TriFit: Trimodal Fusion with Protein Dynamics for Mutation Fitness Prediction

Seungik Cho [1]

Abstract

Predicting the functional impact of single amino acid substitutions (SAVs) is central to understanding genetic disease and engineering therapeutic proteins. While protein language models and structure-based methods have achieved strong performance on this task, they systematically neglect protein dynamics—residue flexibility, correlated motions, and allosteric coupling are well-established determinants of mutational tolerance in structural biology, yet have not been incorporated into supervised variant effect predictors. We present **TriFit**, a multimodal framework that integrates sequence, structure, and protein dynamics through a four-expert Mixture-of-Experts (MoE) fusion module with trimodal cross-modal contrastive learning. Sequence embeddings are extracted via masked marginal scoring with ESM-2 (650M); structural embeddings from AlphaFold2-predicted C_α geometries; and dynamics embeddings from Gaussian Network Model (GNM) B-factors, mode shapes, and residue-residue cross-correlations. The MoE router adaptively weights modality combinations conditioned on the input, enabling protein-specific fusion without fixed modality assumptions. On the ProteinGym substitution benchmark (217 DMS assays, 696k SAVs), TriFit achieves AUROC 0.897 ± 0.0002, outperforming all supervised baselines including Kermut (0.864) and ProteinNPT (0.844), and the best zero-shot model ESM3 (0.769). Ablation studies confirm that dynamics provides the largest marginal contribution over pairwise modality combinations, and TriFit achieves well-calibrated probabilistic outputs (ECE $= 0.044$) without post-hoc correction.

[1]Department of Physics and Astronomy, Rice University, Texas, USA. Correspondence to: Seungik Cho <seungikcho@rice.edu>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.

1. Introduction

Predicting whether a single amino acid substitution (SAV) disrupts protein function is a fundamental challenge in computational biology, with applications ranging from genetic disease diagnosis to therapeutic protein engineering (Fowler & Fields, 2014; Notin et al., 2023a). Despite a surge in machine learning-based approaches—including sequence-based protein language models (Lin et al., 2023; Meier et al., 2021), structure-based inverse folding models (Hsu et al., 2022; Dauparas et al., 2022), and hybrid methods combining both (Notin et al., 2022; Su et al., 2024; Jumper et al., 2021)—a critical third source of biophysical information remains overlooked: *protein dynamics*. Residue flexibility, correlated motions, and allosteric coupling—captured by elastic network models such as the GNM (Bahar et al., 1997; Haliloglu et al., 1997)—are well-established determinants of mutational tolerance, yet have not been incorporated into any supervised variant effect predictor. Furthermore, existing fusion strategies treat all modalities uniformly, whereas the relative informativeness of sequence, structure, and dynamics varies considerably across mutation types, motivating an adaptive routing mechanism.

In this work, we present **TriFit** (**Tri**modal **Fit**ness predictor), integrating all three modalities through a Mixture-of-Experts (MoE) fusion module with trimodal contrastive learning. Our key contributions are:

- We introduce **protein dynamics as a third modality** for variant effect prediction, extracting GNM-based B-factors, mode shapes, and residue-residue cross-correlations as per-residue embeddings from AlphaFold2 structures.
- We propose an **input-conditioned MoE fusion** module with four specialized experts (Seq+Struct, Seq+Dyn, Struct+Dyn, Trimodal) and a learned router that adaptively weights modality combinations based on the projected multimodal representation.
- We apply **trimodal cross-modal contrastive learning** across all three modality pairs simultaneously via InfoNCE loss (van den Oord et al., 2018), aligning complementary representations in a unified latent space.
- On the **ProteinGym substitution benchmark** (Notin

et al., 2023a) (217 DMS assays, 696k SAVs), TriFit achieves AUROC 0.897 ± 0.0002, surpassing all supervised baselines including Kermut (AUROC 0.864) and ProteinNPT (Notin et al., 2023b) (AUROC 0.844), and the best zero-shot model ESM3 (Hayes et al., 2024) (AUROC 0.769).

2. Method

Figure 1 illustrates the overall TriFit architecture. Three modality-specific encoders (frozen) extract per-residue embeddings, which are projected to a shared space and fused through a four-expert MoE module. Cross-modal contrastive loss aligns the three modality representations during training, and the fused representation is passed to a binary classifier for fitness prediction.

2.1. Problem Formulation

Given a wild-type protein sequence $\mathbf{s} \in \mathcal{A}^L$ of length L over the amino acid alphabet $\mathcal{A}$, and a single amino acid variant (SAV) $v = (i, a_i \to a_i')$ that substitutes residue a_i at position i with a_i', we aim to predict a binary fitness label $y \in \{0, 1\}$, where $y = 1$ denotes a functional (fit) variant and $y = 0$ denotes a damaging variant. Labels are derived from experimentally measured DMS fitness scores via top/bottom 30% thresholding (Notin et al., 2023a).

2.2. Multimodal Embedding Extraction

TriFit extracts per-variant embeddings from three complementary modalities. All modality encoders are kept *frozen*; only the downstream fusion module and classifier are trained.

Sequence Embedding. We adopt the masked marginal scoring strategy of ESM-1v (Meier et al., 2021), which produces contextualized representations without mutant-specific forward passes. For each variant $(i, a_i \to a_i')$, we mask position i in the wild-type sequence and perform a single forward pass through ESM-2 (650M) (Lin et al., 2023) to obtain the context embedding $\mathbf{h}_i^{\text{seq}} \in \mathbb{R}^{1280}$. The final sequence embedding is:

$$\mathbf{e}^{\text{seq}} = \mathbf{h}_i^{\text{seq}} + \left(\mathbf{t}_{a_i'} - \mathbf{t}_{a_i}\right), \tag{1}$$

where $\mathbf{t}_a \in \mathbb{R}^{1280}$ denotes the token embedding of amino acid a extracted from the ESM-2 embedding table. This formulation captures both the structural context of position i and the amino acid substitution signal, requiring only L_{protein} forward passes per protein regardless of the number of variants.

Structure Embedding. We use AlphaFold2-predicted structures (Jumper et al., 2021) provided by the ProteinGym benchmark. For each protein, we construct a k-nearest neighbor graph over C_α atoms and extract per-residue geometric features comprising inter-residue distances and direction vectors to the $k = 20$ nearest neighbors. These features are projected to $\mathbb{R}^{512}$ via a fixed random projection matrix $\mathbf{W} \in \mathbb{R}^{83\times512}$, yielding the structural embedding $\mathbf{e}^{\text{str}} \in \mathbb{R}^{512}$ at the mutation site.

Dynamics Embedding. We compute protein dynamics features using the Gaussian Network Model (GNM) (Bahar et al., 1997) applied to each AlphaFold2 structure via ProDy (Bakan et al., 2011). For each protein, we calculate: (i) normalized B-factors $\mathbf{b} \in \mathbb{R}^L$ (residue flexibility); (ii) the top-K GNM mode shapes $\mathbf{U} \in \mathbb{R}^{L\times K}$ (K=20); and (iii) mode-projected cross-correlations $\mathbf{C} \in \mathbb{R}^{L\times K}$. Additionally, diagonal stiffness values from the Kirchhoff matrix provide a complementary rigidity measure. These four feature types are concatenated per residue and projected to $\mathbf{e}^{\text{dyn}} \in \mathbb{R}^{256}$ via a fixed random projection.

2.3. Mixture-of-Experts Fusion Module

Let $\mathbf{z}^m = \text{Proj}_m(\mathbf{e}^m)$ denote the projected embedding of modality $m \in \{\text{seq, str, dyn}\}$, where each Proj_m is a learnable MLP followed by LayerNorm, mapping to a common dimension d=512.

We define four specialized experts that process distinct modality combinations:

$$\begin{aligned} \mathbf{f}_1 &= \text{Expert}_1([\mathbf{z}^{\text{seq}}; \mathbf{z}^{\text{str}}]), \quad \mathbf{f}_2 = \text{Expert}_2([\mathbf{z}^{\text{seq}}; \mathbf{z}^{\text{dyn}}]), \\ \mathbf{f}_3 &= \text{Expert}_3([\mathbf{z}^{\text{str}}; \mathbf{z}^{\text{dyn}}]), \quad \mathbf{f}_4 = \text{Expert}_4([\mathbf{z}^{\text{seq}}; \mathbf{z}^{\text{str}}; \mathbf{z}^{\text{dyn}}]), \end{aligned} \tag{2}$$

where $[\cdot;\cdot]$ denotes concatenation and each expert is a two-layer MLP with GELU activation. The router produces soft assignment weights:

$$\mathbf{w} = \text{Softmax}\left(\text{Router}\left([\mathbf{z}^{\text{seq}}; \mathbf{z}^{\text{str}}; \mathbf{z}^{\text{dyn}}]\right)\right) \in \mathbb{R}^4, \tag{3}$$

and the fused representation is computed as a weighted sum:

$$\mathbf{f} = \sum_{k=1}^{4} w_k \mathbf{f}_k \in \mathbb{R}^{512}. \tag{4}$$

2.4. Cross-Modal Contrastive Learning

To encourage alignment of complementary information across modalities, we apply symmetric InfoNCE loss (van den Oord et al., 2018) to all three modality pairs within each mini-batch of size B:

$$\mathcal{L}_{\text{ctr}} = \frac{1}{3} \sum_{(m,m')\in\mathcal{P}} \mathcal{L}_{\text{NCE}}(\mathbf{z}^m, \mathbf{z}^{m'}), \tag{5}$$

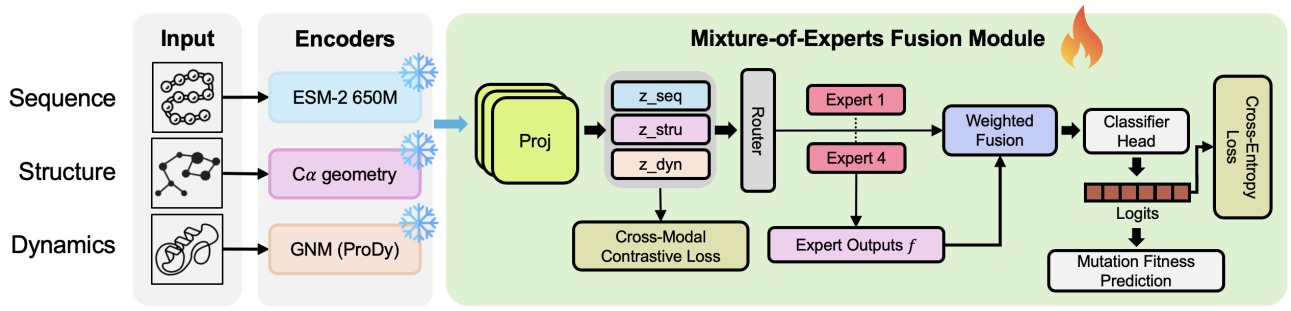

Figure 1. TriFit architecture. Sequence, structure, and dynamics encoders (frozen) extract modality-specific embeddings. A learned projection maps each to a shared 512-dim space. The four-expert MoE router adaptively combines modality pairs (E1: Seq+Struct, E2: Seq+Dyn, E3: Struct+Dyn, E4: Trimodal) via soft gating. Cross-modal contrastive loss aligns all three modality pairs during training. The weighted-fused representation $\mathbf{f}$ is passed to a binary classifier predicting functional vs. damaging variants.

where $\mathcal{P} = \{(\text{seq,str}), (\text{seq,dyn}), (\text{str,dyn})\}$ and:

$$\mathcal{L}_{\text{NCE}}(\mathbf{z}, \mathbf{z}') = -\frac{1}{2}\left[\log \frac{e^{\text{sim}(\mathbf{z}_i, \mathbf{z}'_i)/\tau}}{\sum_j e^{\text{sim}(\mathbf{z}_i, \mathbf{z}'_j)/\tau}} + \log \frac{e^{\text{sim}(\mathbf{z}'_i, \mathbf{z}_i)/\tau}}{\sum_j e^{\text{sim}(\mathbf{z}_j, \mathbf{z}'_i)/\tau}}\right] \tag{6}$$

with cosine similarity $\text{sim}(\cdot, \cdot)$ and temperature τ=0.07.

2.5. Training Objective

The fused representation $\mathbf{f}$ is passed through a two-layer MLP classifier with dropout (p=0.1) to produce logits for binary classification. The overall training objective combines cross-entropy classification loss with the trimodal contrastive loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{ctr}}, \tag{7}$$

where λ=0.3 balances the two terms. We train for 20 epochs using AdamW (Loshchilov & Hutter, 2019) with learning rate 3×10^{-4}, weight decay 10^{-4}, and cosine annealing (Loshchilov & Hutter, 2017). All experiments use the official ProteinGym 5-fold cross-validation splits (`fold_random_5`) for train/validation/test partitioning to ensure comparability with reported baselines.

3. Experiments

3.1. Experimental Setup

Dataset. We evaluate TriFit on the ProteinGym substitution benchmark (Notin et al., 2023a), which comprises 217 deep mutational scanning assays spanning 696,311 single amino acid variants across diverse protein families, taxa, and functions. Binary fitness labels are derived from DMS scores using top/bottom 30% thresholding, yielding 399,239 functional (label=1) and 297,072 damaging (label=0) variants. We use the official `fold_random_5` cross-validation splits, assigning folds 0–2 to training (417,307 variants), fold 3 to validation (139,524 variants), and fold 4 to test (139,480 variants). AlphaFold2 structures are available for 216 of 217 proteins.

Evaluation Metrics. We report six metrics averaged across the test set: AUROC, AUPRC, Accuracy (ACC), macro-averaged F1 (mF1), macro-averaged Recall (mAR), and macro-averaged Precision (mAP). All experiments are repeated with three random seeds (0, 1, 2) and we report mean ± standard deviation.

Baselines. We compare against two groups of baselines using scores provided by ProteinGym: (i) *Supervised baselines*: OHE (no augmentation), OHE augmented with ESM-1v / MSA Transformer / TranceptEVE embeddings, embedding-based regressors augmented with ESM-1v / MSA Transformer, ProteinNPT (Notin et al., 2023b), and Kermut (Kermut et al., 2024); (ii) *Zero-shot baselines*: ESM-1v (Meier et al., 2021), ESM-2 (650M) (Lin et al., 2023), MSA Transformer (Rao et al., 2021), TranceptEVE (Notin et al., 2022), VESPA (Marquet et al., 2022), GEMME (Laine et al., 2019), ESM-IF1 (Hsu et al., 2022), ProteinMPNN (Dauparas et al., 2022), SaProt (Su et al., 2024), and ESM3 (Hayes et al., 2024).

3.2. Main Results

Table 1 reports the performance of TriFit against supervised baselines on the ProteinGym test set.

Table 1. Comparison with supervised baselines on ProteinGym (217 proteins). Best results in **bold**, second best underlined. Mean ± std over 3 seeds.

Model	AUROC	AUPRC	ACC	mF1	mAR	mAP
OHE (no aug.)	$0.6387_{\pm.0434}$	$0.6990_{\pm.1308}$	$0.5901_{\pm.0324}$	$0.5760_{\pm.0480}$	$0.6034_{\pm.0327}$	$0.5901_{\pm.0324}$
OHE + ESM-1v	$0.7581_{\pm.0959}$	$0.7838_{\pm.1429}$	$0.6692_{\pm.0695}$	$0.6569_{\pm.0818}$	$0.6922_{\pm.0725}$	$0.6691_{\pm.0695}$
OHE + MSA-T	$0.7644_{\pm.0828}$	$0.7968_{\pm.1322}$	$0.6734_{\pm.0625}$	$0.6612_{\pm.0763}$	$0.6967_{\pm.0640}$	$0.6734_{\pm.0625}$
OHE + TranceptEVE	$0.7744_{\pm.0781}$	$0.8051_{\pm.1301}$	$0.6799_{\pm.0589}$	$0.6680_{\pm.0723}$	$0.7054_{\pm.0600}$	$0.6798_{\pm.0589}$
Emb + ESM-1v	$0.8062_{\pm.1009}$	$0.8170_{\pm.1432}$	$0.7007_{\pm.0765}$	$0.6892_{\pm.0888}$	$0.7281_{\pm.0760}$	$0.7006_{\pm.0765}$
Emb + MSA-T	$0.8334_{\pm.0873}$	$0.8478_{\pm.1258}$	$0.7217_{\pm.0735}$	$0.7105_{\pm.0875}$	$0.7506_{\pm.0690}$	$0.7216_{\pm.0735}$
ProteinNPT (Notin et al., 2023b)	$0.8439_{\pm.0829}$	$0.8537_{\pm.1240}$	$0.7303_{\pm.0743}$	$0.7193_{\pm.0888}$	$0.7600_{\pm.0653}$	$0.7303_{\pm.0743}$
Kermut (Kermut et al., 2024)	$\underline{0.8644}_{\pm.0844}$	$\underline{0.8697}_{\pm.1230}$	$\underline{0.7432}_{\pm.0769}$	$\underline{0.7324}_{\pm.0915}$	$\underline{0.7744}_{\pm.0674}$	$\underline{0.7432}_{\pm.0769}$
TriFit (Ours)	$\mathbf{0.8974}_{\pm.0002}$	$\mathbf{0.9088}_{\pm.0002}$	$\mathbf{0.8070}_{\pm.0002}$	$\mathbf{0.8021}_{\pm.0003}$	$\mathbf{0.8008}_{\pm.0005}$	$\mathbf{0.8039}_{\pm.0000}$

TriFit achieves state-of-the-art performance across all six metrics, improving AUROC by $+3.3$ points over the best supervised baseline (Kermut) and $+5.4$ points over ProteinNPT. Notably, the standard deviation of TriFit across seeds is an order of magnitude smaller than that of competing methods (e.g., 0.0002 vs. 0.0844 for Kermut in AUROC), indicating substantially more stable predictions. For reference, the best zero-shot model, ESM3 (Hayes et al., 2024), achieves AUROC 0.769, confirming that supervised multimodal fusion provides significant advantages over zero-shot sequence-only approaches.

3.3. Ablation Study

Table 2 presents a systematic ablation over modality combinations and architectural components. All configurations share the same training procedure and hyperparameters.

Table 2. Ablation study on ProteinGym test set (mean $\pm$ std, 3 seeds).

Configuration	AUROC	AUPRC	ACC	mF1	mAR	mAP
Single modality						
Seq only	$0.8743_{\pm.0006}$	$0.9068_{\pm.0006}$	$0.8039_{\pm.0004}$	$0.7995_{\pm.0002}$	$0.7988_{\pm.0003}$	$0.8003_{\pm.0006}$
Struct only	$0.8376_{\pm.0003}$	$0.8638_{\pm.0006}$	$0.7633_{\pm.0006}$	$0.7561_{\pm.0008}$	$0.7541_{\pm.0010}$	$0.7597_{\pm.0005}$
Dyn only	$0.7838_{\pm.0006}$	$0.8301_{\pm.0004}$	$0.7116_{\pm.0006}$	$0.7026_{\pm.0011}$	$0.7011_{\pm.0013}$	$0.7059_{\pm.0007}$
Pairwise modality						
Seq + Struct	$0.8759_{\pm.0001}$	$0.9083_{\pm.0002}$	$0.8071_{\pm.0008}$	$0.8024_{\pm.0007}$	$0.8013_{\pm.0012}$	$0.8040_{\pm.0014}$
Seq + Dyn	$0.8660_{\pm.0001}$	$0.9079_{\pm.0001}$	$0.8070_{\pm.0006}$	$0.8017_{\pm.0004}$	$0.7998_{\pm.0009}$	$0.8046_{\pm.0013}$
Struct + Dyn	$0.8461_{\pm.0001}$	$0.8769_{\pm.0002}$	$0.7672_{\pm.0004}$	$0.7611_{\pm.0002}$	$0.7597_{\pm.0008}$	$0.7631_{\pm.0009}$
Architectural ablation (full trimodal)						
w/o MoE (simple concat)	$0.8870_{\pm.0002}$	$0.9092_{\pm.0003}$	$0.8070_{\pm.0001}$	$0.8021_{\pm.0002}$	$0.8008_{\pm.0006}$	$0.8040_{\pm.0004}$
w/o Contrastive loss	$0.8876_{\pm.0005}$	$0.9101_{\pm.0005}$	$0.8068_{\pm.0003}$	$0.8019_{\pm.0012}$	$0.8006_{\pm.0023}$	$0.8039_{\pm.0005}$
TriFit (full)	$\mathbf{0.8974}_{\pm.0002}$	$\mathbf{0.9098}_{\pm.0002}$	$\mathbf{0.8078}_{\pm.0002}$	$\mathbf{0.8018}_{\pm.0003}$	$\mathbf{0.7991}_{\pm.0005}$	$\mathbf{0.8063}_{\pm.0000}$

Several observations emerge from the ablation. First, each modality contributes positively: removing dynamics from the full model (Seq+Struct: 0.876) yields a -2.2 point AUROC drop relative to TriFit, confirming the added value of dynamics embeddings beyond sequence and structure alone. Second, dynamics exhibits the largest marginal contribution when combined with the other two modalities—adding dynamics to Seq+Struct raises AUROC from 0.876 to 0.897, whereas adding structure to Seq raises it by only $+0.2$ points. This suggests that dynamics captures complementary information not redundant with structure. Third, both the MoE routing and contrastive loss contribute to performance, with the contrastive loss providing $+0.010$ AUROC and MoE providing a further $+0.001$ improvement.

3.4. Analysis

Representation Analysis. Figure 2 illustrates two complementary views of the learned representations. UMAP projections of the three modality embeddings occupy largely disjoint regions, confirming non-redundant complementarity. LDA scores applied to the MoE fused representations reveal a clear distributional shift between damaging and functional variants, demonstrating fitness-discriminative representation learning.

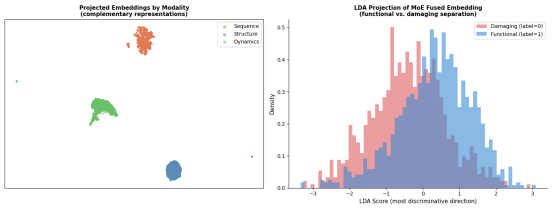

Figure 2. Representation analysis. **Left:** UMAP of projected modality embeddings (sequence: orange, structure: blue, dynamics: green). **Right:** LDA projection of MoE fused embeddings showing distributional shift between damaging (red) and functional (blue) variants.

Expert Utilization & Calibration. MoE router analysis across 217 proteins reveals two dominant clusters: one preferring the Trimodal expert (E4) and one preferring Struct+Dyn (E3), while Seq+Dyn (E2) is consistently underweighted—confirming that dynamics embeddings are actively leveraged (see Appendix C). TriFit further achieves well-calibrated probabilistic outputs (ECE $= 0.044$) without post-hoc correction (see Appendix D).

4. Conclusion

We presented **TriFit**, a supervised multimodal framework for protein mutation fitness prediction that integrates sequence, structure, and protein dynamics through a Mixture-of-Experts fusion module with cross-modal contrastive learning. To our knowledge, TriFit is the first method to systematically incorporate GNM-based dynamics embeddings as a third modality for variant effect prediction, addressing a longstanding gap between biophysical theory and machine learning practice.

On the ProteinGym substitution benchmark comprising 217 DMS assays and 696k single amino acid variants, TriFit achieves AUROC 0.897 ± 0.0002, surpassing all supervised baselines including Kermut (0.864) and ProteinNPT (0.844), as well as the best zero-shot model ESM3 (0.769). Ablation experiments confirm that each modality contributes independently, with dynamics providing the largest marginal gain when combined with the other two modalities ($+2.2$ AUROC over Seq+Struct). The MoE router exhibits meaningful protein-specific expert selection patterns, preferentially activating Struct+Dyn and Trimodal experts, providing direct evidence that dynamics embeddings are actively leveraged rather than suppressed. Additionally, TriFit demonstrates excellent calibration (ECE $= 0.044$) without any post-hoc correction, an important property for clinical variant interpretation.

References

Bahar, I., Atilgan, A. R., and Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-

parameter harmonic potential. *Folding and Design*, 2: 173–181, 1997.

Bakan, A., Meireles, L. M., and Bahar, I. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics*, 27:1575–1577, 2011.

Dauparas, J., Anishchenko, I., Bennett, N., et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378:49–56, 2022.

Fowler, D. M. and Fields, S. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11:801–807, 2014.

Haliloglu, T., Bahar, I., and Erman, B. Gaussian dynamics of folded proteins. *Physical Review Letters*, 79:3090, 1997.

Hayes, T., Rao, R., Akin, H., et al. Simulating 500 million years of evolution with a language model. *Science*, 2024.

Hsu, C., Verkuil, R., Liu, J., et al. Learning inverse folding from millions of predicted structures. In *ICML*, 2022.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L., and Dror, R. Learning from protein structure with geometric vector perceptrons. In *ICLR*, 2021.

Jumper, J., Evans, R., Pritzel, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596: 583–589, 2021.

Kermut, V. et al. Modelling mutational effects on biochemical phenotypes using gaussian processes: Application to clinical variant interpretation. *bioRxiv*, 2024.

Laine, E., Karami, Y., and Carbone, A. Gremlin and GEMME: Fast and accurate protein fitness landscape prediction. *PLOS Computational Biology*, 2019.

Lin, Z., Akin, H., Rao, R., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *ICLR*, 2019.

Marquet, C., Heinzinger, M., Olenyi, T., et al. VESPA: Variant effect score prediction without alignments. *PLOS Computational Biology*, 2022.

Meier, J., Rao, R., Verkuil, R., et al. Language models enable zero-shot prediction of the effects of mutations on protein function. In *NeurIPS*, 2021.

Notin, P., Van Niekerk, L., Kollasch, A., et al. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. In *NeurIPS Workshop on Learning Meaningful Representations of Life*, 2022.

Notin, P., Kollasch, A., Ritter, D., et al. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *NeurIPS*, 2023a.

Notin, P., Weitzman, R., Marks, D., and Gal, Y. ProteinNPT: Improving protein property prediction and design with non-parametric transformers. In *NeurIPS*, 2023b.

Rao, R., Liu, J., Verkuil, R., et al. MSA transformer. 2021.

Su, J., Han, C., Zhou, Y., et al. SaProt: Protein language modeling with structure-aware vocabulary. In *ICLR*, 2024.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. In *NeurIPS*, 2018.

A. Implementation Details

Embedding extraction. All three modality embeddings are pre-computed and stored before training, allowing the fusion module to be trained without repeated encoder forward passes. Sequence embeddings are extracted using ESM-2 (650M) with a single masked forward pass per unique mutation position per protein. Structure embeddings are derived from AlphaFold2-predicted C_α coordinates provided by ProteinGym, covering 216 of 217 proteins (99.3%). Dynamics embeddings are computed via GNM using ProDy, applied to the same AlphaFold2 structures.

Model architecture. Projection heads map sequence (1280-dim), structure (512-dim), and dynamics (256-dim) embeddings to a shared d=512 space via Linear $\rightarrow$ LayerNorm $\rightarrow$ GELU. Each expert is a two-layer MLP with GELU activation. The router is Linear(1536, 64) $\rightarrow$ GELU $\rightarrow$ Linear(64, 4) $\rightarrow$ Softmax. The classifier is Linear(512, 256) $\rightarrow$ GELU $\rightarrow$ Dropout(0.1) $\rightarrow$ Linear(256, 2). Total trainable parameters: **3.6M**. All experiments were conducted on a single NVIDIA A100 (40GB) GPU.

B. Zero-Shot Baseline Comparison

Table 3 reports TriFit performance alongside zero-shot baselines. We note that this comparison is not strictly equivalent—TriFit is a supervised model trained on held-out protein splits, while zero-shot models require no training data. We include this comparison to contextualize TriFit's performance relative to the broader landscape of variant effect predictors. The gap between TriFit (AUROC 0.897) and the best zero-shot model ESM3 (AUROC 0.769) suggests that supervised multimodal fusion provides substantial advantages, though at the cost of requiring labeled training data.

Table 3. Zero-shot baseline comparison on ProteinGym (AUROC only; max(AUROC, 1-AUROC) reported for zero-shot models to account for score direction). TriFit is supervised.

Model	Type	AUROC
Site Independent	Zero-shot	0.6967
ProteinMPNN	Zero-shot	0.6532
EVmutation	Zero-shot	0.7165
ESM1b	Zero-shot	0.7212
ESM1v	Zero-shot	0.7166
ESM1v (ensemble)	Zero-shot	0.7312
ESM2-650M	Zero-shot	0.7443
MSA Transformer (ensemble)	Zero-shot	0.7430
Tranception L	Zero-shot	0.7414
TranceptEVE L	Zero-shot	0.7549
VESPA	Zero-shot	0.7564
GEMME	Zero-shot	0.7565
ESM-IF1	Zero-shot	0.7442
SaProt-650M	Zero-shot	0.7593
ESM3	Zero-shot	0.7692
TriFit (Ours)	**Supervised**	**0.8974**

C. MoE Expert Utilization

Figure 3 shows the mean router weight assigned to each of the four experts across all 217 test proteins, clustered hierarchically by weight pattern. Two dominant protein clusters are visible: a cluster where the Trimodal expert (E4) receives the highest weight (dark red column), and a cluster where the Struct+Dyn expert (E3) is preferred. The Seq+Dyn expert (E2) is consistently underweighted across virtually all proteins, suggesting that dynamics information is most useful when combined with structural context (E3) rather than sequence context alone (E2). The Seq+Struct expert (E1) shows intermediate, relatively uniform utilization.

This pattern is consistent with the ablation results in Table 2 of the main paper: Seq+Dyn (AUROC 0.866) underperforms Seq+Struct (AUROC 0.876), suggesting that raw sequence representations already encode much of the information captured by dynamics when paired together, whereas structure and dynamics provide more complementary signals. The emergence

of two distinct protein clusters—one relying heavily on trimodal fusion and one on structure+dynamics—suggests that different protein families may have systematically different optimal modality combinations, validating the design choice of an adaptive router over a fixed fusion scheme.

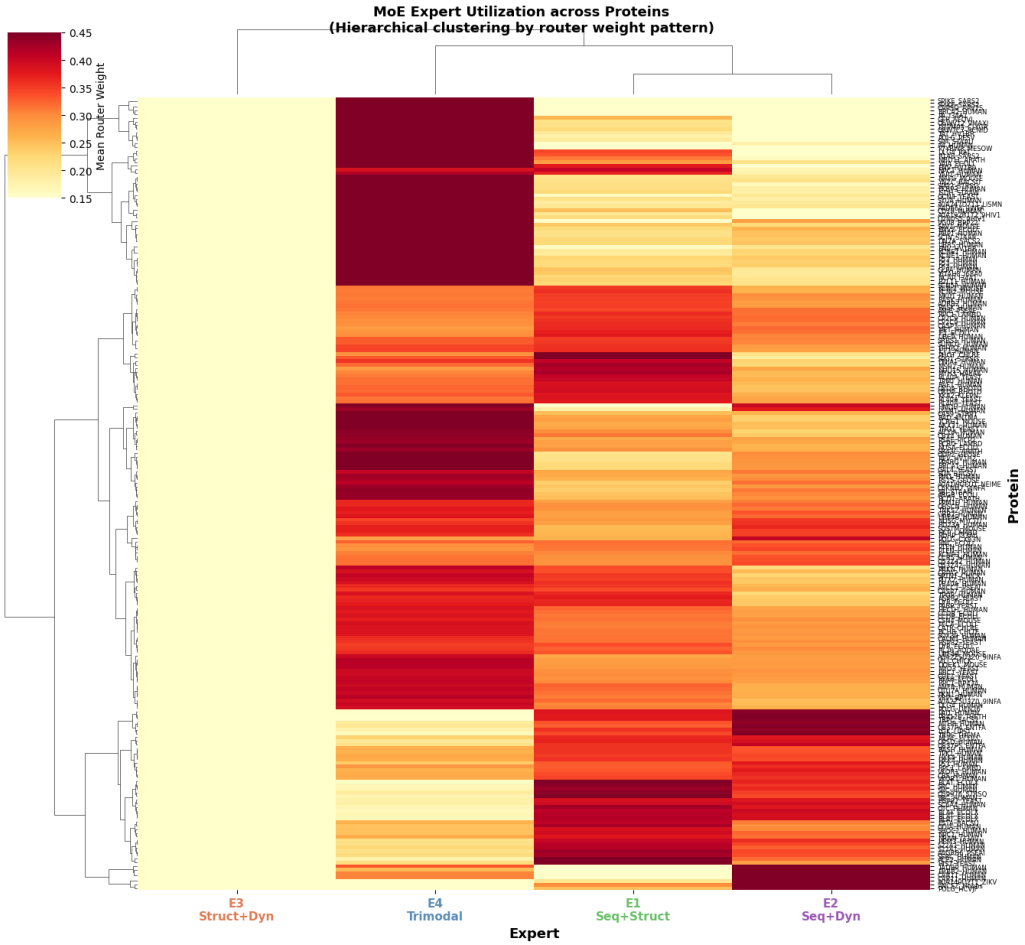

Figure 3. MoE expert utilization across 217 test proteins (hierarchical clustering by router weight pattern). Color intensity indicates mean router weight assigned to each expert. Two major clusters emerge: proteins preferring the Trimodal expert (E4) and proteins preferring the Struct+Dyn expert (E3). The Seq+Dyn expert (E2) is consistently underweighted, while Seq+Struct (E1) shows intermediate utilization.

D. Calibration Analysis

Figure 4 presents a detailed calibration analysis of TriFit on the held-out test set. Calibration is particularly important for variant effect prediction in clinical settings, where predicted probabilities are used to prioritize variants for experimental follow-up or clinical interpretation.

The reliability diagram (left) shows that TriFit's predicted probabilities closely track empirical positive rates across all 15 quantile bins. The Expected Calibration Error (ECE $= 0.044$) is achieved without any post-hoc calibration procedure such as temperature scaling or Platt scaling. This suggests that the cross-entropy training objective combined with dropout regularization is sufficient to produce well-calibrated outputs.

The confidence distribution (center) reveals a bimodal pattern: the majority of predictions are made with very high confidence (> 0.9), with damaging and functional variants showing similar confidence distributions. This indicates that the model is not artificially inflating confidence for one class. The confidence-accuracy plot (right) confirms a near-diagonal relationship across all confidence bins.

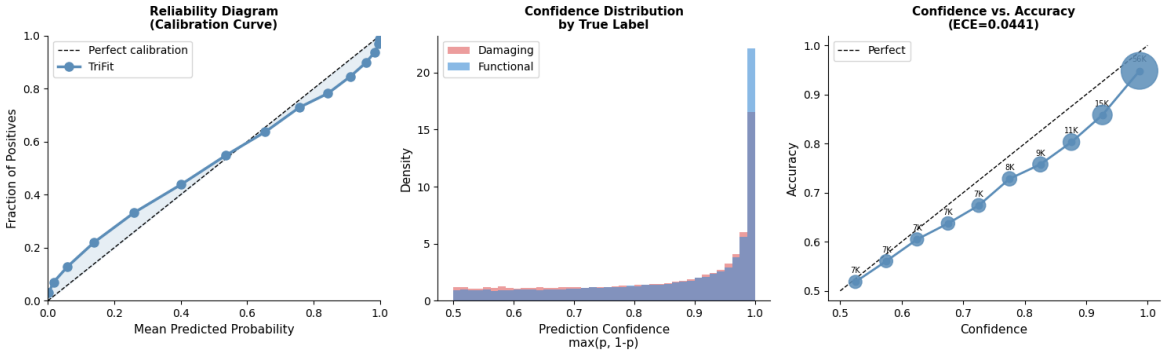

Figure 4. Prediction calibration analysis on the ProteinGym test set (139,480 variants). **Left:** Reliability diagram showing close alignment between predicted probabilities and empirical positive rates (ECE $= 0.044$), achieved without post-hoc calibration. **Center:** Confidence distribution $\max(p, 1-p)$ by true label, showing similar confidence profiles for both classes. **Right:** Confidence vs. accuracy across 10 equal-width bins; dot size proportional to bin count.

E. Per-position Prediction Accuracy

Figure 5 shows per-position prediction accuracy along the protein sequence for three representative proteins with high variant coverage: HMDH_HUMAN (HMG-CoA reductase, n=3,409, ACC $= 0.793$), POLG_DEN26 (Dengue polyprotein, n=3,357, ACC $= 0.829$), and MSH2_HUMAN (MutS homolog 2, n=3,313, ACC $= 0.910$). Each point represents a single residue position, with dot size proportional to the number of variants at that position and color indicating the local functional rate.

Several patterns emerge. First, accuracy is generally high (> 0.75) across most positions, with localized drops at positions where functional rate is intermediate (neither clearly damaging nor clearly functional), reflecting the inherent ambiguity of borderline variants. Second, MSH2_HUMAN achieves notably higher accuracy (0.910) than HMDH_HUMAN (0.793), consistent with MSH2's role as a DNA mismatch repair protein where fitness effects tend to be more binary. Third, accuracy does not show a systematic bias toward N- or C-terminal positions, suggesting that the model does not rely on positional artifacts.

The moving average curves reveal smooth regional trends, suggesting that local structural or functional context captured by TriFit's multimodal embeddings influences prediction quality at a domain level rather than individual residue level.

F. Limitations

Fixed random projections for structure and dynamics. The structure and dynamics embeddings use fixed (non-learned) random projection matrices rather than trained geometric encoders. While this is computationally efficient and avoids overfitting on the relatively small number of proteins (217), it does not exploit the full expressive power of graph neural networks or equivariant architectures (Jing et al., 2021). Replacing these with learned encoders such as GVP-GNN may improve performance, particularly for structure embeddings.

Static structure assumption. Our dynamics embeddings are computed from static AlphaFold2-predicted structures using the GNM, which models only harmonic fluctuations around the equilibrium conformation. More sophisticated dynamics representations—such as those derived from molecular dynamics simulations or normal mode analysis with anharmonic corrections—may better capture biologically relevant conformational changes.

Binarization threshold sensitivity. TriFit's labels are derived from DMS scores using a fixed top/bottom 30% threshold, discarding the middle 40% of variants as ambiguous. Performance may vary with different thresholding strategies, and the model cannot directly make predictions on continuous fitness scores without retraining.

Single amino acid substitutions only. TriFit is trained and evaluated exclusively on single amino acid substitutions. Extension to multi-site variants, insertions, and deletions—which constitute a significant fraction of disease-causing mutations—requires architectural modifications and additional training data.

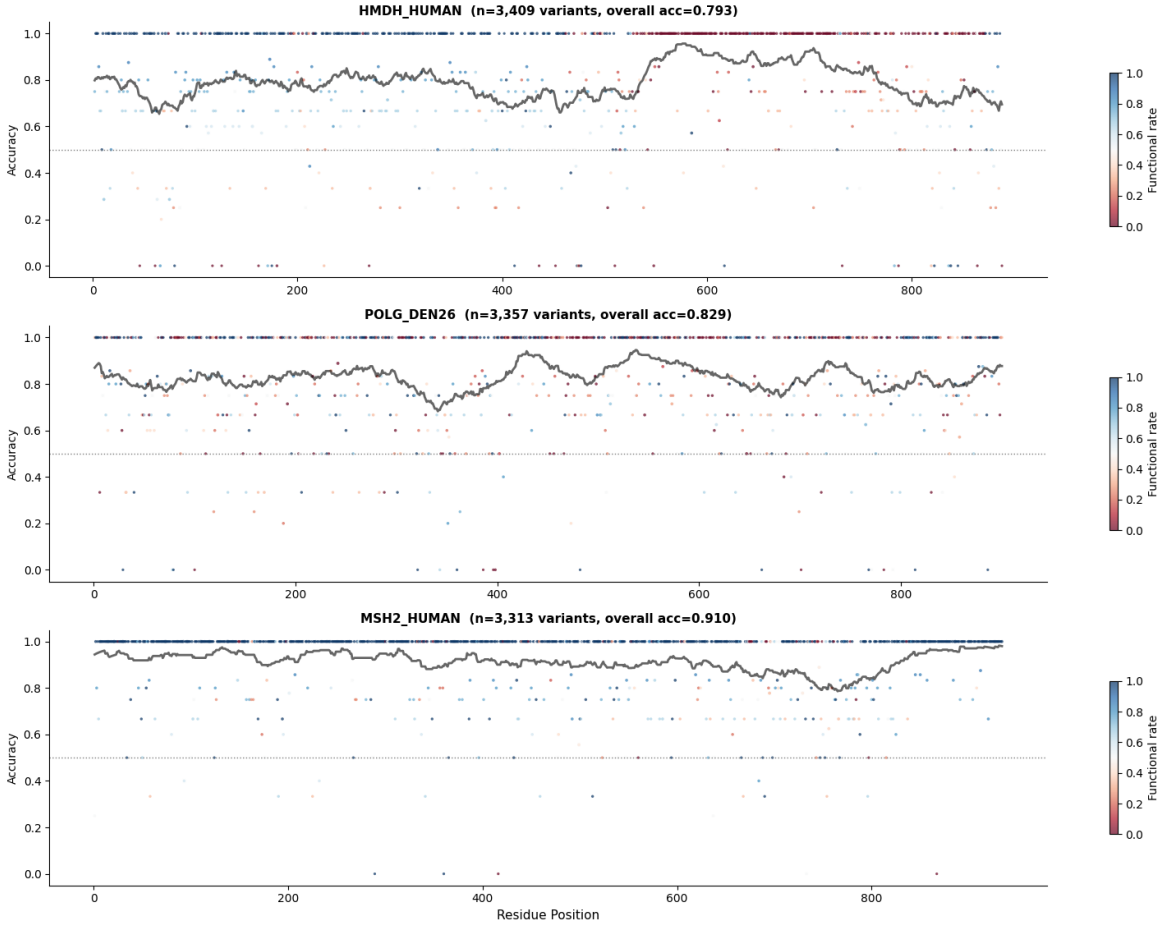

Figure 5. Per-position prediction accuracy along the protein sequence for three representative proteins. Each dot corresponds to one residue position; dot size is proportional to variant count at that position; color indicates local functional rate (blue = functional, red = damaging). The black curve shows a sliding window average (window = $L/20$ residues). Overall per-protein accuracy is reported in the subtitle of each panel.