Genome Biology

SOFTWARE

Open Access

catGRANULE 2.0: accurate predictions of liquid-liquid phase separating proteins at single amino acid resolution

Michele Monti[1,2†], Jonathan Fiorentino[1,2†], Dimitrios Miltiadis-Vrachnos[2,3], Giorgio Bini[2,4], Tiziana Cotrufo[5], Natalia Sanchez de Groot[6], Alexandros Armaos[1,2] and Gian Gaetano Tartaglia[1,2*]

†Michele Monti and Jonathan Fiorentino contributed equally to this work.

*Correspondence: gian.tartaglia@iit.it

[1] Center for Life Nano- & NeuroScience, Fondazione Istituto Italiano di Tecnologia, Viale Regina Elena 291, 00161 Rome, Italy
[2] RNA Systems Biology Lab, Centre for Human Technologies, Fondazione Istituto Italiano di Tecnologia, Via Enrico Melen 83, 16152 Genoa, Italy
[3] Department of Biology and Biotechnologies, University of Rome Sapienza, Piazzale Aldo Moro 5, 00185 Rome, Italy
[4] Physics Department, University of Genoa, Via Dodecaneso 33, 16146 Genoa, Italy
[5] Departament de Biologia Cellular, Fisiologia i Immunologia, Universitat de Barcelona, Avenida Diagonal 643, 08028 Barcelona, Spain
[6] Department of Biochemistry and Molecular Biology, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), 08193 Barcelona, Spain

Abstract

Liquid-liquid phase separation (LLPS) enables the formation of membraneless organelles, essential for cellular organization and implicated in diseases. We introduce catGRANULE 2.0 ROBOT, an algorithm integrating physicochemical properties and AlphaFold-derived structural features to predict LLPS at single-amino-acid resolution. The method achieves high performance and reliably evaluates mutation effects on LLPS propensity, providing detailed predictions of how specific mutations enhance or inhibit phase separation. Supported by experimental validations, including microscopy data, it predicts LLPS across diverse organisms and cellular compartments, offering valuable insights into LLPS mechanisms and mutational impacts. The tool is freely available at https://tools.tartaglialab.com/catgranule2 and https://doi.org/10.5281/zenodo.14205831.

Keywords: Liquid-liquid phase separation, Machine learning, Subcellular compartmentalization, Protein features, Mutations

Background

Liquid-liquid phase separation (LLPS) is a molecular phenomenon that brings molecules together to form membraneless condensates [1–5]. Recent studies evidenced the key role of LLPS in human health and disease, especially in protein condensation and neurodegenerative disorders [6]. Nucleic acids are known to play a central role in LLPS. Indeed, many proteins undergo phase separation in the presence of RNA [7, 8], although some can phase separate independently due to their intrinsically disordered domains [1, 9].

Contrary to the process of liquid to solid phase transition (LSPT), in which proteins go toward an irreversible aggregation state [10, 11], LLPS is a reversible process [12]. The reversibility of LLPS has a dual function: while the increase in protein concentration enhances the enzymatic activity [8, 13, 14], RNA accumulation in organelles such

as P-bodies can inhibit protein translation [15]. Condensates formed through phase separation function as dynamic hubs, catalyzing essential biochemical processes by concentrating and compartmentalizing specific proteins and biomolecules at precise subcellular locations [16].

Despite recent advances in the experimental probing and characterization of LLPS [17] and the construction of large-scale databases of LLPS-prone proteins [18–23], a wide coverage of the proteome of different species in light of their LLPS properties is still lacking. For this reason, several computational methods have been developed to predict the propensity of a protein to undergo LLPS [24–28]. However, most of these methods lack the ability to predict the LLPS propensity at the amino acid level and none of them has been extensively tested for the prediction of the effect of single and multiple mutations of the protein sequence on their capability to undergo LLPS. One of the first LLPS predictors, catGRANULE 1.0, computes the protein propensity for granule formation based on structural disorder and nucleic acid-binding propensities [27]. Following this, the MaGS method was developed using a variety of features including protein abundance, phosphorylation site annotations, and the occurrence of specific amino acids [25, 29]. A more recent method is PICNIC, which uses both sequence-based and structure-based features derived from AlphaFold2 models, focusing on sequence complexity, disorder scores, and amino acid co-occurrences [26]. An extended version, PICNICGO, adds Gene Ontology terms to provide deeper insights into functions like RNA-binding [26]. Lastly, PSPHunter broadens the feature set further by implementing word2vec for sequence analysis, alongside Position-Specific Scoring Matrix (PSSM) and Hidden Markov Model (HMM) to capture evolutionary and structural insights, encompassing a wide array of functional traits like protein modifications and network properties [28].

In this work, we introduce an advanced predictor of LLPS proteins called catGRANULE 2.0 ROBOT (R—Ribonucleoprotein, O—Organization, in B—Biocondensates, O—Organelle, T—Types). This new version significantly enhances the capabilities of its predecessor, catGRANULE 1.0, and it is based on a curated database of phase-separating proteins and their mutants. catGRANULE 2.0 ROBOT integrates a comprehensive set of features that include structural and sequence-based data derived from AlphaFold2 models, specifically targeting properties relevant to phase separation. This version has undergone testing against a wide array of mutations, which were compiled through an exhaustive literature search. catGRANULE 2.0 ROBOT does not rely on protein sequence feature encoders, instead integrating predictions based on physico-chemical properties from sequence and structural data. This strategy enhances the interpretability of predictions while maintaining robust performance.

The extensive training dataset used in catGRANULE 2.0 ROBOT comprises human proteins documented to undergo LLPS, sourced from various databases and resources [18–23, 29, 30]. It also includes a selection of negative proteins—those highly unlikely to undergo LLPS, specifically excluding known interactors of LLPS proteins [31]. In this manuscript, we first report the characterization of proteins belonging to the training dataset compared to the rest of the human proteome [32, 33]. Following this, we describe each protein in the dataset using a list of features that considers both sequence-based physico-chemical properties of the protein and structural properties, which we extract based on the AlphaFold Structure Database [34, 35]. After we encode sequence- and

structure-based features into a vector, we train multiple binary classifiers and we select the best performing model, which we call ROBOT, to define a LLPS propensity score for a protein. We tested the algorithm on proteins from various condensates in humans and other organisms, demonstrating that it performs better than earlier methods that rely on sequence [25, 27–29] and structural features [26]. We also provide an orthogonal validation of our predictions using thousands of antibody-based immunofluorescence (IF) confocal microscopy images obtained from the Human Protein Atlas [36]. catGRANULE 2.0 ROBOT can be employed to predict profiles of LLPS propensity along protein sequences and accurately identifies regions experimentally confirmed as LLPS drivers. Furthermore, catGRANULE 2.0 ROBOT assesses the impact of amino acid mutations on LLPS propensity, determining whether mutations will increase or decrease it. To this end, we employed mutations identified through a comprehensive literature review, including a deep mutational scanning of TDP-43 [37].

To make our algorithm easily usable by the scientific community, we developed a user-friendly web server (https://tools.tartaglialab.com/catgranule2).

Results

Construction and biological characterization of the training dataset

With the aim of building a robust machine learning method to predict the LLPS propensity of proteins at the amino acid level, we defined training and test datasets with the following workflow. We first collected human proteins known to be involved in LLPS from several publicly available databases [18–20, 22, 23, 29, 30], obtaining 5656 LLPS-prone proteins in total (Fig. 1A–B; see Methods section). We built the negative set by removing these proteins and their first interactors from the human proteome (see Methods section) [31]. To prevent overfitting during the training phase, we utilized CD-HIT [38] to filter both positive and negative sets, ensuring sequence similarity was below 50%. Subsequently, we divided the data into training and test sets, as detailed in Additional file 1: Fig. S1A–B (refer to Methods section for more information).

Since it is known that protein length is a relevant feature in LLPS prediction [27], we compared the distribution of the length and abundance [39] of the LLPS proteins from the training set with the negative set (Additional file 1: Fig. S1C–D) and we observed that the differences are consistent with a comparison against the rest of the human proteome (Additional file 1: Fig. S1E–F), highlighting the absence of a bias in the construction of the negative set.

Next, we performed a Gene Ontology (GO) term enrichment analysis using Panther [32], which revealed that the LLPS proteins in our training set are significantly enriched in protein classes linked to RNA-related activities, translation, protein binding, and metabolic processes (Fig. 1C and Additional file 1: Fig. S2), compared to the rest of the proteome, while the proteins in the negative set are enriched for transporters and transmembrane receptors proteins (Additional file 1: Fig. S2). These findings are consistent with the literature. RNA metabolism proteins such as TIA-1 and G3BP1 facilitate stress granule formation through LLPS, emphasizing the role of RNA-binding proteins in cellular stress responses [40, 41]. Moreover, enzymes influence LLPS via various post-translational modifications that alter protein interactions and stability [42]. By contrast, transporter and membrane proteins are known to form irreversible aggregates, as they

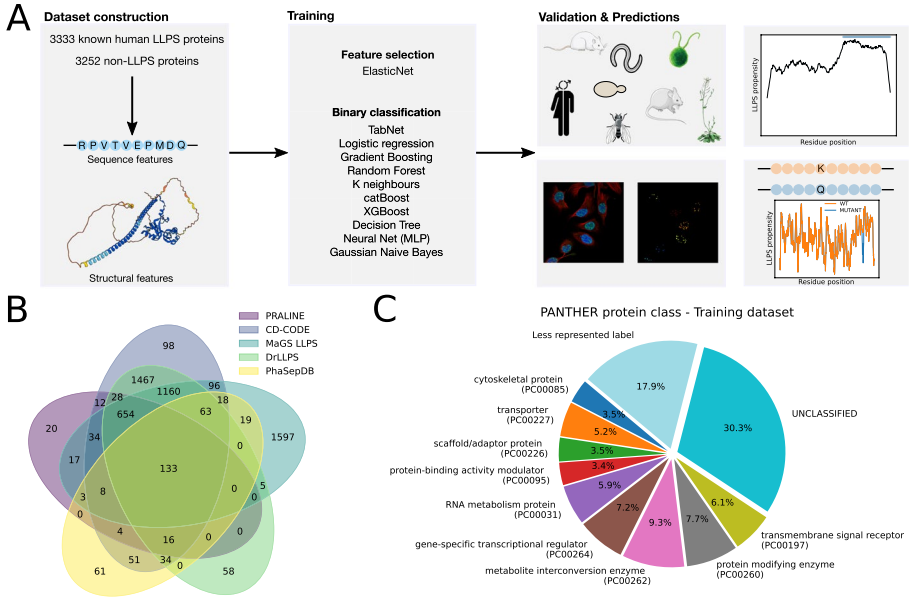


Fig. 1 **A** Schematics of the catGRANULE 2.0 ROBOT workflow. A training dataset is constructed consisting of 3333 known human LLPS proteins and 3252 non-LLPS proteins. The proteins are then encoded in a set of 128 features, including sequence-based physico-chemical and Alphafold2-derived structural features. Next, a subset of relevant features is selected using ElasticNet and ten different classifiers are trained on the dataset; MLP is the selected classifier according to its superior performance on the test dataset. catGRANULE 2.0 ROBOT predictions are then validated on sets of known LLPS-prone proteins from different species [22] and on immunofluorescence microscopy images from the Human Protein Atlas. LLPS propensity profiles are predicted with a sliding window approach and validated on experimentally known LLPS driving regions of proteins belonging to different species, obtained from the PhaSepDB database [19]. Finally, catGRANULE 2.0 ROBOT predicts the effect of single and multiple amino acid mutations on LLPS propensity. **B** Venn diagram showing the overlap of LLPS-prone proteins collected from different databases. **C** Composition of the training dataset in terms of Panther protein class categories. Protein classes with less than 3% have been aggregated in the "Less represented label" category

have fewer disordered regions and increased hydrophobicity [43, 44]. In this regard, it should be mentioned that defense and immunity proteins tend to also undergo LSPT, leading to protein aggregates associated with diseases such as ALS [45] (see Methods section and Additional file 2: Table S1).

Although we found that proteins belonging to the positive and negative training sets are enriched in different biological features, we do not observe a strong separation of the two sets from a principal components analysis (PCA) (Additional file 1: Fig. S1G), motivating us to rely on non linear machine learning methods for the classification task. To this aim, we characterized each protein in the dataset with a set of prioritized features, reported in Additional file 3: Table S2. We incorporated 80 physico-chemical features derived from protein sequence analysis and 2 phenomenological sequence patterns (see Methods section) [27, 46], along with 28 structural features extracted using predicted protein structures from the AlphaFold Structure Database [34, 35]. This approach enabled us to identify features related to both the surface and the inner parts of the protein. Additionally, we incorporated 18 features based on the compositional similarity of protein sequence windows to experimentally determined RNA-binding patches, enhancing our ability to identify potential RNA-interacting regions in the protein sequences [47] (see Methods section). In Additional file 1: Fig. S3, we show a cluster map of the

correlation matrix between the set of 128 features for the training dataset. We observe the presence of large clusters, especially for subsets of the physico-chemical features, as expected given the higher redundancy in their collection, compared to the structural features (see Methods section).

catGRANULE 2.0 ROBOT accurately classifies LLPS prone proteins

After the construction of robust training and test datasets, we developed a machine learning pipeline to predict the LLPS propensity of a protein. We employed ElasticNet [48] to identify the most relevant features for LLPS classification, then we trained 10 different binary classifiers, performing a grid search over the hyper-parameters of each classifier and employing 5-fold cross validation during training to avoid overfitting (see Methods section and Supplementary Information). Finally, we tested their performance on an independent test dataset using the area under the receiver-operating characteristic curve (AUROC) as scoring metric.

We found that the trained classifiers yield comparable AUROC scores and we selected the multi-layer perceptron (MLP) as the optimal one, based on its superior performance on the training dataset in 5-fold cross validation (Additional file 1: Fig. S4A). We compared the performance of our trained model on the test dataset with catGRANULE 1.0 [27], and the top performing state-of-the-art methods that are MaGS [25, 29], PICNIC, PICNIC-GO [26], and PSPHunter [28] (see Methods section), observing that our model (catGRANULE 2.0 ROBOT) emerges as the best one (Fig. 2A). Additionally, we found that catGRANULE 2.0 ROBOT outperforms the other predictors using other performance metrics, such as the accuracy, the F1-score, the Matthew's correlation coefficient (MCC), and the recall, while MaGS, PSPHunter, and PICNIC-GO perform better only for the precision (Supplementary Fig. S4B). Even considering proteins that belong to the test dataset and have $< 20\%$ sequence identity with those of the training dataset, we observe a good overall performance of catGRANULE 2.0 ROBOT, compared to the other algorithms (Additional file 1: Fig. S4C; see Methods section). We highlight that the other algorithms are advantaged in the comparison of the performance, since due to the iterative sampling of negatives from the human proteome during training or to the lack of availability of the full training datasets—especially regarding the negative sets—it is likely that subsets of proteins of our independent test dataset belong to the training sets of the other algorithms. We provide the LLPS score computed using catGRANULE 2.0 ROBOT for the whole human proteome in Additional file 4: Table S3.

Next, we compared the performance of catGRANULE 2.0 ROBOT to the other tools on 41 experimentally annotated LLPS proteins from different species, not including human, which belong to the LLPSDB database [23] and have $< 20\%$ sequence identity with the proteins of our training dataset. We found that catGRANULE 2.0 ROBOT retrieves the highest fraction of LLPS proteins (Additional file 1: Fig. S4D; see Methods section).

As an additional validation of the good performances of our model, we tested its capability to predict the LLPS propensity of proteins other than human, which was the only organism considered during the training (Fig. 2B, Additional file 1: Fig. S4E, and Additional file 5: Table S4). In Fig. 2B, we show the fraction of correctly predicted proteins for several species, and we compare the result with PICNIC when

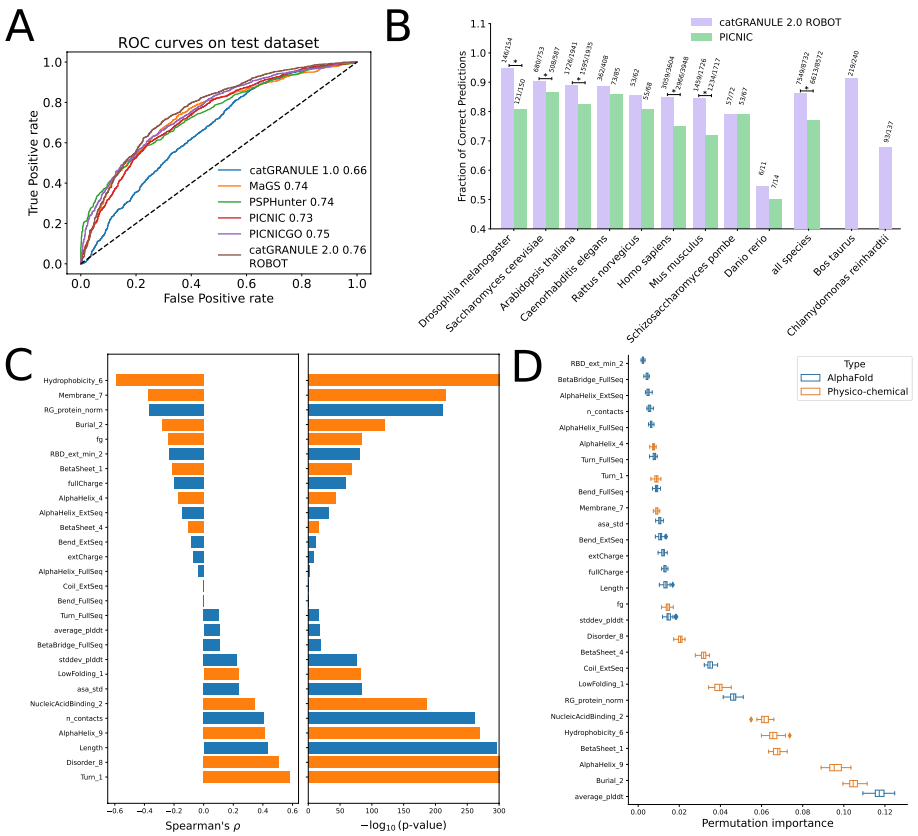


Fig. 2 **A** Receiver-operating characteristic (ROC) curves obtained from the test dataset, for catGRANULE 2.0 ROBOT and other LLPS prediction algorithms (see Methods section for details). The area under the ROC curve (AUROC) for each algorithm is indicated in the legend. **B** Bar plot of the fraction of correctly predicted LLPS proteins for different species. The annotation of LLPS proteins was obtained from the DrLLPS database [22]. A star above a bar indicates a *p* value smaller than 0.05 from a Fisher's exact test between the fraction of correctly predicted LLPS proteins in catGRANULE 2.0 ROBOT and in PICNIC. **C** Bar plot showing the Spearman's correlation coefficient between the 28 features selected during the training step using ElasticNet and the predicted LLPS score, for the proteins belonging to the training dataset. The bar plot on the right shows the −log10(*p* value) of the correlation coefficient. **D** Box plot of the permutation importance computed on the training dataset for the 28 features selected during the training step using ElasticNet

available [26]. We observe that catGRANULE 2.0 ROBOT significantly outperforms PICNIC in most of the species considered and it performs comparably in the others.

Using ElasticNet, we identified 28 features that are the most relevant to distinguish LLPS proteins from the negative set. Interestingly, while some of these features strongly separate the two classes individually, others do not show such strong differences between the classes, justifying our choice of a multivariate and non linear feature selection method (Additional file 1: Fig. S5A, Methods section, and Additional file 1: Supplementary Methods). This result was further corroborated by a comparison of the performances obtained with the 28 features selected by ElasticNet with those achieved by a linear model trained on all the 128 features, or with a MLP model trained adding features iteratively using a univariate feature selection method (Additional file 1: Fig. S5B; see Methods section).

Physico-chemical determinants of LLPS

To quantify the impact of the features to the discrimination of LLPS-prone proteins, we computed the Spearman's correlation coefficient of each selected feature with the predicted LLPS score, using the proteins belonging to the training set (Fig. 2C). We found, as expected and in agreement with the results from the catGRANULE 1.0 algorithm [27], that features associated to hydrophobicity (e.g., Hydrophobicity_6) are negatively associated with the LLPS propensity score [17], while those related to the nucleic acid binding propensity (e.g., NucleicAcidBinding_2) [8], to the disorder (e.g., Disorder_8) [49], or to the protein length [27] are positively associated. Yet, the contribution of specific classes of amino acids is contained in multiple physico-chemical features: for instance, aliphatic amino acids contribute positively to LLPS prediction in the low folding propensity [50, 51] (Fig. 2C), which is in agreement with experimental evidence [52–54].

We found that the radius of gyration normalized by protein length (RG_protein_norm) exhibits a negative correlation with the LLPS score, while the standard deviation of the accessible surface area (asa_std) shows a positive correlation (Fig. 2C). This suggests that more compact proteins are less prone to liquid-liquid phase separation (LLPS). Indeed, the power-law exponent < 1.0 for soluble species (monomers and oligomers, [55]) suggests that phase-separating proteins display consistent scaling behavior in terms of size and shape, a fundamental characteristic of protein structure. So, proteins with variable surface exposure, which facilitates interactions with other proteins and nucleic acids, are more likely to undergo LLPS. Coiled coils deviate from this typical behavior by exhibiting a linear scaling of the radius of gyration with the number of residues, further emphasizing the influence of specific protein structures on LLPS propensity. In this regard, it could be hypothesized that specific features of a protein sequence, computed on the internal or solvent-exposed residues, may be more informative in predicting the protein's LLPS propensity, compared to those computed on the full sequence. To test this, we analyzed the ability of separating the dataset on specific physico-chemical features computed exclusively for the buried or solvent-exposed residues, and compared the performance to the same features derived from the entire sequence. We defined the exposed residues as those with a solvent-accessible surface area (ASA) greater than 50%, using the PDB structure of each protein. In Additional file 1: Fig. S6, we show the ROC curves obtained for the features Hydrophobicity_6 and Disorder_8 considering the exposed or buried residues versus the whole sequence. The full sequence consistently provided more informative results than the exposed or buried residues alone.

Then, we computed the permutation importance to quantify the contribution of each feature to the overall score (Fig. 2D; see Methods section). We found that the average pLDDT is the most relevant feature. This finding aligns with expectations, as pLDDT is a measure of protein disorder [56], which corroborates our analysis that protein disorder significantly contributes to phase separation [27].

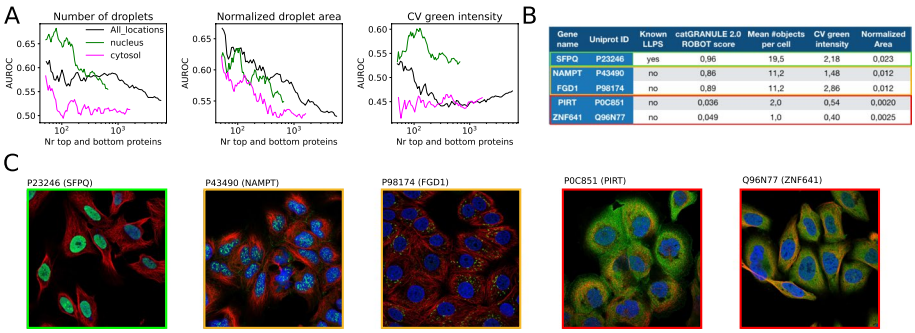
We further validated our feature selection strategy conducting an iterative feature elimination analysis (Additional file 1: Fig. S5 and Supplementary Methods). It highlighted the importance of properties such as nucleic acid binding in LLPS. Specifically, these nucleic acid binding predictions are based on the electrostatic charge, which is known to facilitate RNA contact [7, 11] and prevent protein aggregation [57].

Independent validation of catGRANULE 2.0 ROBOT

As an additional validation of catGRANULE 2.0 ROBOT predictions, we used 10757 antibody-based images obtained by immunofluorescence (IF) confocal microscopy in human cell lines obtained from the Human Protein Atlas (https://www.proteinatlas.org/humanproteome/cell) [36]. After cell segmentation, we computed the coefficient of variation (CV) of the green fluorescence per cell and we considered the maximum of this quantity over the cells, for each protein. Next, we identified puncta of the green fluorescent protein and we computed the area, normalized by the average area of the nuclei per image, and the average number of puncta per protein (see Methods section and Additional file 6: Table S5). We chose these quantities since we hypothesized that proteins undergoing LLPS would have more and larger droplets compared to other proteins, and a more compartmentalized expression.

By ranking the proteins based on their LLPS propensity scores predicted by catGRANULE 2.0 ROBOT and selecting an equal number of proteins from the top and bottom of the ranking, we observe that the AUROC scores initially range from 0.6 to 0.7, depending on the sample size under consideration. As the number of proteins increases, the AUROC scores gradually decrease toward 0.5, as shown in Fig. 3A. This trend supports the hypothesis that proteins with higher predicted LLPS scores exhibit droplets with the expected features in immunofluorescence (IF) images. Our predictions generally align well with the microscopy images, though not perfectly, due to the considerable variability in IF images of the same protein across different cell lines, the heterogeneity in antibody specificity, and the variability of experimental conditions.

In Fig. 3B, we report the values of the features computed from the IF images shown in Fig. 3C for some example proteins well known to undergo LLPS, such as SFPQ [58, 59], predicted to undergo LLPS by catGRANULE 2.0 ROBOT but not known, such as NAMPT and FGD1, and predicted to not perform LLPS (PIRT and ZNF641). We notice that the proteins predicted to undergo LLPS by catGRANULE 2.0 ROBOT show clear



Gene name	Uniprot ID	Known LLPS	catGRANULE 2.0 ROBOT score	Mean #objects per cell	CV green intensity	Normalized Area
SFPQ	P23246	yes	0,96	19,5	2,18	0,023
NAMPT	P43490	no	0,86	11,2	1,48	0,012
FGD1	P98174	no	0,89	11,2	2,86	0,012
PIRT	P0C851	no	0,036	2,0	0,54	0,0020
ZNF641	Q96N77	no	0,049	1,0	0,40	0,0025

Fig. 3 **A** AUROC versus the number of top and bottom proteins, ranked according to the predicted catGRANULE 2.0 ROBOT LLPS propensity score, for the average number of droplets (i.e., green puncta, left), area of the green puncta normalized by the average area of the nuclei (center) and coefficient of variation (CV) of the green intensity over the cell (right), computed from approximately 11k antibody-based images obtained by immunofluorescence (IF) confocal microscopy from the Human Protein Atlas (HPA). Line colors indicate the selection of proteins from different sub-cellular locations. See the Methods section and Additional File 6: Table S5. **B** Table showing the values of the quantities computed from the IF images for five example proteins, together with the LLPS propensity score predicted by catGRANULE 2.0 ROBOT and whether the protein was previously known to undergo LLPS. **C** IF images of the proteins reported in **B**. Note that the edge color matches those in **B**

droplets of the green fluorescence in the nucleus (SFPQ and NAMPT) or in the cytoplasm (FGD1) (Fig. 3C). Notably, FGD1 has been recently predicted between the top dosage sensitive proteins, a feature strongly associated with the ability to undergo LLPS [60]. Meanwhile NAMPT ensures the inactivation of ASK3, a protein that gets inactive after forming condensates via LLPS under hyperosmotic stress [61].

LLPS predictions and subcellular compartmentalization

Next, we studied how the predicted LLPS score varies for proteins belonging to different subcellular locations. We found that nucleolar proteins have the highest LLPS propensity, on average, followed by cytoplasmic and nuclear proteins (Fig. 4A). As expected, secreted, extracellular and membrane proteins are not predicted to undergo LLPS (Fig. 4A). For the nucleolus, nucleus, cytoplasm, and mitochondrion, we collected annotations for proteins from different types of liquid-like condensates from the DrLLPS database [22]. We show the distributions of the predicted LLPS score stratified by condensate in Additional file 1: Fig. S7A. We observe that proteins belonging to the Sam68 nuclear body show the highest LLPS scores, on average, followed by other nuclear condensates, such as the DNA damage foci, and the nucleolus. Cytoplasmic condensates, like stress granules and P-bodies, also show high LLPS score, while the mitochondrial RNA granules are at the bottom of the ranking. These results are in agreement with recent data from filtration chromatography and dilution experiments [62]. We noticed that the composition of certain condensates largely overlaps, e.g., for stress granules and P-bodies, while others display a more unique composition (e.g., postsynaptic density) (Additional file 1: Fig. S7B).

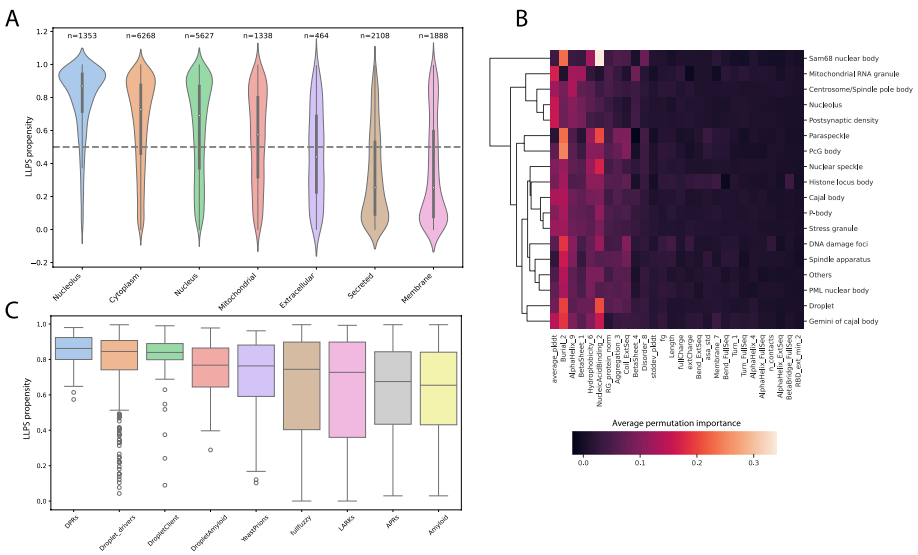


Fig. 4 **A** Violin plot showing the predicted LLPS score for proteins belonging to different subcellular locations, obtained from Uniprot [65], sorted according to descending median LLPS propensity score. The number of proteins for each subcellular location is indicated above each violin. **B** Cluster map of the average permutation importance for each condensate. We show the condensates in the rows and we clustered them, we show the 28 features selected by our model ordered according to the descending permutation importance obtained from the full training dataset (see Fig. 2D). **C** Box plot showing the predicted LLPS propensity score for different classes of LLPS-prone proteins [66], sorted according to the median

We then investigated the relevance of the features selected in our model for each condensate, by averaging the permutation importance over random sub-samplings of the proteins belonging to each condensate (see Methods section). While for most of the condensates we found a ranking of the selected features, according to the average permutation importance, similar to the one obtained on the training dataset (Fig. 2D), for the Sam68 nuclear body we observe that NucleicAcidBinding_2 is the most important feature, while structural features such as AlphaHelix_9 and Beta Sheet_1 are the top scoring for the mitochondrial RNA granule (Fig. 4B and Additional file 3: Table S2). This pattern corresponds well with our analysis of RNA-binding proteins documented in UniProt, where 11 out of 12 proteins in the Sam68 nuclear body are RNA-binding, in contrast to 30 out of 42 in the mitochondrial RNA granule. Specifically, the Sam68 nuclear body proteins exhibited 73 instances of beta strand regions and 38 instances of compositional bias, compared to 293 beta strand occurrences and 10 instances of compositional bias in the mitochondrial RNA granule. We further confirmed the result on the RNA binding propensity by computing the catRAPID signature score for the proteins in each condensate [63], which show that proteins belonging to the Sam68 nuclear body have the highest propensity for RNA binding and there is a wide difference between condensates (Additional file 1: Fig. S7C). Meanwhile looking at the DisProt disorder score [64] (feature Disorder_10, see Additional file 3: Table S2), we see less accentuated differences between condensates, although proteins belonging to the mitochondrial RNA granule have the smallest disorder score, on average (Additional file 1: Fig. S7D).

Finally, we divided proteins in different classes according to their role in condensate formation that have been previously defined [66] (see Methods section). Amyloid proteins are those found exclusively inside solid aggregates. Proteins belonging to the amyloid-promoting region (APR) and droplet-promoting region (DPR) classes have specific domains that can initiate amyloid or droplet formation under certain physical conditions [66]. Droplet drivers and clients are proteins that facilitate or participate in LLPS formation [66]. The FullFuzzy class includes typically intrinsically disordered proteins whose interaction behavior is dependent on the cellular context [67]. Low-complexity aromatic-rich kinked segments (LARKs) proteins are a class of proteins containing RNA binding domains [66]. Using this categorization, we first computed the predicted LLPS score for each protein and grouped the predictions by class (Fig. 4C), observing a clear trend from DPRs as the highest LLPS propensity class to amyloid as the lowest LLPS propensity class, in line with our expectations. Moreover, LARKs proteins also tend to have lower LLPS propensity scores, in line with the known role of aromatic chains in protein aggregation [68, 69]. The DropletAmyloid class, capable of both LLPS and LSPT, ranked intermediately, suggesting these phenomena might not be exclusive but could either compete or synergize, leading to a more thermodynamically stable structure. Indeed, while LLPS involves multivalent macromolecular interactions, LSPT also refers to specific changes in physicochemical properties. Both processes are concentration-dependent, yet intrinsic sequence features and the actual folded state of the proteins critically influence whether they undergo LLPS or LSPT [70, 71]. Additionally, we observe that some classes show a large overlap in their composition, especially the set of fullFuzzy proteins with the LARKs and DropletDrivers (Additional file 1: Fig. S8).

Although our analysis does not show substantial separation between the LLPS scores of drivers and clients (Fig. 4C), we further investigated their classification using the DrLLPS database, in which proteins are divided in scaffolds, clients, and regulators [22]. catGRANULE 2.0 ROBOT performs better than other LLPS predictors at distinguishing scaffolds from regulators and clients (Additional file 1: Fig. S9), which is in agreement with the fact that scaffolds drive LLPS independently from other proteins; notably, also PICNIC shows similar results [26].

LLPS profile and mutation score

Next we investigated the capability of our method to identify experimentally annotated LLPS driving regions, collected from the PhaSepDB database [19], in Fig. 5. Specifically, we studied how the AUROC score varies when considering sets of top and bottom scores of increasing size (see Methods section). We find that both the MLP classifier trained on structural and physico-chemical features and a Random Forest trained only on physico-chemical features achieve a AUROC ~ 0.9 when considering the top predictions, while the performance decreases to AUROC ~ 0.6 when taking into account all the predicted scores (Fig. 5A). We noticed that the predictions of the two classifiers have a good correlation (Additional file 1: Fig. S10A). Interestingly, we obtain better performance on proteins belonging to different organisms compared to the subset of human proteins. From this analysis, we chose to adopt the Random Forest classifier, trained only on physico-chemical features, as the preferred model for the computation of LLPS propensity profiles. Our choice was further supported by the superior performance, on average, of the Random Forest classifier trained on the set of physico-chemical features over all the other classifiers, even when trained on the full set of features, in

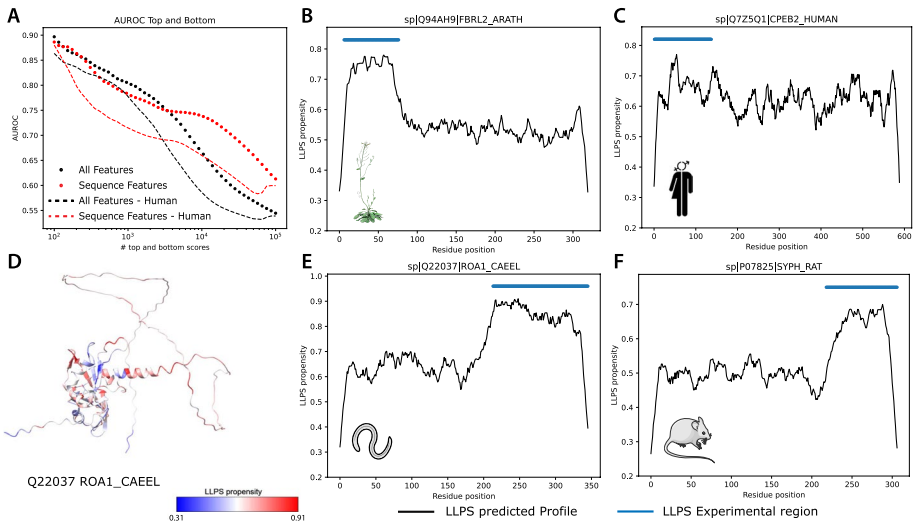


Fig. 5 **A** AUROC vs number of top and bottom scores for the MLP classifier trained on structural and physico-chemical features (black) and a Random Forest classifier trained only on physico-chemical features (red). Dots and dashes indicate proteins from all organisms or only from human, respectively. **B**-**C**-**E**-**F** LLPS propensity profiles predicted by the Random Forest classifier trained only on physico-chemical features (black curve) and experimentally annotated LLPS driving regions (blue lines) obtained from the PhaSepDB database [19] for four proteins from different organisms. **D** Protein structure colored according to the predicted LLPS propensity profile

the prediction of the experimental LLPS-driving regions (Additional file 1: Fig. S10B). We conducted the same analysis of Fig. 5A using the PSPHunter algorithm [28] and we found that catGRANULE 2.0 ROBOT strongly outperforms it in the prediction of experimentally annotated LLPS regions from the LLPS propensity profiles (Additional file 1: Fig. S10D). In Fig. 5B, C, E, and F, we show the predicted LLPS propensity profiles for four proteins from different organisms, together with the experimental LLPS driving region, finding a strong agreement between our predictions and the experimental annotation [19]. Finally, in Fig. 5D we show the structure of the protein (PDB) from panel B and its LLPS propensity colored at single amino acid resolution (see Methods section). In this case, the regions with higher LLPS propensity tend to match with structural elements of low complexity such as loops.

Finally, we used our method to compute LLPS propensity profiles and score mutations affecting condensate formation, a task rarely explored by previous approaches. Given current limitations of AlphaFold in predicting single amino acid mutation effects [72, 73], we employed a Random Forest classifier trained on physico-chemical features. This simplified model performs similarly to the full model in predicting LLPS propensity profiles (Fig. 5A). To validate the capability of our method in identifying the effect of mutations on LLPS, we collected a list of 24 distinct mutations of 9 proteins that undergo LLPS in the WT form but show increased or reduced LLPS propensity when mutated (see Methods section and Additional file 7: Table S6), and we compare the performance in scoring mutations of catGRANULE 2.0 ROBOT with those achieved by catGRANULE 1.0 [27] and PSPHunter [28].

First, we computed the LLPS score of the WT proteins and we notice that only catGRANULE 2.0 ROBOT correctly predicts all of them to undergo LLPS (Fig. 6A). Next, we computed a mutation score (see Methods section) for the 24 mutations for the three algorithms and we evaluated the fraction of correctly predicted mutations separately for the set of mutations increasing or decreasing the LLPS propensity. We found that catGRANULE 2.0 ROBOT and catGRANULE 1.0 correctly predict the 80% of the mutations with a negative effect on LLPS, while PSPHunter correctly calculates only the 50%. For mutations with a positive effect on LLPS, catGRANULE 2.0 ROBOT outperforms the other algorithms (Fig. 6B).

Next, to assess the ability of catGRANULE 2.0 ROBOT to predict the effect of mutations on LLPS propensity under consistent environmental conditions, we analyzed a mutational scanning dataset of TDP-43 [37], a protein known to undergo LLPS and that is implicated in neurodegenerative diseases. This large-scale screening highlights a strong correlation between the formation of LLPS and cellular toxicity. Specifically, mutations that facilitate LLPS in TDP-43 increase cellular toxicity due to interactions with other cellular molecules. Conversely, mutations leading to LSPT result in less toxic, more inert protein aggregates [37]. The mutational scanning includes approximately 60,000 mutations of TDP-43, including both single and double mutations [37] (see Methods section). In Fig. 6C, we show the distribution of the mutation score (Eq. (1)) predicted by catGRANULE 2.0 ROBOT, considering the full mutational scanning or restricting the absolute value of the experimental phase separation score to a certain threshold. We observe that the separation between the distributions of LLPS decreasing and LLPS increasing mutations, represented by the red and black curves, respectively,

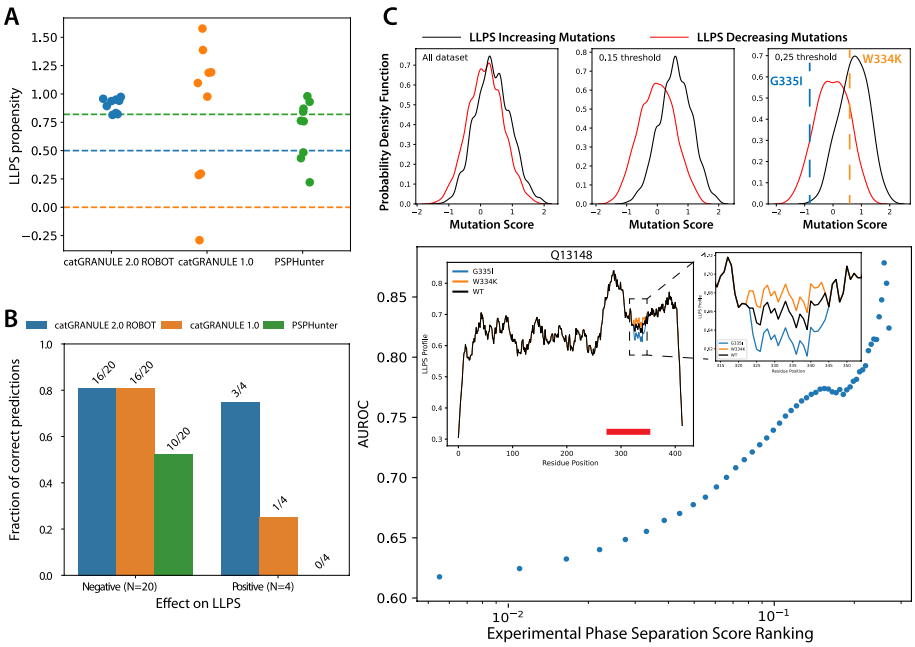


Fig. 6 **A** LLPS propensity score computed by catGRANULE 2.0 ROBOT, catGRANULE 1.0, and PSPHunter on the WT sequence of 9 proteins for which mutations affecting LLPS were collected. Colored dashed lines indicate the threshold to discriminate LLPS from non-LLPS proteins, with the color matching the algorithm. **B** Bar plot showing the fraction of correctly predicted mutation scores by catGRANULE 2.0 ROBOT, catGRANULE 1.0, and PSPHunter, for a set of 24 mutations including 20 mutations with a negative effect on LLPS and 4 mutations with a positive effect. **C** Distributions of the catGRANULE 2.0 ROBOT mutation score for mutations decreasing or increasing LLPS (red and black curves, respectively) from a mutational scanning of TDP-43 [37], at different thresholds on the experimental phase separation score. In the rightmost panel, the colored dashed lines show the predicted mutation score of two selected mutations. **D** AUROC computed on the catGRANULE 2.0 ROBOT mutation scores for the mutational scanning of TDP-43, as a function of the threshold on the experimental phase separation score. Increasing the threshold corresponds to selecting more restricted sets of mutations, with stronger positive and negative experimental effect on LLPS. The inset shows the LLPS propensity profiles predicted by catGRANULE 2.0 ROBOT for the WT sequence of TDP-43 and the two mutations shown in **C**. The red line indicates the experimental LLPS region

increases with the threshold on the experimental score, and that catGRANULE 2.0 ROBOT correctly predicts the sign of the mutation score for a LLPS decreasing (G335I) and a LLPS increasing (W334K) mutations whose physical properties were studied in detail [37] (Fig. 6C). In Fig. 6D, we show the AUROC achieved by catGRANULE 2.0 ROBOT, computed on sets of top and bottom mutations selected at different thresholds of the experimental phase separation score. While at low values of the threshold catGRANULE 2.0 ROBOT achieves an AUROC score slightly higher than 0.6, increasing the threshold, which corresponds to selecting sets of mutations with stronger negative or positive effect on TDP-43 LLPS, it reaches AUROC close to 0.9 (Fig. 6D). The inset shows the profiles of the TDP-43 WT protein, which nicely agrees with the experimental LLPS region (red line) [74], and of the two mutations already shown in Fig. 6C. Furthermore, we observed that the histograms of the catGRANULE 2.0 ROBOT mutation score for single and double mutations are both approximately Gaussian centered at zero (Additional file 1: Fig. S10C). Nevertheless, for double mutations the tails of the distribution are heavier, suggesting that they can have a stronger effect on the LLPS propensity compared to single mutations, as expected [37].

Overall, we showed that catGRANULE 2.0 ROBOT can accurately predict LLPS propensity profiles and LLPS-prone regions of proteins from different species. Moreover, it outperforms previous methods in scoring the effect of single and multiple amino acids mutations on LLPS, making it an appealing tool for the design of proteins or peptides with tunable LLPS properties.

Discussion

The emergence of LLPS as a central molecular process governing membrane-less organelle formation underscores its implications in cellular organization and human health, particularly in the context of neurodegenerative disorders. Unlike the irreversible aggregation characteristic of LSPT [57, 75], LLPS condensates present a reversible but less thermodynamically stable state. This reversibility facilitates dynamic protein and RNA compartmentalization, influencing critical cellular activities [8].

In this study, we developed a LLPS protein predictor, catGRANULE 2.0 ROBOT, which integrates structural and sequence-based features. The algorithm enables detailed profiling of phase-separation properties and computation of mutants, offering an intuitive web interface and pre-calculated scores for various model systems. A series of enhancements positions catGRANULE 2.0 ROBOT at the forefront of predictive tools for studying LLPS outperforming current methodologies by integrating comprehensive biophysical data and cutting-edge computational predictions. Consistent with prior findings, our results demonstrate that nucleic acid binding, intrinsic disorder [27, 49], and specific amino acid properties, such as aliphaticity [53, 54], play a positive role in influencing LLPS. Furthermore, structural insights derived from AlphaFold2 [34], particularly variability in pLDDT-based disorder, highlight the critical balance between structured and disordered regions in regulating LLPS behavior.

A significant contribution of this study lies in the strategic construction of training and testing datasets. We constructed our training and testing sets using diverse and well-curated sources of LLPS-relevant data, incorporating proteins known to undergo LLPS [18–20, 22, 23, 29, 30]. This strategy ensured comprehensive coverage, integrating both frequently observed cases and rarer instances, thereby enhancing the diversity of the training dataset and significantly expanding the LLPS atlas. Additionally, we placed strong emphasis on the construction of a stringent negative set. Specifically, we excluded proteins documented in the analyzed studies and their direct interactors. By adopting this approach, we mitigated biases arising from known non-physical interactions, ultimately improving the specificity and robustness of our predictions [76]. Our training dataset relates to proteins forming phase-separated assemblies in near physiological conditions (e.g., nucleolar proteins) and we focused on the intrinsic determinants of LLPS, namely those sequence and structural properties that favor LLPS under these conditions. However, in the future we plan to include context-dependent features such as ions and cellular-specific chemical modifications, which greatly impact phase separation through mechanisms like ionic strength, pH, and concentration. In this regard, a neural network-based model that predicts LLPS given a set of specific experimental conditions has been recently introduced, although a broad applicability of such class of models will require the curation of large databases with standardized annotations of experimental conditions in which proteins undergo LLPS [77].

While our collection of proteins represents a bona fide set of LLPS-prone candidates, we acknowledge that LLPS is a complex and multifaceted process that requires further refinement in its definition. Traditionally, LLPS describes the spontaneous formation of distinct phases in supersaturated solutions, driven by intrinsic molecular interactions [8]. However, as mechanisms such as phase separation coupled with percolation [78] and nanoclustering [79] are being explored, it becomes evident that the phenomenon may be more nuanced than previously understood. Therefore, future efforts must incorporate more detailed definitions of protein assemblies. These efforts will be critical to fully understanding the diverse mechanisms driving LLPS and refining the properties and definitions associated with this phenomenon.

Through a combination of direct and indirect experimental approaches [36, 37], we corroborated our predictions, affirming the robustness and reliability of catGRANULE 2.0 ROBOT. The analysis of deep mutational scanning, particularly in the case of TDP-43 [37], further strengthens the alignment between our predictions and experimental observations, underscoring the clinical relevance of our study. Indeed, an essential aspect of our method is the capability to identify and predict the effects of mutations on LLPS propensity. This holds promise for elucidating the molecular mechanisms underlying disease-associated mutations and guiding precision protein engineering efforts [80, 81]. We stress that evaluating the impact of mutations on the propensity for LLPS presents significant challenges, primarily due to the dependency of experimental validation on varying environmental and cellular conditions. High-quality in vitro data is notably difficult to procure, as exemplified by studies such as [82, 83]. In contrast, numerous in-cell experiments, such as those reported in [84, 85], provide evidence of phase separation. These studies also include assessments ensuring that the mutations do not interfere with cellular processes or protein interactions that could sequester the protein into stress granules. Currently, the absence of a comprehensive database for extensive validation remains a major limitation in testing predictive methods. Furthermore, it is crucial to acknowledge that environmental factors such as pH, ionic strength, and concentration can significantly influence the conditions under which phase separation occurs, even in mutants. Additionally, it is important to consider that cellular robustness may mitigate the impact of mutations. The presence of multiple proteins supporting the LLPS organelle and the influence of multivalency can buffer the effects of individual mutations. This suggests that in a cellular context, mutations might have a lower impact due to the collective interaction of various molecules involved in the LLPS process.

A key enhancement in catGRANULE 2.0 ROBOT is the use of AlphaFold2 [34], with plans to upgrade to AlphaFold3 [86] in future iterations for even more precise structural predictions. Additionally, we aim to integrate it with algorithms such as catRAPID [87] and scRAPID [88] to better predict specific RNAs that contribute to protein crowding, enhancing the model's ability to simulate complex biological environments, and the cell type specific expression of the RNAs interacting with proteins undergoing LLPS. Moreover, further experimental validation through techniques such as FRAP and FCS [89] will be essential to link LLPS-related properties with specific structural features now predictable with AlphaFold3 [86] and to better define the events leading to LLPS. These developments will position the new generation of

algorithms at the cutting edge of predictive tools for studying protein phase separation, offering unparalleled accuracy and comprehensive biophysical data integration.

Conclusions

Here, we built comprehensive datasets of LLPS proteins and developed catGRANULE 2.0 ROBOT, a method for predicting LLPS propensities. By integrating sequence-based features with structural insights from AlphaFold2, our approach provides reliable LLPS profiles and robust predictions for the effects of mutations on LLPS propensity. To support the research community, we developed a web server for catGRANULE 2.0 ROBOT (https://tools.tartaglialab.com/catgranule2) enabling users to explore LLPS predictions and design mutants for various applications.

We applied catGRANULE 2.0 ROBOT to proteins localized in various cellular compartments, including stress granules, nuclear bodies, Cajal bodies, and P-bodies. In all our analysis, we focused on the intrinsic determinants of LLPS encoded within protein sequences, though future developments should incorporate environmental factors, such as protein concentration, and the conditions under which these proteins are expressed [77]. Further exploration of these conditions, along with chemical modifications of proteins [79] and RNAs [80], could reveal critical interactions essential for organelle formation. Moreover, to enhance our understanding of LLPS assemblies, we plan to integrate data on protein-protein and protein-RNA interactions [31, 90], providing a clearer picture of their composition and dynamics. Such an approach, inspired by efforts to map interactions involving RNA-binding proteins [91], has already revealed critical characteristics of LLPS in our recent work [76, 92].

In conclusion, our study not only deepens the understanding of LLPS but also paves the way for new engineering applications. With the ability to predict and manipulate LLPS propensity, we open avenues for designing proteins with tailored behavior, offering, for instance, potential for therapeutic innovation [93, 94].

Methods

Construction of the training dataset

To build the positive set of LLPS-prone proteins, we collected data for human proteins from several databases and resources. Specifically, we used:

- 929 proteins annotated as "Droplet state" in the PRALINE database [18];
- 3876 proteins from the CD-CODE database [20];
- 3807 proteins used in a recent study on LLPS predictors (provided in table S1 from [29]);
- 117 proteins defined in [30] (obtained from Table S3 in [25], where they are defined as the "Gingras gold standard");
- 3633 proteins from the DrLLPS database [22];
- 833 proteins from the PhaSepDB database (v2.1) [19];
- 59 proteins from the PhaSePro database [21];
- 92 proteins from LLPSDB [23].

We considered the union of these sets since some of them have a large overlap, as shown by the Venn diagram in Fig. 1B, obtaining 5656 proteins. Next, we used CD-HIT [38] to filter proteins from the positive set for 50% sequence identity, obtaining 4807 proteins. To generate the negative set, we removed the proteins belonging to the positive set from the human proteome, and we also removed their first interactors based on protein-protein interactions collected from the BioGRID database (v3.5.175) [31]. Finally, we filtered proteins in the negative set using CD-HIT, as described above. To ease the comparison with the top performing state-of-the-art methods, which are MaGS [25, 29], PICNIC, and PICNIC-GO [26], we included in our training positive set the LLPS-prone proteins on which those algorithms were trained. Since the positive training set for PICNIC was extracted from the CD-CODE database [20], but the authors do not provide the proteins that have been used in the training, we included in the positive training set all the proteins belonging to CD-CODE, after the CD-HIT filtering described above. Next, we randomly sampled the same amount of negative proteins. Considering the intersection with the proteins for which a pdb file is present in the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/) [35], the final training dataset is made of 3333 positive and 3252 negative proteins. We used the remaining LLPS-prone proteins as an independent positive test set and we sampled the negatives as for the training dataset. We obtained a test dataset with 1422 positive and 1376 negative proteins. More details are in Additional file 1: Supplementary Methods.

From the training dataset, we computed the correlation matrix of the features and we represented it in a hierarchical cluster map using the function "clustermap" of the "seaborn" Python package.

Definition of model features

Physico-chemical features

We use a list of 80 experimental scales encoding physico-chemical properties of proteins that describe aggregation, hydrophobicity, membrane, nucleic acid binding, disorder, burial, alpha helix, beta sheet, turn propensities, and two phenomenological sequence patterns [27, 46]. These features are computed solely based on the protein sequence and they were already used in the cleverSuite [46] and in the training of the catGRANULE 1.0 algorithm for LLPS propensity prediction [27]. Each protein sequence is transformed in a list of 82 numbers, representing the average of each physico-chemical feature over the sequence.

AlphaFold2-derived features

We downloaded the structures predicted by AlphaFold2 [34] for the whole human proteome from the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/) [35]. For the proteins having multiple predicted models, we select the model with highest average predicted local-distance difference test (pLDDT), a measure of local accuracy of the AlphaFold2 prediction. Next, we used the "Bio.PDB" submodule of the "Biopython" (version 1.81) python package to extract structural features from the pdb file for each protein. Specifically, we extract the number of contacts, the radius of gyration (RG), and the accessible surface area (ASA), from which we define the exposed residues of the protein as those with ASA 50%, while the rest are considered as internal. Next, for the

internal and the exposed residues we extract the secondary structure properties (helix, strand, coil, turn, accessibility, disorder) and the charge. More details are in Additional file 1: Supplementary Methods.

For each protein structure, we obtain a vector of structural and RNA-binding features of length 46; adding the 82 physico-chemical features we have 128 features per protein. In Additional file 3: Table S2, we report a mapping of the feature names and their categorization in feature families. We also report the original references from which the physico-chemical scales were obtained.

Model training

We use the Python library scikits-learn (version 1.1.1) [95] to train multiple classifiers for LLPS propensity prediction. Specifically, we use Decision Tree, Random Forest, XGBoost, Gradient Boosting, Gaussian Naive Bayes, TabNet, Logistic Regression, catBoost, K-neighbors, and a multilayer perceptron (MLP). For each classifier, we define a scikits-learn pipeline including feature scaling ("StandardScaler" function), feature selection, and training of the classifier. We use ElasticNet as the feature selection method since it performs multivariate feature selection, i.e., it takes into account the (possibly non-linear) statistical dependence between features when selecting them during training [48]. To avoid overfitting, we use a 5-fold cross validation for training ("gridsearchCV" function), including a grid search for hyperparameter optimization on both the feature selection method and the classifiers. To this end, we defined a parameter search space specific to each classifier. We trained each classifier using the area under the receiver operating characteristic curve (AUROC) as the scoring function. We selected the MLP as the best model based on the AUROC obtained on the training set. In the feature selection step, we obtained a set of 28 features selected by ElasticNet. We computed the Spearman's correlation coefficient between each feature and the predicted LLPS propensity score on our training dataset, to obtain a sign for the effect of a feature increase or decrease on the LLPS propensity score. To quantify the contribution of each selected feature to the final AUROC score, we used the "permutation_importance" function from the "sklearn.inspection" submodule, which randomly permutes a feature in the dataset and computes the AUROC using the trained model. We represent the distributions of the values of the permutation importance obtained from 50 repetitions of the described procedure using a box plot. More details are in Additional file 1: Supplementary Methods.

Model validation

We compared the performance of catGRANULE 2.0 ROBOT on the independent test set with those of catGRANULE 1.0 [27] and of the four top performing state-of-the-art algorithms, MaGS [25, 29], PICNIC, PICNIC-GO [26], and PSPHunter [28]. We discarded other LLPS predictors that were shown to perform worse than these state-of-the-art algorithms. For each algorithm, we computed the ROC curves using the function "roc_curve" from "sklearn.metrics" and the AUROC using the function "roc_auc_score" on the test dataset.

As a further validation of catGRANULE 2.0 ROBOT, we considered the proteins in our independent test dataset with sequence identity, computed using the mmseqs2 algorithm [96], smaller than 20% with the training set: we found 716 positive and 590 negative

proteins. From these sets, we randomly sampled a balanced test dataset of 100 proteins 50 times and we computed the mean and standard deviation of different performance metrics for all the LLPS predictors. Since each metric quantifies different strengths and weaknesses of the tools, we also computed a normalized overall score, obtained by computing a z-score for each metric, then aggregating by taking the average over the samples for each tool and metric, computing the mean of the z-score over the metrics for each tool and normalizing these values between 0 and 1. Moreover, we considered proteins from the LLPSDB database [23] with less than 20% sequence identity with our training set, and we found 41 proteins belonging to different species, which did not include *Homo sapiens*. We compared the performance of catGRANULE 2.0 ROBOT with the other tools by considering the predicted LLPS score and evaluating the fraction of correctly predicted LLPS proteins at increasing values of a threshold on the LLPS score. MaGS was not included in this analysis since it works only for human and yeast proteins.

For the computation of the catGRANULE 2.0 ROBOT LLPS propensity score in other species, we retrieved proteins annotated as LLPS-prone from the DrLLPS database [22], we downloaded the pdb files for those proteins from the AlphaFold Structure Database [35], we encoded each protein into the vector of 128 features described above, and we predicted the LLPS propensity score using our pre-trained MLP classifier. We retrieved the numbers of correctly predicted proteins by PICNIC from [26] and, for the species that are predicted both in our study and by PICNIC, we tested, for each species separately, if the fraction of correctly predicted LLPS-prone proteins by catGRANULE 2.0 ROBOT was significantly larger than the one predicted by PICNIC using a Fisher's exact test (function "fisher_exact" in scipy.stats).

Regarding the analysis of the LLPS propensity per condensate, first we retrieved the annotation of the protein sub-cellular locations from Uniprot (https://www.uniprot.org/) [65] and the annotation of LLPS condensates from the DrLLPS database [22]. Since proteins can be found in multiple condensates, we represented the intersections using an upset plot ("upset" function from the "UpsetR" package, version 1.4.0, in R version 4.0.3). The violin plots are obtained using the function "violinplot" from the seaborn Python package. More details are in Additional file 1: Supplementary Methods.

Validation on immunofluorescence images from the Human Protein Atlas

For the analysis of antibody-based images obtained by immunofluorescence (IF) confocal microscopy from the Human Protein Atlas (https://www.proteinatlas.org/humanproteome/cell) [36], we retrieved a curated list of 11,608 images from [97]. We used a CellProfiler3 [98] pipeline provided in [97], which we adapted to compute additional quantities from the images and to use it with CellProfiler4.2.6 [99].

Specifically, we perform cell segmentation from the IF image through the Otsu's thresholding method using the red and blue channels, which quantify the microtubules and the DAPI, respectively. Then, we compute the standard deviation and the mean of the green intensity per cell, whose ratio defines the coefficient of variation (CV). Next, we segment the nuclei using the blue channel, we compute the area of each nucleus, and we take the average over each image. Finally, we segment the droplets (i.e., puncta of the green fluorescent protein) for each cell using the robust background thresholding method, and we compute the area, measured in pixels, of each droplet, which is made adimensional by dividing

it by the average area of the nuclei for the corresponding image. To obtain measurements at the protein level and compare them to the LLPS propensity scores predicted by catGRANULE 2.0 ROBOT, we computed, for each protein, the average number of droplets, the maximum normalized area, and the maximum CV of the green signal. In this way, we obtained the measurements for 10,757 IF proteins, each corresponding to one IF image. We provide all the computed scores in in Additional file 6: Table S5.

Computation of the LLPS propensity profiles and prediction of the effect of mutations on LLPS

catGRANULE 2.0 ROBOT predicts the LLPS propensity of a protein based on a set of sequence- and structural-based features. However, to compute a LLPS propensity profile and to allow the usage of our model on deep mutational scanning of proteins, which generate tens of thousands of mutations, we chose to train the model using only physico-chemical features. This choice allows a fast analysis of deep mutational scanning of proteins and it is supported by previous studies that showed that AlphaFold2 cannot predict reliably the structure of proteins subjected to single-point mutations [72, 73].

Prediction and validation of LLPS propensity profiles

We generate a LLPS propensity profile for a protein by applying a sliding window to a protein sequence and scoring each segment with the trained model. In this way, we obtain a LLPS propensity score at single amino acid resolution. We trained different classifiers using both the full set of features or only the set of physico-chemical features. Then, we collected approximately 250 proteins from different organisms from the PhaSepDB database [19], where regions responsible for LLPS are annotated over the sequence. Furthermore, to increase the accuracy and the sensitivity of the prediction, we filtered out proteins with more than 90% of the sequence annotated as the LLPS-prone region from the PhaSepDB database. Using this dataset, we found that the optimal size of the sliding window for the computation of the LLPS propensity profiles is 21 amino acids. To quantify the performance of the trained models over the PhaSepDB database, we concatenated all the protein sequences and we ranked the amino acids according to the predicted LLPS propensity. Next, we selected subsets of top and bottom LLPS propensity scores and we computed the AUROC score, where the true classes are obtained from the PhaSepDB database (see Fig. 5A). To select a model for the computation of the profiles, we compared the average AUROC obtained from the top-bottom scores approach and we found that a Random Forest classifier, trained only on the set of physico-chemical features, achieves the best performance (Additional file 1: Fig. S10B). Moreover, the average of the LLPS propensity profiles obtained with the Random Forest classifier shows a good correlation, quantified by the Spearman's correlation coefficient, with the LLPS propensity score predicted by the full model, which is the MLP classifier trained on the full set of features (Additional file 1: Fig. S10A).

Prediction of the effect of mutations on LLPS propensity

To score the effect of a mutation on the LLPS propensity of a protein, we employ the sum of the difference between the LLPS profiles of the mutant and wild-type (WT) proteins, indicated as $mut(i)$ and $WT(i)$, respectively, divided by the average of the profile of the WT protein

$$m = N \frac{\sum_{i=0}^{N} mut(i) - WT(i)}{\sum_{0}^{N} WT(i)}, \quad (1)$$

where N is the sequence length.

We collected a list of 24 mutations of 9 proteins, including single and multiple amino acid mutations, from the literature and the Uniprot database [65]. Specifically, we searched for "phase separation" in the "mutagenesis" field. We did not include other search strings (e.g., "condensation") given the variability in experimental conditions and cellular assays. The mutations are reported in Additional file 7: Table S6, where we also indicate the references from which the mutations have been retrieved. These mutations were categorized according to their annotated effect of increasing or decreasing LLPS propensity and/or affecting the protein localization in SGs and PBs. Next, we predicted the LLPS propensity profiles of the wild-type (WT) and mutated proteins, and we computed a mutation score as defined in Eq. (1). We also computed the LLPS propensity of the WT and mutated proteins using the PSPHunter [28] and catGRANULE 1.0 [27] web servers, and we compared the fraction of mutations for which the effect on LLPS propensity was correctly predicted, separately for mutations decreasing and increasing the LLPS propensity, between catGRANULE 2.0 ROBOT and these two algorithms (Fig. 6B). The predicted scores for the WT and mutated proteins for the three algorithms are reported in Additional file 7: Table S6.

Finally, we considered a mutational scanning of TDP-43 where approximately 60,000 mutations of the prion-like domain were generated and their toxicity was quantified in yeast cells [37]. The authors showed that mutations that increase protein aggregation strongly decrease toxicity, while toxic mutations promote LLPS. Thus, the toxicity score can be used as a proxy of experimental LLPS propensity. We predicted a LLPS propensity profile for each mutation and we computed a mutation score as described above. Then, we employed a top-bottom approach as we did for the validation of the LLPS propensity profiles. Specifically, we ranked the mutations according to the experimental phase separation score and we set different thresholds on this score. For each threshold, we computed the AUROC, as shown in Fig. 6C.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03497-7.

Additional file 1: Supplementary Methods and Supplementary Figures: Figs. S1-S10.

Additional file 2: Table S1. Results of a functional enrichment analysis on the positive and negative proteins belonging to the training dataset using gProfiler.

Additional file 3: Table S2. Mapping of feature names and categorization in feature families.

Additional file 4: Table S3. LLPS scores predicted by catGRANULE 2.0 ROBOT for the whole human proteome, with annotation of proteins belonging to the training and test datasets and if the proteins were already known to undergo LLPS.

Additional file 5: Table S4. LLPS scores predicted by catGRANULE 2.0 ROBOT for the sets of proteins annotated as LLPS-prone in the DrLLPS database for different species.

Additional file 6: Table S5. Features computed from immunofluorescence microscopy images obtained from the Human Protein Atlas, aggregated at the protein level.

Additional file 7: Table S6. List of 24 mutations affecting LLPS curated from the literature with predicted mutation and WT scores by catGRANULE 2.0 ROBOT, catGRANULE 1.0 and PSPHunter.

Acknowledgements
The authors acknowledge Andrea Vandelli and Jakob Rupert for useful discussions.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
MM, JF, and GGT conceived the study. MM and JF developed the method and performed the computational analyses. DMV and GB contributed to the computational validation of the method. AA developed the catGRANULE 2.0 ROBOT web server. MM, JF, and GGT wrote the manuscript. TC and NSdG provided critical insights and edited the manuscript. GGT supervised the study. All authors read and approved the final manuscript.

Funding
The research leading to these results have been supported through ERC [ASTRA_855923 (to G.G.T.), H2020 Projects IASIS_727658 and INFORE_825080 and IVBM4PAP_101098989 (to G.G.T.)] and National Center for Gene Therapy and Drug based on RNA Technology (CN00000041), financed by NextGenerationEU PNRR MUR - M4C2 - Action 1.4- Call "Potenziamento strutture di ricerca e di campioni nazionali di R&S" (CUP J33C22001130001) (to G.G.T.). Funding for open access charge: ERC ASTRA_855923 (to G.G.T.).

Data availability
To train, test, and validate catGRANULE 2.0 ROBOT, we used the following public databases: CD-CODE [100], DrLLPS [101], MaGS [102], PhaSepDB [103], LLPSDB [104], PRALINE [105], PhasePRO [106]. Fluorescence microscopy images were obtained from the Human Protein Atlas [107]; interactors of proteins undergoing LLPS were taken from the BioGRID database [108]. Uniprot [109] was used to retrieve protein sequences and mutations affecting LLPS, while 3D predicted protein structures were retrieved from the AlphaFold protein structure database [110]. All the data needed to reproduce the analysis in this manuscript are available from the supplementary material. The code associated with this study is available at the GitHub repository https://github.com/tartaglialabIIT/catGRANULE2.0 [111] and released to Zenodo https://doi.org/10. 5281/zenodo.14205831 [112] under MIT license. The catGRANULE 2.0 ROBOT web server is available at https://tools.tartaglialab.com/catgranule2.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 19 July 2024 Accepted: 6 February 2025
Published online: 20 February 2025

References

1. Martin EW, Holehouse AS. Intrinsically disordered protein regions and phase separation: sequence determinants of assembly or lack thereof. Emerg Top Life Sci. 2020;4(3):307–29.
2. Shapiro DM, Ney M, Eghtesadi SA, Chilkoti A. Protein phase separation arising from intrinsic disorder: first-principles to bespoke applications. J Phys Chem B. 2021;125(25):6740–59.
3. Boeynaems S, Alberti S, Fawzi NL, Mittag T, Polymenidou M, Rousseau F, et al. Protein phase separation: a new phase in cell biology. Trends Cell Biol. 2018;28(6):420–35.
4. Nandana V, Schrader JM. Roles of liquid-liquid phase separation in bacterial RNA metabolism. Curr Opin Microbiol. 2021;61:91–8.
5. Nesterov SV, Ilyinsky NS, Uversky VN. Liquid-liquid phase separation as a common organizing principle of intracellular space and biomembranes providing dynamic adaptive responses. Biochim Biophys Acta (BBA) Mol Cell Res. 2021;1868(11):119102.
6. Mathieu C, Pappu RV, Taylor JP. Beyond aggregation: pathological phase transitions in neurodegenerative disease. Science. 2020;370(6512):56–60.
7. Gotor NL, Armaos A, Calloni G, Torrent Burgas M, Vabulas RM, De Groot NS, et al. RNA-binding and prion domains: the yin and yang of phase separation. Nucleic Acids Res. 2020;48(17):9491–504.
8. Zacco E, Broglia L, Kurihara M, Monti M, Gustincich S, Pastore A, et al. RNA: the unsuspected conductor in the orchestra of macromolecular crowding. Chem Rev. 2024;124(8):4734–77.

9. Posey AE, Holehouse AS, Pappu RV. Phase separation of intrinsically disordered proteins. Methods Enzymol. 2018;611:1–30.
10. Zanzoni A, Marchese D, Agostini F, Bolognesi B, Cirillo D, Botta-Orfila M, et al. Principles of self-organization in biological pathways: a hypothesis on the autogenous association of alpha-synuclein. Nucleic Acids Res. 2013;41(22):9987–98.
11. Rupert J, Monti M, Zacco E, Tartaglia GG. RNA sequestration driven by amyloid formation: the alpha synuclein case. Nucleic Acids Res. 2023;51(21):11466–78.
12. Qamar S, Wang G, Randle SJ, Ruggeri FS, Varela JA, Lin JQ, et al. FUS phase separation is modulated by a molecular chaperone and methylation of arginine cation-π interactions. Cell. 2018;173(3):720–34.
13. O'Flynn BG, Mittag T. The role of liquid-liquid phase separation in regulating enzyme activity. Curr Opin Cell Biol. 2021;69:70–9.
14. Gil-Garcia M, Benítez-Mateos AI, Papp M, Stoffel F, Morelli C, Normak K, et al. Local environment in biomolecular condensates modulates enzymatic activity across length scales. Nat Commun. 2024;15(1):3322.
15. Luo Y, Na Z, Slavoff SA. P-bodies: composition, properties, and functions. Biochemistry. 2018;57(17):2424–31.
16. Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. Nat Rev Mol Cell Biol. 2017;18(5):285–98.
17. Alberti S, Gladfelter A, Mittag T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. Cell. 2019;176(3):419–34.
18. Vandelli A, Arnal Segura M, Monti M, Fiorentino J, Broglia L, Colantoni A, et al. The PRALINE database: protein and Rna humAn singLe nucleotIde variaNts in condEnsates. Bioinformatics. 2023;39(1):btac847.
19. You K, Huang Q, Yu C, Shen B, Sevilla C, Shi M, et al. PhaSepDB: a database of liquid-liquid phase separation related proteins. Nucleic Acids Res. 2019;48(D1):D354–9.
20. Rostam N, Ghosh S, Chow CFW, Hadarovich A, Landerer C, Ghosh R, et al. CD-CODE: crowdsourcing condensate database and encyclopedia. Nat Methods. 2023;20(5):673–6.
21. Mészáros B, Erdős G, Szabó B, Schád É, Tantos Á, Abukhairan R, et al. PhaSePro: the database of proteins driving liquid-liquid phase separation. Nucleic Acids Res. 2020;48(D1):D360–7.
22. Ning W, Guo Y, Lin S, Mei B, Wu Y, Jiang P, et al. DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. Nucleic Acids Res. 2019;48(D1):D288–95.
23. Wang X, Zhou X, Yan Q, Liao S, Tang W, Xu P, et al. LLPSDB v2.0: an updated database of proteins undergoing liquid-liquid phase separation in vitro. Bioinformatics. 2022;38(7):2010–4.
24. Pancsa R, Vranken W, Mészáros B. Computational resources for identifying and describing proteins driving liquid-liquid phase separation. Brief Bioinforma. 2021;22(5):bbaa408.
25. Kuechler ER, Budzyńska PM, Bernardini JP, Gsponer J, Mayor T. Distinct features of stress granule proteins predict localization in membraneless organelles. J Mol Biol. 2020;432(7):2349–68.
26. Hadarovich A, Singh HR, Ghosh S, Scheremetjew M, Rostam N, Hyman AA, et al. PICNIC accurately predicts condensate-forming proteins regardless of their structural disorder across organisms. Nat Commun. 2024;15(1):1–16.
27. Bolognesi B, Gotor NL, Dhar R, Cirillo D, Baldrighi M, Tartaglia GG, et al. A concentration-dependent liquid phase separation can cause toxicity upon increased protein expression. Cell Rep. 2016;16(1):222–31.
28. Sun J, Qu J, Zhao C, Zhang X, Liu X, Wang J, et al. Precise prediction of phase-separation key residues by machine learning. Nat Commun. 2024;15(1):2662.
29. Kuechler ER, Huang A, Bui JM, Mayor T, Gsponer J. Comparison of biomolecular condensate localization and protein phase separation predictors. Biomolecules. 2023;13(3):527.
30. Youn JY, Dyakov BJA, Zhang J, Knight JDR, Vernon RM, Forman-Kay JD, et al. Properties of stress granule and P-body proteomes. Mol Cell. 2019;76(2):286–94.
31. Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021;30(1):187–200.
32. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013;8(8):1551–66.
33. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47(W1):W191–8.
34. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
35. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596(7873):590–6.
36. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. Science. 2017;356(6340):eaal3321.
37. Bolognesi B, Faure AJ, Seuma M, Schmiedel JM, Tartaglia GG, Lehner B. The mutational landscape of a prion-like domain. Nat Commun. 2019;10(1):4162.
38. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
39. Huang Q, Szklarczyk D, Wang M, Simonovic M, von Mering C. PaxDb 5.0: curated protein quantification data suggests adaptive proteome changes in yeasts. Mol Cell Proteomics. 2023;22(10):100640.
40. Riggs CL, Kedersha N, Ivanov P, Anderson P. Mammalian stress granules and P bodies at a glance. J Cell Sci. 2020;133(16):jcs242487.
41. Monti M, Guiducci G, Paone A, Rinaldo S, Giardina G, Liberati FR, et al. Modelling of SHMT1 riboregulation predicts dynamic changes of serine and glycine levels across cellular compartments. Comput Struct Biotechnol J. 2021;19:3034–41.
42. Liu Y, Feng W, Wang Y, Wu B. Crosstalk between protein post-translational modifications and phase separation. Cell Commun Signal. 2024;22(1):110.
43. Tartaglia GG, Caflisch A. Computational analysis of the S. cerevisiae proteome reveals the function and cellular localization of the least and most amyloidogenic proteins. Proteins Struct Funct Bioinforma. 2007;68(1):273–278.

44. Li J, Fang Y, Zhang Y, Wang H, Yang Z, Ding D. Supramolecular self-assembly-facilitated aggregation of tumor-specific transmembrane receptors for signaling activation and converting immunologically cold to hot tumors. Adv Mater. 2021;33(16):2008518.
45. García-González P, Cabral-Miranda F, Hetz C, Osorio F. Interplay between the unfolded protein response and immune function in the development of neurodegenerative diseases. Front Immunol. 2018;9:423355.
46. Klus P, Bolognesi B, Agostini F, Marchese D, Zanzoni A, Tartaglia GG. The cleverSuite approach for protein characterization: predictions of structural properties, solubility, chaperone requirements and RNA-binding abilities. Bioinformatics. 2014;30(11):1601–8.
47. Castello A, Fischer B, Frese CK, Horos R, Alleaume AM, Foehr S, et al. Comprehensive identification of RNA-binding domains in human cells. Mol Cell. 2016;63(4):696–710.
48. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–20.
49. Ibrahim AI, Khaodeuanepheng N, Amarasekara D, Correia JJ, Lewis K, Fitzkee NC, et al. Intrinsically disordered regions that drive phase separation form a robustly distinct protein class. Biophys J. 2023;122(3):207a.
50. Tartaglia GG, Vendruscolo M. Proteome-level interplay between folding and aggregation propensities of proteins. J Mol Biol. 2010;402(5):919–28.
51. Monti M, Armaos A, Fantini M, Pastore A, Tartaglia GG. Aggregation is a context-dependent constraint on protein evolution. Front Mol Biosci. 2021;8:678115.
52. Iglesias V, Santos J, Santos-Suárez J, Pintado-Grima C, Ventura S. SGnn: a web server for the prediction of prion-like domains recruitment to stress granules upon heat stress. Front Mol Biosci. 2021;8:718301.
53. Villegas JA, Levy ED. A unified statistical potential reveals that amino acid stickiness governs nonspecific recruitment of client proteins into condensates. Protein Sci. 2022;31(7):e4361.
54. Mohanty P, Shenoy J, Rizuan A, Mercado-Ortiz JF, Fawzi NL, Mittal J. A synergy between site-specific and transient interactions drives the phase separation of a disordered, low-complexity domain. Proc Natl Acad Sci. 2023;120(34):e2305625120.
55. Tanner JJ. Empirical power laws for the radii of gyration of protein oligomers. Acta Crystallogr D Struct Biol. 2016;72(10):1119–29.
56. Wilson CJ, Choy WY, Karttunen M. AlphaFold2: a role for disordered protein/region prediction? Int J Mol Sci. 2022;23(9):4591.
57. Buell AK, Tartaglia GG, Birkett NR, Waudby CA, Vendruscolo M, Salvatella X, et al. Position-dependent electrostatic protection against protein aggregation. ChemBioChem. 2009;10(8):1309–12.
58. West JA, Mito M, Kurosaka S, Takumi T, Tanegashima C, Chujo T, et al. Structural, super-resolution microscopy analysis of paraspeckle nuclear body organization. J Cell Biol. 2016;214(7):817–30.
59. Marshall AC, Cummins J, Kobelke S, Zhu T, Widagdo J, Anggono V, et al. Different low-complexity regions of SFPQ play distinct roles in the formation of biomolecular condensates. J Mol Biol. 2023;435(24):168364.
60. Yang L, Lyu J, Li X, Guo G, Zhou X, Chen T, et al. Phase separation as a possible mechanism for dosage sensitivity. Genome Biol. 2024;25(1):17.
61. Watanabe K, Morishita K, Zhou X, Shiizaki S, Uchiyama Y, Koike M, et al. Cells recognize osmotic stress through liquid-liquid phase separation lubricated with poly (ADP-ribose). Nat Commun. 2021;12(1):1353.
62. Keber FC, Nguyen T, Mariossi A, Brangwynne CP, Wühr M. Evidence for widespread cytoplasmic structuring into mesoscale condensates. Nat Cell Biol. 2024;26(3):346–52.
63. Livi CM, Klus P, Delli Ponti R, Tartaglia GG. catRAPID signature: identification of ribonucleoproteins and RNA-binding regions. Bioinformatics. 2016;32(5):773–5.
64. Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept Lett. 2008;15(9):956–63.
65. UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):D523–31.
66. Vendruscolo M, Fuxreiter M. Sequence determinants of the aggregation of proteins within condensates generated by liquid-liquid phase separation. J Mol Biol. 2022;434(1):167201.
67. Fuxreiter M. Fuzzy protein theory for disordered proteins. Biochem Soc Trans. 2020;48(6):2557–64.
68. Tartaglia GG, Cavalli A, Pellarin R, Caflisch A. The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. Protein Sci. 2004;13(7):1939–41.
69. Cukalevski R, Boland B, Frohm B, Thulin E, Walsh D, Linse S. Role of aromatic side chains in amyloid β-protein aggregation. ACS Chem Neurosci. 2012;3(12):1008–16.
70. Wang B, Zhang L, Dai T, Qin Z, Lu H, Zhang L, et al. Liquid-liquid phase separation in human health and diseases. Signal Transduct Target Ther. 2021;6(1):290.
71. Hong K, Song D, Jung Y. Behavior control of membrane-less protein liquid condensates with metal ion-induced phase separation. Nat Commun. 2020;11(1):5554.
72. Pak MA, Markhieva KA, Novikova MS, Petrov DS, Vorobyev IS, Maksimova ES, et al. Using AlphaFold to predict the impact of single mutations on protein stability and function. PLoS ONE. 2023;18(3):e0282689.
73. Buel GR, Walters KJ. Can AlphaFold2 predict the impact of missense mutations on structure? Nat Struct Mol Biol. 2022;29(1):1–2.
74. Hallegger M, Chakrabarti AM, Lee FC, Lee BL, Amalietti AG, Odeh HM, et al. TDP-43 condensation properties specify its RNA-binding and regulatory repertoire. Cell. 2021;184(18):4680–96.
75. Brorsson AC, Bolognesi B, Tartaglia GG, Shammas SL, Favrin G, Watson I, et al. Intrinsic determinants of neurotoxic aggregate formation by the amyloid β peptide. Biophys J. 2010;98(8):1677–84.
76. Cerase A, Armaos A, Neumayer C, Avner P, Guttman M, Tartaglia GG. Phase separation drives X-chromosome inactivation: a hypothesis. Nat Struct Mol Biol. 2019;26(5):331–4.
77. Raimondi D, Orlando G, Michiels E, Pakravan D, Bratek-Skicki A, Van Den Bosch L, et al. In silico prediction of in vitro protein liquid-liquid phase separation experiments outcomes with multi-head neural attention. Bioinformatics. 2021;37(20):3473–9.
78. Mittag T, Pappu RV. A conceptual framework for understanding phase separation and addressing open questions and challenges. Mol Cell. 2022;82(12):2201–14.

79. Heinkel F, Abraham L, Ko M, Chao J, Bach H, Hui LT, et al. Phase separation and clustering of an ABC transporter in Mycobacterium tuberculosis. Proc Natl Acad Sci. 2019;116(33):16326–31.
80. Vandelli A, Broglia L, Armaos A, Ponti RD, Tartaglia GG. Rationalizing the effects of RNA modifications on protein interactions. Mol Ther Nucleic Acids. 2024;35(4):102391.
81. Giambruno R, Zacco E, Ugolini C, Vandelli A, Mulroney L, D'Onghia M, et al. Unveiling the role of PUS7-mediated pseudouridylation in host protein interactions specific for the SARS-CoV-2 RNA genome. Mol Ther Nucleic Acids. 2023;34:102052.
82. Xu G, Liu C, Zhou S, Li Q, Feng Y, Sun P, et al. Viral tegument proteins restrict cGAS-DNA phase separation to mediate immune evasion. Mol Cell. 2021;81(13):2823–37.
83. Yang P, Mathieu C, Kolaitis RM, Zhang P, Messing J, Yurtsever U, et al. G3BP1 is a tunable switch that triggers phase separation to assemble stress granules. Cell. 2020;181(2):325–45.
84. Gwon Y, Maxwell BA, Kolaitis RM, Zhang P, Kim HJ, Taylor JP. Ubiquitination of G3BP1 mediates stress granule disassembly in a context-specific manner. Science. 2021;372(6549):eabf6548.
85. Shen C, Li R, Negro R, Cheng J, Vora SM, Fu TM, et al. Phase separation drives RNA virus-induced activation of the NLRP6 inflammasome. Cell. 2021;184(23):5759–74.
86. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):493–500.
87. Armaos A, Colantoni A, Proietti G, Rupert J, Tartaglia GG. catRAPID omics v2.0: going deeper and wider in the prediction of protein-RNA interactions. Nucleic Acids Res. 2021;49(W1):W72–9.
88. Fiorentino J, Armaos A, Colantoni A, Tartaglia GG. Prediction of protein-RNA interactions from single-cell transcriptomic data. Nucleic Acids Res. 2024;52(6):e31–e31.
89. Perego E, Zappone S, Castagnetti F, Mariani D, Vitiello E, Rupert J, et al. Single-photon microscopy to study biomolecular condensates. Nat Commun. 2023;14(1):8224.
90. Zhao W, Zhang S, Zhu Y, Xi X, Bao P, Ma Z, et al. POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. Nucleic Acids Res. 2022;50(D1):D287–94.
91. Brannan KW, Jin W, Huelga SC, Banks CA, Gilmore JM, Florens L, et al. SONAR discovers RNA-binding proteins from analysis of large-scale protein-protein interactomes. Mol Cell. 2016;64(2):282–93.
92. Cerase A, Calabrese JM, Tartaglia GG. Phase separation drives X-chromosome inactivation. Nat Struct Mol Biol. 2022;29(3):183–5.
93. Zacco E, Kantelberg O, Milanetti E, Armaos A, Panei FP, Gregory J, et al. Probing TDP-43 condensation using an in silico designed aptamer. Nat Commun. 2022;13(1):3306.
94. Spence H, Waldron FM, Saleeb RS, Brown AL, Rifai OM, Gilodi M, et al. RNA aptamer reveals nuclear TDP-43 pathology is an early aggregation event that coincides with STMN-2 cryptic splicing and precedes clinical manifestation in ALS. Acta Neuropathol. 2024;147(1):50.
95. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
96. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8.
97. Yu C, Shen B, You K, Huang Q, Shi M, Wu C, et al. Proteome-scale analysis of phase-separated proteins in immunofluorescence images. Brief Bioinforma. 2021;22(3):bbaa187.
98. McQuin C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW, et al. CellProfiler 3.0: next-generation image processing for biology. PLoS Biol. 2018;16(7):e2005970.
99. Stirling DR, Swain-Bowden MJ, Lucas AM, Carpenter AE, Cimini BA, Goodman A. Cell Profiler 4: improvements in speed, utility and usability. BMC Bioinformatics. 2021;22:1–11.
100. Rostam N, Ghosh S, Chow CFW, Hadarovich A, Landerer C, Ghosh R, et al. CD-CODE (CrowDsourcing COndensate Database and Encyclopedia). 2023. https://cd-code.org/. Accessed 22 June 2023.
101. Ning W, Guo Y, Lin S, Mei B, Wu Y, Jiang P, et al. DrLLPS. 2019. http://llps.biocuckoo.cn/. Accessed 22 June 2023.
102. Kuechler, Erich R and Budzyńska, Paulina M and Bernardini, Jonathan P and Gsponer, Jörg and Mayor, Thibault. 2019_StressGranuleFeatures. Github. 2020. https://github.com/ekuec/2019_StressGranuleFeatures/. Accessed 22 June 2023.
103. You K, Huang Q, Yu C, Shen B, Sevilla C, Shi M, et al. PhaSepDB. 2019. http://db.phasep.pro/. Accessed 22 June 2023.
104. Wang X, Zhou X, Yan Q, Liao S, Tang W, Xu P, et al. LLPSDB v2.0. 2022. http://bio-comp.org.cn/llpsdbv2. Accessed 22 June 2023.
105. Vandelli A, Arnal Segura M, Monti M, Fiorentino J, Broglia L, Colantoni A, et al. PRALINE. 2023. http://praline.tartaglialab.com. Accessed 22 June 2023.
106. Mészáros B, Erdős G, Szabó B, Schád É, Tantos Á, Abukhairan R, et al. PhaSePro. 2020. https://phasepro.elte.hu/. Accessed 22 June 2023.
107. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. The Human Protein Atlas. 2017. https://www.proteinatlas.org/. Accessed 22 June 2023.
108. Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, et al. BioGRID. 2021. https://thebiogrid.org/. Accessed 29 June 2023.
109. UniProt. 2023. https://www.uniprot.org/. Accessed 22 June 2023.
110. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. AlphaFold Protein Structure Database. 2021. https://alphafold.ebi.ac.uk/. Accessed 15 June 2023.
111. Fiorentino J, Monti M. tartaglialabIIT/catGRANULE2.0: v1.0.0. Github. 2024. https://github.com/tartaglialabIIT/catGRANULE2.0.git. Accessed 22 Jan 2025.
112. Fiorentino J, Monti M. tartaglialabIIT/catGRANULE2.0: v1.0.0. Zenodo. 2024. https://doi.org/10.5281/zenodo.14205832.

Publisher's Note