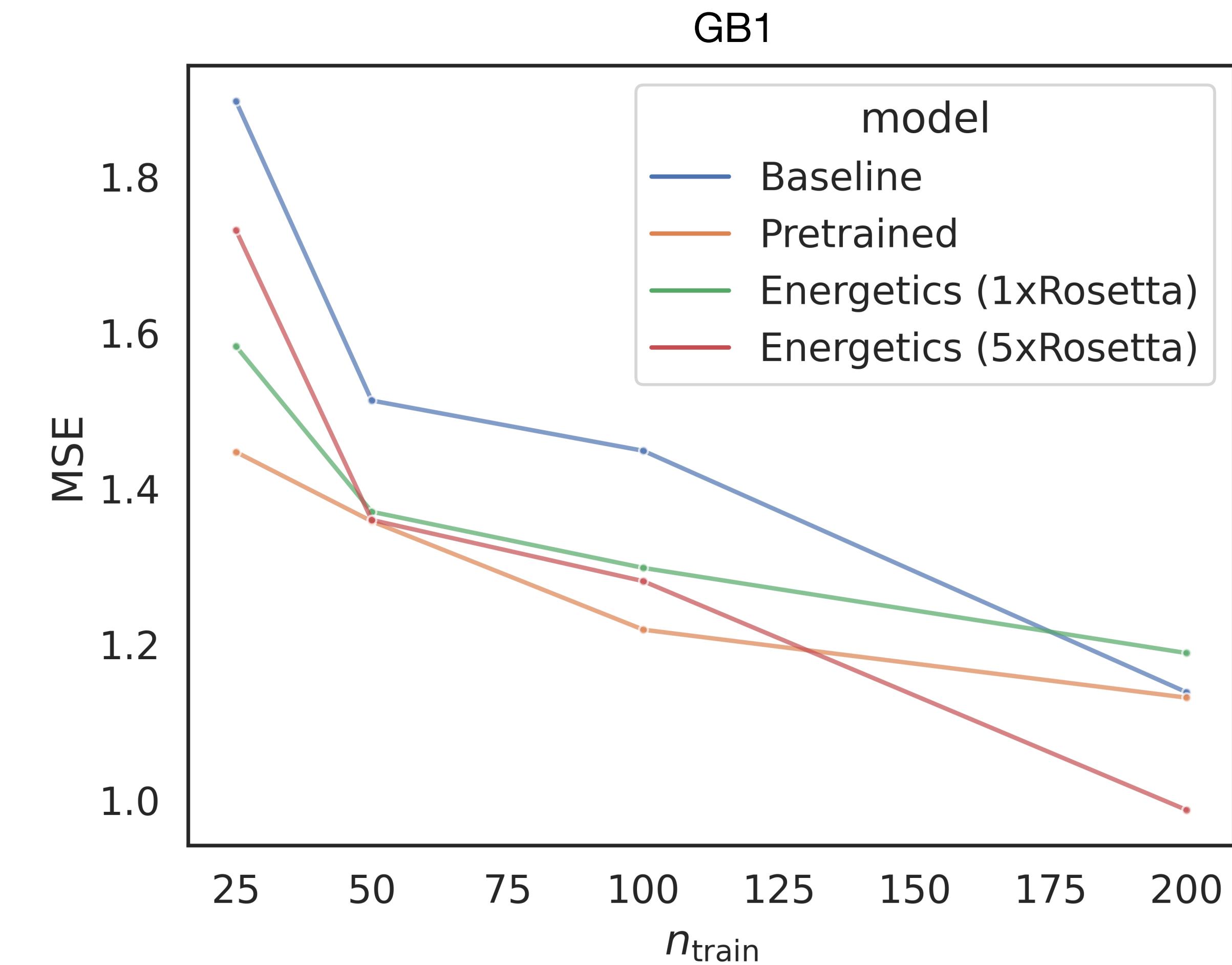
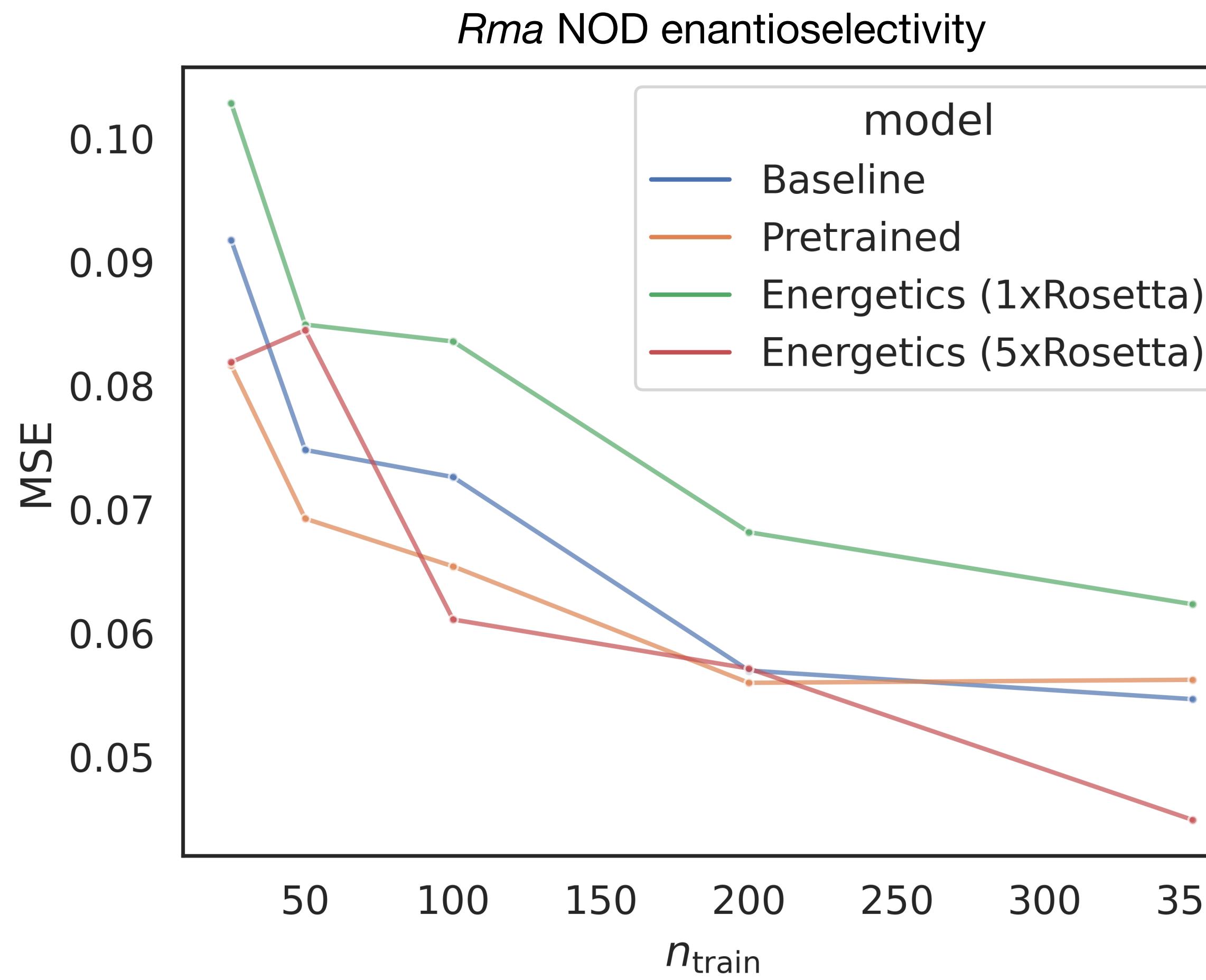
Energetics features improve performance in low-data regimes



Energetics features improve performance in low-data regimes



Energetics features improve performance in low-data regimes

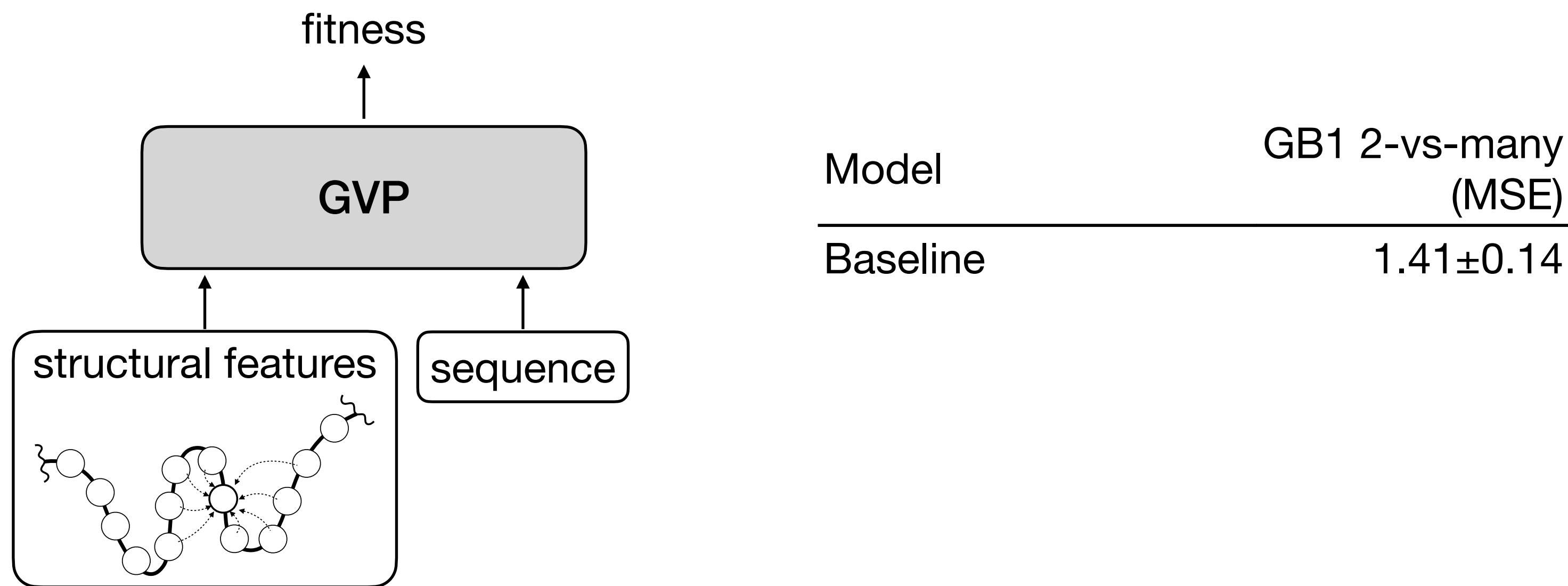


Energetics features match pretraining on OOD

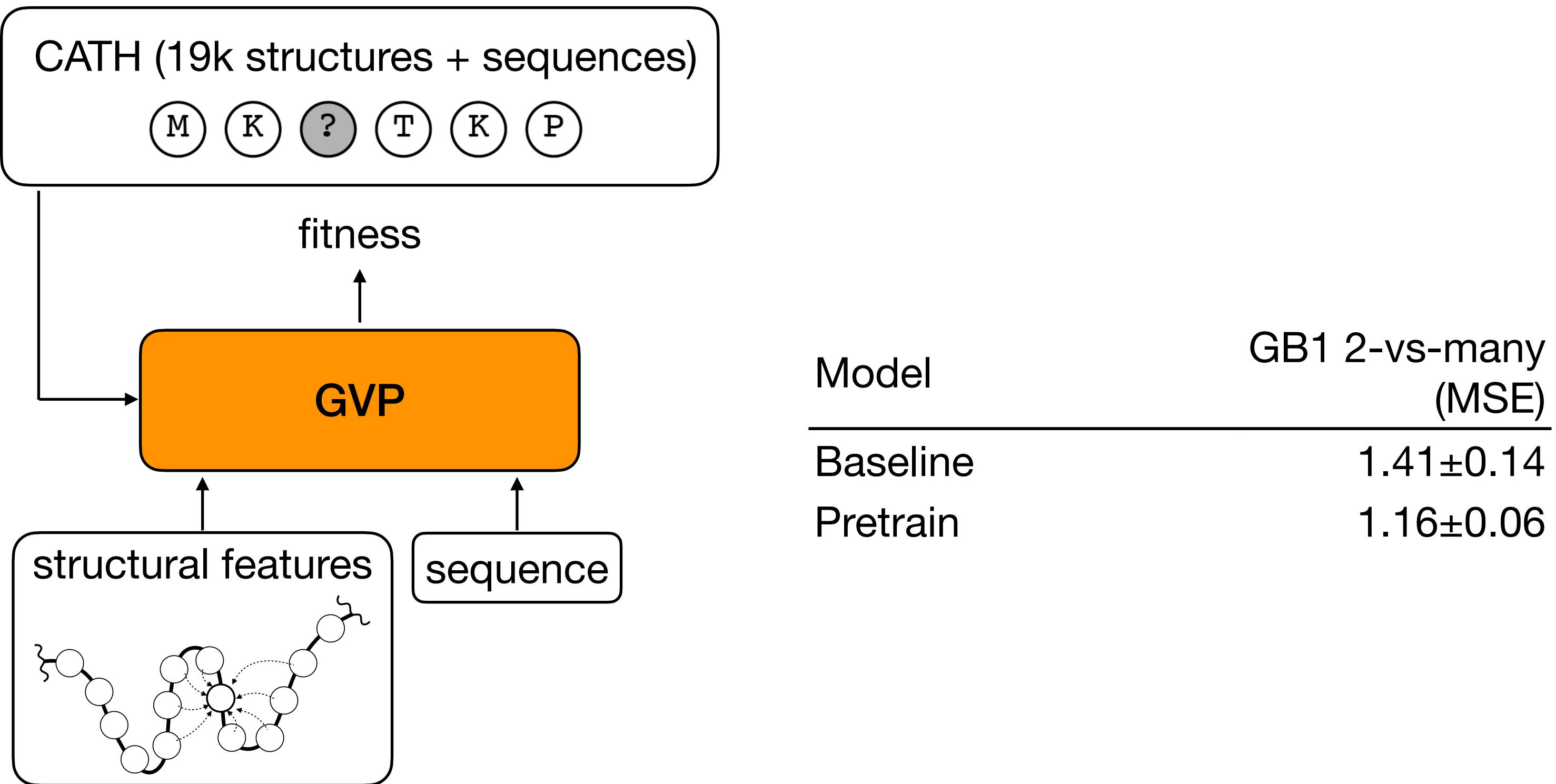
Energetics features match pretraining on OOD

Model	GB1 2-vs-many (MSE)
-------	------------------------

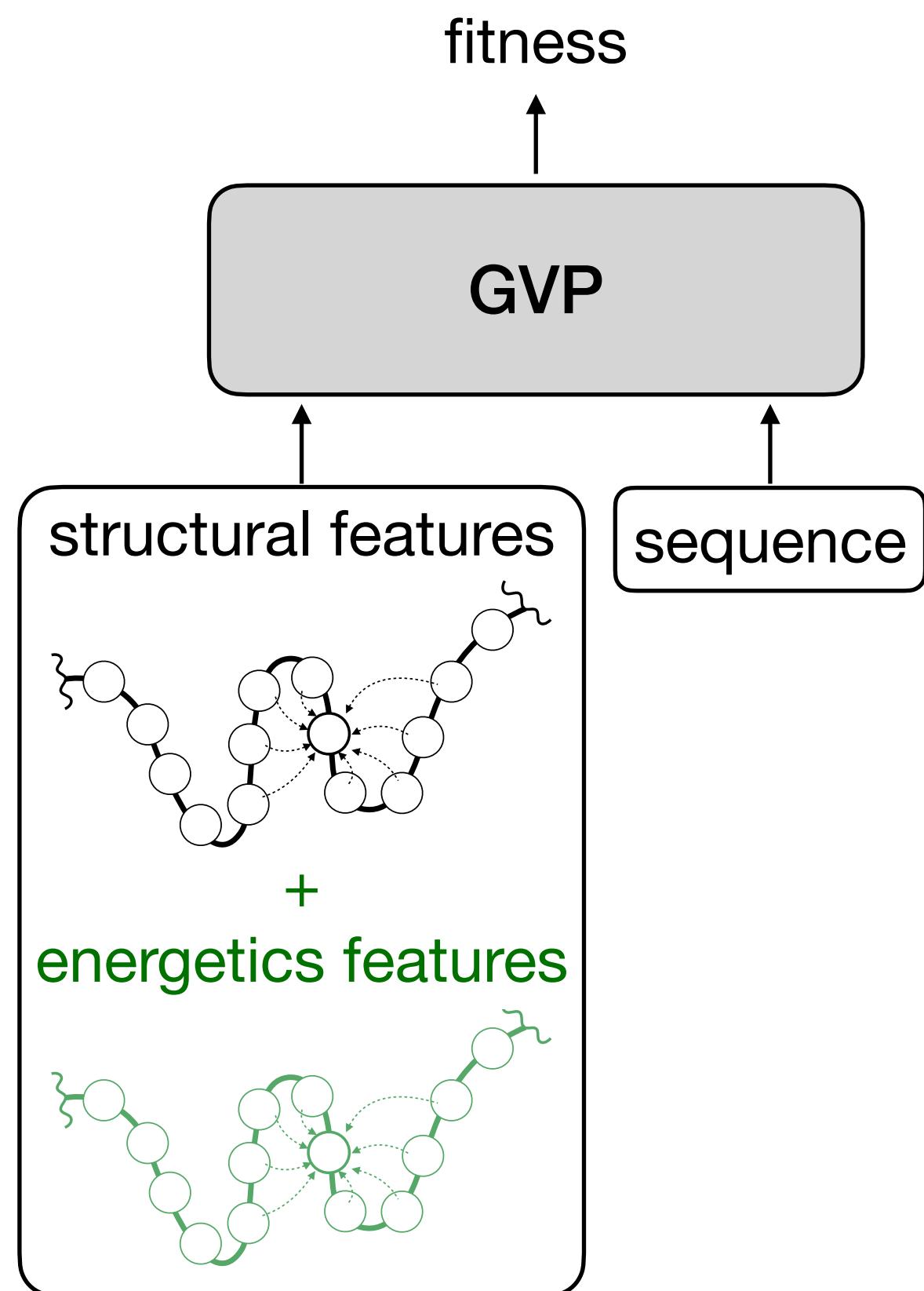
Energetics features match pretraining on OOD



Energetics features match pretraining on OOD

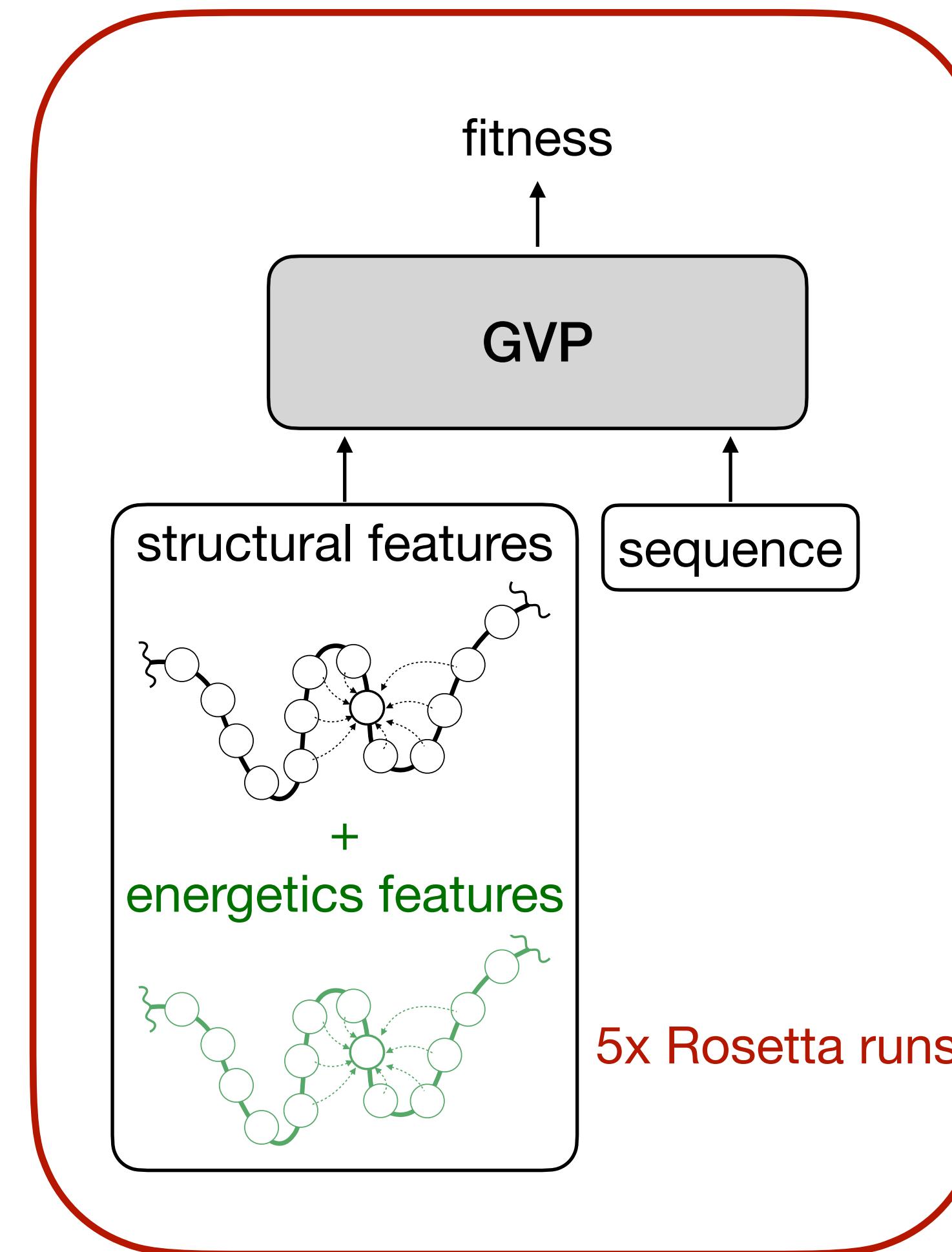


Energetics features match pretraining on OOD



Model	GB1 2-vs-many (MSE)
Baseline	1.41 ± 0.14
Pretrain	1.16 ± 0.06
Energetics (1xRosetta)	1.32 ± 0.13

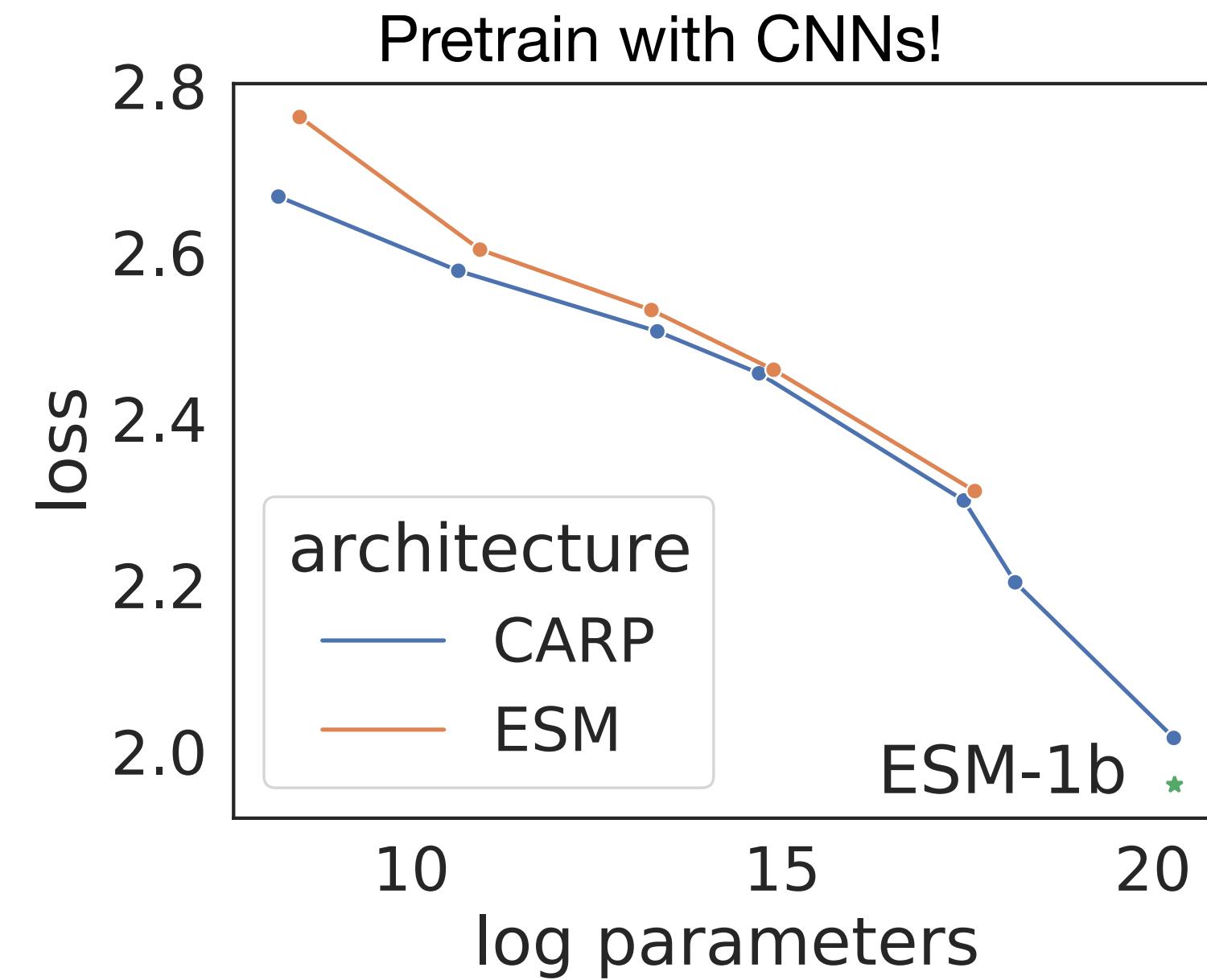
Energetics features match pretraining on OOD



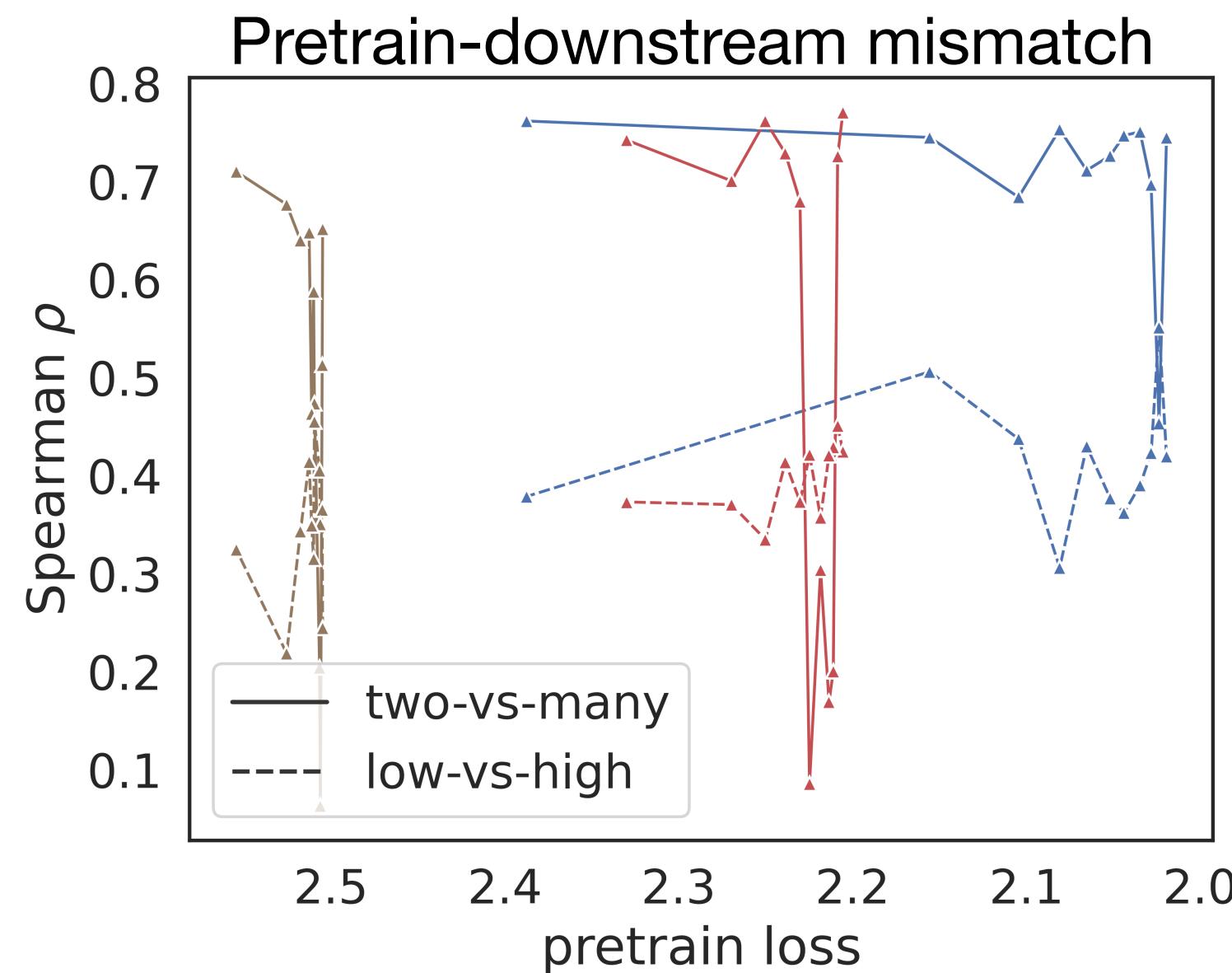
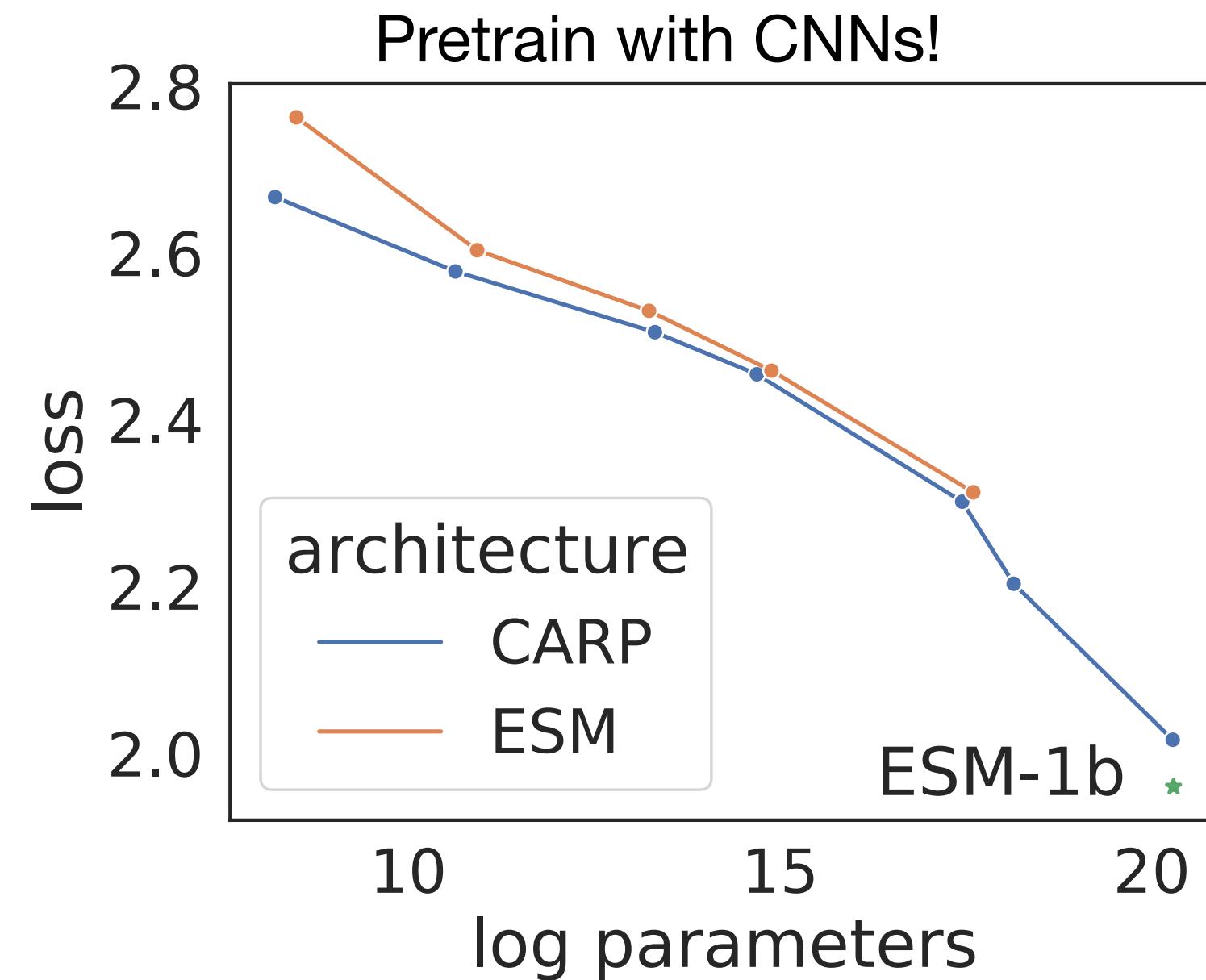
Model	GB1 2-vs-many (MSE)
Baseline	1.41 ± 0.14
Pretrain	1.16 ± 0.06
Energetics (1xRosetta)	1.32 ± 0.13
Energetics (5xRosetta)	1.16 ± 0.05

Use multimodal data to reduce pretrain-downstream mismatch

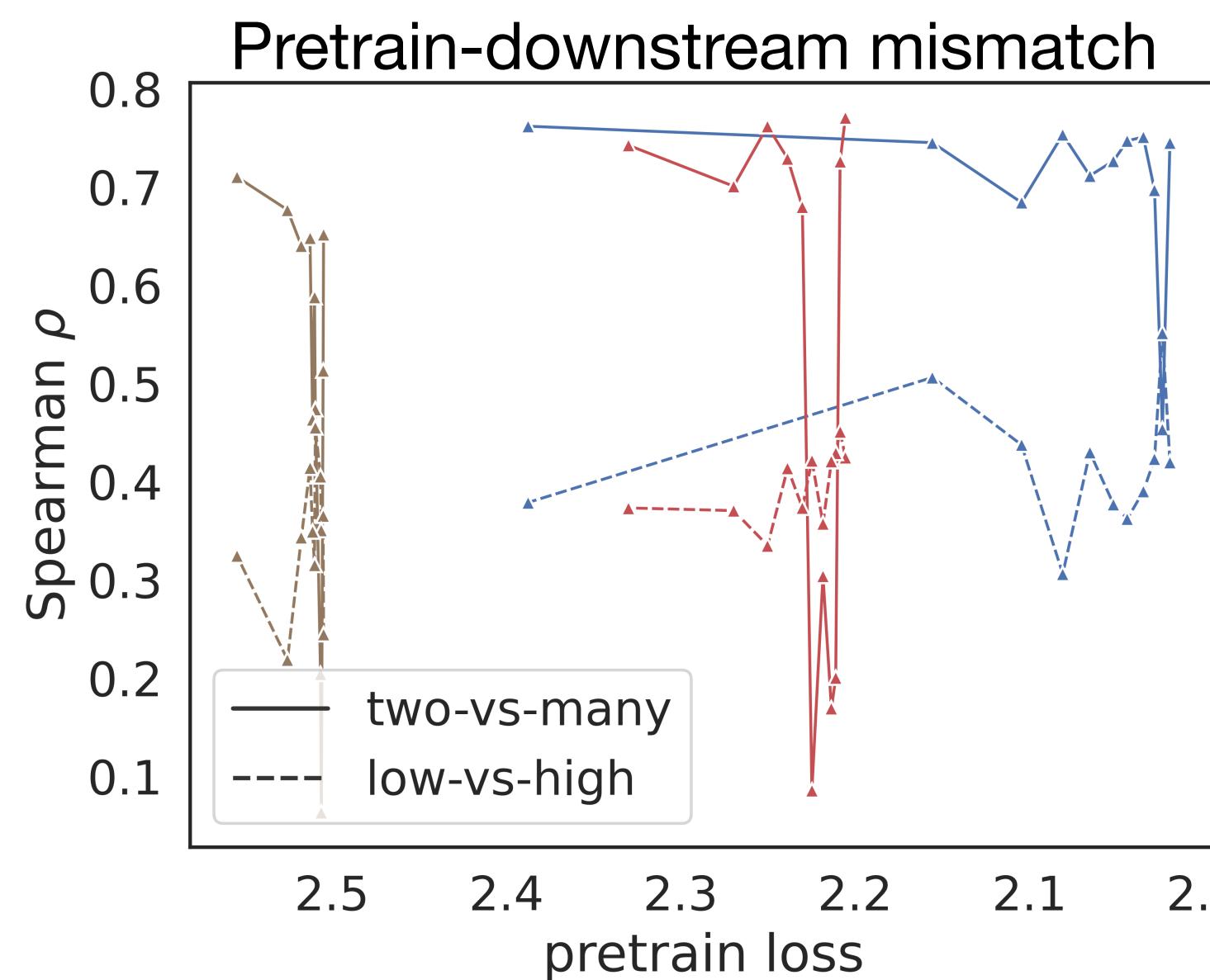
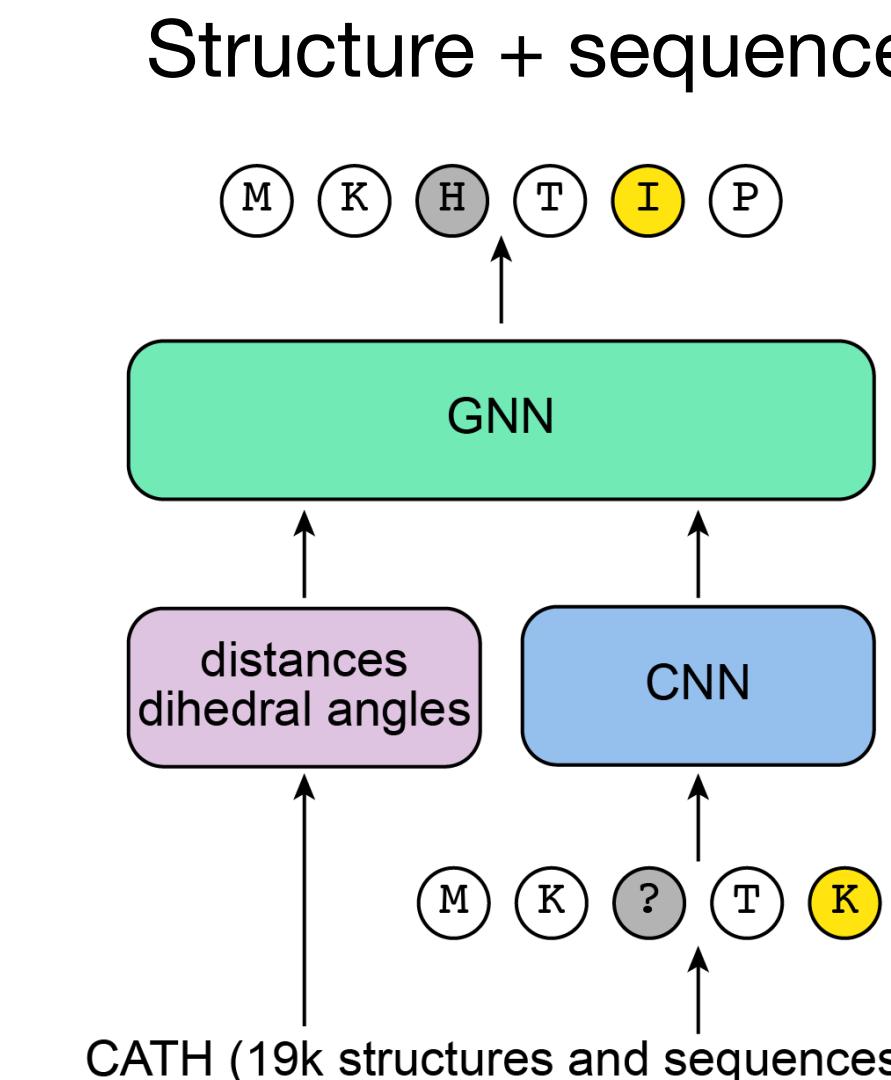
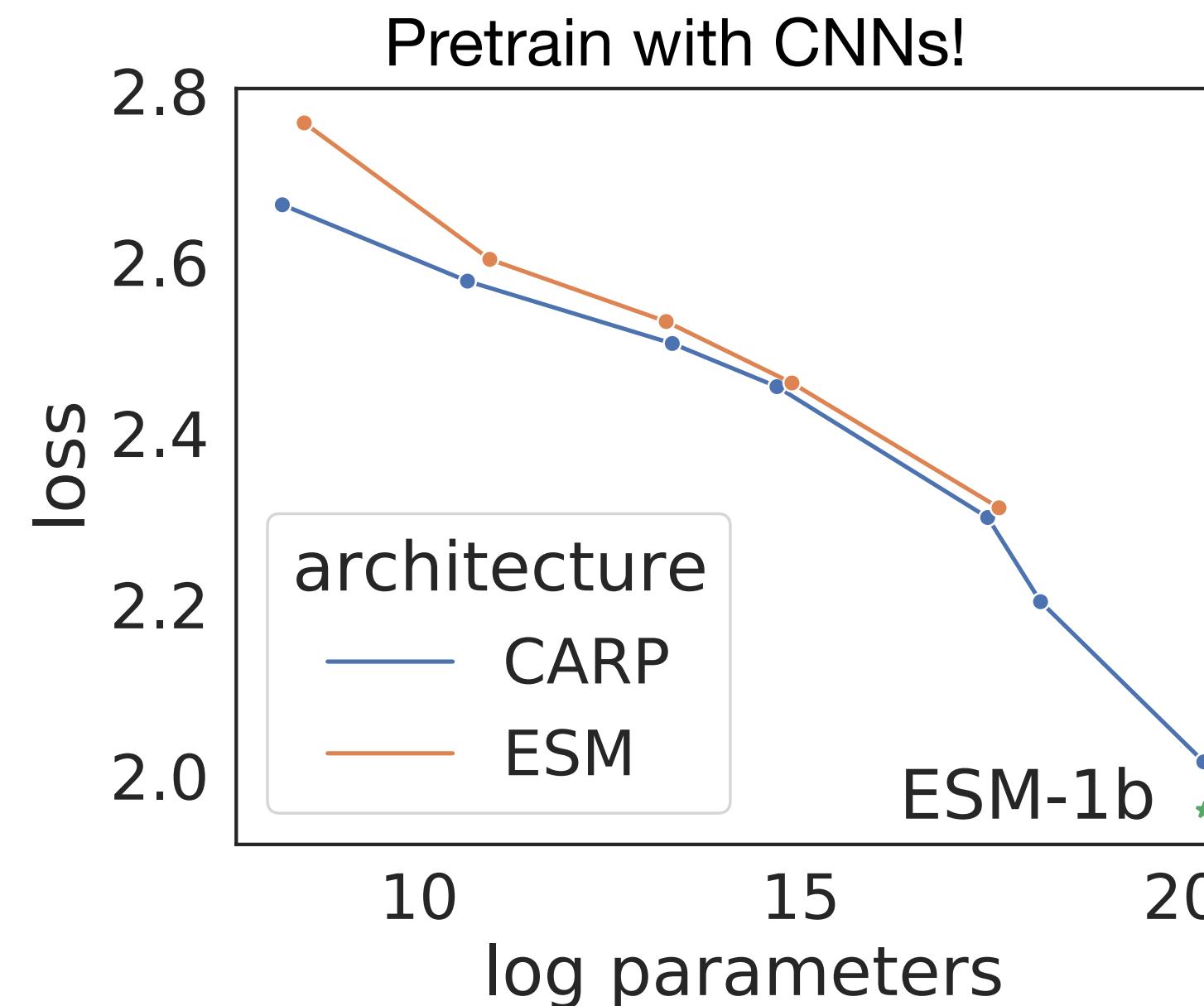
Use multimodal data to reduce pretrain-downstream mismatch



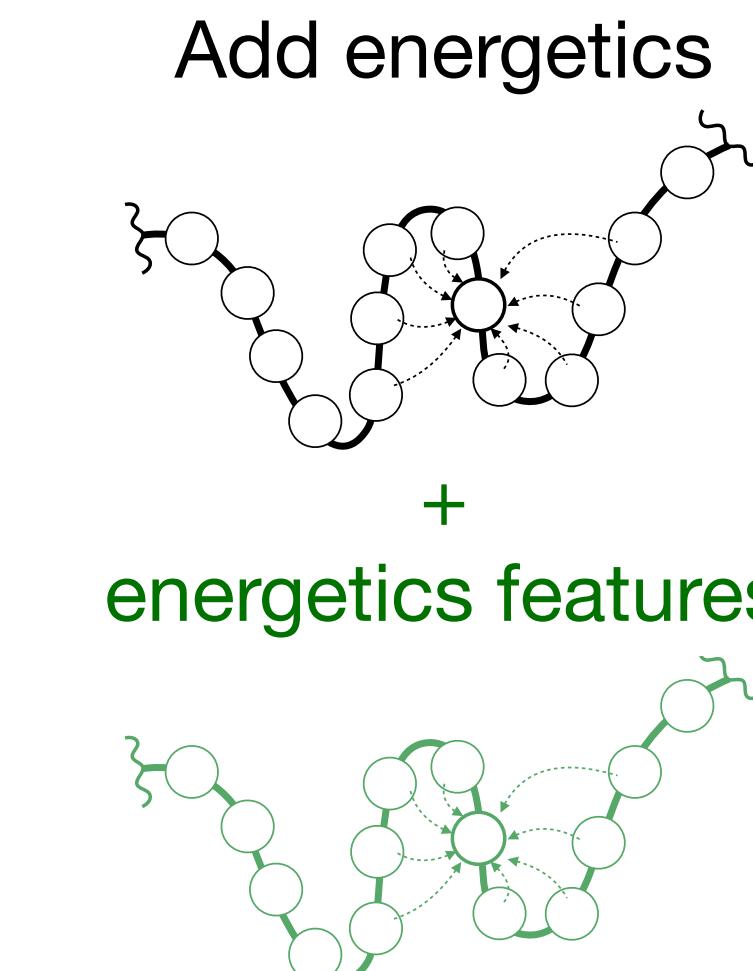
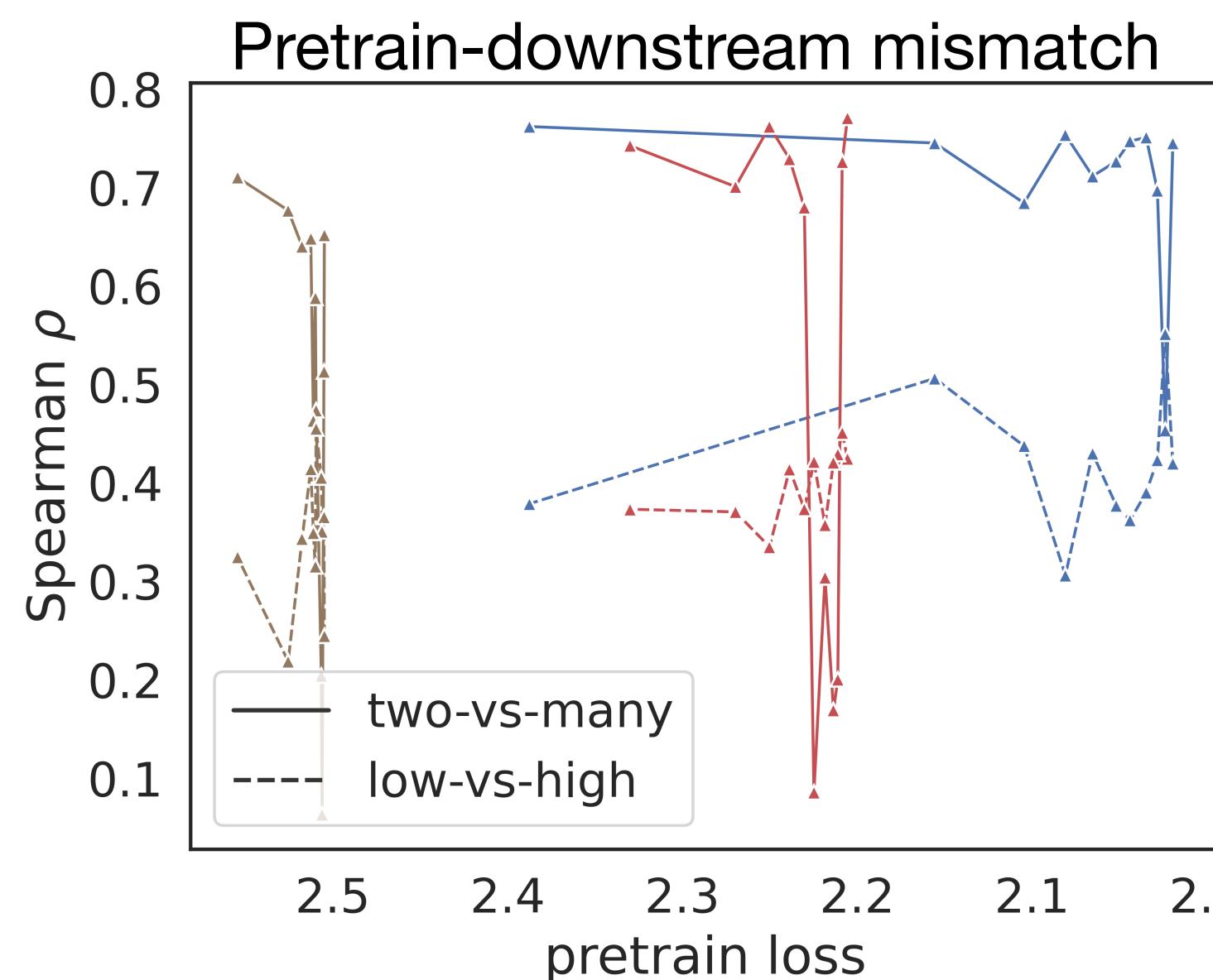
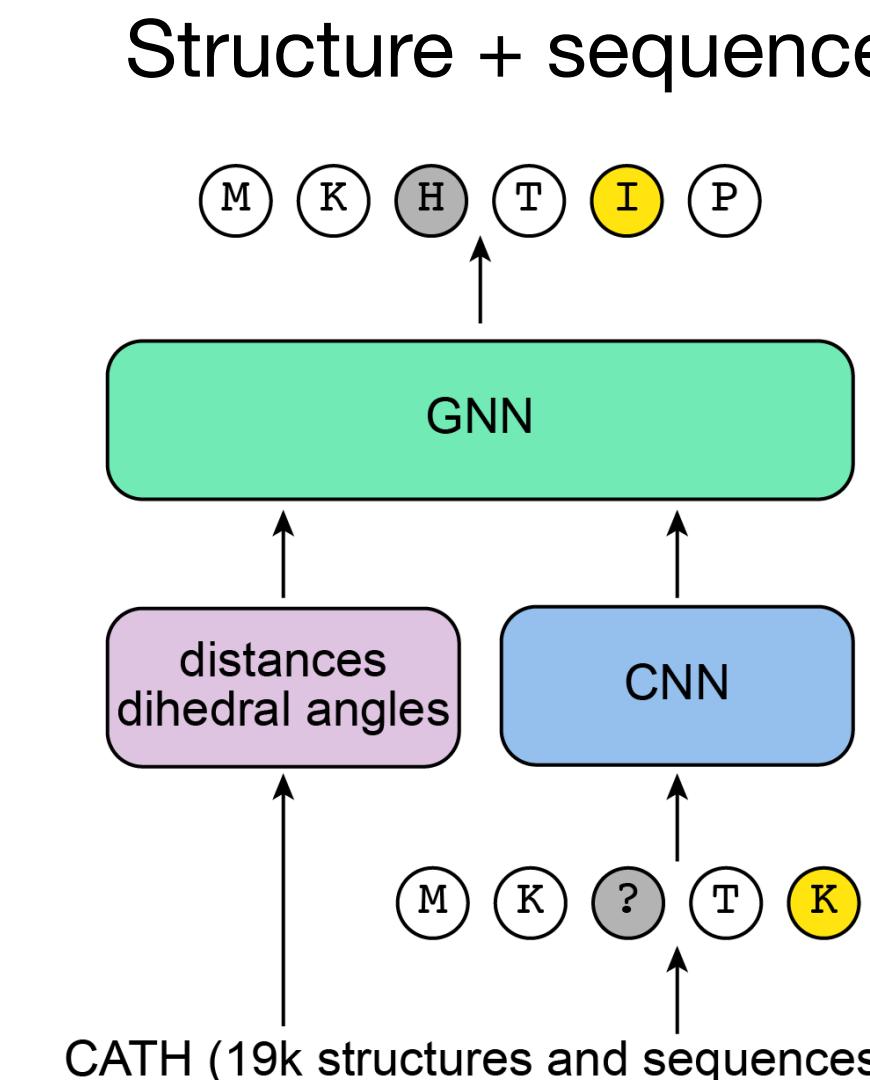
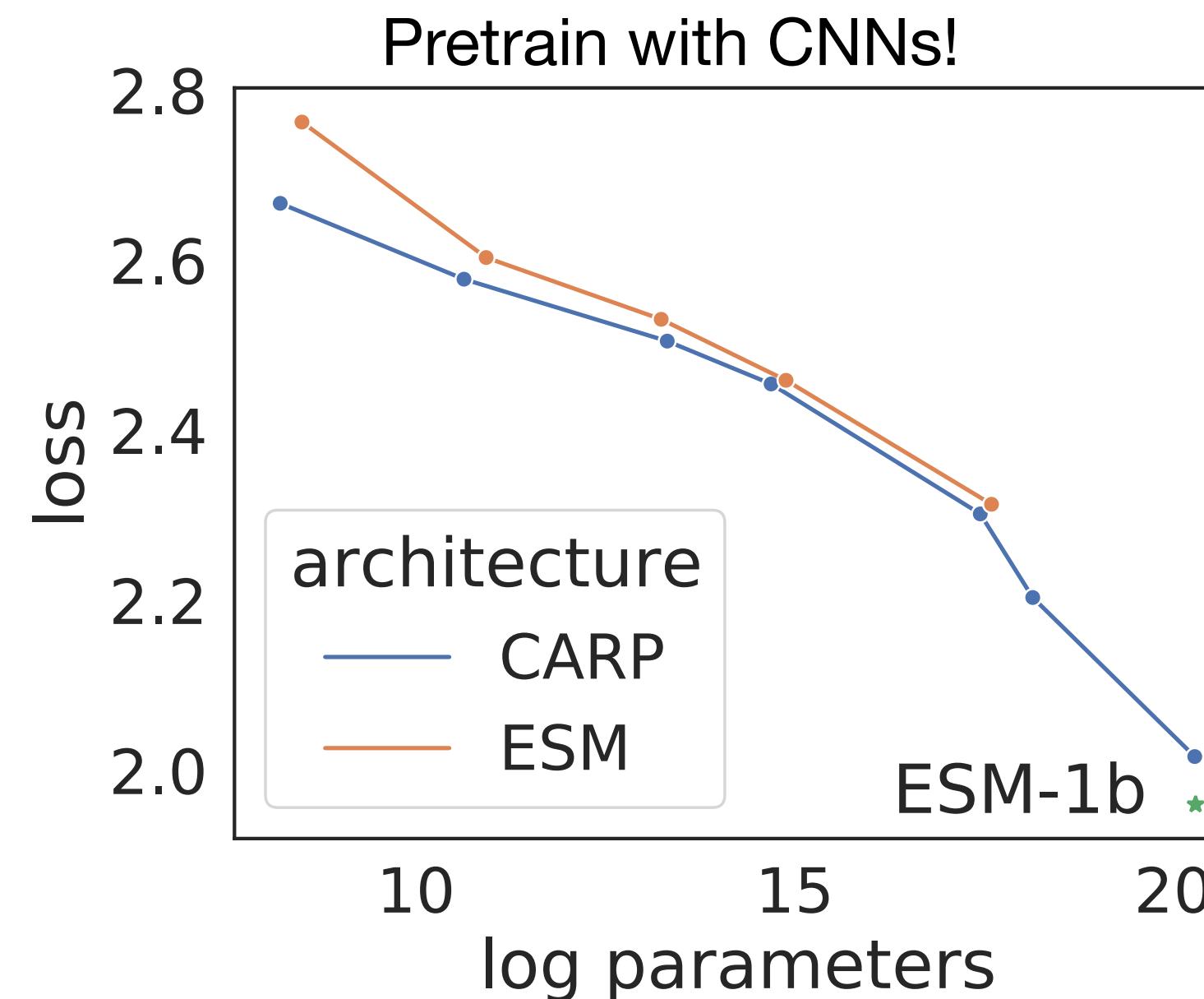
Use multimodal data to reduce pretrain-downstream mismatch



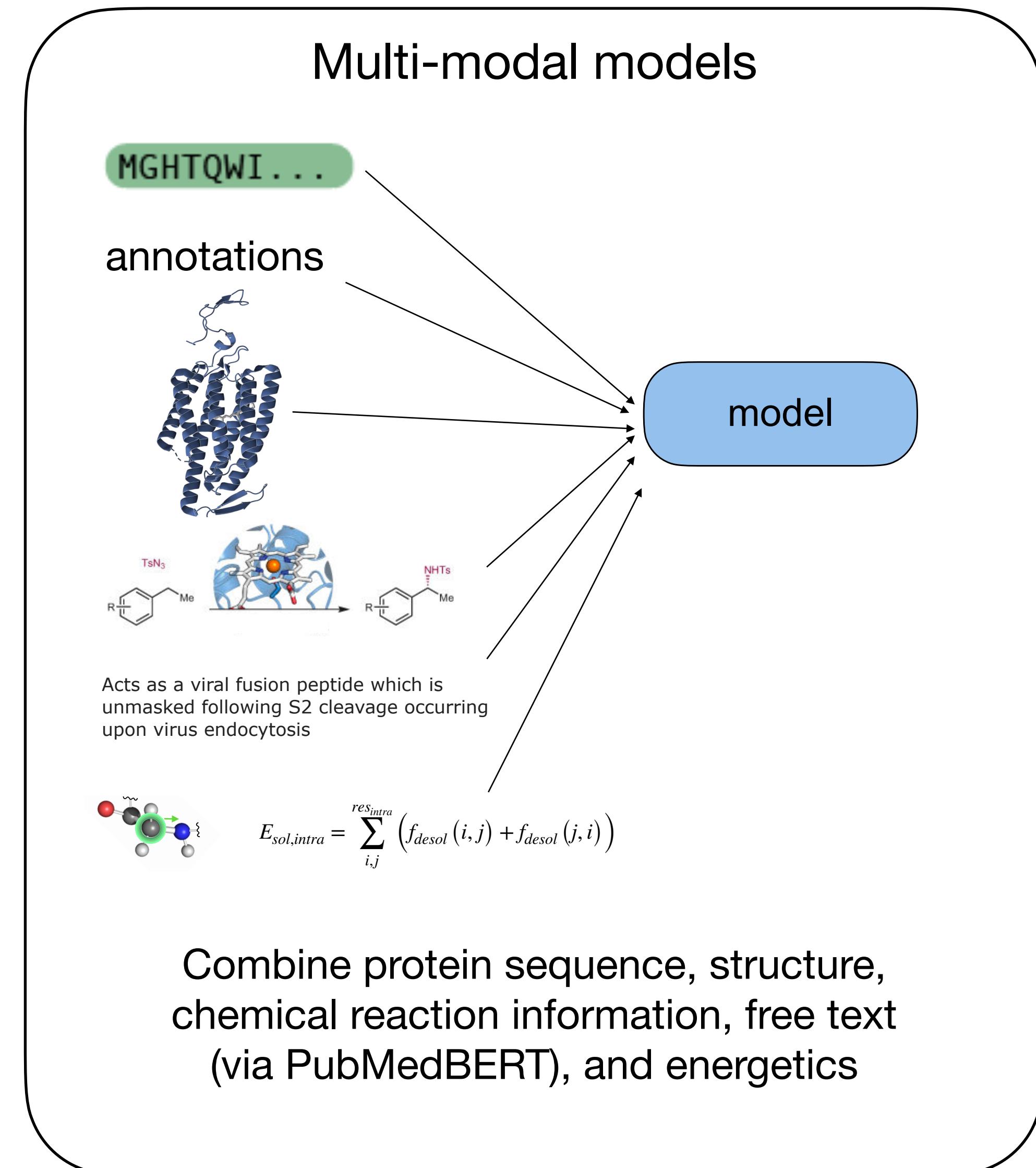
Use multimodal data to reduce pretrain-downstream mismatch



Use multimodal data to reduce pretrain-downstream mismatch



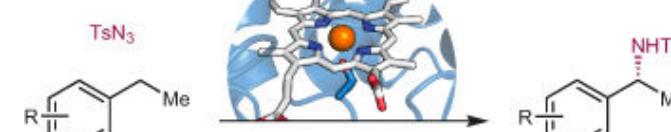
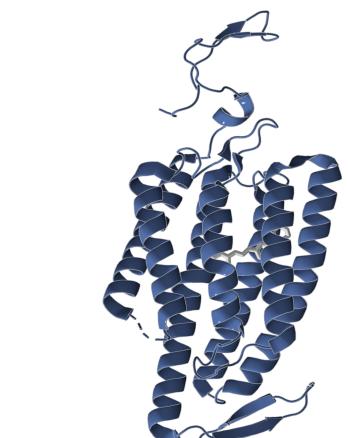
Can different modalities and pretraining tasks do better?



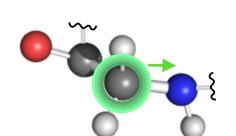
Multi-modal models

MGHTQWI . . .

annotations



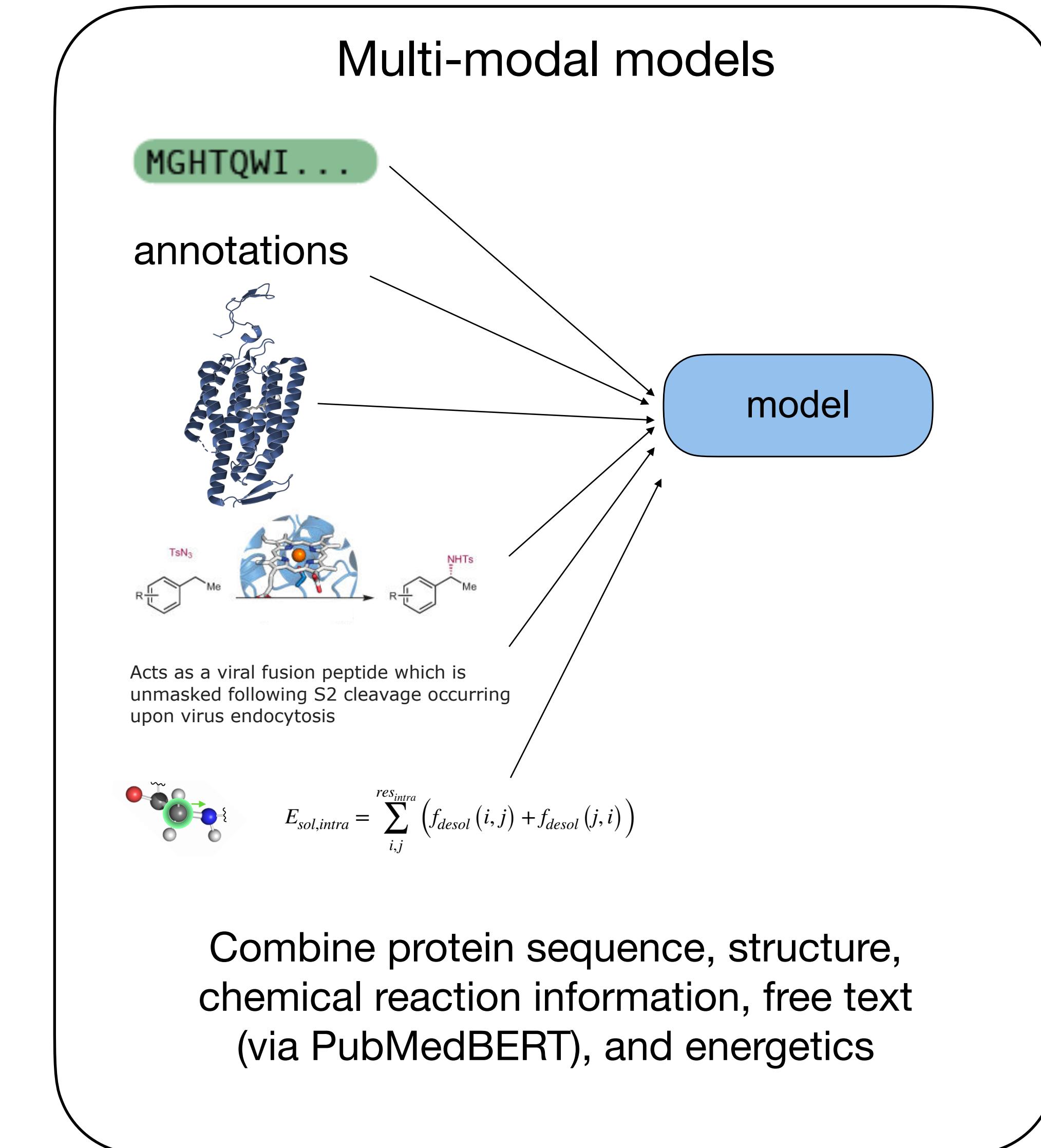
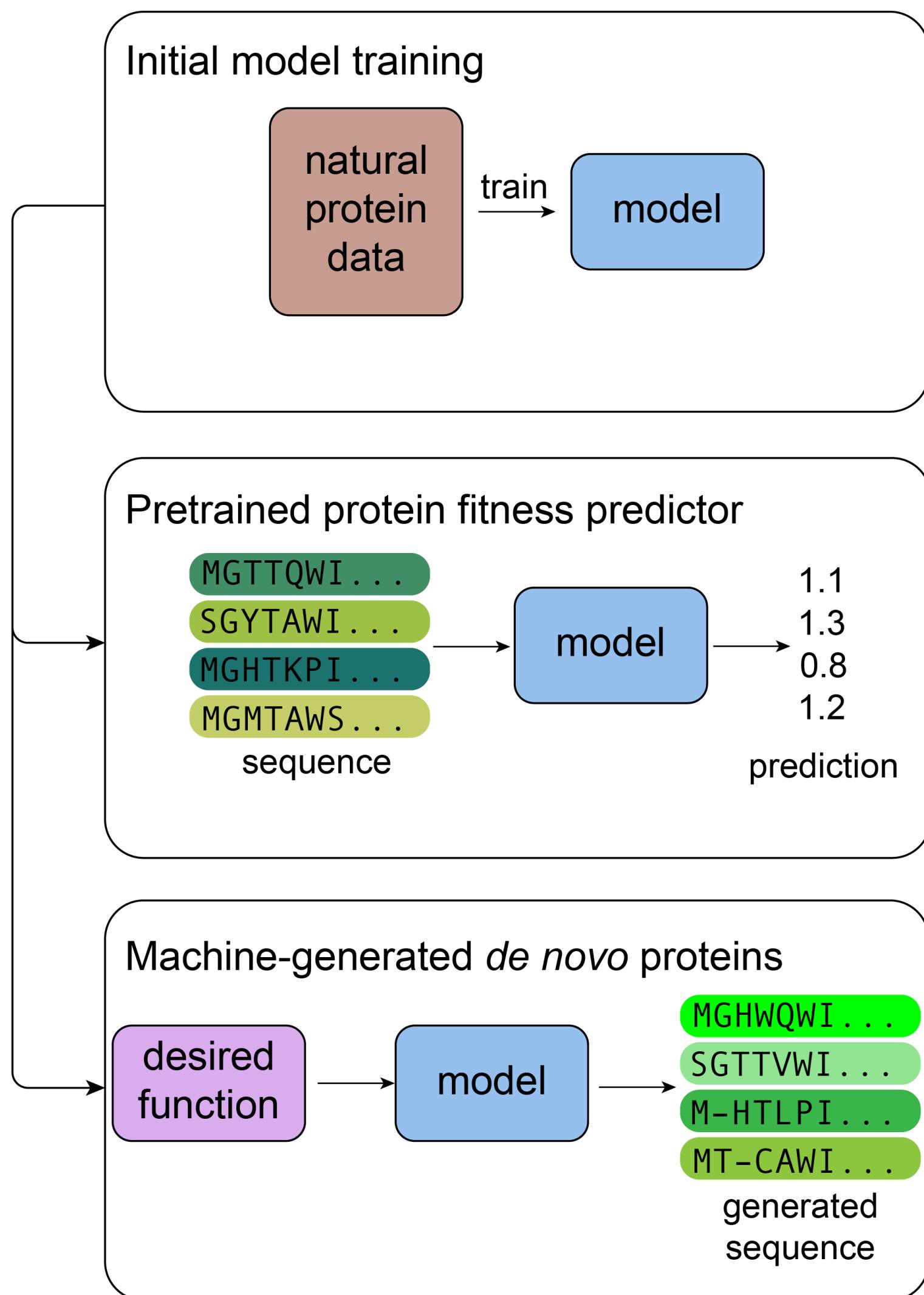
Acts as a viral fusion peptide which is unmasked following S2 cleavage occurring upon virus endocytosis



$$E_{sol,intra} = \sum_{i,j}^{res_{intra}} (f_{desol}(i,j) + f_{desol}(j,i))$$

Combine protein sequence, structure, chemical reaction information, free text (via PubMedBERT), and energetics

Can we generate functional proteins?





BioML at MSR New England

Acknowledgments



BioML at MSR New England