Brief Communication

CD-CODE: crowdsourcing condensate database and encyclopedia

Nadia Rostam [1,2,5], Soumyadeep Ghosh [1,2,5], Chi Fung Willis Chow [1,2,3], Anna Hadarovich[1,2], Cedric Landerer [1,2], Rajat Ghosh[1,2], HongKee Moon [1], Lena Hersemann [1], Diana M. Mitrea[4], Isaac A. Klein[4], Anthony A. Hyman [1,2,3] & Agnes Toth-Petroczy [1,2,3] ✉

The discovery of biomolecular condensates transformed our understanding of intracellular compartmentalization of molecules. To integrate interdisciplinary scientific knowledge about the function and composition of biomolecular condensates, we developed the crowdsourcing condensate database and encyclopedia (cd-code.org). CD-CODE is a community-editable platform, which includes a database of biomolecular condensates based on the literature, an encyclopedia of relevant scientific terms and a crowdsourcing web application. Our platform will accelerate the discovery and validation of biomolecular condensates, and facilitate efforts to understand their role in disease and as therapeutic targets.

Biomolecular condensates are membraneless organelles that selectively concentrate biomolecules (for example, proteins and nucleic acids) in the cell, with spatial and temporal precision[1]. In recent years, their role was implicated in several biochemical processes, in physiology and disease[2]. Consequently, biomolecular condensates are now leveraged as a new class of therapeutic targets[3,4].

Basic science and drug discovery advances build upon published reports and the rate of new discoveries depends on timely accessibility to relevant data. However, as with every novel paradigm, new terms and concepts emerge and evolve as the field develops. Accordingly, currently available databases which catalog proteins involved in condensate formation use various definitions and criteria to define condensates and their constituent proteins and RNAs[5–8]. These are excellent databases curating proteins that phase separate. Specifically, LLPSDB[7] and PhaSePro[6] collect proteins that are thought to drive liquid–liquid phase separation, with the former curating exclusively in vitro data.

However, these databases do not answer the following questions regarding biomolecular condensates: What are the biomolecular condensates discovered and verified to date? What are their known protein components? Which condensates is a given protein known to belong to? What are the experimental evidences supporting the existence of a particular condensate? Our goal is to generate answers for these and other important questions, and to create a community-editable database to facilitate the dynamic data updates. Therefore, we designed a condensate-centric database, which is based on the scientific literature, and provides experimental evidences, scores and references for each condensate–protein relationship (Extended Data Figs. 1 and 2). This database is updated dynamically by contributors to keep up with the growing knowledge in the field. We call our platform CD-CODE, which consists of three main parts: (1) a database of biomolecular condensates and their protein constituents; (2) an encyclopedia for the scientific terms used in condensate biology; and (3) a crowdsourcing web application (Extended Data Fig. 3).

CD-CODE is a 'living database' designed for dynamic and rapid addition and review of information about condensates and proteins by users (Fig. 1) and is open to any expert researcher who wishes to contribute. Our user management system supports three types of users: viewers, contributors and maintainers. Viewers can read and download the curated information. Contributors can suggest edits and propose new condensate and protein entries (Extended Data Figs. 4 and 5). Maintainers are part of the development team, who curate the changes and accept or reject suggestions by contributors, who are then notified about the status of their suggestions and can engage in further discussion. To keep up with the rapidly evolving definitions, nomenclature and growing scientific evidence, the crowdsourcing

[1]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. [2]Center for Systems Biology Dresden, Dresden, Germany. [3]Cluster of Excellence Physics of Life, TU Dresden, Dresden, Germany. [4]Dewpoint Therapeutics, Boston, MA, USA. [5]These authors contributed equally: Nadia Rostam, Soumyadeep Ghosh. ✉e-mail: toth-petroczy@mpi-cbg.de
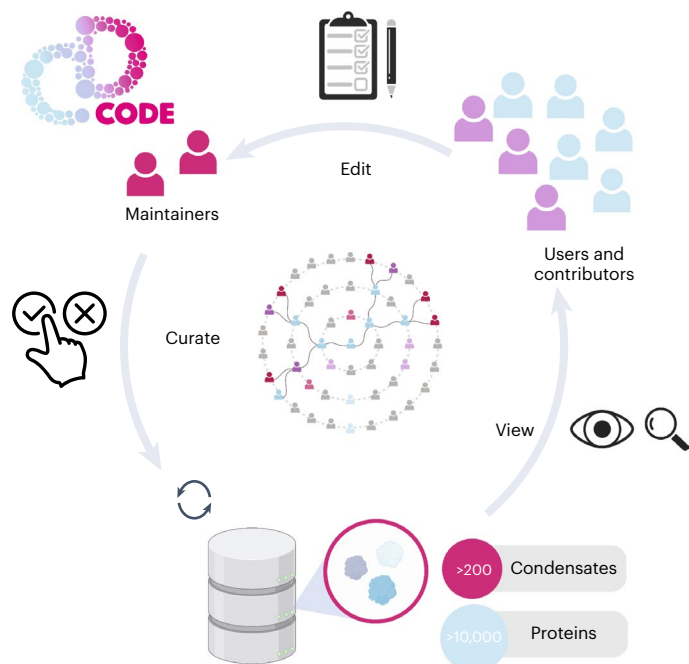
Fig. 1 | CD-CODE information flow. Users can view and search the data, or become contributors after registration and edit the content of the database via the community-editable web application. The maintainers assure quality control and only approved edits will be part of the dynamically updated database. Figure created with BioRender.com.

platform allows the community to aggregate scientific findings in condensate biology.

At the time of this report, CD-CODE (cd-code.org) contains 9,861 proteins linked to 244 unique biomolecular condensates (and 375 in vitro synthetic condensates) across 49 different organisms. Notably, these numbers are continuously changing as contributors add and review more data. CD-CODE, as a semi-manually curated and annotated resource, aggregates information from the primary literature (to date, PubMed references published until 1 June 2022 were manually curated) and other databases[5–8] (Extended Data Fig. 6 and Extended Data Tables 1 and 2). To promote easy integration with other resources, protein entries are cross-referenced with UniProt[9], Ensembl[10] and the Human Protein Atlas (proteinatlas.org)[11]. Common sequence properties of condensate proteins are also displayed graphically, such as disorder score[12] and amino acid composition (Extended Data Fig. 7), facilitating the identification of regions that may drive condensate partitioning.

We standardized the names of condensates by creating an ontology from the literature and grouped condensates by functional categories (Supplementary Table 1) to reveal the evolutionary history of condensates. Most known condensates are found in mammals and many are clade-specific (Fig. 2a). Since our current knowledge is sparse and likely biased, the evolutionary origin of condensates remains an open future research direction that CD-CODE can facilitate.

While many proteins undergo liquid–liquid phase separation in vitro, it is unclear which proteins form condensates in cells and which condensates they partition into. To facilitate our understanding of condensate-specificity of proteins, we collected all known condensates a given protein was found in, and we curated the experimental evidence for association of each protein with a given condensate (confidence score, corresponding to zero to five stars: 1 star: literature evidence, PubMed identifier (ID); 2 stars, high-throughput; 3 stars, in vitro; 4 stars, in cellulo; and 5 stars, in vivo evidence). Condensates and proteins that have zero or one star rating have not been manually curated yet.

As expected, for dynamic cellular compartments, many proteins partition into different condensates and the overlap between

condensate proteomes is substantial (Fig. 2b). While proteins may localize to multiple condensates (members), a few are obligate and essential components (drivers). We annotated 205 driver proteins in specific condensates, providing the corresponding experimental evidences. Our database revealed that several proteins that are drivers in one condensate are nonessential members of another (for example, G3BP1, a driver of stress granules, is also present in processing bodies (P-bodies) and neuronal ribonucleoprotein particle granules). CD-CODE will aid our understanding of the determinants of condensate-specific driver behavior, and whether a driver protein can be used as a 'marker' of a condensate in experiments.

Marker proteins are used to define the identity of the condensates and inform designing of condensate-targeting drug screening pipelines[3]. They are thought to be uniquely associated with a given condensate, and are commonly used to visualize condensates using microscopy, for example, in colocalization experiments to prove the localization of proteins into condensates. Our database revealed that several known marker proteins are not specific to a condensate. For example, whilst DCP1A is used as a marker for P-bodies, it also localizes to stress granules and nucleoli. Knowing specific protein components will facilitate the experimental design for accurate, specific identification of condensates.

CD-CODE enables us to answer the questions posed at the beginning: (1) there are currently 136 unique biomolecular condensates documented in the literature; (2) as an example, P-granules, which are the germ granules of *Caenorhabditis elegans*, have 190 documented protein components: one of them, pgl-3 (PGL3_CAEEL), is a driver for P-granule formation, and its presence within P-granules is supported by in vivo experimental evidence (5 star). Pgl-3 is exclusively reported to be associated with P-granules; thus, it is a P-granule-specific marker protein.

Databases that curate proteins undergoing liquid–liquid phase separation have facilitated the development of machine learning algorithms to predict phase separation[13–15] and the discovery of what protein properties drive phase separation[15]. The next open question is which biomolecular condensate does a specific protein belong to. Our database contains a curated list of condensate proteomes (Fig. 2c), which can facilitate investigations of protein recruitment into specific condensates. Our resource can provide high-quality benchmarking data for machine learning algorithms aimed at predicting the protein components of condensates.

Furthermore, our comprehensive curation of condensate types and their respective composition in multiple species, and the level of experimental support, provides a valuable resource for drug hunters, which can inform the design of assays and screening pipelines. For example, in high-content imaging phenotypic screens, it is desired that the protein or protein combination chosen to be monitored is/ are selective for the target condensate[3]. Additionally, through regular updating of the database by the community and via curation of new publications, CD-CODE supports and accelerates nomination of new condensate-associated drug targets.

The field of biomolecular condensates is highly transdisciplinary and ever-developing, where definitions and terms keep changing, creating a need for constant updates that require consensus within the community. The encyclopedia, as a standalone wiki, serves as a platform to aggregate knowledge about condensate research. In the future, we are planning weekly updates to integrate new data from the users, and yearly updates with new features and data points that become relevant to store, as the research field develops.

The main feature of CD-CODE is that it contains experimentally validated entries. However, caution should be exercised by users when interpreting lack of data on a particular protein, condensate or species, as this may simply reflect the biased interest of the community towards particular model systems and biological pathways. Any missing information could mean that (1) the protein or condensate has not
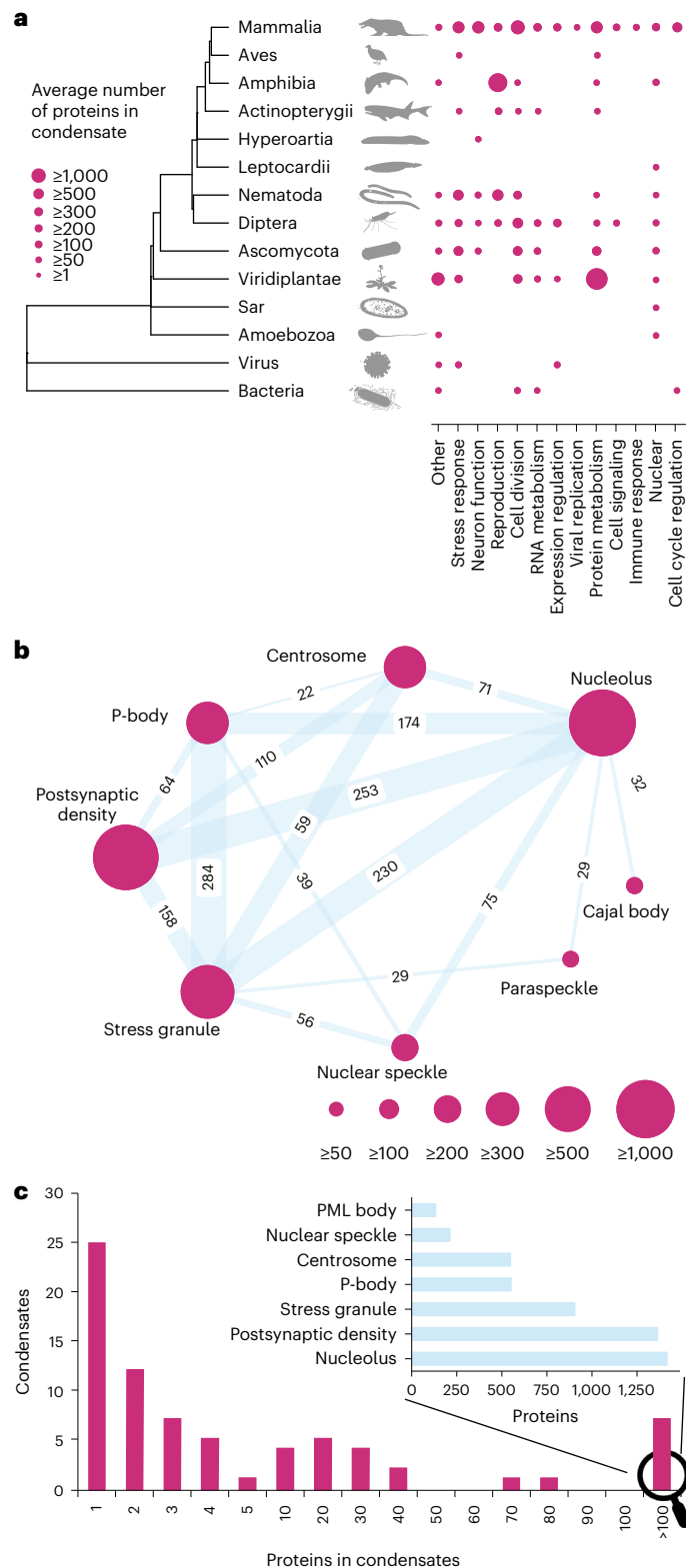
Fig. 2 | Selected features available at CD-CODE. a, Biomolecular condensates across the tree of life. CD-CODE contains information about 244 condensates across 49 species. Here, only major clades are shown for clarity and condensates were grouped into functional categories. **b**, Many proteins localize to multiple condensates. There is a large overlap between the proteomes of different biomolecular condensates in humans. The largest condensates in humans are represented as circles and the shared proteins between every two condensates are shown (only condensates with >20 connections are shown). **c**, The distribution of condensate proteome sizes in humans. Most biomolecular condensates have a few known protein members. The largest condensates contain >1,000 different proteins (inset).

been studied yet; (2) there is a research paper but the information has not been added to the database yet; (3) the condensate truly does not exist; or (4) the protein truly does not belong to a given condensate. As such, CD-CODE aims to highlight the unknowns in the field to guide future research questions to fill the gaps. These gaps in experimental evidence can be bridged by computational predictions[16,17], which are beyond the scope of CD-CODE. Evolution of the CD-CODE database through ongoing curation of new experimental evidence will lead to a progressive increase in high-scoring condensate entries.

In summary, we present CD-CODE, a semi-manually curated condensate database, and a community-editable web application. The crowdsourcing platform allows the community to further scrutinize definitions and evidence as the field evolves. This will ensure that the ever-growing knowledge on condensate research is integrated into the database and into the encyclopedia in a timely manner.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-023-01831-0.

References

1. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).
2. Alberti, S. & Dormann, D. Liquid–liquid phase separation in disease. *Annu. Rev. Genet.* **53**, 171–194 (2019).
3. Mitrea, D. M., Mittasch, M., Gomes, B. F., Klein, I. A. & Murcko, M. A. Modulating biomolecular condensates: a novel approach to drug discovery. *Nat. Rev. Drug Discov.* **21**, 841–862 (2022).
4. Conti, B. A. & Oppikofer, M. Biomolecular condensates: new opportunities for drug discovery and RNA therapeutics. *Trends Pharmacol. Sci.* **43**, 820–837 (2022).
5. Ning, W. et al. DrLLPS: a data resource of liquid–liquid phase separation in eukaryotes. *Nucleic Acids Res.* **48**, D288–D295 (2020).
6. Mészáros, B. et al. PhaSePro: the database of proteins driving liquid–liquid phase separation. *Nucleic Acids Res.* **48**, D360–D367 (2020).
7. Li, Q. et al. LLPSDB: a database of proteins undergoing liquid–liquid phase separation in vitro. *Nucleic Acids Res.* **48**, D320–D327 (2020).
8. You, K. et al. PhaSepDB: a database of liquid–liquid phase separation related proteins. *Nucleic Acids Res.* **48**, D354–D359 (2020).
9. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
10. Cunningham, F. et al. Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
11. Thul, P. J. et al. A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).
12. Mészáros, B., Erdos, G. & Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018).
13. Saar, K. L. et al. Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proc. Natl Acad. Sci. USA* **118**, e2019053118 (2021).
14. van Mierlo, G. et al. Predicting protein condensate formation using machine learning. *Cell Rep.* **34**, 108705 (2021).
15. Hardenberg, M., Horvath, A., Ambrus, V., Fuxreiter, M. & Vendruscolo, M. Widespread occurrence of the droplet state of proteins in the human proteome. *Proc. Natl Acad. Sci. USA* **117**, 33254–33262 (2021).

16. Hatos, A., Tosatto, S. C. E., Vendruscolo, M. & Fuxreiter, M. FuzDrop on AlphaFold: visualizing the sequence-dependent propensity of liquid-liquid phase separation and aggregation of proteins. *Nucleic Acids Res.* **50**, W337–W344 (2022).
17. Chen, Z. et al. Screening membraneless organelle participants with machine-learning models that integrate multimodal features. *Proc. Natl Acad. Sci. USA* **119**, e2115369119 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Technical architecture of the web application

The web application is internally divided into four distinct components (Extended Data Fig. 3).

(1) The main database contains the condensates, proteins and other related datasets. The database service used for this component is MongoDB, where the interlinking of resources is done at the logical layer. The Application Programming Interface (API) layer exposes consumable data for the frontend to visualize data. This can also be used for programmatic access to the data by statisticians and bioinformaticians. It also facilitates filter, search, sort and some basic aggregate functionalities. We used Flask, a simple lightweight Python framework, for this.

(2) The frontend allows visualization of data—list and detail pages for condensates and proteins. It provides user-interactable controls for the addition/modification of data. The frontend is built using a Javascript framework called Vue.js. A subcomponent embedded within the frontend is the Content Management System (CMS), which facilitates user management and workflow of each update item submitted by users. The CMS also helps to maintain a history of contribution for each data record.

(3) The contribution database stores the history of data from the crowdsourcing effort and helps versioning the database. All data edits submitted by contributors and review actions performed by maintainers are stored here to aid transparency.

(4) The Sync script is a scheduled process that runs at regular time intervals (daily/weekly) to copy the update items safely to the main database. This python program interacts with both PostgreSQL from CMS and MongoDB from the main database in a producer–consumer paradigm.

Data aggregation

The four most popular protein phase-separation databases are PhaSePro[6], PhaSepDB[8], LLPSDB[7] and data resource of liquid–liquid phase separation (DrLLPS)[5], which provide an excellent resource for phase-separating proteins along with their possible localization (membraneless organelles) and also synthetic condensates (in vitro experiments). However, all these databases are protein-centric and not condensate-centric, that is, a protein may or may not be part of a biomolecular condensate. To fill this gap and augment our current knowledge, we have built a condensate-centric database. At CD-CODE, we aggregated data from the four databases to create a seed dataset of condensates and their protein members (Extended Data Fig. 6). This was followed by a manual curation of the data and annotation of up-to-date evidences (PubMed references, experimental evidences, confidence scores explained below).

Literature curation

All the 674 condensate literature-related PubMed references which were published after the release of the last four databases were checked manually, and new condensates and proteins involved in condensates were identified and added to CD-CODE along with the relevant PubMed IDs. In total, 26 new biomolecular condensates and 224 new proteins were added to CD-CODE which were not part of the previous databases (Extended Data Fig. 6 and Extended Data Tables 1 and 2).

Data curation

Condensate names were standardized and condensates were merged based on synonyms that were manually assembled from the literature (Supplementary Table 1). In total, 9,916 protein–condensate relationships are supported by a PubMed ID as evidence.

Annotation of driver proteins

We defined the 'Role in condensate' of a protein as the role it plays in a condensate formation and integrity. Based on their role in the formation of the condensate, proteins can be categorized into one of two classes—'driver' and 'member'.

Drivers are defined based on the following criteria:

(1) They induce the formation of a condensate.
(2) They are essential for the integrity of a condensate.

Driver proteins are also called 'scaffolds' since they bind many other proteins. It is important to note that these annotations are valid in the context of a specific condensate. A protein acting as a driver for one condensate can be a member in another condensate.

Confidence scores

We designed the confidence score criteria and started manually adding the experimental evidences. 1 star, PubMed reference-annotated; 2 stars, high-throughput experiment (for example, mass spectrometry); 3 stars, in vitro evidence; 4 stars, in cellulo evidence; 5 stars, in vivo evidence. We also collected the experimental methodology, for example, fluorescence recovery after photobleaching or microscopy colocalization.

Crowdsourcing functionalities

Registered users can contribute to data curation, including any kind of attribute data update for any protein/condensate we already have at CD-CODE, such as marker addition, name change, description update, add/remove PubMed evidence and so on (Extended Data Fig. 4). Additionally, users can add new proteins to existing condensates and create new condensates by filling in a form.

Encyclopedia

The encyclopedia (wiki.cd-code.org), another crowdsourcing element of CD-CODE, contains descriptive data related to definitions, synonyms and terminologies in the world of biomolecular condensates and liquid–liquid phase separation in biology. Contributors have the necessary credentials to create and edit content at the encyclopedia. It is powered by Wiki.js and provides easy content management functionalities to users who are not equipped with programming or HTML skills.

Data analysis

Condensate counts for species were taken from CD-CODE v.1.01 and summarized by functional group annotation (Supplementary Table 1). The tree was downloaded from timetree.org (ref. 18) and scaled with the chronos function in the ape (5.6–2) R (v.4.2.0) package for better visualization. The tree was plotted using ggtree (v.3.4.4) and ggplot (v.3.4.0). Figure 2b,c was created using custom Python scripts (pandas: 1.2.4; matplotlib: 3.3.4; networkx: 2.5).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Our database is available at cd-code.org and the encyclopedia at wiki.cd-code.org. Rolling feature update releases and data dumps will be noted here: https://cd-code.org/release. Source data are provided with this paper.

Code availability

Software and documentation: https://git.mpi-cbg.de/dd-code-team/dd-code-docs.

References

18. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).

Author contributions

A.T.-P., A.A.H. and D.M.M. conceived the project with input from C.F.W.C., C.L. and S.G. S.G., H.M. and L.H. conceived the code concept and design. Backend coding development and implementation was done by S.G. Frontend coding development and implementation was done by H.M. and R.G. N.R. and C.F.W.C. created the content for the encyclopedia. N.R. curated the data manually and served as the maintainer of the data. S.G. collected data and designed the data structure. D.M.M. and I.A.K. provided resources. S.G., A.H. and C.L. analyzed the data. A.H., C.L., A.T.-P., S.G., H.M. and N.R. visualized data and prepared the figures. A.T.-P., L.H. and S.G. supervised the project. A.T.-P. wrote the main manuscript with input from all the coauthors. S.G. prepared supplementary information and documentation of the code. A.T.-P. acquired the funding and administered the project.

Competing interests

A.A.H. is a founder and shareholder of Dewpoint Therapeutics. I.A.K. and D.M.M. are employees and shareholders of Dewpoint Therapeutics. A.T.-P. is a recipient of research support from Dewpoint Therapeutics. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-023-01831-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-023-01831-0.

Correspondence and requests for materials should be addressed to Agnes Toth-Petroczy.

Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Arunima Singh, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.

Leucocyte nuclear body

General Information

Species	Homo sapiens (9606)
Description	Nuclear bodies are membraneless, RNA-rich organelles in the cell nucleus that concentrate certain nuclear proteins and RNA-protein complexes.
Also Known As	Nuclear Body, Nuclear Compartments, Nuclear Protein Granule, Nuclear Bodies That Occur At Super Enhancers In Mescs
Markers ❶	Q13342
Regulators	Unknown
No. of Proteins	21
Confidence Score	★★★★★

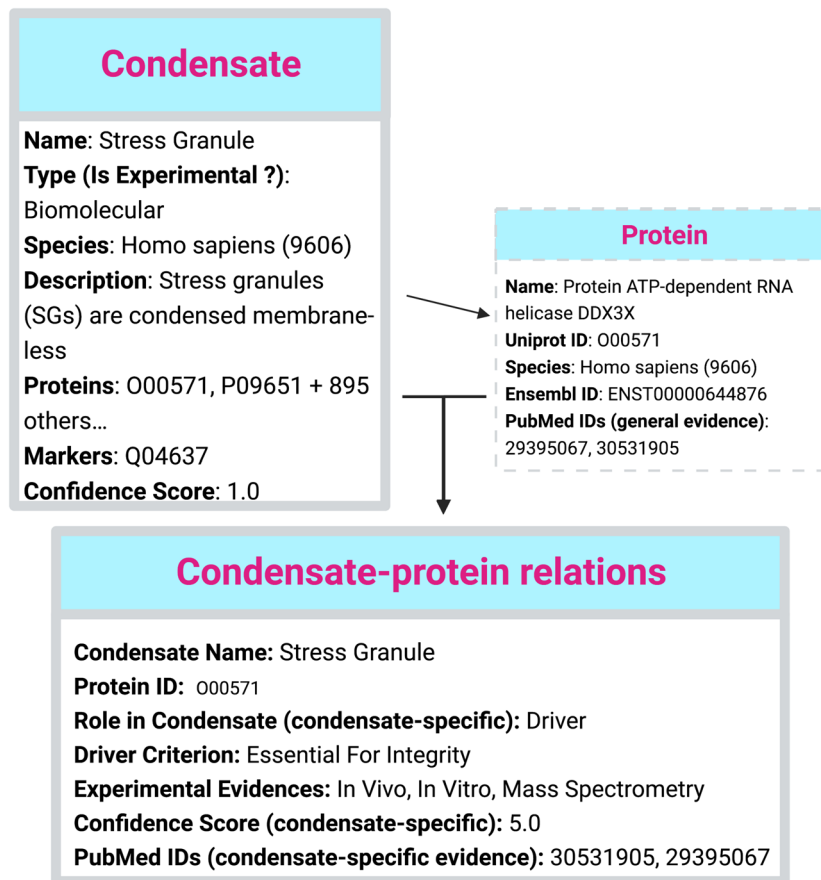### Proteins

Show 10 entries   Filter [ ]   Previous 1 2 3 Next   **Download CSV**

| Gene Name	Name	Pubmed ❶	Role in Condensate	Experimental Evidence	Driver Criterion	Confidence Score
SPOP	Speckle-type POZ protein		Member	In Vivo		★★★★★
BRD4	Bromodomain-containing protein 4		Member	In Vivo		★★★★★
LMNA	Prelamin-A/C		Member			★★★★★
ESR1	Estrogen receptor		Member	In Vivo		★★★★★
PGR	Progesterone receptor		Member			★★★★★
SNRNP70	U1 small nuclear ribonucleoprotein 70 kDa		Member			★★★★★
HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1 OS=Homo sapiens OX=9606 GN=HNRNPA1 PE=1 SV=5		Member			★★★★★
GATA3	Trans-acting T-cell-specific transcription factor GATA-3		Member	In Vivo		★★★★★
MATR3	Matrin-3		Member	In Vivo		★★★★★
EWSR1	RNA-binding protein EWS		Member	In Vivo		★★★★★

Previous 1 2 3 Next

Extended Data Fig. 1 | Detail page of the biomolecular condensate, Leucocyte nuclear body in humans. The top section displays 'General Information' of the condensate whereas the section below shows 'condensate-specific protein' data in a tabular format.

Condensate

Name: Stress Granule
Type (Is Experimental ?):
Biomolecular
Species: Homo sapiens (9606)
Description: Stress granules
(SGs) are condensed membrane-
less
Proteins: O00571, P09651 + 895
others…
Markers: Q04637
Confidence Score: 1.0

Protein

Name: Protein ATP-dependent RNA
helicase DDX3X
Uniprot ID: O00571
Species: Homo sapiens (9606)
Ensembl ID: ENST00000644876
PubMed IDs (general evidence):
29395067, 30531905

Condensate-protein relations

Condensate Name: Stress Granule
Protein ID: O00571
Role in Condensate (condensate-specific): Driver
Driver Criterion: Essential For Integrity
Experimental Evidences: In Vivo, In Vitro, Mass Spectrometry
Confidence Score (condensate-specific): 5.0
**PubMed IDs (condensate-specific evidence):** 30531905, 29395067

Extended Data Fig. 2 | CD-CODE is a condensate centric database. Example data structure of a condensate and one of its member proteins. We defined two distinct entity collections – one for condensates and one for proteins, the latter comprising of general details regarding the members of the former. We also store mappings collection of condensate-protein relations, comprising of data points that relate to specific condensate-protein pairs.
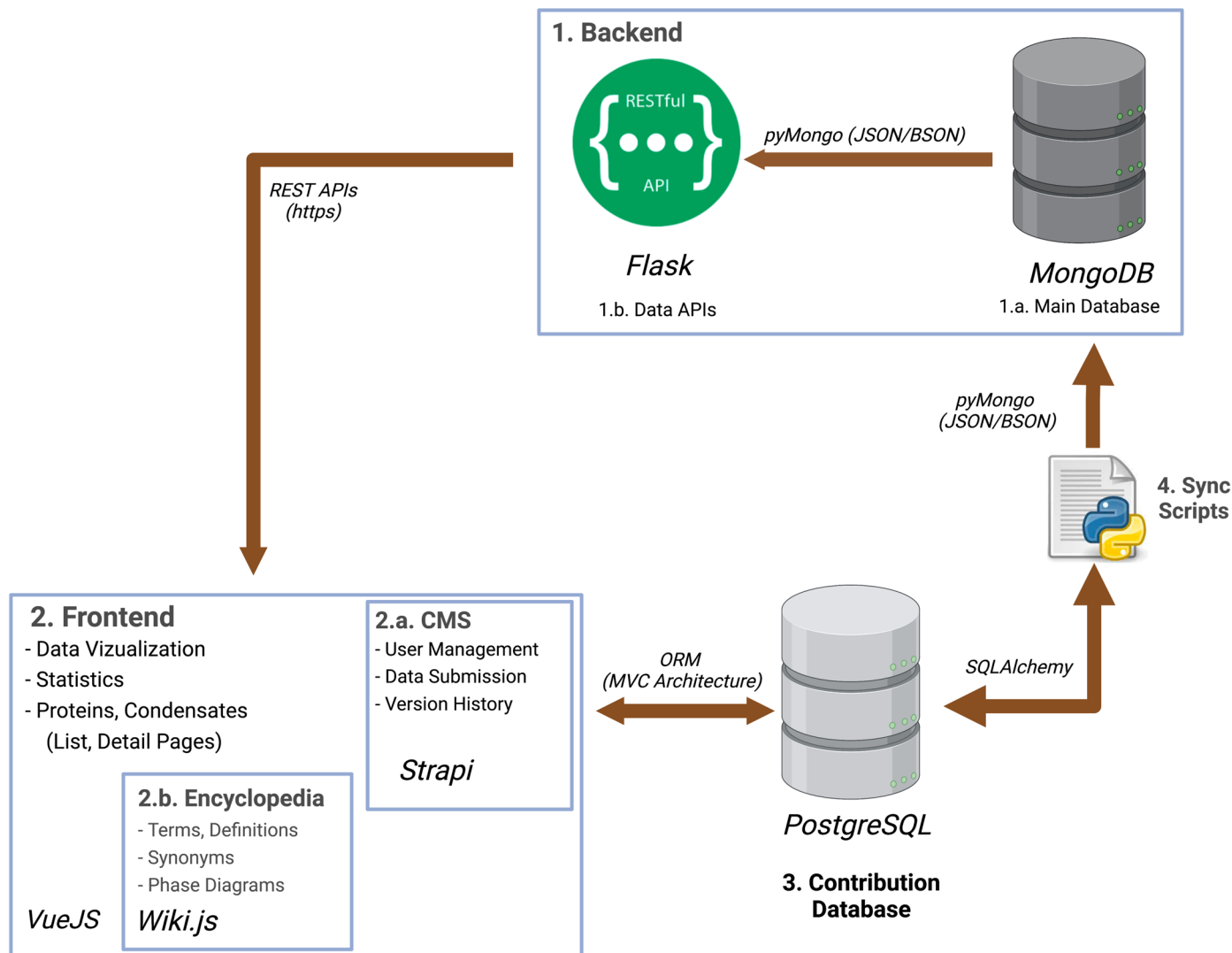
Extended Data Fig. 3 | Schematics of the technical architecture of CD-CODE web application. The web application is internally divided into four distinct components. 1. The main database contains the condensates, proteins, and other related datasets. The API layer exposes consumable data for the frontend to visualize data. This can also be used for programmatic access to the data by statisticians and bioinformaticians. It also facilitates filter, search, sort, and some basic aggregate functionalities. 2. The frontend allows visualization of data, for example detail pages for condensates and proteins (Extended Data Figs. 1 and 7). A sub-component embedded within the frontend is the Content Management System (CMS) which facilitates user management, and workflow of each update item submitted by users. The CMS also helps maintain a history of contribution for each data record. 3. The contribution database stores the history of data from the crowdsourcing effort and helps versioning the database. All data edits submitted by contributors and review actions performed by maintainers are stored here to aid transparency. 4. The Sync script is a scheduled process that runs at regular time intervals (daily/weekly) to copy the update items safely to the main database. This python program interacts with both PostgreSQL from CMS and MongoDB from the main database in a producer-consumer paradigm.
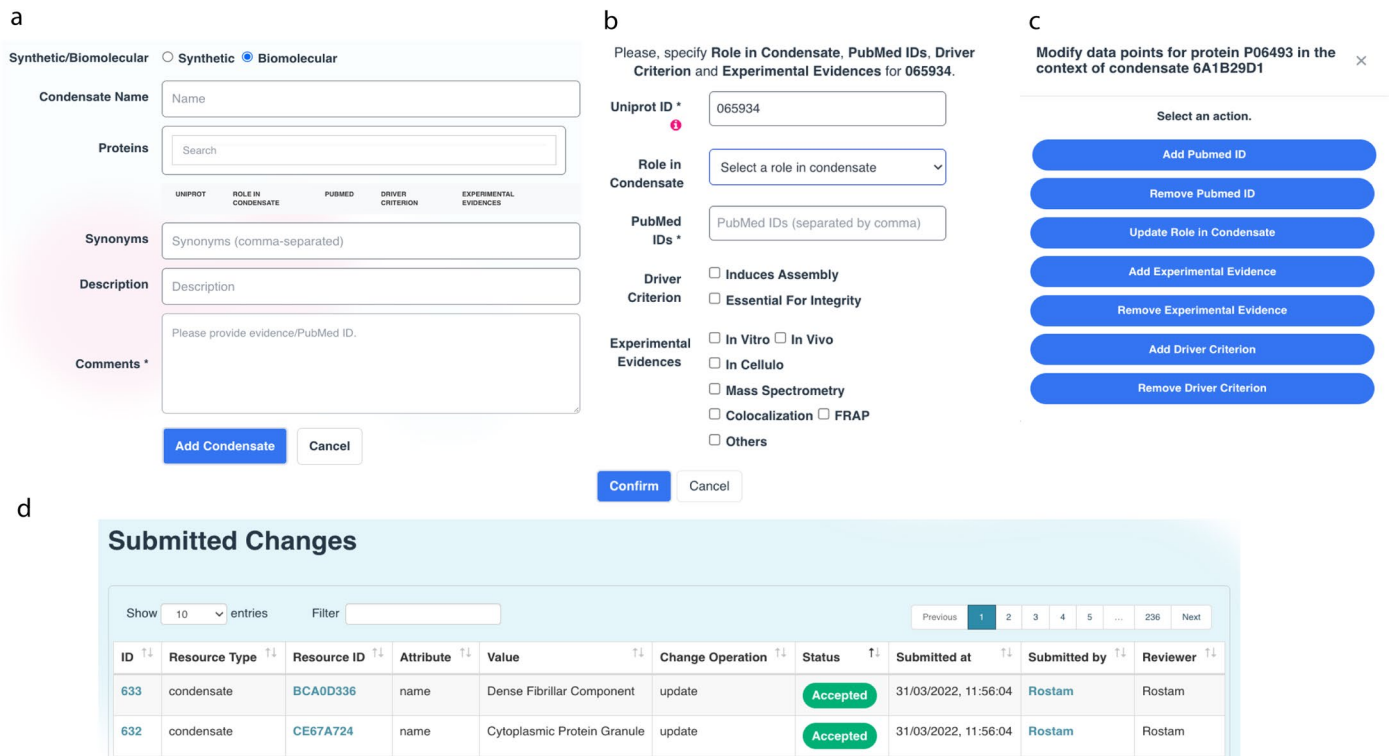
Extended Data Fig. 4 | Data curation from the user interface. a) Add new condensate form. **b**) Add new protein form. **c**) Edit data actions. **d**) History of entries is recorded.
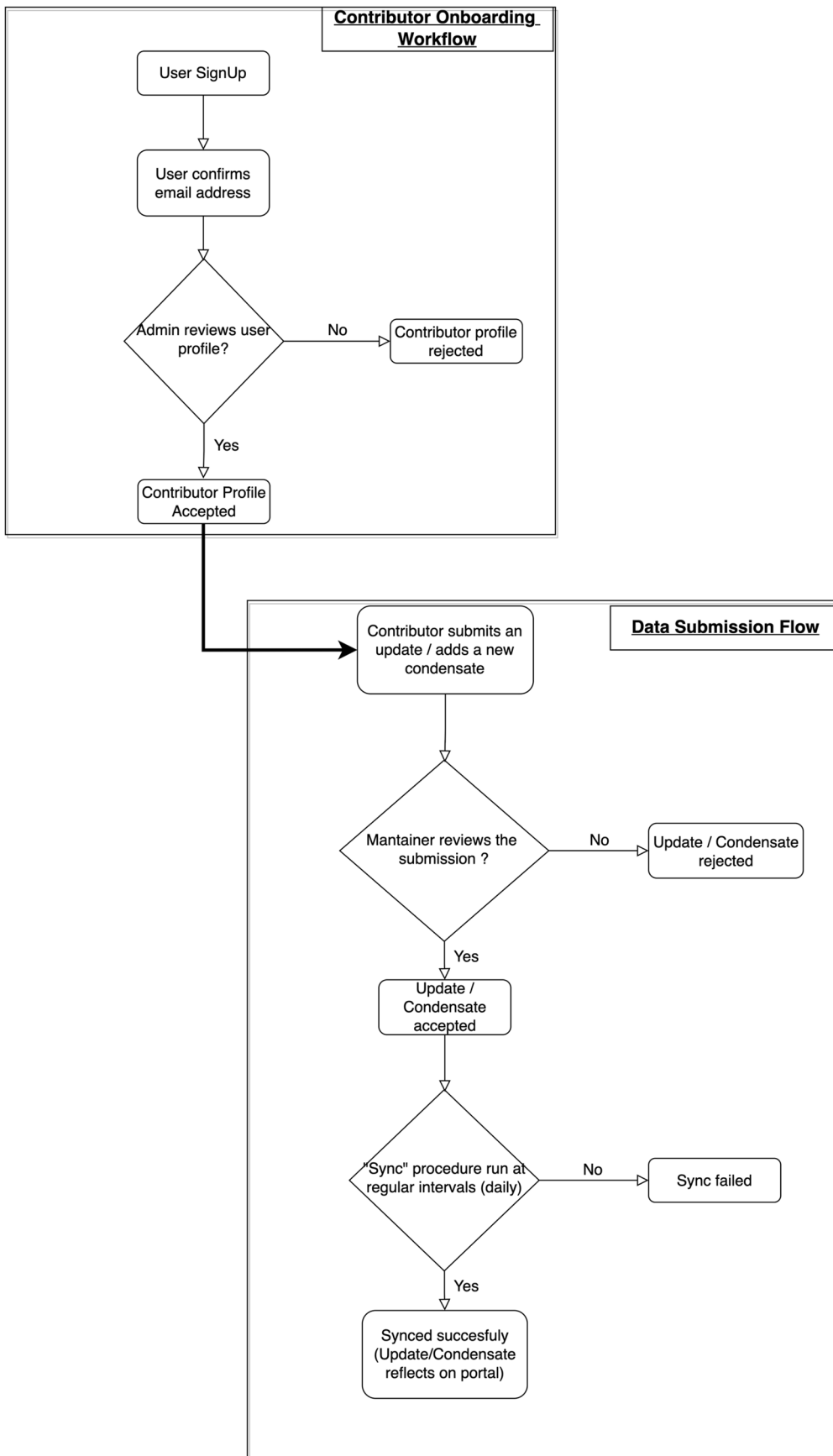
Extended Data Fig. 5 | Flowchart of contributions. After submitting the registration form, the user receives an email to confirm the given address by clicking a link. After the user's email address is confirmed, an administrator checks the pending contributor signup requests and approves them. A signed-in contributor can see editable attributes besides selected data points on detail pages, in addition to a new webpage form for submission of novel condensates. The new submissions are then moderated by maintainers first and later by the Sync script.

Extended Data Fig. 6 | Venn diagram to compare protein entries in condensate/liquid-liquid phase separation (LLPS) databases. The Venn diagram displays the overlap of the protein entries in CD-CODE and four other LLPS/condensate databases (LLPSD[1], DrLLPS[2], PhaSepDB[3], PhaSePro[4]). CD-CODE incorporates data from all four databases and additionally, there are 224 new condensate-forming proteins in v1.01 that were added to CD-CODE from curating the literature.

a LAT

General Information

Name	**Protein Linker for activation of T-cells family member 1**
Species	Homo sapiens (9606)
Ensembl ID	**ENST00000566177** 🔗
Ensembl Gene ID	**ENSG00000213658** 🔗
UniProt	**O43561 (LAT_HUMAN)** 🔗
Antibodies	**ENSG00000213658** 🔗
Verified in Biomolecular Condensates	2
Verified in Synthetic Condensates	No
Pubmed ⓘ	**27056844** 🔗 **29424691** 🔗 **31268421** 🔗 **25321392** 🔗 **33080222** 🔗 **30846600** 🔗 ✏️
Sequence	Copy 📋

10	20	30	40	50
MEEAILVPCV	LGLLLLPILA	MLMALCVHCH	RLPGSYDSTS	SDSLYPRGIQ
60	70	80	90	100
FKRPHTVAPW	PPAYPPVTSY	PPLSQPDLLP	IPRSPQPLGG	SHRTPSSRRD
110	120	130	140	150
SDGANSVASY	ENEGASGIRG	AQAGWGVWGP	SWTRLTPVSL	PPEPACEDAD
160	170	180	190	200
EDEDDYHNPG	YLVVLPDSTP	ATSTAAPSAP	ALSTPGIRDS	AFSMESIDDY
210	220	230	240	250
VNVPESGESA	EASLDGSREY	VNVSQELHPG	AAKTEPAALS	SQEAEEVEEE
260				
GAPDYENLQE	LN			

b

Disordered Content
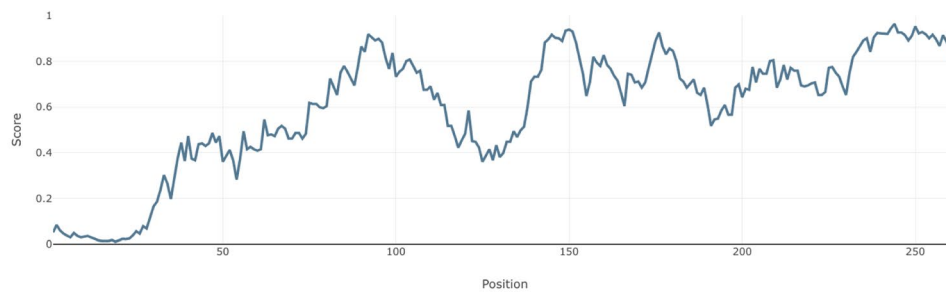
c

Amino Acid Composition
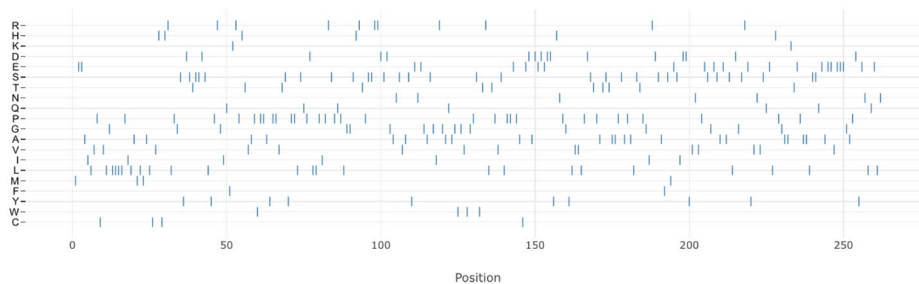
d

PTMs affecting condensate formation

Name	Type	Enzyme	Position	Description
phosphorylation	promotes	ZAP70	200	none
phosphorylation	promotes	ZAP70	220	none

e

Biomolecular Condensates — Count- 2

Name	No. of Proteins	Species	Role in Condensate
T-cell signalosome	3	Homo sapiens	member
Signaling cluster	22	Homo sapiens	member

Extended Data Fig. 7 | Example detail page to illustrate the general information and data points stored in CD-CODE. a) General information about a protein is shown for the LAT Linker for activation of T-cells family member 1 protein (Uniprot ID: LAT_HUMAN). Links are provided to other popular protein-related databases (Uniprot, Ensembl, Human Protein Atlas). **b**) Disorder score computed by IUPred[5]; **c**) Barcode plot depicting amino acid composition bias; **d**) Table of PTMs affecting condensate behavior; **e**) List of biomolecular condensates of which the given protein is a member.

Extended Data Table 1 | Biomolecular condensates across databases

	CD-CODE	DrLLPS	PhaSepDB	PhaSePro
Total no. of Proteins	**9861**	9285	2831	121
No. of Proteins taken in CD-CODE	**9722**	8323	2474	120
No. of Condensates (Only biomolecular)	**244**	111	45	112
No. of Protein-Condensate Mappings	**9724**	10706	7680	213

Comparison of total number of data points presented by CD-CODE in comparison to other databases in the context of biomolecular condensates or membrane-less organelles (MLOs). We count the condensates from different species as separate counts of biomolecular condensates. The other three databases here also report phase-separating proteins without any condensate or with condensates named 'Droplets' and 'Others'. Such entries are removed from CD-CODE by manual curation.

Extended Data Table 2 | Synthetic condensates across databases

	CD-CODE	LLPSDB	PhaSepDB
Total no. of Proteins	**343**	273*	245
No. of Proteins taken in CD-CODE	**343**	188	245
No. of Protein-Condensate Mappings	**343**	190	182

Comparison of total number of proteins presented in CD-CODE to other databases in the context of synthetic condensates or in *vitro* experiments. Since LLPSDB only reports *in vitro* experiments, the condensate-related counts are not reported for that. *Out of the 273 proteins reported by LLPSDB, 198 are natural proteins, and 75 are designed proteins (not all experiments result in phase separation).

Corresponding author(s): Agnes Toth-Petroczy

Last updated by author(s): Jan 2, 2023

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected using custom software: https://git.mpi-cbg.de/dd-code-team/ddcode-db-creator/
Data analysis	Data analysis was done using custom Python scripts (pandas: 1.2.4, matplotlib: 3.3.4, networkx: 2.5) and R (4.2.0) scripts using ape (5.6-2), ggtree (3.4.4) and ggplot (3.4.0). Figure 1 was created using BioRender.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Figure 2 has associated raw data.
Data is available at cd-code.org, and web links are provided to other databases (Uniprot, Enseml, Human Protein Atlas). We collected data from:
1. PhaSePro (https://phasepro.elte.hu), LLPSDB (http://bio-comp.org.cn/llpsdb/), DrLLPS (http://llps.biocuckoo.cn), PhaSepDB (http://db.phasep.pro) - Collect seed data for phase separating proteins and respective PubMed IDs
2. IUPred2A (https://iupred2a.elte.hu/) - Annotate disordered regions in protein
3. UniProt (uniprot.org) - Fetch general details of proteins

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   ☐ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	NA
Data exclusions	NA
Replication	NA
Randomization	NA
Blinding	NA

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒ ☐	Antibodies
☒ ☐	Eukaryotic cell lines
☒ ☐	Palaeontology and archaeology
☒ ☐	Animals and other organisms
☒ ☐	Human research participants
☒ ☐	Clinical data
☒ ☐	Dual use research of concern

Methods

n/a	Involved in the study
☒ ☐	ChIP-seq
☒ ☐	Flow cytometry
☒ ☐	MRI-based neuroimaging