Multimodal deep learning for protein engineering

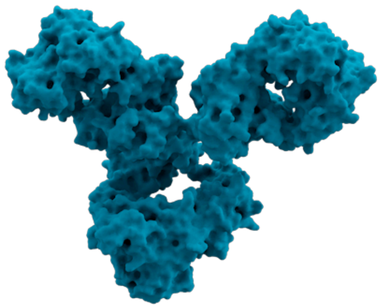
Kevin Kaichuang Yang
Microsoft Research New England
@KevinKaichuang

Proteins are biology’s actuators

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

Proteins are biology’s actuators

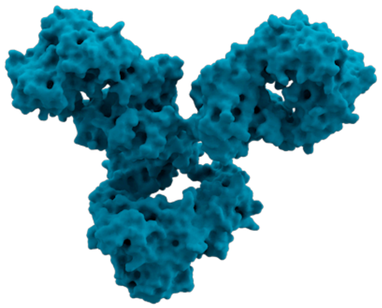
- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling



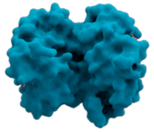
IgG

Proteins are biology’s actuators

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



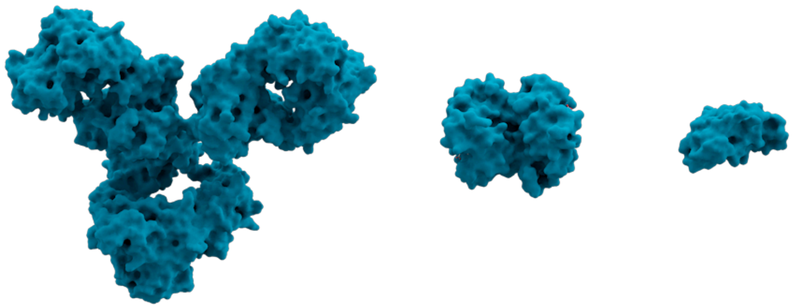
IgG



hemoglobin

Proteins are biology’s actuators

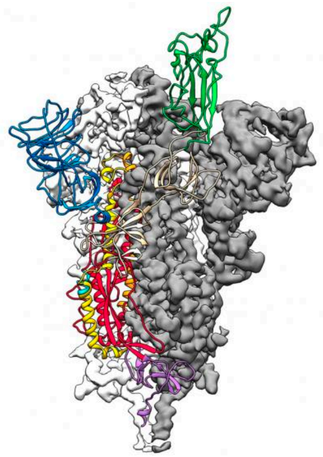
- Human cells contain 1-3 billion proteins each
- Structure, metabolism, and signaling



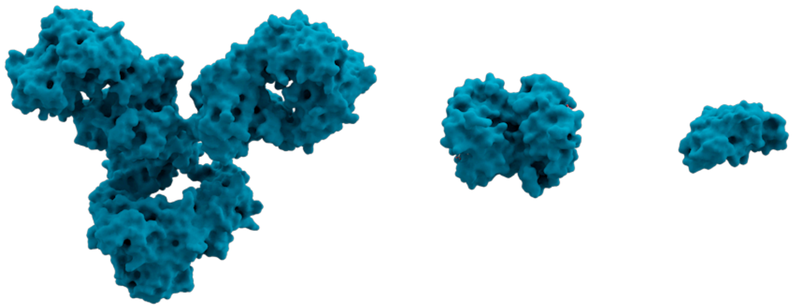
IgG hemoglobin insulin

Proteins are biology’s actuators

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



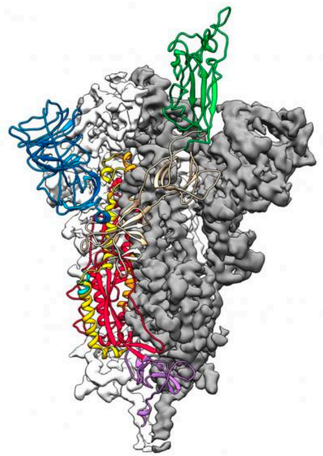
coronavirus spike protein



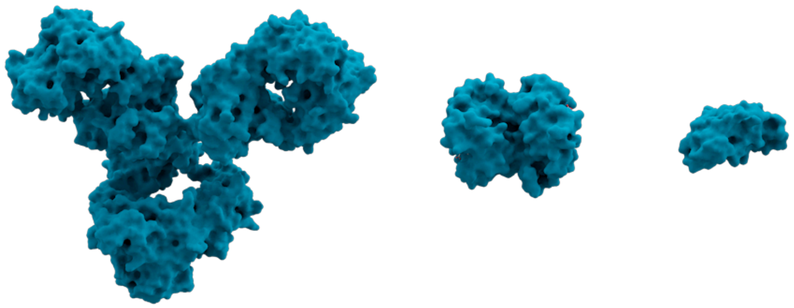
IgG hemoglobin insulin

Proteins are biology’s actuators

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



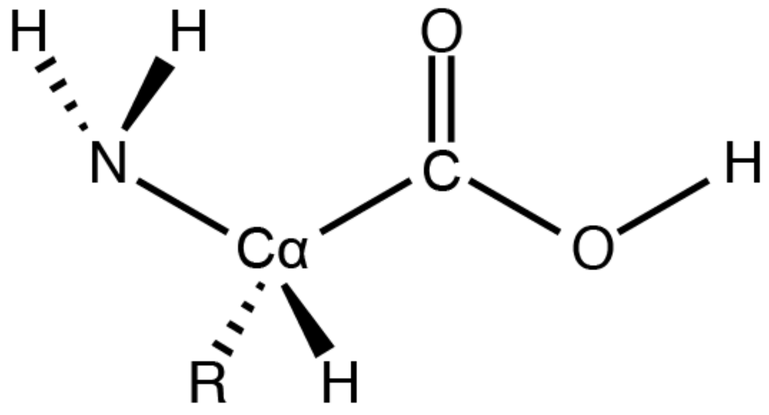
coronavirus spike protein



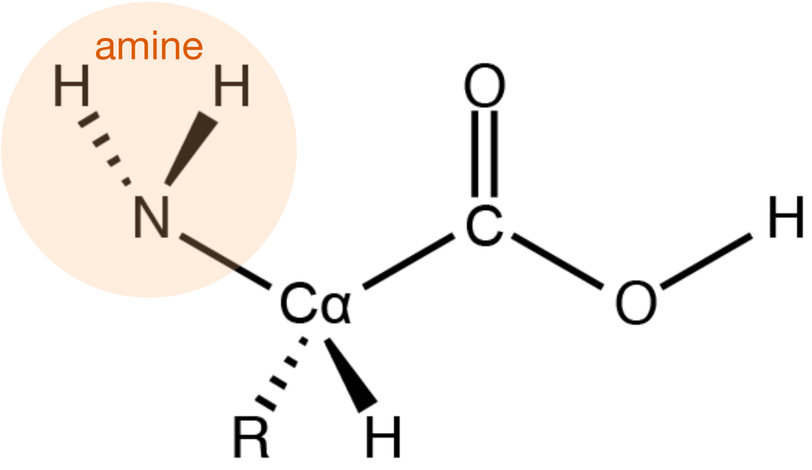
IgG hemoglobin insulin

luciferase

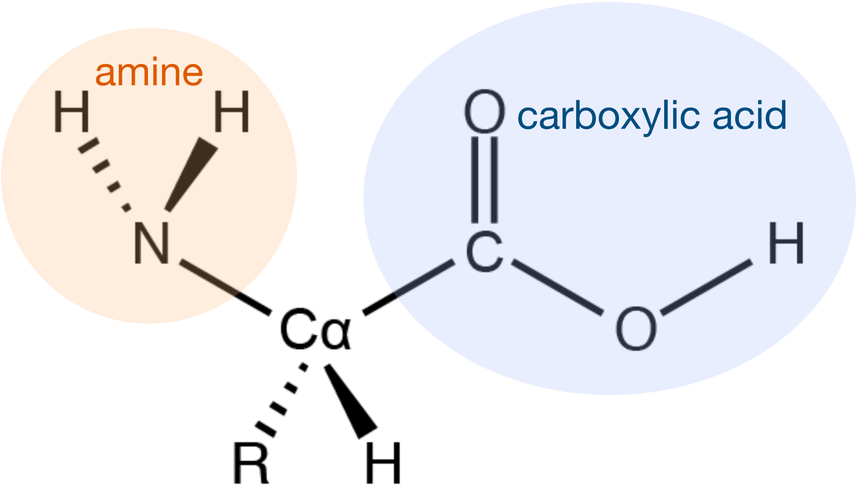
Diversity arises from 20 building blocks



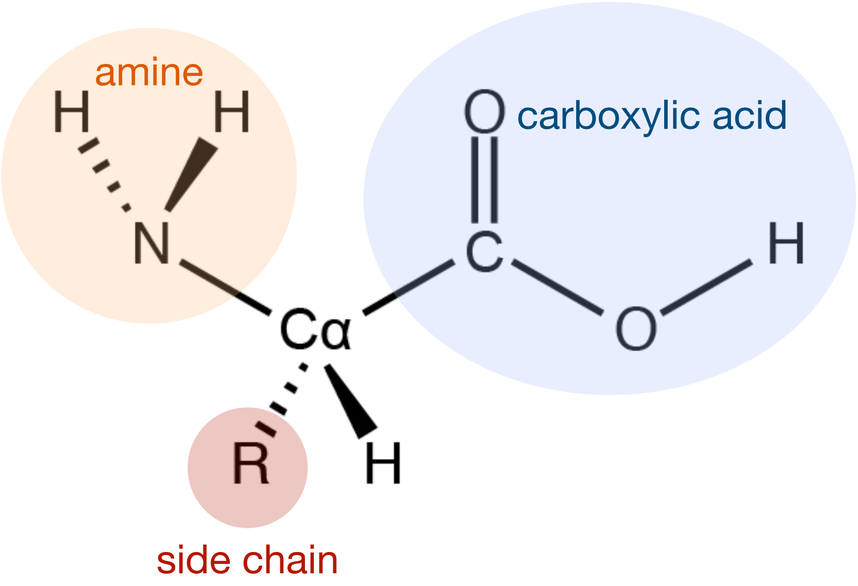
Diversity arises from 20 building blocks



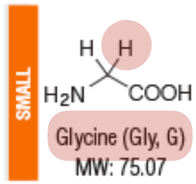
Diversity arises from 20 building blocks

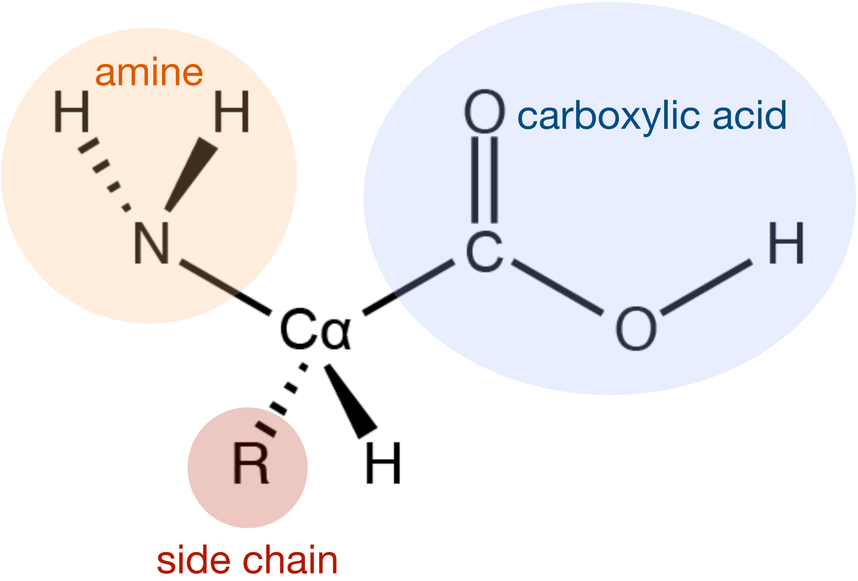


Diversity arises from 20 building blocks

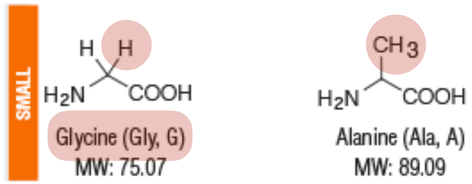


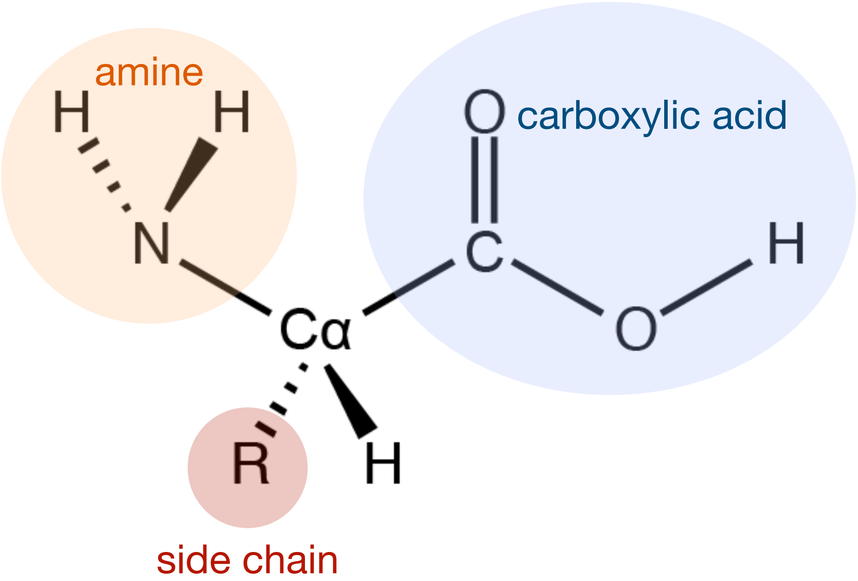
Diversity arises from 20 building blocks



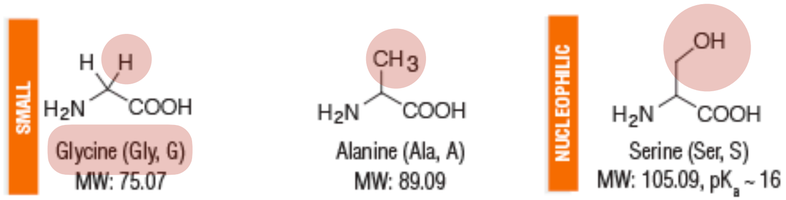


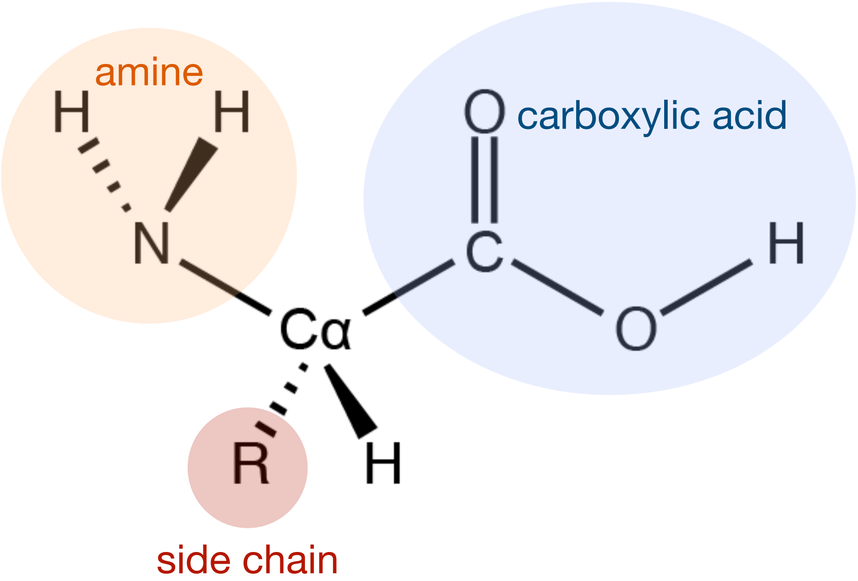
Diversity arises from 20 building blocks



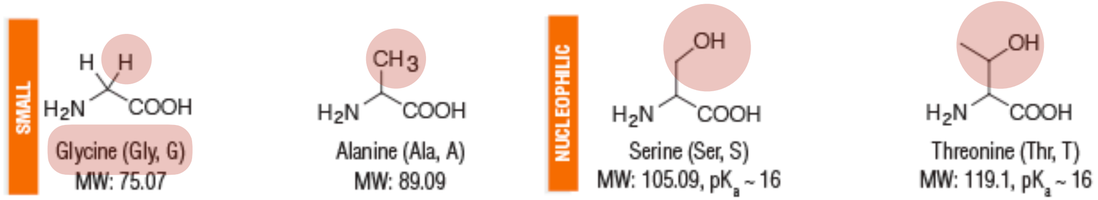


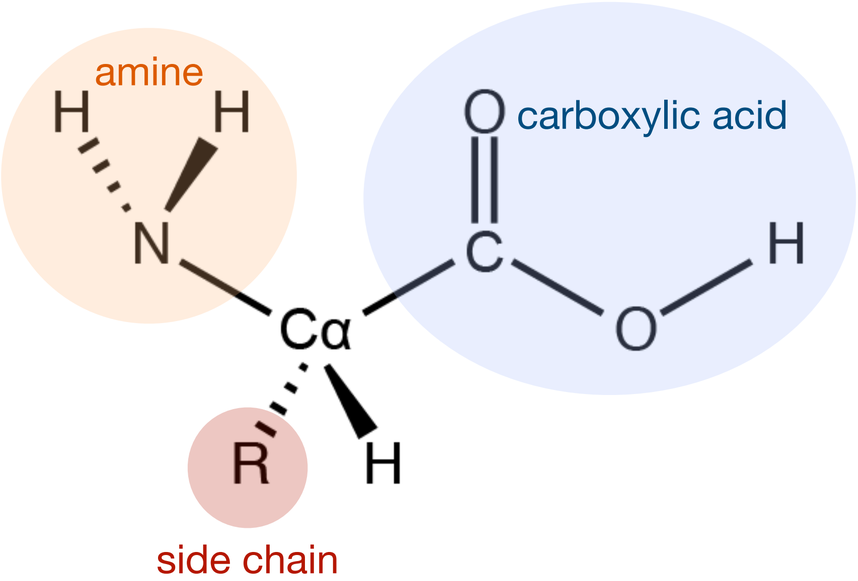
Diversity arises from 20 building blocks



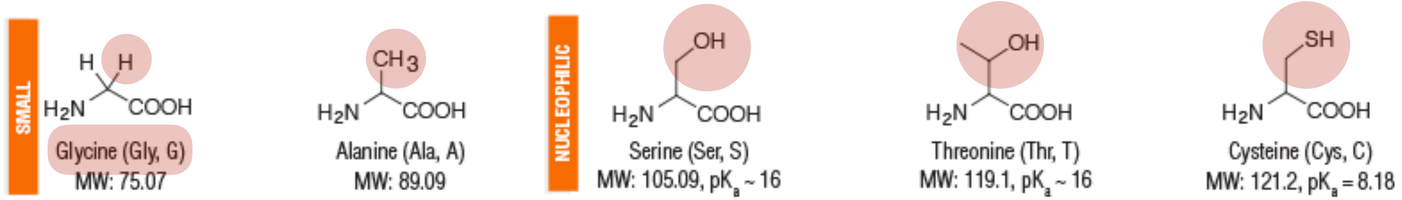


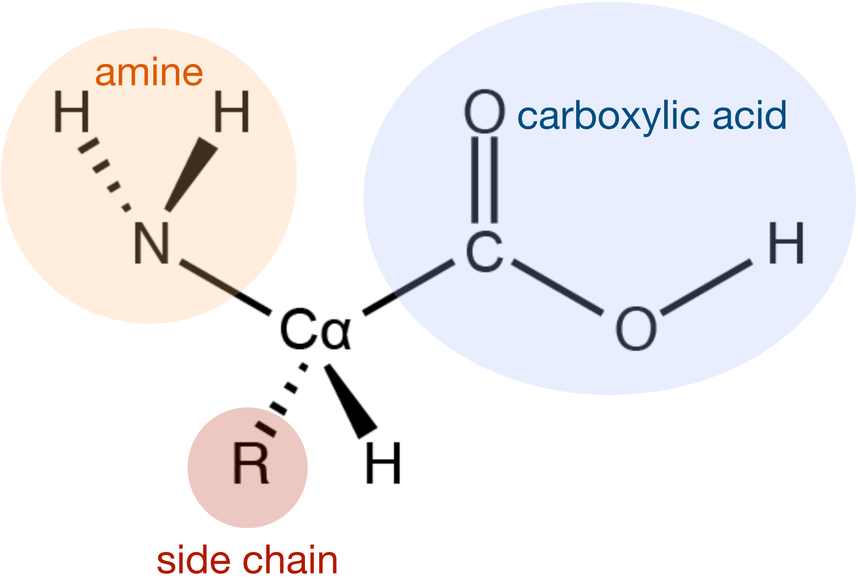
Diversity arises from 20 building blocks



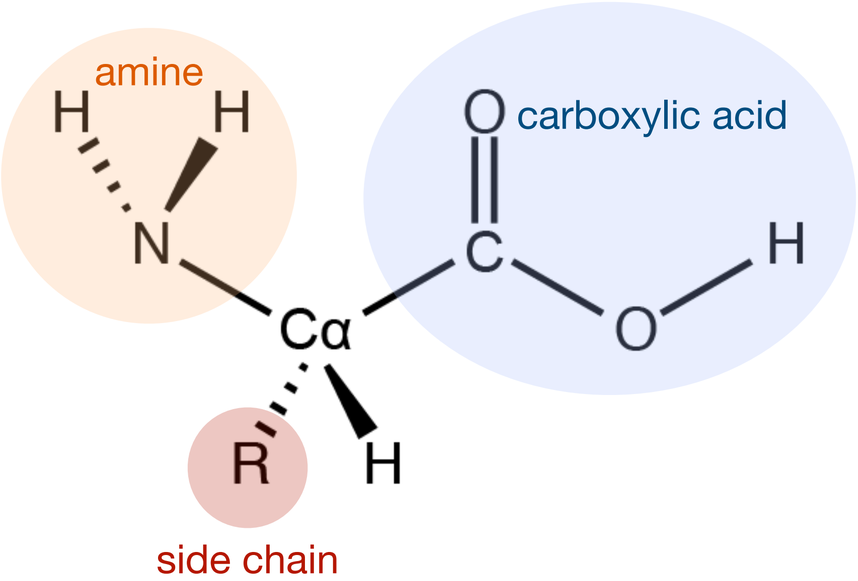


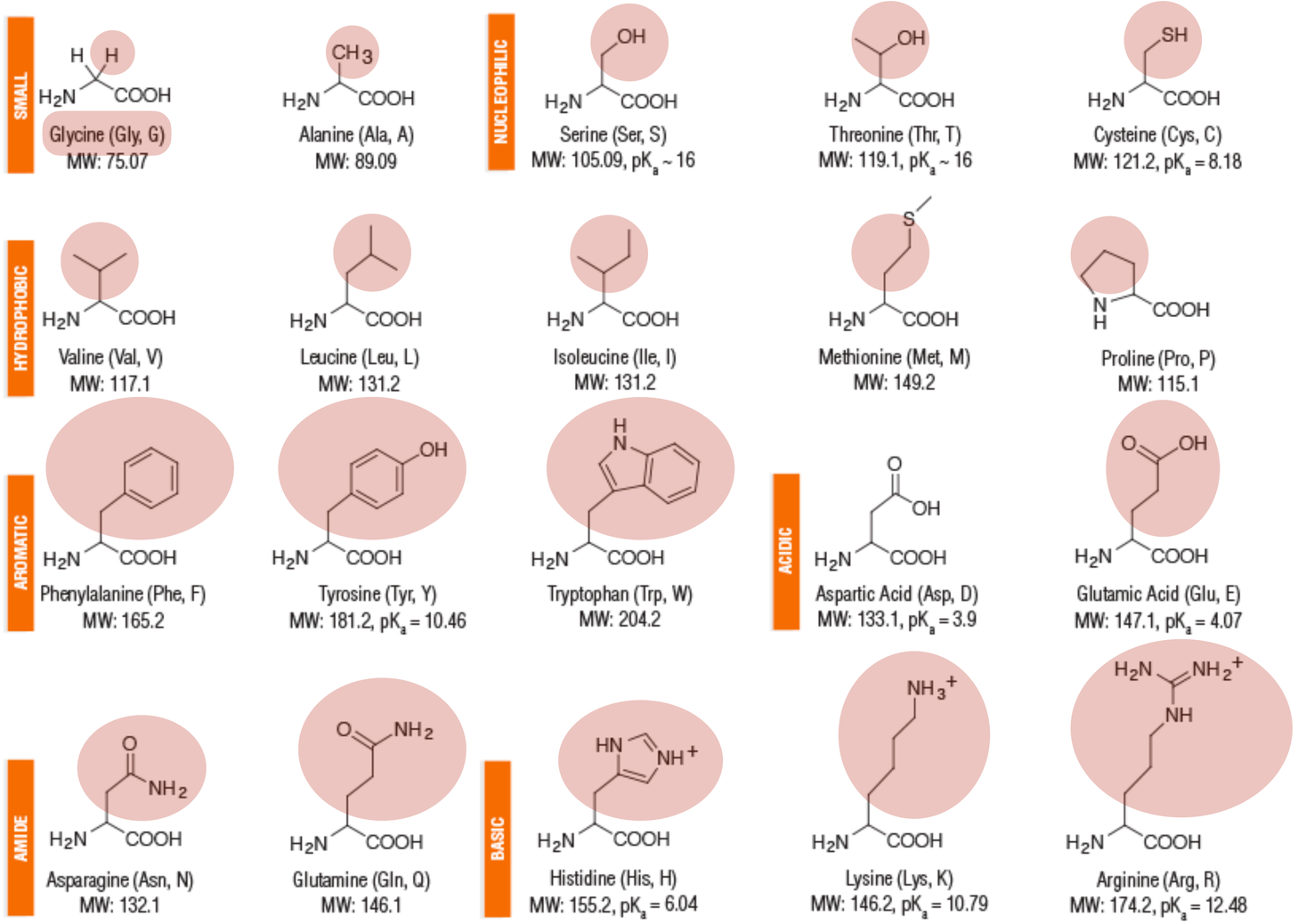
Diversity arises from 20 building blocks





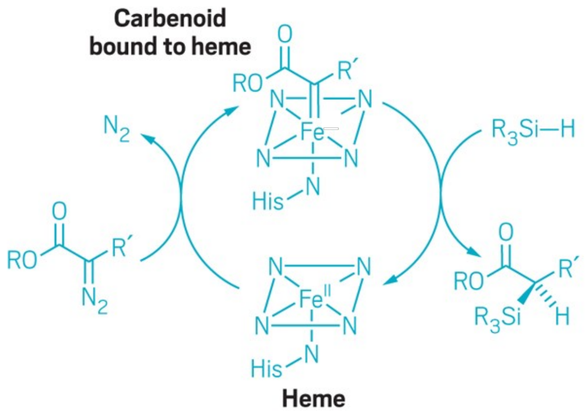
Diversity arises from 20 building blocks





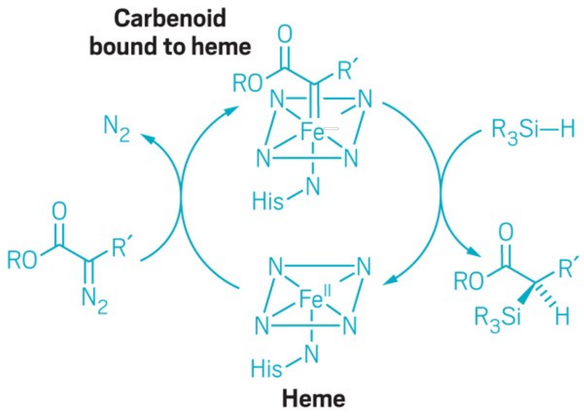
Why design proteins?

Why design proteins?

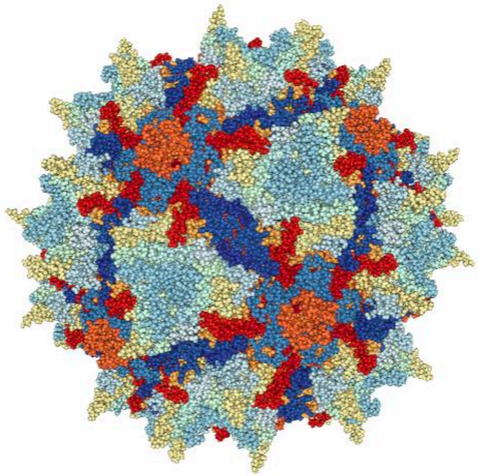


new chemistry

Why design proteins?

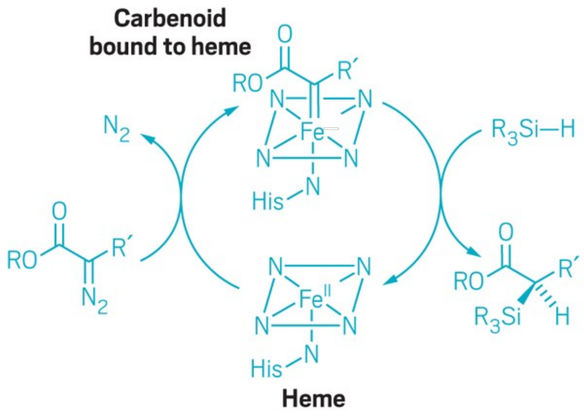


new chemistry

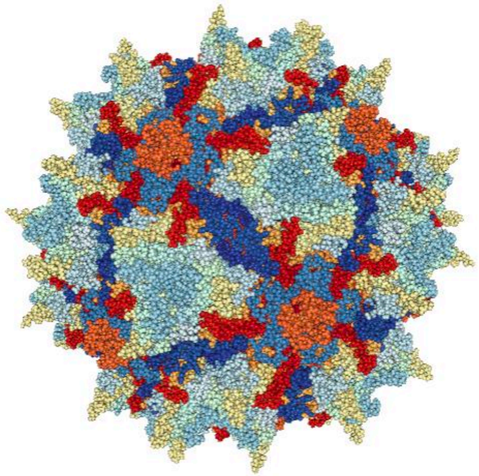


understand function

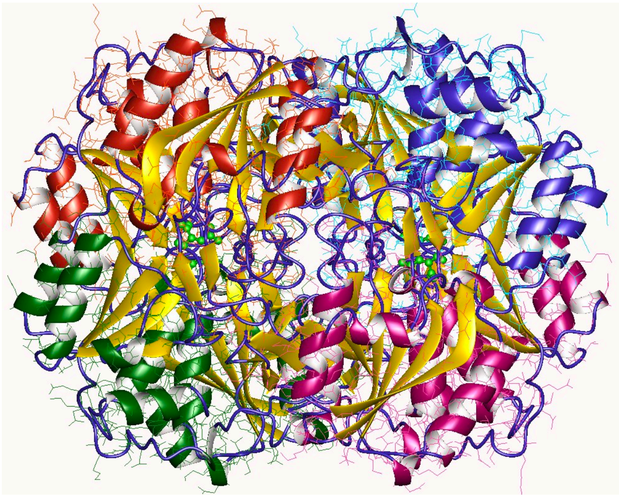
Why design proteins?



new chemistry

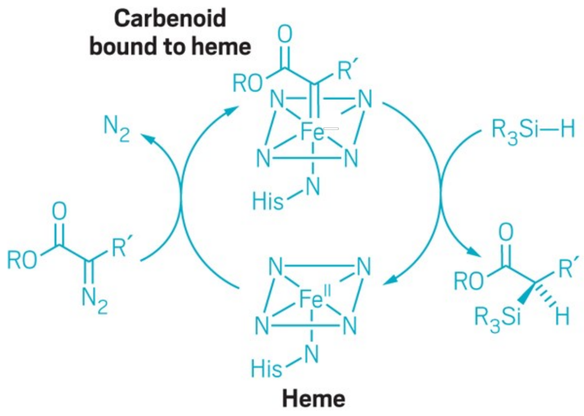


understand function

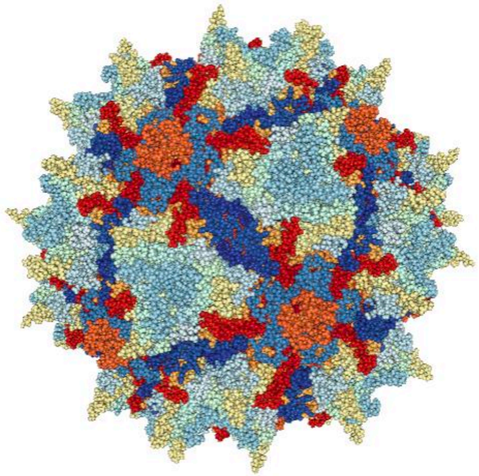


therapeutics

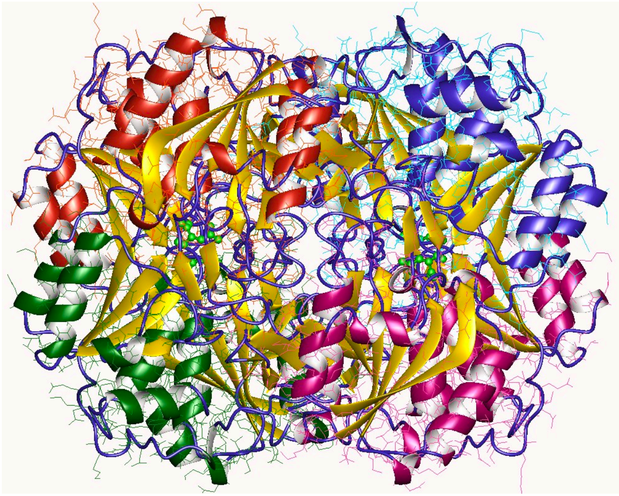
Why design proteins?



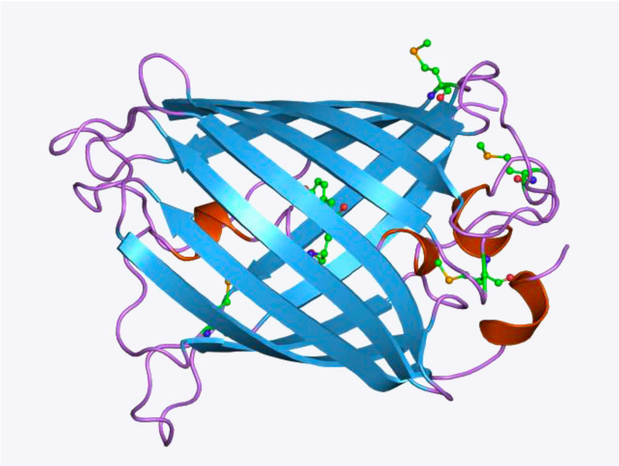
new chemistry



understand function

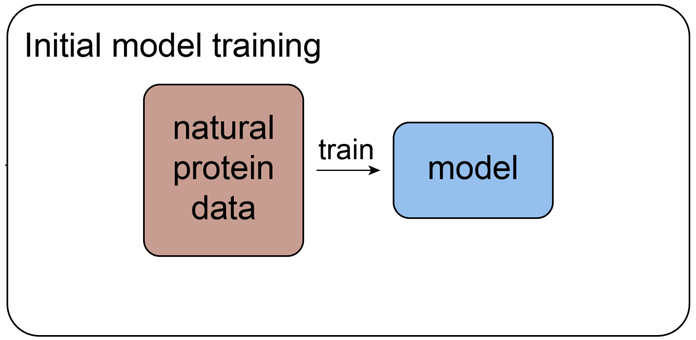


therapeutics

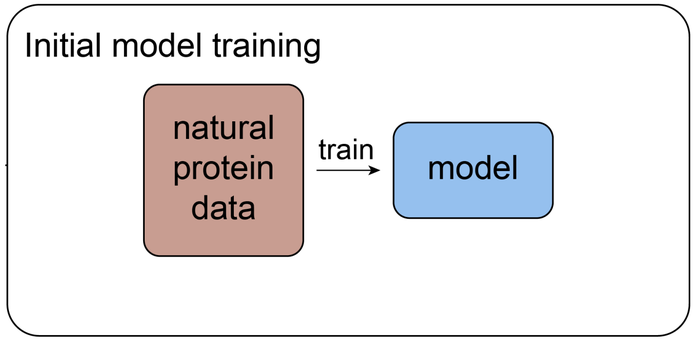


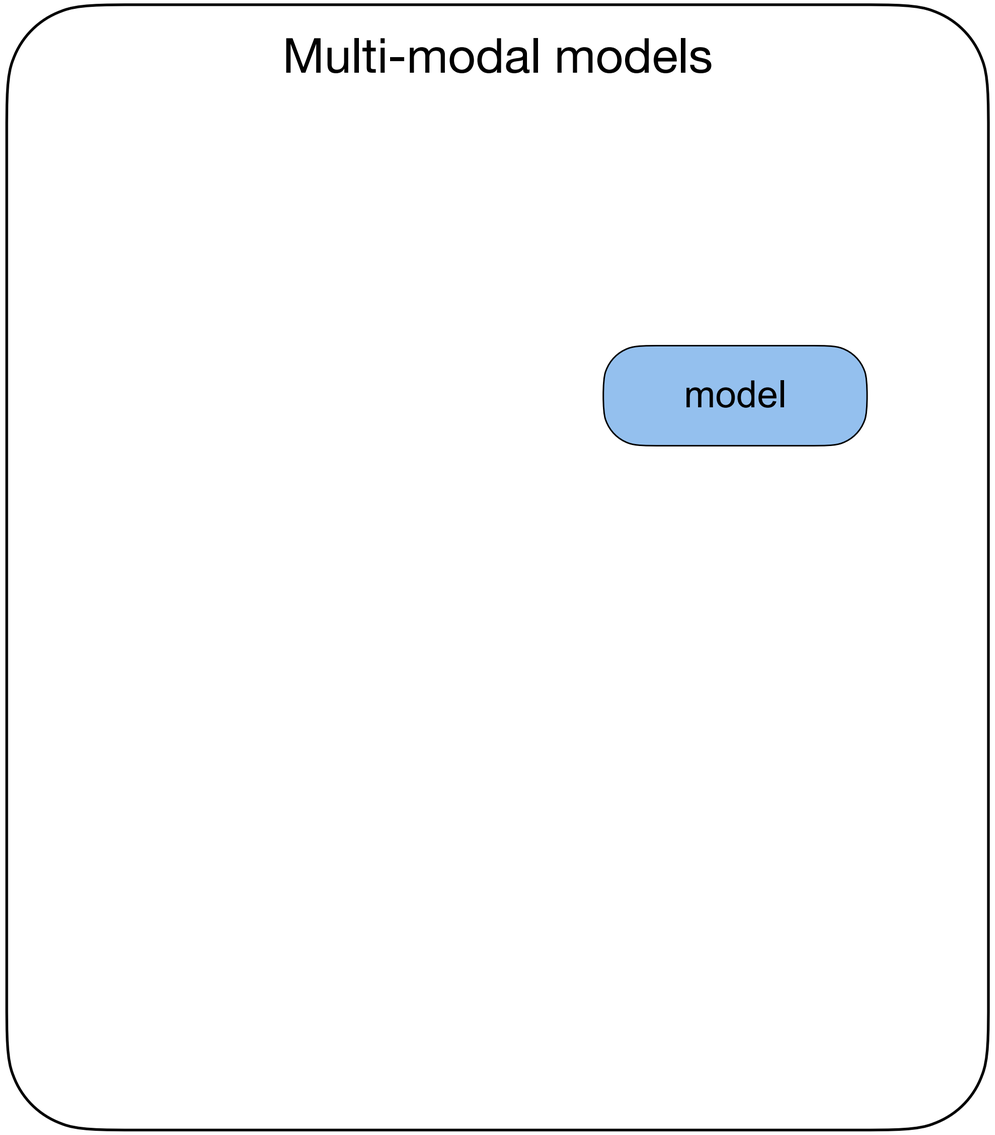
molecular tools

Use multiple data modalities to discover and design proteins

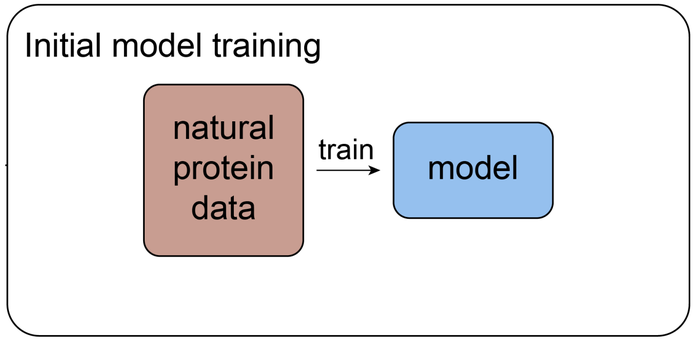


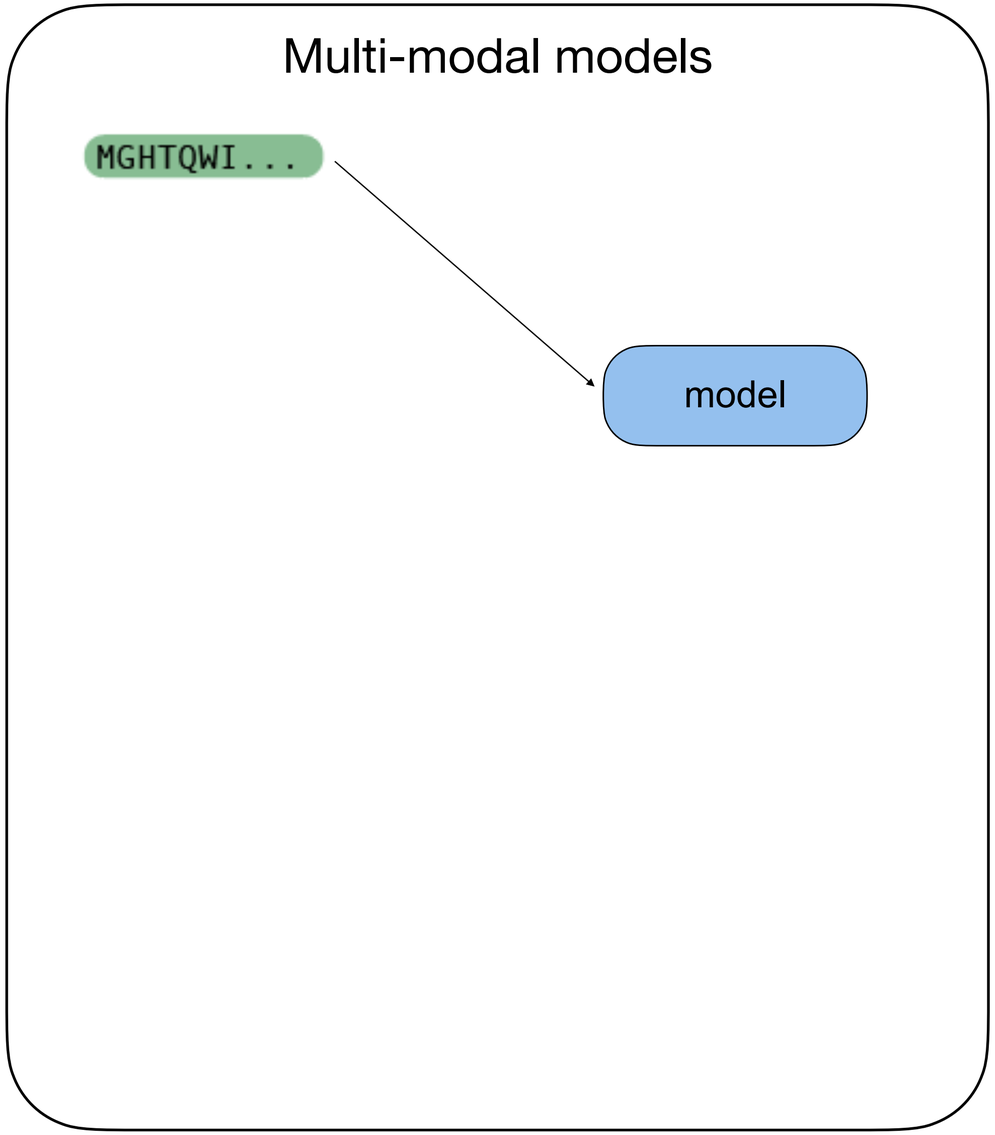
Use multiple data modalities to discover and design proteins



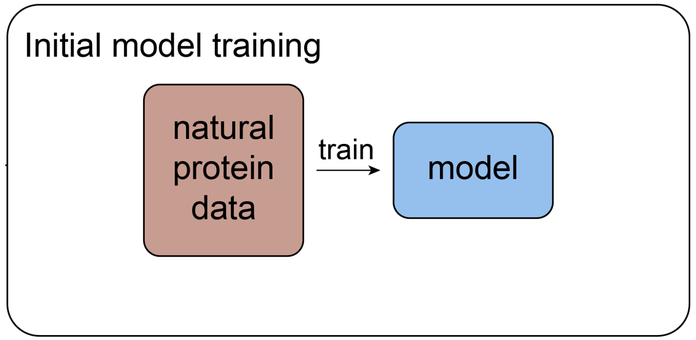


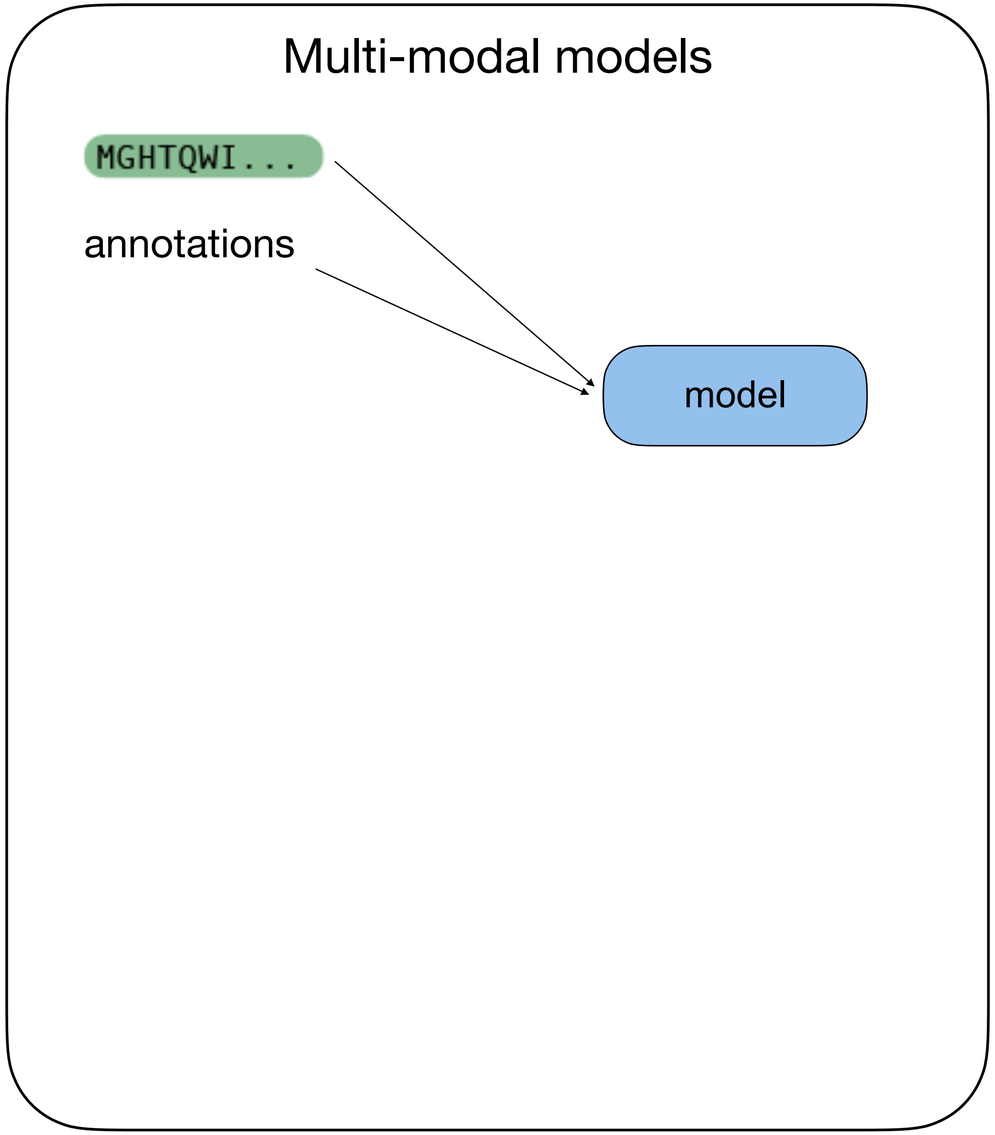
Use multiple data modalities to discover and design proteins



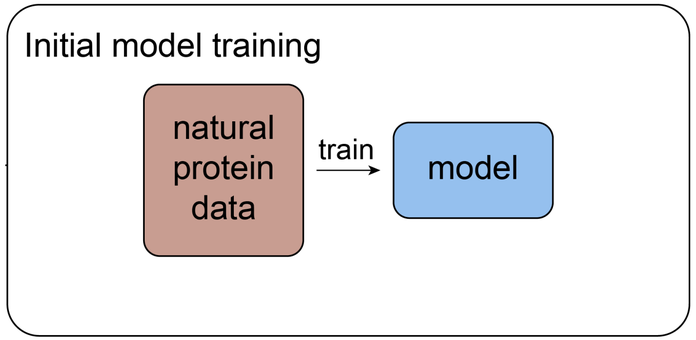


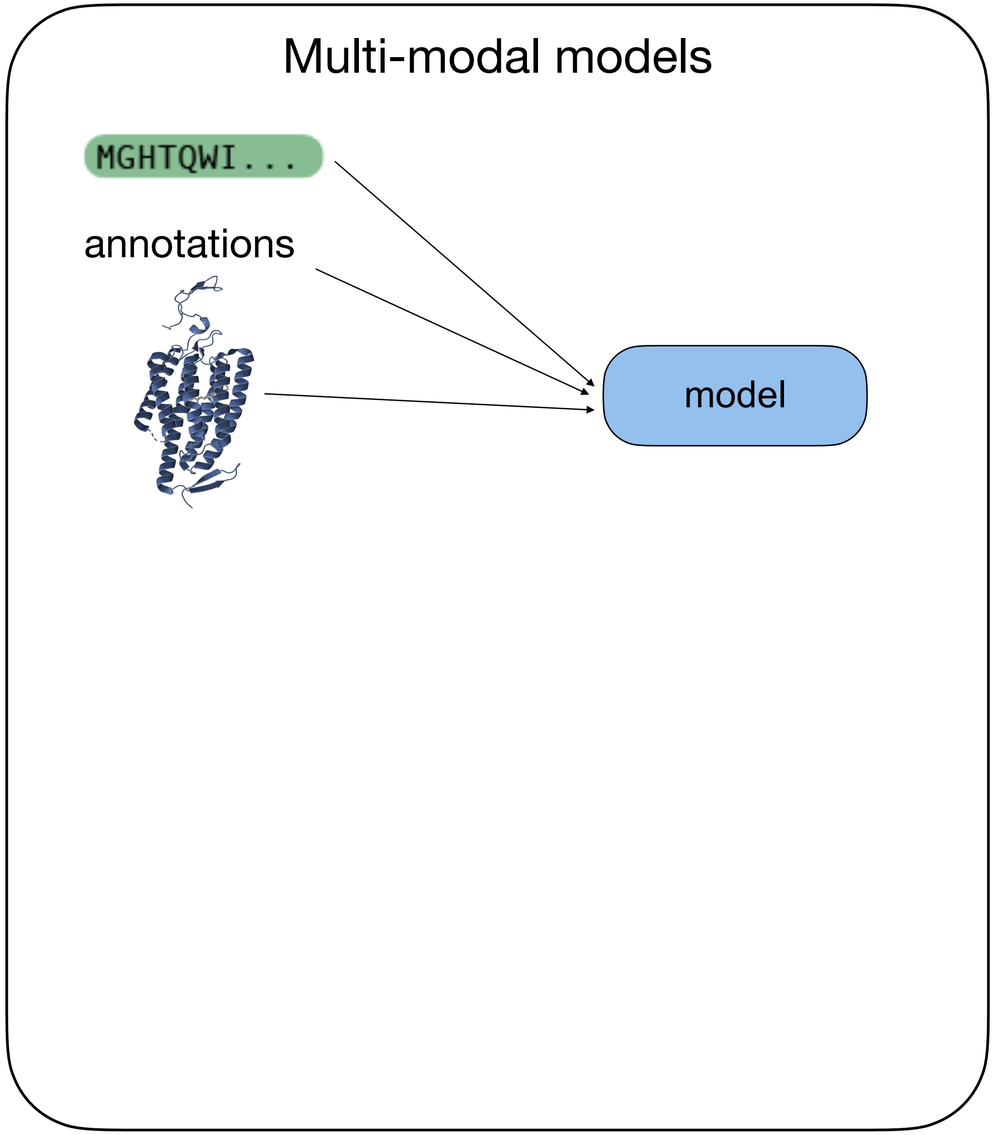
Use multiple data modalities to discover and design proteins



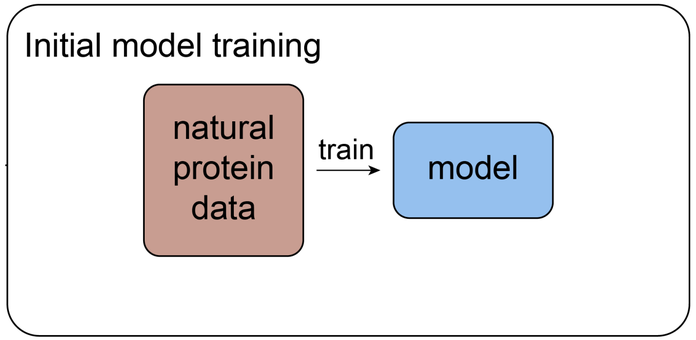


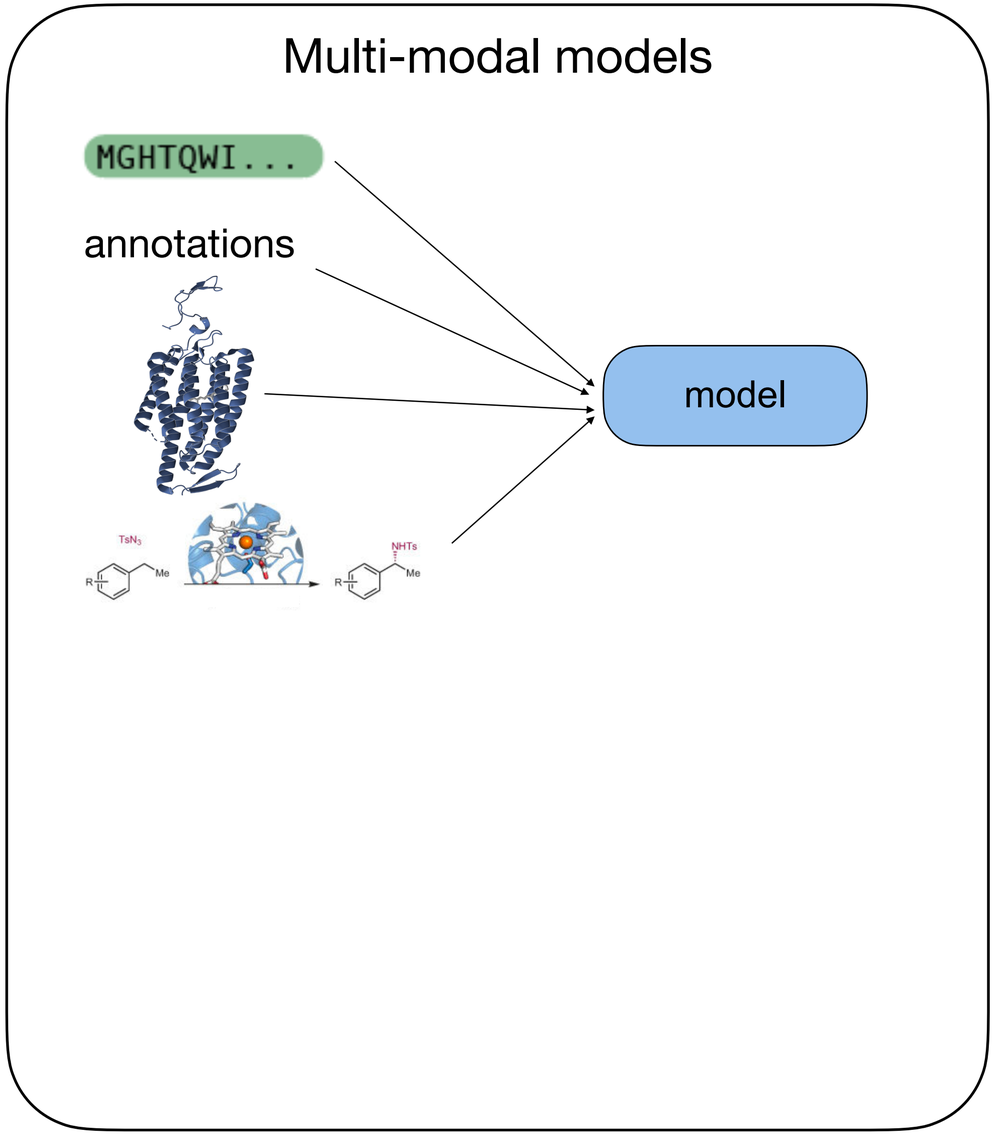
Use multiple data modalities to discover and design proteins



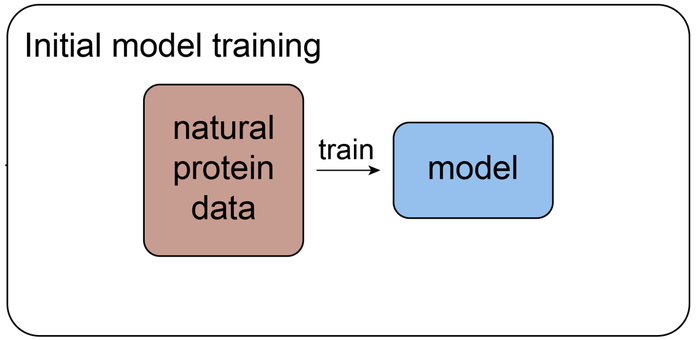


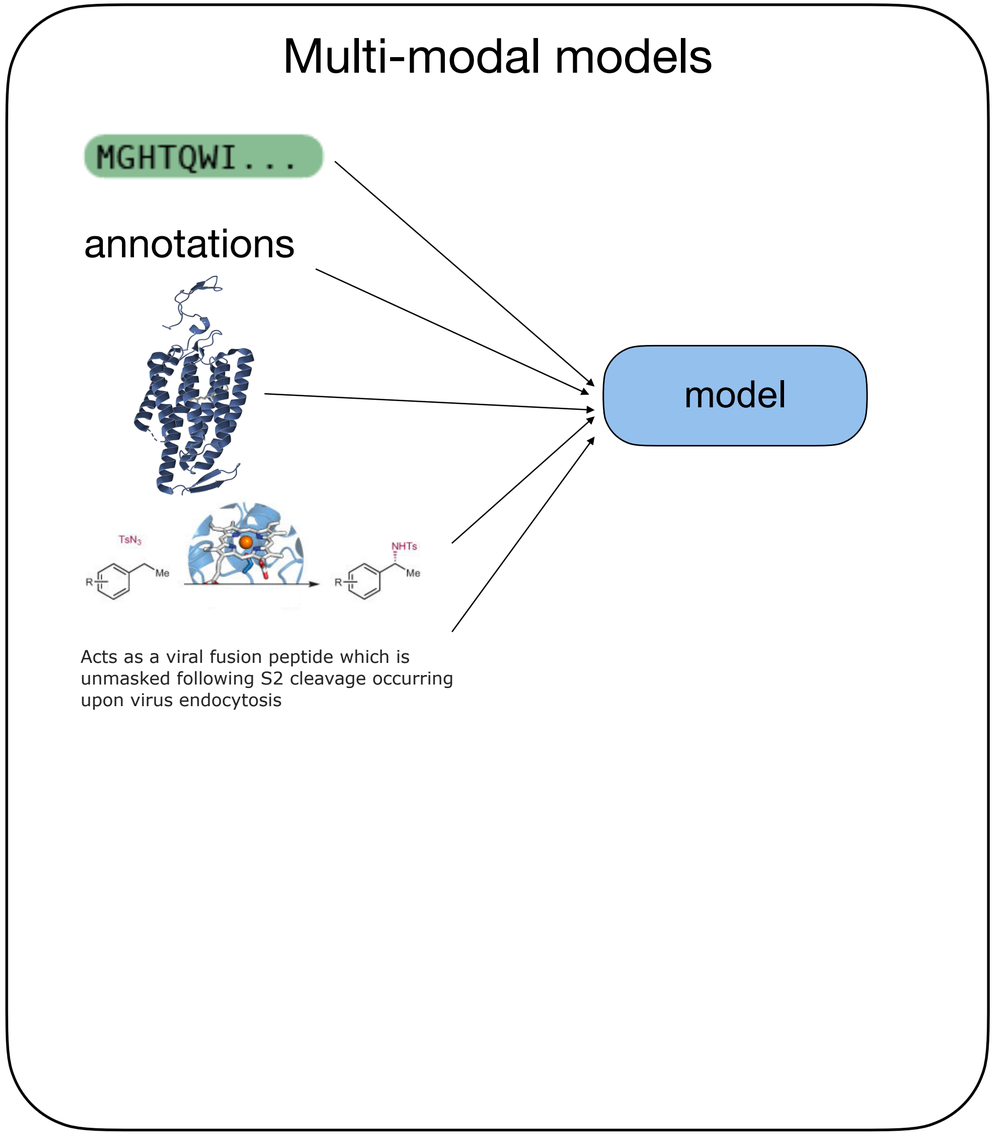
Use multiple data modalities to discover and design proteins



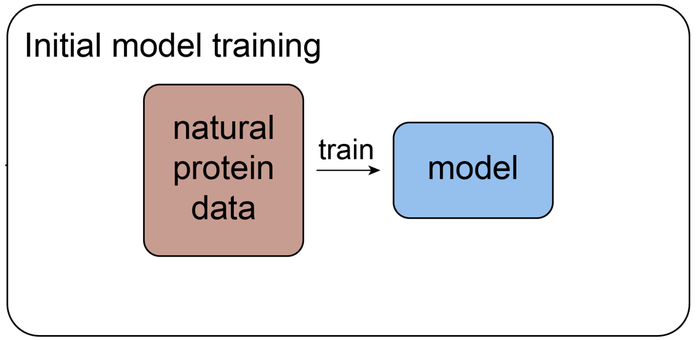


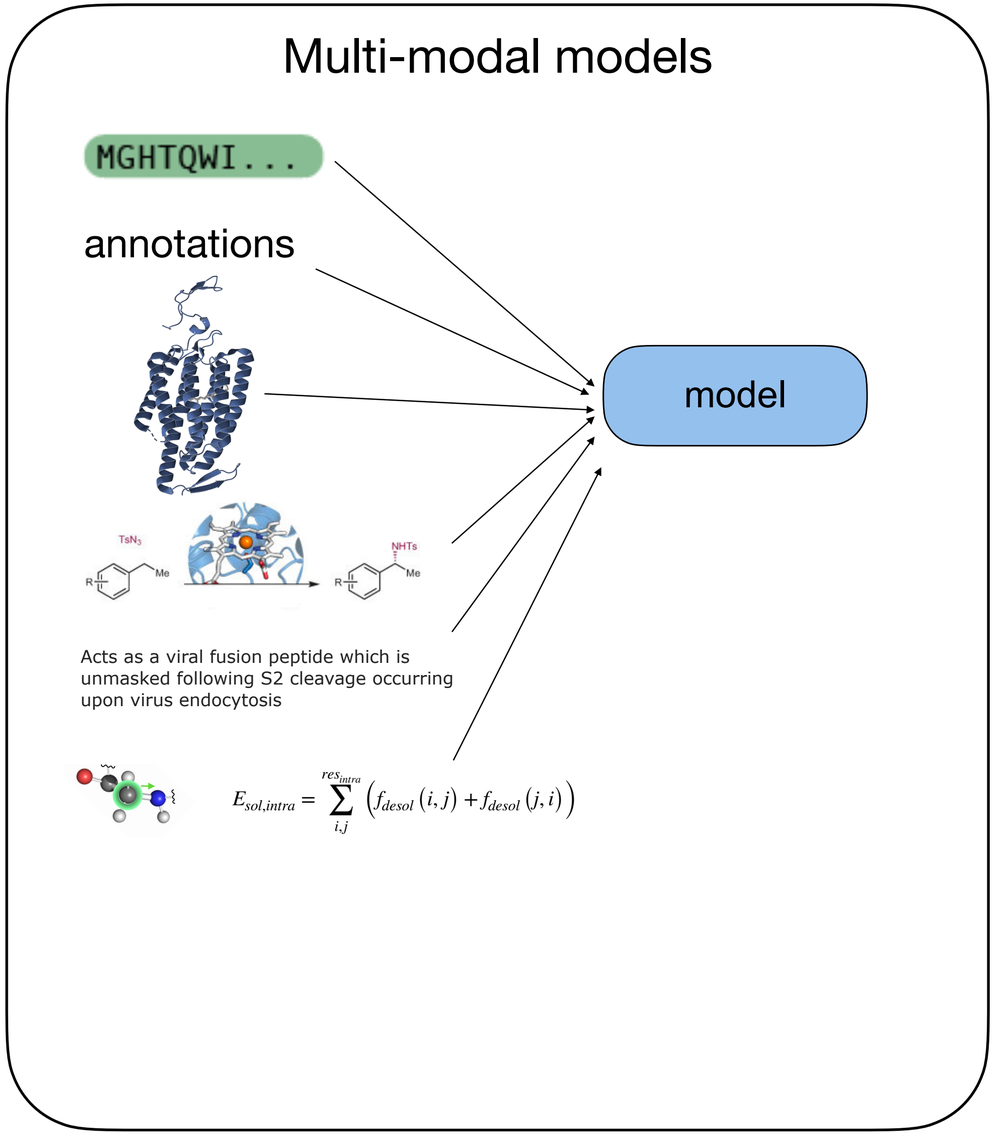
Use multiple data modalities to discover and design proteins



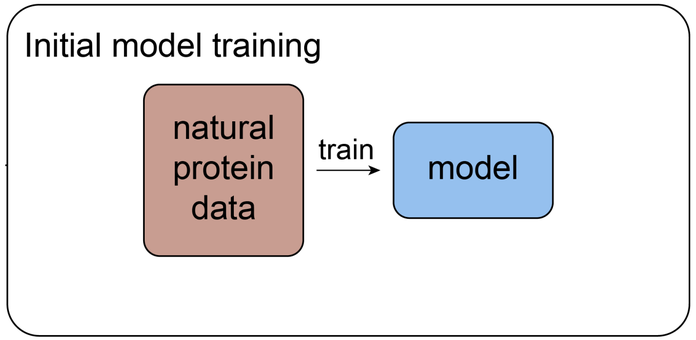


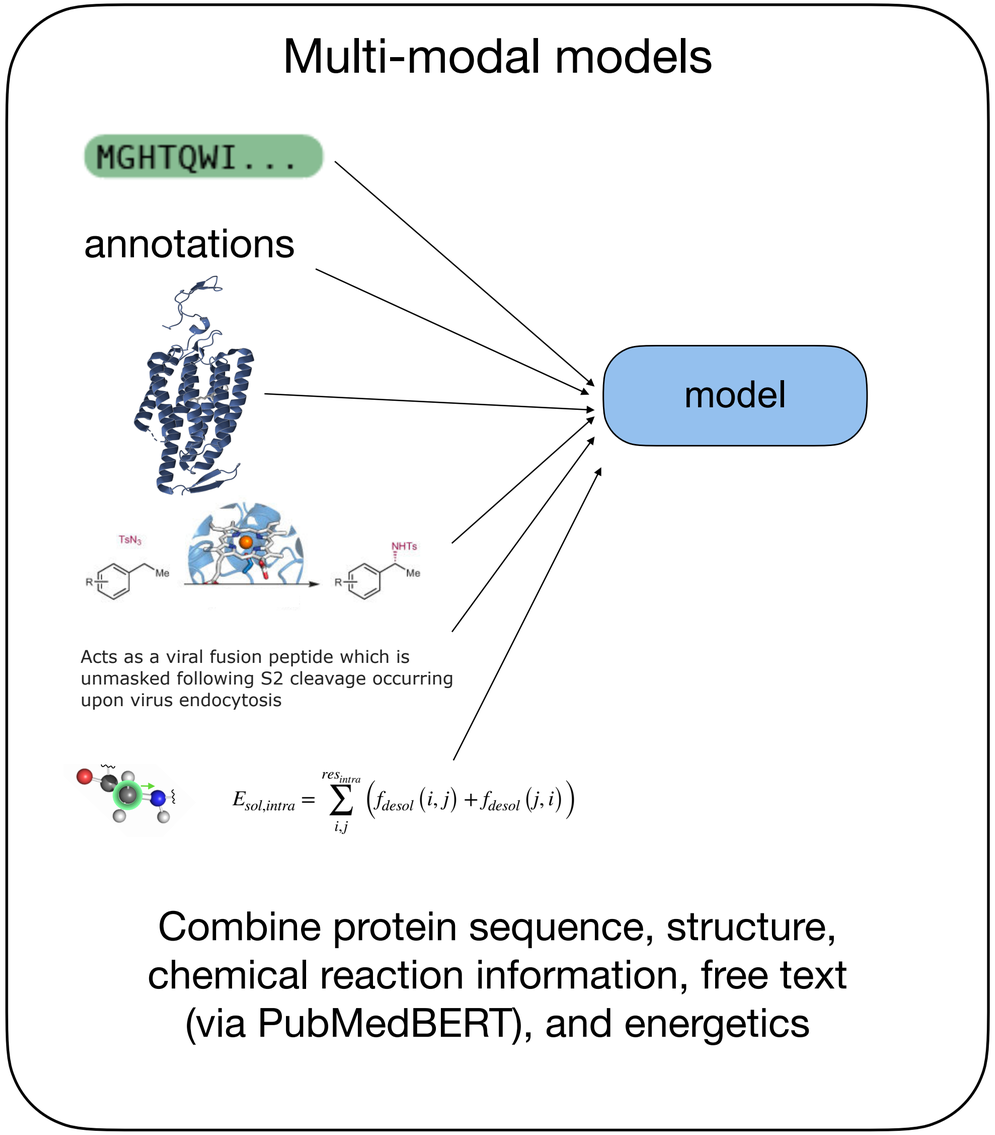
Use multiple data modalities to discover and design proteins



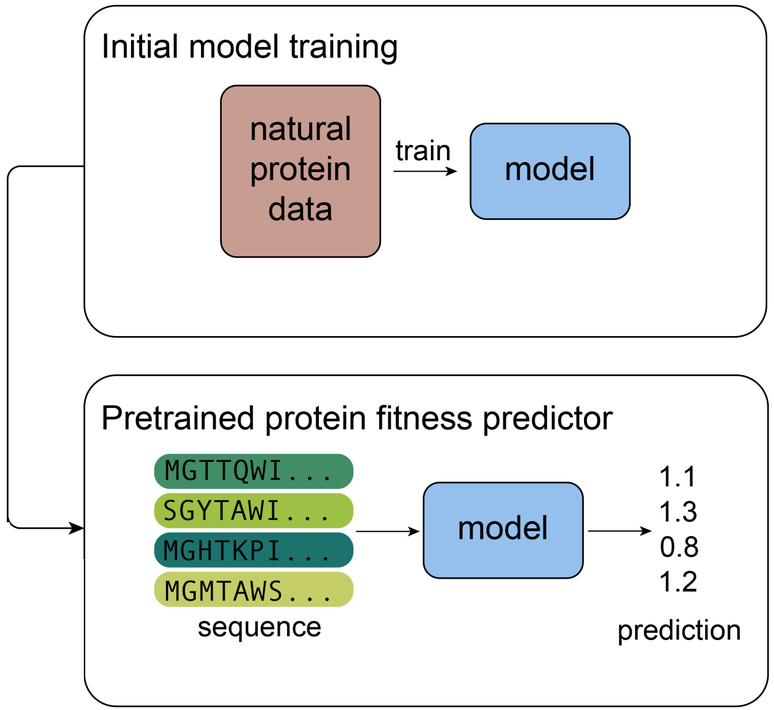


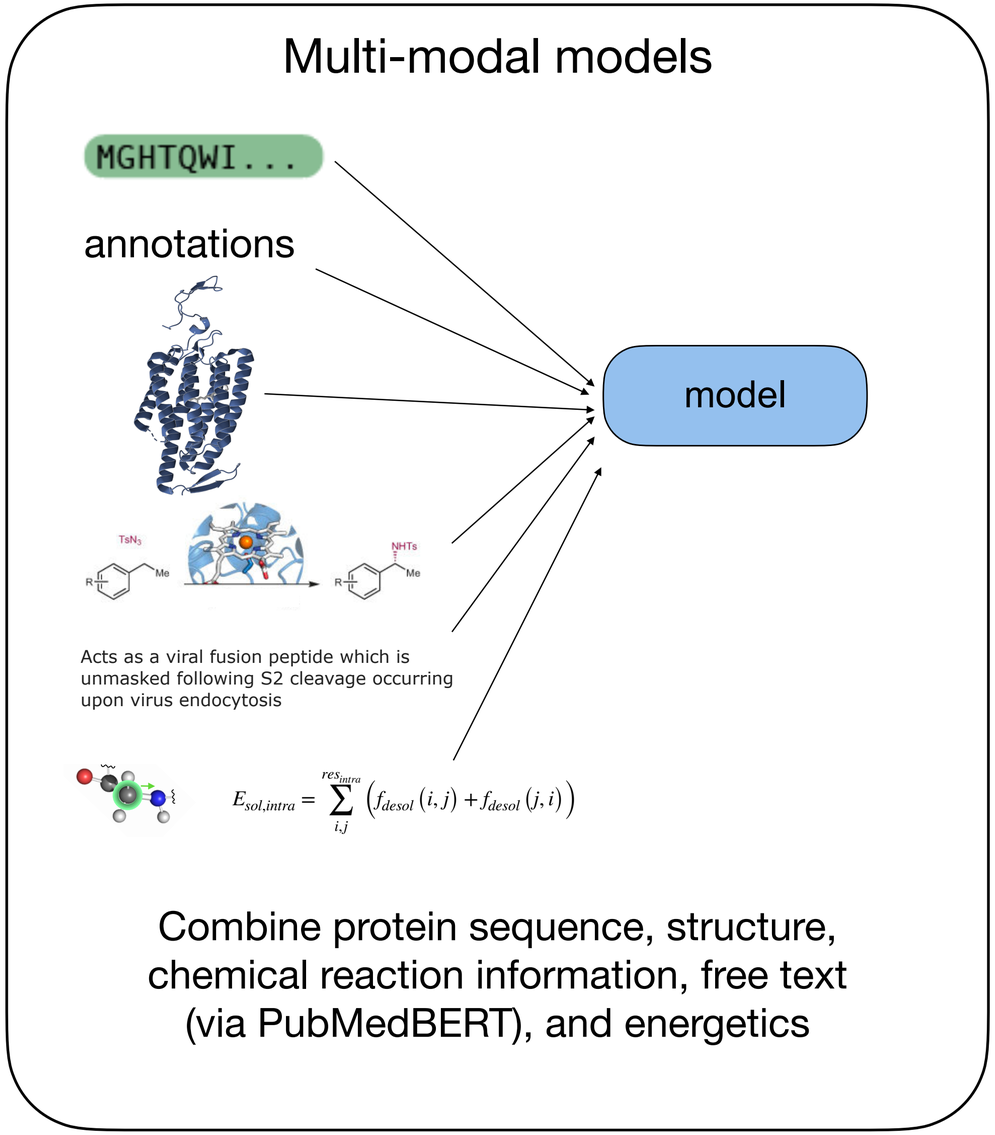
Use multiple data modalities to discover and design proteins



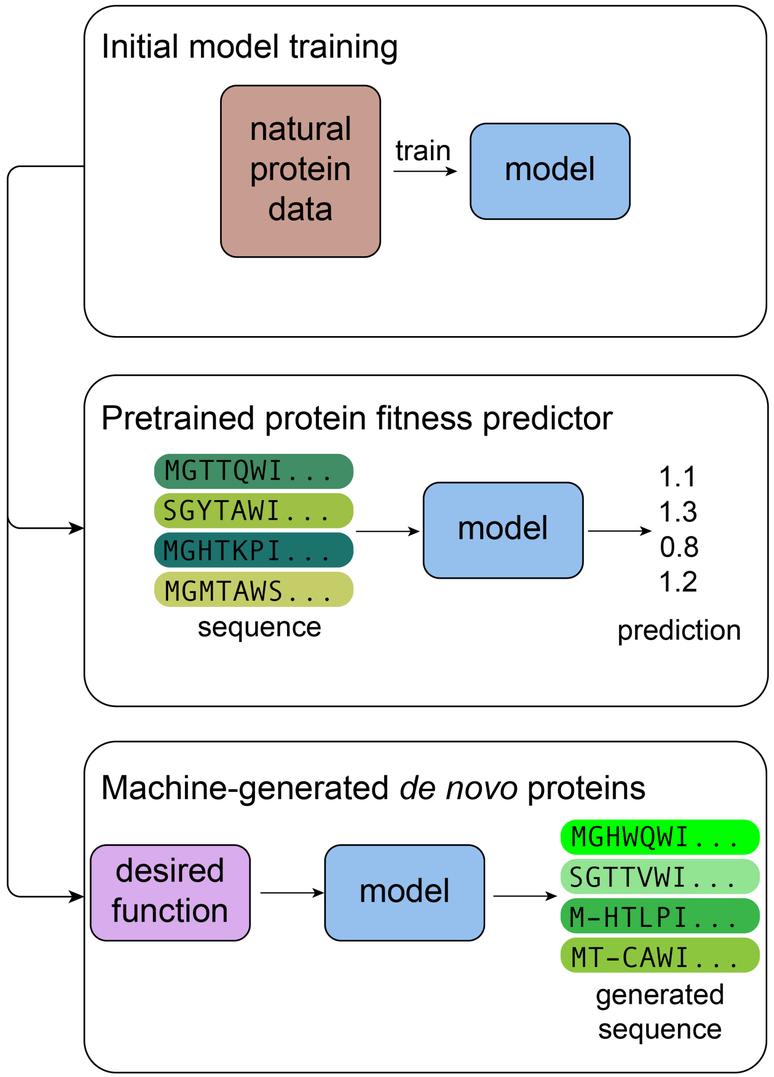


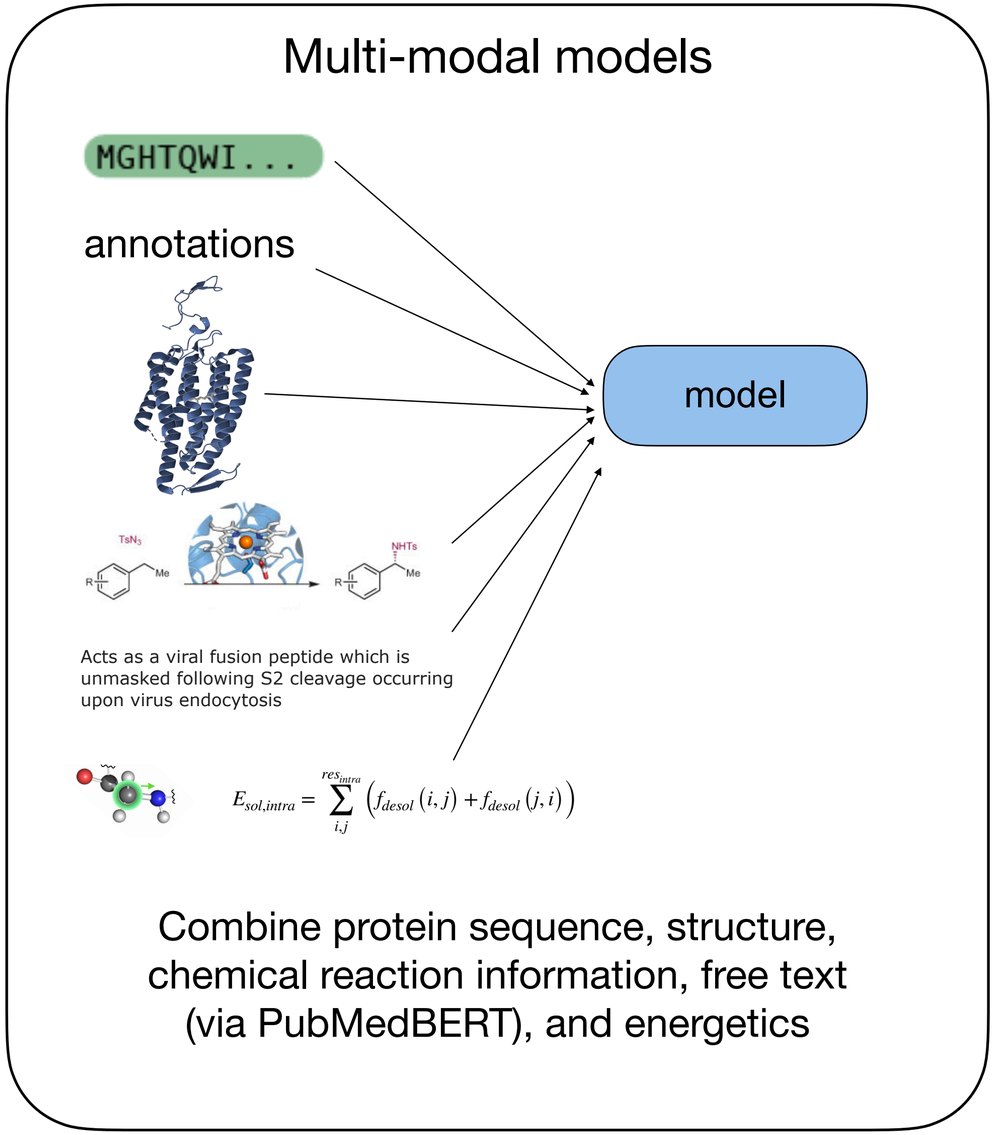
Use multiple data modalities to discover and design proteins



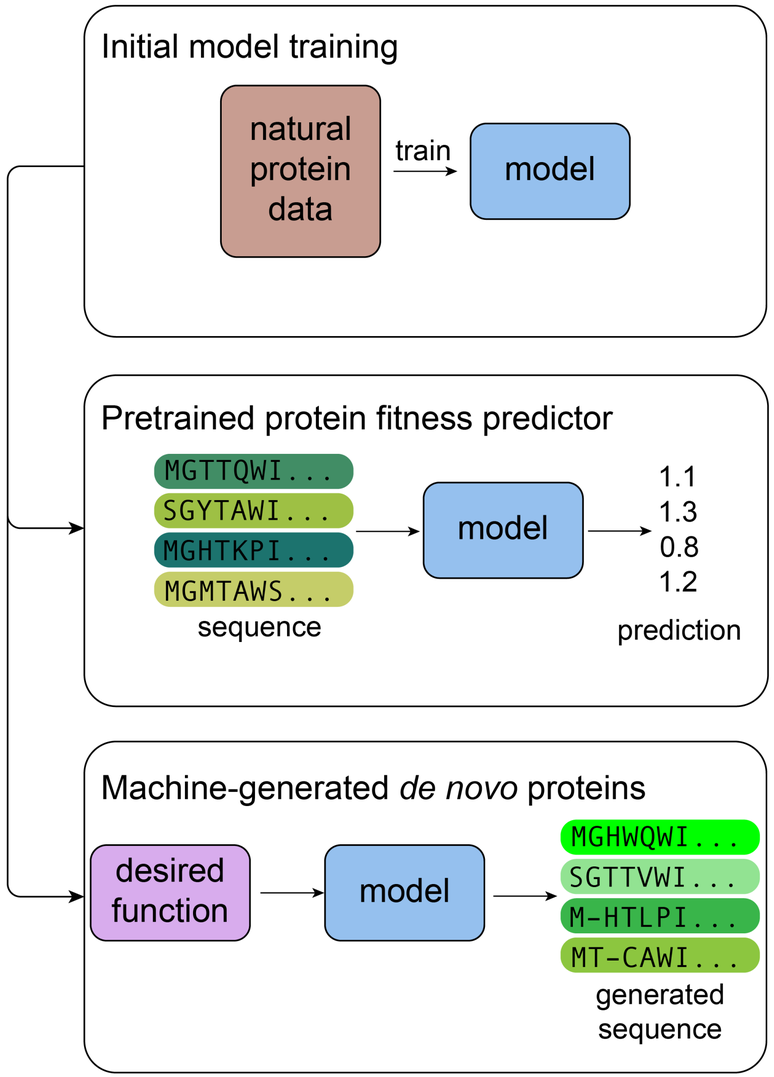


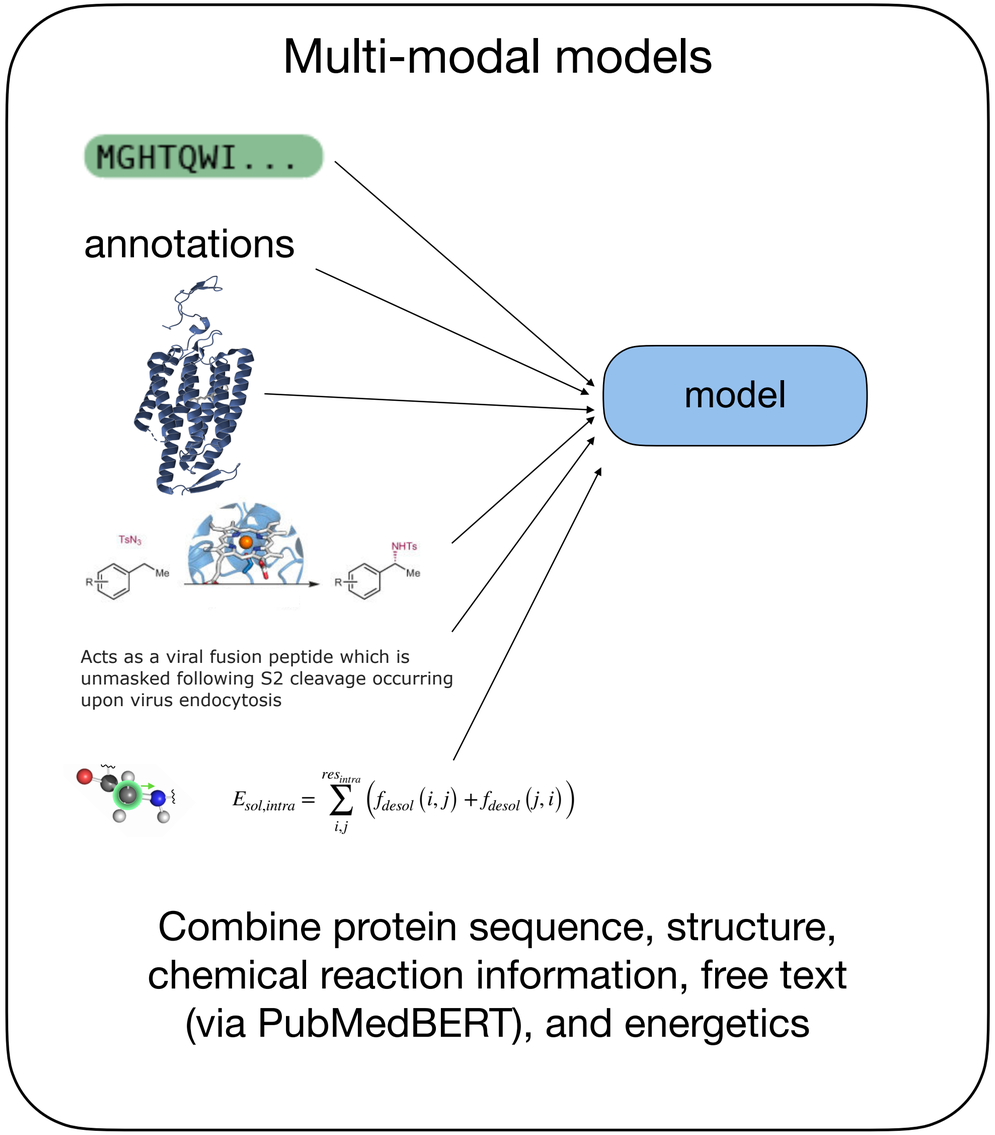
Use multiple data modalities to discover and design proteins



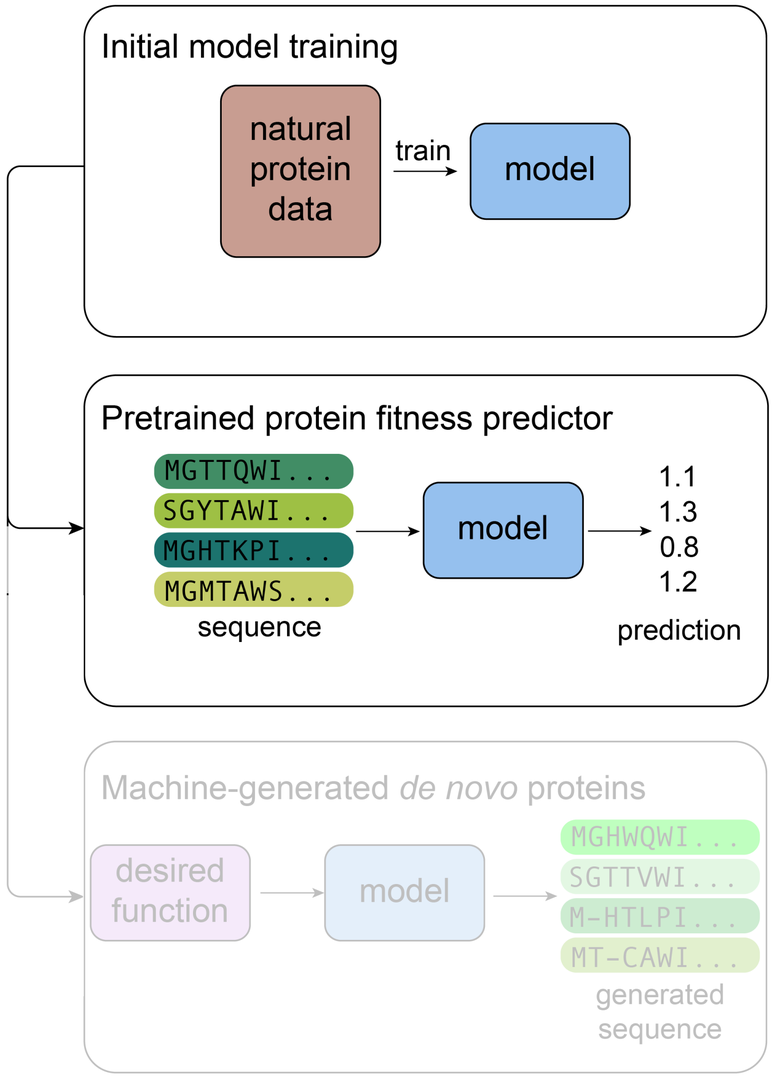


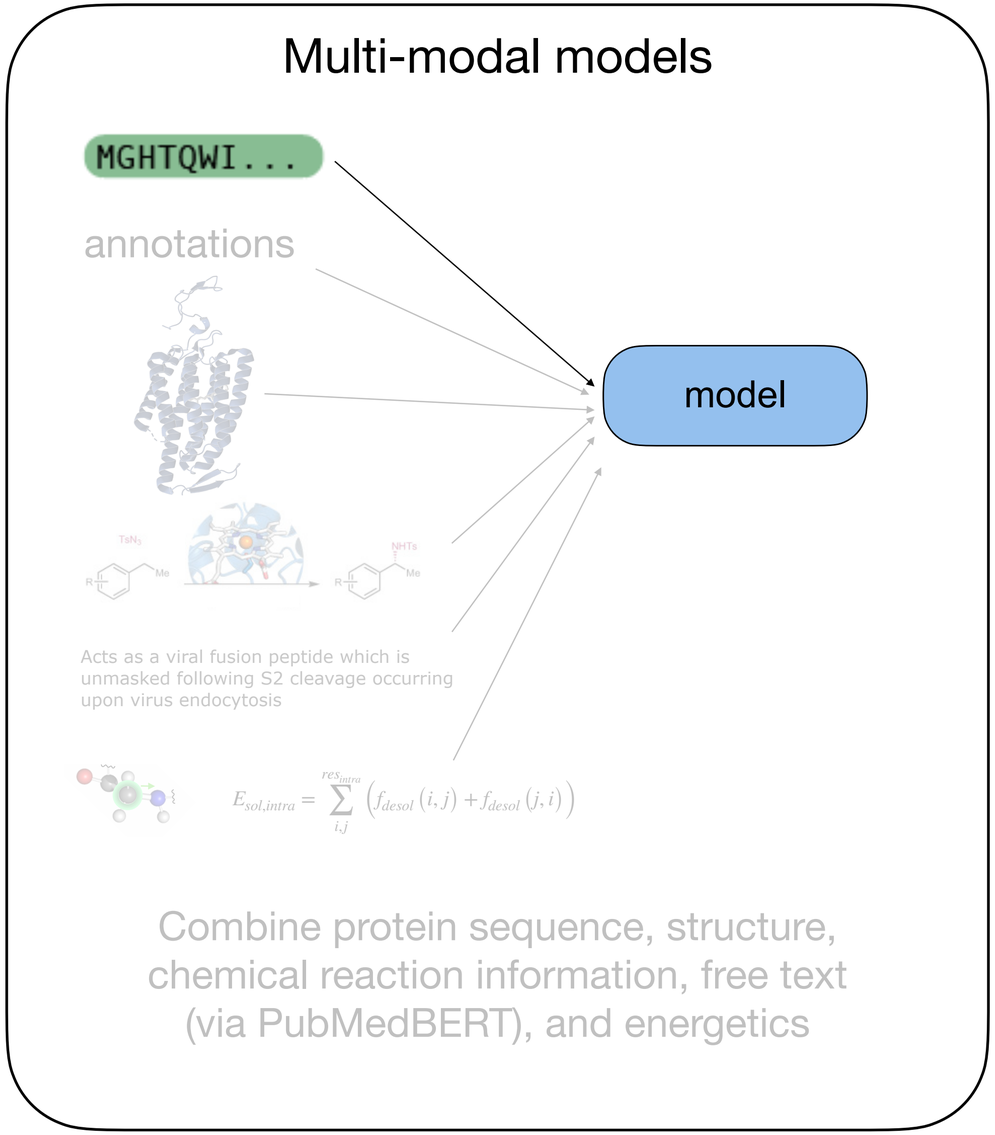
Use multiple data modalities to design proteins



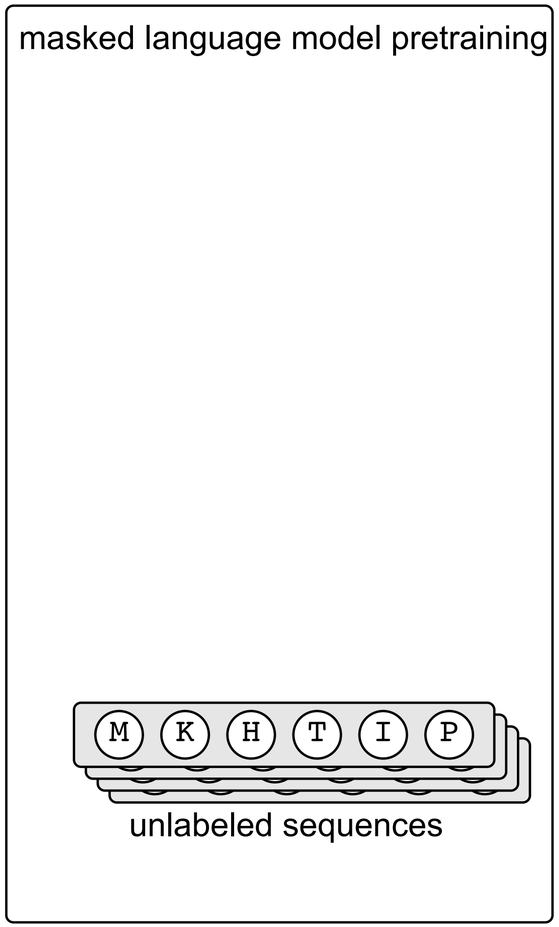


Use multiple data modalities to design proteins

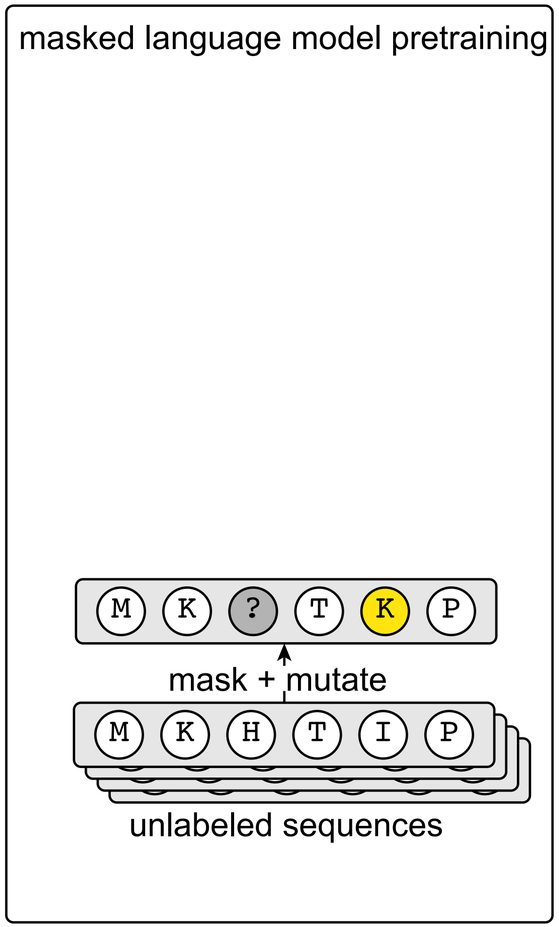




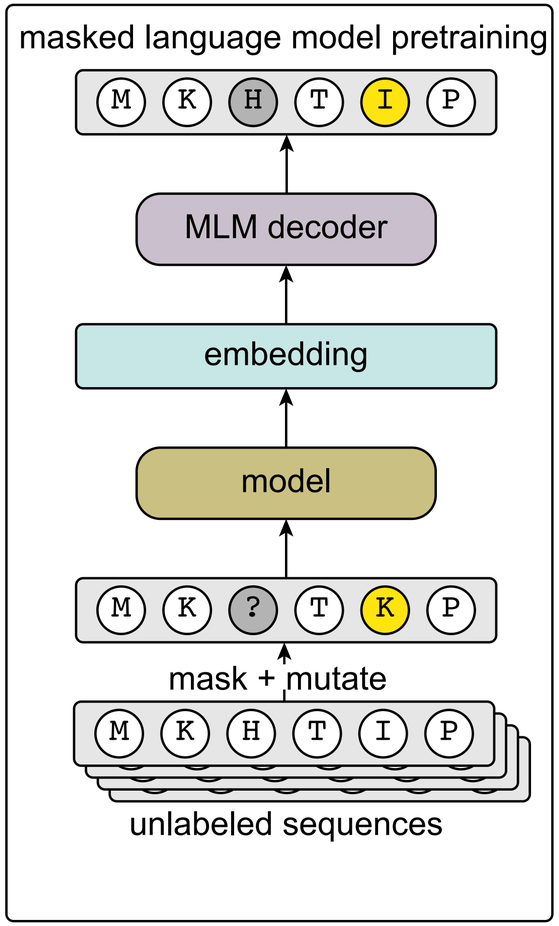
Use pretraining to leverage unlabeled sequences



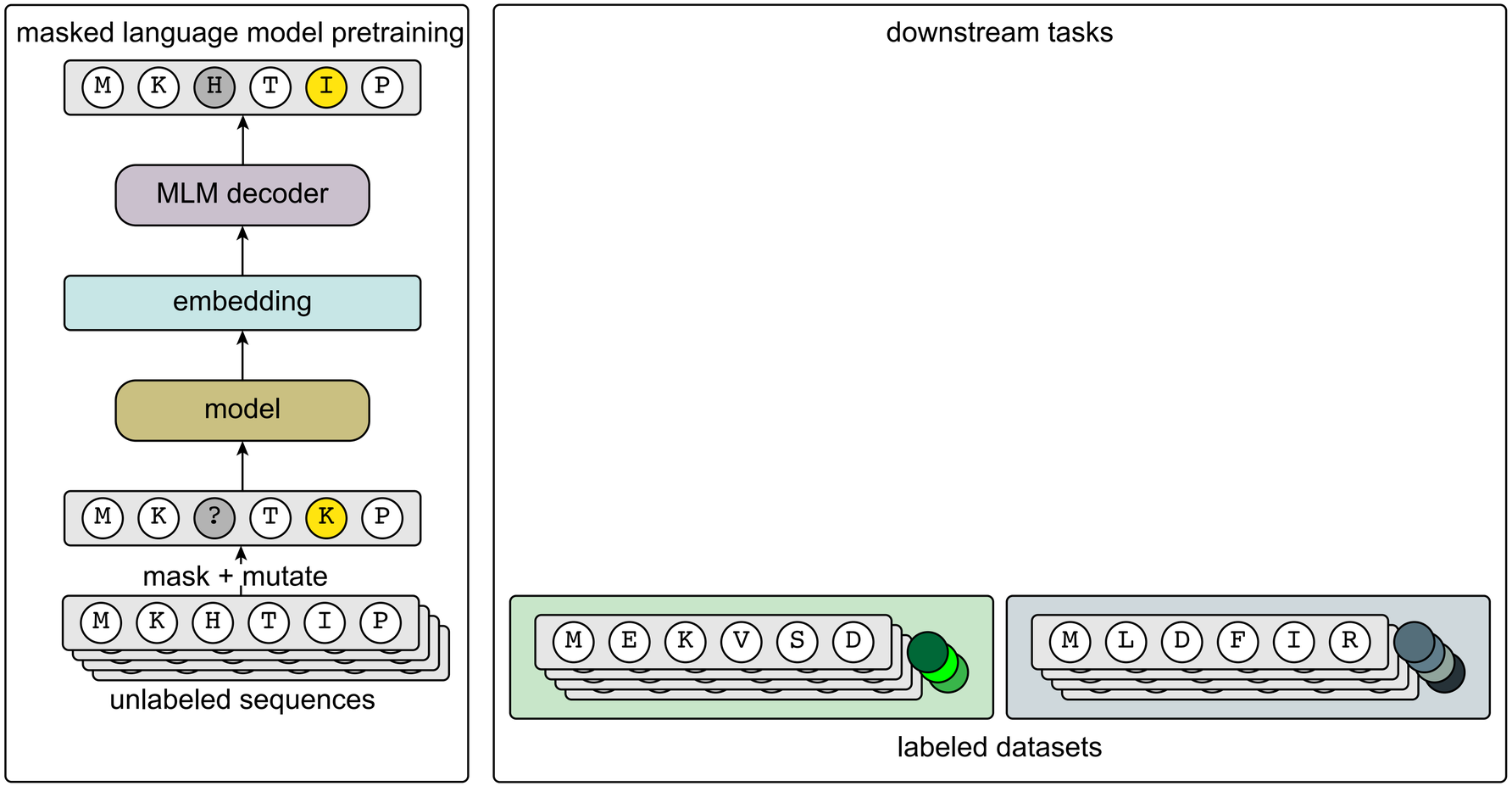
Use pretraining to leverage unlabeled sequences



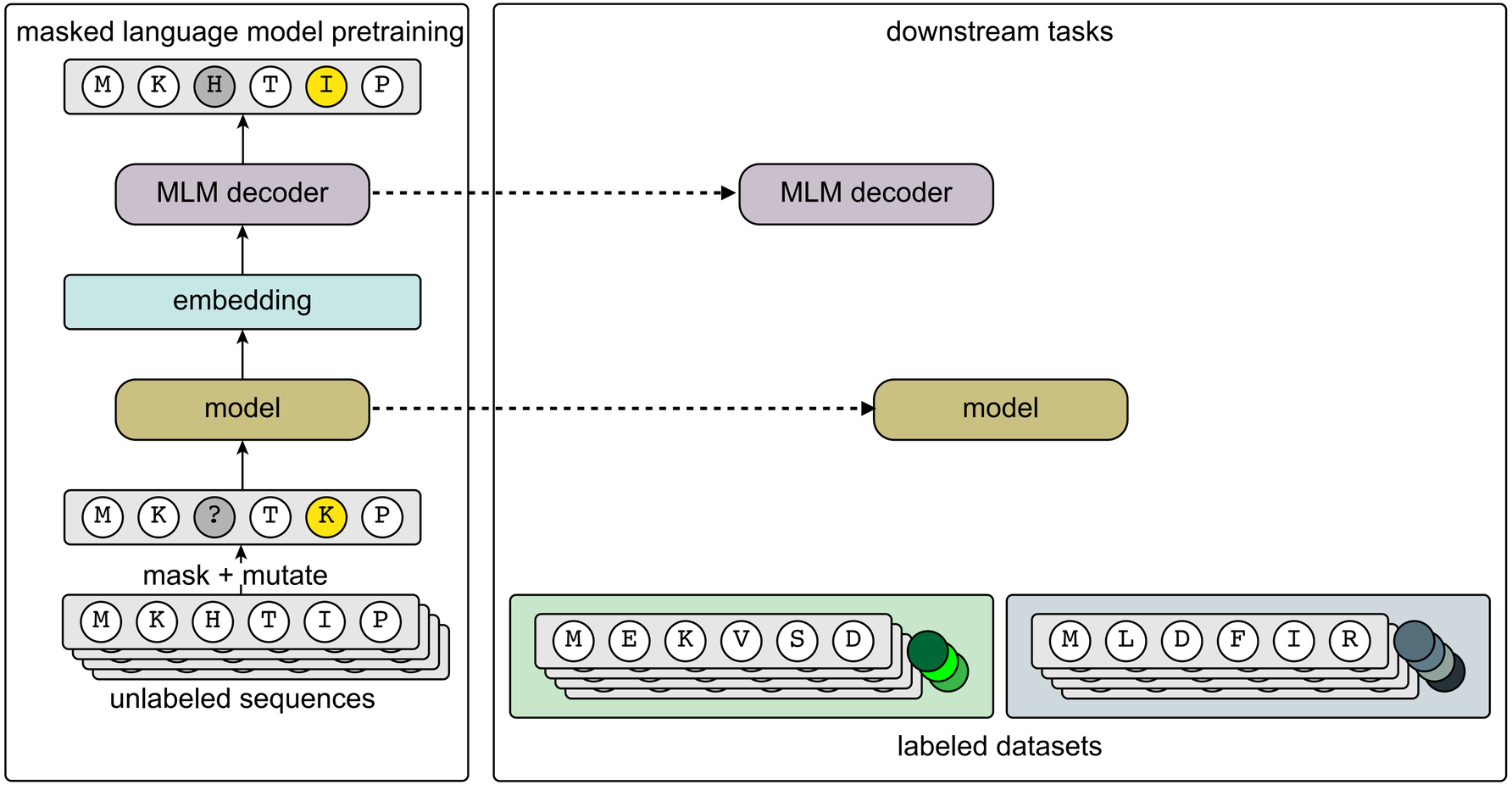
Use pretraining to leverage unlabeled sequences



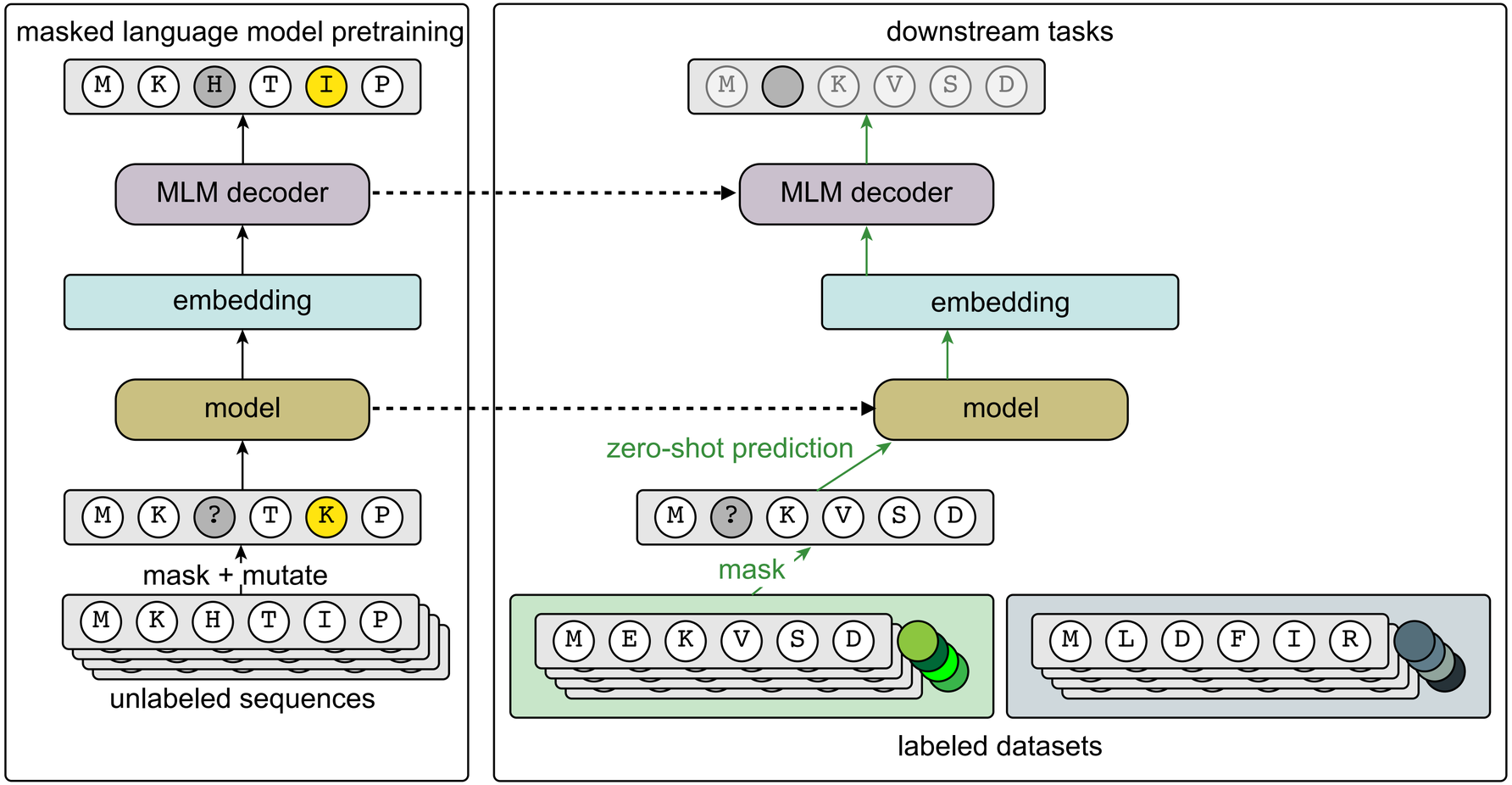
Use pretraining to leverage unlabeled sequences



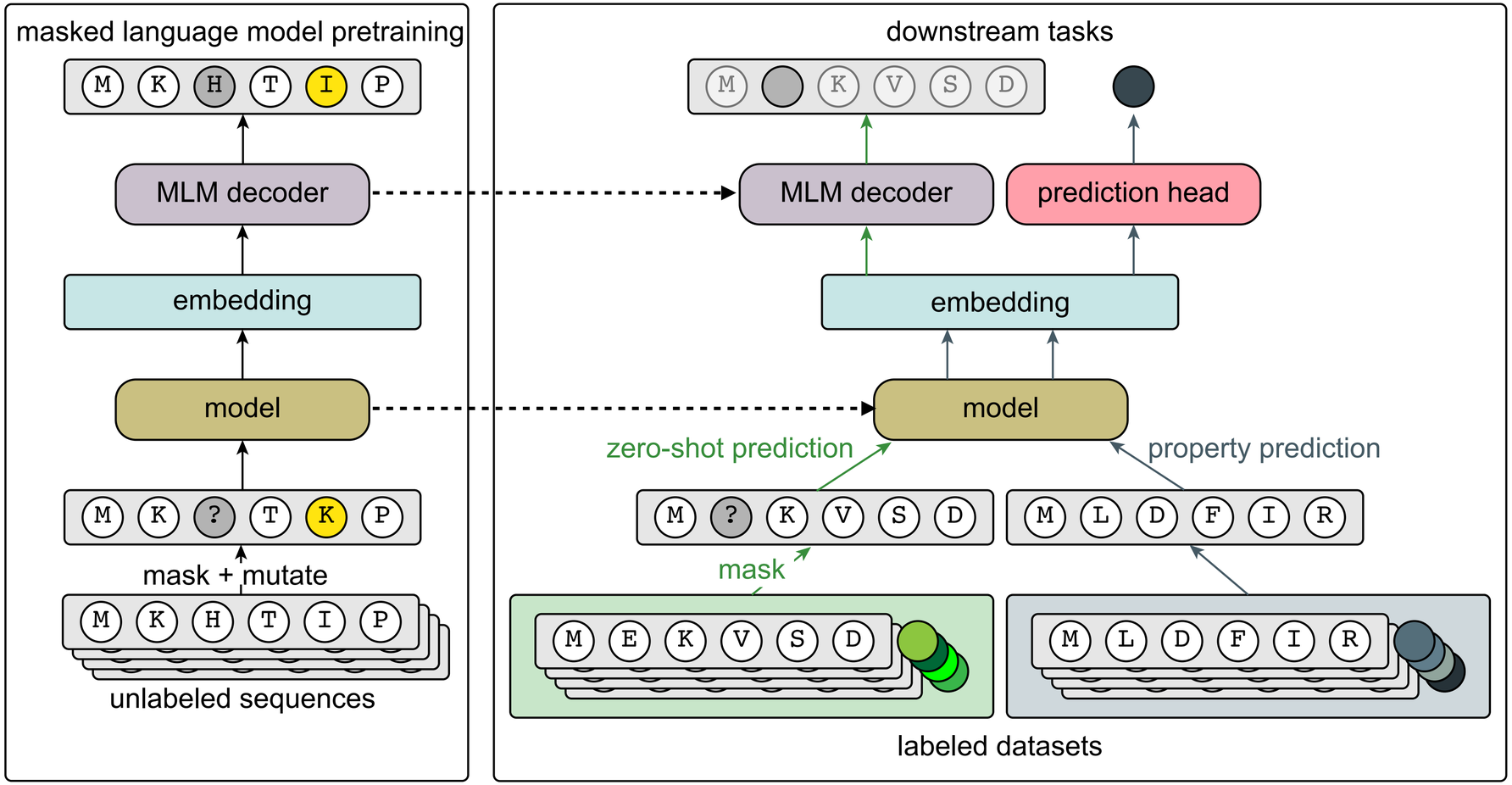
Use pretraining to leverage unlabeled sequences



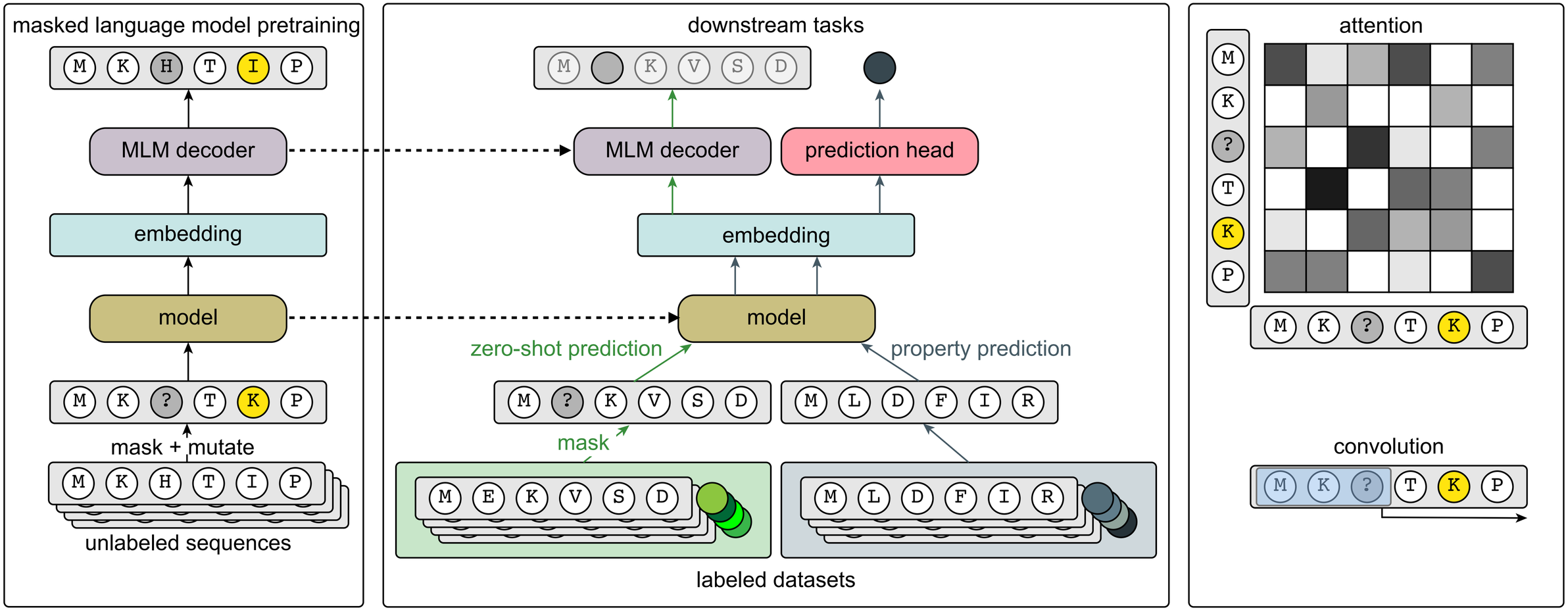
Use pretraining to leverage unlabeled sequences



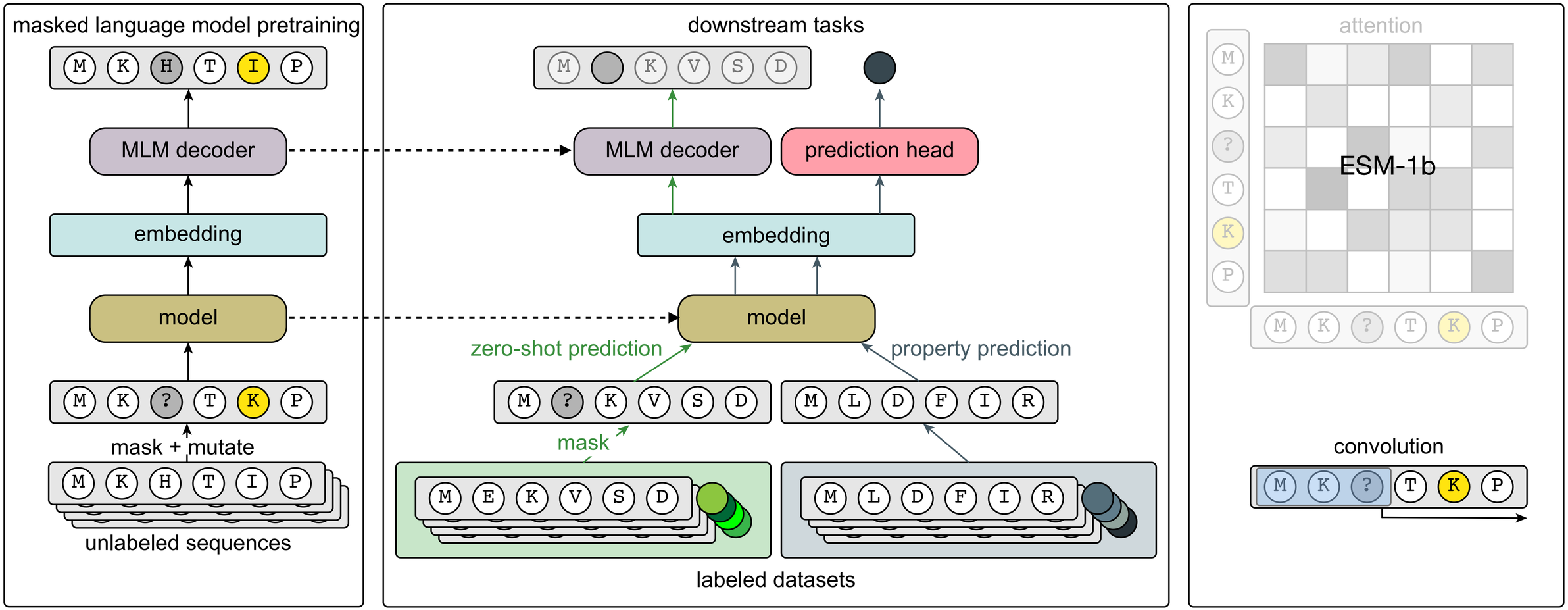
Use pretraining to leverage unlabeled sequences



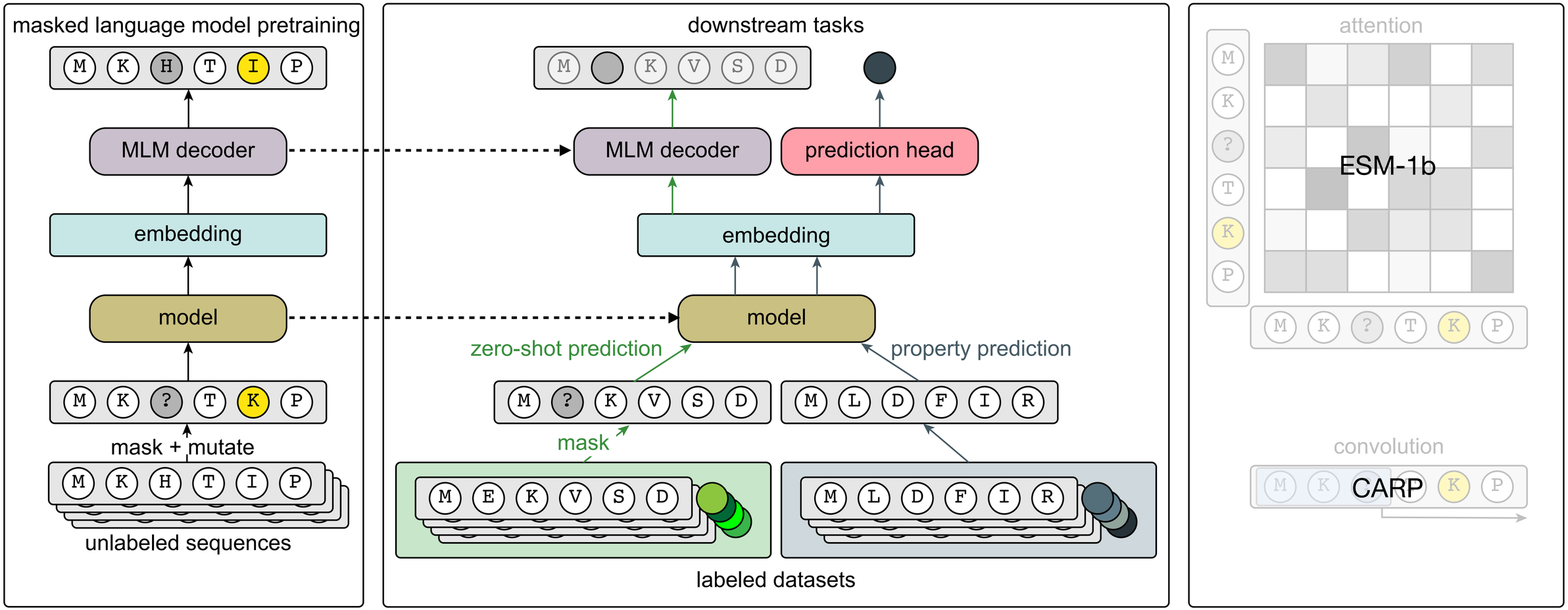
Use pretraining to leverage unlabeled sequences



Use pretraining to leverage unlabeled sequences

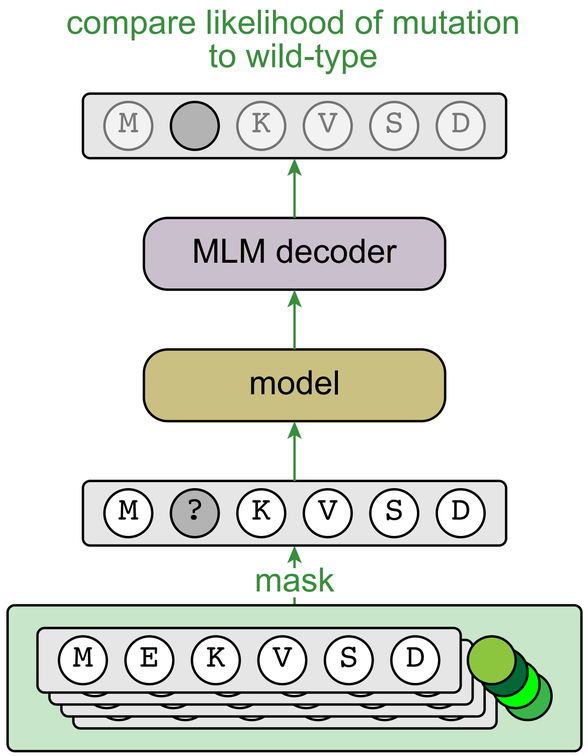


Use pretraining to leverage unlabeled sequences

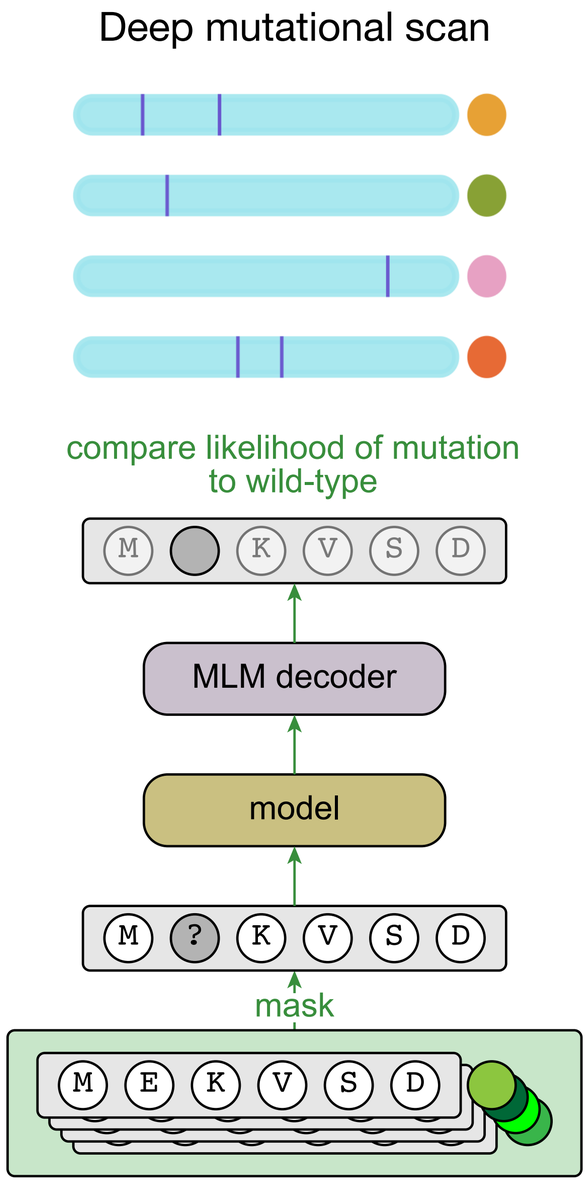


Pretrained models are zero-shot fitness predictors

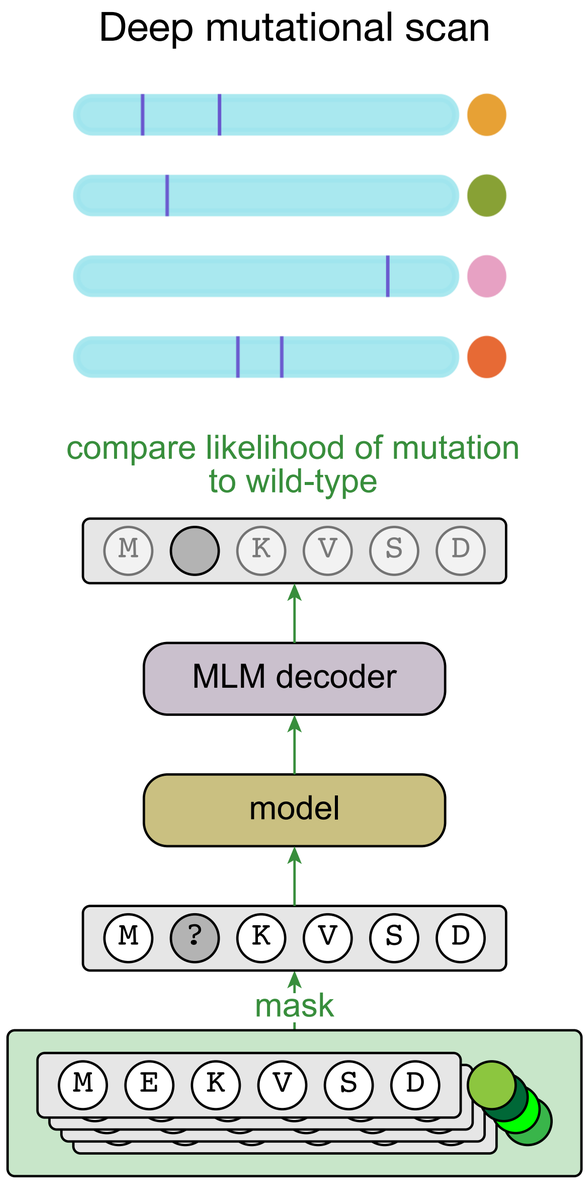
Pretrained models are zero-shot fitness predictors

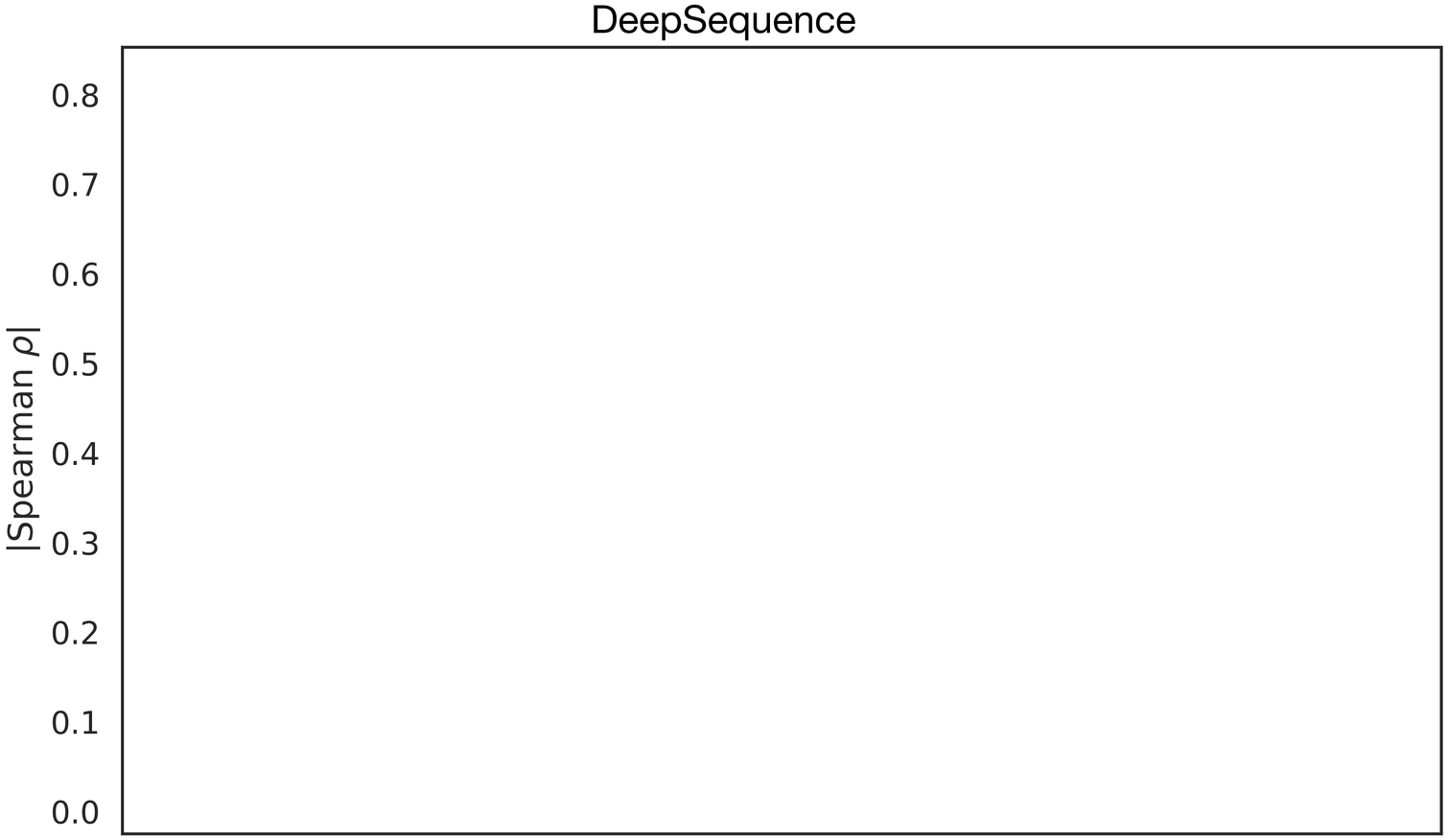


Pretrained models are zero-shot fitness predictors

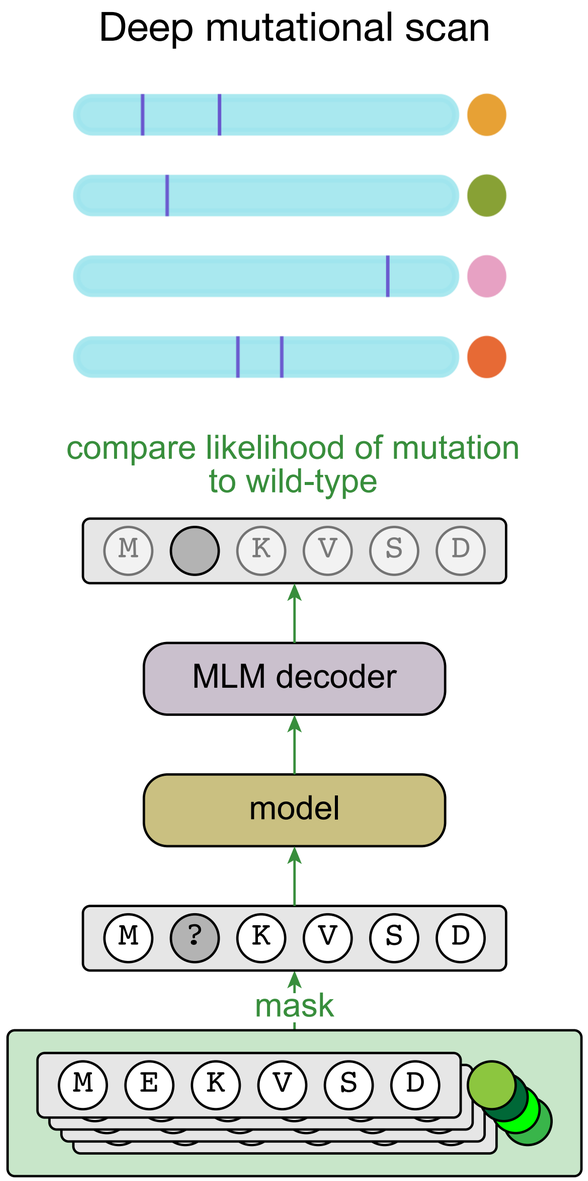


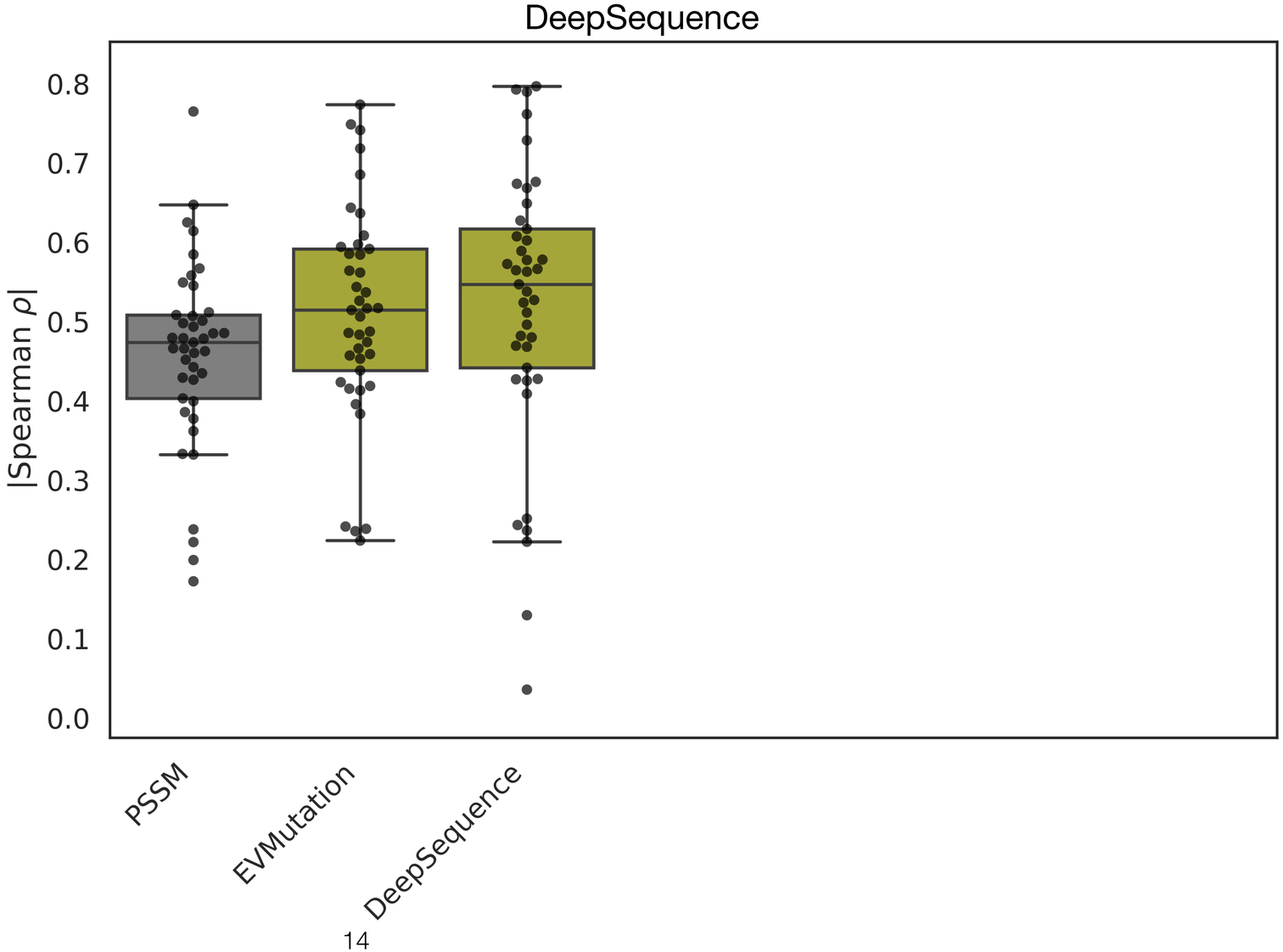
Pretrained models are zero-shot fitness predictors



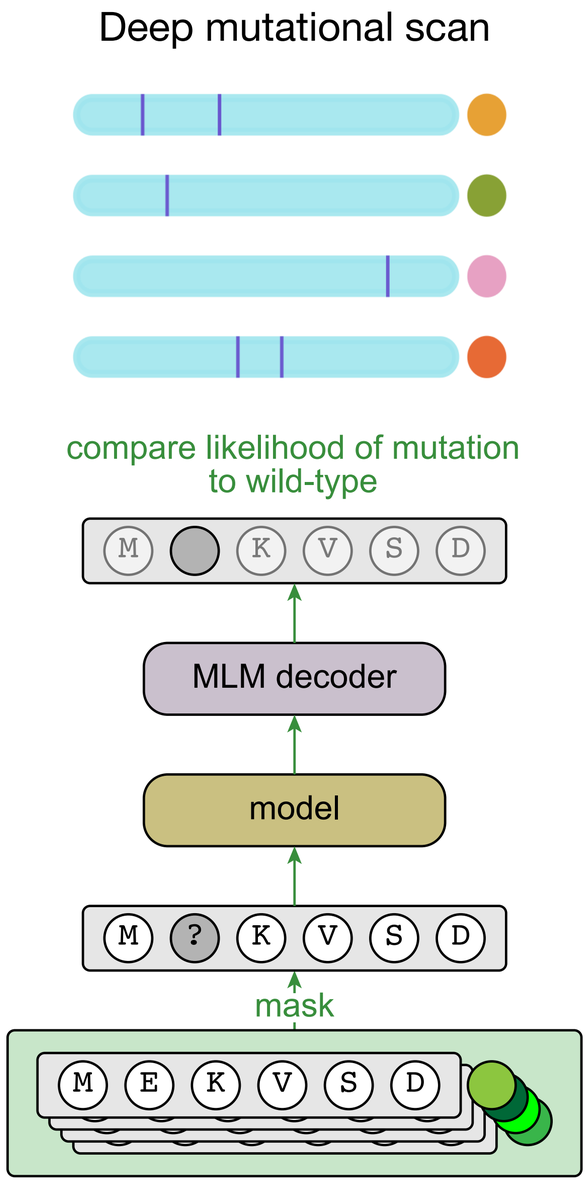


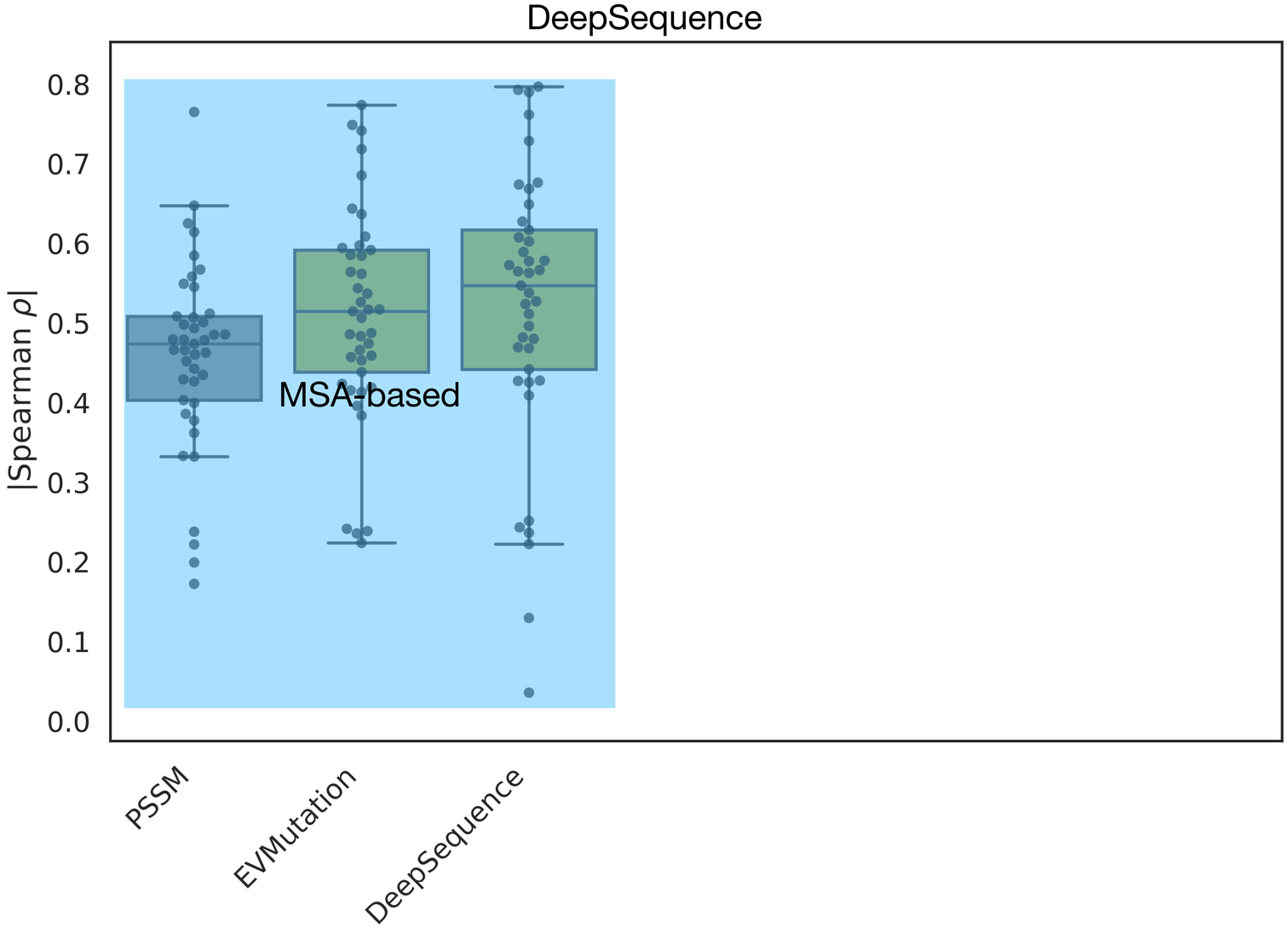
Pretrained models are zero-shot fitness predictors



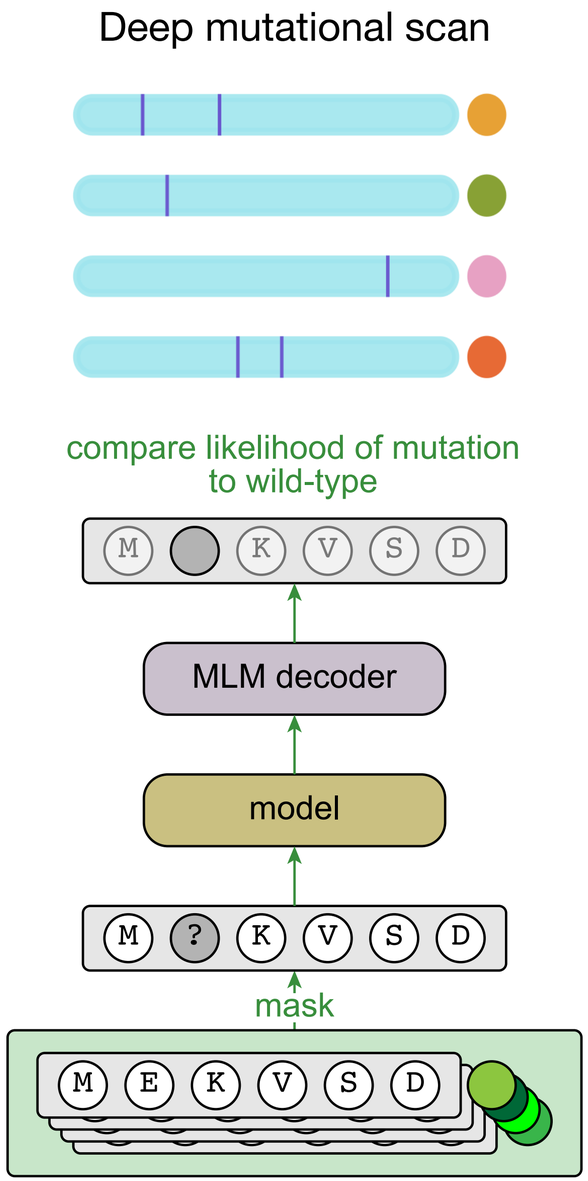


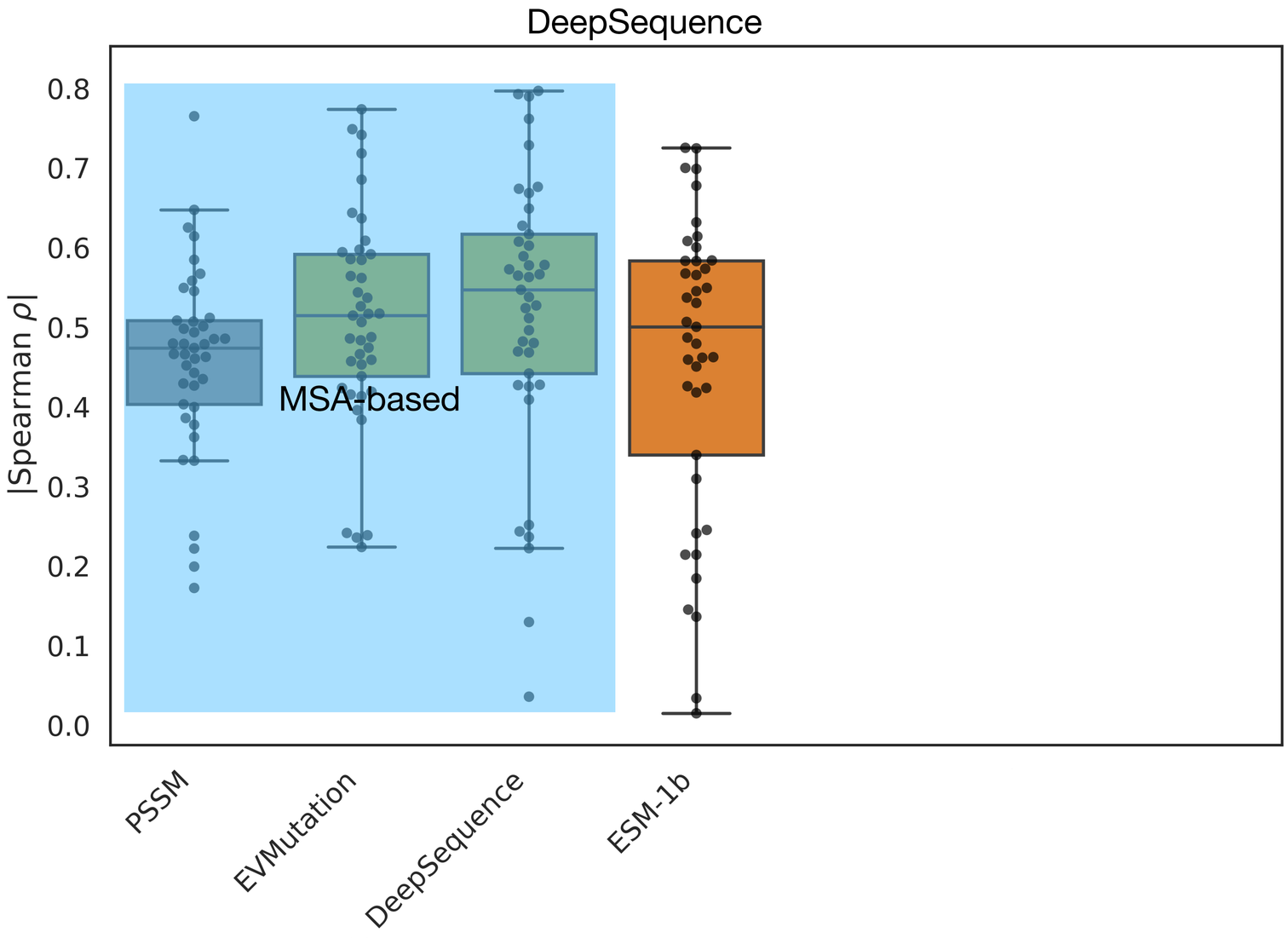
Pretrained models are zero-shot fitness predictors



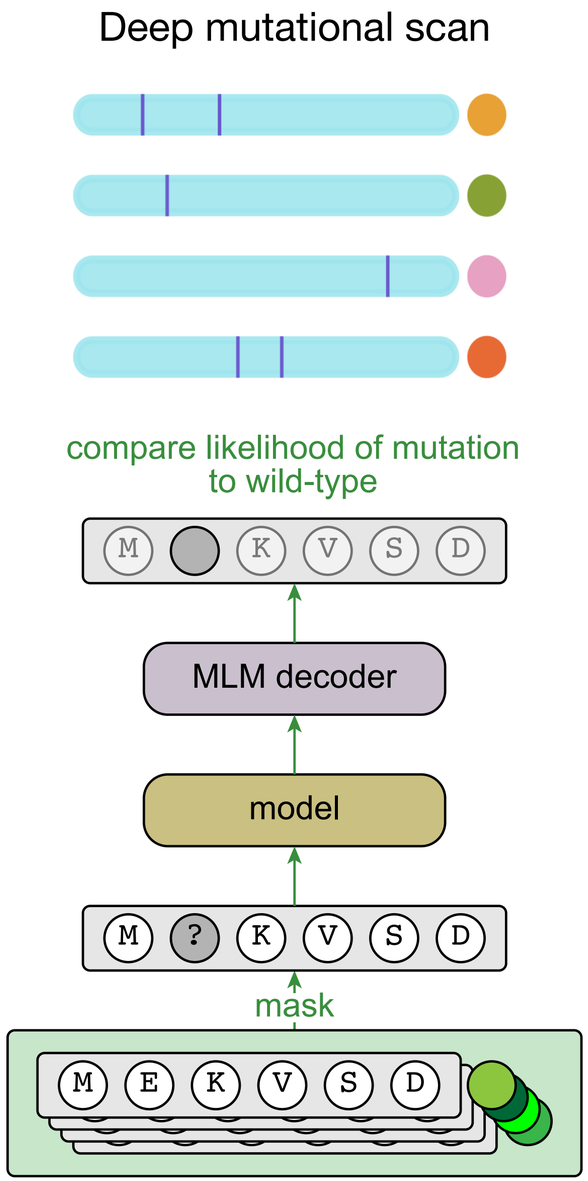


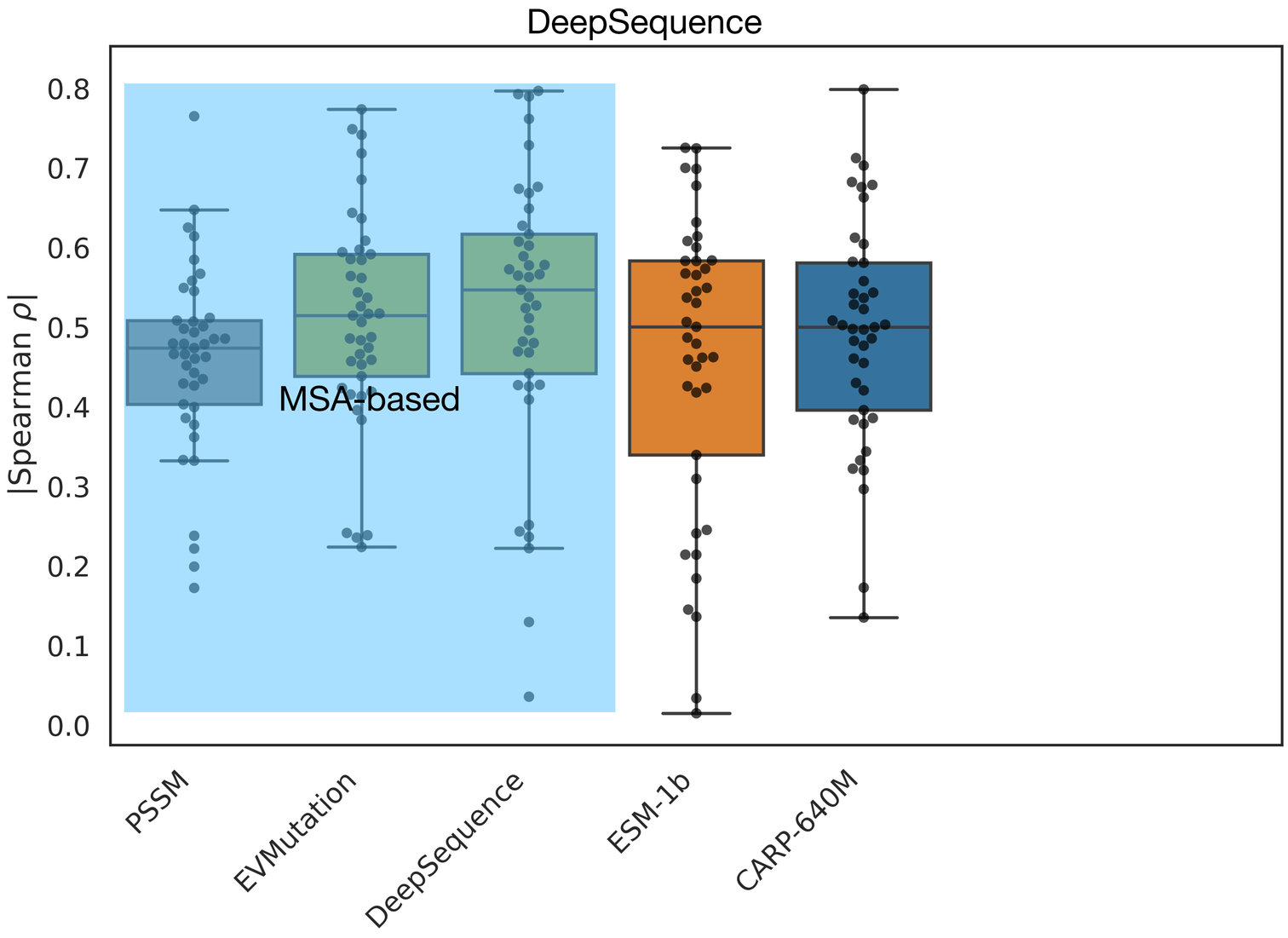
Pretrained models are zero-shot fitness predictors





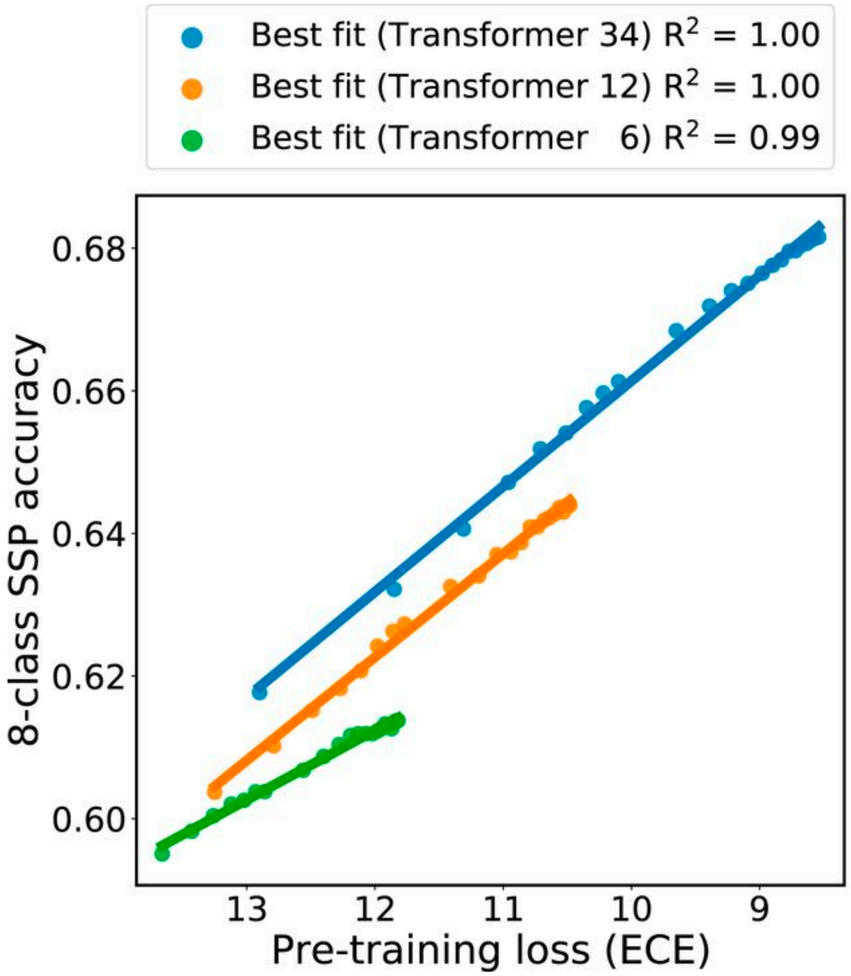
Pretrained models are zero-shot fitness predictors



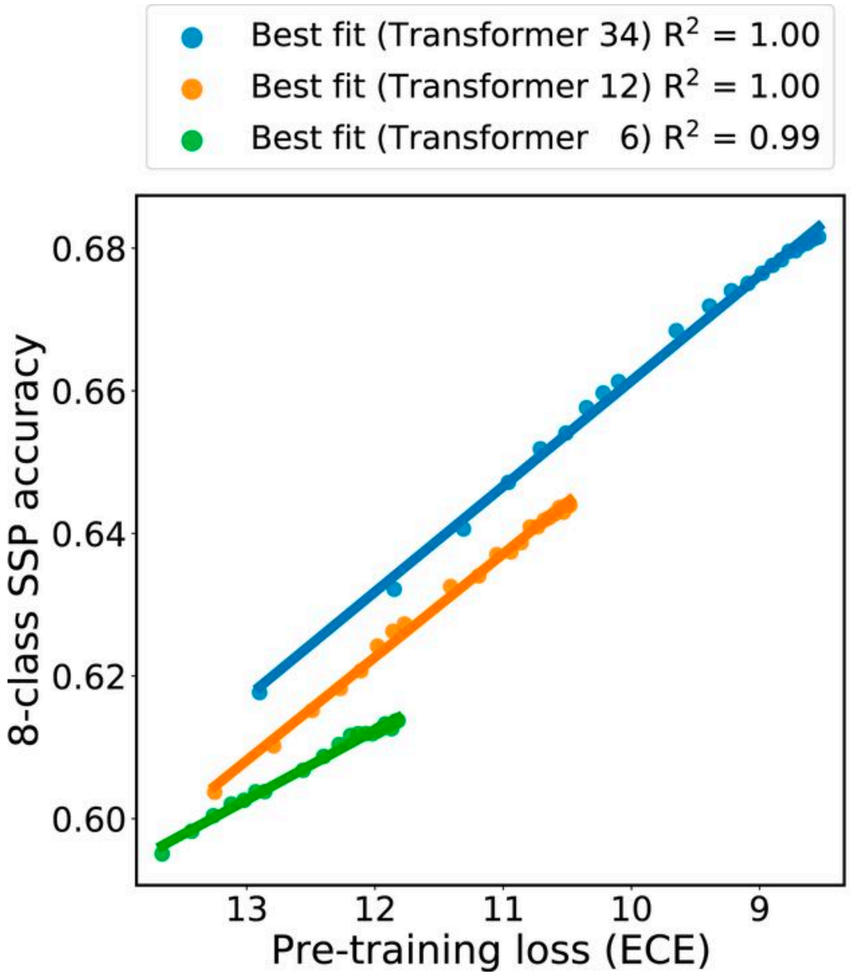


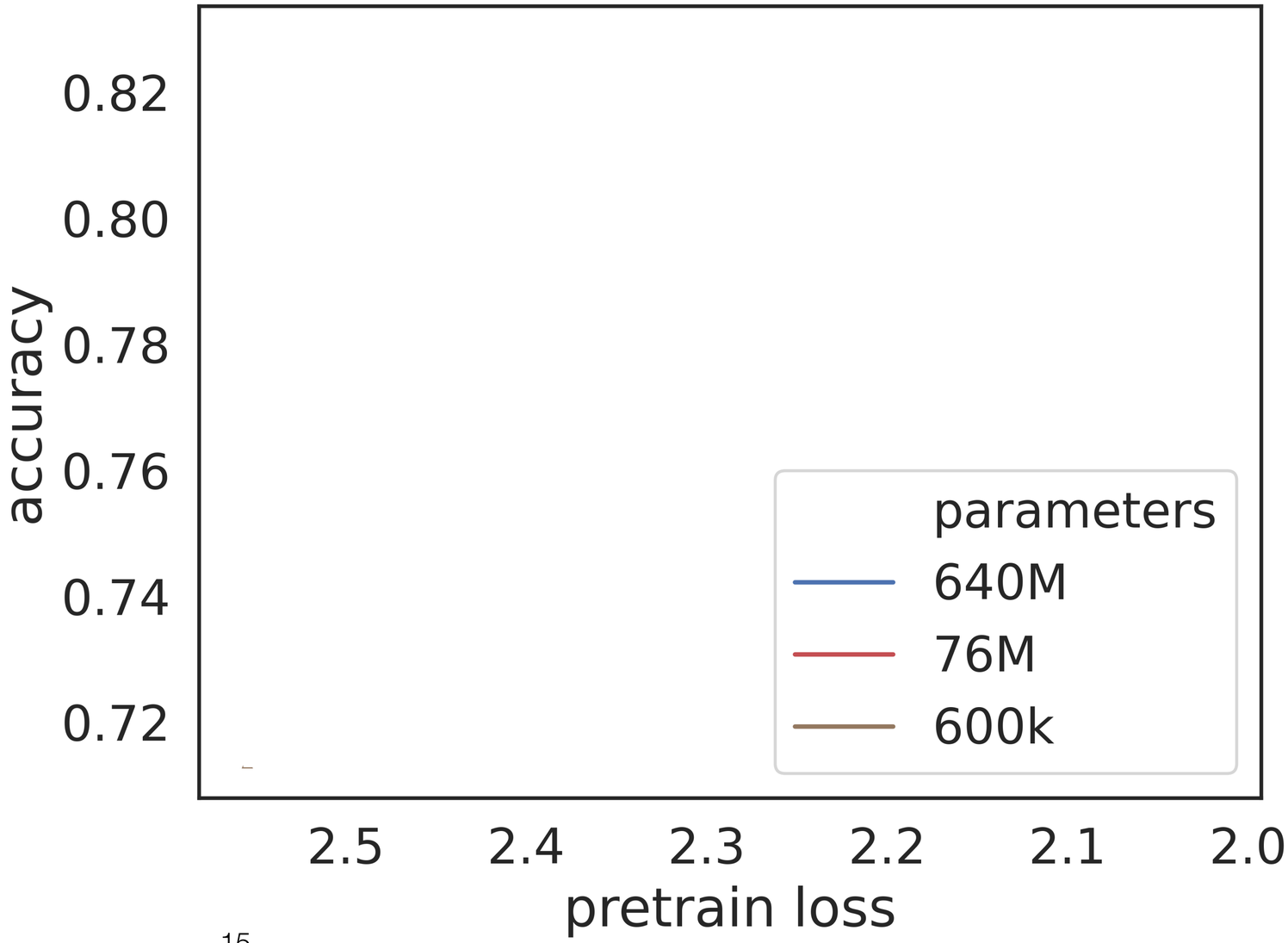
Structure predictions improve smoothly with pretraining

Structure predictions improve smoothly with pretraining

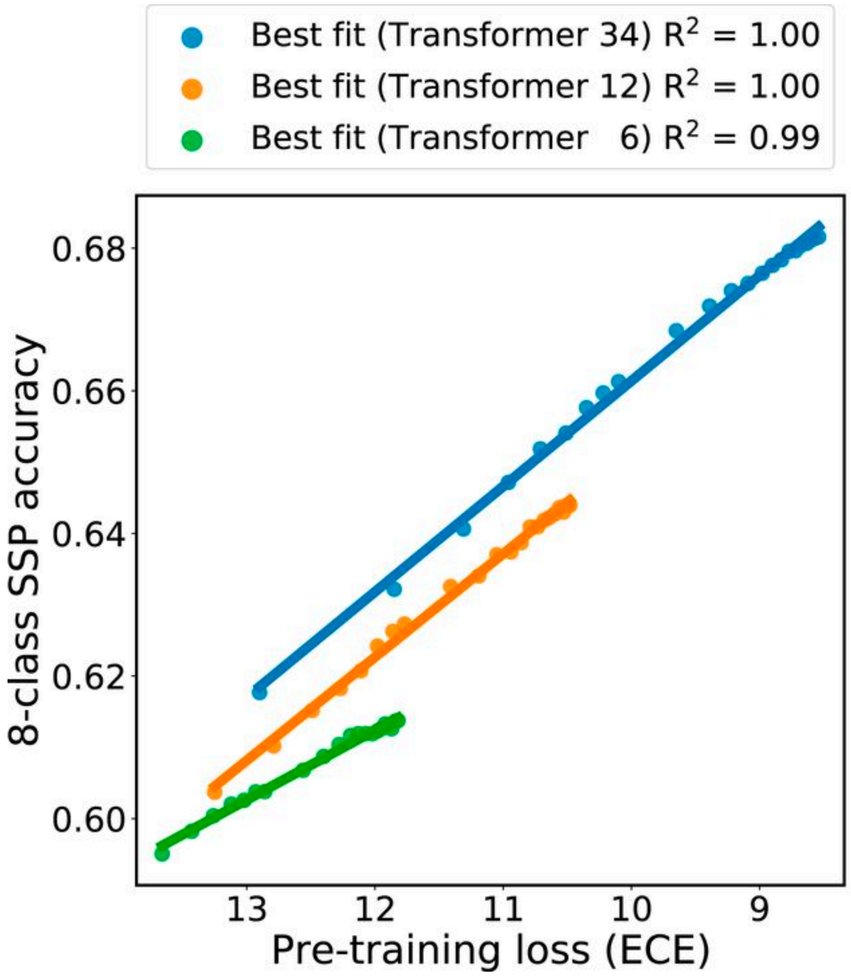


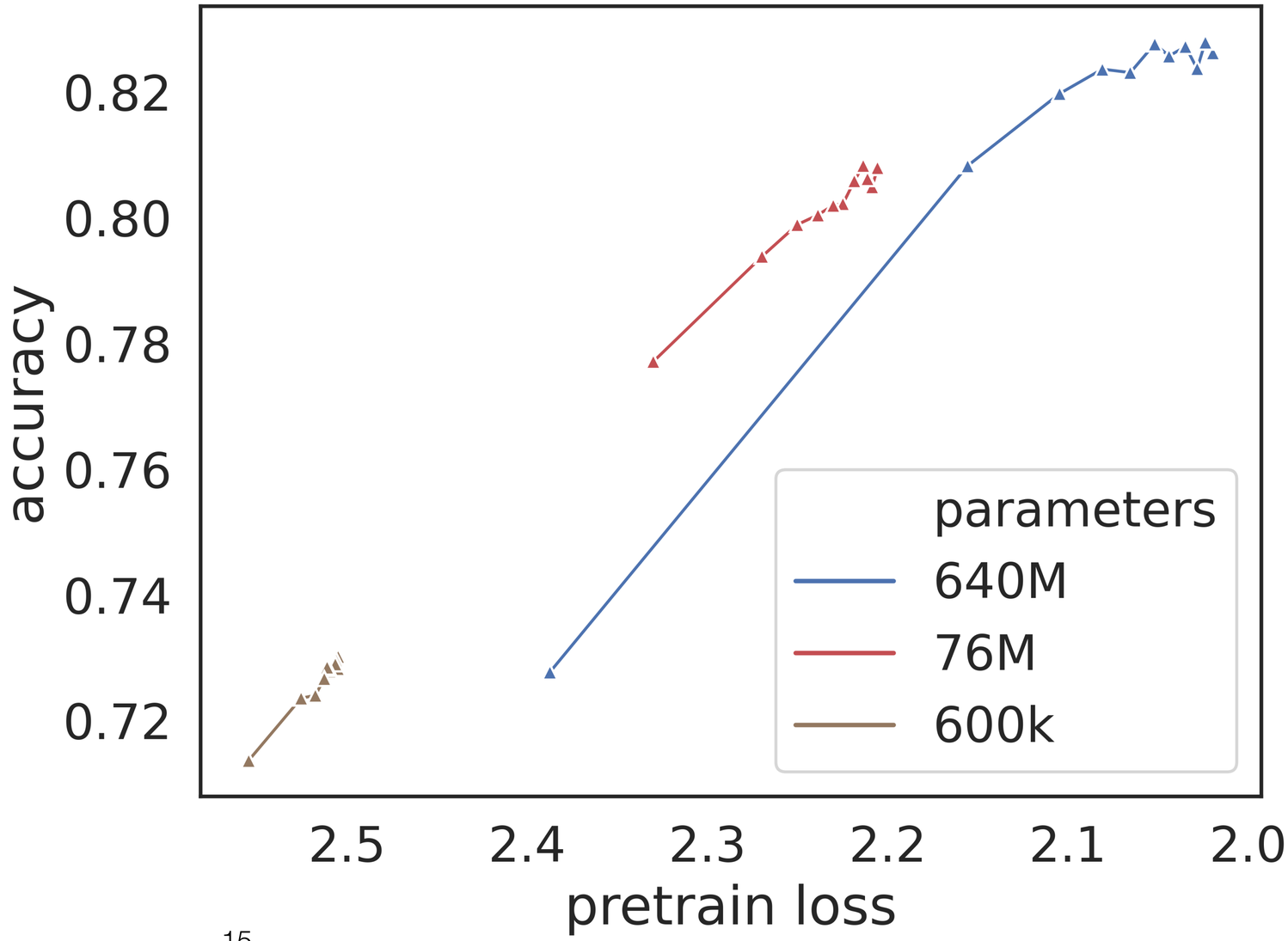
Structure predictions improve smoothly with pretraining



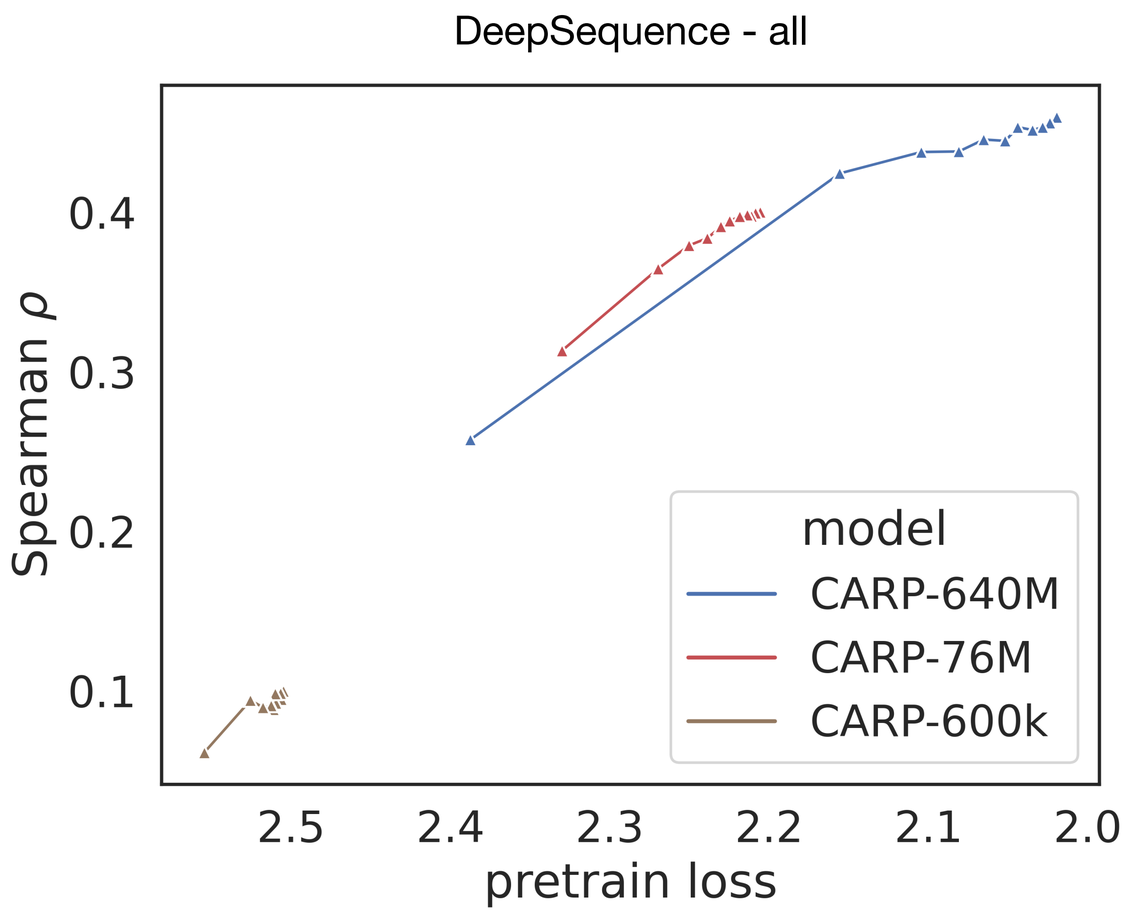


Structure predictions improve smoothly with pretraining

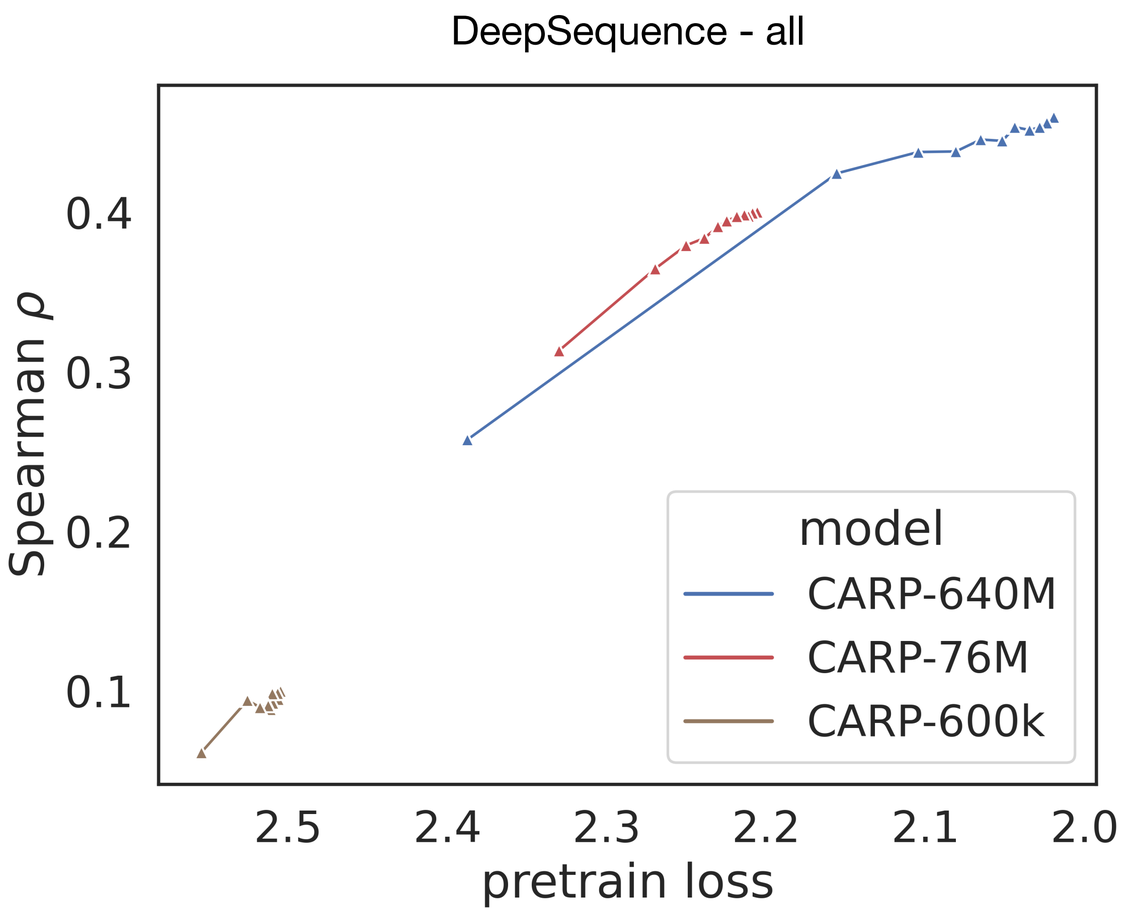


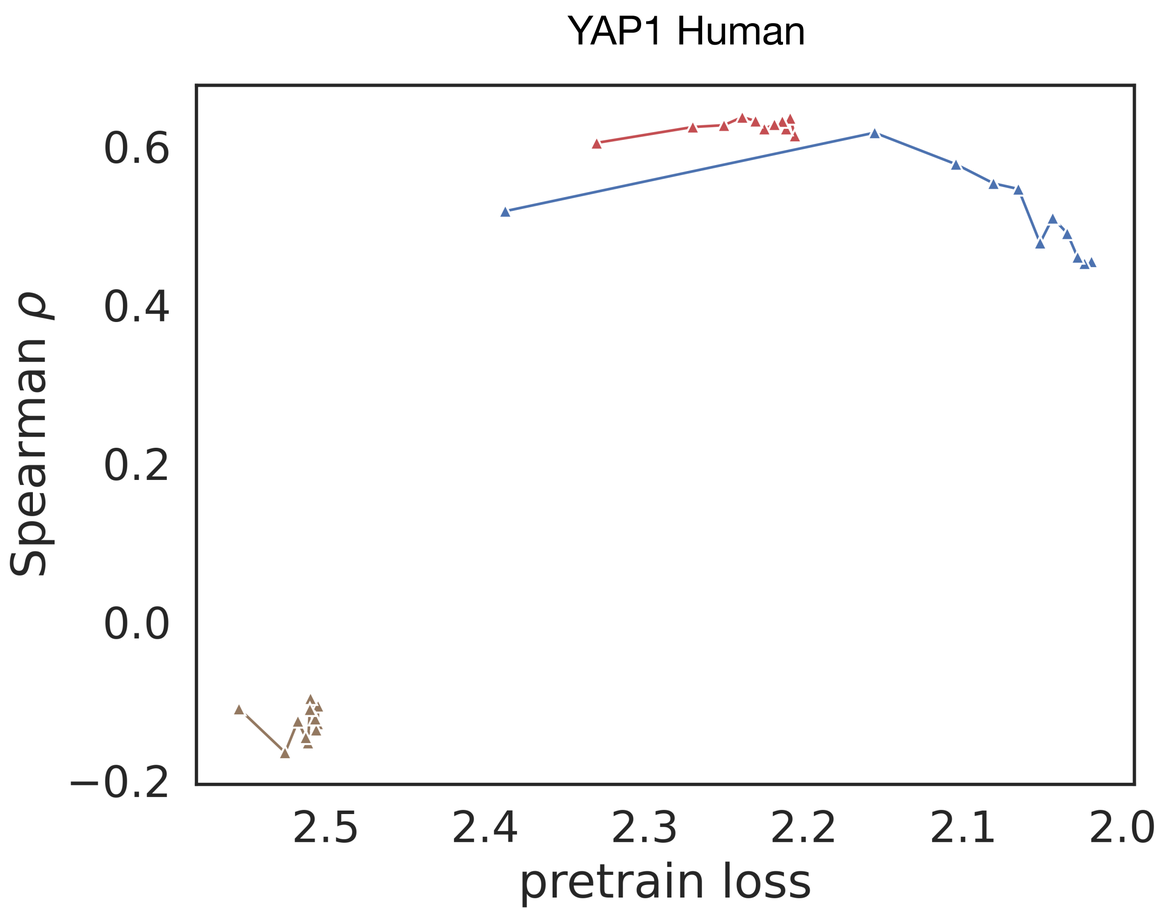


Zero-shot performance mostly improves with pretraining



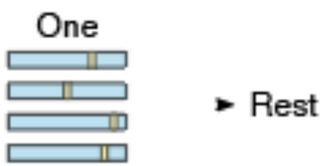
Zero-shot performance mostly improves with pretraining



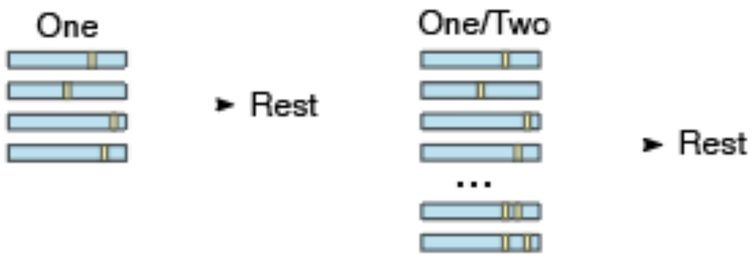


Protein engineering campaigns require out-of-domain generalization

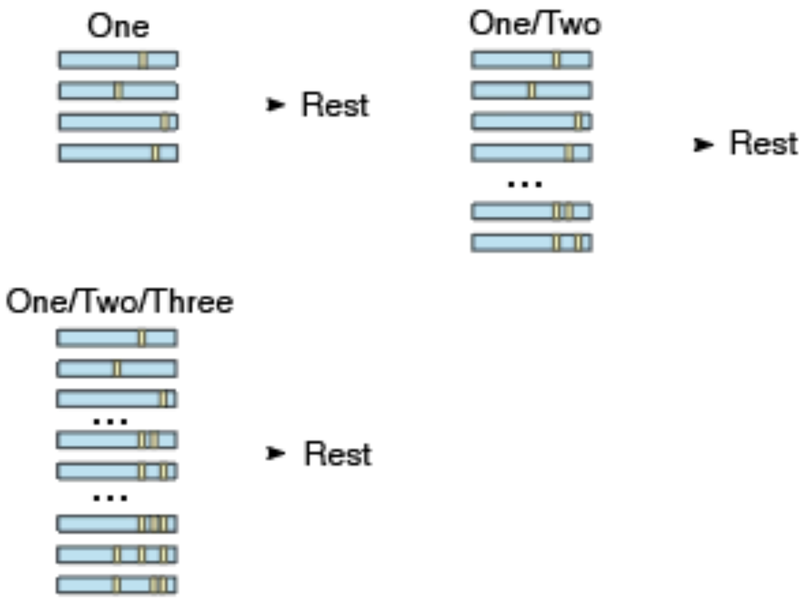
Protein engineering campaigns require out-of-domain generalization



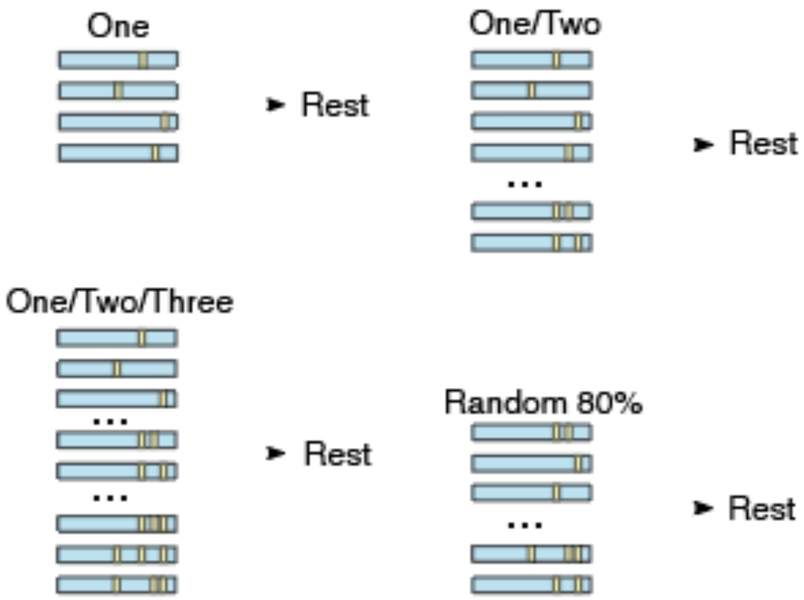
Protein engineering campaigns require out-of-domain generalization



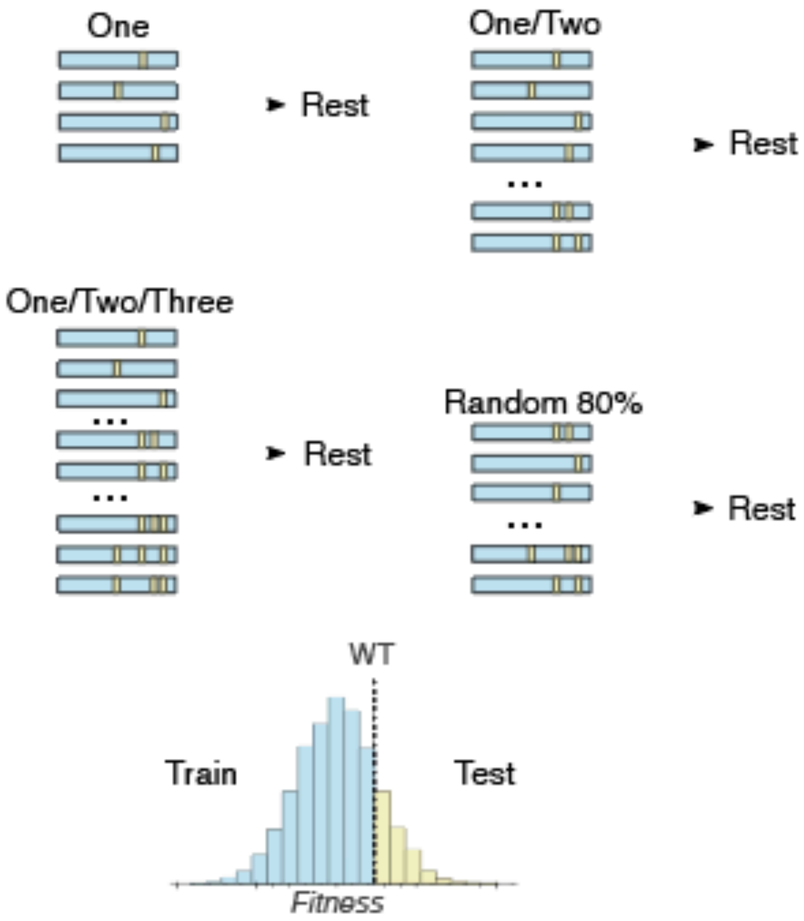
Protein engineering campaigns require out-of-domain generalization



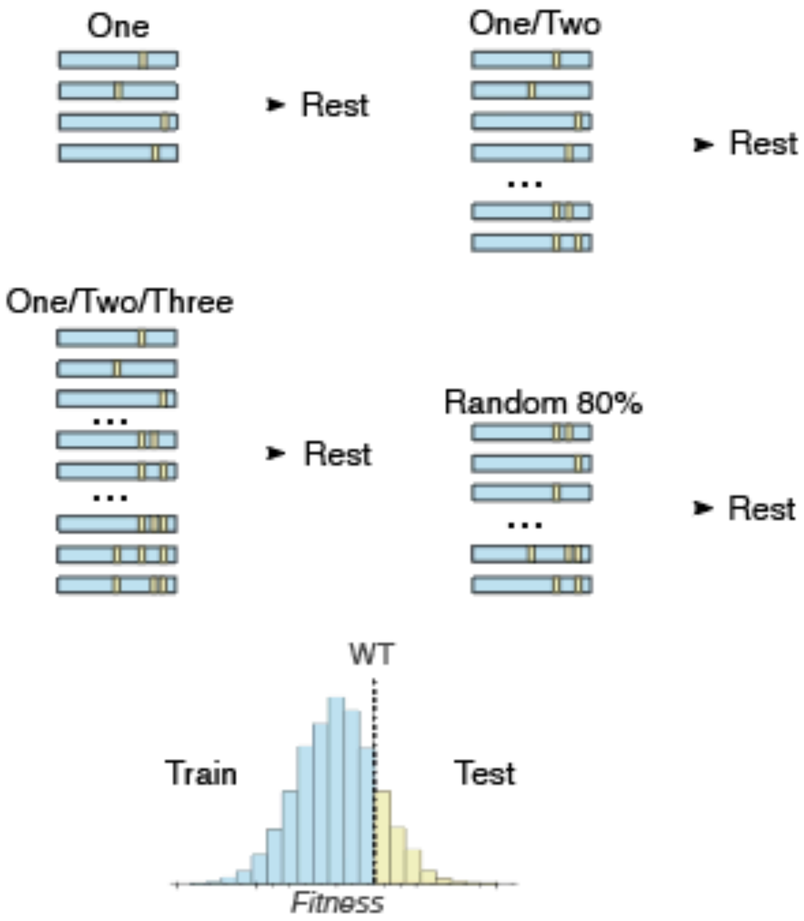
Protein engineering campaigns require out-of-domain generalization



Protein engineering campaigns require out-of-domain generalization

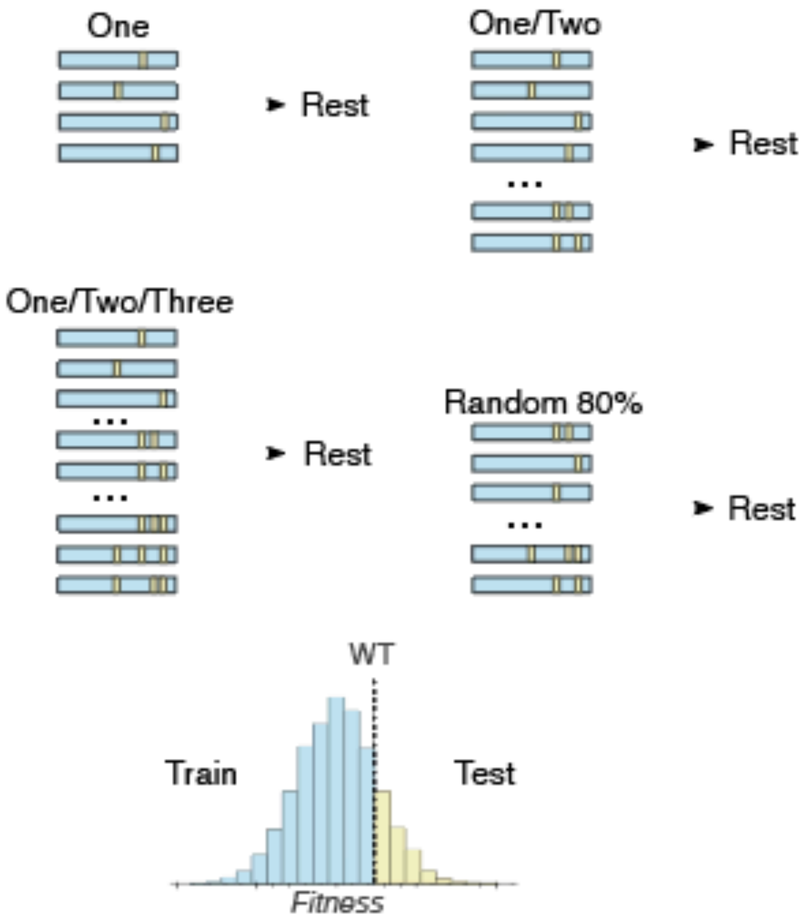


Protein engineering campaigns require out-of-domain generalization



Can usually measure all single mutants

Protein engineering campaigns require out-of-domain generalization

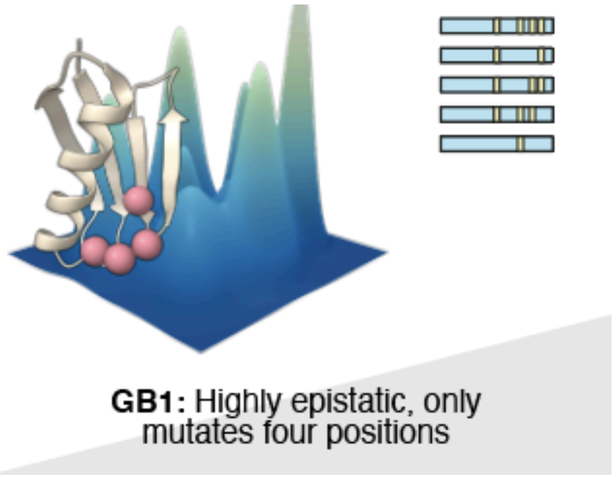


Can usually measure all single mutants

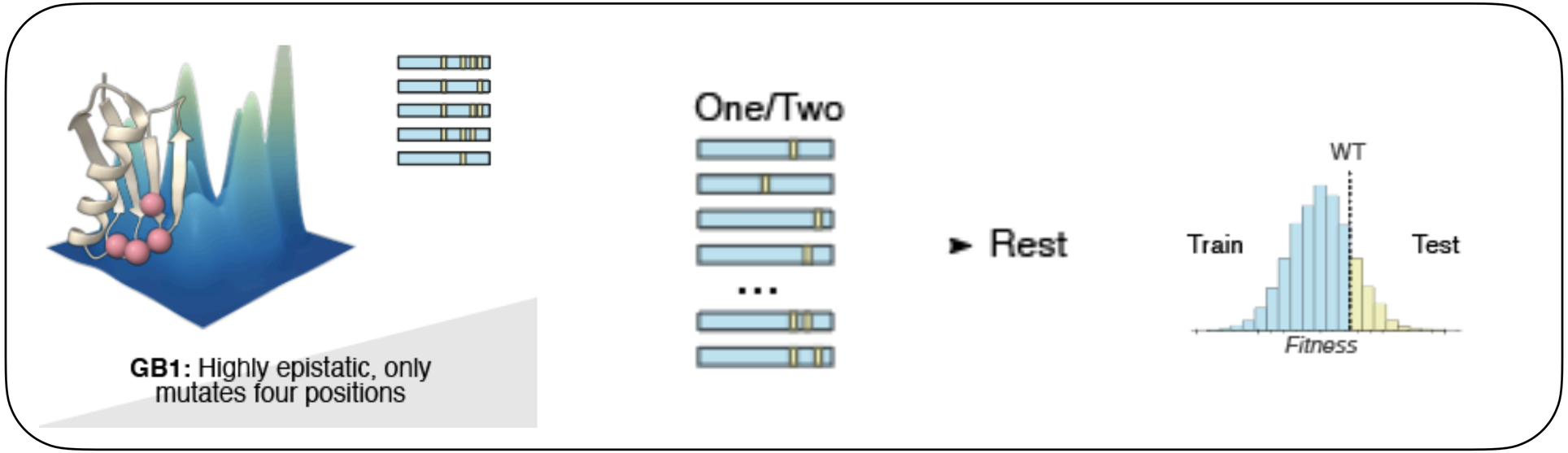
Naively screen bigger spaces
->
mostly non-functional

Protein engineering campaigns require out-of-domain generalization

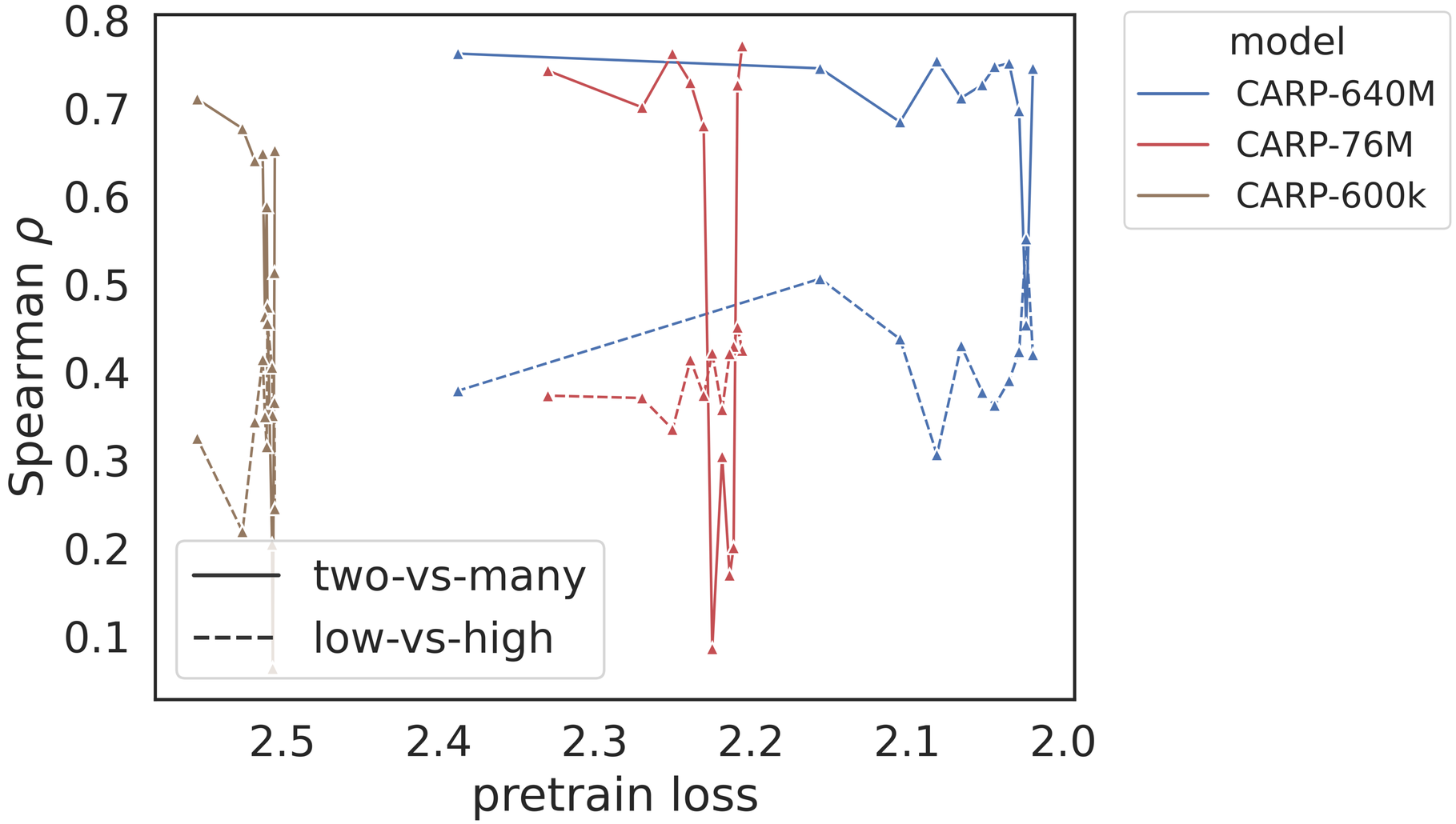
Protein engineering campaigns require out-of-domain generalization



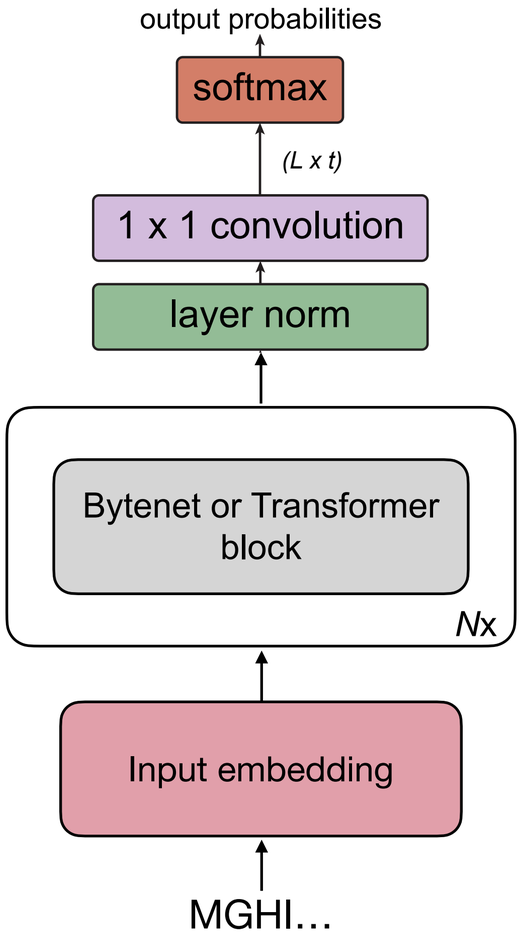
Protein engineering campaigns require out-of-domain generalization



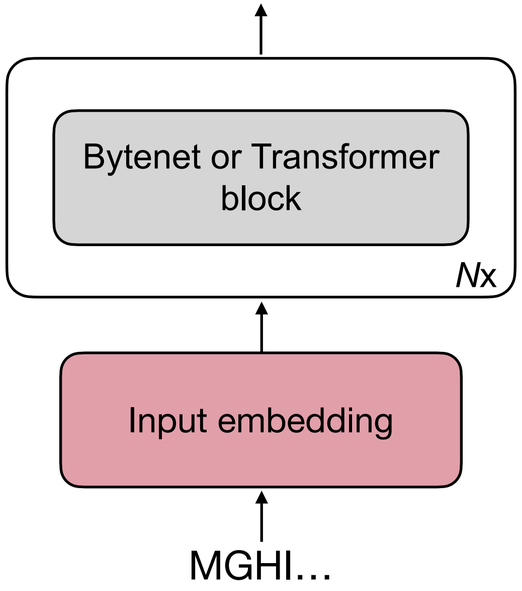
OOD performance does not improve with more pretraining



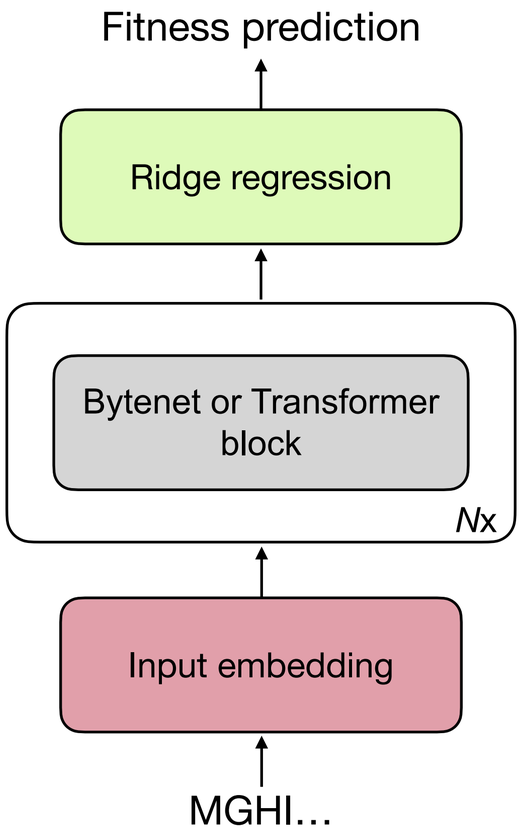
Not all layers are necessary for transfer



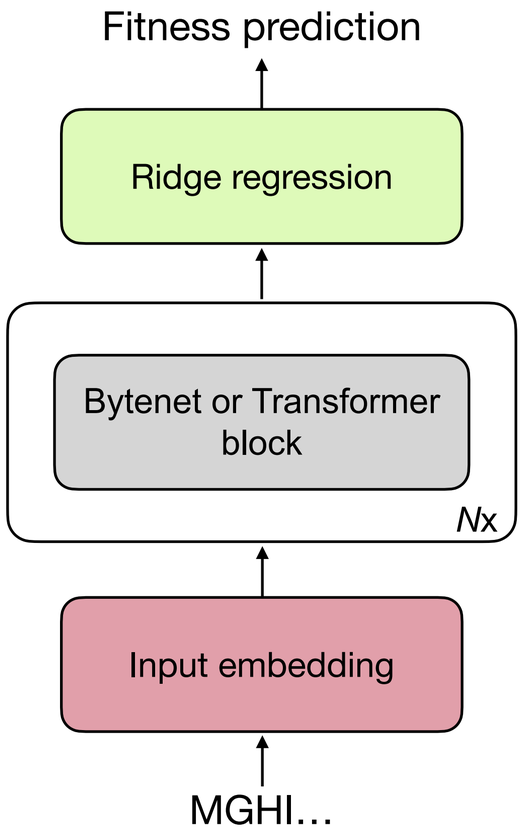
Not all layers are necessary for transfer



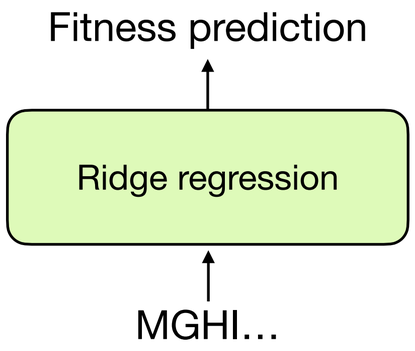
Not all layers are necessary for transfer



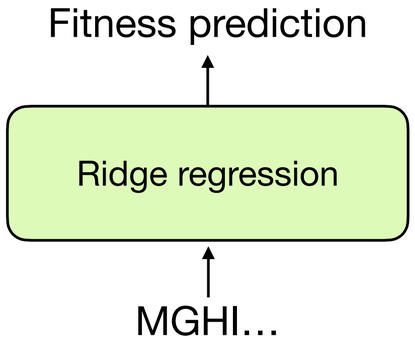
Not all layers are necessary for transfer

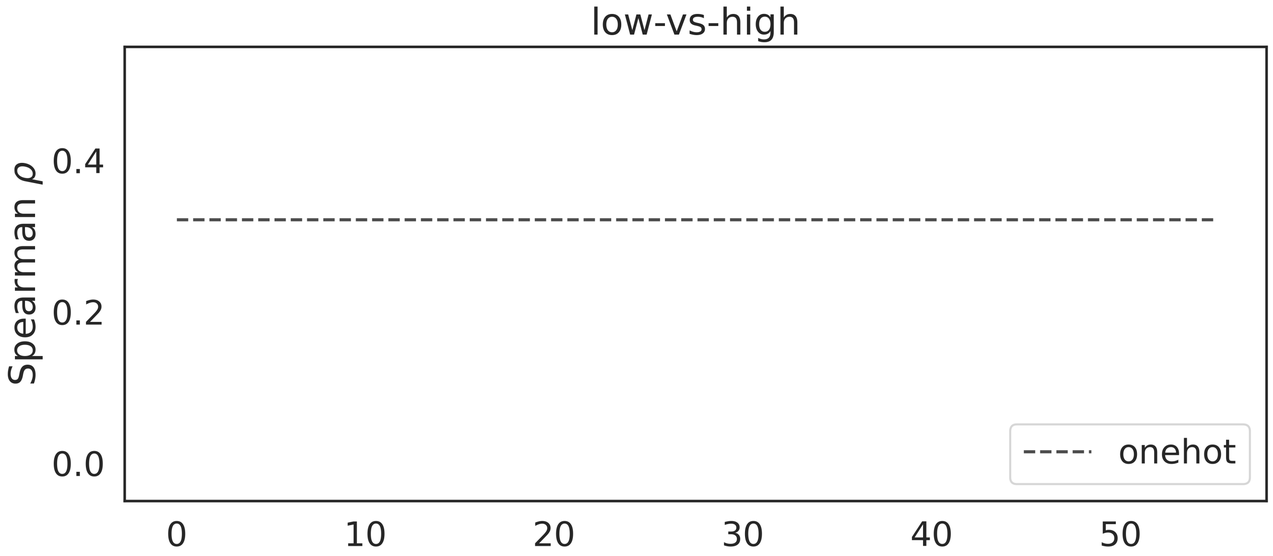


Not all layers are necessary for transfer

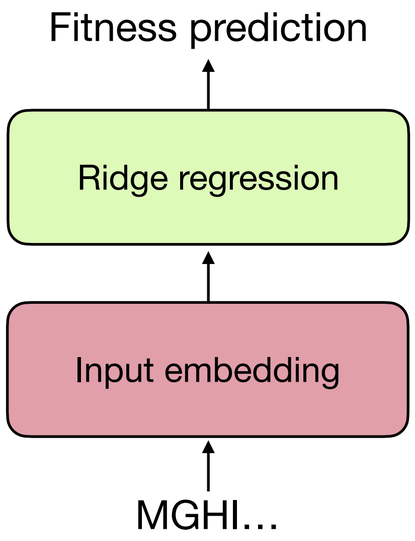


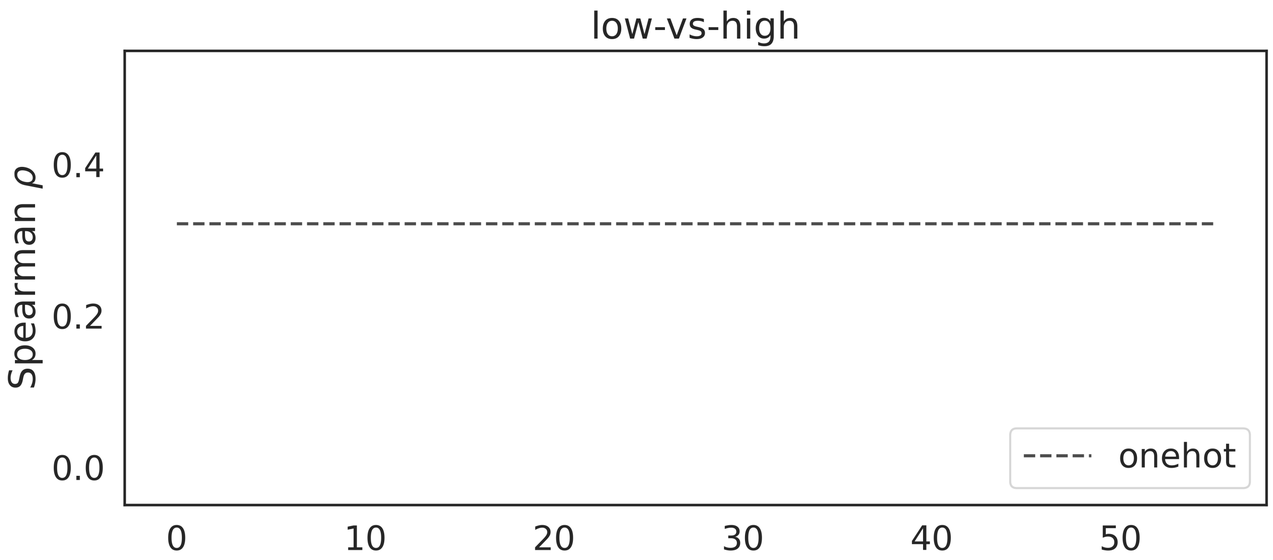
Not all layers are necessary for transfer



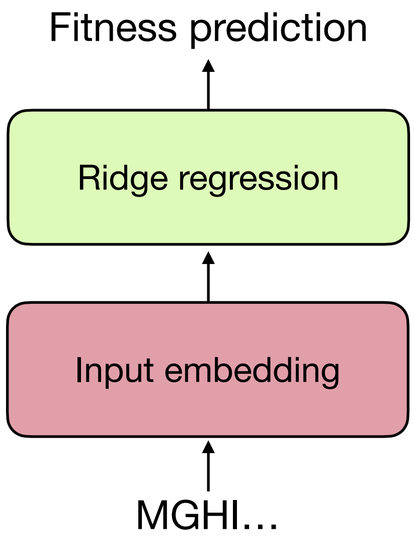


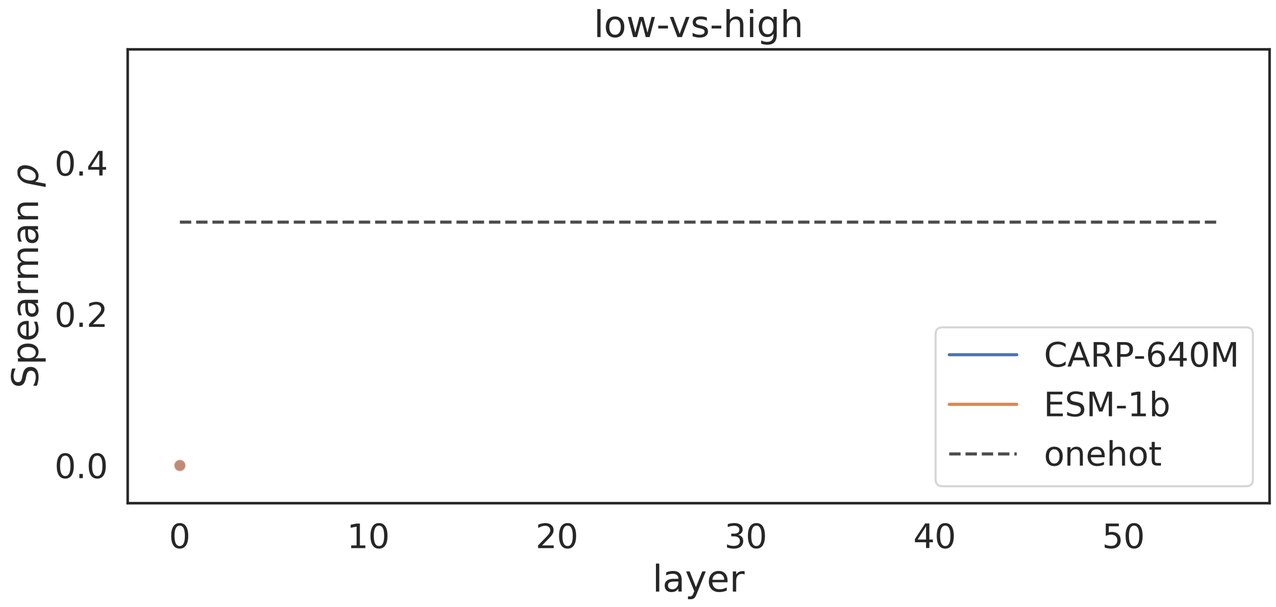
Not all layers are necessary for transfer



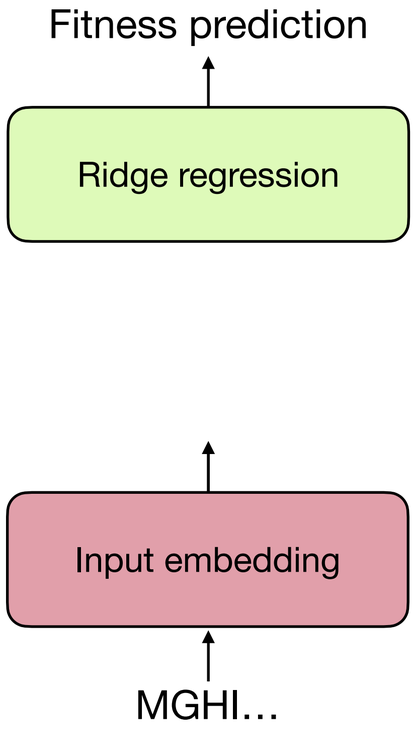


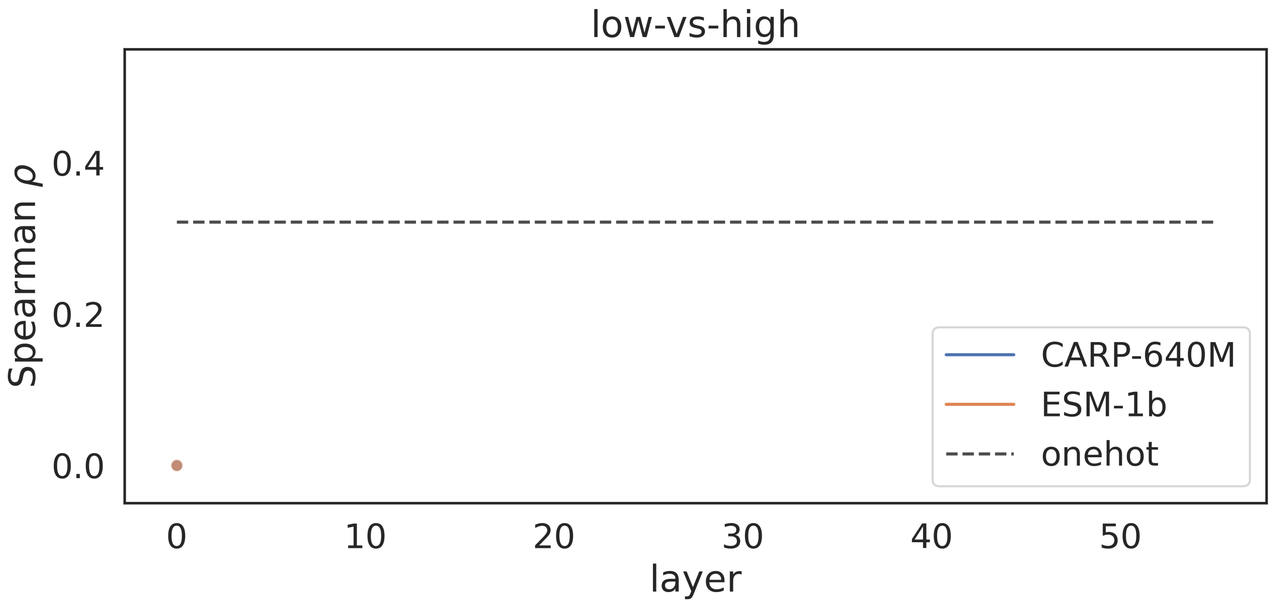
Not all layers are necessary for transfer



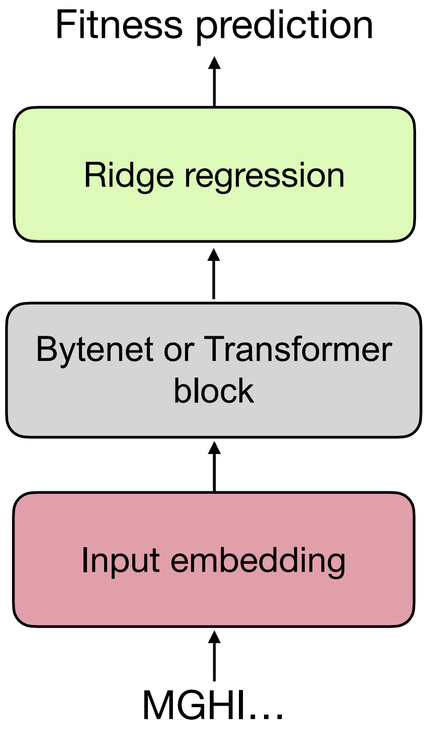


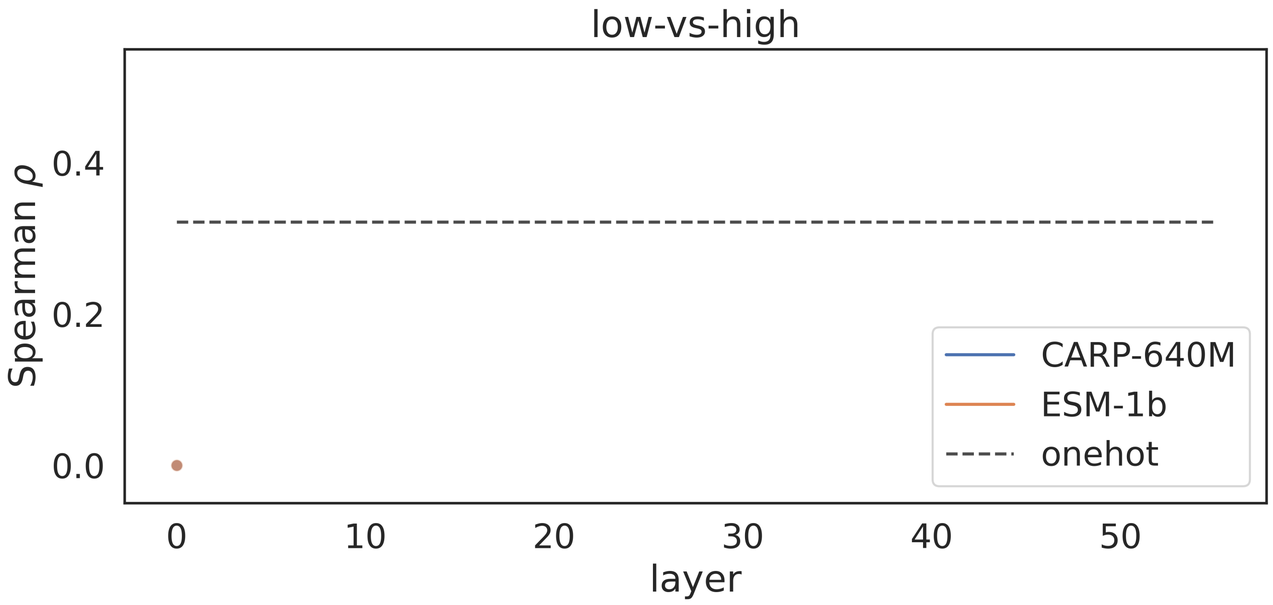
Not all layers are necessary for transfer



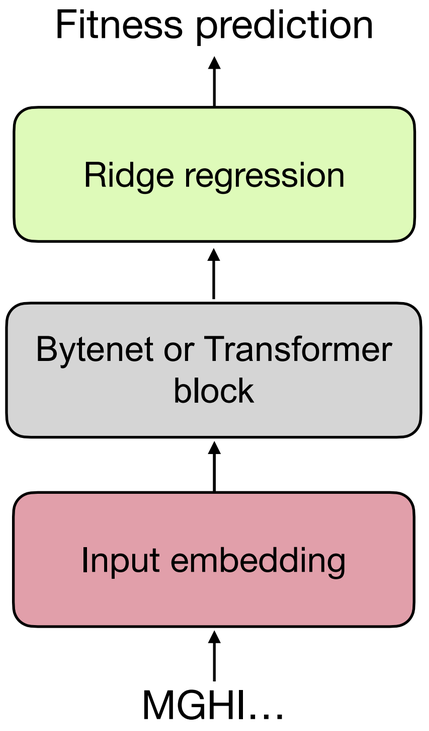


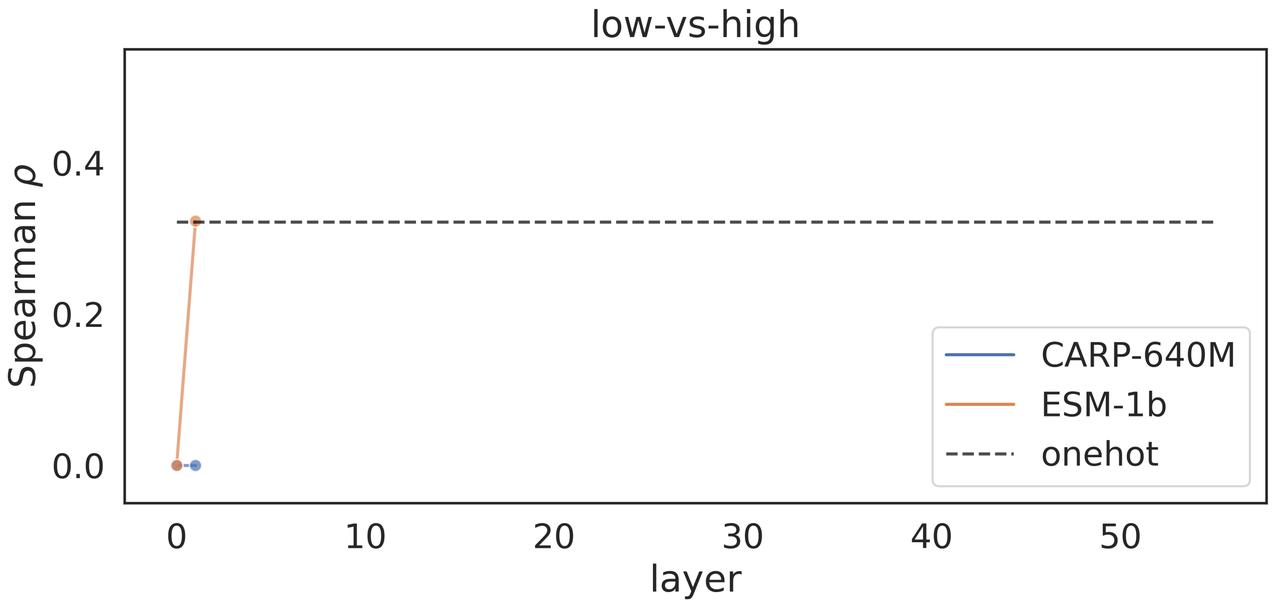
Not all layers are necessary for transfer



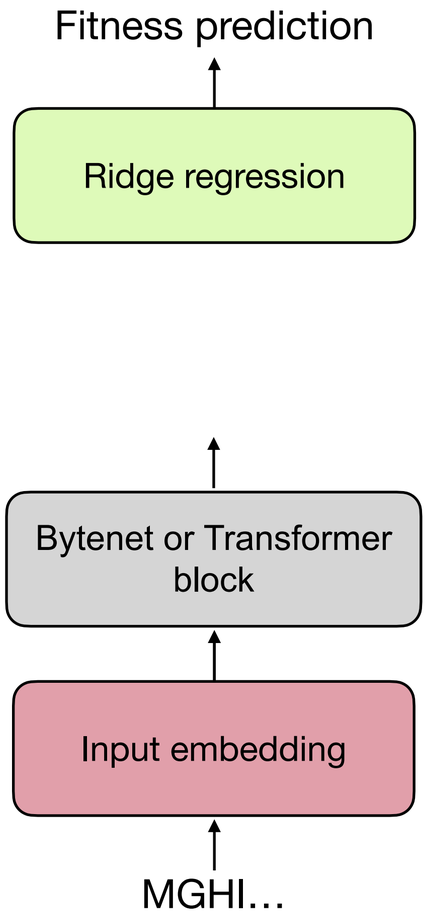


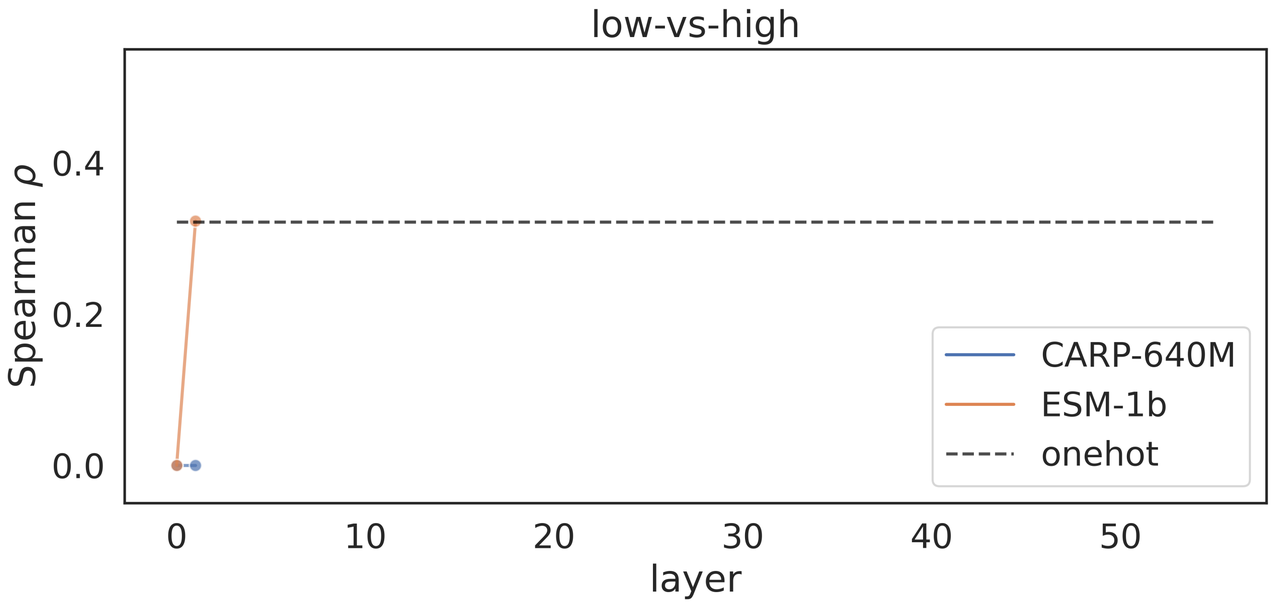
Not all layers are necessary for transfer



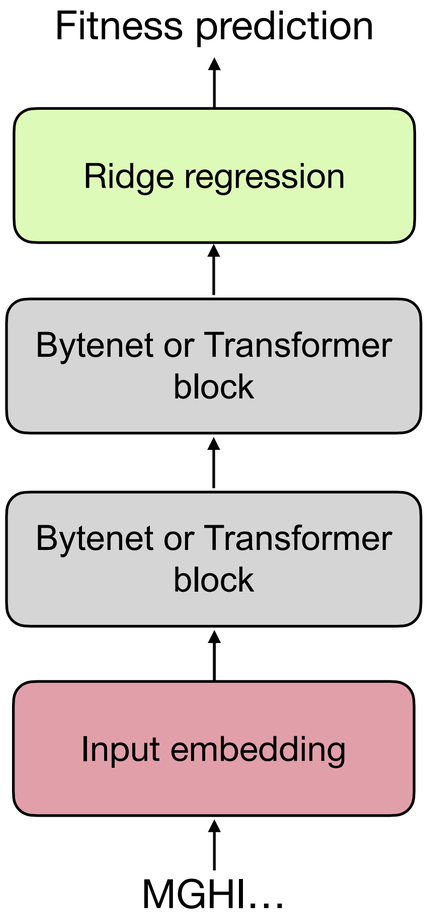


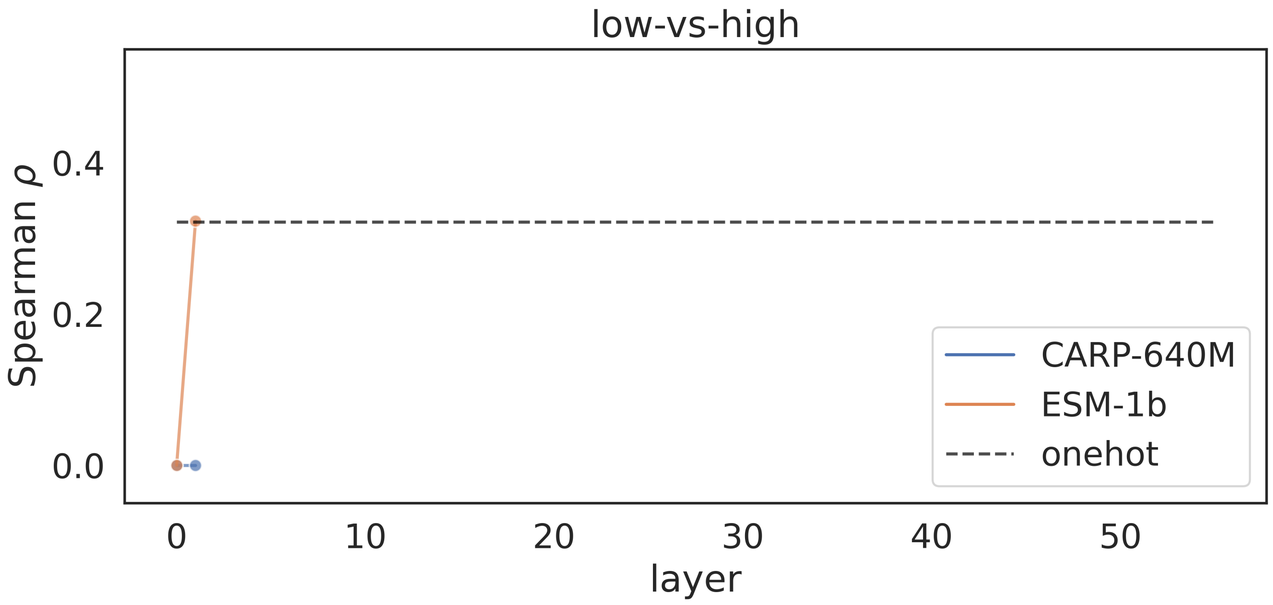
Not all layers are necessary for transfer



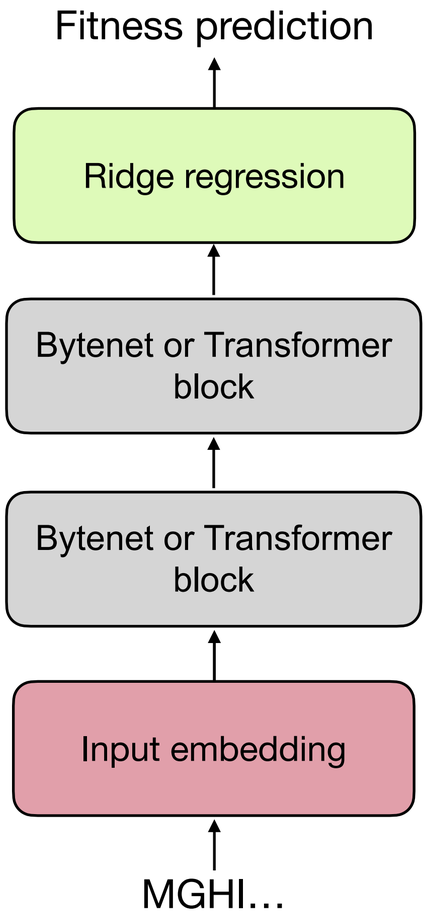


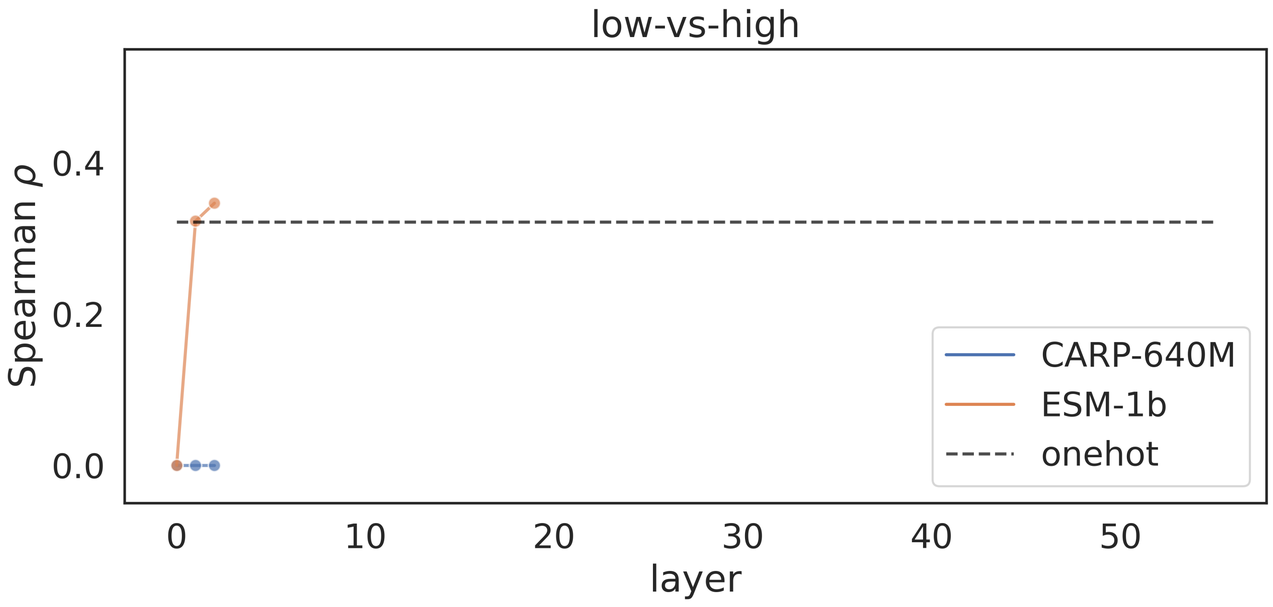
Not all layers are necessary for transfer



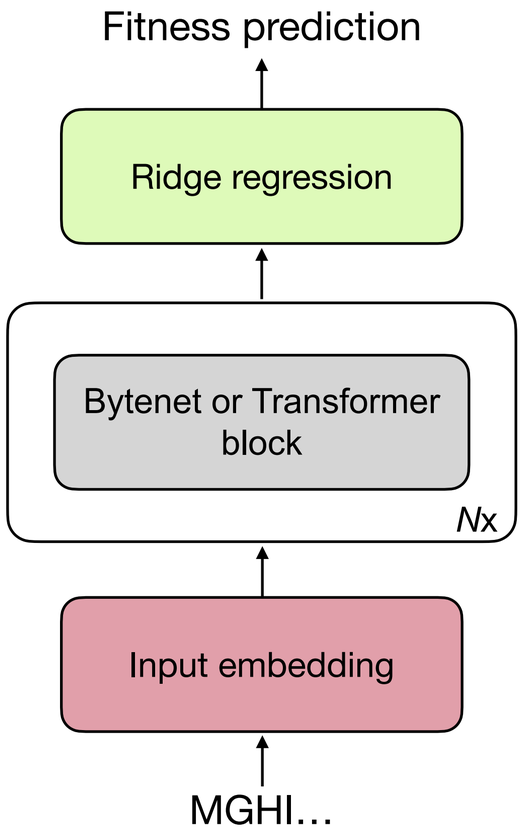


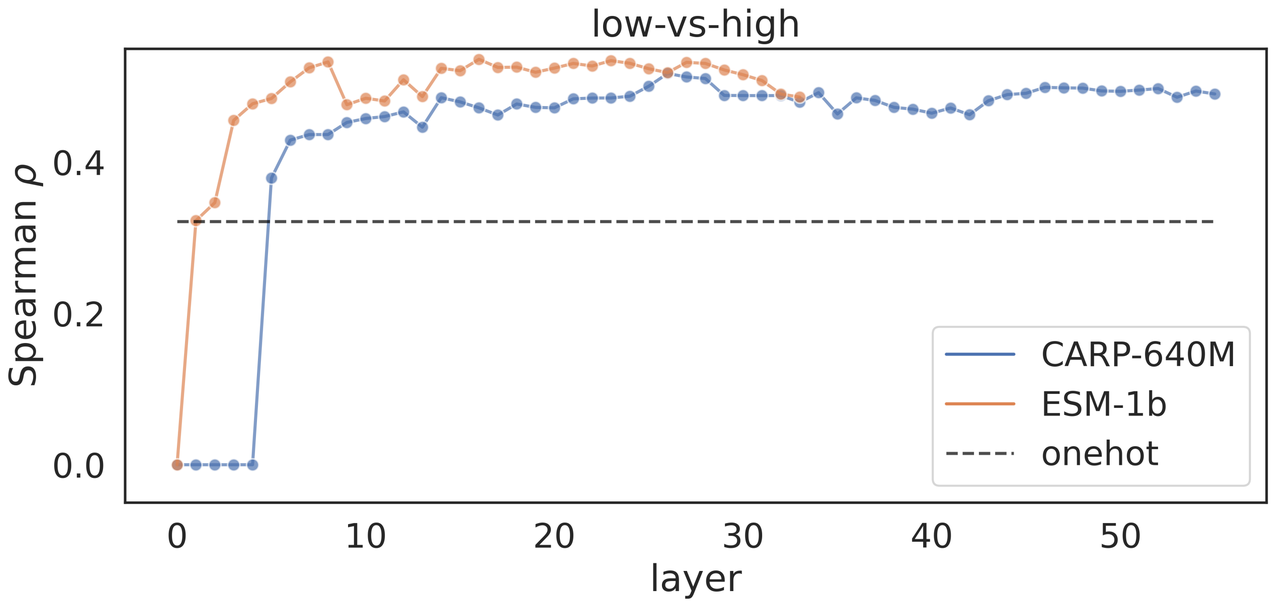
Not all layers are necessary for transfer



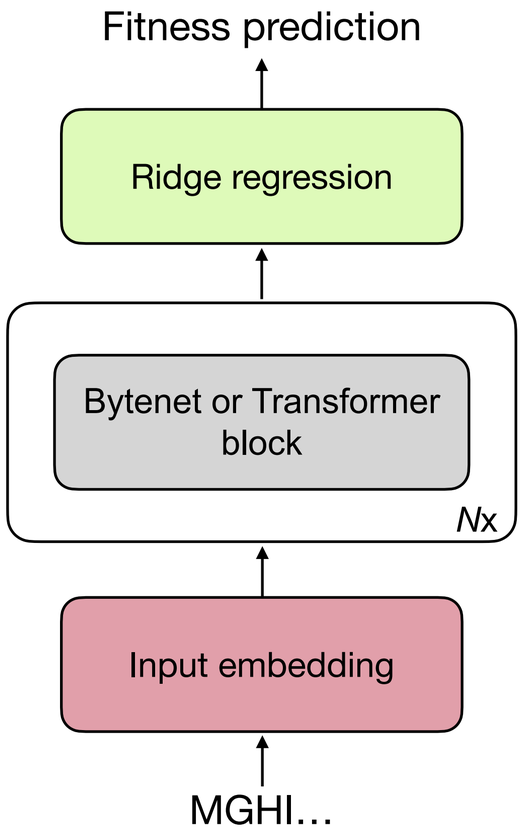


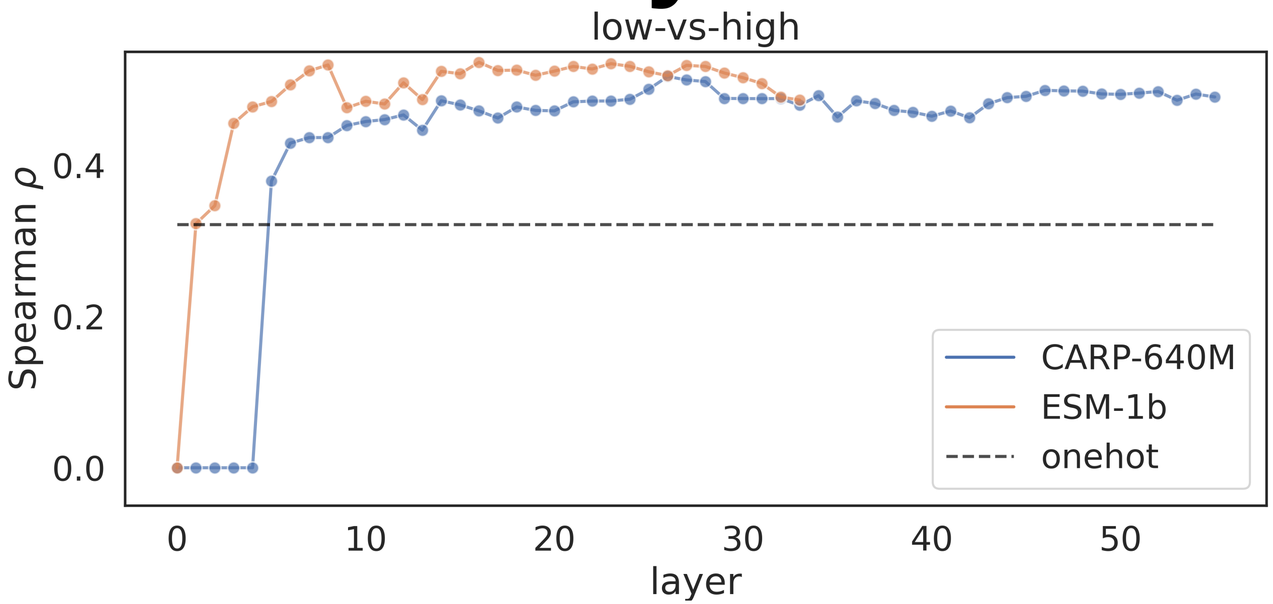
Not all layers are necessary for transfer

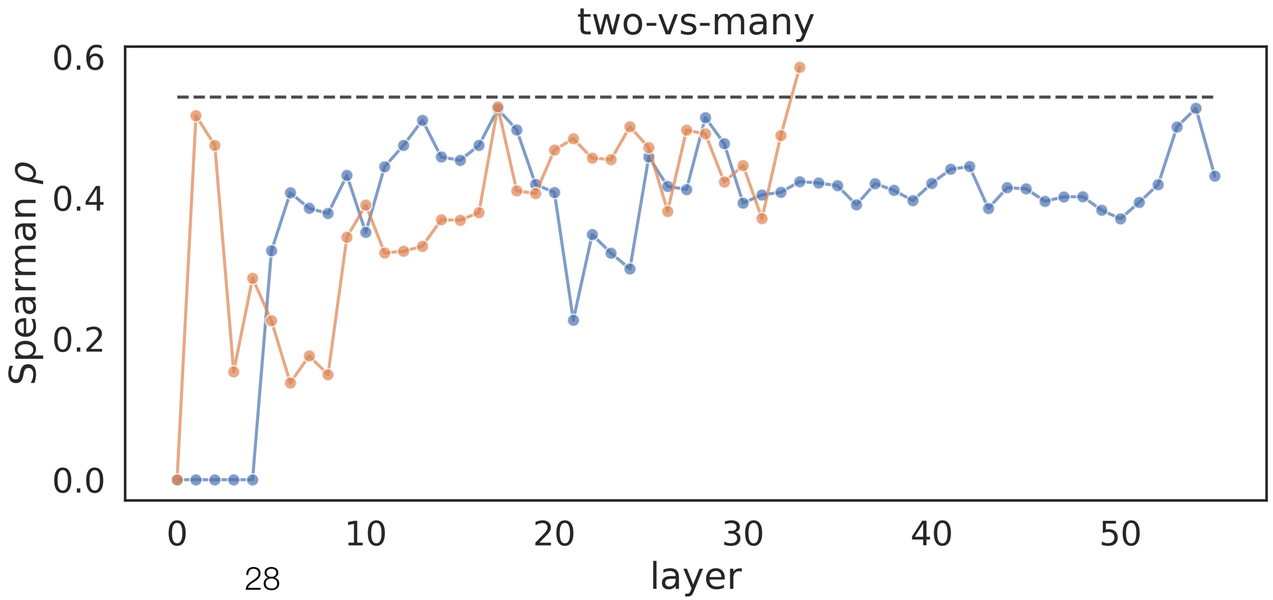




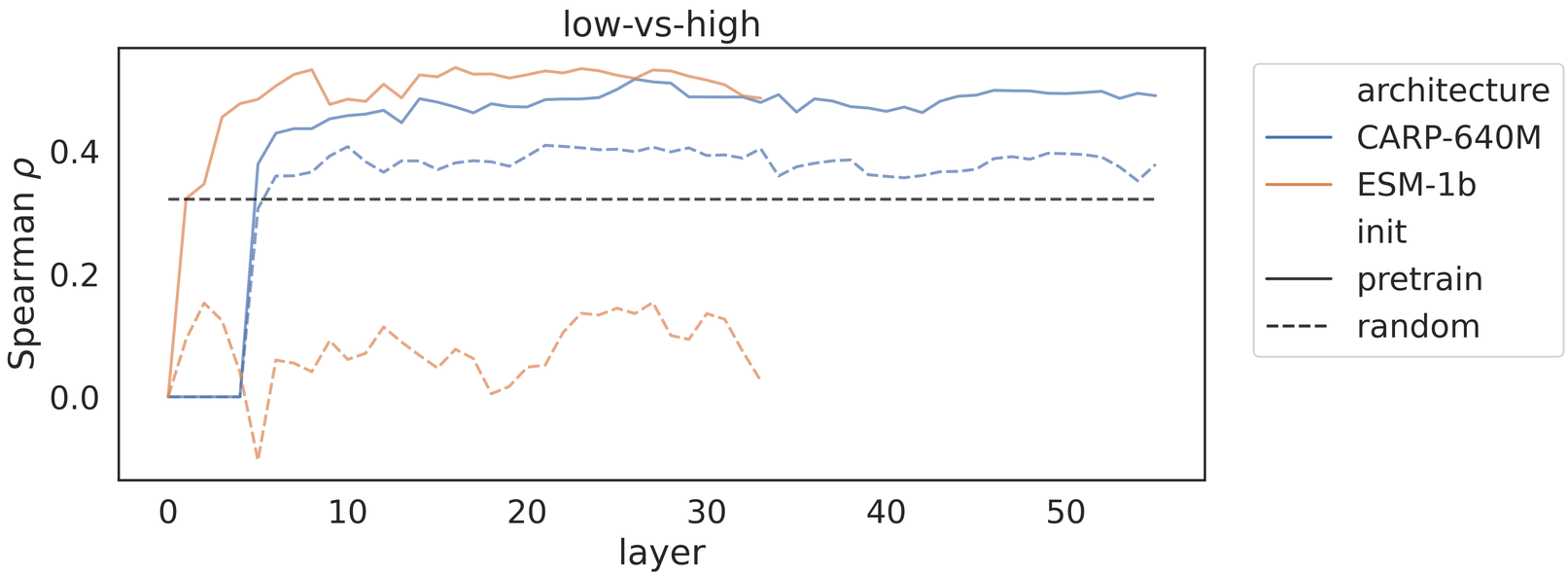
Not all layers are necessary for transfer



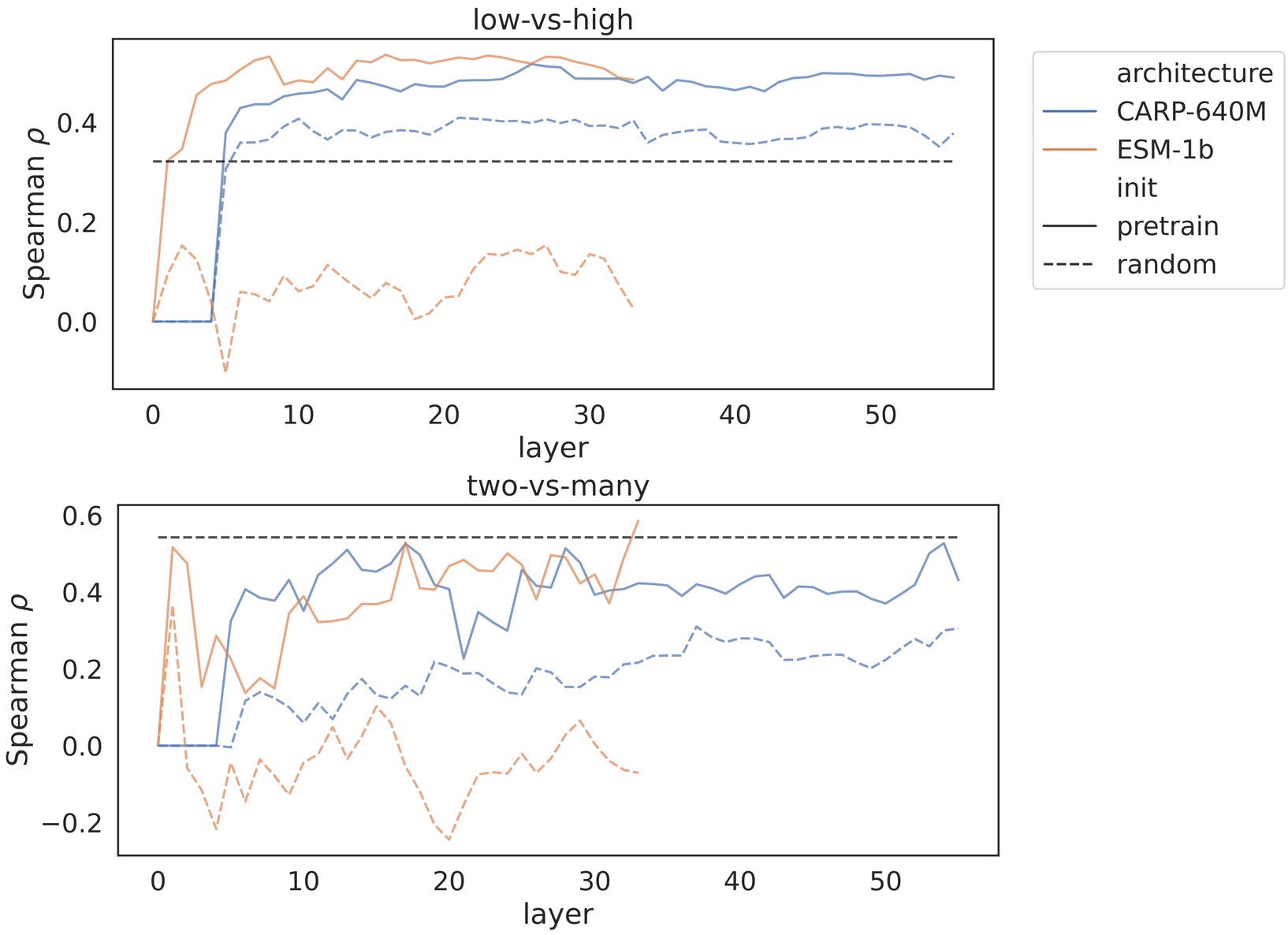




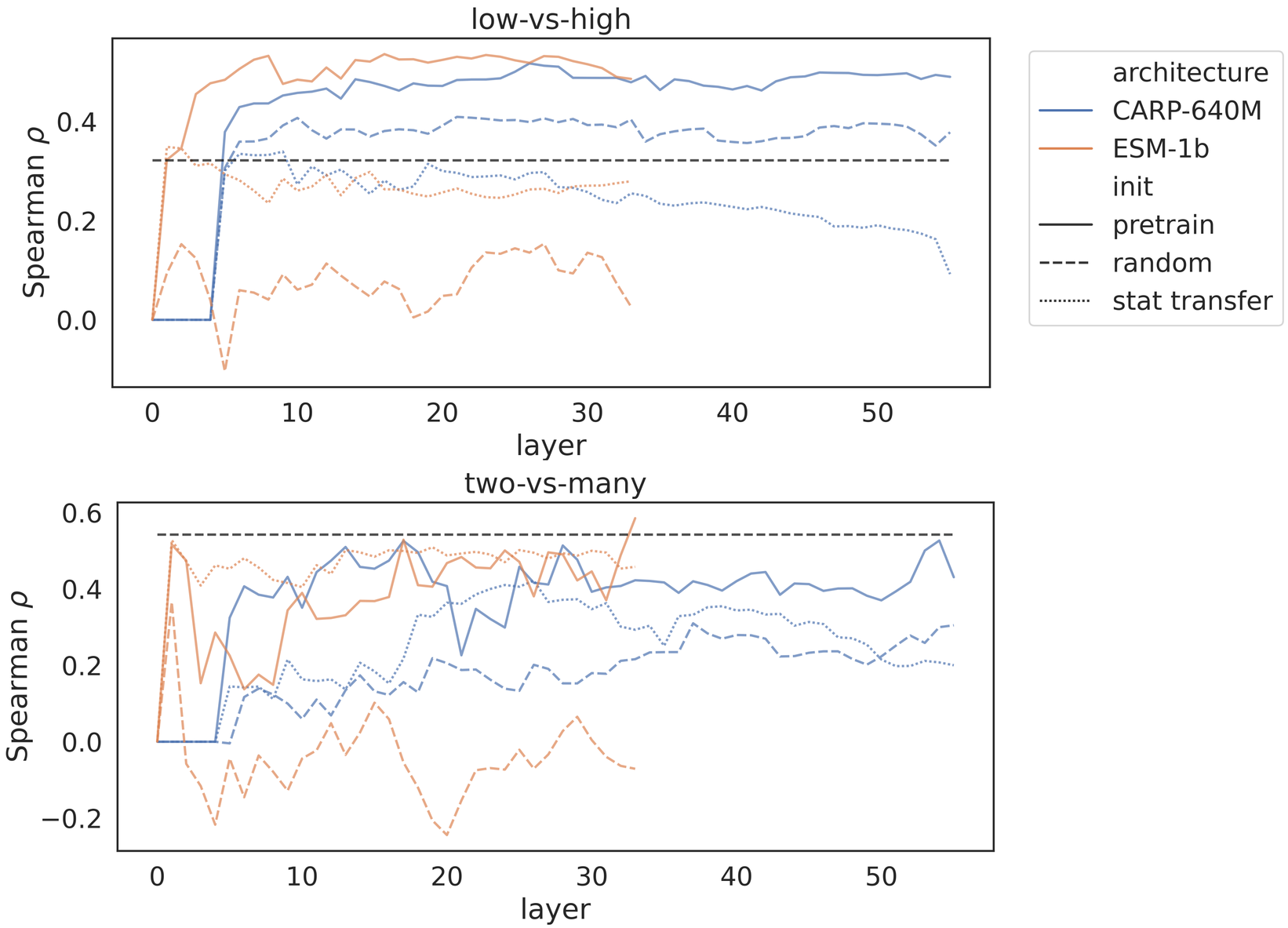
Pretraining helps beyond the architecture



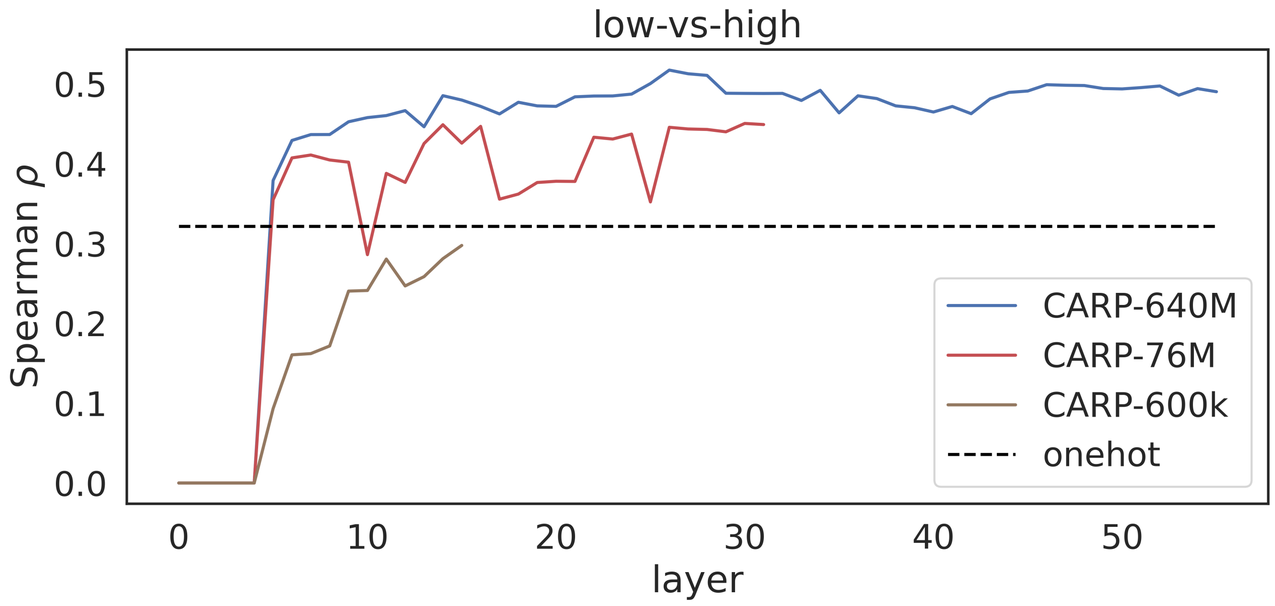
Pretraining helps beyond the architecture



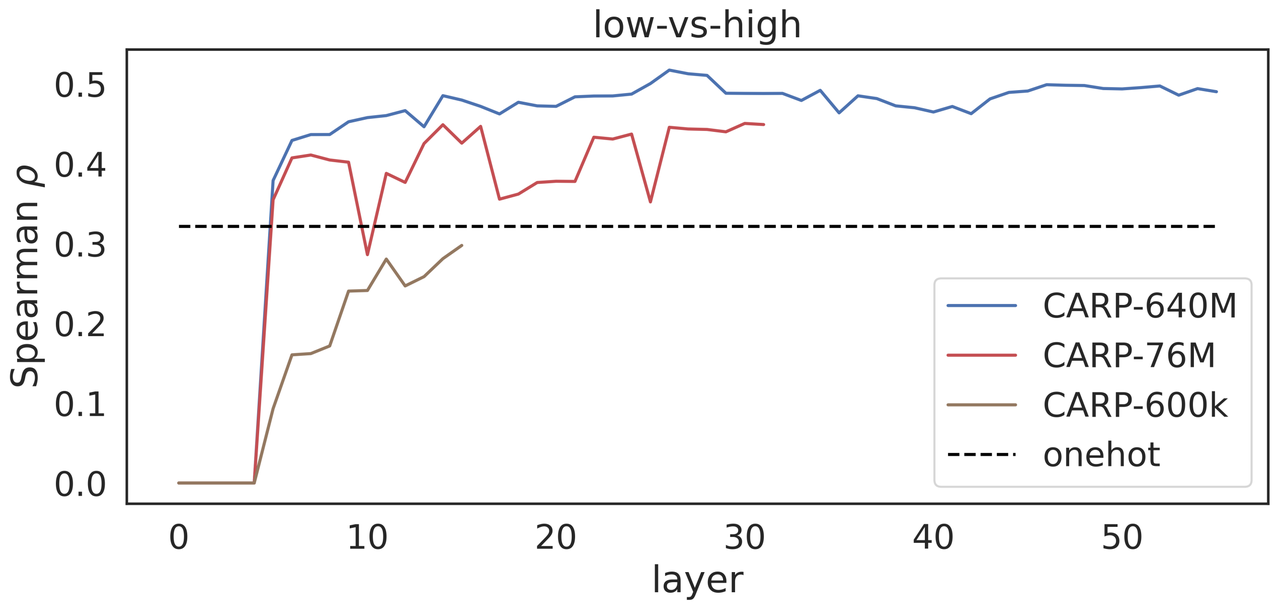
Sometimes, shuffled weights work too

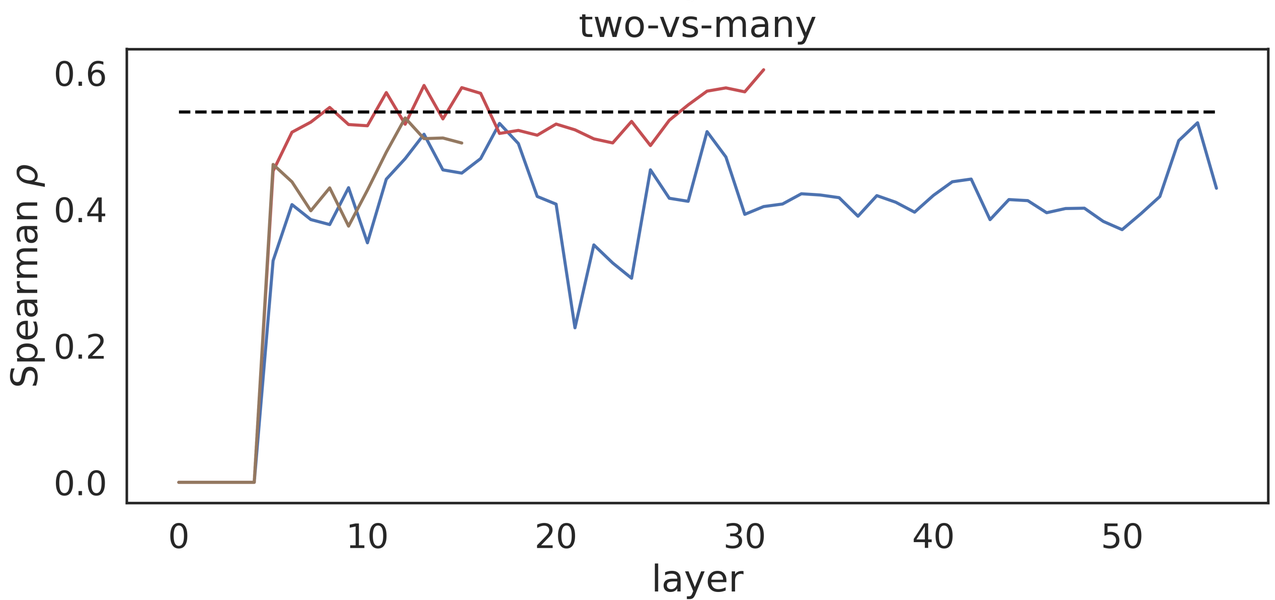


Bigger models do not always help

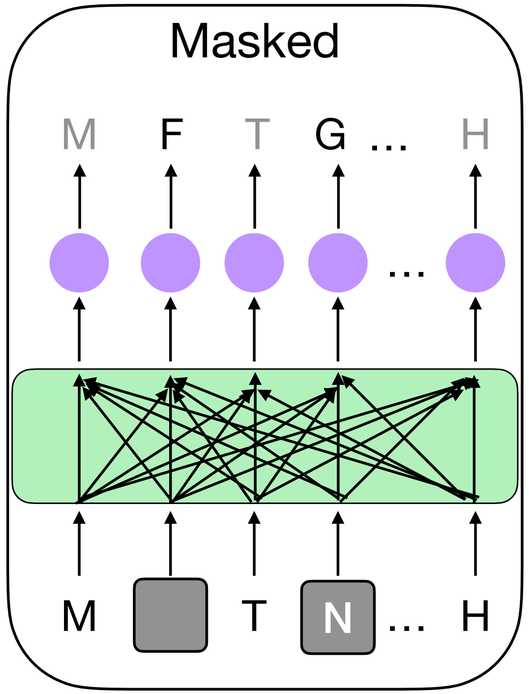


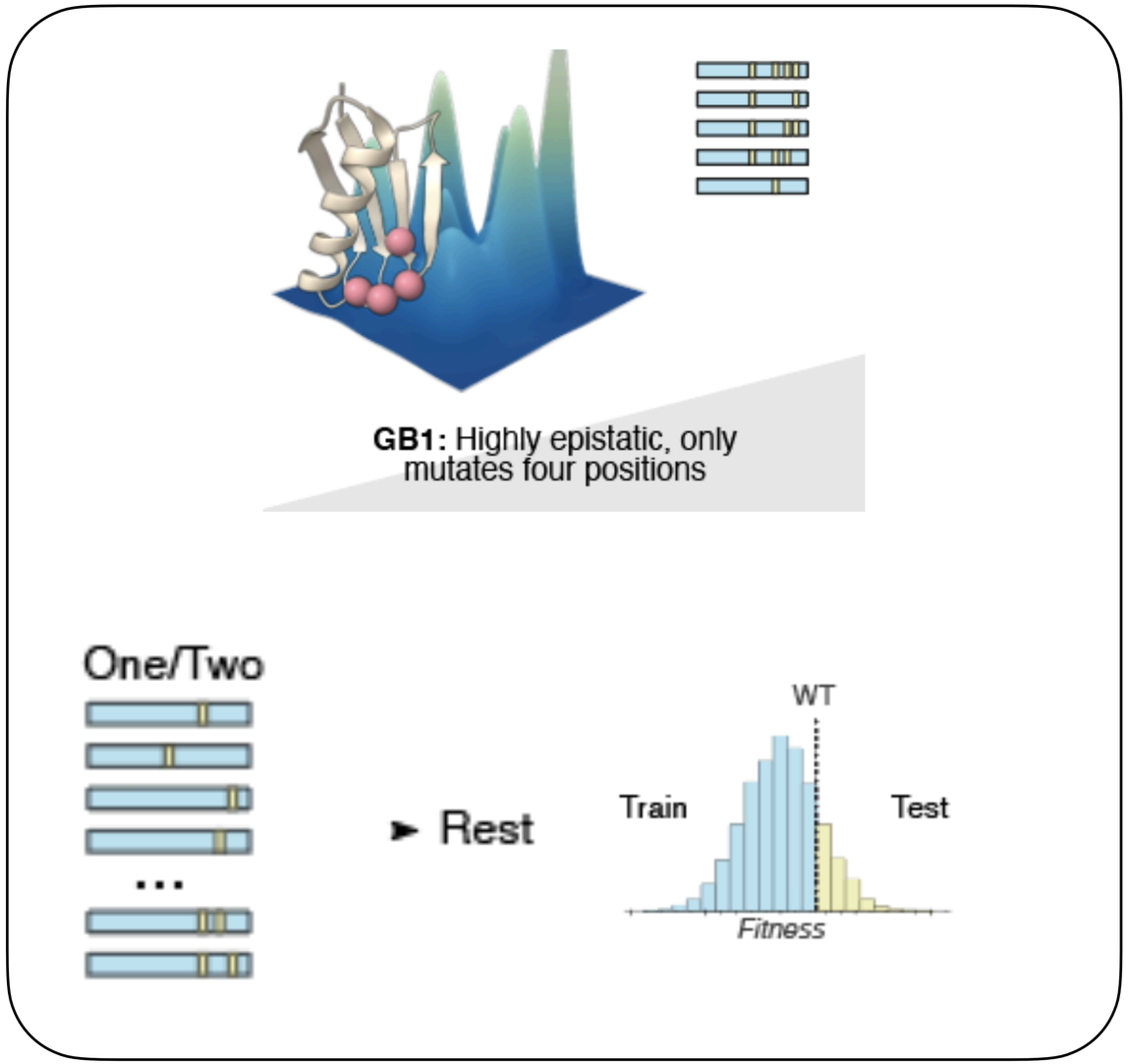
Bigger models do not always help



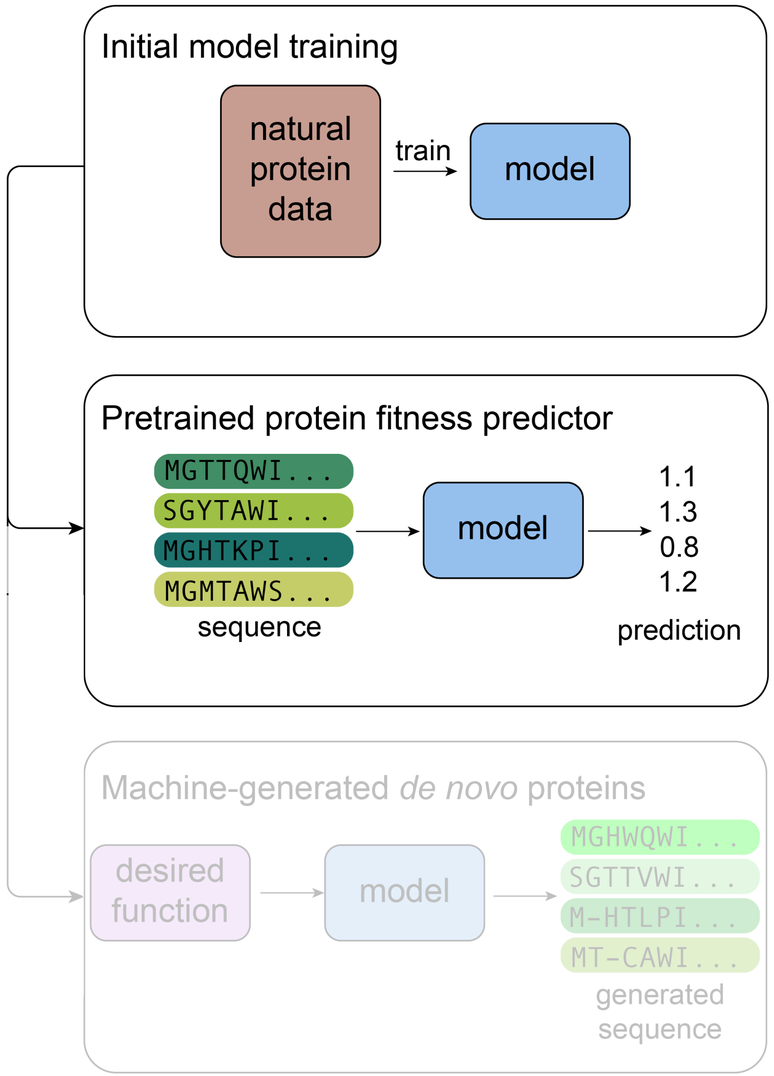


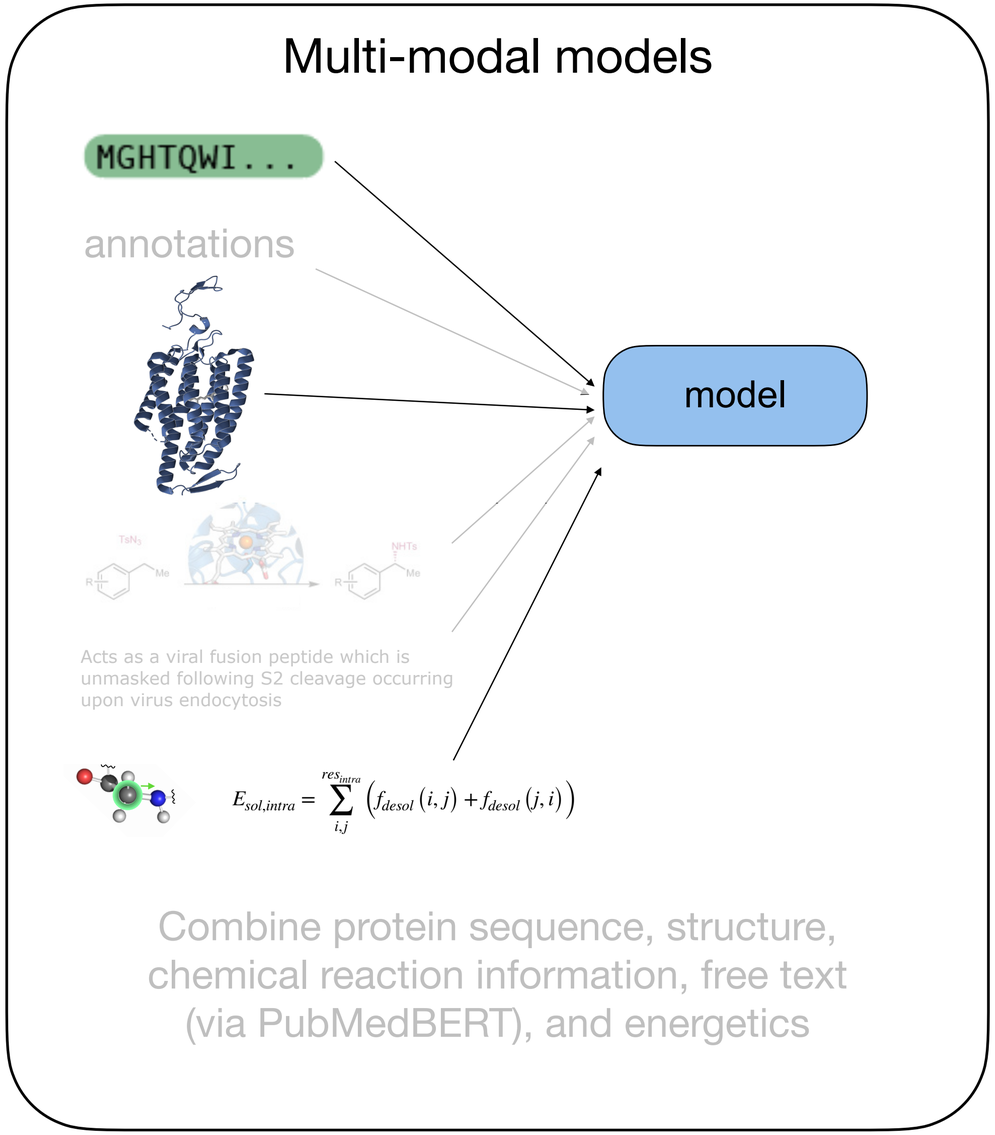
Pretrain and downstream tasks are mismatched



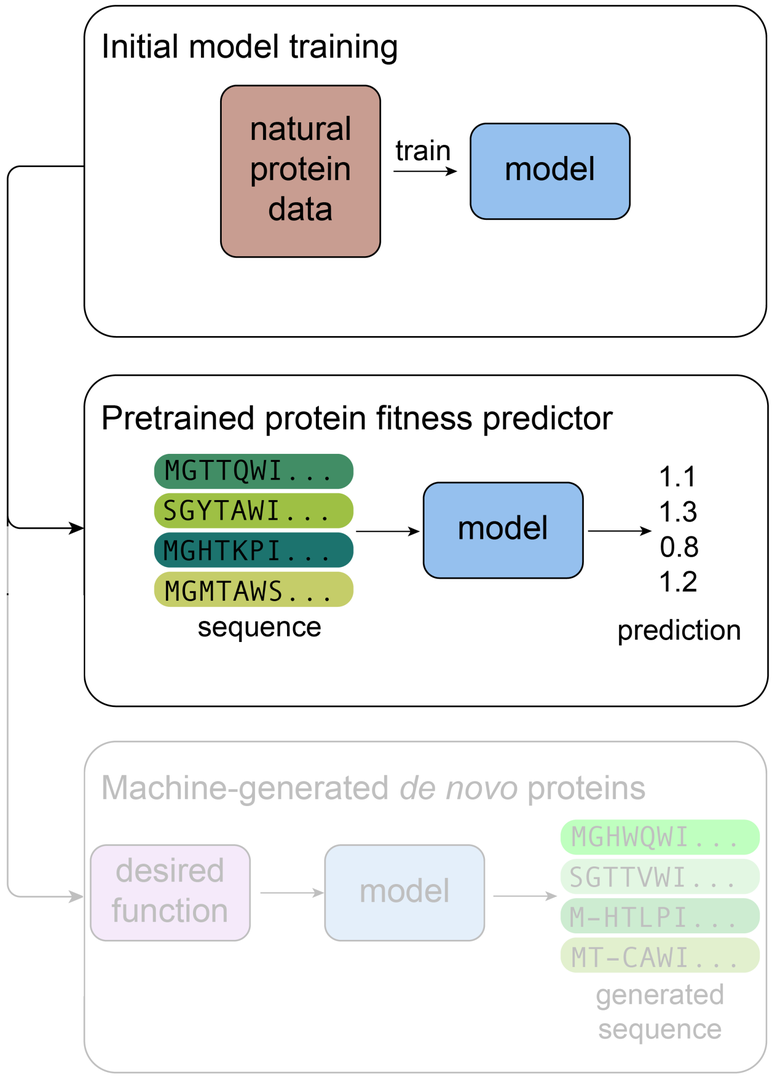


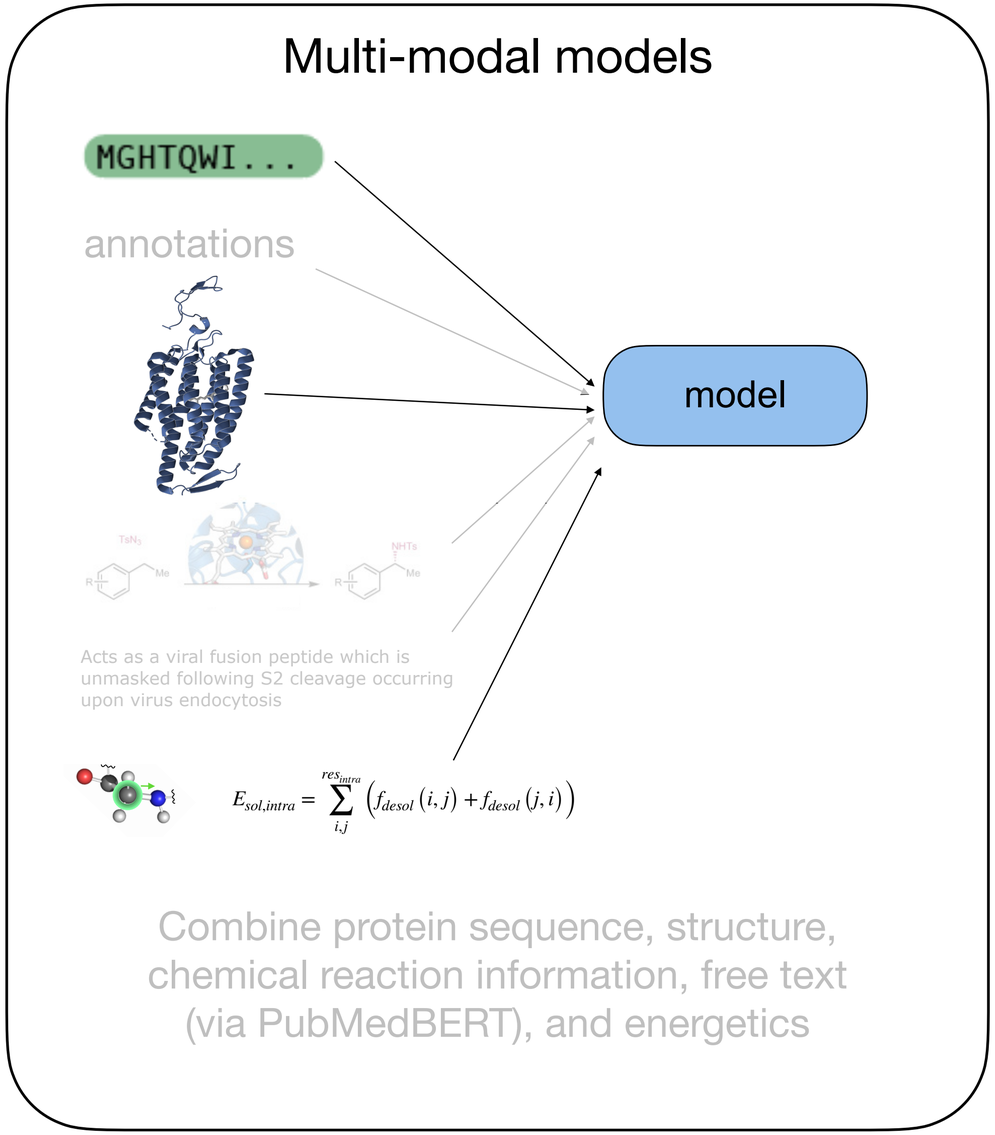
Use energetic features as inputs





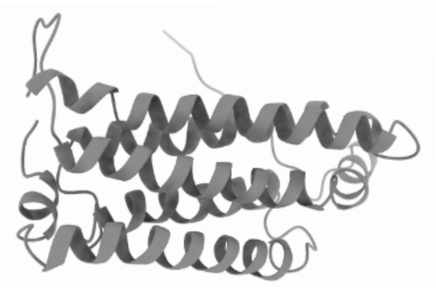
Use energetic features as inputs





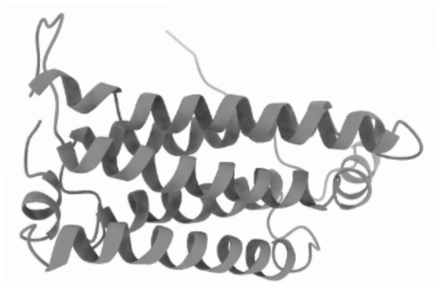
Use precomputed physics-based energies as input features

Use precomputed physics-based energies as input features

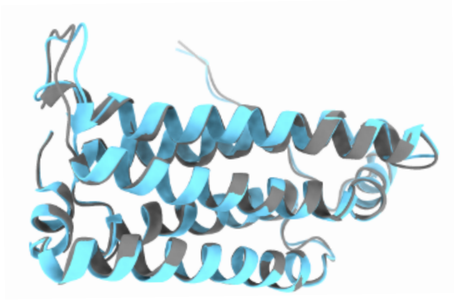


Reference protein

Use precomputed physics-based energies as input features

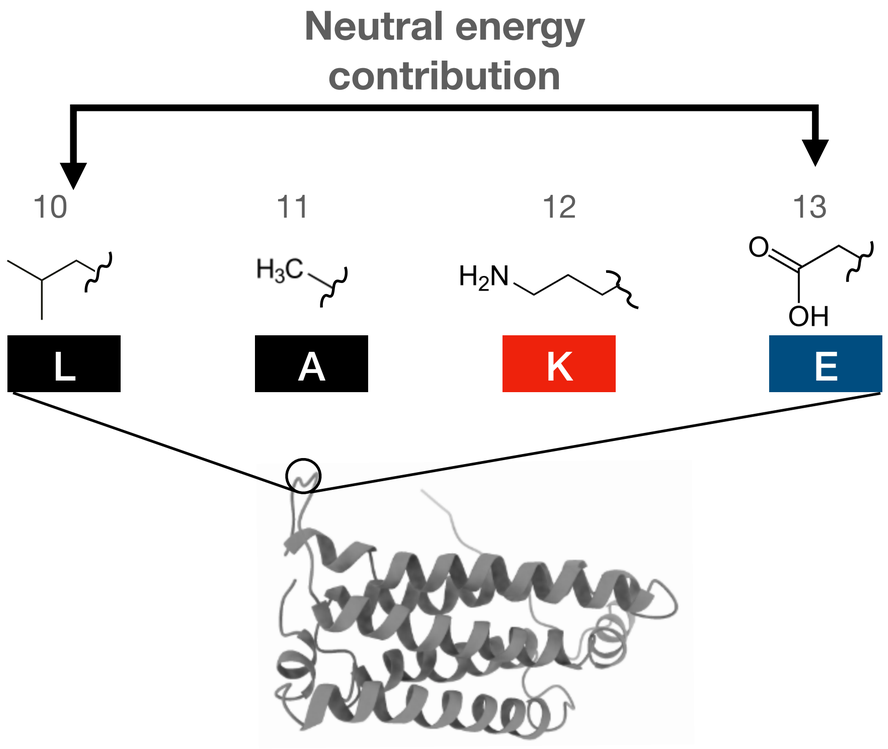


Reference protein

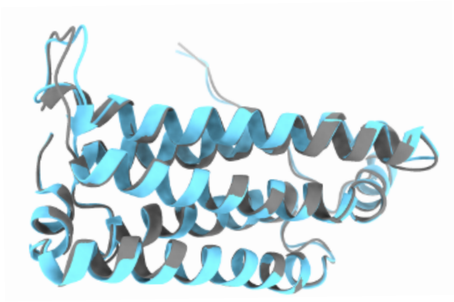


Protein variant: L10E

Use precomputed physics-based energies as input features

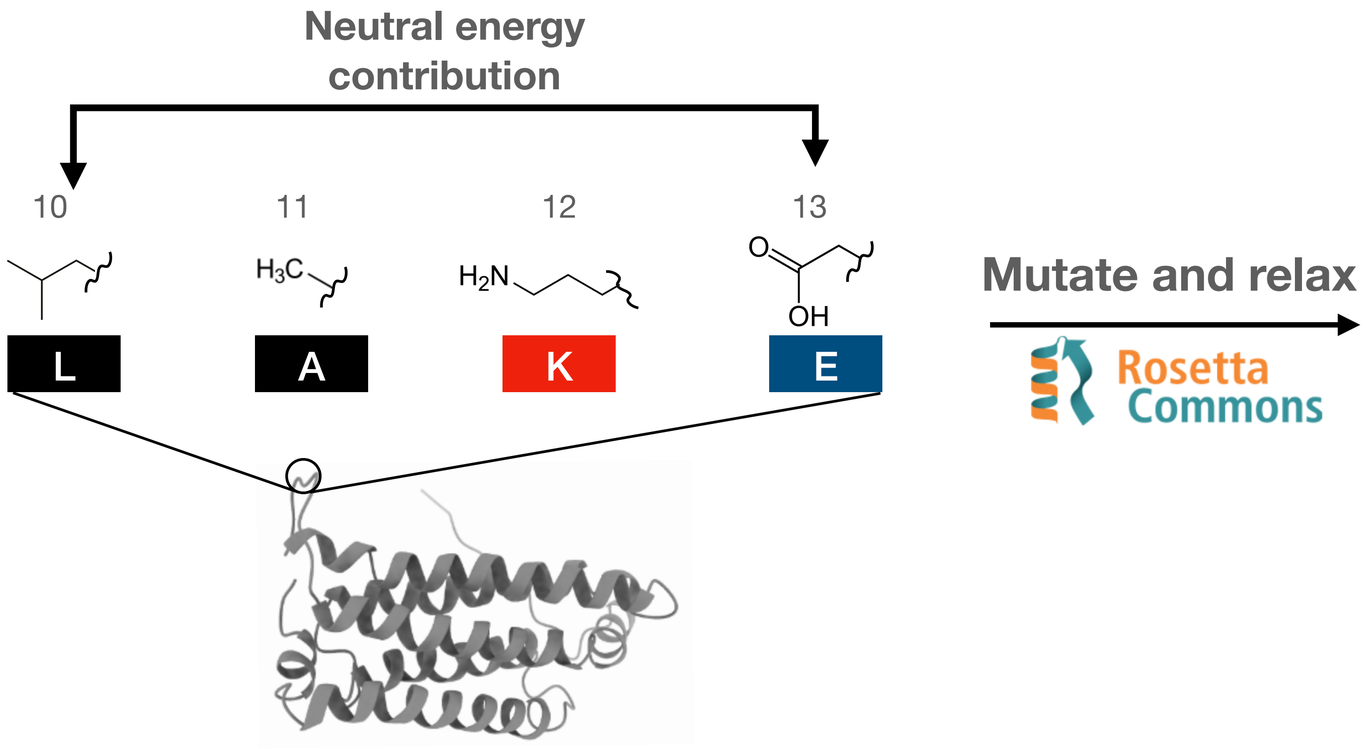


Reference protein

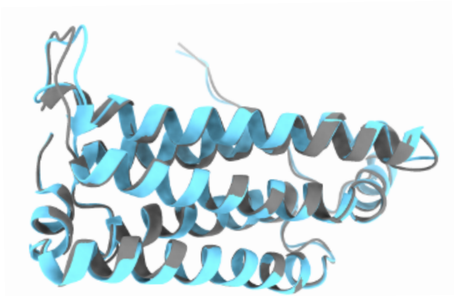


Protein variant: L10E

Use precomputed physics-based energies as input features

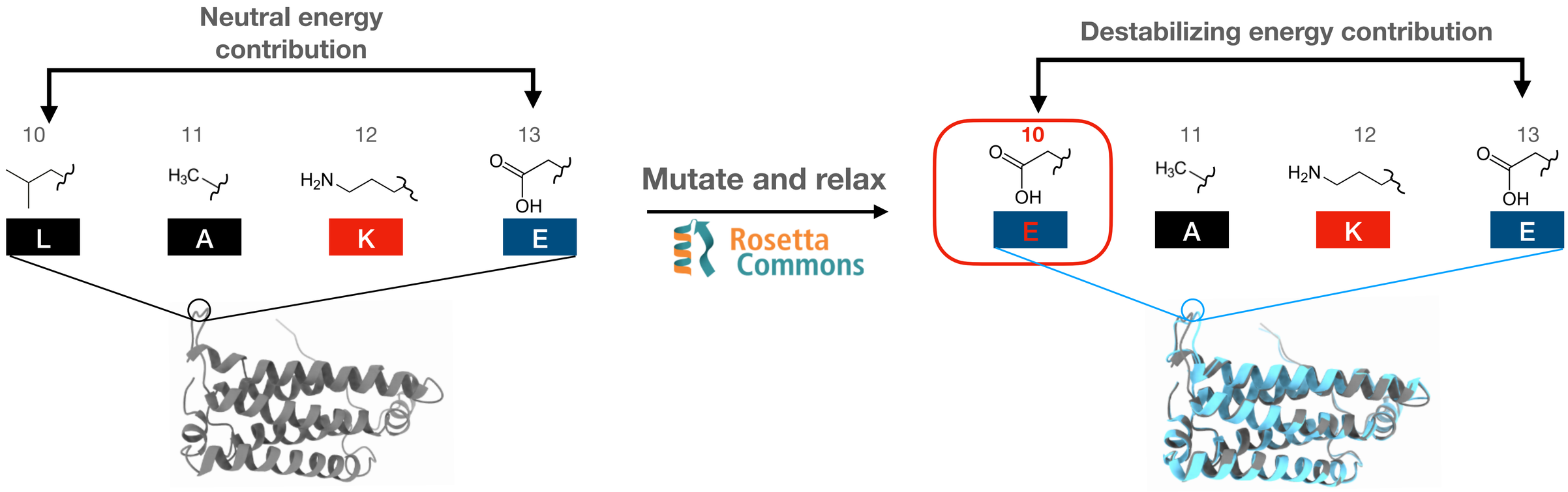


Reference protein



Protein variant: L10E

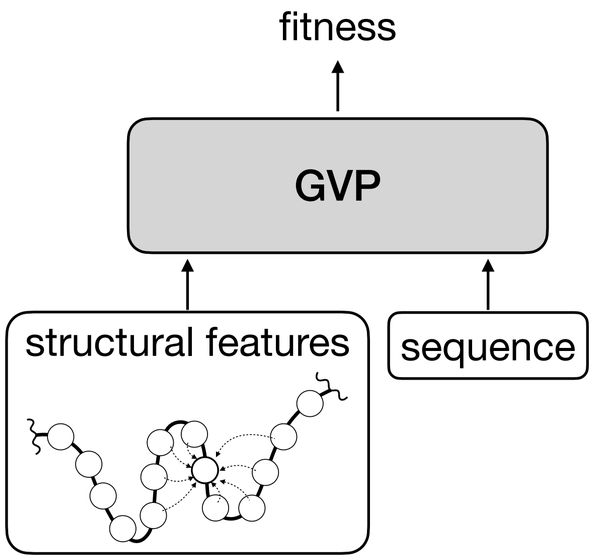
Use precomputed physics-based energies as input features



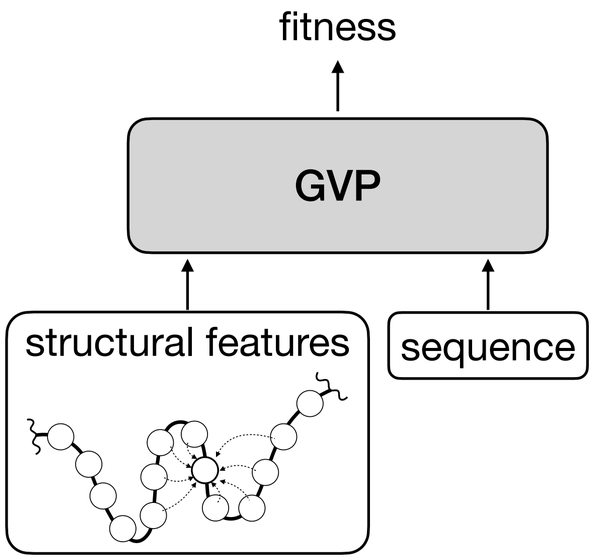
Reference protein

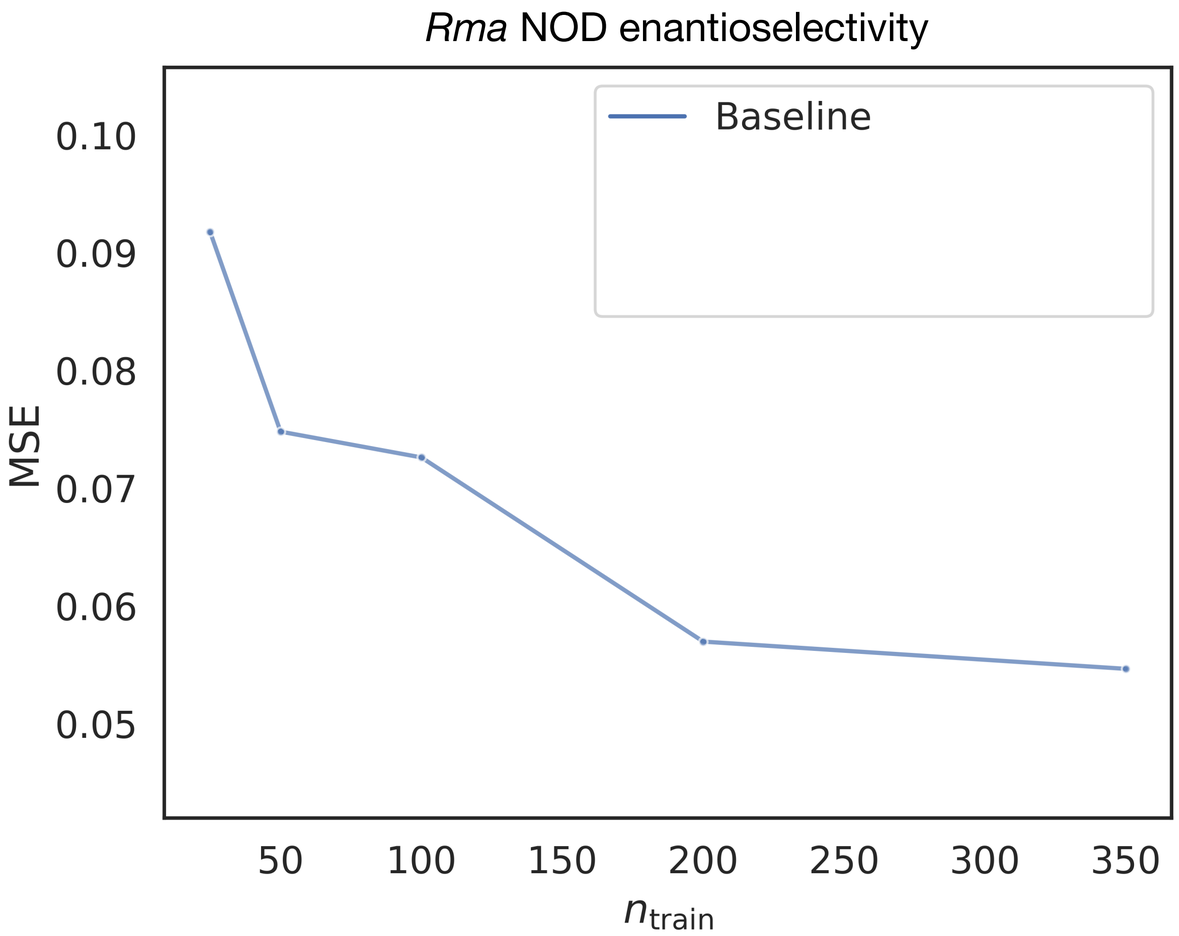
Protein variant: L10E

Energetics features improve performance in low-data regimes

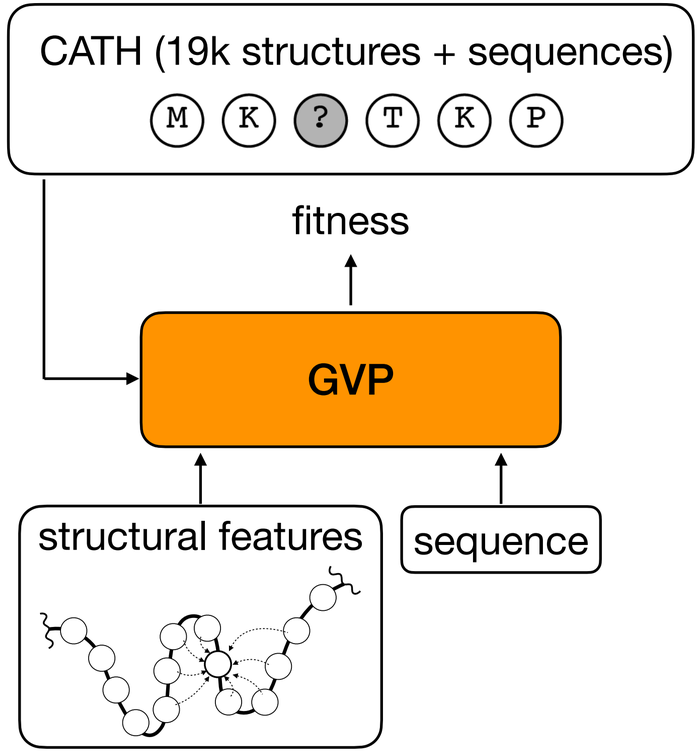


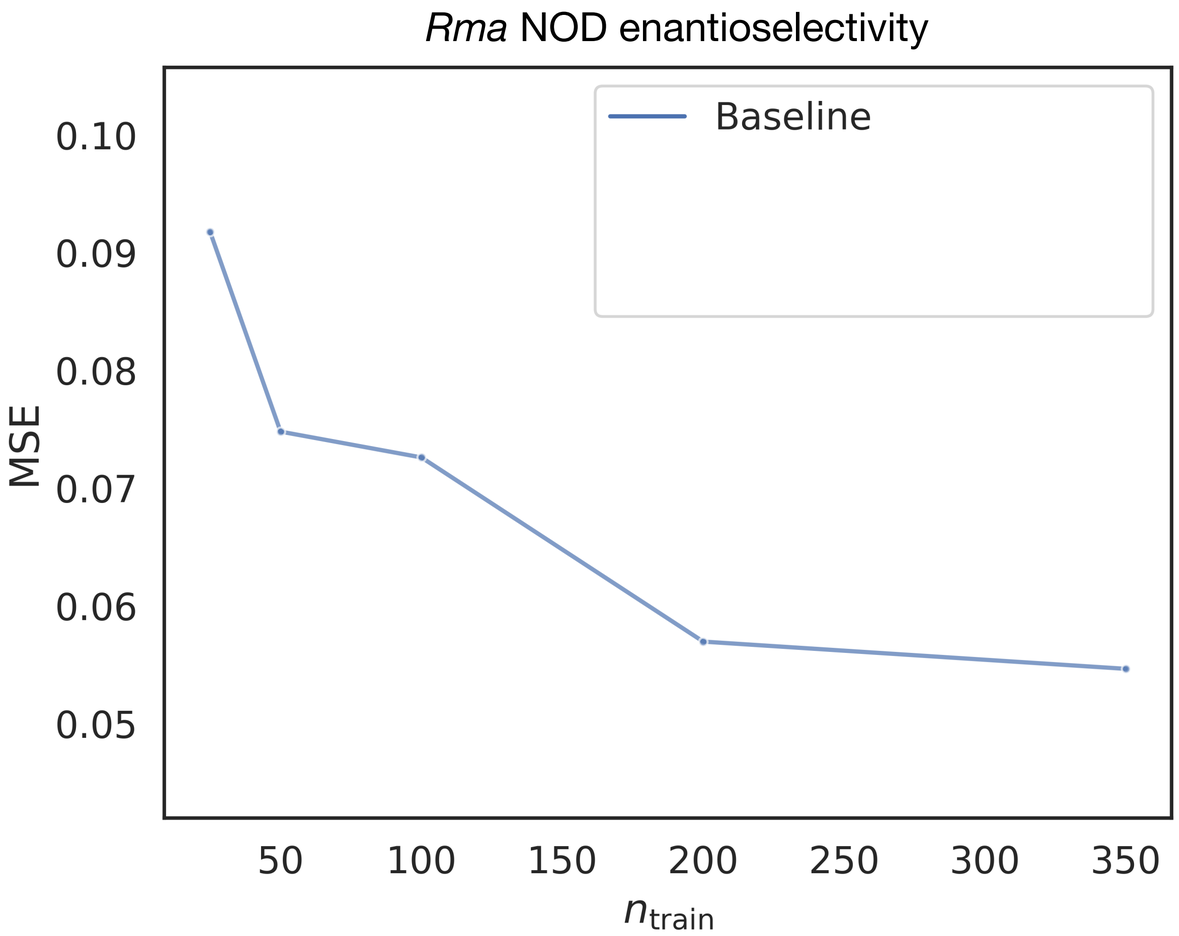
Energetics features improve performance in low-data regimes



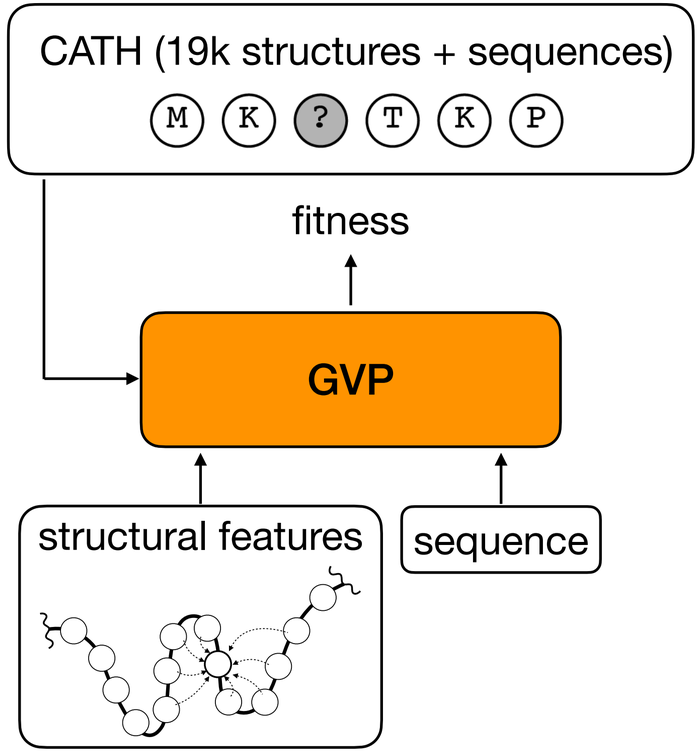


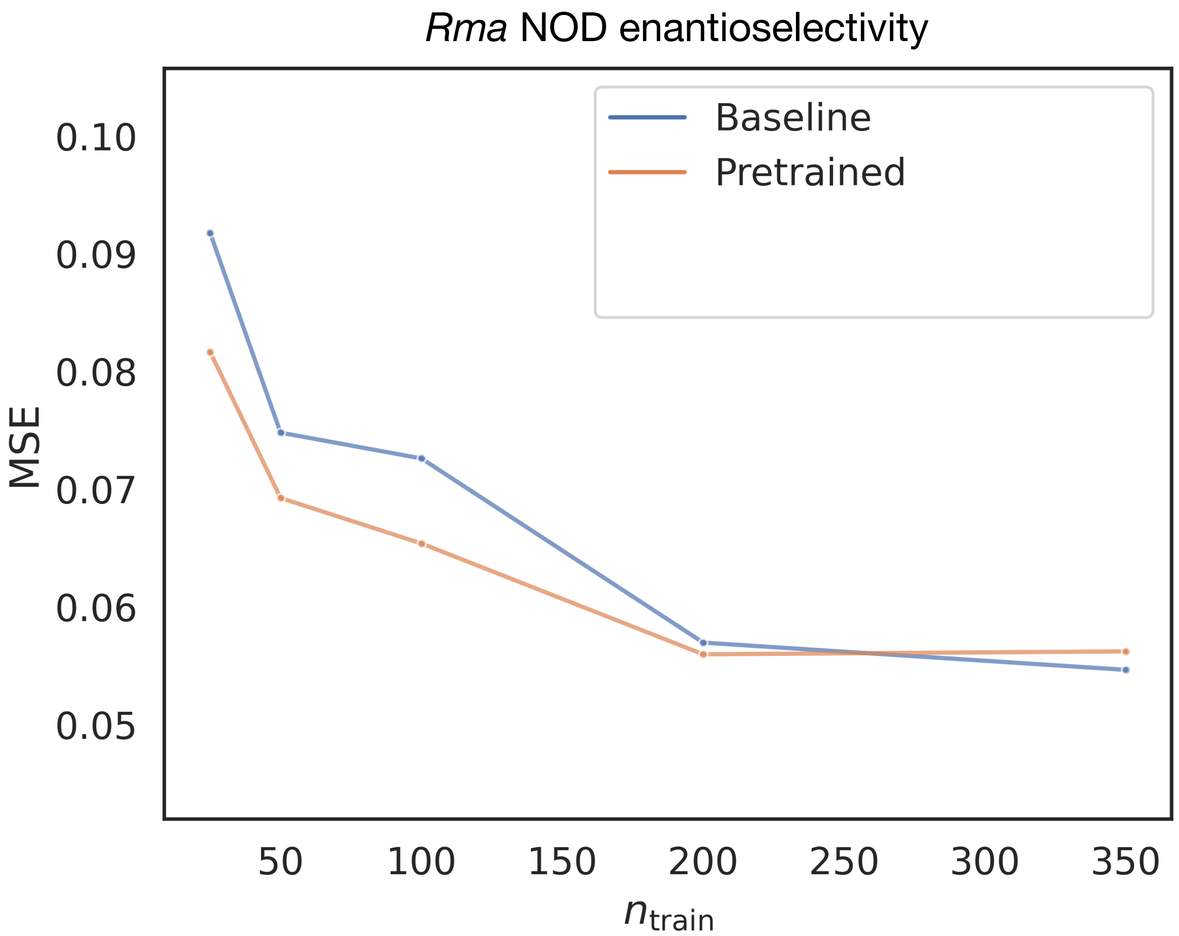
Energetics features improve performance in low-data regimes



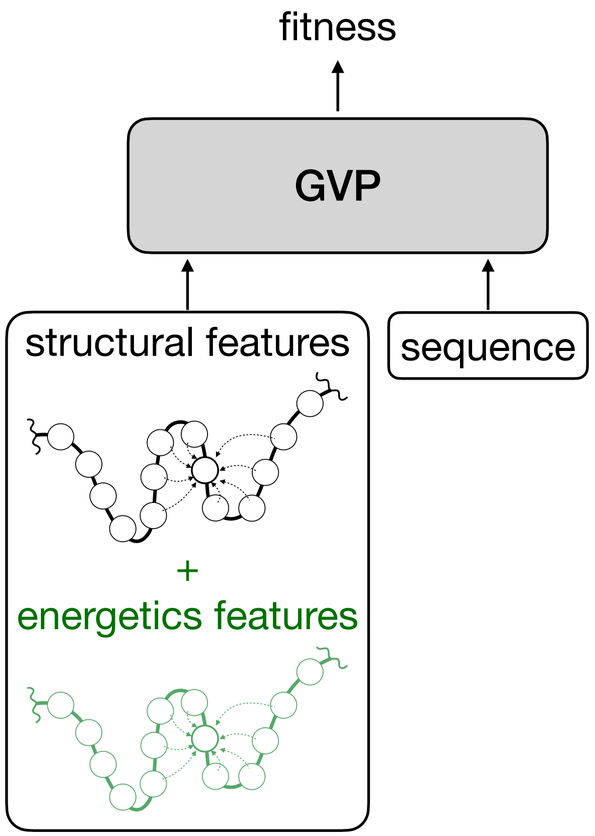


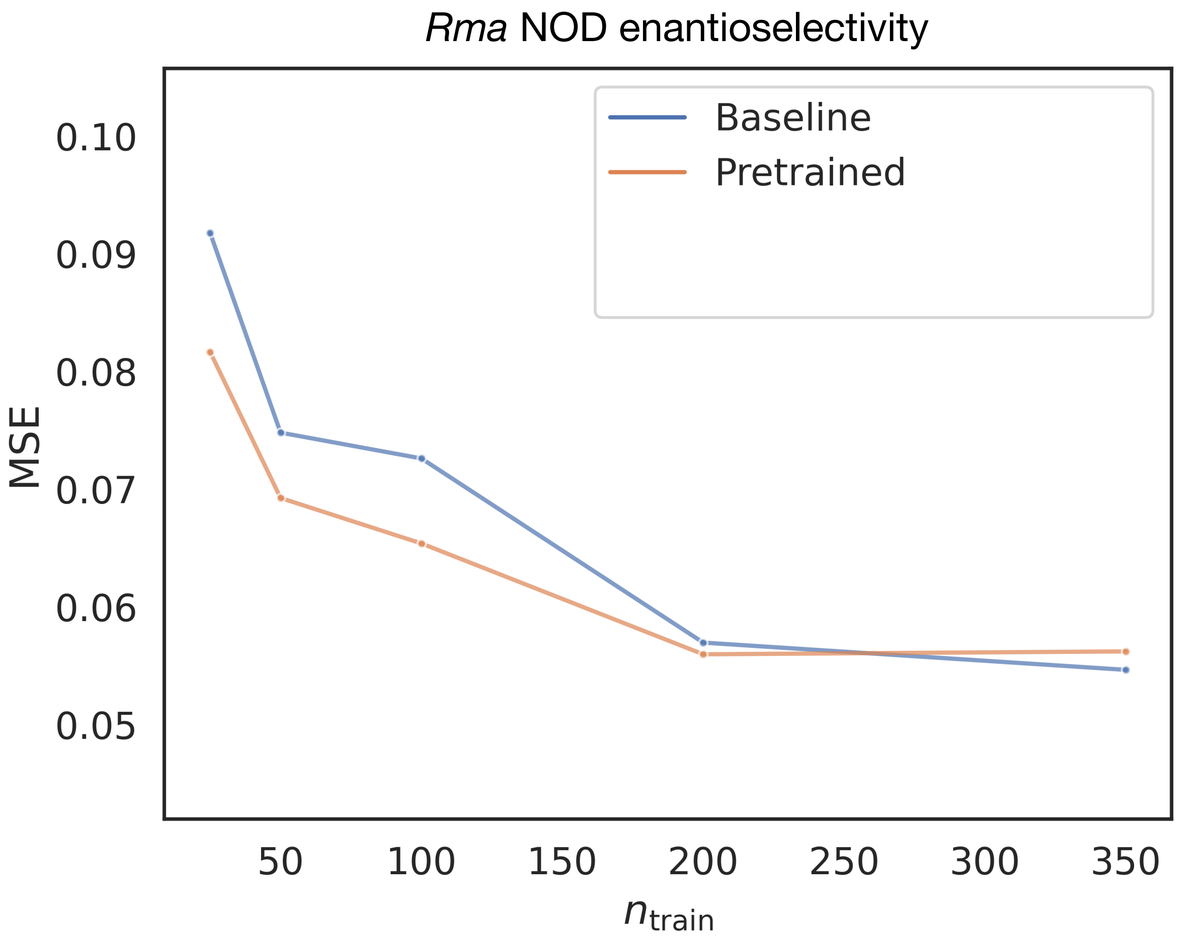
Energetics features improve performance in low-data regimes



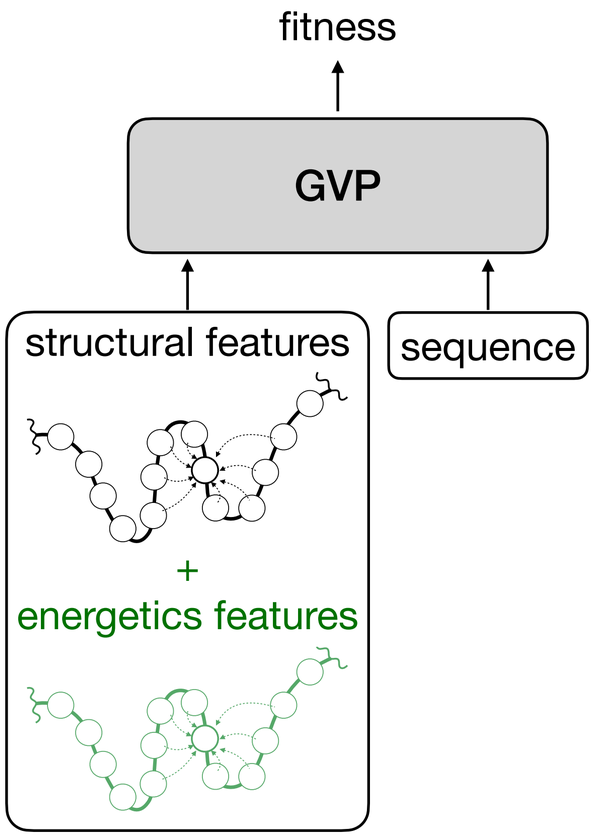


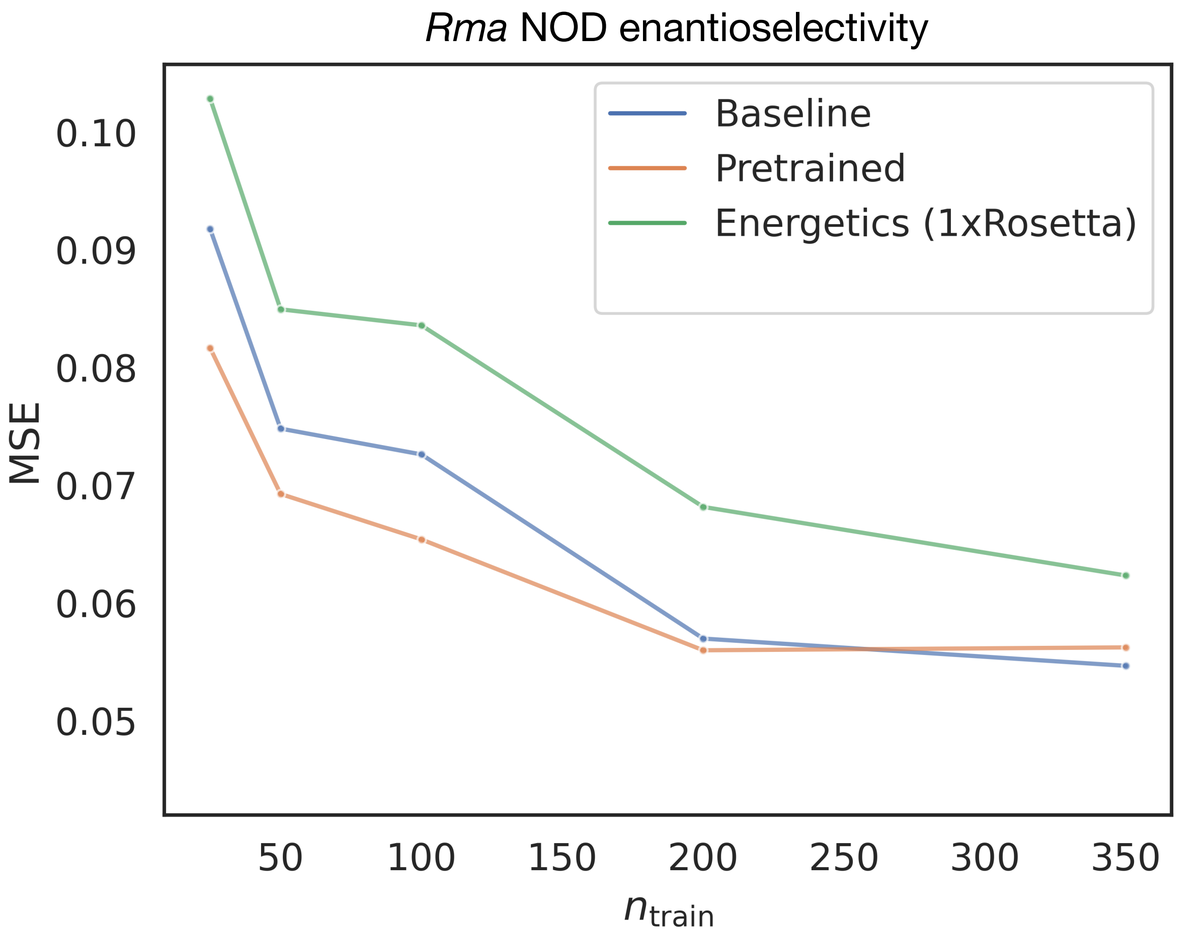
Energetics features improve performance in low-data regimes



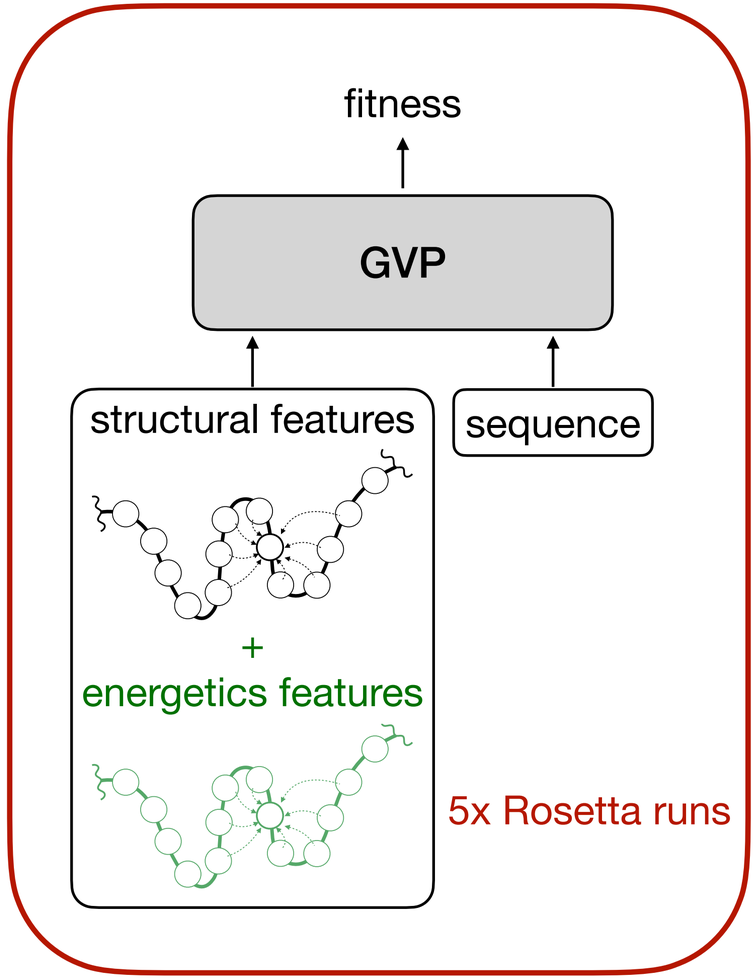


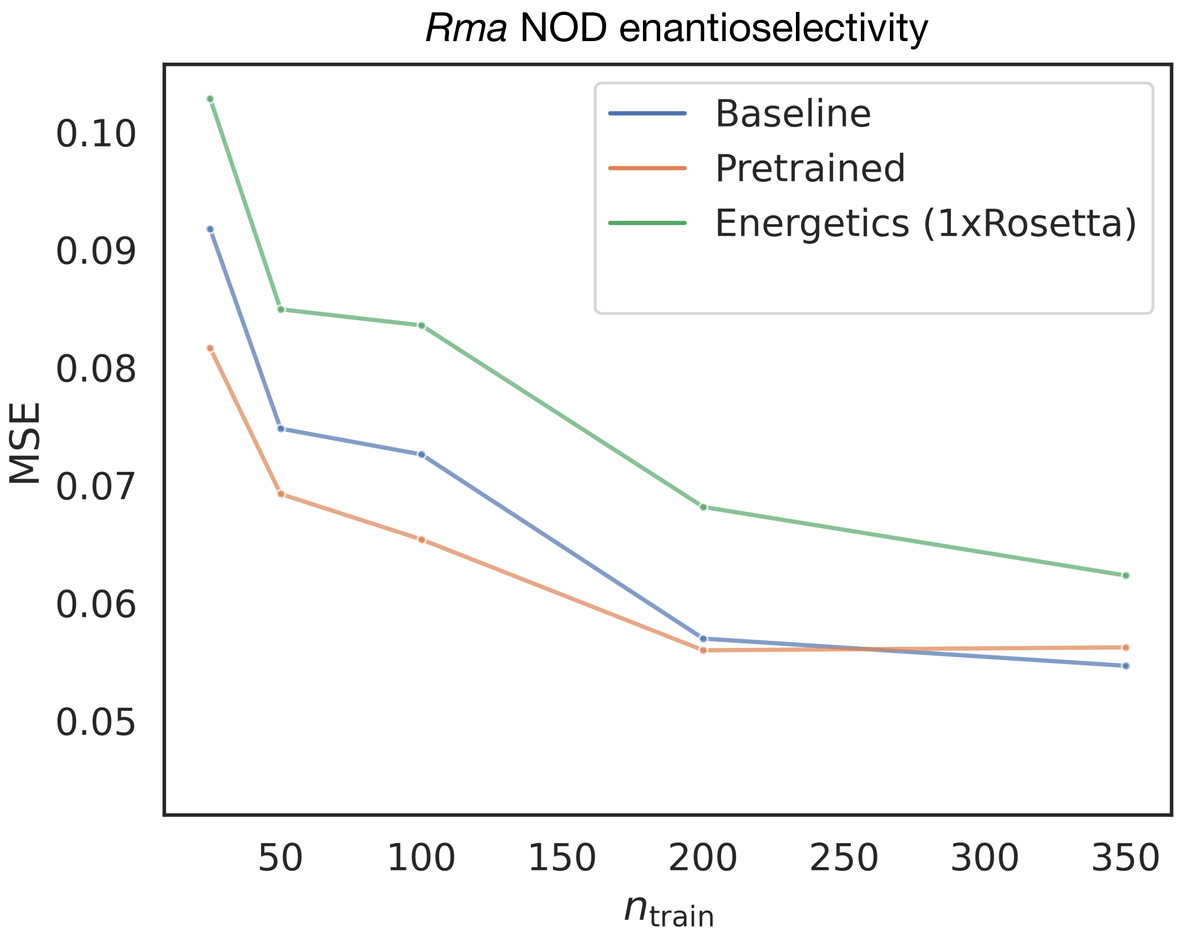
Energetics features improve performance in low-data regimes



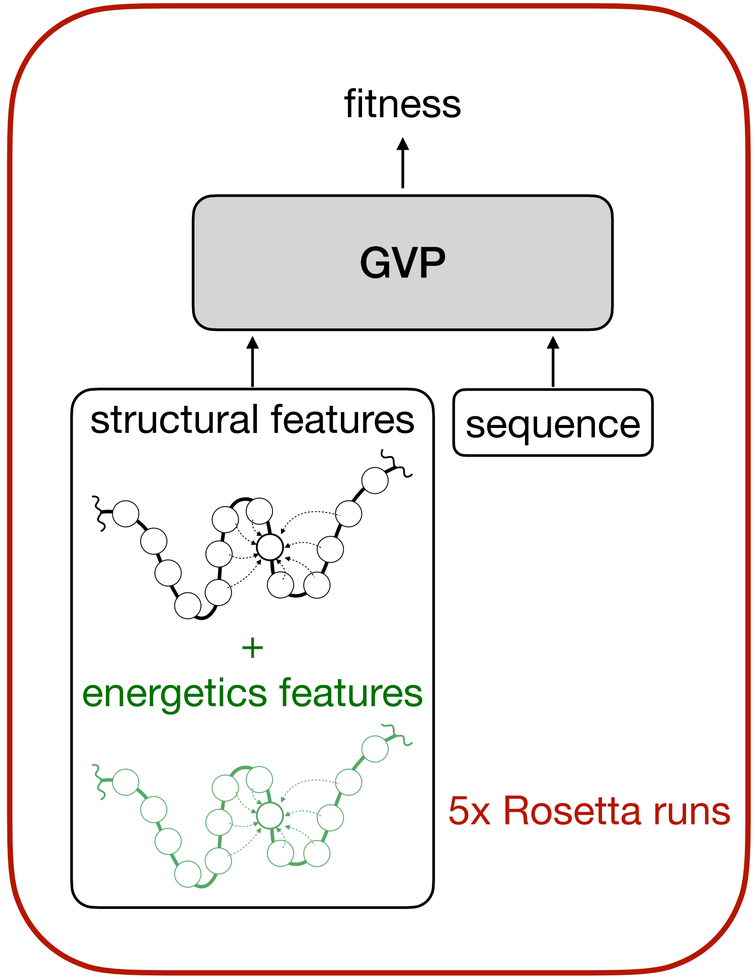


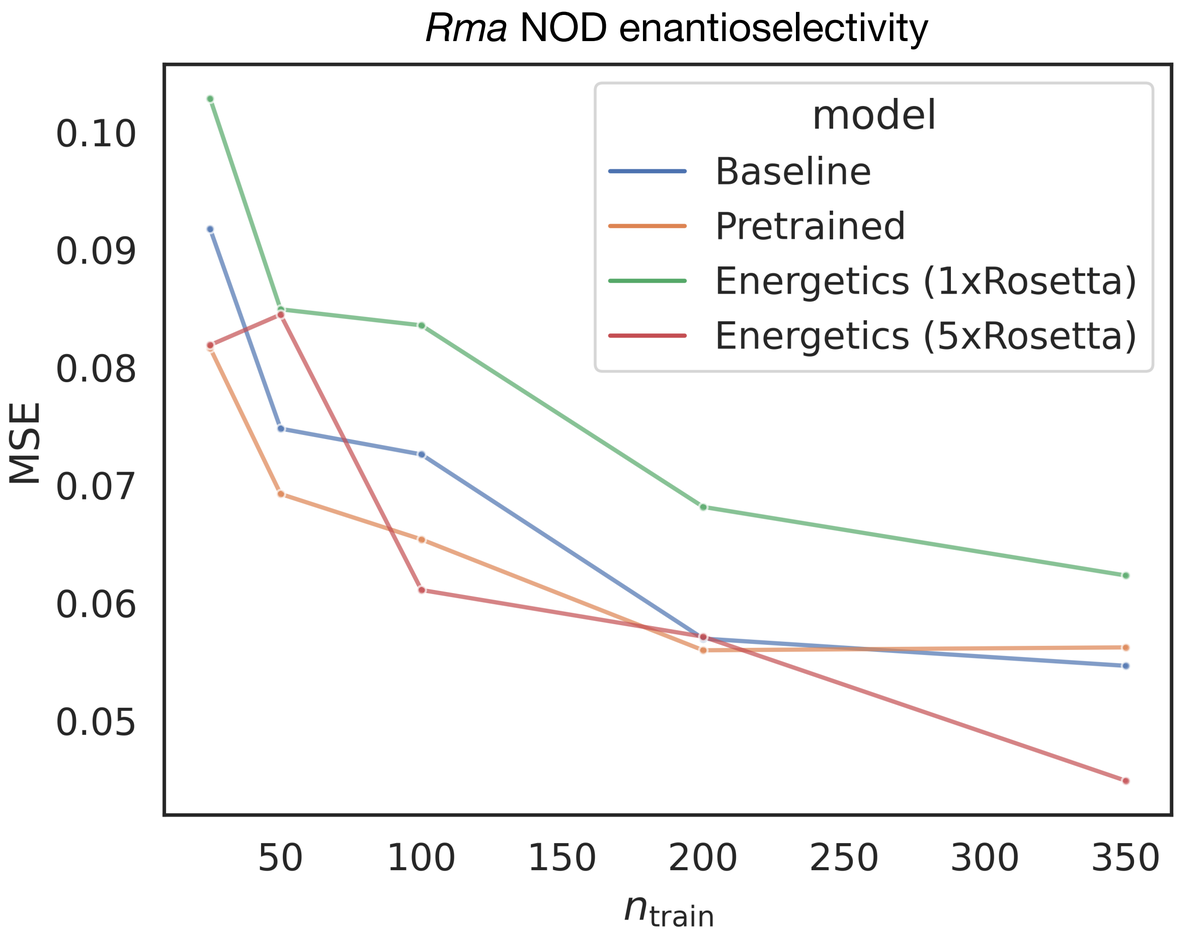
Energetics features improve performance in low-data regimes



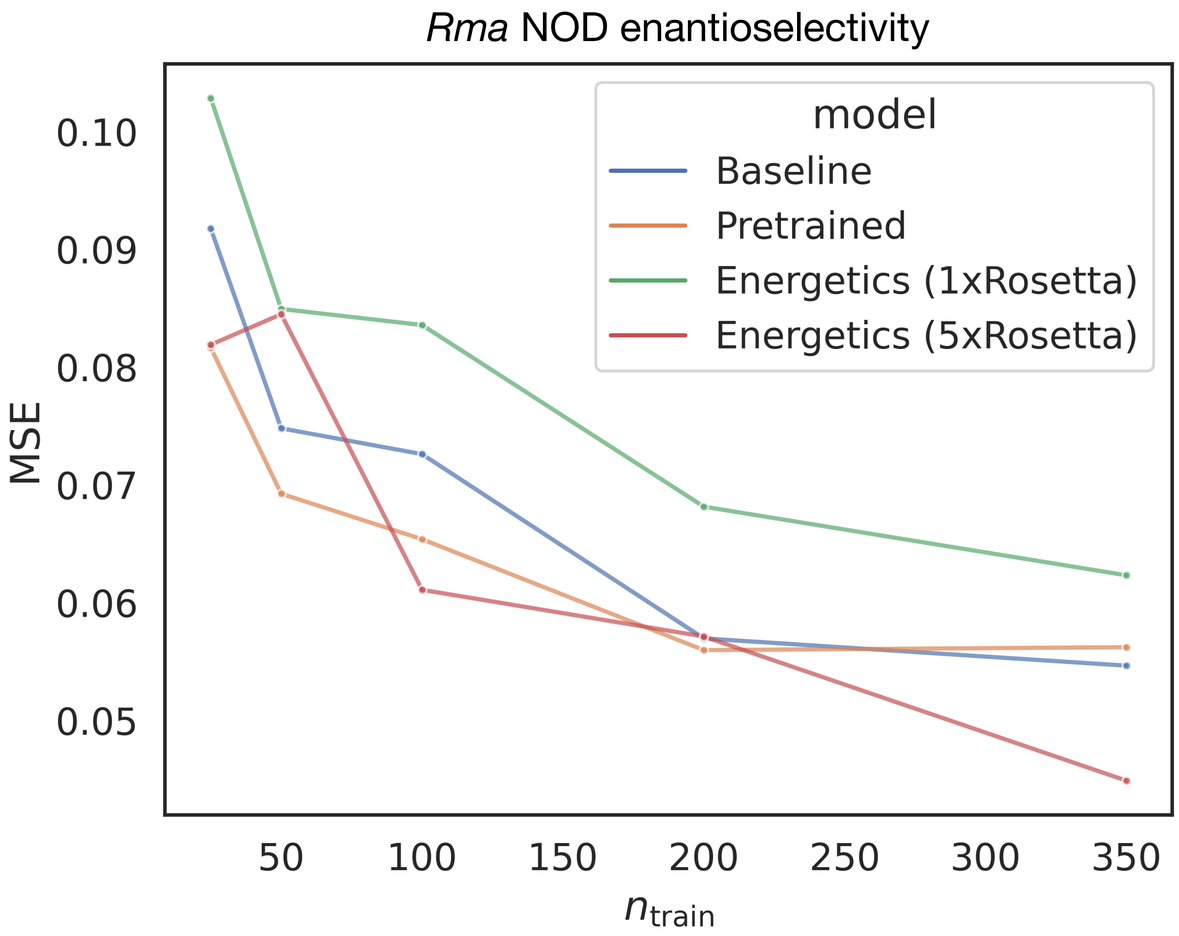


Energetics features improve performance in low-data regimes

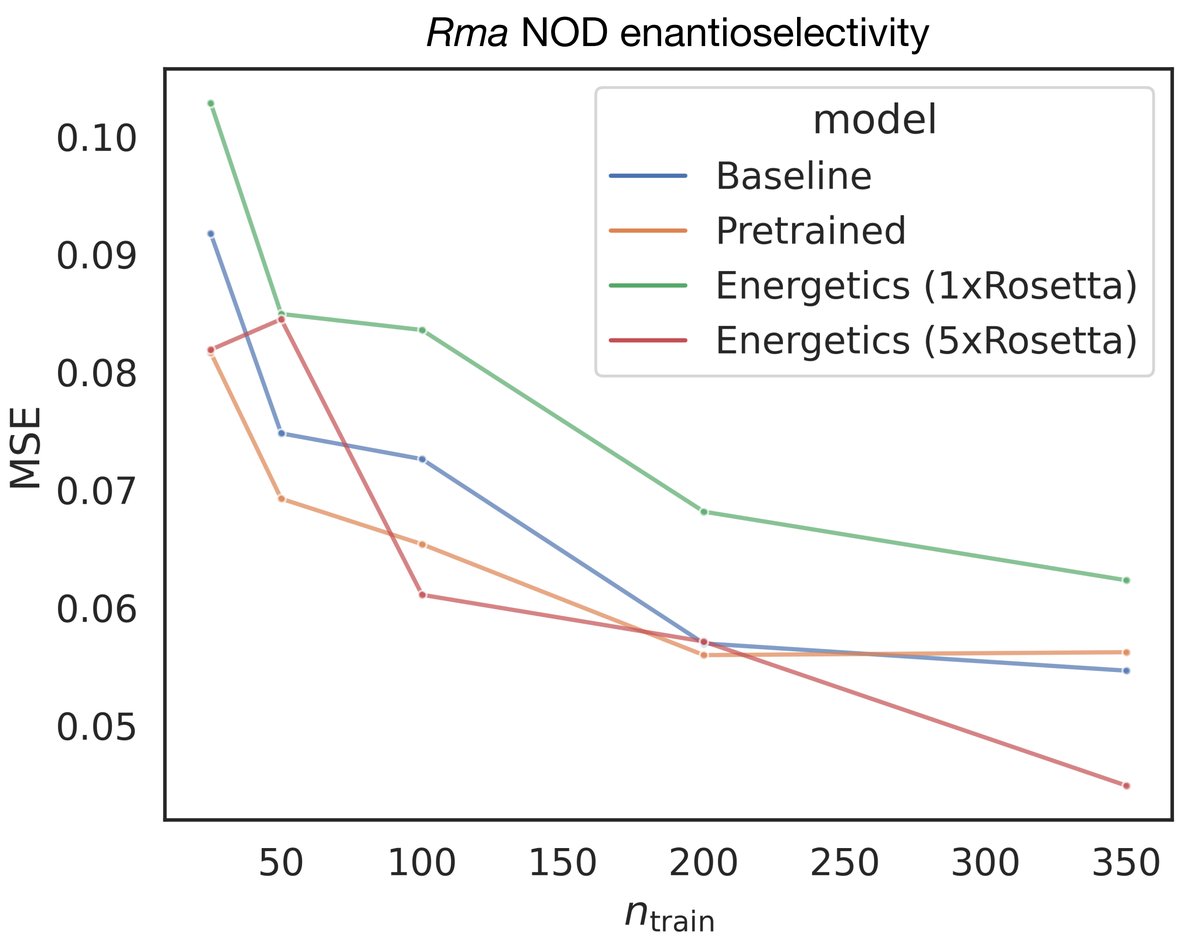


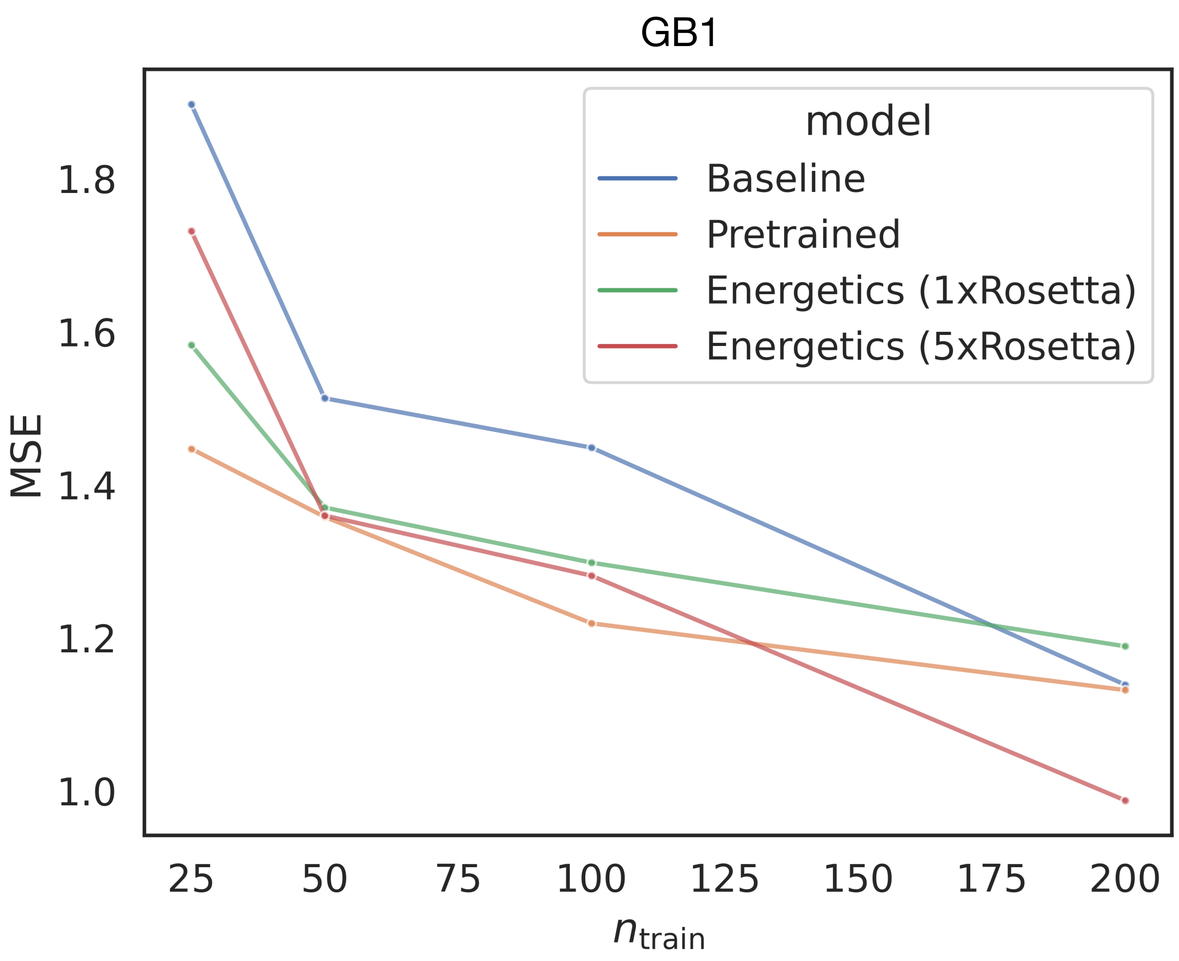


Energetics features improve performance in low-data regimes

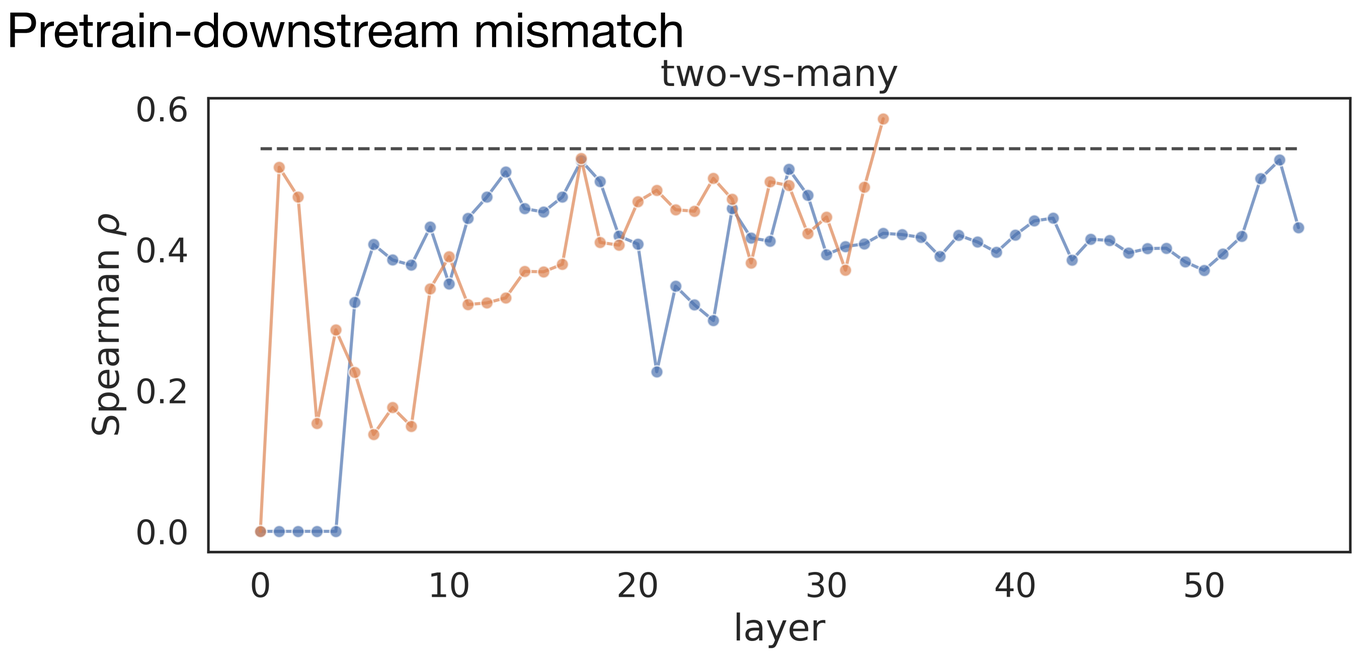


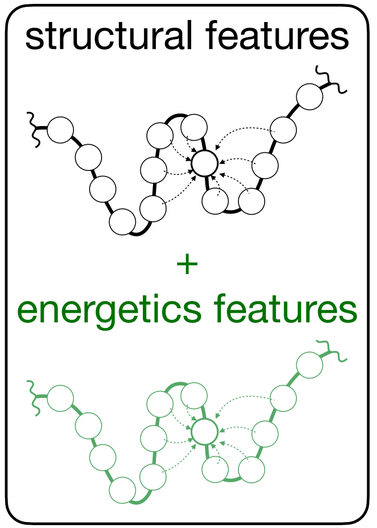
Energetics features improve performance in low-data regimes



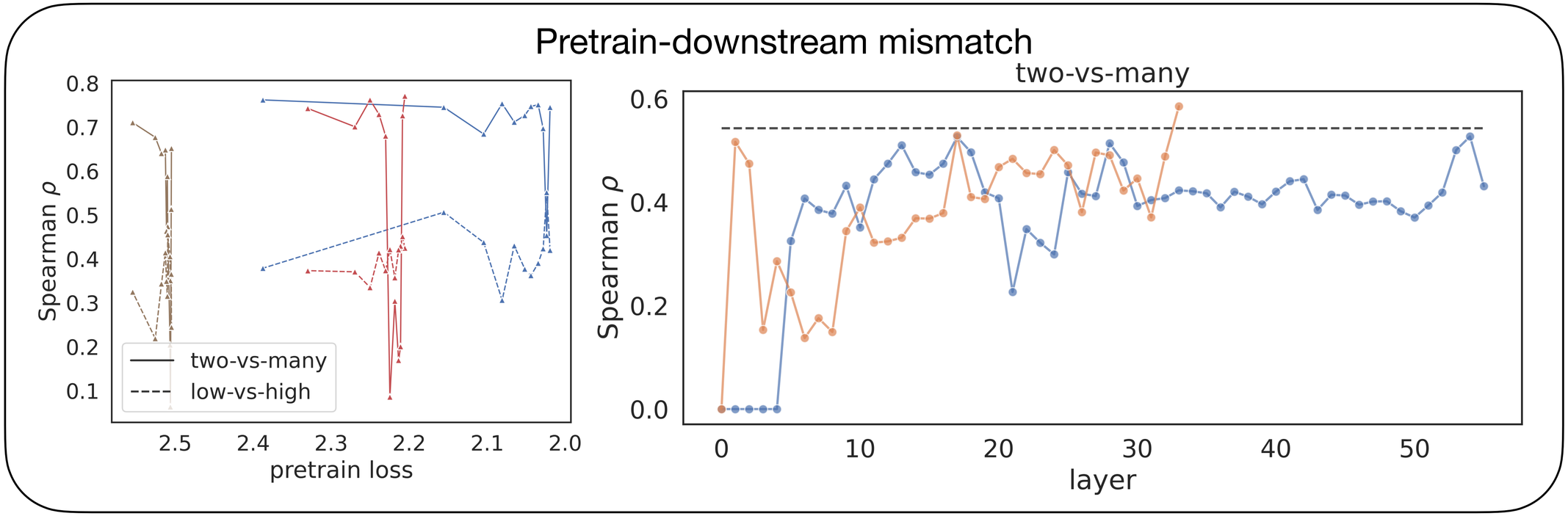


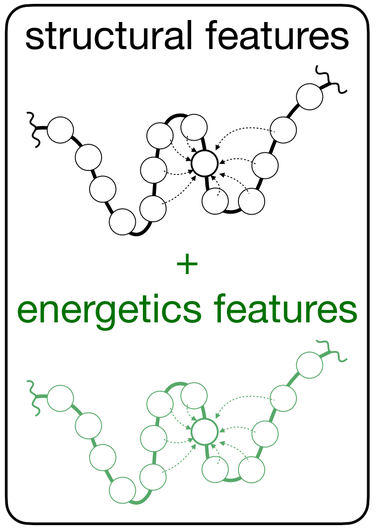
Use energetics and structure to reduce pretrain-downstream mismatch



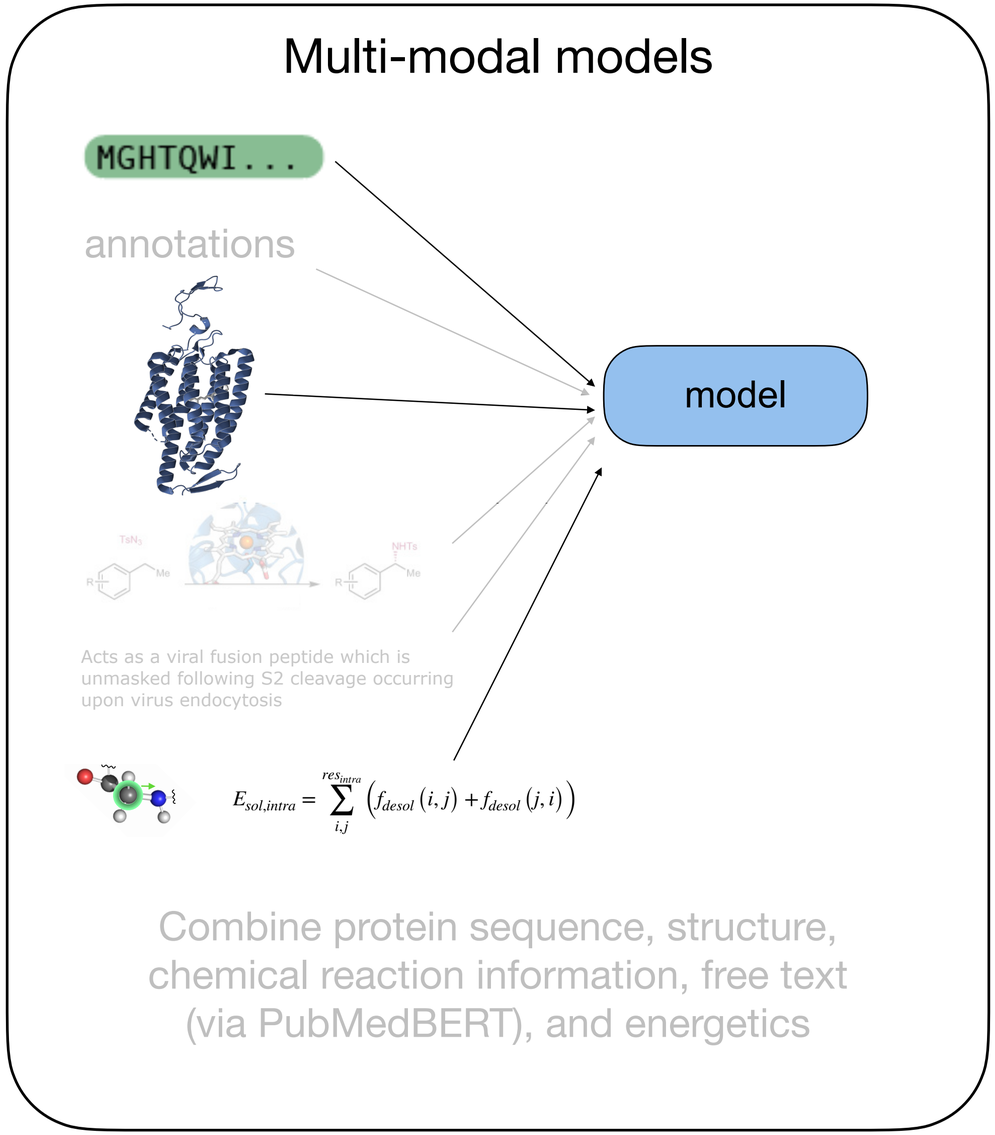


Use energetics and structure to reduce pretrain-downstream mismatch

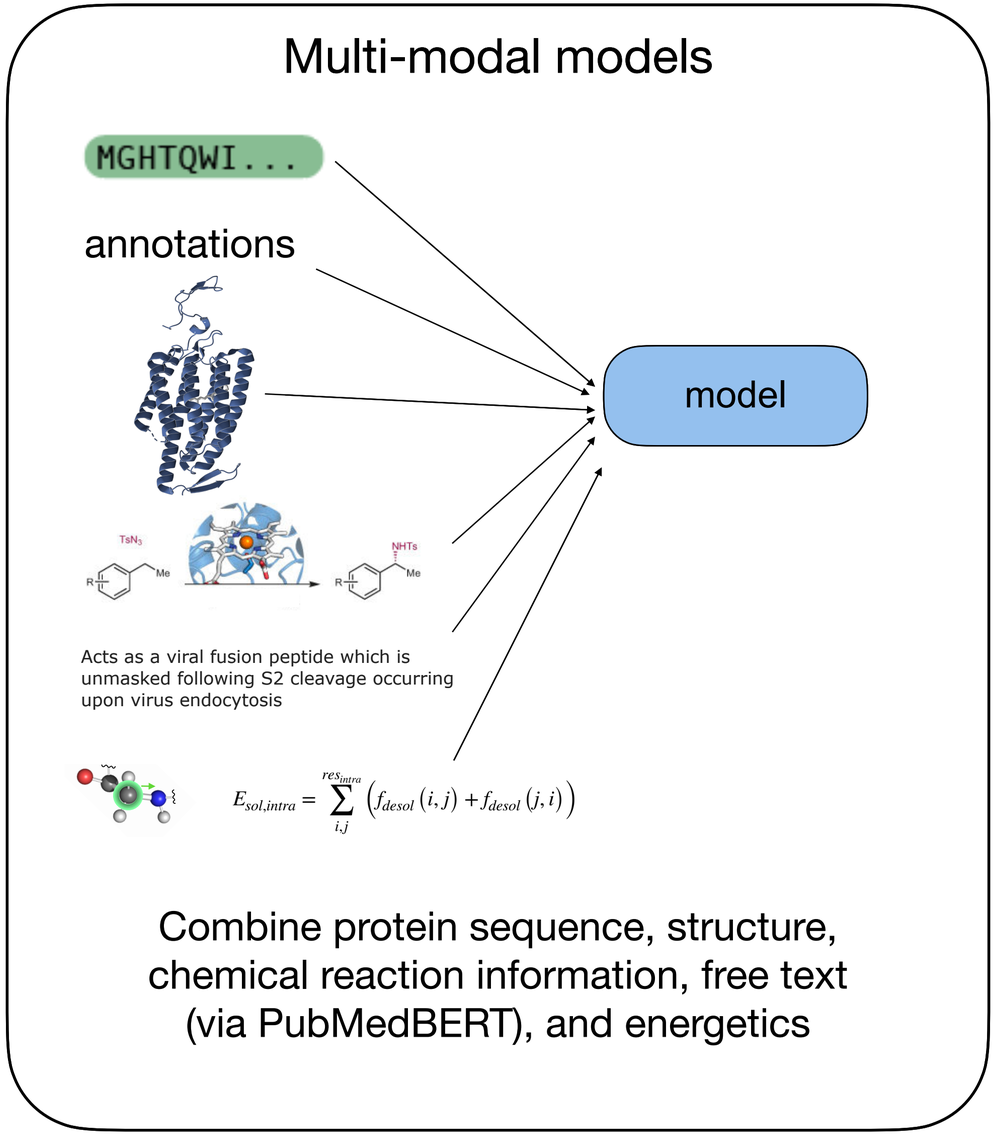




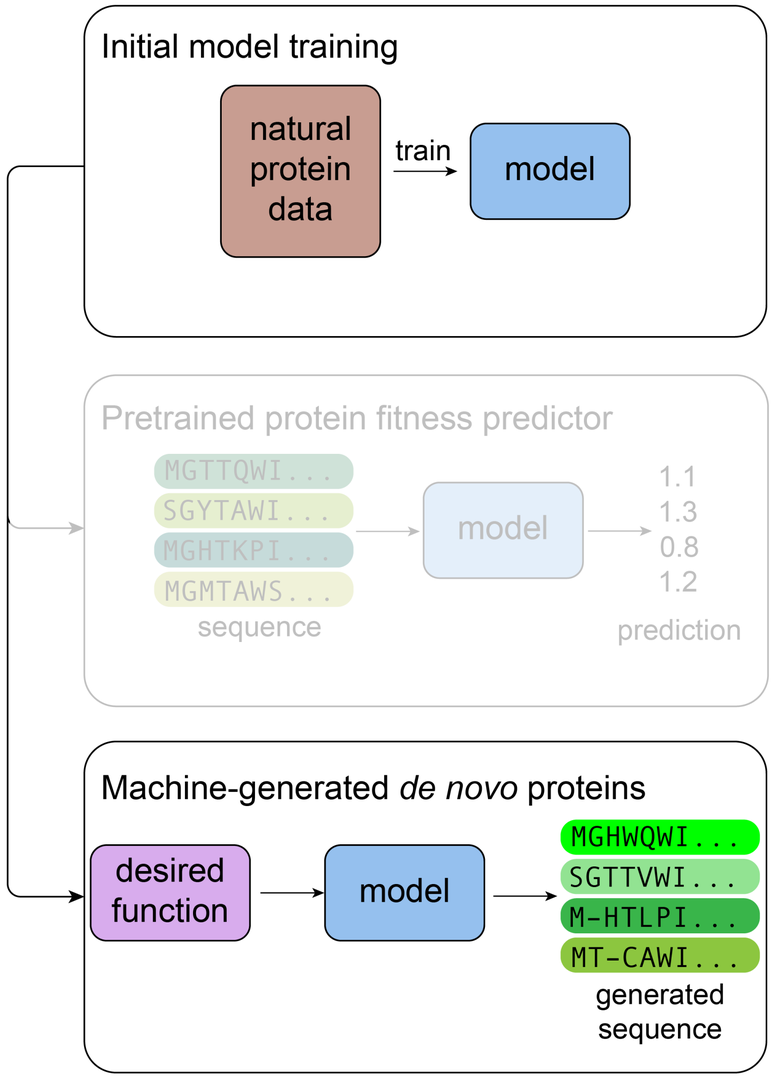
Can different modalities and pretraining tasks do better?

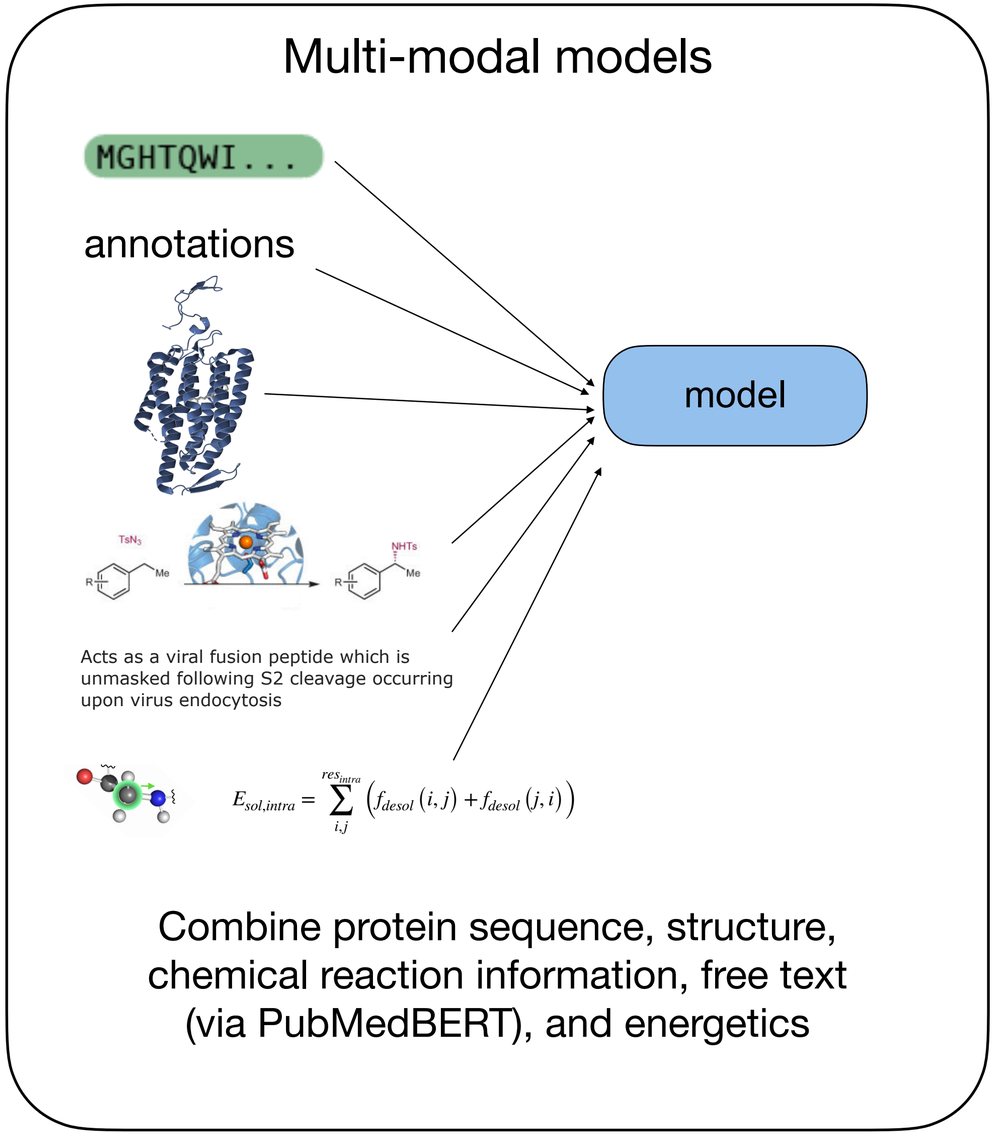


Can different modalities and pretraining tasks do better?

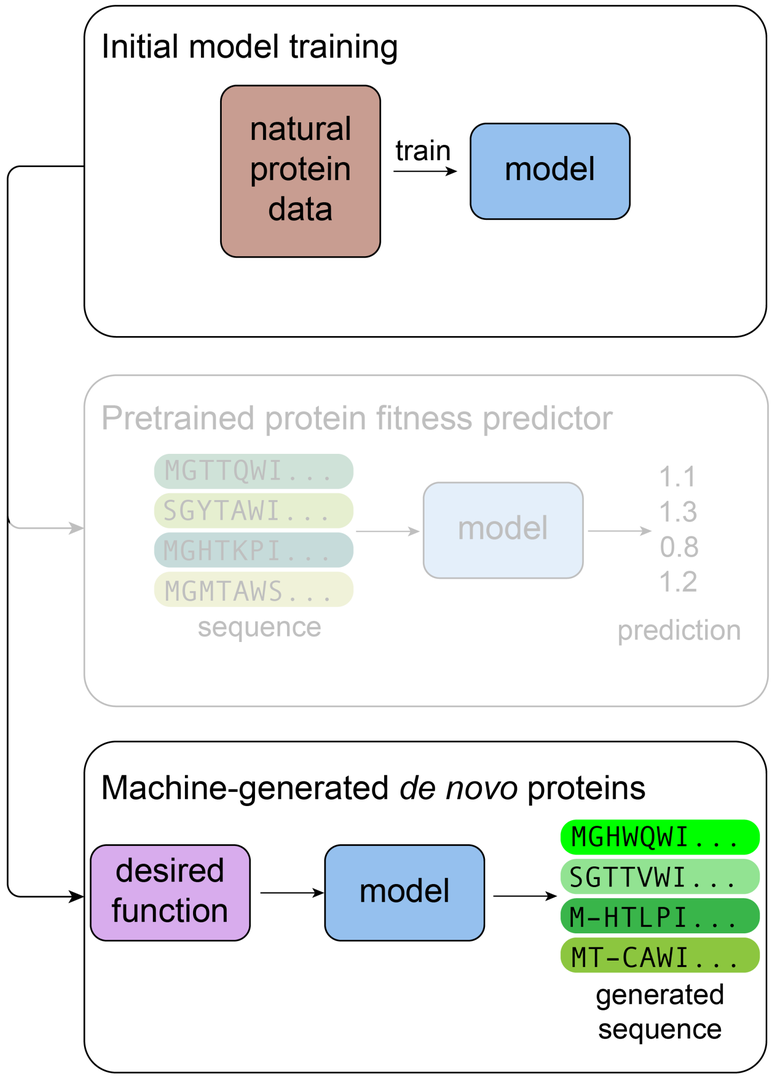


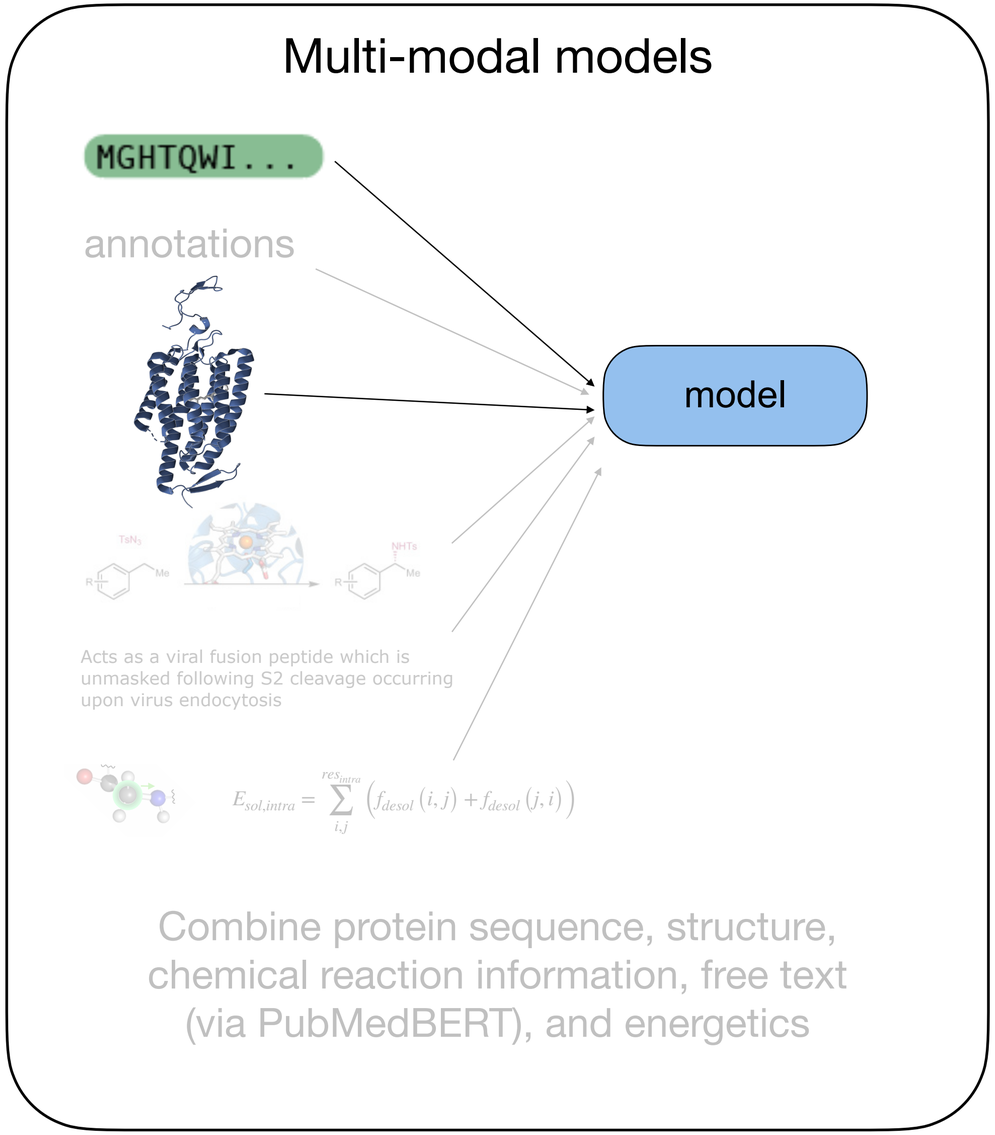
Use multiple data modalities to evaluate proteins





Use multiple data modalities to evaluate proteins





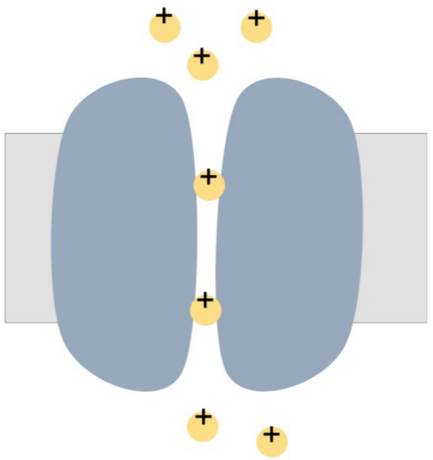
Generative models can expand functional space

Generative models can expand functional space

MGTGDHDD...

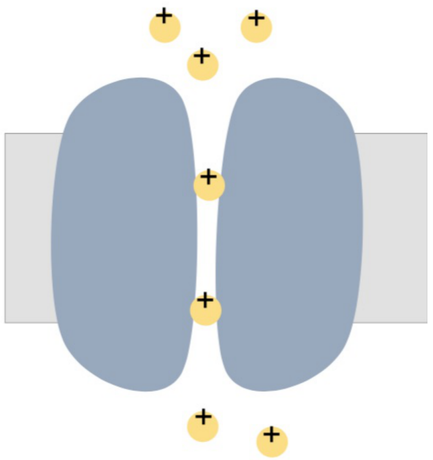
Generative models can expand functional space

MGTGDHDD...



Generative models can expand functional space

MGTGDHDD...

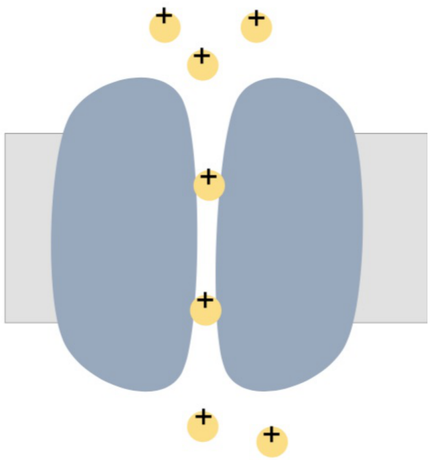


All natural proteins

Members of a protein family

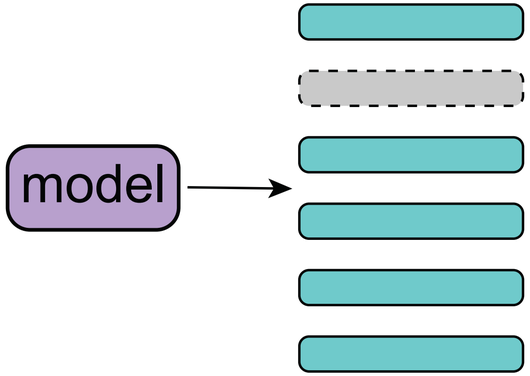
Generative models can expand functional space

MGTGDHDD...



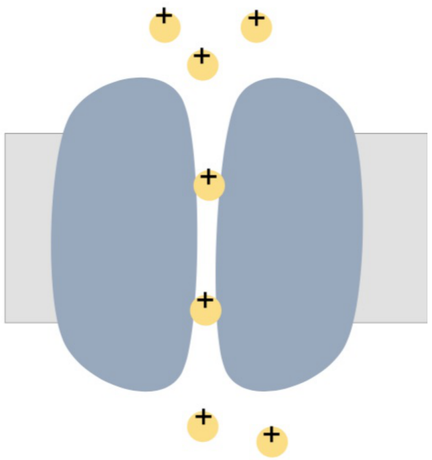
All natural proteins

Members of a protein family



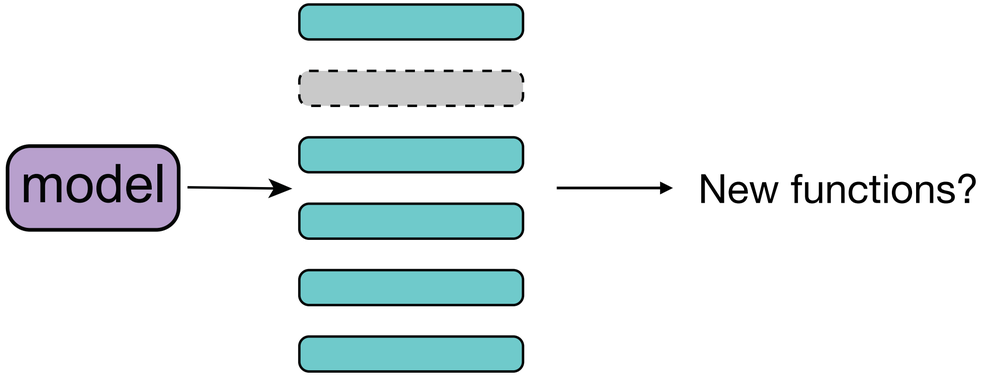
Generative models can expand functional space

MGTGDHDD...



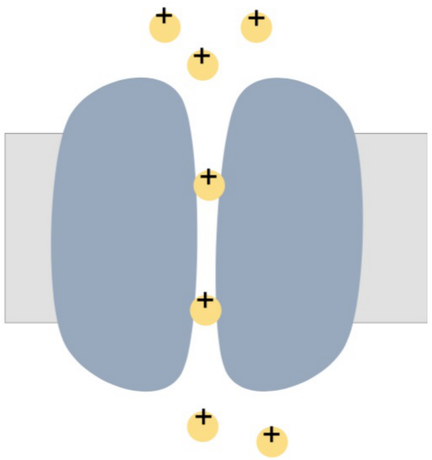
All natural proteins

Members of a protein family



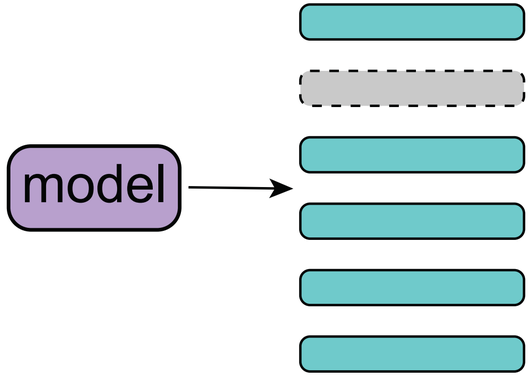
Generative models can expand functional space

MGTGDHDD...



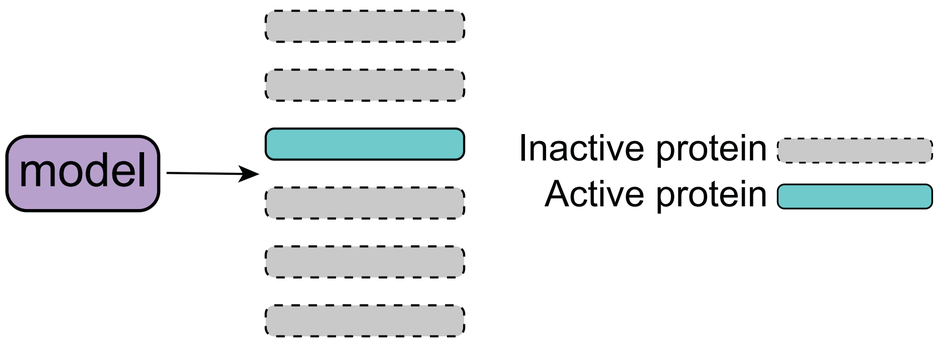
All natural proteins

Members of a protein family

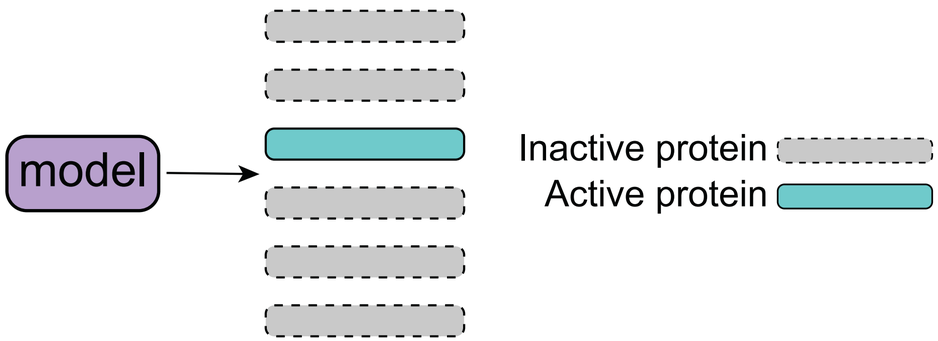


New functions?

Most generated sequences are inactive

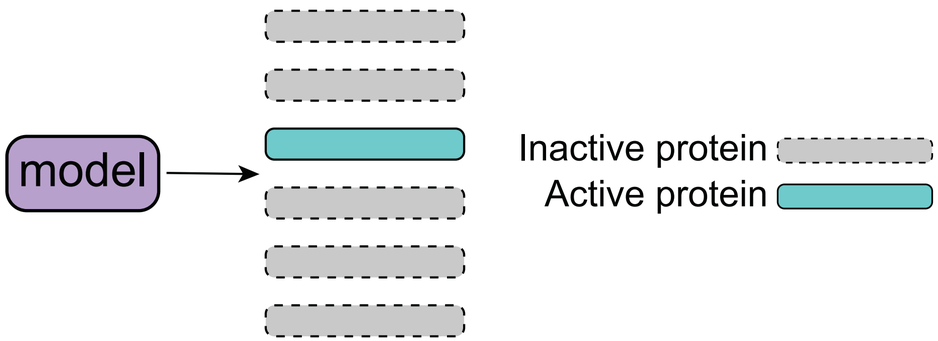


Most generated sequences are inactive



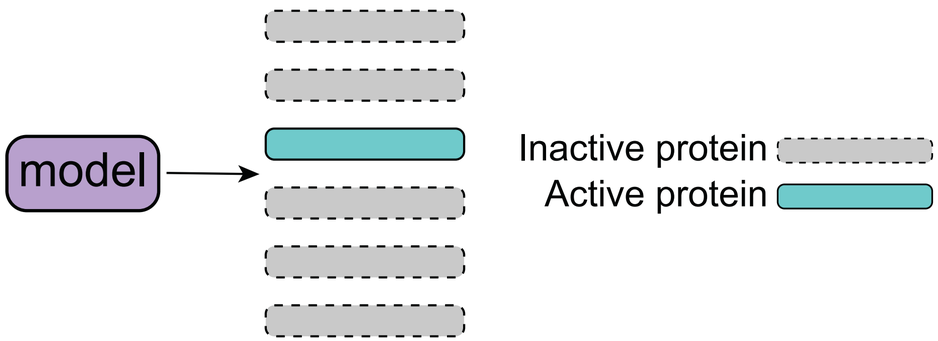
Model
GAN
ESM-MSA
Total

Most generated sequences are inactive



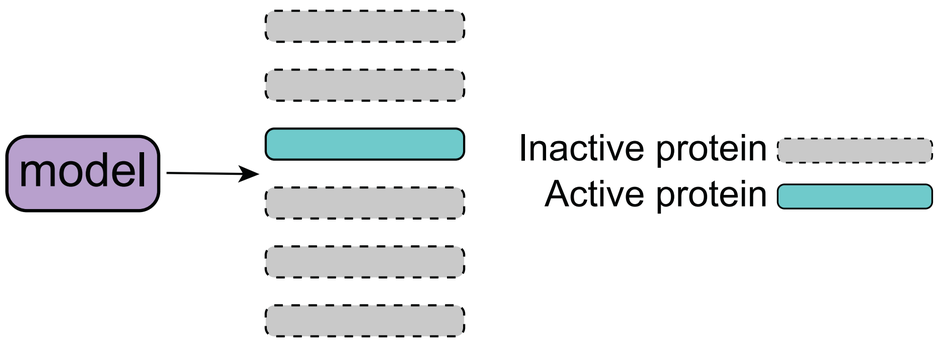
Model	Enzyme
GAN	MDH
	CuSOD
ESM-MSA	MDH
	CuSOD
Total	

Most generated sequences are inactive



Model	Enzyme	Naive
GAN	MDH	0 / 18
	CuSOD	2 / 18
ESM-MSA	MDH	0 / 17
	CuSOD	0 / 18
Total		2 / 71

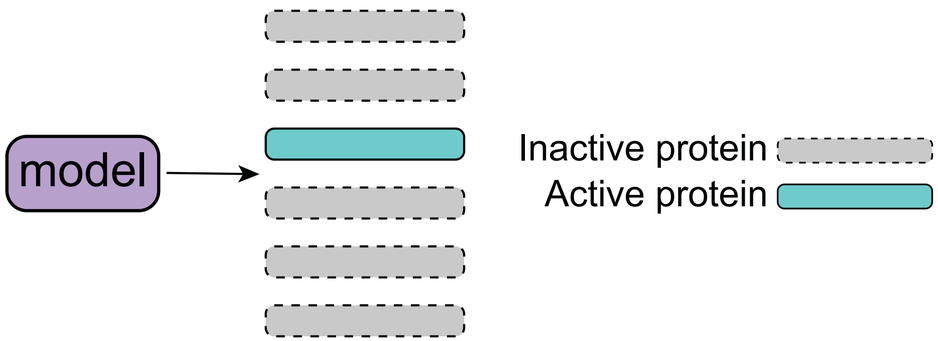
Most generated sequences are inactive



- Destabilized

Model	Enzyme	Naive
GAN	MDH	0 / 18
	CuSOD	2 / 18
ESM-MSA	MDH	0 / 17
	CuSOD	0 / 18
Total		2 / 71

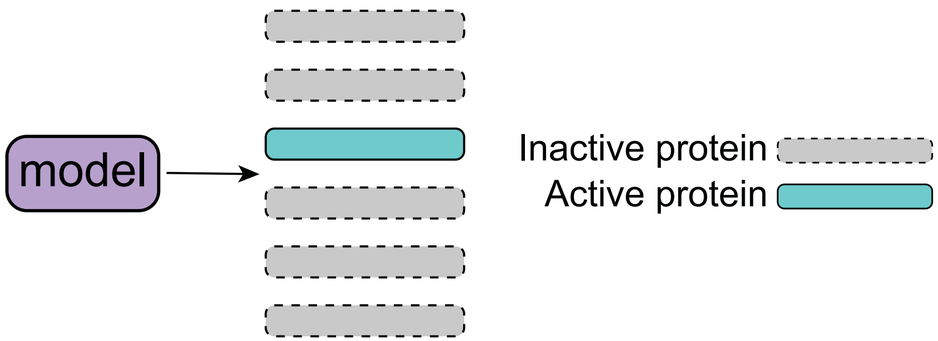
Most generated sequences are inactive



- Destabilized
- Mismatched signal peptide

Model	Enzyme	Naive
GAN	MDH	0 / 18
	CuSOD	2 / 18
ESM-MSA	MDH	0 / 17
	CuSOD	0 / 18
Total		2 / 71

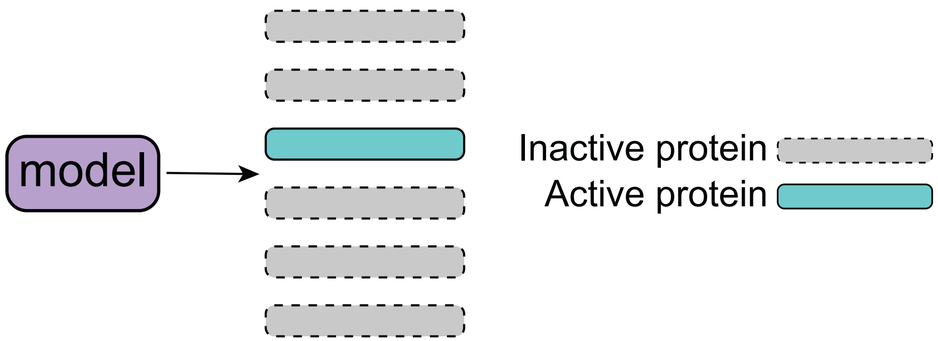
Most generated sequences are inactive



Model	Enzyme	Naive
GAN	MDH	0 / 18
	CuSOD	2 / 18
ESM-MSA	MDH	0 / 17
	CuSOD	0 / 18
Total		2 / 71

- Destabilized
- Mismatched signal peptide
- Transmembrane domain

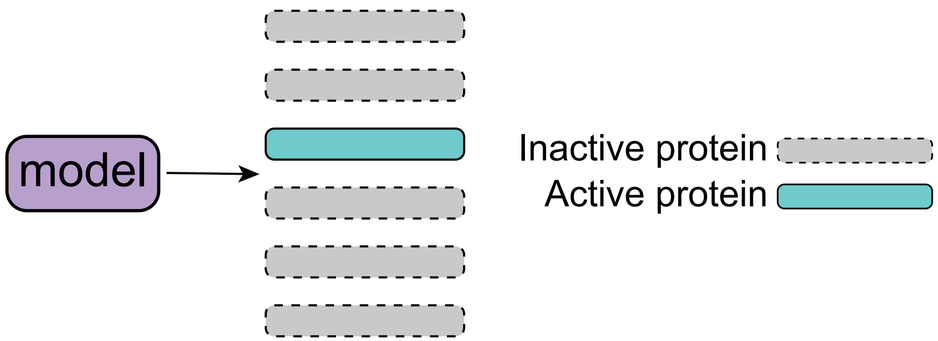
Most generated sequences are inactive



Model	Enzyme	Naive
GAN	MDH	0 / 18
	CuSOD	2 / 18
ESM-MSA	MDH	0 / 17
	CuSOD	0 / 18
Total		2 / 71

- Destabilized
- Mismatched signal peptide
- Transmembrane domain
- Long repeats

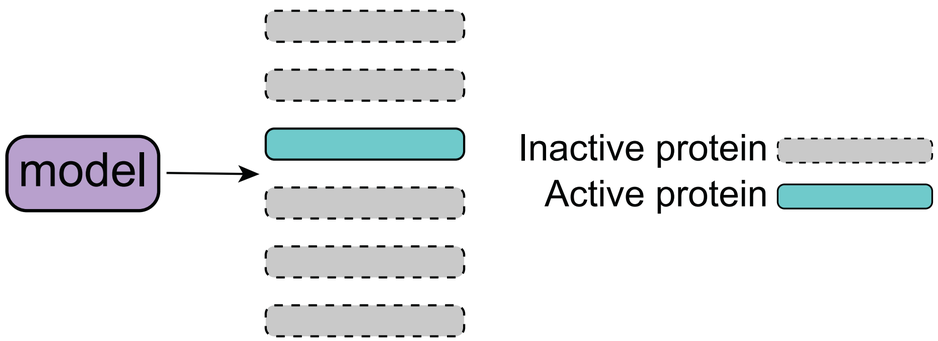
Most generated sequences are inactive



Model	Enzyme	Naive
GAN	MDH	0 / 18
	CuSOD	2 / 18
ESM-MSA	MDH	0 / 17
	CuSOD	0 / 18
Total		2 / 71

- Destabilized
- Mismatched signal peptide
- Transmembrane domain
- Long repeats
- No catalytic residue

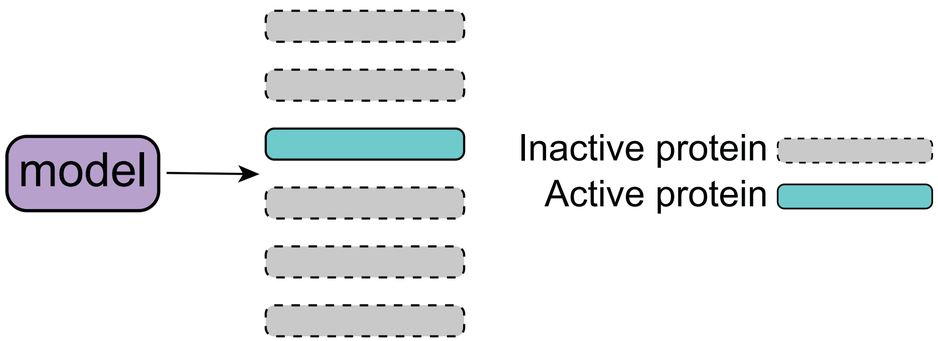
Most generated sequences are inactive



Model	Enzyme	Naive
GAN	MDH	0 / 18
	CuSOD	2 / 18
ESM-MSA	MDH	0 / 17
	CuSOD	0 / 18
Total		2 / 71

- Destabilized
- Mismatched signal peptide
- Transmembrane domain
- Long repeats
- No catalytic residue
- …

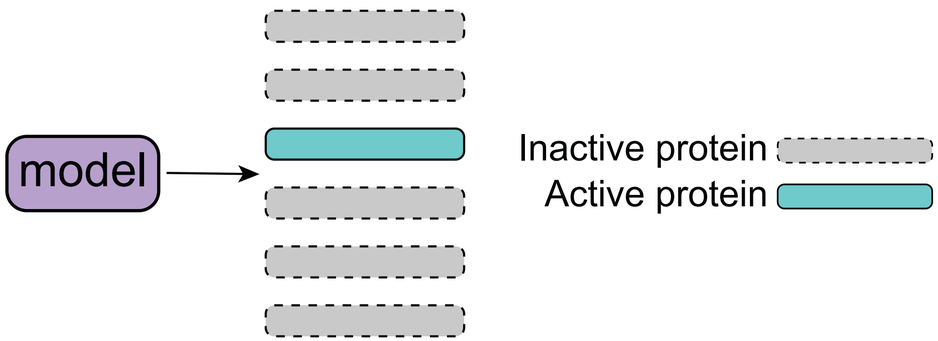
Most generated sequences are inactive



Model	Enzyme	Naive
GAN	MDH	0 / 18
	CuSOD	2 / 18
ESM-MSA	MDH	0 / 17
	CuSOD	0 / 18
Total		2 / 71

- Destabilized
- Mismatched signal peptide
- Transmembrane domain
- Long repeats
- No catalytic residue
- …

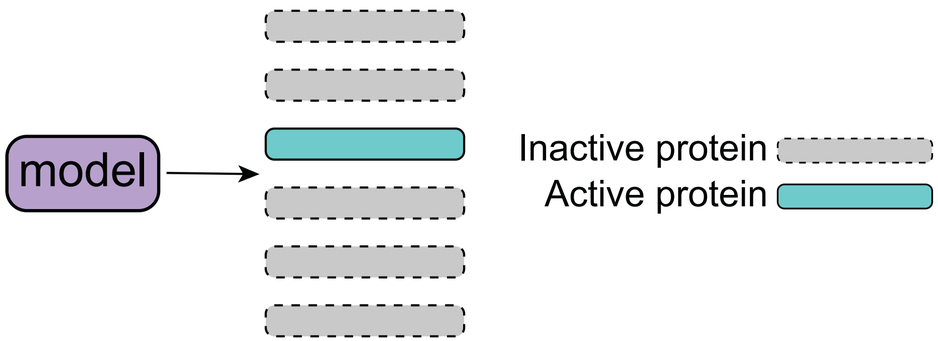
Most generated sequences are inactive



Model	Enzyme	Naive	Checks
GAN	MDH	0 / 18	2 / 18
	CuSOD	2 / 18	12 / 18
ESM-MSA	MDH	0 / 17	9 / 18
	CuSOD	0 / 18	12 / 18
Total		2 / 71	35 / 72

- Destabilized
- Mismatched signal peptide
- Transmembrane domain
- Long repeats
- No catalytic residue
- …

Most generated sequences are inactive

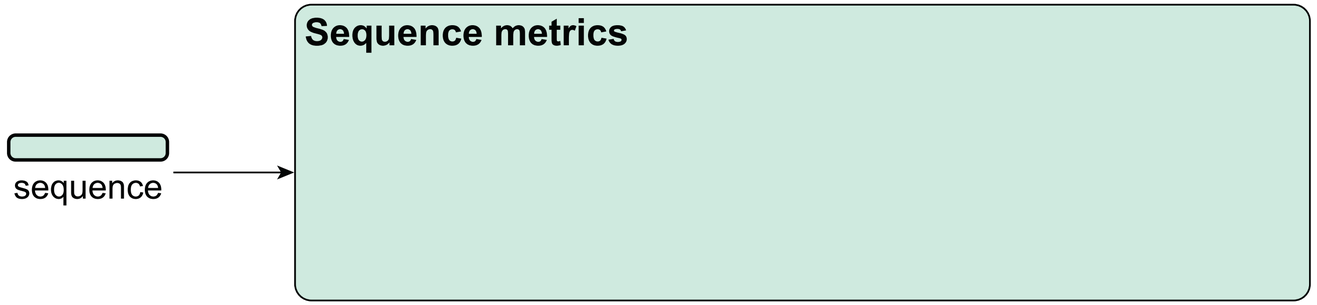


Model	Enzyme	Naive	Checks
GAN	MDH	0 / 18	2 / 18
	CuSOD	2 / 18	12 / 18
ESM-MSA	MDH	0 / 17	9 / 18
	CuSOD	0 / 18	12 / 18
Total		2 / 71	35 / 72

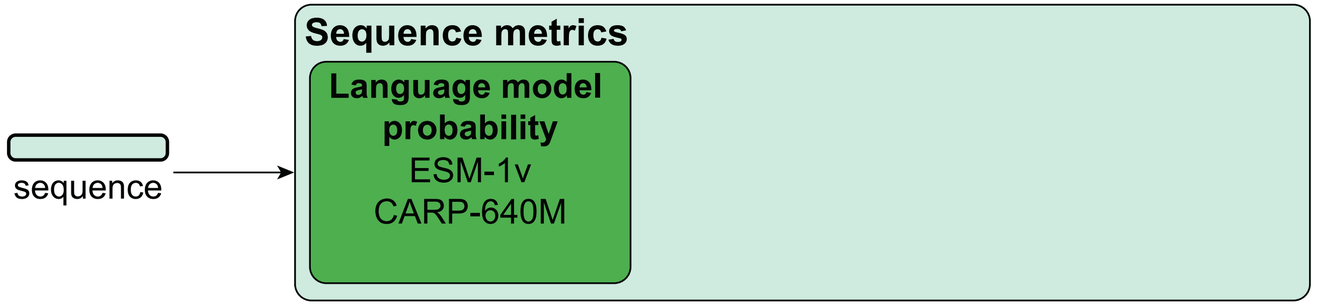
- Destabilized
- Mismatched signal peptide
- Transmembrane domain
- Long repeats
- No catalytic residue
- …

How can we do a better job selecting sequences?

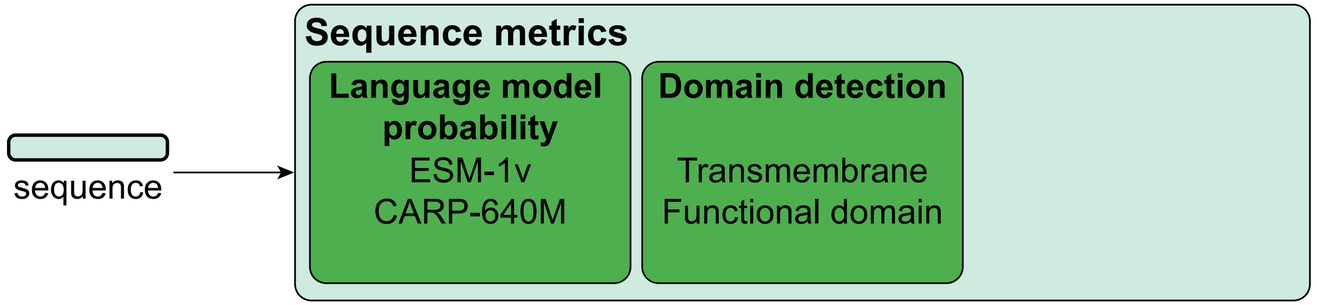
Many possible metrics for selection



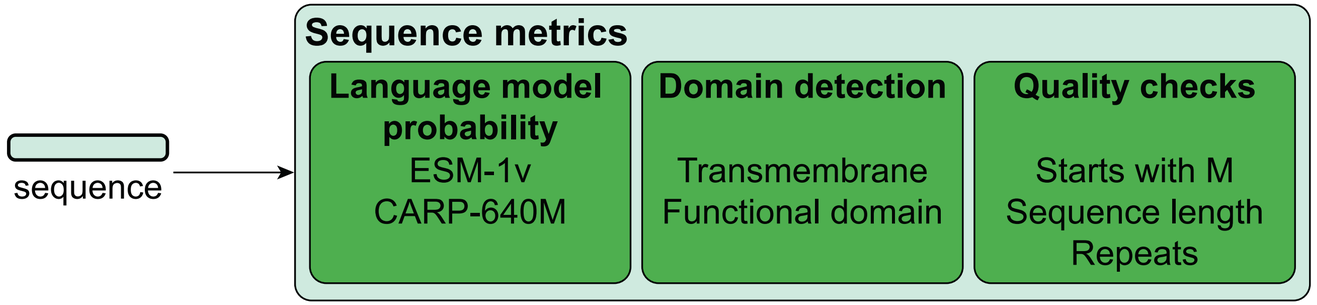
Many possible metrics for selection



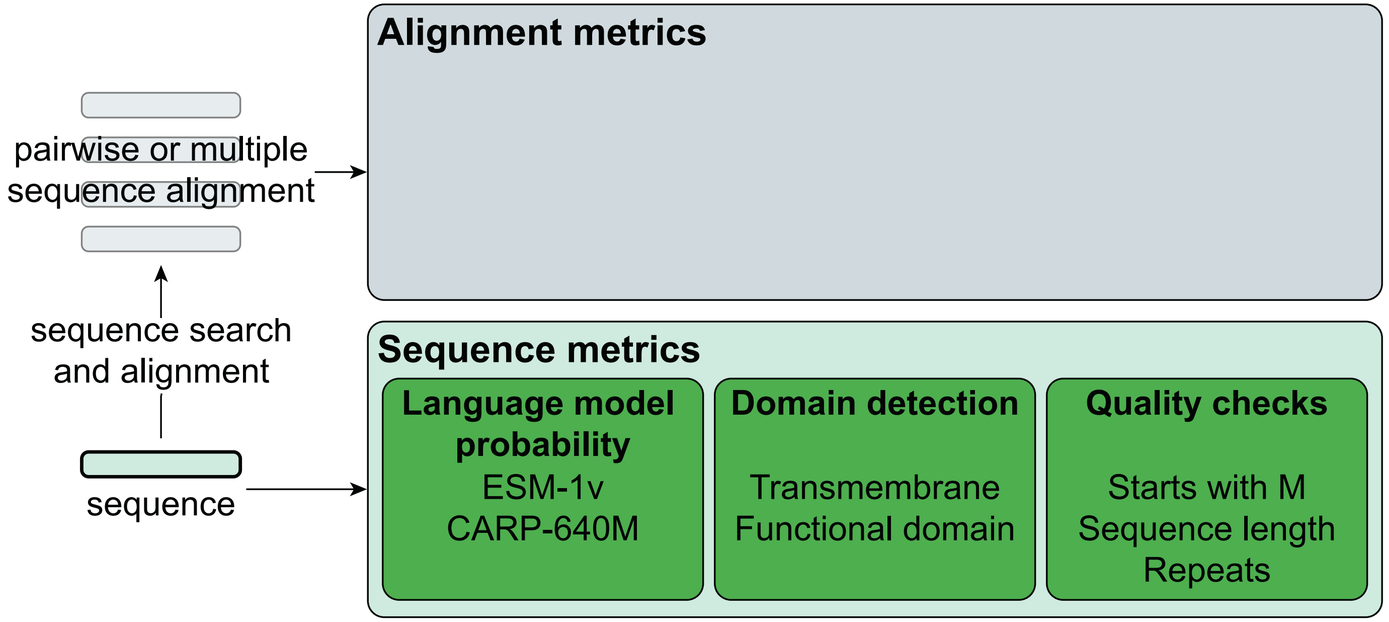
Many possible metrics for selection



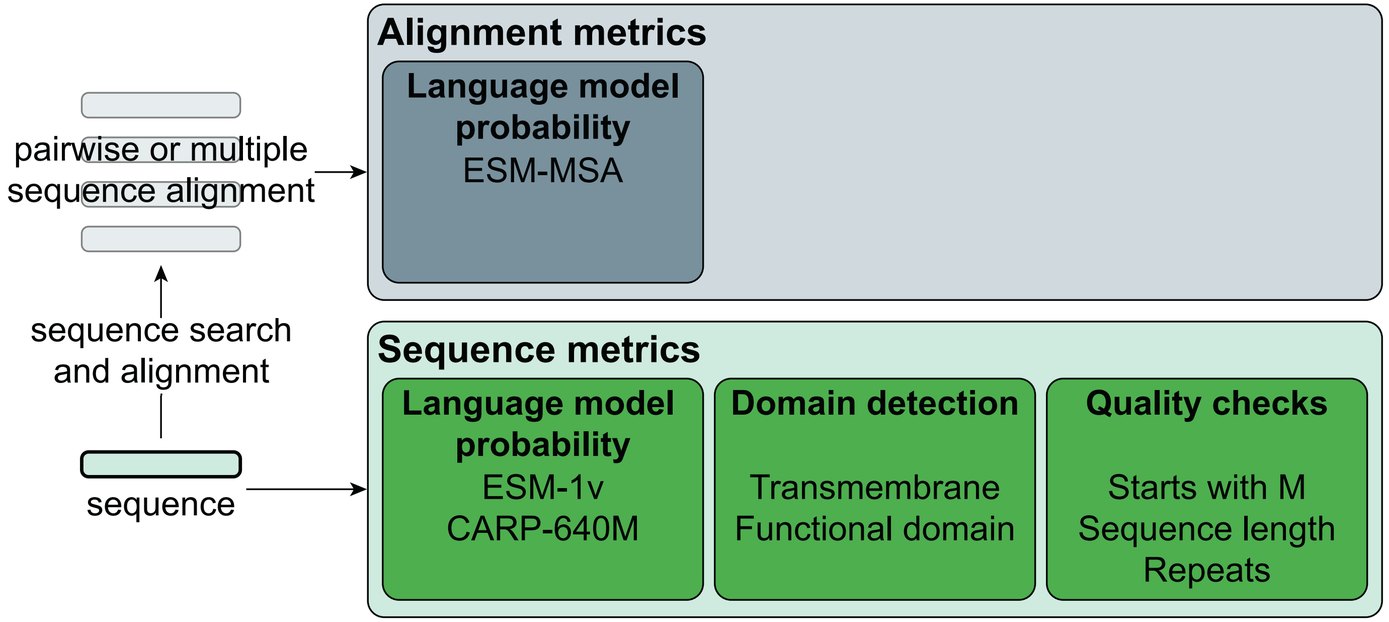
Many possible metrics for selection



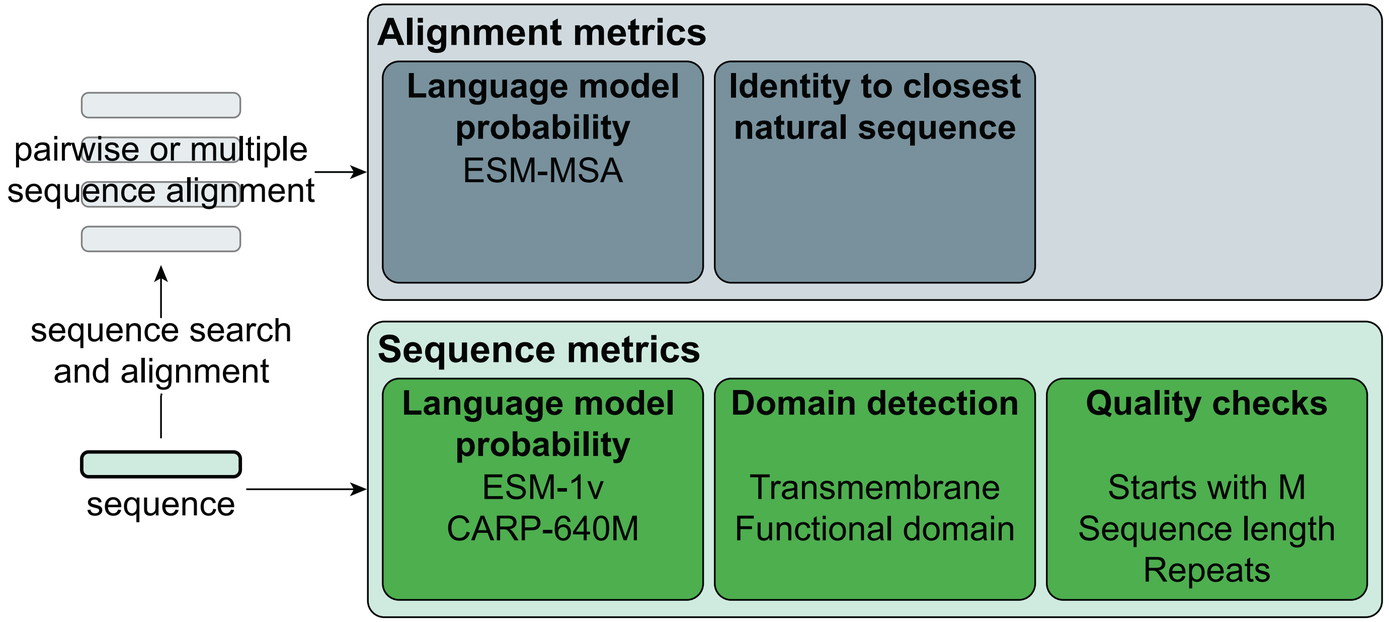
Many possible metrics for selection



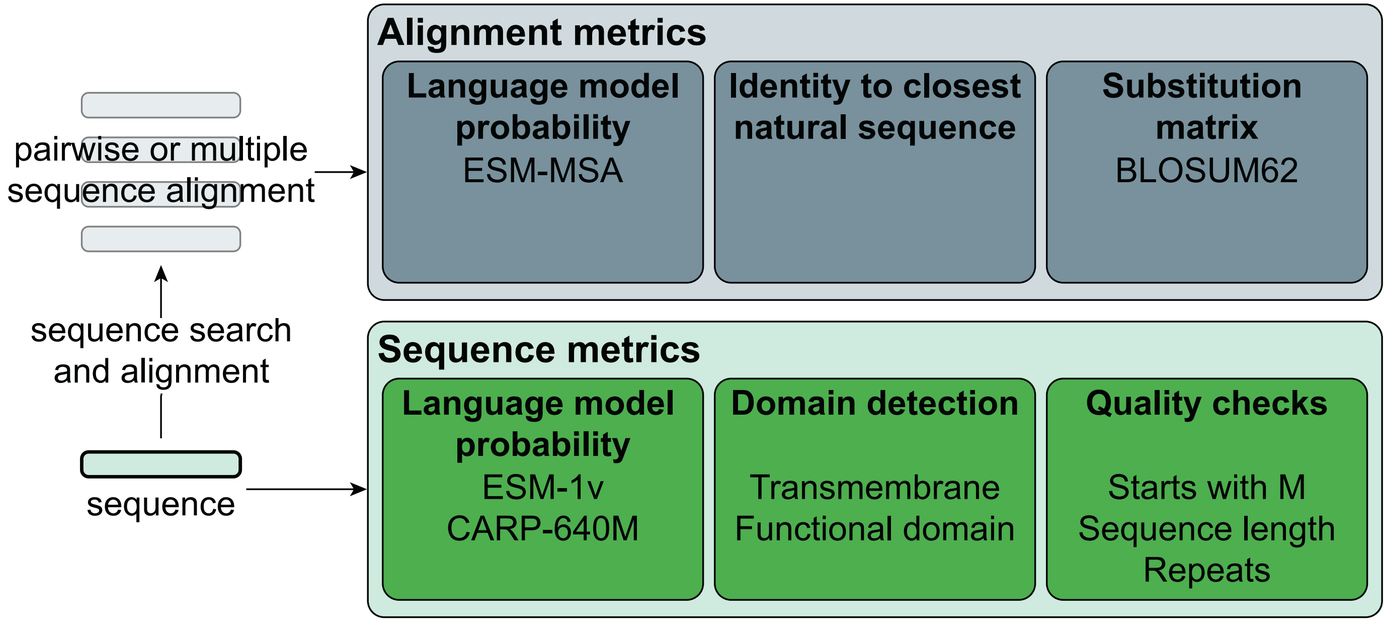
Many possible metrics for selection



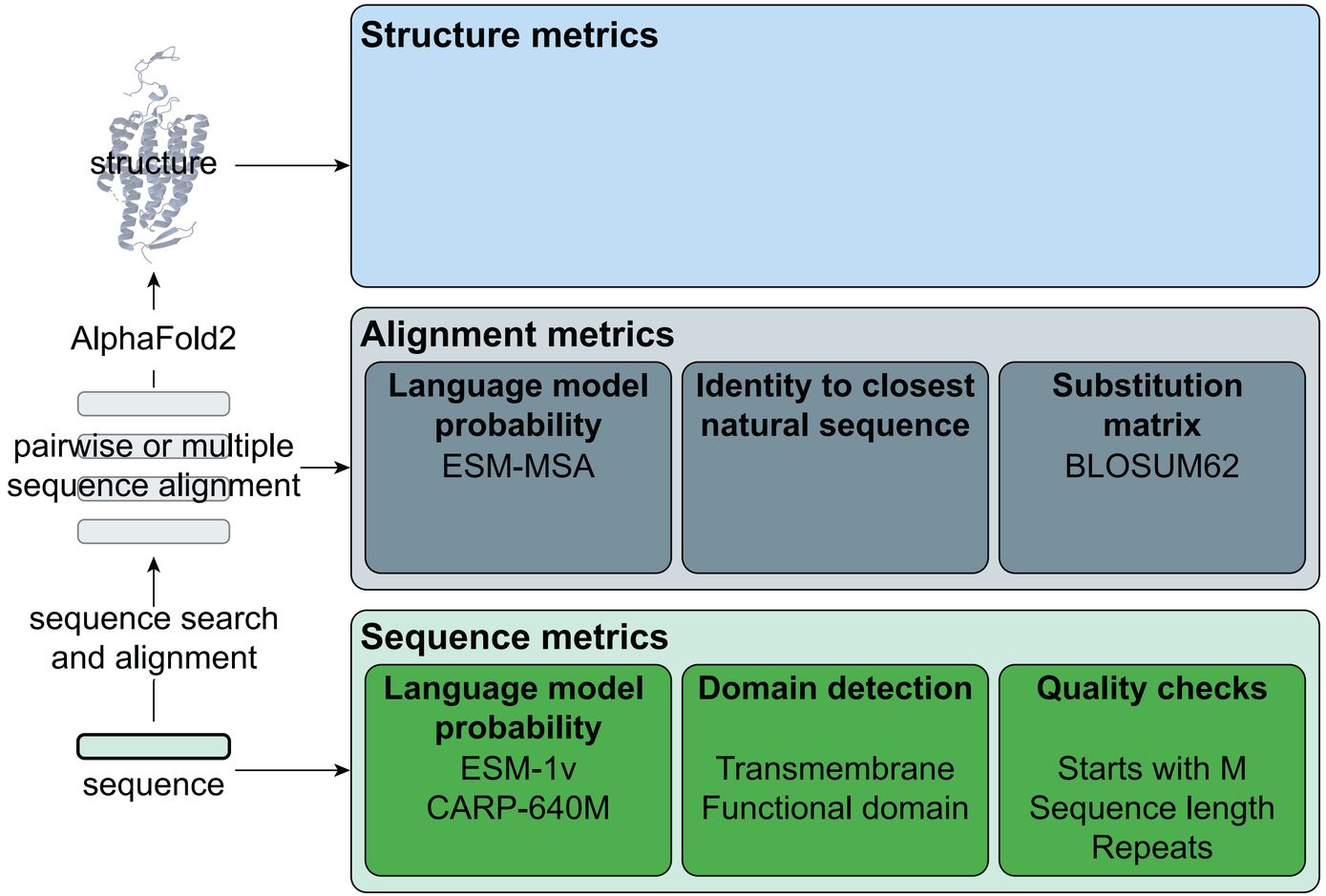
Many possible metrics for selection



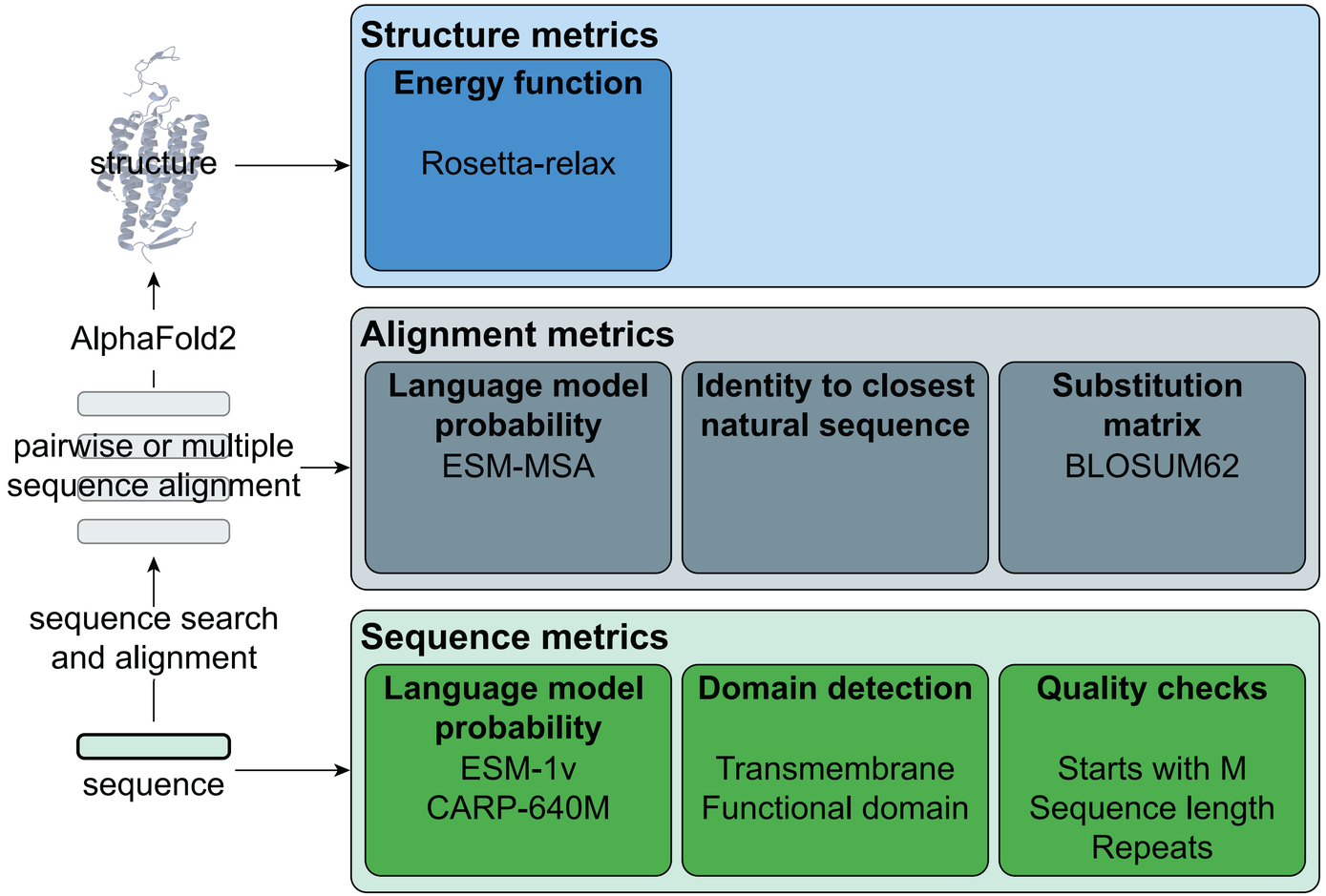
Many possible metrics for selection



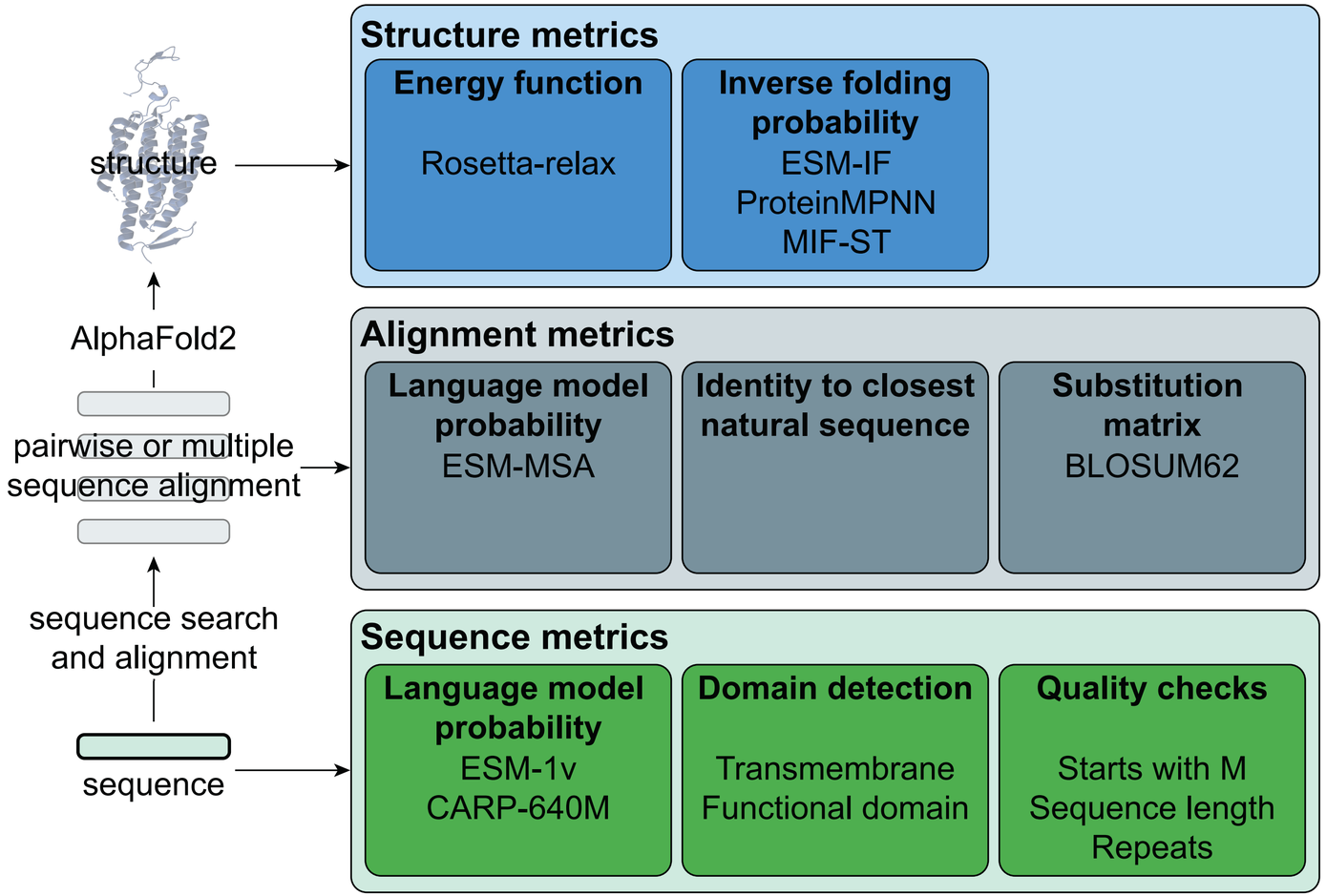
Many possible metrics for selection



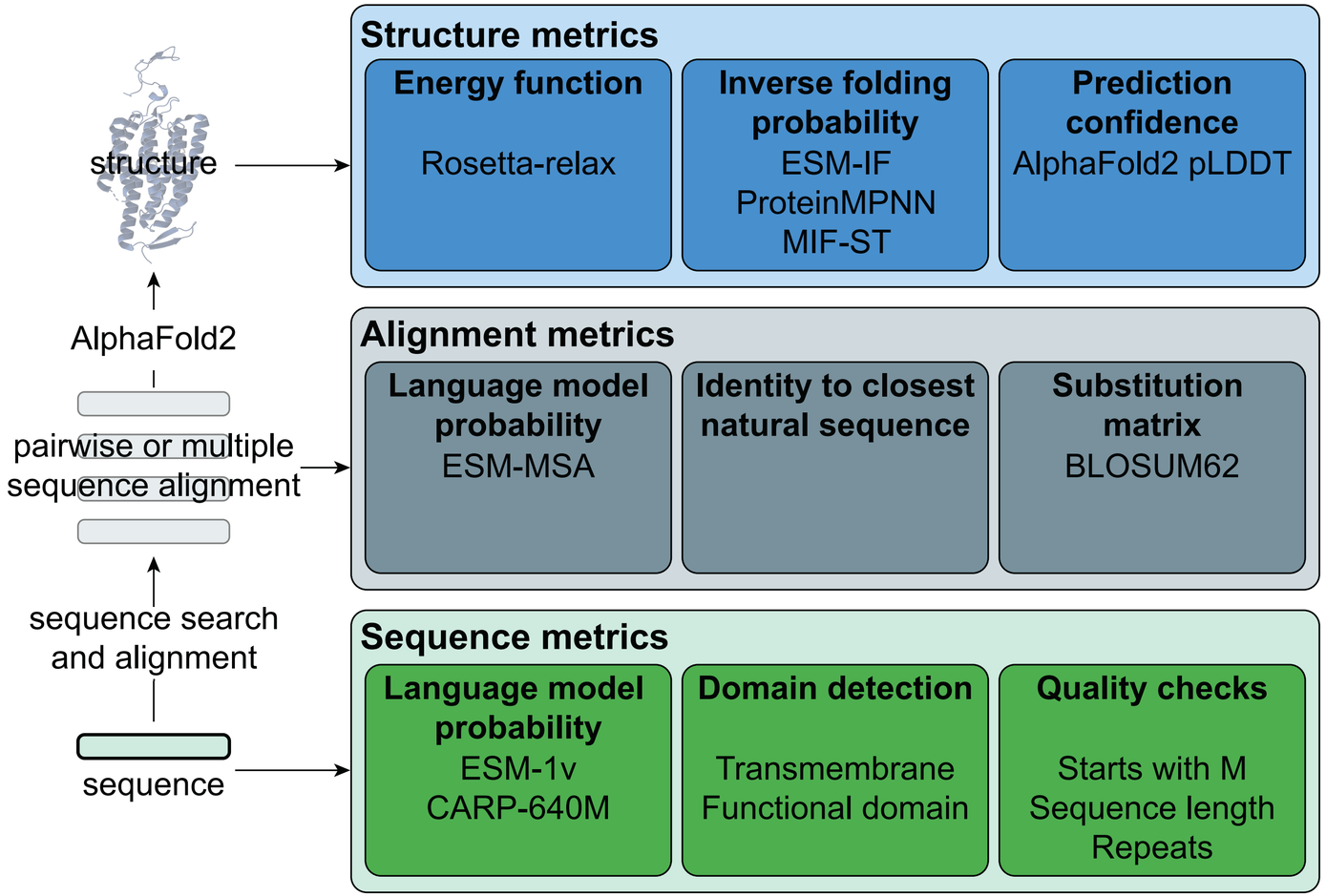
Many possible metrics for selection



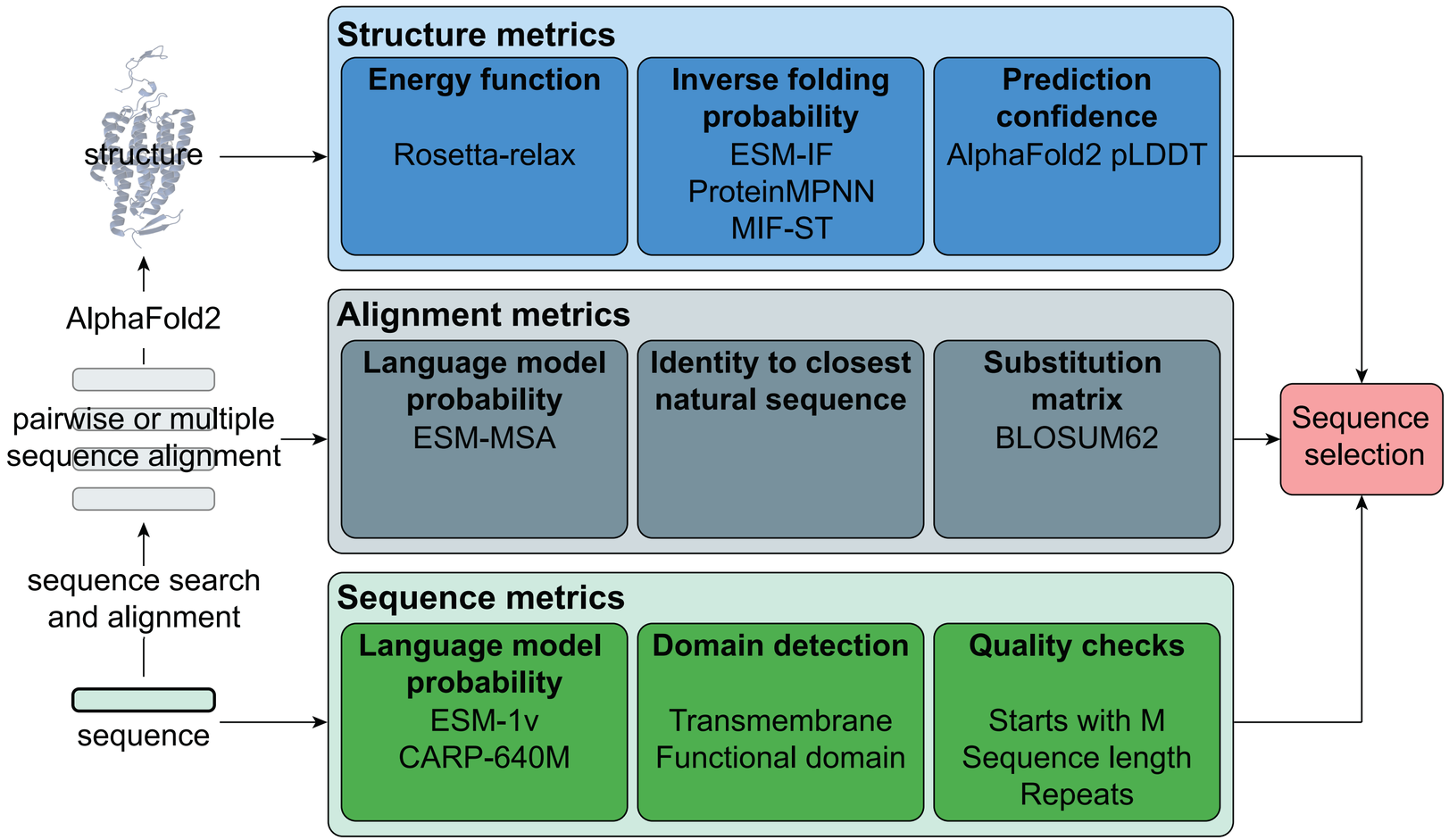
Many possible metrics for selection



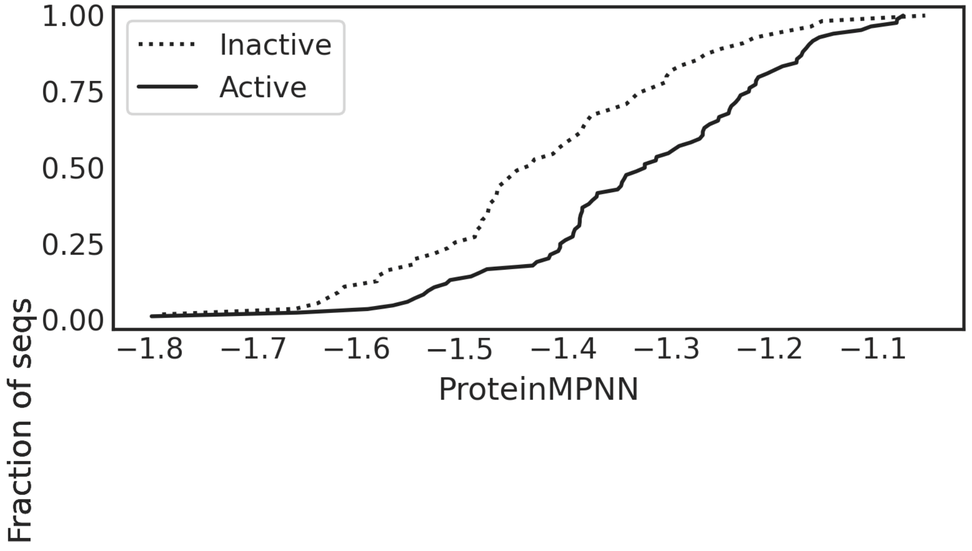
Many possible metrics for selection



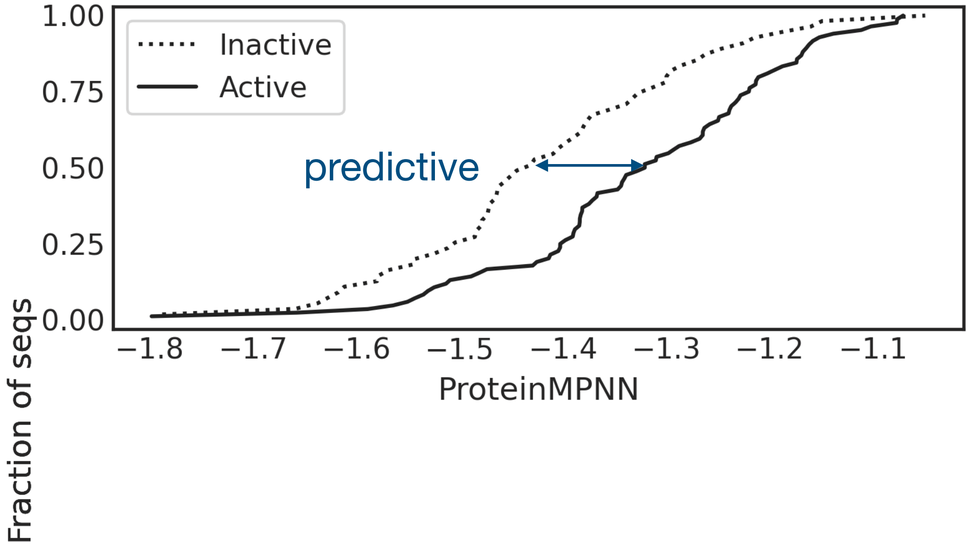
Many possible metrics for selection



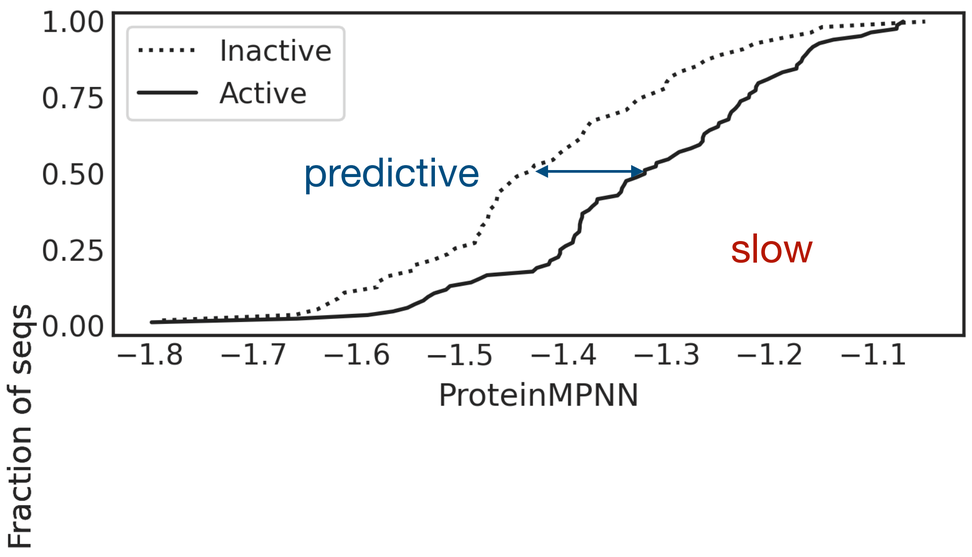
Choose metrics that are predictive, fast, and uncorrelated



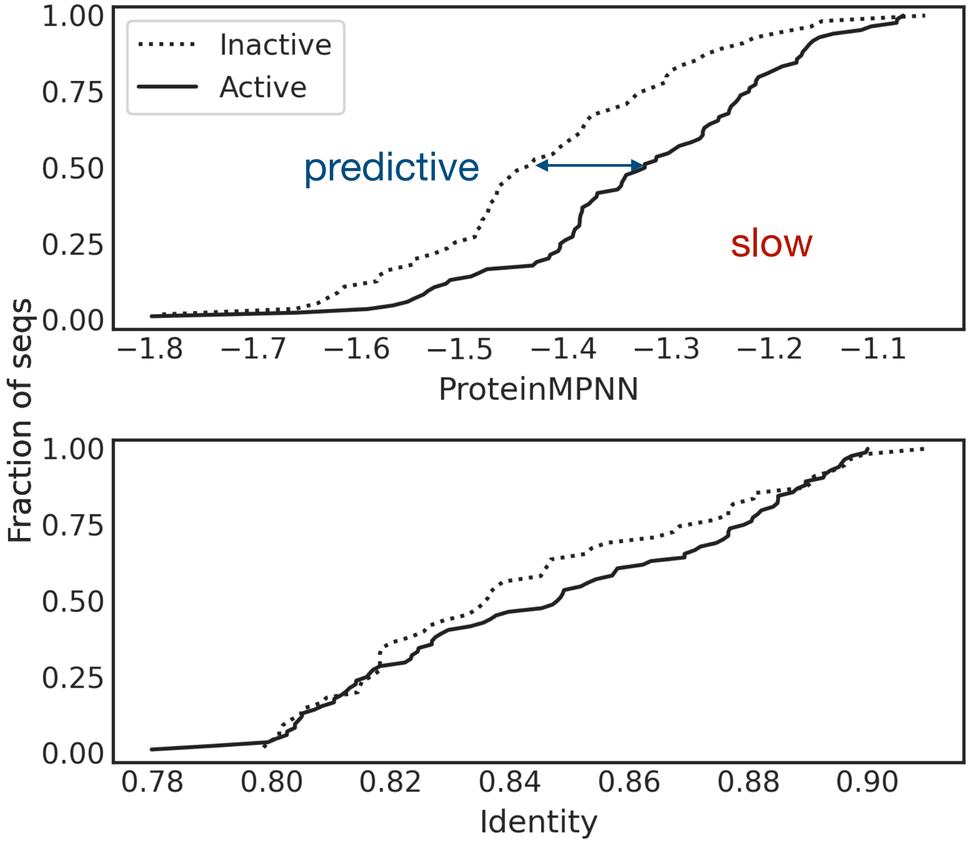
Choose metrics that are predictive, fast, and uncorrelated



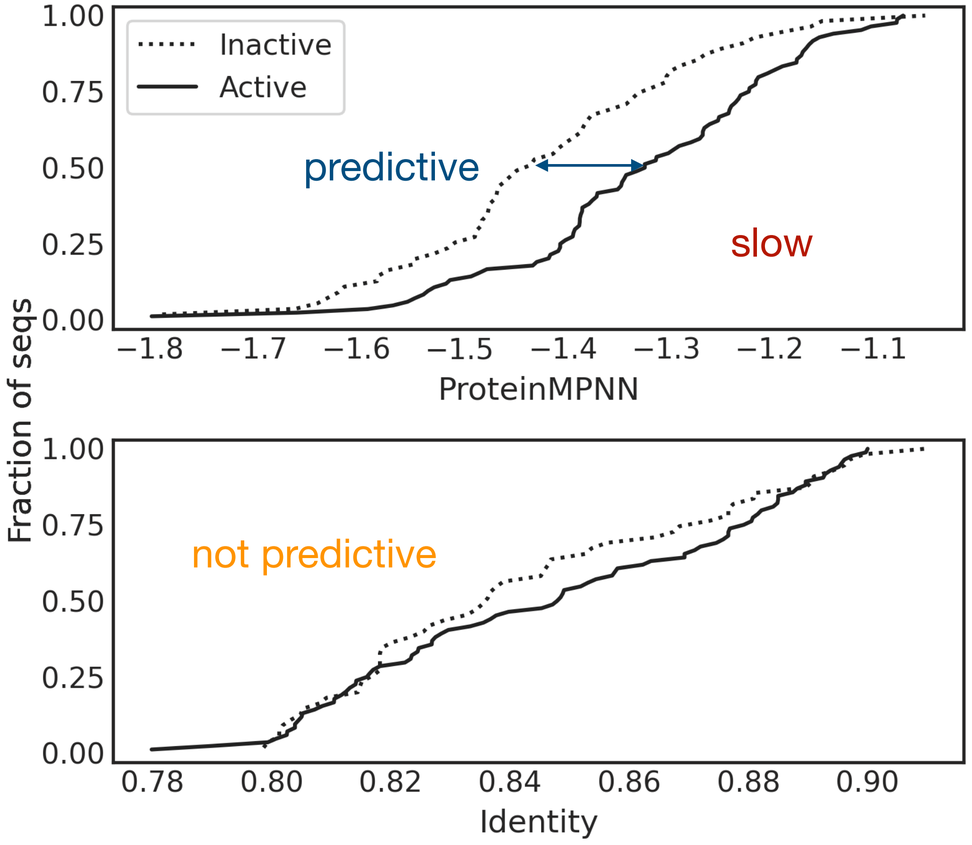
Choose metrics that are predictive, fast, and uncorrelated



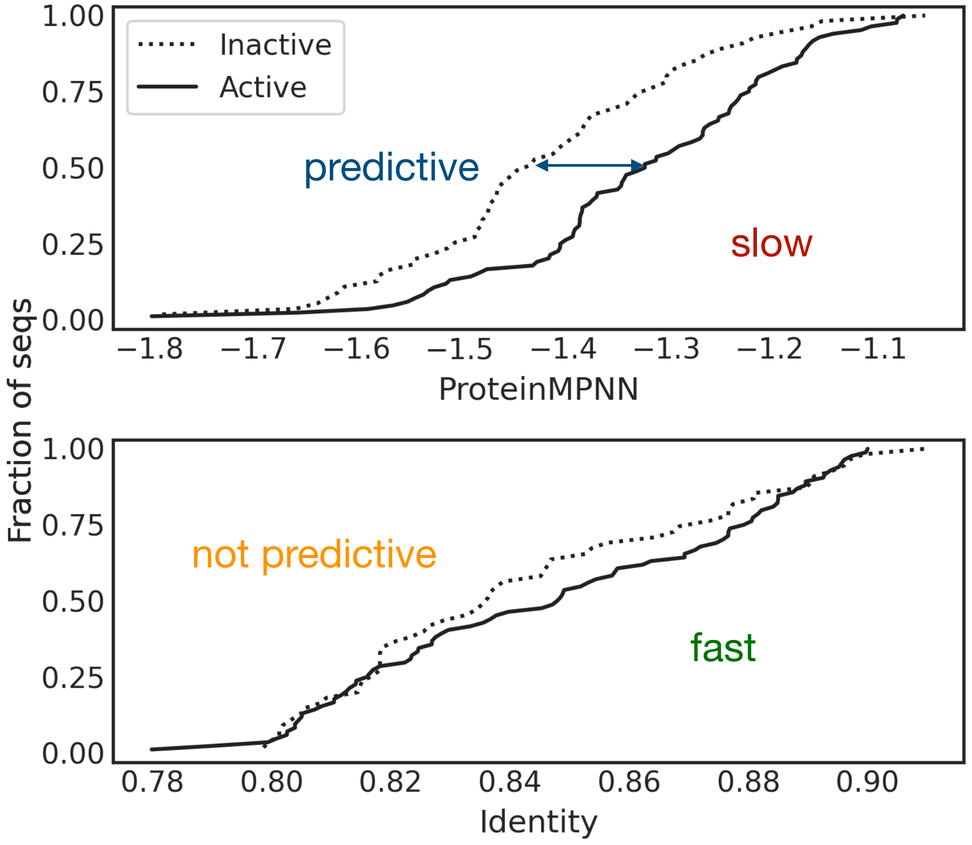
Choose metrics that are predictive, fast, and uncorrelated



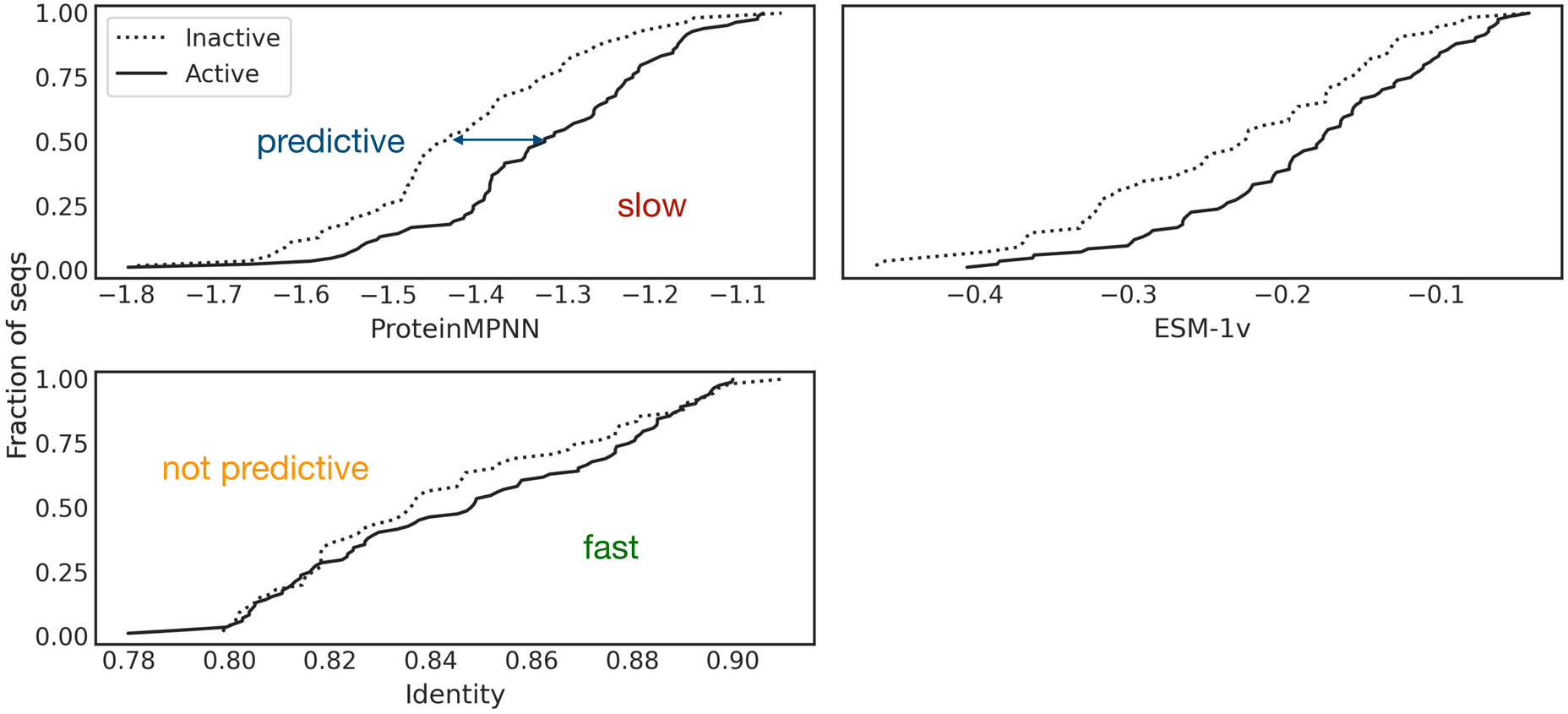
Choose metrics that are predictive, fast, and uncorrelated



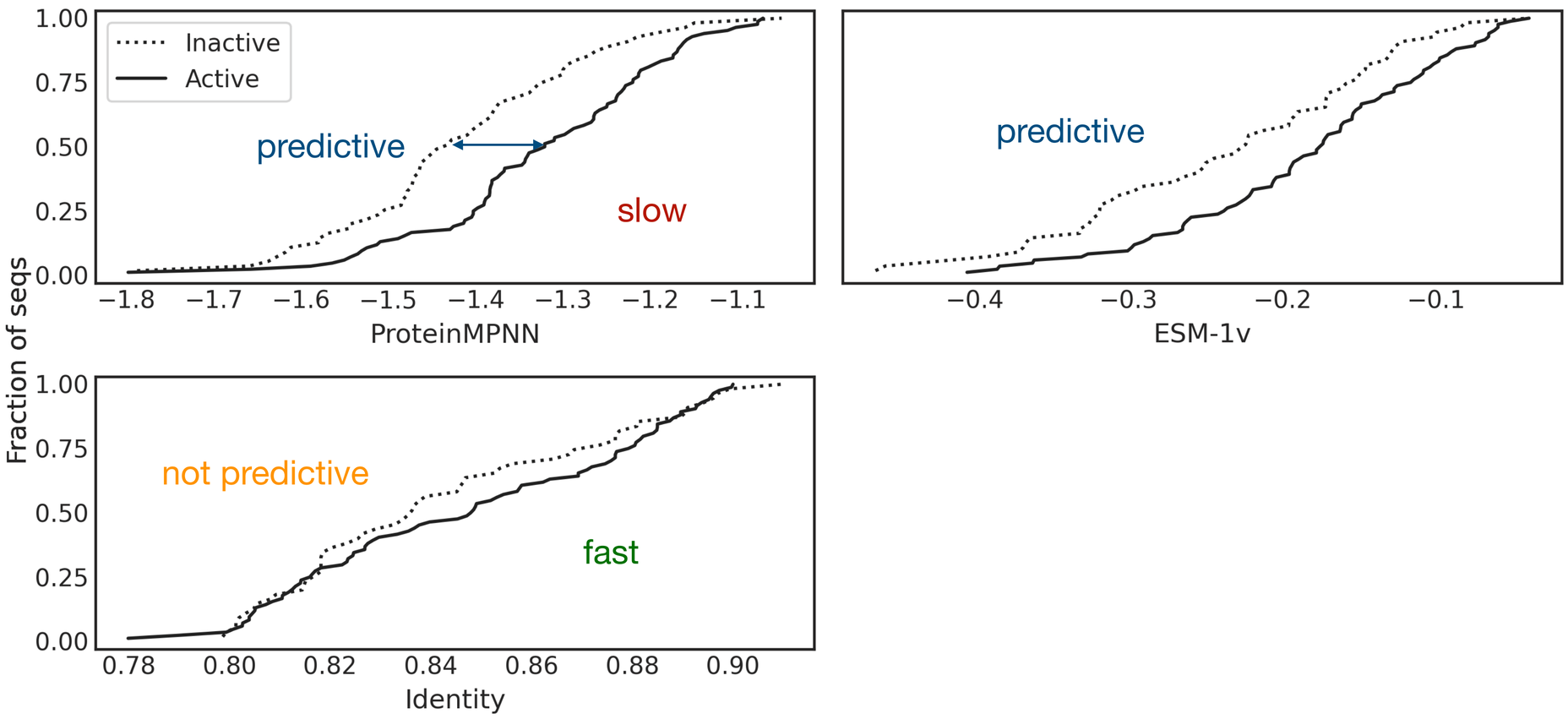
Choose metrics that are predictive, fast, and uncorrelated



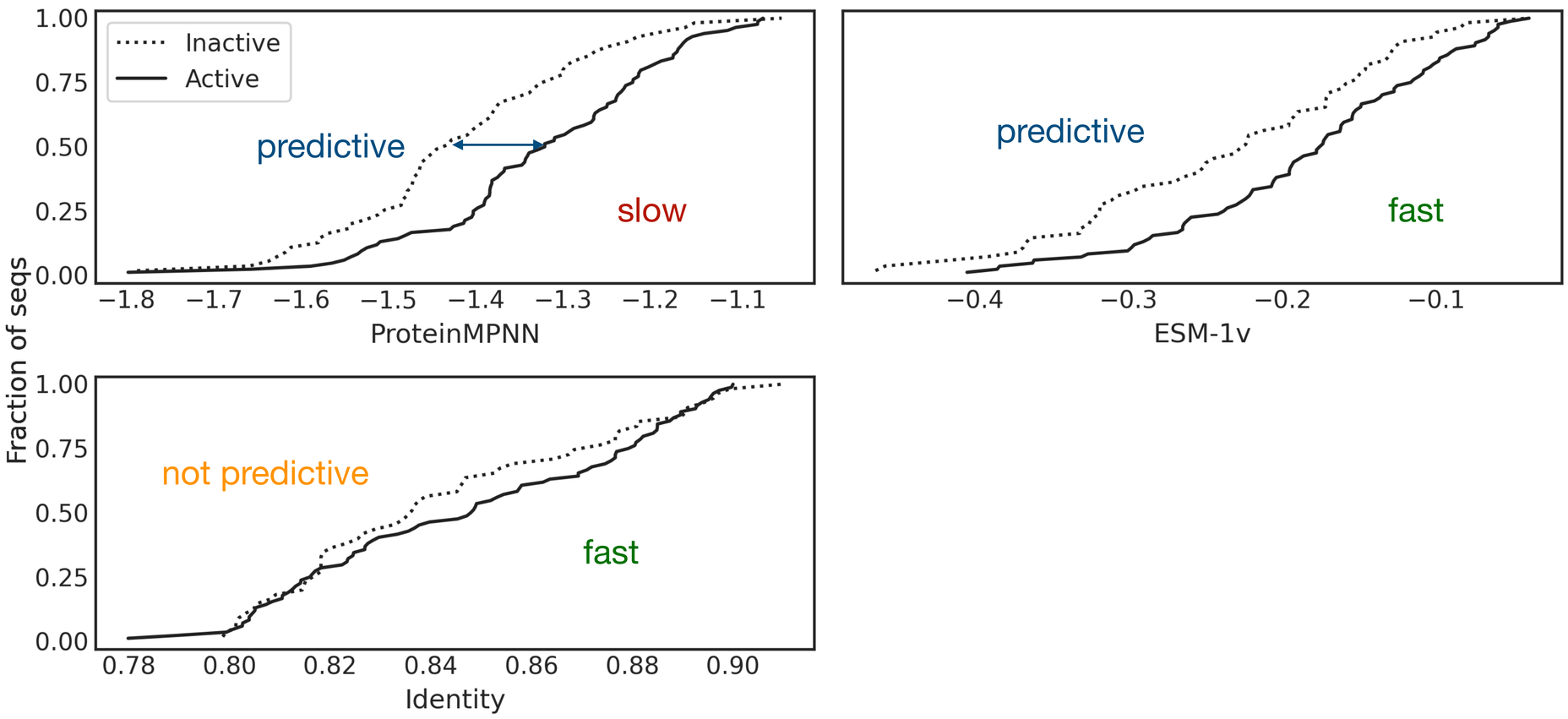
Choose metrics that are predictive, fast, and uncorrelated



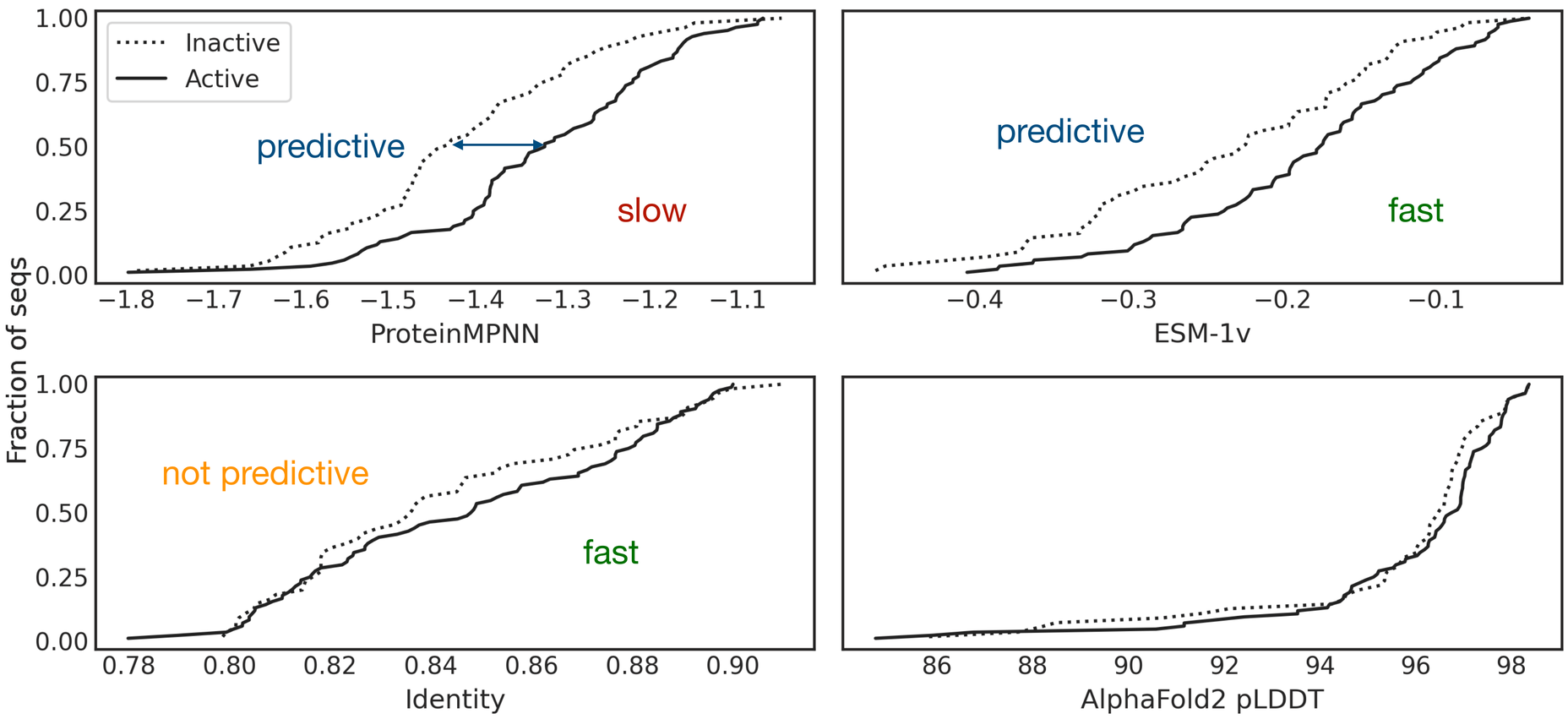
Choose metrics that are predictive, fast, and uncorrelated



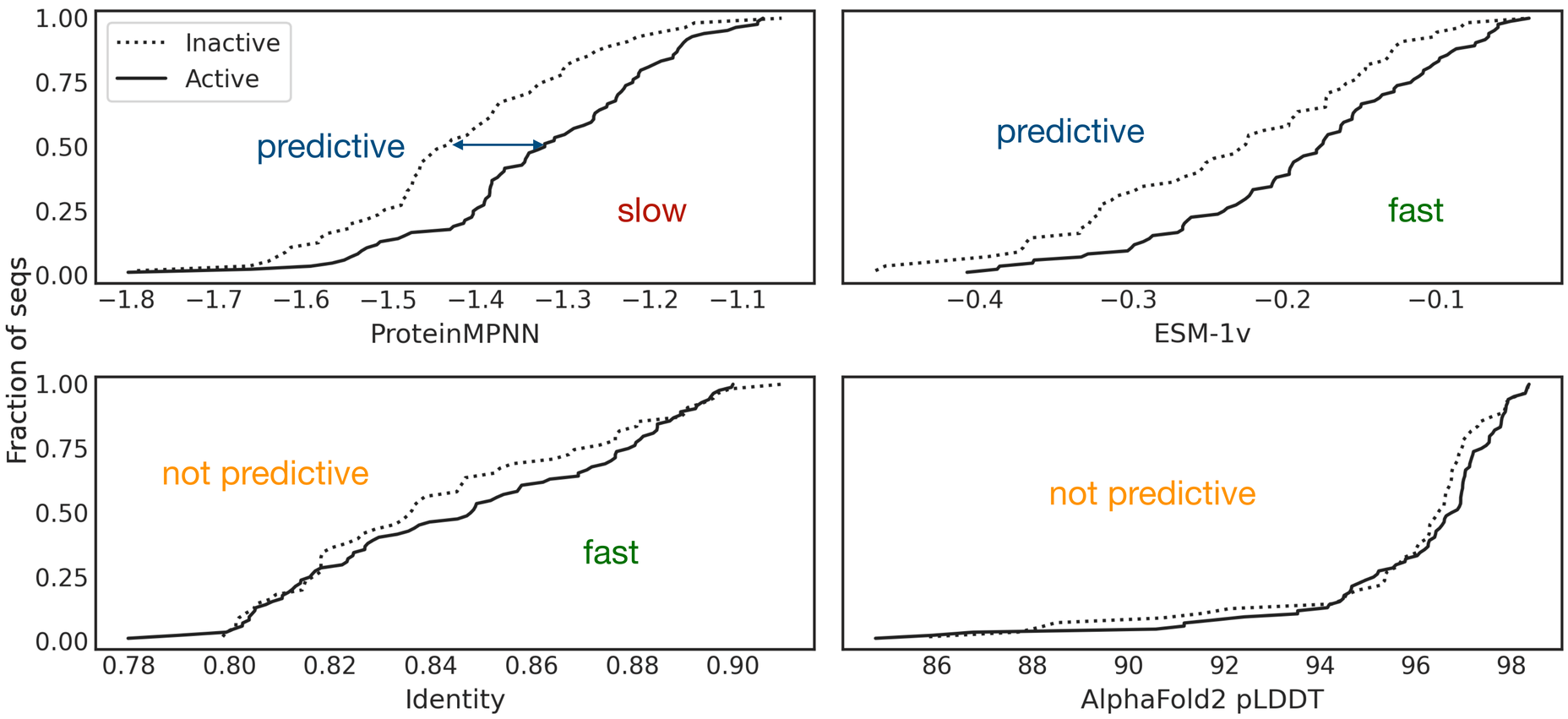
Choose metrics that are predictive, fast, and uncorrelated



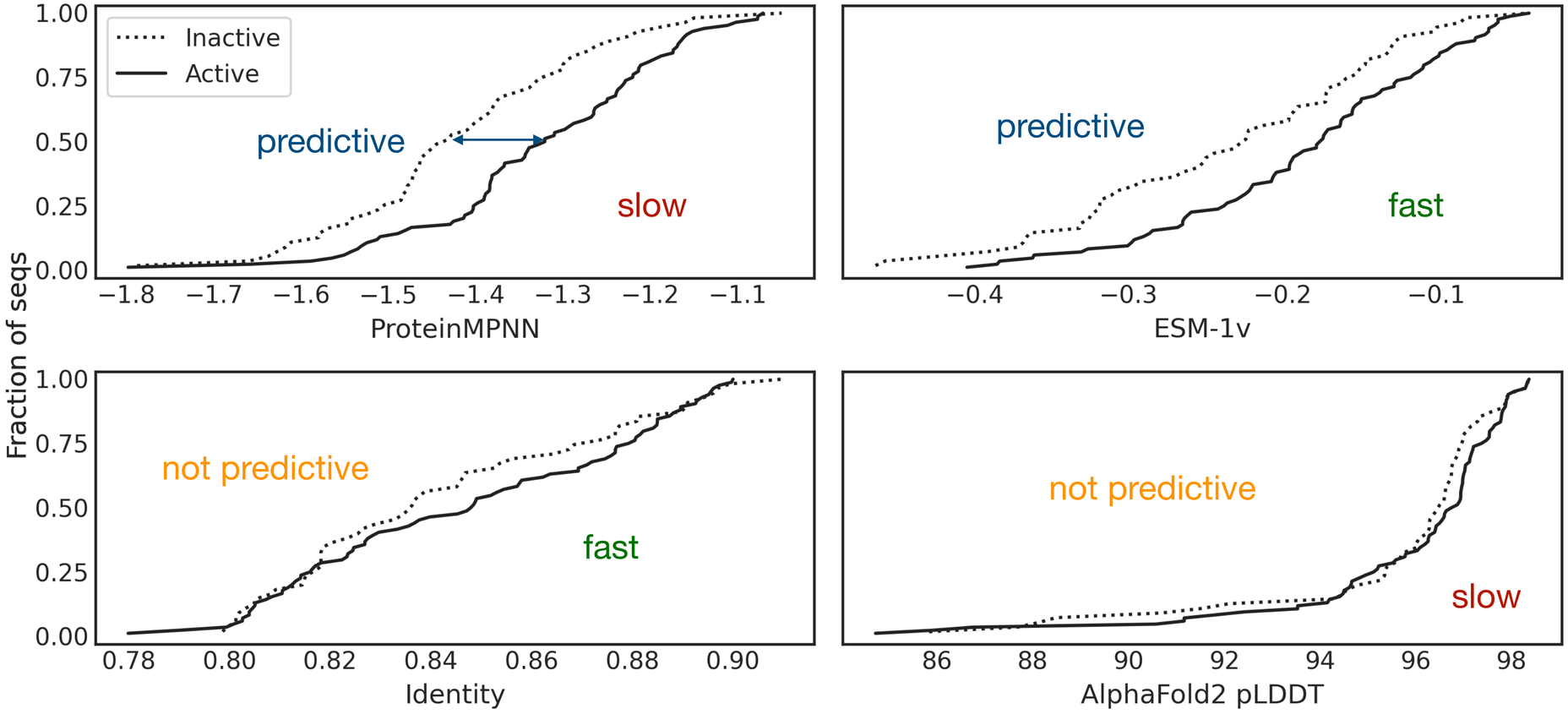
Choose metrics that are predictive, fast, and uncorrelated



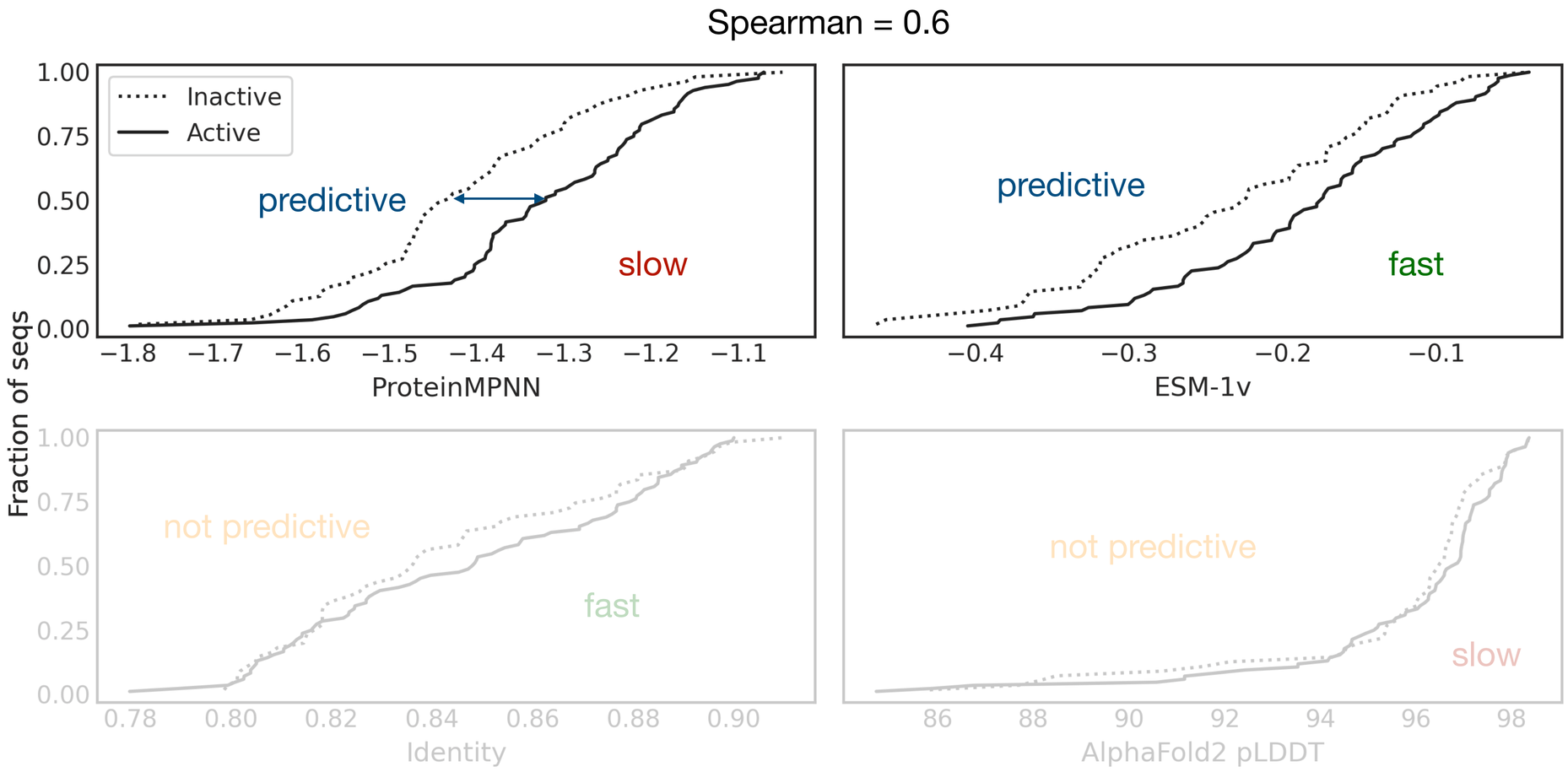
Choose metrics that are predictive, fast, and uncorrelated



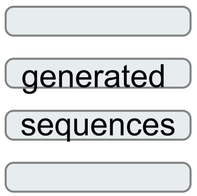
Choose metrics that are predictive, fast, and uncorrelated



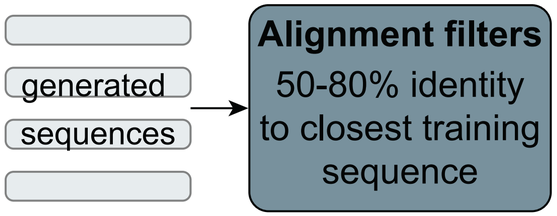
Choose metrics that are predictive, fast, and uncorrelated



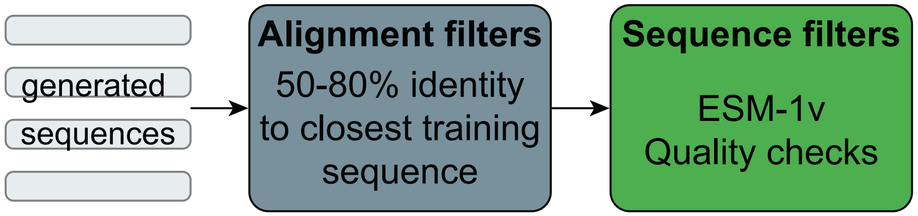
Combine many models into a filtering pipeline



Combine many models into a filtering pipeline

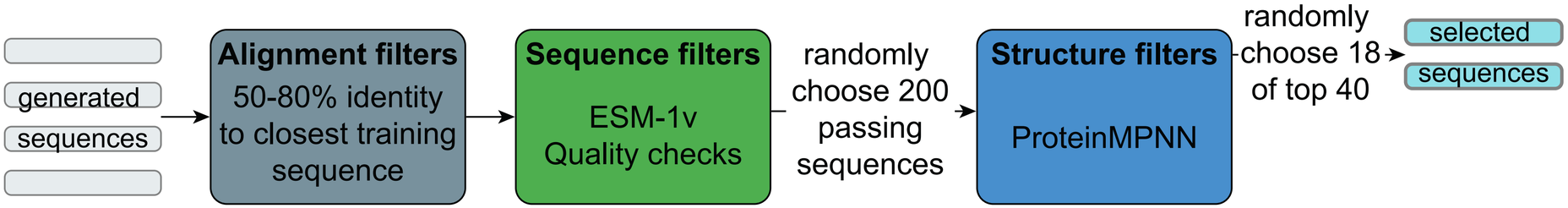


Combine many models into a filtering pipeline

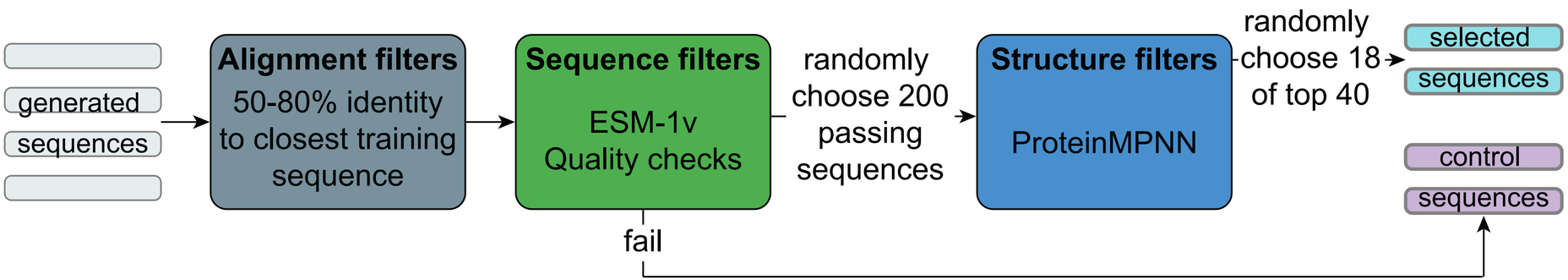


Combine many models into a filtering pipeline

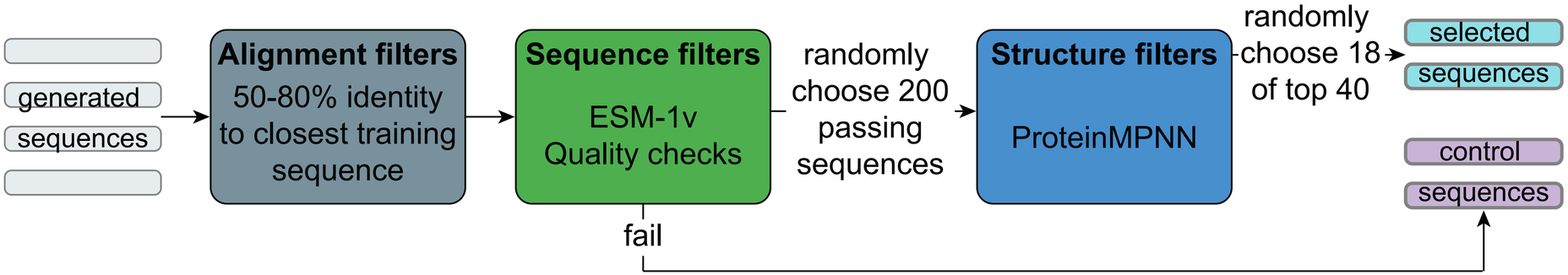
Combine many models into a filtering pipeline



Combine many models into a filtering pipeline

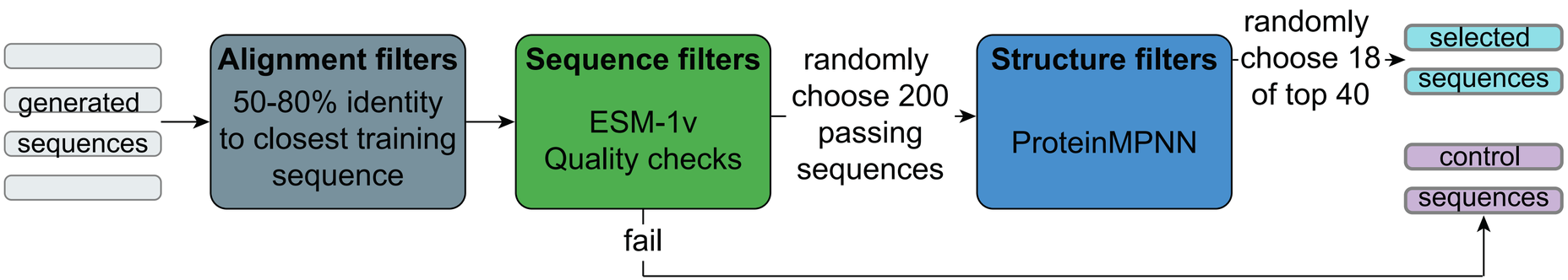


Pipeline increases number of active variants found



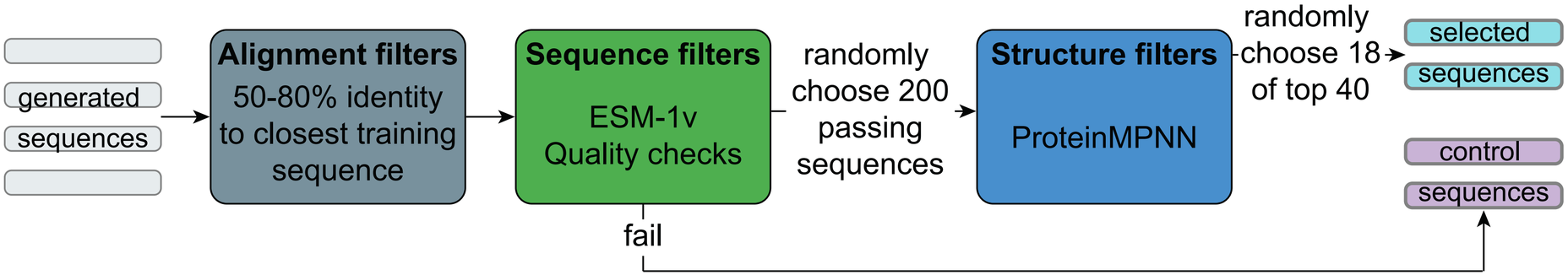
	Naive	Checks
Active	2 / 71	35 / 72

Pipeline increases number of active variants found

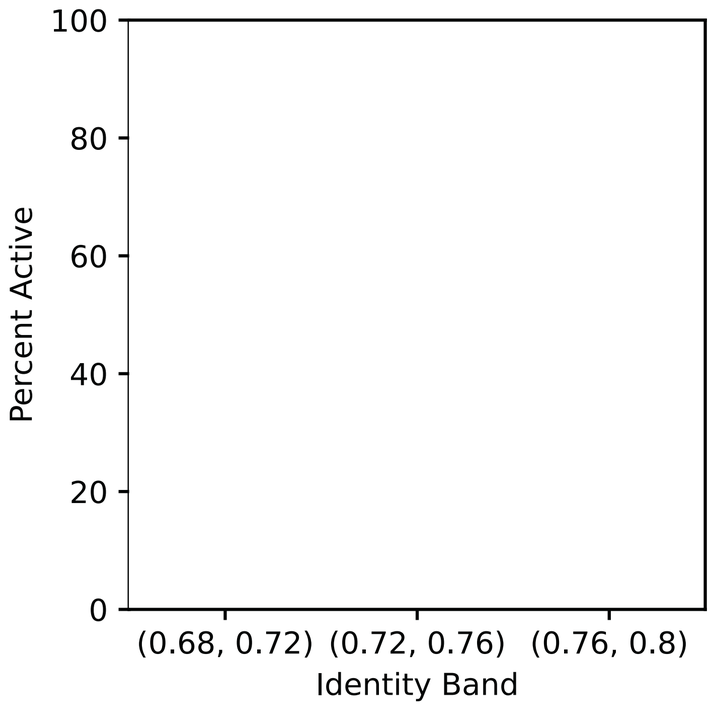


	Naive	Checks	Control
Active	2 / 71	35 / 72	32 / 72

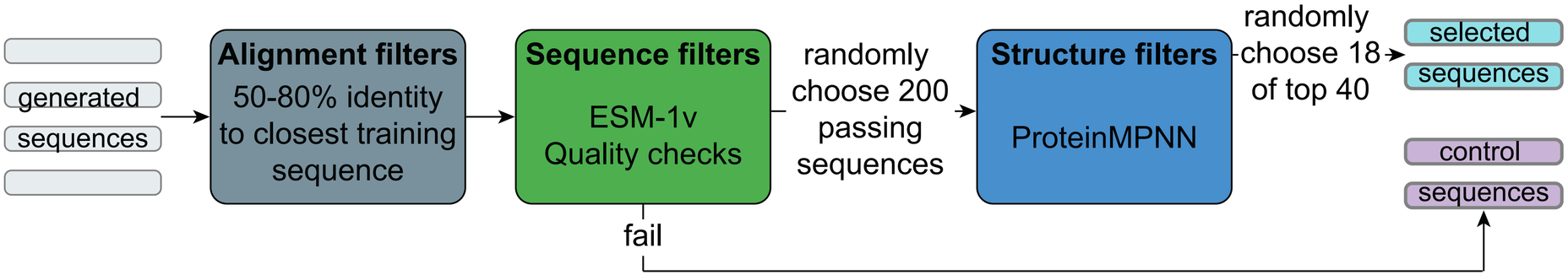
Pipeline increases number of active variants found



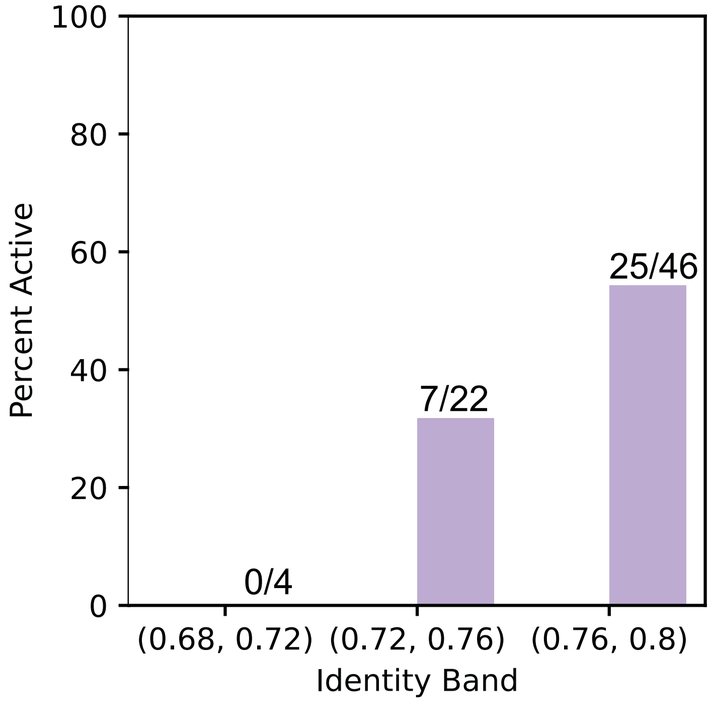
	Naive	Checks	Control
Active	2 / 71	35 / 72	32 / 72



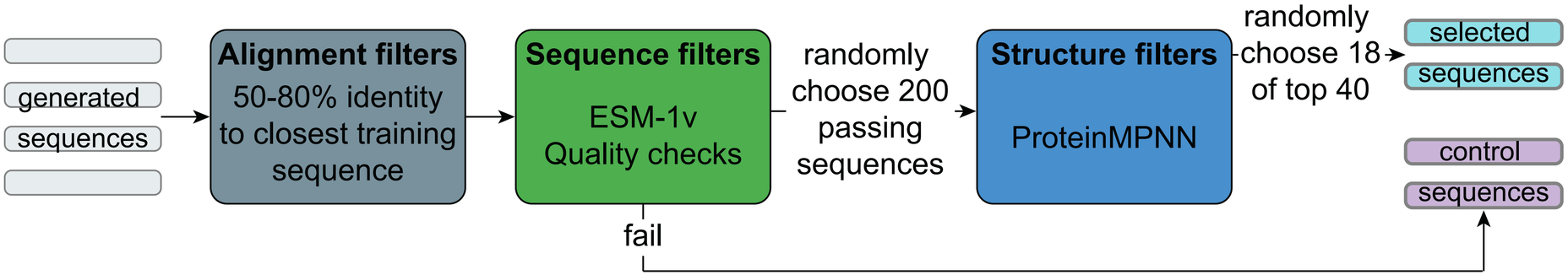
Pipeline increases number of active variants found



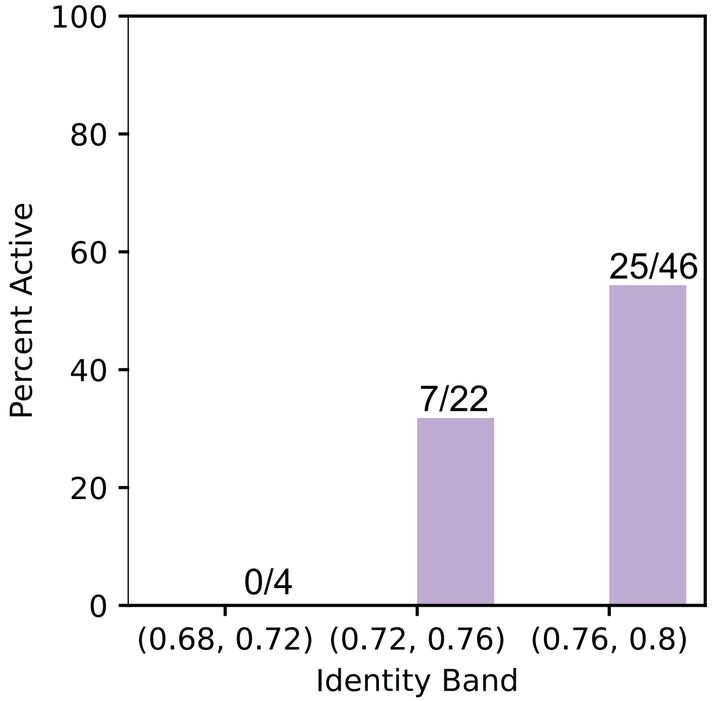
	Naive	Checks	Control
Active	2 / 71	35 / 72	32 / 72



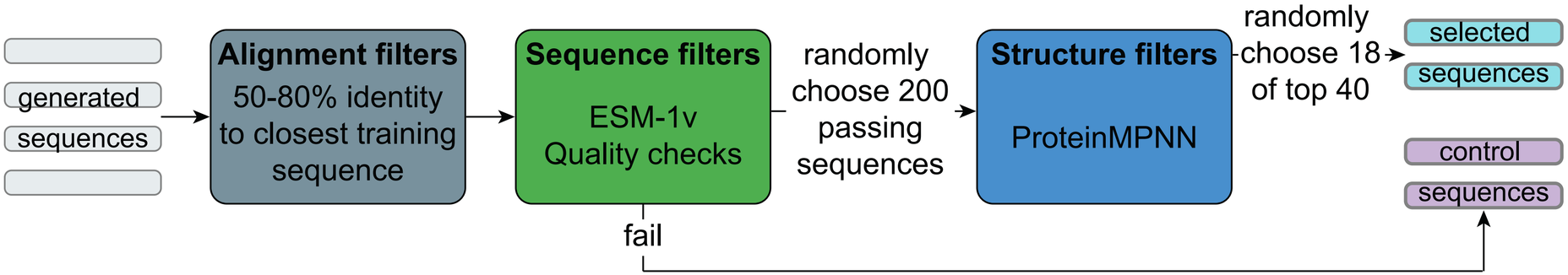
Pipeline increases number of active variants found



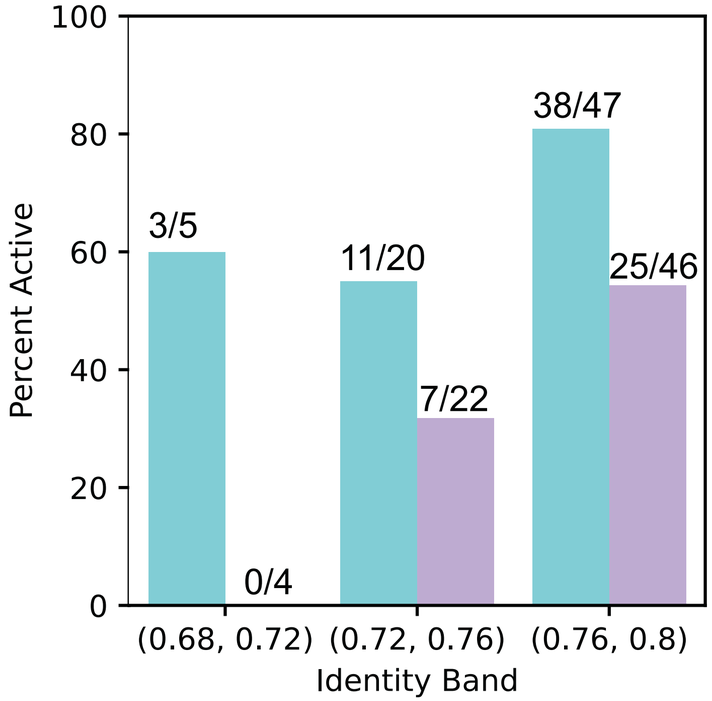
	Naive	Checks	Control	Selected
Active	2 / 71	35 / 72	32 / 72	52 / 72



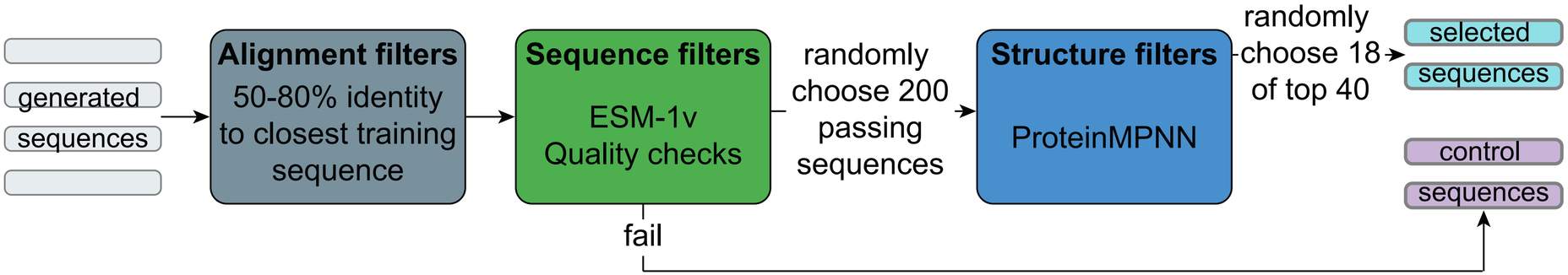
Pipeline increases number of active variants found



	Naive	Checks	Control	Selected
Active	2 / 71	35 / 72	32 / 72	52 / 72

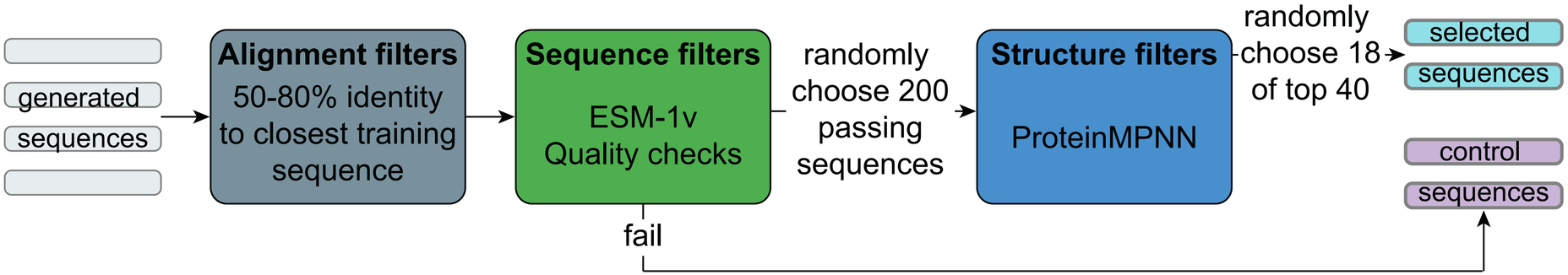


Pipeline increases number of active variants found



Model	Enzyme	Naive	Checks	Control	Selected
GAN	MDH	0 / 18	2 / 18		
	CuSOD	2 / 18	12 / 18		
ESM-MSA	MDH	0 / 17	9 / 18		
	CuSOD	0 / 18	12 / 18		
Total		2 / 71	35 / 72	32 / 72	52 / 72

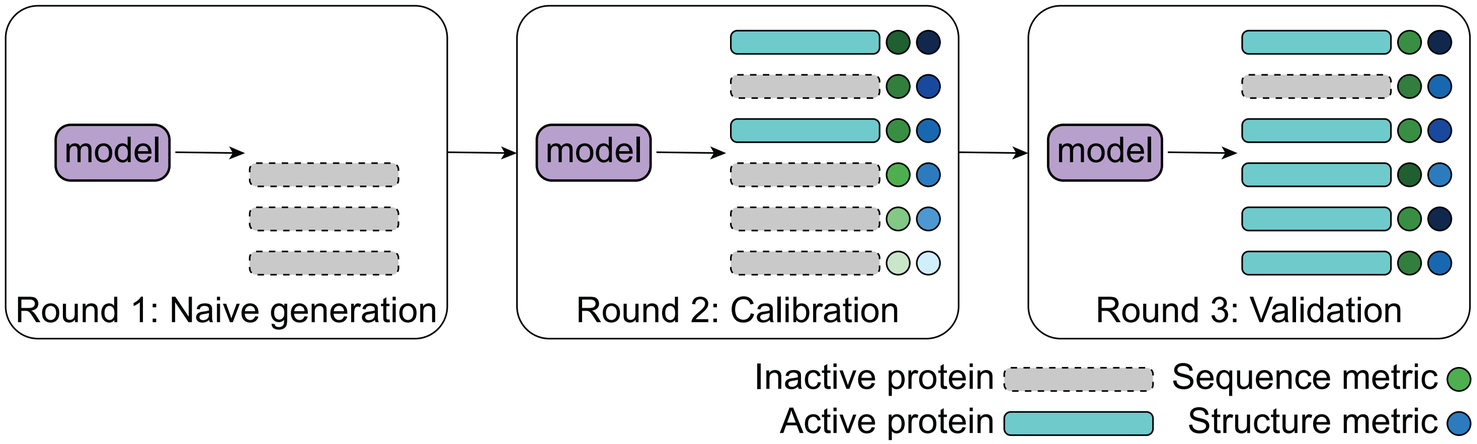
Pipeline increases number of active variants found



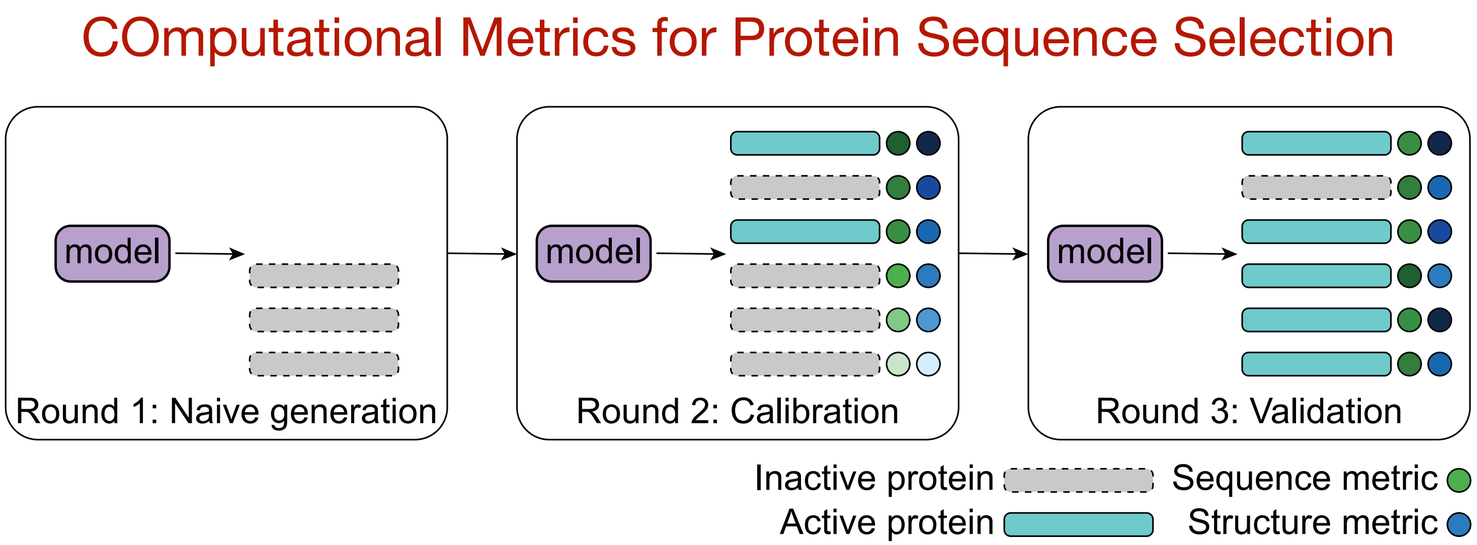
Model	Enzyme	Naive	Checks	Control	Selected
GAN	MDH	0 / 18	2 / 18	2 / 18	4 / 18
	CuSOD	2 / 18	12 / 18	9 / 18	13 / 18
ESM-MSA	MDH	0 / 17	9 / 18	12 / 18	18 / 18
	CuSOD	0 / 18	12 / 18	9 / 18	17 / 18
Total		2 / 71	35 / 72	32 / 72	52 / 72

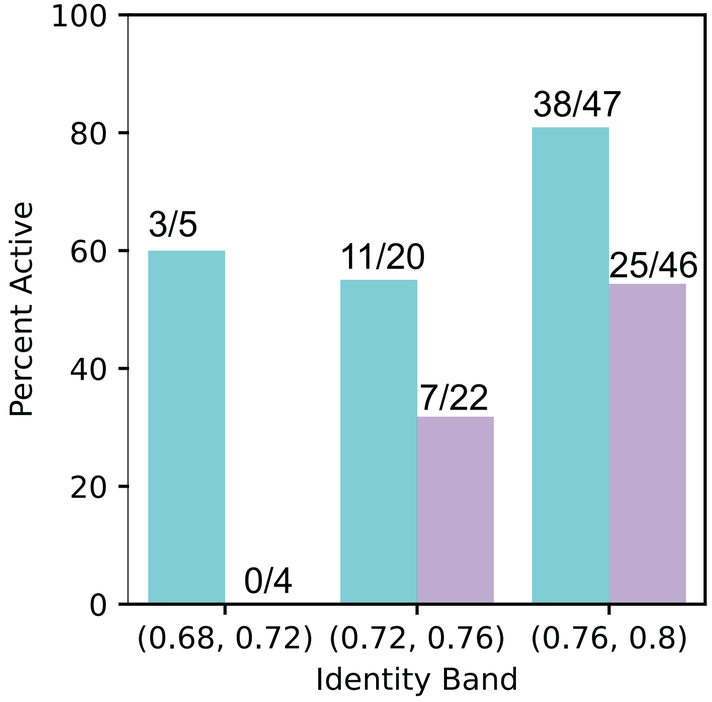
COMPSS is a tool for selecting generated proteins

COmputational Metrics for Protein Sequence Selection

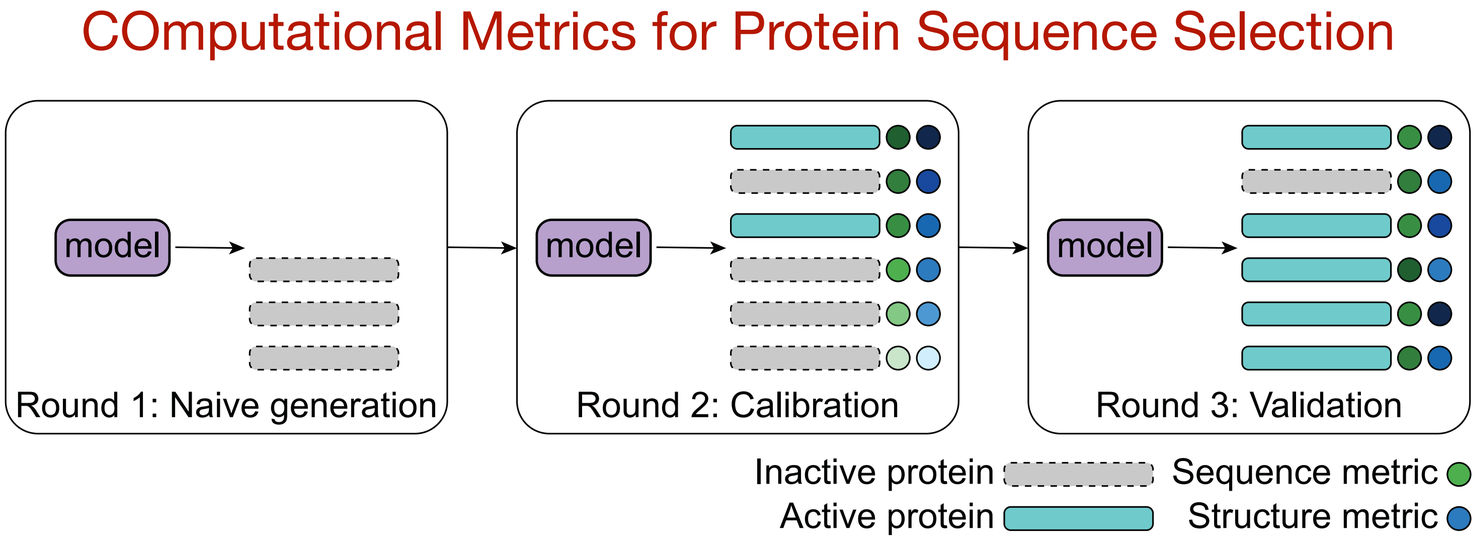


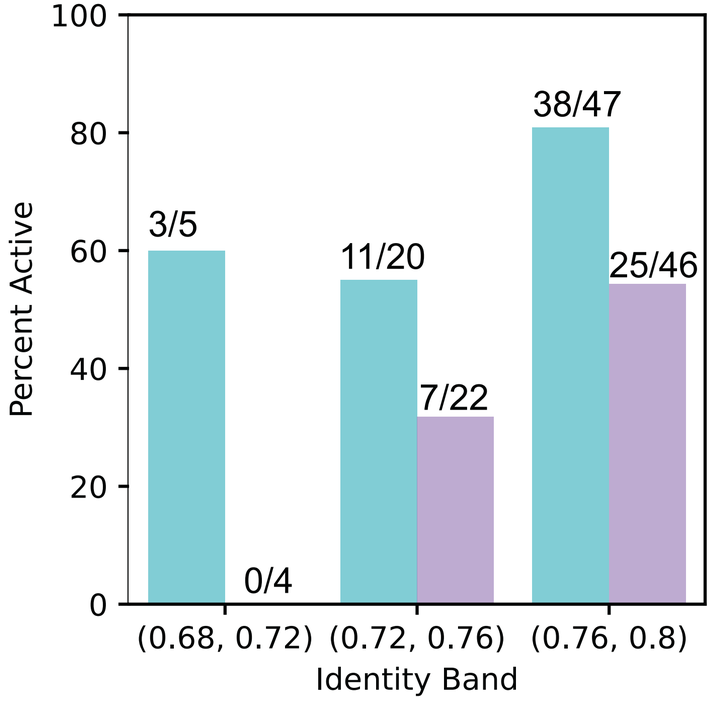
COMPSS is a tool for selecting generated proteins





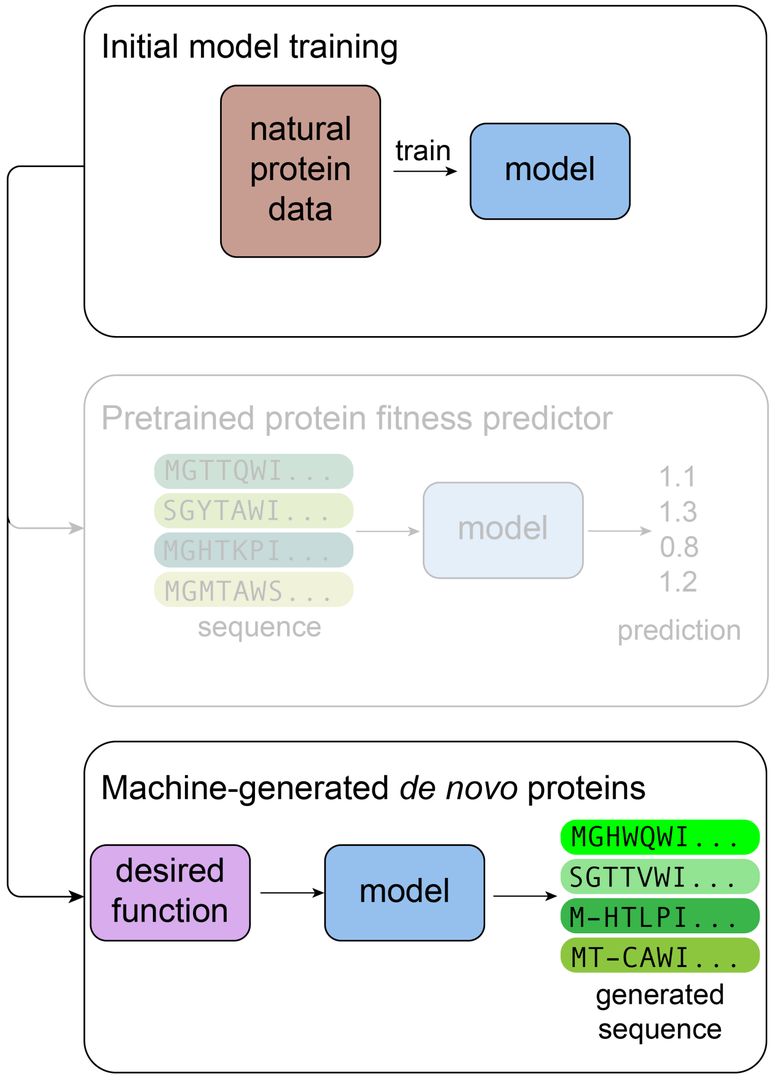
COMPSS is a tool for selecting generated proteins

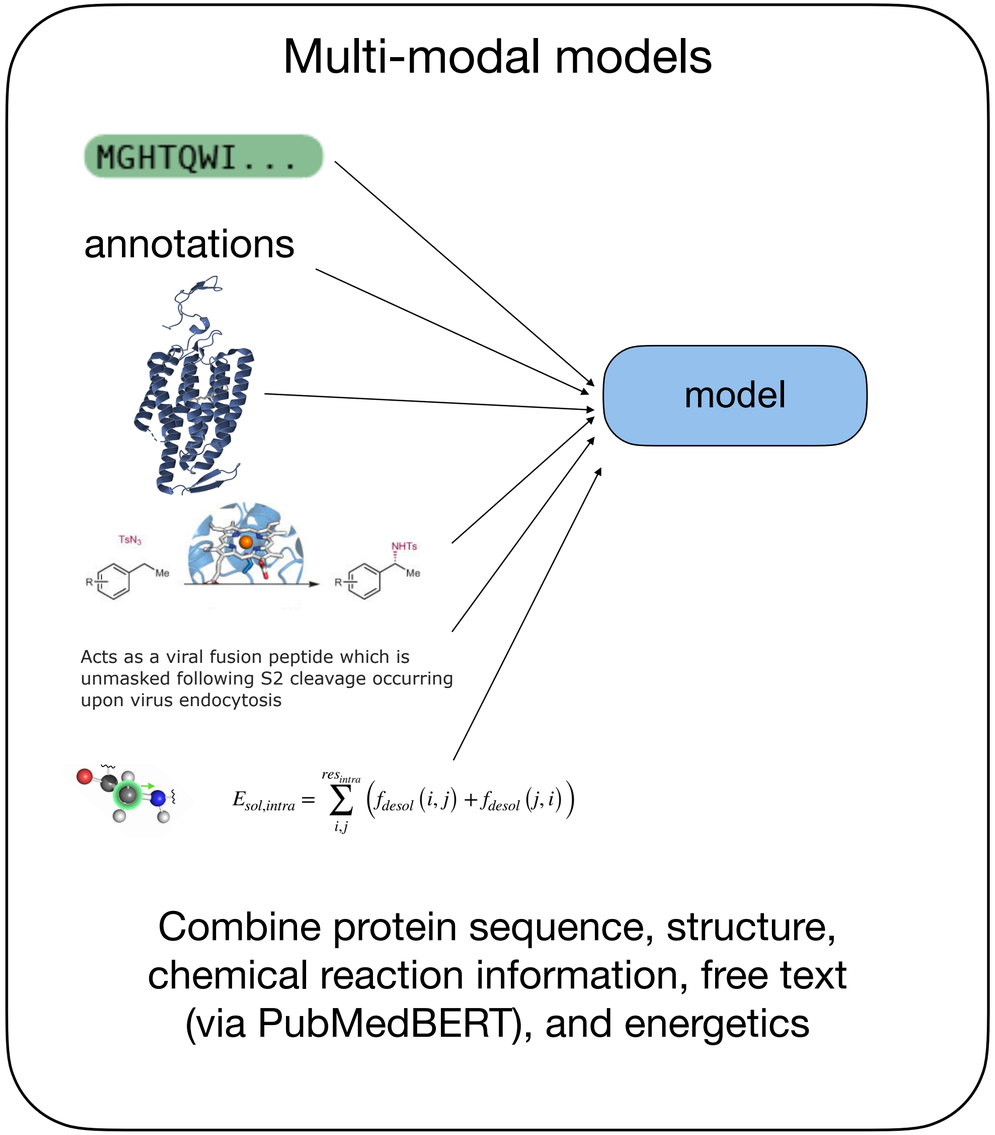




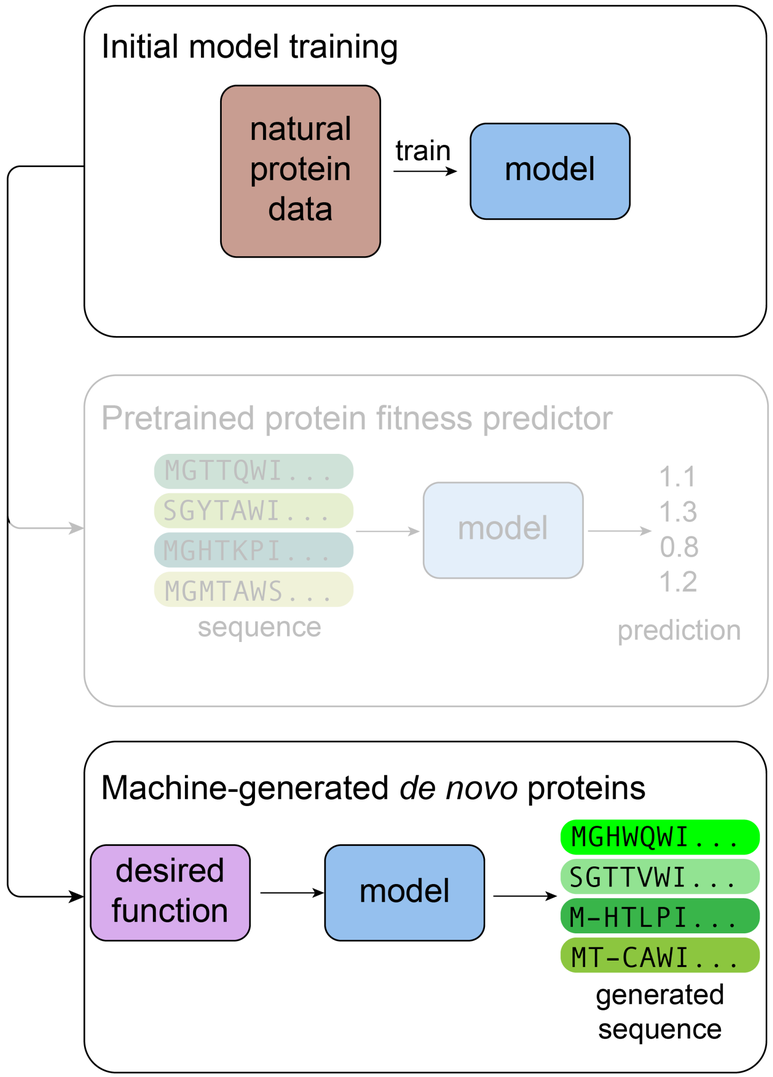
https://github.com/seanrjohnson/protein_scoring

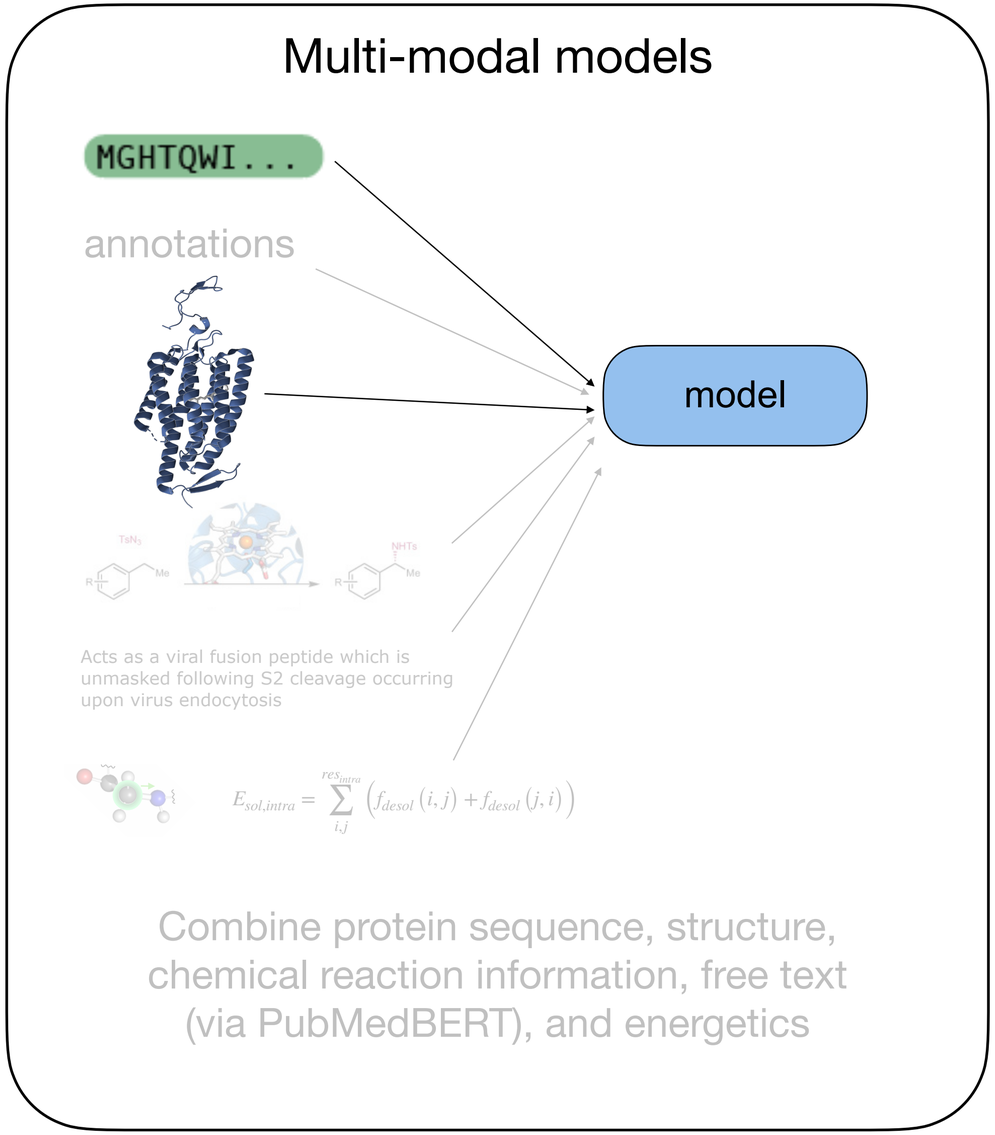
Use multiple data modalities to discover proteins





Use multiple data modalities to discover proteins





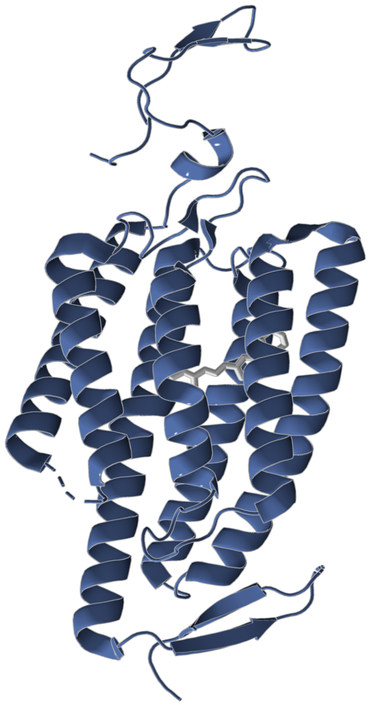
Sequence determines structure determines function

Sequence determines structure determines function

MGTGDHDD...

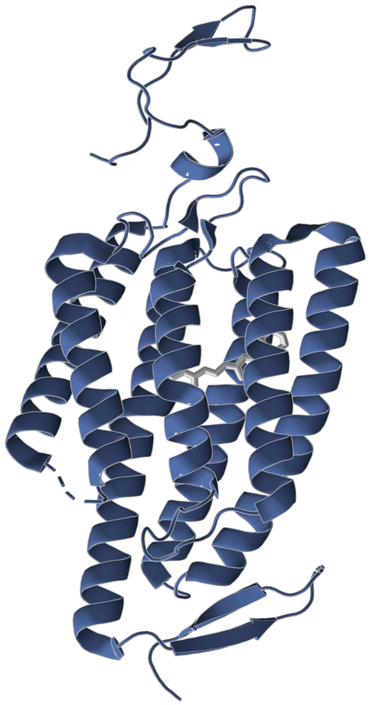
Sequence determines structure determines function

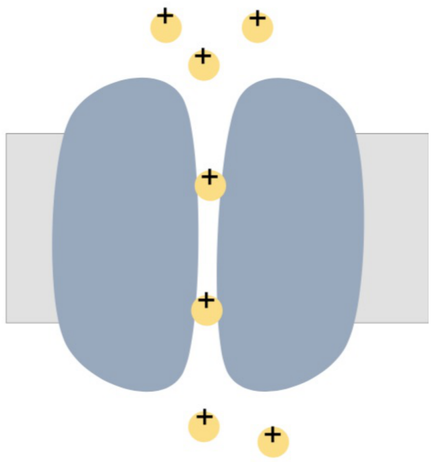
MGTGDHDD... →



Sequence determines structure determines function

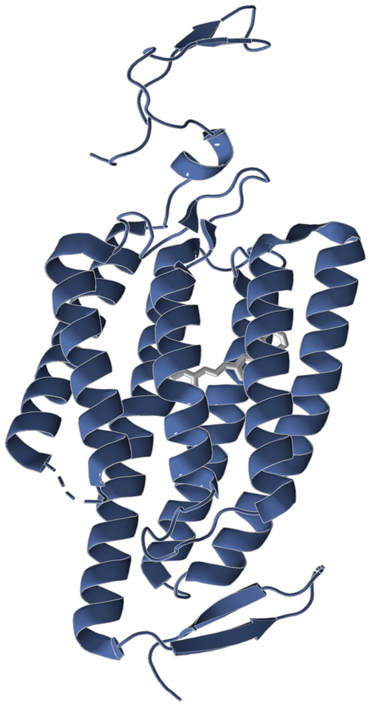
MGTGDHDD...

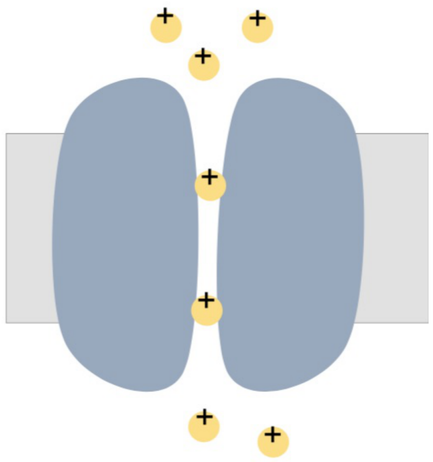




Generating new, designable structures expands functional space

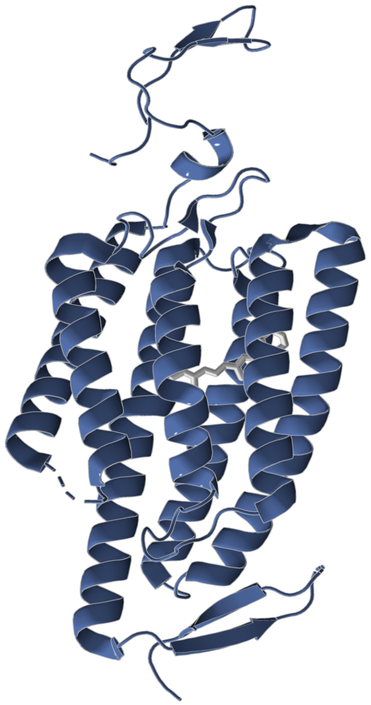
MGTGDHDD...

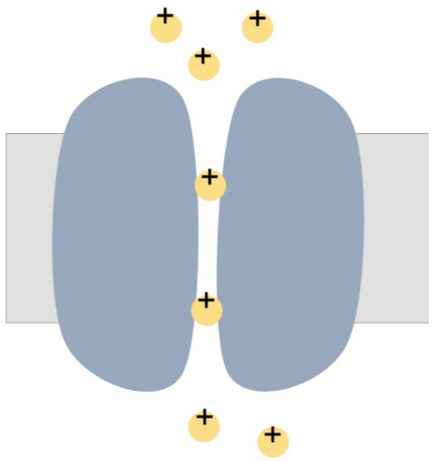




Generating new, designable structures expands functional space

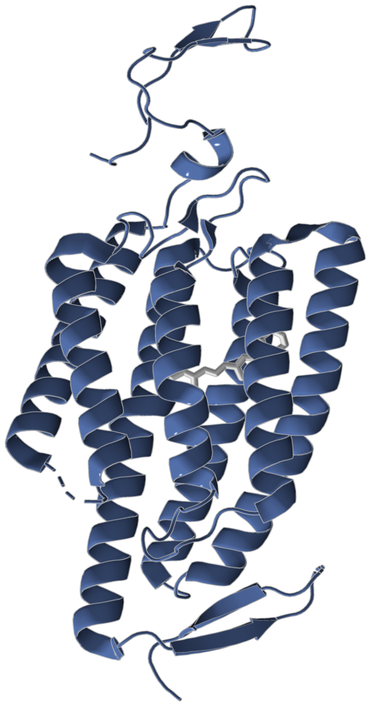
MGTGDHDD...

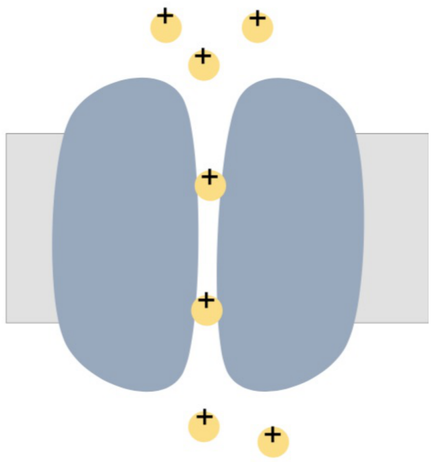




Generating new, designable structures expands functional space

MGTGDHDD...

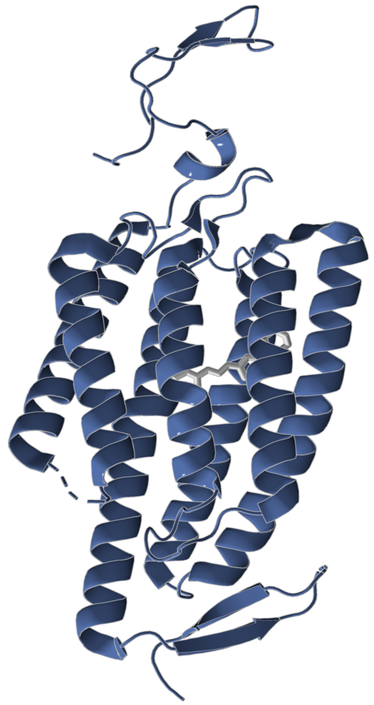


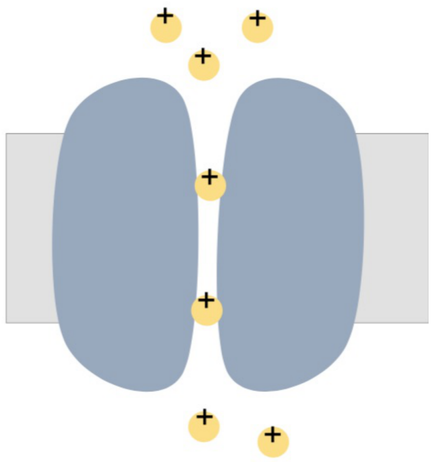


Challenge: Generate diverse and designable structures

Generating new, designable structures expands functional space

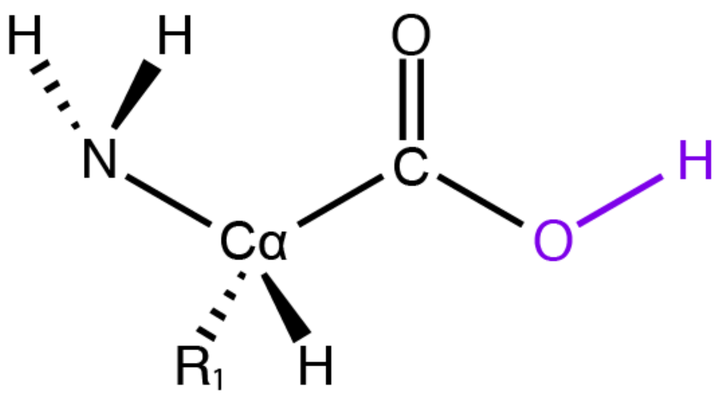
MGTGDHDD...



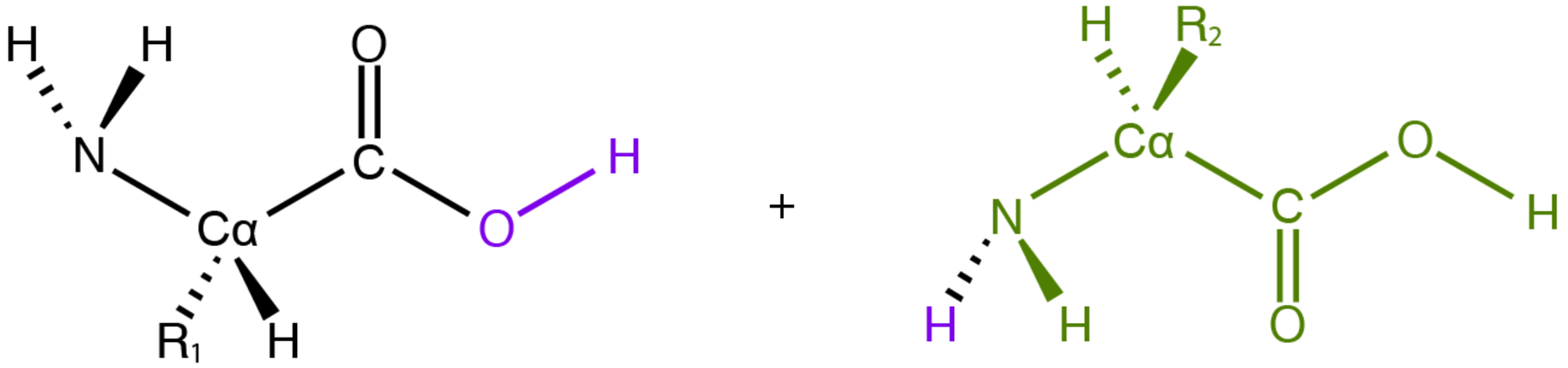


Challenge: Generate diverse and designable structures

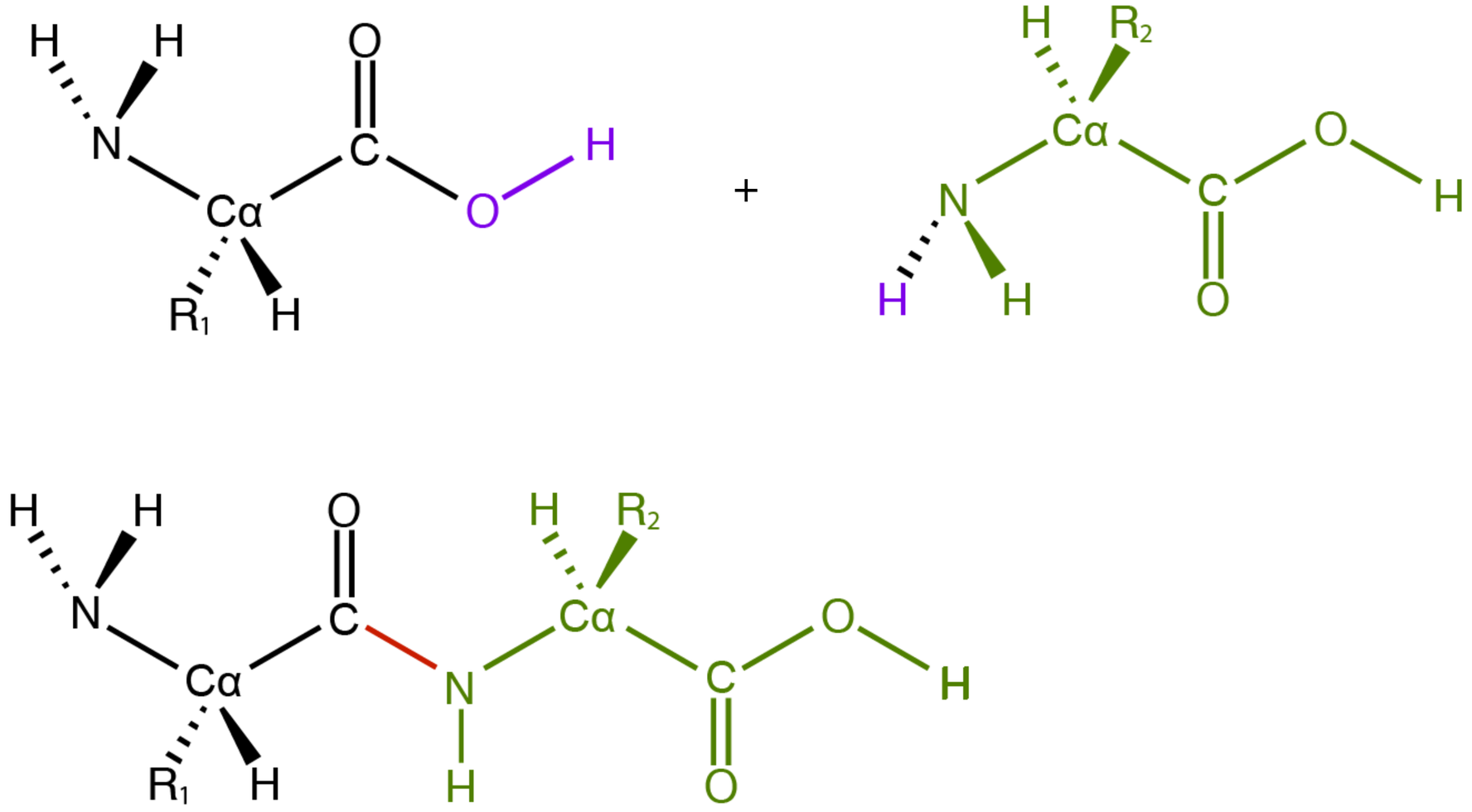
Proteins are polypeptide chains



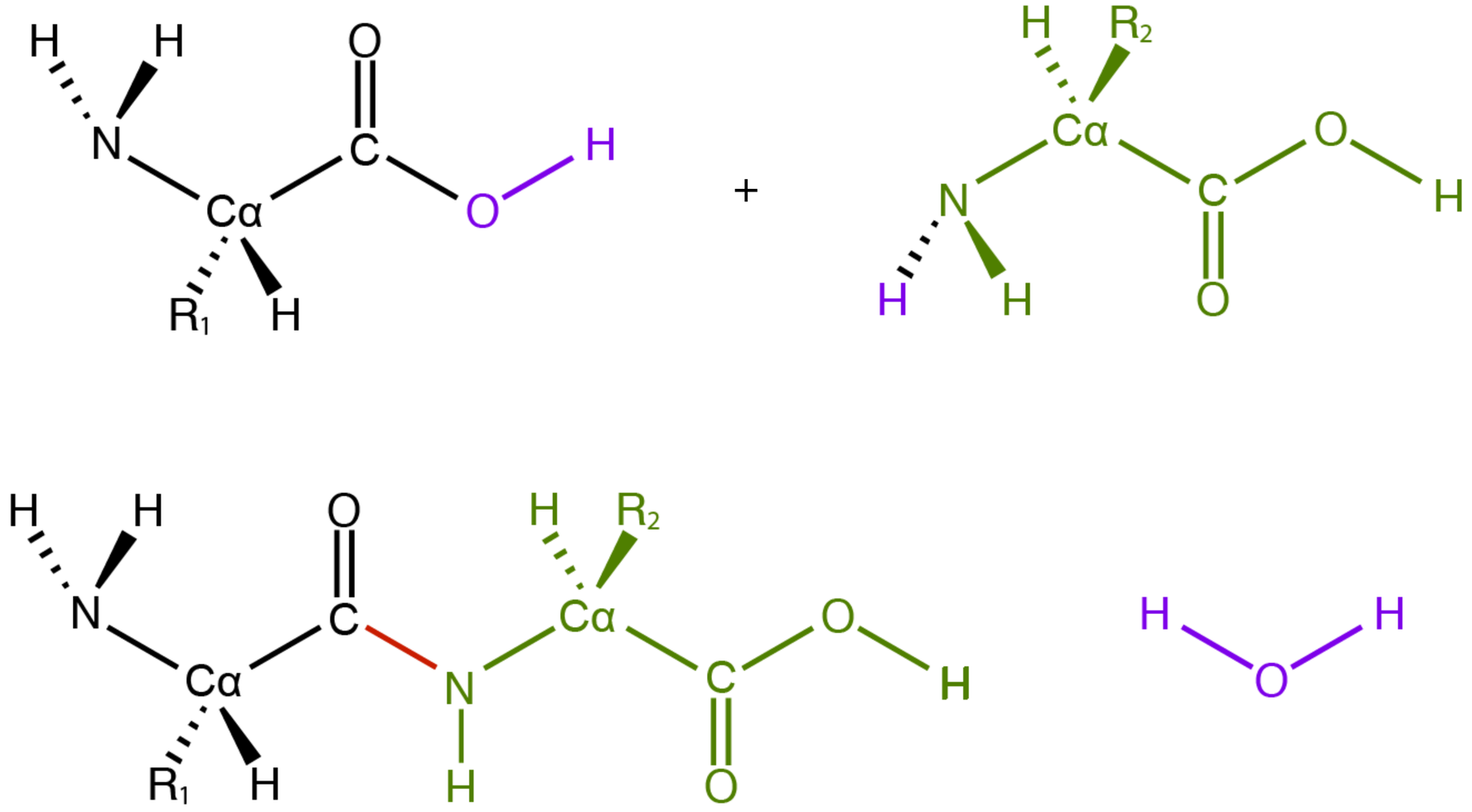
Proteins are polypeptide chains



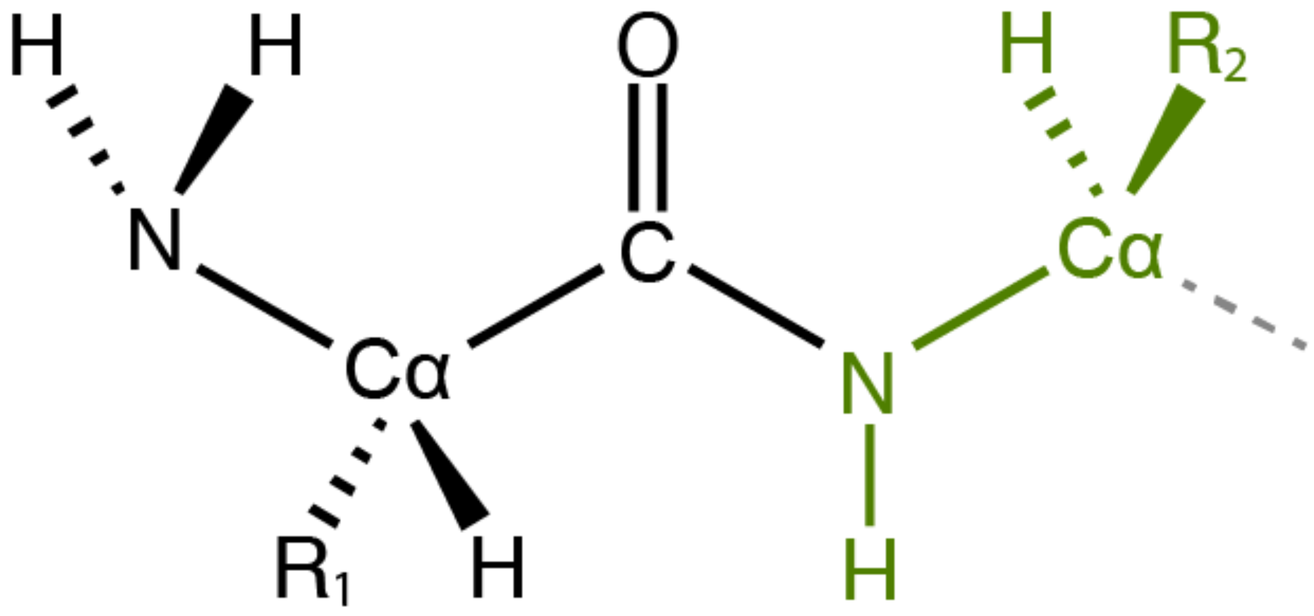
Proteins are polypeptide chains



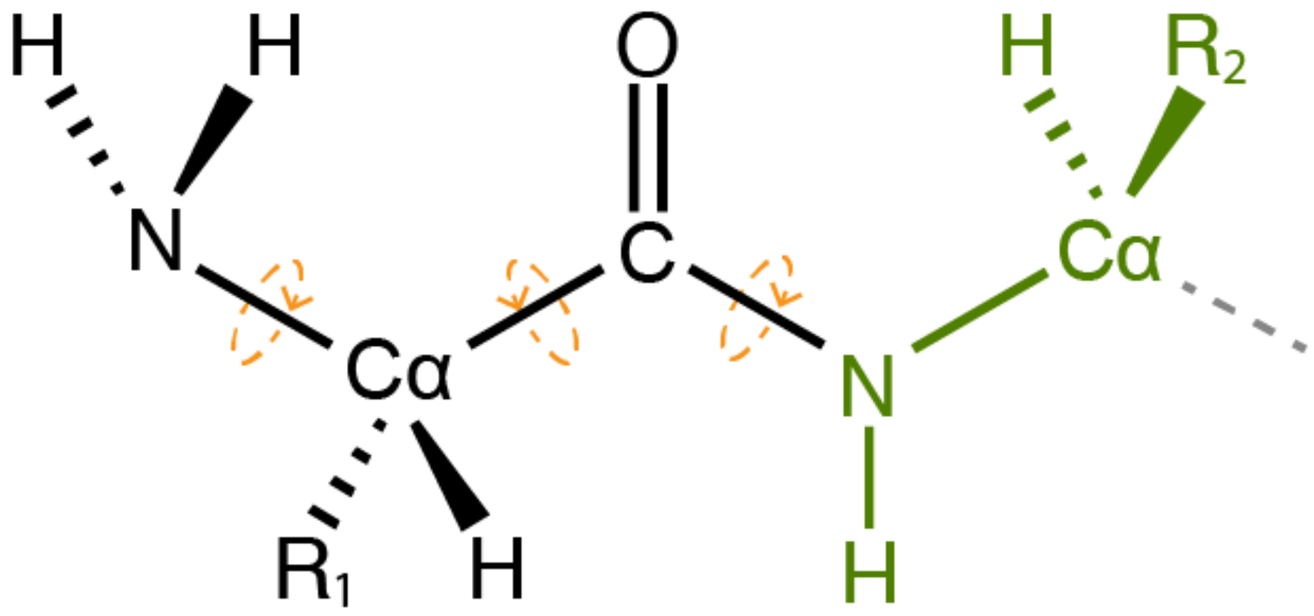
Proteins are polypeptide chains



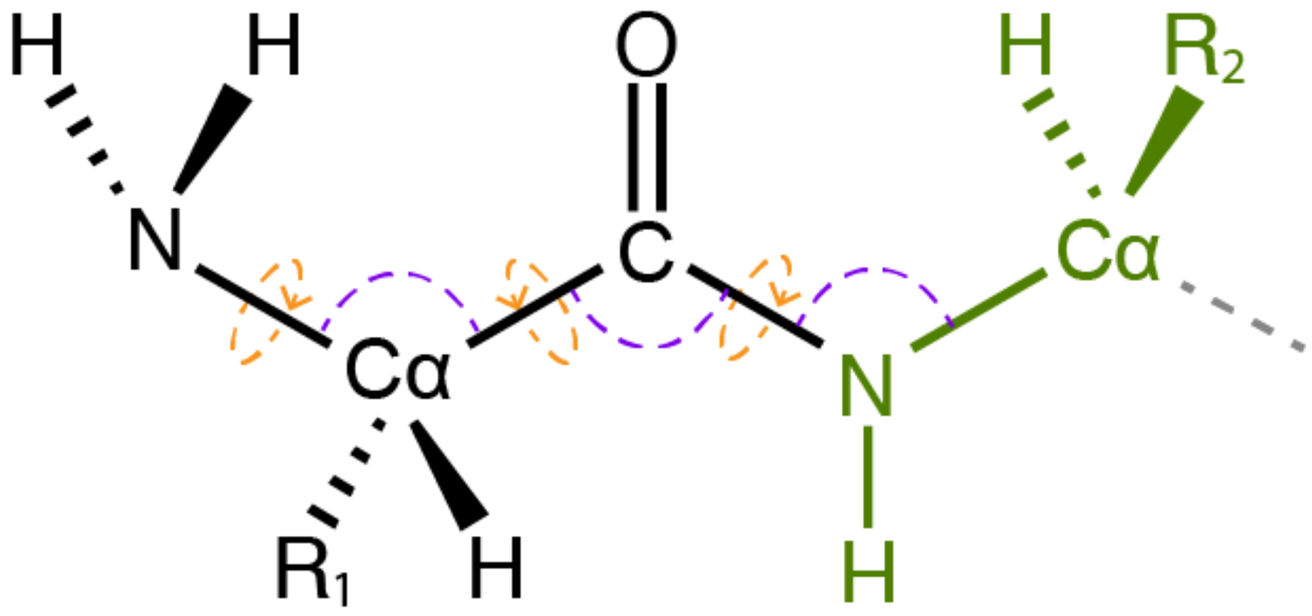
Protein structure is determined by bond angles



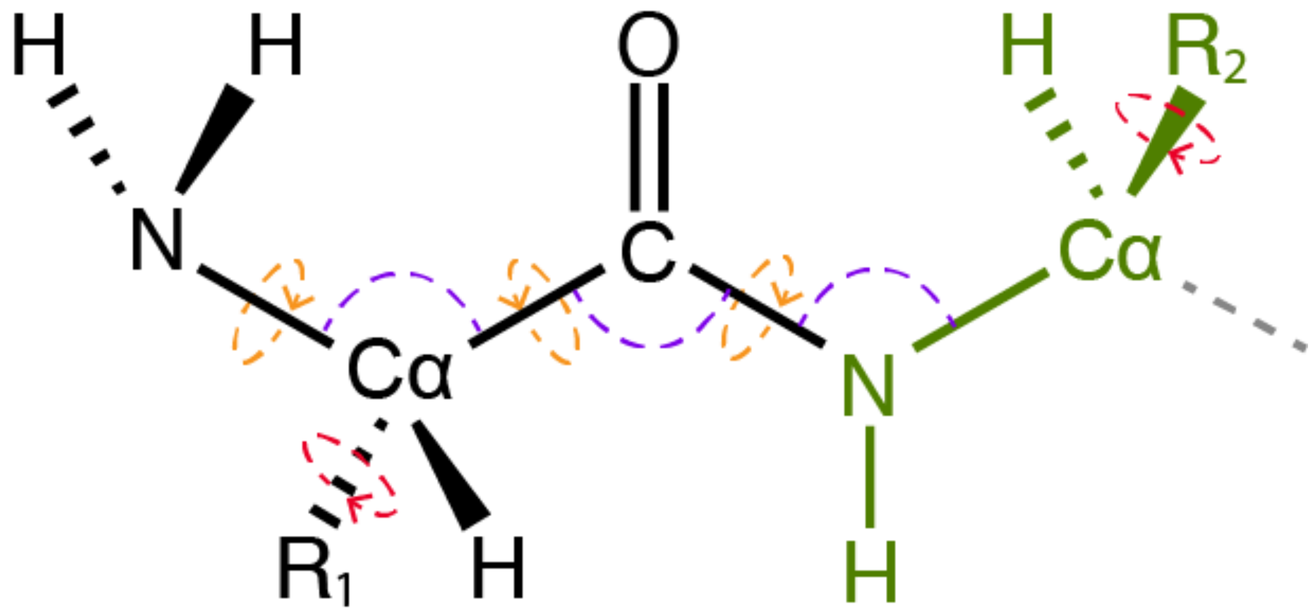
Protein structure is determined by bond angles



Protein structure is determined by bond angles

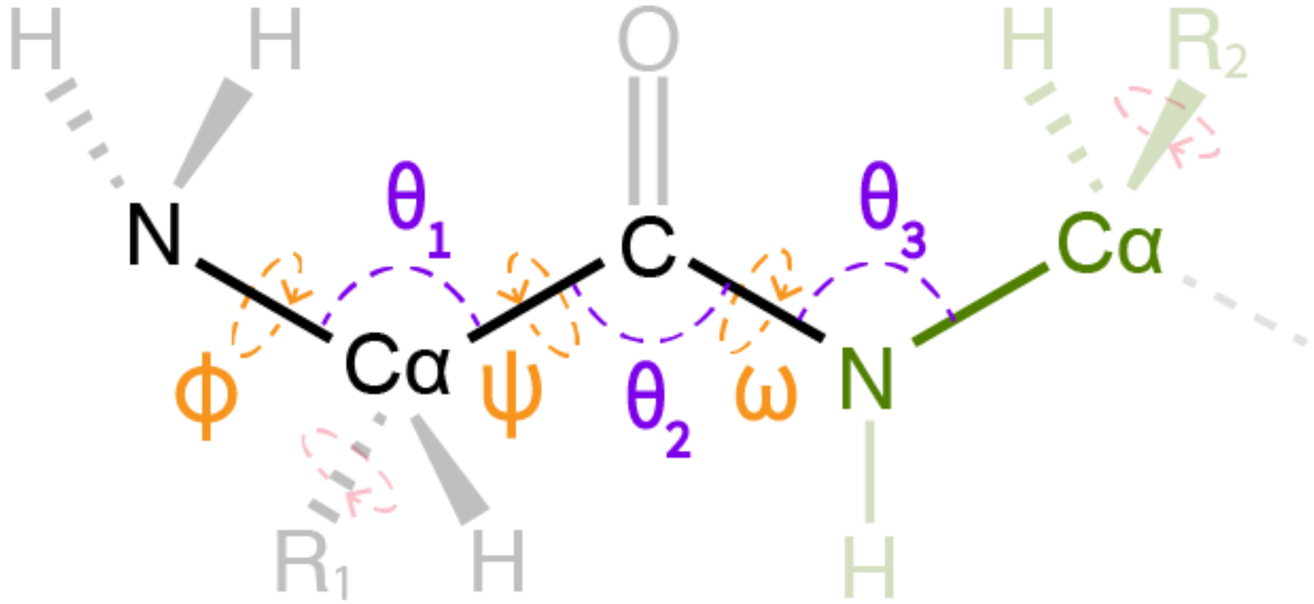


Protein structure is determined by bond angles



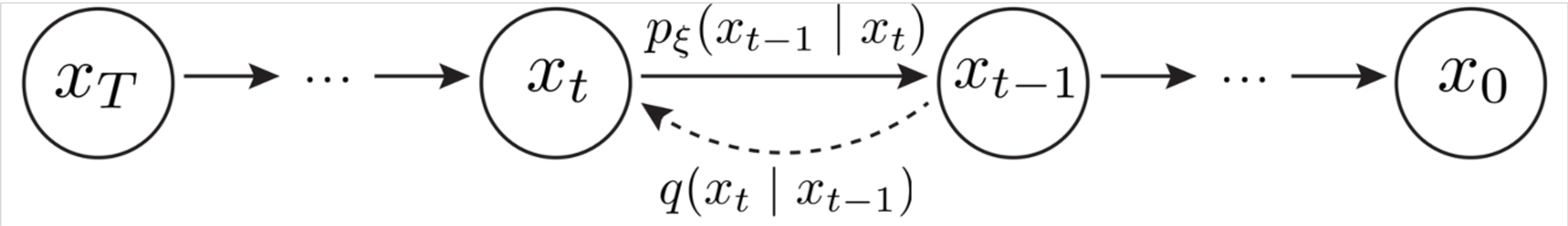
These six angles (for every consecutive pair of amino acids) fully determine the structure

We generate backbone structures represented by dihedral and bond angles

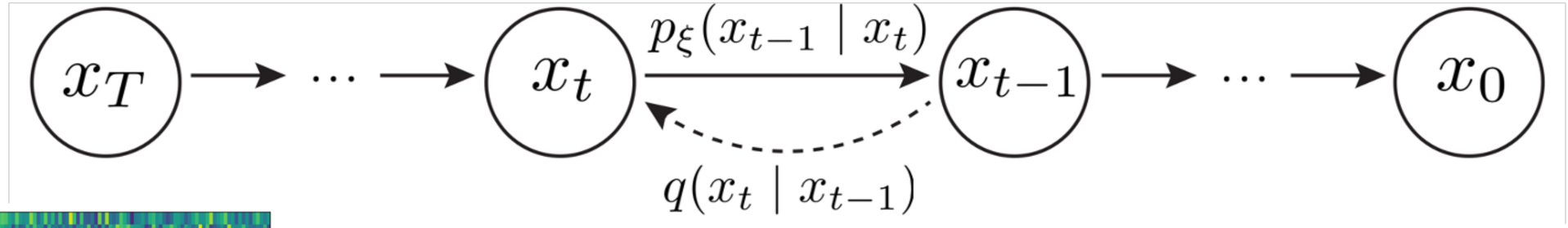


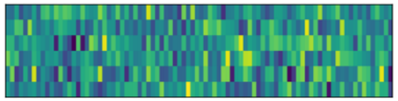
These six angles (for every consecutive pair of amino acids) fully determine the backbone structure

FoldingDiff uses diffusion to generate angles

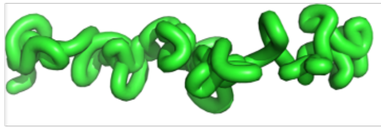


FoldingDiff uses diffusion to generate angles

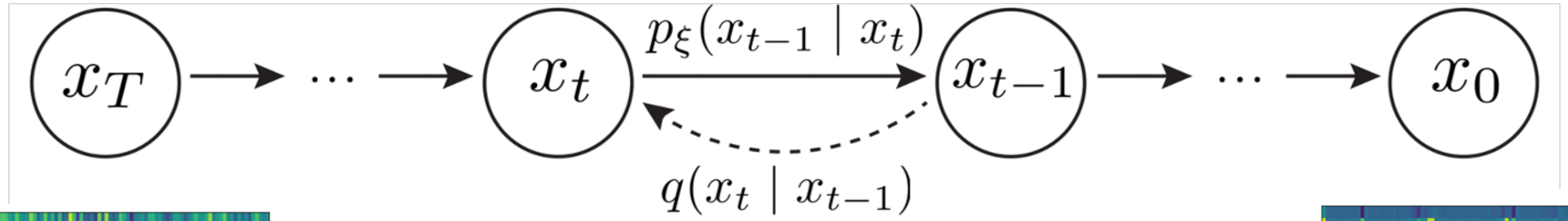


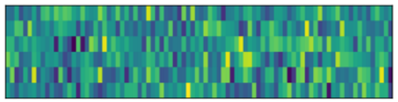


random angles

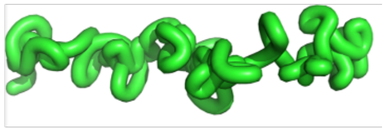


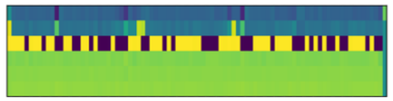
FoldingDiff uses diffusion to generate angles



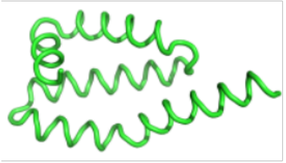


random angles





generated angles

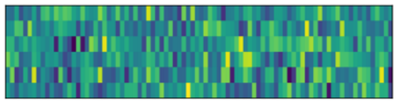


FoldingDiff uses diffusion to generate angles

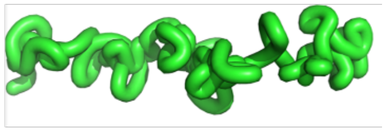
Vanilla transformer

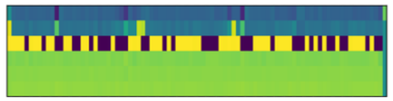
$$x_T \rightarrow \cdots \rightarrow x_t \xrightarrow{p_\xi(x_{t-1} \mid x_t)} x_{t-1} \rightarrow \cdots \rightarrow x_0$$

$$q(x_t \mid x_{t-1})$$

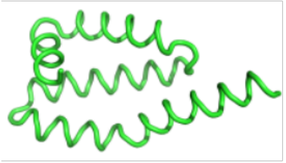


random angles



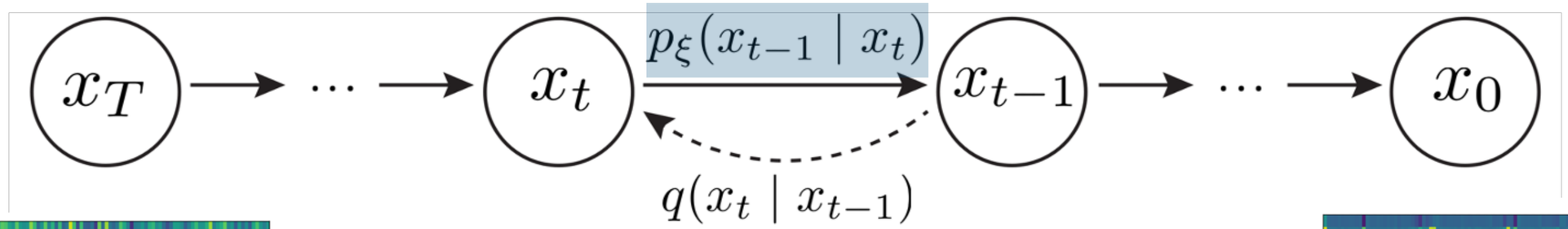


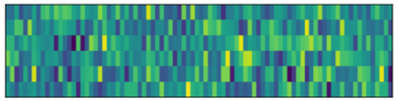
generated angles



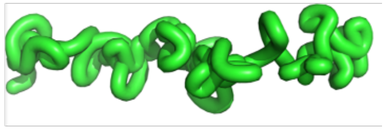
FoldingDiff uses diffusion to generate angles

Vanilla transformer



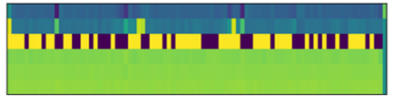


random angles

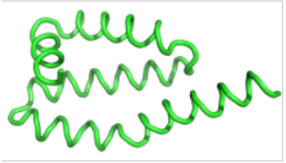


30k structures from CATH

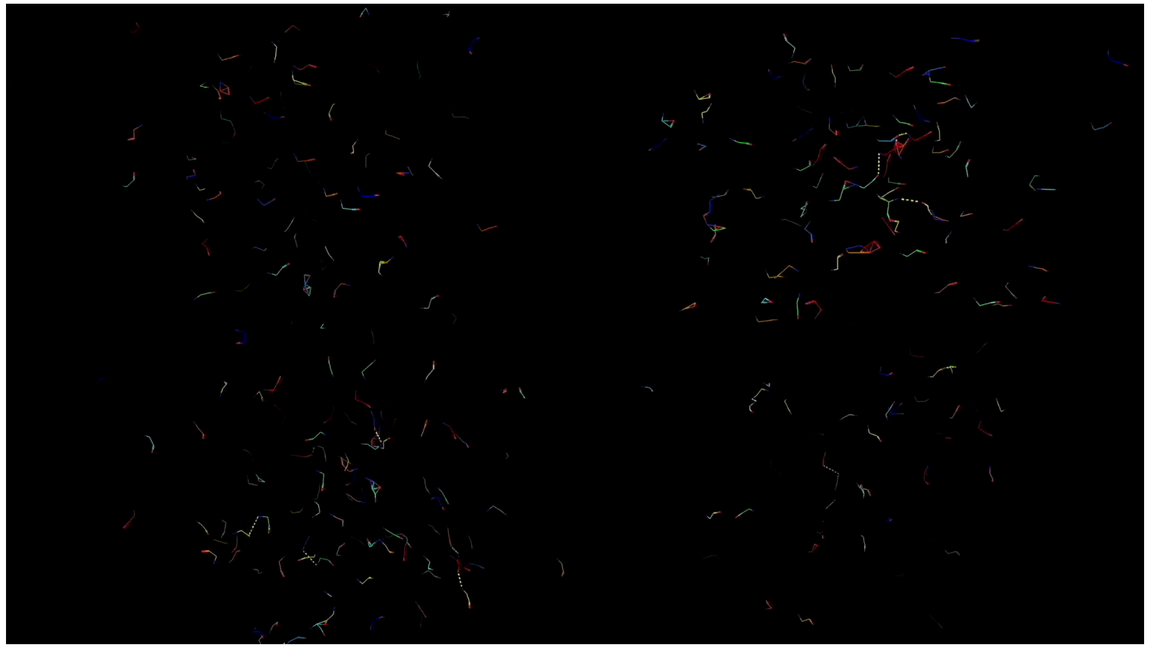
40 - 128 residues



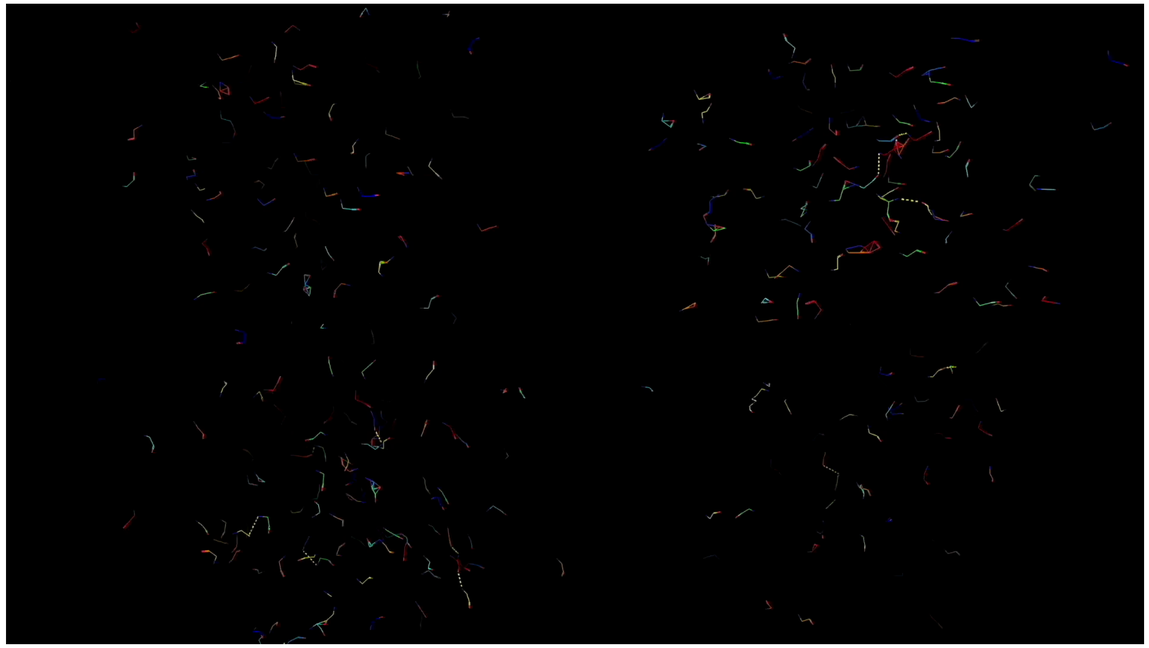
generated angles



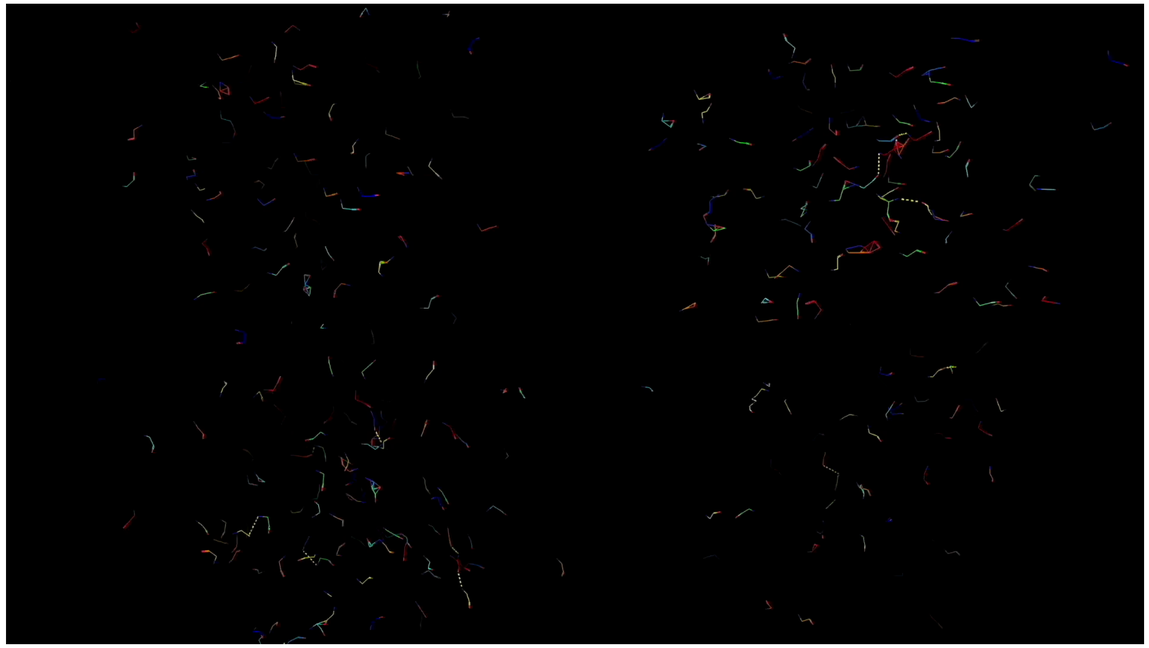
Diffusion on 3D coordinates requires equivariances



Diffusion on 3D coordinates requires equivariances

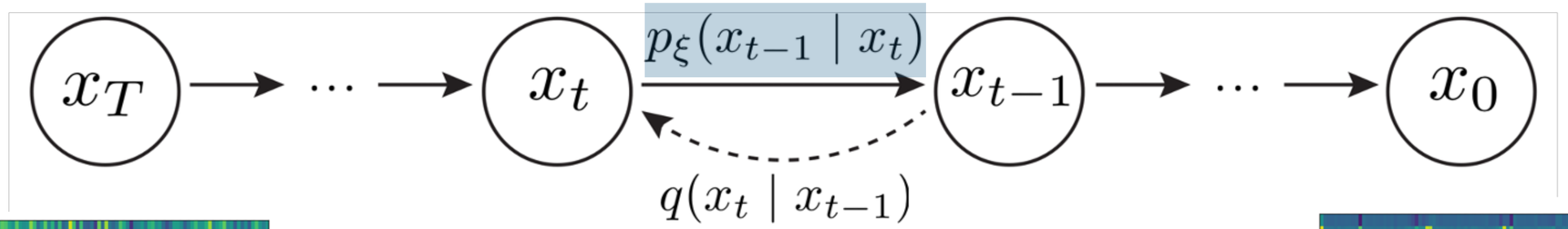


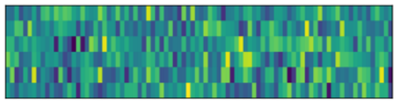
Diffusion on 3D coordinates requires equivariances



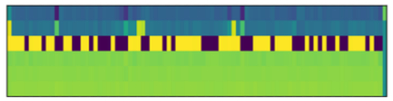
Evaluate generations at 3 levels

Vanilla transformer





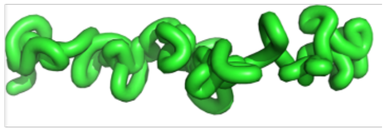
random angles

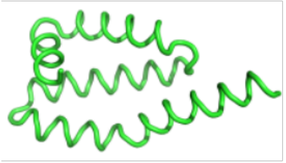


generated angles

30k structures from CATH

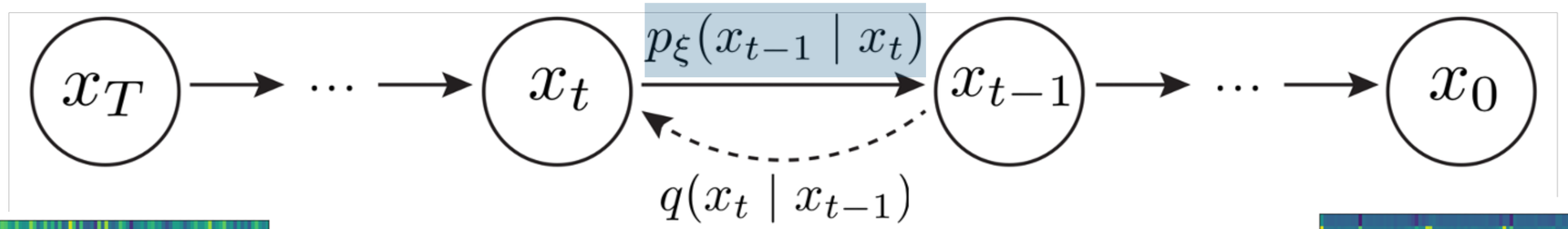
40 - 128 residues

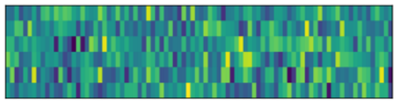




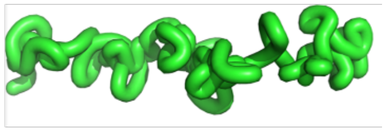
Evaluate generations at 3 levels

Vanilla transformer



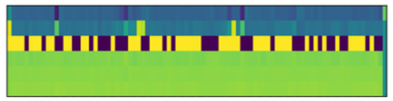


random angles

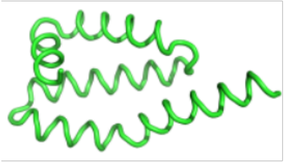


30k structures from CATH

40 - 128 residues



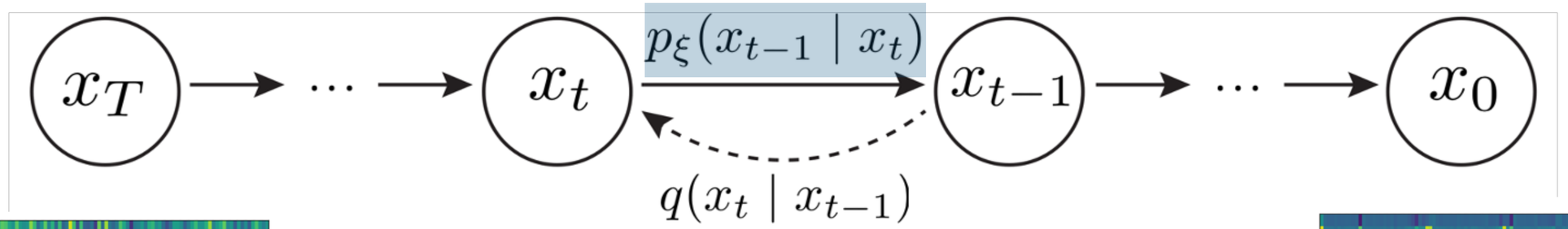
generated angles

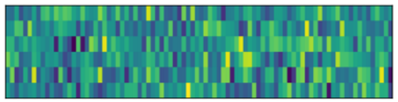


generated angles

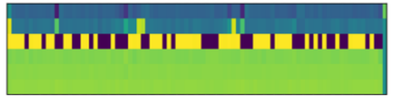
Evaluate generations at 3 levels

Vanilla transformer





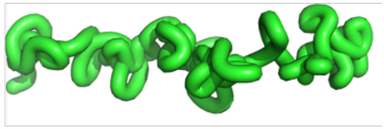
random angles

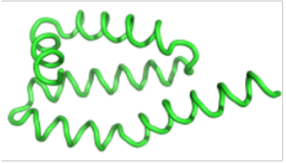


generated angles

30k structures from CATH

40 - 128 residues



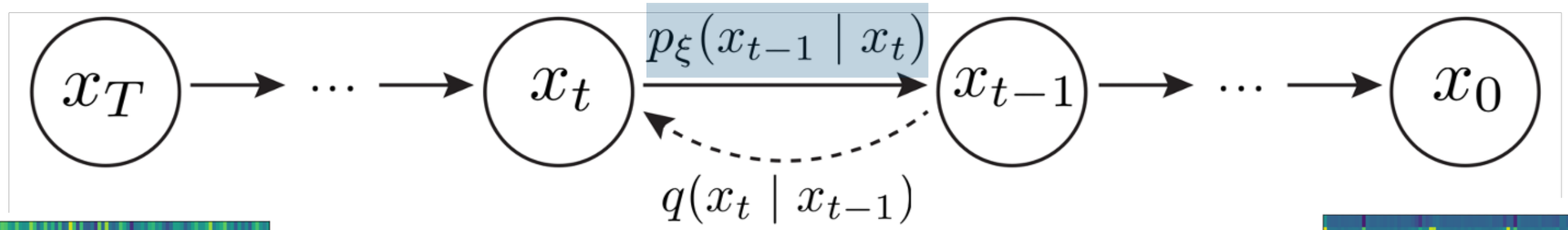


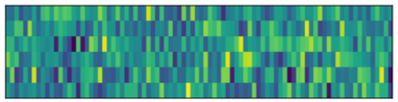
generated angles

structural motifs

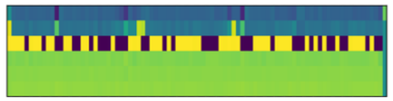
Evaluate generations at 3 levels

Vanilla transformer





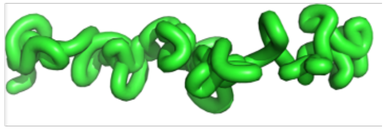
random angles

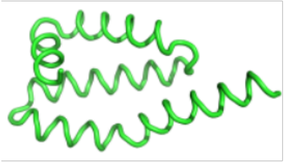


generated angles

30k structures from CATH

40 - 128 residues





generated angles

structural motifs

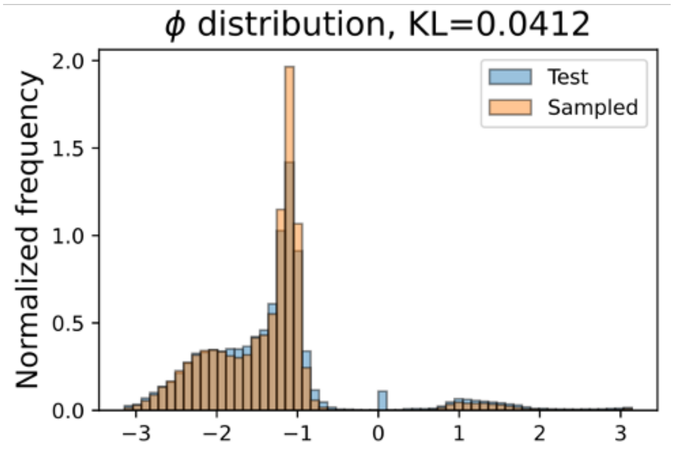
overall structures

Generated angles match test distribution

noise → sample → compare to true distribution

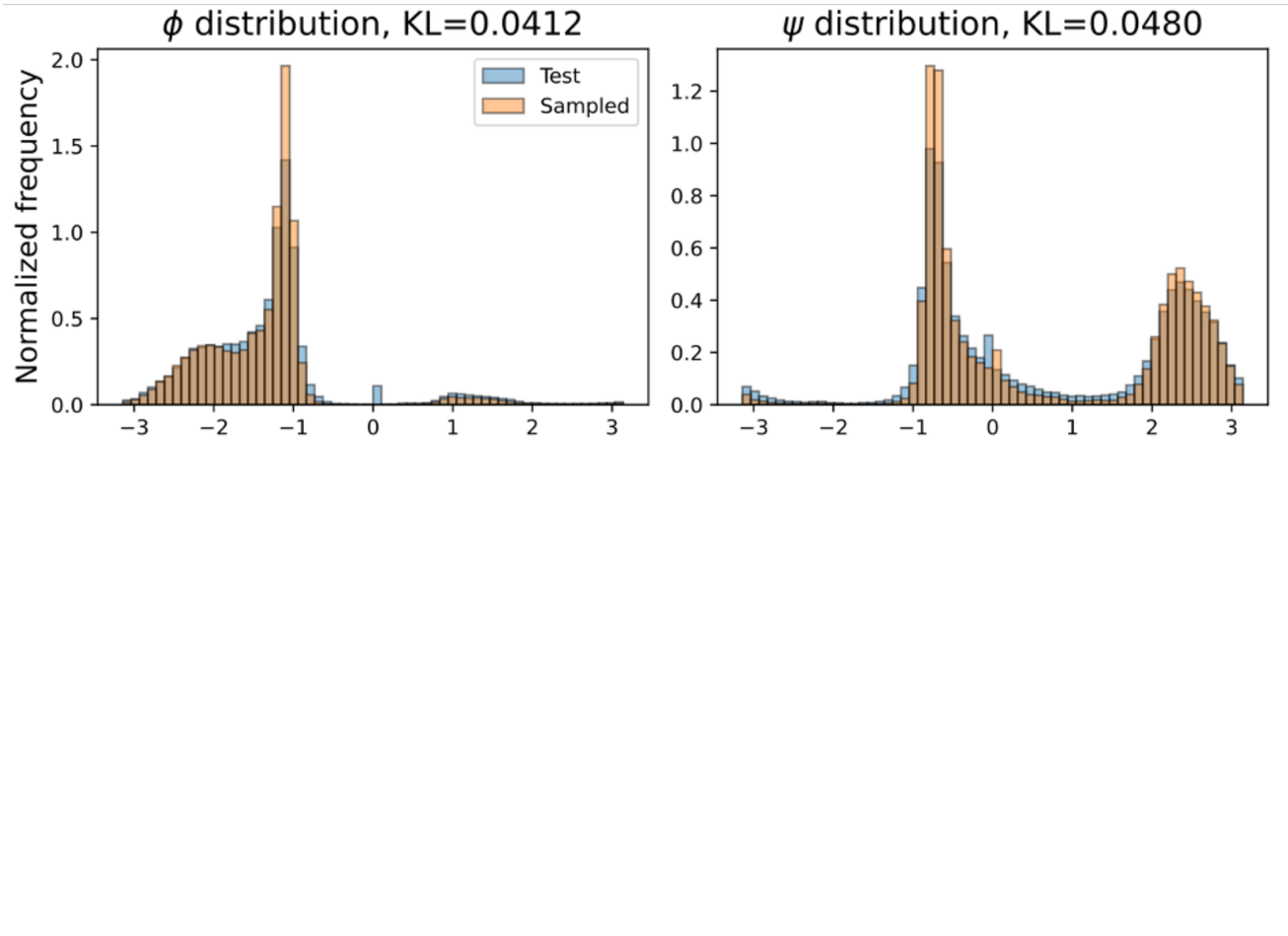
Generated angles match test distribution

noise → sample → compare to true distribution



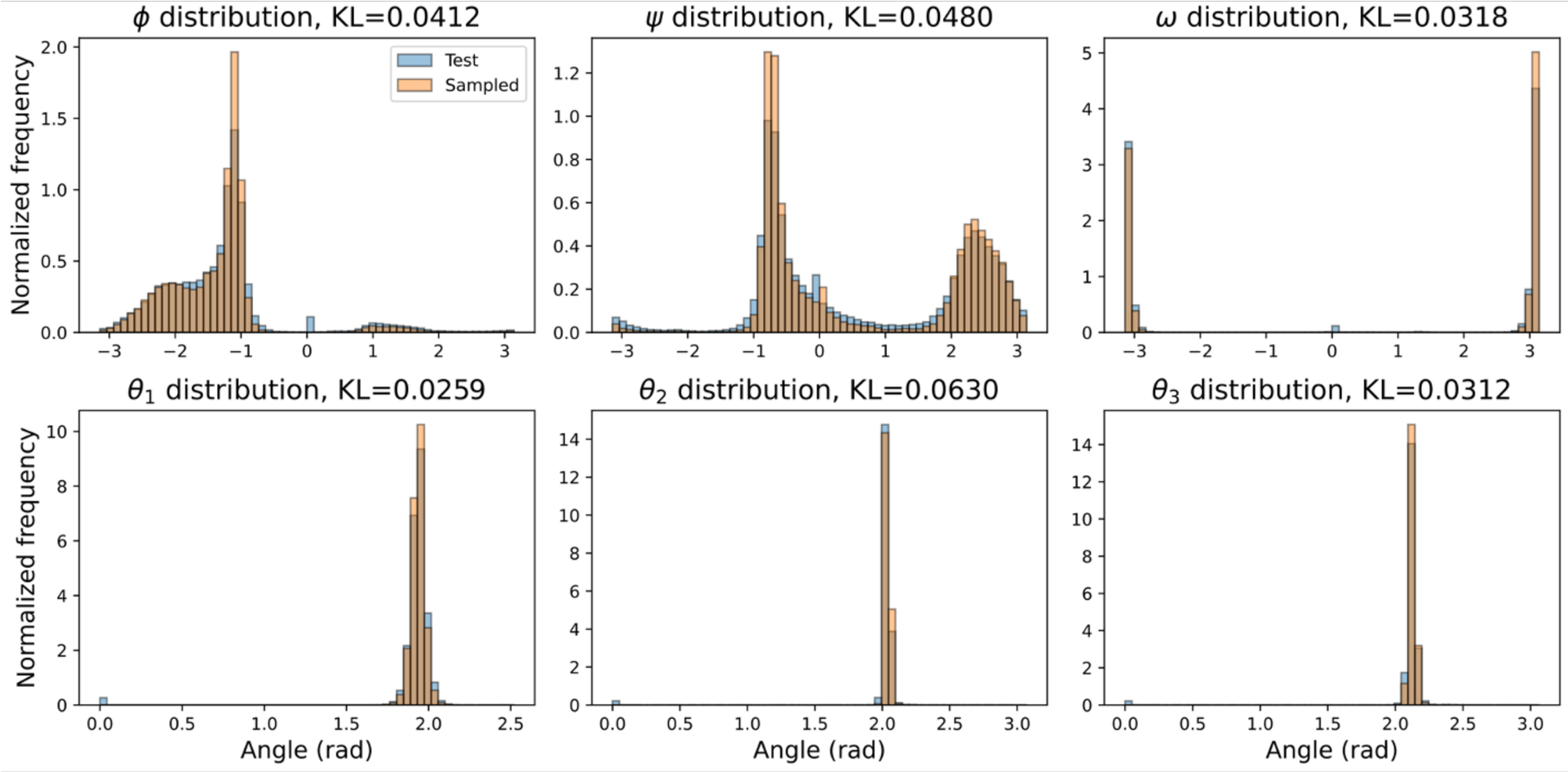
Generated angles match test distribution

noise → sample → compare to true distribution



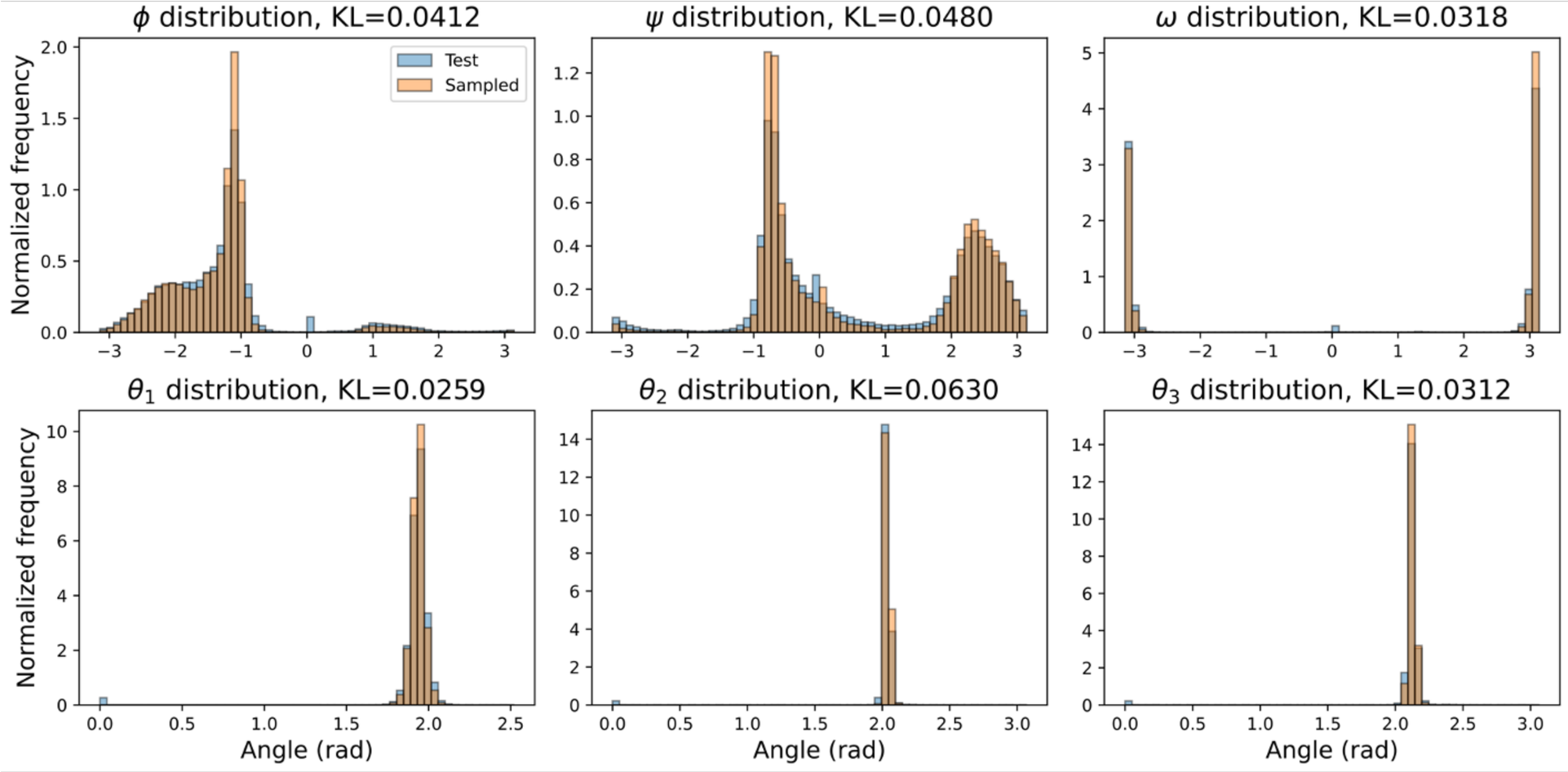
Generated angles match test distribution

noise → sample → compare to true distribution



Generated angles match test distribution

noise → sample → compare to true distribution



Generated distributions match natural distribution of individual angles

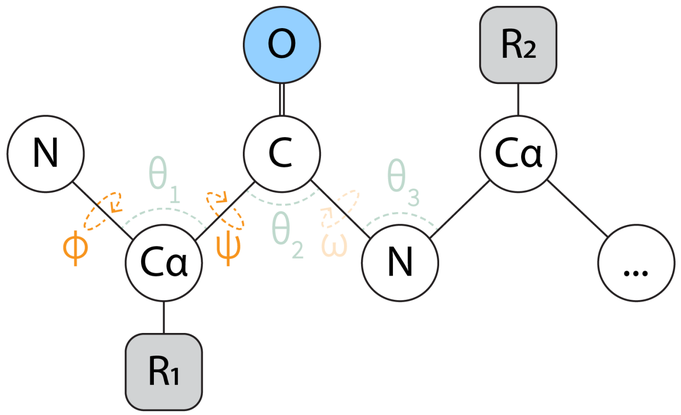
FoldingDiff captures correlations between angles

noise → sample → compare (ϕ, ψ) co-occurence

FoldingDiff generates correlations that define common structural motifs

FoldingDiff captures correlations between angles

noise → sample → compare (ϕ, ψ) co-occurence

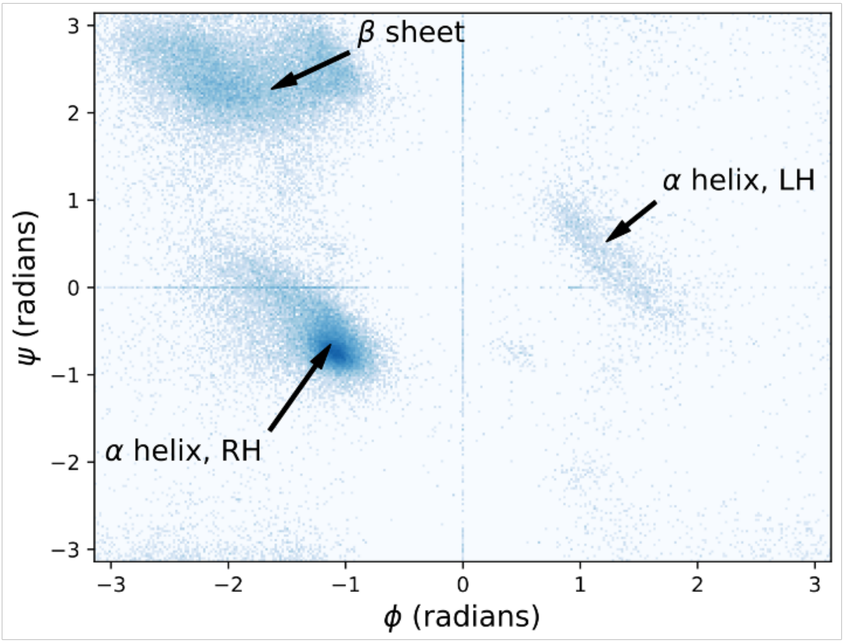


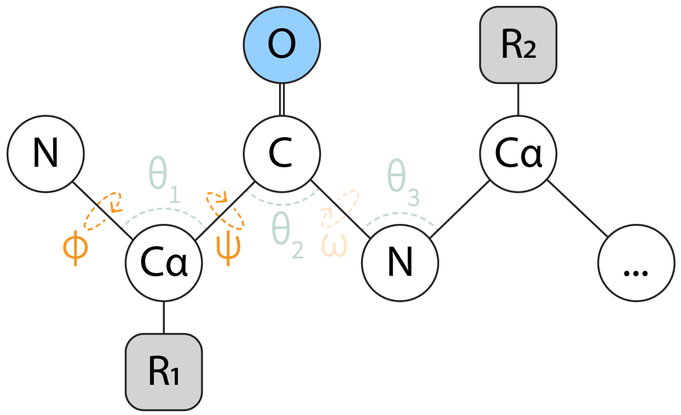
FoldingDiff generates correlations that define common structural motifs

FoldingDiff captures correlations between angles

noise → sample → compare (ϕ, ψ) co-occurence

True



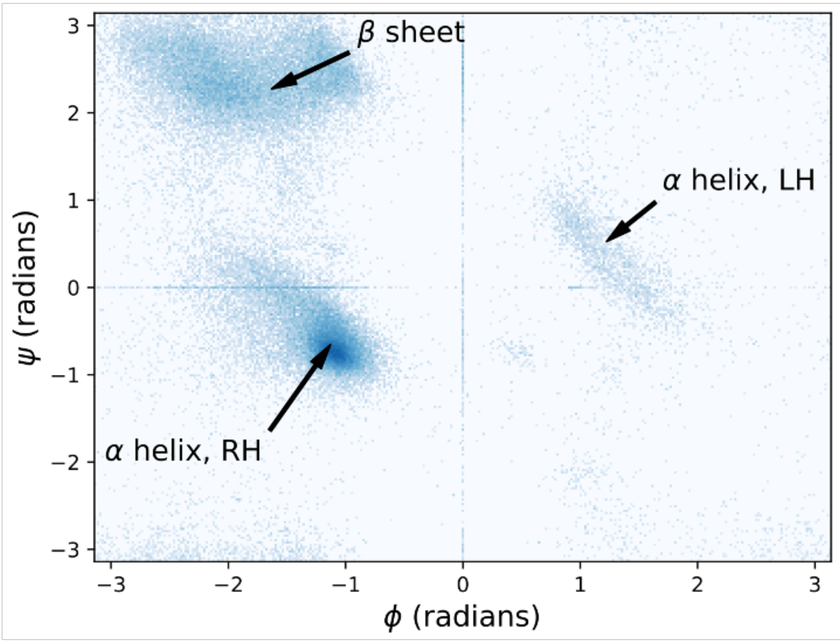


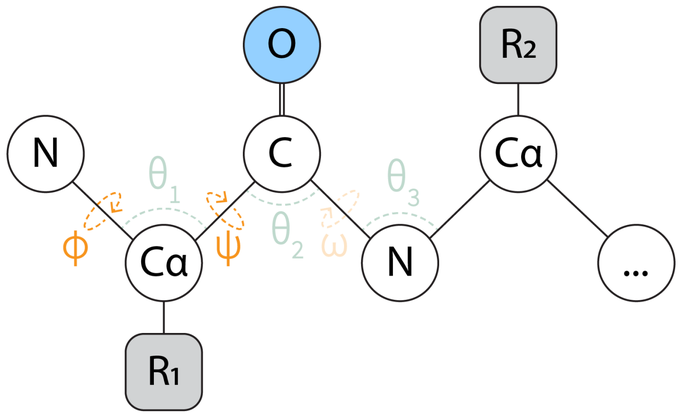
FoldingDiff generates correlations that define common structural motifs

FoldingDiff captures correlations between angles

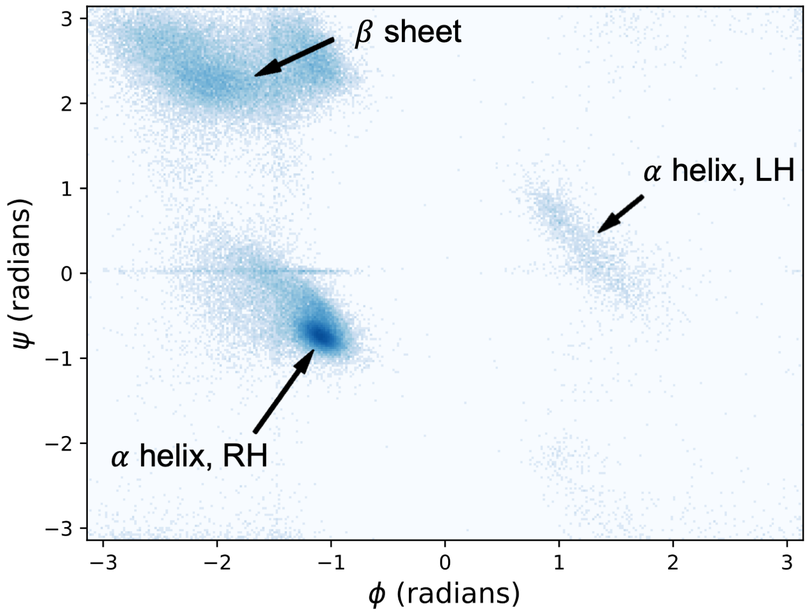
noise → sample → compare (ϕ, ψ) co-occurence

True





Generated

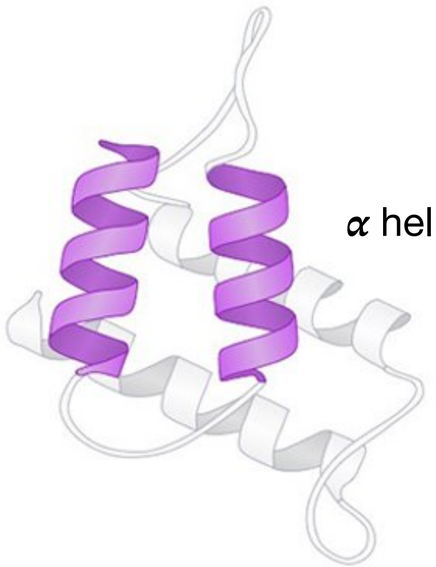


FoldingDiff generates correlations that define common structural motifs

Generated secondary structures match test structures

noise → sample → measure helix, sheet structures

Generated secondary structures match test structures

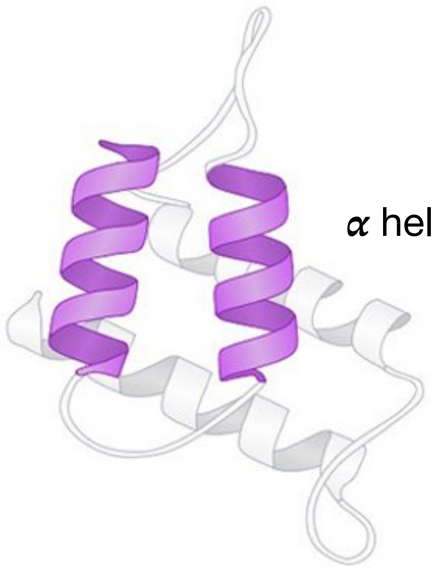


α helix

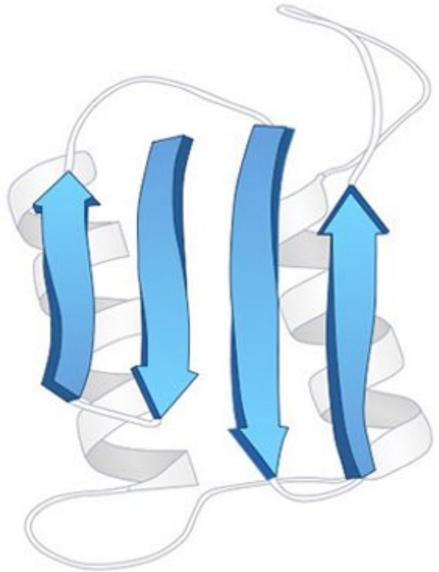
noise → sample → measure helix, sheet structures

Generated secondary structures match test structures

noise → sample → measure helix, sheet structures



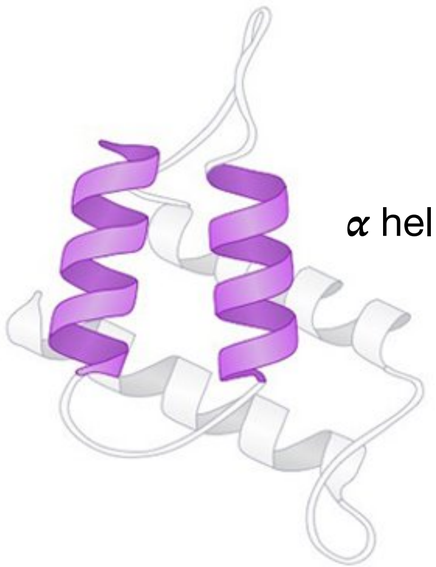
α helix



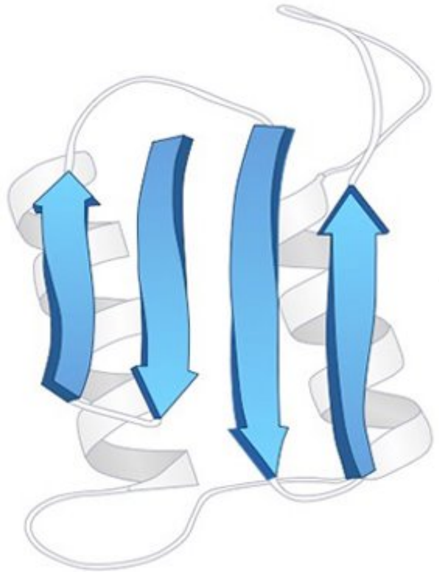
β sheet

Generated secondary structures match test structures

noise → sample → measure helix, sheet structures

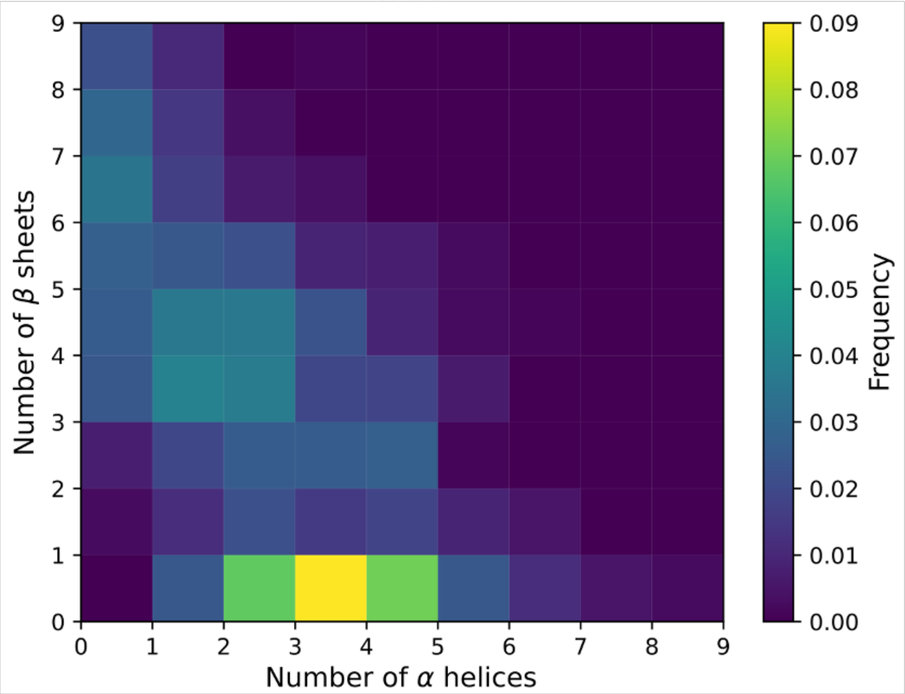


α helix



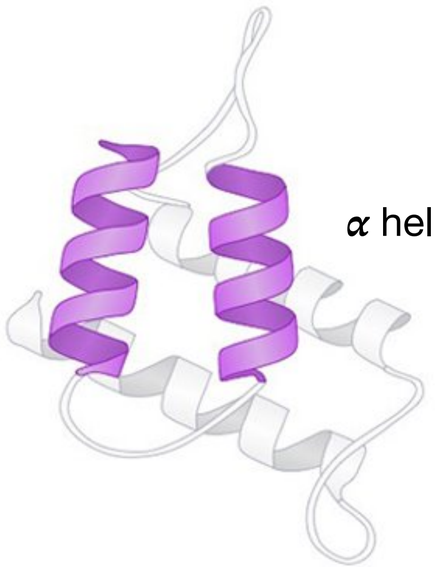
β sheet

True

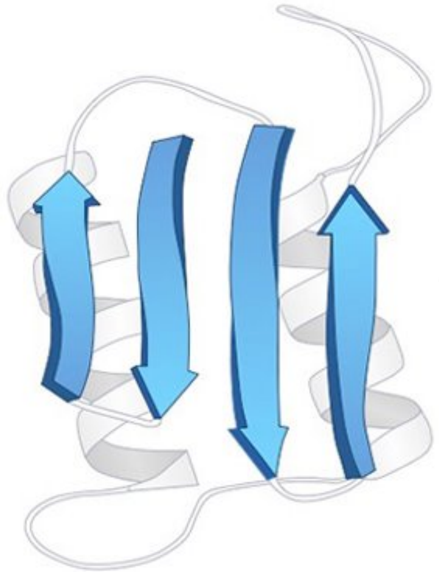


Generated secondary structures match test structures

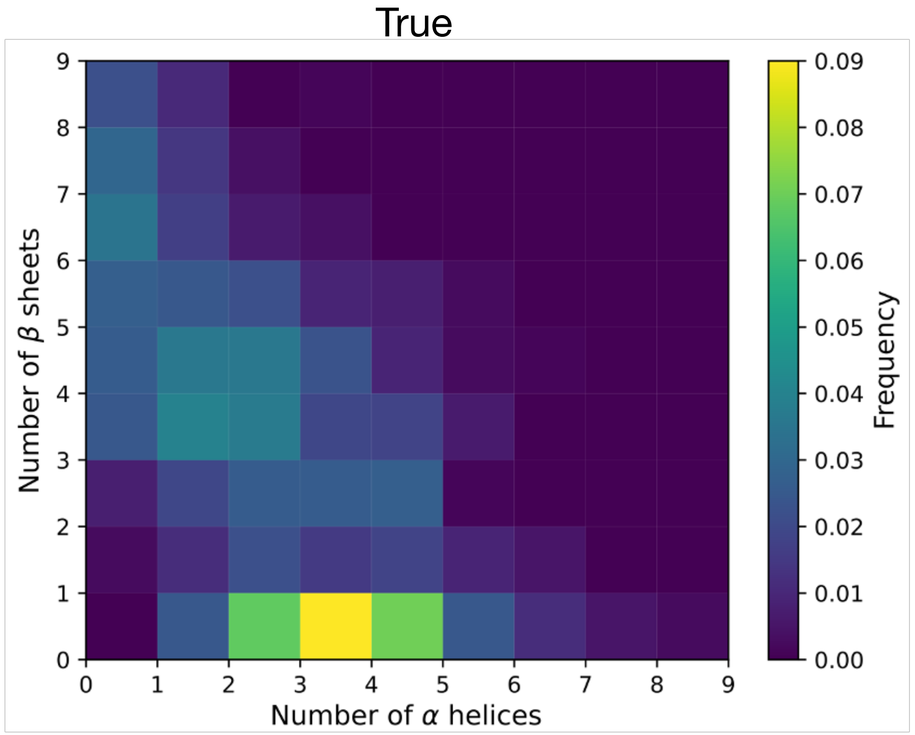
noise → sample → measure helix, sheet structures

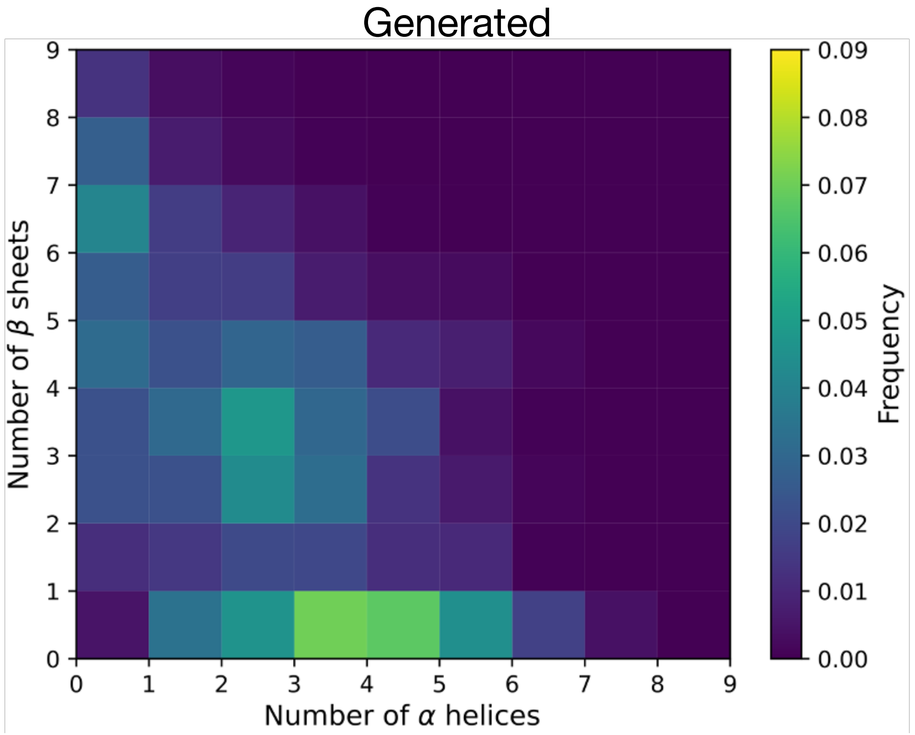


α helix



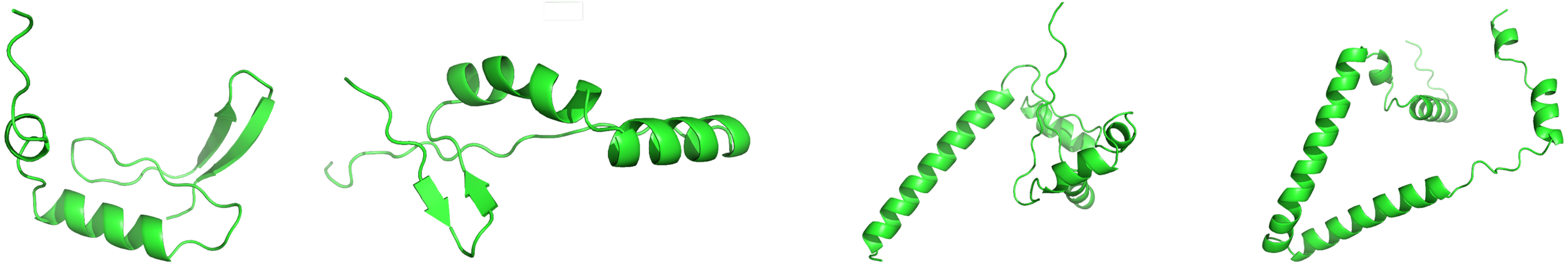
β sheet



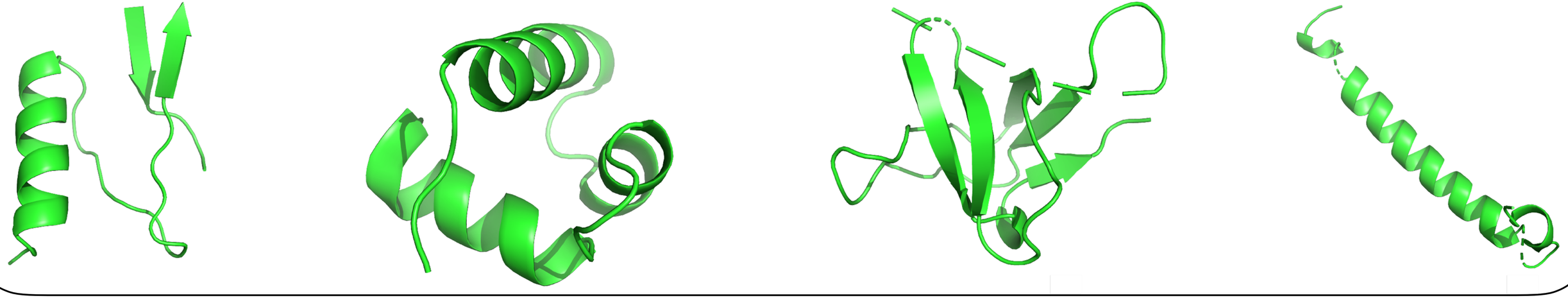


Generated structures look reasonable

Generated structures



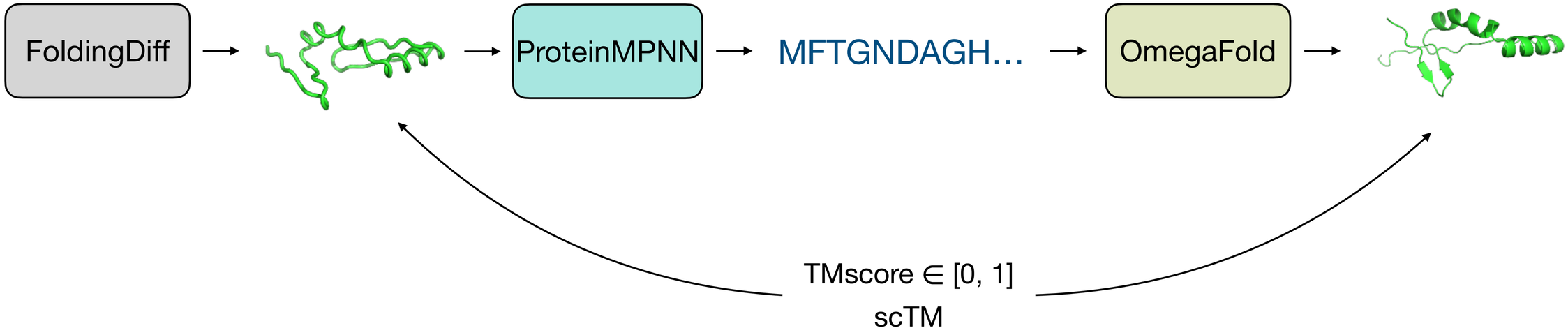
Training set structures



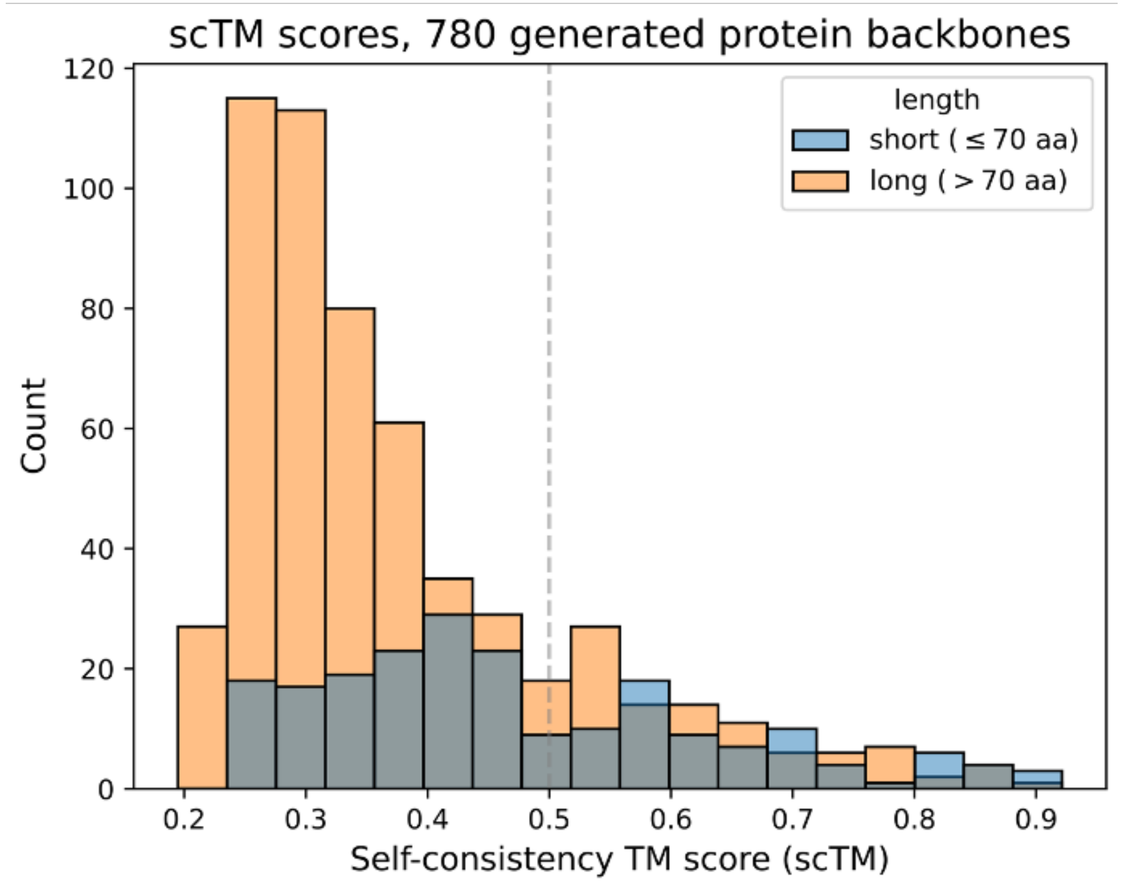
Measure designability of structures with self-consistency TMscore

Measure designability of structures with self-consistency TMscore

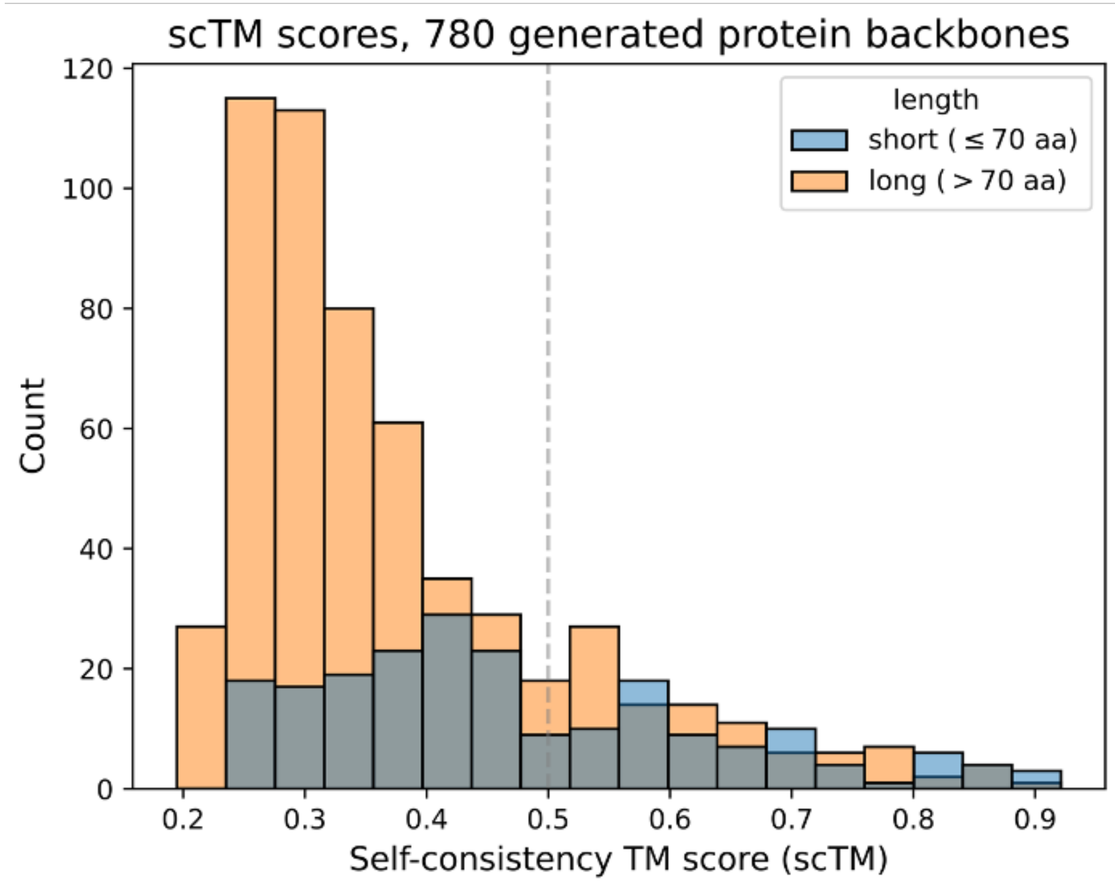
Measure designability of structures with self-consistency TMscore



Many generated structures are designable

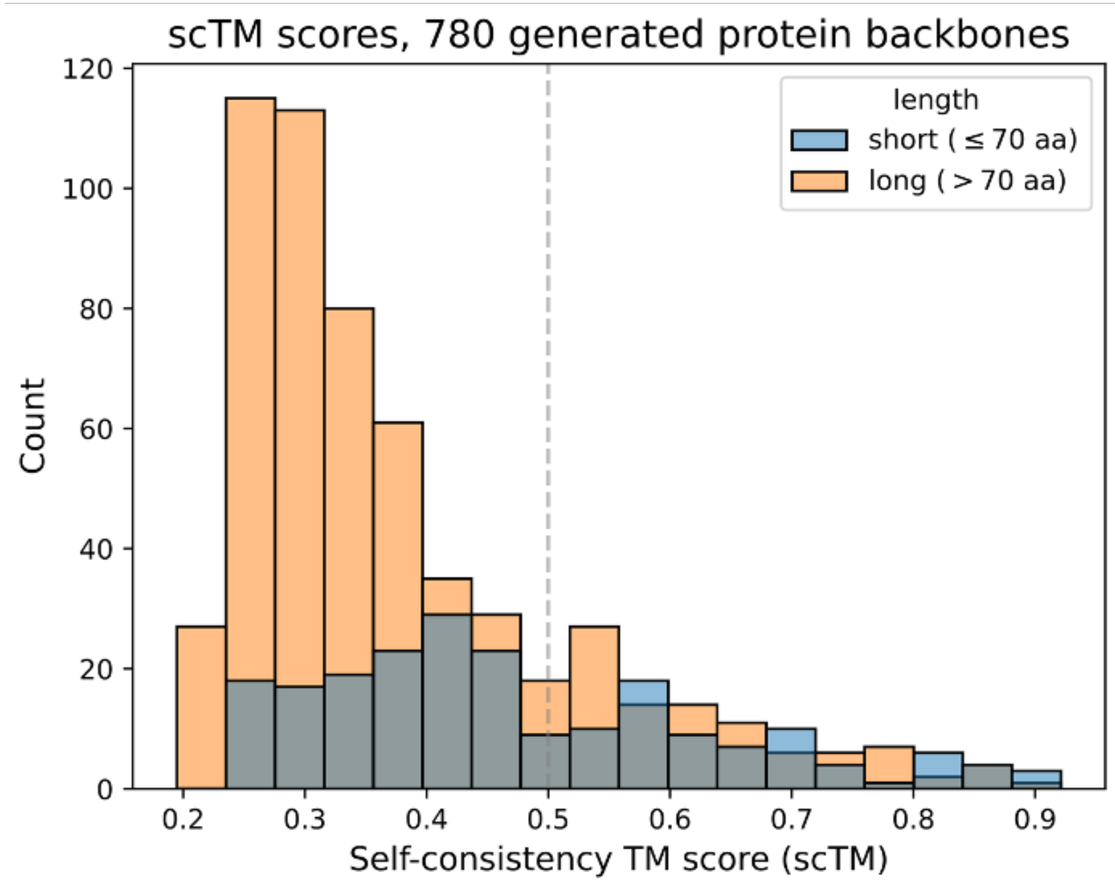


Many generated structures are designable



scTM > 0.5	≤ 70 aa	> 70 aa
FoldingDiff	76/210	87/570
ProtDiff (Trippe *et al.*)	36/210	56/570

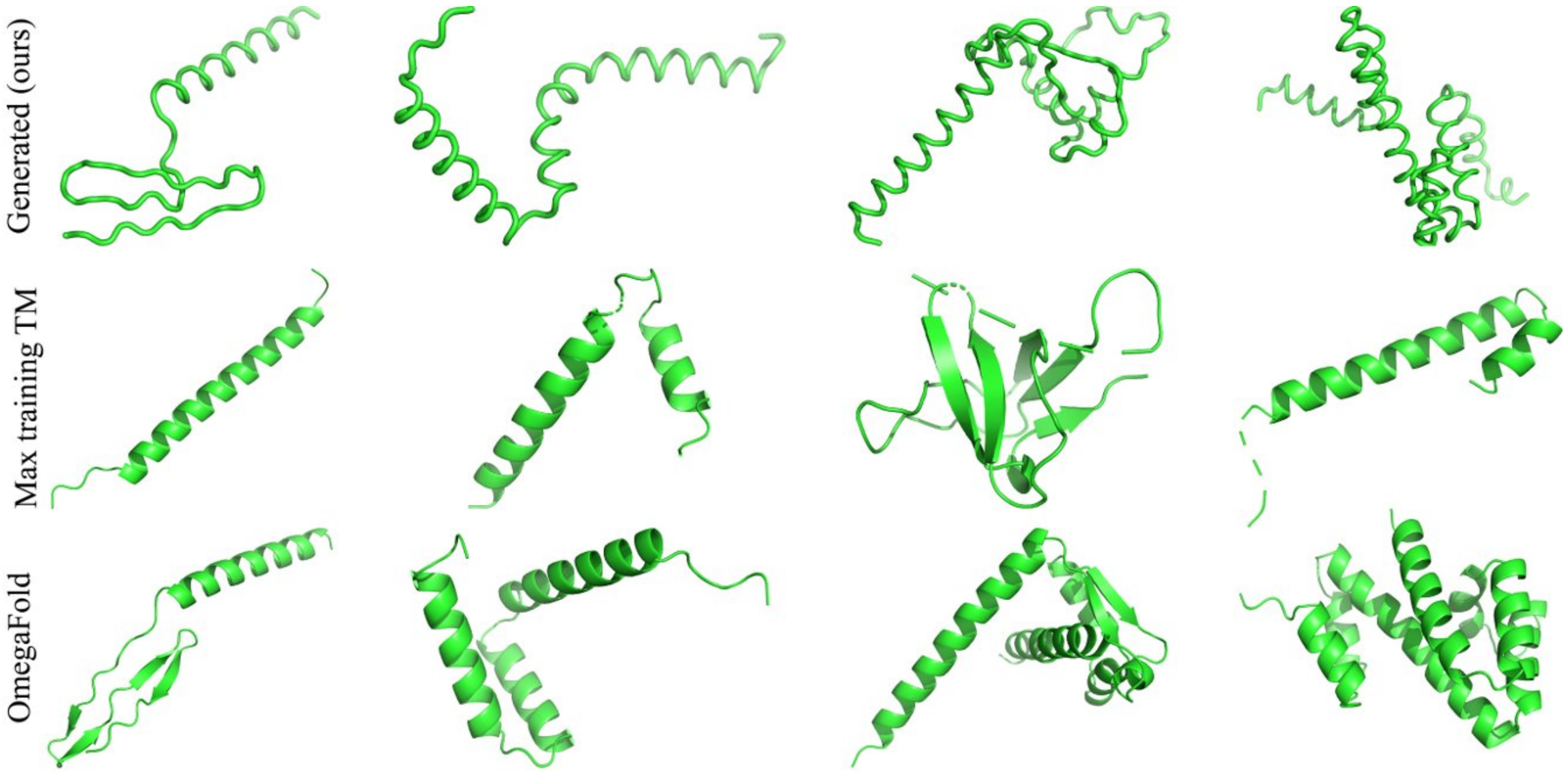
Many generated structures are designable



scTM > 0.5	≤ 70 aa	> 70 aa
FoldingDiff	76/210	87/570
ProtDiff (Trippe *et al.*)	36/210	56/570

Significant improvements over point cloud diffusion model

Generated structures are diverse



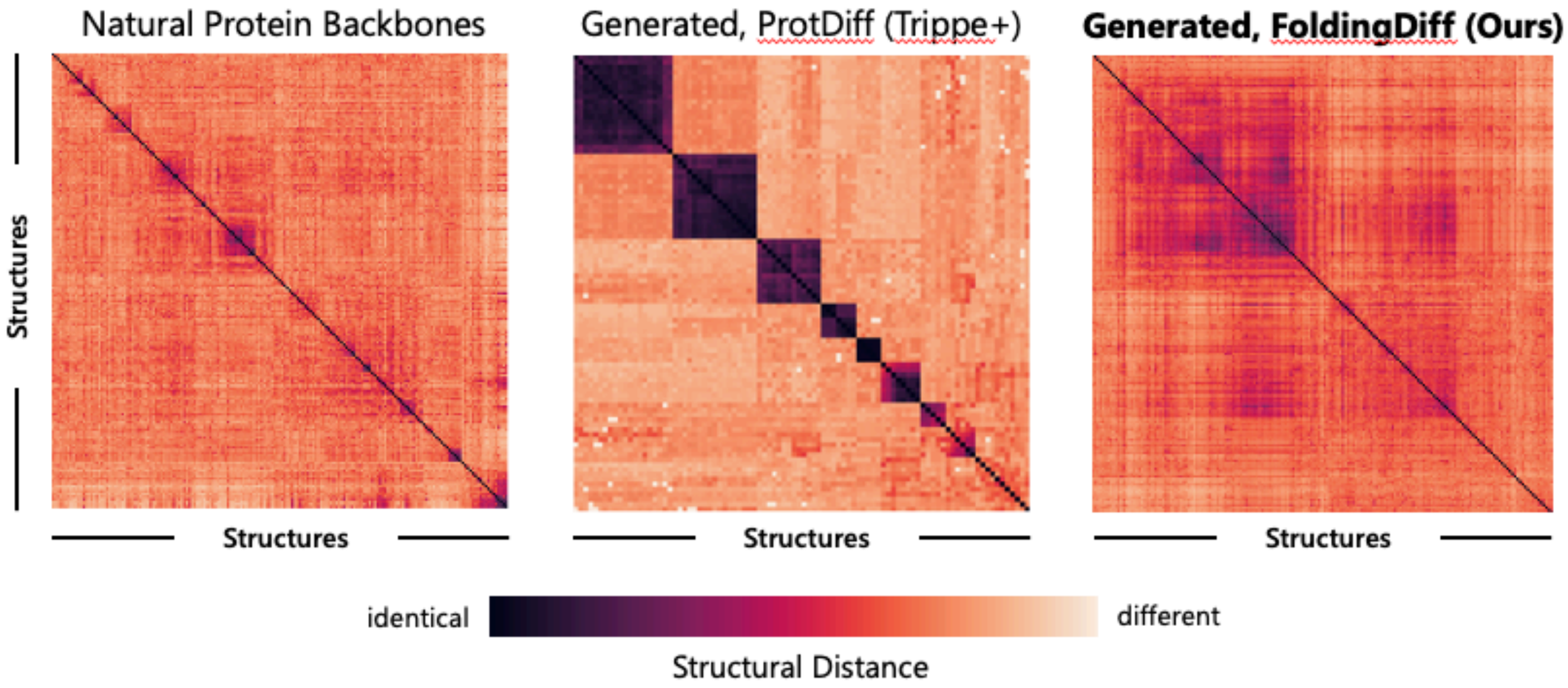
59 residues
scTM = 0.46
max training TM = 0.63

68 residues
scTM = 0.54
max training TM = 0.60

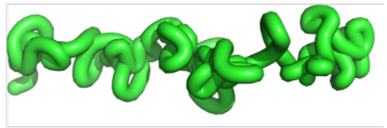
107 residues
scTM = 0.46
max training TM = 0.71

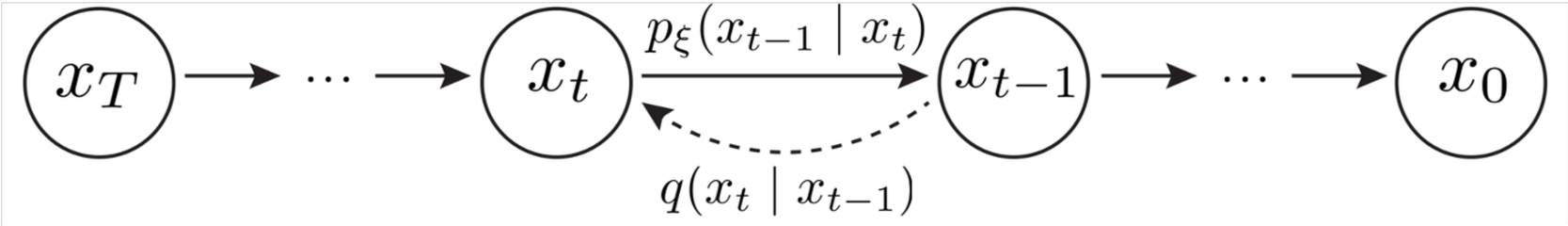
126 residues
scTM = 0.46
max training TM = 0.69

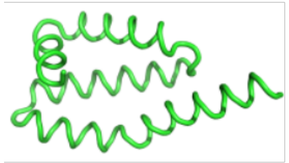
Generated structures are diverse



FoldingDiff is first step towards generating new functions

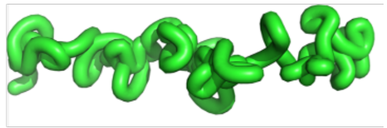


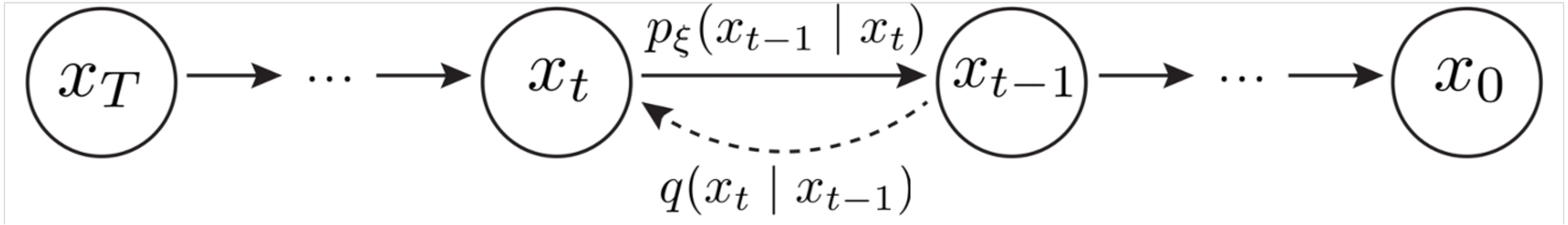


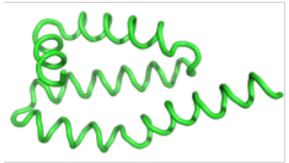


Generate protein backbones by mirroring the folding process

FoldingDiff is first step towards generating new functions

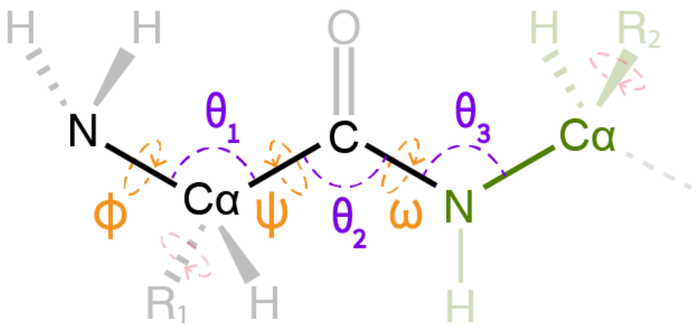




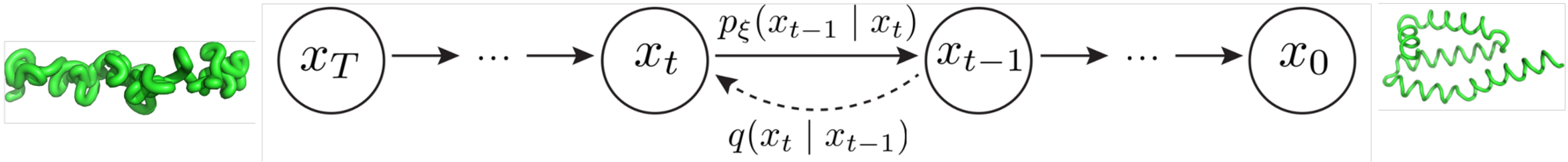


Generate protein backbones by mirroring the folding process

Internal coordinates

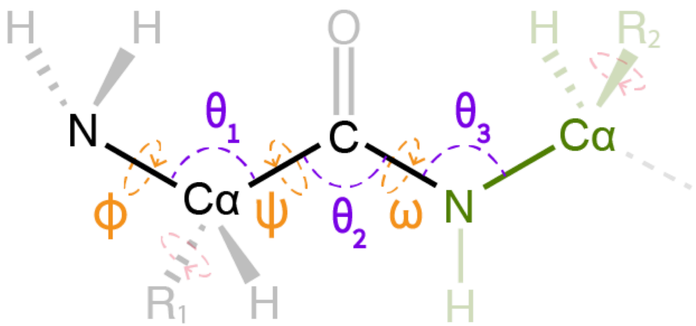


FoldingDiff is first step towards generating new functions

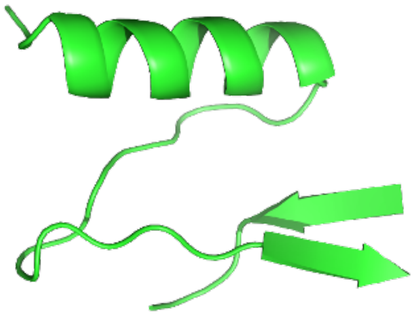


Generate protein backbones by mirroring the folding process

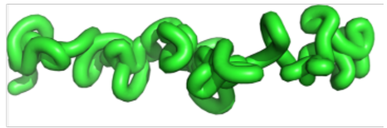
Internal coordinates



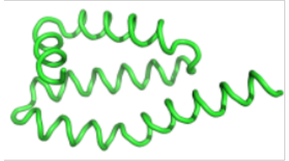
Structural motifs



FoldingDiff is first step towards generating new functions

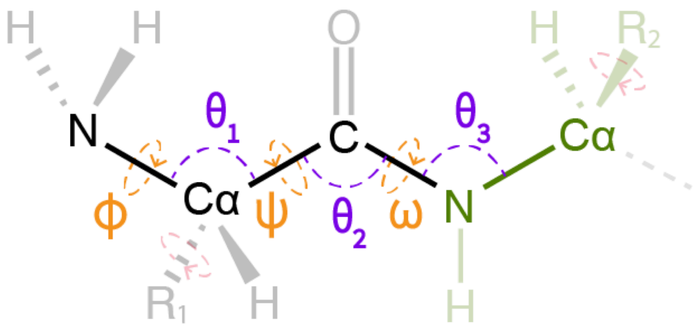


$$x_T \rightarrow \cdots \rightarrow x_t \underset{q(x_t \mid x_{t-1})}{\overset{p_\xi(x_{t-1} \mid x_t)}{\rightleftarrows}} x_{t-1} \rightarrow \cdots \rightarrow x_0$$

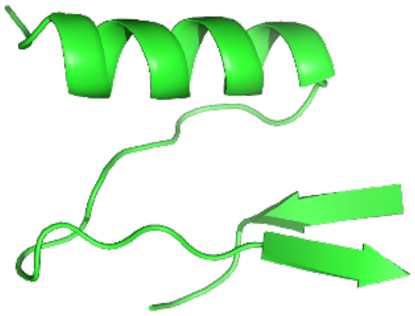


Generate protein backbones by mirroring the folding process

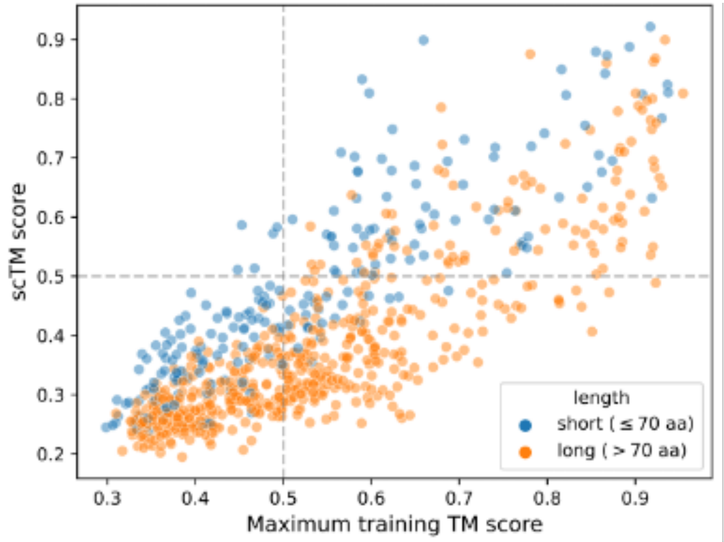
Internal coordinates



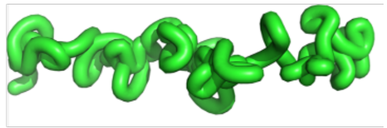
Structural motifs



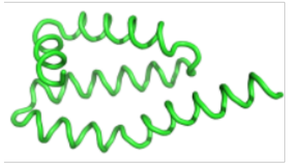
Realistic, diverse samples



FoldingDiff is first step towards generating new functions

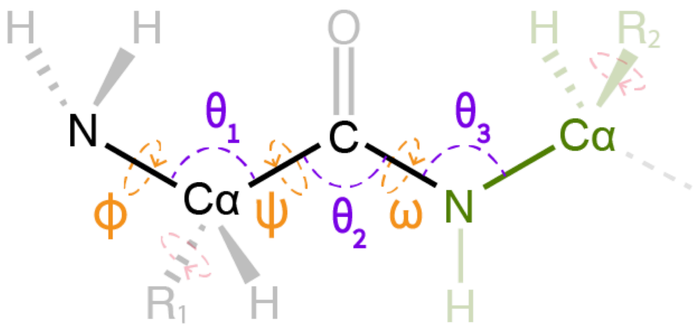


$$x_T \rightarrow \cdots \rightarrow x_t \underset{q(x_t \mid x_{t-1})}{\overset{p_\xi(x_{t-1} \mid x_t)}{\rightleftarrows}} x_{t-1} \rightarrow \cdots \rightarrow x_0$$

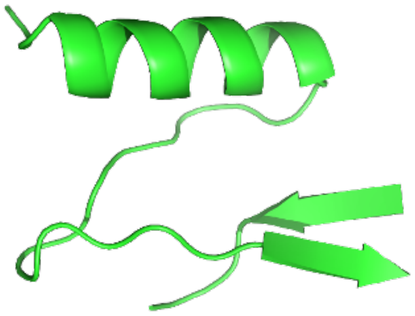


Generate protein backbones by mirroring the folding process

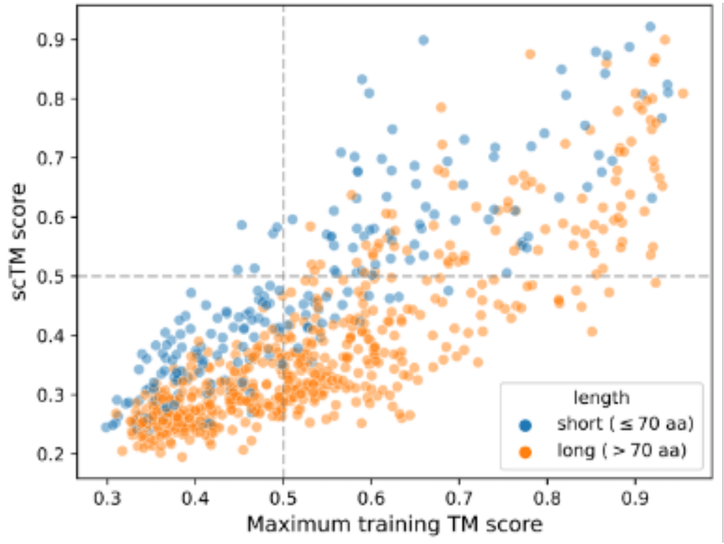
Internal coordinates



Structural motifs

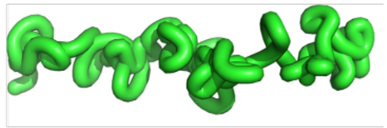


Realistic, diverse samples

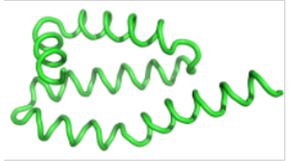


github.com/microsoft/foldingdiff

FoldingDiff is first step towards generating new functions

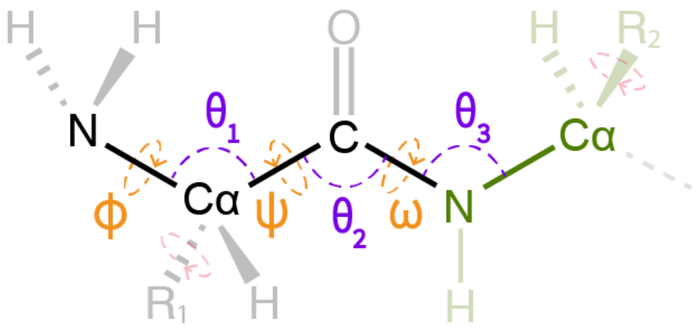


$$x_T \rightarrow \cdots \rightarrow x_t \underset{q(x_t \mid x_{t-1})}{\overset{p_\xi(x_{t-1} \mid x_t)}{\rightleftarrows}} x_{t-1} \rightarrow \cdots \rightarrow x_0$$

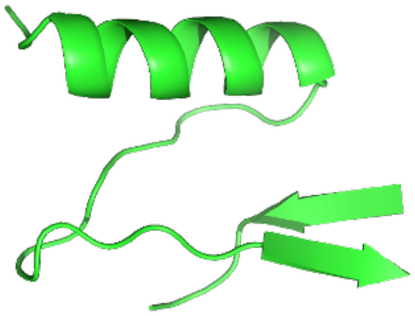


Generate protein backbones by mirroring the folding process

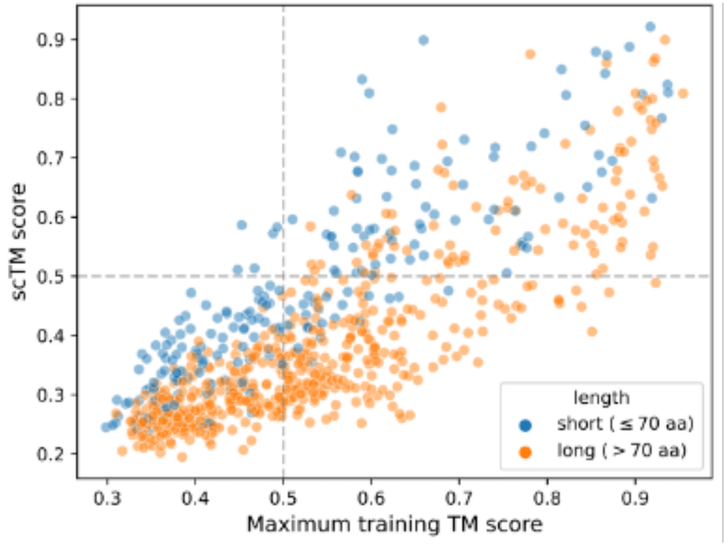
Internal coordinates



Structural motifs



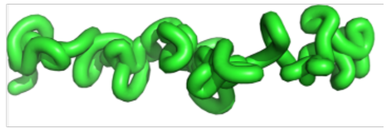
Realistic, diverse samples



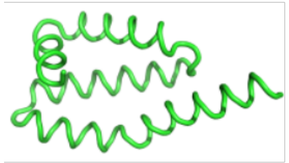
github.com/microsoft/foldingdiff

huggingface.co/spaces/wukevin/foldingdiff

FoldingDiff is first step towards generating new functions

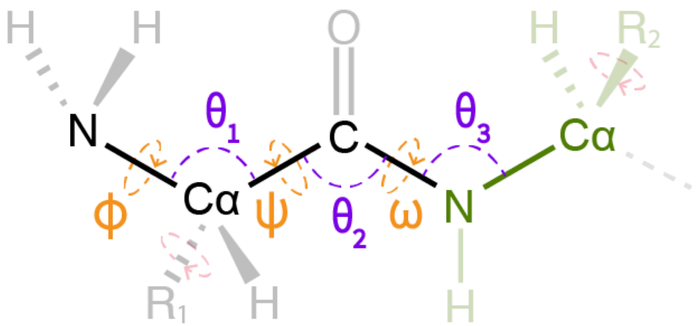


$$x_T \rightarrow \cdots \rightarrow x_t \underset{q(x_t \mid x_{t-1})}{\overset{p_\xi(x_{t-1} \mid x_t)}{\rightleftarrows}} x_{t-1} \rightarrow \cdots \rightarrow x_0$$

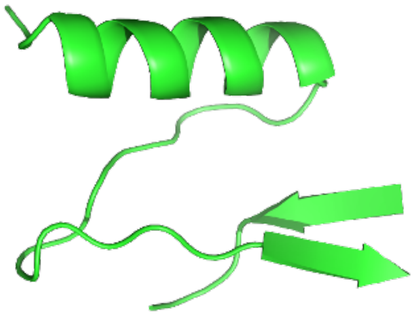


Generate protein backbones by mirroring the folding process

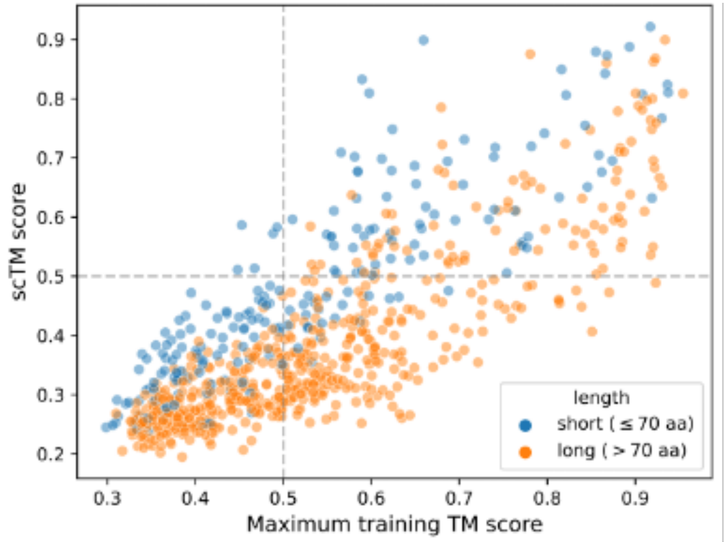
Internal coordinates



Structural motifs



Realistic, diverse samples

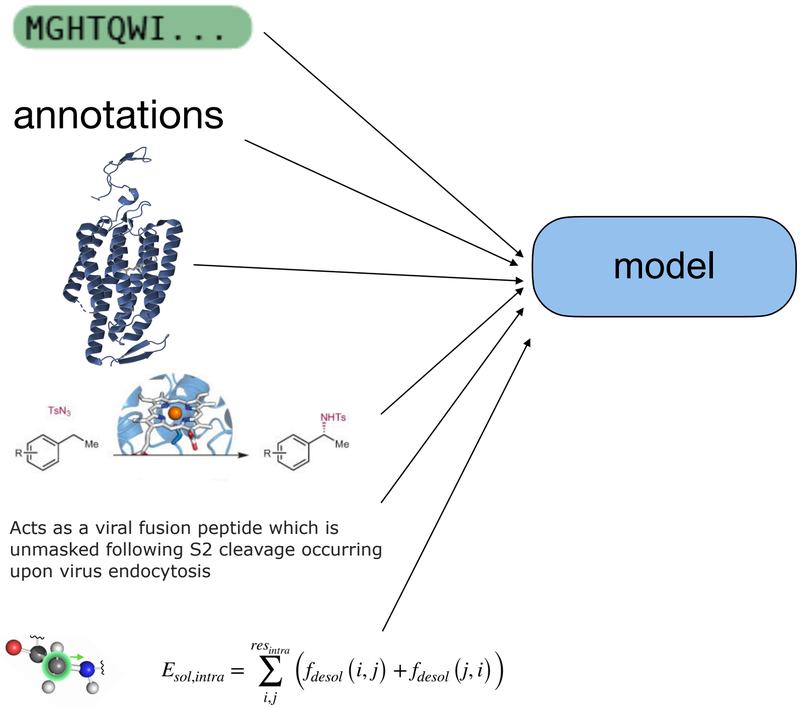


github.com/microsoft/foldingdiff

huggingface.co/spaces/wukevin/foldingdiff

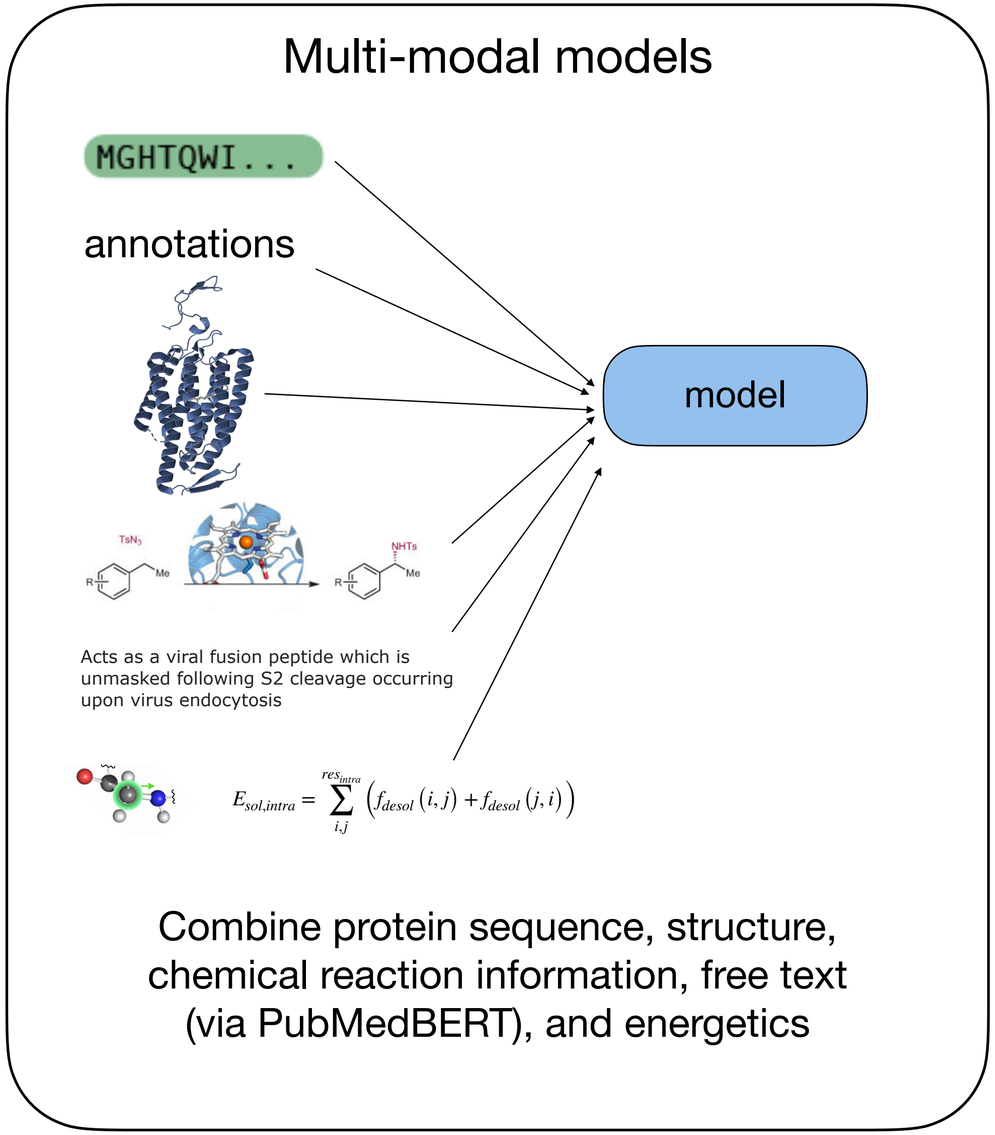
BioNeMo?

Multi-modal models

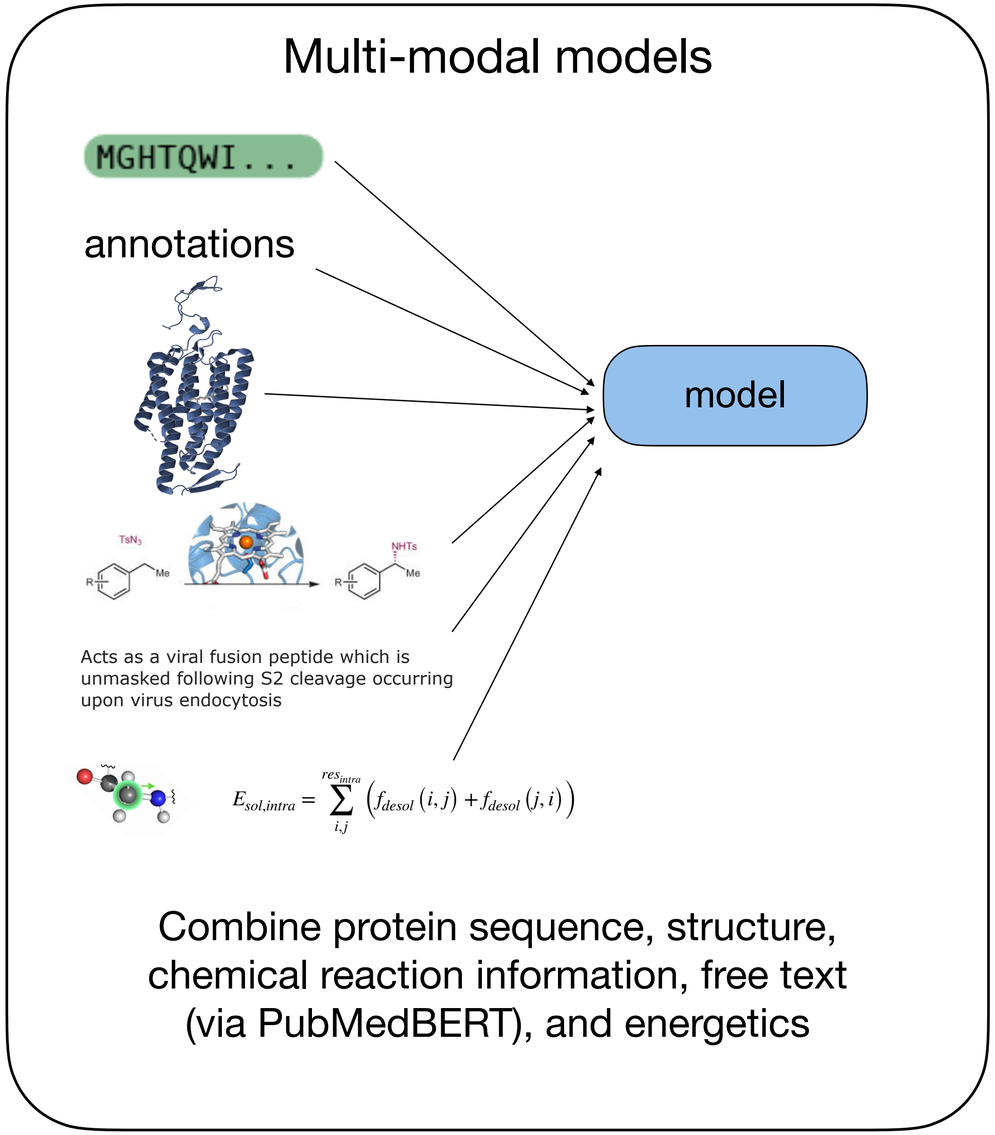


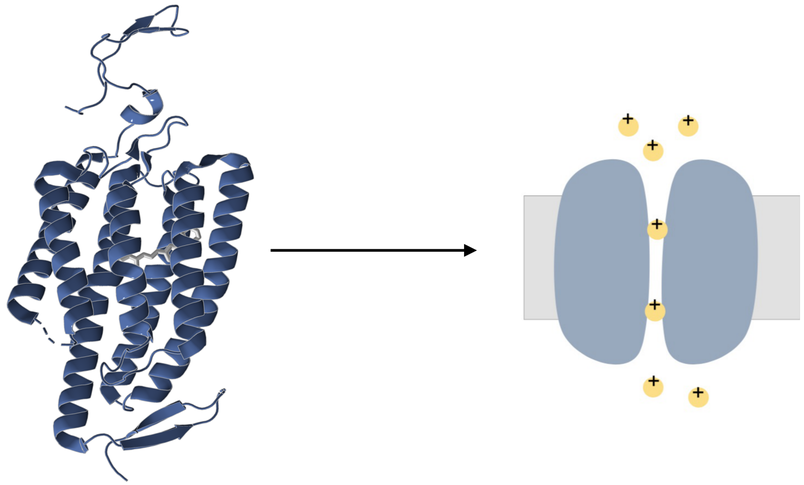
Combine protein sequence, structure, chemical reaction information, free text (via PubMedBERT), and energetics

Can we predict and generate proteins with new functions?



Can we predict and generate proteins with new functions?





Acknowledgments

BioML at MSR New England