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.

Multi-Scale Sequence Encoding Distinguishes Long-Lived and Short-Lived Proteins Revealed by Protein Language Model Embeddings

Tangilal Dihan Chowdhury
Military Institute of Science and Technology
Fasiha Tanzeem Taiba
Military Institute of Science and Technology
Md Ushama Shafoyat
Military Institute of Science and Technology
Maruf Hasan
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

Research Article

Keywords: PLM, long lived protein, short lived protein, PLM embeddings, sequence grammar, motif

Posted Date: May 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9298535/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

Protein stability and turnover are fundamental determinants of proteome regulation, yet how protein lifetime is specified by amino-acid sequences remains incompletely understood. Here, we identify a previously uncharacterized, multi-scale organization of sequence features associated with protein stability using protein language model (PLM) representations. Using experimentally derived half-life data from four human cell lines, we uncover a conserved stability-associated axis in embedding space that separates long-lived proteins (LLP) from short-lived proteins (SLP) (ROC–AUC = 0.80–0.82) and generalizes across species (Pearson r up to 0.465). Systematic decomposition of sequence representations shows that protein stability is not explained by amino-acid composition alone but reflects contributions from multiple sequence levels. While composition provides baseline predictive signal (AUC = 0.64–0.70), sequence grammar (0.65–0.68) and motif features (0.71–0.73) contribute additional information, with PLM representations integrating these signals into a unified framework. Disrupting residue order reduces predictive performance (AUC ~ 0.82 to ~ 0.62), supporting a role for residue organization beyond composition alone. Analysis of sequence features reveals consistent organization patterns associated with stability, with LLP proteins exhibiting tighter lysine spacing, enhanced charge clustering, and increased proline periodicity, whereas SLP proteins show more dispersed residue organization. Perturbation analyses further support the contribution of these features, with disruption of charge organization producing the largest decrease in predicted stability ($\Delta \approx -1.53$). Together, these findings support a model in which protein stability is associated with distributed, multi-scale sequence organization rather than isolated motifs. By systematically resolving stability-related signals across sequence scales, this work highlights the potential of PLM-based representations to uncover biologically meaningful principles governing proteome dynamics.

1. Introduction

Proteins within cells are continuously synthesized and degraded, and the balance between these processes determines protein lifetime, a key regulator of signaling, metabolism, and protein quality control [1]. A central pathway governing targeted degradation is the ubiquitin–proteasome system, which selectively labels proteins for destruction [2]. Despite extensive research, predicting protein lifetime directly from amino-acid sequence remains a major challenge in molecular biology and computational proteomics.

Although proteomics approaches enable large-scale measurement of protein turnover, they do not explain the substantial variability in half-life observed across the proteome. In particular, it remains unclear to what extent intrinsic sequence features contribute to degradation behavior. Existing studies have identified sequence elements such as degron motifs, PEST regions, and compositional biases associated with protein stability [3–5]. However, these features do not fully account for observed variability: many unstable proteins lack known degrons, while many proteins containing such motifs remain stable [4, 6]. This discrepancy highlights an open question as to whether protein stability is

primarily associated with localized sequence motifs or with broader, distributed patterns of residue organization across entire sequences.

Recent advances in protein language models (PLMs) provide a framework to investigate this question. Trained on large-scale sequence data, PLMs capture complex statistical relationships between amino acids and encode diverse structural and functional properties [7–9]. However, whether these representations capture stability-associated signals, and how such signals are organized within sequences, remains incompletely understood.

Here, we examine this problem using experimentally derived half-life datasets from four human cell lines [10]. We identify a stability-associated axis in PLM embedding space that separates long-lived proteins (LLP) from short-lived proteins (SLP) and shows consistent behavior across datasets and species. By systematically decomposing sequence representations, we find that protein stability is not explained by amino-acid composition alone but reflects contributions from multiple sequence levels, including composition, sequence grammar, and local motifs. Disrupting residue order reduces predictive performance, supporting a contribution of residue organization beyond composition alone.

Analysis of sequence features reveals consistent organization patterns associated with stability, including tighter lysine spacing, enhanced charge clustering, and increased proline periodicity in LLP proteins relative to SLP proteins. Perturbation analyses further support the contribution of these features, with disruption of charge organization producing the largest decrease in predicted stability.

Together, these results support a model in which protein stability is associated with distributed, multi-scale sequence organization rather than isolated motifs. This framework provides a basis for interpreting variability in proteome-wide turnover and highlights the potential of PLM-based representations to uncover biologically meaningful sequence principles underlying protein stability [11].

2. Materials and Methods

2.1 Protein Datasets and Sequence Processing

Protein half-life data were obtained from a large-scale quantitative proteomics study [10] that systematically mapped short-lived proteins in human cells using cycloheximide-chase assays combined with multiplexed mass spectrometry. The dataset contains measurements from four human cell lines (U2OS, HEK293T, HCT116, and RPE1). Each cell line dataset was curated by removing redundant protein sequences to reduce sequence redundancy and minimize similarity bias. For stability classification, proteins were categorized as long-lived (LLP; 25–32 h) or short-lived (SLP; 0.2–2.5 h). For each cell line, 300 LLP and 300 SLP proteins were selected to construct balanced datasets (600 proteins per cell line; 2400 proteins total across four cell lines). For continuous stability prediction analyses, the full half-life datasets (~ 1500 proteins per cell line; ~6000 proteins total) were used, spanning a wide range of degradation rates.

Protein sequences were retrieved in FASTA format from UniProt [12] and used for subsequent computational analyses. To assess stability-axis similarity and cross-dataset generalization, an independent turnover dataset obtained using Stable Isotope Labeling by Amino Acids in Cell Culture mass spectrometry [13, 14] was also analyzed. This dataset includes half-life measurements for 4,106 proteins in the human HeLa cell line and 3,528 proteins in the mouse C2C12 cell line (7,634 proteins total). Because experimental methodologies differ substantially between cycloheximide-chase and SILAC approaches [13], the datasets were analyzed separately to avoid methodological bias in half-life estimates. Overall, the analyses include more than 13,000 proteins across multiple independent proteomic datasets. All datasets, including half-life measurements and FASTA sequences, are provided in **Supplementary File 1**.

2.2 Protein Language Model Embeddings

Protein sequences were encoded using the ESM-2 protein language model with 650 million parameters (ESM2-t33-650M) [6–8]. For each protein sequence, token-level representations were extracted from the final 33rd transformer layer [7, 8]. Residue-level embeddings [15, 16] were averaged across all sequence positions to produce a single fixed-length vector representation for each protein. If E_i represents the embedding vector of residue i and L is the sequence length, the final protein embedding (X) was computed as:

$$X = \frac{1}{L} \sum_{i=1}^{L} E_i$$

2.3 Dimensionality Reduction

Because the original embedding vectors are high-dimensional, dimensionality reduction was performed prior to modeling [17]. First, all features were standardized using z-score normalization:

$$Z = \frac{X - \mu}{\sigma}$$

where X represents the original embedding feature, u is the feature mean, and σ is the standard deviation. Principal component analysis (PCA) was then applied to the normalized embeddings to capture the major sources of variation in the data. The top 30 principal components (PCs) were retained and used as the feature space for all downstream analyses. The number of principal components (30) was selected based on cumulative explained variance 81.7%. The embedding features and the python script of extracting embedding features from PLM model (ESM 2) for all datasets are provided in **Supplementary File 2**.

2.4 Stability Score Computation and LLP–SLP Classification

Protein stability was inferred from protein language model embeddings [15, 16] using a logistic regression classifier trained on PCA-transformed embeddings (top 30 PCs). Logistic regression was implemented using scikit learn, with L2 regularization (C = 1). The model estimates the probability that a protein belongs to the LLP class as:

$$P(y = 1|X) = \frac{1}{1 + e^{-(w.X+b)}}$$

where X is the PCA feature vector, w the weight vector, and b the intercept. All preprocessing steps, including normalization and PCA, were performed within each training fold to prevent data leakage. Model performance was evaluated using ROC–AUC. The classifier defines a linear stability axis in embedding space. Each protein was projected onto this direction to compute a Stability Score (SSS) [18]:

$$SSS = w.\,X + b$$

Higher SSS values correspond to proteins predicted to be more stable. To avoid overfitting, SSS values were computed using out-of-fold predictions during cross-validation [19, 20]. To assess potential protein length bias, ROC–AUC values were compared between the original and length-matched datasets using five-fold cross-validation [20], with differences evaluated using a paired t-test [21] and Cohen's d [22]. The datasets and the python scripts of these analysis are provided in **Supplementary File 3**.

2.5 Sequence shuffling and sliding-window analysis

To evaluate whether stability prediction depends on sequence order rather than amino acid composition, protein sequences were randomly shuffled while preserving composition [27]. Embeddings were recomputed from the shuffled sequences and classification performance was reassessed; a decrease in performance suggests that residue order contributes to the stability signal.

To determine whether stability signals arise from localized motifs or distributed sequence patterns, a sliding-window analysis was performed [28]. Each protein was divided into overlapping 25-residue windows, embeddings were computed for each window, and projected onto the stability axis. For each protein, the maximum, mean, and variance of local stability scores were calculated to assess whether stability signals are concentrated in specific regions or distributed across the sequence. All the python analysis scripts are provided in **Supplementary File 6**.

2.6 Sequence grammar and charge clustering analysis

Sequence grammar features were analyzed for long-lived proteins (LLP) and short-lived proteins (SLP) across four cell lines using protein sequences obtained from FASTA files and custom Python scripts. Charge clustering features included lysine cluster size (K_max_cluster) and acidic cluster size (DE_max_cluster), defined as the longest consecutive runs of K or D/E residues within a sequence. The number of lysine clusters (K_cluster_count) was calculated as the total number of contiguous lysine segments, while charge blockiness was measured as the fraction of adjacent charged residues sharing the same charge sign [29].

Additional interpretable sequence grammar features were computed to investigate stability-related organization patterns. This included lysine spacing (mean distance between consecutive lysines), proline periodicity estimated using Fourier-based spectral analysis [30], PEST enrichment [5], proline-rich fraction, and low-complexity regions identified using sliding-window scans [3, 4]. Disorder positioning was calculated from IUPred scores [31] as the difference between C-terminal and N-terminal disorder fractions [4]. Feature distributions between LLP and SLP proteins were compared using the Mann–Whitney U test [32], and mean values were calculated for each cell line. The IUPred scores (. result file) for all the datasets and the python scripts for sequence grammar and charge clustering analysis are provided in **Supplementary File 7**.

2.7 Feature-specific perturbation analysis

Feature-specific perturbation analysis was performed to evaluate the contribution of sequence organization features identified in the study. For each protein sequence, selected sequence organization features were disrupted by residue substitution while preserving overall sequence length and amino acid composition where applicable. Perturbed sequences were re-embedded using the same protein language model, and stability scores were recomputed using the previously defined stability axis. Changes in stability were quantified as the difference between original and perturbed scores (Δ stability score), and mean effects were calculated across all proteins. The python script of this analysis is provided in **Supplementary File 9**.

2.8 Axis similarity and continuous half-life prediction

To assess whether the stability-associated signal [23] captured by protein language model embeddings is conserved across biological contexts, stability axes derived from LLP and SLP proteins were compared across cell lines. Axis similarity was quantified using cosine similarity between logistic regression weight vectors [24], where higher similarity indicates a conserved direction separating stable and unstable proteins in embedding space. Generalization of this stability axis was further evaluated using full proteome datasets. Models trained on one cell line were used to predict stability in other cell lines, and predictions were compared with experimental half-life measurements using Pearson correlation [25]. Cross-experiment and cross-species generalization were also tested using independent mouse and human protein turnover datasets [13], which were analyzed separately due to differences in experimental protocols and half-life distributions. To examine whether embeddings capture continuous degradation rates, linear regression models were used to predict log-transformed protein half-life:

$$\hat{y} = w.X + b$$

where X represents the PCA-transformed embedding vector and y = log (half-life + 1). Model performance was evaluated using Pearson correlation, Spearman correlation, and coefficient of determination (R^2). The python scripts of these experiments are provided in **Supplementary File 4**.

2.9 Interpretation of embedding components and sequence feature correlations

To interpret the embedding dimensions contributing to stability prediction, principal components with the largest logistic regression weights were identified and analyzed. Correlations were computed between these components and sequence-derived properties [26] including protein length, lysine density, charge density, and hydrophobic residue fraction, enabling identification of sequence features underlying the stability grammar encoded in the embeddings. In addition, correlations were calculated between the stability score (SSS) and several sequence-level features across cell lines. Pearson correlations were computed for protein length, lysine density, and intrinsic disorder fraction to evaluate whether basic sequence properties contribute to the embedding-derived stability signal. Finally, correlations with structural sequence properties, including hydrophobic fraction, helix propensity residues, and beta-sheet propensity residues, were examined to assess whether classical structural features explain the observed stability patterns. All the python analysis scripts are provided in **Supplementary File 5**.

2.10 Grammar-based classification and perturbation analysis

To evaluate whether protein stability signals can be captured using interpretable sequence features, a classification model was trained using sequence grammar–derived features, including residue periodicity, clustering, spacing, and related metrics. Model performance was evaluated using ROC–AUC [33], and the approach was further tested on an independent dataset containing human and mouse proteins measured under the same experimental protocol. In this dataset, LLP and SLP classes were defined using the top and bottom 5% of the half-life distribution, and ROC–AUC and precision were used to assess the ability of grammar features to distinguish stability classes across species.

To test whether sequence organization contributes to stability prediction, in silico grammar perturbation experiments were performed. For each LLP sequence, residues at regular intervals (every 10th position) were randomly shuffled, disrupting periodic sequence organization while largely preserving amino acid composition and sequence length. Embeddings were recomputed for the perturbed sequences, and predicted stability values were compared with the original sequences using paired t-tests to determine whether disruption of sequence organization affects predicted stability. The python script of grammar-based classification for Cell 1–4, Human-external, Mouse and permutation analysis are provided in **Supplementary File 8**.

2.11 Functional enrichment analysis

Functional enrichment analysis was performed to identify biological processes associated with long-lived proteins (LLP) and short-lived proteins (SLP). Protein sequences were parsed from FASTA files, and gene names were extracted from UniProt headers (GN field). Gene lists were analyzed using the g:Profiler tool [34] (organism: *Homo sapiens*) to identify enriched Gene Ontology biological process (GO:BP) terms [35]. Enrichment results were filtered at $p < 0.05$, and top-ranking terms were used for interpretation and visualization. The analysis result is provided in **Supplementary File 10**.

2.12 Motif enrichment and motif-based classification

All possible tripeptide motifs ($20^3 = 8000$) were scanned across LLP and SLP protein sequences. Motif frequencies were computed for each sequence and compared between classes using the Mann–Whitney U test, with effect sizes estimated using Cohen's d. The top 20 motifs showing the largest effect sizes were selected as features for classification. A logistic regression model using these motif frequencies was trained to distinguish LLP and SLP proteins, and performance was evaluated using five-fold cross-validation with ROC–AUC and precision. To test whether global sequence organization contributes additional predictive information, the motif features were combined with previously defined sequence grammar features, and model performance was re-evaluated. All the dataset and python scripts are provided in **Supplementary File 11**.

2.13 Sequence representation benchmarking

Amino acid composition features were computed as the frequency of each of the 20 standard residues in a protein sequence. Stability prediction was performed using logistic regression to classify long-lived (LLP) and short-lived proteins (SLP), and performance was evaluated using ROC–AUC with five-fold stratified cross-validation. Models based on sequence grammar, motif features, and protein language model embeddings were constructed as described above. These representations were compared to assess their relative contribution to protein stability prediction. The python scripts of this analysis are provided in **Additional Information 12**.

3. Results

3.1 Identification of a Stability Axis in Protein Language Model Embeddings

Projection of protein embeddings revealed clear separation between LLP and SLP proteins along a derived stability score (SSS) axis (**Supplementary Figure S1A–C**). LLP proteins showed substantially higher scores (mean SSS = 0.8816) than SLP proteins (− 0.9317), with cross-validated means of 0.8199 and − 0.7749, respectively (Table 1). The separation was highly significant (Mann–Whitney $p = 3.84 \times 10^{-35}$; validated $p = 1.35 \times 10^{-26}$) with a large effect size (Cohen's d = 1.345; validated d = 1.118) (Table 1). Cross–cell line validation showed consistent performance when the model trained on Cell1 was applied to other datasets, achieving ROC–AUC values of 0.802, 0.811, and 0.817 for Cell2–Cell4, respectively (**Supplementary Table S1;** Fig. 1A–F). Control analysis confirmed that predictions were not driven by protein length. Model performance remained unchanged after length matching (AUC = 0.781 vs 0.787), with no significant difference (t = − 0.363, p = 0.735; Cohen's d = − 0.18) (**Supplementary Figure S2A–B; Supplementary Table S2**). These results indicate that protein language model embeddings capture a stability-associated direction that separates LLP and SLP proteins and generalizes across cell lines, independent of sequence length. These results indicate that protein language model embeddings capture a stability-associated direction that separates LLP and SLP proteins and generalizes across cell lines, independent of sequence length.

Table 1
Stability score comparison between LLP and SLP proteins.

Group	Mean SSS	Cross-validated-mean SSS
LLP	0.881652634863405	0.8199233676051194
SLP	−0.9317443608705182	−0.7749137369997359
Statistical significance of stability score		
Mann–Whitney p-value	3.835×10^{-35}	1.346×10^{-26}
Cohen's d	1.345	1.118

3.2 Stability Signals Depend on Sequence Organization

Sequence shuffling markedly reduced classification performance across all datasets (Table 2). While the original models achieved AUC = 0.802–0.823, shuffled sequences produced substantially lower values (0.615–0.639). Because amino acid composition was preserved during shuffling, this reduction suggests that residue order contributes to the stability-associated signal. Sliding-window analysis further showed that stability signals are distributed across the protein sequence rather than confined to short motifs (Table 3**)**. LLP proteins exhibited much higher maximum local stability scores (2.78 vs 0.69, p = 1.71×10^{-16}) and mean local scores (0.758 vs − 1.149, p = 6.60×10^{-18}) compared with SLP proteins, while variance remained similar (1.12 vs 1.07, p = 0.426). Together, these results indicate that protein stability is associated with distributed sequence organization, rather than being explained by amino acid composition alone or localized motifs.

Table 2
Effect of sequence shuffling

Cell Line	Original AUC	Shuffled AUC
Cell 1	0.823	0.632
Cell 2	0.802	0.639
Cell 3	0.811	0.621
Cell 4	0.817	0.615

Table 3
Stability window analysis

Cell	Feature	LLP Mean	SLP Mean	p-value
1	max_local_score	2.7777	0.6943	1.71×10^{-16}
	var_local_score	1.1195	1.0701	0.426
	mean_local_score	0.758	-1.1489	6.608×10^{-18}
2	max_local_score	2.491	0.5311	9.351×10^{-16}
	var_local_score	1.2108	1.2556	0.50708
	mean_local_score	0.5514	-1.2764	2.118×10^{-17}
3	max_local_score	2.819	0.6416	7.045×10^{-18}
	var_local_score	1.1254	1.2881	0.783
	mean_local_score	0.7754	-1.244	4.009×10^{-19}
4	max_local_score	2.6625	1.1451	5.232×10^{-11}
	var_local_score	1.0909	1.1901	0.995
	mean_local_score	0.7105	-0.7897	4.296×10^{-19}

3.3 Sequence-Level Correlates of the Stability Axis

Decoding of the dominant embedding components revealed a consistent charge–hydrophobic stability grammar across cell lines (**Supplementary Tables S5–S6**). The stability signal was primarily associated with lysine density, charge distribution, and hydrophobic composition, rather than sequence length. Strong relationships with charge organization were observed in Cell3 (PC7: lysine density r = − 0.450; charge density r = − 0.324), while hydrophobic fraction contributed positively in components such as Cell1 PC17 (r = 0.246) and Cell4 PC26 (r = 0.142). Direct correlations between sequence features and the stability score supported this interpretation (Table 4). Stability scores showed moderate associations with lysine density (r = 0.070–0.319) and protein length (r = 0.268–0.329), whereas intrinsic disorder showed weak correlations (r = − 0.086–0.171) (Fig. 2A–D). Correlations with predicted structural propensities were generally small (**Supplementary Table S7**). Together, these results suggest that the embedding-derived stability axis is associated with sequence-level organization of charged and hydrophobic residues, rather than classical structural features or simple compositional effects.

From a biological perspective, these features are consistent with known properties associated with protein degradation. Lysine residues serve as primary ubiquitin attachment sites, and their density and spatial distribution may influence modification patterns. Similarly, clustering of charged residues may affect local electrostatic environments relevant to protein–protein interactions. Hydrophobic composition may influence folding stability and exposure of sequence features. While not directly measuring degradation mechanisms, these observations indicate that the stability axis captures biologically meaningful sequence properties rather than purely abstract embedding patterns.

Table 4
Correlation between stability score and sequence features across cell lines

Cell Line	Length (r)	Lysine Density (r)	Disorder Fraction (r)
Cell1	0.326	0.180	−0.086
Cell2	0.329	0.070	−0.040
Cell3	0.268	0.319	0.012
Cell4	0.278	0.216	0.171

3.4 Sequence Grammar of Long-Lived and Short-Lived Proteins

Charge clustering analysis revealed consistent differences between LLP and SLP proteins across all cell lines (Table 5; Fig. 3A–D). LLP proteins showed larger lysine clusters (K_max_cluster = 2.18–2.41 vs 1.94–2.06) and larger acidic clusters (DE_max_cluster = 3.86–4.08 vs 3.26–3.34), together with substantially more lysine clusters overall (44.01–47.35 vs 28.33–33.68) ($p = 10^{-4}$–10^{-9}). In contrast, charge blockiness showed no significant differences between LLP and SLP proteins. These results indicate that the stability signal reflects enhanced charge clustering in long-lived proteins.

Additional sequence analysis further highlighted differences in lysine spacing and proline periodicity (Table 5). LLP proteins showed shorter lysine spacing (15.27–17.62 vs 17.27–20.20) and higher proline periodicity (15.14–16.08 vs 10.73–11.94) across cell lines, whereas features such as PEST score, proline-rich fraction, and low-complexity fraction showed no significant differences. The distributions of these patterns are shown in Fig. 4A–H and Fig. 5A–H.

Analysis of degron-like sequence patterns indicated that such motifs are broadly present in both LLP and SLP proteins. However, LLP proteins exhibited a modest but significant increase in motif density compared to SLP proteins (mean density: 0.027 vs 0.024; Mann–Whitney U test, $p = 2.48 \times 10^{-11}$) (**Supplementary Fig. 4**), suggesting that differences in stability are associated with motif distribution rather than simple presence.

Together, these results suggest that protein stability is associated with distributed residue organization, where LLP proteins exhibit stronger lysine and charge clustering together with periodic proline

organization, whereas SLP proteins display larger lysine spacing and weaker clustering patterns.

.

Table 5
Stability-associated sequence organization features across four cell lines

Category	Cell	Feature	LLP Mean	SLP Mean	p-value
Charge Clustering	1	K_max_cluster	2.24	1.9467	4.00×10^{-5}
		DE_max_cluster	3.8567	3.30	1.88×10^{-4}
		K_cluster_count	47.3467	28.33	3.14×10^{-9}
		charge_blockiness	0.5514	0.5629	0.263
	2	K_max_cluster	2.18	1.9433	2.00×10^{-4}
		DE_max_cluster	4.08	3.26	1.64×10^{-6}
		K_cluster_count	45.1067	29.22	4.06×10^{-8}
		charge_blockiness	0.5563	0.5550	0.98
	3	K_max_cluster	2.30	2.0033	2.82×10^{-6}
		DE_max_cluster	3.9367	3.29	2.20×10^{-6}
		K_cluster_count	46.0167	30.21	2.02×10^{-11}
		charge_blockiness	0.5624	0.5555	0.558
	4	K_max_cluster	2.4134	2.06	1.80×10^{-6}
		DE_max_cluster	3.9867	3.34	8.07×10^{-6}
		K_cluster_count	44.01	33.68	1.59×10^{-6}
		charge_blockiness	0.56523	0.5592	0.722
Sequence Grammar	1	PEST score	0.0502	0.0559	0.066
		Proline-rich fraction	0.0082	0.0074	0.805
		Low complexity fraction	0.0299	0.0355	0.065
		K spacing	17.619	20.199	0.0045
		Proline periodicity	16.075	11.227	1.34×10^{-4}
		Disorder shift	-0.0137	-0.0537	0.149
	2	PEST score	0.0588	0.0602	0.815
		Proline-rich fraction	0.0066	0.0075	0.240
		Low complexity fraction	0.0326	0.0371	0.733
		K spacing	16.168	17.270	0.272
		Proline periodicity	15.276	10.732	1.62×10^{-5}

Category	Cell	Feature	LLP Mean	SLP Mean	p-value
		Disorder shift	0.0224	-0.0380	0.0297
	3	PEST score	0.0569	0.0578	0.920
		Proline-rich fraction	0.0081	0.0078	0.375
		Low complexity fraction	0.0344	0.0360	0.492
		K spacing	15.273	18.498	3.57×10^{-4}
		Proline periodicity	15.139	11.865	0.0187
		Disorder shift	-0.0082	-0.0513	0.100
	4	PEST score	0.0588	0.0548	0.144
		Proline-rich fraction	0.0073	0.0067	0.891
		Low complexity fraction	0.0366	0.0343	0.548
		K spacing	16.550	20.194	0.0129
		Proline periodicity	15.708	11.938	0.0094
		Disorder shift	0.0266	-0.0134	0.181

3.5 Feature-specific perturbation analysis reveals dominant contribution of charge organization

To further dissect the contribution of individual sequence grammar components, feature-specific perturbation analysis was performed by selectively disrupting lysine distribution, charge organization, and proline periodicity (Table 6**)**. Disruption of charge organization resulted in the largest decrease in predicted stability, followed by lysine redistribution and proline perturbation, indicating a hierarchical contribution of sequence features (charge > lysine > proline) to the stability signal. These results provide computational support that the embedding-derived stability signal is influenced by specific sequence organization features. These findings identify charge organization as the dominant contributor to the stability signal, highlighting its prominent contribution to the stability-associated signal within sequence representations. The Python script for this analysis is provided in **Supplementary File 13**.

Table 6
Feature-specific perturbation analysis of sequence grammar components

Perturbation Type	Mean Stability Score	Δ Stability Score
Original	−1.03	—
Lysine disruption	−1.47	−0.44
Charge disruption	−2.56	−1.53
Proline disruption	−1.18	−0.15

3.6 Conservation of the Stability Axis Across Cell Lines and Species

Stability axes derived from LLP and SLP proteins were highly similar across the four cell lines, with cosine similarity values of 0.791–0.908 (Table 7; Fig. 6A), indicating a conserved direction separating stable and unstable proteins in embedding space. When models trained on one cell line were applied to predict stability in the remaining datasets using all proteins, cross-cell correlations remained positive (r = 0.412–0.467) (**Supplementary Table S3;** Fig. 6B), demonstrating that the stability rule generalizes beyond the LLP–SLP subsets. The stability axis also captured continuous variation in protein half-life, showing moderate correlations with measured half-life values (Pearson r = 0.374–0.396; Spearman r = 0.371–0.404) (Table 8), indicating that it reflects a gradual stability landscape rather than a purely binary classification boundary. Finally, cross-species evaluation showed that models trained on mouse proteins predicted human stability (r = 0.359), while the reciprocal analysis yielded r = 0.465 (**Supplementary Table S4;** Fig. 6C). Together, these results suggest that the stability-associated signal captured by protein language model embeddings is conserved across cell types and species. This conservation also indicates that the sequence may reflect conserved sequence features associated with protein stability across species and not specific to individual experimental systems.

Table 7
Stability axis cosine similarity matrix derived using LLP and SLP proteins

Train	Cell1	Cell2	Cell3	Cell4
Cell1	1.000	0.908	0.804	0.869
Cell2	0.908	1.000	0.871	0.896
Cell3	0.804	0.871	1.000	0.791
Cell4	0.869	0.896	0.791	1.000

Table 8
Continuous half-life prediction

Cell	Pearson r	Spearman r	R^2
Cell 1	0.374	0.371	0.13
Cell 2	0.395	0.398	0.147
Cell 3	0.385	0.404	0.142
Cell 4	0.396	0.392	0.146

3.7 Functional Role of Stability-Associated Sequence Grammar

Sequence grammar features provided measurable predictive power for distinguishing LLP and SLP proteins. A logistic regression model using grammar-derived features showed consistent performance across cell lines (Table 9), with ROC–AUC values of 0.650–0.677 and similar results in external human and mouse datasets (Fig. 7). The model indicated that LLP proteins tend to exhibit higher proline periodicity, tighter lysine spacing, and stronger lysine/acidic residue clustering, whereas SLP proteins show more dispersed residue organization. Importantly, these results demonstrate that sequence grammar alone captures a substantial portion of the stability signal, independent of full embedding representations. These sequence organization patterns are consistent with properties associated with protein degradation. Differences in lysine spacing and clustering suggest variation in the distribution of potential ubiquitination sites, while charge clustering may influence electrostatic interaction contexts. Periodic proline organization may affect local structural environments and accessibility of sequence features. Although these analyses do not directly measure degradation mechanisms, they provide biologically consistent interpretations of the sequence patterns identified. Perturbation analysis further supports the functional relevance of these features, as disrupting sequence organization consistently reduced predicted stability across all cell lines (Table 10). These results suggest that sequence grammar features contribute to the stability-associated signal captured by the model.

Table 9
Performance of the grammar-based logistic regression model.

Cell Line	ROC-AUC	Precision
Cell 1	0.678	0.629
Cell 2	0.656	0.635
Cell 3	0.666	0.617
Cell 4	0.676	0.639
Human External	0.654	0.61
Mouse	0.722	0.697

Table 10
Perturbation analysis results

Cell	Mean Δ Predicted Log-Half-Life	Paired t-test p-value
1	-0.069	0.00077
2	-0.078	1.048×10^{-6}
3	-0.024	0.0213
4	-0.069	1.1661×10^{-5}

3.8 Functional enrichment of stability-associated proteins

Functional enrichment analysis revealed that both long-lived proteins (LLPs) and short-lived proteins (SLPs) are predominantly associated with regulatory biological processes, particularly those related to transcription and RNA metabolism. LLPs showed relatively stronger enrichment in chromatin organization and transcription-associated processes, indicating a bias toward nuclear gene regulatory functions (Fig. 8). In contrast, SLPs were enriched in broader regulatory categories, including cellular and metabolic processes, reflecting less functionally specific enrichment profiles. Although both protein groups participate extensively in regulatory networks, these results suggest a shift in functional emphasis rather than a strict functional separation. This pattern is consistent with the sequence-derived stability signal identified in this study, indicating that features captured by protein language model embeddings may reflect underlying functional tendencies associated with protein stability. The analysis and results are provided in CSV form in **Supplementary File 9**.

3.9 Decomposition of sequence determinants of protein stability

To dissect sequence contributions to protein stability, we evaluated composition, sequence grammar, motif features, and protein language model embeddings. As shown in Table 11, composition showed moderate performance (ROC–AUC = 0.64–0.70), indicating a strong baseline contribution of amino acid frequencies. Grammar features, capturing residue organization (spacing and clustering), showed comparable performance (0.65–0.68). Motif-based features performed better (0.71–0.73), while embeddings achieved the highest accuracy (0.80–0.82), indicating integration of multi-level sequence information. These results support a hierarchical contribution of sequence features to the stability-associated signal across sequence scales, with higher-order representations integrating and amplifying lower-level sequence signals. Combined analysis across all datasets revealed compositional biases, with LLP enriched in D, Q, K, G, and T, and SLP enriched in C, R, S, N, and W. Together, these results support a multi-scale model of protein stability, in which composition, motif-level signals, and higher-order sequence grammar jointly contribute to stability, and protein language model embeddings integrate these signals into a unified representation. Top motif features are provided in **Supplementary File 10**.

Table 11
Contribution of different sequence representations

Cell Line	Composition (AUC)	Grammar (AUC)	Motif (AUC)	Embedding (AUC)
Cell 1	0.642	0.678	0.712	0.823
Cell 2	0.667	0.656	0.721	0.802
Cell 3	0.699	0.666	0.711	0.811
Cell 4	0.696	0.676	0.726	0.817

3.10 Summary of Stability-Associated Sequence Grammar

As summarized in Fig. 9, protein stability is associated with multiple sequence scales across multiple sequence scales, including amino acid composition, short motif patterns, and higher-order sequence grammar. Long-lived proteins exhibit tighter lysine spacing, enhanced charge clustering, and increased proline periodicity, whereas short-lived proteins display more dispersed residue organization (Fig. 9A). Comparative analysis across sequence representations reveals a clear hierarchy of predictive power, with protein language model embeddings outperforming composition-, motif-, and grammar-based features, indicating integration of multi-scale sequence information (Fig. 9B).

Projection onto the embedding-derived stability axis separates LLP and SLP proteins, demonstrating distinct stability score distributions (Fig. 9C). Validation analyses further show that stability signals depend on sequence organization rather than composition alone, as sequence shuffling reduces predictive performance, stability signals are distributed across sequences, and perturbation of sequence grammar features decreases predicted stability (Fig. 9D). Together, these results support a model in which protein stability arises from distributed sequence organization captured by embedding-based representations and associated with stability-related sequence properties (Fig. 9E). The features and the

python scripts for cohen's d analysis are provided in **Supplementary File 13** and analysis fig is provided in **Supplementary Fig. 3**.

4. Discussion

Protein degradation is central to proteome homeostasis [11], yet the sequence-level principles governing intrinsic protein stability remain incompletely understood. Existing models largely emphasize localized degradation signals such as degrons and PEST regions [3–5], which account for only part of the observed variability in protein lifetimes [4, 6]. The results presented here support an alternative perspective in which protein stability is associated with distributed, multi-scale sequence organization rather than being explained solely by localized motifs.

A central observation of this study is the presence of a stability-associated axis in protein sequence representation space that separates long-lived (LLP) and short-lived proteins (SLP), shows consistent behavior across cell lines and species, and relates to continuous variation in protein half-life. The reproducibility of this signal across independent datasets suggests that it reflects a conserved sequence-associated property rather than a dataset-specific effect.

Decomposition of sequence representations further indicates that protein stability is not explained by amino-acid composition alone but reflects contributions from multiple sequence levels. While composition provides a baseline signal (ROC–AUC ≈ 0.64–0.70), sequence grammar (≈ 0.65–0.68) and motif features (≈ 0.71–0.73) contribute additional information, with protein language model embeddings (≈ 0.80–0.82) integrating these signals into a unified representation. This progressive increase in predictive performance is consistent with a multi-level organization of sequence features contributing to the stability-associated signal.

Perturbation and shuffling analyses provide additional support for the role of sequence organization. Disrupting residue order while preserving composition reduces predictive performance, and sliding-window analyses indicate that stability-associated signals are distributed across sequences rather than confined to localized motifs [27, 28]. Together, these observations are consistent with a framework in which residue arrangement contributes to stability-associated information beyond compositional effects alone.

Analysis of sequence features reveals consistent organization patterns associated with stability, including tighter lysine spacing, enhanced charge clustering, and increased proline periodicity in LLP proteins compared with SLP proteins. Perturbation analyses further support the contribution of these features, with disruption of charge organization producing the largest decrease in predicted stability. These findings suggest that electrostatic organization and residue spacing are associated with the stability signal captured by sequence representations.

From a biochemical perspective, these patterns are consistent with known aspects of protein degradation, including the role of lysine residues in ubiquitination and the influence of charge distribution

on protein–protein interactions [2, 11]. However, these interpretations remain inferential and require direct experimental validation. Motif-level analysis further indicates that degron-like patterns are broadly present across proteins, whereas differences in their distribution and organization provide more discriminative information, supporting the importance of sequence context.

Functional enrichment analysis provides additional biological context, with LLP proteins enriched in chromatin organization and transcription-related processes, and SLP proteins associated with broader regulatory and metabolic functions. These patterns are consistent with the sequence-derived stability signal and suggest a link between sequence organization, protein function, and turnover.

Beyond mechanistic interpretation, these findings have broader implications. Sequence-level organization may contribute to variability in experimentally measured protein half-life across proteomics workflows, indicating that intrinsic sequence features partially shape observed turnover rates. In addition, the identified sequence grammar provides a basis for predicting protein stability directly from sequence and may inform the design of proteins with controlled degradation properties [36, 37].

This study has several limitations. The analyses rely on existing proteomics datasets and do not include direct experimental validation of the proposed mechanisms. Although the consistency of results across datasets, cell lines, and species supports their generality, future work incorporating targeted experimental perturbations will be required to directly test the functional role of the identified sequence features.

Taken together, these results support a shift from strictly motif-centric views of protein degradation toward a framework in which protein stability is associated with distributed, multi-scale sequence organization. This study illustrates how protein language model representations can be used to resolve sequence-level patterns linked to biological function and provides a foundation for further investigation of regulatory information encoded within protein sequences.

4. Conclusion

In this study, we show that protein stability is associated with distributed, multi-scale sequence organization rather than being explained solely by localized sequence motifs. Using protein language model representations, we identify a stability-associated axis that distinguishes long-lived and short-lived proteins and shows consistent behavior across datasets and species. By systematically decomposing sequence representations, we find that stability-related signals reflect contributions from amino-acid composition, local motif patterns, and higher-order sequence organization. In particular, sequence grammar features, including lysine spacing, charge clustering, and proline periodicity, capture interpretable aspects of the stability-associated signal beyond composition alone. These findings support a framework in which protein stability is linked to the integration of sequence features across multiple scales, providing a basis for understanding variability in protein turnover directly from sequence. More broadly, this work highlights the potential of protein language model–based representations to

uncover biologically meaningful principles embedded in protein sequences and to enable systematic analysis of sequence-function relationships in proteome dynamics.

Declarations

Ethics declaration

Not applicable

Competing interests

The authors declare no competing interests.

Author Contribution

Conceptualization: Tangilal Dihan Chowdhury, Md. Shabiul Islam, Fasiha Tanzeem TaibaMethodology: Tangilal Dihan Chowdhury, Md Ushama Shafoyat, Kazy Noor e Alam SiddiqueeSoftware: Tangilal Dihan Chowdhury, Md Ushama Shafoyat, Maruf HasanFormal analysis: Tangilal Dihan Chowdhury, Maruf Hasan, Fasiha Tanzeem TaibaInvestigation: Tangilal Dihan Chowdhury, Kaiissar MannoorData curation: Tangilal Dihan Chowdhury, Md Ushama Shafoyat, Maruf Hasan, Kaiissar MannoorVisualization: Tangilal Dihan Chowdhury, Fasiha Tanzeem TaibaValidation: Md Ushama Shafoyat, Kaiissar MannoorSupervision: Md. Shabiul IslamProject administration: Md. Shabiul IslamWriting – original draft: Tangilal Dihan ChowdhuryWriting – review & editing: All authorsAll authors read and approved the final manuscript. All authors read and approved the final manuscript.

Acknowledgement

We gratefully acknowledge the Department of Biomedical Engineering, 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.

Data Availability

All data generated or analyzed during this study are included in this published article and its Additional Information files. The complete datasets, protein language model (PLM) outputs, and all analysis scripts

are provided in Additional Information 1–13. Additional data are available from the corresponding author upon reasonable request.

References

1. Alber AB, Suter DM. Dynamics of protein synthesis and degradation through the cell cycle. Cell Cycle. 2019;18:784–94. https://doi.org/10.1080/15384101.2019.1598725.
2. Finley D. Recognition and processing of ubiquitin-protein conjugates by the proteasome. Annu Rev Biochem. 2009;78:477–513. https://doi.org/10.1146/annurev.biochem.78.081507.101607.
3. Zhang Z, Mena EL, Timms RT, et al. Degrons: defining the rules of protein degradation. Nat Rev Mol Cell Biol. 2025;26:868–83. https://doi.org/10.1038/s41580-025-00870-z.
4. Varshavsky A. N-degron and C-degron pathways of protein degradation. Proc Natl Acad Sci U S A. 2019;116:358–66. https://doi.org/10.1073/pnas.1816596116.
5. Rogers S, Wells R, Rechsteiner M. Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. Science. 1986;234:364–8. https://doi.org/10.1126/science.2876518.
6. Fishbain S, Inobe T, Israeli E, et al. Sequence composition of disordered regions fine-tunes protein half-life. Nat Struct Mol Biol. 2015;22:214–21. https://doi.org/10.1038/nsmb.2958.
7. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021;118:e2016239118. https://doi.org/10.1073/pnas.2016239118.
8. Leclercq M, Droit A. Protein language models: applications and perspectives. J Proteome Res. 2026;25:507–24. https://doi.org/10.1021/acs.jproteome.5c00506.
9. Lamb KD, Hughes J, Lytras S, et al. From single-sequences to evolutionary trajectories: protein language models capture the evolutionary potential of SARS-CoV-2. Nat Commun. 2026. https://doi.org/10.1038/s41467-026-69569-9.
10. Li J, Cai Z, Vaites LP, et al. Proteome-wide mapping of short-lived proteins in human cells. Mol Cell. 2021;81:4722–e47355. https://doi.org/10.1016/j.molcel.2021.09.015.
11. Zhao L, Zhao J, Zhong K, et al. Targeted protein degradation: mechanisms, strategies and application. Signal Transduct Target Ther. 2022;7:113. https://doi.org/10.1038/s41392-022-00966-4.
12. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43:D204–12. https://doi.org/10.1093/nar/gku989.
13. Ong SE, Blagoev B, Kratchmarova I, et al. Stable isotope labeling by amino acids in cell culture as a simple and accurate approach to expression proteomics. Mol Cell Proteom. 2002;1:376–86. https://doi.org/10.1074/mcp.M200025-MCP200.
14. Cambridge SB, Gnad F, Nguyen C, et al. Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover. J Proteome Res. 2011;10:5275–84. https://doi.org/10.1021/pr101183k.

15. Villegas-Morcillo A, Gomez AM, Sanchez V. An analysis of protein language model embeddings for fold prediction. Brief Bioinform. 2022;23:bbac142. https://doi.org/10.1093/bib/bbac142.
16. Ali S, Chourasia P, Patterson M. When protein structure embedding meets large language models. Genes. 2024;15:25. https://doi.org/10.3390/genes15010025.
17. Hung C, Tsai CF, Wu MH. Dimensionality reduction strategies for classification: ML versus DL approaches and their combinations. Expert Syst. 2025. https://doi.org/10.1111/exsy.70140.
18. Mykhailiuk I, Schäfer K, Büskens C. Parametric stability score and its application in optimal control. IFAC-PapersOnLine. 2022. https://doi.org/10.1016/j.ifacol.2022.09.019.
19. Allgaier J, Pryss R. Cross-validation visualized: a narrative guide to advanced methods. Mach Learn Knowl Extr. 2024;6:1378–88. https://doi.org/10.3390/make6020065.
20. Satria A, Sitompul OS, Mawengkang H. 5-fold cross validation on supporting k-nearest neighbour accuracy of making consimilar symptoms disease classification. In: Proc 2021 Int Conf Comput Sci Eng (IC2SE). IEEE; 2021. https://doi.org/10.1109/IC2SE52832.2021.9792094.
21. Xu M, Fralick D, Zheng JZ, et al. The differences and similarities between two-sample t-test and paired t-test. Shanghai Arch Psychiatry. 2017;29:184–8. https://doi.org/10.11919/j.issn.1002-0829.217070.
22. Diener MJ. Cohen's d. In: The Corsini Encyclopedia of Psychology. 2010. https://doi.org/10.1002/9780470479216.corpsy0200
23. Chu SKS, Narang K, Siegel JB. Protein stability prediction by fine-tuning a protein language model on a mega-scale dataset. PLoS Comput Biol. 2024. https://doi.org/10.1371/journal.pcbi.1012248.
24. Lahitani AR, Permanasari AE, Setiawan NA. Cosine similarity to determine similarity measure: study case in online essay assessment. Proc 2016 Int Conf Cyber IT Serv Manag. IEEE; 2016. https://doi.org/10.1109/CITSM.2016.7577578.
25. Abounaima MC, El Mazouri FZ, Lamrini L et al. The Pearson correlation coefficient applied to compare multi-criteria methods: case the ranking problematic. In: Proc 2020 Int Conf Innov Res Appl Sci Eng Technol (IRASET). IEEE; 2020. https://doi.org/10.1109/IRASET48871.2020.9092242
26. Blakeley-Ruiz JA, Kleiner M. Considerations for constructing a protein sequence database for metaproteomics. Comput Struct Biotechnol J. 2022;20:937–52. https://doi.org/10.1016/j.csbj.2022.01.018.
27. Jiang M, Anderson J, Gillespie J, et al. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinformatics. 2008;9:192. https://doi.org/10.1186/1471-2105-9-192.
28. Mokhtari F, Akhlaghi MI, Simpson SL, et al. Sliding window correlation analysis: modulating window shape for dynamic brain connectivity in resting state. NeuroImage. 2019;189:655–66. https://doi.org/10.1016/j.neuroimage.2019.02.001.
29. Shi WH, Marciel AB. Influence of charge block length on conformation and cluster formation of atactic peptide polyampholytes. Macromolecules. 2026. https://doi.org/10.1021/acs.macromol.5c02873.

30. Palaniappan A, Jakobsson E. Fourier analysis of conservation patterns in protein secondary structure. Comput Struct Biotechnol J. 2017;15:265–70. https://doi.org/10.1016/j.csbj.2017.02.002.
31. Dosztányi Z. Prediction of protein disorder based on IUPred. Protein Sci. 2018;27:331–40. https://doi.org/10.1002/pro.3334.
32. Chicco D, Sichenze A, Jurman G. A simple guide to the use of Student's t-test, Mann–Whitney U test, chi-squared test, and Kruskal–Wallis test in biostatistics. BioData Min. 2025;18:56. https://doi.org/10.1186/s13040-025-00465-6.
33. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol. 2022;75:25–36. https://doi.org/10.4097/kja.21209.
34. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, Vilo J. g:Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016;44(W1):W83–9. https://doi.org/10.1093/nar/gkw199.
35. Chagoyen M, Pazos F. Quantifying the biological significance of gene ontology biological processes —implications for the analysis of systems-wide data. Bioinformatics. 2010;26(3):378–84. https://doi.org/10.1093/bioinformatics/btp663.
36. Holman SW, Hammond DE, Simpson DM, et al. Protein turnover measurement using selected reaction monitoring-mass spectrometry (SRM-MS). Philos Trans Math Phys Eng Sci. 2016;374:20150362. https://doi.org/10.1098/rsta.2015.0362.
37. McGinness KE, Baker TA, Sauer RT. Engineering controllable protein degradation. Mol Cell. 2006;22:701–7. https://doi.org/10.1016/j.molcel.2006.04.027.

Figures

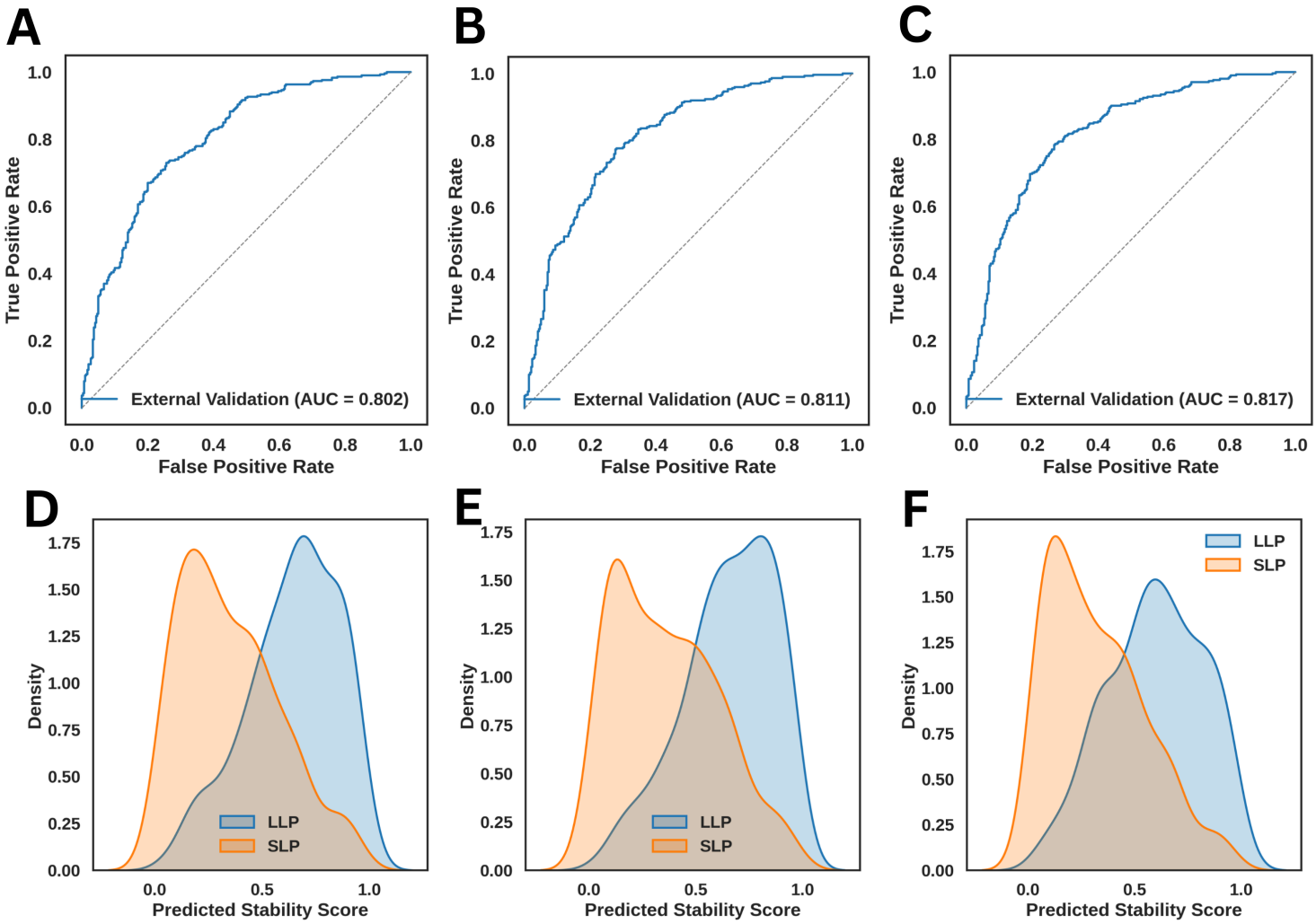


Figure 1

Cross–cell line validation performance of the stability prediction model. ROC curves showing model performance when trained on Cell 1 and tested on independent datasets: (A) Cell 2 (AUC = 0.802), (B) Cell 3 (AUC = 0.811), and (C) Cell 4 (AUC = 0.817). Density distributions of predicted stability scores for LLP and SLP peptides are shown in (D–F) for the corresponding validation sets, illustrating clear separation between the two classes.

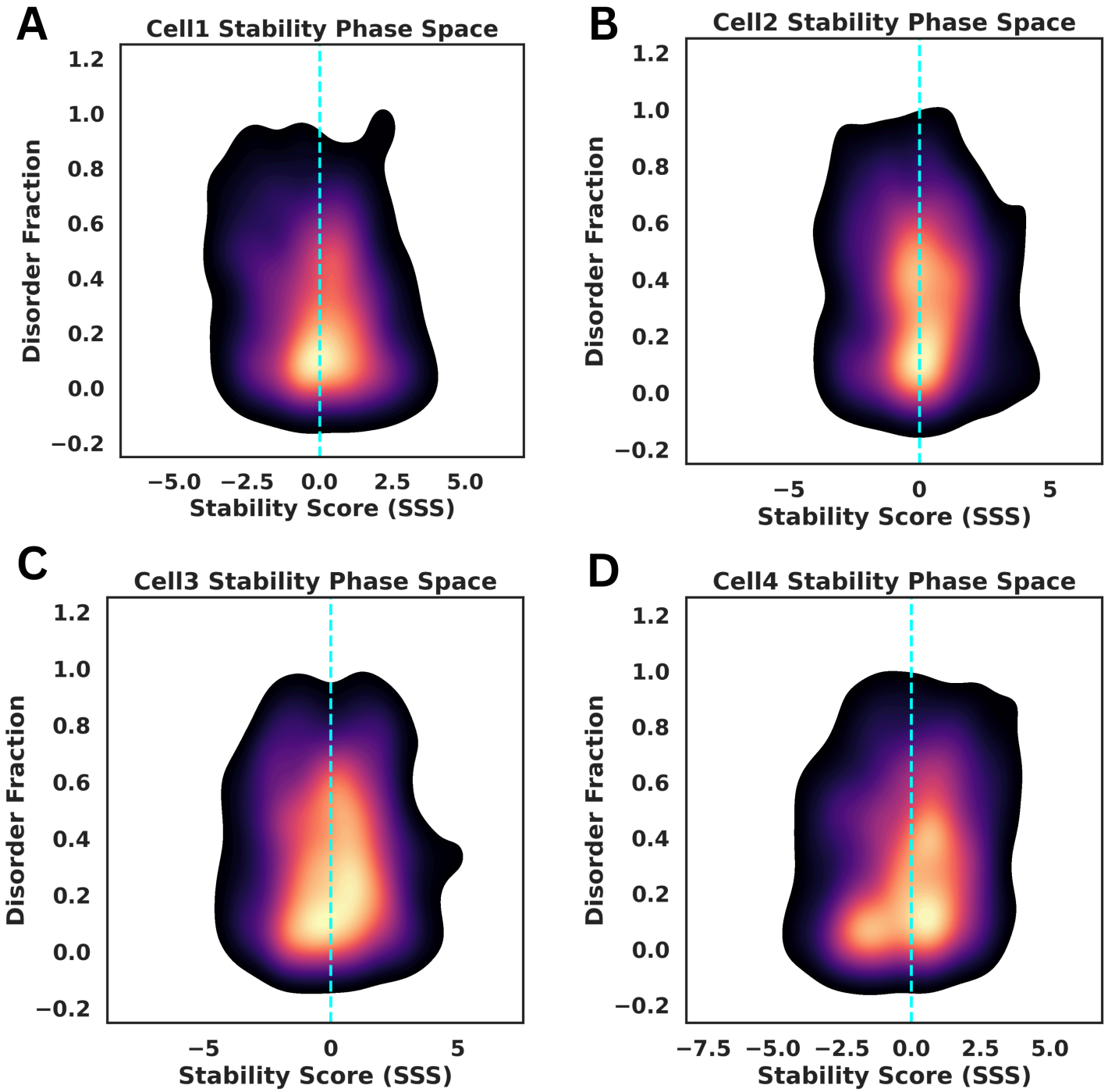


Figure 2

Stability phase-space distribution across cell lines. Density plots showing the distribution of proteins in stability phase space defined by stability score (SSS) and intrinsic disorder fraction for (a) Cell1, (b) Cell2, (c) Cell3, and (d) Cell4. The dashed vertical line indicates the stability score threshold (SSS = 0). The consistent distributions across cell lines illustrate the relationship between predicted stability and disorder content

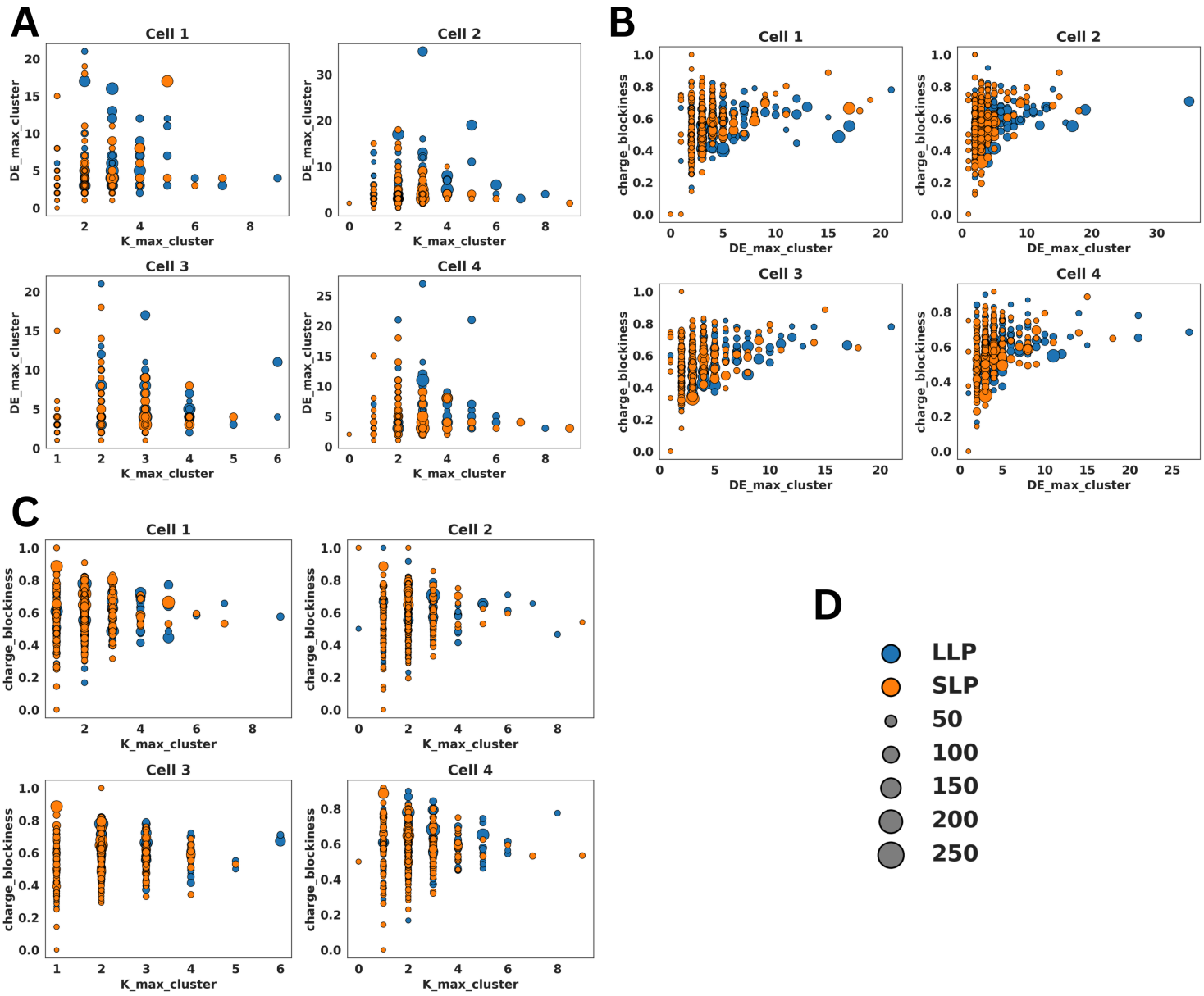


Figure 3

Charge clustering characteristics of LLP and SLP proteins across cell lines.(A) Relationship between negatively charged residue clusters (DE_max_cluster) and lysine clusters (K_max_cluster). (B) Association between charge blockiness and DE_max_cluster. (C) Relationship between charge blockiness and K_max_cluster. (D) Legend indicating LLP and SLP classes and point size scaling by protein count. Each panel shows results for the four cell lines, highlighting differences in charge clustering patterns between LLP and SLP proteins.

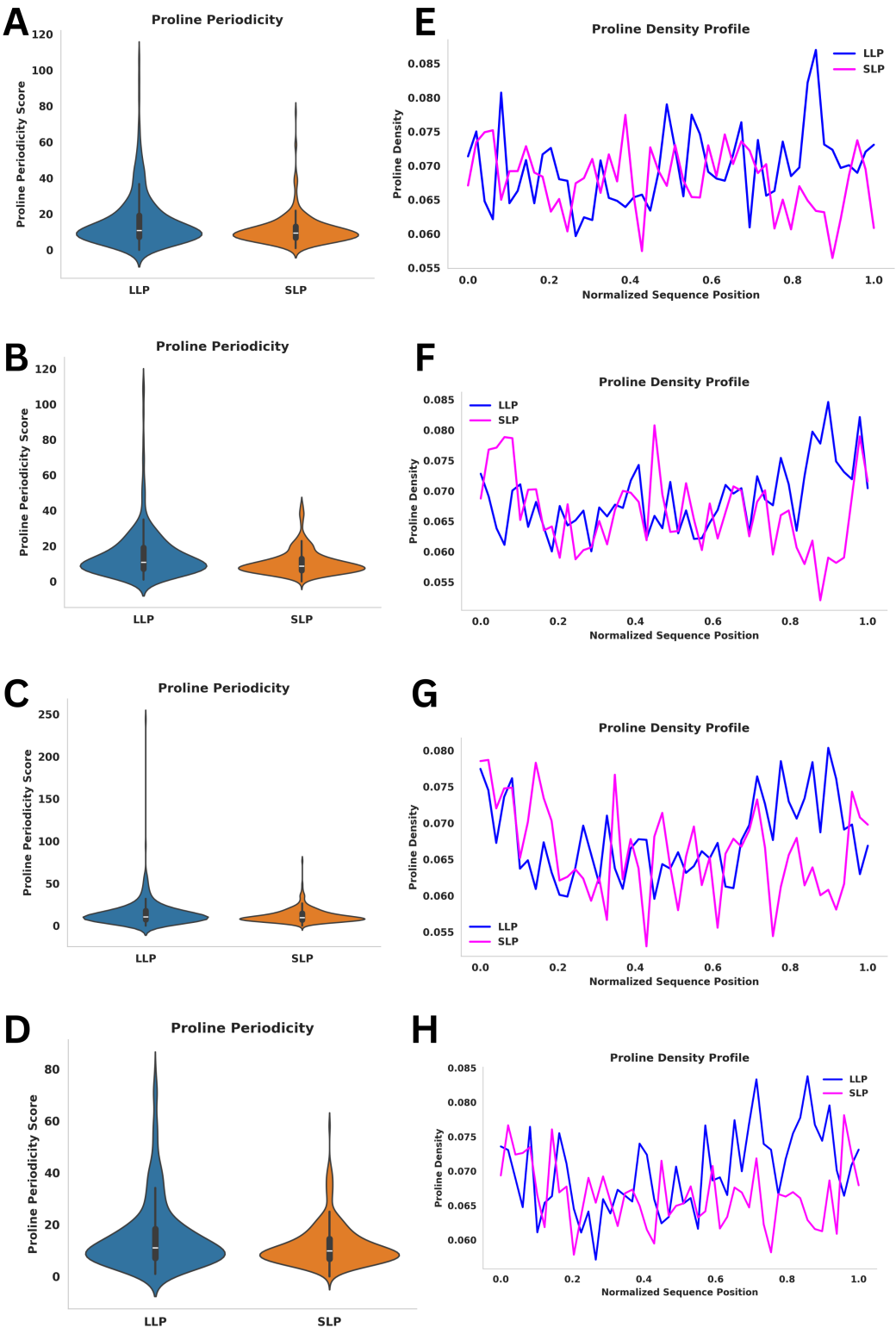


Figure 4

Proline distribution and periodicity patterns in stable and unstable proteins.(A–D) Violin plots showing the distribution of proline periodicity scores for LLP and SLP proteins across the four cell lines. (E–H) Position-normalized proline density profiles along protein sequences for LLP and SLP proteins.

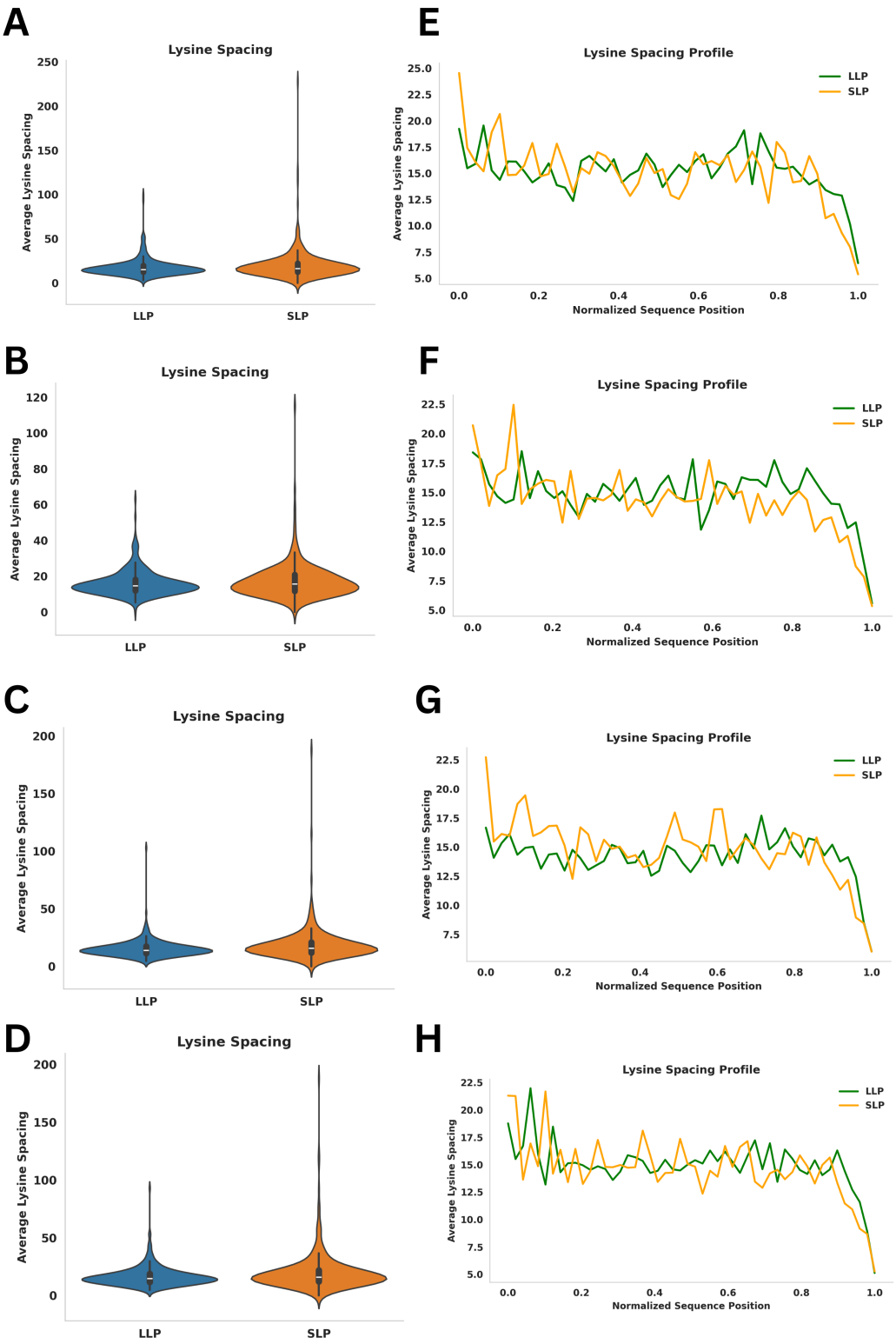


Figure 5

Lysine spacing patterns in stable and unstable proteins. (A–D) Violin plots comparing average lysine spacing between LLP and SLP proteins across the four cell lines. (E–H) Position-normalized lysine spacing profiles along protein sequences.

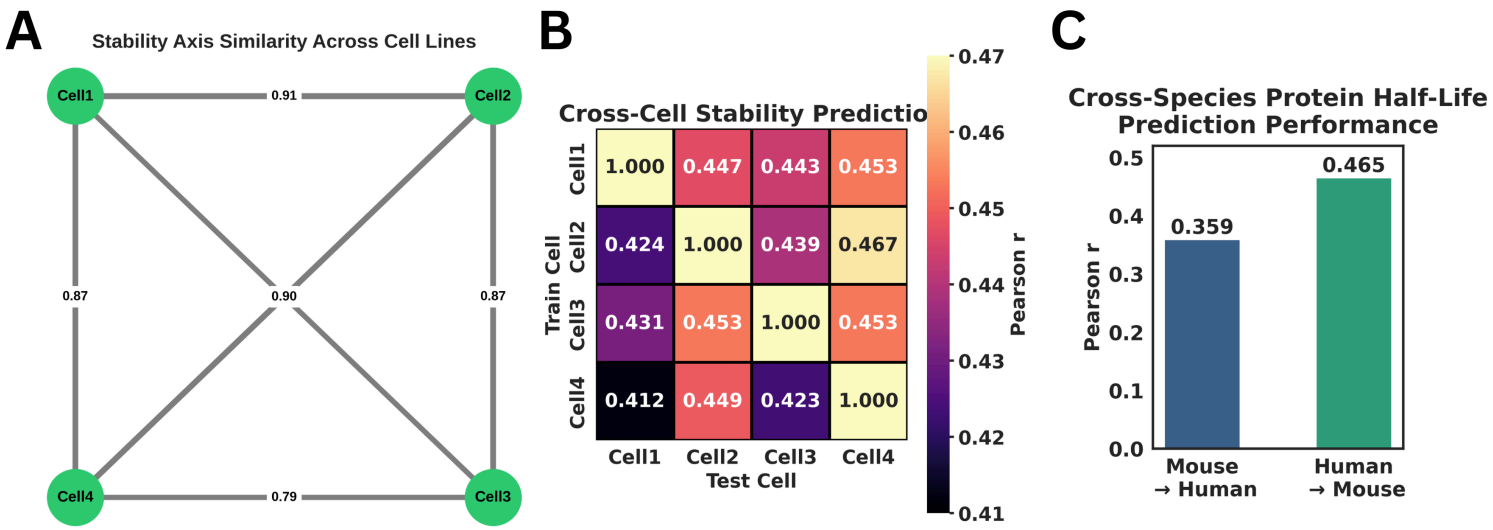


Figure 6

Conservation and generalization of the stability axis across cell lines and species. (A) Network representation of cosine similarity between stability axes derived from LLP and SLP proteins across four cell lines, showing strong similarity (0.791–0.908). (B) Heatmap of cross–cell line stability prediction performance using all proteins, with Pearson correlations ranging from 0.412 to 0.467. (C) Cross-species protein half-life prediction performance showing generalization of the stability signal between mouse and human datasets (Pearson r = 0.359 for mouse → human and r = 0.465 for human → mouse).

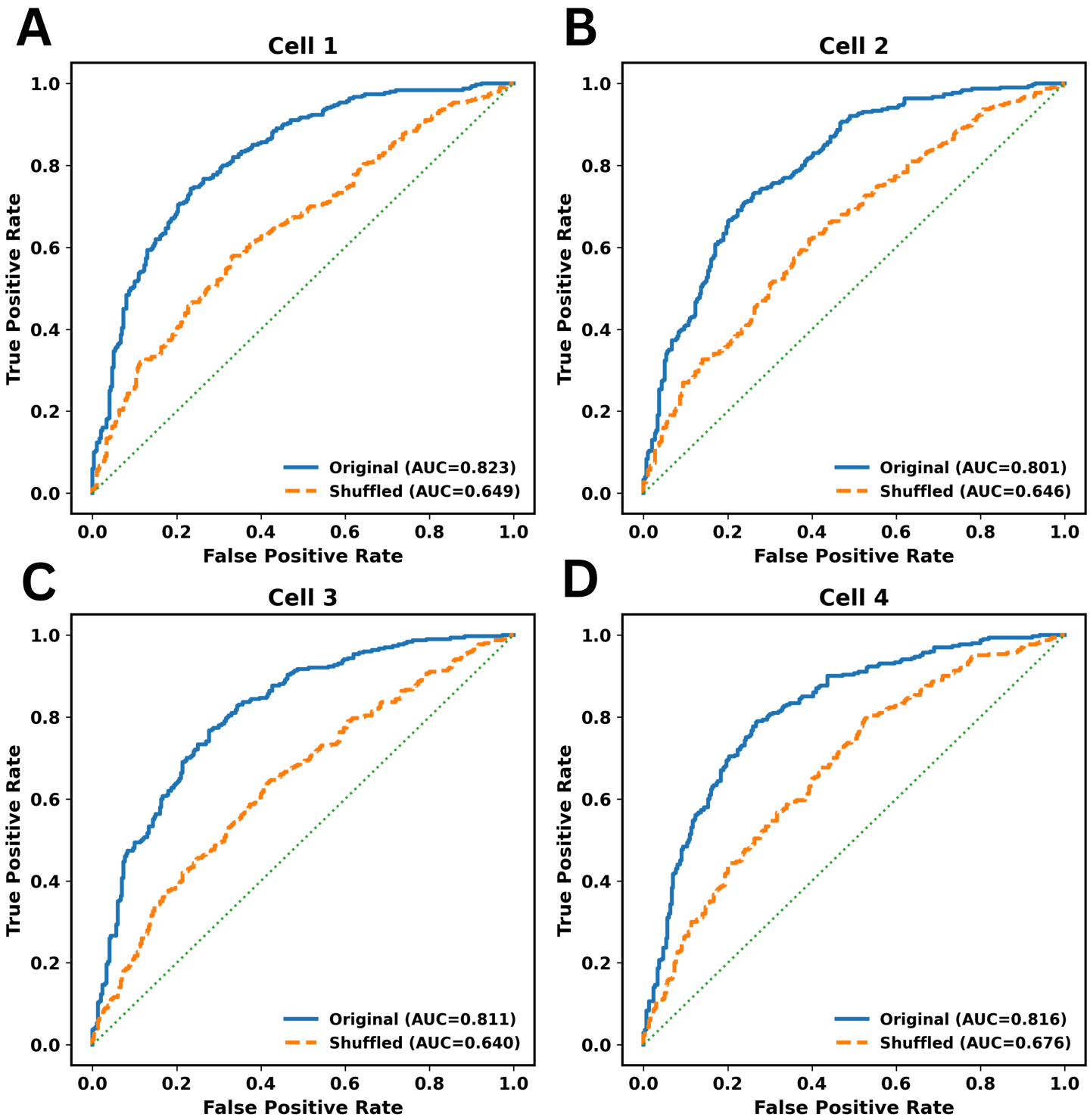


Figure 7

Effect of sequence shuffling on stability prediction across cell lines. Receiver operating characteristic (ROC) curves comparing model performance using original protein sequences (blue; embedding-based model) and shuffled sequences (orange) for (A) Cell 1, (B) Cell 2, (C) Cell 3, and (D) Cell 4. Shuffling preserves amino acid composition but disrupts residue order. Across all datasets, predictive performance decreases substantially after shuffling (AUC ≈ 0.64–0.68) compared to original sequences

(AUC ≈ 0.80–0.82), indicating that protein stability signals depend strongly on sequence organization rather than composition alone. The diagonal dashed line represents random classification.

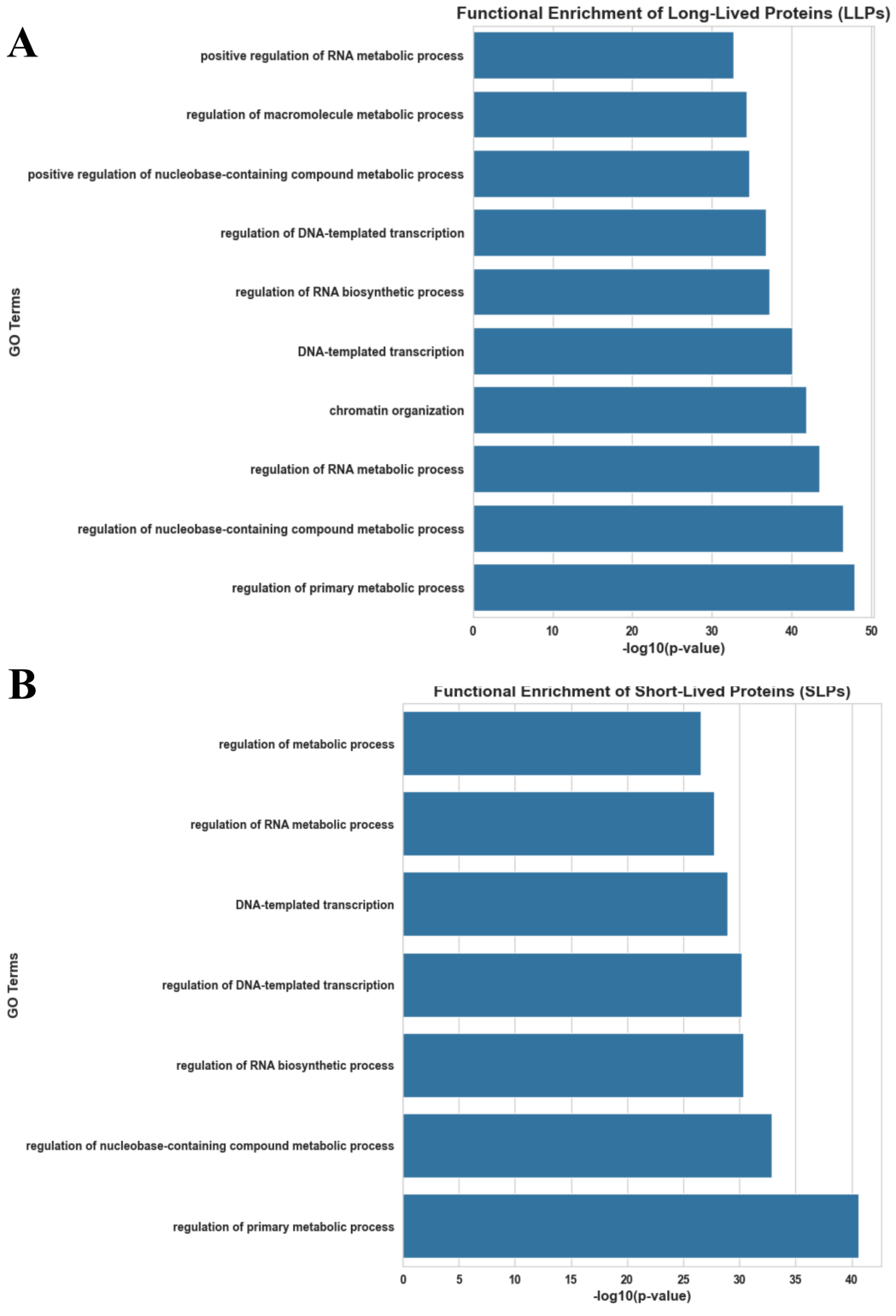


Figure 8

Functional enrichment of long-lived and short-lived proteins.(A) Gene Ontology (GO) enrichment analysis of long-lived proteins (LLPs), showing significant enrichment in transcriptional regulation, RNA metabolic processes, and chromatin organization. (B) Enrichment analysis of short-lived proteins (SLPs),

highlighting broader regulatory and metabolic processes. Bars represent −log10(p-value), indicating the significance of enrichment.

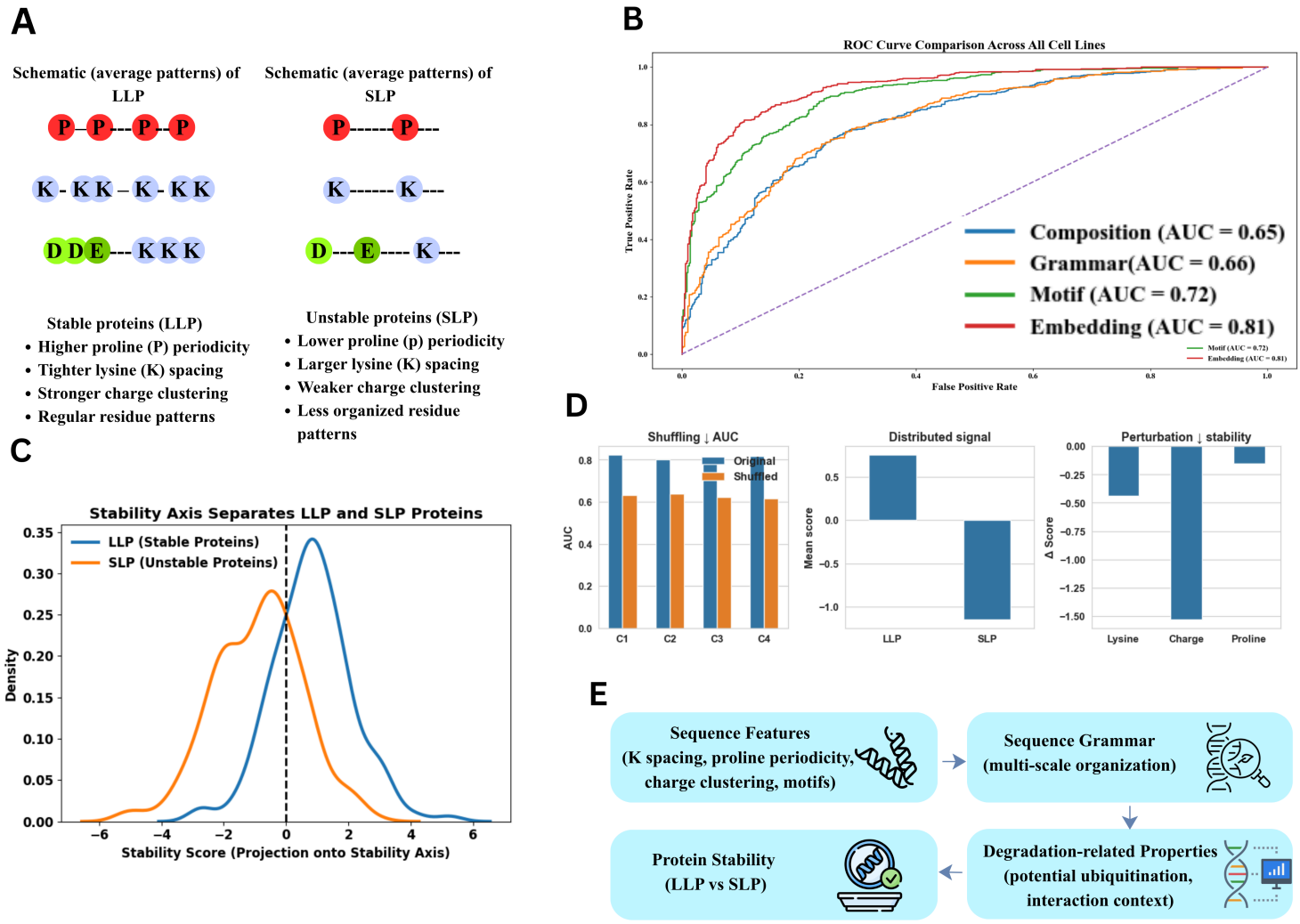


Figure 9

Multi-scale sequence determinants of protein stability and their integration in protein language model embeddings. (A) Schematic representation of average sequence organization patterns in long-lived proteins (LLP) and short-lived proteins (SLP), highlighting differences in proline periodicity, lysine spacing, and charge clustering. (B) Comparison of predictive performance across sequence representations, showing increasing accuracy from amino acid composition to motif features and sequence grammar, with protein language model embeddings achieving the highest performance, indicating integration of multi-scale sequence information. (C) Projection onto the stability axis derived from embeddings separates LLP and SLP proteins, demonstrating distinct stability score distributions. (D) Validation analyses showing that stability signals depend on sequence organization: shuffling reduces predictive performance, stability signals are distributed across sequences, and perturbation of sequence grammar features decreases predicted stability. (E) Conceptual model illustrating how sequence features and multi-scale sequence grammar contribute to stability-associated sequence properties and overall protein stability.

Supplementary Files

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

- Tables.docx
- SupplementaryFile13.zip
- SupplementaryFile9.zip
- SupplementaryFile4.zip
- SupplementaryFile8.zip
- SupplementaryFile10.zip
- SupplementaryFile12.zip
- SupplementaryFile6.zip
- SupplementaryFile5.zip
- SupplementaryFile1.zip
- SupplementaryFile3.zip
- SupplementaryFile11.zip
- SupplementaryFile2.zip
- SupplementaryFile7.zip