SCIENCES SORBONNE UNIVERSITÉ

LCQB

ROSTLAB.

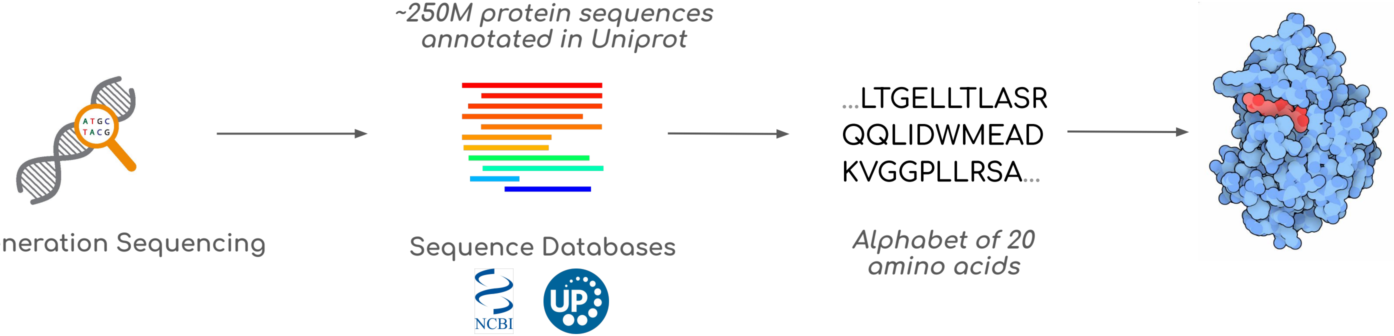
Guided by an Expert Teacher

Accurate and blazingly fast variant effect prediction using protein language model embeddings

Julius Schlensok, Céline Marquet, Marina Abakarova, Burkhard Rost & Elodie Laine

Motivation

Understanding the impact of single amino acid variants (SAVs) on protein function



~250M protein sequences annotated in Uniprot

Next Generation Sequencing

Sequence Databases

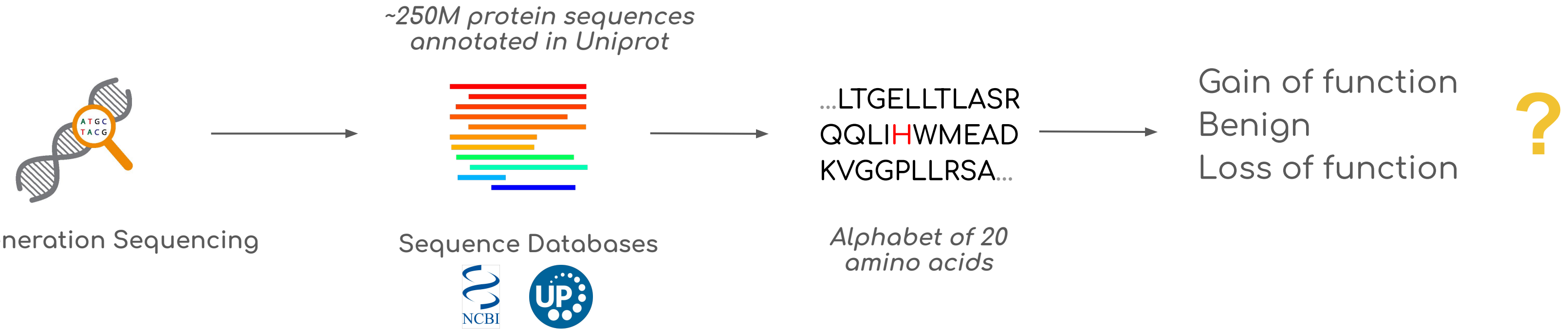
...LTGELLTLASR
QQLIDWMEAD
KVGGPLLRSA...

Alphabet of 20 amino acids

Function:
New Dehli metallo-ß-lactamase breaking down an antibiotics

Motivation

Understanding the impact of single amino acid variants (SAVs) on protein function



Next Generation Sequencing

~250M protein sequences annotated in Uniprot

Sequence Databases

...LTGELLTLASR
QQLIHWMEAD
KVGGPLLRSA...

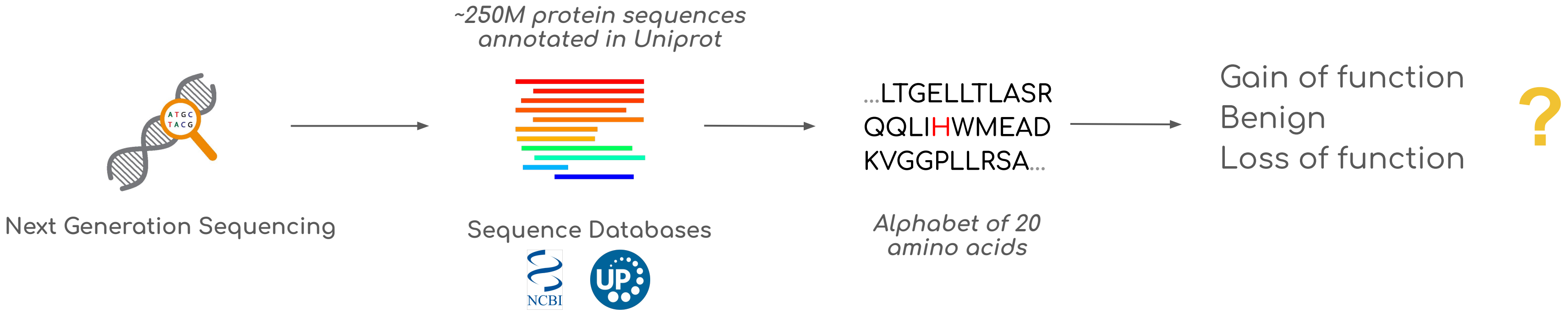
Alphabet of 20 amino acids

Gain of function
Benign
Loss of function

?

Motivation

Understanding the impact of single amino acid variants (SAVs) on protein function



Next Generation Sequencing

~250M protein sequences annotated in Uniprot

Sequence Databases

...LTGELLTLASR
QQLIHWMEAD
KVGGPLLRSA...

Alphabet of 20 amino acids

Gain of function
Benign
Loss of function

?

- ❑ Fundamental biology
- ❑ Bioengineering
- ❑ Drug design

Experimental answer

Deep Mutational Scanning (DMS)

Quantification of mutational outcomes on a large scale

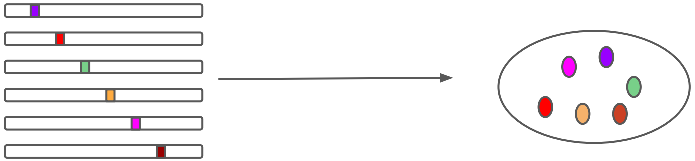
Protocol

Library of mutants

All possible substitutions at all positions

Phenotype

localization, growth, enzyme function, binding...



Experimental answer

Deep Mutational Scanning (DMS)

Quantification of mutational outcomes on a large scale

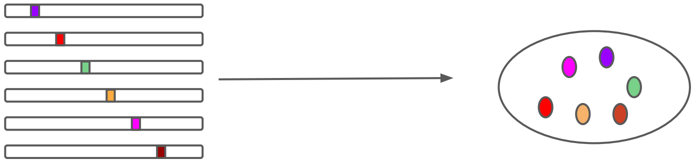
Protocol

Library of mutants

All possible substitutions at all positions

Phenotype

localization, growth, enzyme function, binding...

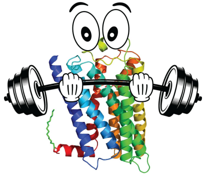


The largest collection of DMS datasets

ProteinGym substitution benchmark

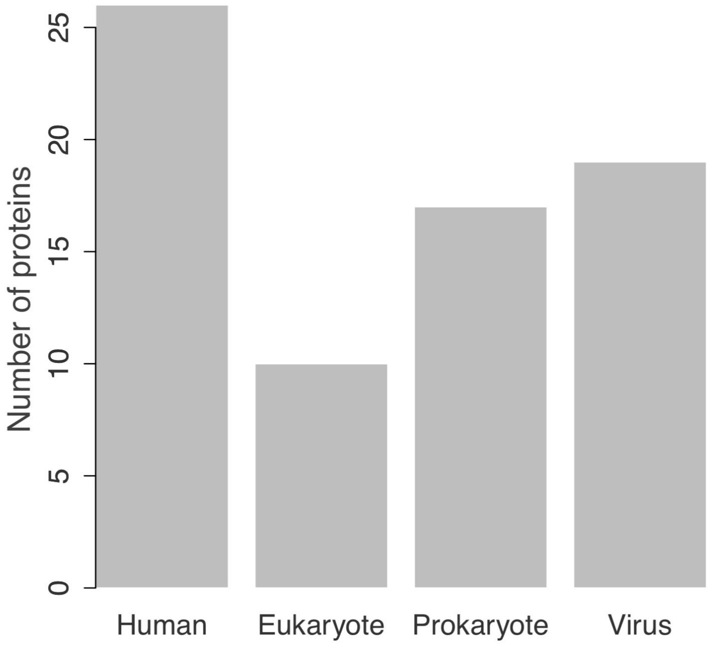
~1.5M SAVs across **72** protein families

https://www.proteingym.org



Notin *et al.* 2022

ProteinGym substitution benchmark



A wide variety of proteins…

- between 70 and 3500 residues
- kinases, ion channels, g-protein coupled receptors, polymerases, transcription factors, tumor suppressors…

… and phenotypes

- thermostability, ligand binding, aggregation, viral replication, and drug resistance

Between 1 and 4 DMS assays per protein

Multiple mutation assays for 11 proteins

DMS (or MAVE) experiments remain too costly for proteome scanning.

ProteinGym substitution benchmark

Number of proteins
0
5
10
15
20
25
Human
Eukaryote
Prokaryote
Virus

A wide variety of proteins…

- between 70 and 3500 residues

coupled
tion

d drug

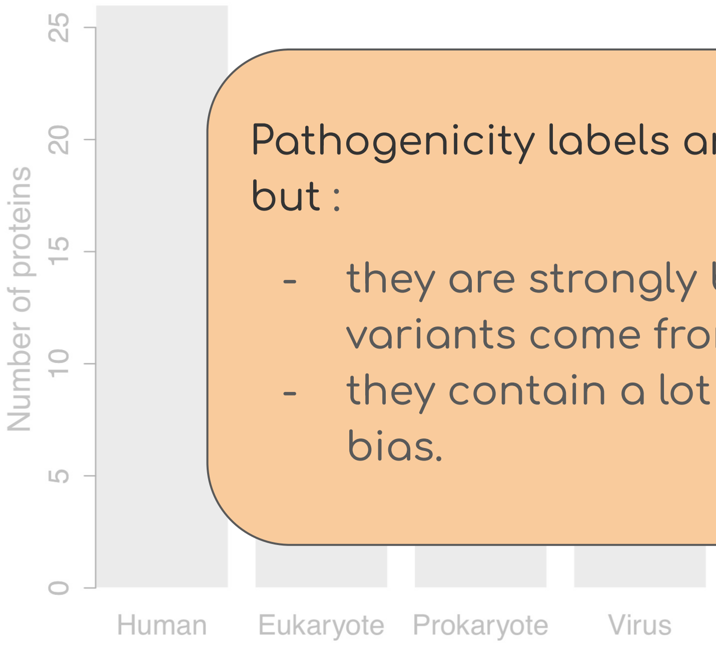
Multiple mutation assays for 11 proteins

DMS experiments:

- may not provide the FULL mutational landscape
- may be noisy, same protein same phenotype but $\rho = 0.58$
- do not agree on a “consensus” phenotype

DMS (or MAVE) experiments remain too costly for proteome scanning.

ProteinGym substitution benchmark



A wide variety of proteins...

between 70 and 3500 residues

coupled
iption

nd drug

Multiple mutation assays for 11 proteins

Pathogenicity labels are also available, e.g. through ClinVar, but :

- they are strongly biased toward few proteins (75% of the variants come from 10% of human genes),
- they contain a lot of uncertainty and human expertise bias.

DMS (or MAVE) experiments remain too costly for proteome scanning.

Computational predictive methods

Supervised

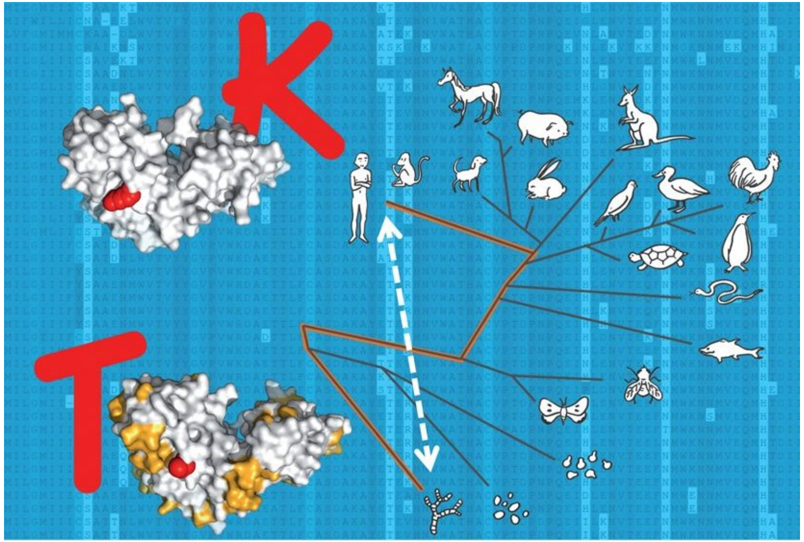
Polyphen-2 (Adzhubei *et al.* 2013)
Envision (Gray *et al.* 2018)
Song *et al.* 2021
VESPA (Marquet *et al.* 2022)
FiTMuSiC (Tsishyn *et al.* 2023)
...

SOTA methods leverage protein sequence information across species.
A few also exploit population data.

Weakly or Un-supervised

CADD (Kircher *et al.* 2014)
DCA (Figliuzzi *et al.* 2016)
DeepSequence (Riesselman *et al.* 2018)
GEMME (Laine *et al.* 2019)
PrimateAI (Sundaram *et al.* 2019)
EVE (Frazer *et al.* 2021)
ESM (Meier *et al.* 2021)
Tranception (Notin *et al.* 2022)
PoET (Truong Jr and Bepler 2023)
AlphaMissense (Cheng *et al.* 2023)
...

Explicitly exploiting natural sequences evolutionary history



http://www.lcqb.upmc.fr/GEMME/Home.html

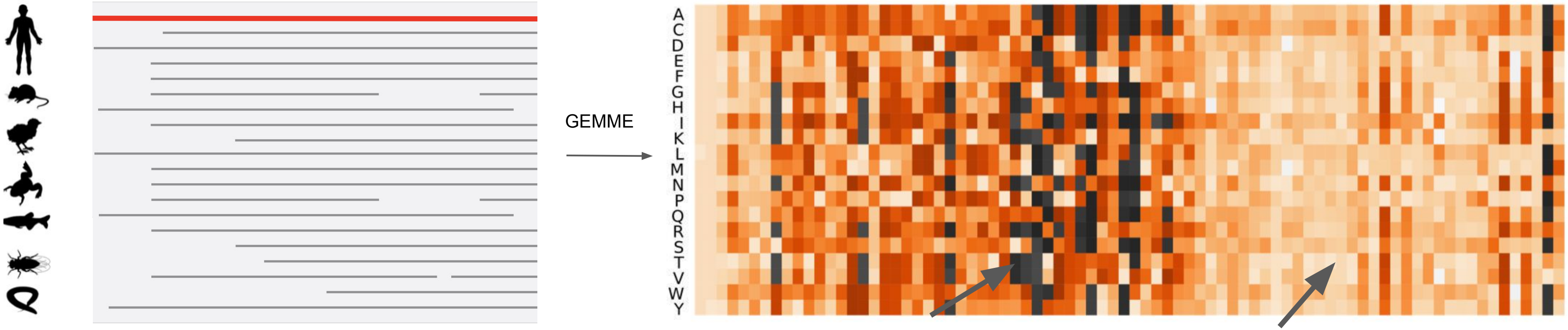
GEMME - an evolutionary-informed predictor

Input

Query-centered multiple sequence alignment (MSA)

Output

Complete single-mutational landscape of the query



high impact

neutral

Aligned homologous sequences

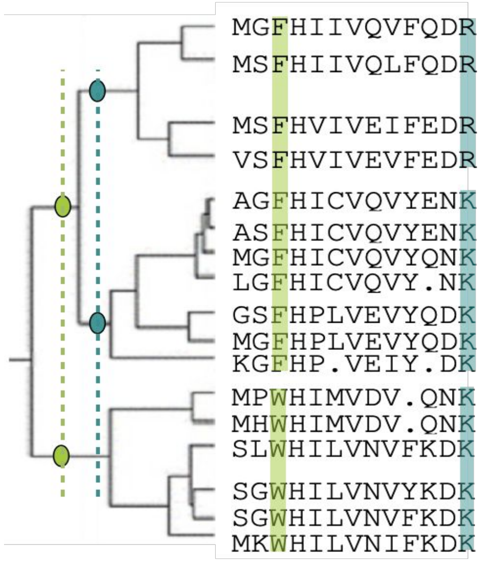
E.Laine *et al.* MBE 2019

http://www.lcqb.upmc.fr/GEMME/Home.html

GEMME - an evolutionary-informed predictor

Main hypotheses: - **conservation** is an indicator of mutational sensitivity

- **epistasis**: positions interact with each other



Joint Evolutionary Trees

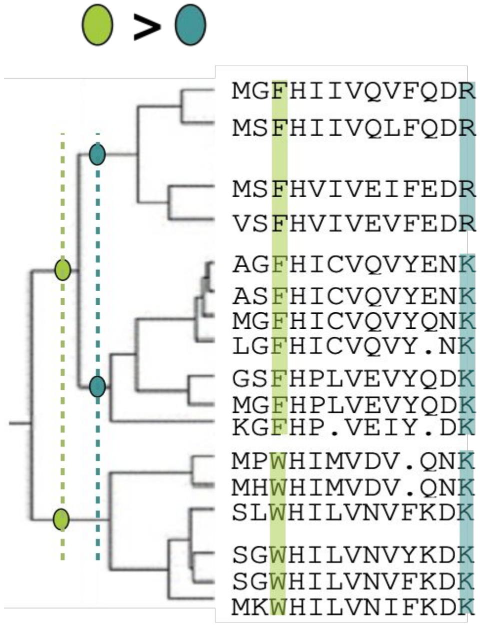
S. Engelen *et al.* PLOS CB 2009

E.Laine *et al.* MBE 2019

http://www.lcqb.upmc.fr/GEMME/Home.html

GEMME - an evolutionary-informed predictor

Main hypotheses: - **conservation** is an indicator of mutational sensitivity
- **epistasis**: positions interact with each other



A measure of conservation accounting for the global context

Joint Evolutionary Trees

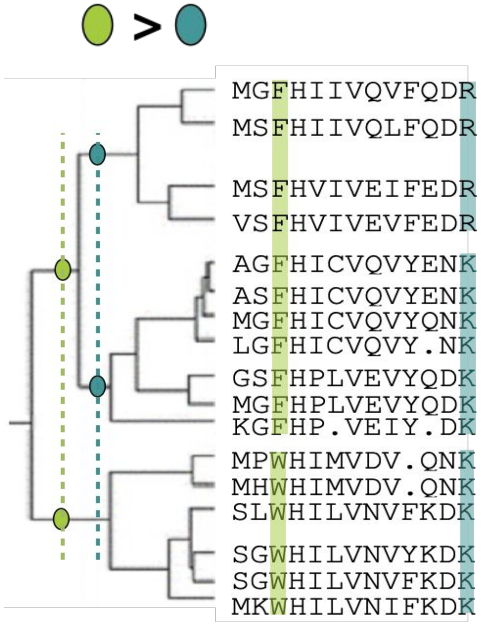
S. Engelen *et al.* PLOS CB 2009

E.Laine *et al.* MBE 2019

http://www.lcqb.upmc.fr/GEMME/Home.html

GEMME - an evolutionary-informed predictor

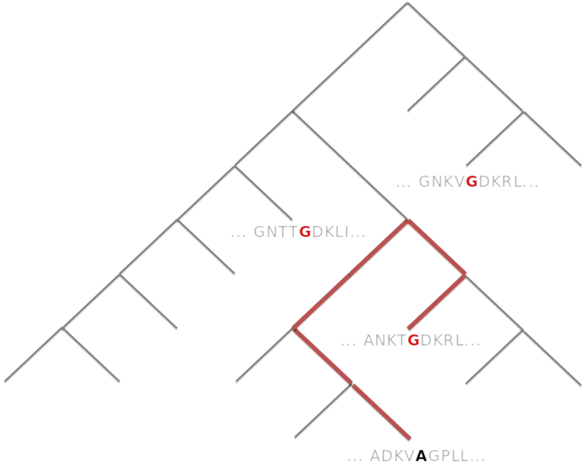
Main hypotheses: - **conservation** is an indicator of mutational sensitivity
- **epistasis**: positions interact with each other



Joint Evolutionary Trees

S. Engelen *et al.* PLOS CB 2009

A measure of conservation accounting for the global context

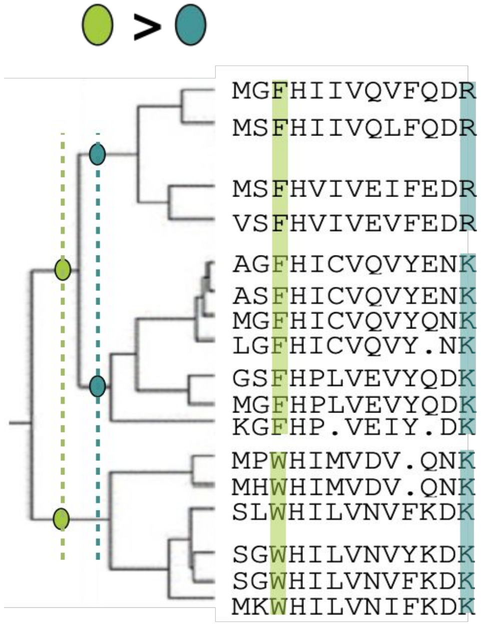


E.Laine *et al.* MBE 2019

http://www.lcqb.upmc.fr/GEMME/Home.html

GEMME - an evolutionary-informed predictor

Main hypotheses: - **conservation** is an indicator of mutational sensitivity
- **epistasis**: positions interact with each other

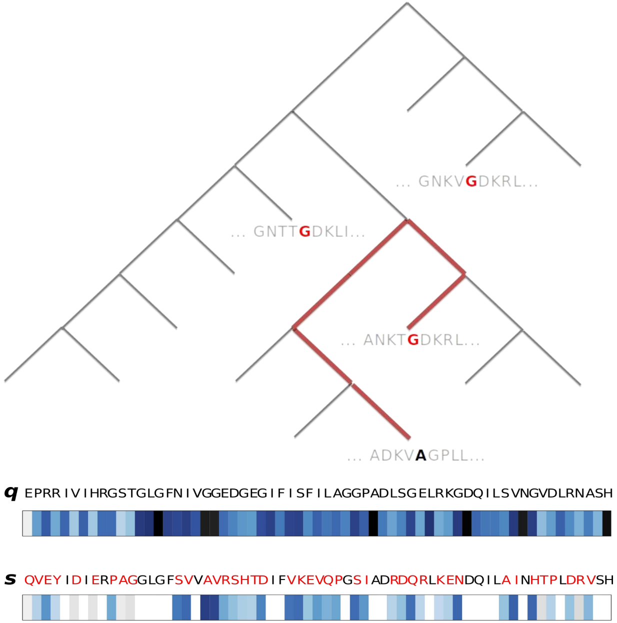


Joint Evolutionary Trees

S. Engelen *et al.* PLOS CB 2009

A measure of conservation accounting for the global context

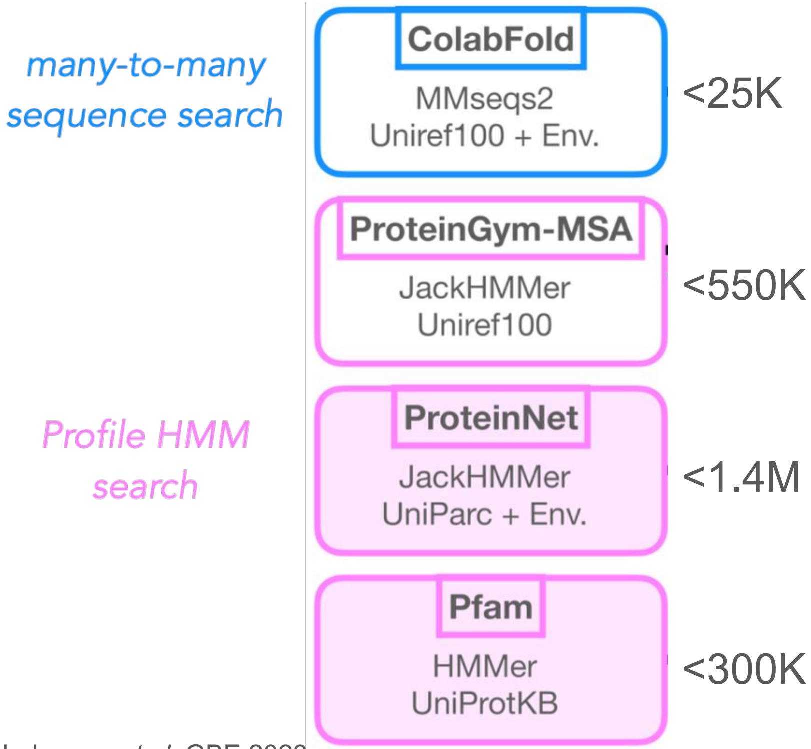
Evolutionary distance to a natural sequence with the mutation


... GNKVGDKRL...
... GNTTGDKLI...
... ANKTGDKRL...
... ADKVAGPLL...
q EPRRIVIHRGSTGLGFNIVGGEDGEGIFISFILAGGPADLSGELRKGDQILSVNGVDLRNASH
s QVEYIDIERPAGGLGFSVVAVRSHTDIFVKEVQPGSIADRDQRLKENDQILAINHTPLDRVSH

E.Laine *et al.* MBE 2019

GEMME - scaling to entire proteomes

GEMME provides a clear readout
of the input alignment.



many-to-many sequence search

ColabFold
MMseqs2
Uniref100 + Env.
<25K

ProteinGym-MSA
JackHMMer
Uniref100
<550K

Profile HMM search

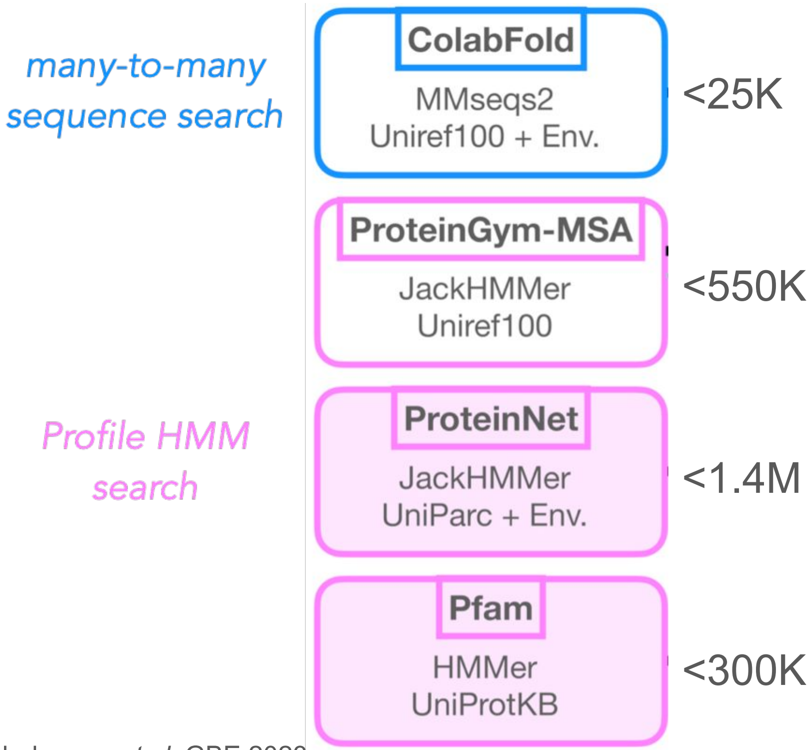
ProteinNet
JackHMMer
UniParc + Env.
<1.4M

Pfam
HMMer
UniProtKB
<300K

M. Abakarova *et al.* GBE 2023

GEMME - scaling to entire proteomes

GEMME provides a clear readout of the input alignment.



many-to-many sequence search

ColabFold
MMseqs2
Uniref100 + Env.
<25K

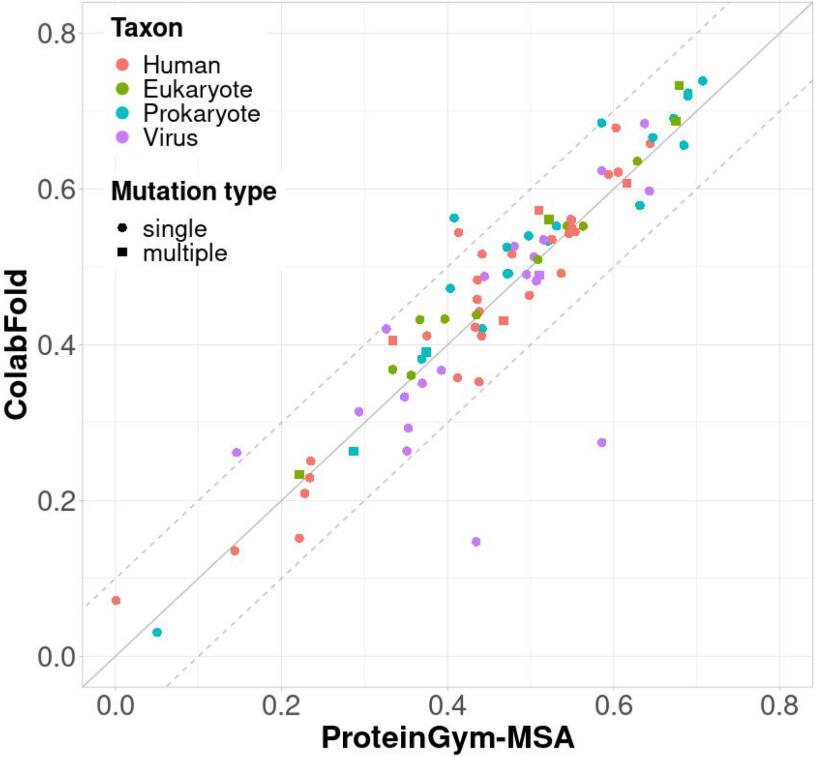
ProteinGym-MSA
JackHMMer
Uniref100
<550K

Profile HMM search

ProteinNet
JackHMMer
UniParc + Env.
<1.4M

Pfam
HMMer
UniProtKB
<300K

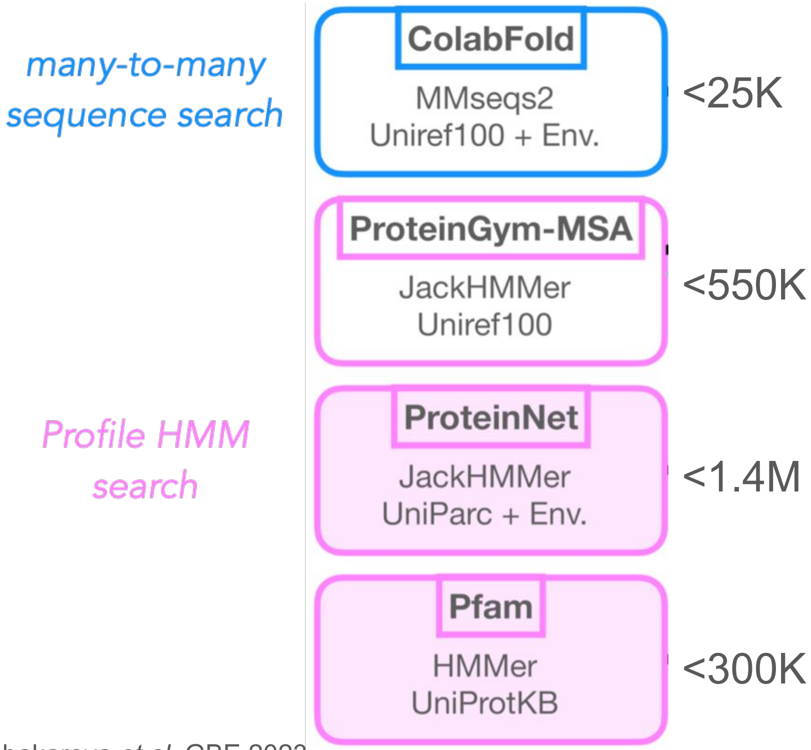
The same prediction accuracy can be attained with much cheaper alignments.



M. Abakarova *et al.* GBE 2023

GEMME - scaling to entire proteomes

GEMME provides a clear readout of the input alignment.



many-to-many sequence search

ColabFold
MMseqs2
Uniref100 + Env.
<25K

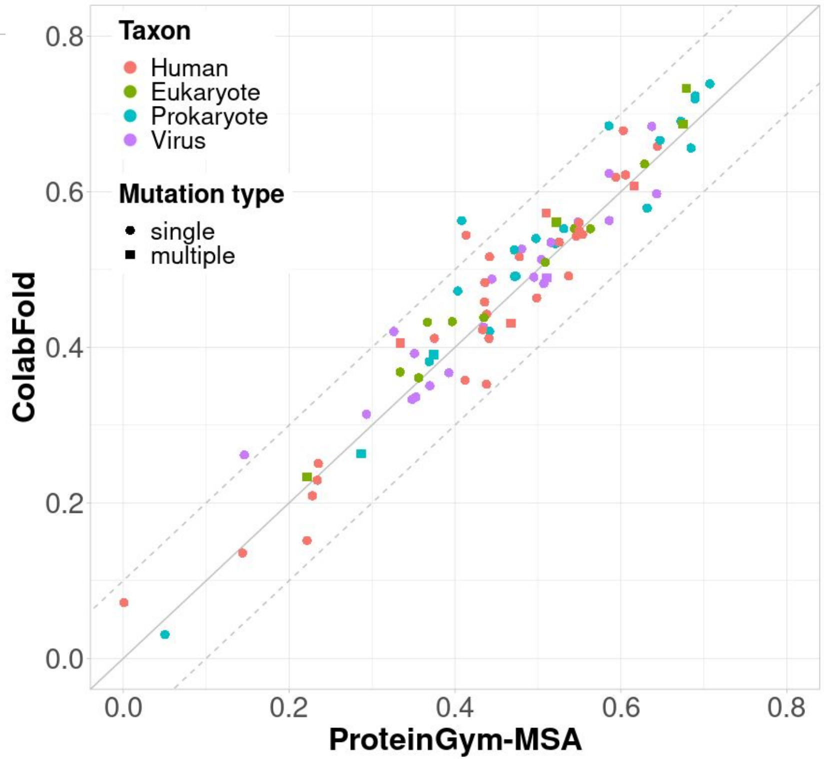
ProteinGym-MSA
JackHMMer
Uniref100
<550K

Profile HMM search

ProteinNet
JackHMMer
UniParc + Env.
<1.4M

Pfam
HMMer
UniProtKB
<300K

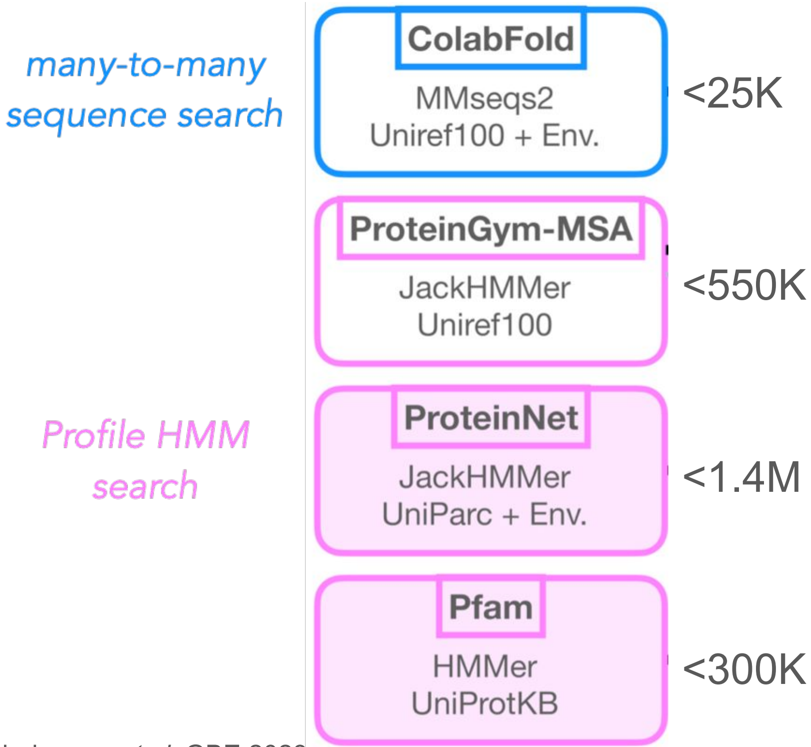
The same prediction accuracy can be attained with much cheaper alignments.



M. Abakarova *et al.* GBE 2023

GEMME - scaling to entire proteomes

GEMME provides a clear readout of the input alignment.



many-to-many sequence search

ColabFold
MMseqs2
Uniref100 + Env.
<25K

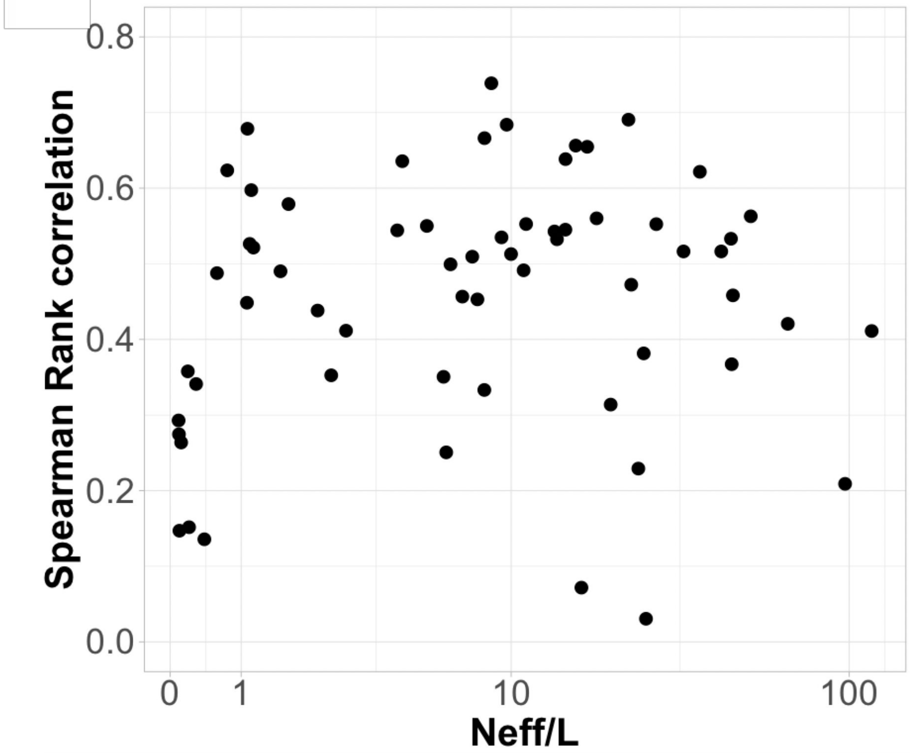
ProteinGym-MSA
JackHMMer
Uniref100
<550K

Profile HMM search

ProteinNet
JackHMMer
UniParc + Env.
<1.4M

Pfam
HMMer
UniProtKB
<300K

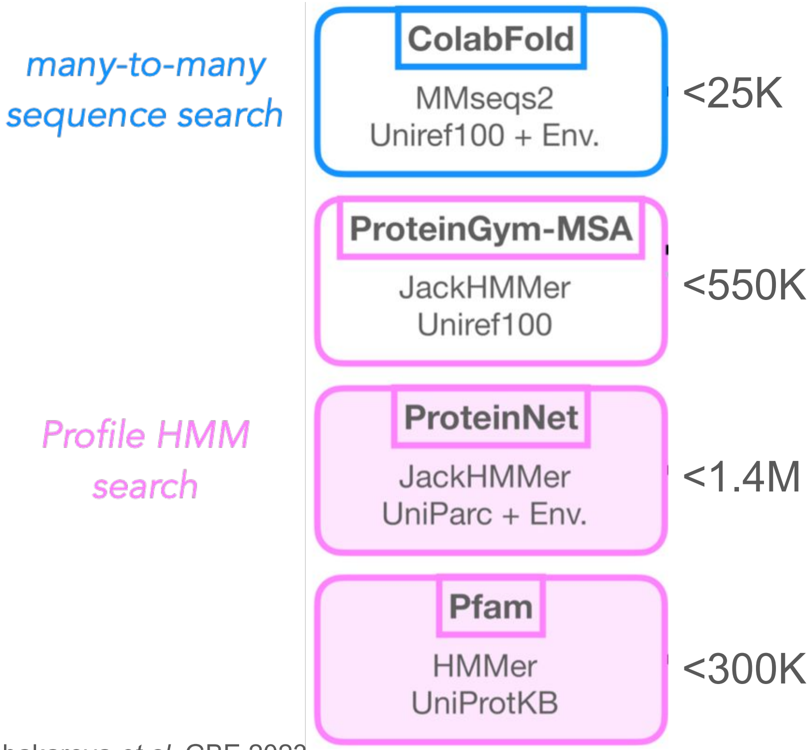
The alignment depth is not as good an indicator of prediction accuracy as one might expect.



M. Abakarova *et al.* GBE 2023

GEMME - scaling to entire proteomes

GEMME provides a clear readout of the input alignment.



many-to-many sequence search

ColabFold
MMseqs2
Uniref100 + Env.
<25K

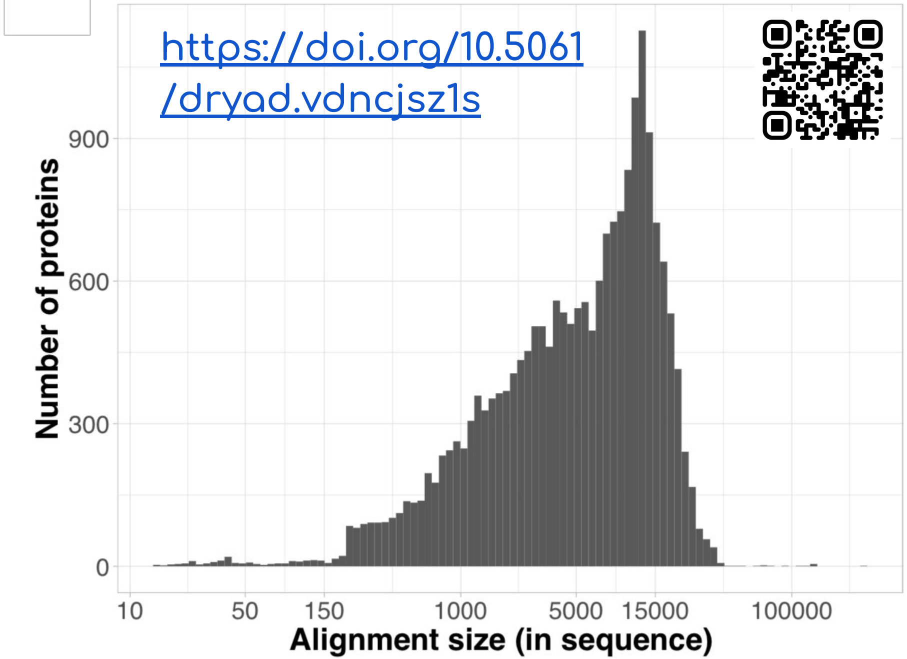
ProteinGym-MSA
JackHMMer
Uniref100
<550K

Profile HMM search

ProteinNet
JackHMMer
UniParc + Env.
<1.4M

Pfam
HMMer
UniProtKB
<300K

Combining ColabFold & GEMME, it takes only a few days to generate the complete single-mutational landscape of the human proteome.



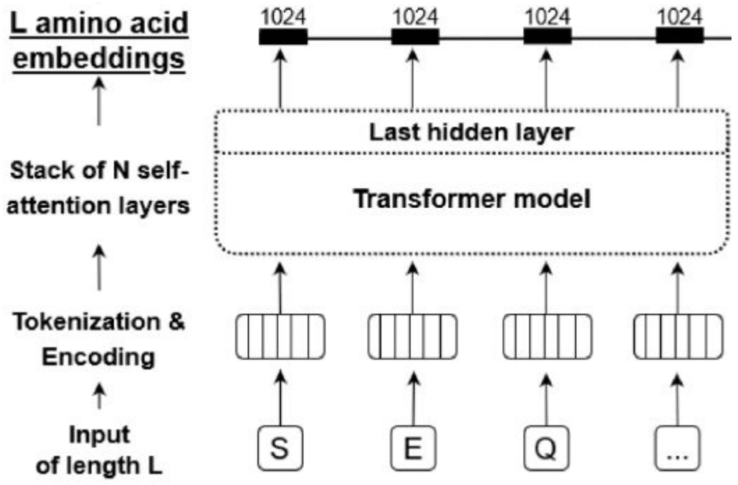
https://doi.org/10.5061/dryad.vdncjsz1s

M. Abakarova *et al.* GBE 2023

Modeling raw protein sequence data at scale

J. Searle's Chinese Room thought experiment

Large language models for proteins

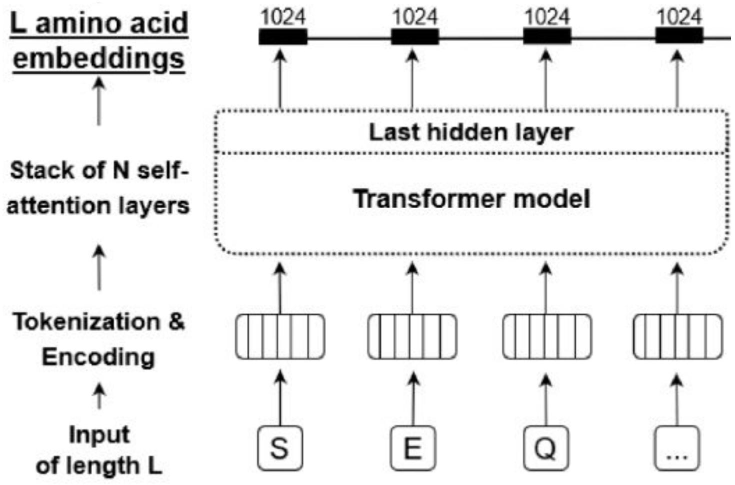


- High capacity transformers
- Input: single sequence (length L)
  Output: high dimensional embedding d x L
- Trained on hundreds of millions of protein sequences to reconstruct masked tokens

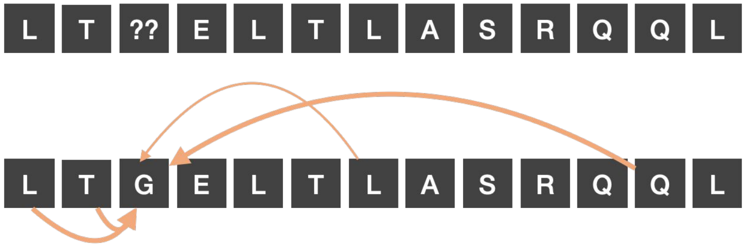
L	T	??	E	L	T	L	A	S	R	Q	Q	L

A. Elnaggar *et al.* IEEE Tr Pat. 2021

Large language models for proteins

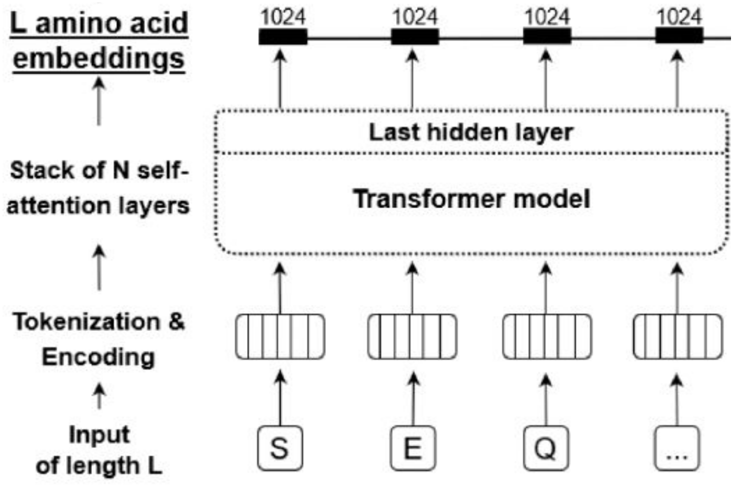


- High capacity transformers
- Input: single sequence (length L)
  Output: high dimensional embedding d x L
- Trained on hundreds of millions of protein sequences to reconstruct masked tokens



A. Elnaggar *et al.* IEEE Tr Pat. 2021

Large language models for proteins



- High capacity transformers
- Input: single sequence (length L)
  Output: high dimensional embedding d x L
- Trained on hundreds of millions of protein sequences to reconstruct masked tokens
- They can be used a zero-shot variant effect predictors through their log-odd ratios.

$$\log \frac{P(x^{mut})}{P(x^{wt})}$$

But they do not reach the state of the art.

A. Elnaggar *et al.* IEEE Tr Pat. 2021

Large language models for proteins

Limitation: they do not explicitly account for the evolutionary relationships between natural sequences.
Ways to overcome it:

- augmenting the input with alignments,
- extracting features from embeddings with supervision (3D structure, conservation, binary variant effect).

g d x L

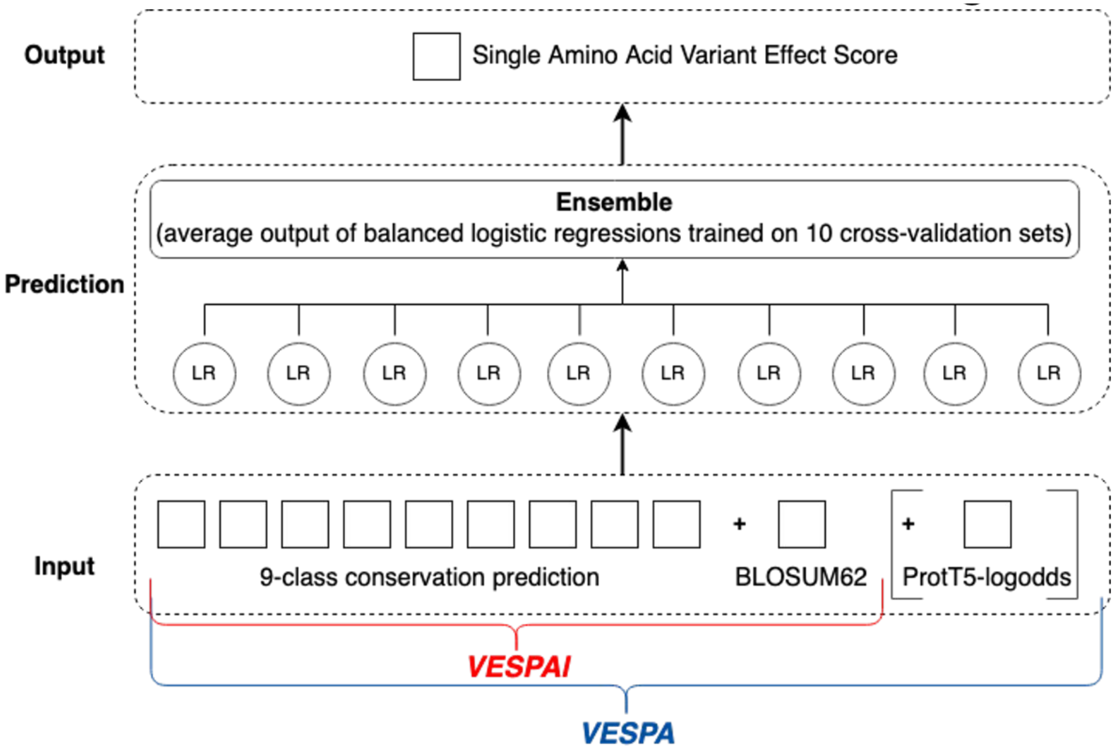
rotein
kens

t effect
ios.

But they do not reach the state of the art.

A. Elnaggar *et al.* IEEE Tr Pat. 2021

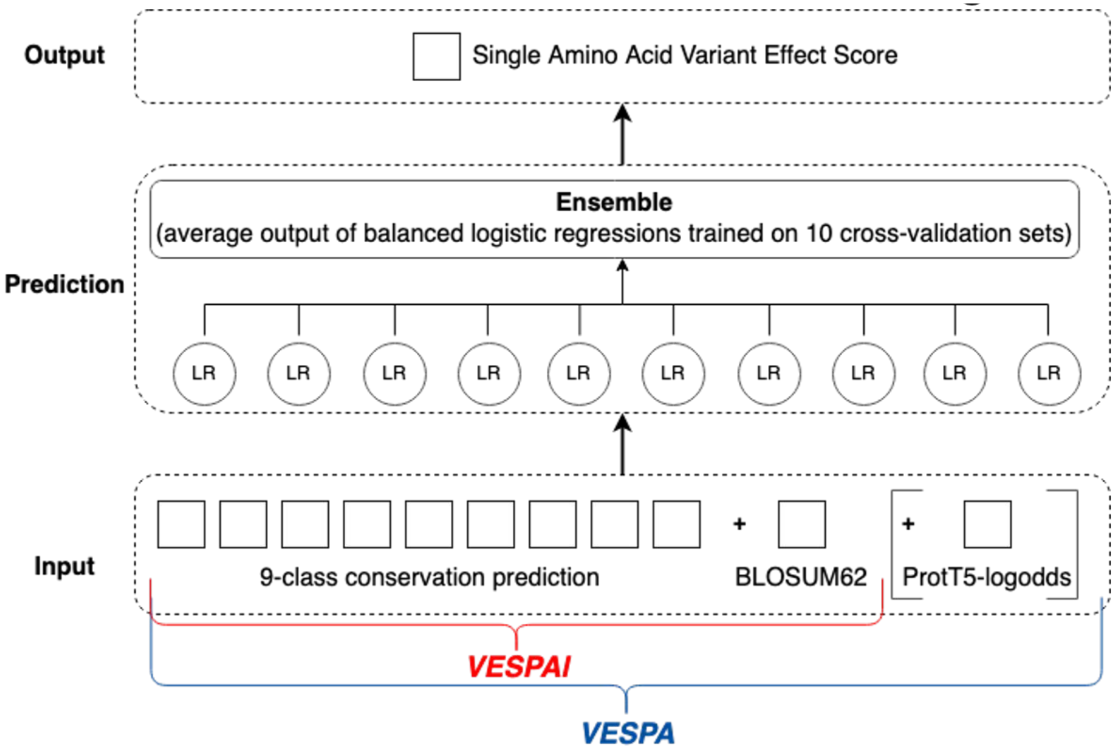
Variant Effect Score Prediction without Alignments



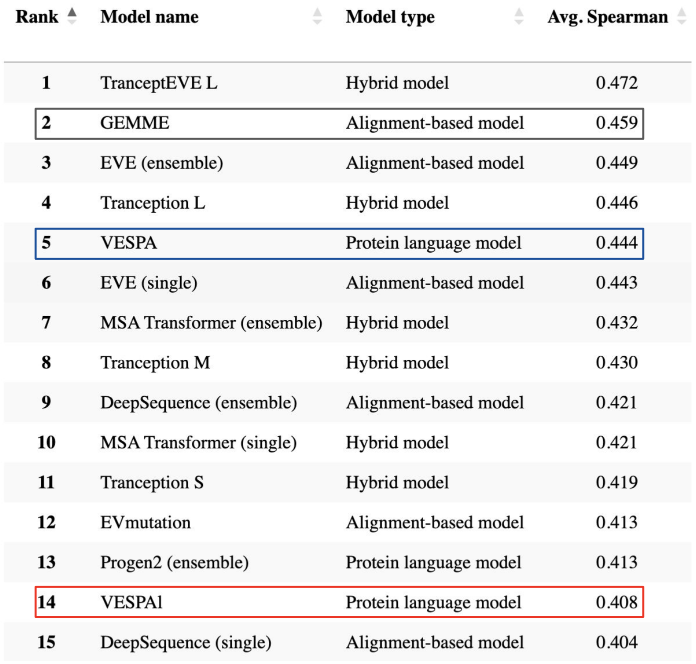
Output
Single Amino Acid Variant Effect Score
Prediction
Ensemble
(average output of balanced logistic regressions trained on 10 cross-validation sets)
LR
LR
LR
LR
LR
LR
LR
LR
LR
LR
Input
+
+
9-class conservation prediction
BLOSUM62
ProtT5-logodds
VESPAl
VESPA

C. Marquet *et al.* Human Genetics 2022

Variant Effect Score Prediction without Alignments


Output
Single Amino Acid Variant Effect Score
Prediction
Ensemble
(average output of balanced logistic regressions trained on 10 cross-validation sets)
LR
LR
LR
LR
LR
LR
LR
LR
LR
LR
Input
+
+
9-class conservation prediction
BLOSUM62
ProtT5-logodds
VESPAl
VESPA

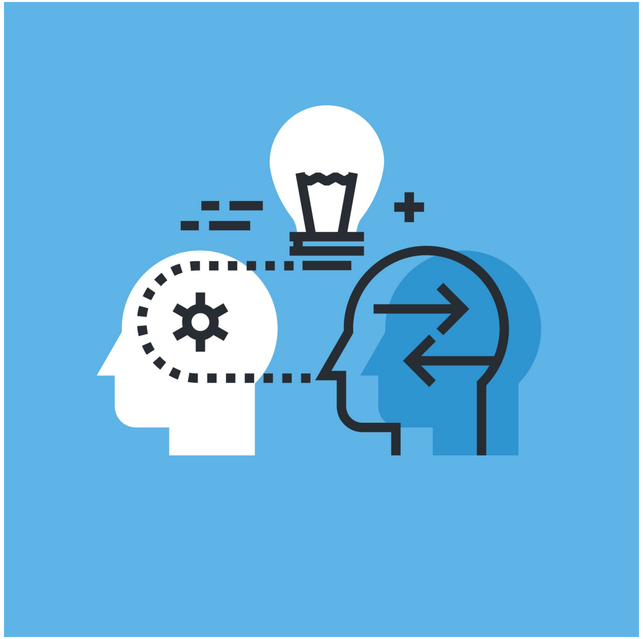
ProteinGym leaderboard



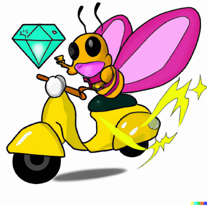
Rank	Model name	Model type	Avg. Spearman
1	TranceptEVE L	Hybrid model	0.472
2	GEMME	Alignment-based model	0.459
3	EVE (ensemble)	Alignment-based model	0.449
4	Tranception L	Hybrid model	0.446
5	VESPA	Protein language model	0.444
6	EVE (single)	Alignment-based model	0.443
7	MSA Transformer (ensemble)	Hybrid model	0.432
8	Tranception M	Hybrid model	0.430
9	DeepSequence (ensemble)	Alignment-based model	0.421
10	MSA Transformer (single)	Hybrid model	0.421
11	Tranception S	Hybrid model	0.419
12	EVmutation	Alignment-based model	0.413
13	Progen2 (ensemble)	Protein language model	0.413
14	VESPAl	Protein language model	0.408
15	DeepSequence (single)	Alignment-based model	0.404

C. Marquet *et al.* Human Genetics 2022

Mapping learnt representations to mutational landscape with an expert teacher



VespaG



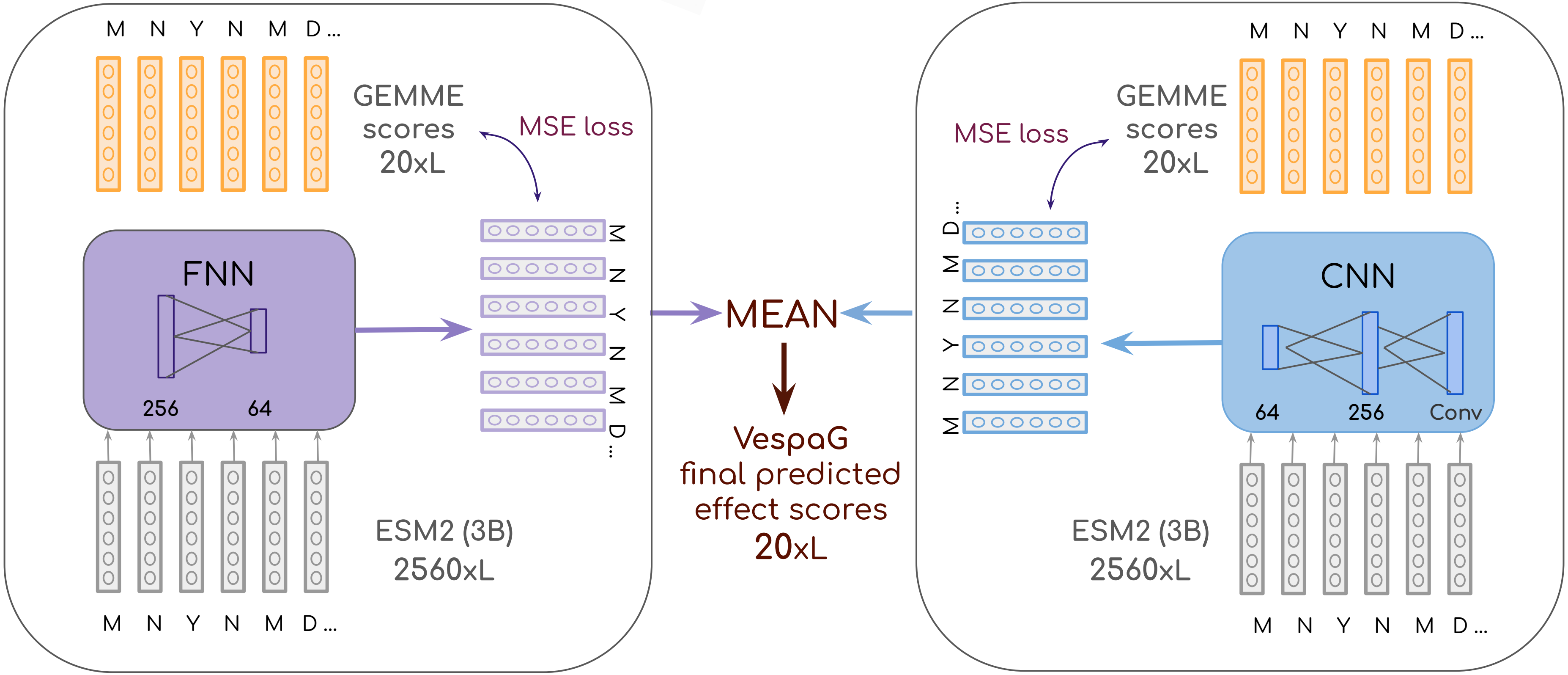
Main idea:

Directly mapping protein language model (pLM) embeddings to mutational landscapes, using an evolutionary-informed model (GEMME) as a teacher.

Advantages:

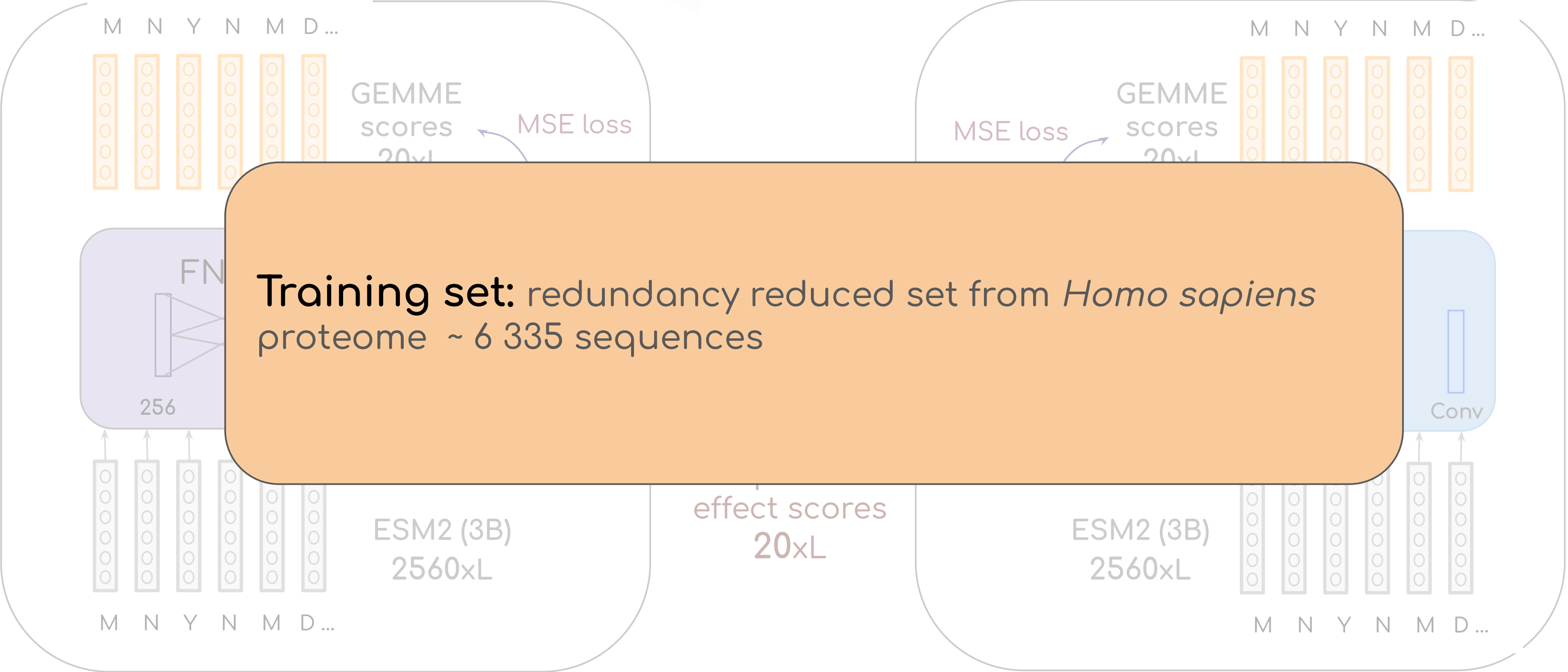
- Avoids the costly computation of log-odd ratios for all substitutions
- Largely increases the body of annotations
- Improves annotations' consistency

Architecture



Lin *et al.* Science 2023

Architecture

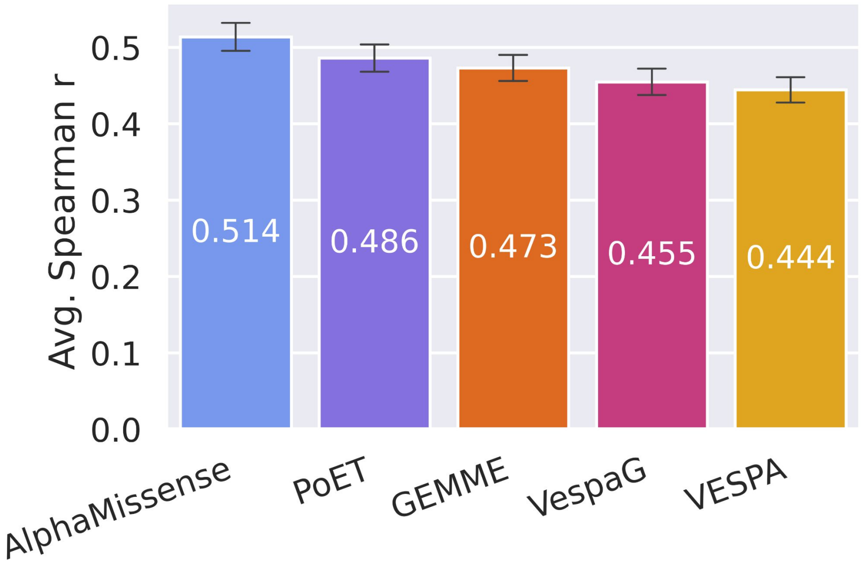


Training set: redundancy reduced set from *Homo sapiens* proteome ~ 6 335 sequences

Lin *et al.* Science 2023

Predictive performances

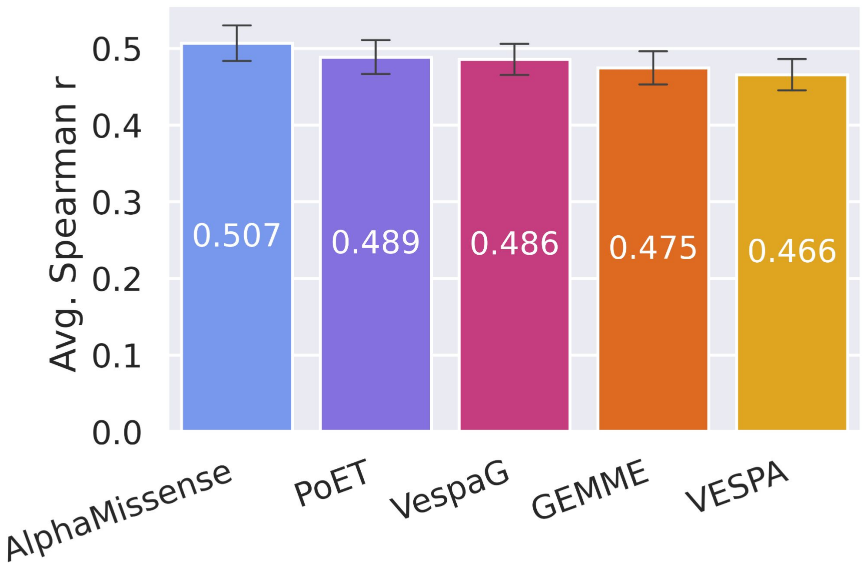
ProteinGym set (1.5M missense mutations)



VespaG achieves results similar to state-of-the-art methods.

Predictive performances

<u>ProteinGym non-viral</u> set (~1.4M missense mutations)



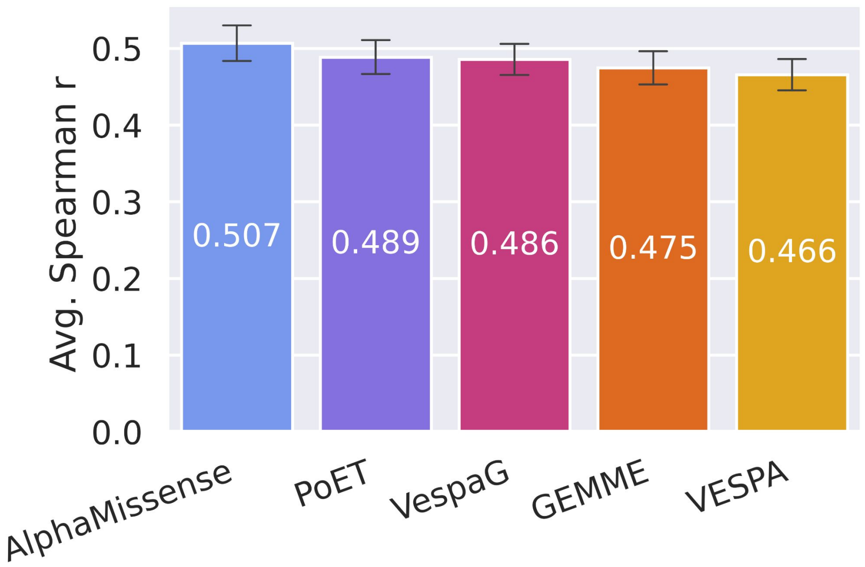
Performance increases when we disregard viral proteins.

In line with previous observations that pLMs do not behave well with viral sequences.

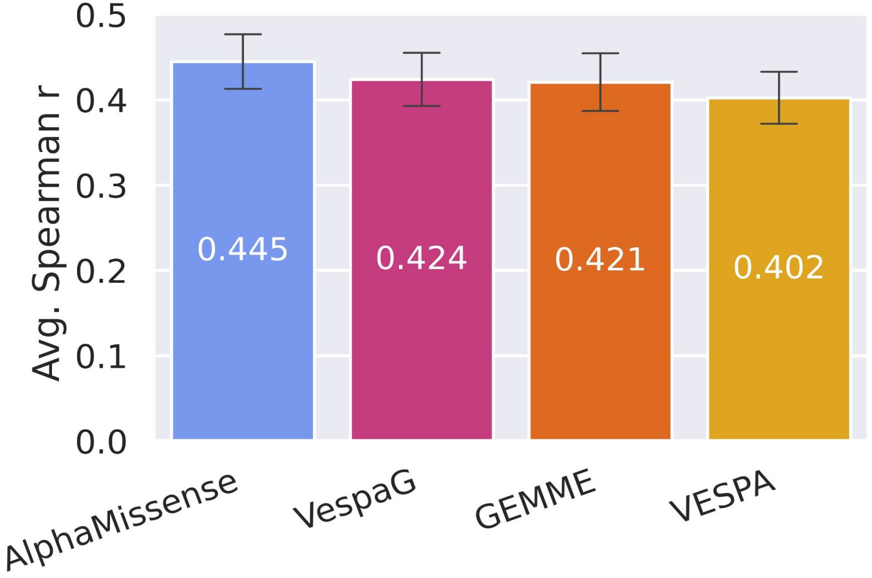
Cheng et *al.*

Predictive performances

ProteinGym non-viral set (~1.4M missense mutations)

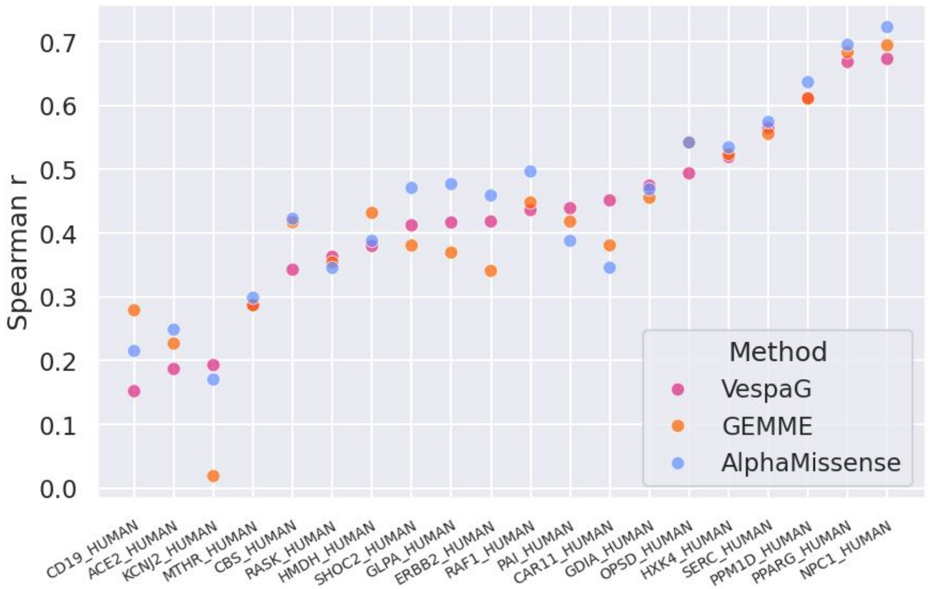


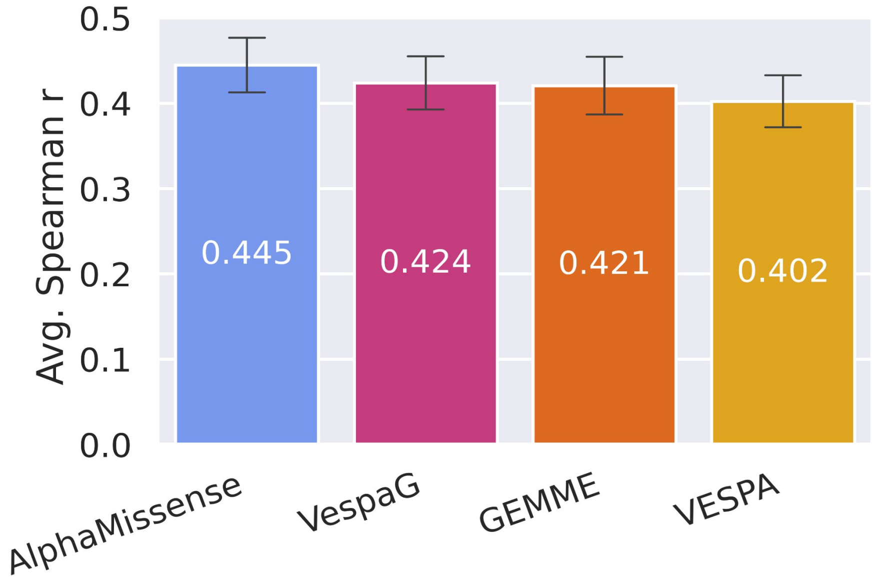
MaveHum23 23 DMS exp for 20 Human proteins (~266k SAVs) from Cheng et al. 2023



Predictive performances

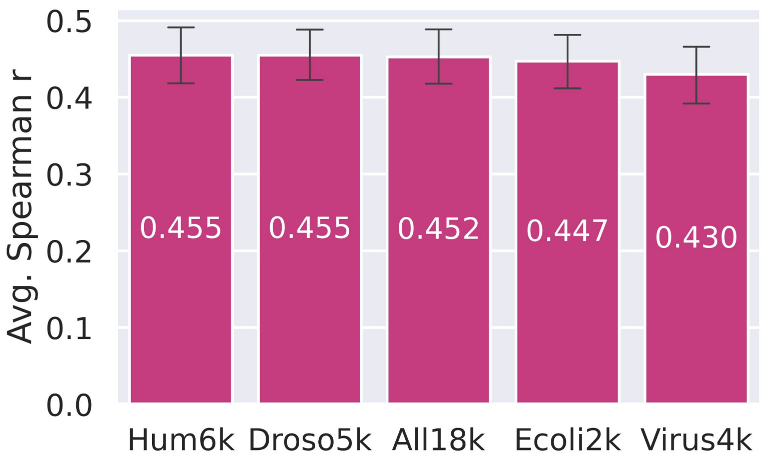
MaveHum23 23 DMS exp for 20 Human proteins (~266k SAVs) from from Cheng et al. 2023





Influence of the training set

Dataset	Hum6k	Droso5k	Ecoli2k	Virus4k	All18k
Organism	*H.sapiens*	*D.melanogaster*	*E.coli*	All viral in SwissProt[1]	All
#(proteins)	6 294	5 650	2 108	4 027	18 079

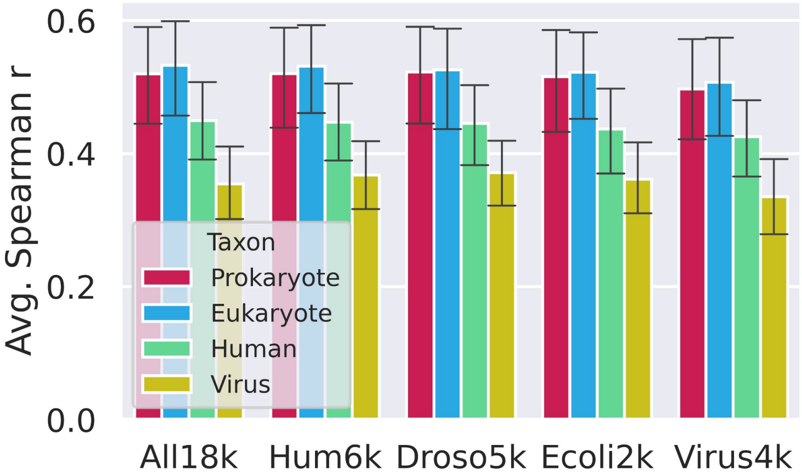


- The performance saturate after a few thousands training proteins.
- Training on a high-quality proteome from a model species suffices.

[1] UniProt Consortium 2023

Influence of the training set

Dataset	Hum6k	Droso5k	Ecoli2k	Virus4k	All18k
Organism	*H.sapiens*	*D.melanogaster*	*E.coli*	All viral in SwissProt[1]	All
#(proteins)	6 294	5 650	2 108	4 027	18 079

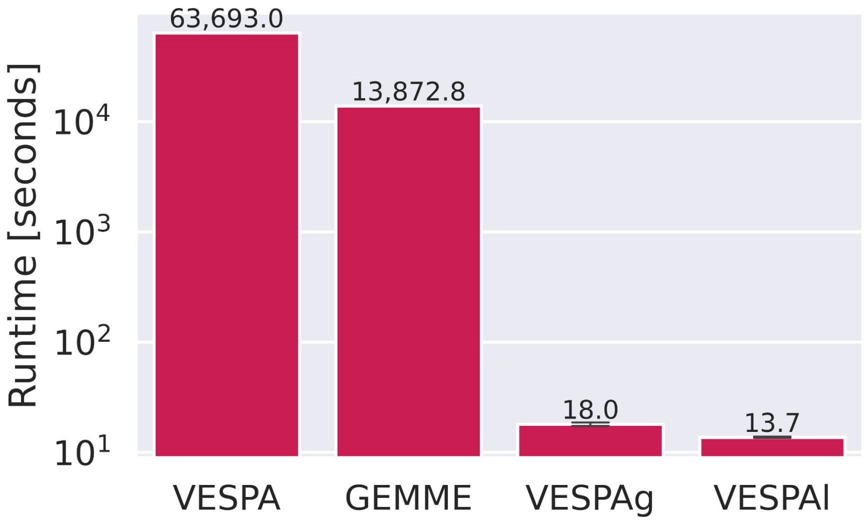


Training on viral sequences does not help for predicting viral variant effects.

[1] UniProt Consortium 2023

Runtime

runtime on ProtionGym benchmark
(87 proteins)



VespaG provides blazingly fast state-of-the-art variant effect predictions from single-sequence-derived pLM embeddings.

Measured @ 64G RAM & 32 CPU cores (+46G VRAM for VESPA), excluding embedding/MSA generation

Conclusions and perspectives

VespaG can…

- directly map pLM embeddings to mutational landscapes
- transfer knowledge across organisms
- produce accurate predictions of variant effects
- scan entire protomes within an hour

VespaG does not…

- deal well with viral sequences

=> needs further investigation to understand the relationship between predictive performance and the availability of homologous sequences.

TUM

ROSTLAB

LCQB

SCIENCES SORBONNE UNIVERSITÉ

IUF institut universitaire de France

Thank you!

SACADO
SERVICE D'AIDE AU CALCUL ET À L'ANALYSE DE DONNÉES
SORBONNE UNIVERSITÉ

cnrs

anr agence nationale de la recherche
AU SERVICE DE LA SCIENCE

ANR-20-CE44-0010

erc
European Research Council
Established by the European Commission