

Rethinking bioinformatics in liquid–liquid phase separation: data resources, predictive models, and an event-centric perspective

Zi-long Yuan^{1,†}, Bo Wang^{1,†}, Yu-lu Chen², Hao-qi Huang², Bin-hao Li¹, Ahmed Zahoor¹, Liping Ren³, Mengze Du^{2,3,*},
Rui-qin Fang^{2,*}, Lin Ning^{1,2,4,*} 

¹School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, 548 Binwen Road, Binjiang District, Hangzhou 310053, Zhejiang, China

²School of Life Science and Technology, University of Electronic Science and Technology of China, No. 2006 Xiyuan Avenue, West Hi-Tech Zone, Chengdu 611731, Sichuan, China

³School of Healthcare and Technology, Chengdu Neusoft University, No. 1 Dongruan Avenue, Qingchengshan Town, Dujiangyan District, Chengdu 611844, Sichuan, China

⁴Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, No. 1 Chengdian Road, Kecheng District, Quzhou 324003, Zhejiang, China

*Corresponding authors. Mengze Du, School of Healthcare and Technology, Chengdu Neusoft University, Chengdu, Sichuan Province 611844, China. E-mail: dumengze@nsu.edu.cn; Rui-qin Fang, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, Sichuan Province 611731, China. E-mail: fangrq@uestc.edu.cn; Lin Ning, School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, Hangzhou, Zhejiang Province 310053, China. E-mail: ninglin3000@uestc.edu.cn.

[†]Zi-long Yuan and Bo Wang contributed equally to this work.

Abstract

Liquid–liquid phase separation (LLPS) has emerged as a fundamental mechanism underlying the formation and regulation of membrane-less cellular compartments and is increasingly implicated in diverse physiological processes and diseases. Alongside rapid experimental and high-throughput advances, bioinformatics data resources and computational models have expanded substantially, enabling systematic cataloguing of LLPS-associated components and prediction of phase-separation behavior from molecular features. However, the resulting computational landscape remains highly fragmented. In this review, we provide a comprehensive and critical synthesis of bioinformatics resources and predictive modelling approaches for LLPS. We examine and compare major LLPS databases, highlighting differences in evidence types, curation strategies, coverage, and cross-resource inconsistencies that limit integrative analysis. We then survey computational models across core LLPS prediction tasks, encompassing more than 40 representative algorithms and tracing methodological evolution from classical machine learning to deep learning and large language model-based frameworks. By integrating these advances, we identify a fundamental mismatch between molecule-centric data abstractions and the inherently multicomponent, context-dependent organization of LLPS phenomena. We argue that future progress may benefit from event-centric frameworks that explicitly represent molecular assemblies, contextual conditions and observable phase behaviors, thereby providing a coherent foundation for next-generation LLPS datasets and computational models with improved mechanistic interpretability and translational relevance.

Keywords liquid–liquid phase separation, bioinformatics, database, predictive model, LLPS events

Introduction

Liquid–liquid phase separation (LLPS) has emerged over the past decade as a fundamental physical principle underlying intracellular organization [1–3]. Since the seminal observation by Brangwynne and colleagues [4] in 2009 that P granules in *Caenorhabditis elegans* exhibit liquid-like behavior, phase separation has been recognized as a widespread mechanism by which cells organize biochemical reactions without the need for membrane-bound compartments [5–8]. Through LLPS, biomacromolecules spontaneously demix into dense and dilute phases once local concentrations exceed critical thresholds, giving rise to dynamic assemblies commonly referred to as biomolecular condensates [9, 10].

A growing number of cellular structures—including stress granules, processing bodies, nucleoli, and Cajal bodies—have been shown to form through phase separation [11–13]. These condensates selectively enrich proteins, RNAs and other macromolecules, thereby establishing spatially confined yet highly dynamic microenvironments that regulate gene expression, signal transduction, RNA metabolism, and stress responses [14, 15]. Unlike classical organelles, phase-separated condensates are typically reversible and responsive to environmental cues [16], enabling rapid cellular adaptation. Increasing evidence further links dysregulated phase behavior to pathological conditions such as neurodegenerative diseases, cancer and viral infection, underscoring the biological and clinical relevance of LLPS [17–20]. Importantly, LLPS is often not determined by an isolated

Received: January 6, 2026. **Revised:** March 24, 2026. **Accepted:** April 21, 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

biomolecule alone, but emerges from coordinated interactions among multiple components under specific physicochemical and cellular conditions. In this review, we refer to such context-dependent, multicomponent manifestations of phase separation as LLPS events, a perspective that helps connect the biological organization of condensates with the computational abstractions currently used to describe them.

At the molecular level, proteins involved in LLPS are frequently enriched in low-complexity domains and intrinsically disordered regions (IDRs) [21, 22]. The conformational plasticity and compositional heterogeneity of these regions complicate structure-centric analyses and challenge traditional experimental approaches [23, 24]. While techniques such as fluorescence recovery after photobleaching (FRAP) [25] and *in vitro* droplet reconstitution are widely used assays for validating phase separation, they are inherently low-throughput and labor-intensive [26]. As a consequence, experimental characterization alone is insufficient [27] to explore the rapidly expanding space of candidate molecules, interactions and conditions associated with LLPS [28]. In parallel, advances in high-throughput sequencing [29], proteomics [30], and systems biology [31] have generated vast amounts of molecular data relevant to phase separation [32]. This data-rich environment has catalyzed a shift in LLPS research from predominantly experiment-driven discovery toward data-driven and computationally assisted analysis [33, 34]. Bioinformatics approaches have become indispensable for integrating heterogeneous datasets, identifying potential phase-separating components and developing predictive models of LLPS-related behavior at scale [35–37].

Over the past decade, a diverse ecosystem of LLPS-related databases and computational prediction models has emerged [38, 39]. These resources have enabled systematic screening of proteins and RNAs, characterization of sequence features associated with phase behavior and large-scale comparative analyses across species and conditions [33]. However, the rapid expansion of data and methods has also exposed conceptual limitations [40]. Most existing computational studies implicitly treat LLPS as an intrinsic property of individual biomolecules, abstracting phase separation into molecule-centric prediction tasks rather than representing it as a structured, context-dependent biological event. Such abstractions, while practical, only partially capture the conditional, multi-component, and context-dependent nature of LLPS *in vivo* [41, 42].

In this review, we provide a comprehensive and critical synthesis of bioinformatics applications in LLPS research, focusing on data resources and computational prediction models. We systematically survey major public databases cataloging LLPS-related molecules, condensates and disease associations, and we review more than 40 computational approaches spanning traditional machine learning, deep learning and protein language model-based methods. Beyond summarizing existing efforts, we articulate a unifying event-centric perspective that reframes LLPS as a structured biological event. By aligning data representation and modeling strategies with the multi-component organization of phase separation, we aim to outline future directions for integrative, mechanistically informed and translationally relevant LLPS bioinformatics.

Landscape and motivation for a comprehensive review of liquid–liquid phase separation computational research

Over the past decade, research on LLPS has experienced sustained and rapid growth, driven by continuous experimental discoveries and

conceptual advances. Bibliometric analyses based on PubMed and Web of Science reveal a steady increase in LLPS-related publications from January 2014 to February 2026 (Fig. 1a and b), reflecting the consolidation of phase separation as a central paradigm in cell biology, biophysics, and disease research. This expansion also mirrors the growing recognition of biomolecular condensates as dynamic organizational principles that can respond to cellular state and environmental perturbations. Importantly, the development of the field has been both quantitative and conceptual, with LLPS now implicated in an increasingly broad spectrum of cellular processes and physiological contexts.

Early LLPS studies primarily focused on protein- and RNA-driven condensates, emphasizing sequence features such as low complexity and intrinsic disorder. More recent work has substantially broadened this view, demonstrating that additional classes of biomolecules—including glycogen and other polysaccharides—can participate in or modulate phase separation [43–45]. These findings underscore LLPS as a collective and context-dependent phenomenon shaped by multiple interacting components and environmental conditions. As mechanistic understanding becomes increasingly complex, analytical frameworks capable of integrating heterogeneous molecular species and contextual information are required. In parallel, computational and bioinformatics research on LLPS has continued to evolve. Curated databases have been developed to collect experimentally validated phase-separating proteins, RNAs, condensates and disease links, while predictive models have progressed from early physicochemical feature-based approaches to machine learning, deep learning and, more recently, large pretrained protein language models. Together, these developments have enabled large-scale screening and hypothesis generation, substantially accelerating LLPS research.

To date, approximately 15 review articles have examined bioinformatics and computational aspects of LLPS, encompassing database construction, predictive modelling, and benchmarking efforts (Fig. 2). However, closer inspection reveals several recurring limitations. First, a number of early surveys have become outdated as LLPS datasets and computational tools have expanded rapidly in both scale and experimental modalities [46]. Second, many more recent reviews remain narrowly centered on molecule-centric tasks—most commonly protein-level LLPS propensity prediction—without systematically accounting for partner dependence, environmental factors such as temperature, pH and ionic strength, or multicomponent competition that governs *in vivo* partitioning and context specificity [47–51]. Third, benchmark-oriented studies [40, 52, 53] that apply stringent filtering criteria often yield small datasets, limiting practical utility and rendering existing benchmarks ill-suited to represent the conditional and multicomponent nature of LLPS.

These review-level limitations reflect deeper challenges within the computational landscape itself. LLPS databases differ substantially in curation criteria, evidence types, update cadence and annotation granularity, while extensive cross-resource redundancy complicates harmonization [54–57]. Inconsistent identifier mapping and heterogeneous metadata schemas further undermine interoperability, making it difficult to trace the same molecule, condensate or experiment across resources. On the modelling side, prediction methods vary widely in task formulations, training data and evaluation protocols; even when addressing ostensibly similar questions, studies frequently adopt divergent definitions of positive and negative samples, representations and performance metrics [58–62]. As a consequence, computational outputs are often not directly comparable or readily integrated, and methodological advances do not consistently translate

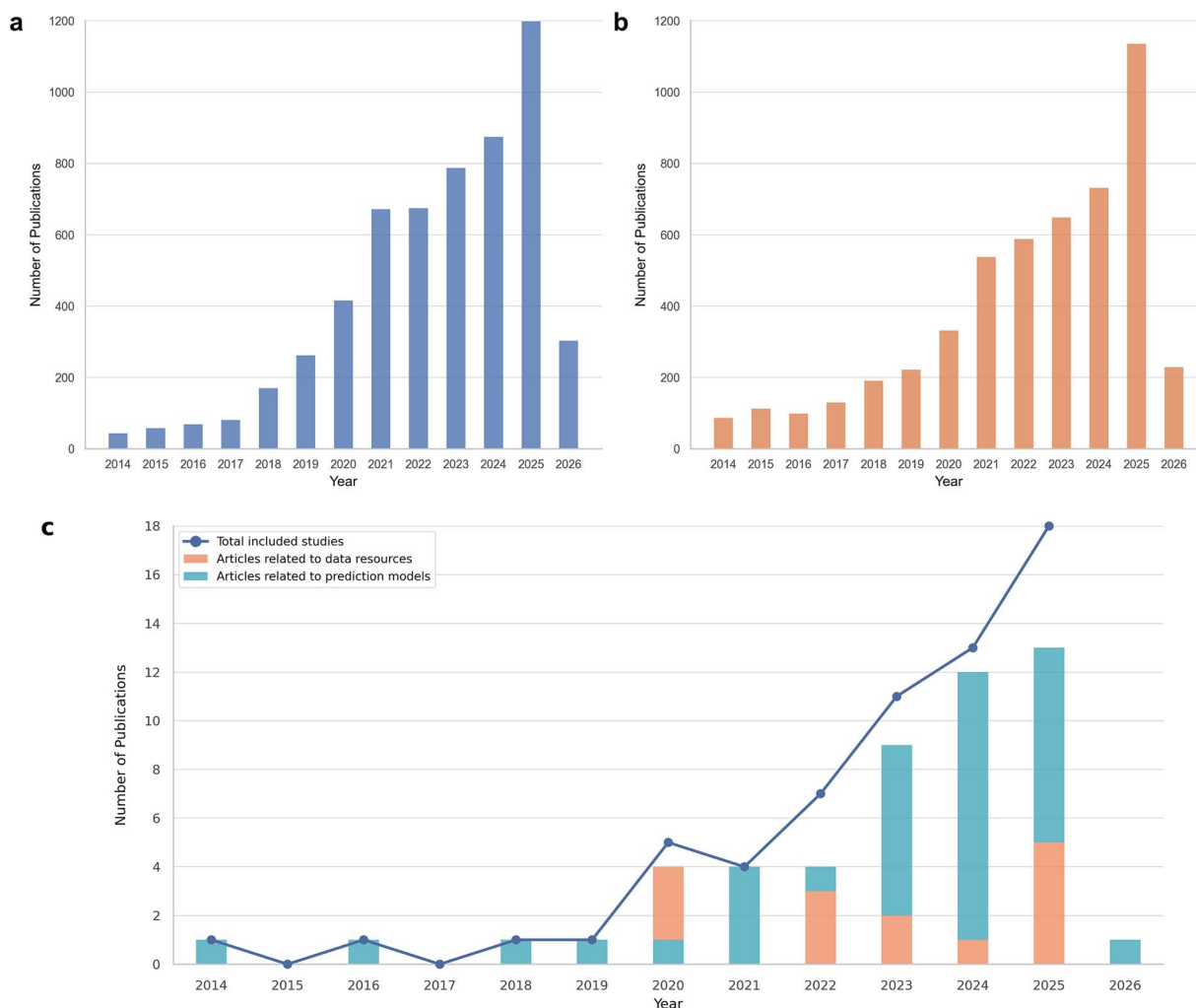


Figure 1 Longitudinal trends in LLPS-related publications and temporal distribution of reviewed literature. Panels a and b present the annual number of LLPS related research articles retrieved from PubMed and Web of Science, respectively, using LLPS-related keywords over the period January 2014–February 2026. Panel c depicts the yearly distribution of the review articles included in this study. The stacked bars represent articles related to data resources and prediction models, whereas the overlaid line (total included studies) indicates the total number of included reviews per year.

into improved biological interpretability or robust, context-relevant hypotheses [63].

Collectively, these observations highlight a widening gap between the pace of LLPS research and the scope of existing computational syntheses. The absence of a comprehensive and up-to-date review limits researchers' ability to navigate an increasingly complex ecosystem of data resources and predictive models, leading to duplicated efforts, inconsistent terminology, and uneven adoption of best practices. Moreover, the prevailing molecule-centric perspective constrains how computational problems are formulated and evaluated, frequently overlooking the conditional, multicomponent, and context-dependent organization of LLPS *in vivo*. Motivated by these challenges, the present review aims to systematically organize LLPS-related data resources and computational models, critically assess their assumptions and limitations, and establish a foundation for event-centric and integrative computational frameworks that more faithfully reflect the biological organization of LLPS. The scope of the computational literature examined in this review, including temporal publication trends of LLPS-related data resources and predictive models, is summarized in Fig. 1c, while detailed search strategies,

inclusion and exclusion criteria and full query terms are provided in Supplementary Fig. S1 and Supplementary Table S1.

Liquid–liquid phase separation data resources: scope, curation strategies, and structural limitations

Overview of liquid–liquid phase separation data resources and organizational paradigms

Over the past decade, a diverse and rapidly expanding collection of public databases has been developed to support data-driven studies of LLPS. These resources reflect both the growing recognition of LLPS as a fundamental biological process and the increasing demand for systematic organization of experimentally derived knowledge. Despite differences in scope and design, most existing LLPS databases can be broadly grouped into three categories based on their primary annotation focus and dominant data presentation form: molecule-centric datasets, condensate- or organelle-oriented resources, and disease-associated LLPS databases; notably, the molecule-centric category can be further subdivided into protein-centric and RNA-related

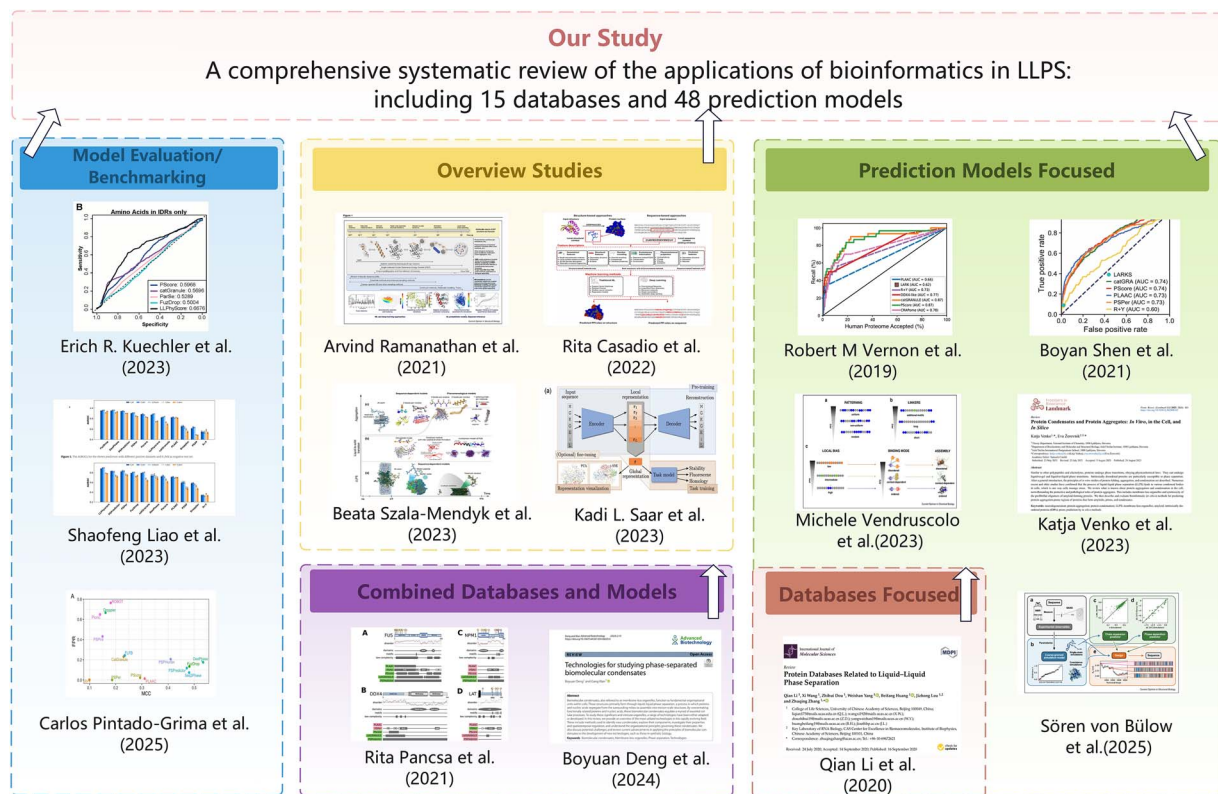


Figure 2 Positioning of this review relative to existing bioinformatics-focused LLPS surveys. This figure compares fifteen previously published reviews on LLPS bioinformatics, grouped according to their primary scope: five focusing exclusively on LLPS prediction models, one dedicated to data resources, two integrating both resources and predictive models, four providing broad descriptive overviews, and three centered on benchmarking or systematic evaluation of LLPS predictors.

resources. Table 1 provides an overview of representative LLPS-related databases, including their database name, version, resource type, number of entries, associated journal publication, release date, URL, and brief notes. Because some resources contain information relevant to more than one category, classification in this review follows the main organizational unit of each database, while cross-category features are indicated in the Notes column of Table 1. For databases with multiple released versions, the release dates of different versions are listed separately to facilitate comparison and statistical analysis across database updates.

Protein-centric databases constitute the earliest and most extensively developed class of LLPS resources. Databases such as LLPSDB [73, 74] and PhaSePro [76] curate experimentally validated phase-separating proteins and, in some cases, delineate LLPS-driving regions at the sequence level. These resources typically emphasize low-complexity domains, IDRs and physicochemical features associated with phase behavior, providing high-confidence reference datasets that have been widely used for computational model development. To improve coverage across species, integrative resources such as DrLLPS [77] and BAV-LLPS [68] further expand protein-centric collections through homology-based inference, substantially increasing dataset size while introducing heterogeneous evidence confidence. Complementing protein-focused resources, a smaller but growing set of databases addresses the role of RNA in LLPS, such as RNAPhaSep [75] and RPS [70, 71]. These databases collect information on RNAs that participate in or regulate phase separation, highlighting the importance of protein–RNA interactions in condensate formation. In parallel, condensate- or organelle-oriented databases, exemplified

by PhaSepDB [64–66], CD-CODE [72], and MLOsMetaDB [38], organize LLPS-related information at the level of membraneless organelles, linking molecular components to specific cellular structures and sub-cellular localizations. This higher-level abstraction provides a bridge between molecular annotations and cellular phenotypes. Finally, PhaSeDis [69], the disease-oriented LLPS resources integrate phase separation information with pathological contexts, cataloging associations between LLPS-related molecules, condensates and human diseases. Such databases support translational studies by connecting LLPS dysregulation to disease mechanisms and clinical phenotypes.

Collectively, these databases provide an essential foundation for computational LLPS research by aggregating dispersed experimental evidence into accessible resources. However, despite differences in annotation scope and granularity, the dominant organizational paradigm across all three categories remains molecule-centric. Individual proteins or RNAs typically serve as the primary units of annotation, while cooperative interactions, contextual dependencies, and event-level outcomes are only partially represented. This shared abstraction underlies both the strengths and limitations of current LLPS data resources, highlighting the importance of evidence-aware annotation and curation.

Evidence hierarchy and curation strategies

The reliability and utility of LLPS data resources are fundamentally shaped by the types of experimental evidence they curate and the strategies used to organize such evidence. Existing LLPS databases integrate information derived from a wide spectrum of experimental

Table 1 Overview of LLPS related databases.

Database	Version	Type	Entry	Journal	Date	URL	Notes
PhaSepDB [64–66]	V3.0	Protein	3484	Nucleic Acids Research	2025/10/8	https://db.phasep.pro/	MLOs
	V2.1	Protein	9492	Nucleic Acids Research	2023/1/6	http://db2.phasep.pro/	MLOs
	V1.0	Protein	2914	Nucleic Acids Research	2020/1/8	/	/
ricePSP [67]	V1.0	Protein	10 654	Genome Biology	2025/11/17	https://ricepsp.github.io/	Proteins in rice
BAV-LLPS [68]	V1.0	Protein	5278	Bioinformatics	2025/10/2	https://bav-llps-db.bioinformatica.org/	Proteins in bacteria, archaea, and viruses
PhaSeDis [69]	V1.0	Disease	931	Genomics, Proteomics & Bioinformatics	2025/5/10	http://mlodis.phasep.pro	LLPS factors and related diseases
RPS [70, 71]	V2.0	RNA	171 301	Nucleic Acids Research	2025/1/6	https://rps.renlab.cn/	/
	V1.0	RNA	42 417	Nucleic Acids Research	2022/1/7	/	/
MLOsMetaDB [38]	V1.0	Protein	12 038	Protein Science	2024/1/1	http://mlos.leloir.org.ar	MLOs
CD-CODE [72]	V1.0	Biomolecular Condensates	9861	Nature Methods	2023/4/6	https://cd-code.org/	/
LLPSDB [73, 74]	V2.0	Protein	2917	Bioinformatics	2022/1/13	http://bio-comp.org.cn/llpsdbv2	Proteins in vitro
	V1.0	Protein	1175	Nucleic Acids Research	2020/1/8	http://bio-comp.org.cn/llpsdb	/
RNAPhaSep [75]	V1.0	RNA	1113	Nucleic Acids Research	2022/1/7	http://www.rnaphasep.cn	/
PhaSePro [76]	V1.0	Protein	121	Nucleic Acids Research	2020/1/8	https://phasepro.elte.hu	Driver proteins in vivo
DrLLPS [77]	V1.0	Protein	437 887	Nucleic Acids Research	2020/1/8	http://llps.biocuckoo.cn/ http://bio-comp.ucas.ac.cn/llpsdb	Proteins in eukaryotes

approaches, ranging from low-throughput, high-confidence assays to indirect or large-scale experimental observations. However, the heterogeneity of evidence types is often only partially reflected in database annotations, complicating downstream interpretation and computational use. For the purpose of this review, we use “LLPS-related” as a working, evidence-aware term for molecules, condensates or records that are explicitly connected in the source literature to phase-separation phenomena or to closely related condensate contexts, while recognizing substantial differences in evidential strength. Records supported by direct experimental observations of phase behavior or material properties *in vitro* and/or *in vivo* are treated as higher-confidence LLPS evidence. By contrast, punctate localization, colocalization with membraneless organelles, or broader MLO association alone are considered indirect but biologically informative evidence, unless accompanied by additional observations consistent with phase separation. Accordingly, homology-based, text-mined, or predictor-linked records are interpreted here as putative or lower-confidence extensions rather than equivalent substitutes for experimentally validated LLPS evidence.

At one end of the spectrum, classical *in vitro* assays provide the strongest direct evidence for phase separation. These include droplet formation assays under controlled conditions and FRAP, which assess the material properties and dynamics of condensates [78]. Databases such as LLPSDB [73, 74] and PhaSePro [76] rely heavily on manual expert curation of such studies, extracting detailed annotations related to experimental conditions, sequence regions driving phase

separation and qualitative phase behavior. While these data offer high confidence and mechanistic insight, they are inherently limited in scale due to the labor-intensive nature of the experiments and the curation process. In contrast, *in vivo* observations—such as the formation of punctate structures in cells or the localization of proteins to membraneless organelles—provide valuable physiological context but often represent indirect evidence of LLPS. These observations may reflect phase separation, but they can also arise from alternative mechanisms, including scaffold-based assembly or stable protein complexes. The extent to which databases distinguish between direct phase separation assays and indirect cellular observations varies substantially, introducing ambiguity in evidence interpretation.

To address scalability, several databases employ hybrid curation strategies that combine manual annotation with homology-based inference or automated text-mining pipelines. Resources such as DrLLPS and BAV-LLPS expand LLPS-associated protein sets by extrapolating experimentally validated cases to homologous proteins, thereby increasing coverage across species [68, 77]. In a related but more prediction-oriented direction, ricePSP extends LLPS resource coverage to rice proteins, but because its entries are derived entirely from computational prediction rather than direct experimental validation, it should be interpreted as a predictor-linked, lower-confidence resource within the present evidence hierarchy [67]. As LLPS-related data have expanded at scale, redundancy and fragmentation across different LLPS databases have become increasingly evident. In response, secondary integration efforts have begun to emerge;

e.g. MLOsMetaDB reorganizes and harmonizes data from PhaSePro, LLPSDB, PhaSepDB, and DrLLPS to establish a resource encompassing both LLPS- and MLO-related entries. More recent database updates, exemplified by advanced versions of PhaSepDB, incorporate artificial intelligence-assisted literature mining via agent-based pipelines that automatically extract information on phase-separation processes from the literature, followed by manual cross-validation, to systematically identify candidate LLPS-related entities. While these approaches enhance data volume and update efficiency, they also introduce variability in evidence confidence and annotation granularity.

From the perspective of database iteration, LLPS data resources have undergone a clear methodological transition—from predominantly manual, expert-driven collection to AI-assisted organization—and a parallel rebalancing of content priorities: early releases were often fragmented and partial, subsequent updates pursued broader coverage by incorporating high-throughput MLO-related entries (and, in some resources, predictor-linked annotations), whereas the most recent releases increasingly re-emphasize evidence stratification and the re-curation of experimentally validated phase-separation records to improve reliability and downstream usability. This trajectory is illustrated by PhaSepDB, which initially compiled 2914 non-redundant LLPS-related proteins (v1.0), expanded in v2.1 to include 1419 PS entries plus 8073 MLO entries (9492 total entries), and most recently released v3.0 with 3484 expert-curated PS entries enabled by an LLM-based agentic extraction workflow followed by expert verification [65, 66, 69]. Collectively, current LLPS databases balance competing demands for data quality, coverage, and scalability. However, inconsistent treatment of experimental evidence types and limited standardization of confidence levels pose challenges for computational modeling, underscoring the need for clearer evidence hierarchies and more transparent curation frameworks.

Redundancy and fragmentation across liquid–liquid phase separation databases

The rapid proliferation of LLPS-related databases has resulted in substantial redundancy and fragmentation across resources, posing challenges for data integration and comparative analysis. Cross-database comparisons based on UniProt identifiers (Fig. 3a) reveal that only a limited core set of LLPS-associated proteins is consistently shared among multiple databases, whereas a large fraction of entries appears in only one or a small subset of resources. This pattern reflects divergent curation criteria, evidence thresholds, and expansion strategies rather than fundamental disagreement regarding LLPS biology. Interestingly, analyses at the literature level reveal a contrasting picture. When overlap is assessed using PubMed identifiers (Fig. 3b), many LLPS databases draw upon a largely shared body of foundational studies. However, these studies are abstracted, filtered, and annotated in different ways, resulting in distinct representations of LLPS knowledge. This discrepancy indicates that fragmentation arises less from differences in experimental evidence itself than from differences in how evidence is interpreted and structured within databases.

Such fragmentation has important implications for computational research. Predictive models trained on data from a single database may inadvertently learn database-specific biases, limiting their generalizability across datasets [79, 80]. Inconsistent inclusion criteria across databases further complicate model development: a protein curated as LLPS-associated in one resource may be missing from others or assigned to a different LLPS role, such as a driver or a

client, creating divergent training sets. As a result, models trained on resource-specific datasets are harder to compare fairly. Moreover, heterogeneous annotation granularity across databases constrains integrative analysis. Some resources provide residue-level annotations or detailed experimental metadata, whereas others offer only protein-level labels. As a result, constructing unified training or evaluation datasets often requires extensive manual harmonization, introducing additional uncertainty and limiting scalability.

Together, redundancy and fragmentation across LLPS databases highlight a structural challenge that extends beyond data volume. While diversity of resources reflects the dynamic growth of the field, the lack of standardized abstraction and comparability constrains both algorithm development and biological interpretation. In practice, this limits reproducibility across studies, complicates cross-database benchmarking, and increases the risk that models capture database-specific labeling conventions rather than transferable biological signals. More fundamentally, when annotations remain anchored to heterogeneous molecule-level entries with incomplete context and provenance, it becomes difficult to assemble coherent, system-level views of condensates or to trace how evidence supports a given label. These limitations point toward the need for higher-level data representations that can reconcile heterogeneous annotations within a coherent framework.

Beyond molecule-centric databases: unmet needs for event-level representation

Taken together, the characteristics of current LLPS databases reveal a fundamental structural limitation rooted in molecule-centric abstraction. Across protein-, RNA-, condensate- and disease-oriented resources, individual biomolecules typically serve as the primary units of annotation, while cooperative interactions, contextual dependencies, and collective outcomes are represented only implicitly or not at all. Although this abstraction has been effective for early discovery and large-scale screening, it increasingly constrains the biological fidelity of LLPS data representation. As LLPS research moves toward mechanistic and system-level questions, the lack of explicit links between molecules, conditions, and phenotypic consequences becomes a primary bottleneck for synthesis and reuse.

LLPS can be conceptualized as an event-level phenomenon that arises from coordinated interactions among multiple molecular components under specific physicochemical and cellular conditions. Critical dimensions of phase separation events—including component composition, interaction stoichiometry, environmental context, and functional outcomes—are often fragmented across separate database entries or absent altogether. As a result, biologically distinct LLPS events may be collapsed into identical molecular annotations, obscuring conditional behaviors and limiting interpretability. For example, the same protein may participate as a scaffold, client, or regulator depending on partners and cellular state, yet such role switching is rarely captured by molecule-level records. Existing databases largely capture static molecular and experimental snapshots, whereas LLPS is a dynamic process. Consequently, current datasets generally lack systematic curation and representation of temporal dynamics and process-level changes.

These limitations may have important implications for computational modeling. When training data are organized solely around individual molecules, predictive tasks are naturally framed in terms of intrinsic propensities rather than conditional event outcomes. Even

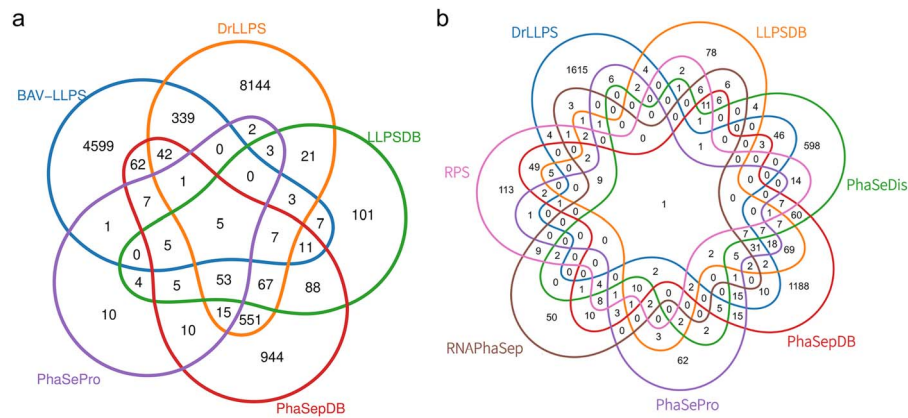


Figure 3 Overlap analysis of phase-separation databases using UniProt-based and PMID-based comparisons. The Venn diagrams illustrate the overlap of LLPS-related resources at two levels using the latest released versions of all databases included in the analysis. Panel a shows the intersections among five phase-separation databases based on UniProt protein identifiers, representing overlap at the protein level. Panel b shows the intersections among nine databases based on PubMed identifiers (PMIDs), reflecting overlap at the literature level; for this analysis, RPS was restricted to entries from Results of reviewed LLPS-associated RNAs (excluding Results of high-throughput LLPS enrichment LLPS-associated RNAs, Results of high-throughput LLPS perturbation LLPS-associated RNAs and Results of predicted LLPS-associated RNAs), and LLPSDB was limited to the Phase_separation_unambiguous dataset within Unambiguous System (excluding entries from ambiguous systems and non-phase-separating datasets). In each panel, the intersecting areas indicate the number of shared entries across the corresponding databases.

highly expressive models are therefore restricted by data abstractions that do not reflect the organization of the underlying biological process. This mismatch helps explain why improvements in predictive accuracy do not always translate into mechanistic insight or robust generalization across contexts. In particular, models may overfit to dataset-specific labeling conventions or proxy features, such as disorder content, that correlate with curation choices but do not capture event-level determinants or mechanistic drivers of phase behavior.

Addressing these challenges requires a shift toward event-centric LLPS data models that explicitly represent multi-component assemblies, contextual variables, and observable outcomes within unified structures. Such representations would enable more faithful integration of heterogeneous evidence and, importantly, reshape how computational problems are formulated. When LLPS is encoded at the level of structured events rather than isolated molecules, predictive modeling naturally shifts toward reasoning about conditional behaviors, interaction dependencies, and context-specific outcomes, creating a more appropriate foundation for advancing LLPS computational analysis. Practically, this also facilitates interoperable benchmarking by enabling consistent definitions of positives/negatives at the event level and supporting cross-database harmonization without discarding contextual detail.

Computational modelling for liquid–liquid phase separation: tasks, modelling paradigms and emerging directions landscape of liquid–liquid phase separation predictive modelling

Computational modelling has become an indispensable component of LLPS research, enabling systematic analysis and large-scale screening beyond the reach of experimental approaches alone [81, 82]. Over the past decade, a growing number of predictive models have been developed to infer LLPS-related properties from molecular data, reflecting both the increasing availability of curated datasets and advances in machine learning and artificial intelligence [83, 84]. Our survey suggests a rapidly expanding yet uneven methodological

landscape, largely shaped by task formulation, data availability, and modelling paradigms.

From a task-oriented perspective, current LLPS prediction models can be broadly grouped into four categories: (i) identification of phase-separating proteins, (ii) prediction of LLPS-driving regions or sequence determinants, (iii) assessment of mutation effects on phase separation behavior, and (iv) classification of LLPS types, roles, or dependency patterns, such as scaffold–client relationships. Protein-level LLPS identification dominates the field, whereas mutation-oriented and type/classification tasks remain comparatively under-explored, likely reflecting both limited high-quality labelled data and increased biological complexity (Table 2, Fig. 4b). Methodologically, the distribution of models across machine learning, deep learning and hybrid paradigms further illustrates how modelling choices have evolved in parallel with dataset growth (Fig. 4a). A complete catalogue of the models included in this review is provided in the Supplementary file. Notably, despite the breadth of protein-centric predictors, dedicated models for LLPS-related RNA entities or RNA-driven phase-separation behavior remain scarce, highlighting an important opportunity to incorporate protein–RNA co-regulation into future predictive frameworks.

From a methodological standpoint, LLPS predictive modelling has followed a clear trajectory. Early studies predominantly relied on handcrafted physicochemical and sequence-derived features coupled with classical machine learning algorithms, such as support vector machines and random forests [132–139]. As larger datasets became available, deep learning architectures—including convolutional and recurrent neural networks [140–143]—were increasingly adopted to capture higher-order sequence patterns. More recently, Transformer-based protein language models have emerged as powerful representation learners, often used as embedding backbones combined with lightweight classifiers [144–148]. Many recent studies further adopt hybrid frameworks that integrate deep representation learning with traditional machine learning classifiers, reflecting a pragmatic response to limited training data and label noise (Table 3). For the purpose of the methodological taxonomy used in Table 3, the “Deep

Table 2 Categorization of LLPS prediction tasks, number of studies, and representative models.

Task type	Number	Models
Task I. LLPS protein identification	31	catGRANULE [85], FLFB [86], LLPhyScore [79], PSPer [87], PSPire [88], MolPhase [89], Pscore [90], PICNIC [91], PSPHunter [92], PSAP [93], PSPredictor [94], FuzDrop [95], GP-GNN [96], Dropller [97], MambaPhase [98], snLLPS [99], PredLLPS_PSSM [100], PSTP [101], Opt_PredLLPS [102], catGRANULE 2.0 ROBOT [103], DeePhase [104], ML_LM [105], Ka Yin Chin <i>et al.</i> (BMC Bioinformatics, 2024) [106], Soren von Bulow <i>et al.</i> (PNAS,2025) [107], Wesley W. Oliver <i>et al.</i> (The Journal of Physical Chemistry B, 2025) [108], Ashwin Lahorkar <i>et al.</i> (IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022) [109], Pratik Mullick <i>et al.</i> (Biomolecules, 2022) [110], Qinglan Ma <i>et al.</i> (Life (Basel), 2023) [111], Zahoor Ahmed <i>et al.</i> (International Journal of Biological Macromolecules, 2024) [112], Jiyan Wang <i>et al.</i> (International Journal of Biological Macromolecules, 2023) [113], Wenbin Li <i>et al.</i> (Briefings in Bioinformatics, 2025) [114] ParSe [115], dSCOPE [116], ParSe 2.0 [117], PLAAC [118], IFF [119], TIDGN [120], IDR-Puncta ML Model [121], PhaSeMotif [122], Yumeng Zhang <i>et al.</i> (Elife, 2025) [123]
Task II. LLPS-driving regions and sequence determinants	9	PSMutPred [124], PhosLLPS [125]
Task III. Predicting mutation effects on LLPS	2	
Task IV. LLPS types, roles, and dependency	6	PULPS [126], PhaSePred [127], Seq2Phase [128], PhaseNet [129], Zahoor Ahmed <i>et al.</i> (Proteomics, 2024) [130], Zahoor Ahmed <i>et al.</i> (Briefings in Bioinformatics, 2025) [131]

Notes: for studies in which the model name is not explicitly specified in the manuscript, we consistently designate the model using the format “Author (Journal, Year)”, the same convention applies hereafter.

Learning” category is defined broadly to include both conventional neural-network-based architectures and transformer-based protein language/foundation model architectures.

Data dependency constitutes a central constraint shaping the current modelling landscape. Across the surveyed literature, training and evaluation datasets repeatedly draw from a small number of widely used sources, leading to substantial dataset reuse across studies. In practice, the datasets can be grouped into three broad origins: self-constructed datasets, publicly released LLPS-related databases and “other” auxiliary datasets; their usage patterns are consistent across Fig. 4c–e. Importantly, a portion of the “other datasets” are themselves inherited from earlier model-development efforts—e.g. PICNIC [91], PSTP [101], Opt_PredLLPS [102], PhaSeMotif [122], and PhaseNet [129] adopt the dataset released with PhaSePred [127] for training and/or testing—creating a cascade of dataset reuse that can propagate upstream curation choices and reduce independence across evaluations. Evaluation protocols also vary considerably, ranging from random splits to more stringent sequence-similarity-aware partitioning, which complicates direct comparison of reported performances. Together, these data-related factors influence not only model accuracy but also generalizability and reproducibility, underscoring the tight coupling between data resource design and algorithmic development in LLPS research.

Task I: Liquid–liquid phase separation protein identification

The identification of phase-separating proteins represents the earliest and most extensively explored task in computational LLPS research [85, 86]. Given a protein sequence, this task aims to predict whether the protein has the intrinsic potential to undergo LLPS [88, 90]. Historically, this formulation has dominated the field, not necessarily because it best captures the biological nature of LLPS, but because it aligns well

with available data and modelling conventions [96, 98, 99]. Early LLPS annotations were predominantly protein-centric, and the resulting labels could be readily cast as a binary classification problem, making this task particularly amenable to computational treatment [100–105]. As a consequence, the majority of existing LLPS predictors focus on protein-level propensity estimation [107, 108].

Despite the large number of published models, most approaches for LLPS protein identification can be broadly grouped into a small number of recurring modelling strategies. One prominent line of work is grounded in explicit biophysical assumptions, leveraging features related to intrinsic disorder, low-complexity regions, hydrophobicity, and specific intermolecular interactions such as π – π stacking. Representative models in this category, including LLPhyScore [79], MolPhase [89], and FuzDrop [95], explicitly encode physicochemical principles believed to stabilize biomolecular condensates. These methods often offer good interpretability and perform well for proteins conforming to known LLPS mechanisms, but their reliance on predefined interaction rules may limit generalizability to newly emerging or atypical phase separation behaviors.

A second major strategy adopts a more data-driven perspective, focusing on sequence-derived statistical patterns with minimal prior assumptions. Models such as PSPredictor [94], PICNIC [91], PSAP [93], and PSPHunter [92] exploit amino acid composition, residue-level features or sequence embeddings to discriminate LLPS proteins from non-phase-separating counterparts. In practice, many of these predictors are implemented with classical machine-learning backbones, such as random forests [149], support vector machines [150], and gradient-boosting methods [151, 152], reflecting a pragmatic balance between performance and limited labelled data. By emphasizing pattern recognition over mechanistic interpretation, these approaches can capture subtle sequence signatures and achieve robust performance across diverse datasets [109, 110]. However, their predictive power is tightly coupled to training data composition, and they are

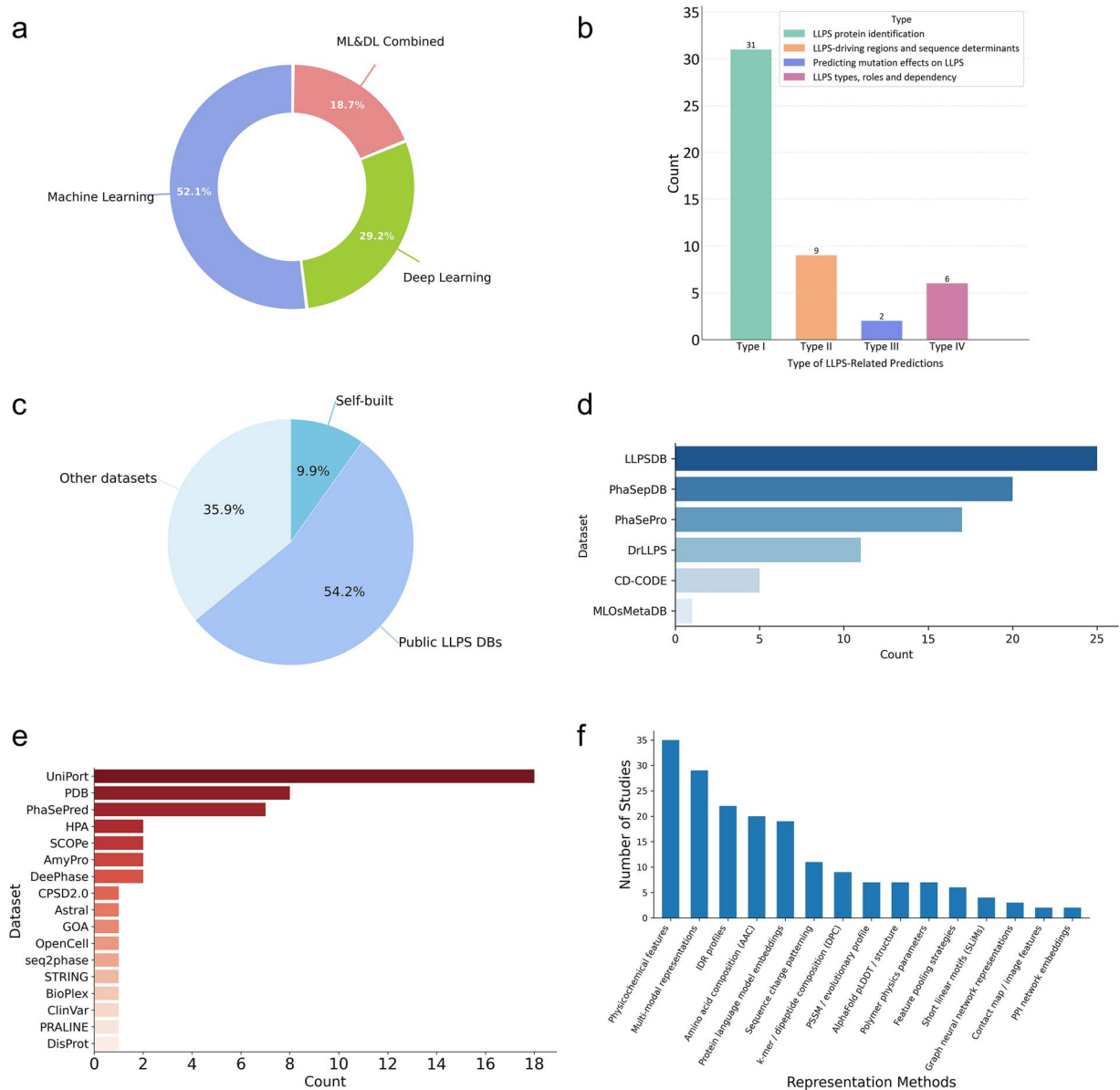


Figure 4 Comprehensive methodological characterization of LLPS protein prediction models with respect to algorithmic paradigms, predictive tasks, dataset utilization, and feature representation strategies. Panels a–f summarize the methodological characteristics of 48 representative LLPS prediction models. Panel a illustrates the proportional distribution of models according to their algorithmic paradigm, classified into machine learning-based, deep learning-based, and hybrid approaches. Panel b categorizes the models based on their primary predictive tasks. Panel c presents the distribution of dataset sources used for model development and evaluation, grouped into three categories: publicly available LLPS databases, self-constructed datasets, and other external resources. Panels d and e further quantify the frequency of dataset utilization, distinguishing between the usage of LLPS-specific datasets and other datasets. Panel f provides a summary of feature representation strategies, showing the number of studies employing each type of feature descriptor.

susceptible to dataset bias and limited biological interpretability [111–113].

A third, more recent direction attempts to move beyond purely protein-centric representations by incorporating contextual information relevant to phase separation. Examples include predictors that explicitly model protein–RNA interactions, protein–protein interaction networks, or experimental conditions, such as PSPer [87], Droptler [97], and the RNAPSEC team’s model [106]. By introducing interaction partners or environmental variables into the modelling framework, these approaches begin to approximate LLPS as a conditional and multi-component process rather than an isolated protein

property. Although still relatively few in number, such models highlight the limitations of simplified propensity-based predictions and point toward more biologically faithful formulations.

Advances in deep learning and protein language models have further enhanced LLPS protein identification by providing powerful sequence representations. Architectures based on convolutional, graph-based, or attention mechanisms, as well as Transformer-derived embeddings, have consistently improved prediction accuracy [153, 154]. Recent studies also increasingly combine deep representation learning with classical classifiers in hybrid pipelines, aiming to stabilize training under small, noisy datasets while retaining

Table 3 Categorization of LLPS prediction models by methodological framework and task coverage.

Technology scope	Algorithm	Models	Task type
Machine Learning	Random Forest	FLFB [86], catGRANULE 2.0 ROBOT [103], PSMutPred [124], dSCOPE [116], DeePhase [104], ML_LM [105], PSPHunter [92], PSAP [93], PSTP [101], Seq2Phase [128], PhaseNet [129], Ka Yin Chin <i>et al.</i> (BMC Bioinformatics, 2024) [106], Ashwin Lahorkar <i>et al.</i> (IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022) [109], Qinglan Ma <i>et al.</i> (Life (Basel), 2023) [111], Zahoor Ahmed <i>et al.</i> (Proteomics, 2024) [130]	I / II / III / IV
	XGBoost / CatBoost / GBDT / AdaBoost	Opt_PredLLPS [102], PSPIre [88], MolPhase [89], PICNIC [91], PSPredictor [94], PhaSePred [127], Seq2Phase [128], PhaseNet [129], Ka Yin Chin <i>et al.</i> (BMC Bioinformatics, 2024) [106], Qinglan Ma <i>et al.</i> (Life (Basel), 2023) [111]	I / IV
	Logistic Regression	PULPS [126], PSMutPred [124], PSTP [101], Seq2Phase [128], FuzDrop [95], Ka Yin Chin <i>et al.</i> (BMC Bioinformatics, 2024) [106]	I / III / IV
	SVM / SVR	PSMutPred [124], Seq2Phase [128], Ka Yin Chin <i>et al.</i> (BMC Bioinformatics, 2024) [106], Ashwin Lahorkar <i>et al.</i> (IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022) [109]	I / III / IV
Deep Learning	HMM	PSPer [87], PLAAC [118]	I / II
	Stacked Ensemble	IDR-Puncta ML model [121]	II
	MLP	catGRANULE 2.0 ROBOT [103], Seq2Phase [128], PhosLLPS [125], PhaseNet [129], Soren von Bulow <i>et al.</i> (PNAS, 2025) [107], Wesley W. Oliver <i>et al.</i> (The Journal of Physical Chemistry B, 2025) [108], Zahoor Ahmed <i>et al.</i> (Briefings in Bioinformatics, 2025) [131]	I / III / IV
	Attention/Transformer / Mamba	IFF [119], Dropller [97], ML_LM [105], PSTP [101], MambaPhase [98], PhaSeMotif [122], PhaseNet [129], Wenbin Li <i>et al.</i> (Briefings in Bioinformatics, 2025) [114], Yumeng Zhang <i>et al.</i> (Elife, 2025) [123]	I / II / IV
	GNN/KNN	TIDGN [120], PhosLLPS [125], Ka Yin Chin <i>et al.</i> (BMC Bioinformatics, 2024) [106]	I / II / III
Others	CNN	Opt_PredLLPS [102], PredLLPS_PSSM [100], PhaSeMotif [122], Zahoor Ahmed <i>et al.</i> (International Journal of Biological Macromolecules, 2024) [142]	I / II
	RNN / LSTM / GRU	Opt_PredLLPS [102], Dropller [97], PredLLPS_PSSM [100]	I
	Contrastive Learning / Siamese Network	IFF [119], MambaPhase [98], snLLPS [99]	I / II
	Rule-based / Heuristic	ParSe [88], ParSe 2.0 [117], Jiyan Wang <i>et al.</i> (International Journal of Biological Macromolecules, 2023) [113]	I / II
	Monte Carlo / GA / SBO	catGRANULE [85], GP-GNN [96], LLPhyScore [79], Pscore [90], Pratik Mullick <i>et al.</i> (Biomolecules, 2022) [110]	I

strong feature expressivity. A ProtT5-based predictor integrating protein language model embeddings with KmerConv and multi-head attention has also been reported for LLPS protein identification, with evaluation performed on image-filtered negative samples across multiple independent datasets [114]. Importantly, these developments have primarily strengthened feature representation rather than redefining the task itself. Even highly accurate models remain constrained by the protein-level abstraction, answering whether a protein “resembles” known LLPS proteins rather than predicting when, where and, with which partners phase separation occurs. This inherent limitation motivates the exploration of alternative task definitions and modelling paradigms, providing the rationale for the event-centric perspective introduced next.

Task II: Liquid–liquid phase separation-driving regions and sequence determinants

While [Task I](#) focuses on identifying whether a protein has the potential to undergo LLPS, a more mechanistically oriented question concerns which specific sequence regions within a protein drive this behavior. [Task II](#) addresses this problem by aiming to localize LLPS-promoting segments and uncover sequence determinants that contribute to

condensate formation. This task reflects an important conceptual shift: rather than treating LLPS as a global protein property, it recognizes phase separation as an emergent outcome of localized sequence features and interaction motifs [118, 123]. Motivated by the central role of low-complexity regions, IDRs and prion-like domains in driving or modulating LLPS, a small set of studies has specifically focused on predicting LLPS-driving segments at the region level; to date, nine dedicated predictors in this category have been reported.

Early approaches to identifying LLPS-driving regions were closely linked to established biochemical insights, particularly the enrichment of IDRs, low-complexity domains and prion-like sequences in phase-separating proteins. Methods such as ParSe [115, 117] and dSCOPE [116] employed sliding-window strategies combined with physicochemical descriptors to score local sequence propensity for phase separation. These approaches offered interpretable connections between sequence composition and LLPS behavior, highlighting the contributions of charge patterning, aromatic residues, and disorder-associated features.

Subsequent models adopted more data-driven strategies to capture higher-order sequence patterns beyond predefined feature sets. By leveraging residue-level embeddings and deep learning architectures, tools such as IFF [119] and related predictors learned discriminative

representations of LLPS-driving segments directly from annotated data. Attention-based mechanisms further enabled the identification of long-range dependencies and cooperative effects between distant residues, which are difficult to encode using handcrafted features alone. More recent work has also begun to refine this task from broad region localization toward motif-level characterization. For instance, PhaSeMotif focuses on essential motifs within phase-separating IDRs, whereas protein language model-based analyses have suggested that LLPS-relevant determinants may also appear as conserved continuous motifs containing both sticker-like and spacer-like residues [122]. Together, these methods improved localization accuracy and expanded the repertoire of detectable LLPS-associated sequence motifs. In addition, transfer-learning-based frameworks such as TIDGN [120] have been proposed to predict interaction/binding sites in intrinsically disordered proteins, providing complementary region-level signals that can be relevant to LLPS-driving or partner-dependent segments.

Despite these advances, predicting LLPS-driving regions remains inherently challenging. Boundaries between driving and non-driving segments are often diffuse, context-dependent, and sensitive to experimental conditions. Moreover, the functional impact of a given region frequently depends on its interaction with other biomolecules and its conformational dynamics, factors that are not fully captured by sequence information alone. Importantly, region-level predictions naturally lead to questions about how local sequence alterations influence phase separation behavior. In a related direction, the IDR-Puncta ML model [121] considered the condensate-forming behavior of individual IDRs in cells, and subsequent proteome-wide analyses linked predicted condensate-forming IDRs to proteins enriched in RNA processing and splicing functions. This consideration provides a direct bridge to [Task III](#), where computational models aim to assess the effects of specific mutations on LLPS propensity and condensate properties.

Task III: Predicting mutation effects on liquid–liquid phase separation

Building upon region-level identification of LLPS-driving sequences, [Task III](#) focuses on predicting how specific sequence variations modulate phase separation behavior. Rather than asking whether a protein or region can promote LLPS, this task addresses a more mechanistic and functionally relevant question: how mutations alter condensate formation, stability, or material properties. This formulation is particularly important for linking LLPS to genotype–phenotype relationships and disease-associated variants. In practice, mutation-focused predictors typically consider single-point or multi-point substitutions within a protein sequence and aim to infer how such perturbations change LLPS ability and behavior; however, systematic modelling studies dedicated to mutation effects remain relatively scarce compared with protein-level identification tasks, despite their importance for pinpointing regulatory sequence features and elucidating molecular mechanisms.

Early observations revealed that point mutations affecting charge distribution, aromatic residues, or disorder-promoting motifs can profoundly influence phase separation propensity. Computational models developed for this task aim to quantify such effects by comparing wild-type and mutant sequences, often integrating sequence context and physicochemical perturbations. Representative approaches, such as PSMutPred [124], frame mutation impact prediction as a

supervised learning problem, leveraging curated mutation datasets to estimate whether a given variant enhances, disrupts or qualitatively alters LLPS behavior. A further development in this area is the incorporation of post-translational regulation at the residue level. PhosLLPS [125] predicts phosphorylation sites that regulate LLPS by combining protein language model embeddings with a graph-based deep learning framework, while the accompanying PTMPaSe resource curates experimentally supported PTM sites with regulatory effects on LLPS. This is particularly relevant because residue-level perturbations in LLPS are not limited to sequence substitutions, but can also arise from reversible modifications that promote or inhibit condensate formation. These models enable systematic screening of variants that may not abolish phase separation entirely but subtly shift condensate dynamics.

The predictive scope of mutation-focused models extends beyond binary outcomes. In some cases, mutations may convert scaffold-like proteins into client-like roles, alter interaction specificity or bias condensates toward aberrant material states. Consequently, mutation effect prediction provides a direct computational route to understanding how LLPS dysregulation contributes to disease mechanisms, particularly in neurodegenerative disorders and cancer, where pathogenic mutations often cluster within intrinsically disordered or low-complexity regions. In addition to dedicated mutation-effect predictors, some region-oriented frameworks, such as PSPHunter [92], provide residue-level signals that can be leveraged to interpret how local substitutions may perturb LLPS-relevant sequence determinants.

Despite their promise, mutation-based predictors face substantial challenges. Available training data remain sparse and heterogeneous, with experimental measurements varying widely in assay type and resolution. Moreover, mutation effects are frequently context-dependent, influenced by interaction partners, expression levels, and cellular conditions. These limitations highlight the need to integrate mutation-level predictions with higher-order information about molecular roles and interaction patterns. While [Task III](#) focuses on quantifying variant-induced effects on phase separation, [Task IV](#) extends the analysis by classifying LLPS types and functional roles, thereby providing a higher-level interpretative framework for these perturbations.

Task IV: Liquid–liquid phase separation types, roles, and dependency

Beyond predicting LLPS propensity or mutation-induced perturbations, a more biologically informative task is to characterize how proteins participate in phase separation. [Task IV](#) addresses this challenge by aiming to classify LLPS types, functional roles, and dependency patterns, such as distinguishing scaffold and client proteins or identifying partner- and condition-dependent phase separation behaviors. This task moves closer to a systems-level understanding of condensate organization, bridging molecular properties with functional outcomes.

Several computational approaches have been proposed to infer LLPS roles based on sequence features, interaction patterns, or learned representations. Models such as Seq2Phase [128] and PhaSePred [127] classify proteins according to their functional involvement in condensates, exploiting differences in sequence composition, disorder content, and interaction propensity between scaffolds and clients. These methods provide a structured interpretation of LLPS participation, enabling predictions that extend beyond binary

phase separation labels [126]. By explicitly modelling functional categories, role-based predictors offer greater explanatory power for interpreting experimental observations and variant effects. More recent studies have further expanded this task by incorporating dependency relationships and contextual variables. Some predictors aim to distinguish spontaneous LLPS from partner-dependent or condition-specific phase separation, recognizing that many proteins only form condensates in the presence of specific interaction partners or under particular environmental conditions [131]. A similar tendency toward finer categorization is reflected in PhaseNet [129], which adopts a dual-task framework in which LLPS proteins are first separated from non-LLPS proteins and then further classified into self-assembling and partner-dependent categories. By combining protein language model embeddings with sequence-derived features for the first stage and an ensemble-based classifier for the second, it provides a more explicit operationalization of dependency patterns within LLPS prediction. By integrating protein–protein or protein–RNA interactions, as well as experimental context, these approaches begin to model LLPS as a conditional process rather than an intrinsic protein attribute.

Despite their conceptual significance, role- and type-based predictors remain relatively scarce. This limitation primarily reflects the lack of comprehensive, standardized annotations for LLPS roles and dependency patterns [130]. Moreover, functional roles are often dynamic, varying across cellular states and molecular compositions, posing additional challenges for static sequence-based models. Nevertheless, Task IV represents a critical step toward biologically faithful LLPS modelling. By framing phase separation in terms of roles, interactions and dependencies, it bridges molecular-level prediction and event-level representation, enabling the cross-cutting modelling paradigms and integrative frameworks outlined below.

Cross-cutting modelling paradigms: from feature engineering to foundation representations

Across the diverse prediction tasks discussed above, a unifying theme in LLPS computational research is the continuous evolution of modelling paradigms driven by advances in protein representation. Rather than changes in classifier architecture alone, it is the transformation of how protein sequences are encoded that has fundamentally reshaped LLPS prediction strategies [101]. This evolution reflects a broader shift from explicit, hypothesis-driven feature engineering toward implicit, data-driven representation learning. Consistent with this trend, Fig. 4f shows the feature representation strategies used in the surveyed studies by reporting the number of models adopting each type of feature descriptor.

Early LLPS predictors were largely built upon handcrafted features derived from physicochemical principles and sequence statistics. Amino acid composition, charge patterning, intrinsic disorder, low-complexity regions, and interaction-related descriptors such as π – π stacking propensity were explicitly encoded to capture known drivers of phase separation [90]. These representations offered strong interpretability and were well aligned with prevailing biophysical models of LLPS. However, they also relied on predefined assumptions about sequence grammar and interaction mechanisms, limiting their ability to generalize across heterogeneous LLPS behaviors.

With the adoption of deep learning, representation learning emerged as a central component of LLPS modelling. Convolutional, recurrent, and attention-based architectures enabled models to learn task-specific embeddings directly from sequence data, reducing

reliance on manual feature design [98]. More recently, the introduction of protein language models based on Transformer architectures has further decoupled representation learning from downstream prediction tasks. Pretrained on large-scale protein sequence corpora, these foundation models provide rich, context-aware embeddings that capture evolutionary, structural, and functional information [105]. In LLPS prediction, such embeddings are frequently combined with lightweight classifiers rather than end-to-end fine-tuning, reflecting both practical constraints and data characteristics.

Notably, hybrid modelling frameworks that integrate deep representations with classical machine learning classifiers have become particularly prevalent in LLPS research. This trend is not merely transitional but can be viewed as a rational response to current data limitations in the field. In many existing LLPS datasets, annotations remain relatively sparse, heterogeneous, and often noisy, with positive and negative labels defined under diverse experimental conditions. Under these dataset constraints, fully end-to-end deep models can be more prone to overfitting, whereas classical machine learning models benefit from robust optimization on limited datasets. Hybrid approaches leverage the expressive power of pretrained representations while maintaining stability and generalizability through simpler classifiers.

Despite substantial improvements in representation quality, an important mismatch persists between modelling capacity and task definition. Most LLPS predictors continue to operate under simplified, protein-centric formulations, using increasingly sophisticated embeddings to estimate intrinsic phase separation propensity. As a result, enhanced representations primarily improve discrimination within existing task boundaries rather than redefining what is being predicted. This limitation underscores that representation advances alone are insufficient to capture the conditional, multi-component and context-dependent nature of LLPS.

Finally, the evolution of modelling paradigms cannot be separated from concurrent advances in LLPS data resources. The growth of curated databases, expansion of sequence annotations and increasing integration of interaction and contextual information have directly enabled the adoption of foundation models and more expressive representations. Together, these trends highlight the co-evolution of LLPS data resources and algorithms, and motivate a shift toward event-level, multimodal modelling frameworks, as described below.

Toward event-level and multimodal modelling of liquid–liquid phase separation

Despite substantial progress in LLPS prediction driven by improved representations and modelling strategies, a fundamental limitation persists across most existing approaches: the abstraction of phase separation as an intrinsic, protein-centric property. As discussed in previous sections, current predictive tasks—ranging from LLPS protein identification to region localization, mutation effect assessment and role classification—largely operate on isolated molecular entities. While these formulations have enabled methodological advances, they only partially reflect the biological reality of LLPS as a conditional, multi-component, and context-dependent process.

At the biological level, LLPS manifests as discrete events arising from cooperative interactions among multiple biomolecules, including proteins, nucleic acids and, in some cases, metabolites or polysaccharides, under specific physicochemical and cellular conditions [155]. Event outcomes are shaped not only by intrinsic sequence features but also by interaction partners, stoichiometry,

environmental parameters, and spatiotemporal context [156]. Protein-centric predictors, even when powered by advanced representations, are inherently limited in their ability to capture such emergent behavior [157]. As a result, high predictive accuracy at the sequence level does not necessarily translate into mechanistic insight or functional interpretability.

These limitations point to the need for a paradigm shift from molecule-level prediction toward event-level modelling of LLPS. In this framework, the prediction target is no longer a single protein's propensity to phase separate, but a structured LLPS event defined by participating components, interaction networks, contextual variables, and functional outcomes. Such a formulation aligns more closely with experimental observations and enables the integration of diverse data modalities, including protein–protein and protein–RNA interactions, expression dynamics, cellular localization, and environmental perturbations.

Recent advances in representation learning and foundation models provide a timely opportunity to support this transition [158]. Multimodal architectures capable of integrating sequence embeddings, interaction graphs, structural proxies, and contextual metadata offer a computational substrate for modelling LLPS events in a holistic manner [159]. Importantly, this shift necessitates corresponding changes in data resources, annotation standards, and evaluation protocols. Event-centric datasets with explicit representation of component composition and experimental context will be essential for training and validating next-generation models.

Ultimately, moving toward event-level and multimodal LLPS modelling represents more than a technical refinement; it reflects a conceptual realignment of computational objectives with biological complexity. By redefining prediction targets and embracing integrative data representations, future computational frameworks may not only improve predictive performance but also enable deeper mechanistic understanding of LLPS and its roles in physiology and disease.

Event-centric frameworks for integrative liquid–liquid phase separation modelling and applications

Redefining liquid–liquid phase separation as a computational event

Most existing computational studies conceptualize LLPS as an intrinsic property of individual biomolecules, predominantly proteins [81]. Under this molecule-centric paradigm, LLPS is formulated as a binary or probabilistic attribute inferred from sequence-derived features, with predictive efforts focused on identifying phase-separating proteins, their driving regions or mutation-sensitive residues [160]. While such formulations have enabled methodological advances and large-scale screening, they only partially capture the biological nature of phase separation.

From a biological perspective, LLPS does not arise from isolated molecular entities but emerges as a collective phenomenon driven by cooperative interactions among multiple components under specific conditions. Proteins, RNAs and, in some contexts, polysaccharides or metabolites jointly contribute to condensate formation, with outcomes shaped by stoichiometry, interaction networks, physicochemical environments, and spatiotemporal context [161–163]. The same molecule may participate in distinct phase separation behaviors depending on its partners, expression level or cellular state [164, 165]. Consequently,

treating LLPS as a static, molecule-intrinsic attribute obscures the conditional and emergent properties that define phase separation *in vivo*.

This mismatch highlights a fundamental limitation in current computational formulations: prediction targets are misaligned with biological reality. Even highly expressive models and advanced sequence representations remain constrained when tasked with predicting an inherently context-dependent process using molecule-level abstractions [40]. Improved accuracy under simplified task definitions does not necessarily translate into mechanistic understanding, nor does it readily support functional interpretation or translational applications.

To address this gap, we propose that LLPS should be redefined, at the computational level, as a structured biological event rather than a molecular property. More specifically, an LLPS event may be defined as a context-bounded phase-separation process in which a particular set of molecular components, linked by interaction relationships and relative abundance constraints, gives rise under defined physicochemical and cellular conditions to an observable condensate state and associated functional consequences. In this framework, the term “event” does not merely denote the static existence of a condensate as an object; rather, it refers to a specific assembly–condition–behavior instance of phase separation. Accordingly, the minimal descriptive unit is not a single protein label but a component–condition–outcome unit that explicitly records who participates, how they are related, under what context phase separation occurs, and what material, functional or phenotypic outcome is observed [166]. This redefinition shifts the computational objective from estimating intrinsic propensities to modelling conditional behaviors arising from multi-component systems.

This definition is consistent with broader views of biomolecular condensates as multicomponent, non-stoichiometric, and condition-sensitive assemblies whose composition, material properties, and functions vary across spatial and temporal scales [11, 167]. From a computational perspective, it also provides a tractable abstraction for integrating heterogeneous evidence and formulating predictive tasks that move beyond intrinsic propensity toward conditional, event-level reasoning. Figure 5 outlines key directions enabled by this event-centric perspective, spanning data representation, modelling paradigms and downstream applications. Importantly, such a perspective does not discard molecule-level predictions but embeds them within a higher-order framework that more faithfully reflects the biology of phase separation. By redefining LLPS as a computational event, subsequent data representation strategies, modelling approaches and application scenarios can be coherently aligned with biological complexity, providing a principled foundation for integrative and multimodal LLPS research.

Event-level data representation: integrating multi-molecular and contextual dimensions

Redefining LLPS as a computational event necessitates a corresponding rethinking of data representation strategies. Current LLPS-related data resources are largely organized around individual entities, such as proteins or RNAs, with annotations describing intrinsic properties, sequence features, or experimental evidence of phase separation [168]. While such entity-centric representations have supported molecule-level prediction tasks, they are insufficient for modelling LLPS as a conditional and multi-component process. In an event-centric perspective, the fundamental data unit is no longer a single biomolecule but a structured LLPS event arising from specific

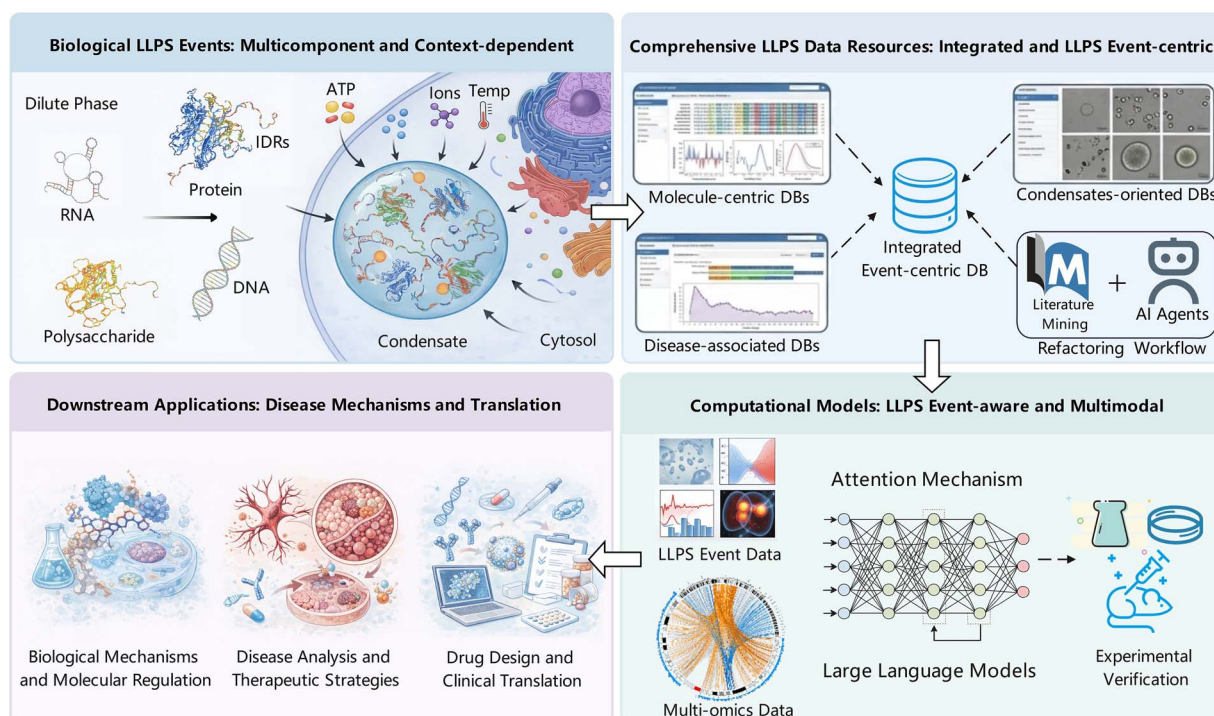


Figure 5 A roadmap for the future of LLPS research. The schematic outlines a four-module roadmap: (i) multicomponent, context-dependent LLPS events and their multi-molecular regulation; (ii) comprehensive, event-centric LLPS data resources that integrate molecule-centric, disease-associated, and condensate-focused databases via text mining and AI-agent-assisted curation; (iii) event-aware, multimodal computational models that combine integrated LLPS event data with multi-omics and large language model-enabled representations, supported by experimental validation; and (iv) downstream applications in biological mechanisms and molecular regulation, disease analysis and therapeutic strategies, and drug design and clinical translation.

combinations of components and contexts. Capturing such events requires data representations that explicitly encode not only what molecules are involved, but also how, under which conditions and with what functional consequences they interact. This shift transforms LLPS data from flat collections of annotated entities into relational and contextualized structures.

An event-level LLPS representation can be conceptualized as a structured schema comprising several interdependent dimensions. First, components define the participating biomolecules, including proteins, RNAs and, where relevant, polysaccharides or metabolites, together with their stoichiometric relationships. Second, interaction layers describe physical or functional associations, such as protein–protein or protein–RNA interactions, interaction strengths and network topology. Third, contextual variables capture experimental and cellular conditions, including *in vitro* versus *in vivo* settings, cellular compartment, stress states, and physicochemical parameters. Finally, outcomes characterize the observable properties of the event, such as condensate type, material state, dynamics, and functional relevance.

Importantly, such structured representations enable explicit modelling of conditionality and variability. The same molecule may participate in multiple LLPS events with distinct partners and outcomes, a reality that cannot be faithfully represented in molecule-centric databases. By contrast, event-level data structures naturally accommodate multiplicity, context dependence, and compositional diversity, aligning computational representations with biological observations. From an implementation perspective, event-level representations are well suited to relational data models and knowledge graph frameworks, where entities, interactions, and contexts can be integrated into a unified structure. This approach not only facilitates data

integration across heterogeneous sources but also provides a foundation for downstream modelling strategies that explicitly reason over multi-component and multimodal information. As such, restructuring LLPS data around events is a critical prerequisite for advancing from descriptive resource compilation to truly integrative and mechanistic computational modelling.

Event-aware and multimodal modelling strategies

The transition from molecule-centric to event-centric LLPS representations has profound implications for computational modelling strategies. When LLPS is framed as a structured event involving multiple components and contextual variables, traditional single-input predictors become inherently inadequate. Models designed to map individual sequences to scalar propensity scores are not equipped to reason over heterogeneous inputs, conditional dependencies, or emergent behaviors arising from component interactions [169]. Event-aware LLPS modelling requires algorithms capable of jointly integrating multiple information streams. At a minimum, such models must accommodate diverse molecular inputs, such as protein and RNA representations, relational information (interaction networks or co-participation patterns) and contextual descriptors (cellular state, experimental conditions, or perturbations). This shift transforms LLPS prediction from a single-task classification problem into a multi-input, multi-factor inference problem, where the outcome depends on coordinated patterns rather than isolated features.

Multimodal modelling frameworks offer a natural solution to these challenges. Architectures that combine sequence-based embeddings, interaction graphs, and contextual metadata enable the learning of

representations that explicitly encode relationships among components. Graph-based models can capture interaction topology and dependency structure, while attention mechanisms facilitate flexible integration of heterogeneous modalities [170–173]. Importantly, multimodal fusion allows models to learn conditional rules [174], such as partner-dependent phase separation or context-specific condensate formation, which are difficult to express under molecule-centric formulations. Large pretrained models further expand the modelling toolkit for event-level LLPS analysis. Protein language models provide rich representations of individual components [146, 148, 175, 176], but their true potential emerges when they are embedded within higher-order frameworks that aggregate multiple component embeddings and contextual signals. In this setting, foundation models act as representation engines rather than end-to-end predictors, supplying semantically meaningful inputs to event-aware architectures. This design aligns well with current data realities, where limited event-level annotations favor modular and hybrid approaches over fully end-to-end training. Multi-omics profiles can further serve as complementary modalities for event-level LLPS modelling, including proteomics, transcriptomics, and genomics [177, 178]. Fused with sequence embeddings, interaction graphs and contextual metadata, these signals encode cellular state and perturbation context, enabling models to learn conditional rules such as cell-type-specific condensate formation and stress- or variant-dependent phase behaviors.

Crucially, event-aware modelling reframes predictive objectives. Instead of asking whether a molecule can phase separately, models aim to infer whether a specific combination of components under defined conditions gives rise to an LLPS event, and to characterize its properties. This perspective supports richer outputs, including probabilistic event occurrence, role assignments, and condition-dependent behavior. By aligning modelling strategies with event-level data structures, computational frameworks can move beyond improved discrimination toward mechanistic reasoning and hypothesis generation, laying the groundwork for meaningful biological and translational insights.

Implications for disease mechanisms and translational applications

Reframing LLPS as an event-level phenomenon has important implications for understanding disease mechanisms and advancing translational applications. Many disease-associated perturbations linked to LLPS cannot be adequately explained by changes in the intrinsic properties of single molecules [179, 180]. Instead, pathological outcomes often arise from altered interactions, disrupted stoichiometry or context-dependent failures of condensate regulation [181]—features that are naturally captured at the level of LLPS events rather than isolated biomolecules.

In neurodegenerative disorders [182], e.g. disease-associated mutations frequently do not abolish phase separation per se but instead shift condensate material properties, alter partner preferences or promote aberrant maturation from liquid-like droplets to pathological aggregates [183, 184]. Event-level modelling enables these effects to be interpreted as transitions between distinct LLPS states driven by specific combinations of molecular components and conditions. By explicitly representing interaction context and event outcomes, computational frameworks can distinguish between benign, functional condensates, and disease-associated aberrant assemblies, providing a more nuanced mechanistic interpretation than molecule-centric

predictors. Similarly, in cancer and other complex diseases, LLPS often acts as a regulatory layer that modulates signaling, transcriptional control, and stress adaptation [185, 186]. Perturbations in expression levels, post-translational modifications or interaction networks can rewire phase separation events without fundamentally altering individual protein sequences. Event-centric representations allow such regulatory rewiring to be modelled as shifts in event composition and context, offering a systems-level perspective on how LLPS contributes to disease progression and cellular heterogeneity.

From a translational standpoint, event-level modelling also reshapes strategies for therapeutic intervention. Traditional drug discovery paradigms focus on targeting individual proteins or disrupting specific binding sites [83]. However, LLPS-mediated functions frequently depend on collective behaviors and mesoscale material properties [187, 188]. Event-aware frameworks enable the exploration of intervention strategies aimed at modulating event composition, interaction strength, or environmental conditions, rather than simply inhibiting single molecular targets. This perspective aligns with emerging efforts to regulate condensate dynamics, stability, or reversibility as therapeutic endpoints. Beyond disease, event-centric LLPS modelling has implications for synthetic biology and biomolecular engineering. Designing synthetic condensates or programmable phase behaviors requires predicting outcomes of multi-component systems under defined conditions—precisely the type of problem addressed by event-level representations and multimodal models. By providing a unified framework that links molecular components, context and emergent properties, event-based approaches facilitate rational design and functional control of phase-separated systems.

Together, these considerations highlight that treating LLPS as a computational event is not merely a conceptual refinement, but a practical necessity for translating computational insights into mechanistic understanding and real-world applications. By aligning data, models and biological questions around a shared event unit, this perspective makes it easier to trace explicit component–condition–outcome relationships and to perform comparable inference across cellular states and disease contexts. Moreover, event-centric representations naturally support actionable outputs—such as state transitions, role assignments, and context-dependent intervention points—thereby narrowing the gap between computational prediction, experimentally testable hypotheses, and translational strategies.

Challenges and outlook: from event modelling to translational liquid–liquid phase separation biology

Despite the conceptual and practical advances achieved in LLPS data curation and predictive modelling, several fundamental challenges continue to limit the mechanistic interpretability and translational impact of current computational approaches. At the core of these challenges lies the intrinsic nature of LLPS as a multicomponent, condition-dependent phenomenon that emerges from collective molecular behavior rather than from isolated biomolecules. Although molecule-centric abstractions have enabled scalable dataset construction and efficient model training, they remain poorly aligned with the biological reality of phase separation, particularly when contextual modulation and conditional outcomes are central to function.

A primary bottleneck for advancing beyond molecule-centric frameworks is the limited availability of high-quality, event-level annotations. Most existing LLPS datasets remain organized around

individual proteins or nucleic acids, with heterogeneous definitions of phase separation, inconsistent experimental conditions and incomplete contextual metadata. In many cases, records lack a consistent schema specifying which molecular components co-participate, under what physicochemical or cellular regimes, and with which observable outcomes—precisely the information required to treat LLPS as a structured computational event. Addressing this limitation will require coordinated community efforts in data curation, reporting standards and annotation practices, including shared vocabularies, explicit evidence codes, and transparent provenance tracking. Minimal reporting requirements for component stoichiometry, perturbations, and assay readouts would further enable event definitions that are interoperable and machine-actionable.

A second major challenge arises at the modelling level. Event-centric representations necessarily introduce high-dimensional and heterogeneous inputs that encompass multiple biomolecules, interaction networks, and contextual variables. While multimodal and graph-based learning architectures provide powerful tools for such integration, they also impose increased demands on data volume, computational resources, and interpretability [189]. Moreover, event-level data will inevitably be incomplete: interaction graphs may be partial, contextual descriptors missing, and outcome labels coarse or assay-dependent. Robust modelling therefore requires principled strategies for handling uncertainty and missing modalities, such as modular architectures that combine pretrained encoders with event-level aggregation layers, weakly or semi-supervised learning, and interpretability methods that attribute predictions to specific components, interaction motifs, or contextual factors rather than opaque latent features.

Validation constitutes a further bottleneck that becomes particularly pronounced at the event level. Unlike molecule-centric predictions, which can often be benchmarked against curated positive and negative sets, event-level predictions demand context-aware experimental validation. This requirement complicates performance assessment and underscores the necessity of close integration between computational modelling and experimental design. Prospective validation may involve targeted reconstitution assays, cellular perturbation experiments or controlled manipulation of component stoichiometry and environmental parameters to test predicted conditional behaviors and state transitions [190]. In this setting, iterative feedback loops between prediction and validation are essential, and active-learning-style strategies that prioritize experiments with maximal information gain may help focus experimental effort where it is most informative.

Despite these challenges, the outlook for event-level LLPS research remains highly promising. Advances in high-throughput experimental technologies, improved reporting of interaction and condition metadata, and the continued expansion of curated LLPS resources are gradually enriching the available data landscape. In particular, emerging multi-omics datasets, spatially resolved measurements and systematic perturbation screens are beginning to provide the contextual richness required to parameterize LLPS beyond single-molecule labels [191]. In parallel, rapid progress in foundation models and multimodal learning architectures offers an increasingly powerful computational substrate for modelling complex biological events. Together, these developments suggest that many of the technical prerequisites for event-aware LLPS modelling are now coming into place. However, the availability of richer data and more powerful models does not, by itself, guarantee a corresponding shift in conceptual abstraction or analytical framing.

Nevertheless, current progress in LLPS bioinformatics remains largely incremental at the level of data abstraction and task formulation. Although expanding data resources and increasingly powerful learning architectures have improved coverage and predictive capacity, most existing datasets and models are still organized around individual molecules or simplified propensity labels [192]. At the same time, recent predictive studies have begun to move beyond coarse-grained protein-level identification toward more mechanistically informative questions, including the delineation of LLPS-driving regions, the assessment of mutation or post-translational modification effects, and the classification of dependency patterns or functional roles. This shift suggests a growing effort to extract deeper biological meaning from LLPS prediction, even though most current formulations remain grounded in molecule-centric abstractions. As a result, multicomponent assembly logic, environmental dependence, and context-specific phase behaviors remain insufficiently represented in current computational frameworks. This continuing abstraction gap reinforces our central argument that next-generation LLPS bioinformatics should move beyond molecule-centric formulations toward explicitly event-centric frameworks capable of encoding assembly-condition-behavior relationships and supporting integrative, context-aware modelling.

Taken together, these observations reinforce a central conclusion of this review: continued refinement of molecule-centric predictors alone is unlikely to resolve the core conceptual and practical challenges of LLPS modelling. Instead, progress will increasingly depend on a decisive shift toward event-centric frameworks that explicitly encode molecular assemblies, contextual conditions, and observable phase behaviors. Treating LLPS as a structured computational event offers a unifying abstraction that can bridge molecular features, interaction networks, and cellular context, align computational outputs with experimental observables, and support more interpretable and actionable predictions. Such a perspective not only advances fundamental understanding of phase separation but also provides a stronger foundation for disease mechanism analysis and translational applications, where conditional and context-specific behaviors are paramount. Ultimately, broad adoption of event-centric LLPS modelling will require sustained community alignment around shared definitions, transparent evidence hierarchies, and interoperable standards, enabling integrative and reproducible analysis to become routine rather than bespoke.

Conclusion

LLPS has emerged as a fundamental principle underlying cellular organization, regulation, and disease. In parallel with rapid experimental advances, bioinformatics and computational biology have become indispensable for extracting patterns, formulating hypotheses, and enabling large-scale analysis of LLPS-related data. In this review, we systematically examined current bioinformatics resources and predictive models for LLPS, providing a comprehensive synthesis of available databases and computational tasks while highlighting both major methodological advances and persistent conceptual limitations.

Our analysis demonstrates that, despite increasingly sophisticated representations and modelling techniques, most existing computational studies remain anchored to molecule-centric formulations that treat LLPS as an intrinsic property of individual biomolecules. While such abstractions have enabled scalable data integration and efficient

predictor development, they only partially capture the conditional, multicomponent, and context-dependent nature of phase separation. As a result, improvements in predictive performance do not always translate into deeper mechanistic insight or translational relevance. To address this disconnect, we proposed a unifying event-centric perspective that reframes LLPS as a structured biological event defined by participating components, interaction relationships, contextual conditions, and functional outcomes. Building on this definition, we outlined how data resources, modelling strategies, and application scenarios can be coherently reorganized around LLPS events.

Within this framework, next-generation bioinformatics in LLPS should treat the event—rather than the isolated molecule—as the primary unit of representation, curation, and prediction. On the resource side, this entails the development of event-level datasets that explicitly encode environmental dependence, including temperature, pH, ionic strength, macromolecular crowding, subcellular localization, and stress states; molecular dependence, spanning partner identity, stoichiometry, competition or cooperativity, post-translational modification, and mutation effects; and dynamic evolution, encompassing assembly and disassembly kinetics, material-state transitions, maturation trajectories, and aberrant aggregation. On the modelling side, predictor development should likewise proceed from an event-based formulation, leveraging multimodal architectures that integrate sequence and structure representations with interaction graphs and contextual metadata, and further incorporating AI-enabled foundation models together with multi-omics signals from proteomic, transcriptomic, epigenomic, and spatially resolved measurements to infer conditional behaviors and emergent properties. Importantly, such an approach provides a more faithful bridge between computational prediction and biological interpretation, particularly in the study of disease mechanisms and translational applications.

Looking ahead, the realization of next-generation LLPS bioinformatics will require coordinated progress in data standardization, event-level annotation, and integrative modelling. By aligning computational abstractions with biological complexity, event-centric frameworks have the potential to shift LLPS bioinformatics from descriptive prediction toward mechanistic understanding and rational intervention. We anticipate that this paradigm will play an increasingly central role in shaping future LLPS research across basic, biomedical, and applied domains.

Key Points

- LLPS is a biological event, not a molecule-level trait—it is inherently multicomponent and condition-dependent.
- The current LLPS bioinformatics ecosystem remains fragmented and predominantly molecule-centric: resources vary in evidence standards and annotation consistency, and existing predictive models often underrepresent event-level context and mechanistic structure.
- Next-generation bioinformatics in LLPS: make events the unit of curation and modelling—standardize evidence, annotate event states, and build integrative models that advance from prediction to mechanism to intervention.
- For data resources, a central priority is to move from molecule lists to interoperable event graphs that explicitly link components/stoichiometry, physicochemical conditions, spatiotemporal dynamics

and material outcomes, supported by unified standards, evidence tiers, and clear provenance.

- In predictive modelling, move beyond propensity scores to event-aware multimodal AI integrating sequence/structure, interaction networks and context, augmented by foundation-model representations and multi-omics to learn conditional rules and emergent behaviors.

Acknowledgements

We would like to thank the anonymous reviewers for valuable suggestions.

Author contributions

Lin Ning, Rui-qin Fang, and Mengze Du conceived and supervised the project. Zi-long Yuan, Bo Wang, Yu-lu Chen, Hao-qi Huang and Bin-hao Li were responsible for literature collection and curation. Zi-long Yuan drafted the initial version of the manuscript and prepared the figures. Bo Wang substantially revised and edited the manuscript. Ahmed Zahoor and Liping Ren provided critical revisions to the text.

Supplementary material

[Supplementary material](#) is available at *Briefings in Bioinformatics* online.

Funding

This work was supported by the National Natural Science Foundation of China (62372088, 62501117, and 62501110); China Postdoctoral Science Foundation (2024M750367) and the Research Project of Zhejiang Chinese Medical University (2025RCZXK44).

Data availability

This study does not produce or analyze new data.

References

1. Ghosh A, Mazarakos K, Zhou HX. Three archetypical classes of macromolecular regulators of protein liquid-liquid phase separation. *Proc Natl Acad Sci USA* 2019;**116**:19474–83. <https://doi.org/10.1073/pnas.1907849116>
2. Kumar A, Tanaka K, Schwartz MA. Focal adhesion-derived liquid-liquid phase separations regulate mRNA translation. *elife* 2025;**13**:RP96157. <https://doi.org/10.7554/eLife.96157>
3. Lee DSW, Choi CH, Sanders DW *et al.* Size distributions of intracellular condensates reflect competition between coalescence and nucleation. *Nat Phys* 2023;**19**:586–96. <https://doi.org/10.1038/s41567-022-01917-0>
4. Brangwynne CP, Eckmann CR, Courson DS *et al.* Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* 2009;**324**:1729–32. <https://doi.org/10.1126/science.1172046>
5. Höllmüller E, Geigges S, Niedermeier ML *et al.* Site-specific ubiquitylation acts as a regulator of linker histone H1. *Nat Commun* 2021;**12**:3497. <https://doi.org/10.1038/s41467-021-23636-5>

6. Lin Y. Active regulation of amyloidosis. *Proc Natl Acad Sci USA* 2024;**121**:e2409665121. <https://doi.org/10.1073/pnas.2409665121>
7. Alberti S, Arosio P, Best RB *et al.* Current practices in the study of biomolecular condensates: a community comment. *Nat Commun* 2025;**16**:7730. <https://doi.org/10.1038/s41467-025-62055-8>
8. Ambadi Thody S, Clements HD, Baniyadi H *et al.* Small-molecule properties define partitioning into biomolecular condensates. *Nat Chem* 2024;**16**:1794–802. <https://doi.org/10.1038/s41557-024-01630-w>
9. Hoang Y, Azaldegui CA, Dow RE *et al.* An experimental framework to assess biomolecular condensates in bacteria. *Nat Commun* 2024;**15**:3222. <https://doi.org/10.1038/s41467-024-47330-4>
10. Zhu Q, Raza Z, Do-Ha D *et al.* Biomolecular condensates as emerging biomaterials: functional mechanisms and advances in computational and experimental approaches. *Adv Mater* 2025;**37**:e10115. <https://doi.org/10.1002/adma.202510115>
11. Banani SF, Lee HO, Hyman AA *et al.* Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol* 2017;**18**:285–98. <https://doi.org/10.1038/nrm.2017.7>
12. Shimobayashi SF, Ronceray P, Sanders DW *et al.* Nucleation landscape of biomolecular condensates. *Nature* 2021;**599**:503–6. <https://doi.org/10.1038/s41586-021-03905-5>
13. Youn JY, Dyakov BJA, Zhang J *et al.* Properties of stress granule and P-body proteomes. *Mol Cell* 2019;**76**:286–94. <https://doi.org/10.1016/j.molcel.2019.09.014>
14. Hirose T, Ninomiya K, Nakagawa S *et al.* A guide to membraneless organelles and their various roles in gene regulation. *Nat Rev Mol Cell Biol* 2023;**24**:288–304. <https://doi.org/10.1038/s41580-022-00558-8>
15. Liu T, Qiao H, Wang Z *et al.* CodLncScape provides a self-enriching framework for the systematic collection and exploration of coding LncRNAs. *Advanced. Science* 2024/06/01 2024;**11**:e2400009. <https://doi.org/10.1002/advs.202400009>
16. Garg M, Lamba S, Rajyaguru PI. Emerging trends in the disassembly of stress granules and P-bodies. *Cell Rep* 2025;**44**:116340. <https://doi.org/10.1016/j.celrep.2025.116340>
17. Bivona TG. Phase-separated biomolecular condensation in cancer: new horizons and next Frontiers. *Cancer Discov* 2024;**14**:630–4. <https://doi.org/10.1158/2159-8290.Cd-23-1551>
18. Hurtle BT, Xie L, Donnelly CJ. Disrupting pathologic phase transitions in neurodegeneration. *J Clin Invest* 2023;**133**:e168549. <https://doi.org/10.1172/JCI168549>
19. Mehta S, Zhang J. Liquid-liquid phase separation drives cellular function and dysfunction in cancer. *Nat Rev Cancer* 2022;**22**:239–52. <https://doi.org/10.1038/s41568-022-00444-7>
20. Silva JL, Foguel D, Ferreira VF *et al.* Targeting biomolecular condensation and protein aggregation against cancer. *Chem Rev* 2023;**123**:9094–138. <https://doi.org/10.1021/acs.chemrev.3c00131>
21. 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. <https://doi.org/10.1016/j.cell.2012.04.017>
22. Wang J, Choi JM, 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. <https://doi.org/10.1016/j.cell.2018.06.006>
23. Lotthammer JM, Ginell GM, Griffith D *et al.* Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat Methods* 2024;**21**:465–76. <https://doi.org/10.1038/s41592-023-02159-5>
24. Zhao B, Kurgan L. Compositional bias of intrinsically disordered proteins and regions and their predictions. *Biomolecules* 2022;**12**:888. <https://doi.org/10.3390/biom12070888>
25. Arter WE, Qi R, Erkamp NA *et al.* Biomolecular condensate phase diagrams with a combinatorial microdroplet platform. *Nat Commun* 2022;**13**:7845. <https://doi.org/10.1038/s41467-022-35265-7>
26. Li P, Chen P, Qi F *et al.* High-throughput and proteome-wide discovery of endogenous biomolecular condensates. *Nat Chem* 2024;**16**:1101–12. <https://doi.org/10.1038/s41557-024-01485-1>
27. Leurs YHA, van den Hout W, Gardin A *et al.* Automated navigation of condensate phase behavior with active machine learning. *Nat Commun* 2025;**16**:9598. <https://doi.org/10.1038/s41467-025-64617-2>
28. Wang Y, Yu C, Pei G *et al.* Dissolution of oncofusion transcription factor condensates for cancer therapy. *Nat Chem Biol* 2023;**19**:1223–34. <https://doi.org/10.1038/s41589-023-01376-5>
29. Wu J, Xiao Y, Liu Y *et al.* Dynamics of RNA localization to nuclear speckles are connected to splicing efficiency. *Sci Adv* 2024;**10**:eadp7727. <https://doi.org/10.1126/sciadv.adp7727>
30. Hu S, Zhang Y, Yi Q *et al.* Time-resolved proteomic profiling reveals compositional and functional transitions across the stress granule life cycle. *Nat Commun* 2023;**14**:7782. <https://doi.org/10.1038/s41467-023-43470-1>
31. Schaffer LV, Hu M, Qian G *et al.* Multimodal cell maps as a foundation for structural and functional genomics. *Nature* 2025;**642**:222–31. <https://doi.org/10.1038/s41586-025-08878-3>
32. Saar KL, Scrutton RM, Blozelyte K *et al.* Protein condensate atlas from predictive models of heteromolecular condensate composition. *Nat Commun* 2024;**15**:5418. <https://doi.org/10.1038/s41467-024-48496-7>
33. Kappel K, Streibinger D, Edmonds KK *et al.* Characterizing protein sequence determinants of nuclear condensates by high-throughput pooled imaging with CondensSeq. *Nat Methods* 2025;**22**:1464–75. <https://doi.org/10.1038/s41592-025-02726-y>
34. Pan CR, Knutson SD, Huth SW *et al.* μ Map proximity labeling in living cells reveals stress granule disassembly mechanisms. *Nat Chem Biol* 2025;**21**:490–500. <https://doi.org/10.1038/s41589-024-01721-2>
35. Morris R, Ali R, Cheng F. Drug repurposing using FDA adverse event reporting system (FAERS) database. *Curr Drug Targets* 2024;**25**:454–64. <https://doi.org/10.2174/0113894501290296240327081624>
36. Wu X, Wang K, Hu G *et al.* Empirical assessment of sequence-based predictions of intrinsically disordered regions involved in phase separation. *Biomolecules* 2025;**15**:1079. <https://doi.org/10.3390/biom15081079>
37. Wessén J, De La Cruz N, Lyons H *et al.* Sequence-based prediction of condensate composition reveals that specificity can emerge from multivalent interactions among disordered regions. *Commun Chem* 2025;**8**:304. <https://doi.org/10.1038/s42004-025-01687-9>
38. Orti F, Fernández ML, Marino-Buslje C. MLOsMetaDB, a meta-database to centralize the information on liquid-liquid phase separation proteins and membraneless organelles. *Protein Sci* 2024;**33**:e4858. <https://doi.org/10.1002/pro.4858>
39. Xiao Q, McAtee CK, Su X. Phase separation in immune signalling. *Nat Rev Immunol* 2022;**22**:188–99. <https://doi.org/10.1038/s41577-021-00572-5>
40. Pintado-Grima C, Bárcenas O, Arribas-Ruiz E *et al.* Comprehensive protein datasets and benchmarking for liquid-liquid phase separation studies. *Genome Biol* 2025;**26**:198. <https://doi.org/10.1186/s13059-025-03668-6>

41. Ausserwöger H, de Csilléry E, Qian D *et al.* Quantifying collective interactions in biomolecular phase separation. *Nat Commun* 2025;**16**:7724. <https://doi.org/10.1038/s41467-025-62437-y>
42. Farahi N, Lazar T, Wodak SJ *et al.* Integration of data from liquid-liquid phase separation databases highlights concentration and dosage sensitivity of LLPS drivers. *Int J Mol Sci* 2021;**22**:3017, 1–26. <https://doi.org/10.3390/ijms22063017>
43. Zhou Y, Zhang C, He L *et al.* Glucose-1-phosphate promotes compartmentalization of glycogen with the pentose phosphate pathway in CD8(+) memory T cells. *Mol Cell* 2025;**85**:2535–2549.e10. <https://doi.org/10.1016/j.molcel.2025.05.019>
44. Thappeta Y, Cañas-Duarte SJ, Wang H *et al.* Glycogen phase-separation drives macromolecular rearrangement and asymmetric division in *E. Coli*. *EMBO J* 2025;**44**:7434–76. <https://doi.org/10.1038/s44318-025-00621-y>
45. Klobučar T, Novljan J, Iosub IA *et al.* Integrative profiling of condensation-prone RNAs during early development. *Cell Genom* 2026;**6**:101065. <https://doi.org/10.1016/j.xgen.2025.101065>
46. Li Q, Wang X, Dou Z *et al.* Protein databases related to liquid-liquid phase separation. *Int J Mol Sci* 2020;**21**:6796. <https://doi.org/10.3390/ijms21186796>
47. Shen B, Chen Z, Yu C *et al.* Computational screening of phase-separating proteins. *Genomics Proteomics Bioinformatics* 2021;**19**:13–24. <https://doi.org/10.1016/j.gpb.2020.11.003>
48. Vendruscolo M, Fuxreiter M. Towards sequence-based principles for protein phase separation predictions. *Curr Opin Chem Biol* 2023;**75**:102317. <https://doi.org/10.1016/j.cbpa.2023.102317>
49. Venko K, Žerovnik E. Protein condensates and protein aggregates: In vitro, in the cell, and In Silico. *Front Biosci (Landmark Ed)* 2023;**28**:183. <https://doi.org/10.31083/j.fbl2808183>
50. Vernon RM, Forman-Kay JD. First-generation predictors of biological protein phase separation. *Curr Opin Struct Biol* 2019;**58**:88–96. <https://doi.org/10.1016/j.sbi.2019.05.016>
51. von Bülow S, Tesei G, Lindorff-Larsen K. Machine learning methods to study sequence-ensemble-function relationships in disordered proteins. *Curr Opin Struct Biol* 2025;**92**:103028. <https://doi.org/10.1016/j.sbi.2025.103028>
52. Kuechler ER, Huang A, Bui JM *et al.* Comparison of biomolecular condensate localization and protein phase separation predictors. *Biomolecules* 2023;**13**:527. <https://doi.org/10.3390/biom13030527>
53. Liao S, Zhang Y, Qi Y *et al.* Evaluation of sequence-based predictors for phase-separating protein. *Brief Bioinform* 2023;**24**:bbad213. <https://doi.org/10.1093/bib/bbad213>
54. Deng B, Wan G. Technologies for studying phase-separated biomolecular condensates. *Advanced biotechnology* 2024;**2**:10. <https://doi.org/10.1007/s44307-024-00020-0>
55. Pancsa R, Vranken W, Mészáros B. Computational resources for identifying and describing proteins driving liquid-liquid phase separation. *Brief Bioinform* 2021;**22**:bbaa408. <https://doi.org/10.1093/bib/bbaa408>
56. Zhang Y, Pan X, Shi T *et al.* P450Rdb: a manually curated database of reactions catalyzed by cytochrome P450 enzymes. *J Adv Res* 2024/09/01/ 2024;**63**:35–42. <https://doi.org/10.1016/j.jare.2023.10.012>
57. Liu T, Qiao H, Ren L *et al.* Deciphering cancer therapy resistance via patient-level single-cell transcriptomics with CellResDB. *Commun Biol* 2025/07/15 2025;**8**:1049. <https://doi.org/10.1038/s42003-025-08457-2>
58. Casadio R, Martelli PL, Savojardo C. Machine learning solutions for predicting protein–protein interactions 2022;**12**:e1618. <https://doi.org/10.1002/wcms.1618>
59. Ramanathan A, Ma H, Parvatikar A *et al.* Artificial intelligence techniques for integrative structural biology of intrinsically disordered proteins. *Curr Opin Struct Biol* 2021;**66**:216–24. <https://doi.org/10.1016/j.sbi.2020.12.001>
60. Saar KL, Qian D, Good LL *et al.* Theoretical and data-driven approaches for biomolecular condensates. *Chem Rev* 2023;**123**:8988–9009. <https://doi.org/10.1021/acs.chemrev.2c00586>
61. Szala-Mendyk B, Phan TM, Mohanty P *et al.* Challenges in studying the liquid-to-solid phase transitions of proteins using computer simulations. *Curr Opin Chem Biol* 2023;**75**:102333. <https://doi.org/10.1016/j.cbpa.2023.102333>
62. Xing JX, Yang SQ, Liang YC *et al.* Deciphering sequence determinants of zygotic genome activation genes: insights from machine learning and the ZGAExplorer platform. *Cell Prolif* 2025;**58**:e70039. <https://doi.org/10.1111/cpr.70039>
63. Dai BJ, Li HS, Wang PZ *et al.* DualNetM: an adaptive dual network framework for inferring functional-oriented markers. *BMC Biol* 2025;**23**:254. <https://doi.org/10.1186/s12915-025-02367-9>
64. You K, Li R, Lian R *et al.* PhaSepDB 3.0: a comprehensive knowledgebase of phase separation-related proteins from AI-assisted curation. *Nucleic Acids Res* 2025;**54**:D445–50. <https://doi.org/10.1093/nar/gkaf973>
65. Hou C, Wang X, Xie H *et al.* PhaSepDB in 2022: annotating phase separation-related proteins with droplet states, co-phase separation partners and other experimental information. *Nucleic Acids Res* 2023;**51**:D460–d465. <https://doi.org/10.1093/nar/gkac783>
66. You K, Huang Q, Yu C *et al.* PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic Acids Res* 2020;**48**:D354–d359. <https://doi.org/10.1093/nar/gkz847>
67. Gao R, Guo M, Luan S *et al.* ricePSP: a database of rice phase separation-associated proteins. *Genome Biol* 2025;**26**:390. <https://doi.org/10.1186/s13059-025-03842-w>
68. Rodriguez CB, Tunque Cahui RR, Demitroff N *et al.* BAV-LLPS: a database of bacterial, archaea, and virus liquid-liquid phase separation proteins. *Bioinformatics* 2025;**41**:btaf550. <https://doi.org/10.1093/bioinformatics/btaf550>
69. Chen T, Tang G, Li T *et al.* PhaSeDis: a manually curated database of phase separation-disease associations and corresponding small molecules. *Genomics Proteomics Bioinformatics* 2025;**23**:qzaf014. <https://doi.org/10.1093/gpbjnl/qzaf014>
70. He Y, Bao X, Chen T *et al.* RPS 2.0: an updated database of RNAs involved in liquid-liquid phase separation. *Nucleic Acids Res* 2025;**53**:D299–d309. <https://doi.org/10.1093/nar/gkae951>
71. Liu M, Li H, Luo X *et al.* RPS: a comprehensive database of RNAs involved in liquid-liquid phase separation. *Nucleic Acids Res* 2022;**50**:D347–d355. <https://doi.org/10.1093/nar/gkab986>
72. Rostam N, Ghosh S, Chow CFW *et al.* CD-CODE: crowdsourcing condensate database and encyclopedia. *Nat Methods* 2023;**20**:673–6. <https://doi.org/10.1038/s41592-023-01831-0>
73. Wang X, Zhou X, Yan Q *et al.* LLPsDB v2.0: an updated database of proteins undergoing liquid-liquid phase separation in vitro. *Bioinformatics* 2022;**38**:2010–4. <https://doi.org/10.1093/bioinformatics/btac026>
74. 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–d327. <https://doi.org/10.1093/nar/gkz778>
75. Zhu H, Fu H, Cui T *et al.* RNAPhaSep: a resource of RNAs undergoing phase separation. *Nucleic Acids Res* 2022;**50**:D340–d346. <https://doi.org/10.1093/nar/gkab985>
76. Mészáros B, Erdős G, Szabó B *et al.* PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res* 2020;**48**:D360–d367. <https://doi.org/10.1093/nar/gkz848>

77. 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–d295. <https://doi.org/10.1093/nar/gkz1027>
78. Li P, Banjade S, Cheng HC *et al.* Phase transitions in the assembly of multivalent signalling proteins. *Nature* 2012;**483**:336–40. <https://doi.org/10.1038/nature10879>
79. Cai H, Vernon RM, Forman-Kay JD. An interpretable machine-learning algorithm to predict disordered protein phase separation based on biophysical interactions. *Biomolecules* 2022;**12**:1131. <https://doi.org/10.3390/biom12081131>
80. Zheng L, Liang P, Long C *et al.* EmAtlas: a comprehensive atlas for exploring spatiotemporal activation in mammalian embryogenesis. *Nucleic Acids Res* 2023;**51**:D924–32. <https://doi.org/10.1093/nar/gkac848>
81. Hatos A, Tosatto SCE, Vendruscolo M *et al.* FuzDrop on AlphaFold: visualizing the sequence-dependent propensity of liquid-liquid phase separation and aggregation of proteins. *Nucleic Acids Res* 2022;**50**:W337–w344. <https://doi.org/10.1093/nar/gkac386>
82. Mathur A, Ghosh R, Nunes-Alves A. Recent progress in Modeling and simulation of biomolecular crowding and condensation inside cells. *J Chem Inf Model* 2024;**64**:9063–81. <https://doi.org/10.1021/acs.jcim.4c01520>
83. Mohapatra M, Sahu C, Mohapatra S. Trends of artificial intelligence (AI) use in drug targets, discovery and development: current status and future perspectives. *Curr Drug Targets* 2025;**26**:221–42. <https://doi.org/10.2174/0113894501322734241008163304>
84. Prašnikar M, Žiberna MB, Ahlin GP. Targeting intermolecular interactions to reduce viscosity in monoclonal antibody formulations: a review. *Int J Biol Macromol* 2025;**327**:147515. <https://doi.org/10.1016/j.ijbiomac.2025.147515>
85. 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. <https://doi.org/10.1016/j.celrep.2016.05.076>
86. Liao S, Zhang Y, Han X *et al.* A sequence-based model for identifying proteins undergoing liquid-liquid phase separation/forming fibril aggregates via machine learning. *Protein Sci* 2024;**33**:e4927. <https://doi.org/10.1002/pro.4927>
87. 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. <https://doi.org/10.1093/bioinformatics/btz274>
88. Hou S, Hu J, Yu Z *et al.* Machine learning predictor PSPire screens for phase-separating proteins lacking intrinsically disordered regions. *Nat Commun* 2024;**15**:2147. <https://doi.org/10.1038/s41467-024-46445-y>
89. Liang Q, Peng N, Xie Y *et al.* MolPhase, an advanced prediction algorithm for protein phase separation. *EMBO J* 2024;**43**:1898–918. <https://doi.org/10.1038/s44318-024-00090-9>
90. Vernon RM, Chong PA, Tsang B *et al.* Pi-pi contacts are an overlooked protein feature relevant to phase separation. *elife* 2018;**7**:e31486. <https://doi.org/10.7554/eLife.31486>
91. Hadarovich A, Singh HR, Ghosh S *et al.* PICNIC accurately predicts condensate-forming proteins regardless of their structural disorder across organisms. *Nat Commun* 2024;**15**:10668. <https://doi.org/10.1038/s41467-024-55089-x>
92. Sun J, Qu J, Zhao C *et al.* Precise prediction of phase-separation key residues by machine learning. *Nat Commun* 2024;**15**:2662. <https://doi.org/10.1038/s41467-024-46901-9>
93. van Mierlo G, Jansen JRG, Wang J *et al.* Predicting protein condensate formation using machine learning. *Cell Rep* 2021;**34**:108705. <https://doi.org/10.1016/j.celrep.2021.108705>
94. Chu X, Sun T, Li Q *et al.* Prediction of liquid-liquid phase separating proteins using machine learning. *BMC Bioinformatics* 2022;**23**:72. <https://doi.org/10.1186/s12859-022-04599-w>
95. Hardenberg M, Horvath A, Ambrus V *et al.* Widespread occurrence of the droplet state of proteins in the human proteome. *Proc Natl Acad Sci USA* 2020;**117**:33254–62. <https://doi.org/10.1073/pnas.2007670117>
96. Wang G, Warrell J, Zheng S *et al.* A variational graph-partitioning approach to modeling protein liquid-liquid phase separation. *Cell Rep Phys Sci* 2024;**5**:102292. <https://doi.org/10.1016/j.xcrp.2024.102292>
97. Raimondi D, Orlando G, Michiels E *et al.* In silico prediction of in vitro protein liquid-liquid phase separation experiments outcomes with multi-head neural attention. *Bioinformatics* 2021;**37**:3473–9. <https://doi.org/10.1093/bioinformatics/btab350>
98. Huang J, Zhang Y, Ren S *et al.* MambaPhase: deep learning for liquid-liquid phase separation protein classification. *Brief Bioinform* 2025;**26**:bbaf230. <https://doi.org/10.1093/bib/bbaf230>
99. Yang YH, Liu Q, Liu JF *et al.* Prediction of liquid-phase separation proteins using Siamese network with feature fusion. *Brief Bioinform* 2025;**26**:bbaf393. <https://doi.org/10.1093/bib/bbaf393>
100. Zhou S, Zhou Y, Liu T *et al.* PredLLPS_PSSM: a novel predictor for liquid-liquid protein separation identification based on evolutionary information and a deep neural network. *Brief Bioinform* 2023;**24**:bbad299. <https://doi.org/10.1093/bib/bbad299>
101. Feng M, Liu L, Xian ZN *et al.* PSTP: accurate residue-level phase separation prediction using protein conformational and language model embeddings. *Brief Bioinform* 2025;**26**:bbaf171. <https://doi.org/10.1093/bib/bbaf171>
102. Zhou Y, Zhou S, Bi Y *et al.* A two-task predictor for discovering phase separation proteins and their undergoing mechanism. *Brief Bioinform* 2024;**25**:bbae528. <https://doi.org/10.1093/bib/bbae528>
103. Monti M, Fiorentino J, Miltiadis-Vrachnos D *et al.* catGRANULE 2.0: accurate predictions of liquid-liquid phase separating proteins at single amino acid resolution. *Genome Biol* 2025;**26**:33. <https://doi.org/10.1186/s13059-025-03497-7>
104. Saar KL, Morgunov AS, Qi R *et al.* Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proc Natl Acad Sci USA* 2021;**118**:e2019053118. <https://doi.org/10.1073/pnas.2019053118>
105. Frank M, Ni P, Jensen M *et al.* Leveraging a large language model to predict protein phase transition: a physical, multiscale, and interpretable approach. *Proc Natl Acad Sci USA* 2024;**121**:e2320510121. <https://doi.org/10.1073/pnas.2320510121>
106. Chin KY, Ishida S, Sasaki Y *et al.* Predicting condensate formation of protein and RNA under various environmental conditions. *BMC Bioinformatics* 2024;**25**:143. <https://doi.org/10.1186/s12859-024-05764-z>
107. von Bülow S, Tesei G, Zaidi FK *et al.* Prediction of phase-separation propensities of disordered proteins from sequence. *Proc Natl Acad Sci USA* 2025;**122**:e2417920122. <https://doi.org/10.1073/pnas.2417920122>
108. Oliver WW, Jacobs WM, Webb MA. When B(2) is not enough: evaluating simple metrics for predicting phase separation of intrinsically disordered proteins. *J Phys Chem B* 2025;**129**:9551–65. <https://doi.org/10.1021/acs.jpcc.5c04955>
109. Lahorkar A, Bhosale H, Sane A *et al.* Identification of phase separating proteins with distributed reduced alphabet representations of sequences. *IEEE/ACM Trans Comput Biol Bioinform* 2023;**20**:410–20. <https://doi.org/10.1109/tcbb.2022.3149310>

110. Mullick P, Trovato A. Sequence-based prediction of protein phase separation: the role of Beta-pairing propensity. *Biomolecules* 2022;**12**:1771. <https://doi.org/10.3390/biom12121771>
111. Ma Q, Huang F, Guo W *et al.* Identification of phase-separation-protein-related function based on gene ontology by using machine learning methods. *Life (Basel)* 2023;**13**:1306. <https://doi.org/10.3390/life13061306>
112. Ahmed Z, Shahzadi K, Temesgen SA *et al.* A protein pre-trained model-based approach for the identification of the liquid-liquid phase separation (LLPS) proteins. *Int J Biol Macromol* 2024;**277**:134146. <https://doi.org/10.1016/j.ijbiomac.2024.134146>
113. Wang J, Chang H, Quan X *et al.* A model for identification of potential phase-separated proteins based on protein sequence, structure and cellular distribution. *Int J Biol Macromol* 2023;**243**:125196. <https://doi.org/10.1016/j.ijbiomac.2023.125196>
114. Li W, Deng X, Xiang C *et al.* Prediction of liquid-liquid phase separation proteins based on protein language model. *Brief Bioinform* 2025;**26**:bbaf681. <https://doi.org/10.1093/bib/bbaf681>
115. Paiz EA, Allen JH, Correia JJ *et al.* Beta turn propensity and a model polymer scaling exponent identify intrinsically disordered phase-separating proteins. *J Biol Chem* 2021;**297**:101343. <https://doi.org/10.1016/j.jbc.2021.101343>
116. Yu K, Liu Z, Cheng H *et al.* dSCOPE: a software to detect sequences critical for liquid-liquid phase separation. *Brief Bioinform* 2023;**24**:bbac550. <https://doi.org/10.1093/bib/bbac550>
117. Wilson C, Lewis KA, Fitzkee NC *et al.* ParSe 2.0: a web tool to identify drivers of protein phase separation at the proteome level. *Protein Sci* 2023;**32**:e4756. <https://doi.org/10.1002/pro.4756>
118. 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. <https://doi.org/10.1093/bioinformatics/btu310>
119. Ho WL, Huang HC, Huang JR. IFF: identifying key residues in intrinsically disordered regions of proteins using machine learning. *Protein Sci* 2023;**32**:e4739. <https://doi.org/10.1002/pro.4739>
120. Xiao J, Hu G, Zhou X *et al.* TIDGN: a transfer learning framework for predicting interactions of intrinsically disordered proteins with high conformational dynamics. *J Chem Inf Model* 2025;**65**:4866–77. <https://doi.org/10.1021/acs.jcim.5c00422>
121. Maiti S, Tripathi S, Baggett DW *et al.* Proteome-wide computational analyses reveal links between protein condensate formation and RNA biology. *Sci Adv* 2025;**11**:eady1420. <https://doi.org/10.1126/sciadv.ady1420>
122. Yang H, You K, Ma L *et al.* Interpretable and generative deep learning models explicate phase separating intrinsically disordered motifs. *Nat Commun* 2026;**17**:2571. <https://doi.org/10.1038/s41467-026-69252-z>
123. Zhang Y, Zheng J, Zhang B. Protein language model identifies disordered, conserved motifs implicated in phase separation. *elife* 2025;**14**:RP105309. <https://doi.org/10.7554/eLife.105309>
124. Feng M, Wei X, Zheng X *et al.* Decoding missense variants by incorporating phase separation via machine learning. *Nat Commun* 2024;**15**:8279. <https://doi.org/10.1038/s41467-024-52580-3>
125. Hong X, Lv J, Li Z *et al.* Deep learning model of post-translational modification regulating liquid-liquid phase separation. *Commun Chem* 2025;**8**:393. <https://doi.org/10.1038/s42004-025-01773-y>
126. Jiang P, Cai R, Lugo-Martinez J *et al.* A hybrid positive unlabeled learning framework for uncovering scaffolds across human proteome by measuring the propensity to drive phase separation. *Brief Bioinform* 2023;**24**:bbad009, 1–11. <https://doi.org/10.1093/bib/bbad009>
127. Chen Z, Hou C, Wang L *et al.* Screening membraneless organelle participants with machine-learning models that integrate multimodal features. *Proc Natl Acad Sci USA* 2022;**119**:e2115369119. <https://doi.org/10.1073/pnas.2115369119>
128. Miyata K, Iwasaki W. Seq2Phase: language model-based accurate prediction of client proteins in liquid-liquid phase separation. *Bioinformatics advances* 2024;**4**:vbad189. <https://doi.org/10.1093/bioadv/vbad189>
129. He X, Guan J, Xie P *et al.* PhaseNet: a computational framework for identifying phase-separating proteins based on protein language model. *Int J Biol Macromol* 2025;**334**:149044. <https://doi.org/10.1016/j.ijbiomac.2025.149044>
130. Ahmed Z, Shahzadi K, Jin Y *et al.* Identification of RNA-dependent liquid-liquid phase separation proteins using an artificial intelligence strategy. *Proteomics* 2024/11/01 2024;**24**:e2400044. <https://doi.org/10.1002/pmic.202400044>
131. Ahmed Z, Shahzadi K, Li R *et al.* An artificial intelligence-based approach for identifying the proteins regulating liquid-liquid phase separation. *Brief Bioinform* 2025;**26**:bbaf313. <https://doi.org/10.1093/bib/bbaf313>
132. Ding CH, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 2001;**17**:349–58. <https://doi.org/10.1093/bioinformatics/17.4.349>
133. Li ZR, Lin HH, Han LY *et al.* PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res* 2006;**34**:W32–7. <https://doi.org/10.1093/nar/gkl305>
134. Cai CZ, Han LY, Ji ZL *et al.* SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res* 2003;**31**:3692–7. <https://doi.org/10.1093/nar/gkg600>
135. BE Boser, IM Guyon, VN Vapnik. A training algorithm for optimal margin classifiers. In: Haussler D (ed.), *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory (COLT '92)*; 27–29 July 1992; Pittsburgh, PA, USA, pp. 144–52. New York, NY: Association for Computing Machinery/ACM Press; 1992. <https://doi.org/10.1145/130385.130401>
136. Liu T, Chen JM, Zhang D *et al.* ApoPred: identification of Apolipoproteins and their subfamilies with multifarious features. *Front Cell Dev Biol* 2020;**8**:621144. <https://doi.org/10.3389/fcell.2020.621144>
137. Nguyen NDH, Pham NT, Seo H *et al.* Mulaqua: an interpretable multimodal deep learning framework for identifying PMT/vPvM substances in drinking water. *J Hazard Mater* 2025;**500**:140573. <https://doi.org/10.1016/j.jhazmat.2025.140573>
138. Nguyen NDH, Pham NT, Tran DT *et al.* xBitterT5: an explainable transformer-based framework with multimodal inputs for identifying bitter-taste peptides. *J Chem* 2025;**17**:127. <https://doi.org/10.1186/s13321-025-01078-1>
139. Tran DT, Pham NT, Nguyen NDH *et al.* HyPepTox-fuse: an interpretable hybrid framework for accurate peptide toxicity prediction fusing protein language model-based embeddings with conventional descriptors. *J Pharm Anal* 2025;**15**:101410. <https://doi.org/10.1016/j.jpha.2025.101410>
140. Lecun Y, Bottou L, Bengio Y *et al.* Gradient-based learning applied to document recognition. *Proc IEEE* 1998;**86**:2278–324. <https://doi.org/10.1109/5.726791>
141. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80. <https://doi.org/10.1162/neco.1997.9.8.1735>
142. Zulfiqar H, Guo Z, Ahmad RM *et al.* Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word

- embeddings. *Original Research Frontiers in Medicine* 2024-January-17 2024;**10**:10. <https://doi.org/10.3389/fmed.2023.1291352>
143. Zhang HQ, Arif M, Thafar MA *et al.* PMPred-AE: a computational model for the detection and interpretation of pathological myopia based on artificial intelligence. *Front Med (Lausanne)* 2025;**12**:1529335. <https://doi.org/10.3389/fmed.2025.1529335>
 144. Baek M, DiMaio F, Anishchenko I *et al.* Accurate prediction of protein structures and interactions using a three-track neural network 2021;**373**:871–6. <https://doi.org/doi:10.1126/science.abj8754>
 145. Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds.), *Advances in Neural Information Processing Systems 30: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5998–6008. Long Beach, CA, USA: Curran Associates, Inc.; 2017.
 146. Rives A, Meier J, Sercu T *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* 2021;**118**:e2016239118. <https://doi.org/10.1073/pnas.2016239118>
 147. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021/08/01 2021;**596**:583–9. <https://doi.org/10.1038/s41586-021-03819-2>
 148. Elnaggar A, Heinzinger M, Dallago C *et al.* ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:7112–27. <https://doi.org/10.1109/tpami.2021.3095381>
 149. Breiman L. Random forests. *Mach Learn* 2001/10/01 2001;**45**:5–32. <https://doi.org/10.1023/A:1010933404324>
 150. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995/09/01 1995;**20**:273–97. <https://doi.org/10.1007/BF00994018>
 151. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001;**29**:1189–232. <https://doi.org/10.1214/aos/1013203451>
 152. Zhang HQ, Liu SH, Li R *et al.* MIBPred: ensemble learning-based metal ion-binding protein classifier. *ACS Omega* 2024;**9**:8439–47. <https://doi.org/10.1021/acsomega.3c09587>
 153. Hammond DK, Vanderheynt P, Gribonval R. Wavelets on graphs via spectral graph theory. *Appl Comput Harmon Anal* 2011/03/01/ 2011;**30**:129–50. <https://doi.org/10.1016/j.acha.2010.04.005>
 154. Sandryhaila A, Moura JMF. *Discrete Signal Processing on Graphs* 2013;**61**(7 %J Trans. Sig. Proc.):1644–56. <https://doi.org/10.1109/tsp.2013.2238935>
 155. Alberti S, Gladfelter A, Mittag T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell* 2019;**176**:419–34. <https://doi.org/10.1016/j.cell.2018.12.035>
 156. Feng Z, Jia B, Zhang M. Liquid-liquid phase separation in biology: specific stoichiometric molecular interactions vs promiscuous interactions mediated by disordered sequences. *Biochemistry* 2021;**60**:2397–406. <https://doi.org/10.1021/acs.biochem.1c00376>
 157. Kuechler ER, Jacobson M, Mayor T *et al.* GraPES: the granule protein enrichment server for prediction of biological condensate constituents. *Nucleic Acids Res* 2022;**50**:W384–w391. <https://doi.org/10.1093/nar/gkac279>
 158. Blaabjerg LM, Jonsson N, Boomsma W *et al.* SSEmb: a joint embedding of protein sequence and structure enables robust variant effect predictions. *Nat Commun* 2024/11/07 2024;**15**:9646. <https://doi.org/10.1038/s41467-024-53982-z>
 159. Zhou C, Cai C-P, Huang X-T *et al.* TarKG: a comprehensive biomedical knowledge graph for target discovery. *Bioinformatics* 2024;**40**:btac598. <https://doi.org/10.1093/bioinformatics/btac598>
 160. Ginell GM, Emenecker RJ, Lotthammer JM *et al.* Sequence-based prediction of intermolecular interactions driven by disordered regions. *Science* 2025;**388**:eadq8381. <https://doi.org/10.1126/science.adq8381>
 161. Cardona AH, Ecsedi S, Khier M *et al.* Self-demixing of mRNA copies buffers mRNA:mRNA and mRNA:regulator stoichiometries. *Cell* 2023;**186**:4310–4324.e23. <https://doi.org/10.1016/j.cell.2023.08.018>
 162. Kim T, Yoo J, Do S *et al.* RNA-mediated demixing transition of low-density condensates. *Nat Commun* 2023;**14**:2425. <https://doi.org/10.1038/s41467-023-38118-z>
 163. Peydayesh M, Kistler S, Zhou J *et al.* Amyloid-polysaccharide interfacial coacervates as therapeutic materials. *Nat Commun* 2023;**14**:1848. <https://doi.org/10.1038/s41467-023-37629-z>
 164. De La Cruz N, Pradhan P, Veettil RT *et al.* Disorder-mediated interactions target proteins to specific condensates. *Mol Cell* 2024;**84**:3497–3512.e9. <https://doi.org/10.1016/j.molcel.2024.08.017>
 165. Yang L, Lyu J, Li X *et al.* Phase separation as a possible mechanism for dosage sensitivity. *Genome Biol* 2024;**25**:17. <https://doi.org/10.1186/s13059-023-03128-z>
 166. Shin Y, Brangwynne CP. Liquid phase condensation in cell physiology and disease. *Science* 2017;**357**:eaaf4382. <https://doi.org/10.1126/science.aaf4382>
 167. Lyon AS, Peebles WB, Rosen MK. A framework for understanding the functions of biomolecular condensates across scales. *Nat Rev Mol Cell Biol* 2021;**22**:215–35. <https://doi.org/10.1038/s41580-020-00303-z>
 168. Vandelli A, Arnal Segura M, Monti M *et al.* The PRALINE database: protein and Rna humAn singLe nucleotide varianTs in condEnsates. *Bioinformatics* 2023;**39**:btac847. <https://doi.org/10.1093/bioinformatics/btac847>
 169. Zhang S, Lim CM, Occhetta M *et al.* AlphaFold2-based prediction of the co-condensation propensity of proteins. *Proc Natl Acad Sci USA* 2024;**121**:e2315005121. <https://doi.org/10.1073/pnas.2315005121>
 170. Daza D, Alivanistos D, Mitra P *et al.* BioBLP: a modular framework for learning on multimodal biomedical knowledge graphs. *J Biomed Semantics* 2023;**14**:20. <https://doi.org/10.1186/s13326-023-00301-y>
 171. Li MM, Huang Y, Sumathipala M *et al.* Contextual AI models for single-cell protein biology. *Nat Methods* 2024;**21**:1546–57. <https://doi.org/10.1038/s41592-024-02341-3>
 172. Jiang L, Yang X, Guo X *et al.* Graph neural network integrated with pretrained protein language model for predicting human-virus protein-protein interactions. *Brief Bioinform* 2025;**26**:bbaf461. <https://doi.org/10.1093/bib/bbaf461>
 173. Wu F, Wu L, Radev D *et al.* Integration of pre-trained protein language models into geometric deep learning networks. *Commun Biol* 2023;**6**:876. <https://doi.org/10.1038/s42003-023-05133-1>
 174. Chen RJ, Lu MY, Wang J *et al.* Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imaging* 2022;**41**:757–70. <https://doi.org/10.1109/tmi.2020.3021387>
 175. Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. <https://doi.org/10.1126/science.ade2574>
 176. Nijkamp E, Ruffolo JA, Weinstein EN *et al.* ProGen2: exploring the boundaries of protein language models. *Cell Syst* 2023;**14**:968–978.e3. <https://doi.org/10.1016/j.cels.2023.10.002>

177. Argelaguet R, Velten B, Arnol D *et al.* Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018;**14**:e8124. <https://doi.org/10.15252/msb.20178124>
178. Zhang B, Wang J, Wang X *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014;**513**:382–7. <https://doi.org/10.1038/nature13438>
179. Oka M, Otani M, Miyamoto Y *et al.* Phase-separated nuclear bodies of nucleoporin fusions promote condensation of MLL1/CRM1 and rearrangement of 3D genome structure. *Cell Rep* 2023;**42**:112884. <https://doi.org/10.1016/j.celrep.2023.112884>
180. Dos Passos PM, Hemamali EH, Mamede LD *et al.* RNA-mediated ribonucleoprotein assembly controls TDP-43 nuclear retention. *PLoS Biol* 2024;**22**:e3002527. <https://doi.org/10.1371/journal.pbio.3002527>
181. Gonzalez-Martinez D, Roth L, Mumford TR *et al.* Oncogenic EML4-ALK assemblies suppress growth factor perception and modulate drug tolerance. *Nat Commun* 2024;**15**:9473. <https://doi.org/10.1038/s41467-024-53451-7>
182. Dai Y, Lei C, Zhang Z *et al.* Amyloid-beta targeted therapeutic approaches for Alzheimer's disease: Long road ahead. *Curr Drug Targets* 2022;**23**:1040–56. <https://doi.org/10.2174/1389450123666220421124030>
183. Patel A, Lee HO, Jawerth L *et al.* A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* 2015;**162**:1066–77. <https://doi.org/10.1016/j.cell.2015.07.047>
184. Hallegger M, Chakrabarti AM, Lee FCY *et al.* TDP-43 condensation properties specify its RNA-binding and regulatory repertoire. *Cell* 2021;**184**:4680–4696.e22. <https://doi.org/10.1016/j.cell.2021.07.018>
185. Farahi N, Lazar T, Tompa P *et al.* Phase-separating fusion proteins drive cancer by upsetting transcription regulation. *Genome Biol* 2025;**26**:330. <https://doi.org/10.1186/s13059-025-03787-0>
186. Kitajima H, Maruyama R, Niinuma T *et al.* TM4SF1-AS1 inhibits apoptosis by promoting stress granule formation in cancer cells. *Cell Death Dis* 2023;**14**:424. <https://doi.org/10.1038/s41419-023-05953-3>
187. Zhu L, Richardson TM, Wacheul L *et al.* Controlling the material properties and rRNA processing function of the nucleolus using light. *Proc Natl Acad Sci USA* 2019;**116**:17330–5. <https://doi.org/10.1073/pnas.1903870116>
188. Riback JA, Eeftens JM, Lee DSW *et al.* Viscoelasticity and advective flow of RNA underlies nucleolar form and function. *Mol Cell* 2023;**83**:3095–3107.e9. <https://doi.org/10.1016/j.molcel.2023.08.006>
189. Duan J, Xiong J, Li Y *et al.* Deep learning based multimodal biomedical data fusion: an overview and comparative review. *Inf Fusion* 2024/12/01/ 2024;**112**:102536. <https://doi.org/10.1016/j.inffus.2024.102536>
190. Shin Y, Berry J, Pannucci N *et al.* Spatiotemporal control of intracellular phase transitions using light-activated optoDroplets. *Cell* 2017;**168**:159–171.e14. <https://doi.org/10.1016/j.cell.2016.11.054>
191. Binan L, Jiang A, Danquah SA *et al.* Simultaneous CRISPR screening and spatial transcriptomics reveal intracellular, intercellular, and functional transcriptional circuits. *Cell* 2025;**188**:2141–2158.e18. <https://doi.org/10.1016/j.cell.2025.02.012>
192. Pinette NC, Terrado M, Bui JM *et al.* Next-generation predictors of protein phase behavior. *Curr Opin Struct Biol* 2025;**96**:103197. <https://doi.org/10.1016/j.sbi.2025.103197>