Pretraining and Representation Learning on Proteins

Kevin Kaichuang Yang
Microsoft Research New England
@KevinKaichuang

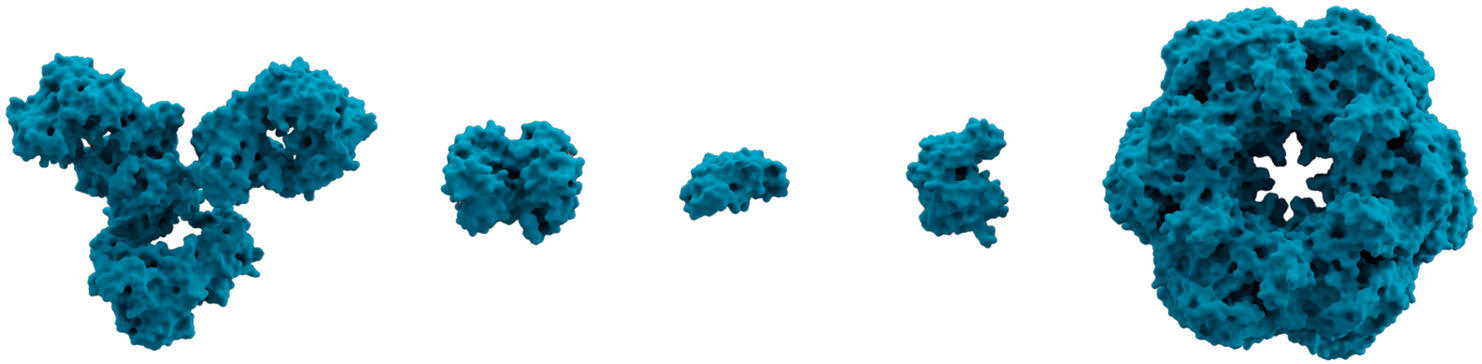
Proteins are biology’s actuators

Proteins are biology’s actuators

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

Proteins are biology’s actuators

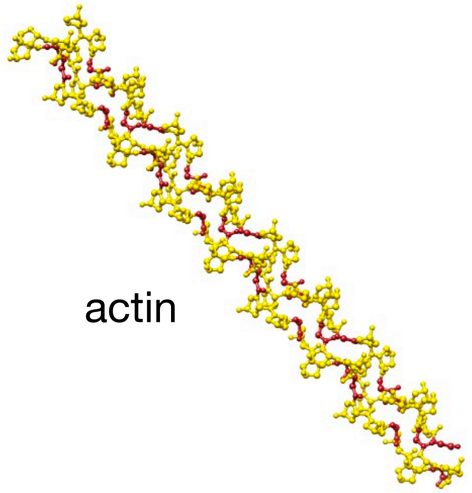
- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling



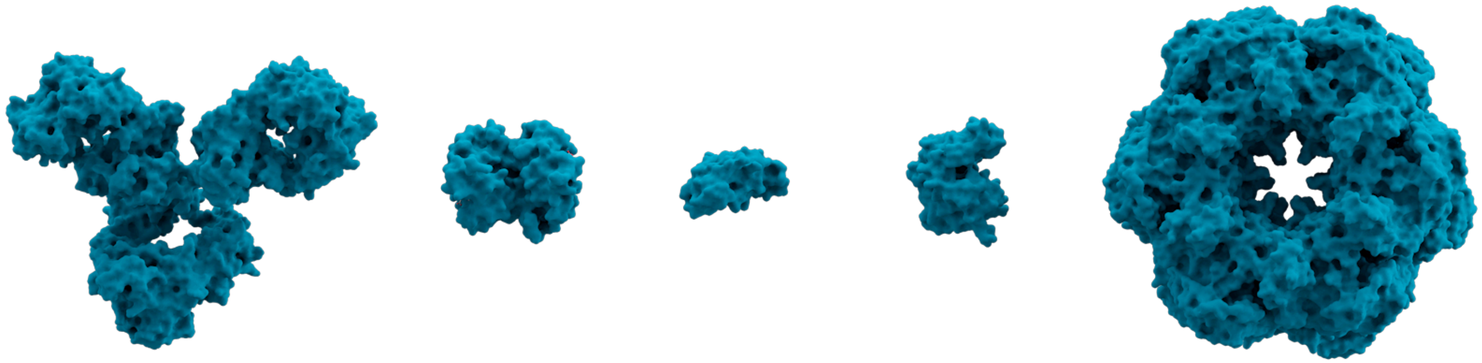
IgG hemoglobin insulin adenylate kinase glutamine synthetase

Proteins are biology's actuators

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



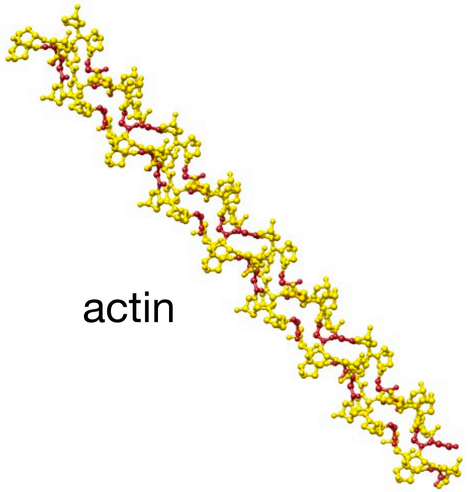
actin



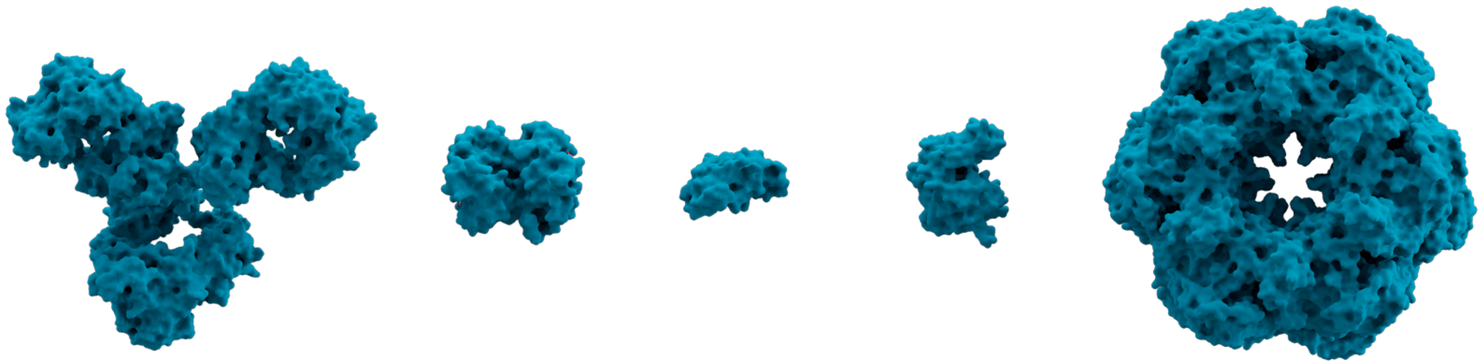
IgG hemoglobin insulin adenylate kinase glutamine synthetase

Proteins are biology's actuators

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



actin

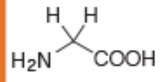


IgG hemoglobin insulin adenylate kinase glutamine synthetase

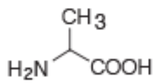
luciferase

Diversity arises from 20 building blocks

SMALL

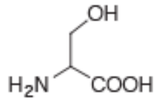


Glycine (Gly, G)
MW: 75.07

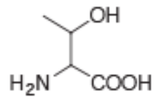


Alanine (Ala, A)
MW: 89.09

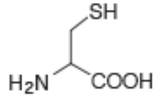
NUCLEOPHILIC



Serine (Ser, S)
MW: 105.09, $pK_a \sim 16$

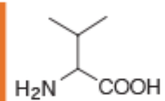


Threonine (Thr, T)
MW: 119.1, $pK_a \sim 16$

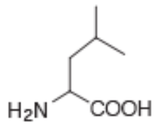


Cysteine (Cys, C)
MW: 121.2, $pK_a = 8.18$

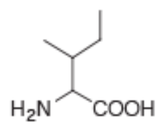
HYDROPHOBIC



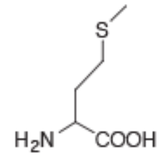
Valine (Val, V)
MW: 117.1



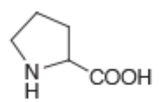
Leucine (Leu, L)
MW: 131.2



Isoleucine (Ile, I)
MW: 131.2

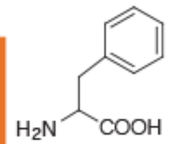


Methionine (Met, M)
MW: 149.2

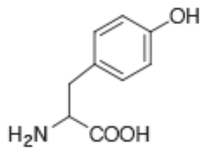


Proline (Pro, P)
MW: 115.1

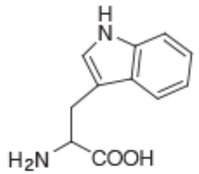
AROMATIC



Phenylalanine (Phe, F)
MW: 165.2

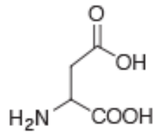


Tyrosine (Tyr, Y)
MW: 181.2, $pK_a = 10.46$

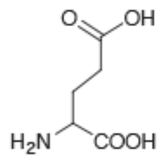


Tryptophan (Trp, W)
MW: 204.2

ACIDIC

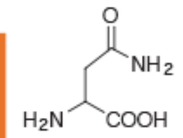


Aspartic Acid (Asp, D)
MW: 133.1, $pK_a = 3.9$

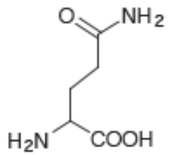


Glutamic Acid (Glu, E)
MW: 147.1, $pK_a = 4.07$

AMIDE

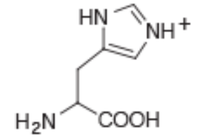


Asparagine (Asn, N)
MW: 132.1

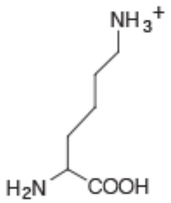


Glutamine (Gln, Q)
MW: 146.1

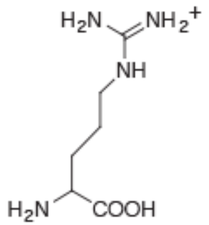
BASIC



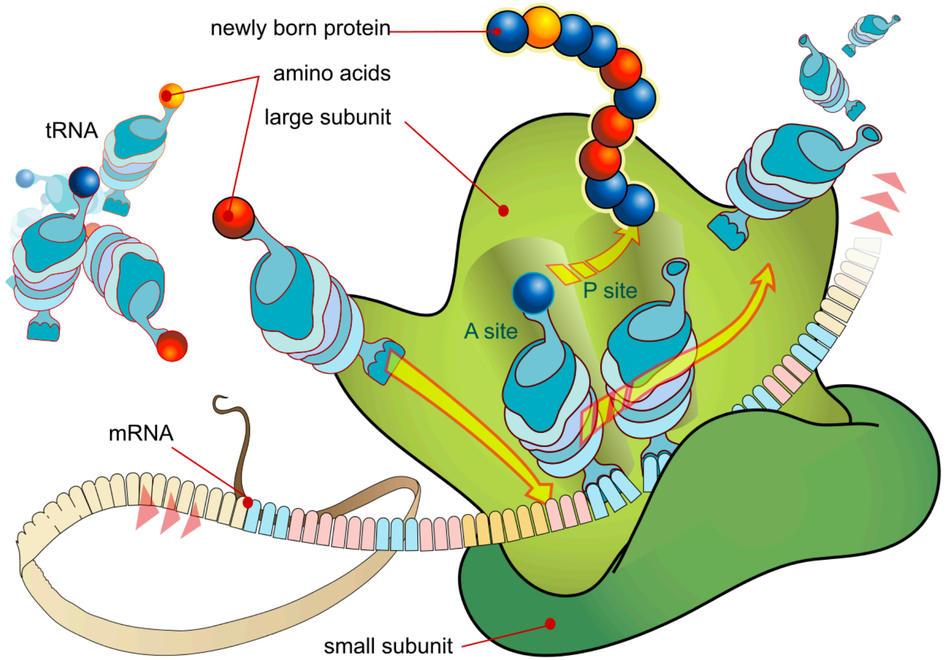
Histidine (His, H)
MW: 155.2, $pK_a = 6.04$



Lysine (Lys, K)
MW: 146.2, $pK_a = 10.79$



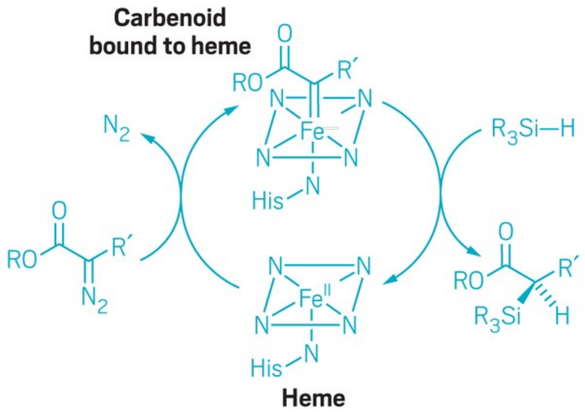
Arginine (Arg, R)
MW: 174.2, $pK_a = 12.48$



Why design proteins?

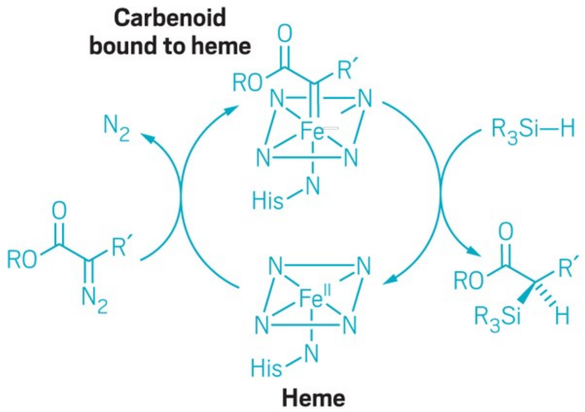
Why design proteins?

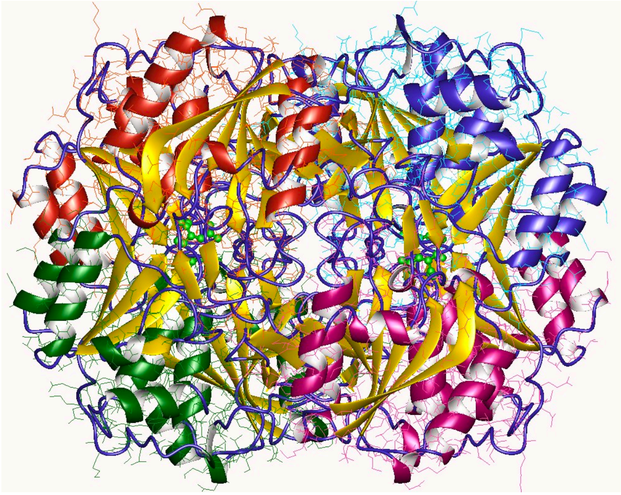
- New chemistry



Why design proteins?

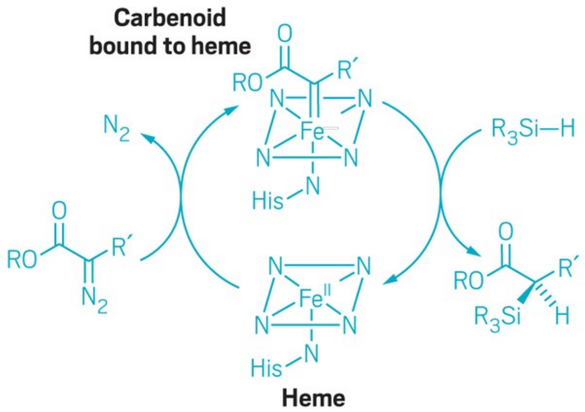
- New chemistry
- Therapeutics

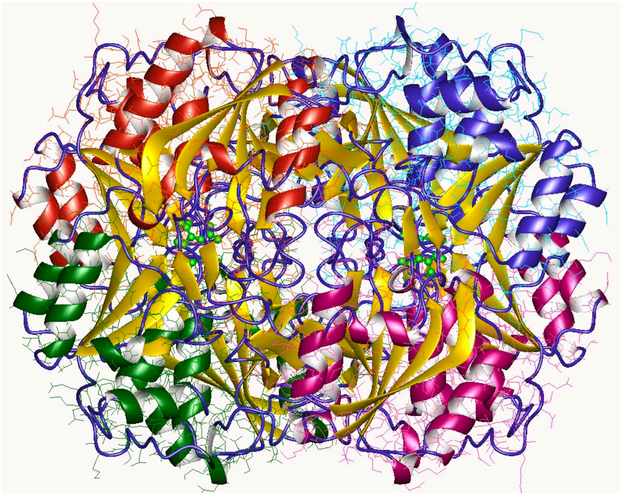


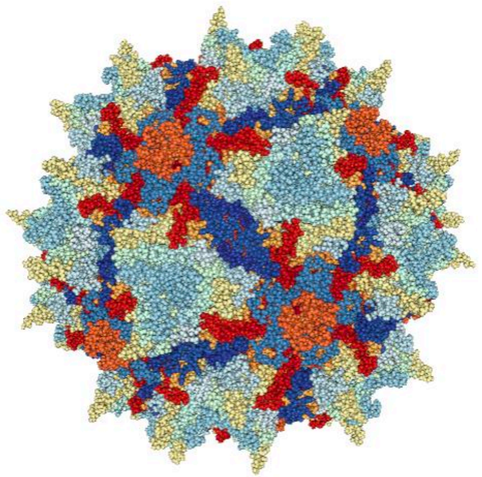


Why design proteins?

- New chemistry
- Therapeutics
- To learn how they work

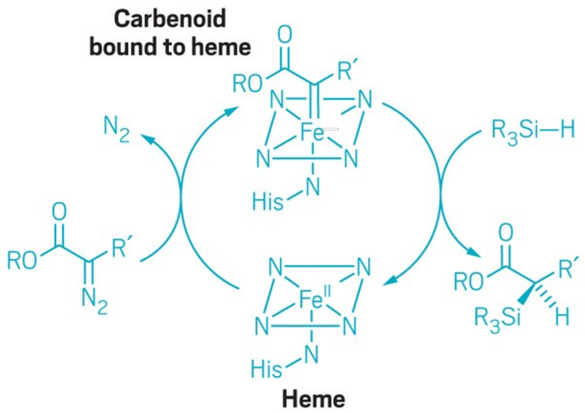


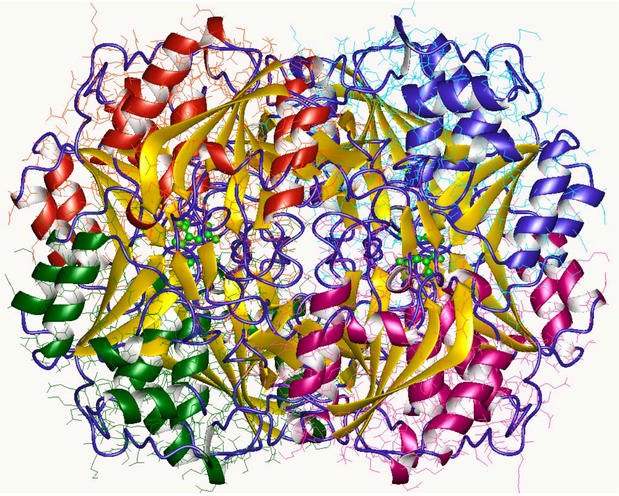


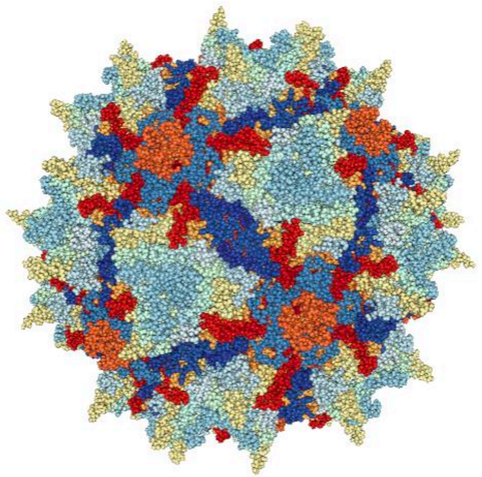


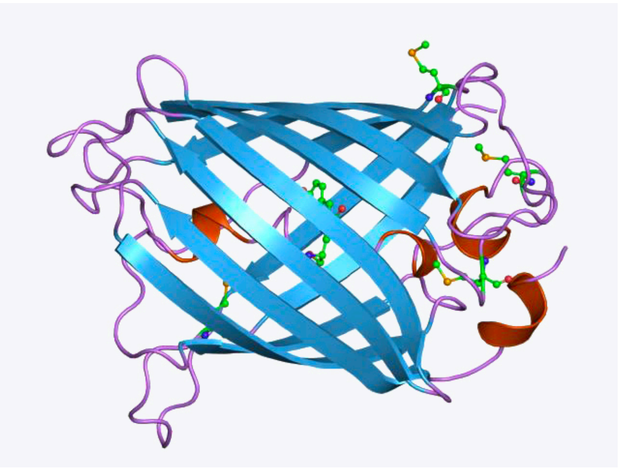
Why design proteins?

- New chemistry
- Therapeutics
- To learn how they work
- Molecular tools









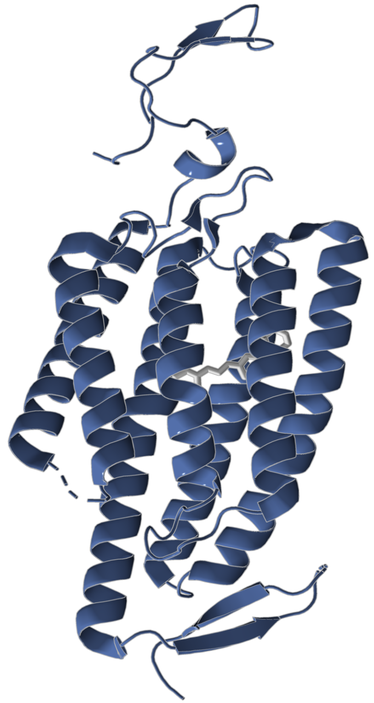
The protein design problem

The protein design problem

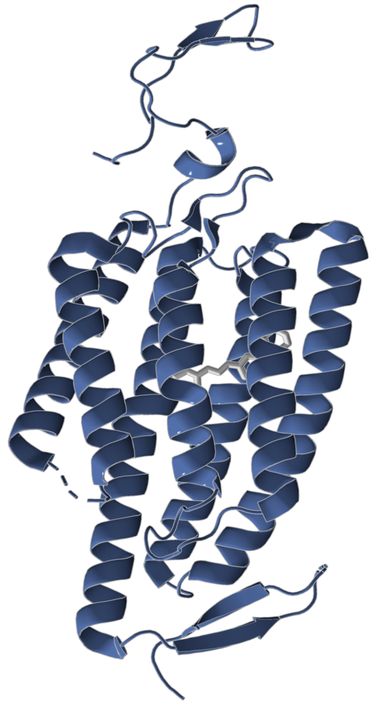
MGTGDHDD...

The protein design problem

MGTGDHDD...



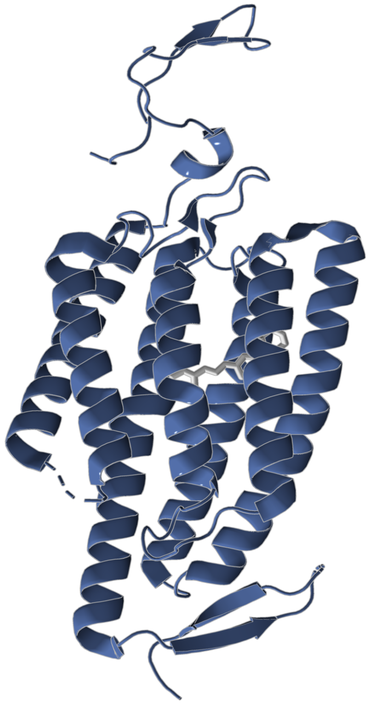
The protein design problem

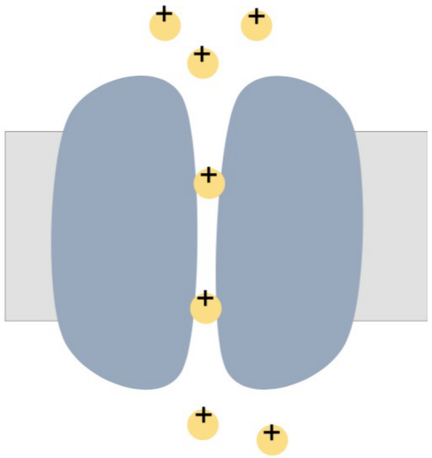
MGTGDHDD... ? 

The protein design problem

MGTGDHDD...

?

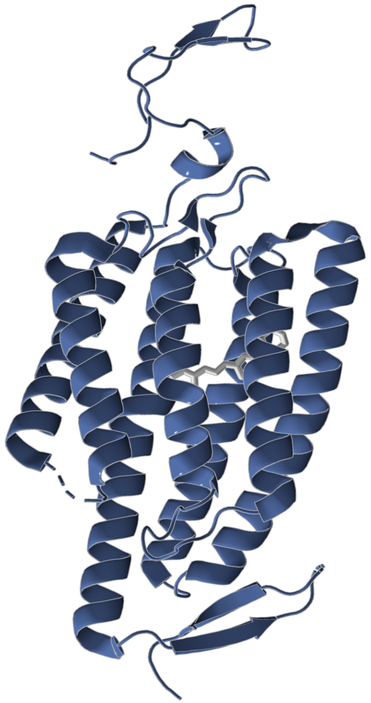




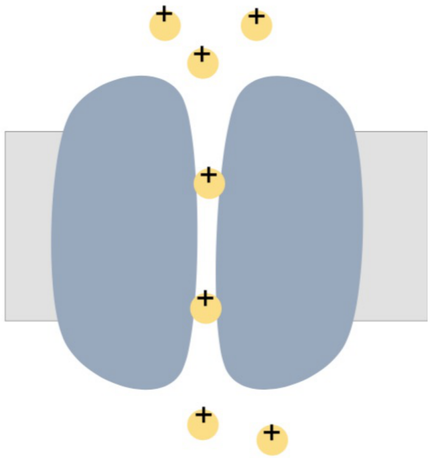
The protein design problem

MGTGDHDD...

?

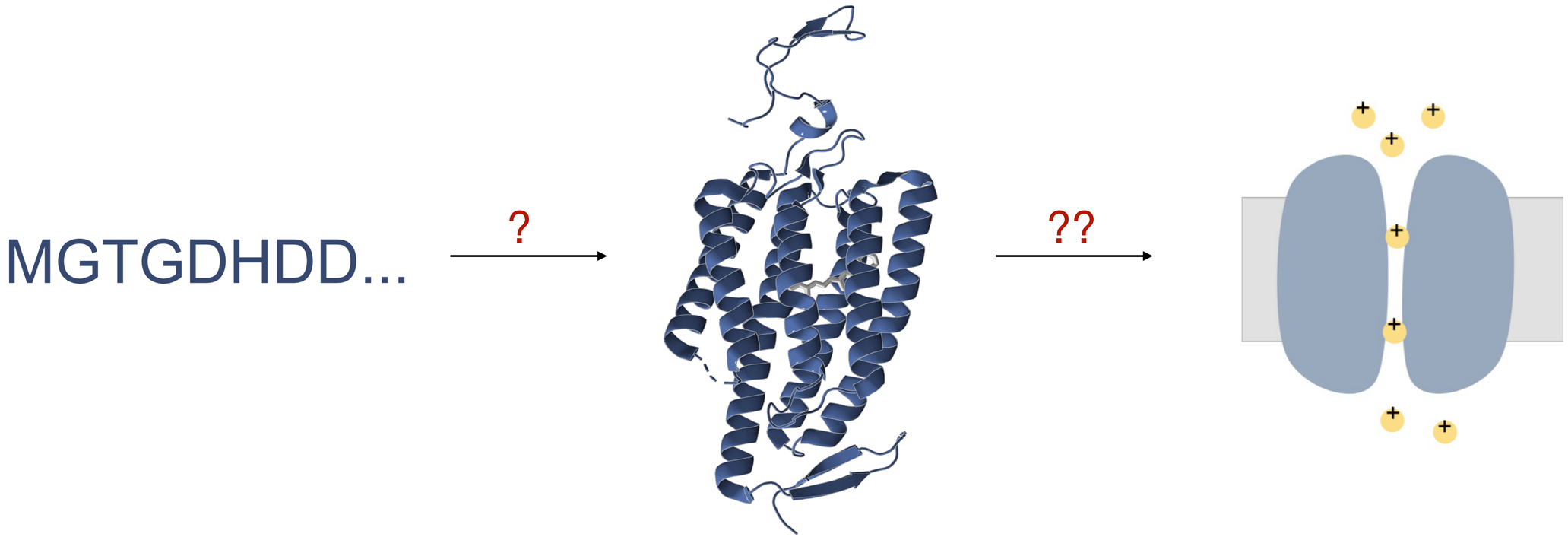


??



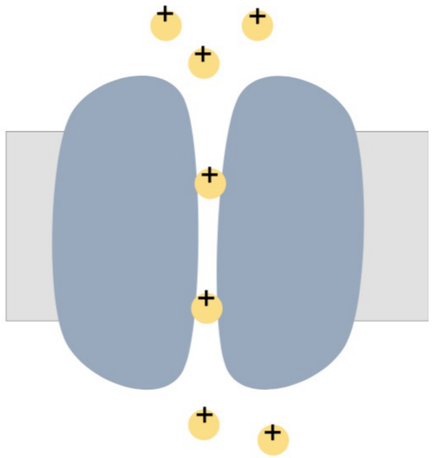
The protein design problem

MGTGDHDD...



What sequence will give the desired function?

The protein design problem

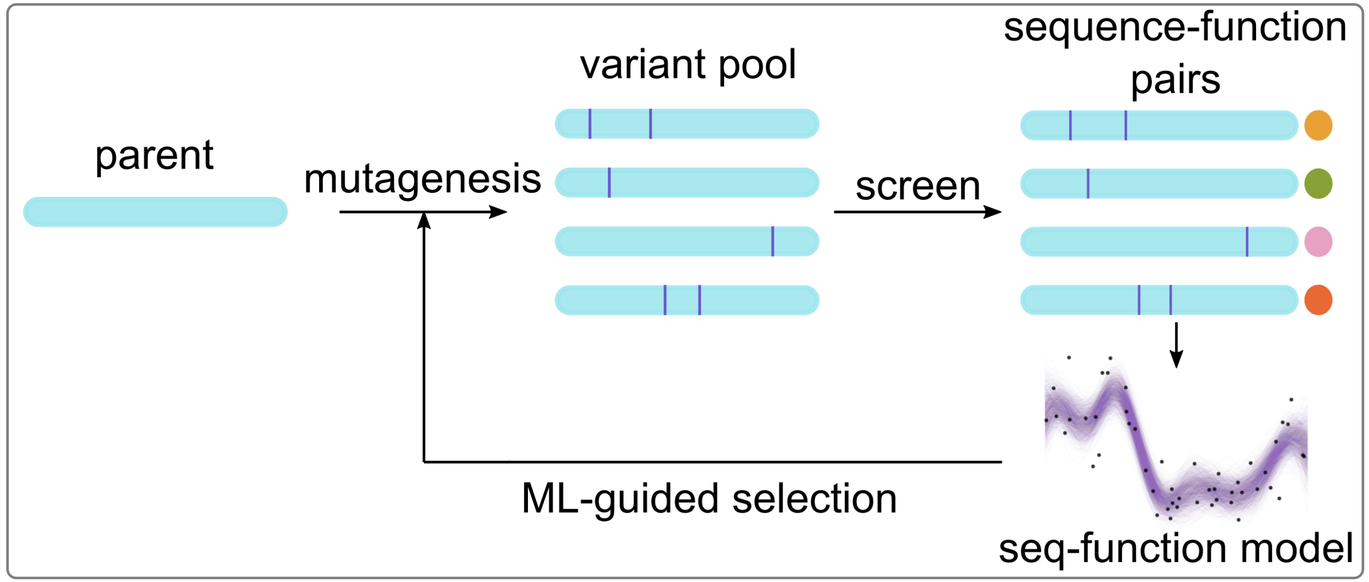


???

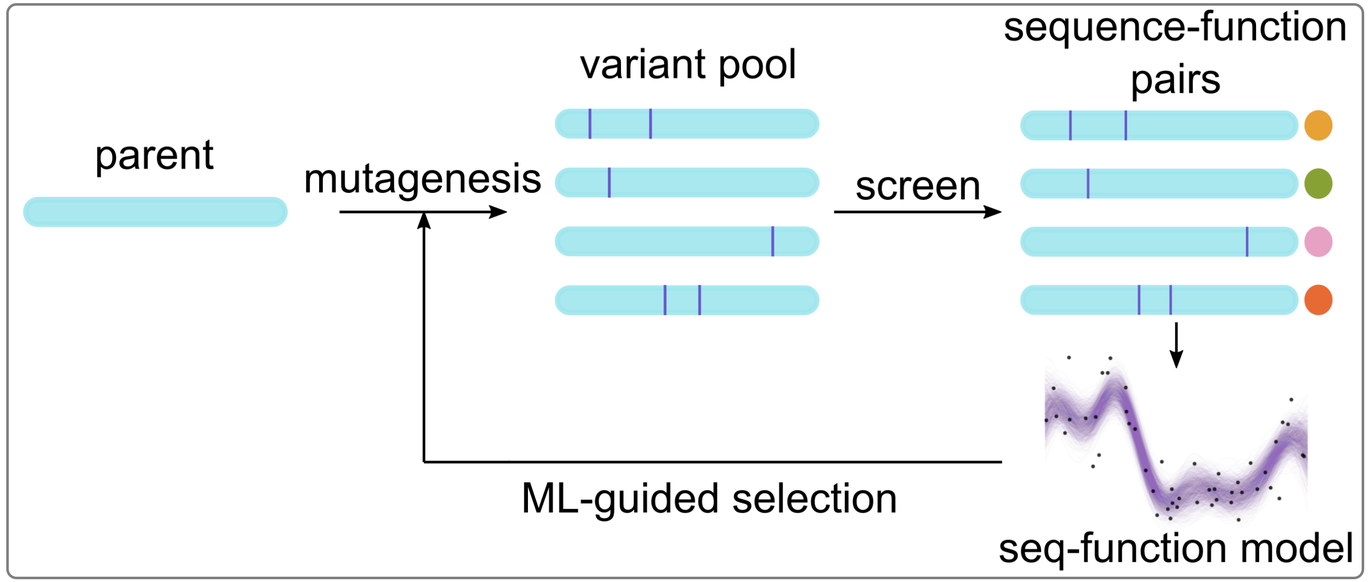
MGTGDHDD...

What sequence will give the desired function?

Machine learning enables optimization with fewer measurements

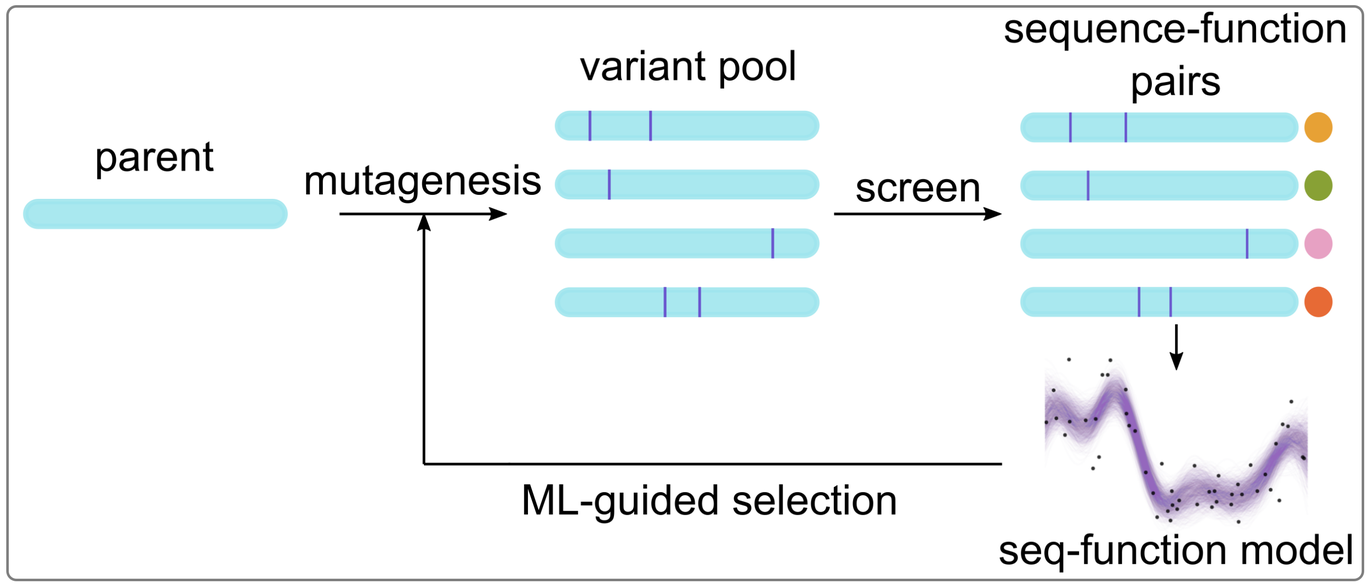


Machine learning enables optimization with fewer measurements

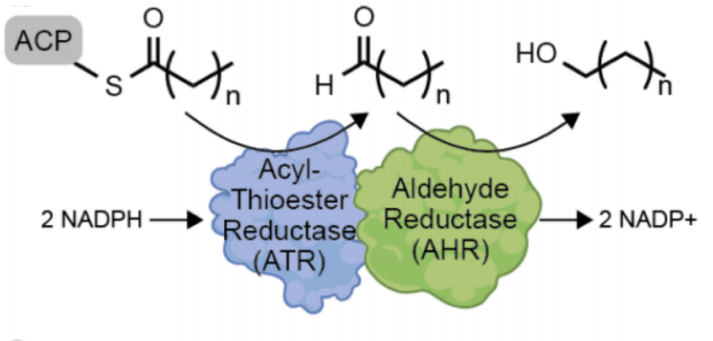


Channelrhodopsin current
~200 measurements (Bedbrook 2019)

Machine learning enables optimization with fewer measurements



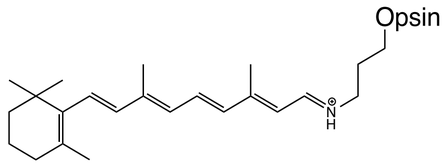
Channelrhodopsin current
~200 measurements (Bedbrook 2019)



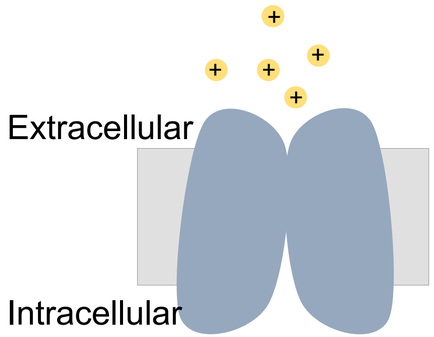
acyl-ACP reductase fatty alcohol synthesis
~100 measurements (Greenhalgh 2021)

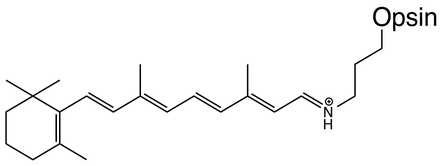
Channelrhodopsins: light-gated ion channels

Channelrhodopsins: light-gated ion channels

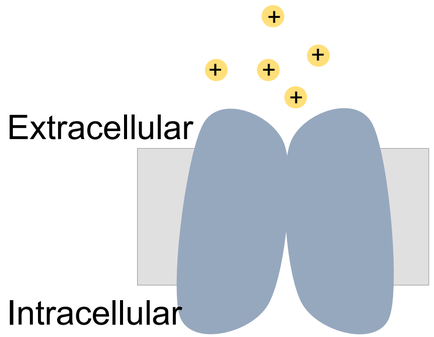


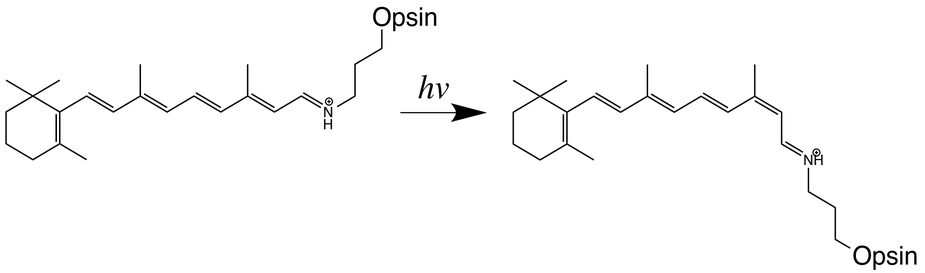
Channelrhodopsins: light-gated ion channels



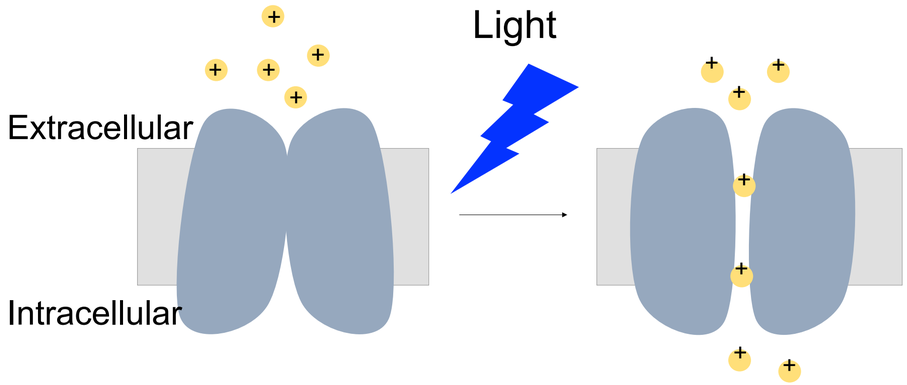


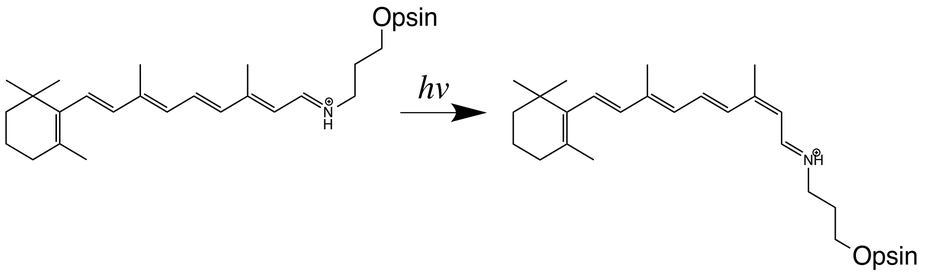
Channelrhodopsins: light-gated ion channels



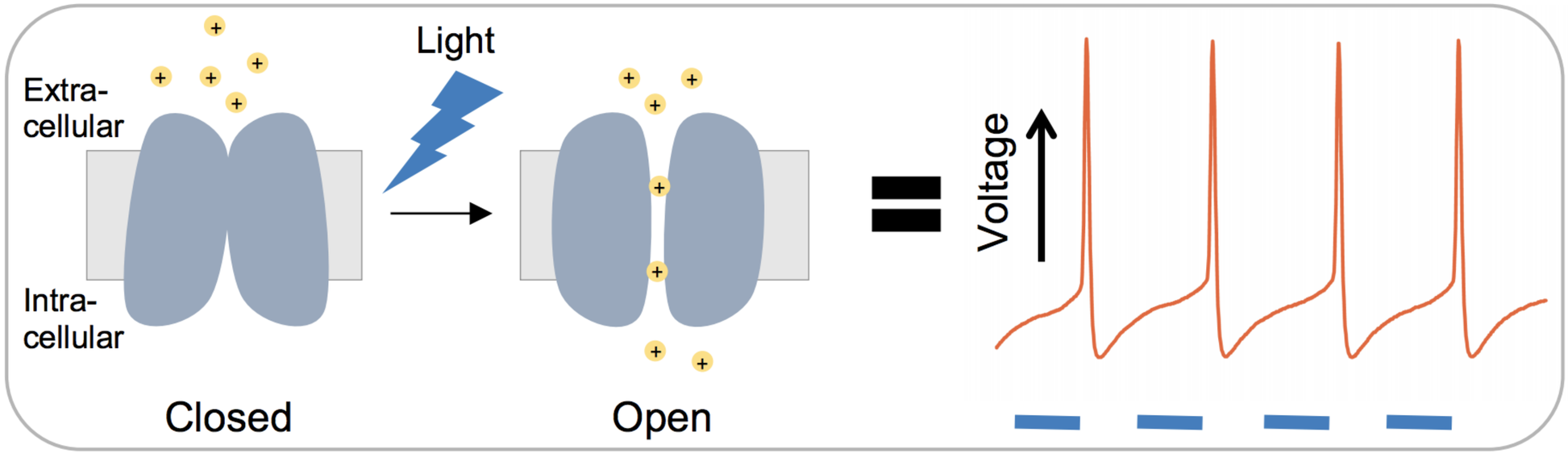


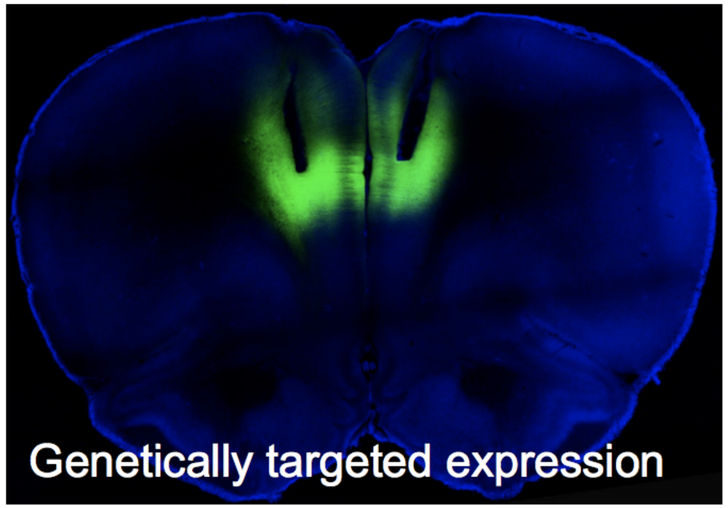
Channelrhodopsins: light-gated ion channels





ChRs are optogenetic tools



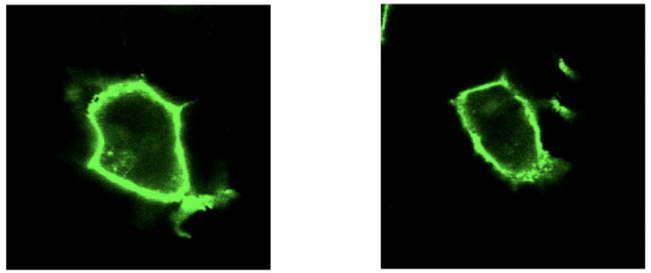


Optogenetics, Nature Method of the year, 2010

Multiple engineering goals

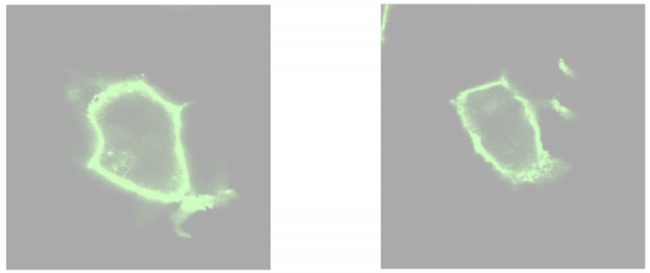
Multiple engineering goals

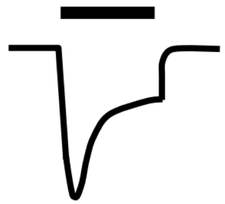
- Heterologous membrane localization



Multiple engineering goals

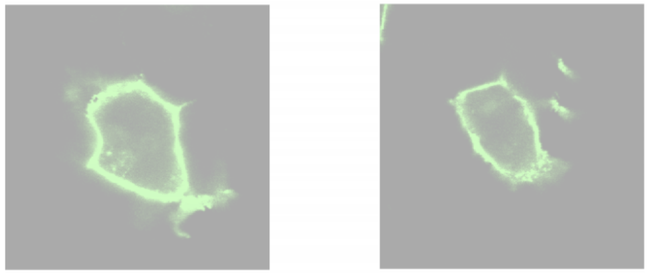
- Heterologous membrane localization
- Increased sensitivity

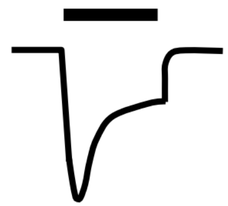




Multiple engineering goals

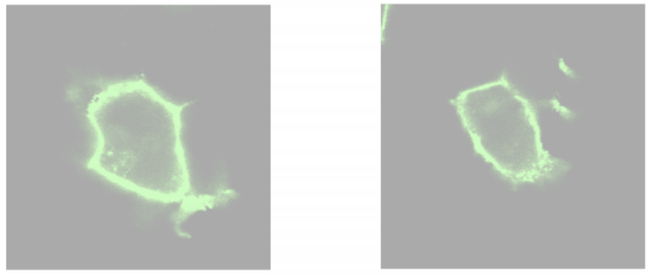
- Heterologous membrane localization
- Increased sensitivity
- Different activation wavelengths



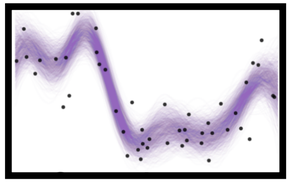


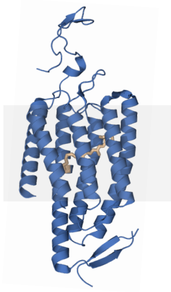
Multiple engineering goals

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



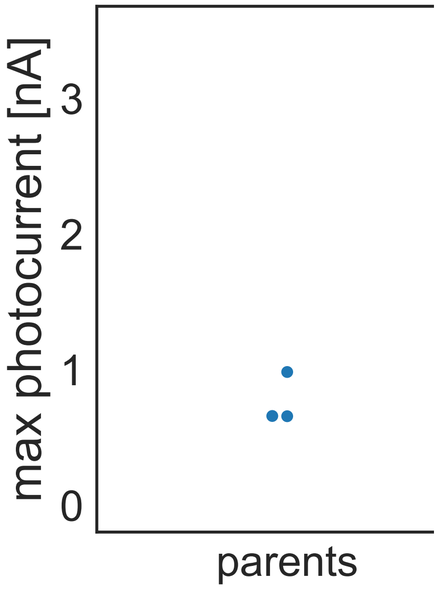
Machine learning enables optimization with fewer measurements

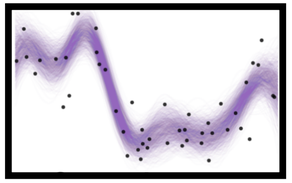


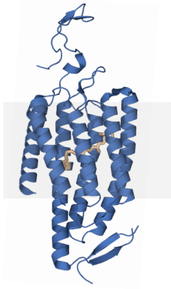


Channelrhodopsin current
~200 measurements (Bedbrook 2019)

Machine learning enables optimization with fewer measurements

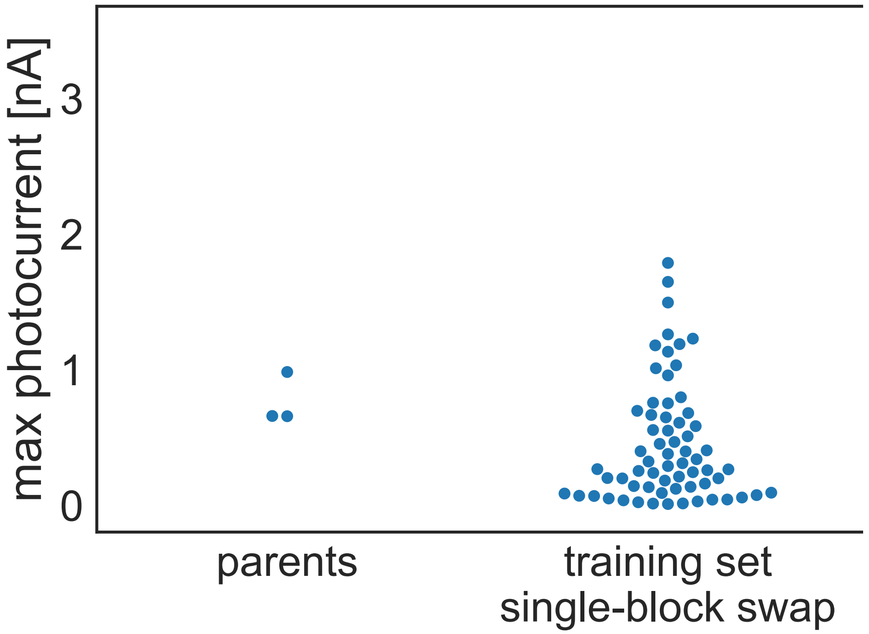


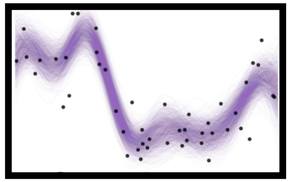


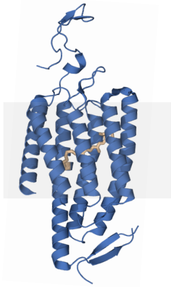


Channelrhodopsin current
~200 measurements (Bedbrook 2019)

Machine learning enables optimization with fewer measurements

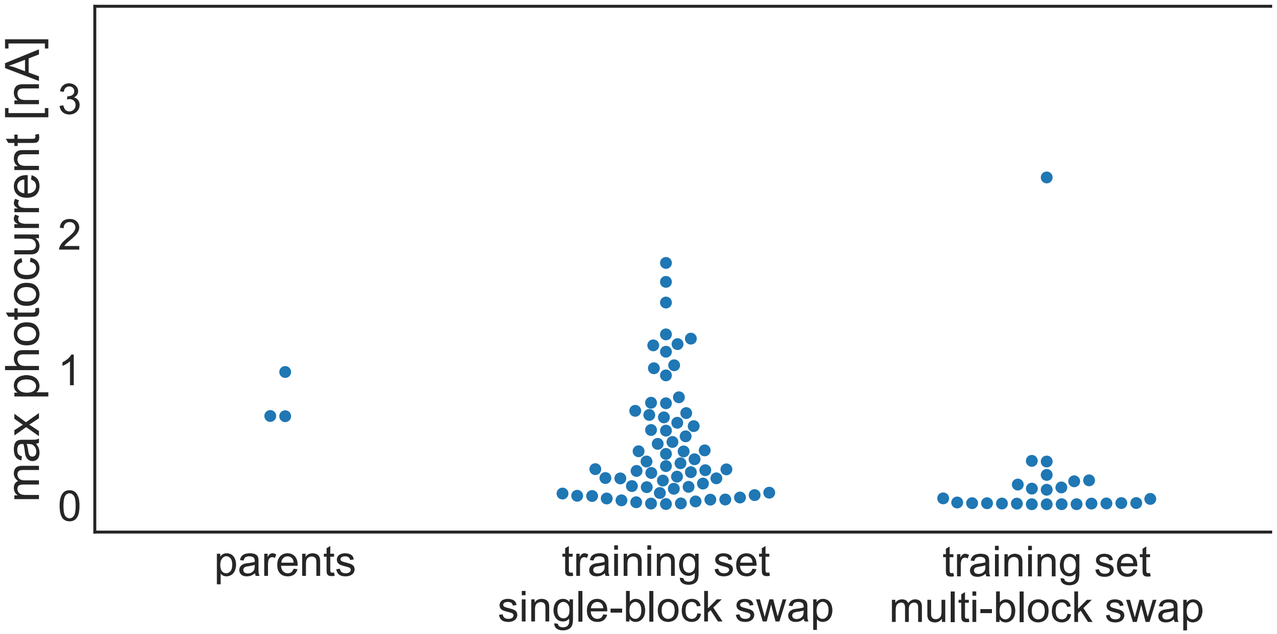


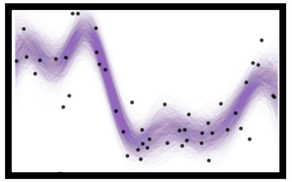


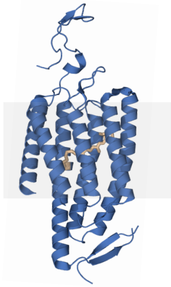


Channelrhodopsin current
~200 measurements (Bedbrook 2019)

Machine learning enables optimization with fewer measurements

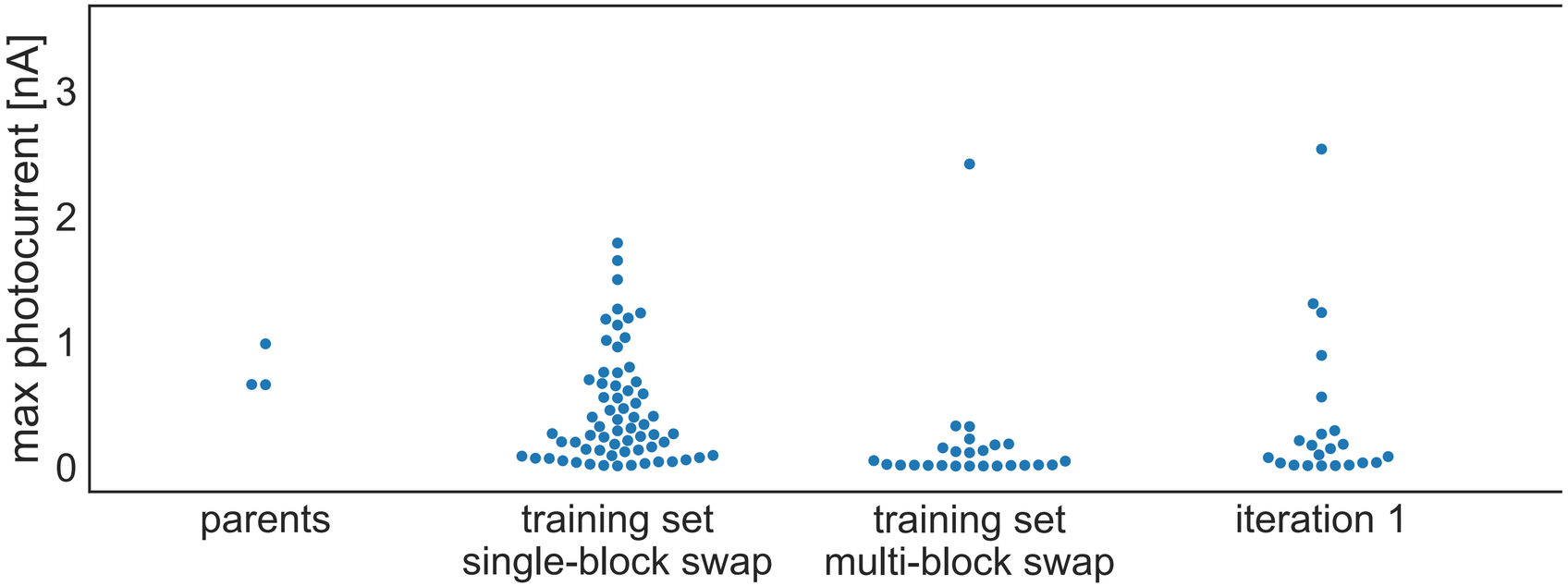


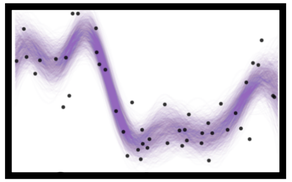


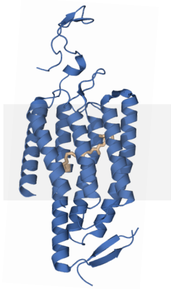


Channelrhodopsin current
~200 measurements (Bedbrook 2019)

Machine learning enables optimization with fewer measurements

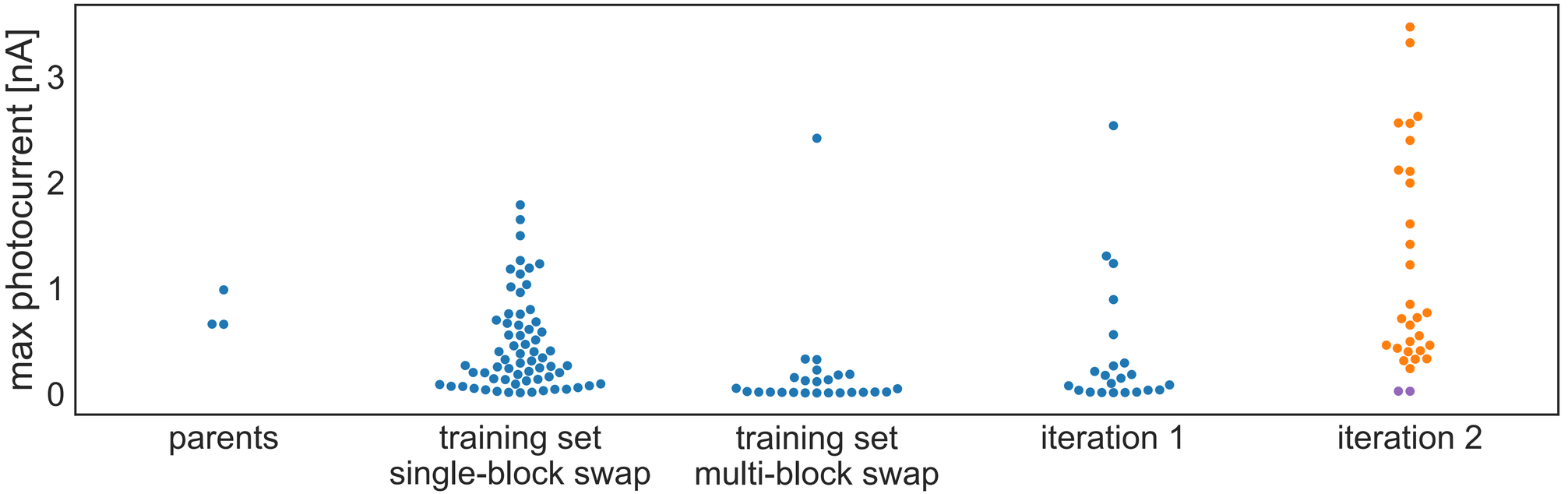


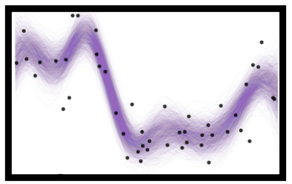




Channelrhodopsin current
~200 measurements (Bedbrook 2019)

Machine learning enables optimization with fewer measurements

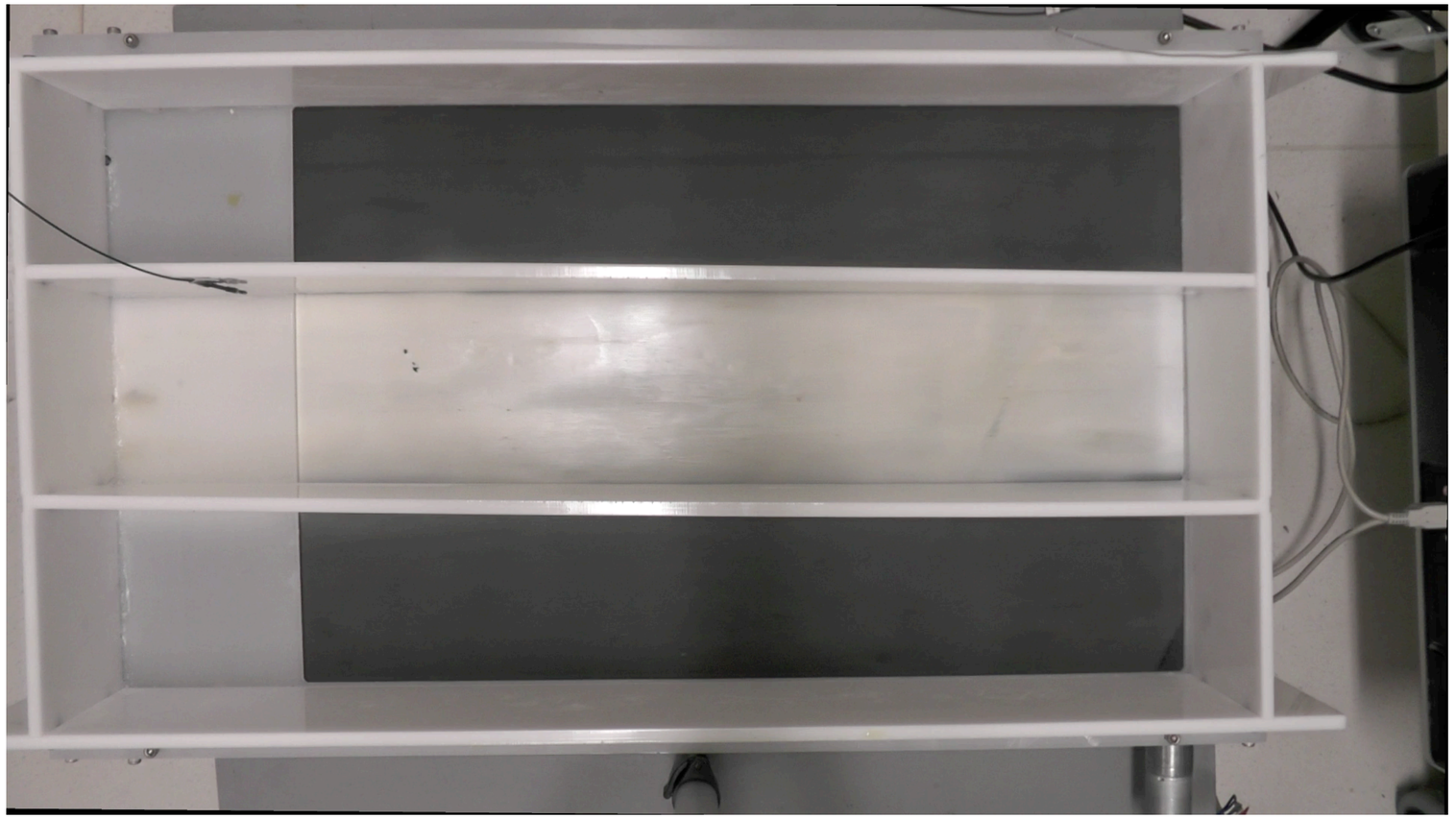




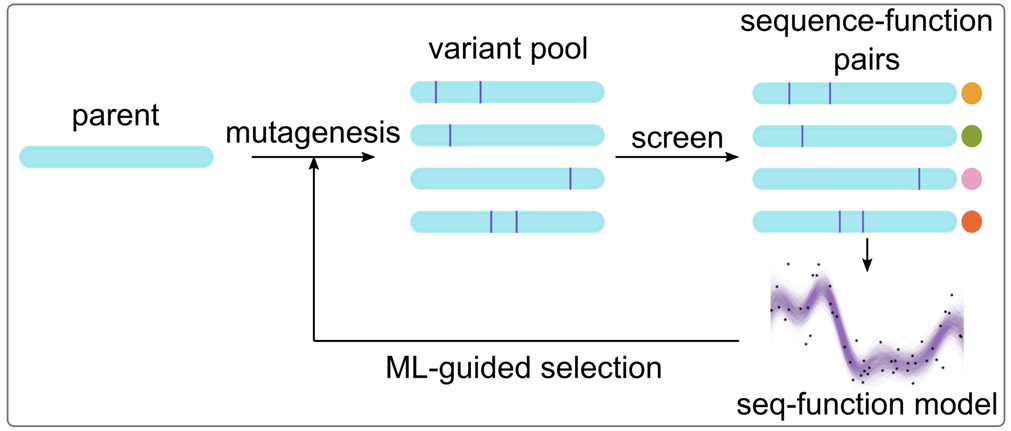
Channelrhodopsin current
~200 measurements (Bedbrook 2019)

Engineered ChRs control mouse neurons without skull removal

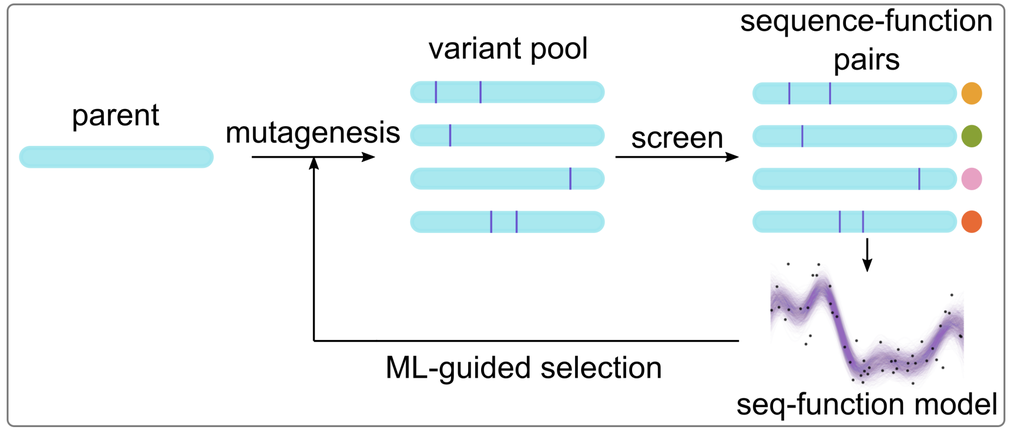
Engineered ChRs control mouse neurons without skull removal



MLDE: successes and limitations

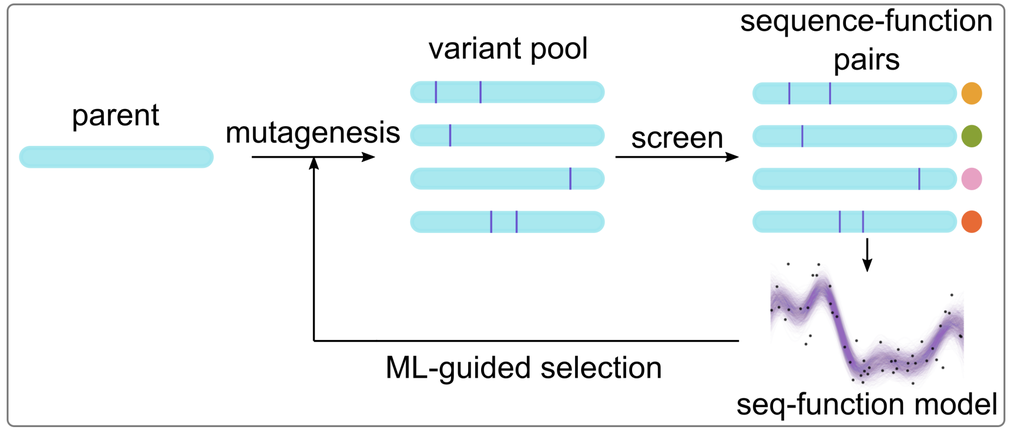


MLDE: successes and limitations



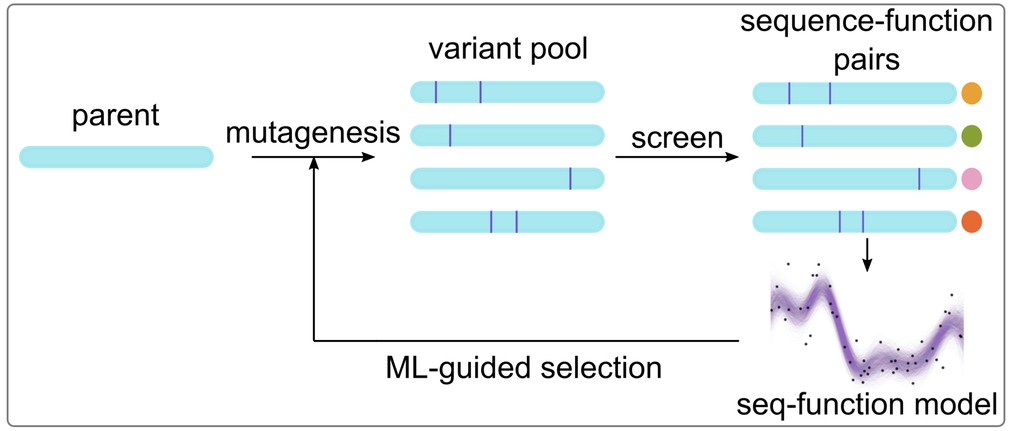
- Accurate models with < 200 measurements

MLDE: successes and limitations



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

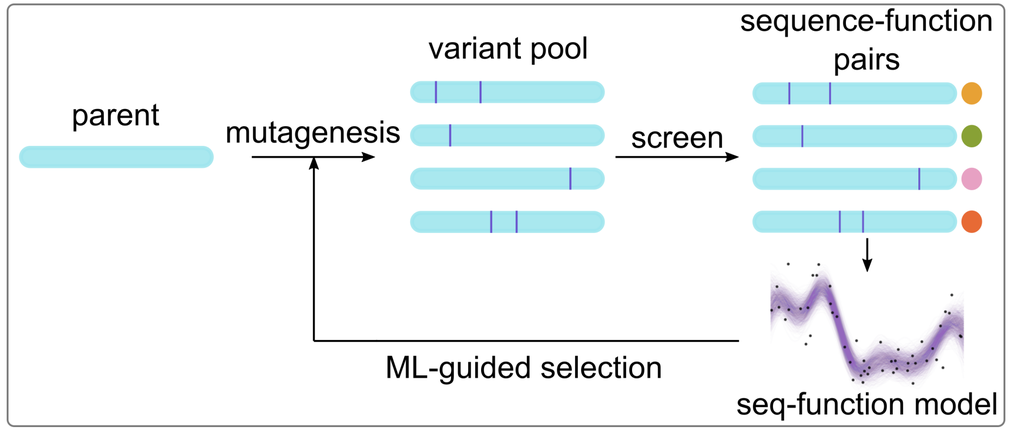
MLDE: successes and limitations



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

Still requires starting point!

MLDE: successes and limitations



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

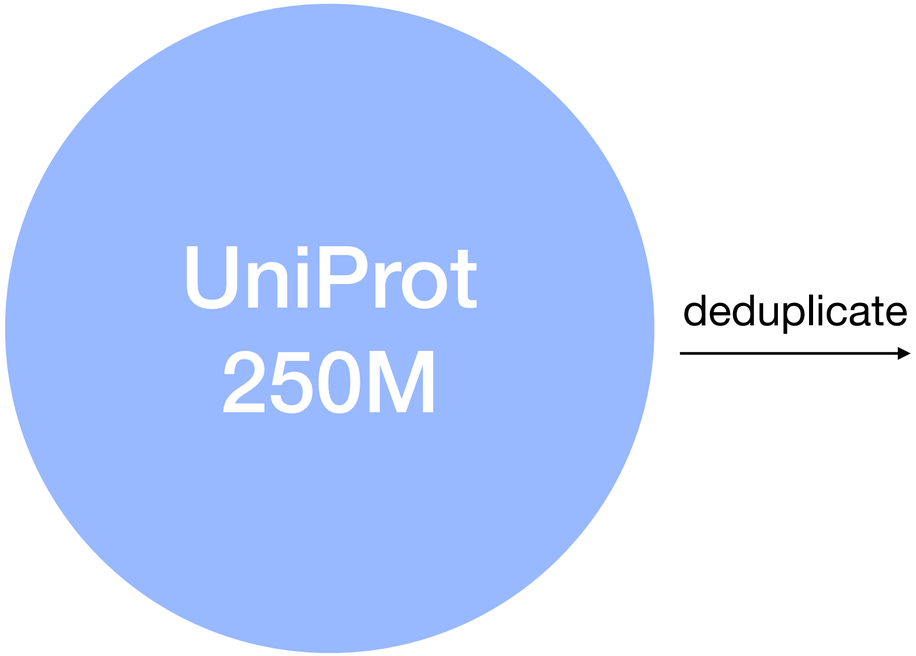
Still requires starting point!

Ignores mountains of protein data

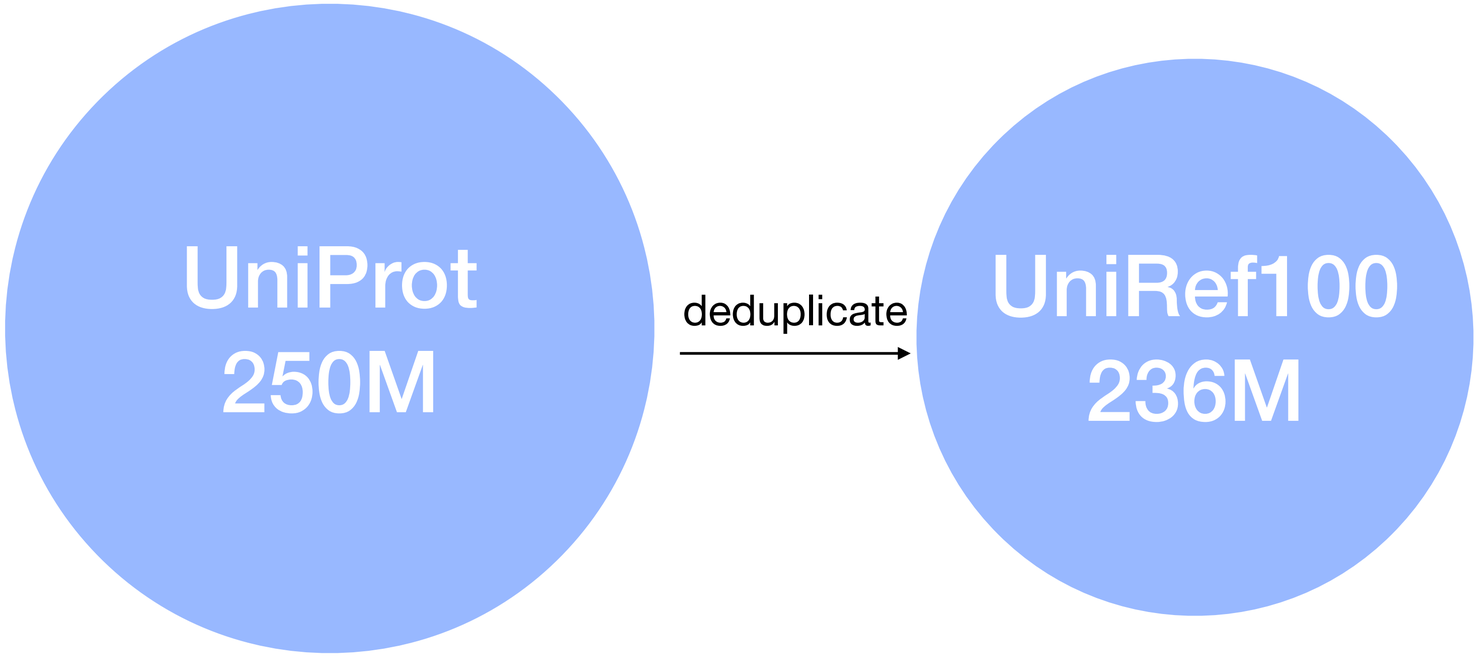
Pretraining can leverage large sequence and structure databases

Pretraining can leverage large sequence and structure databases

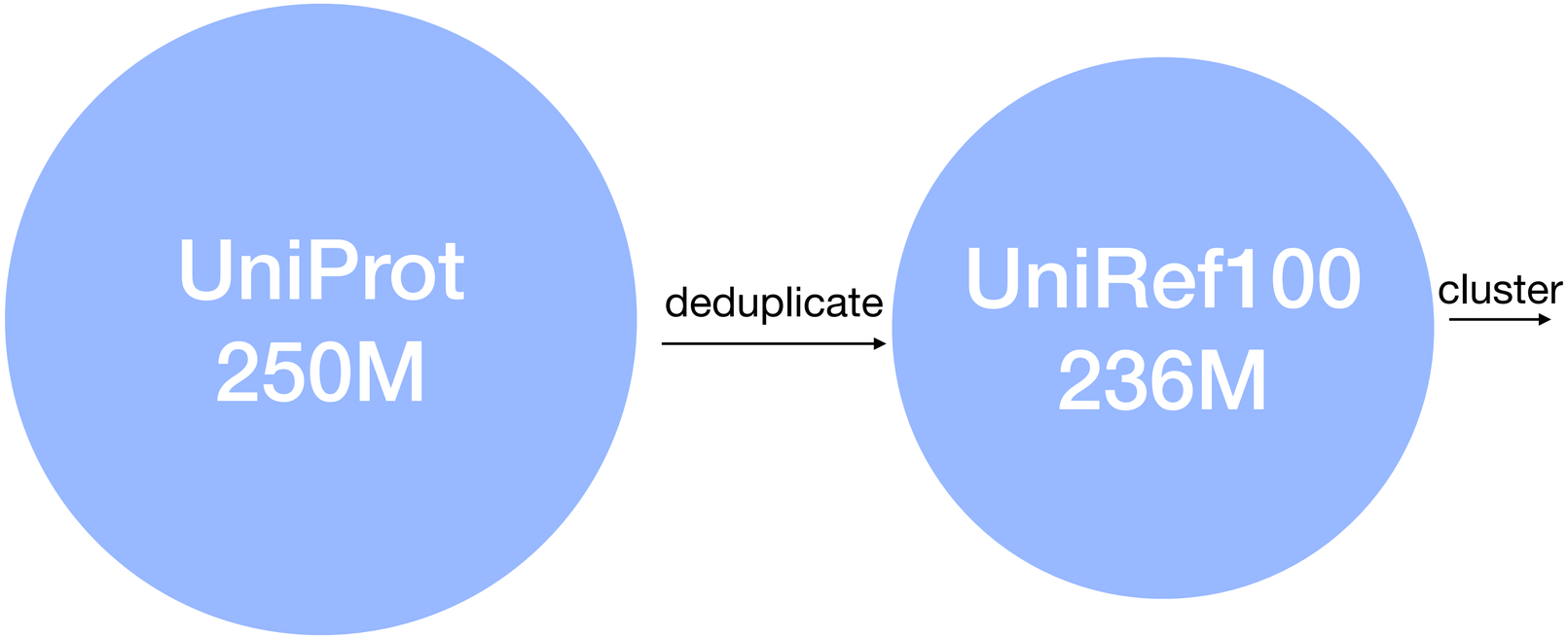
Pretraining can leverage large sequence and structure databases



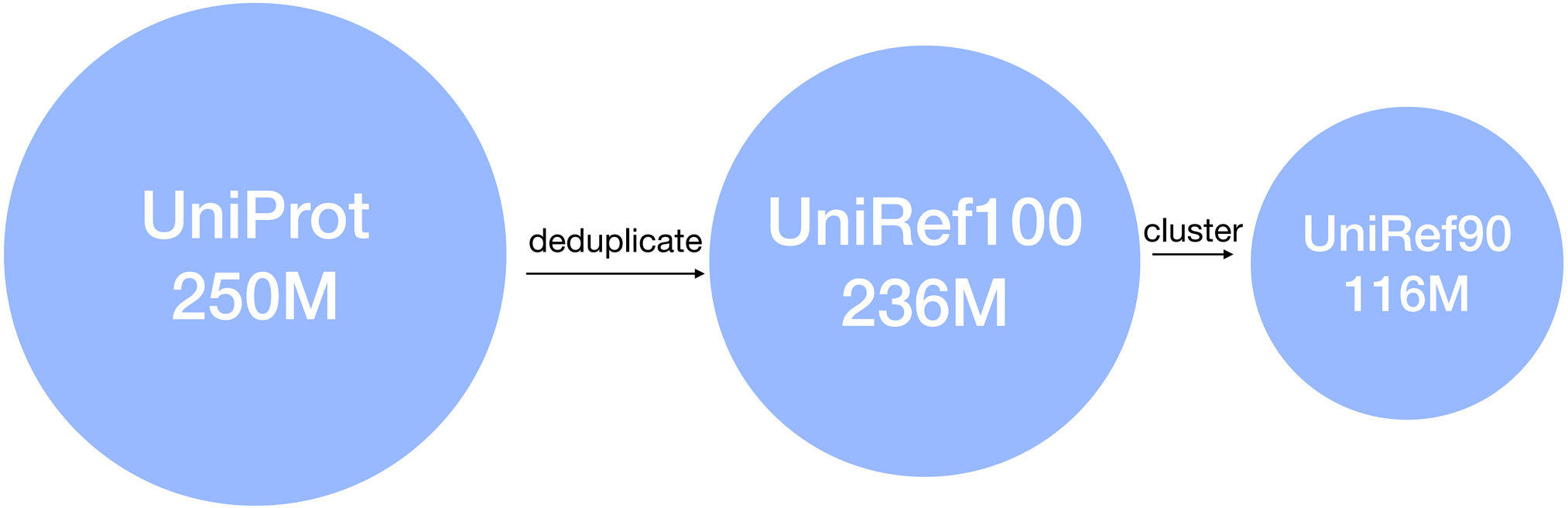
Pretraining can leverage large sequence and structure databases



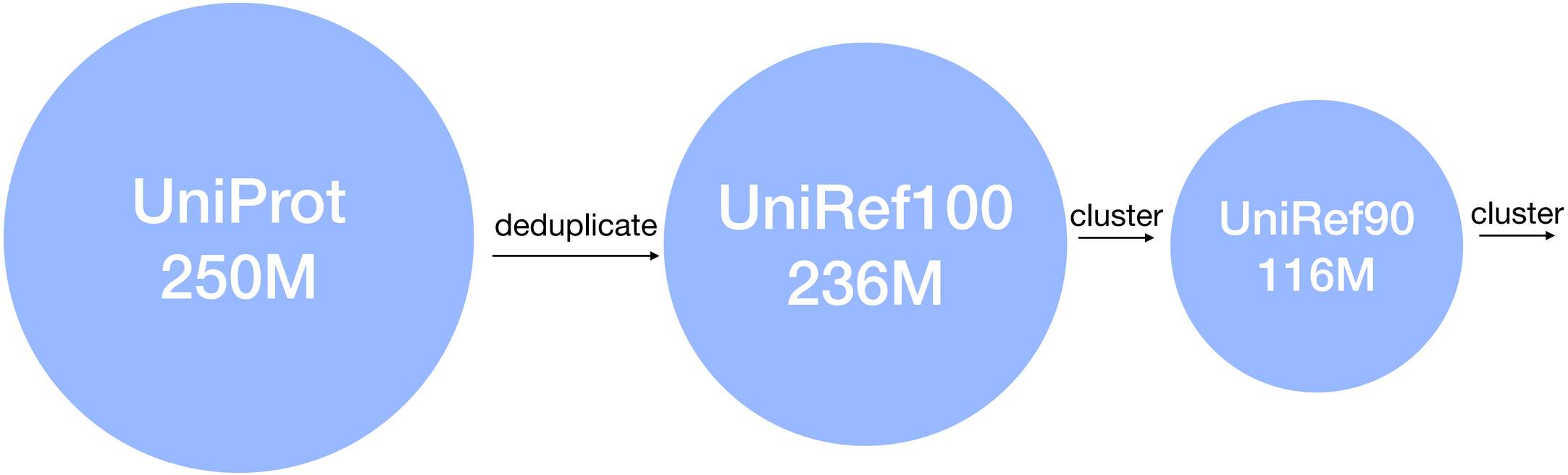
Pretraining can leverage large sequence and structure databases



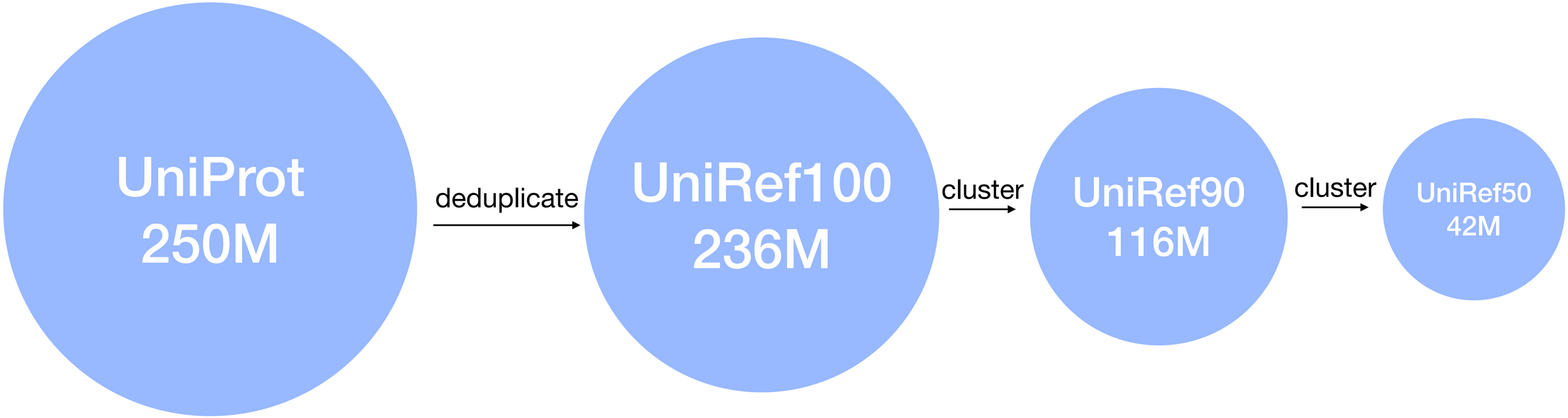
Pretraining can leverage large sequence and structure databases



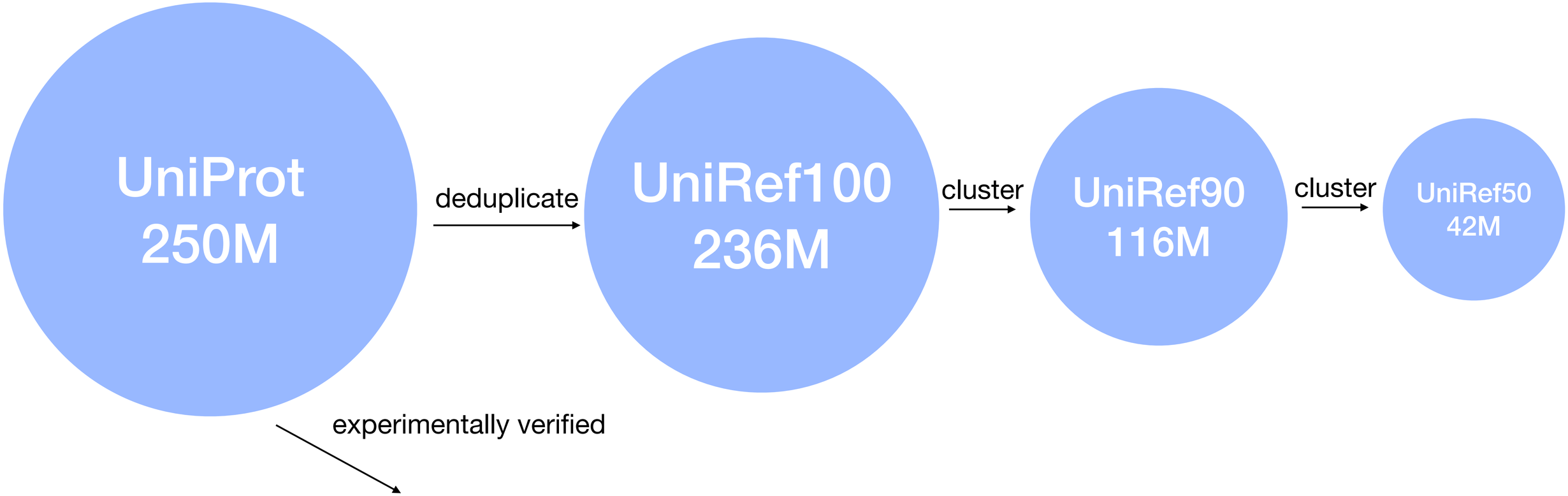
Pretraining can leverage large sequence and structure databases



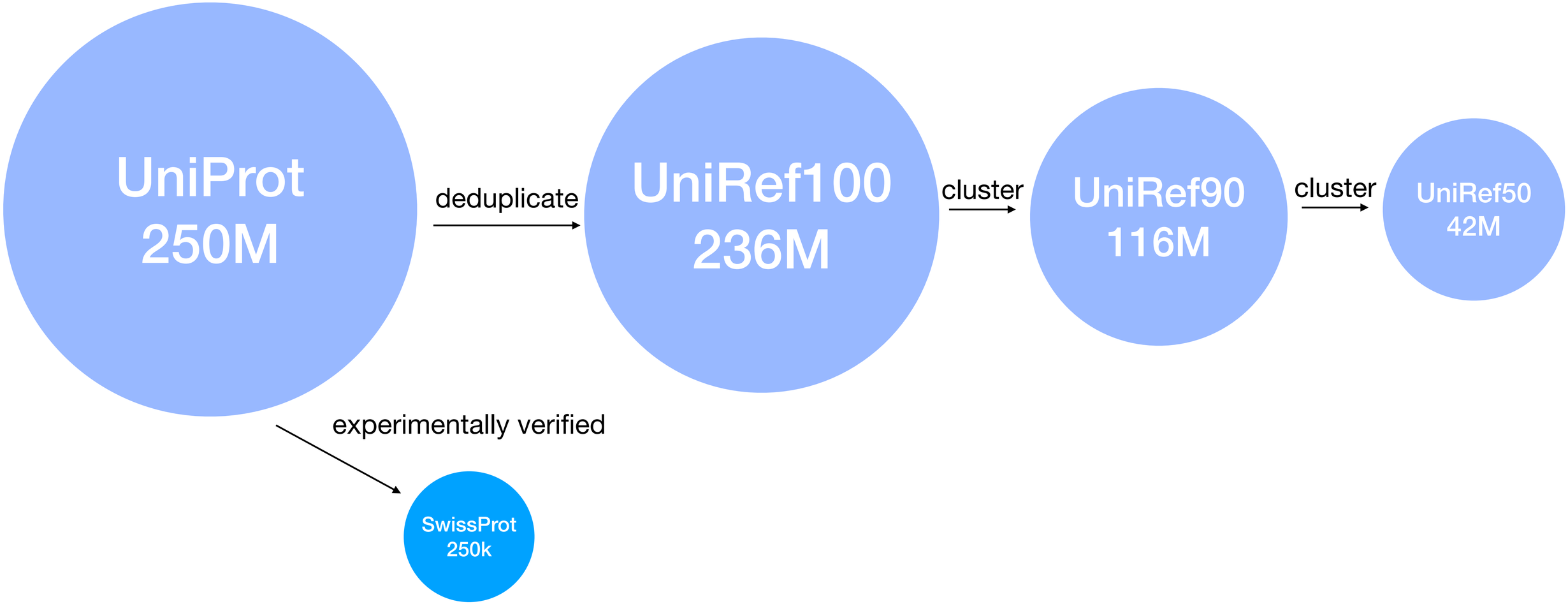
Pretraining can leverage large sequence and structure databases



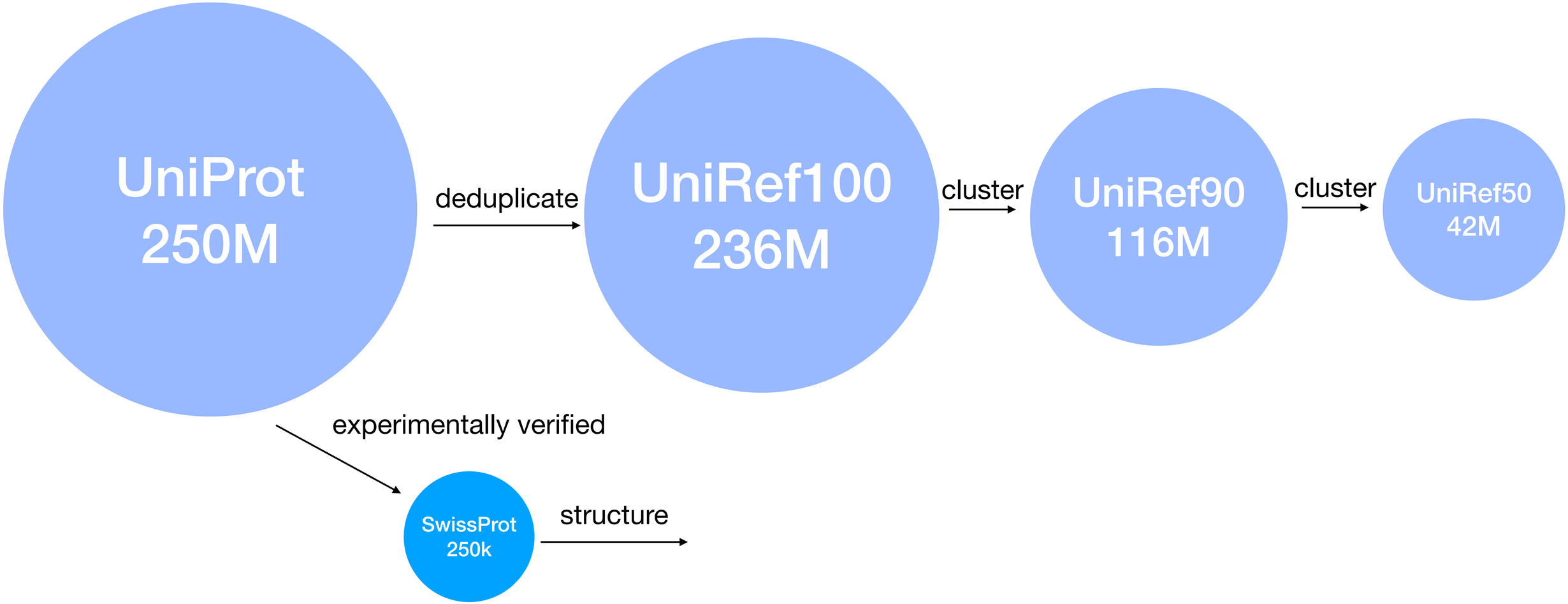
Pretraining can leverage large sequence and structure databases



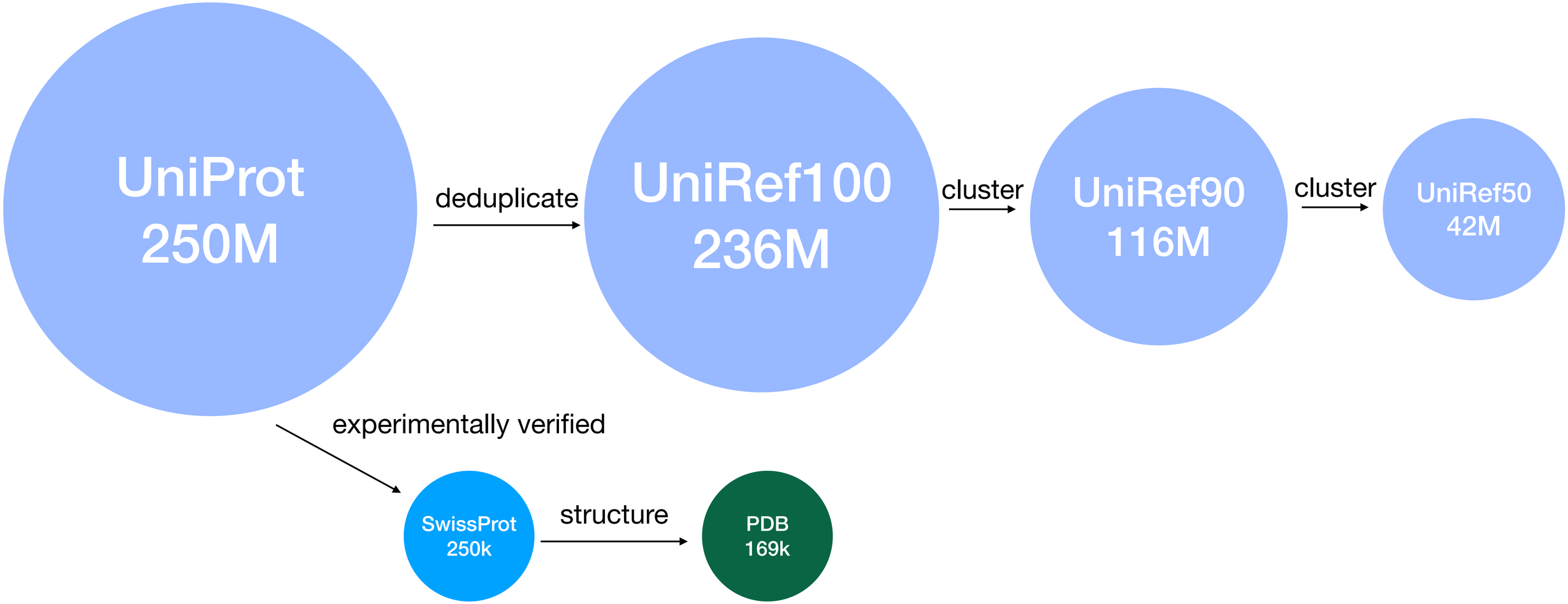
Pretraining can leverage large sequence and structure databases



Pretraining can leverage large sequence and structure databases



Pretraining can leverage large sequence and structure databases



Pretraining can leverage large sequence and structure databases

UniProt 250M

deduplicate

UniRef100 236M

cluster

UniRef90 116M

cluster

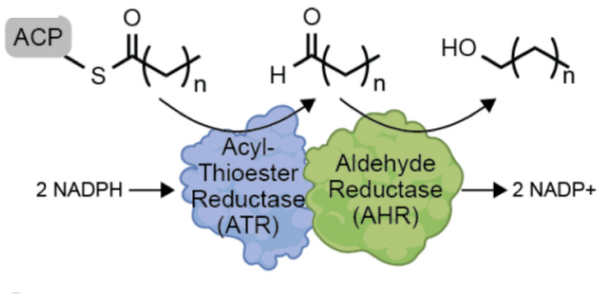
UniRef50 42M

experimentally verified

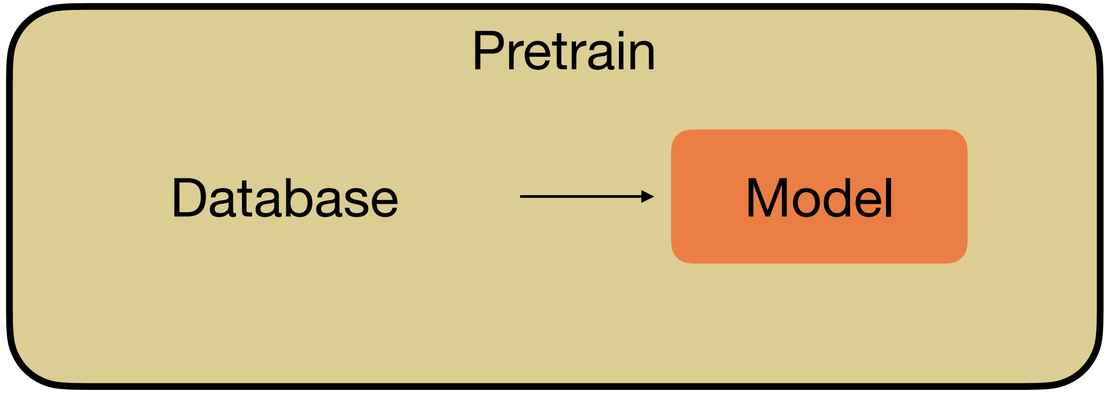
SwissProt 250k

structure

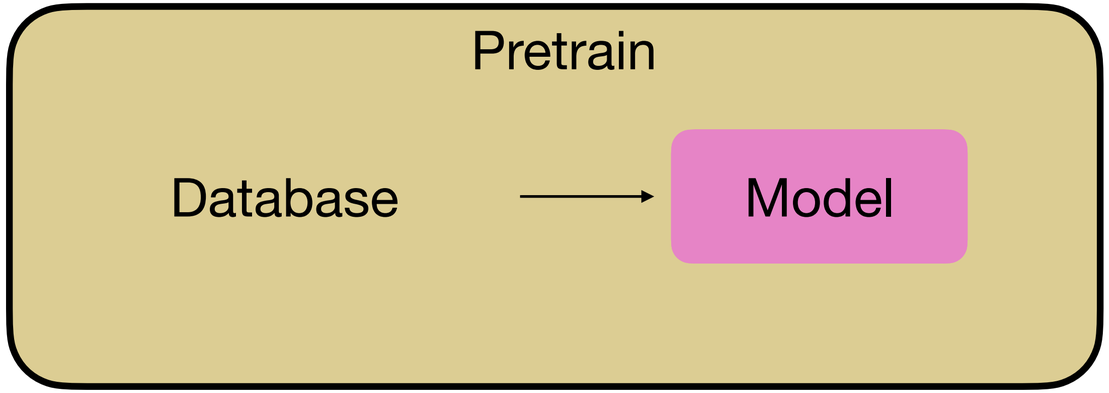
PDB 169k



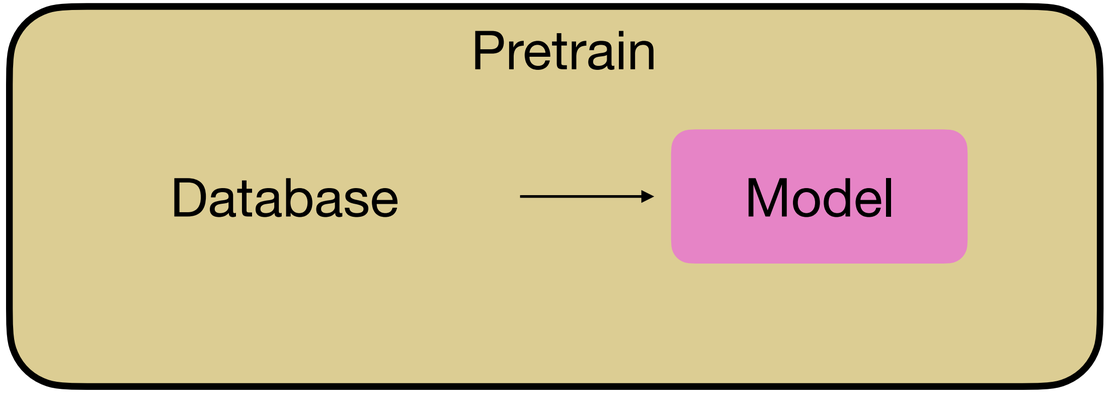
Pretraining allows us to transfer knowledge



Pretraining allows us to transfer knowledge

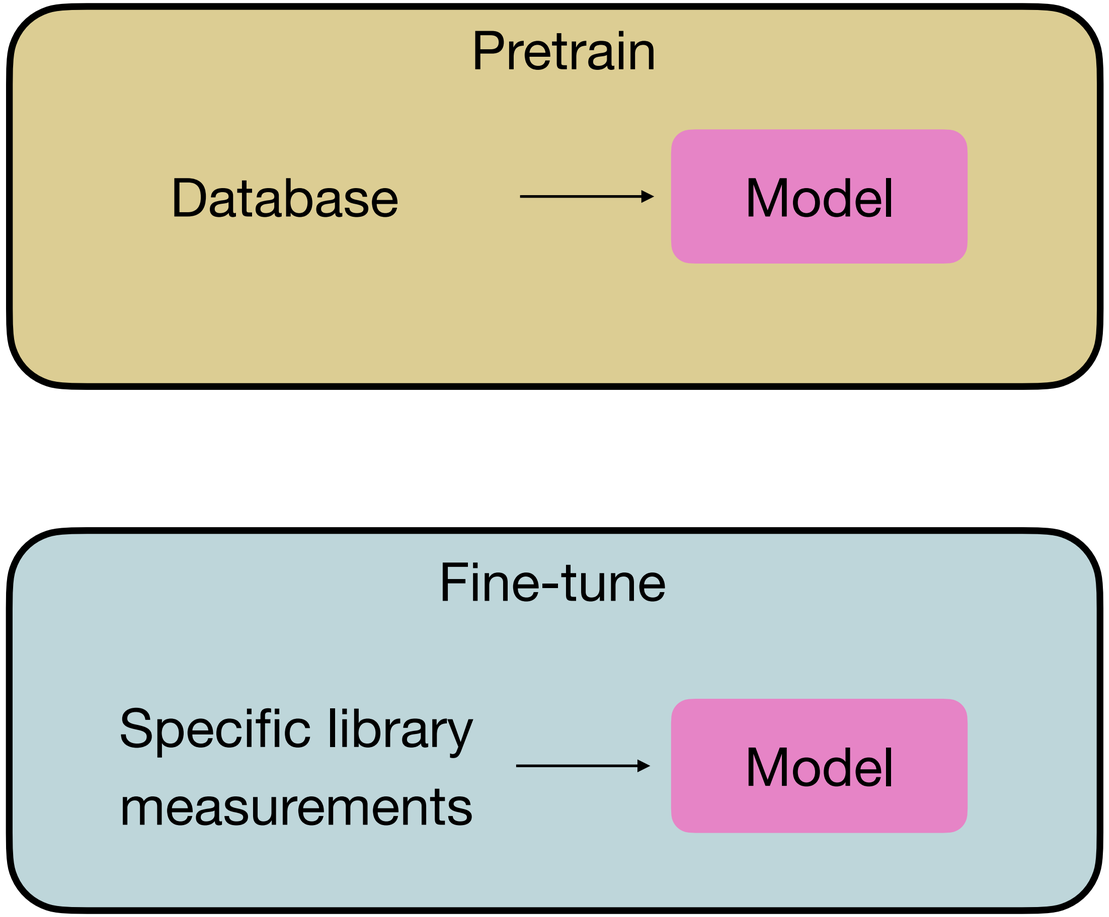


Pretraining allows us to transfer knowledge

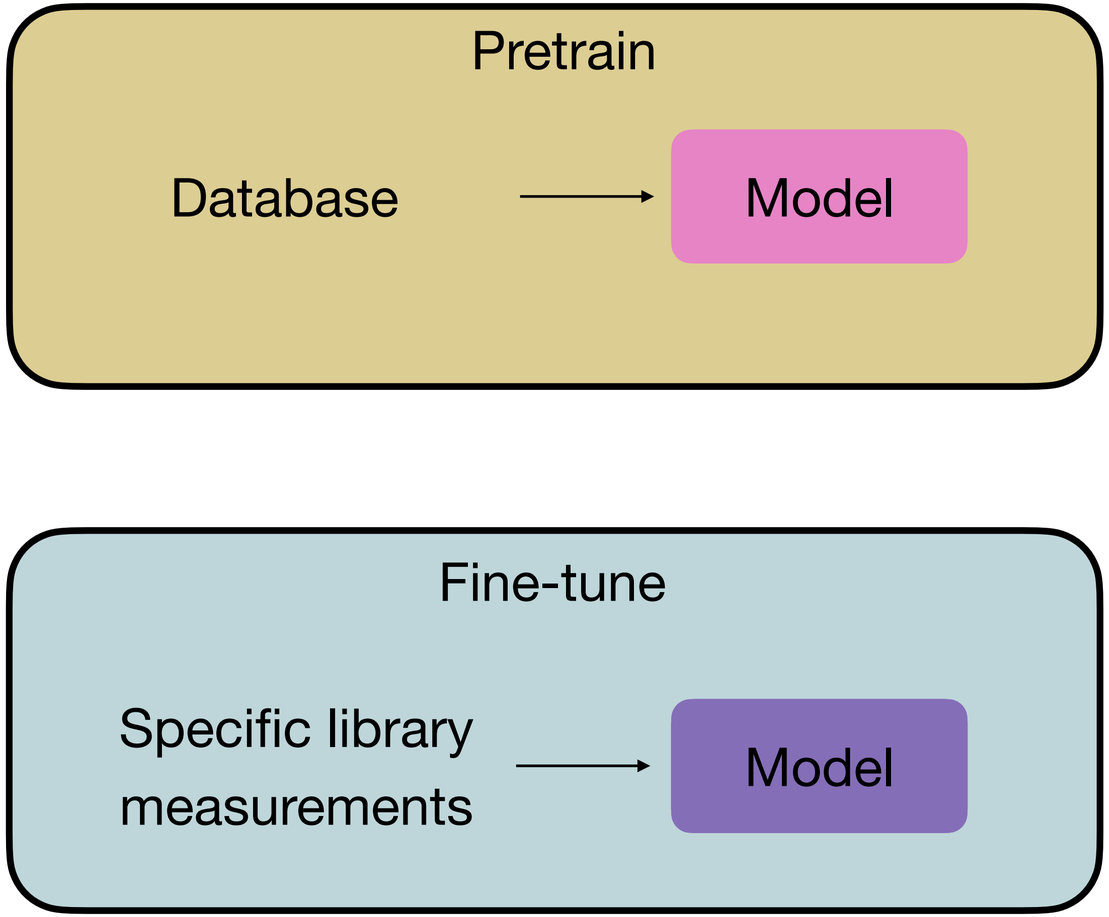


Model

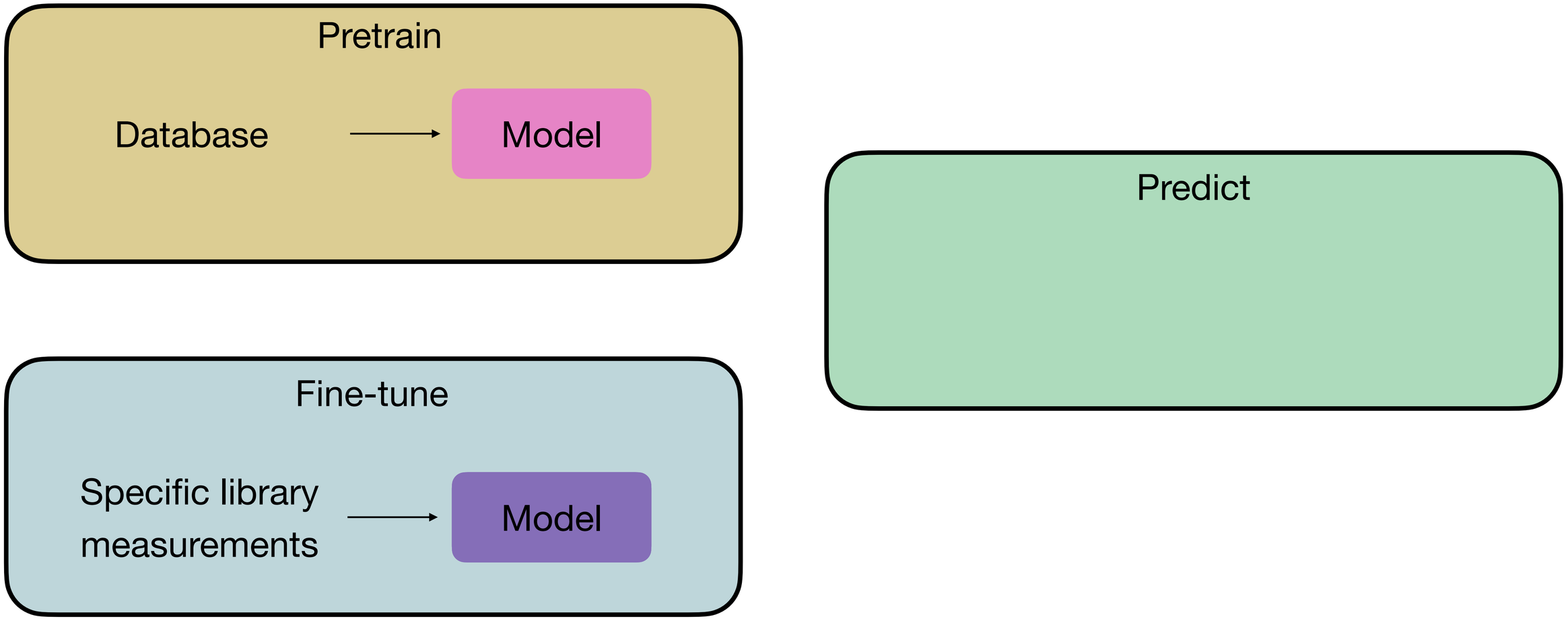
Pretraining allows us to transfer knowledge



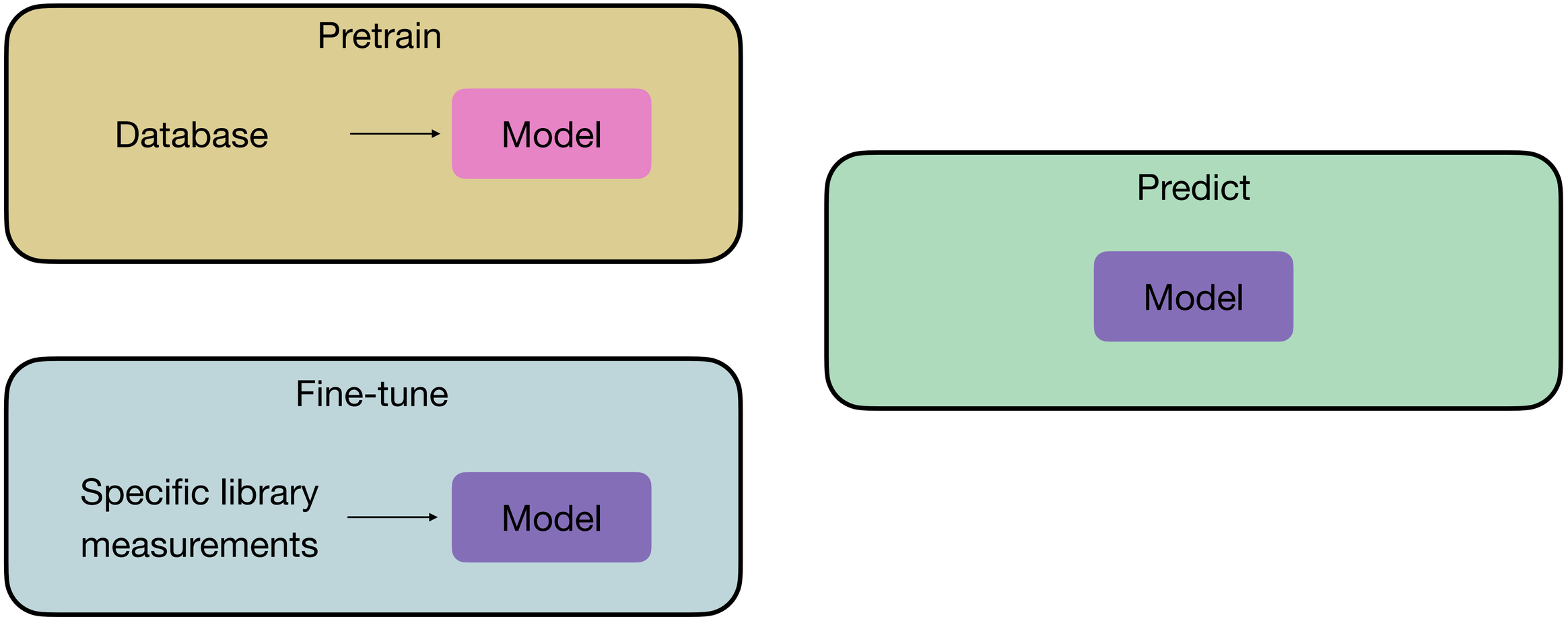
Pretraining allows us to transfer knowledge



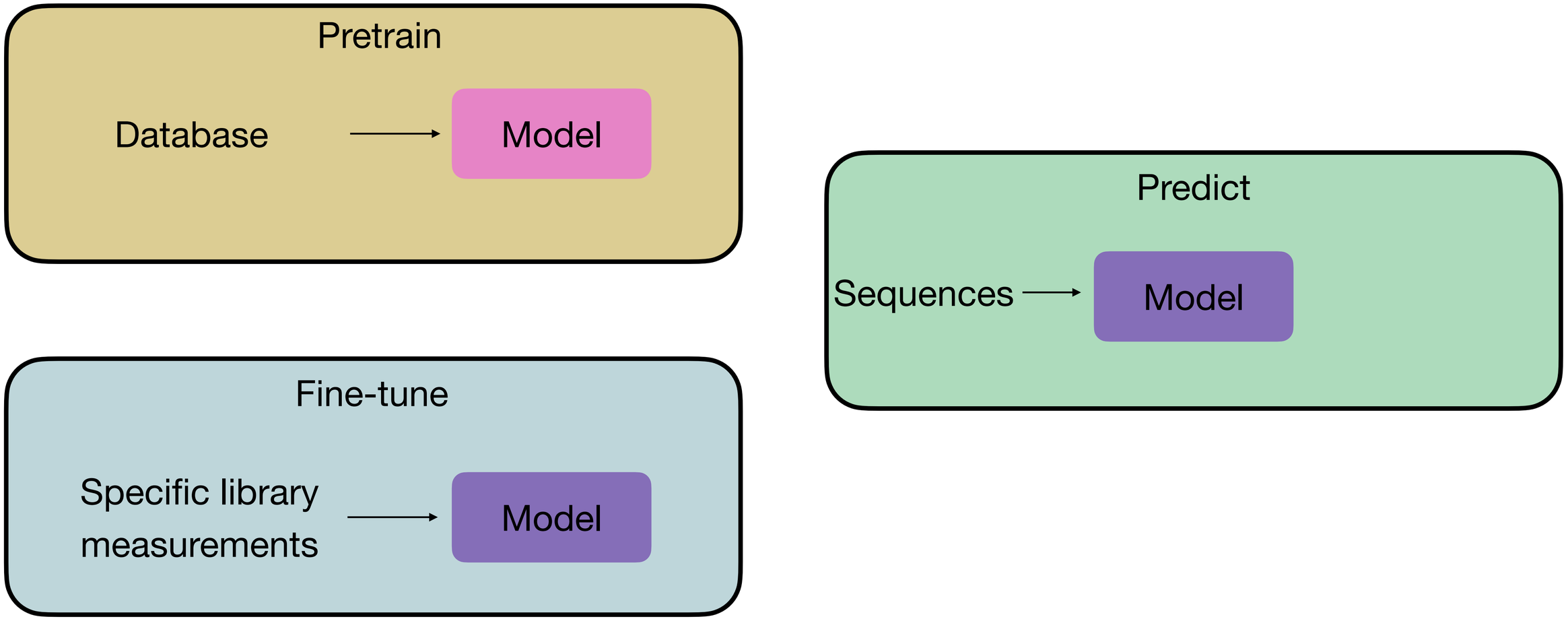
Pretraining allows us to transfer knowledge



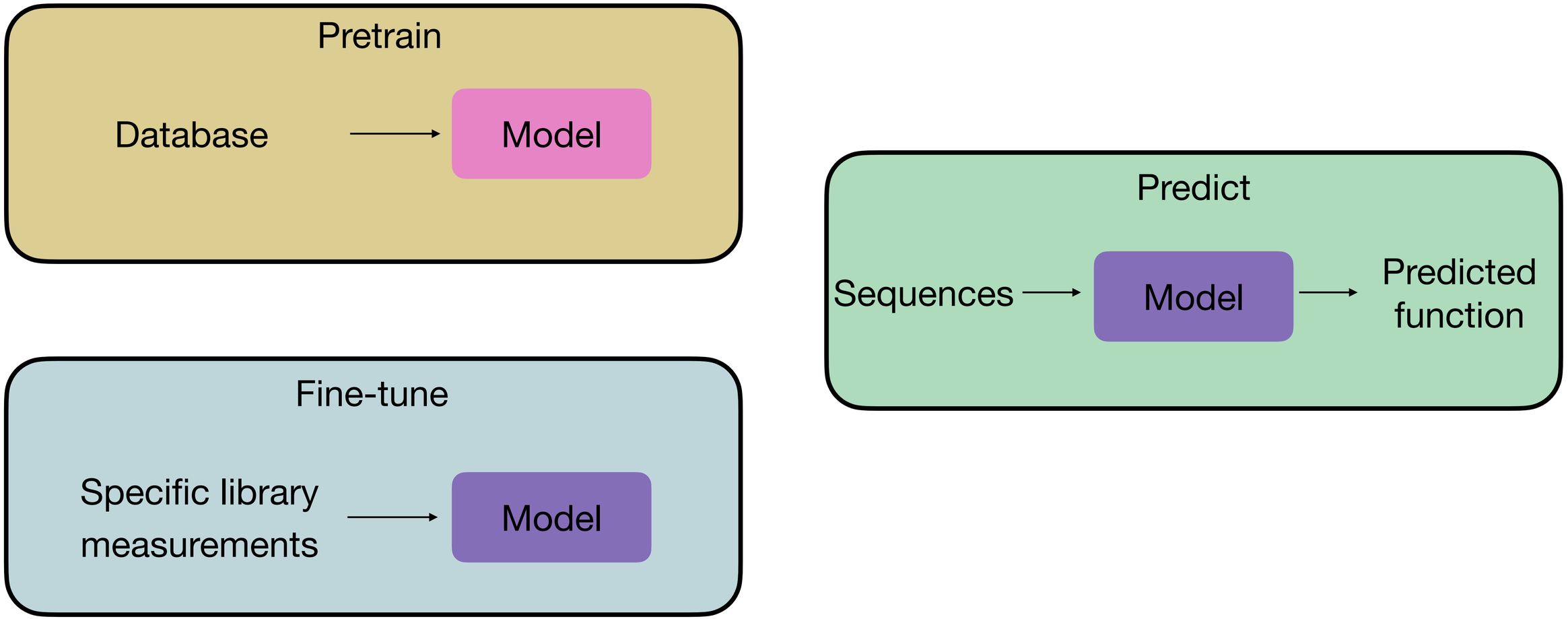
Pretraining allows us to transfer knowledge



Pretraining allows us to transfer knowledge



Pretraining allows us to transfer knowledge



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

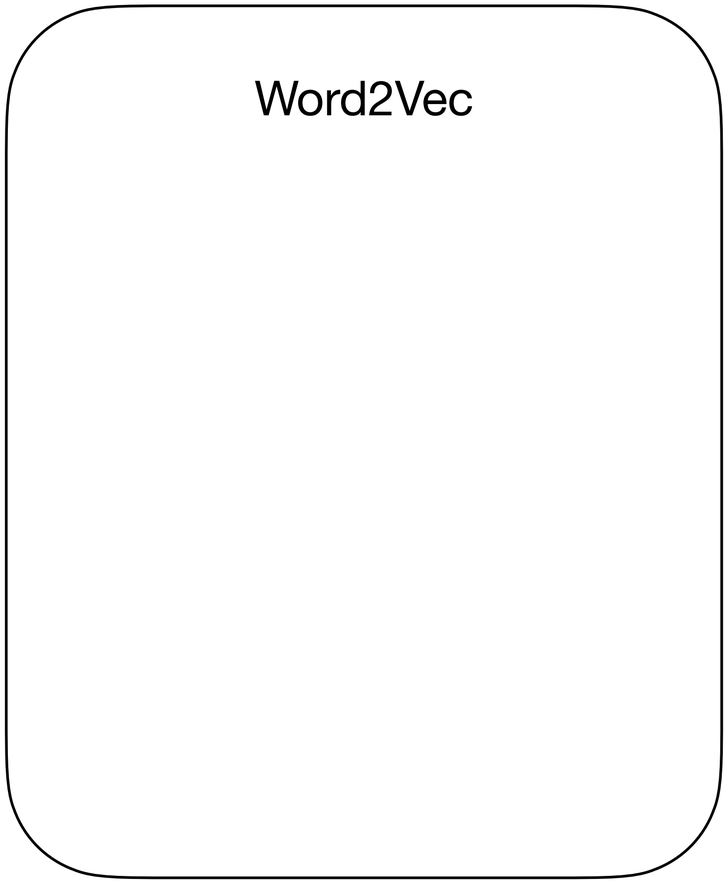
Many methods pretend proteins are language

MFTGNDAGH

Many methods pretend proteins are language

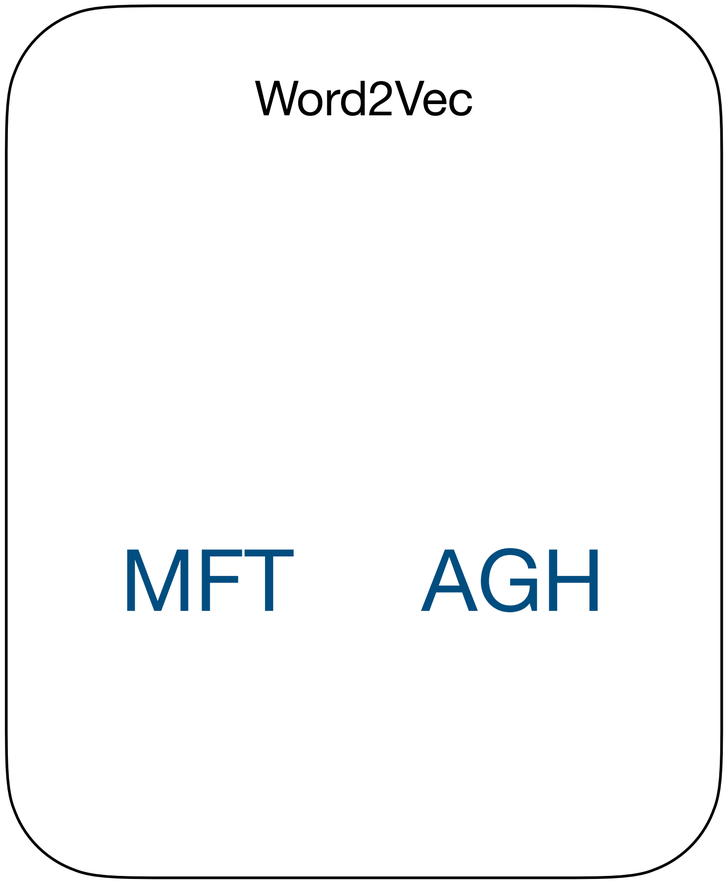
MFTGNDAGH

Many methods pretend proteins are language



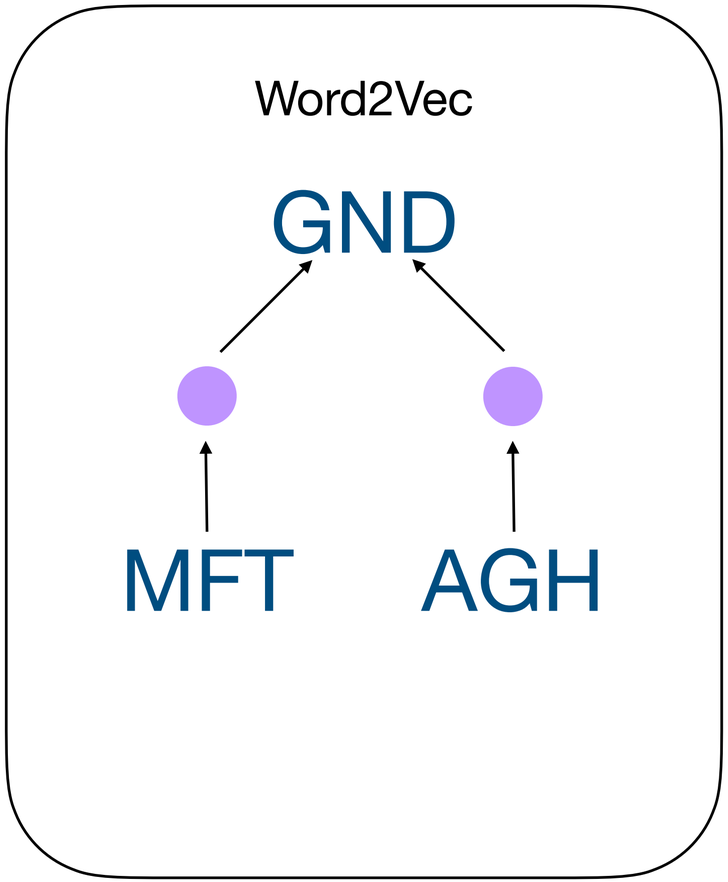
MFTGNDAGH

Many methods pretend proteins are language



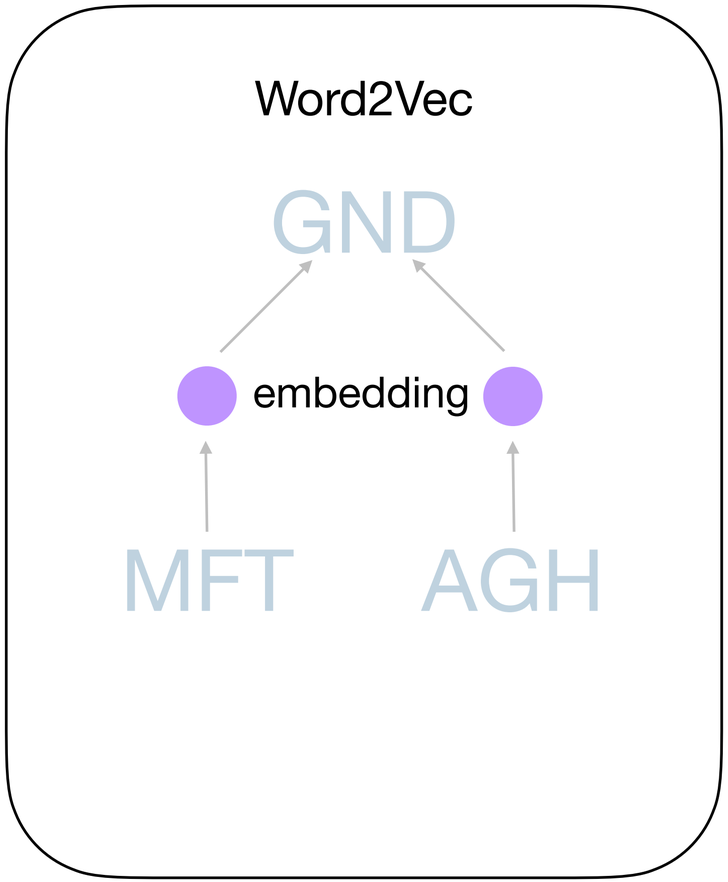
MFTGNDAGH

Many methods pretend proteins are language



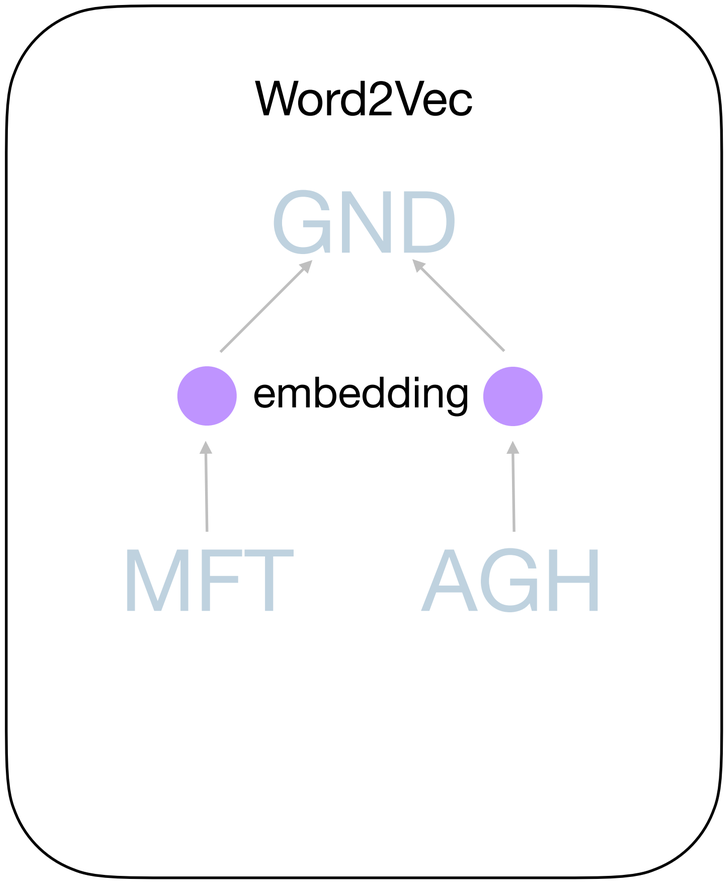
MFTGNDAGH

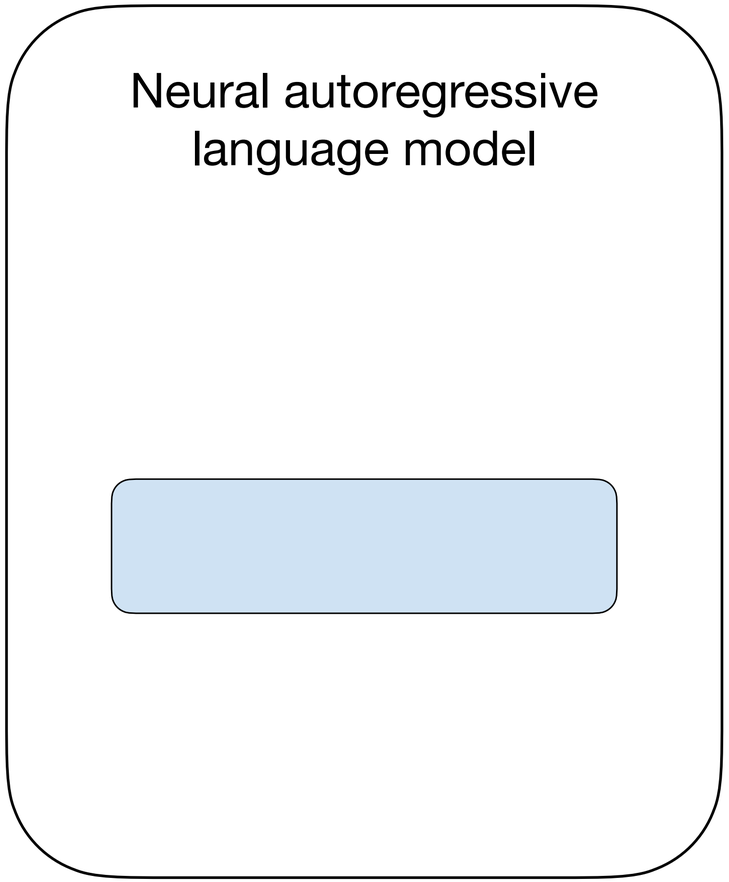
Many methods pretend proteins are language



MFTGNDAGH

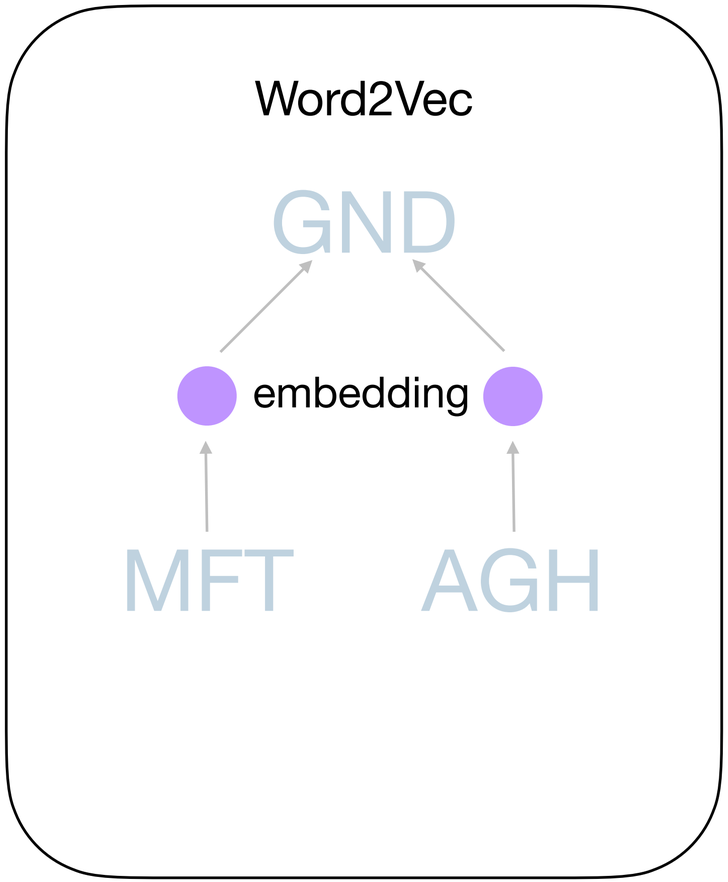
Many methods pretend proteins are language

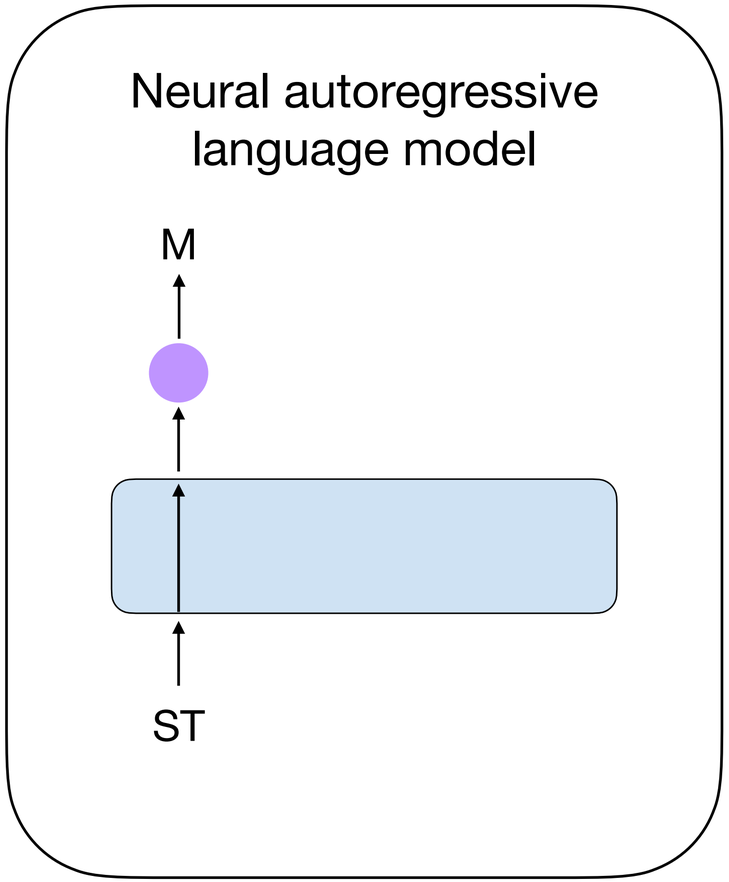




MFTGNDAGH

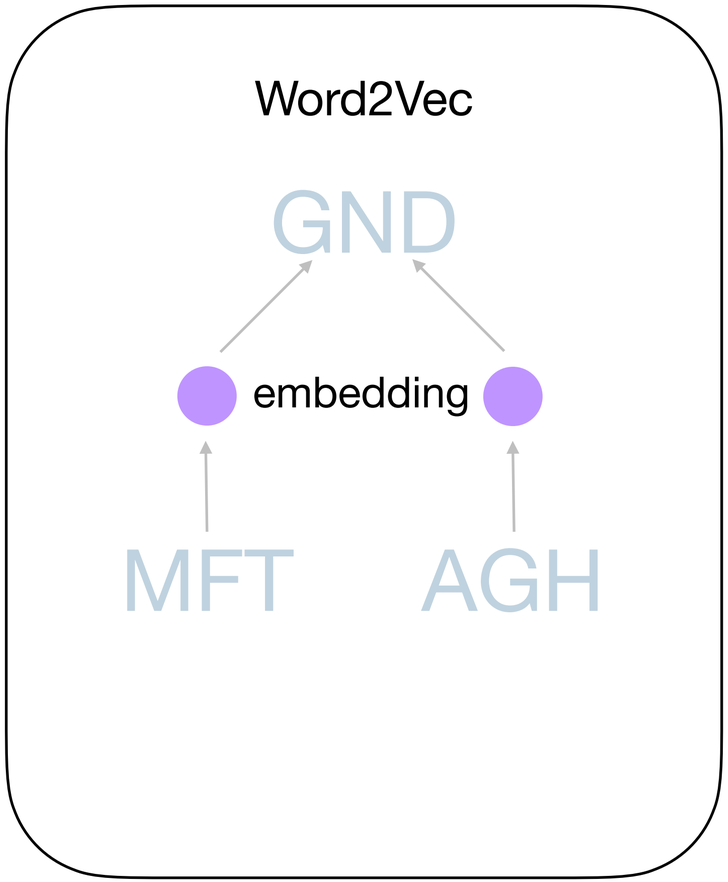
Many methods pretend proteins are language

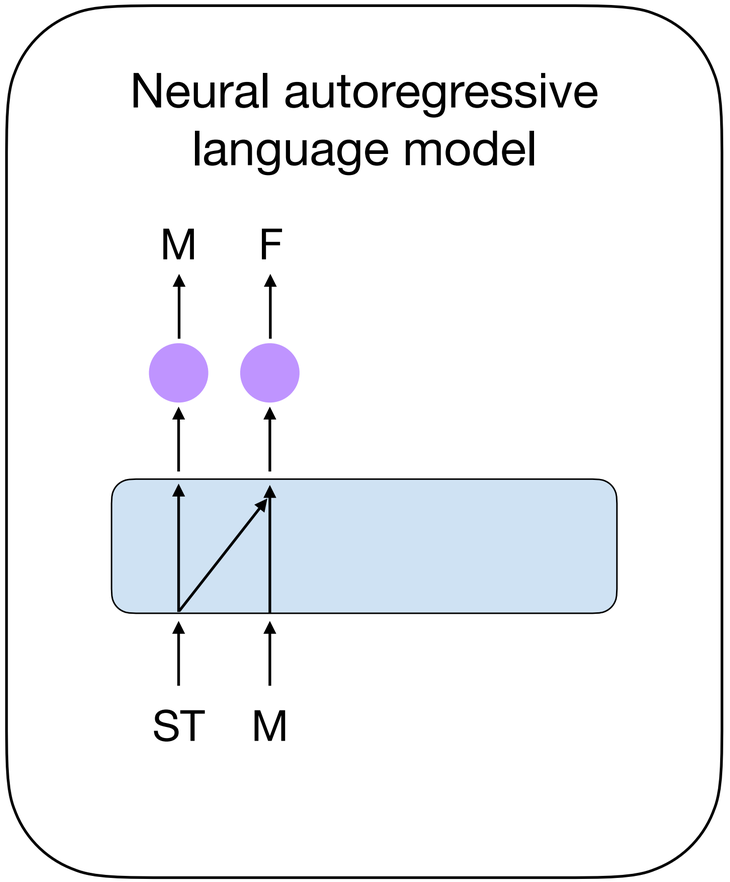




MFTGNDAGH

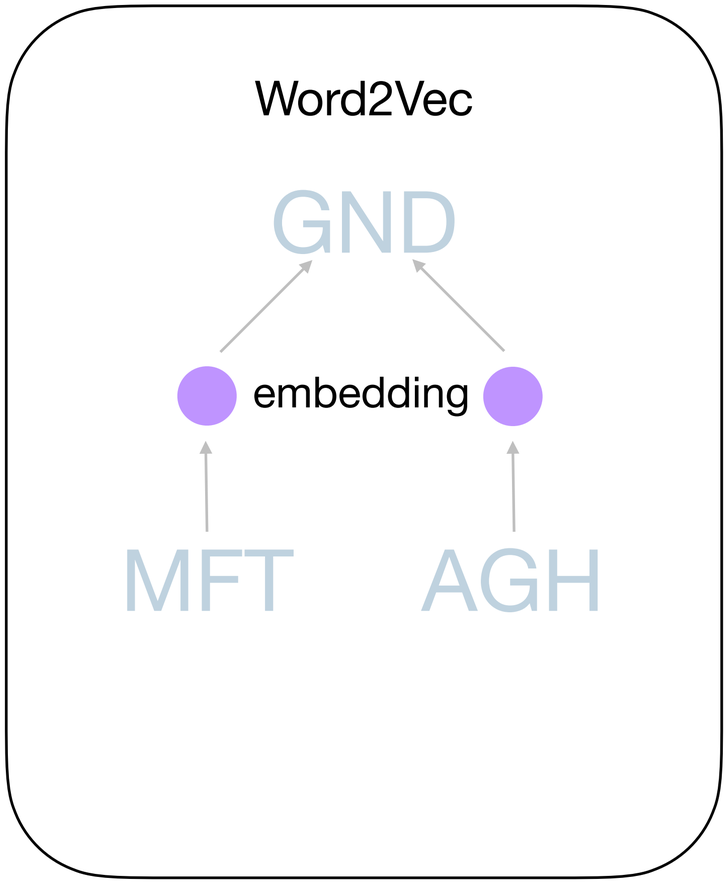
Many methods pretend proteins are language

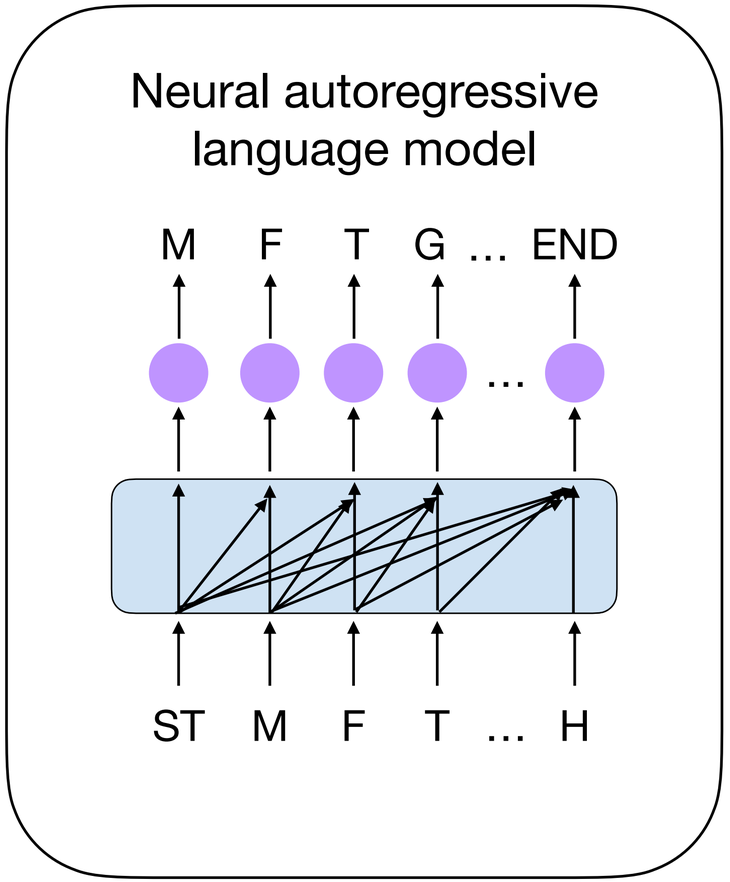




MFTGNDAGH

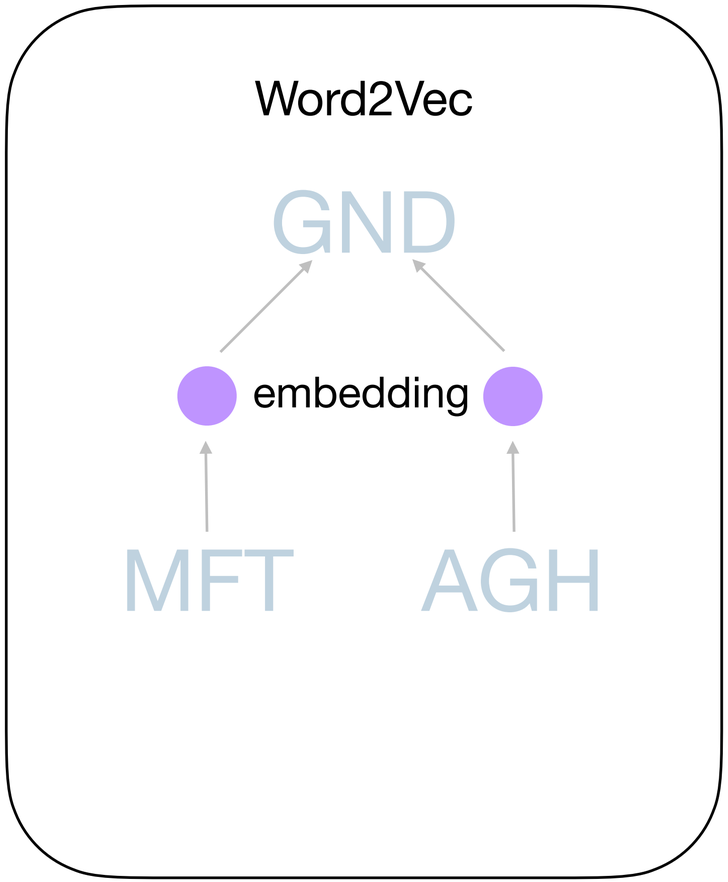
Many methods pretend proteins are language

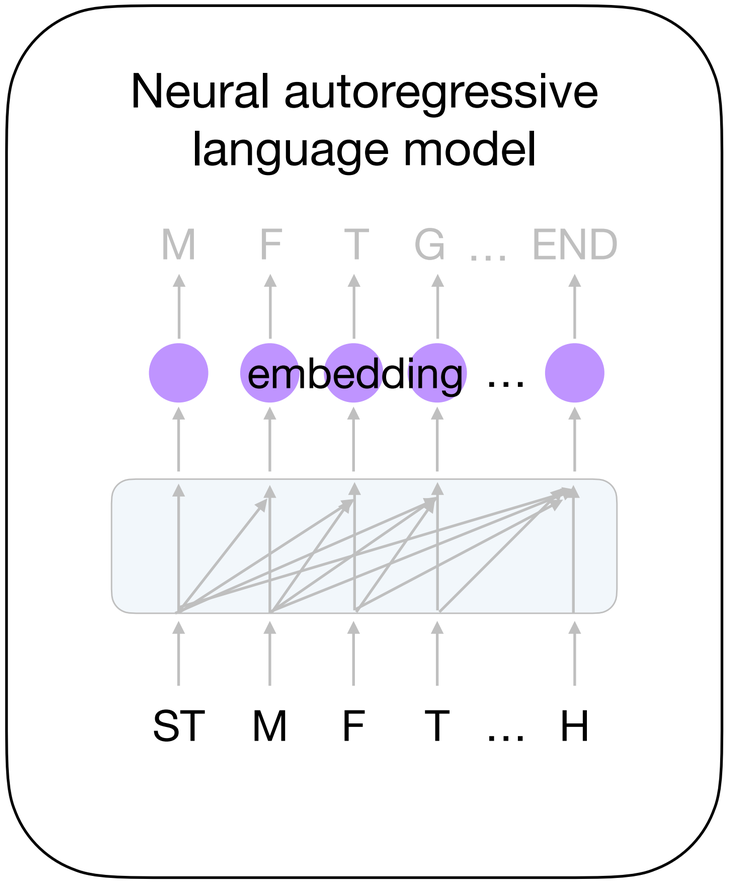




MFTGNDAGH

Many methods pretend proteins are language

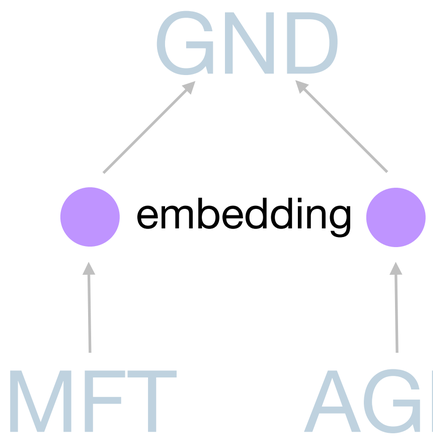




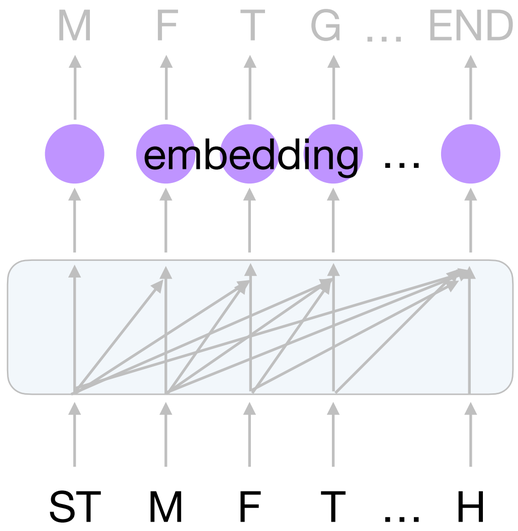
MFTGNDAGH

Many methods pretend proteins are language

Word2Vec



Neural autoregressive language model

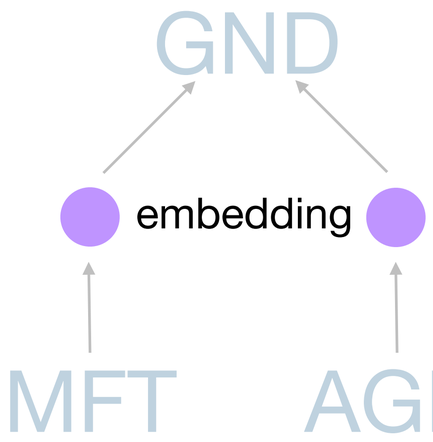


Neural masked language model

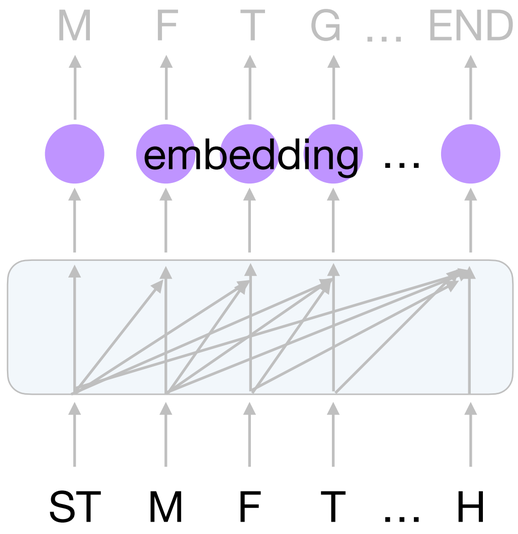
MFTGNDAGH

Many methods pretend proteins are language

Word2Vec



Neural autoregressive language model



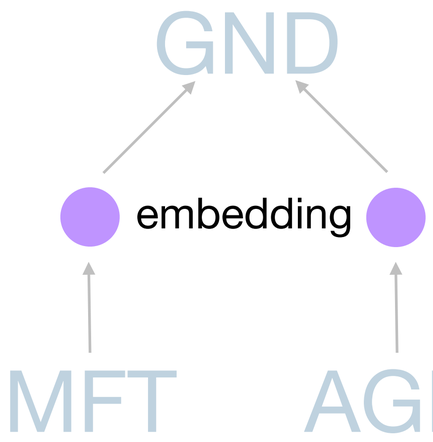
Neural masked language model

M F T G … H

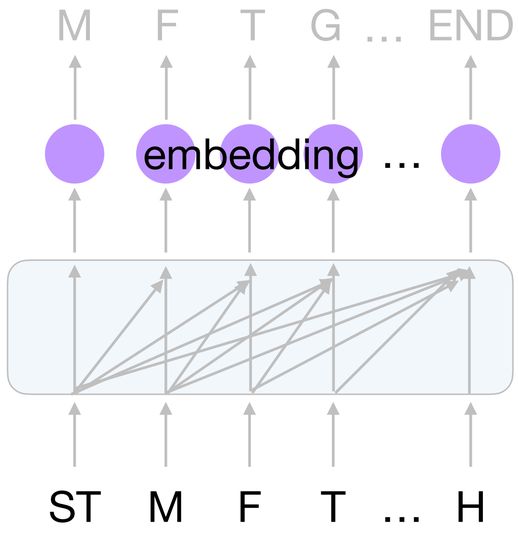
MFTGNDAGH

Many methods pretend proteins are language

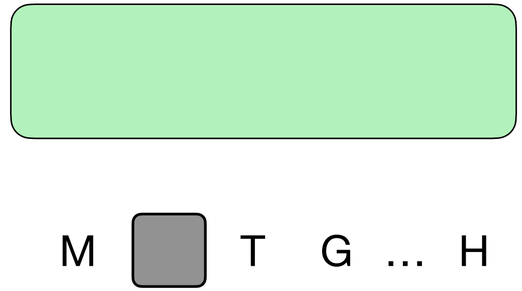
Word2Vec



Neural autoregressive language model



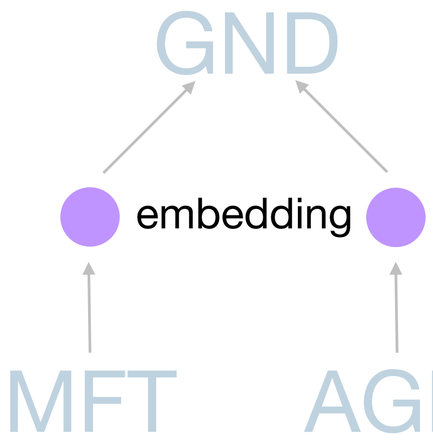
Neural masked language model



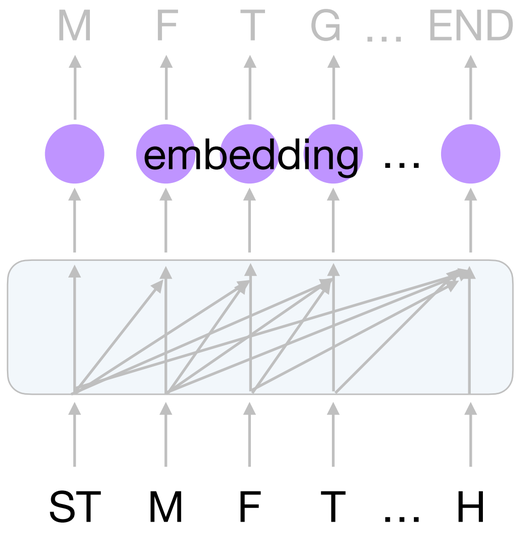
MFTGNDAGH

Many methods pretend proteins are language

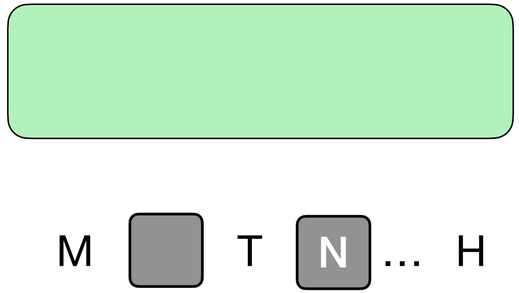
Word2Vec



Neural autoregressive language model



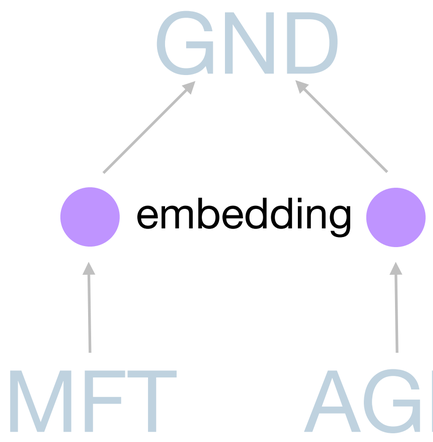
Neural masked language model



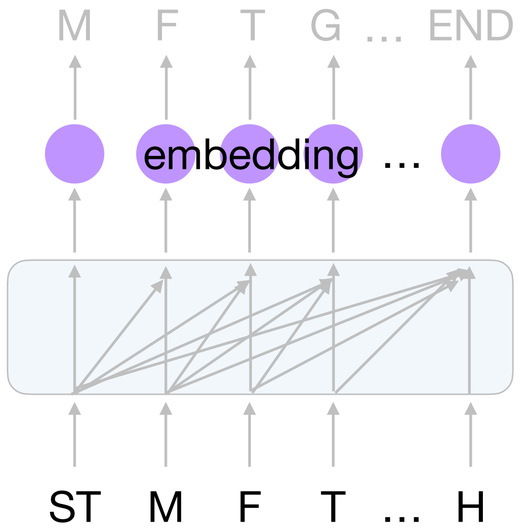
MFTGNDAGH

Many methods pretend proteins are language

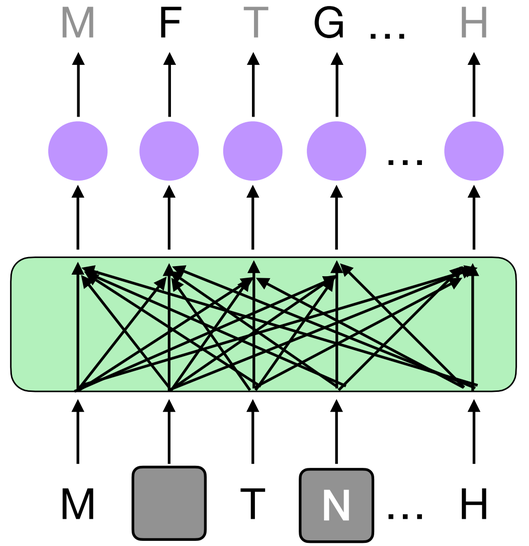
Word2Vec



Neural autoregressive language model



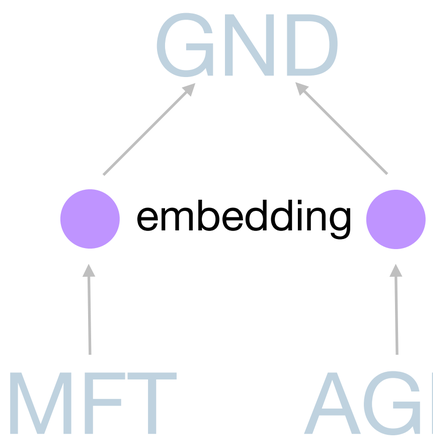
Neural masked language model



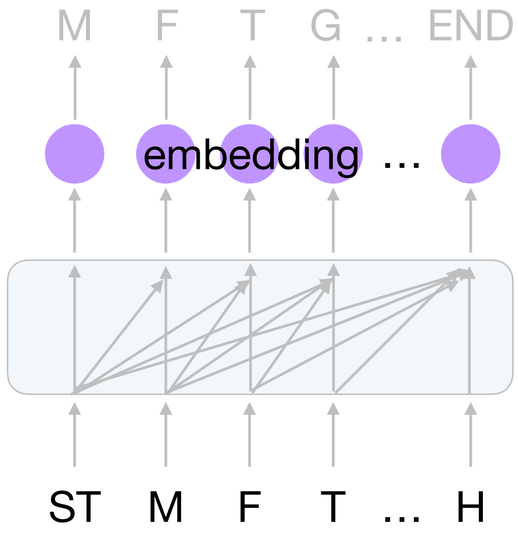
MFTGNDAGH

Many methods pretend proteins are language

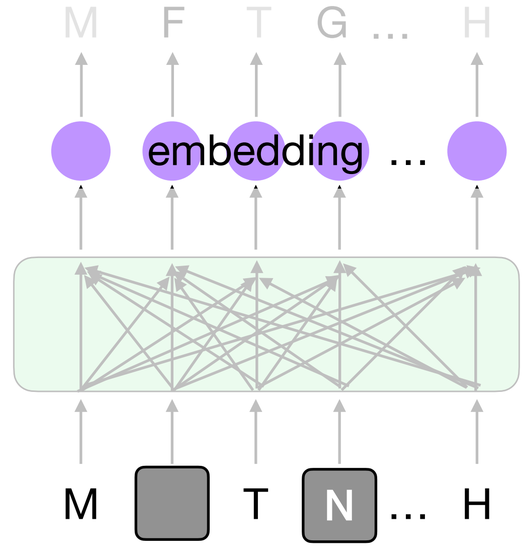
Word2Vec



Neural autoregressive language model



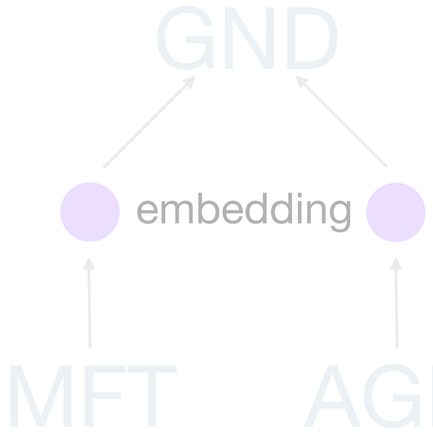
Neural masked language model



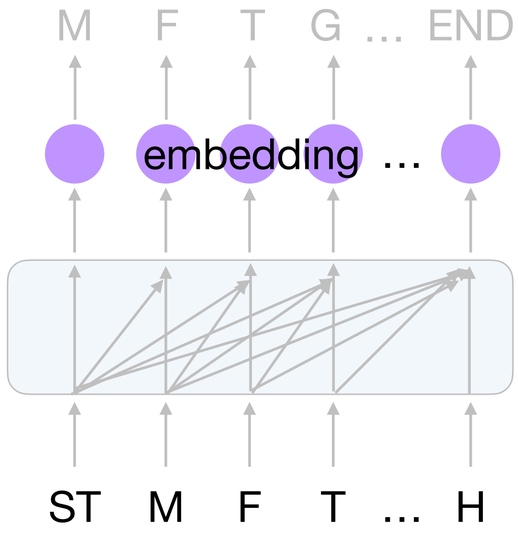
MFTGNDAGH

Many methods pretend proteins are language

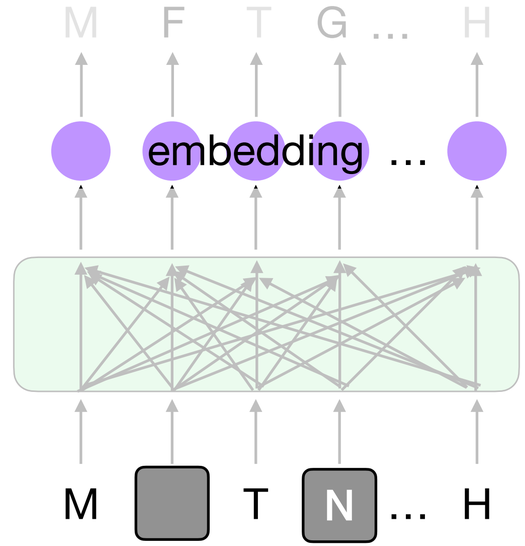
Word2Vec



Neural autoregressive language model



Neural masked language model

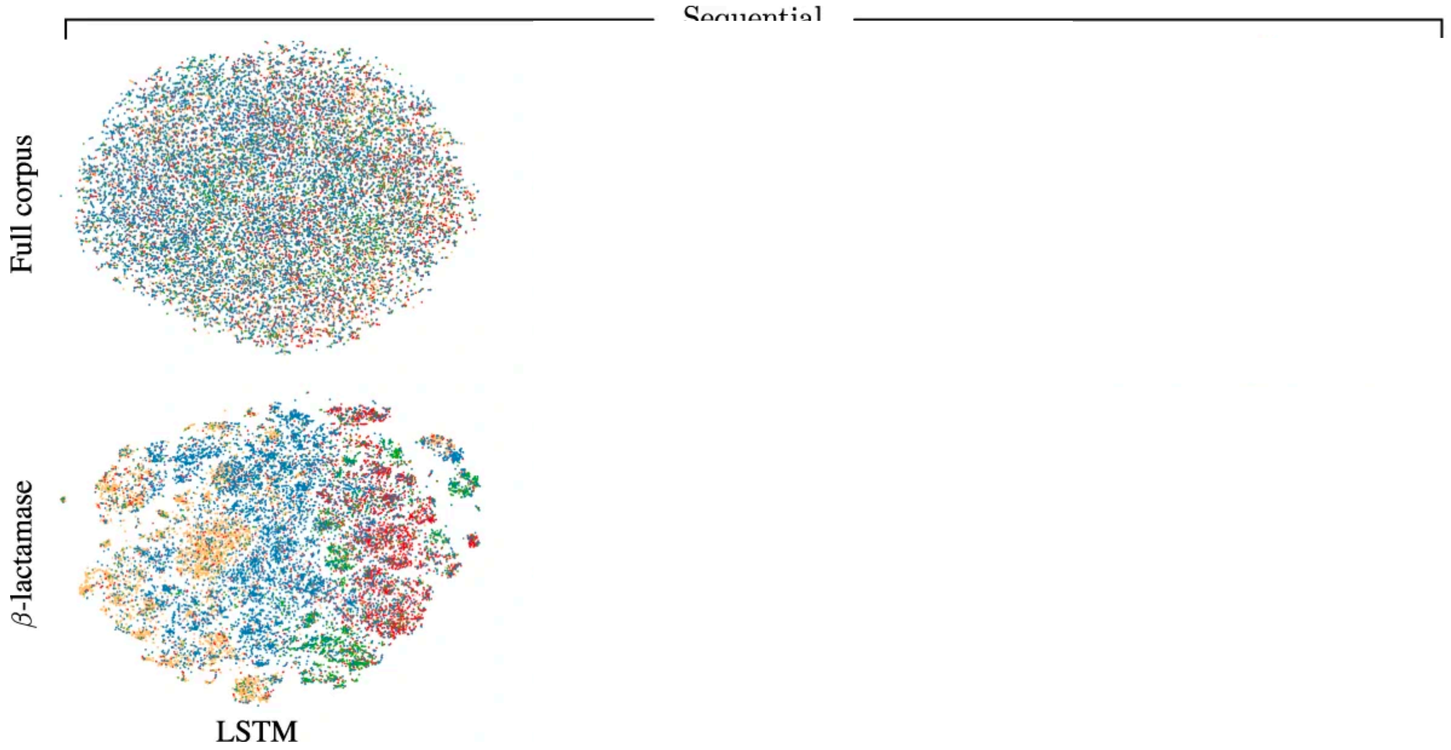


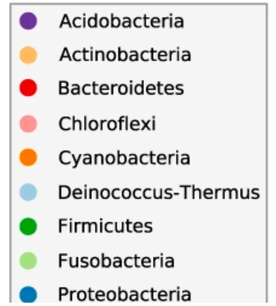
MFTGNDAGH

Architecture, dataset, and task all influence representation quality

Detlefsen *et al.* 2022

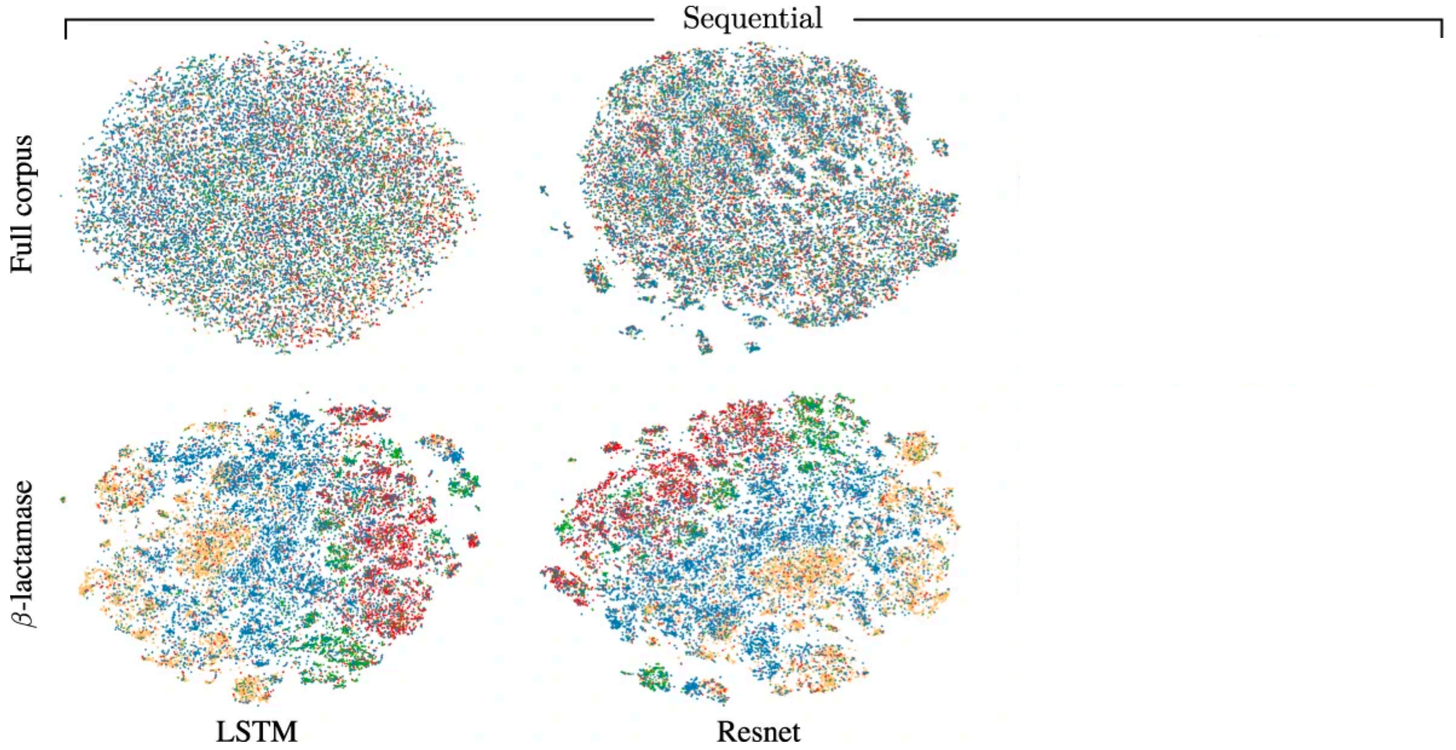
Architecture, dataset, and task all influence representation quality

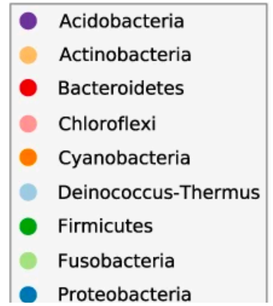




Detlefsen *et al.* 2022

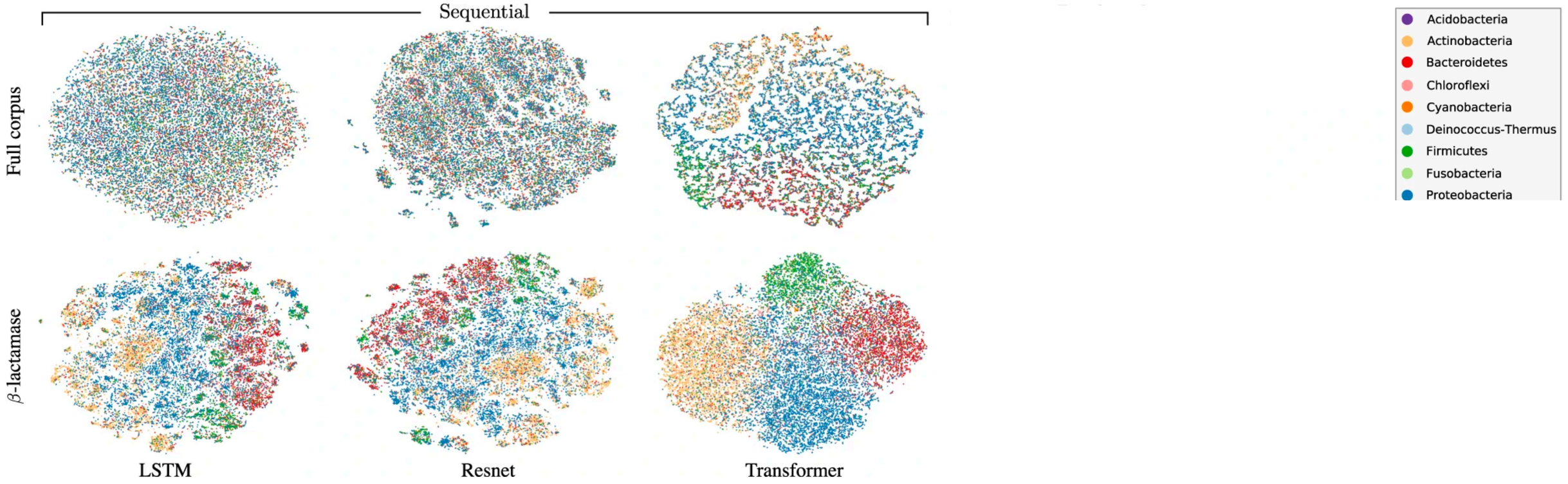
Architecture, dataset, and task all influence representation quality





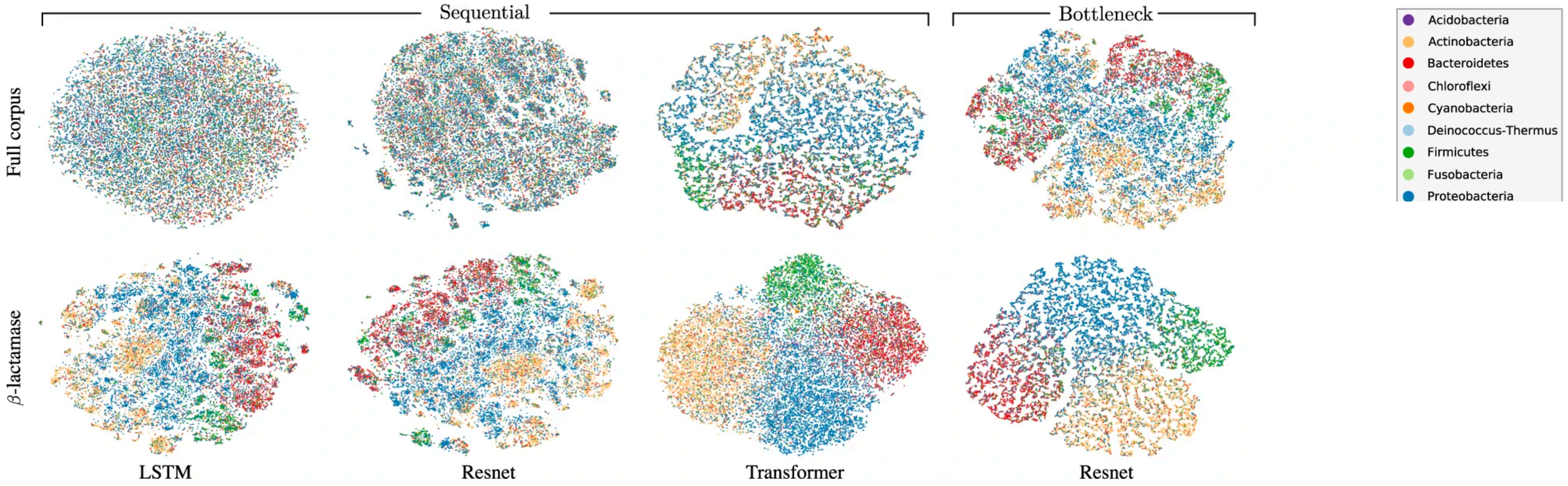
Detlefsen *et al.* 2022

Architecture, dataset, and task all influence representation quality



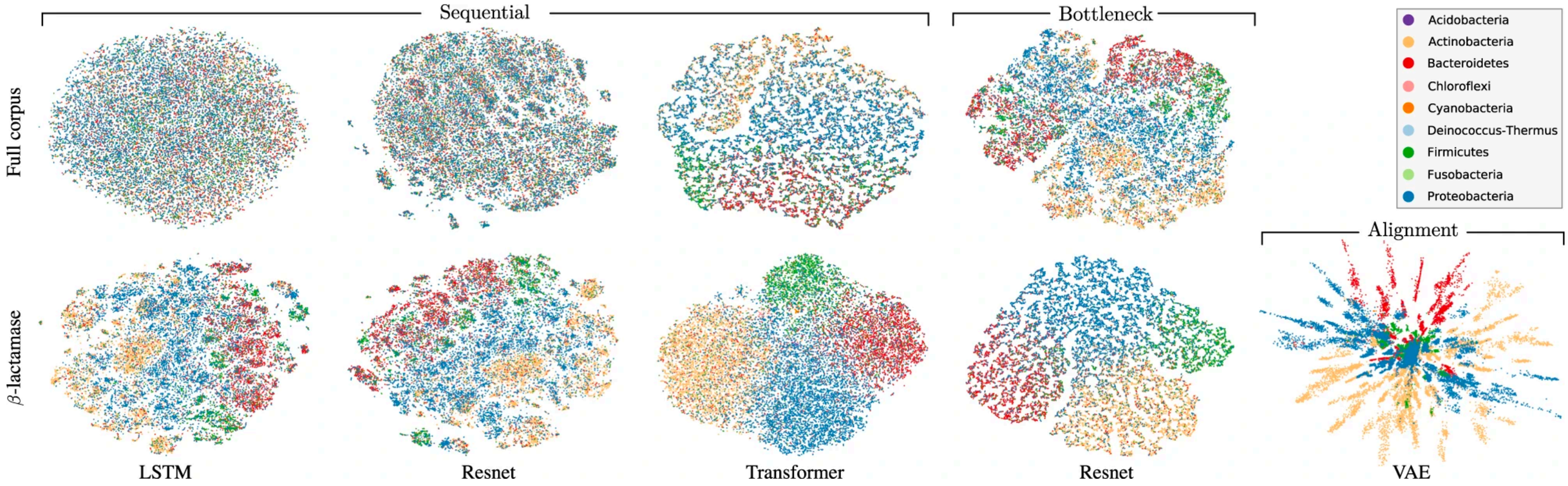
Detlefsen *et al.* 2022

Architecture, dataset, and task all influence representation quality



Detlefsen *et al.* 2022

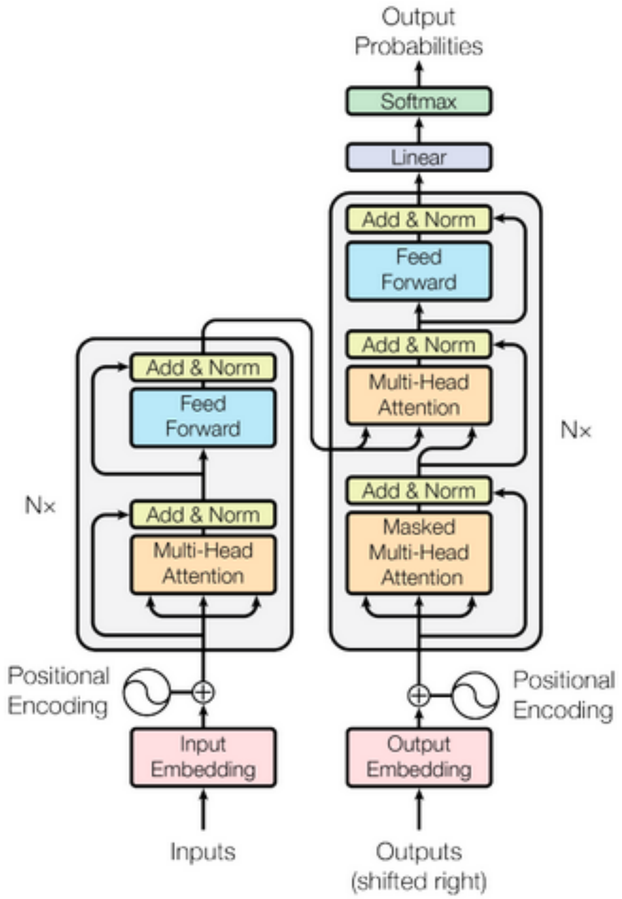
Architecture, dataset, and task all influence representation quality



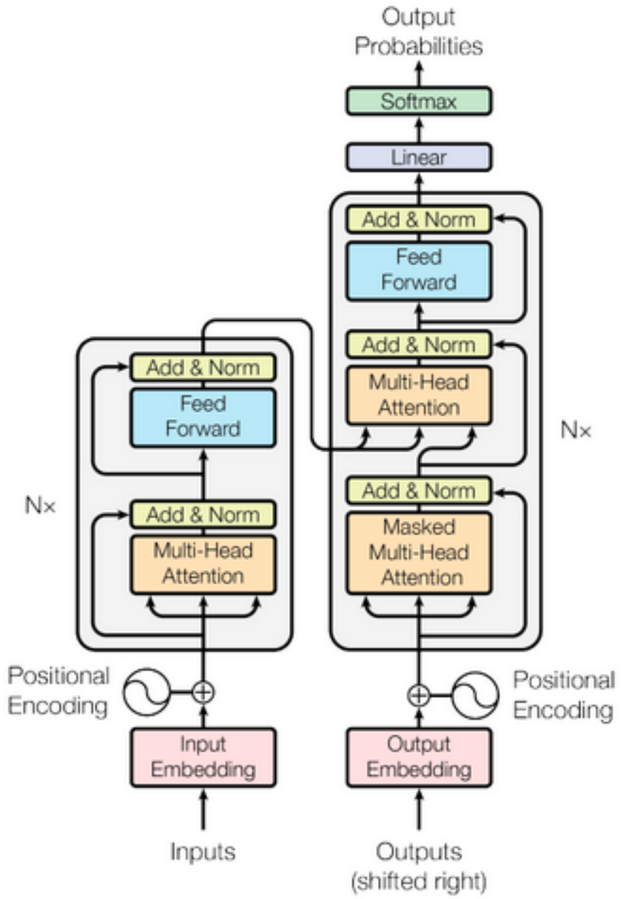
Detlefsen *et al.* 2022

Transformer masked language models are currently popular

Transformer masked language models are currently popular

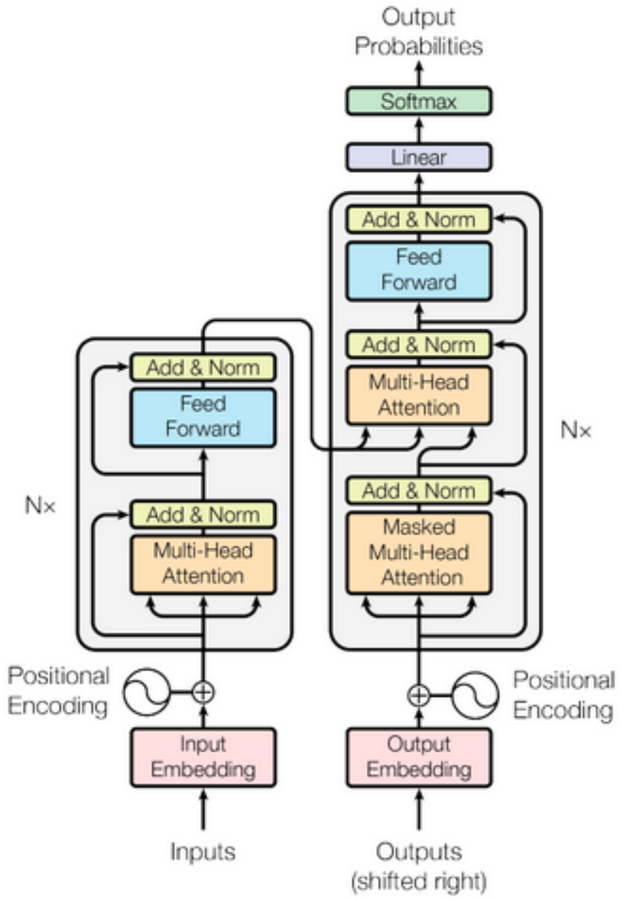


Transformer masked language models are currently popular



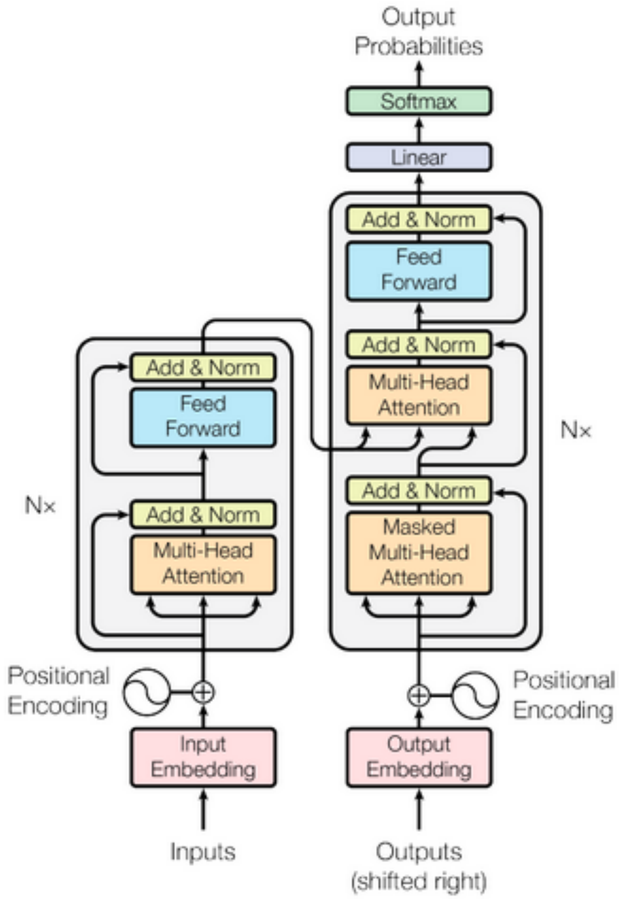
UniRef50
42M

Transformer masked language models are currently popular



UniRef50 42M → ESM-1b

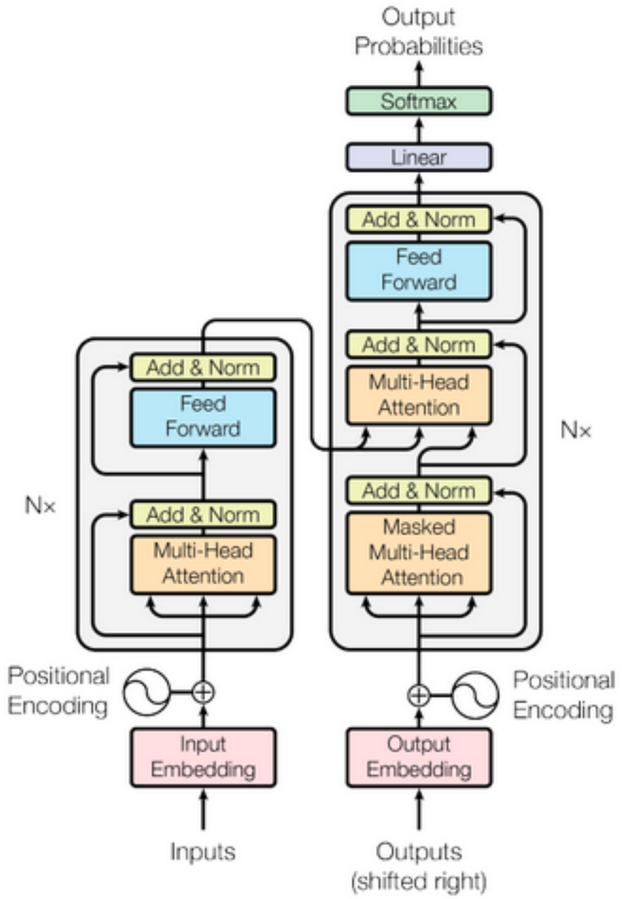
Transformer masked language models are currently popular



UniRef50 42M → ESM-1b

Best at structure

Transformer masked language models are currently popular

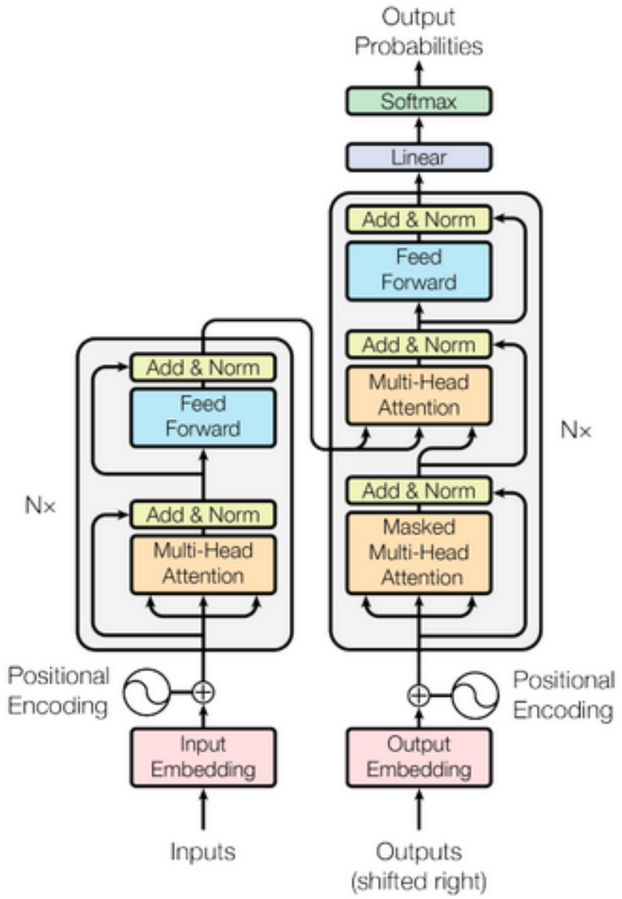


UniRef50 42M → ESM-1b

Best at structure

UniRef90 116M

Transformer masked language models are currently popular

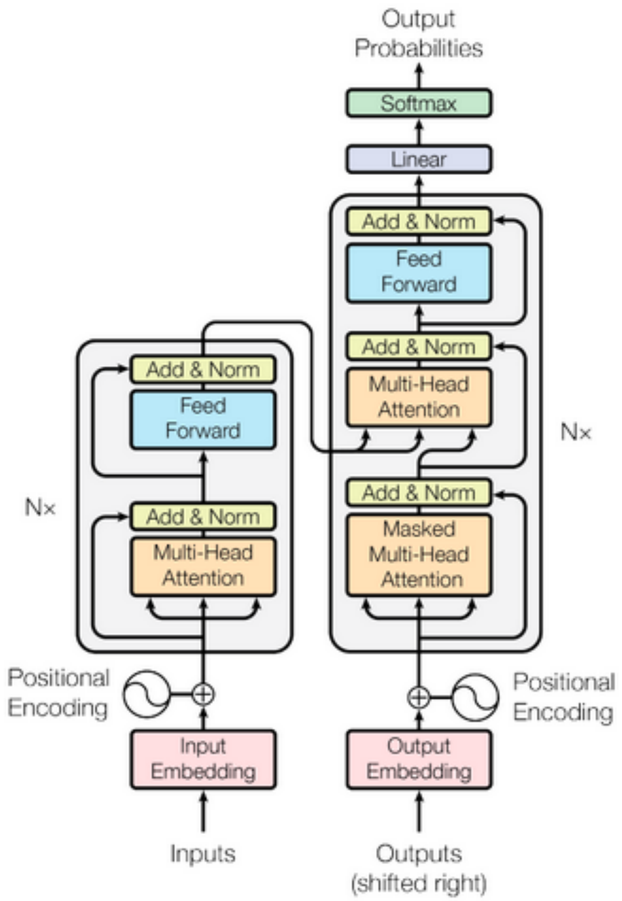


UniRef50 42M → ESM-1b

Best at structure

UniRef90 116M → ESM-1v

Transformer masked language models are currently popular



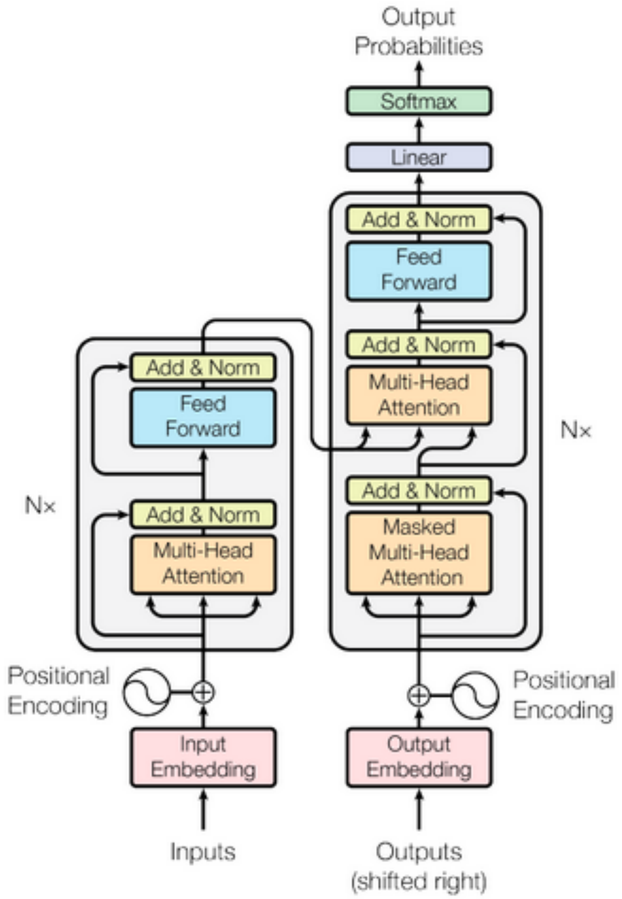
UniRef50 42M → ESM-1b

Best at structure

UniRef90 116M → ESM-1v

Best at variant effect prediction

Transformer masked language models are currently popular



UniRef50 42M → ESM-1b

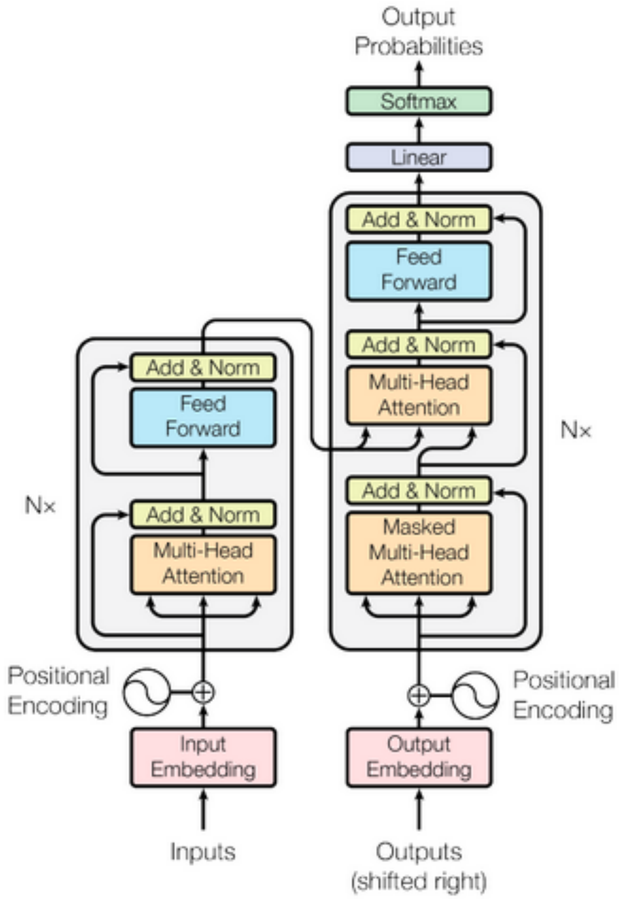
Best at structure

UniRef90 116M → ESM-1v

Best at variant effect prediction

BFD 1.2B

Transformer masked language models are currently popular



UniRef50 42M → ESM-1b

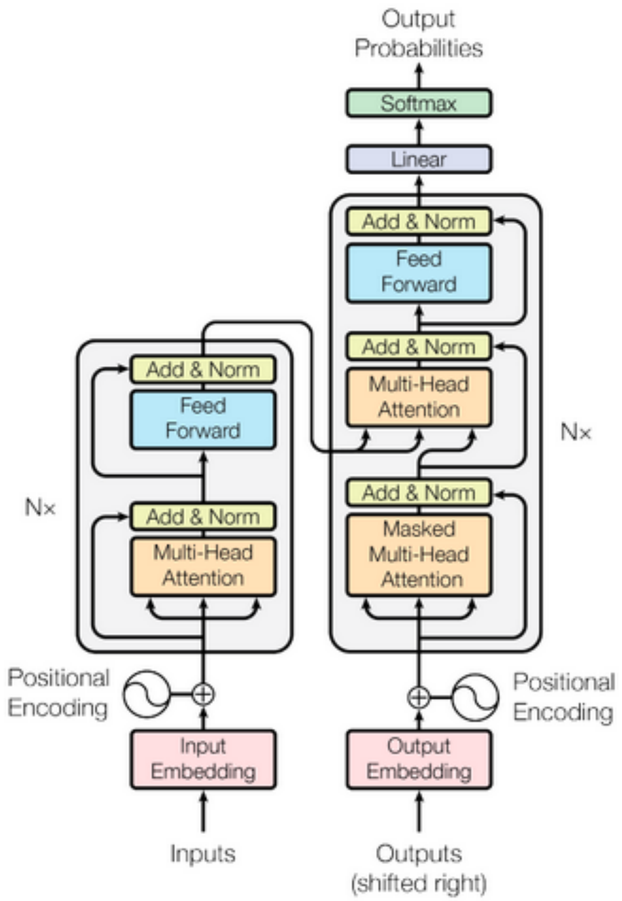
Best at structure

UniRef90 116M → ESM-1v

Best at variant effect prediction

BFD 1.2B → ProtBERT

Transformer masked language models are currently popular



UniRef50 42M → ESM-1b

Best at structure

UniRef90 116M → ESM-1v

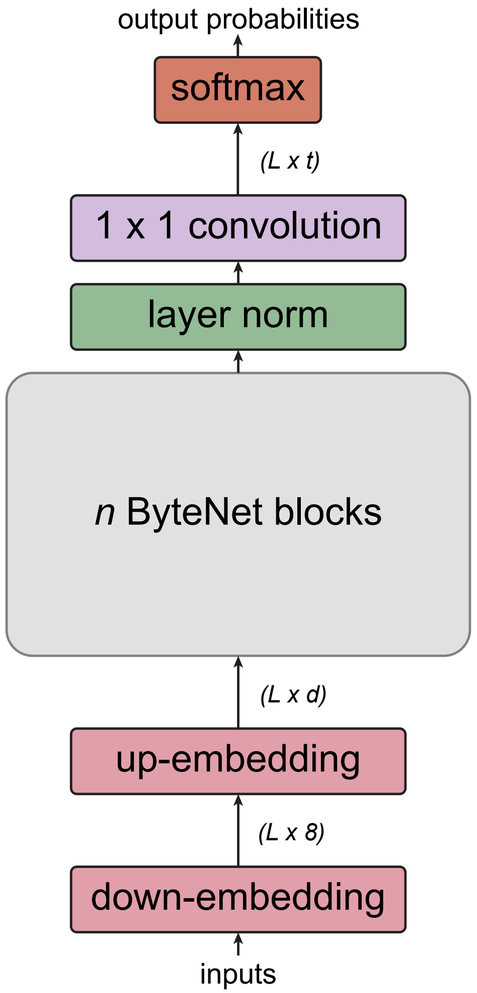
Best at variant effect prediction

BFD 1.2B → ProtBERT

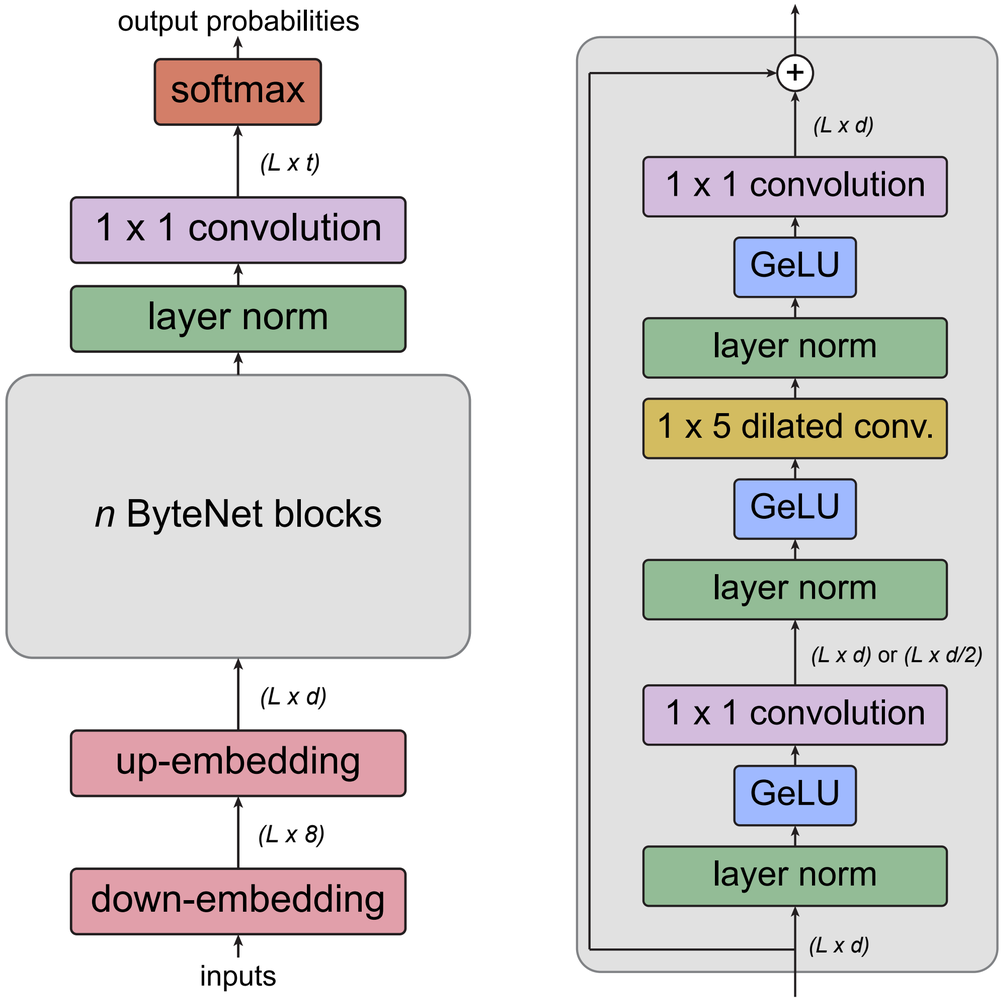
Having many more (redundant) sequences doesn't help

CNNs are competitive with transformers for pretraining

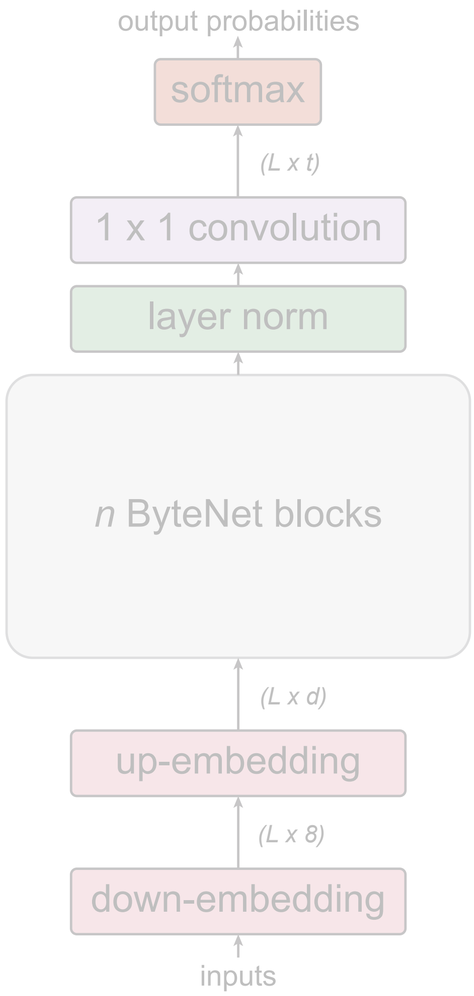
CNNs are competitive with transformers for pretraining

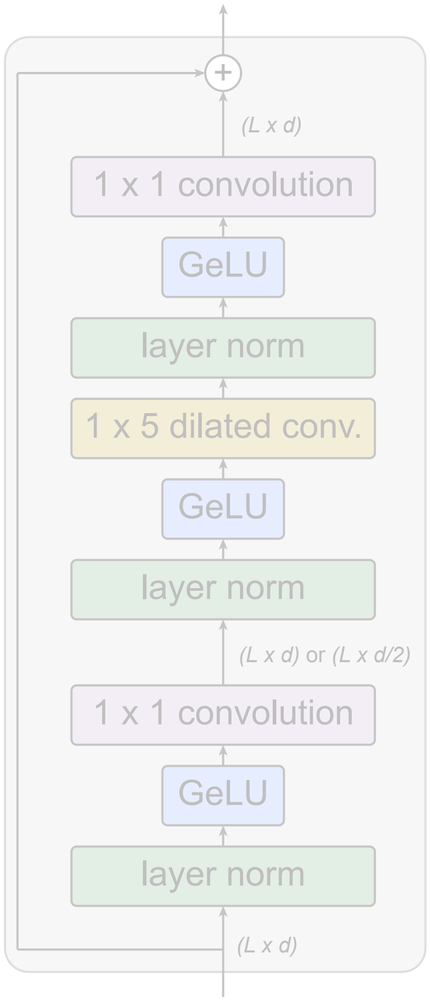


CNNs are competitive with transformers for pretraining

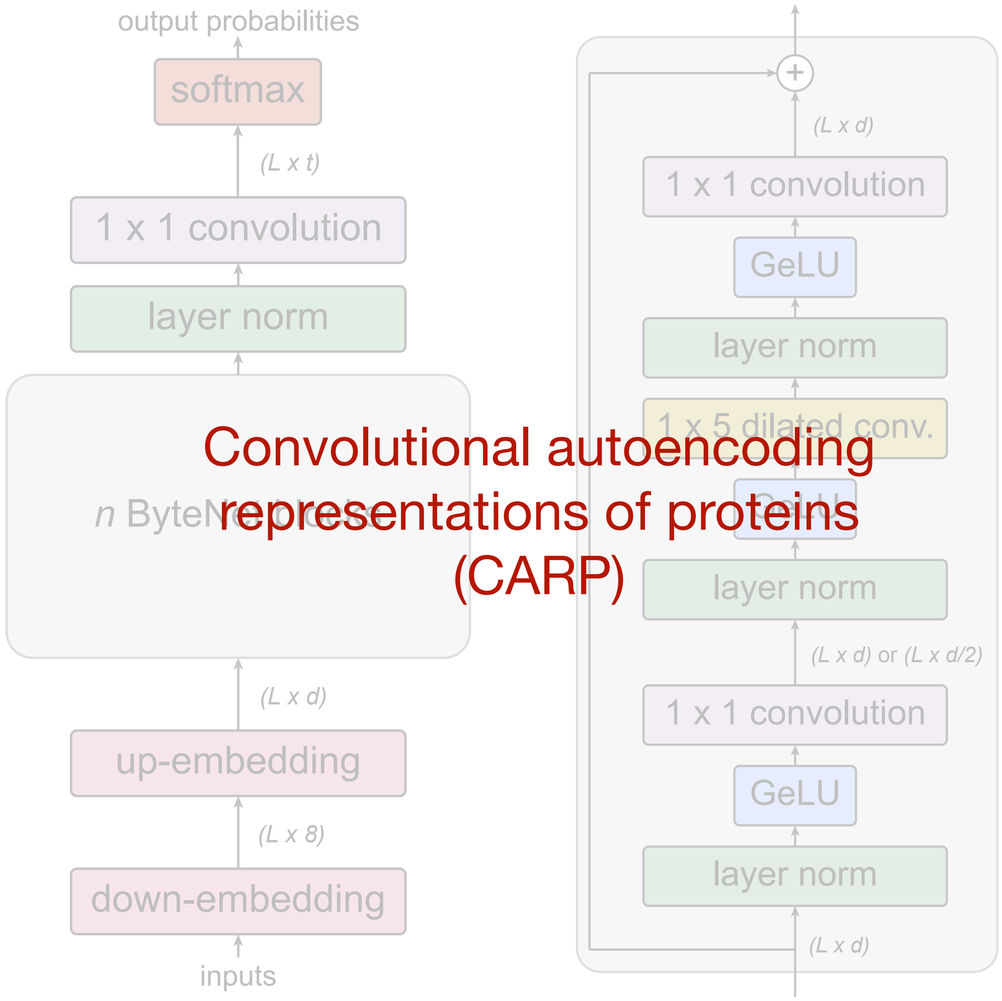


CNNs are competitive with transformers for pretraining



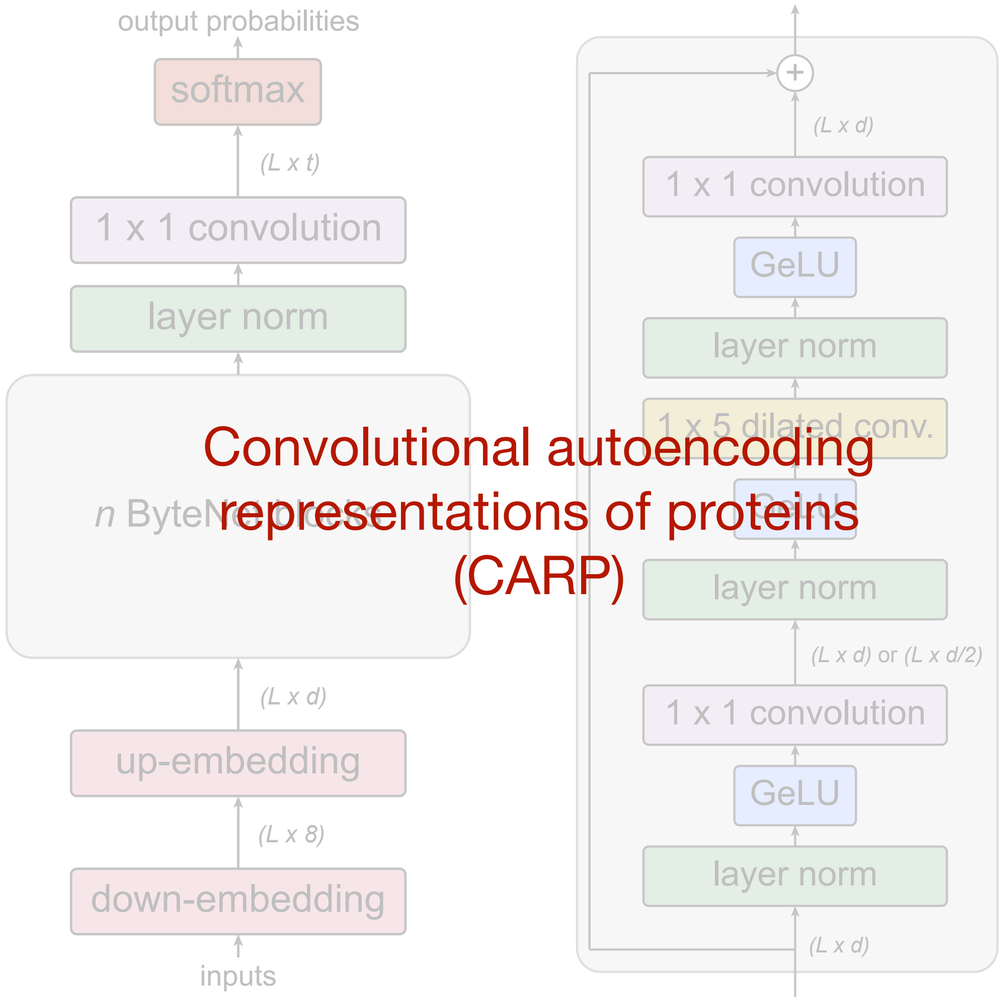


CNNs are competitive with transformers for pretraining

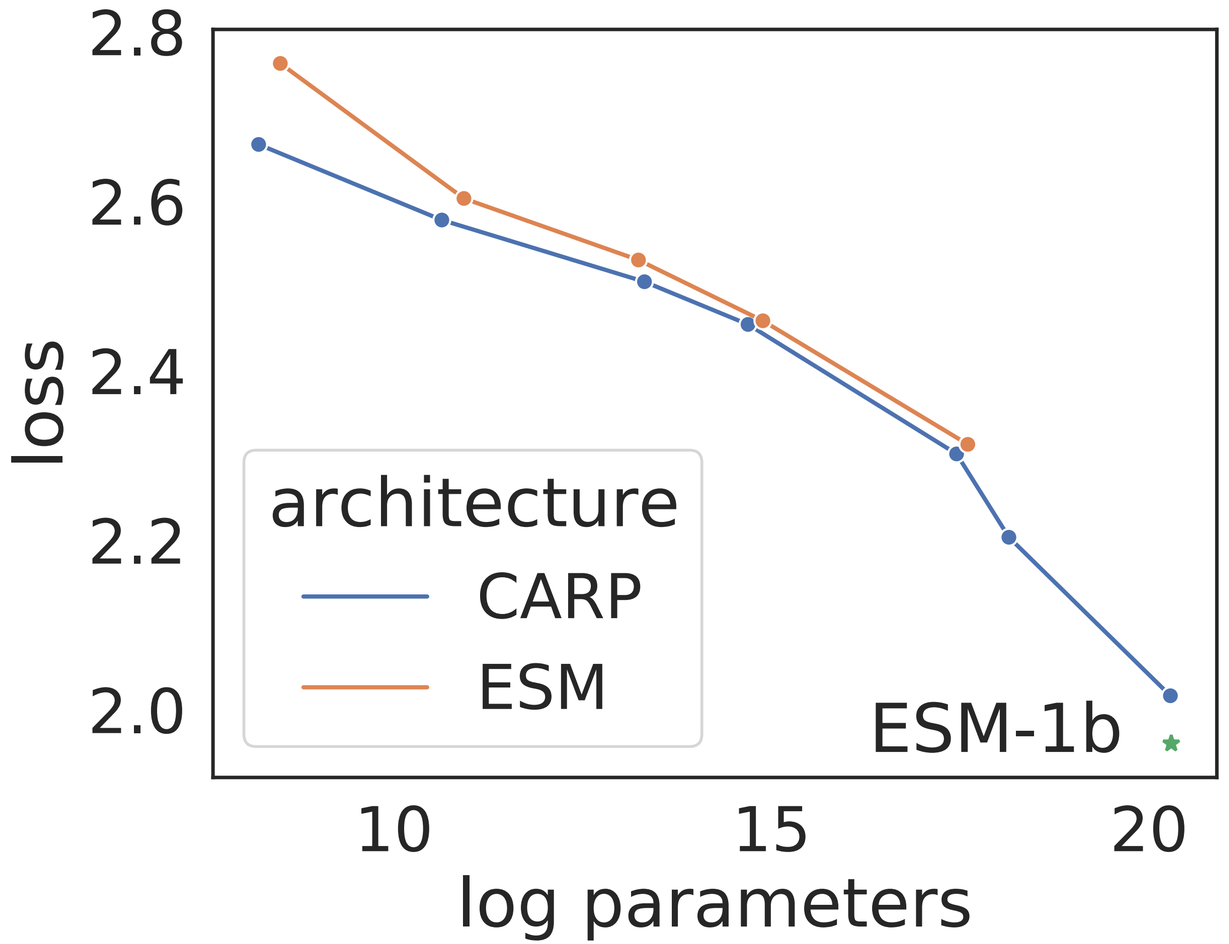


Convolutional autoencoding representations of proteins (CARP)

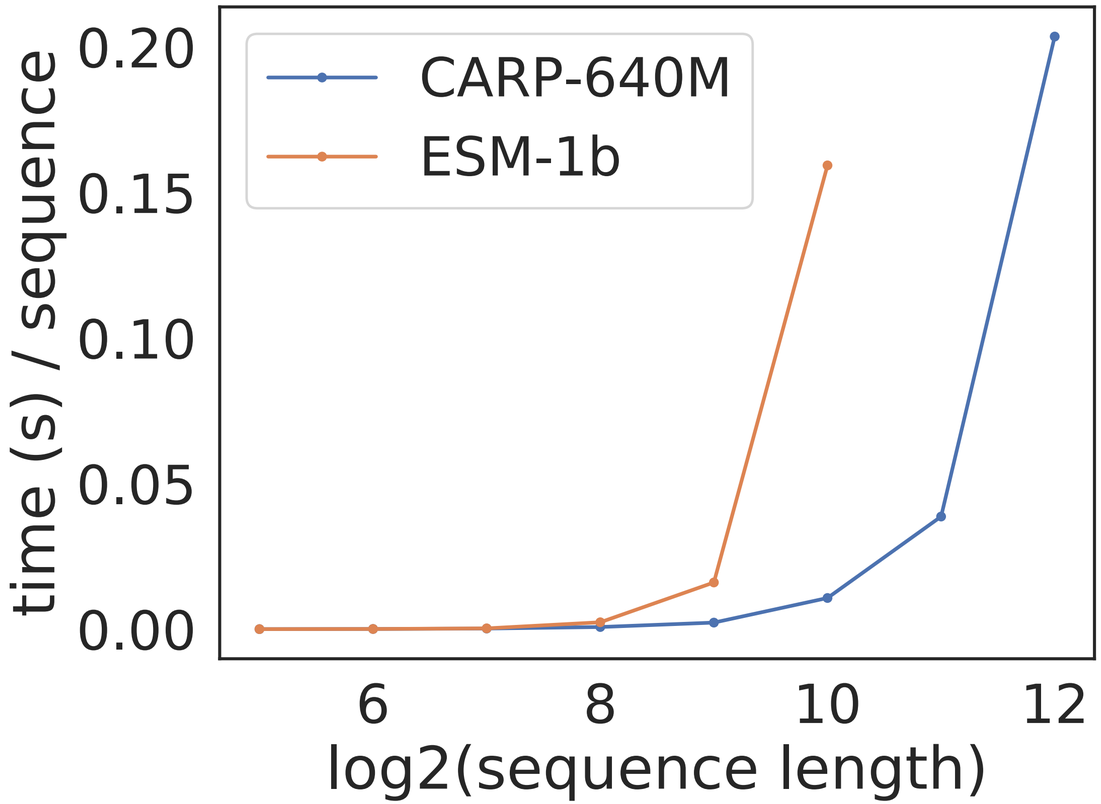
CNNs are competitive with transformers for pretraining



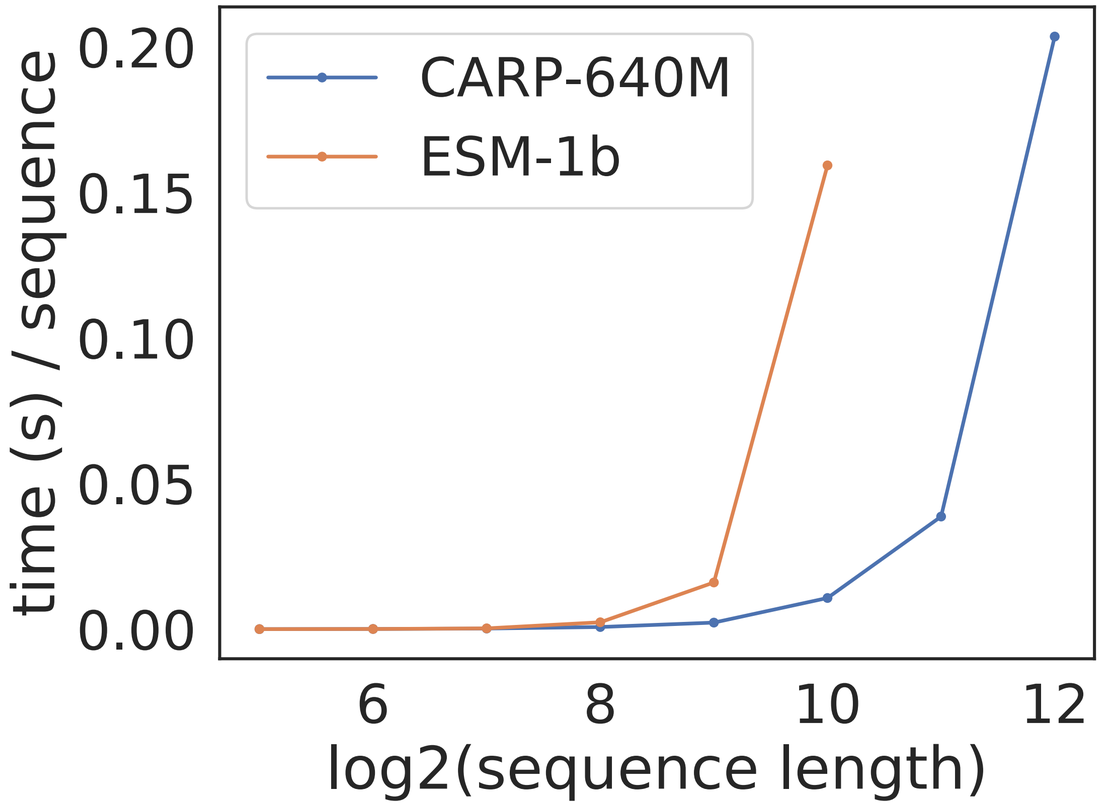
Convolutional autoencoding representations of proteins (CARP)

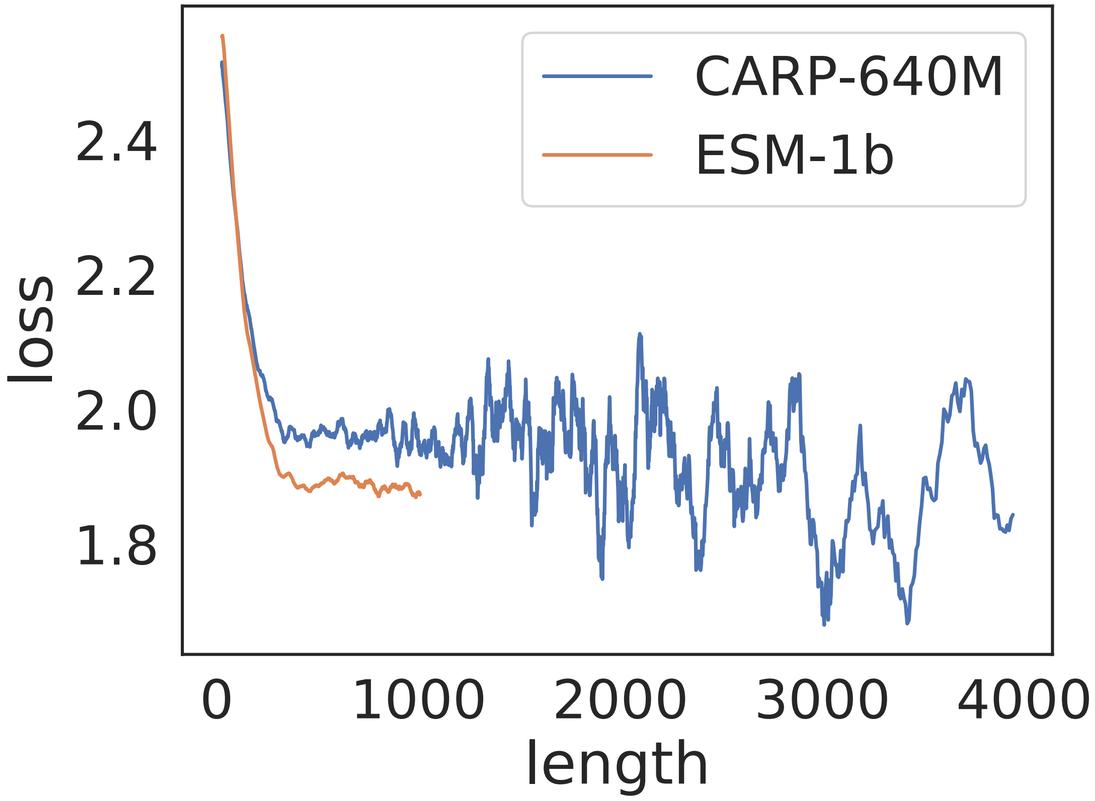


CNNs scale better with length

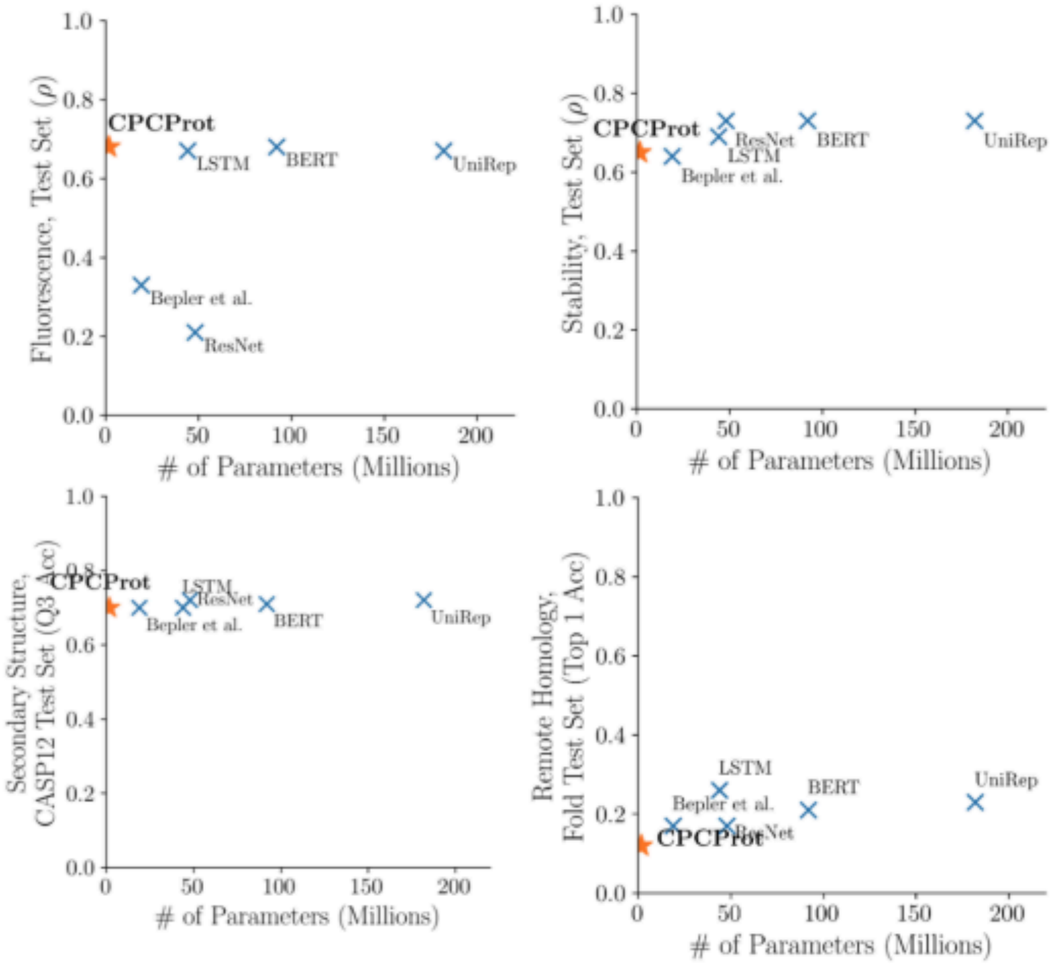


CNNs scale better with length



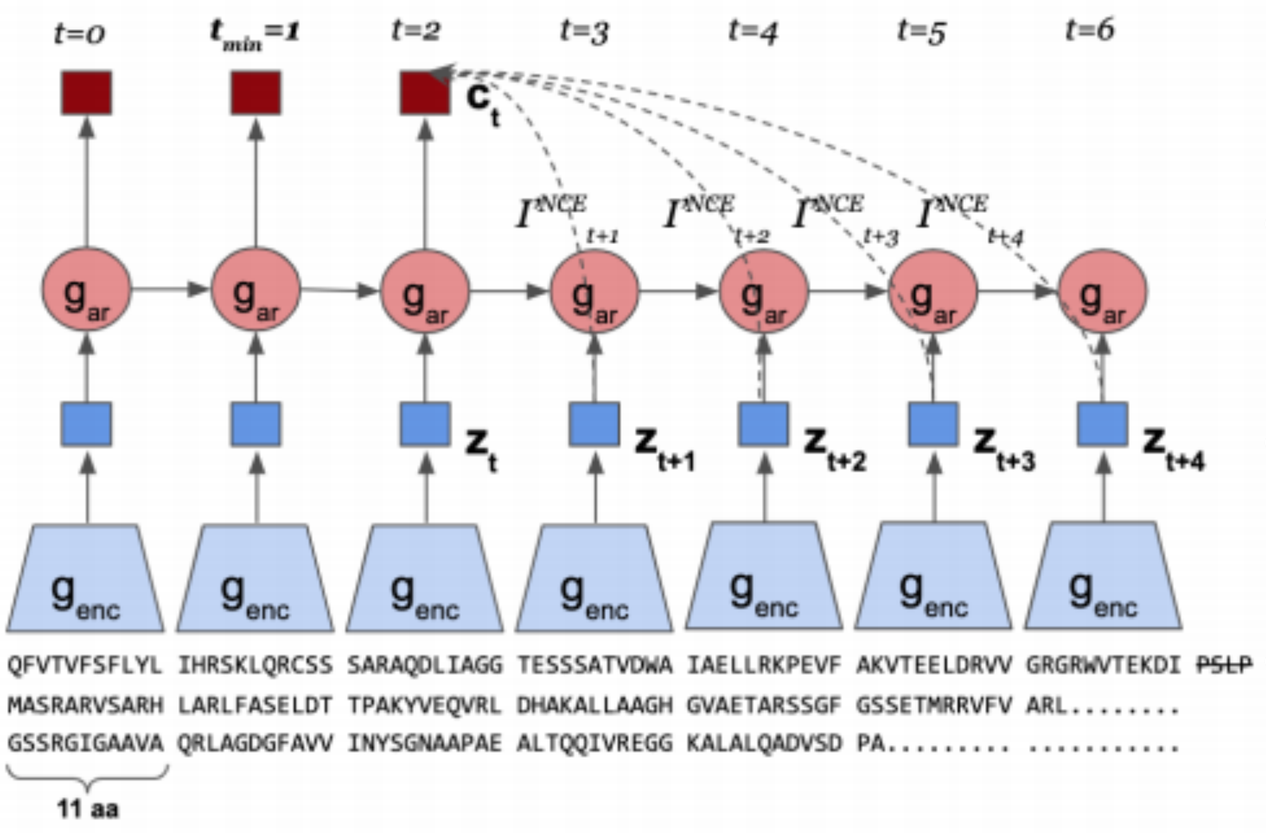


Contrastive learning may require fewer parameters



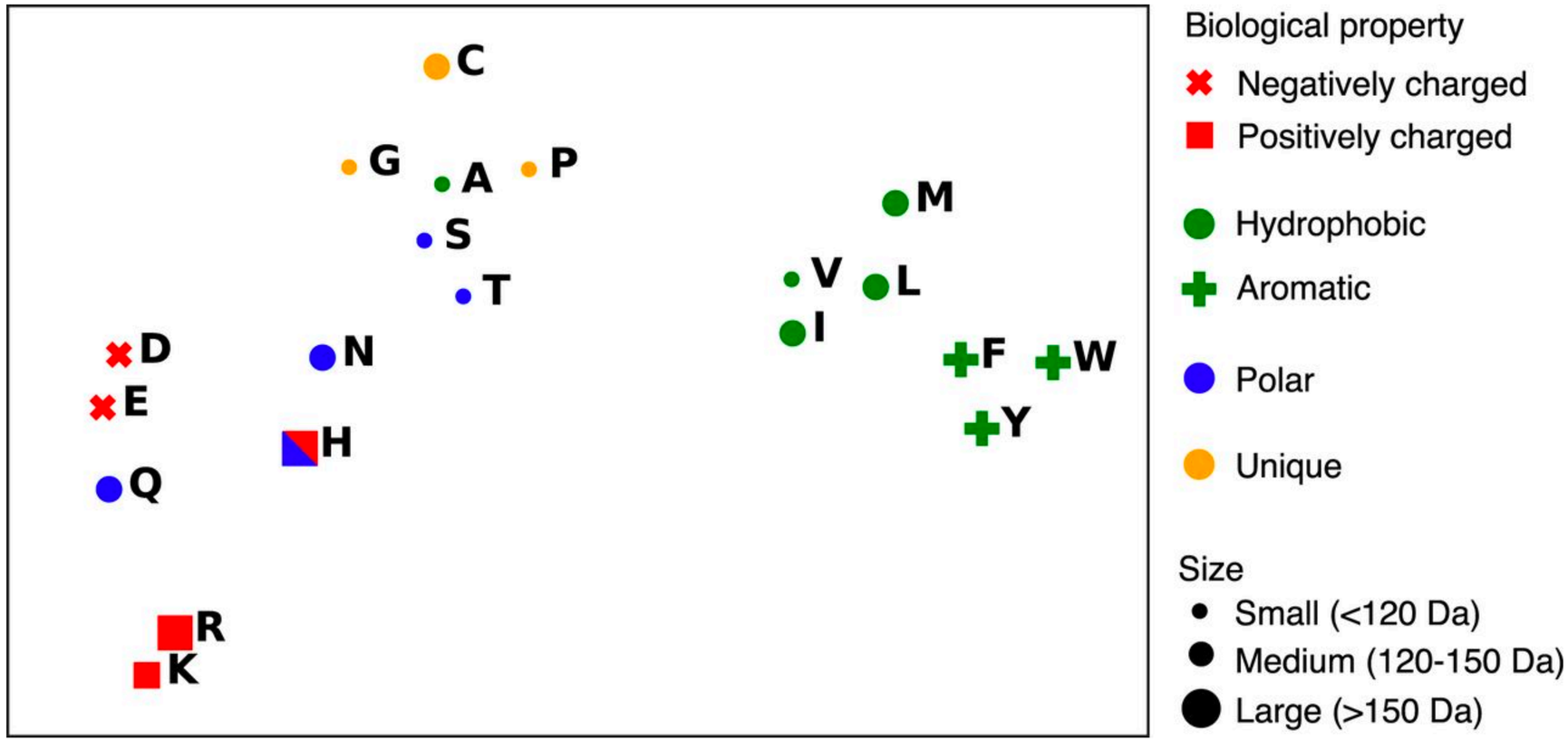
	# of Embedding Parameters	Remote Homology			Secondary Structure			Stability	Fluorescence
		Fold	Superfamily	Family	CB513	CASP12	TS115		
Unirep	182M	0.23	0.38	0.87	0.73	**0.72**	0.77	**0.73**	0.67
BERT	92M	0.21	0.34	0.88	0.73	0.71	0.77	**0.73**	**0.68**
ResNet	48M	0.17	0.31	0.77	**0.75**	**0.72**	**0.78**	**0.73**	0.21
LSTM	44M	**0.26**	**0.43**	**0.92**	**0.75**	0.70	**0.78**	0.69	0.67
Bepler et al.	19M	0.17	0.20	0.79	0.73	0.70	0.76	0.64	0.33
One Hot	0	0.09	0.08	0.39	0.69	0.68	0.72	0.19	0.14
CPCProt	1.7M	0.12	0.12	0.48	0.69	0.70	0.73	0.65	**0.68**
$\text{CPCProt}_{\text{GRU\_large}}$	8.4M	0.13	0.14	0.52	0.70	0.70	0.73	0.65	**0.68**
$\text{CPCProt}_{\text{LSTM}}$	71M	0.11	0.11	0.47	0.68	0.66	0.70	0.68	**0.68**

But is it really the contrastive learning that helps?



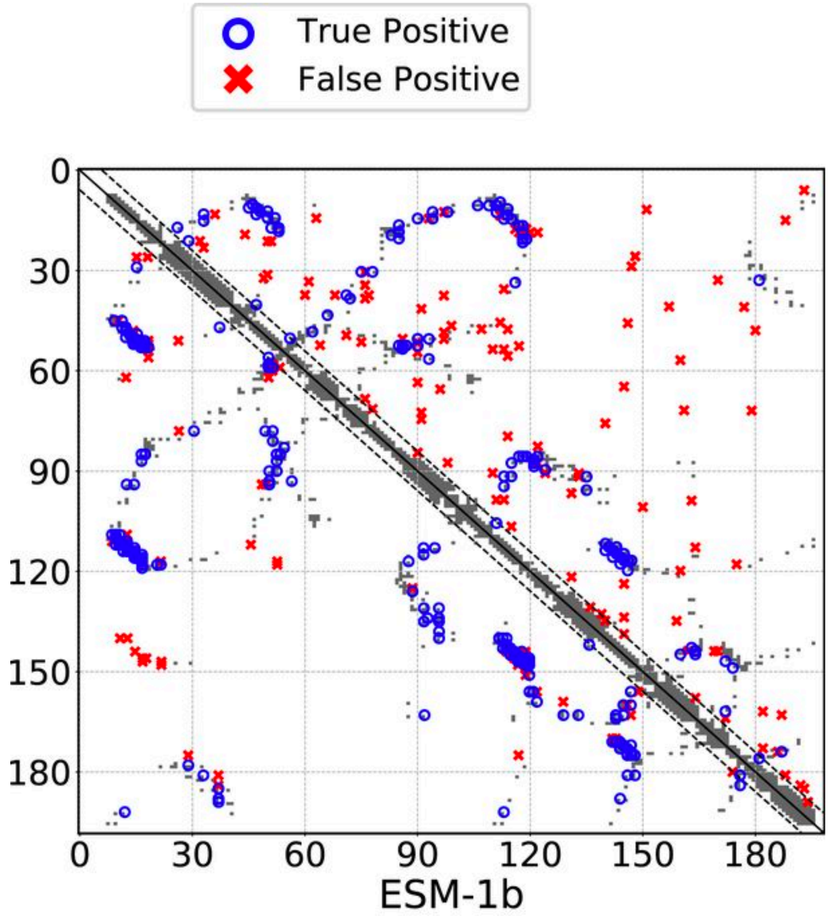
- In images, the augmentation pipeline is most important
- Maybe we just need to be predicting larger blocks?
- Or have better negatives

Deep sequence embeddings recapitulate biophysical properties



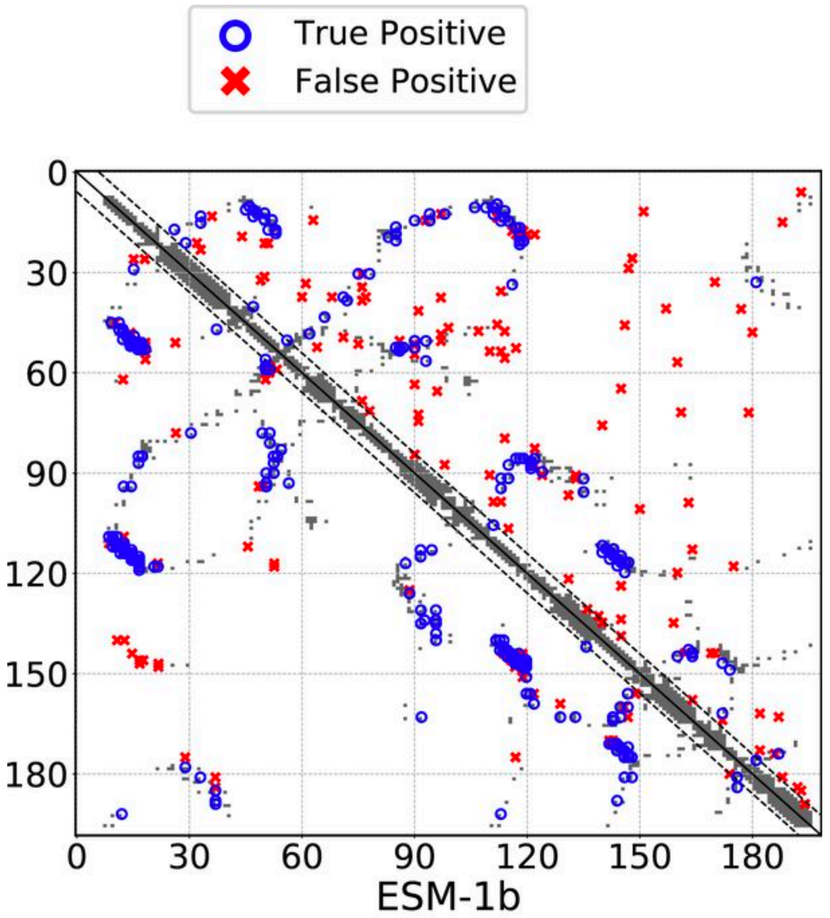
Rives *et al. PNAS* 2021

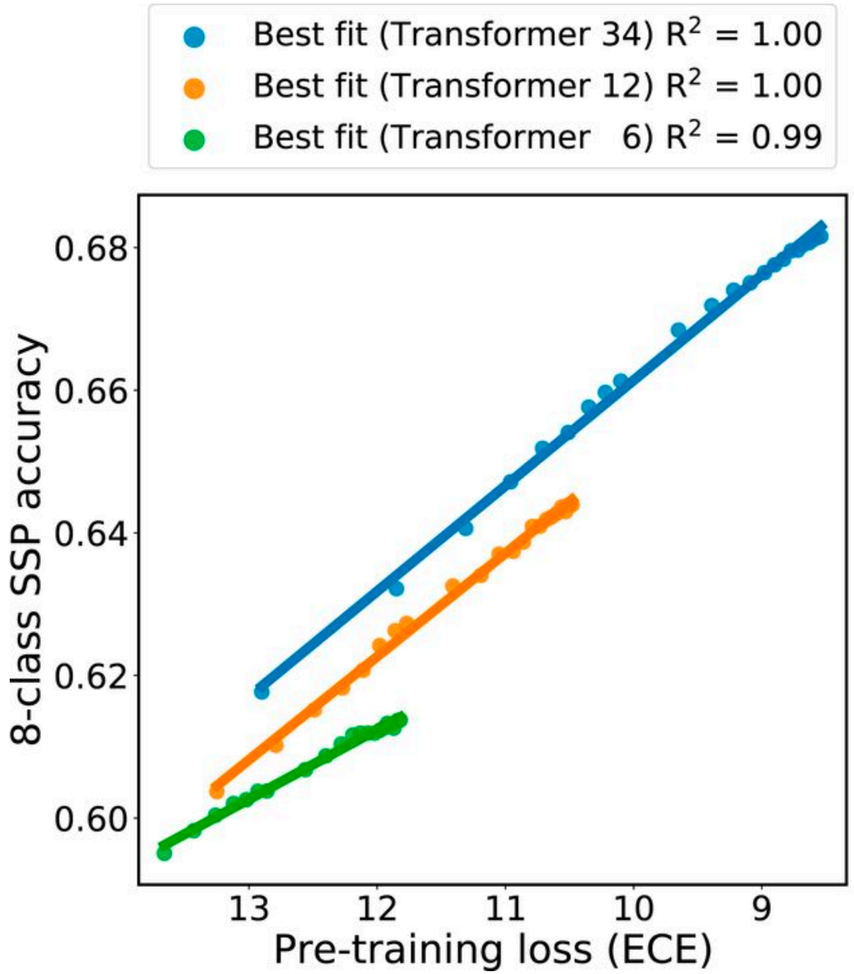
Deep sequence embeddings contain structural information



Rives *et al. PNAS* 2021

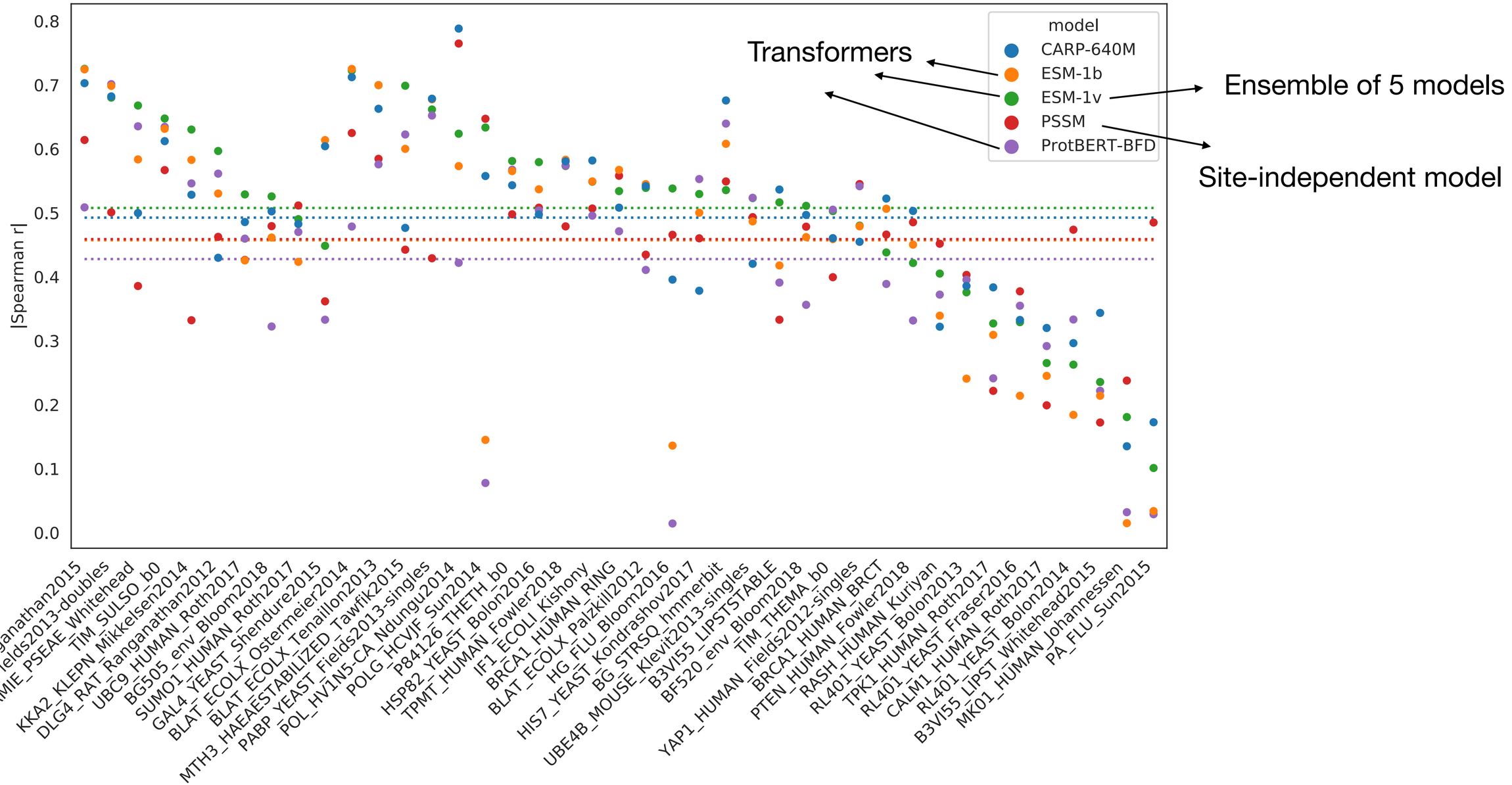
Deep sequence embeddings contain structural information





Rives *et al. PNAS* 2021

MLM pseudo likelihood predicts mutant fitness



Pretrained models contain evolutionary information

Hie *et al.* 2022

Pretrained models contain evolutionary information

Pseudolikelihood correlates with mutational tolerance

Hie *et al.* 2022

Pretrained models contain evolutionary information

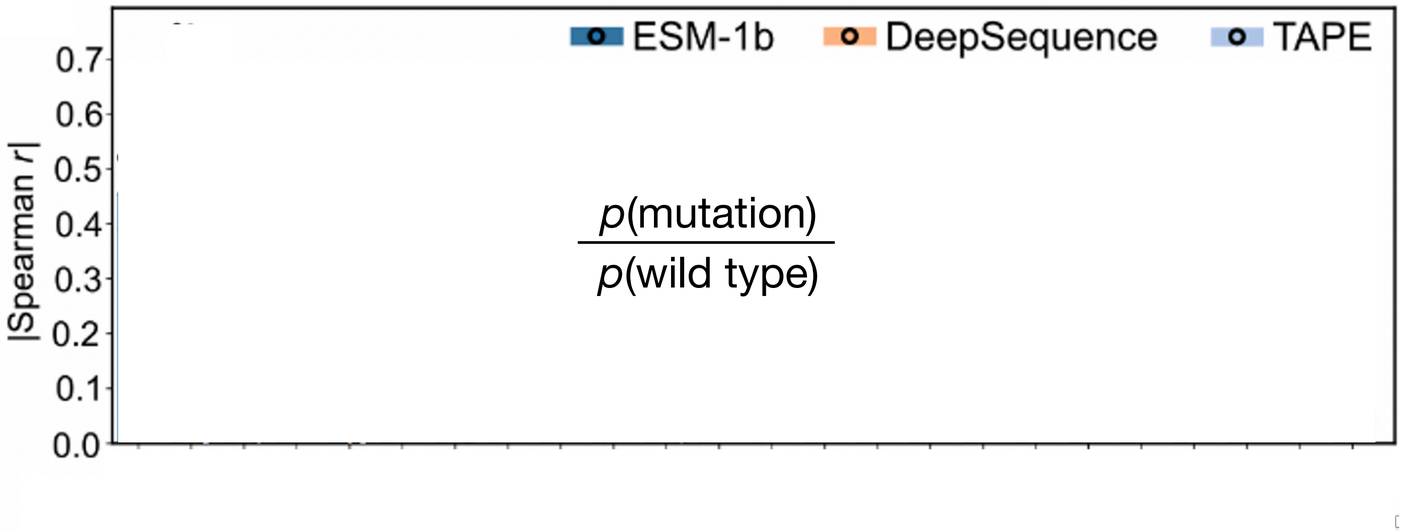
Pseudolikelihood correlates with mutational tolerance

$$\frac{p(\text{mutation})}{p(\text{wild type})}$$

Hie *et al.* 2022

Pretrained models contain evolutionary information

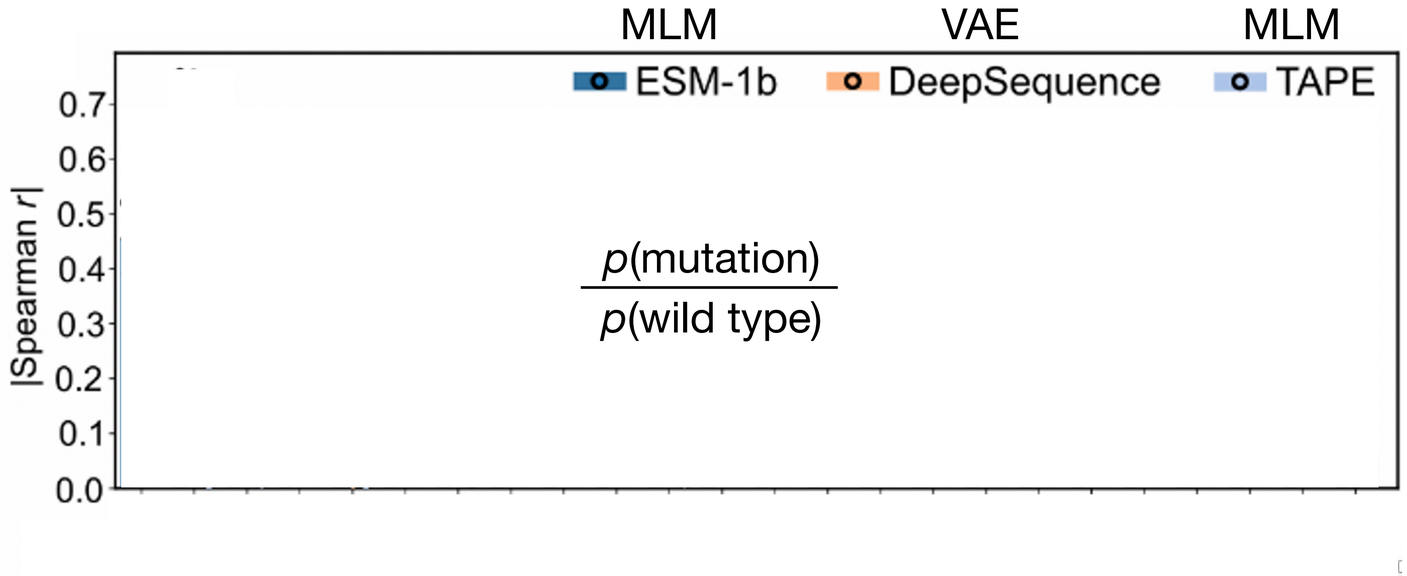
Pseudolikelihood correlates with mutational tolerance



Hie *et al.* 2022

Pretrained models contain evolutionary information

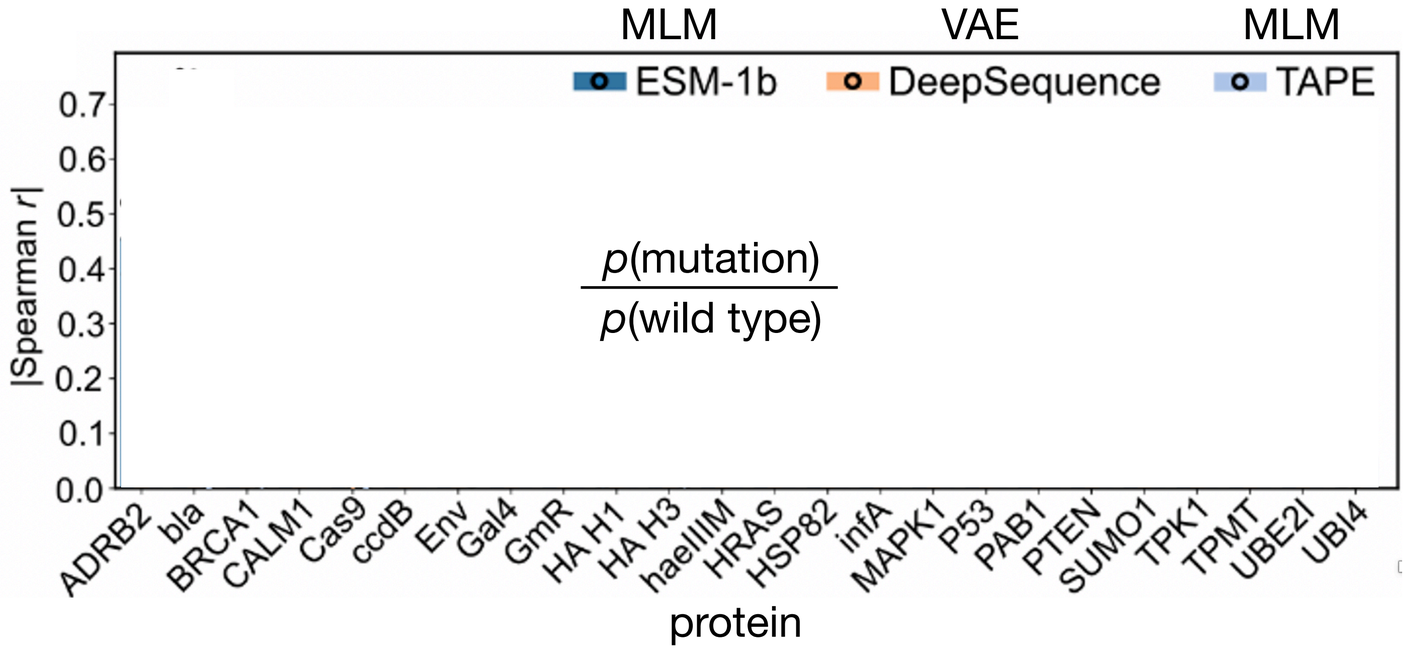
Pseudolikelihood correlates with mutational tolerance



Hie *et al.* 2022

Pretrained models contain evolutionary information

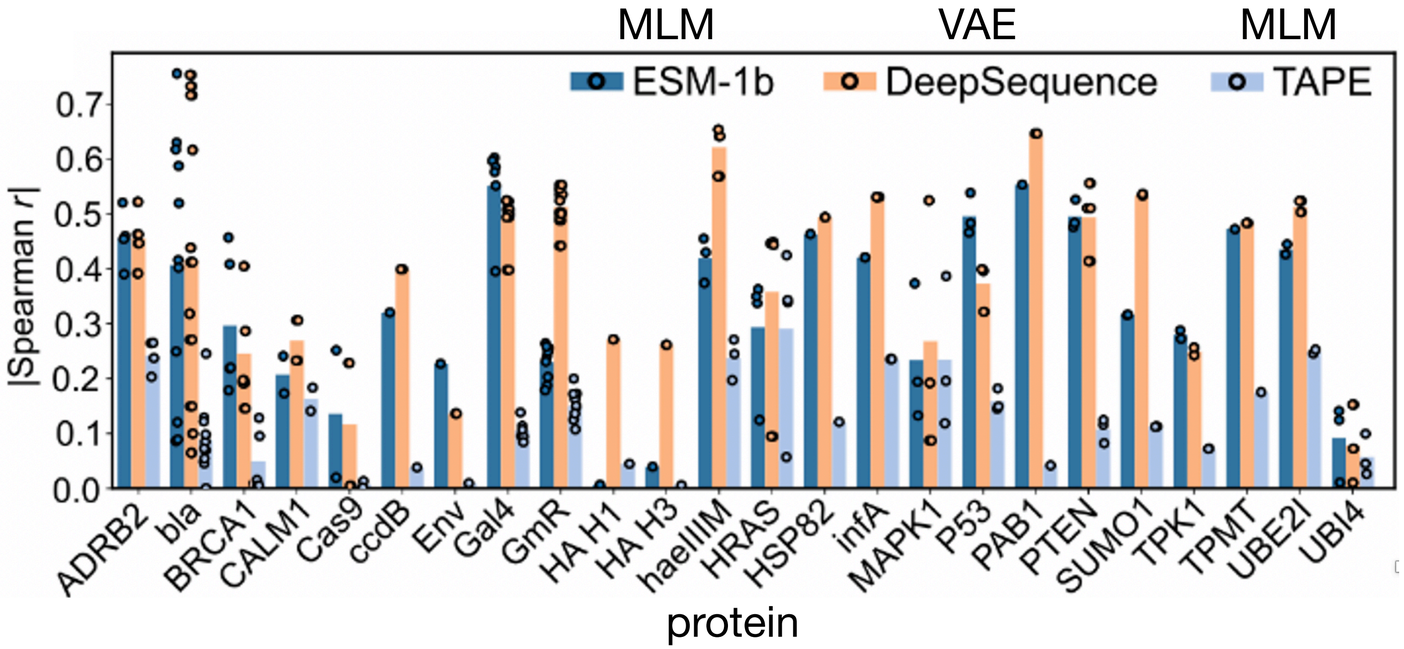
Pseudolikelihood correlates with mutational tolerance



Hie *et al.* 2022

Pretrained models contain evolutionary information

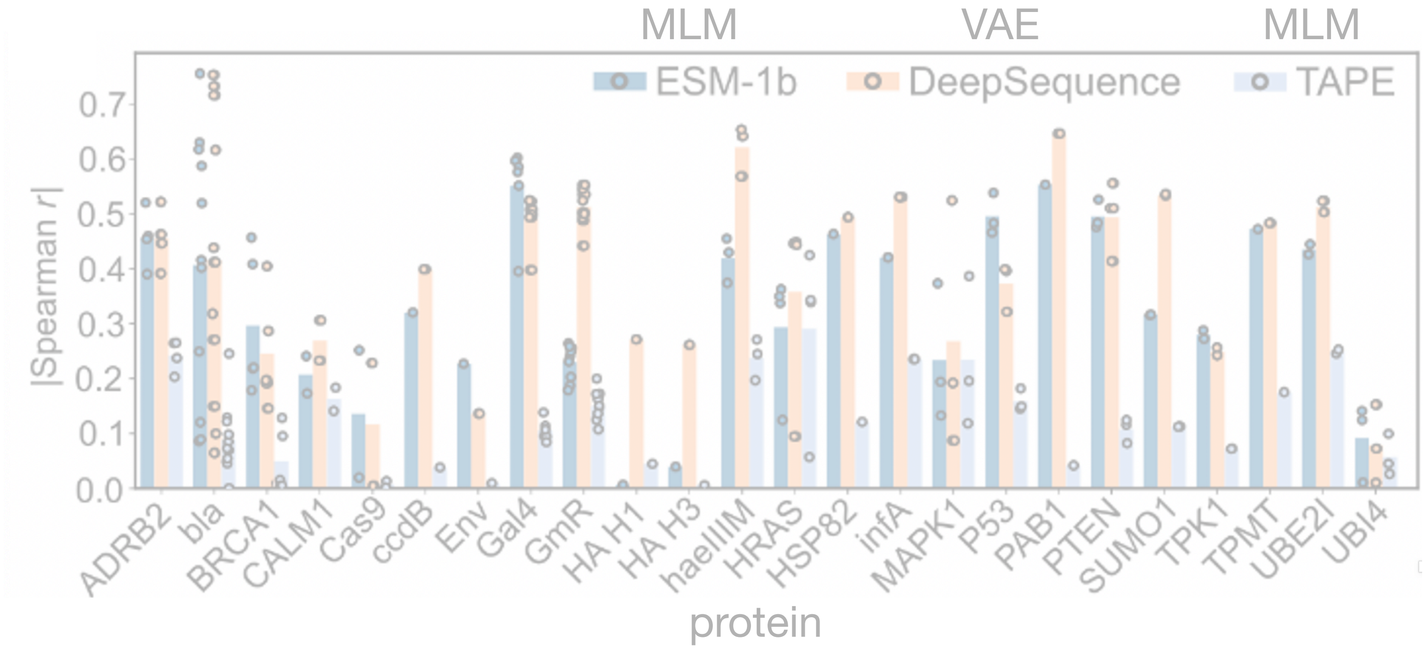
Pseudolikelihood correlates with mutational tolerance



Hie *et al.* 2022

Pretrained models contain evolutionary information

Pseudolikelihood correlates with mutational tolerance

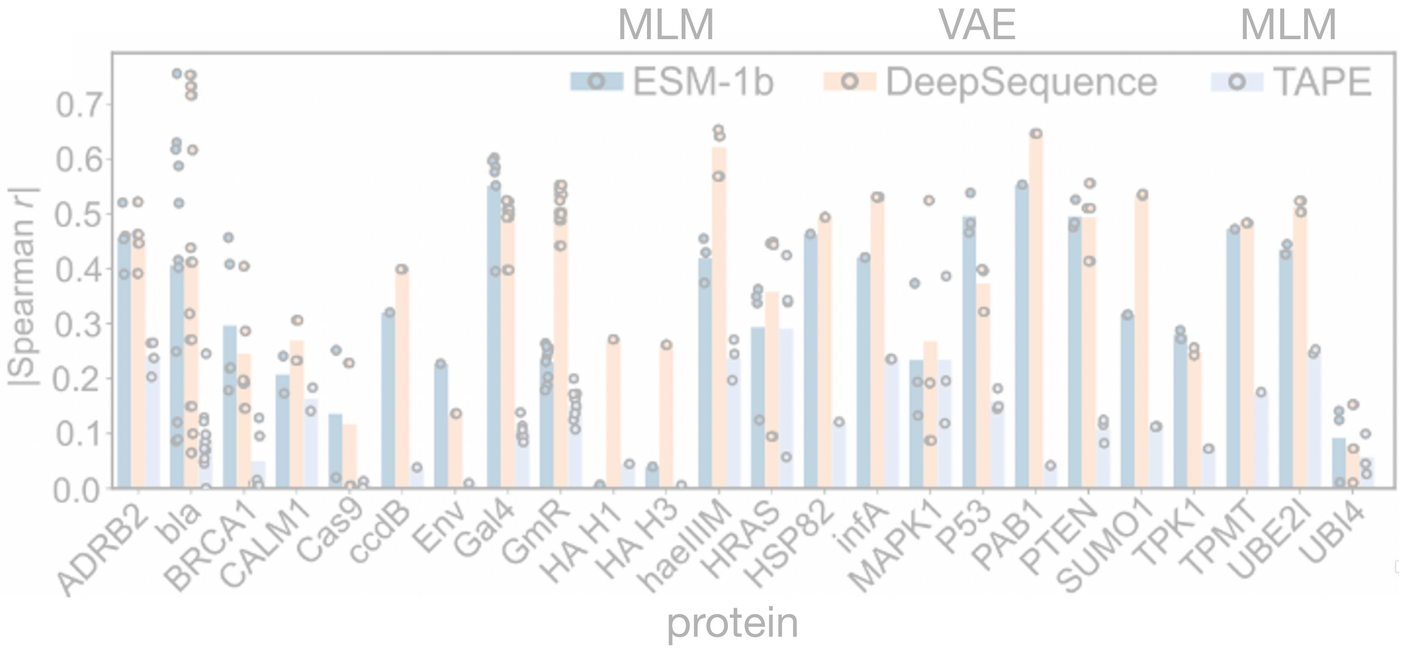


Use pseudolikelihood to predict evolutionary vector field

Hie *et al.* 2022

Pretrained models contain evolutionary information

Pseudolikelihood correlates with mutational tolerance

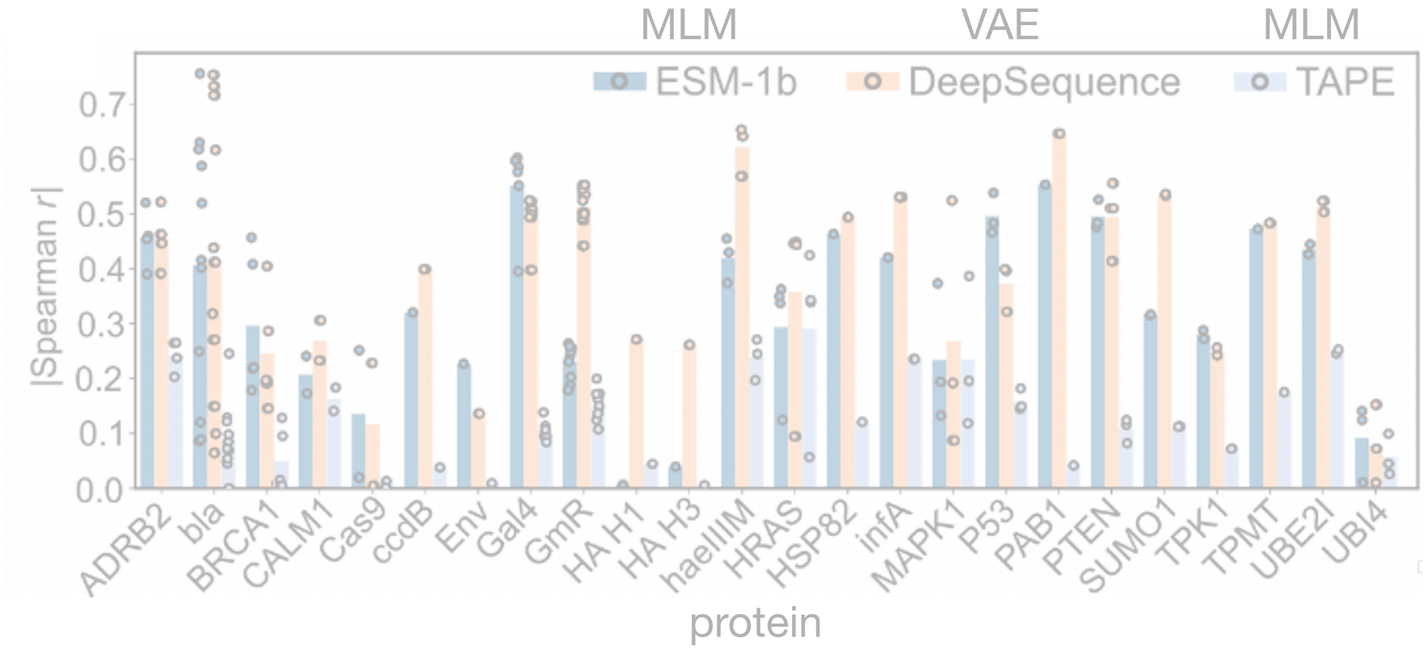


Use pseudolikelihood to predict evolutionary vector field

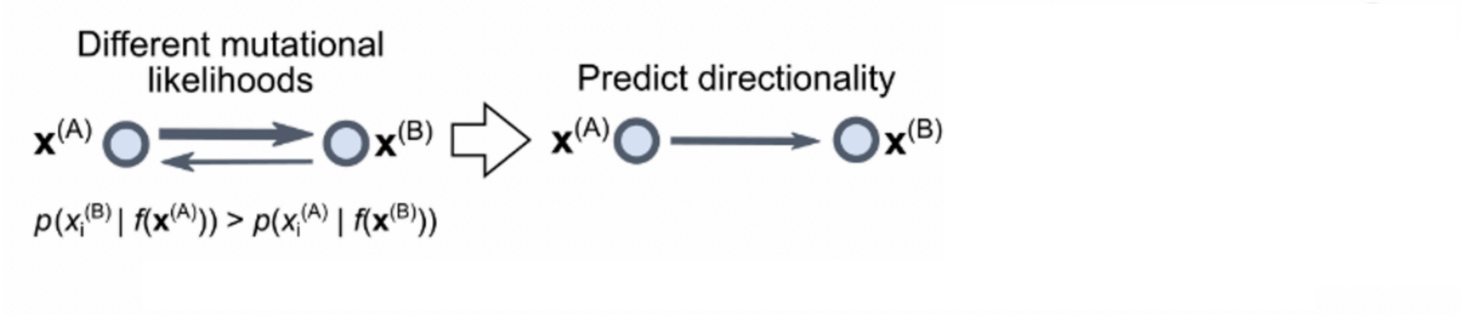
Hie *et al.* 2022

Pretrained models contain evolutionary information

Pseudolikelihood correlates with mutational tolerance



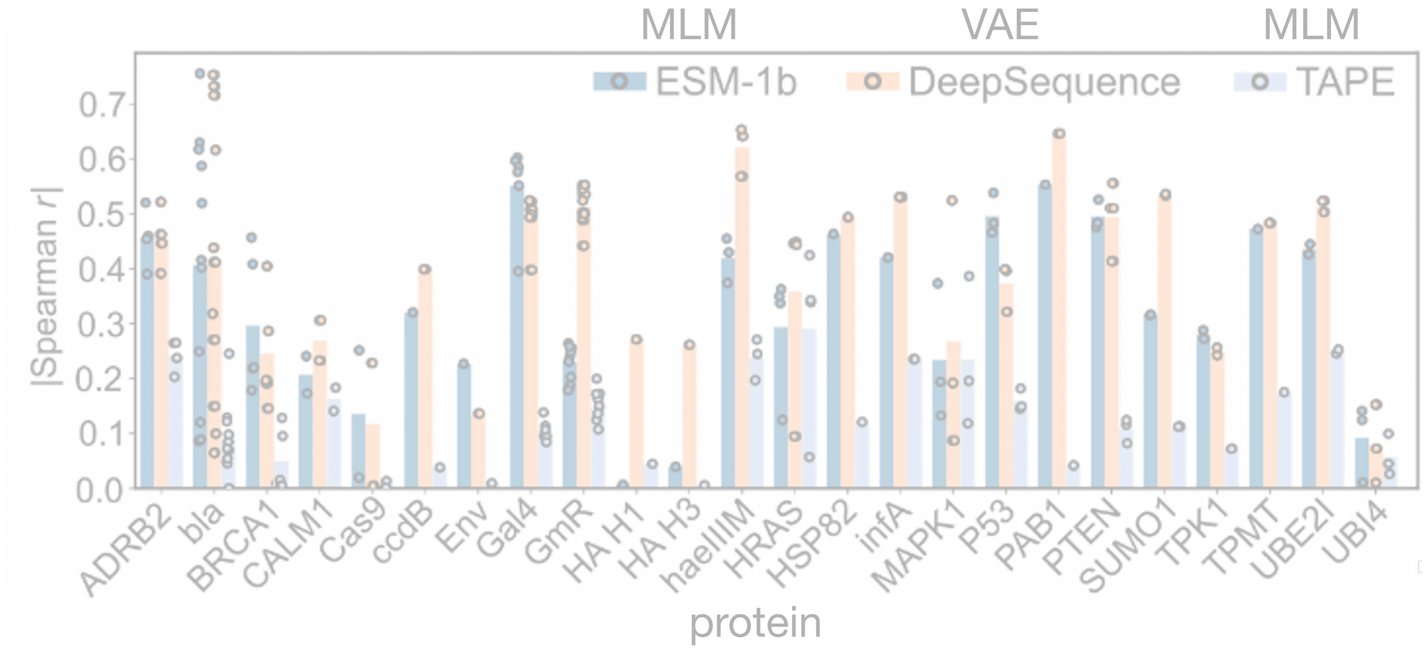
Use pseudolikelihood to predict evolutionary vector field



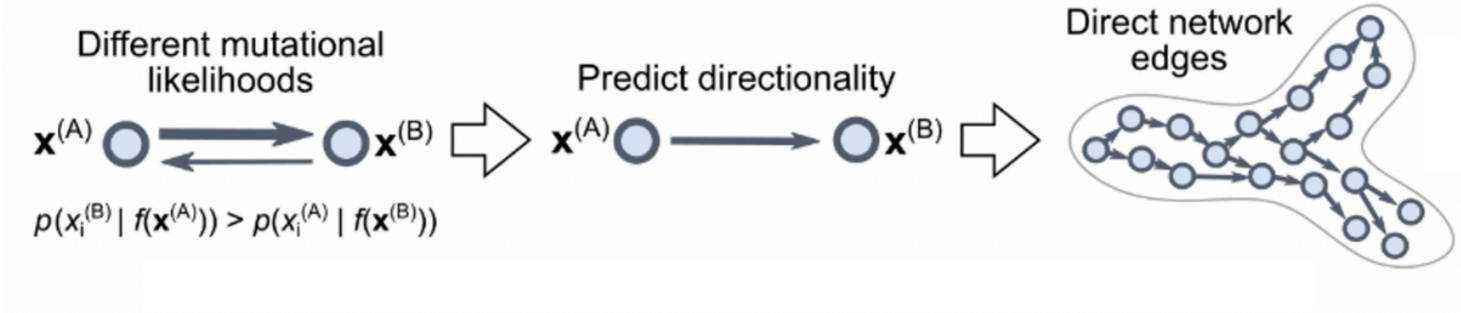
Hie *et al.* 2022

Pretrained models contain evolutionary information

Pseudolikelihood correlates with mutational tolerance



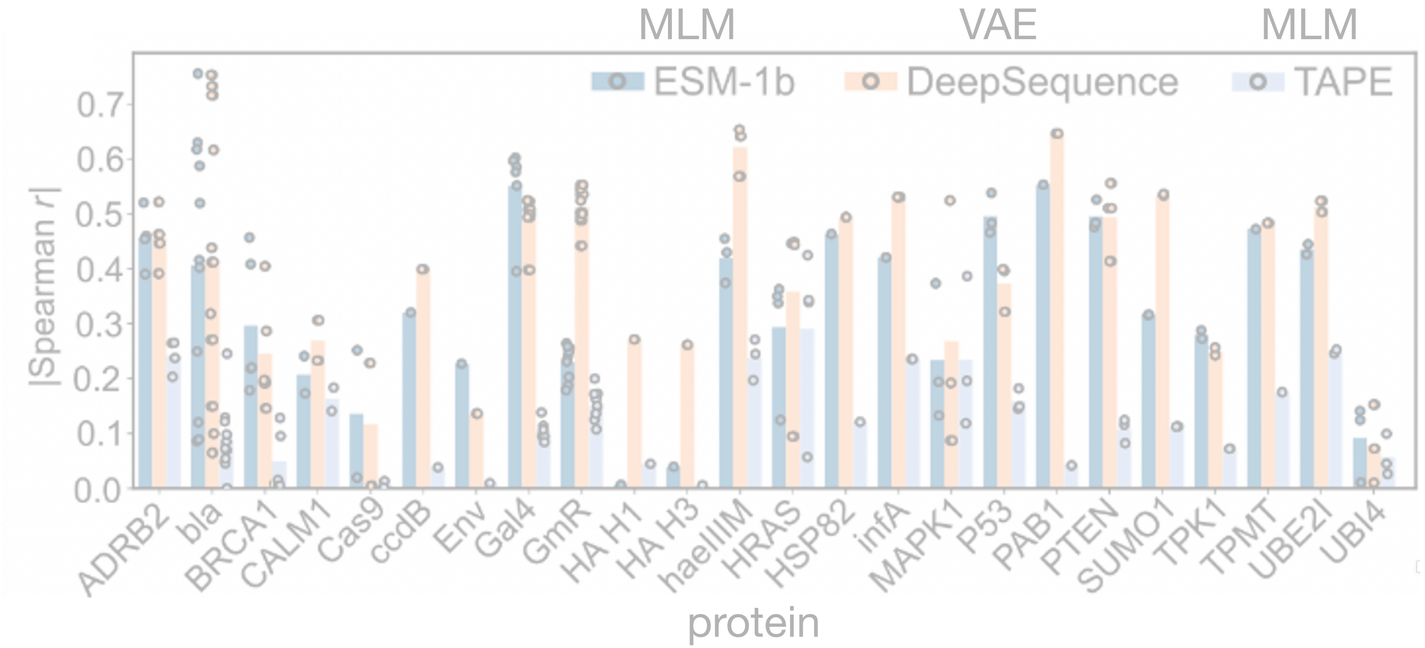
Use pseudolikelihood to predict evolutionary vector field



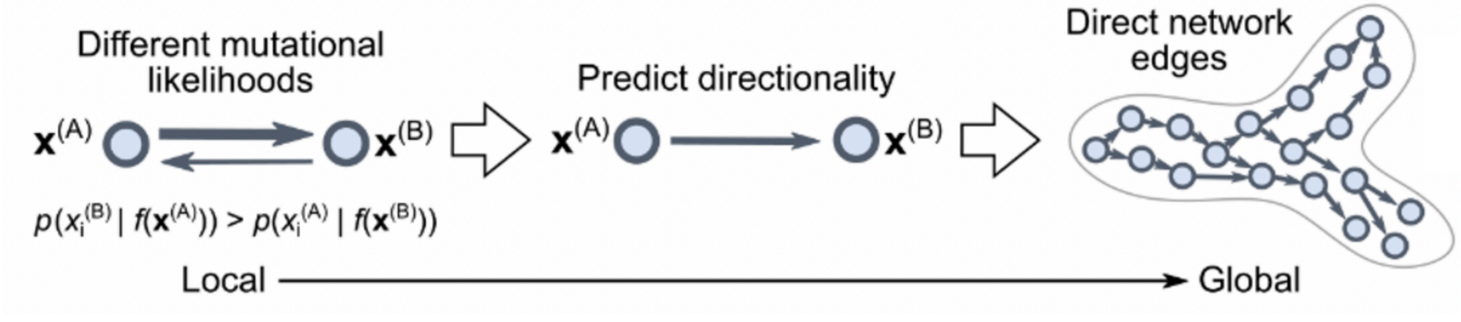
Hie *et al.* 2022

Pretrained models contain evolutionary information

Pseudolikelihood correlates with mutational tolerance



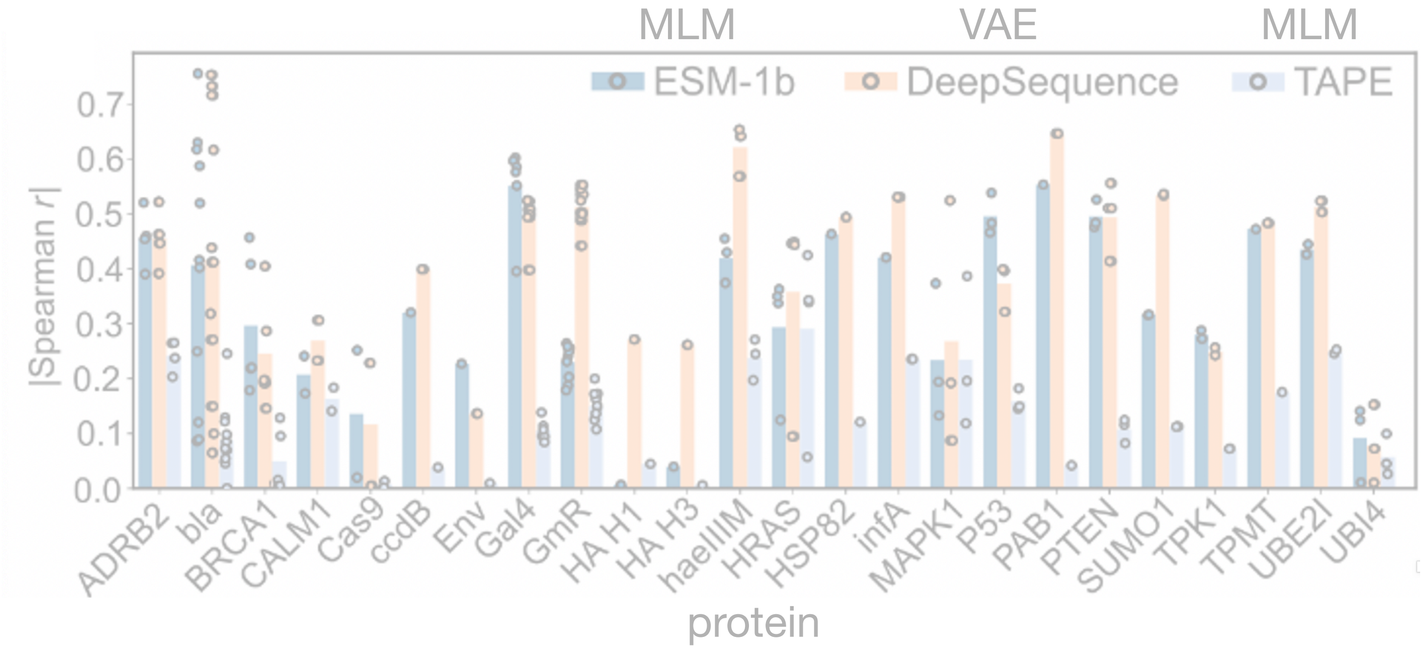
Use pseudolikelihood to predict evolutionary vector field



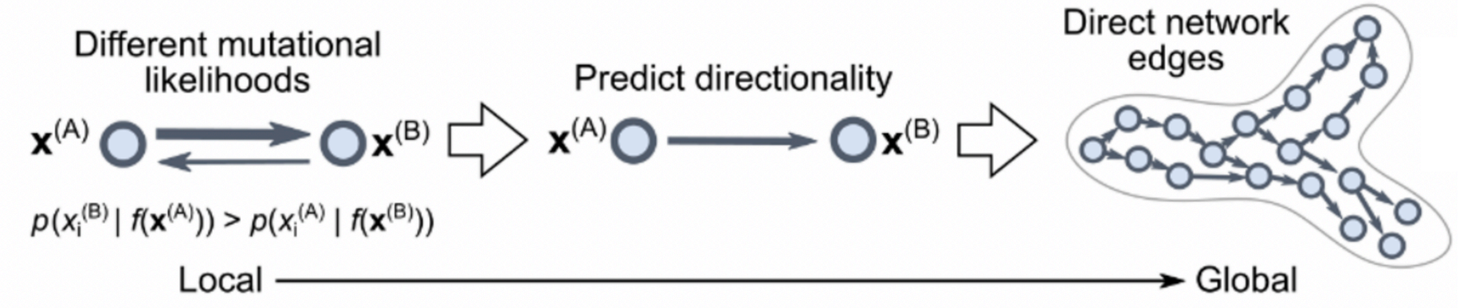
Hie *et al.* 2022

Pretrained models contain evolutionary information

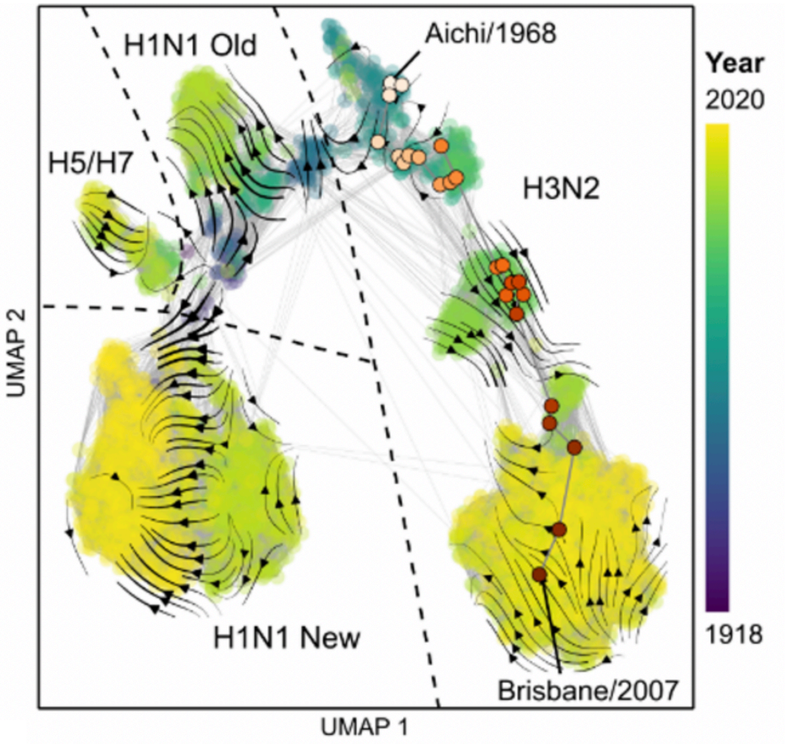
Pseudolikelihood correlates with mutational tolerance



Use pseudolikelihood to predict evolutionary vector field



Influenza A nucleoprotein

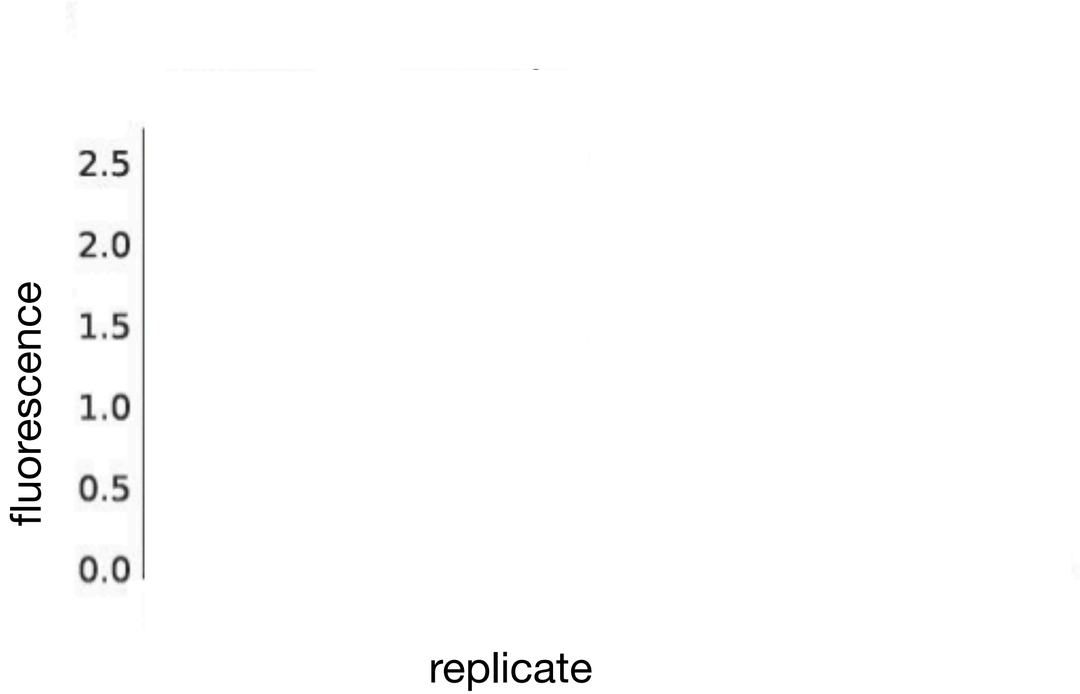


Hie *et al.* 2022

Pretraining guides search away from loss-of-function sequences

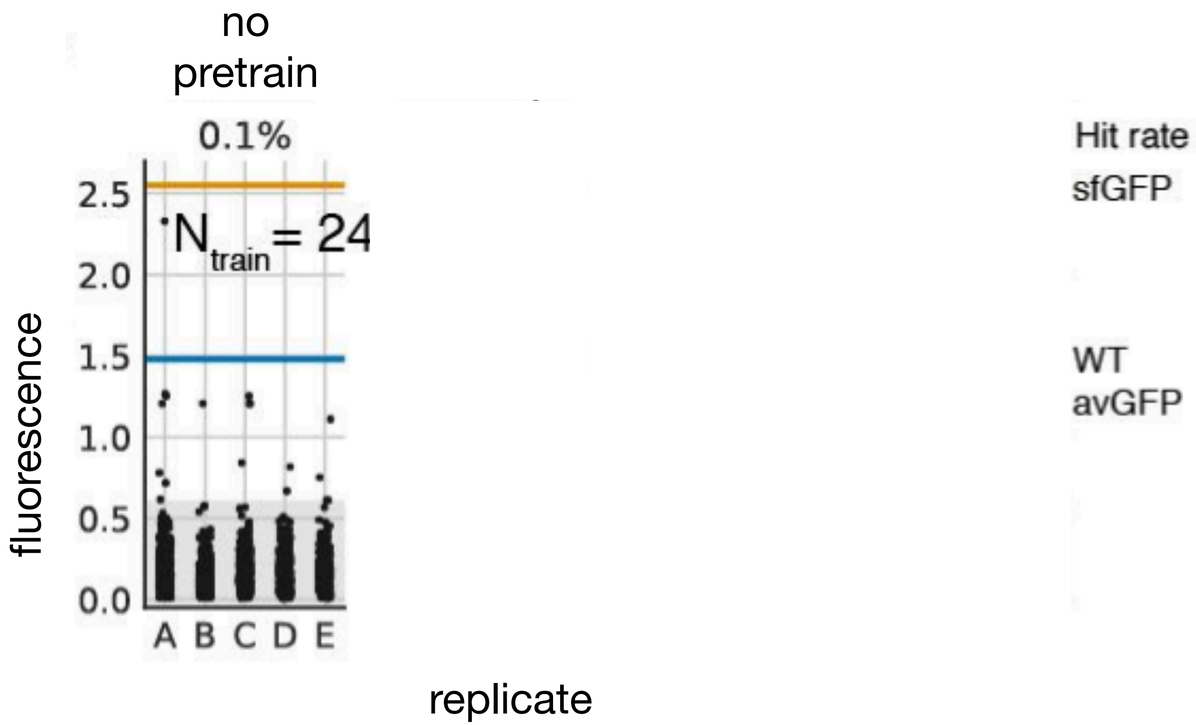
Biswas *et al. Nature Methods* 2021

Pretraining guides search away from loss-of-function sequences



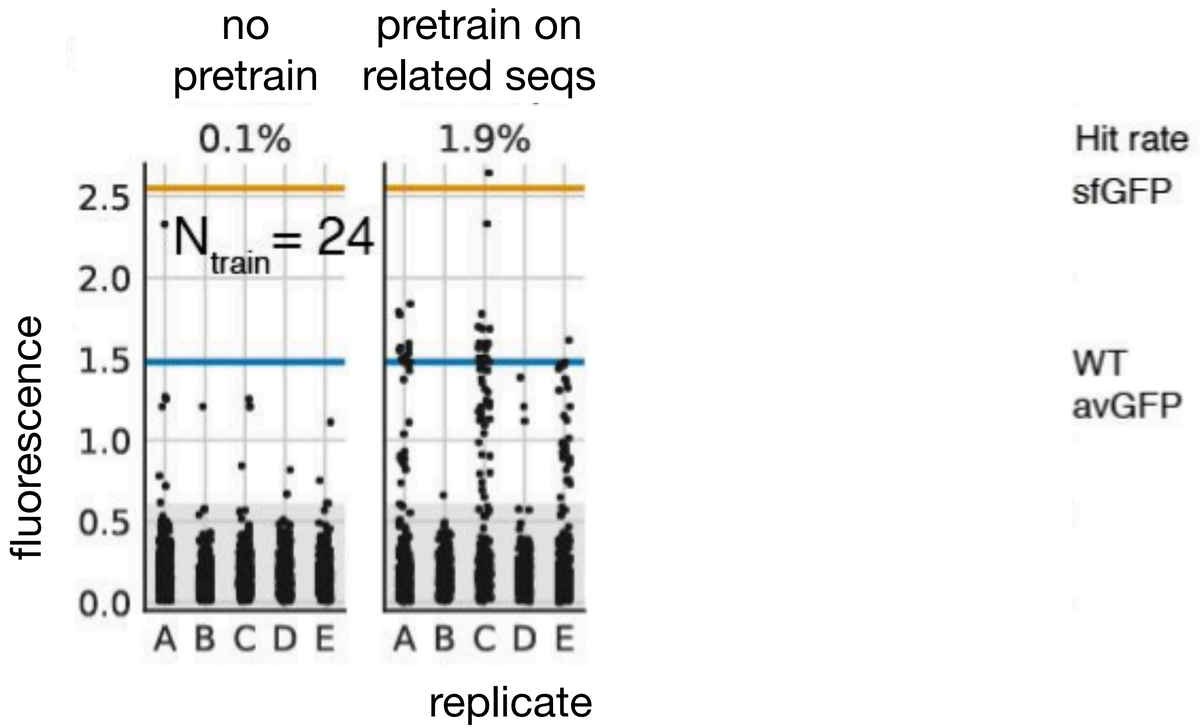
Biswas *et al. Nature Methods* 2021

Pretraining guides search away from loss-of-function sequences



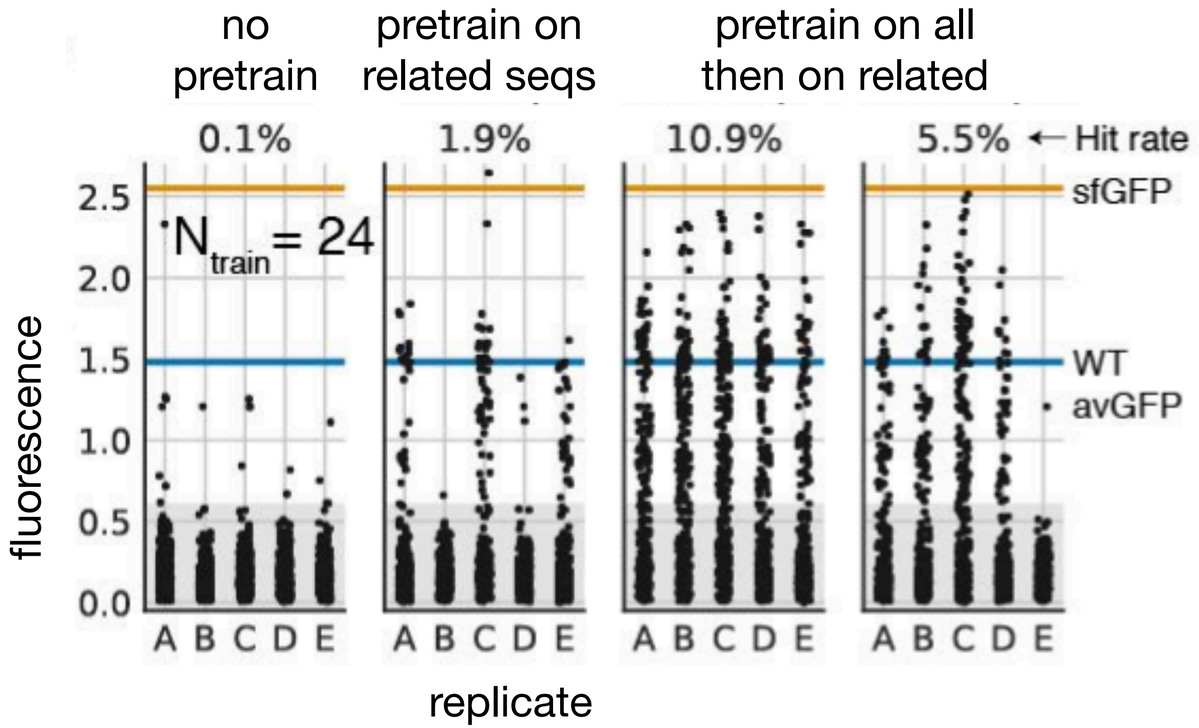
Biswas *et al. Nature Methods* 2021

Pretraining guides search away from loss-of-function sequences



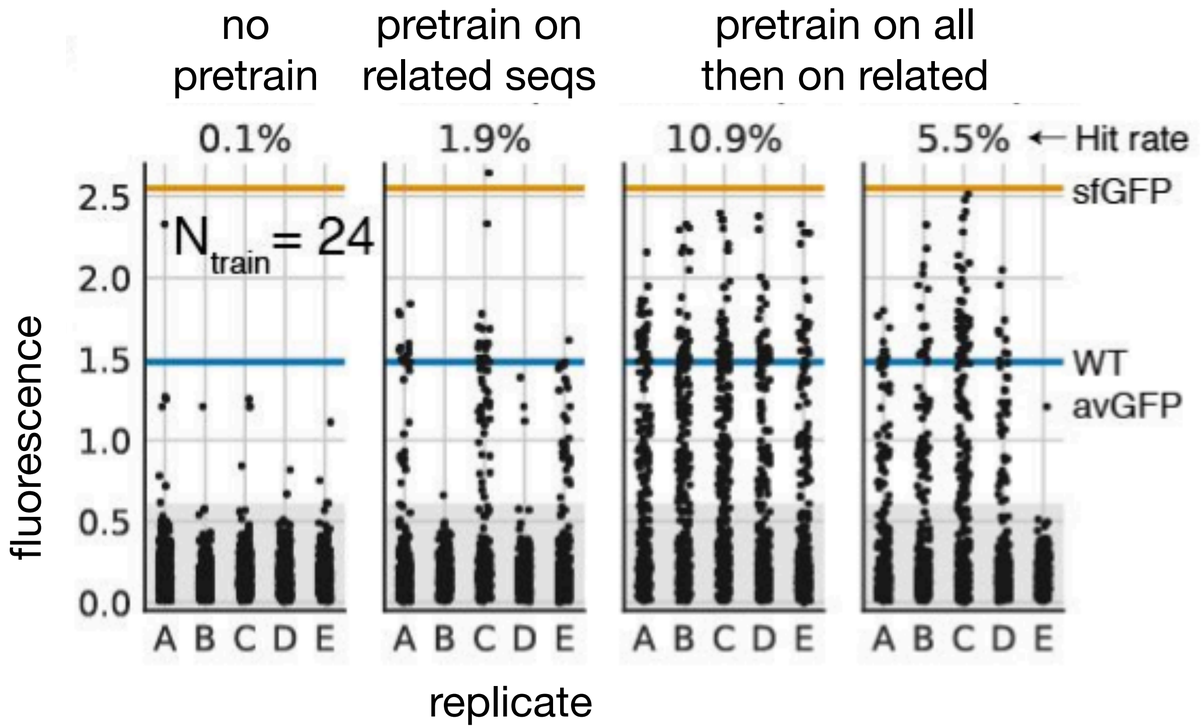
Biswas *et al. Nature Methods* 2021

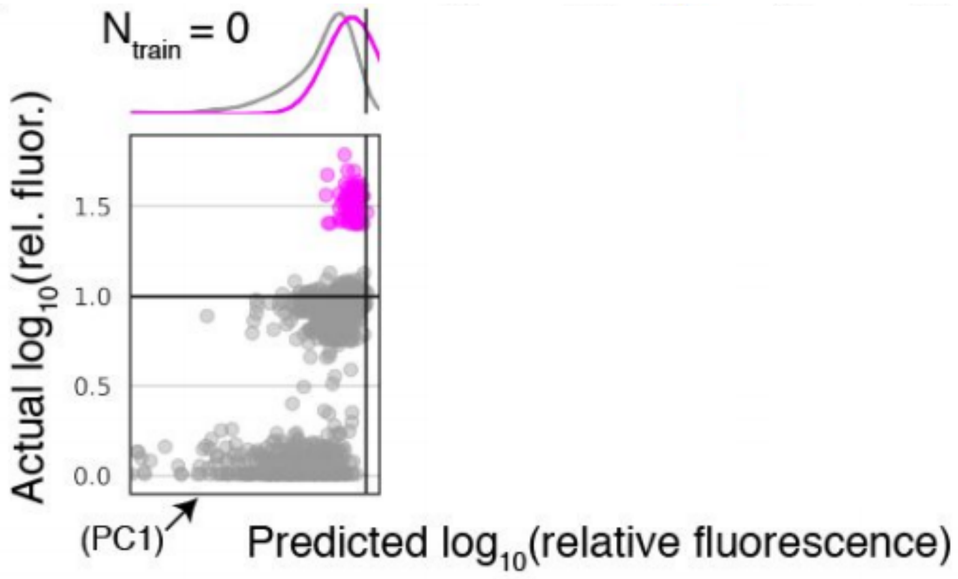
Pretraining guides search away from loss-of-function sequences



Biswas *et al. Nature Methods* 2021

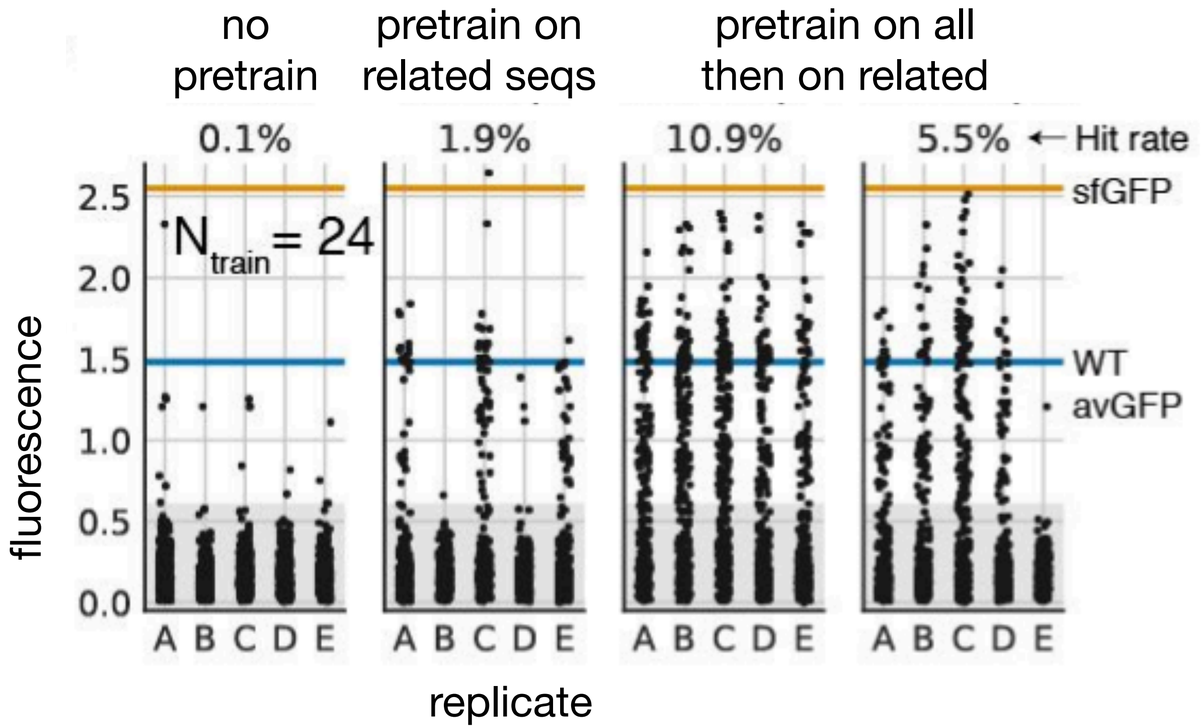
Pretraining guides search away from loss-of-function sequences

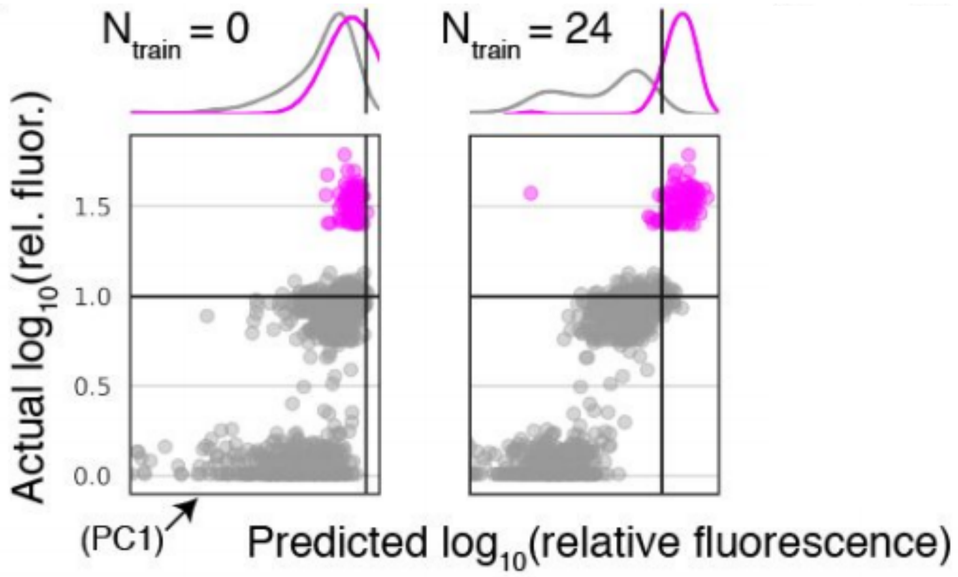




Biswas *et al. Nature Methods* 2021

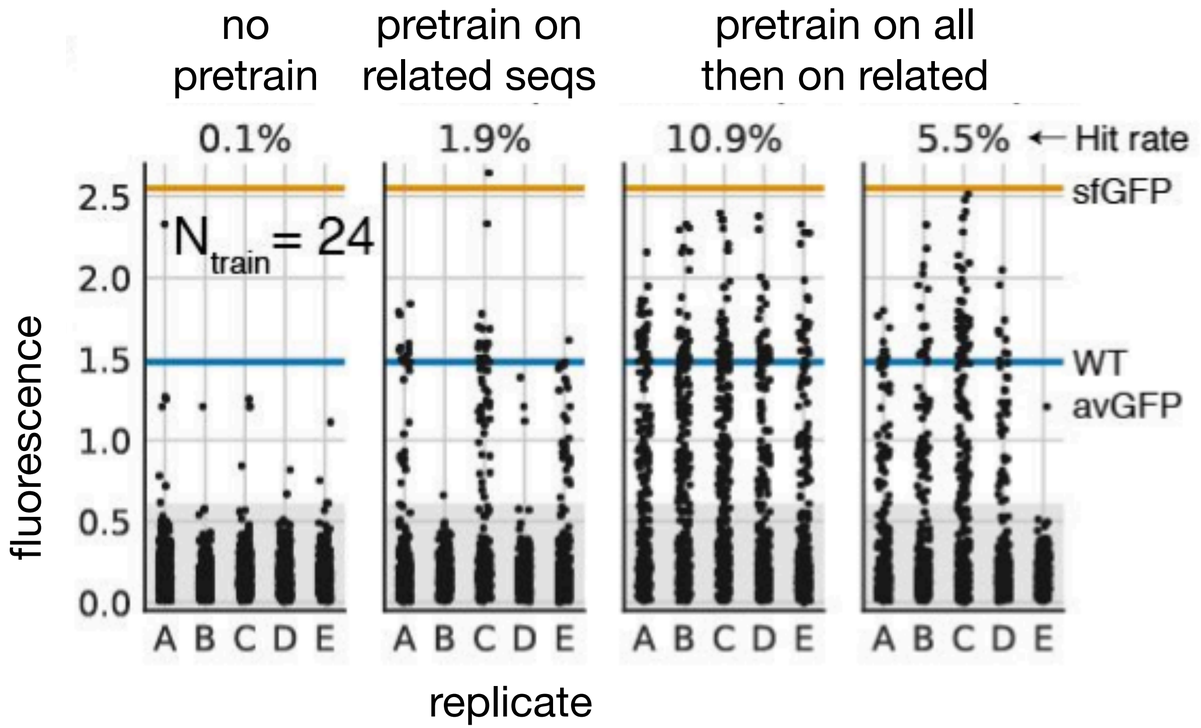
Pretraining guides search away from loss-of-function sequences

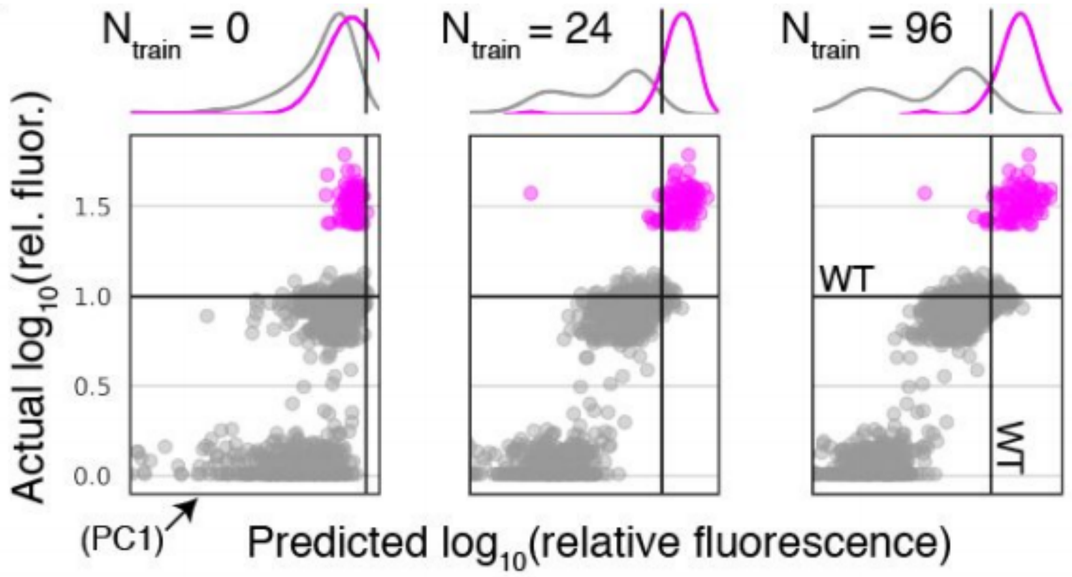




Biswas *et al. Nature Methods* 2021

Pretraining guides search away from loss-of-function sequences



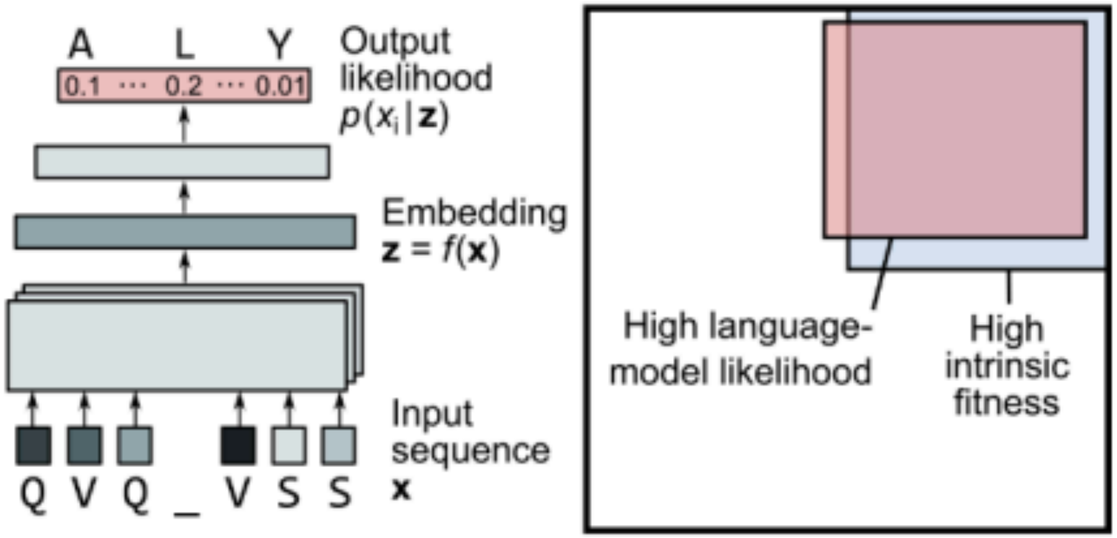


Biswas *et al. Nature Methods* 2021

Pretrained models propose high-fitness sequences

Hie *et al.* 2022

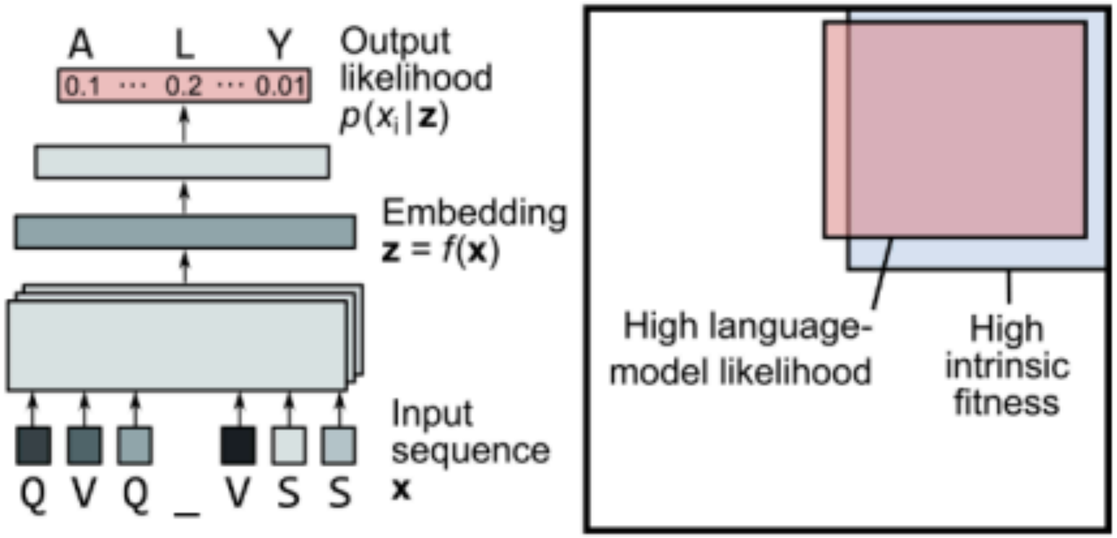
Pretrained models propose high-fitness sequences



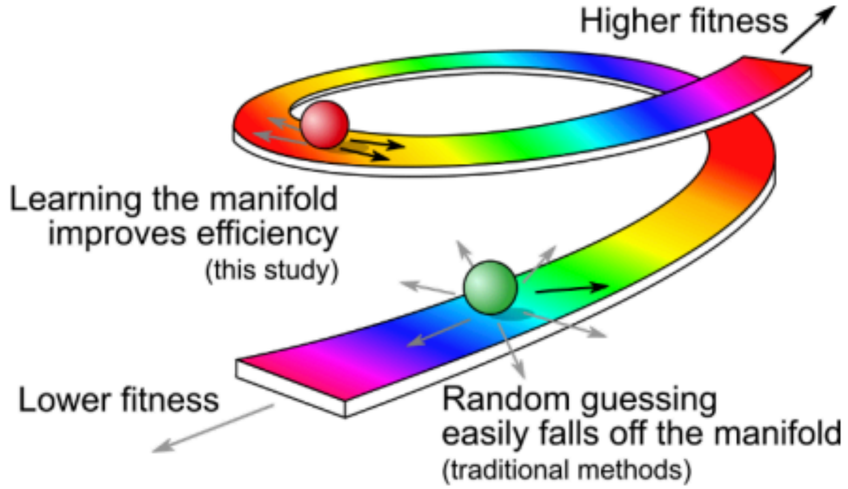
Experiment: Approximate intrinsic fitness with language-model likelihood, measure extrinsic fitness

Hie *et al.* 2022

Pretrained models propose high-fitness sequences

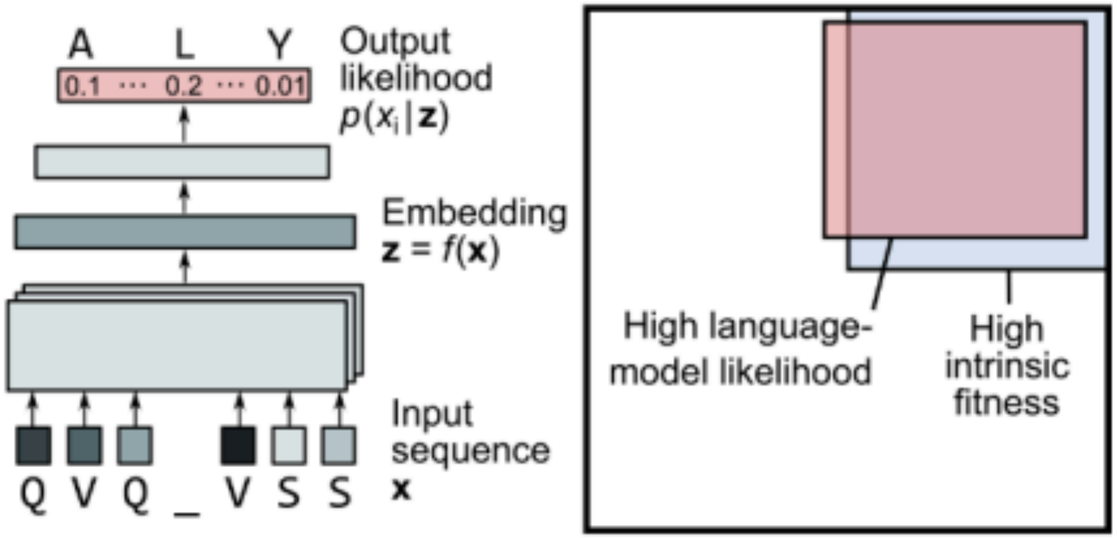


Experiment: Approximate intrinsic fitness with language-model likelihood, measure extrinsic fitness

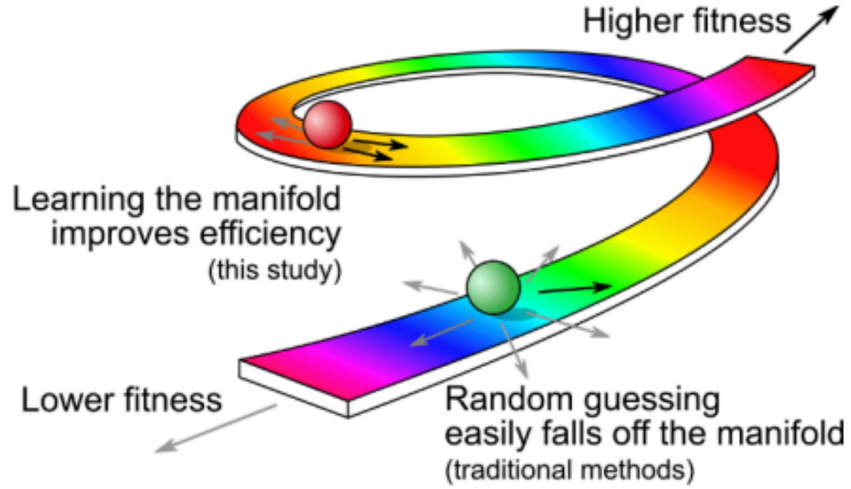


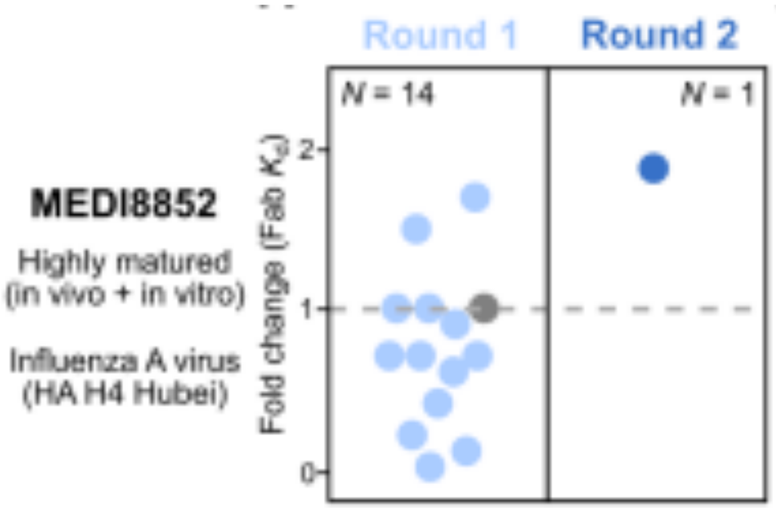
Hie *et al.* 2022

Pretrained models propose high-fitness sequences



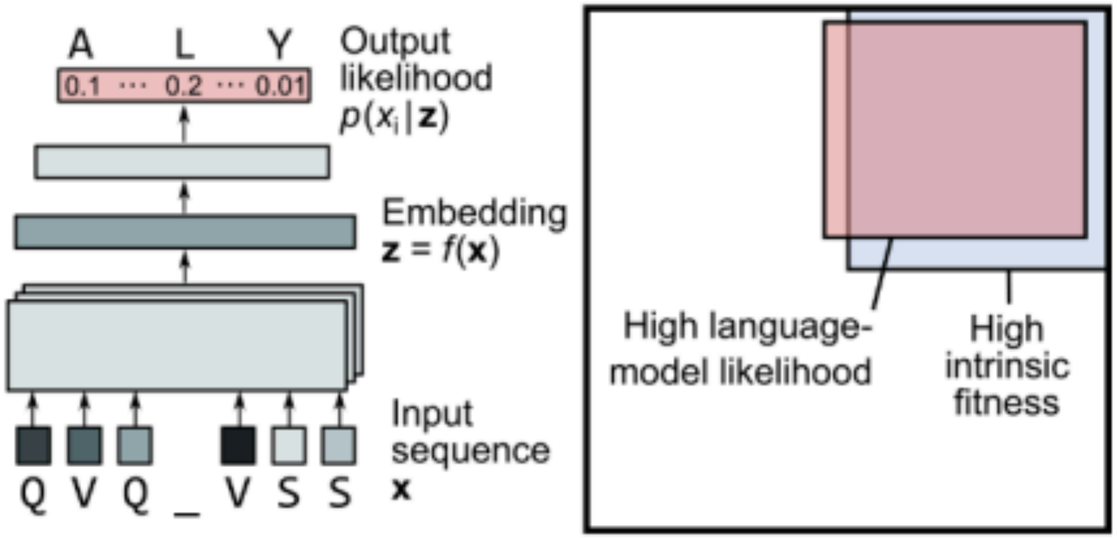
Experiment: Approximate intrinsic fitness with language-model likelihood, measure extrinsic fitness



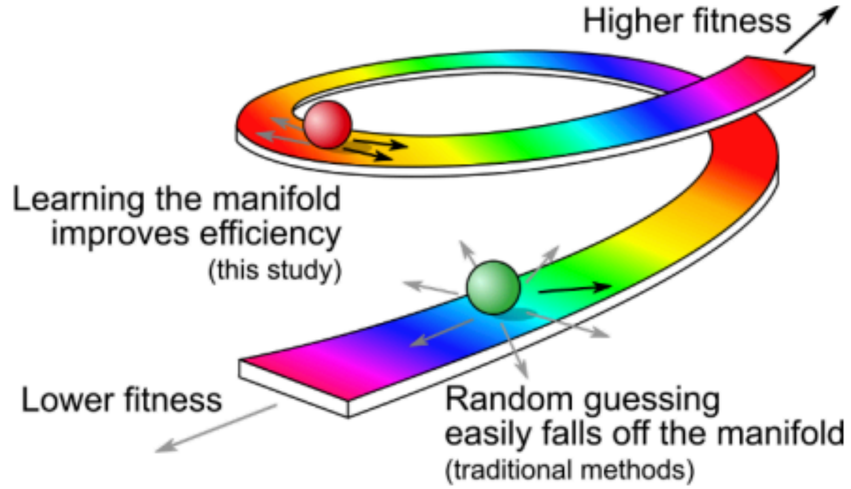


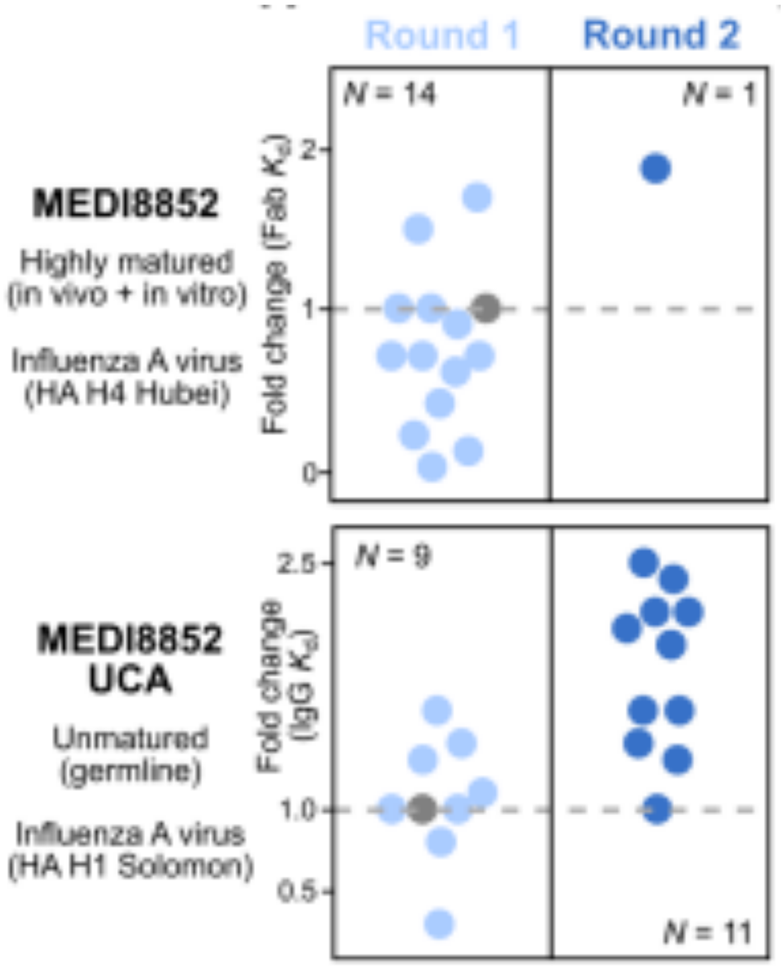
Hie *et al.* 2022

Pretrained models propose high-fitness sequences



Experiment: Approximate intrinsic fitness with language-model likelihood, measure extrinsic fitness





Hie *et al.* 2022

Need benchmarks for protein function ML

Need benchmarks for protein function ML

Need benchmarks for protein function ML

Need benchmarks for protein function ML

IMAGENET

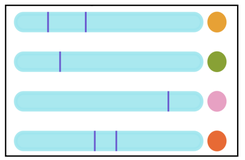
SuperGLUE

Need benchmarks for protein function ML

IMAGENET

SuperGLUE

Focus on sequence -> fitness

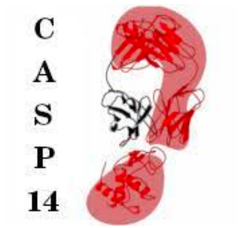


Need benchmarks for protein function ML

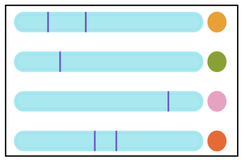
IMAGENET

SuperGLUE

CASP14



Focus on sequence -> fitness

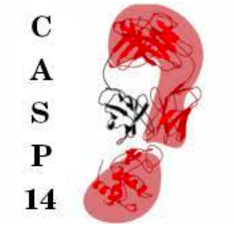


TAPE function datasets do not discriminate well

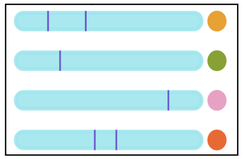
Need benchmarks for protein function ML

IMAGENET

SuperGLUE



Focus on sequence -> fitness



TAPE function datasets do not discriminate well

Model	Pretraining	Fluorescence	Stability

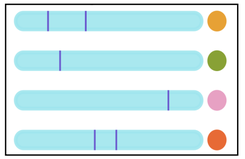
Need benchmarks for protein function ML

IMAGENET

SuperGLUE

CASP14

Focus on sequence -> fitness



TAPE function datasets do not discriminate well

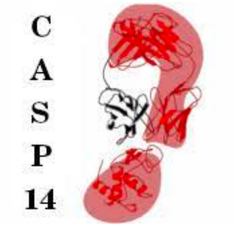
Model	Pretraining	Fluorescence	Stability
ESM (Rives *et al.*)	MLM	0.68	0.71

Need benchmarks for protein function ML

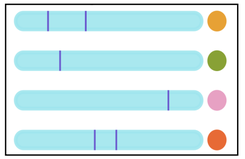
IMAGENET

SuperGLUE

CASP14



Focus on sequence -> fitness



TAPE function datasets do not discriminate well

Model	Pretraining	Fluorescence	Stability
ESM (Rives *et al.*)	MLM	0.68	0.71
TAPE transformer	MLM	0.68	0.73

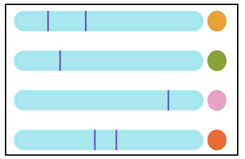
Need benchmarks for protein function ML

IMAGENET

SuperGLUE

CASP14

Focus on sequence -> fitness



TAPE function datasets do not discriminate well

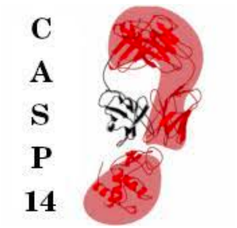
Model	Pretraining	Fluorescence	Stability
ESM (Rives *et al.*)	MLM	0.68	0.71
TAPE transformer	MLM	0.68	0.73
CPCProt (Lu *et al.*)	contrastive	0.68	0.65

Need benchmarks for protein function ML

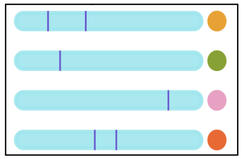
IMAGENET

SuperGLUE

CASP14



Focus on sequence -> fitness



TAPE function datasets do not discriminate well

Model	Pretraining	Fluorescence	Stability
ESM (Rives *et al.*)	MLM	0.68	0.71
TAPE transformer	MLM	0.68	0.73
CPCProt (Lu *et al.*)	contrastive	0.68	0.65
Linear regression	none	0.68	0.48

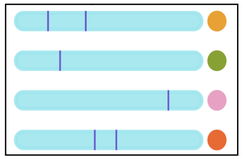
Need benchmarks for protein function ML

IMAGENET

SuperGLUE

CASP14

Focus on sequence -> fitness



TAPE function datasets do not discriminate well

Model	Pretraining	Fluorescence	Stability
ESM (Rives *et al.*)	MLM	0.68	0.71
TAPE transformer	MLM	0.68	0.73
CPCProt (Lu *et al.*)	contrastive	0.68	0.65
Linear regression	none	0.68	0.48
CNN	none	0.67	0.51

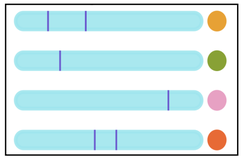
Need benchmarks for protein function ML

IMAGENET

SuperGLUE

CASP14

Focus on sequence -> fitness



TAPE function datasets do not discriminate well

Model	Pretraining	Fluorescence	Stability
ESM (Rives *et al.*)	MLM	0.68	0.71
TAPE transformer	MLM	0.68	0.73
CPCProt (Lu *et al.*)	contrastive	0.68	0.65
Linear regression	none	0.68	0.48
CNN	none	0.67	0.51
CARP-640M	MLM	0.68	0.72

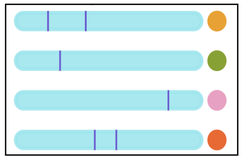
Need benchmarks for protein function ML

IMAGENET

SuperGLUE

CASP14

Focus on sequence -> fitness



TAPE function datasets do not discriminate well

Model	Pretraining	Fluorescence	Stability
ESM (Rives *et al.*)	MLM	0.68	0.71
TAPE transformer	MLM	0.68	0.73
CPCProt (Lu *et al.*)	contrastive	0.68	0.65
Linear regression	none	0.68	0.48
CNN	none	0.67	0.51
CARP-640M	MLM	0.68	0.72

Need benchmarks for protein function ML

IMAGENET

SuperGLUE

CASP14


Christian Dallago


Jody Mou


Kadina Johnston

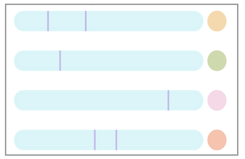

Bruce Wittmann


Nick Bhattacharya


Samuel Goldman


Ali Madani

Focus on sequence -> fitness



TAPE function datasets do not discriminate well

Model	Pretraining	Fluorescence	Stability
ESM (Rives *et al.*)	MLM	0.68	0.71
TAPE transformer	MLM	0.68	0.73
CPCProt (Lu *et al.*)	contrastive	0.68	0.65
Linear regression	none	0.68	0.48
CNN	none	0.67	0.51
CARP-640M	MLM	0.68	0.72

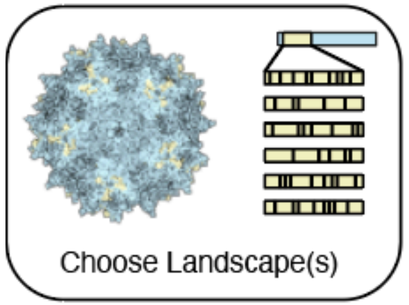
Need benchmarks for protein function ML

IMAGENET

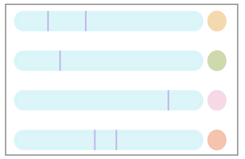
SuperGLUE

CASP14

Christian Dallago Jody Mou Kadina Johnston Bruce Wittmann Nick Bhattacharya Samuel Goldman Ali Madani



Focus on sequence -> fitness



TAPE function datasets do not discriminate well

Model	Pretraining	Fluorescence	Stability
ESM (Rives *et al.*)	MLM	0.68	0.71
TAPE transformer	MLM	0.68	0.73
CPCProt (Lu *et al.*)	contrastive	0.68	0.65
Linear regression	none	0.68	0.48
CNN	none	0.67	0.51
CARP-640M	MLM	0.68	0.72

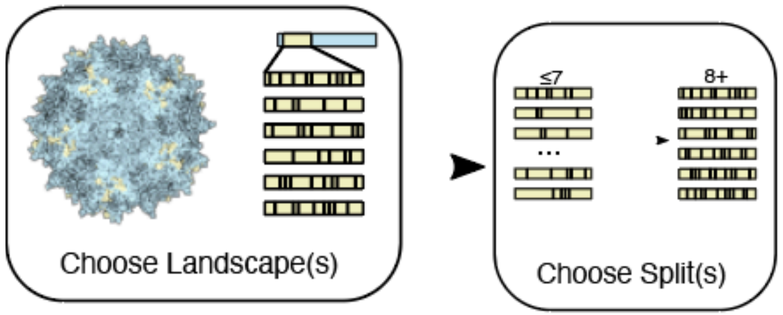
Need benchmarks for protein function ML

IMAGENET

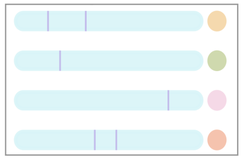
SuperGLUE

CASP14

Christian Dallago Jody Mou Kadina Johnston Bruce Wittmann Nick Bhattacharya Samuel Goldman Ali Madani



Focus on sequence -> fitness



TAPE function datasets do not discriminate well

Model	Pretraining	Fluorescence	Stability
ESM (Rives *et al.*)	MLM	0.68	0.71
TAPE transformer	MLM	0.68	0.73
CPCProt (Lu *et al.*)	contrastive	0.68	0.65
Linear regression	none	0.68	0.48
CNN	none	0.67	0.51
CARP-640M	MLM	0.68	0.72

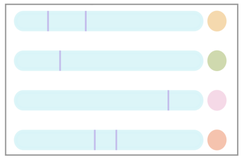
Need benchmarks for protein function ML

IMAGENET

SuperGLUE

CASP14

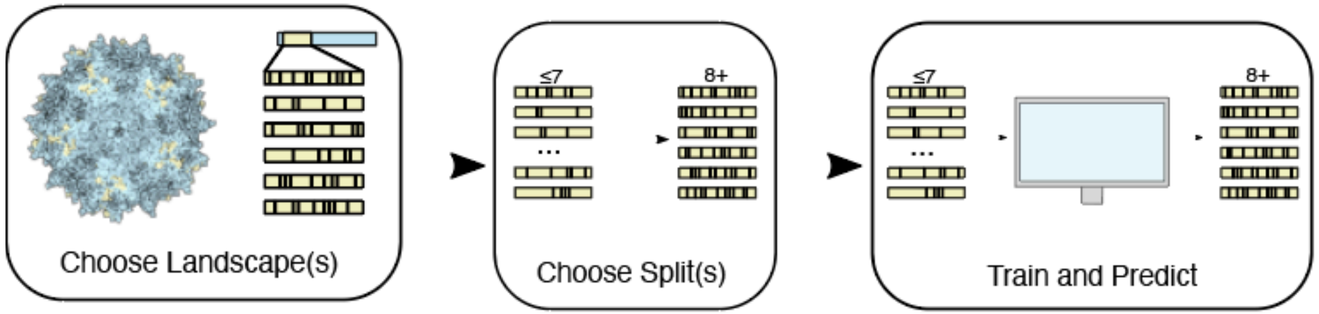
Focus on sequence -> fitness



TAPE function datasets do not discriminate well

Model	Pretraining	Fluorescence	Stability
ESM (Rives *et al.*)	MLM	0.68	0.71
TAPE transformer	MLM	0.68	0.73
CPCProt (Lu *et al.*)	contrastive	0.68	0.65
Linear regression	none	0.68	0.48
CNN	none	0.67	0.51
CARP-640M	MLM	0.68	0.72

Christian Dallago Jody Mou Kadina Johnston Bruce Wittmann Nick Bhattacharya Samuel Goldman Ali Madani



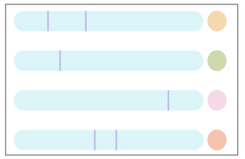
Need benchmarks for protein function ML

IMAGENET

SuperGLUE

CASP14

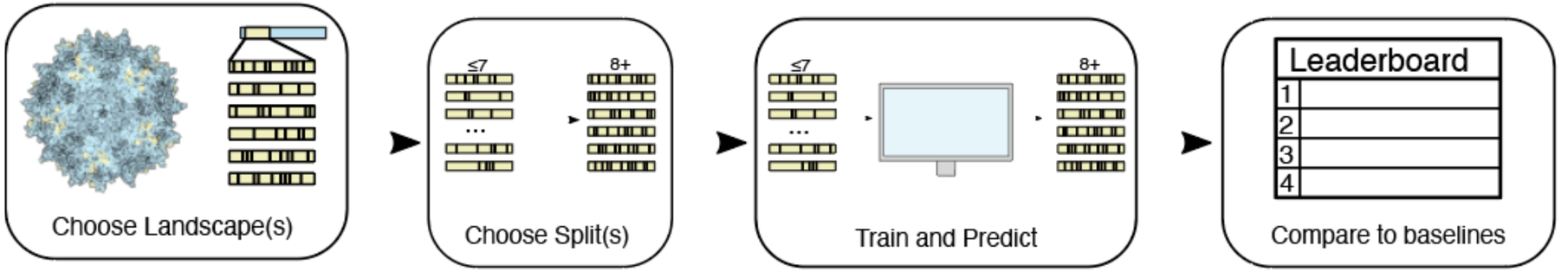
Focus on sequence -> fitness



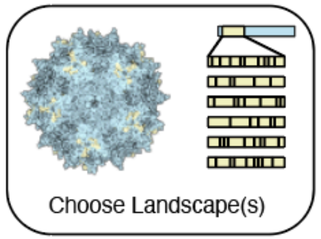
TAPE function datasets do not discriminate well

Model	Pretraining	Fluorescence	Stability
ESM (Rives *et al.*)	MLM	0.68	0.71
TAPE transformer	MLM	0.68	0.73
CPCProt (Lu *et al.*)	contrastive	0.68	0.65
Linear regression	none	0.68	0.48
CNN	none	0.67	0.51
CARP-640M	MLM	0.68	0.72

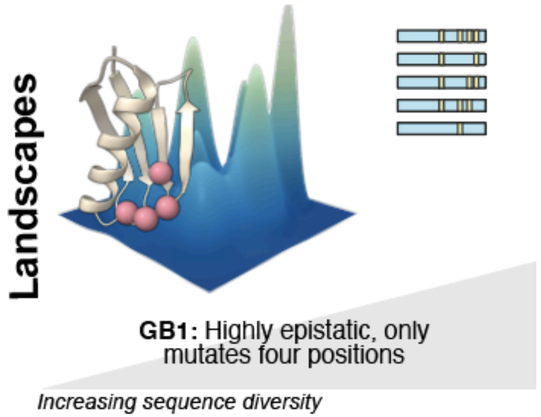
Christian Dallago, Jody Mou, Kadina Johnston, Bruce Wittmann, Nick Bhattacharya, Samuel Goldman, Ali Madani

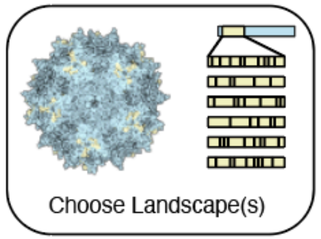


Choose three landscapes

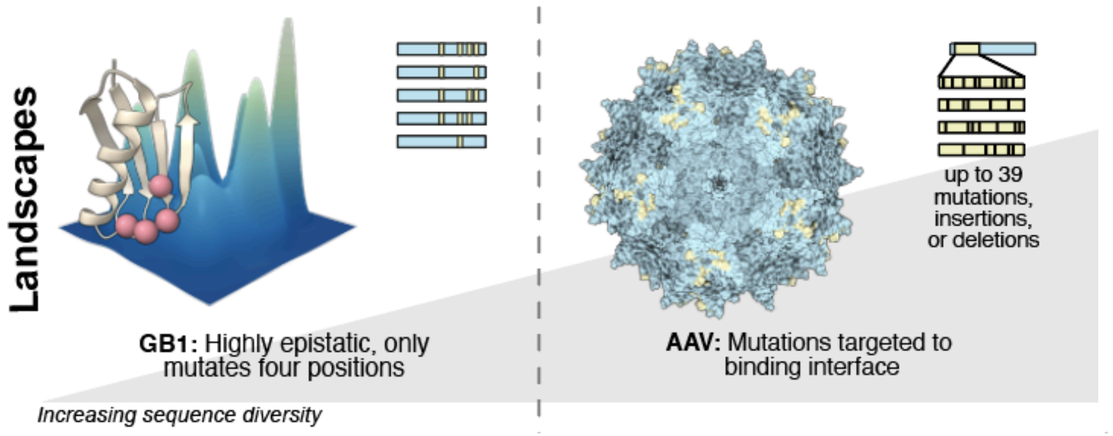


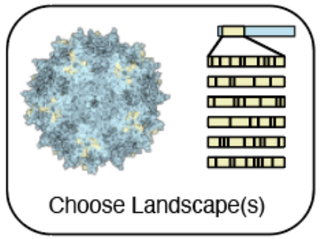
Choose three landscapes



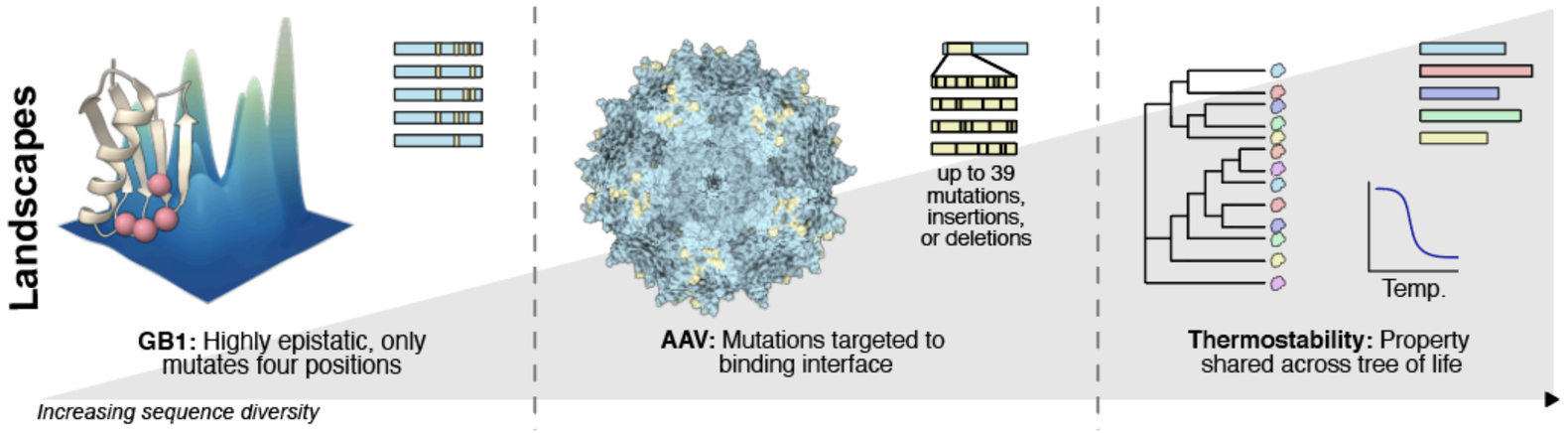


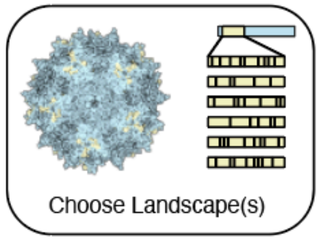
Choose three landscapes



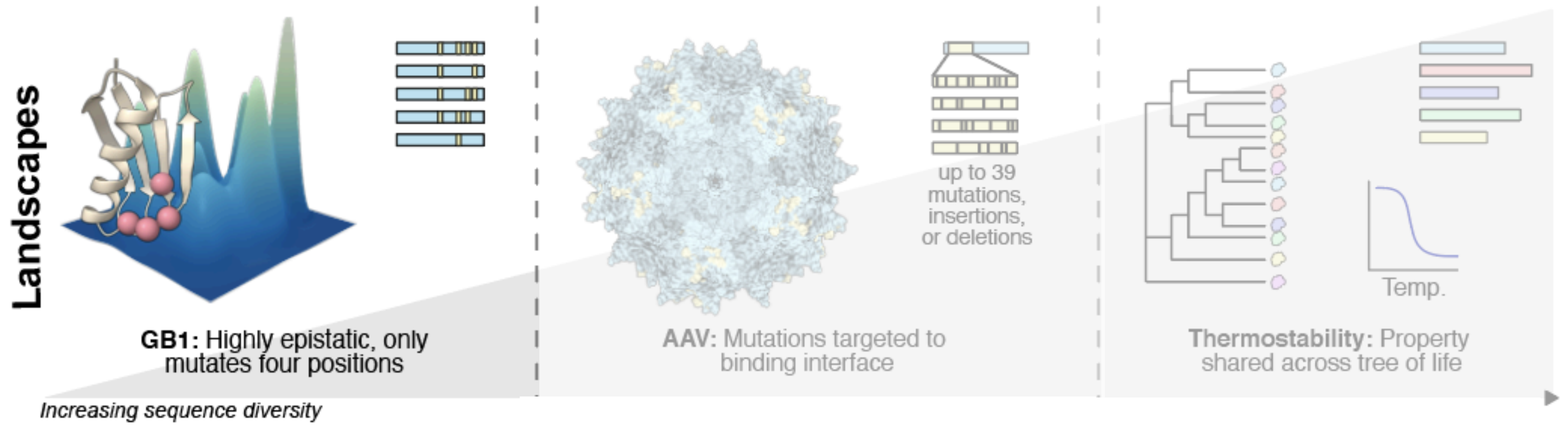


Choose three landscapes

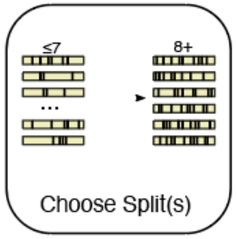




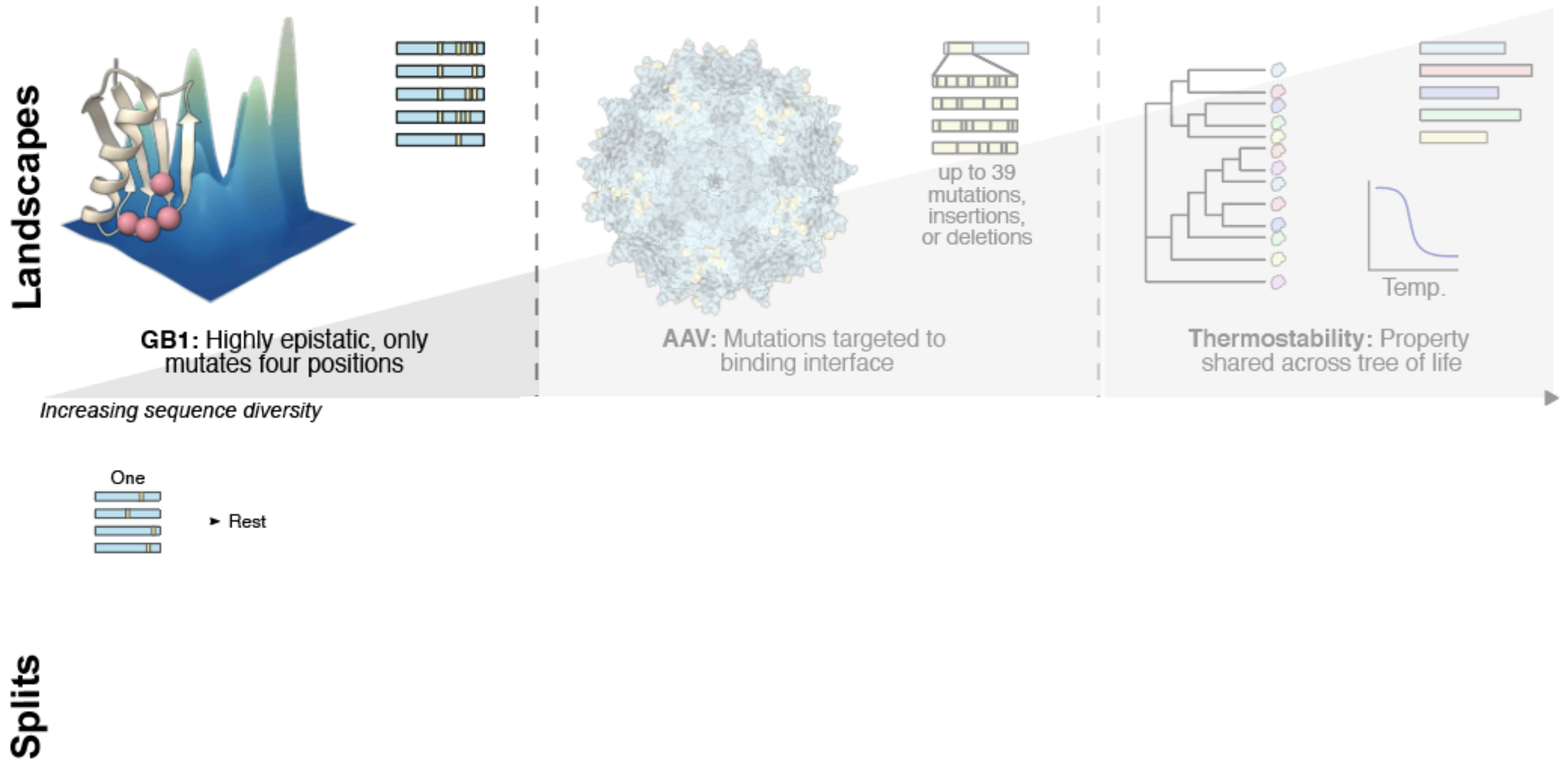
Make biologically-relevant train/test splits

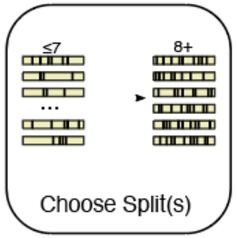


Splits

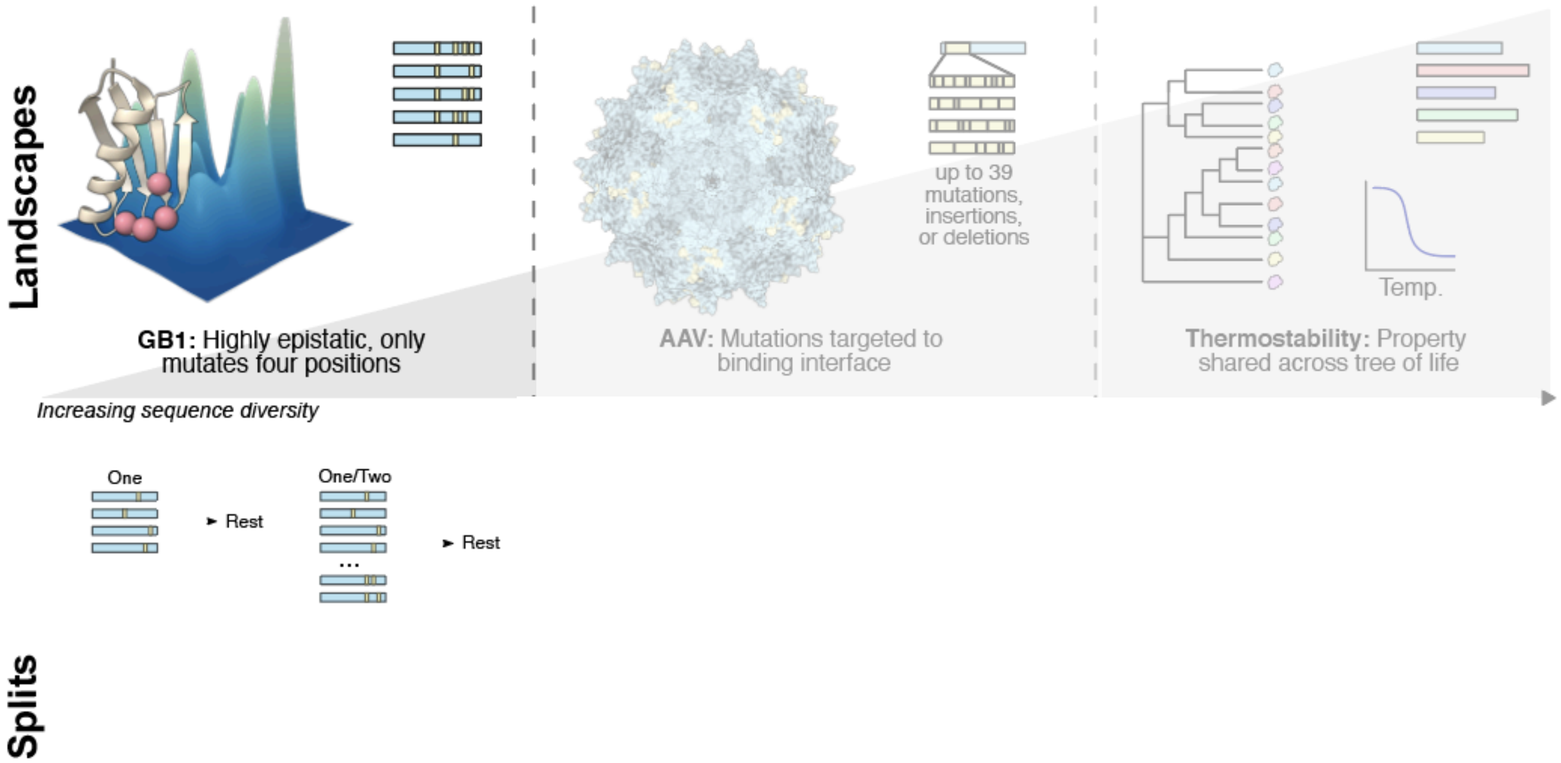


Make biologically-relevant train/test splits

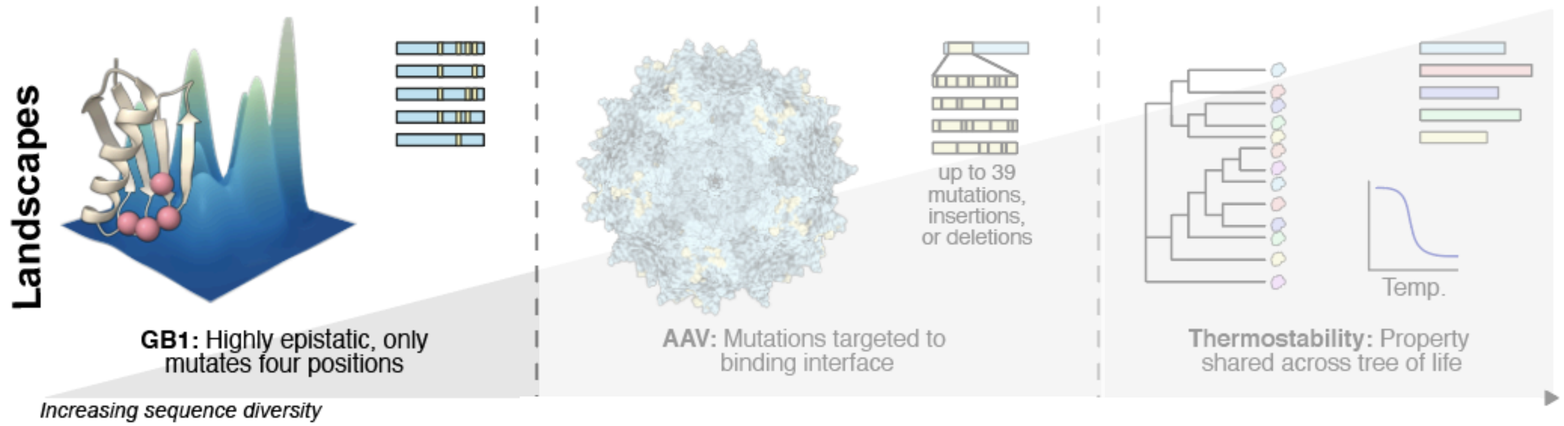


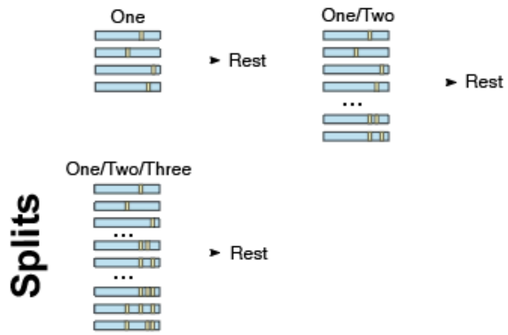


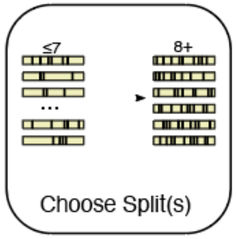
Make biologically-relevant train/test splits



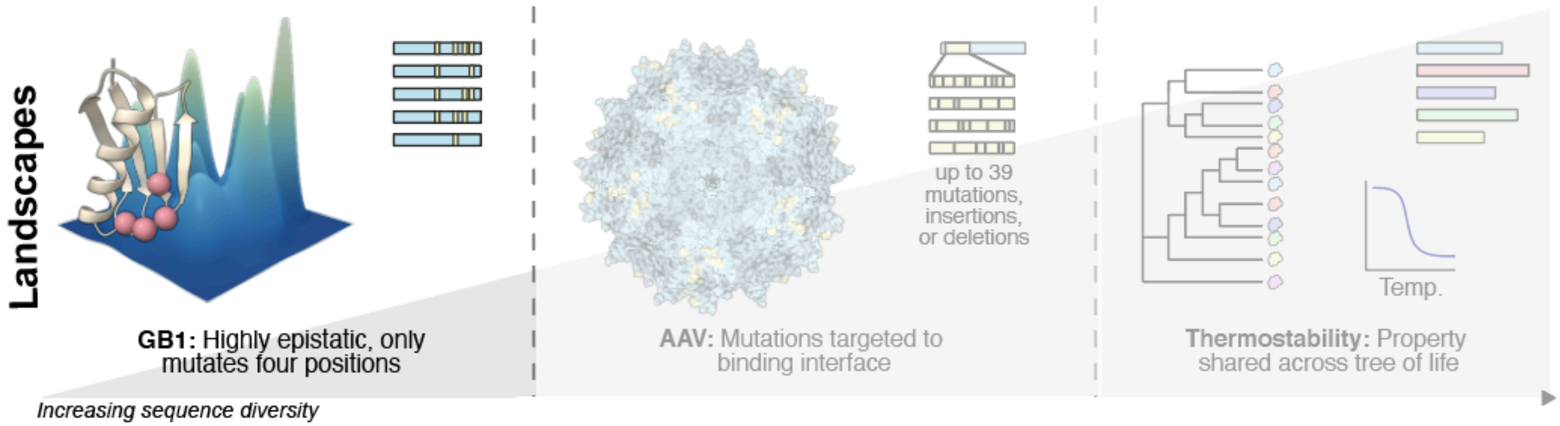
Make biologically-relevant train/test splits

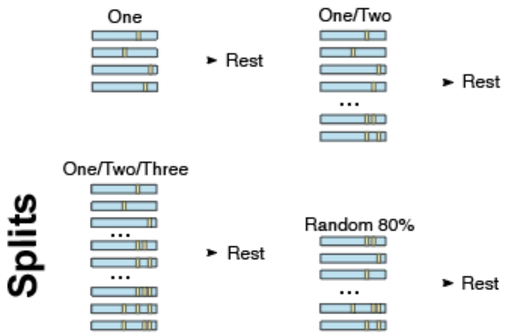


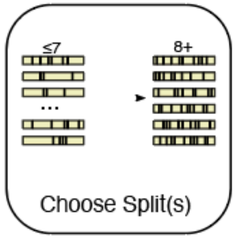




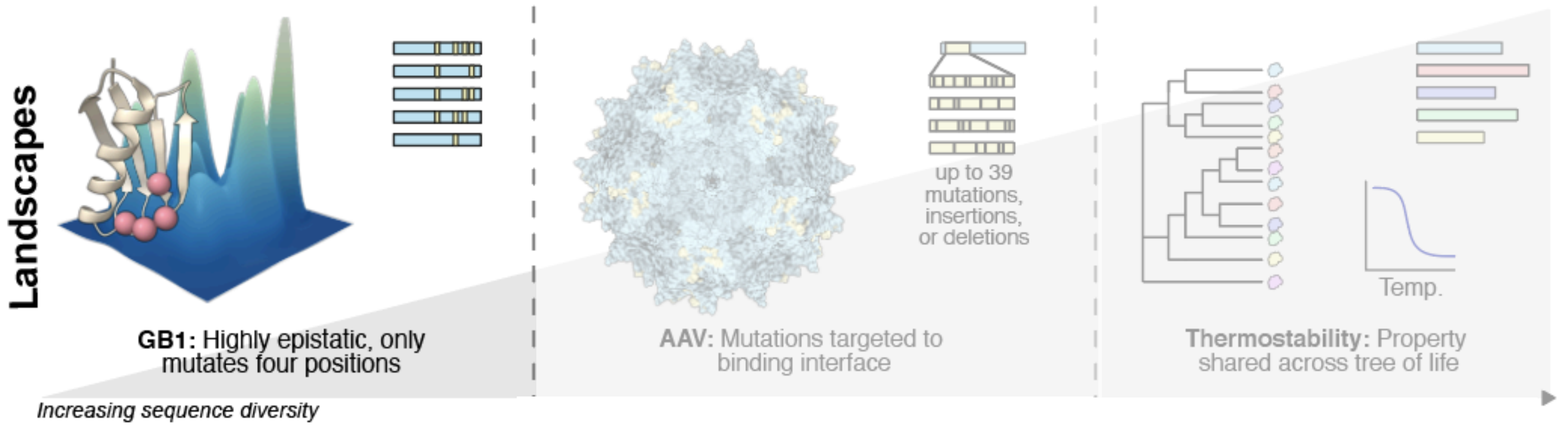
Make biologically-relevant train/test splits

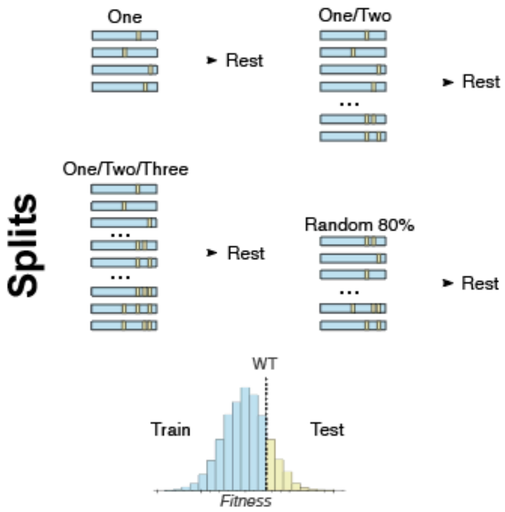


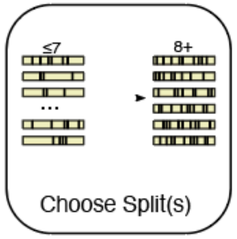




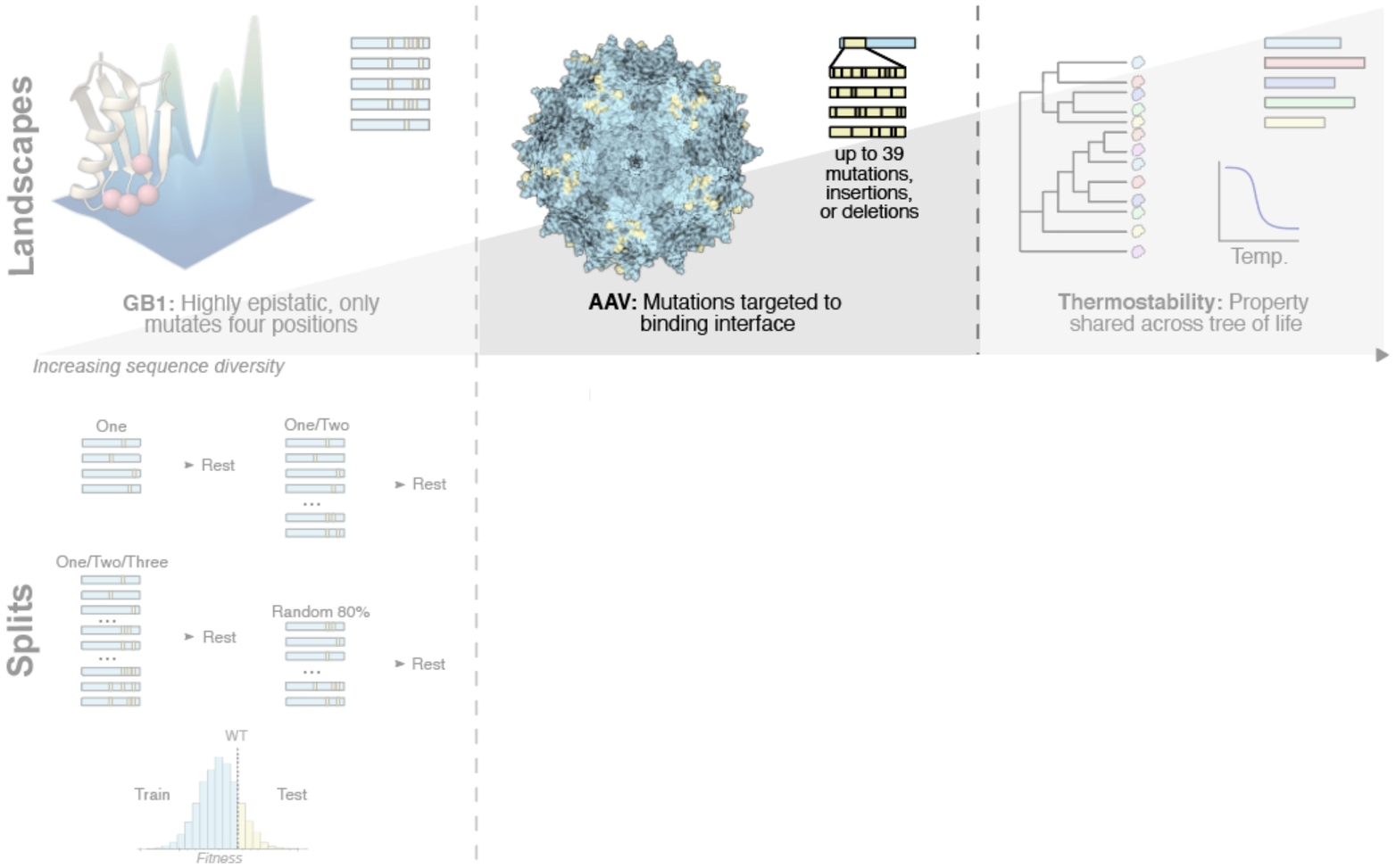
Make biologically-relevant train/test splits

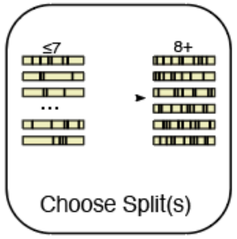




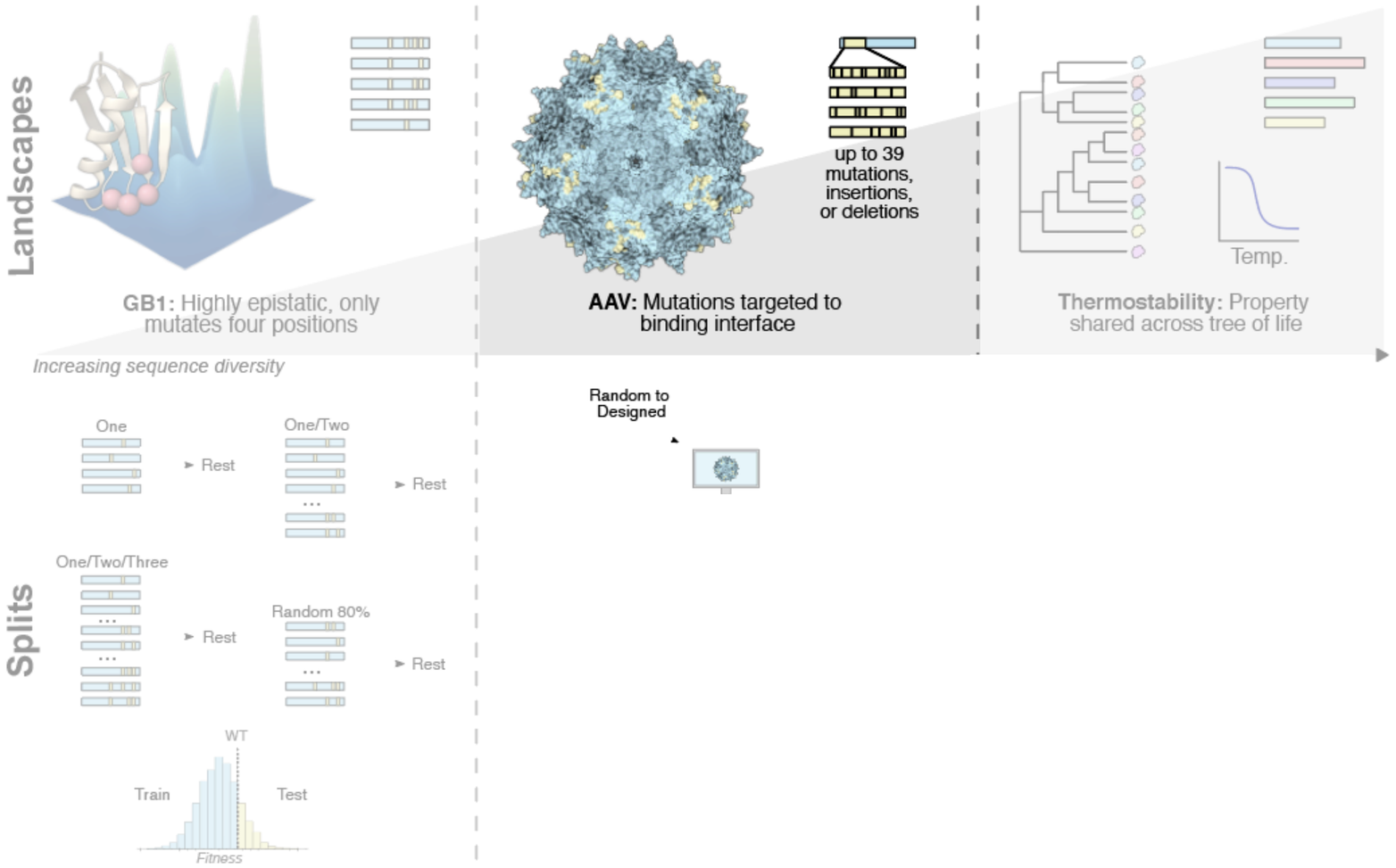


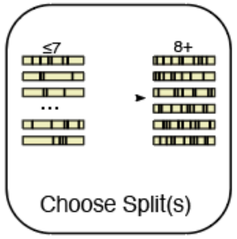
Make biologically-relevant train/test splits



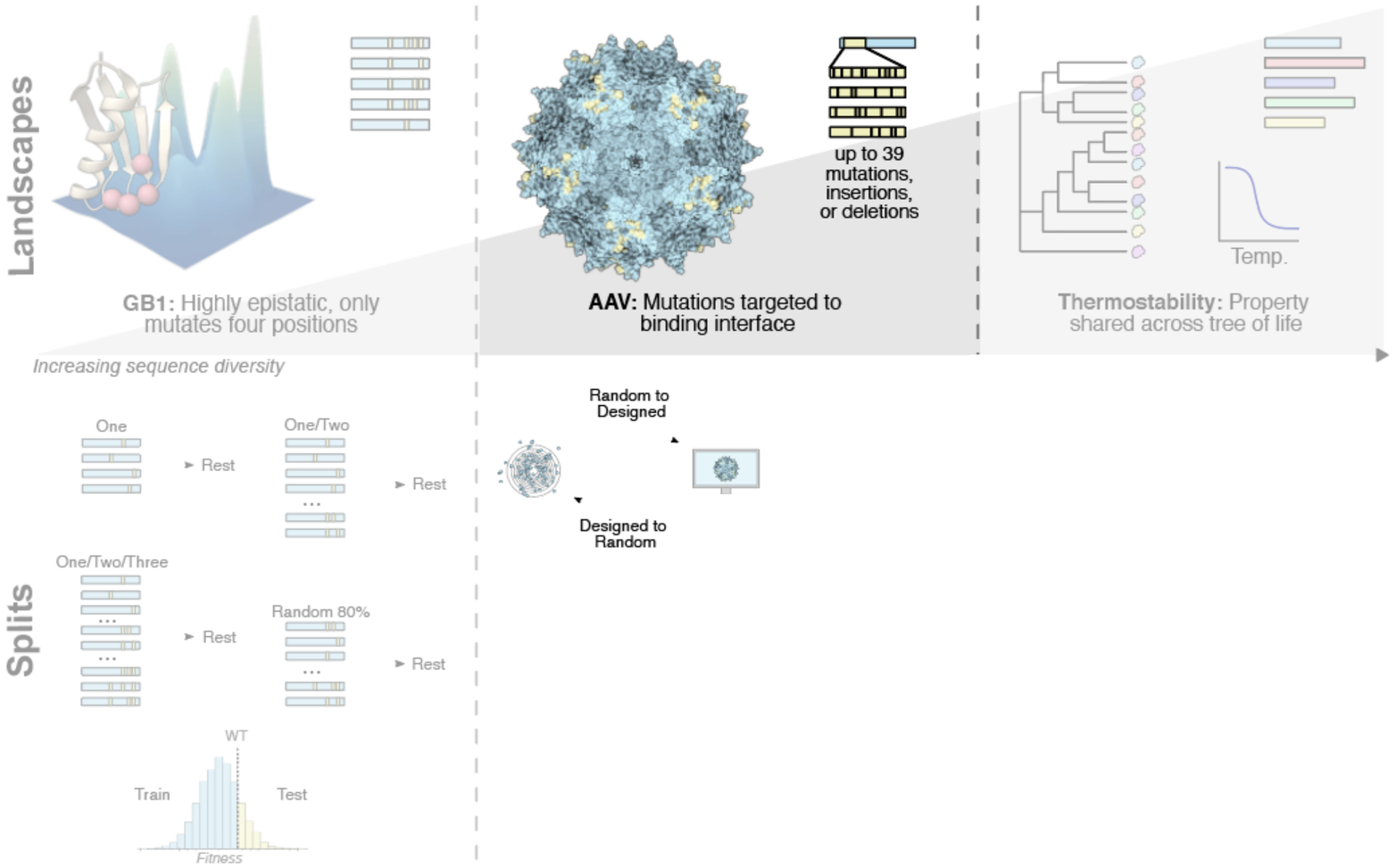


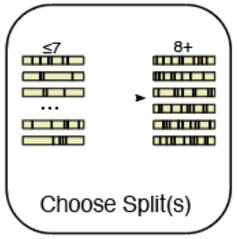
Make biologically-relevant train/test splits



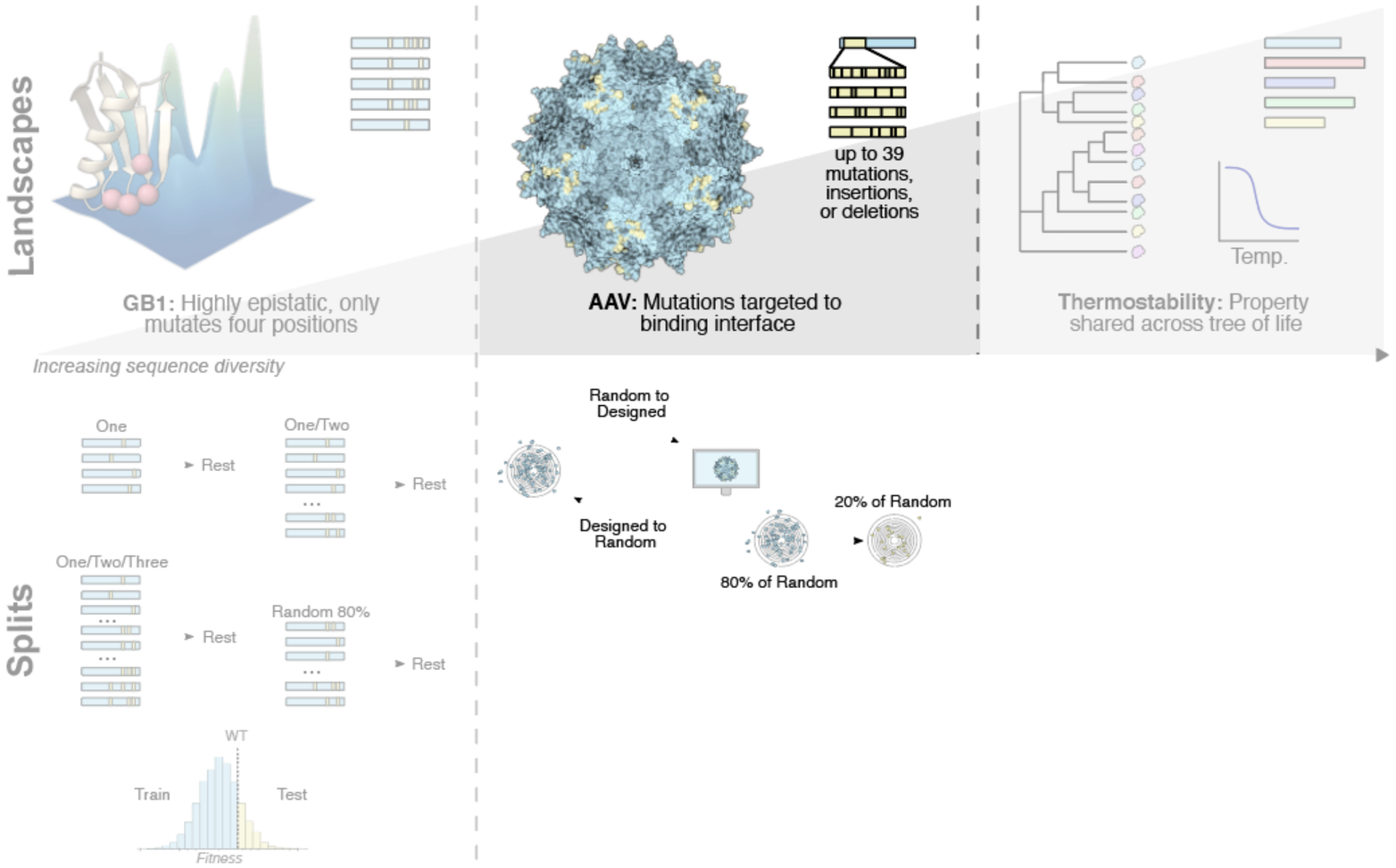


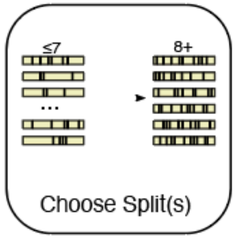
Make biologically-relevant train/test splits



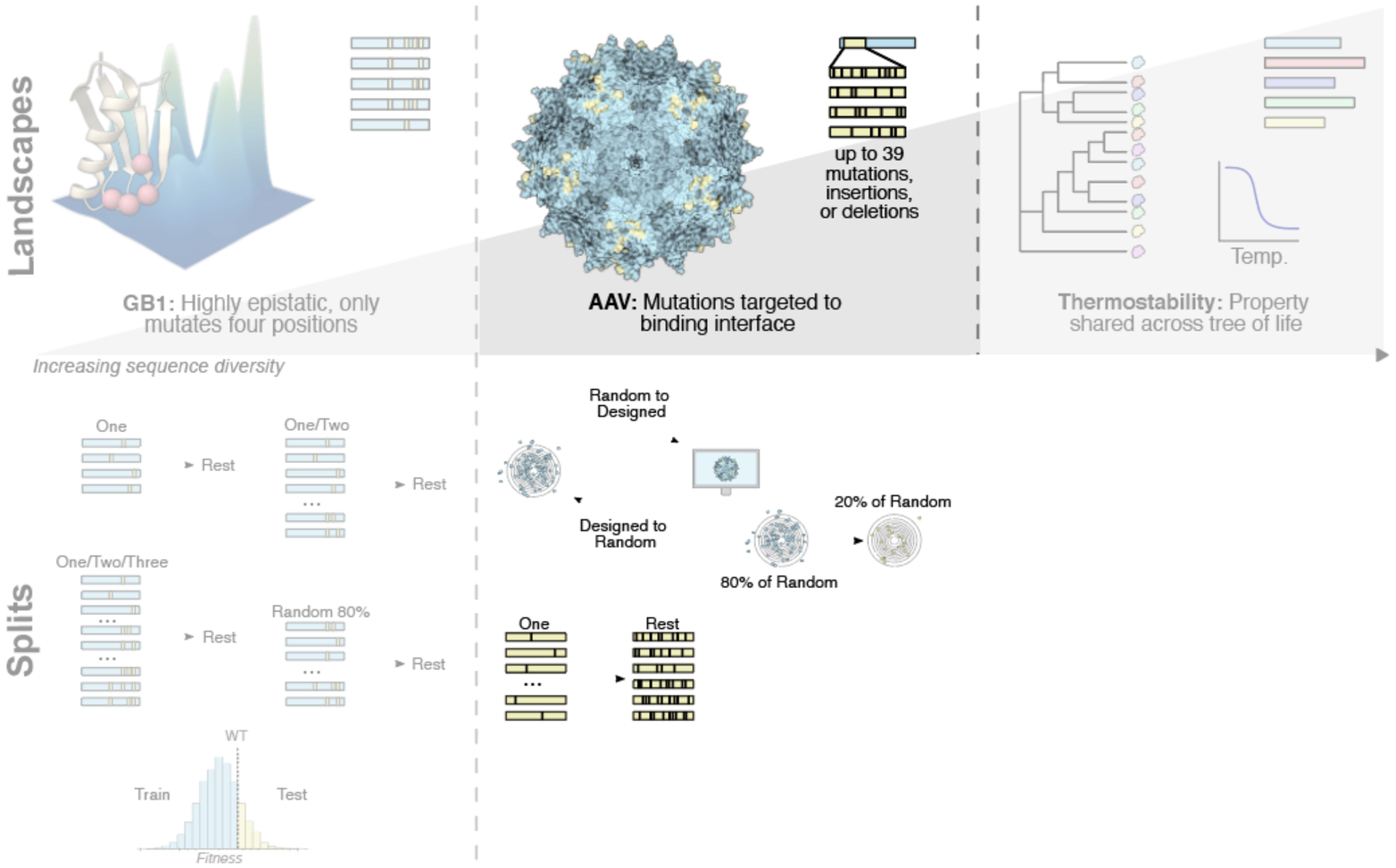


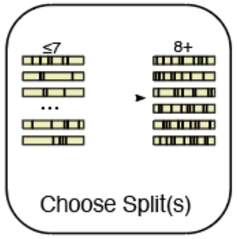
Make biologically-relevant train/test splits



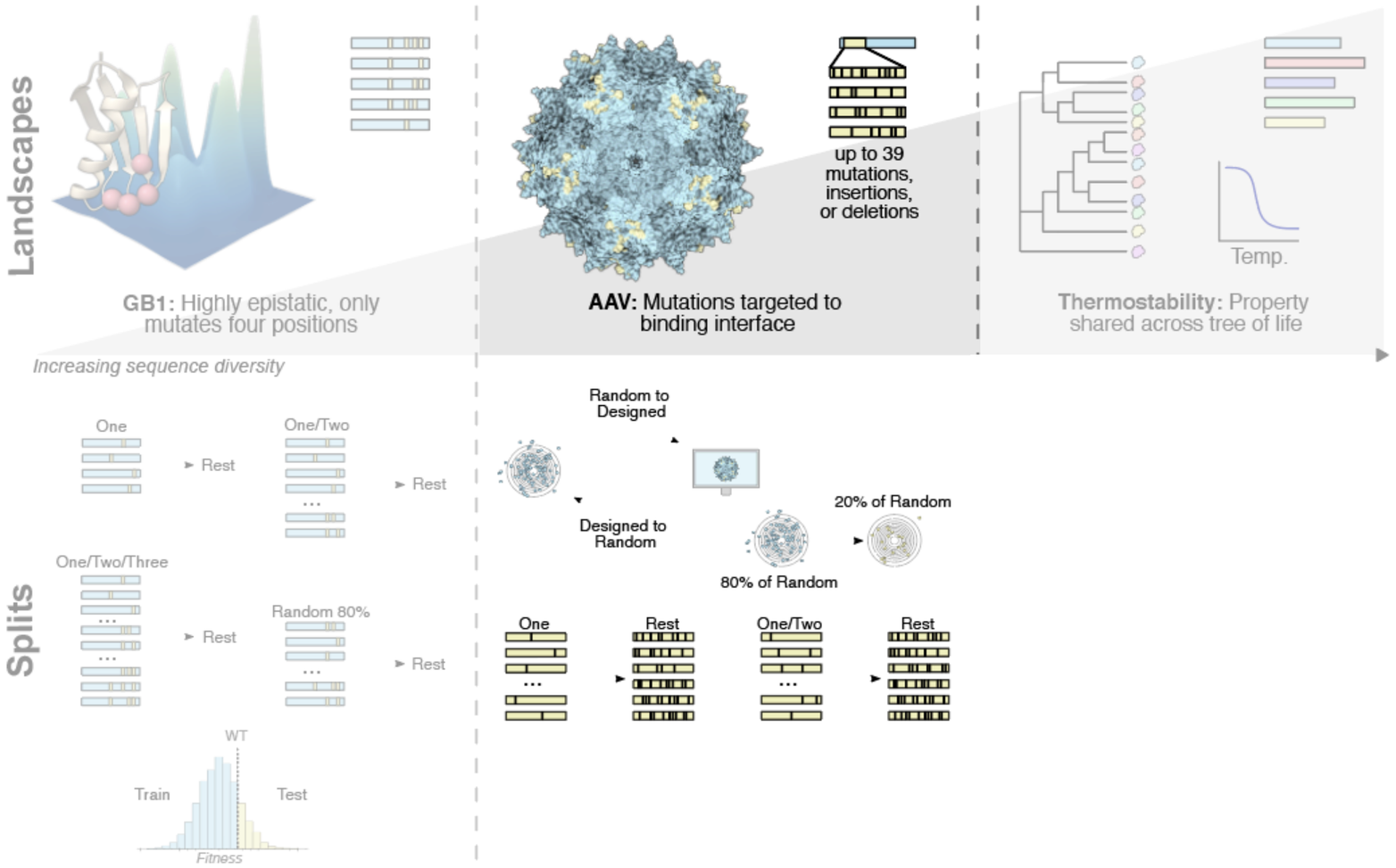


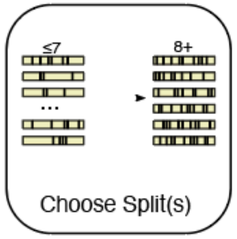
Make biologically-relevant train/test splits



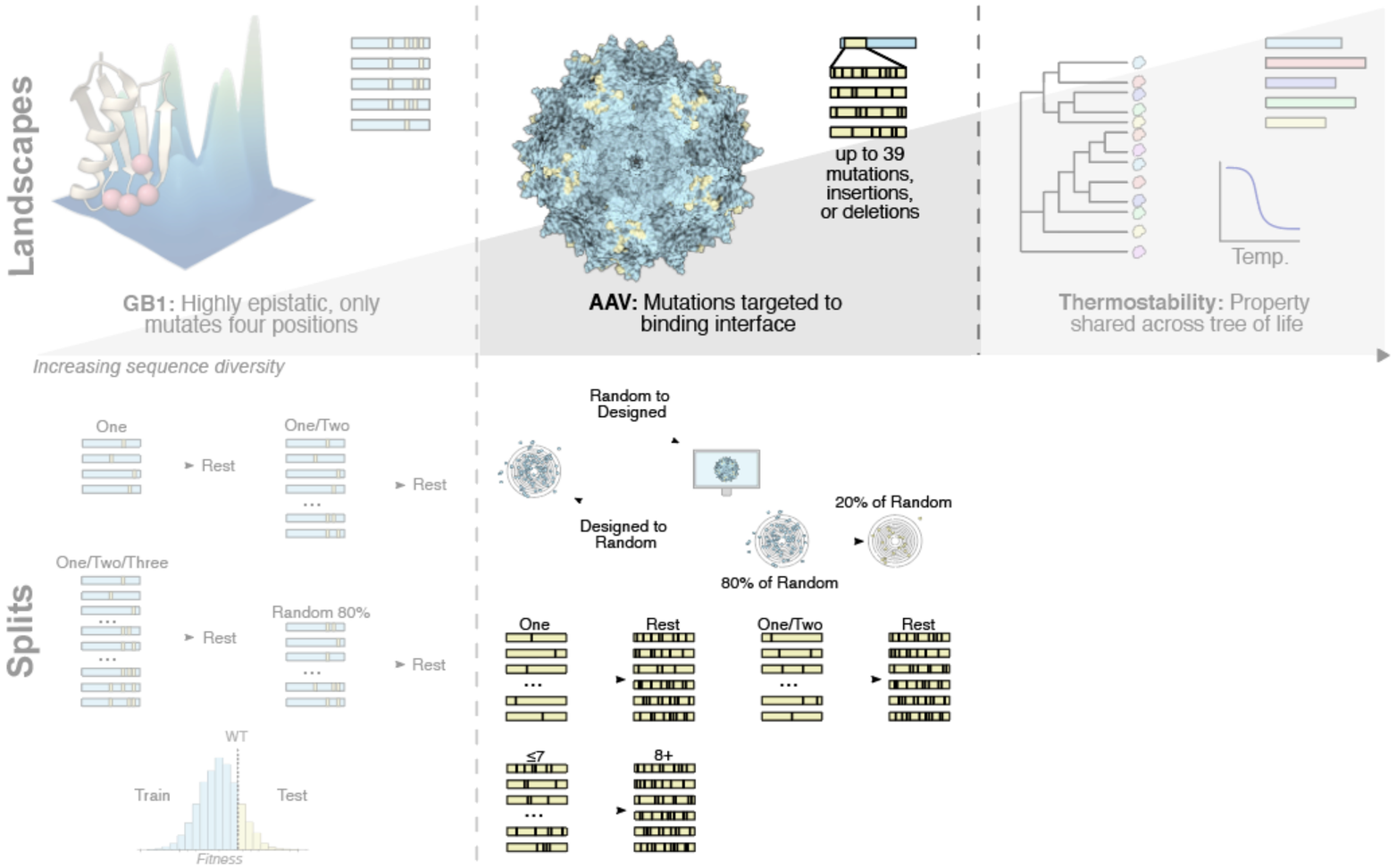


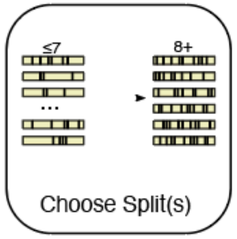
Make biologically-relevant train/test splits



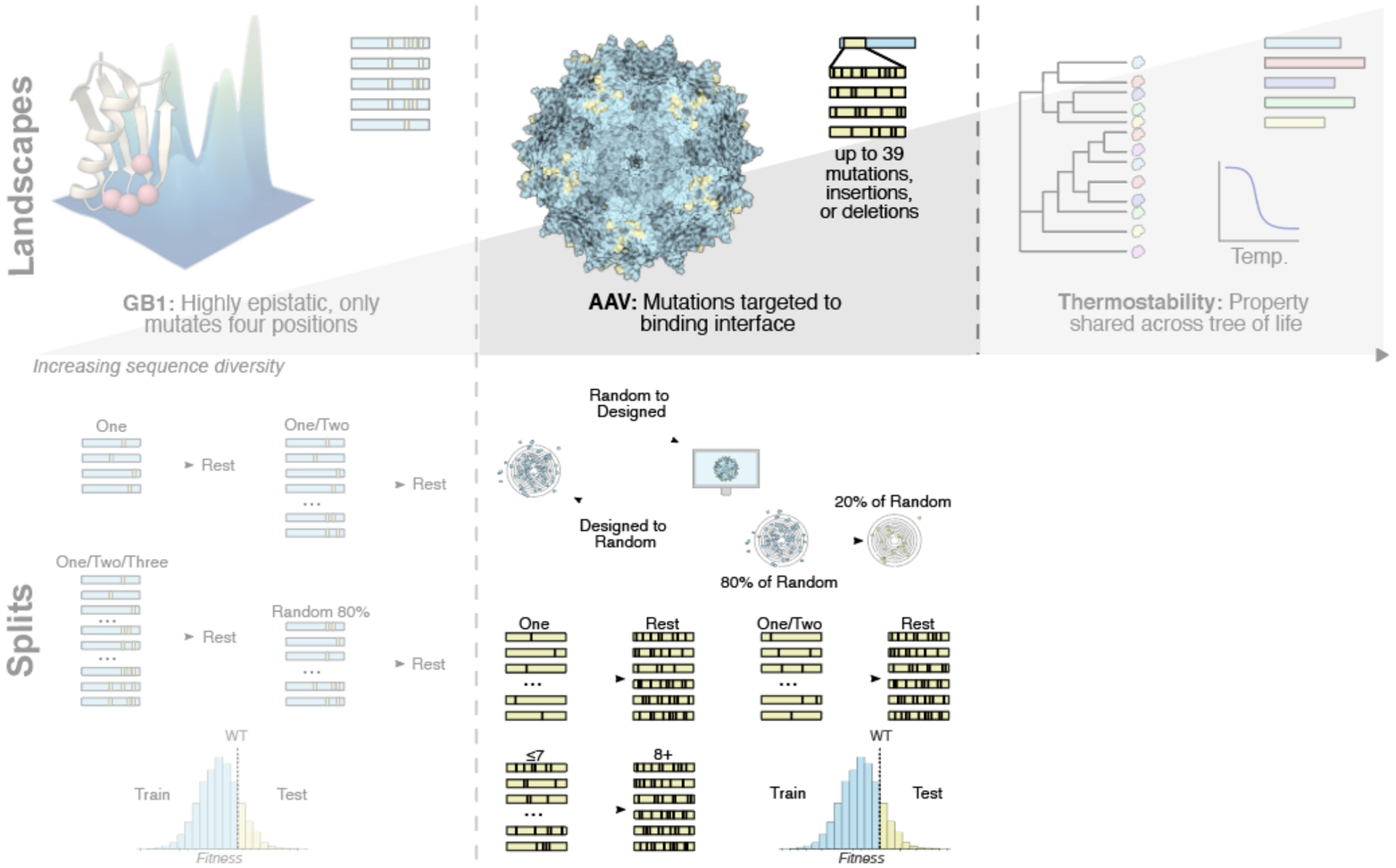


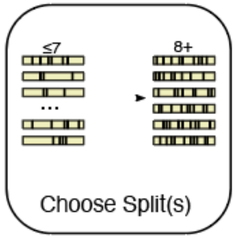
Make biologically-relevant train/test splits



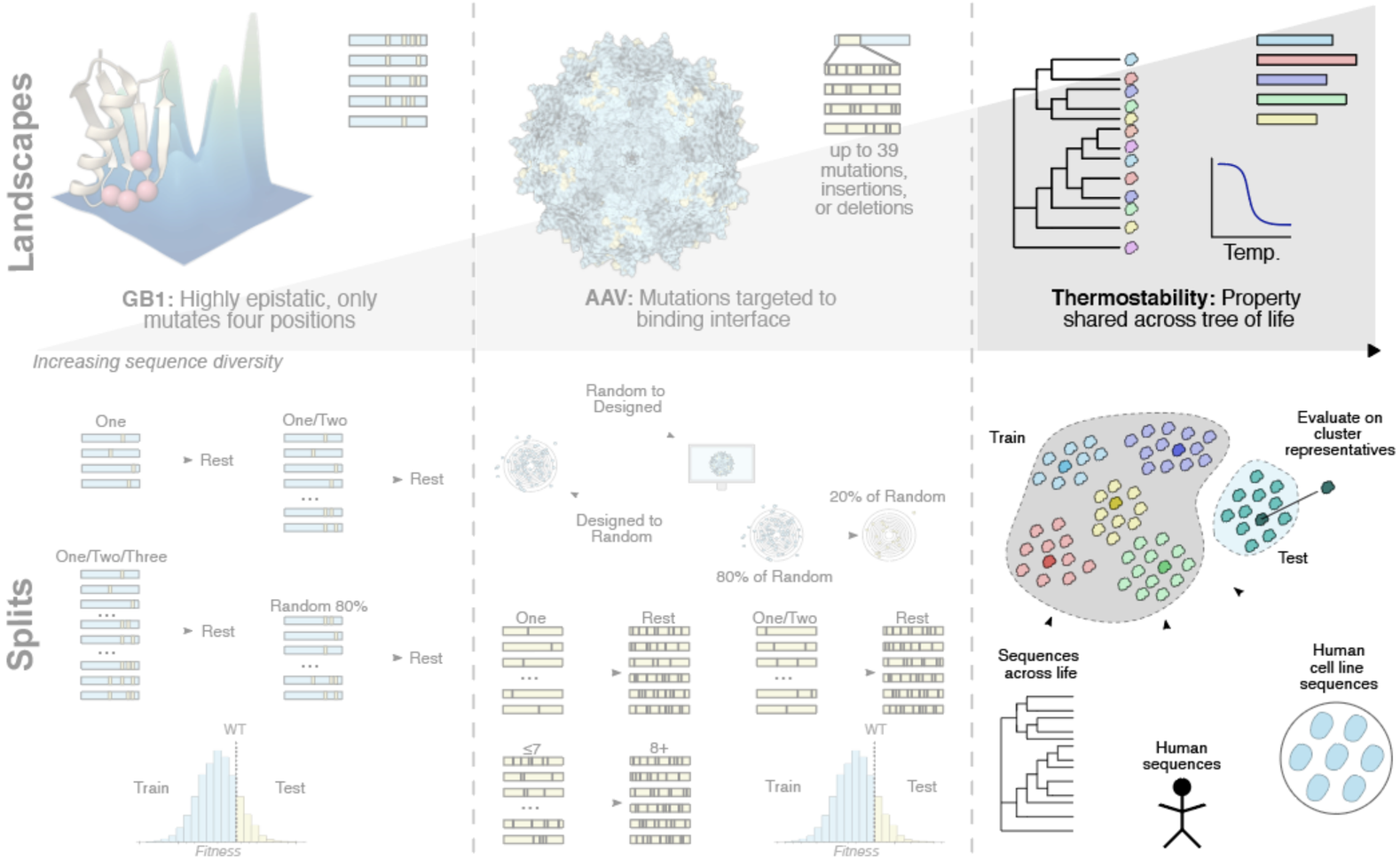


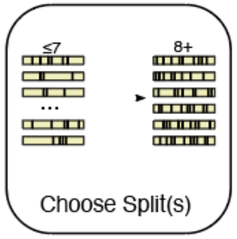
Make biologically-relevant train/test splits



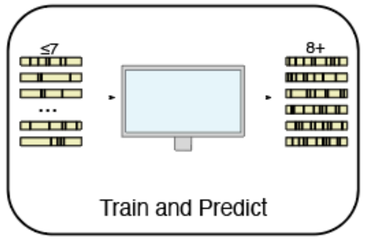


Make biologically-relevant train/test splits





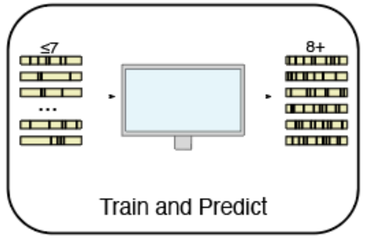
Evaluate many baseline models



Dallago *et al.* 2021

Evaluate many baseline models

Sequence distance

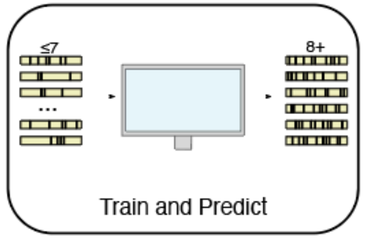


Dallago *et al.* 2021

Evaluate many baseline models

Sequence distance

Levenshtein distance



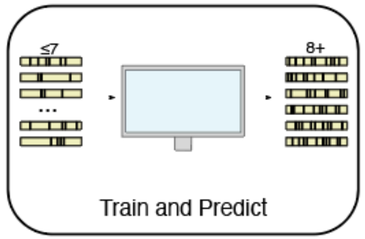
Dallago *et al.* 2021

Evaluate many baseline models

Sequence distance

Levenshtein distance

BLOSUM62



Dallago *et al.* 2021

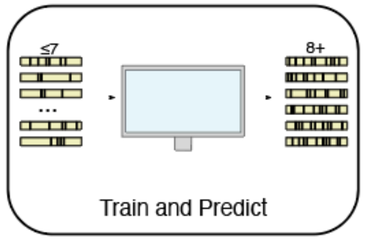
Evaluate many baseline models

Sequence distance

Levenshtein distance

BLOSUM62

Supervised models



Dallago *et al.* 2021

Evaluate many baseline models

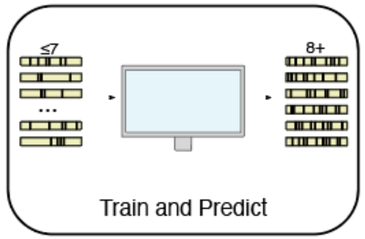
Sequence distance

Levenshtein distance

BLOSUM62

Supervised models

Ridge regression



Dallago *et al.* 2021

Evaluate many baseline models

Sequence distance

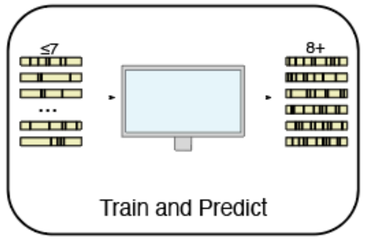
Levenshtein distance

BLOSUM62

Supervised models

Ridge regression

small CNN



Dallago *et al.* 2021

Evaluate many baseline models

Sequence distance

Levenshtein distance

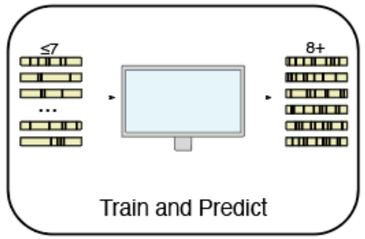
BLOSUM62

Supervised models

Ridge regression

small CNN

ESM-untrained
(750M par. transformer)



Dallago *et al.* 2021

Evaluate many baseline models

Sequence distance

Levenshtein distance

BLOSUM62

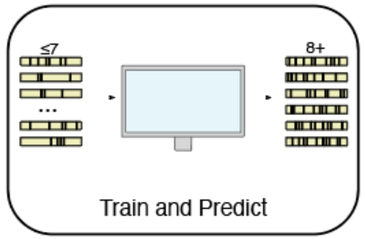
Supervised models

Ridge regression

small CNN

ESM-untrained
(750M par. transformer)

Pretrained models



Dallago *et al.* 2021

Evaluate many baseline models

Sequence distance

Levenshtein distance

BLOSUM62

Supervised models

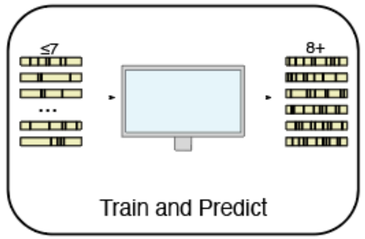
Ridge regression

small CNN

ESM-untrained
(750M par. transformer)

Pretrained models

ESM-1b
(pretrained on UniRef50)



Dallago *et al.* 2021

Evaluate many baseline models

Sequence distance

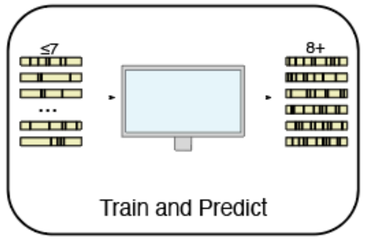
- Levenshtein distance
- BLOSUM62

Supervised models

- Ridge regression
- small CNN
- ESM-untrained (750M par. transformer)

Pretrained models

- ESM-1b (pretrained on UniRef50)
- ESM-1v (pretrained on UniRef90)



Dallago *et al.* 2021

Evaluate many baseline models

Sequence distance

- Levenshtein distance
- BLOSUM62

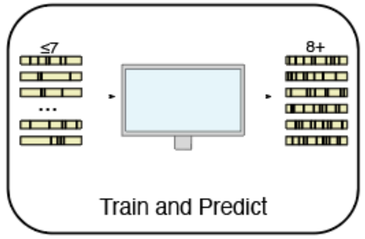
Supervised models

- Ridge regression
- small CNN
- ESM-untrained (750M par. transformer)

Pretrained models

- ESM-1b (pretrained on UniRef50)
- ESM-1v (pretrained on UniRef90)

Best on highly-diverse landscapes (thermostability) or when data is very limited



Dallago *et al.* 2021

Evaluate many baseline models

Sequence distance

- Levenshtein distance
- BLOSUM62

Supervised models

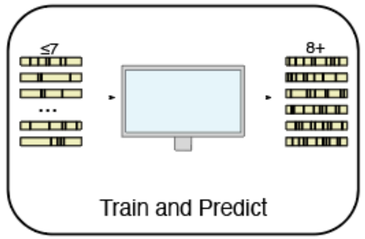
- Ridge regression
- small CNN
- ESM-untrained (750M par. transformer)

Strong performance on most landscapes

Pretrained models

- ESM-1b (pretrained on UniRef50)
- ESM-1v (pretrained on UniRef90)

Best on highly-diverse landscapes (thermostability) or when data is very limited

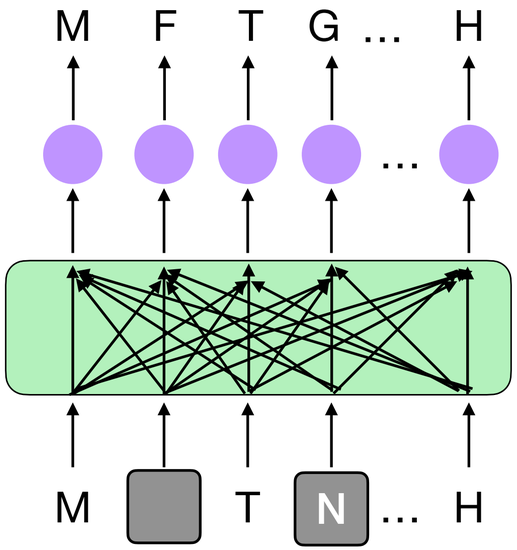


Dallago *et al.* 2021

Try your models!

https://benchmark.protein.properties/home

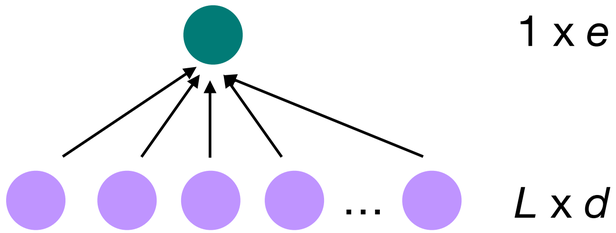
Local-to-global aggregation affects model performance



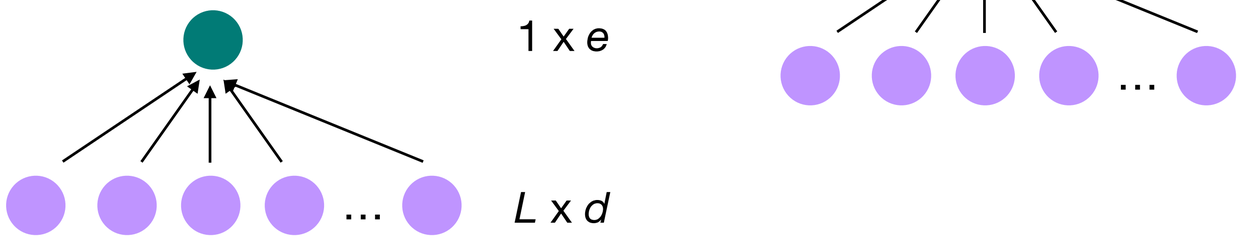
Local-to-global aggregation affects model performance

Local-to-global aggregation affects model performance

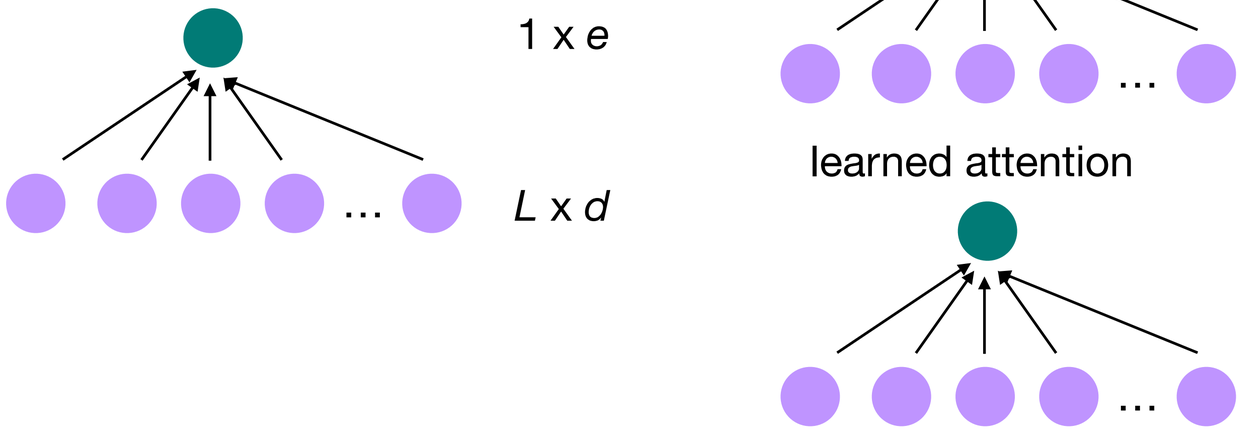
Local-to-global aggregation affects model performance



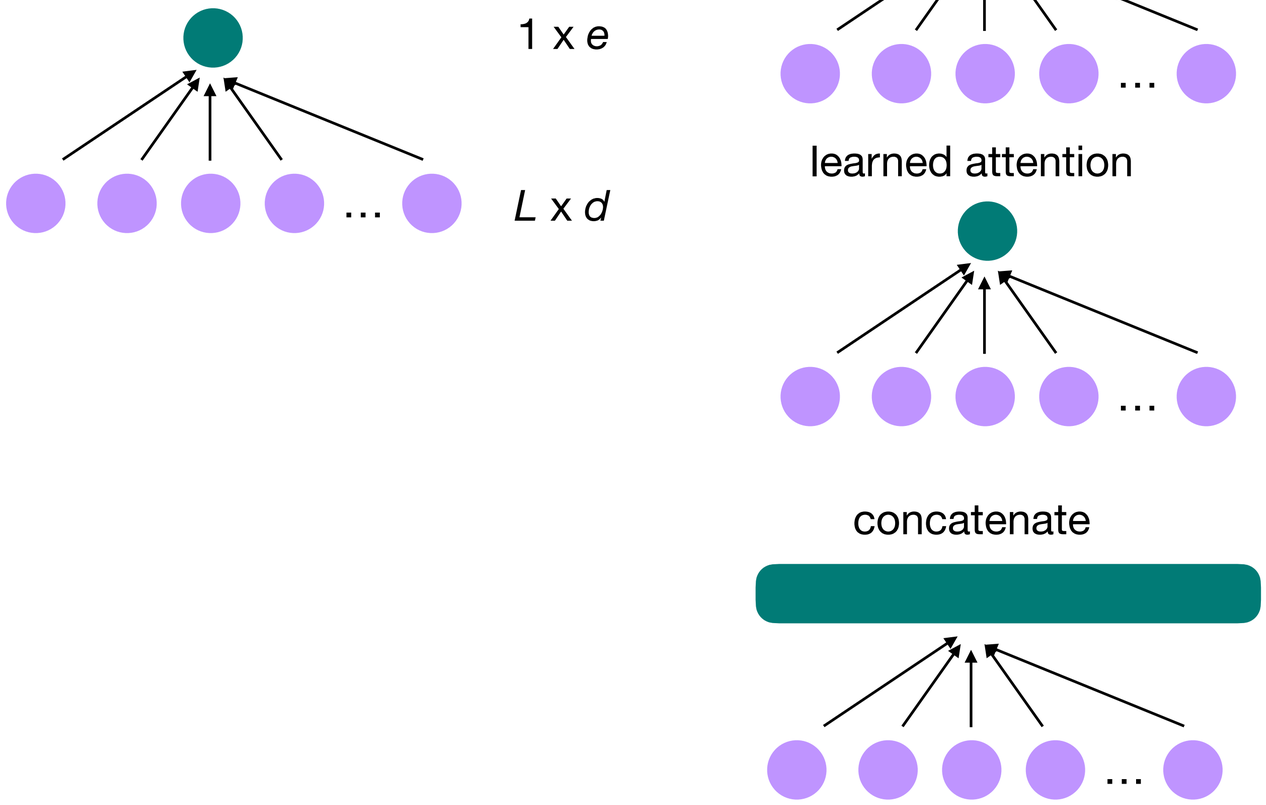
Local-to-global aggregation affects model performance



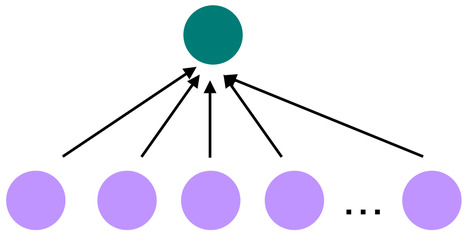
Local-to-global aggregation affects model performance



Local-to-global aggregation affects model performance



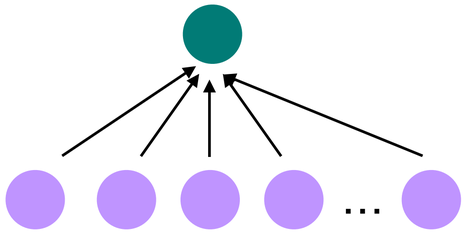
Local-to-global aggregation affects model performance



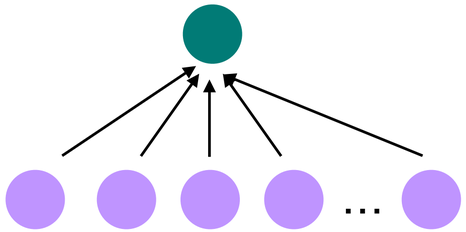
1 x *e*

L x *d*

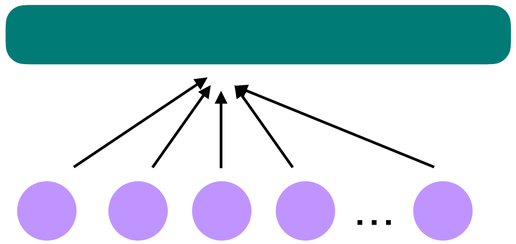
mean



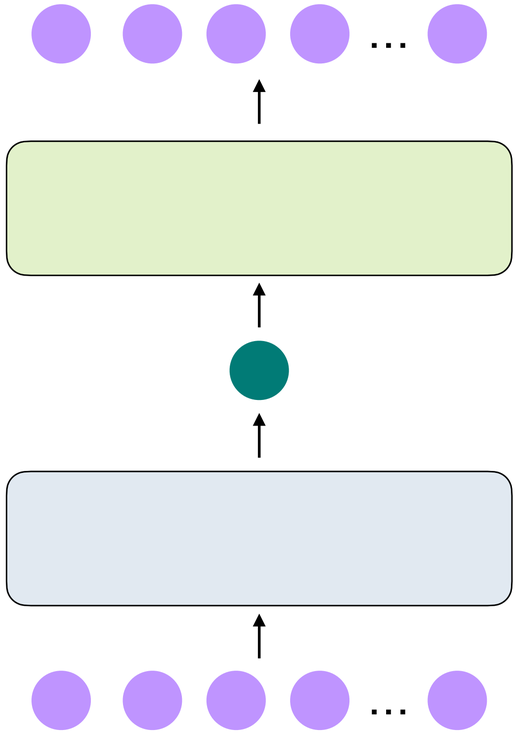
learned attention



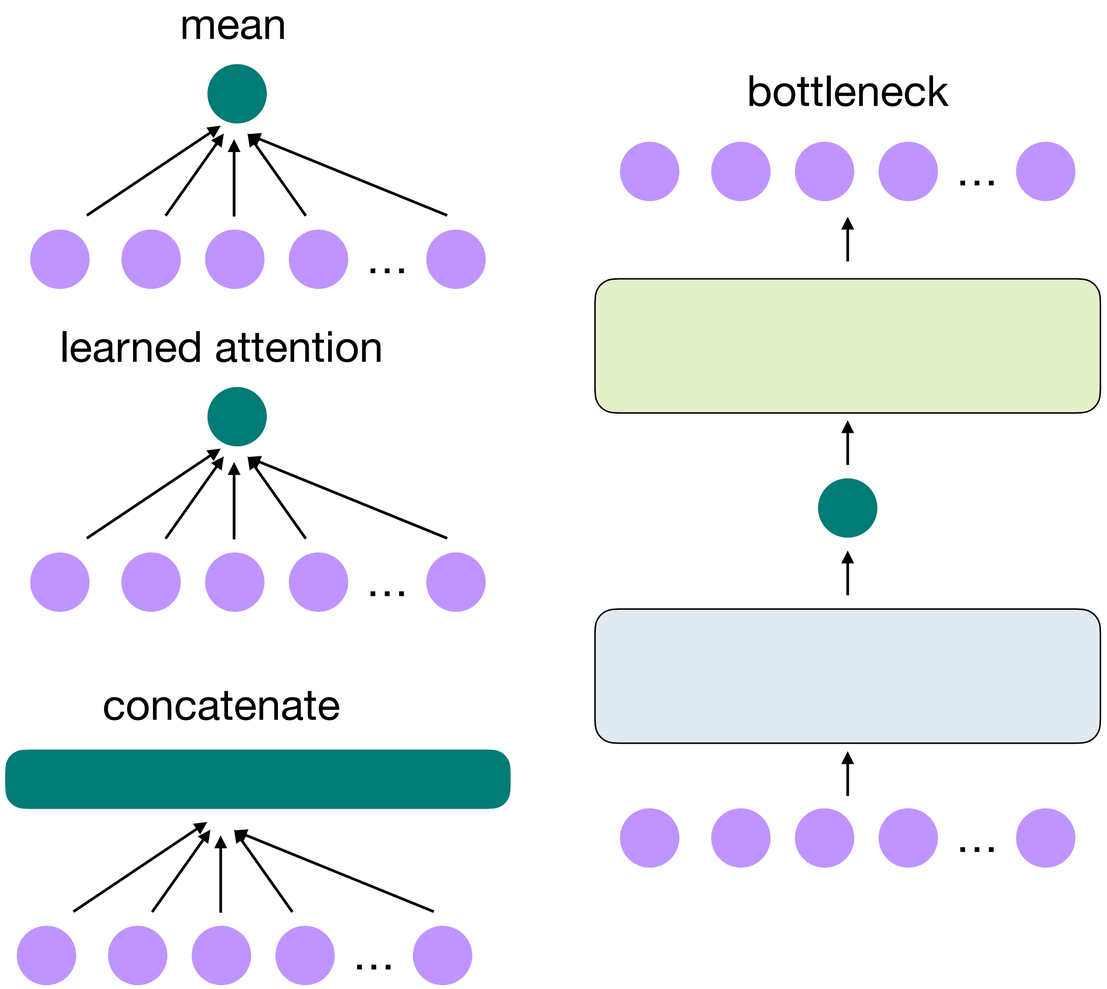
concatenate



bottleneck



Local-to-global aggregation affects model performance



	Stability (Corr.)	Fluorescence (Corr.)	Homology (Acc.)
Mean	0.42	0.19	0.27
Attention	0.65	0.23	0.27
Light Att.	0.66	0.23	0.27
Maximum	0.02	0.02	0.28
MeanMax	0.37	0.15	0.26
KMax	0.10	0.11	0.27
Concat	0.74	0.69	0.34
Bottleneck	**0.79**	**0.78**	**0.41**

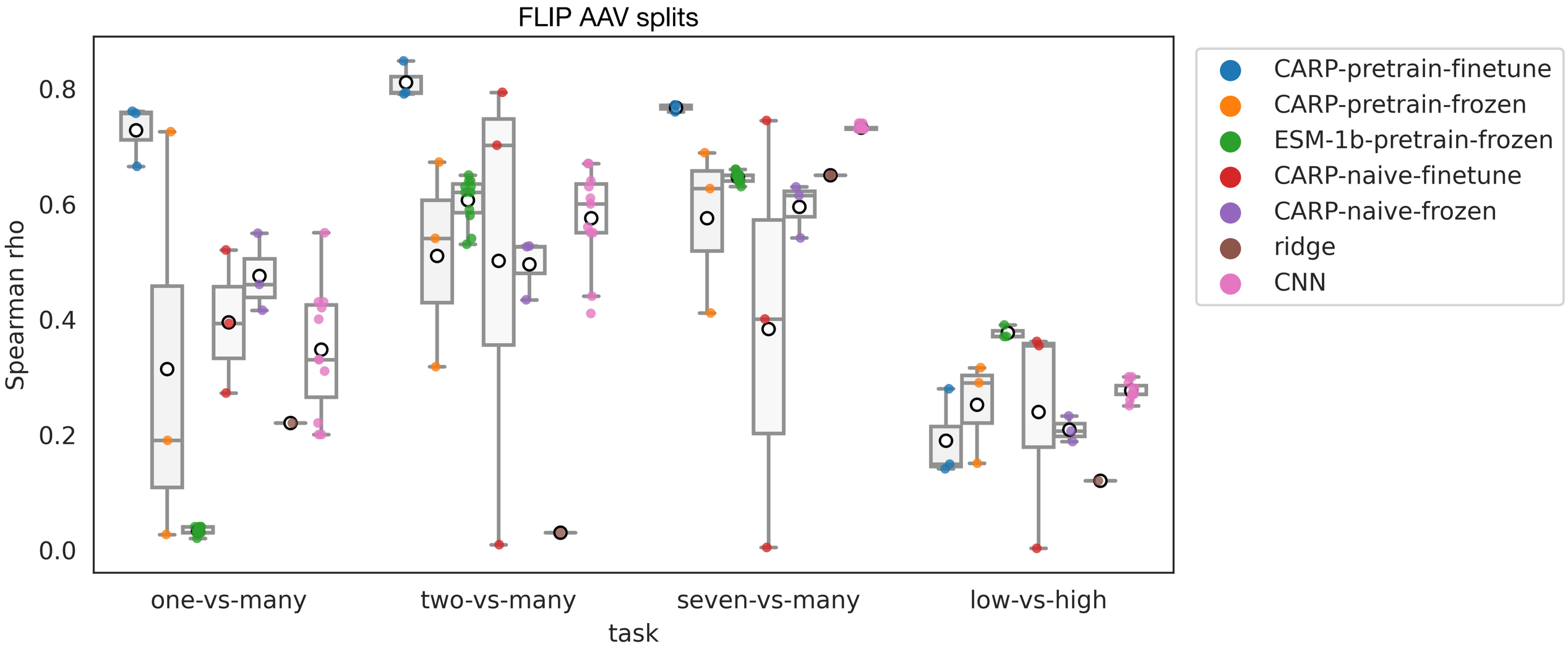
Detlefsen *et al.* 2022

Finetuning/fixing pretrained weights also matters

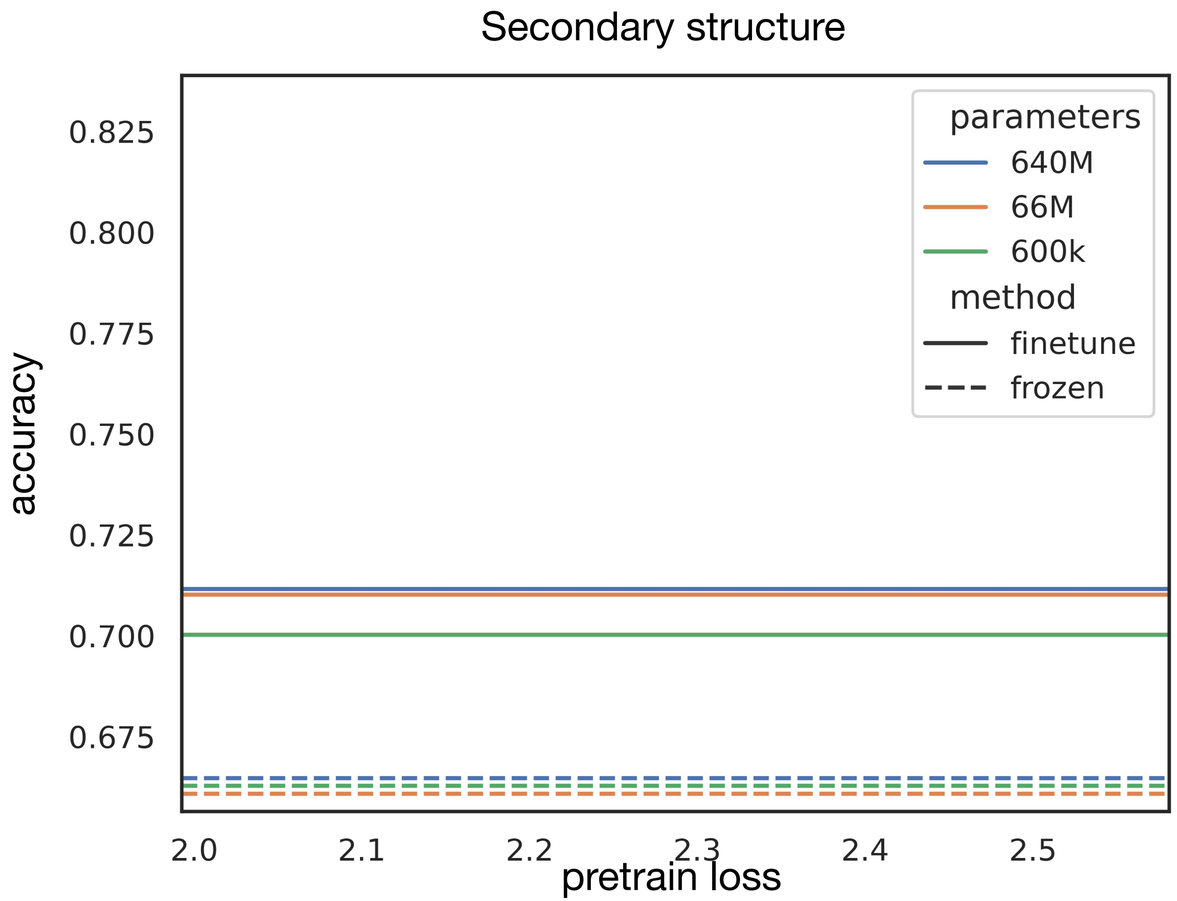
	Remote Homology			Fluorescence			Stability		
	Resnet	**LSTM**	**Trans**	**Resnet**	**LSTM**	**Trans**	**Resnet**	**LSTM**	**Trans**
PRE+FIX	0.27	**0.37**	0.27	0.23	**0.74**	0.48	0.65	0.70	0.62
PRE+FIN	0.17	0.26	0.21	0.21	0.67	0.68	**0.73**	0.69	**0.73**
RNG+FIX	0.03	0.10	0.04	0.25	0.63	0.14	0.21	0.61	–
RNG+FIN	0.10	0.12	0.09	−0.28	0.21	0.22	0.61	0.28	−0.06
Baseline	0.09 (Accuracy)			0.14 (Correlation)			0.19 (Correlation)		

Detlefsen *et al.* 2022

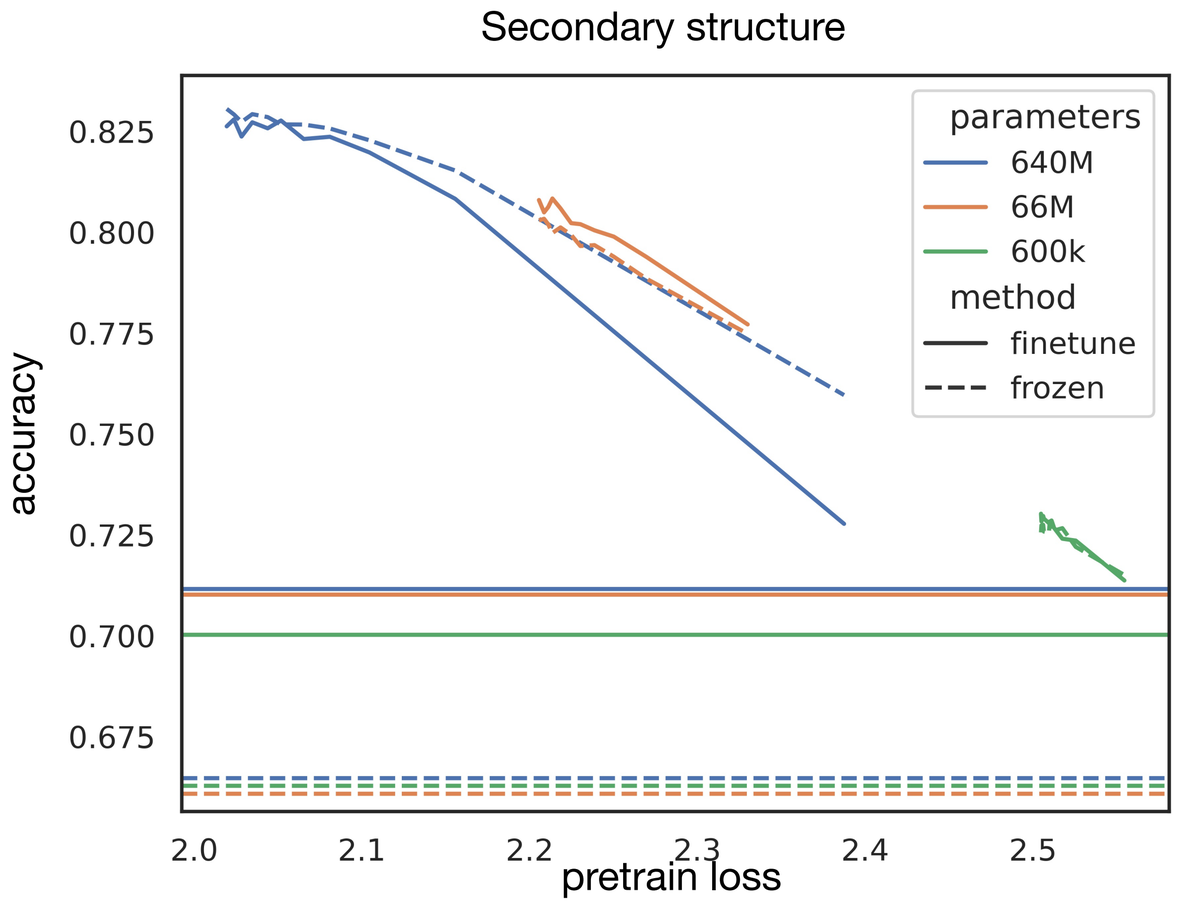
Finetuning/fixing pretrained weights also matters



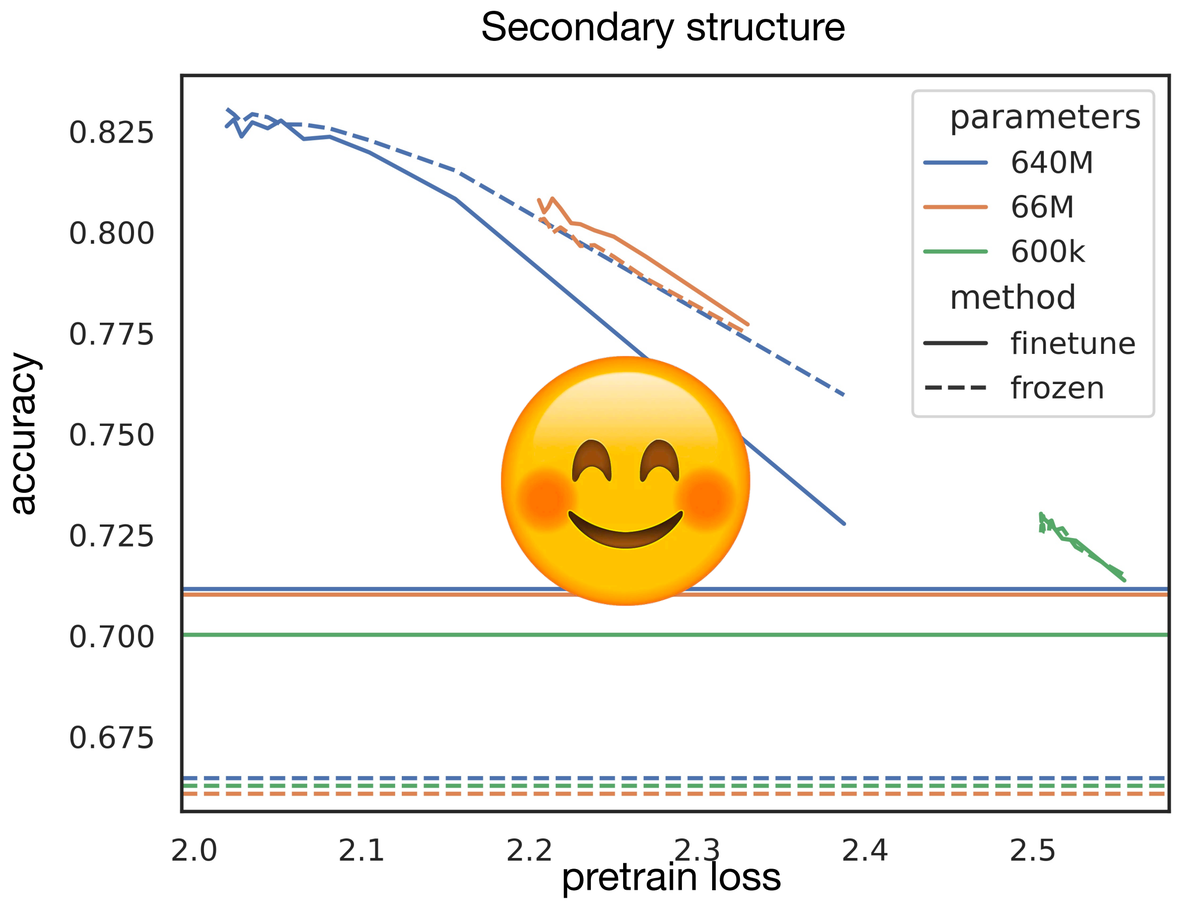
When does pretraining help?



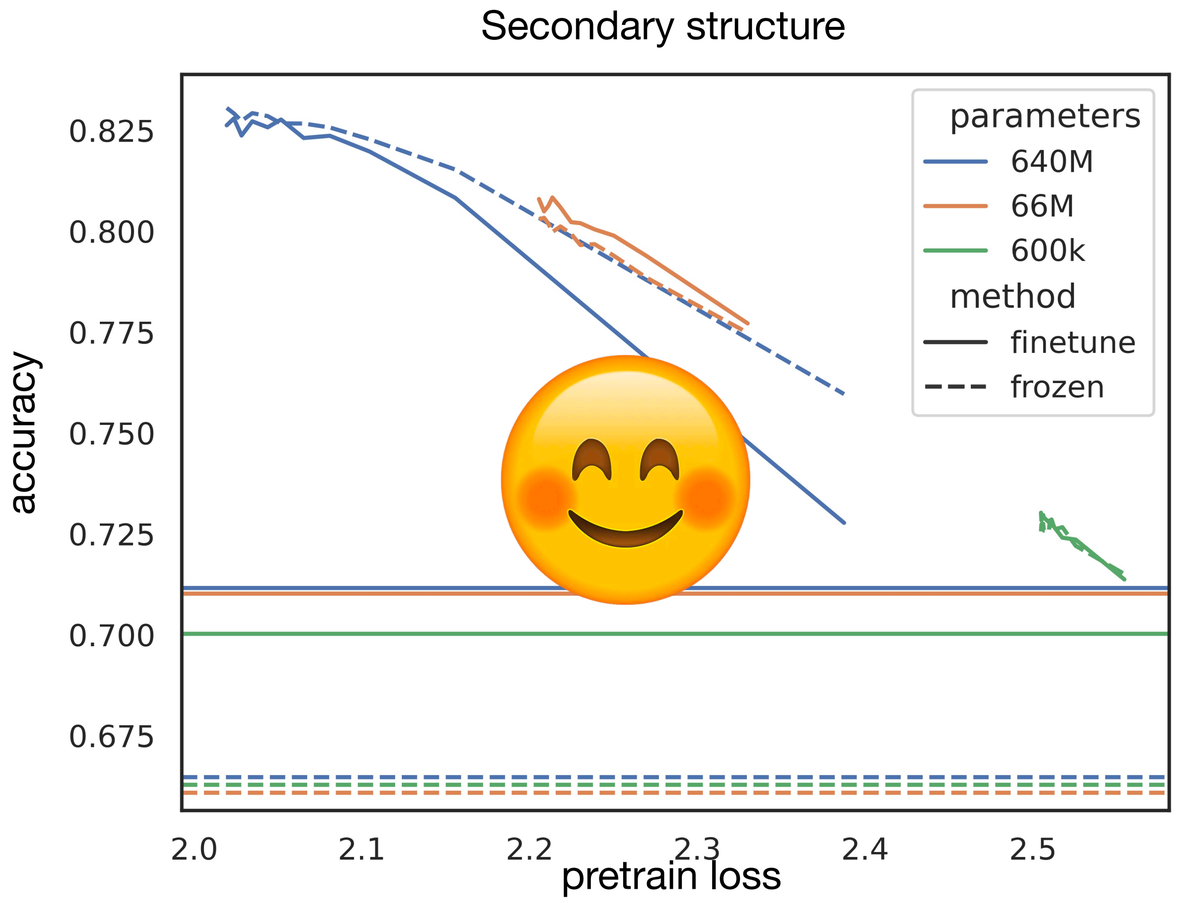
When does pretraining help?

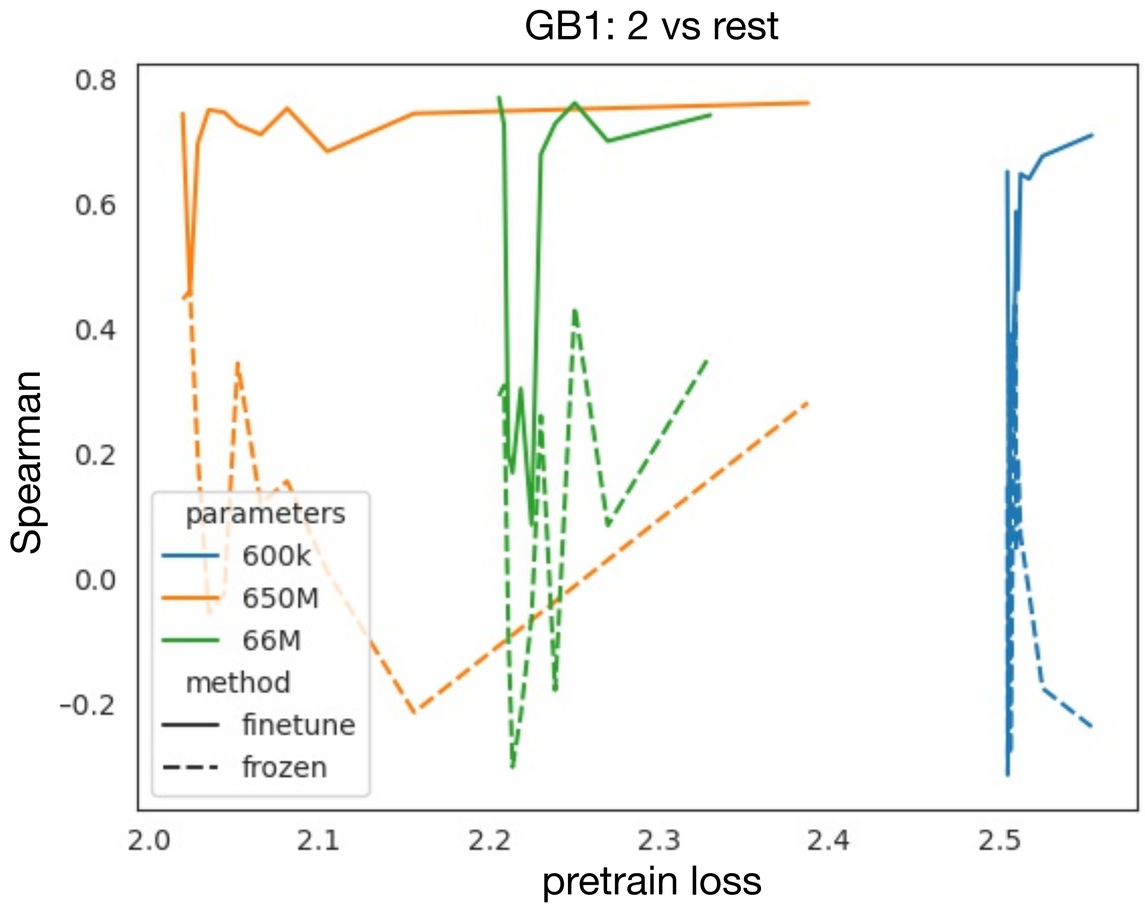


When does pretraining help?

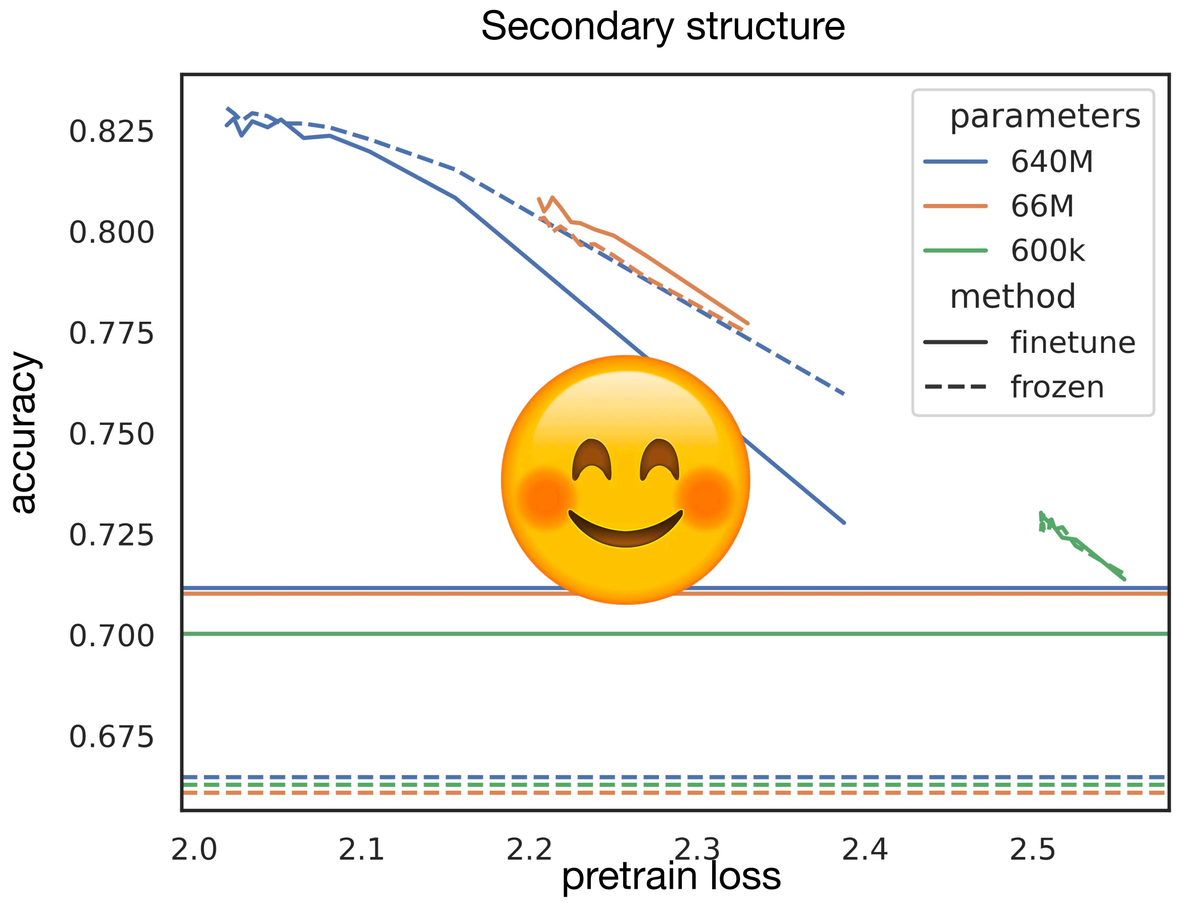


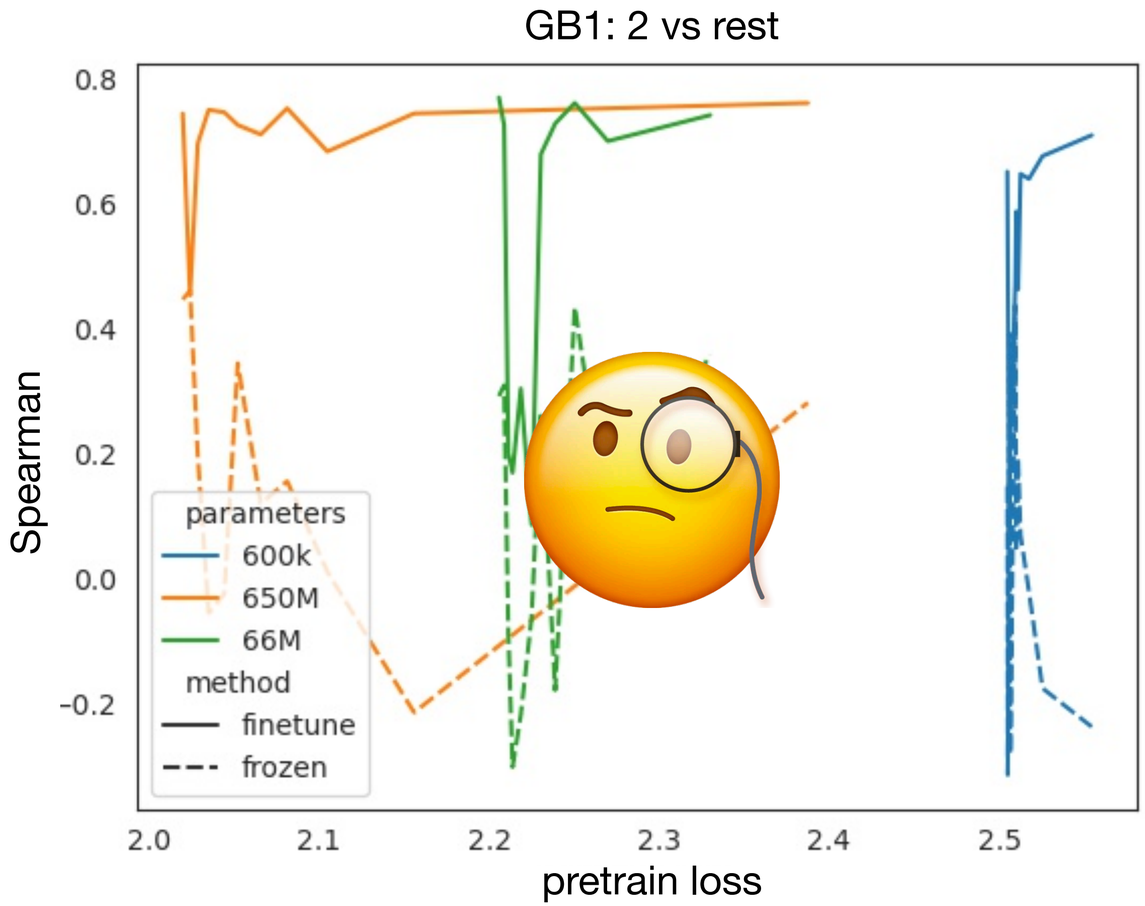
When does pretraining help?



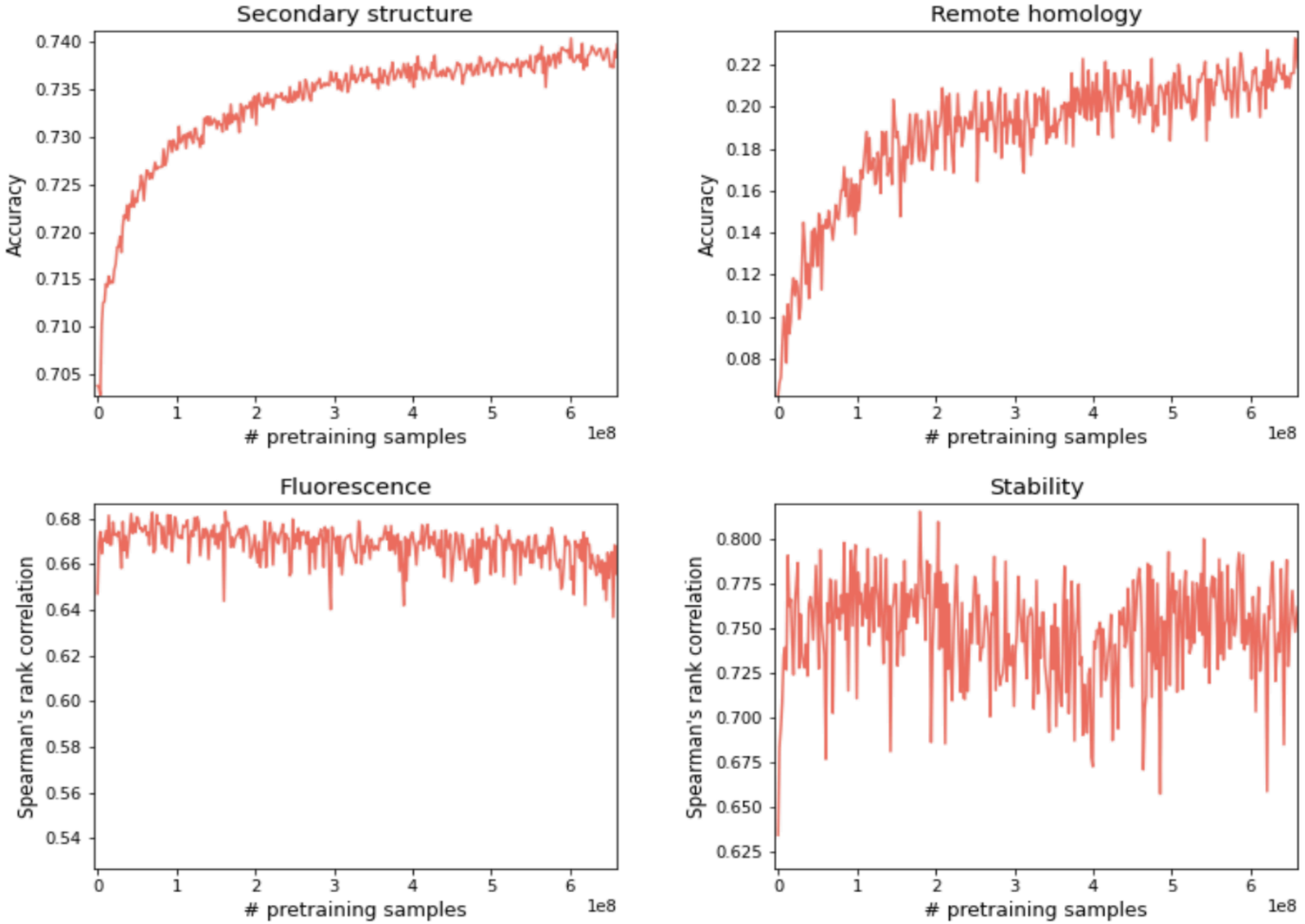


When does pretraining help?





When does pretraining help?



Brandes *et al.* 2021

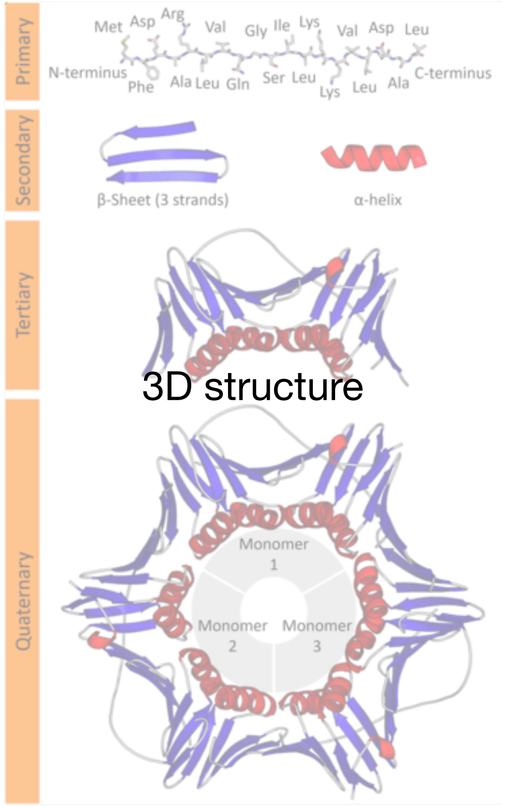
Protein sequences are not language

Protein sequences are not language

Differences between proteins and NL

Protein sequences are not language

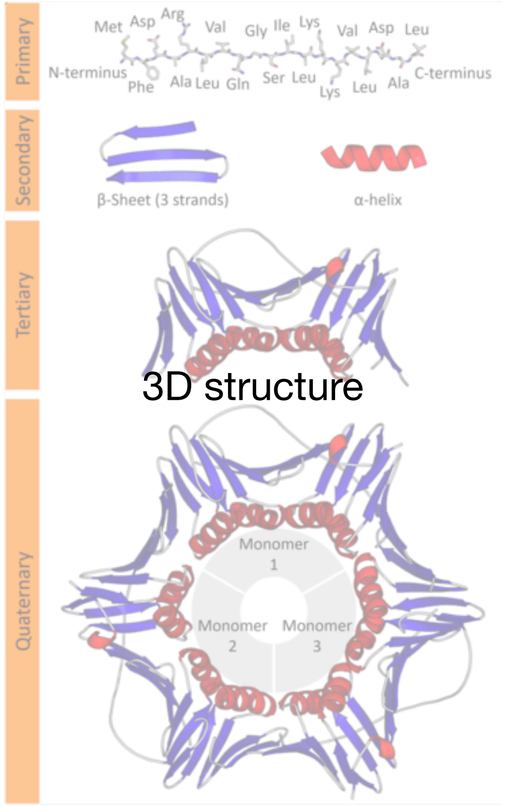
Differences between proteins and NL



Thomas Shafee

Protein sequences are not language

Differences between proteins and NL

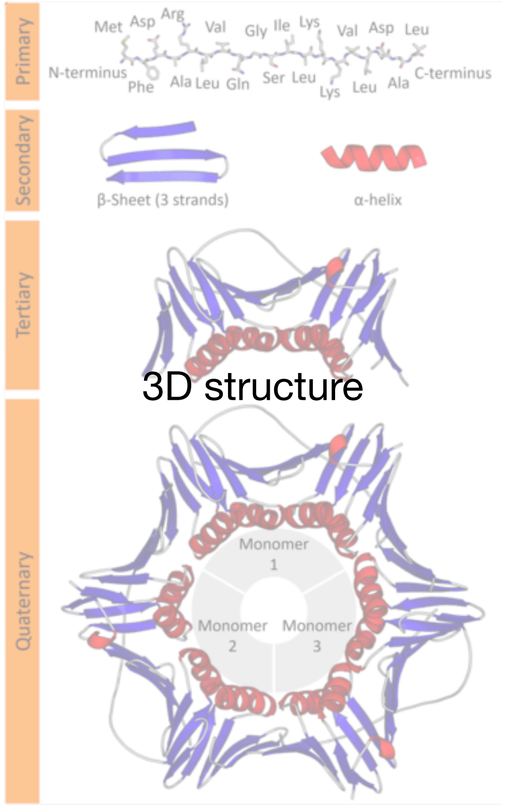


Thomas Shafee

ACDEFGHIKLMNPQRSTVWY
vs
words and subwords

Protein sequences are not language

Differences between proteins and NL

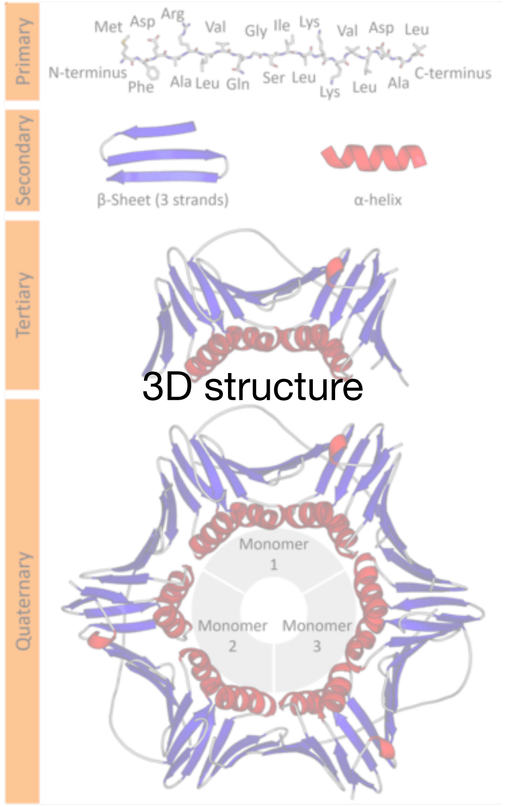


Thomas Shafee

ACDEFGHIKLMNPQRSTVWY
vs
words and subwords

Protein sequences are not language

Differences between proteins and NL



Thomas Shafee

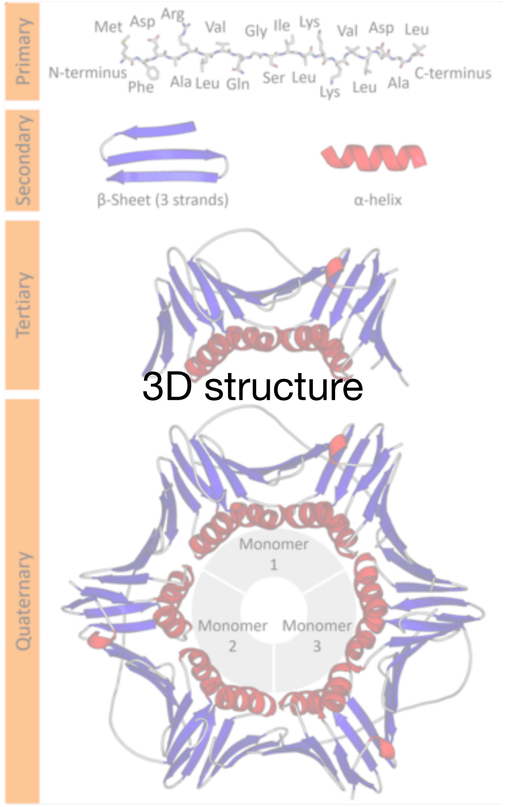
ACDEFGHIKLMNPQRSTVWY
vs
words and subwords

evolution + sequencing

Multi-modal models

Protein sequences are not language

Differences between proteins and NL



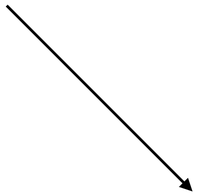
Thomas Shafee

ACDEFGHIKLMNPQRSTVWY
vs
words and subwords

evolution + sequencing

Multi-modal models

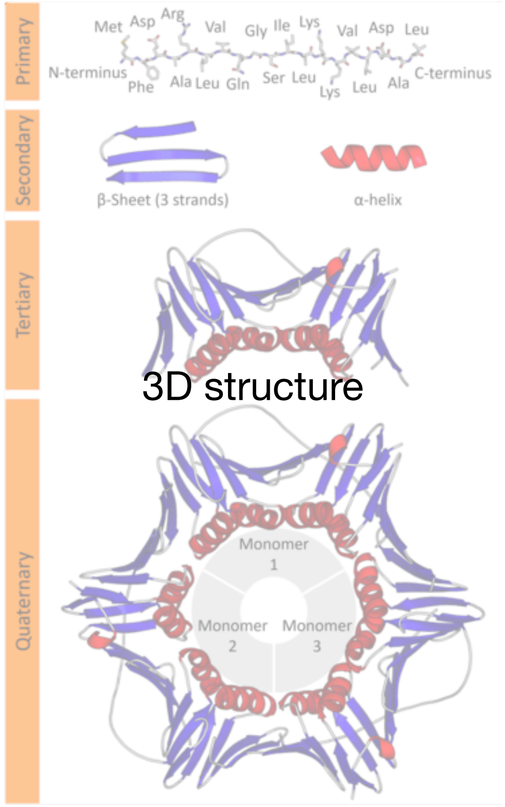
sequence `MGHTQWI...`



model

Protein sequences are not language

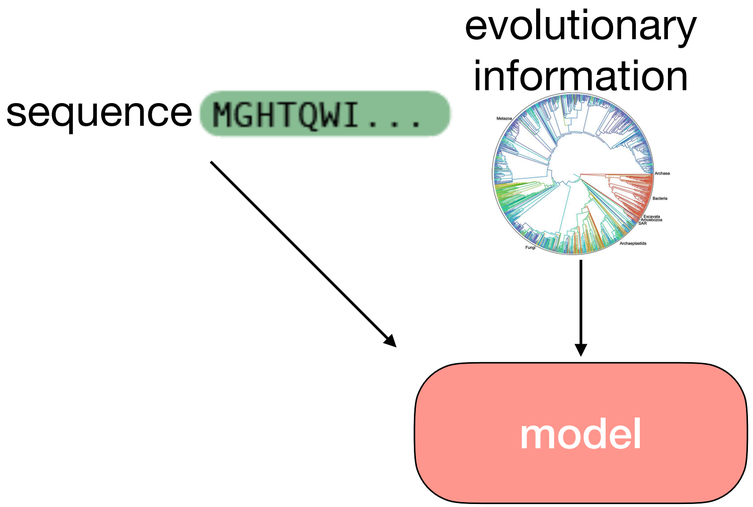
Differences between proteins and NL



Thomas Shafee

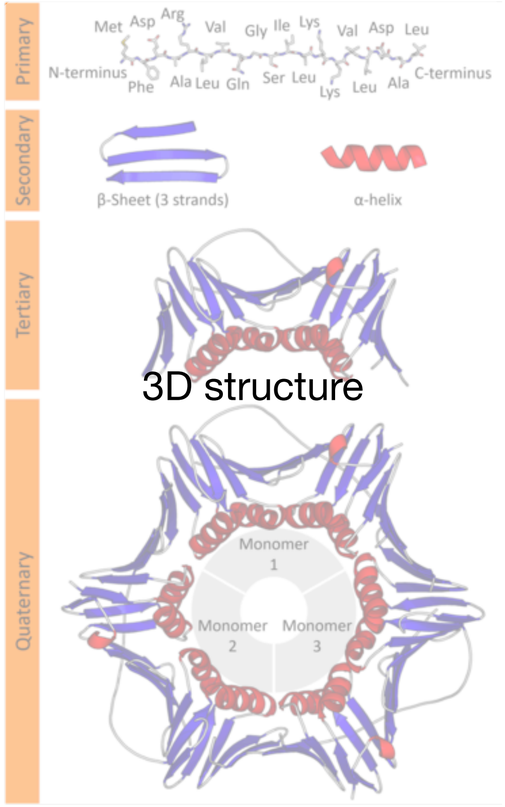
ACDEFGHIKLMNPQRSTVWY
vs
words and subwords

Multi-modal models



Protein sequences are not language

Differences between proteins and NL

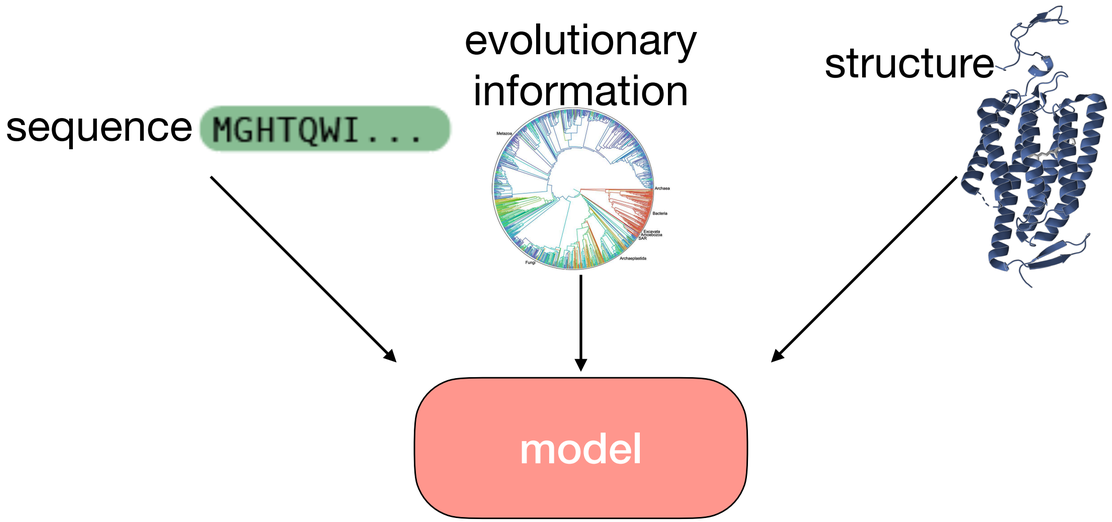


Thomas Shafee

ACDEFGHIKLMNPQRSTVWY
vs
words and subwords

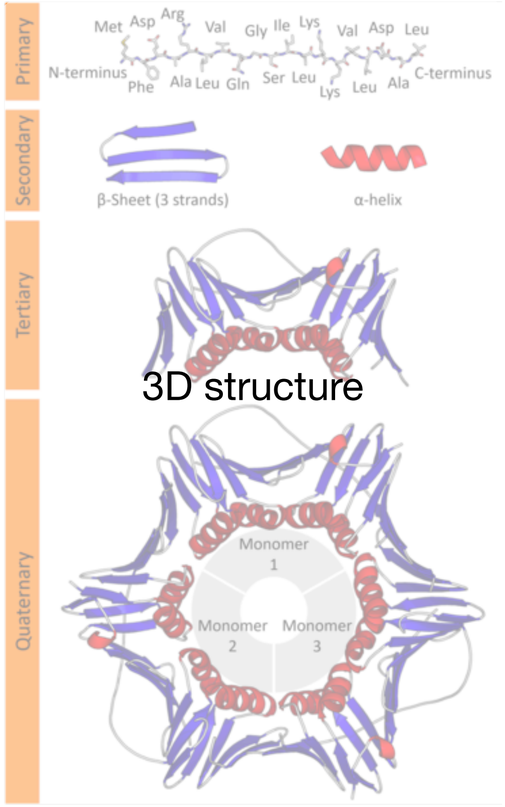
evolution + sequencing

Multi-modal models



Protein sequences are not language

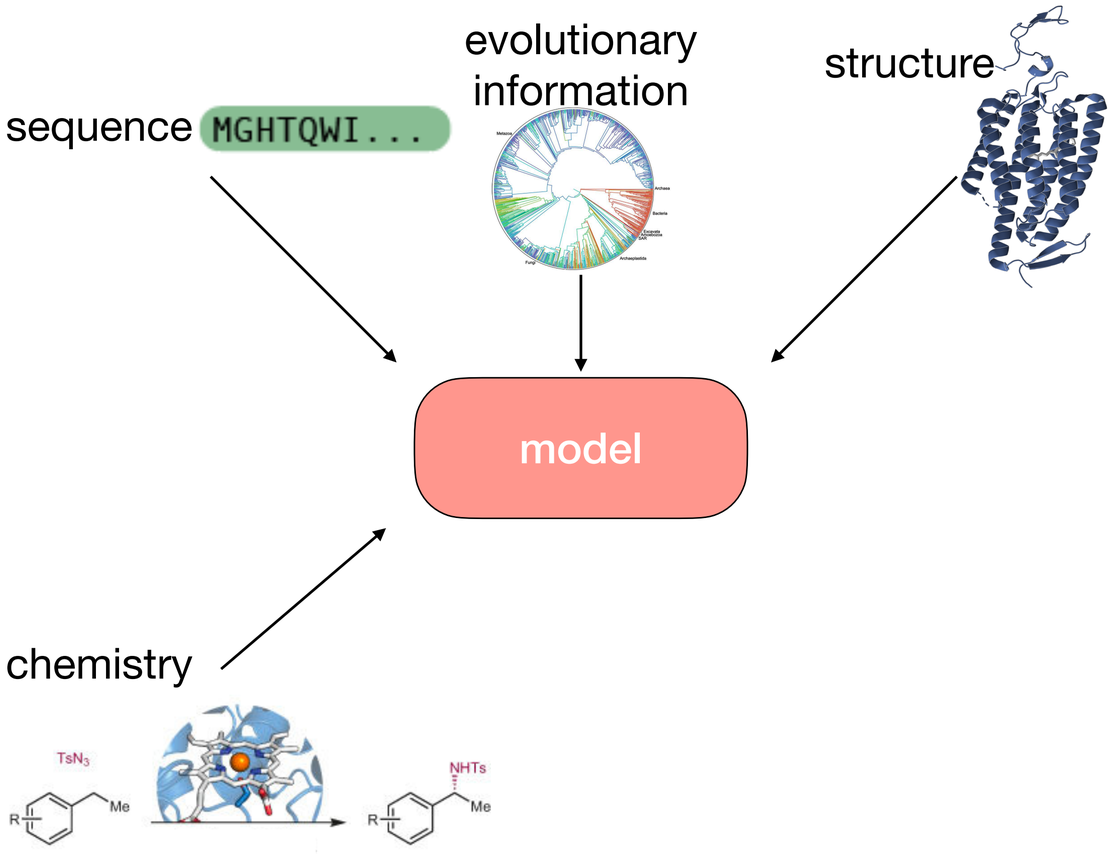
Differences between proteins and NL



Thomas Shafee

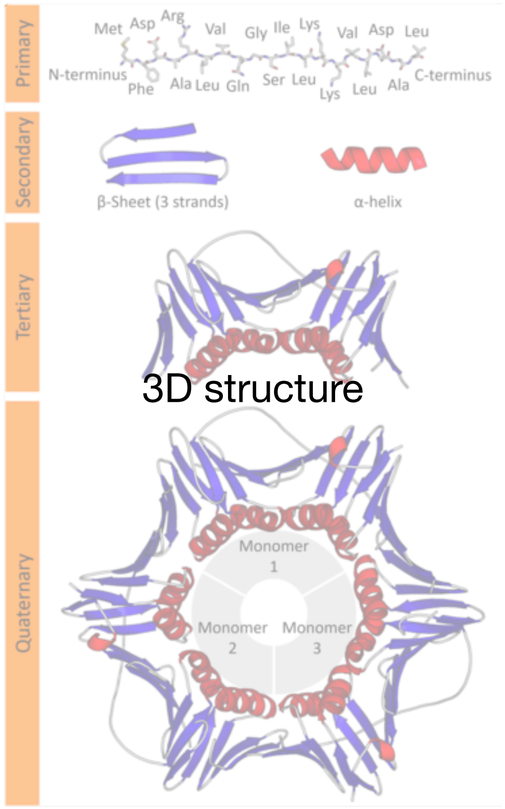
ACDEFGHIKLMNPQRSTVWY
vs
words and subwords

Multi-modal models



Protein sequences are not language

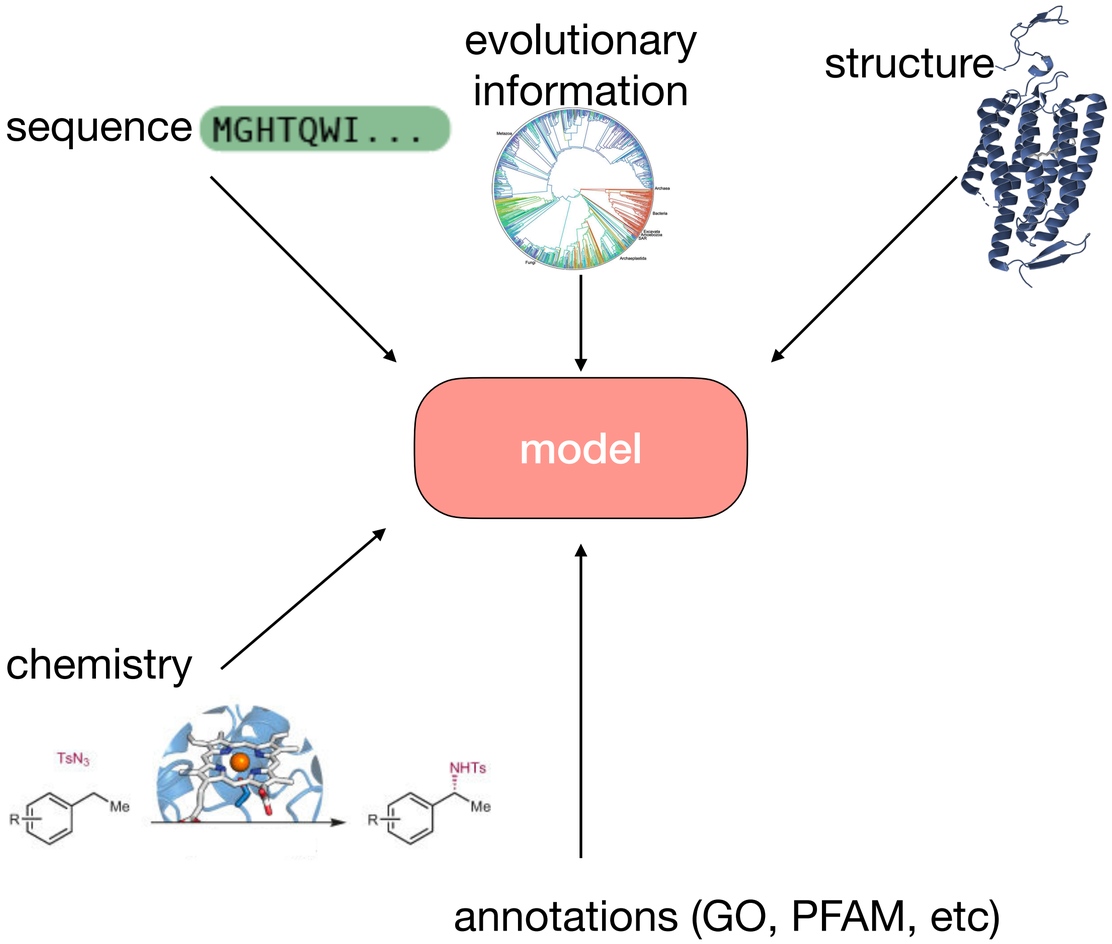
Differences between proteins and NL



Thomas Shafee

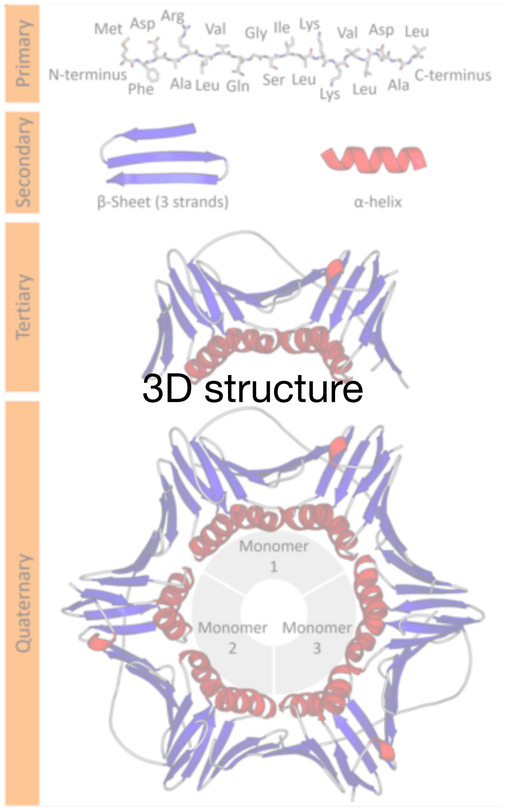
ACDEFGHIKLMNPQRSTVWY
vs
words and subwords

Multi-modal models



Protein sequences are not language

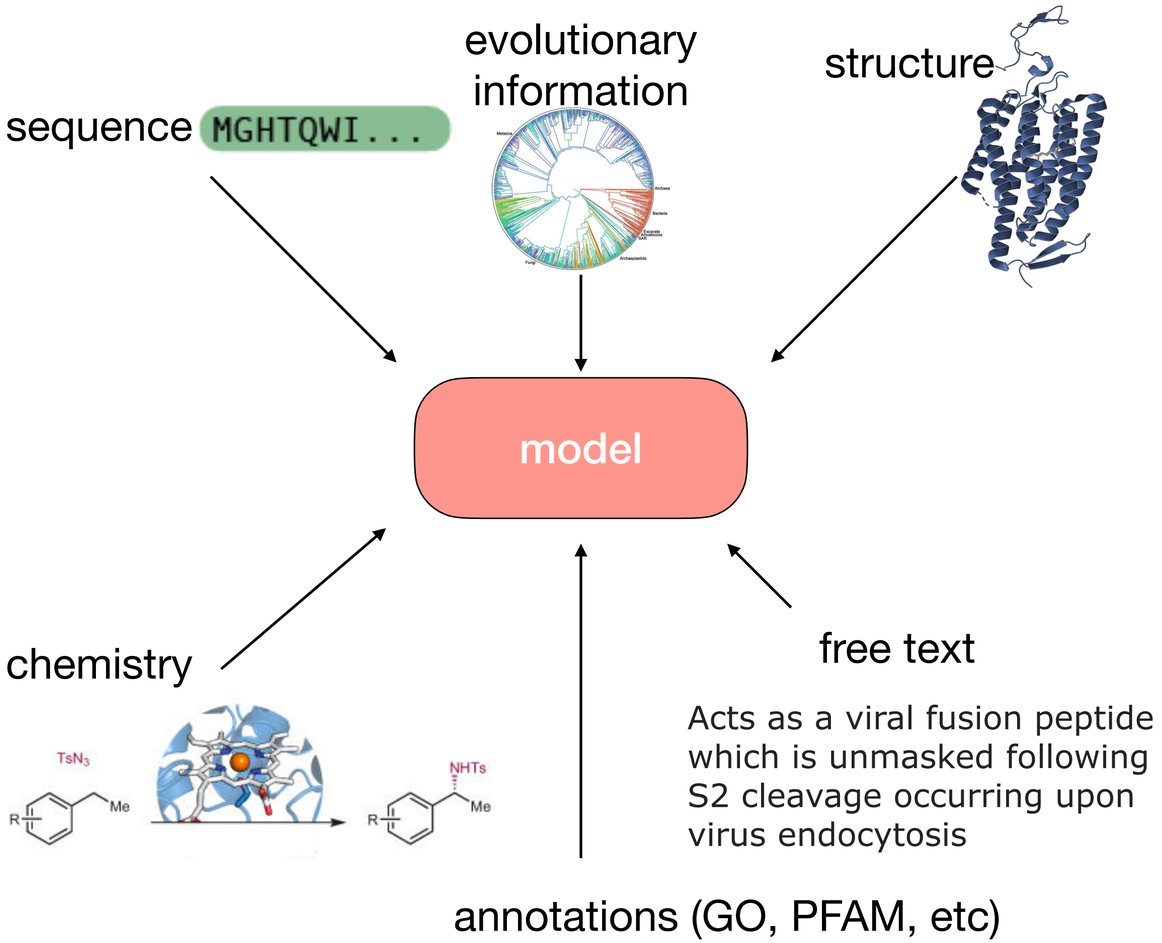
Differences between proteins and NL



Thomas Shafee

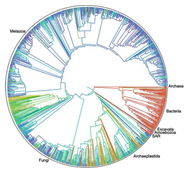
ACDEFGHIKLMNPQRSTVWY
vs
words and subwords

Multi-modal models



Multiple sequence alignments encode structural information

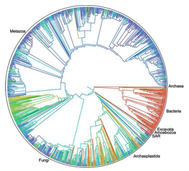
evolutionary information



Multiple sequence alignments encode structural information

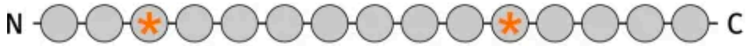
Alignments contain information about contacts

evolutionary information

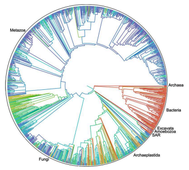


Multiple sequence alignments encode structural information

Alignments contain information about contacts

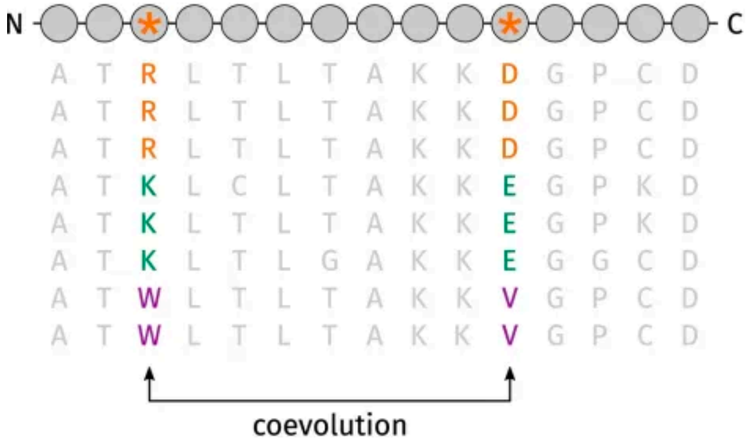


evolutionary information

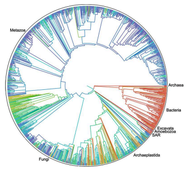


Multiple sequence alignments encode structural information

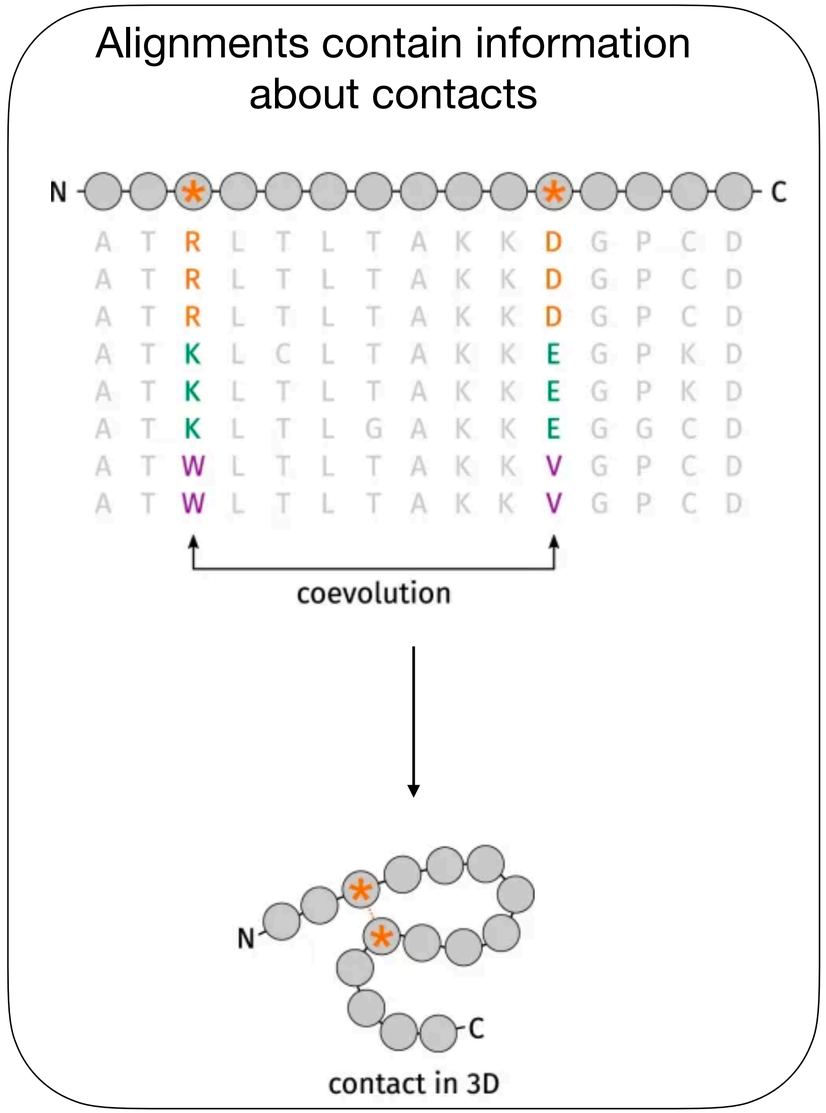
Alignments contain information about contacts



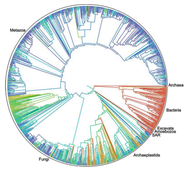
evolutionary information



Multiple sequence alignments encode structural information

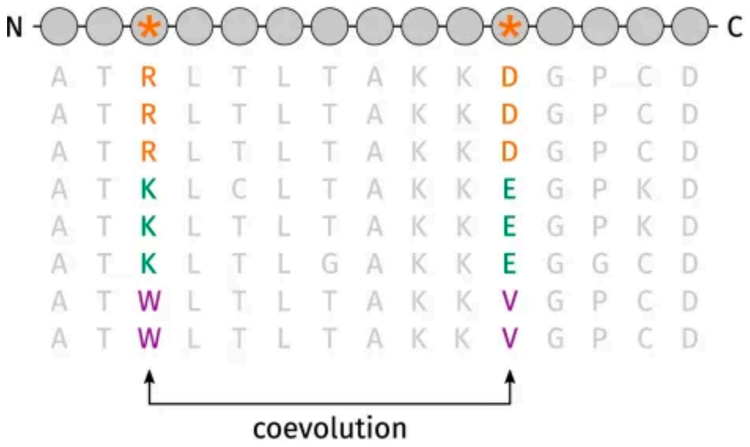


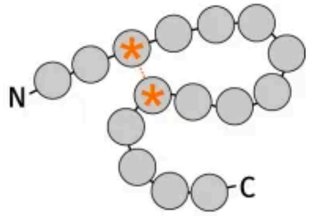
evolutionary information



Multiple sequence alignments encode structural information

Alignments contain information about contacts

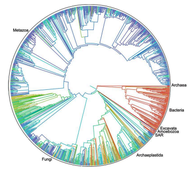




contact in 3D

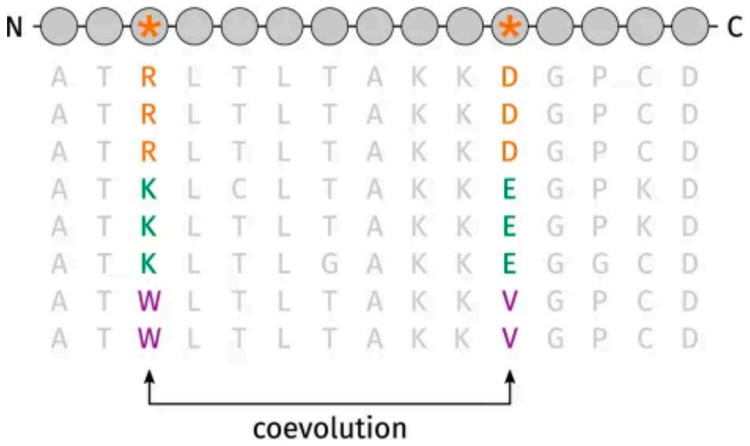
Contacts constrain fold

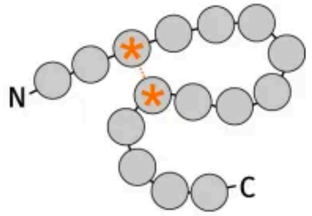
evolutionary information



Multiple sequence alignments encode structural information

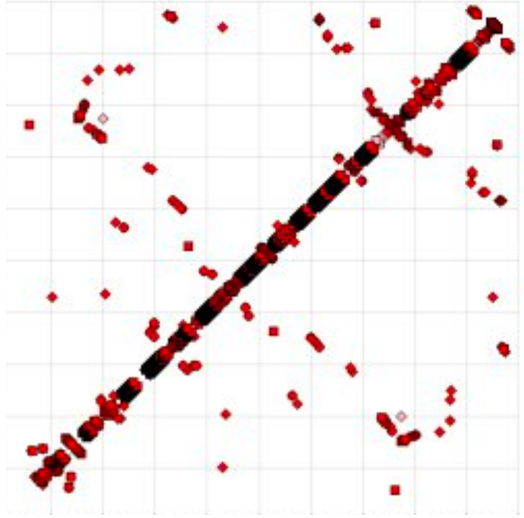
Alignments contain information about contacts



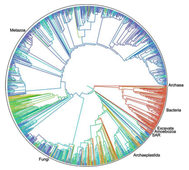


contact in 3D

Contacts constrain fold

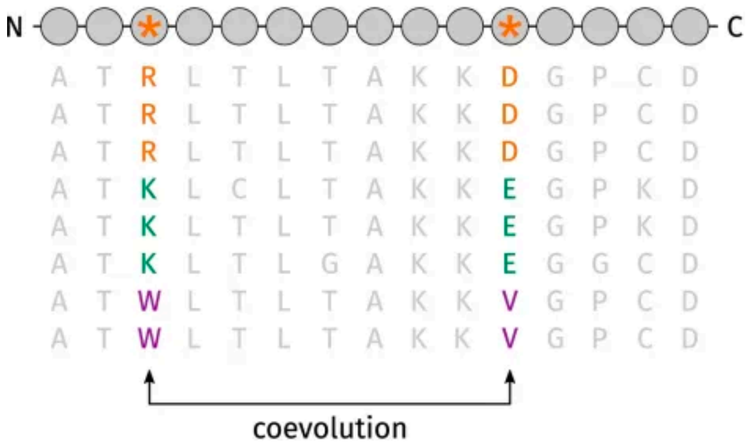


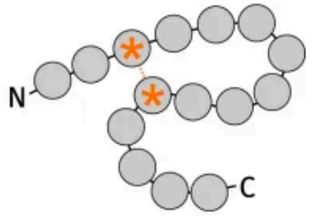
evolutionary information



Multiple sequence alignments encode structural information

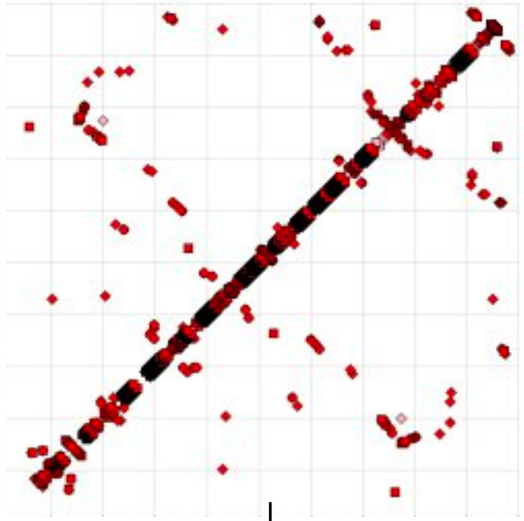
Alignments contain information about contacts

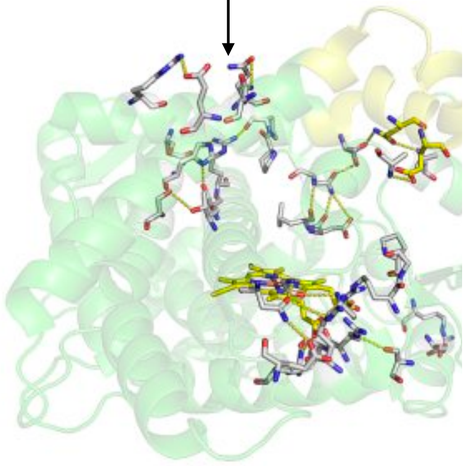




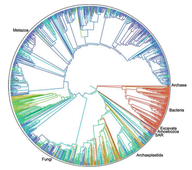
contact in 3D

Contacts constrain fold

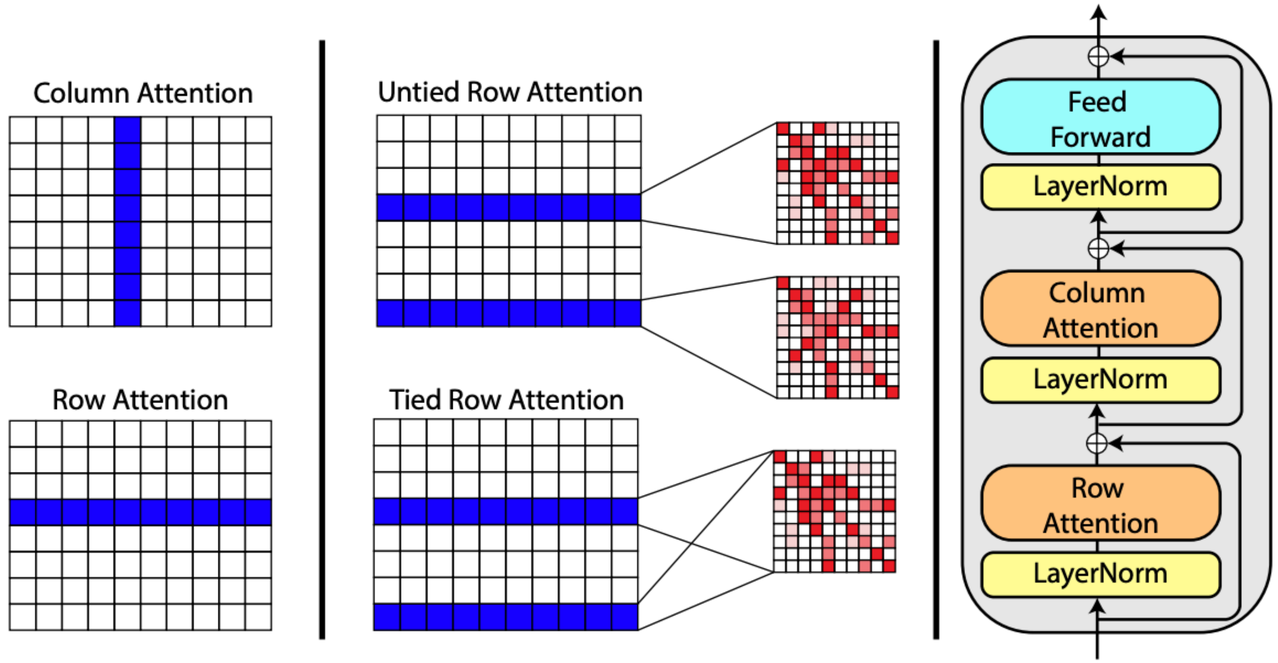




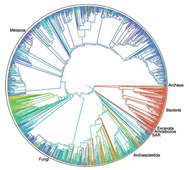
evolutionary information



MSA-Transformer extracts information from MSAs

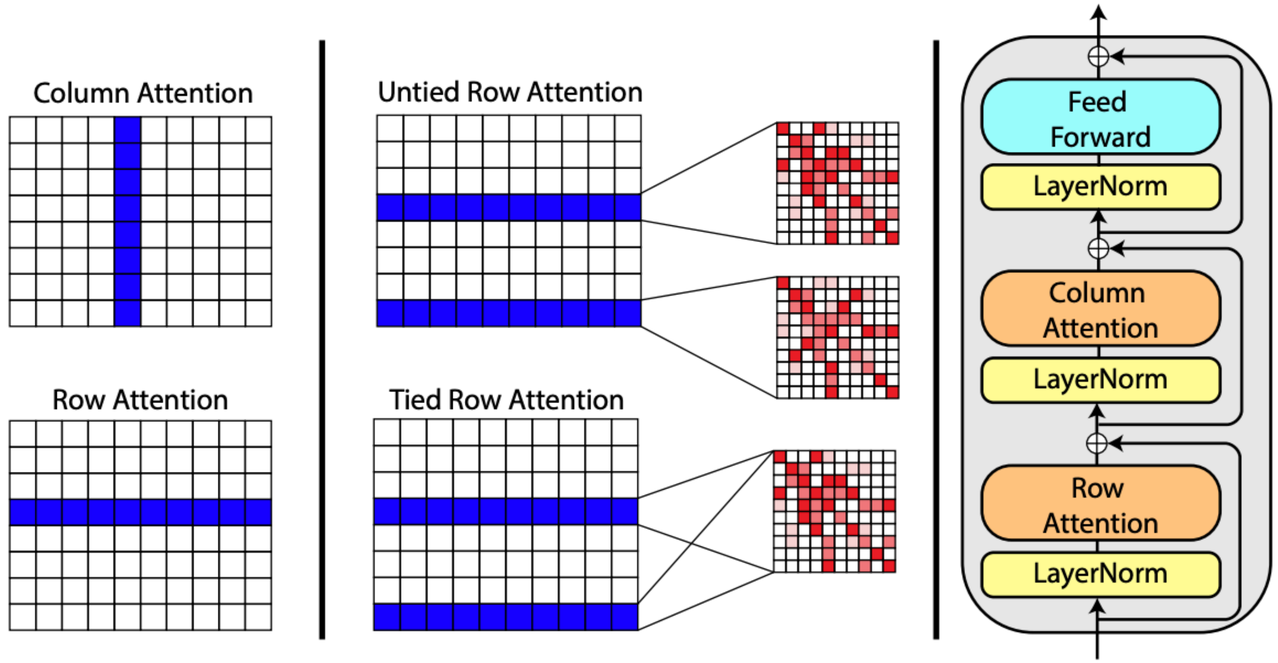


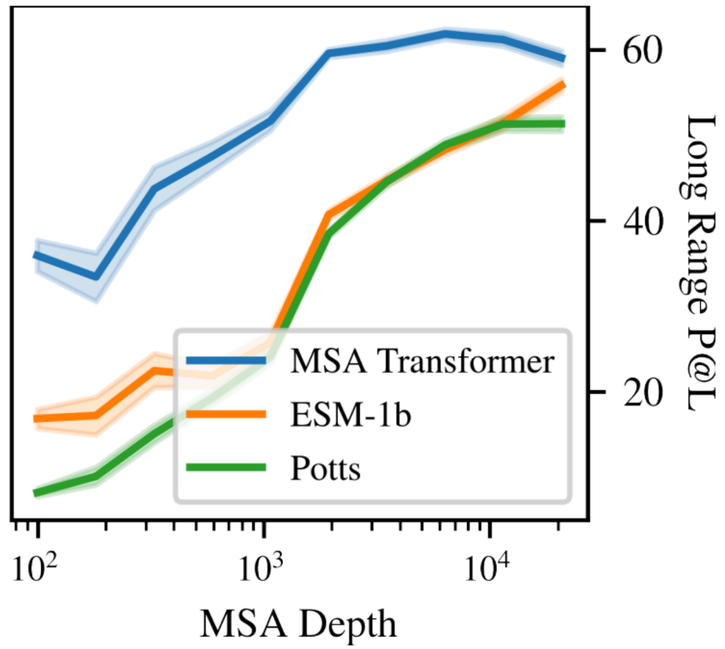
evolutionary information



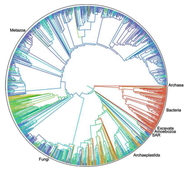
Rao *et al.* 2021

MSA-Transformer extracts information from MSAs



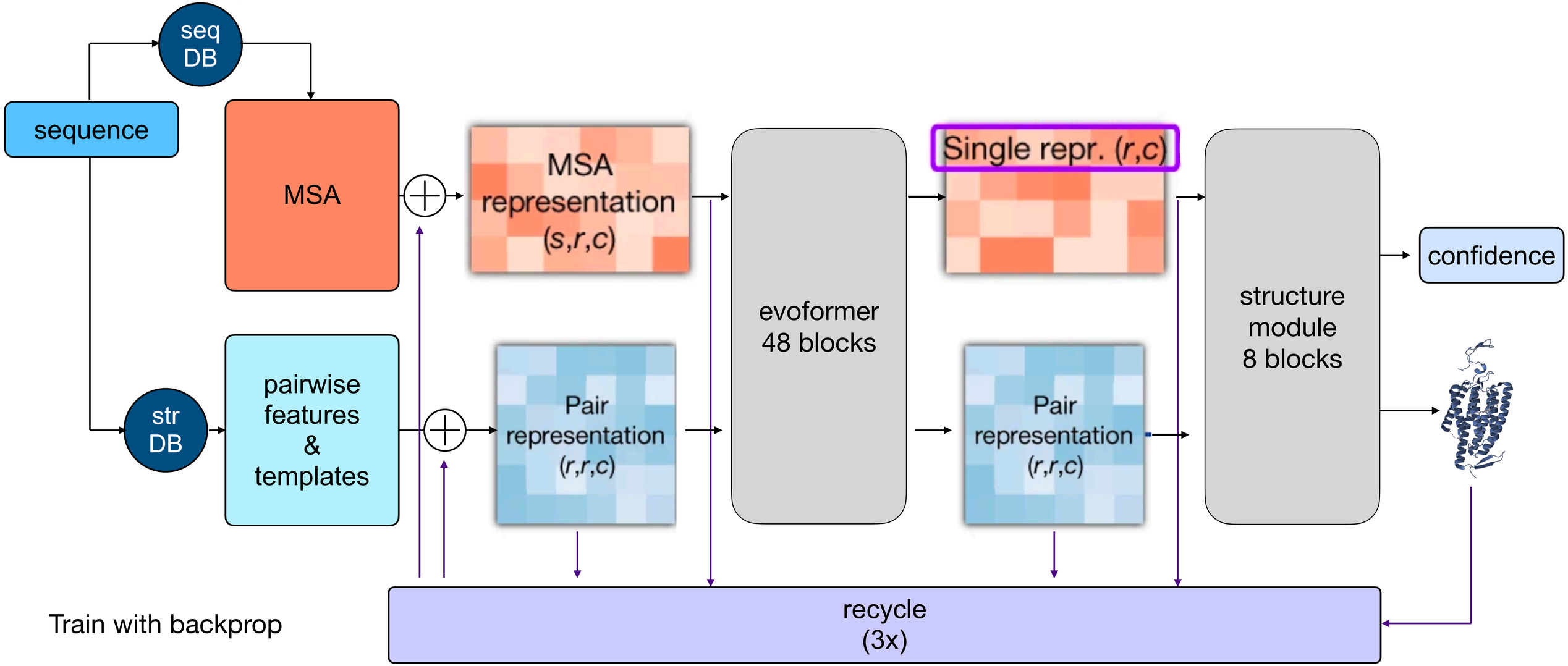


evolutionary information



Rao *et al.* 2021

AlphaFold uses MSAs and transformers

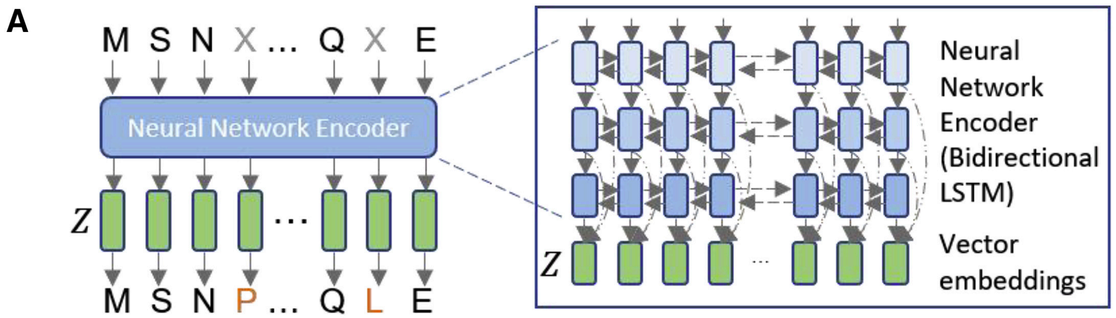


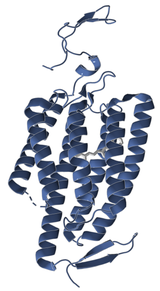
Train with backprop

16 TPUv3s for several weeks

Jumper *et al.* 2021

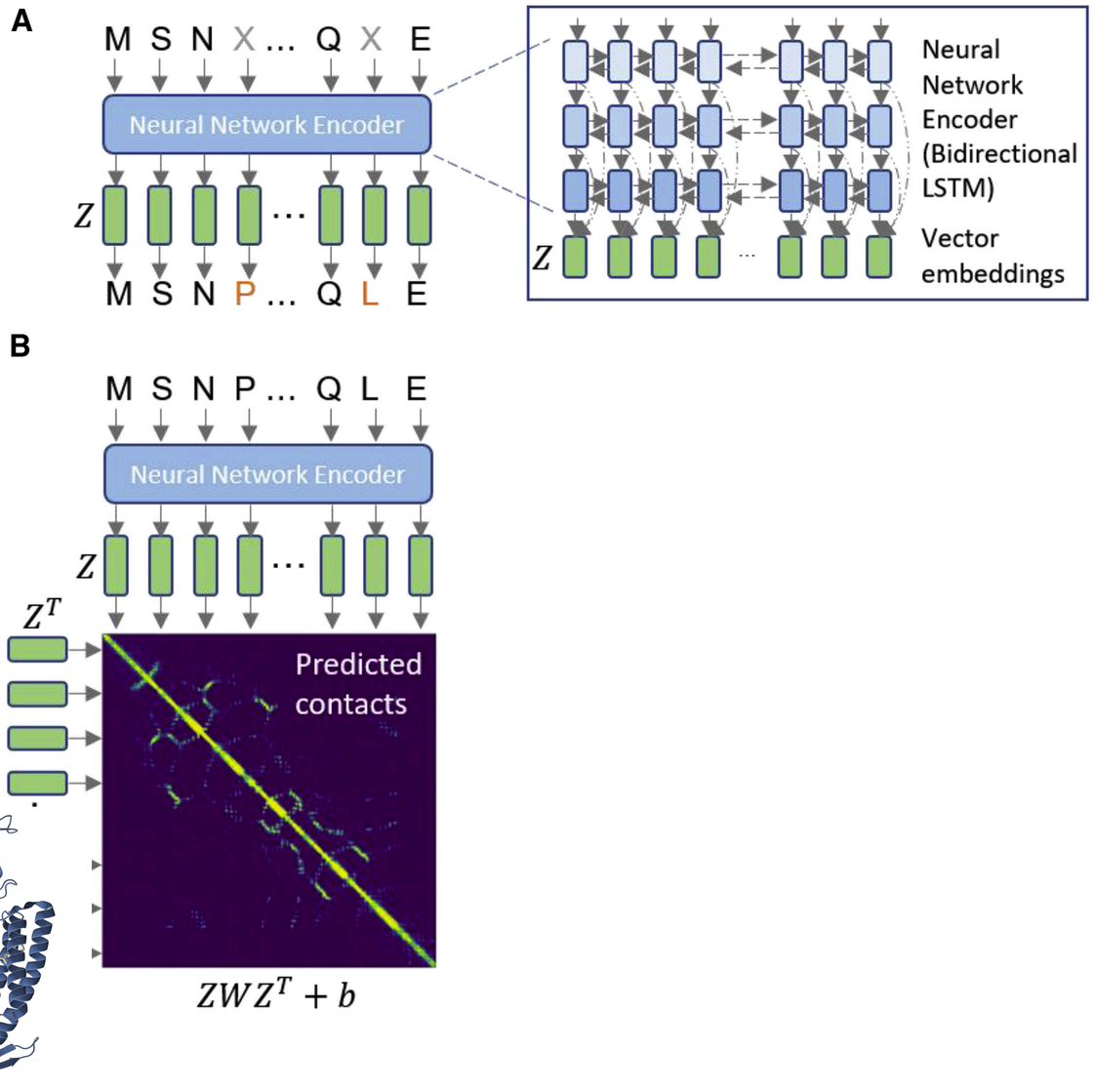
Can use structure as a pretext task





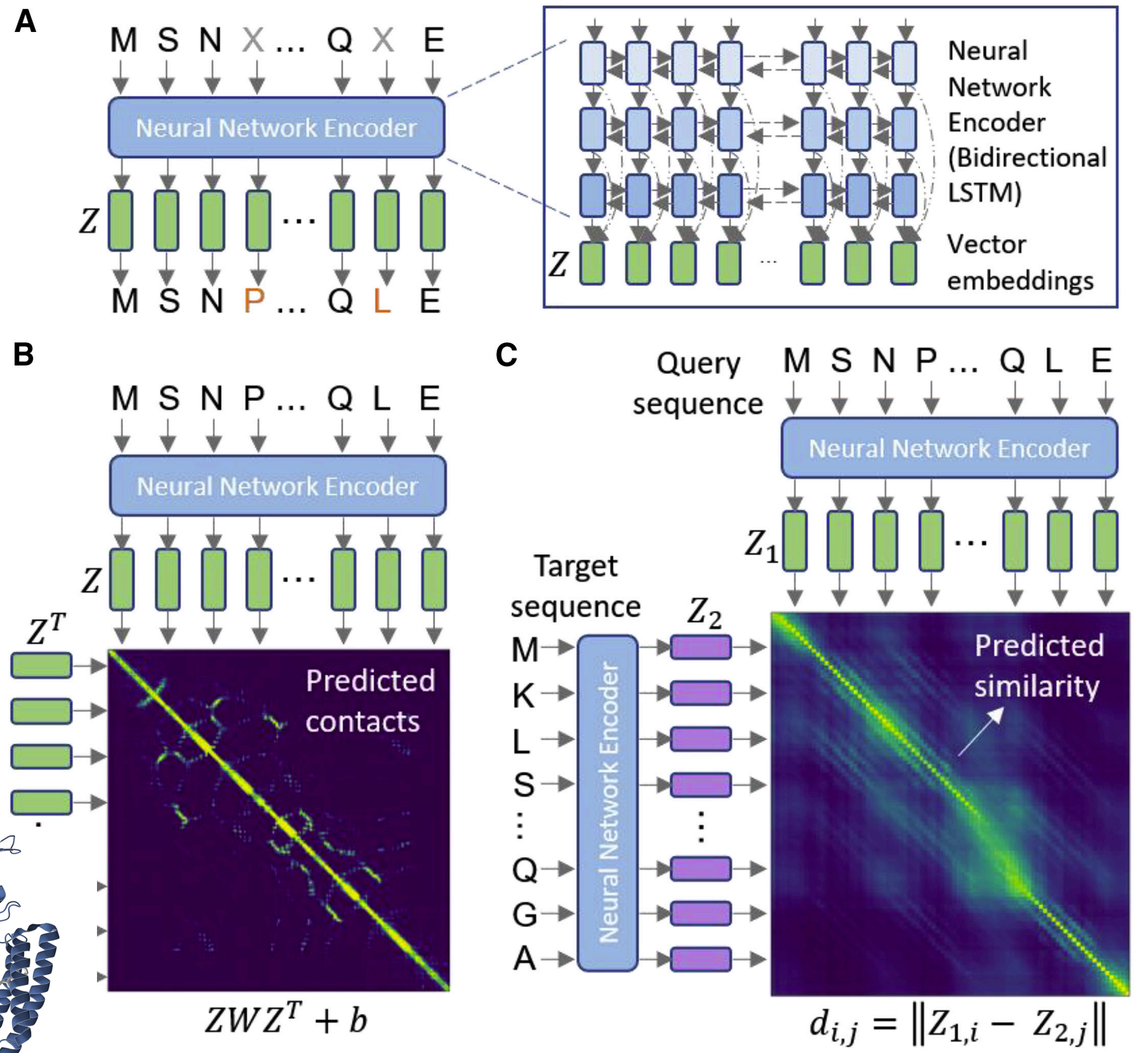
Bepler *et al. Cell Systems* 2021

Can use structure as a pretext task



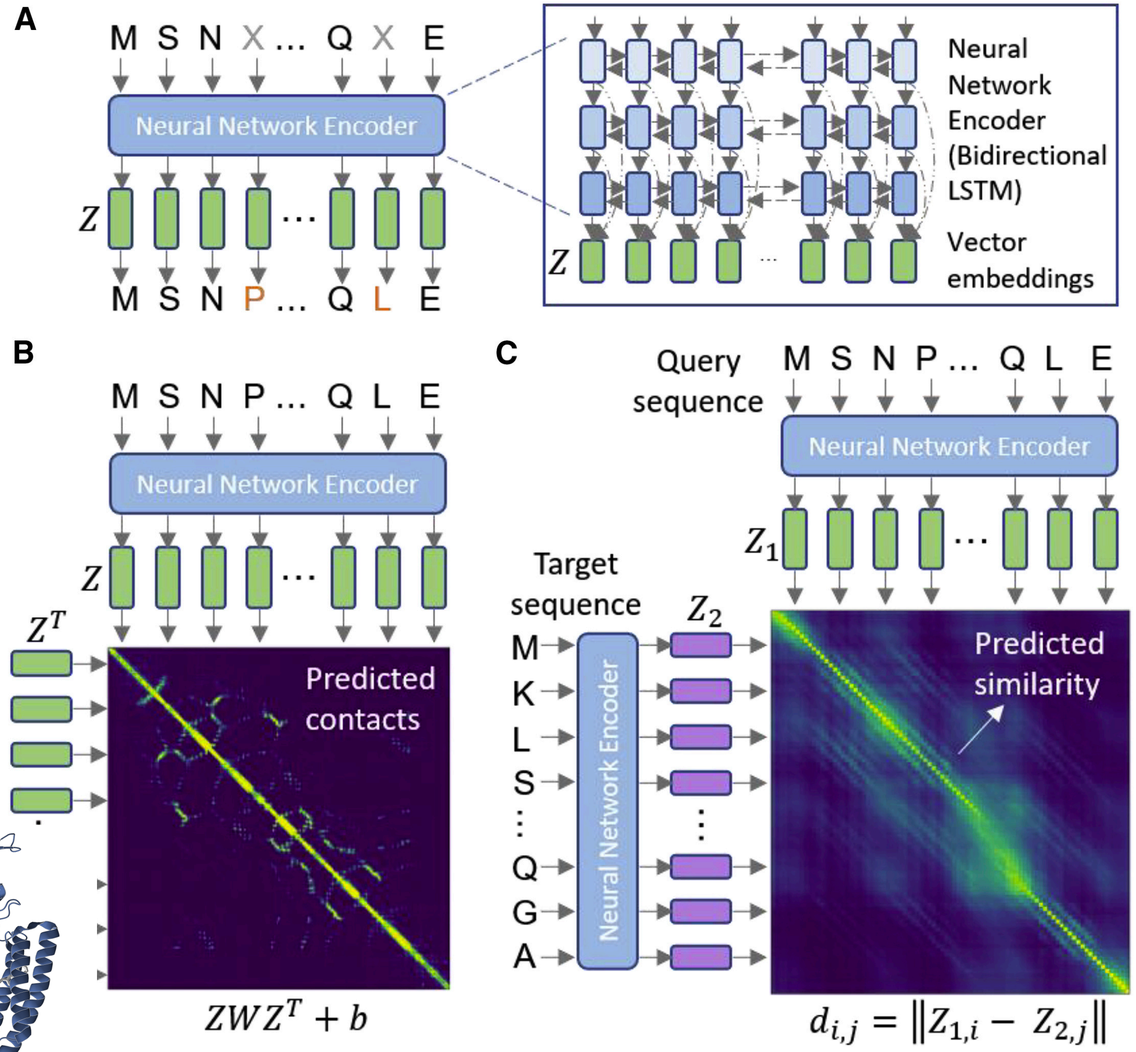
Bepler *et al. Cell Systems* 2021

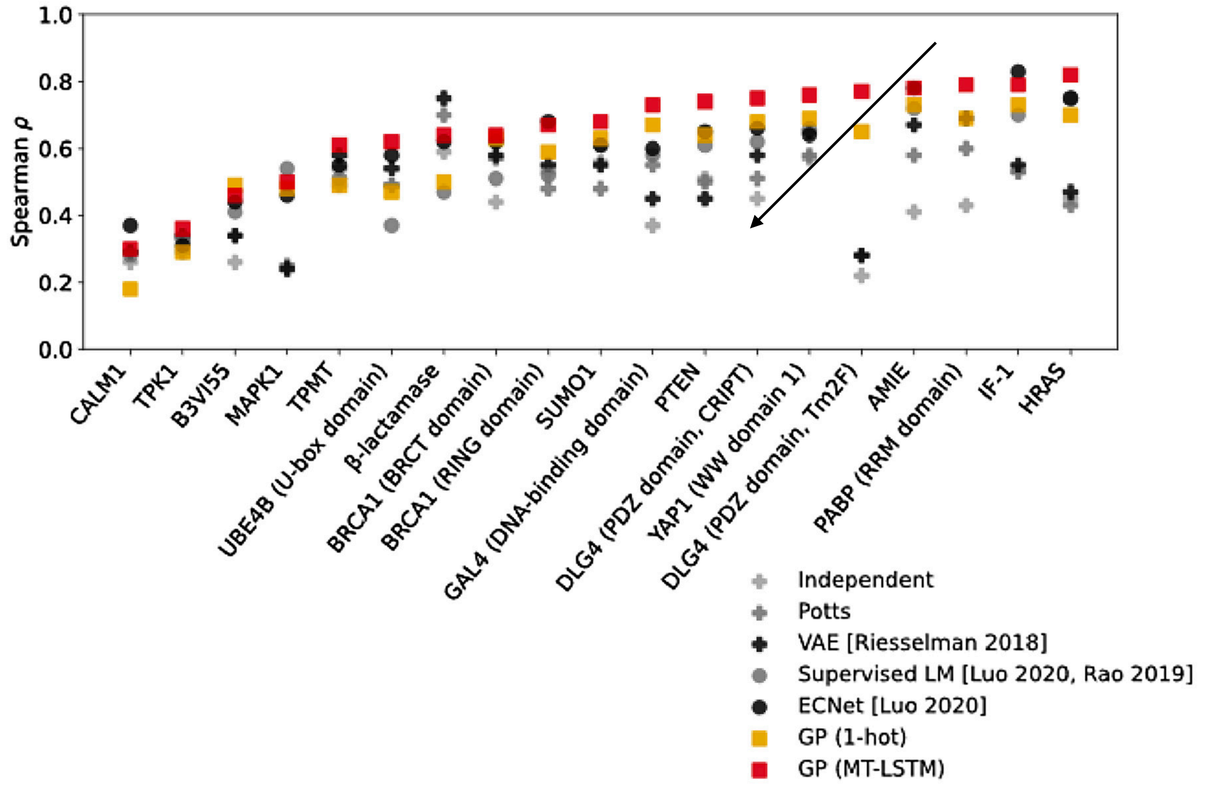
Can use structure as a pretext task



Bepler *et al. Cell Systems* 2021

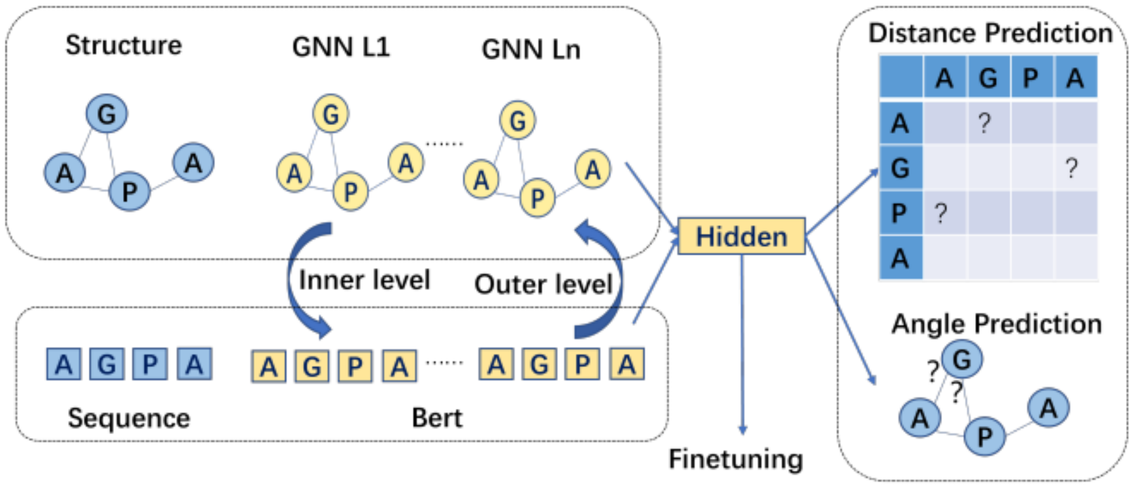
Can use structure as a pretext task





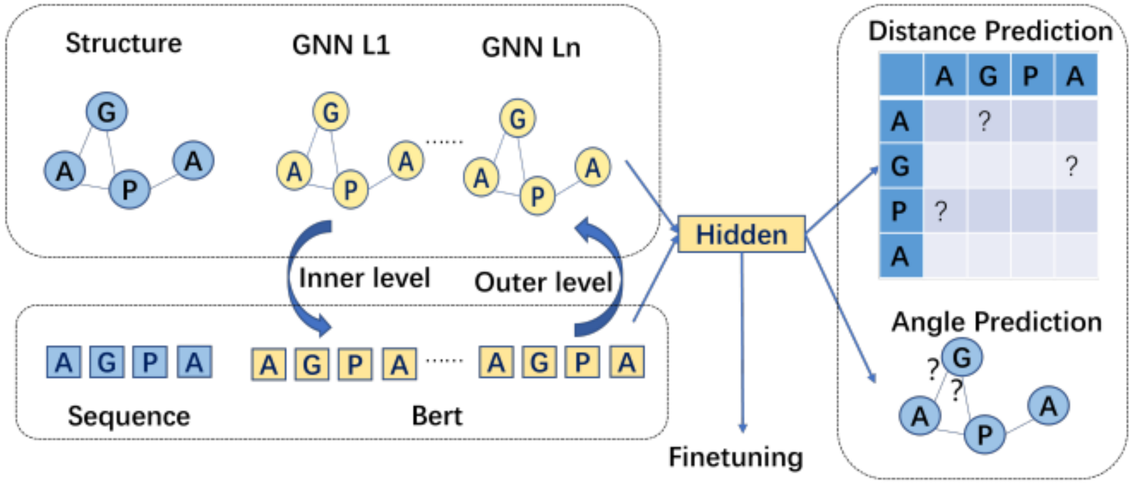
Bepler *et al. Cell Systems* 2021

Can use structure as a pretext task

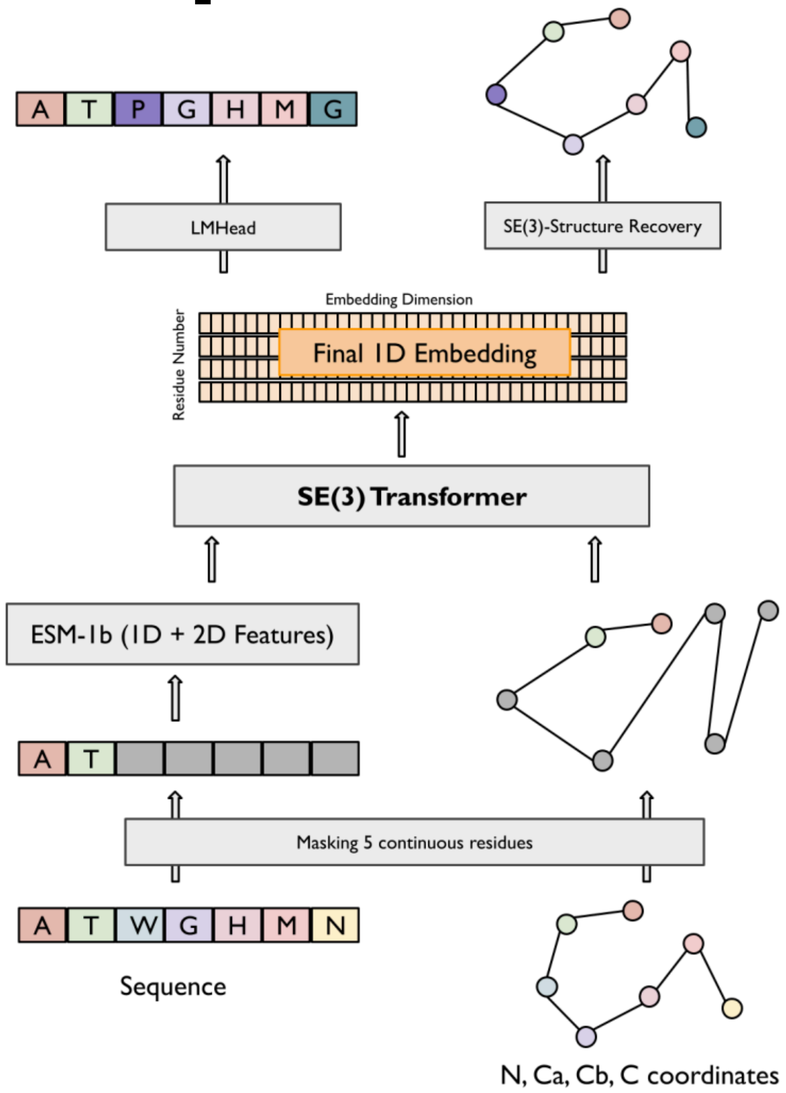


Chen *et al.* 2022

Can use structure as a pretext task



Chen *et al.* 2022



Mansoor *et al.* 2021

Conditioning on structure improves sequence reconstruction

Conditioning on structure improves sequence reconstruction

Structure-conditioned masked language model

Conditioning on structure improves sequence reconstruction

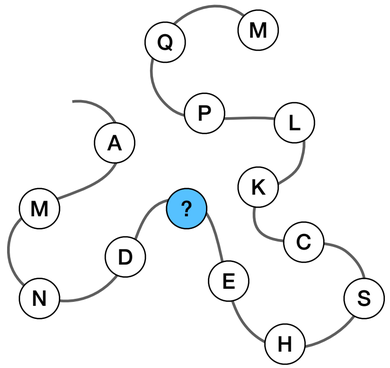
Structure-conditioned masked language model

Goal: Predict P(s_{masked} |structure, $s_{unmasked}$)

Conditioning on structure improves sequence reconstruction

Structure-conditioned masked language model

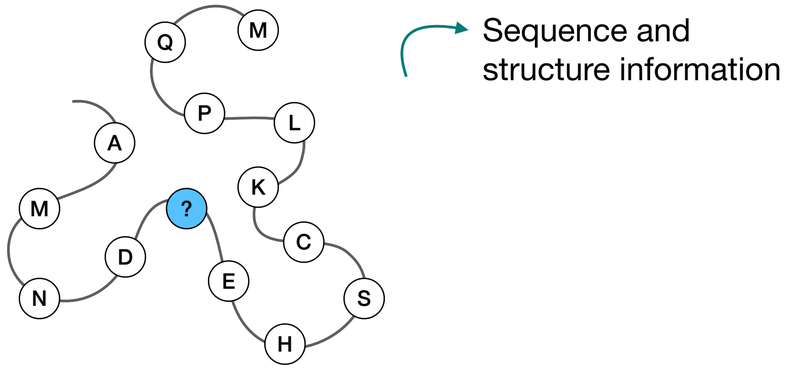
Goal: Predict P(s_{masked} |structure, $s_{unmasked}$)



Conditioning on structure improves sequence reconstruction

Structure-conditioned masked language model

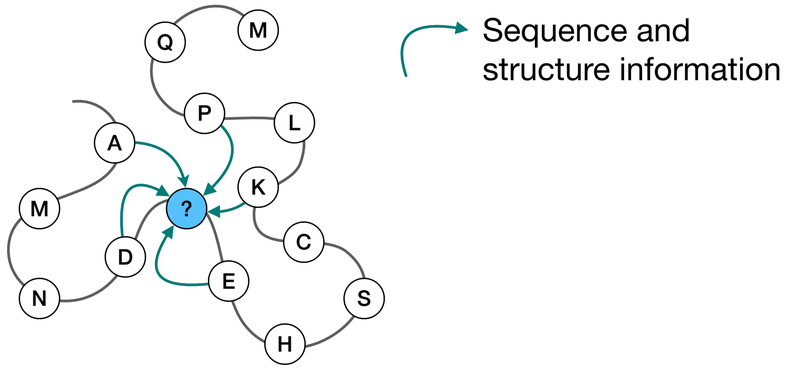
Goal: Predict P(s_{masked} |structure, $s_{unmasked}$)



Conditioning on structure improves sequence reconstruction

Structure-conditioned masked language model

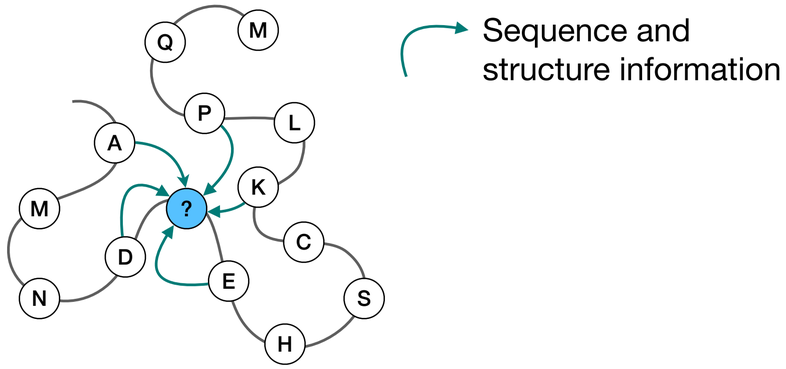
Goal: Predict P(s_{masked} |structure, $s_{unmasked}$)



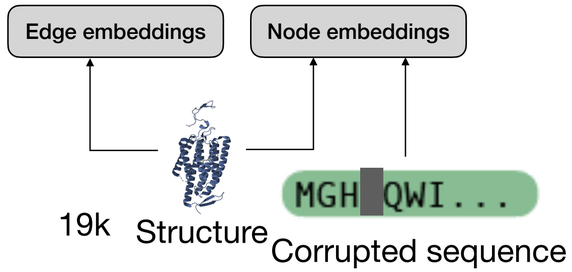
Conditioning on structure improves sequence reconstruction

Structure-conditioned masked language model

Goal: Predict P(s_{masked} |structure, $s_{unmasked}$)



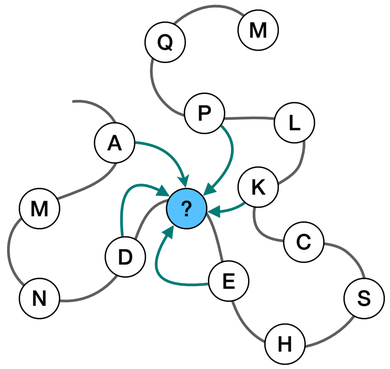
Sequence and
structure information



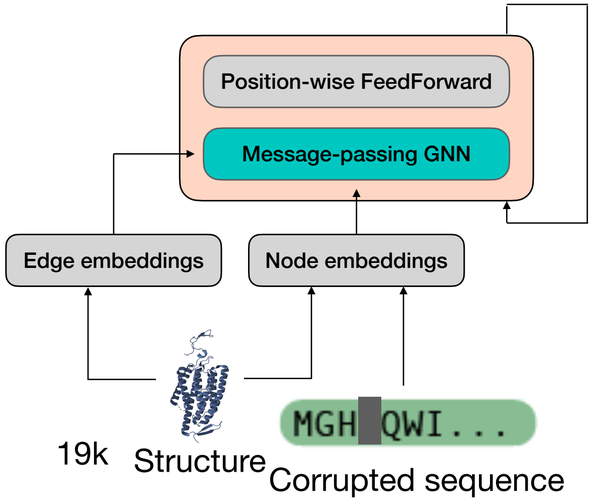
Conditioning on structure improves sequence reconstruction

Structure-conditioned masked language model

Goal: Predict P(s_{masked} |structure, $s_{unmasked}$)



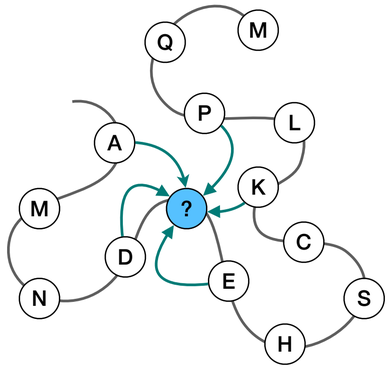
Sequence and structure information



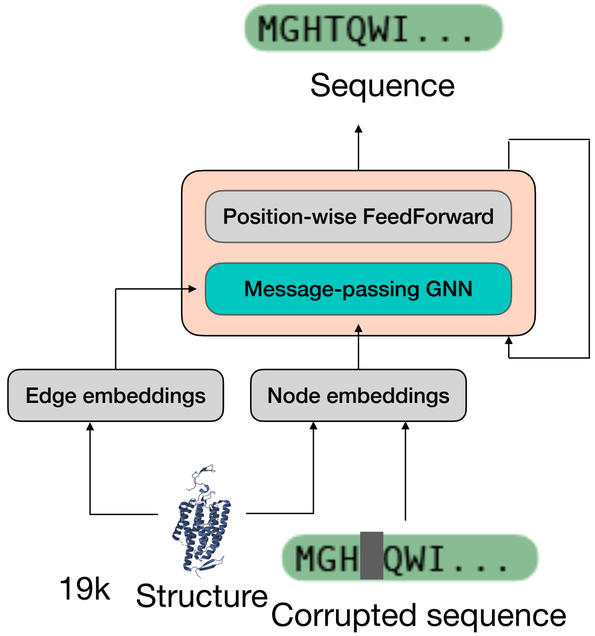
Conditioning on structure improves sequence reconstruction

Structure-conditioned masked language model

Goal: Predict P(s_{masked} |structure, $s_{unmasked}$)



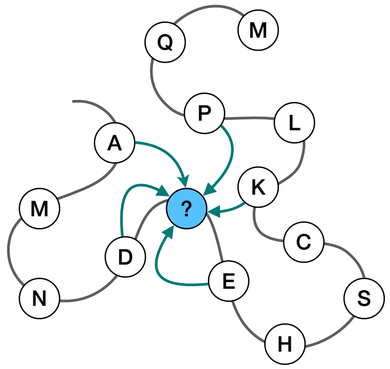
Sequence and structure information



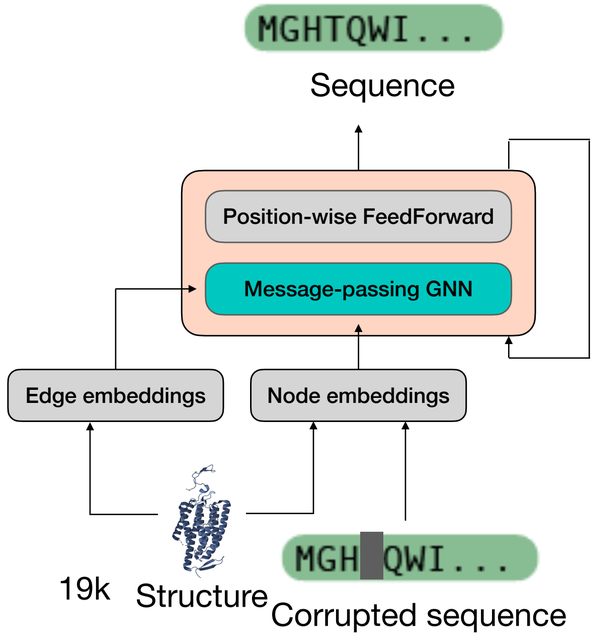
Conditioning on structure improves sequence reconstruction

Structure-conditioned masked language model

Goal: Predict P(s_{masked} |structure, $s_{unmasked}$)



Sequence and structure information

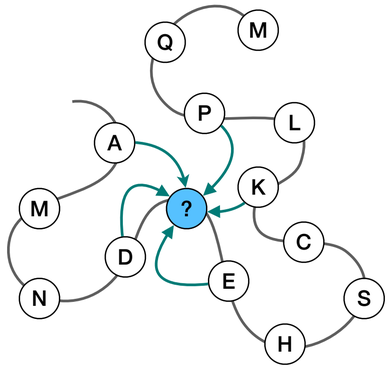


Model	Perplexity
GNN	4.96

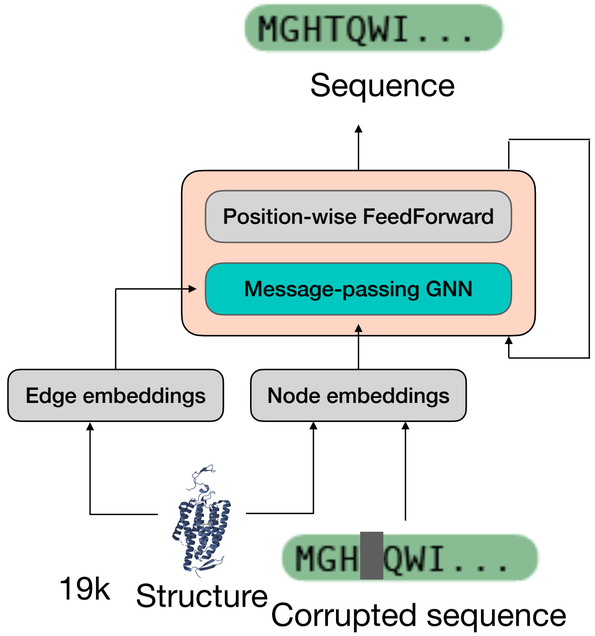
Conditioning on structure improves sequence reconstruction

Structure-conditioned masked language model

Goal: Predict P(s_{masked} |structure, $s_{unmasked}$)

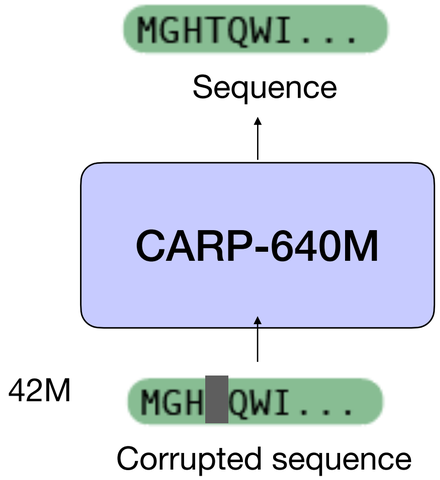


Sequence and structure information



Model	Perplexity
GNN	4.96

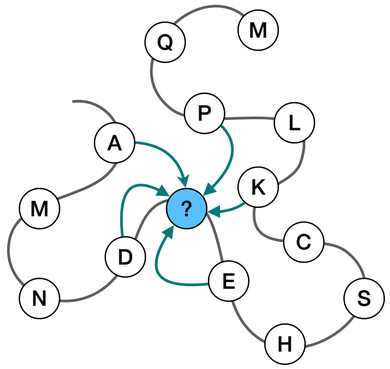
Masked language model



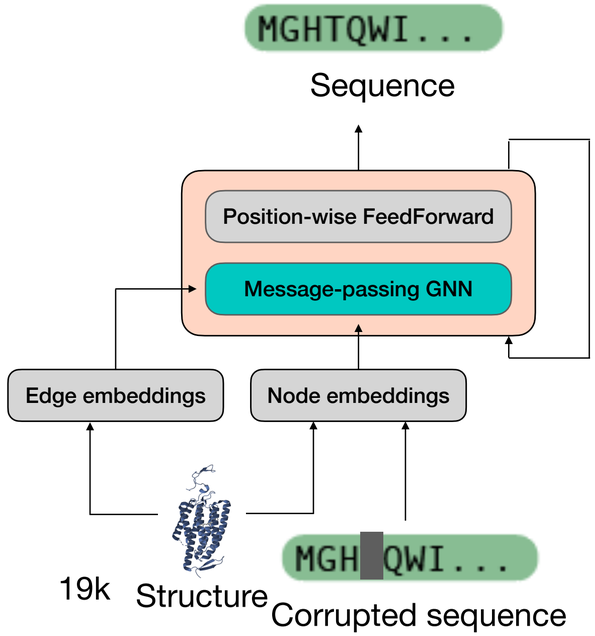
Conditioning on structure improves sequence reconstruction

Structure-conditioned masked language model

Goal: Predict P(s_{masked} |structure, $s_{unmasked}$)

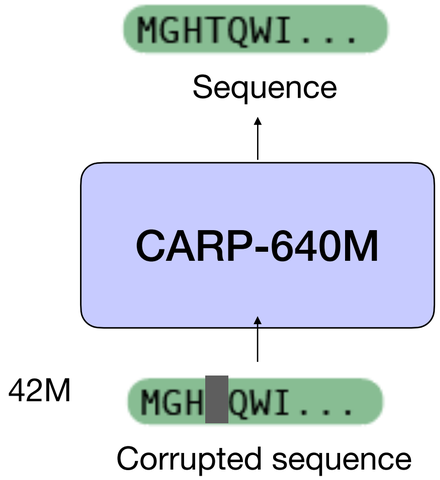


Sequence and structure information



Model	Perplexity
GNN	4.96
Sequence only	7.06

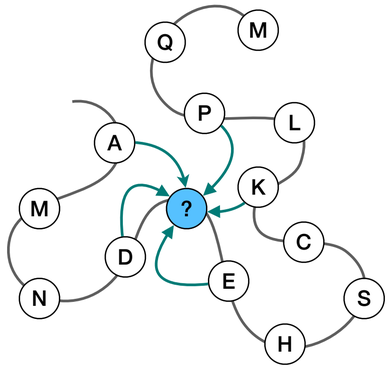
Masked language model



Conditioning on structure improves sequence reconstruction

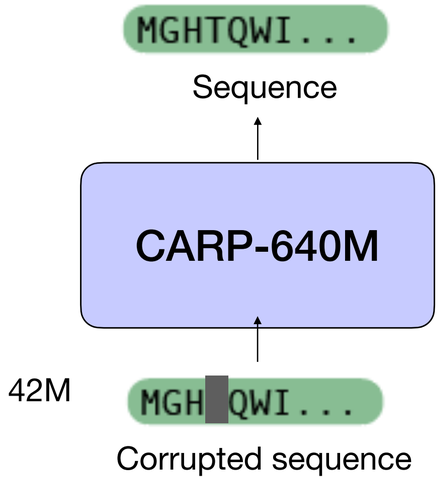
Structure-conditioned masked language model

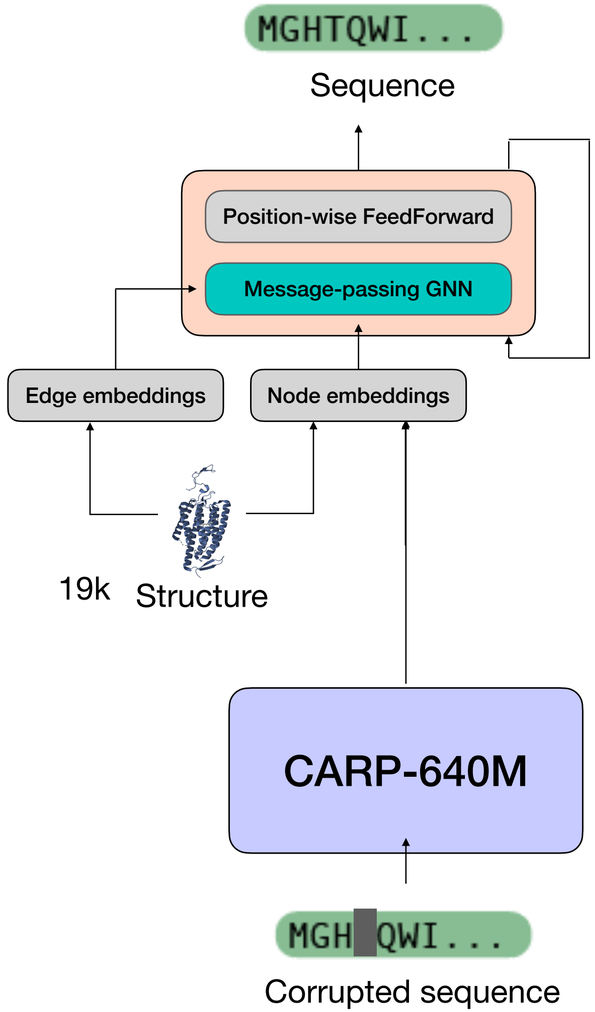
Goal: Predict P(s_{masked} |structure, $s_{unmasked}$)



Sequence and structure information

Masked language model



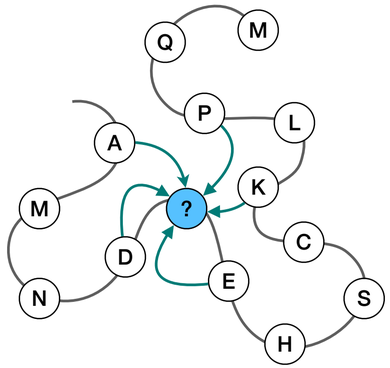


Model	Perplexity
GNN	4.96
Sequence only	7.06

Conditioning on structure improves sequence reconstruction

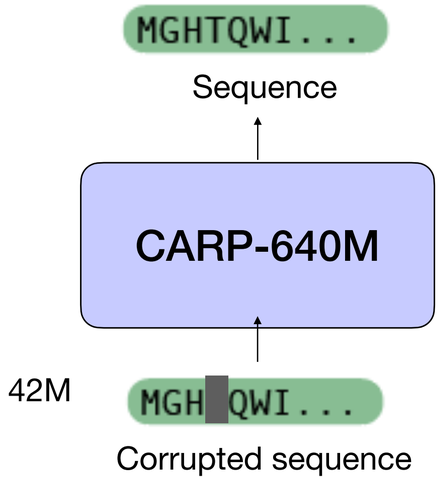
Structure-conditioned masked language model

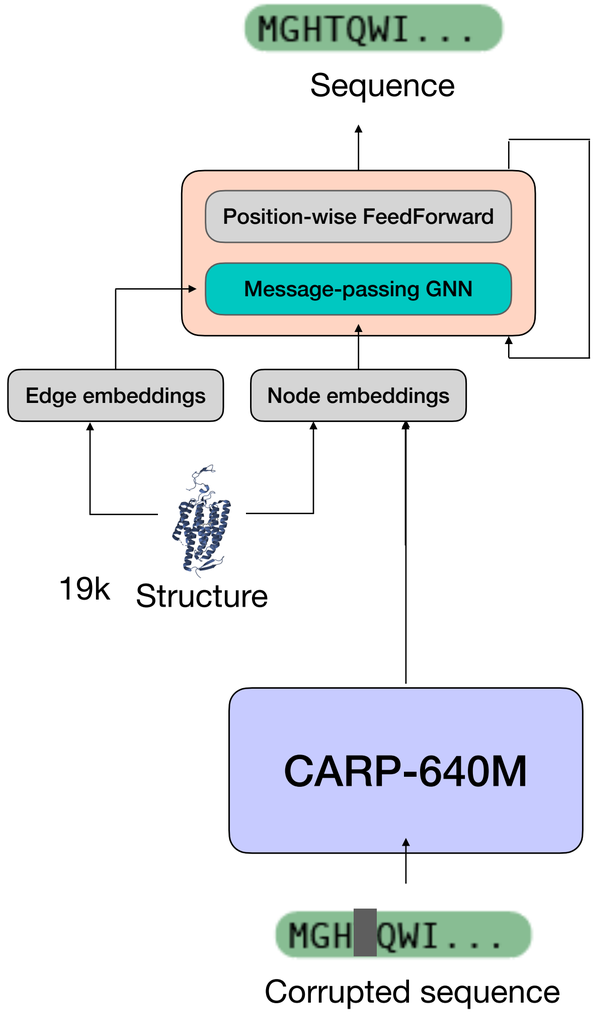
Goal: Predict P(s_{masked} |structure, $s_{unmasked}$)



Sequence and structure information

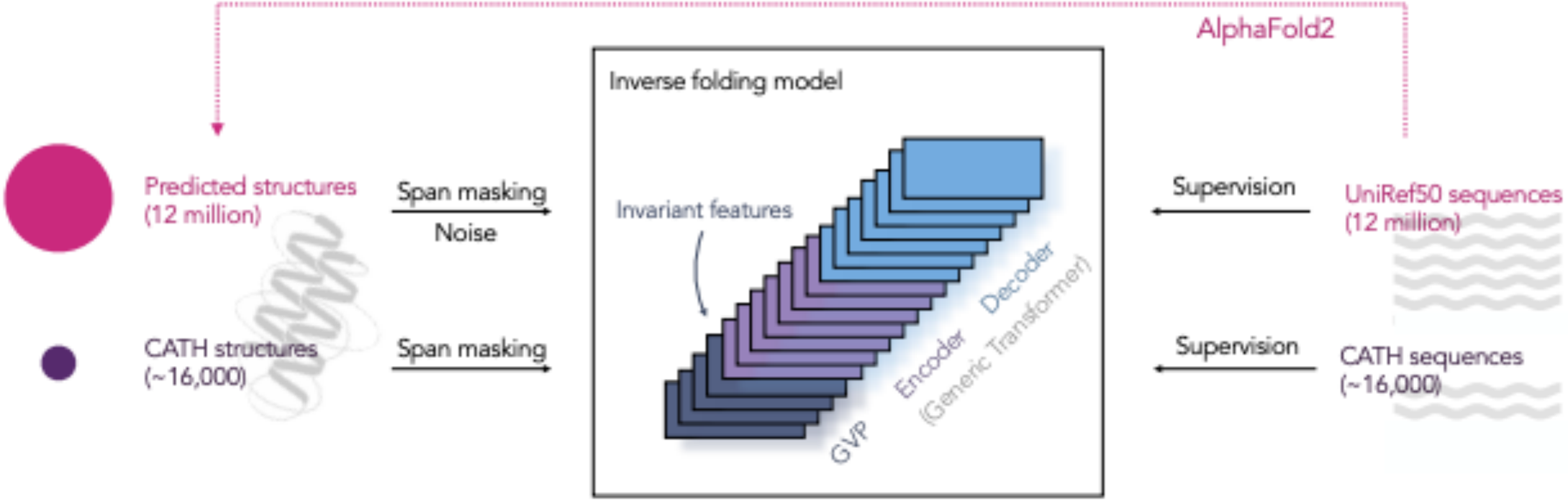
Masked language model





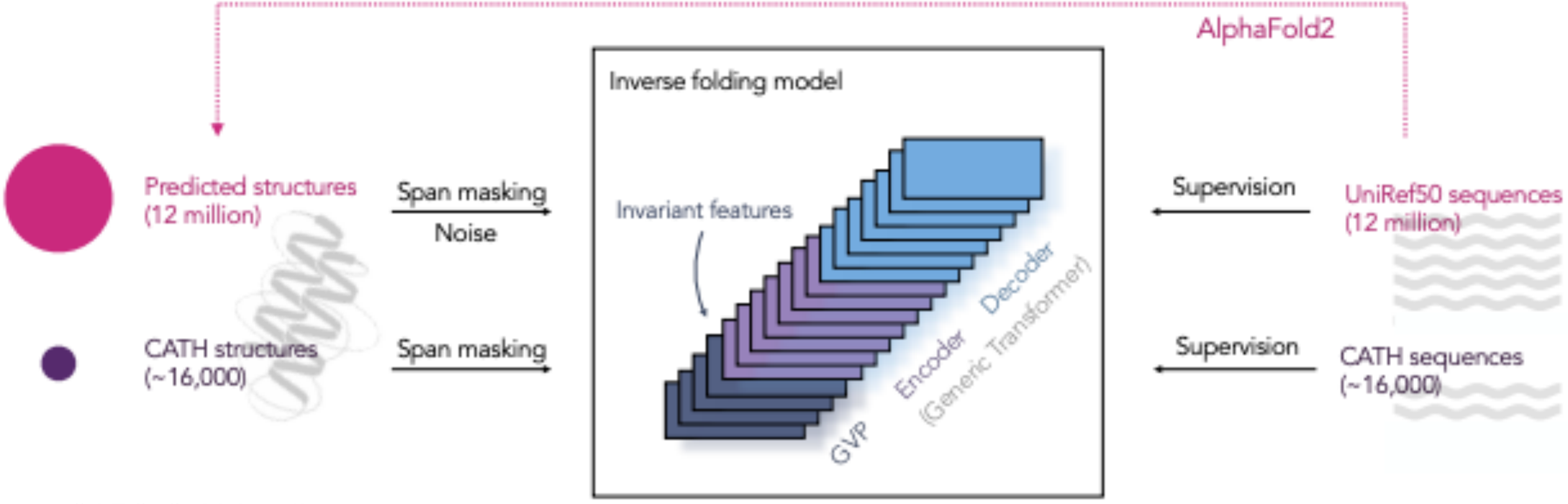
Model	Perplexity
GNN	4.96
Sequence only	7.06
Pretrained logits -> GNN	4.08

Adding predicted structures improves sequence reconstruction

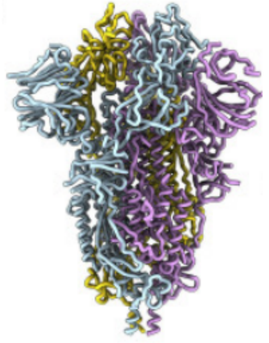


Hsu *et al.* 2022

Adding predicted structures improves sequence reconstruction

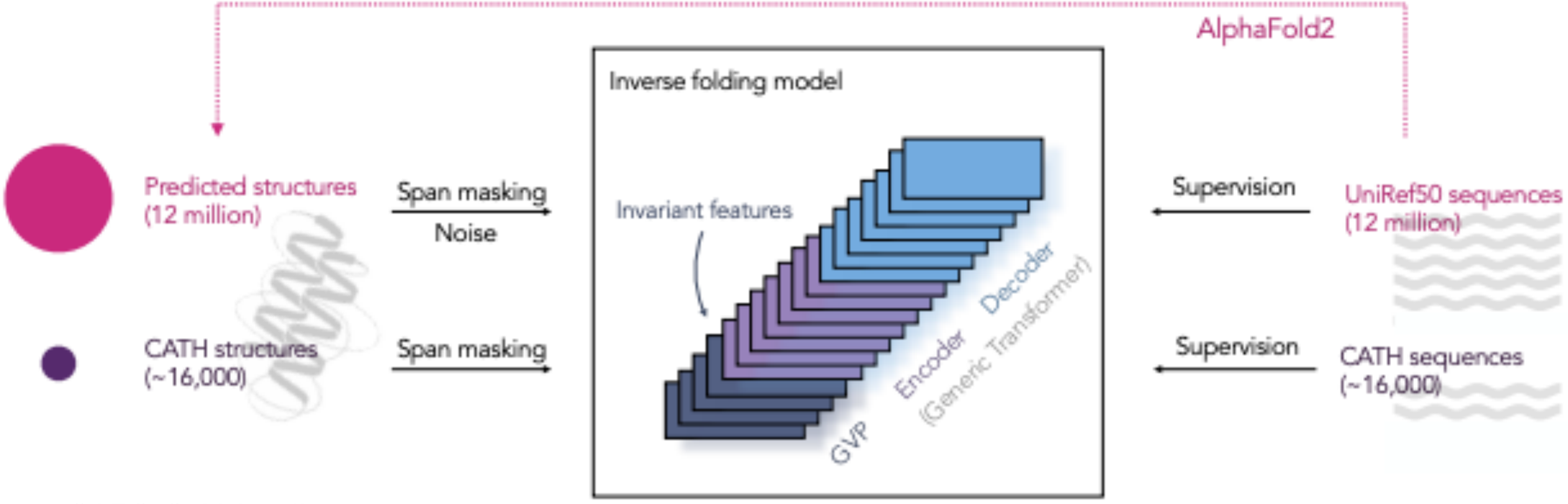


SARS-CoV-2 spike receptor-binding domain

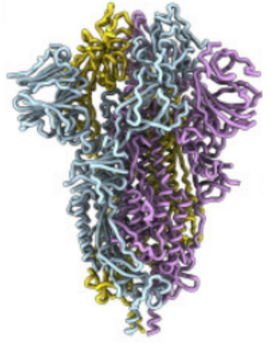


Hsu *et al.* 2022

Adding predicted structures improves sequence reconstruction



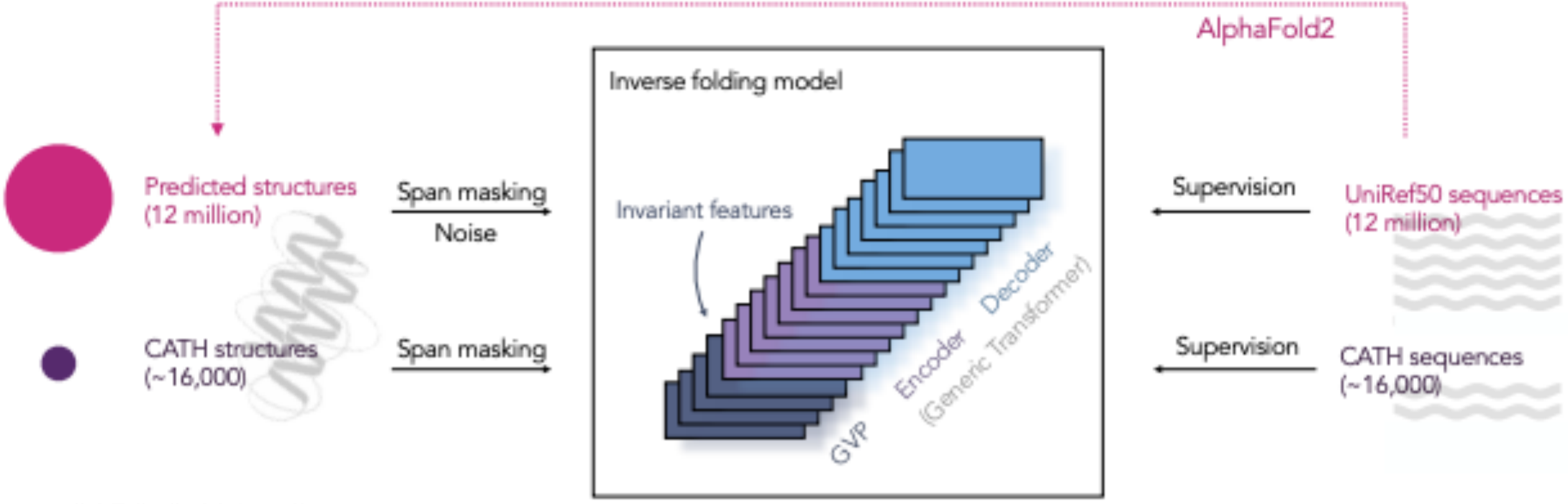
SARS-CoV-2 spike receptor-binding domain



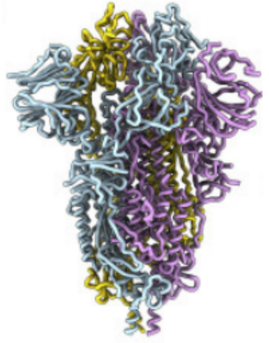
	Spearman correlation			
Model	No coords	No RBM coords	No ACE2 coords	All coords
ESM-1v	0.03			
ESM-1b	0.02			
ESM-MSA-1b (few-shot)	**0.51**			

Hsu *et al.* 2022

Adding predicted structures improves sequence reconstruction



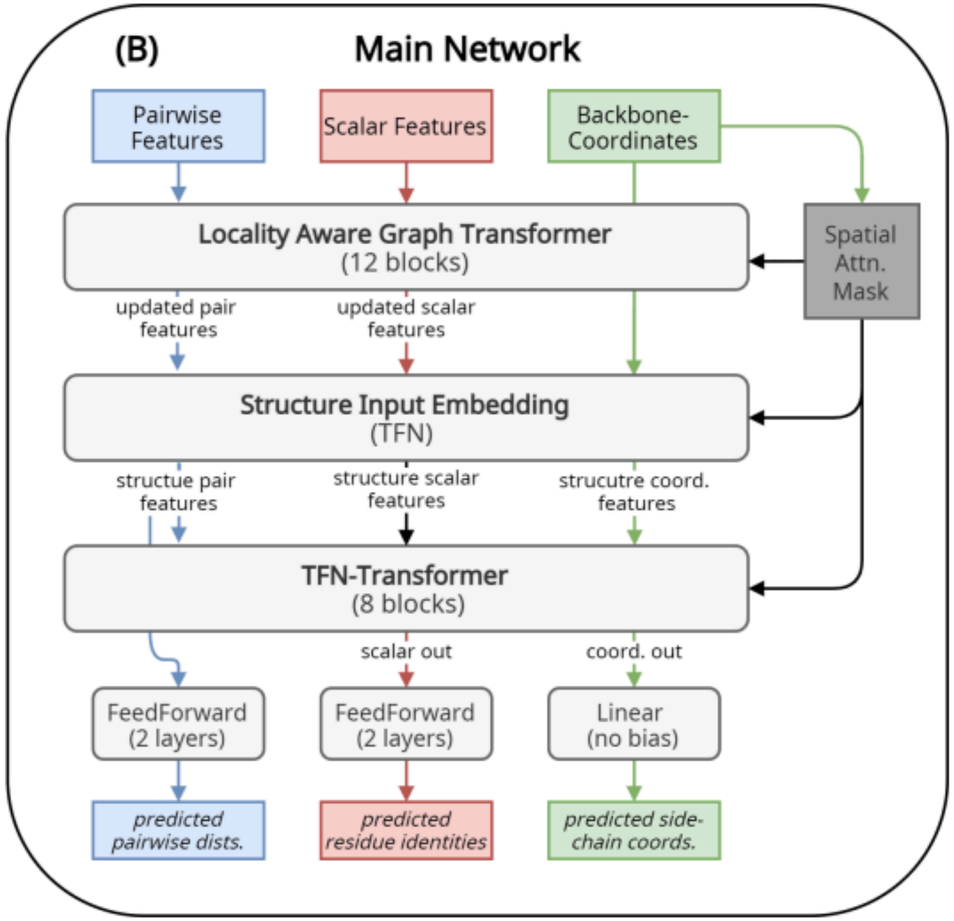
SARS-CoV-2 spike receptor-binding domain



Model	Spearman correlation			
	No coords	No RBM coords	No ACE2 coords	All coords
ESM-1v	0.03			
ESM-1b	0.02			
ESM-MSA-1b (few-shot)	**0.51**			
GVP-GNN		-0.10	0.50	0.60
GVP-GNN-large+AF2		-0.05	0.52	**0.69**
GVP-Transformer+AF2		-0.06	**0.53**	0.64

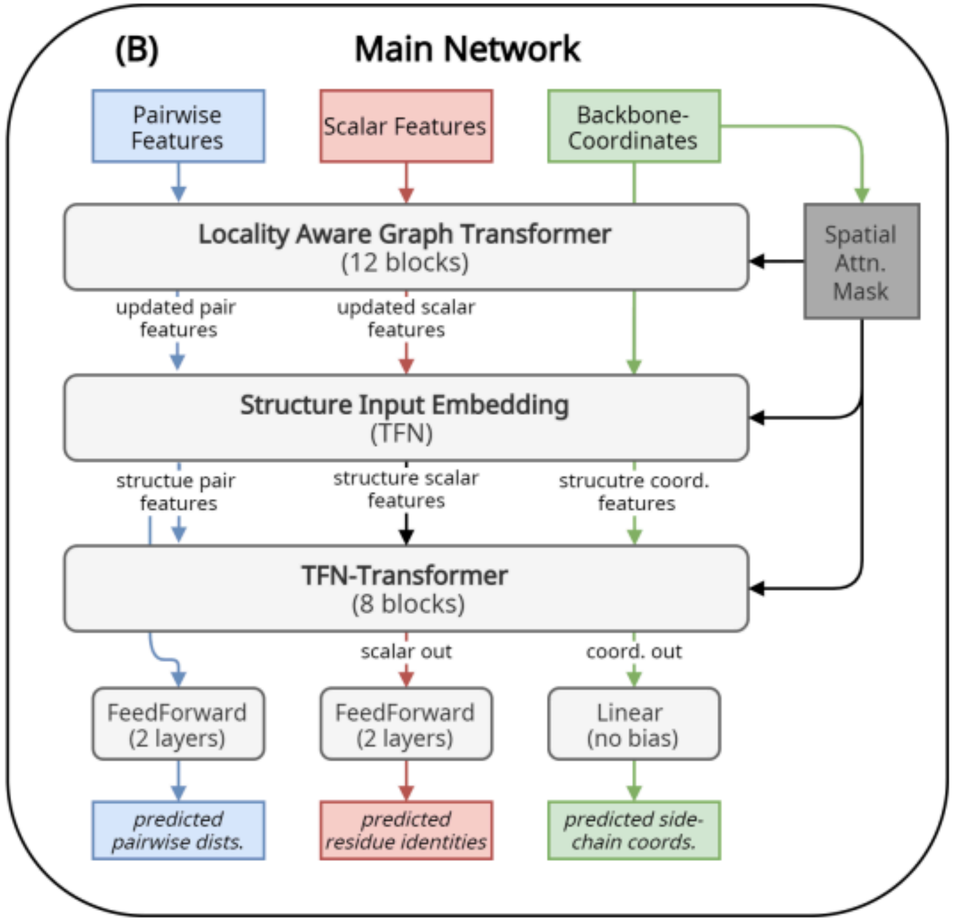
Hsu *et al.* 2022

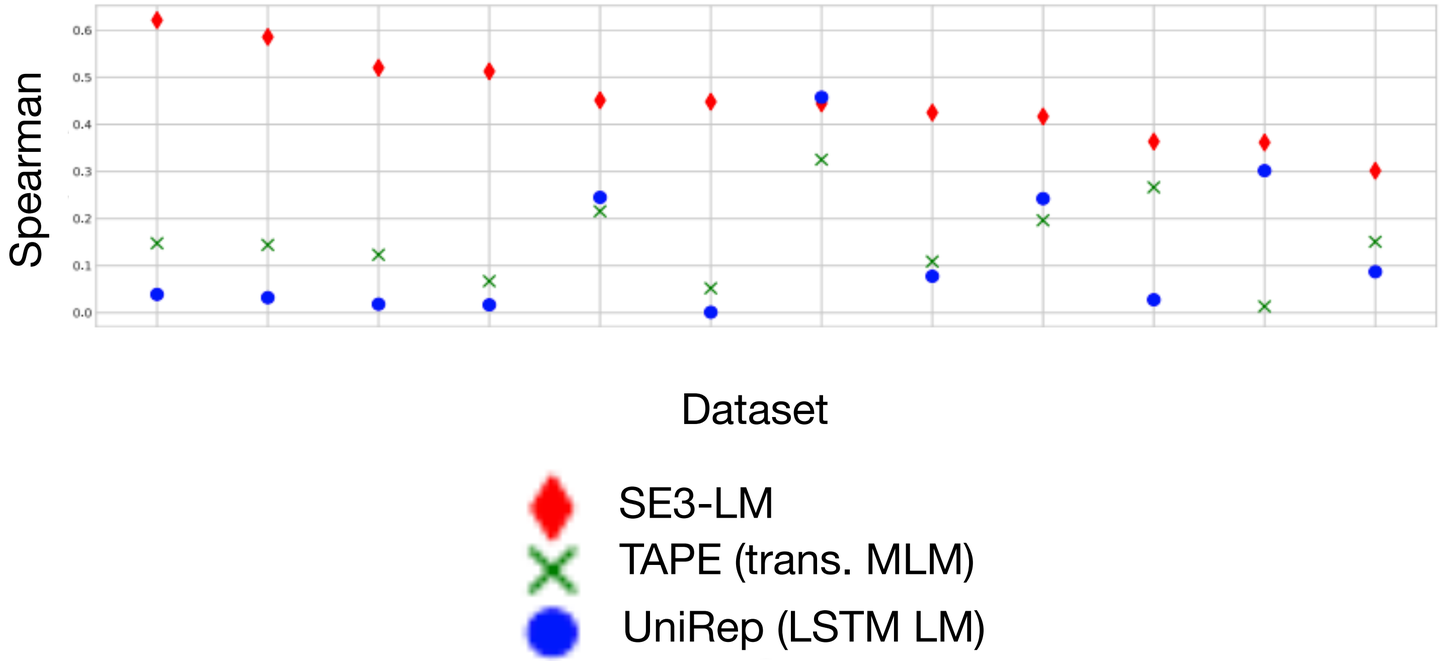
Adding predicted structures improves sequence reconstruction



McPartlon *et al.* 2022

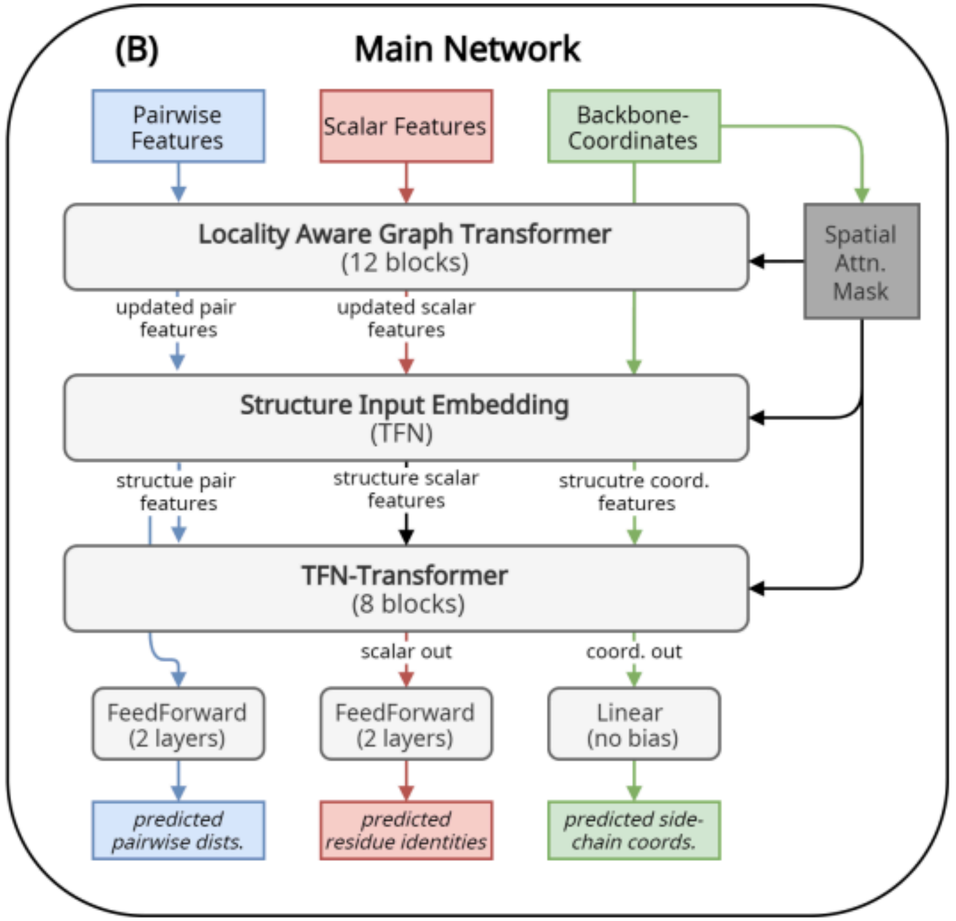
Adding predicted structures improves sequence reconstruction

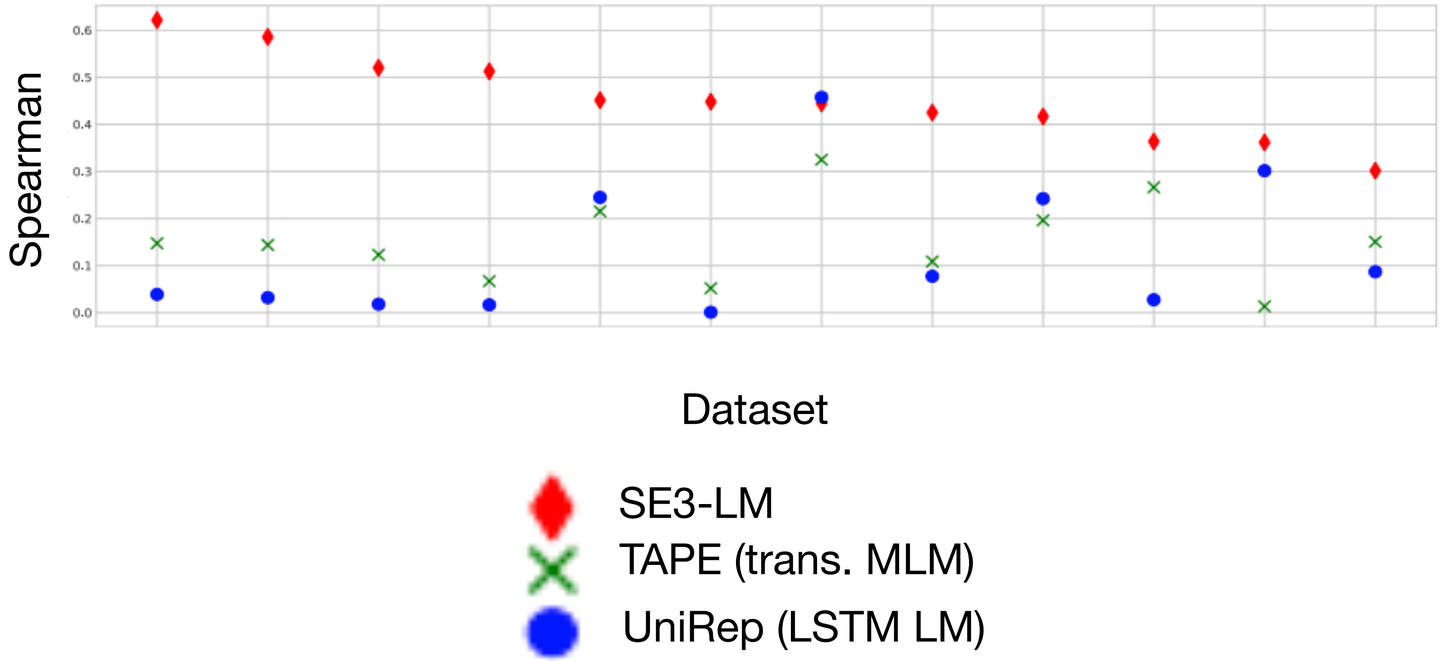




McPartlon *et al.* 2022

Adding predicted structures improves sequence reconstruction

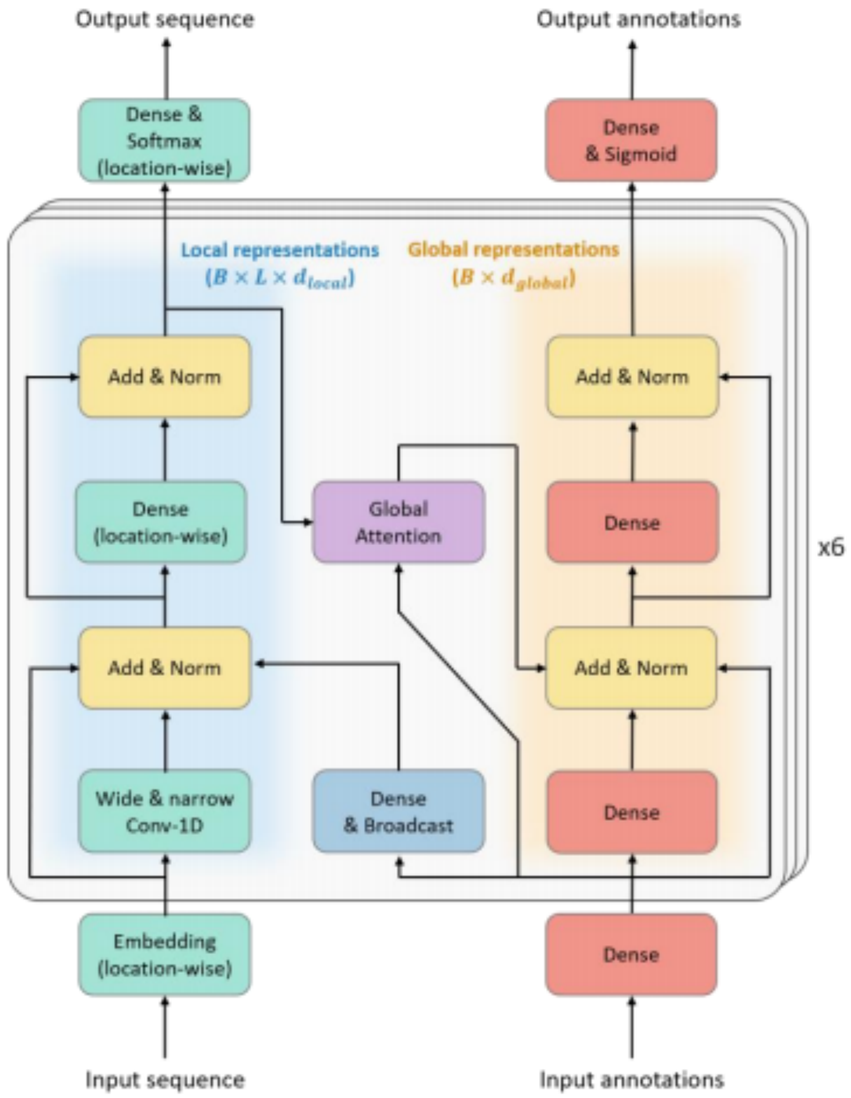




This is all very new — evaluations are very inconsistent!

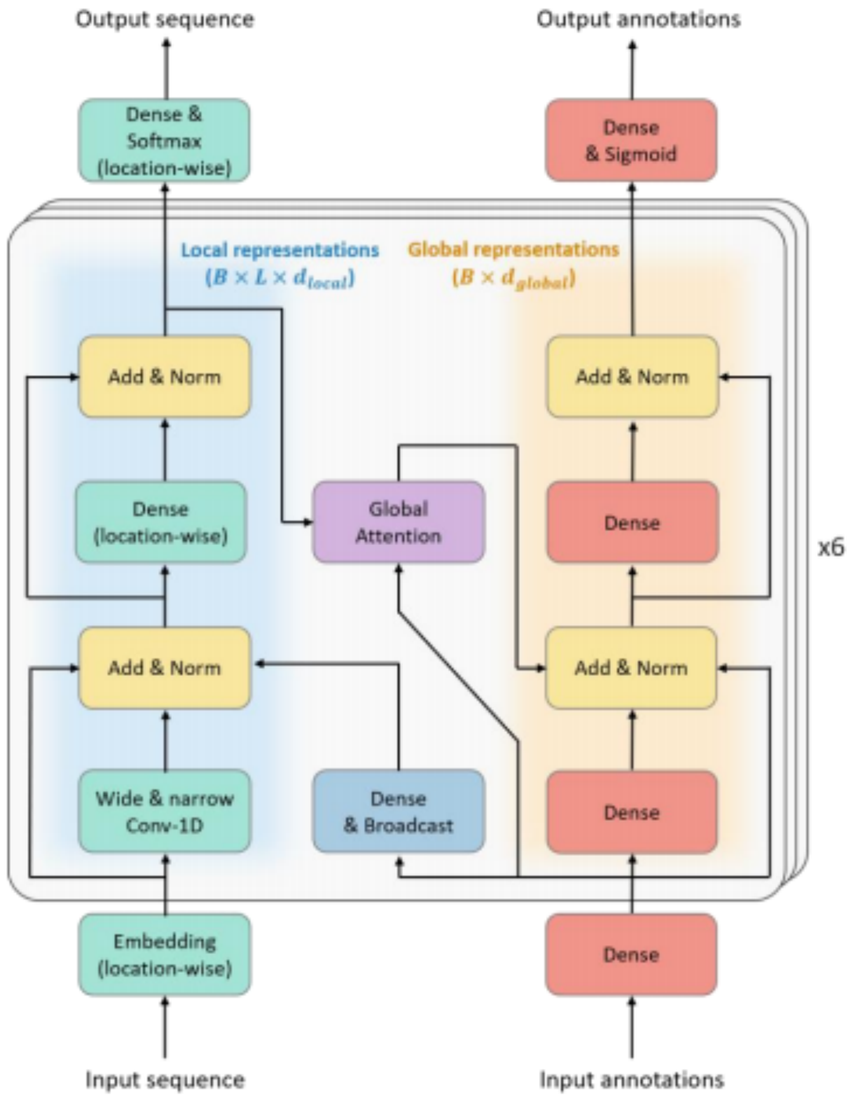
McPartlon *et al.* 2022

Combine sequence and annotations



Brandes *et al.* 2021

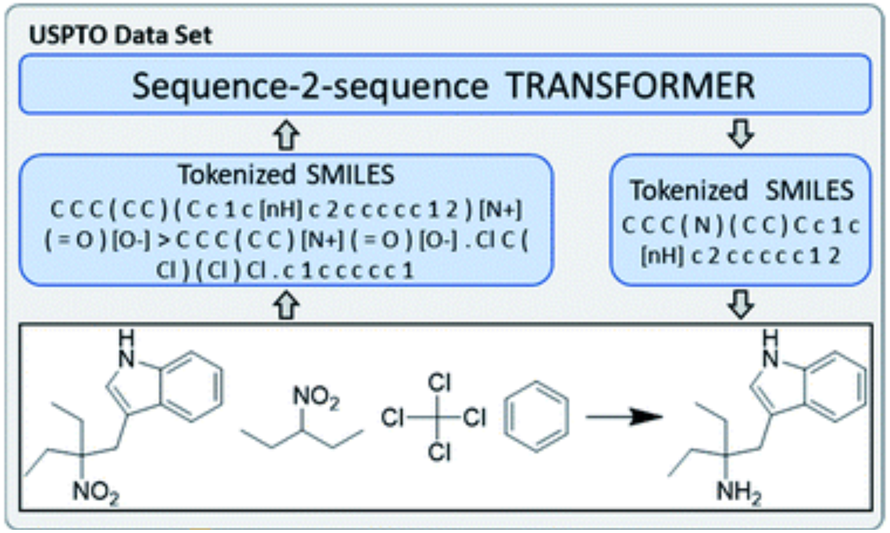
Combine sequence and annotations



	Method	Structure	Evolutionary	Engineering	
		Secondary structure	Remote homology	Fluorescence	Stability
Without Pretraining	TAPE Transformer	0.70	0.09	0.22	-0.06
	LSTM	0.71	0.12	0.21	0.28
	ProteinBERT	0.70	0.06	0.65	0.63
With Pretraining	TAPE Transformer	0.73	0.21	0.68	0.73
	LSTM	0.75	0.26	0.67	0.69
	UniRep mLSTM	0.73	0.23	0.67	0.73
	ProteinBERT	0.74	0.22	0.66	0.76

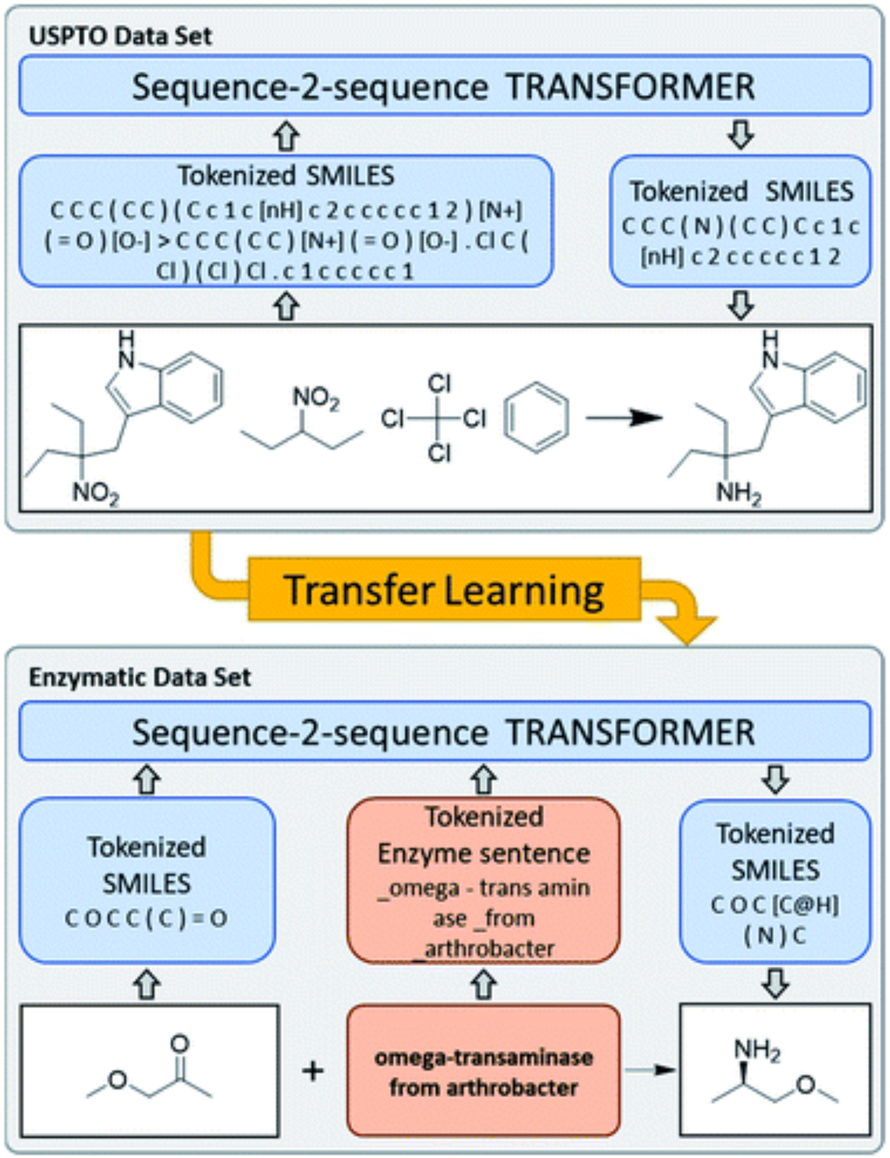
Brandes *et al.* 2021

Predict product from substrates and text



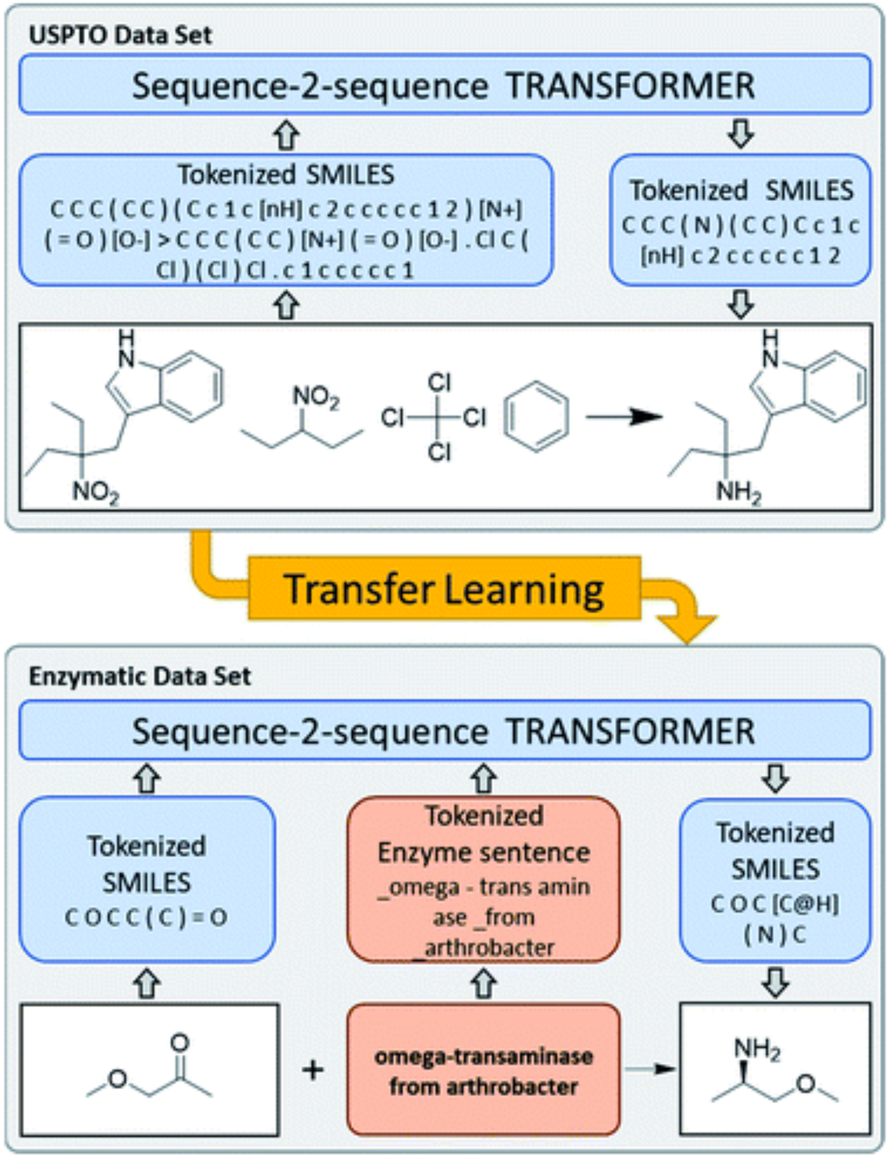
Kreutter *et al. Chem. Sci.* 2021

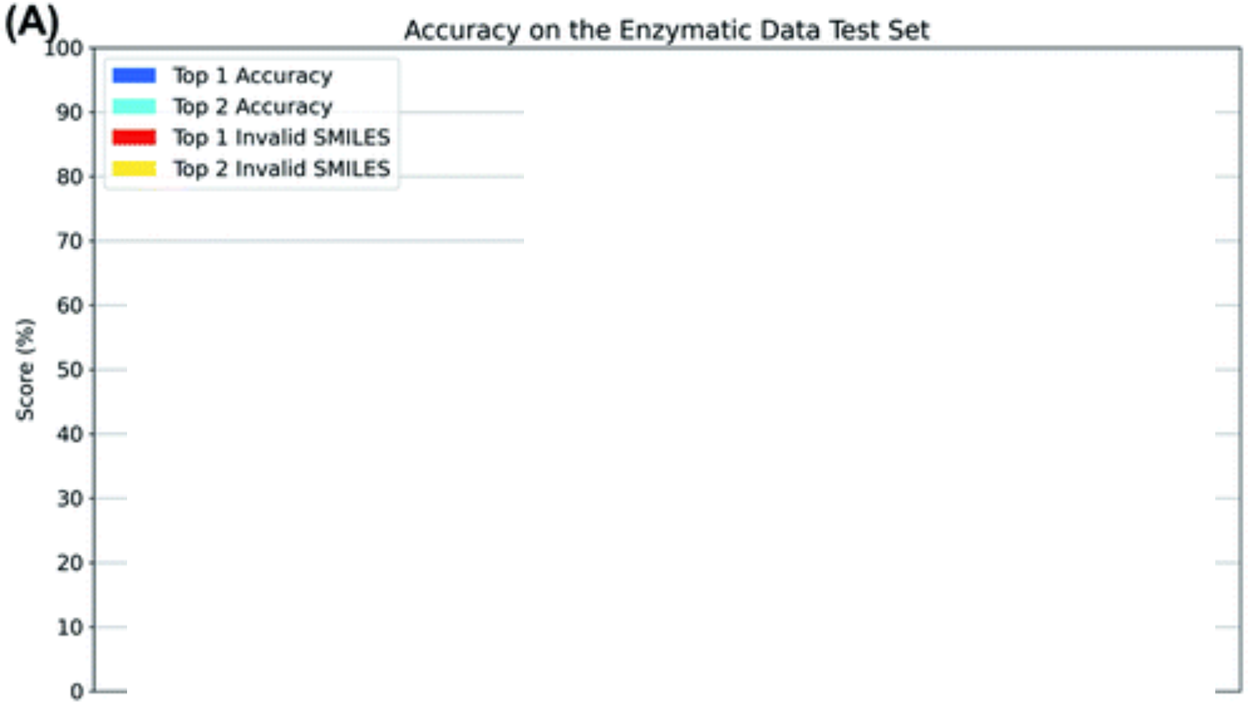
Predict product from substrates and text



Kreutter *et al. Chem. Sci.* 2021

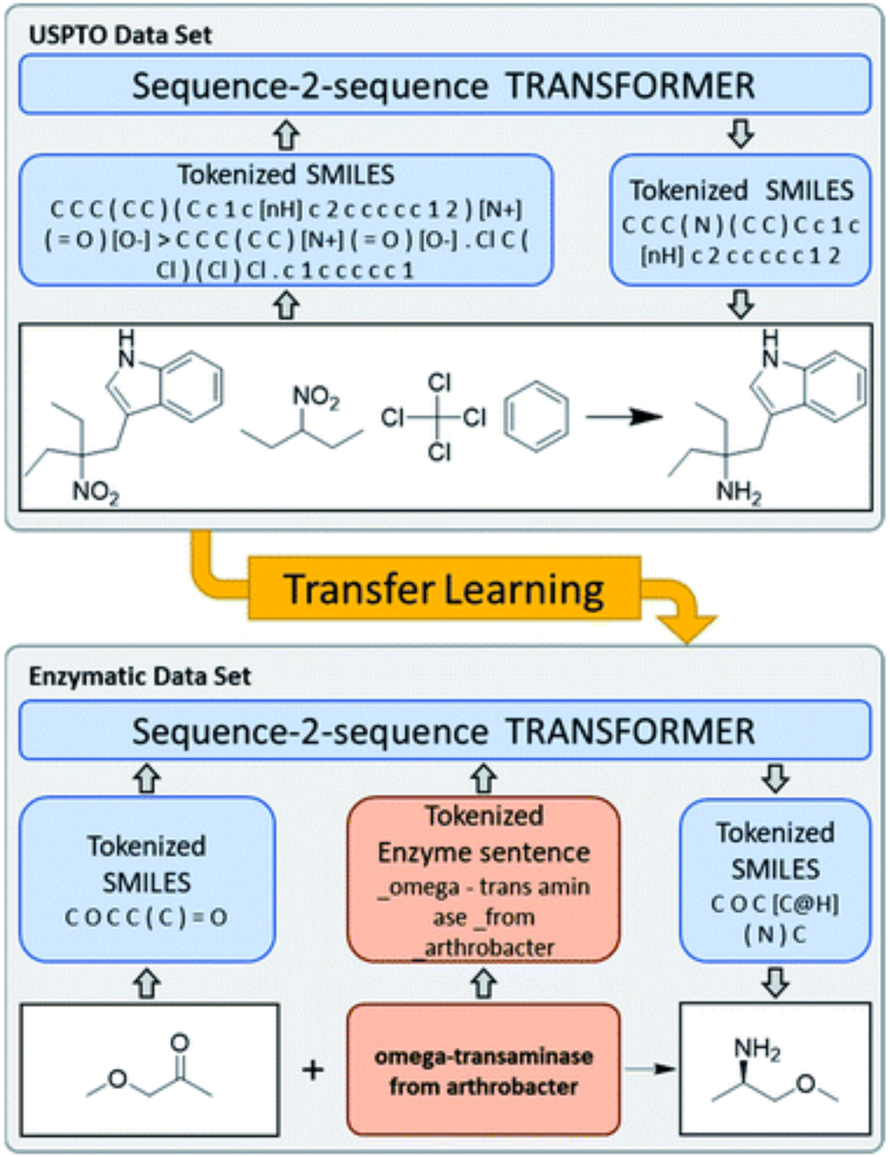
Predict product from substrates and text

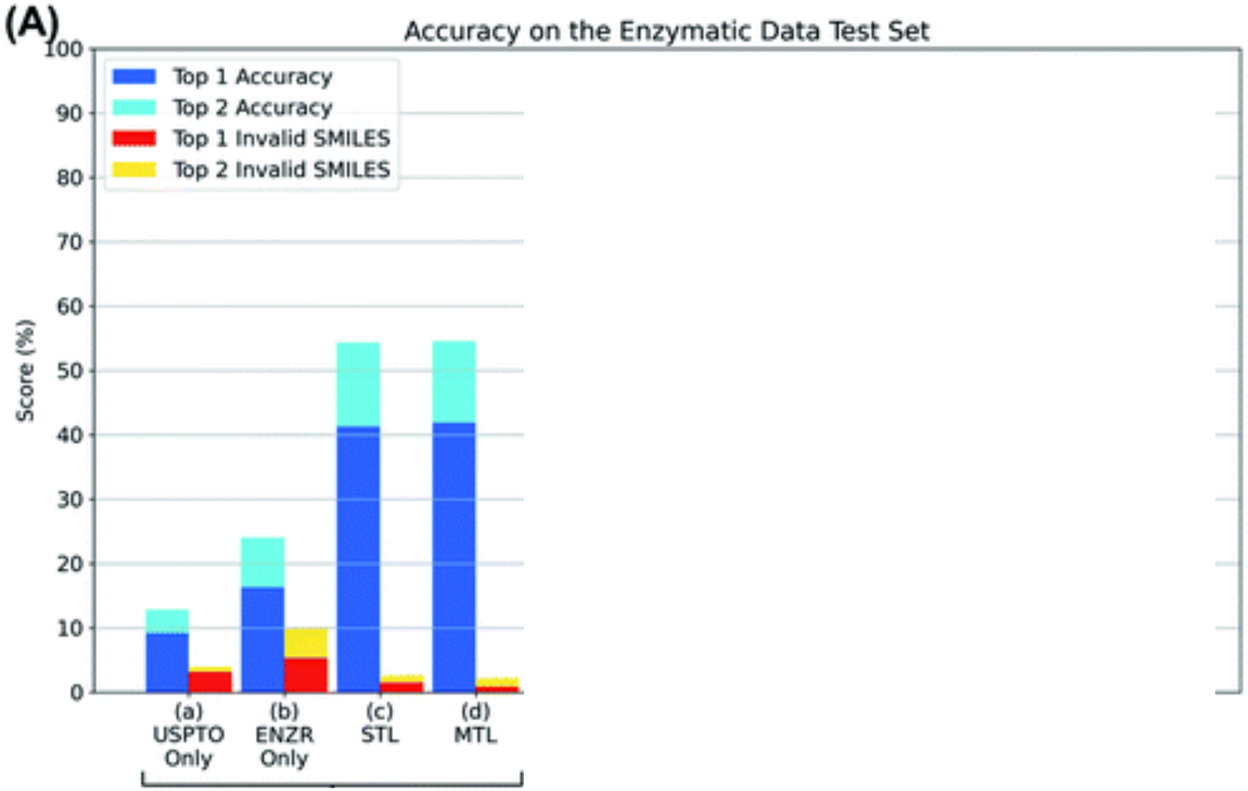




Kreutter *et al. Chem. Sci.* 2021

Predict product from substrates and text

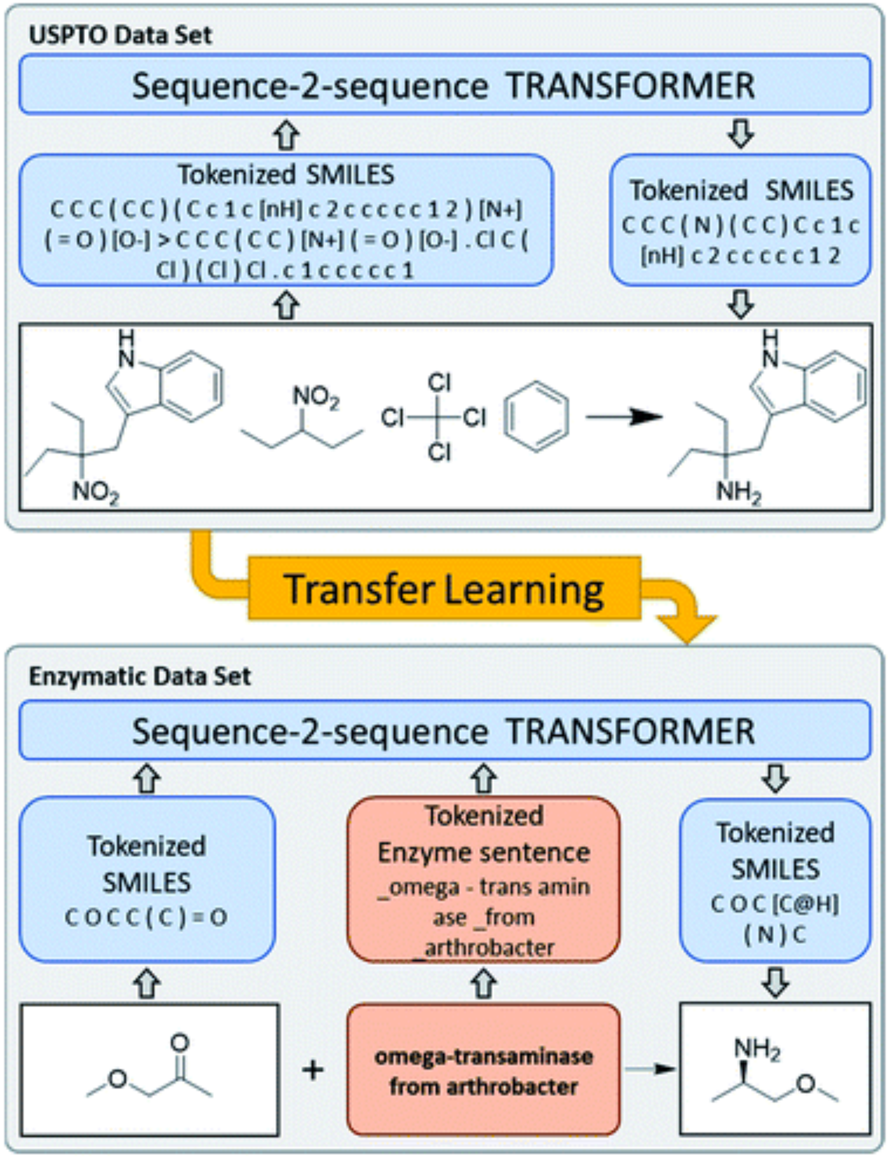


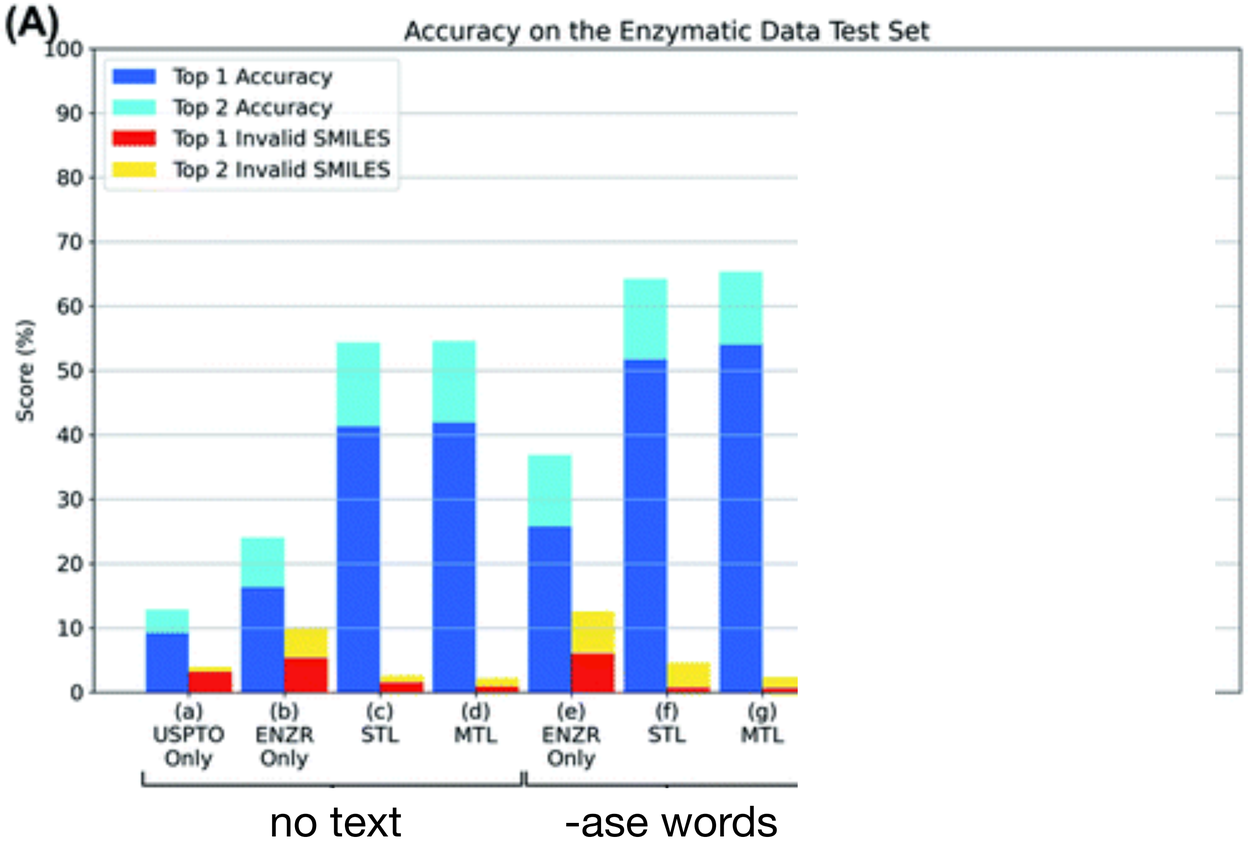


no text

Kreutter *et al. Chem. Sci.* 2021

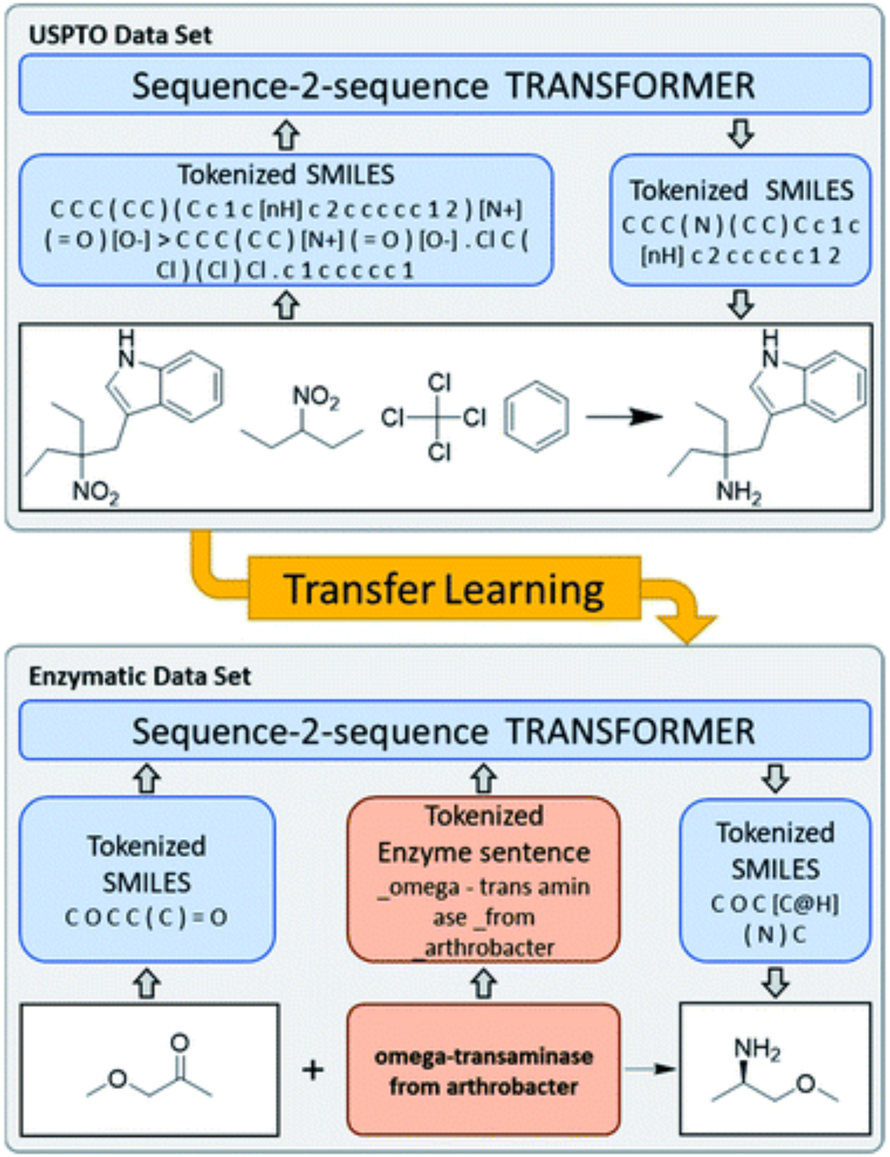
Predict product from substrates and text

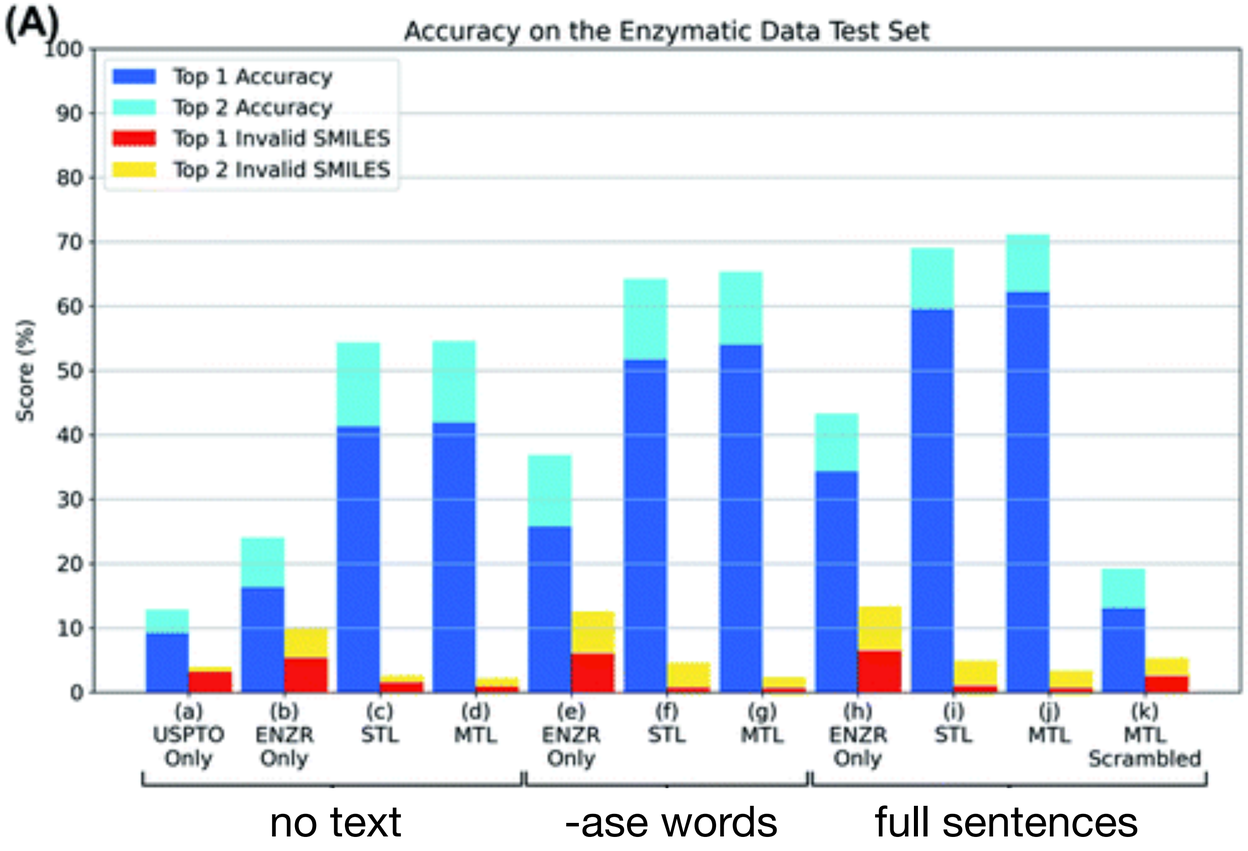




Kreutter *et al. Chem. Sci.* 2021

Predict product from substrates and text





Kreutter *et al. Chem. Sci.* 2021

Questions

@KevinKaichuang
yang.kevin@microsoft.com
https://www.ml4proteinengineering.com/