Briefings in Bioinformatics, 22(5), 2021, 1–20

https://doi.org/10.1093/bib/bbaa408
Method Review

Computational resources for identifying and describing proteins driving liquid–liquid phase separation

Rita Pancsa, Wim Vranken and Bálint Mészáros

Corresponding author: Bálint Mészáros, Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg 69117, Germany. Tel.: +49 151 2206 9730; E-mail: balint.meszaros@embl.de

Abstract

One of the most intriguing fields emerging in current molecular biology is the study of membraneless organelles formed via liquid–liquid phase separation (LLPS). These organelles perform crucial functions in cell regulation and signalling, and recent years have also brought about the understanding of the molecular mechanism of their formation. The LLPS field is continuously developing and optimizing dedicated *in vitro* and *in vivo* methods to identify and characterize these non-stoichiometric molecular condensates and the proteins able to drive or contribute to LLPS. Building on these observations, several computational tools and resources have emerged in parallel to serve as platforms for the collection, annotation and prediction of membraneless organelle-linked proteins. In this survey, we showcase recent advancements in LLPS bioinformatics, focusing on (i) available databases and ontologies that are necessary to describe the studied phenomena and the experimental results in an unambiguous way and (ii) prediction methods to assess the potential LLPS involvement of proteins. Through hands-on application of these resources on example proteins and representative datasets, we give a practical guide to show how they can be used in conjunction to provide *in silico* information on LLPS.

Key words: liquid–liquid phase separation; LLPS; membraneless organelles; condensation

Introduction

Compartmentalization is essential for living cells to provide the spatial regulation of biochemical reactions and interactions. In addition to the classical membrane-bounded organelles, cells also contain a variety of dynamic liquid condensates called membraneless organelles (MLOs). MLOs are specialized cellular compartments that host a variety of cellular functions. Nucleoli, stress granules, P-bodies, neuronal and germ granules, postsynaptic densities, heterochromatin and many other condensates belong to this category [1]. One of the most exciting and most intensively researched recent discoveries in the field of molecular cell biology is that MLOs form through liquid–liquid phase separation (LLPS), an often reversible process generally driven by multivalent weak interactions between proteins and, optionally,

Rita Pancsa is a postdoctoral researcher in the Enzymology Institute of the Research Centre for Natural Sciences, Budapest, Hungary. Her research focuses on the computational investigation of the structure–function relationship of intrinsically disordered proteins, as well as on topics related to protein folding, amyloid formation, liquid–liquid phase separation, protein–protein interactions and the development of associated biological databases and prediction methods.
Wim Vranken is an interdisciplinary research professor in bioengineering, computer science, chemistry and biomedical sciences at the Vrije Universiteit Brussel. He leads the BIo2Byte group (http://bio2byte.be), with computational research focusing on how the dynamics, conformational states and available experimental data of proteins relate to their amino acid sequence and biological function.
Bálint Mészáros is a senior postdoctoral fellow in the Structural and Computational Biology Unit at the European Molecular Biology Laboratory, Heidelberg. His interests are the computational and experimental study of interactions mediated by intrinsically disordered proteins and short linear motifs, and the development of prediction methods and databases for the identification and dissemination of interactions in a structured way.
Submitted: 26 August 2020; **Received (in revised form):** 23 November 2020

other macromolecules [2]. Due to an avalanche of high-impact publications reporting on novel MLOs formed through LLPS in all kingdoms of life—as well as some viruses—it is now considered as a fundamental, generally employed mechanism for the effective spatiotemporal organization of cellular space [3, 4].

MLOs distinctly differ from classical organelles as they are not bounded by a phospholipid membrane but are only defined by a phase boundary that allows the dynamic exchange of constituent molecules with their surroundings [2]. MLOs also differ from well-defined stoichiometric protein complexes as they are dynamic, non-stoichiometric supramolecular assemblies. They have unique material properties [5], with their functions emerging from the collective behaviour of their constituent molecules [6]. By selectively compartmentalizing and recruiting certain macromolecules and excluding others, MLOs confer diverse benefits on cells [6–8]: they can serve as (1) activators of reactions [9, 10], (2) inactivators of reactions [11], (3) biomolecular shields [4] or filters [12], (4) sensors of changes in environmental factors [13, 14], (5) reservoirs of temporarily inactivated macromolecules [15, 16], (6) determinants of cell polarity and asymmetric cell divisions [17, 18] and (7) concentration buffering systems of their constituent macromolecules [19]. MLOs are also highly variable regarding their shapes, sizes and compositions: While certain MLOs act as specialized reservoirs of only a single protein [16] or constituents of a specific pathway [15], others orchestrate major steps of the RNA life cycle (e.g. P-bodies) or cellular stress response and may host hundreds of proteins and mRNAs [20, 21]. Also, while certain MLOs are universally present in eukaryotic cells (e.g. nucleoli, stress granules), others are specific to cell types (e.g. postsynaptic densities, germ granules) [1].

The ability to undergo LLPS may be a universal property of proteins and nucleic acids under specific conditions, i.e. almost any macromolecule can be pushed to undergo LLPS under the right environmental conditions; however, most of these conditions will never be encountered in a living cell. In other words, similar to the formation of amyloids, only particular protein sequences appear to have the ability to phase separate to form MLOs under physiologically relevant conditions [8]. Importantly, MLOs represent a cell-level phenomenon with a very heterogeneous molecular background, as the molecular forces/interaction types contributing to their formation through LLPS are very diverse. LLPS can be driven by protein–protein as well as protein–nucleic acid interactions, and the former could primarily rely on electrostatics, hydrophobic coacervation, cation-π, π–π, domain–motif, domain–domain or PTM-controlled molecular interactions, just to list the major subtypes [7, 8]. Even though the plethora of LLPS systems that have been characterized by various experimental techniques [22] have gradually uncovered these different subtypes, determining the LLPS-responsible protein sequence signatures is far from straightforward as the underlying molecular mechanisms are heterogeneous with complex partner dependencies. This could be the main reason why currently available computational prediction methods have not yet reached maturity, as they are typically only able to recognize proteins driving certain types of phase separation via one or a few of the many molecular driving forces. In addition, several experimentally described subtypes of biomolecular phase separation are as of yet without dedicated prediction methods. Therefore, the first generation of LLPS prediction methods [23] has very specific areas of utility and serves as the stepping stones for the future development of generic LLPS prediction algorithms.

The experimental studies of recent years have generated an immense amount of valuable knowledge about phase separation in cell biology. Enabling this information for computational approaches through data storage necessitates the unambiguous description of the experimental results by structuring these data based on controlled vocabularies and ontologies. In other words, the maximal exploitation of these results requires well-structured databases that store primary data as well as knowledge generated from them. Such databases can in turn provide high-quality training data for the development of sophisticated LLPS prediction methods, which can in turn provide new targets for experimental validation. In parallel with LLPS experiments, computational approaches that provide structured descriptions, databases and prediction methods have also been developed immensely. Recent years have seen the publication of several LLPS-specific databases and publicly available prediction methods accessible via dedicated public web servers. These resources are highly diverse and reflect the extremely heterogeneous nature of LLPS in terms of molecular mechanisms, cellular functions and regulation. In this review, we provide an overview of the resources available for users interested in data access for, or computational study of, LLPS. We describe the terms used to define components of LLPS and the formed MLOs, and how these descriptions are used in describing experiments. We give a comprehensive assessment of databases containing LLPS-related proteins and we survey dedicated and auxiliary prediction methods. In the closing chapter, we show the utility of these resources through selected examples showcasing the heterogeneity of LLPS. As a follow-up to the assessment of the first generation of LLPS methods [23], we give a comprehensive and hands-on guide to accessing and using the computational resources at the disposal of LLPS savvy researchers in 2021.

Towards a common language: nomenclature, controlled vocabularies and ontologies

Similar to most areas of biology, the rapid expansion of the LLPS field has brought about several new concepts and terms that are used to describe the phenomena being studied. Some of these terms have several alternatives being used interchangeably, and many of them were taken from other disciplines including cell biology, polymer physics and thermodynamics. Table 1 provides an overview of the definition of the most commonly used terms in order to reduce ambiguity in terminology throughout this paper. With LLPS nomenclature evolving organically, future standardization efforts will be needed to ensure the best representation of the knowledge being generated by the increasing number of experiments. However, as the LLPS field overlaps with other existing fields, some of their controlled vocabularies and ontologies developed can already be used to describe various aspects of phase separation.

Biocuration, i.e. the description of knowledge gained from biological experiments in standardized ways, is essential for the optimal exploitation of the results of measurements [24]. Biocuration becomes even more essential as a field matures and develops its own vocabulary in parallel with the generation of immense amounts of data—which is exactly the case for the LLPS field. Biocuration, however, relies on the existence of controlled vocabularies (CVs) and ontologies to encode the knowledge being generated. Most of the main biological ontologies developed for various applications are available via the Ontology Lookup Service [25], and several of these might be useful in describing the source organisms (e.g. Fission Yeast Phenotype Ontology or the *Caenorhabditis elegans* Ontology) and cells that were used in the experiment (e.g. Cell Ontology), defining the

Table 1. The most commonly used terms in the LLPS field

Term	Explanation
Liquid-liquid phase separation (LLPS)	A process through which a solution transitions from a single-phase state where the solute or solutes are mixed with the solvent, to a state where the solute(s) form two or more distinct phases with liquid-like properties. The process is also often referred to as liquid demixing, coacervation (simple or complex coacervation, depending on if LLPS requires a single or multiple proteins) or simply condensation.
Gelation	A complex term, with various meanings depending on the field. In its original sense introduced in polymer physics, gelation refers to the transition of a macromolecule solution to a gel phase via interactions between the polymers leading to the dramatic increase of viscosity and loss of fluidity. In the LLPS field, gelation is often used to refer to the loss of droplet dynamics, usually measured in experiments such as fluorescent bleaching. However, the term gelation lacks a unified definition and can refer to several poorly defined observations connected to condensation/LLPS.
Aggregation	Aggregation is the non-reversible interaction between proteins leading to a large, non-stoichiometric assembly. Aggregates can be formed by disordered or misfolded proteins and are commonly associated with disease emergence. Aggregation is defined simply via the end state of the assembly (i.e. a non-soluble permanent assembly of proteins) without defining the process of formation. Several phase-separated systems can transition into a more solid aggregate phase, but aggregation does not necessarily require LLPS. Similar to gelation, aggregation is a loosely defined term used in slightly different meanings depending on the field of research or even the exact experiment.
Biomolecular condensate	A non-stoichiometric assembly of biological molecules, most often proteins, RNA, DNA or a mixture of these molecules that clearly separates from the solvent. This term describes the observed state of molecules without specifying the underlying biophysical processes. While many condensates form via LLPS, they can arise via several other mechanisms as well.
Membraneless organelle (MO or MLO)	A distinct compartment in the cell that is not bounded by a membrane, typically formed via LLPS. Most MLOs are transient structures (such as stress granules), while others are permanent (such as the nucleolus). In a more generic meaning, cellular structures formed via LLPS are often called granules, condensates, foci or puncta, while those formed outside the cell in an *in vitro* environment are usually called droplets.
System	A set of molecules that together are sufficient for LLPS. This might be a single protein for single-component systems or a well-defined set of proteins and other macromolecules for multicomponent systems.
LLPS driver	A protein (or a protein region) or a set of proteins (or protein regions) that can drive LLPS on its/their own. Also referred to as scaffolds, although several scaffold proteins assemble stoichiometric macromolecular complexes without phase separation. In the driver definition, small molecules and ions are most often disregarded, and a protein is called a driver even if it requires a certain concentration of these accessory molecules. In some cases, even large molecules, such as RNA, DNA, polyphosphate or polyubiquitin are disregarded, with only linear polypeptides (i.e. 'regular' proteins) considered.
Clients and regulators	Clients are proteins and other macromolecules that can partition into MOs but do not influence their formation. Regulators are proteins that can switch the phase separation on or off or can modify its properties. Regulators often include enzymes catalysing PTMs of the scaffold proteins.
Intrinsically disordered protein/region (IDP/IDR)	A protein or a protein region that has no stable tertiary structure in isolation under a set of physiological conditions. Many IDRs, especially ones involved in LLPS, have low sequence complexity and contain repeats; however, this is not a universal feature of protein disorder. Most LLPS drivers incorporate IDPs/IDRs; however, the presence of disorder is not a strict prerequisite for LLPS.
Low complexity region/domain (LCD)	A protein region that has a highly unbalanced residue composition, which can be quantified using information theory principles. Low complexity regions are often, but not always, disordered. Low complexity regions are often referred to by naming the residue(s) in which they are enriched, such as Arg-rich or acidic regions, or by naming the repeat it contains (see 'repeats').
Repeat	Multiple tandem copies of a single amino acid are called homorepeats, while a small protein region of at least two residues that occurs in multiple tandem copies in the sequence is called a (hetero)repeat. The region that contains the repeats is usually named by the repeating element(s), such as RG-repeat or FG-repeat region. Repeat regions are often low complexity, but low sequence complexity is very often achieved without a repeating multi-residue element.
RNA-binding domain (RBD)	RBD is an umbrella term for any region in proteins that is able to interact with RNA. Many domain types could constitute the RBDs of LLPS-associated proteins, RNA-recognition motifs (RRMs) being the most frequently occurring ones, but RGG-rich regions also often play a role. While the name RRM includes the term 'motif', RRMs are folded domains, in contrast to SLiMs (see next point).
Short linear motif (SLiM)	A short, usually 3–10 residue long region in a protein, which mediates an interaction with a partner domain. SLiMs can convey various functions, being target sites for PTMs, localization signals, signals for degradation or various recognition sites. SLiMs are most often located in IDRs and their function is largely independent from the rest of the protein.
Prion-like domain (PLD or PrLD)	Prion-like domains are protein regions that typically have a highly biased sequence composition, being enriched in Q/N, as well as aromatic residues, prolines and glycines, thus sharing a resemblance to yeast prion proteins. PLDs are disordered, and due to their compositional bias, often have a low sequence complexity.

Notes: Several terms, such as 'system', 'motif' or 'regulator', have other uses outside the field of LLPS; however, here we only give a definition of the use that is common to the field, and throughout the article, we will use these terms in the sense defined here.

compounds used in a measurement (e.g. ChEBI) and defining the types of post-translational modifications (PSI-MOD Ontology) involved in the regulation of an LLPS event. A more central problem in the LLPS field is the unambiguous definition of the experiments used to assert the phase separation. Currently, there are two developed ontologies for experimental methods, each with their own strengths and caveats. The Evidence and Conclusion Ontology (ECO) [26] is a largely field-independent description of methods used in biology that can provide a framework for describing the technical side of various LLPS measurements. However, currently several techniques central to LLPS are missing from ECO, such as absorbance/turbidity measurements or 1,6-hexanediol treatment. An alternative ontology is provided by the PSI-MI ontology [27], developed by the Molecular Interactions (MI) workgroup of the HUPO Proteomics Standards Initiative (PSI), to describe experiments aimed at protein interaction detection. As LLPS is driven by protein interactions, the PSI-MI ontology is a good candidate for the future standardization of LLPS experiments. However, in addition to detecting the interaction between participants of LLPS, assessing the liquid nature of the resulting condensates is also crucial in LLPS studies, and the techniques routinely applied for this—such as photobleaching or describing droplet morphological traits—are missing from the PSI-MI ontology. Therefore, current ontologies can only serve as a starting point and not as an as-is solution for describing the experimental setup aspects of LLPS measurements.

Arguably, the most widely used structured description of proteins is provided by Gene Ontology (GO) [28]. GO is composed of three separate sub-ontologies called namespaces. The 'biological process' namespace describes the processes the protein is involved in, from the molecular to the organism level. The uncovering of this aspect of proteins is not directly tied to the study of phase separation, and the biological process terms have limited utility in encoding the results of LLPS experiments. The 'molecular function' terms offer a way of describing the mechanistic actions of the protein. This part of the ontology already includes several terms that can describe the cellular functions of various phase separation events. For example, phase separation of Ddx3 sequesters eIF4E/PABP1 to stress granules, shutting down translation [29], which can be described by 'protein sequestering activity' (GO:0140311). Upon stress-induced Zn^{2+} release, TIA1 binds Zn^{2+}, which induces the phase transition into stress granules [30], and this process is well captured by the 'zinc ion sensor activity' term (GO:0106219). Thus, the molecular function namespace of GO can be adapted to describe several cellular level LLPS functions. However, full coverage of known LLPS functions will definitely require the expansion of this namespace of GO.

The 'cellular component' namespace of GO is probably the most well suited to describe a crucial aspect of LLPS, namely the type and location of the MLOs emerging via phase separation. Several well-studied MLOs, such as P-bodies, paraspeckles or PML bodies, are already included in GO, and these terms correctly represent the entities that are described in the LLPS literature. In addition, there are several terms with broader coverage, such as 'intracellular non-membrane bounded organelle', that capture condensates for which the exact cellular location could not be defined or which were only observed *in vitro*. However, as in any field, there are cases that cannot be well described within existing frameworks. For example, GW bodies [31] (named after the Gly-Trp dipeptide containing proteins integral to them) and TIS granules [32] (named after its constituent protein TIS11B) are both well characterized in the literature, which warrants the establishment of dedicated GO terms under the existing 'cytoplasmic ribonucleoprotein granule' term (GO:0036464) for their exact description. Galectin has been shown to form non-stoichiometric lattices with liquid properties [33]. Galectin was long known to homodimerize, which can be described with the existing 'galectin complex' term (GO:1990724). However, this term explicitly requires the dimeric nature of the complex in the definition. Thus, to enable the encoding of the liquid-like lattice, a separate term could be introduced under the existing 'protein complex involved in cell–cell adhesion' term. Figure 1 shows an overview of the current LLPS-relevant GO ontology, together with additional points where the knowledge accumulated in the LLPS field could be harnessed to expand the GO cellular component sub-ontology.

While existing ontologies can describe certain aspects of LLPS, there are undoubtedly several types of information that do not fit into any existing description and will prompt the development of dedicated CVs and ontologies. This effort has already been undertaken with publications setting conceptual frameworks for standardization [5–8]. The most crucial LLPS-specific aspects that require standardized descriptions are the functional roles of phase-separated compartments in the cell, the dominant interactions contributing to the phase separation [such as electrostatic-, π–π- or short linear motif (SLiM)-mediated interactions], the molecular determinants of LLPS (such as the requirement of PTMs or the presence of a membrane) and the key observations in the experiments that are the basis of assessing the liquid property of the condensate (such as recovery after photobleaching or rapid exchange with the solvent). Based on available literature, the development of specific CVs for these four applications has already begun in parallel with database development [34] (see next chapter). It is still an open question whether these CVs will serve as the expansion for existing ontologies (which might be the case for molecular functions with respect to GO) or will be developed into their own complete ontologies.

Databases of proteins undergoing liquid-liquid phase separation

Resources of MLO-related proteins

The development of a common nomenclature for the LLPS field and the adoption of these terms into existing ontologies provided a way to more rigorously collect data on LLPS-related proteins. Databases such as UniProt that link to GO terms provide an indirect way of filtering for proteins described as being associated to MLOs, albeit without the definition of the protein's exact role in MLO formation or maintenance. Other databases, such as CRAPome [35], also provide indirect ways of pinpointing LLPS-related proteins. This notion is rooted in the observation that LLPS-related proteins tend to routinely show up in mass spectrometry experiments with high spectral counts largely irrespective of the exact experimental setup [36]. This observation was found valid for proteins that phase separate via weak but highly multivalent self-interactions dominated by π–π interactions, as about 60% of the examined LLPS drivers showed up in over 10% MS-AP measurements done with non-specific affinity purification steps performed without the specific affinity tag. However, as these databases have a fundamentally different focus, their application in LLPS research is limited.

In 2020, several dedicated LLPS-associated databases were finally published, filling a long-standing gap in this fast developing field. Since the original publications and dedicated web interfaces of these resources give a detailed explanation on their

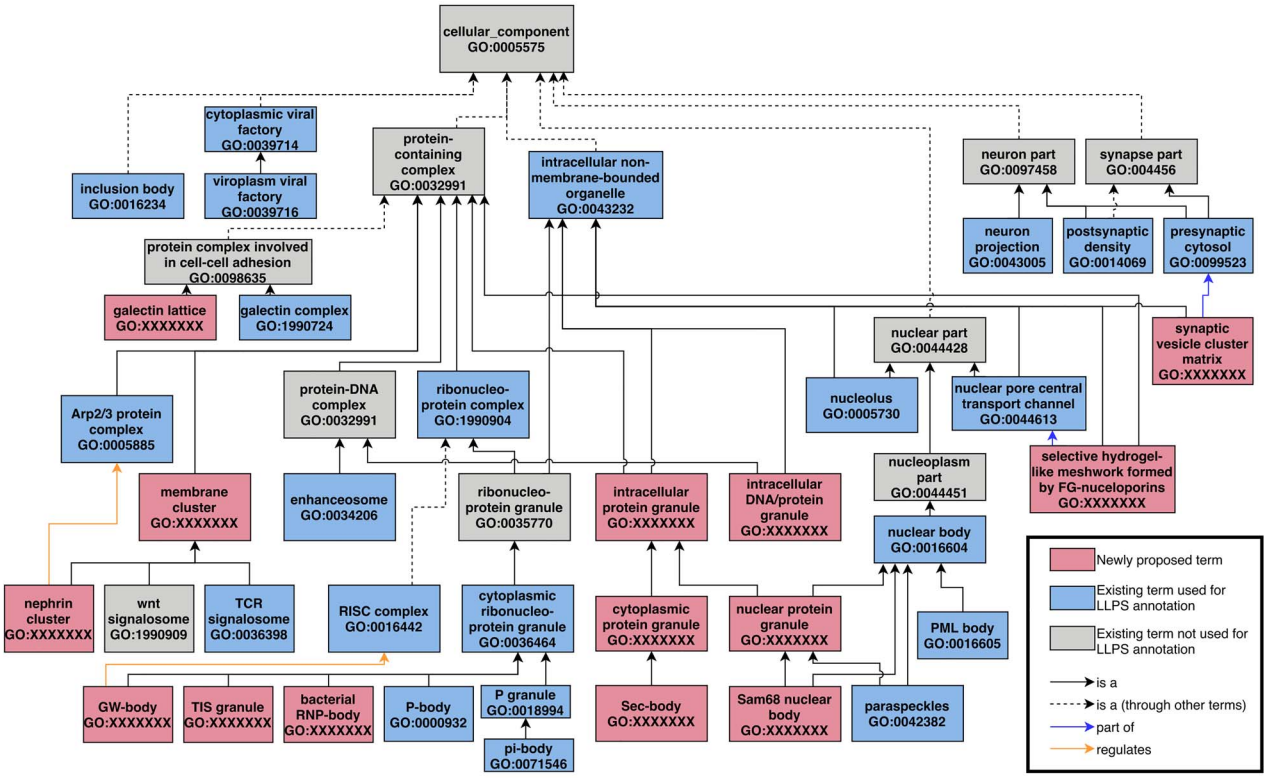

Figure 1. Existing and missing membraneless organelle types in GO. Boxes represent terms in the cellular component namespace of GO, with arrows marking the relationships between them. Blue boxes mark terms that correspond to MLOs used in the literature. Red boxes mark MLO names that are used in the literature but have no corresponding GO terms as of yet. Grey boxes mark existing terms in GO that are not suitable for MLO description but are shown to highlight the hierarchy of and relationships between MLO-specific terms. The top box corresponds to the root term of the ontology ('cellular component').

content and usage, we here aim to provide a guide for users on the strengths and limitations of these resources, explain how they are complementary to each other and explain particular use cases when they could be very beneficial. The main features of the databases discussed are summarized in Table 2. We collected features that highlight the basic principles on which each of the five databases is built, represent the amount and focus of their data content and highlight their characteristic differences.

PhaSepDB (http://db.phasep.pro/) is a resource that aims to collect all proteins that were reported to reside in MLOs and groups them according to the respective MLOs [37]. The proteins were collected based on UniProt localization annotations, literature reviews and high-throughput experiments. Importantly, in contrast to other resources, PhaSepDB does not distinguish between drivers and clients/regulators of LLPS but contains basically all MLO-resident proteins without providing a structured annotation for their role in LLPS. However, the database does provide publication and curator notes, describing LLPS cases. Therefore, this resource is an excellent choice if one wants to know if a particular protein has ever been reported to be localized in any MLO, and how that was evidenced. However, it does not provide queryable information on the mechanisms and regulation of the formation of MLOs.

The RNA granule database (http://rnagranuledb.lunenfeld.ca/) collects primary literature evidence (either cell biological, physical or genetic) from high-throughput and low-throughput approaches supporting protein (or gene) association with mammalian stress granules and P-bodies [38]. The obvious distinctive feature of this database is that it is specific for these two MLOs and is restricted to human, mouse and rat proteins. Similar to PhaSepDB, this database provides evidence on the subcellular localization of the collected proteins, but no experimental evidence on their ability to undergo LLPS, even though predicted LLPS propensities are provided. The resource is of interest for those focusing on stress granule and P-body proteomes, including if a particular protein was reported to reside in these two MLOs (under any conditions) and what evidence supports the reported association.

PhaSepDB and the RNA granule database are focused on subcellular localization, the latter relying on experimental results in the isolation and comprehensive proteomic/transcriptomic analyses of stress granules [21] and P-bodies [20], carried out recently [38]. Also, in a recent study, Yu *et al.* aimed to define the phase-separated subset of the human proteome through the systematic analysis of immunofluorescence images of 12 073 proteins in the Human Protein Atlas and describe important distinctive features of the identified phase-separated candidate proteins [41]. Owing to their localization-centric framework, these resources do not explicitly take into account the mechanisms underlying the formation of MLOs. The study of LLPS mechanisms represents a separate sub-field that has seen major advances in recent years, also fostering the discovery of some hitherto unrecognized, smaller MLOs. The results of these mechanistic studies are necessary in distinguishing drivers from clients, and databases that are explicitly built on these concepts are essential for the development of refined LLPS prediction methods. To date, three such resources have been published, each of which has a primary focus on LLPS as opposed to MLO localization, and therefore, they have a characteristically different architecture and scope.

Table 2. LLPS-related databases

	Features	PhaSepDB [37]	RNA granule DB [38]	LLPSDB [39]	DrLLPS [40]	PhaSePro [34]
Scope	General coverage of several MLO types and species	√	x	√	√	√
	Based on MLO localisation/association	√	√	x	√	x
	Based on the ability to undergo LLPS	x	x	√	√	√
	Based on *in vitro* LLPS experiments	x	x	√	√	√
	Based on *in vivo* LLPS experiments	x	x	x	√	√
	Number of reviewed entries/scaffolds	352/–	371[a]/–	1192/184	9285/150	121/121
	Includes unreviewed entries	√	x	x	√	x
Structure and depth of information on the players of LLPS	Drivers (scaffolds) distinguished	x[b]	x	x	√	√
	Regulators and clients distinguished	x	x	x	√	x
	Annotation of multi-protein driver systems	x	x	√	x	√
	Detailed information on RNA components (if any)	x	x	√	x	√
	Residue boundaries for driver regions	x	x	√[c]	x	√
	Information on the structural properties, compositions of driver regions	x	x	√	x	√
	Isoform-specific information	x	x	√	x	√
Depth of information on LLPS mechanisms, regulation, function of the resulting condensates	Molecular-level annotation of LLPS mechanism	x	x	P[d]	P	√
	Annotation of LLPS regulatory processes	x	x	P[d]	P	√
	Effects of PTMs and mutations on LLPS stored in a structured, computationally processable form	x	x	x	x	√
	Textual descriptions produced by curators on the functional, mechanistic, regulatory features of the LLPS system	P	x	x	x	√
	Functional classification of MLOs	x	x	x	x	√
Experimental details	Details on LLPS experiments	P	x	√	x	P[e]
	Evidences supporting the liquid nature of the formed condensates	P	x	x	x	√
Extra features	Novel, LLPS-specific CVs used	x	x	x	x	√
	Homology extension applied	x	x	x	√	x
	LLPS prediction results shown	√	√	x	x	x
	Driver regions reflected onto PDB structures	x	x	x	x	√
	Domain/disorder information reflected onto the protein chains	√	x	x	√	√

Notes: Check marks and x show if a feature is present or missing for a given database. P marks features where the information is partial or not provided for all the entries and may not be structured well enough for automated processing.

[a]Number of entries in the most confident tier 1 dataset of RNA granule DB is shown.

[b]PhaSepDB does not distinguish drivers in a queryable way; however, for certain proteins, '*in vitro* droplet formation' is indicated as an experimental detail derived from the associated references.

[c]Residue boundaries are provided for the protein components of *in vitro* experiments, but the components are not annotated as drivers, and information on the minimal system is still difficult to derive.

[d]Although a comprehensive overview of the *in vitro* experiments listed by LLPSDB provides some insights into the molecular mechanism and regulation of LLPS by themselves, the database does not provide additional information (or categorization) on the mechanism or regulation of LLPS besides listing the *in vitro* experiments with all their components and measurement parameters provided. Also, as the outcomes of experiments are evaluated on a binary scale (phase separation happened or not), the effects of partners, conditions, modifications or mutations that promote or decrease LLPS by influencing the number and size of droplets formed are unfortunately not possible to derive from LLPSDB.

[e]A textual description of the most important LLPS experiments that support the LLPS driver annotation of the given protein is provided, but these descriptions are not fully comprehensive and measurements cannot be reproduced from them.

Databases for LLPS-specific proteins

LLPSDB (http://bio-comp.org.cn/llpsdb/) is a highly detailed collection of *in vitro* experiments designed to study LLPS [39]. Contrary to the other resources, its entries are not proteins but *in vitro* experiments. All components of the mixtures used in the experiment and relevant measurement parameters are provided for almost 1200 experiments, together with the outcome of the experiment in a binary form (was LLPS observed or not?). LLPSDB also provides details on the protein constructs and their proteoforms, as well as the nucleic acids (if present) used in the experiments. This is especially important, as *in vitro* LLPS measurements are often carried out using non-natural systems and conditions, and thus, LLPSDB also contains designed protein chains and physiologically irrelevant conditions (such as extremely high protein concentrations). While these measurements do not directly describe cellular processes, they can provide valuable insights into the biophysical aspects of LLPS. As such, LLPSDB is the ideal resource for studying the generic polymer physics/biophysics background of LLPS, providing a wide range of data for possible underlying molecular mechanisms. Also, it is an excellent resource for those who want to gain a good overview of the *in vitro* experiments performed in proof of LLPS of a given protein. However, due to the *in vitro* focus, the entries often lack *in vivo* biological context, with no information on associated MLOs and their functional relevance or on the underlying molecular mechanisms or accompanying *in vivo* experiments. Since the outcomes of experiments are evaluated on a binary scale (phase separation happened or not), the effects of partners, conditions, modifications or mutations that promote or decrease LLPS by influencing the number and size of droplets formed cannot be derived from this database. However, there is enough detail for experts to judge if the described experiments followed a reasonable design, if they represent physiological-like conditions, and to see if a protein's phase diagram was sufficiently covered or additional experiments would be required to fully explore its LLPS behaviour.

DrLLPS (http://llps.biocuckoo.cn) is a collection of LLPS-related proteins that employs a three-way classification into scaffolds, regulators and clients, providing rich annotations and cross-references to other data resources [40]. The database is a result of an extensive automated text mining approach, containing reviewed, unreviewed and predicted entries. In addition, potential LLPS-related proteins in a wide range of organisms were identified through homology searches. Assignments of clients, regulators and scaffolds are based on high-throughput and low-throughput association, knockout and other evidence as well as LLPS experiments directly quoted from the literature. Owing to the heavy use of text mining and automation, the approach behind DrLLPS is rather to aggregate data in the form it is presented in the source literature, as opposed to distilling high-level integrated metadata by manual curation. As a result, proteins undergoing LLPS under physiologically irrelevant conditions are also part of the scaffold dataset, and the molecular mechanisms of LLPS and regulatory mechanisms of proteins annotated as regulators are not specifically addressed. Therefore, DrLLPS is a useful resource in applications where high coverage and the ability to discriminate between scaffolds, regulators and clients are important, but not the differences in the underlying molecular mechanisms of MLO formation; it also provides users with an extensive list of cross-references to other resources.

PhaSePro (https://phasepro.elte.hu/) is a manually curated resource of experimentally validated LLPS drivers [34]. Although it has the lowest number of entries among the listed databases, PhaSePro provides a highly reliable set of genuine LLPS drivers that were proven to undergo LLPS alone, or as parts of well-defined LLPS systems with a few co-drivers, in *in vitro* or *in vivo* experimental studies. The strict annotation protocol ensures that only LLPS cases supported by sufficient amount of physiologically relevant experimental evidence are included. Another feature of PhaSePro is that it aims to annotate the minimal set of components required for LLPS; therefore, information on protein region boundaries is provided whenever available. Compared to the other databases, PhaSePro contains a more limited set of experiments; however, it represents these in a structured way using ECO. PhaSePro also links to other ontologies, such as GO, to define MLOs. In addition, information outside the scope of existing CVs and ontologies—including the classification of MLOs according to cellular function and the classification of LLPS systems according to the main biophysical interactions/molecular driving forces involved—are represented with customized LLPS-specific CVs. While the most limited in size and coverage, the highly structured data representation and the careful manual curation make PhaSePro a promising candidate to provide training sets for future high-resolution LLPS prediction methods. Also, PhaSePro is ideal for users who wish to fully explore any included LLPS system, including its structural and functional features, the underlying molecular driving forces, partners or environmental determinants influencing the process, regulation by PTMs or alternative splicing and associations to disease.

It is important to note here that *in vitro* LLPS under physiologically irrelevant conditions (such as those applied for lysozyme [42] or gamma crystalline [43], etc.) or *in vivo* condensate formation under a strong promoter [44] says nothing about physiologically relevant LLPS. Many proteins phase separate at high concentration and low temperature, but they should not be classified as LLPS drivers based on such observations. LLPSDB and DrLLPS include proteins without considering the physiological relevance of the applied experimental conditions, and thus, the information they provide is not uniformly useful for those who are only interested in physiologically relevant LLPS. While LLPSDB at least refrains from assigning driver/scaffold roles to the included proteins and provides the associated measurement conditions, DrLLPS does categorize several proteins as scaffolds, which were measured under non-physiological conditions *in vitro* (e.g. lysozyme [42]), were only observed to form cellular puncta *in vivo* when heavily overexpressed (e.g. Mip6p [44]), were only demonstrated to partition into the condensates formed by other proteins (e.g. CIRBP and CPEB2 [45]) or were only used as donors of smaller protein modules that were used in tandem repeats with (artificially) high multivalency to study the importance of multivalency in LLPS systems (e.g. ABL1 and PIAS2 [46]). Since there clearly is not sufficient evidence that these proteins could function as LLPS scaffolds in cells, categorizing them as such is rather misleading. The users of the LLPS resources should be aware of these caveats.

Overlap of LLPS-specific databases

Owing to the different underlying concepts of the various databases, the sheer volume of data offered by them also varies widely (see Table 2). From a user viewpoint, it is important to know how the data contained in these resources relate to one another, so users can make informed decisions on choosing the right dataset (or a combination of datasets) for their application. For this end, Figure 2 shows the overlap between the reviewed

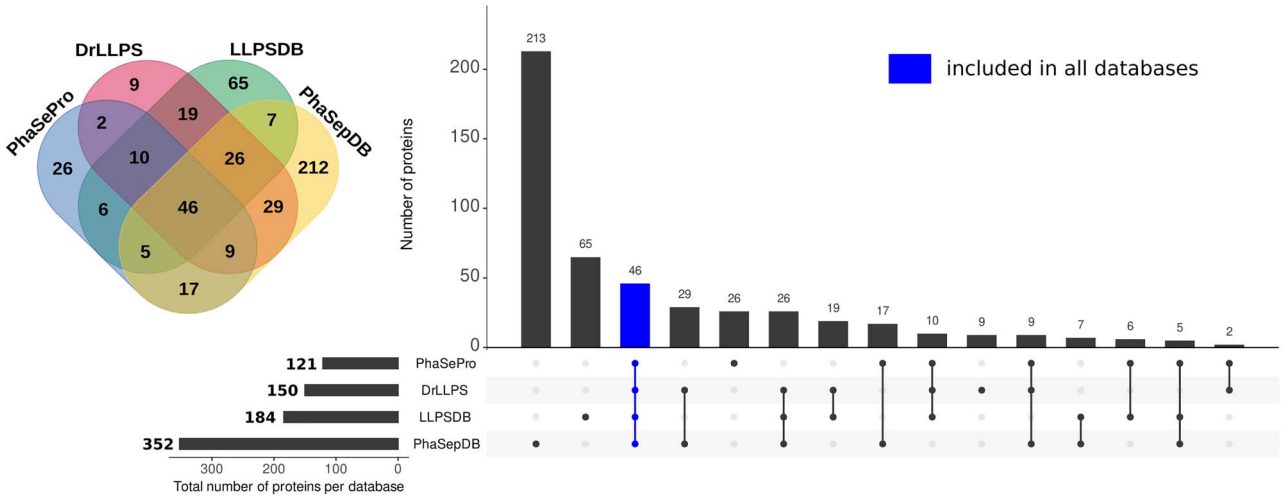

Figure 2. Overlap between various LLPS-specific databases. From databases containing both driver and client/regulator proteins, only the drivers were used. For versions and other database details, see Data and Methods. UniProt accessions for all proteins from each database are shown in Supplementary Table S1, available online at https://academic.oup.com/bib.

entries of PhaSepDB, LLPSDB and scaffolds/drivers from DrLLPS and PhaSePro. The RNA Granule Database is omitted due to its restricted scope of only two MLO types. We only included naturally occurring proteins that can be identified with UniProt accessions (see Data and Methods and Supplementary Table S2 available online at https://academic.oup.com/bib for UniProt accessions), which limits the information represented from LLPSDB.

Only 46 proteins are included in all four databases, reflecting the different inclusion criteria of the various databases. These 46 proteins can be considered to be the core LLPS dataset, and the users interested in the study of any of these have access to a wide range of information. In contrast, there are over 200 proteins that are only included in PhaSepDB and an additional 65 that are only included in LLPSDB. This highlights the wider scope of these databases, aiming at high coverage with more liberal inclusion criteria. Interestingly, the vast majority of data in DrLLPS is also included in other databases, showing the overall reliability of DrLLPS entries and showing that carefully tuned automated annotation methods can be efficiently utilized, ideally as input for later manual curation. As a counterpoint, while PhaSePro contains the smallest number of proteins in total, over 20% of its data is missing from all other databases, showing the advantages of labour-intensive manual curation efforts in capturing cases that are missed by other approaches. In addition, PhaSePro was the last resource to be released and updated (as of writing this paper), which gave it the possibility to include several newly described cases.

Apart from different scope, the amount of overlap between various databases might have technical reasons as well, such as the inclusion of different isoforms or homologues in various resources. In order to assess this, the data shown in Figure 2 exclude isoforms (discarding this information from LLPSDB and PhaSePro), mapping each protein to its canonical UniProt accession. Supplementary Figure S1 available online at https://academic.oup.com/bib shows the overlap between the four databases after mapping all constituent proteins to their UniRef50 clusters (see Supplementary Table S2 available online at https://academic.oup.com/bib for UniRef50 accessions), to decrease sequence redundancy. While the actual numbers change slightly, the overall patterns of overlap are very similar, indicating that the limited overlap between current LLPS databases does in fact mirror the different scopes of the data included in them.

Prediction methods for identifying unknown LLPS drivers

Dedicated prediction methods

Despite the advances in the field, few bioinformatics predictors of phase-separation proteins exist [23], likely due to several reasons. First, there is a variety of mechanisms by which phase separation might occur, for example via interacting with RNA or with oppositely charged proteins. In several known molecular systems, LLPS is driven by a combination of mechanisms, and our knowledge of mechanisms contributing to the process is often incomplete. This translates into the second hindrance, which is more tangible from a method development viewpoint: the presence and sequential order of the compositional elements mediating the LLPS-driving interactions—such as low-complexity regions or RNA binding domains—are highly variable and exhibit irregular distributions in various proteins. This complicates sequence-based prediction with traditional bioinformatics approaches. Third, the ability to undergo LLPS is not an intrinsic property of proteins, such as being disordered, but it is context dependent, usually being a highly complex function of environmental parameters, with stimuli from temperature, salt type and concentration, pH, macromolecular interactions and post-translational modifications (PTMs) all being important [8]. A simple binary classification of the protein as phase separating is therefore too limiting, and it becomes essential to include environmental parameters in the prediction. With the development of context-dependent LLPS prediction methods, it would be possible to calculate a score that implies how readily the protein phase separates within certain parameters, or with a reversed logic, to predict values of environmental parameters under which the protein undergoes phase separation. In addition, methods can differ based on the role of

the protein being identified, as some methods aim to recognize drivers of phase separation, while others aim to recognize all proteins that are localized to MLOs. Therefore, when using phase separation prediction methods, it is crucial to take note of the type of proteins being identified for correct interpretation of the results, and the 'molecular grammar' that the method addresses. For example, FUS-like proteins contain prion-like domains (PLDs), RNA-recognition motifs (RRMs) and disordered, typically arginine (Arg)-rich regions [47], interspersed with regions that have little expected impact. Another example are the DDX4-type proteins, which contain disordered regions containing FG and RG groups arranged in a distinct pattern, governing overall charge patterning in interchanging blocks of positive and negative charge [48].

Given the complexity of the phase separation process, most existing approaches identify individual protein characteristics that were identified as part of the molecular grammar elements identified in phase separation proteins. This can take the form of databases, such as LARK [49], where low-complexity aromatic-rich kinked regions are identified and provided. The users can approach using these databases as a pseudo-prediction method by checking if their specific protein or protein region is included therein. In the case of LARKS, the authors are implementing an automated prediction server to aid this [50]; however, at the time of publishing this review, the server is not yet functional. More accessible approaches enable direct prediction of phase separation-related characteristics directly from protein sequence. PLAAC [51] predicts polar-rich PLDs using a hidden Markov model (HMM). It was developed much prior to the realization of phase separation being occasionally driven by PLDs; therefore, while PLAAC might find PLDs implicated in LLPS, we should definitely not expect it to identify the majority of phase separation driving regions. PScore predicts π–π interactions [36] based on statistically expected long-range interactions between amino acid side chains. ZipperDB collects protein regions that have been identified to have a tendency to form fibrils using structural profiling [52] and also enables the analysis of user input proteins. While these methods aim to capture specific protein regions that are known to be able to drive phase separation, catGRANULE [44] is a more generic method, predicting the propensity of a protein to be localized in 'granules' by combining self-properties of amino acids, such as their tendency towards disorder. A review of such first-generation predictors showed limited overlap between their results, with RNA-binding proteins typically well predicted, but with low performances for ones that require protein–protein interactions or PTMs [23].

These methods in essence use statistical scoring based on the primary sequence composition, which is very useful to identify and classify phase separation proteins. In accord, their output is a score reflecting the predicted tendency of the input protein to drive (or be associated with) phase separation, often together with a definition of the protein region responsible for this behaviour. Further refinement becomes possible by integrating all known molecular grammar elements for a particular class of phase separation proteins, enabling the recognition of multiple driver regions contributing to LLPS. The approach taken by the PSPer method [53] is to use a variety of bioinformatics approaches to first separately identify the elements important for phase separation, in this case for the FUS-like proteins mentioned earlier. These characteristics are predicted from the sequence and are subsequently encoded in a HMM-like model, which provides an overall score for the protein that depends on whether regions with these characteristics are present. The advantage of such an approach is that the relative positions of these regions can be encoded without enforcing a specific organization within the protein. The method achieved a −0.87 Spearman correlation between its HMM score and the experimentally determined saturation concentration of FUS-like proteins. PSPer was trained using a negative dataset containing ordered proteins and hence is expected to have increased performance on LLPS driven by disordered proteins.

A parallel direction of method development is the use of supervised machine learning approaches based on a 'learning set' of proteins involved in phase separation, as implemented by PSPredictor [54]. However, the complexity of the molecular mechanisms involved in LLPS and the different roles that can be adopted by its molecular components (driver, client, regulator) result in a wide variety of relevant proteins and protein features for which we still have little data, and particular care has to be taken to avoid overfitting the machine learning to this currently very limited but highly complex learning set. For all methods, but especially for machine learning algorithms, negative controls will be essential in this respect, especially given that up to 20% of the human proteome could be involved in phase separation [23]. However, identifying proteins that never undergo LLPS under physiological conditions is a huge challenge given the diversity of protein concentration, pH and other environmental parameters found in various cell compartments. The lack of a unified, gold standard negative LLPS dataset produces implicit and hard-to-quantify biases between various methods, as their efficiency will depend on the features of the proteins in the used dataset.

Approaches that take the route of molecular modelling have also been developed. These range from coarse-grained residue-based models of disordered protein condensates [55] that can handle contextual changes such as phosphorylation events, to more coarse approaches, including lattice models [56]. These approaches are showing their potential in elucidating possible mechanisms of phase separation but remain low resolution and do not allow detailed investigation of atomic interactions. Full-atom simulations are also being explored but are currently limited by computational costs and force field issues. Finally, polymer physics-based methods excel at capturing global behaviour of disordered regions and describing the electrostatic interaction-driven effects. However, LLPS is often driven by folded domains, coil-coiled regions and domain–motif interactions, which are out of the scope of these methods. Overall, these approaches have limited applicability for large-scale prediction purposes due to their computational costs, but they are already making significant contributions to understanding specific LLPS mechanisms and will be instrumental in informing the next generation of sequence-based predictors.

In summary, to identify proteins with a particular phase separation mechanism, prediction models integrating its full molecular grammar are likely to be the most useful. On the other hand, to explore proteins for which the mechanism of phase separation is unclear, the most relevant way to proceed is likely a bioinformatics analysis of characteristics related to phase separation, such as π–π interactions, disordered and aggregation-prone regions, which can then be analysed to understand which mechanism might be responsible for driving phase separation [8].

Methods for detecting LLPS-related protein sequence features

LLPS often involves sequence regions that possess some unique quality or certain types of functional modules, apart from the most apparent PLDs and RRMs. The presence of such regions might be indicative of the protein's involvement in LLPS by

highlighting molecular grammar elements connected to LLPS. There are several methods that have been developed for the identification of such regions, and while they were not explicitly developed to aid the identification of LLPS driver proteins, their use might provide valuable information on proteins under study, either reinforcing the prediction of dedicated LLPS prediction methods or providing complementary information.

PLDs have been frequently described to be central drivers of LLPS [57], and methods such as PrionW [58] and PrionScan [59] can highlight such driver regions. In a more generalized approach, one of the characteristic features of PLDs is their low sequence complexity. The lack of sequence complexity can be assessed using various methods such as SEG at the protein level [60] and TRF at the nucleotide sequence level [61]. Apart from the information content and sequence features of protein regions, biochemical and structural features can also be indicative of LLPS drivers. Electrostatic interactions involving charged and aromatic residues inside disordered proteins are hallmark features of several LLPS drivers, such as FUS and TAF15 [47]. Polymer physics-based methods, such as CIDER [62] quantifying the charge patterning in protein regions, can pinpoint regions with molecular behaviour governing physical properties such as radius of gyration, consistent with phase separating disordered regions. Disorder prediction methods and databases, such as IUPred [63] and MobiDB [64], can reach high accuracies, especially when used in combination [65], and thus can provide additional information in identifying LLPS drivers.

While the protein regions targeted by these methods are clearly significant contributors to LLPS in several cases, caution should be exercised in their use in LLPS driver identification. PLDs, low complexity regions and polyampholytes represent much wider functional classes of proteins than being exclusive to LLPS. Not all such regions are connected to LLPS, and conversely, not all LLPS drivers display these features. This notion is especially true for protein disorder, as the lack of a stable structure is a feature of roughly one third of the human proteome [66]. Hence, using these methods can only indicate the protein's potential involvement in LLPS, but additional information is needed to properly assess it.

Detecting conserved functional modules

Given the biological importance of non-membrane-bounded organelles, the function of LLPS drivers is expected to be well conserved. In many cases, LLPS is driven by low complexity and/or disordered regions, where the conserved function does not necessarily require sequence conservation. Therefore, sequences of LLPS drivers that are highly disordered, containing repetitive, highly charged or prion-like regions, are usually difficult to align, hindering conservation score calculations. However, ordered domains—such as RNA-binding domains—also often contribute to LLPS formation, and the presence of these domains can be captured by protein domain prediction methods relying on sequence alignments. Conservation can also be detected for regions of 'constrained disorder', where the flexible character of the protein region is preserved via strong sequence conservation [67]. Constrained, flexible and non-conserved disorders are characteristic of different functional classes of IDPs; however, establishing a connection between these classes and LLPS are yet to be studied.

One of the most widely used methods for identifying conserved protein modules from the protein sequence is Pfam [68]. Pfam generates HMMs built on the alignment of known protein sequences, and the recognized conserved modules are annotated and stored in the database. These HMM profiles also serve as a prediction tool, being able to recognize the annotated modules in any input sequence. While Pfam regions are mostly referred to as domains, they are not necessarily structured domains, as their definition relies on sequence conservation alone. Therefore, certain IDRs and even SLiMs can be identified if their sequence is well conserved throughout evolution.

While certain domain types have been connected to LLPS drivers in individual cases, to date no systematic survey has assessed the overlap between Pfam regions and LLPS driving regions, or the utility of Pfam predictions in identifying LLPS drivers. Figure 3A shows how much on average Pfam regions cover the known protein regions responsible for driving LLPS. While for some proteins Pfam captures some, or even all of the driver regions, for over 40% of known drivers, the driver region does not overlap with any Pfam region. The correlation between the length of the LLPS driver region and the coverage by Pfam is negligible with $r = -0.06$, showing that the poor overall coverage is not a size effect, and longer LLPS driver regions escape identification by Pfam just as easily as short ones. This sets a—fairly low—upper boundary for the utility of Pfam in LLPS driver prediction. Figure 3B shows the type of Pfam regions found to overlap with LLPS driver regions (for a detailed list, see Supplementary Tables S3 and S4 available online at https://academic.oup.com/bib). The most common Pfam region type covers various nucleic acid binding domains, with most of them corresponding to conserved RRMs. The second largest class is given by protein-binding domains and regions, with a large portion of these being SH2 and SH3 domains and modules mediating interactions with histone tails (such as chromo domains). A smaller portion of driver regions contains zinc finger modules, and regions tethered to membranes, reflecting the roles of LLPS in transmembrane signalling. However, over one-third of detected Pfam regions cannot be easily described by a single well-defined function.

In light of these results, the use of the identified Pfam regions as indicators of LLPS driver proteins should be used with extreme caution. On the one hand, only a bit more than half of true drivers overlap with any Pfam region. On the other hand, the analysed LLPS drivers overlap with 83 different types of Pfam regions, and therefore, there is no good answer to questions such as 'which Pfam regions are responsible for driving LLPS?' While Pfam can provide valuable information in the search for LLPS drivers, it should only be used in conjunction with other indicators to achieve meaningful coverage and precision.

The potential of the combination of methods

The available methods for the identification of LLPS driver proteins typically employ different underlying concepts, different architectures and different training and testing sets. This can make it difficult to use them in common settings, such as trying to determine if a protein of interest is a true LLPS driver. To date, no systematic comparative studies have been published, and thus, there are a range of very justified questions on the user side that are difficult—if not impossible—to answer. Which method is the best choice for my protein of interest? How likely are various methods to recognize a true LLPS driver, and how likely are they to give false-positive predictions? What should I expect when using these methods in large-scale proteome-wide studies? Does the efficiency of methods depend on the sequence-level, structural and functional characteristics of the protein studied? Is there a benefit of using these methods in combination, and if so, what can we expect from this approach?

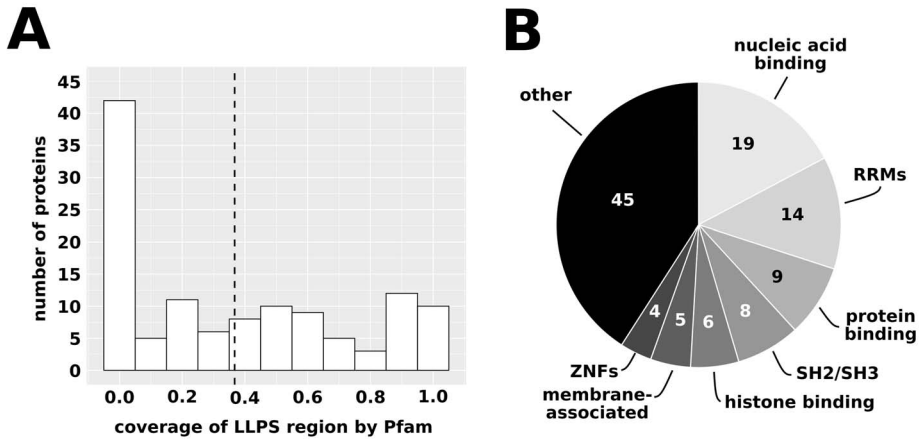

Figure 3. Pfam regions overlapping with protein regions known to drive LLPS. (**A**) The coverage of LLPS driver protein regions by Pfam. The dashed line marks the coverage averaged over all LLPS driver regions taken from PhaSePro. Found Pfam regions are shown in Supplementary Table S3, available online at https://academic.oup.com/bib. (**B**) Types of Pfam regions overlapping with LLPS driving regions. Groups were established assessing GO [28] terms attached to Pfam regions, taken from InterPro mappings [69]. Nucleic acid binding excludes RRMs, and protein binding excludes histone tail binding and SH2/SH3 domains. For the classification of Pfam objects into functional categories, see Supplementary Table S4 available online at https://academic.oup.com/bib.

In order to rigorously answer these questions, the best way would be to test all methods on standardized datasets and calculate standard prediction method evaluation metrics, such as Matthews correlation coefficients or area under the receiver operating characteristic curves. Such standardized positive datasets containing verified LLPS drivers have just recently appeared. Unfortunately, negative datasets that would contain proteins verified to not drive or engage in LLPS formation under any relevant physiological condition are absent. Considering the challenge of assembling such a dataset that represents the sequence heterogeneity of the known protein universe, the construction of such datasets will take considerable time and effort.

In lieu of rigorous testing, we outline the properties of five LLPS prediction methods by comparing their results on a small high-quality LLPS driver dataset taken from PhaSePro. We chose this as a testing set as PhaSePro is limited to cataloging LLPS drivers as opposed to client proteins, as well as defining the residue boundaries of minimally required LLPS driver regions. It also contains additional annotations of LLPS drivers, which enables us to assess the utility of each method as a function of source organism, molecular driving forces of LLPS and other features. We further removed all proteins that do not have *in vivo* evidence for driving LLPS to assess the biologically truly relevant drivers (see Supplementary Table S5 available online at https://academic.oup.com/bib). In total, this left 109 proteins in the testing set: roughly 40% of these proteins are included in all four LLPS databases (see Figure 2 and Supplementary Table S1 available online at https://academic.oup.com/bib), and only 22% is unique to PhaSePro. However, the majority of these PhaSePro-unique proteins have been published after the last curation round of the other databases, and hence the lack of overlap does not mean lack of reliability. It is important to note here that PhaSePro is definitely not an independent testing set for the five methods compared below; such an independent set is currently not available. Each method did have proteins in their respective training sets that are present in PhaSePro or are close homologs of those (in varying numbers), so the comparison provided below is not aiming to provide a fair benchmarking or critical performance assessment of the methods and therefore should not be regarded as such.

In terms of methods, we chose the ones outlined above that (1) are built to specifically recognize LLPS proteins or have a very close focus (in the case of PLAAC) and (2) are accessible via public web servers for all users. In accord, we ran PScore [36], PSPer [53], PLAAC [70], catGRANULE [44] and PSPredictor [54] on the positive dataset. All methods were run with their default settings, except for PSPer, where the protein-level cutoff was lowered from 0.56 to 0.38 to improve coverage. In addition, catGRANULE protein-level scores were evaluated with a cutoff value of 0.75. While catGRANULE does not define a default cutoff, this value corresponds to the third quartile in the distribution of scores calculated in the positive training set [44]. PLAAC does not assign a protein-level score; instead, it defines the region responsible for driving LLPS. Thus, cases with at least one predicted region of any length were considered as positive predictions, while cases with no predicted regions were considered as negative.

In addition to studying one protein of interest, another typical use for such methods is in large-scale studies, often conducted at the proteome level. To assess the utility of these five methods, we also ran them on the 20 350 proteins of the full human proteome (see Supplementary Table S6 available online at https://academic.oup.com/bib) taken from UniProt. This can indicate potential overprediction problems. The best methods are expected to have a large coverage on the positive set and in comparison a limited coverage of the full human proteome. While the traditional evaluation measures are not applicable here, the ratio of the fraction of proteins predicted on the two sets is an approximate indicator of the utility of these methods in real-life applications.

Figure 4A and B shows the results of the runs on the positive dataset and the full human proteome (for full lists of proteins with prediction results and annotations for the positive set, see Supplementary Tables S5 and S6 available online at https://academic.oup.com/bib). The methods show large variations in terms of the number of proteins predicted. In both settings, the order of methods is the same, showing that PLAAC is the most conservative of the five methods, recognizing only a

well-defined set of proteins. While true LLPS drivers recognized by PLAAC are all predicted by at least one other method as well, PLAAC can be excellent at avoiding overprediction when identifying prion-like LLPS drivers. Accordingly, all of the 17 *in vivo* LLPS drivers recognized by all methods contain low complexity PLDs, with 7 of them belonging to the well-studied FUS-like protein family [47] (see Supplementary Table S5 available online at https://academic.oup.com/bib). PSPer and PScore both predict comparable numbers of proteins, albeit with a restricted overlap, since PSPer mostly recognizes PLDs and RNA-driven phase separation, while PScore captures LLPS cases driven by π–π type inter-residue interactions. Thus, PSPer and PScore have good synergy, and their combined use extends the coverage of predictions. catGRANULE and PSPredictor have noticeably larger coverage, probably due to their intentionally broader scope. catGRANULE takes into account disorder propensity, RG and FG content, as well as RNA-binding propensity, and PSPredictor is a machine learning approach trained on proteins from LLPSDB. Accordingly, these two methods achieve the widest coverage of true positives, both giving hits that are missed by all other methods. However, they also predict a high number of proteins in the full human proteome, with over 4200 and 5000 hits for catGRANULE and PSPredictor, respectively. Given that at least some of the methods are quite conservative, it seems reasonable to assume that the 60 proteins in the human proteome identified by all methods are likely to be enriched in LLPS drivers. Twenty four of these proteins are already included in at least one of the LLPS-specific databases we presented earlier, while the rest of these proteins would be reasonable choices for further targeted experimental studies. In order to enable a more refined combination of methods, Supplementary Tables S7 and S8 available online at https://academic.oup.com/bib show a quantified similarity between the outputs of the five methods on the *in vivo* LLPS protein set and the full human proteome. While it would be useful to give objective guidance on which methods should be combined for maximum efficiency, the calculated Jaccard indices are mostly dominated by the difference in the sheer number of proteins being predicted. Therefore, the best course of action for combining methods is to be aware of the characteristics of each method and make an informed decision based on the task at hand.

Table 3 shows the overview of the performance of the five tested methods, together with their observed strengths and weaknesses. In general, the higher coverage a method achieves on the positive dataset, the less precise it gets, judging by the increasing fraction of proteins predicted in the full human proteome (marked as fold enrichment in Table 3). Thus, PLAAC seems to be the most precise method, while predicting the most limited set of true drivers. PSPer and PScore both have a more even balance between coverage and enrichment, while catGRANULE and PSPredictor both achieve high coverage with low enrichment. With the exception of PLAAC, all methods assign a score and the threshold can be adjusted by the user to increase coverage, albeit probably at the expense of losing precision. Various features of the methods shown in Table 3 make them suitable for different applications, and their combined use can further enhance their utility. In addition, in Table 3, we also detail where and how the users can access the methods. For use on single proteins, the simplest way is to use the online servers. However, for large-scale applications, it is greatly beneficial if the users can automate the runs by installing a local copy of the method, using APIs to automate queries or at least having the option of uploading sequences in reasonably large batches.

As a general notion, we point out that by testing on the positive dataset 94 out of 109 proteins are identified by at least one method. This means that some feature(s) of around 86% of known LLPS drivers are captured by available methods. Therefore, current methods are able to describe the main driving forces behind LLPS, and hence their combination—possibly by developing meta-prediction approaches—might significantly increase overall performance. Although studies relying on the combination of the above methods are still scarce, the thoughtful combination of different methods (PLAAC for detecting PLDs, catGRANULE for LLPS propensities and other methods for physicochemical properties) along with experimental validation led to important new insights into the interplay between RNA-binding domains and PLDs in the formation of MLOs recently [72]. While LLPS-prone RBDs and PLDs can already be relatively successfully identified by multiple methods, several specific types of LLPS drivers still pose challenges to all methods. These include viral LLPS driving proteins, as five out of seven viral drivers in the positive dataset were not identified by any of the methods, possibly due to their markedly different sequence composition. The molecular interactions driving LLPS have a profound effect on method efficiency as well: LLPS driven by SLiM–domain interactions or those that require phosphorylation or the presence of a membrane are generally poorly recognized by all methods. In addition, LLPS driver systems that require more than one protein for condensate formation are largely missed by most methods. These features are not independent but largely overlap as SLiM-mediated interactions require more than one protein, are often regulated by phosphorylation events as switches and are often involved in the formation of membrane-associated receptor clusters. The efficient prediction of these cases will require the development of novel methods, integrating data and relying on approaches missing from currently available prediction services.

Examples highlight the use and utility of LLPS resources

In this section, we demonstrate the utility and possible limitations of computational resources on four examples of LLPS drivers, chosen to represent different molecular mechanisms and different cellular roles. All four of these proteins have been extensively studied and can be considered textbook cases in the field. Table 4 details the main features of each protein, together with the molecular details and mechanisms characteristic of their driver role. Table 4 also shows which protein is included in which LLPS-related database. All four of the available generic LLPS databases contain all examples, and in addition, FUS and nucleophosmin are contained in the RNA granule database as well, in accord with their RNA-binding roles. Incidentally, these two proteins are also marked as having high average spectral counts in CRAPome.

Figure 5 shows the output of LLPS-related prediction methods on these cases. In the previous section, we showed on large-scale data that the five LLPS-specific methods are largely complementary, with each having use in true LLPS detection. We also emphasized that additional methods targeting the recognition of domains, SLiMs, disordered regions or low complexity regions—even though they were not specifically developed for studying LLPS—can provide useful information in detecting various types of LLPS drivers. Figure 5 demonstrates these concepts by visualizing the output of the five LLPS predictions, together with Pfam, SEG, ELM (the most comprehensive resource for SLiMs) and IUPred2A for the four example drivers.

Table 3. Summary of the five LLPS prediction methods

Method name	PubMed ID/DOI of source publication	Server address	Availability	Proteins identified in positive set		Proteins identified in the human proteome		Fold enrichment for predicting *in vivo* LLPS drivers	Strengths	Limitations
				Number	Fraction	Number	Fraction			
PLAAC	19345193	http://plaac.wi.mit.edu/	Online: batch upload of even full proteomes API: - local: downloadable in Java	29	26.6%	195	1.0%	27.8	• Potentially very low false-positive rate • Tuneable for organisms • Reasonable at detecting phosphorylation-driven LLPS	• Low coverage • No protein-level prediction score
PSPer	30994888	http://bio2byte.com/psp/	Online: batch upload of sequences API: available Local: downloadable in Python	47	43.1%	1523	7.5%	5.8	• Potentially low false-positive rate • Excellent at RNA-dependent LLPS driver prediction • Reasonable at detecting multi-component systems • Very detailed output assigning various functional regions	• Moderate coverage • Relatively long run times
PScore	29424691	http://abragam.med.utoronto.ca/~JFKlab/Software/psp.htm	Online: single sequence only API: - Local: downloadable in Python	49	45.0%	2118	10.4%	4.3	• Excellent at RNA-dependent LLPS driver prediction • Good at predicting LLPS via discrete oligomerization	• Moderate coverage • Low coverage of SLiM-driven interactions
catGRANULE	27320918	http://s.tartaglialab.com/new_submission/catGRANULES	Online: batch upload of max 500 sequences API: - Local: -	62	56.9%	4248	20.9%	2.7	• High coverage • Good at RNA-dependent LLPS driver prediction • Good at predicting LLPS via discrete oligomerization • Reasonable at detecting SLiM-driven LLPS	• Potentially high false-positive rate
PSPredictor	http://dx.doi.org/10.2139/ssrn.3515387	http://www.pkumdl.cn:8000/PSPredictor/	Online: batch upload of max 100 sequences API: - Local: -	85	78.0%	5021	24.7%	3.2	• High coverage • Reasonable at detecting multi-component systems • Reasonable at detecting SLiM-driven LLPS • Reasonable at detecting partner-dependent LLPS	• Potentially high false-positive rate • No regions are defined

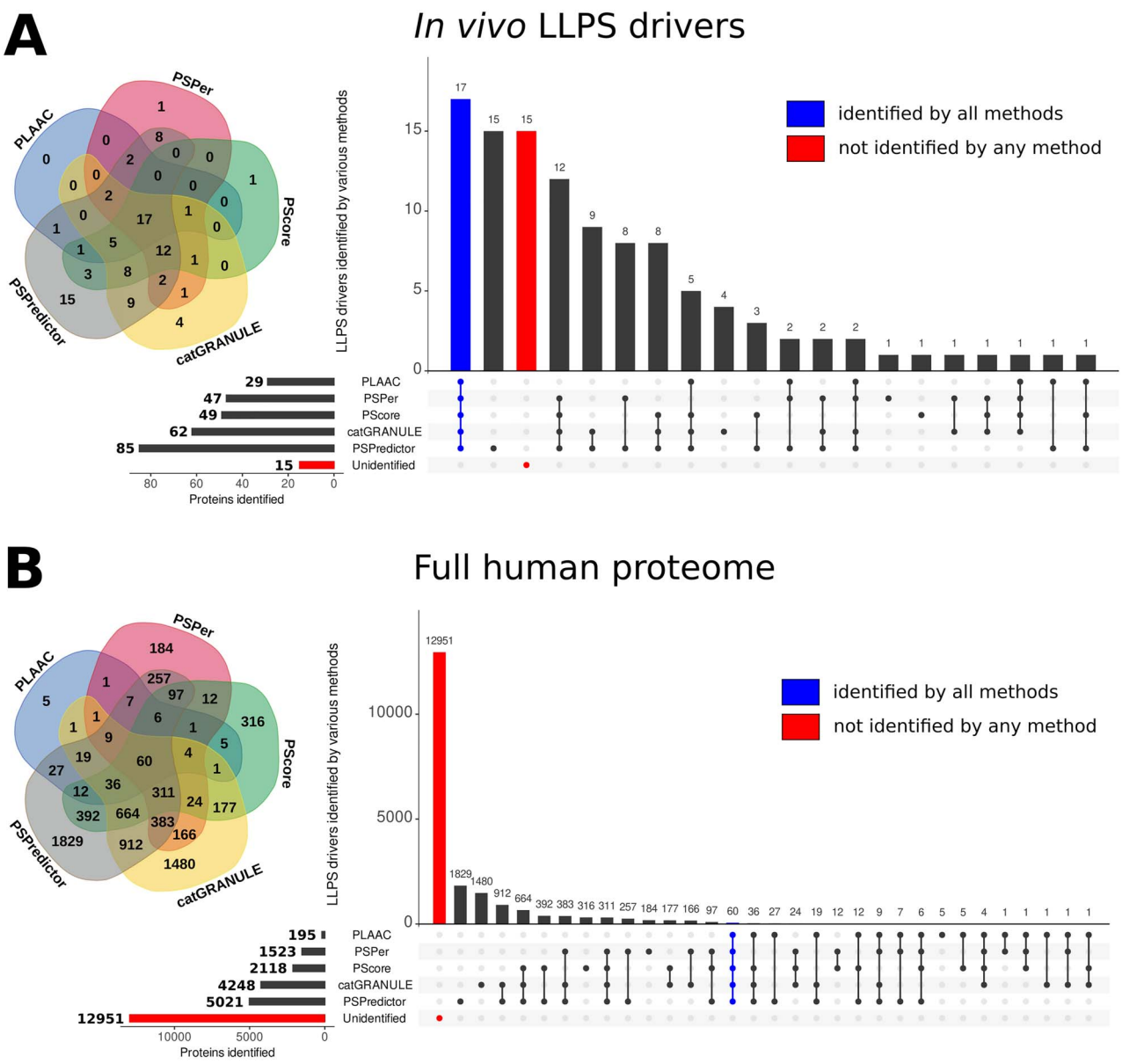

Figure 4. Results and overlap of five selected LLPS prediction methods. (**A**) Performance of the methods on 109 LLPS driver proteins taken from PhaSePro, using only cases that have *in vivo* experimental support. (**B**) Performance of the methods on the full human proteome. In both scenarios, the left side displays results in the form of a Venn diagram, while the right side displays the same data in UpSet [71] presentation. For lists of proteins identified by each method and for annotations of *in vivo* LLPS drivers, see Supplementary Tables S5 and S6 available online at https://academic.oup.com/bib.

FUS is one of the best known examples of LLPS drivers [47]. It is composed of large stretches of disordered regions displaying various compositional biases for residues or groups of residues. FUS also contains an RNA-recognition motif (RRM), and while FUS readily phase separates without RNA as well, the whole protein contributes to granule formation. The low complexity disordered regions are well captured by SEG and IUPred, and Pfam detects the RRM, as well as a RanBP2-type zinc finger (zf), which serves as an additional RNA-binding element [77]. The most characteristic region in FUS is the N-terminal PLD that harbours a large number of phosphorylation sites. This region is correctly recognized by all five LLPS predictors. PSPer correctly identifies the RRM as a separate region, as well as marking surrounding Arg-rich disordered 'spacer' regions. These regions are also highlighted as LLPS-prone by both catGRANULE and PScore. PSPredictor assigns a high score (0.99 out of 1), marking high confidence in the positive prediction and reflecting its FUS-like protein prediction focus.

DDX4 is a DEAD-box helicase, and thus, similar to FUS, it binds RNA. However, for DDX4, the phase separation does not depend on the full protein; rather it is driven by the N-terminal disordered segment rich in repeats of FG and RG [48]. The overall structure is well captured by IUPred, and Pfam clearly recognizes the two helicase domains. In spite of the compositional bias, SEG does not predict large stretches of low complexity regions, showing the clear difference of the N-terminal driver region in comparison with the N-terminal of FUS. As the N-terminal driver of DDX4 is not prion like, PLAAC is unable to identify DDX4 as

Table 4. Annotations of the four examples

	Protein name	RNA-binding protein FUS	RNA helicase DDX4	Nucleophosmin	Linker for activation of T-cells family member 1
Basic data	Gene name	FUS	DDX4	NPM1	LAT
	Organism	Human	Human	Human	Human
	UniProt accession	P35637	Q9NQI0	P06748	O43561
Inclusion in databases	PhaSePro	Yes	Yes	Yes	Yes[a]
	PhaSepDB	Yes	Yes	Yes	Yes
	LLPSDB	Yes	Yes	Yes	Yes[a]
	DrLLPS	Yes[b]	Yes[b]	Yes[b]	Yes[b]
	RNA Granule Database	Yes	No	Yes	No
	CRAPome spectral count	High	Low	High	Low
Organelle name	GO	Perinuclear region of cytoplasm	P granule	Nucleolus/ ribonucleoprotein complex	Immunological synapse
	Literature	Cytoplasmic stress granule	P granule	Nucleolus	TCR signalosome/LAT signalosome
Driver region(s)	Region(s) driving LLPS	PLD, RNA-binding regions (RRMs and RGGs)	Highly charged flexible region	Oligomerization domain, acidic motifs, RNA-binding domain	SLiMs with phosphorylated tyrosines
	Structure	Ordered + IDR	IDR	Ordered + IDR	IDR
	Reference	[47]	[48]	[73]	[74]
Molecular background of LLPS	RNA-dependent LLPS?	No	No	Yes	No
	Multi-protein system?	No	No	No	Yes
	Membrane cluster?	No	No	No	Yes
	Dominant interactions	Cation–π, π–π, electrostatic interactions	Electrostatic, cation–π interactions	Discrete oligomerization, protein–RNA interaction, multivalent domain–motif interactions	Multivalent domain–motif interactions, multivalent domain–PTM interactions

[a] Annotated as part of a multi-protein system.
[b] Annotated as a driver ('scaffold').

LLPS-prone. PSPer correctly detects the second helicase domain as an RNA-binding module, as well as detecting a short PLD-like segment between the N-terminal driver and the Q-motif of the first helicase; however, it assigns a low overall score to the protein. On the other hand, PScore and catGRANULE both pick up on the sequence signatures of the N-terminal driver region and correctly identify DDX4 as an LLPS driver. PSPredictor assigns a reasonably high score (0.64/1), correctly detecting the LLPS tendency.

Nucleophosmin (NPM1) is a constituent of the nucleolus, aiding its molecular organization. NPM1 is able to drive phase separation; however, in contrast to previous examples, it does so in a heavily partner dependent manner [73]. There are three main factors crucial for phase separation: oligomerization of NPM1 via its N-terminal core domain; interaction with arginine-rich protein partners via the three acidic patches, the first of which is inside the core domain with the second and third in a central disordered region and finally interaction with rRNA via its C-terminal nucleotide-binding domain. The overall structure and the locations of the two ordered domains are captured by Pfam and IUPred, and the disordered low complexity acidic tracts are captured correctly by SEG. As NPM1 does not incorporate a PLD, PLAAC is unable to identify it. PSPer does indicate some similarity of the central disordered region to PLDs and RRMs (probably driven by the negative charges in the acidic regions), but these signals are not strong enough to yield an overall positive prediction. PScore misses NPM1 as well, as the phase separation in this case does not involve π–π or cation–π interactions. However, catGRANULE reacts to the similarity of the disordered regions to yeast LLPS drivers in general and gives a positive prediction. PSPredictor gives a very high-confidence positive prediction with a 0.99/1 score.

LAT is a distinctively different LLPS driver in comparison to previous examples. It is part of a multi-protein system, wherein interactions with its partners are mediated via SLiM–domain interactions [74]. LAT is anchored to the plasma membrane, where it is able to form liquid-like membrane clusters amplifying incoming signals. This phase transition is driven by its

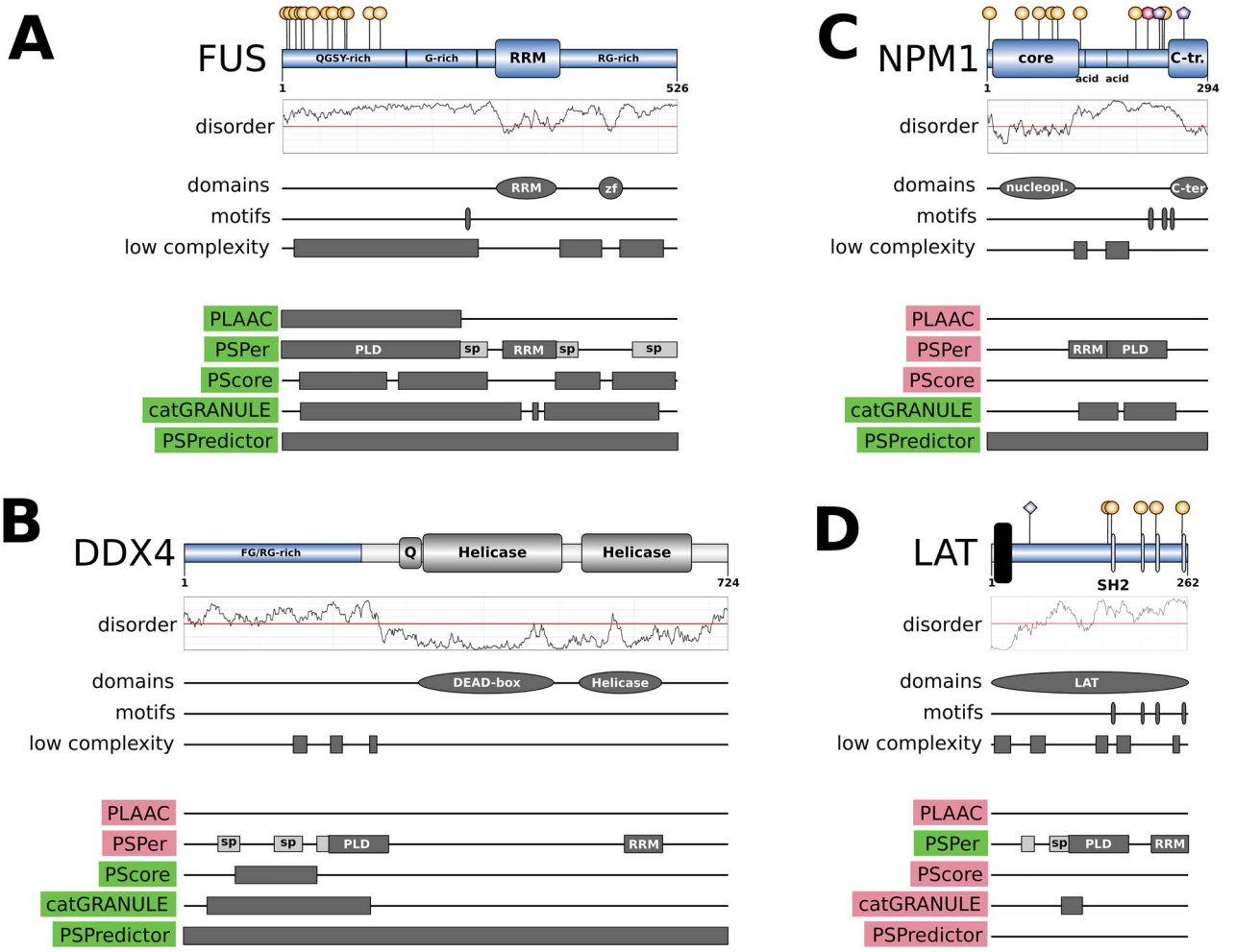

Figure 5. (A) RNA-binding protein FUS, (B) ATP-dependent RNA helicase DDX4, (C) Nucleophosmin (NPM1) and (D) Linker for activation of T-cells family member 1 (LAT). Examples of LLPS drivers. Blue regions in the protein schematics designate LLPS driver regions. Ordered domains are represented by rounded rectangles, SH2-binding SLiMs for LAT are marked with ovals and the transmembrane region is marked in black. Lollipops above the sequences represent phosphorylation (circles), ubiquitination (squares) and SUMOylation sites (pentagons). The colour of the background of prediction method names marks positive (green) and negative (red) predictions, based on the overall score, where applicable (see Data and Methods). Regions recognized by methods are shown with boxes. As PSPredictor does not assign regions only an overall score, positive predictions are represented as a box covering the full protein sequence. Domains and motifs were taken from Pfam [68] and ELM [75], low complexity and disorder were calculated using SEG [60] and IUPred2A [63] and PTMs were taken from PhosphoSitePlus [76] (see Data and methods).

interaction with SH2 domains of partner proteins, such as GRB2. The four SH2-binding SLiMs in LAT become functional only upon phosphorylation of their tyrosine residues. In addition to the LAT:GRB2 interaction, phase separation also requires SOS1, which harbours several proline-rich motifs that are able to bind to the SH3 domains in GRB2. This way LAT, GRB2 and SOS1 form a highly intertwined, non-stoichiometric network held together by a large number of transient and reversible SLiM–domain interactions. The modest affinity, large number and multivalency of these interactions provide the dynamic nature and robustness of the condensate. Figure 5 shows that the disordered nature of LAT is well captured by IUPred. The single, dedicated region predicted by Pfam and the lack of large low complexity regions marked by SEG all indicate the well-conserved nature of LAT. ELM is able to identify the SH2-binding SLiMs, which coincide with phosphorylation sites. However, since LAT is only one of the three protein components of LLPS, containing no classical PLDs and not relying on π–π or cation–π interactions, nearly all methods are unable to identify it as an LLPS driver. A surprising exception is PSPer, which is able to recognize regions that share a similarity with characteristic protein modules required for LLPS in its training model.

These four examples show that there is not one superior method for LLPS driver detection, and the combination of methods can provide insights that no single method can. For easy cases, such as FUS, virtually any method is sufficient for successful prediction. For RNA-dependent cases, such as NPM1, methods trained on such datasets are superior, while in cases where a specific type of interaction is at play, methods explicitly building on the abundance of corresponding residues (such as modelling π–π interactions by PScore for DDX4) will be extremely useful. Methods, such as PSPer, that assign types to identified regions can provide valuable information on possible functions of protein regions even if the overall prediction score is low. Testing methods on large-scale datasets (see Table 3 and Figure 4) and on specific examples (see Figure 5) clearly show that the most

challenging cases are those where phase separation requires the interaction of multiple proteins. For NPM1, the nature of these partners is reflected in the sequence of the driver with acidic patches bearing complementary charges to the partner Arg-rich motifs, and the RNA-binding domains hinting at the presence of RNA in the phase-separated system. In these cases, methods have a fair chance of taking all functional modules present in the system into account based on the driver sequence alone. However, cases like LAT, where the presence of domains and motifs present in other auxiliary drivers are not reflected in the driver sequence at all, pose the greatest challenge to LLPS prediction methods.

Conclusions and perspectives

The recent explosion in the number of proteins experimentally identified to be participating in LLPS has paved the way for the development of computational methods and resources in the LLPS field. The *in silico* counterpart of any field has the potential to work in synergy with experimental efforts, with the data generated by experiments being stored in a structured way in databases, serving as the foundation for prediction method development, in turn providing novel candidates for further experimental validation. In addition, reliable prediction methods allow for the large-scale assessment of the extent, biological roles and evolution of the biological phenomenon being studied. Thus, for every new field, the three most important computational aspects required to achieve synergy are the common language we use to describe the phenomena and observations of the measurements, structured databases to provide access to this knowledge and the development of dedicated prediction methods.

The development of CVs and ontologies to unambiguously describe LLPS-related observations are past the first steps, with the community arriving at a common language. However, as the LLPS field works with observations at several different levels—such as molecular interactions, types of functional protein modules, cellular components and high level biological processes—having a single CV or ontology is neither realistic nor desirable. Therefore, standardization efforts will need to be divided and preferably be interfaced with already existing efforts at developing standards, such as GO, the PSI initiative or ECO. In addition, several dedicated ontologies will surely be required to describe features not captured by existing descriptions.

Recently published databases already utilize these emerging concepts to describe, organize, interpret and provide access to the immense knowledge generated by experiments. As of yet, available databases are built on different concepts, focusing on either the molecular drivers of LLPS or the constituents of the formed MLOs. Thus, existing resources have different content, and the overlap between them is limited. These resources also have differing levels of description they utilize, and future works consolidating their content into a common framework will be immensely valuable. In addition, a future objective of LLPS-related database construction should be the assembly of a gold standard negative dataset—with proteins known to not undergo phase separation under any physiologically relevant set of conditions—as this would enable the proper training and assessment of current and future prediction methods.

Interestingly, prediction method development of the LLPS field was primarily not data-driven, as most current methods predate the publication of LLPS-specific datasets. Early methods were rather done as auxiliary works of experimental surveys, with the methods concentrating on capturing a single feature of sequences through an algorithmic model-based approach. Many of these first-generation methods have been published as supplementary materials, and several of them have no publicly accessible web servers. However, several user friendly methods have emerged as well, based on a limited number of known LLPS cases at the time. Contrary to expectations, their predictions show a much larger overlap than current databases. This might be an indication that even our current methods capture the main molecular backgrounds driving LLPS. That being the case, assessing these methods on a large-scale dataset shows that their false-positive rate might not make them ideal for sound proteome-wide studies just yet.

Our current predictors are based on different principles, and they mostly excel at the detection of low-complexity LLPS driver regions, especially of RNA-binding proteins. Thus, the future of LLPS prediction method development should primarily focus on specific types of LLPS drivers that are currently the most challenging, including drivers incorporating PTMs, forming membrane clusters and relying on oligomerization or multivalent domain–motif interactions between several drivers that act in concert. Therefore, methods common in bioinformatics method development, such as machine learning algorithms, will most likely need to be complemented with knowledge-based approaches. This will be hugely aided by the exploitation of the finally available, recently published databases, enabling the transition of LLPS bioinformatics into the next generation.

Data and methods

Protein sets for assessing overlap between databases

PhaSepDB [37] (version 1.3, October 2019) was represented with the ‘Reviewed’ dataset, containing 352 proteins. LLPSDB [39] (version 1 July 2019) was represented with the ‘Natural proteins’ dataset in the downloaded ‘protein.xls’ table. This dataset was filtered for proteins with a valid UniProt accession and at least one corresponding experiment indicating involvement in LLPS in the ‘LLPS.xls’ table. This filtered dataset contains 184 proteins. DrLLPS [40] was represented by the 150 proteins that have a ‘scaffold’ annotation in the downloadable 1.0 version of the database. PhaSePro [34] was represented with the downloadable set of 121 driver proteins from version v.1.1.0.

Protein sets for testing LLPS predictions

LLPS drivers annotated in PhaSePro [34] were used as the positive set in testing LLPS prediction methods. Only those 109 proteins were used that have *in vivo* support for driving phase separation.

The full human proteome was taken from UniProt [78] using the accession UP000005640 on 11 May 2020. The dataset consists of 20 350 reviewed proteins.

LLPS prediction methods

In total, five LLPS-specific prediction methods were used in large-scale analysis and for predictions. PLAAC [51] was used by downloading the Java package from the PLAAC server. PScore [36] was run using the downloadable Python script locally. PSPer [53] was run locally. catGRANULE [44] and PSPredictor [54] were run on the web servers, uploading the sequences in batches of the largest allowed size. All methods were run using default settings, except for PLAAC where the residue background frequencies were set to 100% human.

PLAAC was considered to give a positive prediction if it returned a non-zero length region. PSPer predictions were evaluated with a cutoff of 0.38, regardless of the regions predicted. PScore predictions were considered as positive if they assigned a score greater than or equal to 4 to any position in the sequence. catGRANULE predictions were evaluated with a cutoff of 0.75, roughly corresponding to the third quartile of the distribution of scores on the testing set described in the source publication. PSPredictor was evaluated using the binary value the method assigns to input proteins.

Other prediction methods and annotations

Pfam was run locally using release 32 [68]. GO terms were connected to Pfam predictions using the service offered by GO [28, 79] via InterPro mappings [80] at http://current.geneontology.org/ontology/external2go/pfam2go, using release 18 April 2020.

Low complexity regions were detected using SEG [60] with default settings. Disorder predictions were calculated using IUPred2A [63]. SLiMs were identified using the verified instances on the ELM server [75], release of 12 March 2020. PTM sites were taken from PhosphoSitePlus v6.5.9.2 [76], using only positions where there are at least two low throughput papers supporting the existence of the PTMs.

Visualization of data

UpSet presentations in Figures 2 and 4 were created using the UpSetR package [71]. Venn diagrams, also in Figures 2 and 4, were created using the web server provided by VIB/UGent at http://bioinformatics.psb.ugent.be/webtools/Venn/. Protein diagrams in Figure 5 were created using the Illustrator for Biological Sequences [81].

Key Points

- The vocabulary of the rapidly growing LLPS field is slowly nearing a point of consensus, enabling the unambiguous description of findings and their inclusion into dedicated databases.
- The vast amount of experimental data has spurred the development of dedicated databases and prediction methods.
- LLPS databases cover the majority of LLPS experimental results; however, various resources are built on markedly different underlying concepts and hence differ in coverage, scope and annotation level.
- Given the complexity of LLPS, there are no prediction methods that are clearly superior to others, as each method describes different underlying molecular driving forces, being able to recognize a different set of driver proteins.
- The combination of current dedicated prediction methods is able to identify the majority of known LLPS drivers; however, there is a tradeoff between coverage versus specificity and level of detail

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Acknowledgement

The authors are grateful for the helpful comments of Peter Tompa, which significantly improved the manuscript.

Data Availability

The data underlying this article are available in the article and in its online supplementary material.

Funding

European Union's Horizon 2020 research and innovation programme (Marie Skłodowska-Curie grant agreement no. 842490 (MIMIC) to B.M.); Hungarian Academy of Sciences (grant PREMIUM-2017-48 to R.P.); National Research, Development and Innovation Office (Fund FK-128133 to R.P.); Research Foundation Flanders (FWO) (project no. G.0328.16N to W.V.).

References

1. Banani SF, Lee HO, Hyman AA, *et al.* Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol* 2017;**18**:285–98.
2. Shin Y, Brangwynne CP. Liquid phase condensation in cell physiology and disease. *Science* 2017;**357**:eaaf4382.
3. Al-Husini N, Tomares DT, Bitar O, *et al.* α-Proteobacterial RNA degradosomes assemble liquid-liquid phase-separated RNP bodies. *Mol Cell* 2018;**71**:1027–1039.e14.
4. Nikolic J, Le Bars R, Lama Z, *et al.* Negri bodies are viral factories with properties of liquid organelles. *Nat Commun* 2017;**8**(1):58.
5. Kaganovich D. There is an inclusion for that: material properties of protein granules provide a platform for building diverse cellular functions. *Trends Biochem Sci* 2017;**42**:765–76.
6. Pancsa R, Schad E, Tantos A, *et al.* Emergent functions of proteins in non-stoichiometric supramolecular assemblies. *Biochim. Biophys Acta Proteins Proteomics* 2019;**1867**:970–9.
7. Alberti S. The wisdom of crowds: regulating cell function through condensed states of living matter. *J Cell Sci* 2017;**130**:2789–96.
8. Alberti S, Gladfelter A, Mittag T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell* 2019;**176**:419–34.
9. Sheu-Gruttadauria J, MacRae IJ. Phase transitions in the assembly and function of human miRISC. *Cell* 2018;**173**:946–957.e16.
10. Banjade S, Rosen MK. Phase transitions of multivalent proteins can promote clustering of membrane receptors. *Elife* 2014;**3**:e04123.
11. Yap K, Mukhina S, Zhang G, *et al.* A short tandem repeat-enriched RNA assembles a nuclear compartment to control alternative splicing and promote cell survival. *Mol Cell* 2018;**72**:525–540.e13.
12. Schmidt HB, Görlich D. Nup98 FG domains from diverse species spontaneously phase-separate into particles with nuclear pore-like permselectivity. *Elife* 2015;**4**:e04251.
13. Yoo H, Triandafillou C, Drummond DA. Cellular sensing by phase separation: using the process, not just the products. *J Biol Chem* 2019;**294**:7151–9.

14. Jung J-H, Barbosa AD, Hutin S, *et al*. A prion-like domain in ELF3 functions as a thermosensor in Arabidopsis. *Nature* 2020;**585**:256–60.
15. Zacharogianni M, Aguilera-Gomez A, Veenendaal T, *et al*. A stress assembly that confers cell viability by preserving ERES components during amino-acid starvation. *Elife* 2014;**3**:e04132.
16. Guillén-Boixet J, Buzon V, Salvatella X, *et al*. CPEB4 is regulated during cell cycle by ERK2/Cdk1-mediated phosphorylation and its assembly into liquid-like droplets. *Elife* 2016;**5**:e19298.
17. Wen W. Phase separation in asymmetric cell division. *Biochemistry* 2020;**59**:47–56.
18. Shan Z, Tu Y, Yang Y, *et al*. Basal condensation of Numb and Pon complex via phase transition during Drosophila neuroblast asymmetric division. *Nat Commun* 2018;**9**:737.
19. Klosin A, Oltsch F, Harmon T, *et al*. Phase separation provides a mechanism to reduce noise in cells. *Science* 2020;**367**:464–8.
20. Hubstenberger A, Courel M, Bénard M, *et al*. P-body purification reveals the condensation of repressed mRNA regulons. *Mol Cell* 2017;**68**:144–157.e5.
21. Khong A, Matheny T, Jain S, *et al*. The stress granule transcriptome reveals principles of mRNA accumulation in stress granules. *Mol Cell* 2017;**68**:808–820.e5.
22. Mitrea DM, Chandra B, Ferrolino MC, *et al*. Methods for physical characterization of phase-separated bodies and membrane-less organelles. *J Mol Biol* 2018;**430**:4773–805.
23. Vernon RM, Forman-Kay JD. First-generation predictors of biological protein phase separation. *Curr Opin Struct Biol* 2019;**58**:88–96.
24. International Society for Biocuration. Biocuration: distilling data into knowledge. *PLoS Biol* 2018;**16**:e2002846.
25. Côté R, Reisinger F, Martens L, *et al*. The ontology lookup service: bigger and better. *Nucleic Acids Res* 2010;**38**:W155–60.
26. Chibucos MC, Siegele DA, Hu JC, *et al*. The evidence and conclusion ontology (ECO): supporting GO annotations. *Methods Mol Biol* 2017;**1446**:245–59.
27. Sivade Dumousseau M, Alonso-López D, Ammari M, *et al*. Encompassing new use cases—level 3.0 of the HUPO-PSI format for molecular interactions. *BMC Bioinformatics* 2018;**19**:134.
28. The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res* 2019;**47**:D330–8.
29. Shih J-W, Wang W-T, Tsai T-Y, *et al*. Critical roles of RNA helicase DDX3 and its interactions with eIF4E/PABP1 in stress granule assembly and stress response. *Biochem J* 2012;**441**:119–29.
30. Rayman JB, Karl KA, Kandel ER. TIA-1 self-multimerization, phase separation, and recruitment into stress granules are dynamically regulated by Zn. *Cell Rep* 2018;**22**:59–71.
31. Patel PH, Barbee SA, GW-Bodies BJT. P-bodies constitute two separate pools of sequestered non-translating RNAs. *PLoS One* 2016;**11**:e0150291.
32. Ma W, Mayr CA. Membraneless organelle associated with the endoplasmic reticulum enables 3′UTR-mediated protein-protein interactions. *Cell* 2018;**175**:1492–1506.e19.
33. Lin Y-H, Qiu D-C, Chang W-H, *et al*. The intrinsically disordered N-terminal domain of galectin-3 dynamically mediates multisite self-association of the protein through fuzzy interactions. *J Biol Chem* 2017;**292**:17845–56.
34. Mészáros B, Erdős G, Szabó B, *et al*. PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res* 2020;**48**:D360–7.
35. Mellacheruvu D, Wright Z, Couzens AL, *et al*. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods* 2013;**10**:730–6.
36. Vernon RM, Chong PA, Tsang B, *et al*. Pi–Pi contacts are an overlooked protein feature relevant to phase separation. *Elife* 2018;**7**:e31486.
37. You K, Huang Q, Yu C, *et al*. PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic Acids Res* 2019;**48**:D354–9.
38. Youn J-Y, Dyakov BJA, Zhang J, *et al*. Properties of stress granule and P-body proteomes. *Mol Cell* 2019;**76**:286–94.
39. Li Q, Peng X, Li Y, *et al*. LLPSDB: a database of proteins undergoing liquid-liquid phase separation in vitro. *Nucleic Acids Res* 2020;**48**:D320–7.
40. Ning W, Guo Y, Lin S, *et al*. DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. *Nucleic Acids Res* 2020;**48**:D288–95.
41. Yu C, Shen B, You K, *et al*. Proteome-scale analysis of phase-separated proteins in immunofluorescence images. *Brief Bioinform* 2020;bbaa187.
42. Iwashita K, Handa A, Shiraki K. Coacervates and coaggregates: liquid-liquid and liquid-solid phase transitions by native and unfolded protein complexes. *Int J Biol Macromol* 2018;**120**:10–8.
43. Cinar S, Cinar H, Chan HS, *et al*. Pressure-sensitive and osmolyte-modulated liquid-liquid phase separation of eye-lens γ-crystallins. *J Am Chem Soc* 2019;**141**:7347–54.
44. Bolognesi B, Lorenzo Gotor N, Dhar R, *et al*. A concentration-dependent liquid phase separation can cause toxicity upon increased protein expression. *Cell Rep* 2016;**16**:222–31.
45. Kato M, Han TW, Xie S, *et al*. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* 2012;**149**:753–67.
46. Banani SF, Rice AM, Peeples WB, *et al*. Compositional control of phase-separated cellular bodies. *Cell* 2016;**166**:651–63.
47. Wang J, Choi J-M, Holehouse AS, *et al*. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell* 2018;**174**:688–699.e16.
48. Nott TJ, Petsalaki E, Farber P, *et al*. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol Cell* 2015;**57**:936–47.
49. Hughes MP, Sawaya MR, Boyer DR, *et al*. Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks. *Science* 2018;**359**:698–701.
50. Hughes MP, Goldschmidt L, Eisenberg DS. The prevalence and distribution in genomes of low-complexity, amyloid-like, reversible, kinked segment (LARKS), a common structural motif in amyloid-like fibrils. *BioRxiv* 2020. doi: 10.1101/2020.12.08.415679.
51. Lancaster AK, Nutter-Upham A, Lindquist S, *et al*. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics* 2014;**30**:2501–2.
52. Castillo V, Graña-Montes R, Sabate R, *et al*. Prediction of the aggregation propensity of proteins from the primary sequence: aggregation properties of proteomes. *Biotechnol J* 2011;**6**:674–85.
53. Orlando G, Raimondi D, Tabaro F, *et al*. Computational identification of prion-like RNA-binding proteins that form liquid phase-separated condensates. *Bioinformatics* 2019;**35**:4617–23.
54. Sun T, Li Q, Xu Y, *et al*. Prediction of liquid-liquid phase separation proteins using machine learning. *BioRxiv* 2020;10.1101/842336.

55. Das S, Amin AN, Lin Y-H, *et al*. Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters. *Phys Chem Chem Phys* 2018;**20**:28558–74.
56. Choi J-M, Dar F, Pappu RVLASSI. A lattice model for simulating phase transitions of multivalent proteins. *PLoS Comput Biol* 2019;**15**:e1007028.
57. Martin EW, Holehouse AS, Peran I, *et al*. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* 2020;**367**:694–9.
58. Zambrano R, Conchillo-Sole O, Iglesias V, *et al*. PrionW: a server to identify proteins containing glutamine/asparagine rich prion-like domains and their amyloid cores. *Nucleic Acids Res* 2015;**43**:W331–7.
59. Espinosa Angarica V, Angulo A, Giner A, *et al*. PrionScan: an online database of predicted prion domains in complete proteomes. *BMC Genomics* 2014;**15**:102.
60. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* 1993;**17**:149–63.
61. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999;**27**:573–80.
62. Holehouse AS, Ahad J, Das RK, *et al*. CIDER: classification of intrinsically disordered ensemble regions. *Biophys J* 2015;**108**:228a.
63. 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* 2018;**46**:W329–37.
64. Piovesan D, Tabaro F, Paladin L, *et al*. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res* 2018;**46**:D471–6.
65. Dosztányi Z, Mészáros B, Simon I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief Bioinform* 2010;**11**:225–43.
66. Tompa P, Davey NE, Gibson TJ, *et al*. A million peptide motifs for the molecular biologist. *Mol Cell* 2014;**55**:161–9.
67. Bellay J, Han S, Michaut M, *et al*. Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol* 2011;**12**:R14.
68. El-Gebali S, Mistry J, Bateman A, *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res* 2019;**47**:D427–32.
69. Mitchell AL, Attwood TK, Babbitt PC, *et al*. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res* 2019;**47**:D351–60.
70. Alberti S, Halfmann R, King O, *et al*. A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. *Cell* 2009;**137**:146–58.
71. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 2017;**33**:2938–40.
72. Gotor NL, Armaos A, Calloni G, *et al*. RNA-binding and prion domains: the Yin and Yang of phase separation. *Nucleic Acids Res* 2020;**48**:9491–504.
73. Mitrea DM, Cika JA, Guy CS, *et al*. Nucleophosmin integrates within the nucleolus via multi-modal interactions with proteins displaying R-rich linear motifs and rRNA. *Elife* 2016;**5**:13571.
74. Su X, Ditlev JA, Hui E, *et al*. Phase separation of signaling molecules promotes T cell receptor signal transduction. *Science* 2016;**352**:595–9.
75. Kumar M, Gouw M, Michael S, *et al*. ELM-the eukaryotic linear motif resource in 2020. *Nucleic Acids Res* 2020;**48**:D296–306.
76. Hornbeck PV, Zhang B, Murray B, *et al*. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* 2015;**43**:D512–20.
77. Loughlin FE, Lukavsky PJ, Kazeeva T, *et al*. The solution structure of FUS bound to RNA reveals a bipartite mode of RNA recognition with both sequence and shape specificity. *Mol Cell* 2019;**73**:490–504.e6.
78. Consortium UP. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**:D506–15.
79. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
80. Mitchell A, Chang H-Y, Daugherty L, *et al*. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* 2015;**43**:D213–21.
81. Liu W, Xie Y, Ma J, *et al*. IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics* 2015;**31**:3359–61.