Decoding Missense Variants by Incorporating Phase Separation via Machine Learning

Mofan Feng[1,2+], Xiaoxi Wei[1+], Xi Zheng[1,2], Liangjie Liu[1,2], Lin Lin[1], Manying Xia[1], Guang He[1,2*], Yi Shi[1,2*], Qing Lu[3*]

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.

3 Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Shanghai Jiao Tong University, Shanghai, China. Ear Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China. Shanghai Key Laboratory of Translational Medicine on Ear and Nose Diseases, Shanghai, China.

+ These authors contributed equally as co-first authors.

*** Corresponding author at: Bio-X Institutes, Shanghai Jiao Tong University, 1954 Huashan Road, Shanghai 200030, P.R.China. Tel & Fax: +86 (0) 21 62822491 Email:**

Guang He: heguang@sjtu.edu.cn,

Yi Shi: yishi@sjtu.edu.cn

Qing Lu: luqing67@sjtu.edu.cn

Abstract

Computational models have made significant progress in predicting the effect of protein variants. However, deciphering numerous variants of unknown significance (VUS) located within intrinsically disordered regions (IDRs) remains challenging. To address this issue, we introduced liquid-liquid phase separation (LLPS), which is tightly linked to IDRs, into the analysis of missense variants. LLPS is vital for multiple physiological processes. By leveraging missense variants that alter phase separation propensity, we developed an innovative machine-learning approach called PSMutPred to predict the impact of missense mutations on phase separation. PSMutPred shows robust performance in predicting missense variants that affect natural phase separation. In vitro experiment results further underscore its validity. By applying PSMutPred to over 522,000 ClinVar missense variants, we identified that PSMutPred significantly contributes to decoding the pathogenesis of disease variants, especially those in IDRs. Our work provides invaluable insights into the understanding of a vast number of VUSs in IDRs, thereby accelerating clinical interpretation and diagnosis of disease variants.

Keywords: phase separation, missense mutation, variant effect prediction, machine learning, intrinsically disordered regions

Introduction

Approximately 25% of disease-associated missense variants are located in intrinsically disordered regions (IDRs)[1], present in around 63% of human proteins. However, studying variants in IDRs is challenging due to the lack of a fixed tertiary structure and limited evolutionary conservation[2]. As a result, these IDR variants are commonly annotated as variants of unknown significance (VUSs), making it difficult to evaluate and predict their functional impact on diseases.

IDRs, particularly low-complexity IDRs, are crucial in regulating liquid-liquid phase

separation (LLPS), a physicochemical process where molecules separate into a confined fluid phase[3]. Phase separation driven by IDRs leads to the formation of membrane-less organelles called condensates which play essential roles in various biological processes. An increasing number of proteins are recognized for their roles via the formation of phase separation condensates. These condensates include the nucleolus and transcription factories in the nucleus[4], stress granules[5], and protein densities at neuronal synapses[6] and inner-ear hair cells[7-9].

Accumulating studies in IDRs reveal that missense variants in IDRs can perturb protein functions by altering phase separation. The extent of this impact is determined by the specific residues altered and the number of mutations[4]. These mutations can affect the IDR conformations, intra-molecular interactions, and inter-molecular protein-protein interactions, leading to abnormal condensate formations[4]. This can cause either loss[10,11] or gain[12-14] of natural phase separation, leading to misplacement and related gain or loss-of-function outcomes. Notably, the gain of unnecessary phase separation can contribute to disease phenotypes like neurodegenerative disorders such as Alzheimer's and Parkinson's diseases [15-17]. These pioneering works have confirmed the association between phase separation and disease variants in several cases.

Existing missense variant prediction algorithms typically rely on protein structural features or evolutionary features derived from multiple sequence alignments (MSA). Using such features, prediction algorithms of disease variants have made significant progress in predicting protein clinical relevance[18-23], bridging the variant interpretation gap left by experiments due to the cost and labor constraints. However, for numerous disease variants located in IDRs which often lack a fixed structure[4] and show poor evolutionary conservation[2], encoding variants using these traditional features becomes unsuitable, making accurate prediction challenging. To solve this obstacle, we consider phase separation, which has been widely recognized as a common property of IDRs associated with various diseases, making it a promising feature for accurately predicting IDR disease variants.

Machine learning algorithms have shown remarkable performance in predicting the propensity of proteins to undergo phase separation. Phase separation is a complex

biophysical process that can be influenced by multiple types of properties. Besides solution properties like temperature and electrostatic property, sequence-related features have been extensively studied in the context of phase separation[24]. Several such features have been identified that made phase separation prediction effective[25-31] including pi-pi[27], cation-pi interactions[28,29,32], electrostatic interactions[31,33], hydrophobic interactions[30,32], and the valency and patterning of the low-complexity region[5,28,34]. Protein-protein interactions (PPIs) and post-translational modifications (PTMs) were also found to be promising features recently[35-37]. Missense mutations can cause pathogenic changes by altering phase separation, however, there are currently no available algorithms to predict the impact of missense variants on phase separation.

To enhance comprehension of the correlation between IDR missense mutations and diseases, we approached the issue from the phase separation standpoint and devised features to represent the underlying phase separation alterations resulting from the mutation. Using experimentally validated missense mutations that alter phase separation propensity in proteins naturally undergoing this process, we constructed a computational tool named PSMutPred to predict the effect of missense mutation on phase separation. We demonstrated that missense mutations that impact the normal phase separation propensity can be accurately predicted, and the direction of the shifts in the PS threshold caused by these mutations are also predictable.

We next explored whether PS can be integrated into the pathogenicity prediction of disease variants. By analyzing PSMutPred scores for 520,000 + missense variants, we observed a strong positive correlation between the variant's tendency to impact phase separation and its pathogenicity, especially in proteins prone to phase separation. We also observed that in proteins related to neurodegenerative diseases, disease variants that were likely to enhance phase separation were proportionally more prevalent compared to those that might weaken phase separation. By integrating PSMutPred scores and other PS-related features into EVE [22], an unsupervised high-accuracy pathogenicity prediction method for missense variants, significant prediction improvement can be achieved, especially for those within IDRs. Our findings show that PSMutPred serves as an informative encoding for IDR variants, which fills the existing

void in mutation prediction within the IDRs, offering fresh perspectives and opportunities for pre-screening studies on the pathogenicity of numerous VUSs.

Results

Collection of Phase Separation (PS)-related missense variants

To investigate relationships between missense mutations and protein phase separation (PS) properties, we reviewed missense variants with altered PS propensity that were documented in PhaSepDB[38] and LLPSDB[39,40] databases. To minimize the noise effect caused by multiple mutations within a single sequence, we narrowed our selection to mutation records with a limited number of mutations in the sequence that alter the normal PS threshold (annotated as 'Impact' mutations). Examples include P22L in *Ape1*, which solidifies semi-liquid *Ape1* Droplets[41], and S48E in *TDP-43,* which disrupts PS[42]. We limited our analysis to variants influencing proteins' spontaneous PS, excluding partner-dependent PS.

Our compilation yielded a list of 307 experimentally validated 'Impact' mutation records from 70 proteins (Fig. S1A), including 79 that strengthened the PS properties (annotated as 'Strengthen') and 228 that weakened or disabled them (annotated as 'Weaken/Disable'). The PScore[27] and PhaSePred[35] scores indicate that the proteins from which these missense variants originate are predominantly proteins undergoing PS spontaneously (Fig. 2A). The PS propensity scores predicted for these proteins are significantly higher compared to those of the human proteome (Fig. 2A). Due to the limited availability of variants that have no PS effect, for each of the 70 proteins, we randomly generated 500 single amino acid (AA) variants, resulting in a total of 35,000 random 'Background' mutations, which were further used as negative samples (i.e., 'Background' mutations) for analyses.

Properties of variants impacting phase separation

We observed that mutations impacting PS (annotated as 'Impact' mutations) are predominantly located in intrinsically disordered regions (IDRs) rather than structured domain areas (Domains) (Fig. 2B). Existing variant effect prediction models primarily

rely on evolutionary features[18,19,21,22,43] and structural features[23,44-46]. However, these features are inadequate for representing variants in IDRs and their effects on PS. We found that current advanced pathogenicity prediction models, such as AlphaMissense[23] and EVE[22], both fail to distinguish 'Impact' mutations from background mutations (Fig. S1C). This underscores the urgent need for developing specialized models to predict the effects of mutations on PS.

Among these 'Impact' mutations, those involving serine, tyrosine, arginine, and glutamine are most prevalent (Fig. 2C). This aligns with prior findings that tyrosine and arginine govern the saturation concentration of PS, while glutamine and serine promote it[29].

Interestingly, our analysis revealed that 'Impact' mutations tend to occur near domain boundaries. Specifically, the amino acid distance from each mutation site to its closest domain boundary was calculated. We noted that, whether within Domains or IDRs, 'Impact' mutations are significantly closer to these boundaries compared to random mutation sites ($P <= 0.0001$, Fig. 2D). In addition, we assessed pi-contact values at mutation sites using the prediction function provided by PScore[27], as pi-pi interactions are vital in determining PS. For 6 out of the 8 predicted pi-contact values, 'Impact' mutations have higher predicted values than random mutation sites (Fig. 2E). This suggests that mutations affecting PS are likely to occur at sites with high pi-contact frequencies.

Next, we compared the changes in amino acid properties before and after mutation between the 'Strengthen' and 'Weaken/Disable' groups. We observed that missense mutations that strengthen PS propensity are prone to have higher amino acid mass ($P = 0.0097$), hydrophobicity ($P = 0.0024$), and decreased polarity ($P < 0.0001$) compared to 'Weaken/Disable' mutations (Fig. 1F). Notably, the increase in hydrophobicity supports the previous discovery that hydrophobicity serves as an important PS driver [30,32].

Together, these findings show the distinct properties of missense mutations impacting PS and the varying characteristics between mutations that either strengthen or weaken PS. Based on these properties, we have developed tools to predict the impact of mutations on PS, which will be discussed in the following section.

PSMutPred for predicting the impact of missense mutation on phase separation tendency

To predict the impact of missense mutations on protein PS properties, we developed PSMutPred composed of two machine-learning (ML) approaches. The first approach termed the 'Impact Prediction' (IP) task trained ML models to predict missense mutations that impact PS. In the second approach termed the 'Strengthen/Weaken Prediction' (SP) task, ML models were trained to predict the direction of the shifts in the normal cellular PS threshold induced by 'Impact' mutations (Methods). The process is depicted in Figure 1, indicated by the green section (Fig. 1).

To encode such changes caused by missense mutations for quantification and model learning, we considered the physicochemical properties of both the wild-type amino acid (AA) and mutant AA at each mutation site, as well as the properties of amino acids within IDRs. We applied pi-pi contact frequency[27] to encode the mutation site's underlying significance for PS. Additionally, features such as IUpred[47] score and mutation distance to domain boundary were used to encode the position of these mutations. Each mutation sample was converted into a 39-dimension feature vector (Methods). Neither MSA features nor structural features widely adopted in existing mutation-related machine learning prediction[19-22,43,44,48] were considered due to the nature of IDRs.

The dataset of missense mutations was divided into a cross-validation dataset (246 'Impact' samples and 23,500 'Background' samples from 47 proteins) and an independent test set (61 'Impact' samples and 11,500 'Background' samples from 23 proteins) (Methods). We trained prediction models including Logistic Regression (LR), Random Forest (RF), and Support Vector Regression (SVR) for both tasks to explore the discriminative power of predicting missense mutations' effect on PS.

Performance evaluation of PSMutPred

Although there is currently no computational algorithm for missense mutations on predicting the effect on PS, we attempted to determine if existing PS prediction methods are capable of discerning alterations in PS propensity caused by missense mutations.

Five high-performing PS methods were selected, including DeePhase[31], PSAP[24], PScore[27], catGRANULE[30], and FuzDrop[49]. We found that the prediction score differences of 'Impact' mutations were higher than those of 'Background' mutations, by comparing the absolute differences of prediction scores (Fig. 3A). However, the AUROCs were unsatisfactory (Fig 3B). Except for FuzDrop[49], none of the methods could differentiate between 'Strengthen' mutations and 'Weaken' mutations (Fig. S2 C and D), as their prediction scores did not accurately reflect the increase or decrease of PS propensity caused by mutations.

To test whether PSMutPred-IP can identify mutations that impact PS, we implemented leave-one-source-out cross-validation (LOSO CV) (Method). LOSO is used to evaluate the models' predictive capabilities for variants from unseen proteins, ensuring their generalizability. AUROCs and AUPRs on LOSO (Fig. 3 C and E) showed the accuracy of our models, especially IP-SVR and IP-RF, have better discriminative power than existing PS algorithms. Evaluation results on the independent test set showed that our models have stable prediction performance when facing mutations from different PS protein categories (Fig. 3C and E).

To ensure that the superior performance of the algorithm was not caused by the distribution bias of 'Impact' mutations, which tend to be located in IDRs (Fig. 2 B and D), we conducted an additional LOSO CV and an independent validation for the IP task. Here we generated 'Background' samples maintaining the same IDRs: Domains ratio as observed in the 'Impact' samples. LOSO CV results and independent validation results (Fig. 3 D and F) were as promising as those from the previous dataset (Fig. 3C and E), showing that our methods understand this type of mutation. Moreover, a small proportion of 'Impact' mutations are multi-point mutations (Fig. S1B), the proposed models still exhibited efficient predictive power and outperformed existing PS models when analyzing the results without considering multi-point mutations (Fig. S2 A and B).

We next investigated the performance of PSMutPred-SP models. The significant discriminative power to distinguish between 'Strengthen' samples and 'Weaken' mutation samples from unseen proteins under LOSO CV (SP-LR, $P < 0.0001$) (Fig. S2

E and F) and AUROCs on the independent test set (Fig. S2G) indicated that SP-LR, SP-RF models can identify the direction of the shifts in PS caused by missense mutations.

Feature importance indices were calculated for random forest models of both tasks and grouped into feature types to discover potentially key features (Fig. S2 G and H). In both tasks, the pi-contact frequency of the mutation site ranked first, suggesting that the pi-pi interaction at the mutation's specific location significantly influences the effect of mutations on protein phase separation. Beyond Pi-contact, no single feature was identified as a dominating one, suggesting the non-linear nature among features and a multi-factor causal relationship between features and the PS outcome.

Overall, PSMutPred models not only exhibit impressive performance in predicting missense mutations that impact PS but also predict the direction of the PS-threshold shift. Our final model generates predictions made by the IP-SVR, IP-RF, IP-LR, SP-LR, SP-RF models as well as their corresponding rank scores (Methods). These models can be employed as effective tools for assessing the tendency of missense mutations to affect PS, enhancing the interpretation of disease variants' pathogenicity.

Experimental validation on the novel PS-related mutations identified by PSMutPred

Aberrant phase separation (PS) or aggregation processes have been implicated in the pathogenesis of various diseases, including neurodegenerative disorders, autism[4], and hearing loss[7-9]. In this study, we employed PSMutPred to predict the PS impact ('Impact' prediction and 'Strengthen/Weaken' prediction) of missense variants from genes associated with PS-related diseases. Specifically, we selected *EPS8*, known for its association with deafness, and analyzed the PS impact of its 'Uncertain' missense mutations from the ClinVar[50,51] database to assess the accuracy of our prediction model.

Epidermal growth factor receptor pathway substrate 8 (*EPS8*) is a multifunctional protein involved in cell mitosis and differentiation[52-55], in capping proteins through side-binding, in bundling of actin filaments[56-58] as well as in the elongation of actin in hair cell stereocilia[52]. Prior research has shown that *EPS8* localizes to the tips of stereocilia and contributes to the formation of PS-mediated condensates at the

262 stereocilia tip complex[8,59]. These findings suggest that *EPS8* has the capacity for self-
263 phase separation and to interact with other molecules.

264 To validate our predictions, we selected missense mutations, including R265C,
265 D586G, and K676R, from the 8 candidate mutations. These candidates were predicted
266 by PSMutPred-IP to impact PS, either by strengthening or weakening it, with rank
267 scores above 0.5 across all models (IP-SVR, IP-RF, and IP-LR). Among these
268 candidates, D586G had the highest scores for both SP-LR and SP-RF, suggesting a
269 stronger propensity to enhance PS, while K676R scored the lowest for both metrics
270 implying an impairment of PS capacity. As negative controls, we selected R702W and
271 E728V from the 15 candidate mutations (all with IP rank scores below 0.5 for IP-LR,
272 IP-SVR, and IP-RF; at least one had a score below 0.1).

273 For experimental validation, we overexpressed the wild-type and mutants of mouse
274 *Eps8* (which is highly conserved with human *EPS8*, as shown in Fig. S3D) fused with
275 a GFP tag in HEK293 cells. Observations made using Olympus fluorescence
276 microscopy highlighted distinct changes in puncta formation. Specifically, R265C and
277 D585G exhibited a notable increase in the number of puncta, while K675R showed a
278 significant reduction (Fig. 4A and B, Fig. S3 A and B). Notably, D586G (equivalent to
279 D585G in mice) demonstrated enhanced PS capacity, aligning with its high SP-RF and
280 SP-LR scores. In contrast, K676R (equivalent to K675R in mice) showed diminished
281 PS ability, supported by its low SP-RF and SP-LR scores.

282 *Eps8* comprises six domains, encompassing both well-structured regions and
283 intrinsically disordered regions (IDRs) (Fig. S3C). D585 in mice (equivalent to D586
284 in humans) is situated in the SH3 domain. It forms stable hydrogen bonds with the
285 sidechains or backbones of K536 and K538 in mice (equivalent to K537 and K538 in
286 humans), maintaining structural stability through either a β-sheet-loop conformation in
287 mice or β-sheet-β-sheet conformation in humans (Fig. 4 D and E, Fig. S3 E and F).
288 Additionally, D585 and its neighboring residues, as well as interaction residues, exhibit
289 high conservation. However, the substitution of glycine for aspartate disrupts the
290 formation of these hydrogen bonds, leading to a destabilized structure. This change
291 might enhance PS by increasing flexibility and modifying the β-sheet-loop structure.

Conversely, K675 in mice (equivalent to K676 in humans) is located in an IDR at the C-terminus of *EPS8* (Fig. 4C, Fig. S3D). Sequence alignment underscores the high conservation of K675 and its adjacent residues (Fig. 4F, Fig. S3G). Lysine's characteristic long and hydrophobic side chain increases the likelihood of extending and 'capturing' polar residues from self-proteins or other proteins. This substitution may alter these interactions, potentially impairing PS.

In summary, our study demonstrates that the changes in PS propensity, as predicted by our model, accurately corresponds to a distinct number of puncta formed in live cells. This serves to underscore the utility and accuracy of PSMutPred.

Phase separation effect prediction of disease variants by PSMutPred

Liquid-liquid phase separation (LLPS) turns out to be a general mechanism for protein condensate assembly in various physiological processes. Mutations that change phase separation (PS) are likely causes of disease[10,11,14,60]. To exam this theory, we divided missense variants corresponding to 8,611 human genes from ClinVar[50,51] (522,016 variants in total, downloaded in 2022.12) into PS-prone and low-PS-prone propensity groups for between-group analyses.

By comparing the PS-prone group defined by 83 PS ClinVar proteins with the low-PS-prone group (other proteins), we found that in the PS-prone group, the PSMutPred-IP scores of variants (n = 1,451) were skewed to pathogenicity compared to those from low-PS-prone group (n = 86,291) (Fig. 5A). To make the sizes of the two groups more comparable and improve the efficiency of the comparison, we also defined the PS-prone group by collecting PS proteins predicted by PScore[27] and a meta-predictor PhaSePred[35] (1,276 proteins, 30,889 variants) and grouped variants from other proteins as low-PS-prone group (56,853 variants). The between-group analysis showed consistent results (Fig. 5B).

We found that the IP scores of variants within IDRs (n = 15,427) are more positively correlated with the pathogenicity of the variants than variants within Domains (n = 15,462) when analyzing PS-prone group defined by algorithms (Fig. 5C). Additionally,

321 we isolated variants likely in IDRs (unmapped by PfamScan[61,62], with IUpred3[47]score >
322 0.5) from the PS-prone group and directly using PhaSePred-IP scores to predict
323 pathogenicity ((Likely) pathogenic or (Likely) benign). Separate analyses were
324 conducted for the PS-prone group as defined by PS proteins and by algorithms (Fig. 5
325 E and F). The resulting AUROC scores indicated that PSMutPred-IP can identify
326 disease variants that lead to PS alterations.

327 Studies indicate that neurodegenerative lesions may be associated with excessive
328 PS[15,63]. Within the predicted PS-prone group, 19 proteins were identified to be highly
329 associated with neurodegenerative disease (Methods). SP prediction scores for the
330 variants (n = 252) of these proteins showed different patterns compared to other variants
331 (n = 19,266). Specifically, among these 252 variants, those predicted to strengthen PS
332 were more inclined towards pathogenicity compared to those predicted to weaken PS.
333 This observation was reflected by the positive Pearson coefficient values for both SP-
334 LR and SP-RF models, contrasting with the negative coefficient values for variants
335 from other PS proteins (Fig. 5D).

336 Additionally, we observed that mutations within IDRs of PS proteins are more likely
337 to affect PS than those in proteins less likely to undergo PS. This conclusion was based
338 on evaluating the proportion of variants of 'Uncertain Significance' (VUSs) predicted
339 to impact PS for each protein (see Methods). Using Kolmogorov's D statistic to
340 compare these proportions, we observed that the PS-prone group had a higher incidence
341 of PS-affecting variants than the low-PS-prone group, as shown in Figure 5G.

342 In summary, our comprehensive analyses revealed a notable clustering of pathogenic
343 missense mutations with an impact on PS propensity in proteins with inherently higher
344 PS propensity. Additionally, our findings indicate that compared to mutations that might
345 weaken PS, gain-of-PS mutations tend to aggregate specifically in disease mutations
346 associated with neurodegenerative-related genes. These results not only provide
347 valuable insights into the relationship between PS and disease but also underscore the
348 reliability and validity of PSMutPred.

PSMutPred improves pathogenicity prediction of disease variants

Computational pathogenicity models primarily rely on evolutionary features[18,19,21,22,43] and structural features[23,44-46] for missense variant pathogenicity predictions. However, IDRs, which often lack a fixed structure and display poor evolutionary conservation[2], challenge the effectiveness of these traditional features[4].To solve this, we explored PS as a potential supplementary feature for variant interpretation. We combined PS-related features with a pathogenicity prediction algorithm (Fig. 1, orange part), to predict missense variants' pathogenicity.

For each variant, we generated a feature vector combining PS-related features such as the scoring of the PSMutPred algorithm with the EVE[22] score (Methods). We chose EVE due to its unsupervised nature, which can offer an unbiased baseline, making integration and testing with our features straightforward. In contrast, combining with other supervised pathogenicity models needs retraining, adding to the risk of data leakage. We trained prediction models including Random Forest (RF), Support Vector Regression (SVR), and Logistic Regression (LR), using ClinVar[51] significances as labels. Their performance was evaluated using a blocked 3-fold cross-validation, aiming at testing the models' ability to predict variants from unseen proteins. Moreover, an independent test set was also used (Methods). All models demonstrated improved pathogenicity prediction, with RF showing the best performance in terms of AUROC and AUPR scores (Fig. S4 A, B, and Table. S1). The inclusion of PS-related features notably enhanced prediction accuracy. Based on these results, RF was selected as the final model.

We investigated the effect of including PSMutPred scores within various feature combinations. RF models were trained on various random feature sets, and their performance on the test set was quantified by AUROC and AUPR (Fig. 6E, Fig. S4C). The comparative analysis highlighted a significant enhancement in the prediction accuracy when PSMutPred scores were incorporated. Additionally, based on the feature importance generated by the final RF model, we found that the probability of the mutation site within IDRs (IUpred[47] Score) and the residue distance of the mutation site

to its' closest domain boundary were identified as two key features to improve pathogenicity prediction, and PSMutPred scores also played important roles in model prediction (Fig. S4D).

Given that IDRs typically exhibit poorer evolutionary conservation than Domains[4], it is unsurprising that we found evolutionary information[22] to be less effective in predicting IDR variants compared to Domain variants (Fig. 6 A and B, Table. S1). Furthermore, since IDRs are closely linked to PS, the incorporation of PS-related features into prediction algorithms led to a more pronounced improvement in identifying IDR disease mutations than those within Domains. Specifically, the AUROC of IDR variants in the independent test set improved by 4.3%, compared to a 2.6% improvement for Domain variants (Fig. 6A, Table. S1). Similarly, the AUPR for IDR variants rose from 0.802 to 0.878, while there was only a 1.7% improvement in AUPR for Domain variants (Fig. 6B, Table. S1). The Kolmogorov–Smirnov test further indicated a significant improvement in the prediction of both pathogenic variants as well as benign variants in IDRs (Fig. 6 C and D).

Using the PS features, combined with EVE scores, we predicted pathogenicity scores for 358,504 ClinVar[51] variants of uncertain significance. Among them, 179,785 are IDR variants (Fig. S4F), 13.3% were predicted pathogenic, and 55.2% were predicted benign (Fig. 6F).

Our findings reveal that the effect of missense mutations on PS serves as a valuable encoding for IDR mutations, thereby complementing evolutionary features in deciphering the effects of mutations. This enhances our understanding of mutational pathogenicity and provides valuable insights in this context.

Discussion

Missense mutations in IDRs are overlooked compared to those in folded regions. Pathogenicity prediction models, such as AlphaMissense[23], typically rely on evolutionary information and protein structural features. However, since IDRs often lack a consistent structure and show limited evolutionary conservation, using these traditional features to encode variants in IDRs turns out to be unsuitable, making

interpreting the numerous VUSs in IDRs difficult[2]. To address the challenge, we turn to phase separation (PS), a widely acknowledged property of IDRs linked to various diseases. As recent studies emphasize, mutations within IDRs can disrupt regular PS, which can lead to disease loss[10,11], highlighting its potential as a predictive feature.

Although a comprehensive understanding of the specific mechanisms by which missense mutations alter PS remains elusive[4], we considered differences in physicochemical properties, viewing them as features that quantify mutation-induced changes. Additionally, we used the predicted pi-pi contact frequency[27] at the mutation site as a position-specific encoding. These considerations were used to develop models predicting the effect of missense mutation on PS. Separate models were constructed to predict the 'Impact' and 'Strengthen/Weaken' effects of missense variants. The reliability of the model predictions was validated through prediction results on variants from unseen proteins and further experimental validations. Moreover, subsequent analyses conducted on a larger-scale dataset further affirmed the accuracy and effectiveness of our models.

It is crucial to note that, compared to in-frame insertions or deletions, the occurrence of a single or limited number of missense mutations poses a greater challenge in causing significant shifts in normal cellular PS[4]. Given this, we utilized experimentally validated missense mutations that alter PS propensity in proteins naturally undergoing this process, mainly involving single amino acid variants, to develop machine learning models. The strong performance of our algorithms indicates an inherent pattern among missense variants impacting PS, allowing for their prediction. Including more high-quality, experimentally validated data, especially regarding the effects of missense mutations on PS-Part protein[35], and integrating additional PS-related features like electrostatic interactions[31,33] and phosphorylation at mutation sites[35] could both refine the prediction of mutation impact on PS.

Through analysis of PSMutPred scores using ClinVar[51] variants, we discovered that proteins prone to PS have a higher proportion of disease variants altering PS, compared to proteins less likely to undergo this process. Among these mutations, those likely to enhance PS were proportionally more prevalent in neurodegenerative-disease-related

genes than those that weaken PS, which supports the current discoveries[15,63]. Therefore, we can confidently assert that PSMutPred scores serve as valuable encodings for assessing the potential impact of missense mutations on PS, making them applicable to a broader range of research.

We further combined PS-related features, including PSMutPred scores, with evolutionary scores[22] to predict pathogenicity and assessed the performance using ClinVar labels. We didn't build a new model from scratch; instead, we highlighted the contribution of PS to improving the accuracy of pathogenicity predictions, offering a fresh perspective in this field. Looking ahead, leveraging diverse data such as weak labels from population frequency and variants from HGMD[64] and gnomAD[65] may further refine these predictions. Additionally, advanced machine learning methods, including deep learning, hold the potential to boost accuracy.

Ultimately, our study not only presents methods to predict the effects of missense mutations on PS but also contributes to the prediction and improved understanding of Variants of Uncertain Significance (VUS) occurring in IDRs, enhancing the diagnostic accuracy for rare genetic disorders.

Methods

Data acquisition

Experimentally validated missense mutations that impact phase separation (PS) and their corresponding experimental sequences were curated from the PhaSepDB[38] and LLPSDB[39,40] databases up to 2022.11. We retained single amino acid mutation samples as well as multi amino acid mutations with a number of mutation sites less than or equal to five. We got 214 single amino acid mutation samples, 43 two-mutation-sites-samples, 26 three-mutation-sites-samples, 18 four-mutation-sites-samples and 6 five-mutation-sites-samples (Fig. S1B). We finally obtained 307 samples corresponding to 70 proteins including 79 'Strengthen' samples that strengthened PS, and 228 'Weaken/Disable' samples that weakened or diminished PS.

As a limited number of experimentally validated mutations that do not affect PS are

available for reference, we created ‘Background’ samples by generating random single amino acid (AA) missense mutations for each of the 70 proteins. Mutations already included in the ‘Impact’ samples were excluded. For each gene, we selected 500 random mutations, leading to a total of 35,000 mutations. It should be noted that this does not imply that background mutations cannot alter PS under any conditions, instead, it allowed us to learn missense mutations that impact PS.

Statistical analysis of phase separation-related missense mutations

We used PfamScan[61,62] to predict the structural domain regions, abbreviated as ‘Domains’, within the experimental sequences. Any segment of the sequence not identified as a domain by PfamScan was annotated as an IDR. Multi-amino acid (AA) mutation samples were broken down into single AA mutations for subsequent statistical analysis.

We estimated the pi-contact frequency at the mutation site using the pi-pi interaction prediction function in PScore[27]. These values correspond to 8 sp^2 groups. The categorization process divides pi-contacts based on: 1) short-range (4 sequence separation) vs. long-range (>4), 2) sidechain vs. backbone, and 3) absolute predicted frequency vs. relative difference from sp2 groups with the same identity respectively (See PScore[27] for more details).

To analyze the changes in the amino acid properties pre and post-mutation, we calculated the index differences for properties including hydrophobicity[66], polarity[67], and mass[68]. The Kolmogorov–Smirnov test was then applied to compare these index differences between the ‘Strengthen’ and ‘Weaken/Disable’ mutation groups.

Development of PSMutPred

Training and testing dataset

We divided the mutation samples into a cross-validation dataset (47 human proteins) and an independent test set (23 non-human proteins). 'Non-human' in this context specifically refers to experimental protein sequences originating from species other than humans.

In the case of the 'Impact Prediction' (IP) models (‘IP task’ in short), we grouped

‘Strengthen’ and ‘Weaken/Disable’ mutations into one group and set their labels to 1, while the labels of ‘Background’ mutations were set to 0. For the IP task, we got 246 ‘Impact’ mutation samples, and 23,500 ‘Background’ samples in the cross-validation dataset, 61 ‘Impact’ samples, and 11,500 ‘Background’ samples in the independent test set. In the case of the 'Strengthen/Weaken Prediction' (SP) models (‘SP task’ in short), labels of ‘Strengthen’ mutations were set to 1, and labels of ‘Weaken/Disable’ mutations were set to 0. For the SP task, we got 174 ‘Weaken/Disable’ samples and 72 ‘Strengthen’ samples in the cross-validation dataset; 7 ‘Strengthen’ samples, and 54 ‘Weaken/Disable’ samples in the independent test set.

Machine learning features

We used a set of simple features (39 dimensions) to encode each mutation entry. The two tasks share a common set of feature encoding. For multi-point mutations, certain features were adjusted to account for the presence of multiple mutation sites. The details of the features for each sample are described below:

IDR-related features of the mutation site(s) (5 dimensions): The structural domains of the experimental sequences corresponding to the sample were predicted using PfamScan[61]. Segments not predicted as Domains were designated as IDRs. 1) One binary value indicating whether the site is located in IDRs. For multi-point mutation, this value is set to 0 if any of the mutation sites are located within Domains; 2) One value quantifies the residue distance of the mutation site to its nearest domain boundary. For multi-point mutation, this value reflects the average distance of all mutation sites; 3) One value representing the likelihood of the mutation site being within IDRs predicted by IUpred3[47], For multi-point mutations, we use the average of the predicted values for all mutation sites; 4) One value representing the average of the full sequence IUpred3[47] scores; 5) One binary value indicating whether the protein sequence has IDRs or not.

The predicted pi-contact frequency at the mutation site (8 dimensions): We derived 8 predicted values for pi-pi interactions at the mutation site from the wild-type sequence using the pi-pi interaction prediction function in PScore[27] (See PScore[27] for more details). In the case of multi-point mutation, we computed the 8 values as the mean

525 scores across all mutation positions

526 Physicochemical feature encoding for the mutation site (26 dimensions): We targeted
527 the changes in AA characteristics at the mutation site whenever possible by employing
528 five physicochemical indices that were previously used for protein feature encoding[69-
529 72]. We selected hydrophobicity[73], polarity[74], volumes of side chains (VSC) of amino
530 acids[75], solvent-accessible surface area (SASA)[76], and net charge index (NCI) of side
531 chains of amino acids[69,77]. For each index, we computed four feature values: 1) Index
532 value for the wild-type AA at the mutation site. For multi-point mutation, we used the
533 mean value; 2) Numerical difference in index value between mutant and wild-type AA,
534 we summed the difference values for multi-point mutations; 3) Average index values
535 for all AAs in the wild-type sequence' IDRs (defined as areas outside of Domains
536 predicted by PfamScan[61,62]), and 4) the difference between the index value of mutant
537 AA and the average index value (from point 3), for multi-point mutation, we
538 accumulated the difference values for each mutant AA. Additionally, six binary values
539 (6 dimensions) were used to capture the presence of three properties (positively charged,
540 negatively charged, and hybridized) in both the mutant AA and wild-type AA. For
541 multi-point mutations, an element is set to 1 if at least one AA exhibits the respective
542 property.

543 **Machine learning algorithm and model performance evaluation**

544 For both the IP task and the SP task, to evaluate the generalizability of our models on
545 variants from unseen proteins, we implemented a blind test called 'leave-one-source-
546 out cross-validation' (LOSO CV). In this approach, for each validation iteration, we
547 held out variants from a single protein from the total set of proteins (variants from cross-
548 validation dataset; 47 proteins), these variants were reserved solely for model
549 evaluation, while variants from the other proteins were used for model training. This
550 process was repeated 100 times for both the IP task, and SP task. Especially for the IP
551 task, we randomly selected 'Background' mutations twice the size of the positive
552 sample from the negative sample during each iteration. We added existing PS predictors
553 including DeePhase[31], PSAP[24], PScore[27], catGRANULE[30] as well as FuzDrop[49] for
554 comparison in the IP task. For FuzDrop[49], we applied the residue-level PS scores for

calculation. Expecting greater score changes for 'Impact' mutations than 'Background' mutations, we calculated the absolute difference in scores before and after each mutation to represent the predicted values.

We applied the Python Scikit learn (sklearn) package to develop machine learning models. Data was normalized by the MinMaxScaler from sklearn before fitting the models. We trained models including a RandomForestClassifier (with parameters: class_weight = 'balanced', and max_depth = 10), an SVR (with default parameters), and a LogisticRregression (with parameters: class_weight = 'balanced', kernel = 'liblinear', and penalty = 'l1') for both the IP and the SP task. We calculated the area under the curve of the receiving operating characteristics (AUROC) with the sklearn roc_auc_score, and roc_curve metric to evaluate the performance. Additionally, the precision and recall were calculated with sklearn precision_recall_curve with the area under the curve of the precision-recall curve (PRC) (AUPR) using precision_recall_curve. The discriminative power of models was evaluated using the Mann-Whitely test.

The cross-validation dataset and independent test set were combined to train the final PSMutPred models for both tasks. According to the models' performance, we produced IP-RF, IP-SVR, and IP-LR scores and SP-RF, SP-LR scores. To prevent the overuse of negative samples in the IP task, for each model, 10 different subsets of samples were randomly sampled from the 'Background' mutations with twice the size of the collected 'Impact' mutations to train 10 sub-models. The averaged prediction scores of the 10 trained models as well as the ranks for these scores in all ClinVar[50,51] variants were used as the final prediction scores for each IP-RF, IP-SVR, and IP-LR model. These models were used to predict PSMutPred scores for each missense variant in ClinVar.

Between-group analysis

Missense mutation data was collected from ClinVar up to 2022.12, which contains 522,016 variants corresponding to 8,611 human proteins. Of these, 39,166 are pathogenic or likely pathogenic, 48,576 are benign or likely benign and the remainder 434,274 are 'Uncertain Significance'.

After excluding the training data points for PSMutPred, we divided missense variants corresponding to 8,611 proteins into a PS-prone group and a low-PS propensity group in two ways: 1. By grouping variants from experiment-verified PS protein to the PS-prone group and the rest to the low-PS-prone group. Specifically, we referenced 155 human PS proteins (59 PS-Self proteins, 96 PS-Part proteins) from PhaSepDB[35,38], among which 83 were found in ClinVar[50]. This resulted in 1,451 variants in the PS-prone group, leaving 86,291 variants for the low-PS-prone group. 2. Variants from predicted PS proteins, based on either a PScore[27] higher than 4 or a PhaSePred[35] rank (PdPS-10fea_rnk) higher than 0.9, were allocated to the PS-prone group. This resulted in 30,889 variants from 1,276 proteins in the PS-prone group and leaving 56,853 variants for the low-PS-prone group.

We first performed analysis excluding variants with 'Uncertain significance'. For each group, the Pearson correlation score was computed using the Scipy package by comparing the PSMutPred scores of its variants against their respective ClinVar pathogenicity labels (pathogenic or likely pathogenic coded as 1 and benign or likely benign coded as 0). For SP-RF and SP-LR, we only considered variants being predicted 'Impact' by PSMutPred-IP (all IP-RF, IP-SVR, and IP-LR above 0.5). The significance of each computed Pearson value is annotated with corresponding P-values, indicating the robustness of the correlation within each group. We compared the coefficient values between the PS-prone group and the low-PS-prone group. We also computed and compared the coefficient values between IDR variants (n = 15,427) and Domain variants (n = 15,462) within the PS-prone group defined by predictions (a variant was classified by Domain variants if identified in a structured domain by PfamScan[61,62] and as an IDR variant if it is unmapped).

In addition, we divided the PS-prone group into a neurodegenerative disease (ND) group and those that are not into a non-ND group. To do that, we collected disease genes related to common neurodegenerative diseases including Alzheimer's disease, Amyotrophic lateral sclerosis, Frontotemporal dementia, Huntington's disease, Multiple Sclerosis, and Parkinson's disease from DisGeNet[78]. After removing duplicates, we screened genes with a gene-disease association score (GDA score)

higher than 0.5, resulting in a subset of 91 genes, 19 of which were both matched in ClinVar genes and the PS-prone group (363 variants, 252 being predicted 'Impact' mutation by PSMutPred-IP). We compared the coefficient values between variants corresponding to the ND genes and those that were not within the predicted PS-prone group.

We then performed between-group analysis to test whether proteins that have a higher propensity to undergo PS have a higher proportion of variants that are predicted to impact PS. For each PSMutPred-IP model (IP-RF, IP-LR, IP-SVR), we calculated the proportion of IDR variants of 'Uncertain Significance' (VUSs) predicted as 'Impact' (PSMutPred-IP rank score > 0.8) for each protein. The proportion values of the PS-prone group and the low-PS-prone group were compared. We compared the PS-prone group (1,276 proteins) with the low-PS-prone group (7,335 proteins) defined by algorithm prediction and compared the PS-prone group (83 proteins) with the low-PS-prone group (8,528 proteins) defined by experimental-verified PS proteins.

Pathogenicity prediction model for missense variant

Dataset acquisition

EVE[22] predicts the pathogenicity of variants, covering variants of 3,219 disease-associated human genes. We collected EVE scores for all possible missense variants corresponding to 3,219 disease-associated genes. Among them, 47,870 ClinVar[50] variants (by annotations up to 2021.4.) were used to evaluate the EVE model, 30,758 pathogenic or likely pathogenic were marked as positive samples and 17,112 'Benign' or 'Likely benign' variants were marked as negative samples. We applied these samples for model cross-validation (cross-validation set). To construct an independent test dataset, we screened new variants (ClinVar[50] annotations up to 2022.12) that were not used to evaluate the EVE model and removed variants in the same amino acid position seen in the cross-validation set to avoid data leakage. This resulted in 15,394 variants including 8,709 pathogenic or likely pathogenic variants and 6,685 benign or likely benign variants.

Pathogenicity prediction combining PS-related features

To test whether the integration of PS-related properties can improve the accuracy of pathogenicity prediction. We built a model based on the following features: PSMutPred-IP model scores predicting the propensity of the missense mutation to impact PS (3 dim, predicted values of IP-LR, IP-RF, and IP-SVR); PSMutPred-SP model scores predicting the direction of the shift in the normal PS threshold induced by the mutation (2 dim, SP-LR, SP-RF); PScore[27] representing the PS tendency of the wild-type protein (1 dim); IUpred[47] score of the mutation site representing the probability of the site to locate in IDRs (1 dim); binary encoding whether the mutation is in a Domain predicted by PfamScan[61,62] (1 dim); the residue distance of the mutation site to the nearest IDR (1 dim); one-hot encoding for wild-type and mutant aa (40 dim); Evolutionary score computed by EVE of the variant (1 dim). We trained models including a RandomForestClassifier (with parameters: n_estimators = 200, class_weight = 'balanced', and max_depth = 15), an SVR (with default parameters), and a LogisticRregression (with parameters: class_weight = 'balanced', kernel = 'liblinear', and penalty = 'l1'). The dataset was standardized using the MinMaxScaler from Scikit Learn before fitting each model. Based on the performance, we selected the RandomForestClassifier as the optimal model.

Model performance benchmarks

We first evaluated our models on the cross-validation set using a blocked 3-fold cross-validation where we strictly separate variants from the same gene into the same group. We then trained the models on the cross-validation set and evaluated the models using the independent test set. For both 3-fold cross-validation and validation on the independent test set, AUROC, AUPR, and Pearson correlation coefficient were computed to evaluate the performance.

The data points were combined to train a final pathogenicity predictor. And to transform the continuous pathogenicity scores into 'Pathogenic', 'Likely benign', and 'Uncertain' categories, we determined 0.55 as the initial threshold to distinguish between 'Likely pathogenic' and 'Likely benign' based on the test set's F1 score. To define the uncertainty of prediction, we considered applying an offset around this

threshold. For instance, with a 0.1 offset, predictions below 0.45 are categorized as 'Likely benign', while those above 0.65 are classified as 'Likely pathogenic'. We observed that starting from the initial 0.55 threshold, the AUROC and Accuracy improved as the offset increased (Fig. S4E). At an offset of 0.15, the AUROC reached 0.94, and at 0.25, it climbed to 0.95. Considering balancing performance and data size, we settled on 0.15 as the offset, marking predictions between 0.4 and 0.7 as 'Uncertain'. And we got 23,857 variants predicted as 'Likely pathogenic', 56,220 variants predicted as 'Uncertain', and 98,642 variants predicted as 'Likely benign' for 178,719 IDR variants among 358,504 ClinVar[50] variants of 'Uncertain significance' (Fig. 5G).

Experimental Materials

Plasmid preparation and cell culture

The full-length coding sequences of mouse *Eps8* (NM_007945.4) and its mutations were PCR amplified and cloned into the pEGFP-C3 vector. These recombinant plasmids were then transiently transfected into HEK293 cells using the Lipofectamine 3000 Kit (Invitrogen), with 1-2 μg of plasmid used per well in a 12-well plate (Costar, Corning) for each transfection. Before fixation, HEK293 cells were cultured for 16-24 hours in DMEM medium (Gibco) supplemented with 10% FBS (Gibco) and 1% Penicillin-Streptomycin Solution (Gibco), under 5% CO2 conditions. Cells were washed 3 times with PBS for 3 minutes each, then fixed in 4% PFA for 20 minutes at room temperature. Detailed validation information and references for these products can be found on the respective manufacturers' websites.

Image analysis and quantification

HEK293 cells were visualized using an IX73 Inverted Fluorescence Microscope (Olympus) equipped with a 60× 1.42 NA Plan Apochromat oil objective. Images were captured with an Iris 9 sCMOS camera (Teledyne Photometrics) and the setup was controlled by the cellSens imaging system (Olympus). After subtracting the background and setting an identical fluorescence threshold, ImageJ software was applied to quantify the number of puncta per cell, with modules including 'Analysis Particles', 'Find Maxima', and 'Set Measurement'. The statistical analysis applied a two-tailed Student's

t-test for experimental comparisons of puncta numbers using GraphPad Prism 9. Each assay was performed at least three times.

Protein structure analysis of *Eps8*

We performed the sequence alignment of *Eps8* with the ClustalX tool. The structures of human and mouse *Eps8* were predicted using AlphaFold2[79]. The interaction among residues was analyzed using the crystal structure of the SH3 domain of *Eps8* (PDB: 7TZK). The protein structure analysis was performed using PyMol version 2.5.0.

References

1. Vacic, V. & Iakoucheva, L.M. Disease mutations in disordered regions--exception to the rule? *Mol Biosyst* **8**, 27-32 (2012).
2. Colak, R. *et al.* Distinct types of disorder in the human proteome: functional implications for alternative splicing. *PLoS Comput Biol* **9**, e1003030 (2013).
3. Alberti, S. Phase separation in biology. *Current Biology* **27**, R1097-R1102 (2017).
4. Tsang, B., Pritisanac, I., Scherer, S.W., Moses, A.M. & Forman-Kay, J.D. Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations. *Cell* **183**, 1742-1756 (2020).
5. Edward, G. & James, S. The molecular language of membraneless organelles. *Journal of Biological Chemistry* **294**, 7115-7127 (2019).
6. Banani, S.F., Lee, H.O., Hyman, A.A. & Rosen, M.K. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol* **18**, 285-298 (2017).
7. Wang, H. *et al.* Temporal and spatial assembly of inner ear hair cell ankle link condensate through phase separation. *Nat Commun* **14**, 1657 (2023).
8. Lin, L. *et al.* Phase separation-mediated condensation of Whirlin-Myo15-Eps8 stereocilia tip complex. *Cell Rep* **34**, 108770 (2021).
9. He, Y., Li, J. & Zhang, M. Myosin VII, USH1C, and ANKS4B or USH1G Together Form Condensed Molecular Assembly via Liquid-Liquid Phase Separation. *Cell Rep* **29**, 974-986.e4 (2019).
10. Niaki, A.G. *et al.* Loss of Dynamic RNA Interaction and Aberrant Phase Separation Induced by Two Distinct Types of ALS/FTD-Linked FUS Mutations. *Molecular Cell* **77**, 82-94.e4 (2020).
11. Hofweber, M. *et al.* Phase Separation of FUS Is Suppressed by Its Nuclear Import Receptor and Arginine Methylation. *Cell* **173**, 706-719.e13 (2018).
12. Patel, A. *et al.* A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell* **162**, 1066-1077 (2015).
13. Murakami, T. *et al.* ALS/FTD Mutation-Induced Phase Transition of FUS Liquid Droplets and Reversible Hydrogels into Irreversible Hydrogels Impairs RNP Granule Function. *Neuron* **88**, 678-690 (2015).
14. Molliex, A. *et al.* Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123-33 (2015).
15. Silva, J.L. *et al.* Targeting Biomolecular Condensation and Protein Aggregation against Cancer.

Chem Rev **123**, 9094-9138 (2023).
16. Xiang, J. *et al.* Development of an α-synuclein positron emission tomography tracer for imaging synucleinopathies. *Cell* **186**, 3350-3367.e19 (2023).
17. Fan, Y. *et al.* Generic amyloid fibrillation of TMEM106B in patient with Parkinson's disease dementia and normal elders. *Cell Res* **32**, 585-588 (2022).
18. Brandes, N., Goldman, G., Wang, C.H., Ye, C.J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet* **55**, 1512-1522 (2023).
19. 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).
20. Fariselli, P., Martelli, P.L., Savojardo, C. & Casadio, R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics* **31**, 2816-21 (2015).
21. 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).
22. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91-95 (2021).
23. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
24. van Mierlo, G. *et al.* Predicting protein condensate formation using machine learning. *Cell Rep* **34**, 108705 (2021).
25. 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).
26. Chong, P.A., Vernon, R.M. & Forman-Kay, J.D. RGG/RG Motif Regions in RNA Binding and Phase Separation. *J Mol Biol* **430**, 4650-4665 (2018).
27. Vernon, R.M. *et al.* Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife* **7**(2018).
28. Martin, E.W. *et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science (New York, N.Y.)* **367**, 694-699 (2021).
29. 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).
30. Bolognesi, B. *et al.* A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity upon Increased Protein Expression. *Cell Rep* **16**, 222-231 (2016).
31. 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**(2021).
32. Dignon, G.L., Best, R.B. & Mittal, J. Biomolecular Phase Separation: From Molecular Driving Forces to Macroscopic Properties. *Annual Review of Physical Chemistry* **71**, 53-75 (2020).
33. Brangwynne, C.P., Tompa, P. & Pappu, R.V. Polymer physics of intracellular phase transitions. *Nature Physics* **11**, 899-904 (2015).
34. Martin, E.W. & Mittag, T. Relationship of Sequence and Phase Separation in Protein Low-Complexity Regions. *Biochemistry* **57**, 2478−2487 (2018).
35. 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).
36. Shen, B. *et al.* Computational Screening of Phase-separating Proteins. *Genomics Proteomics*

Bioinformatics **19**, 13-24 (2021).

37. Monahan, Z. *et al.* Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *Embo j* **36**, 2951-2967 (2017).
38. You, K. *et al.* PhaSepDB: a database of liquid–liquid phase separation related proteins. *Nucleic Acids Research* **48**, D354-D359 (2019).
39. Li, Q. *et al.* LLPSDB: a database of proteins undergoing liquid-liquid phase separation in vitro. *Nucleic Acids Res* **48**, D320-d327 (2020).
40. Wang, X. *et al.* LLPSDB v2.0: an updated database of proteins undergoing liquid-liquid phase separation in vitro. *Bioinformatics* **38**, 2010-2014 (2022).
41. Yamasaki, A. *et al.* Liquidity Is a Critical Determinant for Selective Autophagy of Protein Condensates. *Mol Cell* **77**, 1163-1175.e9 (2020).
42. Koehler, L.C. *et al.* TDP-43 Oligomerization and Phase Separation Properties Are Necessary for Autoregulation. *Front Neurosci* **16**, 818655 (2022).
43. 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).
44. Quan, L., Lv, Q. & Zhang, Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics* **32**, 2936-46 (2016).
45. Geng, C., Vangone, A., Folkers, G.E., Xue, L.C. & Bonvin, A. iSEE: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins* **87**, 110-119 (2019).
46. Iqbal, S. *et al.* Assessing the performance of computational predictors for estimating protein stability changes upon missense mutations. *Brief Bioinform* **22**, bbab184 (2021).
47. 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).
48. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).
49. 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).
50. Landrum, M.J. & Kattman, B.L. ClinVar at five years: Delivering on the promise. *Hum Mutat* **39**, 1623-1630 (2018).
51. Landrum, M.J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res* **48**, D835-d844 (2020).
52. Manor, U. *et al.* Regulation of stereocilia length by myosin XVa and whirlin depends on the actin-regulatory protein Eps8. *Curr Biol* **21**, 167-72 (2011).
53. Frittoli, E. *et al.* The signaling adaptor Eps8 is an essential actin capping protein for dendritic cell migration. *Immunity* **35**, 388-99 (2011).
54. Yap, L.F. *et al.* Upregulation of Eps8 in oral squamous cell carcinoma promotes cell migration and invasion through integrin-dependent Rac1 activation. *Oncogene* **28**, 2524-34 (2009).
55. Menna, E. *et al.* Eps8 regulates axonal filopodia in hippocampal neurons in response to brain-derived neurotrophic factor (BDNF). *PLoS Biol* **7**, e1000138 (2009).
56. Hertzog, M. *et al.* Molecular basis for the dual function of Eps8 on actin dynamics: bundling and capping. *PLoS Biol* **8**, e1000387 (2010).

57. Disanza, A. *et al.* Regulation of cell shape by Cdc42 is mediated by the synergic actin-bundling activity of the Eps8-IRSp53 complex. *Nat Cell Biol* **8**, 1337-47 (2006).
58. Disanza, A. *et al.* Eps8 controls actin-based motility by capping the barbed ends of actin filaments. *Nat Cell Biol* **6**, 1180-8 (2004).
59. Shi, Y., Lin, L., Wang, C. & Zhu, J. Promotion of row 1-specific tip complex condensates by Gpsm2-Gαi provides insights into row identity of the tallest stereocilia. *Sci Adv* **8**, eabn4556 (2022).
60. Sanders, D.W. *et al.* Competing Protein-RNA Interaction Networks Control Multiphase Intracellular Organization. *Cell* **181**, 306-324.e28 (2020).
61. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).
62. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412-d419 (2021).
63. Zbinden, A., Pérez-Berlanga, M., De Rossi, P. & Polymenidou, M. Phase Separation and Neurodegenerative Diseases: A Disturbance in the Force. *Dev Cell* **55**, 45-68 (2020).
64. Stenson, P.D. *et al.* The Human Gene Mutation Database (HGMD(®)): optimizing its use in a clinical diagnostic or research setting. *Hum Genet* **139**, 1197-1207 (2020).
65. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
66. Kyte, J. & Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105-32 (1982).
67. Zimmerman, J.M., Eliezer, N. & Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol* **21**, 170-201 (1968).
68. Nelson, D.L. & Cox, M.M. *Lehninger Principles of Biochemistry*, (Lehninger principles of biochemistry., 2005).
69. Guo, Y., Yu, L., Wen, Z. & Li, M. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res* **36**, 3025-30 (2008).
70. Li, X. *et al.* SDNN-PPI: self-attention with deep neural network effect on protein-protein interaction prediction. *BMC Genomics* **23**, 474 (2022).
71. Guo, Z. *et al.* 3D genome assisted protein–protein interaction prediction. *Future Generation Computer Systems* **137**, 87-96 (2022).
72. You, Z.H., Lei, Y.K., Zhu, L., Xia, J. & Wang, B. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinformatics* **14 Suppl 8**, S10 (2013).
73. Tanford, C. Contribution of Hydrophobic Interactions to the Stability of the Globular Conformation of Proteins. *J.am.chem.soc* **84**, 4240-4247 (1962).
74. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-4 (1974).
75. Pontius, J., Richelle, J. & Wodak, S.J. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J Mol Biol* **264**, 121-36 (1996).
76. Lee, B. & Richards, F.M. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**, 379-400 (1971).
77. Zhou, P., Tian, F.F., Li, B., Wu, S.R. & Li, Z.L. Genetic algorithm-based virtual screening of combinative mode for peptide/protein. *Acta Chimica Sinica* **64**, 691-697 (2006).

78. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* **48**, D845-d855 (2020).
79. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

Acknowledgement

This project is supported by the National Key Research and Development Program (2022YFE0125300), Innovation Program of Shanghai Municipal Education Commission (2023ZKZD16), the National Natural Science Foundation of China (82071262, 32300464, 81671326, 19Z103150073), Natural Science Foundation of Shanghai (20ZR1427200, 20511101900, 21ZR1433000), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01, 20JC1418600), the Shanghai Leading Academic Discipline Project (B205), China Postdoctoral Science Foundation (2023M732266) and Shanghai Jiao Tong University STAR Grant (YG2023ZD26, YG2022ZD024, YG2022QN111, YG2021QN135).

Author contributions

QL, GH, YS, and MF conceived the concept; MF, XW, QL, and YS designed experiments; QL led the project with assistance from MF, GH, YS, and XW; XW, MX, and QL collected samples; MF, XW performed experiments with assistance from QL, GH, YS, and LL; MF, XW and QL analyzed data; MF, XW, XZ, QL, and YS wrote the manuscript with input from all co-authors.

Competing Interests Statement

The authors declare no competing interests

Figures

Figure1

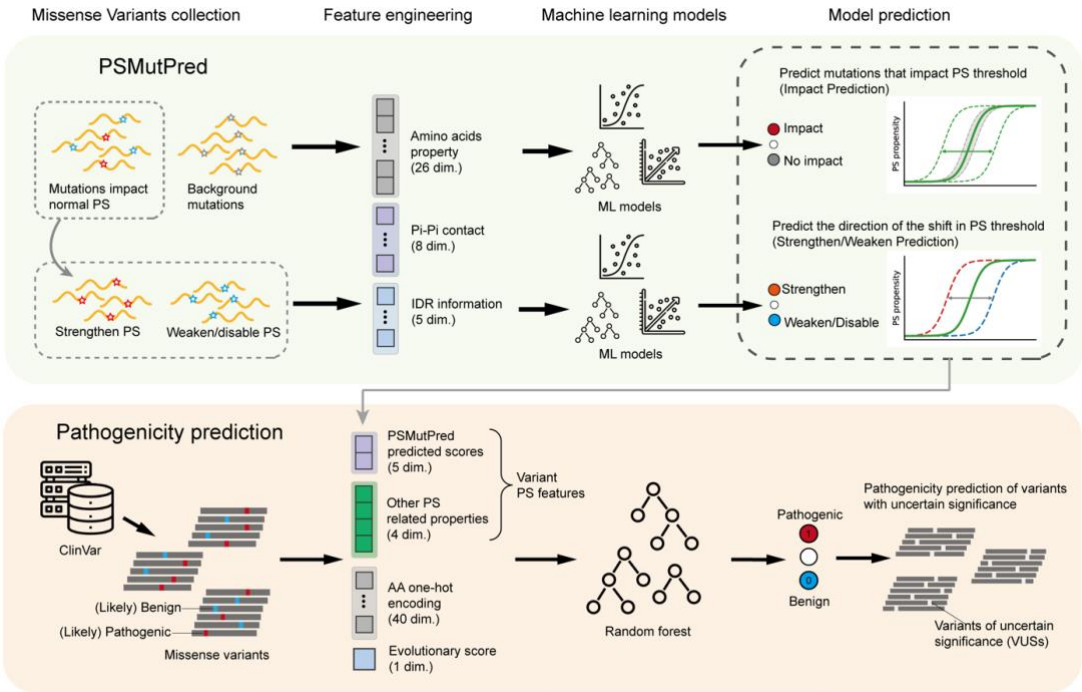


Fig. 1. Depiction of PSMutPred, a machine-learning approach designed to predict the effect of missense mutations on natural phase separation (PS) (upper green panel). Each mutation is converted into a 39-dimensional feature vector. Distinct models were employed for two main tasks: Identifying mutations that impact PS (termed 'Impact Prediction') and determining whether a mutation strengthens or weakens the PS threshold (labeled as 'Strengthen/Weaken Prediction'). Additionally, the pathogenicity of missense variants is predicted by integrating PS features, including the output from PSMutPred (lower orange panel).

Figure2

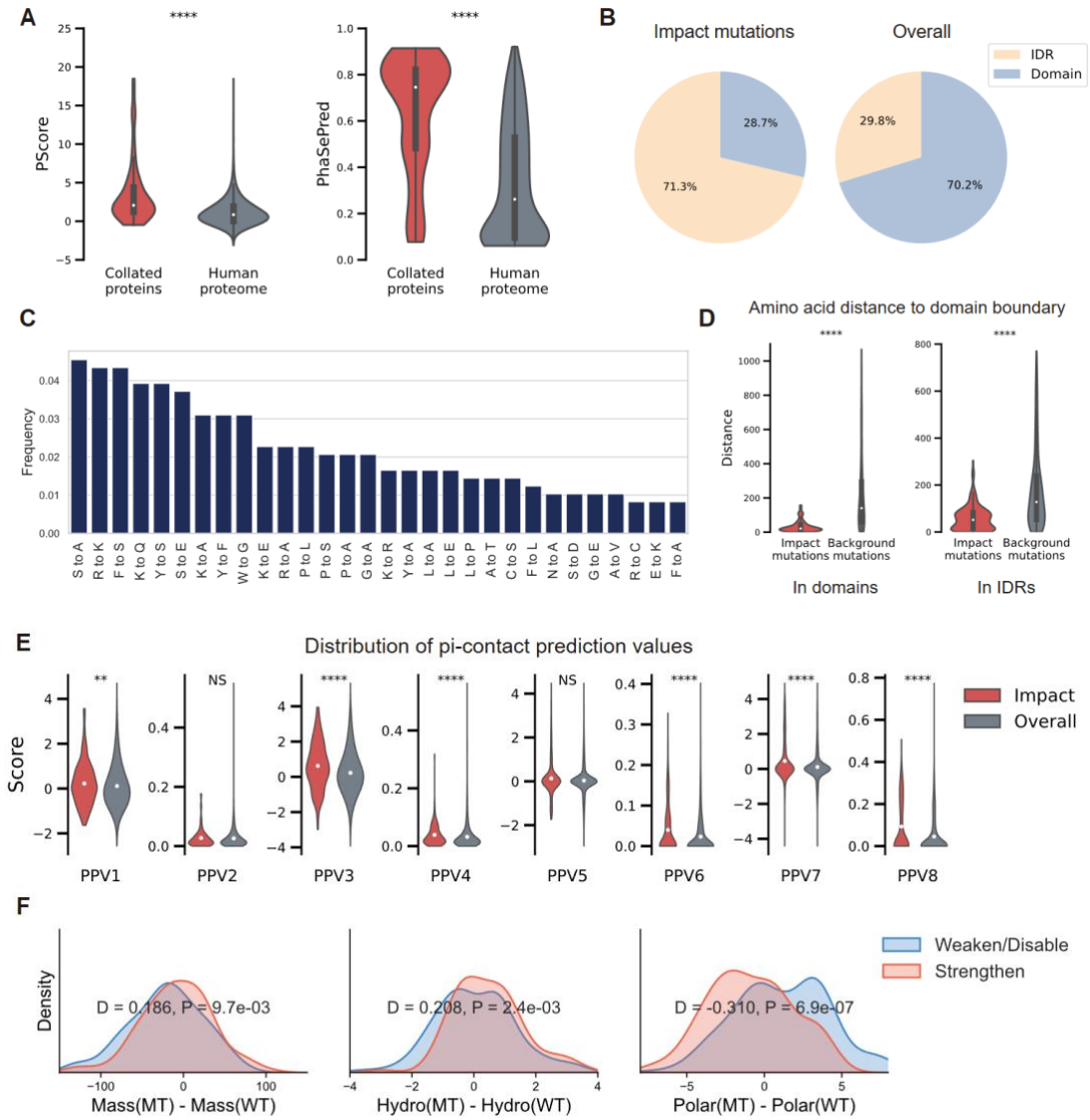


Fig. 2. Analyses of mutations that impact phase separation (PS).

(A) Comparison of the PS propensity of proteins corresponding to collected mutations (70 proteins) with the PS propensity of the human proteome. (PScore[27], Left; PhaSePred[35], Right; ****P < 0.0001, two-sided Mann–Whitney U test).

(B) The proportion of mutations affecting phase separation ('Impact' mutations) (Left) located in IDRs and Domains, compared with the total proportion of IDRs and Domains (Right). 'Impact' mutations are more likely to be located in IDRs rather than Domains.

(C) Top 30 high-frequency mutations among collated 'Impact' mutations.

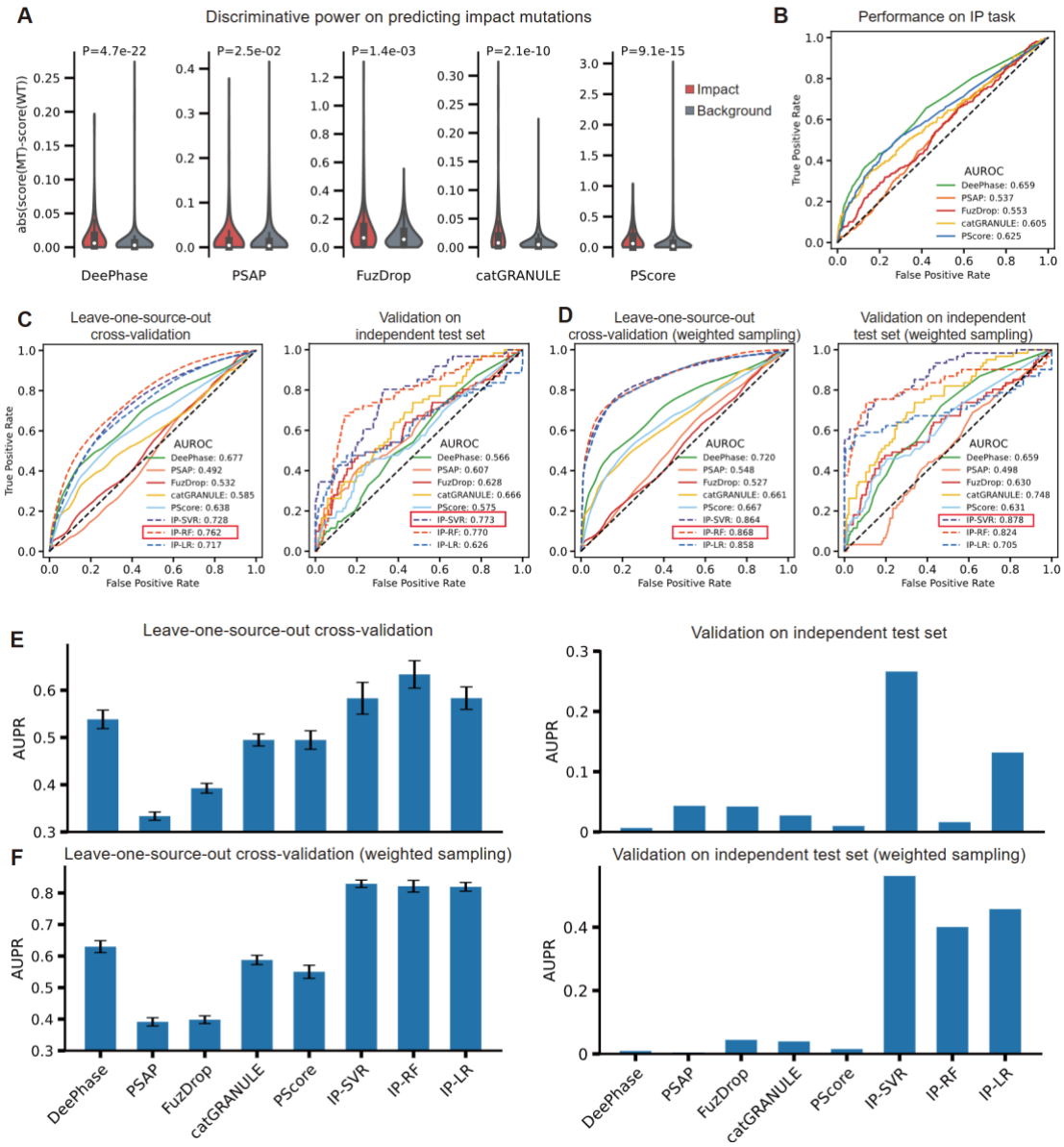
(D) Distribution of amino acid distances from each mutation site to the nearest domain boundary. Distances of 'Impact' mutations (in red) and 'Background' mutations (in grey)

were compared within Domains (Left) and within IDRs (Right). 'Impact' mutations exhibit significantly smaller distances in both cases (****$P < 0.0001$, two-sided Mann–Whitney U test).

(E) Distribution of eight pi-contact prediction values (PPVs) for mutation sites predicted by PScore[27]. Values of 'Impact' mutations (in red) and 'Background' mutations (in grey) were compared. The dot in each violin represents the average of values. Mutations impacting PS exhibit a higher frequency of pi-contact.

(F) Statistical comparison of the changes of amino acid property index before and after mutation between collected ‘Strengthen’ (n = 79, orange) and ‘Weaken/Disable’ groups (n = 228, blue) under Kolmogorov–Smirnov D test (WT, wild-type amino acid; MT, mutant amino acid). The direction of the D statistic was set as positive when the mean value of the ‘Strengthen’ group was higher and as negative when that of the ‘Weaken’ group was higher. Significant differences were observed in amino acid mass (Left), hydrophobicity (Middle) as well as polarity (Right).

941 **Figure3**



942

943

944 **Fig. 3.** Evaluation of model performance on predicting missense mutations that impact
945 natural phase separation (PS) ('Impact' mutation).
946 (A) Discriminative power evaluation of existing PS prediction methods, the absolute
947 difference in scores before and after each mutation were computed and visualized. For
948 each method, a paired violin plot displays score distributions (Left for 'Impact'
949 mutations and right for 'Background' mutations. The dot in each violin represents the
950 median of the values).
951 (B) Performance evaluation of existing PS prediction methods on discerning 'Impact'
952 mutations (IP task). AUROC is derived using the absolute difference in scores before

and after each mutation.

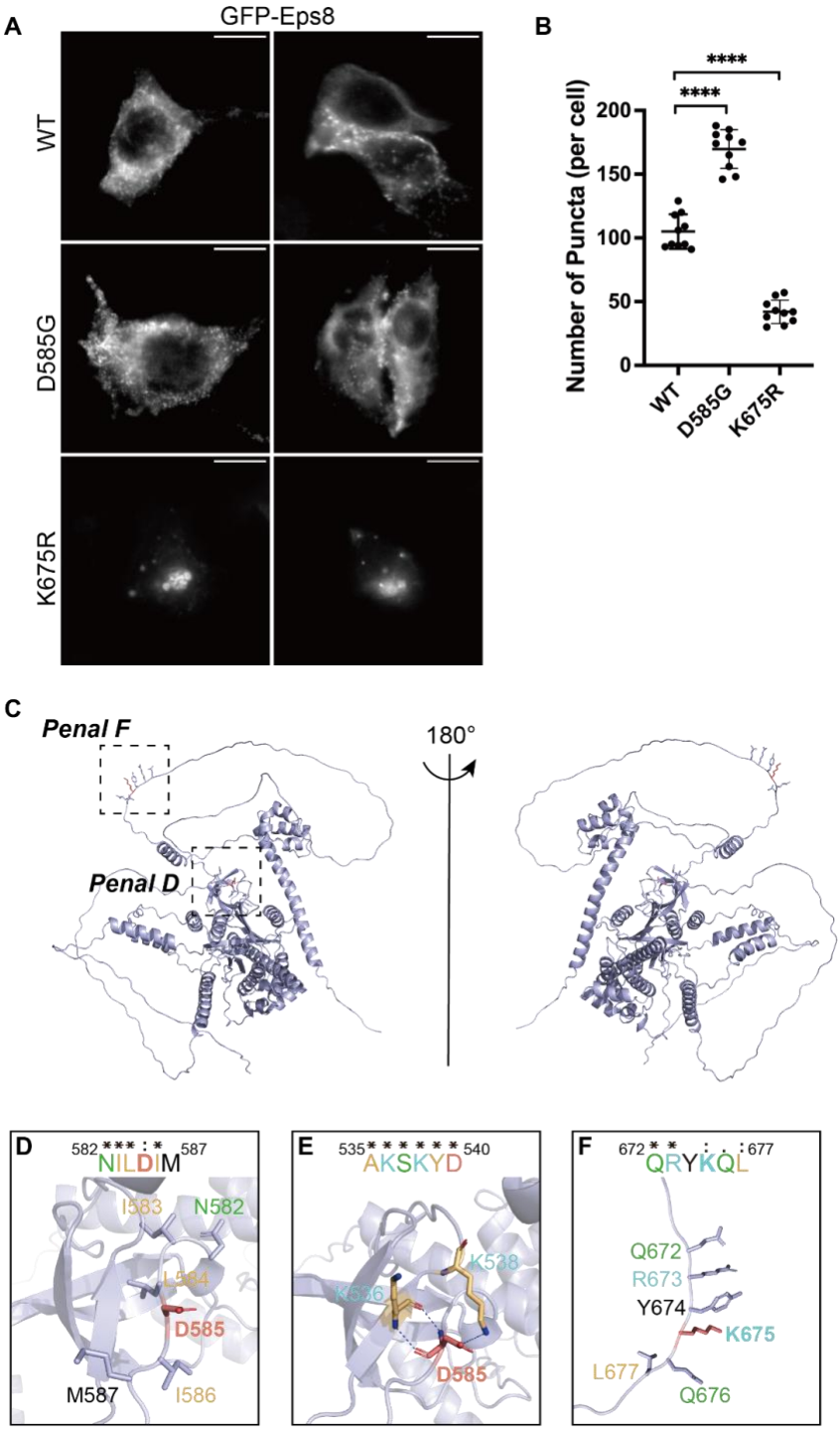
(C) Evaluation of models on discerning 'Impact' mutations using leave-one-source-out (LOSO) (Left) and independent test set (Right).

(D). A parallel evaluation similar to (C) but the 'Background' mutations were generated following the same IDRs: Domains ratio as the collected 'Impact' samples (weighted sampling).

(E) AUPR scores of the leave-one-source-out (LOSO) cross-validation (Left) and independent test set (Right) on the IP task, corresponding to (C).

(F) A parallel evaluation as (E) but the 'Background' mutations were generated following the same IDRs: Domains ratio as the collected 'Impact' samples (weighted sampling), corresponding to (D).

967 **Figure4**



968

969 **Fig. 4.** Experimental validation of *Eps8* missense mutations predicted by PSMutPred

970 to impact PS.

971 (A) Representative images of overexpressed GFP-*Eps8* and its mutants in HEK293

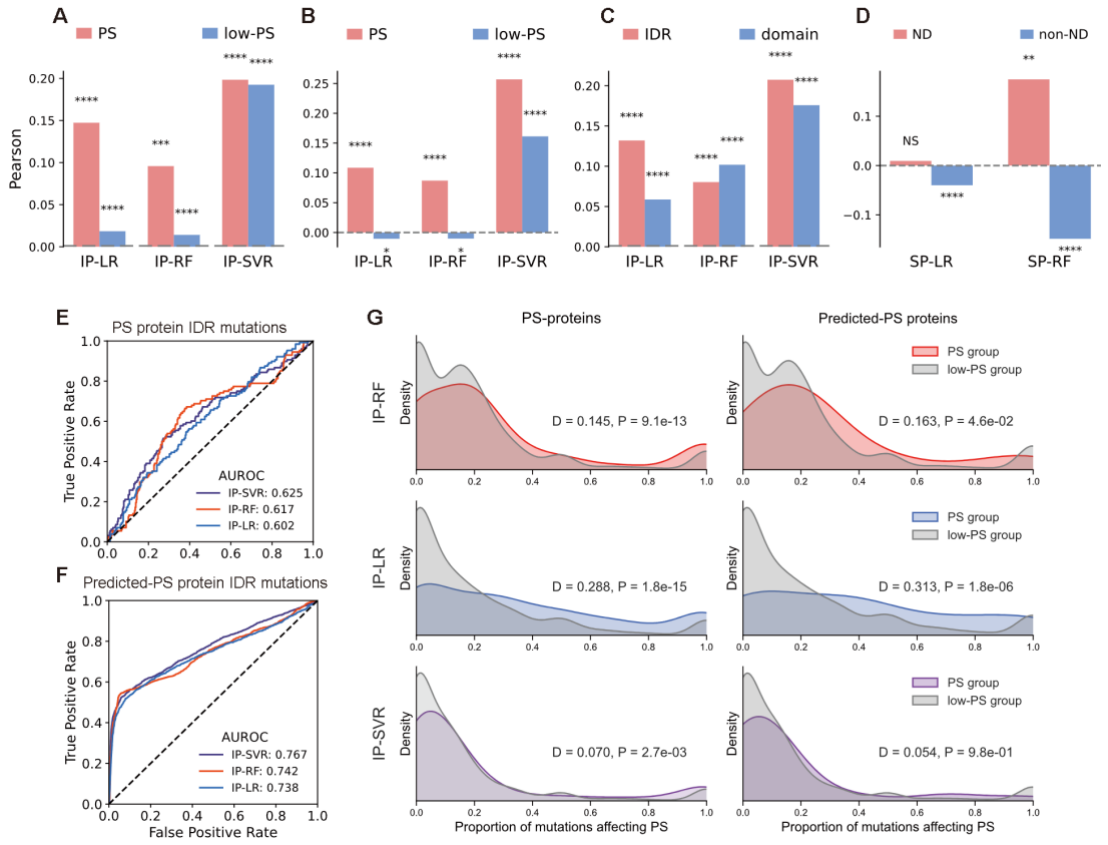
972 cells (scale bars: 10 μm; n = 10 randomly picked cells). WT denotes wild type.

(B) Quantification of puncta within the wild type and mutants of *Eps8* in HEK293 cells (****$P < 0.0001$ by two-tailed Student's t-test, $n = 10$ randomly picked cells). Error bars represent standard deviation (s.d.).

(C) Ribbon diagram representation of mouse *Eps8* structure predicted by AlphaFold2, showing both front (left) and back (right) views.

(D-F) Detailed regions involving missense mutations with their neighboring residues (D, F), and interaction analysis (E). The mutations are shown with the stick mode in red while hydrogen bonds are shown as blue dashed lines. Sequence alignments within critical residues are shown in bold.

985 **Figure5**



986

987 **Fig. 5.** Evaluation of PSMutPred scores across ClinVar[51] variants.

988 (A-D) Pearson correlations comparison between groups. For each group, the Pearson
989 correlation was computed by comparing the PSMutPred scores of its variants against
990 their respective ClinVar pathogenicity labels (pathogenic or likely pathogenic coded as
991 1 and benign or likely benign coded as 0). For SP-RF and SP-LR, variants being
992 predicted 'Impact' by PSMutPred-IP were considered. P-values of Pearson correlation
993 (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001; NS = not significant).

994 (A) Comparison between variants from the PS-prone group defined by phase separation
995 proteins (83 known PS proteins, 1,451 variants) and low-PS-prone group variants
996 (8,528 proteins, 84,840 variants).

997 (B) Comparison between variants from the predicted PS-prone group (1,276 proteins,
998 30,889 variants) and low-PS-prone group variants (7,335 proteins, 56,853 variants).

999 (C) Comparison between variants located in IDRs (n = 15,427) and Domains within the
1000 predicted PS-prone group (n = 15,462).

(D) Comparison between variants from neurodegenerative disease (ND) related proteins (19 proteins) and variants from other proteins (non-ND) (within the computationally predicted PS-prone group), scores of PSMutPred-SP models (n = 252) show a positive correlation with the pathogenicity of ND-related proteins, which differed from the overall pattern.

(E) Performance of PSMutPred-IP models on pathogenicity prediction of missense variants. Variants were selected from the PS-prone group defined by known PS proteins and further narrowed to those prone to locate in IDRs (489 variants).

(F) A parallel evaluation of (E) but focuses on variants from the PS-prone group predicted by algorithms (8,188 variants).

(G) Comparison of the proportion of variants predicted as ‘Impact’ by different PSMutPred-IP models between the PS-prone group and low-PS-prone group. The proportion of 'Uncertain Significance' (VUSs) IDR variants predicted as 'Impact' (PSMutPred-IP rank score > 0.8) for each protein was computed. The distributions of these proportion values are presented in a top-to-bottom sequence: IP-RF (top), IP-LR (middle), and IP-SVR (bottom). We compared the PS-prone group (83 proteins) with the low-PS-prone group (8,528 proteins) defined by known PS proteins (PS proteins, Left) and compared the PS-prone group (1,276 proteins) with the low-PS-prone group (7,335 proteins) predicted by algorithms (Predicted-PS proteins, Right). The differences were calculated based on Kolmogorov’s D statistic, the direction of the D statistic was set as positive when the mean proportion of the PS-prone group was higher and as negative when that of the low-PS-prone group was higher. PS-prone proteins have a higher proportion of missense mutations predicted to ‘Impact’ PS.

Figure6

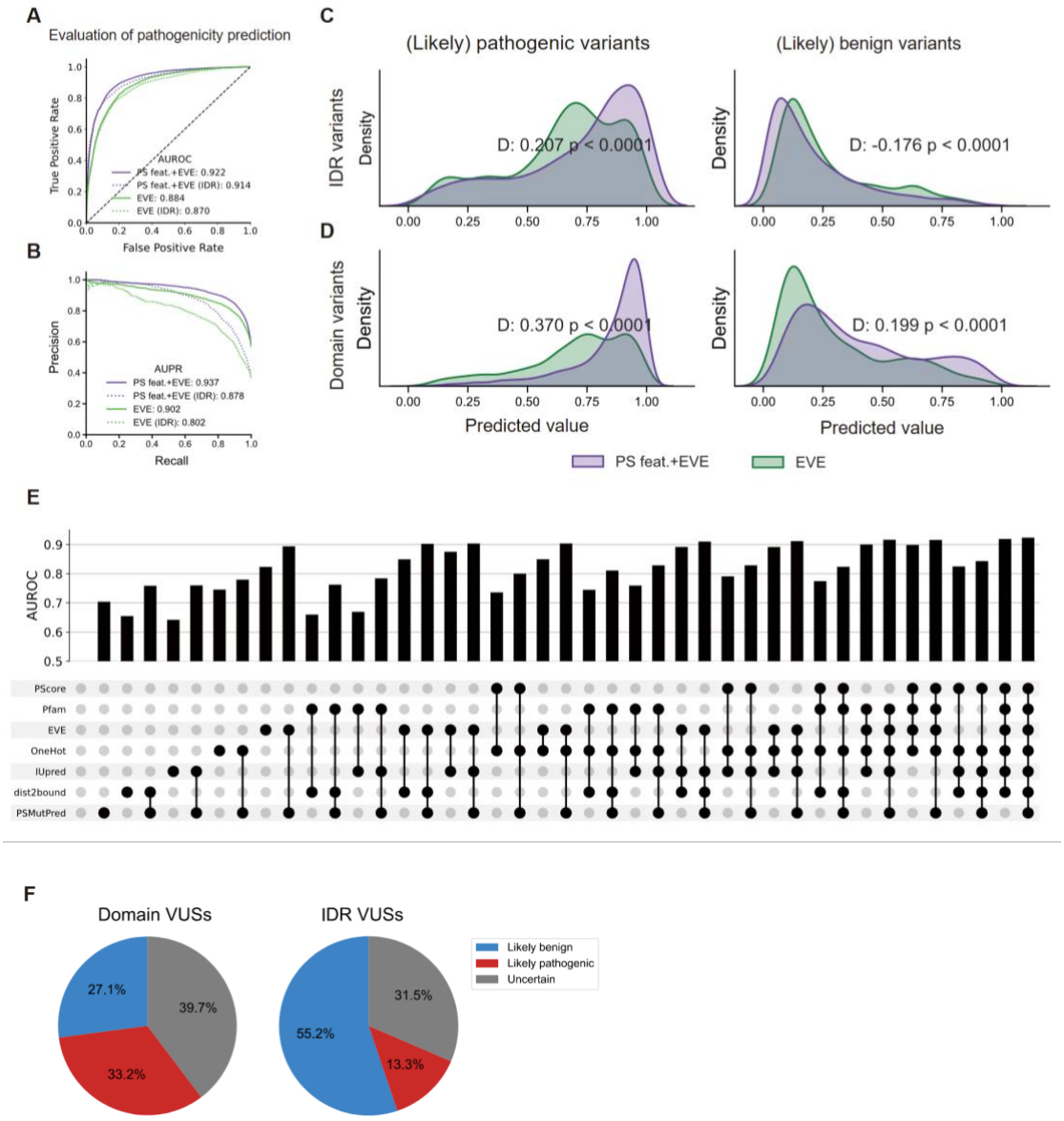


Fig. 6. Enhancement of missense variants' pathogenicity prediction by incorporating PS features including PSMutPred scores.

(A, B) AUROC (A) and AUPR (B) evaluation on the independent test set (n = 15,394). The dotted lines represent performance exclusively considering IDR variants (n = 5,656), revealing notable improvements when integrating PS-associated features into EVE.

(C, D) The divergence of predicted scores distributions between the standalone EVE (green) and the combined EVE with PS features model (purple), quantified using Kolmogorov's D statistic on the independent test set. The direction of the D statistic

was set as positive when the mean predicted value of EVE was lower and as negative when that of the model combining PS features was lower.

(C) The differential score distributions for pathogenic-prone variants (pathogenic and likely pathogenic, n = 2,044) (left graph) and benign-prone variants (benign and likely benign, n = 3,612) (right graph) with a focus on variants located in IDRs.

(D) A parallel evaluation of (C) but focuses on variants located in Domains (6,665 pathogenic or likely pathogenic and 3,073 benign or likely benign).

Except for benign variants located in Domains, the pathogenicity power can be significantly improved by incorporating variant PS features.

(E) Assessment of feature set impact on predictive performance. The bar chart displays the AUROC values achieved on the test set by Random Forest models trained with various random combinations of features. The predictive performance is especially compared with and without the inclusion of PSMutPred scores among the features. The inclusion of PSMutPred scores significantly influences the AUROC.

(F) Combining the PS feature with EVE scores, pathogenicity scores for 358,504 ClinVar variants of uncertain significance were predicted. Left: IDR variants (n = 178,719); Right: Domain variants (n = 179,785).