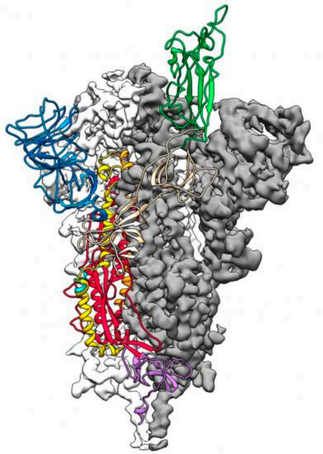
Multimodal deep learning for protein engineering

Kevin Kaichuang Yang
Microsoft Research New England
@KevinKaichuang

Proteins are biology’s actuators

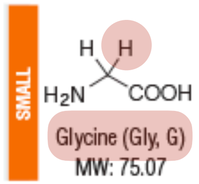
- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling

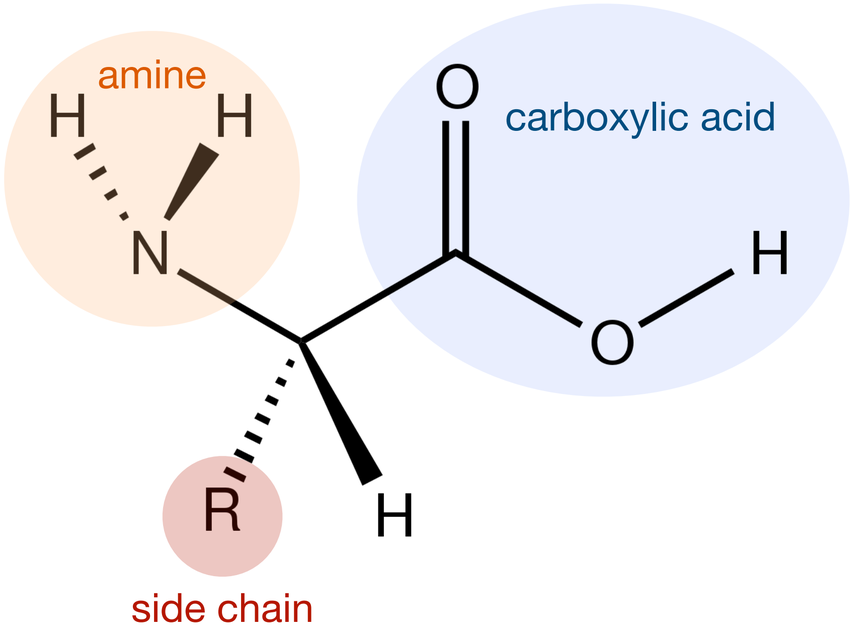


coronavirus spike protein

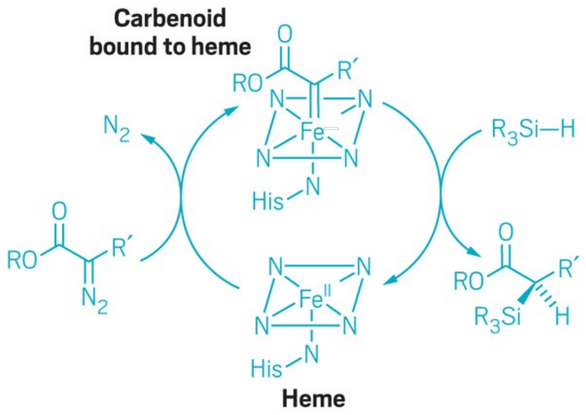
luciferase

Diversity arises from 20 building blocks

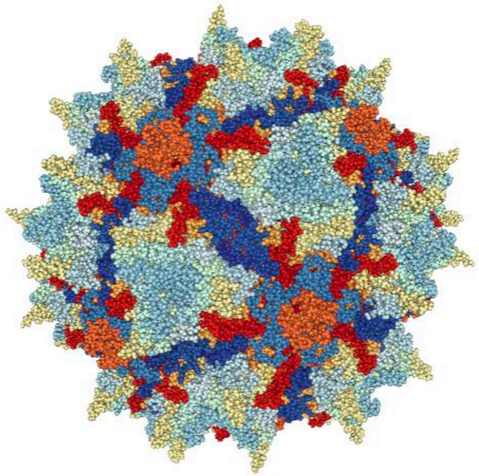




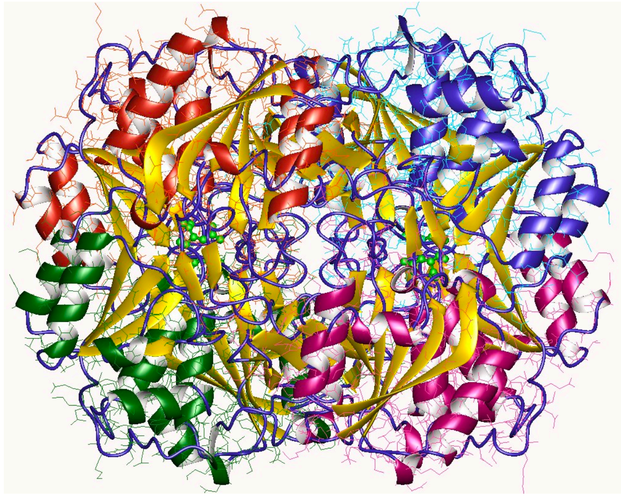
Why design proteins?



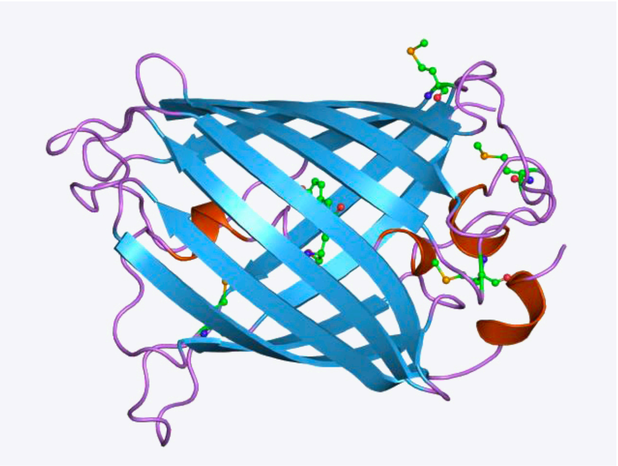
new chemistry



understand function

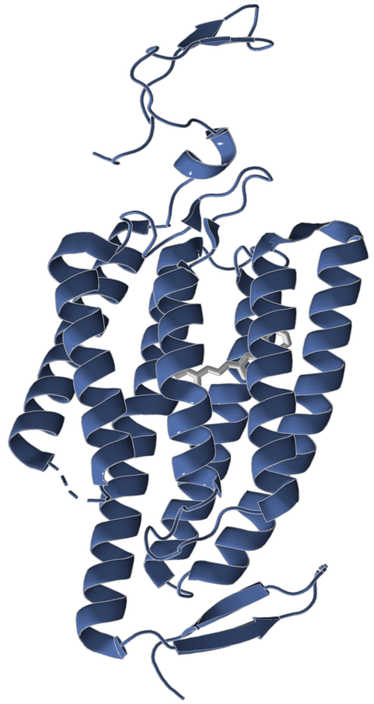
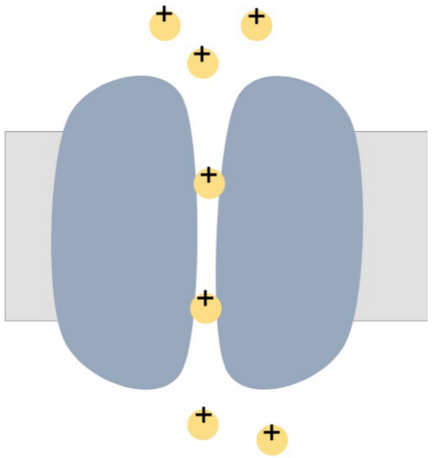


therapeutics



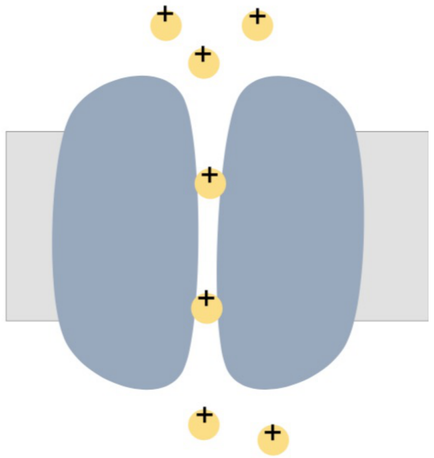
molecular tools

Protein engineering requires going from function to sequence

MGTGDHDD... → ? →  → ?? → 

What sequence will give the desired function?

Protein engineering requires going from function to sequence

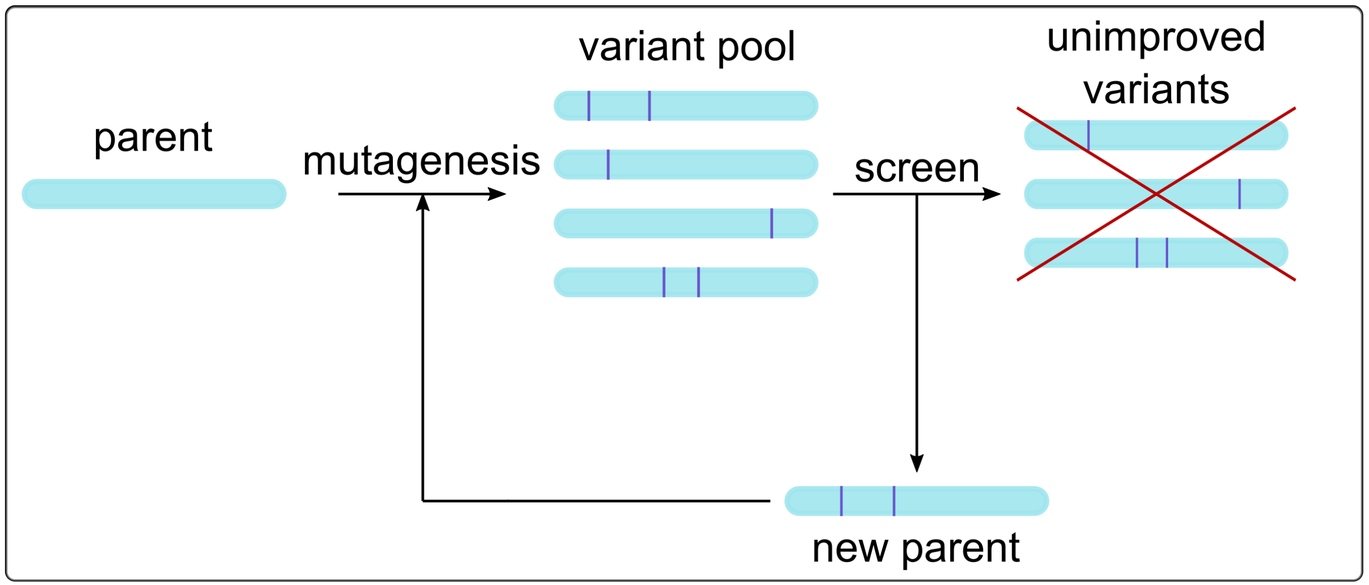


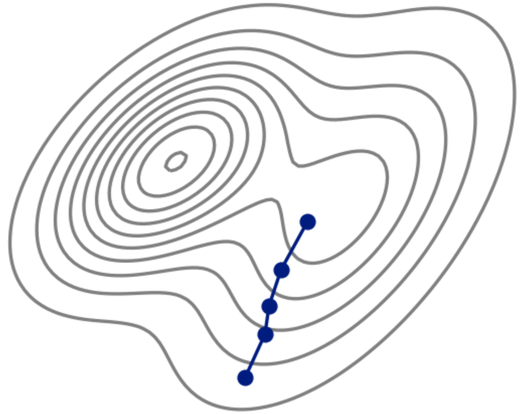
???

MGTGDHDD...

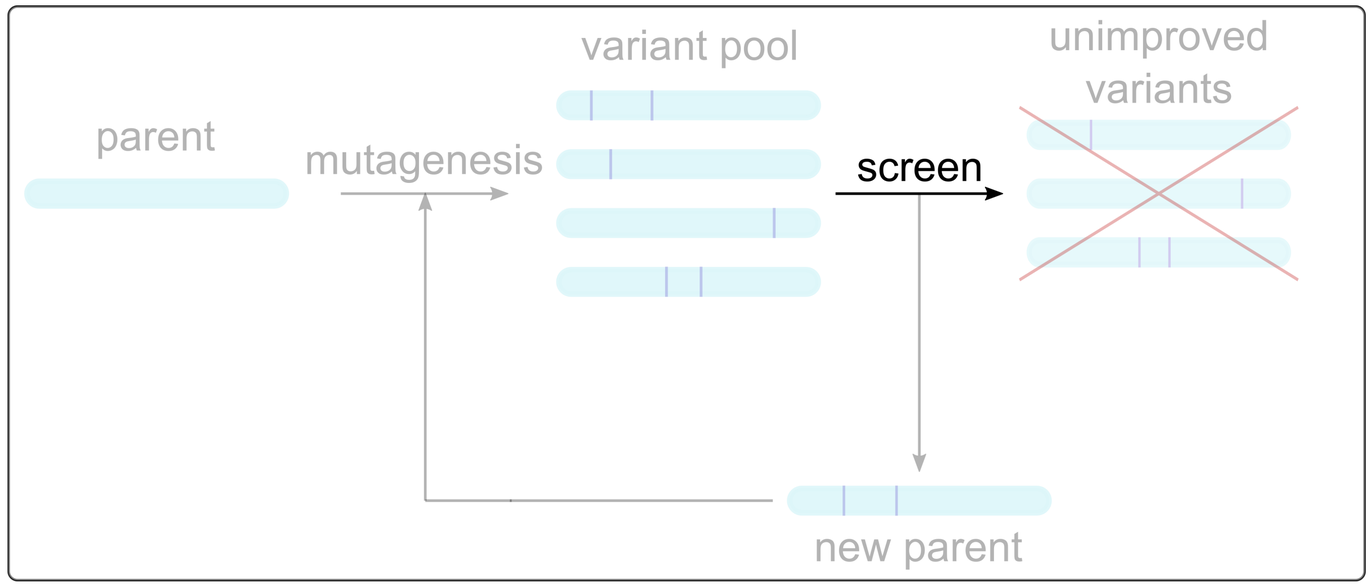
What sequence will give the desired function?

Directed evolution sidesteps the problem



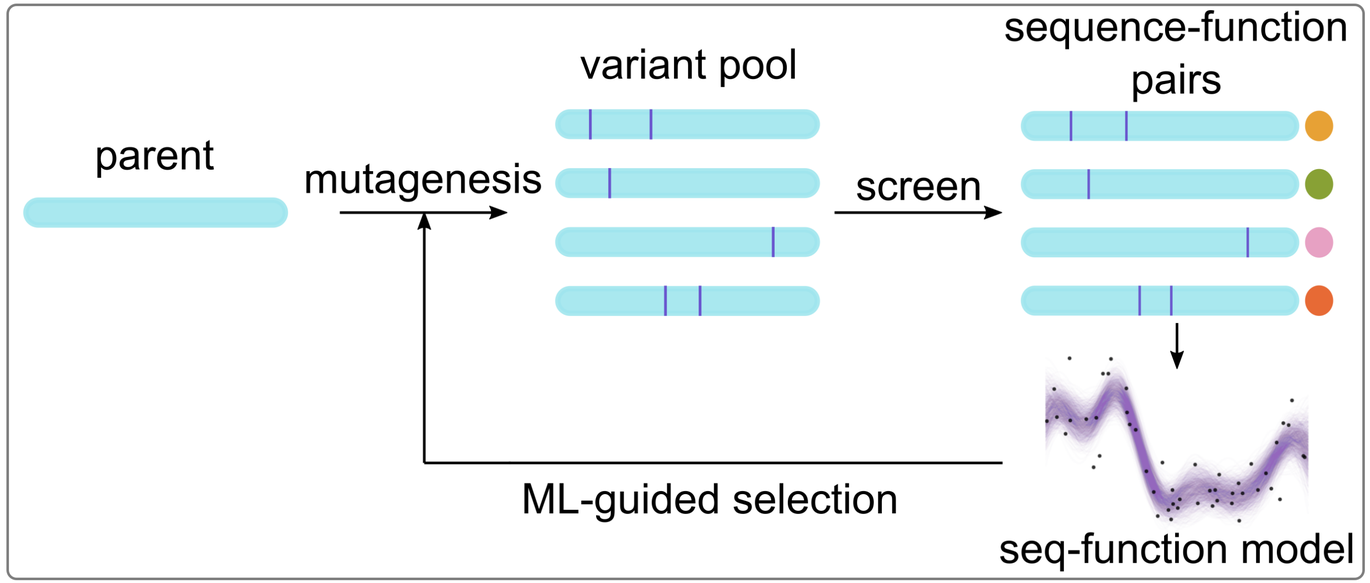


Requirements for directed evolution

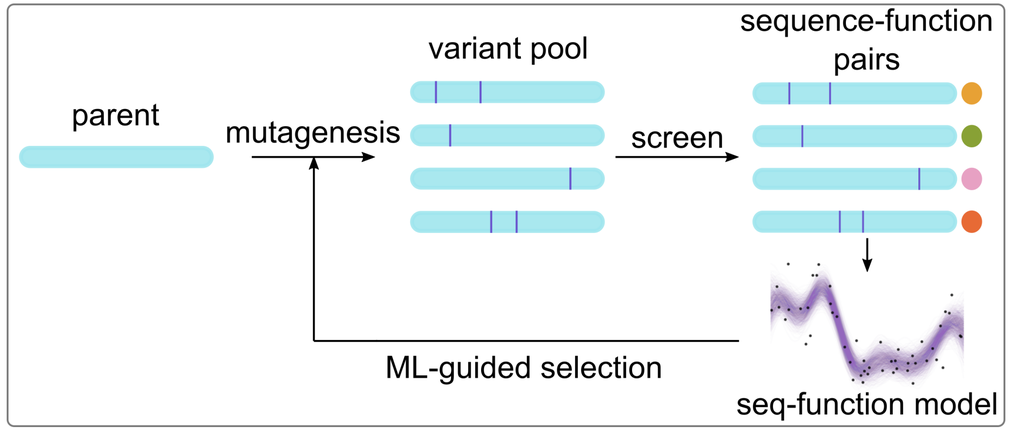


- Parent
- High-throughput screen (>100 / wk)

Machine learning enables optimization with fewer measurements



But still requires starting points

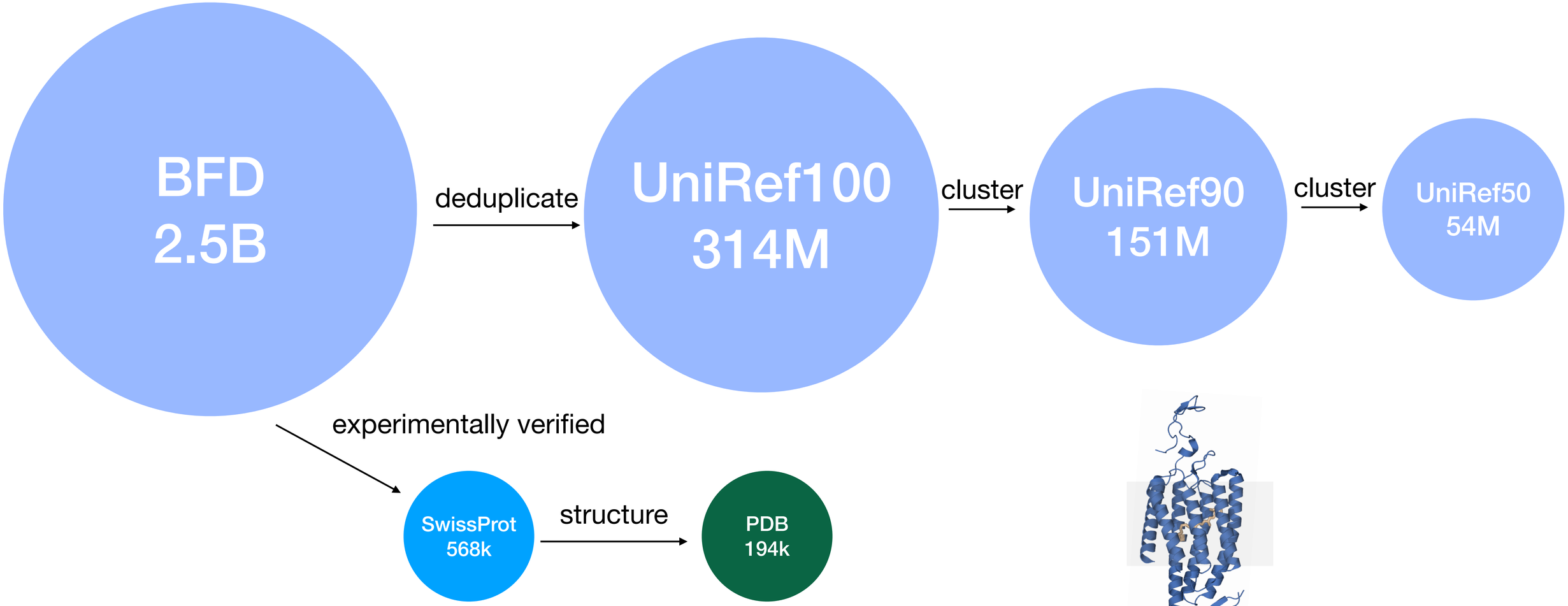


- Accurate models with < 200 measurements
- Find rare, improved sequences with < 200 measurements

Still requires starting point!

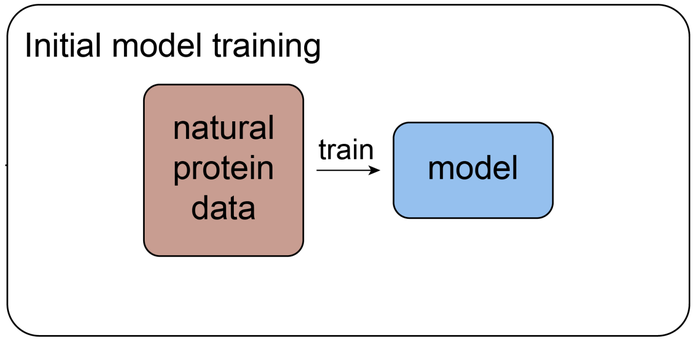
Ignores mountains of protein data

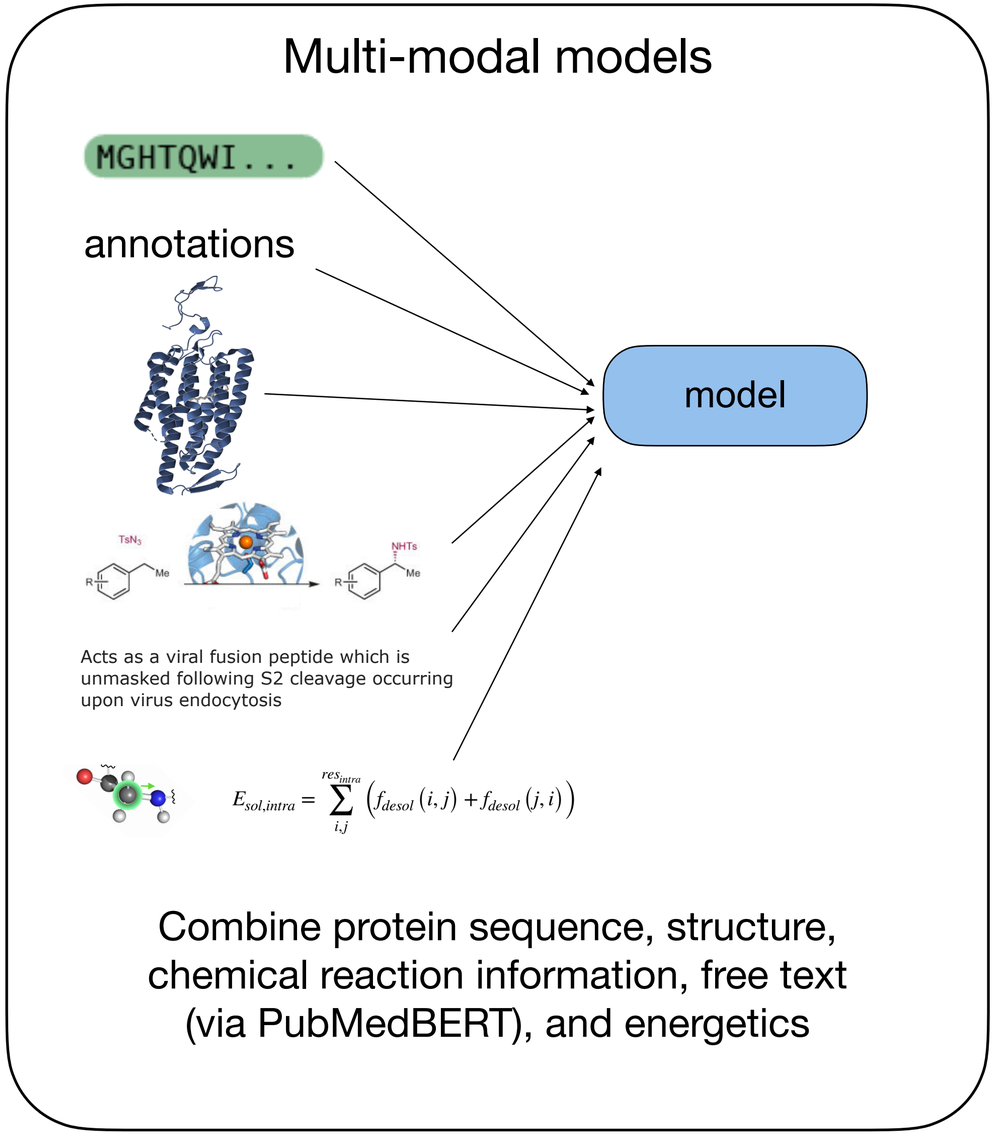
We have access to large protein databases



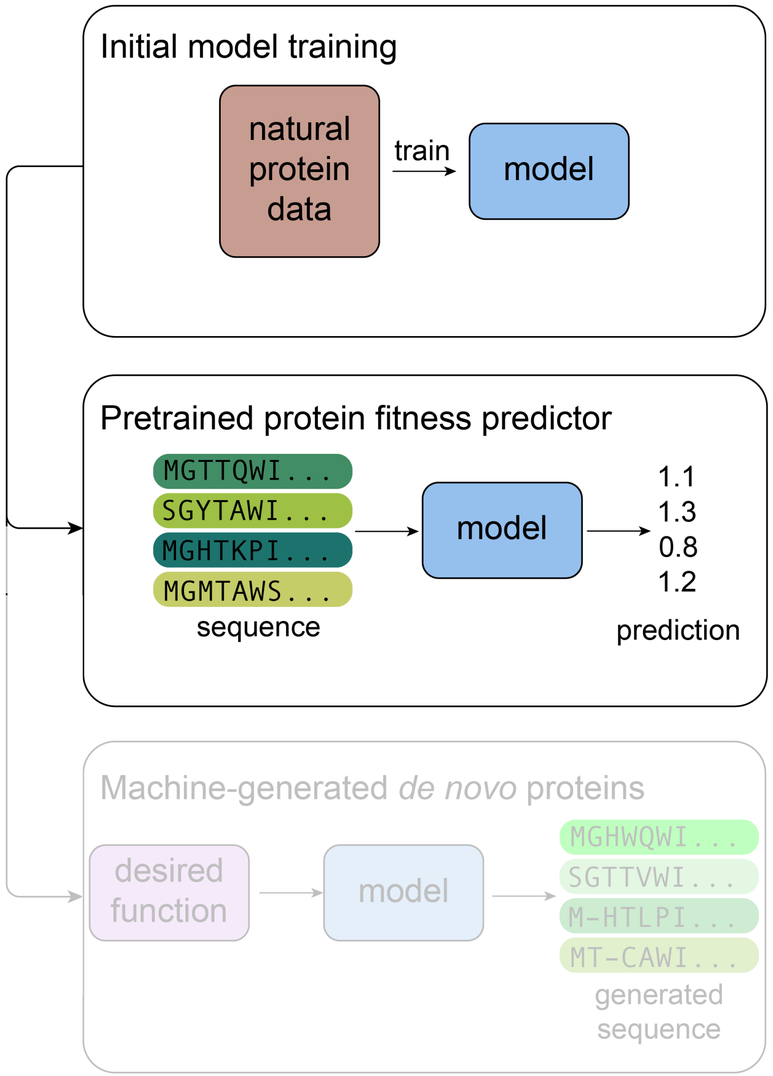
200 measurements

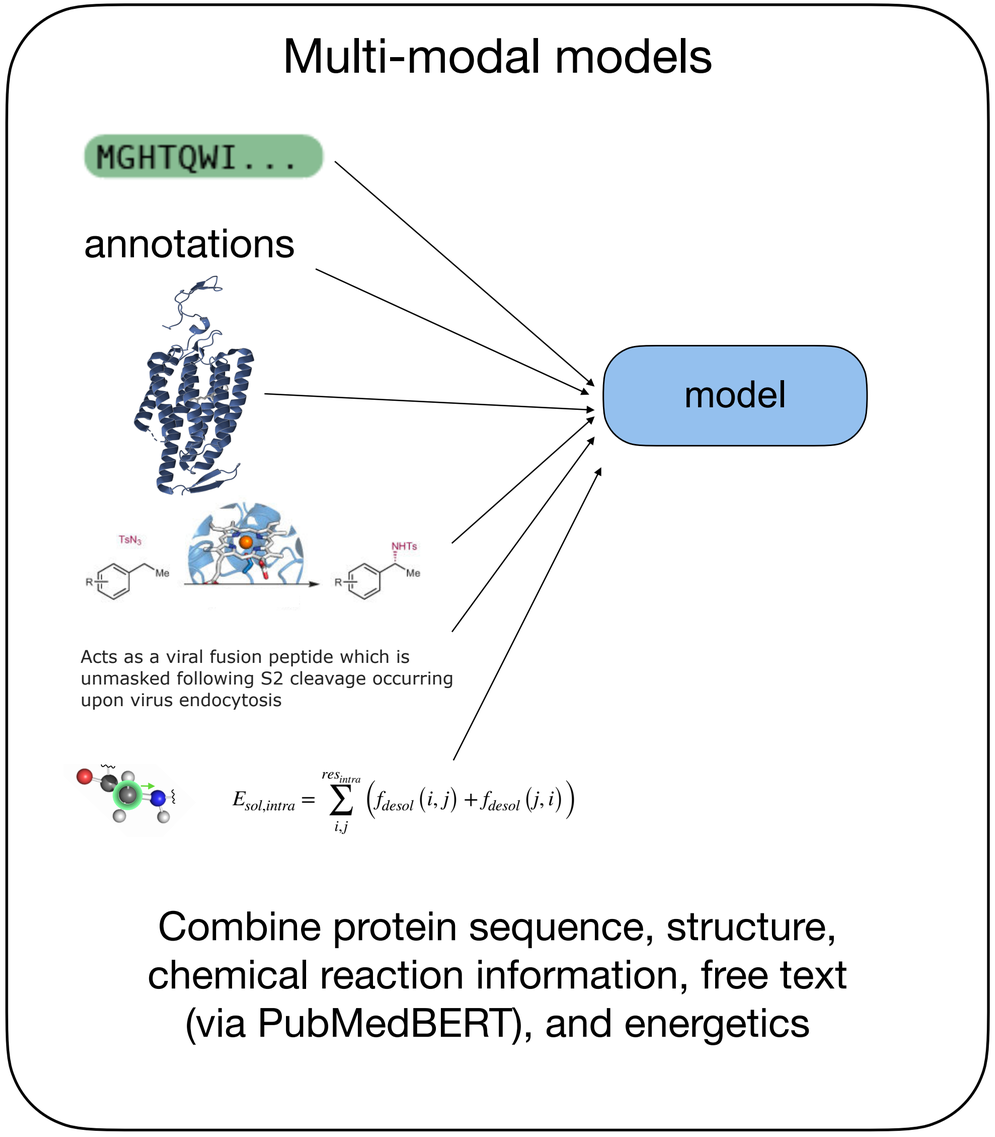
Use multiple data modalities to design proteins





Use multiple data modalities to design proteins





Many methods pretend proteins are language

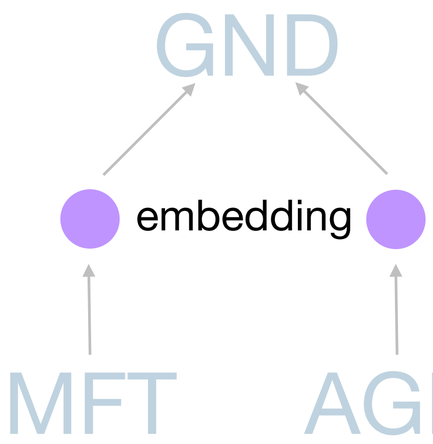
MFTGNDAGH

Treat amino acids like tokens

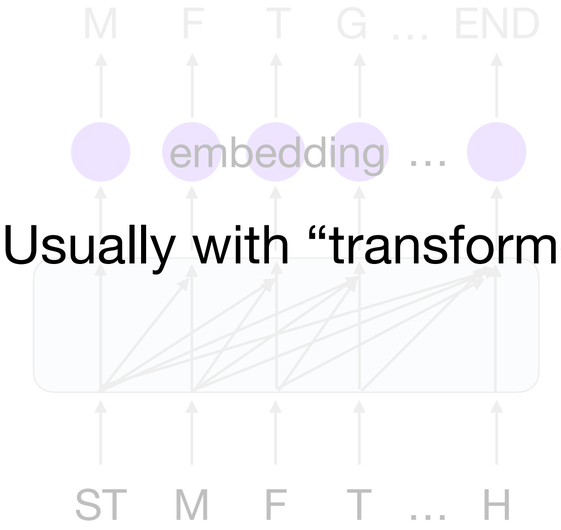
Use tools originally developed for language

Many methods pretend proteins are language

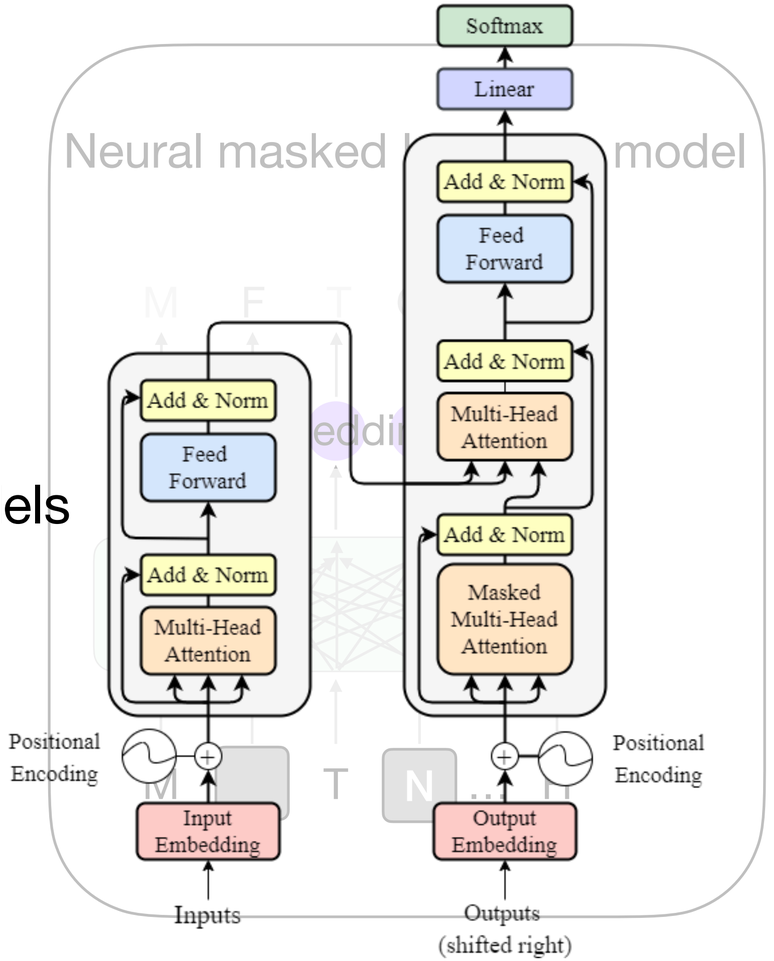
Word2Vec



Neural autoregressive language model



Usually with “transformer” models

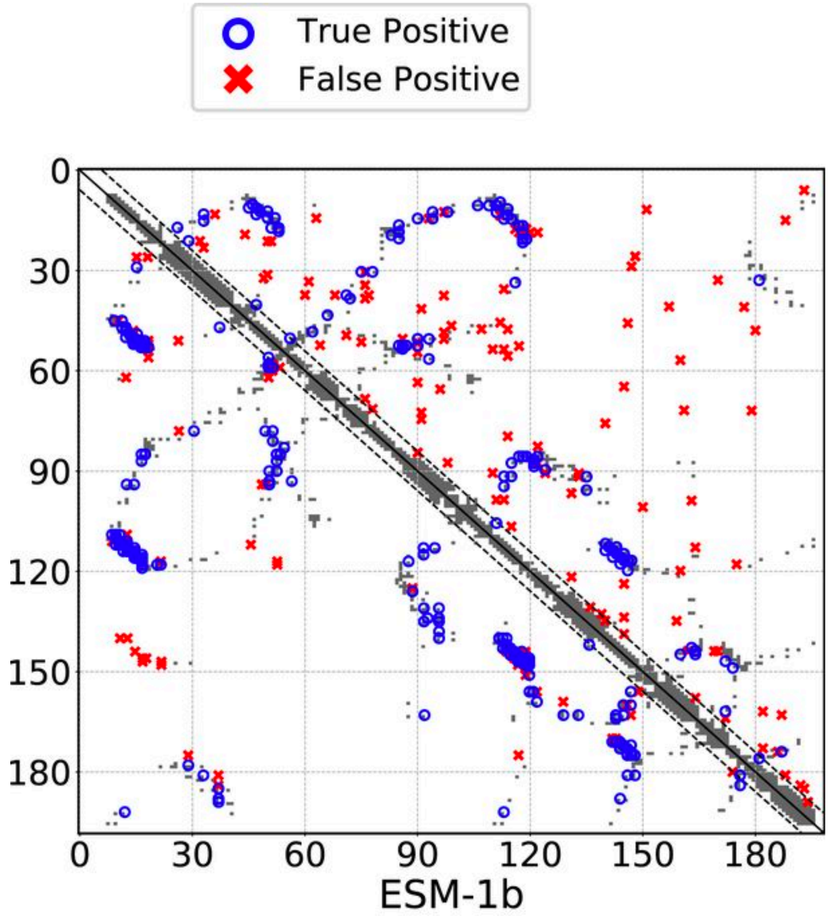


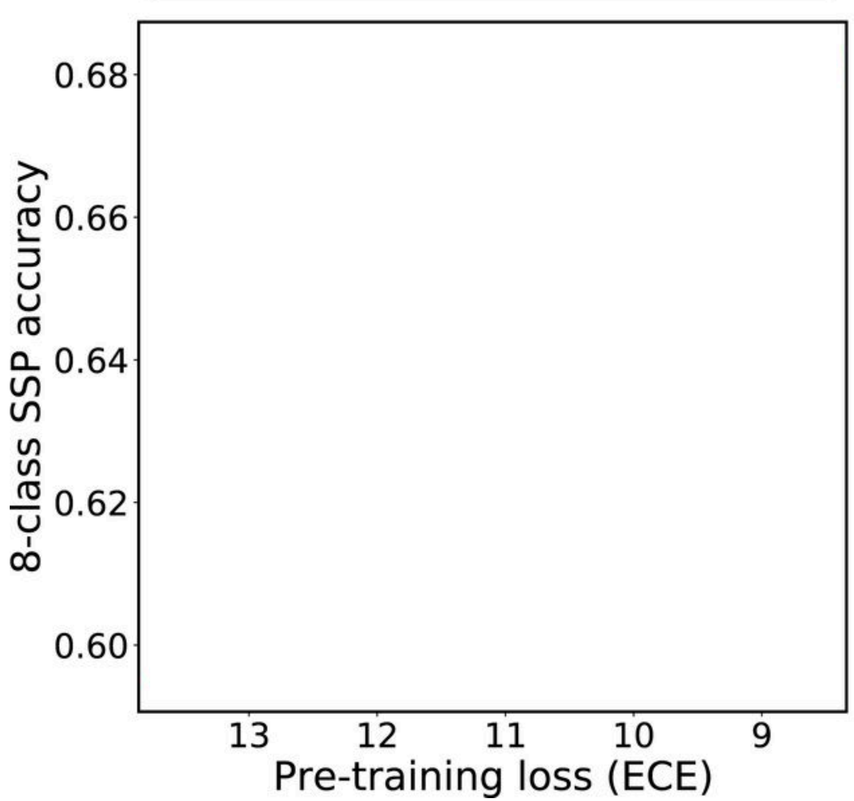
MFTGNDAGH

Pretrained transformers recapitulate biophysical properties

Rives *et al. PNAS* 2021

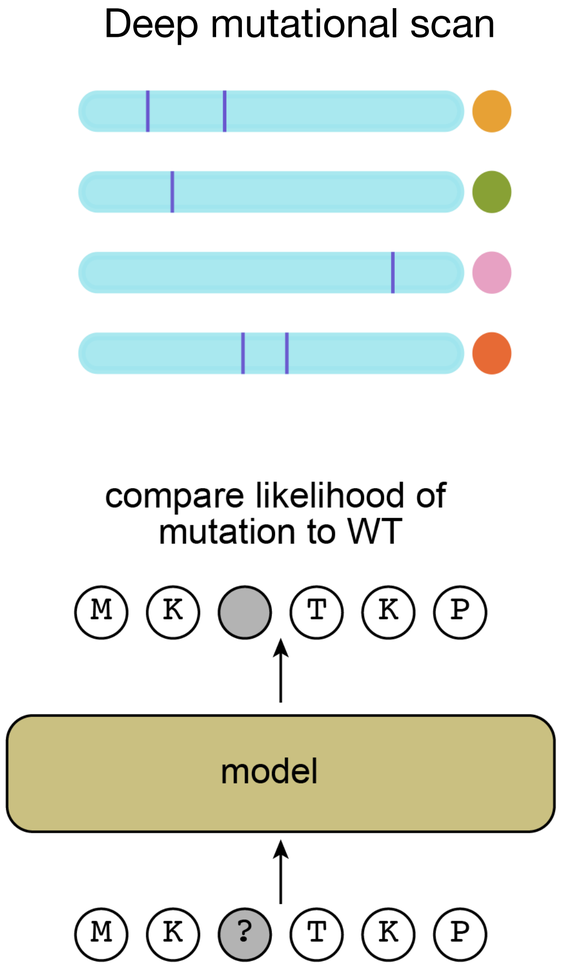
Pretrained transformers contain structural information

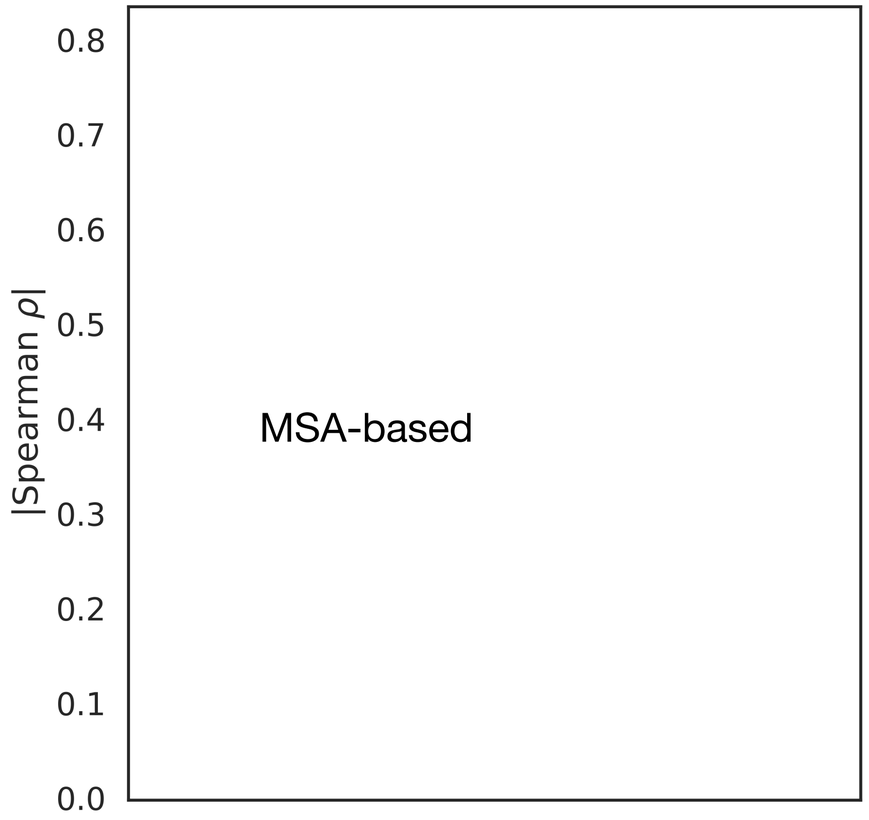




Rives *et al. PNAS* 2021

Pretrained transformers are zero-shot fitness predictors

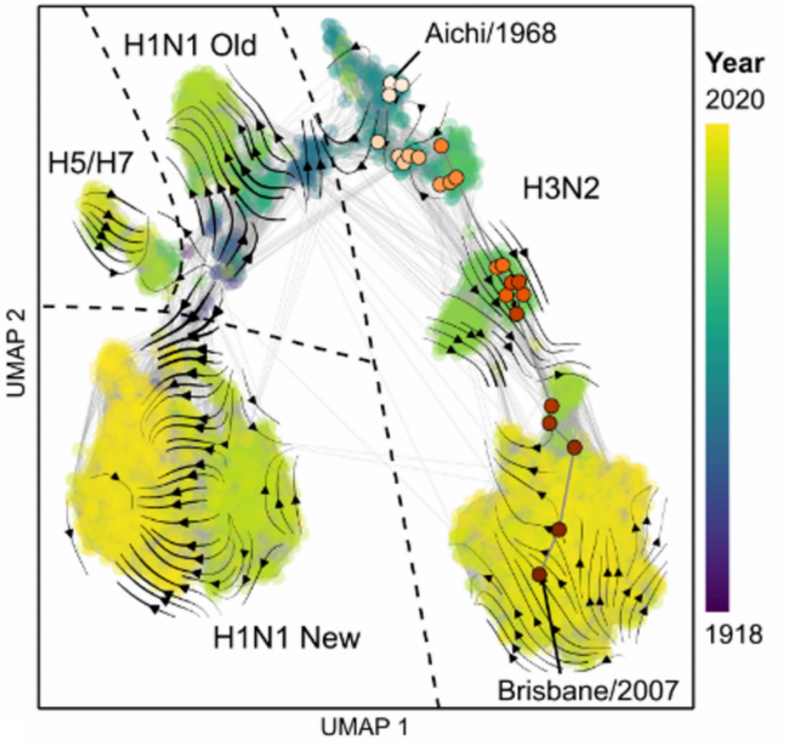




Meier *et al. NeurIPS* 2021

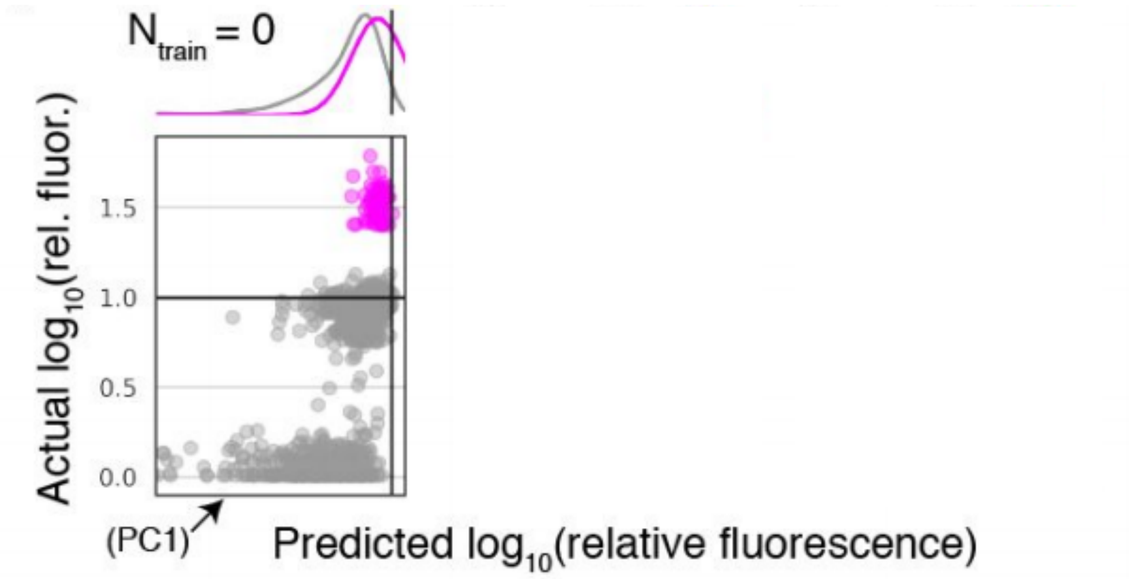
Pretrained transformers reconstruct evolutionary trajectories

Influenza A nucleoprotein



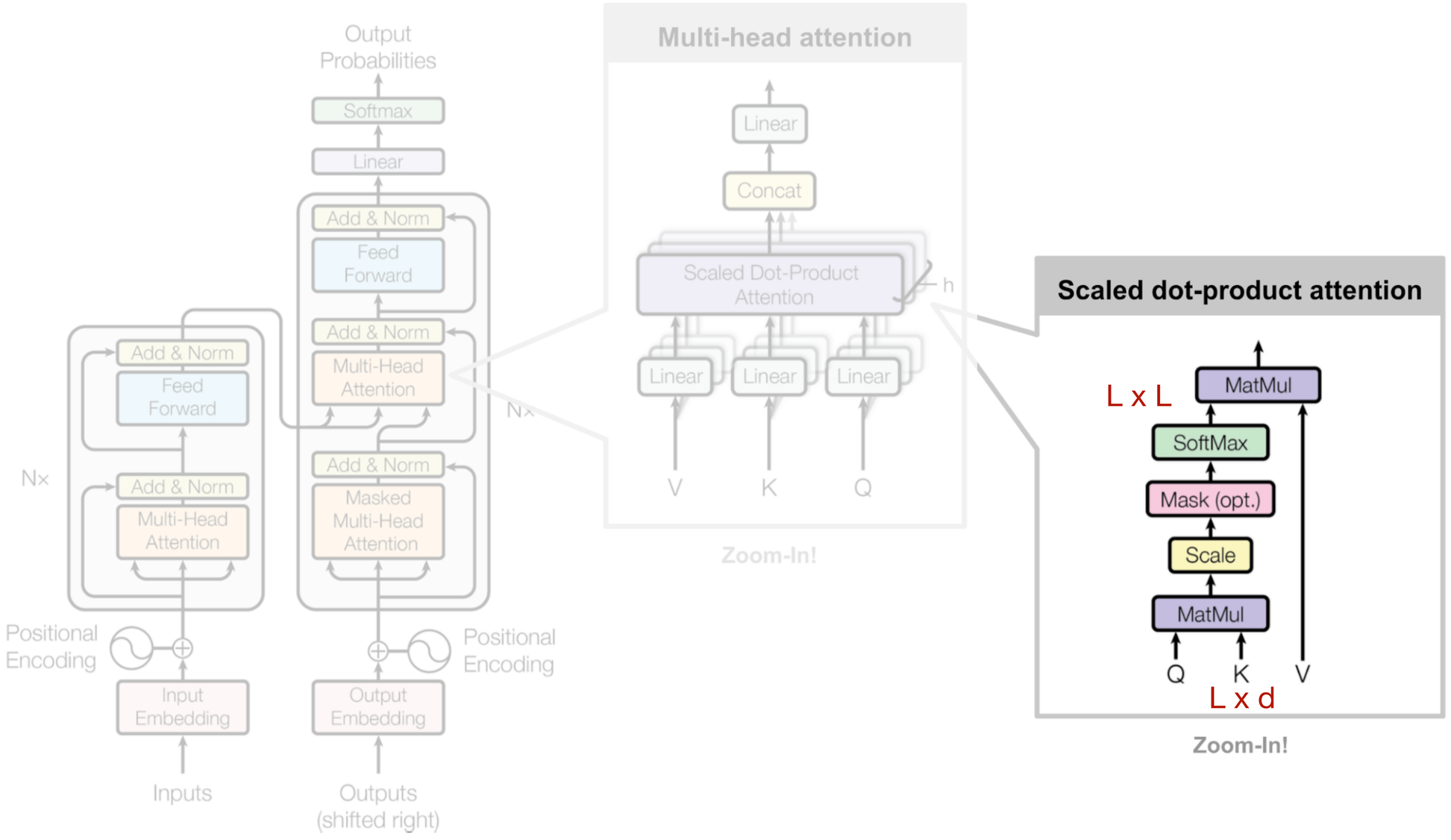
Hie *et al. Cell Systems* 2022

Pretraining guides search away from loss-of-function sequences

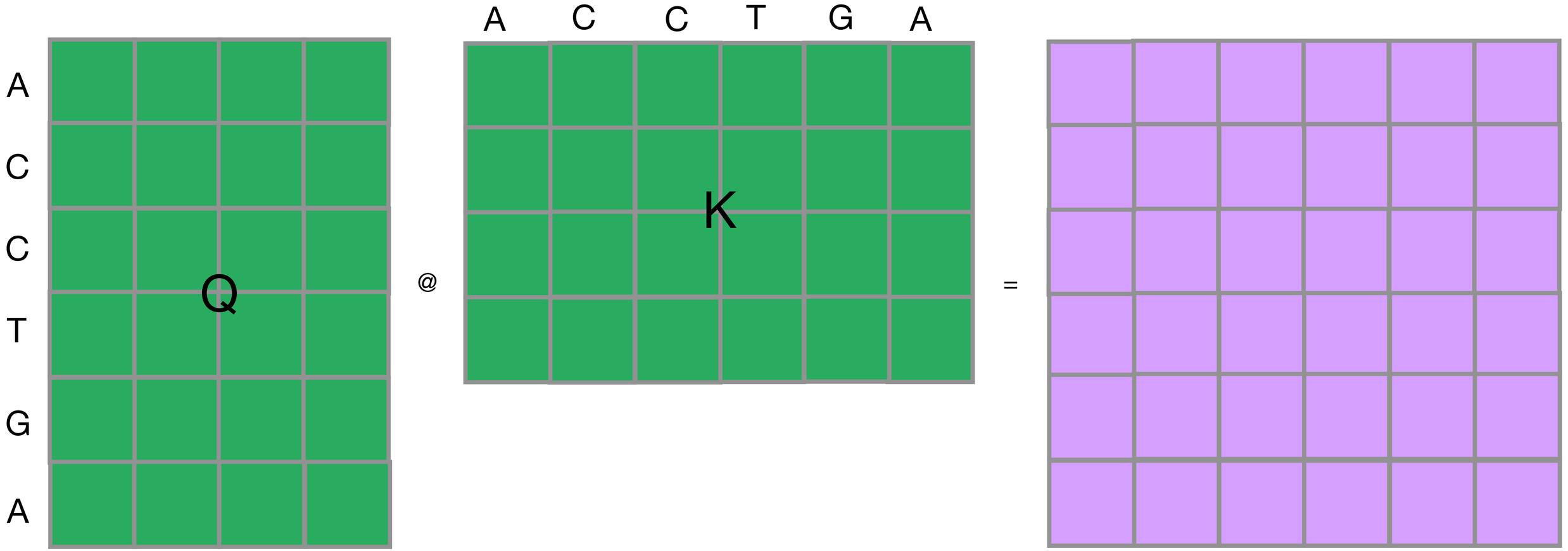


Biswas *et al. Nature Methods* 2021

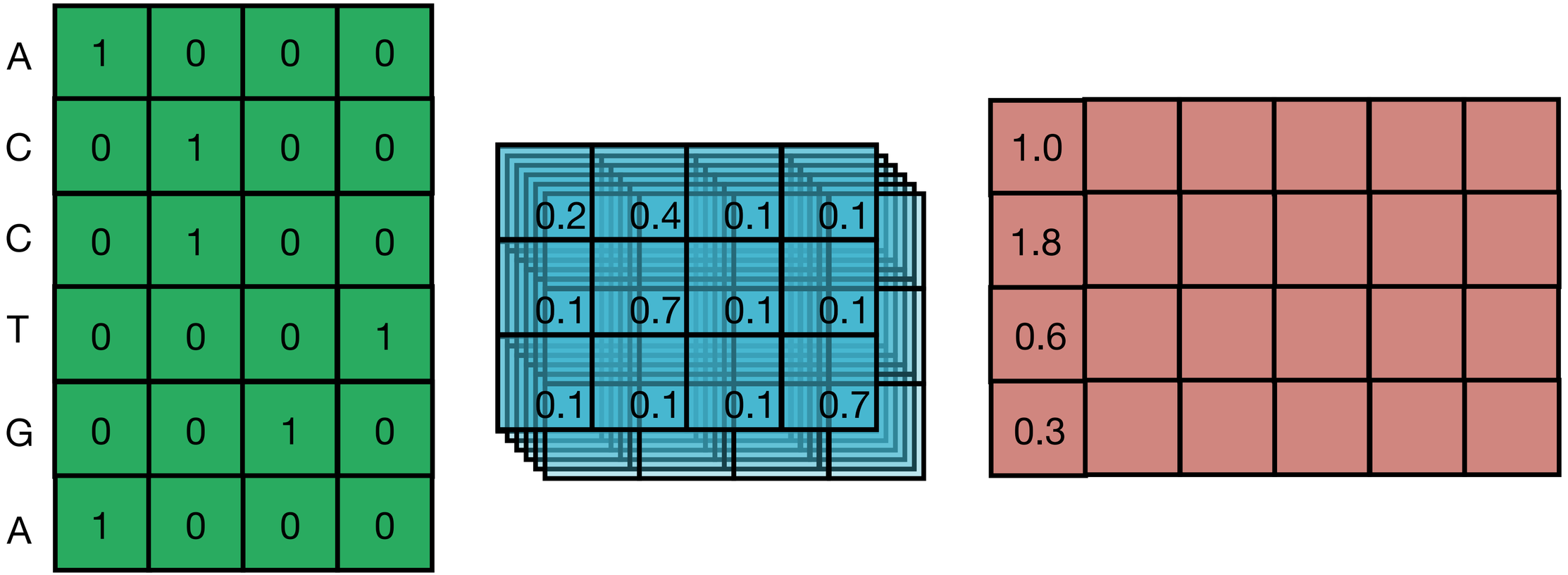
Transformers scale quadratically with length



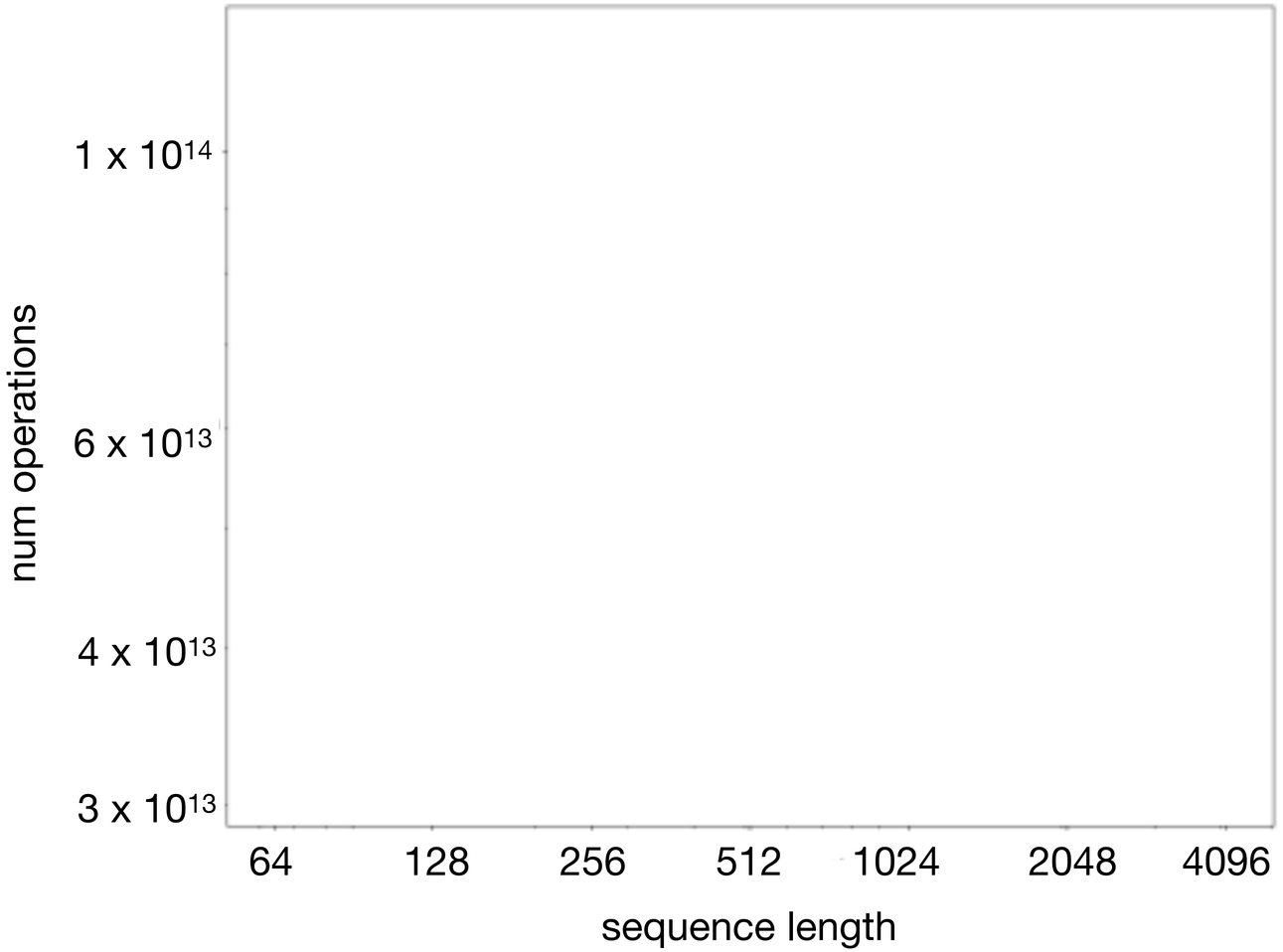
Transformers scale quadratically with length



Convolution scales linearly with length

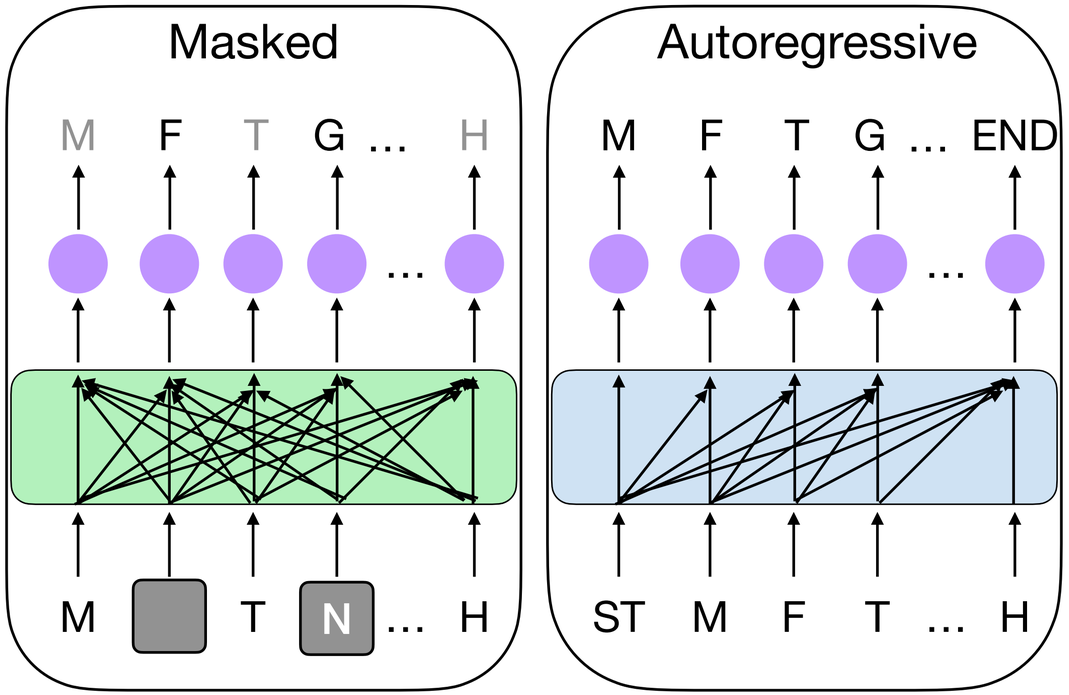


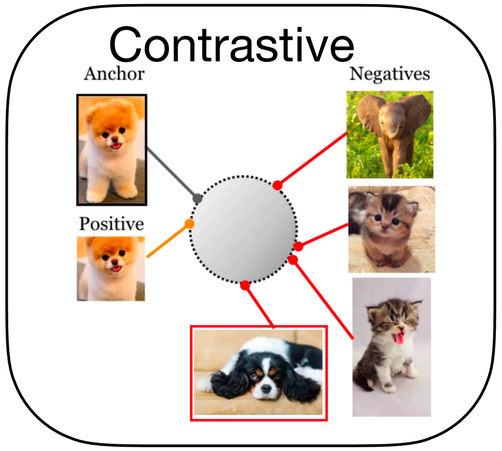
Convolution scales linearly with length



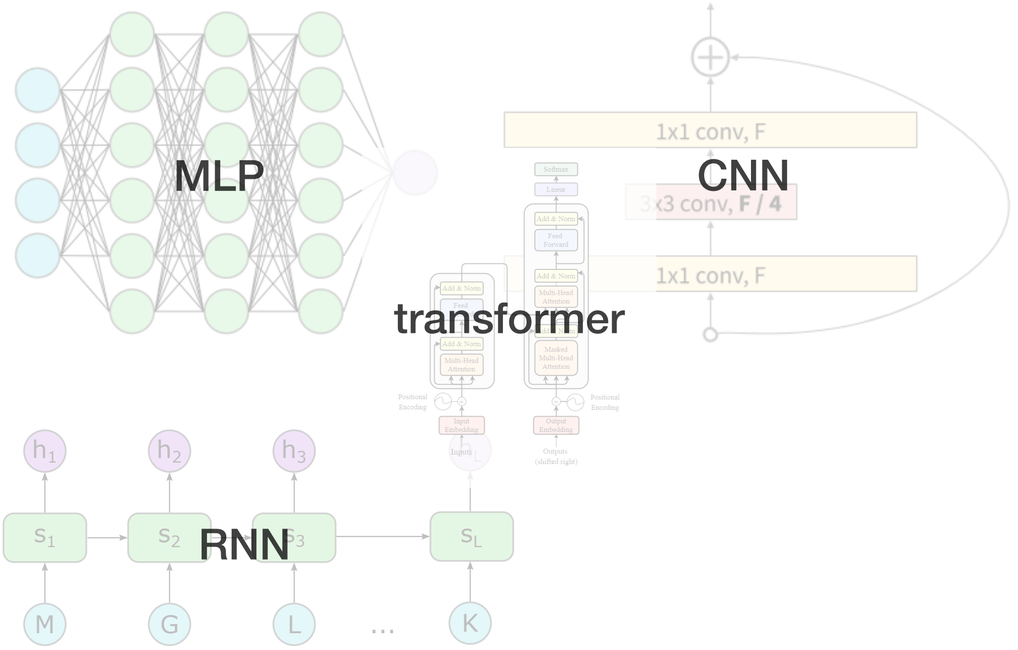
Tay *et al.* 2022

Separate pretraining task

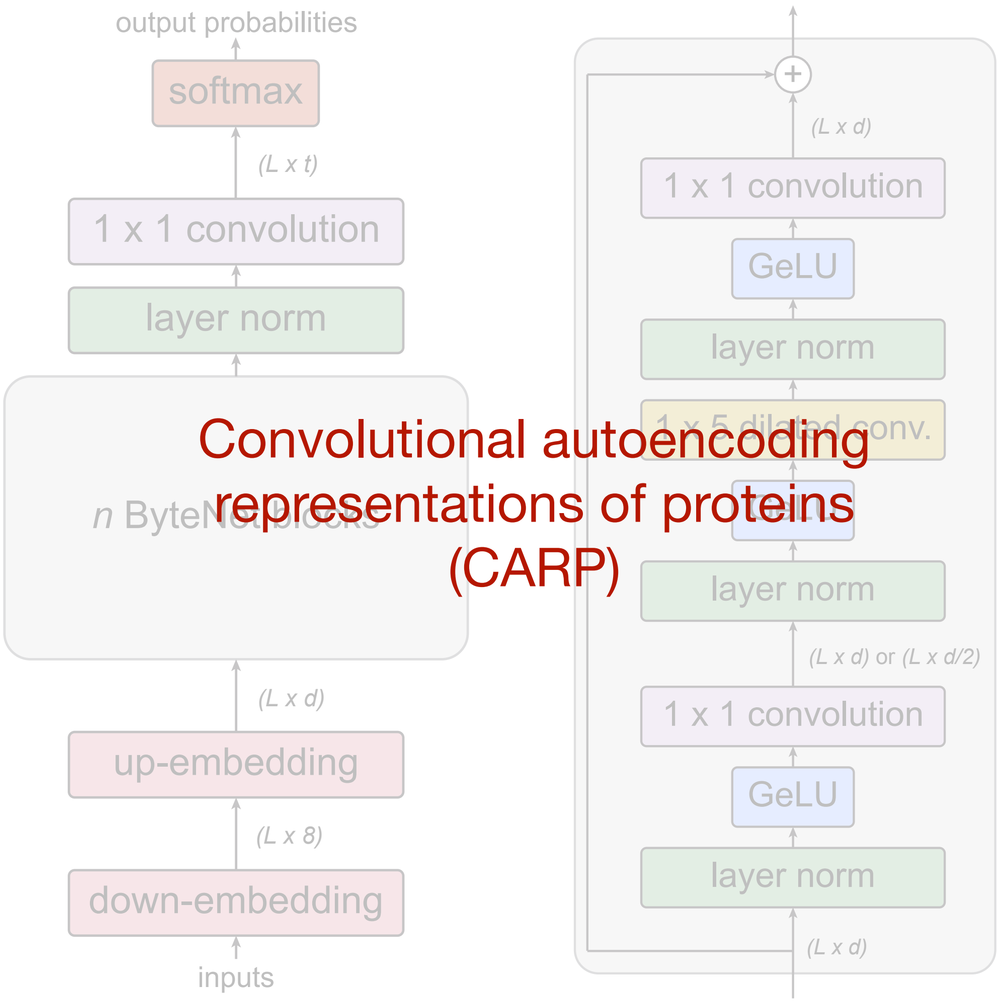




?

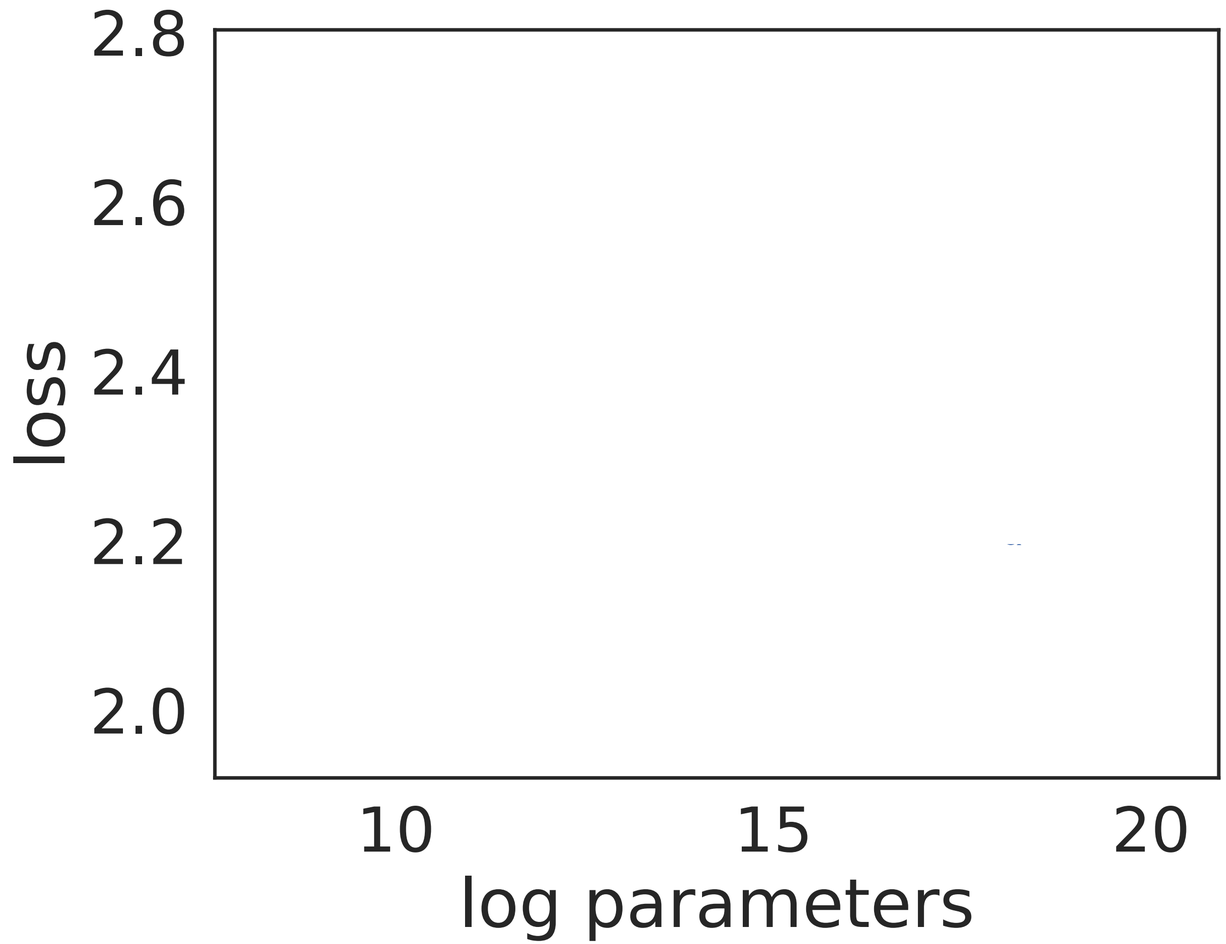


We pretrain CNNs to reconstruct sequences



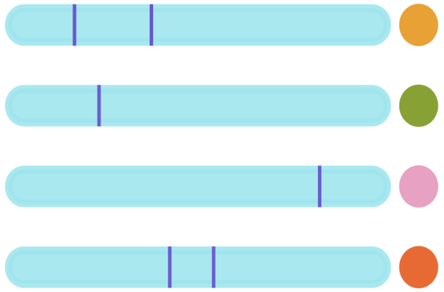
Convolutional autoencoding representations of proteins (CARP)

CNNs are competitive with transformers for pretraining

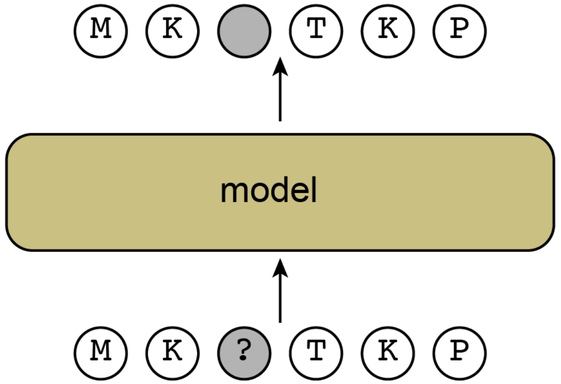


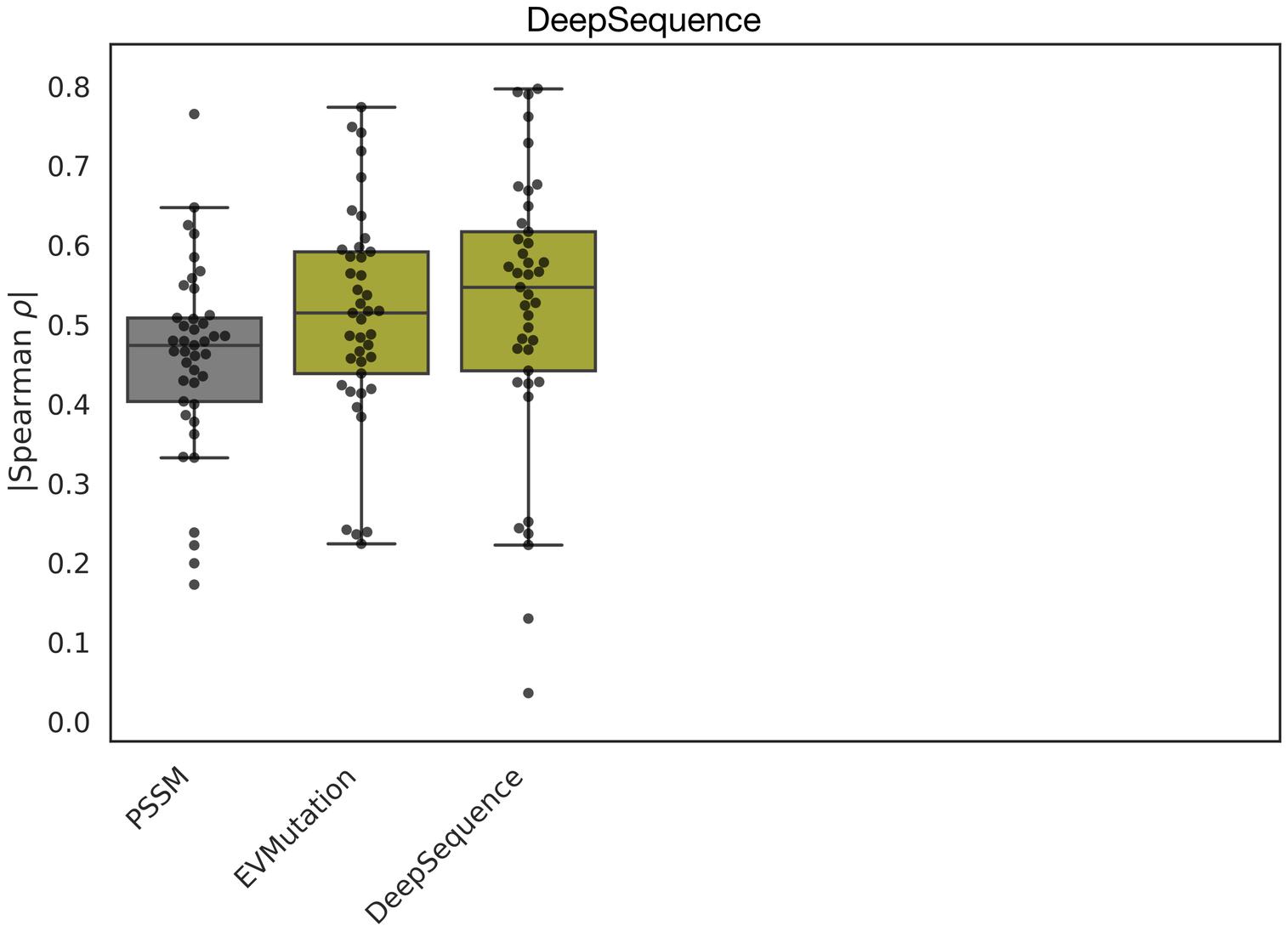
CARP is a zero-shot fitness predictor

Deep mutational scan

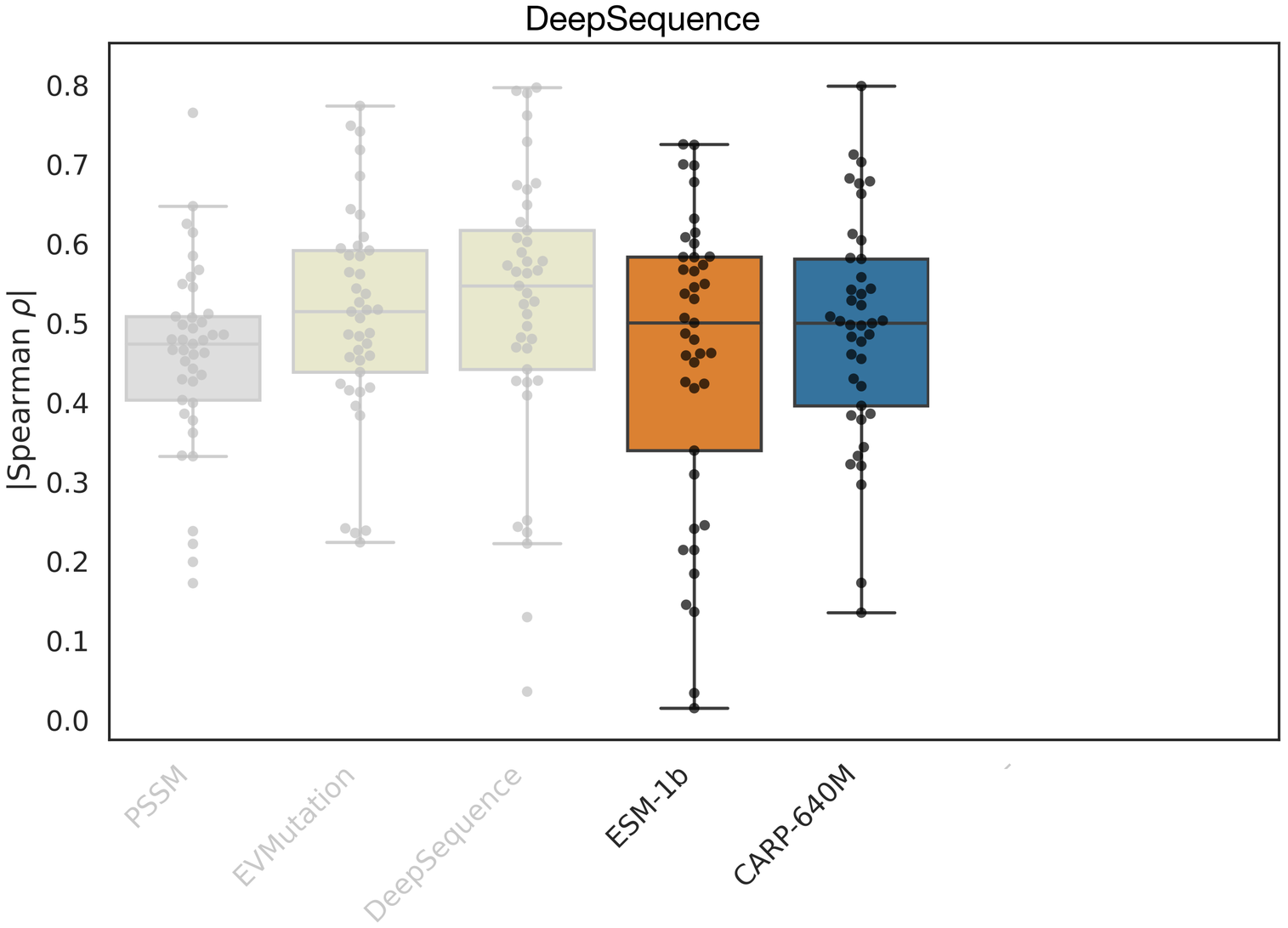


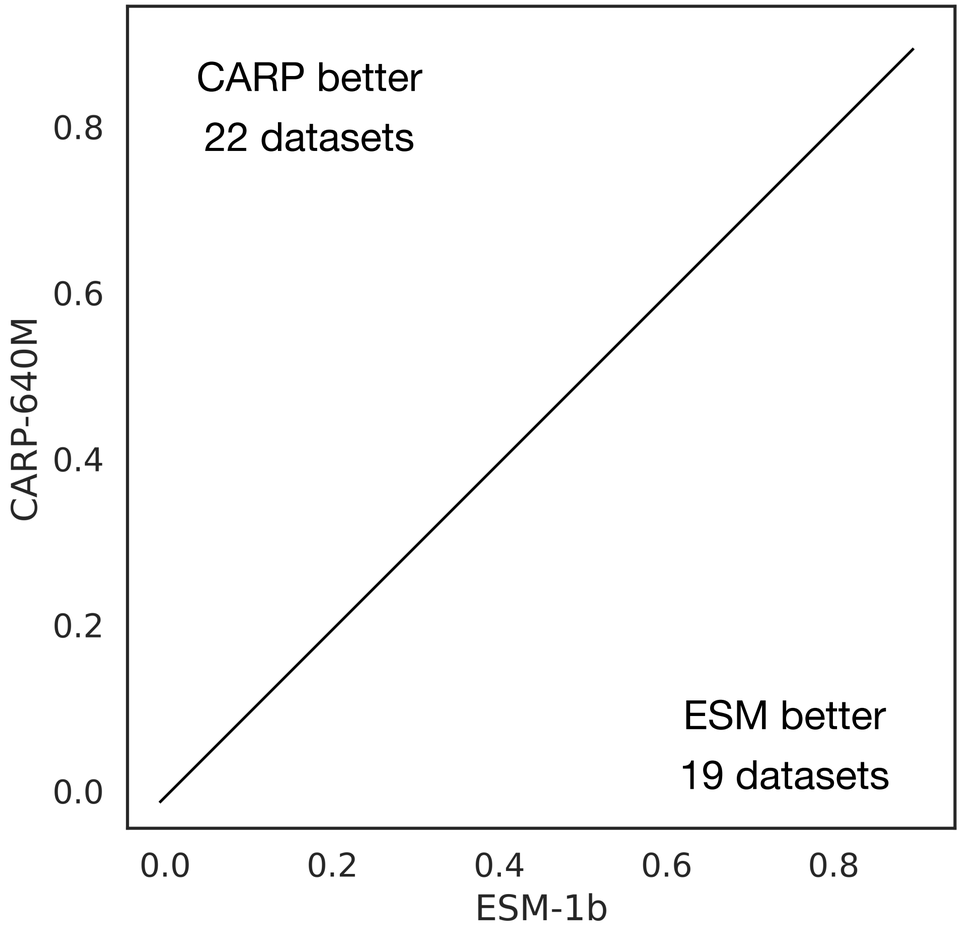
compare likelihood of mutation to WT





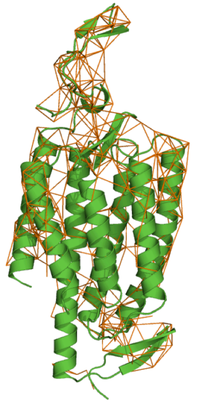
CARP is a zero-shot fitness predictor





CARP learns structure

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



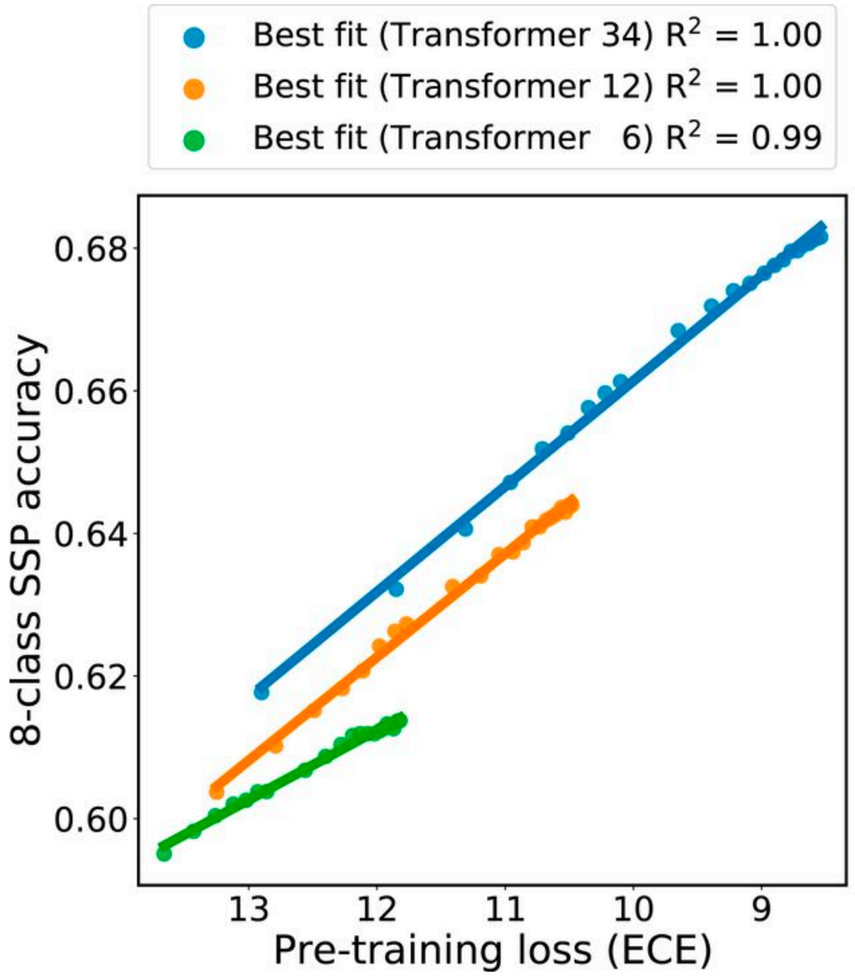
long-range contacts

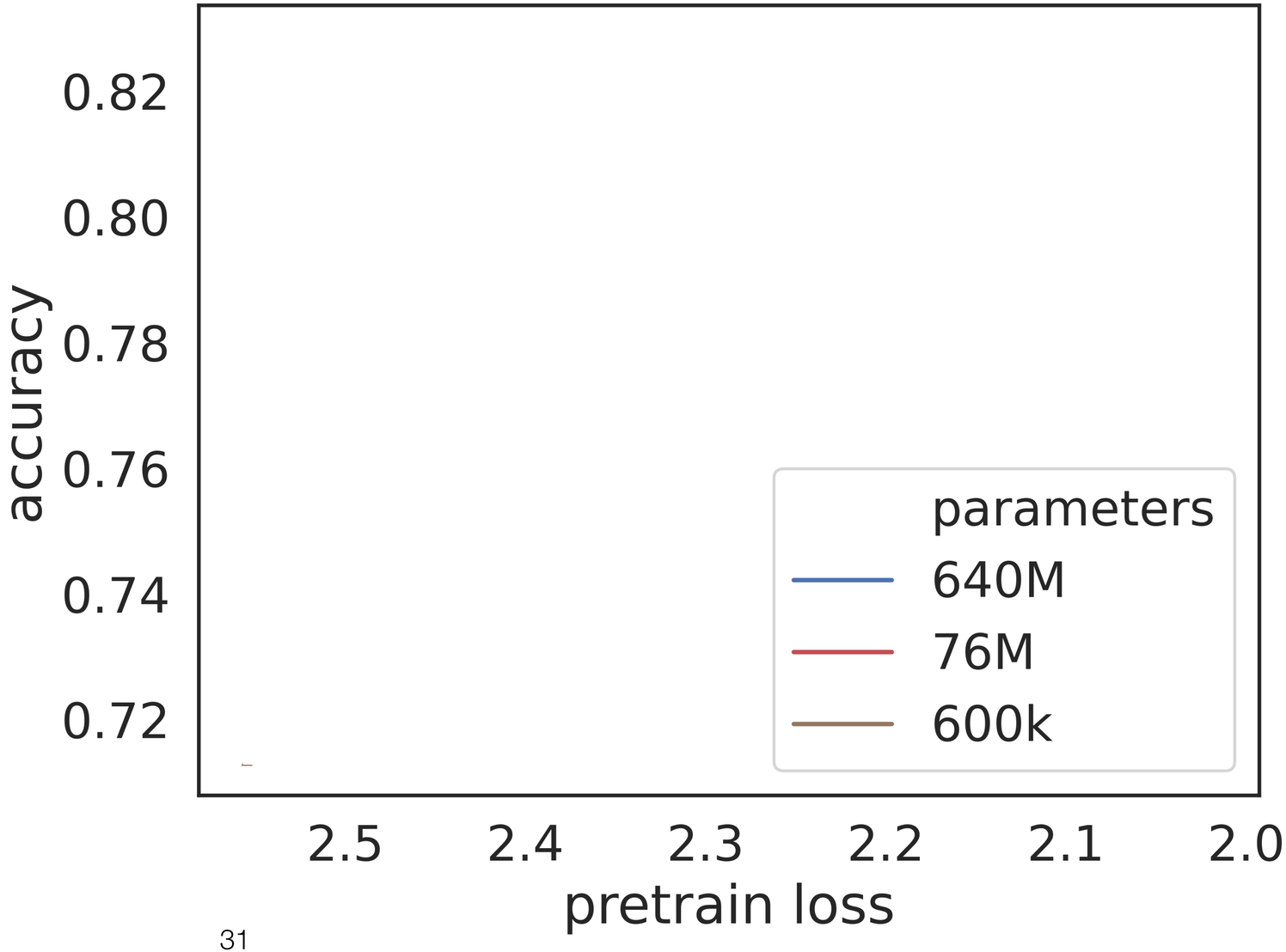
β-Sheet (3 strands)

α-helix

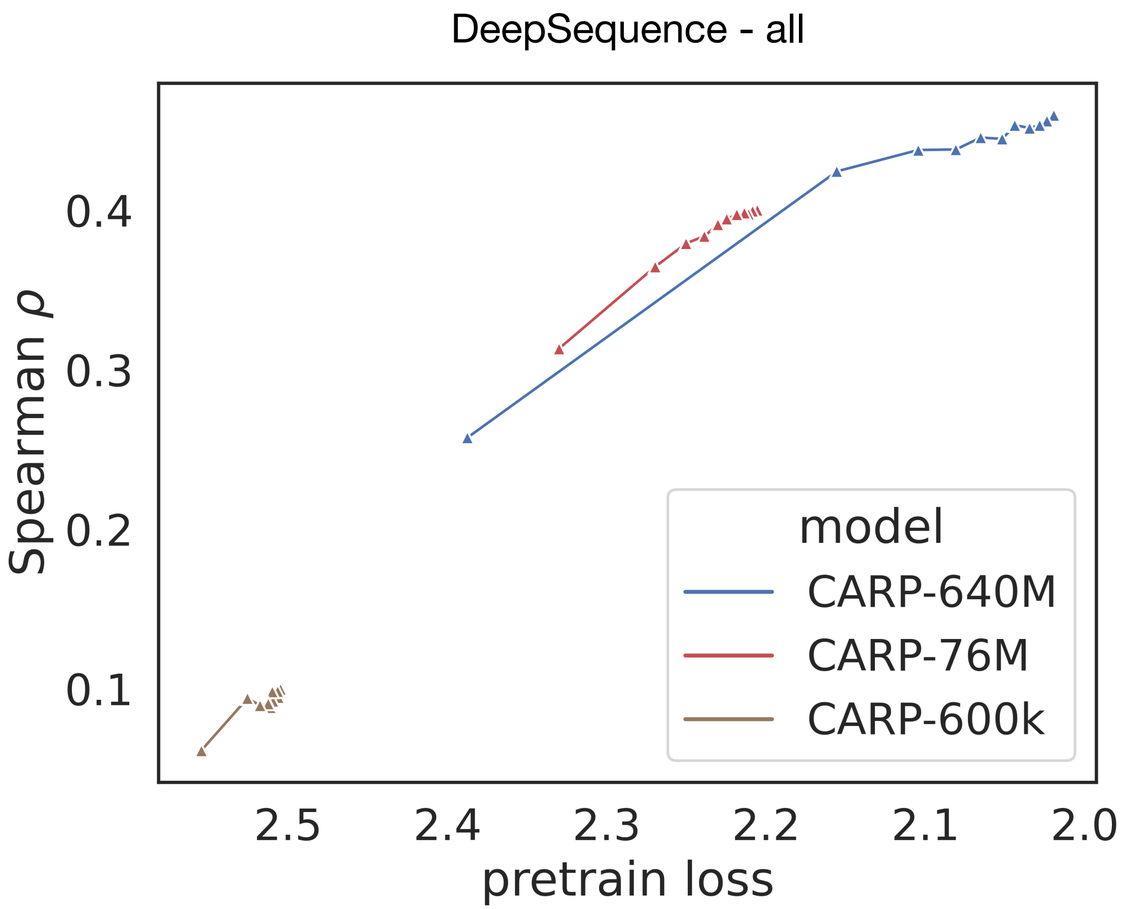
Thomas Shafee

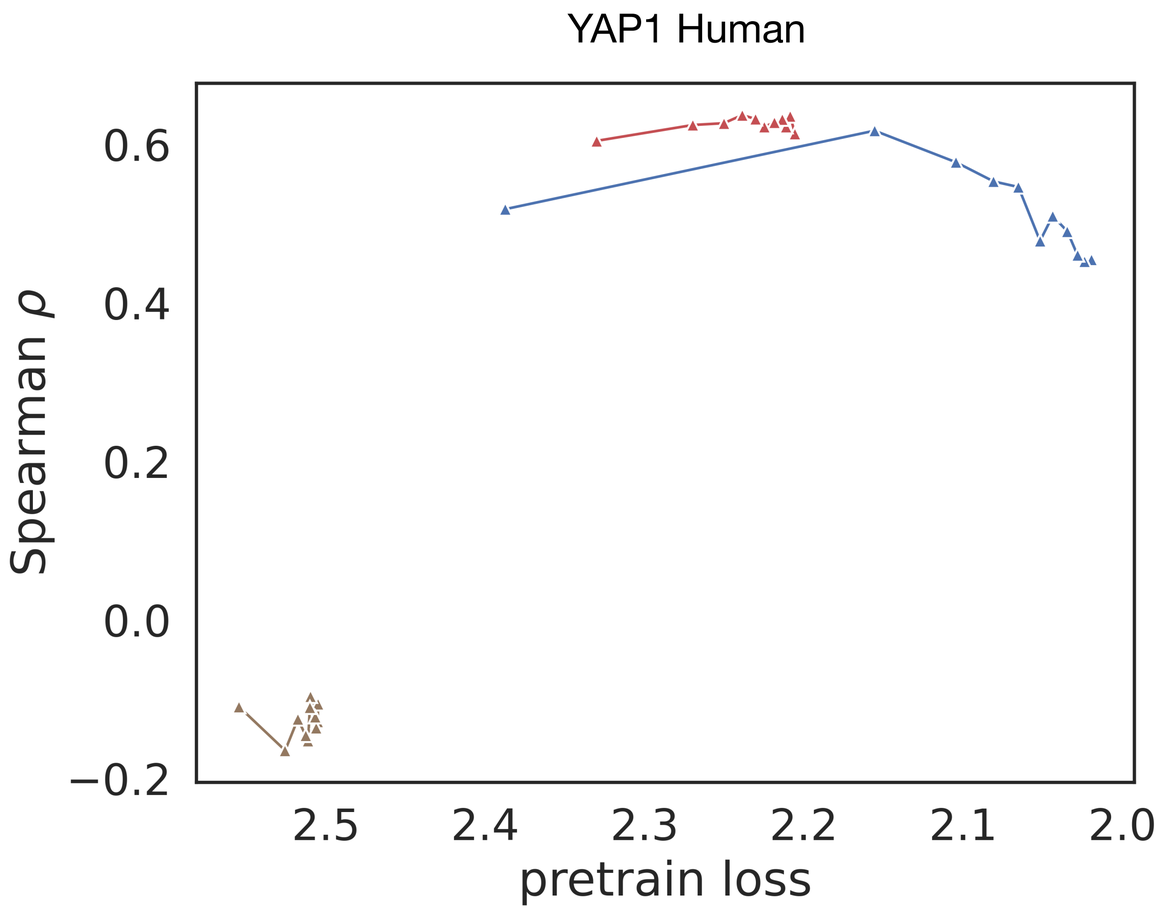
Structure predictions improve smoothly with pretraining





CARP zero-shot performance mostly improves with pretraining



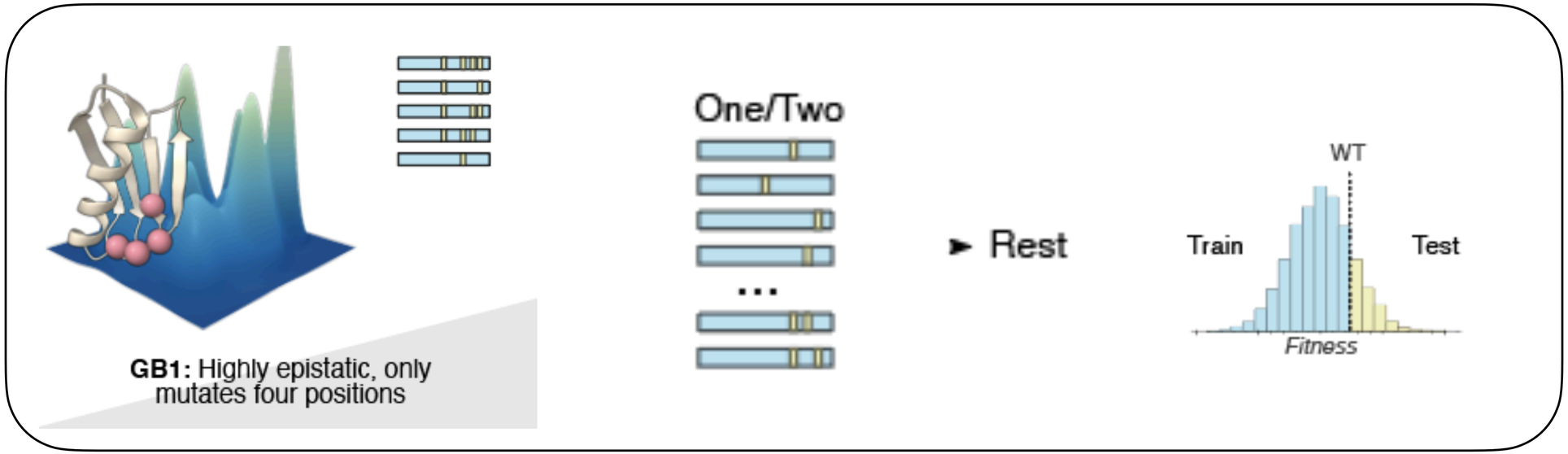


Protein engineering campaigns require out-of-domain generalization

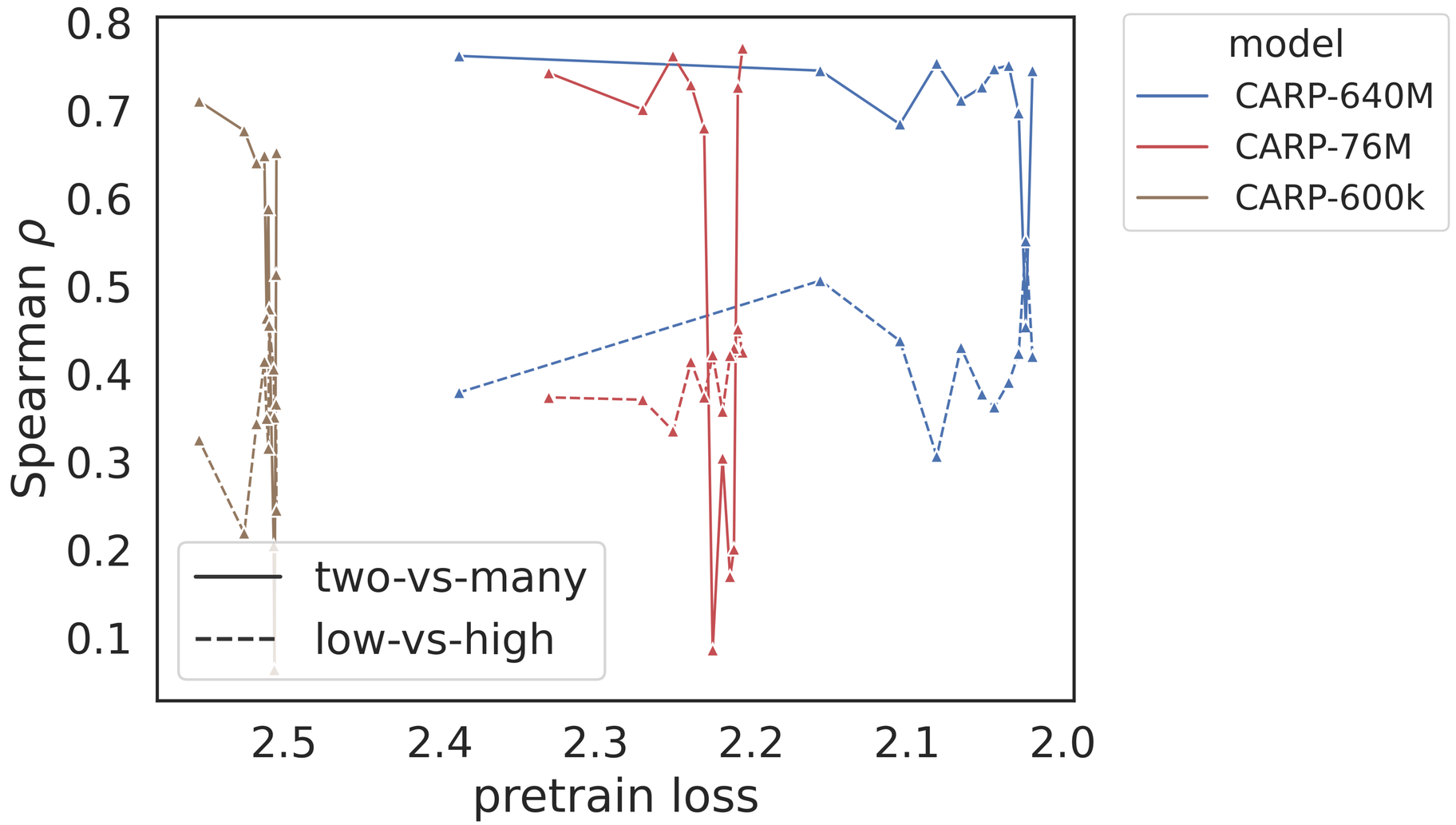
Can usually measure all single mutants

Naively screen bigger spaces
->
mostly non-functional

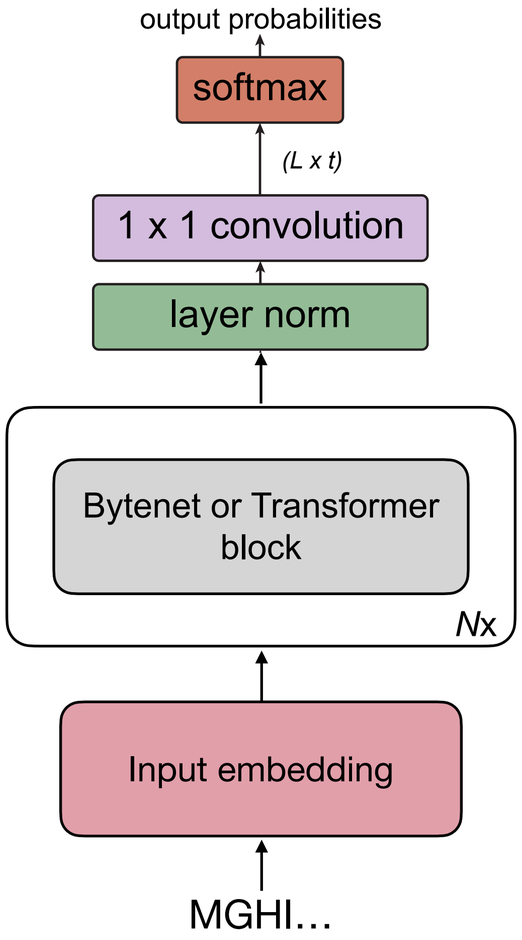
Protein engineering campaigns require out-of-domain generalization



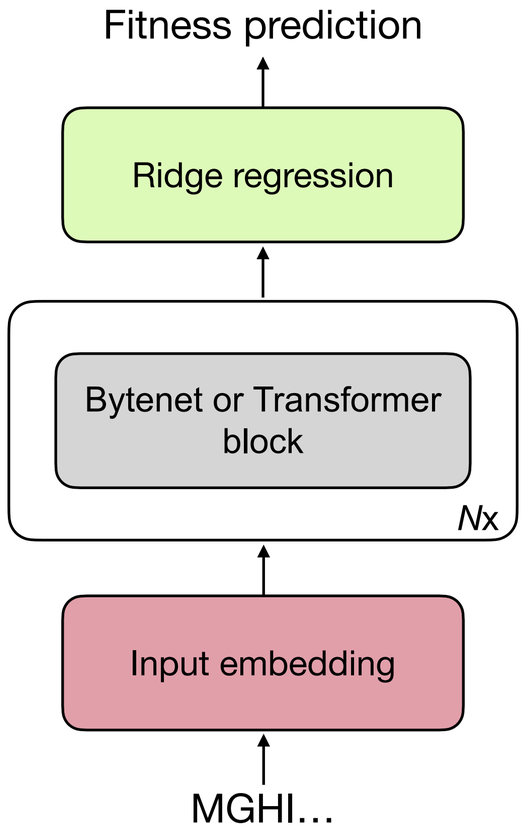
OOD performance does not improve with more pretraining



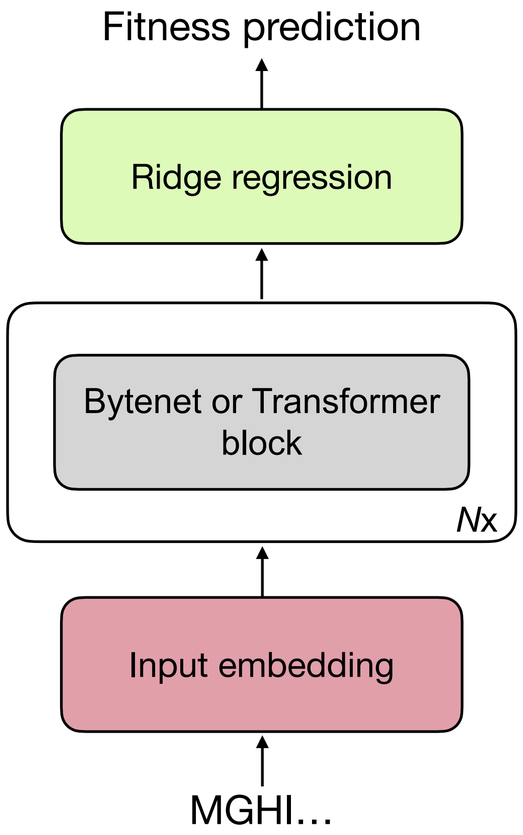
Not all layers are necessary for transfer



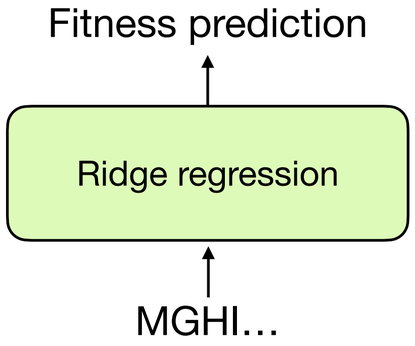
Not all layers are necessary for transfer

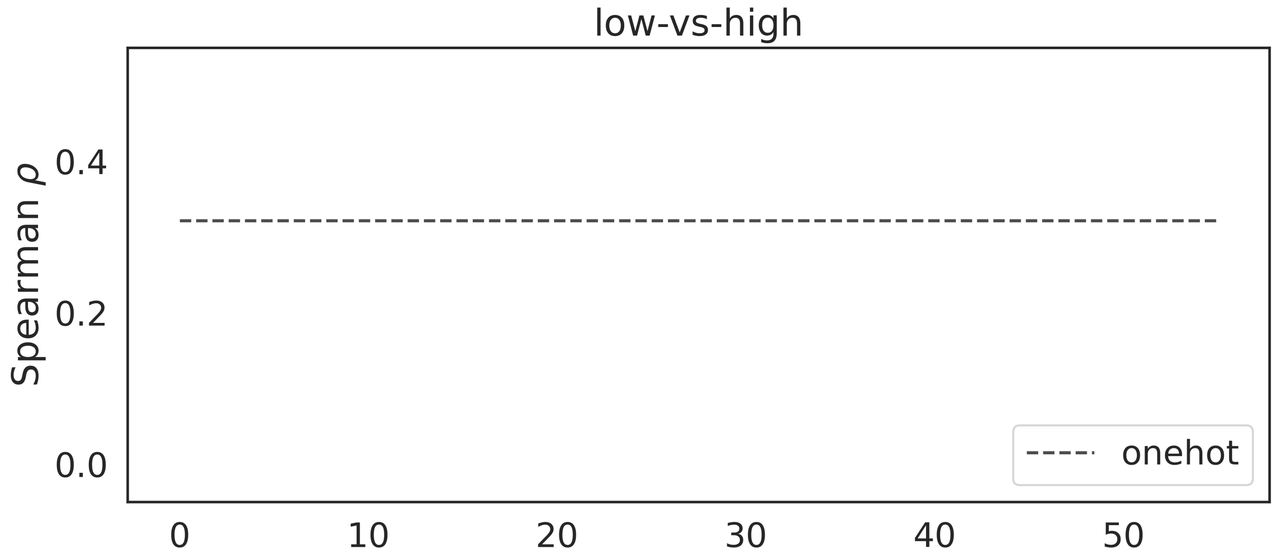


Not all layers are necessary for transfer

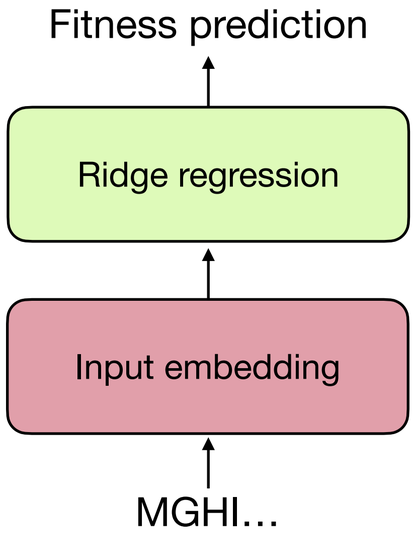


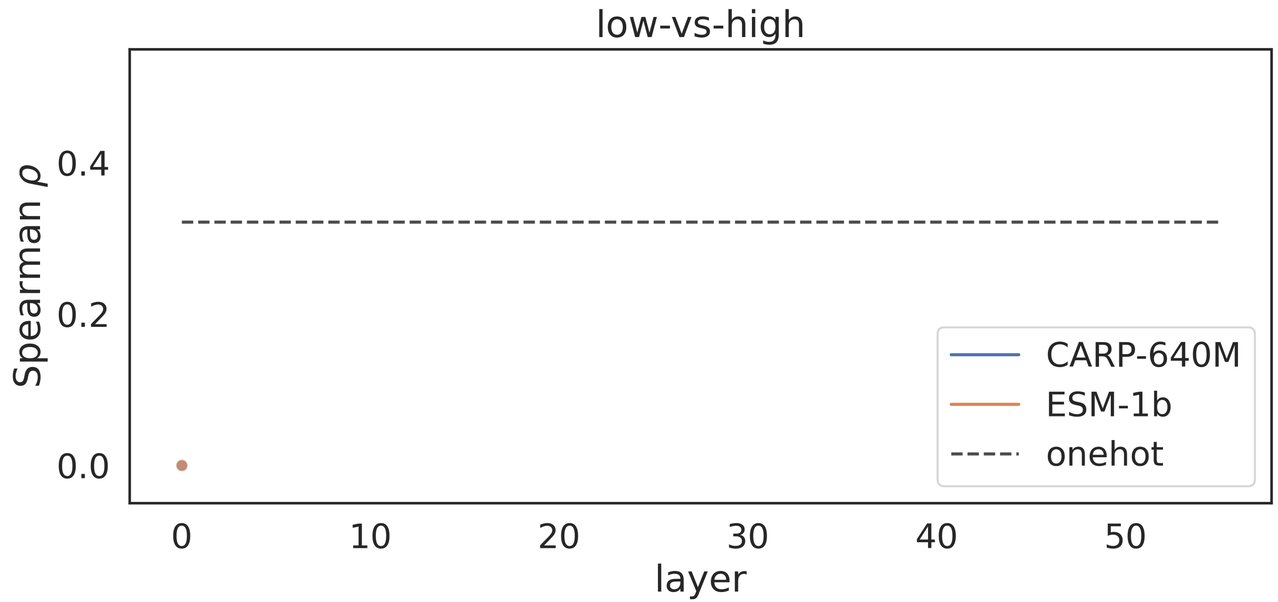
Not all layers are necessary for transfer



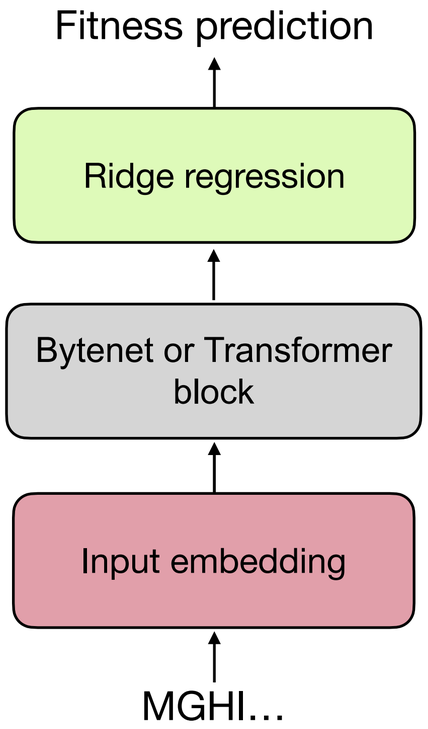


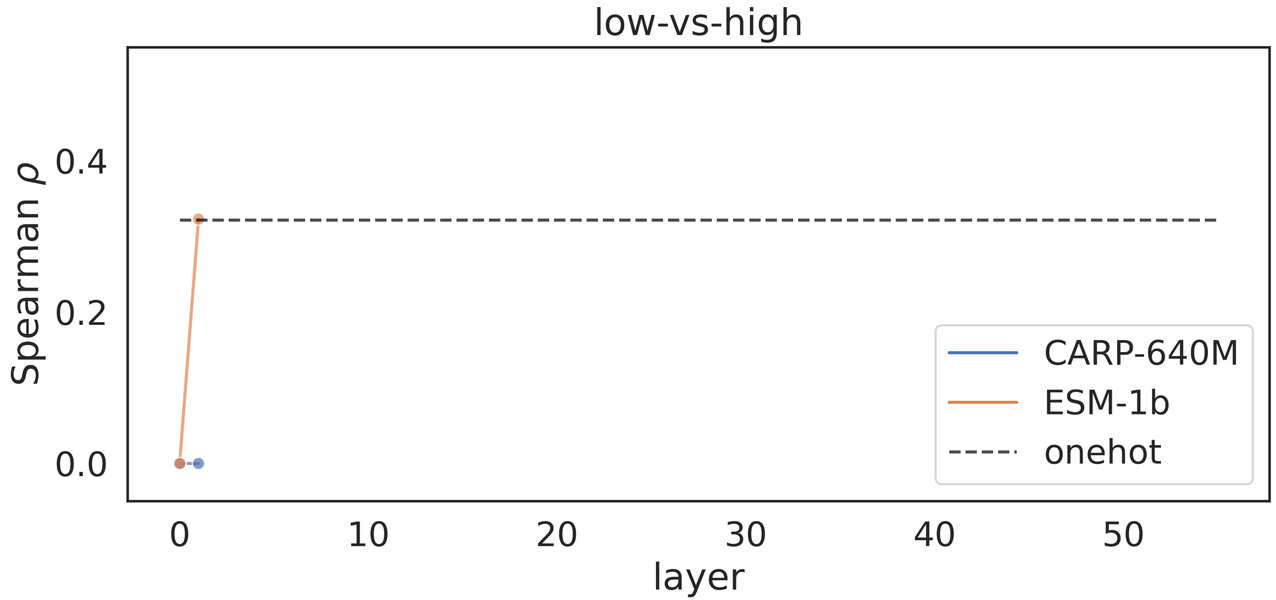
Not all layers are necessary for transfer



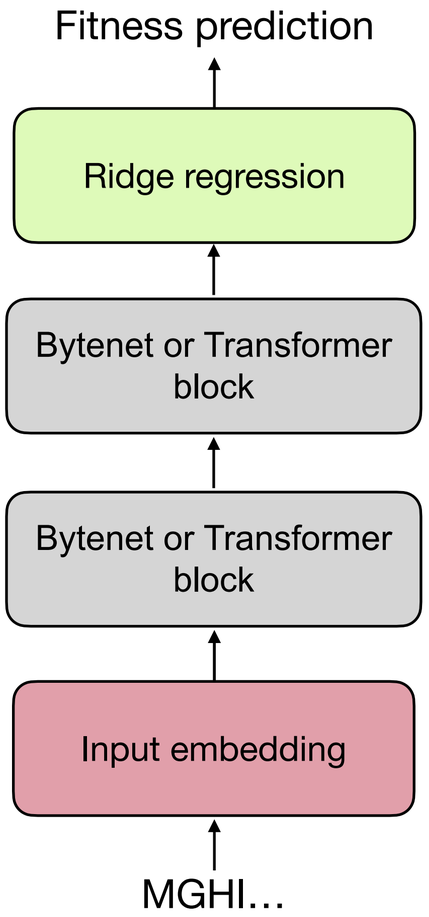


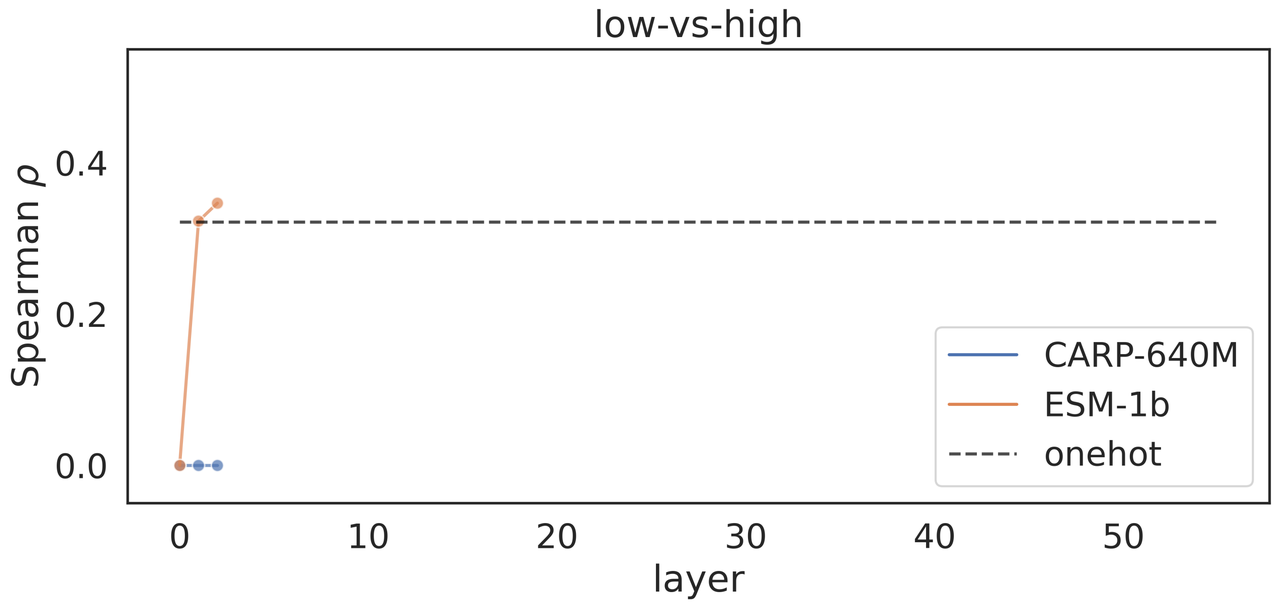
Not all layers are necessary for transfer



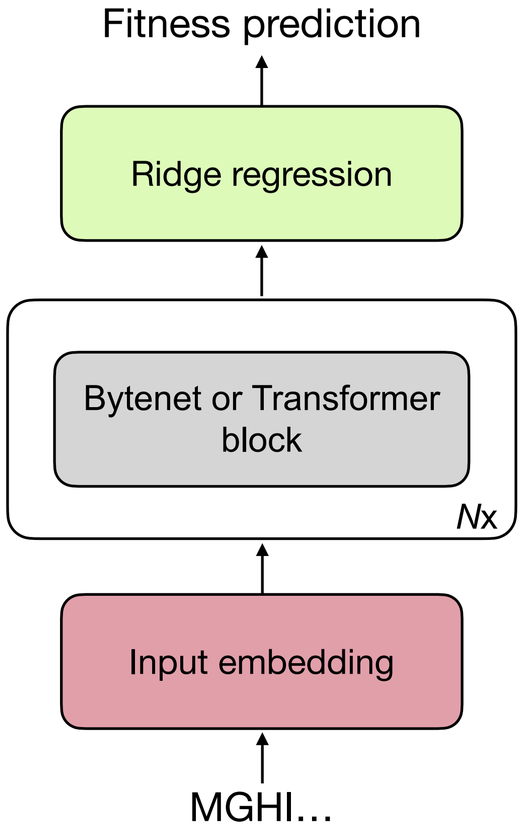


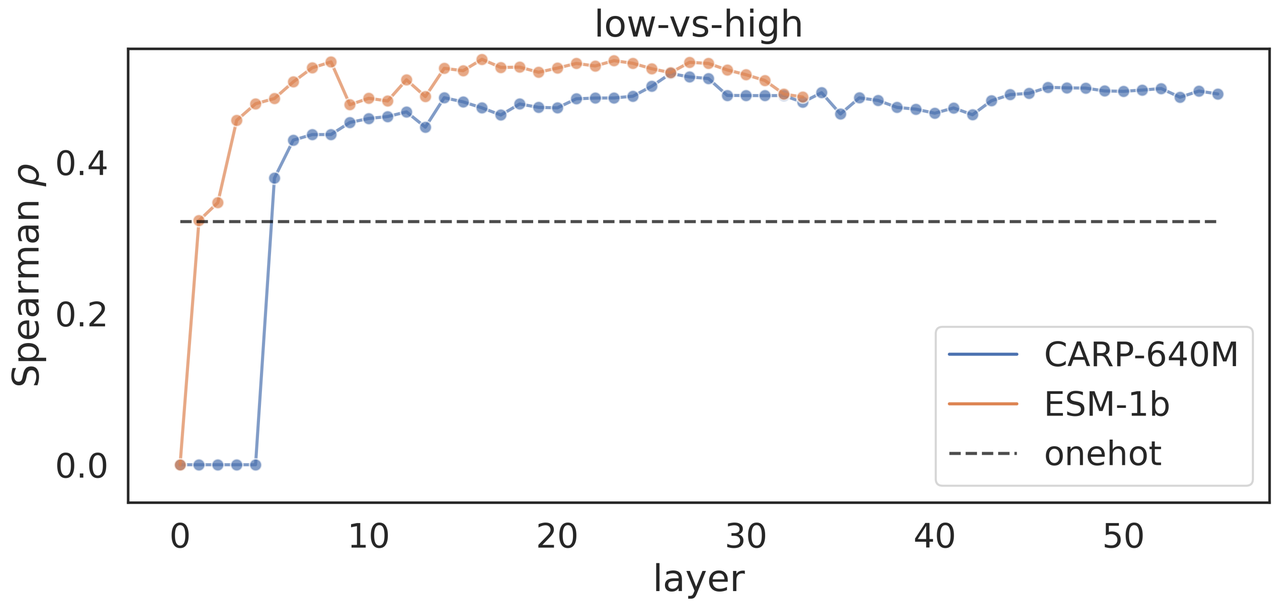
Not all layers are necessary for transfer



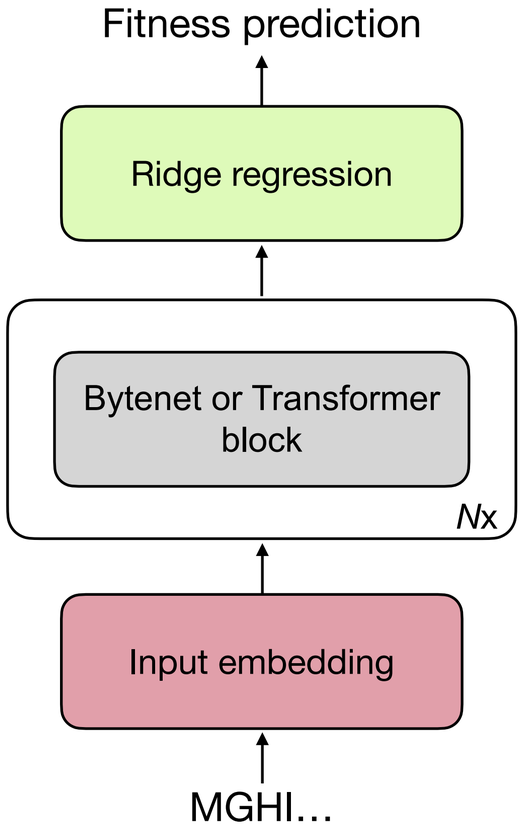


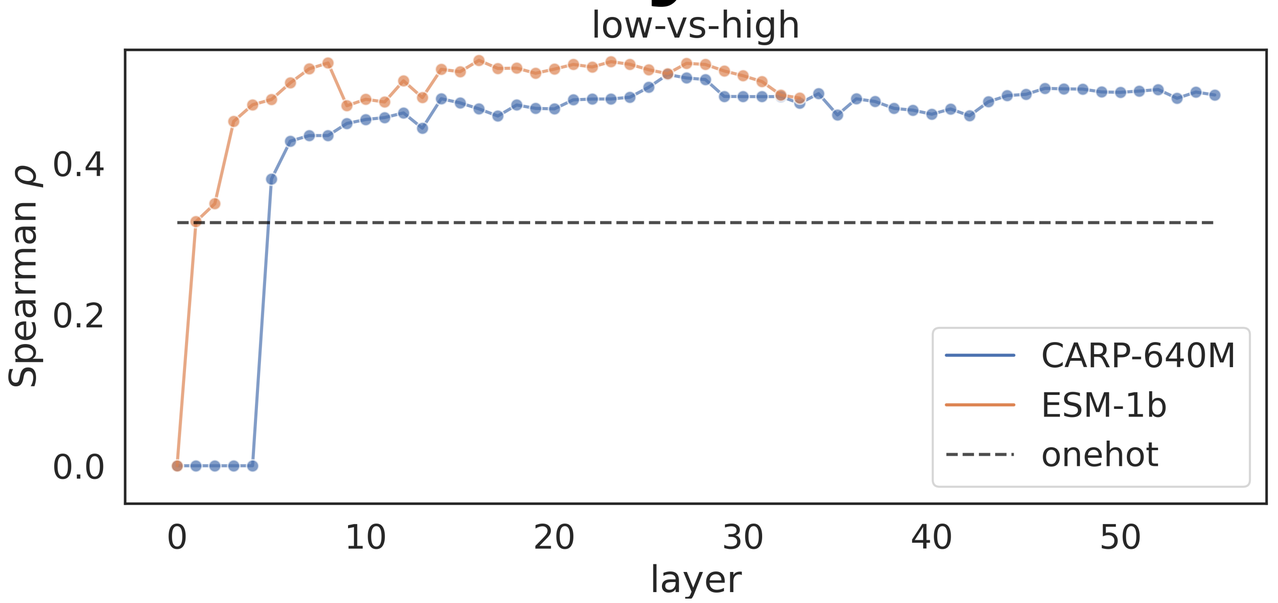
Not all layers are necessary for transfer

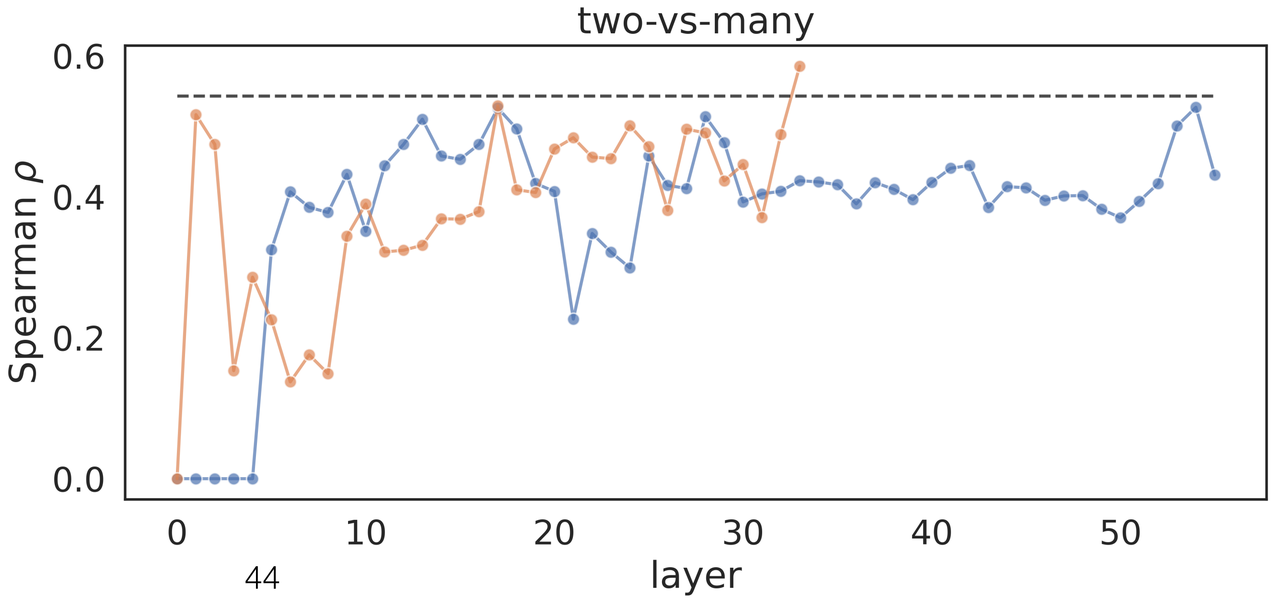




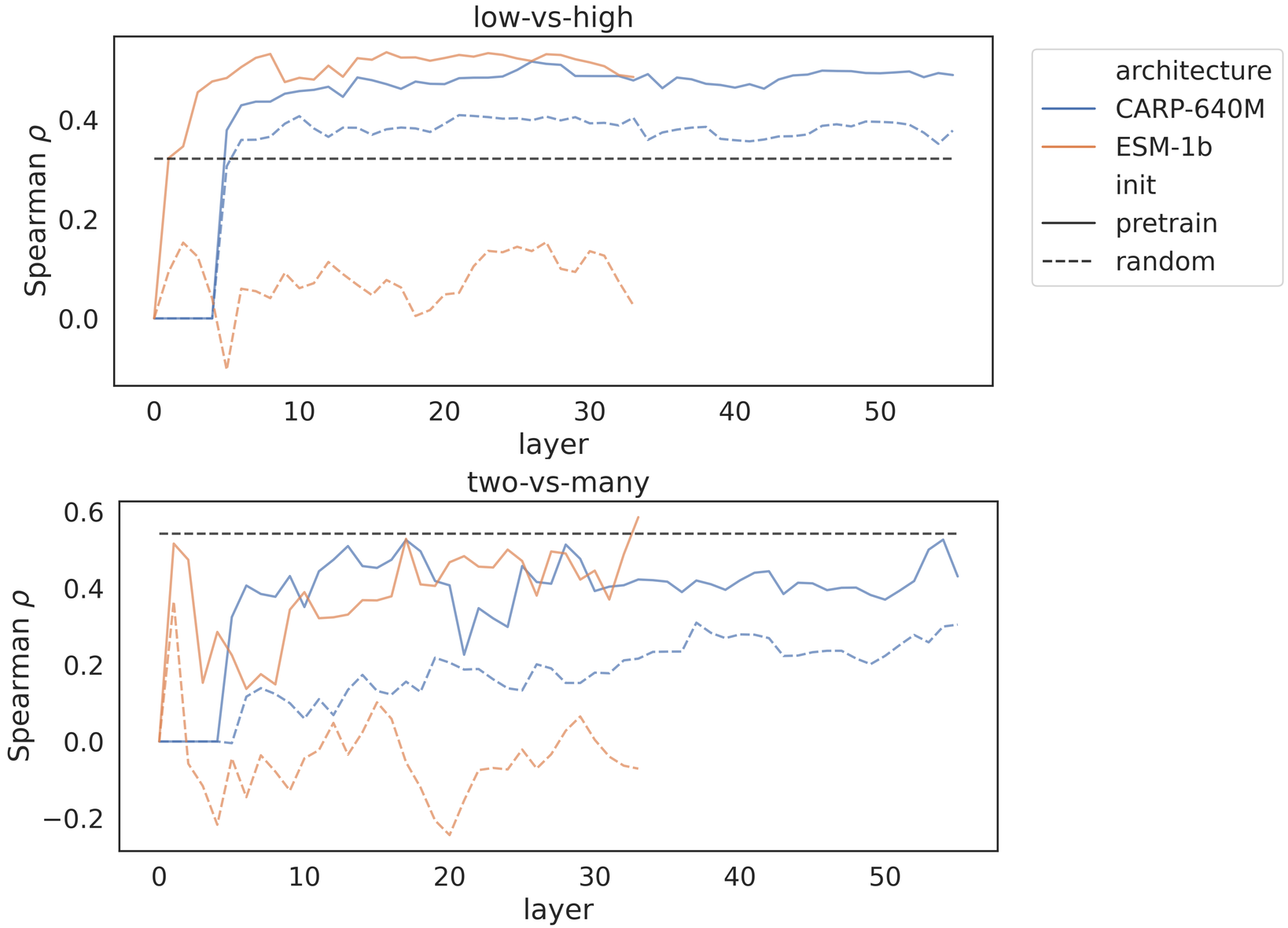
Not all layers are necessary for transfer



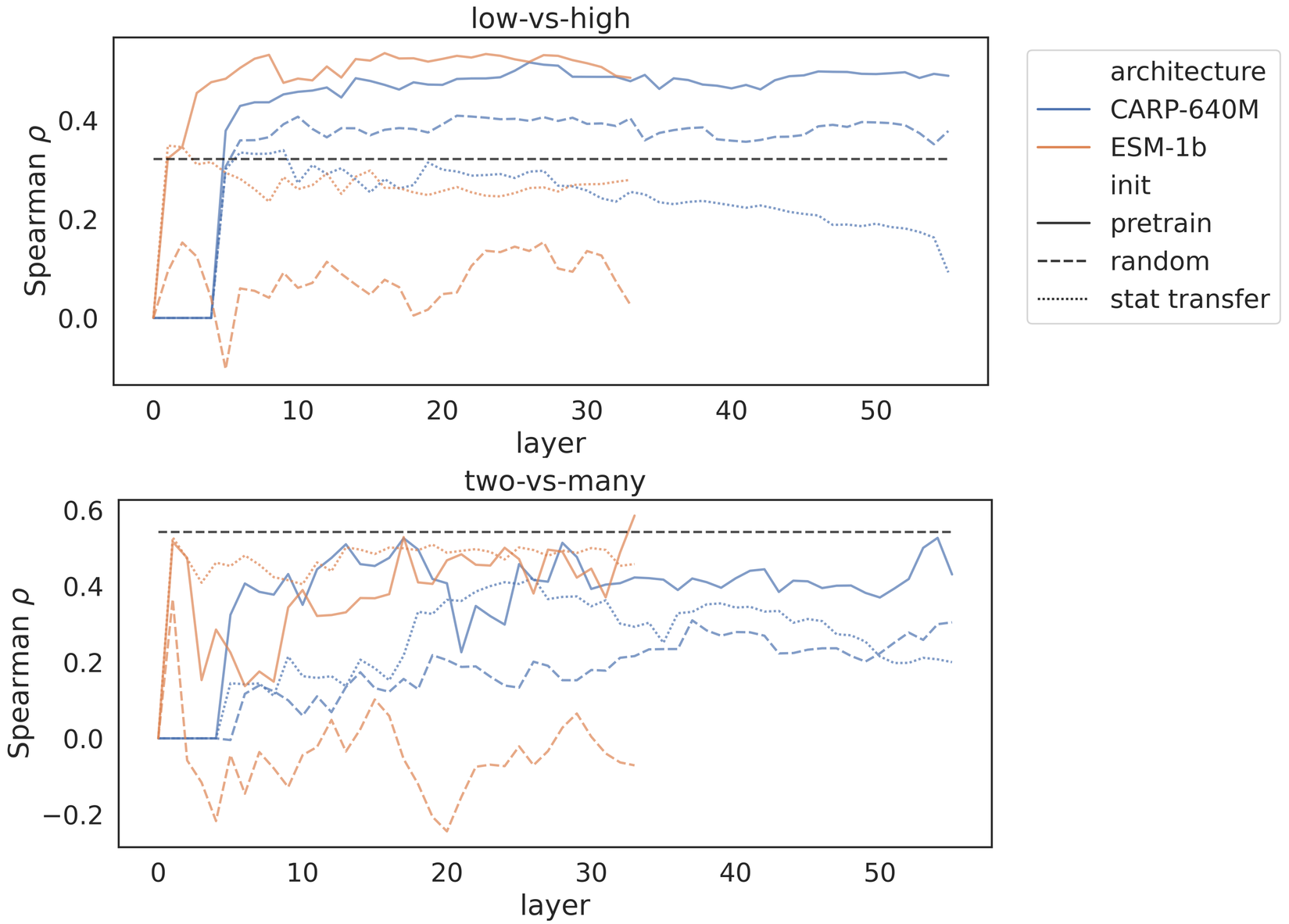




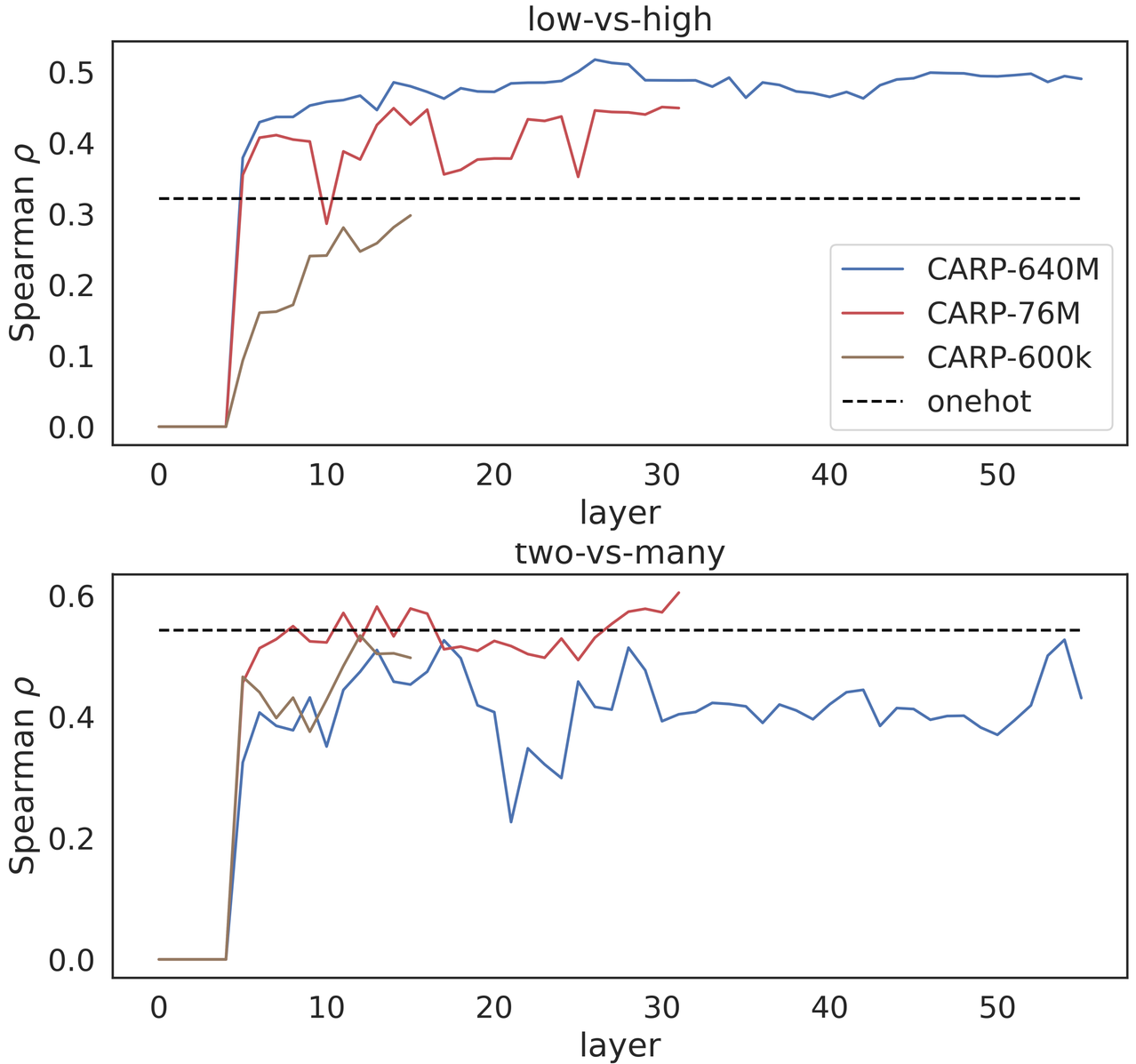
Pretraining helps beyond the architecture



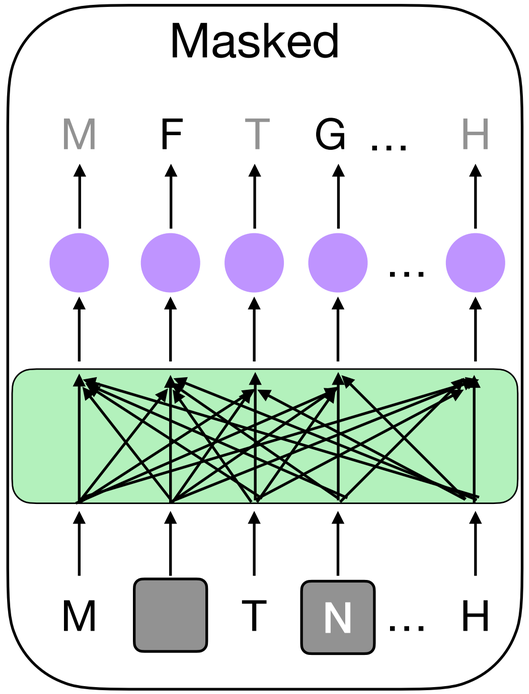
Sometimes, shuffled weights work too

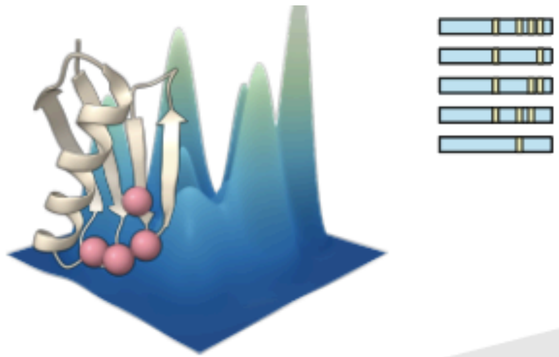


Bigger models do not always help



Pretrain and downstream tasks are mismatched

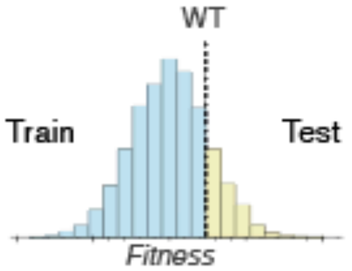




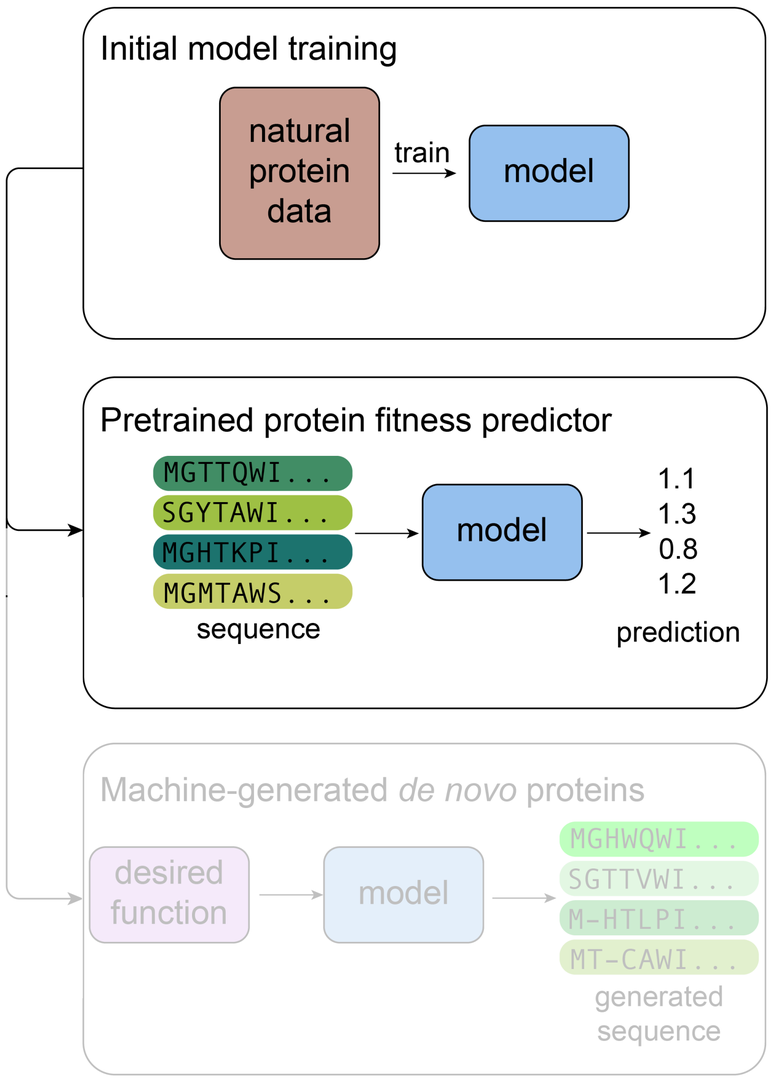
GB1: Highly epistatic, only mutates four positions

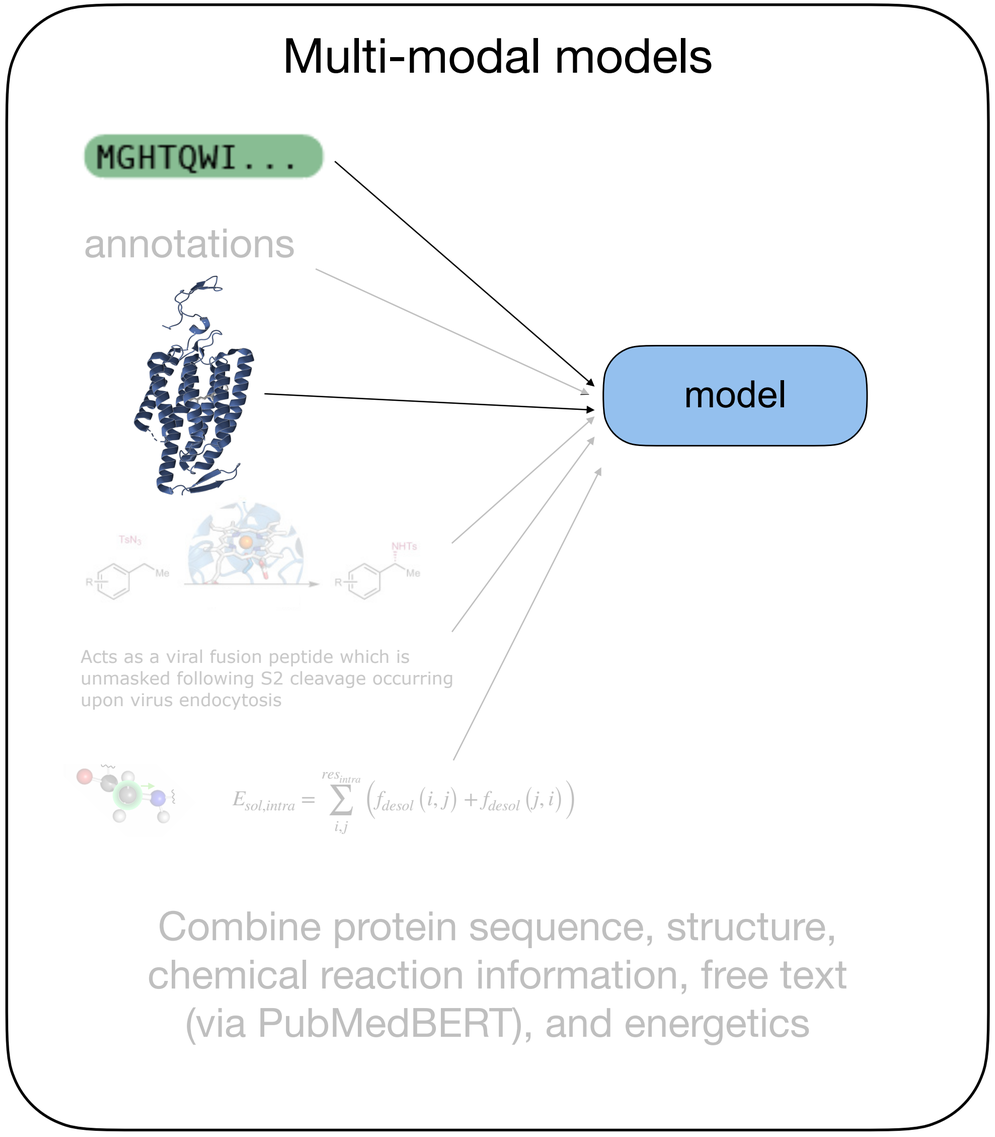
One/Two

► Rest



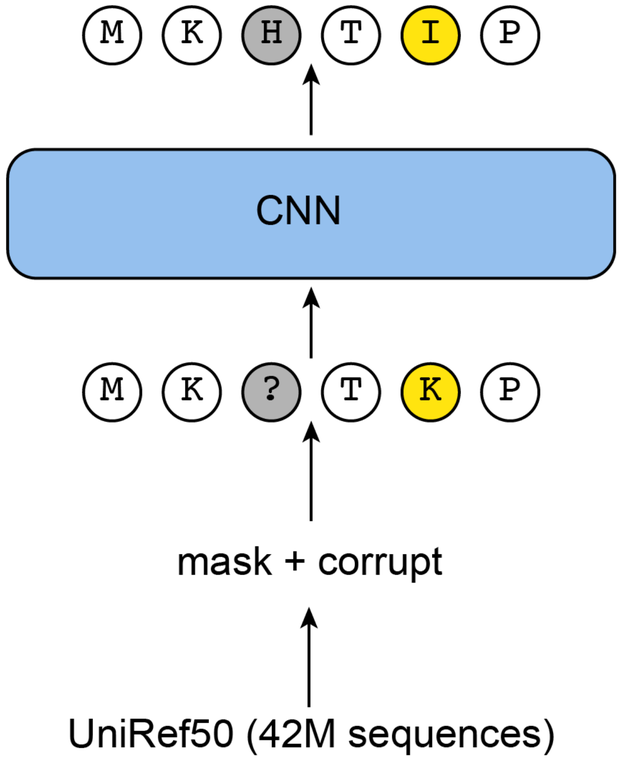
Pretrain with structure and sequence





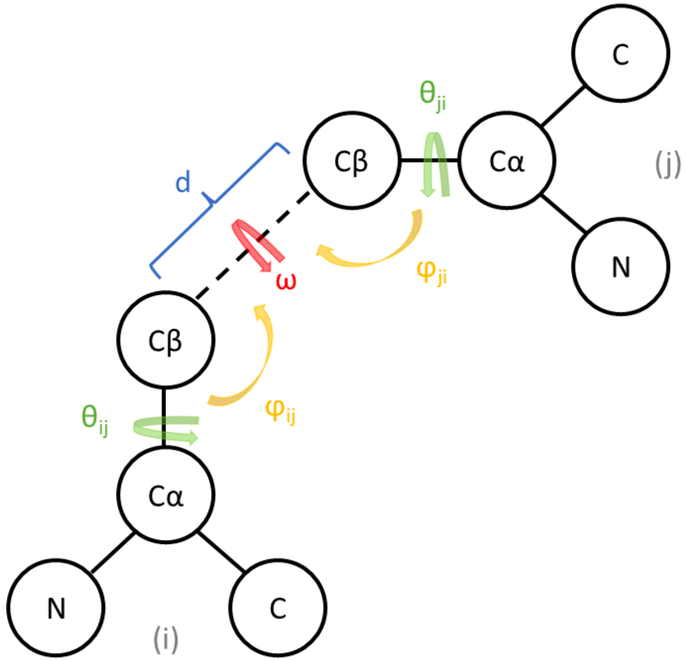
Structural information improves pretraining

Convolutional autoencoding representations of proteins



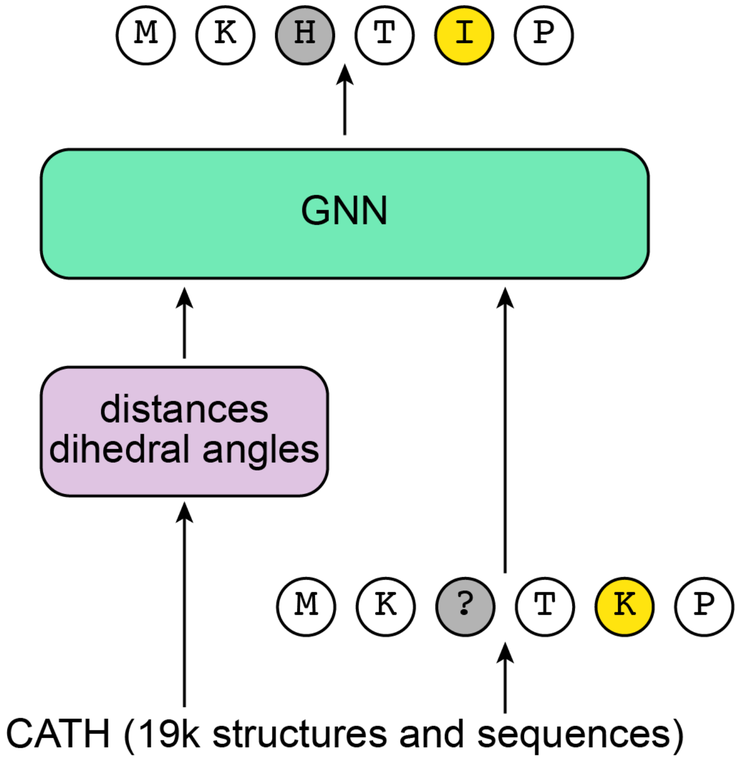
Masked inverse folding

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




Structural information improves pretraining

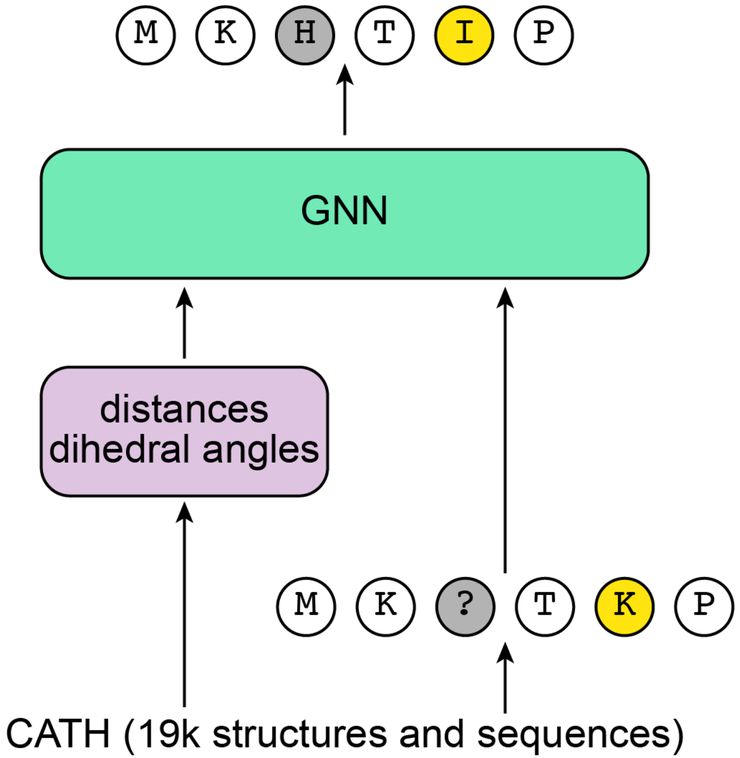
Masked inverse folding 



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

Sequence transfer improves pretraining more

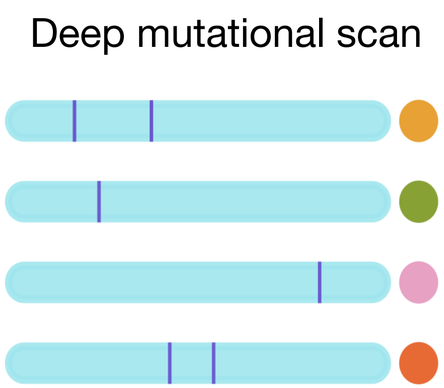
Masked inverse folding 

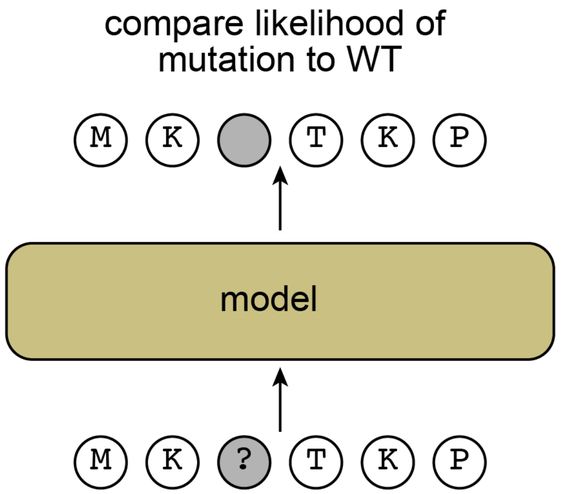


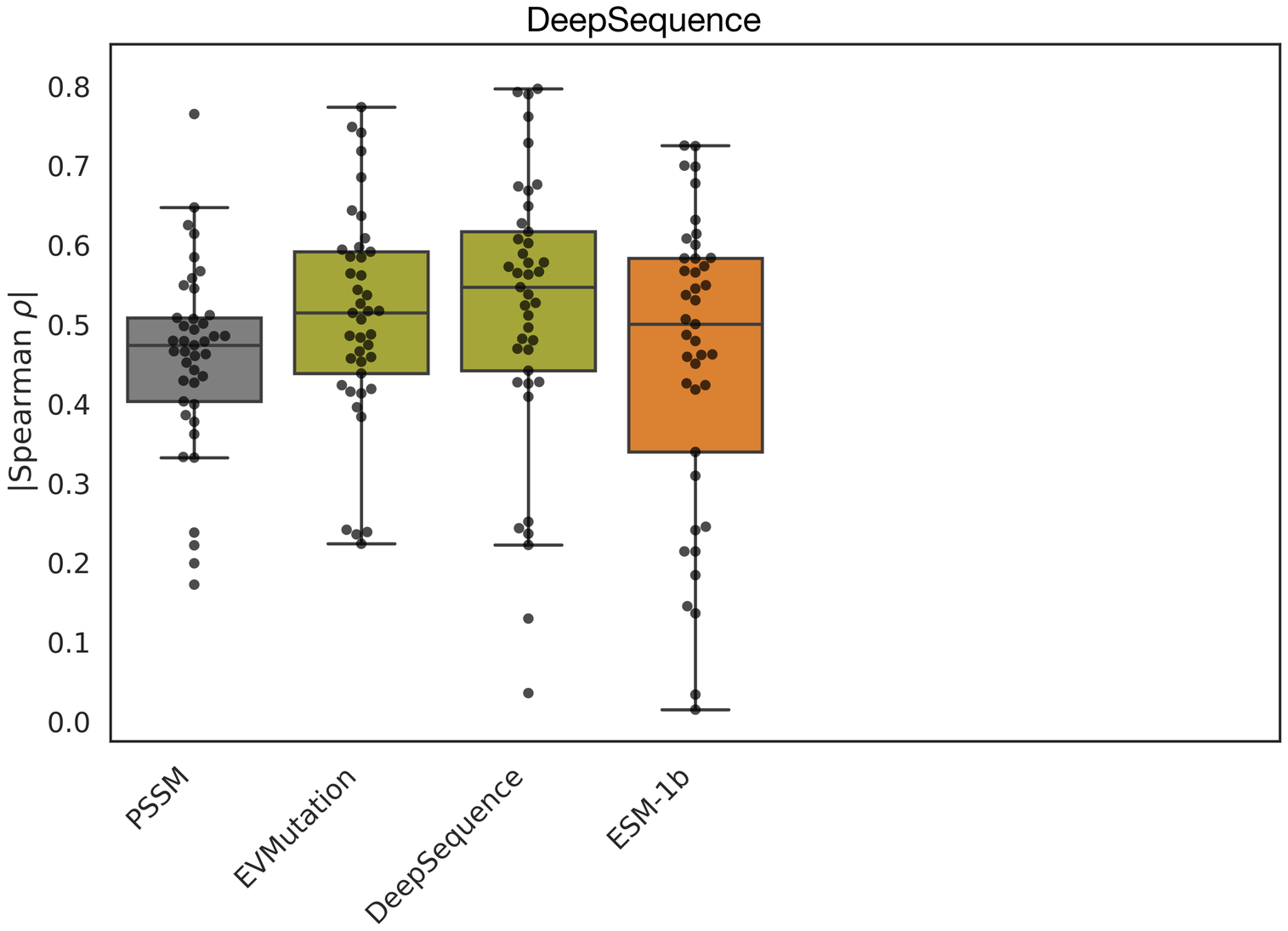
Masked inverse folding with sequence transfer

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

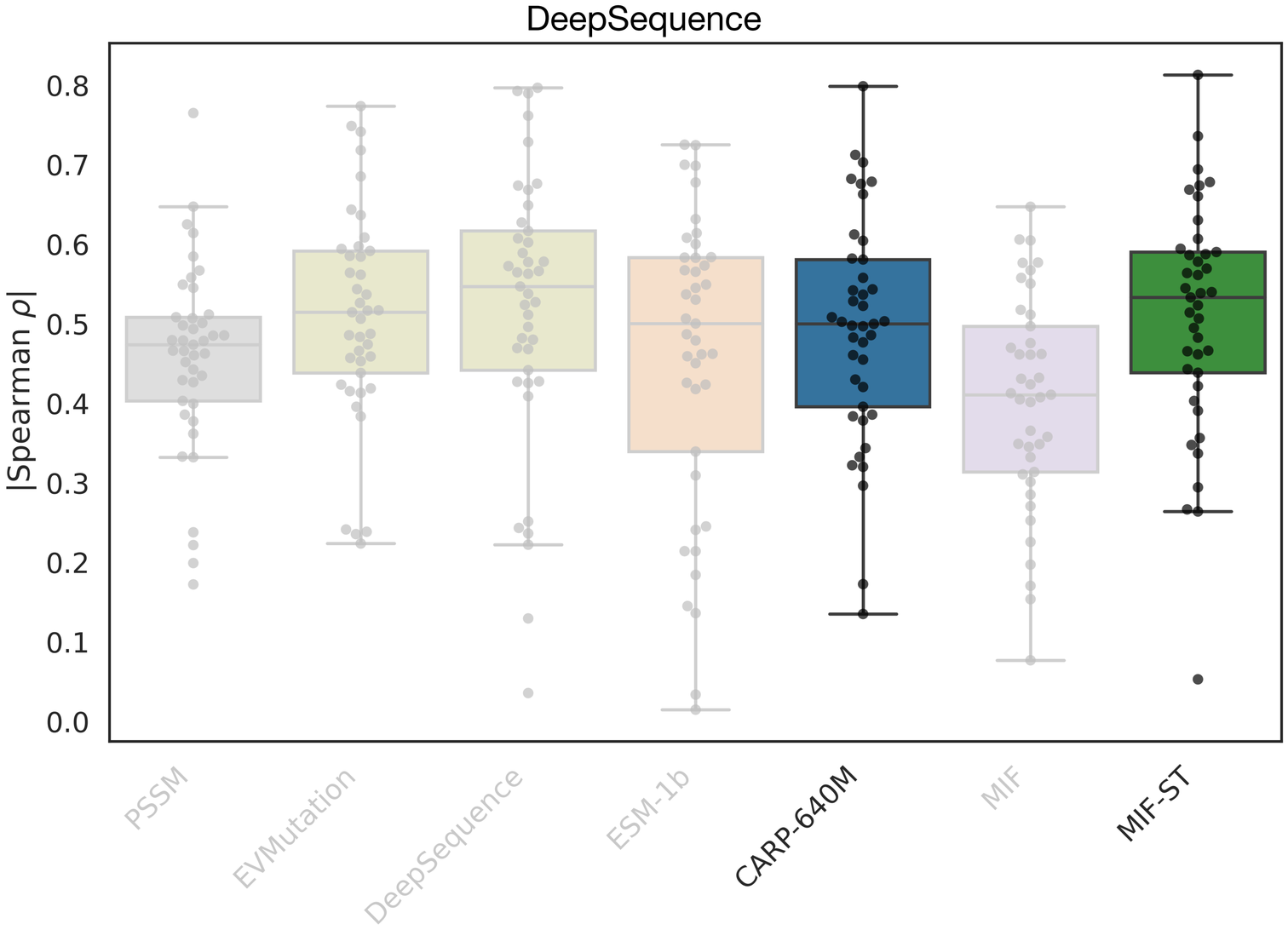
MIF, and MIF-ST are zero-shot fitness predictors

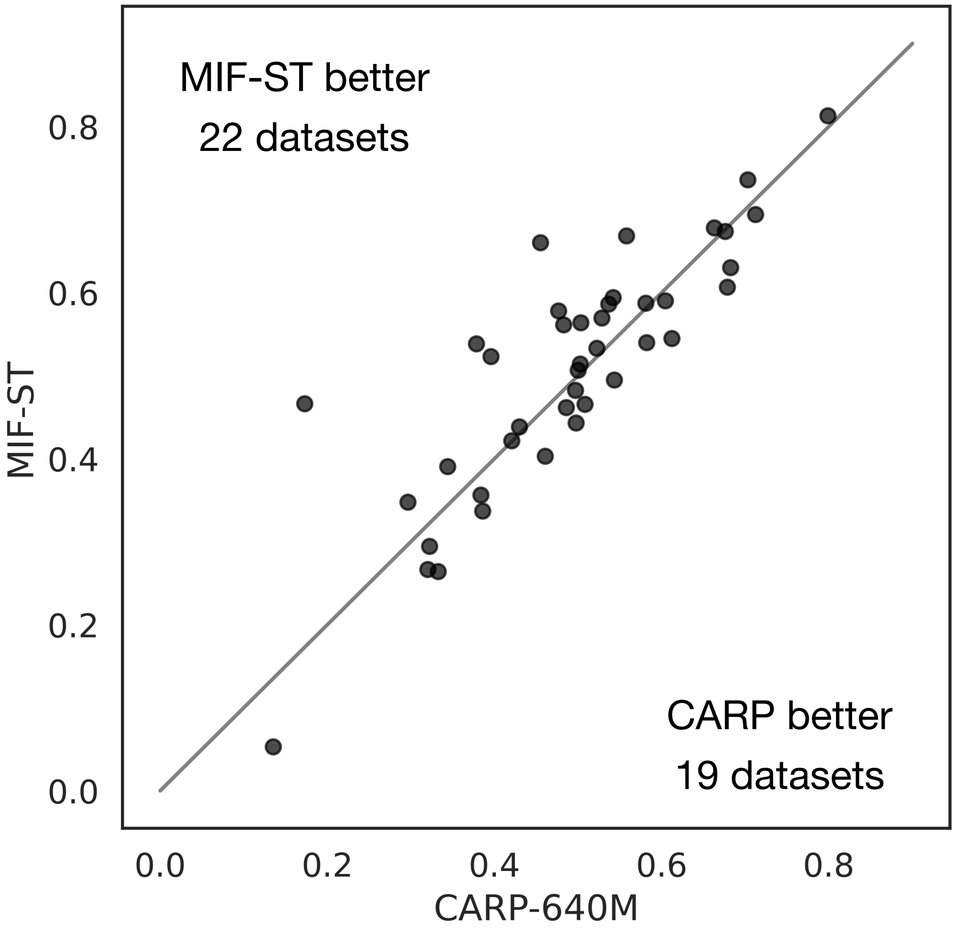




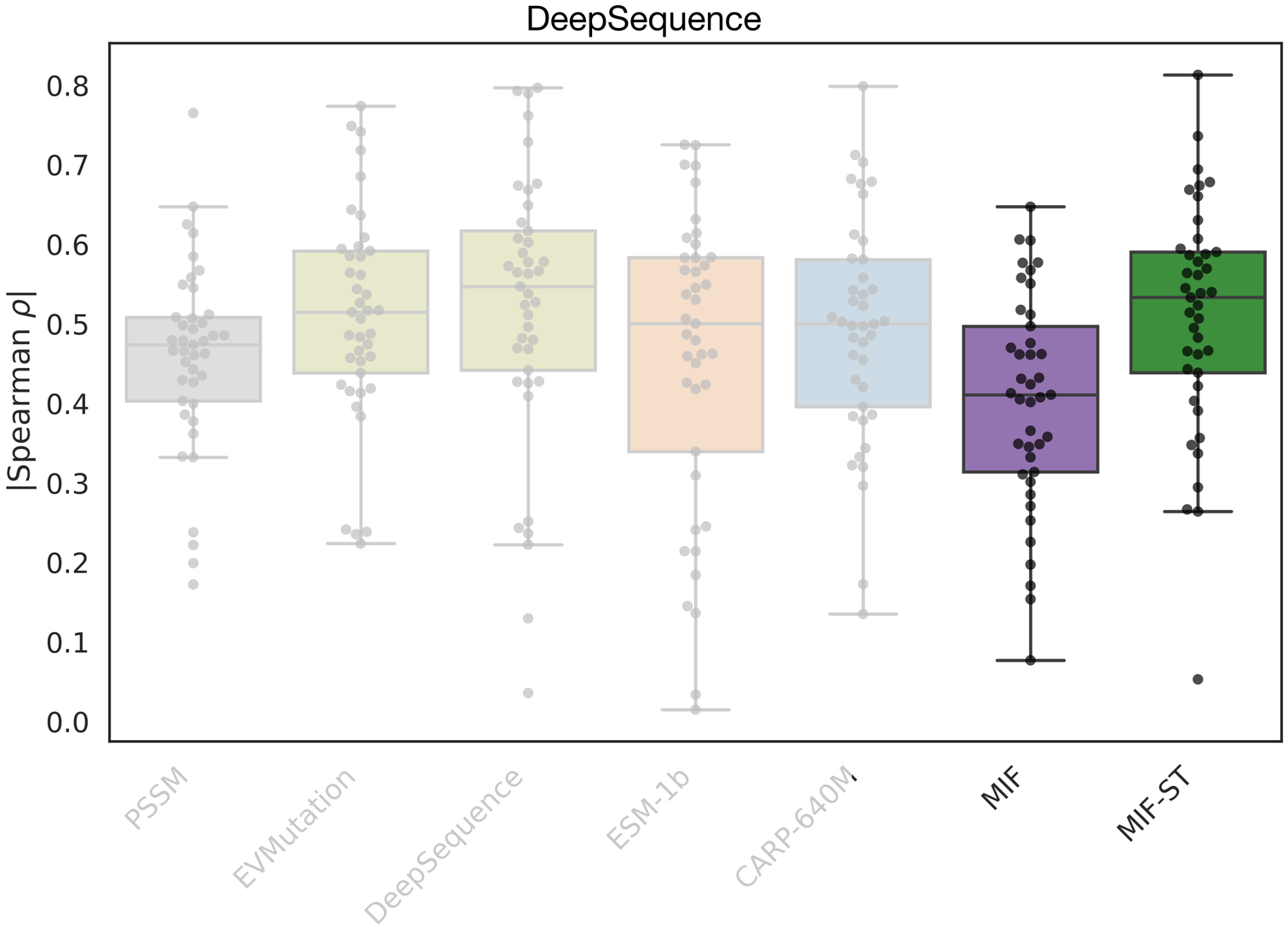


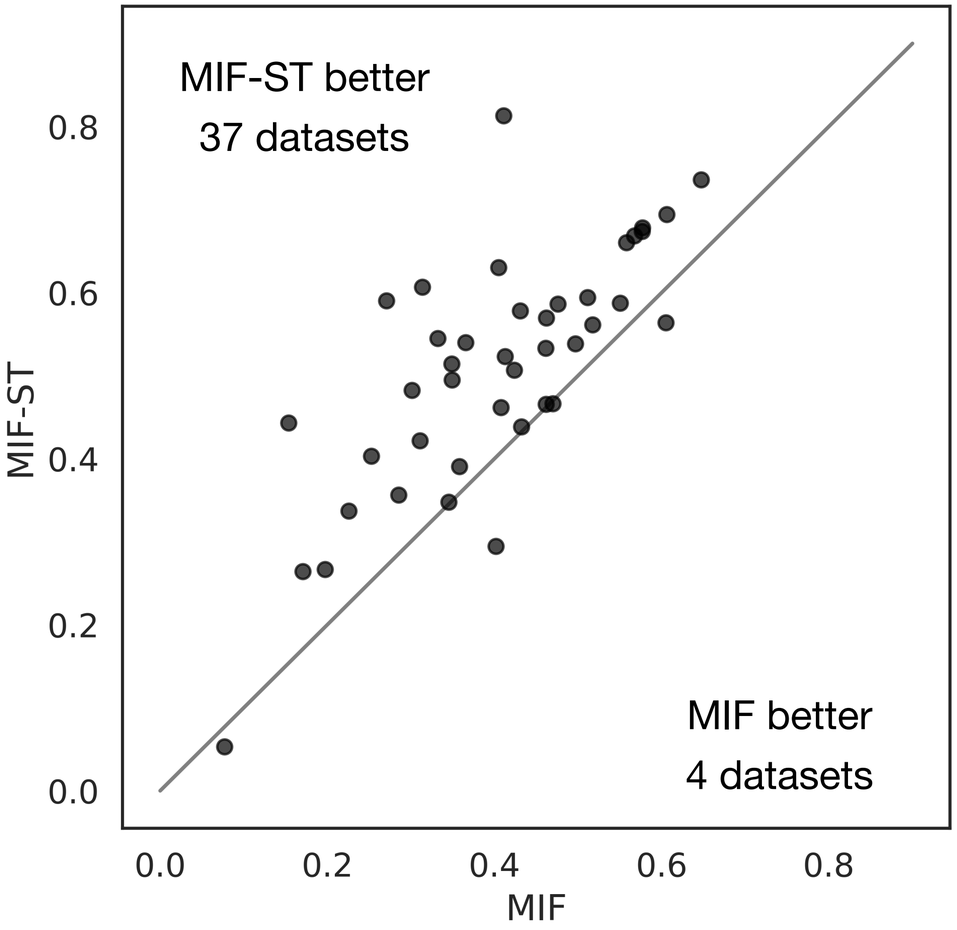
MIF-ST outperforms CARP and MIF on DeepSequence



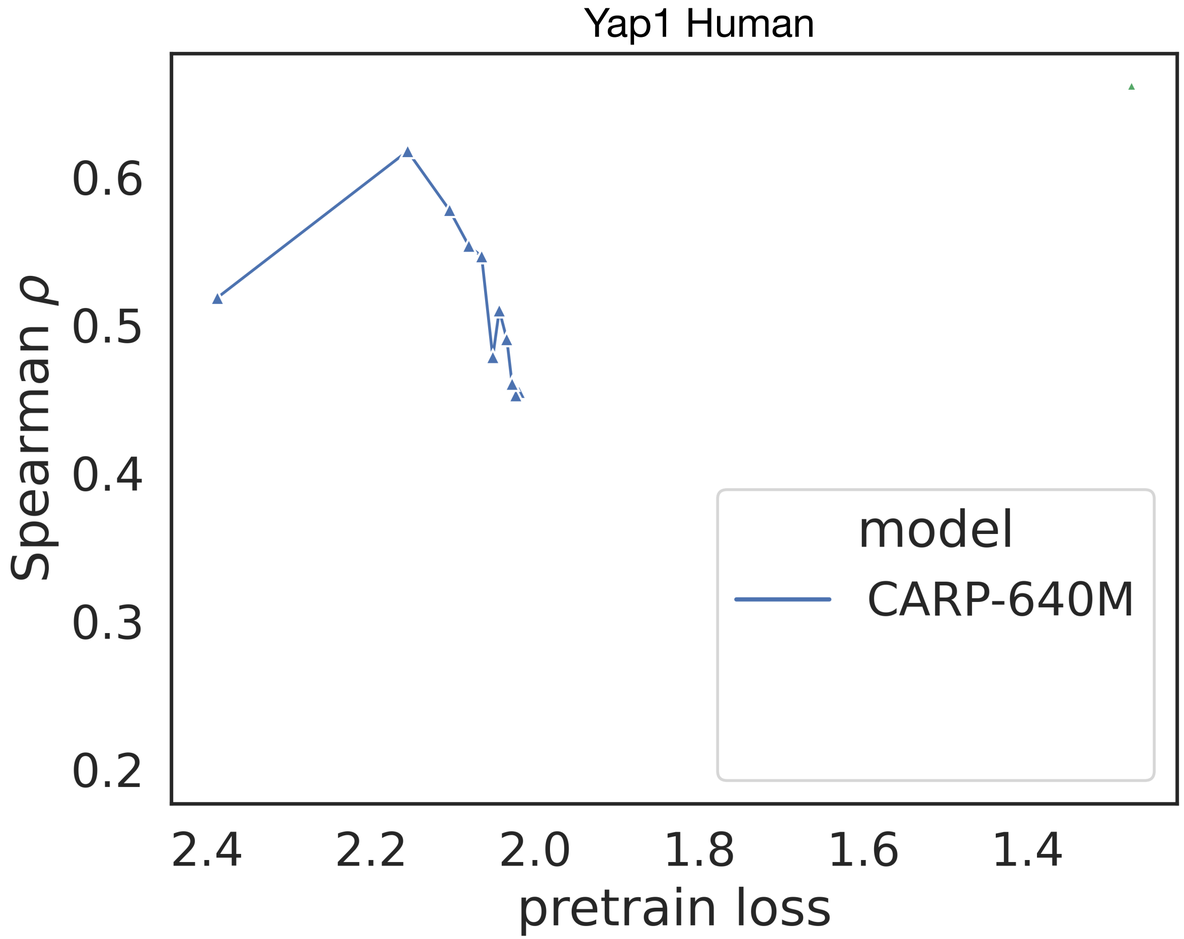


MIF-ST outperforms CARP and MIF on DeepSequence

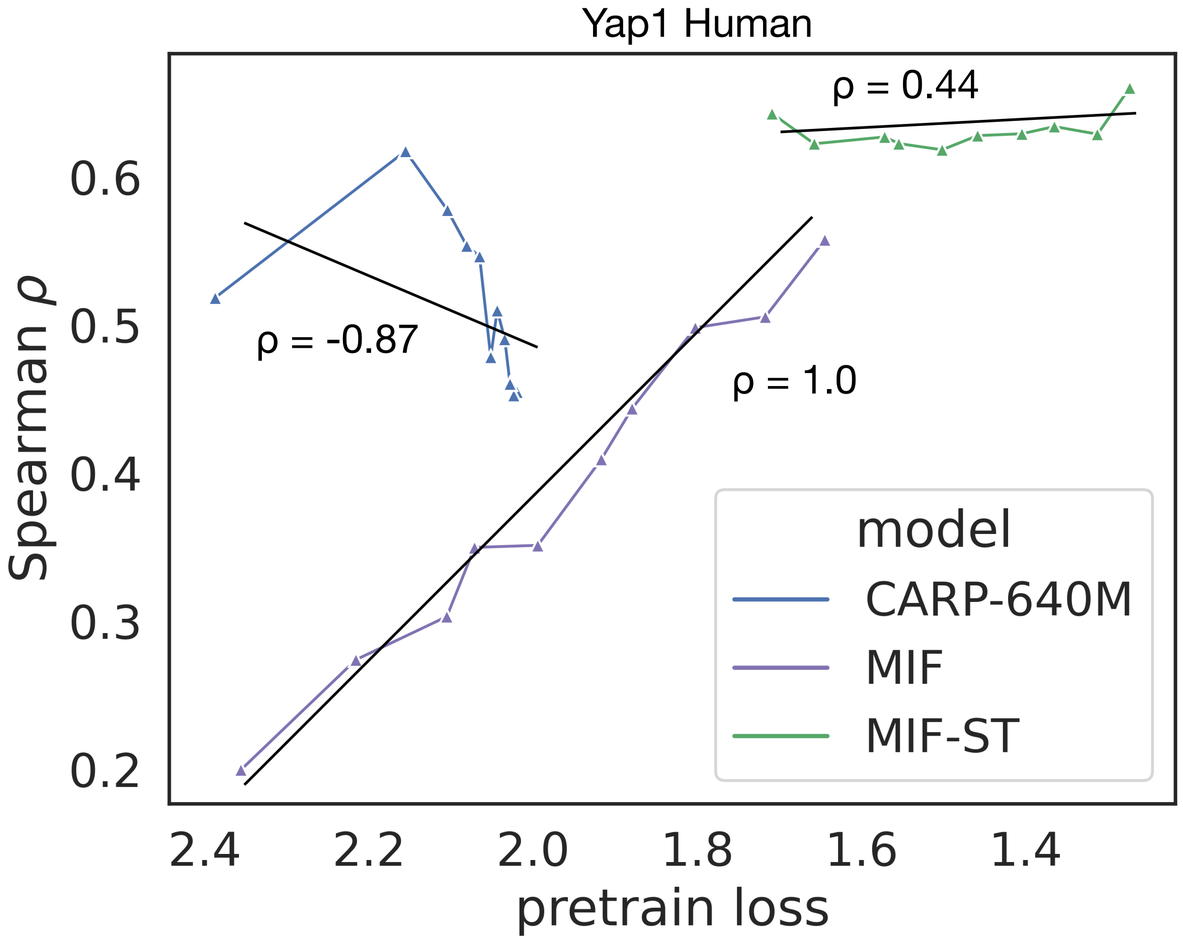


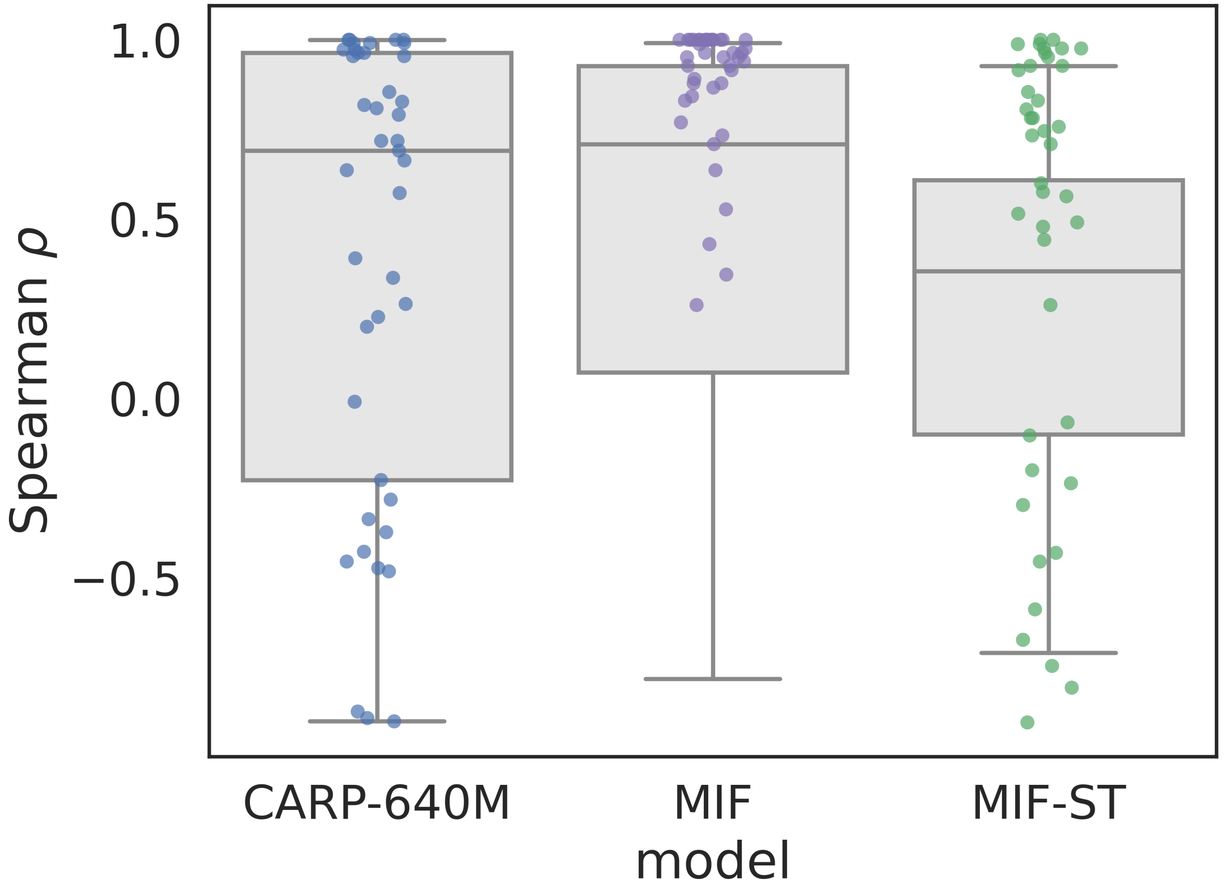


Structure-conditioned zero-shot improves more consistently with pretraining

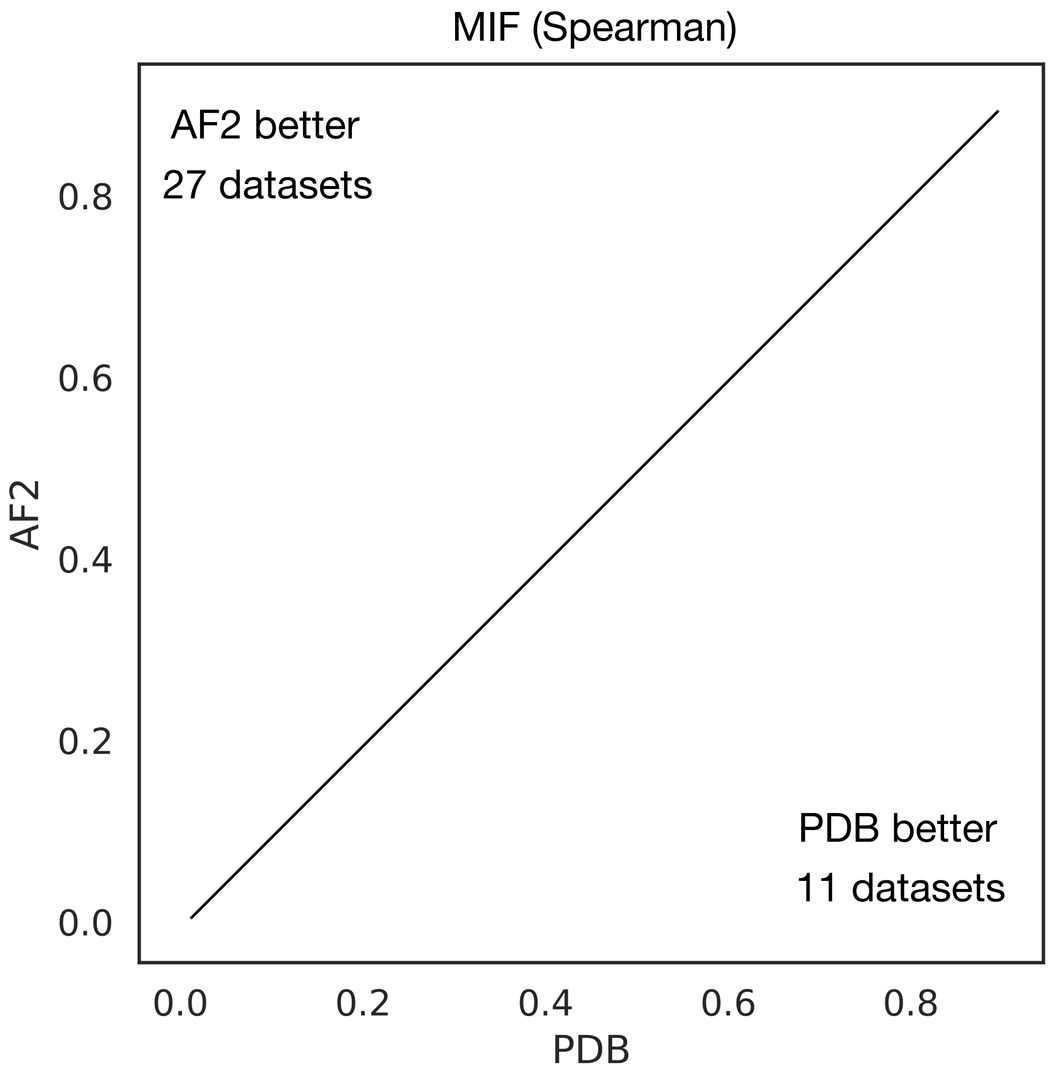


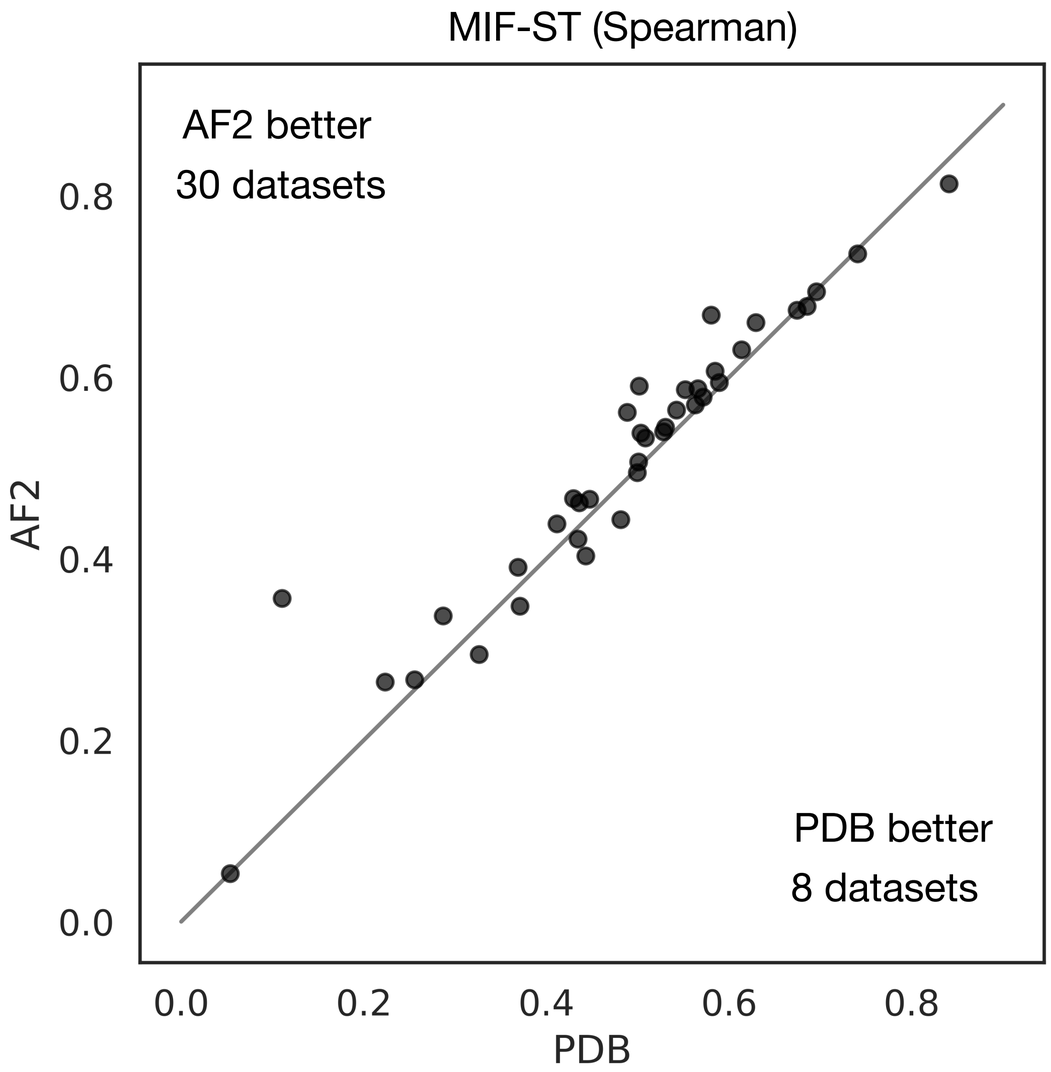
Structure-conditioned zero-shot improves more consistently with pretraining





Predictions often better using AF2 structures





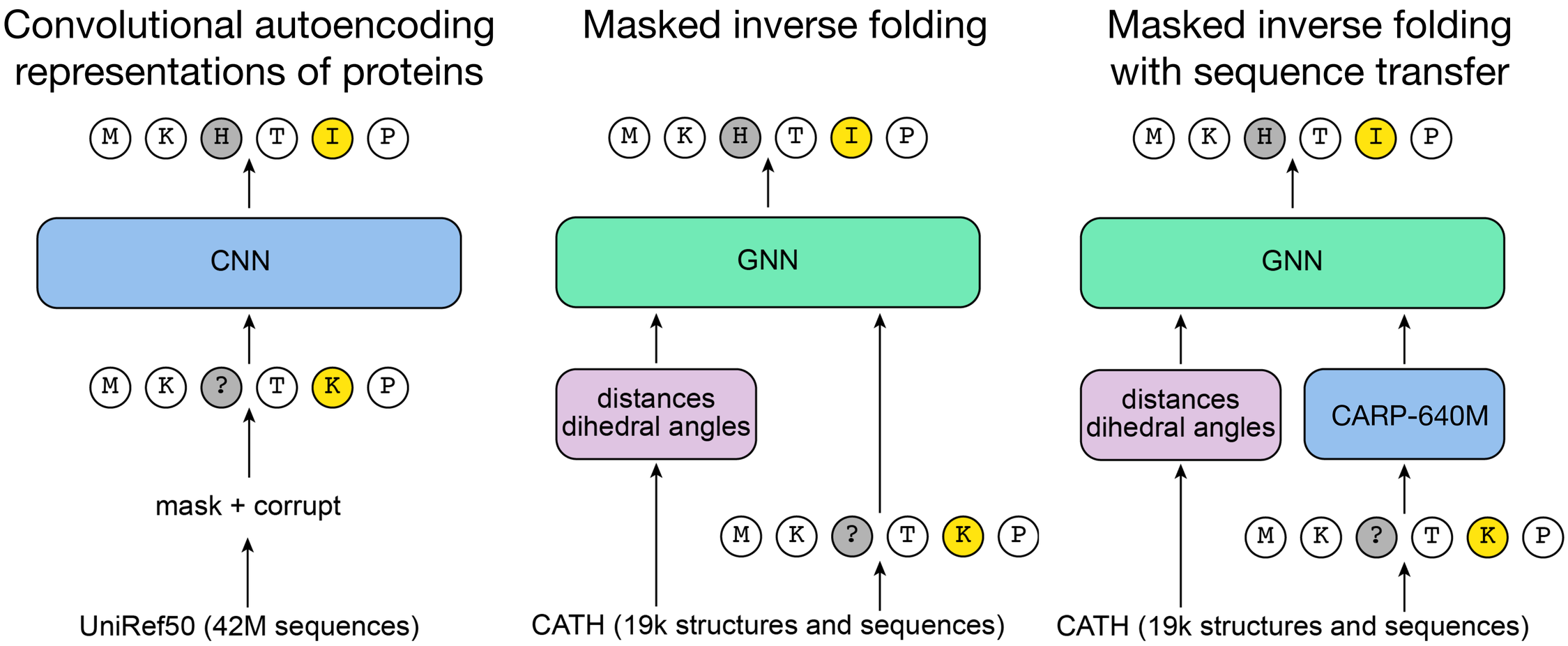
Pretraining improves OOD generalization on GB1

	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

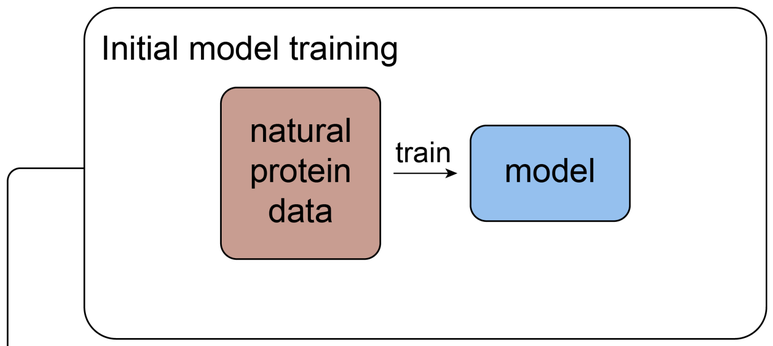
Naive MIF-ST predicts all the same values

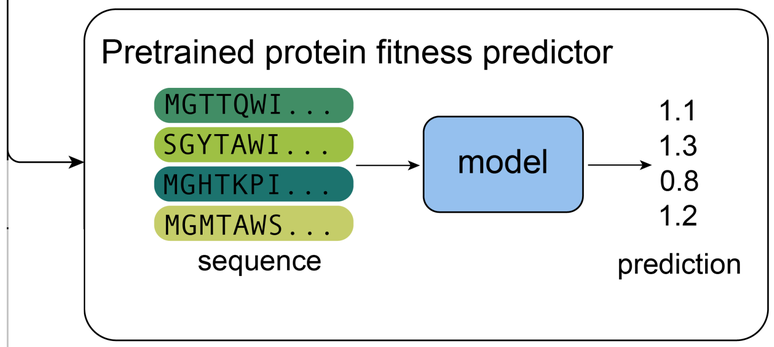
Try CARP, MIF, and MIF-ST!

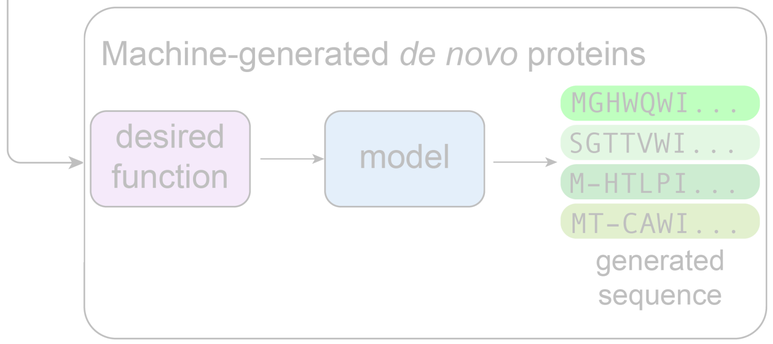


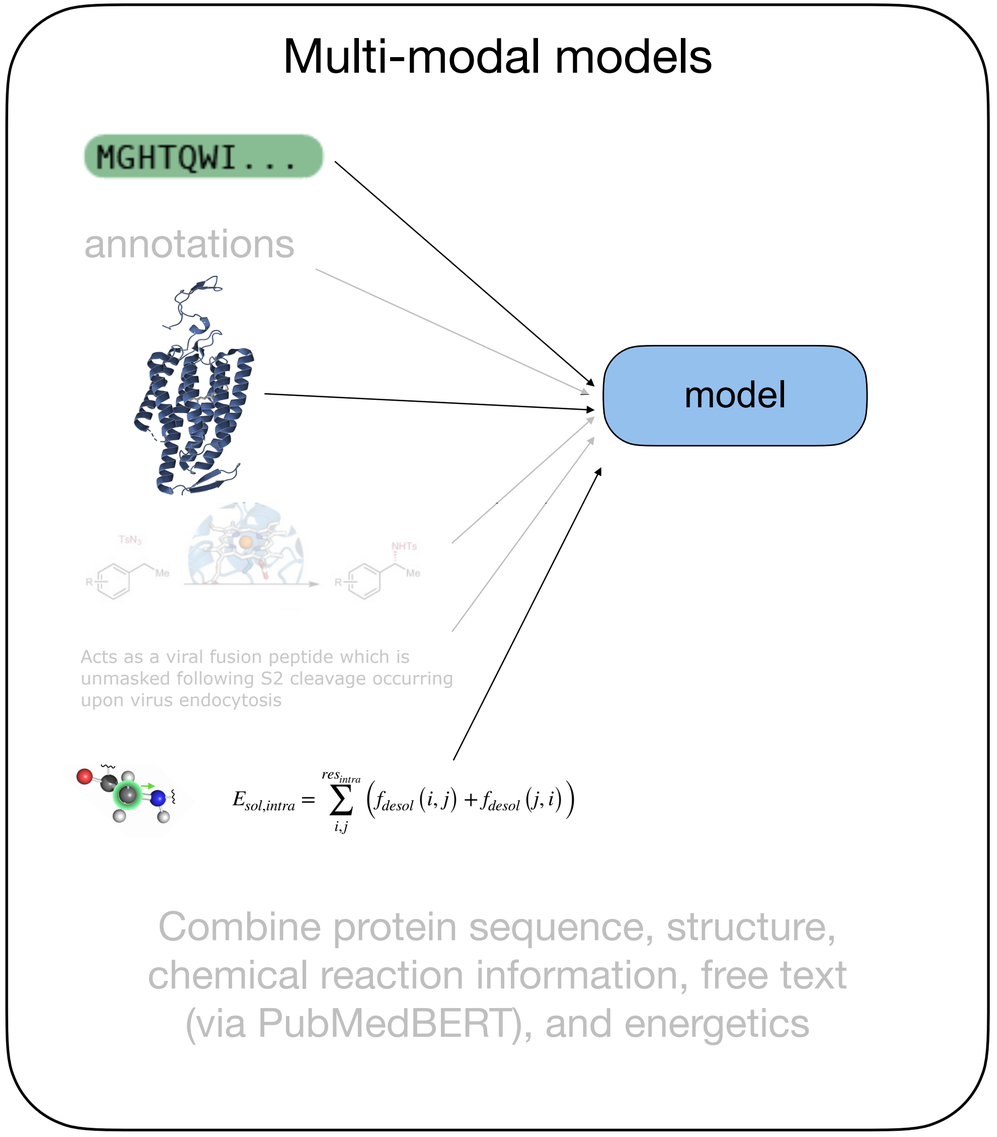
https://github.com/microsoft/protein-sequence-models

Use energetic features as inputs

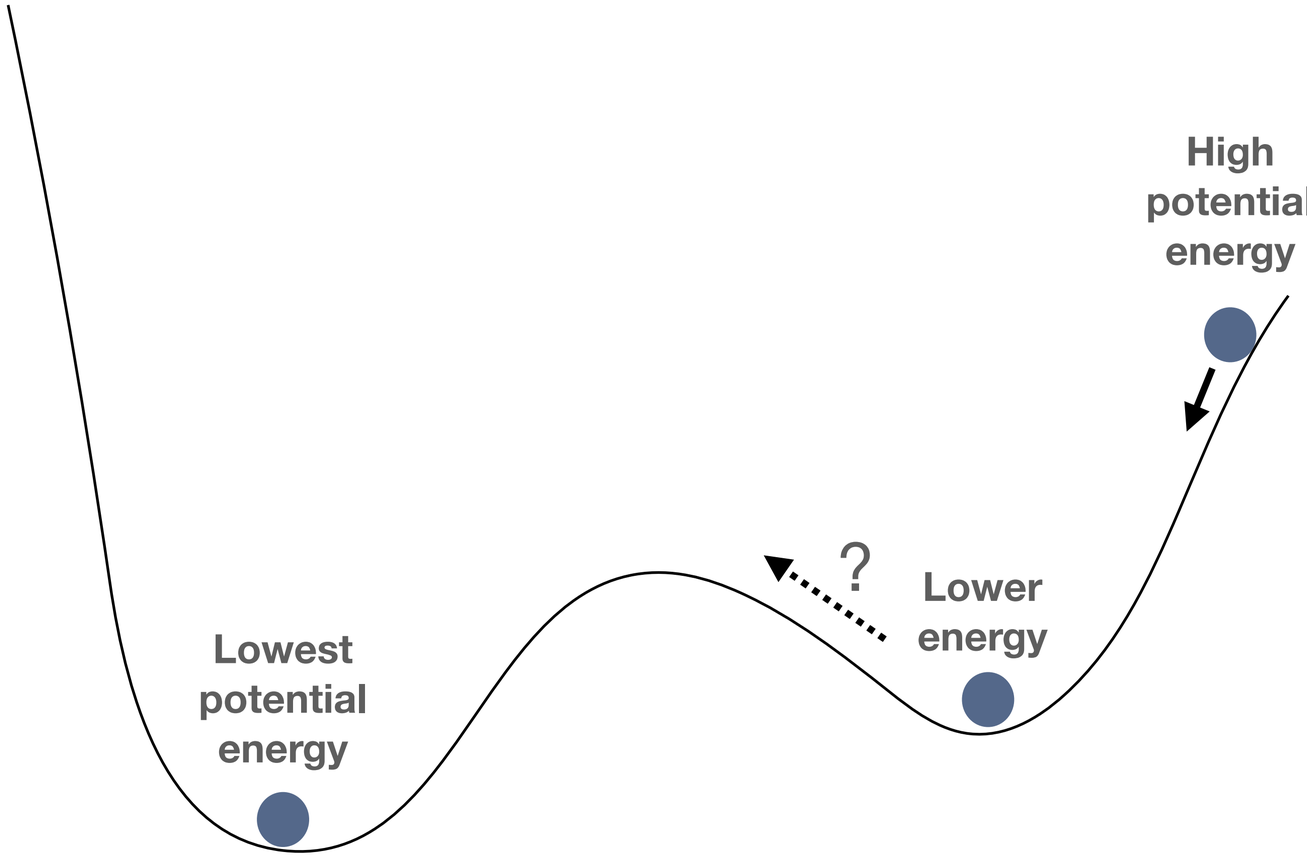


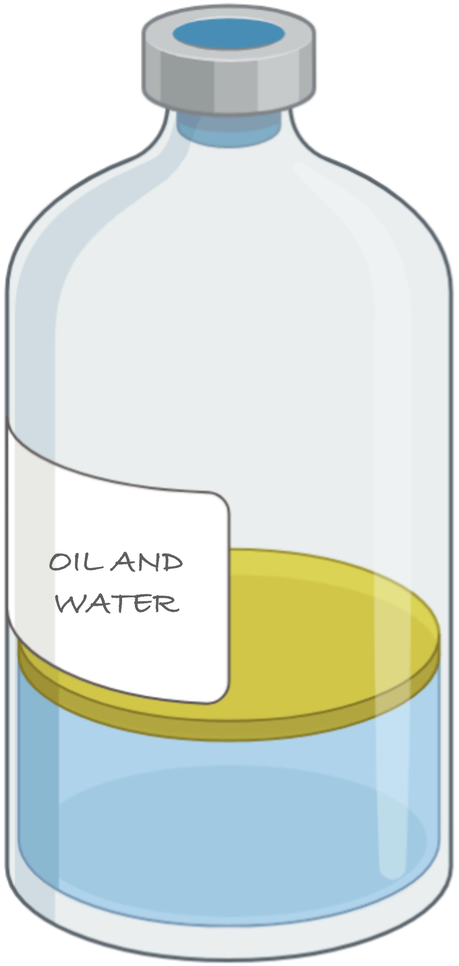




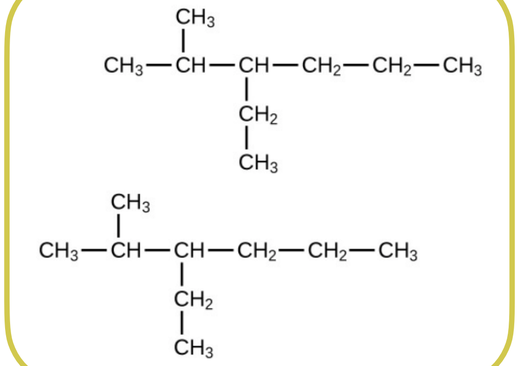


Energy is a physics concept that describes natural processes

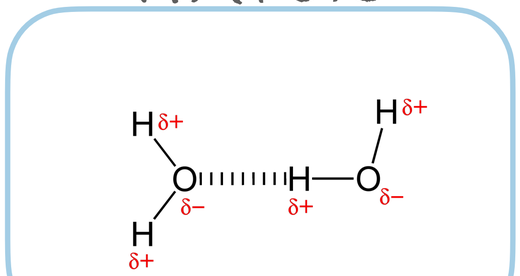




OIL

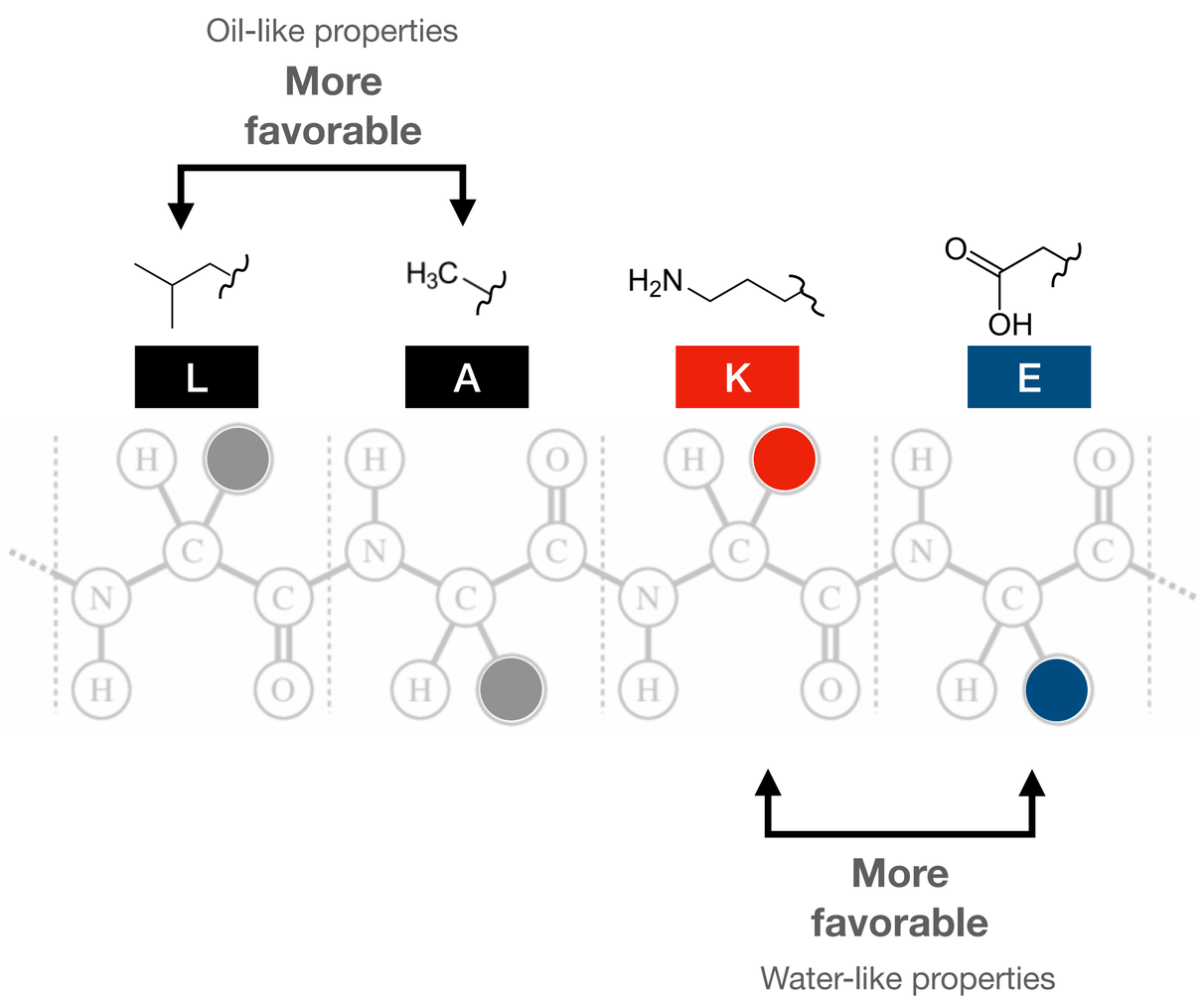


WATER

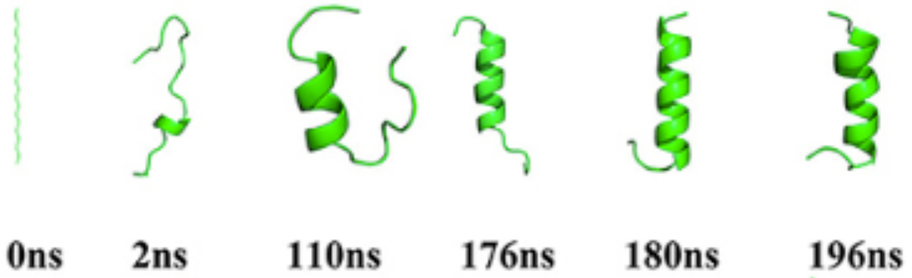


Principle of minimum energy

Protein properties are bound by physical laws

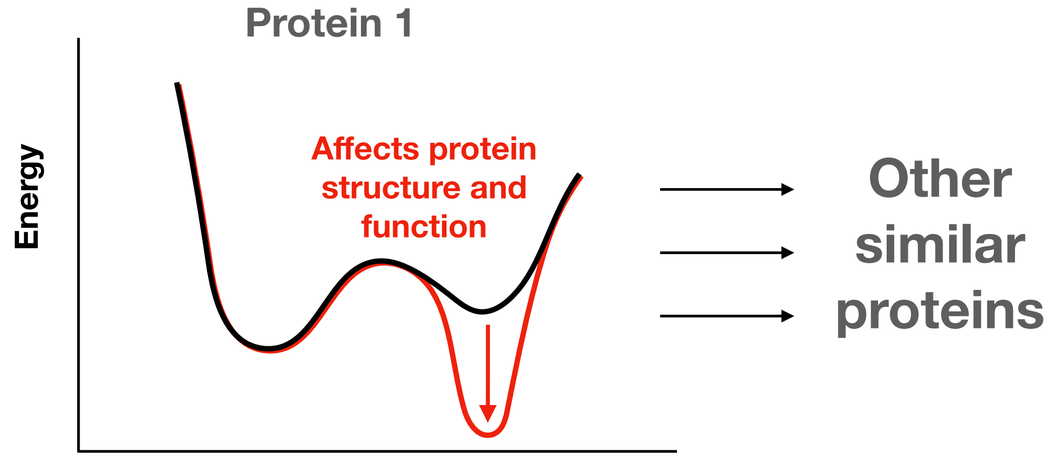


All atom physics-based simulation



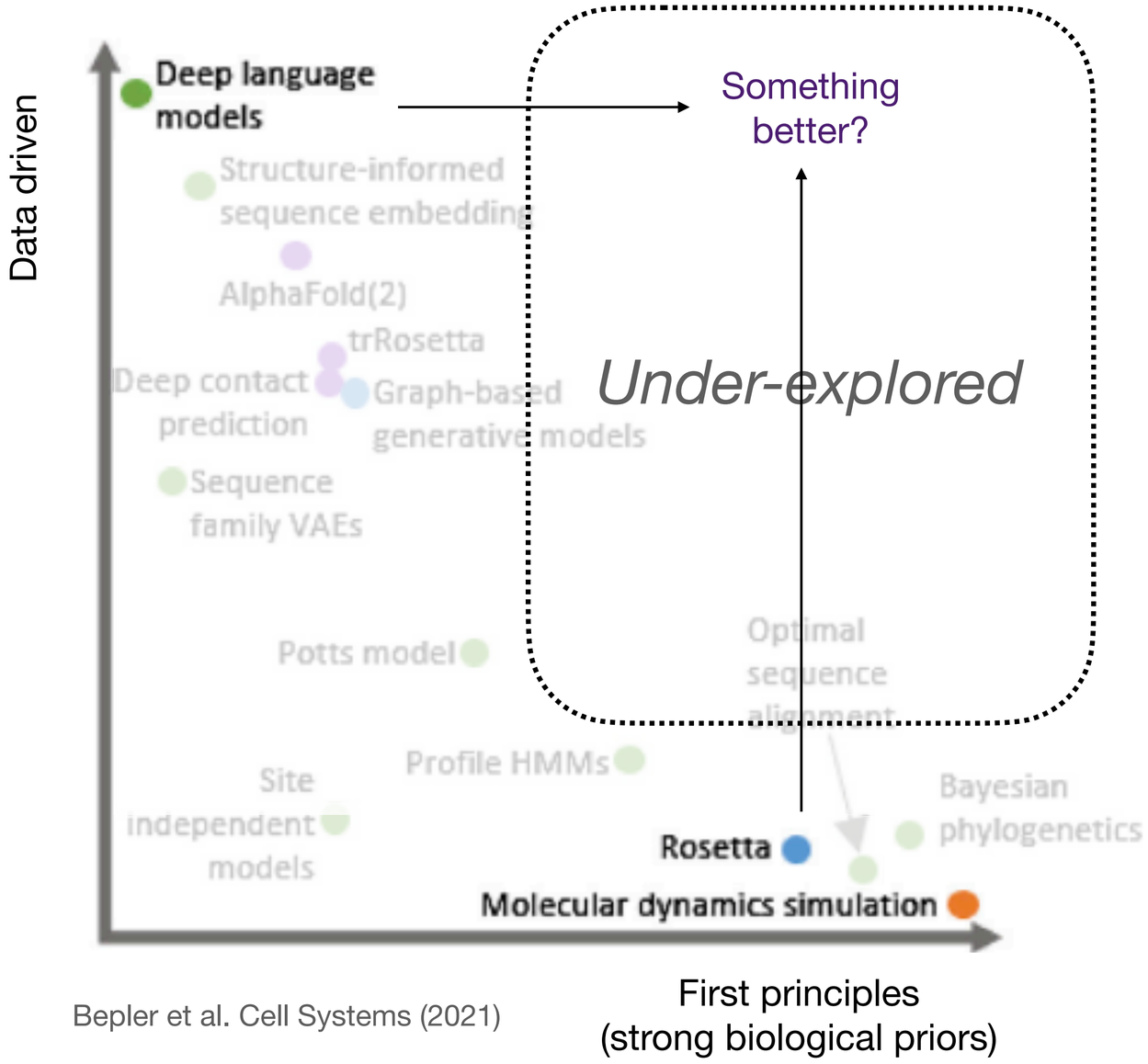
Duan et al. Front Chem (2019)

Protein functional mechanism

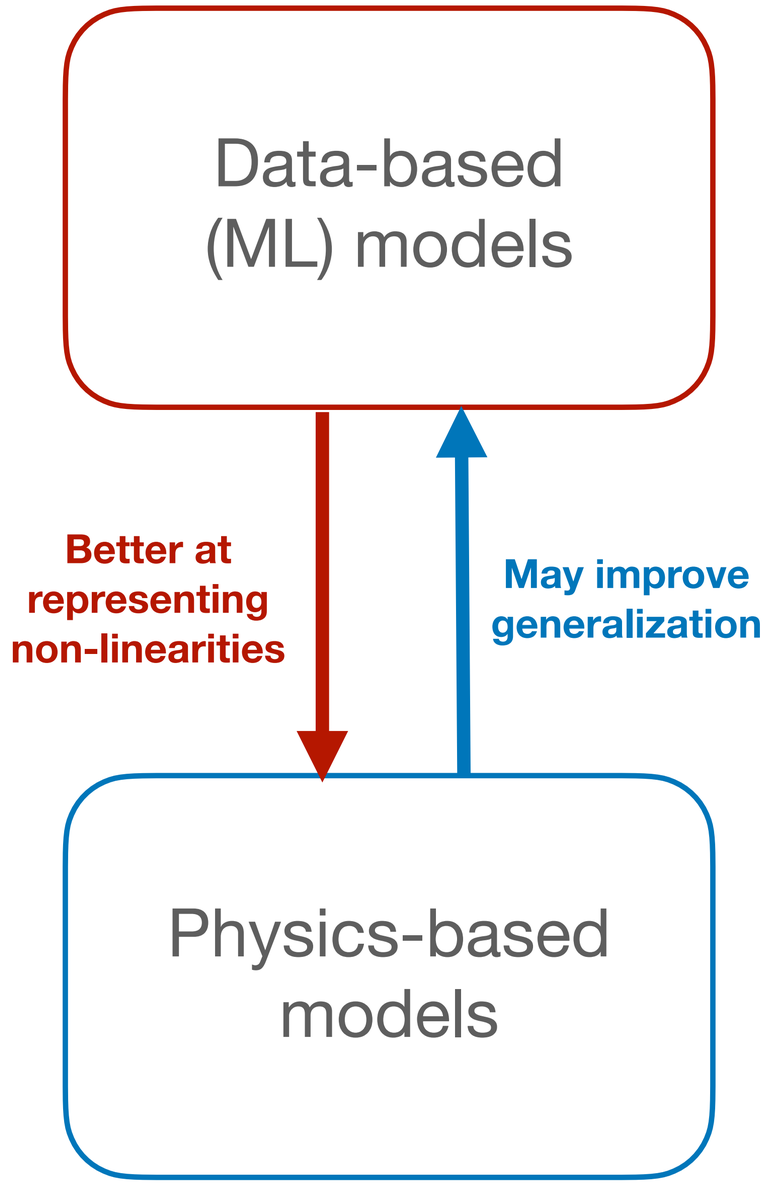


Predictive power limited by computational efficiency and accuracy

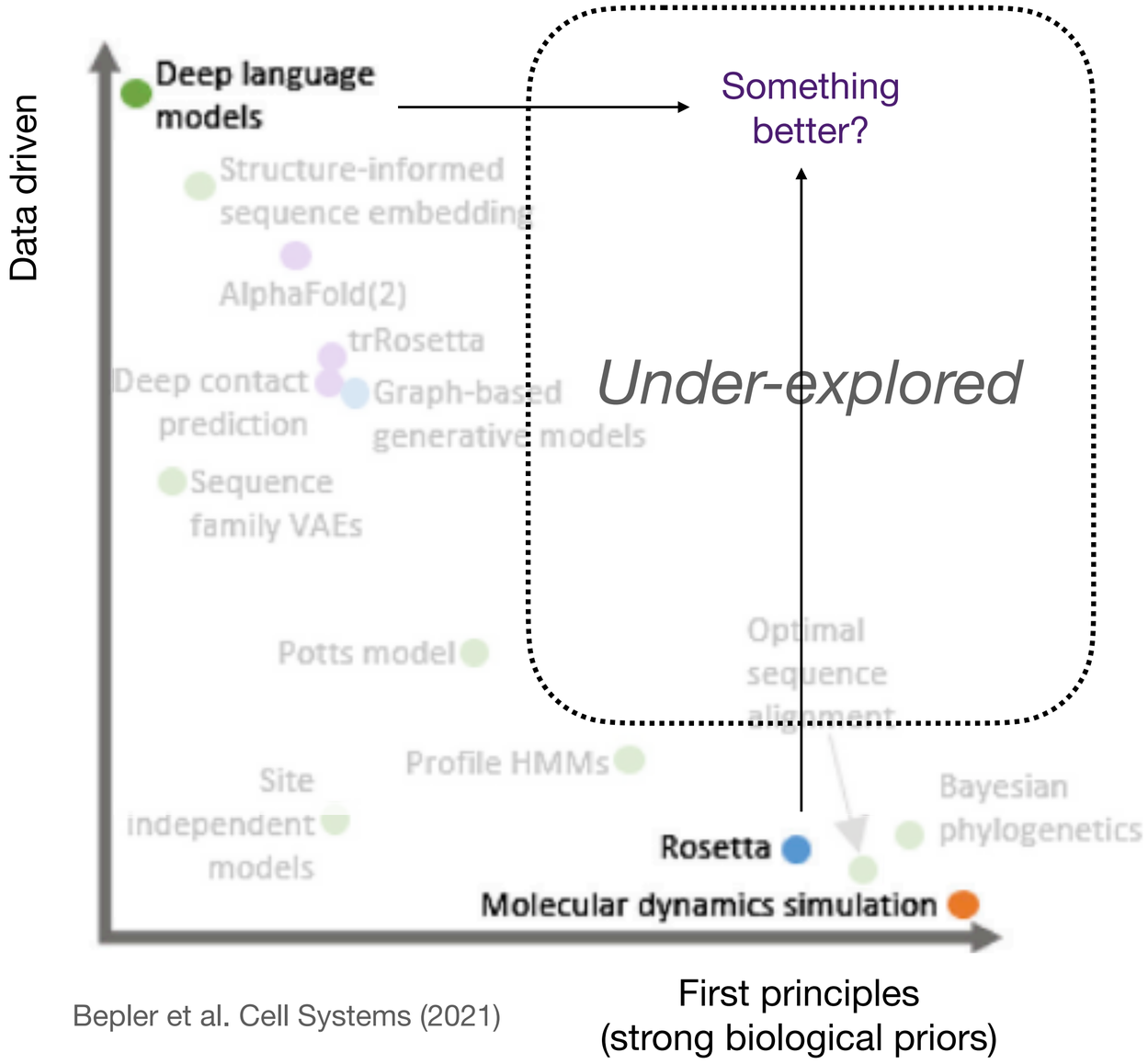
Physics and ML may be complementary



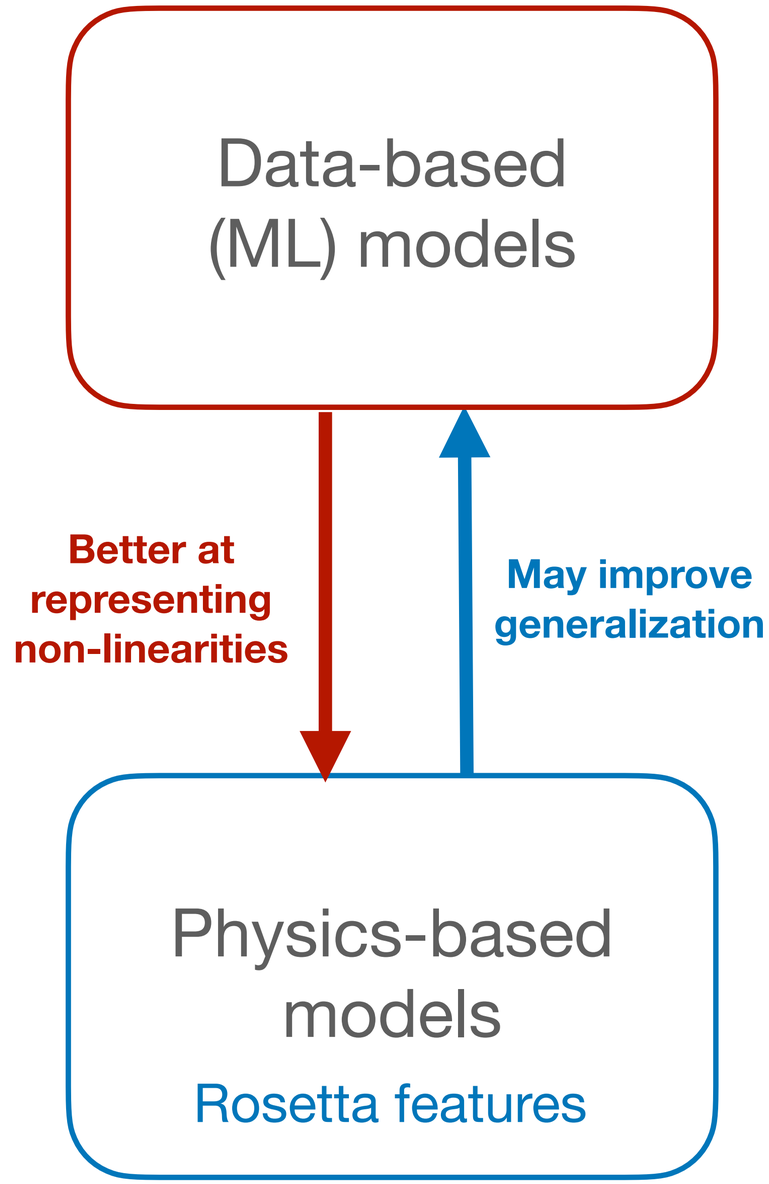
Bepler et al. Cell Systems (2021)



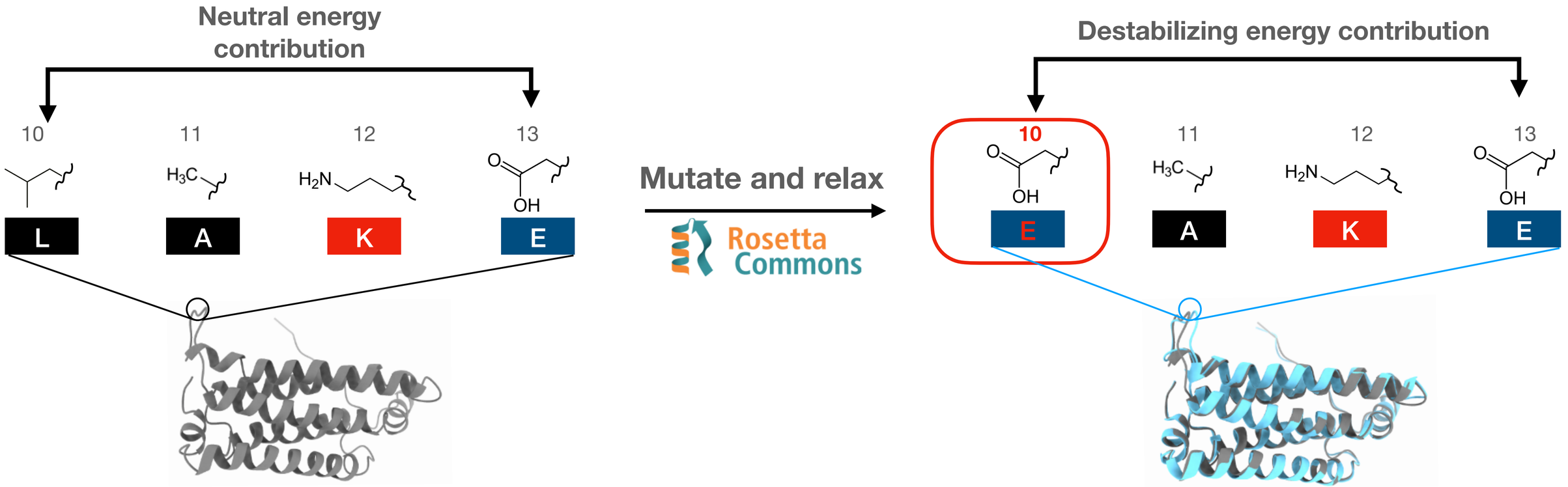
Physics-based features may improve ML models



Bepler et al. Cell Systems (2021)



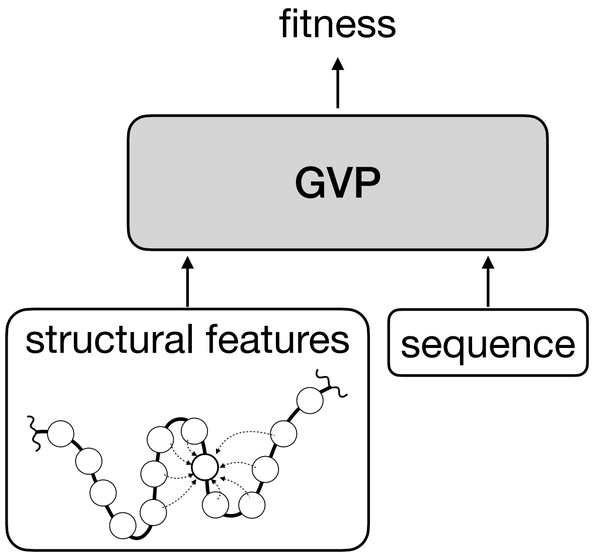
Use precomputed physics-based energies as input features

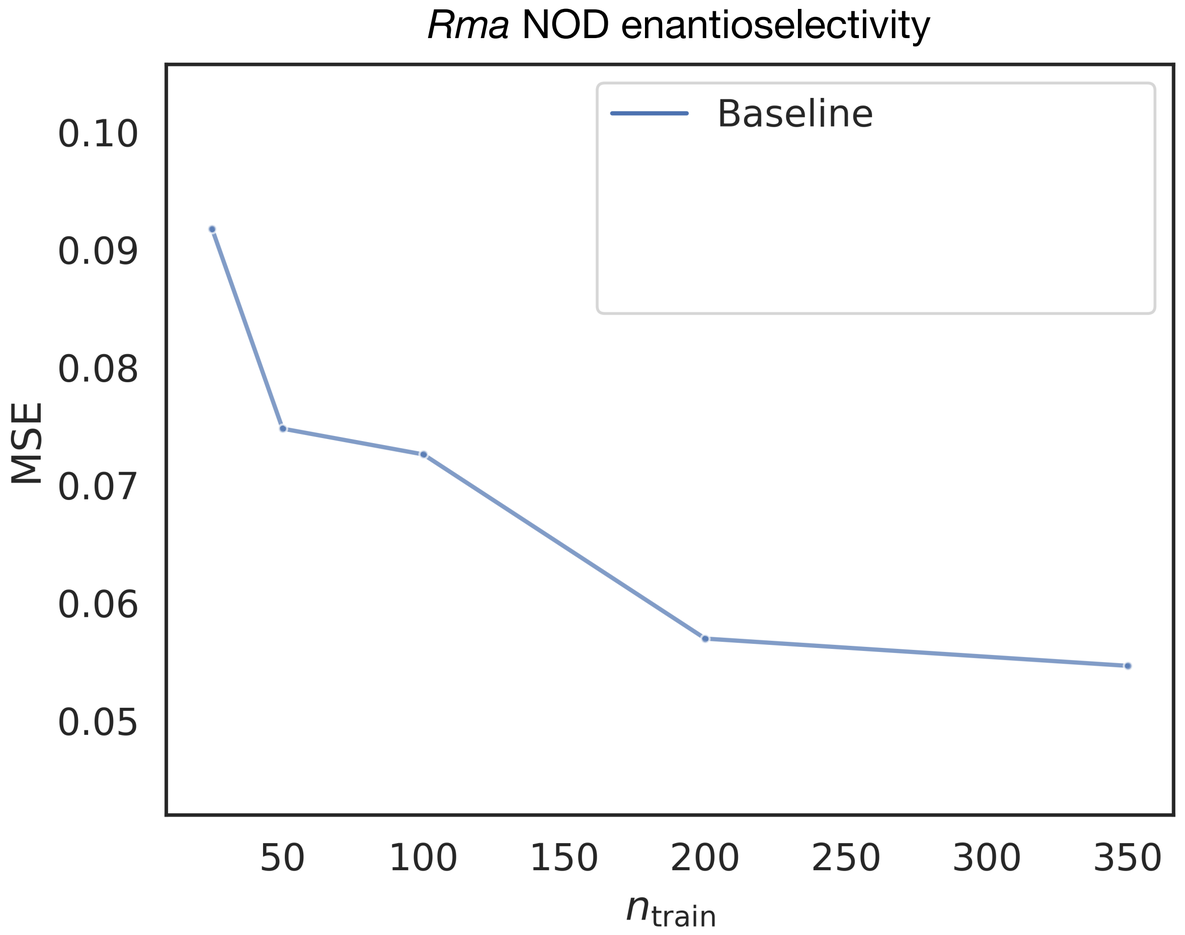


Reference protein

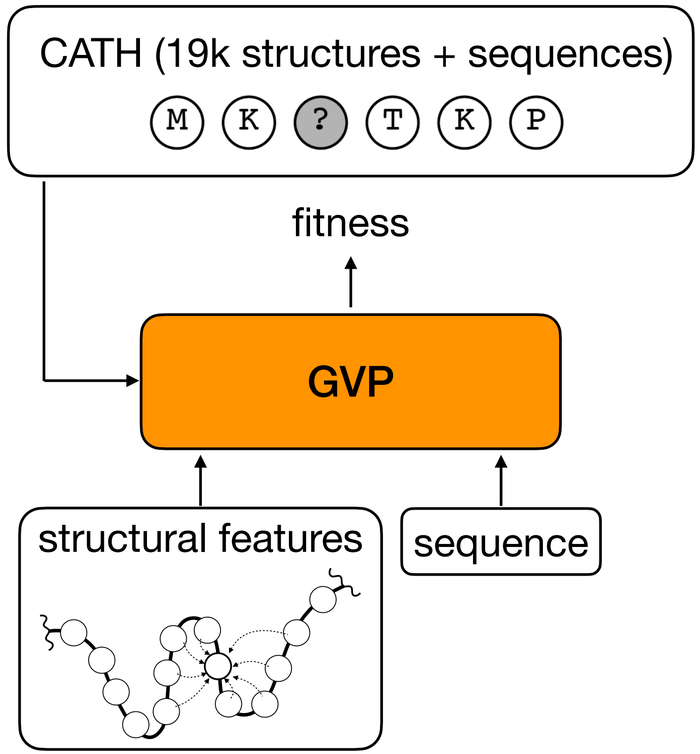
Protein variant: L10E

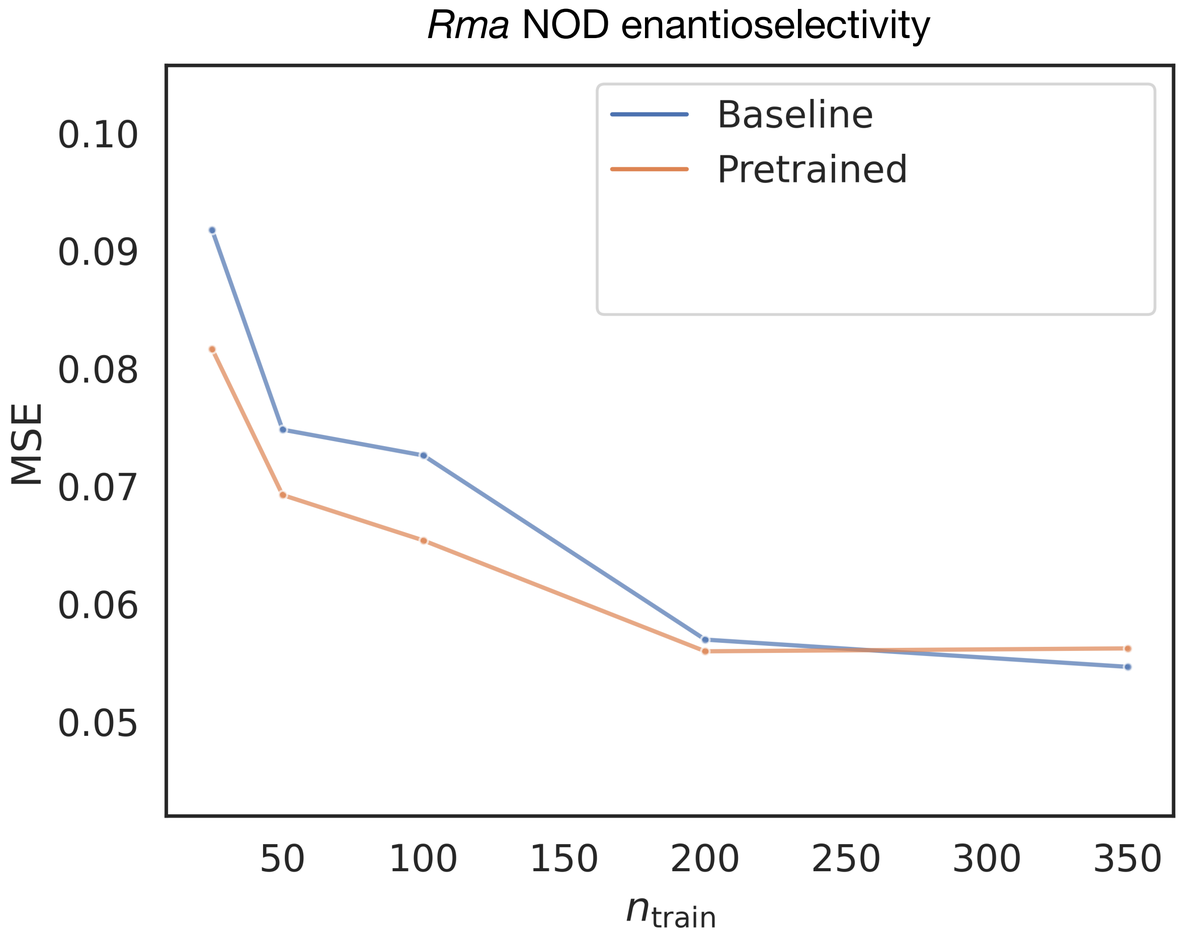
Energetics features improve performance in low-data regimes



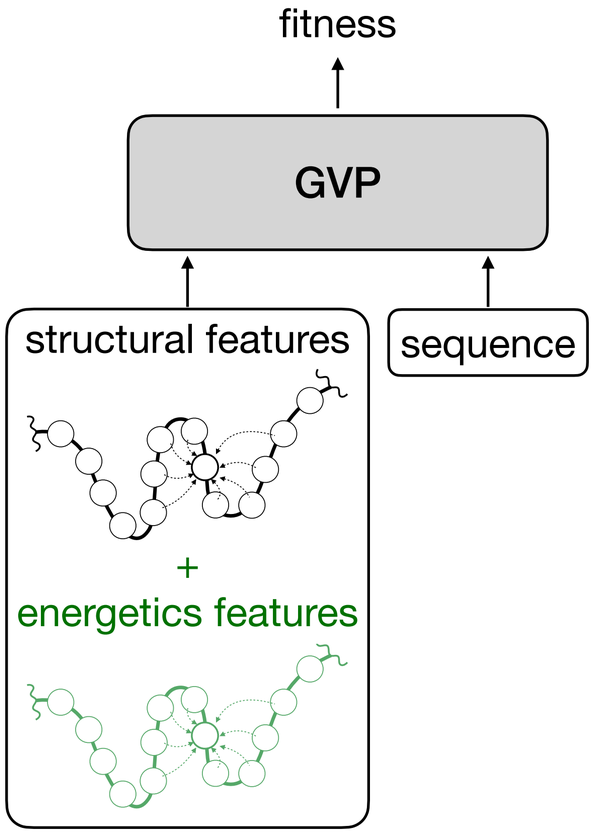


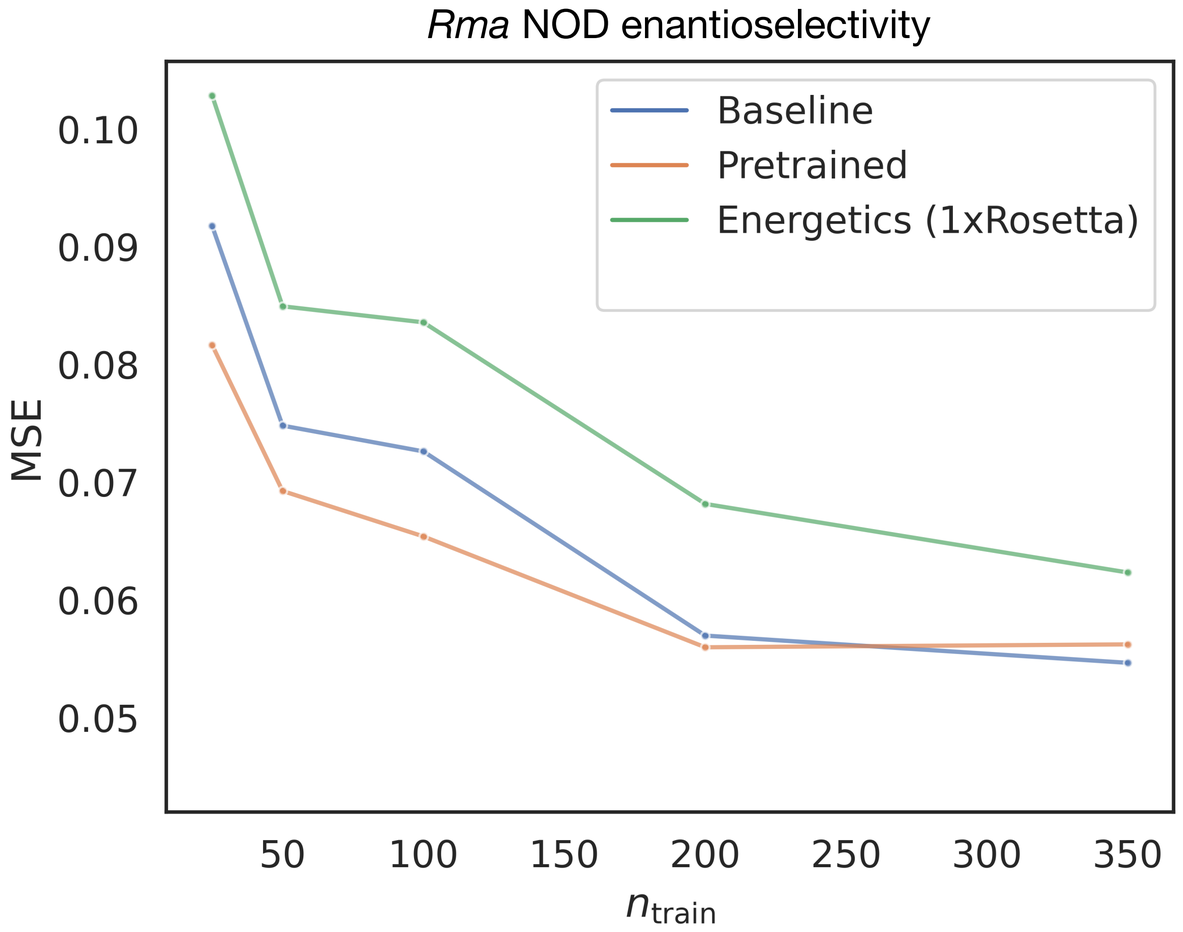
Energetics features improve performance in low-data regimes



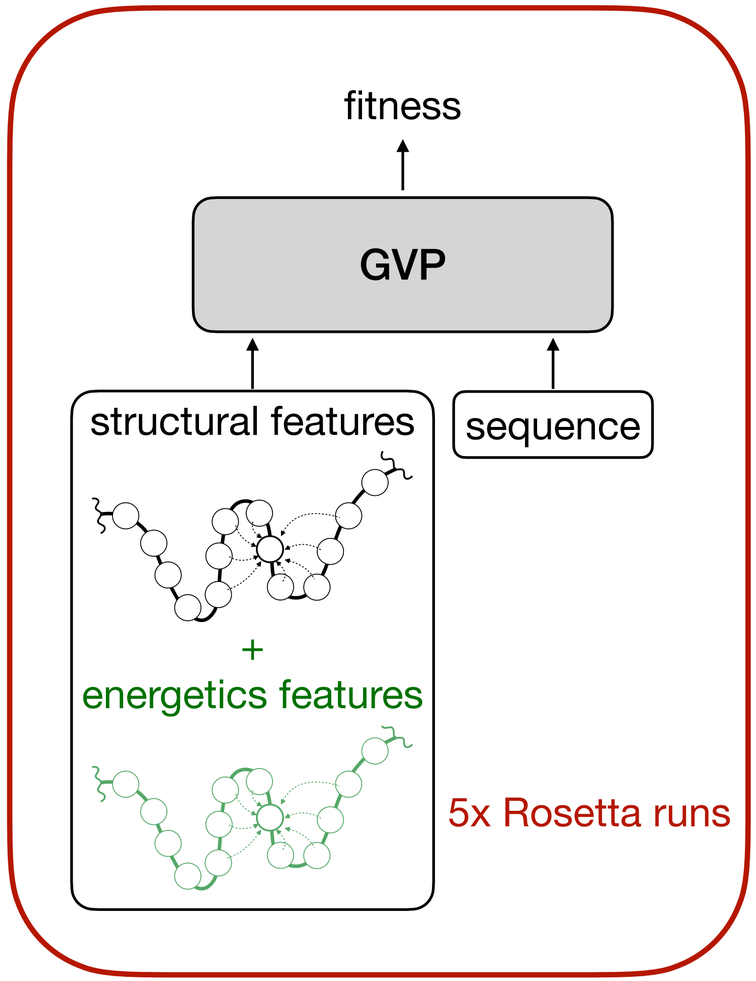


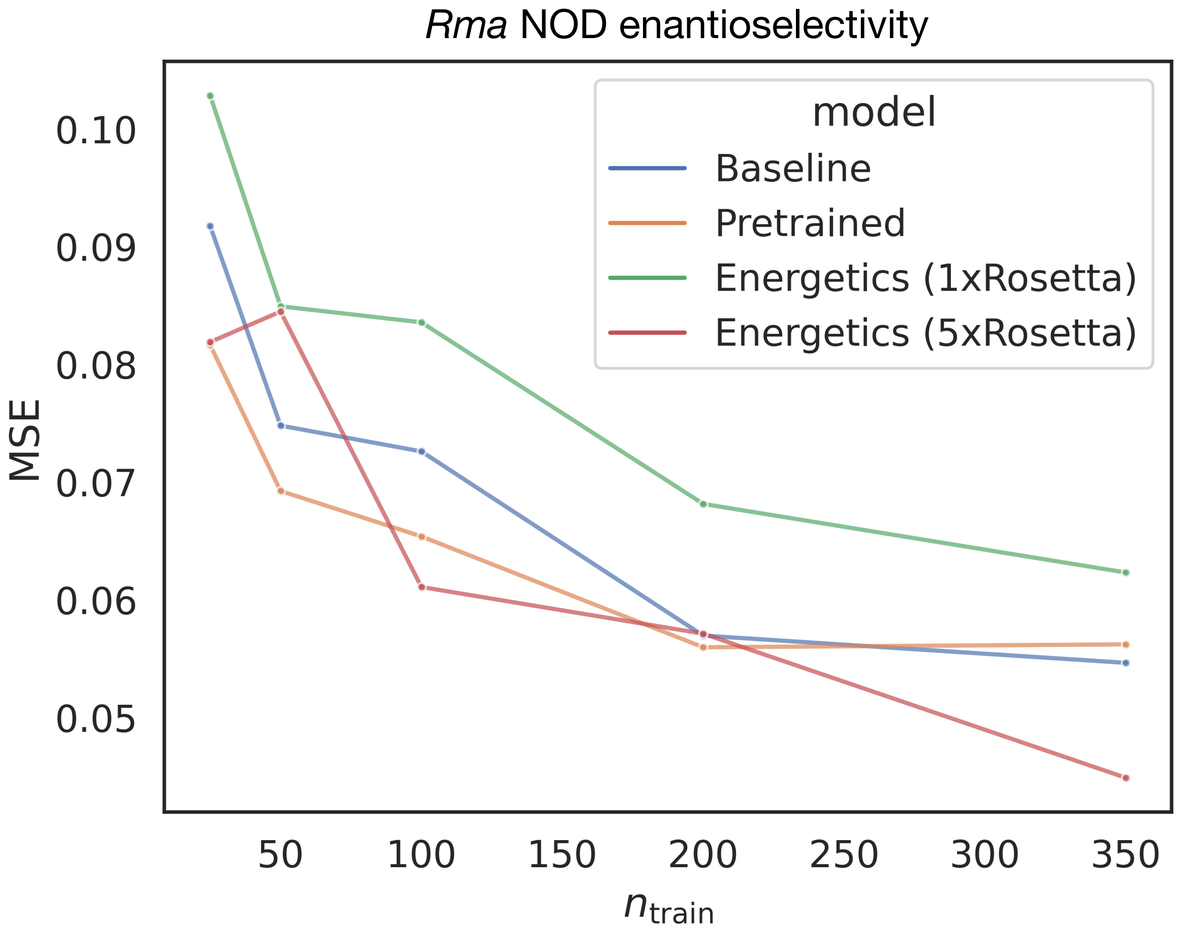
Energetics features improve performance in low-data regimes



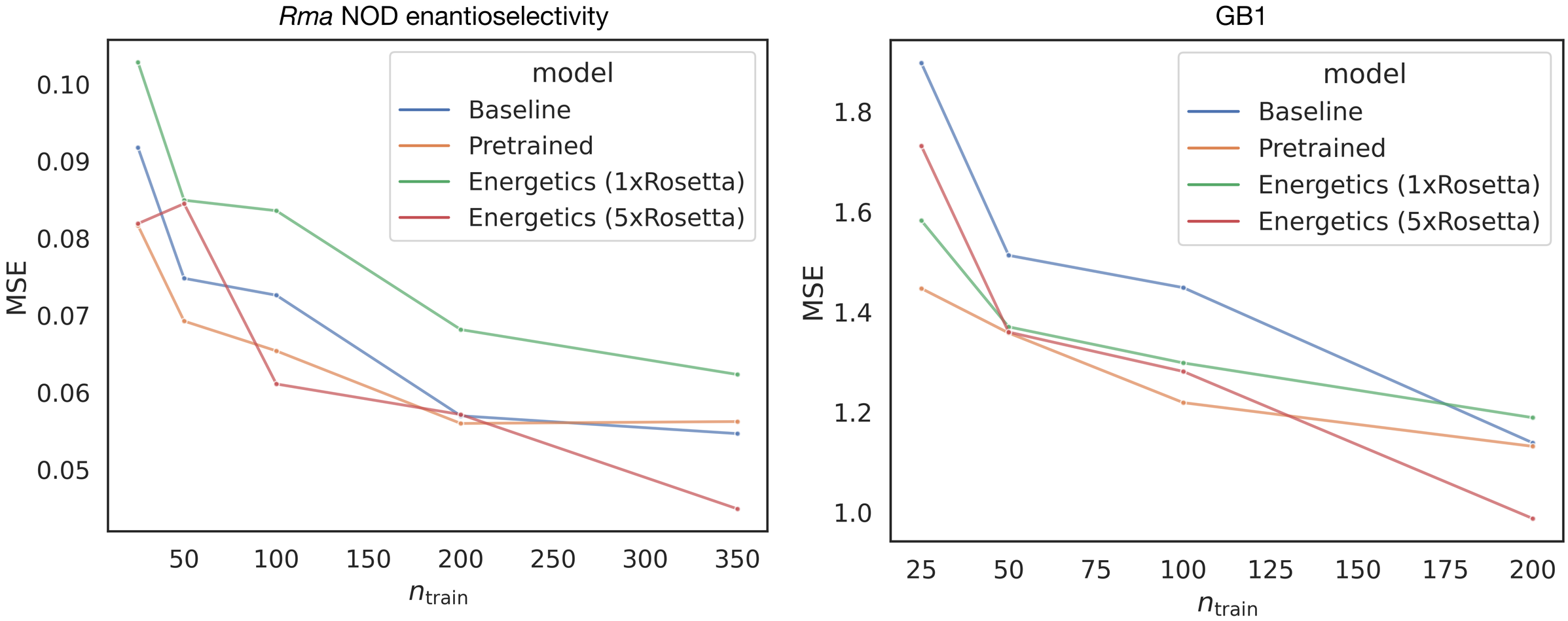


Energetics features improve performance in low-data regimes

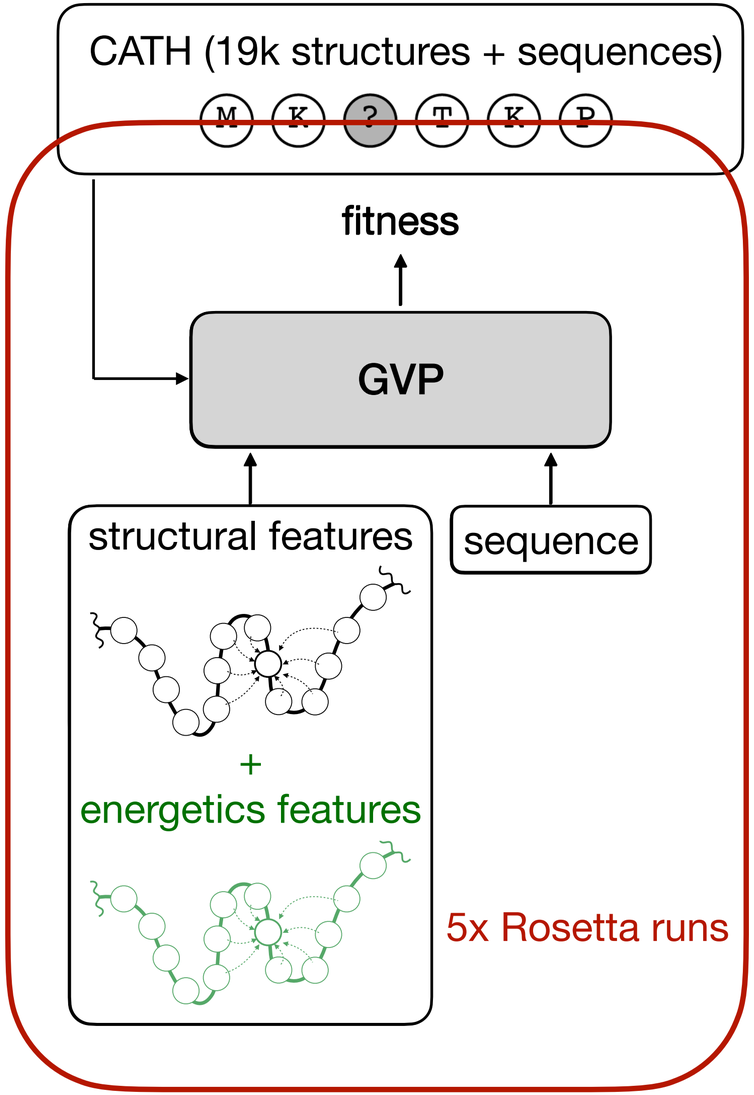




Energetics features improve performance in low-data regimes

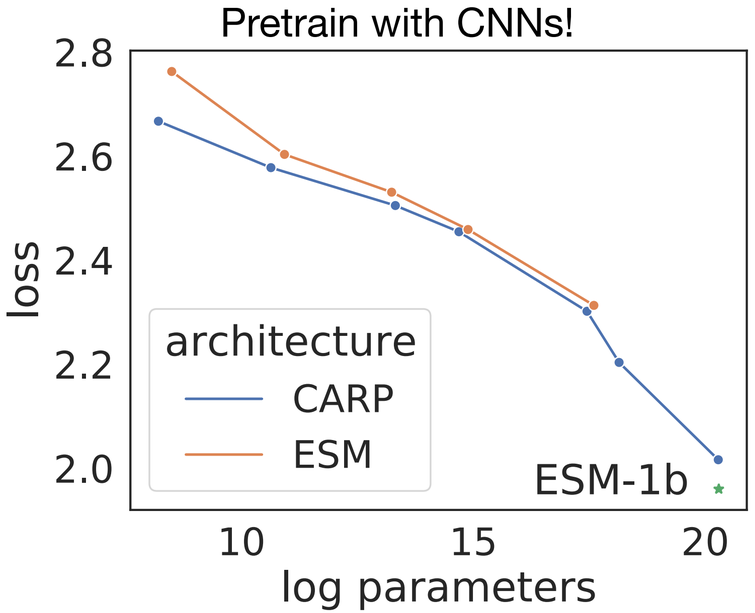


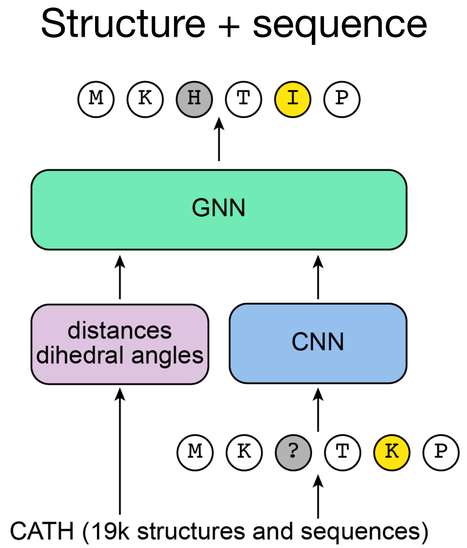
Energetics features match pretraining on OOD

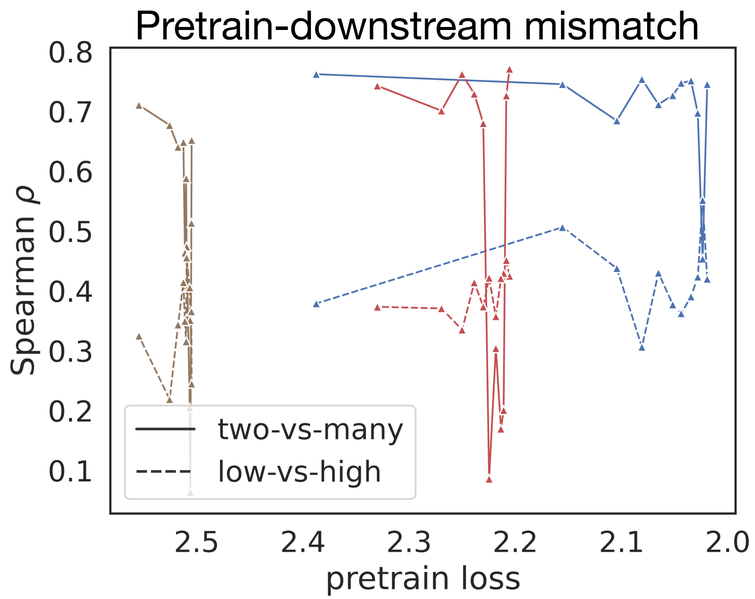


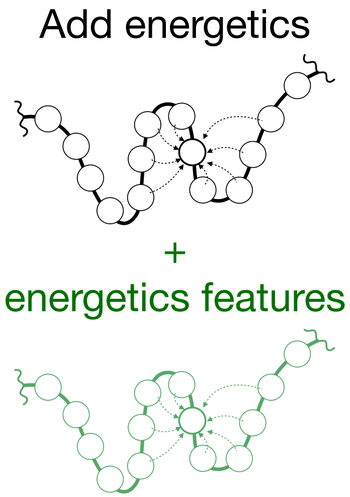
Model	GB1 2-vs-many (MSE)
Baseline	1.41±0.14
Pretrain	1.16±0.06
Energetics (1xRosetta)	1.32±0.13
Energetics (5xRosetta)	1.16±0.05

Use multimodal data to reduce pretrain-downstream mismatch

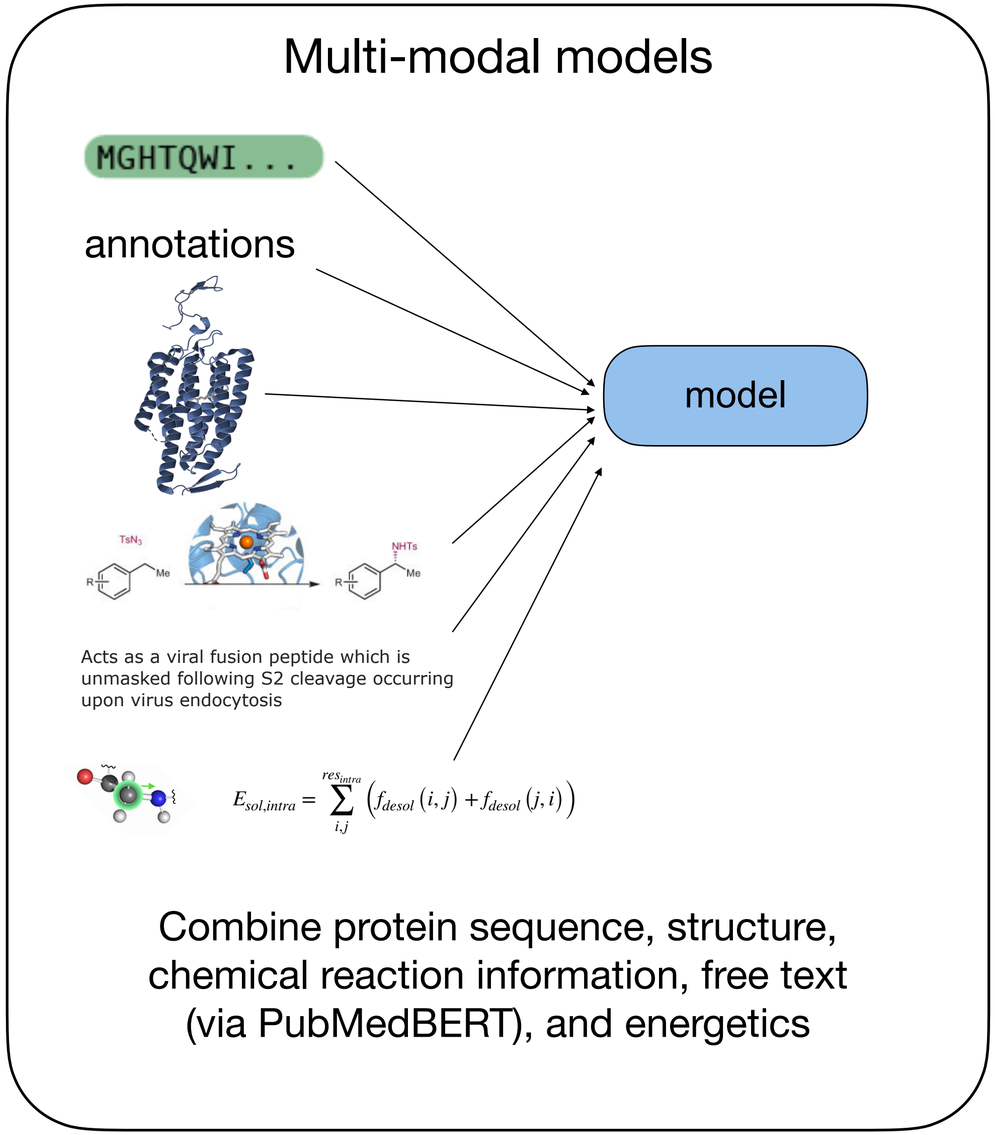




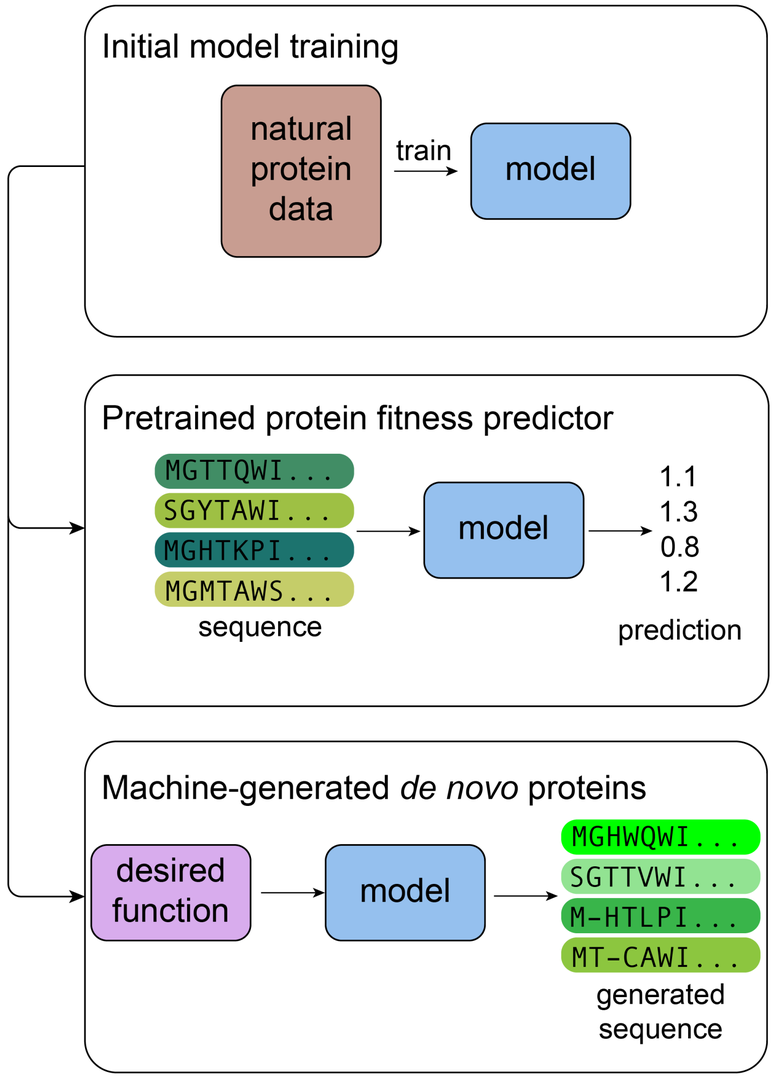


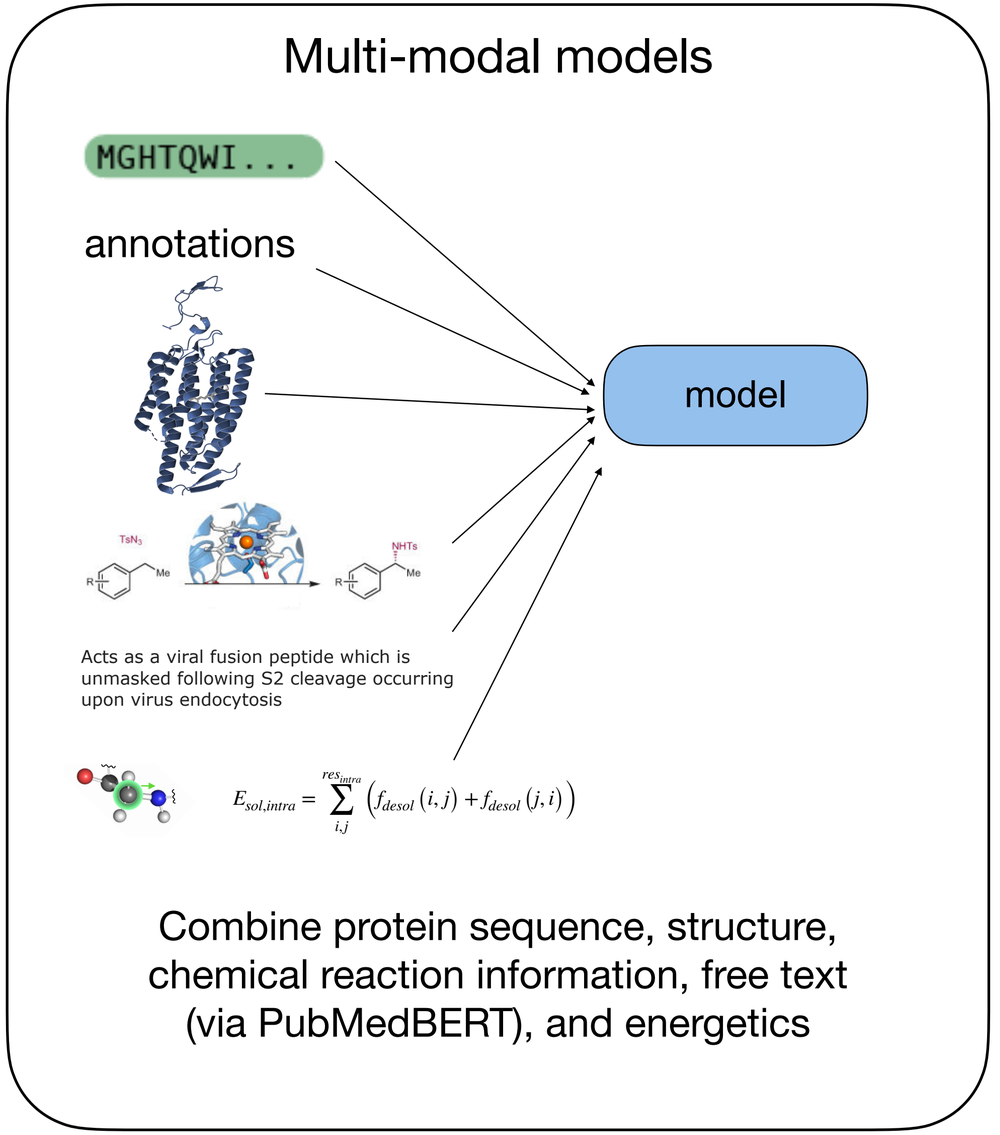


Can different modalities and pretraining tasks do better?



Can we generate functional proteins?





Acknowledgments








BioML at MSR New England