– Independent Work Report 2025 –

# Optimizing ESM-2 for Better Prediction of Mutation Effects in Viral Proteins

Aylin Hadzhieva
Advisor: Mona Singh
Contributors: James Fife

## Abstract

*Large protein language models have achieved strong mutation prediction performance but require massive computational resources and favor well-sampled protein families. We fine-tune small ESM-2 models (8M and 35M parameters) on selectively augmented datasets from UniRef100 to improve predictions for underrepresented proteins. Our approach closes the performance gap with much larger models, often matching or exceeding 650M and 15B parameter baselines. Fine-tuned small models maintain better biological plausibility by achieving moderate entropy levels, rather than overfitting through memorization. These results show that targeted fine-tuning enables lightweight models to deliver accurate meaningful predictions without the need for large-scale architectures.*

## 1. Introduction

Proteins are molecular machines that perform most of the essential tasks within living cells. They are made of linear chains of amino acids, represented by 20 different letters of the alphabet, which fold into complex three-dimensional structures (Fig.1a). This structure, shaped by the specific order of amino acids, determines the protein's function. Understanding how proteins behave — and how changes in their sequence affect that behavior — is critical for studying biological processes and designing therapies.

Mutations, or changes to a protein's amino acid sequence, can alter its stability, function, or interactions with other molecules (Fig.1b). Depending on the location and type of change, a mutation may be harmless, or it may destabilize the protein and impair its function. When a protein fails to work properly, the consequences can include disease progression, drug resistance, or loss of cellular control.

Predicting the effects of mutations is important not only for understanding existing biological systems but also for anticipating how they may evolve. For example, predicting evolutionary outcomes by modeling the probability of sequence variants has become a key approach in studying viral evolution and immune escape [5] .

Traditionally, researchers have relied on experimental techniques to assess mutation. Deep mutational scanning (DMS) is one such method, which allows researchers to calculate the probability or relative impact of different variants [8]. However, DMS is limited by experimental constraints: it is expensive, labor-intensive, and feasible only for a small subset of proteins. Even when successful, DMS datasets often cover only a narrow range of mutation types and miss rare but biologically important variants. Other experimental techniques, such as mutagenesis studies, functional assays, and spectroscopic methods, provide detailed information but are similarly restricted in scale and generalizability.

To address the limitations and inefficiencies inherent in experimental methodologies, computational approaches have emerged as valuable alternatives, offering more scalable and resource-efficient solutions for protein variant analysis. Recent strategies use machine learning and deep learning to learn from large datasets of protein sequences and structures. However, all these methods remain limited by the quantity and diversity of labeled training data.

The emergence of protein language models (PLMs) offers a promising new direction by treating protein sequences as a form of language. Similar to how language models learn grammar and meaning from text, PLMs learn patterns and relationships between amino acids in protein sequences without requiring explicitly labeled mutation effects. These models are trained on large databases of naturally occurring protein sequences and can predict whether a mutation is likely to occur in nature based on evolutionary patterns [9].

Even the most sophisticated PLMs while successful at zero-shot variant prediction, still struggle in certain type of predictions. They perform poorly when analyzing protein families or regions that are underrepresented in training data, or when predicting effects of subtle but important mutations in these less-studied regions [2].

Prior studies [2] [1] have focused primarily on building increasingly larger PLMs or incorporating structural information into existing models to improve performance. These approaches often assumed that simply scaling up model size or training data would capture all necessary biological details. Large PLMs like ESM1b have shown strong performance in predicting mutation effects, but their success relies heavily on massive model sizes (often hundreds of millions of parameters) and training on datasets like UniRef50 [2]. This approach works well for common protein families but often fails to accurately model less-represented proteins.

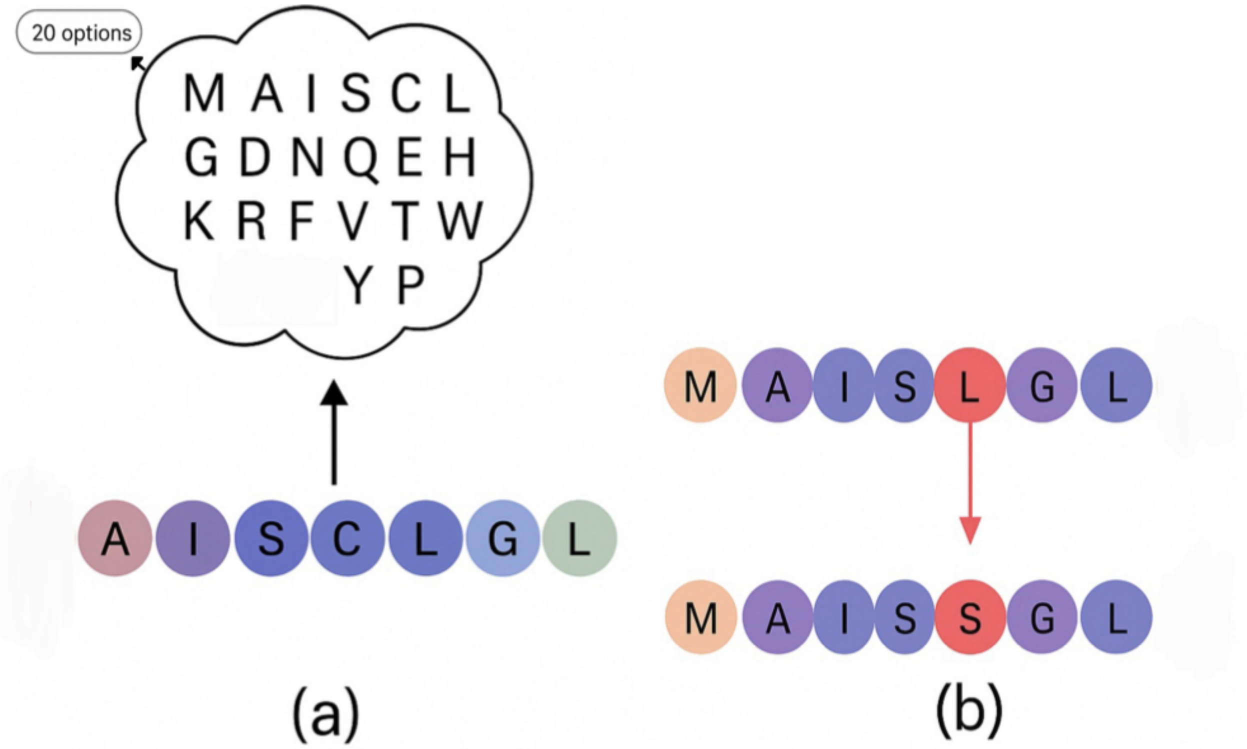


**Figure 1: Protein Sequences and Mutations. (A)** A protein sequence is a long string of smaller units called amino acids. There are only 20 different amino acids, and we represent each one with a single letter. Each letter here is one amino acid. The order of these letters determines how the protein folds into its 3D shape — and that shape determines what the protein can do in the body. **(B)** A small change like “L” to “S” is a mutation — like a typo in a sentence. Depending on where it happens and what it changes, it might be harmless or it might completely disrupt the protein’s function.

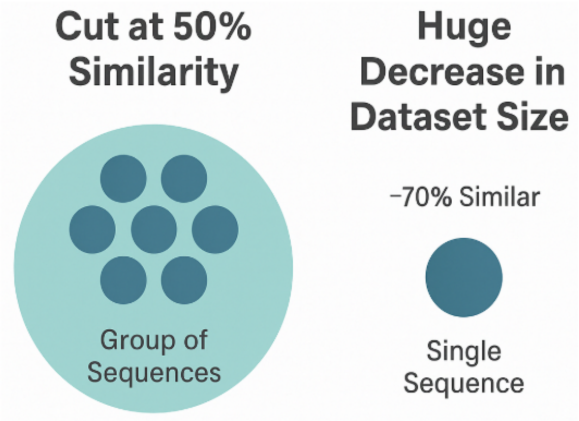


**Figure 2: UniRef50 clustering** A sequence can be part of a group of sequences (up to tens of thousands of seqeunces!) that are all 70% similar and these will get a "representative" sequence in UniRef50, meaning that that cluster of tens of thousands of sequences gets shrunk to a single sequence. This means that when training ESM2 models they only see one example.

Smaller PLMs, while more computationally accessible and efficient, typically perform worse than larger models—especially when analyzing underrepresented proteins. Improving mutation prediction for these cases is crucial to making PLMs more biologically accurate, computationally efficient, and broadly applicable across the protein universe.

Our research explores whether small, computationally efficient PLMs (8M and 35M parameters) can achieve comparable mutation prediction performance to much larger models while maintaining biological relevance. Specifically, we fine-tune small ESM-2 models using an augmented dataset specifically targeted at underrepresented protein sequences, aiming to improve their predictive capabilities for these challenging cases without requiring massive computational resources.

## 2. Methods

### 2.1. Background

Language models are neural networks designed to learn patterns in sequences — whether composed of words, amino acids, or other tokens. A major advance in language modeling came with BERT (Bidirectional Encoder Representations from Transformers), introduced in 2018. BERT is trained to predict a hidden ("masked") word in a sentence by considering the context on both sides of the blank. This training objective, called Masked Language Modeling (MLM), proved highly effective for capturing the structure of natural language.

The same strategy applies naturally to biological sequences. In proteins, the identity of one amino acid often depends on its neighbors, making MLM an ideal training objective. ESM-2 (Evolutionary Scale Modeling 2), developed by Meta AI, adapts the BERT architecture to protein sequences. It is trained by masking an amino acid at random and predicting the correct residue based on the surrounding sequence.

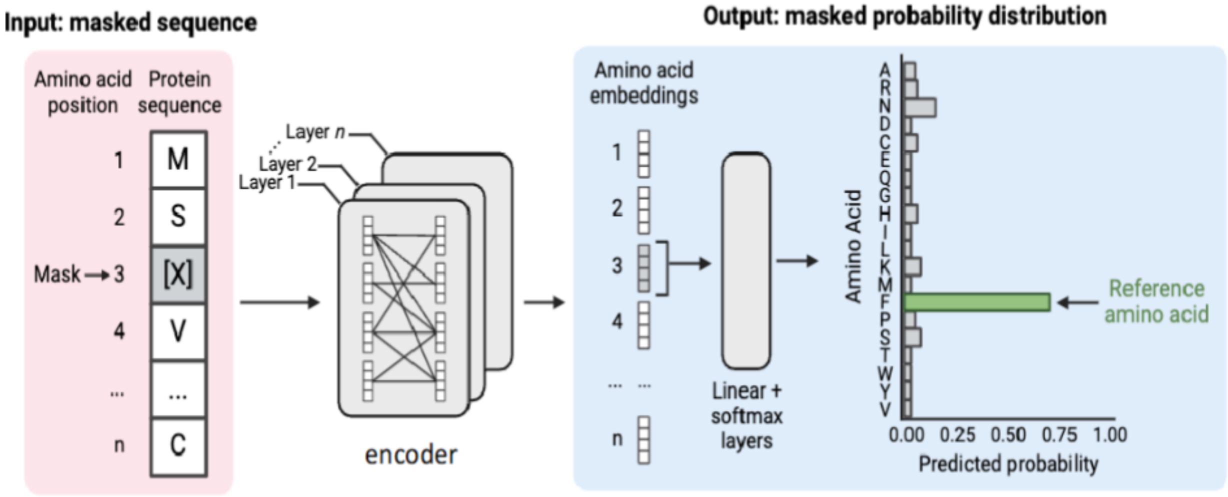
This masked prediction ability is particularly useful for studying the effects of mutations. By masking a position in a protein and examining the model's predicted probabilities for possible substitutions at that specific masked position (Fig.3), researchers can infer how surprising or disruptive a mutation might be — much like checking whether a replacement word makes sense in a sentence.

ESM-2 is trained on the UniRef50 dataset, a curated collection of protein sequences clustered at 50% sequence identity. In UniRef50, sequences that are more than 50% identical are collapsed into a single representative sequence (Fig.2). While this reduces redundancy and speeds up training, it also dramatically reduces the diversity of sequence space seen by the model. A cluster may contain tens of thousands of similar sequences, yet only a single representative is used during training. As a result, certain rare variants — including many viral proteins — are underrepresented, limiting the model's ability to generalize to poorly sampled or rapidly evolving proteins.

These constraints in ESM-2's design and training data directly motivate our fine-tuning approach, which aims to improve mutation effect predictions specifically for underrepresented proteins in the trained dataset without requiring massive computational resources.

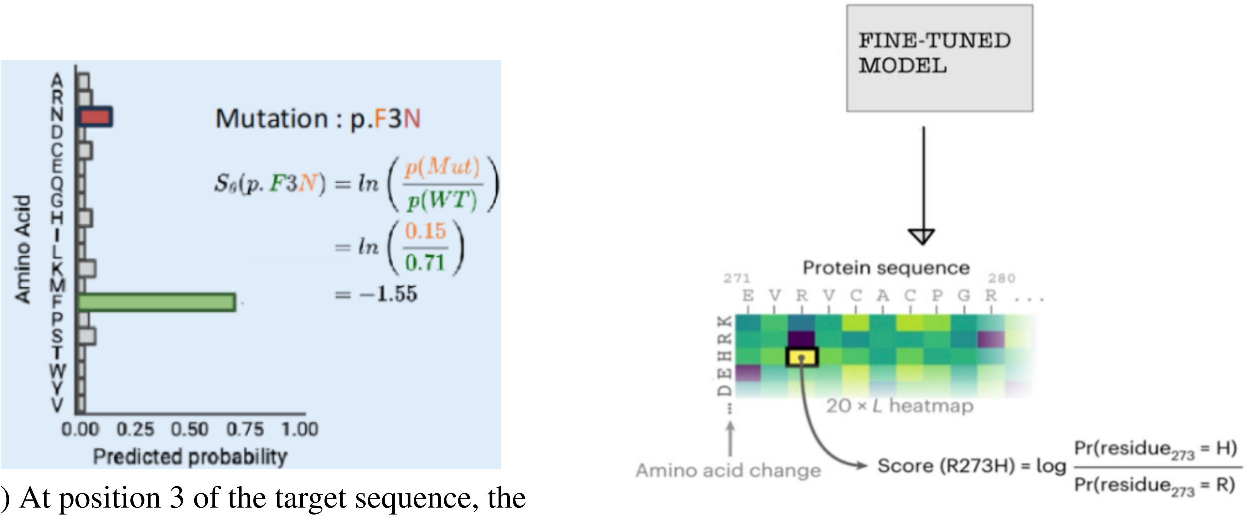
### 2.2. Dataset

This project utilizes two types of data: experimentally measured mutation effects and protein sequence collections. For evaluation, we use deep mutational scanning (DMS) datasets from ProteinGym, which provide experimental measurements of mutation effects across various proteins.



James Fife & Ann Cirincione

**Figure 3: Architecture of ESM-2 for masked amino acid prediction**


Amino Acid
A R N D C E Q G H I L K M F P S T W Y V
0.00 0.25 0.50 0.75 1.00
Predicted probability
Mutation : p.F3N
$S_6(p.F3N) = ln\left(\frac{p(Mut)}{p(WT)}\right)$
$= ln\left(\frac{0.15}{0.71}\right)$
$= -1.55$
FINE-TUNED MODEL
Protein sequence
271
280
E V R V C A C P G R . . .
K R H E D ...
20 × L heatmap
Amino acid change
Score (R273H) = log $\frac{Pr(residue_{273} = H)}{Pr(residue_{273} = R)}$

(a) At position 3 of the target sequence, the probability of a specific mutation to amino acid F is evaluated. The model's predicted probabilities for the wildtype (F) and mutant (N) residues are compared using a log odds ratio.

(b) The fine-tuned model produces a log odds score for every possible amino acid substitution across every position at the protein sequence, displayed as a 20-by-L heatmap where L is the sequence length.

**Figure 4: Calculation of mutation score using log odds ratio.**

These measurements serve as ground truth for assessing model performance. For training, we employ protein sequences from two complementary UniProt-derived databases:

1. UniRef50 - A non-redundant database where sequences with >50% identity are clustered, with only one representative retained per cluster
2. UniRef100 - A more comprehensive database that preserves all unique sequences, including closely related variants filtered out in UniRef50

We selected six target proteins for our experiments based on two criteria: (1) availability of comprehensive DMS data and (2) documented underrepresentation in UniRef50, as evidenced by poor performance of small-parameter models on these sequences. Our evaluation focuses exclusively on mutation effect predictions for these target proteins, allowing us to precisely measure improvements in handling underrepresented protein families.

### 2.3. Data Augmentation and Preprocessing

For each target protein, we employed a two-stage data collection strategy:

- For each target protein, we include its single representative sequence from UniRef50
- We supplement this with additional homologous sequences from UniRef100 using BLASTp with stringent criteria: e-value threshold of $10^{-5}$, minimum sequence coverage of 70%, and minimum percent identity of 70%. This approach ensured we captured biologically relevant sequence variations while maintaining structural and functional similarity to our target proteins.

Table1 presents our six target proteins with their identifiers, associated DMS studies, and the number of homologous sequences retrieved from UniRef100.

All sequences were tokenized using the HuggingFace ESM-2 tokenizer, converting each amino acid to a corresponding input ID. During training, we randomly masked 15% of amino acids in each sequence, following the standard masking procedure for protein language models while excluding special tokens. This masking strategy forces the model to predict amino acids based on surrounding context, improving its understanding of sequence-structure-function relationships specifically for our target protein families. This process was done separately for each target sequence to ensure task-specific fine-tuning.

### 2.4. Model and Fine-Tuning Framework

We fine-tuned two variants of the ESM-2 model: 8M and 35M parameters, both pre-trained on UniRef50. These models represent different efficiency-performance trade-offs in the protein language model space:

**Table 1: Target proteins from Uniref50 with associated sequence counts from UniRef100**

| Protein ID | DMS Study | Length | Sequences |
|---|---|---|---:|
| PA_I34A1 | Wu_2015 | 716 | 22,991 |
| A0A2Z5U3Z0_9INFA | Wu_2014 | 564 | 48,605 |
| SC6A4_HUMAN | Young_2021 | 630 | 32,795 |
| NRAM_I33A0 | Jiang_2016 | 469 | 33,912 |
| MTH3_HAEAE | RockahShmuel_2015 | 374 | 27,483 |
| NCAP_I34A1 | Doud_2015 | 498 | 11,603 |

1. ESM-2 (8M): The most parameter-efficient model, suitable for deployment in resource-constrained environments
2. ESM-2 (35M): An intermediate-sized model offering a balance between computational efficiency and predictive power

For each target protein, we fine-tuned both the 8M and 35M parameter models independently on its protein-specific augmented dataset (Fig.5). This targeted approach allowed each model to specialize in the particular patterns of a single protein family.

Cross-entropy loss was used as the training objective, measuring how accurately the model predicts the original amino acid at each masked position. Each fine-tuning experiment ran for 3-5 epochs on a single GPU using PyTorch and the HuggingFace Transformers library.

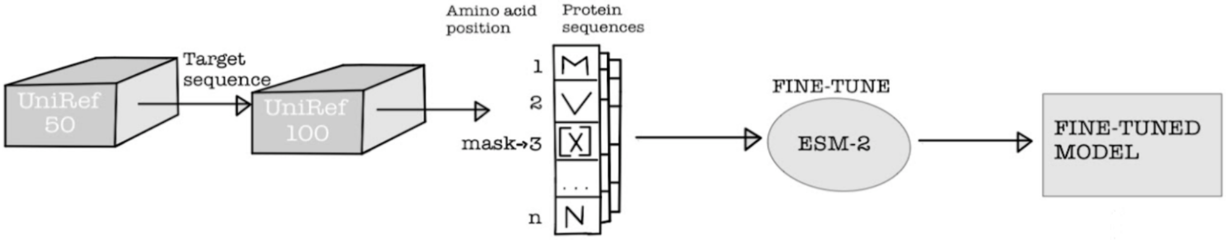


**Figure 5: Fine-tuning approach**

### 2.5. Evaluation Metrics

**2.5.1. Spearman Rank Correlation** We evaluate predictive performance using Spearman rank correlation, calculated between the model's predicted mutation effects and DMS experimentally measured values for each target (wildtype) sequence. Crucially, we obtain these predictions in a zero-shot manner after fine-tuning is complete. We obtain them only for that target sequence with whose augmented dataset the corresponding model was fine-tuned, and we do this independently for all target sequences.

To obtain predictions, we mask one amino acid at a time in the target (wildtype) sequence and compute the model's probability distribution over all 20 amino acids at that position (Fig.3). This distribution reflects the model's belief about which residues are most likely or biologically plausible at that location, given the surrounding sequence context.

We extract two probabilities from the predicted distribution (Fig.4).

- $p_{\text{wt}}$: the probability assigned to the wildtype amino acid,
- $p_{\text{mut}}$: the probability assigned to the mutated amino acid.

We then compute the log odds ratio:

$$\text{score}_{i,j} = \log\left(\frac{p_{\text{mut}}}{p_{\text{wt}}}\right) \tag{1}$$

where $i$ represents the position in the sequence and $j$ represents the specific mutation at that position. This value quantifies how much the model "favors" the mutation over the original residue. A negative score implies the mutation is unlikely, while a positive score suggests it may be tolerated or beneficial.

Because each protein sequence typically has many possible mutations, we collect the log odds ratio scores for all experimentally measured mutations across the sequence. Spearman rank correlation is then computed between these predicted scores and the corresponding experimentally measured fitness values. Thus, Spearman correlation provides a **single value for the entire sequence**, summarizing how well the model ranks mutation effects overall. We do not compute Spearman per position; instead, the ranking quality across all mutations defines a single metric.

We calculate Spearman correlation $\rho$ between the model's log odds predictions and the corresponding DMS fitness scores:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{2}$$

where $d_i$ is the difference in ranks between predicted and experimental values, and $n$ is the number of mutations. This correlation is computed only on mutations from the target (wildtype) sequence, as mutation effects are experimentally measured for that sequence alone. The fine-tuning dataset from UniRef100 is used exclusively during training. Spearman correlation is used as a post-training comparison between fine-tuned models and baselines.

**2.5.2. Entropy** To evaluate how confident and focused the model's predictions are, we compute the Shannon entropy of the output distribution at each masked position:

$$H_i = -\sum_{a \in \mathscr{A}} p_i(a) \log_2 p_i(a) \tag{3}$$

where $H_i$ is the entropy at position $i$, $\mathscr{A}$ is the set of 20 amino acids, and $p_i(a)$ is the model's predicted probability for amino acid $a$ at position $i$.

Entropy captures how spread out the model's belief is across the 20 possibilities. For example, if all amino acids are predicted with equal probability ($p_i(a) = \frac{1}{20}$ for all $a$), the entropy is high (maximum value of $\log_2(20) \approx 4.32$), indicating that the model assigns nearly equal unimportance to all possibilities. If one amino acid has high probability and the rest are low, entropy is low, indicating that the model confidently identifies a preferred amino acid at that site.

We calculate entropy individually for each masked position and then average these entropy values across all positions in the sequence. This provides a **single average entropy value per sequence**, summarizing the model's overall prediction sharpness.

A well-trained model should have low entropy at biologically constrained positions, where only a few substitutions are likely to be tolerated. Extremely low entropy everywhere would suggest overconfidence, ignoring the natural variation seen in proteins, while extremely high entropy would suggest the model is uncertain even at well-constrained regions. We used entropy as both a training diagnostic (measuring change over epochs) and a post-training comparison between fine-tuned models and baselines.

**2.5.3. Baselines** To evaluate the impact of our fine-tuning approach, we compared both the Spearman correlation and entropy metrics of our fine-tuned models against several baselines:

- Zero-shot ESM-2 models (8M, 35M, and 15B parameters) evaluated without fine-tuning
- ESM-2 650M pre-trained on UniRef100, which had access to all sequences but was not specifically adapted to each target protein

These comparisons allow us to isolate the effect of targeted data augmentation and fine-tuning, especially on small parameter-efficient models. [1]

# 3. Results

## 3.1. Fine-tuning Improves Mutation Effect Prediction

We first evaluated how fine-tuning small ESM-2 models (8M and 35M parameters) on augmented UniRef100 datasets affects their ability to predict mutation effects. Fine-tuning consistently improved Spearman rank correlation compared to zero-shot baselines.

Figure 6 shows the training trajectory for the 8M model on a representative target protein, NCAP. Spearman correlation between model predictions and DMS experimental mutation effects increased steadily during fine-tuning. Concurrently, the output entropy of the model decreased, indicating that the model became more confident in its amino acid predictions, as the baseline entropy score was around 4.0, which means that the model had no idea what the sequence was as it was giving unimportance to everything.

## 3.2. Comparative Performance Against Baselines

Our experiments reveal that targeted fine-tuning of small protein language models significantly improves their ability to predict mutation effects, especially for underrepresented proteins. Figure 7 shows the Spearman rank correlation between experimental DMS measurements and model predictions across our six target proteins.

As shown in Figure 7, fine-tuned ESM-2 models (8M and 35M parameters) consistently outperform their baseline not fine-tuned zero-shot counterparts across all targets. For every target protein, fine-tuning leads to substantial improvements compared to the pre-trained model of the same size.

[1] ..

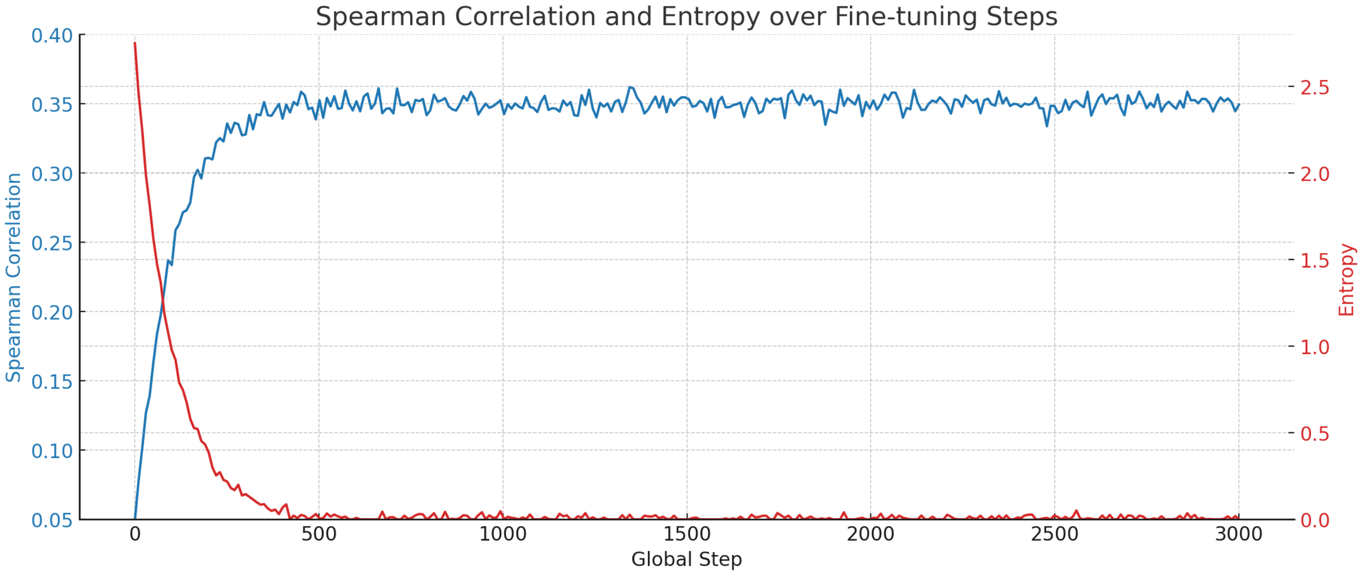


**Figure 6: Metrics over Fine-Tuning Steps** Spearman correlation (left y-axis) and entropy (right y-axis) during fine-tuning for the 8M ESM-2 model on a target protein NCAP. Fine-tuning improves prediction quality while reducing output uncertainty.

Moreover, the fine-tuned small models often match or surpass the performance of much larger models. In many cases, the fine-tuned 8M model achieves Spearman correlations close to or higher than the zero-shot 15B model — despite the 15B model having over 1875 times more parameters.

When compared against the ESM-2 650M model trained on the full UniRef100 dataset, the fine-tuned small models also perform competitively. In most cases, the fine-tuned models achieve similar Spearman correlations, and for some targets — such as SC6A4 — the fine-tuned model even slightly outperforms the 650M baseline. This is notable because the 650M model had access to the full UniRef100 during pretraining, while our fine-tuned models used only a carefully selected subset focused on the target protein family.

These results demonstrate that targeted data augmentation combined with lightweight fine-tuning is sufficient to recover much of the predictive power otherwise gained only through massive model scaling and pretraining on broader datasets. Fine-tuning small models enables efficient, biologically grounded mutation effect prediction without the computational burden of training and deploying billion-parameter models.

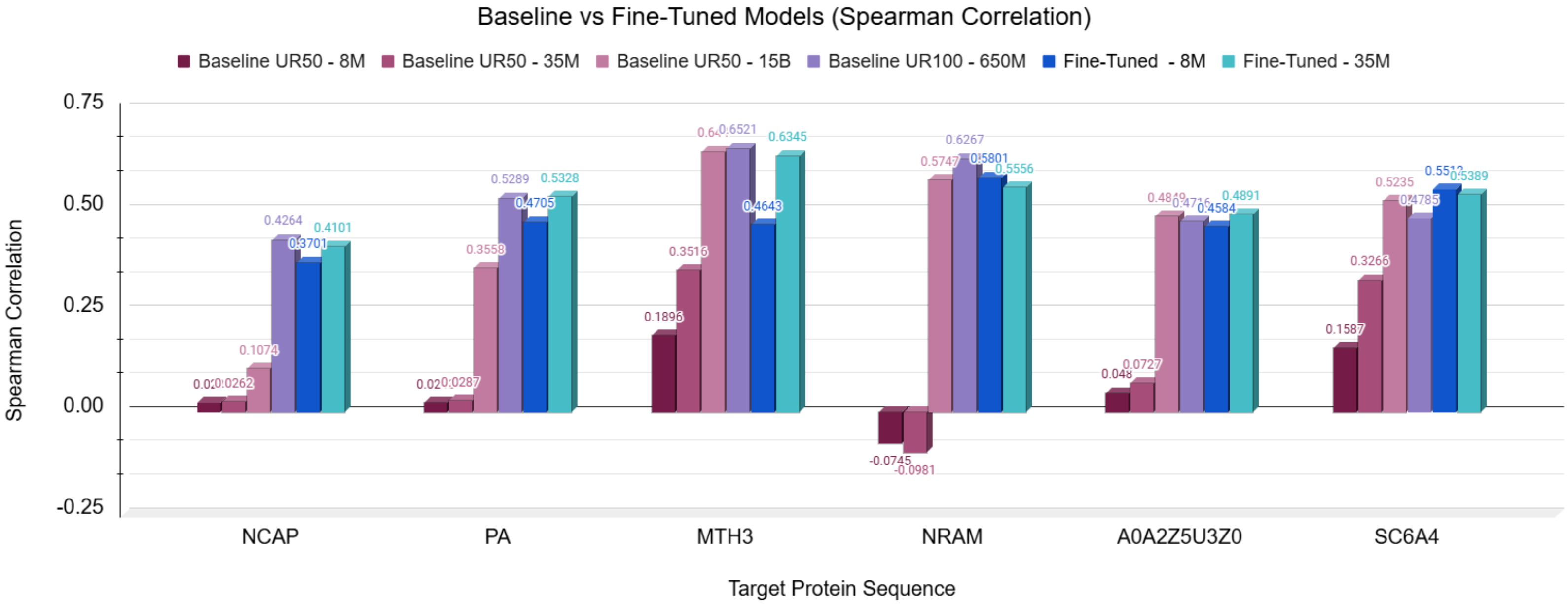


**Figure 7: Comparison of Spearman rank correlations across target proteins.** Fine-tuned small models outperform their zero-shot counterparts, close the gap with 15B models, and achieve performance comparable to the 650M UniRef100 model.

### 3.3. Entropy Dynamics During Fine-Tuning

We next examined how fine-tuning affected the entropy of the model's output distributions, providing insight into the model's prediction sharpness and biological confidence.

Figure 6 shows the evolution of entropy and Spearman correlation as a function of training steps. The model rapidly reduces the entropy of its predictions, reaching values close to zero within the first 500 steps. Correspondingly, Spearman correlation improves sharply during this early phase, suggesting that the model quickly learns to favor a narrow set of likely amino acid substitutions.

In contrast, for proteins where entropy remains around 1 throughout training , the learning curve is more gradual (Fig. 8). Spearman correlation improves steadily over a larger number of steps, indicating that maintaining moderate uncertainty might allow the model to make finer adjustments over time.

These two examples illustrate that Spearman correlation tends to mirror changes in entropy: rapid entropy reduction aligns with early gains in predictive accuracy, while stable entropy supports slower but sustained learning. To highlight the broader trends, we selected these two examples as

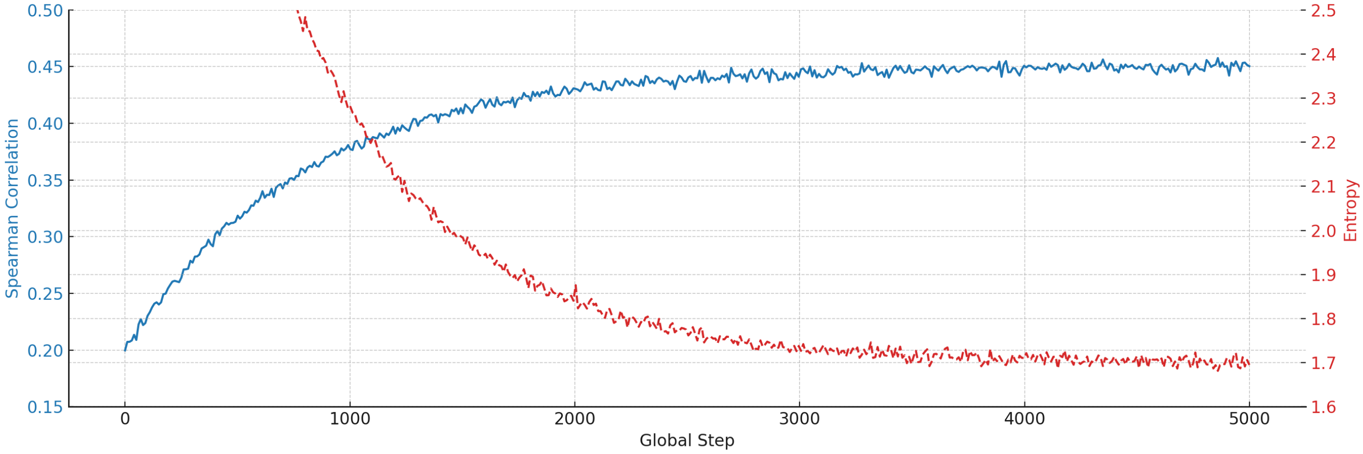


**Figure 8: Training dynamics for a protein sequence where entropy remains moderate.** Learning progresses more smoothly over training steps compared to cases with rapid entropy collapse. The target protein sequence portrayed here is MTH3.

representative cases; the majority of other target sequences follow similar patterns depending on whether their entropy converges toward 0 or stabilizes above 1.

### 3.4. Relationship Between Entropy and Prediction Accuracy

To investigate further the relationship between entropy and mutation prediction performance, we analyzed the relationship between average entropy and Spearman correlation across all fine-tuned models and proteins.

Figure 9 plots Spearman correlation against average entropy for the fine-tuned 8M models for each target sequence, while Figure 10 shows the same for 35M models.

While no strong or monotonic relationship is observed between entropy and predictive accuracy, we notice a weak trend where moderately higher entropy values are associated with somewhat improved Spearman correlations. However, the effect is neither consistent nor statistically strong. This suggests that achieving higher entropy may help, but the current dataset does not provide enough diversity to fully explore this relationship.

### 3.5. Relationship Between Number of Augmented Sequences and Predictive Performance

We also investigated whether the number of augmented sequences retrieved from UniRef100 influences the final predictive performance of the fine-tuned models. Specifically, we plotted the

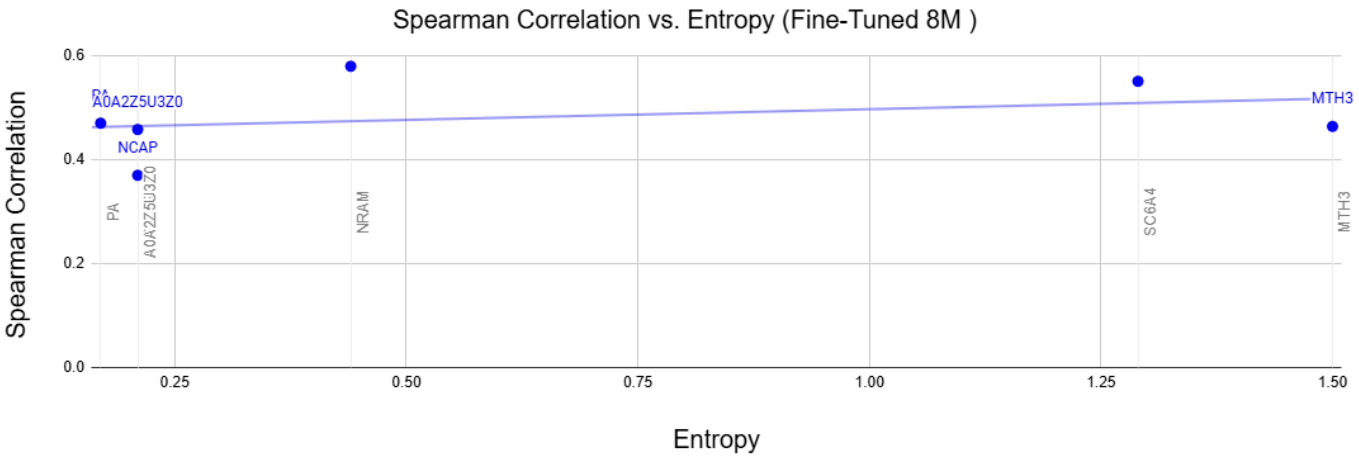


**Figure 9: Relationship between entropy and Spearman correlation for fine-tuned 8M models.** No strong systematic trend is observed, though many sequences with moderate entropy (around 1) achieve good correlation.

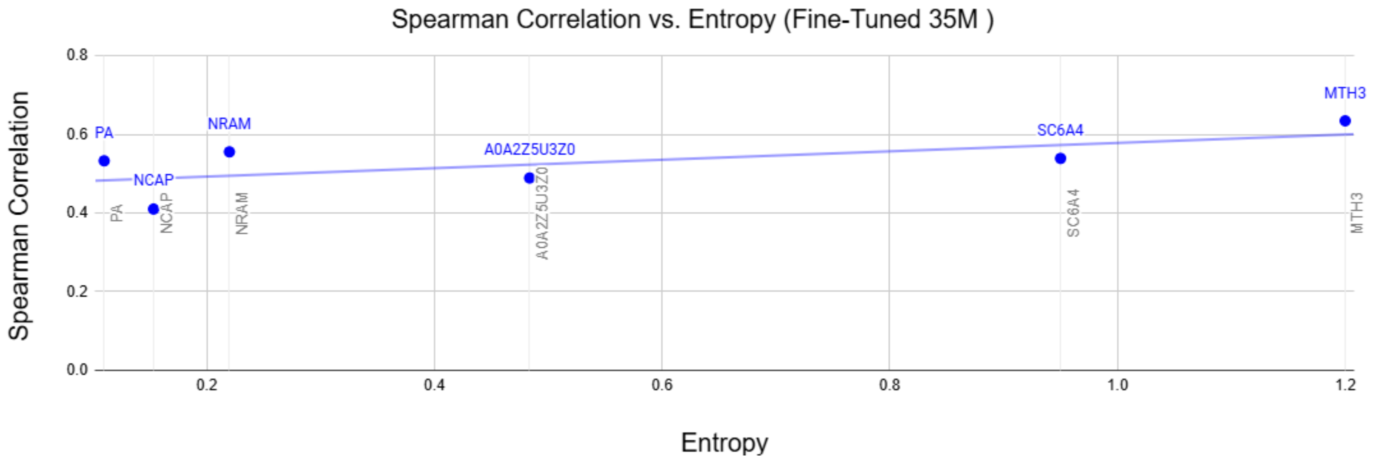


**Figure 10: Relationship between entropy and Spearman correlation for fine-tuned 35M models.** Similar to the 8M case, no strong monotonic relationship emerges.

Spearman correlation achieved after fine-tuning against the number of sequences used for each target protein (Fig. 11).

As shown in Figure 11, no clear trend emerges between the amount of augmented data and the resulting model performance. While a minimal number of sequences is likely necessary to provide sufficient evolutionary context, simply increasing the number of sequences does not guarantee improved predictive accuracy.

This suggests that the quality and diversity of the augmented sequences — not merely their quantity — play a critical role in fine-tuning success. Future work could explore smarter sequence

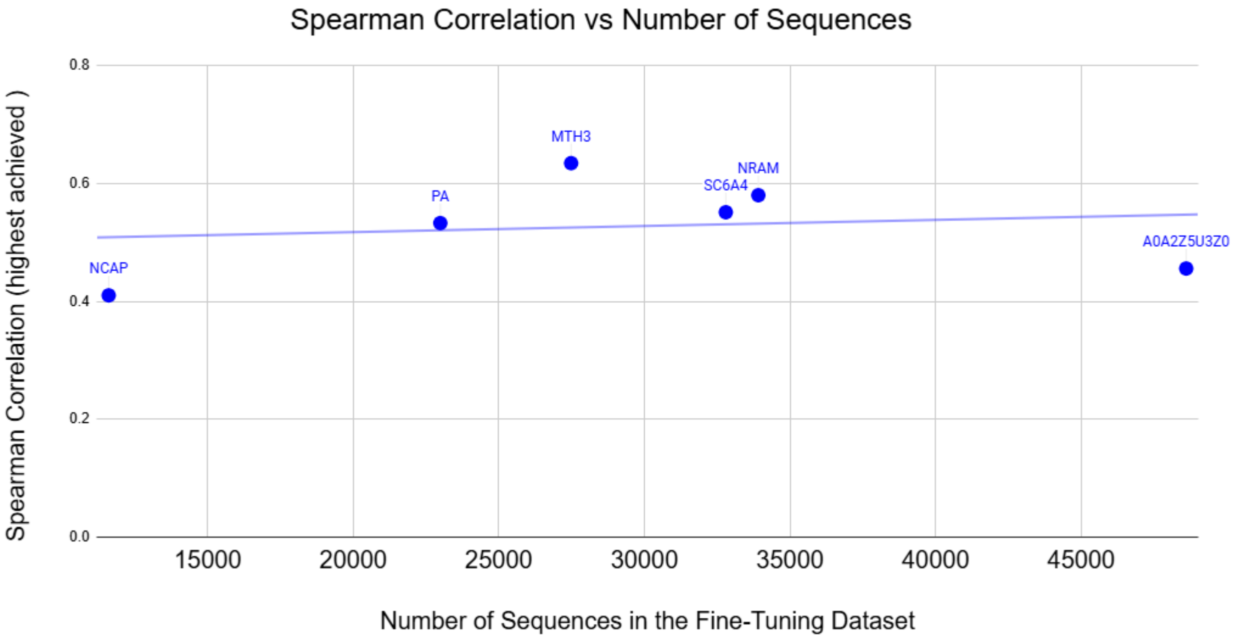


**Figure 11: Relationship between the number of augmented sequences and Spearman correlation after fine-tuning.** Each point corresponds to a different target protein.

selection strategies that prioritize maximizing evolutionary information over sheer sequence count.

## 4. Discussion

Our experiments highlight the effectiveness of task-specific fine-tuning for improving protein mutation prediction. Fine-tuning small ESM-2 models (8M and 35M parameters) with augmented datasets substantially enhances their ability to predict mutation effects, particularly for proteins underrepresented in standard training corpora like UniRef50.

By supplementing training with homologous sequences from UniRef100, we demonstrated that smaller models can match or exceed the predictive performance of substantially larger counterparts. The fine-tuned 8M and 35M models not only outperform their zero-shot versions but often approach or surpass the performance of a 650M parameter model trained on UniRef100. These results suggest that targeted data augmentation offers a computationally efficient alternative to model scaling for improving biological generalization.

Entropy dynamics during fine-tuning provided additional insights into the learning process. For some target proteins, rapid entropy reductions corresponded with early accuracy improvements, while for others, maintaining moderate entropy around 1 enabled more gradual but sustained learning. However, we observed no consistent relationship between average entropy and final predictive accuracy across all targets. Similarly, we found no clear correlation between the number of augmented sequences and performance improvement, suggesting that sequence quality and relevance outweigh quantity.

These findings point to a broader conclusion: achieving biological generalization in protein language models depends critically on the quality of evolutionary information available during training. Carefully designed fine-tuning strategies can close the performance gap even for models with drastically fewer parameters. Our results demonstrate that models over 1500 times smaller can achieve comparable predictive accuracy while maintaining greater biological realism—evidenced by moderate entropy levels rather than pure memorization.

The results underscore several key points:

- Model size alone is insufficient for robust biological generalization; training dataset composition plays a critical role.
- Moderate output entropy appears to support biologically plausible predictions, though further analysis is needed to fully characterize this relationship.

As machine learning applications in protein design, disease prediction, and evolutionary modeling continue to grow, these findings suggest that smaller, specialized models—adapted with attention to biological context—may offer a practical alternative to massive general-purpose models.

## 5. Future Work

Several promising directions emerge from this study:

### 5.1. Diversity-Based Data Augmentation

While we augmented target sequences with homologs meeting strict similarity thresholds, future work could explore augmentation strategies that prioritize functional diversity — selecting sequences that capture broader mutational landscapes while preserving biological relevance. Sampling strategies that maximize evolutionary variety without introducing excessive noise may further improve model generalization across diverse protein families.

### 5.2. Position-Specific Entropy Analysis

Our entropy analyses were averaged across entire sequences. A finer-grained approach that investigates entropy at individual sequence positions could reveal important insights, particularly at biologically constrained regions such as active sites, transmembrane domains, or intrinsically disordered regions. Understanding how uncertainty varies locally could help design better training objectives or adaptive masking strategies.

### 5.3. Extending to More Complex Targets

This study focused on relatively well-characterized viral and human proteins. Future work should test this fine-tuning strategy on more difficult targets — such as membrane proteins, highly dynamic disordered proteins, or rapidly evolving pathogen proteins — to assess its robustness and generalizability across a broader biological space.

### 5.4. Entropy-Guided Training and Model Diagnostics

While no strong or monotonic relationship is observed between entropy and predictive accuracy, we notice a weak trend where moderately higher entropy values are associated with somewhat improved Spearman correlations. Additional research, particularly with proteins exhibiting greater sequence variability, could better clarify the role of entropy in biological generalization. In particular, future studies could investigate whether an optimal entropy threshold exists that could serve as a stopping criterion during training — signaling when sufficient biological learning has been achieved to maximize mutation effect prediction accuracy.

## 6. Data Availability

Resources that guided the development of this project, including mutation scoring methods [7], UniRef50 sequence data [4], and the pretrained ESM-2 model [6], are publicly available on Hugging

Face.

All code developed for this work, including model architecture, fine-tuning framework, and evaluation scripts, is available on GitHub at:

https://github.com/AylinHad/IW09_Spring_2025

## 7. Acknowledgements

I would like to thank Professor Mona Singh for her support and guidance throughout the term, both during weekly seminar meetings and outside of class. I would also like to thank James Fife for his help throughout the research process from idea selection and dataset collection to suggesting implementation resources.

## References

[1] T. Bepler and B. Berger, "Learning the protein language: Evolution, structure, and function," *Cell Systems*, vol. 12, no. 6, pp. 654–669, 2021.

[2] N. Brandes, D. Ochoa, N. R. Lemoine, B. Bolognesi, K. Bryson, A. B. Gussow, S. Leo, T. L. Blundell, B. Rost, E. Petsalaki, and et al., "Genome-wide prediction of disease-variant effects with a deep protein-language model," *Nature Genetics*, vol. 55, no. 9, pp. 1527–1537, 2023.

[3] ChatGPT, "ChatGPT Responses," 2024. [Online]. Available: https://openai.com/chatgpt

[4] H. Face, "Uniref50 protein sequences dataset," https://huggingface.co/datasets/agemagician/uniref50, 2023, accessed: 2025-04-26.

[5] P. A. Gunnarsson and M. M. Babu, "Predicting evolutionary outcomes through the probability of accessing sequence variants," *Science Advances*, vol. 9, no. 30, p. eade2903, 2023.

[6] F. A. Research, "Esm-2 8m ur50d pretrained model," https://huggingface.co/facebook/esm2_t6_8M_UR50D, 2023, accessed: 2025-04-26.

[7] A. Schreiber, "Mutation scoring with protein language models," https://huggingface.co/blog/AmelieSchreiber/mutation-scoring, 2023, accessed: 2025-04-26.

[8] M. Sourisseau, D. J. P. Lawrence, M. C. Schwarz, C. H. Storrs, E. C. Veit, J. D. Bloom, and M. J. Evans, "Deep mutational scanning comprehensively maps how zika envelope protein mutations affect viral growth and antibody escape," *Journal of Virology*, vol. 93, no. 23, pp. e01 291–19, 2019.

[9] Y. Sun and Y. Shen, "Structure-informed protein language models are robust predictors for variant effects," *Human Genetics*, vol. 142, no. 6, pp. 693–707, 2023.