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‡], Amir Pandi[1,3‡]

[1] Systems Engineering and Evolution Dynamics INSERM U1338, Laboratoire de Biologie Computationnelle et Quantitative CNRS UMR 7238, Sorbonne Université, Paris, France
[2] University of Washington, Seattle, WA, USA
[3] Paris Biofoundry, 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 allows the dynamic compartmentalization of biomolecules to regulate and enhance 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 diverse LLPS-prone proteins, as well as key regions and effects of point mutations within them. Proteome-wide analysis across eighteen species suggests that LLPS is evolutionary conserved and involved in multiple 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 versatile LLPS predictor for advancing fundamental and applied research in related fields.

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 in *Caenorhabditis elegans* were observed as biomolecular condensates made up of RNA and proteins playing a crucial role in germ cell development[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]. For example, 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[15–17]. LLPS also plays an important role in coordinating RNA metabolism by forming condensates that contain splicing factors and specific RNA molecules, which enhances the efficiency of RNA processing[13,18–21]. Importantly, LLPS plays a crucial role in human health, with dysregulation of phase separation processes implicated in various pathological conditions, including neurodegenerative disorders, cancer, and viral infections[22–24]. Thus, condensates may serve as potential therapeutic targets[25,26]. Photosynthesis as an ancient and one of the most fundamental biological processes also 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 the rest of the proteins involved in the process[27–31]. These condensates enable carbon-concentrating mechanisms and streamline the CO_2 fixation pathway. Such ubiquitous and important roles of LLPS in biological processes, from bacteria to humans, are reshaping our understanding of cellular organization and function[9,32,33]. Therefore, the prediction of phase-separating proteins can be impactful for both fundamental and applied research.

Over the course of evolution, a vast number of proteins have developed intricate sequence patterns that have enabled them to acquire diverse structures and functions. Taking this as an opportunity, machine learning models are increasingly being used for classification, regression, and generation tasks for various applications[34–36]. Given the widespread LLPS examples in biology, predictive models have been developed accordingly[37–41,42]. These models are based on pre-defined physicochemical and sequence features and often face limitations in generalizability across diverse proteins. To address this limitation, we develop Phaseek, a full sequence-based model, by combining transformers 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 also show Phaseek's ability to predict changes in LLPS scores due to point mutations. We then predict phase-separating proteins across eighteen organisms and investigate their conservation among orthologs as well as their

enrichment for Gene Ontology terms. Finally, we study the proteome-level structural and physicochemical features with respect to predicted LLPS.

We provided the weights of the Phaseek model and a user-friendly Google Colab notebook (**Code availability**).

Results

Phaseek, a model for prediction of LLPS from protein sequence

The core module of our model is an encoder based on Generative Pre-training Transformer (GPT)[43] (**Fig. 1a**). We compiled a dataset of ~1500 positive and ~5000 negative phase-separating sequences from major LLPS datasets and published literature (**Methods**). To address data imbalance, we augmented the positive training data to 3558 sequences using a Markov chain (**Methods**). The data was split into training and test datasets with a 9:1 ratio, and 300 sequences from the test set, along with a manually curated dataset of synthetic sequences, were used for validation.

We incorporated a second module of statistical matrices derived from proteins using FEGS (Feature Extraction based on Graphical and Statistical Features) method[44]. The FEGS method generates 158 unique graphical representations to encode proteins[44]. To optimize the time complexity of using these representations as statistical matrices, we used a random forest model and calculated SHAP (SHapley Additive exPlanations) values (**Methods**). This allowed us to assess the predictive importance of each matrix and select them accordingly (**Fig. 1a**). We selected and integrated 10 statistical matrices with the highest SHAP values in parallel to our GPT architecture (**Fig. 1a**).

Next, we added a third module, an XGBoost bottleneck module to boost the contribution of LLPS key regions to the final sequence score (**Supplementary Fig. 1**). We define key regions as sub-sequences with amino acids having high predicted scores (**Methods**). Key regions carry the required multivalency and interactions for phase separation, that are able to promote LLPS by fusing to other proteins[42,45,46]. These made up our final model combining the (i) encoder, (ii) protein graph, and (iii) bottleneck. From the first version only with the encoder to the model with three modules, AUC score (area under the curve of receiver operating characteristic, ROC) for our validation set was improved from 0.84 to 0.93 (**Supplementary Fig. 2**). The combined model named Phaseek, was used throughout this study for the prediction of LLPS proteins and their key regions from amino acid sequence.

Finaly, we compared our model with other available LLPS predictors, including PSPHunter[42], CatGranule 1.0[47], CatGranule ROBOT 2.0[41], PScore[37], PSPredictor[39], and PSTP[48](see **Methods** for information about each model). We evaluated the performances of these models on the validation dataset which consists of positive and negatives (**Methods**). Phaseek resulted in the highest AUC score (0.93) among all (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**).

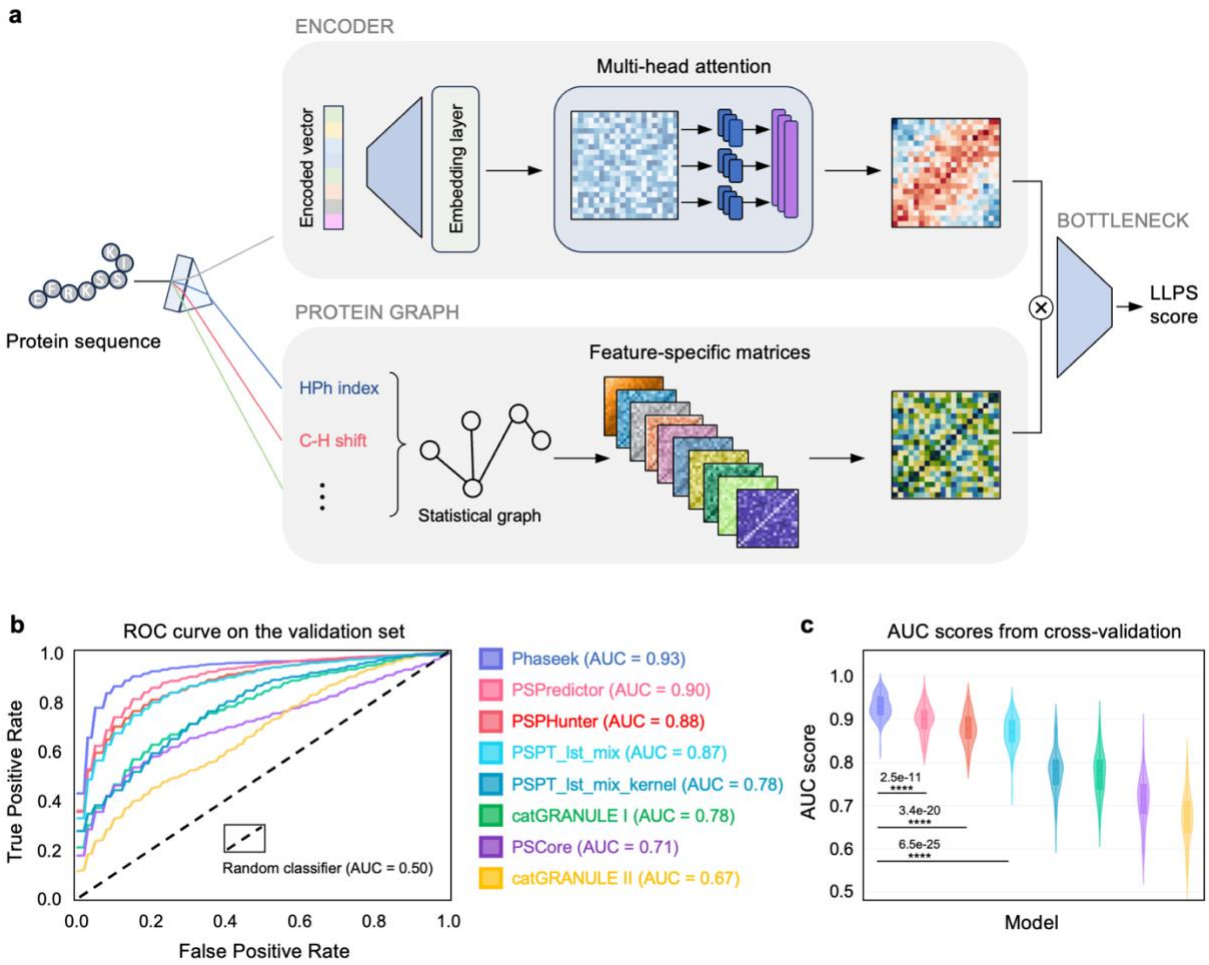

Fig. 1. Schematic representation of Phaseek architecture and model evaluation. (a) With protein sequences as input, Phaseek incorporates the contextual information (Encoder, based on Generative Pre-training Transformer, GPT) and statistical protein graph for LLPS prediction. **(b)** Performance of Phaseek and other models as ROC (Receiver Operating Characteristic) curves on the validation set comprising phase-separating (n = 300) and non-phase-separating (n = 337) sequences compiled from LLPS datasets along with a few studies, excluding our train and test datasets. **(c)** Violin plots illustrating the ROC area under the curve (AUC) scores for Phaseek and other models, evaluated on cross-validation tests. Batches containing 40 phase-separating and 40 non-phase-separating sequences from the validation set were selected and tested over 100 iterations. The p-values are for one-sided Mann-Whitney U test. Violin plots represent the distribution of data, and boxes show the interquartile range between the first and third quartiles, with the median as the middle line. Models are assigned with particular colors in the figure; Phaseek: periwinkle blue, PSPhunter: deep salmon, catGRANULE I: caribbean green, PScore: medium purple, catGRANULE II: saffron yellow, PSPredictor: coral pink, PSPT-LST-mix: bright sky blue, PSPT-LST-kernel: deep cerulean).

Phaseek is a generalizable model outperforming other LLPS predictors

To investigate the versatility and generalizability of Phaseek compared to other models, we benchmarked on four diverse biological datasets, not seen by our model. We assessed the classification accuracy of each model with an arbitrary threshold of 0.5 as models predict values between 0 and 1 or we scaled them in this range. The first set consisted of 97 phase-separating and 27 non-phase-separating synthetic proteins which have been experimentally tested for LLPS[49–53]. 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 set of LLPS proteins corresponded to a proteome study on cell extracts from *Arabidopsis thaliana* seedlings and inflorescences[54]. In reference work[54], intrinsically disordered proteins were precipitated using biotinylated isoxazole (B-isox) and their LLPS property was experimentally validated (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 consisted of fusion oncoproteins (FOs) derived from chromosomal translocations reported in ~17% of cancers which can facilitate oncogenesis through LLPS[45]. This data consists of 115 phase-separating and 63 non-phase-separating FOs. 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**).

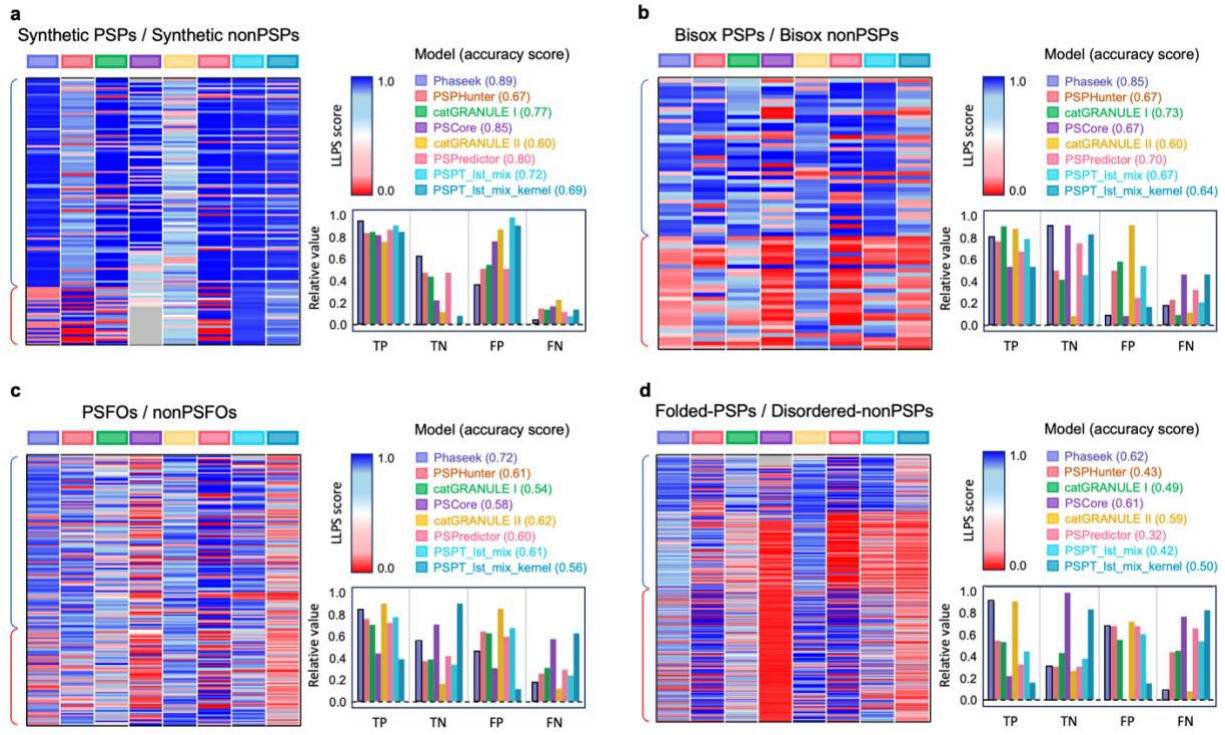

Fig. 2. **Phaseek outperformed state-of-the-art LLPS predictors.** Phaseek was compared to six other phase separation predictor models. Four datasets were used to ensure model comparison in terms of versatility and generalizability, **(a)** synthetic proteins (artificially designed proteins) containing synthetic phase-separating (synthetic PSPs, n=97) and synthetic non-phase-separating (synthetic nonPSPs, n=27) proteins, **(b)** biotinylated Isoxazole (B-isox)-determined proteins, phase-separating proteins (Bisox PSPs, n=43) and non-phase-separating proteins (Bisox nonPSPs, n=24), **(c)** fusion oncoproteins, phase-separating (PSFOs, n= 115) non-phase-separating (nonPSFOs, n=63), and **(d)** folded phase-separating proteins (Folded-PSPs, n=148) and disordered non-phase-separating proteins (Disordered-nonPSPs, n=185). Portions associated with the phase-separating and non-phase-separating sequences are determined by blue and red brackets. The predicted scores by catGRANULE I and PScore were scaled to the 0–1 range using the Sklearn standardscaler module to enable the comparison with other models. The bar charts illustrate the relative values of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) for the classification performance for each model (threshold = 0.5 is considered). The bar charts associated with Phaseek are highlighted. Models are assigned with particular colors in the figure; Phaseek: periwinkle blue, PSPhunter: deep salmon, catGRANULE I: caribbean green, PScore: medium purple, catGRANULE II: saffron yellow, PSPredictor: coral pink, PSPT-LST-mix: bright sky blue, PSPT-LST-kernel: deep cerulean). Protein sequences shorter than 140 amino acids, which cannot be classified by PScore, are represented in gray.

Finally, for the fourth set, we used 185 LLPS-negative proteins predicted as disordered and 148 LLPS-positive folded (non-disordered) proteins (**Methods**). Although phase separation principally occurs as a result of multivalent interactions in IDRs, folded proteins can also promote LLPS, and not all IDRs promote LLPS[55,56]. Phaseek performed the classification of phase-separating and non-phase-separating proteins with an accuracy of 0.62, more efficient than 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**).

Throughout these diverse biological data, Phaseek achieved generalizable superior performance in classifying phase-separating and non-phase-separating proteins.

Phaseek identifies key regions and mutation effects associated with LLPS

Multiple diseases and pathological conditions are related to mutations occurring in key regions of phase separation proteins[23,46,57–60]. Thus, the identification of LLPS key regions can have important implications in pathological conditions and in hypothesis-driven study of LLPS. To challenge the ability of Phaseek to detect LLPS key regions, we corroborated the results of four case studies (unseen to Phaseek) with experimental validation[61–63,64].

Comparing the residue-level LLPS score profiles of those well-studied sequences, Phaseek (followed by PSTP) outperformed other models in distinguishing LLPS from non-LLPS regions (**Fig. 3a**) 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. 3a, b**). Notably, removing residues 47-53 (segment A) and 132-152 (segment B) from the protein sequence reduced the overall LLPS score to 0.43 and 0.37, respectively, *i.e.*, below the classification threshold of 0.5 (**Fig. 3b**, **Supplementary Fig. 3**). This finding recapitulates experimental results that the deletion either of these segments disrupts LLPS phenotype[61].

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)[65], 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) (**Fig. 3c**). For the fused sequence of an artificially designed peptide (HERD 2,2), mCherry and flavin-containing monooxygenases (FMO)[49], 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 (**Fig. 3d**).

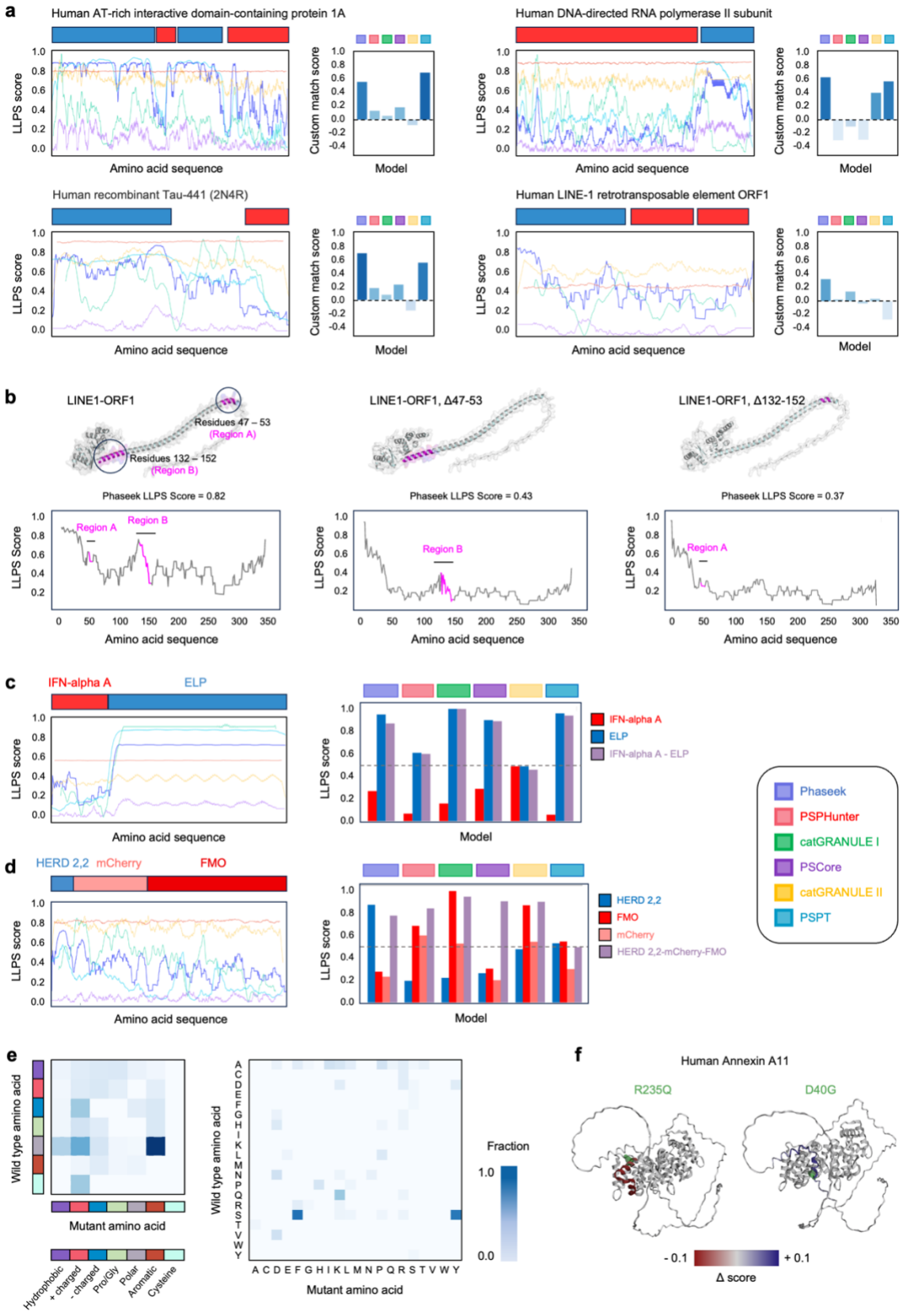
a
Human AT-rich interactive domain-containing protein 1A
Human DNA-directed RNA polymerase II subunit
Human recombinant Tau-441 (2N4R)
Human LINE-1 retrotransposable element ORF1
LLPS score
Amino acid sequence
Custom match score
Model
b
LINE1-ORF1
Residues 47 – 53
(Region A)
Residues 132 – 152
(Region B)
LINE1-ORF1, Δ47-53
LINE1-ORF1, Δ132-152
Phaseek LLPS Score = 0.82
Phaseek LLPS Score = 0.43
Phaseek LLPS Score = 0.37
Region A
Region B
LLPS Score
Amino acid sequence
c
IFN-alpha A
ELP
IFN-alpha A - ELP
d
HERD 2,2
mCherry
FMO
HERD 2,2-mCherry-FMO
Phaseek
PSPHunter
catGRANULE I
PSCore
catGRANULE II
PSPT
e
Wild type amino acid
Mutant amino acid
Fraction
Hydrophobic
+ charged
- charged
Pro/Gly
Polar
Aromatic
Cysteine
f
Human Annexin A11
R235Q
D40G
- 0.1
+ 0.1
Δ score

Fig. 3. Phaseek identifies LLPS key regions and mutation effects. (a) The residue-level LLPS prediction by different models and model comparison using a custom match score, bar charts, (**Methods**) on four human proteins unseen to Phaseek and with experimental identification of their phase-separating (blue) and non-phase-separating (red) regions. These proteins are AT-rich interactive domain-containing protein 1A (ARID1A), DNA-directed RNA polymerase II subunit, recombinant Tau protein, and LINE-1 retrotransposable element ORF-1. **(b)** Investigation of LINE1-ORF1 with two experimentally-determined segments, residues 47-53 (Region A) and residues 132-152 (Region B). The residue-level LLPS prediction of LINE1-ORF1 in 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). The structures were predicted by ESMFold[100]. **(c, d)** Whole sequence score variations along with the LLPS profiles as a result of the fusion of protein sequences to a sequence of interest. The residue-level and overall (bar charts) LLPS prediction for the fusion of **(c)** an elastin-like protein (ELP) and Interferon-alpha A (IFN-alpha A), and **(d)** synthetic phase-separating peptide HERD 2,2,mCherry and FMO (flavin-containing monooxygenases). **(e)** The frequency of mutations in terms of amino acid groups and individual amino acids for impactful mutations (see **Supplementary Fig. 3** for comparison with random mutation). **(f)** A case study (human Annexin A1, UniProtKB #P50995) on the predicted effect of two point mutations, one weakening (R235Q) and one strengthening (D40G) LLPS, as experimentally reported. Changes in the LLPS score of residues are shown in the structures for reduced (red) or increased (blue) LLPS probabilities. The mutated amino acid is shown in green. Structures were predicted by AlphaFold2[83]. The predicted scores by catGRANULE I and PScore **(a, c, d)** were scaled to the 0–1 range using the Sklearn standardscaler module to enable the comparison with other models. Models not used for the comparison **(a, c, d)** do not produce residue-level LLPS profiles. Models are assigned with particular colors in the figure; Phaseek: periwinkle blue, PSPhunter: deep salmon, catGRANULE I: caribbean green, PScore: medium purple, catGRANULE II: saffron yellow, PSPT-LST-kernel: deep cerulean).

Next, we sought to investigate the effect of point mutations on LLPS. We collected 2411 experimentally validated impactful mutations on 76 phase-separating proteins from multiple studies[66–68] (**Data availability**). Additionally, we introduced 40 random mutations per protein, in either random or in key regions, totaling 3040 mutations. For each mutation, we calculated the LLPS score difference between the mutant and the wild type. Phaseek discriminated impactful from Ineffective mutations (**Supplementary Fig. 4a**), and predicted whether an impactful mutation promotes or weakens LLPS (**Supplementary Fig. 4b**). Among the impactful mutations, polar to aromatic amino acids, specifically serine to both phenylalanine and tyrosine, showed the highest frequency (**Fig. 3e, Supplementary Fig. 4c, d**). As an example, we studied human Annexin II with experimentally reported mutations affecting its LLPS[69]. Annexin II tethers RNA granules to lysosomes facilitating long-distance RNA transport in neurons with mutational implications associated with amyotrophic lateral sclerosis (ALS). In line with the experimental results, Phaseek could capture the effect of both weakening (R235Q) and strengthening (D40G) LLPS (**Fig. 3f**)

Together these results demonstrate that, despite being trained on whole sequences without labeled regions, Phaseek can identify LLPS key regions and predict the effect of point mutations.

Phaseek reveals LLPS conservation among orthologs

Few studies have investigated the conservation of phase separation among the tree of life[70,71]. One study reported that the fused in sarcoma (FUS) protein has conserved key residues among orthologs required for phase separation[72]. However, whether LLPS is generally conserved among orthologs[71], *i.e.*, proteins among different organisms with the same function, has remained unexplored. To this end, we examined the proteomes of eighteen 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 trogglodytes*, and *Homo sapiens*. We reasoned that if phase-separation is a conserved feature among orthologs, all members of an orthogroup undergo LLPS in an all-or-none manner. Natobly, variations could arise if orthologs adapted during evolution to meet the specific physiological conditions of a species[70].

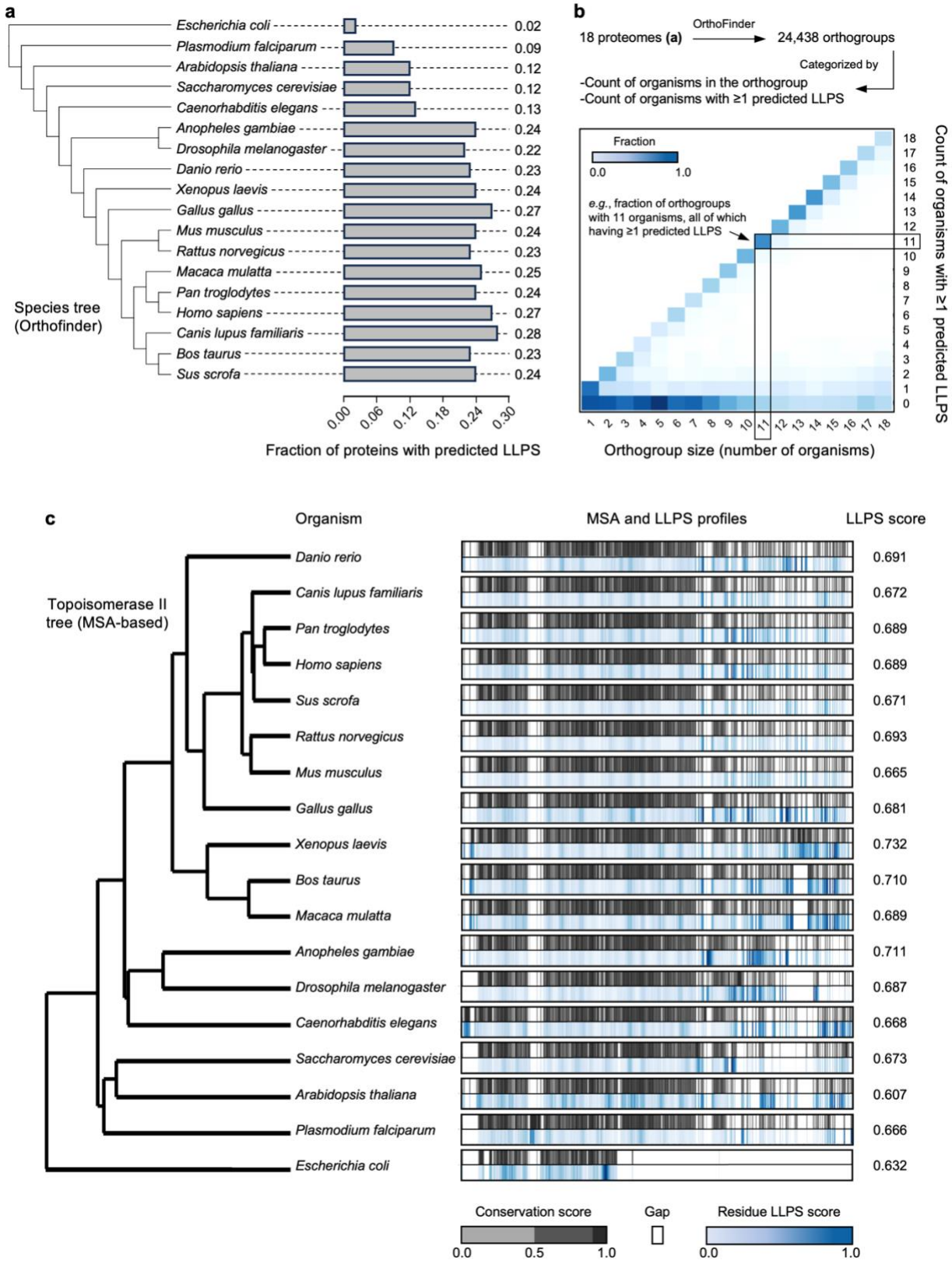

Fig. 4. Phaseek reveals LLPS conservation among orthologs among eighteen proteomes. (a) Fractions of proteins predicted to be phase-separating by Phaseek are shown along with the species tree generated by OrthoFinder[74–76]. **(b)** Orthologs extracted using OrthoFinder across

the eighteen proteomes to study the evolutionary conservation of LLPS. 24,438 orthogroups were achieved with sizes of 1 to 18, representing the number of organisms that each contains. Categorized orthogroups by (i) size count of organisms in each (orthogroup size) and (ii) count of organisms with at least one LLPS-predicted protein. For each category, the heatmap shows the fraction of groups with ≥1 predicted LLPS proteins for each organism (example provided in the figure). **(c)** The phylogenetic tree, generated using Dendroscope[112], based on Multiple sequence alignments (MSA) for topoisomerase II, across eighteen orthologs along with their Phaseek-predicted LLPS profiles and scores.

We quantified the fraction of predicted phase-separating proteins in each of the eighteen proteomes. *E. coli* was predicted to have the lowest fraction of LLPS proteins, 2%, while *S. cerevisiae* and *A. thaliana* resulted in the same fraction of 12%, and *H. sapiens* reached 27% (**Fig. 4a**). These results are consistent with the fact that eukaryotes are significantly more enriched in disordered proteins compared to prokaryotes[73], and thus are expected to contain relatively more phase-separating proteins. Using OrthoFinder, we categorized the proteome sequences into orthogroups[74–76] resulting in a total of 24,438 orthogroups among which 216 had ortholog contributions from every organism. All orthogroups were then sorted by orthogroup size and further categorized by the number of organisms contributing at least one LLPS-predicted ortholog (**Fig. 4b**). Notably, a species might contain more than one ortholog in an orthogroup**.** The marked increase in the fraction of LLPS-positive orthogroups near the distribution margins suggests an all-or-none conservation pattern, indicating that LLPS holds functional relevance across orthologs.

Among the orthogroups with contributions from all organisms was the Topoisomerase II enzyme (Topo II), which resolves DNA supercoiling and entanglements involved in transcription, replication, and DNA repair[77]. Topo II from *S. cerevisiae* and *H. sapiens* (TOP2α/TOP2β) have been reported to undergo LLPS with conserved ability to couple catenation to condensation[77]. Notably, the disordered LLPS-driving C-terminal domain (CTD) is less conserved in amino acid composition[77]. Low sequence similarity in disordered regions of orthologs has been previously reported[78]. For all of the eighteen organisms, Phaseek predicted the LLPS score consistently above 0.60 for Topo II orthologs (**Fig. 4c**). Multiple sequence alignment (MSA) analysis further revealed that the N-terminal ATPase and DNA cleavage domains of the orthologs are conserved among the species, despite their phase-separating CTDs (**Fig. 4c**). Phaseek assigned a high LLPS score to less conserved CTDs across all organisms indicating that the LLPS behavior can be maintained despite changes in the amino acid sequence.

Phaseek identifies biological functions enriched for LLPS

To associate LLPS with potential biological functions, we used proteomes of the eighteen organisms except *Plasmodium Falciparum* as its annotation package (org.Pf.plasmo.db on Bioconductor) is not maintained. For each of the remaining seventeen organisms, we enriched the set of LLPS genes identified by Phaseek for terms belonging to the three Gene Ontology (GO) categories: (i) molecular function, (ii) cellular component and (iii) biological process (**Methods**). We then applied semantic similarity to the results to reduce the list of enriched terms of each organism (**Supplementary Figs. 5-21).** Finally, we ranked common results according to the number of organisms in which they were enriched.

Among the three categories, biological process corresponded to the greatest number of annotated terms, followed by molecular function, and cellular compartment (**Fig. 5**). Interestingly, we found a greater number of enriched terms common to six or more organisms for molecular function than the two other

categories. Among all organisms, *E. coli* had the minimum enriched GO terms (**Supplementary Table 1**) consistent with this organism having the smallest fraction of the proteome predicted for LLPS (**Fig. 4a**).

In the category molecular function, the top-result ATP hydrolysis activity (GO:0016887), was shared across the seventeen organisms followed by ATP binding (GO:0005524), and nucleoside-triphosphatase regulator activity (GO:0060589), both shared across fifteen organisms. These results are in line with recent findings regarding the role of ATP in LLPS regulation[79–81]. Ranked in ninth position and shared across twelve organisms, the GO term catalytic activity acting on DNA (GO:0140097) covers the topoisomerase activity described in the previous section. Interestingly, in this category the term Organic cyclic compound binding (GO:0097159) was associated with the highest statistical significance but was found enriched in only six organisms (*E. coli*, *S. cerevisiae*, *D. melanogaster*, *C. elegans*, *M. musculus* and *H. Sapiens*) potentially reflecting some biases in the annotations. Overall, terms in this category were related to binding, coherent with the phenomenon of LLPS.

In the cellular compartment category, the term nucleus (GO:0005634) was enriched in fifteen organisms, with high statistical significance for *X. laevis*, *R. norvegicus*, *M. musculus*, *B. taurus* and *H. sapiens* while transcription regulator complex (GO:0005667) and MCM complex (GO:0042555) were found enriched across fourteen organisms with lower statistical significance. The other term to display high statistical significance is non-membrane-bounded organelle (GO:0043228) which is enriched in twelve organisms and with high statistical significance in *R. norvegicus*, *M. musculus* and *H. sapiens*. Finally, among the less conserved results are some specialized structures such as synaptic cleft (GO:0043083) or brush borders (GO:0005953).

In the category biological process, the top-result cellular component organization (GO:0016043) is enriched in thirteen organisms followed by homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156) which is enriched in twelve organisms. Interestingly, RNA biosynthetic process (GO:0032774), enriched in nine organisms, is associated with high statistical significance in *D. melanogaster*, *R. norvegicus*, *B. taurus* and *H. sapiens* while amino acid starvation (GO:0043038) is found in those organisms but associated with lower significance. Among less conserved terms are those associated with processes specific to higher organisms such as integrin-mediated signaling pathway (GO:0007229) but also generic ones such as nucleic acid transport (GO:0050657), or negative chemotaxis (GO:0050919). Finally, the last result DNA conformation change (GO:0071103) is enriched in *E. coli*, *S. cerevisiae*, *A. thaliana*, *A. gambiae*, *D. melanogaster*, *C. elegans* but was not found enriched in any vertebrate.

Biological functions identified by enrichment analysis recapitulate processes known to involve LLPS and suggest novel areas of investigation.

a
Number of Gene Ontology terms
Ontology
Molecular function
Cellular compartment
Biological process
15000
10000
5000
0
3848
5346
6697
4434
7924
6213
9317
5190
6561
15658
15784
9267
7019
9864
10738
12302
15541
b
ATP hydrolysis activity (GO:0016887)
ATP binding (GO:0005524)
nucleoside-triphosphatase regulator activity (GO:0060589)
ligase activity (GO:0016874)
ABC-type transporter activity (GO:0140359)
microtubule binding (GO:0008017)
chromatin binding (GO:0003682)
nuclear localization sequence binding (GO:0008139)
catalytic activity, acting on DNA (GO:0140097)
calmodulin binding (GO:0005516)
binding (GO:0005488)
beta-catenin binding (GO:0008013)
SH3 domain binding (GO:0017124)
small GTPase binding (GO:0031267)
purine ribonucleotide binding (GO:0032555)
mRNA binding (GO:0003729)
histone binding (GO:0042393)
modification-dependent protein binding (GO:0140030)
cadherin binding (GO:0045296)
adenylate cyclase activity (GO:0004016)
SMAD binding (GO:0046332)
RNA helicase activity (GO:0003724)
semaphorin receptor binding (GO:0030215)
phosphoprotein binding (GO:0051219)
molecular adaptor activity (GO:0060090)
microtubule motor activity (GO:0003777)
dynein light intermediate chain binding (GO:0051959)
DNA-binding transcription factor activity, RNA polymerase II-specific (GO:0000981)
protein self-association (GO:0043621)
polypeptide N-acetylgalactosaminyltransferase activity (GO:0004653)
phosphatidylinositol binding (GO:0035091)
dynein light chain binding (GO:0045503)
ubiquitin protein ligase activity (GO:0061630)
ubiquitin binding (GO:0043130)
sequence-specific DNA binding (GO:0043565)
protein kinase A binding (GO:0051018)
p53 binding (GO:0002039)
nuclear export signal receptor activity (GO:0005049)
inositol 1,4,5 trisphosphate binding (GO:0070679)
histone demethylase activity (GO:0032452)
transcription corepressor activity (GO:0003714)
protein demethylase activity (GO:0140457)
protein binding (GO:0005515)
organic cyclic compound binding (GO:0097159)
nuclear import signal receptor activity (GO:0061608)
metalloendopeptidase activity (GO:0004222)
ion binding (GO:0043167)
extracellular matrix structural constituent (GO:0005201)
extracellular matrix binding (GO:0050840)
carbohydrate derivative binding (GO:0097367)
basal transcription machinery binding (GO:0001098)
nucleus (GO:0005634)
transcription regulator complex (GO:0005667)
MCM complex (GO:0042555)
mismatch repair complex (GO:0032300)
intracellular anatomical structure (GO:0005622)
collagen trimer (GO:0005581)
chromatin (GO:0000785)
non-membrane-bounded organelle (GO:0043228)
exocyst (GO:0000145)
integrin complex (GO:0008305)
catenin complex (GO:0016342)
perinuclear region of cytoplasm (GO:0048471)
cytosol (GO:0005829)
cell cortex (GO:0005938)
basement membrane (GO:0005604)
adherens junction (GO:0005912)
synaptic cleft (GO:0043083)
supramolecular complex (GO:0099080)
plasma membrane (GO:0005886)
nuclear protein-containing complex (GO:0140513)
midbody (GO:0030496)
plasma membrane bounded cell projection (GO:0120025)
lamellipodium (GO:0030027)
clathrin-coated pit (GO:0005905)
brush border (GO:0005903)
apical plasma membrane (GO:0016324)
cellular component organization (GO:0016043)
homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156)
extracellular matrix organization (GO:0030198)
cellular process (GO:0009987)
intracellular signal transduction (GO:0035556)
amino acid activation (GO:0043038)
RNA biosynthetic process (GO:0032774)
sodium ion transmembrane transport (GO:0035725)
rhythmic process (GO:0048511)
negative regulation of biosynthetic process (GO:0009890)
establishment of cell polarity (GO:0030010)
cell division (GO:0051301)
biological regulation (GO:0065007)
protein-containing complex localization (GO:0031503)
post-transcriptional regulation of gene expression (GO:0010608)
nucleic acid transport (GO:0050657)
microtubule-based movement (GO:0007018)
intermediate filament-based process (GO:0045103)
endocytosis (GO:0006897)
cytoskeleton organization (GO:0007010)
centrosome localization (GO:0051642)
phosphorylation (GO:0016310)
phosphatidylinositol-mediated signaling (GO:0048015)
organic cyclic compound biosynthetic process (GO:1901362)
negative chemotaxis (GO:0050919)
intracellular receptor signaling pathway (GO:0030522)
integrin-mediated signaling pathway (GO:0007229)
cytoplasm organization (GO:0007028)
DNA recombination (GO:0006310)
DNA conformation change (GO:0071103)
-log (p-value)
120
80
40
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
Homo sapiens

Fig. 5. Gene Ontology (GO) terms enriched for predicted LLPS across seventeen organisms. (a) Number of annotated GO terms for each organism belonging in blue to the molecular function category, in green to the cellular compartment category and in red for the biological process category. **(b)** GO term enriched for each organism in the fraction of the proteome predicted for LLPS by Phaseek.

Phaseek unravels physicochemical and structural determinants of LLPS at proteome level

The LLPS phenotype is driven by amino acid sequences with particular physicochemical and structural properties[82]. We aimed to study these features on the entire human proteome predicted by Phaseek. We extracted secondary structures as well as a set of different physicochemical features from human proteome sequences to investigate their correlations with LLPS propensity (**Methods**).

First, we compared IDRs and secondary structures (helices, sheets, turns and bends extracted from AlphaFold[83]-predicted structures) between key and random regions (Methods). Random regions contained significantly more ordered structures such as helices than key regions (median fraction ~0.40 and ~0.03 respectively) (**Fig. 6a**). On the other hand, key regions contained significantly more disordered structures (median of ~0.92 compared to ~0.20 for random regions). This supports the widely accepted notion that IDRs are important drivers of LLPS.

Second, we computed Spearman correlations between 13 physicochemical features and the LLPS score of protein sequences. Some of the features used were specifically tailored to analyze IDRs, *e.g.*, disorder score and ABTscore[84]. The disorder score is the average of the disorder propensity for individual residues predicted by AIUPred[85]. ABTscore measures the extent of continuous sequences of either acidic or basic residues referred to as tracts found within IDRs.

ABTscore showed the highest correlation with LLPS scores (Spearman $\rho = 0.64$). Hydropathy, the fraction of hydrophobic residues and the fraction of aromatic residues showed inverse correlations ($\rho = -0.45$, -0.43, and -0.38 respectively) while polyproline II (PPII) helix propensity and glycine/proline fraction were positively correlated ($\rho = 0.38$ and 0.30) (**Fig. 6b**). We investigated whether these findings were consistent with previous studies. Disordered regions facilitate effective multivalent inter-/intra-molecular interactions[86] supported by the relatively high positive correlation between ABTscore and LLPS score. The lower number of hydrophobic and aromatic residues helps maintain protein disorder and liquid-like properties, while hydrophobic and aromatic interactions favor protein folding and interactions and (in extreme frequency) solid-like assemblies[7]. PPII helix propensity has also been reported to contribute to protein-protein interactions[87] and protein-nucleic acid interactions[88], acting as hubs for inter-/intra-molecular interactions. Glycine and proline residues have been experimentally found to be essential in phase separation[42].

Aromatic clustering, net charge, and charged amino acid patterning *e.g.*, Kappa (representing charge blockiness), had the lowest correlations ($\rho = -0.10$, 0,04 and 0.02, respectively). Mixed positive and negative residues in a sequence (*e.g.*, KEKEKEKEKE), result in lower Kappa (minimum of 0) and clustered charged residues (*e.g.*, KKKKKEEEEE) increase Kappa (maximum of 1). These results seemed counterintuitive as the distribution of charged and aromatic residues have been associated with phase separation[54,89,90], nonetheless, LLPS often depends only on specific regions and not the entire sequence.

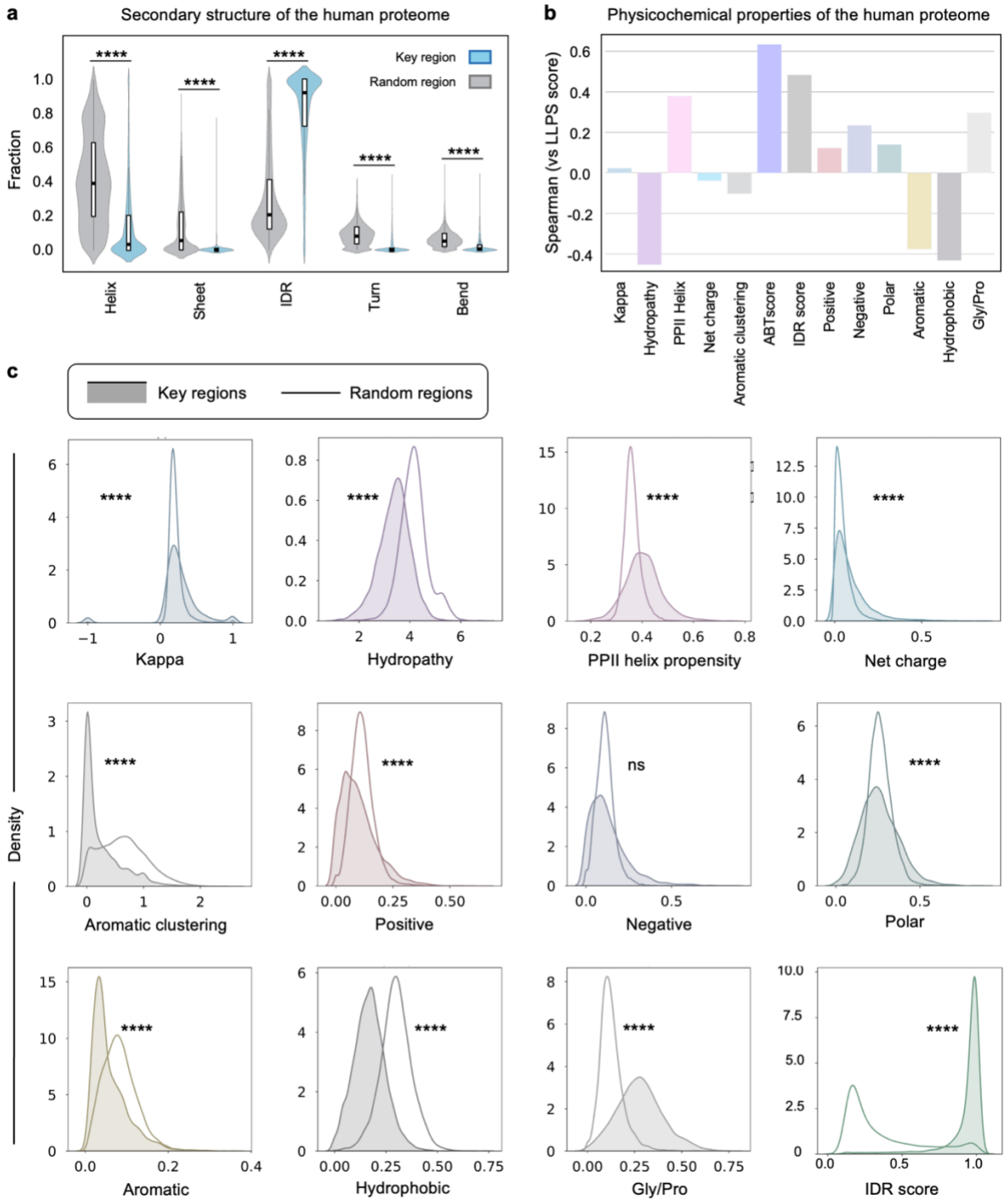

Fig. 6. Structure analysis and physicochemical properties of the human proteome with respect to LLPS. (a) Violin plots representing for random and key regions, the distribution of secondary structures of the human proteome (AlphaFold-predicted structures). The one-sided Mann-Whitney U test was used with ****p-value < 0.000 (~0) for all five. The boxes show the interquartile range between the first and third quartiles, with the median as the middle line. **(b)** Spearman correlation coefficients for the correlations of physicochemical features derived from the human proteome and their corresponding LLPS score predicted by Phaseek. **(c)** Distribution density of different physicochemical features for random and key regions of the human proteome. The one-sided Mann-Whitney U test was used with p-values as ****P < 0.0001 (n.s. = not significant).

Third, we therefore focused on the physicochemical analysis of LLPS-promoting regions (key regions) in the human proteome and compared them to random regions (**Methods**). Kappa and aromatic clustering showed significantly different distributions between key and random regions (**Fig. 6c**). Key regions exhibited higher Kappa values, indicating greater charge blockiness, which can enhance inter-/intramolecular interactions[89,91]. The differences in aromatic clustering suggest that key regions display a greater dispersion of aromatic residues, balancing between LLPS and aggregation[7,90,92].

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

Discussion

We introduce Phaseek, a sequence-based LLPS prediction model integrating transformer architecture, physicochemical statistical matrices and a scoring strategy (**Fig. 1**). We benchmarked the performance of Phaseek against existing LLPS prediction models in four distinct biological contexts with datasets unseen to Phaseek; synthetic proteins (**Fig. 2a**), *A. thaliana* B-isox assay (**Fig. 2b**), fusion oncoproteins (**Fig. 2c**), and phase separation with/without IDRs (**Fig. 2d**). While other models could classify a subset of these data with accuracy comparable to Phaseek, our model showed a generalizable superior performance. Particularly, Phaseek is more efficient in predicting negative datasets (**Fig. 2**).

Phaseek has been trained on amino acid sequences without annotation of local patterns, nonetheless, it can identify LLPS key regions and predict the effect of point mutations (**Fig. 3**). For the key regions, we challenged our model with four human proteins for which key LLPS regions had been experimentally determined. Phaseek outperformed other models in distinguishing key phase-separating regions from the rest of the protein sequence. For point mutations, Phaseek could significantly distinguish impactful mutations and whether they strengthen or weaken the LLPS propensity. This ability makes Phaseek a valuable tool for hypothesis-driven study of LLPS.

The generalizability of Phaseek allows for the evolutionary analysis of LLPS among orthologs and for the study of LLPS-enriched molecular functions, biological processes and cellular compartments. Our results suggest that LLPS can be evolutionarily conserved, following an all-or-none pattern across orthologs (**Fig. 4**). Variations can occur, and an ortholog might have been modified during evolution to meet species-specific physiological conditions. For example, *D. melanogaster*'s BuGZ protein, unlike its orthologs, forms hydrogel-like material rather than condensates through LLPS *in vitro*, while *in vivo*, the specific physiological conditions attenuate or enhance LLPS, although still less significantly than its orthologs[70]. Biological functions identified by enrichment analysis recapitulate processes known to involve LLPS and suggest novel areas of investigation.

It must be noted that the scores produced by Phaseek should not be interpreted as 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 can undergo LLPS. Therefore, a sequence with a higher score is not necessarily expected to require less specific conditions for phase separation than the one with a lower score. However, if the scores in comparison are for the mutated variants of the

same sequence, the scores can be correlated with shifts in LLPS propensity. The ever-growing size and diversity of phase-separating proteins in the future can further improve the predictive ability of Phaseek as a tool in both fundamental and applied research.

Methods

Train, test and validation data

To construct our positive dataset, we leveraged LLPS dataset resources, including LLPSDB[93,94], DrLLPS[93,94], PhaSeDB[95], and PhaSePro[96], along with a manually curated dataset of 611 sequences, totaling 1275 polypeptides. As for our negative dataset, we used the negative dataset collected for PSPredictor[39], which included 5258 sequences.

We filtered out sequences exceeding a length of 512 and applied CD-Hit[97] (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 4858 negative sequences (**Data availability**). Both datasets were then split into training and test sets in a 9:1 ratio.

Due to the limited availability of phase-separating sequences, we augmented the positive training dataset using a Markov Chain-based sequence generation approach, increasing the total to 3558 sequences.

For validation, the positive dataset included sequences not used during training, as well as synthetic sequences compiled from five studies[49–53]. The negative validation set comprised unused sequences from the training split, supplemented with the PSPHunter[42] negative test set to increase the diversity of the data.

Folded/disordered (non-)phase-separating data

For collecting folded phase-separating (folded PSPs) and disordered non-phase-separating (disordered-nonPSPs) we utilized AIUPred[85] to determine the average disorder score associated with each protein sequence across DrLLPS, PhaSePro and PhaSeDB databases. By considering the threshold as 0.8, we filtered out the non-phase-separating sequences with the average IDR score exceeding 0.8. For Folded-nonPSPs we considered 0.2 as threshold and selected the phase-separating sequences with IDR score below 0.2. Finally, all the sequences filtered by CD-hit[97] and the common sequences with our train and test datasets were removed.

Mutations data

2411 missense mutations impactful on LLPS were selected from 76 distinct proteins using data from DisPhaseDB[66], DrLLPS, and other sources[68,98]. For background mutations, 3040 random mutations were generated by introducing 40 random amino acid substitutions per sequence across the same 76 proteins. To evaluate the model in distinguishing LLPS-strengthening from LLPS-weakening mutations, the dataset collected by PSMutPred[67] was used, consisting of 79 strengthening and 166 weakening point mutations. Human Annexin A11 (P50995 · ANX11_HUMAN)[69,99] was selected to illustrate the effect of point mutations on residue-level scoring **(Fig. 3f**). One strengthening and one weakening mutation were chosen for this demonstration

Model architecture

Phaseek consists of three modules; encoder, protein graph and bottleneck. The Encoder Module processes protein sequence, encoded as an array of one-dimensional token indices, and outputs an array of token embeddings. The encoder module is built on a classical GPT architecture with 6 layers and 6 attention heads per layer. We employed the token embedding size of 192 and 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[44] (See matrix selection and protein graph sections below). After padding, these vectors are merged through weighted summation, producing a single extended matrix. In the final pooling step, outputs from these two modules are integrated. The resulting output is then passed through a bottleneck module comprising a fully connected network that reduces the dimensionality of each token to the latent size, followed by a median operation to yield a 2-dimensional probability vector.

Matrix selection

Each protein sequence in our dataset was encoded into a 158-dimensional feature vector using the FEGS method[44], using which a random forest classifier was subsequently trained on the training set (**Code availability**). To evaluate the contribution of each matrix in the scoring process, we employed the Shapley Additive Explanations (SHAP) method, a game theoretic method for interpreting the outputs of machine learning algorithms. This analysis enabled us to quantify the relative contribution of each matrix by calculating its SHAP value. From this, we identified and selected the ten most impactful matrices during the scoring process (**Supplementary Table 2**).

Protein graph

We used the FEGS method[44] to construct the corresponding statistical matrices for each protein sequence. To exploit the physicochemical properties of amino acids, each amino acid is mapped to a unique point on cartesian coordinates based on specific physicochemical properties selected from the AAIndex database[44] according to Equation 1. The amino acids are then sorted and assigned in ascending order. For a given protein containing N amino acids as $S = S_1, S_2, S_3, \ldots, S_n$, the relative position for each amino acid is calculated using Equation 2. Where $\psi_{(0,0)} = (0, 0, 0)$ and $f_{\Omega 1,\Omega 2}$ is the frequency of the amino acid pair $\Omega 1 \Omega 2$ in the subsequence of the first i amino acids of the protein sequence.

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 with a variable size that adjusts based on protein length, ranging from a minimum of 5 to a maximum of 50 amino acids. This window scans the full sequence, centering at each position so that the amino acid of interest is positioned at the midpoint. Subsequently, the window is divided into segments, with a length of half the window size. Finally, the LLPS scores associated with

these segments are calculated and passed through a weighted summation layer to produce the final score. Weights are generated and assigned using an adjusted Gaussian distribution to prioritize regions closer to the target amino acid (**Supplementary Fig. 1**).

To assess the performance of the models in accurately identifying key LLPS-promoting regions and low-LLPS propensity regions we defined a custom match scoring function focused on two key criteria: (i) the model's ability to classify the LLPS-promoting regions and regions with low LLPS propensity, and (ii) the model's capacity 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 is calculated according to the experimentally validated LLPS profile at the residue level.

Additionally, to validate the significance of the differences between the assigned LLPS scores for the two types of regions, we calculate the difference between the average scores designated to LLPS-promoting regions and regions with low LLPS propensity. Finally, the average of the two scores is calculated.

Scoring method

We introduced a boosted scoring mechanism, emphasizing the influences of the key regions on the LLPS of proteins. A gradient boosting classifier (XGBoost) was employed and trained on residue-level LLPS scores generated by the model to leverage residue-level LLPS propensities.

The final score is then obtained by combining the scores of Phaseek and the XGBoost model. This strategy enhances the final LLPS score of proteins by identifying and emphasizing the contributions of key regions involved in phase separation (**Supplementary Fig. 1**).

Other LLPS predictors used in this study

PSPHunter is a random forest model that utilizes 60 sequence-derived features including RNA-DNA binding site affinity and post-translational modification sites, and is specialized in identifying essential key residues in LLPS[42]. CatGranule 1.0 is a statistical algorithm that considers the contributions of structural disorder, nucleic acid binding propensities, sequence length, and amino acid content[47]. CatGranule ROBOT 2.0 can predict LLPS at single amino acid resolution using a multi-layer perceptron model (MLP) based on physicochemical and structural features[41]. PScore is a machine learning model trained to specifically capture pi-pi and pi-cation interactions that promote LLPS[37]. PSPredictor is a word2vec-coded model trained using a gradient-boosting decision tree[39]. Finally, PSTP (with two versions available) is an MLP model that uses the ESM-2 protein language model[100], and ALBATROSS IDR prediction model[101] to encode the input sequences[48].

Key region and random region derivation

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

Random regions were selected from the human proteome, with each protein contributing one randomly chosen segment. The only constraint applied to these regions was a length between 20 and 250 residues.

Evaluation parameters

To evaluate the performance of the model and interpretability of the results, we employed several standard metrics (TN, TP, FN and FP denote true negative, true positive, false negative and false positive, respectively):

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

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

- Recall is a relative metric that depicts the capability of the model to identify the actual positive data. It is defined as the proportion of the correctly predicted true cases out of all positive data (Equation 4).

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

- Precision, also known as positive predictive value, measures the proportion of predicted positive cases that are actually positive (Equation 5).

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

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

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

Secondary structure derivation

The predicted structures of human proteome proteins were obtained from the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/download) in PDB format. Secondary structure information was extracted using DSSP[102,103]. The structural elements were categorized into five groups: helices (H, G, I, P), sheets (B, E), turns (T), bends (S), and irregular or flexible regions (-), with the latter classified as disordered regions. (**Supplementary Table 3**).

Ortholog analysis

The proteome of eighteen organisms, including *E. coli*, *P. falciparum*, *S. cerevisiae*, *A. thaliana*, *A. gambiae*, *D. melanogaster*, *C. elegans*, *D. rerio*, *X. laevis*, *G. gallus*, *R. norvegicus*, *M. musculus*, *C. lupus familiaris*, *B. taurus*, *S. scrofa*, *M. mulatta*, *P. trogglodytes*, *H. sapiens* was obtained from the UniProt proteome database (**Supplementary Table 4**). For each organism, we used a single protein

sequence per gene. The LLPS scores of these sequences were predicted by Phaseek and assigned as phase-separating with a threshold of 0.5. The species three in **Fig. 4a** was generated by OrthoFinder[74–76].

Using OrthoFinder[74–76] 328,887 genes (93.7% of the total) were grouped into 24,438 orthogroups. Half of the genes belonged to orthogroups containing 18 or more genes (G50 = 18) (**Data availability**). Among these, 216 orthogroups contained sequences from all species, including 9 orthogroups composed exclusively of single-copy genes.

Functional enrichment

To associate LLPS phenomenon with biological processes, we predicted the scores for all proteins of seventeen organisms (**Data availability**). For each organism, we further enriched the set of positive proteins for Gene Ontology (GO) terms using gostats[104] v2.70.0 and the annoDb package corresponding to each model organism. We then used rrvgo[105] v1.16.0 and semantic similarity to further reduce the list of enriched terms for each organism (**Code availability**). Enrichment results for each organism are available (**Supplementary Figs. 5-21**, **Data availability**)

Physicochemical features

ABTscores were directly adopted from its corresponding study. ABTscore is calculated by first identifying the disordered regions of the sequence of interest using IUPred[106,107]. For each disordered region being longer than 30 residues, the average net-charge-per-residue (NCPR) is calculated. Charge blocks within these regions, where the product of the number of residues and the average NCPR exceeds 1, are identified. The area of each charge block (area = number of residues X average NCPR) is then multiplied by (0.6 + κ), where κ is the charge blockiness parameter for the respective disordered region. The resulting value is termed RegionScore, and the sum of all RegionScores across the disordered regions yields the ABTscore.

Aromatic clustering[90] was calculated as

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

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

Disorder Scores were calculated using AIUPred[85]

We used localCIDER[108] to calculate the rest of thel features, including Kappa (κ)[109], Hydropathy, Polyproline II Helix (PPII) Propensity, Net Charge, Negative Residue Fraction, Positive Residue Fraction.

The charge blockiness parameter Kappa (κ)[109] quantifies the extent of charged amino acid mixing (R, K, D, E) and is defined as

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

where δ defines the average deviation from charge symmetry across sequence regions (blobs) using a sliding window (size of 5), and δ_{max} is the maximum δ for the sequence given its composition. A κ value of 0 indicates charge disparity, while 1 represents maximum charge blockiness. A negative κ value (*e.g.*, -1, observed as a minor peak in **Fig. 6c**) corresponds to uncharged sequences.

Hydropathy is based on the Kyte-Doolittle hydrophobicity scale, where each amino acid is assigned a hydropathy index reflecting its water-vapor transfer free energy change and its side-chain distribution between the protein's interior and exterior[110].

PPII Propensity is based on the Hisler's scale, whereby each amino acid is assigned a polyproline propensity score measured using a peptide host-guest system and isothermal titration calorimetry[111].

Code and data availability

Will be available with the revised manuscript.

Acknowledgments

This work was supported by ATIP-Avenir young group leader program by INSERM-CNRS and MOPGA (Make Our Planet Great Again) young researcher fellowship by the French government (AP), and the Fondation Bettencourt Schueller (AP and ABL). The authors wish to thank Mostafa Elraies, Negin Hadisadegh, and Roxana Ghasemi for their help in collecting initial training data.

Author contributions

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

Competing interests

The authors declare no competing interests.

References

1. Hyman, A. A. & Simons, K. Cell biology. Beyond oil and water--phase transitions in cells. *Science* **337**, 1047–1049 (2012).
2. Martin, E. W. & Holehouse, A. S. Intrinsically disordered protein regions and phase separation: sequence determinants of assembly or lack thereof. *Emerg Top Life Sci* **4**, 307–329 (2020).
3. Hyman, A. A., Weber, C. A. & Jülicher, F. Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.* **30**, 39–58 (2014).
4. Guo, Q., Shi, X. & Wang, X. RNA and liquid-liquid phase separation. *Noncoding RNA Res* **6**, 92–99 (2021).
5. Rippe, K. Liquid-Liquid Phase Separation in Chromatin. *Cold Spring Harb. Perspect. Biol.* **14**, (2022).
6. Su, Q., Mehta, S. & Zhang, J. Liquid-liquid phase separation: Orchestrating cell signaling through time and space. *Mol. Cell* **81**, 4137–4146 (2021).
7. Dignon, G. L., Best, R. B. & Mittal, J. Biomolecular Phase Separation: From Molecular Driving Forces to Macroscopic Properties. *Annu. Rev. Phys. Chem.* **71**, 53–75 (2020).
8. 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).
9. Azaldegui, C. A., Vecchiarelli, A. G. & Biteen, J. S. The emergence of phase separation as an organizing principle in bacteria. *Biophys. J.* **120**, 1123–1138 (2021).
10. Saeki, S., Kuwahara, N., Nakata, M. & Kaneko, M. Phase separation of poly(ethylene glycol)-water-salt systems. *Polymer* **18**, 1027–1031 (1977).
11. Koningsveld, R., Kleintjens, L. A. & Shultz, A. R. Liquid–liquid phase separation in multicomponent polymer solutions. IX. Concentration-dependent pair interaction parameter from critical miscibility data on the system polystyrene–cyclohexane. *Journal of Polymer Science Part A-2: Polymer Physics* **8**, 1261–1278 (1970).

12. Schaaf, P., Lotz, B. & Wittmann, J. C. Liquid-liquid phase separation and crystallization in binary polymer systems. *Polymer* **28**, 193–200 (1987).
13. Brangwynne, C. P. *et al.* Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **324**, 1729–1732 (2009).
14. Hou, C. *et al.* PhaSepDB in 2022: annotating phase separation-related proteins with droplet states, co-phase separation partners and other experimental information. *Nucleic Acids Res.* **51**, D460–D465 (2023).
15. Altmeyer, M. *et al.* Liquid demixing of intrinsically disordered proteins is seeded by poly(ADP-ribose). *Nat. Commun.* **6**, 8088 (2015).
16. Levone, B. R. *et al.* FUS-dependent liquid-liquid phase separation is important for DNA repair initiation. *J. Cell Biol.* **220**, (2021).
17. Wang, Y.-L. *et al.* Liquid-liquid phase separation in DNA double-strand breaks repair. *Cell Death Dis.* **14**, 746 (2023).
18. Brangwynne, C. P., Mitchison, T. J. & Hyman, A. A. Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4334–4339 (2011).
19. Li, P. *et al.* Phase transitions in the assembly of multivalent signalling proteins. *Nature* **483**, 336–340 (2012).
20. Duchemin, A. *et al.* DHX15-independent roles for TFIP11 in U6 snRNA modification, U4/U6.U5 tri-snRNP assembly and pre-mRNA splicing fidelity. *Nat. Commun.* **12**, 6648 (2021).
21. Liao, S. E. & Regev, O. Splicing at the phase-separated nuclear speckle interface: a model. *Nucleic Acids Res.* **49**, 636–645 (2021).
22. 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).
23. Mehta, S. & Zhang, J. Liquid-liquid phase separation drives cellular function and

dysfunction in cancer. *Nat. Rev. Cancer* **22**, 239–252 (2022).

24. Boeynaems, S. *et al.* Protein Phase Separation: A New Phase in Cell Biology. *Trends Cell Biol.* **28**, 420–435 (2018).
25. Niu, X. *et al.* Biomolecular condensates: Formation mechanisms, biological functions, and therapeutic targets. *MedComm (2020)* **4**, e223 (2023).
26. Delmore, J. E. *et al.* BET bromodomain inhibition as a therapeutic strategy to target c-Myc. *Cell* **146**, 904–917 (2011).
27. Wang, H. *et al.* Rubisco condensate formation by CcmM in β-carboxysome biogenesis. *Nature* **566**, 131–135 (2019).
28. Freeman Rosenzweig, E. S. *et al.* The Eukaryotic CO-Concentrating Organelle Is Liquid-like and Exhibits Dynamic Reorganization. *Cell* **171**, 148–162.e19 (2017).
29. Wunder, T. & Mueller-Cajar, O. Biomolecular condensates in photosynthesis and metabolism. *Curr. Opin. Plant Biol.* **58**, 1–7 (2020).
30. Atkinson, N. *et al.* Introducing an algal carbon-concentrating mechanism into higher plants: location and incorporation of key components. *Plant Biotechnol. J.* **14**, 1302–1315 (2016).
31. Wang, Y., Stessman, D. J. & Spalding, M. H. The CO2 concentrating mechanism and photosynthetic carbon assimilation in limiting CO2 : how Chlamydomonas works against the gradient. *Plant J.* **82**, 429–448 (2015).
32. Noble, D. It's time to admit that genes are not the blueprint for life. *Nature* **626**, 254–255 (2024).
33. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **357**, (2017).
34. Esmaili, F., Pourmirzaei, M., Ramazi, S., Shojaeilangari, S. & Yavari, E. A Review of Machine Learning and Algorithmic Methods for Protein Phosphorylation Site Prediction. *Genomics Proteomics Bioinformatics* **21**, 1266–1285 (2023).
35. Idhaya, T., Suruliandi, A. & Raja, S. P. A Comprehensive Review on Machine Learning

Techniques for Protein Family Prediction. *Protein J* **43**, 171–186 (2024).

36. Xu, Y. *et al.* Deep Dive into Machine Learning Models for Protein Engineering. *J Chem Inf Model* **60**, 2773–2790 (2020).
37. Vernon, R. M. *et al.* Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife* **7**, (2018).
38. 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).
39. Chu, X. *et al.* Prediction of liquid-liquid phase separating proteins using machine learning. *BMC Bioinformatics* **23**, 72 (2022).
40. Hou, S. *et al.* Machine learning predictor PSPire screens for phase-separating proteins lacking intrinsically disordered regions. *Nat. Commun.* **15**, 2147 (2024).
41. Monti, M. *et al.* Accurate predictions of liquid-liquid phase separating proteins at single amino acid resolution. *bioRxiv* (2024) doi:10.1101/2024.07.19.602785.
42. Sun, J. *et al.* Precise prediction of phase-separation key residues by machine learning. *Nat. Commun.* **15**, 2662 (2024).
43. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv [cs.CL]* (2020) doi:10.48550/ARXIV.2005.14165.
44. Mu, Z. *et al.* FEGS: a novel feature extraction model for protein sequences and its applications. *BMC Bioinformatics* **22**, 297 (2021).
45. Tripathi, S. *et al.* Defining the condensate landscape of fusion oncoproteins. *Nat Commun* **14**, 6008 (2023).
46. Boulay, G. *et al.* Cancer-Specific Retargeting of BAF Complexes by a Prion-like Domain. *Cell* **171**, 163–178.e19 (2017).
47. Bolognesi, B. *et al.* A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity upon Increased Protein Expression. *Cell Rep* **16**, 222–231 (2016).

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

59. Chandra, B. *et al.* Phase Separation Mediates NUP98 Fusion Oncoprotein Leukemic Transformation. *Cancer Discov* **12**, 1152–1169 (2022).
60. Mahapatra, A. & Newberry, R. W. Liquid-liquid phase separation of α-synuclein is highly sensitive to sequence complexity. *bioRxiv* (2023) doi:10.1101/2023.08.03.551831.
61. Newton, J. C. *et al.* Phase separation of the LINE-1 ORF1 protein is mediated by the N-terminus and coiled-coil domain. *Biophys. J.* **120**, 2181–2191 (2021).
62. Kim, Y. R. *et al.* Prion-like domain mediated phase separation of ARID1A promotes oncogenic potential of Ewing's sarcoma. *Nat. Commun.* **15**, 6569 (2024).
63. Boehning, M. *et al.* RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nat. Struct. Mol. Biol.* **25**, 833–840 (2018).
64. Kanaan, N. M., Hamel, C., Grabinski, T. & Combs, B. Liquid-liquid phase separation induces pathogenic tau conformations in vitro. *Nat. Commun.* **11**, 2809 (2020).
65. Hu, J., Wang, G., Liu, X. & Gao, W. Enhancing Pharmacokinetics, Tumor Accumulation, and Antitumor Efficacy by Elastin-Like Polypeptide Fusion of Interferon Alpha. *Adv Mater* **27**, 7320–7324 (2015).
66. Navarro, A. M. *et al.* DisPhaseDB: An integrative database of diseases related variations in liquid-liquid phase separation proteins. *Comput Struct Biotechnol J* **20**, 2551–2557 (2022).
67. Feng, M. *et al.* Decoding Missense Variants by Incorporating Phase Separation via Machine Learning. *Nat Commun* **15**, 8279 (2024).
68. Mensah, M. A. *et al.* Aberrant phase separation and nucleolar dysfunction in rare genetic diseases. *Nature* **614**, 564–571 (2023).
69. Liao, Y.-C. *et al.* RNA Granules Hitchhike on Lysosomes for Long-Distance Transport, Using Annexin A11 as a Molecular Tether. *Cell* **179**, 147–164.e20 (2019).
70. Chin, A. F., Zheng, Y. & Hilser, V. J. Phylogenetic convergence of phase separation and mitotic function in the disordered protein BuGZ. *Protein Sci* **31**, 822–834 (2022).
71. Babl, L. *et al.* CTP-controlled liquid-liquid phase separation of ParB. *J Mol Biol* **434**, 167401

(2022).

72. Dasmeh, P. & Wagner, A. Natural Selection on the Phase-Separation Properties of FUS during 160 My of Mammalian Evolution. *Mol Biol Evol* **38**, 940–951 (2021).

73. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**, 635–645 (2004).

74. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).

75. Emms, D. M. & Kelly, S. STRIDE: Species Tree Root Inference from Gene Duplication Events. *Mol Biol Evol* **34**, 3267–3278 (2017).

76. Emms, D. M. & Kelly, S. STAG: Species tree inference from all genes. *bioRxiv* (2018) doi:10.1101/267914.

77. Jeong, J., Lee, J. H., Carcamo, C. C., Parker, M. W. & Berger, J. M. DNA-Stimulated Liquid-Liquid phase separation by eukaryotic topoisomerase ii modulates catalytic function. *Elife* **11**, (2022).

78. Ho, W.-L. & Huang, J.-R. The return of the rings: Evolutionary convergence of aromatic residues in the intrinsically disordered regions of RNA-binding proteins for liquid-liquid phase separation. *Protein Sci* **31**, e4317 (2022).

79. Liu, F. & Wang, J. ATP Acts as a Hydrotrope to Regulate the Phase Separation of NBDY Clusters. *JACS Au* **3**, 2578–2585 (2023).

80. Dang, M., Li, T. & Song, J. ATP and nucleic acids competitively modulate LLPS of the SARS-CoV2 nucleocapsid protein. *Commun Biol* **6**, 80 (2023).

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

82. Badaczewska-Dawid, A. E., Uversky, V. N. & Potoyan, D. A. BIAPSS: A Comprehensive

Physicochemical Analyzer of Proteins Undergoing Liquid-Liquid Phase Separation. *Int J Mol Sci* **23**, (2022).

83. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
84. Somjee, R., Mitrea, D. M. & Kriwacki, R. W. Exploring Relationships between the Density of Charged Tracts within Disordered Regions and Phase Separation. *Pac Symp Biocomput* **25**, 207–218 (2020).
85. Erdős, G. & Dosztányi, Z. AIUPred: combining energy estimation with deep learning for the enhanced prediction of protein disorder. *Nucleic Acids Res* **52**, W176–W181 (2024).
86. Uversky, V. N., Kuznetsova, I. M., Turoverov, K. K. & Zaslavsky, B. Intrinsically disordered proteins as crucial constituents of cellular aqueous two phase systems and coacervates. *FEBS Lett* **589**, 15–22 (2015).
87. Siligardi, G. & Drake, A. F. The importance of extended conformations and, in particular, the PII conformation for the molecular recognition of peptides. *Biopolymers* **37**, 281–292 (1995).
88. Hicks, J. M. & Hsu, V. L. The extended left-handed helix: a simple nucleic acid-binding motif. *Proteins* **55**, 330–338 (2004).
89. Miyagi, T. *et al.* The Patterning and Proportion of Charged Residues in the Arginine-Rich Mixed-Charge Domain Determine the Membrane-Less Organelle Targeted by the Protein. *Int J Mol Sci* **23**, (2022).
90. Holehouse, A. S., Ginell, G. M., Griffith, D. & Böke, E. Clustering of Aromatic Residues in Prion-like Domains Can Tune the Formation, State, and Organization of Biomolecular Condensates. *Biochemistry* **60**, 3566–3581 (2021).
91. Lin, Y.-H. & Chan, H. S. Phase Separation and Single-Chain Compactness of Charged Disordered Proteins Are Strongly Correlated. *Biophys J* **112**, 2043–2046 (2017).
92. Martin, E. W. *et al.* Valence and patterning of aromatic residues determine the phase

behavior of prion-like domains. *Science* **367**, 694–699 (2020).

93. Li, Q. *et al.* LLPSDB: a database of proteins undergoing liquid-liquid phase separation in vitro. *Nucleic Acids Res* **48**, D320–D327 (2020).
94. Ning, W. *et al.* DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. *Nucleic Acids Res* **48**, D288–D295 (2020).
95. You, K. *et al.* PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic Acids Res* **48**, D354–D359 (2020).
96. Mészáros, B. *et al.* PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res* **48**, D360–D367 (2020).
97. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
98. Xie, Q. *et al.* Phase separation in cancer at a glance. *J Transl Med* **21**, 237 (2023).
99. Nahm, M. *et al.* mutations in ALS cause dysregulation of calcium homeostasis and stress granule dynamics. *Sci Transl Med* **12**, (2020).
100. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
101. Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat Methods* **21**, 465–476 (2024).
102. Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res* **39**, D411–9 (2011).
103. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
104. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007).
105. Sayols, S. rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms.

MicroPubl Biol **2023**, (2023).

106. Erdős, G. & Dosztányi, Z. Analyzing Protein Disorder with IUPred2A. *Curr Protoc Bioinformatics* **70**, e99 (2020).

107. Mészáros, B., Erdos, G. & Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* **46**, W329–W337 (2018).

108. Ginell, G. M. & Holehouse, A. S. Analyzing the Sequences of Intrinsically Disordered Regions with CIDER and localCIDER. *Methods Mol Biol* **2141**, 103–126 (2020).

109. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci U S A* **110**, 13392–13397 (2013).

110. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105–132 (1982).

111. Elam, W. A., Schrank, T. P., Campagnolo, A. J. & Hilser, V. J. Evolutionary conservation of the polyproline II conformation surrounding intrinsically disordered phosphorylation sites. *Protein Sci* **22**, 405–417 (2013).

112. Huson, D. H. *et al.* Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**, 460 (2007).