Evaluating the Functional Information Content of Protein Structure Prediction Model Embeddings using Deep Mutational Scanning Data

Laura McDonnell[1] & David Antolick[2]

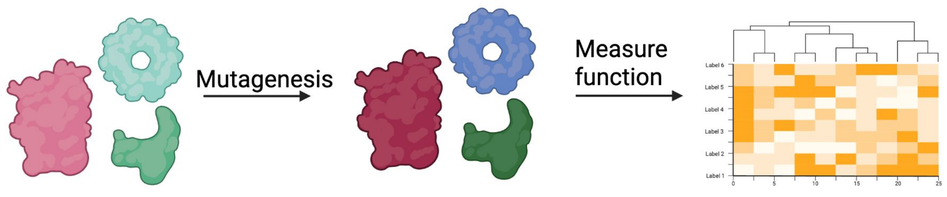
[1]Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University

[2]Department of Computational and Systems Biology, University of Pittsburgh

Background

Deep Mutational Scanning (DMS) provides rich datasets measuring the effects of single-point mutations on protein function. And, protein language models such as ESM-2 learn high-dimensional representations (embeddings) of protein sequences.

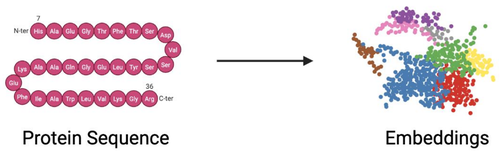
Deep Mutational Scanning (DMS) Concept



Do ESM-2 embeddings contain information predictive of the functional consequences of mutations?

Understanding this relationship could inform future work in protein engineering and mutation effect prediction.

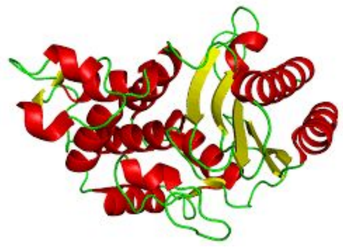
ESM-2 Embeddings



Data

We retrieved DMS scores measuring the functional impact of single amino acid substitutions. Wild-type embeddings were extracted using the pretrained ESM-2 model.

Protein	TEM-1 β-lactamase
Mutation Data	MaveDB **mavedb:00000070-a-1** (18,081 total variants)
Model	ESM-2 Via HuggingFace Transformers



Wild-type structure of TEM-1 β-lactamase[4]

Methods

Raw Data

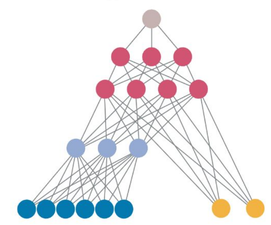
MAVEDB

hgvs_nt	hgvs_splice	hgvs_pro	score	err	ambler
c.861A>T	NA	p.Ter287Tyr	NA	NA	291
c.861A>C	NA	p.Ter287Tyr	0.851	0.377	291
c.860_861delinsGG	NA	p.Ter287Trp	NA	NA	291
c.859_861delinsGTT	NA	p.Ter287Val	NA	NA	291

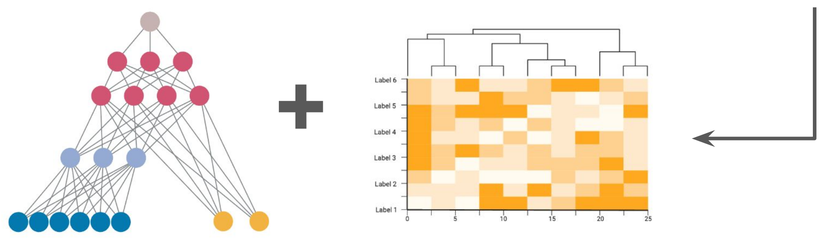
Parse and Filter

```
>mut_L146D_score0.001
AHPETLVKVKDAEDQLGARVGYIELDLNSGKILESFRPEERFPMMSTFKVL
>mut_A132R_score0.001
AHPETLVKVKDAEDQLGARVGYIELDLNSGKILESFRPEERFPMMSTFKVL
>mut_F58N_score3.111
```

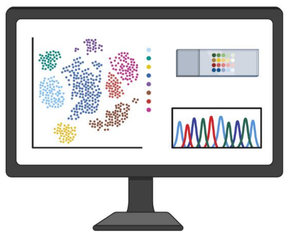
Embedding Extraction



Merge with Mutation Data



Feature Derivation

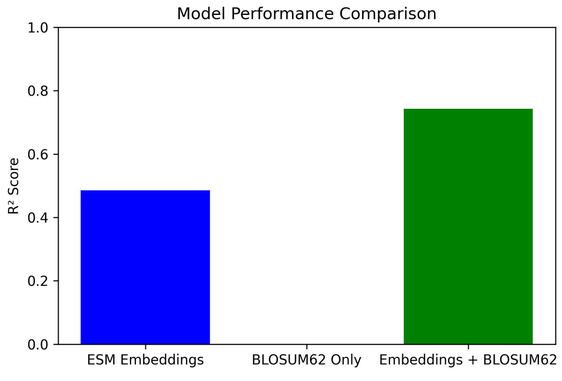


Correlation Analysis & ML Modeling

Results

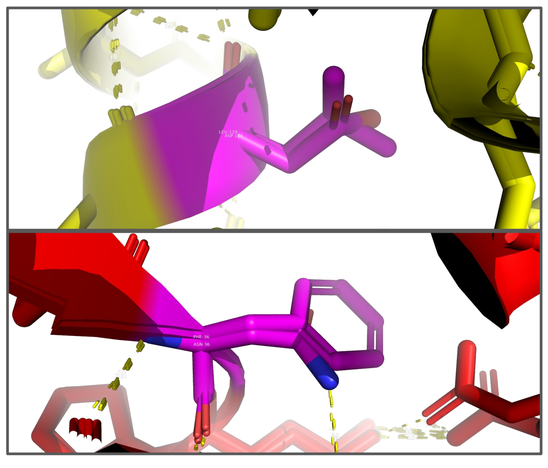
Model Performance in Predicting Mutation Functional Scores

Task	Description	Value
A	Spearman correlation (embedding norm vs DMS score)	~0.27
B	Random Forest R^2 (embeddings only)	~0.45
C	Spearman correlation (BLOSUM62 vs DMS score)	~0.46
D	Random Forest R^2 (embeddings + BLOSUM62)	~0.73



Random Forest models trained on embeddings or BLOSUM62 substitution scores alone achieve moderate performance. However, combining both feature sets significantly improves predictive accuracy, suggesting complementary predictive signals in learned embeddings and evolutionary substitution scores.

Structural Visualization of Mutations with Reduced pLDDT Confidence



L146D
Leucine > Aspartate

F58N
Phenylalanine > Asparagine

Conclusions

- Protein language model embeddings, even extracted from the wild-type sequence, encode information related to mutation functional impact.
- Although embedding features alone offer moderate predictive power, combining them with classical features like BLOSUM62 scores significantly improves performance.
- Future directions include embedding mutant sequences directly, leveraging structural embeddings, and exploring fine-tuning approaches.

References

[1] Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*.
[2] Lin, Z., et al. (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*.
[3] Esposito, D., et al. (2019). MaveDB: an open-source platform for massive assay data. *Genome Biology*.
[4] Bidmon, Katrin & Reina, Guido & Bös, Fabian & Pleiss, Juergen & Ertl, Thomas. (2007). Time-Based Haptic Analysis of Protein Dynamics. 537-542. 10.1109/WHC.2007.115.