Coupling codon and protein constraints decouples drivers of variant pathogenicity

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

Predicting the functional impact of genetic variants remains a fundamental challenge in genomics. Existing models focus on protein-intrinsic defects yet overlook regulatory constraints embedded within coding sequences. Here, we couple a codon language model (CaLM) with a protein language model (ESM-2) to dissect the drivers of variant pathogenicity. On ClinVar data, both modalities contribute near-equally to distinguishing pathogenic from benign variants. Evaluation across Deep Mutational Scanning and CRISPR-Based Genome Editing platforms in ClinMAVE reveals that loss-of-function variants are governed primarily by residue-level features, whereas gain-of-function variants show a greater relative contribution from codon-level constraints, albeit in a gene-specific manner. A controlled comparison of identical variants in *BRCA1* and *TP53* further suggests that codon-level signals are elevated in the endogenous genomic context. Together, these findings indicate that pathogenicity reflects both the "product" and the "process," and that the experimental platform may influence which dimension is observable.

Keywords: Protein language models, Codon language models, Mutation effects, Bayesian optimisation, Variant interpretation, Codon degeneracy

1 Introduction

Predicting the functional impact of genetic variants remains a fundamental challenge in genomics. To address this, deep learning models have been deployed to extract evolutionary constraints directly from biological sequences via "zero-shot learning" [1–3]. These models can score variants by estimating how much a mutation disrupts the probability of observing a token (e.g., an amino acid) in its native sequence context. This disruption can be quantified as a Log-Likelihood Ratio (LLR) between the mutant and wild-type (wt) state, yielding effect scores that correlate strongly with experimental fitness and clinical pathogenicity [1, 4].

Existing models are broadly categorised into three classes based on their training modalities: protein sequence [1, 5], structure [6, 7], and alignment-by-homology [2, 8]. Although integrating evolutionary and structural signals improves predictive performance, these proteocentric models treat coding sequences primarily as generators of proteins, overlooking the regulatory syntax embedded in the genomic context [9–12]. Notably, Boshar and colleagues demonstrated that a joint genomic/proteomic model can outperform specialised protein-only models on specific tasks [13].

Building on this perspective, we consider the relationship between DNA and protein as analogous to distinct natural languages. For example, English and German may convey largely similar semantics yet differ subtly in interpretation (Fig. 1A) [14]. The flow of genetic information from DNA to mRNA to proteins (where naively "translation" appears to be a matter of table look-up) illustrates how core meanings can transcend languages. We reasoned that large language models trained on sequences, expressed in distinct "languages," such as protein-coding DNA (cDNA) and their corresponding protein sequences, are likely to capture nuances derived from, and specific to the cDNA *or* the protein super-sets, respectively (Fig. 1B). The integration of variant-level information from both cDNA and protein sequences is thus hypothesised to provide *complementary* biological signals, thereby revealing pathogenic mechanisms that cannot be explained by protein defects alone.

Here, we employ a dual-probe framework to identify pathogenic drivers that escape proteocentric models but still shape clinical phenotypes. Specifically, we use a Codon Language Model (CLM; CaLM [15]) and a Protein Language Model (PLM; ESM-2 [3]) to independently calculate the LLRs of mutations at the codon and amino acid levels, and integrate these scores via Bayesian optimisation (Fig. 1F). Validation on the *in vivo* ClinVar (Fig. 1C; [16]) and ClinMAVE datasets spanning both Deep Mutational Scanning (DMS; Fig. 1D) and CRISPR-Based Genome Editing (CBGE; Fig. 1E) [17] reveals that Loss-of-Function (LoF) variants are predominantly governed by residue-level features, whereas Gain-of-Function (GoF) variants exhibit a greater, albeit gene-specific, contribution from codon-level constraints. Direct comparison of identical variants assayed by both DMS and CBGE provides evidence that codon-level signals are elevated in the endogenous genomic context, suggesting that exogenous expression systems may underestimate codon-level constraints relevant to pathogenicity. Together, these results highlight that pathogenicity emerges from both the "product" and the "process," a duality effectively captured by a duality of large language models.

2 Results

2.1 Dual-modality model captures pathogenicity missed by single-modality models

To investigate whether and to what extent distinct biological constraints contribute to variant pathogenicity, we curated a comprehensive dataset of missense variants from ClinVar, comprising 137,350 variants (39,280 pathogenic and 98,070 benign) across 13,791 genes (Fig. 1C and Supplementary Table S1).

We benchmarked CaLM and ESM-2 individually and as a hybrid to quantify the information gained by reconciling the signals in the two modalities. While single-modality models generate LLRs independently, we used Bayesian optimisation to determine weights that balance the influence of each modality's LLR (see "Materials and methods – Missense effect scores"). Despite distinct training data and model architectures, CaLM and ESM-2 displayed compatible LLR profiles, with nearly identical standard deviations, both globally and across comparison groups (Supplementary Table S3). This compatibility suggests that both models capture concordant evolutionary landscapes and can be combined directly as raw LLRs.

We employed 10-fold cross-validation, partitioning variants by gene to control for ascertainment bias. Bayesian optimisation yielded a mean CaLM weight of 0.49 across training folds (Fig. 2A), with fold-specific details provided in Supplementary Fig. S1 (see "Materials and methods - Model training and validation"). This near-even balance suggests that, in aggregate, codons and amino acids information jointly define the pathogenicity landscape.

For each fold, we applied the weight optimised on the training set to generate predictions for the corresponding validation set. Aggregating these predictions yielded composite effect scores for all three models (Figs. 2B and C). Notably, although CaLM effectively differentiates benign from pathogenic variants, its discriminative power is marginally lower than that of the protein-based ESM-2. Integrating both modalities achieved an AUROC of 0.862, significantly outperforming both ESM-2 (0.831) and CaLM (0.822) (Fig. 2B; two-tailed paired t-test, $p < 0.001$). Furthermore, the hybrid model maintained this advantage in the Area Under the Precision-Recall Curve (AUPRC) (Supplementary Fig. S2; two-tailed paired t-test, $p < 0.001$), confirming that the weighted signal is robust to class imbalance and independent of the primary optimisation metric. These performance gain indicate that the biological signals captured by two models are complementary rather than redundant.

2.2 Divergent contributions of codon and protein constraints to functional variant phenotypes

To further investigate the relative contributions of codon- and residue-level constraints to variant function prediction, we applied the dual-modality model to the ClinMAVE dataset across two independent experimental platforms: DMS and CBGE. DMS-derived data largely decouple a variant's biophysical impact from its native genomic context through exogenous expression systems, whereas CBGE preserves the endogenous regulatory environment [17]. Comparing the two platforms thus allows us to

assess both the consistency of modality contributions across independent experimental systems and whether the detectability of codon-level constraints is modulated by genomic context. For each platform, we stratified missense variants into three functional categories: functionally normal, LoF, and GoF (Figs. 1D and E and Supplementary Table S2), and formulated two parallel comparisons: functionally normal vs. LoF, and functionally normal vs. GoF. For each comparison, LLRs were computed using CaLM and ESM-2 independently, followed by gene-stratified 10-fold cross-validation to determine the optimal weight of their hybrid. Results from random-split cross-validation and per-fold optimisation landscapes are provided in Supplementary Figs. S3-S5.

In the "functionally normal vs. LoF" comparison, protein features play the dominant role across both platforms. For DMS, the AUROC peaks at a CaLM weight of 0.14 (Fig. 3A), and this protein-dominant pattern is further reinforced in CBGE, where the optimal weight decreases to 0.05 (Fig. 3E). The corresponding ROC curves confirm that ESM-2 provides superior discriminative power for LoF variants on both platforms, while CaLM-derived scores contribute minimally (Figs. 3B and F). The convergence of the hybrid model with the ESM-2 baseline across both experimental systems indicates that protein features are sufficient to capture LoF pathogenicity (Supplementary Fig. S4).

Conversely, the "functionally normal vs. GoF" comparison reveals a shift towards codon-level signals, although the magnitude of this shift differs between platforms. In DMS, model performance peaks at a CaLM weight of 0.77 (Fig. 3C). However, this estimate is unstable: the DMS GoF subset comprises only 12 genes, and per-fold weights vary from 0.0 to 1.0 (Supplementary Fig. S5B), suggesting that a few genes drive the high weight rather than a generalisable trend. CBGE data (391 genes) yields a more robust estimate, with an optimal CaLM weight of 0.19 (Fig. 3G and Fig. S5D). Although modest in absolute terms, this is nearly four-fold the LoF counterpart (0.05), indicating a greater codon-level contribution to GoF classification than to LoF classification. The ROC curves show that all three models achieve comparable but modest performance for GoF discrimination (Figs. 3D and H), consistent with the inherent difficulty of this variant class.

Taken together, the directional trend of modality contributions is consistent across platforms: LoF is protein-dominated, whereas GoF exhibits a greater codon-level contribution. However, the substantial difference in optimal GoF weights between DMS and CBGE likely reflects a combination of factors, including differences in gene composition and potential platform-specific effects on the detectability of codon-level signals. Disentangling these contributions requires controlled comparisons of identical variants within the same genes across platforms, which we address in Section 2.6.

2.3 Codon-level constraints reflect nonsense and synonymous variant effects

We next evaluated the utility of codon-level information from CaLM in characterising *nonsense* and *synonymous* mutations across ClinVar and both ClinMAVE platforms. Because PLMs operate solely on amino acid sequences, they are inherently unable to

detect nucleotide-level variant effects that do not alter the protein sequence or that result in premature truncation.

As illustrated in Fig. 4, CaLM successfully captured the broad distinctions between mutation categories, assigning significantly lower LLR scores to nonsense variants compared to synonymous ones across all three datasets. However, distinguishing functional impact within these categories proved more challenging. Specifically, the LLR distributions for pathogenic and benign synonymous variants exhibited substantial overlap (Fig. 4A), and the separation between functionally normal and LoF/GoF phenotypes was less distinct in both DMS and CBGE (Fig. 4B). Notably, the distributional patterns were highly consistent across the two platforms, suggesting that CaLM's capacity to distinguish nonsense from synonymous variants reflects a fundamental nucleotide-level constraint that is largely independent of experimental context.

This pattern may be partially driven by the inherent evolutionary frequencies of these variants: because natural selection strictly penalises nonsense mutations while tolerating synonymous ones, the CaLM model learns a strong prior for the latter. This implicitly aligns with the clinical observation that nonsense variants are predominantly pathogenic and synonymous ones benign (Figs. 1C-E). Crucially, although the discriminative signal for synonymous variants specifically is subtle, CaLM captures a unique dimension of nucleotide-level constraint (evident in the nonsense vs. synonymous split) that remains completely invisible to PLMs.

2.4 Divergence between codon and protein language models reveals the impact of codon degeneracy

To investigate the biological drivers of model discordance, we focused on the ClinVar dataset, where the hybrid model significantly outperformed both single-modality baselines, providing sufficient contrast to stratify variants by modality preference (Fig. 2B). We isolated "conflict zones" among ClinVar missense variants where CaLM and ESM-2 predictions were maximally divergent. This yielded 2,748 mutations exhibiting extreme LLR discrepancies, comprising 1,281 variants in the *CLM better* group and 1,467 in the *PLM better* group (see "Materials and methods - Identification of missense variant with divergent model predictions" and Fig. 5A). To characterise the specific substitution biases of each model, we quantified the enrichment of wt-to-mutant substitutions in each subset (e.g., *CLM better*) relative to the background distribution of the full missense dataset and assessed statistical significance using two-tailed Fisher's exact tests (see "Materials and methods - Characterisation of amino acid substitution preferences").

We found that the discordance between CLM and PLM predictions correlates strongly with shifts in codon degeneracy. Significant discrepancies are particularly pronounced in regions where mutations induce a transition between low- and high-codon degeneracy states (e.g., from Tryptophan (W), encoded by one codon, to Serine (S), encoded by six, or vice versa; highlighted in blue boxes in Figs. 5B and C).

This suggests that distinct evolutionary signals are being prioritised: substitutions that preserve degeneracy (i.e., synonymous-like flexibility) tend to yield convergent effect scores, whereas dramatic shifts in degeneracy state expose the orthogonality

of the models. The CLM appears to weight the information loss in the nucleotide landscape, while the PLM focuses on the physicochemical impact of the residue change.

This divergence captures complementary biological constraints rather than simple model error. For instance, the PLM significantly outperforms the CLM in predicting Glycine (G) to Glutamic acid (E) substitutions (as shown by the enrichment in Fig. 5C and depletion in Fig. 5B). In this context, the PLM likely captures the specific structural intolerance (or tolerance) of the residue change, whereas the CLM may be overly sensitive to the reduction in codon state space (from 4 codons to 2). Taken together, these findings indicate that variant pathogenicity constitutes a composite function of residue fitness and codon optimality, modulated by the genetic code's structure.

2.5 Distinct constraints dictate variant pathogenicity across gene classes

Moving beyond individual amino acid substitutions, we asked whether the different predictive strengths of CLM and PLM are discernible at the gene level and associated with specific biological functions.

We computed the absolute difference in AUROC (ΔROC_g) for each gene to quantify the relative discriminative capacity. Based on this metric, we filtered the dataset and identified 29 CaLM-superior genes (*CLM better*) and 66 ESM-2-superior genes (*PLM better*), which were used for subsequent biological characterisation (see "Materials and methods - Gene-level performance evaluation via bootstrapping" and Fig. 5D). Within the *CLM better* subset, *PCYT1A* exhibited the most pronounced performance advantage for CaLM, followed by *DNMT1* and *MEF2C* (Fig. 5E). Notably, many of these genes, including *EYA1* [18], *GATA1* [19], and *EZH2* [20], function as transcriptional regulators or chromatin modifiers. Conversely, the genes favouring ESM-2 were led by *SUMF1* and *KCNT1* (Fig. 5F). This group predominantly encodes components of multiprotein complexes (e.g., *HBA1* [21], *RAG1* [22], and *TP53* [23]), which are properties manifested at the amino acid level, such as structural stability.

To evaluate constraint profiles for these gene subsets, we used two gnomAD metrics: probability of loss-of-function intolerance (pLI) and loss-of-function observed/expected upper bound fraction (LOEUF) [24]. We observed that genes in the *CLM better* subset exhibited significantly higher pLI scores and correspondingly lower LOEUF scores compared to the *PLM better* group, with both comparisons demonstrating a moderate effect size (two-sided Mann-Whitney U test, $p < 0.05$; rank-biserial r = -0.352 for pLI, and $r = 0.300$ for LOEUF; Fig. 5G and Supplementary Fig. S6). The consistently near-maximal pLI scores (≈ 1.0) and highly depleted LOEUF values within the *CLM better* subset indicate a strong biological trend toward haploinsufficiency. Consequently, the proper function of these genes is highly sensitive to precise gene dosage - a quantitative trait heavily regulated by translational kinetics. Unlike the PLM, the CLM uniquely captures codon-level signals that contribute to maintaining critical expression thresholds.

To further clarify the biological basis of this stratification, we performed functional enrichment analysis using Gene Ontology Molecular Function (GO:MF) terms.

The *CLM better* genes were preferentially enriched in transcriptional regulatory functions, particularly transcription coregulator binding and chromatin binding (Fig. 5H). This concordance with near-maximal pLI scores suggests that the CLM captures the sequence constraints underlying gene expression regulation – where maintaining precise protein levels (dosage sensitivity) is as important as preserving protein function. We further stratified *PLM better* genes by pLI scores, yielding 26 high-pLI and 34 low-pLI genes. In the high-pLI subset, PLM-associated enrichment was dominated by complex membrane signalling machinery, including NMDA glutamate receptor and calcium channel activities. In contrast, the low-pLI subset showed enrichment for more localised enzymatic and ligand-binding functions, notably biotin binding (Fig. 5I). These findings underscore the PLM's capacity to model both the complex structural dependencies of membrane proteins and highly specific, localised biochemical interactions.

To concretely illustrate this mechanistic difference in dosage-sensitive genes, we examined a pathogenic variant in *MEF2C* (p.Leu38Pro, c.113T>C) (Fig. 5J). Here, we observed a conflict: while the PLM assigned a moderate score (LLR_{y_k} = -4.77) reflecting structural risk, the CLM flagged a strong signal (LLR_{c_k} = -14.22). This discrepancy suggests that for *MEF2C*, the primary pathogenic driver may not be solely structural destabilisation, but rather a critical failure in translational efficiency required to maintain gene dosage.

Collectively, these results indicate that CLM and PLM capture complementary layers of biological information: the CLM is better suited to detecting disruptions in translational regulation and dosage sensitivity, whereas ESM-2 primarily resolves the biophysical constraints governing protein structure and stability.

2.6 Cross-platform validation reveals context-dependent codon constraints

Our gene-level analysis revealed that *CLM better* genes are strongly enriched for dosage-sensitive functions (high pLI), whereas *PLM better* genes are predominantly structure-constrained. To explore whether the detectability of these codon-level constraints is modulated by experimental context, we performed a controlled cross-platform comparison. Owing to the limited availability of genes with a dense overlap of identical variants assayed by both platforms — at the time of analysis, only *BRCA1* and *TP53* had sufficient variant overlap to support stable 10-fold cross-validation — we selected these two genes as well-characterised archetypes for dosage-sensitive and structure-constrained mechanisms, respectively. By evaluating identical variants within each gene, we isolated the experimental platform as the primary variable, providing an initial test of whether codon-level signal detectability depends on experimental context.

Specifically, for each gene we compiled the set of missense variants with functional annotations from both platforms, classified them as functionally normal or LoF, and applied 10-fold cross-validation coupled with Bayesian optimisation to determine the optimal CaLM weight independently under DMS and CBGE contexts. Platform-level differences in the optimised weights were then assessed using a Wilcoxon signed-rank test on paired fold-level estimates.

For BRCA1, 618 variants were assessed under both platforms, with functional annotations showing 74.8% concordance between DMS and CBGE (Fig. 6A). The optimal CaLM weight increased from 0.02 under DMS to 0.19 under CBGE (Wilcoxon signed-rank test, $p < 0.01$; Fig. 6B), indicating a substantially greater codon-level contribution in the endogenous context. This result aligns with *BRCA1*'s known dosage sensitivity [25], and suggests that exogenous expression systems, by bypassing endogenous regulatory control, may attenuate codon-level constraints that help maintain functional protein levels. In contrast, *TP53* (1,195 variants, 77.7% concordance) exhibited negligible codon-level effects. The CaLM weight was 0.0 across all DMS folds (Fig. 6C) and only marginally increased to 0.01 under CBGE (Wilcoxon signed-rank test, $p < 0.01$; Fig. 6D). Despite reaching statistical significance, this difference is biologically negligible, consistent with the understanding that *TP53* pathogenicity is largely driven by structural disruption [26].

While limited to two genes, these results suggest that exogenous expression systems may systematically attenuate codon-level signals. This points to a potential platform-dependent bias: functional classifications based solely on DMS data may underestimate the pathogenic contribution of codon-level disruptions in haploinsufficient genes.

3 Discussion

Traditional variant predictors often reduce the coding sequence to a mere precursor for amino acids, blind to codon identity. Moving beyond this view, we propose that pathogenic potential emerges from the combination of distinct biological modalities spanning mRNA and protein levels. To interrogate pathogenicity through sequence constraints, we prioritised single-sequence models, specifically ESM-2 and CaLM, over MSA-dependent architectures like AlphaMissense [27]. This circumvents the confounding influence of alignment-based evolutionary bias, enabling us to cleanly isolate and quantify the independent contributions of PLM and CLM. As such, our framework serves not as a competitor to clinical diagnostic tools, but as an analytical approach for dissecting variant pathogenicity.

We leveraged the differences between LoF and GoF to validate the biological fidelity of this framework across two independent experimental platforms. Given that LoF pathogenicity is predominantly mediated by structural disruption that compromises protein stability [28–30], we postulated that codon-level signals would contribute less than biophysical constraints. The model supported this pattern across both DMS and CBGE, with minimal CaLM contribution to LoF classification. For GoF variants, the CaLM weight was consistently elevated relative to LoF across both platforms, although the magnitude was substantially lower in CBGE, likely reflecting the greater gene diversity and more robust estimation afforded by this dataset. This shift, albeit moderate, is consistent with prior observations: while GoF mutations often exert milder effects on protein structure compared to LoF variants [30], they frequently implicate process-dependent constraints, such as dosage regulation [31, 32] or the precise kinetic control required for adopting active conformations [33, 34]. Thus, rather than merely ensembling two scores, the hybrid framework modulates information sources to recapitulate the underlying biological context, suggesting that a more complete pathogenicity assessment may benefit from integrating information from both protein and mRNA levels.

A key finding of this study is that the detectability of codon-level constraints is modulated by experimental context. The controlled *BRCA1*/*TP53* comparison demonstrates that exogenous expression systems may attenuate codon-level signals, with the magnitude of this attenuation correlating with gene dosage sensitivity. This observation, while limited to two genes, suggests that functional classifications based solely on DMS data may underestimate the pathogenic contribution of codon-level disruptions in haploinsufficient genes, a hypothesis that warrants broader validation as additional cross-platform data become available.

Our findings also imply that codon-level constraints remain clinically significant in missense contexts, extending beyond synonymous variation [35, 36]. We conceptualise drastic optimality shifts as "translational shock," a phenomenon in which abrupt changes in translational kinetics reduce protein output [37–39]. Crucially, our pan-cell line analysis dissociates this phenomenon from intrinsic background transcript turnover. We observed no significant difference in mRNA half-lives between *CLM better* and *PLM better* genes across 20 diverse cell types (Supplementary Table S4). This indicates that the effect is not driven by gene instability, but rather by variant-specific

disruptions in translational kinetics that compromise expression efficiency. Mechanistically, this supports a composite fitness model wherein missense mutations compromise fitness through two complementary mechanisms: structural impairment (captured by PLM) and dosage reduction (captured by CLM) (Fig. 7). This model is strongly supported by the enrichment of *CLM better* variants in dosage-sensitive high-pLI genes (Fig. 5G), such as *EZH2* [40] and *MEF2C* [41], and independently corroborated by the BRCA1 cross-platform analysis, where the amplification of codon-level signals in the endogenous context directly links dosage sensitivity to translational constraints.

Synonymous variants have historically been difficult to classify owing to the absence of protein alteration. Our results suggest that pure sequence-based modelling of synonymous effects remains challenging (Fig. 4). This limitation likely stems from dependencies on mRNA secondary structure [42] or splicing regulation [43]. While CaLM may capture local splicing motifs embedded within the coding sequence (e.g., exonic splicing enhancers), accurate prediction likely requires the intronic context or global structural awareness that lies beyond model's scope. Nevertheless, CaLM establishes a foundational baseline where PLMs provide no signal, opens a new avenue for codon-aware prediction in non-coding or silent variant spaces.

Ultimately, the integration of CLM and PLM establishes a proof-of-concept for combining complementary biological signals. Beyond clinical applications, we envision this compositional strategy as a useful template for integrating complementary foundation models to address multi-layered biological questions.

4 Materials and methods

4.1 ClinVar variant dataset

ClinVar [16] is a comprehensive database that aggregates genetic variants associated with disease, together with clinical phenotypes and supporting evidence. For this study, we downloaded the complete set of variant annotations from `variant_summary.txt` (release through October 8, 2024) available via the ClinVar FTP site (https://ftp.ncbi.nlm.nih.gov/pub/clinvar). We restricted the dataset to single-nucleotide substitutions classified as *missense*, *nonsense*, or *synonymous*. To ensure data reliability, the following filtering criteria were applied:

1. restricted to variants annotated with a definitive clinical significance of *benign* (including likely benign) or *pathogenic* (including likely pathogenic);
2. required a ClinVar review status of at least one star;
3. excluded variants located in extremely large genes (e.g., *TTN*); and
4. for genes with annotations across multiple transcripts, only the transcript with the largest number of annotated variants was retained, establishing a robust one-to-one gene-transcript mapping.

The final ClinVar 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 Supplementary Table S1).

4.2 ClinMAVE variant dataset

ClinMAVE [17] is a curated database intended to support the clinical translation of high-throughput functional evidence generated by MAVEs. ClinMAVE implements a multi-layer functional annotation framework that applies reviewed score thresholds to MAVE-derived measurements, assigning each variant to one of three functional categories: functionally normal, LoF, or GoF.

For this study, we compiled a comprehensive dataset of *missense*, *nonsense*, and *synonymous* variants from the ClinMAVE database (accessed March 15, 2026), excluding those mapping to coding sequences longer than 10,000 nucleotides. To facilitate cross-platform comparisons, variants were stratified by their experimental technique into DMS and CBGE subsets. The resulting DMS cohort comprised 223,385 missense, 11,117 nonsense, and 12,980 synonymous variants. In parallel, the CBGE cohort contained 213,775 missense, 13,068 nonsense, and 68,495 synonymous variants. A detailed breakdown of variant and gene counts across functional categories is provided in Supplementary Table S2.

4.3 Sequence collection

For each single-point variant compiled from ClinVar and ClinMAVE, we retrieved the corresponding full GenBank records from the NCBI GenBank database. We then parsed these records to identify coding DNA sequence features and extracted the associated cDNA sequences and their translated protein products.

4.4 CaLM and ESM-2

To ensure a fair comparison across modalities, we employed both CaLM (86 million parameters; downloaded from https://github.com/oxpig/CaLM) [15] and ESM-2 (150 million parameters; downloaded from https://huggingface.co/facebook/esm2_t30_150M_UR50D) [3], two transformer-based language models trained via masked language modelling objectives. The models differ in three key respects:

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 9 million cDNA sequences sourced from the European Nucleotide Archive, whereas ESM-2 leverages 65 million 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

Neither model enforces sequence length constraints and we did not specifically design our approach to handle long sequences.

4.5 Missense effect scores

Language models can estimate the probability of a mutation at a given position by leveraging only the surrounding sequence context, without the need for additional training. The wt cDNA sequence was input into CaLM, while its corresponding wt 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 (e.g., $c_1 = (\mathrm{A}, \mathrm{T}, \mathrm{G})$). 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 $\tilde{x}_i$, altering the wt 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 codon and residue levels from the model's per-position log-likelihoods over all amino acids or codons, including the wt amino acid and wt codon, following the methods in [4]:

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

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

To quantify the relative contribution of each modality rather than obscuring it within a complex non-linear head (e.g., a neural network), we define a missense effect score as a linear combination of the codon- and residue-level LLRs:

$$\mathrm{LLR}_{\mathrm{hybrid}} = w \cdot \mathrm{LLR}_{c_k} + (1 - w) \cdot \mathrm{LLR}_{y_k} \tag{3}$$

where $w \in [0, 1]$ is a weighting parameter that controls the relative contribution of the codon model

4.6 Model training and validation

To evaluate model generalisation, we employed gene-wise 10-fold cross-validation for all classification tasks using `Scikit-Learn` (v1.6.1) [44]. In this strategy, all variants of a given gene are assigned exclusively to either the training or the test set within each fold, preventing data leakage from gene-specific sequence dependencies. This partitioning was applied consistently across the ClinVar "benign vs. pathogenic" task and the ClinMAVE "functionally normal vs. LoF" and "functionally normal vs. GoF" comparisons on both DMS and CBGE platforms. For the controlled cross-platform comparison of *BRCA1* and *TP53* (Section 2.6), where all variants belong to a single gene, we instead used a random split. For completeness, random split results for the other tasks are provided in the Supplementary materials.

Within each cross-validation fold, the weighting parameter w (search range from 0 to 1) was optimised solely on the training data via Bayesian optimisation to maximise the AUROC [45, 46]. The optimised w was subsequently evaluated on the corresponding held-out test fold. This optimisation procedure was implemented using the `BayesianOptimization` function from the `bayesian-optimization` package (v2.0.3), configured with 10 initial exploration steps followed by 20 optimisation iterations.

4.7 Metrics for variant classification

For the ClinVar dataset, *pathogenic* variants were defined as the positive class and *benign* variants as the negative class. In the ClinMAVE dataset, *functionally normal* variants served as the negative class, whereas LoF and GoF variants were assigned to the positive class for their respective classification tasks.

The AUROC 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, AUROC may not fully capture nuanced differences. To address potential insensitivity to class imbalance, we additionally employed the AUPRC metric to assess the precision-recall trade-off. For both ROCs and PRs, we calculated the AUROC and AUPRC for each validation fold, respectively, and assessed the statistical significance of the differences using a two-tailed paired t-test via `scipy.stats` (v1.13.1) [47].

4.8 Identification of missense variants with divergent model predictions

To quantify the modality-specific discrepancy in mutational likelihoods learned by the CLM and PLM, we formulated a differential metric based on LLRs, denoted as ΔL_k, for each missense variant k:

$$\Delta \mathrm{L}_k = \mathrm{LLR}_{y_k} - \mathrm{LLR}_{c_k} \tag{4}$$

We isolated variants exhibiting the most extreme discrepancies by selecting the top and bottom 1% of the ΔL distribution. This selection criterion resulted in a dataset of 2,748 variants (representing the 2% most divergent ΔL), which ensures a robust sample size for the characterisation of specific amino acid substitution preferences. We further stratified these variants into two subsets based on which model's prediction was more consistent with the ClinVar's label:

1. *PLM better* (S_P): Variants where the PLM provided a superior score. Specifically, for pathogenic mutations, the PLM assigned a lower LLR ($\mathrm{LLR}_{y_k} < \mathrm{LLR}_{c_k}$), whereas for benign mutations, it assigned a higher LLR ($\mathrm{LLR}_{y_k} > \mathrm{LLR}_{c_k}$).
2. *CLM better* (S_C): Variants where the CLM provided the superior score, satisfying the inverse conditions (i.e., $\mathrm{LLR}_{c_k} < \mathrm{LLR}_{y_k}$ for pathogenic and $\mathrm{LLR}_{c_k} > \mathrm{LLR}_{y_k}$ for benign variants).

Based on these criteria, we finally identified 1,281 missense variants as *CLM better* (S_C) and 1,467 missense variants as *PLM better* (S_P).

4.9 Characterisation of amino acid substitution preferences

To assess whether the PLM and CLM exhibit distinct preferences for specific types of missense variants, we compared the frequency distributions of wt to mutant amino acid substitutions in each subset (S_P and S_C) against their overall distribution in the complete missense dataset (A).

The frequency of a given amino acid substitution p (e.g., $W \rightarrow S$) in subset S_i ($i \in \{\mathrm{P}, \mathrm{C}\}$) is defined as:

$$f_{S_i}(p) = \frac{n_{S_i}(p)}{N_{S_i}} \tag{5}$$

where $n_{S_i}(p)$ denotes the count of substitution p in subset S_i, and N_{S_i} represents the total count of all missense variants in S_i

Similarly, the background frequency of the same substitution in the complete dataset A, denoted as $f_A(p)$, was calculated using the total counts from A. To quantify the relative enrichment or depletion of specific substitutions, we computed the Bias Ratio (BR). For a substitution p in subset S_i, the BR is defined as:

$$\mathrm{BR}_{S_i}(p) = \frac{f_{S_i}(p)}{f_A(p)} \tag{6}$$

A substitution was classified as *enriched* in subset S_i if $\mathrm{BR}_{S_i}(p) > 1$, indicating a higher prevalence relative to the overall dataset, whereas substitutions with $\mathrm{BR}_{S_i}(p) < 1$ were considered *depleted*.

To assess the statistical significance of these enrichment patterns, we employed two-tailed Fisher's exact tests. This method was chosen for its robustness in handling small sample sizes associated with specific rare amino acid substitutions. For each substitution p and subset S_i, we constructed a 2×2 contingency table comparing the counts in the subset (S_i) against the counts in the remainder of the dataset ($A \setminus S_i$) to ensure independence. Tests were performed using the `fisher_exact` function from `scipy.stats` (v1.13.1). To account for multiple hypothesis testing across all unique substitution types, we applied the Benjamini-Hochberg (BH) procedure using the `multipletests` function from `statsmodels.stats.multitest` (v0.14.6) to control the False Discovery Rate (FDR) at $\alpha = 0.01$. Substitutions with adjusted p-values below this threshold were considered statistically significant.

4.10 Gene-level performance evaluation via bootstrapping

To quantify performance at the gene level, we computed the ROC for CaLM and ESM-2 across genes (g), separately. We restricted the analysis to genes with at least 20 variants and imposed a maximum class imbalance ratio of 5:1 (i.e., the ratio of pathogenic to benign variants, or vice versa, did not exceed 5). We defined the performance difference, $\Delta\mathrm{ROC}_g$, as:

$$\Delta\mathrm{ROC}_g = \mathrm{ROC}_\mathrm{P}(g) - \mathrm{ROC}_\mathrm{C}(g) \tag{7}$$

where $\mathrm{ROC}_\mathrm{P}(g)$ and $\mathrm{ROC}_\mathrm{C}(g)$ denote the ROC scores of the ESM-2 (PLM) and CaLM (CLM) models for gene g, respectively

To rigorously assess the statistical significance of $\Delta\mathrm{ROC}_g$, we employed a stratified bootstrap approach. For each gene, we performed 1,000 iterations of resampling with replacement, stratified by the ClinVar's labels to maintain the original class distribution. A 95% Confidence Interval (CI) for the performance difference was constructed using the percentile method. Genes were categorised as having a significant model preference only if the 95% CI did not overlap with zero. This filtering resulted in the identification of 29 *CLM better* genes and 66 *PLM better* genes.

4.11 Gene-level constraint analysis

Evolutionary constraints were assessed using two complementary metrics retrieved from the gnomAD database (v2.1.1) [24]: pLI and LOEUF. The pLI score quantifies the probability of intolerance to heterozygous LoF variation, serving as a proxy for haploinsufficiency and dosage sensitivity [48]. Specifically, genes are stratified into two distinct categories to reflect their mutational constraint: high pLI (pLI ≥ 0.9), which identifies genes highly sensitive to dosage changes, and low pLI (pLI ≤ 0.1), which comprises genes generally more tolerant of LoF mutations. Conversely, LOEUF provides a continuous and robust measure of mutational constraint, where lower values signify stronger purifying selection against LoF variants [24].

To ensure robustness, we included only genes with available high-confidence constraint metrics, resulting in a finalised dataset of 28 (out of 29) *CLM better* and 64 (out of 66) *PLM better* genes. Within the *PLM better* group, 26 genes were classified as high pLI and 34 as low pLI. Differences in the score distributions between these two cohorts were evaluated using a two-sided Mann-Whitney U test implemented in `scipy.stats` (v1.13.1). Additionally, the magnitude of these differences was quantified by calculating the rank-biserial correlation coefficient (r) as a measure of effect size [49].

4.12 mRNA stability analysis

To rigorously assess whether the performance gain of the CLM is driven by transcript instability, we downloaded mRNA half-life data from the Transcriptome Turnover Database [50]. We used 20 diverse human cell lines, including epithelial, hepatic, immune, and neural progenitor cells, to exclude tissue-specific biases. For each cell line, multiple half-life measurements associated with a single gene (e.g., from technical replicates) were aggregated by calculating the median value to derive a robust gene-level estimate. Cell lines with sufficient coverage (more than 5 genes in both the *CLM better* and *PLM better* groups) were included to ensure statistical validity. We evaluated the difference in stability between *CLM better* and *PLM better* gene groups using a one-sided Mann-Whitney U test implemented in `scipy.stats` (v1.13.1).

4.13 Functional enrichment analysis

Functional enrichment analysis for GO:MF was performed in R (v4.2.3) using the `gprofiler2` package [51]. The `gost` function was used to identify significantly enriched terms for the input gene list against the annotated *Homo sapiens* background. *P*-values were adjusted for multiple testing using the `g_SCS` correction method, with a significance threshold of 0.05. To focus on specific biological mechanisms, broad terms were manually excluded.

4.14 Cross-platform controlled comparison

To disentangle platform-specific effects from gene composition confounds, we selected *BRCA1* and *TP53* as representative genes. For each gene, the same set of missense variants was independently classified as functionally normal or LoF by each platform. The analysis comprised 618 variants for *BRCA1* (74.8% label concordance) and 1,195 variants for *TP53* (77.7% label concordance). To determine the optimal CaLM weight, we employed a 10-fold cross-validation with a random split of variants, coupled with Bayesian optimisation. Finally, the statistical significance of weight differences between platforms was evaluated using a Wilcoxon signed-rank test on paired fold-level estimates via `scipy.stats.wilcoxon` (v1.16.3).

4.15 Data analysis and visualisation

Data analysis and figure preparation were performed in Jupyter Notebook v6.5.5 using Python v3.11.11 and the following libraries: numpy v1.26.4, pandas v2.2.2, scipy.stats v1.13.1, matplotlib v3.10.0, seaborn v0.13.2, and ggplot2 v4.0.1.

4.16 Use of large language models

During the preparation of this manuscript, the authors used Claude 3 Opus (Anthropic) to enhance the clarity of the text.

Acknowledgements

Figure 1 and 7 were created with BioRender (https://www.biorender.com/).
Author contributions: R.C. and M.B. designed the study. R.C. implemented code and methods. M.B., G.F. and N.P. supervised the project. All authors contributed to the writing, reviewed, and approved the final manuscript.

Conflict of interest

None declared.

Funding

This work was supported by University of Queensland Research Training Scholarships to R.C. N.P. and M.B. received support from Medical Research Future Fund (2016033).

Data availability

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

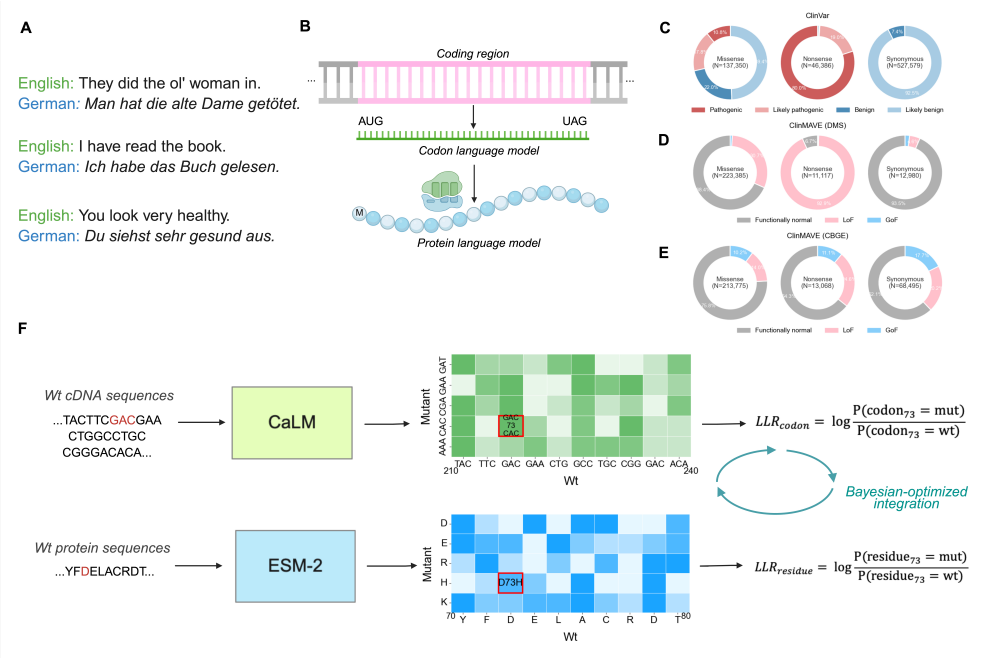

Fig. 1: Linguistic duality reveals orthogonal evolutionary constraints across the central dogma.

(A) Semantic alignment across natural languages. English and German sentences convey identical core meanings yet differ in syntactic structure and expression. **(B)** An analogy between natural languages and biological language models. Just as different languages convey similar meanings with subtle interpretational differences, two language models (i.e., CLM and PLM) trained on the same coding region may develop nuanced variations in their outputs. **(C–E)** Dataset composition and label distribution. The distribution of ground-truth labels for synonymous, missense, and nonsense variants in the (C) ClinVar dataset, (D) ClinMAVE dataset using DMS technique, and (E) ClinMAVE dataset using CBGE technique. **(F)** Schematic of the hybrid model architecture. The framework employs two parallel inference streams: CaLM processes wt cDNA sequences to compute codon-level effect scores (LLR_{c_k}), while ESM-2 independently processes protein sequences to derive residue-level scores (LLR_{y_k}). The two scores are then unified through a Bayesian-optimised integration strategy to predict variant pathogenicity. This figure was created using BioRender.com.

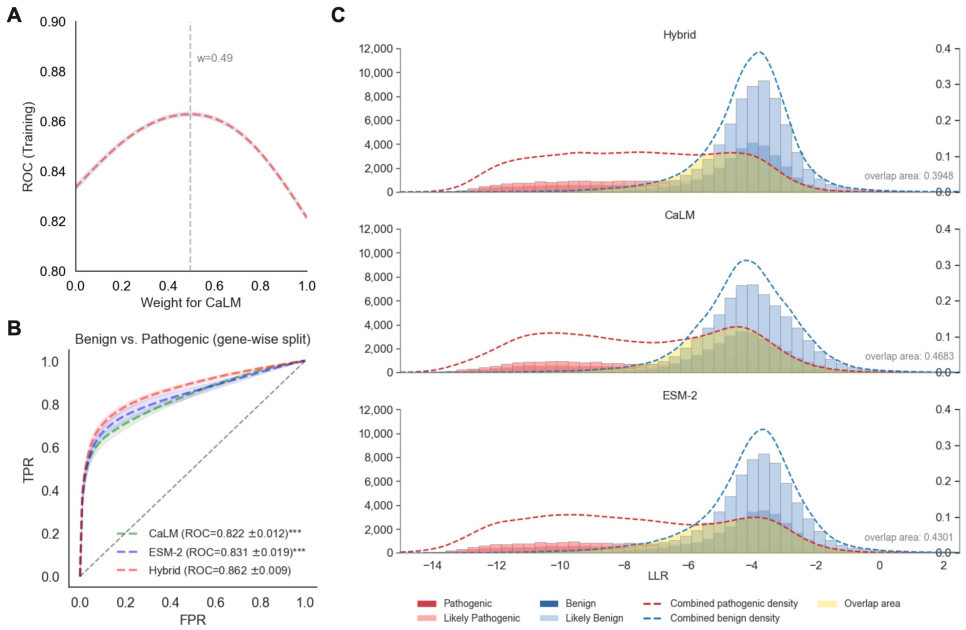

Fig. 2: Codon and protein constraints define a composite landscape of variant pathogenicity.

(A) Hyperparameter tuning for the combination weight (w) of CaLM for the "pathogenic vs. benign" task. The vertical dashed line indicates the optimal weight ($w = 0.49$) determined by maximizing AUROC on the training set using a gene-stratified split. **(B)** ROC curves benchmarking model performance on the held-out test set. Shaded areas represent the standard deviation across 10 folds. The statistical significance of the performance difference between the hybrid model and each single-modality baseline was evaluated using a two-tailed paired t-test ($^{***}p < 0.001$). **(C)** Distributions of LLR for benign (blue shades: benign, likely Benign) and pathogenic (red shades: pathogenic, likely pathogenic) variants in ClinVar, as predicted by ESM-2, CaLM, and the hybrid model. The dashed lines represent the combined density estimates for pathogenic (red) and benign (blue) classes. The yellow shaded region indicates the overlap area between these densities.

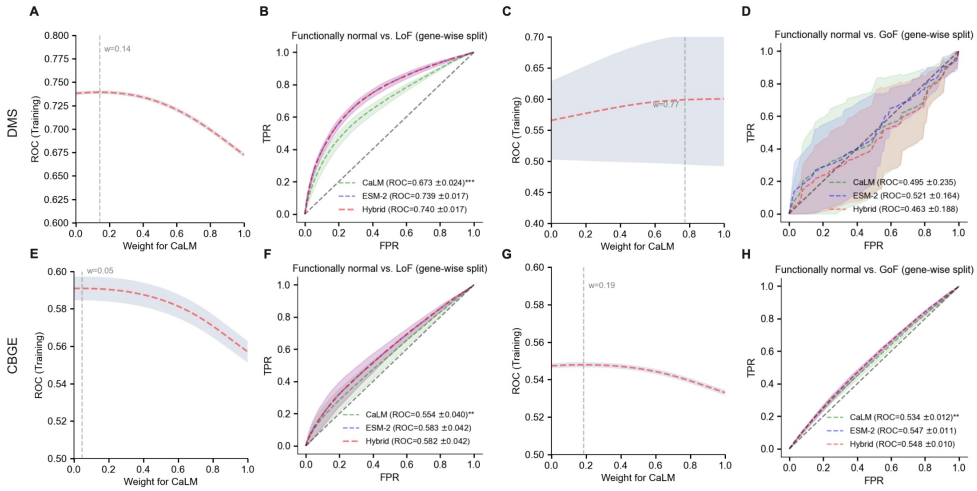

Fig. 3: Decoupling evolutionary drivers of LoF and GoF mechanisms across DMS and CBGE platforms.

(A) Hyperparameter tuning for the weight (w) of CaLM for the "functionally normal vs. LoF" task using DMS data. The vertical dashed line indicates the optimal weight ($w = 0.14$) determined by maximising AUROC on the training set using a gene-stratified split. **(B)** ROC curves benchmarking model performance on the held-out test set for the DMS "functionally normal vs. LoF" comparison. **(C-D)** Same as (A-B), but applied to the DMS "functionally normal vs. GoF" task. **(E)** Hyperparameter tuning for the weight of CaLM for the "functionally normal vs. LoF" task using CBGE data. **(F)** ROC curves for the CBGE "functionally normal vs. LoF" comparison. **(G-H)** Same as (C-D), but applied to the CBGE "functionally normal vs. GoF" task. Shaded areas in ROC curves represent the standard deviation across 10 folds. Statistical significance of the performance difference between CaLM and the hybrid model was evaluated using a two-tailed paired t-test ($^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$).

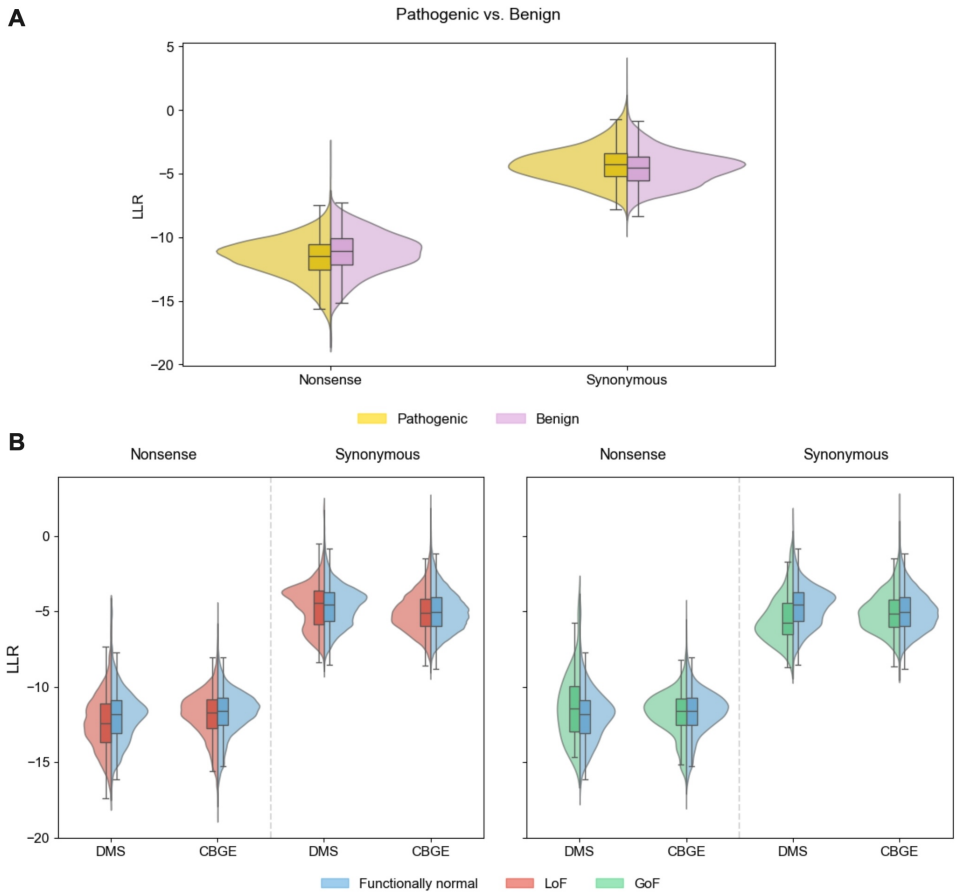

Fig. 4: Codon-level constraints reflect nonsense and synonymous variant effects.

(A) Comparison of LLR distributions between pathogenic and benign variants in the ClinVar dataset, grouped by nonsense and synonymous subtypes. **(B)** Comparison of LLR distributions across functional categories from the ClinMAVE dataset. The left sub-panel contrasts functionally normal variants with LoF variants, while the right sub-panel contrasts functionally normal variants with GoF variants. Within each sub-panel, variants are stratified by sequence consequence (nonsense vs. synonymous) and further grouped by experimental platform (DMS vs. CBGE). Data are presented as split violin plots with internal box plots representing the median and inter-quartile ranges.

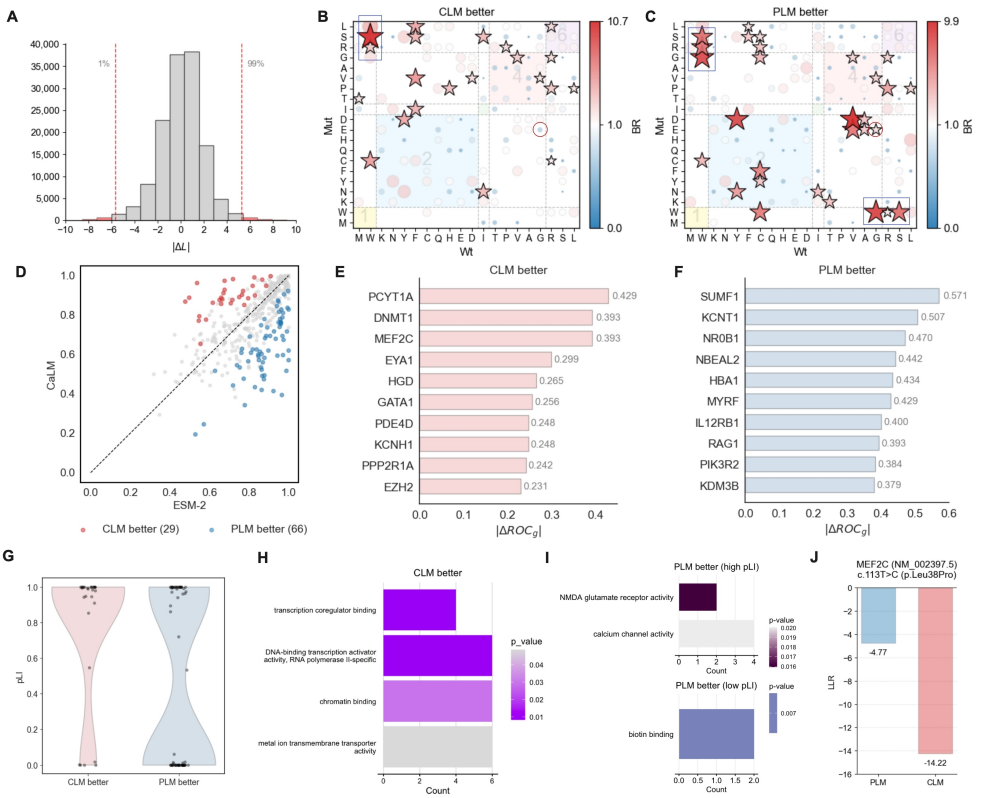

Fig. 5: Dissecting the biological complementarity of CLM and PLM.

(A) Distribution of the difference in LLRs (ΔL) between the two models for ClinVar missense variants. The dashed blue lines mark the 1th and 99th percentiles, isolating variants where one modality captures pathogenic features significantly better than the other. **(B-C)** Heatmaps depicting amino acid substitution biases associated with superior performance of (B) CaLM and (C) ESM-2. Axes represent wt (x-axis) and mutant (y-axis) residues, ordered by increasing codon degeneracy; coloured background blocks along the diagonal distinguish amino acid groups sharing the same codon degeneracy count. Circle size and colour intensity quantify the relative enrichment (red) or depletion (blue) of specific substitutions compared to background distribution of all missense variants. Statistically significant enrichment are denoted by stars using two-tailed Fisher's exact tests. **(D)** Gene-level performance comparison. Scatter plot of AUROC values for individual genes predicted by CaLM (y-axis) versus ESM-2 (x-axis). Genes where CLM performs substantially better are coloured in red, while genes favouring PLM are in blue. **(E-F)** The top 10 genes exhibiting the largest performance gap favouring (E) CLM and (F) PLM. **(G)** Violin plots showing the distribution of pLI scores for genes in the *CLM better* and *PLM better* groups. Black dots represent individual genes. Statistical significance was determined by a two-sided Mann-Whitney U test ($p = 0.007$), with the effect size quantified by the rank-biserial correlation coefficient ($r = -0.352$). **(H-I)** Enriched GO:MF terms for (H) *CLM better* genes and (I) *PLM better* genes. Bar lengths represent the number of genes (count) associated with each term. The colour gradient indicates the significance level (adjusted p-value). For *PLM better* genes, the enrichment results are further stratified into high pLI (top) and low pLI (bottom) gene sets. **(J)** Case study of effect score divergence in *MEF2C* (p.Leu38Pro). Bar chart showing LLRs for that pathogenic variant.

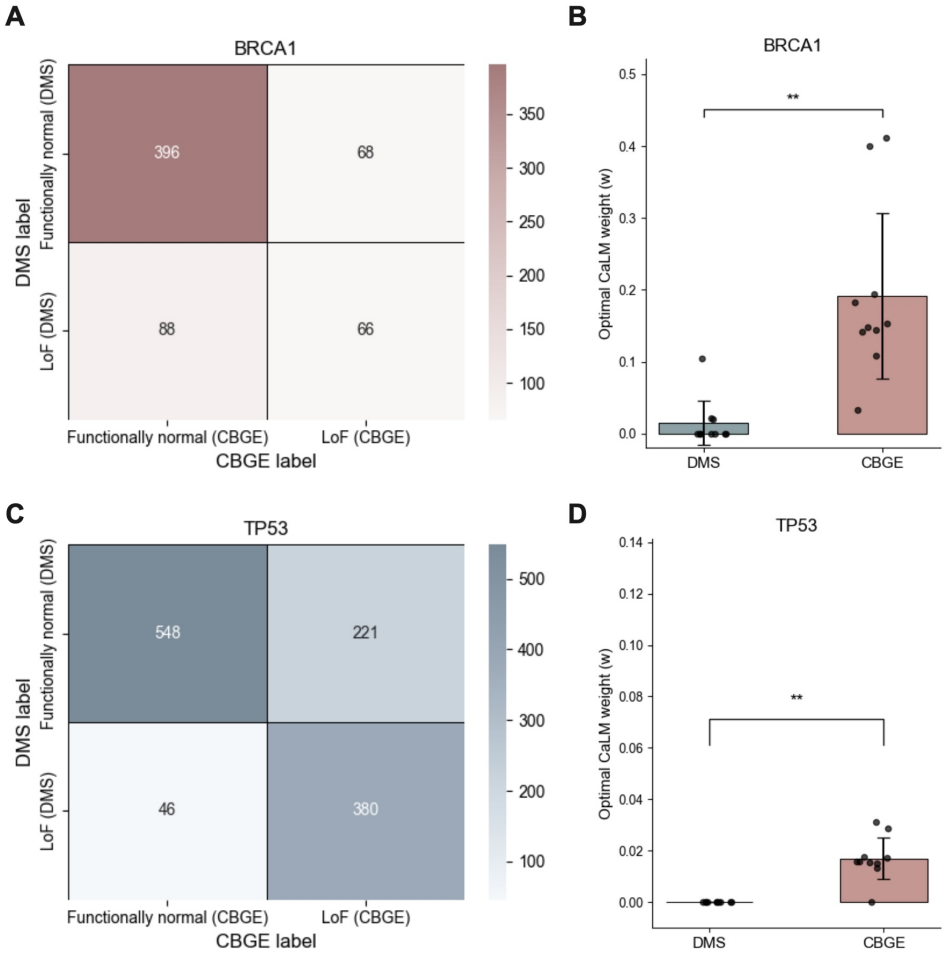

Fig. 6: Cross-platform comparison reveals context-dependent codon constraints in *BRCA1* and *TP53*.

(A) Confusion matrix showing the concordance of functional annotations between DMS and CBGE for *BRCA1*. **(B)** Optimal CaLM weight (w) across 10-fold cross-validation for *BRCA1* under DMS and CBGE. **(C)** Same as (A), but for *TP53*. **(D)** Same as (B), but for *TP53*. Bar heights represent the mean across 10 folds, error bars indicate standard deviation, and individual points represent fold-level estimates. Statistical significance was determined by a Wilcoxon signed-rank test ($^{**}p < 0.01$).

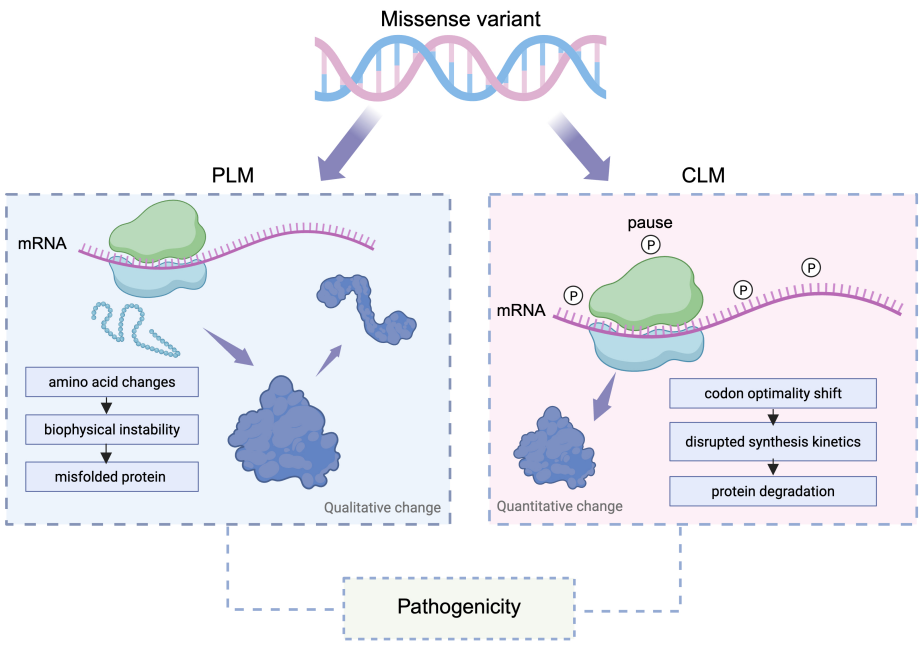

Fig. 7: Schematic of the proposed mechanism.

Missense variants exert deleterious effects via dual mechanisms: (1) structural impairment (left), where amino acid substitutions disrupt protein stability, captured by the PLM; and (2) dosage attenuation (right), where suboptimal codons trigger ribosome stalling and reduce protein expression, modeled by the CLM. This figure was created using BioRender.com.

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] 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)

[3] 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)

[4] Brandes, N., Goldman, G., Wang, C.H., Ye, C.J., Ntranos, V.: Genome-wide prediction of disease variant effects with a deep protein language model. Nature Genetics **55**(9), 1512–1522 (2023)

[5] 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)

[6] 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

[7] 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)

[8] 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

[9] Notin, P., Van Niekerk, L., Kollasch, A.W., Ritter, D., Gal, Y., Marks, D.S.: Trancepteve: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. bioRxiv, 2022–12 (2022)

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

[11] 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)

[12] Zvyagin, M., Brace, A., Hippe, K., Deng, Y., Zhang, B., Bohorquez, C.O., Clyde, A., Kale, B., Perez-Rivera, D., Ma, H., *et al.*: Genslms: Genome-scale language

models reveal sars-cov-2 evolutionary dynamics. The International Journal of High Performance Computing Applications **37**(6), 683–705 (2023)

[13] Boshar, S., Trop, E., Almeida, B.P., Copoiu, L., Pierrot, T.: Are genomic language models all you need? exploring genomic language models on protein downstream tasks. Bioinformatics **40**(9), 529 (2024)

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

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

[16] 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)

[17] Ma, C., Li, Z., Tang, X., Li, P., Li, L., Wang, S., Wu, J., Luo, L., Liu, Y., Zhang, Z., et al.: Clinmave: a curated database for clinical application of data from multiplexed assays of variant effect. Nucleic Acids Research, 1102 (2025)

[18] Xu, P.-X., Adams, J., Peters, H., Brown, M.C., Heaney, S., Maas, R.: Eya1-deficient mice lack ears and kidneys and show abnormal apoptosis of organ primordia. Nature genetics **23**(1), 113–117 (1999)

[19] Crispino, J.D., Horwitz, M.S.: Gata factor mutations in hematologic disease. Blood, The Journal of the American Society of Hematology **129**(15), 2103–2110 (2017)

[20] Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.-M., *et al.*: The polycomb group protein ezh2 directly controls dna methylation. Nature **439**(7078), 871–874 (2006)

[21] Barrera-Reyes, P.K., Tejero, M.E.: Genetic variation influencing hemoglobin levels and risk for anemia across populations. Annals of the New York Academy of Sciences **1450**(1), 32–46 (2019)

[22] Kim, M.-S., Chuenchor, W., Chen, X., Cui, Y., Zhang, X., Zhou, Z.H., Gellert, M., Yang, W.: Cracking the dna code for v (d) j recombination. Molecular cell **70**(2), 358–370 (2018)

[23] Joerger, A.C., Fersht, A.R.: The tumor suppressor p53: from structures to drug discovery. Cold Spring Harbor perspectives in biology **2**(6), 000919 (2010)

[24] Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang,

Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., *et al.*: The mutational constraint spectrum quantified from variation in 141,456 humans. Nature **581**(7809), 434–443 (2020)

[25] Li, C.M.-C., Cordes, A., Oliphant, M.U., Quinn, S.A., Thomas, M., Selfors, L.M., Silvestri, F., Girnius, N., Rinaldi, G., Zoeller, J.J., *et al.*: Brca1 haploinsufficiency promotes early tumor onset and epigenetic alterations in a mouse model of hereditary breast cancer. Nature genetics **56**(12), 2763–2775 (2024)

[26] Joerger, A.C., Fersht, A.R.: The p53 pathway: origins, inactivation in cancer, and emerging therapeutic approaches. Annual review of biochemistry **85**(1), 375–404 (2016)

[27] Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., *et al.*: Accurate proteome-wide missense variant effect prediction with alphamissense. Science **381**(6664), 7492 (2023)

[28] Matreyek, K.A., Starita, L.M., Stephany, J.J., Martin, B., Chiasson, M.A., Gray, V.E., Kircher, M., Khechaduri, A., Dines, J.N., Hause, R.J., *et al.*: Multiplex assessment of protein variant abundance by massively parallel sequencing. Nature genetics **50**(6), 874–882 (2018)

[29] Cagiada, M., Johansson, K.E., Valanciute, A., Nielsen, S.V., Hartmann-Petersen, R., Yang, J.J., Fowler, D.M., Stein, A., Lindorff-Larsen, K.: Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. Molecular Biology and Evolution **38**(8), 3235–3246 (2021)

[30] Gerasimavicius, L., Livesey, B.J., Marsh, J.A.: Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. Nature communications **13**(1), 3895 (2022)

[31] Lampson, B.L., Pershing, N.L., Prinz, J.A., Lacsina, J.R., Marzluff, W.F., Nicchitta, C.V., MacAlpine, D.M., Counter, C.M.: Rare codons regulate kras oncogenesis. Current Biology **23**(1), 70–75 (2013)

[32] Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., Lehner, B.: Synonymous mutations frequently act as driver mutations in human cancers. Cell **156**(6), 1324–1335 (2014)

[33] Kimchi-Sarfaty, C., Oh, J.M., Kim, I.-W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V., Gottesman, M.M.: A" silent" polymorphism in the mdr 1 gene changes substrate specificity. Science **315**(5811), 525–528 (2007)

[34] Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M.V., Komar, A.A.: Synonymous codons direct cotranslational folding toward

different protein conformations. Molecular cell **61**(3), 341–351 (2016)

[35] Bonekamp, F., Dalbøge, H., Christensen, T., Jensen, K.F.: Translation rates of individual codons are not correlated with trna abundances or with frequencies of utilization in escherichia coli. Journal of bacteriology **171**(11), 5812–5816 (1989)

[36] Akashi, H.: Synonymous codon usage in drosophila melanogaster: natural selection and translational accuracy. Genetics **136**(3), 927–935 (1994)

[37] Yu, C.-H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M.S., Liu, Y.: Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. Molecular cell **59**(5), 744–754 (2015)

[38] Liu, Y., Yang, Q., Zhao, F.: Synonymous but not silent: the codon usage code for gene expression and protein folding. Annual review of biochemistry **90**(1), 375–401 (2021)

[39] Walsh, I.M., Bowman, M.A., Soto Santarriaga, I.F., Rodriguez, A., Clark, P.L.: Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. Proceedings of the National Academy of Sciences **117**(7), 3528–3534 (2020)

[40] Suri, T., Dixit, A.: The phenotype of ezh2 haploinsufficiency—1.2-mb deletion at 7q36. 1 in a child with tall stature and intellectual disability. American Journal of Medical Genetics Part A **173**(10), 2731–2735 (2017)

[41] Rocha, H., Sampaio, M., Rocha, R., Fernandes, S., Leão, M.: Mef2c haploinsufficiency syndrome: Report of a new mef2c mutation and review. European journal of medical genetics **59**(9), 478–482 (2016)

[42] Duan, J., Wainwright, M.S., Comeron, J.M., Saitou, N., Sanders, A.R., Gelernter, J., Gejman, P.V.: Synonymous mutations in the human dopamine receptor d2 (drd2) affect mrna stability and synthesis of the receptor. Human molecular genetics **12**(3), 205–216 (2003)

[43] Sauna, Z.E., Kimchi-Sarfaty, C.: Understanding the contribution of synonymous mutations to human disease. Nature Reviews Genetics **12**(10), 683–691 (2011)

[44] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)

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

[46] Hie, B., Zhong, E.D., Berger, B., Bryson, B.: Learning the language of viral

evolution and escape. Science **371**(6526), 284–288 (2021)

[47] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.*: Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods **17**(3), 261–272 (2020)

[48] Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.*: Analysis of protein-coding genetic variation in 60,706 humans. Nature **536**(7616), 285–291 (2016)

[49] Kerby, D.S.: The simple difference formula: An approach to teaching nonparametric correlation. Comprehensive Psychology **3**, 11 (2014)

[50] Jiang, H., Xu, Z., Li, T., Wang, J., Xie, Z.: Ttdb: a comprehensive transcriptome turnover database for exploring mrna stability. Nucleic Acids Research **54**(D1), 350–357 (2026)

[51] Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., Peterson, H.: g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic acids research **51**(W1), 207–212 (2023)