Multilingual model improves zero-shot prediction of disease effects on proteins

Ruyi Chen[1], Nathan Palpant[2], Gabriel Foley[1], Mikael Bodén[1,*]

[1]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia.

[2]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia.

[*]To whom correspondence should be addressed: `m.boden@uq.edu.au`.

Abstract

Models for mutation effect prediction in coding sequences rely on sequence-, structure-, or homology-based features. Here, we introduce a novel method that combines a *codon* language model with a *protein* language model, providing a *dual* representation for evaluating effects of mutations on disease. By capturing contextual dependencies at both the genetic and protein level, our approach achieves a 3% increase in ROC-AUC classifying disease effects for 137,350 ClinVar missense variants across 13,791 genes, outperforming two single-sequence-based language models. Obviously the codon language model can uniquely differentiate synonymous from nonsense mutations at the genomic level. Our strategy uses information at complementary biological scales (akin to human multilingual models) to enable protein fitness landscape modeling and evolutionary studies, with potential applications in precision medicine, protein engineering, and genomics.

Keywords: Protein language models, Codon language models, Zero-shot prediction, Disease variant effect

1 Main

Single-point mutations in amino acid sequence can significantly impact protein function and stability, underscoring the necessity of accurate mutation effect predictions for both protein engineering and evolutionary biology. Computational approaches can expedite the classification of such variants. Unlike traditional supervised learning that depends on labels, which in turn tend to be noisy and imbalanced, Large Language Models (LLMs) are label-free. Biological Language Models (BLMs) are trained to capture dependencies inherent in biological sequence data, and have been shown to predict the mutation effect without being trained to do so ("zero-shot learning"). In this setting, inference is performed via a forward pass of the wild-type sequence, estimating the probability of a token (e.g., an amino acid) at a given position based on its surrounding sequence context (e.g., within a protein). The effect score is then measured by the differences in Log-Likelihood Ratios (LLRs) between the mutated and reference tokens [1]. Based on the nature of the training data, these models broadly fall into three categories: protein sequence-based [1, 2], structure-based [3, 4], and alignment-based [5, 6]. The integration of diverse types of information (e.g., sequences and structures) enhances the reliability of LLR estimates and consequently leads to improved prediction performance [7, 8]. Here, we explore the use of complementary sources of information to enhance predictive robustness.

Natural languages (e.g., English and German) may convey semantics that are largely similar yet subtly different in interpretation (Fig.1a) [9]. Analogously, genetic information flows from DNA to mRNA to proteins, mirroring how core meanings transcend languages. We reason that language models trained on equivalent areas but expressed in distinct "languages", such as protein-coding DNA (cDNA) and their corresponding protein sequences, are likely to capture analogous linguistic properties (Fig.1b). If this holds true, integrating mutation information from both cDNA and protein sequences could offer complementary insights, thereby improving prediction performance.

As a proof of principle, we tested combining two BLMs, CaLM, a Codon Language Model (CLM) [10], with ESM-2, a Protein Language Model (PLM) [11], representing coding sequences at two distinct scales, to evaluate their ability to enhance zero-shot disease effect prediction (Methods and Fig.1d). We curated three types of single-point mutations from ClinVar, each with validated mutation effects (Methods and Fig.1c). We then computed and compared missense effect scores generated by CaLM, ESM-2, and a late-integrating hybrid of the two using a benchmark dataset comprising 137,350 missense variants (Methods and Fig.2a). The codon-level effect scores are effective in distinguishing benign (blue) from pathogenic (red) variants. However, its discriminative power is slightly lower than that of ESM-2 and the hybrid model, as evidenced by a density curve overlap of 0.4684. Among the three models evaluated as classifiers for distinguishing 39,280 pathogenic variants from 98,070 benign variants, our hybrid model achieved the highest receiver operating characteristic area under the curve (ROC-AUC) score of 0.8629, significantly outperforming the single models ESM-2 (0.8336) and CaLM (0.8214) (Fig.2b, two-tailed DeLong test, $p < 0.01$). Moreover, it also surpassed in weighted area under the precision-recall curve (AUPRC) scores (Extended Data Fig.1). Different ways for incorporating CLM and PLM effect

scores were explored (Extended Data Table 2); however, the simpler approach yielded comparable performance and is reported in detail below.

We also evaluated CaLM's performance in classifying *nonsense* and *synonymous* mutations, which are undetectable by PLMs. We integrated three types of single-point mutations ("*All*" label in Fig.1c and and Extended Data Table 1) and computed their effect scores, resulting in an overall ROC-AUC of 0.9017 and AUPRC of 0.8330 (Extended Data Fig.2). This is primarily due to the imbalance in variant distribution: nonsense mutations are predominantly pathogenic, whereas synonymous mutations are generally benign. The CaLM classifies these categories effectively, assigning nonsense mutations lower effect scores compared to synonymous mutations. However, it struggles to differentiate benign nonsense from pathogenic synonymous mutations.

Finally, we investigated the underlying reasons why combining PLM and CLM enhances disease effect prediction by analysing missense mutations that render significantly different effect scores. In particular, we identified 20,423 mutations divided into two groups: *CLM better* and *PLM better* (Methods and Fig.2c). For each group, we compared the frequency distribution of wild-type and mutated amino acid pairs with the overall distribution and identified statistically enriched pairs (see Methods). We found that the differences in CLM and PLM effect scores centred on whether the missense mutation involved a transition from a high-codon to a low-codon amino acid (say, an amino acid with six codons to one with two codons) or vice versa (see Fig.2d,e).

Using a linguistic analogy, high-codon amino acids (like words with many synonyms) flexibly permit substitutions, while low-codon amino acids (like unique words) are less tolerant to change. In essence, the PLM focuses on amino acid properties (semantics of words), whereas the CLM also incorporates codon redundancy and translation efficiency (word frequency). Consequently, mutations involving shifts between high-codon and low-codon amino acids (in either direction) lead to significant differences in effect prediction. Interestingly, the two models differ in predicting specific mutation types; for example, the CLM more accurately identifies threonine (T) to methionine (M) substitutions, whereas the PLM performs better for the reverse (M to T). This asymmetry suggests that the models capture different aspects of codon bias and may thus complement each other. Thus, the enhanced predictive performance may be linked to codon usage, though not exclusively. Other factors, including mutation positional context and evolutionary conservation may also play a role.

Linguistics serves as a key inspiration in this study, particularly the notion that the same phenomenon (e.g., missense variants) can be expressed in different languages (e.g., BLMs), where integrating information enhances semantic understanding (e.g., disease effect prediction). Our hybrid model surpasses both single models in ROC-AUC and AUPRC on ClinVar missense variants, highlighting the complementary strengths of language models. This improvement partly stems from codon usage during translation, reflecting the lexical specificity of different languages. Beyond disease effect prediction, our strategy could extend to protein fitness landscape modeling and evolutionary studies by incorporating language models at different scales, with potential applications in precision medicine, protein engineering, and genomics.

2 Methods

2.1 Single-point mutation acquisition

ClinVar [12] is a comprehensive, publicly accessible database that aggregates genetic variants associated with diseases along with their clinical phenotypes and supporting evidence. In this study, we downloaded the complete set of variant annotations from the file `variant_summary.txt` (up to October 8, 2024) available on ClinVar's FTP site (https://ftp.ncbi.nlm.nih.gov/pub/clinvar). To facilitate analysis, we filtered the dataset to include only three types of single-point variants: *missense*, *nonsense*, and *synonymous*. They were further selected using the following criteria:

1. Mutations based on clinical significance are labelled as benign (including *Benign* and *Likely Benign*) or pathogenic (including *Pathogenic* and *Likely Pathogenic*)
2. Mutations must have at least one reviewer star
3. Mutations associated with extremely large genes (e.g., *TTN*) were removed
4. For genes with mutations across multiple transcripts, only the transcript with the highest number of annotated mutations was retained, ensuring that each gene is matched by a single transcript

The final benchmark dataset comprised 137,350 missense variants across 13,791 genes, 46,386 nonsense variants across 3,291 genes, and 527,579 synonymous variants across 11,593 genes (Fig.1c and Extended Data Table 1).

2.2 Wild-type sequence collection

For each provided gene accession, we retrieved complete GenBank records from the NCBI GenBank database. We then parsed these records to identify coding sequence features, extracting the corresponding cDNA sequences and their protein products.

2.3 CaLM and ESM-2

We employed both CaLM (86M; downloaded from https://github.com/oxpig/CaLM) [10] and ESM-2 (150M; downloaded from https://huggingface.co/facebook/esm2_t30_150M_UR50D) [11], two transformer-based language models. They were trained via masked language modeling objectives; however, they differ in tokenization strategies, training datasets, and output representations. Specifically:

1. Tokens: CaLM tokenizes cDNA sequences into codons, whereas ESM-2 tokenizes protein sequences into residues
2. Training data: CaLM is trained on a corpus of 9M cDNA sequences sourced from the European Nucleotide Archive, whereas ESM-2 leverages protein sequences derived from the UniRef database
3. Output logits: CaLM generates position-specific effect scores for all 64 codons at each codon position, while ESM-2 produces effect scores for each of the 20 canonical amino acids at each position

Although neither model enforces sequence length constraints, processing long sequences remains computationally prohibitive.

2.4 The hybrid model

Language models estimate the probability of a mutation at a given position using only its surrounding context without additional training. The wild-type cDNA sequence was input into CaLM, while its corresponding wild-type protein sequence was processed using ESM-2. A softmax function was then applied to generate an uncalibrated probability distribution over codons or residues.

Consider a cDNA sequence $\mathbf{x} = (x_1, x_2, \ldots, x_{3L+3})$, where each nucleotide $x_i \in \{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$ and L denotes the number of residues encoded by the cDNA sequence. It comprises a start codon, a coding region, and a stop codon, resulting in a total length of $3L + 3$ nucleotides. A codon $c_k = (x_{3k-2}, x_{3k-1}, x_{3k})$, for $k = 1, 2, \ldots, L + 1$, where $k = L + 1$ corresponds to the stop codon. During translation, each codon c_k is mapped to a residue according to the genetic code. We thus define the protein sequence as $\mathbf{y} = (y_1, y_2, \ldots, y_L)$, where each y_i represents one of the 20 standard amino acids (typically denoted by their one-letter codes, e.g., $\{\mathrm{A}, \mathrm{R}, \mathrm{N}, \mathrm{D}, \mathrm{C}, \ldots\}$). (Note that although the cDNA sequence contains $L + 1$ codons, the last codon c_{L+1} is a stop codon and does not encode a residue.) Only the first L triplet of three nucleotides are encoded into amino acids.

A missense mutation at position i in a cDNA sequence, denoted by $\tilde{x}_i$, may alter the codon c_k^{wt} to a mutated form $\tilde{c}_k$, thereby replacing the residue y_k^{wt} with $\tilde{y}_k$ in the protein sequence. To quantify the effect of this mutation, we compute LLRs at both the codon and residue levels:

$$\mathrm{LLR}_{c_k} = \log \frac{\Pr(c_k = \tilde{c}_k \mid \theta)}{\Pr(c_k = c_k^{\mathrm{wt}} \mid \theta)}, \tag{1}$$

$$\mathrm{LLR}_{y_k} = \log \frac{\Pr(y_k = \tilde{y}_k \mid \theta)}{\Pr(y_k = y_k^{\mathrm{wt}} \mid \theta)}, \tag{2}$$

where θ represents a set of evolutionary constraint parameters optimized to reproduce the statistical properties of observed sequence evolution

Our hybrid model integrates effect scores at two levels through a linear combination:

$$\mathrm{LLR} = \frac{\alpha \cdot \mathrm{LLR}_{c_k} + \beta \cdot \mathrm{LLR}_{y_k}}{2}, \tag{3}$$

where weighting factors α and β regulate the relative contributions of codon- and residue-level effects

The Bayesian optimization [13, 14] was conducted using a Gaussian process-based search over the weight parameters $\alpha, \beta \in [-5, 5]$, aiming to maximize the ROC-AUC score. The optimization process involved an initial exploration of 10 iterations followed by 100 further iterations, implemented via the `BayesianOptimization` function from bayesian-optimization v.2.0.3. Ultimately, equal weights were assigned to both terms.

For other mutation types, only LLR_{c_k} is used.

2.5 Metrics for pathogenicity variant classification

The ROC-AUC metric quantifies the trade-off between the true-positive rate (TPR) and the false-positive rate (FPR), thereby providing a robust overall assessment of classifier performance. However, in datasets with significant class imbalance, ROC-AUC

may not fully capture nuanced performance differences. To address this limitation, we employed a balanced AUPRC metric to assess the precision-recall trade-off, as its sensitivity to class imbalance makes it a more reliable performance measure. These metrics were computed via Scikit-Learn v.1.6.1.

We performed the DeLong test to compare AUCs from different models and assess statistical significance ($p < 0.01$) as a two-tailed test.

2.6 Significantly different mutations identification

To quantify the differences in effect scores assigned by CLM and PLM for each missense variant k, we define a difference metric, *Diff*:

$$\textit{Diff} = \alpha\,\mathrm{LLR}_{y_k} - \beta\,\mathrm{LLR}_{c_k}, \tag{4}$$

(We set $\alpha = \beta = 1$ based on our prior calculations.)

To identify mutations with statistically significant differences between LLR_{y_k} and LLR_{c_k}, we performed a two-tailed hypothesis test based on the standard normal distribution. Specifically, we utilized scipy.stats v.1.13.1 to compute the critical values at a significance level of $\alpha_{\mathrm{level}} = 0.01$. The thresholds were determined using the `scipy.stats.norm.ppf` function. Mutations with *Diff* values exceeding these thresholds ($\textit{Diff} < -2.58$ or $\textit{Diff} > 2.58$, corresponding to $p < 0.01$) were classified as *statistically significant*, yielding a total of 20,423 missense mutations in our following analysis.

We further categorized these mutations into two datasets based on the performance of a PLM and a CLM:

1. **PLM better**: For pathogenic mutations, LLR_{y_k} was lower, whereas for benign mutations, LLR_{y_k} was higher compared to LLR_{c_k}
2. **CLM better**: LLR_{c_k} followed the opposite trend relative to LLR_{y_k}

Using this standard, we finally identified 11,536 mutations as *CLM better* and 8,887 mutations as *PLM better*.

2.7 Missense variant preferences characterization

To evaluate whether a PLM or a CLM exhibits a preference for specific types of missense mutations, we analyzed the frequency distributions of wild-type and mutated amino acid pairs within each group and compared them to their overall distribution across all missense mutations. The frequency of a given amino acid pair p (e.g., M→E) in a dataset D is defined as:

$$f_D(p) = \frac{N_D(p)}{\sum_{p' \in D} N_D(p')}, \tag{5}$$

where $N_D(p)$ denotes the occurrence count of an amino acid pair p in dataset D, and the denominator represents the total occurrences of all amino acid pairs in D

Similarly, the overall frequency of a pair p across the whole dataset is defined as:

$$f_{\text{overall}}(p) = \frac{N_{\text{overall}}(p)}{\sum_{p' \in \text{overall}} N_{\text{overall}}(p')}, \tag{6}$$

where $N_{\text{overall}}(p)$ represents the total occurrences of a pair p in the overall dataset

To quantify the relative enrichment or depletion of specific amino acid pairs in the *PLM better* and *CLM better* datasets, we computed the Bias Ratio (BR) as the ratio of subset pair frequency to overall frequency. For a pair p in dataset D, BR is defined as:

$$\text{BR}_D(p) = \frac{f_D(p)}{f_{\text{overall}}(p)}, \tag{7}$$

A pair was classified as *enriched* in the dataset D if $\text{BR}_D(p) > 1$, indicating a higher occurrence relative to the overall dataset, whereas pairs with $\text{BR}_D(p) < 1$ were considered *depleted*, signifying a lower occurrence compared to the overall distribution

We then employed chi-square tests to assess the statistical significance of the observed enrichment and depletion patterns. They were performed using the `chi2_contingency` function from scipy.stats v.1.13.1. To account for multiple hypothesis testing, the Benjamini-Hochberg procedure was implemented via the `multipletests` function from statsmodels.stats.multitest v.0.14.4 to control the false discovery rate of 0.01. Amino acid pairs with adjusted p-values below this threshold were considered statistically significant.

2.8 Data analysis and visualization

Data analysis and plot preparation were performed in Jupyter Notebook v.6.5.5 using Python v.3.11.11 and the following libraries: numpy v.1.26.4, pandas v.2.2.2, matplotlib v.3.10.0, and seaborn v.0.13.2.

3 Data availability

Data and results are available at: https://github.com/Cassie818/Viral-mut.

4 Code availability

All the code and data necessary to reproduce the analyses, figures, and results are available on our GitHub repository: https://github.com/Cassie818/Viral-mut.

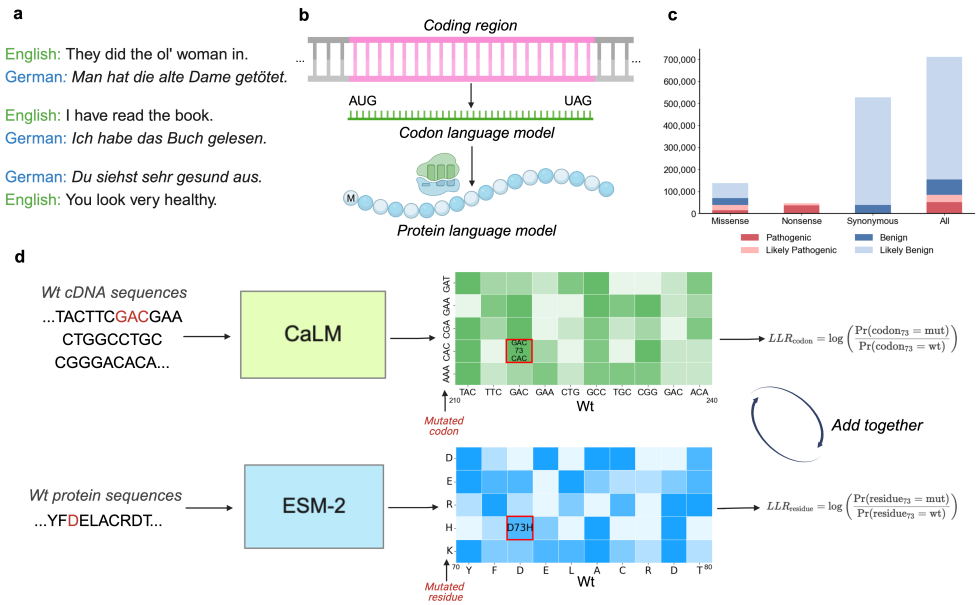


Fig. 1: The overall workflow of the proposed method.

a, Different natural languages (e.g., English and German) convey similar semantics with subtle differences. **b,** An analogy between natural languages and biological language models. Just as different languages (e.g., English and German) convey similar meanings with subtle interpretational differences, two language models trained on the same coding region may develop nuanced variations in their outputs. **c,** Three types of single-point mutations are used in this study. **d,** The proposed hybrid model. CaLM operates directly on cDNA sequences to compute codon-level effect scores ($\text{LLR}_{\text{codon}}$) for all possible missense variants relative to wild-type codons. Then, a PLM accordingly (e.g., ESM-2) generates residue-level effect scores ($\text{LLR}_{\text{residue}}$). Our hybrid model integrates effect scores at two levels. This figure was created using BioRender.com.

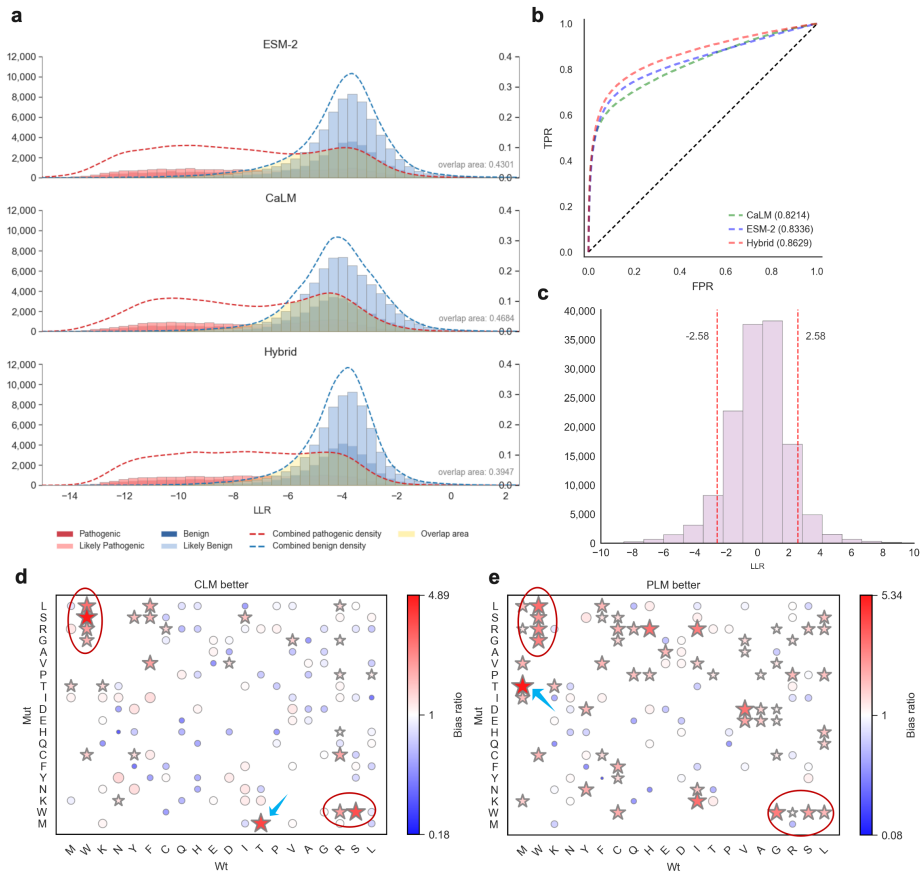


Fig. 2: The hybrid method improves disease effects prediction.

a, The distribution of effect scores for three models: ESM-2, CaLM, and Hybrid. Each panel also shows the density of pathogenic (combined pathogenic and likely pathogenic) and benign (combined benign and likely benign) variants. The overlap areas between pathogenic and benign densities are marked by yellow. **b,** Comparison of three models in distinguishing pathogenic from benign variants via global ROC-AUC scores. **c,** The distribution of effect score differences between two models. Differences exceeding ± 2.58 constitute the reject region, indicating a statistically significant divergence in effect scores between the two models. **d,** The bias ratio heatmap of wild-type (x-axis) to mutated (y-axis) amino acid substitutions observed in *CLM better* mutations. **e,** The bias ratio heatmap of wild-type (x-axis) to mutated (y-axis) amino acid substitutions observed in *PLM better* mutations. Both axes were arranged in ascending order based on codon numbers associated with each amino acid, ranging from the least to the most. Each circle's color and size reflect the degree to which a given substitution is overrepresented (red, larger circles) or underrepresented (blue, smaller circles) compared to the average. Significantly preferred pairs for each model are marked with a star.

References

[1] Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., Rives, A.: Language models enable zero-shot prediction of the effects of mutations on protein function. Advances in neural information processing systems **34**, 29287–29303 (2021)

[2] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., *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), 2016239118 (2021)

[3] Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., Rives, A.: Learning inverse folding from millions of predicted structures. In: International Conference on Machine Learning, pp. 8946–8970 (2022). PMLR

[4] Yang, K.K., Zanichelli, N., Yeh, H.: Masked inverse folding with sequence transfer for protein representation learning. Protein Engineering, Design and Selection **36**, 015 (2023)

[5] Rao, R.M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., Rives, A.: Msa transformer. In: International Conference on Machine Learning, pp. 8844–8856 (2021). PMLR

[6] Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K., Brock, K., Gal, Y., Marks, D.S.: Disease variant prediction with deep generative models of evolutionary data. Nature **599**(7883), 91–95 (2021)

[7] Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., Yuan, F.: Saprot: Protein language modeling with structure-aware vocabulary. bioRxiv, 2023–10 (2023)

[8] Li, M., Tan, Y., Ma, X., Zhong, B., Yu, H., Zhou, Z., Ouyang, W., Zhou, B., Hong, L., Tan, P.: Prosst: Protein language modeling with quantized structure and disentangled attention. bioRxiv, 2024–04 (2024)

[9] Laubitz, Z.: The gain and loss of information during translation (1983)

[10] Outeiral, C., Deane, C.M.: Codon language embeddings provide strong signals for use in protein engineering. Nature Machine Intelligence **6**(2), 170–179 (2024)

[11] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., *et al.*: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023)

[12] Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., *et al.*: Clinvar: public archive of interpretations of clinically relevant variants. Nucleic acids research **44**(D1), 862–868 (2016)

[13] Auer, P.: Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research **3**(Nov), 397–422 (2002)

[14] Hie, B., Zhong, E.D., Berger, B., Bryson, B.: Learning the language of viral evolution and escape. Science **371**(6526), 284–288 (2021)

Supplemental information

Multilingual language model improves zero-shot prediction of disease effects on proteins

Ruyi Chen, Nathan Palpant, Gabriel Foley, and Mikael Boden

Extended Data Tables

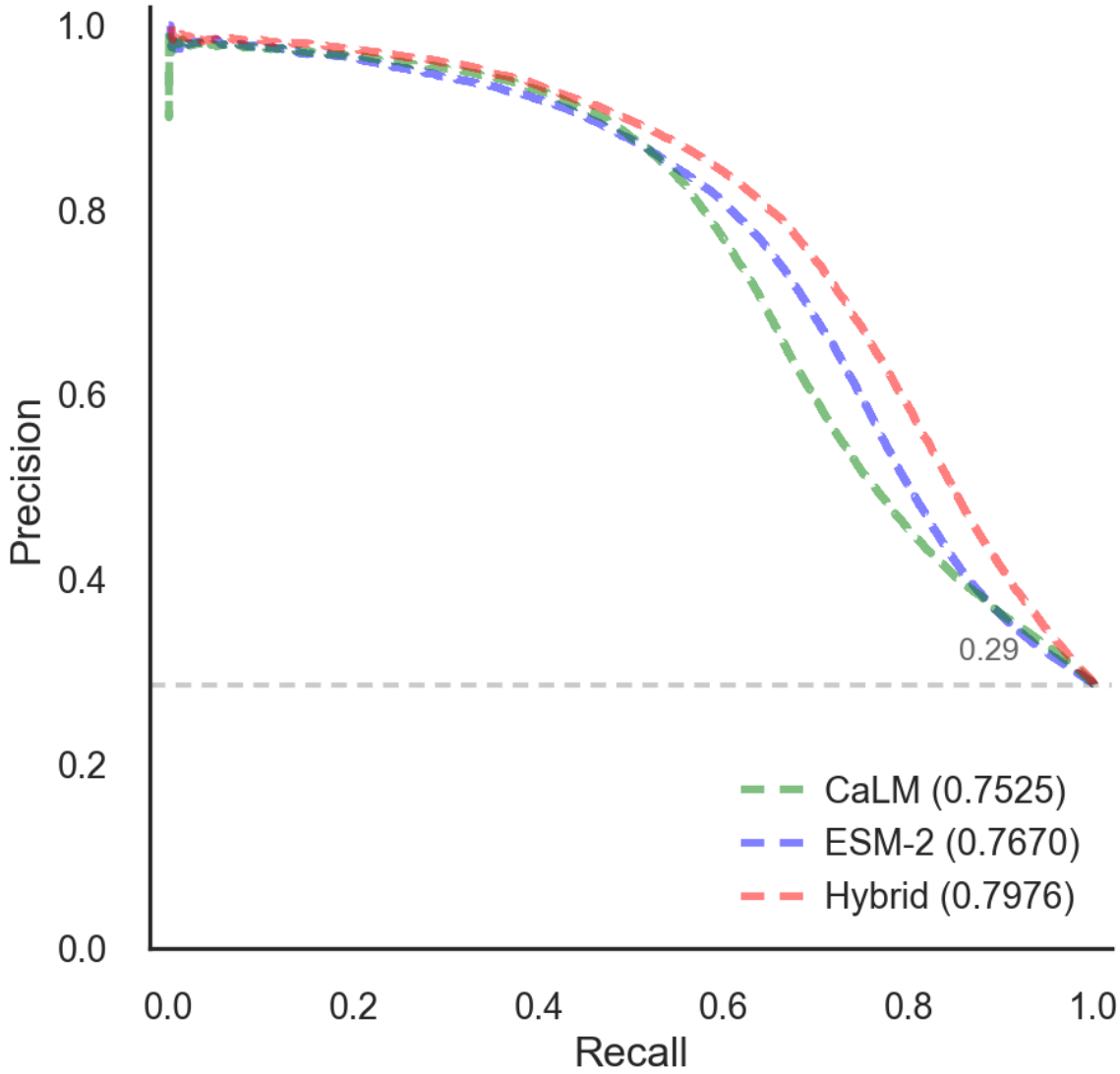
	Missense mutation	**Nonsense mutation**	**Synonymous mutation**	**All**
Pathogenic	14,798	37,113	212	52,123
Likely pathogenic	24,482	8,834	219	33,535
Benign	30,225	134	38,903	69,262
Likely benign	67,845	305	488,245	556,395
Total	137,350	46,386	527,579	711,315
Genes	13,791	3,291	11,593	\

Extended Data Table 1: Three types of single-point mutations collected from ClinVar.

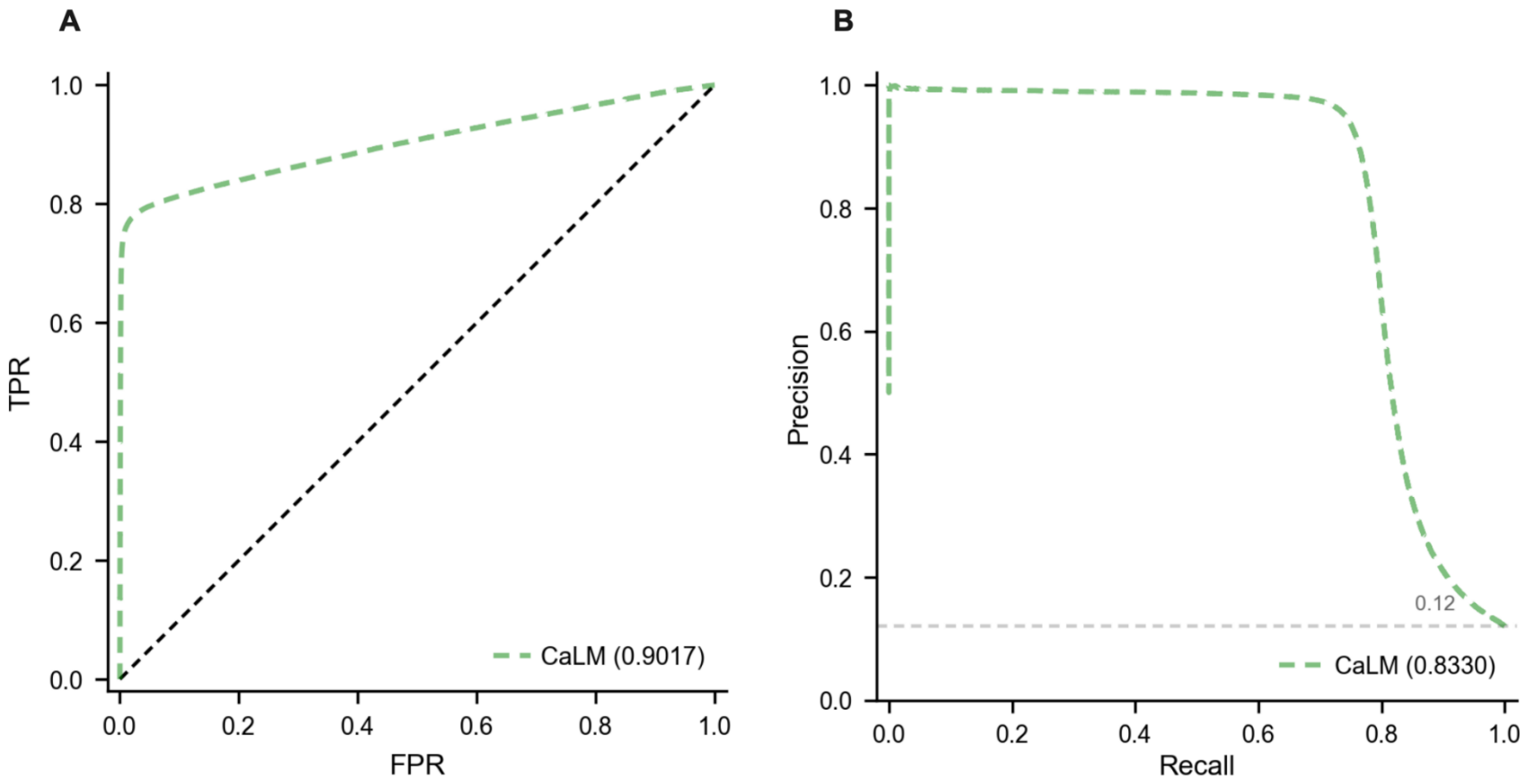
	Formulas	Values	ROC-AUC	AUPRC
1	$LLR = \alpha\, LLR_{c_k} + \beta LLR_{y_k}$	$\alpha = 4.07$, $\beta = 4.15$	0.8629	0.7976
2	$LLR = \alpha\, LLR_{c_k} + \beta {LLR_{y_k}}^3$	$\alpha = 5.00$, $\beta = 0.17$	0.8615	0.7570
3	$LLR = \alpha\, {LLR_{c_k}}^3 + \beta LLR_{y_k}$	$\alpha = 0.21$, $\beta = 4.97$	0.8552	0.7710
4	$LLR = \alpha\, {LLR_{c_k}}^3 + \beta {LLR_{y_k}}^3$	$\alpha = 4.98$, $\beta = 4.90$	0.8677	0.7976
5	$LLR = \alpha\, LLR_{c_k} + \beta LLR_{y_k} + \gamma\, LLR_{c_k} * LLR_{y_k}$	$\alpha = 4.98$, $\beta = 4.85$, $\gamma = 0.26$	0.8644	0.8005

Extended Data Table 2: Different approaches for combining effect scores. Drawing inspiration from Bayesian optimization, we systematically optimized the weighting parameters (α, β, γ) in each combinatorial formula to maximize the ROC-AUC.

Extended Data Figures



Extended Data Fig.1 Precision-Recall curve comparison of three models in missense mutations classifications. The y-axis represents precision, while the x-axis represents recall. The Hybrid model achieves the best overall performance with PR-AUC of 0.7976, followed by ESM-2 (0.7670) and CaLM (0.7525)



Extended Data Fig.2 ROC-AUC and AUPRC plots in single-point mutations classifications by CaLM.