Proteome-scale quantification of the interactions driving condensate formation of intrinsically disordered proteins

Rasmus K. Norrild[1], Sören von Bülow[2], Einar Halldórsson[1], Kresten Lindorff-Larsen[2], Joseph M. Rogers[3,*], and Alexander K. Buell[1,*]

[1]Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs. Lyngby, Denmark
[2]Structural Biology and NMR Laboratory & the Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[3]Department of Drug Design and Pharmacology, University of Copenhagen, Copenhagen, Denmark
[*]To whom correspondence should be addressed: joseph.rogers@sund.ku.dk and alebu@dtu.dk

December 21, 2024

Abstract

The formation of biomolecular condensates is involved in compartmentalisation, regulation, and signalling across most living organisms. Condensation can be driven by phase separation of proteins, and involves highly dynamic, multivalent, and weak interactions of their intrinsically disordered regions (IDRs). Understanding these interactions has been limited by the lack of large-scale experimental methods capable of generating thermodynamic data. Here, we develop and use Condensate Partitioning by mRNA-Display (CPmD) to directly measure the energetics of the partitioning of almost 100,000 peptides of 14 to 40 amino acid length into condensates formed by an IDR, that of the germ granule protein Dead-box helicase 4 (DDX4). This large set of peptides contained many that were derived from DDX4 itself, as well as a tiling library of a significant fraction of all experimentally characterised IDRs. We show that the results of such large scale partitioning studies yield a uniquely quantitative picture of the driving forces that govern IDR condensate formation of DDX4, as well as partitioning of other sequences into its condensates. Furthermore, we demonstrate that the derived set of interactions is also able to quantitatively explain condensate formation of unrelated sequences.

Introduction

Biomolecular condensates have been shown to be ubiquitous across cell biology and act as membrane-less compartments for spatial control of cellular functions[1]. Malfunction of such condensates is linked to aggregation diseases[2], cancer[3], and autism spectrum disorders[4], which are often caused by genetic variation. Therefore, mechanistic and sequence-level understanding is important in order to predict effects of genetic variation and to modulate the behaviour of condensates with drugs[5].

Biomolecular condensates are assemblies that can concentrate various biomolecules, primarily proteins and RNA, and are currently thought to occur via phase separation coupled to percolation[6]. This mechanism depends on multivalent interactions between biomolecules, as exemplified by the co-condensation of tandem repeats of SH3 domains and their peptide ligands[7]. In addition to such motif-driven and stoichiometric interactions based on folded domains, IDRs[8] often contribute significantly to phase separation[9–11].

The conserved RNA helicase DDX4 is an example of a protein containing an IDR necessary and sufficient for phase separation[9]. DDX4 is highly specific to germ line cells and is a core component of ribonucleoprotein germ granules, which are exemplary biomolecular condensates that contain essential factors for germ cell determination and function[12–15]. Deletion of DDX4 orthologs causes detrimental effects in either male or

female reproduction, and deletions of the N-terminal IDR (DDX4N1) causes defects consistent with functions as a condensate scaffold[16–18].

DDX4N1 is a representative IDR from the dominant type of condensate-driving regions[19], and the purified protein domain remains disordered both in dilute buffer and within the condensed phase and does not convert into amyloid fibrils[20]. The presence of adequately spaced phenylalanines and arginines in the IDR is evolutionarily conserved and removal of these residues ablates phase separation[9, 20]. This is consistent with features that explain phase behaviour of prion-like low complexity domains[21] and elastin-like peptides[22, 23] and is sometimes referred to as a "sequence grammar"[24].

However, if one "sequence grammar" exists for condensate formation of disordered regions, it is unclear how multiple condensates can coexist without mixing inside the cell[25, 26]. A promising experimental strategy has been to use mutagenesis to cover the sequence space more broadly but it is limited by the need to obtain many different highly pure protein variants in substantial quantities[27, 28]. High throughput methods collectively termed Multiplexed Assays of Variant Effects (MAVEs)[29] have greatly increased the scale at which similar experiments can be conducted for other protein properties. Prominently, cDNA-display or mRNA-display coupled to massively parallel DNA sequencing has been used to obtain high-quality data on the energetics of protein folding and binding at a large scale, up to hundreds of thousands of sequences[30–32].

Here, we introduce CPmD, a novel method in which we leverage mRNA-display techniques[33, 34] coupled to high-throughput sequencing to study the interactions leading to condensate formation. We show that partitioning of overlapping peptide tiles can provide a highly resolved profile of the contributions of different regions of an IDR sequence to form condensates. Furthermore, the unprecedented throughput of CPmD allows us to quantify the partitioning of close to 100 000 peptides into DDX4N1 condensates. These peptides included a large mutational library of a specific 40 amin acid tile of DDX4N1, as well as a tiled library of a significant fraction of all experimentally characterised IDRs. The unbiased nature of these experiments yields a fully consistent set of interaction strengths that is able to account quantitatively for DDX4N1 wild type (WT) and mutant condensate formation. Surprisingly, this set of interactions is able to quantify the condensate formation of some unrelated sequences, thereby demonstrating that CPmD can reveal generic features of IDR interactions even if applied to only a single IDR scaffold. CPmD is generally applicable to reconstituted condensates *in vitro* and through its scale and quantitative nature has the potential to provide insight into the interactions between biomolecules that drive the formation of diverse condensates.

CPmD measures both mRNA and peptide partitioning

Large-scale screening of proteins using display methods coupled to massively-parallel DNA sequencing technologies are usually restricted to screening and selection of properties that are conferred by individual molecules (e.g. binding, folding, stability or enzymatic activity). Condensate formation or phase separation is instead a collective property and *in vitro* evaluation of protein phase separation usually requires significant quantities of highly pure protein[27]. Therefore, a parallelizable proxy measurement capable of generating data at very high throughput is strongly desirable.

We reasoned that mRNA-display[33, 34] would be a suitable display technologies for biomolecular condensates, as both RNA and protein are core components of these assemblies and therefore the presence of the mRNA would not be disruptive to the condensates. Additionally, mRNA-display is the minimal construction that links the phenotype (peptide) to the genotype (RNA). The experiments are carried out using a purified protein to form reconstituted condensates, which circumvents challenges with complex effects of similar experiments performed inside living cells[35].

We employed an established protocol that uses puromycin to attach the mRNA to the nascent peptide via a flexible phosphate-containing polyethylene glycol (PEG) linker[36] (Methods, Supplementary Fig. 1). As a means for selecting or screening, we initiated phase separation of reconstituted DDX4 condensates in the presence of the mRNA libraries, and then separated the dilute and condensed phases by centrifugation (Fig. 1a).

mRNA-displayed libraries were observed to enter the DDX4N1 condensed phase, and we used confocal microscopy to verify that they distributed homogeneously within the condensate (Fig. 1b). mRNA alone,

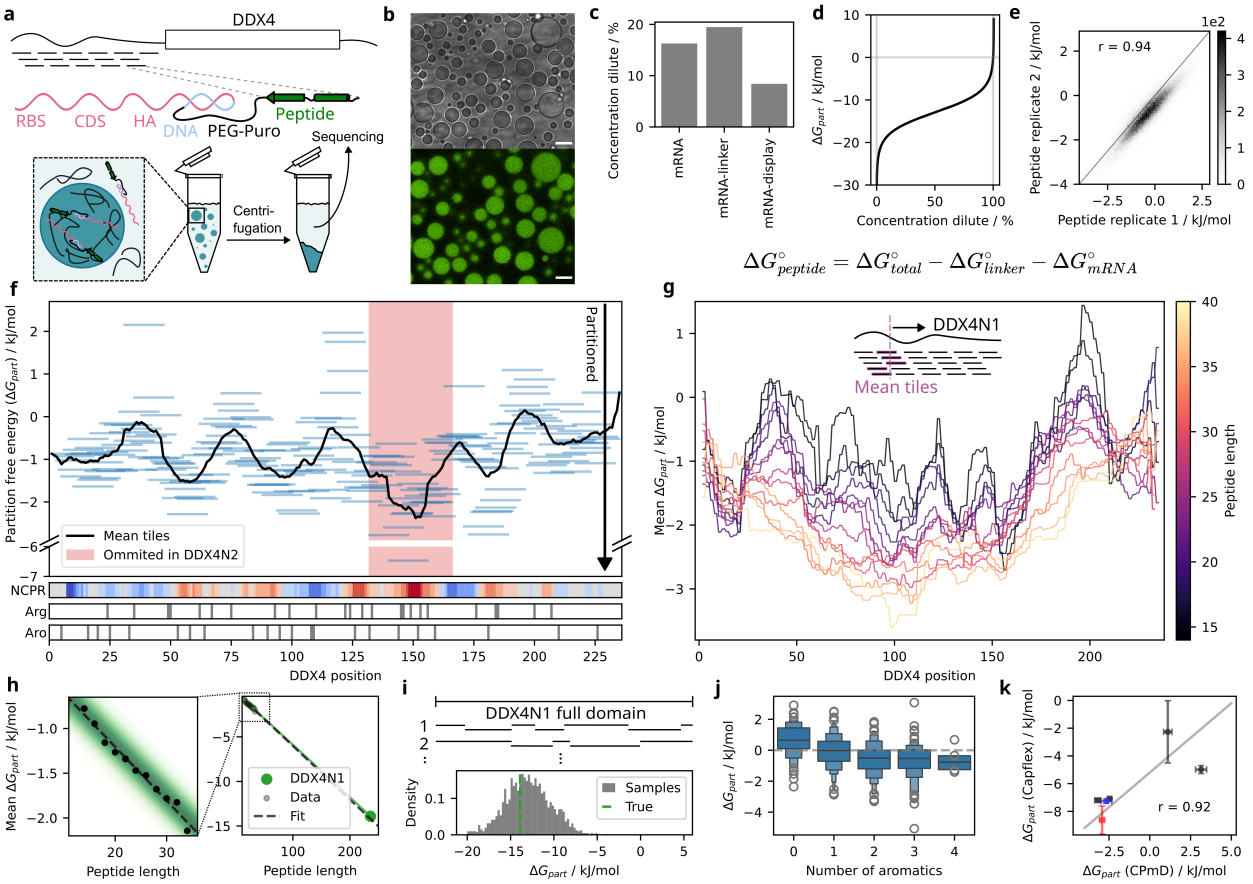

Figure 1: CPmD tiling of DDX4N1 quantitatively explains condensate formation. a, Peptides from the N-terminal IDR of DDX4 were encoded in a DNA library for mRNA display and transcribed using purified T7 RNA polymerase. The resulting mRNA contains a ribosomal binding site (RBS), a coding sequence (CDS) for peptide of interest, and a HA-tag for purification via the translated peptide. The synthetic mRNA (pink) is ligated to a puromycin-containing PEG-linker (black) containing a DNA sequence (blue) that forms a hetero duplex with the mRNA to facilitate single stranded RNA ligation. The resulting mRNA display molecule is purified using the HA-tag and added to an excess of DDX4N1 to measure partitioning of all library members through RNA/DNA sequencing. **b**, Brightfield and fluorescence micrograph of DDX4N1 condensates mixed with 1 nM mRNA-display library using confocal microscopy, where the puromycin linker contained FITC for fluorescence detection. Scale bar is 10 μm. **c**, qPCR estimation of the concentration of the mRNA-display library in the dilute phase after centrifugation for the 16 residue library. **d**, The partition free energy is calculated by using the known dense phase volume fraction (0.5%). **e**, The measured partition free energies is reproducible between 2 repeats. **f**, The partition free energies of 16 amino acid tiles of DDX4N1 shows that interaction prone regions are distributed throughout the protein with a periodicity of ∼40 amino acids. The splice isoform Q9NQI0-2 of DDX4N1 (DDX4N2) lacks an exon (in red) containing the most partitioned region of the IDR and, consistently, does not form condensates under the examined conditions[9] (Supplementary Fig. 5). **g**, The moving average of peptide tiles, as shown in black in panel f, for different peptide lengths. **h**, The mean partition free energy of DDX4N1 peptides increases linearly with length. The expected length dependence obtained from concentration measurements in the dense and dilute phase of the full domain is indicated in green. **i**, The partition free energy of DDX4N1 re-sampled as the sum of different sets of peptide tiles. **j**, The number of aromatic residues strongly correlates with peptide partitioning. **k**, Selected peptides were synthesised with a Oregon Green fluorophore (OG488). The resulting measurements of partition free energy (including favorable partitioning of the OG488 dye itself) are shown against the values obtained with mRNA-display. Negatively (blue) and positively (red) charged peptides are highlighted.

without displayed peptide, was also capable of significant partitioning into the condensed phase (Fig. 1c). However, peptide partition free energies for individual sequences, in principle, could still be obtained by measuring the partition free energies of mRNA-display constructs and subtracting the contribution from the mRNA-puromycin linker without any translated peptide attached.

We note that the typical concentration of purified mRNA-display libraries (1-100 nM) is three to five orders of magnitude lower compared to the 120 μM DDX4N1 used. Even with strongly partitioned libraries with an average partition free energy of more than 20 kJ/mol, the library concentration inside the condensates is only tens to hundreds of μM (Fig. 1d, Supplementary Fig. 2) whereas the DDX4N1 has a concentration of tens of mM inside the condensates[20]. Any interactions between library members or perturbations to the overall phase separation of DDX4N1 will therefore be minimal and can be neglected.

Peptide partitioning explains DDX4N1 phase separation

Tiling approaches, where a protein region is sequentially cut into overlapping peptide fragments, have proven powerful to map important regions of IDRs, for example in activator domains of transcription factors[37]. We therefore started by exploring if CPmD could identify peptide regions driving DDX4N1 phase separation. We reasoned that a peptide fragment of DDX4N1 partitioning at low concentration into an excess of the full DDX4N1 domain would experience the same types of intermolecular interactions in the condensate as it would do when part of the longer IDR[38].

We therefore designed libraries with peptides of different lengths tiled through DDX4N1 starting at every second position, and one higher-resolution tiling using only peptides with 16 residues but starting from every single position in the sequence. The synthetic mRNA encoding the library partitioned strongly, which agreed well with earlier reports that single stranded nucleic acids were strongly partitioned into DDX4N1 condensates[9]. Partitioning was stronger for mRNA with a higher fraction of purine content, lower secondary structure propensity, and longer length (Supplementary Fig. 3), however, the main energetic contribution was independent of mRNA length as recently seen for stress granule formation in yeast cells[39]. Despite that, we were able to isolate the partition free energy of the peptide by subtracting that of the mRNA (or mRNA-puromycin linker fusion), to yield highly reproducible data (r=0.94, Fig. 1e).

The peptide contributions to the partition free energy were about 10% of that of the mRNA, corresponding well to their relative molecular weights (Supplementary Fig. 1). The 16 amino acid tiles of DDX4N1 partitioned with an energy of -1 to -2 kJ/mol corresponding to only an approximately 2-fold enrichment within the condensed phase of such an isolated peptide (Fig. 1f). The average partition free energy as a function of peptide length revealed a linear relationship in the range where we had reliable data, i.e. 14 to 36 amino acids (Fig. 1g). Extrapolation of this trend to the full length of the domain (236 amino acids) resulted in -14.2 kJ/mol, which is approximately the energy expected based on measurement of the coexisting dilute and dense phase concentrations of the DDX4N1 domain alone under these solution conditions, -13.9 kJ/mol (Fig. 1h,i, Supplementary Fig. 4). This result showed that the sum of the measured small partition free energies from the peptide tiles were fully sufficient to explain the phase separation of the entire DDX4N1 domain, and thereby established a direct link between partitioning into condensates and the ability to undergo condensate formation.

Specific regions of DDX4N1 drive condensation

By taking a moving mean of the tiles through the domain such that all tiles containing an amino acid at a given position were averaged, a profile along the IDR is obtained. This partition profile of DDX4N1 shows a periodicity of $\sim$40 amino acids, resulting in a block copolymer-like separation of regions driving partitioning and, by proxy, phase separation (Fig. 1f). Strong autocorrelation in the moving mean is expected, but only below a length scale of 16 amino acids, which is confirmed by auto-correlation analysis of simulated peptides with random partition free energies (Supplementary Fig. 4). With regard to sequence composition, we found an enrichment of aromatic amino acids in the most strongly partitioned peptides with an average effect of 0.43 kJ/mol per aromatic residue (Fig. 1j), consistent with results of simulations that showed that areas rich in aromatic residues and arginine drive intermolecular interactions in DDX4N1[40]. A central region of the IDR around position 150 was particularly strongly partitioned. We note that the second natural isoform of the protein, Q9NQI0-2, lacks the exon encoding this region and has been reported not to form condensates[9]. We also did not see condensate formation under the examined conditions when purifying the isoform-specific IDR,

called DDX4N2 (Supplementary Fig. 5). When the same averaging procedure was carried out with peptides of increasing lengths, the features became less pronounced (Fig. 1g). This indicated that the variation in the propensity to form condensates is more clearly visible if averaging over shorter peptide tiles, however, the data quality also declined with length due to sequencing biases (Supplementary Fig. 4).

As DDX4N1 has a predicted net charge of -3 at pH = 6.5, positively charged peptides could be expected to partition more strongly. In addition, the separation of charged patches has been described as a driving force for phase separation of DDX4N1 and other IDRs[9, 40, 41]. The predicted charge of the peptides anticorrelated with partition free energy (-0.13 kJ/mol per positive charge), however, this effect was mostly driven by arginine and not lysine (Supplementary Fig. 6). The guanidine group of arginine has delocalised electrons that can form stacking interactions with other arginines and aromatic amino acids[42], and we found that the peptides partitioned on average 0.28 kJ/mol stronger for each arginine (vs. 0.04 kJ/mol for lysines). The potential effect of charges on partitioning could be masked by the highly negatively charged mRNA, however we did not find particular disagreement for the partitioning of oppositely charged peptides between mRNA-display and measurements on solid-state-synthesised peptides, as shown by good agreement with measurement of fluorescently labeled peptides using our Capflex method, which allows accurate quantification of dilute phase concentrations[43] (r=0.92, Fig. 1k).

In summary, the CPmD experiments yield a detailed, quantitative, and consistent view on which sequence regions of DDX4N1 drive its condensate formation.

Unbiased mutagenesis probes grammar of phase separation

Having shown that we could probe the interactions driving DDX4N1 phase separation, we next asked what features of the protein sequence facilitated partitioning into its condensates. Firstly, interactions could be driven by certain motifs[44], and thus we measured the partitioning of two peptides with scrambled amino acid sequences for each tile in the WT sequence (Fig. 2a). The scrambled data overall showed good agreement with the WT tiles, suggesting that amino acid composition rather than motifs was the main determinant of partitioning. This largely agrees with a model where the content and equal distribution of aromatic amino acids in the sequence is sufficient to drive condensate formation[21], and with a recently described machine learning model that predicts phase separation mostly from features that depend on amino acid composition[45]. Non-equal distribution and patterning of aromatic residues can significantly change the properties, but natural IDR sequences appear to be mostly well mixed[46].

Although the scrambled peptides were mostly similar to WT, differences were observed in the central region of DDX4N1. However, we could not find statistical significance for any features reported to correlate with the extent of condensate formation [45] (e.g. the content and mixing of charged and hydrophobic residues) that could robustly explain the differences between WT and scrambled peptides. The partitioning of any given peptide was primarily defined by whether or not the corresponding non-scrambled peptide has a particularly strong or weak intrinsic partitioning (Supplementary Fig. 6).

We then created a combinatorial mutant library of a 40 residues long peptide fragment in the center of DDX4N1 (100-139). The scale of the CPmD experiment allowed us to do large-scale unbiased mutagenesis resulting in reliable data on 27,269 peptides (20,165 without stop codon) with on average 8.1 substitutions per peptide (20% residues mutated) (Fig. 2b and Supplementary Fig. 7). Using Global Multi-Mutant Analysis (GMMA)[47] we disentangled the effects of all single substitutions in the peptide (represented as changes in partition free energy: $\Delta\Delta G_{\mathrm{part}}$). GMMA was developed to study folded proteins, but we used data generated by molecular dynamics simulation to show that it also works well for phase separation of disordered proteins (Supplementary Fig. 8). GMMA showed pronounced sequence-specific effects where substitutions in the C-terminal part of the region had less pronounced effects (Fig. 2b). However, as expected from the scrambled data, the composition of the mutated peptides was the major determinant of their partitioning, as shown by the correlation between predictions made by taking the sum of amino acid transfer free energies compared with the data (Fig. 2c). Adding the information on sequence position, such that the model used $\Delta\Delta G_{\mathrm{part}}$-values, the predictions came close to a complete description of the data given the experimental noise.

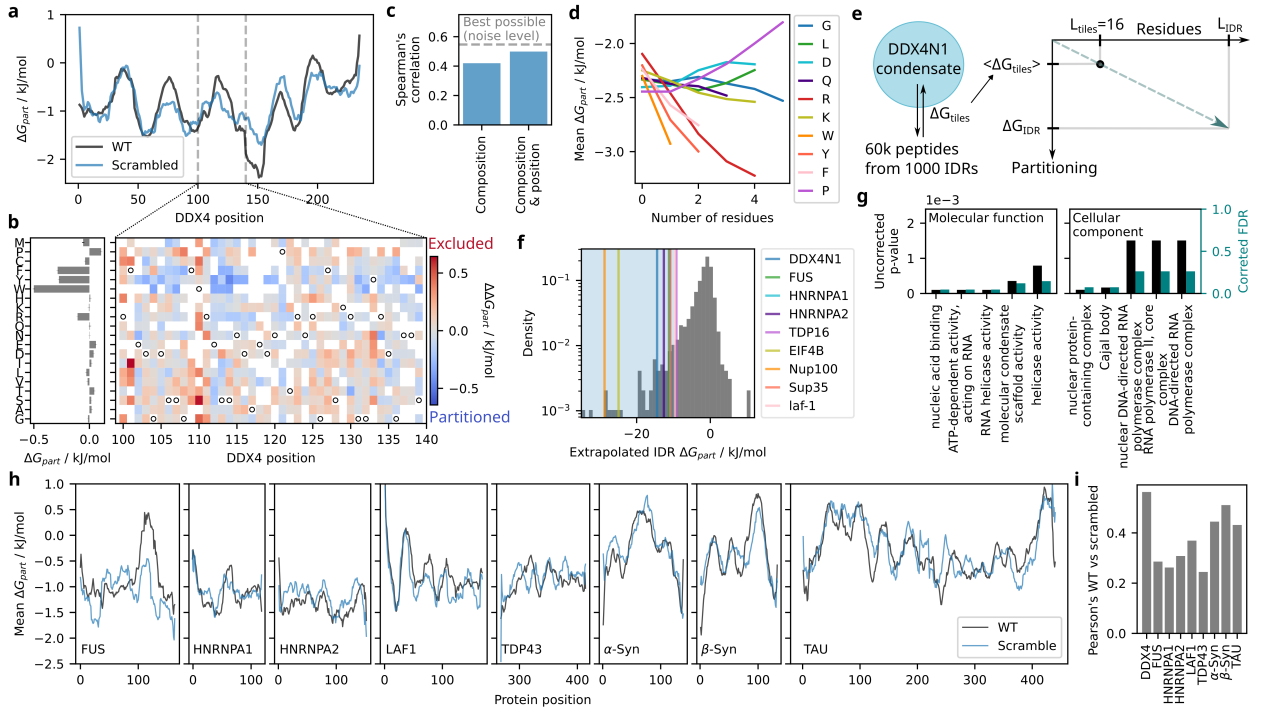

Figure 2: The same "sequence grammar" governs disordered proteome partitioning and DDX4N1 condensate formation. a A high resolution tiling experiment of partitioning of 16 residue peptides of DDX4N1 into full length condensates. For every peptide tile, two scrambled peptides were included to probe the dependence on sequence rather than composition. The running mean was calculated either for the WT or for the scrambled peptides. **b**, The effect on partition free energy $\Delta\Delta G_{\text{part}}$ of a large fraction of all possible substitutions of amino acids in a peptide comprising residues 100-139 of DDX4N1 was measured via the generation of more than 20,000 multi-mutants (Supplementary Fig. 7). **c**, The capability of linear models based on the amino acids composition (left, 20 fit parameters) and amino acid composition and position (right, 800 fit parameters) to explain the multi-mutant data. **d**, Dependence of partition free energy of the 16 residues long peptides on the number of different amino acids they contain. **e**, Peptide tiles from a library of the experimentally verified intrinsically disordered proteome (DisProt database) were measured. The partition free energies of the full length IDRs were then obtained from the average partition free energy of the 16 residues peptide tiles from a given IDR and extrapolated to the length of that IDR. **f**, Distribution of the extrapolated IDR partition free energies from the DisProt database. Only 2.6% partitioned similarly or more strongly than DDX4N1 itself, and about 5% would be expected to partition more than 50-fold. Yeast proteins are denoted by a "p"-prefix. **g**, GO-term enrichment analysis on genes with IDRs more partitioned than -5 kJ/mol (13% of data) shows that favourable partitioning into the DDX4N1 condensates correlates with functions in nucleic acid biology. **h** Comparison of the condensation propensity profiles calculated from averaging of tiles of 16 residues over WT and scrambled sequences for a range of other proteins known to form condensates. **i**, Correlation of the average partition free energy between wild-type (WT) peptides and the two scrambled peptides with the same amino acid composition.

Other phase separating IDRs partition to DDX4N1 condensates

The partitioning of peptides from DDX4N1 and closely related sequence variants into WT DDX4N1 condensates reveals a specific "sequence grammar" of that IDR. Other than specific interaction motifs, a particular overall composition needed for interaction could be encoded in DDX4N1, which could differ for other phase separating IDRs[19]. We therefore wondered which sequence regions of the IDRs of other proteins, if any, would enable them to partition into DDX4N1 condensates, and if the amino acid composition of those regions would be different from those that governed partitioning of intrinsic DDX4N1 peptides.

Because of the unprecedented scale of CPmD, we were able to measure IDR partitioning at the proteome-scale. We measured ∼55,500 tiles of 16 amino acid length with a four-fold overlap (4 amino acids skip sites) of all IDRs from the manually curated database of intrinsically disordered proteins (DisProt)[48]. In addition to this set of peptides, we added ∼4,700 high resolution (16-fold overlap, 1 amino acid skip sites)

tiles of select IDRs that are known to be involved in condensate formation *in vivo* and/or *in vitro* including scrambled versions of the peptides, together with β-synuclein as a control that has been reported not to form condensates[49].

On average, this library partitioned significantly less than peptides from DDX4N1 itself, but the dependence on amino acid composition was very similar, with increasing hydrophobic amino acid content favoring partitioning (Fig. 2d, Supplementary Fig. 6). Assuming linearity of partitioning with respect to length as demonstrated for DDX4N1 (Fig. 1h), we estimated the partition free energy of all the IDRs in the library by extrapolation (Fig. 2e,f). The obtained values showed that ∼10% of the IDRs would be more than 10-fold enriched (ΔG_{part} < -5.6 kJ/mol) in DDX4N1 condensates, and this subset of the database was enriched in gene ontology (GO)-terms related to RNA-biology and biomolecular condensates (Fig. 2g). Several IDRs known to form biomolecular condensates were found to be strongly partitioned into DDX4N1 condensates. In line with recent observations[50], we only found a weak correlation with proteins found in biomolecular condensates listed in the CD-CODE database[51] and this correlation was driven primarily by proteins listed to be present in many different condensates (Supplementary Fig. 6). Several different kinds of macromolecular interactions can recruit protein to condensates, (e.g. interactions between folded domains, protein and DNA, protein and RNA). Driving forces for condensate formation encoded in an IDR thus seem to either cause a protein to be a scaffold or be recruited to many different condensates with low specificity.

Restricting the data further to IDRs which would be more than 50-fold enriched in DDX4N1 condensates (ΔG_{part} < -9.5 kJ/mol) led to the identification of 21 human proteins. Of these, 8 have been shown able to form condensates in their purified forms (AFF4, DDX3X, EIF4B, EWSR1, FUS, HNRNPA1, HNRNPA2B1, MAPT), 4 participate in condensate formation (BRCA1, LAT, NOLC1, and SRRM1), ATG9A interacts with lipids, the annotated RICTOR IDR likely contains folded domains, and 7 have no described condensate behaviour but have no obvious reasons for dismissal (AHR, AHCTF1, FGA, IBSP, LOX, NBS, RBM7, and SECISBP2). Several proteins from other organisms known to phase separate, e.g. Nup100 and Sup35 from yeast and laf-1 from *Caenorhabditis elegans*, were also present in this subset (Fig. 2g).

Unexpectedly, data on the proteins included in the high resolution tiles (Fig. 2h) showed a larger disagreement between WT and scrambled tiles than DDX4N1 itself (Fig. 2i). The reason for this is unclear, but it supports the conclusion above that sequence motif-specific interactions are not a main driver of DDX4N1 phase separation. DDX4N1 has been shown to be completely disordered[20], but it is possible that several of the other proteins have transient secondary structure elements that influence partitioning and which are disturbed upon scrambling. Such subtle IDR features are uniquely captured by CPmD, and as many of these other IDRs form condensates that mature into amyloid fibrils, these data might help further understand such reactions that are often related pathology.

These results showed that many IDRs unrelated to DDX4 with known phase separation capabilities were strongly partitioned into DDX4N1 condensates.

DDX4N1 partitioning rationalises condensate formation of unrelated IDRs

We have shown that the overall tendency of DDX4N1 to form condensates can be extrapolated from partitioning of a tiled library of peptide fragments of this protein. Furthermore, sequence regions of particular importance for condensate formation were identified in such tiling experiments. Building on these insights, we set out to explore in more detail the idea that partitioning of mutated peptides is indeed a direct proxy for the effect of those mutations on condensate formation of the full-length IDR. We created mutated versions of DDX4N1 carrying some of the substitutions quantified in the GMMA-analysis, as well as some deletion variants, and measured their phase separation using Taylor Dispersion Induced Phase Separation (TDIPS)[56] (Supplementary Fig. 5). There was an overall good agreement, but the effects on intrinsic phase separation were in general stronger than for partitioning. We reasoned that this is likely to be caused by substitutions causing a double effect, whereby both the intrinsic propensity of the peptide to partition into a condensate and the properties of the resulting condensate itself are modified. Therefore, we developed a theoretical framework for phase separation that separated these two effects (Fig. 3a and Supplementary Fig. 9). A sim-

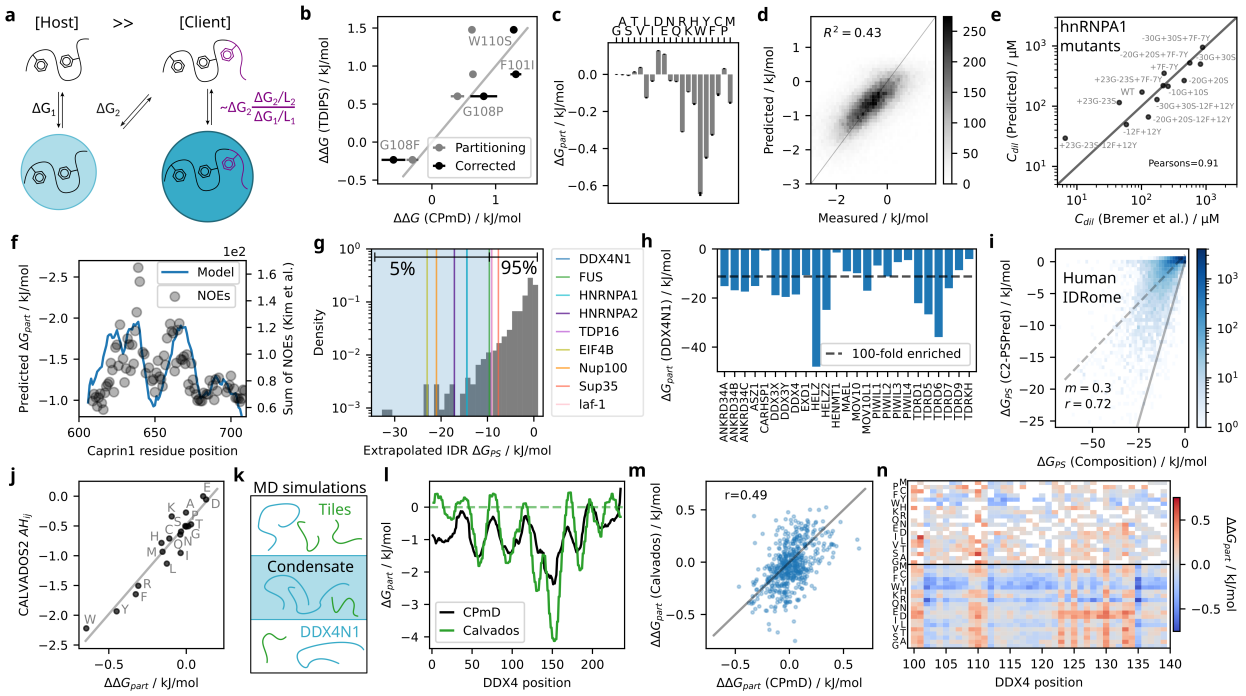

Figure 3: Peptide partitioning into DDX4N1 captures general properties of IDR condensate formation. **a**, We extrapolate the intrinsic phase separation propensity from data on the free energy of partitioning of a client to DDX4N1 (host) condensates ΔG_2. The extrapolation rest on the assumption that the ability of a condensate formed by a host peptide to recruit a client peptide is proportional to the energy density of the host peptide $\Delta G_{host}/L_{host}$. This results in a scaling of ΔG_2 by the energy density of the two condensates to obtain the intrinsic phase separation propensity of the client (see Supplementary Information). **b**, Correlation between changes in phase separation free energy as measure on purified mutants using TDIPS and the energies obtained from CPmD, with and without the correction described in a. **c**, Free energies of transfer for each amino acid as fitted using a composition model to explain partitioning of the DisProt library. Pairwise terms are shown in Supplementary Fig. 6. **d**, Performance of the model. **e**, The peptide partitioning model based on composition is capable of explaining differences in dilute phase concentrations of charge-conservative mutations of hnRNPA1-LCD[27]. **f**, A tiling experiment (as in Fig. 1f) was simulated of CAPRIN1-LCD using 10 residues long tiles, which agrees with interaction hot spots as measured by dense phase Nuclear Overhauser Effect (NOE) signals obtained by Kim et al.[52]. **g**, Correction for condensate energy density allows for estimation of the intrinsic phase separation propensity, ΔG_{PS} of the IDRs measured in the DisProt library (Fig. 2). **h**, This model was used to predict partitioning of known germ granule proteins to DDX4N1 condensates. Several of the components are predicted to partition more strongly than 100-fold (interrupted black line) exclusively based on the disordered regions. **i**, Prediction of the phase separation propensity of the human IDRome[53] based on the composition model compared with predictions by CALVADOS2-PSPred[45]. The predictions correlate but the composition model predicts 3-times stronger energy to form condensates. **j**, Correlation between amino acid partition free energies from the model based on CPmD data and the volume-integrated non-electrostatic potential energies ($AH_{i,j}$[45]) in a coarse-grained simulation model optimised to describe IDRs[54, 55]. **k**, The CALVADOS2 force field was used to perform coarse-grained molecular dynamics simulations of the CPmD experiments, in the absence of mRNA and linker. **l**, Correlation between single-substitution effects in the central 40 residues long peptide of DDX4N1 using CPmD and CALVADOS. **m**, Overlay of partition profile for 14 residues tiles of DDX4N1 from CPmD and CALVADOS simulations. The top of the dynamic range for the simulations is indicated by a horizontal interrupted line. **n**, Heatmap of data in l showing CPmD (top) and CALVADOS simulations (bottom).

ple assumption, that the propensity of a condensate formed by a scaffold IDR to recruit a client peptide/IDR is directly proportional to the phase separation energy density (ΔG_{PS} per residue), was sufficient to reconcile the differences between partitioning and intrinsic phase separation (Fig. 3b and Supplementary Fig. 5).

This quantitative link between partitioning and intrinsic condensate formation propensity suggested that a model capable of predicting peptide partitioning to DDX4N1 directly from a sequence of amino acids would define a molecular grammar for phase separation in the context of DDX4N1. Inspired by the good performance of a composition model for explaining the DDX4N1 mutant peptide partitioning, we fitted a similar model to predict partitioning of all peptides from the DisProt database (Fig. 3c,d). We also included terms for

pairwise interaction of neighbouring amino acids, but this extended model only mildly improved performance (Supplementary Fig. 6). The model was able to explain 43% of the variance in the data (Fig. 3d), which was enough to get good correlation between the extrapolated IDR partition free energy from the DisProt library and predictions directly from sequence (Pearson's=0.85, Supplementary Fig. 6), and was highly similar to the model fitted to the mutant-library data (Supplementary Fig. 6). Combining the predicted partitioning with the correction describe above to convert to predicted phase separation improved the ability to rationalise the effect of deletions in DDX4N1 compared to partitioning alone (Supplementary Fig. 6), supporting the idea that the model could predict phase separation. Furthermore, we argued that this combined model could in principle be used to predict phase separation directly from the sequence of other unrelated IDRs, if they shared common types of molecular interactions responsible for condensate formation with DDX4N1. The chemical environment within condensates has recently been shown to be well described as akin to a hydrophobic solvent with different strengths to partition small molecules[57]. Additionally, the parameters for the 20 natural amino acids can be interpreted as the free energy of transfer for moving that amino acid from an aqueous phase to a DDX4N1 condensate phase (Fig. 3c). These energies were recently shown to vary little between different simulated condensates[58], which supports our hypothesis that we might be able to extrapolate from the DDX4N1 partitioning data. To explore this, we compared the predictions of this model with experimental data of the dilute phase concentrations of mutants of the unrelated hnRNPA1 low complexity domain (LCD) with similar charge measured at 20°C [27], which showed good agreement (Pearson's=0.91, Fig. 3e, Supplementary Methods). However, data at 4°C with varying charge were less well explained (Pearsons's=0.48, Supplementary Fig. 6), which could be caused by temperature effects that are currently not part of our model. As a further demonstration of the plausibility of our model, we simulated a tiling experiment with peptides of 10 amino acid-length of CAPRIN1-LCD partitioning into condensates formed by that IDR. The observed interaction profile was strongly correlated with the interaction hot spots defined by Nuclear Overhauser Effect (NOE) signals in NMR measurements on the condensed phase[52] (Fig. 3f).

The good performance of our partitioning model for DDX4N1, hnRNPA1-LCD and CAPRIN1-LCD, made us confident in the assumptions needed to extrapolate from partitioning to intrinsic phase separation. We therefore used the energy density (as described in Fig. 3a) to correct the extrapolated IDR partition free energies measured in the DisProt library to represent homotypic condensate formation. This analysis showed a similar overall distribution of energies compared to that for partitioning into DDX4N1 condensates; however, the IDRs were ordered differently because of their differing predicted energy densities of the condensed phase (Fig. 3g). Interestingly, these data can explain the phase separation of prion-like LCDs, such as FUS, hnRNPA1, hnRNPA2, TDP43, which have an amino acid composition very different from that of DDX4N1. Even the yeast prion protein Sup35, which is also known to phase separate in addition to form a prion state, has an energy of ∼-8 kJ/mol, sufficient to support condensate formation with addition of molecular crowding as was previously shown[59]. The partitioning model also allowed us to predict partitioning to DDX4N1 of IDRs that were not included in the DisProt library, and we found that several of the proteins expected to co-localise with DDX4N1 in germ granules contained IDRs predicted to cause strong partitioning (Fig. 3h). To explore the capabilities of the model more broadly, we predicted the phase separation propensity of the entire human IDRome[53], and correlated it with the predictions of a model trained from coarse-grained molecular dynamics simulations with CALVADOS, PSPred[45] (Fig. 3i). We found a strong correlation, albeit with a slope different from unity. Our composition-based model, which does not contain electrostatic effects, overestimates the driving force for condensate formation by approximately three-fold compared with PSPred. Nevertheless, these results demonstrate that the partitioning data can, to a substantial extent, be generalised to explain homotypic phase separation.

We argue that the amino acid specific free energies of transfer in the partition model represent a hydrophobicity scale analogously to earlier work on the coacervation of mutated elastin-like peptides[22], despite having significant differences (r=0.71, Supplementary Fig. 6). Instead, the correlation was much greater with the "stickiness" (λ) parameters, as well as an integrated form of the pair potential (AH[45]), of the coarse-grained model of intrinsically disordered proteins, CALVADOS, that was optimised using experimental data [54] (Fig. 3j). This strong correlation further substantiates that the experimentally determined partitioning parameters are general and relatively transferable beyond the DDX4N1 system. In order to explore further

the similarity between our fitted contributions of the amino acids to the free energy of partitioning and the coarse-grained molecular simulation model, we carried out simulations of the partitioning of tiled peptides into an excess of full DDX4N1 (Fig. 3k and Supplementary Fig. 8). Data on the single peptide level showed low correlations (Supplementary Fig. 8) but this improved significantly when taking the running average along DDX4N1 (Fig. 3l). We also simulated experiments with the multi-mutant peptides of DDX4N1 and this showed over-all good agreement (Fig. 3m), particularly on the importance of aromatic amino acids, but, as expected, the coarse-grained simulations showed a stronger dependence on the effect of charged amino acids compared to the CPmD data (Fig. 3n, Supplementary Fig. 8).

Conclusions

The evolution of genes that encode folded proteins is largely constrained by the need to have a sufficient free energy difference between unfolded and folded states to robustly fold[60]. The realisation that IDRs can function via phase separation to form biomolecular condensates raises the question how such function is encoded and maintained in the sequence[61]. To address these questions that require significantly larger data than currently available, we have adapted mRNA-display to measure the partitioning of peptides, and their corresponding synthetic mRNAs, at proteome-scale into reconstituted biomolecular condensates *in vitro* in a novel workflow that we call CPmD. Applying CPmD to a prototypical IDR capable of forming liquid condensates, DDX4N1, allowed us to obtain data of unprecedented resolution on its ability to form condensates. We find that systematically tiling the IDR into short peptides allows for the identification of regions that have stronger propensity to partition into condensates of the full-length IDR, and that this propensity is governed to a large degree by amino acid composition. We also find that the phase separation of DDX4N1 can largely be explained by the affinity of its constituent peptides for the condensates, establishing the link between partitioning and intrinsic phase separation propensity. Exploiting this link, we use random mutagenesis to probe the effect of mutations in a core region of the domain. This data, together with molecular dynamics simulations, reveal that the effect of substitutions are largely additive for phase separation, as is the case for protein folding when considering two-state folding[62, 63]. We further measured partitioning of tiled peptides across all experimentally characterised IDRs in the Disprot database, which showed selective recruitment of other phase separating IDRs, in particular from proteins involved in RNA biology, into DDX4N1 condensates. We show that simple linear models perform surprisingly well to predict peptide partitioning, but we expect that more advanced models could also be trained on our large-scale data.

Our results indicate that energetic effects from peptides can be added together to explain phase separation of long IDRs, but also that amino acid-specific partition energies can to some degree add together to explain peptide partitioning. To show the consequence of this, we present a relatively simple model that nonetheless can explain quantitatively the phase separation of several unrelated IDRs. However, we were not able to detect contributions from peptide net charge or charged-patch separations likely because of the strong negative charge of the mRNA. Protein net charge is important for complex coacervation[64], and close-to-neutral proteins are generally less soluble which also applies to homotypic phase separation[27, 54]. While charge patterning can also affect homotypic phase separation[9, 41, 45, 65], the charges in naturally occurring IDRs are generally well mixed[46, 65] and so likely only play a role in a smaller number of proteins. Further studies should aim to evaluate the relative contribution of these interactions for different systems.

The results obtained by applying CPmD to one particular type of condensate-forming protein as a scaffold, DDX4N1, proved to be surprisingly general. This is most clearly highlighted by the strong correlation between the partition free energies of the individual amino acids and the CALVADOS 'stickiness parameter' λ derived from a Bayesian optimisation using SAXS data of multiple different IDRs, and using all known hydrophobicity scales as priors. However, in order to achieve full generality of the derived parameters, CPmD needs to be applied to additional condensates-forming IDPs with distinct sequence compositions. As the cost of DNA synthesis and sequencing becomes ever cheaper, two-to-three orders of magnitude larger scale experiments should be accessible using CPmD. DNA synthesis and sequencing is currently also a bottleneck for evaluating longer IDRs or full proteins, which is in principle possible with CPmD. This should allow for investigation of reconstituted biomolecular condensates at a complexity approaching that of the cell, and

could also include folded protein domains. This could help understanding how specificity is encoded such that multiple condensates coexist. Furthermore, we envisage that CPmD might be capable of providing data at the scale needed to allow machine learning-enabled protein engineering to cover also disordered proteins, which could create peptides capable of perturbing condensates or directing cargo to partition to a given condensate.

Acknowledgements

We are grateful for the assistance with performing Illumina sequencing runs received from PhD Marlene Danner Dalgaard from DTU Multi Assay Core and Malene Taarngaard Nørager and Anne Meldgaard Hansen from the Kennedy center at Rigshospitalet, Copenhagen. We also thank Soumik Ray for assisting with confocal microscopy. We thank Gerhard Steger for help with the RNA structure prediction. A.K.B. and R.K.N. would like to acknowledge the Villum Foundation (grant number: 35823), the Novo Nordisk Foundation (grant number: NNFSA170028392), and the European Union (ERC CoG 101088163 EMMA) for funding. J.M.R. was supported by the Novo Nordisk Foundation (grant number: NNF19OC0054441). S.v.B. acknowledges support by the European Molecular Biology Organisation through Postdoctoral Fellowship grant ALTF 810-2022. K.L.-L. acknowledges support by the Novo Nordisk Foundation via the PRISM (Protein Interactions and Stability in Medicine and Genomics) centre (NNF18OC0033950). We acknowledge access to computational resources from the Biocomputing Core Facility at the Department of Biology, University of Copenhagen, from the Resource for Biomolecular Simulations (ROBUST; supported by the Novo Nordisk Foundation; NNF18OC0032608), and the Danish National Supercomputer for Life Sciences (Computerome).

Competing interests

R.K.N., J.M.R., and A.K.B. share a part in a patent application based on the technology developed in this manuscript. K.L.-L. holds stock options in and is a consultant for Peptone. A.K.B. is a member of the scientific advisory board of Nuage Therapeutics.

Data availability

Data and code will become available on Zenodo at `https://doi.org/10.5281/zenodo.14534153`.

Methods

DDX4N1 CtoA expression and purification

A mutant of the N-terminal domain of Dead-box helicase 4 (DDX4N1) was used in which all cysteine residues were mutated to alanines (CtoA), as this simplifies handling of the protein and removes the potential for disulphide bridge formation and has little or no effect on LLPS[56]. Purification was performed through the recombinant expression of DDX4N1 CtoA fused to a *de novo*-designed thioredoxin-tag, eMM9 ([47]). Briefly, eMM9 is thought to facilitate protein expression through its high expression level and its high thermodynamic stability. The stability confers resistance to both heat and high denaturant concentrations useful for purification strategies used for intrinsically disordered proteins. A pET29(a) plasmid was constructed using Golden gate cloning containing an open reading frame coding for C-terminally His-tagged eMM9 followed by codon-optimized DDX4N1 CtoA linked by a 10-residue GS-linker and a TEV-cleavage site (ENLYFQ/G). The plasmid was used to transform *E. coli* strain BL21 (DE3) (New England Biolabs), and a clone was selected on kanamycin LB-agarose plates and Sanger-sequenced to confirm the correct sequence. 10 ml overnight culture was prepared in LB-medium supplied with 50 mg/l kanamycin and was used to inoculate 1l AB-LB (LB-medium with 5 g/l NaCl supplemented with 40 mM sodium/potassium phosphate (K/NaPi) buffer, pH 7.0, 15 mM ammonium sulphate, 50 mM NaCl, 2 mM MgCl2, 0.1 mM CaCl2, 3 nM FeCl3, 50 mg/l kanamycin).

The culture was grown at 37°C to OD600 1.6 and induced with 1 mM IPTG before incubating overnight at 20°C. The next day, the cells were harvested by centrifugation at 7000 xg for 20 min at 4°C and the pellet was frozen and stored at -20°C. Next, 30 ml lysis buffer (50 mM NaPi buffer, pH 6.5 with 500 mM NaCl) was used to resuspend the cell pellet and 8 μl Benzonase (purity > 90%, Merck-Millipore) were added. Cells were lysed by 30s/30s sonication pulses for a total time of 30 min on ice. Insoluble debris was then removed by centrifugation at 20000 xg for 20 min at 4°C and the supernatant was then heat-treated by incubation at 80°C for 15 min. Heat-induced protein aggregates were cleared by centrifugation at 4500 xg for 30 min at 4°C and passed through a 0.22 μm filter. 1 M imidazole was added to a final concentration of 10 mM followed by quick mixing. The solution was then loaded onto 5 ml pre-equilibrated Ni-NTA resin (ThermoFisher Scientific) on a gravity flow column and washed with 10 column volumes of a buffer containing 50 mM NaPi, pH 6.5, 500 mM NaCl, and 20 mM imidazole, then 10 column volumes denaturing buffer containing 50 mM NaPi pH 6.5, 500 mM NaCl, 20 mM imidazole, and 3M GdnHCl. Finally, the denaturing salt was washed out with 10 column volumes of the non-denaturing wash buffer. A protease solution (9.5 ml lysis buffer with 2 mM TCEP and 500 μl His-tagged TEV protease (10KU, Genscript or in-house purified)) was then added to release DDX4N1 CtoA into the solution during over-night incubation at room temperature with gentle mixing. TEV protease stays bound, and the eluted solution was concentrated using a 10 kDa centrifugal filter to 2 ml and SEC purified on a HiLoad Superdex 75 16/600 column equipped to an ÄKTA Pure system in a cooled cabinet at 8°C. Pooled fractions were concentrated to 500-600 μM and aliquots were flash frozen in liquid nitrogen before storing at -80°C. The protein purity was checked at different stages of purification using SDS-PAGE and the protocol yielded ∼14 mg protein per litre culture.

Plate based purification of mutated protein variants

Primers for generating mutants via Golden Gate cloning were ordered from IDT (Germany) at 100 μM concentration TE buffer in a 96-well plate. The primers targeted 30 different substitutions and deletions of DDX4N1. They were then diluted to 1 μM in MilliQ to use for PCR amplification of the pET29a plasmid encoding eMM9-tagged DDX4N1 CtoA. 20 μL PCR reaction was set up using 1x Phusion master mix (F531L, ThermoFisher Scientific), 0.04 ng/μL plasmid, and 50 pM primers. The PCR reaction was done in a Opentrons pipetting robot equipped with a thermal cyler (Opentrons, USA). 30 cycles were run with initial denaturation for 30 s at 98°C, final extension for 10 min at 72°C, and cycling through: 5 s at 98°C and 240 s at 72°C. The PCR reaction was assessed on 1% agar gel and the successfully produced PCR products were pooled together, 3.7 μL of 27 products, totaling 100 μL. The pooled PCR product was cleaned with QIAquick PCR clean-up kit(Qiagen, Netherlands). The product was then ligated via Golden Gate Assembly in an overnight ligation reaction at 37°C with the following final concentrations: 1000U T4 DNA ligase (M0202, NEB, USA), 1x T4 DNA ligase buffer, 30U Bsal-HFv2 (R3733, NEB, USA), 192 ng PCR product. The reaction was then inactivated by heating to 60°C for 5 min. 1 μL of Dpn1 (FD1703, Thermo, USA) was added and incubated for 1 hour at 37°C, and then inactivated at 80°C for 5 min. 50 μL of BL21 (DE3) cells (NEB, USA) were transformed with the product and colonies formed were transferred to a 96-well agar kanamycin plate to be sequenced by Eurofins Genomics (Germany) and to a deep well plate for glycerol stocks.

The sequencing resulted in 23 unique sequences of mutants that were then used for expression together with DDX4N1 CtoA as a control. This was done in two 24 deep-well plates (2x2 ml of each variant). 20 μL of overnight cultures made from the glycerol stocks were transferred to 2 mL of Studier ZYM-5052 medium (Teknova, USA) with kanamycin. The cultures grew in a Eppendorf ThermoMixer shaking at 800 RPM at 25°C for 24 hours. Cells were harvested in 2 mL tubes by centrifugation at 7000 g for 10 min. The supernatant was discarded, and the 2 mL from the other plate was added and centrifuged again. 500 mL wash buffer (20 mM NaPi, 500 mM NaCl, 20 mM Imidazole, pH = 6.5) was added to resuspended cells. The lysis was then performed at 80°C for 20 min, followed by centrifugation for 20 min at 20,000 g to remove debris and aggregates. The supernatant was purified using 200 μL NTA-beads (88222, Thermo, USA) in a filter plate using a vacuum manifold. After washing, 500 μL of the supernatant was loaded onto the plate, allowed to sit for 10 min before vacuum was applied, and then washed 3x with 200 μL wash buffer + 3M GdnHCl. Then, several more wash steps were performed with wash buffer without GdnHCl. Finally, 100 μL of TEV

cleavage solution (CombiTEVp 11 μM (in-house purified) in wash buffer supplemented with 0.5 μM TCEP) was loaded and incubated overnight. The protein was then eluted using three additional 100 μL washes with wash buffer and concentrated with 30 kDa filters (UFC503096, Merck Millipore, USA) to 30 μL. The produced proteins were quality checked using SDS-PAGE (Supplementary Fig. 5) and their hydrodynamic radius was measured using FIDA.

TIDPS measurements of critical concentration

All measurements were performed on the Fida 1 instrument (Fida Biosystems, Denmark) using a 50 cm long capillary with a 75 μm diameter for TDIPS, and the standard capillary of 1 m. Samples were injected from a compatible V-bottom 96-well plate and from a standard 50-vial holder, and detected using a UV detector with a 280 nm LED for excitation and a 300 nm long pass filter for emission. The signal detection was performed with 450 V detector gain with a recording frequency of 90 Hz.

Tray	Vial	Pressure (mbar)	Time (s)	Outlet	Measure	Comment
2	1	3500	45	Variable	No	1M NaOH Wash
1	Analyte	3500	45	Variable	No	Buffer wash
1	Indicator	50	15	Variable	No	Protein loading
1	Analyte	300	90	Variable	Yes	Mobilize and measure

Table 1: FIDA method used for measuring TDIPS curves of DDX4N1.

Partition coefficients of solid-state synthesised peptides

To validate the mRNA-display approach, we bought six peptides with 14 residues containing and additional N-terminal cysteine residues with a maleamide Oregon Green 488 fluorophore (Schafer-N, Denmark). 100 μM stock solutions were obtained by solubilising the peptides in MilliQ water, and were subsequently diluted to 100 nM for the experiment in a buffer containing 20 mM NaPi and 0.05%(V/V) Tween-20 pH 6.5. A 300 μM stock of DDX4N1 was prepared by mixing 50:50 with a buffer containing 20 mM NaPi, 500 mM NaCl, 0.05%(V/V) Tween-20 pH 6.5. 16 μL peptide stocks were thoroughly mixed with 4 μL 300 μM DDX4N1 for a final concentration of 80 nM peptide, 50 μM DDX4N1, and 100 mM NaCl. The solution was immediately measured on the Fida 1 (Fida Biosystems, Denmark) at 20°C in a standard capillary (1 m and 75 μm diameter) using the 480 nm LED detector with a gain of 420 V and a measuring frequency of 90 Hz. Control samples without protein were also measured and the change in the dilute phase was quantified. All samples were measured in duplicates. Using a condensed phase volume fraction of 0.22%, the partition free energy of the peptides was calculated. The quantification of spikes (signal from condensates) was in agreement with the data from the dilute phase concentration but was not used for direct quantification.

Tray	Vial	Pressure (mbar)	Time (s)	Outlet	Measure	Comment
2	1	3500	45	Variable	No	1M NaOH Wash
2	2	3500	20	Variable	No	MilliQ wash
2	5	3500	30	Variable	No	Buffer wash
1	Indicator	300	300	Variable	Yes	Measure sample

Table 2: FIDA method used for measuring Capflex of DDX4N1 with fluorescently labelled peptides.

Microscopy

1 nM mRNA-display libraries containing a FITC dye were mixed with 120 μM DDX4N1 at 100 mM NaCl in NaPi buffer pH 6.5 with 0.05% Tween-20. Brightfield and confocal micrographs where recorded using a

Leica SP8 confocal microscope (Germany) using an 63x/1.40 oil immersion objective at the DTU bio-imaging core facility. Confocal images were obtained with a 16-bit depth. An excitation wavelength of 488 nm and emission channel of 500–600 nm was used. The gain was set to 1121V for the fluorescent channel and the sensitivity was set to 85% to detect the faint signal from the mRNA-display libraries. Images were recorded on rectangular glass coverslips 24x50 mm Menzel gläser which where plasma cleaned for 1 minute using the PDC-002-CE plasma cleaner from Harrick Plasma (USA). Plasma cleaned PDMS gaskets with 3 μL circular wells were attached to the coverslips and used for imaging.

mRNA-display

Fragment library from oligo pool

To investigate which parts of the DDX4N1 sequence participated most strongly in the condensation, we designed a tiling library consisting of many overlapping fragments of the protein domain. The library was designed with fragments that ranged from 14 amino acids in length and increased by two to 50 amino acids (l = 14,16,18... 50). For each length, the overlapping fragments were placed to start from every other amino acid, so that for 14 amino acid length fragments, each amino acid in DDX4N1 sufficiently far from the N and C-terminus was present in 7 different fragments. The library was synthesised by Twist Bioscience (USA) with 5' and 3' regions designed for PCR amplification. This PCR step is recommended by the supplier, but also allowed the attachment of the remaining 5' and 3' regions needed for mRNA-display[36]. The library was solubilised to 2.5 ng/μL in TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH=8) and 1 μL was added for a total of 100 μL PCR solution (0.5 μM `Fivep_flank`, 0.5 μM `Threep_flank_HA`, 1X Phusion HF buffer, 0.2 mM dNTPs, and 0.02 U Phusion (NEB M0530, USA)). 14 cycles were run with initial denaturation for 30 s at 98°C, final extension for 60 s at 72°C, and cycling through: 10 s at 98°C, 30 s at 56°C, and 40 s at 72°C.

Mutation library from doped oligo nucleotides

Doped oligos are oligo nucleotides synthesised with error, by adding a small amount of each of the other three nucleotides to each vial for synthesis. The resulting synthesised products will be highly diverse, with almost all molecules being unique, given a sufficiently high mutagenesis rate. We designed two such oligos, `Pep6_lib_fw` and `Pep6_lib_rev`, to generate a randomly mutated library of a central region in DDX4N1 positions [100:139]. The two oligos were designed with constant regions to anneal to `Fivep_flank` and `Threep_flank_HA` respectively for PCR amplification. Additionally, the two oligos overlapped with 15 base pairs at their 3'-ends allowing them to anneal and be extended to a double stranded product using the Klenow fragment polymerase (NEB M0210, USA). The extension reaction was performed in 100 μL with 1 μM of each oligo in NEBuffer 2 (50 mM NaCl, 10 mM Tris-HCl, 10 mM MgCl2, 1 mM DTT, pH 7.9). The oligos were annealed in the buffer but without dNTPs and polymerase by heating the solution to 94°C for 2 minutes and then ramping down the temperature at 0.2 °C/s down to 50°C. The solution was kept for 2 minutes and then ramped down to 37°C at the same speed. Then 2.5 μL 10 mM dNTPs and 1 μL Klenow fragment was added and the reaction was incubated at 37°C for 30 minutes. To limit the number of variants in the library, the 10 nM double-stranded product from the Klenow extension was diluted 10-fold (5 μL in 45 μL 0.05% triton X-100) 6 times to a theoretical maximum amount of molecules in the solution in 1 μL of ∼600,000. A PCR reaction similar to the one described above for the tiling library was performed using GC buffer instead of HF buffer, which was run for 33 cycles.

Disprot library

A 60,000 member tiling library of the Disprot database (release 2022_12) containing 16 residue-long peptides with 4-fold overlap was ordered from Agilent (G7270A, USA). The library was solubilised in 40 μL TE buffer yielding ∼1 μM concentration. It was then diluted 1000-fold into a 100 μL PCR solution as described above, and split in 2x50 μL during PCR amplification. 13 cycles were run with initial denaturation for 30 s at 98°C, final extension for 60 s at 72°C, and cycling through: 5 s at 98°C, 15 s at 66°C, and 30 s at 72°C. The two samples were combined after the PCR reaction.

Transcription

Before transcription, the template DNA was cleaned using phenol-chloroform extraction. Briefly, 3M NaCl was added to the solution to a final concentration of 300 mM. For a 100 μL solution, 100 μL (same volume) phenol-cholorform-isoamyl alcohol (PCI) at a 25:24:1 ratio pH=6.5 was then added before vortexing and centrifuging at 10,000 g for 4 minutes. The top aqueous phase was then transferred to a new tube, and cleaned by adding 100 μL chloroform-isoamyl alcohol (CI) at 24:1 ratio. Again, the solution was vortexed and centrifuged to separate the two phases. Finally, the aqueous phase was then transferred to a new tube and precipitated by adding ethanol to a final 70% (v/v). The pellet was washed with fresh 25 μL of 70% ethanol, dried for a couple of minutes, and then re-solubilised in 10 μL RNAse free water, for a theoretical maximum of 5 μM (assuming complete PCR and no loss in PCI/CI extraction). Transcription was performed using the Ribomax kit (Promega, USA) in a 50 μL reaction: 10 μL 5X buffer, 15 μL 100 mM rNTPs, 2.5 μL DNA template, 0.5 μL 1 mM DDT, 0.3 μL RNAsin Plus (Promega, USA), 5 μL Enzyme mix, and 16.7 μL RNAse free water. The reaction was incubated at 37°C overnight ($\sim$16 hours). To remove template DNA, 5.7 μL DNAse buffer (0.4M Tris pH 8.0, 0.1M MgSO4, 0.01M CaCl2) and 1 μL DNAse 1 (Thermo EN0525, USA) were added and the solution was incubated for a maximum of 30 minutes at 37°C. To stop the reaction, EDTA pH=8 was added to a final concentration of 50 mM and NaCl was added to a final concentration of 300 mM. The RNA was then purified using PCI/CI as described above using a PCI solution with pH=4.5. Instead of precipitation, the aqueous phase after CI treatment was transferred to a NAP-5 column (Cytiva, USA) equilibrated with 10 mL RNAse-free water. If adding a 70 μL sample, 430 μL of RNAse-free water was then added. Finally, 500 μL RNAse-free water was added to elute the mRNA. 0.3 μL RNAsin was added to the tube before storing at -80°C.

Puromycin ligation

Ligation to the puromycin-containing linker was done in 50 μL reaction containing 7.5 μM `FPur-Linker` and 5 μM mRNA. A 25 μL solution with 2x concentration of the two (15 and 10 μM, respectively) was heated to 95°C and slowly cooled to room temperature. Then, 5 μL 10x T4 RNA ligase buffer and 5 μL T4 ssRNA ligase 1 (NEB, USA) were added and the reaction was incubated for 1 hour at 37°C. 3 μL of the reaction was loaded onto a PAGE gel containing TBE (Tris-Borate-EDTA), 8M Urea, 20%(v/v) formamide, and 8% acrylamide 19:1 to evaluate ligation. EDTA was added to the remainder of the reaction to a final concentration of 50 mM and NaCl to 300 mM, and the solution was purified using PCI/CI as described above using pH 6.5 PCI solution. For the Disprot library, additional measures were taken to purify the ligated product from unligated mRNA. 2 μL 10% acetic acid was added to decrease the pH of the solution, together with 3 μL formamide to a final 5% to inhibit RNA-RNA interactions. Using pH 6.5 PCI solution, ligated mRNA ws found to be located between the lower phenol and upper aqueous phase. The aqueous phase was removed and the remainder was washed with 300 mM NaCl, which was then removed again after centrifugation. The purified product was released using 100 mM Tris-HCl pH 8.0 with 300 mM NaCl by vortexing. The aqueous phase was recovered after centrifugation and cleaned using CI as described above. The aqueous phase was then precipitated in 70% ethanol, washed with 70% ethanol, and solubilised in 7.5 μL RNAse free water. This solution also contained un-ligated mRNA, however, this was removed in the HA-tag purification step.

Translation

Translation of the ligated mRNA was performed using the PureExpress ΔRF1,2,3 kit (NEB, USA). A 25 μL reaction was performed using 10 μL solution A, 7.5 μL solution B, 5 μL template, and 2.5 μL RNAse free water. The solution was incubated for 1 hour at 37°C for translation to proceed. 3.25 μL 100 mM Mg(OAc)2 pH=8 were added to the solution to improve the efficiency of the puromycin reaction during 30 min incubation at 37°C. Finally, 8.25 μL 100 mM EDTA pH=8 were added to release the ribosomes during 30 minutes incubation at 37°C. The translation solution was then mixed 1:1 with 2X blocking buffer (100 mM HEPES-KOH pH 7.6, 400 mM NaCl, 0.1%(v/v) Tween-20, 0.1 g/L yeast RNA (Thermo, USA), 0.2 g/L acetylated BSA (Thermor, USA), and 5 μM `GS3an_R36`) and cooled on ice. 40 μL Pierce Anti-HA beads (Thermo, USA) was washed 3 times in ice cold wash buffer (50 mM HEPES-KOH pH 7.6, 200 mM NaCl,

0.05%(v/v) Tween-20, 1 μM `GS3an_R36`) using magnetic separation. The blocked translation mixture was used to re-suspend the pelleted beads and incubated for 1 hour at 4°C with rotation to allow binding. The supernatant was then removed after magnetic separation and two rounds of washing were swiftly performed using ice-cold washing buffer. A final wash was performed with elution buffer (20 mM NaPi, 500 mM NaCl, 0.05%(v/v) Tween-20, and trace amounts of RNAsin (Promega, USA)), before eluting the bound molecules by adding 20 μL elution buffer and incubating the beads at 80°C for 10 minutes. The solution was aliquoted and flash-frozen in liquid nitrogen. Quantification was performed using qPCR or fluorescence from the fluorescein on the puromycin linker yielding a typical concentration of ∼20 nM.

Partitioning experiments

An aliquot of the library to be screened for partitioning was diluted 10-fold with low salt buffer (20 mM NaPi and 0.05%(v/v) Tween-20) such that this solution could induce coacervation of DDX4N1 CtoA. In parallel, the mRNA used to create the library was diluted to the same concentration, usually ∼2 nM, to be used as a control. Both of these solutions were cooled on ice and then added to two different PCR tubes, one containing 1.5 μL 600 μM DDX4N1 in elution buffer and one with 1.5 μL elution buffer without protein, resulting in four tubes in total. All tubes were incubated for 1 minute at room temperature and then centrifuged for 10 minutes at 10,000 g in a pre-equilibrated centrifuge at 20°C. After centrifugation, 1 μL of the supernatant from each tube was immediately added to the reverse transcription solution (1x SS-IV buffer, 0.25 mM dNTPs, 0.75 μM `GS3an_R36`, 1.5 units of SS-IV reverse transcriptase (Thermo, USA), and trace amount of RNAsin (Promega,USA)). The reaction was proceeded at 50°C for 1 hour to ensure completion. The resulting solution was stored at -20°C.

Sequencing

Library preparation for sequencing

Attachment of Illumina adapter sequences was done through two consecutive PCR reactions. But first, the concentration of the template in each of the reverse transcription reactions was estimated using qPCR, such that over-cycling could be avoided. The first PCR reaction was performed in 50 μL using 0.15 μM of `Rd1T7g10M_F71` and `an13Rd2_R49`, 1X HF buffer, 0.2 mM dNTPs, 0.5 μL reverse transcription solution, and 0.5 μL Phusion polymerase (NEB, USA). A varying amount of cycles (usually 10-20) were run based on the number needed to get 0.05 μM double-stranded product. The PCR program included an initial denaturation for 30 s at 98°C, final extension for 60 s at 72°C, and cycling through: 10 s at 98°C, 30 s at 52°C, and 40 s at 72°C. For NovaSeq analysis, phasing version of the two primers with added random nucleotides to frame-shif the reads were used, such that all PCR products had the same length, e.g. 2 extra bases at one end and 6 extra bases at the other for a total of 8. For the second reaction, 0.15 μM of `Rd2N(701)P7_R52` and `P5S(502)Rd1_F57` were used in a similar PCR reaction using 2.5 μL of the former PCR reaction as a template in a 50 μL reaction. Each sample was assigned a unique combination of multiplexing barcodes in the two primers. Samples of FragLib2 were cleaned with QIAquick PCR clean-up kit (Qiagen, Netherlands) and sequenced using MiSeq v3 chip for 2x300 bps read at the DTU Multi Assay Core. Samples for the mutation library, Pep6Lib, were cleaned using Mag-Bind TotalPure NGS (Omega Bio-tek, USA) and sequenced using an SP Flow-cell on NovaSeq at the Kennedy center, Rigshospitalet Copenhagen.

Processing of data

NGMerge[66] was used to merge paired-end reads using the default configuration. For the fragment library, all merged reads were mapped to variants expected in the library, allowing a single base pair discrepancy originating from sequencing errors, using a custom Python script. For the mutation library, the reads were cut to only contain the mutated area and unique sequences were then counted using custom code written in C++. As the library variation is similar to sequencing errors, the only way to remove sequencing noise is to require a minimum amount of reads. This was implemented by requiring all variants to have at least ten reads and be present in all samples.

Analysis and modelling

Calculation of partition energy

The model rests on a number of assumptions which are expanded on in the supplementary materials but briefly mentioned here. 1) The relative abundance of sequencing reads is directly proportional to the concentration of each variant in the library. 2) The concentration in the condensate phase is exactly the molecules removed from the dilute phase divided by the volume fraction of the condensed phase. To avoid contamination between phases, we chose to measure only the concentrations of library members in the dilute phase (Fig. 1c) and infer the concentrations in the condensed phase using information on the total concentrations in the library and the dense phase volume fraction. 3) The system is at equilibrium, and the decrease in the concentration of a molecule in the dilute phase is directly due to its partition constant and, thus the free energy of partitioning. 4) The free energy of partitioning is a sum of the contribution from mRNA, linker, and peptide. Derivations of the formulae used to calculate the partition energy can be found in the Supplementary Information.

Fitting of linear models

Weighted ordinary least square regressions where used to explain partitioning of both peptides and RNA. Sequences were encoded either using their amino acid or nucleic acid composition in 20-member vectors or using one-hot encoding. These were all fitted in Python using linear algebra in Numpy. When needed, L2 regularisation (also called Tickhonov or Ridge regression) was used and regularisation strength was tuned using 5-fold cross validation. Potential over-fitting was evaluated on 10% held out data in all cases. Model performance and regularisation parameters are all reported in Supplementary Table 5.

GO-term enrichment analysis

To obtain the extrapolated partition free energy of all IDRs in the Disprot library, the average partition energy of the tiling peptides within one IDR was calculated. This value was then devided by 16 (the number of amino acids in the peptides) and multiplied by the length of the IDR. To obtain a single value for proteins with multiple IDRs, the one with the highest affinity for the DDX4N1 condensates was choosen. GO-term enrichment was performed using GOATOOLS[67] in Python, using genes from species with more than 100 entries in the disprot database. TaxIDs: 9606 (Human), 10090 (Mouse), 10116 (Rat), 4932 (Yeast), 3702 (mouse-ear cress), 7227 (fruit fly), and 6239 (*Caenorhabditis elegans*). Enrichment p-values where calculated using Fishers exact t-test and correction for multiple testing was done using the Benjamini and Hochberg procedure (Corrected false discovery rate (FDR)).

Coarse-grained molecular dynamic simulations

We performed coarse-grained molecular dynamics (MD) simulations with openMM 8.0[68] using the CALVADOS 2 force field, with force field parameters described elsewhere[54, 55]. Briefly described, the CALVADOS 2 force field represents IDRs using one bead per residue, with neighboring amino acid residues bonded via a harmonic potential. Salt-screened electrostatic interactions are represented via a Debye-Hückel potential. Non-electrostatic short-ranged interactions are described by a modified Lennard-Jones potential with residue type-specific sizes and stickiness values λ.

Each simulation was performed at constant molecule number, volume and temperature (Langevin integrator, T=293 K, friction coefficient=0.01 ps^{-1}) at 175 mM ionic strength and pH 6.5 in a tetragonal, periodic simulation box with edge lengths 20 nm x 20 nm x 300 nm. For WT tiling simulations, WT fragment tiles of 14 residue length were generated along the DDX4N1 sequence, beginning from every second residue (res. 1-14, 3-16, etc). We simulated a slab of 200 molecules of DDX4N1 together with two copies of all WT tiles in the same simulation box. The partitioning free energy ΔG_{part} for a given fragment tile was calculated using the ratio of average number of molecules in the dilute phase of the simulation box and the average number of molecules in the DDX4N1 condensate (dense phase) across the trajectory. The boundaries between dense and dilute phase were determined as described previously[55].

To estimate single substitution effects on fragment partitioning into DDX4N1, all possible single substitution variants of a peptide corresponding to DDX4N1 residues 100-139 were generated. Five copies of each of the 20 fragments corresponding to the WT and all single substitutions at a given residue position were simulated together with a slab of 200 DDX4N1 molecules. Each setup was simulated in four replicates, resulting in 4x40=160 simulations. Reported $\Delta\Delta G$ values are calculated as the mean partitioning ΔG of all 5x4 copies of a given variant fragment minus the mean partitioning ΔG of all 5x4x20 control WT fragments in the simulations.

We simulated a subset of 200 of the multi-mutant fragments for DDX4N1 residues 100-139 that only contained a defined subset of substitutions in batches of 20 different fragments together with a slab of 200 DDX4N1 molecules, as above. GMMA analysis was performed on these simulated data using the linear models described above similarly to how they were used for CPmD data.

References

1. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. en. *Nature Reviews Molecular Cell Biology* **18,** 285–298 (May 2017).
2. Kim, H. J. *et al.* Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. en. *Nature* **495,** 467–473 (Mar. 2013).
3. Tulpule, A. *et al.* Kinase-mediated RAS signaling via membraneless cytoplasmic protein granules. en. *Cell* **184,** 2649–2664.e18 (May 2021).
4. Garcia-Cabau, C. *et al.* Mis-splicing of a neuronal microexon promotes CPEB4 aggregation in ASD. en. *Nature* (Dec. 2024).
5. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. en. *Science* **357,** eaaf4382 (Sept. 2017).
6. Mittag, T. & Pappu, R. V. A conceptual framework for understanding phase separation and addressing open questions and challenges. en. *Molecular Cell* **82,** 2201–2214 (June 2022).
7. Li, P. *et al.* Phase transitions in the assembly of multivalent signalling proteins. en. *Nature* **483,** 336–340 (Mar. 2012).
8. Wright, P. E. & Dyson, H. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. en. *Journal of Molecular Biology* **293,** 321–331 (Oct. 1999).
9. Nott, T. J. *et al.* Phase Transition of a Disordered Nuage Protein Generates Environmentally Responsive Membraneless Organelles. en. *Molecular Cell* **57,** 936–947 (Mar. 2015).
10. Molliex, A. *et al.* Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization. en. *Cell* **163,** 123–133 (Sept. 2015).
11. Elbaum-Garfinkle, S. *et al.* The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. en. *Proceedings of the National Academy of Sciences* **112,** 7189–7194 (June 2015).
12. Schupbach, T. & Wieschaus, E. Maternal-effect mutations altering the anterior-posterior pattern of the Drosophila embryo. en. *Roux's Archives of Developmental Biology* **195,** 302–317 (July 1986).
13. Castrillon, D. H., Quade, B. J., Wang, T. Y., Quigley, C. & Crum, C. P. The human *VASA* gene is specifically expressed in the germ cell lineage. en. *Proceedings of the National Academy of Sciences* **97,** 9585–9590 (Aug. 2000).
14. Brangwynne, C. P. *et al.* Germline P Granules Are Liquid Droplets That Localize by Controlled Dissolution/Condensation. en. *Science* **324,** 1729–1732 (June 2009).
15. Thomas, L., Putnam, A. & Folkmann, A. Germ granules in development. en. *Development* **150,** dev201037 (Jan. 2023).

16. Tanaka, S. S. *et al.* The mouse homolog of *Drosophila Vasa* is required for the development of male germ cells. en. *Genes & Development* **14,** 841–853 (Apr. 2000).

17. Dehghani, M. & Lasko, P. *In vivo* mapping of the functional regions of the DEAD-box helicase Vasa. en. *Biology Open* **4,** 450–462 (Apr. 2015).

18. Yamazaki, H. *et al.* Bombyx Vasa sequesters transposon mRNAs in nuage via phase separation requiring RNA binding and self-association. en. *Nature Communications* **14,** 1942 (Apr. 2023).

19. Sabari, B. R., Hyman, A. A. & Hnisz, D. Functional specificity in biomolecular condensates revealed by genetic complementation. en. *Nature Reviews Genetics* (Oct. 2024).

20. Brady, J. P. *et al.* Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. en. *Proceedings of the National Academy of Sciences* **114,** E8194–E8203 (Sept. 2017).

21. Martin, E. W. *et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. en. *Science* **367,** 694–699 (Feb. 2020).

22. Urry, D. W. *et al.* Temperature of polypeptide inverse temperature transition depends on mean residue hydrophobicity. en. *Journal of the American Chemical Society* **113,** 4346–4348 (May 1991).

23. Quiroz, F. G. & Chilkoti, A. Sequence heuristics to encode phase behaviour in intrinsically disordered protein polymers. en. *Nature Materials* **14,** 1164–1171 (Nov. 2015).

24. Wang, J. *et al.* A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. en. *Cell* **174,** 688–699.e16 (July 2018).

25. De La Cruz, N. *et al.* Disorder-mediated interactions target proteins to specific condensates. en. *Molecular Cell* **84,** 3497–3512.e9 (Sept. 2024).

26. Pei, G. *et al. Sequence Composition Dictates Condensate Miscibility* en. Nov. 2024.

27. Bremer, A. *et al.* Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. en. *Nature Chemistry* **14,** 196–207 (Feb. 2022).

28. Rekhi, S. *et al.* Expanding the molecular language of protein liquid–liquid phase separation. en. *Nature Chemistry* **16,** 1113–1124 (July 2024).

29. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. en. *Nature Methods* **11,** 801–807 (Aug. 2014).

30. Olson, C. A., Wu, N. C. & Sun, R. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. en. *Current Biology* **24,** 2643–2651 (Nov. 2014).

31. Rogers, J. M., Passioura, T. & Suga, H. Nonproteinogenic deep mutational scanning of linear and cyclic peptides. en. *Proceedings of the National Academy of Sciences* **115,** 10959–10964 (Oct. 2018).

32. Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. en. *Nature* **620,** 434–444 (Aug. 2023).

33. Nemoto, N., Miyamoto-Sato, E., Husimi, Y. & Yanagawa, H. In vitro virus: Bonding of mRNA bearing puromycin at the 3-terminal end to the C-terminal end of its encoded protein on the ribosome in vitro. en. *FEBS Letters* **414,** 405–408 (Sept. 1997).

34. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. en. *Proceedings of the National Academy of Sciences* **94,** 12297–12302 (Nov. 1997).

35. Bolognesi, B. *et al.* The mutational landscape of a prion-like domain. en. *Nature Communications* **10,** 4162 (Dec. 2019).

36. Yamagishi, Y. *et al.* Natural Product-Like Macrocyclic N-Methyl-Peptide Inhibitors against a Ubiquitin Ligase Uncovered from a Ribosome-Expressed De Novo Library. *Chemistry & Biology* **18,** 1562–1570 (Dec. 2011).

37. DelRosso, N. *et al.* Large-scale mapping and mutagenesis of human transcriptional effector domains. en. *Nature* (Apr. 2023).

38. Ng, S. C. & Görlich, D. A simple thermodynamic description of phase separation of Nup98 FG domains. en. *Nature Communications* **13,** 6172 (Oct. 2022).

39. Glauninger, H. *et al. Transcriptome-wide mRNA condensation precedes stress granule formation and excludes stress-induced transcripts* en. Apr. 2024.

40. Paloni, M., Bailly, R., Ciandrini, L. & Barducci, A. Unraveling Molecular Interactions in Liquid–Liquid Phase Separation of Disordered Proteins by Atomistic Simulations. en. *The Journal of Physical Chemistry B* **124,** 9009–9016 (Oct. 2020).

41. Pak, C. W. *et al.* Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. en. *Molecular Cell* **63,** 72–85 (July 2016).

42. Vernon, R. M. *et al.* Pi-Pi contacts are an overlooked protein feature relevant to phase separation. en. *eLife* **7,** e31486 (Feb. 2018).

43. Stender, E. G. P. *et al.* Capillary flow experiments for thermodynamic and kinetic characterization of protein liquid-liquid phase separation. en. *Nature Communications* **12,** 7289 (Dec. 2021).

44. Hughes, M. P. *et al.* Atomic structures of low-complexity protein segments reveal kinked sheets that assemble networks. en. *Science* **359,** 698–701 (Feb. 2018).

45. Von Bülow, S., Tesei, G. & Lindorff-Larsen, K. *Prediction of phase separation propensities of disordered proteins from sequence* en. June 2024.

46. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. & Pappu, R. V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. en. *Biophysical Journal* **112,** 16–21 (Jan. 2017).

47. Norrild, R. K. *et al.* Increasing protein stability by inferring substitution effects from high-throughput experiments. en. *Cell Reports Methods,* 100333 (Nov. 2022).

48. Quaglia, F. *et al.* DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. en. *Nucleic Acids Research* **50,** D480–D487 (Jan. 2022).

49. Li, X. *et al.* -synuclein regulates the phase transitions and amyloid conversion of -synuclein. en. *Nature Communications* **15,** 8748 (Oct. 2024).

50. Hadarovich, A. *et al.* PICNIC accurately predicts condensate-forming proteins regardless of their structural disorder across organisms. en. *Nature Communications* **15,** 10668 (Dec. 2024).

51. Rostam, N. *et al.* CD-CODE: crowdsourcing condensate database and encyclopedia. en. *Nature Methods* **20,** 673–676 (May 2023).

52. Kim, T. H. *et al.* Interaction hot spots for phase separation revealed by NMR studies of a CAPRIN1 condensed phase. en. *Proceedings of the National Academy of Sciences* **118,** e2104897118 (June 2021).

53. Tesei, G. *et al.* Conformational ensembles of the human intrinsically disordered proteome. en. *Nature* **626,** 897–904 (Feb. 2024).

54. Tesei, G., Schulze, T. K., Crehuet, R. & Lindorff-Larsen, K. Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. en. *Proceedings of the National Academy of Sciences* **118,** e2111696118 (Nov. 2021).

55. Tesei, G. & Lindorff-Larsen, K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. en. *Open Research Europe* **2,** 94 (Jan. 2023).

56. Norrild, R. K. *et al.* Taylor Dispersion-Induced Phase Separation for the Efficient Characterisation of Protein Condensate Formation. en. *Angewandte Chemie International Edition* **63,** e202404018 (June 2024).

57. Ambadi Thody, S. *et al.* Small-molecule properties define partitioning into biomolecular condensates. en. *Nature Chemistry* (Sept. 2024).

58. Rekhi, S. & Mittal, J. *Amino Acid Transfer Free Energies Reveal Thermodynamic Driving Forces in Biomolecular Condensate Formation* en. Dec. 2024.

59. Franzmann, T. M. *et al.* Phase separation of a yeast prion protein promotes cellular fitness. en. *Science* **359,** eaao5654 (Jan. 2018).

60. Bloom, J. D. *et al.* Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences* **102,** 606–611 (2005).

61. Zarin, T. *et al.* Proteome-wide signatures of function in highly diverged intrinsically disordered regions. en. *eLife* **8,** e46883 (July 2019).

62. Wells, J. A. Additivity of mutational effects in proteins. en. *Biochemistry* **29,** 8509–8517 (Sept. 1990).

63. Faure, A. J. *et al.* The genetic architecture of protein stability. en. *Nature* (Sept. 2024).

64. Aumiller, W. M. & Keating, C. D. Phosphorylation-mediated RNA/peptide complex coacervation as a model for intracellular liquid organelles. en. *Nature Chemistry* **8,** 129–137 (Feb. 2016).

65. Pesce, F. *et al.* Design of intrinsically disordered protein variants with diverse structural properties. en. *Science Advances* (2024).

66. Gaspar, J. M. NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. en. *BMC Bioinformatics* **19,** 536 (Dec. 2018).

67. Klopfenstein, D. V. *et al.* GOATOOLS: A Python library for Gene Ontology analyses. en. *Scientific Reports* **8,** 10872 (July 2018).

68. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. en. *PLOS Computational Biology* **13** (ed Gentleman, R.) e1005659 (July 2017).