Genome Biology

RESEARCH

Open Access

Comprehensive protein datasets and benchmarking for liquid–liquid phase separation studies

Carlos Pintado-Grima[1], Oriol Bárcenas[1,2], Eva Arribas-Ruiz[1], Valentín Iglesias[3], Michał Burdukiewicz[1,3,4*] and Salvador Ventura[1,5*]

*Correspondence: michalburdukiewicz@gmail.com; Salvador.Ventura@uab.cat

[1] Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, Barcelona 08193, Spain
[2] Institute of Advanced Chemistry of Catalonia (IQAC), CSIC, Barcelona, Spain
[3] Clinical Research Centre, Medical University of Białystok, Kilińskiego 1, Białystok 15-369, Poland
[4] Institute of Biotechnology, Vilnius University, Saulėtekio Al. 7, Vilniaus 10257, Lithuania
[5] Hospital Universitari Parc Taulí, Institut d'Investigació i Innovació Parc Taulí (I3PT-CERCA), Sabadell, Spain

Abstract

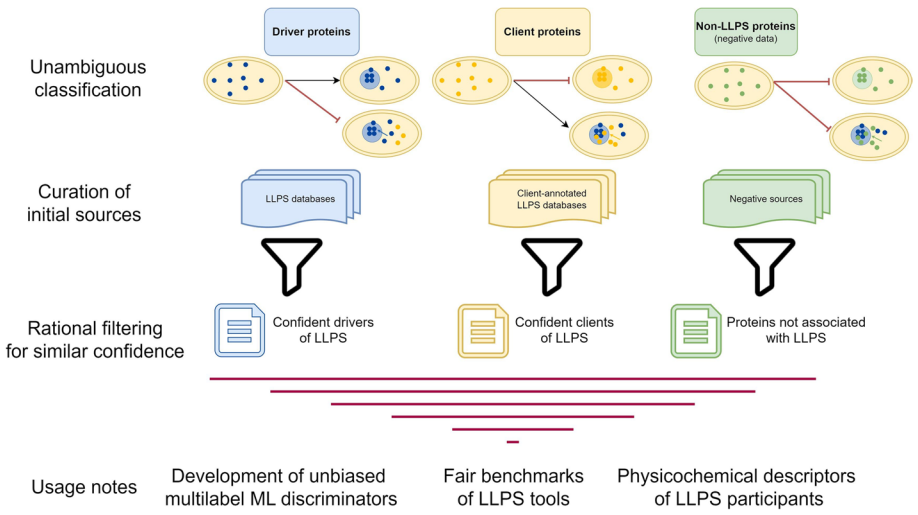
Background: Proteins self-organize in dynamic cellular environments by assembling into reversible biomolecular condensates through liquid–liquid phase separation (LLPS). These condensates can comprise single or multiple proteins, with different roles in the ensemble's structural and functional integrity. Driver proteins form condensates autonomously, while client proteins just localize within them. Although several databases exist to catalog proteins undergoing LLPS, they often contain divergent data that impedes interoperability between these resources. Additionally, there is a lack of consensus on selecting proteins without explicit experimental association with condensates under physiological conditions (non-LLPS proteins or negative proteins). These two aspects have prevented the generation of reliable predictive models and fair benchmarks.

Results: In this work, we use an integrated biocuration protocol to analyze information from all relevant LLPS databases and generate confident datasets of client and driver proteins. We introduce standardized negative datasets, encompassing both globular and disordered proteins. To validate our datasets, we investigate specific physicochemical traits related to LLPS across different subsets of protein sequences and benchmark them against 16 predictive algorithms. We observe significant differences not only between positive and negative instances but also among LLPS proteins themselves. The datasets from this study are available as a website at https://llpsdatasets.ppmclab.com and as a data repository at https://doi.org/10.5281/zenodo.15118996.

Conclusions: Our datasets offer a reliable means for confidently assessing the specific roles of proteins in LLPS and identifying key differences in physicochemical properties underlying this process. Moreover, we describe limitations in classical and state-of-the-art predictive algorithms by providing the most comprehensive benchmark to date.

Keywords: Liquid–liquid phase separation, Datasets, Integration, Driver, Client, Negative, Proteins, Disorder, Machine learning, Benchmark

Graphical Abstract



Background

The discovery of intracellular membraneless organelles (MLOs) has marked a paradigm shift in our understanding of spatiotemporal cellular organization [1]. These condensates are dynamic supramolecular structures that can concentrate different biomolecules, including proteins and nucleic acids. They act as central hubs for interactions that enable rapid and reversible compartmentalization, critical for diverse biological functions [2–5].

Although it is increasingly evident that numerous proteins can undergo liquid–liquid phase separation (LLPS), the heterogeneous composition of these condensates complicates our understanding of the precise role played by each particular protein in a given MLO. Therefore, the introduction of specific controlled vocabularies for categorizing LLPS participants has been instrumental in the progression of the field [6, 7]. *Driver* proteins can undergo LLPS on their own, without any partner—either protein, DNA, or RNA. In contrast, *client* proteins are recruited into pre-existing condensates and are not essential for their integrity. Other proteins act as regulators and can influence the behavior of drivers and clients, but they are not physically part of condensates. It is important to highlight that these roles are not mutually exclusive; a driver of a specific condensate can also be a client in another molecular scenario. Similarly, proteins can behave as clients in a given condensate but also phase-separate individually under different conditions. This duality stems from the high context-dependency of LLPS, which can be modulated by environmental conditions [8], crowding agents [9], and additional partners [10]. Indeed, for any multivalent protein, there likely exists a solute condition regime under which self-assembly into condensates will occur [11]. Therefore, an unequivocal categorization of LLPS proteins into drivers and clients requires a cautious examination of both attributes.

Given the biological relevance of LLPS in physiology, aging, and disease [12–14], several databases have been deployed to annotate proteins observed in biomolecular

condensates. However, the conceptual strategies followed to build such databases vary significantly. Consequently, the number of entries, their annotations, and the level of experimental evidence seen on each repository are highly divergent [7]. For instance, the PhaSePro database [6] collects only experimentally validated driver proteins or regions. PhaSepDB [15] contains regions with the potential to drive LLPS (*psself*) but also others that require protein or nucleic acid partners (*psother*). LLPSDB [16] annotates several protein components and solute conditions across different LLPS experiments. CD-CODE [17] is oriented toward biomolecular condensates and their constituents, making a specific distinction between driver and member proteins for each MLO. Finally, DrLLPS [18], while more protein-centric, also collects the associated condensates for each protein and the role it plays, either as a scaffold, client, or regulator. Other related databases contain condensate information but were not specifically developed to collect LLPS proteins. For example, FuzDB [19] was conceptualized to inform on fuzzy interactions between proteins and, consequently, it should not be strictly considered as an LLPS database.

Despite the efforts of curators to annotate proteins involved in LLPS, it is clear that different databases are built aiming for different objectives and collecting distinct types of data, eventually diluting important information across sources. Considering this, efforts to unify LLPS data sources are needed for a better understanding of proteins' role in condensates, as well as to train and benchmark machine learning (ML) models. MLOsMetaDB constitutes a first attempt at centralizing annotations from most LLPS databases while enriching them with external information (disorder, globular domains, function, orthologs) [20]. Still, little attempts have been made to maintain a comparable level of experimental evidence while integrating proteins from different sources, a fact that has hindered data interoperability and noiseless data annotation [21]. Besides, the evident lack of biologically relevant proteins that do not phase separate under physiological conditions (*negative proteins*) and an unambiguous distinction between clients and drivers pose significant challenges for benchmarking predictive algorithms. This situation motivated us to carefully inspect and process the data collected by LLPS databases to generate reliable datasets of client, driver, but also potential negative proteins that should not undergo phase separation (*negative datasets*), which are fundamental for building more accurate predictive tools and standardized benchmarks.

LLPS predictive tools such as FuzDrop [22] and catGRANULE [23] are designed to detect protein regions driving the formation of MLOs under standard conditions. In many instances, intrinsically disordered regions (IDRs) [24, 25] or prion-like domains (PrLDs) [26] overlap with these predicted LLPS-promoting regions. However, not all IDRs or PrLDs necessarily engage in LLPS, leading to potential biases in predictions that favor these features over actual domains with multivalent potential to establish the weak interactions necessary for LLPS [27–29]. In an effort to alleviate this issue, beyond the full-length protein, in the present study we also annotated disorder-related sequential elements, including IDRs and PrLDs**.** We illustrate how the analysis of relevant features commonly linked to LLPS can be applied to identify significant differences between datasets and mitigate sequential overlaps.

Based on current knowledge of the LLPS phenomenon and the harmonization of curation criteria, we have developed high-quality datasets of client and driver proteins

involved in LLPS. These datasets should allow a better understanding of the physicochemical properties that distinguish proteins participating in different condensates from proteins that do not. Additionally, they should help in distinguishing the specific roles played by participant proteins in LLPS reactions.

Results

Integrated dataset generation of client, driver, and potential negative proteins in LLPS

To integrate LLPS proteins into complete specific categorical datasets, we compiled data from the most recognized LLPS resources. Since different databases provide varying levels of evidence for the collected data, our first step implied the design of standardized filters aligned with LLPS vocabulary definitions to generate a curated group of proteins with consistent levels of confidence for all protein categories.

First, for databases that collect general LLPS proteins but do not specifically differentiate between clients and driver/scaffold proteins, entries were retrieved by applying filters that ensure that those proteins are actually drivers. This means that they indeed have no partner dependency—nor protein or RNA/DNA—or require further modifications such as PTM or mutations to phase separate. This distinction is crucial because even databases specifically developed to collect driver proteins with associated experimental evidence, such as PhaSePro, include partner-dependent proteins.

For databases that already consider both driver and client labels, the first stage involved distinguishing them from one another (drivers from clients) and then classifying only those proteins with at least in vitro experimental evidence, thus ensuring a higher confidence level.

Considering the high context-dependency of LLPS, a critical aspect of this kind of study involves integrating specific negative datasets of proteins not involved in LLPS. These datasets should include disordered proteins (DisProt), which are mostly overlooked in current negative datasets, in addition to globular proteins (PDB), which are often taken as the naive and only negative set (Fig. 1).

The description of confident negative datasets of proteins not involved in LLPS is challenging because of the condition-dependent nature of the process and the lack of dedicated studies on this specific protein trait. However, having well-defined negative datasets is crucial for effective training and benchmarking of unbiased predictive methods [30]. To address this need, here we implemented two independent datasets: ND (DisProt) and NP (PDB). Filters applied to the original DisProt and PDB databases involved selecting negative entries with no current evidence of association with LLPS, not present in any of the original source LLPS databases and without annotations of potential LLPS interactors, ensuring the robustness of these negative datasets. To provide a more fine-grained description of the degree of disorder, we provide the level of annotated order and disorder found in the proteins.

When specific category classifications were applied in each independent dataset we generated, the number of final entries was significantly reduced compared to the source databases due to the stringency of the applied filters (Fig. 2). In this sense, the datasets are unique not because they contain any new protein absent in other LLPS databases but because they are thoroughly created to guarantee data interoperability and data confidence. These results suggest that predictive bioinformatics tools trained with generic raw data

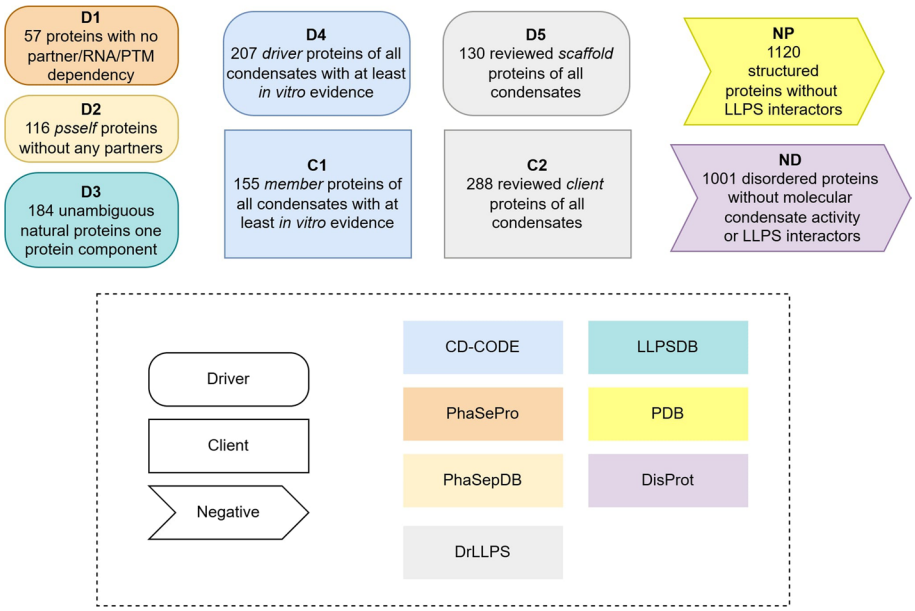


Fig. 1 Scheme of dataset generation. The nature of each dataset (driver, client, or negative) is described by the shape of the box, whereas the original source dataset can be identified by its color. A description of the filters applied to each dataset is briefly described inside each box

directly sourced from the original LLPS databases might produce nonspecific models of LLPS due to the lack of rational data filtering.

Given the multilabel condition of some LLPS participants, unambiguously distinguishing LLPS proteins as either drivers or clients is not trivial. To address this, here we attempt to provide lists of specific and confident datasets of clients and drivers by cross-checking the information from previous datasets (Fig. 3). Exclusive clients (CE) are proteins that appear only in CD-CODE or DrLLPS as clients/members and not as drivers in the rest of the positive datasets. Exclusive drivers (DE) only appear with the scaffold/driver tag and never as clients. Finally, a protein is both a client and a driver if it is tagged with both terms (C_D). The confidence of each category is also assessed by counting the number of appearances of clients and drivers in the original databases. Thus, intersecting clients (C+) are proteins found in both client databases (CD-CODE and DrLLPS), whereas intersecting drivers (D+) are those observed in at least 3 out of the 5 driver databases. All dataset records are deposited into an interactive, user-oriented website (https://llpsdatasets.ppmclab.com), intended to provide an accessible platform for users to browse and filter data intuitively.

Finally, additional annotations of disordered-related sequential elements (IDRs and PrLDs) have been precalculated. Predicting such sequences from full-length proteins could help detach existing biases in LLPS predictions and reveal how certain physicochemical features may vary between datasets. Detailed descriptions of these annotations are provided in Additional file 1: Supplementary methods.

LLPS-positive proteins and DisProt negative dataset display a similarly low proportion of ordered residues

The generation of the DisProt negative dataset (ND) was paramount as it adds a necessary subset of negative proteins beyond the naive PDB dataset (NP). Proteins in

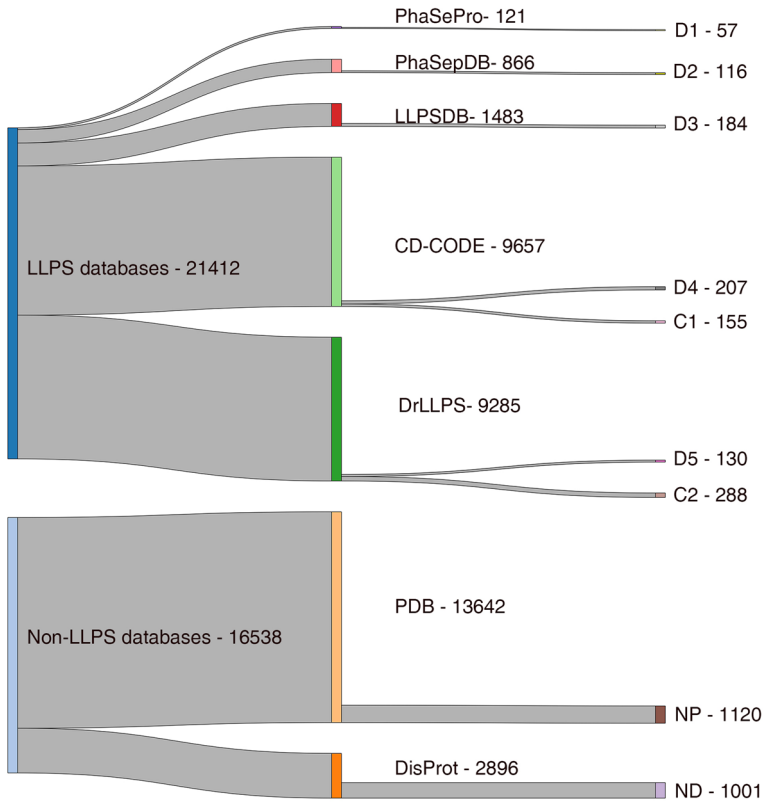


Fig. 2 Alluvial plot of original source database proteins and the final number seen in datasets after general integrative filters are applied. Depending on the original data of each database, different criteria were applied to obtain drivers (D1 from PhasePro; D2 from PhaSepDB; D3 from LLPSDB), drivers and clients (D4 and C1 from CD-CODE; D5 and C2 from DrLLPS) or negative data from non-LLPS databases (NP from PDB; ND from DisProt) with similar levels of confidence

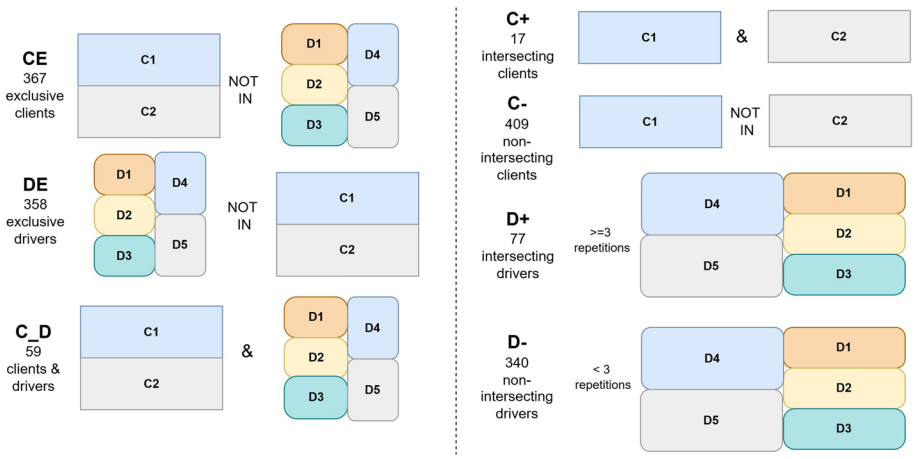


Fig. 3 Scheme of dataset crosstalk for unambiguous and confident protein annotation. On the left side, category specificity combinations are generated to obtain proteins that are exclusive clients, exclusive drivers, or both. On the right side, category confidence combinations are designed to obtain intersecting clients and intersecting drivers

ND are absent from the condensates-related specific thematic dataset of DisProt or within our positive data.

LLPS proteins often contain a considerable degree of disorder [21] that facilitates multivalent interactions [27] to the point where disorder predictors have turned out to be acceptable LLPS predictors [31]. In other words, there exists an intrinsic bias toward the prediction of IDRs rather than genuine multivalent sequences when forecasting LLPS propensities [32, 33]. This becomes evident when comparing the fraction of ordered residues in ND with that in LLPS-positive proteins, with both datasets displaying a very similar distribution profile (Fig. 4). The same trend is seen in the score distributions observed between disordered negative proteins and positive entries obtained by classical and state of the art LLPS predictive methods (Additional file 1: Fig. S1). Considering this unavoidable bias, annotating the fraction of order and disorder for every protein becomes instrumental to uncovering possible stratifications of disorder that could help to identify protein regions contributing the most to condensate formation.

We acknowledge that some proteins in ND might possess LLPS properties that have not yet been evaluated, and thus, future studies might reveal their potential to undergo LLPS under certain conditions. Still, different ML strategies are readily prepared to handle potentially noisy data, for example, by applying weakly supervised classifications that go beyond the one-instance, one-label standard [34]. In this sense, models could learn from proteins that have partial annotations and consider this feature for the final prediction.

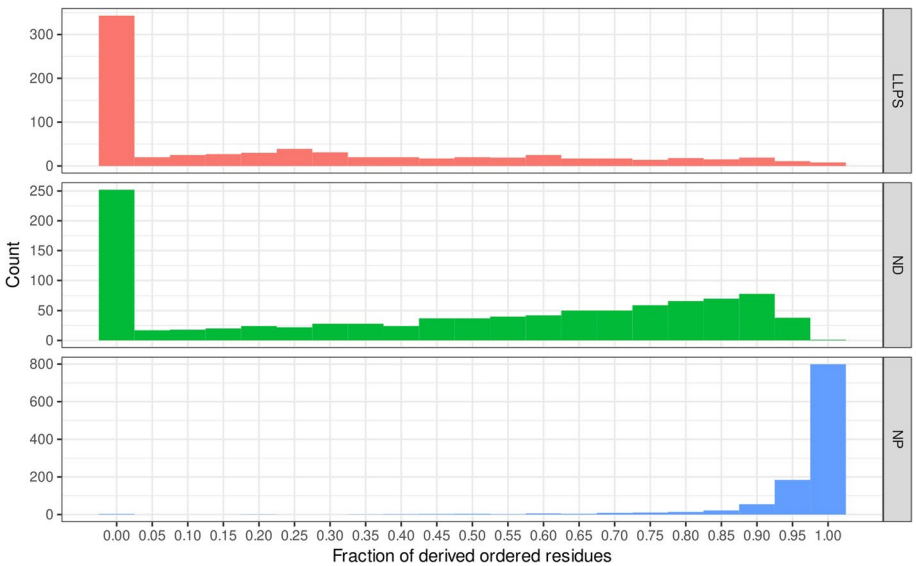


Fig. 4 Distribution of the fraction of ordered residues in positive datasets (DE, CE, C_D) in comparison with both negative DisProt (ND) and PDB (NP). A very similar distribution is observed between positive and DisProt datasets, highlighting the importance of such a negative dataset for fine-grained LLPS prediction and benchmarking. The analysis was performed on the fraction of ordered residues to maximize the number of annotated proteins. Experimentally validated order annotation is applicable to all proteins with an available PDB structure, whereas disorder annotation requires manual curation and is applicable only to proteins in DisProt

Physicochemical analysis of LLPS properties indicates differences between drivers, clients, and negative proteins

The generation of both positive and negative datasets, along with their segmentation into their disordered elements (IDRs and PrLDs), should enable a comparative analysis of how certain physicochemical properties may differ in these protein subsets. To technically validate the utility of our datasets, we evaluate four different physicochemical traits traditionally linked to LLPS sequences: charge distribution (κ), sticker/spacer distribution ($\kappa_{s|s}$), percentage of tyrosines and arginines (%Y + R) and net charge per residue (NCPR). Additionally, we include three more features with evidence in mediating interactions that can promote condensation: aggregation propensity [35, 36], cryptic amyloidogenicity [37, 38], and the presence of conditionally disordered regions capable of undergoing disorder-to-order transition upon partner binding [1, 4].

While NCPR is a key feature for LLPS [35, 39], the distribution of charged amino acids along the sequences also influences this behavior [40, 41]. The κ parameter was first introduced to compute the patterning of positively and negatively charged residues [42]. Beyond charges, the sticker and spacer model of LLPS assumes that sticky residues are responsible for establishing the first weak interactions required for condensation, whereas spacer amino acids are intercalated between stickers to regulate droplet formation and properties [43–45]. In this framework, we have introduced a variant of κ, the κ stickers-spacers ($\kappa_{s|s}$), to evaluate the distribution of sticker (YRF) and spacer (GSQN) residues (Additional file 1: Supplementary methods), thus expanding on previous approaches that just consider stickiness [46].

The %Y + R metric reveals no significant differences between datasets when considering full-length sequences (FLS) or only IDRs (Fig. 5). This indicates that the percentage of sticky amino acids alone cannot distinguish LLPS proteins (C_D, CE, and DE) from those not found in condensates (NP and ND).

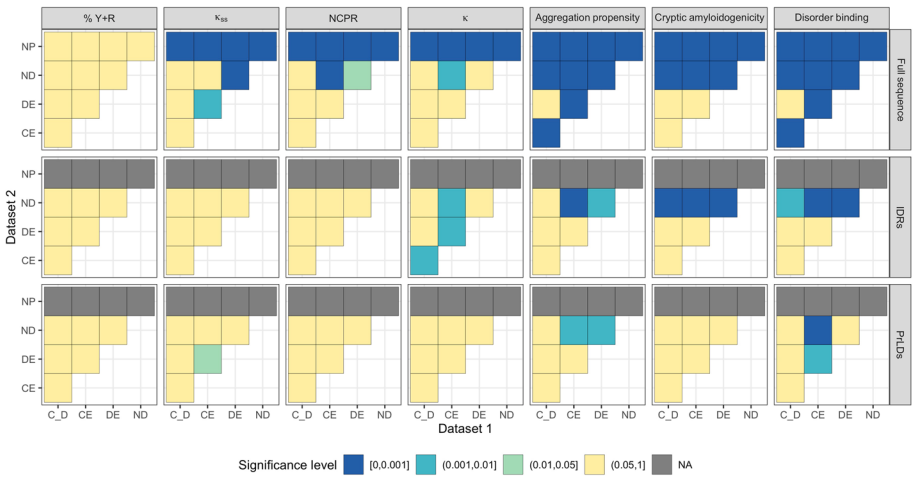


Fig. 5 Physicochemical analysis of seven properties commonly associated with LLPS in the full-length sequences, IDRs and PrLDs of protein datasets. Colors represent different levels of significance. Statistical significance was assessed by Mann–Whitney-Wilcoxon two-sided test with Benjamini correction. IDRs and PrLDs were not computed on the NP dataset (NA)

When considering FLS, NP differs from LLPS proteins across the six other properties analyzed, but it also differs significantly from ND. This finding evidences why state-of-the-art LLPS prediction methods, trained solely against NP, approximate LLPS with intrinsic disorder [27, 32]. This highlights the need for caution when using PDB entries alone as the negative dataset in benchmarking exercises. Accordingly, the NP dataset was not considered in the comparisons we outline below.

The $\kappa_{s|s}$ metric discriminates exclusive drivers (DE) from ND and exclusive clients (CE) when considering FLS, highlighting the importance of stickers and spacer residues distribution along the sequence. However, when analyzing exclusively IDRs, $\kappa_{s|s}$ loses its discriminative property. This implies that the distinction between clients and drivers is not confined, or at least not only to the disordered segments, but is contingent on the entire protein sequence.

NCPR allows discriminating ND from LLPS proteins, particularly for CE in FLS. Again, this differentiation is lost when considering IDRs alone. Similarly, κ (distribution of charged residues) has discriminatory power in FLS for CE but, in contrast with NCPR, it also does for IDRs between CE and DE, as well as between CE and C_D, underscoring the relevance of charge distribution for IDRs condensation [41] and suggesting its potential use not only to discriminate exclusive clients from exclusive drivers but also from ambiguous participants (C_D). These data also indicate that, despite NCPR and κ being obviously related, they convey different information, which can be potentially combined for better discrimination between datasets.

Aggregation shows significant differences between NDs and LLPS-positive datasets for FLS (Table 1). This aligns with the hypothesis that aggregation propensity is one of the driving forces for the reversible assembly of proteins in stress granules [14, 35] and plays a key role in the liquid-to-solid transition of condensates [47, 48]. Importantly, aggregation propensity is significantly different between CE and DE, as well as between CE and C_D, providing strong discrimination between the different roles played by LLPS proteins.

While hydrophobic aggregation-prone regions in IDRs are traditionally considered deleterious due to their likelihood to nucleate toxic aggregate formation [49, 50], cryptic amyloidogenic regions of a polar nature are widespread in both IDRs and PrLDs [38, 51]. These regions endorse disordered proteins with a self-assembly potential to establish interactions while minimizing the risk of pathogenic aggregation. This makes it the property with the highest levels of significance for discriminating IDRs in NDs from the IDRs of LLPS proteins.

The profile of disorder-binding regions (DBRs) mirrors the significance levels of aggregation propensity. This is expected, because these are likely the regions that contribute

Table 1 Significant comparisons observed for specific datasets in their full-length sequences. Physicochemical properties that can significantly discriminate ($p \leq 0.01$) are marked

	% Y + R	κ_{ss}	NCPR	κ	Aggregation propensity	Cryptic amyloidogenicity	Disorder binding
ND-CE			X	X	X	X	X
ND-DE		X			X	X	X
CE-DE		X			X		X

the most to LLPS and, in many instances, DBRs overlap with aggregation-prone regions [37, 52]. Again, differences between datasets are less pronounced or lost when considering IDRs alone.

Interestingly, when considering all the properties together, DE vs C_D is the only pairwise comparison without a significance level in any sequence subset. This implies that C_D proteins are more similar to drivers, and the properties to discriminate them (if any) go beyond the ones considered in this study.

A key observation of our analysis is that, despite PrLDs being often assumed to be a trait of LLPS proteins [2, 43, 53, 54], only aggregation propensity and disorder binding showed some discriminative power. Notably, increased aggregation propensity within PrLDs was previously linked to their recruitment into stress granules upon heat stress [35]. In the same spirit, IDRs alone are less informative than entire protein sequences. Cryptic amyloidogenicity and disorder binding seem less affected by not considering the full sequence context since they can still distinguish NDs from all LLPS-positive datasets in IDRs. Overall, disordered elements per se, when considered individually, bear poor discriminative information. These findings support the notion that multivalency extends beyond IDRs [45], as other sequential traits could be exploited in LLPS prediction to mitigate the intrinsic IDR bias [32].

The selection of these LLPS-relevant properties served as a technical validation of our datasets. Still, we acknowledge that other physicochemical properties could offer further insights about the role of proteins in LLPS. For instance, studies have correlated molecular weight or polymer length with LLPS [55, 56], showing that higher molecular weights tend to display a higher condensate saturation concentration in cells. Other research has indicated that some proteins undergo LLPS driven by hydrophobic interactions [57], while hydrophobicity seems to be positively correlated with increased predicted phase separation propensity [58]. Additionally, specific solvent-exposed hydrophobic regions may act as structural interaction motifs with multivalent potential for phase separation [1], for example, by binding to helical leucine-rich motifs [59].

Despite the statistical significance observed for some properties and dataset pairs, the building of independent feature models based on single physicochemical features could be surpassed by combinatory models that consider all the properties (Additional file 1: Fig. S2). The higher performance of these kinds of models could be explained by the integration of several characteristics that together better balance the intrinsic factors that drive these proteins to phase separate or not.

Benchmarking current LLPS predictive tools with independent protein datasets

The current landscape of LLPS predictors includes more than 20 different tools. However, specific and reliable LLPS prediction remains unsolved, with new models continuously emerging in pursuit of higher performance. Evaluating the performance of these predictors is paramount for identifying their limitations and building better models. Despite this, a comprehensive benchmark built on external independent data has yet to be conducted. This gap can be attributed to the significant heterogeneity in methodologies and the particular challenges in the usability of each tool (such as the lack of standalone apps or limitations in web server functionality).

We benchmarked those protein-level tools available before February 2025 (see methods), spanning from state-of-the-art to classical, using the independent datasets generated in this study (Fig. 6). A holistic assessment of performance metrics is crucial when benchmarking models to identify the most suitable predictors. Since not all models provide probabilistic output, relying solely on a single metric such as AUC could be misleading [60]. Accordingly, we have calculated additional metrics such as MCC, F1 score, FPR, and sensitivity for all those models that provided a clear decision threshold. Moreover, we computed AUC and PRAUC for all models that provided LLPS probabilistic outputs (Additional file 2: Table S1).

Recent ML methods such as PSPredictor [61], DeePhase [62], and PSPHunter [63] outperform other heuristic tools, suggesting that ML models capture essential contextual

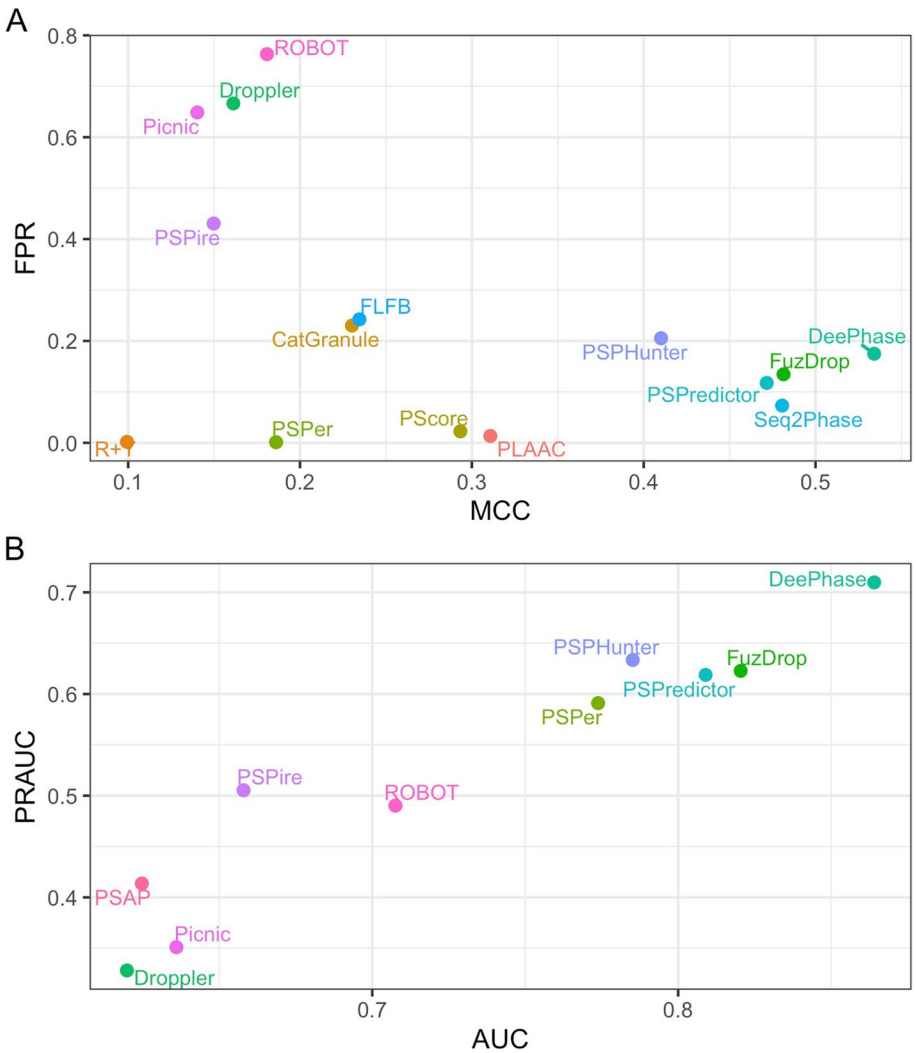


Fig. 6 Benchmark of current LLPS tools with independent protein datasets. Comprehensive evaluation of LLPS performance for all the models with information on decision threshold (**A**). AUC vs PRAUC for models that provide LLPS probabilistic output (**B**). Details about the construction of the benchmark as well as information about additional metrics can be found in methods and the Additional file 1: Supplementary methods. MCC, Matthew's correlation coefficient; FPR, false positive rate; AUC, area under the curve; PRAUC, area under the precision-recall curve

information, leading to more accurate predictions (Fig. 6A). Notably, first-generation tools, such as PLAAC [64] which were not specifically designed to predict phase separation, show a very low rate of false positives and achieve MCC close to dedicated LLPS predictors. Single property methods, such as the %R + Y alone, fail to distinguish LLPS proteins effectively, consistent with the observed lack of discriminative power of this property in the previous section.

Tools with the highest AUC recapitulate the results observed for MCC, with DeePhase, PSPredictor, and PSPHunter emerging again as the best performers (Fig. 6B). Two of the most recently published tools, PicNic [65] and catGRANULE 2.0 ROBOT [66], report AUC values around 0.7 but with a high false positive rate. These commonalities in performance metrics may be attributed to the similar negative datasets used to train the respective algorithms. Thus, despite the efforts of recent tools in generating novel datasets to train their models, the high false positive rates remain a challenge, due to limitations in the prediction of negative proteins without current LLPS association (Additional file 1: Fig. S3).

Beyond the intrinsic limitations of predictors, some authors do not disclose the data used to train their models. Therefore, it is possible that some of the proteins we used to test them were also used for training. This concept is known as information leak and could artificially inflate their performance metrics. This is one of the major factors behind building and releasing our datasets as a data repository and an accessible website. We aim to provide an open resource that scientists can reference when training and testing their algorithms, alleviating this problem in the future.

The present benchmark and datasets are a step forward for the standardization of fair comparisons between tools, suggesting there is still room for improvement in the bioinformatics prediction of LLPS.

Discussion

The datasets generated in this work allow for a confident evaluation of the role of a given protein in LLPS while integrating information from diverse LLPS sources. A total of 2876 different proteins (755 positives—either drivers, clients, or both—and 2121 negatives) are classified in the datasets, aiming to provide a realistic context for the LLPS phenomenon. This is significant given that fully annotated LLPS proteins constitute only a small fraction of the entire protein universe. In fact, a recent model reports that only 5% of the total IDRs in the human proteome are predicted to undergo homotypic phase separation [58]. Proteins not included in the positive datasets either lack sufficient evidence of undergoing LLPS, need additional partners (e.g., a protein co-driver or an RNA), undergo post-translational modifications (e.g., phosphorylation), or simply participate as regulators.

Despite the known context-dependency of the process, efforts were made to select reliable structured and disordered negative proteins. Recent ML models like PSPire [32], PSPHunter [63], PicNic [65], or catGRANULE 2.0 ROBOT [66] have generated negative datasets by selecting model proteomes, excluding positive data and other potential LLPS participants by checking for interactors or specific protein domains (Table 2). The resulting proteins are combined into a single dataset, without accounting for their relative levels of secondary structure. In contrast, here, we rationally split negative datasets

Table 2 Negative datasets built by recent ML models. These models were built by selecting potential non-LLPS proteins from the human proteome using external source databases and excluding positive entries. PSPHunter considers the exclusion of single-domain proteins. PicNic and catGRANULE 2.0 ROBOT exclude proteins with described interactions with LLPS proteins. PSPire obtains human negative proteins from PhaSePred [67]. NA, not applicable

Resource	*N* negatives (train + test)	*N* identities to our negatives	Current availability (21–03-25)	Focus on human-only proteome	External non-LLPS databases	Exclusion of positive proteins	Data stratification
PSPire [32]	10,284	66	Supplementary	Yes	No	Yes	No
PSPHunter [63]	5754	59	GitHub	Yes	PfamScan [68] (domains)	Yes	No
PicNic [65]	1709	29	Supplementary	Yes	InWeb [69] (interactions)	Yes	No
ROBOT [66]	4628	58	GitHub	Yes	BioGRID [70] (interactions)	Yes	No
Ours	2121	NA	Website and GitHub	No	PDB [71], DisProt [72] and BioGRID [70]	Yes	Yes

into structured (NP, derived from PDB) and disordered (ND, derived from DisProt) proteins, providing annotated fractions of order and disorder for all proteins, further defining their conformational attributes. Some of these tools build their negative datasets focusing only on specific organisms, which likely contributes to the low number of identities with our negative datasets. This limited intersection might partially explain the challenges in successfully predicting proteins across species. We are aware that tools trained solely on e.g. human data might perform better on data derived from the human proteome. However, as these tools are generally presented as generic LLPS predictors and not restricted to human proteins, we considered it licit to benchmark them using broader, cross-species datasets.

The level of annotation of the datasets should allow for specific protein stratifications to perform further analyses. For instance, it is possible to work with exclusive clients or exclusive drivers (category specificity; CE, DE) to uncover additional properties that may influence the client-driver distinction [73] such as molecular weight, hydrophobicity, or the presence of particular structured motifs that could enhance the development of integrated feature models. Conversely, working with proteins from ambiguous datasets (e.g. C_D) can prove useful in studying context-dependent LLPS and uncovering possible associated variables [8, 74]. Although IDRs and PrLDs alone are generally insufficient to discriminate LLPS proteins from negative data, specific properties such as cryptic amyloidogenicity or disorder binding provide hints on the features that set these sequences apart from negative disordered proteins.

Importantly, these datasets offer an opportunity to reassess the performance of current LLPS predictive methods and train more accurate models. Our benchmark showcased significant performance differences among available tools, with dedicated ML-based LLPS models outperforming non-specific and heuristic tools. This underscores the

potential of our confident protein datasets to drive the development of future ML architectures capable of recognizing critical contextual features overlooked by existing methods. Moreover, our data allows the development of both single-label and multi-label ML models. Single-label models could address problems such as distinguishing between LLPS and non-LLPS proteins, or even specific client prediction [73]. Multi-label models should allow estimating the probability of a protein acting as a driver, client, both client and driver, or none, thus identifying the most probable role of each protein. This strategy would provide a more precise and protein-centric perspective compared to other tools that combine independent models for predicting self-assembled and partner-dependent LLPS proteins [67]. Finally, the public availability of the datasets via our website facilitates a direct retrieval of proteins to train, test, and benchmark models using independent data.

Machine learning classifiers such as n-grams could be used as a first approach to identify multivalent patterns along the sequences, as they have already proven successful in predicting amyloidogenic motifs in protein sequences [75]. Although the modest size of our datasets might constrain the effective usage of deep models that require large training data [76], they could still be valuable for fine-tuning transformer-based models [77].

The incorporation of expanded and confident negative datasets, in addition to the novel client and driver distinction, should establish the basis for setting up comprehensive benchmarks of specific LLPS proteins built on independent data. Particularly, the generation of a dedicated disordered negative dataset plus the annotation of proteins' disorder fraction is expected to drive the development and refinement of specific models minimizing sequential IDR biases [32], advancing towards the implementation of a new generation of LLPS predictors [31].

Conclusions

In this work, we share holistic and rigorously scrutinized datasets to reevaluate the prediction, distinction, and benchmarking of the client, driver, and negative proteins in LLPS. We highlight a similarly low proportion of ordered residues between positive and negative data and elucidate significant differences between full-length drivers, clients, and negative proteins in specific physicochemical properties connected to LLPS behavior. Finally, we use our data to perform a critical assessment of the LLPS prediction landscape, providing the most comprehensive benchmark to date, including all 16 protein-level available algorithms until February 2025.

Methods

Filtering clients and drivers

To obtain proteins that fulfill the definition of drivers (ability to phase-separate by themselves), we thoroughly filtered the databases to exclude entries with any known partner dependency:

- D1: 57 proteins from PhaSePro v1.1.0 with no partner, RNA or PTM dependency.
- D2: 116 *psself* proteins from PhaSepDB v2.1 without LLPS partners (either proteins, RNA, or DNA) or regulations (PTMs, repeats, mutations or splicing).

- D3: 184 *unambiguous natural* proteins with one protein component without mutations, repetitions, or PTMs obtained from LLPSDB v2.0.
- D4: 207 *driver* proteins from all biomolecular condensates in CD-CODE with in vitro, in cellulo, or in vivo evidence (confidence score ≥ 3).
- D5: 130 *scaffold* proteins from DrLLPS with condensate information and tissue/cell annotations.

To collect client proteins that are recruited into preformed biomolecular condensates, we could only make use of CD-CODE and DrLLPS, since they specifically accommodate the definition of *member* and *client* proteins, respectively.

- C1: 155 *member* proteins from all biomolecular condensates with in vitro, in cellulo, or in vivo evidence were obtained from CD-CODE v1.
- C2: 288 *client* proteins from DrLLPS with condensate information, tissue/cell defined and evidence descriptions. To avoid possible high-throughput annotations, we excluded proteins reported in publications covering more than 10 entries.

We did not include regulator proteins in our datasets because they are not physically associated with condensates and are only considered by DrLLPS, precluding a consensus annotation of these types of proteins.

Obtaining unambiguous clients and unambiguous drivers

Category specificity

- CE: 367 exclusive clients are those collected in CD-CODE as *member* proteins (C1) or DrLLPS as *clients* (C2) which are not present in any of the five driver datasets (D1, D2, D3, D4, D5).
- DE: 358 exclusive drivers are those collected in any of the driver datasets and not present in C1 or C2.
- C_D: 59 clients and driver proteins appear either in C1 or C2 and also in D1, D2, D3, D4 or D5.

Category intersection

- C+: 17 intersecting clients appear in C1 and C2.
- C−: 409 non-intersecting clients appear either in C1 or C2.
- D+: 77 intersecting drivers appear in at least 3 out of 5 driver datasets.
- D−: 340 non-intersecting drivers appear less than 3 times in all driver datasets.

Generation of negative datasets

- NP: 1120 structured proteins from the PDB, with length $\geq$ 50aa and $\leq$ 5000 residues and similarity cutoff > 30% [62, 78], not present in any of the original LLPS source databases or annotated as first-degree interactors of positive LLPS proteins by

BioGRID v4.4 [70]. This was used as the classical "naive" dataset of structured proteins, which is used in many other publications of the field for benchmarking and/or training models. UniProt Accession numbers were obtained from BLASTp. Although specific contacts in globular proteins—many relying on modular interaction domains [1]—have been associated with phase separation, in general terms, they are not that prone to establishing most of the weak multivalent interactions required for LLPS. In light of this, globular domains seem to be the most obvious negative dataset and are represented in this first negative group of proteins.

- ND: 1001 proteins with annotated disorder collected from DisProt (2023_06 release) not present in the "Condensates-related proteins" thematic dataset, not associated with the GO term "molecular condensate scaffold activity," not present in any of the original LLPS source databases or annotated as first-degree interactors of positive LLPS proteins by BioGRID v4.4. DisProt entries are manually curated from the literature by expert biocurators [72].

Protein disorder/order annotation

Proteins in datasets can have different levels of disorder content. Since IDRs can overlap with LLPS regions, two metrics accounting for the fraction of disorder and order were extracted from Mobi-DB [79] for all protein datasets. The "disordered fraction" collects curated and derived annotations whereas the "ordered fraction" collects PDB-derived annotations. These metrics allow for possible further stratifications according to the fraction of disorder/order of well annotated proteins.

General protein annotation with UniProt

The UniProt database was used to collect relevant information, such as the protein cellular location (GO-CC) and the amino acid sequence. The cytoplasmic or nuclear localization of certain proteins involved in LLPS has become pivotal in unveiling the reasons behind their pathogenicity [80, 81]. Therefore, proteins with cytoplasmic (*cyto**) or nuclear (*nucl**) related GO terms were saved. Proteins without GO information, obsolete entries, or isoforms ($n = 168$) were discarded since they are, in most cases, associated with low-annotated proteins/variants. After UniProt annotation, 2876 unique proteins were integrated from all datasets into a single *.tsv* file and included in the final website (Additional file 3: Table S2).

Disordered-related sequential elements: IDRs and PrLDs

IDRs with at least 10 amino acids were obtained by considering the 'disorder consensus' sequences annotated by MobiDB [79]. PrLDs were obtained with the PLAAC algorithm [64] using a core length of 60 amino acids and relative weighting of background probabilities of 100. All sequences from disordered elements are collected in a *.json* file for each unique protein. Length distributions of IDRs and PrLDs for both positive and negative datasets can be checked in Additional file 1: Fig. S4.

Physicochemical property analysis

Each feature was calculated for all independent sequences and disorder-related sequential elements (IDRs and PrLDs) (Additional file 4: Table S3). κ and $\kappa_{s|s}$ were calculated with localCIDER [82] and an adapted version for stickers and spacers. Briefly, positive and negative charged residues calculated for κ were changed for sticker (YRF) and spacer (GSQN) residues. NCPR was calculated with the Henderson-Hasselbalch equation at pH 7.0. The %Y + R was calculated as the percentage of tyrosines and arginines. Aggregation propensity was calculated with AGGRESCAN [83], using the Na4vSS derived score. Cryptic amyloidogenicity was calculated using the Waltz algorithm at threshold 85 [52, 84], averaging the score obtained for each region with at least 7 residues. Disorder binding propensity was calculated with ANCHOR2 [85, 86], averaging the per-residue score obtained for each sequence. Heatmap's statistical significance was assessed by the Mann–Whitney-Wilcoxon two-sided test with Benjamini correction.

For the generation of feature algorithms, a model for each individual physicochemical property (κ, κs|s, %Y + R, NCPR, aggregation, cryptic amyloidogenicity and disorder binding) and data pair was trained as well as a unique combined model including all properties for full-length proteins. Support Vector Machines (SVMs) with a Gaussian kernel were used as classifiers. A fivefold cross-validation to obtain a stable estimate of the performance (AUROCs) was performed. Before the cross-validation, all entries with missing values (~ 1% of the total data) were removed. The complete code necessary to reproduce this analysis can be found in the data repository.

Benchmark analysis

Tools included in the benchmark consist of both classical and state-of-the-art LLPS predictors published until February 2025, excluding non-available, non-reproducible or region-level focused tools. Those include PLAAC [64], R + Y [87], CatGranule [23], PScore [88], PSPer [89], FuzDrop [90], Droppler [91], PSAP [92], DeePhase [62], PSPredictor [61], Seq2Phase [73], FLFB [93], PSPHunter [63], PSPire [32], Picnic [65], and ROBOT [66] (Additional file 5: Table S4). Homology reduction was first applied with the CD-HIT algorithm [94] setting a sequence identity threshold of 0.40 to effectively filter out redundant samples with similarities exceeding 40%. Independent predictors were configured to run multiple protein sequences and extract the output according to the decision thresholds established by each tool. Servers without standalone apps were run by customized R scripts harnessing the RSelenium package. Standard performance metrics were computed to assess the predictive capabilities of each tool. Extended details about the benchmark can be found in Additional file 1: Supplementary methods.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03668-6.

Additional file 1: Supplementary methods and Supplementary Figures: Figs. S1–S4.

Additional file 2: Table S1. Benchmark table with performance metrics.

Additional file 3: Table S2. Dataset statistics for the final annotated proteins.

Additional file 4: Table S3. Physicochemical property calculations for proteins in datasets.

Additional file 5: Table S4. Benchmark tools with selected thresholds.

Acknowledgements
We thank Jakub Kołodziejczyk for his valuable help in data curation.

Peer review information
Gian Gaetano Tartaglia and Andrew Cosgrove were the primary editors 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
C.P-G and O.B generated the datasets and curated the data. C.P-G, O.B, V.I and M.B performed the formal analysis, investigation and designed the figures. E.A-R and M.B designed and developed the website. O.B, E.A-R and M.B designed and conducted the benchmark. C.P-G drafted the manuscript. M.B and S.V supervised the project. S.V acquired the funding. All authors contributed to the study's conceptualization, reviewing and editing of the manuscript. All authors read and approved the final manuscript.

Funding
CP-G was supported by the Secretariat of Universities and Research of the Catalan Government and the European Social Fund (2023 FI_3 00018). OB was supported by the Spanish Ministry of Science and Innovation via a doctoral grant (FPU22/03656). VI was supported by the Polish National Agency for Academic Exchange under the ULAM NAWA Programme (Grant agreement no. BPN/ULM/2023/1/00189/U/00001). MB was supported by the Maria Zambrano grant funded by the European Union-NextGenerationEU. SV was supported by the Spanish Ministry of Science and Innovation (PID2022-137963OB-I00), ICREA, ICREA-Academia 2020 and 2021-SGR-00635 AGAUR (Generalitat de Catalunya) and CERCA Programme (Generalitat de Catalunya).

Data availability
To generate the datasets, we used the following public databases: PhasePro [95], PhaSepDB [96], LLPSDB [97], CD-CODE [98], DrLLPS [99], PDB [100], DisProt [101], MobiDB [102], BioGRID [103] and UniProt [104]. Datasets and coding scripts are available at https://github.com/PPMC-lab/llps-datasets [105] and released to Zenodo https://doi.org/10.5281/zenodo.15118996 [106] under a MIT license. Dataset website can be accessed at https://llpsdatasets.ppmclab.com.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 17 June 2024 Accepted: 25 June 2025
Published online: 08 July 2025

References

1. Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. Nat Rev Mol Cell Biol. 2017;18:285–98.
2. Hutin S, Kumita JR, Strotmann VI, Dolata A, Ling WL, Louafi N, Popov A, Milhiet PE, Blackledge M, Nanao MH, et al. Phase separation and molecular ordering of the prion-like domain of the Arabidopsis thermosensory protein EARLY FLOWERING 3. Proc Natl Acad Sci U S A. 2023;120:e2304714120.
3. Decker CJ, Parker R. P-bodies and stress granules: possible roles in the control of translation and mRNA degradation. Cold Spring Harb Perspect Biol. 2012;4:a012286.
4. Brocca S, Grandori R, Longhi S, Uversky V. Liquid-liquid phase separation by intrinsically disordered protein regions of viruses: roles in viral life cycle and control of virus-host interactions. Int J Mol Sci. 2020;21:9045.
5. Alberti S, Gladfelter A, Mittag T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. Cell. 2019;176:419–34.
6. Mészáros B, Erdős G, Szabó B, Schád É, Tantos Á, Abukhairan R, Horváth T, Murvai N, Kovács OP, Kovács M, et al. PhaSePro: the database of proteins driving liquid-liquid phase separation. Nucleic Acids Res. 2020;48:D360–7.
7. Farahi N, Lazar T, Wodak SJ, Tompa P, Pancsa R. Integration of data from liquid-liquid phase separation databases highlights concentration and dosage sensitivity of LLPS drivers. Int J Mol Sci. 2021;22:3017.
8. Pintado-Grima C, Bárcenas O, Ventura S. In-silico analysis of pH-dependent liquid-liquid phase separation in intrinsically disordered proteins. Biomolecules. 2022;12:974.
9. André AAM, Yewdall NA, Spruijt E. Crowding-induced phase separation and gelling by co-condensation of PEG in NPM1-rRNA condensates. Biophys J. 2023;122:397–407.
10. Zhou H, Song Z, Zhong S, Zuo L, Qi Z, Qu LJ, Lai L. Mechanism of DNA-induced phase separation for transcriptional repressor VRN1. Angew Chem Int Ed Engl. 2019;58:4858–62.
11. Poudyal M, Patel K, Gadhe L, Sawner AS, Kadu P, Datta D, Mukherjee S, Ray S, Navalkar A, Maiti S, et al. Intermolecular interactions underlie protein/peptide phase separation irrespective of sequence and structure at crowded milieu. Nat Commun. 2023;14:6199.
12. Alberti S, Hyman AA. Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. Nat Rev Mol Cell Biol. 2021;22:196–213.

13. Emmanouilidis L, Bartalucci E, Kan Y, Ijavi M, Pérez ME, Afanasyev P, Boehringer D, Zehnder J, Parekh SH, Bonn M, et al. A solid beta-sheet structure is formed at the surface of FUS droplets during aging. Nat Chem Biol. 2024; 20:1044–1052.
14. Batlle C, Yang P, Coughlin M, Messing J, Pesarrodona M, Szulc E, Salvatella X, Kim HJ, Taylor JP, Ventura S. hnRNPDL phase separation is regulated by alternative splicing and disease-causing mutations accelerate its aggregation. Cell Rep. 2020;30:1117-1128.e1115.
15. Hou C, Wang X, Xie H, Chen T, Zhu P, Xu X, You K, Li T. PhaSepDB in 2022: annotating phase separation-related proteins with droplet states, co-phase separation partners and other experimental information. Nucleic Acids Res. 2023;51:D460–5.
16. Li Q, Peng X, Li Y, Tang W, Zhu J, Huang J, Qi Y, Zhang Z. LLPSDB: a database of proteins undergoing liquid-liquid phase separation in vitro. Nucleic Acids Res. 2020;48:D320–7.
17. Rostam N, Ghosh S, Chow CFW, Hadarovich A, Landerer C, Ghosh R, Moon H, Hersemann L, Mitrea DM, Klein IA, et al. CD-CODE: crowdsourcing condensate database and encyclopedia. Nat Methods. 2023;20:673–6.
18. Ning W, Guo Y, Lin S, Mei B, Wu Y, Jiang P, Tan X, Zhang W, Chen G, Peng D, et al. DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. Nucleic Acids Res. 2020;48:D288–95.
19. Hatos A, Monzon AM, Tosatto SCE, Piovesan D, Fuxreiter M. FuzDB: a new phase in understanding fuzzy interactions. Nucleic Acids Res. 2022;50:D509–17.
20. Orti F, Fernández ML, Marino-Buslje C. MLOsMetaDB, a meta-database to centralize the information on liquid-liquid phase separation proteins and membraneless organelles. Protein Sci. 2024;33:e4858.
21. Orti F, Navarro AM, Rabinovich A, Wodak SJ, Marino-Buslje C. Insight into membraneless organelles and their associated proteins: drivers, clients and regulators. Comput Struct Biotechnol J. 2021;19:3964–77.
22. Hatos A, Tosatto SCE, Vendruscolo M, Fuxreiter M. FuzDrop on AlphaFold: visualizing the sequence-dependent propensity of liquid-liquid phase separation and aggregation of proteins. Nucleic Acids Res. 2022;50:W337–44.
23. Bolognesi B, Lorenzo Gotor N, Dhar R, Cirillo D, Baldrighi M, Tartaglia GG, Lehner B. A concentration-dependent liquid phase separation can cause toxicity upon increased protein expression. Cell Rep. 2016;16:222–31.
24. Kato M, Han TW, Xie S, Shi K, Du X, Wu LC, Mirzaei H, Goldsmith EJ, Longgood J, Pei J, et al. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. Cell. 2012;149:753–67.
25. Nott TJ, Petsalaki E, Farber P, Jervis D, Fussner E, Plochowietz A, Craggs TD, Bazett-Jones DP, Pawson T, Forman-Kay JD, Baldwin AJ. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. Mol Cell. 2015;57:936–47.
26. Dorone Y, Boeynaems S, Flores E, Jin B, Hateley S, Bossi F, Lazarus E, Pennington JG, Michiels E, De Decker M, et al. A prion-like protein regulator of seed germination undergoes hydration-dependent phase separation. Cell. 2021;184:4284-4298.e4227.
27. Martin EW, Holehouse AS. Intrinsically disordered protein regions and phase separation: sequence determinants of assembly or lack thereof. Emerg Top Life Sci. 2020;4:307–29.
28. Ibrahim AY, Khaodeuanepheng NP, Amarasekara DL, Correia JJ, Lewis KA, Fitzkee NC, Hough LE, Whitten ST. Intrinsically disordered regions that drive phase separation form a robustly distinct protein class. J Biol Chem. 2023;299:102801.
29. Lin Y, Currie SL, Rosen MK. Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs. J Biol Chem. 2017;292:19110–20.
30. Sidorczuk K, Gagat P, Pietluch F, Kała J, Rafacz D, Bąkała L, Słowik J, Kolenda R, Rödiger S, Fingerhut LCHW, et al. Benchmarks in antimicrobial peptide prediction are biased due to the selection of negative data. Brief Bioinform. 2022;23:bbac343.
31. Vernon RM, Forman-Kay JD. First-generation predictors of biological protein phase separation. Curr Opin Struct Biol. 2019;58:88–96.
32. Hou S, Hu J, Yu Z, Li D, Liu C, Zhang Y. Machine learning predictor PSPire screens for phase-separating proteins lacking intrinsically disordered regions. Nat Commun. 2024;15:2147.
33. Shen B, Chen Z, Yu C, Chen T, Shi M, Li T. Computational screening of phase-separating proteins. Genom Proteom Bioinform. 2021;19:13–24.
34. Hernández-González J, Inza I, Jose AL. Weak supervision and other non-standard classification problems: a taxonomy. Pattern Recogn Lett. 2016;69:49–55.
35. 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.
36. Wallace EW, Kear-Scott JL, Pilipenko EV, Schwartz MH, Laskowski PR, Rojek AE, Katanski CD, Riback JA, Dion MF, Franks AM, et al. Reversible, specific, active aggregates of endogenous proteins assemble upon heat stress. Cell. 2015;162:1286–98.
37. Santos J, Pallarès I, Iglesias V, Ventura S. Cryptic amyloidogenic regions in intrinsically disordered proteins: function and disease association. Comput Struct Biotechnol J. 2021;19:4192–206.
38. Pintado-Grima C, Santos J, Iglesias V, Manglano-Artuñedo Z, Pallarès I, Ventura S. Exploring cryptic amyloidogenic regions in prion-like proteins from plants. Front Plant Sci. 2022;13:1060410.
39. Das S, Lin YH, Vernon RM, Forman-Kay JD, Chan HS. Comparative roles of charge. Proc Natl Acad Sci U S A. 2020;117:28795–805.
40. Hazra MK, Levy Y. Charge pattern affects the structure and dynamics of polyampholyte condensates. Phys Chem Chem Phys. 2020;22:19368–75.
41. Bianchi G, Mangiagalli M, Ami D, Ahmed J, Lombardi S, Longhi S, Natalello A, Tompa P, Brocca S. Condensation of the N-terminal domain of human topoisomerase 1 is driven by electrostatic interactions and tuned by its charge distribution. Int J Biol Macromol. 2024;254:127754.
42. Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc Natl Acad Sci U S A. 2013;110:13392–7.

43. Martin EW, Holehouse AS, Peran I, Farag M, Incicco JJ, Bremer A, Grace CR, Soranno A, Pappu RV, Mittag T. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science. 2020;367:694–9.
44. Wang J, Choi JM, Holehouse AS, Lee HO, Zhang X, Jahnel M, Maharana S, Lemaitre R, Pozniakovsky A, Drechsel D, et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. Cell. 2018;174:688-699.e616.
45. Choi JM, Holehouse AS, Pappu RV. Physical principles underlying the complex biology of intracellular phase transitions. Annu Rev Biophys. 2020;49:107–33.
46. 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:e4361.
47. Vendruscolo M, Fuxreiter M. Protein condensation diseases: therapeutic opportunities. Nat Commun. 2022;13:5550.
48. Garcia-Pardo J, Ventura S. Cryo-EM structures of functional and pathological amyloid ribonucleoprotein assemblies. Trends Biochem Sci. 2024;49:119–33.
49. Langenberg T, Gallardo R, van der Kant R, Louros N, Michiels E, Duran-Romaña R, Houben B, Cassio R, Wilkinson H, Garcia T, et al. Thermodynamic and evolutionary coupling between the native and amyloid state of globular proteins. Cell Rep. 2020;31:107512.
50. Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L. A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. J Mol Biol. 2004;342:345–53.
51. Pintado-Grima C, Bárcenas O, Manglano-Artuñedo Z, Vilaça R, Macedo-Ribeiro S, Pallarès I, Santos J, Ventura S. CARs-DB: a database of cryptic amyloidogenic regions in intrinsically disordered proteins. Front Mol Biosci. 2022;9:882160.
52. Pintado-Grima C, Bárcenas O, Ventura S. Expanding the Landscape of Amyloid Sequences with CARs-DB: A Database of Polar Amyloidogenic Peptides from Disordered Proteins. Methods Mol Biol. 2024;2714:171–85.
53. Gotor NL, Armaos A, Calloni G, Torrent Burgas M, Vabulas RM, De Groot NS, Tartaglia GG. RNA-binding and prion domains: the Yin and Yang of phase separation. Nucleic Acids Res. 2020;48:9491–504.
54. Han TW, Kato M, Xie S, Wu LC, Mirzaei H, Pei J, Chen M, Xie Y, Allen J, Xiao G, McKnight SL. Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. Cell. 2012;149:768–79.
55. Dzuricky M, Rogers BA, Shahid A, Cremer PS, Chilkoti A. De novo engineering of intracellular condensates using artificial disordered proteins. Nat Chem. 2020;12:814–25.
56. Valdes-Garcia G, Gamage K, Smith C, Martirosova K, Feig M, Lapidus LJ. The effect of polymer length in liquid-liquid phase separation. Cell Rep Phys Sci. 2023;4:101415.
57. Sołtys K, Wycisk K, Ożyhar A. Liquid-liquid phase separation of the intrinsically disordered AB region of hRXRγ is driven by hydrophobic interactions. Int J Biol Macromol. 2021;183:936–49.
58. von Bülow S, Tesei G, Zaidi FK, Mittag T, Lindorff-Larsen K. Prediction of phase-separation propensities of disordered proteins from sequence. Proc Natl Acad Sci U S A. 2025;122:e2417920122.
59. Fromm SA, Kamenz J, Nöldeke ER, Neu A, Zocher G, Sprangers R. In vitro reconstitution of a cellular phase-transition process that involves the mRNA decapping machinery. Angew Chem Int Ed Engl. 2014;53:7354–9.
60. Muschelli J. ROC and AUC with a binary predictor: a potentially misleading metric. J Classif. 2020;37:696–708.
61. Chu X, Sun T, Li Q, Xu Y, Zhang Z, Lai L, Pei J. Prediction of liquid-liquid phase separating proteins using machine learning. BMC Bioinformatics. 2022;23:72.
62. Saar KL, Morgunov AS, Qi R, Arter WE, Krainer G, Lee AA, Knowles TPJ. Learning the molecular grammar of protein condensates from sequence determinants and embeddings. Proc Natl Acad Sci U S A. 2021;118:e2019053118.
63. Sun J, Qu J, Zhao C, Zhang X, Liu X, Wang J, Wei C, Wang M, Zeng P, Tang X, et al. Precise prediction of phase-separation key residues by machine learning. Nat Commun. 2024;15:2662.
64. Lancaster AK, Nutter-Upham A, Lindquist S, King OD. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. Bioinformatics. 2014;30:2501–2.
65. Hadarovich A, Singh HR, Ghosh S, Scheremetjew M, Rostam N, Hyman AA, Toth-Petroczy A. PICNIC accurately predicts condensate-forming proteins regardless of their structural disorder across organisms. Nat Commun. 2024;15:10668.
66. Monti M, Fiorentino J, Miltiadis-Vrachnos D, Bini G, Cotrufo T, Sanchez de Groot N, Armaos A, Tartaglia GG. catGRANULE 2.0: accurate predictions of liquid-liquid phase separating proteins at single amino acid resolution. Genome Biol. 2025;26:33.
67. Chen Z, Hou C, Wang L, Yu C, Chen T, Shen B, Hou Y, Li P, Li T. Screening membraneless organelle participants with machine-learning models that integrate multimodal features. Proc Natl Acad Sci U S A. 2022;119:e2115369119.
68. Mistry J, Bateman A, Finn RD. Predicting active site residue annotations in the Pfam database. BMC Bioinformatics. 2007;8:298.
69. Li T, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkowicz G, Workman CT, Rigina O, Rapacki K, Stærfeldt HH, et al. A scored human protein-protein interaction network to catalyze genomic interpretation. Nat Methods. 2017;14:61–4.
70. Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021;30:187–200.
71. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res. 2000;28:235–42.
72. Aspromonte MC, Nugnes MV, Quaglia F, Bouharoua A, Tosatto SCE, Piovesan D, Consortium D. DisProt in 2024: improving function annotation of intrinsically disordered proteins. Nucleic Acids Res. 2024;52:D434–41.
73. Miyata K, Iwasaki W. Seq2Phase: language model-based accurate prediction of client proteins in liquid-liquid phase separation. Bioinform Adv. 2024;4:vbad189.
74. Pintado C, Santos J, Iglesias V, Ventura S. SolupHred: a server to predict the pH-dependent aggregation of intrinsically disordered proteins. Bioinformatics. 2021;37:1602–3.

75. Burdukiewicz M, Sobczyk P, Rödiger S, Duda-Madej A, Mackiewicz P, Kotulska M. Amyloidogenic motifs revealed by n-gram analysis. Sci Rep. 2017;7:12961.
76. García-Jacas CR, Pinacho-Castellanos SA, García-González LA, Brizuela CA. Do deep learning models make a difference in the identification of antimicrobial peptides? Brief Bioinform. 2022;23:bbac094.
77. Chandra A, Tünnermann L, Löfstedt T, Gratz R. Transformer-based deep learning for predicting protein properties in the life sciences. Elife. 2023;12:e82819.
78. Zhou S, Zhou Y, Liu T, Zheng J, Jia C. PredLLPS_PSSM: a novel predictor for liquid-liquid protein separation identification based on evolutionary information and a deep neural network. Brief Bioinform. 2023;24:bbad299.
79. Piovesan D, Del Conte A, Clementel D, Monzon AM, Bevilacqua M, Aspromonte MC, Iserte JA, Orti FE, Marino-Buslje C, Tosatto SCE. MobiDB: 10 years of intrinsically disordered proteins. Nucleic Acids Res. 2023;51:D438–44.
80. Chou CC, Zhang Y, Umoh ME, Vaughan SW, Lorenzini I, Liu F, Sayegh M, Donlin-Asp PG, Chen YH, Duong DM, et al. TDP-43 pathology disrupts nuclear pore complexes and nucleocytoplasmic transport in ALS/FTD. Nat Neurosci. 2018;21:228–39.
81. Tyzack GE, Luisier R, Taha DM, Neeves J, Modic M, Mitchell JS, Meyer I, Greensmith L, Newcombe J, Ule J, et al. Widespread FUS mislocalization is a molecular hallmark of amyotrophic lateral sclerosis. Brain. 2019;142:2572–80.
82. Holehouse AS, Das RK, Ahad JN, Richardson MO, Pappu RV. CIDER: resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. Biophys J. 2017;112:16–21.
83. Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinformatics. 2007;8:65.
84. Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L, et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Methods. 2010;7:237–242.
85. Dosztányi Z, Mészáros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics. 2009;25:2745–6.
86. Erdős G, Pajkos M, Dosztányi Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. Nucleic Acids Res. 2021;49:W297–303.
87. Schwartz JC, Cech TR, Parker RR. Biochemical properties and biological functions of FET proteins. Annu Rev Biochem. 2015;84:355–79.
88. Vernon RM, Chong PA, Tsang B, Kim TH, Bah A, Farber P, Lin H, Forman-Kay JD. Pi-Pi contacts are an overlooked protein feature relevant to phase separation. Elife. 2018;7:e31486.
89. Orlando G, Raimondi D, Tabaro F, Codicè F, Moreau Y, Vranken WF. Computational identification of prion-like RNA-binding proteins that form liquid phase-separated condensates. Bioinformatics. 2019;35:4617–23.
90. 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. 2020;117:33254–62.
91. Raimondi D, Orlando G, Michiels E, Pakravan D, Bratek-Skicki A, Van Den Bosch L, Moreau Y, Rousseau F, Schymkowitz J. In silico prediction of in vitro protein liquid-liquid phase separation experiments outcomes with multi-head neural attention. Bioinformatics. 2021;37:3473–9.
92. van Mierlo G, Jansen JRG, Wang J, Poser I, van Heeringen SJ, Vermeulen M. Predicting protein condensate formation using machine learning. Cell Rep. 2021;34:108705.
93. Liao S, Zhang Y, Han X, Wang T, Wang X, Yan Q, Li Q, Qi Y, Zhang Z. A sequence-based model for identifying proteins undergoing liquid-liquid phase separation/forming fibril aggregates via machine learning. Protein Sci. 2024;33:e4927.
94. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
95. Mészáros B, Erdős G, Szabó B, Schád É, Tantos Á, Abukhairan R, Horváth T, Murvai N, Kovács OP, Kovács M, et al. PhaSePro. Datasets. Gene Expression Omnibus. 2020. https://phasepro.elte.hu/.
96. Hou C, Wang X, Xie H, Chen T, Zhu P, Xu X, You K, Li T. PhaSepDB. Datasets. Gene Expression Omnibus. 2022. http://db.phasep.pro/.
97. Li Q, Peng X, Li Y, Tang W, Zhu J, Huang J, Qi Y, Zhang Z. LLPSDB. Datasets. Gene Expression Omnibus. 2020. http://bio-comp.org.cn/llpsdbv2.
98. Rostam N, Ghosh S, Chow CFW, Hadarovich A, Landerer C, Ghosh R, Moon H, Hersemann L, Mitrea DM, Klein IA, et al. CD-CODE. Datasets. Gene Expression Omnibus. 2023. https://cd-code.org/.
99. Ning W, Guo Y, Lin S, Mei B, Wu Y, Jiang P, Tan X, Zhang W, Chen G, Peng D, et al. DrLLPS. Datasets. Gene Expression Omnibus. 2020. https://llps.biocuckoo.cn/.
100. PDB-contributors. Protein Data Bank (PDB). Datasets. Gene Expression Omnibus. 2024. https://www.rcsb.org/.
101. DisProt-Consortium. DisProt. Datasets. Gene Expression Omnibus. 2024. https://disprot.org/.
102. Piovesan D, Del Conte A, Clementel D, Monzon AM, Bevilacqua M, Aspromonte MC, Iserte JA, Orti FE, Marino-Buslje C, Tosatto SCE. MobiDB. Datasets. Gene Expression Omnibus. 2023. https://mobidb.org/.
103. Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, et al. BioGrid. Datasets. Gene Expression Omnibus. 2021. https://thebiogrid.org/.
104. UniProt-Consortium. UniProt. Datasets. Gene Expression Omnibus. 2024. https://www.uniprot.org/.
105. Pintado-Grima C, Bárcenas O, Arribas-Ruiz E, Iglesias V, Burdukiewicz M, Ventura S. PPMC-lab/llps-datasets: v.1.1. GitHub. 2025. https://github.com/PPMC-lab/llps-datasets.
106. Pintado-Grima C, Bárcenas O, Arribas-Ruiz E, Iglesias V, Burdukiewicz M, Ventura S. PPMC-lab/llps-datasets: v.1.1. Zenodo. 2025. https://doi.org/10.5281/zenodo.15118996.

Publisher's Note