Language models enable zero-shot prediction of the effects of mutations on protein function

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, Alexander Rives

Goal

- To understand how mutation affects the function of the protein.
- Summarize the effects of mutation in a matrix

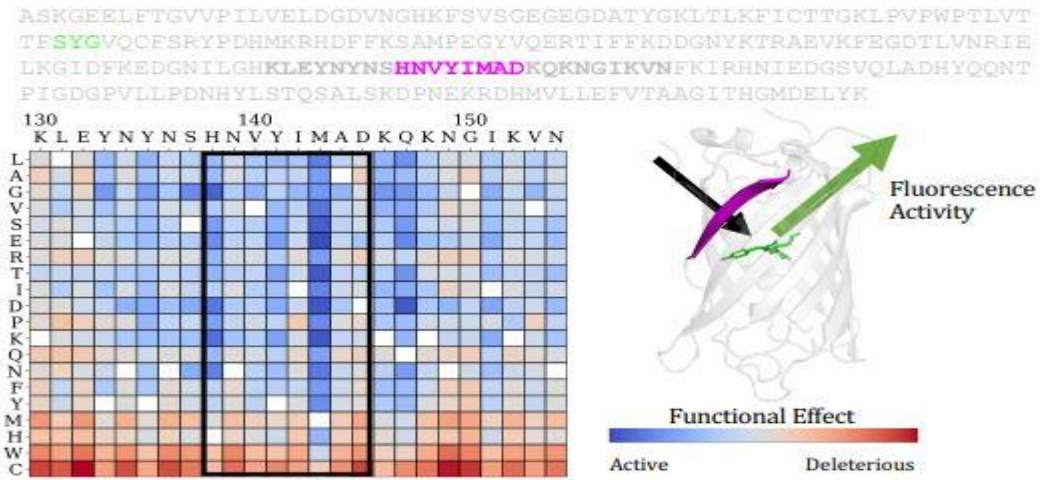


Figure 1: Depiction of a mutational effect prediction task. The objective is to score the effect of sequence mutations on the function of a protein. Deep mutational scanning experiments provide ground truth experimental measurements of the protein's function (fluorescence activity in the example here) for a large set of single mutations or combinations of mutations. For each protein, the prediction task is to score each possible mutation and rank its relative activity. Predictions for single substitutions can be described in a score matrix. The columns are the positions in the sequence. The rows are the possible variations at each position.

Previous works

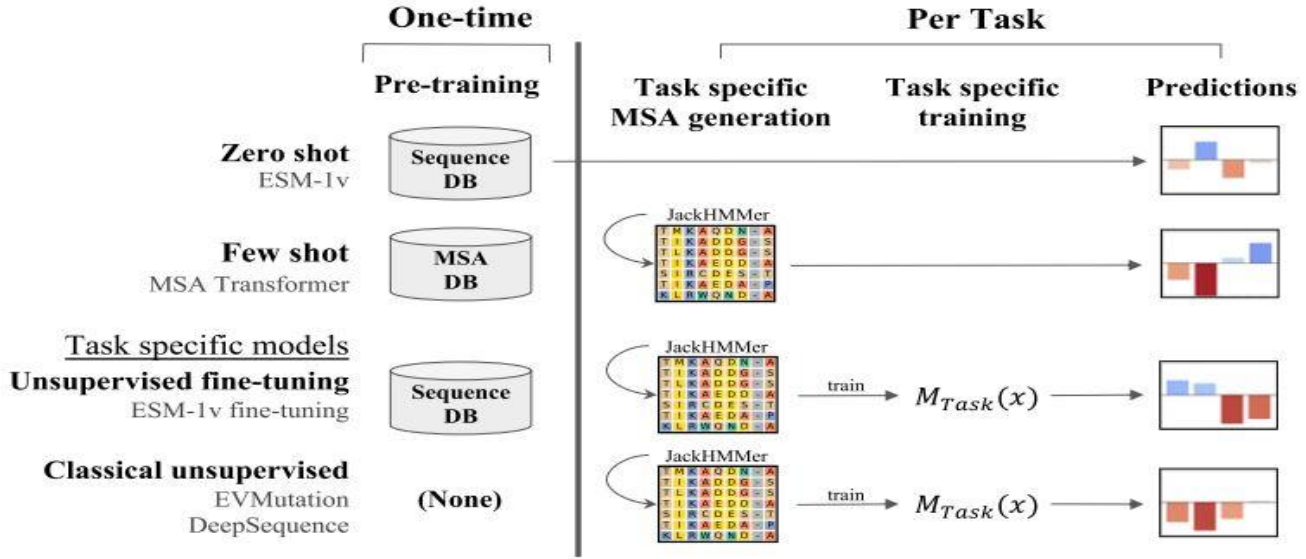


Figure 2: Steps involved in variant effect prediction methods. Compared with EVMutation [4] and DeepSequence [20], MSA Transformer and ESM-1v require no task-specific model training for inference. Moreover, ESM-1v does not require MSA generation.

Zero-shot and few-shot transfer

- Zero-shot is transfer of a model to a new task without any further supervision to specialize the model to the task
- Few-shot setting in which a few examples are given to the model as inputs at inference time
- Few shot and zero-shot setting, no gradient updates are performed to specialize the model
- The assumption is that in the pre-training stage, the model learns information relevant to the tasks to which it will later be transferred.
- The pre-training dataset includes sequences across evolution.
- Model is general purpose and can be applied across a variety of tasks without specialization

Few-shot Setting

Few-shot setting:

- MSA transformer is used
- MSA transformer is trained on a large database of MSAs using masked language modeling

  Original Sequence: K L E Y N Y N S H

  Masked Sequence: K L E [MASK] N Y N S H

- It takes an MSA as input during inference
- They provide the MSA a few example of the related protein from the same family
- Do not perform any additional training
- MSA transformer can perform effectively in this setting

Zero-shot setting

- They train ESM-1v transformer-based protein language model for prediction of variant effects
- ESM-1v contains 650M parameters
- Model is trained only on sequences
- Uniref90 data set which is large databases of unaligned and unrelated protein sequences
- Uniref90 contains 98 million diverse protein sequences across evolution
- Employing the ESM-1b architecture and masked language modeling approach
- Train five models with different seeds to produce an ensemble

Method

- Protein language model learn the information necessary to solve a task from pre-training
- They can be applied directly to new instances of the task, without specialization
- Masked language modeling objective output the probability that an amino acid occurs at a position in a protein given the surrounding context
- Use this capability to score sequence variations
- Score mutations using the log odds ratio at the mutated position
- For a given mutation, consider the amino acid in the wildtype sequence of protein as a reference state
- Comparing the probability assigned to the mutated amino acid with the probability assigned to the wildtype
- Assuming an additive model when multiple mutations T exist in the same sequence:

$$\sum_{t \in T} \log p(x_t = x_t^{mt} | x_{\setminus T}) - \log p(x_t = x_t^{wt} | x_{\setminus T}) \quad (1)$$

Dataset and Evaluation

Dataset:

- Deep mutational scanning experiments provide ground truth experimental measurements of the protein's function
- Deep mutational scanning experiments measure the effects of mutations on a single protein
- A set of 41 deep mutational scanning datasets
- Treat each deep mutational scanning dataset as a separate prediction task
- They study zero-shot and few-shot transfer of protein language models using this data.

Evaluation:

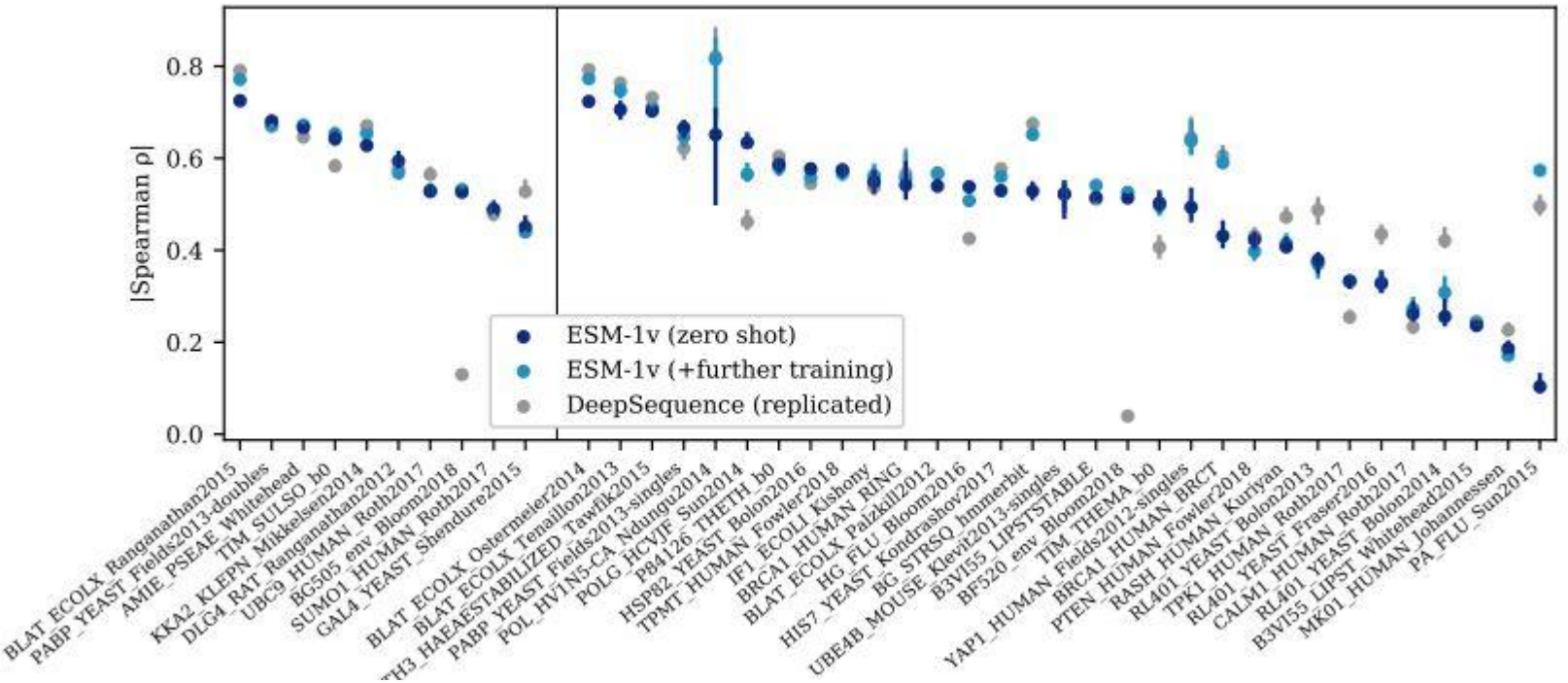
- They use absolute Spearman which captures the correlation between a ranked variable in the prediction and the ground truth

Results

Models	Full	Test
PSSM	0.460	0.460
EVMutation (published)	0.508	0.495
EVMutation (replicated)	0.511	0.498
DeepSequence (published)	0.514	0.499
DeepSequence (replicated)	0.520	0.506
MSA Transformer	0.542	0.524
ESM-1v (zero shot)	0.509	0.482
ESM-1v (+further training)	0.538	0.519

Table 1: Comparison of protein language models to state-of-the-art methods. Average |Spearman ρ| on full and test sets. DeepSequence and ESM-1v models are each ensembles of 5 models. MSA Transformer is a single model, but is ensembled across 5 random samples of the MSA.

Results


|Spearman ρ|
0.0
0.2
0.4
0.6
0.8
ESM-1v (zero shot)
ESM-1v (+further training)
DeepSequence (replicated)
BLAT_ECOLX_Ranganathan2015
PABP_YEAST_Fields2013-doubles
AMIE_PSEAE_Whitehead
TIM_SULSO_b0
KKA2_KLEPN_Mikkelsen2014
DLG4_RAT_Ranganathan2012
UBC9_HUMAN_Roth2017
BG505_env_Bloom2018
SUMO1_HUMAN_Roth2017
GAL4_YEAST_Shendure2015
BLAT_ECOLX_Ostermeier2014
BLAT_ECOLX_Tenaillon2013
HAEAESTABILIZED_Tawfik2015
PABP_YEAST_Fields2013-singles
POL_HV1N5-CA_Ndungu2014
POLG_HCVJF_Sun2014
P84126_THETH_b0
HSP82_YEAST_Bolon2016
TPMT_HUMAN_Fowler2018
IF1_ECOLI_Kishony
BRCA1_HUMAN_RING
BLAT_ECOLX_Palzkill2012
HG_FLU_Bloom2016
HIS7_YEAST_Kondrashov2017
UBE4B_MOUSE_Klevit2013-singles
B3VI55_LIPSTSTABLE
BF5200_env_Bloom2018
TIM_THEMA_b0
YAP1_HUMAN_Fields2012-singles
BRCA1_HUMAN_BRCT
PTEN_HUMAN_Fowler2018
RASH_HUMAN_Kuriyan
RL401_YEAST_Bolon2013
TPK1_HUMAN_Roth2017
RL401_YEAST_Fraser2016
CALM1_HUMAN_Roth2017
RL401_YEAST_Bolon2014
B3VI55_LIPST_Whitehead2015
MK01_HUMAN_Johannessen
PA_FLU_Sun2015

Results

Models	Full	Test
UniRep	0.156	0.151
TAPE	0.171	0.175
ProtBERT-BFD	0.428	0.399
ESM-1b	0.459	0.424
ESM-1v†	0.484	0.457
ESM-1v*	0.509	0.482

Table 2: Zero-shot performance. Average |Spearman ρ| on full and test sets. †Average performance of five ESM-1v models. *Ensemble of the five ESM-1v models.

Pre-training Process

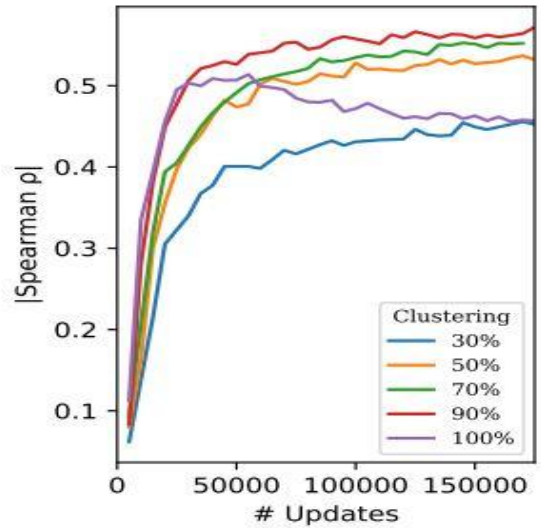


Figure 4: Comparison of pre-training datasets. Average |Spearman ρ| on the single-mutation validation set. While a 50% clustering threshold was used for ESM-1b, training with 90% clustering results in a significant improvement on variant prediction tasks. Notably, models trained on Uniref100, the largest dataset in this figure, appear to deteriorate early in training. These results establish a link between model performance and the data distribution, and highlight the importance of training data in the design of protein language models.

Functional Effects of Mutation

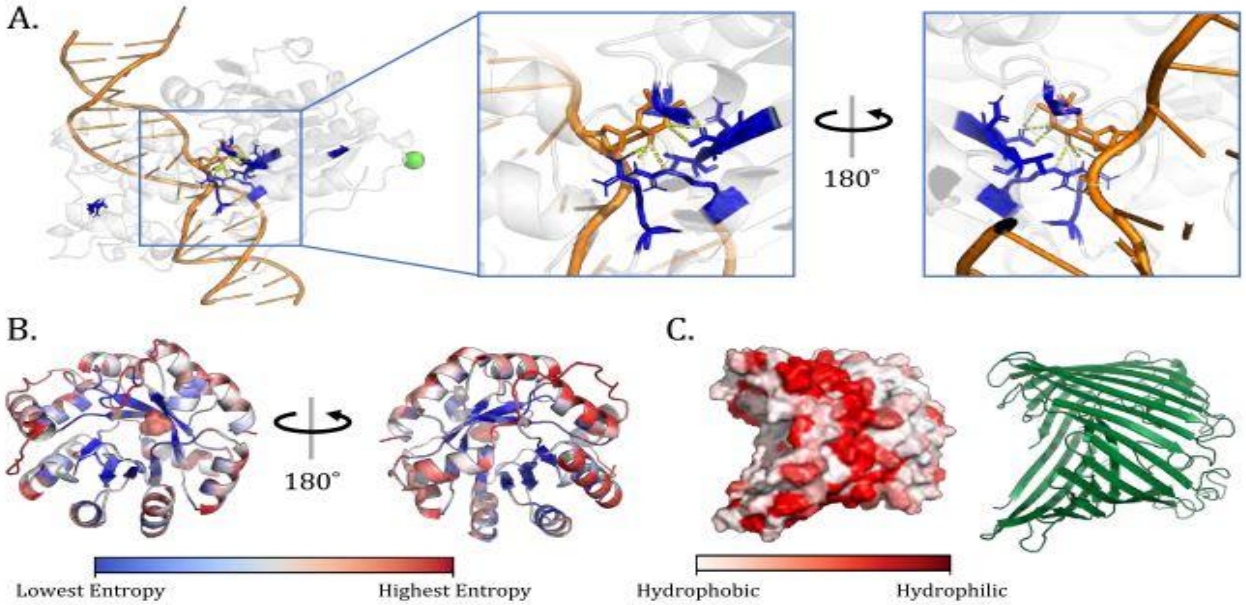


Figure 5: ESM-1v reflects the molecular basis of function in proteins. **(A)** DNA methylase HaeIII (pdbid: 1DCT [29]). Side chains for the top 10 positions with lowest prediction entropy shown in blue. Low-entropy positions cluster in the active site. **(B)** TIM Barrel (pdbid: 1IGS [30]) with residues colored by entropy. The model's predictions for residues on the surface have highest entropy (red) while those in the core have lower entropy (blue). Notably, residues on the alpha helices show a clear gradient from high to low entropy as residues transition from surface-facing to core-facing. **(C)** Sucrose-specific Porin (pdbid: 1A0T [31]), a transmembrane protein. The model predicts a hydrophobic band where the protein is embedded in the membrane.

Calibration

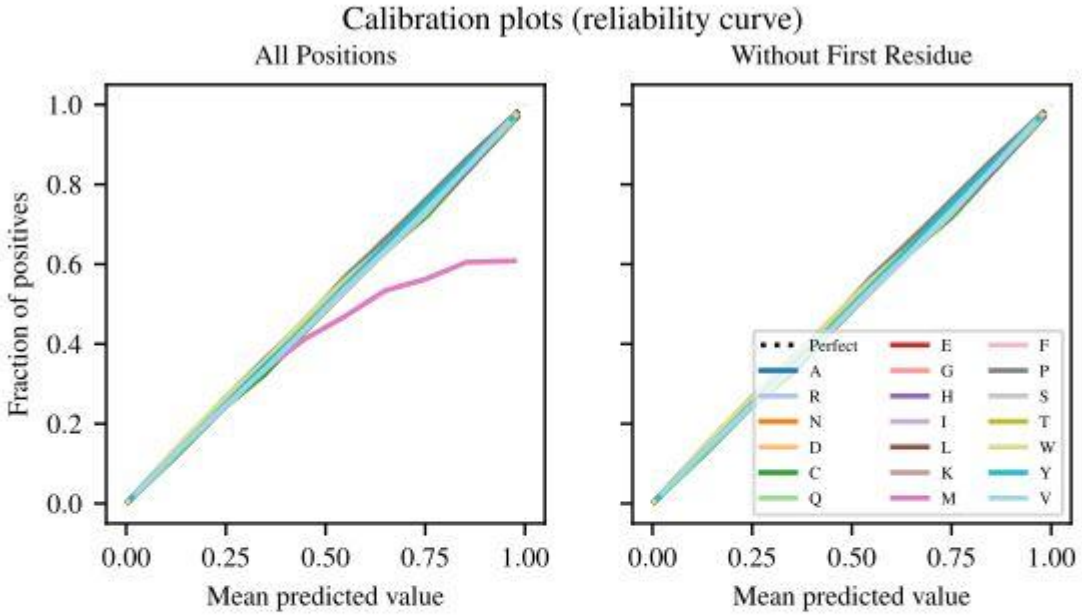


Figure 6: Calibration plot for ESM-1v predictions on each of the 20 naturally occurring amino acids on the trRosetta dataset. The multi-class classification is converted into a set of 20 one-versus-all classifications for the purpose of this analysis. Left and right plots show calibration of all positions and positions excluding the first residue, respectively. Since full sequences always start with Methionine, the model overwhelmingly predicts it in the first position. When evaluating the model on subsequences, such as those in the trRosetta dataset, this causes a miscalibration at the first residue. Including the first residue, the model has an average calibration error (ACE) of 0.011 in the first case and 0.006 in the second.

Thoughts

- They have performed several ablation analysis
- The novelty is that they are the first to propose an unsupervised method for mutation effects prediction
- Very helpful because in many cases the information is not available for proteins
- However a naive fine tuning the model may lead to overfitting
- In supplementary they talk about the spiked fine tuning of the proposed language