

MuTaP: Multi-Task Mutation Predictor via LoRA-Adapted ESM-2

Stanford CS224N Custom Project

Aya Aburous

Department of Computer Science
Stanford University
aaburous@stanford.edu

Jad Bitar

Department of Computer Science
Stanford University
jadbitar@stanford.edu

Abstract

Interpreting missense variants is still a major bottleneck in clinical genetics: most existing predictors focus only on pathogenicity and many rely on multiple sequence alignments (MSAs). In this paper, we present **MuTaP** (Multi-Task Mutation Predictor), a parameter-efficient model that fine-tunes ESM-2 (650M) [1] using Low-Rank Adaptation (LoRA) [2], updating only 0.21% of the model parameters. MuTaP is trained jointly on two supervised tasks: pathogenicity classification and a 5-class molecular mechanism prediction head (LOF, GOF, dominant-negative, haploinsufficiency, neutral), with Kendall uncertainty-weighted multi-task loss [3]. The mechanism labels are weakly supervised and were extracted from 6.2M ClinVar free-text submission descriptions using keyword matching with negation detection. On a gene-level held-out test set of 27,122 ClinVar variants, MuTaP achieves a pathogenicity AUROC of 0.900 and a mechanism macro-F1 of 0.310, compared to a majority-class baseline of 0.195 for mechanism prediction. This performance is largely driven by strong LOF classification (F1 0.977), while the rarer mechanism classes remain difficult. A single-task ablation shows that multi-task training gives a +0.4% AUROC gain in-distribution, but a -1.6% drop on the out-of-distribution ProteinGym benchmark [4], which highlights a tradeoff between task interference and regularization. On ProteinGym, MuTaP reaches a mean per-protein AUROC of 0.833 across all 2,164 evaluated proteins without using MSAs. The architecture also includes heads for stability ($\Delta\Delta G$) and structural location, but we were not able to evaluate them in this study because of the lack of mappable human thermodynamic data.

1 Key Information

- **Mentor:** Fang Wu
- **External Collaborators:** None
- **Sharing project:** No

2 Introduction

Missense variants are single amino-acid substitutions caused by single nucleotide changes. They make up the majority of known disease-causing mutations. ClinVar [5] catalogs hundreds of thousands of these variants, but most of them still lack enough evidence to be classified confidently. This variant interpretation gap remains a central bottleneck in clinical genetics: when a patient carries a Variant of Uncertain Significance (VUS), clinicians are limited in how confidently they can advise on disease risk, treatment decisions or family screening.

State-of-the-art pathogenicity predictors such as REVEL [6] and ClinPred [7] achieve AUROCs near 0.97–0.98 on standardized benchmarks and broader tools such as CADD [8] also remain highly

competitive. However, these methods share two important limitations. First of all, they focus only on pathogenicity scoring, offering no mechanistic insight into how a variant causes disease. And second of all, many rely heavily on conservation-based and other engineered features, often produced through MSA-based pipelines, which can limit scalability.

Protein language models (PLMs) such as ESM-2 [1] capture evolutionary and structural patterns without requiring MSAs. ESM-1v introduced zero-shot variant scoring through masked log-likelihood ratios [9] and AlphaMissense [10] is a strong recent model for missense variant effect prediction. Still, all of these models ultimately give an output of only a single pathogenicity score.

From an NLP perspective, this problem setting is especially interesting because it brings together several core challenges: transfer learning from a self-supervised protein language model to a supervised clinical task, parameter-efficient fine-tuning (PEFT) of a 650M-parameter transformer under GPU memory constraints, multi-task learning with heterogeneous losses and imbalanced label coverage, weak supervision using keyword-based label extraction from unstructured clinical text and out-of-distribution generalization from ClinVar training data to the ProteinGym benchmark.

Our contribution. We present MuTaP, a multi-task architecture that extends ESM-2 with LoRA adapters and task-specific heads, and is trained using uncertainty-weighted loss. Our main contributions are:

1. **Pathogenicity prediction** through LoRA fine-tuning, improving AUROC by +3.0% over ESM-2 zero-shot.
2. **A novel weakly supervised mechanism classifier** (5-class), built using labels extracted from 6.2M ClinVar submission descriptions.
3. **A single-task ablation** that reveals an in-distribution vs. out-of-distribution tradeoff introduced by multi-task training.
4. **ProteinGym evaluation** across 2,164 proteins, including a leaderboard comparison against 30 established models.

The architecture also includes heads for thermodynamic stability ($\Delta\Delta G$) and structural location (core vs. surface), but because mappable human stability data was unavailable and AlphaFold annotation was deferred, these heads had zero label coverage; we discuss them as future extensions in Section 7.

3 Related Work

Pathogenicity prediction. Conservation-based ensemble methods such as REVEL [6] and ClinPred [7] combine dozens of evolutionary and structural features, achieving mean per-protein AUROCs of 0.97–0.98 on the ProteinGym clinical benchmark [4]. These methods rely heavily on conservation and other engineered features, many of which come from MSA-based pipelines. On the other hand, CADD [8] takes a different approach by training on simulated versus observed variants, but it still outputs only a pathogenicity score.

Protein language models for variant effects. ESM-1v [9] showed that masked log-likelihood ratios from a pre-trained PLM can be used to score variants in a zero-shot setting. EVE [11] trains a protein-family-specific VAE on MSAs. PoET [12], which ranks 9th on the ProteinGym clinical leaderboard with an AUROC of 0.920, extends autoregressive generation to variant. AlphaMissense [10] fine-tunes a PLM with structural context. To our knowledge, none of these models produce mechanism predictions alongside pathogenicity, which is the gap MuTaP aims to address.

Parameter-efficient fine-tuning. LoRA [2] adds trainable low-rank matrices to attention projections, making fine-tuning possible at a small fraction of the full parameter cost. Following prior work showing that LoRA is effective for protein language models [13], we adapt the Key and Value projections in all attention layers.

Multi-task learning and uncertainty weighting. The Kendall uncertainty-weighted loss [3] balances heterogeneous tasks by learning per-task noise parameters σ_i , which automatically down-weight noisier objectives. This is especially well suited to our setting, where label covers a wide range from 100% (pathogenicity) to 18% (mechanism).

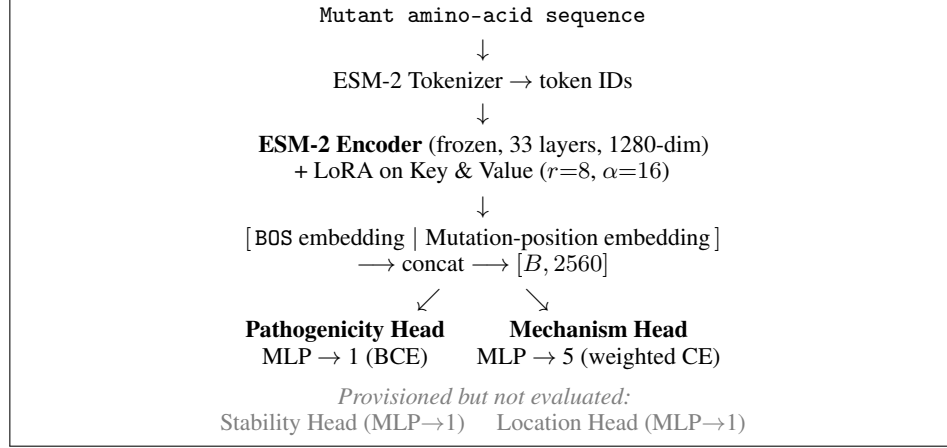


Figure 1: MuTaP architecture. The shared ESM-2 backbone with LoRA adapters feeds two evaluated task heads (pathogenicity and mechanism) and two provisioned heads (stability and location, shown in gray) that received no training signal.

4 Approach

4.1 Model Architecture

Our backbone model is ESM-2 650M (esm2_t33_650M_UR50D) [1], a 33-layer transformer encoder with 1280-dimensional hidden states, pre-trained on $\sim 65\text{M}$ protein sequences from UniRef50. All backbone weights remain frozen during fine-tuning.

LoRA adaptation. We apply LoRA [2] with rank $r=8$ and scaling $\alpha=16$ to the Key and Value projection matrices in all 33 attention layers. Each adapter adds low-rank matrices $\mathbf{B} \in \mathbb{R}^{d \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times d}$ on top of the frozen weight $\mathbf{W}_0$:

$$\mathbf{h} = \left(\mathbf{W}_0 + \frac{\alpha}{r} \mathbf{B} \mathbf{A} \right) \mathbf{x}. \quad (1)$$

This gives us a total of 1,351,680 trainable parameters out of 652,392,341 overall (which is about 0.21%).

Input representation. Each input is the mutant amino-acid sequence. For proteins longer than 512 residues, we extract a centered 512-residue window around the mutation site. After encoding, we take the first-position token to capture global context and the mutation-position token to capture local residue context, then we concatenate them into a shared 2560-dimensional representation.

Task heads. Each task head follows the same architecture: Linear $\rightarrow$ Layer-Norm $\rightarrow$ GELU $\rightarrow$ Dropout $\rightarrow$ Linear.

- **Pathogenicity** (binary): 2560 $\rightarrow$ 512 $\rightarrow$ 1; BCE loss.
- **Mechanism** (5-class): 2560 $\rightarrow$ 512 $\rightarrow$ 256 $\rightarrow$ 5; weighted cross-entropy with inverse-frequency class weights.

The model also includes a stability head (2560 $\rightarrow$ 256 $\rightarrow$ 64 $\rightarrow$ 1; MSE loss, with optional WT+mutant 5120-dim input) and a location head (2560 $\rightarrow$ 256 $\rightarrow$ 1; BCE loss). However, these received zero training signal in the current study because of the data constraints described in Section 5.1. The full architecture is shown in Figure 1.

4.2 Multi-Task Loss

We use the homoscedastic uncertainty weighting method introduced by Kendall et al. [3]. For each active task i , the loss contribution is:

$$\mathcal{L}_{\text{total}} = \sum_{i \in \mathcal{A}(b)} e^{-s_i} \mathcal{L}_i + s_i, \quad (2)$$

Table 1: Dataset split statistics (gene-level, no gene overlap).

Split	Variants	Path%	Mech-labeled	Genes
Train	136,915	42.7%	24,715 (18.1%)	~10,290
Val	29,908	43.5%	5,558 (18.6%)	~2,205
Test	27,122	40.5%	4,716 (17.4%)	~2,205

where $s_i = \log \hat{\sigma}_i^2$ is a learnable scalar and $\mathcal{A}(b)$ denotes the set of tasks with at least one labeled sample in batch b . If no sample in a batch has a label for task i , the entire contribution of that task (both the precision-weighted loss term and the log-variance regularizer) is masked to zero. In practice, because stability and location had zero labels across the full dataset, only pathogenicity and mechanism contributed to the loss during training. The negative total loss observed from epoch 2 onward is expected under this formulation: because as the individual task losses $\mathcal{L}_i$ decrease, the optimal per-task contribution approaches $1 + \log \mathcal{L}_i$, which becomes negative once $\mathcal{L}_i < e^{-1} \approx 0.37$.

4.3 Gene-Level Data Split

We split variants by gene symbol (`GroupShuffleSplit`), ensuring that there is no gene overlap between the train, validation and test sets. This prevents information leakage through gene-level patterns and gives a more rigorous evaluation of how well the model generalizes to genes that are unseen.

5 Experiments

5.1 Data

ClinVar. We use the ClinVar database [5] (`variant_summary.txt`, Feb. 2026; 8.8M rows). We keep SNVs with unambiguous clinical significance (Pathogenic/Likely pathogenic $\rightarrow$ 1; Benign/Likely benign $\rightarrow$ 0) at least 1-star review status, GRCh38 coordinates and germline origin. Gene symbols are mapped to canonical Swiss-Prot sequences (20,431 human entries) and wild-type amino acids are verified against the reference sequence (mismatches are discarded). The final dataset contains 193,945 variants across ~14,700 genes.

Mechanism labels (weak supervision). We extract mechanism labels from 6.2M ClinVar submission free-text descriptions using keyword matching with a 5-word negation window, then resolve labels by majority vote across submissions for each variant. This assigns one of five classes (which are Loss-of-Function (LOF), Gain-of-Function (GOF), Dominant-Negative (DN), Haploinsufficiency (HAPLO) and Neutral) to 34,989 variants, which corresponds to 18.0% coverage. The class distribution is highly skewed: LOF 95.4%, Neutral 2.7%, GOF 0.9%, DN 0.7%, and HAPLO 0.3%. We apply inverse-frequency class weights during training.

These labels are weakly supervised, since they are derived from clinical free text rather than expert-curated gold annotations. The 95.4% dominance of LOF may partly reflect labeling bias, LOF-related keywords likely appear more often in submissions, rather than true biological prevalence. Validation of the top-labeled genes still shows biological coherence: the top LOF genes include *BRCA2*, *NF1*, *ATM*, *BRCA1* and *APC*, all of which are well-established tumor suppressors with documented loss-of-function mechanisms (see Appendix B).

Stability and location (not evaluated). The MegaScale dataset [14] contains around 1.3M $\Delta\Delta G$ measurements for engineered and bacterial proteins, but none of these mapped to human UniProt IDs, resulting in 0% stability coverage. AlphaFold-based structural location annotation was also deferred because of time constraints. As a result, both heads remained part of the architecture but received no training signal. The uncertainty loss (Eq. 2) handles this by fully masking their contributions.

ProteinGym benchmark. For out-of-distribution evaluation, we use the ProteinGym clinical substitution benchmark [4], which includes 2,525 proteins and 2,164 of which have sufficient variant diversity for per-protein AUROC evaluation.

Table 2: Training configuration.

Hyperparameter	Value
Backbone	ESM-2 650M (frozen)
LoRA rank / α / targets	8 / 16 / Key + Value
Trainable parameters	1,351,680 (0.21%)
Batch size (actual / effective)	8 / 64 (gradient accumulation $\times 8$)
Optimizer	AdamW, lr = 10^{-4} , cosine + 5% warmup
Weight decay / grad clip	0.01 / max norm 1.0
Max sequence length	512 (centered window)
Precision	fp16 mixed
Epochs (total)	6 (4 initial + 2 resumed)
Training time	≈ 90 h (≈ 14.5 h/epoch)

5.2 Evaluation Method

In terms of pathogenicity, we report AUROC as the primary metric, along with AUPRC, F1 at a threshold of 0.5 and MCC. For mechanism, we report macro-F1, which treats all 5 classes equally and is the most appropriate metric given the extreme class imbalance. Because our 5-class mechanism dataset is a novel contribution extracted from free-text annotations, there are no established baseline models for this exact task at this scale. Therefore, we compare against a majority-class baseline (always predicting LOF; macro-F1 0.195) and a uniform random baseline (macro-F1 0.088). The ProteinGym protocol computes AUROC separately for each protein and then averages across proteins. This prevents the evaluation from being dominated by proteins with many variants.

5.3 Experimental Details

We performed training on a single GCP NVIDIA L4 GPU (23 GB VRAM). We used the HuggingFace Transformers and PEFT libraries for model loading and LoRA insertion, while all task heads, the training loop and the evaluation code were implemented by us. Table 2 summarizes the key hyperparameters.

We ran training for 6 epochs where each epoch took around 15 hours. A checkpoint-saving bug in the initial run persisted the epoch 1 state (val AUROC 0.901) rather than the true best at epoch 2 (val AUROC 0.904). Epochs 5–6, run after fixing the bug, confirmed that additional training did not recover epoch 2 performance. All reported test results use the epoch 1 checkpoint; the likely impact is small, as the validation gap between epoch 1 and epoch 2 was only 0.003 and later resumed training did not recover or exceed epoch 2 performance.

5.4 Results

Core ablation. Table 3 shows our main comparison. ESM-2 zero-shot (masked log-likelihood ratio, no fine-tuning) achieves an AUROC of 0.870 on a sampled subset of the ClinVar test split (3,029 of 5,000 variants were successfully scored).¹ Single-task LoRA fine-tuning (pathogenicity head only) improves this by +2.6%, reaching 0.896, while multi-task training adds another +0.4% to reach 0.900. On ProteinGym, an out-of-distribution benchmark with different proteins and curation, the pattern reverses: we found that the single-task model generalizes better (0.849 vs. 0.833), which is consistent with the hypothesis that the mechanism head may introduce ClinVar-specific bias and thus, reduce transfer. Mechanism classification requires joint multi-task training and is not applicable to the pathogenicity-only ablation.

Full test-set metrics. Table 4 reports detailed metrics on the 27,122 held-out test variants. The pathogenicity AUROC of 0.900 meets our project proposal target (of ≥ 0.90). For mechanism prediction, LOF achieves an F1 of 0.977, indicating strong performance on the dominant class. GOF (F1 0.221) and Neutral (F1 0.351) suggest that the model is able to learn partially even from relatively small labeled sets (318 and 945 total labeled examples, respectively). By contrast, DN

¹39.4% of sampled variants had missing sequences or lookup failures; the AUROC estimate carries $\sim \pm 0.01$ uncertainty. MuTaP results are reported on the full 27,122-variant test set, so this comparison is approximate.

Table 3: Ablation: pathogenicity and mechanism by model configuration. Zero-shot is estimated on a subset ($\dagger$). Mechanism is N/A for the single-task model, which has no trained mechanism head.

Model	ClinVar AUROC	ProteinGym AUROC	Mech Macro-F1
ESM-2 zero-shot $\dagger$	0.870	—	—
Single-task (path only)	0.896	0.849	—
Multi-task (MuTaP)	0.900	0.833	0.310
Majority baseline	—	—	0.195
Random baseline	—	—	0.088

Table 4: Test-set results (27,122 variants; epoch 1 checkpoint).

Metric	Score
<i>Pathogenicity</i>	
AUROC	0.900
AUPRC	0.877
F1 (threshold 0.5)	0.753
MCC	0.604
Accuracy	0.811
<i>Mechanism</i> (4,716 labeled test variants)	
Macro-F1	0.310
LOF / GOF / DN / HAPLO / Neutral F1	0.977 / 0.221 / 0.0 / 0.0 / 0.351

(F1 0.0, $N=242$) and HAPLO (F1 0.0, $N=100$) receive zero predictions; the extreme imbalance relative to 33,384 LOF examples makes these classes effectively non-learnable at this scale, even with inverse-frequency weighting.

ProteinGym leaderboard. Table 5 places MuTaP in context. Using direction-corrected ProteinGym baseline scores (with details in Appendix A), MuTaP ranks 27th of 31 with a mean AUROC of 0.833. Although this is lower than pathogenicity-specialized ensembles, MuTaP is the only system on this leaderboard that also provides simultaneous mechanism classification. MuTaP evaluates all 2,164 eligible proteins without MSAs, matching PoET and EVE for the most complete coverage.

6 Analysis

6.1 Multi-Task vs. Single-Task Tradeoff

The ablation in Table 3 reveals a nuanced tradeoff. On ClinVar (in-distribution), the multi-task model outperforms the single-task model by +0.4%, suggesting that the mechanism head acts as a mild regularizer and enriches the shared representation. However, if we look at ProteinGym (out-of-distribution) the opposite pattern appears: single-task performs better by +1.6%, indicating that the mechanism head may introduce a ClinVar-specific distributional bias that hurts transfer to novel protein families. This provides clear evidence of task interference: the shared LoRA adapters are optimized jointly for both objectives and the representation that is best for ClinVar mechanism classification does not perfectly align with the one that best supports broad-spectrum pathogenicity prediction. For ClinVar-like in-distribution evaluation, multi-task appears preferable. But for generalization to novel proteins, single-task offers better transfer.

6.2 Training Dynamics and Task Interference

Table 6 shows that the two tasks converge differently over training. Pathogenicity AUROC peaks at epoch 2 (0.904) and then declines monotonically through epoch 6 (0.897). The macro-F1 mechanism, on the contrary, improves steadily from 0.288 (epoch 1) to 0.329 (epoch 6). This is strong evidence that the two objectives operate on different convergence timescales: pathogenicity patterns are captured quickly by the pre-trained representations, whereas mechanism classification (which is a harder task with 18% label coverage and extreme imbalance) requires more updates to separate

Table 5: ProteinGym clinical benchmark (selected models, direction-corrected).

Rank	Model	Mean AUROC	Proteins
1	ClinPred [7]	0.982	2,128
2	MetaRNN	0.979	2,139
3	BayesDel (addAF)	0.971	2,134
9	PoET [12]	0.920	2,164
14	CADD [8]	0.904	2,139
17	ESM-1b [15]	0.891	2,164
27	MuTaP (ours)	0.833	2,164
28	MutationTaster	0.811	2,125
31	FATHMM	0.724	2,134

Table 6: Validation metrics per epoch (multi-task training).

Epoch	Val AUROC	Val Mech F1	Notes
1	0.901	0.288	checkpoint saved (bug)
2	0.904	0.305	true best (path)
3	0.903	0.305	
4	0.898	0.325	
5	0.897	0.320	resumed
6	0.897	0.329	best mechanism F1

minority classes. Future work could explore task-specific learning-rate scheduling or early stopping for each head.

6.3 Distribution Shift: ClinVar vs. ProteinGym

The 6.7-point gap between the ClinVar test AUROC (0.900) and the ProteinGym mean AUROC (0.833) reflects distribution shift rather than data leakage. There are three factors that likely contribute to this. First, ClinVar training variants come with strong expert evidence, whereas ProteinGym is curated differently and distributed across different proteins and variant sets. Second, although gene-level splitting ensures that test proteins are unseen, the protein families in ProteinGym may differ even more substantially from those seen during training. Third, per-protein averaging gives equal weight to proteins with 10 variants and proteins with 10,000 variants, which amplifies the effect of especially difficult proteins. The median per-protein AUROC of 0.896 suggests strong performance on most proteins. The mean is pulled down by a tail of harder cases.

6.4 Qualitative Analysis: Case Studies

To illustrate MuTaP’s predictions in practice, we sampled 15 variants from the held-out test set in well-studied genes such as *TP53*. For each variant, we compared the predicted pathogenicity probability and predicted mechanism class against the ground-truth ClinVar annotations. In this sample, we found that 14 of the 15 pathogenicity predictions were correct. The one error was the *TP53* G293W variant, which is labeled benign but was incorrectly assigned a borderline pathogenic probability of 0.541. For variants with available mechanism labels, the model’s top predicted mechanism was often consistent with the ClinVar-derived label. For example, pathogenic variants known to act through loss-of-function or dominant-negative mechanisms received both high pathogenicity scores and LOF as the top predicted mechanism. This indicates that the model is learning biologically coherent representations beyond what is captured by the aggregate metrics alone.

6.5 Per-Protein Error Analysis

Of the 2,164 proteins, approximately 48% achieve AUROC >0.90, while 119 (5.5%) fall below random performance (AUROC <0.50). In particular, 13 proteins have an AUROC of exactly 0.0. Our analysis shows that all 13 have very small variant sets ($N=5-17$), typically with only 1 pathogenic variant. At that scale, a single misranked variant is sufficient to make AUROC collapse, so this reflects

a limitation in statistical power rather than a systematic model failure. More broadly, the hardest proteins likely depend on structural, post-translational or tissue-context factors that are not captured by sequence alone or on gain-of-function mechanisms in which unusual mutations are pathogenic.

6.6 Mechanism Classification

The mechanism head shows a classic long-tail pattern. LOF dominates the labeled data (95.4%) and achieves an F1 of 0.977. GOF (318 labeled examples) and Neutral (945 labeled examples) show partial performance, with F1 scores of 0.221 and 0.351 respectively. DN (242 labeled examples) and HAPLO (100 labeled examples), by contrast, have zero F1 (no weighting strategy can fully compensate for class ratios of 100 versus 33,384). In practice, the model behaves as a reliable LOF vs. non-LOF discriminator, with partial ability to detect GOF and Neutral cases, which is still useful given that LOF is the dominant disease mechanism. A promising direction for future work would be to merge DN and HAPLO into LOF, because they are both closely related to loss-of-function behavior.

7 Conclusion

We show that LoRA fine-tuning of ESM-2 with only 0.21% trainable parameters can produce meaningful gains for clinical variant interpretation. Our key findings are as follows:

Fine-tuning adds clinical signal. MuTaP achieves an AUROC of 0.900 on the full held-out ClinVar test set, compared with 0.870 for an ESM-2 zero-shot estimate on a sampled subset of the same distribution which suggests an approximately 3-point gain from supervised fine-tuning.

Multi-task learning presents a tradeoff. Joint training with the mechanism head provides a +0.4% in-distribution gain but a −1.6% drop on ProteinGym which discloses task interference between pathogenicity and mechanism objectives that operate on different convergence timescales.

Mechanism prediction is novel and partially successful. The weakly supervised mechanism head achieves a macro-F1 of 0.310, which is 59% above the majority baseline, with particularly strong LOF classification (F1 0.977). DN and HAPLO remain non-functional because of extreme class imbalance.

Full ProteinGym coverage. MuTaP ranks 27th of 31 on the direction-corrected ProteinGym clinical leaderboard (with a mean AUROC 0.833), while evaluating all 2,164 proteins without MSAs.

Limitations. The stability ($\Delta\Delta G$) and structural location heads received zero training signal because mappable human thermodynamic data was unavailable and AlphaFold annotation was deferred, respectively. This means that the model functions as a two-task system in this study. The mechanism classifier also faces extreme class imbalance (95% LOF), and as a result, DN and HAPLO achieve zero F1.

Ethical considerations. We highlight two main ethical considerations. First, population bias: ClinVar has historically been skewed toward populations of European ancestry, which means the model may show higher error rates for variants from under-represented groups. We therefore recommend stratified evaluation whenever ancestry metadata is available. Second, clinical use: MuTaP is strictly a research tool, not a diagnostic device. Its predictions should never be used on their own for clinical decision-making.

Future work. Future work could: (1) integrate human $\Delta\Delta G$ data (for example, ProtaBank) with WT+mutant dual-sequence input for the stability head; (2) add AlphaFold pLDDT-based location labels; (3) perform LoRA rank ablations ($r \in \{4, 16, 32\}$) and target-projection ablations; (4) merge rare mechanism classes (DN + HAPLO $\rightarrow$ LOF); and (5) use task-specific early stopping to take advantage of the observed decoupled convergence of the pathogenicity and mechanism heads.

Team Contributions

Aya Aburous led the data pipeline (ClinVar preprocessing, UniProt mapping, mechanism label extraction via keyword matching with negation detection), designed the multi-task loss and task-head architecture and co-wrote the paper.

Jad Bitar led the training infrastructure (LoRA integration, trainer loop, fp16 support, checkpoint management), performed the full GCP training and single-task ablation runs, ran all evaluations and baseline comparisons (including ProteinGym and zero-shot) and co-wrote the paper.

References

- [1] Zeming Lin and Halil et al. Akin. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 2023.
- [2] Edward J Hu and Yelong et al. Shen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.
- [3] Alex Kendall and Yarin et al. Gal. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [4] Pascal Notin and Mafalda et al. Dias. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 2023.
- [5] Melissa J Landrum and Jennifer M et al. Lee. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 2015.
- [6] Nilah M Ioannidis and Joseph H et al. Rothstein. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics.*, 2016.
- [7] Najmeh Alirezaie and Kristin D et al. Kernohan. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *The American Journal of Human Genetics*, 2018.
- [8] Philipp Rentzsch and Daniela et al. Witten. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 2019.
- [9] Joshua Meier and Roshan et al. Rao. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 2021.
- [10] Jun Cheng and Guido et al. Novati. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664), 2023.
- [11] Jonathan Frazer and Pascal et al. Notin. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599, 2021.
- [12] Timothy F Truong Jr and Tristan Bepler. PoET: A generative model of protein families as sequences-of-sequences. *Advances in Neural Information Processing Systems*, 2023.
- [13] Samuel Sledzieski and Meghana et al. Kshirsagar. Democratizing protein language models with parameter-efficient fine-tuning. *Proceedings of the National Academy of Sciences*, 2024.
- [14] Kotaro Tsuboyama and Justas et al. Dauparas. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 2023.
- [15] Alexander Rives and Joshua et al. Meier. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 2020.

A Full ProteinGym Leaderboard

The full 31-model direction-corrected leaderboard is included in the submitted code archive (`results/full_leaderboard_corrected.csv`). In the raw ProteinGym files, five baseline models had inverted score directions (FATHMM, LRT, SIFT4G, SIFT, and Provean). After correcting for this, MuTaP moved from an originally calculated rank of 22 to its true rank of 27.

B Mechanism Label Validation

The top LOF-labeled genes by count ($N > 150$ labeled variants) are: BRCA2 (627), NF1 (538), ATM (451), DMD (422), BRCA1 (399), USH2A (275), APC (275), FBN1 (252), MSH6 (232) and PKD1 (227). All of these are well-established disease genes with documented loss-of-function mechanisms, which supports that the keyword-based extraction produces biologically coherent labels for the genes it covers.

C Test-Set Task Coverage

Task	Labeled variants	Coverage
Pathogenicity	27,122	100%
Mechanism	4,716	17.4%
Stability ($\Delta\Delta G$)	0	0%
Location (core/surface)	0	0%

Table 7: Label coverage across the four architectural heads on the test split.