ONE STRUCTURE IS NOT ENOUGH: CONFORMATIONAL ENSEMBLE REPRESENTATIONS IMPROVE IN-FRAME INDEL EFFECT PREDICTION ACROSS PROTEIN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting the functional effect of in-frame insertions and deletions (indels) is essential in genome interpretation for clinical diagnosis, genomic research and drug development. Structural information is one of the main sources for variant effect prediction, as variant effects are often mediated by protein structure alteration. A predicted static wild-type structure has enabled structure-informed protein language models (PLMs) to achieve high accuracy in genomic variant effect prediction, especially for missense variants. However, current computational tools show relatively lower prediction performance for in-frame indels, which are also difficult for experts to interpret manually. This limitation may partly stem from relying on a single static structure, since (i) AlphaFold's reliability for structure-disrupting variants remains uncertain, and (ii) a static structure cannot fully capture the protein dynamics linked to indel tolerance, such as local flexibility. Here, we systematically evaluate conformational ensembles generated from indel variant sequences as structural input to six structure-informed PLMs used as frozen feature extractors. Across 4,270 in-frame indels from 1,418 proteins, the PLMs' pathogenicity prediction improved significantly with the ensembles over single static structure inputs in most comparisons, especially for ESM3, whose AUROC rose from 0.8794 to 0.9359. Most PLMs also produced more variant-specific representations, and the improvement was concentrated in regions with low pLDDT, unstructured regions, and long benign variants. Consistent with recent findings on the relationship between intrinsic structural dynamics and protein tolerance to indels, these results suggest a way to incorporate such biological knowledge into PLMs at the representation level.

1 INTRODUCTION

Machine learning has become a core technology in biological research, and tools built on it are now used as evidence in clinical variant classification (Pejaver et al., 2022; Bergquist et al., 2025). For example, AlphaFold (Jumper et al., 2021) predicts protein structures with accuracy competitive with experimental structures, and AlphaMissense (Cheng et al., 2023), an adaptation of AlphaFold, predicts the pathogenicity of all possible human single amino acid substitutions. Unlike single amino acid substitutions, in-frame insertions and deletions (indels) are genetic variants that alter the length of encoded polypeptide chains while preserving the reading frame. Predicting the impact of these indels on protein structure and function is crucial for genomic interpretation in clinical diagnostics, precision medicine, and drug development. However, the functional consequences of in-frame indels span a broad spectrum, ranging from complete loss of function to benign phenotypes. Because they can induce diverse and unpredictable structural alterations compared to wild-type proteins, manual interpretation of their effects is challenging. Despite this need, indels remain notably underrepresented in variant databases like ClinVar and Deep Mutational Scanning (DMS) datasets (Landrum et al., 2025; Fowler & Fields, 2014). Elucidating these consequences traditionally relies on experimental approaches, such as DMS, or computational variant effect predictors (VEPs). However, experimental approaches are highly time consuming and resource intensive. Furthermore, many existing computational tools either do not provide prediction scores for in-frame indels or exhibit

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

lower performance in differentiating pathogenic from benign variants compared to other variant types (Abderrazzaq et al., 2026; Cannon et al., 2023).

Because a protein's function can be dictated by the three-dimensional arrangement of its residues, and the effect of a variant depends on how it alters the non-covalent interactions that stabilize the folded state or maintain the geometry of functional sites, structural information is one of the crucial factors for interpreting variant consequences (Stein et al., 2019). Recently, variant effect predictors based on structure-informed protein language models (PLMs) have demonstrated high performance in interpreting missense variants by leveraging predicted single static wild-type (WT) structures (Su et al., 2024). Building upon the architectural foundation of Transformer-based large language models (LLMs), PLMs are pre-trained on massive corpora of natural protein sequences to decipher the fundamental grammar and evolutionary language of life (Vaswani et al., 2017; Rives et al., 2021). To capture the semantics of this language, recent structure-informed PLMs also tokenize the 3D coordinates of proteins, for instance by vector-quantizing the spatial neighborhood of each position into a learned structural alphabet, and integrate the resulting tokens as contextual features (Van Kempen et al., 2024). Alongside pathogenicity prediction, these models are highly effective in predicting DMS assay outcomes. Structure-informed PLMs occupy the majority of the top-tier positions in the ProteinGym DMS substitution benchmark (Notin et al., 2023). Similarly, emerging studies highlight the importance of structural context in evaluating in-frame indels (Goldman et al., 2025).

However, structural information remains difficult to incorporate into indel interpretation, which limits further improvement in the accuracy of predicting whether indels are pathogenic or benign. Reflecting this challenge, studies are hardly found that utilize structure-informed PLMs to interpret in-frame indels, to our knowledge. This stems from two technical challenges in incorporating structural information for in-frame indel variants.

First, because in-frame indels disrupt the one-to-one mapping between the mutant sequence and the WT structure, the static WT structure cannot be used as a direct proxy. While missense variants allow direct mapping due to unchanged sequence lengths, indels introduce length changes that potentially cause significant local structural deviations. Second, predicting mutant structures directly from variant sequences has inherent limitations. There are conflicting reports on whether static structure prediction tools accurately capture mutation-induced conformational changes of missense variants (Buel & Walters, 2022; Pak et al., 2023). For in-frame indels, this challenge is expected to be even more pronounced due to the more extensive backbone rearrangements they induce.

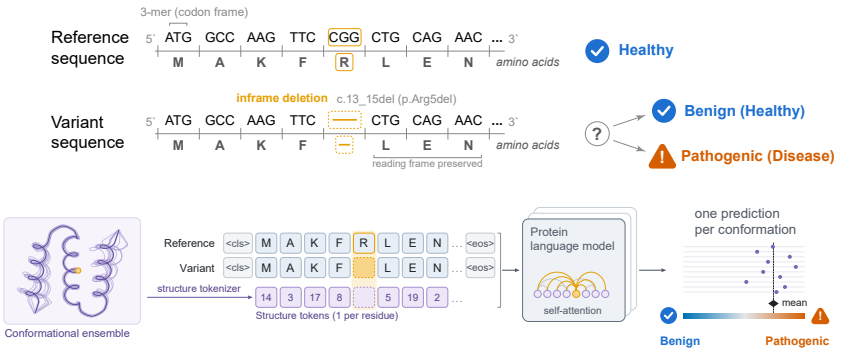

Figure 1: Overview of in-frame indel pathogenicity prediction with conformational ensembles. (Top) An in-frame deletion, c.13_15del (p.Arg5del), removes one codon while preserving the reading frame, and the task is to classify such a variant as benign or pathogenic. (Bottom) Each conformation generated for the indel sequence is converted into structure tokens, one per residue, by the structure tokenizer of the PLM. The structure tokens are provided to a frozen structure-informed PLM together with the amino acid tokens of the indel sequence, and the resulting embedding is pooled with the embedding of the WT sequence and passed to a classifier head. A prediction is made for each conformation, and the predictions are ensemble-averaged to obtain the probability of pathogenicity.

Recent studies emphasize that protein dynamics, the inherent process by which a flexible protein naturally fluctuates and samples an ensemble of 3D conformations over time, play a critical role

in determining the consequences of in-frame indels (Henzler-Wildman & Kern, 2007; Banerjee & Bahar, 2023). Benign and fold-retaining deletions predominantly map to highly flexible regions characterized by high residue fluctuation and low mechanical stiffness; thus, dynamics-based features could be strong predictors of post-mutation structural stability. Indels can also stabilize active conformations by altering local flexibility, as revealed by dynamics-based analyses. For example, the EGFR exon 19 deletion E746 - A750del shortens the β3-αC loop and restrains the alpha helix in its active position, promoting kinase activity (Tamirat et al., 2019; Foster et al., 2016). While molecular dynamics (MD) simulations traditionally provide the conformational ensembles required to analyze these dynamic changes, their high computational cost limits large-scale variant screening. To address this, deep learning generative models have recently emerged to efficiently approximate MD equilibrium distributions (Noé et al., 2019; Zheng et al., 2024; Jing et al., 2024). These models can rapidly generate structural ensembles and accurately predict the thermodynamic stability of mutant proteins, offering a highly scalable alternative to traditional simulations (Lewis et al., 2025). In representational terms, tokenizing a single static structure treats structure as a deterministic function of sequence and collapses an inherently one-to-many mapping onto a point estimate. An ensemble instead induces a per-residue distribution over structure tokens, whose entropy encodes local flexibility and thereby enriches the information a single sequence can carry.

In this study, we used conformational ensembles generated for in-frame indel sequences as the structural input of structure-informed PLMs and examined whether they improve indel pathogenicity prediction (Figure 1). We applied this approach to six structure-informed PLMs that differ in architecture and structure encoding, namely ESM3 (Hayes et al., 2025), SaProt (Su et al., 2024), ProSST (Li et al., 2024a), VenusREM (Tan et al., 2025), S2F (Zhang et al., 2024), and AIDO.RAG (Li et al., 2024b). All PLMs were used as frozen feature extractors with an identical classifier head. Using 4,270 in-frame indels across 1,418 proteins and a sequence similarity-based split, we compared the 10-conformation ensemble with each of the static WT structure, the static indel structure, and a single sampled conformation. The use of the 10-conformation ensemble significantly outperformed the other three inputs in 15 of 18 comparisons, and in ESM3 in particular, this approach increased the median AUROC from 0.8794 with the static indel structure to 0.9359. Utilizing the ensemble also yielded more variant-specific representations and more consistent predictions across PLMs. The variants whose predictions were improved by the ensemble across PLMs were mostly located in regions with low pLDDT and few residue contacts in the static WT structure, and the deletions among them were longer than the remaining deletions. By utilizing the ensemble, all six PLMs achieved higher AUROC than CADD and INDELi-E on the variants scored by each tool. We expect that this approach can serve as a practical way to use structural information for in-frame indels in structure-informed PLMs without modifying the PLMs themselves.

2 METHODS

2.1 MODEL ARCHITECTURE

All PLMs were employed as frozen feature extractors. Regarding structural inputs, conformation ensembles were exclusively utilized for in-frame indels; for missense variants and wild-types (WT), the WT structures generated by ColabFold were used, whereas structures for all other variant types were replaced with padding tokens (Mirdita et al., 2022). Embeddings for both the WT and variant sequences were utilized together to construct a pooled feature. Specifically, the pooled feature was defined as follows.

$$\text{Pooled Feature} = [h_{var}(s) \parallel \text{mean}_{|i-s|\leq 7} h_{var}(i) \parallel (h_{var}(s) - h_{wt}(s)) \parallel \ell]$$

where s denotes the variant position, h represents the hidden state of dimension D, and $\parallel$ denotes concatenation. The length feature ℓ was formulated as $[L_{var}/1000, (L_{var} - L_{wt})/100]$, resulting in a total feature width of $3D + 2$. True pre-padding lengths were strictly used for these length features to preserve the indel signal, which would otherwise be lost if derived directly from padded tensor shapes. To measure how well the PLM naturally organizes the protein representations, a simple, fixed probe was used. The classifier head was designed sequentially with a linear layer (mapping $3D + 2$ to 256 dimensions), a GELU activation function, a Dropout layer (rate = 0.3), and a final linear layer (mapping 256 to 1 dimension).

2.2 Training Procedure

Model training was performed using the BCEWithLogitsLoss function and the AdamW optimizer with a weight decay of 1e-2. The training process was executed for 60 epochs without early stopping, utilizing a batch size of 1024, a learning rate of 1e-4, and random seeds set to 0, 1, 2. When conformation ensembles were applied, a single conformer was randomly sampled for each training iteration, while during inference, predictions were computed for every conformer in the ensemble and subsequently averaged. After all epochs were completed, the model achieving the highest AUROC on the validation set was selected as the final model.

2.3 Multiple Sequence Alignment (MSA) Generation

To generate Multiple Sequence Alignment (MSA) inputs for ColabFold, BioEmu, and specific MSA-based protein language models (PLMs), colabfold_search was employed. The searches were executed locally against the UniRef30 (release 2302) and ColabFoldDB environmental (release 202108) databases (Mirdita et al., 2017). Alignments were constructed independently for each variant sequence. Specifically, insertion/deletion (indel) variants were aligned as their respective indel sequences, and missense variants were aligned as their respective missense sequences.

2.4 Structure Generation

All wild-type (WT) and variant protein structures were generated using ColabFold (v1.6.2) with AlphaFold2-ptm. The structure predictions were executed using a single model with the following defined parameters: three recycles with an early stopping threshold of 0.5 Å, an MSA depth capped at 512:1024, a single random seed, without the use of templates, and without Amber relaxation.

Conformer ensembles for both WT and in-frame indel variants were generated using BioEmu (v1.4.1, checkpoint bioemu-v1.1). For each sequence, exactly 10 conformer samples (n_samples = 10) were produced using the DPM denoiser (Lu et al., 2022). The sampling process utilized the previously generated MSAs and followed a deterministic ODE integration path (noise = 0) starting from an initially seeded_latent space.

2.5 Pathogenicity Dataset

A pathogenicity dataset was constructed by integrating three datasets. First, variants from the ClinVar database (downloaded on August 6, 2026) with a review status of at least one star were retrieved; those classified as "benign" or "likely benign" were defined as benign variants, whereas those classified as "pathogenic" or "likely pathogenic" were defined as pathogenic variants. Variants mapped to pathogenic, likely pathogenic, benign, or likely benign classifications based on the ACMG guidelines using Evidence were incorporated into the dataset for data augmentation (Richards et al., 2015; Seo et al., 2020). Lastly, to augment likely benign variants from the general human population, variants from gnomAD v4.1 exhibiting an allele frequency >0.001 and capable of being expressed in the HGVS protein (HGVSp) format were included (Chen et al., 2024). Only the variants and their corresponding labels derived from ClinVar were utilized as the validation and test set to evaluate model performance, whereas all other datasets were used exclusively for model training.

2.6 Dataset split

The dataset was partitioned using a similarity-based split. Sequences were clustered using MMseqs2 at a threshold of 30% sequence identity and 80% coverage (Steinegger & Söding, 2017). To prevent information leakage, identical sequences and short isoforms contained within longer sequences were subsequently merged into the same clusters. Finally, the resulting clusters were randomly partitioned into five distinct folds. During training, when the ith fold was assigned as the test set, the $(i-1)^{\text{th}}$ fold was used as the validation set.

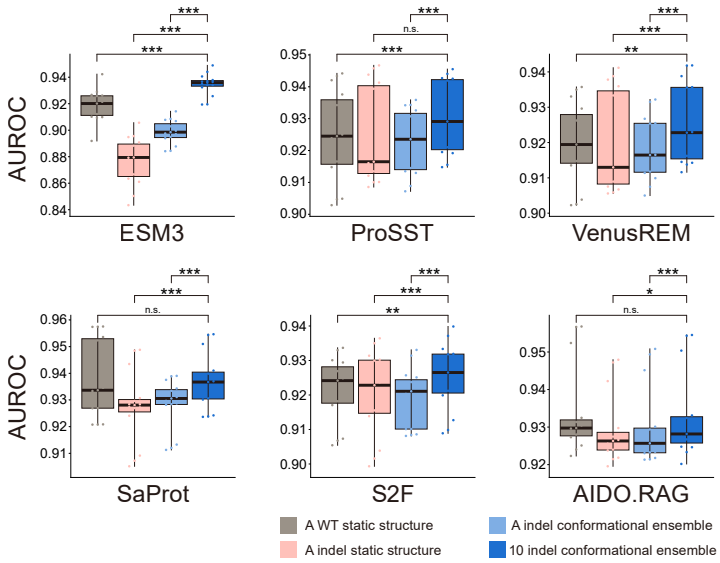

Figure 2: Indel pathogenicity prediction with four structure inputs across six structure-informed PLMs. AUROC on the test folds is shown for the static WT structure (gray), the static indel structure (pink), a single sampled conformation of the indel sequence (light blue), and the 10-conformation ensemble (dark blue). Each point represents one of 15 runs (five folds × three random seeds) under the similarity-based split. Boxes show the median and interquartile range, and whiskers show the full range. Statistical significance was calculated using a paired Wilcoxon signed-rank test. Asterisks indicate p <0.05 (*), p <0.01 (**), and p <0.001 (***), and n.s. indicates not significant.

3 Results

3.1 Evaluating Conformational Ensembles in Structure-Informed PLMs for Indel Pathogenicity Prediction

Conformational ensembles enhance pathogenicity prediction performance for indel variants. Among the six structure-informed PLMs, five exhibited a higher median AUROC when using the 10-conformation ensemble (blue) compared to the static indel structures (pink), with five showing statistically significant improvements (Figure 2). To test whether variant structure tokens derived from conformational ensembles improve in-frame indel pathogenicity prediction over those derived from a static structure, we compared four types of structure tokens across six structure-informed PLMs. The four structural tokens differed only in the structure provided for each indel variant, and all other training conditions were identical. Specifically, the four inputs were (i) the predicted static structure of the wild-type (WT) protein, (ii) the predicted static structure of the indel variant, (iii) a single conformation sampled for the indel sequence, and (iv) an ensemble of 10 conformations sampled for the same indel sequence (hereafter, the 10-conformation ensemble). Across the 15 matched pairs under the similarity split, the 10-conformation ensemble significantly outperformed the other three inputs in 15 of the 18 comparisons (six pLMs × three inputs; paired Wilcoxon signed-rank test, p <0.05). Compared with the static indel structure, the 10-conformation ensemble yielded a significant performance gain in five of the six pLMs: ESM3, SaProt-650M, and S2F ($p = 1.4 \times 10^{-4}$ each), VenusREM ($p = 5.3 \times 10^{-4}$), and AIDO.RAG-16B ($p = 3.0 \times 10^{-2}$). This gain was most pronounced in ESM3, with a median AUROC of 0.9359 for the 10-conformation ensemble versus 0.8794 for the static indel structure. Similarly, compared to the static WT structure (gray), the 10-conformation ensemble achieved a higher median AUROC in five models, four of which demonstrated significant improvements. Across all PLMs, there were no cases where the static structure significantly outperformed the 10-conformation ensemble, nor did the static indel structure perform significantly better than the static WT structure.

Additionally, performance on missense variants was unaffected by which structure was provided for the indel variants (Appendix A). Across all six PLMs, the missense AUROC distributions were nearly identical among the static WT structure, the static indel structure, the single sampled conformation, and the 10-conformation ensemble. We also evaluated conditions in which no indel-specific structure prediction was available (Appendix B). In these conditions, each indel variant received one of four WT-derived structure inputs: (i) the unmodified WT static structure with end padding, (ii) a position-matched WT structure, in which WT structure tokens were projected onto the indel sequence by alignment, (iii) a shuffled WT structure, in which the order of WT structure tokens was randomly permuted, and (iv) a masked WT structure, in which all structure tokens were masked so that no structural information was provided. Removing structural information by shuffling or masking generally reduced performance. The shuffled WT structure yielded a lower median AUROC than the unmodified WT structure in all six PLMs, and the masked WT structure did so in four PLMs (ESM3, ProSST, VenusREM, and SaProt). The position-matched WT structure performed close to the unmodified WT structure, and the unmodified WT structure showed a slight advantage in most models.

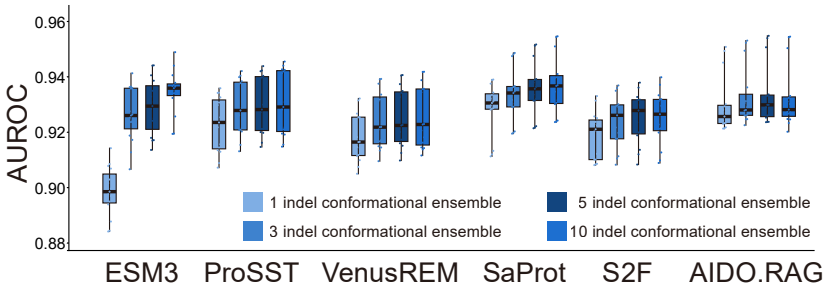

Figure 3: Effect of ensemble size on indel pathogenicity prediction. AUROC on the test folds is shown for ensembles of k = 1, 3, 5, and 10 conformations per indel sequence, with the same k used for training and inference. Each point represents one of 15 runs (five folds × three random seeds).

3.2 Evaluating Indel Pathogenicity Prediction Performance Across Conformational Ensemble Sizes

We evaluated how the size of the conformational ensemble used at training and inference affected model performance by varying the number of sampled conformations (k = 1, 3, 5, and 10). The indel AUROC increased consistently from k = 1 to k = 10 across all six PLMs (Figure 3). In addition, for a single protein with a large number of both benign and pathogenic indel variants, we increased the ensemble size to 50 at inference only, using models trained with k = 10, and observed a steady improvement in performance (Appendix C). These results suggest that performance may not have reached a plateau at ten conformations, and further increasing the ensemble size may be expected to yield additional gains in predictive performance.

3.3 Conformational Ensembles Yield More Variant-Specific Representations and More Consistent Predictions

In five of the six PLMs, representations built from a single sampled conformation captured more variant-specific variation than those built from the static indel structure (Table 1). To examine how structure inputs shape the learned representations, we measured the maximum explainable variance (MEV) of the output embeddings. MEV is the fraction of variance explained by the first principal component of a set of embeddings. A lower MEV indicates that the embeddings are less dominated by a single shared direction and therefore more context-specific. For each protein with at least five indel variants, we embedded every indel variant using either its static indel structure or one sampled conformation per variant, and computed MEV across the variants of that protein. Repeating the latter for each of the ten sampled conformations gave ten MEV values per model (Table 1). In five of the six PLMs, representations built from a single sampled conformation had a lower MEV than those built from the static indel structure (e.g., ESM3, 0.2872 ± 0.0057 vs. 0.3130; ProSST, 0.3886

Table 1: Maximum explainable variance (MEV) of indel variant embeddings under different structure inputs (Sanabria et al., 2024; Ethayarajh, 2019). Lower MEV indicates more context-specific embeddings. For the single-conformation setting, MEV was computed separately for each of the ten sampled conformations and is reported as mean ± SD. The pooled setting combines embeddings from all ten conformations ($10n$ embeddings per protein) and is not size-matched to the other two settings.

Model	**Static structure**	**Single conformation (mean ± SD, $n = 10$)**	**Pooled ($10n$)**
ESM3-open	0.3130	0.2872 ± 0.0057	0.1816
SaProt-650M	0.3885	0.3676 ± 0.0076	0.3064
ProSST-2048	0.4275	0.3886 ± 0.0075	0.3367
VenusREM	0.4285	0.3921 ± 0.0061	0.3459
S2F	0.4416	0.4727 ± 0.0052	0.4240
AIDO.RAG-16B	0.3958	0.3823 ± 0.0040	0.3240

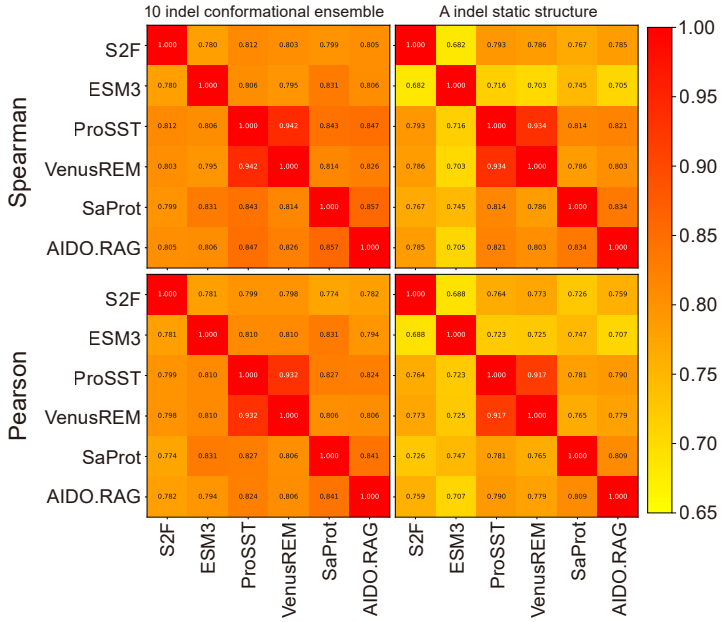

Figure 4: Consistency of predictions across PLMs. Spearman (top) and Pearson (bottom) correlations of variant-level predictions between pairs of PLMs for the 10-conformation ensemble (left) and the static indel structure (right).

± 0.0075 vs. 0.4275). When the embeddings from all ten conformations were pooled (10n embeddings per protein), MEV decreased further in all six PLMs, including S2F (e.g., ESM3, 0.1816; S2F, 0.4240). Because the pooled set contains ten times as many embeddings as the single-structure settings, this comparison is not size-matched. It nonetheless indicates that the ten conformations span representational directions that no single structure captures. We also assessed prediction consistency across the 15 training runs (five folds × three random seeds). With the 10-conformation ensemble, all six PLMs showed high seed-to-seed Spearman correlations of variant-level predictions ($\rho \geq 0.89$; Appendix D). We also compared predictions across PLMs (Figure 4). With the 10-conformation ensemble, inter-model Spearman correlations were higher than with the static indel structure for all 15 model pairs (range, 0.780 to 0.942 vs. 0.682 to 0.934), and Pearson correlations showed the same trend. Taken together, these results suggest that conformational ensembles lead to more variant-specific representations and reduce model-specific variability in the resulting predictions, so that different structure-informed PLMs converge toward more similar pathogenicity estimates.

3.4 Exploration of Improvement by Conformation Ensembles

The consensus variants improved by the conformational ensembles across the top-performing PLMs formed a structurally distinct group characterized by more heterogeneous ensembles, longer deletions, disordered WT environments, and benign variants (Figure 5). Focusing on the five PLMs (SaProt, ESM3, S2F, VenusREM, and AIDO) that exhibited significant improvements with the conformational ensemble over the static indel structure, consensus variants carried significantly more heterogeneous ensembles, with a median ensemble diversity of 0.325 compared to 0.302 in the remaining variants ($P = 1.5 \times 10^{-10}$, Mann-Whitney U test, Figure 5a). Ensemble diversity was defined as the mean normalized Hamming distance between the Foldseek 3Di strings of all conformer pairs in the conformational ensemble of each variant. This trend remained consistent across both insertions and deletions. In the case of deletions, variant length was significantly greater in consensus variants than in the remaining variants, with a median of 4 compared with 2 ($P = 1.5 \times 10^{-9}$, Mann-Whitney U test, Figure 5e). Consensus variants were located in low-contact and low-confidence regions of the WT static structure. We compared consensus variants with the remaining variants at the variant start position and within the ±7-residue window centered on it, which corresponds to the region used for variant embedding. In the WT static structure, the median contact number at the variant start position was 2 in the consensus group versus 9 in the remaining variants ($P = 7.0 \times 10^{-81}$, Mann-Whitney U test, Figure 5f). Within the ±7-residue window, the median of the window-averaged contact number was 1.6 versus 10.7 ($P = 9.8 \times 10^{-84}$, Figure 5g). Consistent with this, pLDDT scores were significantly lower in the consensus group both at the start position (38.7 vs. 86.1, $P = 1.1 \times 10^{-80}$, Figure 5h) and within the window (38.9 vs. 85.5, $P = 1.0 \times 10^{-84}$, Figure 5i). This tendency of lower pLDDT and fewer WT contacts remained consistent across both insertions and deletions. These results suggest that consensus variants tend to occur in regions where the static structure is uncertain, where a single conformation may not adequately represent the local structure and a conformational ensemble may therefore provide more informative structural context. Consensus variants were also overwhelmingly benign, comprising 96.5% of the set compared to 32.6% of the remaining variants ($P = 5.9 \times 10^{-142}$, Figure 5j). The improvements observed across the two labels manifested differently: benign variants shifted uniformly by a small amount towards the true label, while pathogenic variants split in both directions (Appendix E). We also noted that ensemble diversity increased with indel length, suggesting that longer indels drive the generator toward a broader set of structural conformations (Appendix F).

3.5 Performance Comparison with External Models for Indel Pathogenicity Prediction

To benchmark our approach, we compared the predictive performance of the six PLMs against established external baselines, CADD and INDELi-E (Schubach et al., 2024; Topolska et al., 2025). All PLMs consistently outperformed CADD across every evaluation cell (Appendix G). Specifically, on the CADD-scored subsets, the conformational ensemble exceeded CADD's mean AUROC (0.8515), achieving means up to 0.9373. When evaluated on single-residue indels scored by both baselines, INDELi-E proved to be a stronger baseline than CADD (mean AUROC of 0.884 vs. 0.799). Nevertheless, the conformational ensemble still outperformed INDELi-E across all six PLMs.

4 Conclusion

Across six structure-informed PLMs, the 10-conformation ensemble significantly outperformed single-structure inputs in 15 of 18 comparisons. A static structure provides a single structure token per residue, an ensemble provides a distribution of structure tokens whose variability might reflect local flexibility. Embeddings derived from conformational ensembles were more variant-specific, and predictions agreed more closely across PLMs. The variants with improved predictions across PLMs were mostly located in regions with low pLDDT and few residue contacts, where a single structure is unlikely to represent the local structure, and most of these variants were benign. This suggests that the ensemble mainly improves the recognition of indels that are tolerated in flexible regions, consistent with previous reports linking indel tolerance to local flexibility. Because this method focuses on how to utilize structure tokens, it could be naturally applied to other structure-informed PLMs.

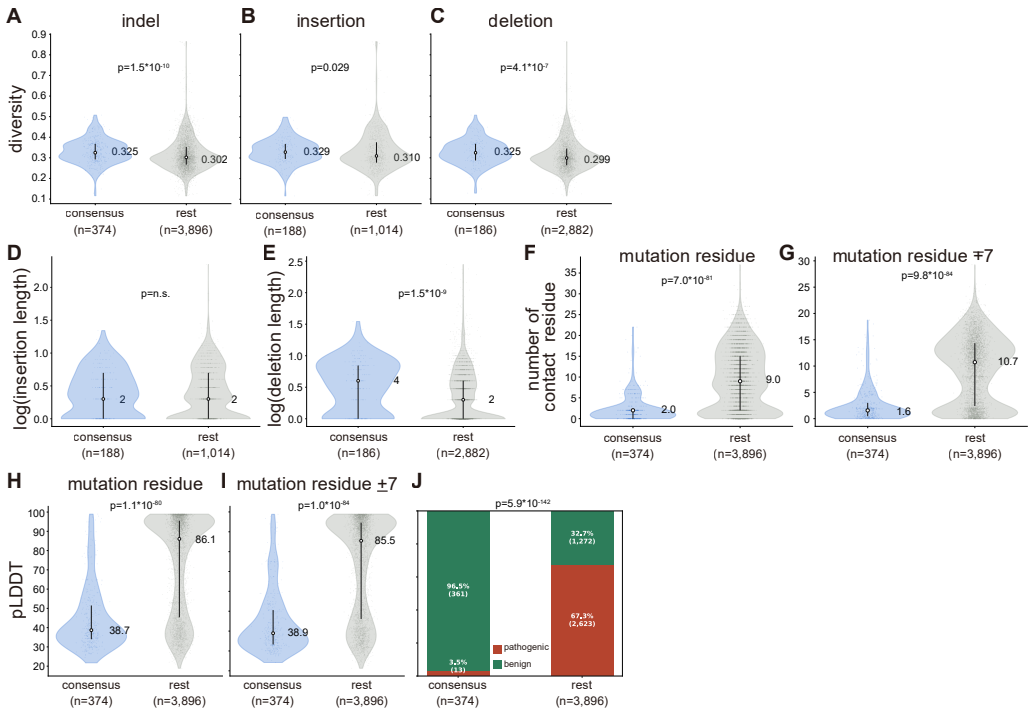

Figure 5: Characteristics of variants consistently improved by the conformational ensemble. Consensus variants are indels whose predictions were improved by the 10-conformation ensemble relative to the static indel structure in all five PLMs with a significant gain, and the remaining variants are labeled rest. (A, B, C) Ensemble diversity for all indels, insertions, and deletions. (D, E) Insertion and deletion lengths on a log10 scale. (F, G) Number of contact residues in the static WT structure at the variant start position and averaged over the ±7-residue window. (H, I) pLDDT of the static WT structure at the variant start position and averaged over the ±7-residue window. (J) Proportions of pathogenic and benign variants. White dots and adjacent numbers indicate medians, and vertical lines indicate interquartile ranges. p values were obtained with the Mann-Whitney U test for A to I and with Fisher's exact test for J.

5 LIMITATION

This study focused on whether the benefit of conformational ensembles holds across structure-informed PLMs with different architectures and structure encodings. All PLMs were used as frozen feature extractors with the same shallow classifier head, a setting close to linear probing, thus absolute performance could be further improved by fine-tuning the PLMs, optimizing the head architecture, or aggregating conformational ensembles with a learned head. To ensure a fair comparison, we used an identical prediction head architecture. The optimal combination of generated embeddings and the classifier head would vary across models, requiring a distinct, optimal head design for each PLM (Hou et al., 2026; Schmirler et al., 2024). Ensembles were generated only with BioEmu, and the optimal ensemble size and the agreement of these ensembles with MD simulations remain to be determined. As structure tokenizers and generative models advance, representations that more precisely capture protein dynamics are expected to emerge, and our framework can readily accommodate these future improvements. Although we explored larger ensemble sizes (up to 50 conformations) on a subset of the data, an exhaustive evaluation across the entire dataset was not feasible due to high computational costs.

AI USE STATEMENT

In this work, we have not used generative AI tools for proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, or implementing methods, and the remaining tasks with required disclosure are not applicable to this work. Additionally, we used generative AI tools for searching for information, editing this paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. The authors personally checked all AI-generated outputs, including the edited text, the retrieved information, and the identified literature, to confirm that they are consistent with the authors' intent. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All details necessary to reproduce our results are provided in the main text. Section 2 describes the model architecture, the training procedure, the datasets and data processing steps, and the evaluation protocol.

REFERENCES

Haneen Abderrazzaq, Mugdha Singh, Larry Babb, Timothy Bergquist, Steven E Brenner, Vikas Pejaver, Anne O'Donnell-Luria, Predrag Radivojac, ClinGen Computational Working Group, ClinGen Variant Classification Working Group, et al. Calibration of in-frame indel variant effect predictors for clinical variant classification. *bioRxiv*, 2026.

Anupam Banerjee and Ivet Bahar. Structural dynamics predominantly determine the adaptability of proteins to amino acid deletions. *International Journal of Molecular Sciences*, 24(9):8450, 2023.

Timothy Bergquist, Sarah L Stenton, Emily AW Nadeau, Alicia B Byrne, Marc S Greenblatt, Steven M Harrison, Sean V Tavtigian, Anne O'Donnell-Luria, Leslie G Biesecker, Predrag Radivojac, et al. Calibration of additional computational tools expands clingen recommendation options for variant classification with pp3/bp4 criteria. *Genetics in Medicine*, 27(6):101402, 2025.

Gwen R Buel and Kylie J Walters. Can alphafold2 predict the impact of missense mutations on structure? *Nature structural & molecular biology*, 29(1):1–2, 2022.

Stuart Cannon, Matthew Williams, Adam C Gunning, and Caroline F Wright. Evaluation of in silico pathogenicity prediction tools for the classification of small in-frame indels. *BMC Medical Genomics*, 16(1):36, 2023.

Siwei Chen, Laurent C Francioli, Julia K Goodrich, Ryan L Collins, Masahiro Kanai, Qingbo Wang, Jessica Alföldi, Nicholas A Watts, Christopher Vittal, Laura D Gauthier, et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625(7993):92–100, 2024.

Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492, 2023.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Scott A Foster, Daniel M Whalen, Ayşegül Özen, Matthew J Wongchenko, JianPing Yin, Ivana Yen, Gabriele Schaefer, John D Mayfield, Juliann Chmielecki, Philip J Stephens, et al. Activation mechanism of oncogenic deletion mutations in braf, egfr, and her2. *Cancer cell*, 29(4):477–493, 2016.

Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801–807, 2014.

Grant Goldman, Prathamesh Chati, and Vasilis Ntranos. Uncovering differential tolerance to deletions versus substitutions with a protein language model. *Cell systems*, 16(9), 2025.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450 (7172):964–972, 2007.

Chao Hou, Di Liu, Aziz Zafar, and Yufeng Shen. Understanding language model scaling for protein fitness prediction. *Nature Computational Science*, pp. 1–11, 2026.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Melissa J Landrum, Shanmuga Chitipiralla, Kuljeet Kaur, Garth Brown, Chao Chen, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Chunlei Liu, Zenith Maddipatla, et al. Clinvar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Research*, 53(D1): D1313–D1321, 2025.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761): eadv9817, 2025.

Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. Prosst: Protein language modeling with quantized structure and disentangled attention. *Advances in Neural Information Processing Systems*, 37:35700–35726, 2024a.

Pan Li, Xingyi Cheng, Le Song, and Eric Xing. Retrieval augmented protein language models for protein structure prediction. *bioRxiv*, pp. 2024–12, 2024b.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Milot Mirdita, Lars Von Den Driesch, Clovis Galiez, Maria J Martin, Johannes Söding, and Martin Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research*, 45(D1):D170–D176, 2017.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in neural information processing systems*, 36: 64331–64379, 2023.

Marina A Pak, Karina A Markhieva, Mariia S Novikova, Dmitry S Petrov, Ilya S Vorobyev, Ekaterina S Maksimova, Fyodor A Kondrashov, and Dmitry N Ivankov. Using alphafold to predict the impact of single mutations on protein stability and function. *Plos one*, 18(3):e0282689, 2023.

Vikas Pejaver, Alicia B Byrne, Bing-Jian Feng, Kymberleigh A Pagel, Sean D Mooney, Rachel Karchin, Anne O'Donnell-Luria, Steven M Harrison, Sean V Tavtigian, Marc S Greenblatt, et al. Calibration of computational tools for missense variant pathogenicity classification and clingen recommendations for pp3/bp4 criteria. *The American Journal of Human Genetics*, 109(12):2163–2177, 2022.

S Richards, N Aziz, S Bale, D Bick, So Das, J Gastier-Foster, WW Grody, M Hegde, E Lyon, E Spector, et al. Acmg laboratory quality assurance committee. standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genet Med*, 17(5): 405–424, 2015.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the national academy of sciences*, 118(15):e2016239118, 2021.

Melissa Sanabria, Jonas Hirsch, and Anna R Poetsch. Distinguishing word identity and sequence context in dna language models. *BMC bioinformatics*, 25(1):301, 2024.

Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15(1):7407, 2024.

Max Schubach, Thorben Maass, Lusiné Nazaretyan, Sebastian Röner, and Martin Kircher. Cadd v1. 7: using protein language models, regulatory cnns and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic acids research*, 52(D1):D1143–D1154, 2024.

Go Hun Seo, Taeho Kim, In Hee Choi, Jung-young Park, Jungsul Lee, Sehwan Kim, Dhong-gun Won, Arum Oh, Yena Lee, Jeongmin Choi, et al. Diagnostic yield and clinical utility of whole exome sequencing using an automated variant prioritization system, evidence. *Clinical genetics*, 98(6):562–570, 2020.

Amelie Stein, Douglas M Fowler, Rasmus Hartmann-Petersen, and Kresten Lindorff-Larsen. Biophysical and mechanistic models for disease-causing protein variants. *Trends in biochemical sciences*, 44(7):575–588, 2019.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, volume 2024, pp. 6987–7009, 2024.

Mahlet Z Tamirat, Marika Koivu, Klaus Elenius, and Mark S Johnson. Structural characterization of egfr exon 19 deletion mutation using molecular dynamics simulation. *PloS one*, 14(9):e0222814, 2019.

Yang Tan, Ruilin Wang, Banghao Wu, Liang Hong, and Bingxin Zhou. From high-throughput evaluation to wet-lab studies: advancing mutation effect prediction with a retrieval-enhanced model. *Bioinformatics*, 41(Supplement_1):i401–i409, 2025.

Magdalena Topolska, Antoni Beltran, and Ben Lehner. Deep indel mutagenesis reveals the impact of amino acid insertions and deletions on protein stability and function. *Nature communications*, 16(1):2617, 2025.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, et al. Attention is all you need [j]. *Advances in neural information processing systems*, 30(1):261–272, 2017.

Zuobai Zhang, Pascal Notin, Yining Huang, Aurélie Lozano, Vijil Chenthamarakshan, Debora Marks, Payel Das, and Jian Tang. Multi-scale representation learning for protein fitness prediction. *Advances in Neural Information Processing Systems*, 37:101456–101473, 2024.

Shuxin Zheng, Jiyan He, Chang Liu, Yu Shi, Ziheng Lu, Weitao Feng, Fusong Ju, Jiaxi Wang, Jianwei Zhu, Yaosen Min, et al. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence*, 6(5):558–567, 2024.

A MISSENSE VARIANT PERFORMANCE UNDER DIFFERENT STRUCTURE TOKENS FOR INDELS.

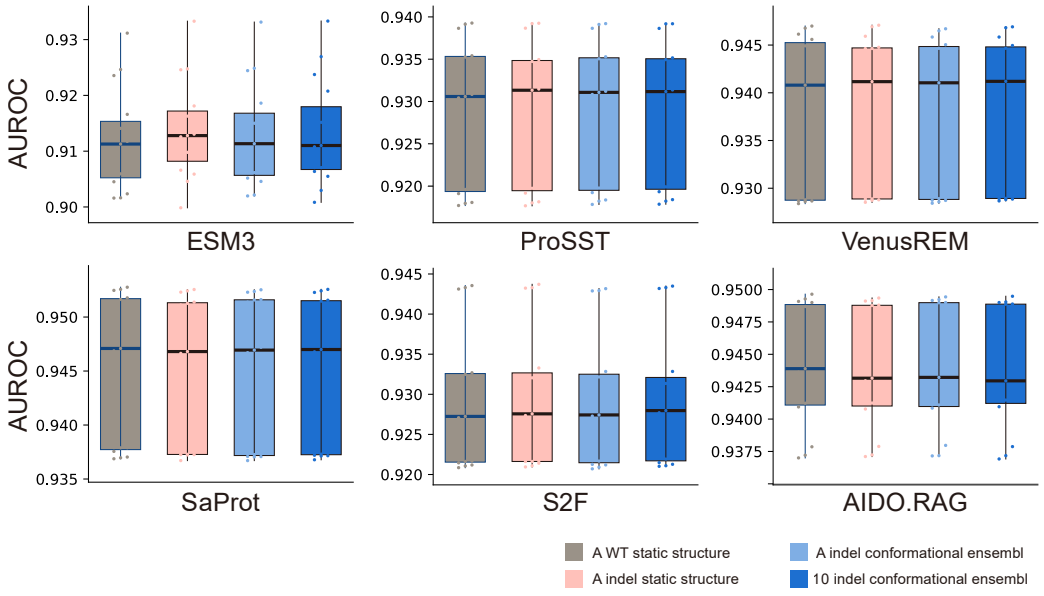

Figure 6: AUROC on missense variants in the test folds is shown when the static WT structure (gray), the static indel structure (pink), a single sampled conformation (light blue), or the 10-conformation ensemble (dark blue) was provided for indel variants. Missense variants received the static WT structure in all settings. Each point represents one of 15 runs, and boxes show the median and interquartile range with whiskers showing the full range.

B INDEL PATHOGENICITY PREDICTION WITH WT-DERIVED STRUCTURE TOKENS.

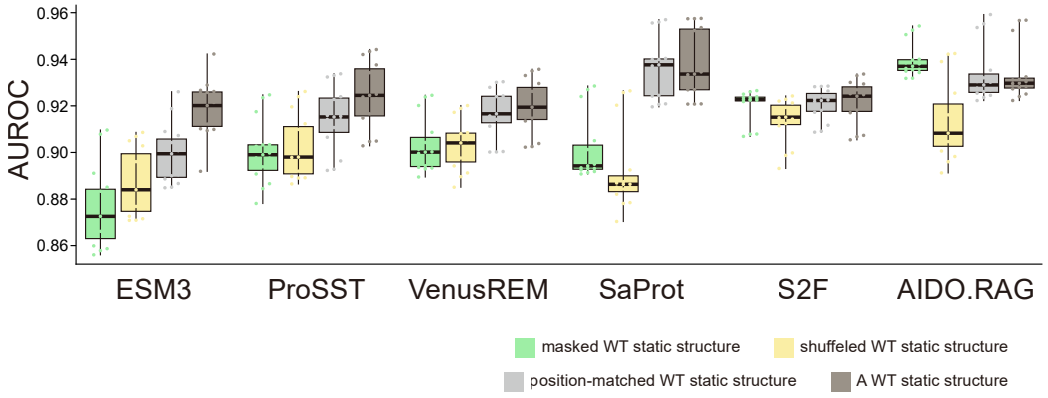

Figure 7: AUROC on the test folds is shown when each indel variant received the unmodified static WT structure with end padding (dark gray), a position-matched WT structure in which WT structure tokens were projected onto the indel sequence by alignment (light gray), a shuffled WT structure in which the order of WT structure tokens was randomly permuted (yellow), or a masked WT structure in which all structure tokens were masked (green). Each point represents one of 15 runs, and boxes show the median and interquartile range with whiskers showing the full range.

C EFFECT OF THE NUMBER OF CONFORMATIONS USED AT INFERENCE FOR A SINGLE PROTEIN.

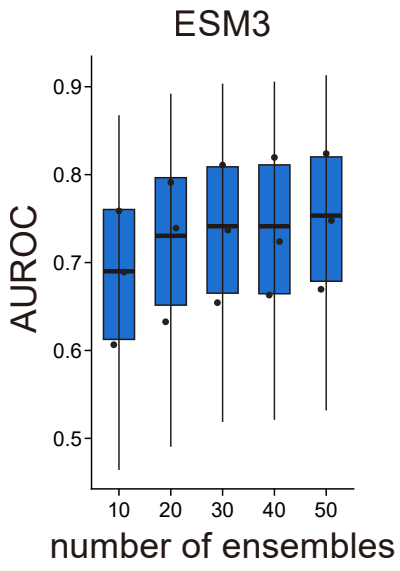

Figure 8: ESM3 models trained with k = 10 were evaluated on the in-frame indels of MECP2 (NP_001104262.1), which has many benign and pathogenic indels, while the number of conformations averaged at inference was increased from 10 to 50. To prevent data leakage under the similarity-based split, only 3 of the 15 trained models were evaluated (shown as points). Boxes summarize 2,000 paired bootstrap resamples of the 43 test indels (20 pathogenic, 23 benign) per model, showing the median and interquartile range, with whiskers spanning the 2.5th to 97.5th percentiles.

D CONSISTENCY OF PREDICTIONS ACROSS RANDOM SEEDS.

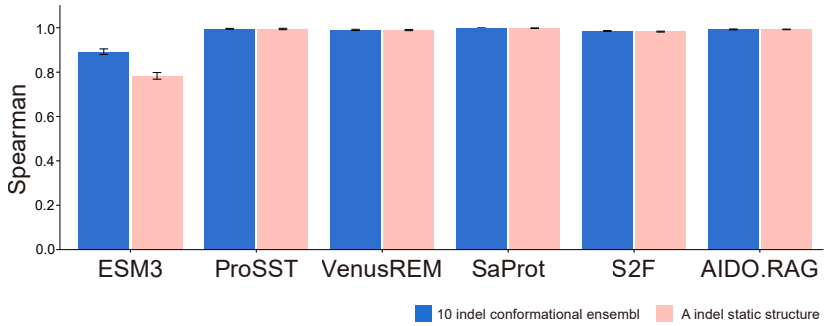

Figure 9: Spearman correlation of variant-level predictions between pairs of random seeds for the 10-conformation ensemble (blue) and the static indel structure (pink). Bars show the mean, and error bars show the standard error across seed pairs.

E DISTRIBUTION OF NORMALIZED IMPROVEMENT BY PATHOGENICITY LABEL.

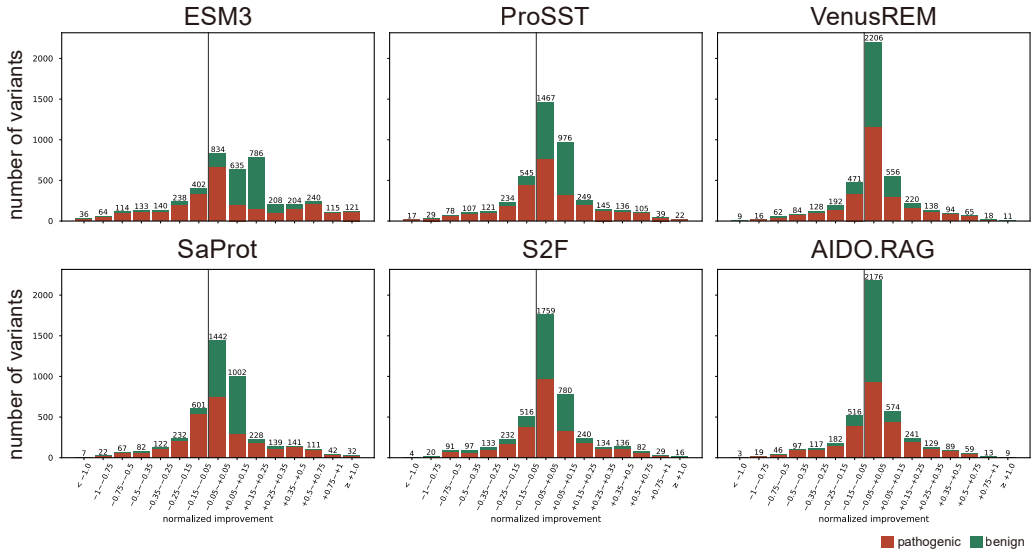

Figure 10: For each PLM, all test indels (n = 4,270) are binned by the normalized improvement of the 10-conformation ensemble, and positive values indicate that the prediction moved toward the true label. Normalized improvement measures the shift in standardized predicted probability toward the true label when comparing the 10-conformation ensemble to the static AlphaFold structure of the same indel. Red and green segments indicate pathogenic and benign variants, respectively, and numbers above the bars indicate variant counts.

F ENSEMBLE DIVERSITY AGAINST VARIANT LENGTH.

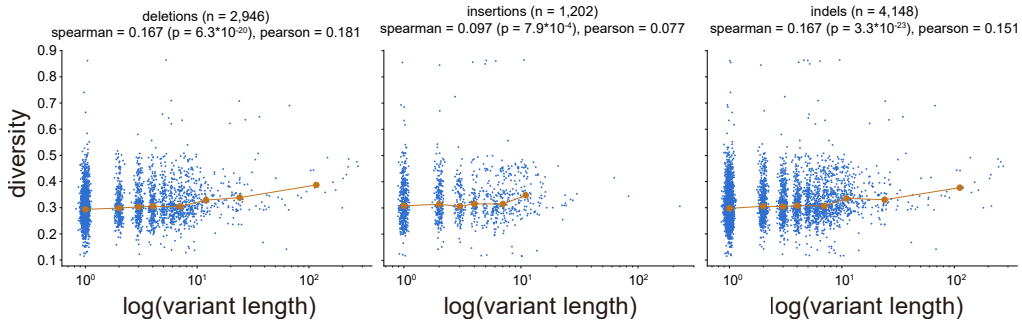

Figure 11: Ensemble diversity is plotted against variant length on a logarithmic scale for deletions (n = 2,946), insertions (n = 1,202), and all indels (n = 4,148). Orange points connected by lines indicate the median diversity within specific length bins ([1, 2), [2, 3), [3, 4), [4, 6), [6, 10), [10, 20), [20, 50), and [50, ∞)). Spearman and Pearson correlation coefficients are shown above each panel.

G COMPARISON WITH EXTERNAL PREDICTORS.

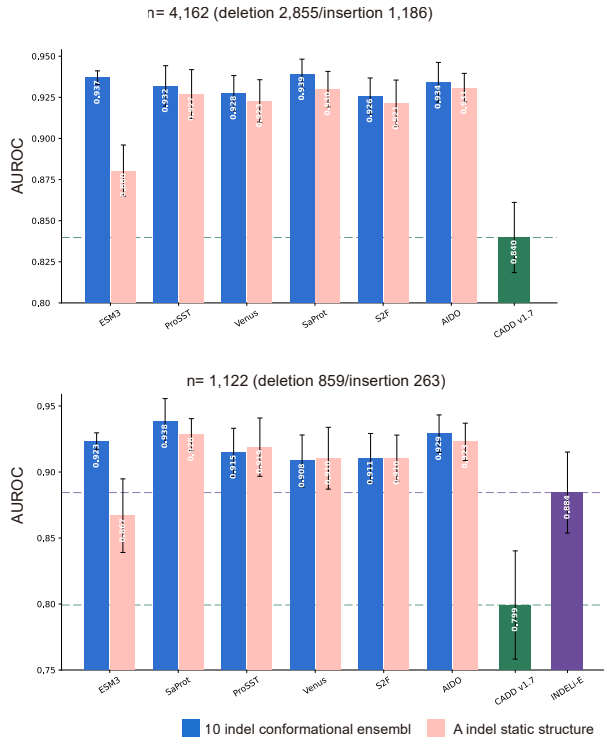

Figure 12: Comparison with external predictors. (Top) AUROC of the six PLMs with the 10-conformation ensemble (blue) or the static indel structure (pink) and of CADD v1.7 (green) on indels scored by CADD (n = 4,162). (Bottom) The same comparison with INDELi-E (purple) added, on single-residue indels scored by both CADD and INDELi-E (n = 1,122). Bars show the mean AUROC across evaluation runs, error bars show the standard deviation, and dashed lines mark the AUROC of the external predictors.