Briefings in Bioinformatics, 2025, **26(3)**, bbaf171
https://doi.org/10.1093/bib/bbaf171
Problem Solving Protocol

PSTP: accurate residue-level phase separation prediction using protein conformational and language model embeddings

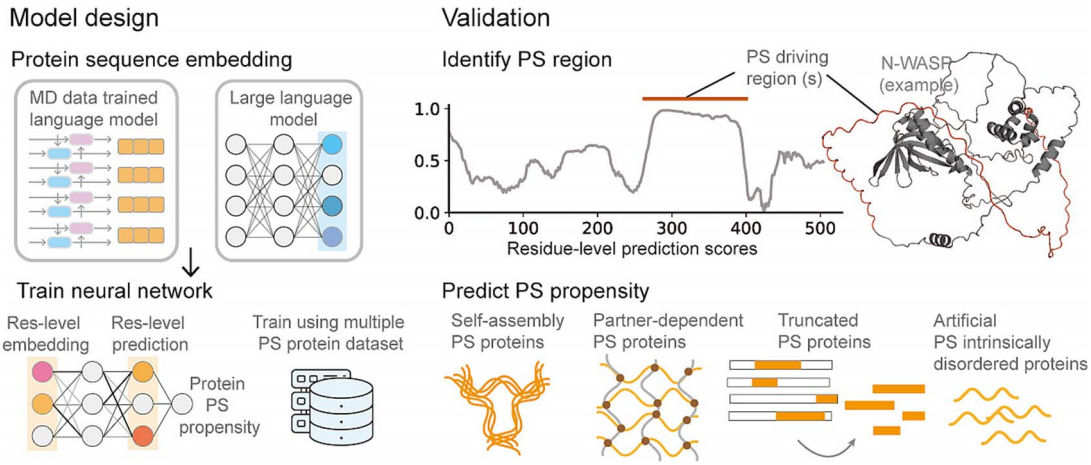
Mofan Feng[1,2], Liangjie Liu[1,2], Zhuo-Ning Xian[3], Xiaoxi Wei[1], Keyi Li[1,2], Wenqian Yan[1,2], Qing Lu[1,*], Yi Shi[1,2,*], Guang He[1,2,*]

[1]Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Shanghai Jiao Tong University, No. 1954 Huashan Road, Xuhui District, Shanghai 200030, China
[2]Shanghai Institute of Medical Genetics, Shanghai Children's Hospital, Shanghai Jiao Tong University School of Medicine, No. 24 Lane 1400 West Beijing Road, Jing'an District, Shanghai 200040, China
[3]School of Environmental Science & Engineering, Shanghai Jiao Tong University, No. 800 Dongchuan Road, Minhang District, Shanghai 200240, China
*Corresponding authors. Qing Lu, Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Shanghai Jiao Tong University, Shanghai 200030, China. E-mail: luqing67@sjtu.edu.cn; Yi Shi, Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Shanghai Jiao Tong University, No. 1954 Huashan Road, Xuhui District, Shanghai 200030, China. E-mail: yishi@sjtu.edu.cn; Guang He, Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Shanghai Jiao Tong University, No. 1954 Huashan Road, Xuhui District, Shanghai 200030, China. E-mail: heguang@sjtu.edu.cn
Biographical note: Dr He's laboratory focuses on identifying candidate genes for bipolar disorder and other mental illnesses, as well as unraveling the century-old mystery of the declining incidence of colorectal cancer in schizophrenia patients.

Abstract

Phase separation (PS) is essential in cellular processes and disease mechanisms, highlighting the need for predictive algorithms to analyze uncharacterized sequences and accelerate experimental validation. Current high-accuracy methods often rely on extensive annotations or handcrafted features, limiting their generalizability to sequences lacking such annotations and making it difficult to identify key protein regions involved in PS. We introduce Phase Separation's Transfer-learning Prediction (PSTP), which combines conformational embeddings with large language model embeddings, enabling state-of-the-art PS predictions from protein sequences alone. PSTP performs well across various prediction scenarios and shows potential for predicting novel-designed artificial proteins. Additionally, PSTP provides residue-level predictions that are highly correlated with experimentally validated PS regions. By analyzing 160 000+ variants, PSTP characterizes the strong link between the incidence of pathogenic variants and residue-level PS propensities in unconserved intrinsically disordered regions, offering insights into underexplored mutation effects. PSTP's sliding-window optimization reduces its memory usage to a few hundred megabytes, facilitating rapid execution on typical CPUs and GPUs. Offered via both a web server and an installable Python package, PSTP provides a versatile tool for decoding protein PS behavior and supporting disease-focused research.

Graphical Abstract



Keywords: language model; protein phase separation; residue-level phase separation prediction; conformation-aware embeddings; intrinsically disordered regions

Received: January 18, 2025. **Revised:** March 7, 2025. **Accepted:** March 19, 2025

Introduction

Biomolecular condensates formed through phase separation (PS) are found throughout eukaryotic cells [1–3] enabling a wide range of functions across multiple scales [4]. Including impacting biochemical reaction rates [5], protein organization [6], and cellular localization [7] and its dysregulation is associated with diseases like cancer and neurodegeneration [8].

PS is a complex and not fully understood process [9, 10], driven by multivalent interactions involving various physicochemical forces [11–15]. Both structured and intrinsically disordered regions (IDRs) contribute to PS, either as a single component or through a "scaffold-client" manner involving multiple components [1, 2]. In addition, IDR chain compaction is closely linked to their PS propensity, and recent studies used molecular dynamics (MD) data to train machine learning (ML) models that predict IDR conformational properties from sequence [16, 17]. Considering the vast amount of uncharacterized sequence data, developing reliable methods to predict PS and identifying the sequences responsible for this process are crucial for advancing our understanding of biomolecular condensates, their functional roles, and the mechanisms underlying disease development.

To predict PS, "first-generation" algorithms were developed, focusing primarily on specific protein sequence characteristics [11, 18–22]. For example, PScore [11] predicts PS using pi–pi contact information, and catGRANULE [20] predicts PS using the frequency of specific amino acid. More advanced, integrated feature-based models [23–27] have since been introduced, combining multiple sequence features with novel attributes. For example, PSAP [23] and MolPhase [26] use a comprehensive collection of biochemical–physical properties and employ random forests for PS prediction. DeePhase, similar to PSAP, enhances its feature set by incorporating trained Skip-Gram Word2Vec vectors. PhaSePred [25] integrates multiple sequence-based predictors with immunofluorescence imaging and post-translational modification information. PSPHunter [24] incorporates Word2Vec and combines various annotations, including protein–protein interactions (PPIs) and functional annotations. In parallel, other models have emerged that differentiate PS proteins from those forming amyloids [28, 29].

Nevertheless, advanced integrated feature-based models rely on engineered features from various sources, such as protein annotations and imaging data, which limits their generalizability to proteins, including variants and isoforms, that lack such detailed annotations. Moreover, current methods that provide residue-level PS scores are mostly "first-generation" algorithms that focus on limited sequence properties, making it challenging to accurately predict PS-driving regions due to the complex and diverse sequence patterns involved. These limitations hinder progress in uncovering key biological mechanisms and highlight the need for more flexible and broadly applicable models.

To address these critical issues, we developed PSTP (Phase Separation's Transfer-learning Prediction) (Fig. 1), employing a dual-language model embedding strategy (Fig. 1A). We utilize a large language model [30] to extracts latent biological features, and we also modify a MD simulations-based language model [16] to encode the sequence grammar underlying the physical and dynamic conformational properties. With this embedding method, state-of-the-art PS predictions are achievable using only protein sequence.

We also developed a lightweight attention-based module PSTP-Scan, to extract local sequence information to identify key regions driving PS (Fig. 1C). Without direct residue-level training, PSTP-Scan identified 120 out of 143 PS regions in PhaSePro [31] and improved the correlation coefficient with experimentally validated regions to ~150% of that of the best existing models.

In addition, we demonstrate the applicability and generalizability of PSTP across various protein types, such as artificial intrinsically disordered proteins (A-IDPs) and truncated proteins that undergo PS. Leveraging the high-throughput capability of PSTP, we characterized the relationship between pathogenicity and protein PS in low-conservation regions, whose variants are much less studied compared to structured domains, using >160 000 variants from ClinVar [32, 33]. We believe that PSTP will facilitate a deeper understanding of cellular processes and aid in disease research and therapeutic development.

PSTP models were designed with ease of use and portability in mind. To ensure low memory consumption, we implemented a sliding-window approach, making the large language model embedding and prediction within seconds on both standard CPUs and GPUs. We provide a locally installable version of PSTP, offering full access to the algorithms. Additionally, a user-friendly web server (www.pstp.online) has been developed to enable easy prediction for sequences of interest.

Materials and methods

Feature engineering

Feature embedding with protein large language model

We adopted two embedding matrices for each protein. The first matrix is the embedding computed by ESM-2 [30]; we applied the esm2_t6_8M_UR50D (ESM2-8M in short) from the python "esm" package (https://github.com/facebookresearch/esm/tree/main). ESM2-8M generates a 320-dimension vector for each position within each sequence (Fig. 1A and Fig. S4B). For each sequence, to reduce excessive memory usage as well as reduce the time required for long-distance attention computing, we apply a sliding-window process similar to AlphaFold2 [34] and AlphaMissense [35] when dealing with long sequences (as detailed in the Supplementary Information).

Feature embedding with ALBATROSS

The second embedding approach we adopted is the ALBATROSS program [16] (https://github.com/idptools/sparrow). To capture as much conformational information as possible, we adopted three ALBATROSS sub-models: the "asphericity" model, the "radius_of_gyration_scaled" model, and the "end_to_end_distance_scaled" model. For each of the three sub-models, we extract the hidden layer output at each residue position to produce a 110-dimensional ALBATROSS embedding per residue (Fig. 1A and Fig. S4A). These embeddings are then concatenated to form a 330-dimensional × sequence-length vector.

Feature engineering methods for comparison

1. Word2Vec Skip-Gram approach, which was adopted in several PS prediction methods such as PSPredictor [36], DeePhase, and PSPHunter [24]. We selected a trained Word2Vec embedding method [37].

2. Engineered feature vectors developed by van Mierlo *et al.* [23], which include 52 features commonly used in previous PS prediction methods. These features encompass amino acid composition, amino acid dimers, and biophysical properties like aromaticity, hydrophobicity, secondary structure propensities,

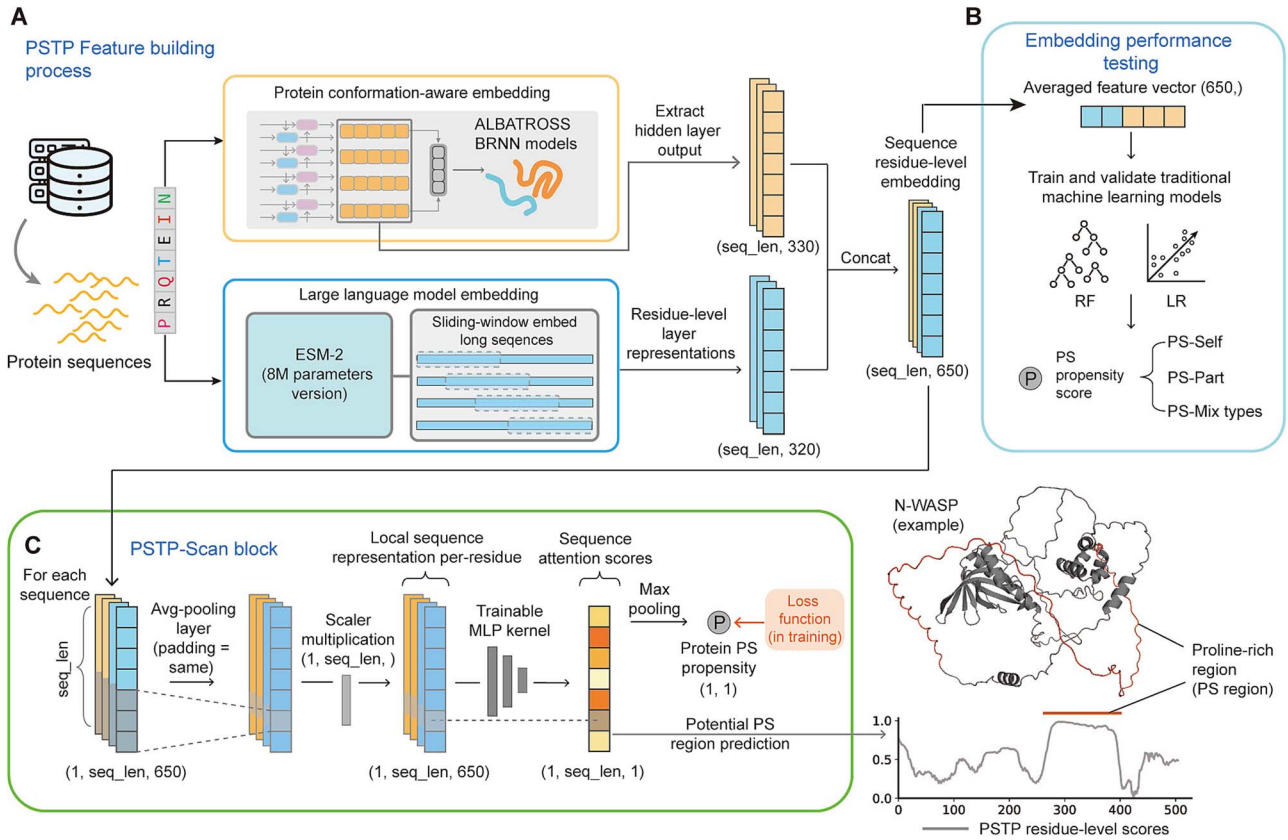


Figure 1. Schematic of PSTP embedding and machine learning (ML) architecture. (A) Protein sequences are converted into feature matrices using embeddings generated by the ALBATROSS LSTM-BRNN and the embeddings from the ESM-2 (8M parameters version) language model. For sequences longer than 256, ESM-2 is applied in a sliding-window manner to reduce time and memory costs. The combined result is a matrix of sequence-length × 650 for each sequence. (B) Training and validation of traditional ML models using the PSTP embedding. The sequence-length × 650 embedding matrix is averaged to a vector of length 650 for each sequence. Four categories of phase separation (PS) proteins are considered: self-assembly PS proteins (PS-Self), partner-dependent PS proteins (PS-Part), and mixed PS proteins (PS-Mix types). (C) Schematic of the PSTP-Scan block. The embedding matrix generated in (A) is used as input to produce attention scores, which serve as residue-level PS scores to detect PS regions. The maximum attention score for each sequence is output as the sequence-level PS propensity score. Here, we present N-WASP as an example, where PSTP-Scan predicts the proline-rich motif that forms multivalent interactions with its PS partner, Nck. MLP, multilayer perceptron.

molecular weight, disorder propensity, and sequence low-complexity property.

PSTP-scan neural network

Model architecture

PSTP-Scan takes the sequence-length × 650-dimension (320 for large language model embedding and 330 for conformation-aware embedding) feature matrix as input for each protein sequence. The model is implemented using the Python "pytorch" package.

Within each PSTP-Scan block, (1) the first layer is an average pool layer with "same" padding. (2) To mitigate the effects of "same" padding, a scaling step adjusts the values at the sequence ends to ensure that they reflect the actual averages of the sequence elements rather than padded zeros. Thus, the representation obtained for each position i can be written as below, where the window size is $2k + 1$, $x_{j,d}$ represents the dth value of the 650-dimensional input vector input for position j, and L represents the sequence length:

$$R_{i,d} = \frac{1}{\min(i+k, L) - \max(1, i-k)} \sum_{j=\max(1,i-k)}^{\min(i+k,L)} x_{j,d} \quad d = 1, 2, \ldots, 650 \tag{1}$$

(3) Next, a shared, trainable multilayer perceptron (MLP) kernel processes the averaged vector R_i. The MLP architecture consists of an initial layer with 20 neurons, followed by a LeakyReLU activation, a hidden layer with five neurons, and a final output layer with one neuron followed by a Sigmoid activation to produce a score between 0 and 1 for each residue position. These residue-level scores for each residue position form the sequence-length attention vector. So, the attention computation using the representation R_i at each position can be written as function (2). (4) The last is a maximum pooling layer that selects the maximum value of the attention vector, providing the sequence-level sequence propensity prediction (used to compute the loss against each binary label), as shown in function (3).

$$a_i = \mathrm{MLP}(R_i), \quad i = 1, 2, \ldots, L \tag{2}$$

$$O = \max_{i=1}^{L} a_i \tag{3}$$

Three parallel blocks with identical architectures but different average pooling window sizes (Fig. S4C) ($256+1$, $128+1$, and $32+1$, the extra 1 ensures symmetric "same" padding) are used to capture PS information from varying sequence fragment lengths. Model performance also remained robust across these window sizes when using a single block (Fig. S3E). The predicted propensities (single-value sequence-level propensity) from these three blocks are averaged to produce the final output, which is optimized through backpropagation to update the MLP parameters.

During the training process, for each iteration (50 epochs in total), we held out a random subset of proteins for the size of the positive dataset from the background dataset as the negative dataset. This approach was applied to address sample imbalance.

Model performance benchmarks

Area under the receiver operating characteristic curve (AUC) and area under the precision–recall curve (AUPR) were applied during each validation. For the head-to-head comparison of PSTP with PhaSePred and other PS predictors when using the validation dataset, we randomly selected proteins twice the size of the PS dataset from the background proteins and merged them with the positive dataset. Performance was evaluated using AUC and AUPR, and this data sample-evaluation process was repeated 50 times to calculate the average performance to ensure robustness. We employed a balanced AUC/AUPR computation strategy, where the evaluation dataset was randomly downsampled to an equal number of positive and negative samples (80% of the minority class variants) before computing AUC/AUPR, with 100 repetitions to obtain average performance. To evaluate residue-level performances, we calculated Spearman correlation coefficients between these residue-level scores and the binary annotation of actual PS regions (where 1 indicates positions within PS regions and 0 indicates positions outside) to assess predictive performance.

PS datasets for evaluation

Self-assembly PS and partner-dependent PS datasets

We utilize the curated datasets used for training and evaluating PhaSePred [25], which were obtained from the supplementary materials provided with its published literature. A total of 201 SaPS proteins, 327 PdPS proteins, and 60 220 NoPS background proteins were collected. The acquired data included training and cross-validation (CV) datasets: SaPS (128 proteins), hSaPS (59 proteins), PdPS (214 proteins), hPdPS (96 proteins), NoPS (48,158 proteins), and hNoPS (8801 proteins). The independent test sets comprised SaPS-test (73 proteins), hSaPS-test (34 proteins), PdPS-test (113 proteins), hPdPS-test (60 proteins), NoPS-test (12 062 proteins), hNoPS-test (2200 proteins), PS-test (53 proteins), and hPS-test (23 proteins).

Additional dataset for independent validation

To build an additional validation set, we utilize a PS dataset curated by a more recent study [24]. Among this data collection, we utilized the "hPS167" datasets containing 167 human-specific PS proteins along with human background proteins as the independent test dataset and the remaining PS proteins along with non-human background proteins as the training dataset. After reducing the sequence redundancy by CD-Hit at a threshold of 0.4, we obtained 136 human PS proteins, 375 non-human PS proteins, and 6280 human background proteins. According to the ratio of non-human to human PS proteins, we randomly selected 2.75 times the size of the human background proteins from the NoPS dataset (60 220 proteins) proteins and obtained 17 270 proteins (PS proteins were filtered out). We next filtered both background proteins to ensure they were not longer than the longest human PS proteins (5537 length), and we finally got 6259 human background proteins and 17 268 non-human background proteins.

Other PS models for comparison

We chose several representative prediction methods for comparison, including PhaSePred [25], which utilizes multidimensional integrated features; DeePhase, which combines a Word2Vec language model with engineered features; PLAAC [18, 19], which predicts prion-like propensities; PSAP [23], which utilizes engineered biochemical and biophysical features; and ParSe [22], a structure-based PS predictor. For comparison with PSPHunter [24], we constructed features following its sequence encoding method. Additionally, we evaluated other sequence-based predictors, including catGRANULE [20], FuzDrop [21, 38], and PScore [11].

Results

Conformational property analysis of PS proteins

Protein IDR chain compaction is closely tied to PS propensity [14, 17, 39]. Compared to the background proteome, IDRs in PS proteins generally exhibit more compact conformations, reflected by lower Flory scaling exponent (ν), conformational entropy per residue (S_{conf}/N), and asphericity (Fig. 2A and B, mix PSP IDRs column). We found that within PS proteins, self-assembly PS-Self types exhibit the greatest compactness, surpassing that of PS-Part proteins needing partners (Fig. 2A and B). In contrast, proteins that participate in PS through structured regions exhibit even higher conformation property values than the background proteome, reflecting more expanded IDRs (Fig. 2A and B, low-IDR PSP column).

We next evaluated four feature sets for capturing conformational properties: (1) ESM-2 embeddings (8M and 650M versions, averaged along the residue dimension), (2) conformational-aware embeddings designed using ALBATROSS [16], a MD-data trained model, (3) Word2Vec Skip-Gram approach used in several PS predictors, and (4) a 52-dimensional engineered feature collected from previous PS methods.

Using the IDRome dataset [17] (28 058 IDRs with corresponding MD simulated properties), we compared extremely compact ($\nu < 0.45$) and expanded ($\nu > 0.60$) IDRs, as well as those classified by conformational entropy per residue ($S_{conf}/N < 10$ versus $S_{conf}/N > 10.25$). UMAP visualizations for each method revealed that both versions of ESM-2 separated compact from expanded IDRs (Fig. 2D and F), with conformational-aware embeddings revealing even stronger distinctions (Fig. 2D and F, "ALBATROSS"). By contrast, "Word2Vec" showed no significant separation (Fig. 2D and F, "Word2Vec"), and engineered features partially differentiated compact and expanded IDRs but still produced overlapping clusters (Fig. S1). These results indicate that MD-based language models and the large language model capture meaningful conformational information about chain compaction.

Robust PS prediction with dual-language embeddings

To assess features' effectiveness in encoding essential information for PS proteins, we utilized curated data from PhaSePred [25] (a total of 201 PS-Self proteins, 327 PS-Part proteins, and 60 220 NoPS background proteins), and trained and evaluated logistic regression (LR) and random forest (RF) models on various feature combinations tested in the previous section.

Cross-validation and independent test results show that combining ESM-2 and conformational embeddings (averaged along the residue dimension) already yielded high performances for both PS-Self and PS-Part protein types, while incorporating engineered features or "Word2Vec" did not improve performance, and in some cases, even reduce it, as shown through both LR (Fig. 3A) and RF (Fig. S2A). Notably, excluding conformational embeddings (using ESM-2 only) resulted in a significant performance drop across both cross-validation and the test set, underscoring the critical role of protein conformational information in predicting

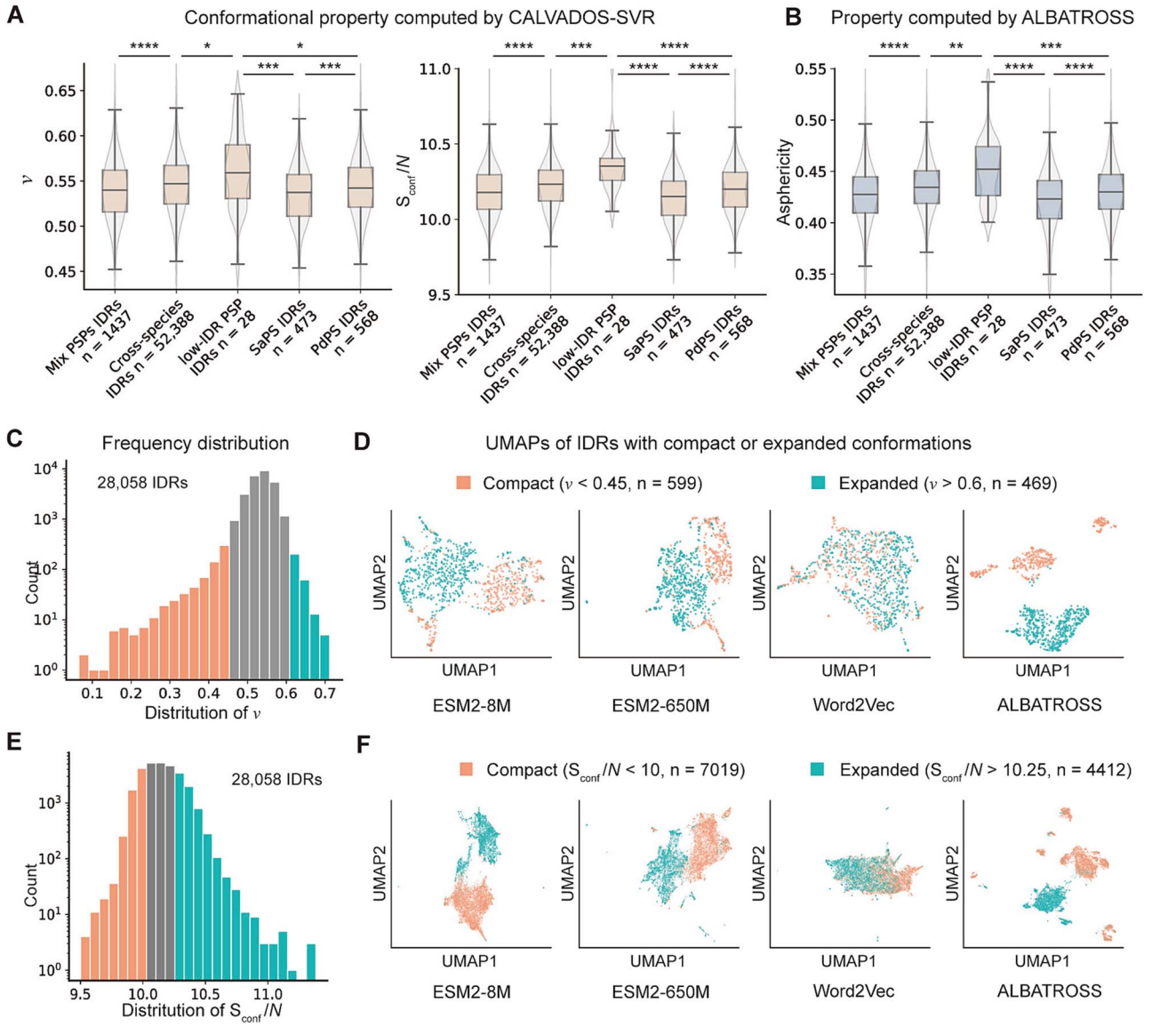


Figure 2. Analysis of IDR chain compaction properties and UMAP visualization of IDRs with varying chain compaction. (A and B) Comparison of predicted conformational property metrics of IDRs across different protein types (two-sided Mann–Whitney *U* test; mix PSPs IDRs: IDRs from PS proteins; cross-species IDRs: IDRs from background proteins; SaPS IDRs: IDRs from PS-Self proteins; PdPS IDRs: IDRs from PS-Part proteins; low-IDR PSPs IDRs: IDRs from PS protein driven by structured regions). (C) Distribution of ν values for IDRs. IDRs with ν <0.45 are classified as extremely compact, while those with ν >0.55 are classified as expanded. To balance data size, IDRs with ν >0.6 were selected for the expanded group. (D) UMAP visualization of IDRs, categorized into compact ($\nu < 0.45$) and expanded groups ($\nu > 0.6$), illustrating the vectorized embeddings produced by different models. ESM2-8M and ESM2-650M represent the averaged layer representations from the 8M and 650M parameter versions of the ESM-2 model, respectively. Word2Vec represents embeddings produced by the Skip-Gram approach. ALBATROSS represents the averaged hidden layer outputs of the ALBATROSS model. (E and F) A parallel analysis to (C and D), considering IDRs with $S_{conf}/N < 10$ as compact and IDRs with $S_{conf}/N > 10.25$ as expanded.

PS (Fig. 3B, Fig. S2B). We refer to the combination of conformational embeddings and ESM-2 as PSTP (Phase Separation's Transfer-learning Prediction embeddings).

We compared the embeddings' performance against PhaSePred, an advanced PS predictor, using the same training and independent validation sets. Although SaPS-10 and PdPS-10 (two PhaSePred versions) integrate multiple annotations and multidimensional features, PSTP-LR—relying solely on sequence input—matched SaPS-10's accuracy for PS-Self proteins and improved AUC for PS-Part proteins (Fig. S3A). It also surpassed PhaSePred-10 on the hPS-test set under both PS-Self and PS-Part training. These results highlight the superiority of the PSTP embedding across PS proteins from various species.

We used a sliding-window approach to handle sequences over 256 residues during ESM-2 embedding (Fig. 1A, Fig. S9A and B, Supplementary Materials and Methods), thereby mitigating the high memory demands of Transformer-based attention for long sequences (O(Sequence length^2) complexity). Expanding this threshold to 1024 residues—maximum length during ESM-2 training [30]—did not improve accuracy (Fig. S2D), affirming our initial configuration's efficiency.

We also assessed PSTP on membraneless organelle (MLO) datasets. Proteins in these datasets likely undergo PS or co-PS with other proteins to form MLOs, and the PS-Part proteins trained PSTP model demonstrated strong performance in identifying these proteins (Fig. S3B).

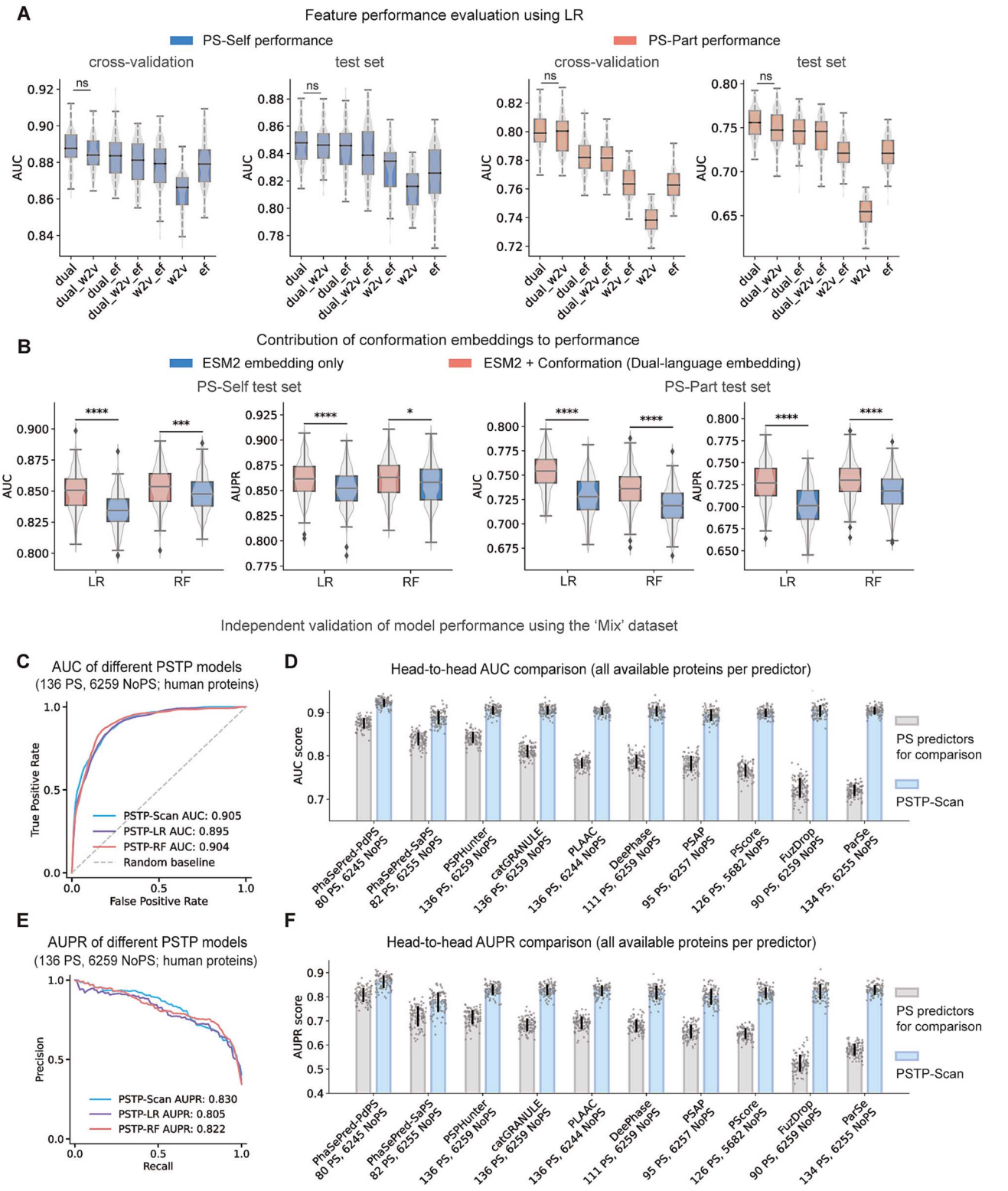


Figure 3. Evaluation of PSTP model performance. (A and B) Model performance of different feature combinations in predicting PS-Self and PS-Part proteins, evaluated through cross-validation and independent testing. Performance metrics were calculated using 50 replicates of subset sampling from the background dataset (LR: logistic regression; RF: random forest; two-tailed Student's *t*-test; boxplot components within each violin, from top to bottom, are maxima, upper quartile, median, lower quartile, and minima). (A) Performance of different feature combinations using LR. "Dual": default PSTP embedding, "w2v": Word2Vec embedding, "ef": engineered features. (B) Comparison of the default PSTP embeddings versus excluding conformational embeddings. Excluding conformational embeddings (using only ESM-2 embeddings) reduced performance, while adding other features did not improve performance. (C) Model performances of different PSTP models on the independent validation dataset of PS-Mix proteins. Performance evaluated using 100 replicates of subset sampling from the background dataset is shown. (D) Head-to-head comparison between PSTP-Scan and other representative PS predictors. PSTP-Scan returns scores for all proteins; the comparison with each method is thus performed on a set of proteins that can be predicted by that method, excluding those used in training for that method. (E and F) Parallel evaluations to (D and E) but focusing on the AUPR (area under the precision–recall curve) values.

Overall, the high predictive accuracy underscores the significance of conformational features in PS and demonstrates the capability of pre-trained language models to capture the latent sequence grammar essential for PS.

PSTP-Scan provides reliable protein-level PS prediction

The previous section showed that the averaged PSTP embedding vectors perform well in protein-level prediction. Since PS is often related to specific regions within a protein, we suggest focusing on the averaged vectors from these local regions instead of the entire protein to capture key PS-prone areas. Drawing inspiration from image-based spatial attention neural networks, which assign attention scores based on local context [40–43], we developed PSTP-Scan, which generates residue-level PS profiles and predicts the protein's overall PS propensity. PSTP-Scan, which generates scores based on potential PS regions, and PSTP-LR, which utilizes the entire sequence, exhibited similar performance in predicting PS-Part and PS-Self regions (Fig. S3C).

To further evaluate PSTP's robustness, we used the "Mix" dataset curated by Sun *et al.*, a more recent study [24], which consists of 892 PS proteins (from LLPSDB [44], PhaSepDB [45], DrLLPS [46], and PhaSePro [31]) and 8897 single-domain human proteins as background samples. Redundancy was removed using CD-Hit [47] (0.4 threshold), yielding a test set of 136 human PS proteins and 6259 background proteins. The "Mix" training set contained 375 PS proteins and 17 288 background proteins from other species.

We first performed a 5-fold cross-validation using the Mix training dataset, examining how different hyperparameter settings, such as learning rate and regularization rate, affect PSTP-Scan performance (see Table S1). The results demonstrate stability across configurations: only extreme learning rates caused marginal performance degradation (~3% AUC drop), while other parameter variations had little impact.

During PSTP-Scan training, in each epoch (50 epochs in total), negative samples were randomly downsampled to match the positive dataset size. To access the impact of negative subset size, we conducted 5-fold cross-validation on the Mix training dataset. Results showed that increasing the negative subset size reduced performance (Fig. S9E, left), but applying a weight-balanced loss function made performance relatively insensitive to subset size (Fig. S9E, right). Our approach of matching negative and positive samples ensured both model performance and computational efficiency while reducing training time compared to larger negative datasets.

PSTP-LR, PSTP-RF, and PSTP-Scan showed strong performance in distinguishing human PS proteins from background human proteins (Fig. 3C and E, Fig. S3D). We also tested a single PSTP-Scan block with different window sizes, observing robust results independent of size (Fig. S3E).

Unlike some predictors that cannot score all proteins, PSTP provides scores for all proteins. In the head-to-head comparison with other PS methods, we included only proteins that were predicted by the compared methods and were absent from their training sets. PSTP-Scan achieves an AUC of ~0.9 for all comparisons, outperforming representative methods (Fig. 3D and F).

These results show that leveraging large language models and conformation-aware embeddings enables efficient PS feature embedding, offering a simplified and reliable PS prediction method using only protein sequences.

When evaluated on UniProt-reviewed human proteins, PSTP-Scan on average processes 100 sequences in 14.9 s, faster than representative methods including DeePhase, FuzDrop, and PScore (Fig. S9C). Although the integration of ESM-2 large language model embeddings results in a memory requirement of ~750 MB (Fig. S9D), PSTP's efficiency remains practical for hardware deployment. We used the full "Mix" dataset to train the PSTP (Mix) model and the full PS-Self and PS-Part datasets to train PSTP (SaPS) and PSTP (PdPS) models, respectively.

PSTP-Scan improves identification of PS-driving region

Using the PhaSePro [31] database of experimentally validated PS regions, derived from proteins across different species, we evaluated the residue-level PS scores (the attention layer output) generated by PSTP-Scan (Fig. 1C). Although no residue-level PS information was included in PSTP-Scan's training, all sub-models (SaPS, PdPS, and Mix) here were specifically trained without including any PhaSePro proteins (Fig. S5A) to ensure unbiased testing.

We found that, out of 143 PhaSePro regions, 120 overlapped with the PSTP-Scan's PS regions, a notable improvement over the 109 overlaps achieved by FuzDrop, which was directly trained on PhaSePro PS regions (Fig. 4B, Fig. S5C and D). PSTP-Scan recovered 28 378 residues in predicted PS regions, compared to FuzDrop's 25 014 (Fig. 4B, Fig. S5C and D). Additionally, PSTP-Scan outperformed PSPHunter [24], a method recently reported to predict key regions for PS (Fig. S5C and D).

The *t*-test statistics showed that PSTP-Scan outperforms other predictors in distinguishing residues in PS-driving regions from non-PS regions within those PS proteins (Fig. 4C). We also calculated the Spearman correlation between residue-level scores and binary annotations of experimentally validated PS regions (1 for PS positions, 0 for non-PS positions) (Fig. 4D). We calculated both the overall Spearman correlation and the correlation score for each protein. PSTP-Scan achieved significantly higher correlations compared to other predictors, reaching ~150% of FuzDrop's performance. In terms of protein-specific correlation, PSTP-Scan also demonstrated much higher performance over other methods (Fig. 4E and Fig. S5B).

Figure 4F and G and Fig. S6 showcase PSTP-Scan prediction results. The PSTP-Scan (SaPS) model, trained on PS-Self proteins, identifies PS-driving regions like the low-complexity domain (LCD) repeats in cortactin [48], and the N terminus IDR of Ddx4, which contains a repeating net charge block and FG, RG blocks [49]. The PSTP-Scan (PdPS) model, trained on PS-Part proteins, outperforms PSTP-Scan (SaPS) in predicting PS regions that rely on partner interactions. This includes the phosphotyrosine motif of the nephrin and the proline-rich domain of N-WASP (Fig. 4G) which form multivalent interactions with the SH2 and SH3 domains of NCK, respectively [50]. For proteins with both RNA-recognition motifs (RRMs) and LCDs, such as PAB1 [39] and Nab3, PSTP-Scan (PdPS) identifies the RRMs, while PSTP-Scan (SaPS) focuses on the LCDs (Fig. 4G).

PSTP-Scan achieved an overall correlation of 0.4 for PS-Self proteins (58 records, Fig. S5E and H); second is PLAAC, a method to screen prion-like regions. All models showed reduced performance for PS-Part proteins, with PSTP-Scan remaining the best performer (50 records, Fig. S5F and I). Lower performance for PS-Part proteins may result from some protein–protein interaction sites that are also denoted as PS regions in the PhaSePro database, the prediction of which requires prior knowledge of specific protein interactions. For instance, PSTP-Scan did not predict the tri-RG motif in Buc that interacts with the regulator Tdrd6a [51], but successfully predicted the prion-like domain responsible for

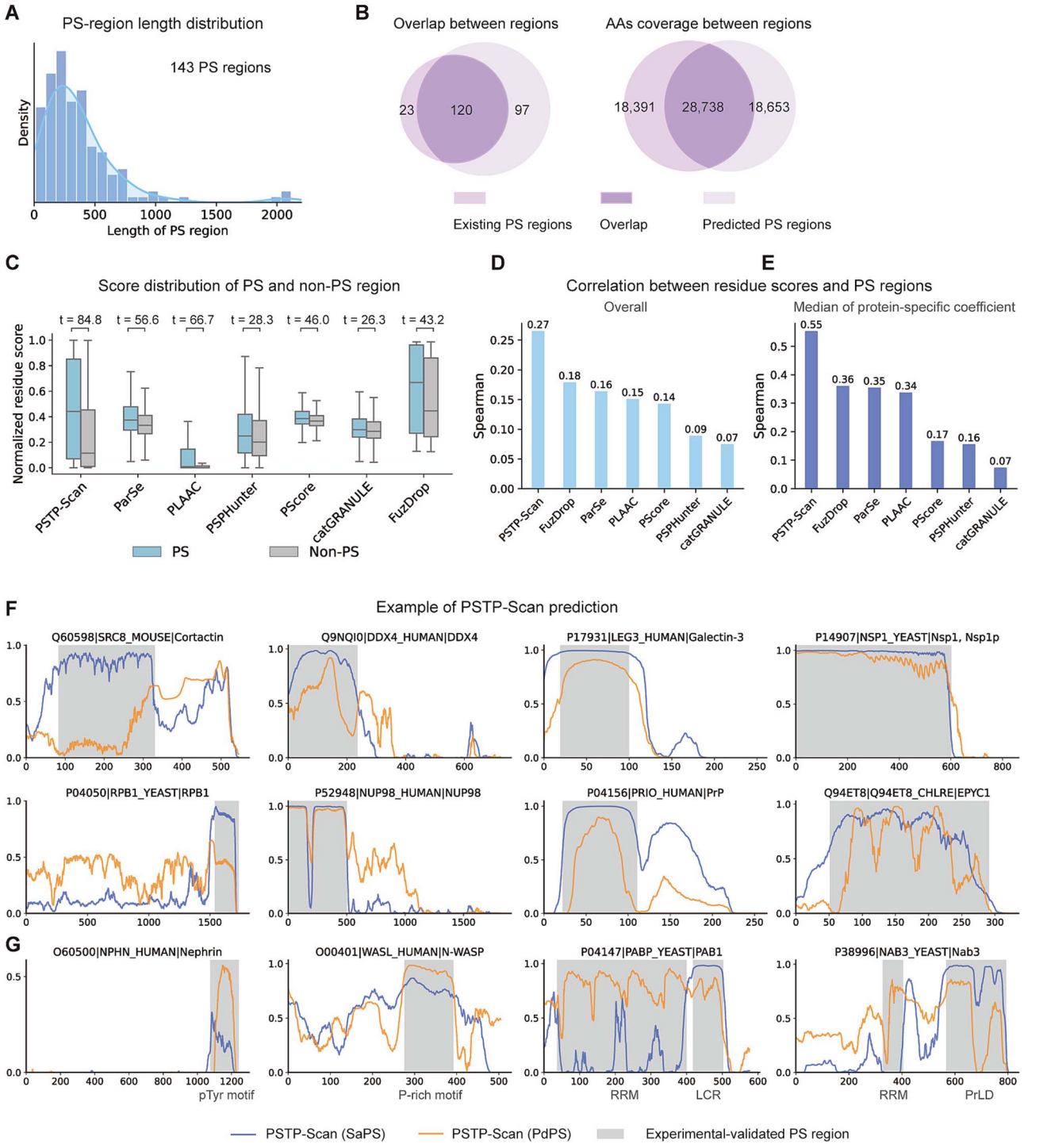


Figure 4. Evaluation of residue-level prediction scores generated by PSTP-Scan and other PS predictors under experimentally validated PS-driving regions in the PhaSePro database (143 PS regions from 121 proteins). (A) Distribution of sequence lengths for regions that drive PS. (B) Comparison of PS regions predicted by PSTP-Scan against those documented in PhaSePro. This panel highlights the overlap between predicted regions and the corresponding overlap at the individual amino acid level. (C) Distribution of the residue-level scores for residues in PS regions versus non-PS regions in proteins documented in PhaSePro. All components have P-values $<.0001$. (D) Spearman correlation coefficient computed between residue-level scores and PS regions documented in PhaSePro for each method. The coefficients are computed between joint residue-level score vectors for each protein and a binary vector, where corresponding positions in PS regions are assigned 1 and others are assigned 0. (E) Median Spearman correlation coefficients at the protein level, computed between residue-level scores and PS regions for each protein. (F and G) Example of residue-level scores predicted by three PSTP-Scan models, with experimentally validated PS regions highlighted in gray. pTyr motif: phosphotyrosine motif; P-rich motif: proline-rich motif; RRM: RNA-recognition motifs; LCD: low-complexity domain; PrLD: prion-like domain.

Buc's self-assembly [52] (Fig. S6B), which is absent in the PhaSePro database. Similarly, while PSTP-Scan predicted the tandem repeats for the scaffold protein EPYC1, neither PSTP-Scan nor FuzDrop predicted the alpha-helices of Rubisco that bind to EPYC1 [53] (Fig. S6C). Incorporating models that focus on predicting protein–protein interactions may provide better insights into these types of PS mechanisms.

The varying performance of current methods may also stem from the highly flexible definitions of PS regions. For example, PSTP-Scan achieves a 0.43 correlation with PS regions identified by NMR experiments (Fig. S5G and J), higher than the overall correlation. Nevertheless, given that PSTP-Scan predicts diverse PS-driving regions without being specifically trained on distinct PS regions, its ability to directly generate these predictions effectively demonstrates its validity, offering detailed insights into protein PS behavior.

PSTP-Scan predicts PS in truncated and non-natural proteins

To evaluate PSTP's generalizability on protein variants, we compared PSTP-Scan and other predictors on a dataset of 93 experimentally truncated PS proteins and 242 non-PS proteins curated by Yang *et al*. [54]. Truncated proteins, often caused by (Nonsense-Mediated Decay) NMD-escaping mutations that bypass nonsense-mediated decay, are critical for understanding protein function and pathogenicity [55].

To avoid information leakage, each PSTP model evaluated here (PSTP-SaPS, PSTP-PdPS, and PSTP-Mix) was specifically trained without using any proteins from the test dataset's UniProt IDs (Fig. S7A). Compared with other representative sequence-based predictors, PSTP achieves the highest performance in both AUC and AUPR (Fig. 5B and Fig. S7B). The PSTP-Scan MLP kernel (Fig. 5A and Fig. 5B), designed specifically for short sequence processing, achieves the highest performance, with an AUC of 0.88 (Fig. 5B). Unlike TruncPS [54], which was tested using a cross-validation approach on these truncated PS proteins, PSTP-Scan achieved high performance without additional training on this protein type. This demonstrates PSTP-Scan's robustness and effectiveness in predicting PS for truncated proteins.

We also explore PSTP's applicability to non-natural proteins using artificial intrinsically disordered proteins (A-IDPs) designed to assess how amino acid sequence patterns regulate PS from a recent study [56]. This study generated 11 synthetic polypeptides based on prior research [57, 58], each comprising a 200-residue sequence formed by repeating an octapeptide 25 times. These A-IDPs exhibited PS at micromolar concentrations; most did so at room temperature.

The PSTP-Scan (Mix) kernel assigned high scores to these peptides. To compare, we generated 10 000 random background proteins by truncating 200-residue segments from UniProt-reviewed proteins (Fig. 5C, "Uniprot random"). All models, except PLAAC, significantly distinguished A-IDPs from these background proteins, highlighting the unique sequence properties of these PS A-IDPs (Fig. 5C). However, PSAP produced relatively low scores, which failed to predict these peptides. Next, we generated control peptides by repeating random octapeptides 25 times (matching the repetition number of the PS A-IDPs), with the octapeptides generated based on the same amino acid composition weights as the A-IDPs (Fig. 5C, "AA-weighted random IDPs"). Interestingly, PSTP-Scan predicted significantly lower scores for these control peptides compared to the carefully designed A-IDPs (*P*-value = .0085), while other predictors did not show any significant differences. This indicates that PSTP-Scan understands these A-IDPs by capturing information on both amino acid composition and the specific order of residues.

Both results indicate PSTP-Scan's sensitivity to sequence information and its flexibility in adapting to different prediction scenarios. Leveraging the high-throughput nature of PSTP, we predict >42 000 proteins in the human proteome, including canonical and isoforms, as well as >570 000 Uniprot-reviewed proteins. We found that ∼35% of human proteome and isoform proteins, and 27% of UniProt-reviewed proteins, contain potential PS regions (PSTP-Scan residue-level score > 0.5) (Fig. 5E). This proportion increases to 39% and 35%, respectively, when restricted to proteins with at least one IDR (Fig. 5F).

Incidence of pathogenic variants

Analyzing mutations in low-evolutionarily conserved IDR regions has been challenging [59], comprising ∼52% (∼530 000) of unresolved ClinVar VUSs. Given the link between IDRs and PS, we investigated the relationship between PS-driving propensity and the occurrence of pathogenic variants in these low-conserved IDRs.

We analyzed PSTP-Scan residue-level scores for ClinVar missense variants located in low-conservation IDRs (Alphafold2 pLDDT <50) (Fig. 6A). pLDDT scores, reflecting AlphaFold2 prediction confidence and correlating with evolutionary conservation, serve as a state-of-the-art IDR predictor [17]. Among 4756 pathogenic and 37 291 benign variants, pathogenic variants tend to be located in positions with higher residue-level PS propensity (Fig. 6B), and all PSTP-Scan sub-models detected this difference (*P*-value $<1\times10^{-10}$, Mann–Whitney test). When we restricted the analysis to variants in PS proteins, this trend still remains, but possibly due to the relatively small number of mapped pathogenic variants, only PSTP-Scan (SaPS) exhibited a statistically significant *P*-value (Fig. S10A). Furthermore, within low-pLDDT IDRs, 119 B/LB and 54 P/LP variants were mapped to PhaSePro PS regions, while the remaining 79 B/LB and 11 P/LP variants were not. A Fisher's exact test on these counts yielded an odds ratio of 3.26 ($P=8\times10^{-4}$), indicating that P/LP variants are ∼3.26 times more likely to be located in PS regions compared to B/LB variants, further supporting our hypothesis.

Examples of pathogenic IDR variants located in high-PS regions (>0.8 in any PSTP-Scan sub-model) include P39L in HSPB1 [60], F89I in DNAJB6 [61], and A315T in TARDBP [62, 63], associated with Charcot–Marie–Tooth disease type 2, limb-girdle muscular dystrophy type 1D, and amyotrophic lateral sclerosis, respectively. These pathogenic variants promote abnormal protein aggregation [60–62], but are predicted as either VUS or benign by EVE [59], an evolutionary-based top-performance pathogenicity predictor. For IUPred3-defined IDRs (IUPred score > 0.5), the distinction was less pronounced (Fig. S7E) compared to pLDDT-defined IDRs. This might be because IUPred-predicted IDRs include regions that are evolutionarily conserved [64].

Next, we evaluated this correlation based on allele frequency (AF), as variants with very low AF are more likely to be pathogenic than highly frequent variants [35, 65]. Using variants from gnomAD V4.1.0, we found that in low-conservation regions (pLDDT <50), variants with low AF ($AF<1\times10^{-5}$, 3 841 613 variants) tend to be located in positions with higher residue-level PS score than variants with higher AF ($AF\geq0.001$, 25 673) (Fig. S7D and F). In contrast, this pattern was not observed for variants located in structured regions (pLDDT ≥70) (Fig. S7C), where other factors like protein evolution and function play a larger role in pathogenicity.

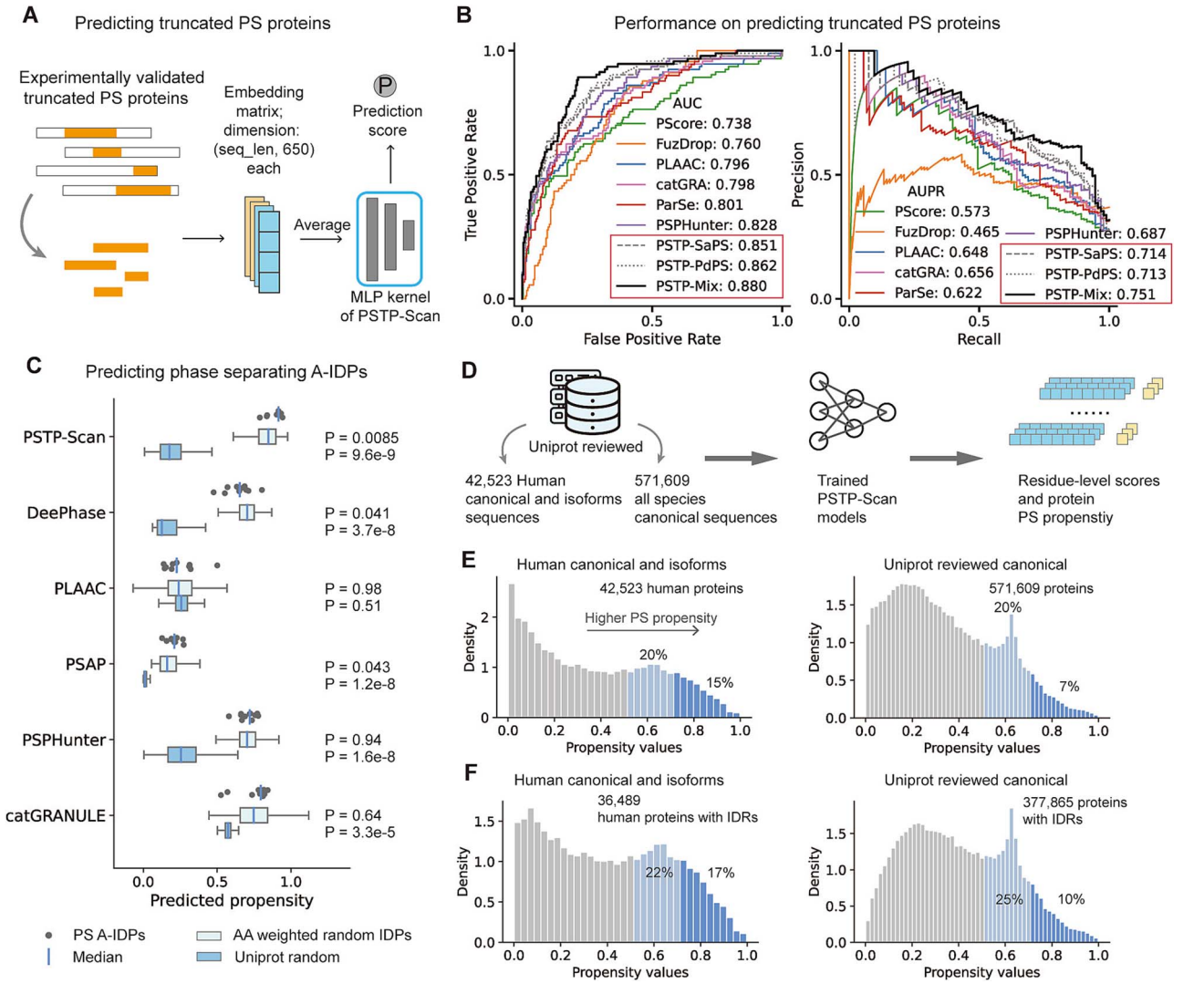


Figure 5. Extended scope of PS protein prediction. (A) Schematic of truncated protein prediction using the trained MLP kernel from PSTP-Scan. (B) Performance evaluation of different models for predicting experimentally validated truncated PS proteins and non-PS truncated proteins, assessed by AUC (left) and AUPR (right). (C) Score distribution comparison across models for predicting 11 artificial intrinsically disordered proteins (A-IDPs) that undergo PS, alongside random proteins. "Uniprot random" represents scores for random 200-residue segments from UniProt-reviewed proteins ($n = 10\,000$), while "AA-weighted random IDPs" refers to scores for control peptides, generated by repeating random octapeptides 25 times (same repeating number as the PS A-IDPs group), with the octapeptides generated based on the same amino acid composition weights as the A-IDPs. (D) PSTP-Scan-generated sequence-level and residue-level predictions for >570 000 proteins from UniProt. (E) Distribution of PSTP-Scan scores across the human proteome, including both canonical proteins and isoforms (left), and for all UniProt-reviewed proteins (right). PSTP-Scan scores are determined by selecting the highest residue-level score from the three sub-models. (F) A parallel evaluation as (E), but with the inclusion of only proteins containing at least one IDR (IUPred >0.5, length unrestricted).

We observed that within disordered PS regions, glycine mutations, such as neurodegeneration-related G294V [66] and G298S [67] in TARDBP, are most common (Fig. 6C and D). These mutations replace flexible glycine residues with more rigid or charged amino acids, potentially altering dynamics and disrupting the multivalent interactions of key PS regions.

The results suggest that alterations in PS-driving regions are, on average, more likely to lead to be pathogenic than alternation in non-PS regions, even though these PS-driving regions are evolutionary less conserved.

In addition to missense variants, we extracted 17 304 protein-altering variants, including nonsense, in-frame deletions, and frameshift variants (Fig. 6E) from ClinVar. We found that pathogenic mutations more significantly alter PS propensity than benign ones, as supported by the two-sample KS test (Fig. 6F). This trend persisted across both the full dataset and the most frequent in-frame deletion mutation type (Fig. S7G–I).

Our findings suggest that altering PS-driving regions, even in less-conserved disordered regions, are factors that contribute to pathogenicity. Combining the effect on PS with traditional structure- and function-based pathogenicity analyses may enhance our understanding of disease mechanisms.

Discussion

Protein language models have made significant progress in recent years. Although language model–based embeddings may be less interpretable than engineered features, they capture comprehensive sequence characteristics more objectively, thus improving generalizability. For broader applicability, we used the

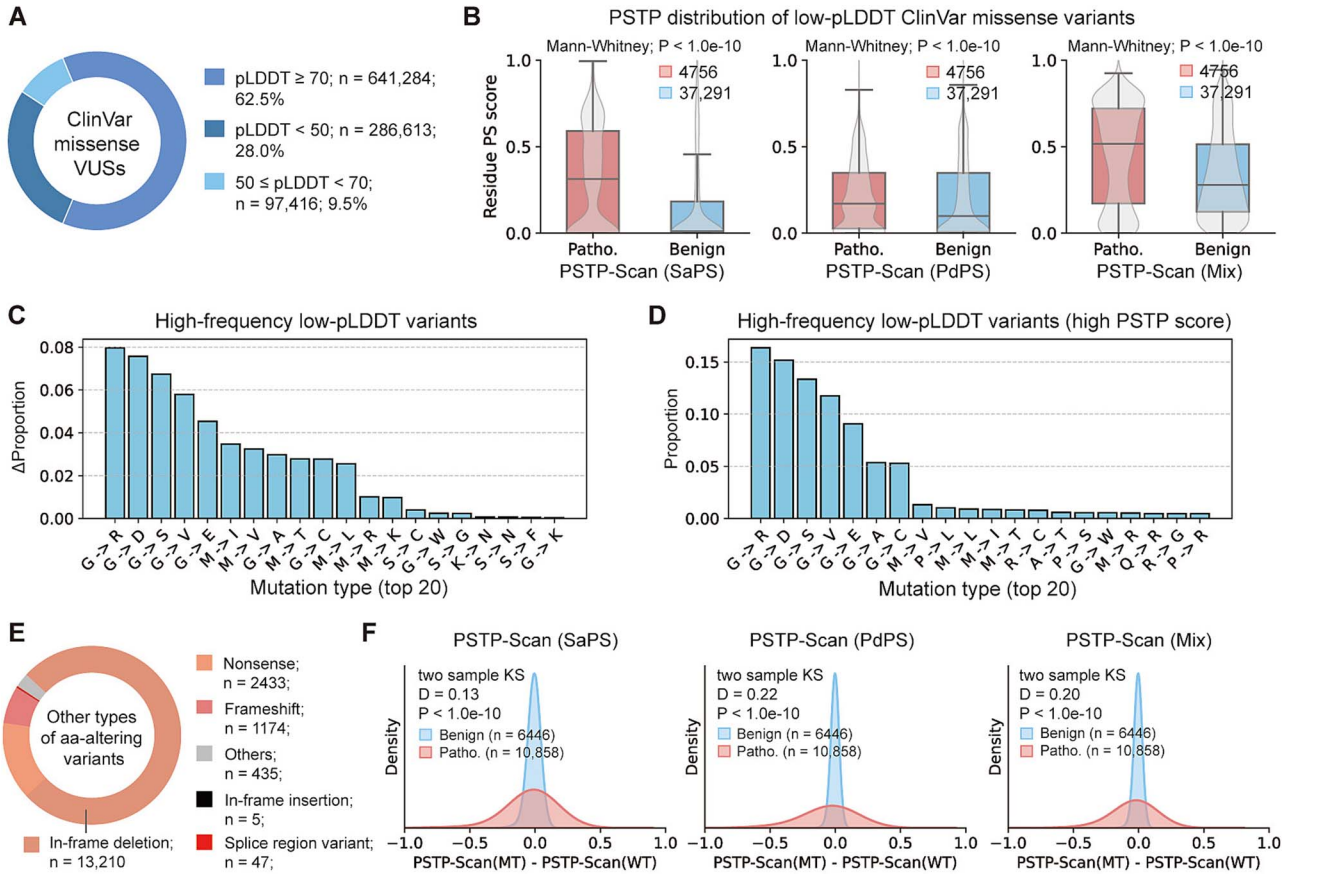


Figure 6. Analysis of the incidence of pathogenic variants and PS propensity. (A) Proportion of variants of uncertain significance (VUSs) based on pLDDT scores. (B) Distribution of PSTP-Scan residue-level scores for ClinVar variant positions with low-pLDDT (pLDDT <50), classified as pathogenic (pathogenic/likely pathogenic) and benign (benign/likely benign). (C) Top 20 high-frequency mutations among low-pLDDT (pLDDT <50) pathogenic variants. For each mutation type, ΔProportion is computed as the difference in proportion between the low-pLDDT (pLDDT <50) pathogenic group and other pathogenic variants (pLDDT ≥50). (D) Top 20 high-frequency mutations among low-pLDDT (pLDDT <50) pathogenic variants with high PSTP-Scan residue-level scores (score > 0.5 from any sub-model). (E) Frequency of ClinVar protein-altering variants, excluding missense variants. (F) Distribution of PSTP-Scan score variations before and after mutations for pathogenic and benign ClinVar variants described in (E). Pathogenic variants show significantly higher PSTP-Scan score variations compared to benign variants.

smallest version of the ESM-2 model and employed a sliding-window approach to reduce memory usage, as its performance already demonstrates the effectiveness of the proposed approach. While language models can extract rich evolutionary information from sequences, IDRs are relatively less conserved. Recent studies suggest that while IDRs may undergo significant sequence variation throughout evolution, they can maintain stable conformational properties [16, 17]. Therefore, we consider that the conformational information of IDRs, which is closely linked to phase separation (PS) propensity, might be better captured by models specifically trained using MD simulation data. To address this, we designed an embedding approach that utilizes a pre-trained long short-term memory (LSTM) neural network to encode the dynamic properties of IDRs. While both types of embeddings alone can provide PS predictions using traditional machine learning models like logistic regression or random forest, ESM-2 outperforms conformational features due to its ability to better capture the full sequence landscape. However, their combination performs significantly better, confirming the effectiveness of our dual-language model embedding.

Unlike existing predictors, such as FuzDrop [21, 38] or PScore [11], which employ supervised learning models trained at each residue position, PSTP-Scan was trained without incorporating residue-level PS information, allowing it to learn autonomously. Due to the highly flexible definitions of PS-driving or regulating regions, the performance of current methods is often limited. However, PSTP-Scan's design, which leverages local sequence information, makes it successfully predict many PS-driving segments, outperforming other models and providing important insights into numerous uncharacterized sequences. PSTP-Scan also provides accurate protein-level PS prediction by focusing on the local region most likely to drive PS, aligning with the role of PS-driving regions in facilitating multivalent interactions with other protein regions [48, 51, 52, 68, 69].

Given that the training data size for phase-separating proteins is relatively small compared to other deep learning fields, to prevent overfitting, we avoided using overly complex deep models like Transformer-based attention layers with query-key-value mechanisms [70]. We also chose not to use attention weights from the ESM-2 model, as they may emphasize structural and functional regions, introducing potential noise. Instead, we constructed an attention calculation method where only the weights of a multilayer perceptron (MLP) network are trainable.

To better handle dataset imbalance caused by using a large "background dataset" as negative samples—a common training approach for current PS predictors [23–25, 71]—we prevented the neural network from fitting the same data in every iteration.

Specifically, in each epoch of PSTP-Scan training, we randomly select a small subset of proteins from the background dataset to serve as the negative dataset. This randomization helps prevent overfitting on background proteins that have not been experimentally confirmed as non-PS.

The disordered nature of IDRs not only makes the interpretation of their variants challenging but also limits the effectiveness of evolutionary or structure-based variant effect predictors like EVE [59] or AlphaMissense [35]. This hinders the discovery and study of pathogenic variants. Our results reveal a link between the pathogenicity of low-conservation IDR variants and PS. Diseases related to PS often involve both abnormal protein aggregation and may also be linked to dosage effects [54]. To better interpret these variants' pathogenicity, advanced methods are needed to capture and understand the evolution of IDRs, given their sequence divergence yet conformational conservation [16, 17].

In addition, for multicomponent PS systems, particularly those involving multi-tandem domains, where protein–protein and protein–nucleic acid interaction play a dominant role, mutations that impair protein–protein interaction binding affinity can disrupt this kind of PS processes, potentially leading to pathogenic outcomes. While our current work focuses on single-component PS systems and partner-dependent PS scaffold proteins, future research integrating co-phase separation experimental data and PPI-informed models [72] will improve our understanding of these systems and the pathogenic impact of mutations.

Biomolecular condensates perform essential cellular functions through complex and diverse interaction patterns. With its high performance, robustness, and sequence-only input nature, PSTP serves as a reliable tool for predicting biomolecular condensate formation. We believe that PSTP will support a deeper understanding of cellular processes, aiding in disease research and therapeutic development.

Key Points

- We introduce PSTP, a dual-language model method that uses conformational-aware embeddings to encode protein sequences, demonstrating that sequence data alone can achieve state-of-the-art performance in phase separation prediction without requiring extensive annotations or handcrafted features.
- We propose a framework that provides accurate residue-level scores, closely correlating with experimentally validated phase separation regions, without requiring additional residue-level task-specific training.
- PSTP reliably predicts various types of phase separation, including self-assembly phase separation proteins, truncated proteins, and partner-dependent phase separation proteins, and shows potential for predicting artificial intrinsically disordered proteins.
- A strong correlation between pathogenic variants and PS propensity in low-conservation regions is observed, offering new insights into variants of uncertain significance.
- PSTP's lightweight design ensures portability and scalability, facilitating rapid execution on normal CPUs and GPUs.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Author contributions

M.F., Q.L., G.H., and Y.S. conceived the concept; M.F. and L.L. curated datasets, and designed computational algorithms and experiments; M.F., L.L., X.W., K.L., and Q.Y. performed computational experiments with assistance from Q.L., G.H., and Y.S.; M.F. and Z.X. analyzed data; G.H. led the project with assistance from M.F., Q.L., Y.S., and X.W.; M.F., Q.L., G.H., and Y.S. wrote the manuscript with input from all co-authors.

Conflict of interest: None declared.

Funding

This work was supported by the National Key Research and Development Program (2024YFC2707002, 2022YFE0125300), Innovation Program of Shanghai Municipal Education Commission (2023ZKZD16), National Natural Science Foundation of China (82071262, 32300464, 82301682, 32450663, 32470742), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01, 20JC1418600), China Postdoctoral Science Foundation (2023M732266), Key Technology Breakthrough Program of Ningbo Sci-Tech Innovation YONGJIANG 2035 (2024Z221), Municipal Public Welfare Project (2022S035), and Shanghai Jiao Tong University STAR Grant (YG2023ZD26, YG2022ZD024, YG2022QN111, YG2023LC14, YG2024QNA59, 23X010300421).

Data and code availability

All data mentioned in this paper are public data and can be obtained through the corresponding description in the results or methods. An installable Python package for the software is available at https://github.com/Morvan98/PSTP. The paper code is available at https://github.com/Morvan98/paper_code_PSTP.

References

1. Tsang B, Pritišanac I, Scherer SW. *et al.* Phase separation as a missing mechanism for interpretation of disease mutations. *Cell* 2020;**183**:1742–56. https://doi.org/10.1016/j.cell.2020.11.050
2. Gao Y, Li X, Li P. *et al.* A brief guideline for studies of phase-separated biomolecular condensates. *Nat Chem Biol* 2022;**18**:1307–18. https://doi.org/10.1038/s41589-022-01204-2
3. Alberti S. Phase separation in biology. *Curr Biol* 2017;**27**:R1097–102. https://doi.org/10.1016/j.cub.2017.08.069
4. Lyon AS, Peeples WB, Rosen MK. A framework for understanding the functions of biomolecular condensates across scales. *Nat Rev Mol Cell Biol* 2021;**22**:215–35. https://doi.org/10.1038/s41580-020-00303-z
5. Oltrogge LM, Chaijarasphong T, Chen AW. *et al.* Multivalent interactions between CsoS2 and rubisco mediate α-carboxysome formation. *Nat Struct Mol Biol* 2020;**27**:281–7.
6. Singatulina AS, Hamon L, Sukhanova MV. *et al.* PARP-1 activation directs FUS to DNA damage sites to form PARG-reversible compartments enriched in damaged DNA. *Cell Rep* 2019;**27**:1809–1821.e1805. https://doi.org/10.1016/j.celrep.2019.04.031

7. Liu X, Shen J, Xie L. *et al.* Mitotic implantation of the transcription factor Prospero via phase separation drives terminal neuronal differentiation. *Dev Cell* 2020;**52**:277–293.e278. https://doi.org/10.1016/j.devcel.2019.11.019
8. Alberti S, Dormann D. Liquid-liquid phase separation in disease. *Annu Rev Genet* 2019;**53**:171–94. https://doi.org/10.1146/annurev-genet-112618-043527
9. Pappu RV, Cohen SR, Dar F. *et al.* Phase transitions of associative biomacromolecules. *Chem Rev* 2023;**123**:8945–87. https://doi.org/10.1021/acs.chemrev.2c00814
10. Mittag T, Pappu RV. A conceptual framework for understanding phase separation and addressing open questions and challenges. *Mol Cell* 2022;**82**:2201–14. https://doi.org/10.1016/j.molcel.2022.05.018
11. Vernon RM, Chong PA, Tsang B. *et al.* Pi-pi contacts are an overlooked protein feature relevant to phase separation. *Elife* 2018;**7**:e31486. https://doi.org/10.7554/eLife.31486
12. Wang J, Choi JM, Holehouse AS. *et al.* A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell* 2018;**174**:688–699.e616. https://doi.org/10.1016/j.cell.2018.06.006
13. Dignon GL, Best RB, Mittal J. Biomolecular phase separation: from molecular driving forces to macroscopic properties. *Annu Rev Phys Chem* 2020;**71**:53–75. https://doi.org/10.1146/annurev-physchem-071819-113553
14. Martin EW, Holehouse AS, Peran I. *et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* 2020;**367**:694–9. https://doi.org/10.1126/science.aaw8653
15. Brangwynne Clifford P, Tompa P, Pappu RV. Polymer physics of intracellular phase transitions. *Nat Phys* 2015;**11**:899–904. https://doi.org/10.1038/nphys3532
16. Lotthammer JM, Ginell GM, Griffith D. *et al.* Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat Methods* 2024;**21**:465–76. https://doi.org/10.1038/s41592-023-02159-5
17. Tesei G, Trolle AI, Jonsson N. *et al.* Conformational ensembles of the human intrinsically disordered proteome. *Nature* 2024;**626**:897–904. https://doi.org/10.1038/s41586-023-07004-5
18. Lancaster AK, Nutter-Upham A, Lindquist S. *et al.* PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics* 2014;**30**:2501–2. https://doi.org/10.1093/bioinformatics/btu310
19. Alberti S, Halfmann R, King O. *et al.* A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. *Cell* 2009;**137**:146–58. https://doi.org/10.1016/j.cell.2009.02.044
20. Bolognesi B, Lorenzo Gotor N, Dhar R. *et al.* A concentration-dependent liquid phase separation can cause toxicity upon increased protein expression. *Cell Rep* 2016;**16**:222–31. https://doi.org/10.1016/j.celrep.2016.05.076
21. Hardenberg M, Horvath A, Ambrus V. *et al.* Widespread occurrence of the droplet state of proteins in the human proteome. *Proc Natl Acad Sci U S A* 2020;**117**:33254–62. https://doi.org/10.1073/pnas.2007670117
22. Paiz EA, Allen JH, Correia JJ. *et al.* Beta turn propensity and a model polymer scaling exponent identify intrinsically disordered phase-separating proteins. *J Biol Chem* 2021;**297**:101343. https://doi.org/10.1016/j.jbc.2021.101343
23. van Mierlo G, Jansen JRG, Wang J. *et al.* Predicting protein condensate formation using machine learning. *Cell Rep* 2021;**34**:108705. https://doi.org/10.1016/j.celrep.2021.108705
24. Sun J, Qu J, Zhao C. *et al.* Precise prediction of phase-separation key residues by machine learning. *Nat Commun* 2024;**15**:2662. https://doi.org/10.1038/s41467-024-46901-9
25. Chen Z, Hou C, Wang L. *et al.* Screening membraneless organelle participants with machine-learning models that integrate multimodal features. *Proc Natl Acad Sci U S A* 2022;**119**:e2115369119. https://doi.org/10.1073/pnas.2203894119
26. Liang Q, Peng N, Xie Y. *et al.* MolPhase, an advanced prediction algorithm for protein phase separation. *EMBO J* 2024;**43**:1898–918. https://doi.org/10.1038/s44318-024-00090-9
27. Saar KL, Morgunov AS, Qi R. *et al.* Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proc Natl Acad Sci U S A* 2021;**118**:e2019053118. https://doi.org/10.1073/pnas.2019053118
28. Liao S, Zhang Y, Han X. *et al.* A sequence-based model for identifying proteins undergoing liquid-liquid phase separation/forming fibril aggregates via machine learning. *Protein Sci* 2024;**33**:e4927. https://doi.org/10.1002/pro.4927
29. Frank M, Ni P, Jensen M. *et al.* Leveraging a large language model to predict protein phase transition: a physical, multiscale, and interpretable approach. *Proc Natl Acad Sci U S A* 2024;**121**:e2320510121. https://doi.org/10.1073/pnas.2320510121
30. Lin Z, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574
31. Mészáros B, Erdős G, Szabó B. *et al.* PhaSePro: the database of proteins driving liquid–liquid phase separation. *Nucleic Acids Res* 2019;**48**:D360–7.
32. Landrum MJ, Kattman BL. ClinVar at five years: delivering on the promise. *Hum Mutat* 2018;**39**:1623–30. https://doi.org/10.1002/humu.23641
33. Landrum MJ, Chitipiralla S, Brown GR. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res* 2020;**48**:D835–d844. https://doi.org/10.1093/nar/gkz972
34. Jumper J, Evans R, Pritzel A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2
35. Cheng J, Novati G, Pan J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 2023;**381**:eadg7492.
36. Chu X, Sun T, Li Q. *et al.* Prediction of liquid-liquid phase separating proteins using machine learning. *BMC Bioinformatics* 2022;**23**:72. https://doi.org/10.1186/s12859-022-04599-w
37. Asgari E, Mofrad MR. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS One* 2015;**10**:e0141287. https://doi.org/10.1371/journal.pone.0141287
38. Hatos A, Tosatto SCE, Vendruscolo M. *et al.* FuzDrop on AlphaFold: visualizing the sequence-dependent propensity of liquid–liquid phase separation and aggregation of proteins. *Nucleic Acids Res* 2022;**50**:W337–44. https://doi.org/10.1093/nar/gkac386
39. Riback JA, Katanski CD, Kear-Scott JL. *et al.* Stress-triggered phase separation is an adaptive, evolutionary tuned response. *Cell* 2017;**168**:1028–1040.e1019. https://doi.org/10.1016/j.cell.2017.02.027
40. Arar M, Shamir A, Bermano AH. Learned queries for efficient local attention. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*. New Orleans, LA, USA: IEEE, 2022, 10831–42. https://doi.org/10.1109/CVPR52688.2022.01057
41. Chu X, Tian Z, Wang Y. *et al.* Twins: revisiting the design of spatial attention in vision transformers. In: Ranzato M, Beygelzimer A,

Dauphin Y, Liang PS, Vaughan JW. (eds), *35th Conference on Neural Information Processing Systems* (NeurIPS 2021). Online Conference, Canada: Curran Associates Inc., 2021;9355–66.
42. Ramachandran P, Bello I, Parmar N. *et al.* Stand-alone self-attention in vision models. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E. (eds), *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. Vancouver, Canada: Curran Associates Inc., 2019;9355–66.
43. Dai Z, Liu H, Le QV. *et al.* CoAtNet: marrying convolution and attention for all data sizes. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW. (eds), *35th Conference on Neural Information Processing Systems (NeurIPS 2021). Online Conference, Canada: Curran Associates Inc.*, 2021;3965–77.
44. Wang X, Zhou X, Yan Q. *et al.* LLPSDB v2.0: an updated database of proteins undergoing liquid-liquid phase separation in vitro. *Bioinformatics* 2022;**38**:2010–4. https://doi.org/10.1093/bioinformatics/btac026
45. You K, Huang Q, Yu C. *et al.* PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic Acids Res* 2020;**48**:D354–9. https://doi.org/10.1093/nar/gkz847
46. Ning W, Guo Y, Lin S. *et al.* DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. *Nucleic Acids Res* 2020;**48**:D288–95. https://doi.org/10.1093/nar/gkz1027
47. Fu L, Niu B, Zhu Z. *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2. https://doi.org/10.1093/bioinformatics/bts565
48. Saito M, Hess D, Eglinger J. *et al.* Acetylation of intrinsically disordered regions regulates phase separation. *Nat Chem Biol* 2019;**15**:51–61. https://doi.org/10.1038/s41589-018-0180-7
49. Nott Timothy J, Petsalaki E, Farber P. *et al.* Phase transition of a disordered Nuage protein generates environmentally responsive membraneless organelles. *Mol Cell* 2015;**57**:936–47. https://doi.org/10.1016/j.molcel.2015.01.013
50. Banjade S, Wu Q, Mittal A. *et al.* Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck. *Proc Natl Acad Sci U S A* 2015;**112**:E6426–35. https://doi.org/10.1073/pnas.1508778112
51. Roovers EF, Kaaij LJT, Redl S. *et al.* Tdrd6a regulates the aggregation of Buc into functional subcellular compartments that drive germ cell specification. *Dev Cell* 2018;**46**:285–301.e289. https://doi.org/10.1016/j.devcel.2018.07.009
52. Boke E, Ruer M, Wühr M. *et al.* Amyloid-like self-assembly of a cellular compartment. *Cell* 2016;**166**:637–50. https://doi.org/10.1016/j.cell.2016.06.051
53. Mackinder LCM, Meyer MT, Mettler-Altmann T. *et al.* A repeat protein links rubisco to form the eukaryotic carbon-concentrating organelle. *Proc Natl Acad Sci U S A* 2016;**113**:5958–63. https://doi.org/10.1073/pnas.1522866113
54. Yang L, Lyu J, Li X. *et al.* Phase separation as a possible mechanism for dosage sensitivity. *Genome Biol* 2024;**25**:17.
55. Supek F, Lehner B, Lindeboom RGH. To NMD or not to NMD: nonsense-mediated mRNA decay in cancer and other genetic diseases. *Trends Genet* 2021;**37**:657–68. https://doi.org/10.1016/j.tig.2020.11.002
56. Rekhi S, Garcia CG, Barai M. *et al.* Expanding the molecular language of protein liquid-liquid phase separation. *Nat Chem* 2024;**16**:1113–24. https://doi.org/10.1038/s41557-024-01489-x
57. Dzuricky M, Rogers BA, Shahid A. *et al.* De novo engineering of intracellular condensates using artificial disordered proteins. *Nat Chem* 2020;**12**:814–25. https://doi.org/10.1038/s41557-020-0511-7
58. Dai Y, Farag M, Lee D. *et al.* Programmable synthetic biomolecular condensates for cellular control. *Nat Chem Biol* 2023;**19**:518–28. https://doi.org/10.1038/s41589-022-01252-8
59. Frazer J, Notin P, Dias M. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* 2021;**599**:91–5. https://doi.org/10.1038/s41586-021-04043-8
60. Clouser AF, Baughman HE, Basanta B. *et al.* Interplay of disordered and ordered regions of a human small heat shock protein yields an ensemble of 'quasi-ordered' states. *Elife* 2019;8:e50259.
61. Sarparanta J, Jonson PH, Golzio C. *et al.* Mutations affecting the cytoplasmic functions of the co-chaperone DNAJB6 cause limb-girdle muscular dystrophy. *Nat Genet* 2012;**44**:450–5. https://doi.org/10.1038/ng.1103
62. Guo W, Chen Y, Zhou X. *et al.* An ALS-associated mutation affecting TDP-43 enhances protein aggregation, fibril formation and neurotoxicity. *Nat Struct Mol Biol* 2011;**18**:822–30. https://doi.org/10.1038/nsmb.2053
63. Moujalled D, Grubman A, Acevedo K. *et al.* TDP-43 mutations causing amyotrophic lateral sclerosis are associated with altered expression of RNA-binding protein hnRNP K and affect the Nrf2 antioxidant pathway. *Hum Mol Genet* 2017;**26**:1732–46. https://doi.org/10.1093/hmg/ddx093
64. Alderson TR, Pritisanac I, Kolaric D. *et al.* Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proc Natl Acad Sci U S A* 2023;**120**:e2304302120. https://doi.org/10.1073/pnas.2304302120
65. Wu Y, Li R, Sun S. *et al.* Improved pathogenicity prediction for rare human missense variants. *Am J Hum Genet* 2021;**108**:1891–906. https://doi.org/10.1016/j.ajhg.2021.08.012
66. Kreiter N, Pal A, Lojewski X. *et al.* Age-dependent neurodegeneration and organelle transport deficiencies in mutant TDP43 patient-derived neurons are independent of TDP43 aggregation. *Neurobiol Dis* 2018;**115**:167–81. https://doi.org/10.1016/j.nbd.2018.03.010
67. Wang W, Wang L, Lu J. *et al.* The inhibition of TDP-43 mitochondrial localization blocks its neuronal toxicity. *Nat Med* 2016;**22**:869–78. https://doi.org/10.1038/nm.4130
68. Lin Y-H, Qiu D-C, Chang W-H. *et al.* The intrinsically disordered N-terminal domain of galectin-3 dynamically mediates multisite self-association of the protein through fuzzy interactions. *J Biol Chem* 2017;**292**:17845–56. https://doi.org/10.1074/jbc.M117.802793
69. Kostylev MA, Tuttle MD, Lee S. *et al.* Liquid and hydrogel phases of PrP-C linked to conformation shifts and triggered by Alzheimer's amyloid-β oligomers. *Mol Cell* 2018;**72**:426–443.e412. https://doi.org/10.1016/j.molcel.2018.10.009
70. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R. (eds), *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*. Long Beach, California, USA: Curran Associates Inc., 2017, 6000–10. https://doi.org/10.1055/s-0044-1791727.
71. Hou S, Hu J, Yu Z. *et al.* Machine learning predictor PSPire screens for phase-separating proteins lacking intrinsically disordered regions. *Nat Commun* 2024;**15**:2147. 10.1038/s41467-024-46445-y
72. Zhang Y, Dong M, Deng J. *et al.* Graph masked self-distillation learning for prediction of mutation impact on protein–protein interactions. *Commun Biol* 2024;**7**:7. https://doi.org/10.1038/s42003-024-07066-9

Briefings in Bioinformatics, 2025, **26(3)**, bbaf171
https://doi.org/10.1093/bib/bbaf171
Problem Solving Protocol