PÁZMÁNY PÉTER CATHOLIC UNIVERSITY

ROSKA TAMÁS DOCTORAL SCHOOL
OF SCIENCES AND TECHNOLOGY

fides et ratio

SZABÓ András László

# Analysis of phase-separating proteins within postsynaptic densities through a combination of computational and experimental approaches

PhD Dissertation

Thesis Supervisor:
GÁSPÁRI Zoltán, PhD
2025

# Table of contents

# 1. List of abbreviations

| Abbreviations | Definition |
|---|---|
| BSA | Bovine Serum Albumin |
| CDR | Charge-Dense Region |
| CRR | Charged Residue Repeat |
| DLS | Dynamic Light Scattering |
| D233 | Drebrin construct |
| EGFP | Enhanced Green Fluorescent Protein |
| FITC | green-fluorescent fluorescein-EX |
| GKAP | Guanylate Kinase-Associated Protein |
| GKAP-PBM | GKAP's PDZ-Binding Motif |
| GKAP-DLC2 | GKAP's two LC8(Dynein Light Chain)-binding motifs |
| GO | Gene Ontology |
| IDP | Intrinsically Disordered Protein |
| LC8 | dynein light chain LC8 protein |
| LLPS | Liquid-Liquid Phase Separation |
| MLO | Membraneless Organelle |
| PSD | Postsynaptic Density |
| PSD-95 | Postsynaptic Density protein 95 |
| RNABP | RNA-Binding Protein |
| ROC | Receiver Operating Characteristic analysis |
| SAH | Single $\alpha$-Helices |
| sCDR | signed Charge-Dense Region |
| SynGAP | Synaptic Ras GTPase-Activating Protein 1 |
| uCDR | unsigned Charge-Dense Region |

# 2. Introduction

## 2.1. Protein phase separation

There is a plethora of complex biochemical processes in living cells that must be efficiently conducted and finely regulated in space and time. Most processes employ membrane-bound organelles that provide an ideal environment for a specific mechanism, such as lysosomes providing an acidic environment for degradative enzymes. However, there are some cellular mechanisms that involve so called membraneless organelles (MLOs), which provide reversible and finely tuned compartmentalization of certain biochemical processes. The formation of MLOs, or biochemical condensates in the case of *in vivo* studies, is referred to as protein phase separation. This complex molecular phenomenon has been shown to have a critical role in cellular processes such as chromatin regulation, RNA transcription, and the organization of postsynaptic densities. [1-3] It can also yield different states of MLOs from liquids through gels to solids, the latter usually depriving the phenomenon of its reversibility, ultimately leading to aggregation. [4]

Protein phase separation is typically initialized by the interactions of multivalent proteins with multiple modular domains and disordered regions, often of low sequence complexity. [5] These proteins fall into the category of “scaffolds”, molecules that are essential to the structural integrity of MLOs. Additional components may participate in functionalities, but only under certain circumstances. These are referred to as “clients”. [6] As a rule of thumb, scaffold-scaffold interactions are more persistent than scaffold-client interactions, and the composition of MLOs changes according to several factors such as stress and the cell cycle. [1, 7] P bodies perfectly illustrate these general characteristics of phase separation, as they are scaffolded by a few critical RNA-binding proteins (RNABPs) but store a large variety of mRNAs and additional protein components as clients.

Due to the diverse attributes observed in various cases of phase-separating systems, multiple types of the phenomenon have been distinguished from each other. Maybe the most prominent category is liquid-liquid phase separation (LLPS), characterized by solutions transitioning into distinct phases where certain solutes are present in highly elevated concentrations, while the phases exhibit liquid-like properties. This specific type of the phenomenon has its own terminology, in which “scaffolds” are replaced by “drivers”, referring to sets of proteins that are able to drive LLPS on their own. Small molecules and ions are not considered drivers, even if they are required for the initiation of the process. In this context, clients are molecules that may partition into MLOs without participating in their formation. [8] During LLPS, liquid-like droplets may grow up to a few microns in diameter through additional molecules partitioning into them, multiple MLOs fusing with one another, or a so-called coarsening process in some cases. The latter way of droplet growth is also referred to as Ostwald ripening, and it is caused by larger droplets being inherently more stable, while smaller droplets being more prone to dissolution, resulting in a net movement of components towards larger condensates. This coarsening is time-dependent and while it is not necessarily accompanied by the formation of aggregates, there are systems with aggregation-prone proteins where it correlates with pathological conditions. Generally, however, the final stage of droplet maturation is an arrested, gel-like state. [9]

There are several attributes potentially present in a protein's sequence and structure that have been shown to increase propensity towards phase separation. Intrinsically disordered proteins (IDPs) are especially prone towards the phenomenon, due to their low-complexity, prion-like sub-sequences that often govern LLPS. Systems involving such components are prone to undergo material state transitions, such as the liquid-solid transition of the RNABP Fused in sarcoma (FUS) or the TAR DNA-binding protein 43 (TDP-43). Liquid-solid phase transitions are also called aggregation in specific cases where they are often associated with severe diseases such as amyotrophic lateral sclerosis (ALS). [10, 11] In case of the intracellular domain of Nephrin (NICD), its intrinsically disordered nature leads to the formation of dense liquid droplets called coacervates through associative interactions between multiple soluble (macro)molecules. Therefore, the phase separation of this particular protein has been described as complex coacervation where a polymer-dense and a polymer-depleted phase are in equilibrium with one another. [12] In contrast, simple coacervation would only require one type of polymer. NICD was observed to undergo LLPS *in vitro* when mixed with positively charged GFP. No specific motifs were proven responsible for the phenomenon, but multiple shuffle and deletion mutants were able to drive LLPS, as long as blocks of negatively charged residues were retained in a specific pattern, implying the robustness of phase separation against mutations. The presence of structural components rich in arginine and aspartic acid, also referred to as mixed-charged domains (MCDs), was observed to be important in nuclear speckle condensation, and the formation of cell-to-cell channels for certain fungi. [13]

In addition to the above attributes that are intrinsic to phase-separating proteins, there are external factors that regulate the phenomenon. Supersaturated proteins with their low solubility are sensitive to mild physiological fluctuations caused by the cell cycle. [14] Similarly, many proteins aggregate in response to stress conditions, such as glucose depletion. [15] It is well-known that protein concentration tends to act as a switch for phase separation, but in some cases that is regulated through changes in

RNA concentration, thus promoting condensate formation. [16] The solubility of a specific protein is lowest when the pH of its solution is at the protein's isoelectric point, which is in the mildly acidic range for most proteins. [17] Therefore, pH is another environmental factor that significantly affects protein behavior. Finally, temperature has a major role in the formation of high-order structures, especially in case of RNABPs that have been shown to be heat-sensitive. [18] It is important to note that there is some dissonance between *in vitro* and *in vivo* results as to which conditions are impactful. [19, 20]

## 2.2. Postsynaptic densities

Postsynaptic densities (PSDs) are multilayered cellular components largely situated on the internal surface of postsynaptic membranes, with receptor proteins, such as the N-methyl D-aspartate receptor (NMDAR), extending it into the synaptic cleft. [21] These disk-shaped compartments encompass a complex network of proteins and nucleic acids, including actins, RNABPs, and membrane-associated guanylate kinases (MAGUKs). Some of the most essential constituents of this network include the postsynaptic density protein 95 (PSD-95), the guanylate kinase-associated protein (GKAP), the synaptic Ras GTPase-activating protein 1 (SynGAP), and various Shank and Homer proteins. [21-28] PSDs are the primary cellular components that receive synaptic transmissions. Their structural changes exhibit a strong correlation with synaptic strength and plasticity, which are in turn essential mechanisms for higher biological functions such as memory and learning. [3]

Dynamic structural changes within PSDs are regulated through finely tuned biochemical processes, including LLPS. This is supported by a two-component *in vitro* model consisting of SynGAP and PSD-95, proteins abundantly present in PSDs, which have been observed to self-organize into highly condensed, PSD-like droplets. [29] The importance of phase separation in the organization of PSDs has been further reinforced by more complex *in vitro* models, one including GKAP, Shank3, and Homer3, in addition to SynGAP and PSD-95. [30] PSD-95 has also been shown to colocalize with various heterogeneous nuclear ribonucleoproteins (hnRNPs) at PSDs, a phenomenon amplified by synaptic activity. [21]

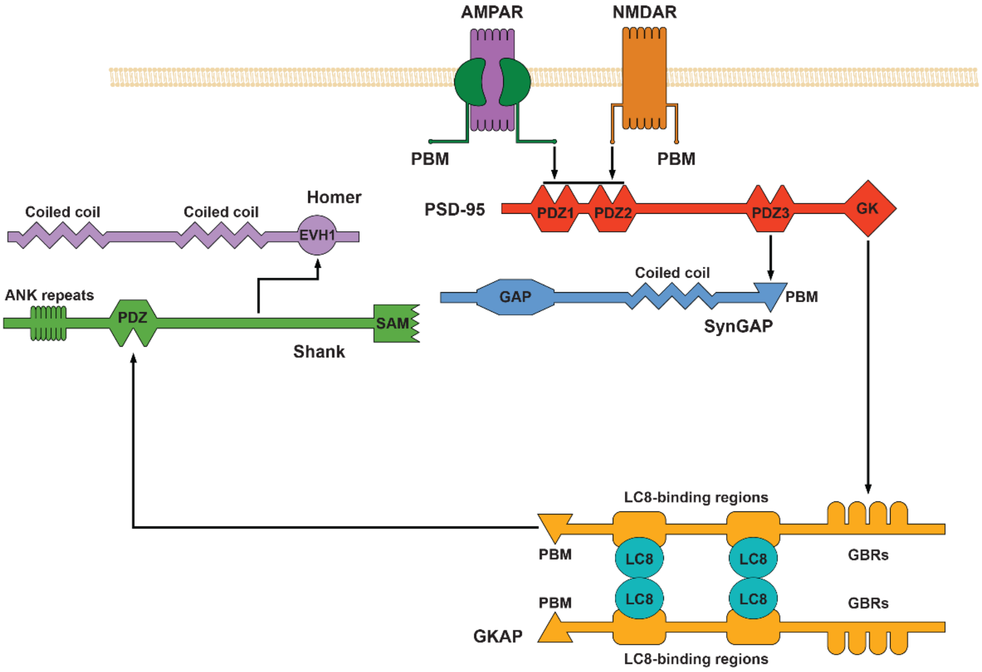


Fig. 1. An *in vitro* model that includes the primary components of PSDs. Proteins are depicted in arbitrary orientations for a simpler illustration of the selected interactions. [II, Fig. 3.]

The experimental analysis presented in this dissertation focused on a two-component system from the PSD (Fig. 1.), consisting of GKAP and one of its binding partners, the dynein light chain LC8 protein (referred to as LC8, Fig. 2.).

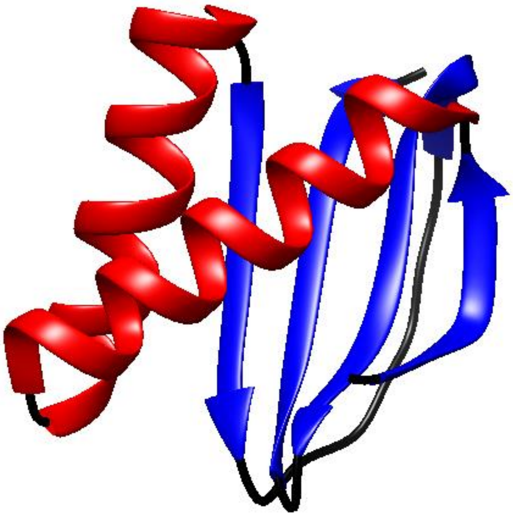


Fig. 2. Structure of a monomeric LC8, measured via X-ray crystallography (RCSB PDB: 7CNU). All helical structures are highlighted in red.

GKAP is a well-documented scaffolding protein with a high ratio of disordered regions and multiple binding sites, two of which are LC8-binding motifs. GKAP also participates in the regulation of NMDA receptors. [31] LC8 on the other hand is known for its multivalent interactions with IDPs and its affinity towards forming dimers that can bind two additional ligands. [32] The combination of LC8's multivalent interactions, the involvement of IDPs, and the amount of disorder within GKAP suggest that these two proteins are likely capable of driving LLPS cooperatively. They are certainly capable of forming hetero-oligomeric complexes (Fig. 3.), the exact stoichiometry of which has only been explored recently. [33, 34]

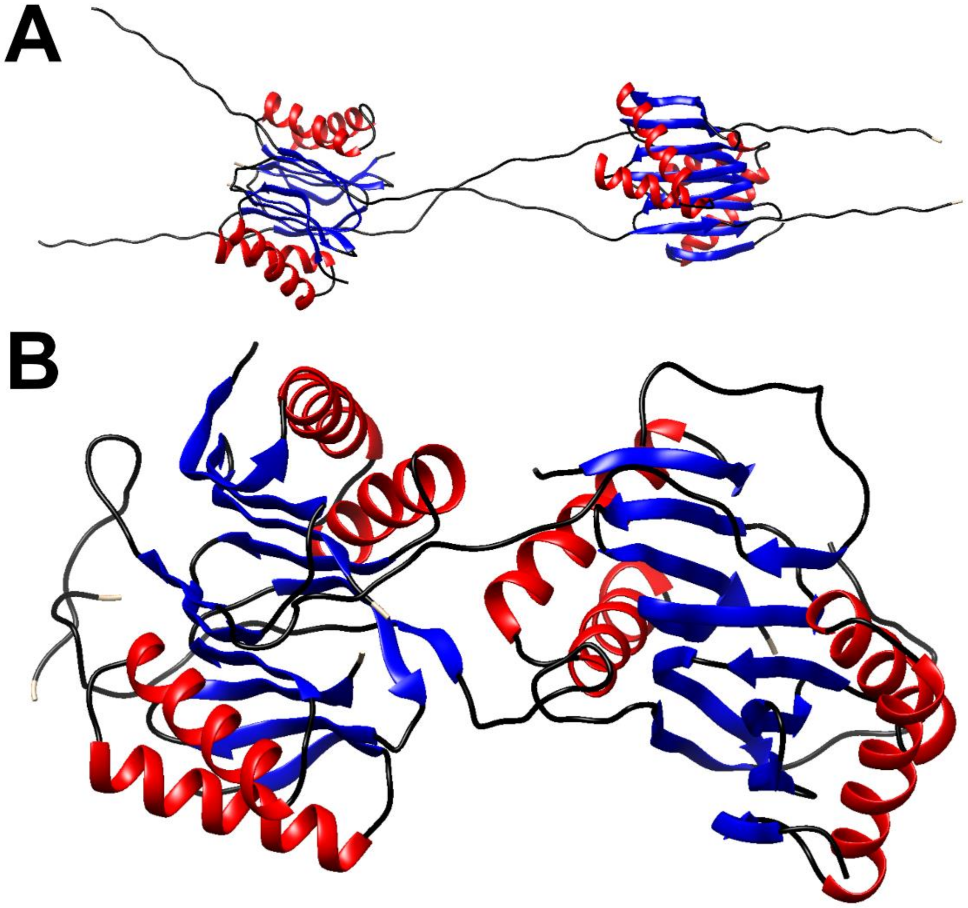


Fig. 3. Simulated structures of the hexameric complex, with the GKAP-DLC2 chains unfolded for a closer resemblance to the schematic arrangement displayed in Fig. 1. (A), and with the folded structure (B). These structures were provided by Zsófia E. Dobson-Kálmán. All helical structures are highlighted in red.

Apart from these two partners, the protein Drebrin has also been touched upon in this study (Fig. 4.), though not as part of the above system, as it does not exhibit direct interactions with GKAP or LC8, but Homer, specifically its EVH1 domain. [35]

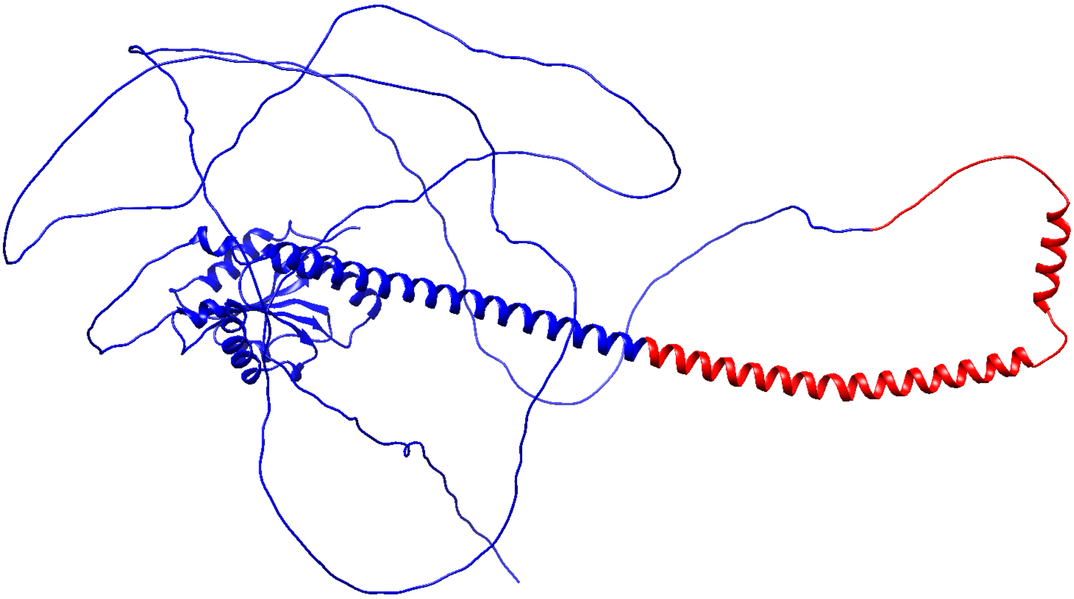


Fig. 4. Predicted structure (AF-Q16643-F1) of the human Drebrin, with its investigated segment (residues 233-317) highlighted in red. A large portion of this AlphaFold structure was predicted to be SAH. However, other models align with this prediction for the first dozen residues only, with the rest of the region identified as disordered.

## 2.3. Experimental approaches to characterize LLPS

A reasonable first approach for investigating PSDs, LLPS and nano- to microscale droplets is to examine protein samples with a microscope. Fluorescent microscopy in particular has the advantage of easy differentiation between MLOs of labelled proteins and other objects and artefacts, enabling *in vitro* as well as *in vivo* studies. It additionally offers non-invasiveness, and even reversibility *in vitro*. [36]

Another popular approach for *in vitro* LLPS studies is dynamic light scattering (DLS), a fast and non-invasive method that can monitor the development of MLOs from the moment of initialization to the typical conclusion of aggregation. It does so by generating detailed size distribution diagrams that shift from protein complexes under 10 nm to particles 10-100 nm in diameter and, depending on the system, even micro-sized objects. [37]

An experimental method that had not been used to investigate PSDs or LLPS employs microfluidic devices in combination with fluorescent microscopy to measure the diffusion coefficients of particles during laminar flow, which can be converted into hydrodynamic radii. This method had been previously utilized to measure the size of green fluorescent polystyrene particles, α-synuclein molecules, antibody fragments, and other nanoparticles, and then compare the results with *a priori* data yielded by DLS.

[38] This technique was further tested on polydisperse mixtures with up to three components of distinct sizes: α-synuclein fibrils, small unilamellar vesicles (SUVs), and SUVs with α-synuclein fibrils bound to their external surface. [39]

## 2.4. Single α-helices

Single α-helices (SAHs) are protein segments that exhibit a characteristic repetitive pattern of oppositely charged residues, which leads to the formation of rigid helical structures that remain stable even in isolation (Fig. 5.). [40, 41] This stability is at least partially owed to intrahelical salt bridges. [42] SAHs are rich in arginine, glutamate, and lysine, and vary from a few dozen up to about 200 residues in length, showing similar characteristics to coiled coils. Approximately 4% of human proteins that were previously predicted to contain coiled coils actually have SAHs instead. [43] Although they are unlikely to directly contribute to any interactions with RNA molecules, SAH domains have been found to be prevalent in RNABPs. [40]

Fig. 5. NMR structure of the SAH domain of myosin-6 (RCSB PDB: 6OBI).

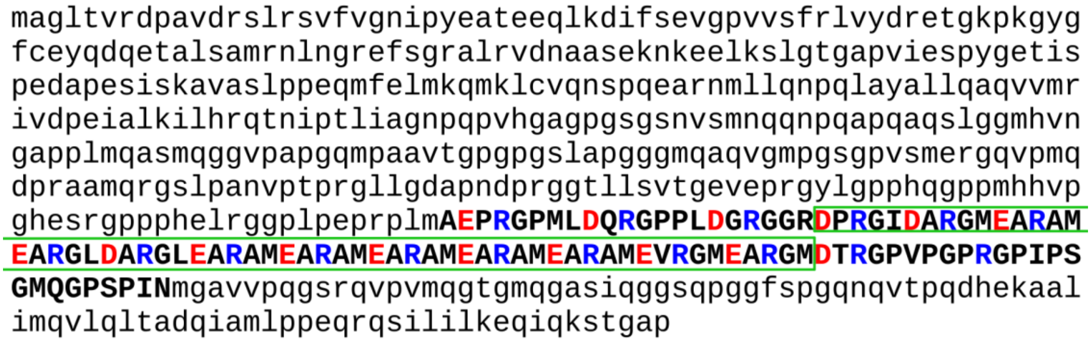
## 2.5. Computational approaches to identify SAHs

An immense number of phase-separating proteins has been identified in the past few decades, and thus it has become a relevant issue to catalogue them. There are various databases that accumulate information about phase separation.

- PhaSePro is a manually curated database that specifically details protein sequences with experimental evidence of driving LLPS. [44]
- DrLLPS contains hundreds of thousands of proteins that are computationally associated with the phenomenon. [45]
- PhaSepDB is a manually curated collection of proteins related to protein phase separation or MLOs. [46]
- CD-CODE relies on contributors in order to gather information about biomolecular condensates. [47]
- LLPSDB contains LLPS-related proteins with experimental evidence, while also including some information about the conditions associated with the phenomenon. [48]
- MSGP is a highly specialized database that only collects information about stress granule proteins. [49]

In comparison, there are only a few reliable assemblies about PSD proteins, such as PSINDB, which specifically collects curated information about postsynaptic protein-protein interactions. [50] Multiple computational approaches have been explored to identify proteins that may undergo phase separation, specifically LLPS. [8] Among these approaches there are those that generate a score based on primary sequences, such as PLAAC and FuzDrop. [51, 52] There are predictive models that use machine learning, such as PSAP, PSPredictor, and PICNIC. [53-55] And there are molecular modelling approaches that usually rely on coarse-grained simulations. [56]

However, one of the most important computational assets to this study is the FT_CHARGE algorithm that can identify SAHs. [40] It calculates the charge correlation function of a protein's sequence and then converts that into its Fourier transform. There, repetitive charge patterns within the sequence are revealed as peaks with an amplitude significantly higher than expected from a random sequence with similar content of positively and negatively charged residues (Fig. 6.). Thus, this approach can be classified as a "pattern recognition strategy". [57] The minimum length of the revealed patterns, referred to as charged residue repeats (CRRs), is defined by the window size (16, 32 or 64) used by FT_CHARGE. The maximum length is not limited as consecutive windows identified as CRRs can be combined. The algorithm covers a discrete spectrum from 1/64 to 1/2, and SAHs have a characteristic frequency of 1/9 to 1/6. Therefore, they can be considered a subset of CRRs.



```
magltvrdpavdrslrsvfvgnipyeateeqlkdifsevgpvvsfrlvydretgkpkgyg
fceyqdqetalsamrnlngrefsgralrvdnaaseknkeelkslgtgapviespygetis
pedapesiskavaslppeqmfelmkqmklcvqnspqearnmllqnpqlayallqaqvvmr
ivdpeialkilhrqtniptliagnpqpvhgagpgsgsnvsmnqqnpqapqaqslggmhvn
gapplmqasmqggvpapgqmpaavtgpgpgslapgggmqaqvgmpgsgpvsmergqvpmq
dpraamqrgslpanvptprgllgdapndprggtllsvtgeveprgylgpphqgppmhhvp
ghesrgppphelrggplpeprplmAEPRGPMLDQRGPPLDGRGGRDPRGIDARGMEARAM
EARGLDARGLEARAMEARAMEARAMEARAMEARAMEVRGMEARGMDTRGPVPGPRGPIPS
GMQGPSPINmgavvpqgsrqvpvmqgtgmqgasiqggsqpggfspgqnqvtpqdhekaal
imqvlqltadqiamlppeqrqsililkeqiqkstgap
```

Fig. 6. The cleavage stimulation factor subunit 2 (CSTF2, UniProtKB ID: P33240) contains 12 X 5 AA tandem repeats on a 60 residue-long segment (green) that was identified to be part of a 105 residue long CRR by FT_CHARGE. The general sequence of the repeats is: [D/E]XRXX where X can be any residue that is not charged (the first residue is either an aspartic or glutamic acid). Positively charged residues are highlighted in blue, while negatively charged residues are highlighted in red. The sequence does not qualify as a SAH due to its 13/64 (~1.2/6) Fourier frequency, which is slightly higher than the upper limit for SAHs (1/6). [I, Fig. 2.]

# 3. Aims of the Study

The Introduction section showed that SAHs are abundant in RNA-binding proteins that are both present in PSDs and have an above average propensity towards phase separation. It has also been shown that certain protein motifs with repeating blocks of charged residues drive LLPS. Furthermore, there are two proteins, GKAP and LC8, that have been shown to exhibit multivalent interactions with each other and additional PSD components, including IDPs that are known to be prone towards phase separation. While neither of these proteins is RNA-binding, GKAP does interact with PSD-95, which colocalizes with various hnRNPs. Taking all of this into consideration, it is possible that the complexes of GKAP and LC8, their indirect association with RNABPs, and the presence of SAHs and other charged sequence motifs all contribute to the phase separation phenomena that partially regulate the dynamic structural changes of PSDs.

The main goal of the study was to evaluate the role of different sequence motifs and interactions of multivalent proteins regarding phase separation, in the context of PSDs. The large-scale investigation of certain sequence motifs and their potential associations to protein phase separation stipulated two specific objectives:

- Assessment of readily available resources for the identification of certain types of charged sequence motifs, with the possibility of having to improve upon existing approaches or developing entirely new methods.
- Confirmation – or refutation – of associations between identified motifs and their host sequences' propensity towards phase separation, with special care taken to assess the robustness of the results.

Having access to *in silico* resources for the identification of SAHs and other CRRs further rationalized that the large-scale investigation of possible association between them and protein phase separation should be a computational study where a dataset of human protein sequences must be compiled and expanded with specific types of charged sequence motifs and regions with experimental evidence of contributing to phase separation. This dataset can then be used to explore associations between the presence of the investigated motifs and the protein's propensity towards phase separation. The robustness of the results could be improved via minimizing redundancy within the dataset.

Meanwhile, examining the role of specific PSD proteins regarding phase separation requires a reliable method that can identify the formation of MLOs and smaller complexes. Experimental researchers at the faculty have already been working on the expression and purification of various PSD proteins, including GKAP, LC8, and

Drebrin. Additionally, faculty technicians specialized in fluorescent microscopy and microfluidics were open to collaboration, making it feasible to adapt a diffusion-based method capable of determining the size of solute particles. Therefore, the following aims were identified regarding the investigation of these PSD proteins:

- Development of an *in vitro* approach capable of determining the size of PSD proteins and their complexes. This approach would combine microfluidics with fluorescent microscopy techniques to monitor the diffusion of solute particles.
- Evaluation of the designed approach regarding the accuracy and precision of approximated particles sizes, comparing it with other methods.

Designing such a method would be an iterative process where an initial experimental setup and microfluidic device would be evaluated and improved upon multiple times. It would also involve the preparation of fluorescent samples for both calibration and analysis. Finally, the approach requires an analytic software that can evaluate the measured data to complete the feedback loop of the iterative development process.

# 4. Methods

## 4.1. Computational analysis of charged sequence motifs

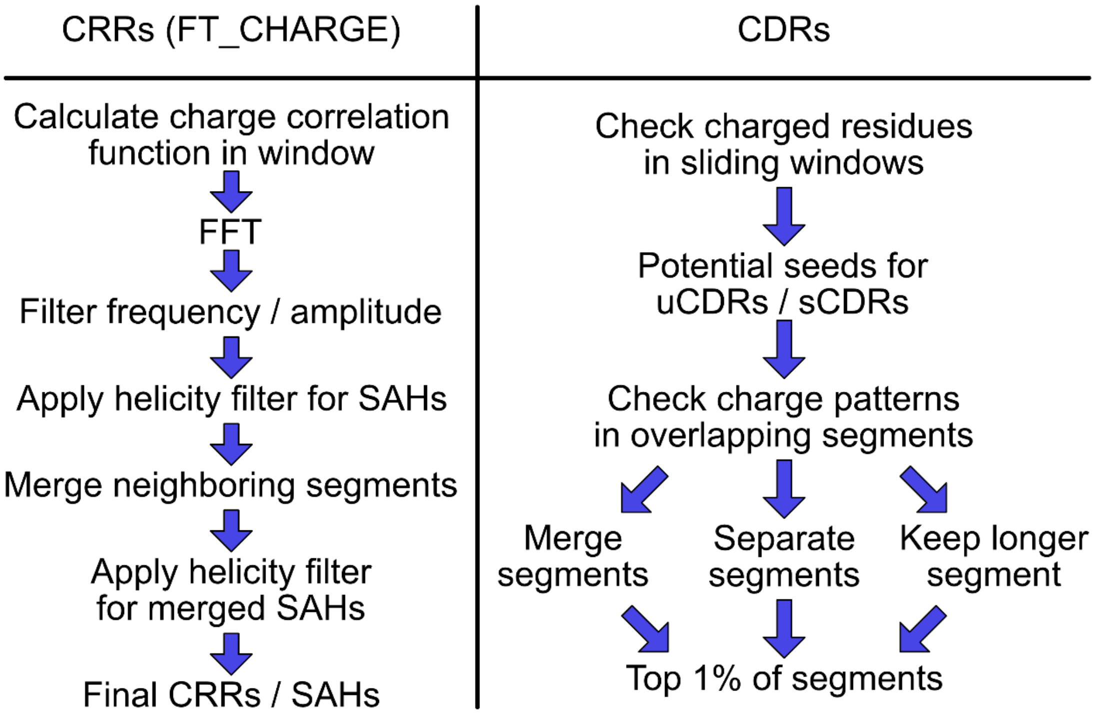
### 4.1.1. Compiling data on human protein sequences and PhaSepDB regions

The human reference proteome used for this study encompassed 20659 human genes with one isoform per gene, which were gathered from UniProtKB, specifically from its subset of manually curated entries called SwissProt. As of writing this dissertation, the number of entries changed to 20417, therefore it is important to mention here that the data for this research was collected in 2020-2021, and since then other resources may have changed as well, such as the CD-HIT webserver shutting down in 2022. Some investigations involved other datasets, such as another UniProtKB query of 41818 sequences annotated as human transmembrane proteins (the exact query was annotation:(type:transmem) AND organism:"Homo sapiens (Human) [9606]"). This was necessary for the exclusion of transmembrane proteins for some of the analyses that would have been distorted by the inclusion of immobile entries. Additional small sequence sets were also generated to match the length distribution of the entries within PhaSepDB. [46] This was achieved by selecting one to ten sequences from the reference proteome for each human PhaSepDB entry, where they had to match in length with ±5% relative error.

Each protein entry of PhaSepDB2.0 falls under one of three categories. Proteins with articles published after January 1st of 2000 constitute the “Reviewed” category, which is the most reliable one. Sequences supported by articles from before the 21st century may be included in the “UniProt reviewed” category, but only if the results have been confirmed by more recent studies. Data yielded by high-throughput methods such as organelle purification, proximity labelling, immunofluorescence image-based screen or affinity purification are assigned to the “High-throughput” category. Complementing the dataset with motifs that had experimental evidence of contributing to phase separation was simply done by integrating the regions of human PhaSepDB entries from the “Reviewed” category.

### 4.1.2. Expanding the dataset with charged sequence motifs

CRRs were identified with FT_CHARGE that detects regularly alternating positively and negatively charged residues based on the Fourier transform of the sequence’s charge correlation function (Fig. 7.). SAHs were highlighted among these motifs by their characteristic frequency of 1/9 to 1/6. [58] The reference proteome was surveyed multiple times with FT_CHARGE using window sizes of 16, 32, and 64. Therefore, the minimum frequency was 1/64, which corresponds to long, repeated segments of identically charged amino acids, such as a polylysine run. In contrast the maximum frequency of 1/2 corresponds to a region entirely composed of residues with alternating charges (e.g., the sequence “KEKEKEKEKE”). [59, 60]



CRRs (FT_CHARGE)
CDRs
Calculate charge correlation function in window
FFT
Filter frequency / amplitude
Apply helicity filter for SAHs
Merge neighboring segments
Apply helicity filter for merged SAHs
Final CRRs / SAHs
Check charged residues in sliding windows
Potential seeds for uCDRs / sCDRs
Check charge patterns in overlapping segments
Merge segments
Separate segments
Keep longer segment
Top 1% of segments

Fig. 7. Flow chart illustrating the identification process for SAHs and other CRRs via FT_CHARGE, and the algorithm developed for CDR detection.

The dataset included additional sub-sequences that contained either a high ratio of charged residues or a high net charge without featuring any specific repeating patterns of charged residues. Two separate survey strategies were developed for such regions, referred to as charge-dense regions (CDRs). One considered a sub-sequence highly charged if its ratio of charged residues had reached a given threshold. The other identified regions the overall charge of which had significantly differed from neutral (zero). Because of this difference between the two approaches, their respective yields of CDRs were denoted as either “signed” or “unsigned”, from now on referred to as sCDRs and uCDRs. Both approaches utilized windowing functions with a window size of either 16, 32, 64, or 128 as well as simple scoring scheme. A specific window of residues was considered to be a CDR if the absolute value of its sum score divided by its length had reached a pre-set cut-off value. For sCDRs, the scoring scheme assigned a score of +1 to arginines, histidines, and lysines, while assigning -1 to aspartic and glutamic acids. The scoring scheme for uCDRs assigned 1 to each of these five residues. Cut-off values were determined respectively for both approaches and individual window sizes. To this end, the algorithms identifying CDRs were applied, with a threshold of zero, to a version of the reference proteome where each sequence has been individually randomized. The results showed the expected scores for random sequences that were identical in residue composition and length distribution to the reference proteome. Two cut-off values were determined for each approach and window size, one that yields sequences within the top 5% of score, and another that results in entries within the top 1%. Because of the discrete nature of the scoring schemes (e.g. scoring a window of 16 residues from 0/16 to 16/16), it was impossible to select thresholds that yielded precise percentages. Therefore, the 1% cut-off values were uniformly lax in the sense that they identified slightly over 1% of all randomized sequences as CDRs, while the 5% cut-off values were all stricter, having a yield slightly under 5%. With these thresholds the wild-type reference proteome was surveyed to identify signed and unsigned CDRs. Identified segments of the same type and window size were merged, which was not a trivial process because the resulting merged regions still had to score above their corresponding threshold. The algorithm written for this purpose made pairwise comparisons between overlapping CDRs of the same variety and merged them if the resulting region still scored above the threshold. Otherwise, one of two operations was carried out. Either the regions could be separated by removing overlapping residues from one or both. If that was not possible without reducing the score of one or both below threshold, then one of them was potentially extended with some of the residues form the other, as long as its score would remain above threshold. The other region was then removed from the dataset. The algorithm was set to maximize the resulting region or regions if two overlapping ones were separated (Fig. 7.). It also iterated through the same sequence multiple times in case more than two windows had overlapping segments. The final step was to merge all sCDRs and all uCDRs, respectively, using the lowest related threshold in each case. It is important to note that many CRRs qualify as at least one type of charge-dense region, but not all of them fall under any of those categories.

### 4.1.3. Minimizing the redundancy of the sequence set

In order to reduce redundancy from the reference proteome, it was clustered with CD-HIT based on sequence similarity. Three degrees of clustering were carried out with thresholds of 0.9, 0.7, and 0.5, which resulted in clustered proteomes with 90%, 70%, and 50% sequence identity, respectively. All of these were done with global sequence identity, a 20-residue bandwidth of alignment, a minimal sequence length of 10 residues, and default alignment coverage parameters. All sequences were assigned to the best cluster that met the cut-off value, and all redundancy-filtered arrays were further processed with MATLAB scripts to make them compatible with the analytic tools down the line.

### 4.1.4. Exploring associations between sequence motifs and phase separation

One of the most straightforward ways of assessing the correlation between variables is Fisher' exact test of independence. In this case there were two variables, the presence of charged sequence motifs within proteins and the propensity of those proteins towards phase separation. Consequently, the reference proteome as well as its three redundancy-filtered variants were categorized into 2x2 contingency tables upon which Fisher's test was applied as implemented in R (www.r-project.org), determining the P-value (Tables 4-5.). In the case of clusters, two approaches were used for categorization. One of those was cluster-wise, meaning that association with protein phase separation was considered positive if any sequence within the given cluster was annotated that way in PhaSepDB. The other approach only considered a cluster to be related to the phenomenon if its representative sequence, determined by CD-HIT, was annotated as such. Therefore, all sequences were categorized in terms of their association to phase separation based on their presence in PhaSepDB2.0, which required experimental evidence of a protein's participation in phase transition before admission. The presence of charged sequence motifs was established similarly. PhaSePro was utilized in the case studies, since some of its annotations described blocks of charged residues as LLPS drivers.

The presence of charged sequence motifs was also evaluated as a possible indicator for a protein's likelihood of participating in phase separation. Receiver operating characteristic (ROC) tests are purpose built for the evaluation of an attribute as a viable indicator for classification. To this end, the reference proteome was also sorted according to the scores obtained from the CRR and CDR detections, after which true positive, true negative, false positive, and false negative rates were calculated based on the association to phase separation.

## 4.2. *In vitro* examination of PSD proteins and complex formation

### 4.2.1. Design of the experimental setup

Solute, unrestricted particles are in perpetual motion via diffusion, the rate of which is determined by their size. Smaller particles have a higher diffusion coefficient, meaning they move faster, as described by the Stokes-Einstein equation:

Eq. 1. $$D = \frac{k_B T}{6\pi\eta r}$$

where D is the diffusion coefficient of the particle, $k_B$ is the Boltzmann constant, T is the absolute temperature, η is the dynamic viscosity of the medium (in this case protein solution), and r is the hydrodynamic radius of the particle that is assumedly spherical. It is important to note that this equation only applies at low Reynolds numbers, which is characteristic of laminar flow (Re < 2000, see section 5.2.1.). Particles must also have a well-defined initial state from which they can freely diffuse along dimensions that can be monitored sufficiently to measure the rate of their motion.

Microfluidic focusers are devices designed to compress analytes into pre-determined sections of a channel. Using such a device with three inlets would allow focusing the analyte into the middle with two buffer streams from the sides (Fig. 9.). And extending the focuser with a long straight channel the width of which is significantly higher than its height would allow the analyte unrestricted diffusion along a horizontal plane that can be monitored through a microscope. Another assumption is that the diffusing particles would approximate normal distribution. Consequently, the distribution of particles along a vector that is perpendicular to the flow could be described with 1D Brownian motion:

Eq. 2. $$\rho(x,t) = \frac{N}{\sqrt{4\pi Dt}}\, e^{-\frac{x^2}{4Dt}}$$

where t is the time particle spent diffusing without restrictions, N is the number of particles that start from their original position, x is the distance from that position, and ρ is the particle’s density at distance x and time t. Similarly, the mathematical description of Gaussian functions is as follows:

Eq. 3. $$f(x) = ae^{-(\frac{x-b}{c})^2}$$

where a is the amplitude, b is the center, c is the standard deviation (STD), and f(x) is the value of the function at position x. Assuming the time component in Eq. 2. is constant and the Gaussian function is centered on the origin (b = 0):

Eq. 4. $$\frac{N}{\sqrt{4\pi Dt}}\, e^{-\frac{x^2}{4Dt}} = ae^{-\frac{x^2}{c^2}}$$

Eq. 5. $$c^2 = 4Dt$$

Eq. 6. $$a = \frac{N}{\sqrt{4\pi Dt}}$$

Therefore, the diffusion coefficient D and the number of particles N are:

Eq. 7. $$D = \frac{c^2}{4t}$$

Eq. 8. $$N = ac\sqrt{\pi}$$

And so, the diffusion coefficient of flowing particles, and thus their size, can be approximated as the incline of the linear function determined by the Gaussian functions' variances and the corresponding time components (Fig. 8.).

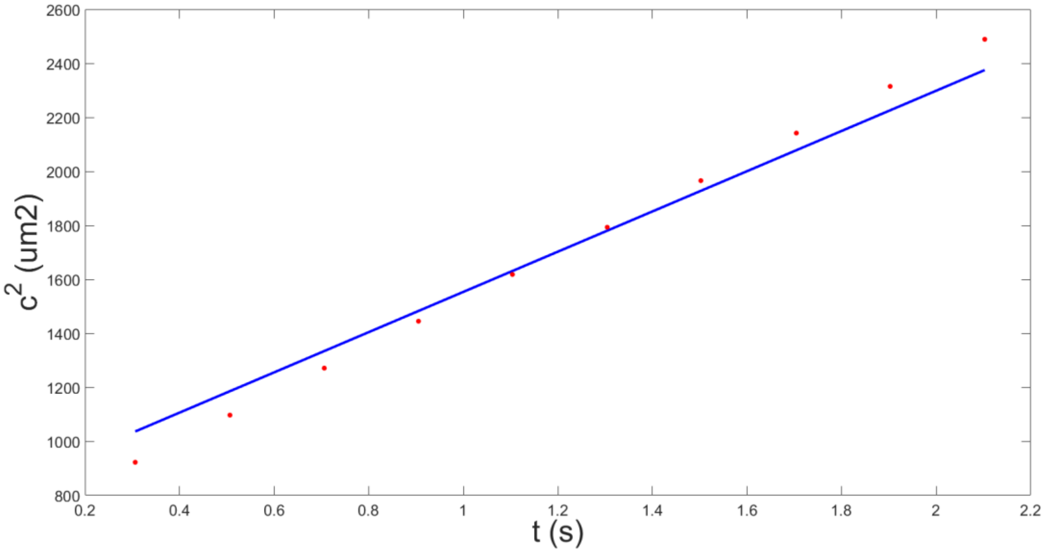


Fig. 8. The variance of Gaussian functions ($c^2$) fitted to fluorescent intensity profiles measured at different points along the microfluidic device. The x-axis shows the time (t) particles take to reach different measurement points at a given flow rate (see Measurement protocol). The resulting datapoints are shown in red, while their approximated incline is shown in blue. This data was measured while calibrating the experimental setup, for which Enhanced Green Fluorescent Protein (EGFP) was used as analyte. [II, Fig. 1.]

Measuring the time-dependent changes in the variance of Gaussian functions requires multiple measurement points set up along the channel at specific intervals.

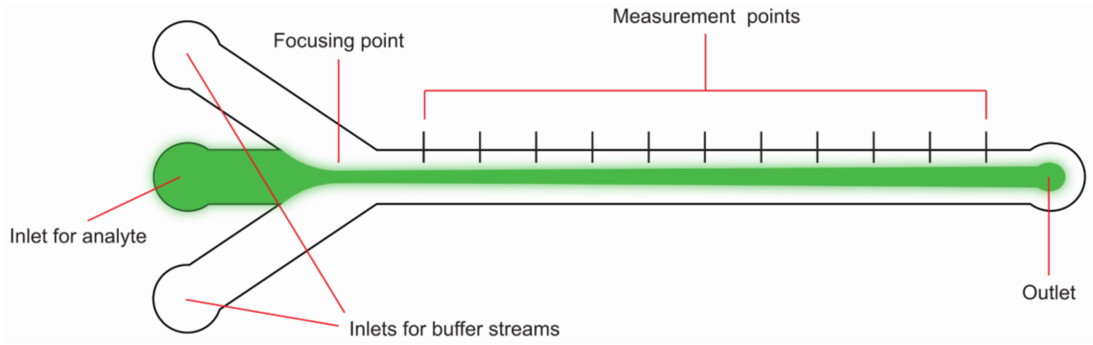


Fig. 9. Conceptual illustration of a three-inlet microfluidic focuser combined with a straight channel with multiple pre-determined measurement points. The fluorescent analyte is highlighted in green. [II, Fig. 2.]

The following protocol is the end product of an iterative process that involved testing different microfluidic devices that all captured the same basic principles detailed above, the reconfiguration of a Nikon Ti-2 E inverted microscope, and the adjustment of the analytic software to yield the best results. First, each inlet on the microfluidic device is connected to its separate syringe pump (Fig. 10.) and the filter turret of the microscope is set to record with a FITC filter (Excitation: 480/30, Dichroic mirror: 505, Barrier filter: 515), an Andor Zyla 4.2 camera, and a Nikon CFI Plan Apochromat Lambda D 20X lens. It is impossible to record the entire 300 µm width of the channel at higher magnifications. The microfluidic device is then taped to the motorized stage of the microscope to minimize the risk of any movement while the stage repositions it to different measurement points.

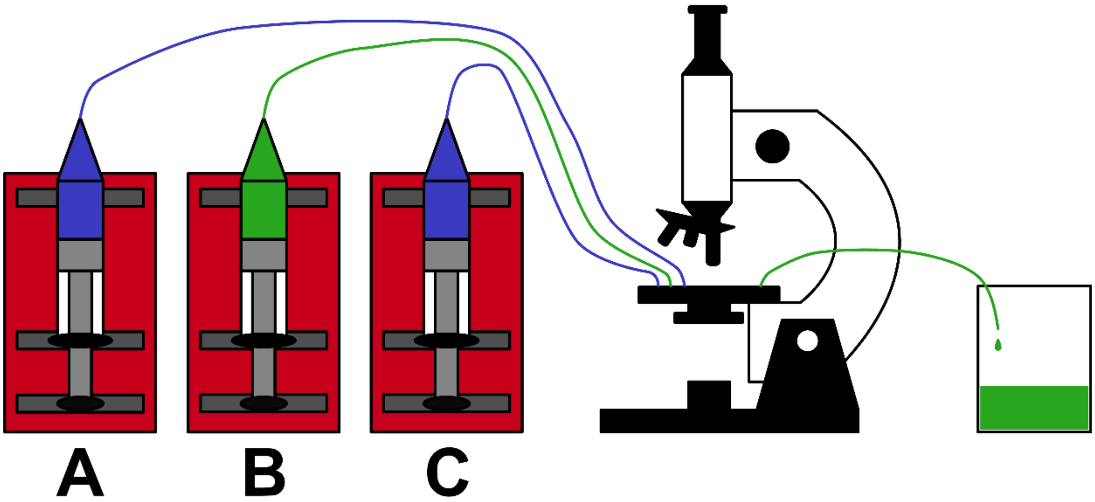


Fig. 10. Illustration of the experimental setup that involves syringe pumps (red), syringes filled with analyte (green) and buffer solutions (blue), a microscope, and a collection cup.

The internal surface of the device is treated with BSA in order to slow down the accumulation of fluorescent particles on it. To this end, each syringe contains 600 µl of 1% BSA solution that is pumped into the device at a flow rate of 20 µl/min. After the solutions have reached the device at all inlets the flow is maintained at the same rate for 15 minutes to properly degas the entire system, followed by an additional 15 minutes at 4 µl/min. This is necessary because the syringes connected to the side inlets will be replaced, which has a lower chance of reintroducing gas into the system when the solution flows slower, and the actual flow rate exhibits a hyperbolic decline after lowering its value on the syringe pumps. This provides ample time to manually focus on each measurement point, saving these vertical positions along with their corresponding horizontal positions. This will allow the automated recording of all measurement points. The brightfield images showing the device boundaries can be taken at this time, certifying that the microscope will record sharp images at the set measurement points.

After 15 minutes the syringes connected to the side inlets are replaced with ones containing 400 µl of PBS buffer solution, and the flow rates of the corresponding pumps are increased to 14 µl/min for 15 minutes, followed by another 15 minutes maintained at 4 µl/min to make sure that the system remains degassed. After that the syringe connected to the middle inlet is replaced by one containing 300 µl of the analyte, this time making sure that the syringe head is completely full of solution at the time of screwing it onto the syringe, since it is no longer viable to degas the system again due to the limited amount of analyte, and its tendency to accumulate on the internal surface of the device over time. Concurrently, all light sources are minimized around the microscope. All three syringe pumps are operated at 4 µl/min until the analyte reaches the device, which is monitored at an exposition rate of 300 ms and 4x gain.

Immediately after the analyte reached the device the flow rate is reduced to 1 µl/min and fluorescent images are recorded at 1x gain 5-15 minutes later with an exposition rate of 1-5 s, depending on signal strength. The temperature around the stage is also recorded at this time. Since fluorescent and brightfield images are recorded precisely at the same positions they will be perfectly aligned when combined with the microscope’s software. However, it is not feasible to place the microfluidic device onto the stage in such a fixed position while also making the channel within appear perfectly horizontal in the recording. Therefore, the combined images usually must be rotated by ± 2°, so that the angle between the flow direction and the vertical vector along which intensity profiles are recorded is 90°. These intensity profiles contain pairs of fluorescent and brightfield datapoints, and they are recorded where the signal-to-noise ratio is the highest, noting their distance from the marker of the given measurement point. The measured profiles are then exported into an Excel file, the first tab of which contains general information about the experiment, including temperature, viscosity, exposition time, analyte and buffer flow rates, the precise position of recorded profiles in relation to the focusing point of the device, and measurement points that should be omitted due to some kind of artefact, such as the laminar flow being disturbed by contamination. Each additional tab contains the profiles at one specific measurement point, organized into three columns: the position along the vertical vector, the intensity of the brightfield image at that position, and the intensity of the fluorescent image at the same position.

### 4.2.2. Fabrication of microfluidic devices

All microfluidic devices were produced by Mária Laki at the microfluidics lab of the faculty. The molds for the devices were fabricated using soft lithography and polydimethylsiloxane (PDMS) replica molding techniques. A negative photoresist height of 20 µm was applied to the top of a silicon wafer by spin-coating. The designed layouts were applied on the surface by laser writing.  The development of the mold was followed by the PDMS base. The curing agent was mixed in a 10:1 ratio, degassed, then poured over the mold and cured at 70°C for 90 min. Following the polymerization process, the PDMS was removed from the mold surface, and the inlets and outlets were processed. Finally, the PDMS slice with the microfluidic channel was bonded to a glass slide via plasma treatment. During measurements PTFE tubing was inserted into each inlet and the outlet, connecting 2 ml syringes and NE-1002X syringe pumps through 27 Gauge needle tips. The PTFE tube on the outlet was 15 cm long, leading into a beaker on the stage of the microscope. The PTFE tubes on the inlets were 25 cm long, which is the minimum length required to keep the tubing from becoming taut as the stage moves the device around.

### 4.2.3. Preparation of fluorescent samples

Protocols for the expression, purification, and fluorescent labelling of GKAP and LC8 protein samples were developed and carried out by Eszter Nagy-Kanta at the proteomics lab of the faculty. There were two different constructs for GKAP, one including both LC8-binding motifs in GKAP with extended flanking regions (10 residues on the N-terminus, 14 residues on the C-terminus), and one limited to the PDZ-binding motif on the sequence’s C-terminal (referred to as GKAP-PBM). For the former (referred to as GKAP-DLC2), the segment 655-711 in the *Rattus norvegicus* GKAP isoform 3 was selected. GKAP-PBM incorporated the PDZ-binding motif EAQTRL and

37 additional residues, characterizing the flanking region of the motif. In both cases, the insert was cloned to an altered pEV vector that contained an N-terminal His-tag and a tobacco etch virus (TEV) protease cleavage site. Both constructs contained four additional residues (GSHM) at the N-terminus that were remnants of the expression tag. The *Rattus norvegicus* DYNLL2 gene contained in the pEV plasmid vector was identical to the human ortholog. The pEV plasmid vector also contained a His-tag, the TEV protease cleavage site and the four residues at the N-terminus.

All three protein constructs were produced in BL21 (DE3) *E. coli* cells, transformed with the vectors, grown in LB media, and induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) at 6 MFU cell density. The recombinant proteins were expressed at 20°C overnight. Cell pellets were lysed by ultrasonic homogenization in 10% cell suspension. The lysis buffer contained 50 mM NaPi and 300 mM NaCl, set to pH 7.4. Denaturing-renaturing IMAC purification was applied to GKAP-DLC2, with 6 M GdnHCl and 50 mM NaPi added to 5 ml Ni-affinity column for denaturing, and native buffer with 50 mM NaPi, 20 mM NaCl, set to pH 7.4 for renaturing. This was followed up by washing and then elution that was performed with 250 mM imidazole. Afterwards, His-tags were removed with TEV protease. LC8 and GKAP-PBM involved the same purification protocol, except after ultrasonic homogenization and centrifugation the supernatant was immediately purified with an IMAC Ni-affinity column without the denaturing-renaturing step.

Samples of both GKAP constructs as well as LC8 were concentrated via ultrafiltration with a 3 kDa molecular weight cut-off value. The buffer was changed to low salt NaPi Buffer with 50 mM NaPi, 20 mM NaCl, set to pH 6.0. Samples, except for GKAP-PBM, were further purified by ion exchange chromatography (IEC), using 5 ml High Q column with the same buffer, collecting recombinant proteins in the flow through fraction. All samples were further purified with size exclusion chromatography (SEC) on a Superdex$^{TM}$ 75 Increase 10/300 GL 24 ml column, with the buffer solution containing 50 mM NaPi and 20 mM NaCl, set to pH 6.0. 5 mM pH 7.4 TCEP was added to GKAP-DLC2 and LC8 samples. The concentration of LC8 and GKAP-PBM was measured by its absorbance at 280 nm, while the concentration of GKAP-DLC2+LC8 complexes was measured with Qubit Protein assay. GKAP-DLC2, LC8, and GKAP-PBM were determined to have a molecular weight of 7.01 kDa, 10.6 kDa, and 5.2 kDa, respectively, validated via SDS-PAGE.

Protocols for the production of Drebrin (D233) samples were developed and carried out by Soma Varga. The full sequence in the *Homo sapiens* Drebrin isoform Q16643 was used as template for cloning the segment 233-317 into NdeI and HindIII sites of a modified pET-15b vector with N-terminal 6xHis-tag and TEV cleavage site

(ENLYFQG). The construct was produced in BL21 (DE3) *E. coli* cells, grown in LB media, and induced with 1 mM IPTG. Cells were incubated for 3 h at 37 °C before harvesting by centrifugation. The lysis buffer contained 50 mM NaPi, 300 mM NaCl, and 5 mM β-mercaptoethanol (BME), as well as 1 mM AEBSF protease inhibitor cocktail, and was set to pH 7.4. IMAC purification was applied to the His-tagged D233, with a modified lysis buffer lacking the protease inhibitor cocktail added to the Ni-affinity column. This was followed by elution, performed with 500 mM imidazole, and the cleavage of the His-tag with TEV protease. D233 was further purified via SEC on the same column as previous samples, with the buffer containing 50 mM NaPi and 20 mM NaCl, set to pH 8.0. Assuming natural isotopic abundance, the molecular weight of the construct was determined to be 10.32 kDa, which was reinforced by SDS-PAGE.

All protein samples were labelled with Green-fluorescent Fluorescein-EX (FITC) using the following protocol: The buffer of GKAP-DLC2, GKAP-PBM, and D233 was changed to 50 mM NaPi and 20 mM NaCl, set to pH 8.0. Based on absorbance measurement, the concentration was 3.5 mg/ml and 2 mg/ml, from which 0.5 mL and 1.5 mL were added to one respective vial of reactive dye. All samples were incubated for one hour while stirring at room temperature, and GKAP-DLC2 samples were additionally stored at 4°C overnight. Any unbound reactive dye was filtered out via SEC. 0.5 mL samples were injected one by one into a 10/300 GL 24 ml column. The flow speed was about 0.8 ml/min with the same buffer as the one used for the labeling. Unlabeled LC8 dimers were added to the fluorescein-labeled GKAP-DLC2 with 2:2 stoichiometry. The final volume of labelled protein solutions was greater than the required minimum of 300 μl. These analytes were focused in the microfluidic devices with low salt NaPi buffer streams containing 50 mM NaPi and 20 mM NaCl, at pH 6.0.

Various layouts were considered for the microfluidic device, all of which were tested with EGFP that was focused with a pH 7.4 PBS buffer containing 137 mM NaCl, 2.7 mM KCl, 10 mM $Na_2HPO_4$, and 1.8 mM $KH_2PO_4$. EGFP analytes were prepared by diluting 2 mg/ml stock solution in equal amount of the same PBS buffer. All measurements involved the surface treatment of the microfluidic device with 1% BSA solution, prepared by dissolving 0.05 g BSA powder in 50 ml of PBS buffer. The preparation of BSA and EGFP samples was carried out by Edit Andrea Jáger.

### 4.2.4. Developing the analytic software

The measured experimental data was processed by an analytic software package that had been written in MATLAB and used Excel files as input, containing pairs of brightfield and fluorescent intensity profiles (dimensionless quantities), their exact position along the microfluidic device (μm), as well as the following general conditions: The width and height of the channel (μm), the thickness of its side walls as displayed in the brightfield image (μm), the flow rate of the analyte and the buffer solutions (μl/min), the absolute temperature during measurement (K), the viscosity of the analyte (Pa*s), the exposition time (s). The script that processed this information omitted measurement points annotated as inadequate due to disruptive conditions such a large contaminant being stuck in the device. It then calculated the mean velocity of particles in the main channel as follows:

Eq. 9. $$v_{particles} = v_{flow} \,/\, (h * w)$$

Where $v_{flow}$ is the sum flow rate of the three inlets, $h$ is the channel height, and $w$ is the channel width. Dividing the distance between measurement points by $v_{particles}$ yielded the average time particles spent between them. Particles were expected to scale between 1 nm and 2 μm, and so limits were drawn to exclude hydrodynamic radii outside this interval, as well as their correspondent inclines in STD.

In some cases, the images had to be rotated (see Designing the experimental setup) before the intensity profiles could be recorded, which led to the inclusion of datapoints outside the image where intensities would drop to zero from the usual 100-200 background noise. These irrelevant segments were removed by a cropping algorithm that defines an intensity band based on the mean value and STD of brightfield datapoints, as seen below. The algorithm iterated from the edges of the profile towards the middle until it came across an intensity value within the band, removing all values up to that point.

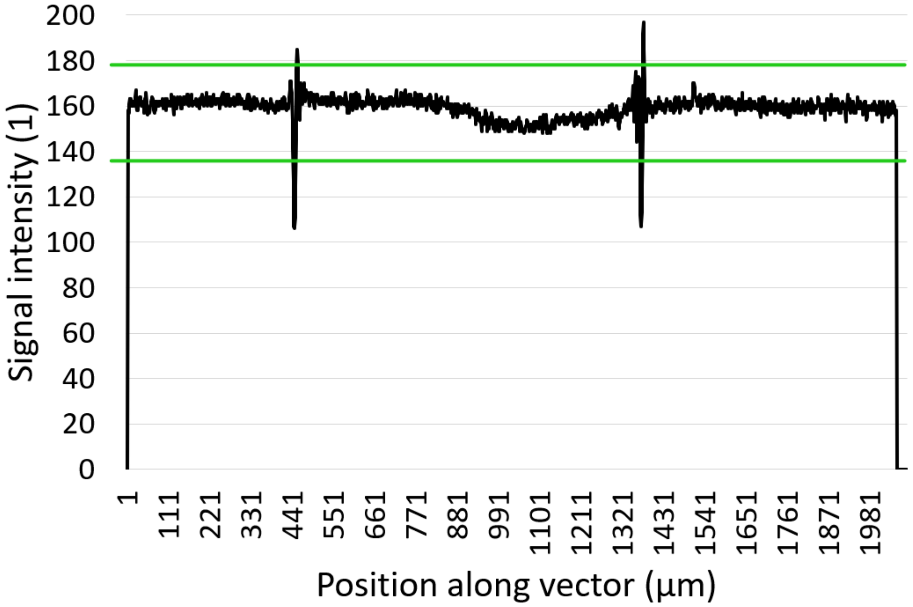


Fig. 11. Brightfield intensity profile of a FITC+GKAP-DLC2 complex sample (black) with horizontal lines (green) denoting the limits of the band used by the cropping function.

The upper limit is defined as the sum of the datapoints' mean value and their STD (177.53). The lower limit is the difference of the same properties (135.66). There are two large spikes extending outside the band in both directions 140-160 µm and 340-360 µm, denoting the positions of the channel's side walls. [II, Fig. 11.]

Cropping was followed by normalization where datapoints outside the microfluidic channel were brought to the same baseline, followed by alignment to the x-axis (intensity = 0) and centering the fluorescent curve on the y-axis (where zero corresponds to the middle of the channel). External datapoints were selected based on their relation to the channel walls, which in turn were identified by their characteristic spikes in the brightfield image (see Fig. 11-12.). The normalization algorithm excluded the sections around these spikes according to the wall thickness defined in the input file, dividing the profiles into one internal and two external segments. Based on the mean value of the external segments brightfield profiles were aligned with the 200-intensity line, which had been observed to be the most common baseline. Fluorescent profiles were aligned to 200 in an analogous manner, aiming to remove any inconsistencies resulting from different noise conditions. However, they were also multiplied by a coefficient in order to even their integrals that assumedly correspond to the number of particles passing through the 2D planes represented by the profiles. Only internal segments were considered for this particular step, followed by aligning the baseline of fluorescent profiles to the x-axis (intensity = 0) and shifting them horizontally to center their peak on the y-axis (distance from the middle of the channel = 0). Associated brightfield profiles were aligned the same way, and while the position of the fluorescent peak does not necessarily correspond to the middle of the channel, this discrepancy is nullified in the next step.

Eight levels of complexity were considered for fitting functions to the measured and processed fluorescent profiles. The level of complexity constituted a single Gaussian function being fit to each profile, while higher levels of complexity meant that a linear combination of two to eight Gaussian functions would be fit. All of this was accomplished using MATLAB's "prepareCurveData" and "fit" functions with the following parameters: the amplitude of the curve must be positive, its center must be between -5 µm and 5 µm, and its STD must be positive but not higher than 100. These parameters were observed to yield the best results. The limits of STD were also adjusted for each consecutive profile based on the minimum and maximum hydrodynamic radii defined at the beginning of the analysis. This ensured that the fitted Gaussian curves would get progressively wider as the profiles got further away from the beginning of the channel. Additionally, each fit was calculated multiple times with different initial parameters until a lower root mean square deviation (RMSE) could not be achieved for five consecutive attempts. Increasing this number did not lead to significantly improved RMSE for the finalized fits.

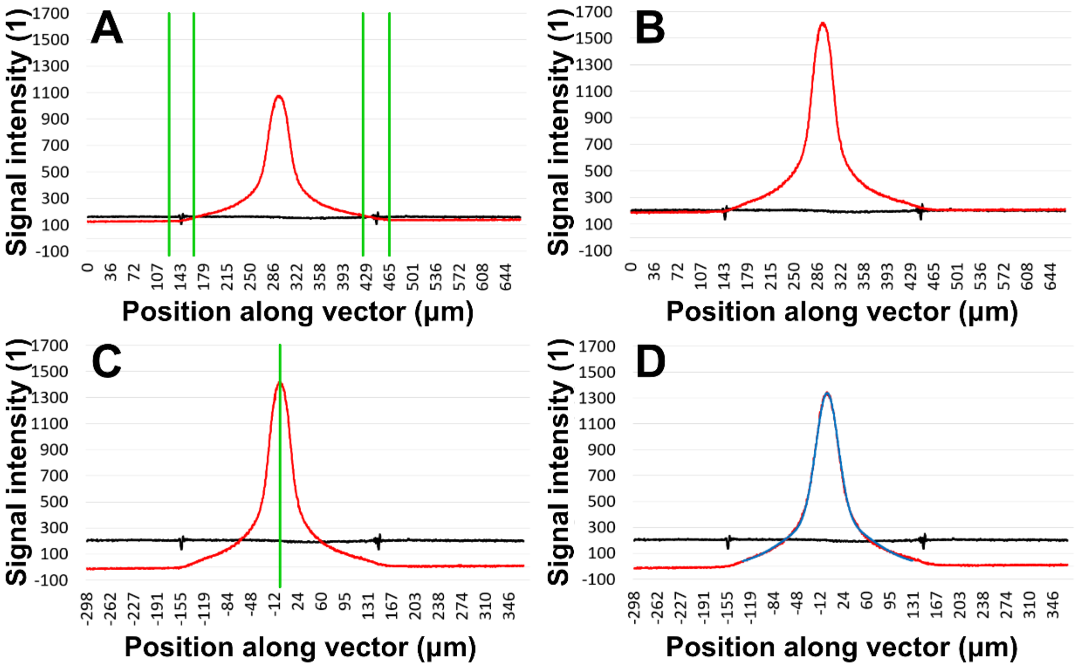


Fig. 12. The main steps of processing measured profiles after they have been cropped. **A)** Sections of manually defined width around spikes in the brightfield profile (black) reveal the position of the channel's sidewalls (green). Their thickness is doubled to account for the uncertainty in the relation between the spikes' position and that of the actual wall. **B)** Both brightfield and fluorescent profiles (red) are aligned with the standard baseline of 200. **C)** The fluorescent profile is adjusted according to its integral, then aligned with the x-axis (signal intensity = 0) as well as centered around its peak value (green). **D)** The internal segment of fluorescent profile is fitted with the linear combination of two Gaussian functions (blue). [II, Fig. 12.]

The algorithm that carries out the fitting process above packaged the Gaussian coefficients and the so-called "Goodness of Fit" metrics into 3D tensors for each measurement point and level of complexity (See The importance of BSA treatment and finding the appropriate fit). If the experimental data consisted of ten measurement points, and the corresponding ten fluorescent profiles were all fitted with the linear combination of eight Gaussian functions, then there would be eighty functions that would be organized into eight sets, each of which would contain a single function belonging to a different measurement point. Their variances would form a linear function of positive incline when paired with the corresponding time components.

Therefore, the eight sets would form eight linear functions, each of which would represent different particle sizes that are supposedly present within the analyte. And so, the penultimate operation carried out by the analytic software was fitting a linear function to each set of datapoints. Finally, the inclines of these linear functions were then converted into diffusion coefficients and hydrodynamic radii.

# 5. Results

## 5.1. Associations between charged sequence motifs and LLPS

### 5.1.1. Survey of the human proteome

Assessment of charged sequence motifs within the human reference proteome revealed that most proteins contain at least one CDR with the more relaxed 5% criterion. Therefore, the analyses below were all carried out using the stricter 1% threshold (see Expanding the dataset with charged sequence motifs and Table 1.). The theorized relations between CDRs and other types of charged sequence motifs are illustrated in Fig. 13. Proteins with transmembrane segments have been excluded from further investigations, as any strong association to phase separation was expected to be characteristic of soluble proteins, and the presence of a large number of transmembrane proteins was expected to skew the analysis.

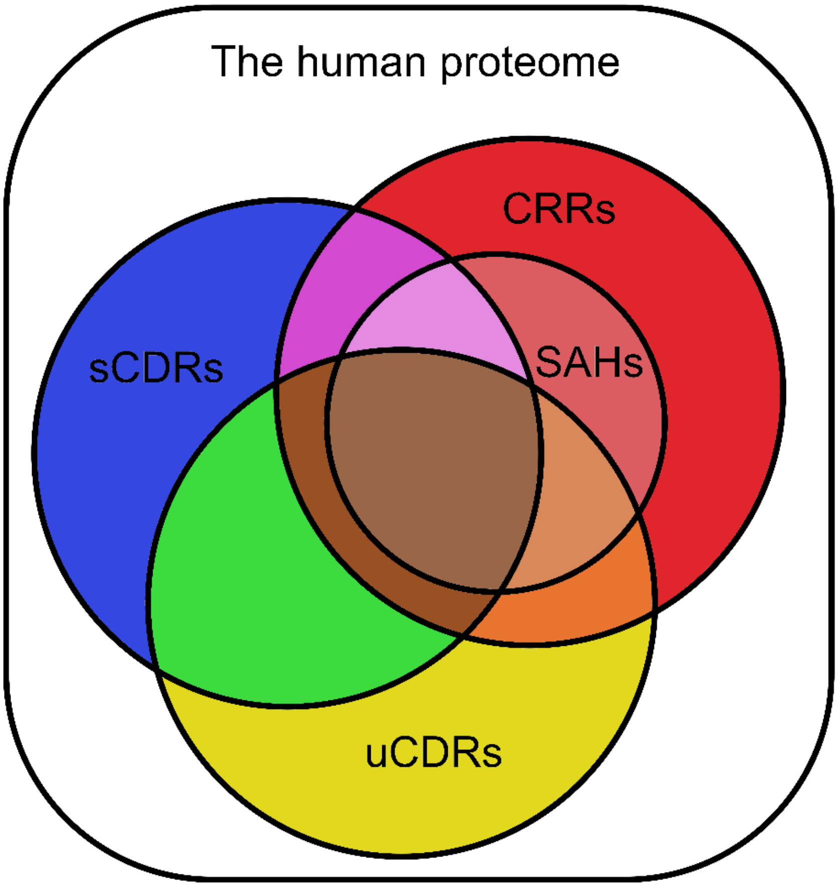


Fig. 13. Illustration of different types of charged sequence motifs within the human proteome. Although, all SAHs qualified as either signed or unsigned CDRs, even with the stricter criterion, it is still possible that there are examples of that motif beyond the definition of charge-dense regions, especially since the reference proteome only included one isoform per UniProtKB entry.

| Sequence motifs | Full proteome | Redundancy-filtered proteomes | | |
|---|---|---|---|---|
| | | 90% | 70% | 50% |
| **All proteins** | 20 659 | 19 638 | 18 294 | 15 672 |
| **uCDRs** | 9 731 | 9 314 | 8 792 | 7 669 |
| **sCDRs** | 14 065 | 13 471 | 12 757 | 11 097 |
| **CRRs** | 1 054 | 1 025 | 985 | 910 |
| **SAHs** | 134 | 131 | 126 | 118 |

Table 1. Number of protein entries that contain at least one type of charged sequence motif in the reference proteome, as well as in its redundancy-filtered variants.

| UniProt ID | Driver motif | uCDRs | sCDRs | CRRs |
|---|---|---|---|---|
| **O60500** | 1077-1241 | 774-807, 1085-1116 | 98-113, 758-791, 1101-1233 | - |
| **Q9NQI0** | 1-236 | - | - | - |
| **P05453** | 136-250 | - | - | - |
| **Q8N884** | 1-146 | 362-444 | 47-64, 347-365 | - |
| **Q8N884** | 161-522 | 362-444 | 47-64, 347-365 | - |
| **Q15648** | 948-1574 | 685-718, 989-1024, 1349-1364, 1442-1577 | 684-704, 831-894, 988-1119, 1285-1300, 1496-1560 | 1459-1565 |
| **O60885** | 674-1351 | 478-747, 1148-1295, 1321-1340 | 276-291, 485-516, 517-599, 701-718, 721-784, 987-1050 | 1220-1253 |
| **Q9A749** | 451-898 | - | - | - |
| **P08287** | 115-225 | - | - | - |
| **Q14781** | 1-532 | 20-167 | 19-121, 123-219, 457-472 | - |
| **Q8L3W1** | 110-220 | - | - | - |
| **Q54VP4** | 146-416 | - | - | - |
| **P06748** | 120-240 | 22-40, 100-292 | 55-205 | 106-234 |
| **Q16082** | 147-182 | - | 164-181 | - |
| **Q07352** | 1-338 | - | 139-154 | - |

Table 2. PhaSePro entries and their motifs with experimental evidence of driving LLPS, as well as any charged sequence motifs within their sequences. None of the CRRs qualified as SAHs. UniProt IDs highlighted in red are not human entries. Charged sequence motifs highlighted in blue overlap with the driver motif of the sequence.

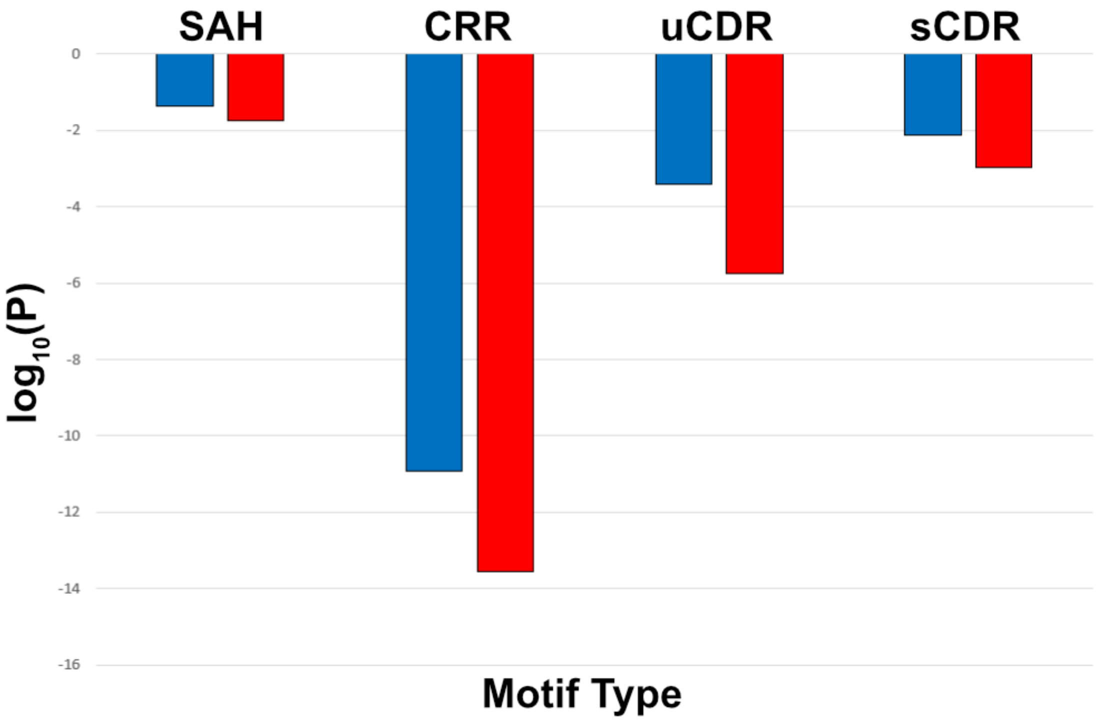


Fig. 14. The common logarithm of P-values from Fisher’s exact test of independence assessing the correlation between human proteins’ propensity towards phase separation and the presence of specific charged sequence motifs (red). All tests were repeated with the exclusion of proteins containing transmembrane segments (blue). [I, Fig. 3.]

Fisher’s tests revealed that CRRs are highly enriched in proteins involved in phase separation, while SAHs showed a weaker but still significant association with the phenomenon (Fig. 14.). These investigations were repeated on the random datasets constructed to match the size distribution and residue composition of the reference proteome. Associations between motifs and the phenomenon proved to be robust.

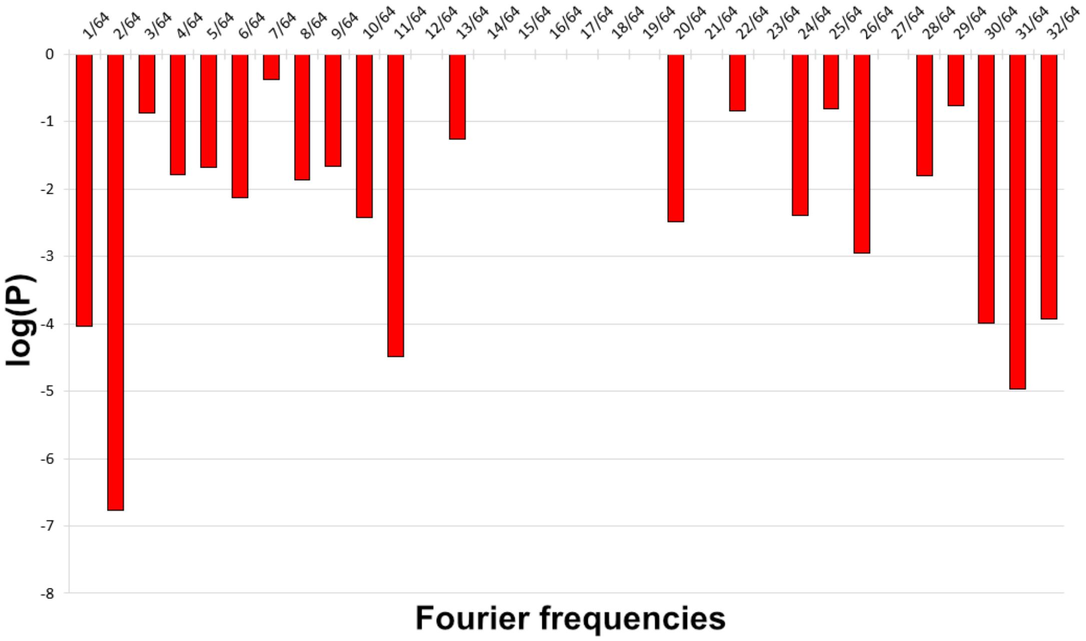


Fig. 15. The association between CRRs and phase separation, broken down into distinct intervals along the Fourier spectrum of charge correlation functions. [I, Fig. 4.] Although weaker than that of CRRs in general, SAHs still exhibited a statistically significant enrichment in phase-separating proteins. They also constitute a subset of CRRs with a characteristic frequency of 1/9 to 1/6, that would translate to an interval between 7.11/64 and 10.67/64 in Fig. 15. Areas of the Fourier spectrum that exhibit stronger associations with LLPS are also the most populated, as shown below.

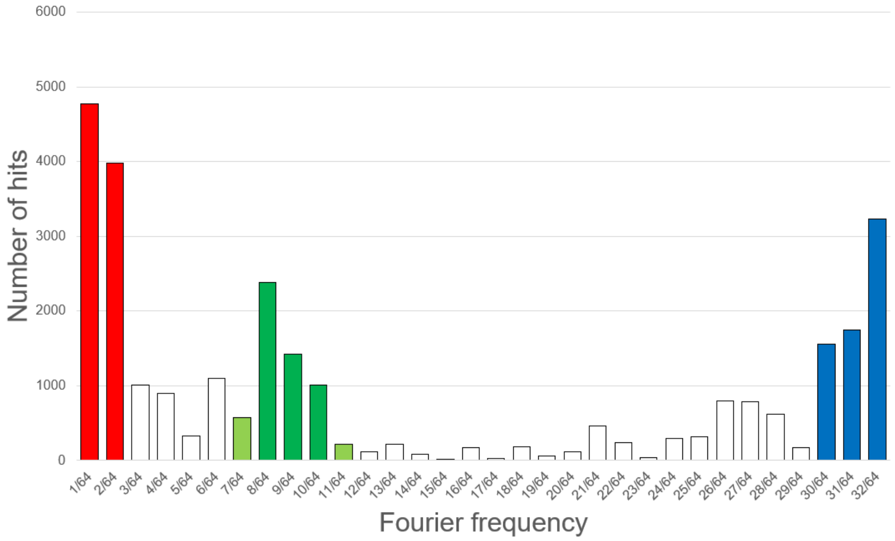


Fig. 16. Distribution of CRRs along the Fourier spectrum of charged correlation functions used by FT_CHARGE to identify these motifs. The three groups with the largest populations are highlighted, among which SAHs are the green group. Changes in residue composition in case of these three groups are detailed in Table 3.

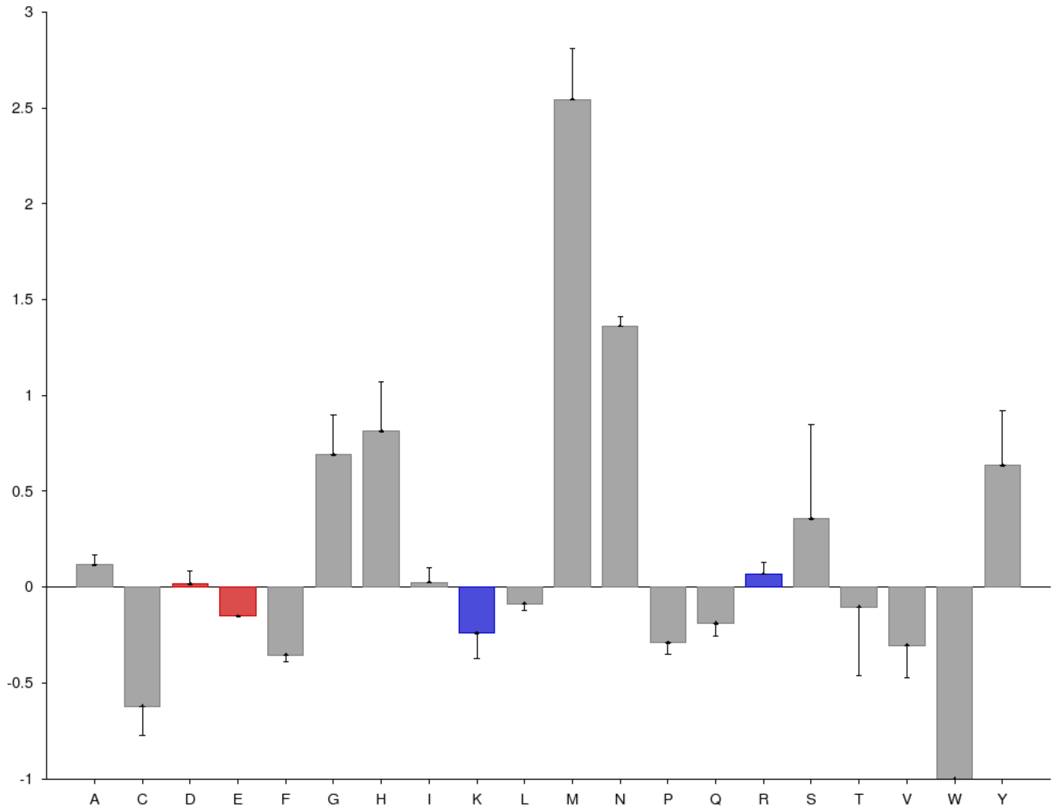


Fig. 17. Shifts in the abundance of residues within SAHs in phase-separating proteins, compared to SAHs in sequences unrelated to the phenomenon. Positively (blue) and negatively (red) charged residues that are essential to SAHs are highlighted.

As expected, the abundance of charged residues shows no significant changes, as they maintain the well-conserved charge pattern characteristic of SAHs (Fig. 17.).

| Residue | CRRs (1/64 to 2/64) | SAHs (CRRs 1/9 to 1/6) | CRRs (30/64 to 32/64) | uCDRs | sCDRs |
|---|---|---|---|---|---|
| A | Enriched | Insignificant | Depleted | Insignificant | Insignificant |
| C | Depleted | Insignificant | Insignificant | Depleted | Depleted |
| D | Insignificant | Insignificant | Enriched | Insignificant | Enriched |
| E | Insignificant | Depleted | Depleted | Depleted | Enriched |
| F | Depleted | Insignificant | Insignificant | Depleted | Depleted |
| G | Insignificant | Enriched | Insignificant | Enriched | Insignificant |
| H | Insignificant | Enriched | Enriched | Depleted | Depleted |
| I | Insignificant | Insignificant | Insignificant | Insignificant | Depleted |
| K | Enriched | Depleted | Depleted | Insignificant | Insignificant |
| L | Depleted | Insignificant | Depleted | Depleted | Depleted |
| M | Depleted | Enriched | Depleted | Insignificant | Insignificant |
| N | Insignificant | Enriched | Insignificant | Insignificant | Insignificant |
| P | Insignificant | Insignificant | Enriched | Enriched | Enriched |
| Q | Insignificant | Insignificant | Depleted | Depleted | Depleted |
| R | Depleted | Insignificant | Enriched | Enriched | Enriched |
| S | Insignificant | Insignificant | Enriched | Enriched | Enriched |
| T | Insignificant | Insignificant | Depleted | Insignificant | Insignificant |
| V | Insignificant | Insignificant | Insignificant | Insignificant | Insignificant |
| W | Depleted | Depleted | Enriched | Insignificant | Depleted |
| Y | Depleted | Insignificant | Insignificant | Depleted | Depleted |

Table 3. Enrichment – or depletion – of residue types in charged sequence motifs within proteins associated with LLPS, compared to their abundance in proteins that do not participate in the phenomenon. The three sub-sets of CRRs correspond to the three groups along the Fourier spectrum with the highest populations, as well as the most significant associations with LLPS (Fig. 16.). The group from 1/9 to 1/6 represents SAHs. Residues highlighted in green or red exhibit the same shift in abundance in at least three different categories.

| Redundancy | Presence of CRRs | Unrelated to LLPS | Related to LLPS | P value |
|---|---|---|---|---|
| **Full reference proteome** | No | 19 418 | 187 | 2.7751*e-14 |
|  | Yes | 1 010 | 44 |  |
| **90% sequence identity** | No | 18 422 | 185 | 5.3264*e-14 |
|  | Yes | 987 | 44 |  |
| **70% sequence identity** | No | 17 118 | 179 | 4.7102*e-15 |
|  | Yes | 951 | 46 |  |
| **50% sequence identity** | No | 14 549 | 166 | 1.2966*e-12 |
|  | Yes | 914 | 43 |  |
| **90% sequence identity (representative isoforms)** | No | 18 431 | 182 | 2.6547*e-14 |
|  | Yes | 981 | 44 |  |
| **70% sequence identity (representative isoforms)** | No | 17 141 | 168 | 6.8380*e-15 |
|  | Yes | 941 | 44 |  |
| **50% sequence identity (representative isoforms)** | No | 14 616 | 146 | 5.2441*e-12 |
|  | Yes | 872 | 38 |  |
|  | **Presence of SAHs** | **Unrelated to LLPS** | **Related to LLPS** | **P value** |
| **Full reference proteome** | No | 20 299 | 226 | 0.0175 |
|  | Yes | 129 | 5 |  |
| **90% sequence identity** | No | 19 283 | 224 | 0.0188 |
|  | Yes | 126 | 5 |  |
| **70% sequence identity** | No | 17 946 | 220 | 0.0211 |
|  | Yes | 123 | 5 |  |
| **50% sequence identity** | No | 15 344 | 206 | 0.2226 |
|  | Yes | 119 | 3 |  |
| **90% sequence identity (representative isoforms)** | No | 19 286 | 221 | 0.0179 |
|  | Yes | 126 | 5 |  |
| **70% sequence identity (representative isoforms)** | No | 17 961 | 207 | 0.0157 |
|  | Yes | 121 | 5 |  |
| **50% sequence identity (representative isoforms)** | No | 15 373 | 181 | 0.1615 |
|  | Yes | 115 | 3 |  |

Table 4. Contingency tables and corresponding P values, yielded by Fisher's exact test of independence for SAHs and CRRs in general. All tests were repeated with proteomes of reduced redundancy, given by two different clustering approaches.

| | Presence of motif | Unrelated to LLPS | Related to LLPS | P value |
|---|---|---|---|---|
| **Top 5% of uCDR hits** | No | 3 561 | 15 | 1.7809*e-6 |
| | Yes | 16 867 | 216 | |
| **Top 1% of uCDR hits** | No | 10 853 | 75 | 4.1511*e-10 |
| | Yes | 9 575 | 156 | |
| **Top 5% of sCDR hits** | No | 2 521 | 13 | 0.0011 |
| | Yes | 17 907 | 218 | |
| **Top 1% of sCDR hits** | No | 6 550 | 44 | 1.3200*e-5 |
| | Yes | 13 878 | 187 | |

Table 5. Contingency tables and corresponding P values, yielded by Fisher's exact test of independence for signed and unsigned CDRs. These tests show how tighter restrictions in the respective scoring schemes of signed and unsigned CDRs provided sequence motifs that occupy a smaller portion of the reference proteome, while having a stronger association with LLPS.

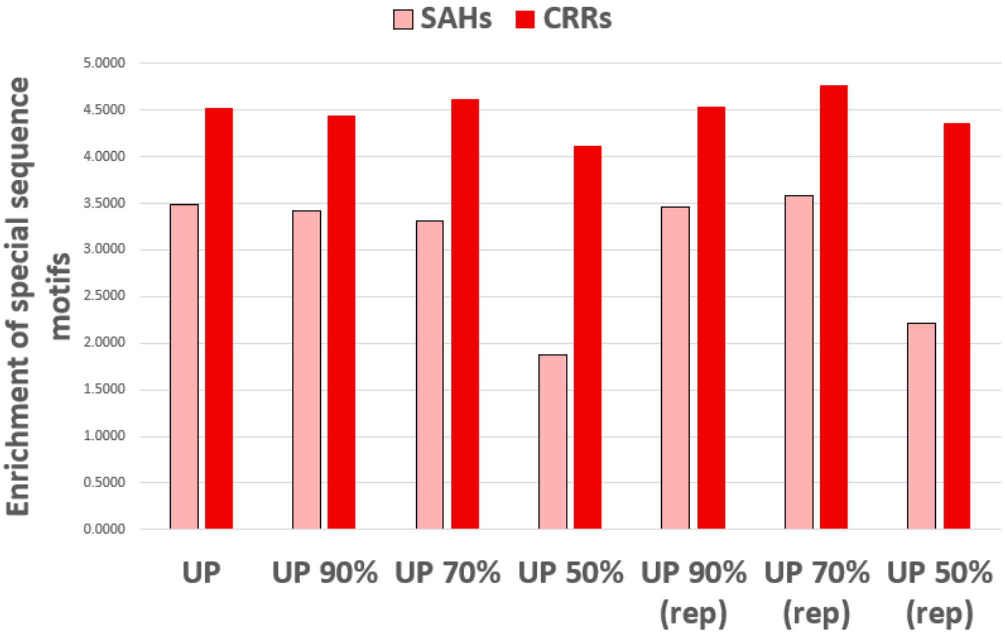


Fig. 18. Enrichment of SAHs and CRRs in phase-separating proteins compared to sequences unrelated to the phenomenon. The assessment was conducted on the human reference proteome constructed from UniProtKB entries (UP), its redundancy-filtered variants (UP 90%, UP 70%, UP 50%), and specifically the representative sequences of those variants (rep). [I, Fig. 5.]

The level of association between structure and function was also investigated through the enrichment of motifs (Fig. 18.). For easier interpretation, the enrichment of CRRs in the entire reference proteome was calculated as follows: The ratio of sequences with CRRs to other entries was 1010:19418 in case of proteins unrelated to phase separation. The same ratio for related proteins was 44:187, marking a 4.52-fold increase between these functional categories.

CDRs were also proven to be prevalent in proteins associated with phase separation. Approximately 47.10% of proteins contained uCDRs, which covered 15.92% of the entire reference proteome sequence-wise. 90.99% of CRRs displayed an above 90% overlap with uCDRs, and 87.80% of CRRs were entirely encompassed by them. Only 2.53% of CRRs were free of any overlap. In contrast, 22.76% of CRRs showed an above 90% overlap with sCDRs, and only 20.25% of them had total coverage. 37.14% of CRRs were completely distinct from any sCDR.

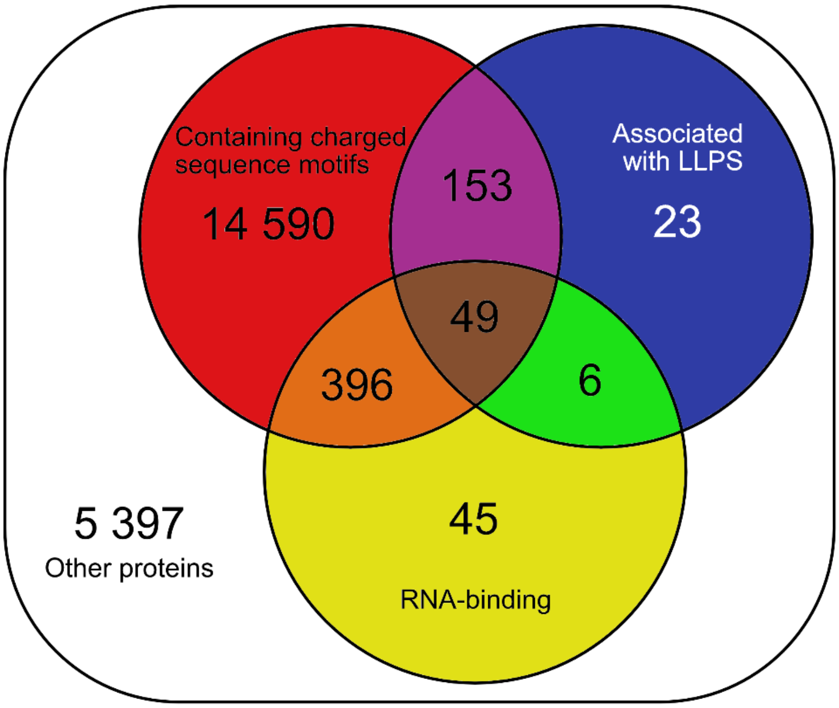


Fig. 19. Number of proteins containing either charged sequence motifs, LLPS-driving regions, or RNA-binding motifs.

While RNA-binding as a functional feature did not exhibit significant association with the presence of any specific type of charged sequence motif, there are only a few RNA-binding entries in the reference proteome that do not contain some type of charged motif (Fig. 19.). The associations between these motifs and a protein’s propensity towards LLPS are further enhanced considering that only a fraction of proteins participate in the phenomenon without containing at least one type of charged motif.

| | Presence of motif | Drivers | Clients |
|---|---|---|---|
| **CRRs** | No | 49 | 138 |
| | Yes | 6 | 38 |
| **SAHs** | No | 52 | 174 |
| | Yes | 3 | 2 |
| **Top 5% uCDRs** | No | 2 | 13 |
| | Yes | 53 | 163 |
| **Top 1% uCDRs** | No | 16 | 59 |
| | Yes | 39 | 117 |
| **Top 5% sCDRs** | No | 2 | 11 |
| | Yes | 53 | 165 |
| **Top 1% sCDRs** | No | 14 | 30 |
| | Yes | 41 | 146 |

Table 6. LLPS-associated sequences annotated in PhaSepDB, categorized based on whether they are also annotated in PhaSePro as drivers of the phenomenon. Sequences are further catalogued based on the presence of charged sequence motifs.

All charged sequence motifs seem to be more abundant on clients except SAHs that are quite evenly distributed, although this is probably the product of their low sample size. CDRs are all about three times more frequent in clients than in drivers, while CRRs are over six times more frequent in clients (Table 6.).

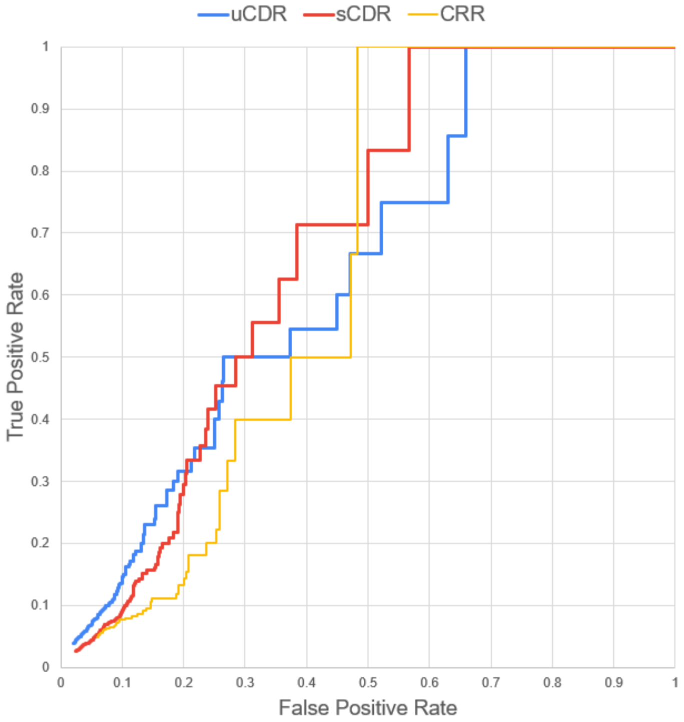


Fig. 20. Receiver operating characteristic (ROC) analysis of the predictive capabilities of different sequence motifs towards phase separation. The corresponding AUC values were 0.5072 (uCDR), 0.4986 (sCDR), and 0.4909 (CRR). [I, Fig. 6.]

ROC analysis shows the rate at which true and false positive predictions are yielded by a model, in this case, using the presence of charged sequence motifs as an indicator that the given protein undergoes phase separation (Fig. 20.). While alone it was proven to be a weak predictor, together with the results of the Fisher tests, their absence was identified as a strong predictor that the given protein would not participate in the phenomenon.

Panther's Gene List Analysis tools were used to carry out overrepresentation tests (OTs), comparing investigated motifs in proteins associated with phase separation versus those in unrelated sequences. [61] These tests were conducted on SAHs, CRRs, sCDRs and uCDRs individually, further dividing them into functional categories, resulting in a total of eight OTs. Each test yielded Gene Ontology (GO) terms associated with the given subset, from which only the ones at the bottom of their respective hierarchy were selected for further analysis, in order to maximize specificity and minimize redundancy. These terms were organized into a 4x4 table based on their enrichment in each of the eight categories, resulting in a gradient table showing the distribution of GO terms from highly relevant to phase separation to highly irrelevant (Table 7.). The table was also used to investigate GO terms containing specific keywords, such as 'response,' 'RNA' or 'metabol'.

| **A)** | **Total number of entries =** | | | | **307** |
|---|---|---|---|---|---|
| Enrich. | 0/4 Unrel. | 1/4 Unrel. | 2/4 Unrel. | 3/4 Unrel. | 4/4 Unrel. |
| 4/4 Rel. | 1 | 0 | 0 | 0 | 0 |
| 3/4 Rel. | 10 | 0 | 0 | 0 | 0 |
| 2/4 Rel. | 53 | 7 | 1 | 0 | 0 |
| 1/4 Rel. | 90 | 5 | 3 | 0 | 0 |
| 0/4 Rel. | 9 | 112 | 16 | 0 | 0 |

| **B)** | **Total number of entries related to ‘RNA’ =** | | | | **55** |
|---|---|---|---|---|---|
| Enrich. | 0/4 Unrel. | 1/4 Unrel. | 2/4 Unrel. | 3/4 Unrel. | 4/4 Unrel. |
| 4/4 Rel. | 0 | 0 | 0 | 0 | 0 |
| 3/4 Rel. | 5 | 0 | 0 | 0 | 0 |
| 2/4 Rel. | 13 | 1 | 0 | 0 | 0 |
| 1/4 Rel. | 26 | 0 | 3 | 0 | 0 |
| 0/4 Rel. | 0 | 7 | 0 | 0 | 0 |

| **C)** | **Total number of entries related to ‘response’ =** | | | | **28** |
|---|---|---|---|---|---|
| Enrich. | 0/4 Unrel. | 1/4 Unrel. | 2/4 Unrel. | 3/4 Unrel. | 4/4 Unrel. |
| 4/4 Rel. | 0 | 0 | 0 | 0 | 0 |
| 3/4 Rel. | 0 | 0 | 0 | 0 | 0 |
| 2/4 Rel. | 10 | 0 | 0 | 0 | 0 |
| 1/4 Rel. | 9 | 0 | 0 | 0 | 0 |
| 0/4 Rel. | 1 | 6 | 2 | 0 | 0 |

| **D)** | **Total number of entries related to ‘metabol’ =** | | | | **22** |
|---|---|---|---|---|---|
| Enrich. | 0/4 Unrel. | 1/4 Unrel. | 2/4 Unrel. | 3/4 Unrel. | 4/4 Unrel. |
| 4/4 Rel. | 0 | 0 | 0 | 0 | 0 |
| 3/4 Rel. | 0 | 0 | 0 | 0 | 0 |
| 2/4 Rel. | 1 | 0 | 0 | 0 | 0 |
| 1/4 Rel. | 2 | 1 | 0 | 0 | 0 |
| 0/4 Rel. | 4 | 14 | 0 | 0 | 0 |

Table 7. Gradient tables displaying the distribution of all GO terms (A), as well as terms related to RNAs (B), responses (C), and metabolism (D). In the latter case the actual keyword was ‘metabol’ to include GO terms containing ‘metabolism’ and ‘metabolic’ alike. The gradient table is color coded from enrichment exclusively in motifs related to

phase separation (blue) to enrichment exclusively in motifs unrelated to the phenomenon (red). [I, Table 2.]

### 5.1.2. Case studies of the phase-separating proteins with charged sequence motifs

Exploring these associations were concluded by conducting case studies that were largely based on the information available in the PhaSePro database about proteins participating in phase separation, specifically LLPS (Table 2.). These case studies revealed that a large variety of proteins may undergo the phenomenon, including but not limited to transport proteins, ribonucleases, splicing factors, transcriptional repressors, translational initiation factors, and Zinc finger proteins. A common feature among these investigated proteins was the molecular function of RNA-binding that 31 out of 44, approximately 68.18% of all cases, possessed (Table 8.). [62-80] This finding reinforced the idea that charged regions such as CRRs promote the process of RNA-binding. It also aligned with the fact that many MLOs formed by LLPS contain RNAs. 5 out of the 44 proteins were also predicted to contain SAHs, all of which constituted more than 10% of their respective sequences. However, apart from these similarities the sequences and functions of the investigated proteins varied to such a degree that only five of them were clustered together into two respective groups. The first group consisted of the probable global transcription activator SNF2L2 and the transcription activator BRG1, showing 59.75% sequence identity. Both are involved in transcriptional activation and repression of select genes by chromatin remodeling, as components in SWI/SNF chromatin remodeling complexes that change chromatin structure through enzymatic activities. [81] The second group contained the proline- and glutamine-rich splicing factor SFPQ as its representative sequence, the non-POU domain-containing octamer-binding protein NONO with 58.39% identity, and the paraspeckle component 1 PSPC1 with 54.49% identity. They cooperatively regulate androgen receptor-mediated gene transcription in the Sertoli cell line [82, 83], and all three are primary components of the paraspeckle. They also share a conserved SAH located at the C-terminal end of a right-handed coiled coil segment, as well as the so-called NOPS region on the N-terminal side of the coiled coil, which has been shown to be responsible for dimerization, and consequently, LLPS. [84] The trinucleotide repeat-containing gene 6B protein (TNRC6C) also contains this NOPS region, acting as a scaffold that simultaneously interacts with argonaute proteins and deadenylase complexes. [85] The E3 ubiquitin-protein ligase RNF168 binds to ubiquitinated histone H2A and H2AX, accumulating repair proteins at sites of DNA-damage. [86] Apart from SNF2L2 and

BRG1, all the above-mentioned cases included at least one alpha-helical structure, and apart from RNF168, all of them exhibited RNA-binding traits.

| UniProtKB ID | Name / description | RNA-binding |
|---|---|---|
| O60832 | H/ACA ribonucleoprotein complex subunit DKC1 | + |
| O60885 | Bromodomain-containing protein 4 | - |
| O95453 | Poly(A)-specific ribonuclease PARN | + |
| O95613 | Pericentrin (centriolar protein) | - |
| P06748 | Nucleophosmin (has diverse chaperonin activities) | + |
| P11387 | DNA topoisomerase 1 | + |
| P11388 | DNA topoisomerase 2-alpha | + |
| P17480 | Nucleolar transcription factor 1 | + |
| P18583 | Protein SON (splicing cofactor) | + |
| P23246 | Splicing factor, proline- and glutamine-rich (part of paraspeckles) | + |
| P23497 | Nuclear autoantigen Sp-100 (part of PML bodies) | - |
| P25490 | Transcriptional repressor protein YY1 | + |
| P33240 | Cleavage stimulation factor subunit 2 (involved in mRNA maturation) | + |
| P38432 | Coilin (part of Cajal bodies) | + |
| P42858 | Huntingtin | - |
| P46100 | Transcriptional regulator ATRX | - |
| P49711 | Transcriptional repressor CTCF | - |
| P51531 | SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 2 | - |
| P51532 | SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 4 | + |
| P61129 | Zinc finger CCCH domain-containing protein 6 | + |
| Q04637 | Eukaryotic translation initiation factor 4 gamma 1 | + |
| Q13428 | Treacle protein (regulator or RNA polymerase I) | + |
| Q14151 | Scaffold attachment factor B2 | + |
| Q14152 | Eukaryotic translation initiation factor 3 subunit A | + |
| Q14676 | Mediator of DNA damage checkpoint protein 1 | - |

| | | |
|---|---|---|
| Q15020 | Spliceosome associated factor 3, U4/U6 recycling protein | + |
| Q15233 | Non-POU domain-containing octamer-binding protein (part of paraspeckles) | + |
| Q15424 | Scaffold attachment factor B1 | + |
| Q15648 | Mediator of RNA polymerase II transcription subunit 1 | - |
| Q16630 | Cleavage and polyadenylation specificity factor subunit 6 | + |
| Q5U5Q3 | RNA-binding E3 ubiquitin-protein ligase MEX3C | + |
| Q6VMQ6 | Activating transcription factor 7-interacting protein 1 | - |
| Q8IYB3 | Serine/arginine repetitive matrix protein 1 (involved in splicing) | + |
| Q8IYW5 | E3 ubiquitin-protein ligase RNF168 | - |
| Q8N684 | Cleavage and polyadenylation specificity factor subunit 7 | + |
| Q8WXF1 | Paraspeckle component 1 | + |
| Q92973 | Transportin-1 | + |
| Q96GM8 | Target of EGR1 protein 1 (regulates cell cycle) | + |
| Q96MU7 | YTH domain-containing protein 1 (regulates splicing) | + |
| Q9H4Z2 | Zinc finger protein 335 (part of histone methyltransferase complexes) | - |
| Q9NR30 | Nucleolar RNA helicase 2 | + |
| Q9NYH9 | U3 small nucleolar RNA-associated protein 6 homolog | + |
| Q9UKY1 | Zinc fingers and homeoboxes protein 1 (transcriptional repressor) | - |
| Q9UPQ9 | Trinucleotide repeat-containing gene 6B protein (involved in RNA interference) | + |

Table 8. List of reference proteome entries that both participate in LLPS and contain at least one type of charged sequence motif.

PSD-specific associations between LLPS and charged sequence motifs included SynGAP1, a component of complex clusters around NMDA receptors in excitatory synapses. Its mouse ortholog has been known to participate in LLPS through its C-terminal segment forming a trimeric coiled coil and interacting with PSD-95 (also known as disk large homolog 4, DLG4) through a PDZ-domain (Fig. 21.). While no CRRs were identified in its sequence, SynGAP as well as PSD-95 contained CDRs. Their interaction contributes to a larger scaffolding protein network in PSDs that involves GKAP, Shank3, Homer3 and the NR2B subunit of NMDAR, which exhibited a complex set of multivalent interactions that led to phase separation *in vitro*. [30] Although none of the proteins contained any CRRs, they all featured CDRs that overlapped with the LLPS-driving region in case of SynGAP and PSD-95, the only two sequences annotated in PhaSePro.

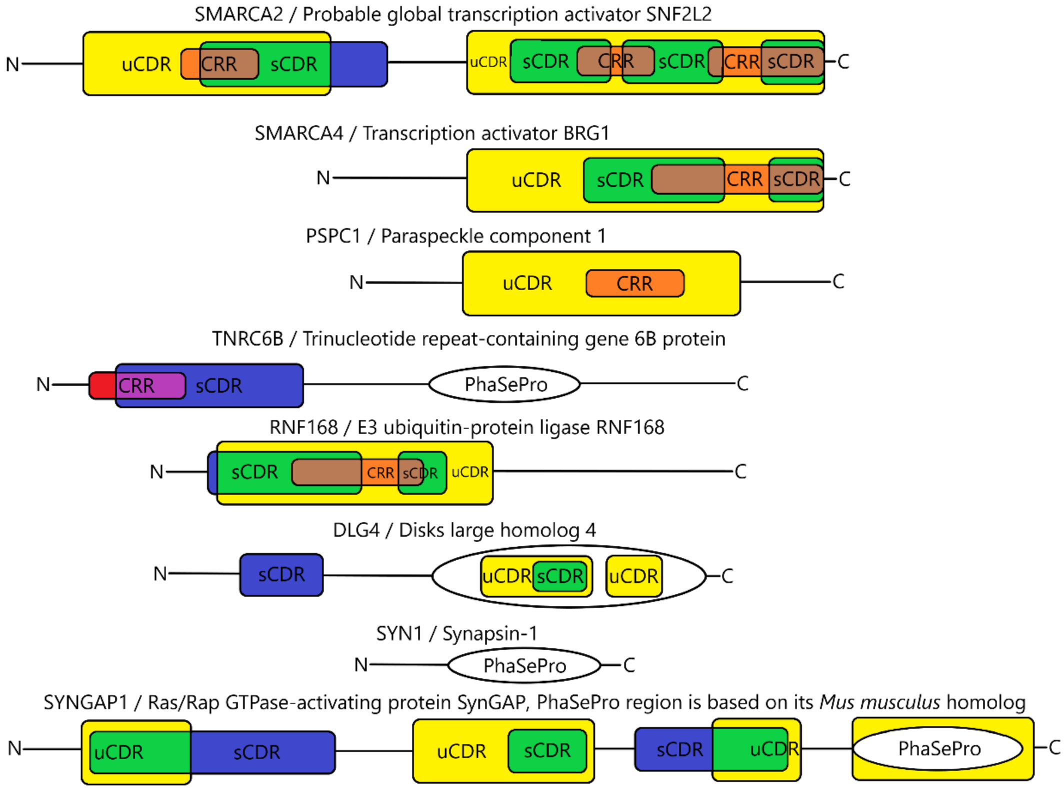

SMARCA2 / Probable global transcription activator SNF2L2
N
uCDR
CRR
sCDR
uCDR
sCDR
CRR
sCDR
CRR
sCDR
C
SMARCA4 / Transcription activator BRG1
N
uCDR
sCDR
CRR
sCDR
C
PSPC1 / Paraspeckle component 1
N
uCDR
CRR
C
TNRC6B / Trinucleotide repeat-containing gene 6B protein
N
CRR
sCDR
PhaSePro
C
RNF168 / E3 ubiquitin-protein ligase RNF168
N
sCDR
CRR
sCDR
uCDR
C
DLG4 / Disks large homolog 4
N
sCDR
uCDR
sCDR
uCDR
C
SYN1 / Synapsin-1
N
PhaSePro
C
SYNGAP1 / Ras/Rap GTPase-activating protein SynGAP, PhaSePro region is based on its Mus musculus homolog
N
uCDR
sCDR
uCDR
sCDR
sCDR
uCDR
PhaSePro
C

Fig. 21. Illustration of the relative positions of different sequence motifs and LLPS-driving regions annotated in PhaSePro. CRRs are highlighted in red, sCDRs are highlighted in blue, and uCDRs are highlighted in yellow, while overlapping regions are color coded according to their constituents. [I, Fig. 7.]

Out of all the case studies that involved both CRRs and LLPS, only the U3 small nucleolar RNA-associated protein 6 homolog was devoid of any predicted disorder. Furthermore, from the remaining 43 sequences only paraspeckle component 1 and the Non-POU domain-containing octamer-binding protein contained disordered segments without any overlaps with charged regions. Not all disordered regions overlapped with CRRs in the other 41 sequences, but each of those proteins contained at least one that did. While this may indicate that some of the association between CRRs and LLPS can be explained by the presence of disorder, there are fourteen sequences in PhaSePro where electrostatic interactions are directly responsible for phase separation, out of which nine are human entries. All of these sequences included either CRRs or CDRs, except the probable ATP-dependent RNA helicase DDX4 protein that contained one region just below the 1% threshold of CDRs. The mediator of RNA polymerase II transcription subunit 1 (MED1) is a great example for how regularly alternating blocks of positively and negatively charged residues can offer a platform for interactions in multiple orientations. It includes a relatively large IDR that shares overlaps with both CRRs and CDRs.  This region of MED1 consists of acidic and basic segments and has also been experimentally proven to facilitate phase separation. [87]

## 5.2. Complex formation investigated in a microfluidic environment

### 5.2.1. Microfluidic device development

The following devices were products of an iterative design process where an initial layout was drafted, a couple of devices using this layout were manufactured and tested, after which the layout was reworked based on the observations. All layouts were designed in Autodesk AutoCAD by Mária Laki.

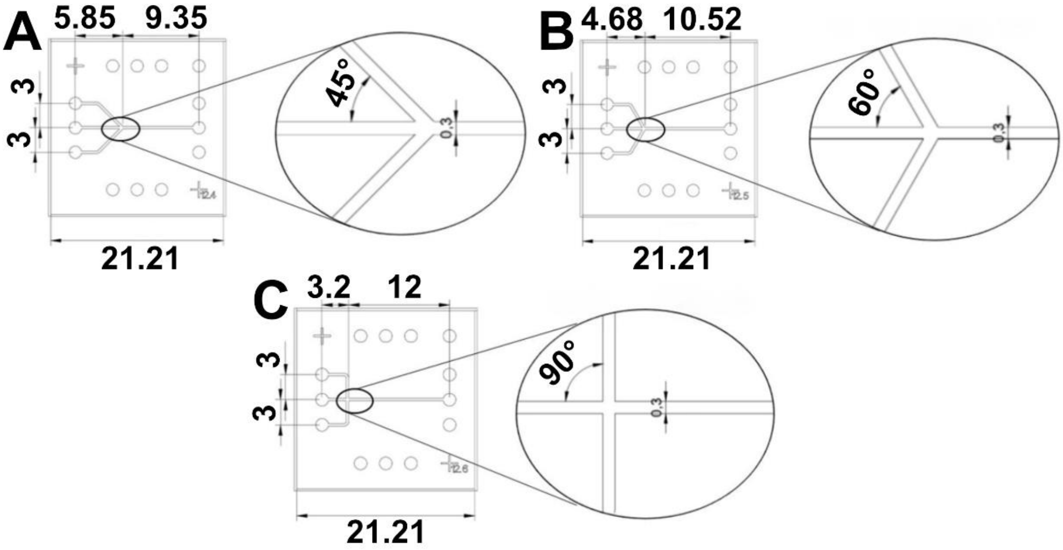


Fig. 22. Three early layouts consisting of one design unit each, featuring microfluidic focuser with three inlets and one outlet. They were used to test flow quality with intersections of different angles: 45° (A), 60° (B), and 90° (C). All size parameters are featured in mm, the channel height is 20 μm. [II, Fig. 4.]

Testing the designs in Fig. 22. led to the conclusion that the angle at the intersection is proportional to the likelihood of backflow, where streams coming from the side inlets would either continue towards the other side or to the middle inlet. Since the analyte had to flow at a lower rate, this usually meant that it would be halted by the side streams. Compared to 90°, the likelihood of backflow was significantly lower in the case of 60° but decreasing it to 45° did not seem to yield further improvements. Placing the inlets closer to the edge of the device greatly increases the probability of leakage, so decreasing the angle inevitably involves placing the intersection further away from the inlets, which results in less space for the channel where diffusion can be observed. Consequently, future layouts featured all included intersections where the angle was about 60° to conserve space for the channel.

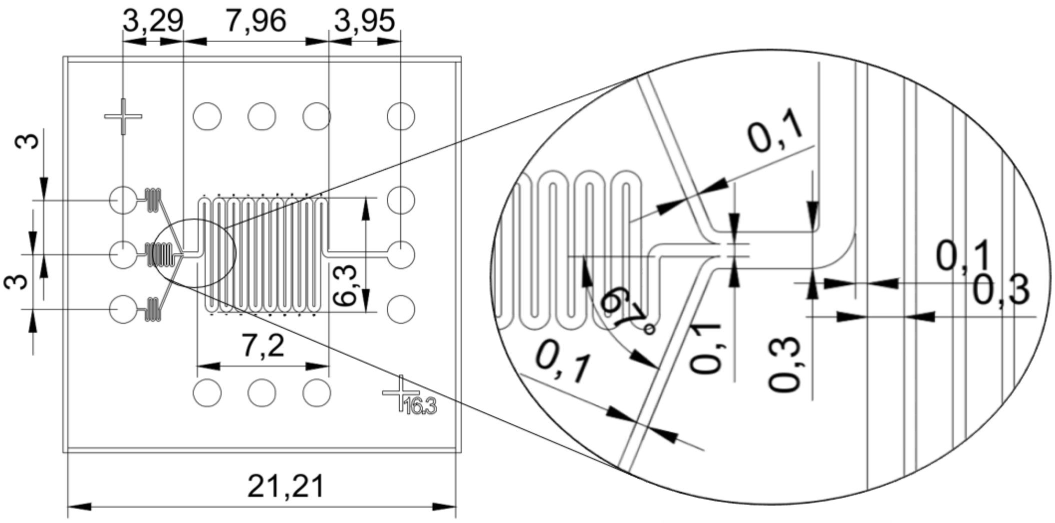


Fig. 23. An improved design where the inlets meet at 67°. All size parameters are featured in mm, the channel height is 20 μm. [II, Fig. 5.]

With a 67° intersection, a wider (300 μm) main channel, and the addition of resistances to all inlets, the likelihood backflow was further diminished, at the cost of leaving even less space for the main channel. To counter this, the design above elongated the main channel with turns but those introduced a centripetal force to the solutions, which meant that particles would be moved laterally by a phenomenon other than diffusion, undermining a basic principle of the measurement. However, the wider main channel proved to be useful for giving particles more space before they reached the sidewalls, allowing lower flow rates, so later designs kept this feature (Fig. 23.).

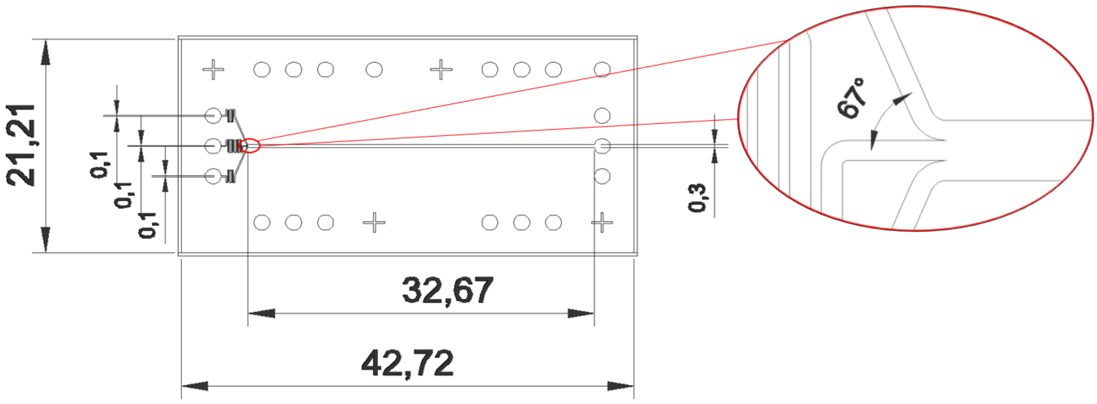


Fig. 24. The first device that occupied two design units. All size parameters are featured in mm, the channel height is 20 µm. [II, Fig. 6.]

Devices larger than one design unit were avoided before, because they cannot be bound to cover plates, only glass slides, limiting the possible magnification to 40x. However, the new main channels could not be recorded in their entire width above 40x magnification anyway, which alleviated this restriction. Testing the layout in Fig. 24. concluded that multi-unit devices with wide main channels do not diminish the quality of the measured data.

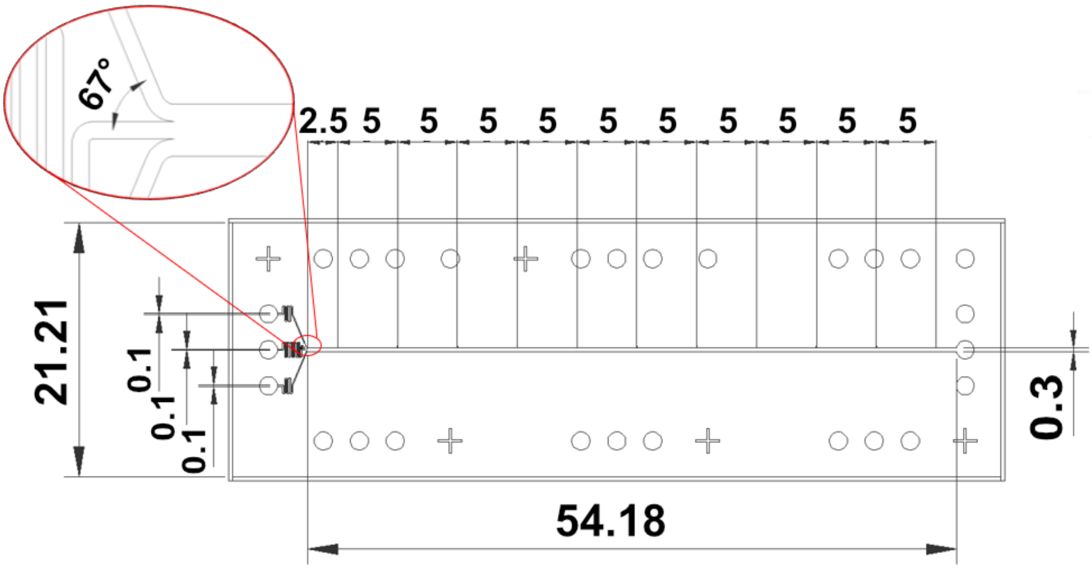


Fig. 25. The latest device, encompassing three design units. All size parameters are featured in mm, the channel height is 20 µm. [II, Fig. 7.]

Three design units is the largest size a device can have and still fit on a single glass slide, therefore the 54 mm main channel in the layout above is the longest possible channel without including turns or compromising the features that minimize the likelihood of backflow. Another important improvement was placing markers at pre-determined measurement points, aiding the location of appropriately distant locations for recording images, and therefore, profiles (Fig. 25.). The Reynolds number calculated for the main channel of this device is $Re = \frac{\rho v L}{\eta} = 9.375 * 10^{-4}$, where the fluid density was approx. $\rho = 1000\ \frac{kg}{m^3}$, the average fluid velocity was $v = 0.025\ \frac{m}{s}$, the characteristic length of the rectangular channel was $L = 37.5 * 10^{-6}\ m$, and the dynamic viscosity of the fluid was approx. $\eta = 1\ \frac{kg}{m*s}$.

### 5.2.2. The importance of BSA treatment and finding the appropriate fit

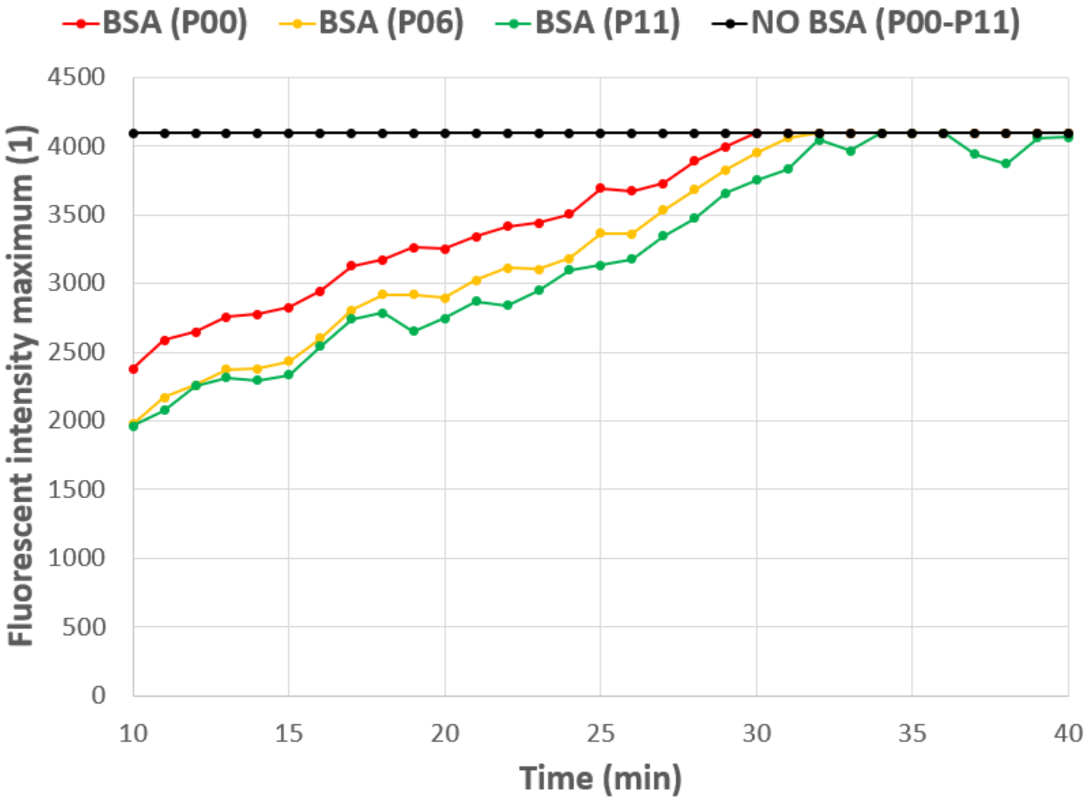


Fig. 26. Peak fluorescent intensity at the intersection of the microfluidic device (P00), about halfway down its main channel (P06), and just before its outlet (P11). Intensity was measured every minute for half an hour in a microfluidic device treated with 1% BSA, and in another one without treatment. In both cases, the analyte consisted of 0.05 µm fluorescent microspheres. The first measurement took place 10 mins after the analyte had reached the intersection, marking the minimum waiting period before the flow would be considered steady (see Designing the experimental setup). [II, Fig. 8.]

The previously described microscopic setup can only record fluorescent signals up to 4000 in intensity. Without treatment, the peak intensity surpasses this limit throughout the entire device before any measurements could take place (Fig. 26.), proving the necessity of coating the internal surface of the channel with BSA.

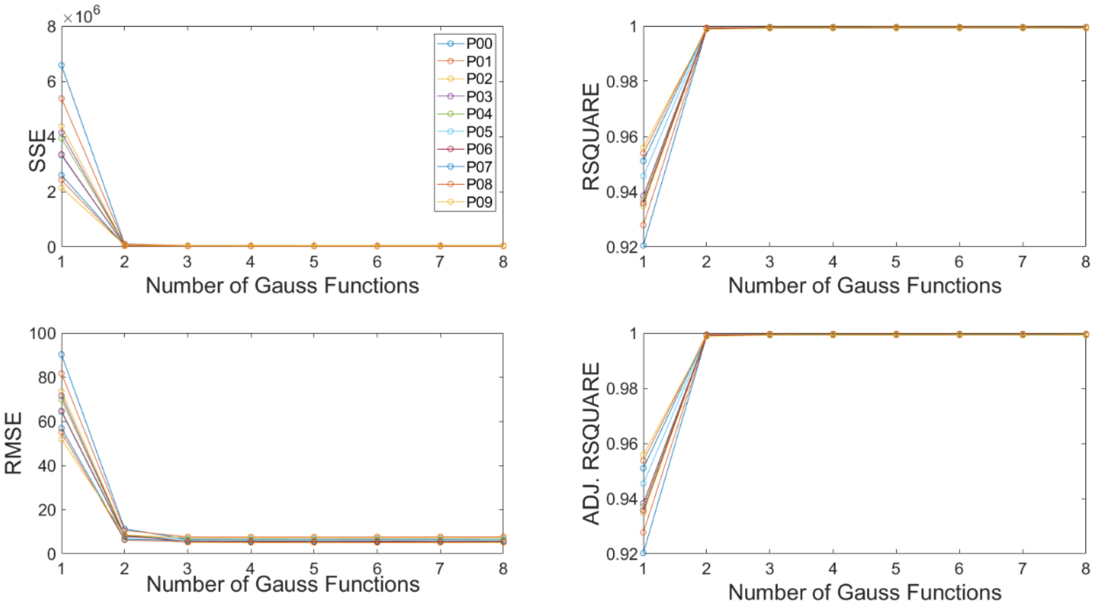


Fig. 27. Goodness of fit (GoF) metrics, calculated for different measurement points (P00-P09) and eight levels of complexity. This measurement used EGFP, but other analytes also exhibited the same trend in all four GoF metrics. [II, Fig. 9.]

The following statistical methods were used to determine the best model:

- Sum of squares due to error (SSE) was used to quantify the variation between measured profiles and their fits. Lower values correspond to less discrepancy between data and model.

  Eq. 10. $$SSE = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

  where $y_i$ are the datapoints, $\hat{y}_i$ are the fits to the datapoints, $w_i$ are the weights of the datapoints, and $n$ is the number of datapoints.

- Root mean squared error (RMSE) was used to determine the average difference between experimental datapoints and values of the model, quantifying how dispersed the measured points were compared to the Gaussian curve. Similarly to SSE, lower values suggest a better fit.

  Eq. 11. $$RMSE = \frac{\sqrt{SSE}}{v}$$

  where $v = n - m$ is the residual degrees of freedom defined as the number of datapoints minus the number of fitted coefficients.

- RSQUARE, or R-squared, is used in statistics to determine how much of the variation of a dependent variable can be explained by an independent variable, from 0 to 1. In this case, the dependent variable was the fluorescent intensity value, and the independent variable was its position, while the R-squared value was maximized to find the line of best fit between the two. Unlike the previous two metrics, lower values are associated with worse fits.

Eq. 12. $$RSQUARE = 1 - \frac{SSE}{\sum_{i=1}^{n} w_i(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean value of the datapoints. The expression in the denominator is also called the sum of squares about the mean (SST).

- The degrees of freedom adjusted R-squared (ADJ. RSQUARE) is quite similar, but it penalizes the inclusion of irrelevant variables, therefore it can take on negative values.

Eq. 13. $$ADJ.RSQUARE = 1 - \frac{SSE(n-1)}{SST(v)}$$

All four metrics followed the same tendency, where using the linear combination of two Gaussian functions yielded significantly better fits, but the results did not improve much at higher levels of complexity (Fig. 27.). Since all analytes were assumed to be monodisperse, low complexity models were considered to have more physical meaning. Therefore, all profiles were fitted with the linear combination of two Gaussian functions.

### 5.2.3. Approximated particle sizes

During development fluorescent microspheres (MS) of various nominal diameters were used in addition to EGFP for testing different microfluidic device layouts. Some of these particles were far outside the scale of the investigated PSD proteins and their complexes; however, they provided insight into the limits of the experimental setup (Table 9.). Control measurements of dynamic light scattering (DLS) were carried out on each particle type by Eszter Nagy-Kanta.

| Analyte | Expected radius (nm) | Approximate radius (nm) | Lower approx. radius (nm) | Higher approx. radius (nm) |
|---|---|---|---|---|
| GKAP-PBM | 1.39 | 1.16 ± 0.10 | 1.06 ± 0.00 | 1.26 ± 0.03 |
| GKAP-DLC2 | 1.27 ± 0.72<br>1.90 ± 1.00 | 1.36 ± 0.30 | 1.15 ± 0.18 | 1.56 ± 0.24 |
| GKAP-DLC2 + LC8 | 2.17 ± 0.44<br>3.11 ± 0.76 | 1.64 ± 0.36 | 1.38 ± 0.27 | 1.91 ± 0.30 |
| D233 | 1.81 | 1.45 ± 0.39 | 1.06 ± 0.00 | 1.83 ± 0.09 |
| EGFP | 2.72 ± 1.01 | 1.72 ± 0.42 | 1.42 ± 0.36 | 2.03 ± 0.20 |
| 50 nm MS | 22.68 ± 2.11 | 9.28 ± 11.09 | 6.78 ± 9.46 | 11.78 ± 12.00 |
| 200 nm MS | 97.09 ± 9.51 | 15.15 ± 13.36 | 5.19 ± 5.11 | 25.12 ± 11.50 |
| 1100 nm MS | 461.61 ± 35.59 | 6.62 ± 5.78 | 1.62 ± 0.47 | 11.62 ± 4.07 |

Table 9. List of the analytes, their expected radii as revealed by DLS (or estimated from molecular size in the case of GKAP-PBM and D233), and their approximate radii given by the diffusion-based approach. Two DLS measurements were carried out for both the FITC-labelled GKAP-DLC2 and GKAP-DLC2+LC8 complex, which are included as separate rows. The mean values and STDs for approximate radii were calculated from multiple diffusion measurements. The analytic software was set to fit the linear combination of two Gaussian functions to the measured fluorescent intensity profiles (see The importance of BSA treatment and finding the appropriate fit). Because of that, the output for each measurement was a pair of hydrodynamic radii, each corresponding to one set of Gaussian functions. The values in column three include both radii, while the values in columns four and five respectively reflect either the larger or smaller radii yielded by the measurements.

The mean values of approximate radii were quite similar to that of their expected values, exhibiting the same trend where GKAP-PBM proves to be the smallest particle type, followed by GKAP-DLC2, D233, GKAP-DLC2+LC8 hexamers and EGFP. However, the diffusion-based approach yielded lower STDs for these samples than DLS. This might be explained by a higher sensitivity towards smaller particles, the diffusion of which is much more significant with the same conditions. This bias was mitigated by discarding the lower hydrodynamic radius from each measurement (Fig. 28.). Another explanation for the higher STD of DLS is that GKAP-PBM, D233, GKAP-DLC2, GKAP-DLC2+LC8 hexamers and EGFP all occupy the lower end of its 1 nm to 10 μm dynamic range, where precision becomes increasingly system-dependent.

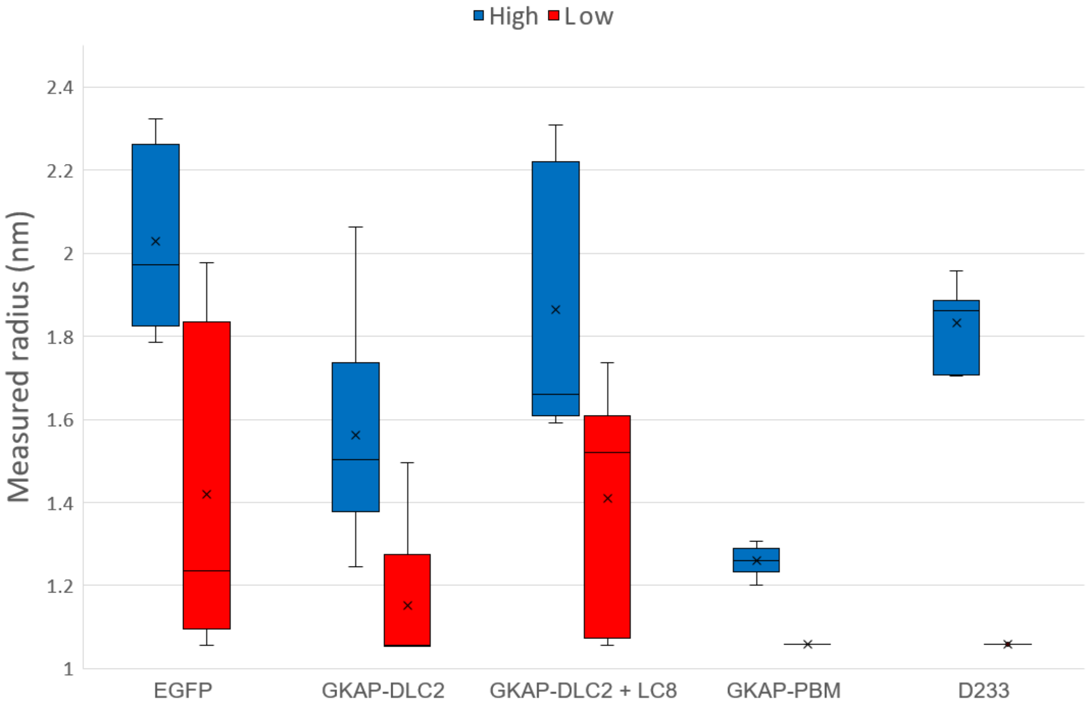


Fig. 28. Box plot of approximated radii for EGFP, GKAP-DLC2, GKAP-DLC2+LC8 hexamers, GKAP-PBM, and D233. The lower limit for approximate radii was 1 nm; however, due to rounding errors, the actual limit turned out to be about 1.05 nm, which is reflected by the boxes containing the lower value from each measurement (red). [II, Fig. 10.]

# 6. Discussion

The associations between charged sequence motifs and phase separation were demonstrated *in silico* and reinforced via multiple approaches. These associations are weakened with the exclusion of transmembrane proteins but remained significant even in the filtered dataset. Although this decision was based on the expectation that soluble proteins are more likely to participate in phase separation, it is important to note that some transmembrane proteins such as Nephrin and the linker for activation of T-cells family member 1 have been associated with the phenomenon. There are also various transmembrane proteins containing at least one SAH motif, including the sodium bicarbonate transporter 3, which is associated with membrane proteins that are in turn associated with vesicles.

The results also clearly showed that all investigated motifs, especially CRRs, are more abundant in LLPS-associated proteins than random datasets, suggesting that the described associations are not by-products of some trivial biophysical constraint. The investigations also concluded that the presence of charged sequence motifs is not a strong indicator for MLO formation. It is also highly unlikely that the motifs analyzed in this study would be directly responsible for LLPS. The enrichment of CRRs and CDRs in LLPS-prone proteins probably involves a more complex explanation, especially since little has been revealed about the structural features of the investigated motifs. Unsurprisingly, many regions rich in charged residues were predicted to be intrinsically disordered. However, considering that SAHs possess a stable well-characterized structure, it is possible that other types of CRRs similarly have specific preferences for either monomeric or oligomeric structures. And with the ever-increasing number of proteins experimentally associated with biomolecular condensates, as well as the improvement of their curation, the available knowledge about phase-separating proteins and involved molecular mechanisms is expected to expand further. For now, charged residues seem to provide structural and dynamical features that can be robustly maintained in multiple material phases, which allows the precise positioning of regions directly responsible for phase separation. Since their publication, these results have been used in research projects about conserved secondary structures that are incorrectly predicted as disordered regions, fungal circadian clocks, and how LLPS affects kinase signaling. [88-90]

Applying microfluidics and fluorescent microscopy to measure the size of solute particles based on their diffusion has been successfully adapted to monitoring constructs of Drebrin, GKAP, LC8, as well as the hexameric complexes of the latter two. Additionally, certain limitations of the technique were revealed that mostly involved larger particles. The approach harbored an inherent bias towards smaller particles that exhibit more prominent motions via diffusion. It yielded accurate information about particles below 10 nm in diameter, which included the primary targets of this study. Furthermore, it did so more precisely than DLS, one of the most widespread techniques for investigating the formation of MLOs and condensates *in vitro*. The precision of DLS on this scale is highly system-dependent; therefore, the relation between the results of these two techniques may differ for other protein samples.

As for possible improvements regarding the diffusion-based approach, uniting the inlets of the side streams would nullify the slight instabilities in the flow that occasionally manifest due to asynchronous pulsing of their syringe pumps. It has also been concluded that results should be cross-referenced with other methods. It is important to note that previous implementations of this technique did not directly measure the diffusion coefficient of solute particles. Instead, they recorded the changes in the fluorescent signal's shape along the channel and assigned them to the given analyte, based on which different particles were distinguished. Circumventing the utilization of the Stokes-Einstein equation makes the maintenance of laminar flow unnecessary and the identification of larger particles feasible. However, the approach described in this study yields more precise results for PSD proteins and performs a step towards understanding their interactions that lead to phase separation.

In summary, charged sequence motifs have been shown to possess robust associations with protein phase separation, and the LLPS-related complex formation of GKAP and LC8 has been observed with a novel *in vitro* approach that has specific advantages and disadvantages compared to dynamic light scattering. The specialized microfluidic device used for this diffusion-based approach is the product of an iterative and highly collaborative effort spanning several years of designing and testing, including multiple revisions of both its layout and the accompanying experimental setup and software. Characterization of complex biomolecular systems like multivalent complexes and condensates requires integrated approaches that combine experimental and computational methods. [91] For the methods described in this thesis, this kind of synergy is not yet straightforward as currently no quantitative modeling approach can link the two in a direct manner. However, the dynamic nature of ionic interactions in charged segments, as well as the structured nature of SAHs is not trivially captured by simulations, especially for multicomponent assemblies. Thus, computational analysis of LLPS-prone regions along with experimental methods capable of capturing heterogeneous systems can still be combined in a meaningful way in the near future. I believe that the insight given from sequence analysis and the microfluidics-based setup will eventually contribute to a better description of condensates, and will be used to characterize specific systems like MLOs in the PSD, as already initiated in our research group.

## 7. Thesis points

**Thesis I.** I developed novel *in silico* methods as well as utilized already existing ones to identify multiple types of charged sequence motifs. [I]

**Thesis II.** I confirmed the existence of robust associations between the presence of different charged sequence motifs and the given protein's propensity towards phase separation. These are mostly negative associations, meaning that the absence of investigated motifs makes protein phase separation unlikely. [I]

**Thesis III.** I developed an *in vitro* approach that determines the size of proteins and their complexes, based on their lateral diffusion during laminar flow. [II]

**Thesis IV.** I proved that the results yielded by this approach are consistent with *a priori* data, and that it is more precise than dynamic light scattering for particles under 10 nm in diameter. [II]

## 8. Data Accessibility

Tables summarizing the predictions and calculations supporting the findings about charged sequence motifs are available in the Supplementary Material of the corresponding article. Detailed prediction outputs are available from the corresponding author [gaspari.zoltan@itk.ppke.hu] upon reasonable request. Measured fluorescent and brightfield intensity profiles that support the findings regarding the developed diffusion-based approach are available in the following Zenodo dataset: doi:10.5281/zenodo.15394328. The analytic software created to process the measured intensity profiles is available in the following Zenodo software package: doi:10.5281/zenodo.15394359. Previous layouts of the microfluidic device along with the description of its gradual development are available as the following collection of figures: doi:10.5281/zenodo.15773654.

# 9. Contributions

Zoltán Gáspári advised me on research direction, assisted in the interpretation of my results, and connected me with experts from different scientific fields who all contributed to this multidisciplinary study. Rita Pancsa provided her insight into the studied sequences motifs and related computational approaches. Anna Sánta contributed to the analysis of the sequence motifs and their correlation to protein phase separation. Eszter Nagy-Kanta, Edit Andrea Jáger, and Soma Varga prepared the samples the diffusion-based size determination approach was tested with. Eszter Nagy-Kanta provided the DLS results the approach was compared to. Csaba István Pongor assisted with the development of the measurement protocol. Mária Laki and András József Laki produced the microfluidic devices and lent their insight into the design process.

The following tasks were my contributions to the collaborative effort: I compiled a proteome-scale dataset from openly available online databases, as well as files about SAHs and other CRRs. I clustered the dataset with CD-HIT and wrote a software package in MATLAB to carry out Fisher's exact test of independence, ROC analyses, and further investigations regarding sequence composition and GO terms. I contributed to the case studies about sequences that contained charged motifs and participated in phase separation. I contributed to the design of the diffusion-based approach, the development of microfluidic devices, and the conducted measurements. I developed the analytic software package that processed the measured data. I contributed to the analysis and interpretation of all results.

# 10. Acknowledgements

This work was supported by the National Research, Development and Innovation Office under grant TKP2021-EGA-42.

# 11. Publications

## 11.1. Journal articles

I. A. L. Szabó, A. Sánta, R. Pancsa, Z. Gáspári, “Charged sequence motifs increase the propensity towards liquid–liquid phase separation,” *FEBS Lett.*, vol. 596, no. 8, pp. 1013-1028, Apr 2022

II. A. L. Szabó, E. Nagy-Kanta, S. Varga, E. A. Jáger, C. I. Pongor, M. Laki, A. J. Laki, Z. Gáspári, “Diffusion-based size determination of solute particles: a method adapted for PSD proteins,” *FEBS Open Bio*, Accepted for publication, preprint doi:10.1101/2025.02.05.636588

III. Z. Harmat, A. L. Szabó, O. Tőke, Z. Gáspári, “Different modes of barrel opening suggest a complex pathway of ligand binding in human gastrotropin,” *PLoS ONE*, vol. 14, no. 5, e0216142, May 2019, doi:10.1371/journal.pone.0216142

## 11.2. Conference presentations

IV. A. L. Szabó. (Jul 2022). Charged sequence motifs increase propensity towards liquid-liquid phase separation. Presented at IUBMB-FEBS-PABMB Young Scientists’ Forum 2022, Vimeiro, Portugal. [Poster]. Available: Young Scientists’ Forum 22 Programme and Abstract Book (pp. 163)

V. A. L. Szabó., A. Sánta, R. Pancsa, Z. Gáspári. (Jul 2022). Charged sequence motifs increase propensity towards liquid-liquid phase separation. Presented at The Biochemical Global Summit 2022, Lisbon, Portugal. [Poster]. Available: https://febs.onlinelibrary.wiley.com/doi/epdf/10.1002/2211-5463.13440 (pp. 321)

VI. A. L. Szabó, A. Sánta, R. Pancsa, E. A. Jáger, C. Pongor, A. J. Laki, Z. Gáspári. (Nov 2022). Phase separation of postsynaptic proteins: insights from bioinformatics and microfluidics. Presented at 1st FEBS-IUBMB-ENABLE Conference, Seville, Spain. [Poster].

VII. A. L. Szabó, E. A. Jáger, C. I. Pongor, A. J. Laki, Z. Gáspári. (Jun 2024). Diffusion-based analysis of phase separating PSD proteins: combining microfluidics with fluorescent microscopy techniques. Presented at 48th FEBS Congress, Milano, Italy. [Poster]. Available: https://febs.onlinelibrary.wiley.com/doi/epdf/10.1002/2211-5463.13837 (pp. 135)

# 12. References

1. X. Liu, X. Liu, H. Wang, Z. Dou, K. Ruan, D. L. Hill, et al., “Phase separation drives decision making in cell division,” *J. Biol. Chem.*, vol. 295, no. 39, pp. 13419-13431, Sep 2020
2. D. M. Mitrea, R. W. Kriwacki, “Phase separation in biology; functional organization of a higher order,” *Cell Commun. Signal.*, vol. 14, article number: 1, Jan 2016
3. Z. Feng, X. Chen, M. Zeng, M. Zhang, “Phase separation as a mechanism for assembling dynamic postsynaptic density signalling complexes,” *Curr. Opin. Neurobiol.*, vol. 57, pp. 1-8, Aug 2019
4. S. Boeynaems, S. Alberti, N. L. Fawzi, T. Mittag, M. Polymenidou, F. Rousseau, et al., “Protein Phase Separation: A New Phase in Cell Biology,” *Trends Cell Biol.*, vol. 28, no. 6, pp. 420-435, Jun 2018
5. Y. Shin, C. P. Brangwynne, “Liquid phase condensation in cell physiology and disease,” *Science*, vol. 357, no. 6357, eaaf4382, Sep 2017
6. S. F. Banani, A. M. Rice, W. B. Peeples, Y. Lin, S. Jain, R. Parker, et al., “Compositional Control of Phase-Separated Cellular Bodies,” *Cell*, vol. 166, pp. 651-663, Jul 2016
7. A. Molliex, J. Temirov, J. Lee, M. Coughlin, A. P. Kanagaraj, H. J. Kim, et al., “Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization,” *Cell*, vol. 163, pp. 123-133, Sep 2015
8. R. Pancsa, W. Vranken, B. Mészáros, “Computational resources for identifying and describing proteins driving liquid-liquid phase separation,” *Brief Bioinform.*, vol. 22, no. 5, pp. 1-20, Sep 2021
9. W. M. Babinchak, W. K. Surewicz, “Liquid–Liquid Phase Separation and Its Mechanistic Role in Pathological Protein Aggregation,” *J. Mol. Biol.*, vol. 432, no. 7, pp. 1910-1925, Mar 2020
10. Y. Lin, S. L. Currie, M. K. Rosen, “Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs,” *J. Biol. Chem.*, vol. 292, no. 46, pp. 19110-19120, Nov 2017
11. H. R. Li, W. C. Chiang, P. C. Chou, W. J. Wang, J. R. Huang, “TAR DNA-binding protein 43 (TDP-43) liquid-liquid phase separation is mediated by just a few aromatic residues,” *J. Biol. Chem.*, vol. 293, no. 16, pp. 6090-6098, Apr 2018

12. C. W. Pak, M. Kosno, A. S. Holehouse, S. B. Padrick, A. Mittal, R. Ali, et al., "Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein," *Mol. Cell*, vol. 63, no. 1, pp. 72-85, Jul 2016
13. J. A. Greig, T. A. Nguyen, M. Lee, A. S. Holehouse, A. E. Posey, R. V. Pappu, et al., "Arginine-Enriched Mixed-Charged Domains Provide Cohesion for Nuclear Speckle Condensation," *Mol. Cell*, vol. 77, no. 6, pp. 1237-1250, Mar 2020
14. I. Becher, A. Andrés-Pons, N. Romanov, F. Stein, M. Schramm, F. Baudin, et al., "Pervasive protein thermal stability variation during the cell cycle," *Cell*, vol. 173, pp. 1495-1507, May 2018
15. R. Narayanaswamy, M. Levy, M. Tsechansky, G. M. Stovall, J. D. O'Connell, J. Mirrielees, et al., "Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, no. 25, pp. 10147-10152, Jun 2019
16. S. Maharana, J. Wang, D. K. Papadopoulos, D. Richter, A. Pozniakovsky, I. Poser, et al., "RNA buffers the phase separation behavior of prion-like RNA binding proteins," *Science*, vol. 360, pp. 918-921, Apr 2018
17. M. C. Munder, D. Midtvedt, T. Franzmann, E. Nüske, O. Otto, M. Herbig, et al., "A pH-driver transition of the cytoplasm from a liquid- to a solid-like state promotes entry into dormancy," *eLife*, vol. 5, e09347, Mar 2016, doi:10.7554/eLife.09347
18. E. W. J. Wallace, J. L. Kear-Scott, E. V. Pilipenko, M. H. Schwartz, P. R. Laskowski, A. E. Rojek, et al., "Reversible, specific, active aggregates of endogenous proteins assemble upon heat stress," *Cell*, vol. 162, pp. 1286–1298, Sep 2015
19. B. A. Gibson, L. K. Doolittle, M. W. G. Schneider, D. W. Gerlich, S. Redding, M. K. Rosen, "Organization of Chromatin by Intrinsic and Regulated Phase Separation," *Cell*, vol. 179, no. 2, pp. 470-484, Oct 2019
20. H. Strickfaden, T. O. Tolsma, A. Sharma, D. A. Underhill, J. C. Hansen, M. J. Hendzel, "Condensed Chromatin Behaves like a Solid on the Mesoscale *In Vitro* and in Living Cells," *Cell*, vol. 183, no. 7, pp. 1772-1784, Dec 2020
21. G. Zhang, T. A. Neubert, B. A. Jordan, "RNA Binding Proteins Accumulate at the Postsynaptic Density with Synaptic Activity," *J. Neurosci.*, vol. 32, no. 2, pp. 599-609, Jan 2012
22. S. Won, J. M. Levy, R. A. Nicoll, K. W. Roche, "MAGUKs: multifaceted synaptic organizers," *Curr. Opin. Neurobiol.*, vol. 43, pp. 94-101, Feb 2017

23. E. Kim, S. Naisbitt, Y. P. Hsueh, A. Rao, A. Rothschild, A. M. Craig, M. Sheng, “GKAP, a novel synaptic protein that interacts with the guanylate kinase-like domain of the PSD-95/SAP90 family of channel clustering molecules,” *J. Cell Biol.*, vol. 136, no. 3, pp. 669-78, Feb 1997
24. P. Monteiro, G. Feng, “SHANK proteins: roles at the synapse and in autism spectrum disorder,” *Nat. Rev. Neurosci.*, vol. 18, no. 3, pp. 147-157, Mar 2017
25. P. R. Brakeman, A. A. Lanahan, R. O’Brien, K. Roche, C. A. Barnes, R. L. Huganir, et al., “Homer: a protein that selectively binds metabotropic glutamate receptors,” *Nature*, vol. 386, no. 6622, pp. 284-288, Mar 1997
26. D. Cheng, C. C. Hoogenraad, J. Rush, E. Ramm, M. A. Schlager, D. M. Duong, et al., “Relative and absolute quantification of postsynaptic density proteome isolated from rat forebrain and cerebellum,” *Mol. Cell. Proteomics*, vol. 5, no. 6, pp. 1158-70, Jun 2006
27. G. J. Iacobucci, G. K. Popesku, “NMDA receptors: linking physiological output to biophysical operation,” *Nat. Rev. Neurosci.*, vol. 18, no. 4, pp. 236-249, Mar 2017
28. G. H. Diering, R. S. Nirujogi, R. H. Roth, P. F. Worley, A. Pandey, R. L. Huganir, “Homer1a drives homeostatic scaling-down of excitatory synapses during sleep,” *Science*, vol. 355, no. 6324, pp. 511-515, Feb 2017
29. M. Zeng, Y. Shang, Y. Araki, T. Guo, R. L. Huganir, M. Zhang, “Phase Transition in Postsynaptic Densities Underlies Formation of Synaptic Complexes and Synaptic Plasticity,” *Cell*, vol. 166, no. 5, pp. 1163-1175, Aug 2016
30. M. Zeng, X. Chen, D. Guan, J. Xu, H. Wu, P. Tong, et al., “Reconstituted Postsynaptic Density as a Molecular Platform for Understanding Synapse Formation and Plasticity,” *Cell*, vol. 174, no. 5, pp. 1172-1187, Aug 2018
31. E. Moutin, F. Raynaud, L. Fagni, J. Perroy, “GKAP-DLC2 interaction organizes the postsynaptic scaffold complex to enhance synaptic NMDA receptor activity,” *J. Cell Sci.*, vol. 125, no. 8, pp. 2030–2040, Apr 2012
32. S. A. Clark, N. Jespersen, C. Woodward, E. Barbar, “Multivalent IDP assemblies: Unique properties of LC8-associated, IDP duplex scaffolds,” *FEBS Lett.*, vol. 589, no. 19, pp. 2543–2551, Jul 2015
33. E. Moutin, V. Compan, F. Raynaud, C. Clerté, N Bouquier, G. Labesse, et al., “The stoichiometry of scaffold complexes in living neurons - DLC2 functions as a dimerization engine for GKAP,” *J. Cell Sci.*, vol. 127, no. 16, pp. 3451–3462, Aug 2014
34. E. Nagy-Kanta, Z. E. Kálmán, H. Tossavainen, T. Juhász, F. Farkas, J. Hegedüs, et al., “Residual flexibility in the topologically constrained multivalent complex

between the GKAP scaffold and LC8 hub proteins," *Preprint*, doi:10.1101/2024.11.25.624264

35. S. Varga, J. M. Kaasen, Z. Gáspári, B. F. Péterfia, F. A. A. Mulder, "Resonance assignment of the intrinsically disordered actin-binding region of Drebrin," *Biomol. NMR Assign.*, Jun 2025, doi:10.1007/s12104-025-10239-0
36. X. Zhang, H. Li, Y. Ma, D. Zhong, S. Hou, "Study liquid-liquid phase separation with optical microscopy: A methodology review," *APL Bioeng.*, vol. 7, no. 2, 021502, May 2023
37. A. R. Titus, P. P. Madeira, L. A. Ferreira, V. Y. Chernyak, V. N. Uversky, B. Y. Zaslavsky, "Mechanism of Phase Separation in Aqueous Two-Phase Systems," *Int. J. Mol. Sci.*, vol. 23, no. 22, 14366, Nov 2022
38. P. Arosio, T. Müller, L. Rajah, E. V. Yates, F. A. Aprile, Y. Zhang, et al., "Microfluidic Diffusion Analysis of the Sizes and Interactions of Proteins under Native Solution Conditions," *ACS Nano*, vol. 10, no. 1, pp. 333-341, Dec 2016
39. H. Gang, C. Galvagnion, G. Meisl, T. Müller, M. Pfammatter, A. K. Buell, et al., "Microfluidic Diffusion Platform for Characterizing the Sizes of Lipid Vesicles and the Thermodynamics of Protein–Lipid Interactions," *Anal. Chem.*, vol. 90, no. 5, pp. 3284–3290, Jan 2018
40. D. Süveges, Z. Gáspári, G. Tóth, L. Nyitrai, "Charged single alpha-helix: a versatile protein structural motif," *Proteins*, vol. 74, no. 4, pp. 905-916, Mar 2009
41. C. A. Barnes, Y. Shen, J. Ying, Y. Takagi, D. A. Torchia, J. R. Sellers, et al., "Remarkable Rigidity of the Single ɑ-Helical Domain of Myosin-VI As Revealed by NMR Spectroscopy," *J. Am. Chem. Soc.*, vol. 141, no. 22, pp. 9004-9017, Jun 2019
42. M. Wolny, M. Batchelor, G. J. Bartlett, E. G. Baker, M. Kurzawa, P. J. Knight, et al., "Characterization of long and stable *de novo* single alpha-helix domains provides novel insight into their stability," *Sci. Rep.*, vol. 7, 44341, Mar 2017
43. M. Peckham, P. J. Knight, "When a predicted coiled coil is really a single α-helix, in myosins and other proteins," *Soft Matter*, vol. 5, pp. 2493-2503, Mar 2009
44. B. Mészáros, G. Erdős, B. Szabó, É. Schád, Á. Tantos, R. Abukhairan, et al., "PhaSePro: the database of proteins driving liquid–liquid phase separation," *Nucleic Acids Res.*, vol. 48, D360-D367, Oct 2020
45. W. Ning, Y. Guo, S. Lin, B. Mei, Y. Wu, P. Jiang, et al., "DrLLPS: a data resource of liquid–liquid phase separation in eukaryotes," *Nucleic Acids Res.*, vol. 48, D288-D295, Jan 2020

46. K. You, Q. Huang, C. Yu, B. Shen, C. Sevilla, M. Shi, et al., “PhaSepDB: a database of liquid–liquid phase separation related proteins,” *Nucleic Acids Res.*, vol. 48, D354-D359, Oct 2020

47. N. Rostam, S. Ghosh, C. F. W. Chow, A. Hadarovich, C. Landerer, R. Ghosh, et al., “CD-CODE: crowdsourcing condensate database and encyclopedia,” *Nat. Methods*, vol. 20, pp. 673-676, Apr 2023

48. Q. Li, X. Peng, Y. Li, W. Tang, J. Zhu, J. Huang, et al., “LLPSDB: a database of proteins undergoing liquid–liquid phase separation in vitro,” *Nucleic Acids Res.*, vol. 48, D320-D327, Jan 2020

49. C. Nunes, I. Mestre, A. Marcelo, R. Koppenol, C. A. Matos, C. Nóbrega, “MSGP: the first database of the protein components of the mammalian stress granules,” *Database*, vol. 2019, baz031, Jan 2019, doi:10.1093/database/baz031

50. Z. E. Kálmán, D. Dudola, B. Mészáros, Z. Gáspári, L. Dobson, “PSINDB: The postsynaptic protein-protein interaction database,” *Database*, vol. 2022, baac007, Mar 2022, doi:10.1093/database/baac007

51. A. K. Lancaster, A. Nutter-Upham, S. Lindquist, O.D. King, “PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition,” *Bioinf.*, vol. 30, no. 17, pp. 2501–2502, May 2014

52. A. Hatos, S. C. E. Tosatto, M. Vendruscolo, M. Fuxreiter, “FuzDrop on AlphaFold: visualizing the sequence-dependent propensity of liquid–liquid phase separation and aggregation of proteins,” *Nucleic Acids Res.*, vol. 50, W337–W344, Jul 2022

53. G. van Mierlo, J. R. G. Jansen, J. Wang, I. Poser, S. J. van Heeringen, M. Vermeulen, “Predicting protein condensate formation using machine learning,” *Cell Rep.*, vol. 34, no. 5, 108705, Feb 2021

54. X. Chu, T. Sun, Q. Li, Y. Xu, Z. Zhang, L. Lai, et al., “Prediction of liquid–liquid phase separating proteins using machine learning,” *BMC Bioinf.*, vol. 23, article number: 72, Feb 2022

55. A. Hadarovich, H. R. Singh, S. Ghosh, N. Rostam, A. A. Hyman, Á. Tóth-Petróczy, “PICNIC accurately predicts condensate-forming proteins regardless of their structural disorder across organisms,” *Nat. Commun.*, vol. 15, article number: 10668, Dec 2024

56. J. A. Joseph, A. Reinhardt, A. Aguirre, P. Y. Chew, K. O. Russell, J. R. Espinosa, A. Garaizar, R. Collepardo-Guevara, “Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy,” *Nat. Comput. Sci.*, vol. 1, no. 11, pp. 732-743, Nov 2021

57. H. Luo, H. Nijveen, “Understanding and identifying amino acid repeats,” *Brief Bioinform.*, vol. 15, no. 4, pp. 582-591, Jul 2014
58. Z. Gáspári, D. Süveges, A. Perczel, L. Nyitray, G. Tóth, “Charged single alpha-helices in proteomes revealed by a consensus prediction approach,” *BBA*, vol. 1824, no. 4, pp. 637-646, Apr 2012
59. Á. Kovács, D. Dudola, L. Nyitray, G. Tóth, Z. Nagy, Z. Gáspári, “Detection of single alpha-helices in large protein sequence sets using hardware acceleration,” *J. Struct. Biol.*, vol. 204, pp. 109-116, Oct 2018
60. D. Dudola, G. Tóth, L. Nyitray, T. Gáspári, “Consensus prediction of charged single alpha-helices with CSAHserver,” *Methods Mol. Biol.*, vol. 1484, pp. 25-34, Oct 2016
61. H. Mi, P. Thomas, “PANTHER Pathway: an ontology-based pathway database coupled with data analysis tools,” *Methods Mol. Biol.*, vol. 563, pp. 123-140, May 2009
62. J. Martínez, Y. G. Ren, A. C. Thuresson, U. Hellman, J. Astrom, A. Virtanen, “A 54-kDa Fragment of the Poly(A)-specific Ribonuclease Is an Oligomeric, Processive, and Cap-interacting Poly(A)-specific 3’ Exonuclease,” *RNA*, vol. 275, no. 31, P24222-P24230, Aug 2000
63. J. H. Kim, D. N. Shinde, M. R. F. Reijnders, N. S. Hauser, R. L. Belmonte, G. R. Wilson, et al., “De Novo Mutations in SON Disrupt RNA Splicing of Genes Essential for Brain Development and Metabolism, Causing an Intellectual-Disability Syndrome,” *Am. J. Hum. Genet.*, vol. 99, no. 3, pp. 711-719, Sep 2016
64. F. Heyd, K. W. Lynch, “Phosphorylation-dependent regulation of PSF by GSK3 controls CD45 alternative splicing,” *Mol. Cell.*, vol. 40, no. 1, pp. 126-137, Oct 2010
65. M. Klar, J. Bode, “Enhanceosome formation over the beta interferon promoter underlies a remote-control mechanism mediated by YY1 and YY2,” *Mol. Cell Biol.*, vol. 25, no. 22, pp. 10159-10170, Nov 2005
66. S. Wang, B. Zhang, D. V. Faller, “Prohibitin requires Brg-1 and Brm for the repression of E2F and cell growth,” *EMBO J.*, vol. 21, no. 12, pp. 3019-3028, Jun 2002
67. A. S. Y. Lee, P. J. Kranzusch, J. H. D. Cate, “eIF3 targets cell-proliferation messenger RNAs for translational activation or repression,” *Nature*, vol. 522, pp. 111-114, Jun 2015
68. P. Gaudet, M. S. Livstone, S. E. Lewis, P. D. Thomas, “Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium,” *Brief Bioinform.*, vol. 12, no. 5, pp. 449-462, Sep 2011

69. R. Pardo, M. Molina-Calavita, G. Poizat, G. Keryer, S. Humbert, F. Saudou, “pARIS-htt: an optimised expression platform to study huntingtin reveals functional domains required for vesicular trafficking,” *Mol. Brain*, vol. 3, 17, Jun 2010

70. J. Liang, J. Wang, A. Azfer, W. Song, G. Tromp, P. E. Kolattukudy, et al., “A Novel CCCH-Zinc Finger Protein Family Regulates Proinflammatory Activation of Macrophages,” *J. Biol. Chem.*, vol. 283, no. 10, pp. 6337-6346, Mar 2008

71. K. Yamada, H. Kawata, K. Matsuura, Z. Shou, S. Hirano, T. Mizutani, et al., “Functional analysis and the molecular dissection of zinc-fingers and homeoboxes 1 (ZHX1),” *Biochem. Biophys. Res. Commun.*, vol. 297, no. 2, pp. 368-374, Sep 2002

72. E. D. Egan, K. Collins, “An Enhanced H/ACA RNP Assembly Mechanism for Human Telomerase RNA,” *Mol. Cell. Biol.*, vol. 32, no. 13, pp. 2428-2439, Jul 2012

73. M. Okuwaki, M. Tsujimoto, K. Nagata, “The RNA Binding Activity of a Ribosome Biogenesis Factor, Nucleophosmin/B23, is Modulated by Phosphorylation with a Cell Cycle-dependent Kinase and by Association with Its Subtype,” *Mol. Biol. Cell*, vol. 13, no. 6, pp. 2016-2030, Jun 2002

74. A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B. M. Beckmann, C. Strein, et al., “Insight into RNA Biology from an Atlas of Mammalian mRNA-binding Proteins,” *Cell*, vol. 149, no. 6, pp. 1393-1406, Jun 2012

75. A. G. Baltz, M. Munschauer, B. Schwanhäusser, A. Vasile, Y. Murakawa, M. Schueler, et al., “The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts,” *Mol. Cell*, vol. 46, no. 5, pp. 674-690, Jun 2012

76. E. Y. Ahn, R. C. DeKelver, M. C. Lo, T. A. Nguyen, S. Matsuura, A. Boyapati, et al., “SON Controls Cell-Cycle Progression by Coordinated Regulation of RNA Splicing,” *Mol. Cell*, vol. 42, no. 2, pp. 185-198, Apr 2011

77. Y. Jeon, J. T. Lee, “YY1 Tethers Xist RNA to the Inactive X Nucleation Center,” *Cell*, vol. 146, no. 1, pp. 119-133, Jul 2011

78. B. Pereira, S. Sousa, R. Barros, L. Carreto, P. Oliveira, C. Oliveira, et al., “CDX2 regulation by the RNA-binding protein MEX3A: impact on intestinal differentiation and stemness,” *Nucleic Acids Res.*, vol. 41, no. 7, pp. 3986-3999, Apr 2013

79. J. Fritz, A. Strehblow, A. Taschner, S. Schopoff, P. Pasierbek, M. F. Jantsch, “RNA-Regulated Interaction of Transportin-1 and Exportin-5 with the Double-Stranded RNA-Binding Domain Regulates Nucleocytoplasmic Shuttling of ADAR1,” *Mol. Cell. Biol.*, vol. 29, no. 6, pp. 1487-1497, Mar 2009

80. Z. Zhang, D. Theler, K. H. Kaminska, M. Hiller, P. de la Grange, R. Pudimat, et al., “The YTH Domain Is a Novel RNA Binding Domain,” *J. Biol. Chem.*, vol. 285, no. 19, pp. 14701-14710, May 2010
81. D. A. Bochar, L. Wang, H. Beniya, A. Kinev, Y. Xue, W. S. Lane, et al., “BRCA1 Is Associated with a Human SWI/SNF-Related Complex,” *Cell*, vol. 102, no. 2, pp. 257-265, Jul 2000
82. M. B. Sewer, V. Q. Nguyen, C. J. Huang, P. W. Tucker, N. Kagawa, M. R. Waterman, “Transcriptional Activation of Human CYP17 in H295R Adrenocortical Cells Depends on Complex Formation among p54(nrb)/NonO, Protein-Associated Splicing Factor, and SF-1, a Complex That Also Participates in Repression of Transcription,” *Endocrinology*, vol. 143, no. 4, pp. 1280-1290, Apr 2002
83. D. M. Passon, M. Lee, O. Rackham, W. A. Stanley, A. Sadowska, A. Filipovska, et al., “Structure of heterodimer of human NONO and paraspeckle protein component 1 and analysis of its role of subnuclear body formation,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 13, pp. 4846-4850, Mar 2012
84. L. Dobson, L. Nyitray, Z. Gáspári, “A conserved charged single $\alpha$-helix with a putative steric role in paraspeckle formation,” *RNA*, vol. 21, pp. 2023-2029, Dec 2015
85. J. T. Zipprich, S. Bhattacharyya, H. Mathys, W. Filipowicz, “Importance of the C-terminal domain of the human GW182 protein TNRC6C for translational repression,” *RNA*, vol. 15, no. 5, pp. 781-793, May 2009
86. F. Mattiroli, J. H. A. Vissers, W. J. van Dijk, P. Ikpa, E. Citterio, W. Vermeulen, et al., “RNF168 Ubiquitinates K13-15 on H2A/H2AX to Drive DNA Damage Signaling,” *Cell*, vol. 150, no. 6, pp. 1182-1195, Sep 2012
87. B. R. Sabari, A. Dall’Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, et al., “Coactivator condensation at super-enhancers links phase separation and gene control,” *Science*, vol. 361, eaar3958, Jul 2018, doi:10.1126/science.aar3958
88. C. G. Triandafillou, R. W. Pan, A. R. Dinner, D. A. Drummond, “Pervasive, conserved secondary structure in highly charged protein regions,” *Preprint*, doi: 10.1101/2023.02.15.528637
89. D. Tariq, N. Maurici, B. M. Bartholomai, S. Chandrasekaran, J. C. Dunlap, A. Bah, et al., “Phosphorylation, disorder, and phase separation govern the behavior of Frequency in the fungal circadian clock,” *eLife*, vol. 12, RP90259, Mar 2024, doi:10.7554/eLife.90259.3

90. T. P. López-Palacios, J. L. Andersen, “Kinase regulation by liquid-liquid phase separation,” *Trends Cell Biol.*, vol. 33, no. 8, pp. 649-666, Aug 2023

91. N. Galvanetto, M. T. Ivanović, A. Chowdhury, A. Sottini, M. F. Nüesch, D. Nettels, et al., “Extreme dynamics in a biomolecular condensate,” *Nature*, vol. 619, pp. 876-883, Jul 2023