Learning protein fitness models from evolutionary and assay-labeled data

Hsu, C., Nisonoff, H., Fannjiang, C. *et al*. Learning protein fitness models from evolutionary and assay-labeled data. *Nat Biotechnol* (2022). https://doi.org/10.1038/s41587-021-01146-5

(CSE 590C WI 22 - Alyssa La Fleur)

UNIVERSITY *of* WASHINGTON

Protein fitness prediction

- **Protein "fitness":** any protein property (stability, enzyme activity, binding strength, etc.)
- **Predicting fitness for protein sequences:** assist with design, potential pathogenicity prediction
  - Pathogenicity prediction task != Fitness prediction task

Protein fitness prediction

- **Protein "fitness":** any protein property (stability, enzyme activity, binding strength, etc.)
- **Predicting fitness for protein sequences:** assist with design, potential pathogenicity prediction
  - Pathogenicity prediction task != Fitness prediction task

This paper evaluates existing fitness prediction methods, and proposes a new one

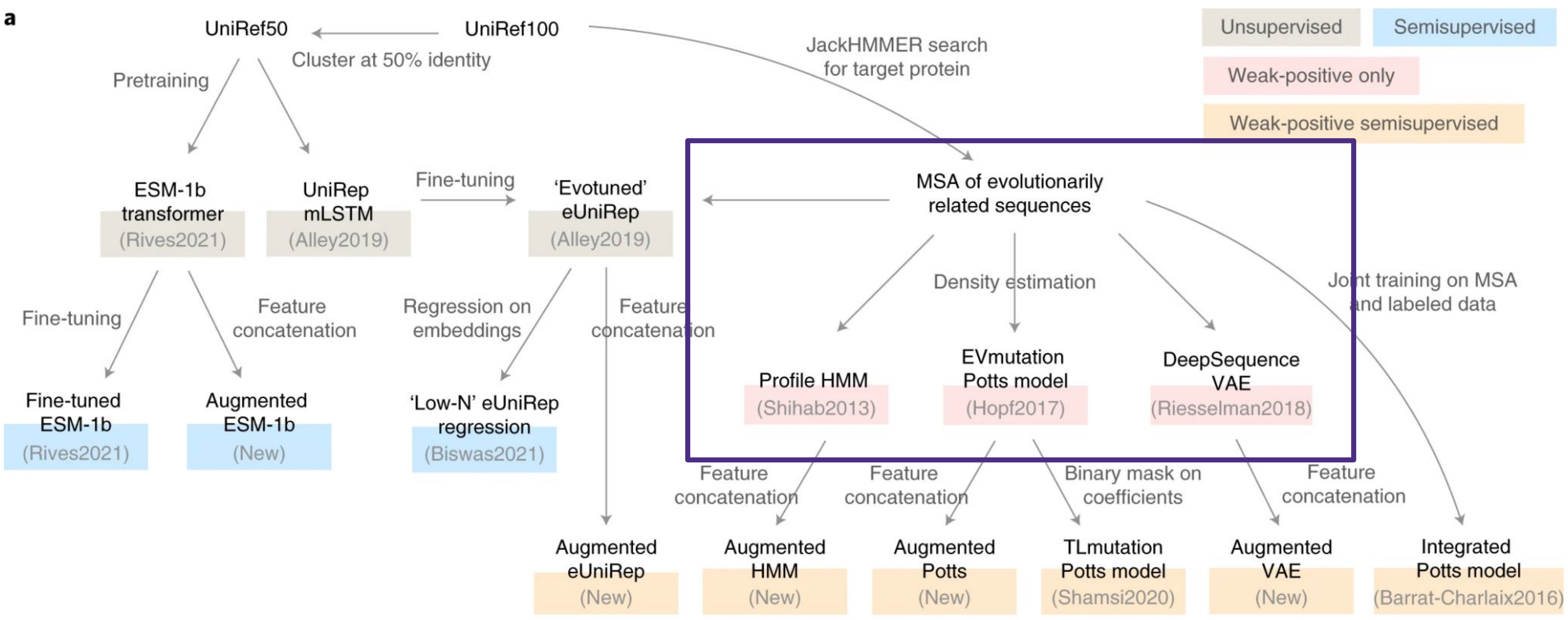
Two main ML strategies

1. **Evolutionary models**
    1. Get a sequence alignment for your target protein
    2. Model the probability density of these sequences
    3. Predict mutant fitness using the probability density model

Two main ML strategies

1. **Evolutionary models**
   a. Get a sequence alignment for your target protein
   b. Model the probability density of these sequences
   c. Predict mutant fitness using the probability density model

'Weak positive' learning - these approaches assume that evolutionary related sequences have similar functions to the target



a

UniRef50

UniRef100

Cluster at 50% identity

Pretraining

JackHMMER search for target protein

Unsupervised

Semisupervised

Weak-positive only

Weak-positive semisupervised

ESM-1b transformer (Rives2021)

UniRep mLSTM (Alley2019)

Fine-tuning

'Evotuned' eUniRep (Alley2019)

MSA of evolutionarily related sequences

Density estimation

Joint training on MSA and labeled data

Fine-tuning

Feature concatenation

Regression on embeddings

Feature concatenation

Profile HMM (Shihab2013)

EVmutation Potts model (Hopf2017)

DeepSequence VAE (Riesselman2018)

Fine-tuned ESM-1b (Rives2021)

Augmented ESM-1b (New)

'Low-N' eUniRep regression (Biswas2021)

Feature concatenation

Feature concatenation

Binary mask on coefficients

Feature concatenation

Augmented eUniRep (New)

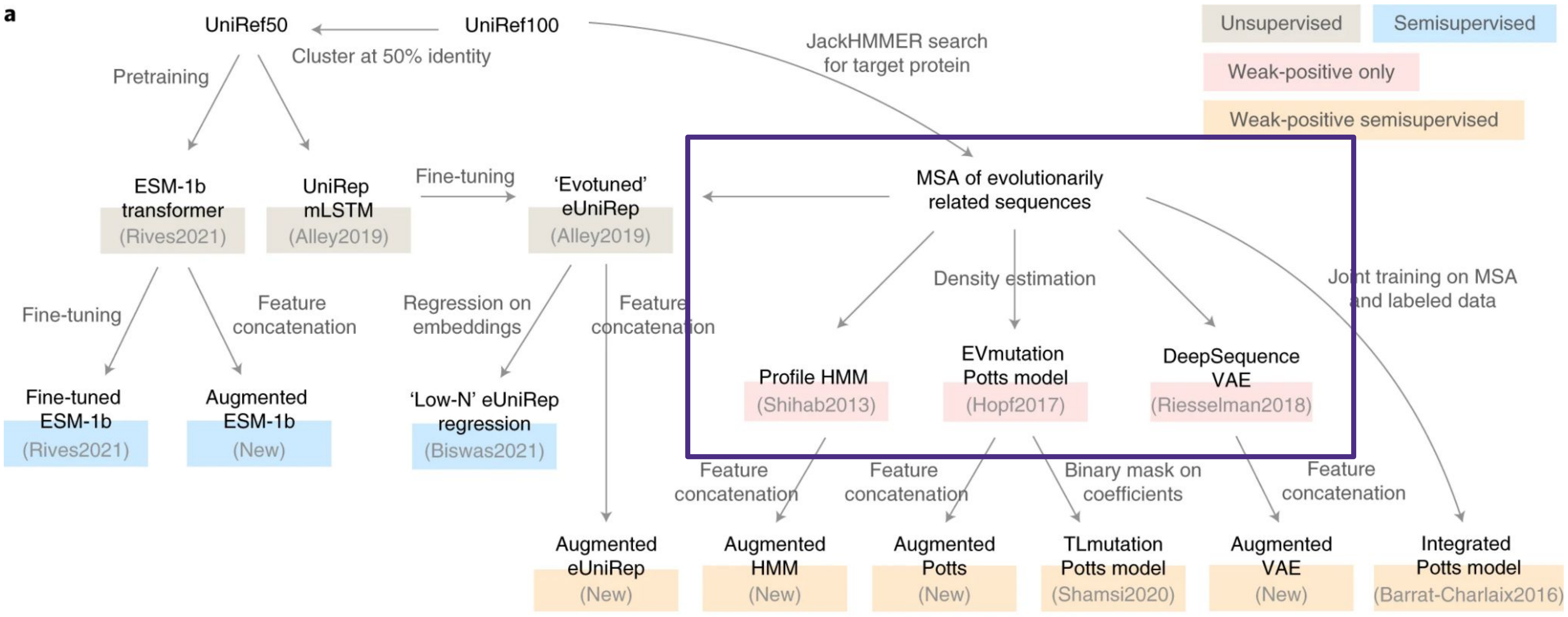
Augmented HMM (New)

Augmented Potts (New)

TLmutation Potts model (Shamsi2020)

Augmented VAE (New)

Integrated Potts model (Barrat-Charlaix2016)



a

Unsupervised

Semisupervised

Weak-positive only

Weak-positive semisupervised

UniRef50

UniRef100

Cluster at 50% identity

Pretraining

JackHMMER search for target protein

ESM-1b transformer (Rives2021)

UniRep mLSTM (Alley2019)

Fine-tuning

'Evotuned' eUniRep (Alley2019)

MSA of evolutionarily related sequences

Fine-tuning

Feature concatenation

Regression on embeddings

Feature concatenation

Density estimation

Joint training on MSA and labeled data

Fine-tuned ESM-1b (Rives2021)

Augmented ESM-1b (New)

'Low-N' eUniRep regression (Biswas2021)

Profile HMM (Shihab2013)

EVmutation Potts model (Hopf2017)

DeepSequence VAE (Riesselman2018)

Feature concatenation

Feature concatenation

Binary mask on coefficients

Feature concatenation

Augmented eUniRep (New)

Augmented HMM (New)

Augmented Potts (New)

TLmutation Potts model (Shamsi2020)

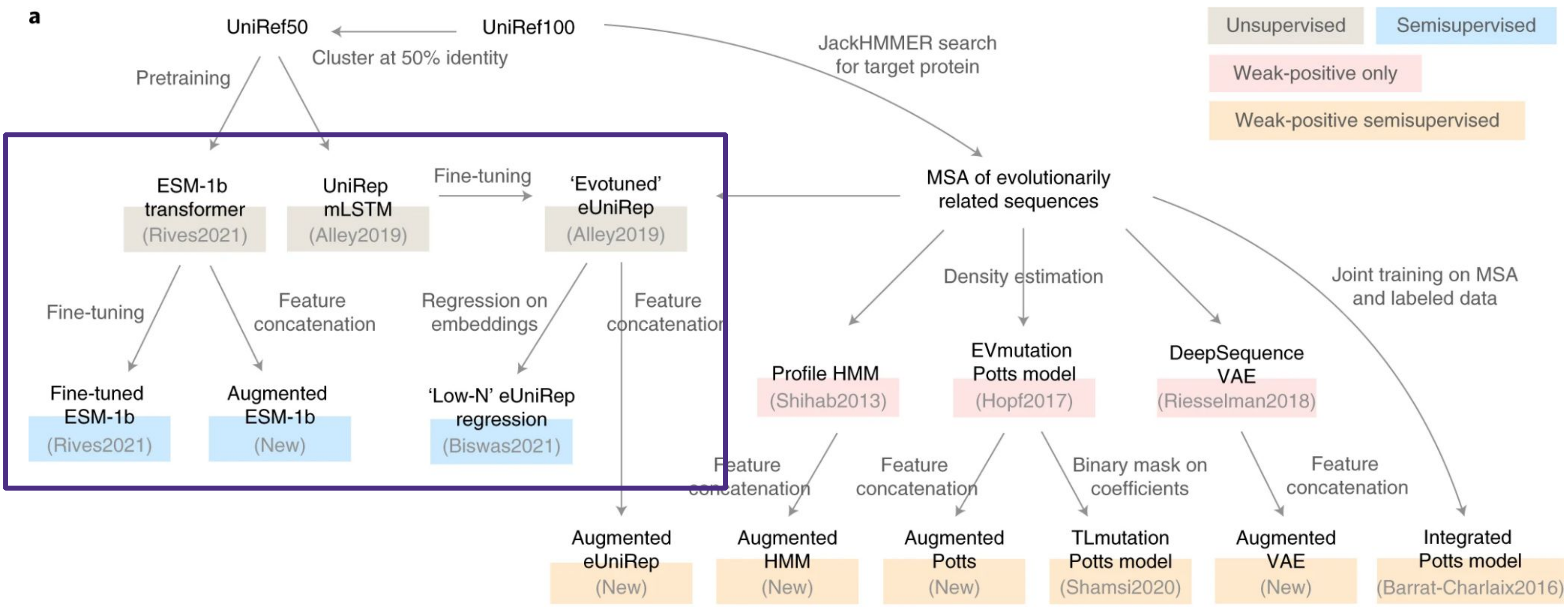
Augmented VAE (New)

Integrated Potts model (Barrat-Charlaix2016)

One limitation is that alignment depth may vary - some targets may only have hundreds of usable sequences in their alignment

Two main ML strategies

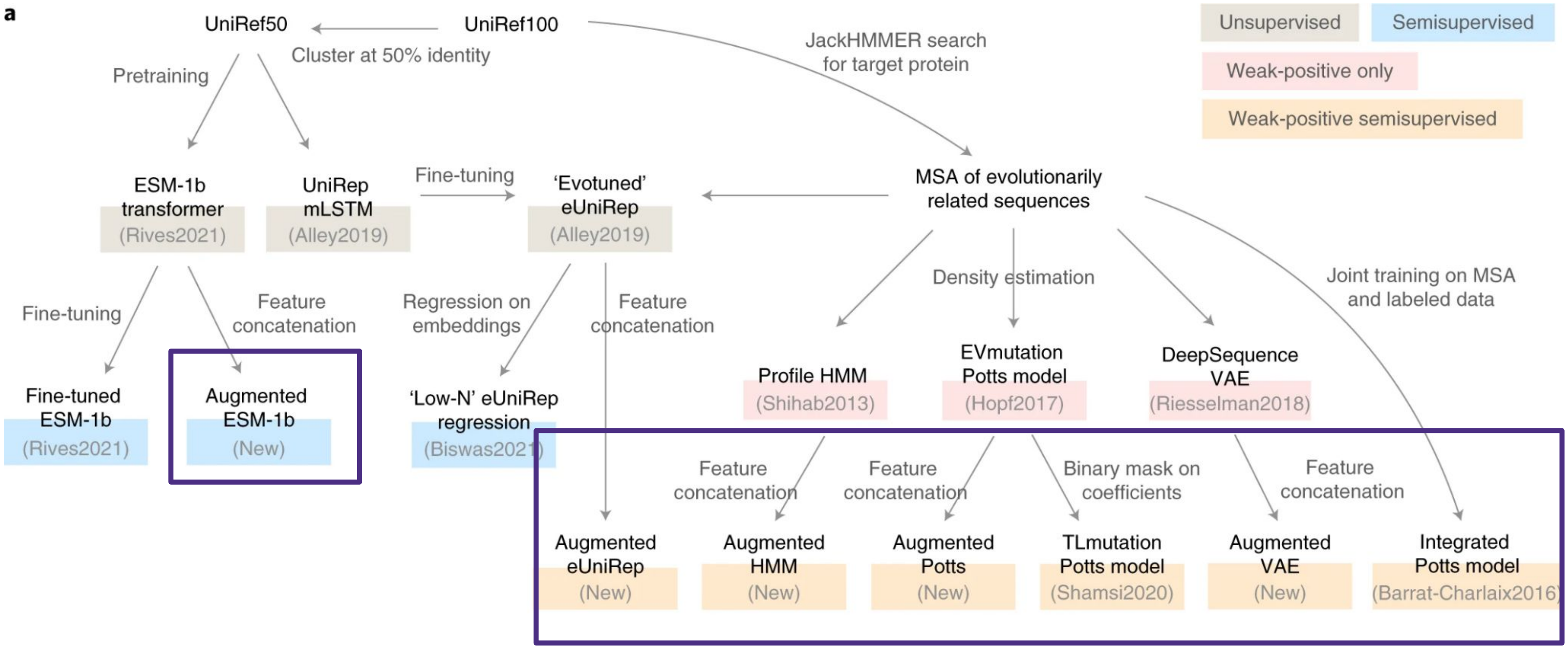
2. **Supervised regression models**
    a. Models range from simple (linear regression) to complex (CNN, LSTM, Transformers, etc.)
    b. **Semi-supervised:** Supervised regression models can also be trained using unsupervised NLP model protein representations



Can be limited by number of mutants in training set, coverage of positions by mutation (few positions vs. many)

A combined strategy

- **Weak-positive semi-supervised learning:** learning a distribution of sequences using alignments, with supervised learning on labelled sequences
- **Their 'baseline' augmentation combined approach (Had max performance in 15/19 test sets)**



a
Unsupervised
Semisupervised
Weak-positive only
Weak-positive semisupervised
UniRef50
UniRef100
Cluster at 50% identity
JackHMMER search for target protein
Pretraining
ESM-1b transformer (Rives2021)
UniRep mLSTM (Alley2019)
Fine-tuning
'Evotuned' eUniRep (Alley2019)
MSA of evolutionarily related sequences
Fine-tuning
Feature concatenation
Regression on embeddings
Feature concatenation
Density estimation
Joint training on MSA and labeled data
Fine-tuned ESM-1b (Rives2021)
Augmented ESM-1b (New)
'Low-N' eUniRep regression (Biswas2021)
Profile HMM (Shihab2013)
EVmutation Potts model (Hopf2017)
DeepSequence VAE (Riesselman2018)
Feature concatenation
Feature concatenation
Binary mask on coefficients
Feature concatenation
Augmented eUniRep (New)
Augmented HMM (New)
Augmented Potts (New)
TLmutation Potts model (Shamsi2020)
Augmented VAE (New)
Integrated Potts model (Barrat-Charlaix2016)

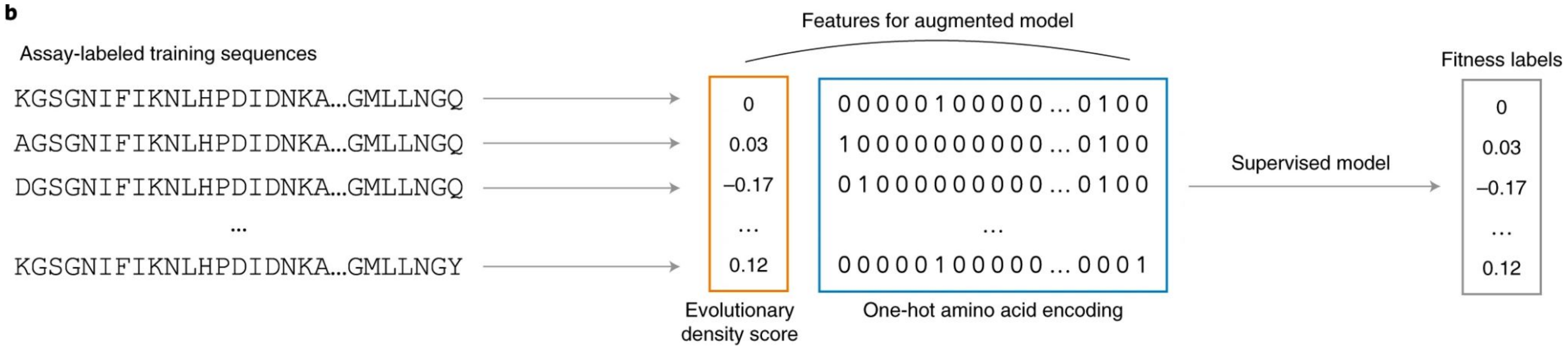
Can be limited by number of mutants in training set, coverage of positions by mutation (few positions vs. many)

Potts models

- TLMutation Potts models SUMMARY GOES HERE

Augmentation combined approach

- Sequence log-likelihoods from a sequence density model & one-hot encoded protein sequences
- Supervised model is ridge regression (L2)


b
Assay-labeled training sequences
KGSGNIFIKNLHPDIDNKA...GMLLNGQ
AGSGNIFIKNLHPDIDNKA...GMLLNGQ
DGSGNIFIKNLHPDIDNKA...GMLLNGQ
...
KGSGNIFIKNLHPDIDNKA...GMLLNGY
Features for augmented model
0
0.03
–0.17
...
0.12
Evolutionary density score
000001000000...0100
100000000000...0100
010000000000...0100
...
000001000000...0001
One-hot amino acid encoding
Supervised model
Fitness labels
0
0.03
–0.17
...
0.12

Deep mutation scanning (DMS) datasets

- They used 19 of the DMS datasets from EVMutation (one of the competitor models they compared against) + a GFP fluorescence data set
- All had mutations throughout a domain or whole protein
- 16/19 had sequences one missense mutation away from WT (single mutants)
- Only evaluated mutants at positions with < 30% gaps in the MSAs generated

Dataset splits

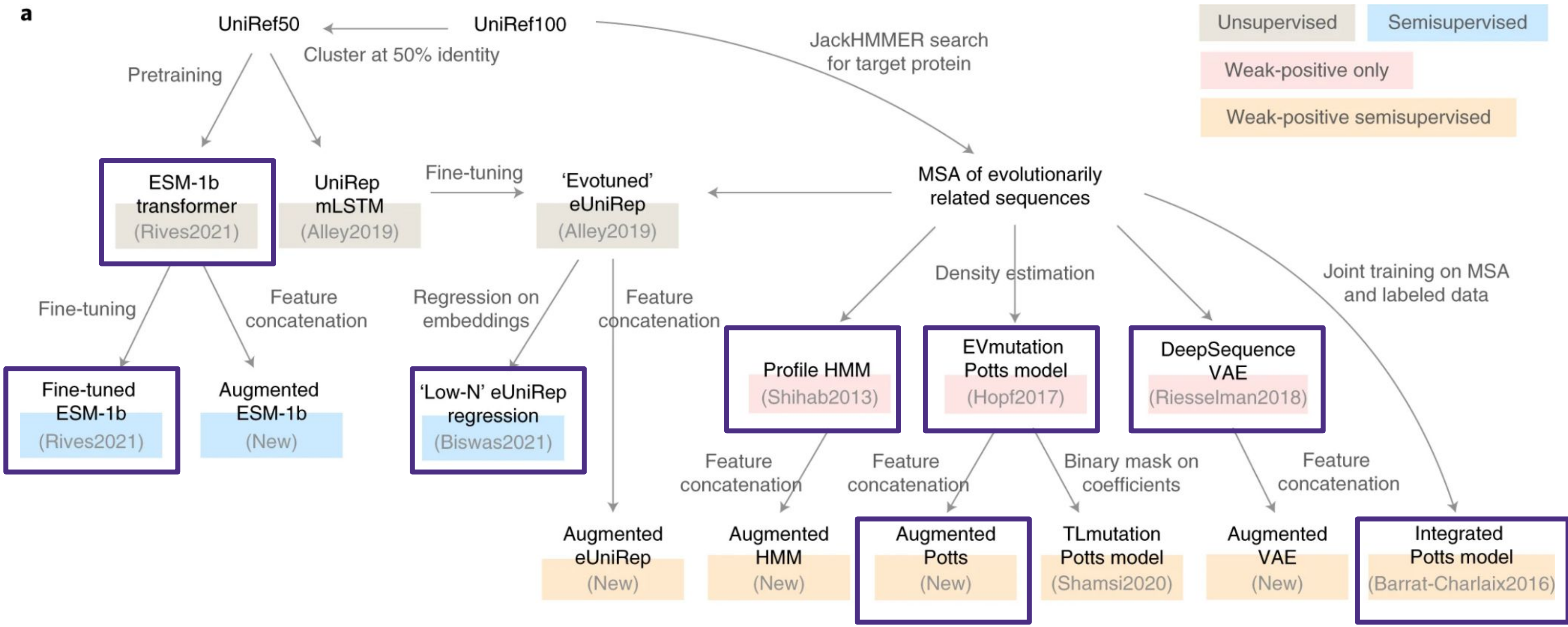
- **20% test**, varying sizes of training sets
- **80/20 train/test**, five fold cross-validation on the 80% where computationally feasible (For comparing to methods like TLMutation, ESM1-b)

20 random seems were used for data partitioning for each approach

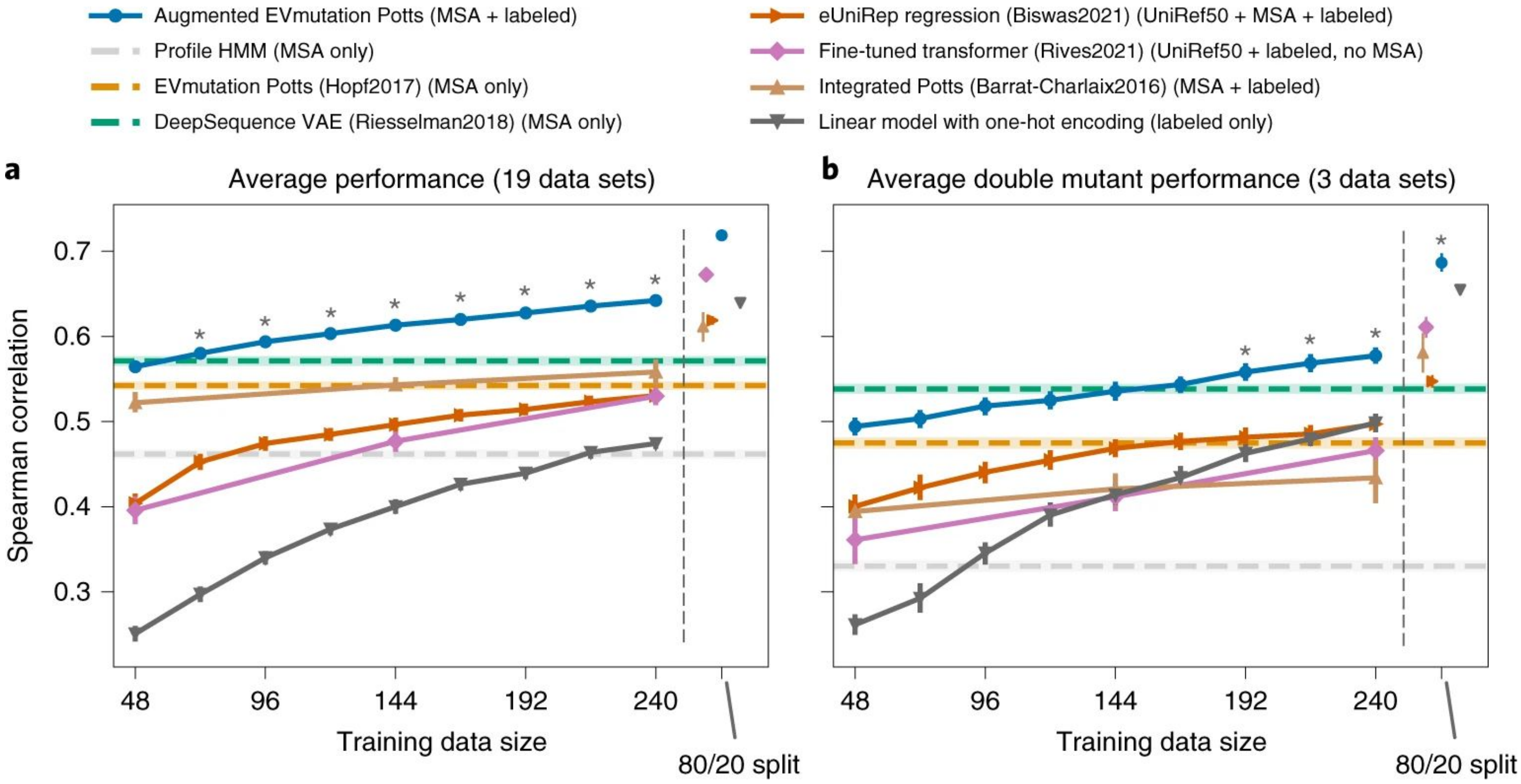
Ranking metrics

- **Spearman rank correlation coefficient:**
- **Normalized discounted cumulative gain (NDCG):** From information retrieval, similar to a weighted Spearman rank which focuses on high value agreement

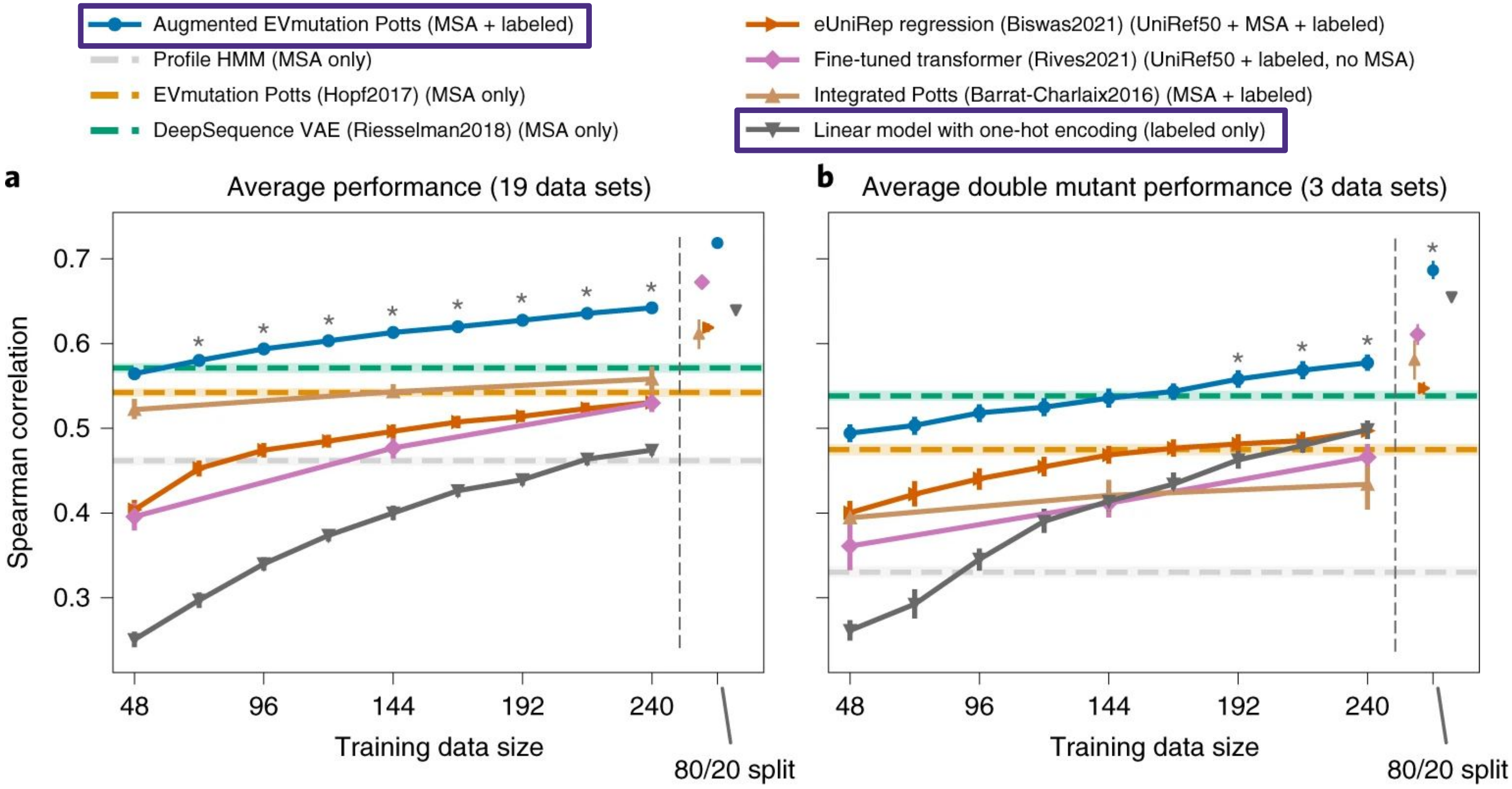
One hot linear model


a
UniRef50
UniRef100
Cluster at 50% identity
Pretraining
JackHMMER search for target protein
Unsupervised
Semisupervised
Weak-positive only
Weak-positive semisupervised
ESM-1b transformer (Rives2021)
UniRep mLSTM (Alley2019)
Fine-tuning
'Evotuned' eUniRep (Alley2019)
MSA of evolutionarily related sequences
Fine-tuning
Feature concatenation
Regression on embeddings
Feature concatenation
Density estimation
Joint training on MSA and labeled data
Fine-tuned ESM-1b (Rives2021)
Augmented ESM-1b (New)
'Low-N' eUniRep regression (Biswas2021)
Profile HMM (Shihab2013)
EVmutation Potts model (Hopf2017)
DeepSequence VAE (Riesselman2018)
Feature concatenation
Feature concatenation
Binary mask on coefficients
Feature concatenation
Augmented eUniRep (New)
Augmented HMM (New)
Augmented Potts (New)
TLmutation Potts model (Shamsi2020)
Augmented VAE (New)
Integrated Potts model (Barrat-Charlaix2016)

Low-N Training Predictions

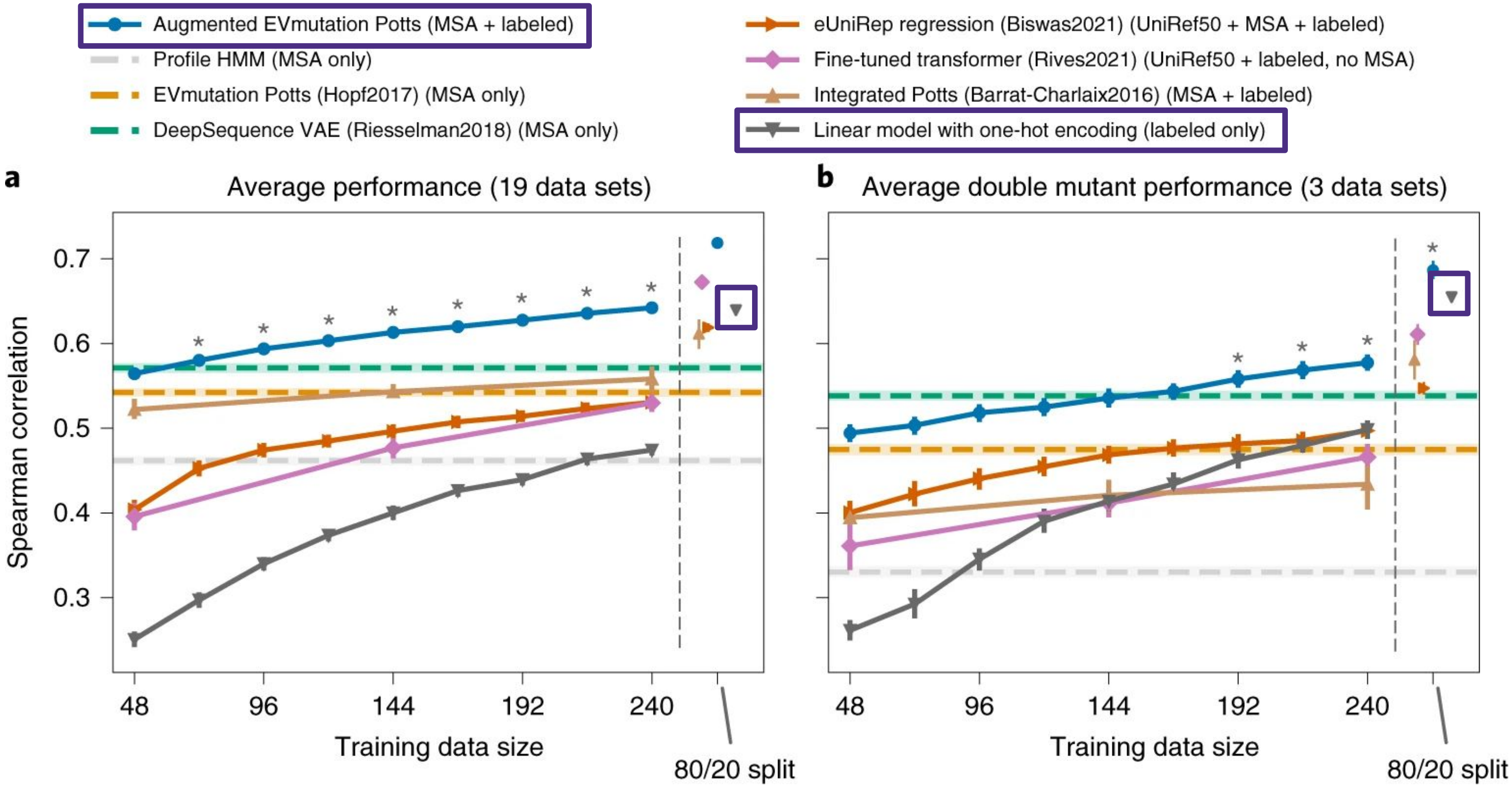


Low-N Training Predictions



All data

Low-N Training Predictions

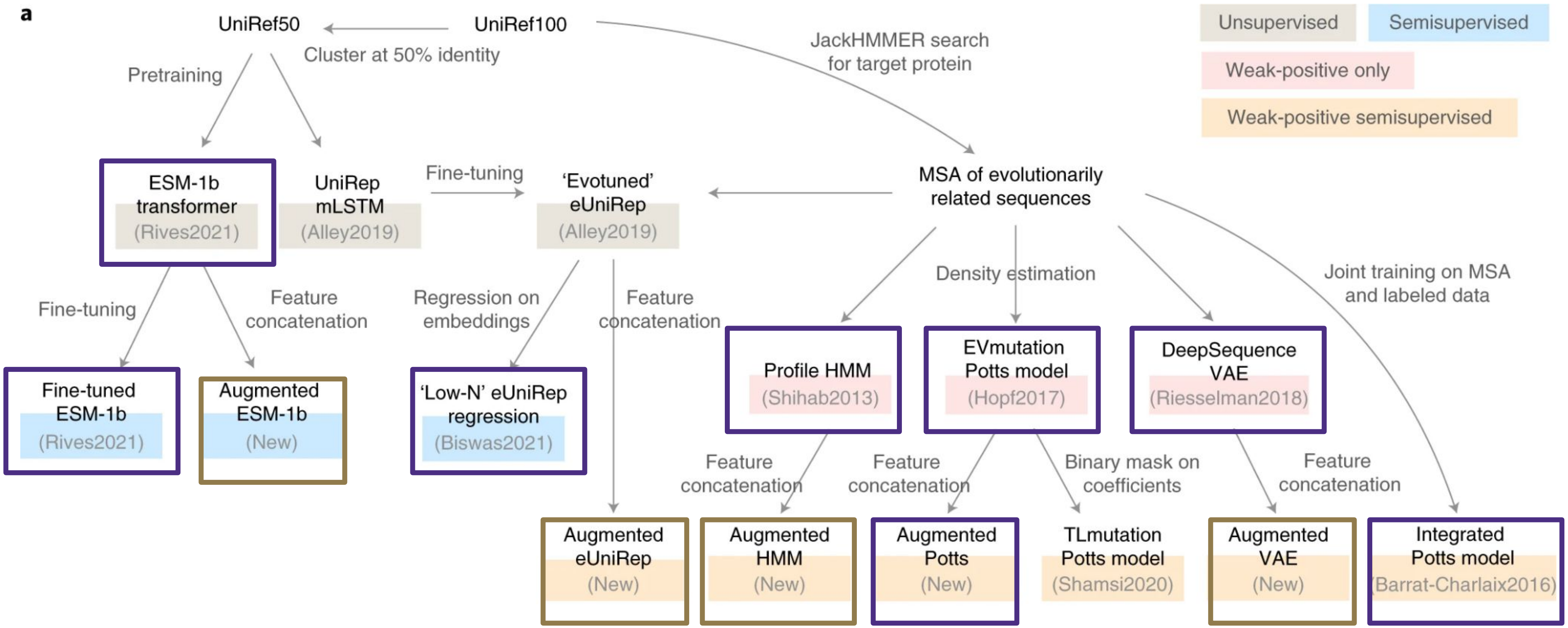


All data

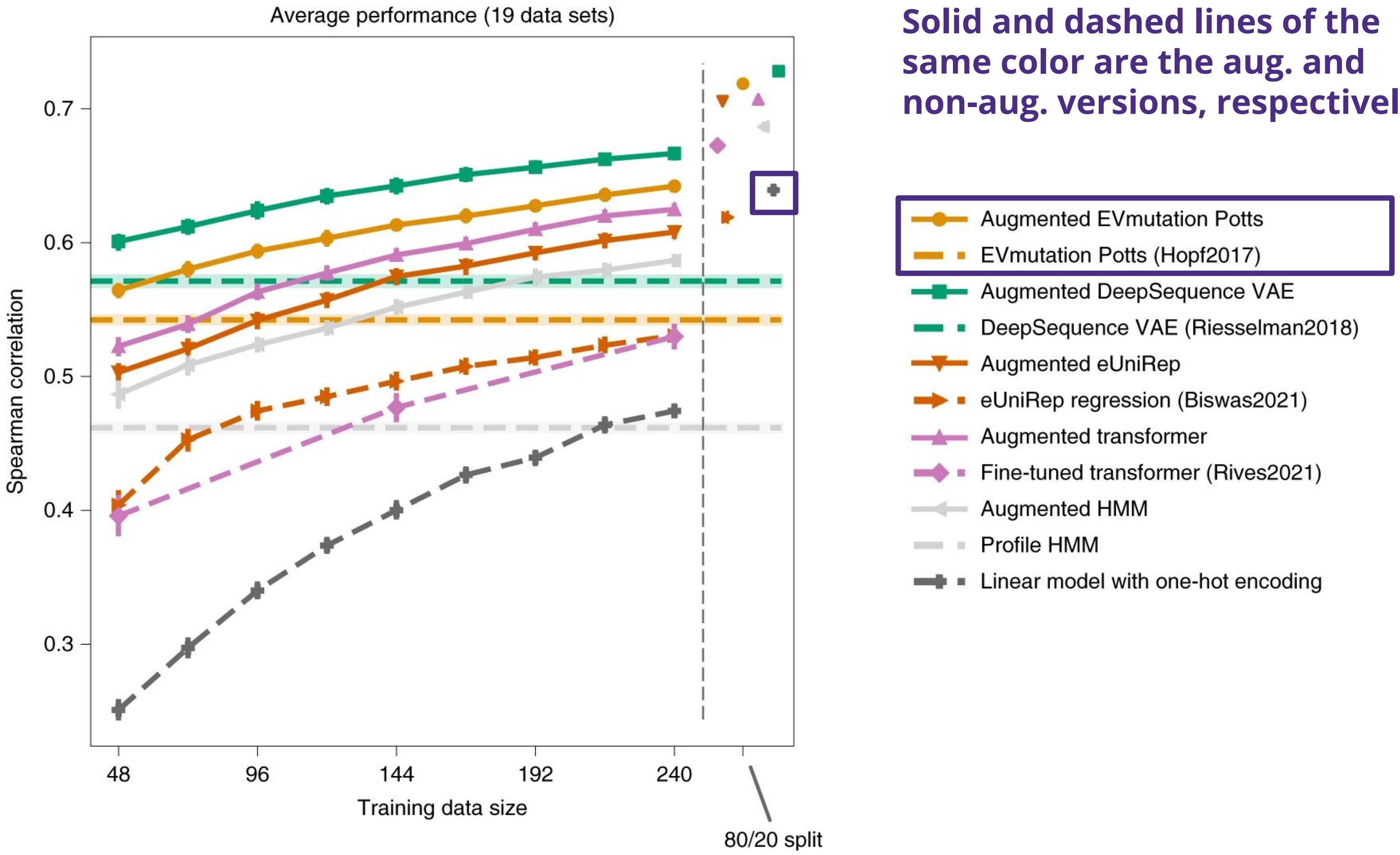
Additional augmented models

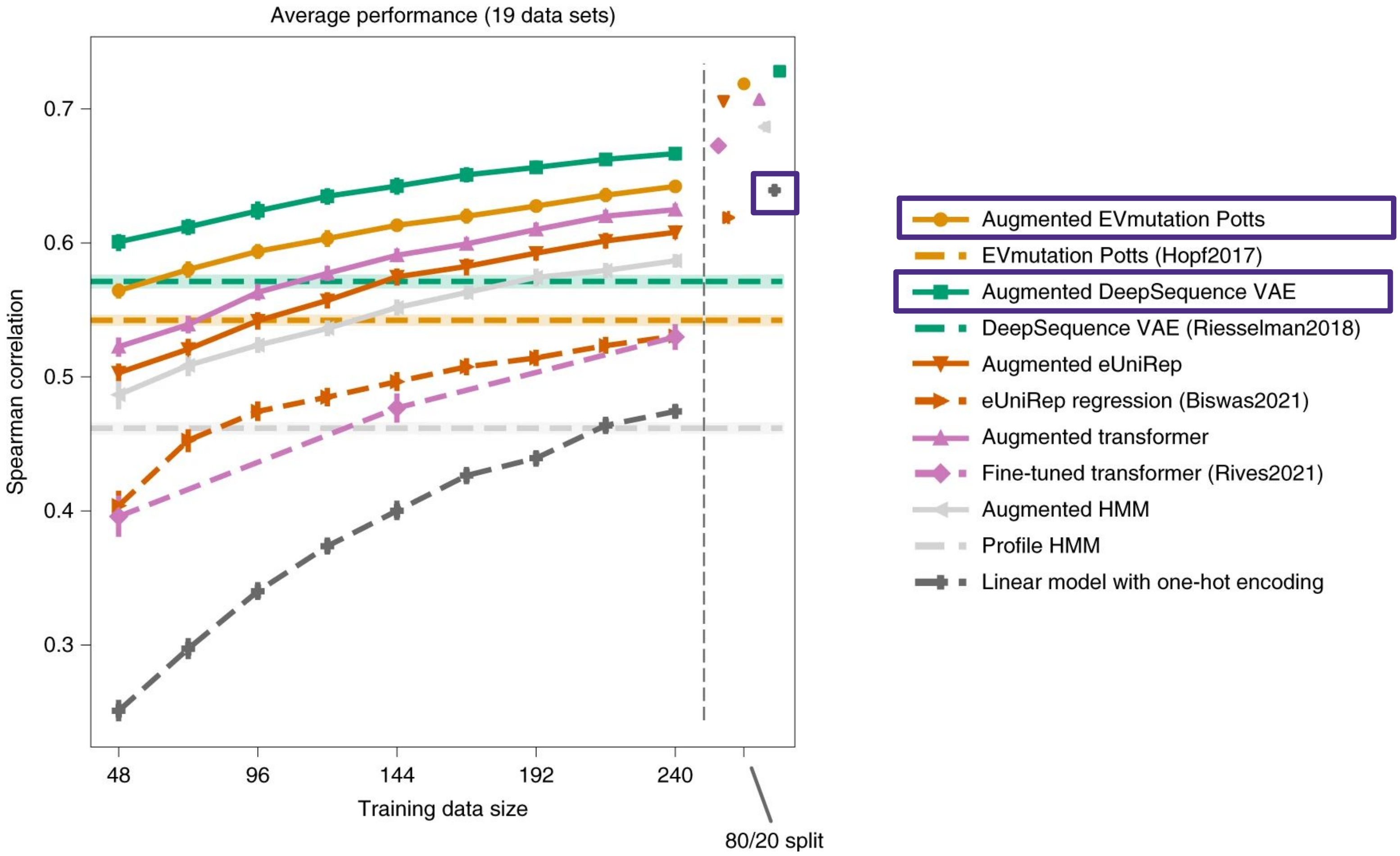
- Augmented other models than just Potts
- Note that the transformer (not eUniRep) is the only method not using any evolutionary data
- Augmented models outperformed non-augmented model, regardless of training set size

One hot linear model

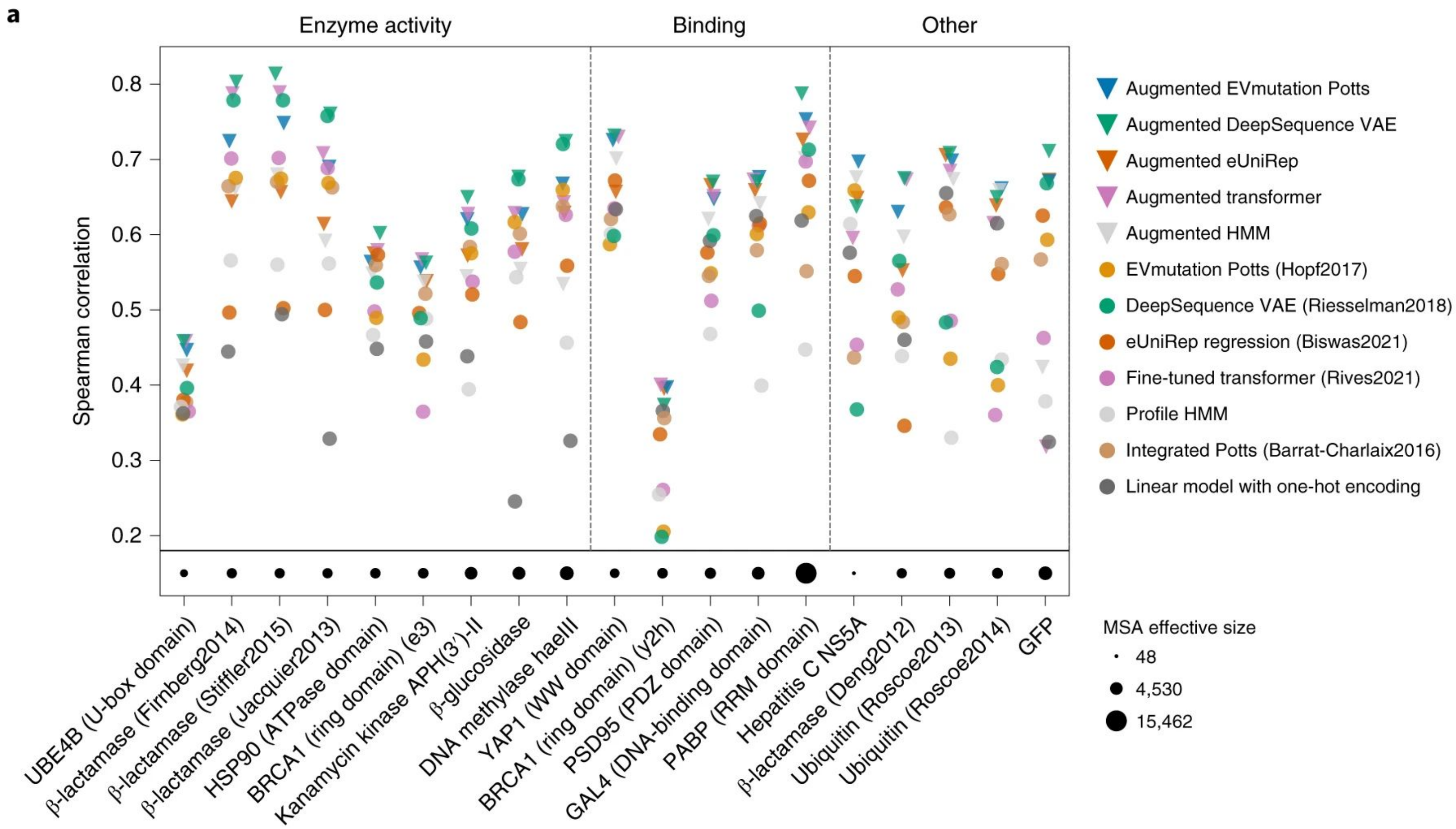

a
UniRef50
UniRef100
Cluster at 50% identity
JackHMMER search for target protein
Unsupervised
Semisupervised
Weak-positive only
Weak-positive semisupervised
Pretraining
ESM-1b transformer (Rives2021)
UniRep mLSTM (Alley2019)
Fine-tuning
'Evotuned' eUniRep (Alley2019)
MSA of evolutionarily related sequences
Fine-tuning
Feature concatenation
Regression on embeddings
Feature concatenation
Density estimation
Joint training on MSA and labeled data
Fine-tuned ESM-1b (Rives2021)
Augmented ESM-1b (New)
'Low-N' eUniRep regression (Biswas2021)
Profile HMM (Shihab2013)
EVmutation Potts model (Hopf2017)
DeepSequence VAE (Riesselman2018)
Feature concatenation
Feature concatenation
Binary mask on coefficients
Feature concatenation
Augmented eUniRep (New)
Augmented HMM (New)
Augmented Potts (New)
TLmutation Potts model (Shamsi2020)
Augmented VAE (New)
Integrated Potts model (Barrat-Charlaix2016)

Solid and dashed lines of the same color are the aug. and non-aug. versions, respectively



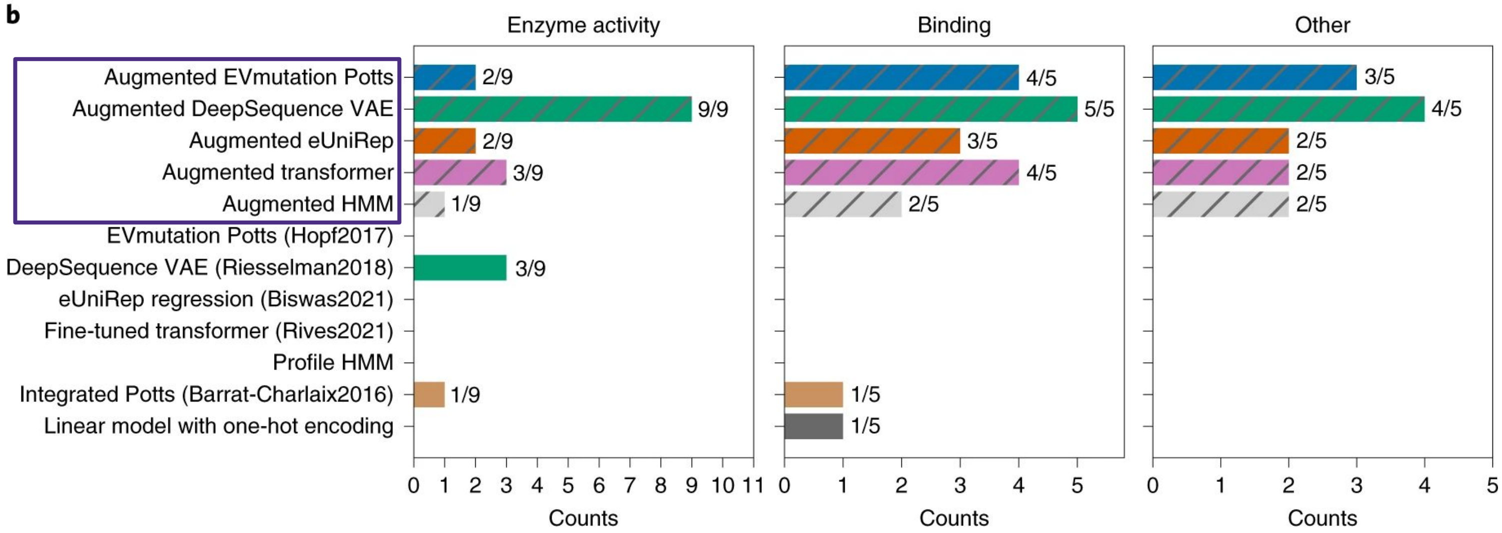

Average performance (19 data sets)
Spearman correlation
0.7
0.6
0.5
0.4
0.3
48
96
144
192
240
Training data size
80/20 split
Augmented EVmutation Potts
EVmutation Potts (Hopf2017)
Augmented DeepSequence VAE
DeepSequence VAE (Riesselman2018)
Augmented eUniRep
eUniRep regression (Biswas2021)
Augmented transformer
Fine-tuned transformer (Rives2021)
Augmented HMM
Profile HMM
Linear model with one-hot encoding

Performance w/ train N = 240



Augmented DeepSequence VAE was the best (esp. enzyme activity)

Maximal Spearman values w/ train N = 240



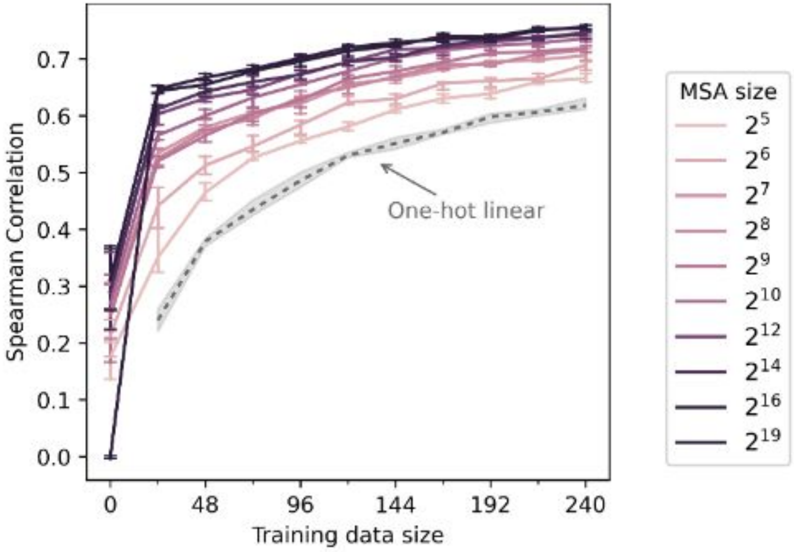
Augmented DeepSequence VAE was the best (esp. enzyme activity)

Performance w/ train N = 240

- Models with evolutionary data had better Spearman correlation w/ larger effective MSA size
- Relative model ranking appeared to not relate to MSA size

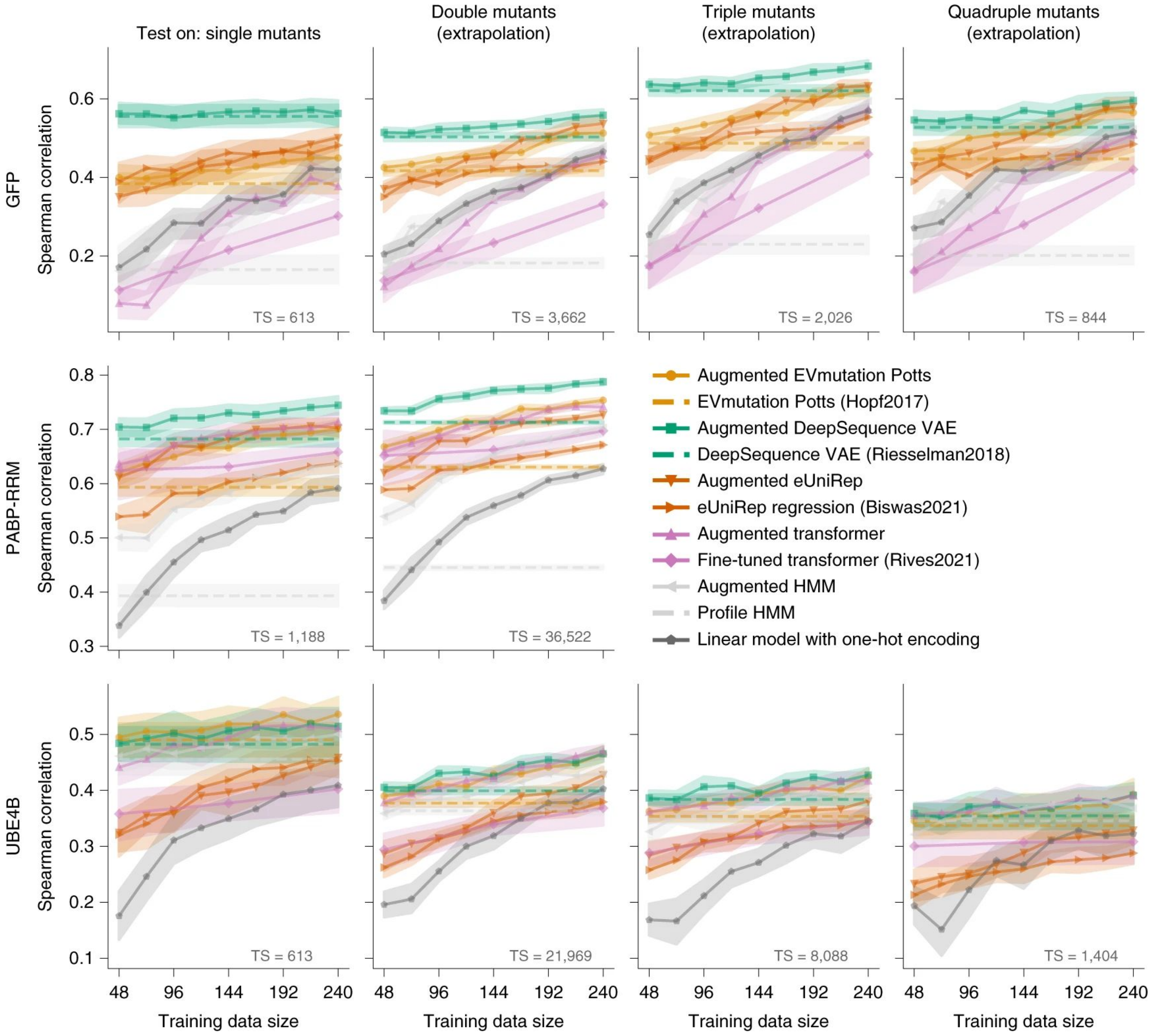
Effect of reducing MSA size

- Chose largest MSA data set (poly(A)-binding protein) and decreased effective size
- Examined aug. Potts model peformance



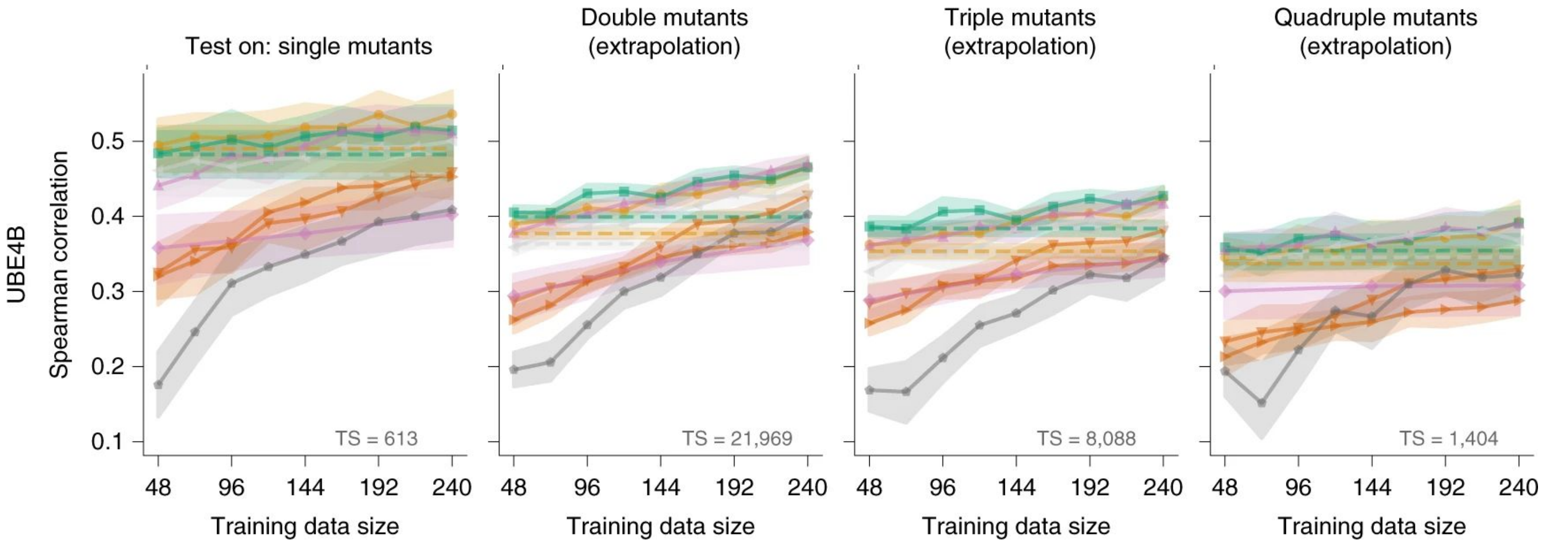
Single to higher order mutant prediction

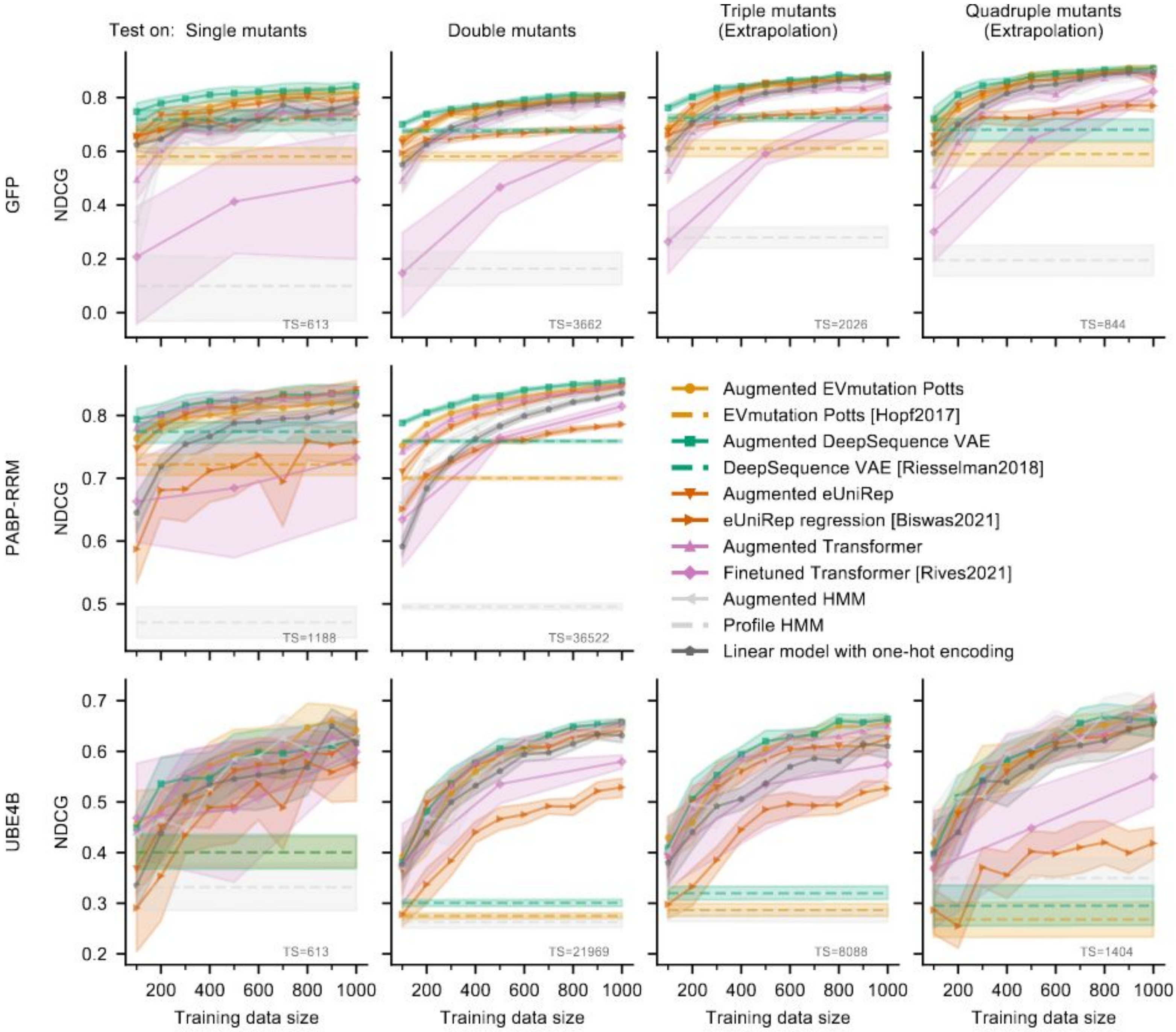
- Trained the model on only single mutant data - tested on single, double, triple, and quadruple mutants
- Should capture how much epistasis contributes to the fitness landscape, and if/how much models capture it
- Only 3 datasets


Test on: single mutants
Double mutants (extrapolation)
Triple mutants (extrapolation)
Quadruple mutants (extrapolation)
GFP
PABP-RRM
UBE4B
Spearman correlation
0.6
0.4
0.2
0.8
0.7
0.5
0.3
0.1
TS = 613
TS = 3,662
TS = 2,026
TS = 844
TS = 1,188
TS = 36,522
TS = 613
TS = 21,969
TS = 8,088
TS = 1,404
48
96
144
192
240
Training data size
Augmented EVmutation Potts
EVmutation Potts (Hopf2017)
Augmented DeepSequence VAE
DeepSequence VAE (Riesselman2018)
Augmented eUniRep
eUniRep regression (Biswas2021)
Augmented transformer
Fine-tuned transformer (Rives2021)
Augmented HMM
Profile HMM
Linear model with one-hot encoding

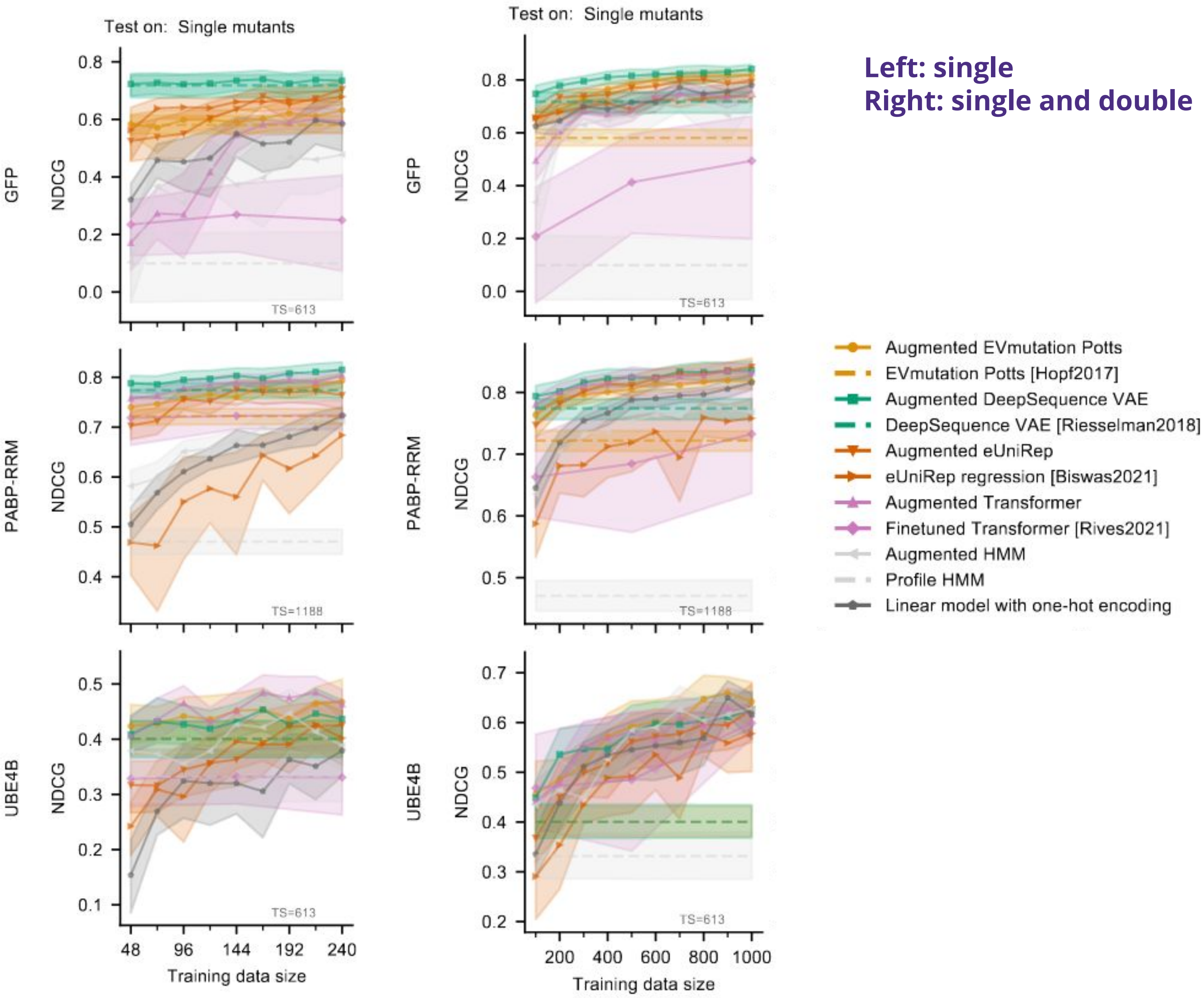
Single to higher order mutant prediction

> Poor performance on ubiquitination factor E4B (UBE4B) may be due to evolutionary data not providing much relevant information to assayed value

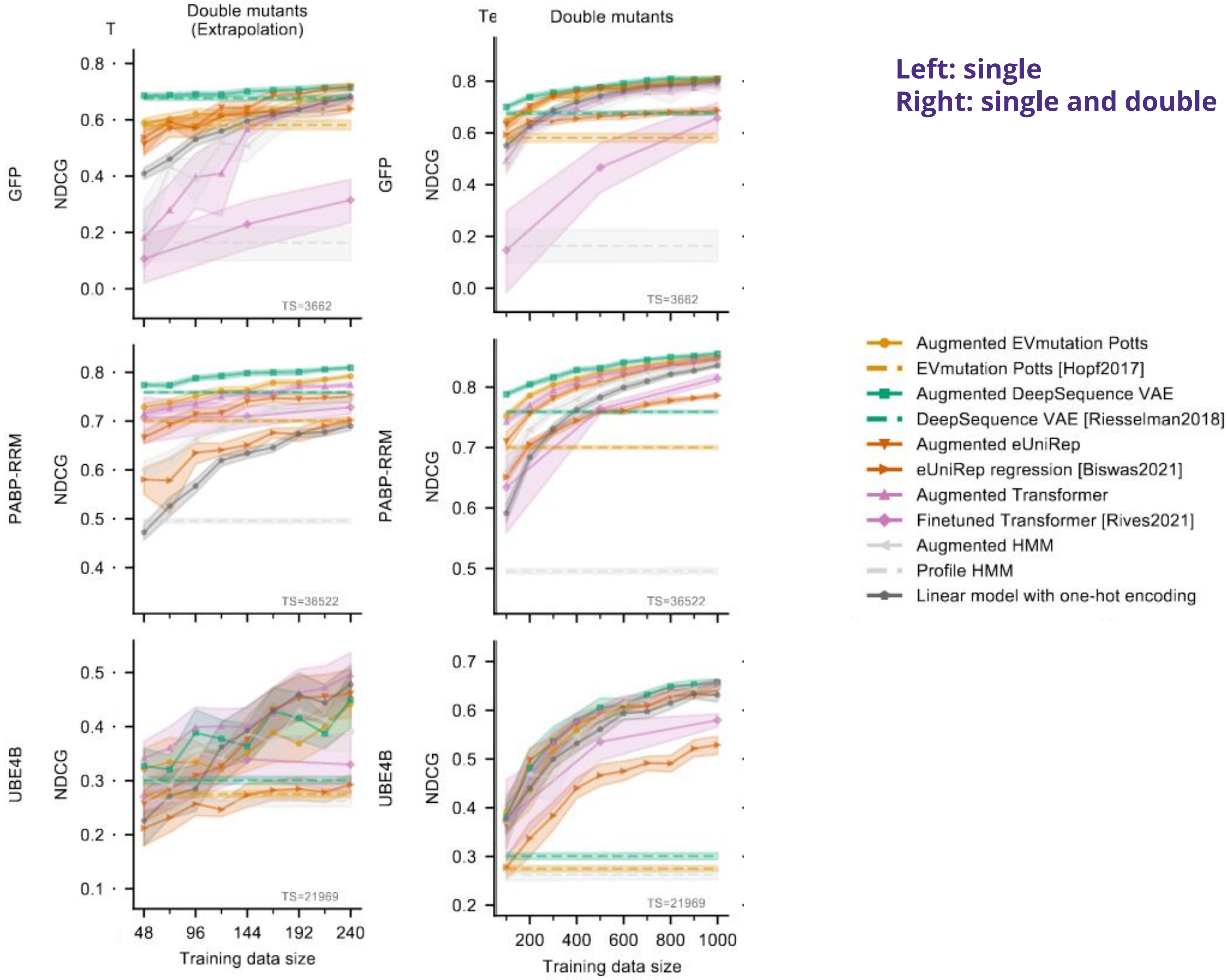



Test on: Single mutants
Double mutants
Triple mutants (Extrapolation)
Quadruple mutants (Extrapolation)
GFP
PABP-RRM
UBE4B
NDCG
0.0
0.2
0.4
0.6
0.8
0.5
0.6
0.7
0.8
0.2
0.3
0.4
0.5
0.6
0.7
TS=613
TS=3662
TS=2026
TS=844
TS=1188
TS=36522
TS=613
TS=21969
TS=8088
TS=1404
200
400
600
800
1000
Training data size
Augmented EVmutation Potts
EVmutation Potts [Hopf2017]
Augmented DeepSequence VAE
DeepSequence VAE [Riesselman2018]
Augmented eUniRep
eUniRep regression [Biswas2021]
Augmented Transformer
Finetuned Transformer [Rives2021]
Augmented HMM
Profile HMM
Linear model with one-hot encoding

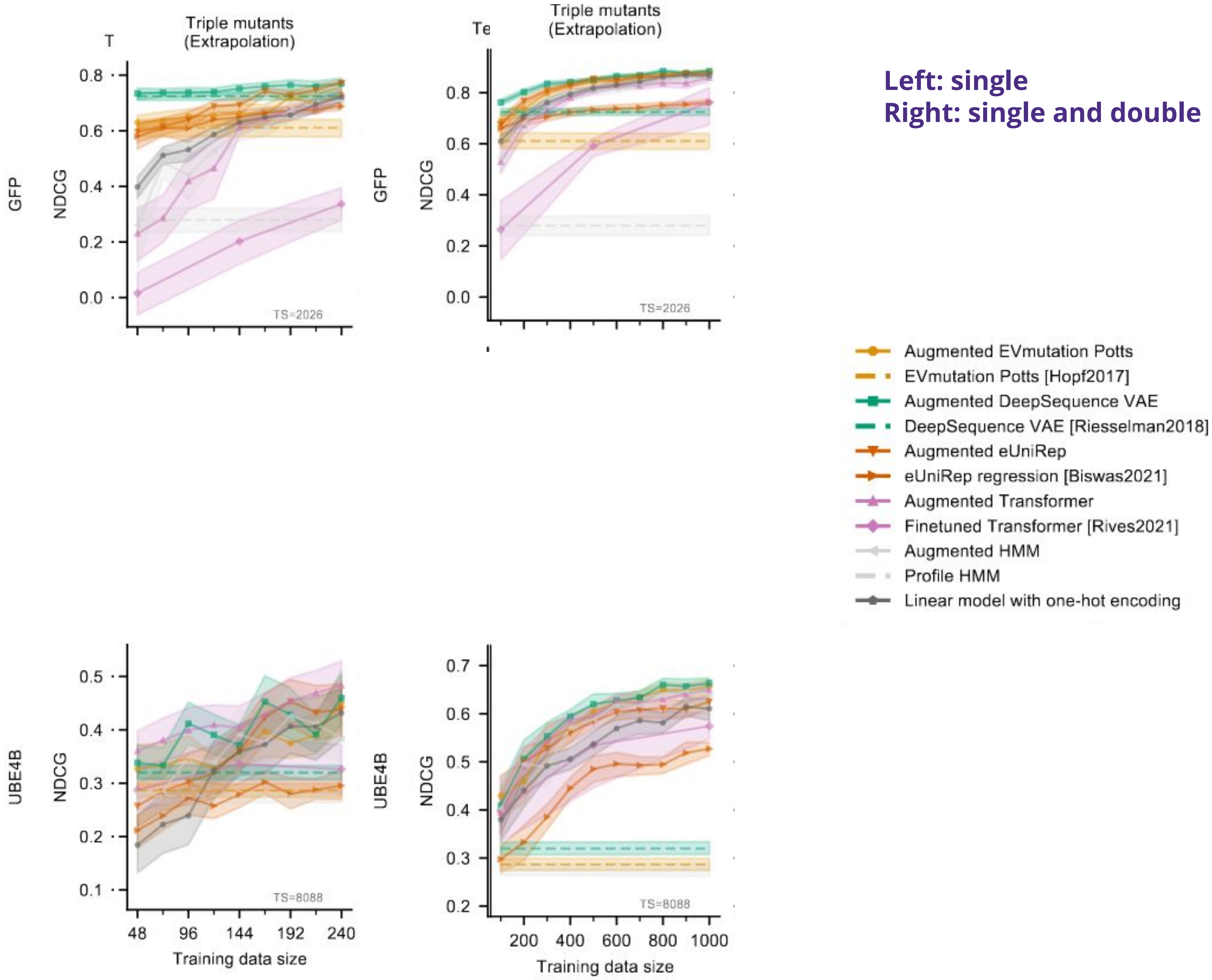
Left: single
Right: single and double



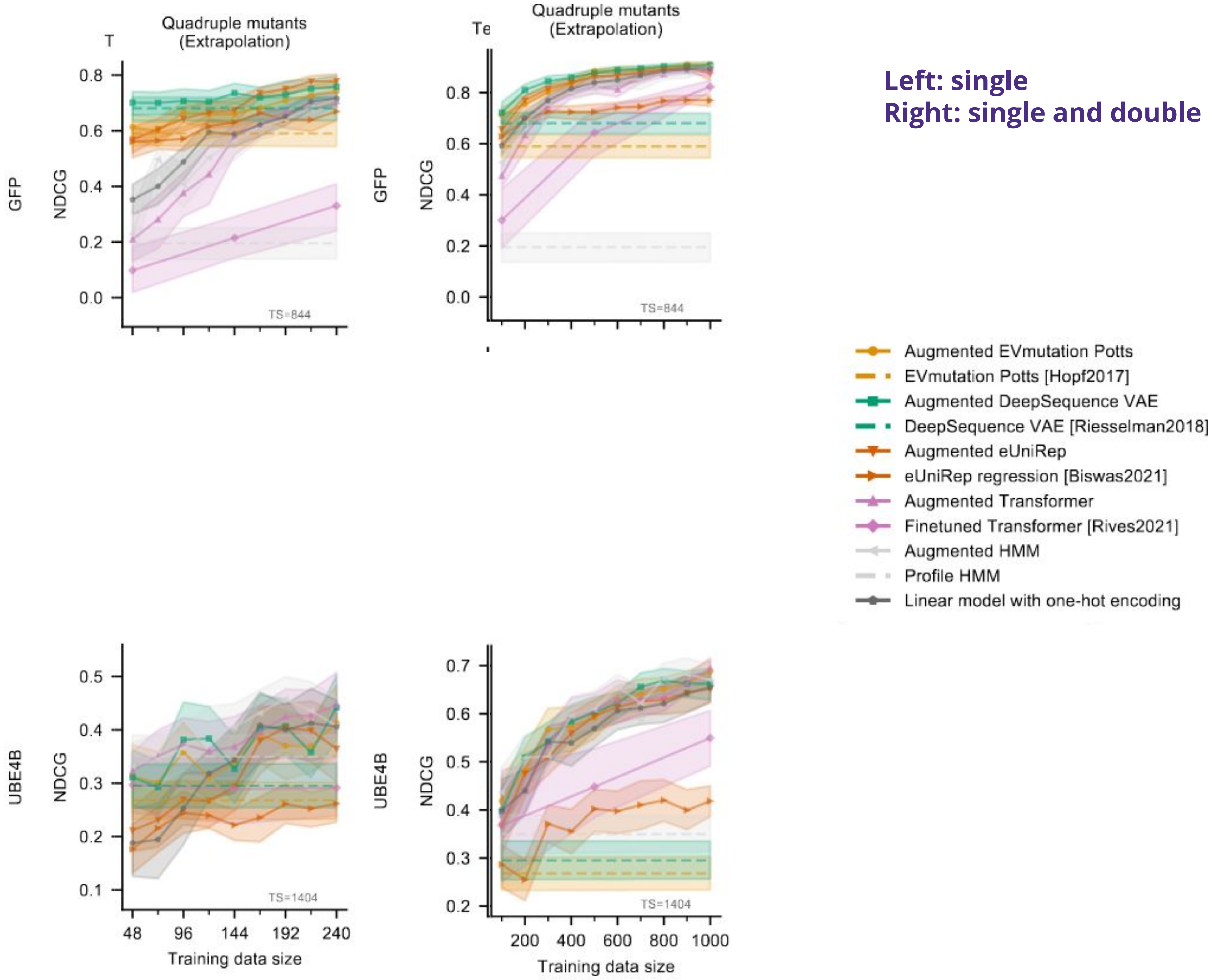
Left: single
Right: single and double



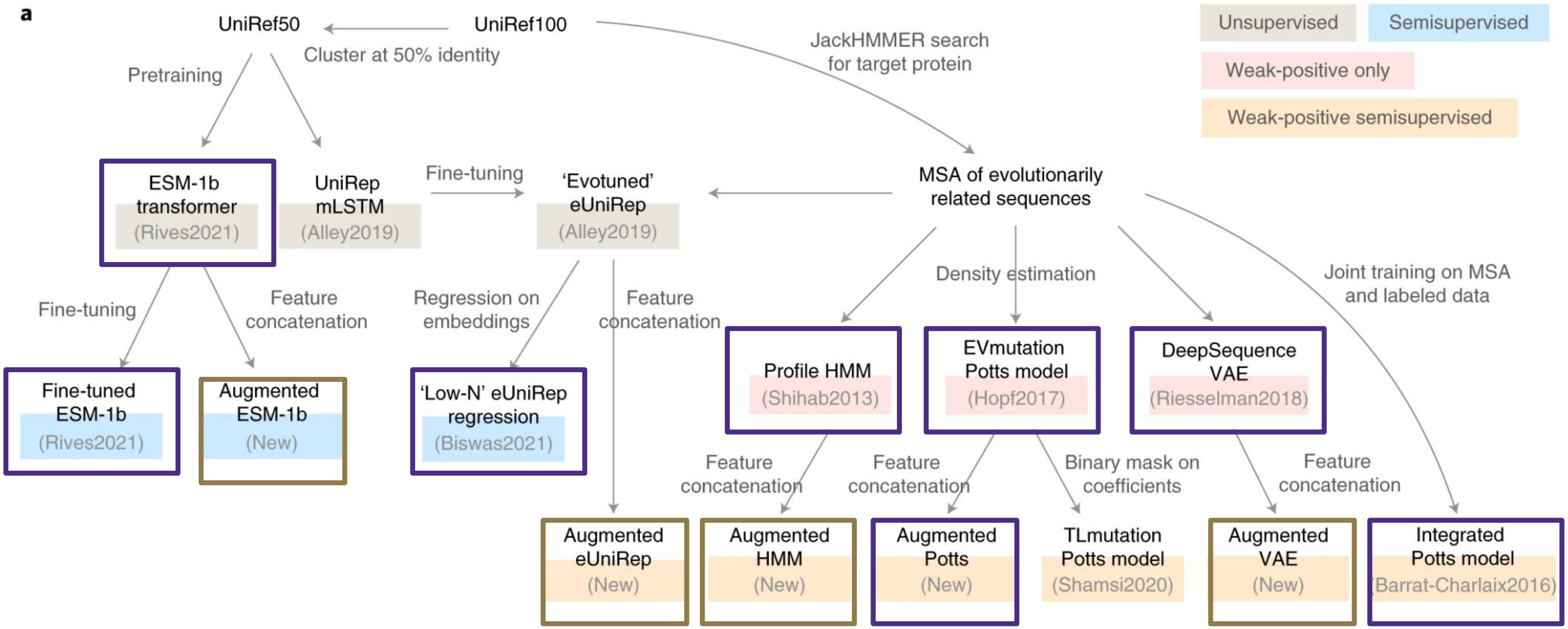
Left: single
Right: single and double



Left: single
Right: single and double



One hot linear model

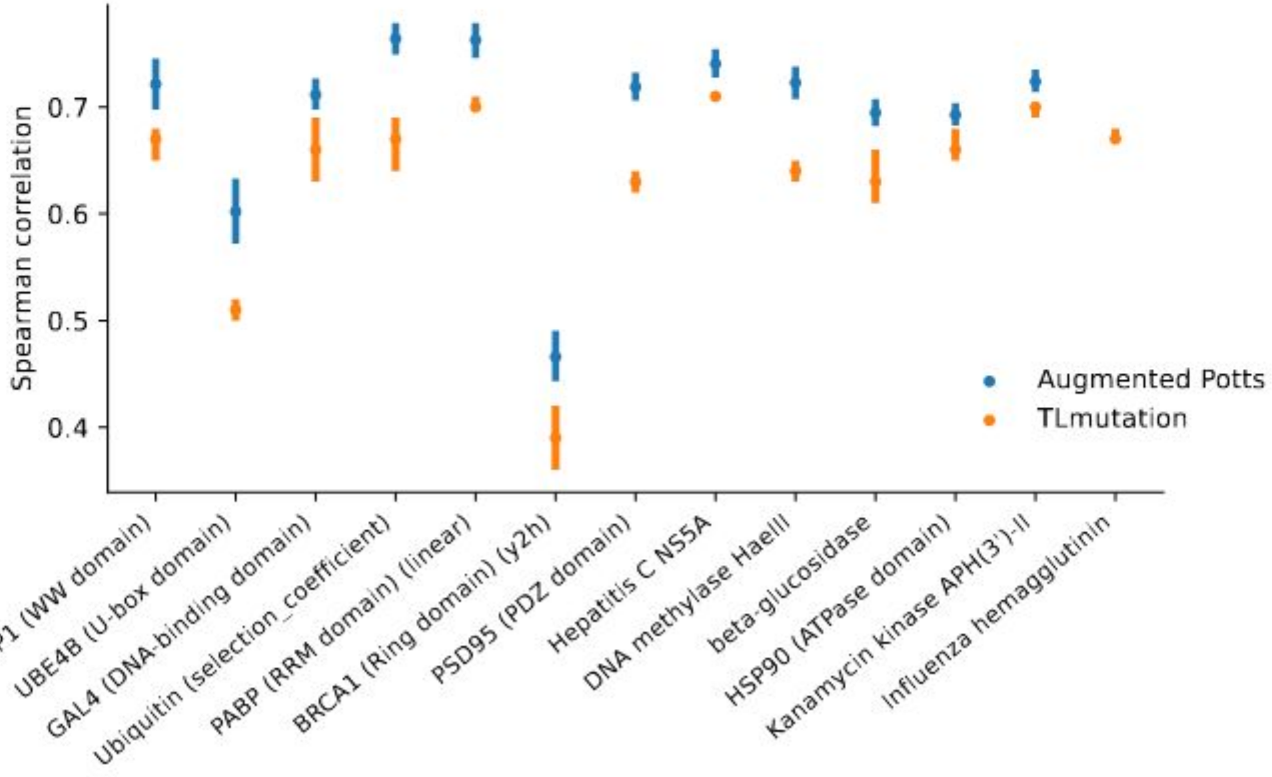

a
UniRef50
UniRef100
Cluster at 50% identity
JackHMMER search for target protein
Unsupervised
Semisupervised
Weak-positive only
Weak-positive semisupervised
Pretraining
ESM-1b transformer (Rives2021)
UniRep mLSTM (Alley2019)
Fine-tuning
'Evotuned' eUniRep (Alley2019)
MSA of evolutionarily related sequences
Fine-tuning
Feature concatenation
Regression on embeddings
Feature concatenation
Density estimation
Joint training on MSA and labeled data
Profile HMM (Shihab2013)
EVmutation Potts model (Hopf2017)
DeepSequence VAE (Riesselman2018)
Fine-tuned ESM-1b (Rives2021)
Augmented ESM-1b (New)
'Low-N' eUniRep regression (Biswas2021)
Feature concatenation
Feature concatenation
Binary mask on coefficients
Feature concatenation
Augmented eUniRep (New)
Augmented HMM (New)
Augmented Potts (New)
TLmutation Potts model (Shamsi2020)
Augmented VAE (New)
Integrated Potts model (Barrat-Charlaix2016)

What about TLMutation?

TLMutation

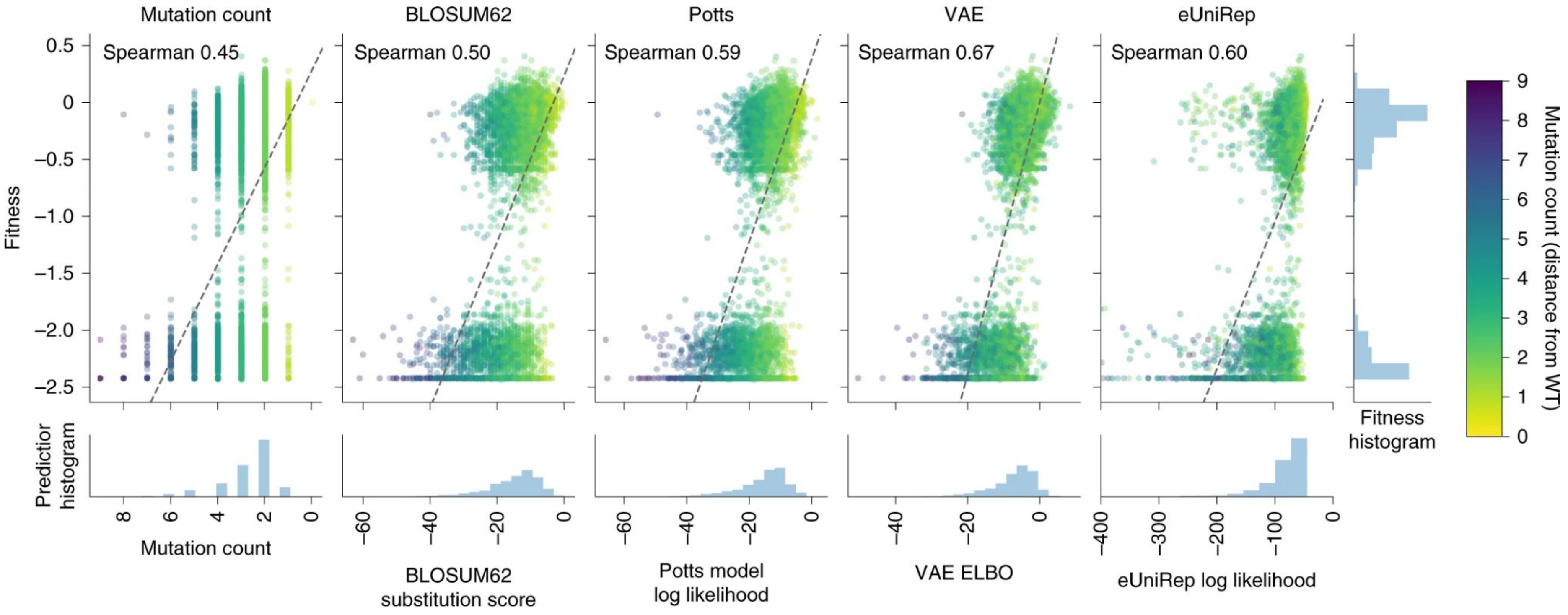
- Conceptually similar to the aug. Potts model (combining density model and supervised learning)
- Allows for zeroing out Potts model parameters with supervised learning - learns a mask.
- More computationally expensive, worse than the aug Potts model

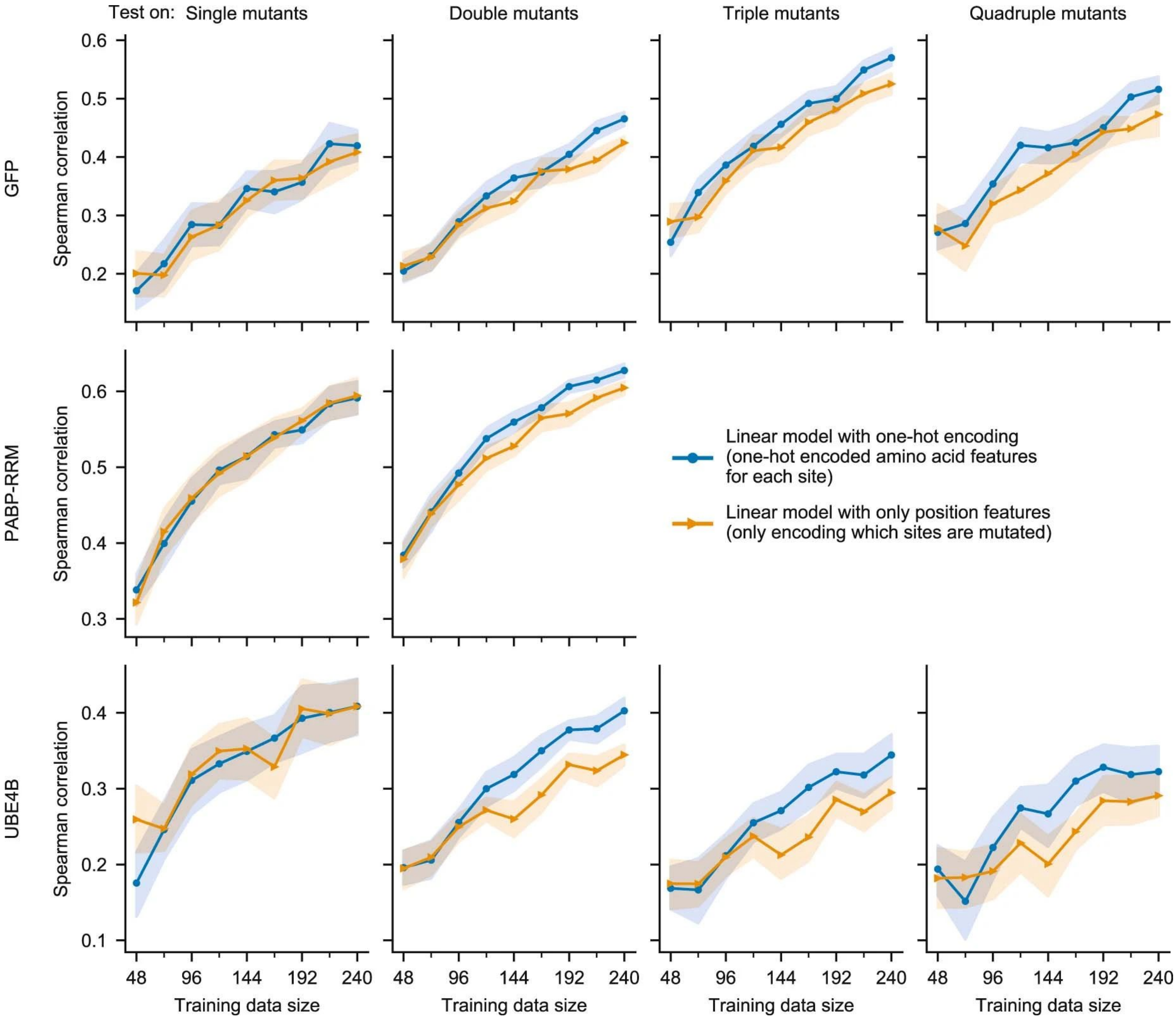
TLMutation comparison



What about really simple models?

- They tried using edit distance from WT to predict fitness & found some correlation with GFP
- Non-aug. predictions were unimodal, but fitness values were bimodal
- Less correlation with UBE4B
- Tried just encoding position information +


Mutation count
BLOSUM62
Potts
VAE
eUniRep
Spearman 0.45
Spearman 0.50
Spearman 0.59
Spearman 0.67
Spearman 0.60
Fitness
0.5
0
−0.5
−1.0
−1.5
−2.0
−2.5
Predictior histogram
8
6
4
2
0
Mutation count
−60
−40
−20
0
BLOSUM62 substitution score
−60
−40
−20
0
Potts model log likelihood
−40
−20
0
VAE ELBO
−400
−300
−200
−100
0
eUniRep log likelihood
Fitness histogram
Mutation count (distance from WT)
9
8
7
6
5
4
3
2
1
0


Test on: Single mutants
Double mutants
Triple mutants
Quadruple mutants
GFP
PABP-RRM
UBE4B
Spearman correlation
Training data size
48
96
144
192
240
0.1
0.2
0.3
0.4
0.5
0.6
Linear model with one-hot encoding (one-hot encoded amino acid features for each site)
Linear model with only position features (only encoding which sites are mutated)

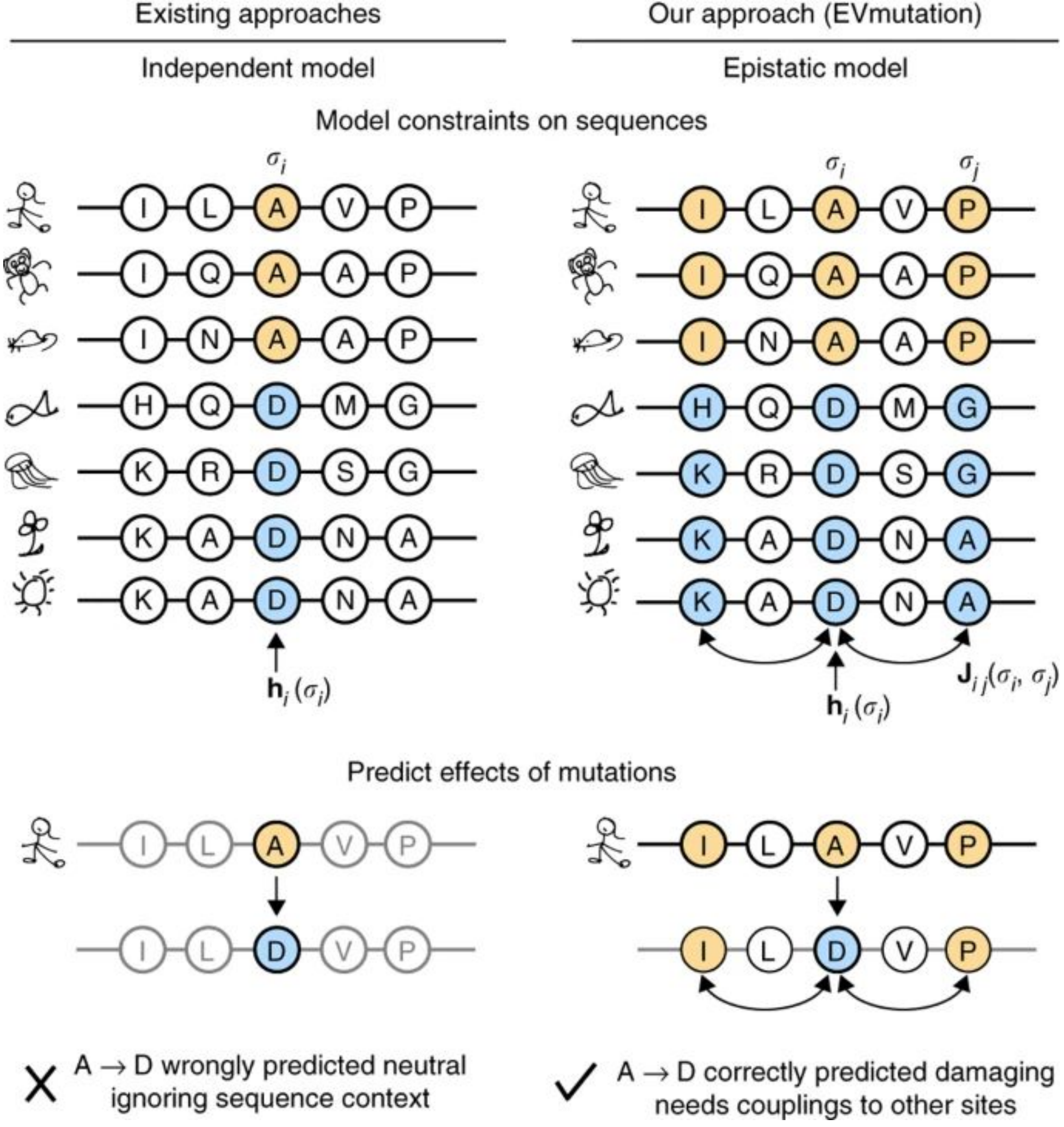
Summary

- Simple linear regression with one-hot encoded amino acid features and a evolutionary density feature from density models outperforms said density models
- Deep learning models may be used with these features instead - but this was not tested
- Aug. transformer could be used for small MSA proteins

Discussion questions

- Do you think using their augmented features with a more complicated regression model would lead to a better predictor?
  - Would it be worth (presumably) trade-offs in requiring higher N training set sizes?
  - Would it be worth it for protein design?
- Do you think their decision to remove TLMutation from their comparison figures throughout was fair to the assessment?

Misc.


Existing approaches
Our approach (EVmutation)
Independent model
Epistatic model
Model constraints on sequences
σ_i
I L A V P
I Q A A P
I N A A P
H Q D M G
K R D S G
K A D N A
K A D N A
$\mathbf{h}_i(\sigma_i)$
σ_i
σ_j
$\mathbf{J}_{ij}(\sigma_i, \sigma_j)$
$\mathbf{h}_i(\sigma_i)$
Predict effects of mutations
I L A V P
I L D V P
✗ A → D wrongly predicted neutral ignoring sequence context
✓ A → D correctly predicted damaging needs couplings to other sites

Protein	UniProt ID	Measurement	# Mutations before exclusion	# Mutations after excluding positions with $\geq$ 30% gaps	MSA size	Reference
β-glucosidase	(Custom sequence)	Enzyme activity	3000 (1)	2634 (1)	28048	Romero et al., PNAS, 2015
β-lactamase	BLAT_ECOLX	Ampicillin resistance Ampicillin resistance Amoxicillin resistance Stability	5199 (1) 4997 (1) 990 (1) 4998 (1)	4611 (1) 4807 (1) 951 (1) 4808 (1)	8403	Firnberg et al., Mol Biol Evol, 2014 Stiffler et al., Cell, 2015 Jacquier et al., PNAS 2013 Deng et al., JMB, 2012
BRCA 1 (RING domain)	BRCA1_HUMAN	E3 ligase activity BARD1 interaction	4872 (1) 1748 (1)	1382 (1) 1335 (1)	25828	Starita et al., Genetics, 2015
PSD95 (PDZ domain)	DLG4_RAT	Peptide binding	1578 (1)	1578 (1)	102410	McLaughlin et al., Nature, 2012
GAL4 (DNA-binding domain)	GAL4_YEAST	Transcriptional activity	1196 (1)	1123 (1)	17521	Kitzmann et al., Nature Methods, 2015
HSP90 (ATPase domain)	HSP82_YEAST	ATPase activity	4324 (1)	4104 (1)	15329	Mishra et al., Cell Reports, 2016
Kanamycin kinase APH(3')-II	KKA2_KLEPN	Kinase activity	4582 (1)	4385 (1)	12861	Melnikov et al., NAR, 2014
DNA methylase HaeIII	(Custom sequence)	DNA methyltransferase activity	1778 (1, filtered)	1634 (1)	14115	Rockah-Shmuel et al., PLOS Comp Bio, 2015
Poly(A)-binding protein (RRM domain)	PABP_YEAST	RNA binding	1188 (1) 36522 (2)	1188 (1) 36522 (2)	152041	Melamed et al., RNA, 2013
Hepatitis C NS5A	POLG_HCVJF	Viral replication	1632 (1)	1632 (1)	8106	Qi et al., PLOS Pathogens, 2014
Ubiquitin	RL401_YEAST	Growth E1 reactivity	1196 (1) 1366 (1)	1161 (1) 1295 (1)	21448	Roscoe et al, JMB, 2013 Roscoe et al, JMB, 2014
UBE4B (U-box domain)	UBE4B_MOUSE	Ligase activity	91031 (1-9 mut.)	613 (1) 21969 (2) 8088 (3) 1404 (4) 216 ($\geq$ 5)	9172	Starita et al., PNAS, 2013
YAP1 (WW domain 1)	YAP1_HUMAN	Peptide binding	363 (1)	319 (1)	40302	Araya et al., PNAS, 2012
Green Fluorescent Protein	(Custom sequence)	Fluorescence	51715 (1-14 mut.)	613 (1) 3662 (2) 2026 (3) 844 (4) 630 ($\geq$ 5)	22535	Sarkisyan et al., Nature, 2016