Fine-tuning protein language models on human spatial constraint yields state-of-the-art variant effect prediction

Gyasu Bajracharya[1] and John A. Capra[2,3,4,*]

[1]Biophysics Graduate Program, University of California, San Francisco, CA
[2]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA
[3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA
[4]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA

*Correspondence to tony@capralab.org

Abstract

Millions of missense variants are present in human genomes, yet the functional consequences of most remain unknown. Here, we introduce Human Spatial Constraint (HSC), a framework for quantifying intraspecies constraint on missense variants that integrates population-scale human genetic variation with 3D protein structures. HSC models the expected frequency of missense variation under neutral evolution and compares it to observed variation, accounting for both variation in mutational processes and structural context. HSC outperforms traditional inter- and intraspecies conservation metrics, as well as unsupervised protein language models (PLMs) such as ESM1b, in predicting pathogenic variants, achieving performance comparable to AlphaMissense. Fine-tuning PLMs on HSC scores improves the prediction of variant fitness across diverse taxa and deep mutational scanning (DMS) functional assay types. Together, these results demonstrate that combining intraspecies constraint with cross-species PLMs improves variant effect interpretation and understanding of protein function.

1 Introduction

The frequency of genetic variants within a population and between species is shaped by neutral processes and selection. Alleles that decrease organismal fitness tend to be lost over time, while fitness-increasing mutations are more likely to be retained. Comparing patterns of mutations and their frequencies to expectations under neutral evolution can highlight evolutionary constraints[1,2]. For example, sites in proteins essential for maintaining structure and function often exhibit little genetic variation due to strong evolutionary constraints. Quantifying these constraints both within and between species has provided valuable insights into proteins' roles in disease and the selective pressures shaping their evolution.

Common methods for inferring evolutionary constraints between species rely on multiple sequence alignments (MSAs), where sequence conservation across species indicates constraint[3–6]. Recently, deep learning approaches such as protein language models (PLMs), trained on diverse protein sequences across the evolutionary tree, have expanded our ability to model constraints on sequence variation[7–10]. These models have achieved strong performance in predicting pathogenic variants identified in rare diseases and estimates of protein fitness from deep mutational scanning (DMS) data[11]. However, these methods do not inherently account for intraspecies sequence variation, which captures patterns of variation that reflect more recent selective pressures. Additionally, the complexity of PLMs makes interpretation challenging; it is not always clear how they arrive at their decisions. This so-called 'black-box' nature may limit trust and broader adoption, especially in clinical settings.

The recent availability of large databases of human genetic variation across hundreds of thousands of individuals has enabled new strategies for quantifying evolutionary constraint within human populations[12–14]. Quantifying evolutionary constraints within the modern human population can provide relevant context for understanding the effects of variants on human disease. These methods have focused on quantifying constraints based on tolerance to missense or loss-of-function (LoF) variants at the protein or region level[15–20]. These methods have been further augmented by leveraging models of human protein 3D structures to consider the spatial context of variation[2,21,22]. These approaches become more powerful and comprehensive with the growth of accurate computational methods for predicting 3D structure from sequence, but have not been competitive with interspecies conservation metrics in the prediction of site-specific variant effects.

Here, we unify these perspectives by improving the quantification of population-scale human genetic variation in protein 3D structural context and integrating it into PLMs. First, we introduce the Human Spatial Constraint (HSC) score, a framework that quantifies deviations in the frequency of missense variation within protein spatial contexts from expectations under neutral evolution. Additionally, we present an approach for incorporating these measurements of intraspecies constraint into PLMs through supervised fine-tuning. We show that HSC outperforms traditional inter- and intraspecies conservation metrics and rivals deep-learning approaches in predicting pathogenicity. Furthermore, integrating HSC into PLMs improves protein fitness predictions beyond the base model, highlighting the complementary roles of inter- and intra-species evolutionary constraints. Our results demonstrate that integrating intraspecies and interspecies constraints to capture both long-

term evolutionary constraint and recent human-specific selection provides a more comprehensive view of a protein's functional landscape and variant effects.

2 Results

The Human Spatial Constraint (HSC) framework maps evolutionary constraint in 3D protein structures

We developed the Human Spatial Constraint (HSC) framework to quantify evolutionary constraint on 3D spatial regions of proteins in the human population. Our approach integrates human genetic variation data across hundreds of thousands of individuals with accurate models of protein 3D structures for ~80% of human protein-coding genes. Many previous approaches, including our own COSMIS model, have quantified patterns of missense variation across human proteins. HSC builds on these previous approaches by integrating strategies for considering protein 3D structural context, variable mutation rates at the nucleotide and protein scales, and the frequency distribution of observed variants.

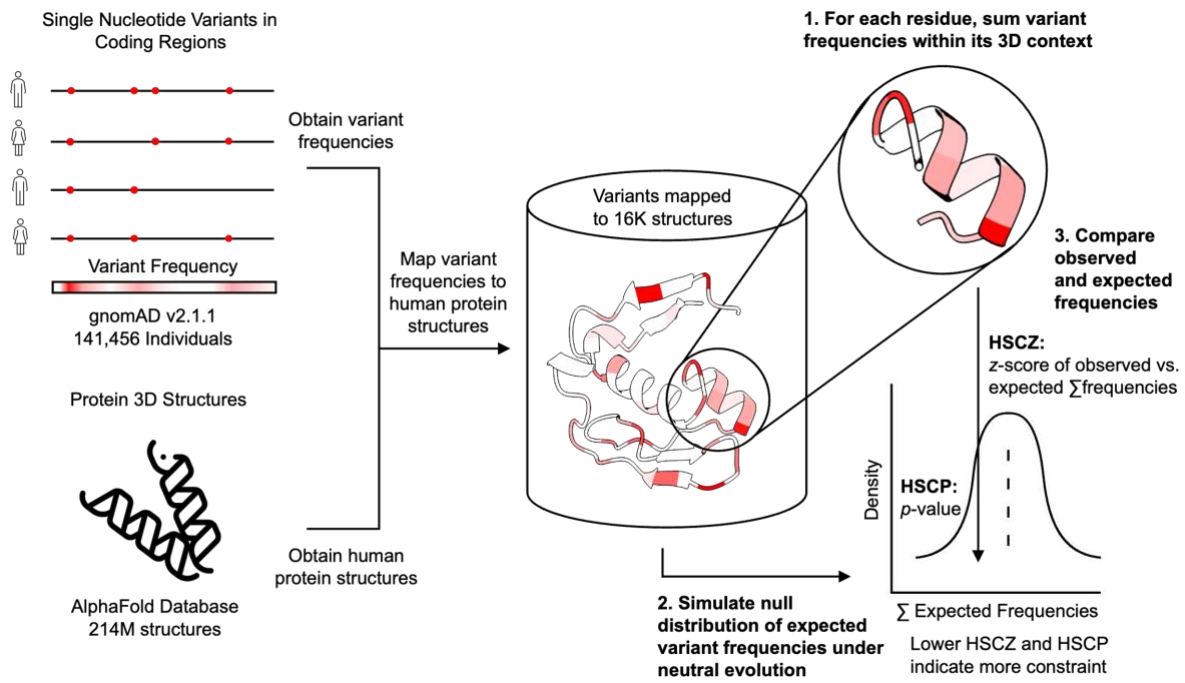

Figure 1: Schematic of the Human Spatial Constraint (HSC) Framework
The HSC framework integrates population genetics data with protein structure information by mapping single-nucleotide variant frequencies from hundreds of thousands of human genomes onto predicted 3D protein structures from the AlphaFold Database. For each protein residue, the framework aggregates the frequencies of variants observed within a 3D spatial region centered on the residue (e.g., an 8 Å sphere). The observed variant frequency within each 3D context for a residue is compared against a null distribution of total expected frequencies, generated using a mutation-spectrum-aware neutral model, to compute a p-value (HSCP) and a log-transformed z-score (HSCZ).

The HSC framework maps missense and synonymous variants and their minor allele frequencies (MAFs) across 141,456 individuals from gnomAD v2.1.1 into 16,138 3D protein structures from the AlphaFold Database (Fig. 1)[23,24]. To estimate the expected number and frequency of missense variants in a given 3D spatial region of a protein under neutral evolution, we built a permutation-based model that incorporates 3D spatial regions, local variation in the mutability of different nucleotide contexts, and global variation between proteins (Methods). Using this model, we calculate the expected frequency of missense variants in variable 3D spatial regions (e.g., spheres of varying radius) centered on each amino acid position in each protein.

To compute the HSC score, we compare the observed frequency of variants in a spatial region to the expected frequency based on permutations reflecting neutral evolution (Methods). Specifically, for a given protein spatial region, we quantify its constraint with a Z-score-based measure called HSCZ that is the logarithm of the observed frequency of variation in the spatial region minus the average of the expected frequency for the region over 10,000 permutations divided by the standard deviation across permutations. We also compute HSCP, an empirical p-value based on comparing the observed frequency of variants to the frequency distribution from the permutations. Thus, lower HSCZ values signify a stronger constraint on a given 3D spatial region.

Human Spatial Constraint (HSC) scores highlight protein regions of functional importance

We applied the HSC framework with an 8 Å radius context region to 8.8 million residues across 16,138 human proteins. As expected, the distribution of HSCZ scores has a primary mode just below zero, reflecting sites experiencing evolutionary constraint, where missense variation is less tolerated than expected under neutral evolution (Fig. 2A). It also has a broader positive mode extending into a long right tail. This wider secondary mode corresponds to sites with more frequent variation than expected. The shorter tail for the negative scores (high constraint) is due to the fact that many of the most constrained regions do not have any variation, limiting the ability to differentiate among the most strongly constrained sites. Of the 8.8 million positions scored, approximately 1.1 million were evolving slower than expected (HSCP < 0.05), whereas around 3.3 million were evolving faster than expected (HSCP > 0.95) (Fig. S1A).

As a qualitative visual example, we show HSC scores, quantified using an 8 Å radius context region, overlaid on a crystal structure (PDB: 3DZU) of peroxisome proliferator-activated receptor γ (PPARG)[25] (Fig. 2B). Mutations in PPARG are linked to numerous diseases, including diabetes[26] and lipodystrophy[27,28]. PPARG functions as an obligate heterodimer with retinoid X receptor alpha (RXRα) and is capable of sequence-specific binding to DNA[29]. The HSC scores reveal constraints that align with functionally important regions of PPARG. Residues in the DNA-binding and ligand-binding domains, especially at the binding interfaces, exhibit strong constraint. This visual example shows that HSC scores are consistent with our expectations of constraint.

Comparing HSC scores across gene sets with known differences in evolutionary constraint and functional importance further supports that the scores perform as intended (Fig. 2C). For example, spatial regions in essential genes have significantly lower HSCZ

scores than non-essential genes, consistent with stronger constraint. Similarly, dominant genes show lower scores than recessive genes. Haploinsufficient genes—where one functional copy is insufficient for normal function—have the lowest scores, reflecting their heightened sensitivity to variation. Conversely, olfactory receptor genes have the highest scores, consistent with relaxed constraint in this gene family.

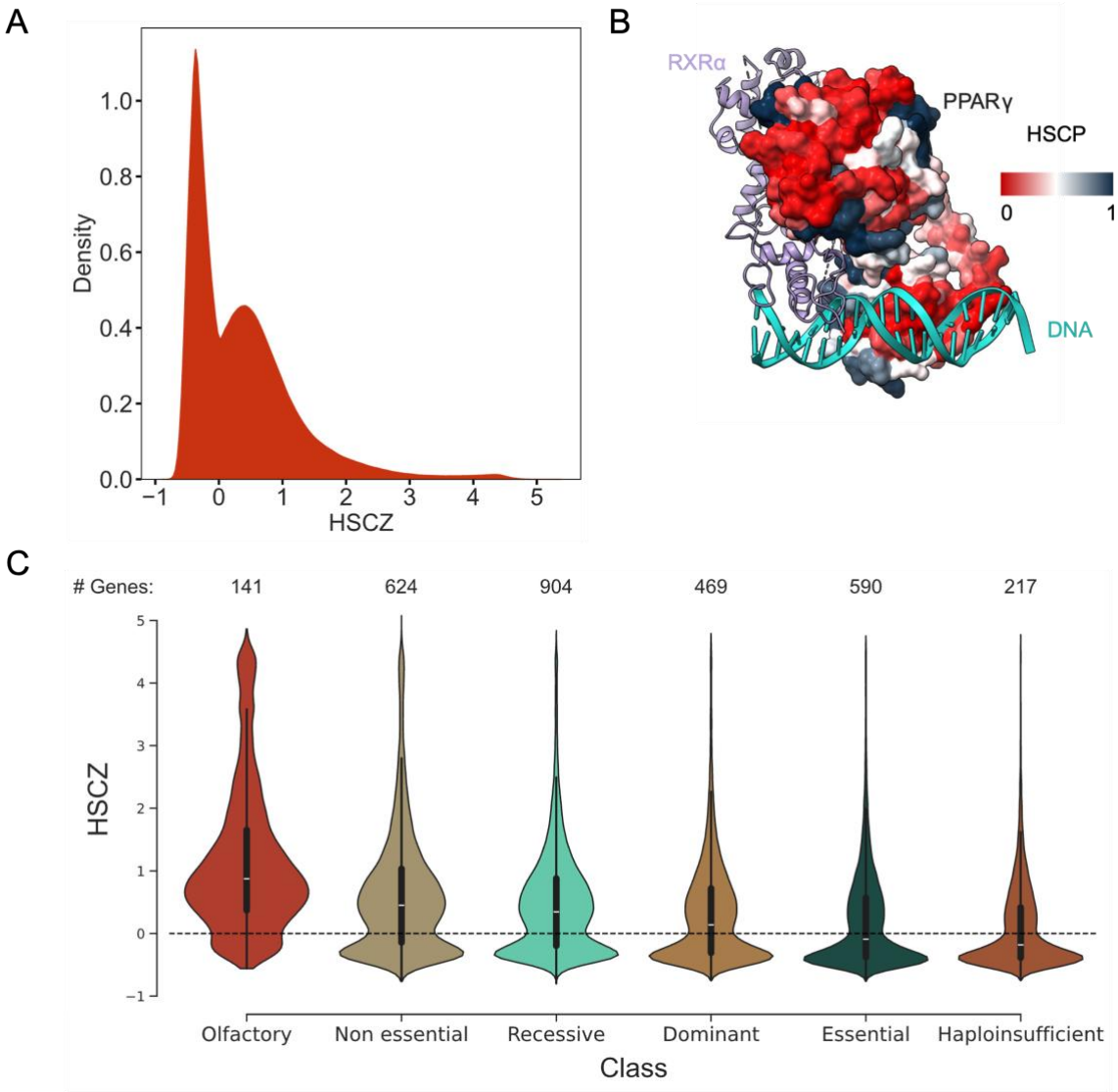

Figure 2: Human Spatial Constraint (HSC) scores reveal functional constraint on protein 3D spatial regions

(A) Distribution of HSCZ scores over 8.8 million residues across 16,138 human proteins. Negative scores indicate greater constraint, while positive scores indicate more variation than expected under neutral evolution. (B) Visualization of HSCP scores on the 3D crystal structure of Peroxisome Proliferator-Activated Receptor γ (PPARG) within the PPARγ–RXR–DNA complex (PDB: 3dzu). Residues are colored with lower HSCP values (red), indicating stronger constraints. The scores highlight strong constraints on the interfaces with RXRa (purple) and DNA (teal). (C) Violin plots of HSCZ scores for all residues within proteins stratified by protein functional annotations. Proteins with essential functions (haploinsufficient and essential genes) exhibit higher constraint compared to proteins unlikely to influence

fitness (olfactory receptors and non-essential genes). All pairwise comparisons between classes were statistically significant (Mann–Whitney U test, FDR-corrected $p < 0.001$).

HSC outperforms other intraspecies and interspecies constraint metrics in pathogenicity prediction

The HSC framework does not directly quantify pathogenicity; rather, it quantifies constraint on missense variation in 3D spatial regions in human populations. Evolutionary constraint correlates with pathogenicity, since pathogenic variants are often found in highly constrained regions of the genome. As a result, constraint metrics are often used alone or in combination with other features to identify pathogenic variants. Thus, we compared HSC's performance at pathogenicity prediction against both intraspecies and interspecies constraint metrics.

We first evaluated the ability of different methods to distinguish pathogenic from benign missense variants in ClinVar using receiver operating characteristic (ROC) and precision–recall (PR) curve area under the curve (AUC) metrics (Fig. 3A). The evaluation included 6,390 pathogenic and 7,099 benign variants curated from ClinVar.

HSCZ, calculated using an 8 Å context window, outperforms all other conservation metrics in our comparison, achieving ROC and PR AUCs of 0.91 and 0.90, respectively. The next best-performing methods are ConSurf (ROC = 0.84, PR = 0.79) and PhyloP (ROC = 0.80, PR = 0.76), both of which rely on interspecies conservation derived from multiple sequence alignments (MSAs). Among intraspecies constraint metrics, COSMIS (ROC = 0.72, PR = 0.74) and MTR3D (ROC = 0.66, PR = 0.68) perform comparably, whereas GERP (ROC = 0.68, PR = 0.60) and PhastCons (ROC = 0.68, PR = 0.67) show weaker performance. This does not indicate that HSC is necessarily more accurate than these other metrics; rather, it suggests that HSC's approach for quantifying constraint is more predictive of pathogenicity.

To assess the similarity of HSC scores compared to these methods, we computed Spearman correlations between HSC and other constraint metrics across approximately 3.6 million residues in the human proteome (Fig. 3B). As expected, HSC is most correlated with COSMIS ($\rho = 0.61$), a predecessor method that also integrates structural context with intraspecies genetic variation patterns. Its correlations with other intraspecies metrics, MTR3D and MTR, are lower ($\rho = 0.34$ and 0.35, respectively). HSC shows comparable correlations with gene-level constraint metrics such as RVIS ($\rho = 0.34$), pLI ($\rho = 0.36$), and missense Z ($\rho = 0.33$). By contrast, its correlations with interspecies conservation metrics are weaker—ConSurf ($\rho = 0.29$), phyloP ($\rho = 0.26$), GERP ($\rho = 0.17$), and PhastCons ($\rho = 0.20$)—consistent with these metrics capturing evolutionary conservation across species rather than population-level constraint. As expected, interspecies conservation metrics exhibit strong mutual correlations, particularly between phyloP and ConSurf ($\rho = 0.79$) and between PhastCons and GERP ($\rho = 0.67$).

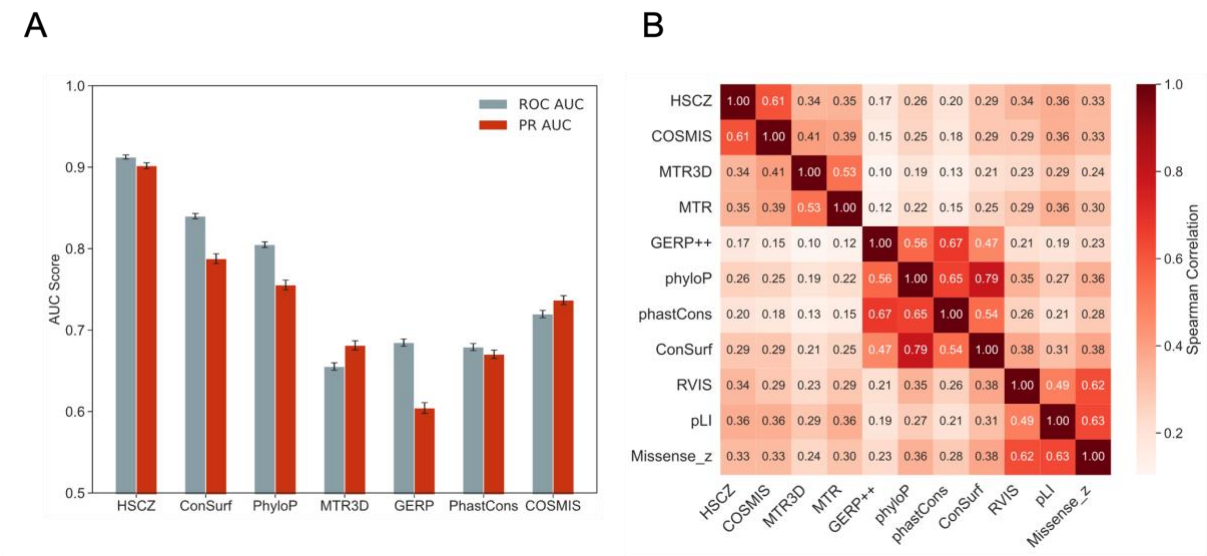

Figure 3. Comparison of HSC with other intraspecies and interspecies conservation metrics.

(A) Receiver operating characteristic (ROC) and precision–recall (PR) area under the curve (AUC) values for classifying 6,390 pathogenic and 7,099 benign missense variants from ClinVar using HSCZ (8 Å context) and other conservation metrics. Error bars represent the standard deviation of AUCs calculated from 1,000 bootstrap resamples of the variant dataset. Pairwise empirical p-values were computed from the bootstrap distributions to assess statistical significance between methods, with HSCZ achieving significantly higher ROC and PR AUCs compared to all other methods (FDR-corrected $p < 0.05$). (B) Spearman correlation matrix comparing HSC scores with other conservation metrics. HCSZ is only modestly correlated with interspecies (GERP++, phyloP, phastCons, ConSurf), intraspecies (MTR3D, MTR), and protein-level constraint scores (RVIS pLI, Missense_z). The correlation was computed over 3.6 million sites across 12,107 proteins.

Impact of the amount of spatial context on constraint quantification

HSC's integration of evolutionary constraint into the context of a 3D protein structure rather than a linear protein sequence context improves identification of pathogenic residues. To explore the influence of different amounts of 3D spatial context on constraint quantification, we applied the HSC framework using regions ranging from no structural context to all positions within 14 Å. We hypothesized that expanding the structural region would improve power to quantify constraint by capturing more sites and biologically relevant residue interactions. However, we also anticipated a trade-off between improved constraint estimation and lack of specificity when incorporating more distant residues, which could obscure constraints on smaller spatial scales.

The distributions of HSCZ and HSCP scores across the proteome shift systematically with increasing context region size (Fig. S1). Without structural context, the HSCZ distribution exhibits a large, sharp peak centered just below zero, reflecting that most positions in a protein experience modest constraint, but considering a single position offers limited ability to differentiate constraint levels (Fig. S1B). The corresponding HSCP distribution indicates low statistical power to identify constraints (Fig. S1A). With increasing

context region size, the HSCZ distribution broadens, suggesting improved sensitivity to variation in constraint. Meanwhile, the HSCP distribution shows more ability to detect extreme deviations from expected levels of variation. At 14 Å, we see a decrease in significant constraint, suggesting that constraint captured by smaller context regions begins to be diluted at this scale.

We investigated how well HSC, quantified using different context region sizes, classifies pathogenic variants from ClinVar. Figure S2A shows the area under the curve (AUC) for both the Receiver Operating Characteristic (ROC) curve and the Precision-Recall (PR) curve for HSCZ scores quantified across various context region sizes. Both ROC AUC and PR AUC scores consistently decrease as the context region increases, indicating a reduction in their ability to distinguish pathogenic variants from benign variants. This suggests a structural context beyond directly contacting residues may introduce more false positives and negatives for this task.

Both no spatial context and 5Å context regions show the highest overall PR AUC (0.92; Fig. S2A), maintaining high precision (>0.9) at high recall values. In contrast, considering more spatial context results in the highest precision in the high confidence predictions (the low recall region of the curve), but at the cost of substantially lower precision at recall levels above 0.5 (Fig. S2B).

For pathogenicity prediction, quantifying constraint in the local spatial context improves precision for high-confidence predictions, but reduces the ability to identify all pathogenic variants. These findings suggest that a 5 Å context window for pathogenicity prediction may offer the best balance for leveraging structural context without leading to false positives. However, for cases where the precision of the top predictions is paramount (even at the cost of overall recall), an 8 Å context window may be preferable.

Combining HSC and deep-learning-based models improves pathogenicity prediction

Given the strong performance of HSC in predicting pathogenicity, we sought to compare its performance against state-of-the-art deep learning methods trained to predict variant effects, such as AlphaMissense and ESM1b. Despite HSC being a position-specific constraint score with no training on this task, we find that its performance in classifying pathogenic variants in ClinVar is nearly as high as these advanced models specifically tailored for pathogenicity prediction (Fig. 4A). AlphaMissense achieves the highest overall performance (ROC AUC of 0.93); however, HSCZ calculated with a 5 Å context window closely follows with a ROC AUC of 0.92, surpassing ESM1b (0.90). HSCZ calculated with an 8 Å window exhibits comparable performance to ESM1b, with ROC and PR AUC values of 0.89 and 0.88, respectively.

Since the AUC values are similar across models, we examined the curves to better understand performance differences (Fig. 4B). AlphaMissense and ESM1b excel in the high-confidence region of the curve, achieving higher true positive rates at low false positive rates. However, HSCZ identifies more true positives at higher false positive rates, suggesting that it may be advantageous in scenarios where maximizing sensitivity is prioritized over minimizing false positives in the top predictions.

AlphaMissense and ESM1b both leverage sequence information across the evolutionary tree, while HSC is human-specific. We investigated whether integrating HSC

scores with these models could enhance pathogenicity prediction (Fig. 4C). We performed logistic regression with five-fold cross-validation using two combined models: (i) AlphaMissense + HSCZ_5 and (ii) ESM1b + HSCZ_5 (Fig. 4C). In both cases, incorporating HSCZ_5 improved pathogenicity classification compared to the individual models, ROC AUC of 0.95 vs. 0.93 for AlphaMissense and 0.92 vs. 0.90 for ESM1b. This highlights the complementary nature of 3D spatial constraint in human populations captured by HSC and existing state-of-the-art pathogenicity predictors.

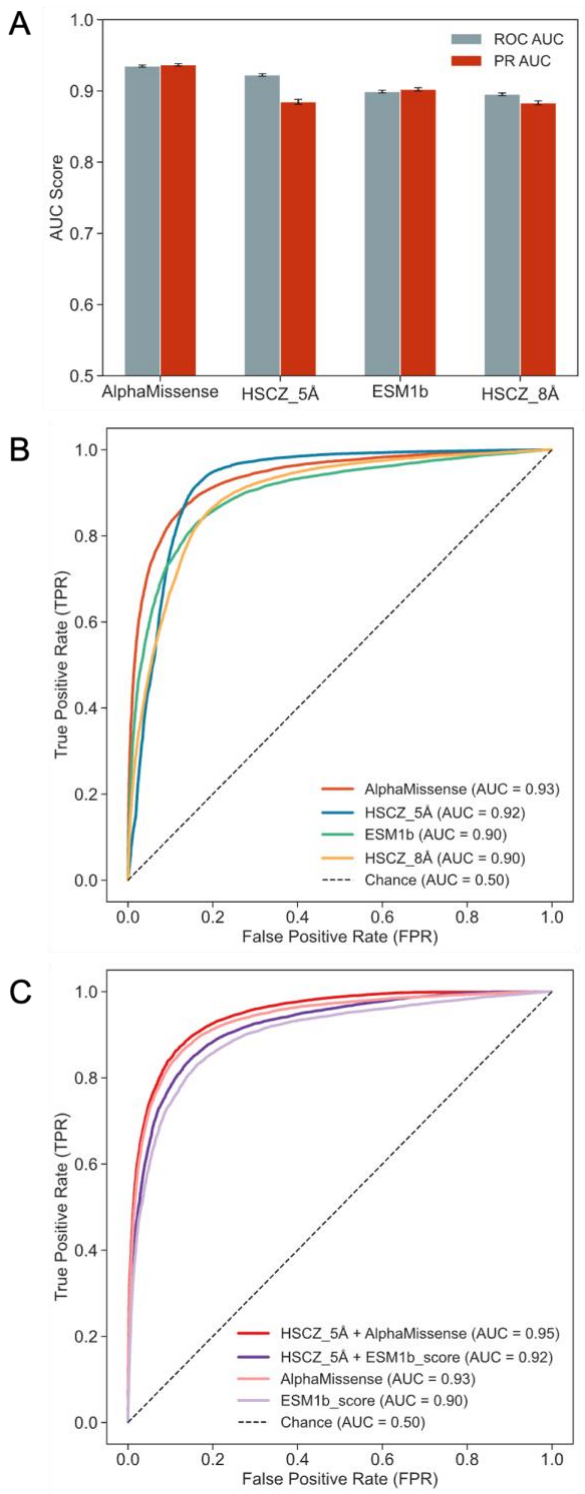

Figure 4. Benchmarking and combining HSC with deep learning-based variant effect predictors.
(A) ROC and PR AUCs for HSCZ with 5 Å (HSCZ_5) and 8 Å (HSCZ_8) context regions, AlphaMissense, and ESM1b in distinguishing pathogenic from benign ClinVar variants. The dataset was balanced to include 14,771 pathogenic and 14,771 benign variants. Error bars represent the standard deviation of AUCs computed from 1,000 bootstrap resamples. Pairwise empirical p-values (FDR-corrected) indicate that HSCZ_5 and HSCZ_8 are statistically significantly different from ESM1b and AlphaMissense for both ROC and PR AUCs. Even though HSCZ was not trained on this task, it performs nearly as well as these trained deep-learning models. (B) ROC curves comparing HSCZ_5, HSCZ_8, AlphaMissense, and ESM1b. AlphaMissense performs best in the high-confidence region, while HSCZ performs better at higher false positive rates. (C) ROC curves for logistic regression models combining HSCZ_5 with AlphaMissense and ESM1b using five-fold cross-validation. Integrating HSCZ with these state-of-the-art models further improves pathogenicity prediction.

Fine-tuning protein language models (PLMs) using Human Spatial Constraint (HSC) improves protein fitness predictions

In this section, we demonstrate that HSC scores are useful for an additional variant effect prediction task and illustrate an approach for incorporating HSC scores into protein language models (PLMs) via fine-tuning.

Variant classifiers trained on existing labels may inherit biases from human curation, limiting their generalizability[30]. In contrast, constraint scores are not trained, and PLMs are pre-trained on large-scale, unlabeled, naturally occurring protein sequences, reducing reliance on curated datasets and mitigating these biases. To move beyond the simple regression approach used above for combining HSC with PLM-derived scores, we hypothesized that directly incorporating intraspecies spatial constraint information into the model parameters of a PLM would improve its performance on downstream tasks.

We fine-tuned ESM2[31], a family of protein language models, using HSC scores as supervision labels. Since we wanted ESM2 to retain its pre-trained knowledge of constraints across species, we applied low-rank adaptation (LoRA)[32], which injects trainable rank decomposition matrices, allowing the original model weights to be frozen, thereby mitigating catastrophic forgetting issues. For each protein sequence passed through ESM2, we first compute the log-likelihood ratio (LLR) of amino acid substitutions using log-probabilities from the model logits. We applied a softmax over these LLRs to obtain the substitution probability distribution and calculated Shannon entropy of each position as a measure of local sequence constraint (Figure 5A; Methods). The model was then trained to minimize a listwise ranking loss between predicted and observed HSC values. We performed this fine-tuning for four ESM2 variants containing 8 M, 25 M, 150 M, and 650 M parameters.

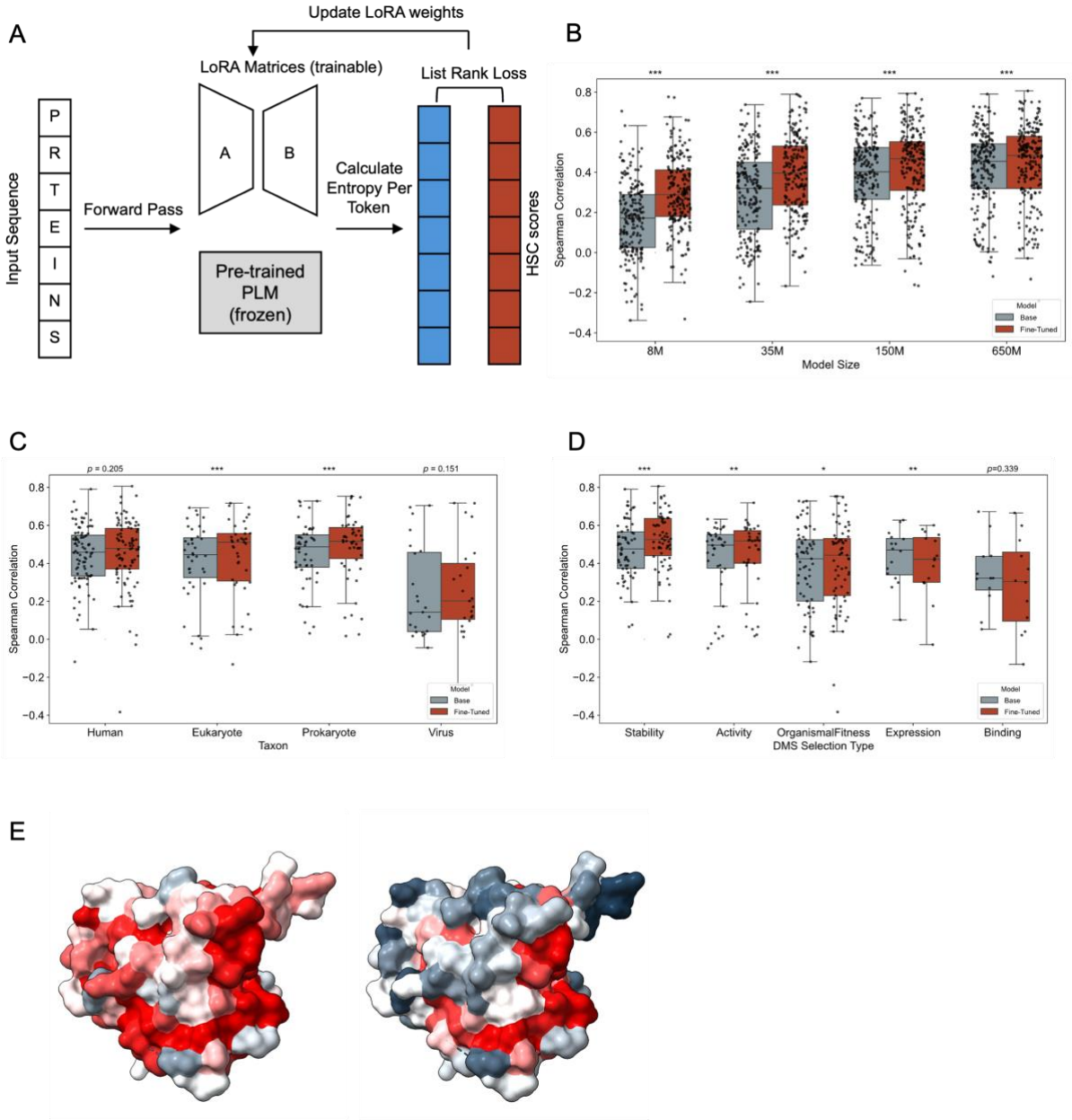

Figure 5. Fine-tuning ESM2 with HSC scores improves protein fitness predictions. (A) Schematic of the fine-tuning strategy, where Human Spatial Constraint (HSC) scores were used to fine-tune ESM2 models while preserving its pretrained evolutionary knowledge via low-rank adaptation (LoRA). (B) Comparison of Spearman correlations between model predictions and experimentally measured deep mutational scanning (DMS) scores for 202 proteins across ESM2 models of varying parameter sizes (8M, 35M, 150M, and 650M) before and after fine-tuning. Fine-tuned models show consistent performance gains across all sizes (Wilcoxon signed-rank test; *** $p < 0.001$). (C) Improvements in Spearman correlation with DMS data extend beyond human proteins for ESM2 650M. Significant improvements are observed for eukaryotic ($n = 39$, $p = 9 \times 10^{-4}$) and prokaryotic ($n = 50$, $p = 1 \times 10^{-4}$) datasets, while human ($n = 89$, $p = 0.21$) and viral ($n = 23$, $p = 0.15$) datasets show no statistically significant change. DMS datasets are stratified by each protein's taxon of origin. (D) Fine-tuned ESM2 650M achieves significant gains for stability ($n = 66$, $p < 1 \times 10^{-4}$), enzymatic activity ($n = 38$, $p = 4.5 \times 10^{-3}$), and organismal fitness ($n = 69$, $p = 0.036$),

whereas expression ($n = 16$, $p = 4.2 \times 10^{-3}$) and binding ($n = 12$, $p = 0.34$) assays show mixed results. (E) Visualization of entropy values derived from softmax-transformed log-likelihood ratio (LLR) scores across the p53 structure, highlighting increased sensitivity of the fine-tuned model to functionally constrained regions compared to the base model.

To illustrate the benefits of our approach for another protein variant effect prediction task, we evaluated the ability of the fine-tuned models to predict the effects of variants on a range of protein properties as estimated by deep mutational scanning (DMS) experiments. Fine-tuning ESM2 with HSC scores improves missense variant effect prediction across all model sizes (Fig. 5B). For example, the median Spearman correlation between model predictions and DMS scores increased from of the predictions of the 8M parameter model with the DMS values improves from 0.17 to 0.29 for the 8M model, 0.32 to 0.40 for the 35M model, 0.40 to 0.47 for the 150M model, and 0.45 to 0.48 for the 650M model (Wilcoxon signed-rank test, $p < 0.001$ for all). As expected, the relative improvement from fine-tuning is smaller for larger models, but the gains remain statistically significant even for the largest (650M) model evaluated.

Notably, this improvement extends beyond human proteins. When stratifying DMS datasets by taxon, we observe significant increases in predictive performance in eukaryotic (median $\rho = 0.45$ to 0.51, $p = 9 \times 10^{-4}$) and prokaryotic (0.49 to 0.52, $p = 1 \times 10^{-4}$) proteins, proteins after fine-tuning (Fig. 5C). This suggests that intraspecies spatial constraints capture fundamental aspects of protein function that generalize across evolutionary lineages. Furthermore, when categorizing DMS datasets by assay type, fine-tuning led to the most significant improvements for stability (0.48 to 0.53, $p < 1 \times 10^{-4}$), enzymatic activity (0.50 to 0.52, $p = 4.5 \times 10^{-3}$), and organismal fitness (0.42 to 0.44, $p = 0.036$), whereas expression and binding assays showed mixed or no improvement (Fig. 5D).

One striking example where the fine-tuned model demonstrates significantly improved performance is on a DMS of p53, a tumor suppressor protein that functions as a transcription factor[33]. This DMS quantifies the effects of variants on cell viability under treatment with Nutlin, a p53-activating compound that disrupts its interaction with an E3 ubiquitin-protein ligase MDM2. To visualize this difference, we overlay the entropy of the softmax-transformed LLR scores across protein positions (Fig. 5E). Compared to the base model, the fine-tuned model exhibits higher entropy at numerous residues, suggesting a refined ability to capture functional variability and constraint within p53.

3 Discussion

Quantifying mutational constraint within the human population can provide valuable insights into the functional importance of genomic regions. Existing methods that quantify such constraints within the proteome show moderately high correlation with pathogenicity. While deep learning models such as protein language models (PLMs) outperform conservation and constraint metrics in pathogenicity classification of variants, they do not directly account for within-human population variation. To address these limitations, we developed the Human Spatial Constraint (HSC), a new metric that integrates large-scale human genome and exome sequencing data with protein structure predictions to quantify constraint in 3D spatial regions. HSC outperforms previous population-based metrics and achieves performance comparable to deep learning models without requiring any costly training. Furthermore, by fine-tuning PLMs with HSC scores, we demonstrate improved model performance, suggesting that this approach reinforces and refines their pretrained evolutionary knowledge.

The HSC framework quantifies evolutionary constraint by integrating a protein's 3D spatial regions, allele frequency information, and background mutation rates. Since spatially proximal residues often share similar selective pressures, incorporating 3D context could provide a more biologically meaningful representation of constraint across the proteome. It also increases power by considering variation patterns across multiple sites. Previous methods, including our earlier approach (COSMIS), primarily considered the presence or absence of variants and thus overlooked valuable information contained in variant frequencies[2]. Since allele frequency reflects the relative tolerance of a site to mutation, we hypothesized that incorporating these distributions ensures that both the occurrence and prevalence of variants contribute to a more representative constraint estimation. Consistent with expectations, HSC scores reveal stronger constraint in essential and disease-associated genes. HSC scores also outperform other conservation metrics in classifying pathogenic variants. Together, these findings support our hypothesis that combining structural context with allele frequency information yields a more informative and robust measure of mutational constraint than existing approaches.

The size of the 3D spatial region used for constraint quantification introduces a tradeoff between power and specificity. Larger spatial regions capture broader structural context, potentially increasing power to detect constrained regions, but may also dilute site-specific signals by averaging across residues under differing selective pressures. Consistent with this tradeoff, we find that incorporating 3D context improves precision for the most highly constrained amino acids, yet scores computed with a 5 Å window perform similarly to those without spatial context for pathogenicity classification. This aligns with previous findings that expanding spatial context yields limited gains in pathogenicity prediction[22]. Overall, while structural information enhances the detection of strongly constrained sites, smaller spatial regions may be required to preserve specificity in constraint estimation.

We also demonstrate that HSC scores can serve as complementary features to existing conservation and constraint metrics in machine learning frameworks. They show only weak correlation with interspecies conservation scores, indicating that they capture distinct aspects of selective pressure. Consistent with this complementarity, a simple logistic

regression model combining HSC with deep learning–based predictions from AlphaMissense or ESM1b yields improved performance over either method alone. These results highlight the value of incorporating population-based structural constraints into predictive models of variant pathogenicity.

We also present a more sophisticated yet straightforward integration of HSC scores into protein language models (PLMs), which are not typically trained to capture patterns of sequence variation within human populations. To incorporate this information, we used HSC scores as a supervision signal to fine-tune PLMs without altering the model architecture. We also applied LoRA-based fine-tuning, making this approach easily transferable to other PLMs. Interestingly, performance gains diminished with increasing model size, suggesting that larger models encode aspects of population-level constraint derived from diverse evolutionary data. Although we anticipated improvements would be confined to human proteins, the benefits generalized across taxa, likely reflecting conserved patterns of constraint that are more easily captured in population-scale variation data. Improvements were most pronounced for functional assays related to stability, organismal fitness, and activity, which may be more directly linked to evolutionary constraint than binding or expression. Since foundation model embeddings are used for numerous downstream applications, embeddings derived from these fine-tuned models may provide richer representations of proteins compared to their base counterparts.

Our work has several limitations. First, HSC currently applies a linear weighting of allele frequencies, even though the relationship between allele frequency and variant effect size is inherently nonlinear. Incorporating a nonlinear weighting scheme (such as logarithmic scaling) could enhance sensitivity to low-frequency variants while reducing the influence of high-frequency outliers. Second, in our neutral model, single-nucleotide polymorphisms (SNPs) are permuted based on mutation rates, and their frequencies are randomly sampled from the observed allele frequency spectrum. A more principled simulation approach that accounts for the site frequency spectrum could improve the accuracy of neutral expectations. Finally, although incorporating structural context provides a biologically relevant representation of the neighborhood of individual residues, HSC only considers static structures and does not currently account for local folding dynamics or conformational flexibility, which can substantially influence the degree of constraint inferred for a given site.

Future iterations of this framework could incorporate an adaptive context window that uses a set of pre-defined local structural units, allowing the size of the 3D region to better reflect the local fold and differences in statistical power. Additionally, tools that generate conformational ensembles rather than static protein structures could further improve constraint estimation. The HSC framework can also be readily extended to additional species, enabling systematic comparisons of intraspecies constraint across the tree of life and providing new insights into the evolution of mutational tolerance. Similarly, the fine-tuning strategy we introduce can be applied to more sophisticated protein language models, offering a flexible approach to integrate population-level information into increasingly powerful architectures. Together, these extensions have the potential to enhance biological resolution and provide improved tools for understanding protein evolution and interpreting variant effects.

4 Methods

Mapping human variants to 3D protein structures

We followed the variant mapping approach developed in our previous COSMIS method[2]. We obtained a JSON file from github.com/capralab/cosmis that provides genetic variant counts and types (missense or synonymous) across 141,456 individuals from gnomAD v2.1.1 for all positions in the human proteome[14]. Briefly, variant statistics were extracted from GRCh38-lifted gnomAD v2.1.1. Using vcftools, sites with a FILTER flag other than PASS were removed, and only single-nucleotide variants were retained. For each human protein, we used its UniProt ID to retrieve the corresponding Ensembl transcript IDs via the UniProt ID Mapping service. Only transcripts with complete and valid coding sequences (CDSs), as confirmed by in silico translation, were retained. When multiple transcripts were valid, the transcript with the largest number of observed variable sites was selected.

3D protein structures were taken from the AlphaFold database version 2022-11-01[24]. For each codon, we took the minor allele frequency (MAF) of the observed variants and aggregated these values per codon position, generating dictionaries of synonymous and missense variant frequencies. 3D spatial neighborhoods for each residue were defined by parsing protein structures using Biopython's PDBParser and computing distances between the Cα atoms for all residues. Each residue was annotated with the number of residues within a given distance (e.g., 8 Å).

Estimating codon-specific synonymous and missense mutability

We estimated sequence context-dependent codon mutability using precomputed trinucleotide mutation rates obtained from github.com/capralab/cosmis. For each codon in the coding sequence, all possible single-nucleotide substitutions were considered. To account for local sequence effects, the substitution probability of each nucleotide was determined by the mutation rates specific to its trinucleotide context, defined as the codon plus its immediately adjacent upstream and downstream nucleotides. Each potential substitution was classified as synonymous or missense, excluding substitutions that generated stop codons. For codons at the termini of the coding sequence, where only two nucleotides could be mutated within the defined context, probabilities were restricted accordingly. Mutability contributions from all three possible substitutions at each nucleotide were summed across the codon, yielding a per-codon pair of synonymous and missense mutation probabilities. This procedure was applied to all codons in the transcript, producing a nucleotide sequence context-dependent estimate of expected synonymous and missense mutability for each amino acid in the protein.

Estimating per-transcript expected synonymous and missense variant counts

We estimated the expected number of synonymous and missense variants per transcript using a regression-based framework that models the relationship between mutability and observed variant counts under minimal selective pressure. Briefly, we computed transcript-level synonymous and missense mutability by summing codon-specific context-dependent mutabilities. We then fitted a linear regression of synonymous variant counts on synonymous mutability across transcripts, leveraging the fact that synonymous variation is largely neutral. The resulting regression model provided an expectation of variant counts under minimal constraint, which we applied to per-transcript missense mutability to obtain expected missense variant counts.

Simulating a null distribution of site-level mutational load

We implemented a permutation-based simulation procedure to derive the null distribution of expected total variant frequencies for a protein site. For each transcript, we normalized the missense mutability for each codon. In each permutation, we simulated *m* missense variants, where *m* is the expected count derived from the regression model, by sampling sites with replacement, with the probability of selecting site *i* proportional to its normalized missense mutability. At each site, a minor allele frequency (MAF) was randomly drawn from the empirical distribution of observed MAFs within the transcript and added to the site's mutational burden. This procedure was repeated 10,000 times to generate an empirical null distribution of site-level mutational loads.

Calculation of Human Spatial Constraint (HSC) scores

To quantify HSC scores for each residue within a protein, we considered its 3D spatial region. We quantified constraint by comparing the total observed frequency of missense variants to a permutation-based null distribution. The observed total was calculated as the total frequency of missense variants across all residues within its 3D spatial region. To generate the null, we performed 10,000 permutations of variant placements as described above.

The empirical *p*-value was then computed as:

$$p = \frac{n+1}{N+1}$$

where (n) is the number of permutations in which the simulated missense count was less than or equal to the observed count, and (N) is the total number of permutations (10,000). Adding one to both the numerator and denominator reflects the fact that under the null the observed count is also a realization of the same random variable as the permutation; this also ensures that *p*-values are never zero. We refer to this as the human spatial constraint p-value (HSCP).

To capture the magnitude of deviation from expectation, we also calculated a standardized Z-score:

$$Z = \frac{O - \mu null}{\sigma null}$$

where (O) is the observed missense count, $\mu null$ is the mean missense count from the null distribution, and $\sigma null$ is its standard deviation. Negative Z-scores indicate fewer observed variants than expected (constraint), whereas positive scores indicate tolerance. To improve interpretability and mitigate extreme values, we applied a log transformation, yielding the human spatial constraint Z-score (HSCZ).

We provide both HSCP and HSCZ scores because they offer complementary advantages in interpreting evolutionary constraints. HSCP, as a p-value, is bounded between 0 and 1, making it well-suited for visual analysis and threshold-based classification. However, since it is derived from 10,000 permutations, many positions share identical p-values, limiting its resolution. In contrast, HSCZ, a log1p-transformed z-score, provides a continuous measure with a broader dynamic range, allowing finer distinctions between constraint levels. Together, these scores enable both intuitive interpretation and precise differentiation of constrained residues.

Protein functional properties and other constraint scores

To evaluate the HSC scores, we obtained a curated gene set representing different levels of essentiality from the MacArthur Lab's *gene_lists* repository. These included haploinsufficient, essential, nonessential, autosomal dominant, autosomal recessive, and olfactory receptor genes, many of which were also used to benchmark the pLI metric. Genes were identified by HGNC symbols and mapped to UniProt accession numbers using UniProt's API. We calculated HSC scores for all positions for these genes. These included 141 olfactory genes, 624 non-essential genes, 904 recessive genes, 469 dominant genes, 590 essential genes, and 217 haploinsufficient genes.

Evaluations and comparisons of methods

We evaluated interspecies and intraspecies conservation/constraint metrics on the same dataset curated by our lab previously. Briefly, this dataset was retrieved from ClinVar[34] in August 2021. The dataset only includes missense variants labeled as "Pathogenic", "Likely pathogenic", or "Pathogenic/Likely pathogenic" for true positive (pathogenic) variants and "Benign", "Likely benign", or "Benign/Likely benign" for true negative (benign) variants. Only variants with at least one-star review status and no conflicting interpretations were included. The ClinVar dataset for comparison with Esm1b and AlphaMissense was obtained from github.com/ntranoslab/esm-variants. This dataset was filtered for unique reference residue within canonical isoforms. The dataset was balanced by random downsampling of the majority class. We also trained logistic regression models using five-fold stratified cross-validation with standardized features. ROC curves were generated for individual predictors and combined models (e.g., HSC + AlphaMissense or HSC + ESM-1b), and mean AUCs were reported to quantify predictive performance. For each metric (ROC AUC and PR AUC), we performed 1,000 bootstrap resamples to estimate the distribution of performance for each score. Pairwise empirical *p*-values were computed by comparing bootstrap distributions, and false discovery rate (FDR) correction was applied using the Benjamini–Hochberg method to account for multiple testing.

We compared the Spearman correlation between HSC and several conservation metrics to assess its similarity in capturing evolutionary constraint. This dataset was curated by our lab previously. Human-variation–based metrics included RVIS[16], missense Z[15,17] score, pLI[15], MTR[1], and MTR3D[21], while inter-species conservation metrics included GERP++[35], phyloP[36], phastCons[37], and ConSurf[38]. ConSurf scores were computed using Rate4Site with default parameters and the 100-way vertebrate alignment from the UCSC Genome Browser. Relative solvent accessibility was estimated using DSSP within Biopython.

Protein language model fine-tuning with HSC scores

All human proteins with available HSC scores were randomly divided into training (75%) and validation (25%) sets. We fine-tuned ESM2 models using Low-Rank Adaptation (LoRA), applied only to the self-attention modules. Specifically, LoRA adapters (rank = 8, $\alpha = 64$, dropout = 0.05) were inserted into the query, key, and value projection layers of the attention blocks, following the PEFT framework (task type: *feature extraction*). LoRA introduces low-rank updates to selected layers instead of modifying all weights, substantially reducing the number of trainable parameters while preserving model capacity. This makes fine-tuning computationally efficient and memory-friendly, especially for large models.

Protein sequences were filtered to include only those with at least 25 constrained residues (HSCZ < 0) and a median HSCZ score < 0, ensuring an adequate signal of constraint for training. The filtered dataset was split into training (75%) and validation (25%)

sets. Fine-tuning was performed with a batch size of 1 using the AdamW optimizer (learning rate = 1×10^{-5}, weight decay = 1×10^{-2}) on an NVIDIA A100 GPU.

For each batch, we computed position-specific log-likelihood ratio (LLR) scores from model logits by normalizing the wild-type amino acid probability at each position and converting these to entropy values across the 20 amino acids. The normalized entropy, representing model uncertainty, was compared to HSCZ scores using a listwise ranking loss applied only to residues with negative COSMIS values. To avoid overfitting, we implemented early stopping with validation checks based on mean validation loss. The model with the lowest validation loss was saved as the final fine-tuned checkpoint.

The base and fine-tuned models were evaluated using deep mutational scanning (DMS) data from the ProteinGym benchmark. We restricted the dataset to 202 proteins with sequence lengths ≤1022 amino acids to ensure compatibility with model input constraints. For each protein, log-likelihood ratio (LLR) scores were computed by passing the wild-type sequence through the model. In cases where multiple variants corresponded to the same residue, their scores were averaged to obtain a single representative value per position.

Data availability

All datasets generated and code used in this study will be made available soon at the HSC repository (GitHub: [github.com/gyasu/HSC]).

4.1 Competing interests

The authors declare no competing interests.

References

1. Silk, M., Petrovski, S. & Ascher, D. B. MTR-Viewer: identifying regions within genes under purifying selection. *Nucleic Acids Res.* **47**, W121–W126 (2019).
2. Li, B., Roden, D. M. & Capra, J. A. The 3D mutational constraint on amino acid sites in the human proteome. *Nat. Commun.* **13**, 3273 (2022).
3. Ng, P. C. & Henikoff, S. Predicting Deleterious Amino Acid Substitutions. *Genome Res.* **11**, 863–874 (2001).
4. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18**, S71–S77 (2002).
5. Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).

6. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
7. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
8. 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. *Nat. Genet.* **55**, 1512–1522 (2023).
9. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. Preprint at https://doi.org/10.1101/2021.07.09.450648 (2021).
10. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Preprint at https://doi.org/10.1101/622803 (2019).
11. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
12. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
13. Sun, K. Y. *et al.* A deep catalogue of protein-coding variation in 983,578 individuals. *Nature* **631**, 583–592 (2024).
14. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
15. Exome Aggregation Consortium *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
16. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet.* **9**, e1003709 (2013).
17. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).

18. Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S. & Goldstein, D. B. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol.* **17**, 9 (2016).
19. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
20. Hayeck, T. J. *et al.* Improved Pathogenic Variant Localization via a Hierarchical Model of Sub-regional Intolerance. *Am. J. Hum. Genet.* **104**, 299–309 (2019).
21. Silk, M. *et al.* MTR3D: identifying regions within protein tertiary structures under purifying selection. *Nucleic Acids Res.* **49**, W438–W445 (2021).
22. Kovacs, A. S., Portelli, S., Silk, M., Rodrigues, C. H. M. & Ascher, D. B. MTR3D - AF2 : Expanding the coverage of spatially derived missense tolerance scores across the human proteome using ALPHAFOLD2. *Protein Sci.* **33**, e5112 (2024).
23. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
24. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
25. Chandra, V. *et al.* Structure of the intact PPAR-γ–RXR-α nuclear receptor complex on DNA. *Nature* **456**, 350–356 (2008).
26. Majithia, A. R. *et al.* Rare variants in *PPARG* with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc. Natl. Acad. Sci.* **111**, 13127–13132 (2014).
27. Barroso, I. *et al.* Dominant negative mutations in human PPARγ associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature* **402**, 880–883 (1999).
28. Jeninga, E. H., Gurnell, M. & Kalkhoven, E. Functional implications of genetic variation in human PPARγ. *Trends Endocrinol. Metab.* **20**, 380–387 (2009).
29. Chawla, A., Repa, J. J., Evans, R. M. & Mangelsdorf, D. J. Nuclear Receptors and Lipid Physiology: Opening the X-Files. *Science* **294**, 1866–1870 (2001).

30. Grimm, D. G. *et al.* The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015).
31. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
32. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2106.09685 (2021).
33. Giacomelli, A. O. *et al.* Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.* **50**, 1381–1387 (2018).
34. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
35. Davydov, E. V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
36. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
37. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
38. Armon, A., Graur, D. & Ben-Tal, N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**, 447–463 (2001).