Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Independent Structural Contribution to Mutation Sensitivity in Protein Language Models Revealed by Variance Decomposition Across Complete Missense Landscapes

Tangilal Dihan Chowdhury
Military Institute of Science and Technology
Md Ushama Shafoyat
Military Institute of Science and Technology
Syed Rashedul Haque
Military Institute of Science and Technology
Md. Tobibul Islam
Military Institute of Science and Technology
Maruf Hasan
Military Institute of Science and Technology
Mohammad Shahjahan Majib
Military Institute of Science and Technology
Md Mahbubur Rahman
Military Institute of Science and Technology
Kazy Noor e Alam Siddiquee
Military Institute of Science and Technology
Kaiissar Mannoor
Military Institute of Science and Technology
Md. Shabiul Islam
shabiul.islam@mmu.edu.my

Multimedia University

Article

Keywords: Protein language models, Missense mutation, Structural constraint, Relative solvent accessibility, Evolutionary modeling

Posted Date: May 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9299039/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Interpreting missense mutations requires separating structural constraints from evolutionary and substitution-driven effects across complete substitution space. Here, we constructed a full missense landscape for human hemoglobin subunit β (HBB) by scoring all 2,774 possible single–amino-acid substitutions using pretrained protein language models (PLMs). Mutation sensitivity, defined as the change in sequence log-likelihood, showed strong structural dependence, with substitutions at buried residues significantly more disruptive than at solvent-exposed positions (median 5.19 vs. 3.72; $p = 4.48 \times 10^{-6}$). PLM scores partially aligned with evolutionary constraints (entropy $\rho = 0.245$; BLOSUM62 $\rho = 0.50$) but showed incomplete overlap with deep evolutionary modeling (EVE; $\rho = -0.694$; $R^2 = 0.459$). Multivariable regression revealed that structural depth (RSA) contributes independent variance (~ 5.1%) beyond substitution effects (adjusted $R^2 = 0.301$), with a significant interaction indicating modulation of substitution tolerance by structural context. This structural component was consistent across models (ESM-1b vs. ESM-2; $\rho = 0.78$) and generalizable across 40 proteins. Layer-wise analysis showed strongest structural encoding in early transformer layers. These findings demonstrate that PLM mutation sensitivity includes a separable structural component beyond evolutionary constraint, reflecting burial-dependent biophysical effects.

Introduction

Interpreting the functional consequences of missense mutations remains a central challenge in human genetics, particularly in clinical variant classification and mechanistic inference [1–3]. Although large-scale sequencing efforts and standardized interpretation frameworks have improved variant assessment [4, 5], most analyses are restricted to clinically observed variants, representing a biased subset of the full substitution space. Systematic evaluation of all possible single–amino-acid substitutions, conceptually aligned with deep mutational scanning and fitness landscape mapping, provides an unbiased framework for quantifying mutational constraint [6–11]. Such complete landscapes mitigate ascertainment bias in variant databases [6–8] and enable mutation-level analysis across full substitution space.

Hemoglobin β (HBB) provides a well-characterized model system for such analysis [12, 13]. Missense mutations in HBB produce diverse phenotypic outcomes, including loss-of-function mutations underlying β-thalassemia [12, 13] and gain-of-function mutations such as the canonical sickle-cell variant (E6V) [14, 15]. The coexistence of deleterious, gain-of-function, and benign variants within a structurally resolved protein [16] enables quantitative investigation of how mutation sensitivity relates to structural context.

Protein language models (PLMs), trained on large-scale sequence data, have demonstrated the ability to capture structural and functional constraints without explicit three-dimensional supervision [17–19]. Transformer-based architectures such as ESM-1b encode contextual amino acid representations reflecting evolutionary regularities [17, 20] and show strong performance in zero-shot variant effect prediction [18, 21]. Prior studies have reported correlations between PLM-derived mutation scores and

experimental fitness measurements [18, 23], suggesting that PLMs implicitly encode biologically relevant constraints.

However, a critical unresolved question is whether PLM-derived mutation sensitivity reflects structural constraints independently of evolutionary and substitution-based effects. Existing studies primarily demonstrate correlations with structure or function but do not disentangle whether these signals are reducible to classical evolutionary substitution patterns—captured by matrices such as BLOSUM or PAM [26, 27] or to generative evolutionary models such as EVmutation and EVE [24, 25]. As a result, the extent to which PLMs encode a separable structural component beyond evolutionary constraint remains unclear.

The central question is therefore not whether PLMs correlate with evolutionary constraint, but whether structural depth quantified by burial explains mutation-level variance independently of substitution chemistry and conservation. Addressing this requires analysis across complete substitution space and explicit variance decomposition.

Here, we construct the complete HBB missense landscape using ESM-1b [17, 20] and perform mutation-level variance decomposition to quantify the independent contribution of structural depth, measured by relative solvent accessibility (RSA), beyond substitution metrics [26, 27] and evolutionary constraint models [24, 25]. We further replicate analyses using ESM-2 [22] to assess robustness, perform layer-resolved analysis to examine how structural encoding is distributed across transformer depth, and extend the analysis to 39 additional proteins to evaluate generalizability. By separating structural, evolutionary, and substitution contributions, this study establishes whether PLM-derived mutation sensitivity contains a statistically independent structural component beyond classical evolutionary constraints.

Results

Global Landscape of HBB Mutation Sensitivity

To characterize mutation sensitivity across the HBB sequence, average Δlog-likelihood scores were computed per residue. As shown in (Fig. 1A**)**, Mutation sensitivity was highly heterogeneous along the sequence, with several pronounced troughs indicating mutation-intolerant regions. While most residues exhibited moderately negative average scores, a subset of positions displayed markedly stronger intolerance, suggesting localized structural or functional constraints. The distribution of mutation effects across all 2,774 possible single amino acid substitutions revealed a broad dynamic range, spanning from mildly tolerated to strongly deleterious substitutions (Fig. 1B**)**. Disease-associated variants were generally skewed toward more disruptive regions of the landscape.

Disease Variant Distribution Along the Disruptiveness Spectrum

Mapping clinically annotated variants onto the complete mutational landscape revealed a continuous severity gradient rather than strict categorical separation (Fig. 1A, B**)**. β-thalassemia-associated mutations were generally skewed toward more disruptive Δlog-likelihood values and lower ranks (more disruptive) relative to benign variants, although substantial overlap between groups was observed (Fig. 1C**)**. Benign variants predominantly occupied intermediate regions of the distribution. The sickle-cell mutation (E6V) exhibited moderate disruptiveness, consistent with its gain-of-function phenotype that preserves global folding while altering surface interactions and promoting hemoglobin polymerization under deoxygenated conditions [45]. These results indicate that PLM-derived mutation sensitivity reflects relative severity within HBB mutation space, rather than enabling binary discrimination between disease and benign variants. The python code of Fig. 1–3 formation is provided in **SI 9 (Supplementary Information 9)**.

Identification and Structural Characterization of Mutation-Intolerant Hotspots

To identify structurally constrained regions, residues were ranked based on average mutation sensitivity across all substitutions. The 15 most intolerant residues (approximately top 10%) were classified as mutation-intolerant hotspots (**Supplementary Table 1**). These positions exhibited strongly negative average Δlog-likelihood values, indicating pronounced intolerance to amino acid substitution.

Mapping these hotspots onto the three-dimensional structure of HBB revealed clear spatial clustering within the protein core and in proximity to the heme-binding pocket (**Supplementary Fig. 1**). Structural characterization showed that hotspot residues are associated with elevated local packing density (higher neighbor counts) and, in several cases, reduced distance to the heme group, indicating functional and structural constraint.

Consistent with these observations, the structural feature heatmap (**Supplementary Fig. 3**) demonstrated that mutation-intolerant residues exhibit increased buriedness and proximity to the heme moiety relative to the protein-wide distribution. Together, these results indicate that mutation intolerance inferred from PLM-derived likelihood shifts is strongly associated with structural packing and functional regions, supporting a link between mutation sensitivity and three-dimensional protein architecture. This spatial clustering demonstrates that PLM-derived mutation sensitivity reflects organized structural constraints rather than isolated residue-level effects.

Core–Surface Separation Reveals Global Structural Encoding

To determine whether PLM-derived mutation sensitivity reflects global structural organization, residues were classified as buried (core) or solvent-exposed (surface) based on relative solvent accessibility (RSA) computed from the HBB crystal structure (PDB 1A3N, chain B). Mutation disruptiveness across all 2,774 substitutions was compared between these structural classes.

Substitutions at core residues exhibited significantly greater disruptiveness than those at surface residues (median ΔLL: 5.19 vs. 3.72; Mann–Whitney U test, $p = 4.48 \times 10^{-6}$; U = 3785.0) (Fig. 2; **Supplementary Table 1**). This separation across the complete substitution space demonstrates that burial-dependent structural depth systematically influences PLM likelihood shifts.

Importantly, this effect cannot be explained solely by amino acid substitution preferences, as it emerges across all substitutions and reflects global organization of the protein fold. Buried residues, which contribute to hydrophobic packing and structural stability, are substantially less permissive to variation than solvent-exposed residues [46, 47]. The emergence of this pattern from unsupervised sequence modeling without explicit structural supervision indicates that PLMs encode biophysically meaningful structural constraints at the mutation level. This demonstrates that PLM likelihood shifts encode global structural organization that cannot be reduced to substitution preferences alone.

Layer-Resolved Analysis Reveals Hierarchical Structural Encoding

To investigate how structural information is represented across model depth, layer-wise mutation sensitivity was analyzed using ESM-1b representations. Spearman correlation between RSA and mutation sensitivity was computed for each transformer layer (Table 1; Fig. 3).

Table 1
Statistical Associations Between PLM-Derived Mutation Sensitivity, Structural Depth, and Evolutionary Constraint Metrics.

Comparison	Metric	Numeric Result	Statistical Significance (P)
Residual entropy vs PLM mutation sensitivity	Spearman correlation	$\rho = 0.17$	1.09×10^{-2}
Core vs surface residues	Mann–Whitney U test	Median(core) = 5.19; Median(surface) = 3.72	4.48×10^{-6}
PLM vs BLOSUM62	Spearman correlation	$\rho = 0.50$	4.30×10^{-197}
Evolutionary Entropy vs PLM mutation sensitivity	Spearman correlation	$\rho = 0.245$	3×10^{-3}

Early layers (1–10) exhibited strong negative correlation ($\rho \approx -0.57$ to -0.59; $p < 10^{-14}$), indicating pronounced mutation intolerance at buried residues. This relationship progressively weakened across intermediate layers (15–25; $\rho \approx -0.53$ to -0.46) and approached zero in deeper layers (29–32). Notably, the final layer (33) showed a positive correlation ($\rho = 0.358$; $p = 4.48 \times 10^{-6}$), indicating inversion of the burial-dependent signal.

This progressive attenuation and inversion demonstrate that structural information is encoded strongly in early transformer representations and subsequently transformed across depth. Rather than being monotonically accumulated, burial-dependent structural signals are redistributed as representations

become increasingly optimized for sequence prediction. These findings indicate that PLM architectures encode structural constraints hierarchically, with early layers capturing local packing effects and deeper layers reflecting higher-order sequence context. This inversion at the final layer indicates that structural information is not merely preserved but transformed into representations optimized for sequence prediction objectives. The analysis python script is provided in **SI 11** (**Supplementary Information 11**).

Structural Distribution of β-Thalassemia Mutations

To assess how structural constraints manifest in disease-associated variation, β-thalassemia mutations (n = 20) were stratified into structural categories: core, heme-proximal, and surface. The majority of mutations (17/20; 85.0%) localized to core or heme-proximal regions, indicating enrichment in structurally constrained environments.

Median mutation disruptiveness showed moderate variation across categories (core: −4.50; heme-proximal: −3.68; surface: −4.84; **Supplementary Fig. 2**), although differences were not statistically significant due to limited sample size. In contrast, benign variants (n = 15) were more broadly distributed, with reduced concentration in structurally constrained regions.

Although not intended as a primary statistical test, this analysis provides illustrative support for the structural interpretation of mutation sensitivity, consistent with established loss-of-function mechanisms underlying β-thalassemia [12, 51]. This analysis serves as a qualitative illustration linking PLM-derived mutation sensitivity to clinically observed structural constraint rather than a primary statistical inference.

Partial Alignment with Evolutionary Constraint

To evaluate the relationship between PLM-derived mutation sensitivity and evolutionary constraint, correlations were computed with multiple metrics (**Table 3**).

PLM-derived residue entropy showed a modest positive correlation with mutation sensitivity ($\rho = 0.17$; $p = 1.09 \times 10^{-2}$), reflecting internal model uncertainty. In contrast, evolutionary entropy derived from multiple sequence alignment exhibited a stronger association ($\rho = 0.245$; $p = 3 \times 10^{-3}$), indicating partial alignment with evolutionary conservation [49].

Mutation sensitivity also showed moderate correlation with BLOSUM62 substitution scores ($\rho = 0.50$; $p = 4.30 \times 10^{-197}$), demonstrating agreement with classical substitution preferences [22]. However, the incomplete correlations across all metrics indicate that PLM-derived mutation sensitivity cannot be fully explained by substitution matrices or conservation alone, demonstrating the presence of additional constraints captured by the model [20].

Variance Decomposition Reveals Independent Structural Contribution

To quantify the relative contributions of structural and substitution-based factors, mutation-level multivariable regression was performed (ΔLL ~ RSA + BLOSUM62 + RSA × BLOSUM62) (**Supplementary Table 3**; Fig. 4).

The full model explained 30.6% of variance (adjusted $R^2 = 0.306$; $p < 10^{-218}$), with BLOSUM62 as the dominant predictor (24.0% unique variance). Crucially, RSA contributed a statistically independent component (5.1% unique variance; $\beta = 0.226$; $p < 0.001$), demonstrating that structural depth explains mutation-level variation beyond substitution chemistry.

The significant interaction term (RSA × BLOSUM62; $\beta = -0.076$; $p < 0.001$; 0.6% unique variance) further indicates that structural context modulates substitution effects. Stratified regression slopes (Fig. 4b) showed stronger substitution effects at solvent-exposed positions ($\beta \approx 0.59$) compared to buried residues ($\beta \approx 0.39$), consistent with attenuation of substitution tolerance in structurally constrained regions.

Variance inflation factors (~ 1.0) confirmed negligible multicollinearity. Together, these results establish that PLM-derived mutation sensitivity is not fully reducible to substitution-based or evolutionary constraints, but includes an independent structural component reflecting burial-dependent effects.

This establishes that structural depth contributes non-redundant information to PLM mutation sensitivity, indicating integration of biophysical constraints beyond sequence-derived substitution patterns.

Comparison with Deep Evolutionary Modeling Reveals Non-Redundant Signal

To further evaluate the relationship between PLM-derived mutation sensitivity and evolutionary models, scores were compared with predictions from EVE.

PLM mutation sensitivity exhibited strong inverse association with EVE scores (Spearman $\rho = -0.694$; Pearson $r = -0.677$; $p < 10^{-300}$), with EVE explaining 45.9% of variance ($R^2 = 0.459$; Fig. 5). While this indicates substantial overlap, more than half of the variance remains unexplained.

This residual variance demonstrates that PLMs encode constraints not captured by evolutionary generative models, supporting the presence of non-redundant structural or biophysical information. Statistical analyses were conducted using Python (SciPy and stats models) **SI 12 (Supplementary Information 12).**

Cross-model consistency of structural encoding

As shown in (**Supplementary Table 4)**, Mutation sensitivity scores derived from ESM-1b and ESM-2 were strongly correlated (Spearman $\rho = 0.78$), indicating substantial preservation of mutational ranking across model generations. Both models demonstrated significant separation between core and surface

residues (ESM-1b: $p = 4.48 \times 10^{-6}$; ESM-2: $p = 7.66 \times 10^{-7}$), consistent with burial-dependent mutation sensitivity. Correlation between relative solvent accessibility (RSA) and mutation sensitivity was moderate and increased in ESM-2 ($\rho = 0.478$) compared to ESM-1b ($\rho = 0.35$). Mutation sensitivity also aligned with substitution chemistry, as reflected by correlations with BLOSUM62 scores (ESM-1b: $\rho = 0.50$; ESM-2: $\rho = 0.5264$). Multivariable regression analysis showed comparable explanatory power across models (adjusted R^2: 0.301 vs. 0.3126). RSA retained independent explanatory contribution in both models (5.1% vs. 6.86% unique variance). This consistency demonstrates that burial-dependent structural encoding is a stable property of PLMs rather than an artifact of a specific model architecture. The Python script used for (ESM − 2) PLM result curation and dataset are provided in **SI 10 (Supplementary Information 10)**.

Cross-Protein Analysis Demonstrates Generality of Structural Encoding

To assess generalizability, the analysis was extended to 39 additional proteins (total n = 40; **Supplementary Table 5**; Table 2).

Table 2
Cross-Protein Structural Encoding Summary (n = 40 proteins)

Metric	Value
Total proteins analyzed	40
Proteins with $p < 0.05$ (core vs surface)	37 / 40 (92.5%)
Proteins significant after FDR correction ($q < 0.05$)	37 / 40 (92.5%)
Median RSA correlation (ρ)	0.435
Mean RSA correlation (ρ)	0.442
Interquartile range (IQR) of RSA ρ	0.14
Median core vs surface p-value	1.08×10^{-7}

Across proteins, core residues exhibited significantly greater mutation sensitivity than surface residues in 37/40 cases (92.5%). RSA correlations ranged from $\rho = 0.22–0.64$ (median ≈ 0.42), indicating consistent structural dependence across diverse folds and sizes.

While evolutionary entropy correlations were variable, BLOSUM62 correlations remained moderately consistent ($\rho = 0.11–0.56$), indicating stable substitution effects. Importantly, burial-dependent mutation intolerance was reproducible across proteins, establishing that structural depth represents a general and reproducible property of protein language models rather than a protein-specific phenomenon.

Conceptual Framework of Structural and Evolutionary Contributions

A conceptual framework summarizing the main findings is shown in Fig. 6. PLM-derived mutation sensitivity, computed across the complete missense landscape, exhibits clear structural dependence, including core–surface separation and hotspot clustering. While mutation sensitivity partially aligns with evolutionary and substitution-based metrics, variance decomposition demonstrates that structural depth contributes a statistically independent component (~ 5–7%) beyond these effects. Validation across model architectures and proteins confirms that this structural contribution is reproducible and generalizable, indicating that PLM mutation sensitivity integrates both evolutionary and burial-dependent structural constraints.

Discussion

This study demonstrates that PLM-derived mutation sensitivity contains a statistically independent structural component associated with burial that is not reducible to substitution chemistry or evolutionary constraint. By analyzing the complete HBB substitution landscape, we directly quantified mutation-level variance contributions rather than relying on curated variant subsets.

Across all 2,774 substitutions, buried residues exhibit significantly greater mutation intolerance than surface residues ($p = 4.48 \times 10^{-6}$), consistent with known structural constraints governing protein stability [50, 51]. Importantly, this effect emerges from unsupervised sequence modeling without explicit structural supervision, indicating that PLMs internalize biophysical constraints from sequence statistics alone [17, 22, 53].

Comparison with evolutionary metrics clarifies the nature of this signal. While PLM mutation sensitivity shows moderate alignment with BLOSUM62 substitution scores [26] and modest correlation with evolutionary entropy [46], a substantial fraction of variance remains unexplained. The strong but incomplete agreement with deep generative evolutionary modeling (EVE; $R^2 = 0.459$) further demonstrates that PLM-derived mutation sensitivity cannot be fully reduced to evolutionary constraint.

Critically, multivariable regression establishes that structural depth (RSA) contributes statistically independent explanatory power beyond substitution chemistry. The presence of a significant RSA × BLOSUM interaction indicates that structural context modulates substitution effects, rather than acting purely additively. Although the unique variance attributable to RSA is moderate in magnitude (5–7%), such effect sizes are expected in systems where evolutionary constraint dominates. The key result is not magnitude alone, but statistical independence, demonstrating that structural information constitutes a distinct component of PLM mutation sensitivity rather than a byproduct of substitution frequency bias.

This structural contribution is reproducible across model generations and protein contexts. Replication using ESM-2 confirms preservation and modest strengthening of the structural signal, while analysis across 39 additional proteins demonstrates that burial-dependent mutation intolerance is a general property of PLMs rather than an HBB-specific observation.

Layer-resolved analysis further reveals that structural encoding is hierarchically distributed, with strong burial-dependent signals in early transformer layers that attenuate and invert toward deeper layers. This pattern indicates that structural information is encoded early and subsequently transformed as representations become increasingly optimized for sequence prediction objectives. Structural constraints are therefore not uniformly accumulated across layers but dynamically redistributed.

Together, these findings support a model in which PLMs integrate multiple sources of constraint, including evolutionary statistics, substitution preferences, and biophysical structural information. The incomplete overlap with evolutionary models indicates that PLMs capture aspects of protein behavior not fully represented in sequence-derived evolutionary frameworks.

Several limitations should be noted. The independently quantifiable structural contribution is modest relative to substitution chemistry, and experimental deep mutational scanning data were not directly incorporated. Future integration with empirical fitness landscapes will be important to further validate the biological interpretation of the structural signal.

Overall, this study establishes that mutation sensitivity in protein language models includes a separable, statistically independent structural component reflecting burial-dependent biophysical constraints. This quantitative decomposition provides a framework for interpreting PLM-derived variant effects beyond purely evolutionary models.

Methods

Dataset Collection

The human hemoglobin β (HBB) protein sequence, comprising 146 amino acids, was used as the reference sequence and obtained from RCSB (PDB ID 1A3E) [28]. To construct a complete missense mutational landscape, all possible single amino-acid substitutions were generated at each residue position, excluding synonymous and non-amino-acid-altering variants. This resulted in a total of 2,774 unique missense mutations, representing every possible amino-acid substitution across the HBB sequence. Curated sets of disease-associated missense variants were obtained from the HbVar database for β-thalassemia (19) and included the canonical sickle-cell mutation (E6V) [29]. Benign missense variants (16) were compiled from ClinVar, retaining only variants annotated as Benign or Likely Benign [30]. The dataset of 2774 synthetic and disease mutations are provided in **SI 1 (Supplementary Information 1)**.

Protein Language Model-Based Mutation Scoring

Mutation sensitivity was quantified using a pretrained ESM1b, a 650-million-parameter protein language model (PLM) by computing the change in sequence log-likelihood (Δ log-likelihood) for each missense substitution relative to the wild-type residue [17, 31, 32]. More negative values indicate greater

disruption. All 2,774 possible HBB missense mutations were ranked based on their Δ log-likelihood values. Disease-associated and benign variants were used only to label subsets of this complete mutation space for interpretation and validation. PLM scores were computed independently of any variant annotations. The distributions of β-thalassemia mutations, benign variants, and the sickle-cell mutation within the ranked mutational landscape were examined to assess disease-related distribution pattern. Clinically annotated β-thalassemia, sickle-cell, and benign variants were curated and mapped to the global mutation landscape. Each variant was assigned a Δlog-likelihood value and corresponding disruptiveness rank relative to all possible substitutions. All PLM results, along with the Python script used for PLM result curation, are provided in **SI 2 (Supplementary Information 2)**.

Identification of Mutation-Intolerant Hotspots

Residues were ranked according to their average mutation disruptiveness. The 15 most intolerant residues were defined as mutation-intolerant hotspots. Local structural features were computed using the crystal structure of human hemoglobin β-chain (PDB ID: 1A3N, chain B; Neighbor count: number of residues within a defined spatial cutoff, Cα–Cα distance threshold; Distance to heme: minimum atomic distance between residue and heme group). Structural mapping and visualization were performed using standard molecular visualization tool PyMOL [33]. The complete Python script for the procedure, along with the list of 146 amino acid hotspots and the CSV file containing the top ~ 10% hotspots, are provided in **SI 3 (Supplementary Information 3)**.

Solvent Accessibility and Core vs Surface Classification

Solvent-accessible surface area (SASA) was computed from the 1A3N crystal structure (chain B) using the Shrake–Rupley algorithm as implemented in Biopython [34–36]. Residue-level SASA values were normalized to obtain relative solvent accessibility (RSA). Residues were classified as: Core (Closer to HEM): RSA < 0.25, Surface (Not closer to HEM): RSA ≥ 0.25. Distributions of mutation disruptiveness between core and surface residues were compared using a two-sided Mann–Whitney U test. The Python script, together with the corresponding input files and the SASA and RSA CSV files, are provided in **SI 4 (Supplementary Information 4)**.

Layer-Resolved Structural Encoding Analysis

Layer-wise mutation sensitivity was computed using ESM-1b (33 layers) [17]. For the HBB sequence, hidden representations from each transformer layer were extracted in a single forward pass. Layer-specific logits were obtained via the shared language modeling head, and log-probabilities were computed using softmax normalization. At each residue and layer, average mutation sensitivity (ΔLL) was calculated as the mean log-likelihood difference between the wild-type residue and all 19 possible substitutions. Relative solvent accessibility (RSA) values were derived from the HBB crystal structure

(PDB 1A3N, chain B). For each layer, Spearman correlation was computed between positional average ΔLL and RSA (n = 145 residues).

Structural Classification of β-Thalassemia Mutations

All 146 amino-acid positions in hemoglobin β (PDB 1A3N) were classified into three structural categories based on atomic contacts and heme proximity: Core, Heme-proximal (positions within 10 Å of heme iron), and Surface (remaining positions). β-thalassemia-associated mutations (n = 20) were partitioned by structural classification. Median Δ log-likelihood (ΔLL) values were computed for each class.

Fitness Landscape Topology Analysis

$$V_{(i)} = \frac{\sum [\Delta \, \mathrm{LL} \, (\mathrm{i}, \, \mathrm{j}) - \mu \, (\mathrm{i} \,)] \, ^{2}}{n}$$

Positions were classified as high-variance (> median) or low-variance (≤ median) using threshold of 5.97 (Corresponds to the median positional variance across all positions). Proportions of β-thalassemia (n = 20) and benign (n = 15) mutations at high-variance positions were compared using Fisher's exact test (α = 0.05). The Python script used for the topology analysis and structural classification along with the corresponding files, is provided in **SI 5 (Supplementary Information 5)**.

Evolutionary Constraint Analyses

Evolutionary Entropy

Evolutionary residue entropy was computed from a multiple sequence alignment of homologous sequences for each protein. Homologous sequences were retrieved from the non-redundant protein sequence database using BLAST with an E-value threshold < 10^{-5} and aligned using MAFFT [37–39]. Shannon entropy was calculated at each position based on amino acid frequency distributions, with lower entropy indicating stronger evolutionary conservation [40]. Spearman rank correlation was used to evaluate the association between evolutionary entropy and the average PLM-derived mutation disruptiveness.

BLOSUM62 Substitution Scores

For each mutation, the corresponding BLOSUM62 substitution score [26] was assigned based on the wild-type and mutant amino acids. Spearman correlation was used to quantify the association between PLM-derived mutation disruptiveness and BLOSUM62 substitution scores [41].

PLM-Derived Residue Entropy

PLM-derived residue entropy was computed directly from the model's predicted amino acid probability distribution at each position using Shannon entropy [42]. Unlike evolutionary entropy derived from multiple sequence alignment, this measure reflects the model's internal distributional uncertainty. Spearman correlation and corresponding p-values were calculated to assess the relationship between PLM-derived residue entropy and PLM mutation disruptiveness.

Statistical Analysis

All statistical analyses were conducted in Python using SciPy and related scientific computing libraries [43]. Correlations were evaluated using Spearman rank correlation. Group comparisons were performed using two-sided Mann–Whitney U tests. Statistical significance was defined as $p < 0.05$. The whole datasets, python scripts for evolutionary entropy, BLOSUM62 [26] and residual entropy are provided in **SI 6 (Supplementary Information 6)**.

Multivariable Regression and Variance Decomposition

To quantify the independent and interactive contributions of structural depth and substitution chemistry to PLM-derived mutation sensitivity, we performed mutation-level ordinary least squares (OLS) regression across all substitutions. Standardized Δlog-likelihood values (z-scored) were modeled as a function of standardized relative solvent accessibility (RSA), standardized BLOSUM62 substitution scores [26], and their interaction term (RSA × BLOSUM62). Adjusted R^2 was used to estimate total explained variance. Semi-partial R^2 values were calculated to quantify unique variance independently attributable to each predictor. Multicollinearity was assessed using variance inflation factors (VIF). The python script is provided in **SI 7 (Supplementary Information 7)**.

Mutation-Level PLM and EVE Comparison

All possible single amino acid substitutions across the HBB sequence (n = 2,774 missense variants) were evaluated. For each mutation, mutation sensitivity was quantified using the change in log-likelihood (Δlog-likelihood) derived from the protein language model (PLM). Variant effect predictions from a deep generative evolutionary model (EVE) were obtained for the same mutation set [44]. Spearman rank correlation and Pearson correlation were computed to assess monotonic and linear relationships between PLM mutation sensitivity and EVE scores. To quantify the proportion of variance shared between the two models, ordinary least squares (OLS) regression was performed using standardized variables (z-score transformation). Standardization was applied to both PLM and EVE scores to enable interpretation of regression coefficients as standardized effect sizes.

Cross-Model Validation

To assess the robustness and reproducibility of mutation sensitivity estimates, the complete mutational landscape analysis was independently repeated using the next-generation protein language model ESM-2 [22]. Mutation-level scores derived from ESM-2 were compared with those obtained from ESM-1b to evaluate cross-model concordance and the preservation of structural encoding signals across model generations.

Additional Proteins for Cross-Protein Validation

To assess generalizability, we extended the mutation landscape analysis to 39 additional structurally diverse proteins (total n = 40; length range 37–597 residues). For each protein, all possible single–amino-acid substitutions were scored using the same PLM pipeline applied to HBB [17, 20]. Relative solvent accessibility (RSA) was computed from crystal structures. Core–surface comparison (Mann–Whitney), Spearman correlation with RSA, evolutionary entropy, and BLOSUM62 substitution scores were calculated uniformly across proteins without retraining **SI 8 (Supplementary Information 8)**.

Declarations

Acknowledgements

We gratefully acknowledge the Department of Biomedical Engineering and the Department of Computer Science and Engineering at the Military Institute of Science and Technology and Multimedia University, Malaysia for their institutional support and facilities provided during the completion of this research. We also acknowledge the use of artificial intelligence–based language tools for assistance with grammar refinement and improvement of formal scientific writing.

Author Contributions

Tangilal Dihan Chowdhury conceived the study, designed the methodology, performed all computational analyses, interpreted the results, and wrote the original manuscript.

Md Ushama Shafoyat and Syed Rashedul Haque contributed to data processing, implementation of analytical pipelines, and assisted in validation of results.

Md. Tobibul Islam and Maruf Hasan contributed to data curation, preprocessing, and computational workflow execution.

Mohammad Shahjahan Majib, Md Mahbubur Rahman, and Kazy Noor e Alam Siddiquee provided domain expertise, contributed to structural and biological interpretation, and critically reviewed the results.

Kaiissar Mannoor contributed to study design, provided conceptual input, and participated in critical revision of the manuscript.

Md. Shabiul Islam supervised the overall project, provided strategic guidance, and critically reviewed and edited the manuscript.

All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Data availability

All data generated or analyzed during this study are included in this published article and its Supplementary Information files. The complete mutation datasets, protein language model (PLM) outputs, and all analysis scripts are provided in Supplementary Information 1–12. Additional data are available from the corresponding author upon reasonable request.

Additional Information

Competing interests

The authors declare no competing interests.

References

1. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
2. Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
3. Findlay, G. M. et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
4. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods*. **11**, 801–807 (2014).
5. Kao, H. J. et al. De novo mutation and skewed X-inactivation in a girl with BCAP31-related syndrome. *Hum. Mutat.* **41**, 1775–1782 (2020).
6. Manrai, A. K. et al. Genetic misdiagnoses and the potential for health disparities. *N Engl. J. Med.* **375**, 655–665 (2016).
7. Frye, M. et al. RNA modifications: what have we learned and where are we headed? *Nat. Rev. Genet.* **17**, 365–372 (2016).
8. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

9. Saad, H. K. M. et al. HBB gene mutations and their pathological impacts on HbE/β-thalassaemia in Kuala Terengganu. *Malaysia Diagnostics*. **13**, 1247 (2023).
10. Firnberg, E. et al. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* **31**, 1581–1592 (2014).
11. Bloom, J. D. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol. Biol. Evol.* **31**, 1956–1978 (2014).
12. Thein, S. L. The molecular basis of β-thalassemia. *Cold Spring Harb Perspect. Med.* **3**, a011700 (2013).
13. Jaing, T. H. et al. Molecular genetics of β-thalassemia: a narrative review. *Medicine* **100**, e27522 (2021).
14. Ingram, V. M. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature* **180**, 326–328 (1957).
15. Saalbach, K. P. Gain-of-function research. *Adv. Appl. Microbiol.* **120**, 79–111 (2022).
16. Perutz, M. F. et al. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution. *Nature* **185**, 416–422 (1960).
17. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* 118, e2016239118 (2021).
18. Cheng, P. et al. Zero-shot prediction of mutation effects with multimodal deep representation learning guides protein engineering. *Cell. Res.* **34**, 630–647 (2024).
19. Elnaggar, A. et al. ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
20. Rao, R. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems 34 (NeurIPS*) (2021).) (2021). (2021).
21. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **54**, 1812–1819 (2022).
22. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
23. Hsu, C. et al. Learning protein fitness models from evolutionary and assay-labeled data. *Nat. Biotechnol.* **40**, 1114–1122 (2022).
24. Hopf, T. A. et al. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
25. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. *Proc. Natl Acad. Sci. USA* 118, e2016239118 (2021).
26. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* 89, 10915–10919 (1992).

27. Kabsch, W. & Sander, C. Dictionary of protein secondary structure. *Biopolymers* **22**, 2577–2637 (1983).
28. Rose, P. W. et al. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.* **39**, D392–D401 (2011).
29. Giardine, B. M. et al. Clinically relevant updates of the HbVar database of human hemoglobin variants. *Nucleic Acids Res.* **49**, D1192–D1196 (2021).
30. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
31. Dhawan, M. et al. Delta variant of SARS-CoV-2: mutations, impact and challenges. *Hum. Vaccin Immunother*. **18**, 2068883 (2022).
32. Shirvani-Dastgerdi, E. et al. Comprehensive analysis of hepatitis delta virus mutations. *Clin. Microbiol. Infect.* **21**, 510e11–510e23 (2015).
33. Yuan, S., Chan, H. C. S. & Hu, Z. Using PyMOL as a platform for computational drug design. *WIREs Comput. Mol. Sci.* **7**, e1298 (2017).
34. Shrake, A. & Rupley, J. A. Environment and exposure to solvent of protein atoms. *J. Mol. Biol.* **79**, 351–371 (1973).
35. Ribeiro, J. et al. Calculation of accurate interatomic contact surface areas. *Bioinformatics* **35**, 3499–3501 (2019).
36. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology. *Bioinformatics* **25**, 1422–1423 (2009).
37. Suzek, B. E. et al. UniRef clusters: improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
38. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
39. Katoh, K. & Standley, D. M. MAFFT version 7. *Mol. Biol. Evol.* **30**, 772–780 (2013).
40. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*. **37**, 145–151 (1991).
41. Ferdous, S. M. et al. BlosumACPred: identification of anticancer peptides. In ICTIS 2025 (Springer, (2026).
42. Feutrill, A. & Roughan, M. A review of entropy rate estimation. *Entropy* **23**, 1046 (2021).
43. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing. *Nat. Methods*. **17**, 261–272 (2020).
44. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
45. Bunn, H. F. Pathogenesis and treatment of sickle cell disease. *N Engl. J. Med.* **337**, 762–769 (1997).
46. Bowie, J. U., Lüthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known structure. *Science* **253**, 164–170 (1991).

47. Tokuriki, N. & Tawfik, D. S. Chaperonin overexpression promotes genetic variation. *Nature* **459**, 668–673 (2009).
48. Cao, A. & Galanello, R. Beta-thalassemia. *Genet. Med.* **12**, 61–76 (2010).
49. Karczewski, K. J. et al. The mutational constraint spectrum in humans. *Nature* **581**, 434–443 (2020).
50. Fermi, G. et al. Crystal structure of human deoxyhaemoglobin. *J. Mol. Biol.* **175**, 159–174 (1984).
51. Perutz, M. F. Stereochemistry of cooperative effects in haemoglobin. *Nature* **228**, 726–734 (1970).

Figures

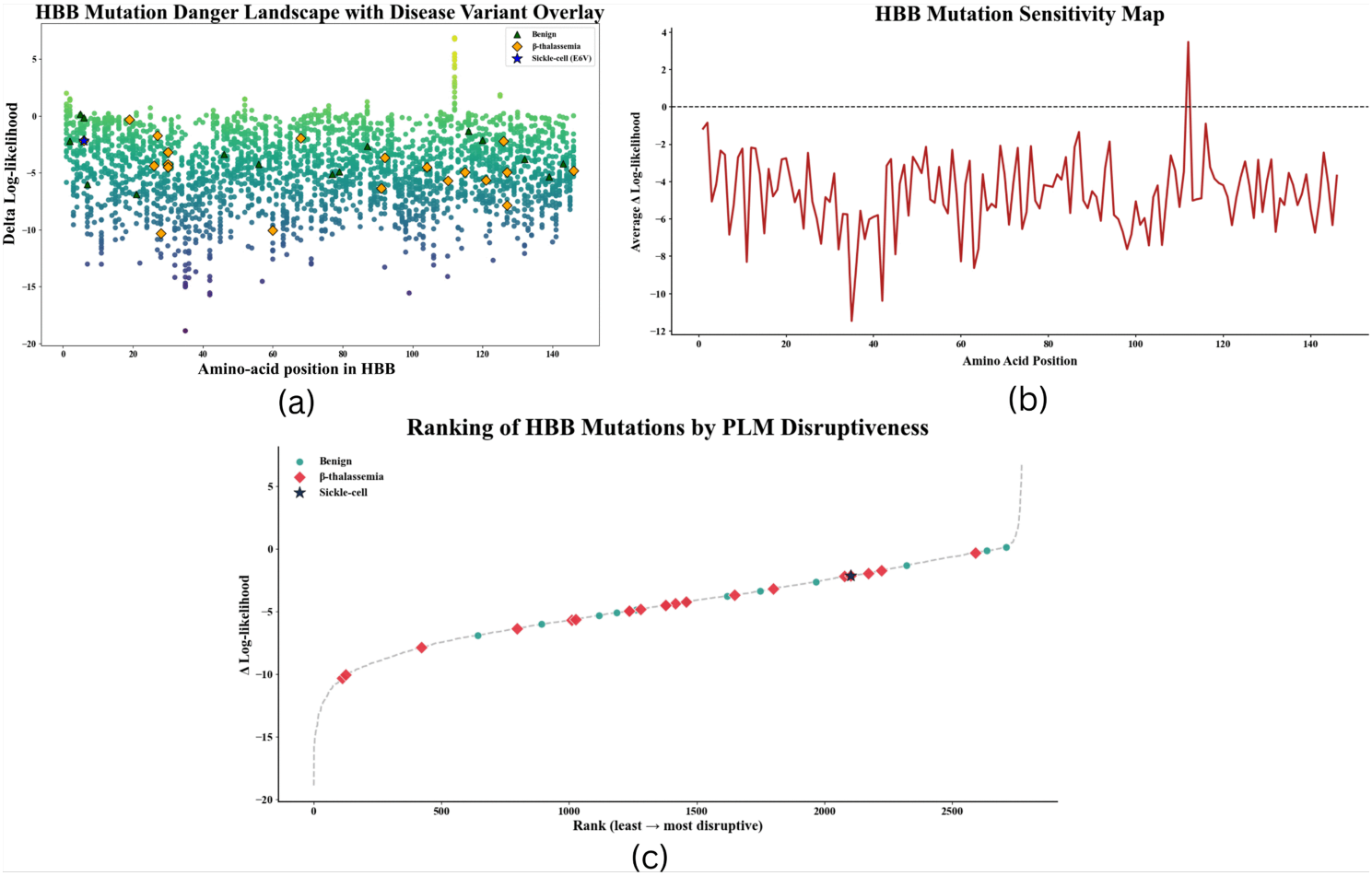


Figure 1

Comprehensive view of the HBB missense mutation landscape. (a) Distribution of PLM-derived mutation sensitivity (Δlog-likelihood) across all 2,774 possible missense substitutions in HBB, with clinically annotated variants overlaid (β-thalassemia, benign, and sickle-cell E6V). (b) Average mutation sensitivity per residue position, highlighting regions of increased intolerance suggestive of structural or functional constraint. (c) Ranking of all possible substitutions by disruptiveness, showing the relative positioning of clinically annotated variants within the global mutation severity spectrum.

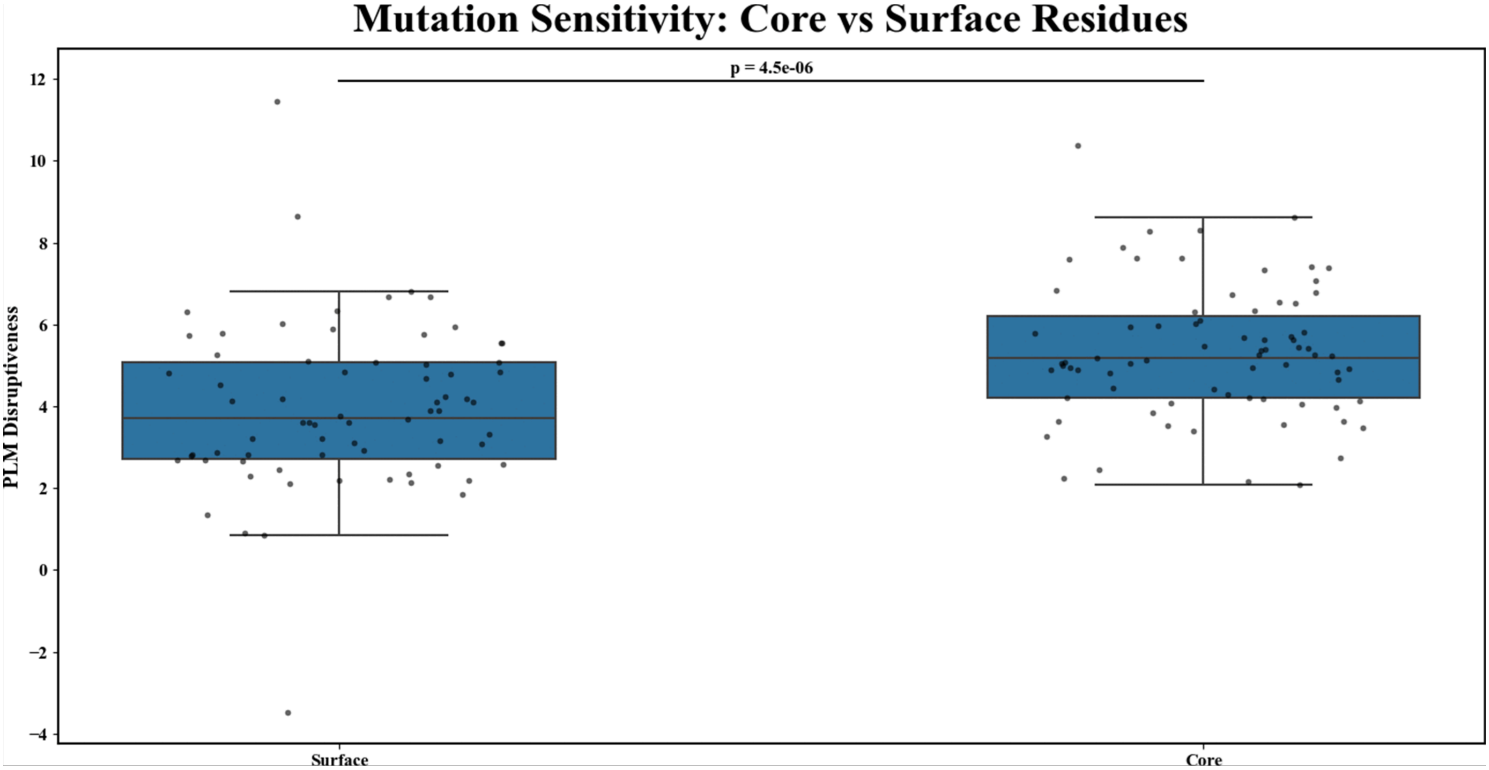


Figure 2

Mutation Sensitivity: Core versus Surface Residues. Comparison of PLM-derived mutation disruptiveness between buried (core) and solvent-exposed (surface) residues across all 2,774 substitutions. Core residues exhibit significantly greater mutation sensitivity (Mann–Whitney U test, $p = 4.48 \times 10^{-6}$). Boxplots show median and interquartile range, with individual residue-level averages plotted as points. Results demonstrate strong structural encoding in the PLM-derived mutation landscape.

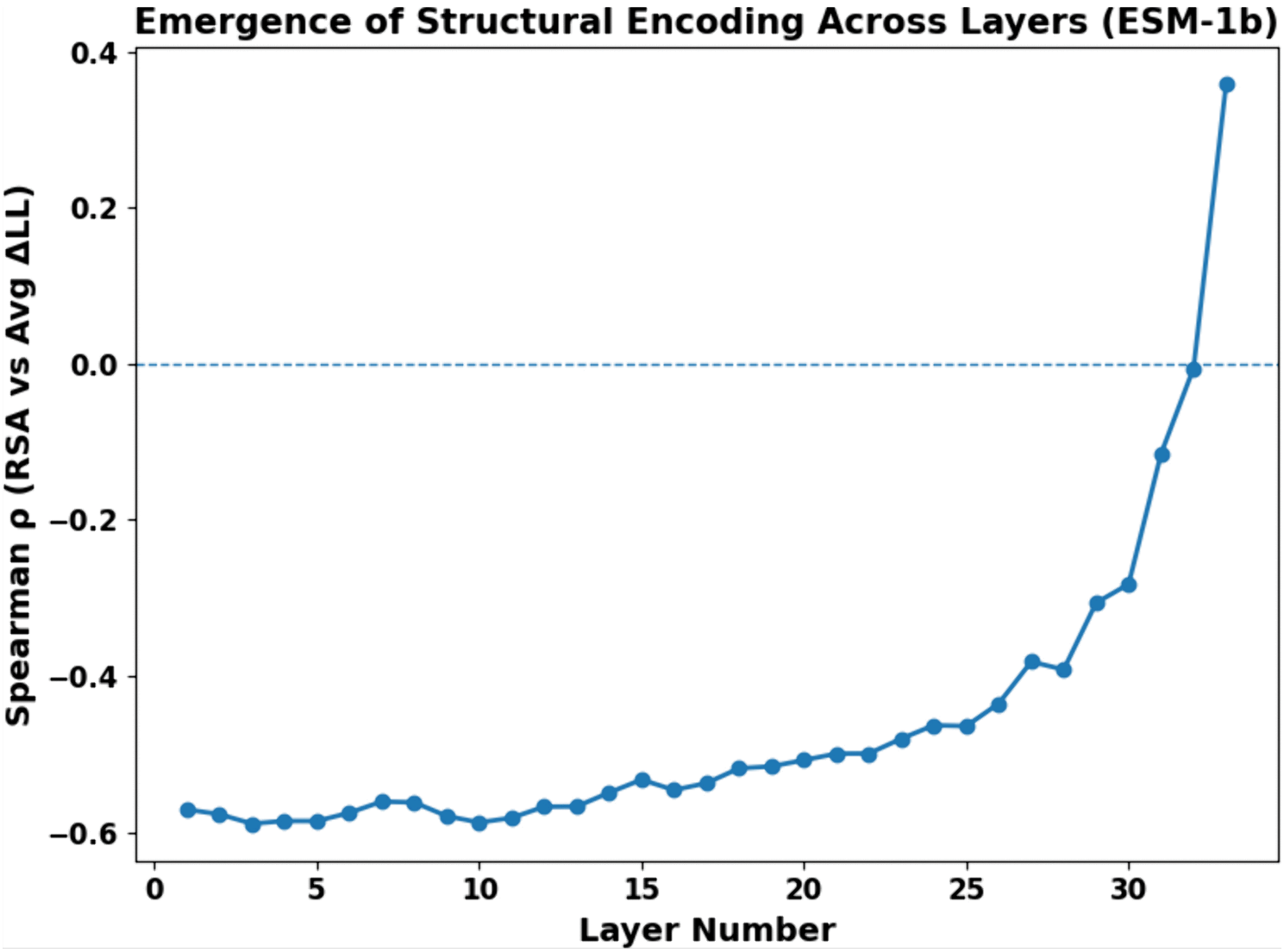


Figure 3

Layer-wise distribution of burial-dependent structural encoding in ESM-1b. Spearman correlation (ρ) between relative solvent accessibility (RSA) and average mutation sensitivity (ΔLL) computed at each transformer layer (1–33) for human hemoglobin β. Early layers exhibit strong negative correlation (ρ ≈ −0.59), indicating pronounced mutation intolerance at buried residues. The magnitude of this association progressively decreases across intermediate layers and approaches zero in deep layers (31–32). Notably, the final output layer (33) shows positive correlation (ρ = 0.358), indicating inversion of the burial-dependent signal.

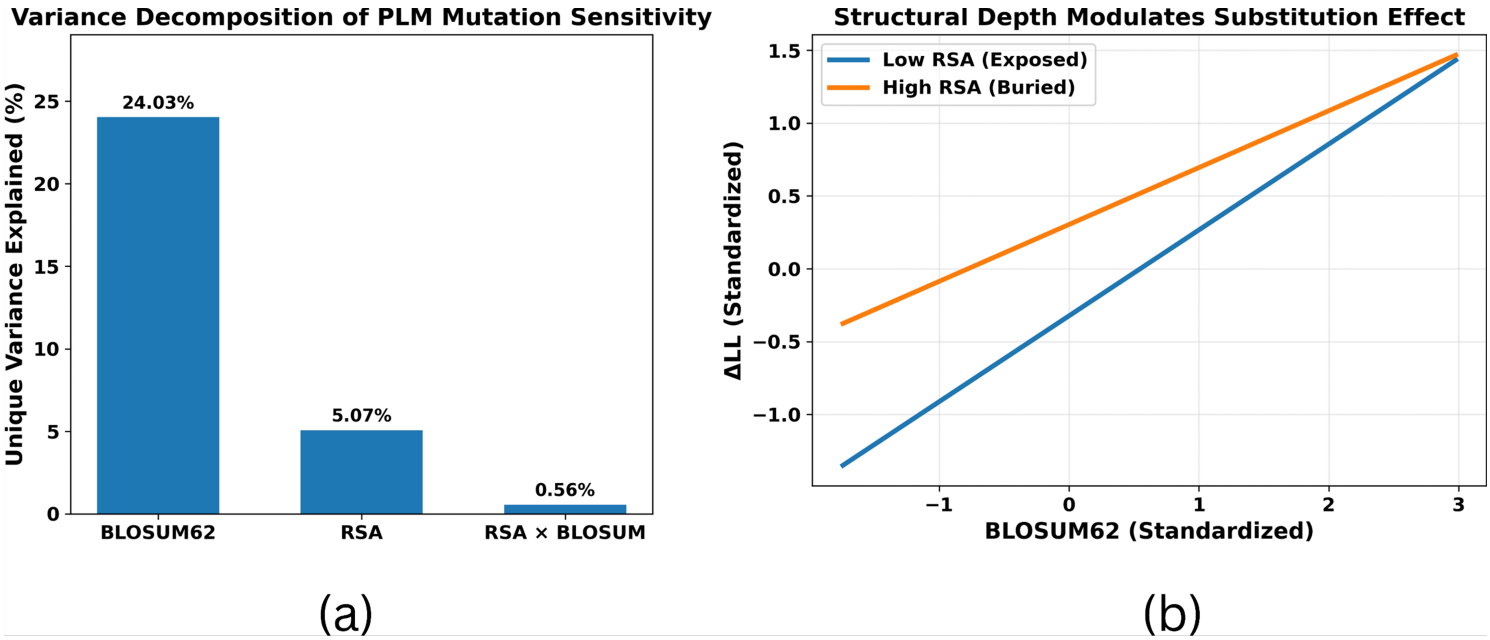


Figure 4

Structural and Evolutionary Determinants of PLM Mutation Sensitivity. (a) Analysis of Protein Language Model (PLM) scores shows that BLOSUM62 explains the majority of unique variance (24.03%), followed by Relative Solvent Accessibility (RSA; 5.07%) and their interaction (0.56%). (b) Comparison of regression slopes for buried (high RSA) vs. exposed (low RSA) residues. The higher intercept for buried residues indicates that PLMs predict greater mutational sensitivity (Δ*LL*) for substitutions within the protein core, particularly when evolutionary favorability (BLOSUM62) is low.

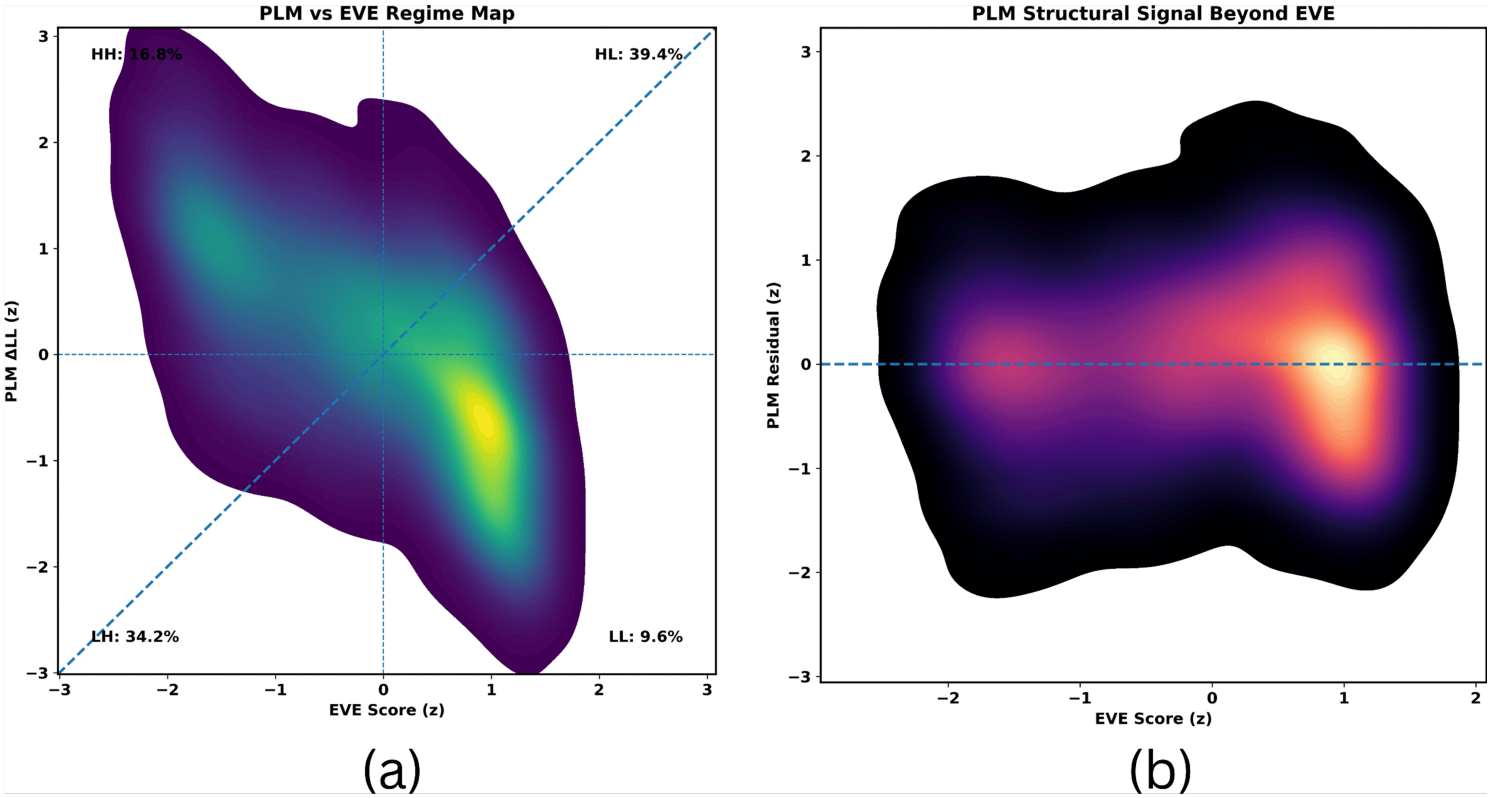


Figure 5

Comparative Analysis of Protein Language Model (PLM) and Evolutionary Model (EVE) Mutational Predictions. (a) Density plot comparing standardized mutational effect scores from PLM (ΔLL) and EVE. While the models generally align (High-High: 16.8%; Low-Low: 9.6%), significant discordance exists where EVE predicts high constraint but PLMs do not (High-Low: 39.4%), or vice versa (Low-High: 34.2%), indicating distinct modeling regimes. (b) Density plot of PLM residuals relative to EVE scores.

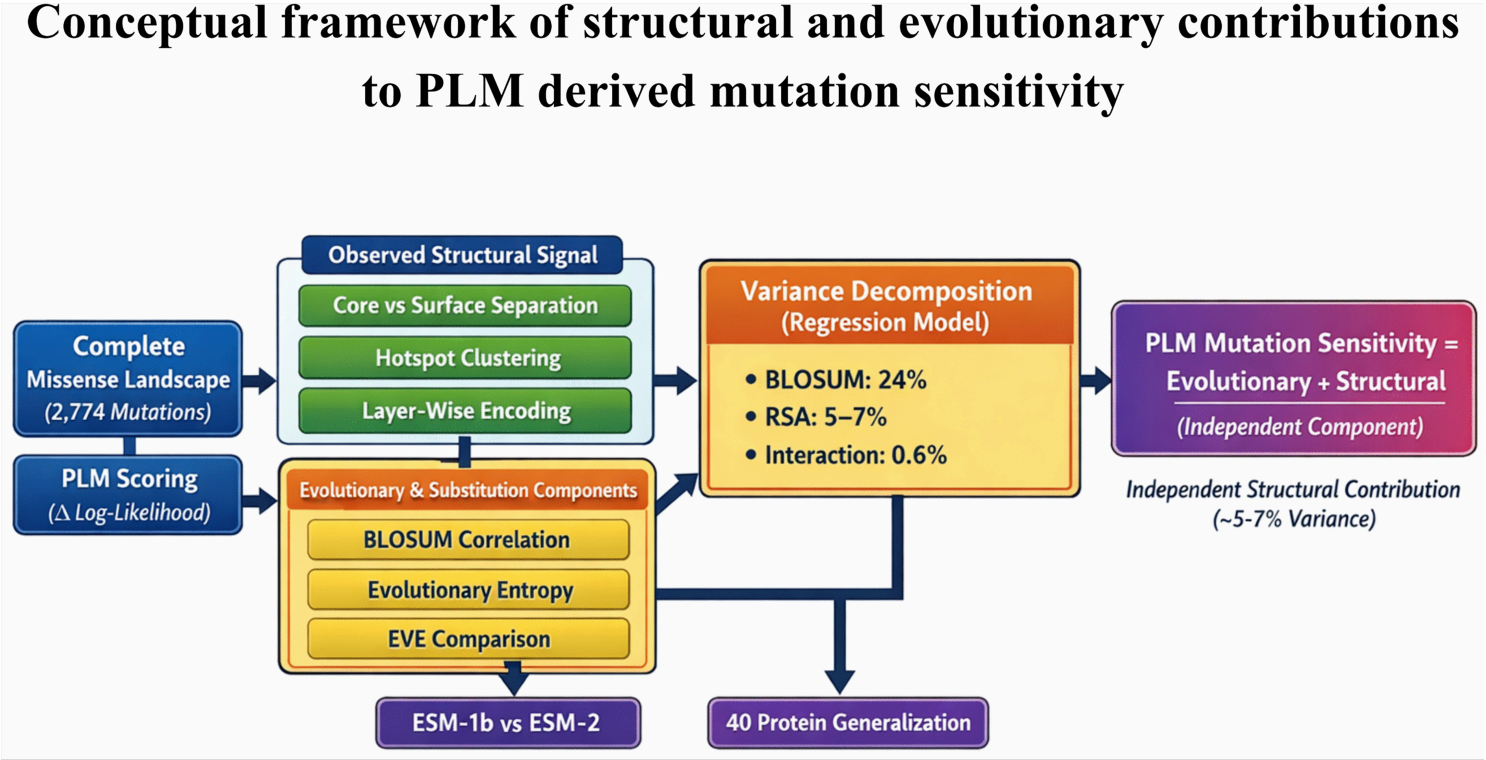


Figure 6

Conceptual framework of PLM mutation sensitivity. PLM-derived mutation sensitivity across the complete missense landscape shows structural dependence (core–surface separation, hotspots, layer effects) and partial alignment with evolutionary constraints. Variance decomposition demonstrates an independent structural contribution (~5–7%) beyond substitution effects. This signal is reproducible across models and generalizable across proteins.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation1.zip
- SupplementaryInformation2.zip
- SupplementaryInformation5.zip
- SupplementaryInformation8.zip

- SupplementaryInformation3.zip
- SupplementaryInformation10.zip
- SupplementaryInformation4.zip
- SupplementaryInformation12.zip
- SupplementaryInformation7.zip
- SupplementaryInformation6.zip
- SupplementaryInformation11.zip
- SupplementaryTables.docx
- SupplementaryFigures.docx
- SupplementaryInformation.docx