Received: 21 September 2023 | Revised: 29 November 2023 | Accepted: 3 December 2023

DOI: 10.1002/pro.4858

TOOLS FOR PROTEIN SCIENCE

THE PROTEIN SOCIETY WILEY

MLOsMetaDB, a meta-database to centralize the information on liquid–liquid phase separation proteins and membraneless organelles

Fernando Orti[1] | **María Laura Fernández**[2,3] | **Cristina Marino-Buslje**[1]

[1]Leloir Institute/IIBBA, Buenos Aires, Argentina

[2]Instituto de física interdisciplinaria y aplicada (IFNIA), Universidad de Buenos Aires, Argentina

[3]Dto. de Física. Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina

Correspondence
Cristina Marino-Buslje, Leloir Institute/IIBBA, Buenos Aires, Argentina.
Email: cmb@leloir.org.ar

Funding information
Consejo Nacional de Investigaciones Científicas y Técnicas

Review Editor: Aitziber L. Cortajarena

Abstract

Over the past few years, there has been a focus on proteins that create separate liquid phases in the intracellular liquid environment, known as membraneless organelles (MLOs). These organelles allow for the spatiotemporal associations of macromolecules that dynamically exchange within the cellular milieu. They provide a form of compartmentalization crucial for organizing key functions in many cells. Metabolic processes and signaling pathways in both the cytoplasm and nucleus are among the functions performed by MLOs, which are facilitated by diverse combinations of proteins and nucleic acids. However, disruptions in these liquid–liquid phase separation processes (LLPS) may lead to several diseases, such as neurodegenerative disorders and cancer, among others. To foster the study of this process and MLO function, we present *MLOsMetaDB* (http://mlos.leloir.org.ar), a comprehensive resource of information on MLO- and LLPS-related proteins. Our database integrates and centralizes available information for every protein involved in MLOs, which is otherwise disseminated across a plethora of different databases. Our manuscript outlines the development and features of MLOsMetaDB, which provides an interactive and user-friendly environment with modern biological visualizations and easy and quick access to proteins based on LLPS role, MLO location, and organisms. In addition, it offers an advanced search for making complex queries to generate customized information. Furthermore, MLOsMetaDB provides evolutionary information by collecting the orthologs of every protein in the same database. Overall, MLOsMetaDB is a valuable resource as a starting point for researchers studying the many processes driven by LLPS proteins and membraneless organelles.

KEYWORDS

biocondensates, drivers, LLPS, membraneless organelles, MLO, molecular condensates, phase separation

1 | INTRODUCTION

Membraneless organelles (*MLOs*) or biomolecular condensates are dynamic subcellular compartments that lack surrounding lipidic membranes (Banani et al., 2017). *MLOs* are formed through a liquid–liquid phase separation phenomenon (*LLPS*) by which proteins and nucleic acid molecules self-assemble into liquid droplets maintained in a condensed phase through multivalent interactions, mostly transients and dynamics (Brangwynne et al., 2009). The proteins or nucleic acids that play a direct role in the formation and/or the stability of the MLOs through the LLPS process, are referred to as *drivers*. The components that do not actively participate in the formation but are present in the MLO under specific conditions, are known as *clients*. Proteins that can alter the formation, composition, dissolution, or stability of MLOs are classified as *regulators* (Farahi et al., 2021).

Recently, several primary LLPS-dedicated databases became available with curated and not curated datasets of proteins. These databases are *PhaSePro* (Mészáros et al., 2020), *PhaSepDB* (You et al., 2020), *DrLLPS* (Ning et al., 2020), *LLPSDB* (Li et al., 2020), *RNAPhaSep* (Zhu et al., 2021), and *RSP* (Liu et al., 2021). The last two also collect RNAs involved in the LLPS process.

In a previous study by our group, it was found that current LLPS-dedicated databases differ in their focus, inclusion criterion, bias in the included protein, annotation, incompleteness, and curation levels, among other characteristics, making the analysis difficult and the generation of new knowledge challenging (Orti et al., 2021).

To address this issue, we developed *MLOsMetaDB* (http://mlos.leloir.org.ar/), a comprehensive and centralized resource of information on MLO- and LLPS-related proteins.

MLOsMetaDB integrates and centralizes protein information of the primary LLPS/MLOs databases and enriches their annotations with functional, structural, and evolutionary information, focusing on molecular features associated with the LLPS process. In such a way, it concentrates available information disseminated in a plethora of different databases for every protein involved in MLOs, which is difficult to embrace when one or a group of proteins are the focus of a study.

Our database provides an interactive and user-friendly environment, modern biological visualizations, and easy and quick access to proteins by LLPS role, MLO location, and organism. It also provides an advanced search to make complex queries and generate customized information.

Also, it allows the scientific community to easily create different subsets of interest to carry out further analyses.

Finally, it provides evolutionary information by collecting the orthologous proteins by protein in the database. This is a key resource for transferring knowledge to other species because the LLPS protein databases' limitation is that they are highly enriched in human proteins and, in total, contain very few organisms, the most studied or model ones.

2 | RESULTS

2.1 | Protein entries

The MLOsMetaDB dataset is composed of *12,038* unique proteins obtained from the four LLPS/MLOs databases. We retrieved *121* proteins from PhaSePro, *373* from LLPSDB, *5191* from PhaSepDB, and *9251* from DrLLPS (Figure 1a). From the total, *8365* proteins are annotated in one database and *2504* and *113* proteins in two and three databases, respectively. It is worth noting that the overlap between DBs is very small. Only *56* proteins of the dataset are present in the four databases, highlighting the importance of this effort to centralize the data (Figure 1b).

The imbalance in the number of annotated proteins between the different databases can be explained by the different inclusion and curation criteria of each database. PhaSePro and LLPSDB only collect manually curated proteins with experimental evidence of being LLPS drivers, while DrLLPS and PhaSepDB also include proteins from high-throughput experiments with no evidence of LLPS behavior.

MLOsMetaDB contains annotations of *780* LLPS driver proteins, *10,887* clients, and *986* regulators. There are *238* proteins annotated as driver and client, *34* as driver and regulator, and *267* as client and regulator, 39 proteins are annotated as driver, client, and regulators (see Figure 1c,d). The overlap between LLPS roles may be due to the fact that a protein can have different roles in different MLOs or be miss classified in different databases. For instance, the human protein PCNT (Uniprot Acc: O95613) is annotated as a client protein within the centrosome in DrLLPS, whereas PhaSepDB designates it as a driver protein within the same MLO. Another example is proteins annotated with different roles in different organelles. Figure 1d illustrates the organisms' distribution in the DB.

Considering the biology of MLOs, it is to be expected that there will be a greater number of client proteins recruited by MLOs than by driver proteins. To characterize a protein as a driver, stronger experimental evidence is required, therefore, some proteins that are currently

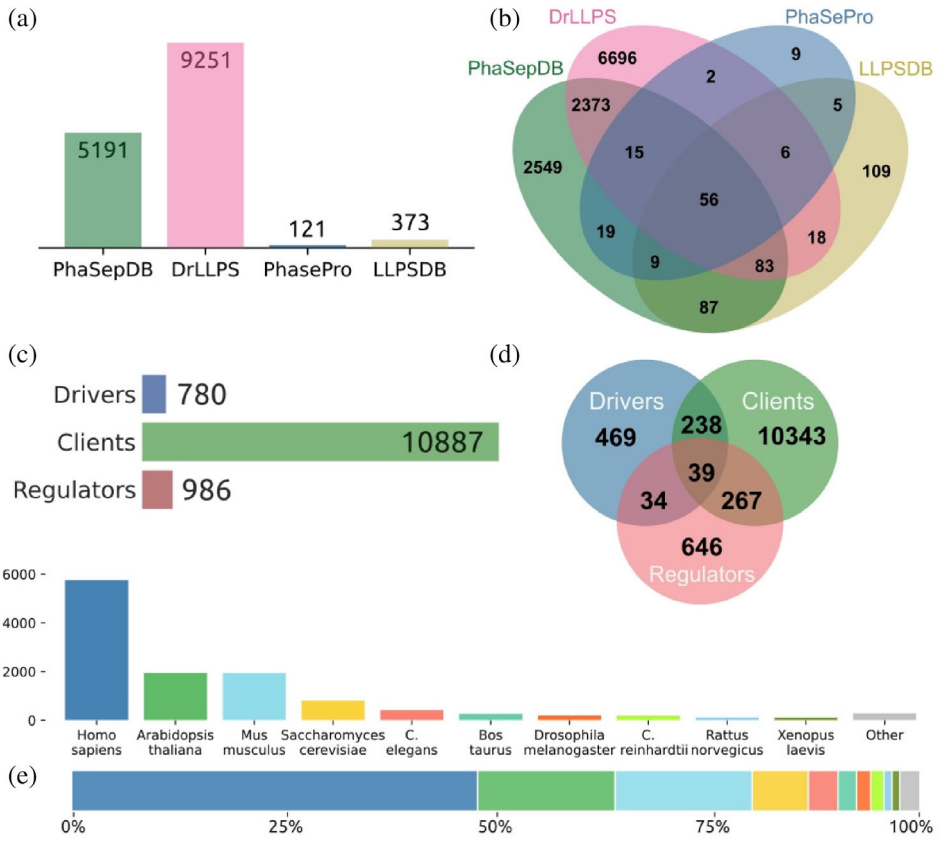

FIGURE 1 MLOsMetaDB statistics. (a) Number of proteins in MLOsMetaDB categorized by their source database. (b) Venn diagram illustrating protein overlap between databases. (c) Number of proteins in MLOsMetaDB classified by LLPS role annotation. (d) Venn diagram showing protein overlap between LLPS roles. (e) Distribution of proteins by organism: number (top) and percentage (bottom).

annotated as clients may, in the future, be reclassified as drivers.

2.1.1 | Organisms distribution

Direct experimental data is available for a limited number of proteins, and there is a strong bias toward the most studied proteins and organisms. Human proteins are 47%, while the rest are distributed among a few organisms, all of them well-studied or model ones (Figure 1e). This is a limiting point since data for less popular organisms or proteins is still lacking. Having experimental support is labor intensive, time-consuming and costly, so for some proteins, it is not possible to have this information unless transferring it by homology.

Instead, by gathering orthologous proteins, starting from *12,038* we reach *118,433* other proteins in *2408* different organisms, in such a way, enriching them with valuable information to better understand their biological roles.

2.2 | Web usage

MLOsMetaDB has been designed to facilitate users' research and navigation.

The basic search is with the Uniprot accession or Gene Name. Also, in the *main page* there are quick accesses to different datasets such as a particular MLO, LLPS role, and the principal model organisms.

The *advanced search* allows more complex queries filtering by protein identifiers, annotation type, PFAM domains, LLPS roles, LLPS source databases, and molecular features related to the LLPS process and all the possible combinations of filters (as an example: mouse driver proteins, present in SG and Cajal Bodies with 30% of disorder content), among other possibilities (Figure 2).

The *results page* offers a quick and nice view of the general features of a protein (LLPS Rol), organism, molecular features, intrinsically disordered regions, low complexity regions, and database sources, among others (Figure 3). Results can be displayed in cards or table mode, filtered by columns and can be downloaded as JSON or TSV formats.

On the *protein page*, users will find detailed information about the protein as its sequence, protein and gene names, biological function (when available), LLPS resource database (crosslinked), an interactive feature viewer, and sequence viewer with molecular features (Figure 4). It also provides a list of MLOs where the protein is associated. Users will also find the predicted structure by AlphaFold2 with an interactive visualization. For

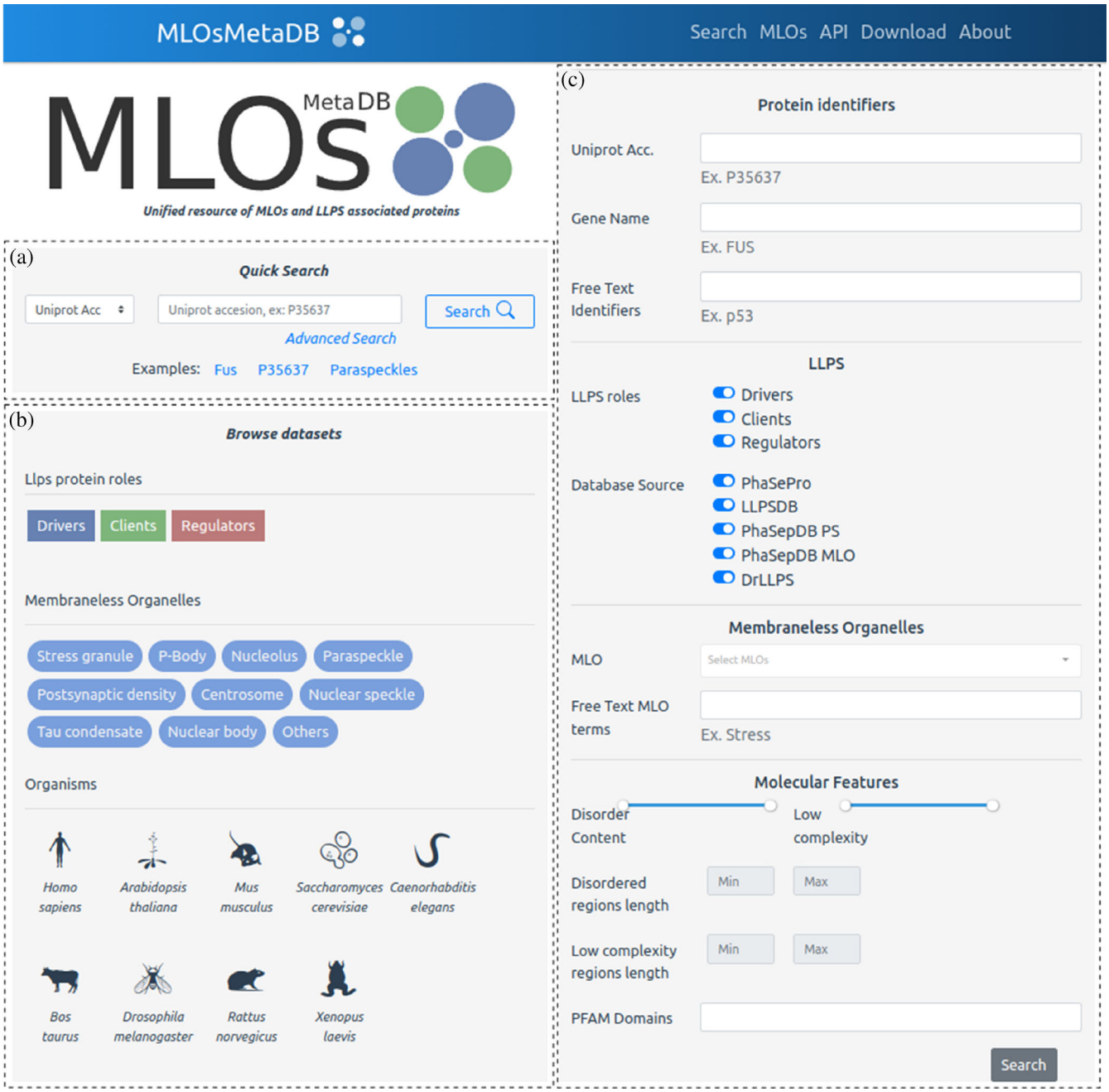

FIGURE 2 MLOsMetaDB's main page. (a) Quick search by Uniprot Acc, Gene Name, and MLO. (b) Browse datasets by LLPS role, MLO, and model organisms. (c) Advanced search: In this section, users can realize more complex searches, combining different fields, including LLPS roles, LLPS source databases, PFAM domains, one or more membraneless organelles (exact or free text search). Also, the sequence percent of disorder and low complexity content and lengths of the segments can be specified.

driver proteins, a table of orthologs is provided. All the information is downloadable in JSON format.

In addition, MLOsMetaDB offers an application programming interface (API) that allows users to programmatically query and download data, enabling complex searches and facilitating integration into analysis pipelines. The API is thoroughly documented, with detailed explanations of available paths and parameters. Swagger UI is used as a tool for testing queries and results.

3 | DISCUSSION

Currently, there are few primary databases associated with LLPS and MLO with differences in methodology,

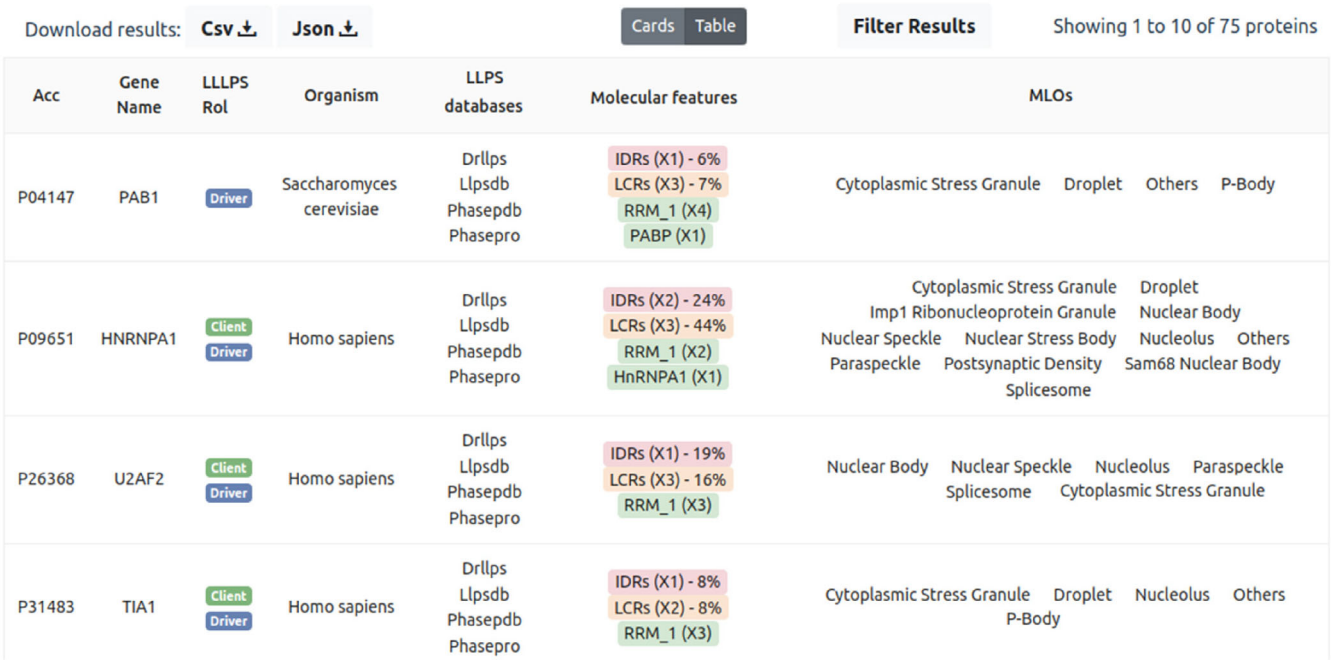

FIGURE 3 Results page. Example of results page for a query of driver proteins containing RRM_1 domains. Each record shows information related to the LLPS role, LLPS dataset source, molecular features (phase separation, intrinsically disordered regions, low complexity regions), and MLOs where the protein can be found.

objectives, level of curation, number of proteins, and nonuniform, poor, or missing annotation for a limited number of proteins, making further analysis difficult.

In response, MLOsMetaDB offers the scientific community centralized and unified information and predictions on LLPS- and MLO-associated proteins, supporting both simple and complex searches. The platform also provides a modern, interactive website for ease of use by non-bioinformatics users. Moreover, MLOsMetaDB includes an API that integrates access, search, and download of LLPS/MLOs protein information into a pipeline, streamlining analyses.

4 | MATERIALS AND METHODS

4.1 | LLPS and MLOs associated proteins

Proteins were collected from four LLPS- and MLOs-related databases: *PhaSePro* (Mészáros et al., 2020), *PhaSepDB* (You et al., 2020), *DrLLPS* (Ning et al., 2020), and *LLPSDB* (Li et al., 2020; Ning et al., 2020). Entries were integrated and unified by Uniprot Accession when available.

Disorder-related annotations and predictions were retrieved from MobiDB (Piovesan et al., 2021), while structural annotations were obtained from *PFAM*, *PDB*, and *AlphaFold2 DB* (Varadi et al., 2021). Orhologs and evolutionary annotations were retrieved from *OmaDB* (Altenhoff et al., 2021).

4.2 | Biological role assessment

Proteins within the MLOs were categorized as drivers, clients, and regulators depending on the annotations in their source DB and/or when clear evidence was found in different datasets or literature.

Drivers are proteins retrieved from PhaSePro, LLPSDB, phase separation dataset from PhaSepDB, and proteins annotated as scaffolds in DrLLPS.

Clients are proteins retrieved from the MLO dataset from PhaSepDB and proteins annotated as Clients in DrLLPS.

Regulators are proteins annotated as regulators in DrLLPS.

4.3 | Data processing

Data was downloaded and processed using Python 3.8 and Pandas package; graphs were prepared with Matplotlib 3.5.1 and Seaborn 3.5.1 libraries.

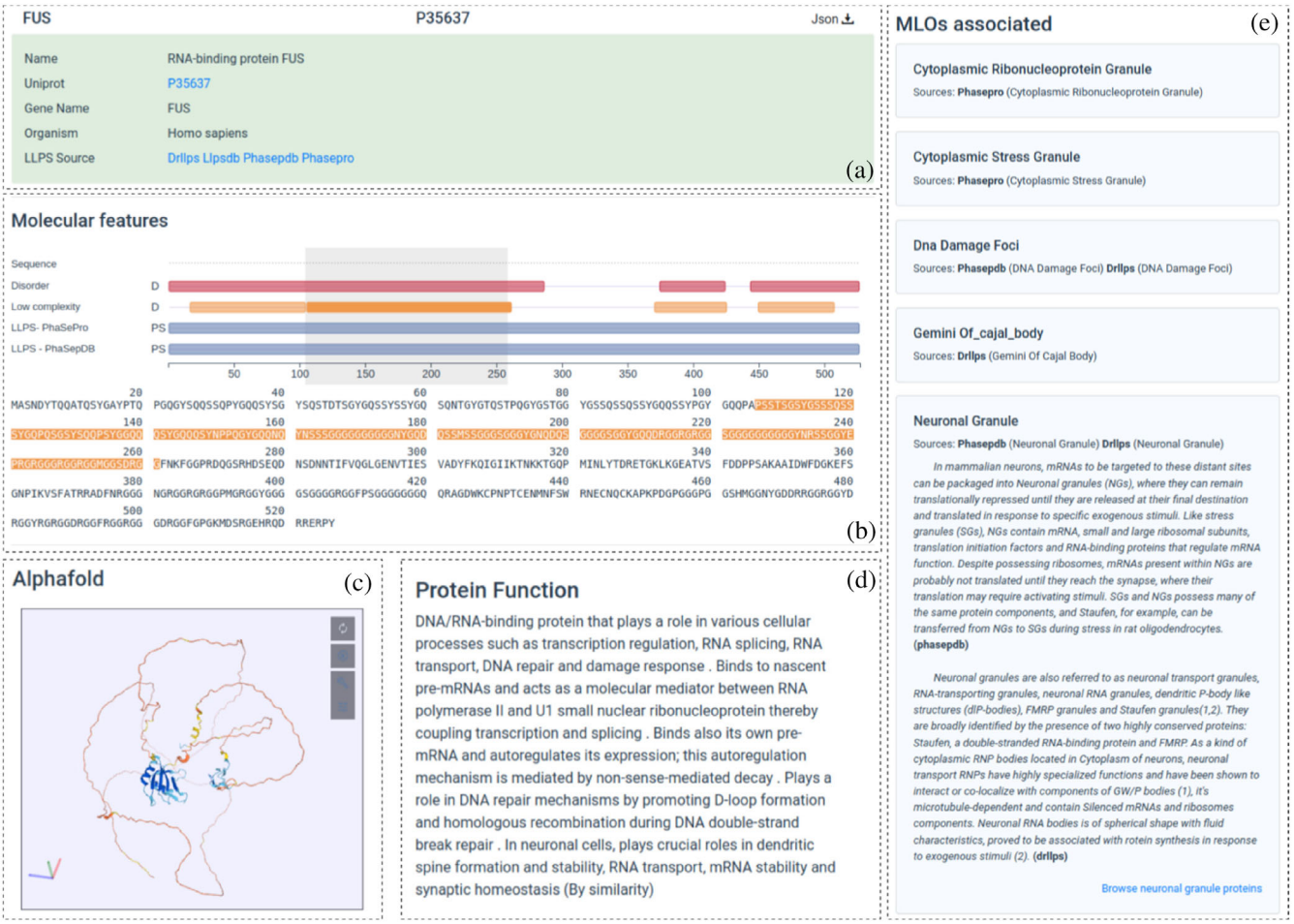

FIGURE 4 Protein page. (a) Protein information and related links to the source databases. (b) Molecular features, interactive visualizations of LLPS-related molecular features (phase separation promoting regions, intrinsically disordered regions, low complexity regions, Prion-like domains, PFAM domains). (c) AlphaFold2 predicted protein structure visualization. (d) Protein function (source UniProt). Protein biological function. (e) Membraneless organelles are associated with the protein. Each MLO has information related to the source database, a brief description of the MLO when available and a link to directly browse proteins associated with the MLO.

4.4 | Web server implementation

MLOsMetaDB is stored in a MongoDB database, and its backend was developed using Flask framework (Version 1.1.2). The website is implemented using VueJS and Bootstrap 5. The RestFul API and its documentation were developed using Swagger Ui Library. Interactive biological visualizations include implementations of ProSeqViewer (Bevilacqua et al., 2021), Feature Viewer (Paladin et al., 2020), and Mol* Viewer (Sehnal et al., 2021).

AUTHOR CONTRIBUTIONS

Cristina Marino-Buslje: Conceptualization; investigation; funding acquisition; writing – original draft; visualization; writing – review and editing; formal analysis; project administration; supervision; resources. **Fernando Orti:** Conceptualization; investigation; methodology; software; data curation; formal analysis; writing – original draft; validation; visualization. **María Laura Fernández:** Validation; data curation; writing – original draft; writing – review and editing; methodology.

ACKNOWLEDGMENTS

Fernando Orti is granted by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), María Laura Fernández and Cristina Marino-Buslje are researchs of CONICET.

ORCID

Fernando Orti https://orcid.org/0000-0002-1757-6693
Cristina Marino-Buslje https://orcid.org/0000-0002-6564-1920

REFERENCES

Altenhoff AM, Train CM, Gilbert KJ, Mediratta I, Mendes de Farias T, Moi D, et al. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. Nucleic Acids Res. 2021;49(D1):D373–9. https://doi.org/10.1093/nar/gkaa1007

Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. Nat Rev Mol Cell Biol. 2017;18(5):285–98.

Bevilacqua M, Paladin L, Tosatto SCE, Piovesan D. ProSeqViewer: an interactive, responsive and efficient TypeScript library for visualization of sequences and alignments in web applications. Bioinformatics. 2021;38(4):1129–30.

Brangwynne CP, Eckmann CR, Courson DS, Rybarska A, Hoege C, Gharakhani J, et al. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. Science. 2009;324(5935):1729–32.

Farahi N, Lazar T, Wodak SJ, Tompa P, Pancsa R. Integration of data from liquid-liquid phase separation databases highlights concentration and dosage sensitivity of LLPS drivers. Int J Mol Sci. 2021;22(6):3017. https://doi.org/10.3390/ijms22063017

Li Q, Peng X, Li Y, Tang W, Zhu J, Huang J, et al. LLPSDB: a database of proteins undergoing liquid-liquid phase separation in vitro. Nucleic Acids Res. 2020;48(D1):D320–7. https://doi.org/10.1093/nar/gkz778

Liu M, Li H, Luo X, Cai J, Chen T, Xie Y, et al. RPS: a comprehensive database of RNAs involved in liquid–liquid phase separation. Nucleic Acids Res. 2021;50(D1):D347–55.

Mészáros B, Erdős G, Szabó B, Schád É, Tantos Á, Abukhairan R, et al. PhaSePro: the database of proteins driving liquid-liquid phase separation. Nucleic Acids Res. 2020;48(D1):D360–7. https://doi.org/10.1093/nar/gkz848

Ning W, Guo Y, Lin S, Mei B, Wu Y, Jiang P, et al. DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. Nucleic Acids Res. 2020;48(D1):D288–95. https://doi.org/10.1093/nar/gkz1027

Orti F, Navarro AM, Rabinovich A, Wodak SJ, Marino-Buslje C. Insight into membraneless organelles and their associated proteins: drivers, clients and regulators. Comput Struct Biotechnol J. 2021;19:3964–77. https://doi.org/10.1016/j.csbj.2021.06.042

Paladin L, Schaeffer M, Gaudet P, Zahn-Zabal M, Michel PA, Piovesan D, et al. The feature-viewer: a visualization tool for positional annotations on a sequence. Bioinformatics. 2020;36(10):3244–5. https://doi.org/10.1093/bioinformatics/btaa055

Piovesan D, Necci M, Escobedo N, Monzon AM, Hatos A, Mičetić I, et al. MobiDB: intrinsically disordered proteins in 2021. Nucleic Acids Res. 2021;49(D1):D361–7. https://doi.org/10.1093/nar/gkaa1058

Sehnal D, Bittrich S, Deshpande M, Svobodová R, Berka K, Bazgier V, et al. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res. 2021;49(W1):W431–7. https://doi.org/10.1093/nar/gkab314

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2021;50(D1):D439–44.

You K, Huang Q, Yu C, Shen B, Sevilla C, Shi M, et al. PhaSepDB: a database of liquid-liquid phase separation related proteins. Nucleic Acids Res. 2020;48(D1):D354–9. https://doi.org/10.1093/nar/gkz847

Zhu H, Fu H, Cui T, Ning L, Shao H, Guo Y, et al. RNAPhaSep: a resource of RNAs undergoing phase separation. Nucleic Acids Res. 2021;50(D1):D340–6.

How to cite this article: 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 Science. 2024;33(1):e4858. https://doi.org/10.1002/pro.4858