PICNIC accurately predicts condensate-forming proteins regardless of their structural disorder across organisms

Anna Hadarovich[1,2], Hari Raj Singh[1], Soumyadeep Ghosh[1,2], Nadia Rostam[1,2,#], Anthony A. Hyman[1,2,3], Agnes Toth-Petroczy[1,2,3]*

1: Max Planck Institute of Molecular Cell Biology and Genetics, Dresden 01307, Germany

2: Center for Systems Biology Dresden, Dresden, 01307 Germany

3: Cluster of Excellence Physics of Life, TU Dresden, 01062 Dresden, Germany

*Corresponding author: Agnes Toth-Petroczy, toth-petroczy@mpi-cbg.de

[#]Current address: Department of Biology, College of Science, University of Sulaimani, Iraq

Keywords: biomolecular condensates, membrane-less organelles, prediction of proteins in condensates, machine learning, structural disorder, IDRs

Abstract

Biomolecular condensates are membraneless organelles that can concentrate hundreds of different proteins to operate essential biological functions. However, accurate identification of their components remains challenging and biased towards proteins with high structural disorder content with focus on self-phase separating (driver) proteins. Here, we present a machine learning algorithm, PICNIC (Proteins Involved in CoNdensates In Cells) to classify proteins involved in biomolecular condensates regardless of their role in condensate formation. PICNIC successfully predicts condensate members by identifying amino acid patterns in the protein sequence and structure in addition to the intrinsic disorder and outperforms previous methods. We performed extensive experimental validation *in cellulo* and demonstrated that PICNIC accurately predicts 21 out of 24 condensate-forming proteins regardless of their structural disorder content. Even though increasing disorder content was associated with organismal complexity, we found no correlation between predicted condensate proteome content and disorder content across organisms. Overall, we applied a novel machine learning classifier to interrogate condensate components at single protein and whole-proteome levels across the tree of life (picnic.cd-code.org).

Main

Introduction

Biomolecular condensates are membrane-less organelles that can selectively concentrate biomolecules[1] and are typically non-stochiometric assemblies of thousands of protein molecules and nucleic acids. The role of condensates has been implicated in several fundamental biochemical processes in physiology and disease[2–4]. Functions exerted by condensates include: i) sequestering molecules and shutting down translation of specific mRNAs[5]; ii) buffering concentration of proteins[6]; iii) reservoir of proteins for fast assembly and disassembly of large complexes, such as the nuclear envelope[7].

Proteins can have two major roles in a condensate: drivers (scaffolds) or clients[8,9]. Drivers can induce the formation of condensates and are essential members of condensates. For example, knock-out of the driver protein can lead to disassembly of a condensate. While a client is recruited to a condensate often via an interaction with a driver protein, and it is neither necessary nor sufficient in driving the condensate formation. Drivers often self-phase separate *in vitro*, nevertheless self-phase separation does not guarantee *in vivo* driver functionality. For most condensate-forming proteins the client or driver status is unknown, therefore we refer to them as condensate members.

While many proteins can phase-separate in the test-tube and form liquid-like condensates *in vitro*, studying if they also form condensates *in viv*o is more challenging. Individual proteins can be labeled using fluorescent tags and imaged for testing droplet formation which exhibit liquid-like properties such as fusion, Oswald ripening, fast dynamics in fluorescent recovery after photobleaching (FRAP)[10,11] assay. However, condensates may contain hundreds and even thousands of different proteins. Systematic detection of condensate proteomes is limited to a few mass spectrometry and proximity labeling studies: purified nucleoli[12,13], P-bodies[14] and stress granules[15,16] were subjected to mass spectrometry analysis for enrichment. Systematically and experimentally testing which proteins are members of condensates remains a bottleneck in the field. Therefore, computational methods can facilitate the process of characterizing proteins involved in biomolecular condensates at proteome-scale.

The main limitation of computational method development is related to the sparse experimental data of verified condensate-forming proteins. The first models were proposed based on properties of a few protein families, such as CatGranule[17] and PScore[18,19]. While these methods aided the discovery of novel condensate-forming proteins with similar properties, they do not generalize well[20].

In recent years, several data driven and machine learning based liquid-liquid phase separation (LLPS) predictors have been developed[21–24] making use of the experimental data aggregated in four LLPS databases[25–28]. While most predictors focused on identifying self-phase separating proteins (i.e. form *in vitro* condensates, that are also often drivers)[22–24] and thus trained on *in vitro* data, a recent metapredictor combined scores of previous methods and microscopy data to identify all condensate members (both drivers and clients)[29]. These methods have excellent performance compared to the first generation of predictors, nevertheless they have several shortcomings. For example a sub-optimal or biased definition of the negative dataset based on a priori assumptions about driving features, such as disorder being the main determinant, and accordingly using structured proteins from PDB as negative data[23].

Here, we focused on predicting proteins involved in biomolecular condensates instead of proteins involved in *in vitro* self-phase separation per se. We developed a new machine learning model called PICNIC (Proteins Involved in CoNdensates In Cells). As amino acid composition bias and patterning of charges were shown to impact the ability of proteins to form condensates[30–32], we developed novel features that represent short and long range co-occurrences of amino acids in the protein sequence and structure (AlphaFold2 models), as well as used features such as sequence complexity[23], disorder score[33] that were previously shown to be successful in identifying drivers. We addressed the short-comings of previous methods by using a larger and highly curated and non-redundant dataset of *in vivo* condensates (derived from CD-CODE)[34], and by defining the non-condensate forming, negative dataset, based on a protein-protein interaction network. Experimental validation of 24 proteins spanning a wide-range structural disorder confirmed that

18 of them localize to condensates with high confidence, while 3 form condensates with low confidence. Further, 8 of them co-localize with known biomolecular condensates. Thus, our experimental validation suggests an ~88% success rate in identifying condensate forming proteins. Moreover, we trained another model with extended set of features which include Gene Ontology terms (PICNIC$_{GO}$), which provides useful insights about specific protein functions that are enriched in proteins of biomolecular condensate, such as RNA-binding.

Although PICNIC was trained on the richest human data, it generalizes well to other organisms tested. Proteome-wide predictions by PICNIC estimate that ~40% of proteins partition into condensates across different organisms, from bacteria to humans, with no apparent correlation with organismal complexity or disordered protein content.

Results

Defining condensate-forming proteins

In order to develop a model to identify condensate-forming proteins, we assembled a ground truth dataset for *H. sapiens*, that has the most experimentally studied condensates of all organisms to date. Since we aimed at developing a binary classifier, we considered two classes of proteins: 1) proteins involved in condensates (positive dataset) and proteins not involved in condensates (negative dataset) (**Figure 1a**). The positive dataset was constructed from a semi-manually curated dataset of biomolecular condensates and their respective proteins, called CD-CODE (CrowDsourcing COndensate Database and Encyclopedia), developed by our labs[34]. CD-CODE compiles information from primary literature and from four widely used databases of LLPS proteins [25–28].

Building the negative dataset is a complicated task as there is no publicly available resource that reports proteins that do not form condensates. Additionally, condensates may form only under specific conditions[35]. Here, we defined the negative dataset based on protein-protein interaction network (InWeb database[36] for human proteins). We excluded all proteins that have direct connections with known condensate proteins. We reasoned that these proteins are potential condensate members that have not yet been studied. The remaining proteins comprised the negative dataset (**Figure 1a**). Of course, this procedure doesn't guarantee the absence of condensate proteins among the negative dataset (false negatives). But exclusion of the proteins that directly interact with proteins that were reported as members of synthetic or biomolecular condensates is lowering the probability of mixing positive and negative data. Overall, our non-redundant dataset (filtered by 50% sequence identity) contained 2142 positive and 1709 negative human proteins, which were divided by 4:1 ratio into training and test datasets.

PICNIC identifies sequence- and structure-determinants of condensate formation

We hypothesized that the ability to form condensates is encoded in the proteins' sequence and structure, and developed a machine learning classifier called PICNIC (Proteins Involved in CoNdensates In Cell) based on sequence-distance-based and structure-based features derived from Alphafold2 models (**Figure 1b**, in total 65 sequence-distance-based and 21 structure-based features).

It has been already shown that many proteins involved in condensates harbor intrinsically disordered regions (IDRs) and low-complexity sequences. Due to their inherent flexibility, multi-valency and ability to sample multiple conformations, they are adept at a wide array of binding-related functions including macro-molecular assemblies[9,37,38]. We also tested several metrics of disorder and sequence complexity as features (Supplementary Methods). Our final model contained several features related to disorder, such as IUPred scores[33], that have a feature importance of 0.5-3%.

Although the presence of highly disordered residues is among most important features (**Figure 1e,** pink), it is not a prerequisite for the protein to have long disordered domains to be a member of the condensate. The proportion of known condensate-forming proteins with no disordered regions in the human proteome is 21% (disordered regions < 10aa. **Figure S1**), while 33% of all human proteins have no disordered regions. For example, Human protein Guanine nucleotide exchange factor C9orf72 is a driver protein in stress granules; Speckle-type POZ protein is a driver in nuclear speckle and SPOP/DAXX body. Both proteins consist of ordered domains only that were experimentally determined by electron microscopy and X-ray crystallography, respectively (**Figure S1**, PDB ids 6LT0 ad 3HU6). Thus, both analysis of experimentally verified condensates and the selected features by our model suggest that disorder is not a necessity for condensate-forming proteins.

Along with overall sequence complexity and disorder scores of a protein, the secondary structure of individual residue types was also found to be important. We used the confidence score of the AlphaFold model prediction, the pLDDT score, that was shown to correlate with sequence disorder[39]. We represented the occurrence of an amino acid (AA) in a given secondary structure element (SSE) with a given model confidence as a triad (aa-SSE-pLDDT).

As amino acid composition bias and patterning of charges were shown to impact the ability of proteins to form condensates[30–32], we developed features that represent short and long range co-occurrences of amino acids in the protein sequence. We represent co-occurrence of amino-acids in the protein sequence within a distance (number of amino acids in linear sequence) by triads (AA_1, distance, AA_2). After feature selection, the long-range distance between charged amino acids, e.g. Lysine and Arginine (K,60, R) and Aspartic acid and Lysine (D,20,K), and short-range distance of Leucine and hydrophobic amino acids (e.g. L,0,W; F,2,L; L,2,L), and the distance between Cysteine and hydrophobic amino acids were shown to be the most important features. Overall Lysine and Leucine amino acids contribute the most to the model (**Figure 1c)**.

RNA-related functions are enriched in condensate proteins

In order to further increase to performance of our model, we integrated functional information that is already known about each protein and described as Gene Ontology (GO) terms. We developed a second classifier, called $\text{PICNIC}_{\text{GO}}$ that combines GO terms and the previously used PICNIC features (**Figure 1b)**. After feature selection only 18 features were included in the final model, 10 new GO features and 8 features of PICNIC described in the previous section. We found, that the most significant GO term is RNA binding, which is superseding the importance of other terms and features by several orders of magnitude (**Figure 1d**).

While the GO annotation feature is biased by existing knowledge, it nevertheless validates the design of the negative dataset: it highlights RNA binding molecular function as one of the of the most important features in $\text{PICNIC}_{\text{GO}}$ model, which is known to play crucial role in biomolecular condensate formation, as RNA molecules are significant constituents of condensates[9,40]. This feature is efficient in discriminating the two classes of proteins because the positive and negative datasets show different distributions of RNA binding annotation. In $\text{PICNIC}_{\text{GO}}$ the following functions were marked as the most important: transferase activity, transferring phosphorus containing groups, enzyme binding, phosphoric ester hydrolase activity, organic cyclic compound binding, and heterocyclic compound binding (**Figure 1d**). The success of including functional annotation demonstrates, that for proteins with unknown cellular compartmentalization, specific set of functional descriptions can provide additional evidence for their tendency to form biomolecular condensates.

PICNIC accurately identifies proteins involved in biomolecular condensate formation

Several data-driven predictors were developed in the last few years, that aim to predict proteins involved in LLPS from protein sequence alone or from sequence and experimental data, such as microscopy images[41]. Here, we compared the performance of PICNIC to sequence based predictors, PSAP[23], DeePhase[42] and two versions of PhaSePred[29], one general model (PdPS-8fea based on 8 features) and one developed for human proteins only which uses existing experimental data, such as fluorescent microscopy images of the proteins as well as experimental data on phosphorylation sites (PdPS-10fea based on 10 features).

We compared the performance of tools on three different datasets: 1) test dataset from the recently published PhaSePred methods[29]; 2) proteins forming nuclear punctae defined by the OpenCell project[43]; 3) test dataset generated from CD-CODE[34] (see Supplementary Methods, **Dataset S1**). Although the CD-CODE test data is not independent and was partially used by existing predictors during their training process, PICNIC has superior performance with a maximum F1-score of 0.81 **(Figure 2c)**. Not surprisingly, the models that use existing experimental information outperform the sequence-based predictors. Specifically, PICNIC$_{GO}$ has the best performance (ROC-AUC=0.91, F1-score=0.84) followed by PdPS-10fea (ROC-AUC=0.89, F1-score=0.83).

To further validate our model, we used microscopy images from Human Protein Atlas (HPA) where fluorescently labeled proteins were imaged and their cellular localization was determined[44]. Specifically, three types of cellular localizations were screened: nucleolus, centrosome and nuclear speckle. We filtered the list of proteins from HPA that were already in our training set that resulted in 484 proteins with known localization. Overall, PICNIC scores were higher for the proteins from HPA than for proteins without known localization (**Figure S3**). 69% of proteins mapped from HPA (with exclusion of the proteins from the training dataset) have a PICNIC score greater than 0.5, meaning that PICNIC correctly identified them as members of biomolecular condensates. It should be noted that HPA doesn't report if a protein does not belong to given condensate (negative examples). Therefore, this dataset can be used only to check model sensitivity (recall, what fraction of true condensate forming proteins were predicted correctly), but not model precision (what fraction of positive predictions are actually true positives).

PICNIC is robust in identifying small sequence perturbations that impact condensate formation

A real challenge of computational predictors is to be sensitive to small sequence perturbations that impact condensate formation. To test if PICNIC can distinguish similar sequences with altered condensate forming properties, we considered the synuclein family, that comprises three paralogs in human. Although they have similar sequences (**Figure 3a** and **c**, 60-70% identity) and structures as predicted by AlphaFold (**Figure 3b**), only α- and γ-synuclein form condensates *in vivo,* and only α-synuclein phase separates *in vitro*. Specifically, FITC-labeled β-synuclein, which lacks the characteristic NAC region of α-synuclein, does not phase separate at high concentrations (200 μM) and under crowding conditions (10% [weight/volume] PEG), whereas FITC-labeled α-synuclein forms condensates under the same conditions[22] . While α- and γ-synuclein can form amyloid-like fibers, β-synuclein does not[45,46]. Moreover, α- and γ-synuclein are part of biomolecular condensates: α-synuclein is reported to be the member of Synaptic vesicle pool condensate[47], γ-synuclein is a member of the Centrosome[48], but β-synuclein has not been found in any biomolecular condensates yet. PICNIC is the only method tested here which accurately predicts the *in vivo* condensate-forming ability of the synuclein family (**Figure 3d**).

Other methods either give the same score for all three paralogs and/or do not predict the correct tendency of condensate formation *in vivo*. We surmise that PICNIC is sensitive to structural rearrangements of proteins, and hypothesize that the bending of alpha-helix in β-synuclein potentially hinders the protein's ability to form condensate.

Interpretation of our machine learning models

To explore the generalisability of our models, we compared the most important features learnt by PICNIC vs. PICNIC$_{GO}$. First, we calculated the most divergent gene ontology features between the dataset of known condensate-forming proteins (CD-CODE) and the dataset of whole proteomes of corresponding organisms (**Figure S4**). Further, we compared these distributions to the distributions of potential condensate proteins predicted by PICNIC. This analysis demonstrated that PICNIC can recognize the protein properties deduced as most important by PICNIC$_{GO}$ model, as well as terms describing cellular localization which were excluded from PICNIC$_{GO}$ features (**Figure S5**).

We also compared the two models in the opposite direction to see whether PICNIC$_{GO}$ model can identify the sequence- and structure-based features considered as important by the PICNIC model. Interestingly, for different species different subsets of features were highlighted, but in concordance with features selected as most significant by PICNIC (**Figure S6**). Thus, PICNIC that was trained without Gene Ontology annotation can detect properties of proteins in biomolecular condensates captured by Gene Ontology terms for different species and vice versa: PICNIC$_{GO}$ detects properties of proteins in biomolecular condensates captured by distance-based and AlphaFold-based features for different species, further validating the generalizability of the model. Overall, the propensity of a protein to be a member of biomolecular condensate seems encoded in the sequence, and machine learning models, such as PICNIC can recognize these encoded properties.

Experimental validation of predicted condensate-forming proteins

In order to experimentally validate our model, we decided to predict the condensate localization of poorly characterized human proteins and sought to validate their condensate-forming behavior inside living cells. To do so, we chose 24 proteins (**Dataset S2**) which: i) cover diverse molecular functions spanning the entire central dogma of molecular biology and regulation of all the major cellular bio-polymers for instance, nucleic acids, proteins and chromatin (**Figure S7a**), ii) represent the average sequence length of human proteins (i.e., around 350 amino acids) by having a range of 125 to 684 amino acids (**Figure S7b**), iii) represent diverse 3D structures from ordered, alpha-helical, beta-stranded to highly disordered (**Figure 4c**) iv) are involved in genetic diseases (*AIMP1, CWC27, RP9, LMOD1*) as well as host-virus interaction (*IF2GL*). Overall, the 24 proteins, we chose for experimentally verifying and benchmarking PICNIC, represent global cellular functions and therefore are suitable to demonstrate how robust our machine learning model is in predicting condensate-forming proteins across entire proteomes.

We cloned 24 transgenes and transfected them in U2OS cells expressing fluorescently labelled proteins (see Supplementary Methods). Using fluorescent imaging, we found, that 21 out of the 24 tested proteins (87.5%) localized to mesoscale foci without any stressors, while 3 proteins (*C1ORF52, SPAG7 and CWC27,* encircled in red) localized to the nucleoplasm without forming any discernible mesoscale foci (**Figure 4a**). Foci were defined based on enrichment in fluorescent intensity, i.e., the intensity ratio inside relative to outside the foci is greater than one (**Figure S8a**). Only three proteins tested have no detectable foci, and 21 form foci (**Figure 4a**).

In order to classify the observed foci as biomolecular condensates, we aimed to define quantitative characteristics and thresholds. We measured four simple characteristics from fluorescent microscopy images (**Figure S8**): area and perimeter, informing on the size and the typical number of proteins in a foci); shape (roundness); number of foci per cell. Next, we decided on a threshold for these characteristics to aid a quantitative definition of condensates. We consider foci as condensates above the diameter of 350 nm (distance between two furthest pixels in one condensate), that is well above the diffraction limit. This would correspond to at least ~1μm perimeter assuming a near round shape (**Figure S8c**). Using back-of-the-envelope calculations, we can consider an average protein size as 10 nm^3, then a 1 μm^3 compartment can contain ca. 1 million protein molecules and a 500 nm^3 compartment can contain 100,000 protein molecules. We note, that other super resolution techniques are required to characterize the size of clusters of proteins below the diffraction limit.

Experiments confirm 87.5% of PICNIC predictions

By applying the above definition, in our dataset, 75 %, *i.e.* 18 proteins (encircled in green) form high confidence condensates, 12.5 % i.e. 3 proteins (encircled in orange) form low confidence condensates (foci with perimeter <1 μm) while other 3 are not forming any condensates (fluorescent intensity ratio ~ 1). We observe most condensates to be round (**Figure S8d**). The number of condensates per cell varies between 1 to 100s depending on the protein of interest. Using co-localization experiments with known nucleolus and P-body markers, we found that 5 proteins (*R51A1, H2A1H, H1T, CWC27, MRPL1*) can localize to the nucleolus, a well-characterised liquid-like nuclear condensate (**Figure 5a**). Further, 2 proteins (*PHP14* and *HBS1L*) localize to processing bodies (P-bodies), another well-characterized cytoplasmic condensate. In addition, 8 other proteins (*RS10L, TYW5, SPA24, AIMP1, ZC3H15*, *IF2GL, LMOD1, RPS4Y2*) localized to cytoplasmic bodies and rest (*KHDC4, CWC25, PolD3, RAMAC, DRC4, RP9*) localize to the nuclear bodies (**Figure 5a**). FRAP recovery profiles for the nuclear body forming proteins revealed, that 6 out of 7 tested proteins show very fast dynamics while one (*RBMY1D*) shows no recovery at the indicated time (**Figure 5c**).

Other popular tools to predict condensate proteins would fail to make correct predictions for many of these proteins as shown by the high misclassification rates (**Figure 4d**). Overall, 87.5% of PICNIC predictions were found to be correct (misclassification rate is 25% for high confidence condensates and 12.5% if we include both high and low confidence condensates) in our experimental assays validating the model.

Proteome-wide predictions detect no correlation of predicted condensate proteome size with disorder content and organismal complexity

In order to see if our model is generalizable, we tested its performance in identifying known condensate-forming proteins of other organisms. The CD-CODE database[34] was screened to evaluate the fraction of proteins that were correctly identified as members of condensates by the developed model. Although PICNIC was trained on human data, it successfully predicted 72% of such proteins in mouse and 86% in Caenorhabditis elegans for example (**Figure 6a**). Thus, PICNIC model is species-independent and can be used in different organisms to assess the ability of proteins to be involved in biomolecular condensates.

To estimate the overall fraction of condensate-proteins, we calculated PICNIC scores for 14 different organisms across the tree of life including bacteria, plants and fungi (**Figure 6c**). We chose organisms that have already known condensate protein members that were experimentally verified in CD-CODE. We excluded organisms from further analysis where the number of known proteins is too small to compute statistics on performance: *Danio rerio* (N=14), *Dictyostelium discoideum* (N=2), *Escherichia coli* (N=6), *Mycobacterium tuberculosis* (N=1), *Oryza sativa* (N=1), *Candida albicans* (N=6). We found that the proportion of the predicted condensate-forming proteome is 40-60%, and is similar across related organisms, e.g., 42% and 39% in human and mouse, respectively (**Figure 6c**). Interestingly, while the fraction of disordered proteins increases with organismal complexity as shown before[49,50], we found no correlation between fraction of predicted condensate proteins in a proteome and the disordered protein content (**Figure 6b**). For example, *E. coli* and *H. sapiens* have both ~40% of their proteome predicted to be involved in biomolecular condensates **(Figure 6c**).

Discussion

Here, we present a machine learning classifier that can learn and decode biomolecular condensate forming behaviour of proteins, that goes beyond protein structural disorder. We developed two models for the prediction of proteins involved in condensates *in vivo*, that reach precision of 77% (81% for recall) and 87% (82% for recall) for PICNIC and PICNIC$_{GO}$, respectively at the suggested score threshold of 0.5 (**Figure 2**). The success of PICNIC models relies on several innovations: 1) novel amino acid co-occurrence features combined with protein structure-based features, that became only possible on a proteome scale since the Alphafold2 revolution[51,52]; 2) gradient boosting classifier that has an appropriate model complexity that fits the size of the training dataset; 3) improved curation and definition of the positive and negative datasets. Specifically, PICNIC models benefit from high quality positive data, that is the manually curated database of biomolecular condensates CD-CODE[34]. Additionally, we designed the negative dataset based on no *a priori* assumptions about protein disorder/structure. Previous predictors used either a i) exclusion of positive dataset from all known proteins[29,42] or ii) proteins with 3D structures (retrieved from PDB) since phase separation is common in disordered proteins that do not have well-defined 3D structures. The simple exclusion does not provide reliable negative dataset, as it may contain many potentially phase separating proteins that have not been discovered yet. The second approach generates a biased negative dataset and is problematic because of two reasons: 1) phase separating proteins can have well-defined structure (**Figure S1), Figure 4c**), 2) such dataset is biased towards set of properties inherited by experimentally solved proteins. E.g., disorder predictors would also use 3D structures as negative dataset. To resolve this issue, here we used a protein-protein interaction network-based approach and excluded proteins that have a connection in the network with known condensate proteins (**Figure 1a**).

Gradient boosting methods (GBM) allow optimization on different loss functions which provides necessary flexibility, but more prone to overfitting. Here, the latter was compensated by choice of tree depth and providing of an evaluation dataset. GBM generally outperforms simpler models such as Random Forest or Support Vector Machine, but at the same time it doesn't require as much data as models based on Neural Networks. Because there is not enough well-annotated data for positive and negative datasets, we chose GBM to mitigate the tradeoff between model complexity and its performance.

With the incorporation of biological knowledge about genes and proteins, PICNIC$_{GO}$ reaches even higher performance (**Figure 2c**). Not surprisingly, it identified, along with the known impact of disordered and low complexity regions, that RNA binding seems to play crucial roles in the ability of proteins to be involved in biomolecular condensates (**Figure 1c**). Although gene ontology terms are valuable resource of information, they could introduce bias due to their nature of annotation.

We would like to emphasize that some of the other tools we benchmarked against were designed to solve a different task, namely to predict the ability of a protein to undergo phase separation or drive phase separation and *in vitro* condensate formation[22,42]. Specifically, PSAP[23] was trained on

90 human proteins that can drive phase separation, and FuzDrop[22] was parametrized based on 67 proteins known to self-phase separate based on in vitro experiments. Therefore, it is not expected that these tools would identify all condensate forming proteins, e.g. clients observed in vivo such as the nuclear punctae from the OpenCell project (**Figure S2b**). In contrast, PICNIC was adapted to recognize the proteins that are present in biomolecular condensates (serving a role of a driver or client) regardless of the mechanism of condensate formation. Therefore, our model does not evaluate if a protein is part of synthetic condensates, i.e., *in vitro* experiments, but rather focuses on if a protein is part of condensates in biologically relevant conditions.

PICNIC scores for clients and drivers show similar distribution (**Figure S9**). Since the same protein can behave as a driver or a client depending on the condensate identity and environment, we surmise that the ability of driving condensate formation is not solely encoded in the sequence and structure of individual proteins. This suggests, that additional input data are needed to be taken into account by the model. Studying the protein in the context of its interacting partners within a condensate may shed light on the properties of driver proteins.

While PICNIC is superior in identifying protein membership in biomolecular condensates, it has several limitations. It cannot predict which sequence features or motifs are responsible for the condensate function in a given protein. FuzDrop is the only method that can recognize specific motifs in protein sequences that promote phase separation behavior[22,53–55] .

The algorithms in this paper are designed to understand whether a protein has the potential to localize to a condensate, based on previous data. We tested 24 predicted proteins and found that 87.5% form foci *in cellulo*, and 75% form high confidence condensates based on a quantitative definition of condensates. In making this analysis, we set a cut-off above the diffraction limit. This does not mean that clusters of proteins under this limit do not form condensates, but that super-resolution techniques would be required to analyze this. It is important to state that this localization makes no claim to biological function in cells. These condensates were observed in cell lines, using GFP-tagged proteins that had varied expression levels. It is possible that some proteins will only form condensates under stress or in certain cell types at certain concentrations. In some cases, a protein for instance might not localize to a condensate under physiological conditions, but might when overexpressed in cancer. Therefore, although each protein can localize to a condensate, it remains to be sorted out by detailed experiments whether any individual protein localizes to a condensate in a specific cell type at a certain concentration and using different tags or antibodies.

We detect that structural disorder is not a prerequisite for condensate member proteins, as many of them have no IDRs both based on analysis of the CD-CODE database (**Figure S1b**) as well as 5 out of the 21 proteins identified here experimentally as condensate members have <30% disordered residues (**Figure 4c**). Thus, a wide-range of structural disorder can lead to condensate partitioning. Accordingly, we found no correlation of condensate proteome size and disorder content of an organism.

The generalizability of our model shows that the predictor learned general features (based solely on sequence information and structures that were also deduced from the sequence) across the tree of life. The provided results can shed light on evolution of biomolecular condensates across different species and by predicting condensate members beyond drivers can aid identifying potential protein targets to modulate biomolecular condensate behaviour and aid drug design[56,57].

Overall, PICNIC accurately predicts proteins involved in biomolecular condensates and provides proteome-wide perspective on proteins involved in condensate formation in different species.

Methods

Construction of positive and negative datasets

The positive dataset of condensate forming proteins was extracted from CD-CODE database v1.00[34]. We used all human proteins with at least 1 evidence star as a positive dataset. The negative dataset was constructed by excluding proteins that interact with known condensate forming proteins based on the InWeb v3 database[36]. After filtering the sequence for 50% sequence identity, our dataset composed of 2142 positive and 1709 negative proteins, that were divided by ratio 4:1 into training and test datasets. We divided the dataset into 20% test dataset (used only for testing the final model) and 80% working dataset, which was randomly divided into 70% training and 30% validation datasets for the 10-fold cross-validation (**Dataset S1**).

Model features

Disorder score and sequence complexity

Intrinsically disordered regions (IDRs) and low-complexity regions of proteins were shown to be an important feature for predicting the ability to phase separate. To estimate protein disorder, we used the IUPred algorithm[58], which assigns a score to each residue in the sequence. We used k^{th} percentile with k equal to [5,25,50,75,95], which is the score below which k percentage of residue scores fall. The 5th and 95th percentiles were chosen instead of minimum/maximum values to exclude the bias due to outliers.

We calculated sequence complexity in order to identify low-complexity regions (LCRs). These regions often contain repeats of single amino acids or short amino acid motifs. We calculated sequence complexity according to the definition suggested by Wootton and Federhen[59] using two different sizes of sliding window for the protein sequence: 40 and 60. Proteins with length less than 60 residues were excluded from this analysis.

Sequence distance-based features

We represented the co-occurrence of amino-acids (AA) in the protein sequence within certain distance by a pair of triads (AA_1, $distance_{short}$, AA_2) and (AA_1, $distance_{long}$, AA_2), where AA_1, AA_2 is one of the 20 amino acids types, $distance_{short}$ э [1,2,3,4,5] and $distance_{long}$ э [[0,20),[20,40),[40,60),[60,80)]. These features represent short and long range co-occurences of amino acids in the protein sequence. Short threshold $distance_{short}$ defines the distance between two types of amino acids (e.g., 1 means neighboring residues). Long threshold, $distance_{long}$ defines a window equal to 20, that is the distance in sequence between the two types of amino acids co-occur. The total number of tested sequence distance-based features was 1890, including both short and long distances.

Secondary structure features based on AlphaFold models

Recent advances in Deep Learning techniques enabled *de novo* modeling of protein structures from sequences with high accuracy that is comparable to experimental methods[51]. Here, we used predicted AlphaFold2 models that were downloaded from the resource created by the EMBL Consortium (second release, Date of access: January, 2022)[52], which contains precomputed structures for proteomes of many organisms. Alongside with atomic protein structures, AlphaFold provides the pLDDT score (predicted lDDT-Cα), that is a per-residue measure of local confidence on a scale from 0 – 100 [60]. pLDDT scores were divided into four classes (according to DeepMind classification): [0, 50) - 'very low', [50,70) - low', [70,90) –'confident'', [90,100] – 'very high'. We used the STRIDE algorithm to annotate the secondary structure based on 3D protein structure [61]. STRIDE assigns one of seven classes to each amino acid in the protein sequence: Alpha helix (H), 3-10 helix (G), PI-helix (I), Extended conformation (E), Isolated bridge (B), Turn (T), Coil (C).

Next, we calculated all possible triads in the form (AA, SSE, pLDDT), where aa belongs to one of 20 types of amino acids, SSE э [H, G, I, E, B, T, C], pLDDT э ['very low, ‘low’, ’confident', ‘very high’]. For longer proteins, AlphaFold2 models consist of overlapping segments of 1400 aa length. In case of discrepancy, when the same amino acid is assigned with different 3D coordinates and pLDDT score, we consolidated the predictions using the following rules: 1) We calculated all possible STRIDE predictions (with different pLDDT scores); 2) we selected the most frequent STRIDE class that had the highest pLDDT score. The total number of features based on structural information provided by Alphafold2 and STRIDE was 560.

Gene Ontology features

Gene Ontology terms are a hierarchal dictionary of annotations describing the function of a particular gene[62,63]. They assign gene characteristics for three directions: molecular function, biological process and cellular component. Each of the direction is represented by a directed acyclic graph, where nodes represent terms (or annotations) and edges represent the relationship of subtype from descendant to ancestor node. There are tens of thousands of terms, therefore one hot encoding for all possible terms is not feasible. To decrease the number of encoded terms, we took only the most frequent terms into account. To estimate the frequency of terms, we used the Swissprot annotations of proteins (after removing redundancy by excluding sequences with more than 30% sequence identity) from Uniprot database[64]. It must be noted, that the frequency of ancestor term t_a also included summation of the frequencies of all descendant terms t_d, calculated by the equation:

$\mathrm{Fr}(t_a) = \mathrm{N_{oc}}(t_a)+\sum_{t_d} N_{oc}(t_d)$, where $\mathrm{N_{oc}}(t)$ – number of occurrences of term t in the corpus.

Only terms with frequencies greater than a threshold were kept for feature calculation. Thresholds were chosen for each of the three directions separately to encompass adequate number of terms in the annotations (2500 for molecular function and biological process). We excluded the cellular component direction from the feature selection, as some gene ontology terms contain information about cellular compartments (based on Human Protein Atlas [44], OpenCell [43]). We used one-hot encoding, where each protein was assigned with a vector of fixed length (equal to the number of chosen terms); 1 was assigned if the considered term in the given position (or any of their descendants) was mentioned in protein annotation, otherwise 0. The total number of features describing molecular function and biological process for the set of proteins was 1002.

Machine learning algorithms

We developed two types of models: one including gene ontology features ($\mathrm{PICNIC_{GO}}$) and one without (PICNIC). Both models have the following structure: they consist of 10 Catboost classifiers with dataset for early stopping (to calculate loss function on the dataset different from training to prevent overfitting over training dataset) and fixed depth[65,66]. Catboost is a classifier based on gradient boosting machine (GBM) – a machine learning technique that gives a prediction model in the form of an ensemble of decision trees[67].

Among other gradient boosting classifiers CatBoostClassifier from catboost library showed consistency across multiple runs (we compared LGBMClassifier from lightgbm library and XGBClassifier from xgboost library). To estimate the overall model performance across multiple runs with different parameters, we used the following metric: we chose the validation score (F1-score) of best iteration of each separate fold, and then computed the mean value across 10 folds. The model training started with all features, that is 2467 for the model without Gene Ontology features (PICNIC), and 3469 for the model with GO features, $\mathrm{PICNIC_{GO}}$. We selected the best features based on feature importance: at each training iteration only features with importance greater than a given threshold were selected for the next run (we took the union of features across different folds). Thus, each subsequent training iteration decreased the number of used features. We chose this feature selection approach because the feature importance did not fluctuate much for different folds (**Figure 1e**). The final models contained 18 and 92 features for PICNIC$_{GO}$ and PICNIC, respectively. The number of features in the model with protein annotations are much lesser in comparison with another model as subset of the sequence- and structure-based features connected to phase separation properties are already directly encoded by Gene Ontology annotations.

Cloning, cell culture and imaging

U2OS cells were transfected with plasmids encoding Human proteins tagged with iRF-670 at the N-terminus. All the 24 tested genes were codon optimized (for synthesis ease as well as to override any cellular regulation involving mRNA degradation of endogenous sequences) and synthesized by Integrated DNA Technologies (IDT) (**Dataset S3**), restriction digested using NotI-HF and AscI enzymes (NEB-R3189 and NEB-R0558) and then ligated into the pre-digested vectors using T4 DNA ligase(M0202) and transformed in *E. coli* DH5-alpha cells. Positive clones were confirmed by insert release and correctness was verified by DNA sequencing.

Cells were grown in high glucose DMEM medium (Gibco-31966021) supplemented with 10% FBS (Sigma- S0615) as well as 100 units/ml Penicillin-Streptomycin (Gibco-15140112). Following trypsinization with Trypsin-EDTA (Gibco-25300054) cells were seeded on the Ibidi 8 well imaging chamber (80826). After overnight growing the cells, cells were then transfected with plasmids encoding the 24 transgenes and for the colocalization purpose with GFP-tagged DCPa1 (a processing-body marker plasmid) and GFP-tagged Fibrillarin (a sub-nucleolar marker specific to the middle dense fibrillar region allowing the potential dissection of all 3 sub-nucleolar localization of the test proteins) using the Fugene HD transfection reagent (Promega-E2311).

Cellular fluorescent images were recorded on a spinning disk confocal microscope with FRAP capability using the 60x/1.2U-Plan-SApo, water immersion objective lens (Olympus) using 488nm and 640nm laser lines, on an Andor-iXon-897-EMCCD camera. Images, colocalization and the FRAP movies were analyzed and representative images were prepared using Fiji[68] (**Dataset S4**).

Fluorescent microscopy images were quantified manually using Fiji and shape, size descriptors (roundness, perimeter and area) as well as fluorescence intensity were measured (**Figure S8)**. Enrichment ratio of the condensates was calculated as the ratio of the mean fluorescence of the condensate foci divided by the mean fluorescence of the background. The data was plotted using GraphPad prism software. All the images were finally prepared in Adobe Illustrator.

Data and Software Availability

The training, validation and test datasets are available as Dataset S1. The list and properties of the 24 proteins selected for experimental validation is provided as Dataset S2. Plasmid vector maps and representative images are available as Dataset S3 and S4, shared as a public repository available at https://edmond.mpdl.mpg.de/dataset.xhtml?persistentId=doi:10.17617/3.0Y9Q8N. The PICNIC code, documentation and examples as jupyter notebooks can be found at https://git.mpi-cbg.de/tothpetroczylab/picnic. Predictions across proteomes of 14 organisms are provided as a web application https://picnic.cd-code.org.

Acknowledgements

A.H. was funded by the ELBE postdoctoral fellowship. H.R.S thanks the Nomis foundation for financial support. A.T.-P. was funded by the MPG RGL funds. N.R. were founded by Dewpoint Therapeutics. We would like to thank the Computer Services and Scientific Computing Facilities of the MPI-CBG for their support, especially to Oscar Gonzales for supporting our HPC and HongKee Moon for developing the PICNIC web application. We thank Andrej Pozniakovski for molecular biology support.

Author Contributions

AH, HRS, AAH and ATP designed research; AH and HRS performed research; SG and NR contributed new reagents or analytic tools; AH, HRS, SG and ATP analyzed data; AH, HRS, AAH and ATP wrote the paper.

Competing Interest statement

A.A.H. is a founder and shareholder of Dewpoint Therapeutics. The remaining authors declare no competing interests.

Main Figures

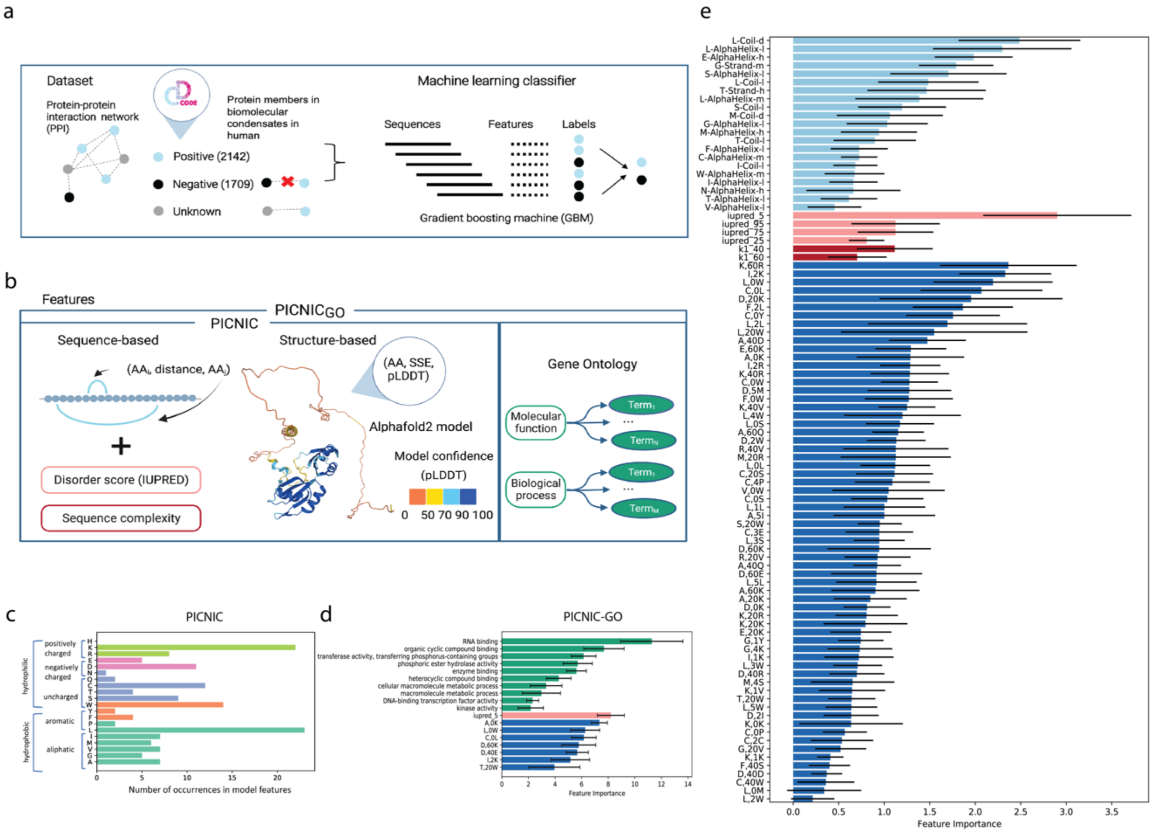


Figure 1. Development of PICNIC (Proteins Involved in CoNdensates In Cells) algorithm.

a) In order to construct a training dataset, we annotated the known condensate-forming proteins from CD-CODE (positive dataset, members of biomolecular condensates) on the protein-protein interaction (PPI) network, and we excluded their first connections (proteins having interactions with condensate proteins). The remaining proteins comprised the negative dataset. Gradient boosting machine was used to distinguish two classes of proteins: members of biomolecular condensates and proteins that are not involved in any type of biomolecular condensate. **b)** Sequence, structure and function-based features of PICNIC. We developed two types of models: with (PICNIC$_{GO}$) and without the use of gene ontology annotation features (PICNIC). Sequence-based features included sequence complexity, disorder score (IUPred), and features based on amino acid co-occurrences. Structure-based features based on Alphafold2 models included the pLDDT score, a per-residue measure of local confidence on a scale from 0 – 100 (colored on the structure). We annotated the secondary structure (SSE) based on 3D protein structures using STRIDE and all possible triads in the form (AA, SSE, pLDDT) were calculated. **c)** Amino acid occurrences in the features of PICNIC model show that Leucine and Lysine contribute the most to the model predictions. **d)** The model including GO terms (PICNIC$_{GO}$) used three types of features, based on gene ontology (in green), disorder (orange), distance-based (violet). The most important features included RNA-binding, disorder and co-occurrences of charged and hydrophobic residues. The feature importance is consistent across different folds of cross-validation (mean values across 10 folds are shown, black lines with upper and lower border show standard deviation for each feature within cross-validation). **e)** Feature importance of PICNIC is consistent across different folds. Mean values across 10 folds are shown, black lines with upper and lower border show standard deviation for each feature within cross-validation. Features constitute three groups: based on AlphaFold2 models (in light blue), disorder (pink), complexity (dark red) and distance-based (blue).

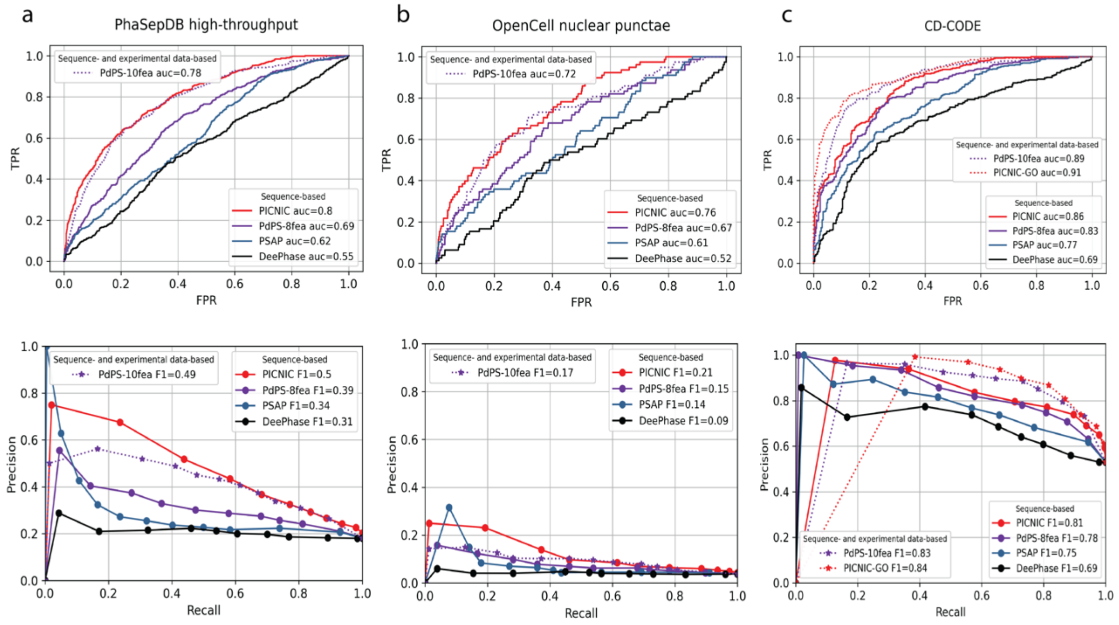
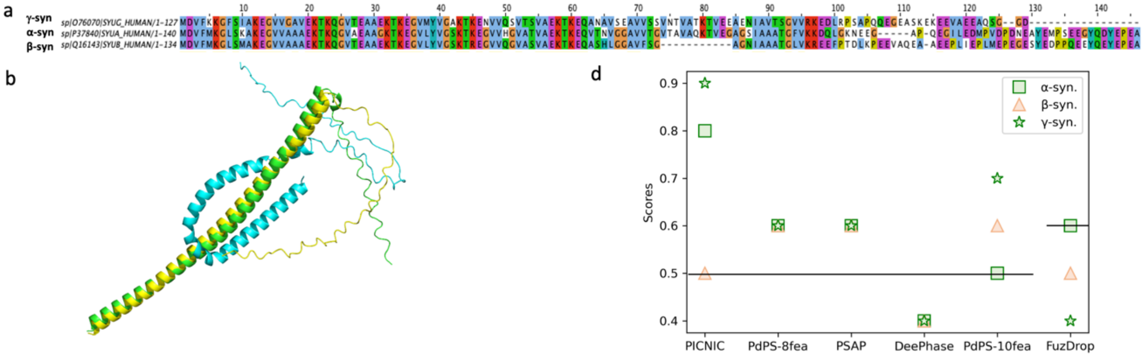


Figure 2. PICNIC models have the best performance in predicting condensate-forming proteins. Comparison of sequence-based predictors (lines, PICNIC, PdPS-8fea, PSAP, and DeePhase) and predictors using experimental data (dotted lines) as features to predict protein condensates. Specifically, PdPS-10-fea uses phosphorylation sites and immunofluorescent microscopy images of the proteins, PICNIC$_{GO}$ uses GO-terms as features. **a)** Test dataset from PhaSepDB high-throughput retrieved from[29] (441 positive and 1998 negative examples, excluding proteins that were part of the PICNIC training set), **b)** test dataset from OpenCell[43] (78 positive and 1998 negative examples excluding proteins that were part of the PICNIC training set), **c)** test dataset from the current study based on CD-CODE[34] (338 positive and 299 negative examples). PICNIC outperforms sequence-based predictors even on the test set that includes training data of previously published predictors, that may inflate their performance.



c

	Sequence identity			PICNIC	PdPS-8fea	PSAP	FuzDrop	DeePhase	PdPS-10fea	*In vivo* evidence	*In vitro* evidence
α-syn.	1.0			0.8	0.6	0.6	0.6	0.4	0.5	In vivo (in Synapsin Condensates [47])	Phase separation, amyloid fiber formation [22,45,46]
β-syn.	0.72	1.0		0.5	0.6	0.6	0.5	0.4	0.6	Not reported in biomolecular condensate	No phase separation, no amyloid fiber formation [22,45]
γ-syn.	0.66	0.63	1.0	0.9	0.6	0.6	0.4	0.4	0.7	In vivo (in Centrosome [48])	No phase separation, amyloid fiber formation [45,46]
	α-syn.	β-syn.	γ-syn.								

Figure 3. PICNIC captures the different phase separation behaviour within the synuclein protein family. a) The three paralogs in human share high sequence identity as depicted in the multiple sequence alignment. **b)** Structural models for α-synuclein (yellow), β-synuclein (cyan) and γ-synuclein (green), predicted by AlphaFold2 reveal that β-synuclein has a bent structure. **c)** Despite the high sequence similarity, only α- and γ-synuclein are part of biomolecular condensates, while β-synuclein has not been found in any biomolecular condensates yet and was shown not to phase separate *in vitro*. **d)** Comparison of prediction scores of different tools in identifying condensate forming (α and γ, green) and non-condensate forming paralog (β, red). PICNIC accurately predicts the condensate-forming ability of the synuclein family, and ranks β-synuclein the lowest, while other tools give equivalent scores to all paralogs or fail to identify the right trend. Vertical lines indicate the threshold used by the various methods to classify condensate-forming proteins.

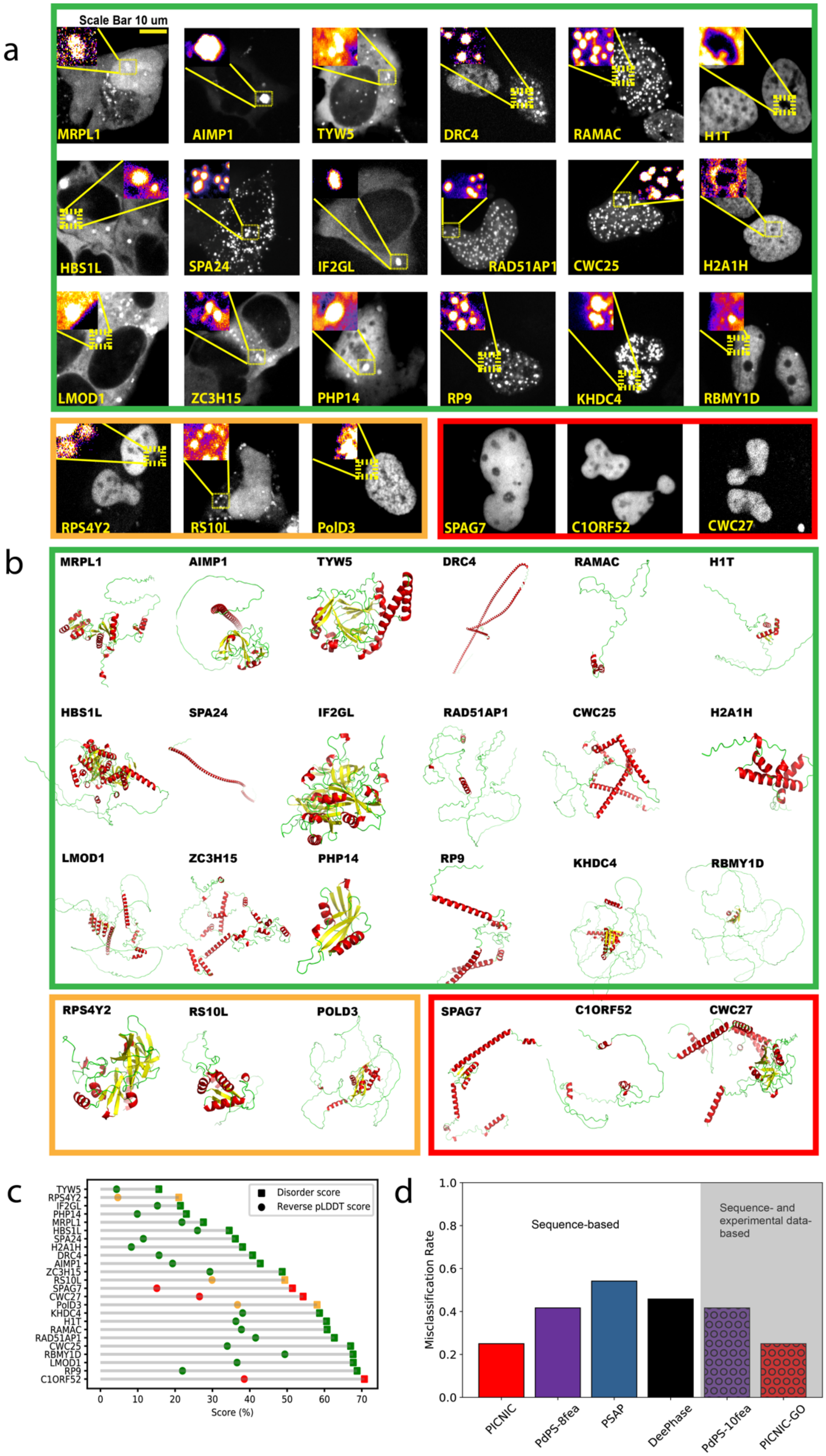

a
Scale Bar 10 um
MRPL1
AIMP1
TYW5
DRC4
RAMAC
H1T
HBS1L
SPA24
IF2GL
RAD51AP1
CWC25
H2A1H
LMOD1
ZC3H15
PHP14
RP9
KHDC4
RBMY1D
RPS4Y2
RS10L
PolD3
SPAG7
C1ORF52
CWC27
b
MRPL1
AIMP1
TYW5
DRC4
RAMAC
H1T
HBS1L
SPA24
IF2GL
RAD51AP1
CWC25
H2A1H
LMOD1
ZC3H15
PHP14
RP9
KHDC4
RBMY1D
RPS4Y2
RS10L
POLD3
SPAG7
C1ORF52
CWC27
c
TYW5
RPS4Y2
IF2GL
PHP14
MRPL1
HBS1L
SPA24
H2A1H
DRC4
AIMP1
ZC3H15
RS10L
SPAG7
CWC27
PolD3
KHDC4
H1T
RAMAC
RAD51AP1
CWC25
RBMY1D
LMOD1
RP9
C1ORF52
■ Disorder score
● Reverse pLDDT score
0
10
20
30
40
50
60
70
Score (%)
d
1.0
0.8
0.6
0.4
0.2
0.0
Misclassification Rate
Sequence-based
Sequence- and experimental data-based
PICNIC
PdPS-8fea
PSAP
DeePhase
PdPS-10fea
PICNIC-GO

Figure 4. Most (18 out of 24) tested proteins form high confidence condensates *in cellulo*.

a) Representative images of the U2OS cells expressing the tested proteins tagged with a fluorescent protein (GFP). Formation of mesoscale cellular condensates are highlighted in the inset. All images are scaled to the scale bar 10 micron (shown on upper left image). We found 21 out of the 24 tested proteins (87.5%) formed mesoscale foci without any stressors, while 3 proteins (*C1ORF52, SPAG7 and CWC27,* encircled in red) localized to the nucleoplasm without forming any detectable foci (foci were defined by exhibiting a fluorescent intensity ratio >1). Notice the presence of rim like structures in case of *H1T* and *H2A1H*. Using size, shape and the fraction of cells forming mesoscale foci as a deciding characteristic, ~75% i.e., 18 proteins (encircled in green) form high confidence condensates, and 3 proteins (encircled in orange) form low confidence condensates (foci with longest diameter <350nm. **Figure S8**). **b)** Wide range of structural motifs covered in the test proteins; AlphaFold2 structural models of the proteins are colored according to secondary structures. Notice the wide range of structural motifs, alpha-helical (red), beta stranded (yellow) to largely disordered (green) proteins.

c) Disorder content (computed as mean IUPred score or reverse pLDDT score (1 - pLDDT)) of the tested protein does not correlate with the ability to form condensates. **d)** Comparison of the predictions provided by sequence-based predictors (bars, PICNIC, PdPS-8fea, PSAP, and DeePhase) and predictors using experimental data (hatched bars) as features to predict protein condensates. Both PICNIC and PICNIC$_{GO}$ exhibit the lowest misclassification rate for the tested 24 proteins.

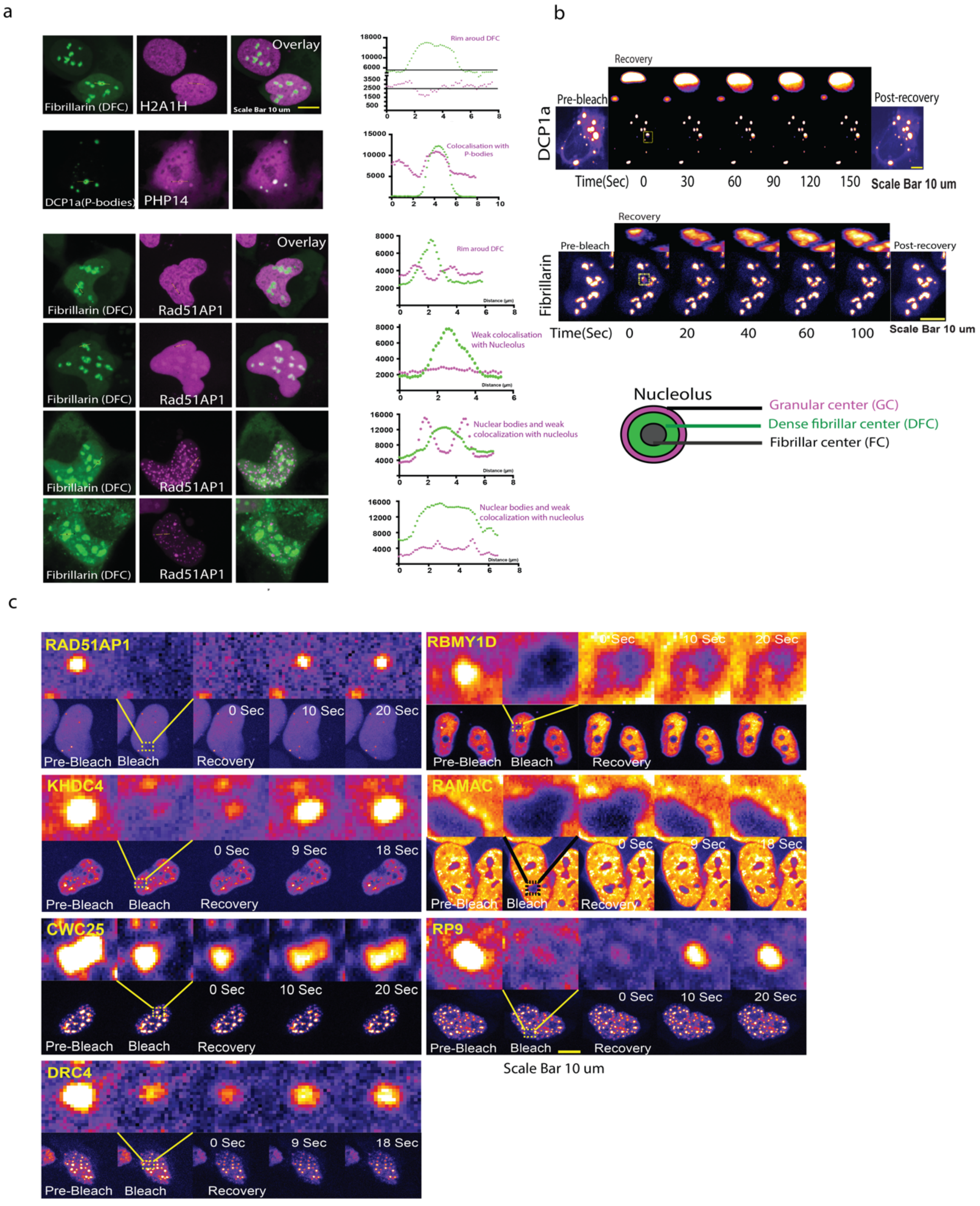


Figure 5. A subset of the tested proteins localizes to known condensates.

a) Colocalization of the cellular condensate-forming proteins with well-characterized liquid like cellular condensates as can be concluded from the fluorescence intensity profiles correlation with the marker protein fluorescence profiles. While *Rad51AP1* (in purple) localizes strongly around the Dense fibrillar center (DFC, in green) forming a rim like structure, *H2A1H* (in purple) show rather weak localization as a rim around the DFC (in green) showing sub-nucleolar localization specific to the outer Granular center (GC). See the cartoon representation of the nucleolar architecture. Further, *PHP14* (purple) strongly co-localizes with the DCP1a (green) labelled processing bodies. RAD15-associated protein 1 (*Rad51AP1*) shows multi-condensate localization that varies from exclusive nuclear bodies, to nuclear bodies abutting the nucleolus, to sub-nucleolar localization (GC) as well as complete nucleolar localization suggesting an interesting regulated role for this protein’s involvement on multiple nuclear-condensates possibly in a cell-cycle stage regulated manner. All images are scaled to the scale bar 10 micron, upper right. **b)** FRAP assays showing the fast recovery dynamics consistent with the liquid like nature of the P-bodies (upper panel) and the Nucleolus (lower panel). Fibrillarin and the DCP1a FRAP recovery profile, inset highlighting the fast recovery dynamics of targeted P-body and the nucleolus. The scale bar is 10 micron each for each. **c)** FRAP recovery profiles for the nuclear body forming proteins. 6 out of 7 tested proteins show very fast dynamics while one (*RBMY1D*) shows no recovery at the indicated time. All the images with indicated whole nuclei (lower panel for each protein) are scaled to the 10-micron scale bar shown in the bottom right.

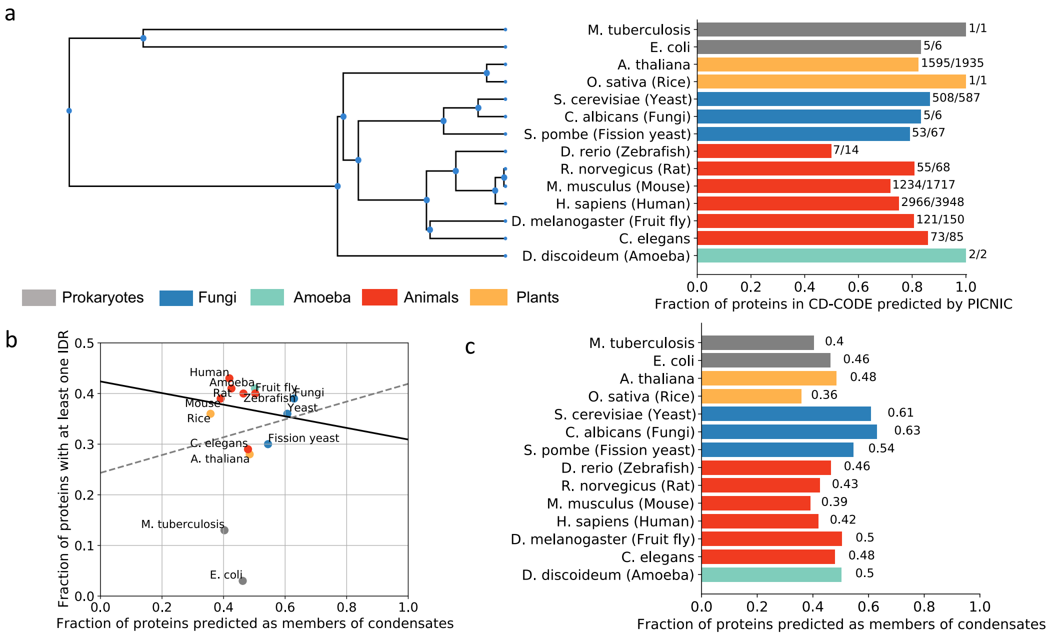


Figure 6. Inferring condensate proteins across the tree of life reveals no correlation with disorder content.

a) PICNIC model is species-independent. We validated the PICNIC model on known condensate proteins from different species (defined by CD-CODE). PICNIC correctly identified 70-100% of known condensate proteins of all species tested, except for zebrafish (50%). **b) Disorder content and fraction of condensate-forming proteins of a proteome are not correlated.** Proteome-wide prediction of proteins in biomolecular condensates by PICNIC predictor (**c**) compared with the fraction of disordered proteins (proteins with at least one disordered region of >=40 residues) in proteomes shows no correlation across many organisms from bacteria to mammals ($R^2 = 0.014$ for all datapoints, grey dotted line, and $R^2 = 0.033$ when excluding *E. coli* and *M. tuberculosis*, black solid line).

References

1. Tsang, B., Pritišanac, I., Scherer, S. W., Moses, A. M. & Forman-Kay, J. D. Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations. *Cell* **183**, 1742–1756 (2020).
2. Alberti, S. & Dormann, D. Liquid-liquid phase separation in disease. *Annu. Rev. Genet.* **53**, 171–194 (2019).
3. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **357**, (2017).
4. Cerase, A., Calabrese, J. M. & Tartaglia, G. G. Phase separation drives X-chromosome inactivation. *Nat. Struct. Mol. Biol.* **29**, 183–185 (2022).
5. Iserman, C. *et al.* Condensation of Ded1p Promotes a Translational Switch from Housekeeping to Stress Protein Production. *Cell* **181**, 818-831.e19 (2020).
6. Klosin, A. *et al.* Phase separation provides a mechanism to reduce noise in cells. *Science* **367**, 464–468 (2020).
7. von Appen, A. *et al.* LEM2 phase separation promotes ESCRT-mediated nuclear envelope reformation. *Nature* **582**, 115–118 (2020).
8. Lyon, A. S., Peeples, W. B. & Rosen, M. K. A framework for understanding the functions of biomolecular condensates across scales. *Nat. Rev. Mol. Cell Biol.* **22**, 215–235 (2021).
9. Ditlev, J. A., Case, L. B. & Rosen, M. K. Who's In and Who's Out—Compositional Control of Biomolecular Condensates. *J. Mol. Biol.* **430**, 4666–4684 (2018).
10. Alberti, S. *et al.* A User's Guide for Phase Separation Assays with Purified Proteins. *J. Mol. Biol.* **430**, 4806–4820 (2018).
11. Alberti, S., Gladfelter, A. & Mittag, T. Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell* **176**, 419–434 (2019).
12. Lyon, C. E. *et al.* Nucleolar proteome dynamics. *Nature* **433**, 77–83 (2005).
13. Ahmad, Y., Boisvert, F.-M., Gregor, P., Cobley, A. & Lamond, A. I. NOPdb: Nucleolar Proteome Database--2008 update. *Nucleic Acids Res.* **37**, D181-4 (2009).
14. Hubstenberger, A. *et al.* P-Body Purification Reveals the Condensation of Repressed mRNA Regulons. *Mol. Cell* **68**, 144-157.e5 (2017).
15. Jain, S. *et al.* ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. *Cell* **164**, 487–498 (2016).
16. Markmiller, S. *et al.* Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell* **172**, 590-604.e13 (2018).
17. Bolognesi, B. *et al.* A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity upon Increased Protein Expression. *Cell Rep.* **16**, 222–231 (2016).
18. Vernon, R. M. *et al.* Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife* **7**, (2018).
19. Wang, J. *et al.* A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* **174**, 688-699.e16 (2018).
20. Vernon, R. M. & Forman-Kay, J. D. First-generation predictors of biological protein phase separation. *Curr. Opin. Struct. Biol.* **58**, 88–96 (2019).
21. Sun, T. *et al.* Prediction of Liquid-Liquid Phase Separation Proteins Using Machine Learning. *SSRN Electronic Journal* (2020) doi:10.2139/ssrn.3515387.
22. Hardenberg, M., Horvath, A., Ambrus, V., Fuxreiter, M. & Vendruscolo, M. Widespread occurrence of the droplet state of proteins in the human proteome. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 33254–33262 (2021).
23. van Mierlo, G. *et al.* Predicting protein condensate formation using machine learning. *Cell Rep.* **34**, 108705 (2021).

24. Chu, X. *et al.* Prediction of liquid-liquid phase separating proteins using machine learning. *BMC Bioinformatics* **23**, 72 (2022).
25. Ning, W. *et al.* DrLLPS: A data resource of liquid-liquid phase separation in eukaryotes. *Nucleic Acids Res.* **48**, D288–D295 (2020).
26. You, K. *et al.* PhaSepDB: A database of liquid-liquid phase separation related proteins. *Nucleic Acids Res.* **48**, D354–D359 (2020).
27. Mészáros, B. *et al.* PhaSePro: The database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res.* **48**, D360–D367 (2020).
28. Li, Q. *et al.* LLPSDB: a database of proteins undergoing liquid-liquid phase separation in vitro. *Nucleic Acids Res.* **48**, D320–D327 (2020).
29. Chen, Z. *et al.* Screening membraneless organelle participants with machine-learning models that integrate multimodal features. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2115369119 (2022).
30. Martin, E. W. & Mittag, T. Relationship of Sequence and Phase Separation in Protein Low-Complexity Regions. *Biochemistry* **57**, 2478–2487 (2018).
31. Martin, E. W. *et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367**, 694–699 (2020).
32. Das, S., Lin, Y.-H., Vernon, R. M., Forman-Kay, J. D. & Chan, H. S. Comparative roles of charge, *π*, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 28795–28805 (2020).
33. Mészáros, B., Erdos, G. & Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018).
34. Rostam, N. *et al.* CD-CODE: crowdsourcing condensate database and encyclopedia. *Nat. Methods* (2023) doi:10.1038/s41592-023-01831-0.
35. Yang, P. *et al.* G3BP1 Is a Tunable Switch that Triggers Phase Separation to Assemble Stress Granules. *Cell* **181**, 325-345.e28 (2020).
36. Li, T. *et al.* A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14**, 61–64 (2017).
37. van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
38. Fuxreiter, M. *et al.* Disordered proteinaceous machines. *Chem. Rev.* **114**, (2014).
39. Akdel, M. *et al.* A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* **29**, 1056–1067 (2022).
40. Roden, C. & Gladfelter, A. S. RNA contributions to the form and function of biomolecular condensates. *Nat. Rev. Mol. Cell Biol.* **22**, 183–195 (2021).
41. Yu, C. *et al.* Proteome-scale analysis of phase-separated proteins in immunofluorescence images. *Brief. Bioinform.* **22**, (2021).
42. Saar, K. L. *et al.* Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
43. Cho, N. H. *et al.* OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science* **375**, eabi6983 (2022).
44. Karlsson, M. *et al.* A single-cell type transcriptomics map of human tissues. *Sci Adv* **7**, (2021).
45. Ghosh, D. *et al.* Structure based aggregation studies reveal the presence of helix-rich intermediate during α-Synuclein aggregation. *Sci. Rep.* **5**, 9228 (2015).
46. Ray, S. *et al.* α-Synuclein aggregation nucleates through liquid-liquid phase separation. *Nat. Chem.* **12**, 705–716 (2020).
47. Hoffmann, C. *et al.* Synapsin Condensates Recruit alpha-Synuclein. *J. Mol. Biol.* **433**, 166961 (2021).

48. Surgucheva, I., McMahon, B. & Surguchov, A. gamma-synuclein has a dynamic intracellular localization. *Cell Motil. Cytoskeleton* **63**, 447–458 (2006).
49. Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. Flavors of protein disorder. *Proteins* **52**, 573–584 (2003).
50. Schad, E., Tompa, P. & Hegyi, H. The relationship between proteome size, structural disorder and organism complexity. *Genome Biol.* **12**, R120 (2011).
51. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
52. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
53. Hatos, A., Tosatto, S. C. E., Vendruscolo, M. & Fuxreiter, M. FuzDrop on AlphaFold: visualizing the sequence-dependent propensity of liquid-liquid phase separation and aggregation of proteins. *Nucleic Acids Res.* **50**, W337-44 (2022).
54. Vendruscolo, M. & Fuxreiter, M. Sequence Determinants of the Aggregation of Proteins Within Condensates Generated by Liquid-liquid Phase Separation. *J. Mol. Biol.* **434**, 167201 (2022).
55. Miskei, M., Horvath, A., Vendruscolo, M. & Fuxreiter, M. Sequence-Based Prediction of Fuzzy Protein Interactions. *J. Mol. Biol.* **432**, 2289–2303 (2020).
56. Mitrea, D. M., Mittasch, M., Gomes, B. F., Klein, I. A. & Murcko, M. A. Modulating biomolecular condensates: a novel approach to drug discovery. *Nat. Rev. Drug Discov.* 1–22 (2022).
57. Conti, B. A. & Oppikofer, M. Biomolecular condensates: new opportunities for drug discovery and RNA therapeutics. *Trends Pharmacol. Sci.* (2022) doi:10.1016/j.tips.2022.07.001.
58. Mészáros, B., Erdős, G. & Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018).
59. Wootton, J. C. & Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17**, 149–163 (1993).
60. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
61. Heinig, M. & Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **32**, W500-2 (2004).
62. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
63. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
64. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
65. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. *arXiv [cs.LG]* (2017).
66. Dorogush, A. V., Ershov, V. & Gulin, A. CatBoost: gradient boosting with categorical features support. *arXiv [cs.LG]* (2018).
67. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *aos* **29**, 1189–1232 (2001).
68. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).