Generalizable prediction of liquid-liquid phase separation from protein sequence

Amir M. MohammadHosseini[1*], Hossein Teimouri[1*], Vincent Gureghian[2], Rayan Najjar[2], Ariel B. Lindner[1,3, 4‡], Amir Pandi[1,3, 4‡]

[1]Sorbonne Université, CNRS, ERL U1338 Inserm, Department of Computational, Quantitative and Synthetic Biology, Paris, France
[2] University of Washington, Seattle, WA, USA
[3]Sorbonne Université, CNRS, Inserm, Institut de Biologie Paris-Seine, Paris, France
[4]Sorbonne Université, CNRS, Université de Technologie de Compiègne, Inserm, Biofoundry Alliance Sorbonne Université, Paris, France
[*] These authors contributed equally to this work.
[‡] To whom correspondence should be addressed: ariel.lindner@inserm.fr, amir.pandi@inserm.fr

Abstract

Liquid-liquid phase separation (LLPS) is a physicochemical process through which a homogeneous liquid solution spontaneously separates into distinct liquid phases with different compositions and properties. Driven by weak multivalent interactions, LLPS in living systems enables dynamic compartmentalization of biomolecules to promote or regulate various cellular processes. Despite recent advances, predicting phase-separating proteins and their key LLPS-driving regions remains limited by the versatility of models. Here, we developed Phaseek, a generalizable LLPS predicting model that combines contextual sequence encoding with statistical protein graph representations. Phaseek accurately predicts LLPS-prone proteins in diverse biological contexts and identifies key regions and effects of point mutations. Proteome-wide analysis across 18 species suggests evolutionary conservation of LLPS among orthologs and associated biological processes. Additionally, predictions by Phaseek highlight the key physicochemical and structural properties associated with LLPS. Provided as an open-access model with a user-friendly implementation, Phaseek serves as a multipurpose LLPS predictor for advancing fundamental and applied research.

Introduction

Liquid-liquid phase separation (LLPS) in living cells is a physicochemical process by which specific biomolecules undergo demixing to form dense and dilute coexisting phases[1,2]. The major driving force of LLPS is weak multivalent intra- and inter-molecular interactions, *e.g.*, charge-charge, cation-π, dipole-dipole, π-π interactions, hydrophobic contacts and hydrogen bonds[2–7]. LLPS creates concentrated hubs of specific biomolecules and facilitates spatial compartmentalization of cellular processes[8,9]. The dynamic nature of such condensates allows for rapid assembly/disassembly to precisely regulate specific processes within cells[8].

LLPS was discovered in aqueous solutions in the early 1900s, and various mixtures have since been shown to exhibit this property[10–12]. In biology, LLPS was first reported in 2009 when P granules, molecular condensates made up of RNA and proteins, were shown to play a crucial role in germ cell development in *Caenorhabditis elegans*[13]. Since this discovery, numerous biomolecules including hundreds of phase-separating proteins have been identified in various cellular contexts, highlighting the ubiquity of LLPS in biology[3–6,14].

Photosynthesis as an ancient and one of the most fundamental biological processes benefits from LLPS. Carboxysomes in cyanobacteria and pyrenoids in photosynthetic algae are formed through LLPS, concentrating the enzyme ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) with other proteins involved in the process[15–19]. These condensates enable carbon-concentrating mechanisms and streamline the CO_2 fixation pathway. In the context of DNA damage repair, LLPS facilitates the formation of foci that concentrate repair machinery at damage sites, thereby enhancing the efficiency of the repair process[20–22]. LLPS also plays an important role in coordinating RNA metabolism by forming condensates that contain splicing factors and specific RNA molecules, enhancing the efficiency of RNA processing[13,23–26]. Such ubiquitous and important roles of LLPS in biological processes, ranging from bacteria to humans, are reshaping our understanding of cellular organization and functions[9,27,28]. Importantly, LLPS plays a crucial role in human health and aging[29]; with dysregulation of phase separation being implicated in pathological conditions such as neurodegenerative disorders, cancer, but also viral infections[30–33]. Thus, condensates constitute potential therapeutic targets[34,35] and the prediction of phase-separating proteins can be impactful for both basic and clinical research.

Over the course of evolution, a large number of proteins developed intricate sequence patterns enabling them to acquire different structures and functions. Taking this as an opportunity, machine learning models are increasingly used for classification, regression, and generation tasks for diverse applications[36–38]. Given the widespread LLPS examples in biology, multiple predictive models have been developed[39–43,44]. These models, based on pre-defined physicochemical and sequence features, often face limitations in generalizability across diverse proteins. To address this limitation, we develop Phaseek, a full sequence-based model combining the transformers architecture and statistical protein graphs. We demonstrate that our model achieves superior performance compared to existing models in predicting phase-separating proteins and their LLPS-driving key regions. We showcase Phaseek's ability to predict changes in LLPS scores due to point mutations. We then predict phase-separating proteins across 18 organisms and investigate their

conservation among orthologs and association with Gene Ontology terms. Finally, we analyze the proteome-wide structural and physicochemical properties associated with predicted phase separation.

We provide the weights of the Phaseek model under an open source license and a user-friendly implementation as a Google Colab notebook allowing users to query the prediction of LLPS score of protein sequences and their key regions (**Code availability**).

Results

Phaseek, a generalizable model for the prediction of LLPS from protein sequence

Our LLPS prediction model, Phaseek, comprises three modules. The first module is an encoder based on Generative Pre-trained Transformer (GPT) architecture[45] (**Fig. 1a**). For training, we compiled and curated previously published data, totaling ~1,500 positive and ~5,000 negative phase-separating sequences. We augmented the set of ~3,500 positive sequences using a trained Hidden Markov Model to reach a total of ~5,000 and split the data between training and test with a 9:1 ratio (**Methods**).

The second module, a protein graph generator, generates statistical protein graphs derived from protein sequences using FEGS (Feature Extraction based on Graphical and Statistical Features) method[46]. The FEGS method generates 158 unique graphical representations to encode proteins[46]. To optimize the time complexity of using these representations as statistical matrices, we used a random forest model and calculated the SHapley Additive exPlanations (SHAP) value for each matrix to assess its predictive importance (**Methods**). We finally used the top 10 statistical matrices with the highest SHAP values in parallel with our GPT architecture to produce residue-level scores (**Fig. 1a, Supplementary Table 1).** These two modules collectively comprise our core model.

The third module is an XGBoost-based bottleneck. LLPS regions can carry the multivalency and interactions required for phase separation and promote LLPS when fused to other proteins[44,47,48]. Defining key regions as sub-sequences with amino acids having high predicted scores, this module boosts the contribution of these regions to the final sequence score (**Methods, Supplementary Fig. 1a**). Transitioning from a decoder-only version of the model to the three-module version, the AUC score improved from 0.84 to 0.93 for the evaluation set (**Methods**, **Supplementary Fig. 2**).

We compared Phaseek with other LLPS predictors, including PSPHunter[44], CatGranule 1.0[49], CatGranule ROBOT 2.0[43], PScore[39], PSPredictor[41], and PSTP[50] (for information about each model, see **Methods**). Using our evaluation set, Phaseek resulted in the highest AUC score: 0.93 (PSPHunter: 0.88, CatGranule 1.0: 0.78, and CatGranule 2.0: 0.67, PScore: 0.71, PSPredictor: 0.90, and PSTP: 0.87 and 0.78) (**Fig. 1b, c**). Next, we investigated the versatility and generalizability of Phaseek by comparing the models' performances in four distinct biological contexts.

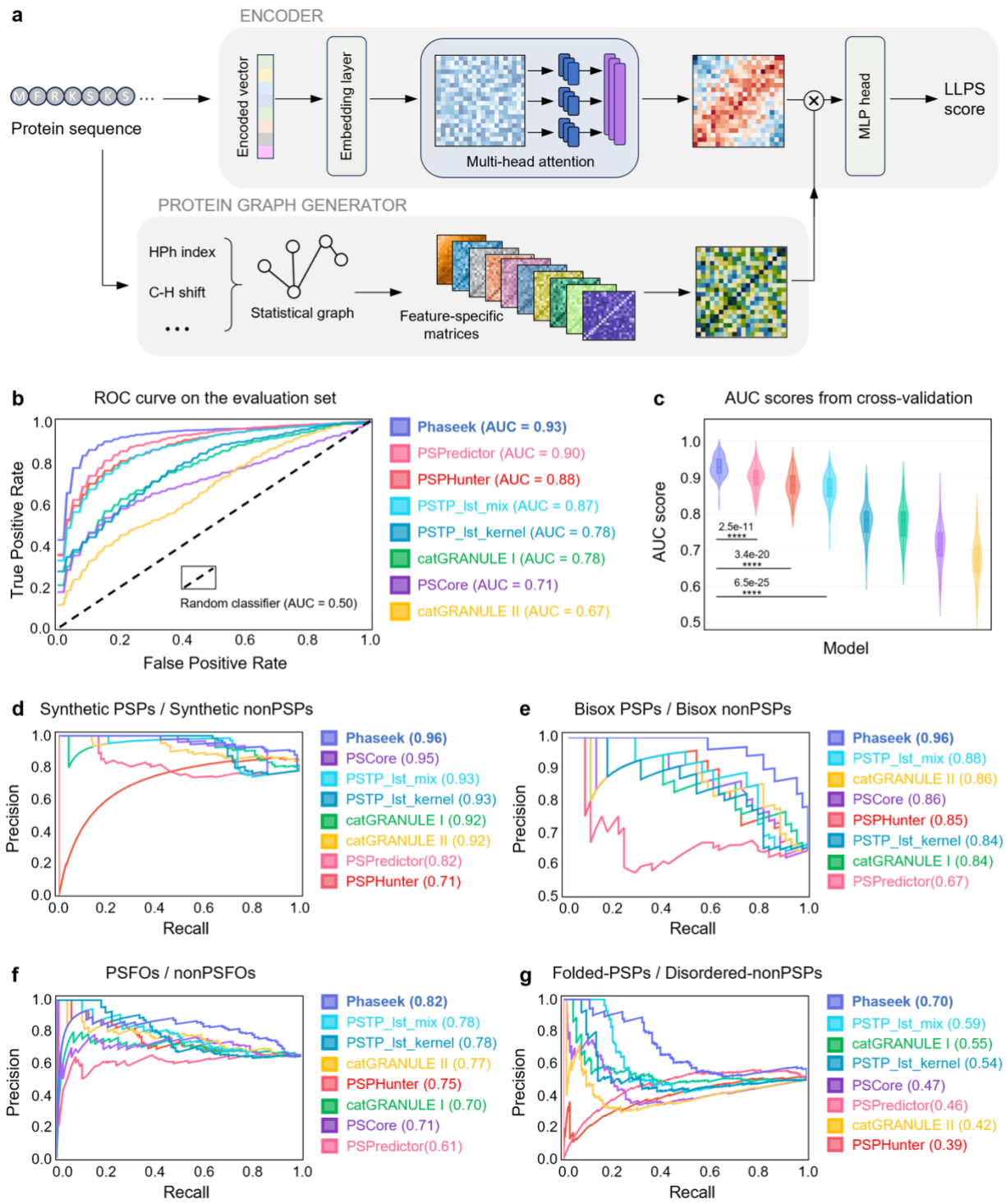

Fig. 1. Phaseek architecture and model evaluation. (a) The core model of Phaseek comprises two modules: a GPT-based encoder that extracts contextual sequence features, and a protein graph generator that encodes statistical representations using the FEGS method. The output from both modules is integrated to produce residue-level LLPS scores. For clarity, the XGBoost-based bottleneck module, used to boost the contribution of LLPS-enriched regions to the final prediction, is not shown. **(b)** Performance of Phaseek and other models shown as ROC (Receiver Operating Characteristic) curves on the evaluation set comprising phase-separating (n = 300) and non-phase-separating (n = 337)

sequences, compiled from LLPS datasets and select studies, excluding sequences used in training. **(c)** Violin plots illustrating ROC area under the curve (AUC) scores for Phaseek and other models, evaluated using Monte Carlo Cross-Validation (MCCV)[123]. Batches containing 40 phase-separating and 40 non-phase-separating sequences from the evaluation set were selected and tested over 100 iterations. The statistical significance is indicated by the one-sided Mann–Whitney U p-values (****$p < 0.0001$). Violin plots represent data distributions, and boxes indicate the interquartile range (Q1–Q3) with the median marked by the center line. Models are color-coded in the figure: Phaseek (periwinkle blue), PSPhunter (deep salmon), catGRANULE I (Caribbean green), PScore (medium purple), catGRANULE II (saffron yellow), PSPredictor (coral pink), PSTP-LST-mix (bright sky blue), and PSTP-LST-kernel (deep cerulean). To ensure a fair comparison of versatility and generalizability, four extra external datasets were used: **(d)** Synthetic proteins, including synthetic phase-separating proteins (synthetic PSPs, n = 97) and synthetic non-phase-separating proteins (synthetic nonPSPs, n = 27). **(e)** Biotinylated isoxazole (B-isox)-identified proteins: phase-separating (Bisox PSPs, n = 43) and non-phase-separating (Bisox nonPSPs, n = 24). **(f)** Fusion oncoproteins: phase-separating (PSFOs, n = 115) and non-phase-separating (nonPSFOs, n = 63). **(g)** Folded phase-separating proteins (Folded-PSPs, n = 148) and disordered non-phase-separating proteins (Disordered-nonPSPs, n = 185). Protein sequences shorter than 140 amino acids, which cannot be classified by PScore, are shown in gray.

The first dataset consisted of 97 phase-separating and 27 non-phase-separating synthetic proteins[51–55]. Phaseek outperformed all models with an accuracy of 0.89 (PSPHunter: 0.67, CatGranule 1.0: 0.77, and CatGranule 2.0: 0.60, PScore: 0.80, PSPredictor: 0.80, and PSTP: 0.72 and 0.69) (**Fig. 2a**).

The second dataset corresponds to a proteome study on cell extracts from *Arabidopsis thaliana* seedlings and inflorescences[56] in which intrinsically disordered proteins were precipitated using biotinylated isoxazole (B-isox) and their LLPS behavior was experimentally confirmed (43 phase-separating and 24 non-phase-separating proteins). Again, Phaseek showed the greatest accuracy with a score of 0.85 (PSPHunter: 0.67, CatGranule 1.0: 0.73, and CatGranule 2.0: 0.60, PScore: 0.67, PSPredictor: 0.70, and PSTP: 0.67 and 0.64) (**Fig. 2b**).

The third data set consists of fusion oncoproteins (FOs), which may facilitate oncogenesis through LLPS[47], in total 115 phase-separating and 63 non-phase-separating FOs. Once again, Phaseek achieved the best accuracy with a score of 0.72 (PSPHunter: 0.61, CatGranule 1.0: 0.54, and CatGranule 2.0: 0.62, PScore: 0.58, PSPredictor: 0.60, and PSTP: 0.61 and 0.56) (**Fig. 2c**).

The fourth dataset contains 185 predicted as disordered but LLPS-negative proteins and 148 folded (non-disordered) but LLPS-positive proteins (**Methods**). Multivalent interactions in intrinsically disordered regions (IDRs) can result in phase separation but not all IDRs promote LLPS and folded proteins can also display this behavior[57,58]. Phaseek classified phase-separating and non-phase-separating proteins with an accuracy of 0.62, higher than all other models (PSPHunter: 0.43, CatGranule 1.0: 0.49, CatGranule 2.0: 0.59, PScore: 0.61, PSPredictor: 0.32, and PSTP: 0.42 and 0.50) (**Fig. 2d**).

Across these diverse biological datasets, Phaseek achieved superior performance in classifying phase-separating and non-phase-separating proteins, establishing it as a versatile and reliable tool for LLPS prediction.

Phaseek identifies LLPS key regions and predicts the effects of point mutations

Multiple diseases and pathological conditions are associated with mutations occurring in key regions of phase separation proteins[31,48,59–62]. Thus, the identification of LLPS key regions is important to both fundamental and translational research. To assess the ability of Phaseek to detect LLPS key regions, we tested it against experimentally annotated sequences from four studies[63–66]. Comparing the residue-level LLPS score profiles of these sequences, Phaseek (followed by PSTP) outperformed other models in distinguishing LLPS from non-LLPS regions (**Fig. 2a, b, Supplementary Fig. 3a, b**) as measured by a custom match score (**Methods**). For instance, for the human LINE-1 retrotransposable element, with an overall LLPS score of 0.82, Phaseek captured the known LLPS key region (**Fig. 2b, c**). Notably, removing residues 47-53 (segment A) and 132-152 (segment B) from the protein sequence, compared to random truncation, reduced the overall LLPS score to 0.43 and 0.37, respectively (**Fig. 2c, Supplementary Fig. 4**), *i.e.*, below the classification threshold of 0.7 (**Methods**). This finding recapitulates experimental results that the deletion of either of these segments disrupts the LLPS phenotype[63].

Next, we analyzed variations in LLPS scores following the fusion of phase-separating peptides to non-phase-separating proteins. For interferon-alpha A (IFN-alpha A) fused to an artificially designed elastin-like protein (ELP)[67], all models except catGRANULE 2.0 correctly predicted LLPS scores for IFN-alpha A (<0.5), ELP (>0.5) and IFN-alpha A-ELP (>0.5) (**Supplementary Fig. 3c**). For the fused sequence of an artificially designed peptide (HERD 2,2), mCherry, and flavin-containing monooxygenases (FMO)[51], only Phaseek predicted correct LLPS scores, *i.e.*, <0.5 for mCherry and FMO, and >0.5 for HERD 2,2 and HERD 2,2-mCherry-FMO (**Supplementary Fig. 3d**). Notably, while we used 0.5 as the arbitrary comparison threshold across different models, this observation remains valid when Phaseek's optimal threshold of 0.7 is applied (**Supplementary Fig. 5**).

Next, we sought to investigate the effect of point mutations on LLPS. We collected 2,617 experimentally validated impactful mutations on 137 phase-separating proteins from multiple studies[68–70] (**Data availability**). Additionally, we introduced 20 random mutations for each of the 137 proteins, yielding a total of 2,740 mutations (**Methods**). To evaluate Phaseek's ability to capture the effects of point mutations at the sequence, amino acid, and regional level, we calculated three custom metrics: Δ(score), Δ(aa_score), and Σ(Δ(aa_score)), respectively (**Methods**). Using these metrics, Phaseek significantly discriminated impactful from random mutations (**Fig. 2d, Supplementary Fig. 6**). Notably, the Δ(score) and Σ(Δ(aa_score)) metrics, which capture sequence and cumulative regional level variations, resulted in significant differences (p-value = 1.1×10^{-19} and 4.6×10^{-12}), in both non-IDR and IDR regions (**Supplementary Fig. 7**). However, the Δ(aa_score) metric, which focuses solely on the mutated residue, showed no significant difference (p-value = 0.83) (**Supplementary Fig. 6, 7, 9)**. These results suggest that Phaseek relies on contextual information in predicting mutation effects. Among impactful mutations, polar to aromatic, particularly serine to phenylalanine and serine to tyrosine, resulted in the highest score increase. Meanwhile, aromatic to hydrophobic mutations, especially tyrosine to isoleucine, were associated with the largest decreases in score (**Supplementary Fig. 8**). These findings are consistent with prior reports[71,72]**.**

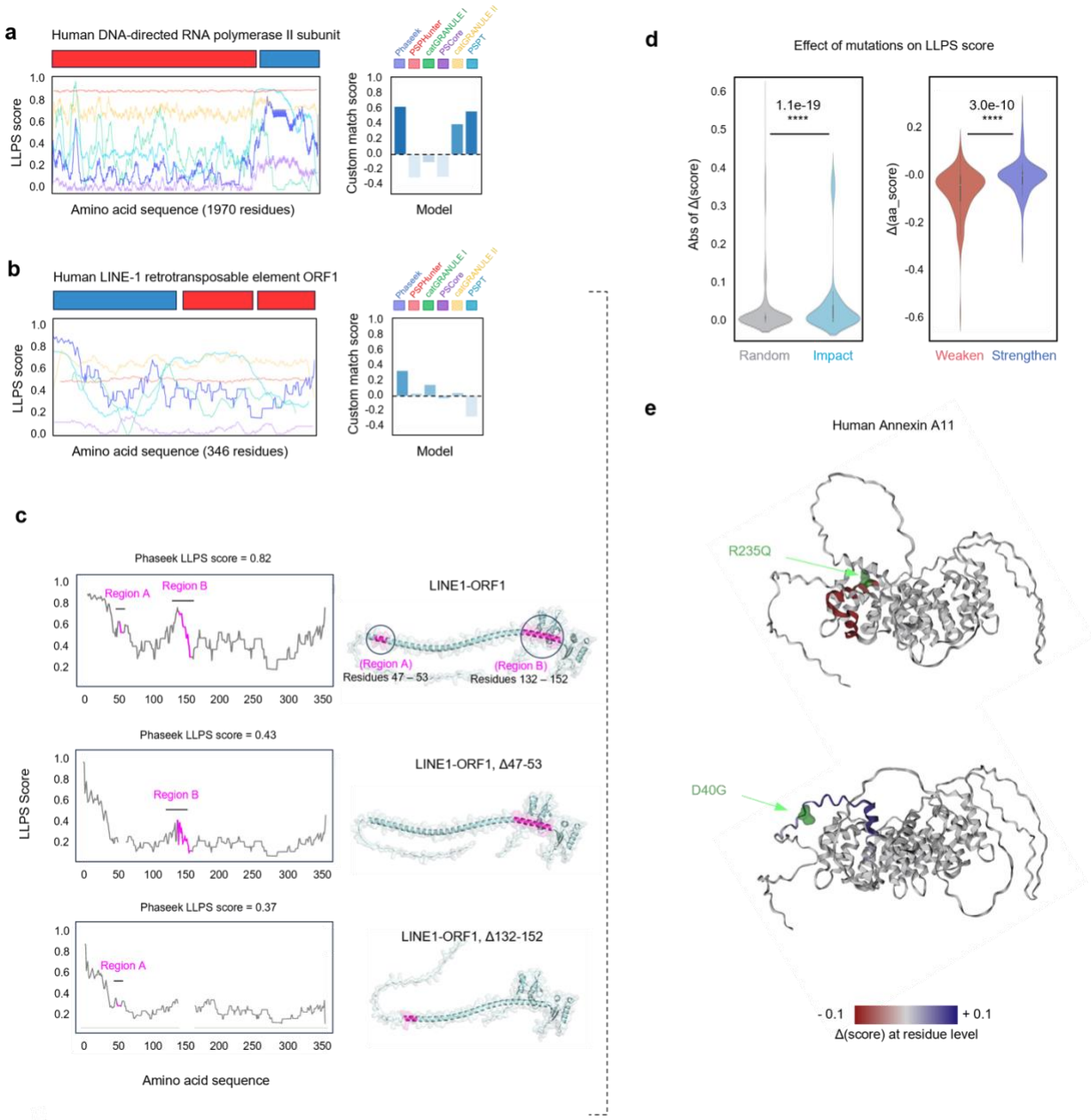

Fig. 2. Phaseek identifies LLPS key regions and mutation effects. (a) Residue-level LLPS predictions from different models and their comparison using a custom match score (**Methods**) on four human proteins not used during Phaseek training, with experimentally validated phase-separating (blue) and non-phase-separating (red) regions. These proteins include AT-rich interactive domain-containing protein 1A (ARID1A), DNA-directed RNA polymerase II subunit, and recombinant Tau protein (see **Supplementary Fig. 3a** for ARID1A and recombinant Tau). **(b, c)** Analysis of LINE1-ORF1, containing two experimentally validated segments: residues 47–53 (Region A) and 132–152 (Region B). Residue-level LLPS predictions from various models and their comparisons are shown for the wild-type form (LLPS score: 0.82), deletion of Region A (LLPS score: 0.43), and deletion of Region B (LLPS score: 0.37). Structures were predicted using AlphaFold2. **(d)** Comparative performance of Phaseek in distinguishing impactful mutations from random mutations (violin plots on the right) and strengthening mutations from weakening mutations (violin plots on the left), using three customized evaluation metrics. See **Supplementary Fig. 6-9** for full mutation performance analysis. **(e)** A case study on human Annexin A1 (UniProtKB: P50995) showing predicted effects of six point mutations, three LLPS-weakening

(R235Q, H390P, R456H) and three LLPS-strengthening (D40G, R346C, G38R), based on experimental evidence. Changes in residue-level LLPS scores are shown on AlphaFold2-predicted structures, with decreased scores in red, increased scores in blue, and mutated residues in green (**Supplementary Fig. 10**). Predictions by catGRANULE I and PScore in panels (a, b) were scaled to a 0–1 range using the StandardScaler module in scikit-learn for comparison purposes. Models not included in panels (a, b) do not generate residue-level LLPS profiles. Model colors: Phaseek (periwinkle blue), PSPhunter (deep salmon), catGRANULE I (Caribbean green), PScore (medium purple), catGRANULE II (saffron yellow), PSTP-LST-kernel (deep cerulean). The statistical significance is indicated by the one-sided Mann–Whitney U p-values (****$p < 0.0001$).

Phaseek was evaluated on a set of sequences with experimentally characterized LLPS-promoting and -weakening mutations, and accurately identified both types (**Fig. 2d, Supplementary Fig. 9**). As an example, we examined such mutations in human Annexin II[73], which facilitates RNA transport in neurons and has been implicated in amyotrophic lateral sclerosis[73]. In line with the experimental observations, Phaseek captured the effects of both weakening (R235Q, H395P, and R456H) and strengthening (D40G, R346C, and G38R) mutations on LLPS (**Fig. 2e and Supplementary Fig. 10**).

Together, these results demonstrate that, despite being trained on whole sequences without labeled regions, Phaseek identifies sub-sequences driving LLPS and predicts the effect of mutations at residue level.

Phaseek reveals LLPS conservation among orthologs and associated biological functions

Few studies have investigated the conservation of phase separation across the tree of life[74,75,76]. Whether LLPS is conserved among orthologs has remained largely unexplored. To address this, we examined the proteomes of 18 organisms: *Escherichia coli*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Anopheles gambiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Danio rerio*, *Xenopus laevis*, *Gallus gallus*, *Rattus norvegicus*, *Mus musculus*, *Canis lupus familiaris*, *Bos taurus*, *Sus scrofa*, *Macaca mulatta*, *Pan troglodytes*, and *Homo sapiens*. If phase separation is a conserved feature, it is more likely that members of an orthogroup coherently show LLPS behavior, with variations arising only if orthologs adapted through evolutionary divergence[74].

In each of the 18 proteomes, we quantified the fraction of phase-separating predicted proteins. Among all organisms, *E. coli* had the lowest fraction (2%), consistent with the observation that eukaryotes have more disordered proteins than prokaryotes[77]. By contrast, *S. cerevisiae* and *A. thaliana* each reached 12%, and *H. sapiens* 27% (**Fig. 3a**). Next, we categorized the proteome sequences using OrthoFinder[78–80], resulting in 24,438 orthogroups, of which 216 had contributions from every organism (**methods**). All orthogroups were then sorted by size and further categorized by the number of organisms contributing at least one ortholog with predicted LLPS (**Fig. 3b**). Notably, a species can encode multiple instances of an ortholog, *e.g.*, as isoforms, within the same orthogroup**.** In the heatmap shown in **Fig. 3b**, the marked increase in the fraction of LLPS-positive orthogroups near the horizontal and diagonal margins suggests that LLPS holds functional relevance across orthologs. Additionally, we tested for outliers among orthogroups in which only one organism was absent versus those in which only one organism was present and found that *D. rerio* and *C. familiaris* as outliers, respectively (**Supplementary Fig. 11a and 11b**).

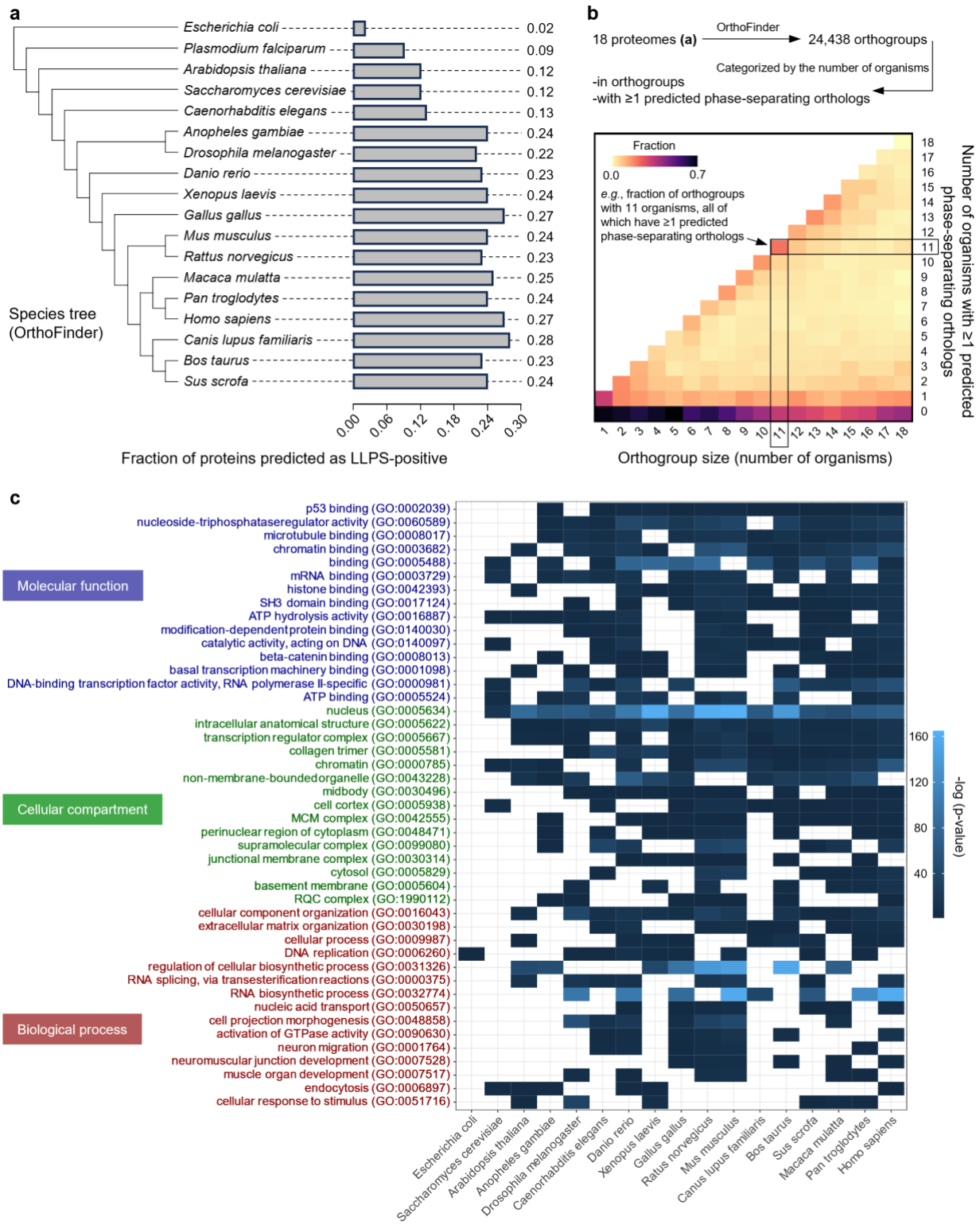

Fig. 3. Phaseek reveals LLPS conservation among orthologs and identifies biological processes involved. (a) Fractions of proteins predicted to be phase-separating by Phaseek are shown along with the species tree generated by OrthoFinder[78–80]. **(b)** Orthologs extracted using OrthoFinder across the 18 proteomes to study the evolutionary

conservation of LLPS. 24,438 orthogroups were identified with sizes of 1 to 18, representing the number of organisms that each contains. Orthogroups were categorized based on: (i) the number of organisms in each group (orthogroup size), and (ii) the number of organisms with at least one LLPS-predicted protein. For each orthogroup size, the heatmap shows the fraction of groups with ≥1 predicted LLPS proteins for each organism (example provided in the figure). **(c)** Gene Ontology (GO) terms enriched for the LLPS predicted fraction of 17 proteomes (all organisms but *P. falciparum*). Semantic similarity was applied to reduce the list of enriched terms.

Among the orthogroups with contributions from all 18 organisms, the one corresponding to the topoisomerase II enzyme family (Topo II) included 14 members with LLPS scores >0.7. Topo II resolves tensions from DNA supercoiling and entanglements[81]. In *S. cerevisiae* and *H. sapiens*, Topo II (TOP2α/TOP2β) undergoes LLPS via its less-conserved disordered C-terminal domain (CTD) to couple catenation with chromatin condensation[81,82]. Multiple sequence alignment further illustrates how phase-separating CTDs can retain high LLPS scores despite low conservation scores, suggesting that LLPS behavior can be conserved despite changes in the amino acid sequence (**Supplementary Fig. 12**).

To identify biological processes involving LLPS, we used proteome predictions of the 18 organisms but *Plasmodium Falciparum,* as its annotation package (org.Pf.plasmo.db on Bioconductor) is not maintained. For each organism, we enriched the set of genes for the LLPS-positive proteins for Gene Ontology (GO) terms and applied semantic similarity to identify the most relevant terms (**Methods and Supplementary Table 2**).

In the Molecular Function category, the majority of enriched terms were related to binding, coherent with the phenomenon of LLPS (**Fig. 3c**). Among the top-results, nucleoside-triphosphatase regulator activity (GO:0060589) is shared across 13 organisms, ATP hydrolysis activity (GO:0016887) is shared across 11 and ATP binding (GO:0005524) across 10 which is in line with the central role of ATP in LLPS regulation[83–85]. Interestingly, the aforementioned Topo II enzyme corresponds to the parental term catalytic activity, acting on a nucleic acid (GO:0140640) enriched in 10 organisms. In the Cellular Compartment category, the term nucleus (GO:0005634) was enriched in 15 organisms while transcription regulator complex (GO:0005667) and MCM complex (GO:0042555) were found enriched across 14 organisms. In the Biological Process category, the top-result is cellular component organization (GO:0016043) enriched in 13 organisms followed by homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156) enriched in 12 organisms (see **Code availability**).

Orthologs display LLPS behavior in a coordinated fashion, suggesting evolutionary conservation. The biological functions enriched across organisms recapitulate processes known to involve LLPS and suggest novel areas of investigation.

Phaseek unravels physicochemical and structural determinants of LLPS at the proteome level

The LLPS phenotype is driven by amino acid sequences with specific physicochemical and structural properties[86]. We studied such features using the proteomes of 18 organisms for which the LLPS behavior of individual proteins was predicted by Phaseek. We extracted the secondary

structures and physicochemical features of these sequences to investigate their correlations with the predicted LLPS propensity (**Fig. 4a**, **Methods**).

First, we compared the secondary structures predicted by AlphaFold[87] (helices, sheets, turns, bends, and IDRs) between random and key regions (**Methods**), where key regions were defined as subsequences predicted to have a high LLPS propensity. Compared to key regions, random regions contained significantly more ordered structures such as helices (median fraction of 0.422 vs 0.037, respectively) (**Fig. 4b**, see **Supplementary Fig. 13** for species-specific plots). In contrast, key regions were significantly enriched in disordered structures (median of 0.913 vs 0.188 in random regions). This supports the widely accepted notion that IDRs are important drivers of LLPS[7].

Second, we examined the relationship between predicted LLPS propensity (the score of the full-length sequence) and 12 physicochemical features across proteomes by computing Spearman correlations (**Fig. 4c**, **Supplementary Fig. 14**). One of the physicochemical features, IDR score, is the average of the disorder propensity for individual residues predicted by AIUPred[88]. Among all features, IDR score exhibited the highest correlation with LLPS score (Spearman $\rho = 0.45$). Indeed, disordered regions facilitate effective multivalent inter-/intra-molecular interactions[89]. In contrast, hydropathy, the fraction of hydrophobic residues, and the fraction of aromatic residues showed inverse correlations ($\rho = -0.41$, -0.40, and -0.35, respectively). Hydrophobic and aromatic interactions favor protein folding and (in extreme cases) solid-like assemblies[7]. Polyproline II (PPII) helix propensity and glycine/proline fraction were positively correlated ($\rho = 0.33$ and 0.27). PPII helix propensity has been reported to contribute to protein-protein interactions[90] and protein-nucleic acid interactions[91]. Moreover, glycine and proline residues have been found to be essential in phase separation[44]. Aromatic clustering, net charge, and charged amino acid patterning (*e.g.*, Kappa, representing the tendency of positive/negative amino acids to form clusters) showed the lowest correlations ($\rho = -0.09$, -0.07 and 0.04, respectively). Although the patterning of charged and aromatic residues has been associated with phase separation[56,92,93], LLPS often depends only on a specific region within a sequence.

We focused on the physicochemical analysis of LLPS-promoting regions (key regions) and compared them to random regions (**Methods**). We compared medians and common-language effect sizes (CLES)[94] to quantify differences between their distributions. Key regions exhibited a median aromatic clustering score of 0.34 lower than random regions (CLES = 0.22, meaning key regions have higher values in 22%, or lower values in 78%, of all pairwise comparisons), and a median Kappa of 0.043 higher (CLES = 0.59) (**Fig. 4d, Supplementary Fig. 15, 16**). Thus, key regions exhibited higher Kappa values, indicating greater charge blockiness, which can enhance inter-/intramolecular interactions[92,95]. The differences in aromatic clustering suggest key regions display a greater dispersion of aromatic residues, balancing between LLPS and aggregation[7,93,96].

Overall, these results show that physicochemical features associated with LLPS hold true at the proteome level.

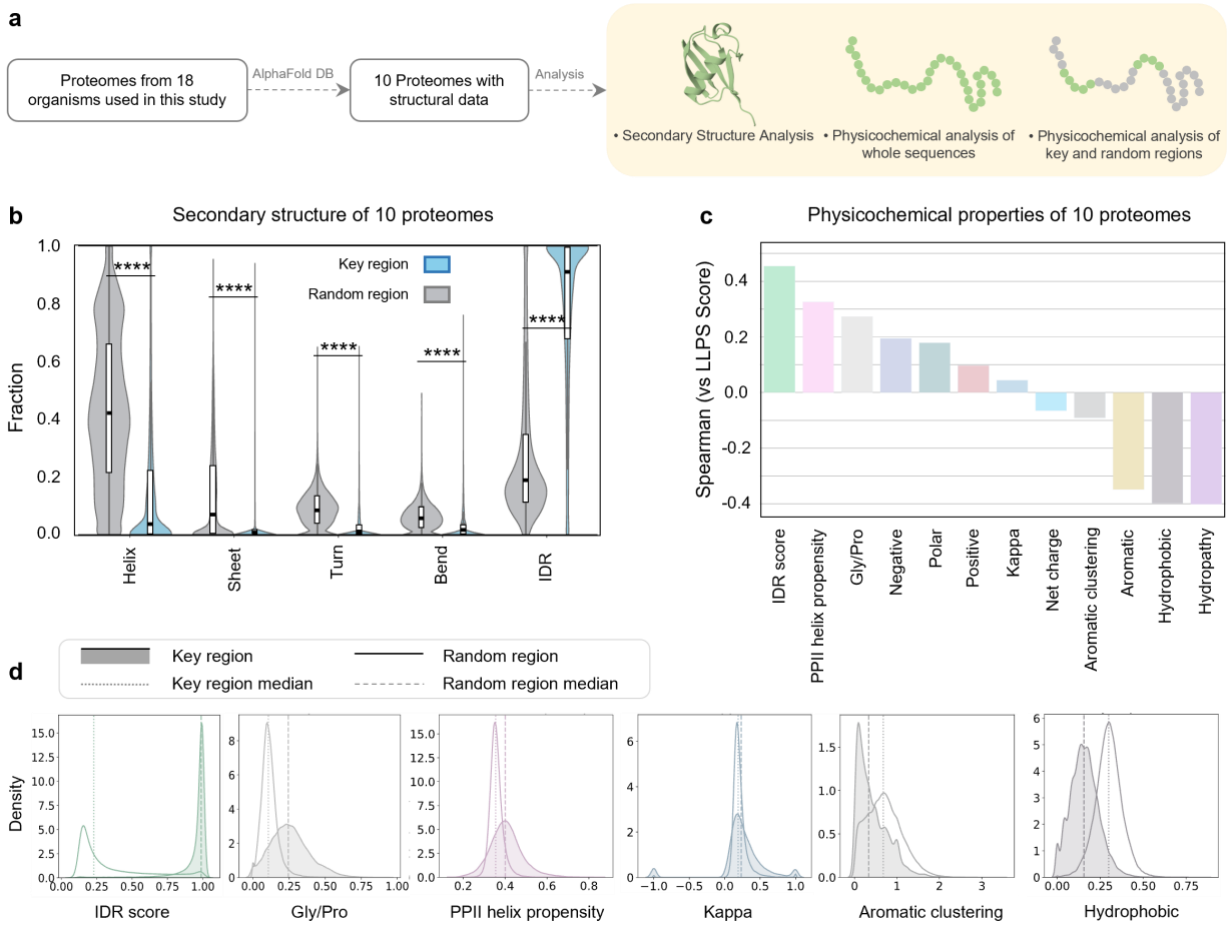

Fig. 4. Structural and physicochemical properties associated with LLPS across 10 proteomes. (a) Only 10 out of the 18 proteomes used in this study had predicted secondary structures available in the AlphaFold Protein Structure Database (AlphaFold DB). We analyzed secondary structures (b), physicochemical features of full-length sequences (c), and physicochemical features of key versus random regions (d) to investigate how these properties correlate with the protein LLPS propensity. **(b)** Violin plots compare secondary structure distributions for random and key regions across the 10 proteomes with AlphaFold data (*E. coli*, *P. falciparum*, *S. cerevisiae*, *A. thaliana*, *D. melanogaster*, *C. elegans*, *D. rerio*, *R. norvegicus*, *M. musculus*, and *H. sapiens*). A one-sided Mann–Whitney U test produced $p < 0.0001$ for all five structural classes. Boxes indicate the interquartile range (Q1–Q3) and medians. See **Supplementary Fig. 13** for secondary structure analysis of individual proteomes. **(c)** Spearman correlation coefficients between physicochemical feature s and Phaseek LLPS scores of the same 10 proteomes. See **Supplementary Fig. 14** for species-specific plots, as well as the combined 18-organism plot. **(d)** Density plots of physicochemical feature distributions in random versus key regions for the same 10 proteomes, with medians shown. One-sided Mann–Whitney U tests yielded $p < 0.0001$ for all comparisons shown. Distributions for all physicochemical features across both the 10 proteomes and the full set of 18 proteomes are shown in **Supplementary Fig. 15** and **Supplementary Fig. 16**, respectively. Feature labels in (c, d) use shorthand: aromatic (fraction of F, W, Y), hydrophobic (A, V, I, L, M), Gly/Pro (G and P), polar (S, T, N, Q, C, H), positive (K, R), and negative (D, E).

Discussion

Here, we introduced Phaseek, a sequence-based LLPS prediction model that integrates the transformers architecture, FEGS statistical matrices, and region-boosting bottleneck (**Fig. 1, Supplementary Fig. 1b**). We benchmarked Phaseek against existing LLPS predicting models

using data unseen by any of the models and corresponding to four distinct biological contexts: synthetic proteins (**Fig. 1d**), *A. thaliana* B-isox assay (**Fig. 1e**), fusion oncoproteins (**Fig. 1f**), and phase separating proteins with/without IDRs (**Fig. 1g**). While other models could discriminate a subset of the data with a comparable accuracy, Phaseek showed overall superior performance.

Despite being trained on amino acid sequences without explicit annotation of subsequences, Phaseek identifies key regions involved in LLPS and predicts the impact of point mutations (**Fig. 2**). For key regions, we challenged Phaseek with four human proteins, unseen to the model, for which LLPS regions had been experimentally determined. Phaseek outperformed other models in distinguishing key phase-separating regions from the rest of the protein sequence (**Fig. 2a-c, Supplementary Fig. 3**). For point mutations, Phaseek significantly distinguished impactful from random mutations and predicted whether they promote or weaken the LLPS propensity. This ability makes Phaseek a valuable tool for hypothesis-driven study of LLPS.

We applied Phaseek to study the evolutionary conservation of LLPS among orthologs and to identify associated molecular functions, biological processes, and cellular compartments across 18 proteomes. Our results suggest that LLPS tends to be evolutionarily conserved (**Fig. 3b**). Variations can occur; an ortholog might have been modified during evolution to meet species physiological conditions. For example, *D. melanogaster*'s BuGZ protein, unlike its orthologs, forms hydrogel-like material rather than liquid condensates[74]. Some biological functions identified by enrichment analysis are shared across all studied organisms and recapitulate processes known to involve LLPS and suggest novel areas of investigation (**Fig. 3c**).

It must be noted that the scores produced by Phaseek are not direct indicators of the extent of phase separation, condensate size, dynamics, or critical concentration across different sequences. Rather, they reflect the likelihood that a sequence harbors phase-separating region(s) and undergoes LLPS. Therefore, a sequence with a higher score is not necessarily expected to require fewer specific conditions for phase separation than the one with a lower score. However, if the scores in comparison are for mutated variants of the same sequence, they can be correlated with shifts in LLPS propensity. Proteins involved in LLPS are categorized into three major groups: scaffolds, regulators, and clients[97,98]. Scaffolds, which drive the formation of condensates, are further divided into self-phase-separating and partner-dependent types[97,98]. Clients are recruited into scaffold-driven condensates but do not independently initiate phase separation. Regulators influence LLPS behavior through mechanisms such as post-translational modifications. While Phaseek's training data included both self-phase-separating and partner-dependent sequences to enhance generalizability, future models may be developed for predicting LLPS within a specific subclass of these proteins. The discovery of diverse phase-separating proteins can lead to improvements in the predictive ability of models that is of interest to both fundamental and applied research.

Methods

Train, test and evaluation data

To construct our positive dataset, we leveraged LLPS dataset resources, including LLPSDB[98,99], DrLLPS[98,99], PhaSeDB[100], and PhaSePro[97], along with a manually curated dataset of 611 sequences. In total, this yielded 1,275 LLPS-positive polypeptides. For our negative dataset, we used 5,258 sequences compiled for PSPredictor[41].

We filtered out sequences exceeding a length of 512 and applied CD-HIT[101] (Cluster Database at High Identity with Tolerance) clustering with a similarity threshold of 0.9 to eliminate highly similar sequences. This resulted in a final dataset of 436 positive and 4,858 negative sequences (**Data availability**). Both datasets were then randomly split into training and test sets in a 9:1 ratio.

Due to the limited availability of phase-separating protein sequences, we employed an oversampling strategy to augment the size and diversity of our positive training set. IDRs are widely recognized as key drivers of LLPS[7]. To preserve these regions during data augmentation, we selectively introduced mutations only within the structured (folded) regions of protein sequences, leaving disordered regions intact. We trained a CategoricalHMM (categorical hidden markov model) using the hmmlearn library[102] on a separated set of folded regions from our phase-separating peptides (**Data availability**), with residues having AIUPred IDR scores of <0.5 considered folded.

The trained HMM was used to sample new amino acid sequences of the same length as each original sequence. We repeated this augmentation process multiple times per sequence, ultimately increasing the size of the positive set from 436 to a total of 4,151 sequences.

The positive evaluation set included sequences not used during training and tuning, as well as synthetic sequences compiled from five studies[51–55]. The negative evaluation set comprised unused sequences from the training split, supplemented with the PSPHunter[44] negative test set to increase diversity.

Folded/disordered (non-)phase-separating data

To identify folded phase-separating (folded PSPs) and disordered non-phase-separating (disordered-nonPSPs), we utilized AIUPred to calculate the average disorder score associated with each protein sequence across DrLLPS, PhaSePro, and PhaSeDB databases. Using a strict cut off of 0.8 on the average AIUPred IDR score, we filtered out the non-phase-separating sequences. For Folded-nonPSPs we considered 0.2 as threshold and selected the phase-separating sequences with IDR score below 0.2. Sequences overlapping with the train/test sets or failing CD-HIT filtering were excluded from further analysis.

Mutation data

We obtained 2,410 mutations assigned as missense mutations from the DisPhaseDB[66] database. These mutations influence LLPS in disease-associated proteins, although they have not been specifically classified as either LLPS-strengthening or LLPS-weakening in the database. Additionally, we gathered 207 mutations, labeled as LLPS-strengthening (n = 93) or LLPS-

weakening (n = 114) from DrLLPS, PSMutPred[69] and literature[60,70,103–108,109]. These resulted in a total of 2,617 impactful mutations from 137 distinct proteins.

For the background mutations used in comparative analysis, 2,740 random mutations were generated by introducing 20 random amino acid substitutions per sequence across the same 137 proteins. To assess the ability of Phaseek to distinguish impactful mutations from random ones, all impactful mutations including the 2,410 missense, 93 LLPS-strengthening, and 114 LLPS-weakening mutations were compared against the random mutation set. To further evaluate whether Phaseek can differentiate LLPS-strengthening from LLPS-weakening mutations, only the 93 strengthening and 114 weakening mutations were used.

Three distinct scores were defined to evaluate model performance to detect impactful mutations:

1. Δ(score): The difference in LLPS score between the mutated sequence and the corresponding wild-type sequence.

2. Δ(aa_score): The change in LLPS score for the single mutated residue before and after mutation.

3. Σ(Δ(aa_score)): The cumulative LLPS score difference across all residues whose individual LLPS propensity values changed as a result of the mutation, calculated only for positions where the absolute score difference exceeds a threshold of 1e-4.

Human Annexin A11 (P50995 · ANX11_HUMAN)[73,110] was selected to illustrate the effect of point mutations on residue-level scoring (**Fig. 2, Supplementary Fig. 10**). One strengthening and one weakening mutation were chosen for this demonstration.

Model architecture and training

Our core model consists of the encoder and the protein graph generator. The encoder module takes as input a protein sequence, as an array of one-dimensional token indices, and outputs an array of token embeddings. It is built based on a classical GPT architecture with 6 attention heads per layer and 6 layers in total. We employed a token embedding size of 192 and a block size of 512. The protein graph module operates in parallel with the encoder to generate graphical space representations of protein sequences. Each sequence is encoded into 10 distinct statistical matrices based on sorting methods derived from the FEGS method[46] (see Matrix selection and Protein graph sections below). After padding, these vectors are merged via weighted summation, producing a single extended matrix. In the final pooling step, outputs from the encoder and protein graph modules are integrated. The resulting output is then passed through a MLP that reduces the dimensionality of each token, followed by a median operation to yield a 2-dimensional probability vector. The integrated core model, comprising the encoder and the protein graph generator, was trained on the augmented training set (see Train, test and evaluation data section). Optimization was performed using the AdamW optimizer, with the cross entropy loss function applied to quantify prediction errors. The model demonstrated convergence after 35 epochs, achieving the minimum test loss at this point (**Code availability**).

Matrix selection

Each protein sequence in our dataset was encoded by the FEGS method[46] into 158 matrices, whose eigenvalues were concatenated into a 158-dimensional feature vector. A random forest classifier was subsequently trained on the training set based on the resulting vectors (**Code availability**). To evaluate the contribution of each of the 158 matrices in the scoring process, we employed the SHAP method, a game theoretic approach for interpreting the outputs of machine learning models. This analysis enabled us to quantify the relative contribution of each matrix by calculating its SHAP value. Based on this, we identified and selected the ten most impactful matrices during the scoring process (**Supplementary Table 1**).

Protein graph

We used the FEGS method[46] to construct the corresponding statistical matrices for each protein sequence. To leverage the physicochemical properties of amino acids, each amino acid is mapped to a unique point in Cartesian space based on the specific physicochemical properties selected from the AAIndex database[46], as described in Equation 1. The amino acids are then sorted and assigned in ascending order. For a given protein containing N amino acids as S = S_1, S_2, S_3, …, S_n, the relative position for each amino acid is calculated using Equation 2, where $\psi_{(0,0)}$ = (0, 0, 0), and $f_{\Omega 1,\Omega 2}$ denotes the frequency of the amino acid pair Ω_1, Ω_2 in the subsequence of the first i amino acids.

Equation 1: $\phi(\Omega_i) = (cos(\frac{2\pi i}{20}), sin(\frac{2\pi i}{20}), 1)$

Equation 2: $\psi(S_i) = \psi(S_{i-1}) + \phi(S_i) + \sum f_{\Omega_1,\Omega_2}.\varphi(\Omega_1, \Omega_2)$

Residue-level LLPS profile of proteins

We implemented a sliding window whose size varies based on protein length, ranging from a minimum of 5 to a maximum of 50 amino acids. This window slides across the full sequence, centering on each residue so that the target amino acid of interest lies at the midpoint. Subsequently, the window is divided into segments, with a length of half the window size. Finally, the LLPS scores for each segment are calculated by the core model and processed by a weighted summation layer to produce the final residue-level score. Weights are generated and assigned using an adjusted Gaussian distribution, giving higher importance to residues closer to the central amino acid (**Supplementary Fig. 1**).

To assess the models' ability to accurately distinguish LLPS-promoting regions from those with low-propensity, we defined a custom match scoring function focused on two key criteria: (i) the model's accuracy to classify regions as LLPS-promoting or low LLPS propensity, and (ii) its ability to produce a significant difference in scores between these two region types.

Residues were classified as LLPS-positive or LLPS-negative based on a score threshold of 0.5. The corresponding accuracy score was calculated according to the experimentally validated LLPS profiles at the residue level.

492 Additionally, to validate the significance of the differences in assigned LLPS scores, we calculated
493 the average score difference between LLPS-promoting and low-propensity regions. Finally, the
494 mean of the two scores was computed.

495 **Scoring method**

496 We introduced a boosted scoring mechanism to better capture the contribution of key regions to
497 LLPS. A gradient boosting classifier (XGBoost) was trained on residue-level LLPS scores
498 generated by the core model to leverage local phase separation propensities.

499 The final score is obtained by combining the scores of the core model and the XGBoost bottleneck
500 module. This strategy enhances the accuracy of LLPS prediction by identifying and emphasizing
501 the contributions of key regions involved in phase separation (**Supplementary Fig. 1**).

502 **Other LLPS predictors used in this study**

503 PSPHunter is a random forest model that utilizes 60 sequence-derived features including RNA-
504 DNA binding site affinity and post-translational modification sites, and is specialized in identifying
505 essential key residues in LLPS[44]. CatGranule 1.0 is a statistical algorithm that considers the
506 contributions of structural disorder, nucleic acid binding propensities, sequence length, and amino
507 acid content[49]. CatGranule ROBOT 2.0 can predict LLPS at single amino acid resolution using a
508 multi-layer perceptron model (MLP) based on physicochemical and structural features[43]. PScore is
509 a machine learning model trained to specifically capture pi-pi and pi-cation interactions that
510 promote LLPS[39]. PSPredictor is a word2vec-coded model trained using a gradient-boosting
511 decision tree[41]. Finally, PSTP (with two versions available) is an MLP model that uses the ESM-2
512 protein language model[111], and ALBATROSS IDR prediction model[112] to encode the input
513 sequences[50].

514 **Key region and random region derivation**

515 To define the key regions, we first smoothed the residue-level scores of each sequence with a
516 sliding window of size 7. Key regions were defined as consecutive segments where the smoothed
517 scores remained at or above 0.7 for a minimum of 20 residues. If the average score within any 7-
518 residue window dropped below 0.7, the region was considered discontinued.

519 Random regions, ranging from 20 to 250 residues in length, were sampled from each protein
520 sequence across the proteomes.

521 **Evaluation parameters**

522 To evaluate the model's performance and interpret the results, we employed several standard
523 metrics (TN, TP, FN and FP denote true negative, true positive, false negative, and false positive,
524 respectively):

525 - AUC (area under the ROC curve) assesses the ability of the classifier to differentiate positive from
526 negative classes regardless of the classification threshold. Higher AUC indicates better
527 performance (Equation 3).

528 Equation 3: $AUC = \int \frac{TP}{TP + FN} d(\frac{FP}{FP + TN})$

529 - Recall measures the capability of the model to identify the actual positive cases. It is defined as the
530 proportion of true-positives out of all actual positive cases (Equation 4).

531 Equation 4: $Recall = \frac{TP}{TP + FN}$

532 - Precision, or positive predictive value, measures the proportion of predicted positive cases that are
533 actually positive (Equation 5).

534 Equation 5: $Precision = \frac{TP}{TP + FP}$

535 - F1-Score combines precision and recall into a single metric, offering a balanced measure of
536 performance, especially useful when dealing with imbalanced data (Equation 6).

537 Equation 6: $F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$

538 **Secondary structure derivation**

539 Predicted structures of proteins for 10 proteomes were obtained in PDB format from the AlphaFold
540 Protein Structure Database[113]. Secondary structure information was extracted using the DSSP
541 (Dictionary of Secondary Structure in Proteins) package[114,115]. Structural elements were
542 categorized into five groups: helices (H, G, I, P), sheets (B, E), turns (T), bends (S), and irregular
543 or flexible regions (-), with the latter classified as disordered regions (**Supplementary Table 3**).

544 **Ortholog analysis**

545 The proteome of 18 organisms was obtained from the UniProt proteome database
546 (**Supplementary Table 4**), including *E. coli*, *P. falciparum*, *S. cerevisiae*, *A. thaliana*, *A. gambiae*,
547 *D. melanogaster*, *C. elegans*, *D. rerio*, *X. laevis*, *G. gallus*, *R. norvegicus*, *M. musculus*, *C. lupus*
548 *familiaris*, *B. taurus*, *S. scrofa*, *M. mulatta*, *P. troglodytes*, and *H. sapiens*.

549 For each organism, we used a single protein sequence per gene. The LLPS scores of these
550 sequences were predicted by Phaseek and assigned as phase-separating with a threshold of 0.7.
551 Using OrthoFinder[78–80], 328,887 genes (93.7% of the total) were grouped into 24,438 orthogroups.
552 Half of the genes belonged to orthogroups containing 18 or more genes (G50 = 18) (see **Data**
553 **availability**). Among these, 216 orthogroups contained sequences from all species, including 9
554 orthogroups composed exclusively of single-copy genes. The tree in Fig. 3a was generated using
555 Dendroscope[116].

556 **LLPS investigation in Topoisomerase II enzyme orthologs:**

557 We explored the orthogroups with members from all 18 organisms to identify protein sequences
558 with conserved LLPS properties across the tree of life. Only orthogroups containing a minimum of
559 nine LLPS-positive members were selected, resulting in 45 candidates. Among these, the
560 Topoisomerase II family stood out (min = 0.64, max = 0.76, median = 0.73). Given the enzyme's
561 fundamental role in DNA supercoiling resolution and genome organization during transcription and
562 replication, we investigated whether the elevated LLPS scores might reflect a conserved role for
563 phase separation in Topoisomerase II function. To investigate further, we conducted a literature
564 search and found that Topoisomerase II orthologs in two evolutionarily distant organisms,
565 *Saccharomyces cerevisiae* and *Homo sapiens*, have been shown to undergo phase separation[81].

566 **Functional enrichment**

567 To associate the LLPS phenomenon with biological processes, we predicted LLPS scores for all
568 proteins of seventeen organisms (**Supplementary Table 2**). For each organism, we filtered
569 proteins with a score of 0.7 or higher and enriched the corresponding gene sets for Gene Ontology
570 (GO) terms, which hierarchically represent biological processes.

571 We used the annoDb package of each organism to map genes and GO terms (**code availability**)
572 and performed conditional hypergeometric tests as implemented in the gostats[117] package
573 (v2.70.0) to ensure that a parent term was only reported if it is more enriched than any of its child
574 term.

575 To reduce the list of enriched terms, we applied semantic similarity using the rrvgo[118] package
576 (v1.16.0) with default method (Rel) and threshold (0.7). Enrichment results for each organism are
577 provided in a notebook together with the code to reproduce this analysis (**Data availability**).

578 **Physicochemical features**

579 Aromatic clustering[93] was calculated as

580 Equation 8: $Aromatic\ Clustering\ = \frac{1}{n}\sum_{i=1}^{n} P_i \quad where \quad P_i\ = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \cdots$

581 a, b, and c are the distances between aromatic residues. Proteome sequences with no aromatic
582 residues present were excluded from the analysis.

583 Disorder scores were calculated using AIUPred[88]

584 We used localCIDER[119] to calculate kappa (κ)[120], hydropathy, polyproline II helix (PPII) propensity,
585 net charge, and amino acid fractions in different categories including negative residue fraction
586 (nagative), positive residue fraction (positive), glycine and proline residue fraction (Gly/Pro), polar
587 residue fraction (polar), aromatic residue fraction (aromatic), and hydropathic residue fractions
588 (hydrophobic).

589 The charge blockiness parameter Kappa (κ)[120] quantifies the extent of charged amino acid mixing
590 (R, K, D, E) and is defined as

591 Equation 7: $\kappa = (\frac{\delta}{\delta_{max}})$

592 where δ defines the average deviation from charge symmetry across sequence regions (blobs)
593 using a sliding window (size of 5), and δ_{max} is the maximum δ for the sequence given its
594 composition. A κ value of 0 indicates charge disparity, while 1 represents maximum charge
595 blockiness. A negative κ value (*e.g.*, -1, see the minor peak in **Fig. 4d**) corresponds to uncharged
596 sequences.

597 Hydropathy is based on the Kyte-Doolittle hydrophobicity scale, where each amino acid is assigned
598 a hydropathy index reflecting its water-vapor transfer free energy change and its side-chain
599 distribution between the protein's interior and exterior[121].

600 PPII helix propensity is based on the Hisler's scale, whereby each amino acid is assigned a
601 polyproline propensity score measured using a peptide host-guest system and isothermal titration
602 calorimetry[122].

603 **Data availability**

604 The training and test data is available on https://github.com/AMIRMOHAMMAD-OSS/Phaseek,
605 along with the detailed pathway enrichment analysis data for each organism (associated with **Fig.**
606 **3c**). The predicted LLPS score of the 18 proteomes (associated with **Fig. 3, 4**), physicochemical
607 and structural data (**Fig. 4, Supplementary Fig. 13-16**) are available on
608 https://zenodo.org/records/16728261. Data underlying **Fig. 1b-g**, **Fig. 2a-e**, **Fig. 3c**, **Fig. 4b**,
609 **Supplementary Fig. 2-4**, **6-12** are provided as **Source Data File**.

610 **Code availability**

611 The code associated with the model, along with the trained model, is provided on
612 https://github.com/AMIRMOHAMMAD-OSS/Phaseek. The code for pathway enrichment analysis
613 is also available on the same GitHub repository. A user-friendly Google Colab notebook of Phaseek
614 is available through this link.

615 **Acknowledgments**

616 This work was supported by ATIP-Avenir young group leader program by INSERM-CNRS and
617 MOPGA (Make Our Planet Great Again) young researcher fellowship by the French government
618 (AP), and the Fondation Bettencourt Schueller (AP and ABL). The authors wish to thank Mostafa
619 Elraies, Negin Hadisadegh, and Roxana Ghasemi for their help in collecting initial training data and
620 Amir Reza Aliakbari and Sadrodin Barikbin for revising the code developed in this work.

621 **Author contributions**

622 AMM and HT gathered data, developed the model and performed the predictions and together with
623 AP analyzed the results. VG performed enrichment analyses. AMM and HT prepared the draft of
624 the manuscript with contributions from other authors. RN, ABL, and AP supervised the work.

625 **Competing interests**

626 The authors declare no competing interests.

627 **References**

628 1. Hyman, A. A. & Simons, K. Cell biology. Beyond oil and water--phase transitions in cells. *Science*
629 **337**, 1047–1049 (2012).

630 2. Martin, E. W. & Holehouse, A. S. Intrinsically disordered protein regions and phase separation:
631 sequence determinants of assembly or lack thereof. *Emerg Top Life Sci* **4**, 307–329 (2020).

632 3. Hyman, A. A., Weber, C. A. & Jülicher, F. Liquid-liquid phase separation in biology. *Annu. Rev.*
633 *Cell Dev. Biol.* **30**, 39–58 (2014).

634 4. Guo, Q., Shi, X. & Wang, X. RNA and liquid-liquid phase separation. *Noncoding RNA Res* **6**, 92–
635 99 (2021).

636 5. Rippe, K. Liquid-Liquid Phase Separation in Chromatin. *Cold Spring Harb. Perspect. Biol.* **14**,
637 (2022).

638 6. Su, Q., Mehta, S. & Zhang, J. Liquid-liquid phase separation: Orchestrating cell signaling through
639 time and space. *Mol. Cell* **81**, 4137–4146 (2021).

640 7. Dignon, G. L., Best, R. B. & Mittal, J. Biomolecular Phase Separation: From Molecular Driving
641 Forces to Macroscopic Properties. *Annu. Rev. Phys. Chem.* **71**, 53–75 (2020).

642 8. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of
643 cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).

644 9. Azaldegui, C. A., Vecchiarelli, A. G. & Biteen, J. S. The emergence of phase separation as an
645 organizing principle in bacteria. *Biophys. J.* **120**, 1123–1138 (2021).

646 10. Saeki, S., Kuwahara, N., Nakata, M. & Kaneko, M. Phase separation of poly(ethylene glycol)-
647 water-salt systems. *Polymer* **18**, 1027–1031 (1977).

648 11. Koningsveld, R., Kleintjens, L. A. & Shultz, A. R. Liquid–liquid phase separation in
649 multicomponent polymer solutions. IX. Concentration-dependent pair interaction parameter from
650 critical miscibility data on the system polystyrene–cyclohexane. *Journal of Polymer Science Part*
651 *A-2: Polymer Physics* **8**, 1261–1278 (1970).

652 12. Schaaf, P., Lotz, B. & Wittmann, J. C. Liquid-liquid phase separation and crystallization in binary

653 polymer systems. *Polymer* **28**, 193–200 (1987).

654 13. Brangwynne, C. P. *et al.* Germline P granules are liquid droplets that localize by controlled
655 dissolution/condensation. *Science* **324**, 1729–1732 (2009).

656 14. Hou, C. *et al.* PhaSepDB in 2022: annotating phase separation-related proteins with droplet
657 states, co-phase separation partners and other experimental information. *Nucleic Acids Res.* **51**,
658 D460–D465 (2023).

659 15. Wang, H. *et al.* Rubisco condensate formation by CcmM in β-carboxysome biogenesis. *Nature*
660 **566**, 131–135 (2019).

661 16. Freeman Rosenzweig, E. S. *et al.* The Eukaryotic CO-Concentrating Organelle Is Liquid-like and
662 Exhibits Dynamic Reorganization. *Cell* **171**, 148–162.e19 (2017).

663 17. Wunder, T. & Mueller-Cajar, O. Biomolecular condensates in photosynthesis and metabolism.
664 *Curr. Opin. Plant Biol.* **58**, 1–7 (2020).

665 18. Atkinson, N. *et al.* Introducing an algal carbon-concentrating mechanism into higher plants:
666 location and incorporation of key components. *Plant Biotechnol. J.* **14**, 1302–1315 (2016).

667 19. Wang, Y., Stessman, D. J. & Spalding, M. H. The CO2 concentrating mechanism and
668 photosynthetic carbon assimilation in limiting CO2 : how Chlamydomonas works against the
669 gradient. *Plant J.* **82**, 429–448 (2015).

670 20. Altmeyer, M. *et al.* Liquid demixing of intrinsically disordered proteins is seeded by poly(ADP-
671 ribose). *Nat. Commun.* **6**, 8088 (2015).

672 21. Levone, B. R. *et al.* FUS-dependent liquid-liquid phase separation is important for DNA repair
673 initiation. *J. Cell Biol.* **220**, (2021).

674 22. Wang, Y.-L. *et al.* Liquid-liquid phase separation in DNA double-strand breaks repair. *Cell Death*
675 *Dis.* **14**, 746 (2023).

676 23. Brangwynne, C. P., Mitchison, T. J. & Hyman, A. A. Active liquid-like behavior of nucleoli
677 determines their size and shape in Xenopus laevis oocytes. *Proc. Natl. Acad. Sci. U. S. A.* **108**,
678 4334–4339 (2011).

679 24. Li, P. *et al.* Phase transitions in the assembly of multivalent signalling proteins. *Nature* **483**, 336–
680 340 (2012).
681 25. Duchemin, A. *et al.* DHX15-independent roles for TFIP11 in U6 snRNA modification, U4/U6.U5
682 tri-snRNP assembly and pre-mRNA splicing fidelity. *Nat. Commun.* **12**, 6648 (2021).
683 26. Liao, S. E. & Regev, O. Splicing at the phase-separated nuclear speckle interface: a model.
684 *Nucleic Acids Res.* **49**, 636–645 (2021).
685 27. Noble, D. It's time to admit that genes are not the blueprint for life. *Nature* **626**, 254–255 (2024).
686 28. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science*
687 **357**, (2017).
688 29. Saqr, B. M. G. A. *et al.* Ageing drop by drop: Disturbance of the membrane-less organelle
689 biogenesis as an aging hallmark. *Biochem Biophys Res Commun* **742**, 151088 (2025).
690 30. Zbinden, A., Pérez-Berlanga, M., De Rossi, P. & Polymenidou, M. Phase Separation and
691 Neurodegenerative Diseases: A Disturbance in the Force. *Dev. Cell* **55**, 45–68 (2020).
692 31. Mehta, S. & Zhang, J. Liquid-liquid phase separation drives cellular function and dysfunction in
693 cancer. *Nat. Rev. Cancer* **22**, 239–252 (2022).
694 32. Boeynaems, S. *et al.* Protein Phase Separation: A New Phase in Cell Biology. *Trends Cell Biol.*
695 **28**, 420–435 (2018).
696 33. Yao, Y. *et al.* LncRNA-MEG3 Regulates Muscle Mass and Metabolic Homeostasis by Facilitating
697 SUZ12 Liquid-Liquid Phase Separation. *Adv Sci (Weinh)* **12**, e2417715 (2025).
698 34. Niu, X. *et al.* Biomolecular condensates: Formation mechanisms, biological functions, and
699 therapeutic targets. *MedComm (2020)* **4**, e223 (2023).
700 35. Delmore, J. E. *et al.* BET bromodomain inhibition as a therapeutic strategy to target c-Myc. *Cell*
701 **146**, 904–917 (2011).
702 36. Esmaili, F., Pourmirzaei, M., Ramazi, S., Shojaeilangari, S. & Yavari, E. A Review of Machine
703 Learning and Algorithmic Methods for Protein Phosphorylation Site Prediction. *Genomics*
704 *Proteomics Bioinformatics* **21**, 1266–1285 (2023).

705 37. Idhaya, T., Suruliandi, A. & Raja, S. P. A Comprehensive Review on Machine Learning
706 Techniques for Protein Family Prediction. *Protein J* **43**, 171–186 (2024).
707 38. Xu, Y. *et al.* Deep Dive into Machine Learning Models for Protein Engineering. *J Chem Inf Model*
708 **60**, 2773–2790 (2020).
709 39. Vernon, R. M. *et al.* Pi-Pi contacts are an overlooked protein feature relevant to phase
710 separation. *Elife* **7**, (2018).
711 40. Chen, Z. *et al.* Screening membraneless organelle participants with machine-learning models that
712 integrate multimodal features. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2115369119 (2022).
713 41. Chu, X. *et al.* Prediction of liquid-liquid phase separating proteins using machine learning. *BMC*
714 *Bioinformatics* **23**, 72 (2022).
715 42. Hou, S. *et al.* Machine learning predictor PSPire screens for phase-separating proteins lacking
716 intrinsically disordered regions. *Nat. Commun.* **15**, 2147 (2024).
717 43. Monti, M. *et al.* Accurate predictions of liquid-liquid phase separating proteins at single amino
718 acid resolution. *bioRxiv* (2024) doi:10.1101/2024.07.19.602785.
719 44. Sun, J. *et al.* Precise prediction of phase-separation key residues by machine learning. *Nat.*
720 *Commun.* **15**, 2662 (2024).
721 45. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv [cs.CL]* (2020)
722 doi:10.48550/ARXIV.2005.14165.
723 46. Mu, Z. *et al.* FEGS: a novel feature extraction model for protein sequences and its applications.
724 *BMC Bioinformatics* **22**, 297 (2021).
725 47. Tripathi, S. *et al.* Defining the condensate landscape of fusion oncoproteins. *Nat Commun* **14**,
726 6008 (2023).
727 48. Boulay, G. *et al.* Cancer-Specific Retargeting of BAF Complexes by a Prion-like Domain. *Cell*
728 **171**, 163–178.e19 (2017).
729 49. Bolognesi, B. *et al.* A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity
730 upon Increased Protein Expression. *Cell Rep* **16**, 222–231 (2016).

731 50. Feng, M. *et al.* PSTP: Decoding latent sequence grammar for protein phase separation through
732 transfer learning and attention. *bioRxiv* (2024) doi:10.1101/2024.10.29.620820.
733 51. Hilditch, A. T. *et al.* Assembling membraneless organelles from de novo designed proteins. *Nat*
734 *Chem* **16**, 89–97 (2024).
735 52. Quiroz, F. G. & Chilkoti, A. Sequence heuristics to encode phase behaviour in intrinsically
736 disordered protein polymers. *Nat Mater* **14**, 1164–1171 (2015).
737 53. Simon, J. R., Carroll, N. J., Rubinstein, M., Chilkoti, A. & López, G. P. Programming molecular
738 self-assembly of intrinsically disordered proteins containing sequences of low complexity. *Nat*
739 *Chem* **9**, 509–515 (2017).
740 54. MacEwan, S. R. *et al.* Phase Behavior and Self-Assembly of Perfectly Sequence-Defined and
741 Monodisperse Multiblock Copolypeptides. *Biomacromolecules* **18**, 599–609 (2017).
742 55. Ramšak, M. *et al.* Programmable de novo designed coiled coil-mediated phase separation in
743 mammalian cells. *Nat Commun* **14**, 7973 (2023).
744 56. Zhang, H. *et al.* Large-scale identification of potential phase-separation proteins from plants using
745 a cell-free system. *Mol Plant* **16**, 310–313 (2023).
746 57. Feng, Z., Jia, B. & Zhang, M. Liquid-Liquid Phase Separation in Biology: Specific Stoichiometric
747 Molecular Interactions vs Promiscuous Interactions Mediated by Disordered Sequences.
748 *Biochemistry* **60**, 2397–2406 (2021).
749 58. Zhang, H. *et al.* Liquid-liquid phase separation in biology: mechanisms, physiological functions
750 and human diseases. *Sci China Life Sci* **63**, 953–985 (2020).
751 59. Farahi, N., Lazar, T., Tompa, P., Meszaros, B. & Pancsa, R. Phase-separating fusion proteins
752 drive cancer by dysregulating transcription through ectopic condensates. (2023)
753 doi:10.1101/2023.09.20.558425.
754 60. Conicella, A. E., Zerze, G. H., Mittal, J. & Fawzi, N. L. ALS Mutations Disrupt Phase Separation
755 Mediated by α-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain. *Structure* **24**,
756 1537–1549 (2016).

757 61. Chandra, B. *et al.* Phase Separation Mediates NUP98 Fusion Oncoprotein Leukemic
758 Transformation. *Cancer Discov* **12**, 1152–1169 (2022).
759 62. Mahapatra, A. & Newberry, R. W. Liquid-liquid phase separation of α-synuclein is highly sensitive
760 to sequence complexity. *bioRxiv* (2023) doi:10.1101/2023.08.03.551831.
761 63. Newton, J. C. *et al.* Phase separation of the LINE-1 ORF1 protein is mediated by the N-terminus
762 and coiled-coil domain. *Biophys. J.* **120**, 2181–2191 (2021).
763 64. Kim, Y. R. *et al.* Prion-like domain mediated phase separation of ARID1A promotes oncogenic
764 potential of Ewing's sarcoma. *Nat. Commun.* **15**, 6569 (2024).
765 65. Boehning, M. *et al.* RNA polymerase II clustering through carboxy-terminal domain phase
766 separation. *Nat. Struct. Mol. Biol.* **25**, 833–840 (2018).
767 66. Kanaan, N. M., Hamel, C., Grabinski, T. & Combs, B. Liquid-liquid phase separation induces
768 pathogenic tau conformations in vitro. *Nat. Commun.* **11**, 2809 (2020).
769 67. Hu, J., Wang, G., Liu, X. & Gao, W. Enhancing Pharmacokinetics, Tumor Accumulation, and
770 Antitumor Efficacy by Elastin-Like Polypeptide Fusion of Interferon Alpha. *Adv Mater* **27**, 7320–
771 7324 (2015).
772 68. Navarro, A. M. *et al.* DisPhaseDB: An integrative database of diseases related variations in liquid-
773 liquid phase separation proteins. *Comput Struct Biotechnol J* **20**, 2551–2557 (2022).
774 69. Feng, M. *et al.* Decoding Missense Variants by Incorporating Phase Separation via Machine
775 Learning. *Nat Commun* **15**, 8279 (2024).
776 70. Mensah, M. A. *et al.* Aberrant phase separation and nucleolar dysfunction in rare genetic
777 diseases. *Nature* **614**, 564–571 (2023).
778 71. Lin, Y., Currie, S. L. & Rosen, M. K. Intrinsically disordered sequences enable modulation of
779 protein phase separation through distributed tyrosine motifs. *J Biol Chem* **292**, 19110–19120
780 (2017).
781 72. Wang, J. *et al.* A Molecular Grammar Governing the Driving Forces for Phase Separation of
782 Prion-like RNA Binding Proteins. *Cell* **174**, 688–699.e16 (2018).

783 73. Liao, Y.-C. *et al.* RNA Granules Hitchhike on Lysosomes for Long-Distance Transport, Using
784 Annexin A11 as a Molecular Tether. *Cell* **179**, 147–164.e20 (2019).
785 74. Chin, A. F., Zheng, Y. & Hilser, V. J. Phylogenetic convergence of phase separation and mitotic
786 function in the disordered protein BuGZ. *Protein Sci* **31**, 822–834 (2022).
787 75. Babl, L. *et al.* CTP-controlled liquid-liquid phase separation of ParB. *J Mol Biol* **434**, 167401
788 (2022).
789 76. Dasmeh, P. & Wagner, A. Natural Selection on the Phase-Separation Properties of FUS during
790 160 My of Mammalian Evolution. *Mol Biol Evol* **38**, 940–951 (2021).
791 77. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional
792 analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**, 635–645
793 (2004).
794 78. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
795 genomics. *Genome Biol* **20**, 238 (2019).
796 79. Emms, D. M. & Kelly, S. STRIDE: Species Tree Root Inference from Gene Duplication Events.
797 *Mol Biol Evol* **34**, 3267–3278 (2017).
798 80. Emms, D. M. & Kelly, S. STAG: Species tree inference from all genes. *bioRxiv* (2018)
799 doi:10.1101/267914.
800 81. Jeong, J., Lee, J. H., Carcamo, C. C., Parker, M. W. & Berger, J. M. DNA-Stimulated Liquid-
801 Liquid phase separation by eukaryotic topoisomerase ii modulates catalytic function. *Elife* **11**,
802 (2022).
803 82. Ho, W.-L. & Huang, J.-R. The return of the rings: Evolutionary convergence of aromatic residues
804 in the intrinsically disordered regions of RNA-binding proteins for liquid-liquid phase separation.
805 *Protein Sci* **31**, e4317 (2022).
806 83. Liu, F. & Wang, J. ATP Acts as a Hydrotrope to Regulate the Phase Separation of NBDY
807 Clusters. *JACS Au* **3**, 2578–2585 (2023).
808 84. Dang, M., Li, T. & Song, J. ATP and nucleic acids competitively modulate LLPS of the SARS-

809 CoV2 nucleocapsid protein. *Commun Biol* **6**, 80 (2023).

810 85. Zhu, Q., Wu, Y. & Luo, R. Understanding and fine tuning the propensity of ATP-driven liquid-
811 Liquid Phase Separation with oligolysine. *ChemRxiv* (2023) doi:10.26434/chemrxiv-2023-2fl3h.

812 86. Badaczewska-Dawid, A. E., Uversky, V. N. & Potoyan, D. A. BIAPSS: A Comprehensive
813 Physicochemical Analyzer of Proteins Undergoing Liquid-Liquid Phase Separation. *Int J Mol Sci*
814 **23**, (2022).

815 87. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–
816 589 (2021).

817 88. Erdős, G. & Dosztányi, Z. AIUPred: combining energy estimation with deep learning for the
818 enhanced prediction of protein disorder. *Nucleic Acids Res* **52**, W176–W181 (2024).

819 89. Uversky, V. N., Kuznetsova, I. M., Turoverov, K. K. & Zaslavsky, B. Intrinsically disordered
820 proteins as crucial constituents of cellular aqueous two phase systems and coacervates. *FEBS*
821 *Lett* **589**, 15–22 (2015).

822 90. Siligardi, G. & Drake, A. F. The importance of extended conformations and, in particular, the PII
823 conformation for the molecular recognition of peptides. *Biopolymers* **37**, 281–292 (1995).

824 91. Hicks, J. M. & Hsu, V. L. The extended left-handed helix: a simple nucleic acid-binding motif.
825 *Proteins* **55**, 330–338 (2004).

826 92. Miyagi, T. *et al.* The Patterning and Proportion of Charged Residues in the Arginine-Rich Mixed-
827 Charge Domain Determine the Membrane-Less Organelle Targeted by the Protein. *Int J Mol Sci*
828 **23**, (2022).

829 93. Holehouse, A. S., Ginell, G. M., Griffith, D. & Böke, E. Clustering of Aromatic Residues in Prion-
830 like Domains Can Tune the Formation, State, and Organization of Biomolecular Condensates.
831 *Biochemistry* **60**, 3566–3581 (2021).

832 94. McGraw, K. O., & Wong, S. P. A common language effect size statistic. Psychological Bulletin,
833 111.2, 361–365. doi:10.1037/0033-2909.111.2.361.

834 95. Lin, Y.-H. & Chan, H. S. Phase Separation and Single-Chain Compactness of Charged

835 Disordered Proteins Are Strongly Correlated. *Biophys J* **112**, 2043–2046 (2017).

836 96. Martin, E. W. *et al.* Valence and patterning of aromatic residues determine the phase behavior of
837 prion-like domains. *Science* **367**, 694–699 (2020).

838 97. Mészáros, B. *et al.* PhaSePro: the database of proteins driving liquid-liquid phase separation.
839 *Nucleic Acids Res* **48**, D360–D367 (2020).

840 98. Ning, W. *et al.* DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. *Nucleic*
841 *Acids Res* **48**, D288–D295 (2020).

842 99. Li, Q. *et al.* LLPSDB: a database of proteins undergoing liquid-liquid phase separation in vitro.
843 *Nucleic Acids Res* **48**, D320–D327 (2020).

844 100. You, K. *et al.* PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic*
845 *Acids Res* **48**, D354–D359 (2020).

846 101. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation
847 sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

848 102. GitHub - hmmlearn/hmmlearn: Hidden Markov Models in Python, with scikit-learn like API. *GitHub*
849 https://github.com/hmmlearn/hmmlearn.

850 103. Wang, L. *et al.* Rett syndrome-causing mutations compromise MeCP2-mediated liquid-liquid
851 phase separation of chromatin. *Cell Res* **30**, 393–407 (2020).

852 104. Li, M. *et al.* LLPS of FXR proteins drives replication organelle clustering for β-coronaviral
853 proliferation. *J Cell Biol* **223**, (2024).

854 105. Xu, B. *et al.* Distinct Effects of Familial Parkinson's Disease-Associated Mutations on α-Synuclein
855 Phase Separation and Amyloid Aggregation. *Biomolecules* **13**, (2023).

856 106. O'Carroll, A. *et al.* Pathological mutations differentially affect the self-assembly and
857 polymerisation of the innate immune system signalling adaptor molecule MyD88. *BMC Biol* **16**,
858 149 (2018).

859 107. Dao, T. P. *et al.* ALS-Linked Mutations Affect UBQLN2 Oligomerization and Phase Separation in
860 a Position- and Amino Acid-Dependent Manner. *Structure* **27**, 937–951.e5 (2019).

861 108. Kim, H. J. *et al.* Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem
862 proteinopathy and ALS. *Nature* **495**, 467–473 (2013).
863 109. Xie, Q. *et al.* Phase separation in cancer at a glance. *J Transl Med* **21**, 237 (2023).
864 110. Nahm, M. *et al.* mutations in ALS cause dysregulation of calcium homeostasis and stress granule
865 dynamics. *Sci Transl Med* **12**, (2020).
866 111. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model.
867 *Science* **379**, 1123–1130 (2023).
868 112. Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. Direct
869 prediction of intrinsically disordered protein conformational properties from sequence. *Nat*
870 *Methods* **21**, 465–476 (2024).
871 113. Database, A. P. S. AlphaFold Protein Structure Database. https://alphafold.ebi.ac.uk/download.
872 114. Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res*
873 **39**, D411–9 (2011).
874 115. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of
875 hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
876 116. Huson, D. H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees
877 and networks. *Syst Biol* **61**, 1061–1067 (2012).
878 117. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association.
879 *Bioinformatics* **23**, 257–258 (2007).
880 118. Sayols, S. rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms. *MicroPubl*
881 *Biol* **2023**, (2023).
882 119. Ginell, G. M. & Holehouse, A. S. Analyzing the Sequences of Intrinsically Disordered Regions
883 with CIDER and localCIDER. *Methods Mol Biol* **2141**, 103–126 (2020).
884 120. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by
885 linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci U S A* **110**,
886 13392–13397 (2013).

887 121. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J*
888 *Mol Biol* **157**, 105–132 (1982).
889 122. Elam, W. A., Schrank, T. P., Campagnolo, A. J. & Hilser, V. J. Evolutionary conservation of the
890 polyproline II conformation surrounding intrinsically disordered phosphorylation sites. *Protein Sci*
891 **22**, 405–417 (2013).
892 123. Picard, R. R. & Cook, R. D. Cross-validation of regression models. *J. Am. Stat. Assoc.* **79**, 575–
893 583 (1984).

894