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

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

Alphabet of 20 amino acids

...LTGELTLASR QQLIDWMEAD KVGGPLLRSA...
Motivation

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

-250M protein sequences annotated in Uniprot

Next Generation Sequencing

Sequence Databases

Alphabet of 20 amino acids

Gain of function
Benign
Loss of function

LTGELLTLASR
QQLIHMEAD
KVGGPLLRSA

?
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 QQLIHWMEAD KVGGPLLRS... → Gain of function
Benign Loss of function

Fundamental biology
Bioengineering
Drug design
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.
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 per protein

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
- 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 per protein

Multiple mutation assays for 11 proteins

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

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.
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)
...

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)
...

SOTA methods leverage protein sequence information across species. A few also exploit population data.
Explicitly exploiting natural sequences evolutionary history
GEMME - an evolutionary-informed predictor

Input

Query-centered multiple sequence alignment (MSA)

Aligned homologous sequences

Output

Complete single-mutational landscape of the query

high impact
neutral

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

GEMME - an evolutionary-informed predictor

A measure of **conservation** accounting for the global context

Joint Evolutionary Trees
S. Engelen et al. PLOS CB 2009

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

Evolutionary distance to a natural sequence with the mutation

**Joint Evolutionary Trees**
S. Engelen *et al.* PLOS CB 2009

[http://www.lcqb.upmc.fr/GEMME/Home.html](http://www.lcqb.upmc.fr/GEMME/Home.html)
GEMME - scaling to entire proteomes

GEMME provides a clear readout of the input alignment.

- **ColabFold**
  - MMseqs2
  - Uniref100 + Env.
  - <25K

- **ProteinGym-MSA**
  - JackHMMer
  - Uniref100
  - <550K

- **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.

- **ColabFold**
  - MMseqs2
  - Uniref100 + Env.
  - \(<25K\)

- **ProteinGym-MSA**
  - JackHMMer
  - Uniref100
  - \(<550K\)

- **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.

**ColabFold**
- MMseqs2
- Uniref100 + Env.
- <25K

**ProteinGym-MSA**
- JackHMMer
- Uniref100
- <550K

**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.

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

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

GEMME provides a clear readout of the input alignment.

- **ColabFold**: MMseqs2 Uniref100 + Env. <25K
- **ProteinGym-MSA**: JackHMMer Uniref100 <550K
- **ProteinNet**: JackHMMer UniParc + Env. <1.4M
- **Pfam**: HMMer UniProtKB <300K

[https://doi.org/10.5061/dryad.vdncjsz1s](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 \( d \times L \)
- Trained on hundreds of millions of protein sequences to reconstruct **masked** tokens

```
L T ?? E L T L A S R Q Q L
```
Large language models for proteins

- High capacity **transformers**
- Input: single sequence (length L)
  Output: high dimensional embedding $d \times L$
- Trained on hundreds of millions of protein sequences to reconstruct **masked** tokens
Large language models for proteins

- High capacity transformers
- Input: single sequence (length $L$)
  Output: high dimensional embedding $d \times L$
- 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.
**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).

But they do not reach the state of the art.
Variant Effect Score Prediction without Alignments

Output
- Single Amino Acid Variant Effect Score

Ensemble
(average output of balanced logistic regressions trained on 10 cross-validation sets)

Prediction
- LR × 10

Input
- 9-class conservation prediction
- BLOSUM62
- ProtT5-logodds

VESPAI + VESPA
Variant Effect Score Prediction without Alignments

### ProteinGym leaderboard

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

C. Marquet et al. Human Genetics 2022
Mapping learnt representations to mutational landscape with an expert teacher
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
Lin et al. Science 2023
Training set: redundancy reduced set from *Homo sapiens* proteome ~ 6,335 sequences
Predictive performances

ProteinGym set (1.5M missense mutations)

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

ProteinGym non-viral 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.
Predictive performances

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Spearman r</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlphaMissense</td>
<td>0.507</td>
</tr>
<tr>
<td>PoET</td>
<td>0.489</td>
</tr>
<tr>
<td>VespaG</td>
<td>0.486</td>
</tr>
<tr>
<td>GEMME</td>
<td>0.475</td>
</tr>
<tr>
<td>VESPA</td>
<td>0.466</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Spearman r</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlphaMissense</td>
<td>0.445</td>
</tr>
<tr>
<td>VespaG</td>
<td>0.424</td>
</tr>
<tr>
<td>GEMME</td>
<td>0.421</td>
</tr>
<tr>
<td>VESPA</td>
<td>0.402</td>
</tr>
</tbody>
</table>
Predictive performances

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

![Graph showing predictive performances with Spearman r values and bar chart showing average Spearman r values for different methods: VespaG, GEMME, AlphaMissense, and VESPA. The graph illustrates the correlation of different methods with the actual data, while the bar chart compares their performance as measured by the average Spearman r values.](image-url)
Influence of the training set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Hum6k</th>
<th>Droso5k</th>
<th>Ecoli2k</th>
<th>Virus4k</th>
<th>All18k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism</td>
<td>H. sapiens</td>
<td>D. melanogaster</td>
<td>E. coli</td>
<td>All viral in SwissProt</td>
<td>All</td>
</tr>
<tr>
<td>#(proteins)</td>
<td>6 294</td>
<td>5 650</td>
<td>2 108</td>
<td>4 027</td>
<td>18 079</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Hum6k</th>
<th>Droso5k</th>
<th>Ecoli2k</th>
<th>Virus4k</th>
<th>All18k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism</td>
<td>H.sapiens</td>
<td>D.melanogaster</td>
<td>E.coli</td>
<td>All viral in SwissProt¹</td>
<td>All</td>
</tr>
<tr>
<td>#(proteins)</td>
<td>6 294</td>
<td>5 650</td>
<td>2 108</td>
<td>4 027</td>
<td>18 079</td>
</tr>
</tbody>
</table>

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

¹UniProt Consortium 2023
Runtime

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.