LLMs for Protein Evolution

Which ESM-derived Statistic Best Approximates Protein Fitness? A Comparative Analysis

Members: Adina Nadeem, Prithvi Rajan, and Fazil Tarlan

Introduction: The Protein Fitness Landscape

Protein Fitness: A measure of how well a protein performs its biological function (e.g., 0-100%). This can be quantified experimentally.

A "Fitness Landscape" is a giant 3D map of all possible protein mutants and their fitness scores.

The Problem: We Can't Test Everything in a Lab

Even a small protein has an astronomical number of possible mutants.

Testing every single mutant in a real lab is physically impossible.

What if we use an AI to predict the fitness score for any mutant we want?

These AIs are called Protein Language Models (PLMs) (like ESM-2). They are trained on billions of real proteins from nature

The Project's Real Goal: Which AI Score is Best?

The AI doesn't give a simple 0-100% score. It gives a probability score.

There are different ways to calculate this probability score.

Our Project's Core Question: Which AI model, combined with which probability statistic, is the best and most accurate predictor of real-world fitness?

The AI Models

1. **ESM-1v**: An older model, but it was specifically trained for this exact job of predicting mutation effects. This is our "specialist" benchmark.
2. **ESM-2 (150M)**: A newer, larger, and more powerful general-purpose model.
3. **ESM-2 (650M)**
4. **ESM-3**: The newest generative model

The Statistics We Will Test

Statistic 1: Pseudo-Log-Likelihood (PLL)

Scores the entire mutant protein by averaging the probability of the true amino acid at each position.

This is a thorough metric but computationally slow.

Statistic 2: Log-Likelihood Ratio (LLR)

Calculates the change in probability: (PLL of Mutant) - (PLL of Original)

This metric controls for the model's inherent biases.

This is the primary method used by the ESM-1v paper.

Statistic 3: Masked Marginal Probability

Only scores the probability of the mutant amino acid at the single position (or positions) that changed.

Very fast, but may be less accurate as it ignores the mutation's effect on the rest of the sequence.

Part - I

Goal: To find the winning combination of model + statistic.

How: We need an real dataset with probability scores to check our AI's predictions against.

Deep Mutational Scanning (DMS) dataset.

This dataset contains two columns:

- Mutant Protein Sequence
- Real Lab Fitness Score

Validation Data: Testing for Epistasis

Epistasis: Non-additive effects where the impact of one mutation is dependent on the presence of another

We will use DMS datasets that provide fitness scores for both:

- Single Mutants (1-mutation)
- **Double Mutants (2-mutations)**

Success Criterion: A model's ability to accurately predict the fitness of *both* single and double mutants is essential to prove it has learned the landscape's complex and non-linear features.

Criterion 1: Correlation

- We will be comparing our AI's predicted scores to the Real Lab Scores.
- The statistic with the highest **correlation score** (closest to 1.0) is the winner. This proves it is the most accurate.

Criterion 2: Computational Speed

- How long does it take to calculate the score? A fast-but-accurate statistic is more useful.

Part II - **Mapping the Broader Fitness Landscape**

Objective: To characterize the landscape structure beyond the local, experimentally-validated neighborhood (i.e., beyond k=1, 2).

We will use the optimized model-statistic combination (the "winner" from Experiment 1) as our calibrated proxy for fitness.

Question: How does predicted fitness decay as we move further from the known, high-fitness sequence?

We will computationally generate libraries of N random mutants.

These libraries will be generated at fixed mutational distances (k) from the original sequence (e.g., k=5, k=10, k=20).

Our validated model from Part 1 will be used to score all generated mutants.

We will plot the distribution (or mean) of predicted fitness scores against the mutational distance (k).

Key Hypotheses

H1: The "Difference Score" (LLR) will be the most accurate statistic across all models.

H2: The specialized ESM-1v model will outperform the generalist ESM-2 models on the 1-mutation tests.

H3: The larger ESM-2 (650M) will outperform the smaller ESM-2 (150M).

H4: Predictive performance for all models will be much higher for single mutants than for double mutants (because epistasis is hard).

H5: The best-performing model on the double-mutant "epistasis test" will be the most trustworthy for exploring the "whole landscape" (e.g., 10+ mutations).

Timeline (8 Weeks)

Weeks 1-2: Setup computing environment, finalize literature review. Download and pre-process ProteinGym data.

Weeks 3-4: Implement the code for all 6 scoring statistics for all 3 models.

Weeks 5-6: Run the full analysis. This is the main computation phase, generating scores for millions of mutants.

Weeks 7-8: Statistical analysis. Calculate all Spearman correlations. Analyze single-mutant vs. double-mutant performance. Make report.

Expected Outcomes

Deliverables:

1. A final ranking of all model/statistic combinations.
2. A decision framework (eg: "For speed, use X. For accuracy, use Y.").
3. A final report analyzing why certain methods performed best.