PSTP: Decoding Latent Sequence Grammar through Transfer Learning and Attention for Protein Phase Separation

Mofan Feng[1,2], Xiaoxi Wei[1], 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, Shanghai, China.

2 The Collaborative Innovation Center for Brain Science, and Brain Science and Technology Research Center, Shanghai Jiao Tong University, Shanghai, China.

*** Corresponding author:**

Qing Lu: luqing67@sjtu.edu.cn

Yi Shi: yishi@sjtu.edu.cn

Guang He: heguang@sjtu.edu.cn

Abstract

Phase separation (PS) is essential in various biological processes, necessitating high-accuracy predictive algorithms to study numerous uncharacterized sequences and accelerate experimental validation. However, many recent prediction methods face challenges in generalizability due to their reliance on engineered features, and accurately identifying protein regions involved in PS remains difficult. We propose PSTP, which employs a dual-language model embedding strategy and a lightweight attention model. PSTP achieves identification of 84% of PS regions in PhaSePro and demonstrates ~50% improvement in correlation coefficient over existing models. It shows robust performance across different types of PS proteins, and shows the potential for predicting artificial proteins. By analyzing 160,000 variants, PSTP characterizes the link between the incidence of pathogenic variants and residue-level PS propensities. PSTP's predictive power and broad applicability make it a valuable tool for accelerating our understanding of biomolecular condensates, and their mechanisms underlying disease development.

Introduction

Biomolecular condensates formed through phase separation (PS) are found throughout eukaryotic cells[1-3]. Their formation enables numerous functions across multiple scales[4]. For example, at the molecular level, they regulate the rates of biochemical reactions such as photosynthesis[5,6] and mRNA degradation[7], and participate in immune signaling[8-10] and macromolecule folding[11-13]. At the mesoscale, condensates organize protein densities at neuronal synapses[14,15] and the nucleolus[16,17], and coordinate DNA damage response[18]. At the cellular level, condensates mediate subcellular localization[19-22], and buffer stochastic cellular noises[23,24]. Dysregulation of PS is linked to numerous diseases, such as cancer and neurodegeneration[25,26].

PS is a complex and not fully understood process[27,28], driven by multivalent interactions involving various physicochemical forces[29-33]. Both structured regions and intrinsically disordered regions (IDRs) contribute to driving PS[1,2]. Recent studies on IDR ensembles have revealed the strong relationship between the conformational properties of IDRs, specifically their chain compaction, and their ability to drive PS[34-36]. With the vast amount of uncharacterized sequence data, developing reliable methods for predicting PS and identifying the sequences driving this process is 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[29,37-41]. More advanced, integrated feature-based models[42-46] have been introduced, combining multiple sequence features with novel attributes such as immunofluorescence imaging, post-translational modifications (PTMs), protein-protein interactions (PPIs), and functional annotations. Recently, models have emerged that predict co-phase separation[47] and differentiating PS proteins from those forming amyloids[48,49].

Nevertheless, many recent high-performing models rely heavily on engineered features from various sources, such as protein annotations and images, which limits their generalizability to proteins, including variants and isoforms, that lack such detailed annotations. Moreover, current methods that provide residue-level PS scores often focus on one or a few specific 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 pressing need for more flexible and broadly applicable models.

To address these limitations, we introduce PSTP (**P**hase **S**eparation **P**rediction via **T**ransfer learning) (Fig. 1). PSTP utilizes a dual-language model embedding strategy (Fig. 1A) to process protein sequences. Specifically, we employ a large language model[50] to capture latent information on evolutionary conservation and functional characteristics; and a modified language model[34] to encode the sequence grammar that underpins the physical and dynamic conformational properties of IDRs, which are generally less conserved. This sequence-based feature embedding shows robust performance in predicting both self-assembly PS proteins and those whose PS behaviors are regulated by partner components, such as other proteins or nucleic acids (PS-Part proteins).

Importantly, recognizing that PS is often driven by specific regions rather than the entire protein[51-56], we developed a lightweight attention-based model PSTP-Scan, to extract local information from sequence embeddings to focus on these key regions (Fig. 1C). Using 143 experimentally validated PS regions in PhaSePro[57], we systematically evaluated existing residue-level PS predictors and found that PSTP-Scan's attention weights significantly improve the identification of PS-driving regions. Impressively, it identified 120 out of 143 PS regions, achieving approximately 50% higher correlation coefficients compared to the best-performing existing residue-level models.

In addition, we demonstrate the applicability and generalizability of our predictor 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 characterize the relationship between pathogenicity and protein PS using 160,000+ variants from ClinVar[58,59]. We believe PSTP will facilitate deeper understanding of cellular processes and aiding 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 (https://github.com/Morvan98/PSTP). A user-friendly webserver has also been developed that enable

86 prediction for easy prediction of sequences of interest.

87 Results

88 Phase separation proteins and IDR conformational properties

89 IDRs play important roles in driving protein phase separation[1,2] (PS), and recent findings show that
90 their ability to drive PS is closely tied to their degree of chain compaction[34,35]. Specifically, compact
91 IDRs favor PS[35,60], while expanded IDRs disfavor PS[36,61,62]. Several sequence features, such as charge
92 distribution[34,63,64], hydrophobic[32,65] patterning and aromatic residues content[32,34,35], significantly
93 influence the compaction of IDRs.

94 We first investigate different types of phase-separating proteins (PS proteins) and their IDRs'
95 conformational properties. We predicted IDRs for various types of PS proteins and assessed their chain
96 compaction. We predicted the Flory scaling exponent (ν)[35], the conformational entropy per residue
97 (S_{conf}/N)[35], and Aspericity[34] (Fig. 2 A and B), using models trained on IDR molecular dynamics
98 simulations data. Higher values of these metrics indicate more expanded conformations, while lower
99 values represent more compact conformations. As expected, IDRs from PS proteins exhibited more
100 compact conformations compared to IDRs from background proteome (Fig. 2 A and B). Interestingly,
101 the IDRs of proteins that self-assemble to form condensate (PS-Self proteins) are more compact than
102 those of partner-dependent PS Proteins (PS-Part proteins), whose PS is regulated by partner
103 components. This increased compaction correlates with the high PS propensity observed in PS-Self
104 proteins. Besides IDRs, PS can also be driven by structured domains[2,66-68], and some proteins undergo
105 PS in the absence of IDRs[69]. We collected PS proteins lacking IDRs from the PSPire[69] training dataset
106 and only identified 28 IDRs from these proteins. These IDRs are significantly more expanded
107 compared to other PS proteins, and even more so than the background proteome (Fig. 2 A and B). This
108 expanded nature may facilitate multivalent interactions between structured regions that drive PS.

109 Protein embeddings understand IDR conformational property

110 In the previous section, we demonstrated the distinct conformational properties of IDRs across
111 different types of PS proteins. Language model-based embeddings offering greater flexibility and
112 generalization to encode protein sequences. Given the relationship between IDR chain compaction and
113 PS, we explored whether language models can effectively capture the conformational properties of
114 IDRs.

We collected the IDRome[35] dataset, which contains 28,058 IDRs along with the conformation ensembles generated using CALVADOS, a Molecular dynamics simulations model that shows highly correlation with experimental results. We then assessed whether any language model-based embedding metrics could capture meaningful latent information related to sequence chain compaction by applying these metrics to the IDRs. We test this through visualizing the embedding representations for these sequences.

We first selected and colored extremely compact IDRs ($\nu < 0.45$, n = 599) and expanded IDRs ($\nu > 0.60$, n = 469) (Fig. 2C). The representations generated by the large language model ESM-2[50] were able to distinguish between compact and expanded IDRs (Fig. 2D). Specifically, visualization of the embedding space clearly shows that sequences are clustered into distinct groups by both the smallest version of ESM-2 (8 million parameters) and the larger 650 million parameter version. More pronounced distinctions were observed in the hidden output layer of ALBATROSS[34], a deep neural network trained on molecular dynamics simulation data (Fig. 2D, 'ALBATROSS'). We next evaluated the word2vec skip grams metric[70], which was adopted in several PS prediction metrics such as PSPredictor[71], DeePhase[43] and PSPHunter[44]. In contrast, we did not observe significant differentiation between compact and expanded IDRs with this metric (Fig. 2D, 'Word2vec').

We then divided the IDRs based on conformational entropy per residue (S_{conf}/N). Using this metric, which is strongly correlated with ν, we classified IDRs as compact ($S_{conf}/N < 10$, n = 7,019) or expanded ($S_{conf}/N > 10.25$, n = 4,412) (Fig. 2E). The results were consistent with those observed using ν (Fig. 2F). These findings suggest that embeddings generated from both molecular dynamics-based models and large language models provide meaningful information about IDRs, particularly with regard to chain conformational properties.

Leverage transfer learning to predicting phase separation

Leveraging the conformational properties of IDRs and the underlying information decoded by protein large language models, we purpose a residue-level protein feature embedding matrix (Fig. 1A). Specifically, for a given protein, to encode the underlying information of IDRs' chain conformation, we extract the hidden layer outputs for three sub-ALBATROSS models that predict radius of gyration (Rg), end-to-end distance (Re), and asphericity, respectively. These outputs, each of dimension sequence-length×110, were concatenated to form a sequence-length×330-dimension matrix (IDR

144 conformation feature). In parallel, we incorporated the residue-level layer representations from the
145 ESM-2 model (8 million parameter version), resulting in a sequence-length×320-dimensional matrix
146 (large language model feature). To reduce memory and computational demands associated with long-
147 distance attention calculations, we applied a sliding window approach for embedding sequences longer
148 than 256 residues (methods). Combining the outputs from ALBATROSS and ESM-2, we generated a
149 sequence-length×650-dimension feature matrix, referred to as PSTP (Phase Separation Prediction via
150 Transfer learning) embeddings.
151 To assess the ability of PSTP embeddings to encode essential information for PS proteins, we first
152 averaged the embedding matrix along the residue dimension (Fig. 1B). Using curated data from
153 PhaSePred[45], we trained and evaluated Logistic Regression (LR) and Random Forest (RF) models with
154 default parameters to evaluate the efficiency of our embeddings (Fig. 1B).
155 Cross-validation results indicated that using only the IDR conformation feature (330-dimension vector)
156 alone yielded an average AUC at 0.81 for predicting PS-Self proteins and 0.72 for PS-Part proteins
157 (Fig. S1B). The integrated embeddings achieved median AUCs of 0.85 and 0.76 with PSTP-LR for
158 PS-Self and PS-Part, respectively, (Fig. 3A) demonstrating predictive ability across PS proteins from
159 different species. Notably, excluding IDR conformation features resulted in a significant performance
160 drop across both cross-validation and the test set, underscoring the critical role of IDR conformational
161 features in predicting PS (Fig. 3A).
162 We next conducted an independent validation of PSTP-LR using the same training and independent
163 validation set split as PhaSePred, one of the state-of-the-art PS predictors. PhaSePred is a meta-
164 predictor that integrates several sequence-based algorithms, and multi-dimensional features such as
165 immunofluorescence imaging and PTMs. Without performing additional feature selection or
166 hyperparameter tuning, our LR model achieved comparable predictive accuracy to PhaSePred (SaPS-
167 10) for PS-Self proteins, while showing an improvement in AUC (a 3% increase to PdPS-10) for
168 predicting human PS-Part proteins (Fig. 3B). PSTP-LR model also outperformed PhaSePred on the
169 hPS-test set, whether trained with PS-Self or PS-Part data (Fig. 3B, right panel).
170 The sliding window embedding approach for long sequences with a length threshold of 256 residues
171 did not lead to a significant reduction in performance. Increasing the threshold to 1024 residues (the

maximum sequence length during ESM-2 training) did not markedly enhance the model's accuracy (Fig. S1C), indicating that our original configuration is robust for handling longer sequences.

We further assessed the performance of PSTP in predicting membrane-less organelles (MLOs) participants localized to various intracellular locations and using datasets of known MLO-associated proteins. These participants are likely to phase separate or co-phase separate with other proteins to form MLOs. PSTP trained to predict PS-Part proteins demonstrated strong predictive performance in identifying these proteins (Fig. S1E).

The high predictive accuracy underscores the significance of conformational features in PS and demonstrates the capability of pretrained language models the latent grammar essential for PS. These results highlight that solely utilizing protein sequences can effectively capture the underlying characteristics of PS proteins.

Model architecture of PSTP-Scan

In the previous section, the residue dimension-averaged PSTP embedding vectors showed robust performance. However, since PS is often driven by specific regions within a protein, averaging the feature matrix may overlook critical information about these PS-prone regions. Drawing inspiration from image-based spatial attention neural networks, which generate attention scores for each pixel using local contextual information[72-75], we developed a model architecture named PSTP-Scan. This architecture provides residue-level PS profiles while simultaneously predicting the overall PS propensity of the protein.

For each sequence, the PSTP-Scan model takes the sequence-length×650-dimension PSTP feature matrix as input. Attention weight for each residue position is generated using a convolution-like operation. This involves processing a local window of features around the position through an average pooling layer, followed by a trainable two-layer multilayer perceptron (MLP) kernel (Fig. 1C, Method). The sequence-length unnormalized vector serves as the residue-level attention score.

The maximum attention weight directly serves as the predicted PS propensity for the protein, while the full vector of attention weights provides residue-level PS predictions. Given that the lengths of PS-driving regions vary from tens to hundreds of residues (Fig. 4A), during training, the model employs multiple window sizes for the MLP kernel to capture information at different scales (Methods, Fig. S2C), enhancing robustness. Although the predicted propensity is primarily influenced by the sequence

fragment with the highest attention score, the PSTP-Scan model achieves performance comparable to PSTP-LR in predicting both PS-Self and PS-Part proteins (Fig. S1D). This demonstrates that focusing on the most influential regions does not compromise overall predictive accuracy, thereby validating the reliability of this architecture.

PSTP-Scan can provide reliable PS prediction

To further evaluate PSTP's robustness, we utilize a PS dataset curated by a more recent study[44]. The datasets consists of three PS sub-datasets, comprising a total of 892 PS proteins curated from various sources, including LLPSDB[76], PhaSepDB[77], DrLLPS[78] and PhaSePro[57], along with a background sub-dataset of 8897 single-domain human proteins.

After applying CD-Hit[79] to remove redundant sequences with a 0.4 threshold, we obtained a human-specific PS protein sub-dataset containing 136 proteins and a background set of 6259 human proteins (Methods), which served as the independent test set. We also obtained 375 PS proteins and 17,288 background proteins from other species, which were utilized as the training dataset (Methods).

PSTP-LR, PSTP-RF and PSTP-Scan all demonstrated strong performance in distinguishing human PS proteins from the human background proteins (Fig. 3C and E, Fig. S1G). We then conducted a head-to-head comparison of PSTP-Scan with current representative PS methods, excluding proteins used in the training datasets of those methods from the evaluation. The methods compared included PhaSePred[45], which utilizes multi-dimensional integrated features; DeePhase[43], which combines a word2vec language model with engineered features; PLAAC[37,38], which predicts prion-like propensities; PSAP[42], which utilizes engineered biochemical and biophysical features; and ParSe[41], a structure-based PS predictor. For comparison with PSPHunter[44], we constructed features following its sequence encoding method. Additionally, we evaluated other sequence-based predictors, including catGRANULE[39], FuzDrop[40] and PScore[29].

PSTP-Scan achieve a ~0.9 AUC in each head-to-head query, better than these representative methods (Fig. 3 D and F). We also evaluated individual PSTP-Scan blocks with different window sizes and observed consistently robust results (Fig. S1G), showing the robustness of PSTP-Scan, and insensitivity to window size variation.

These findings demonstrate that by leveraging large language models and deep learning models that encode IDR conformation information, PS-related features can be efficiently embedded. This approach

230 provides a reliable and simplified method for PS prediction using only protein sequence as input.
231 The training and testing dataset together to form a 'Mix' dataset, that we used to train PSTP (Mix)
232 model. We also trained a PSTP (SaPS) model using the PS-Self dataset as well as a PSTP (PdPS) model
233 using the PS-Part dataset.

234 PSTP-Scan improve identification of PS driving region

235 To evaluate PSTP-Scan's ability to predict PS-driving regions at the residue level, we assessed the
236 residue-level PS scores generated by the attention layer output from PSTP-Scan. We utilize the
237 experimentally validated PS regions documented in the PhaSePro[57] database, derived from proteins
238 across different species. Although PSTP-Scan's training did not involve any residue-level information
239 about PS regions, we trained each sub-model of PSTP-Scan (SaPS, PdPS and Mix) after excluding
240 proteins already included in the PhaSePro database (Fig. S3A) to avoid potential self-validation.
241 We found that, out of 143 experimentally validated PS regions documented in PhaSePro, 120
242 overlapped with the PSTP-Scan's PS regions, which is a notable improvement over the 109 overlaps
243 achieved by FuzDrop, a model directly trained on PhaSePro PS regions (Fig. 4B). In terms of amino
244 acid-level overlap between predicted PS and experimentally validated regions, PSTP-Scan recovered
245 28,378 residues, compared to 25,014 residues recovered by FuzDrop (Fig. 5E). Furthermore, PSTP-
246 Scan outperformed PSPHunter, a method recently reported to predict key regions for PS (Fig. S4 C
247 and D).
248 We next systematically compared PSTP-Scan with other residue-level PS predictors in discerning PS
249 regions. T-test statics showed that, compared to other predictors, PSTP-Scan has the strongest
250 discriminative power in distinguishing residues located in PS driving regions from those in non-PS
251 regions within PS proteins. Additionally, we computed Spearman correlation between the residue-level
252 scores of each predictor and the experimentally validated PS regions in PhaSePro (where a position is
253 assigned 1 if it belongs to a PS region and 0 if not) (Fig. 4D). We calculated both the overall Spearman
254 correlation and protein-specific correlations for each protein. PSTP-Scan achieved significantly higher
255 correlations compared to other predictors, with a 9% improvement observed overall. In terms of
256 protein-specific correlation, PSTP-Scan also demonstrated much higher performance over other
257 methods (Fig. 4E and Fig. S4B).
258 In Fig. 4 F and G, and Fig. S4, we present examples of the PSTP-Scan prediction results. The Scan-

259 SaPS model, trained on PS-Self proteins, identify regions that drive PS, such as the low-complexity
260 domain (LCD) repeats in cortactin[53], and the N terminus IDR of Ddx4, which contains containing
261 repeating net charge block and FG, RG blocks[80]. The Scan-PdPS model, trained on PS-Part proteins,
262 outperforms Scan-SaPS in predicting PS-driving regions that rely on interactions with partners. This
263 includes the phosphotyrosine motif of the Nephrin and the proline-rich domain of N-WASP (Fig. 4G)
264 which form multivalent interactions with the SH2 and SH3 domains of NCK, respectively[55]. For
265 proteins that contain both RNA-recognition motifs (RRMs) and LCDs, such as PAB1[52] and Nab3[54],
266 Scan-PdPS identifies the RRMs, while Scan-SaPS focuses on the LCDs (Fig. 4G).
267 We extracted records by PS-Self and PS-Part proteins, and found that PSTP-Scan achieves an overall
268 0.40 correlation for PS-Self proteins (58 records, Fig. S4 E and H). In contrast, all models showed
269 significantly reduced performance for PS-Part proteins, though PSTP-Scan still performed the best (50
270 records, Fig. S4 F and I). The lower performance for PS-Part proteins may be due to the fact that some
271 protein-protein interaction sites are also denoted as PS-regions in the PhaSePro database, the prediction
272 of which requires prior knowledge of specific protein interactions. For instance, PSTP-Scan did not
273 predict the tri-RG motif in Buc that interacts with the regulator Tdrd6a[81], but successfully predicted
274 the prion-like domain responsible for Buc's self-assembly[51] (Fig. S4B), which is absent in the
275 PhaSePro database. Similarly, while PSTP-Scan predicted the tandem repeats for the scaffold protein
276 EPYC1, neither PSTP-Scan nor FuzDrop predicted the alpha-helices of Rubisco that bind to EPYC1[56]
277 (Fig. S4C). Incorporating models that focus on predicting protein-protein interactions may provide
278 better insights into these types of PS mechanisms.
279 The limited performance of current methods may also stem from the highly flexible definitions of PS-
280 driving segments or key regions. For example, PSTP-Scan achieves a 0.43 correlation with PS regions
281 identified by NMR experiments (Fig. S3E), significantly higher than the overall correlation.
282 Nevertheless, given that PSTP-Scan predicts diverse PS-driving regions without being specifically
283 trained on distinct PS regions, its ability to directly generate these predictions effectively demonstrates
284 its validity, offering detailed insights into protein PS behavior.

285 **PSTP-Scan provide a broader range of prediction**

286 Studying truncated proteins is essential for understanding protein function and pathogenicity, as NMD-
287 escaping mutations that fail to trigger NMD often result in the expression of truncated proteins[82]. We

evaluate PSTP-Scan as well as other sequence-based predictors on a test dataset of 93 experimentally truncated PS proteins and 242 non-PS proteins (Fig. 5A) by collecting data from a recent study[83]. To avoid information leakage, we specially trained each PSTP models (PSTP-SaPS, PSTP-PdPS and PSTP-Mix) after excluding proteins corresponding to the protein fragment in the test dataset from the training dataset (Fig. S5A). We predicted these proteins using PSTP-Scan as well as PSTP-Scan's MLP kernel (Fig. 5A, Fig. S5 C and D). We observed that compared with other representative sequence-based predictors, PSTP achieves the best performance in terms of both AUC and AUPR, while the MLP kernel from the PSTP-Scan (Mix) achieve the best performance at an AUC of 0.88 (Fig. 5A). This demonstrates PSTP-Scan's robustness and effectiveness in predicting PS in truncated proteins.

In addition, a recent study[84] designed artificial intrinsically disordered proteins (A-IDPs) to investigate the influence of amino-acid sequence patterns on the regulation of PS. This study created 11 artificial polypeptides based on prior knowledges from earlier research[85,86], each consisting of a 200-residue sequence formed by repeating an octapeptide 25 times, which exhibited PS at micromolar concentrations. Most of these A-IDPs underwent PS at room temperature, while some required higher temperatures.

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. All models, except PLAAC, significantly distinguished A-IDPs from these background proteins, highlighting the unique sequence properties of the A-IDPs (Fig. 5C). However, PSAP produced relatively low scores, failed to predict these peptides. Next, we generated control peptides by randomly repeating octapeptides 25 times, using the same amino acid composition as the A-IDPs (Fig. 5C, AA-weighted random). Interestingly, PSTP-Scan predicted lower scores for these control peptides (p-value = 0.0085), while other predictors did not show significant differences. This indicates that PSTP-Scan understands these A-IDPs better, captures not only amino acid composition but also the specific order of residues.

Both results indicate PSTP-Scan's sensitivity to sequence information and its predictive power across a broad range of proteins.

Improved prediction of PSP that lack IDRs

We also developed a predictor focused on PS scaffolds and regulators that lack IDRs to address the

317 high disorder propensity (Fig. S6A) of current PS datasets. Utilizing a curated dataset from PSPire[69],
318 which includes 128 PS proteins and 10,251 background proteins, we trained and evaluated a PSTP-
319 LR and a PSTP-RF model. Compared to PSPire, which utilizes structural information from
320 AlphaFold2[87], the PSTP models demonstrated greater robustness and reliability, achieving a ~50%
321 increase in AUPR (Fig. S6B and C).

322 **Incidence of pathogenic variants**

323 Evolutionary conservation is currently a major factor in interpreting disease variants [88-92], and using
324 this feature alone can predict the majority of known pathogenic variants[91]. IDRs tend to be less
325 evolutionary conserved and lack fixed structures in comparison to structured and functionally essential
326 regions[1]. As a result, studies on the pathogenicity of IDR variants have progressed slowly, leaving
327 numerous VUSs (~ 530,000 ClinVar VUSs) unsolved located in low-conservation IDRs. Given the
328 established link between IDRs and PS, we aim to investigate the potential between PS-driving
329 propensity and the incidence of pathogenic variants in these low-evolutionary-conserved IDRs.

330 We analyzed PSTP-Scan scores for ClinVar variants located in low-conservation IDRs, defined by
331 AlphaFold2 pLDDT score (pLDDT < 50) (Fig. 5F). The pLDDT scores, which reflects AlphaFold2
332 prediction confidence and is highly correlated with evolutionary conservation, is considered a state-
333 of-the-art predictor for IDRs[35,93]. We extracted 4756 pathogenic and 37,291 benign variants, and
334 observed that, compared to benign variants, pathogenic variants tend to locate in position with higher
335 residue-level PS propensity (Fig. 5G). This significant difference, with Mann-Whitney p-values below
336 1×10^{-10}, was detected by all sub-models of PSTP-Scan. Additionally, we observed a similar pattern in
337 variants located in disordered regions as defined by IUPred3 (IUPred score > 0.5) (Fig. S5E). However,
338 this distinction was weaker, possibly because IUPred-predicted IDRs also include evolutionary
339 conserved IDRs[93]. Next, we evaluated this correlation based on allele frequency (AF), as variants with
340 very low AF are likely to be pathogenic, while highly frequent variants are less likely to be[94,95]. We
341 collected variants from gnomAD V4.1.0 and found that, within low conservation regions, variants with
342 low AF (AF < 1×10^{-5}, 3,841,613 variants) tend to locate in position with higher residue-level PS score
343 region than variants with higher AF (AF ≥ 0.001, 25,673) (Fig. S5 D and F). In contrast, this correlation
344 between PS and disease was not observed for variants located in structured regions (pLDDT > 70) (Fig.
345 S5C), where pathogenicity is influenced by factors including protein evolution and function. Our

346 findings suggest that alterations in PS driving region, are on average, more likely to lead to be
347 pathogenic than alternation in non-PS region, even these PS driving region are evolutionary un-
348 conserved.
349 In addition to missense variants, we also extracted 17,304 protein-altering variants, including nonsense,
350 in-frame deletions, and frameshift variants (Fig. 5H). We computed the change in PSTP-Scan scores
351 before and after mutation, and compared the distribution of these changes between pathogenic and
352 benign variants using a two-sample KS test. We observed that pathogenic variants are more likely to
353 alter PS propensity than benign variants, as indicated by the significant p-values of the KS test (Fig.
354 5I). This result shows that for variants that induce extensive protein sequence changes, disrupting PS
355 may be a contributing factor to their pathogenicity.

356 **Discussion**

357 Language models have made significant progress in recent years. ESM-2 is a sequence-input-only
358 encoder that provides a much faster and more flexible encoding compared to AlphaFold2[87], which
359 relies on computationally intensive multiple sequence alignments (MSAs). Although language-model-
360 based embeddings may be less interpretable than engineered features, they capture comprehensive
361 sequence characteristics more objectively, expanding the range of proteins that can be predicted. For
362 broader applicability, we used the smallest version of the ESM-2 model, as the existing performance
363 already demonstrates the effectiveness of our approach.
364 Although language models can extract rich evolutionary information from sequences, intrinsically
365 disordered regions (IDRs) are relatively less conserved. Recent studies suggest that while IDRs may
366 undergo significant sequence variation throughout evolution, they can maintain stable conformational
367 properties[34,35]. Therefore, we consider that the conformational information of IDRs, which is closely
368 linked to PS propensity, might be better captured by models specially trained using molecular
369 dynamics simulation data. To address this, we designed an embedding metrics that utilize a pretrained
370 LSTM to encode the dynamic properties of IDRs. While both types of embeddings alone can provide
371 PS predictions using traditional ML models like LR or RF, their combination performs significantly
372 better, confirming the effectiveness of our dual-language model embedding.
373 Unlike existing predictors such as FuzDrop or PScore, which employ supervised learning models

trained at each residue position, our PSTP-Scan training process did not introduce any residue-level information beyond allowing the model to learn autonomously. Due to the highly flexible definitions of PS driving or regulating regions, the performance of current methods is relatively limited. However, PSTP-Scan successfully predicts many PS driving segments, providing important insights into numerous uncharacterized sequences. We believe that PSTP-Scan's prediction scores are more robust than those of existing residue-level prediction algorithms.

Given that the training data for phase-separating proteins is relatively small compared to other deep learning fields, we avoided using overly complex deep models like transformer-based attention layers with query-key-value mechanisms. Instead, we constructed an attention calculation method where only the weights of a multilayer perceptron (MLP) network are trainable. Furthermore, we did not utilize the attention weights from the ESM-2 model itself because its attention scores may focus more on structural and functional regions, potentially introducing additional noise. Our results demonstrate that predicting PS by using only the maximum of local scores as output is effective, aligning with the nature of PS-driving regions responsible for multivalent interactions with other regions.

PSTP demonstrates excellent performance among various phase separation prediction tools and provides the most accurate predictions of phase-separating regions. We believe PSTP will support a deeper understanding of cellular processes, aiding in disease research and therapeutic development.

Methods

IDR conformational property

Visualizing space distribution of IDRs of different properties

The SVR model developed by Tesei et al.[35] (https://github.com/KULL-Centre/_2023_Tesei_IDRome) is used to predict v and S_{conf}/N of IDRs. ALBATROSS developed by Lotthammer et al.[34] (https://github.com/idptools/sparrow) is used to predict the asphericity of IDRs.

We collected curated datasets from several sources: self-assembly PS proteins (SaPS) (201 proteins), partner-dependent PS proteins (PdPS) (327 proteins) as well as non-redundant multi-species background proteins (60,220 NoPS proteins) from PhaSePred[45]. PS proteins that lack IDRs were collected from PSPire[69] (128 proteins). In addition, we grouped PS proteins from several literature[44,45,69] to obtain a 'Mix PSPs' group (811 proteins after removing redundancy). For each group,

402 we adopted IUPred3[96] to predict residue-level disorder scores for each protein, following a process
403 similar to Tesei et al.[35] to filter IDRs. Specifically, residues with IUPred3 scores ≥ 0.5 were first labeled
404 as disordered and those < 0.5 as folded. Regions shorter than ten residues were classified as gaps.
405 These gaps were labeled as disordered if they were either flanked by disordered regions or terminal
406 and adjacent to disordered regions. IDRs shorter than 30 residues or exceeding 1,500 residues were
407 discarded.
408 The predicted conformational properties were compared between different IDR groups using the two-
409 sided Mann–Whitney U test.

410 **Visualizing space distribution of IDRs of different properties**

411 The set of 28,058 IDRs used for visualization, along with their conformational properties generated by
412 CALVADOS[35], were obtained from (https://sid.erda.dk/cgi-sid/ls.py?share_id=AVZAJvJnCO). Each
413 IDR was vectorized using two models from ESM-2[50]: esm2_t6_8M_UR50D and
414 esm2_t33_650M_UR50D, which output metrics of sequence-length×320, sequence-length×1280
415 dimension respectively. Each matrix was then averaged into vectors of lengths 320 and 1280 for each
416 IDR.
417 Additionally, a trained word2vec embedding method[97] based on the skip-gram model was tested. The
418 outputs for all contiguous 3-mer sequences of each IDR were summed to form a 100-length vector,
419 which was also averaged to mitigate the impact of sequence length.
420 To evaluate the embeddings generated by ALBATROSS, we concatenated the hidden layer outputs
421 from three sub-models for each input IDR into a sequence-length×330-dimensional matrix (see the
422 'Feature Engineering' section for more details) and averaged it to obtain a 330-dimensional vector.
423 Dimensionality reduction was performed using the UMAP algorithm, implemented through the Python
424 'umap' package.

425 **PS proteins dataset**

426 **Self-assembly PS and partner-dependent PS datasets**

427 We utilize the curated datasets used for training and evaluating PhaSePred[45], which were obtained
428 from the supplementary materials provided with its published literature. The acquired data included
429 training and cross-validation (CV) datasets: SaPS (128 proteins), hSaPS (59 proteins), PdPS (214
430 proteins), hPdPS (96 proteins), NoPS (48,158 proteins), 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 (2,200 proteins), PS-test (53 proteins), hPS-test (23 proteins). A total of 201 SaPS proteins, 327 PdPS proteins, and 60,220 NoPS proteins were collected.

Additional dataset for independent validation

To build an additional validation set, we utilized PS proteins and background proteins used to train and validate PSPHunter. This dataset contains a total of 604 PS proteins form PS either independently or in conjunction with co-scaffolds (annotated as 'Mix' dataset). 167 human PS proteins, 437 non-human PS proteins, and 8897 non-redundant single-domain human background proteins were included in the dataset.

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 437 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 not filtered). We next filtered both background proteins to ensure they were not longer than the longest human PS proteins (5,537 length), we finally got 6259 human background proteins and 17,268 non-human background proteins.

Feature engineering

Feature engineering with protein language model:

We adopted two embedding matrices for each protein. The first metric is the embedding computed by ESM-2[50], 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 takes sequence as input and the sixth representation layer is output to return a 320-dimension vector for each position within each sequence. For each sequence, to avoid memory leakage as well as reduce the time required for long-distance attention computing, we apply a sliding window process similar to AlphaFold2[87] and AlphaMissense[95] when dealing with long sequences. Specifically, we applied a 256-length sliding

window to scan through the sequence with a step size of 50 (Given a 520-length sequence for example, we generated sliding window as [0, 256], [50, 306], … [250,506], [264, 520]). For each position, the embedding from the sliding window where the position is deepest was applied.

The speed of ESM2-8M is much faster compared to other ESM-2 versions on both CPU and GPU. Take an AMD Ryzen 7 3700X 8-Core Processor 3.7 G Hz CPU combined with 16 G memory as an example, for each 256-length-sequence (single sequence each batch), an average of 3.47 seconds is required for esm2_t6_8M_UR50D, while only an average of 0.11 seconds is required for ESM2-8M. This makes it easier and faster for users to predict sequences using PSTP.

Feature engineering with ALBATROSS:

The second embedding metric we adopted is the ALBATROSS program[34] (https://github.com/idptools/sparrow). ALBATROSS is trained with molecular dynamic simulation data of IDRs, and contains five pretrained LSTM-BRNN models to predict conformation properties of IDRs including radius of gyration (R_g), end-to-end distance, asphericity, the scaling exponent and prefactor.

Considering each model has learned the underlying information of IDRs conformation, we adopted the 'asphericity' model, the 'radius_of_gyration_scaled' model, the 'end_to_end_distance_scaled' model as they share the same model structure.

For each model of the three models, a protein sequence is inputted, and the output of each timestep was adopted as the ALBATROSS embedding for each protein position, which returns a (sequence_length, 110) dimension vector for each input sequence. These values range from -1 to 1. The vectors are then concatenated to a (sequence_length, 330) dimension for each input sequence.

For predicting PS proteins that lack IDRs and to compare with PSPire, we used only large language model features for the feature embeddings without IDR conformational features.

Machine learning models

We applied a 'LogisticRegression' model using the 'linear_model' module and a 'RandomForestClassifier' model using the 'ensemble' module from the python Sci-kit learn package. Both models were set with parameter class_weight = 'balanced' and other parameters were kept on default, and no normalization was applied to the dataset, neither was hyper parameter selection procedure. In each test scenario, to evaluate the LR and RF models during both 5-fold cross-validation

(CV) and independent validation, the PS protein dataset was supplemented with randomly selected NoPS proteins at twice the size of the PS set. To ensure robustness, this sampling and validation process was repeated at least 50 times to calculate the average performance.

PSTP-Scan neural network

Model architecture

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

Within each PSTP-Scan block: 1. The first layer is an average pool layer with 'same' padding, that for each protein position, a local window of feature matrix is averaged along the residue dimension to create a 650-dimensional vector. 2. To mitigate the effects of 'same' padding, a scaling step adjusts the values at the sequence ends to ensure they reflect the actual averages of the sequence elements rather than padded zeros. So, the representation obtained for each position (R_i) can be written as below, were x_j represents the embedding vector for position j and L represents the sequence-length:

$$R_i = \frac{1}{n_i} \sum_{j=max(0,i-k)}^{min(i+k,L-1)} x_j \tag{1}$$

3. Next, a shared, trainable multilayer perceptron (MLP) kernel processes the averaged vectors. The MLP architecture consists of an initial layer with 20 neurons, followed by a LeakyReLU activation, a hidden layer with 5 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 forms the sequence-length attention vector. 4. The last is a max pooling layer that selects the maximum value of the attention vector, providing the sequence-level sequence propensity prediction. The attention computation using the representation at each position can be written as follows.

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

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

Given that the lengths of known phase-separating (PS) regions range from tens to hundreds of residues (Fig. 4A), three parallel blocks with identical architectures but different average pooling window sizes are used to capture PS information from sequence fragments of varying lengths. The pooling sizes are

517 set to 256 + 1, 128 + 1, and 32 + 1, corresponding to powers of 2, with the extra 1 ensuring symmetric
518 'same' padding on both sides. We demonstrated the robustness of model performance across different
519 window sizes (Fig. S1G) when applying a single PSTP-Scan block. The predicted propensities (single-
520 value sequence-level propensity) from these blocks are averaged to produce the final output, which is
521 optimized through backpropagation to update the MLP weights.

522 **Model training**

523 We initialized the network parameters using 'kaiming_normal_' with default settings. The GFS
524 normalization metric we previously developed[98] was applied to independently penalize the parameters
525 for IDR and ESM embeddings in the first layer of the MLP, with the penalty weight set to 0.03. Binary
526 cross-entropy loss was used. The Adam optimizer was applied with a learning rate of 0.003, without
527 additional hyperparameter tuning. During the training process, for each iteration (50 epochs in total),
528 we held out a random subset of proteins for the size of the positive dataset from the background dataset
529 as the negative dataset. This approach was applied to address sample imbalance.

530 Model performance benchmarks

531 AUC, AUPR were applied during each validation. For the head-to-head comparison with PSPire, we
532 also computed F1-Score, Precision, MCC, Accuracy, FNR, FPR, and Sensitivity by setting the cutoff
533 at 0.4 best on the optimized average F1-Score. For the head-to-head comparison with PhaSePred and
534 other PS predictors, we randomly selected proteins twice the size of the PS dataset from the
535 background proteins and merged them with the positive dataset to compute the performance using
536 AUC and AUPR.

537 Evaluation of PS-Driving Region Identification

538 We collected PS proteins and their corresponding experimentally validated PS regions from the
539 PhaSePro[57] database (https://phasepro.elte.hu/download_full.tsv). To prevent self-validation, we
540 trained each PSTP-Scan model (SaPS, PdPS and Mix) after excluding proteins already included in
541 PhaSePro[57]. We set the average pool size at 33 (the minimum value of the average pool size when
542 training PSTP-Scan) when generating the residue level scores for these proteins. We define the PSTP-
543 Scan PS regions using the prediction values of the three models, that windows where scores were
544 continuously over 0.5 more than 20 AAs by the maximum of three models were considered PS regions.
545 To systematically evaluate PSTP-Scan's residue-level PS scores, we assigned residue-level scores to

each protein based on the highest protein-level PS predicted value among the three sub-models: PSTP-Scan-Mix, PSTP-Scan-SaPS, and PSTP-Scan-PdPS. For example, if PSTP-Scan-PdPS yielded the highest predicted propensity for a protein—indicating it is likely to undergo partner-dependent PS—we used the residue-level scores from the PSTP-Scan-PdPS model for that protein.

We calculated Spearman correlation coefficients between these residue-level scores and the binary annotation of actual PS regions (assigning a value of 1 to positions within PS regions and 0 otherwise) to assess predictive performance.

Other PS models

For each of PScore[29], PLAAC[37], PSAP[42], ParSe[41], and PSPHunter[44], the prediction software was obtained from their respective GitHub repositories, as documented in the corresponding literature. PhaSePred[45] scores for self-assembly PS proteins ('SaPS-10fea_rnk') and partner-dependent PS proteins ('PdPS-10fea_rnk') were downloaded from (http://predict.phasep.pro/download/). Prediction scores for FuzDrop[40] were obtained from (https://fuzdrop.bio.unipd.it/). Prediction functions, as well as feature vectors and model parameters for catGRANULE[39], were obtained from its corresponding supplementary information and methods. During the independent testing (Fig. 3C-F), proteins that were included in the PS samples for each method were excluded from both positive and negative datasets.

PS and Pathogenicity

For the pathogenicity analysis, we downloaded variants from ClinVar (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz) as of April 2024, and from gnomAD V4.1.0 (https://gnomad.broadinstitute.org/).

Final PSTP models

For the final PSTP (PSTP-LR, PSTP-RF and PSTP-Scan) models, we trained three versions, using the PS-Self data, PS-Part data and Mix PS data, respectively. The training and testing datasets for each task were combined to train the final model. For each of the three models, 10 sub-models were trained, and the average output of these 10 models was used as the final result at both the residue and protein levels.

Code availability

An installable python package is available at (https://github.com/Morvan98/PSTP).

575 # References

576 1. Tsang, B., Pritišanac, I., Scherer, S.W., Moses, A.M. & Forman-Kay, J.D. Phase Separation as a Missing Mechanism
577 for Interpretation of Disease Mutations. *Cell* **183**, 1742-1756 (2020).
578 2. Gao, Y., Li, X., Li, P. & Lin, Y. A brief guideline for studies of phase-separated biomolecular condensates. *Nat Chem*
579 *Biol* **18**, 1307-1318 (2022).
580 3. Alberti, S. Phase separation in biology. *Curr Biol* **27**, R1097-1102 (2017).
581 4. Lyon, A.S., Peeples, W.B. & Rosen, M.K. A framework for understanding the functions of biomolecular condensates
582 across scales. *Nat Rev Mol Cell Biol* **22**, 215-235 (2021).
583 5. Oltrogge, L.M. *et al.* Multivalent interactions between CsoS2 and Rubisco mediate α-carboxysome formation. *Nat*
584 *Struct Mol Biol* **27**, 281-287 (2020).
585 6. Wang, H. *et al.* Rubisco condensate formation by CcmM in β-carboxysome biogenesis. *Nature* **566**, 131-135 (2019).
586 7. Sheu-Gruttadauria, J. & MacRae, I.J. Phase Transitions in the Assembly and Function of Human miRISC. *Cell* **173**,
587 946-957.e16 (2018).
588 8. Su, X. *et al.* Phase separation of signaling molecules promotes T cell receptor signal transduction. *Science* **352**, 595-
589 9 (2016).
590 9. Wu, J. *et al.* Cyclic GMP-AMP is an endogenous second messenger in innate immune signaling by cytosolic DNA.
591 *Science* **339**, 826-30 (2013).
592 10. Xiao, Q., McAtee, C.K. & Su, X. Phase separation in immune signalling. *Nat Rev Immunol* **22**, 188-199 (2022).
593 11. Frottin, F. *et al.* The nucleolus functions as a phase-separated protein quality control compartment. *Science* **365**,
594 342-347 (2019).
595 12. Yang, P. *et al.* G3BP1 Is a Tunable Switch that Triggers Phase Separation to Assemble Stress Granules. *Cell* **181**,
596 325-345.e28 (2020).
597 13. Nott, T.J., Craggs, T.D. & Baldwin, A.J. Membraneless organelles can melt nucleic acid duplexes and act as
598 biomolecular filters. *Nat Chem* **8**, 569-75 (2016).
599 14. Banani, S.F., Lee, H.O., Hyman, A.A. & Rosen, M.K. Biomolecular condensates: organizers of cellular biochemistry.
600 *Nat Rev Mol Cell Biol* **18**, 285-298 (2017).
601 15. Zeng, M. *et al.* Reconstituted Postsynaptic Density as a Molecular Platform for Understanding Synapse Formation
602 and Plasticity. *Cell* **174**, 1172-1187.e16 (2018).
603 16. Yao, R.W. *et al.* Nascent Pre-rRNA Sorting via Phase Separation Drives the Assembly of Dense Fibrillar
604 Components in the Human Nucleolus. *Mol Cell* **76**, 767-783.e11 (2019).
605 17. Falahati, H. & Wieschaus, E. Independent active and thermodynamic processes govern the nucleolus assembly in
606 vivo. *Proc Natl Acad Sci U S A* **114**, 1335-1340 (2017).
607 18. Singatulina, A.S. *et al.* PARP-1 Activation Directs FUS to DNA Damage Sites to Form PARG-Reversible
608 Compartments Enriched in Damaged DNA. *Cell Rep* **27**, 1809-1821.e5 (2019).
609 19. Maday, S., Twelvetrees, A.E., Moughamian, A.J. & Holzbaur, E.L. Axonal transport: cargo-specific mechanisms of
610 motility and regulation. *Neuron* **84**, 292-309 (2014).
611 20. Liao, Y.C. *et al.* RNA Granules Hitchhike on Lysosomes for Long-Distance Transport, Using Annexin A11 as a
612 Molecular Tether. *Cell* **179**, 147-164.e20 (2019).
613 21. Liu, X. *et al.* Mitotic Implantation of the Transcription Factor Prospero via Phase Separation Drives Terminal
614 Neuronal Differentiation. *Dev Cell* **52**, 277-293.e8 (2020).
615 22. Strom, A.R. *et al.* Phase separation drives heterochromatin domain formation. *Nature* **547**, 241-245 (2017).

23. Riback, J.A. *et al.* Composition-dependent thermodynamics of intracellular phase separation. *Nature* **581**, 209-214 (2020).
24. Klosin, A. *et al.* Phase separation provides a mechanism to reduce noise in cells. *Science* **367**, 464-468 (2020).
25. Alberti, S. & Dormann, D. Liquid-Liquid Phase Separation in Disease. *Annu Rev Genet* **53**, 171-194 (2019).
26. Nedelsky, N.B. & Taylor, J.P. Bridging biophysics and neurology: aberrant phase transitions in neurodegenerative disease. *Nat Rev Neurol* **15**, 272-286 (2019).
27. Pappu, R.V., Cohen, S.R., Dar, F., Farag, M. & Kar, M. Phase Transitions of Associative Biomacromolecules. *Chem Rev* **123**, 8945-8987 (2023).
28. Mittag, T. & Pappu, R.V. A conceptual framework for understanding phase separation and addressing open questions and challenges. *Mol Cell* **82**, 2201-2214 (2022).
29. Vernon, R.M. *et al.* Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife* **7**, e31486 (2018).
30. Wang, J. *et al.* A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* **174**, 688-699.e16 (2018).
31. Dignon, G.L., Best, R.B. & Mittal, J. Biomolecular Phase Separation: From Molecular Driving Forces to Macroscopic Properties. *Annu Rev Phys Chem* **71**, 53-75 (2020).
32. Martin, E.W. *et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367**, 694-699 (2020).
33. Brangwynne, Clifford P., Tompa, P. & Pappu, Rohit V. Polymer physics of intracellular phase transitions. *Nature Physics* **11**, 899-904 (2015).
34. Lotthammer, J.M., Ginell, G.M., Griffith, D., Emenecker, R.J. & Holehouse, A.S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat Methods* **21**, 465-476 (2024).
35. Tesei, G. *et al.* Conformational ensembles of the human intrinsically disordered proteome. *Nature* **626**, 897-904 (2024).
36. Tesei, G., Schulze, T.K., Crehuet, R. & Lindorff-Larsen, K. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc Natl Acad Sci U S A* **118**(2021).
37. Lancaster, A.K., Nutter-Upham, A., Lindquist, S. & King, O.D. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics* **30**, 2501-2 (2014).
38. Alberti, S., Halfmann, R., King, O., Kapila, A. & Lindquist, S. A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. *Cell* **137**, 146-58 (2009).
39. Bolognesi, B. *et al.* A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity upon Increased Protein Expression. *Cell Rep* **16**, 222-231 (2016).
40. Hardenberg, M., Horvath, A., Ambrus, V., Fuxreiter, M. & Vendruscolo, M. Widespread occurrence of the droplet state of proteins in the human proteome. *Proc Natl Acad Sci U S A* **117**, 33254-33262 (2020).
41. Paiz, E.A. *et al.* Beta turn propensity and a model polymer scaling exponent identify intrinsically disordered phase-separating proteins. *J Biol Chem* **297**, 101343 (2021).
42. van Mierlo, G. *et al.* Predicting protein condensate formation using machine learning. *Cell Rep* **34**, 108705 (2021).
43. Saar, K.L. *et al.* Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proc Natl Acad Sci U S A* **118**, e2019053118 (2021).
44. Sun, J. *et al.* Precise prediction of phase-separation key residues by machine learning. *Nat Commun* **15**, 2662 (2024).
45. Chen, Z. *et al.* Screening membraneless organelle participants with machine-learning models that integrate multimodal features. *Proc Natl Acad Sci U S A* **119**, e2115369119 (2022).

46. Liang, Q. *et al.* MolPhase, an advanced prediction algorithm for protein phase separation. *The EMBO Journal* **43**, 1898-1918 (2024).
47. Zhang, S., Lim, C.M., Occhetta, M. & Vendruscolo, M. AlphaFold2-based prediction of the co-condensation propensity of proteins. *Proc Natl Acad Sci U S A* **121**, e2315005121 (2024).
48. Liao, S. *et al.* A sequence-based model for identifying proteins undergoing liquid-liquid phase separation/forming fibril aggregates via machine learning. *Protein Sci* **33**, e4927 (2024).
49. Frank, M., Ni, P., Jensen, M. & Gerstein, M.B. Leveraging a large language model to predict protein phase transition: A physical, multiscale, and interpretable approach. *Proc Natl Acad Sci U S A* **121**, e2320510121 (2024).
50. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
51. Boke, E. *et al.* Amyloid-like Self-Assembly of a Cellular Compartment. *Cell* **166**, 637-650 (2016).
52. Riback, J.A. *et al.* Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. *Cell* **168**, 1028-1040.e19 (2017).
53. Saito, M. *et al.* Acetylation of intrinsically disordered regions regulates phase separation. *Nat Chem Biol* **15**, 51-61 (2019).
54. Loya, T.J., O'Rourke, T.W., Simke, W.C., Kelley, J.B. & Reines, D. Nab3's localization to a nuclear granule in response to nutrient deprivation is determined by its essential prion-like domain. in *PloS one* Vol. 13 e0209195 (2018).
55. Banjade, S. *et al.* Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck. *Proc Natl Acad Sci U S A* **112**, E6426-35 (2015).
56. Mackinder, L.C.M. *et al.* A repeat protein links Rubisco to form the eukaryotic carbon-concentrating organelle. *Proceedings of the National Academy of Sciences* **113**, 5958-5963 (2016).
57. Mészáros, B. *et al.* PhaSePro: the database of proteins driving liquid–liquid phase separation. *Nucleic Acids Research* **48**, D360-D367 (2019).
58. Landrum, M.J. & Kattman, B.L. ClinVar at five years: Delivering on the promise. *Hum Mutat* **39**, 1623-1630 (2018).
59. Landrum, M.J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res* **48**, D835-d844 (2020).
60. Ibrahim, A.Y. *et al.* Intrinsically disordered regions that drive phase separation form a robustly distinct protein class. *Journal of Biological Chemistry* **299**(2023).
61. Bremer, A. *et al.* Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat Chem* **14**, 196-207 (2022).
62. Lin, Y.H. & Chan, H.S. Phase Separation and Single-Chain Compactness of Charged Disordered Proteins Are Strongly Correlated. *Biophys J* **112**, 2043-2046 (2017).
63. Das, R.K. & Pappu, R.V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences* **110**, 13392-13397 (2013).
64. Sawle, L. & Ghosh, K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J Chem Phys* **143**, 085101 (2015).
65. Zheng, W., Dignon, G., Brown, M., Kim, Y.C. & Mittal, J. Hydropathy Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins. *J Phys Chem Lett* **11**, 3408-3415 (2020).
66. Li, P. *et al.* Phase transitions in the assembly of multivalent signalling proteins. *Nature* **483**, 336-340 (2012).
67. Lin, Y., Protter, D.S., Rosen, M.K. & Parker, R. Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. *Mol Cell* **60**, 208-19 (2015).

68. Zeng, M. *et al.* Phase Transition in Postsynaptic Densities Underlies Formation of Synaptic Complexes and Synaptic Plasticity. *Cell* **166**, 1163-1175.e12 (2016).
69. Hou, S. *et al.* Machine learning predictor PSPire screens for phase-separating proteins lacking intrinsically disordered regions. *Nat Commun* **15**, 2147 (2024).
70. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* 3111–3119 (Curran Associates Inc., Lake Tahoe, Nevada, 2013).
71. Chu, X. *et al.* Prediction of liquid-liquid phase separating proteins using machine learning. *BMC Bioinformatics* **23**, 72 (2022).
72. Arar, M., Shamir, A. & Bermano, A.H. Learned Queries for Efficient Local Attention. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10831-10842 (2022).
73. Chu, X. *et al.* Twins: Revisiting the Design of Spatial attention in Vision Transformers. in *35th Conference on Neural Information Processing Systems (NeurIPS 2021)* (2021).
74. Ramachandran, P. *et al.* Stand-Alone Self-Attention in Vision Models. in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (Vancouver, Canada, 2019).
75. Dai, Z., Liu, H., Le, Q.V. & Tan, M. CoAtNet: Marrying Convolution and Attention for All Data Sizes. in *35th Conference on Neural Information Processing Systems* (2021).
76. Wang, X. *et al.* LLPSDB v2.0: an updated database of proteins undergoing liquid-liquid phase separation in vitro. *Bioinformatics* **38**, 2010-2014 (2022).
77. You, K. *et al.* PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic Acids Res* **48**, D354-d359 (2020).
78. Ning, W. *et al.* DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. *Nucleic Acids Res* **48**, D288-d295 (2020).
79. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-2 (2012).
80. Nott, Timothy J. *et al.* Phase Transition of a Disordered Nuage Protein Generates Environmentally Responsive Membraneless Organelles. *Molecular Cell* **57**, 936-947 (2015).
81. Roovers, E.F. *et al.* Tdrd6a Regulates the Aggregation of Buc into Functional Subcellular Compartments that Drive Germ Cell Specification. *Developmental Cell* **46**, 285-301.e9 (2018).
82. Supek, F., Lehner, B. & Lindeboom, R.G.H. To NMD or Not To NMD: Nonsense-Mediated mRNA Decay in Cancer and Other Genetic Diseases. *Trends Genet* **37**, 657-668 (2021).
83. Yang, L. *et al.* Phase separation as a possible mechanism for dosage sensitivity. *Genome Biol* **25**, 17 (2024).
84. Rekhi, S. *et al.* Expanding the molecular language of protein liquid-liquid phase separation. *Nat Chem* **16**, 1113-1124 (2024).
85. Dzuricky, M., Rogers, B.A., Shahid, A., Cremer, P.S. & Chilkoti, A. De novo engineering of intracellular condensates using artificial disordered proteins. *Nat Chem* **12**, 814-825 (2020).
86. Dai, Y. *et al.* Programmable synthetic biomolecular condensates for cellular control. *Nat Chem Biol* **19**, 518-528 (2023).
87. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
88. Raimondi, D. *et al.* DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res* **45**, W201-w206 (2017).
89. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants

throughout the human genome. *Nucleic Acids Res* **47**, D886-d894 (2019).

90. Li, G., Panday, S.K. & Alexov, E. SAAFEC-SEQ: A Sequence-Based Method for Predicting the Effect of Single Point Mutations on Protein Thermodynamic Stability. *Int J Mol Sci* **22**, 606 (2021).
91. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91-95 (2021).
92. Liu, Y., Zhang, T., You, N., Wu, S. & Shen, N. MAGPIE: accurate pathogenic prediction for multiple variant types using machine learning approach. *Genome Med* **16**, 3 (2024).
93. Alderson, T.R., Pritisanac, I., Kolaric, D., Moses, A.M. & Forman-Kay, J.D. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proc Natl Acad Sci U S A* **120**, e2304302120 (2023).
94. Wu, Y., Li, R., Sun, S., Weile, J. & Roth, F.P. Improved pathogenicity prediction for rare human missense variants. *Am J Hum Genet* **108**, 1891-1906 (2021).
95. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
96. Erdős, G., Pajkos, M. & Dosztányi, Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res* **49**, W297-w303 (2021).
97. Asgari, E. & Mofrad, M.R. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLoS One* **10**, e0141287 (2015).
98. Shi, Y. *et al.* DeepAntigen: a novel method for neoantigen prioritization via 3D genome and deep sparse learning. *Bioinformatics* **36**, 4894-4901 (2020).

Acknowledgement

This project is supported by the National Key Research and Development Program (2022YFE0125300 to Y.S, 2020YFA0509700 to Q.L), Innovation Program of Shanghai Municipal Education Commission (2023ZKZD16) to G.H, the National Natural Science Foundation of China (82071262) to G.H, Shanghai Municipal Science and Technology Major Project (2017SHZDZX01 to Y.S, 20JC1418600 to G.H), Key Technology Breakthrough Program of Ningbo Sci-Tech Innovation YONGJIANG 2035 (2024Z221) to G.H, and Shanghai Jiao Tong University STAR Grant (YG2023ZD26, YG2023LC14, and YG2024QNA59 to G.H, YG2022ZD024, YG2022QN111 to Y.S, 23X010300421 to Q.L).

Author contributions

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

Competing Interests Statement

The authors declare no competing interests.

Figures

Fig. 1

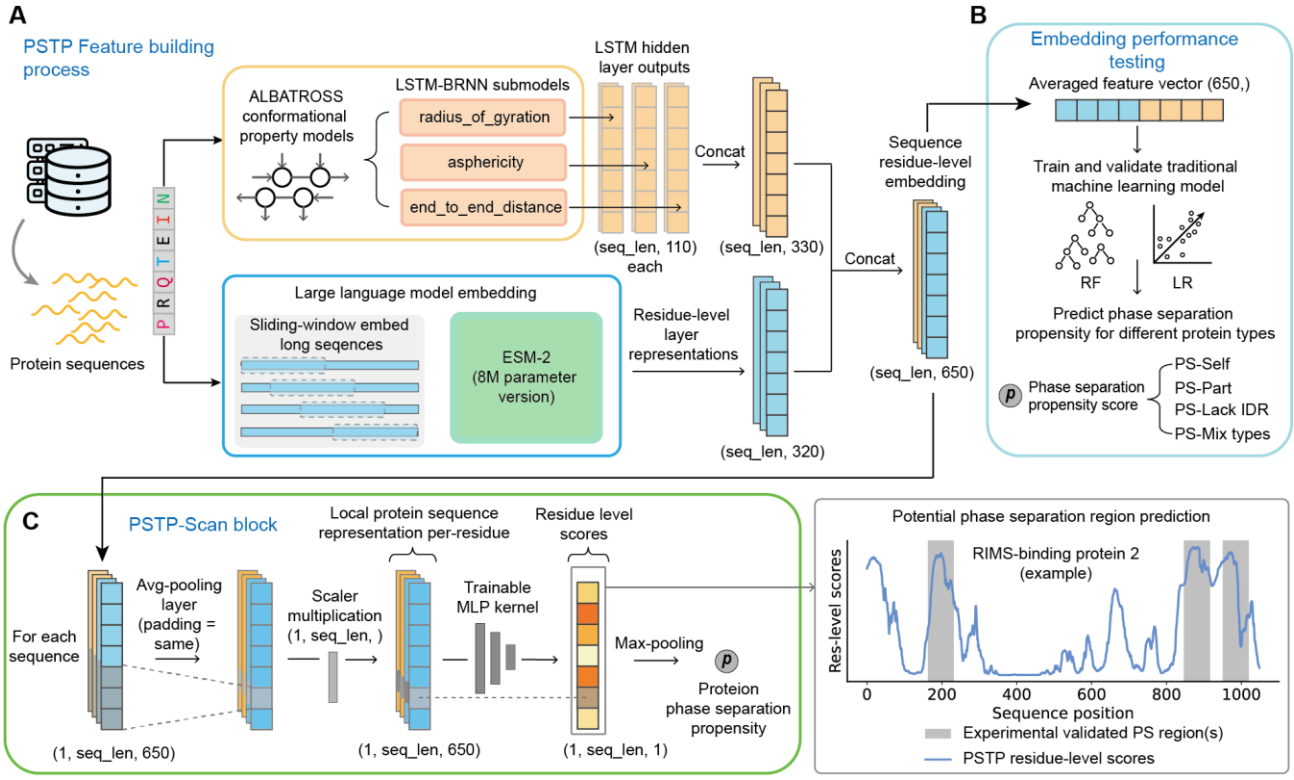


Fig. 1. Schematic of PSTP embedding and machine learning (ML) architecture.

(A) Protein sequences are converted into feature matrices using the hidden layer outputs of the ALBATROSS LSTM-BRNN and the final layer representations from the ESM-2 (8M version) language model. For each sequence, the outputs from three ALBATROSS sub-models are extracted to form a matrix with dimensions of sequence-length×330, while ESM-2 generates a matrix of sequence-length×320. 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. sequence-length: length of protein 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), mixed PS proteins (PS-Mix types), and PS proteins lacking IDRs (PS-Lack IDR).

(C) Schematic of the PSTP-Scan block. The embedding matrix generated in (A) is used as input to

802 produce attention scores, which serve as residue-level PS scores to detect PS regions. The maximum
803 attention score for each sequence is output as the sequence-level PS propensity score. MLP: multilayer
804 perceptron.
805

Fig. 2

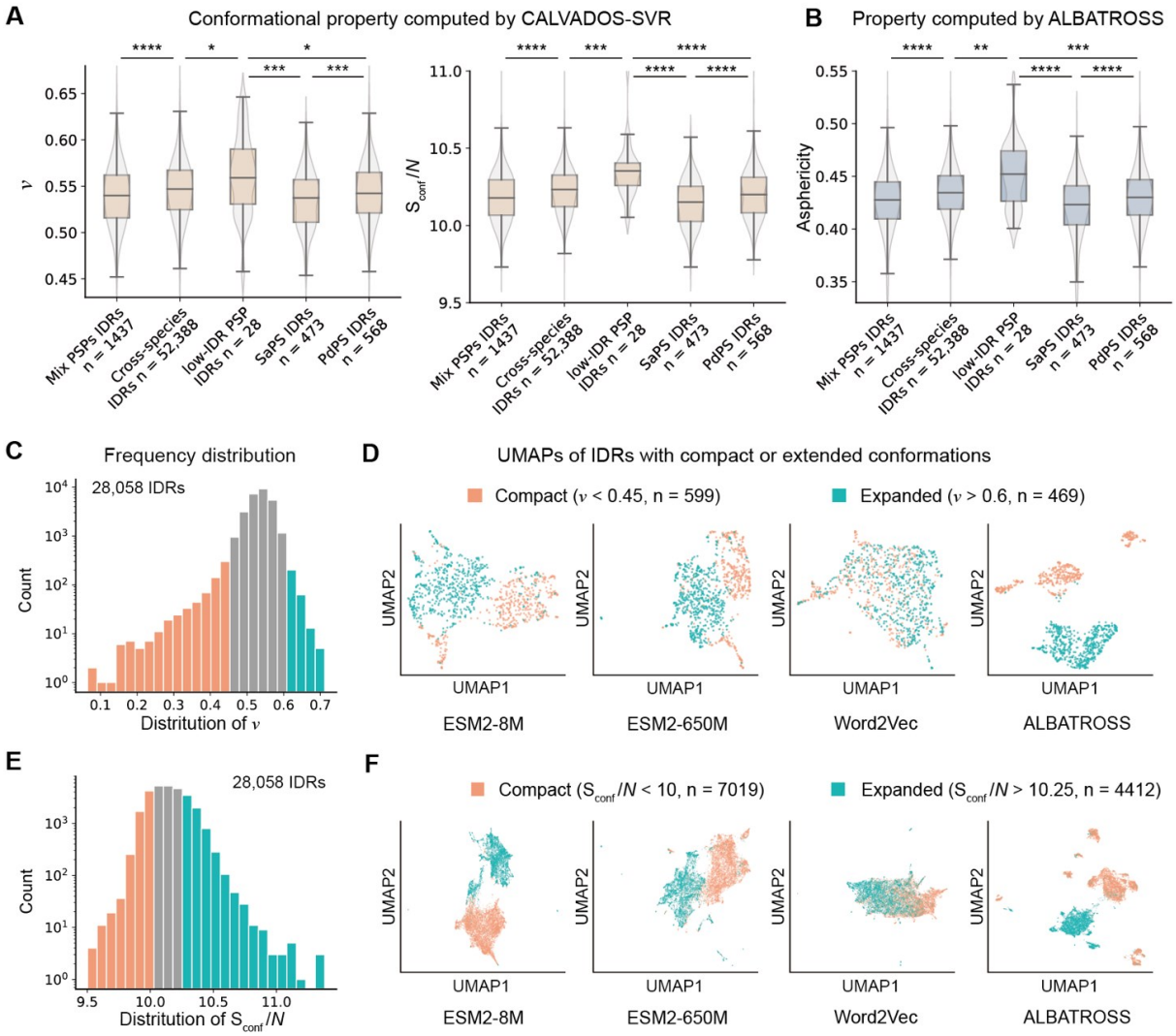


Fig. 2. Analysis of IDR chain compaction properties and UMAP visualization of IDRs with varying chain compaction.

(A and B) Comparison of conformational property metrics of intrinsically disordered regions (IDRs) across different protein types. Metrics reflecting IDR global conformation include v and S_{conf}/N in (A), and asphericity in (B) (PSP: Phase separating proteins; ****P < 0.0001, ***P < 0.001, **P < 0.01, *P < 0.05; two-sided Mann–Whitney U test; boxplot components within each violin, from top to bottom are maxima, upper quartile, median, lower quartile, and minima.).

(C) Distribution of v values for IDRs. IDRs with $v < 0.45$ are classified as extremely compact, while those with $v > 0.55$ are classified as expanded. To balance data size, IDRs with $v > 0.6$ 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 metric. 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.

Fig. 3

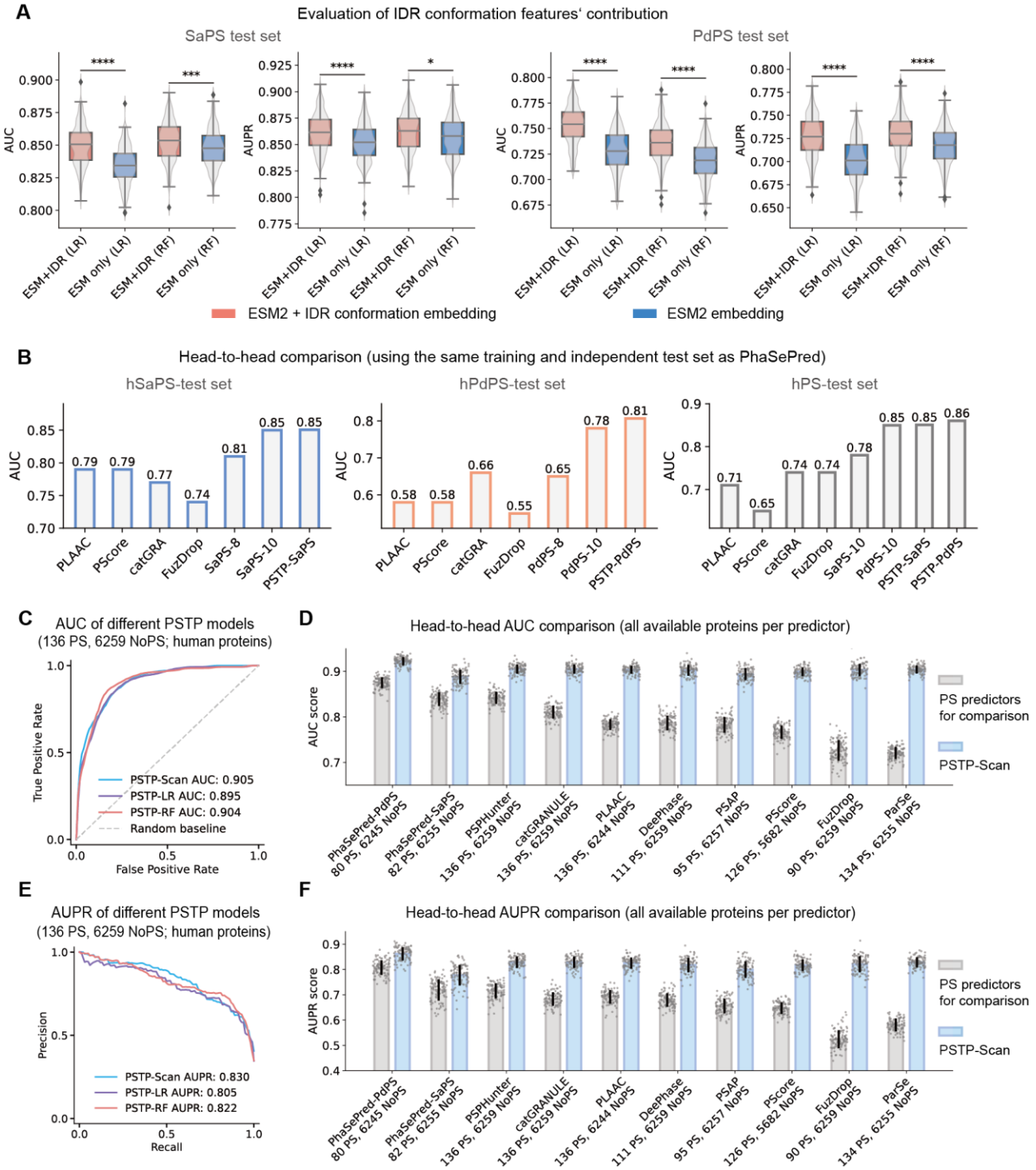


Fig. 3. Evaluation of PSTP Model performance.

(A) Performance of various PSTP models on the SaPS and PdPS test sets. 'LR' represents Logistic Regression and 'RF' represents Random Forest, 'ESM-2 embedding' represents the large language model embedding. Performance was evaluated using 200 replicates of subset sampling from the background dataset. PSTP models trained with the full set of features (including IDR conformational

feature) outperform those trained solely on ESM features (large language model feature). (****P < 0.0001, ***P < 0.001, *P < 0.05; two-tailed Student's t-test; boxplot components within each violin, from top to bottom are maxima, upper quartile, median, lower quartile, and minima.).

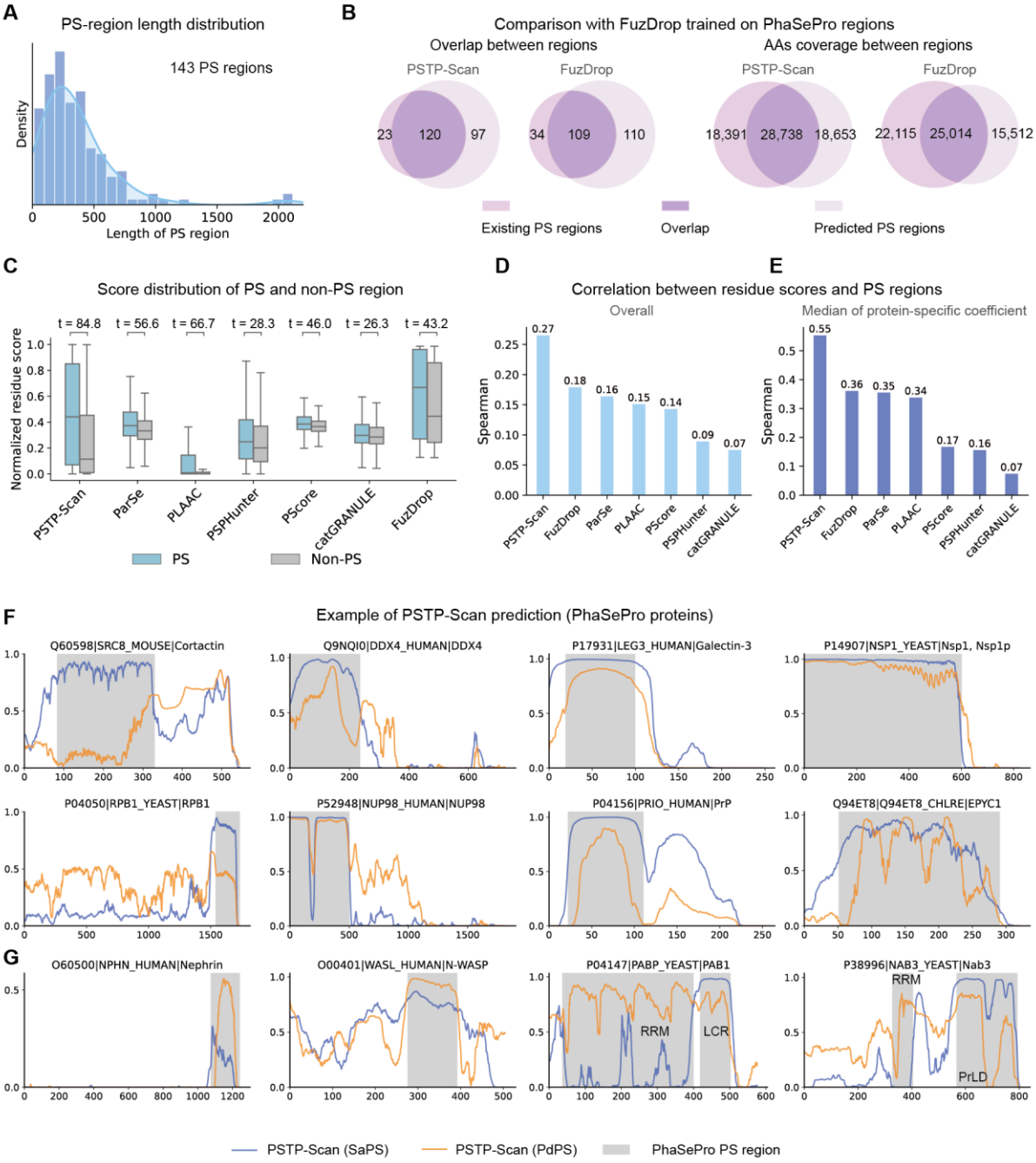
(B) Comparison of PSTP-LR with PhaSePred as well as other representative models using the same independent test set as PhaSePred. PSTP-LR was trained and evaluated on the independent validation set following the same dataset split for PhaSePred. Results for PhaSePred and other PS predictors were sourced from the PhaSePred literature.

(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 score 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. AUC (Area under the Receiver Operating Characteristic curve) values were computed and visualized. Data are shown as mean ± SD (Standard Deviation), with scatter points representing performance across sampling repeats.

(E and F) Parallel evaluations to (C and D) but focusing on the AUPR (Area Under the Precision-Recall Curve) values.

853 **Fig. 4**



854

855 Fig. 4. Evaluation of residue-level prediction scores generated by PSTP-Scan and other PS predictors
856 under experimentally validated PS-driving regions in the PhaSePro database (143 PS regions from 121
857 proteins).
858 (A) Distribution of sequence lengths for regions that drive PS.
859 (B) Comparison of PS regions predicted by PSTP-Scan and FuzDrop against those documented in
860 PhaSePro. FuzDrop was trained using PhaSePro PS regions. 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 PS regions versus non-PS regions in proteins documented in PhaSePro. All components have p-value < 0.0001. The boxplot components within each violin, from top to bottom are maxima, upper quartile, median, lower quartile, and minima.

(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. RRM: RNA-recognition motifs; LCD: Low-complexity domain; PrLD: Prion-like domain

Fig. 5

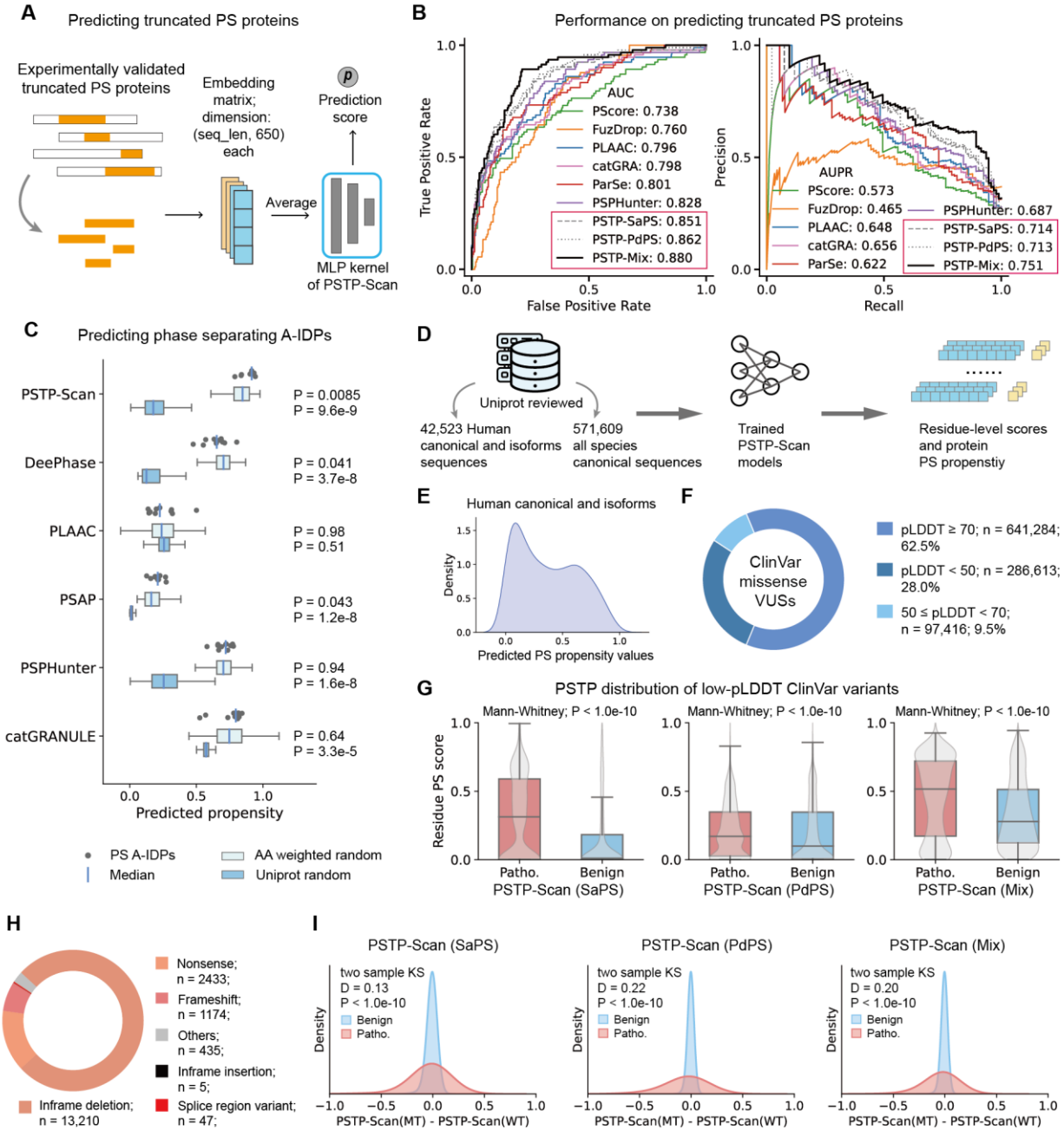


Fig. 5. Extended scope of PS protein prediction and analysis of disease variants.

(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. ‘Random’ represents scores for random 200-residue segments from UniProt-reviewed proteins (n = 10,000), while ‘AA weighted random’ refers to scores for control peptides, generated by repeating octapeptides 25 times as the A-IDPs, matching both the amino acid composition and repeating patterns of the A-IDPs.

(D) PSTP-Scan-generated sequence-level and residue-level predictions for over 570,000 proteins from UniProt.

(E) Score distribution of PSTP-Scan across the human proteome, including both canonical proteins and isoforms.

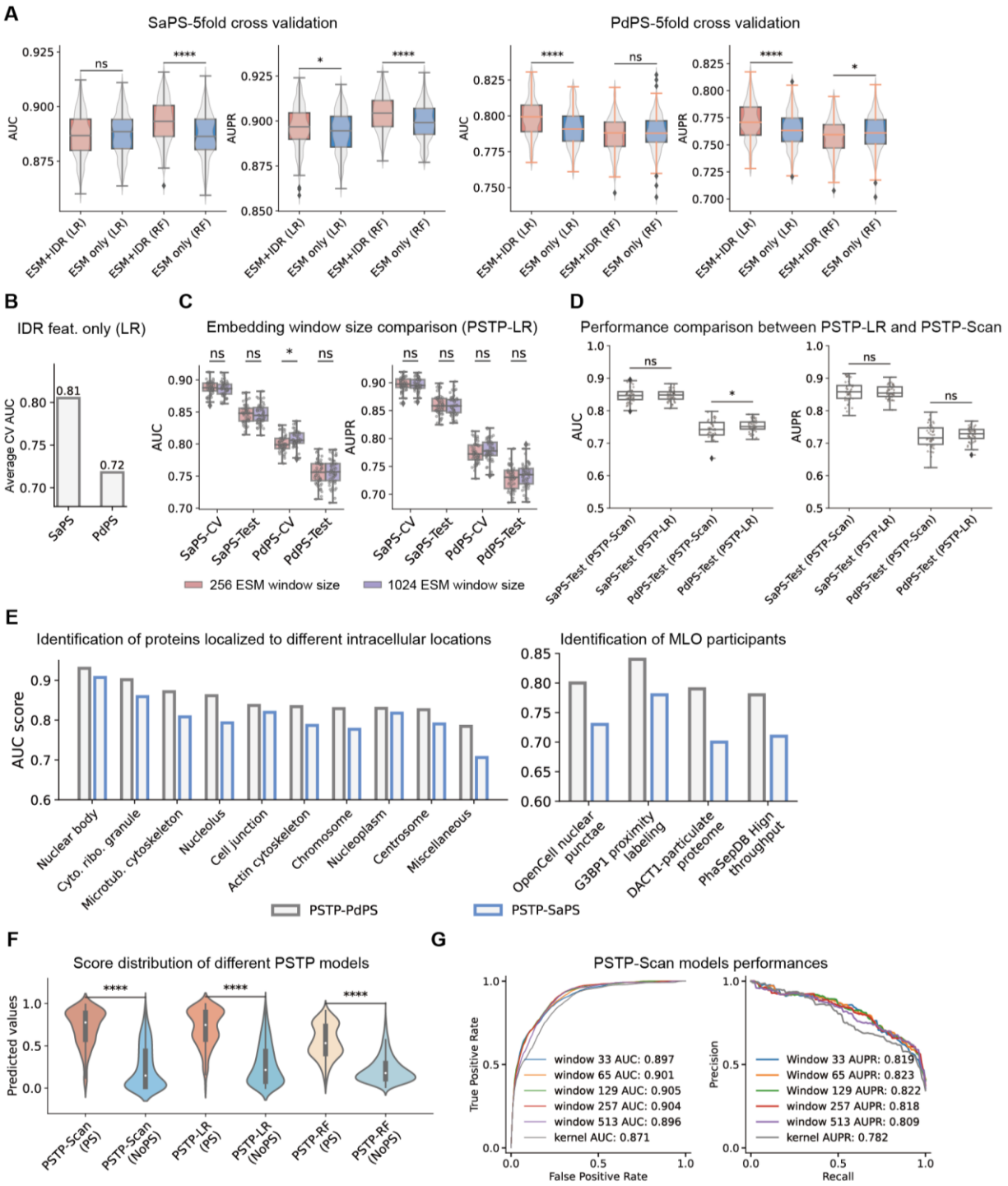
(F) Proportion of variants of uncertain significance (VUSs) based on pLDDT scores. (G) Distribution of PSTP-Scan residue-level scores for pathogenic and benign ClinVar variant positions.

(H) The proportion of ClinVar protein-altering variants, excluding missense variants.

(I) Distribution of PSTP-Scan score differences computed before and after mutations for pathogenic and benign ClinVar variants described in (H).

897 **Supplementary figures**

898 **Fig. S1.**



899

900 Fig. S1. Model performance benchmark, related to Fig. 3.

901 (A) Parallel evaluation to Fig. 3A, performed on the SaPS and PdPS cross-validation sets using 5-fold

902 cross-validation (****P < 0.0001, ***P < 0.001, *P < 0.05; two-tailed Student's t-test; boxplot

903 components within each violin, from top to bottom are maxima, upper quartile, median, lower quartile,

and minima.)

(B) Performance evaluation using only the 330-dimension IDR conformation features using the LR model under 5-fold cross-validation.

(C) Comparison of sliding window embeddings (ESM-2 large language model embeddings) with 256-length and 1024-length thresholds across different cross-validation and independent testing. Scatter points represent the performance scores from repeated sampling. The performance of both window lengths is comparable. (*P < 0.05, ns: not significant; two-tailed Student's t-test; boxplot components within each violin, from top to bottom are maxima, upper quartile, median, lower quartile, and minima.).

(D) Performance comparison of PSTP-LR and PSTP-Scan on the SaPS and PdPS test set. Performance was evaluated using 50 replicates of subset sampling from the background dataset, with scatter points representing performance scores from these repeats. Both models show similar performance (ns: not significant, *P < 0.05; two-tailed Student's t-test).

(E) Prediction performance of PSTP-LR models trained on SaPS data (PSTP-SaPS) and PdPS data (PSTP-PdPS) in predicting MLO participants from those localized to intracellular locations (left) and four MLO datasets (right). AUC values were calculated using proteins in these datasets as positive samples and human NoPS proteins as negative samples, excluding those in the model training data.

(F) Discriminative power of PSTP-Scan models on the independent test set.

(G) Performance of PSTP-Scan block with different window size. Unlike Fig. 3 C-F, where the default three parallel PSTP-Scan block was used, each model here represents a single-block PSTP-Scan with a specific window size. The 'kernel' model refers to the configuration where the averaged vector is directly passed to the MLP kernel to generate model output.

Fig. S2.

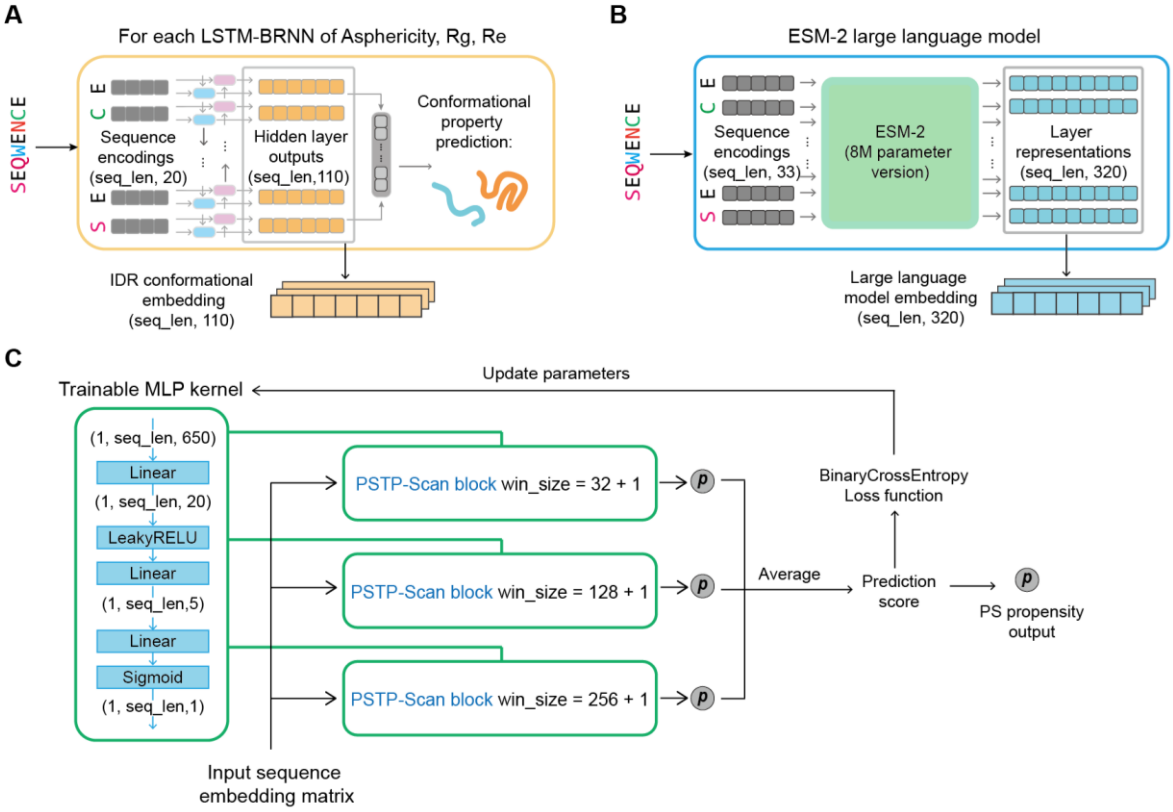


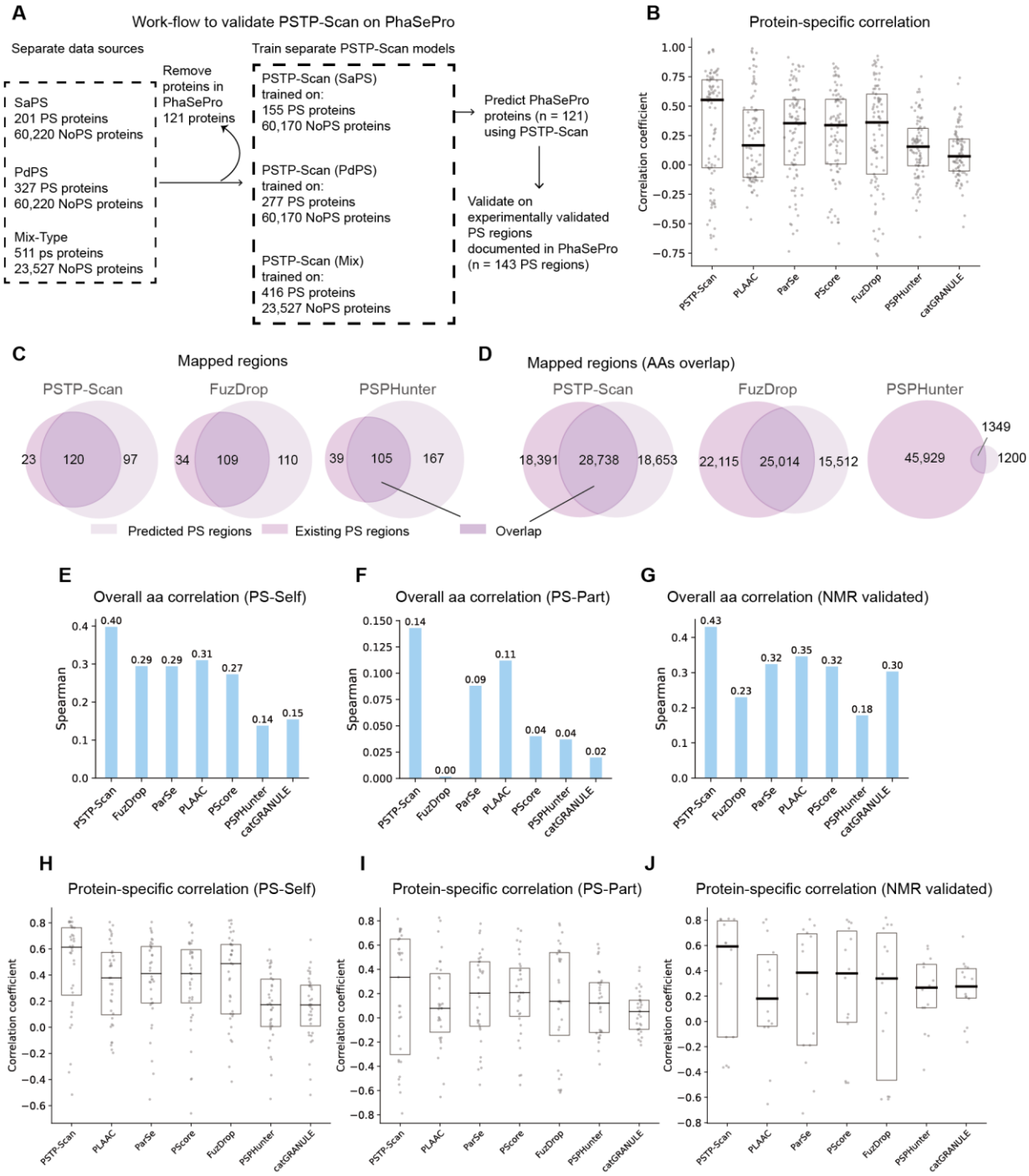
Fig. S2. Schematic of the PSTP model, related to Fig. 2.

(A) IDR conformational embedding: For each LSTM-BRNN sub-model in ALBATROSS, the sequence-length × 110-dimension hidden layer output is extracted prior to predicting the final property value.

(B) Large language model embedding: Residue-level layer representations from the ESM-2 model (8 million parameter version) are extracted, resulting in a sequence-length ×320-dimensional matrix for each input sequence.

(C) PSTP-Scan model training and inference: Three parallel PSTP-Scan blocks with identical architectures but different average pooling window sizes are employed to capture sequence information at varying local lengths. The predicted sequence-level propensities (single-value output from each block) are averaged to produce the final output. This output is optimized via backpropagation to update the weights of the MLP.

943 **Fig. S3**



944

945 Fig. S3. Evaluation of residue-level prediction scores for proteins documented in the PhaSePro

946 database.

947 (A) Illustration of the validation process for PSTP-Scan in predicting PS driving regions. Each PSTP-

948 Scan model was trained excluding proteins documented in PhaSePro to prevent self-validation and

949 data leakage.

950 (B) Distribution of the protein-level Spearman correlation coefficients computed between the residue-

951 level scores and the binary vector representing the PS region(s) for each protein, corresponding to Fig.
952 4E. Scatter points represent individual coefficient values, while the boxplot components within each
953 violin, from top to bottom, represent upper quartile, median, lower quartile.
954 (C and D) Comparison of PS regions predicted by PSTP-Scan, FuzDrop, and PSPHunter against those
955 documented in PhaSePro, related to Fig. 4B. This panel illustrates the overlap between predicted
956 regions and the corresponding overlap at the individual amino acid level (D).
957 (E and H) Parallel evaluations to Fig. 4C (E) and Fig. 4B (H), focusing exclusively on PS regions of
958 PS-Self proteins (58 records).
959 (F and I) Parallel evaluations to Fig. 4C (F) and Fig. 4B (I), focusing exclusively on PS regions of PS-
960 Part proteins (50 records).
961 (G and J) Parallel evaluations to Fig. 4C (G) and Fig. 4B (J), focusing exclusively on NMR-validated
962 PS regions (14 records).
963
964

Fig. S4.

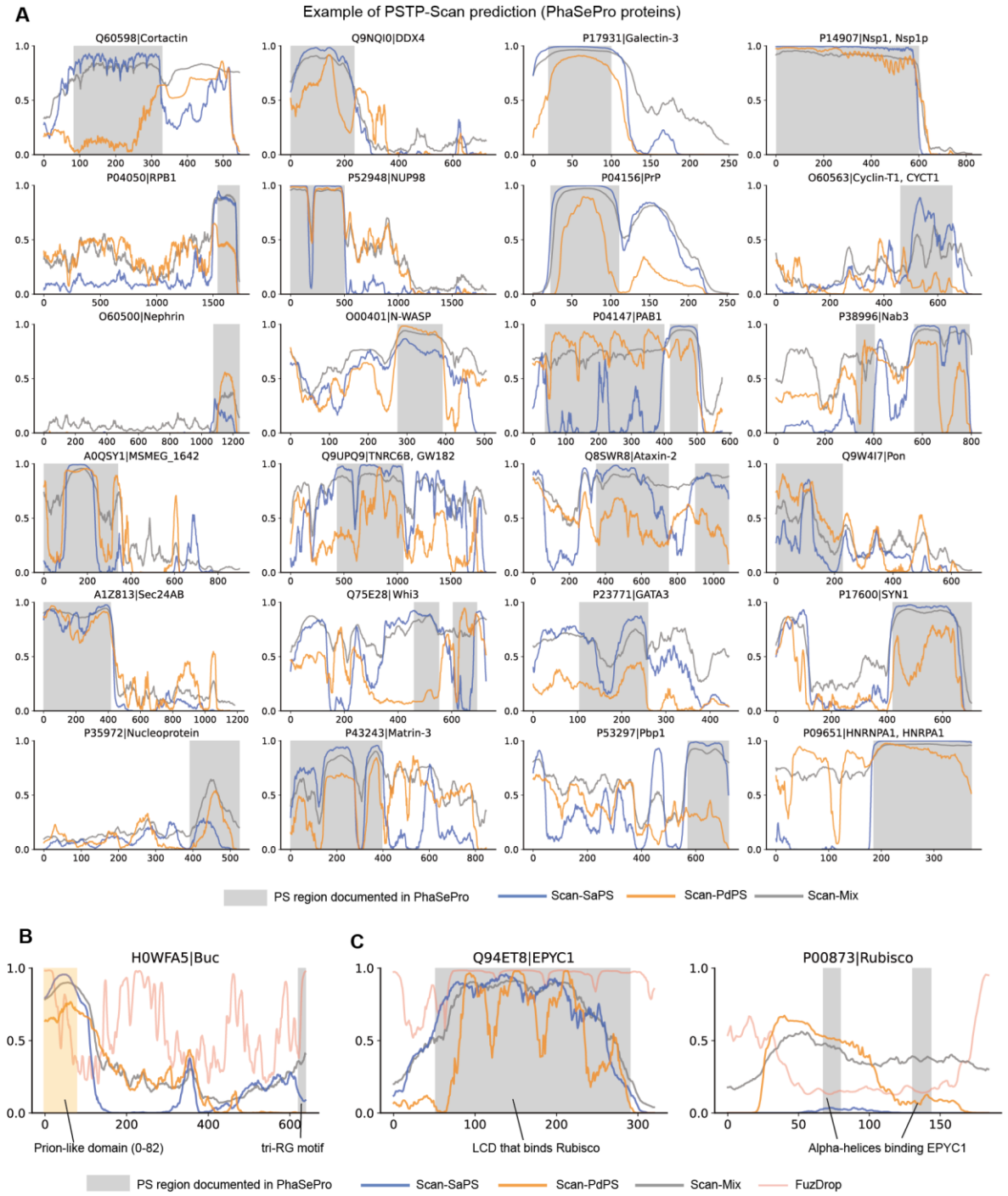
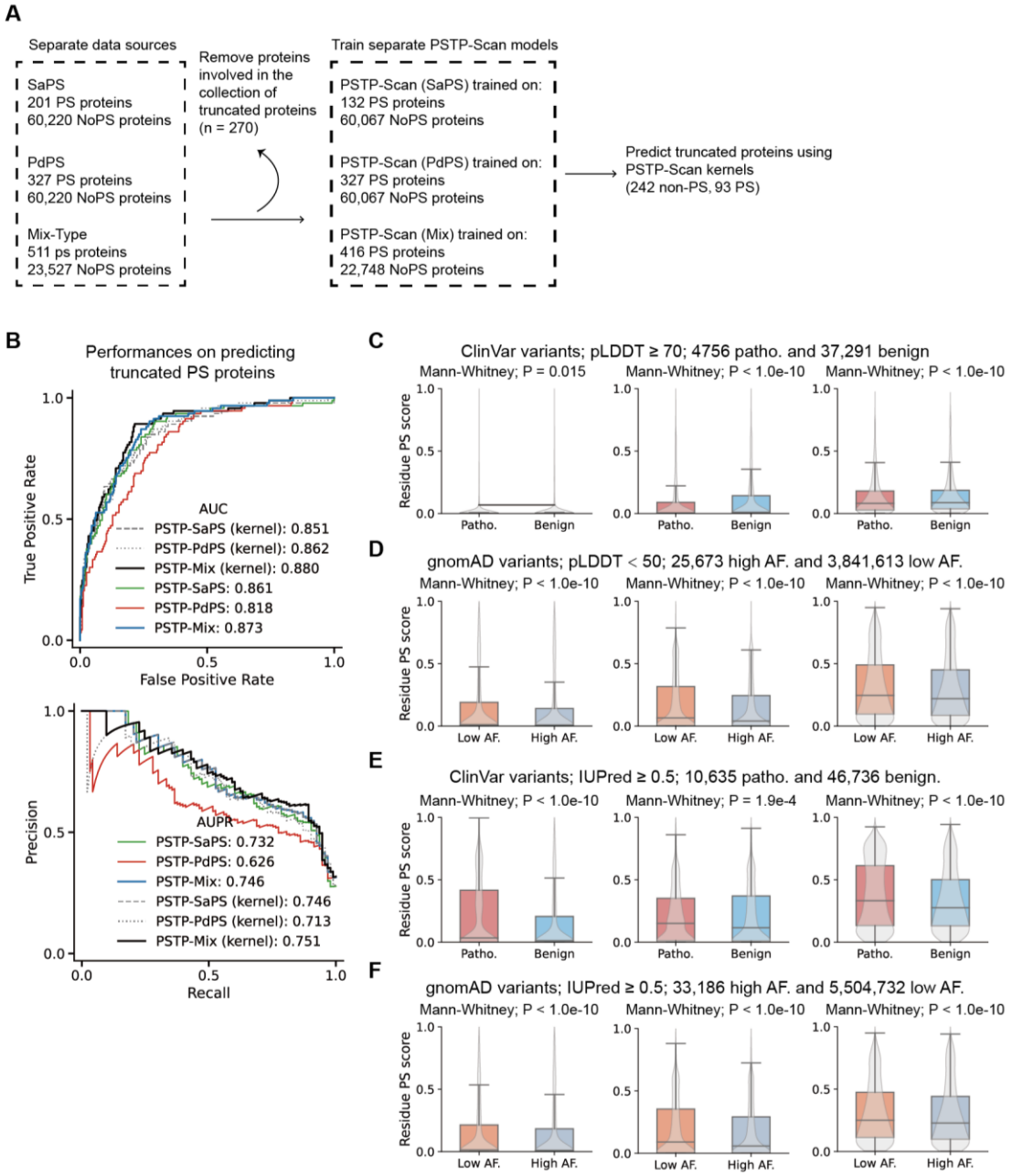


Fig. S4. Example of residue-level scores predicted by three PSTP-Scan models, with experimentally validated PS regions highlighted in gray. RRM: RNA-recognition motifs.

971 **Fig. S5**



972

973 Fig. S5. Extended scope of PS protein prediction and analysis of disease variants, related to Fig. 5

974 (A) Illustration of the validation process for PSTP-Scan in predicting truncated proteins. PSTP-Scan

975 models were trained after excluding proteins that overlap (share the same UniProt entry) with those in

976 the truncated protein validation dataset.

977 (B) AUC and AUPR evaluations of different PSTP-scan model for predicting truncated PS proteins.

978 'kernel' refers to using only the trained MLP kernel from the corresponding 'Kernel' refers to using

979 only the trained MLP kernel from the corresponding PSTP-Scan model to predict based on the

980 averaged feature vector of each protein. PSTP-SaPS, PSTP-PdPS, and PSTP-Mix refer to the three
981 trained PSTP-Scan models, respectively, with a single block applied and the window size set to 32+1.
982 (C) PSTP residue-level scores for high pLDDT (pLDDT ≥ 70) pathogenicity and benign ClinVar
983 variants.
984 (D) PSTP residue-level scores for variants with high allele frequency (AF) and low allele frequency
985 (AF) from gnomAD, located within low pLDDT (pLDDT < 50) regions.
986 (E) PSTP residue-level scores for pathogenicity and benign ClinVar variants located in disordered
987 regions as defined by IUPred3 (IUPred ≥ 0.5).
988 (F) PSTP residue-level scores for high AF and low AF variants from gnomAD, located in disordered
989 regions as defined by IUPred3 (IUPred ≥ 0.5).
990

Fig. S6.

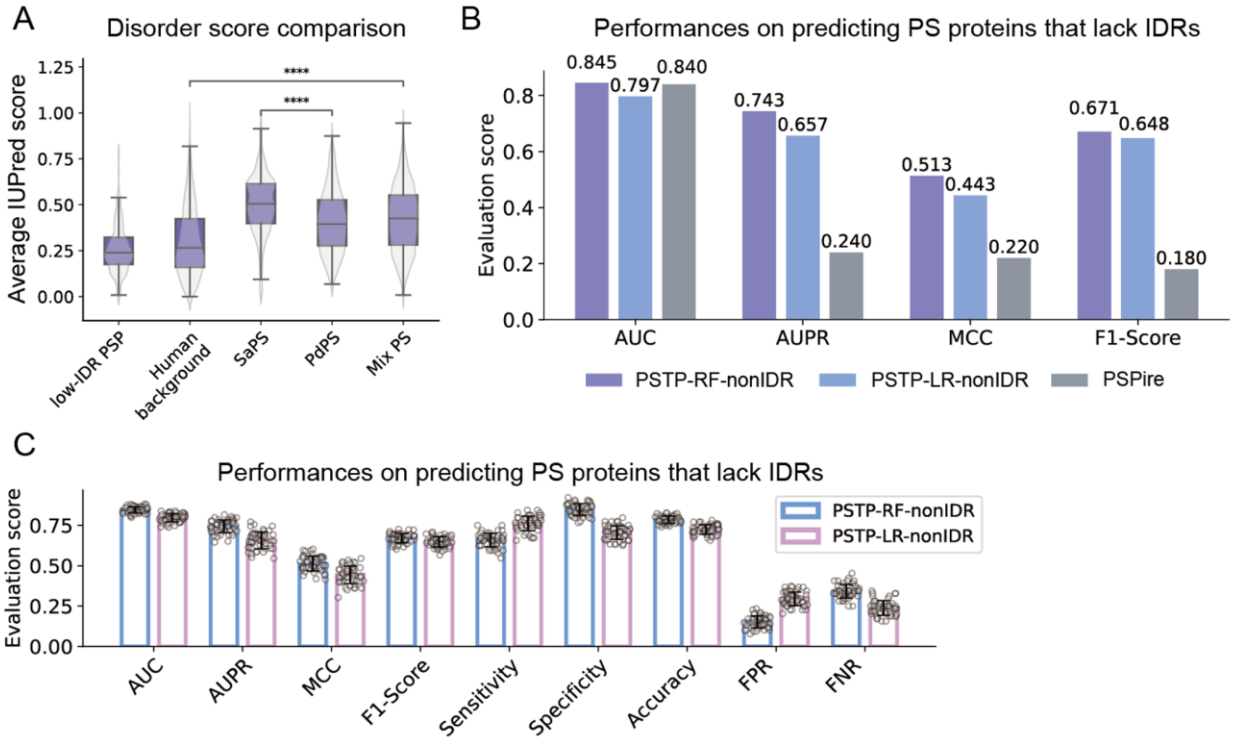


Fig. S6. Model performance in predicting PS proteins lacking IDRs.

(A) Distribution of IUPred3 scores, reflecting protein disorder across different types of proteins. PSP: PS protein; Low-IDR PSP: PS Proteins that lack IDRs; Human background: Proteins from the human proteome; SaPS: Self-assembly PS proteins; PdPS: Partner-dependent PS proteins; Mix PS: Mixture of different types of PS proteins.

(B and C) Performance of PSTP-RF (PSTP-RF-nonIDR) and PSTP-LR (PSTP-LR-nonIDR) in predicting PS proteins. Both models were evaluated using the same training and independent datasets used for validating PSPire[69]. A head-to-head comparison was performed between PSTP models and PSPire, with the performance of PSPire extracted from the literature[69]. PSTP models were trained exclusively using the large language model features.

1005