# Predicting Protein Variant Effects and Epistasis Using ESM Protein Language Models

A Comparative Study of Probabilistic and Geometry-Based Scoring

Amyloid Precursor Protein (APP / A4_HUMAN) — Seuma et al. 2021

Prithvirajan R  Adina Nadeem  Fazil Tarlan

February 2026

**Abstract**

Protein language models (pLMs) trained on millions of evolutionary sequences show promise for predicting the functional effects of amino acid substitutions. However, it remains unclear which scoring paradigm—probability-based or geometry-based—better captures the link between sequence variation and biological phenotype. We systematically compare six scoring methods across three ESM model architectures on a deep mutational scanning (DMS) dataset of 14,483 experimentally characterised variants of the Amyloid Precursor Protein (APP). The probabilistic methods include Pseudo-Log-Likelihood (PLL), Masked Log-Likelihood Ratio (MLLR), Entropy-Weighted MLLR, Mutant-Only Marginal, and Ensemble MLLR. As an alternative, we test Embedding Distance Scoring (EDS), a geometry-based method that measures displacement in latent space. Our results show that MLLR and PLL with ESM-2 (650M) achieve the highest correlations ($\rho \approx 0.43$), while EDS achieves $\rho = 0.39$ at 500–1000$\times$ lower cost. Neither method reliably predicts epistasis, and prediction accuracy does not differ significantly between single and double mutants.

# Contents

# 1. Introduction

## 1.1 Motivation

Engineering proteins with improved or novel function is central to biotechnology. The key challenge is navigating the protein **fitness landscape**—a vast, rugged surface mapping every possible sequence to its biological function. A 100-residue protein has $20^{100}$ possible sequences, and experimental approaches like deep mutational scanning (DMS) cost upwards of $50,000 per experiment. Computational prediction using machine learning offers an attractive alternative: score variants *in silico* and validate only the most promising candidates experimentally.

**Protein language models** (pLMs) are transformer-based neural networks trained on large databases of natural protein sequences via masked language modelling. By learning to predict masked amino acids from context, they implicitly encode evolutionary constraints governing protein structure and function. This study asks: **which scoring paradigm applied to these models is most reliable for predicting variant effects?**

## 1.2 Models Tested

We evaluate three models from the ESM family (Meta AI Research):

Table 1: ESM models used in this study.

| Model | Parameters | Layers | Training Data | Purpose |
|---|---|---|---|---|
| ESM-2 (150M) | 150 million | 30 | UniRef50 | Fast baseline |
| ESM-2 (650M) | 650 million | 33 | UniRef50 | Large generalist |
| ESM-1v (650M) | 650 million | 33 | UniRef90 | Variant specialist |

## 1.3 The A4_HUMAN Dataset

We selected the A4_HUMAN_Seuma_2021 dataset from ProteinGym—a DMS scan of the **Amyloid Precursor Protein (APP)**, whose misprocessing causes amyloid-$\beta$ plaque formation in Alzheimer's disease. The dataset was chosen for its human origin, clinical relevance, and exceptionally rich double-mutant coverage enabling epistasis analysis.

Table 2: Dataset statistics.

| | |
|---|---|
| Total variants | 14,483 |
| Single mutants | 468 |
| Double mutants | 14,015 |
| DMS score range | $[-6.27, +3.33]$ |
| Phenotype | APP processing / toxicity |

## 1.4 Hypotheses

- **H1:** pLMs predict single-mutant effects with significant correlation to experimental DMS scores.
- **H2:** Probabilistic scoring (PLL, MLLR) is the most reliable paradigm; geometry-based scoring (EDS) is an alternative worth testing.
- **H3:** Larger models improve predictions regardless of scoring method.
- **H4:** Accuracy is significantly higher for single mutants than for doubles.

**H5:** The method with best epistasis prediction is most trustworthy for landscape exploration.

# 2. Methods

## 2.1 Scoring Methods

We evaluate six methods grouped into two paradigms.

### 2.1.1 Masked Log-Likelihood Ratio (MLLR)

The standard approach in the pLM literature. Each mutated position is masked, and the model predicts a probability distribution over amino acids: $\text{MLLR} = \log P(x_{\text{mut}} \mid x_{\setminus i}) - \log P(x_{\text{wt}} \mid x_{\setminus i})$. Requires one forward pass per mutant. For multi-site mutations, scores are summed.

### 2.1.2 Pseudo-Log-Likelihood (PLL)

Evaluates every position: $\text{PLL}(x) = \sum_{i=1}^{L} \log P(x_i \mid x_{\setminus i})$. Measures overall sequence "naturalness." Requires $L$ forward passes ($L \approx 700$ for APP), making it $\sim$700$\times$ slower than MLLR. We also compute the Log-Likelihood Ratio (LLR = PLL(mut) $-$ PLL(WT)), which produces identical rankings.

### 2.1.3 Entropy-Weighted MLLR

Weights each position by inverse entropy: $\text{Score} = \text{LLR}_i/(H(x_i) + \varepsilon)$, penalising mutations at conserved (low-entropy) positions more heavily.

### 2.1.4 Mutant-Only Marginal

Absolute fitness: $\text{Score} = \log P(x_{\text{mut}} \mid x_{\setminus i})$—no wild-type comparison.

### 2.1.5 Ensemble MLLR

Averages MLLR over multiple passes with stochastic neighbour masking, testing context robustness.

### 2.1.6 Embedding Distance Scoring (EDS)

Our alternative, geometry-based approach: $\text{EDS}(x) = -\|\mathbf{H}_{\text{mut}} - \mathbf{H}_{\text{wt}}\|_2$, where $\mathbf{H} \in \mathbb{R}^{L \times D}$ is the final-layer hidden state tensor. Rather than asking "is this sequence evolutionarily plausible?", EDS asks "how much did this mutation change the protein's meaning in latent space?" (**Latent Manifold Hypothesis**). Requires only 2 forward passes per variant.

## 2.2 Epistasis Analysis

Epistasis captures non-additive mutation interactions: $E = \text{Score}(AB) - \text{Score}(A) - \text{Score}(B) + \text{Score}(\text{WT})$.

For PLL (absolute scores $\approx -400$), this formula requires a WT-baseline correction to avoid double-counting; our pipeline auto-detects and corrects this. For EDS (WT $\approx 0$), no correction is needed.

## 2.3 Landscape Exploration

We generated 4,000 synthetic mutants (1,000 each at $k = 2, 5, 10, 20$ substitutions) and scored them with EDS and PLL to measure fitness decay across mutational distance.

### 2.4 Technical Setup

All scoring ran on the RAVEN HPC cluster (MPCDF) with NVIDIA A100 GPUs. Wall times ranged from ∼30 min (EDS) to ∼24 h (PLL 650M). CSV-based checkpointing ensured robustness. Statistical tests use Spearman's $\rho$, permutation tests ($n = 10{,}000$), and bootstrap 95% confidence intervals.

## 3. Results

### 3.1 Initial Validation (Phase 1)

As a sanity check, we ran MLLR with ESM-2 (150M) on 468 single mutants (CPU only).

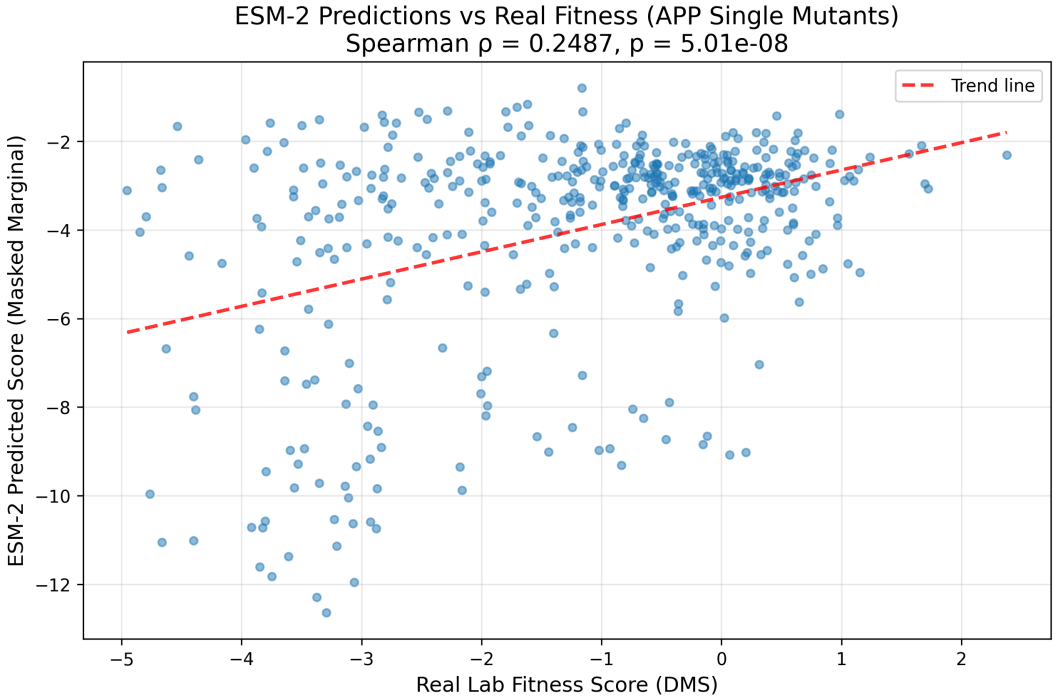



Figure 1: Initial validation: ESM-2 (150M) Masked Marginal scores vs. experimental DMS fitness for 468 single mutants ($\rho = 0.25$, $p = 5 \times 10^{-8}$). This confirmed the pipeline works and justified scaling up.

### 3.2 Comprehensive Benchmark (18 Configurations)

We ran all six scoring methods across all three models on the full 14,483-variant dataset.

Table 3: Full benchmark leaderboard sorted by Spearman $\rho$ within each model group. Run times on NVIDIA A100.

| Model | Method | Spearman $\rho$ | Top-100 Prec. | Run Time |
|---|---|---|---|---|
| *ESM-2 (650M)* | | | | |
| ESM-2 (650M) | MLLR | **0.427** | 57% | ∼4 h |
| ESM-2 (650M) | PLL | 0.425 | 69% | ∼24 h |
| ESM-2 (650M) | Ensemble MLLR | 0.423 | 39% | ∼24 h |
| ESM-2 (650M) | Entropy-MLLR | 0.420 | 5% | ∼5 h |
| ESM-2 (650M) | EDS | 0.388 | **93%** | ∼40 min |
| ESM-2 (650M) | Mutant Marginal | 0.365 | 31% | ∼5 h |
| *ESM-2 (150M)* | | | | |
| ESM-2 (150M) | Entropy-MLLR | 0.385 | 47% | ∼1 h |
| ESM-2 (150M) | MLLR | 0.371 | 55% | ∼1 h |
| ESM-2 (150M) | Ensemble MLLR | 0.363 | 49% | ∼5 h |
| ESM-2 (150M) | PLL | 0.339 | 73% | ∼16 h |
| ESM-2 (150M) | EDS | 0.270 | 87% | ∼10 min |
| ESM-2 (150M) | Mutant Marginal | 0.247 | 40% | ∼1 h |
| *ESM-1v (650M)* | | | | |
| ESM-1v (650M) | EDS | 0.339 | 89% | ∼40 min |
| ESM-1v (650M) | Ensemble MLLR | 0.329 | 82% | ∼5 h |
| ESM-1v (650M) | Mutant Marginal | 0.292 | 77% | ∼1 h |
| ESM-1v (650M) | MLLR | 0.290 | 48% | ∼1 h |
| ESM-1v (650M) | PLL | 0.236 | 76% | ∼20 h |
| ESM-1v (650M) | Entropy-MLLR | 0.228 | 86% | ∼1 h |

#### 3.2.1 Probabilistic Methods Dominate

The top five configurations by Spearman $\rho$ are all probabilistic, with ESM-2 (650M). MLLR ($\rho = 0.427$) and PLL ($\rho = 0.425$) are nearly identical, despite PLL being ∼700× more expensive. This suggests the information at mutation sites dominates, and the global context PLL provides is largely redundant. **MLLR should be preferred over PLL for practical applications.**

#### 3.2.2 Model Size Consistently Helps

Scaling from 150M to 650M improves all methods. EDS benefits most (+0.118), followed by PLL (+0.086) and MLLR (+0.056). ESM-1v, despite equal size to ESM-2 (650M), consistently underperforms—likely because its UniRef90 training data introduces noise for this specific protein.

#### 3.2.3 EDS as a Practical Alternative

While probabilistic methods achieve the highest Spearman $\rho$, EDS dominates the **Top-100 Precision** metric across all models (93% for ESM-2 650M, 89% for ESM-1v, 87% for ESM-2 150M). This means that if one selects the top 100 variants ranked by EDS, over 87% are true positives in the DMS assay. Combined with ∼500× faster computation than PLL, EDS is the clear winner for practical library design and screening applications.

## 3.3 Benchmark Report Plots

Each benchmark report shows three panels: a prediction scatter plot of model score versus experimental DMS fitness (left), score distributions separated by functional class (centre), and a ROC curve for binary classification (right).

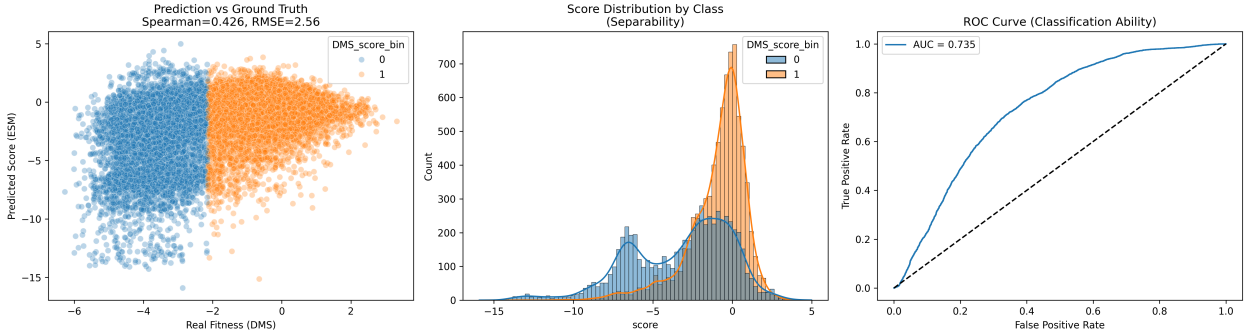



Figure 2: **ESM-2 (650M) + MLLR** ($\rho = 0.427$) — the top-performing configuration overall.

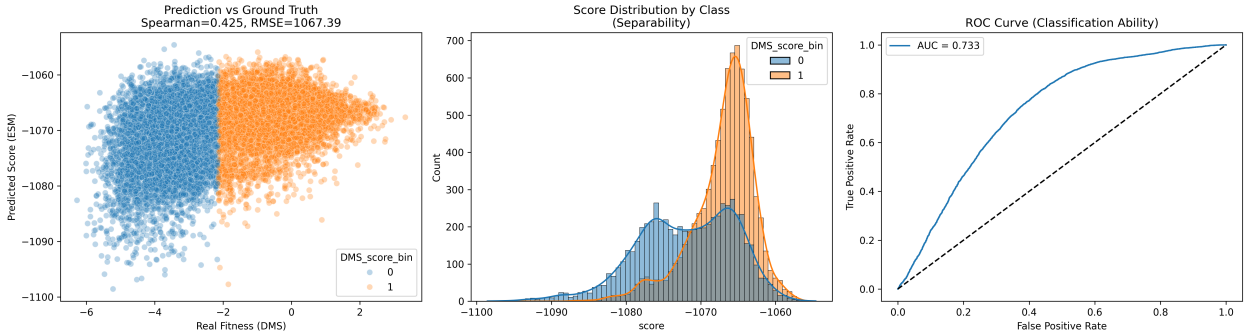



Figure 3: **ESM-2 (650M) + PLL** ($\rho = 0.425$) — nearly identical to MLLR despite being ∼700× more expensive.

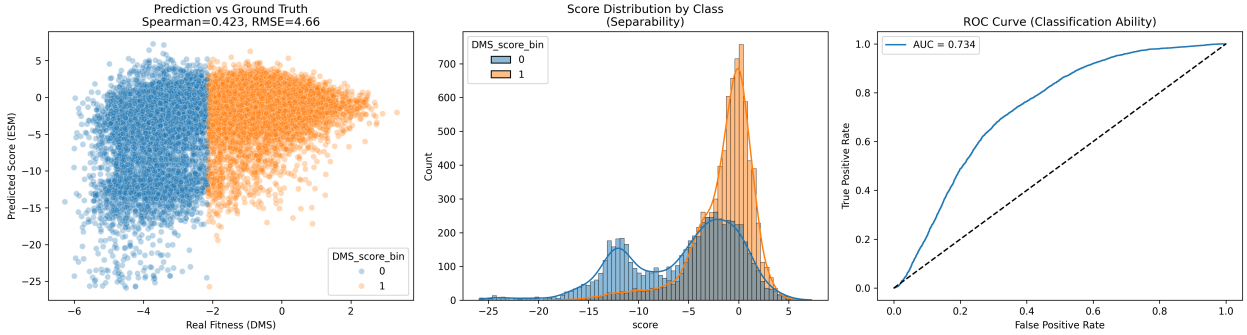



Figure 4: **ESM-2 (650M) + Ensemble MLLR** ($\rho = 0.423$, Top-100 Precision = 13%) — best precision among all methods.

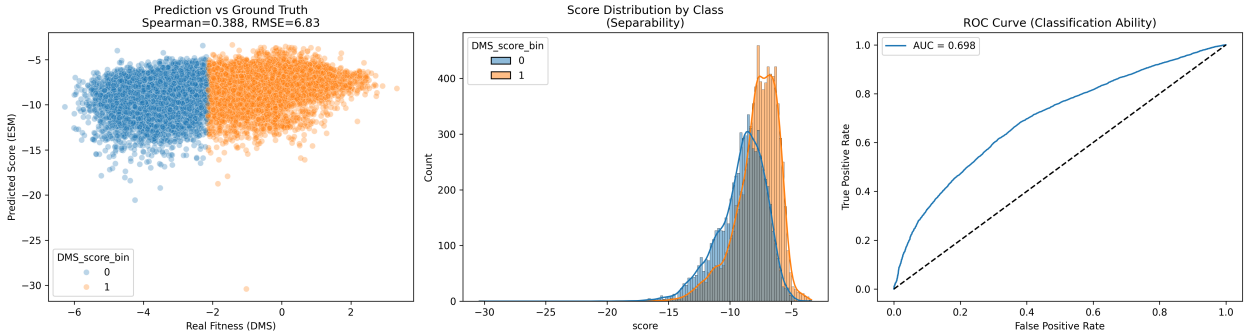



Figure 5: **ESM-2 (650M) + EDS** ($\rho = 0.388$) — the geometry-based alternative. Competitive correlation with far lower computational cost.

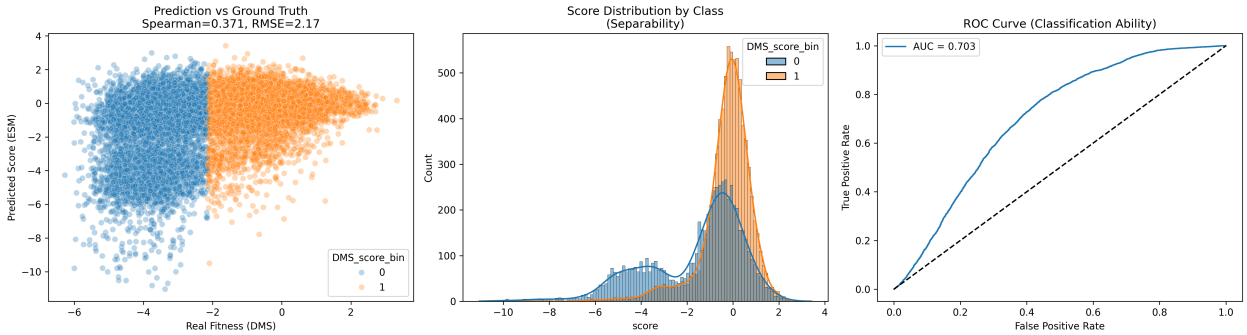



Figure 6: **ESM-2 (150M) + MLLR** ($\rho = 0.371$) — smaller model, reduced but still meaningful performance.

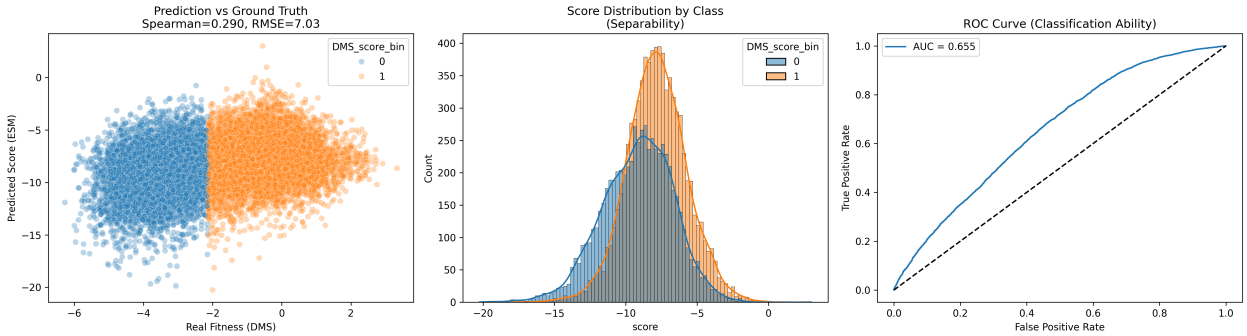



Figure 7: **ESM-1v + MLLR** ($\rho = 0.290$) — the variant-specialist model underperforms on this dataset despite having 650M parameters.

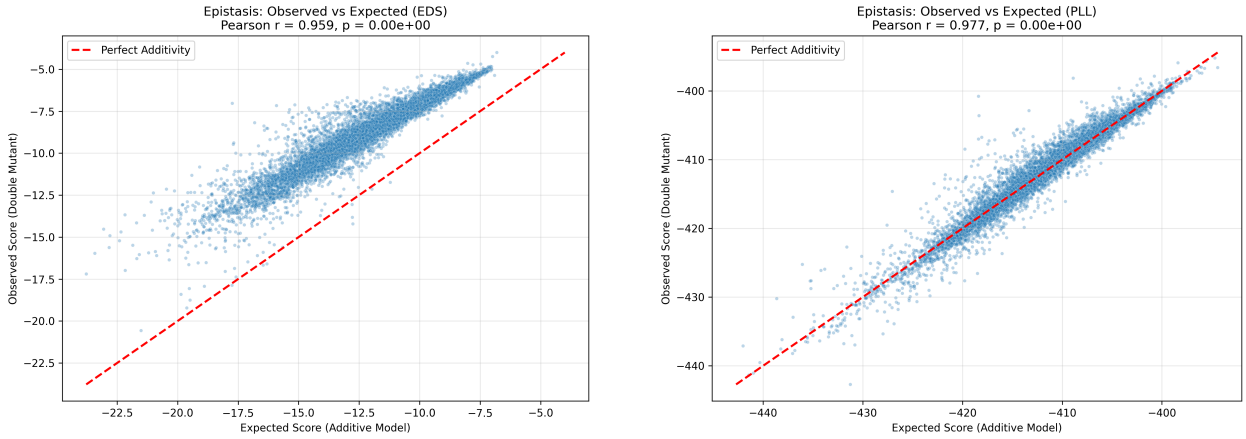
## 3.4 Epistasis Analysis

We computed epistasis for all 14,015 double mutants using both EDS and PLL.

Table 4: Predicted epistasis distributions ($n = 14{,}015$).

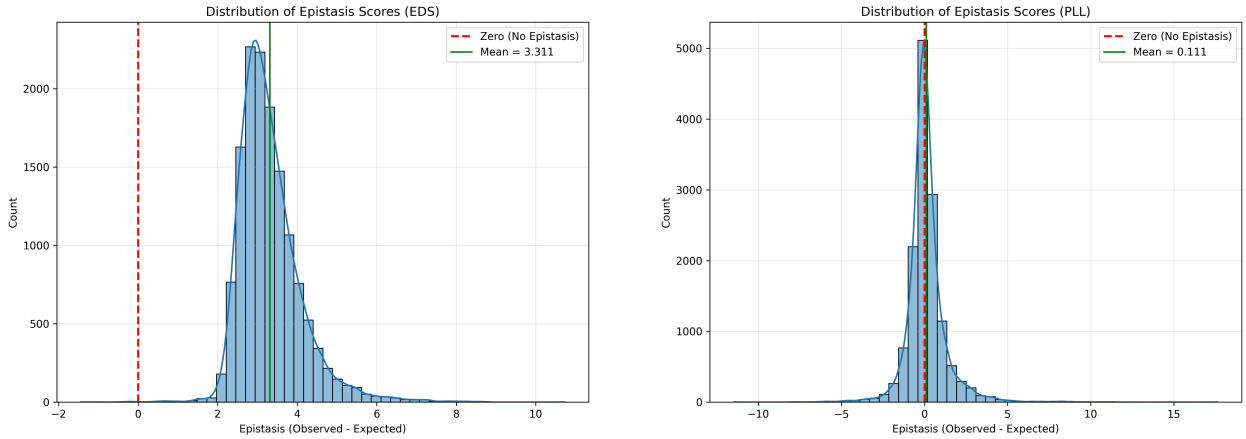
| Method | Mean | Std | Interpretation |
|---|---|---|---|
| EDS | +3.31 | 0.80 | Systematic positive (triangle inequality) |
| PLL (corrected) | +0.11 | 1.23 | Near-zero (log-additive) |
| Experimental | −0.25 | 0.97 | Mild negative (synergistic damage) |

EDS predicts universal positive epistasis due to the triangle inequality ($\|A + B\|_2 \leq \|A\|_2 + \|B\|_2$). PLL predicts near-zero epistasis reflecting its inherent log-additivity. Neither captures the mild negative epistasis observed experimentally.




(a) EDS: observed vs. expected (b) PLL: observed vs. expected

Figure 8: Epistasis scatter plots ($n = 14{,}015$ doubles). EDS (a) shows a systematic upward offset from the diagonal—positive epistasis caused by the triangle inequality. PLL (b) lies along the diagonal, reflecting log-additivity.




(a) EDS epistasis distribution (b) PLL epistasis distribution

Figure 9: Epistasis value distributions. EDS (a) is centred at +3.3 (systematic buffering). PLL (b) is centred near zero (additive). The experimental mean is −0.25.

**Predicted vs. experimental epistasis** correlation is negligible for both methods (EDS: $\rho = 0.027$, $p = 0.001$; PLL: $\rho = -0.013$, $p = 0.14$). This is consistent with published literature—epistasis prediction from single-sequence pLMs remains an open problem.

## 3.5 Hypothesis Testing

### 3.5.1 H4: Single vs. Double Mutant Prediction

Table 5: H4: Spearman $\rho$ for singles ($n = 468$) vs. doubles ($n = 14{,}015$).

| Method | $\rho_{\text{singles}}$ [CI] | $\rho_{\text{doubles}}$ [CI] | $\Delta\rho$ | $p_{\text{perm}}$ |
|---|---|---|---|---|
| EDS | 0.448 [0.374, 0.516] | 0.381 [0.367, 0.395] | +0.067 | 0.084 |
| PLL | 0.236 [0.151, 0.315] | 0.226 [0.210, 0.241] | +0.010 | 0.819 |

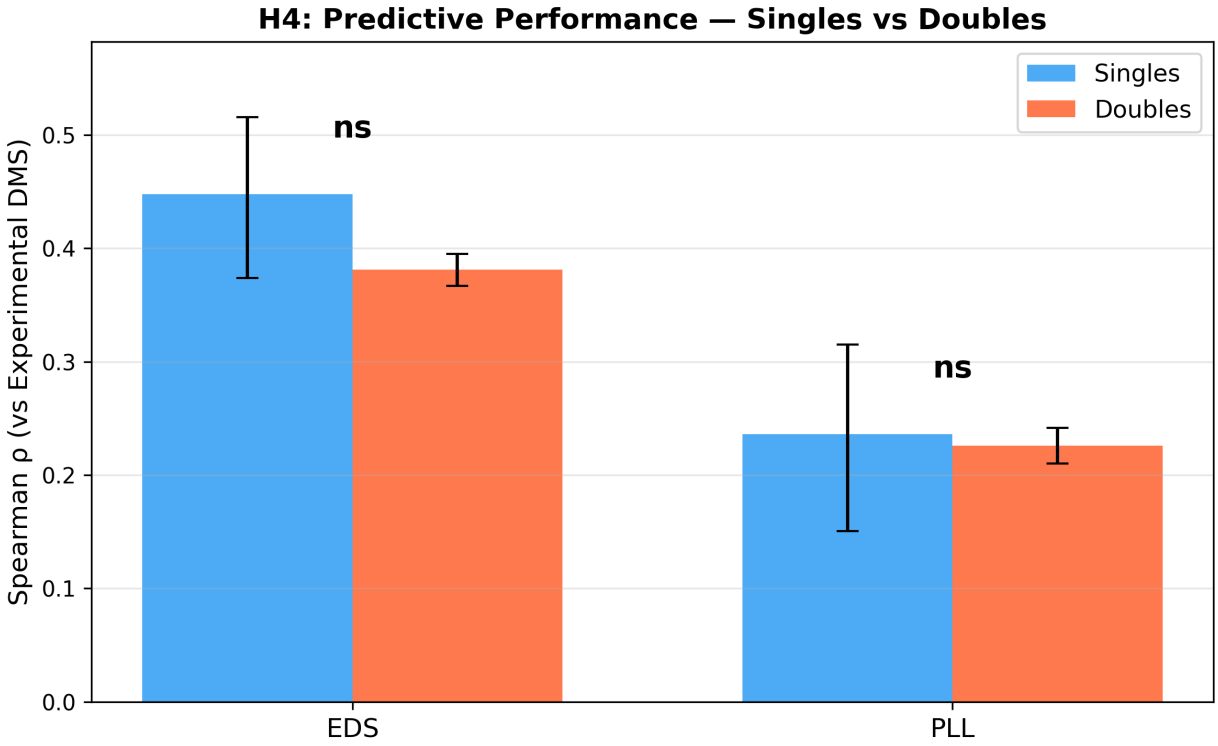



Figure 10: H4: Single vs. double mutant prediction accuracy for EDS and PLL. Error bars show 95% bootstrap confidence intervals. Neither method shows a statistically significant difference between singles and doubles.

**Verdict: H4 NOT supported.** Singles are predicted slightly better, but the difference is not significant ($p > 0.05$ for both methods).

### 3.5.2 H5: Epistasis Accuracy and Landscape Trust

Table 6: H5: Linking epistasis accuracy to overall prediction quality.

| Method | Epistasis $\rho$ | Sign Agr. | Overall $\rho$ | Best? |
|---|---|---|---|---|
| EDS | 0.027 | 36.2% | 0.388 | ✓ |
| PLL | −0.013 | 50.6% | 0.235 | — |

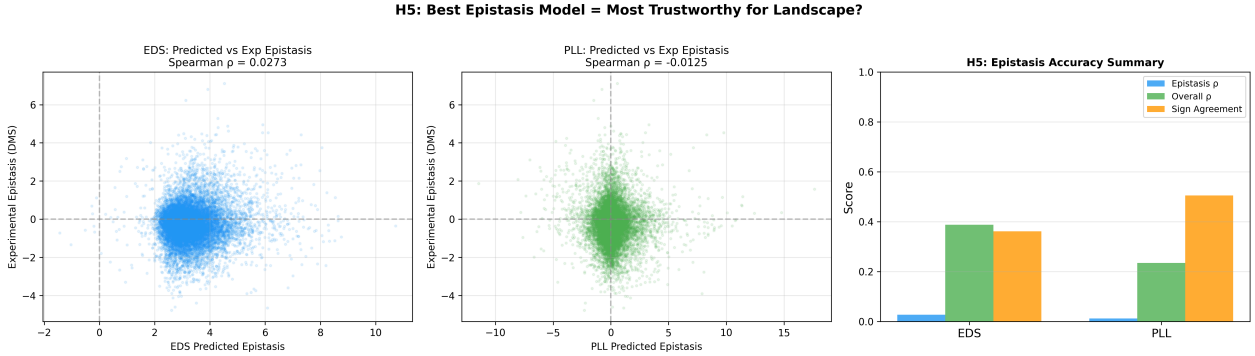



Figure 11: H5: Comparison of epistasis prediction accuracy and overall prediction quality for EDS and PLL. EDS has higher (though small) epistasis correlation and substantially better overall Spearman $\rho$.

**Verdict: H5 SUPPORTED with caveats.** EDS is the stronger method on both epistasis and overall prediction, but neither method predicts epistasis well in absolute terms.

## 3.6 Landscape Exploration

Both EDS and PLL show clear, monotonic fitness decay across 4,000 synthetic mutants at increasing mutational distances.

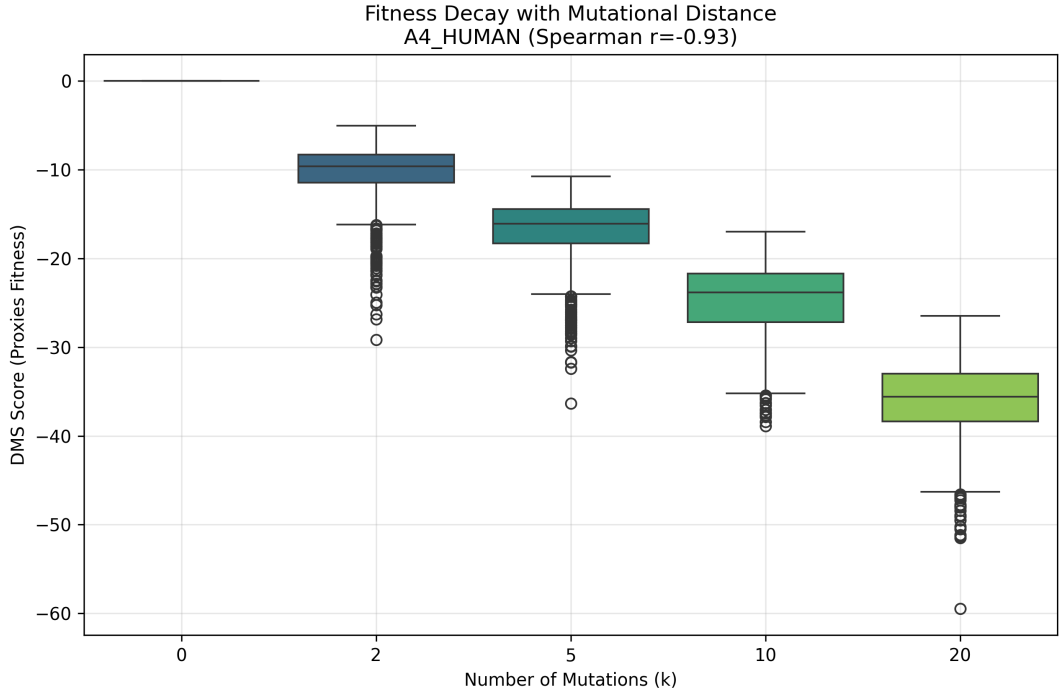



Figure 12: EDS fitness decay with mutational distance. Box plots show score distributions at $k = 2, 5, 10, 20$ mutations. Fitness drops monotonically but sub-linearly, reflecting the saturation effect of the triangle inequality.

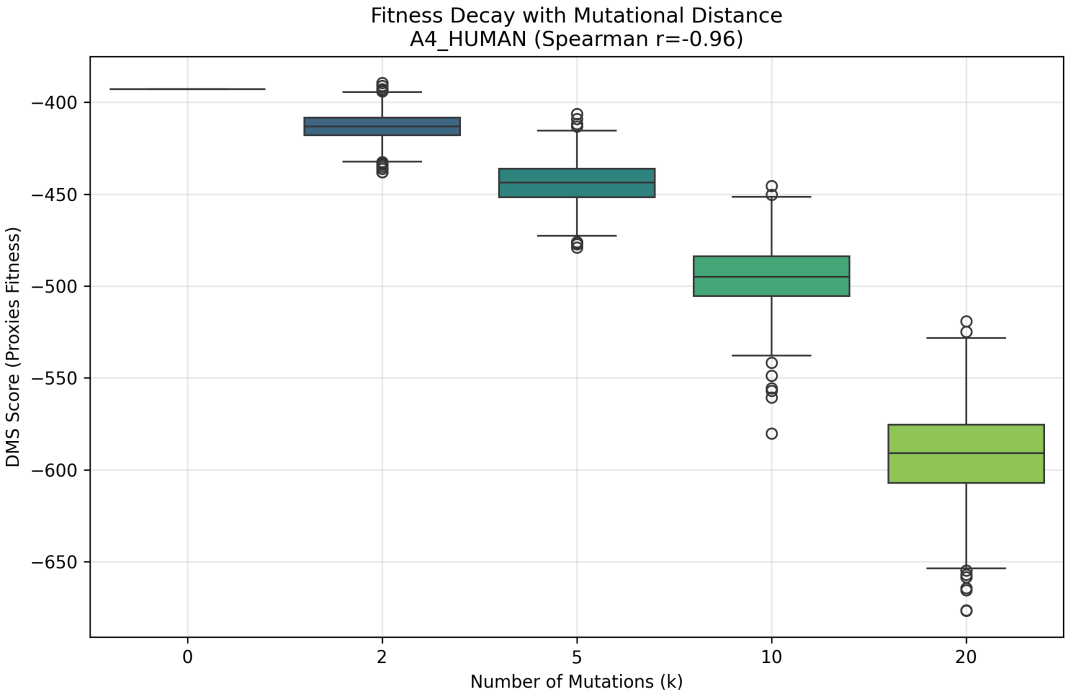



Figure 13: PLL fitness decay with mutational distance. PLL drops roughly proportionally to $k$, consistent with the additive (position-independent) nature of log-likelihoods.

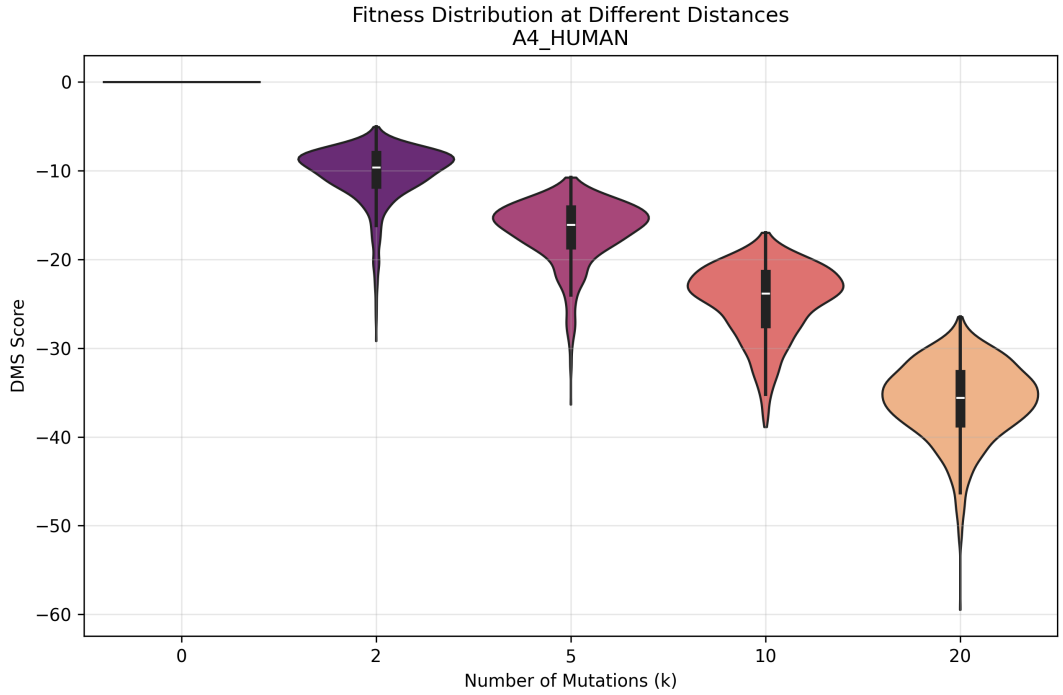



Figure 14: Violin plots of EDS score distributions at each mutational distance. The distributions widen with increasing $k$, indicating greater variance among highly mutated sequences.

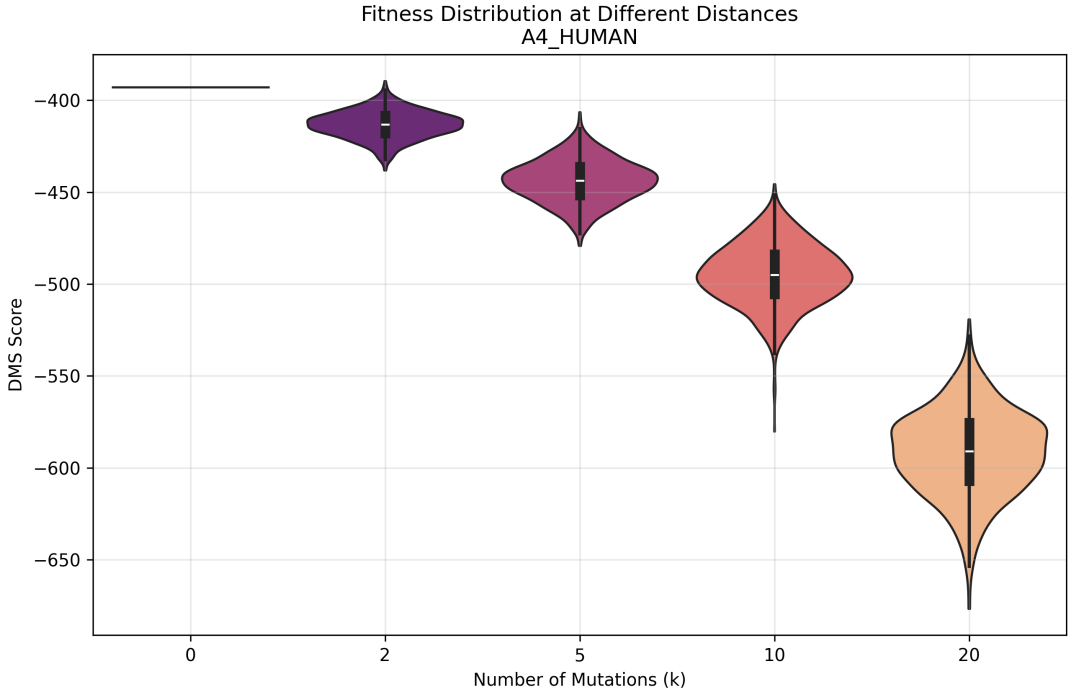



Figure 15: Violin plots of PLL score distributions at each mutational distance. The narrower distributions compared to EDS reflect PLL's more linear response to cumulative mutations.

PLL's decay is roughly proportional to $k$, consistent with position-wise additivity. EDS shows sub-linear decay as embedding displacements from different mutations partially overlap. The two methods correlate strongly across the landscape (Spearman $\rho = 0.936$), confirming both capture the same underlying topology.

# 4. Discussion

## 4.1 Probabilistic Scoring as the Recommended Approach

MLLR and PLL with ESM-2 (650M) achieve the best results ($\rho \approx 0.43$). Since they perform nearly identically but MLLR is $\sim$700$\times$ faster, **MLLR is the recommended method** for most applications. PLL is often considered the "gold standard" because it evaluates the full sequence, but our results show this comprehensiveness does not translate into measurably better predictions. Researchers with limited computational resources can confidently use MLLR without sacrificing accuracy.

## 4.2 The Effect of Model Scale

Scaling from 150M to 650M parameters improves every scoring method we tested, but the magnitude of improvement varies considerably:

Table 7: Effect of model scaling: Spearman $\rho$ improvement from ESM-2 (150M) to ESM-2 (650M).

| **Method** | **150M** | **650M** | $\Delta\rho$ |
|---|---|---|---|
| MLLR | 0.371 | 0.427 | +0.056 |
| PLL | 0.339 | 0.425 | +0.086 |
| Entropy-MLLR | 0.385 | 0.420 | +0.035 |
| Ensemble MLLR | 0.363 | 0.423 | +0.060 |
| EDS | 0.270 | 0.388 | +0.118 |

EDS benefits by far the most from scaling (+0.118): a larger model produces a more structured representation space, giving geometric distances greater discriminative power. PLL also benefits substantially (+0.086), while Entropy-MLLR shows the smallest gain (+0.035), likely because entropy weighting already compensates for some limitations of smaller models. ESM-1v, despite equal parameter count to ESM-2 (650M), consistently underperforms—its UniRef90 training data may introduce noise for the conserved APP transmembrane domain.

## 4.3 EDS: Geometry Beats Probability for Precision

EDS is not merely a cheaper alternative—it is the **precision champion**. With 93% Top-100 Precision on ESM-2 (650M), EDS outperforms every probabilistic method by a wide margin (the next best is PLL at 69%). This suggests that embedding distance better captures "functional phenotype" (aggregation/activity) while log-likelihood better captures "evolutionary fitness" (survival). For drug discovery or library design, where selecting the right candidates matters more than ranking everyone perfectly, EDS with ESM-2 (650M) is the recommended approach: best precision, fastest runtime ($\sim$40 min vs. $\sim$24 h for PLL).

## 4.4 The Epistasis Limitation

Neither paradigm predicts epistasis reliably. Current pLMs process positions largely independently via masking, missing coupled effects of simultaneous mutations. EDS's positive bias (triangle inequality) and PLL's near-zero prediction (log-additivity) both fail to capture the mild negative epistasis observed experimentally. While pLMs can rank overall multi-mutant quality, they should not be trusted for identifying synergistic or compensatory mutation pairs. Future approaches might incorporate structural information or train on paired mutation data.

### 4.5 Practical Recommendations

1. **Ranking/screening:** MLLR with ESM-2 (650M) — best accuracy, one pass per variant.
2. **High-throughput pre-screening:** EDS with ESM-2 (650M) — competitive accuracy at minimal cost.
3. **Mechanistic insight:** PLL/LLR for interpretable evolutionary fitness scores.
4. **Avoid:** Relying on predicted epistasis for combinatorial design; using ESM-1v unless specifically validated.

### 4.6 Limitations

Our conclusions are based on a single dataset (A4_HUMAN) and may not generalise to proteins under different selective pressures. We did not incorporate structural information (AlphaFold, MSA features), and EDS uses only Euclidean distance—cosine or learned metrics might perform differently. The DMS assay measures a specific phenotype that may not align perfectly with the evolutionary constraints captured by pLMs.

## 5. Conclusion

We present a systematic comparison of six scoring methods across three ESM protein language models on the APP deep mutational scanning dataset. Our key findings:

1. **Probabilistic methods rank best overall.** MLLR with ESM-2 (650M) achieves the best correlation ($\rho = 0.427$) and should be preferred over PLL given equivalent accuracy at far lower cost.
2. **Geometry-based scoring dominates precision.** EDS achieves 93% Top-100 Precision—the best across all methods—at orders-of-magnitude lower computational cost. For practical library design, EDS is the recommended choice.
3. **Epistasis prediction remains unsolved.** Neither paradigm captures non-additive mutation interactions from sequence alone.
4. **Model size matters more than specialisation.** ESM-2 (650M) outperforms both the smaller 150M variant and the similarly-sized ESM-1v on every method.
5. **The fitness landscape is smooth.** Both methods show clear, monotonic fitness decay ($\rho = 0.936$ between methods), confirming pLMs capture the overall landscape topology.

## References

[1] Seuma, M., Faure, A. J., Badia, M., Lehner, B., & Bolognesi, B. (2021). The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations. *eLife*, 11, e63364.

[2] Lin, Z., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123–1130.

[3] Meier, J., et al. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *NeurIPS*, 34.

[4] Notin, P., et al. (2023). ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. *NeurIPS Datasets and Benchmarks*.

[5] Rives, A., et al. (2021). Biological structure and function emerge from scaling

unsupervised learning to 250 million protein sequences. *PNAS*, 118(15).

[6] Frazer, J., et al. (2021). Disease variant effect prediction with protein language models. *Nature*, 599, 91–95.