PÁZMÁNY PÉTER CATHOLIC UNIVERSITY

ROSKA TAMÁS DOCTORAL SCHOOL
OF SCIENCES AND TECHNOLOGY

SZABÓ András László

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

Theses of PhD Dissertation

Thesis Supervisor:
GÁSPÁRI Zoltán, PhD

2025

Aims of the study

The main goal of this study is to investigate whether GKAP, LC8, and their complexes, as well as associated motifs such as single α-helices contribute to the phase separation phenomena that shape postsynaptic densities. Testing the latter demands the proteome-scale investigation of charged sequence motifs, which involves:

- Compiling a proteome-scale dataset including charged sequence motifs and regions responsible for phase separation.
- Minimizing redundancy.
- Assessment of plausible associations between the investigated motifs and protein phase separation.

The complex formation of GKAP and LC8 could be examined via measuring the size of solute particles, which can be achieved through a diffusion-based approach combining fluorescent microscopy with microfluidics. Developing such a method requires the following steps:

- Designing the experimental setup.
- Fabricating microfluidic devices with appropriate layouts.
- Preparing fluorescent samples for calibration and analysis.
- Developing the analytic software that evaluates the data.

Methods

Compiling the proteome-scale dataset

The human reference proteome used for this study encompassed 20659 human genes with one isoform per gene, gathered from the manually curated SwissProt database. Complementing the dataset with motifs that had experimental evidence of contributing to phase separation was carried out by integrating regions annotated as such in human PhaSepDB entries. Charged residue repeats (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. SAHs were highlighted among these motifs by their characteristic frequency of 1/9 to 1/6. The dataset was further expanded with 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.

Minimizing redundancy

The complete reference proteome was clustered with CD-HIT based on sequence similarity, with a threshold of 0.9, 0.7, and 0.5, respectively. All sequences were assigned to the best cluster that met the threshold.

Assessment of plausible association between sequence motifs and phase separation

The elements of the reference proteome as well as its clustered variants were organized into 2x2 contingency tables based on two variables. One of the variables was the presence – or absence – of charged sequence motifs. The other one was propensity towards phase separation, indicated by the presence of regions annotated as such in PhaSepDB. The degree of independence between these variables was determined by the P-values yielded by Fisher's exact test of independence carried out on the contingency tables. The presence of charged sequence motifs was also evaluated as a predictive indicator for a protein's likelihood of participating in phase separation via receiver operating characteristic (ROC) tests. To this end, the

reference proteome was sorted according to the scores obtained from motif detections, after which true positive, true negative, false positive, and false negative rates were calculated based on association with phase separation. True positive entries were further investigated in case studies.

Designing the experimental setup

GKAP, LC8, and their complex were distinguished by their hydrodynamic radii according to the Stokes-Einstein equation that only applies at low Reynolds numbers, characteristic of laminar flow:

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

Particles were positioned into a well-defined initial state by microfluidic focusers with three inlets that compressed the analyte into the middle with two buffer streams from the sides. From this position they continued via laminar flow through a flat channel recorded with an inverted fluorescent microscope, where they only moved laterally via diffusion. Their diffusion coefficients were approximated as the incline of the linear function determined by their distribution at specific points along the channel:

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

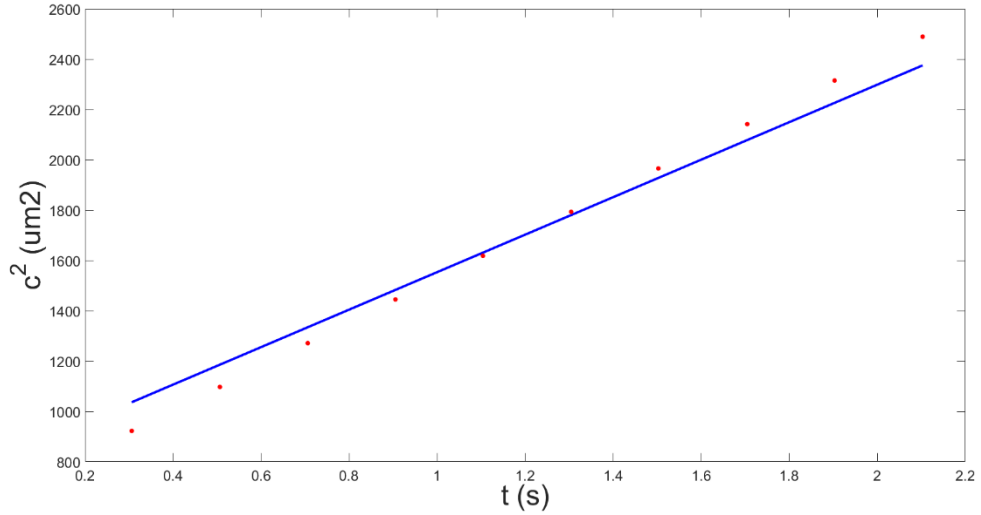

Fig. 1. Distribution of solute particles, approximated by the variance of Gaussian functions (c^2) fitted to fluorescent intensity profiles measured at different points along the microfluidic device. The x-axis shows the time (t) particles spent between measurement points. The analyte was Enhanced Green Fluorescent Protein (EGFP). [2, Fig. 1.]

Device fabrication and sample preparation

Mária Laki fabricated the microfluidic devices the layouts of which were developed iteratively based on feedback from testing. All measurements involved treating microfluidic devices with bovine serum albumin (BSA) to delay the accumulation of particles on their surfaces. Edit Andrea Jáger prepared the samples of enhanced green fluorescent protein (EGFP) and fluorescent microspheres. Eszter Nagy-Kanta expressed, purified, and labelled the samples of GKAP, LC8, and their hexameric complex. Protein samples were labelled with fluorescein (FITC) that was selected after testing various other fluorescent dyes. The approach was also tested on a FITC-labelled construct of Drebrin, provided by Soma Varga.

Developing the analytic software

The measured experimental data was processed by an analytic software package written in MATLAB and using Excel files as input, which contained pairs of brightfield and fluorescent intensity profiles, their exact position along the microfluidic device, and general conditions such as absolute temperature and exposition time. Time spent by particles between measurement points was approximated through their mean velocity, determined by flow rates. Regular artefacts were automatically removed from the fluorescent intensity profiles that were then normalized, aligned, and fitted with the linear combination of two Gaussian functions. Each of these functions were assigned to one of two sets containing one function per measurement point. The variances of Gaussian functions of the same set were displayed with their corresponding time components, as shown in Fig. 1. The linear function fitted to these points directly correlated with the diffusion coefficient of a solute particle.

Results

During my PhD studies I have conducted complex research projects that aimed to investigate different aspects of protein phase separation, specifically in the context of postsynaptic densities. These projects led to various results and conclusions, the most significant ones of which are summarized in the following 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. [1]

a) **Methods:** I compiled a reference proteome from the human protein entries of SwissProt. I complemented this dataset with the positions of sequence motifs that had experimental evidence of contributing to phase separation, as annotated in PhaSepDB2.0. I expanded the dataset with charged residue repeats identified with

the FT_CHARGE algorithm, highlighting motifs that had qualified as single α-helices based on the characteristic frequency of their charge patterns. I further expanded the dataset with charge-dense regions (CDRs), identified with an algorithm of my own making, utilizing windowing functions and two different scoring schemes, yielding two different types of CDRs. I minimized the redundancy of the complete reference proteome using CD-HIT (Table 1.).

b) **Results:** Assessment of charged sequence motifs within the human reference proteome revealed that a significant portion of proteins contains at least one CDR. Proteins with transmembrane segments have been excluded from further investigations, as any strong association to phase separation had been expected to be characteristic of soluble proteins.

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.

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. [1]

a) **Methods:** I converted the reference proteome as well as its redundancy-filtered variants into two-by-two contingency tables upon which I implemented Fisher's exact test of independence, determining whether a protein's participation in phase separation and the presence of a specific type of charged motif within its sequence are independent variables. These investigations were repeated on random datasets constructed to match the size distribution and residue composition of the reference proteome. I also conducted receiver operating characteristic (ROC) tests to evaluate the presence of a specific type of charged sequence motif as an indicator of its host protein's participation in phase separation.

b) **Results:** 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. 2.). The random datasets yielded matching results, proving the associations between motifs and the phenomenon to be robust.

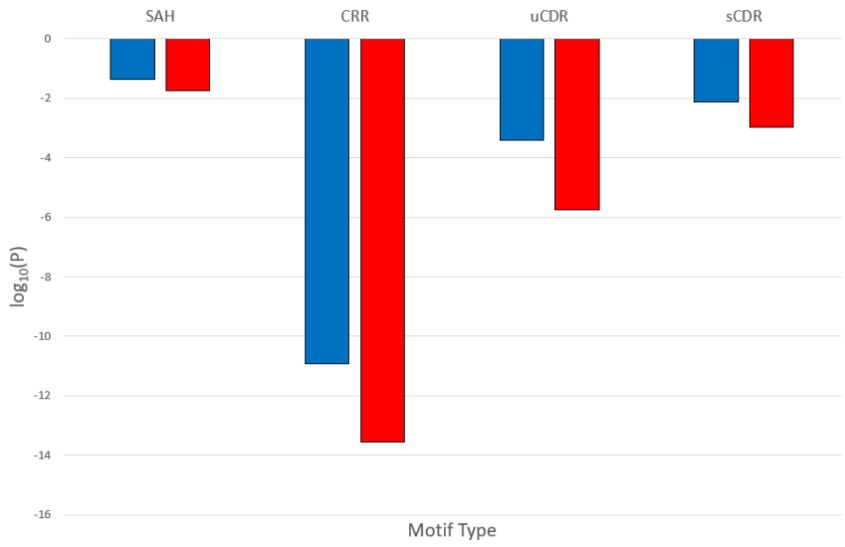

Fig. 2. The common logarithm of P-values from Fisher's exact test of independence assessing the correlation between human

proteins' propensity towards phase separation and the presence of specific charged sequence motifs within their sequence. This assessment has been repeated with the exclusion of transmembrane proteins (blue). [1, Fig. 3.]

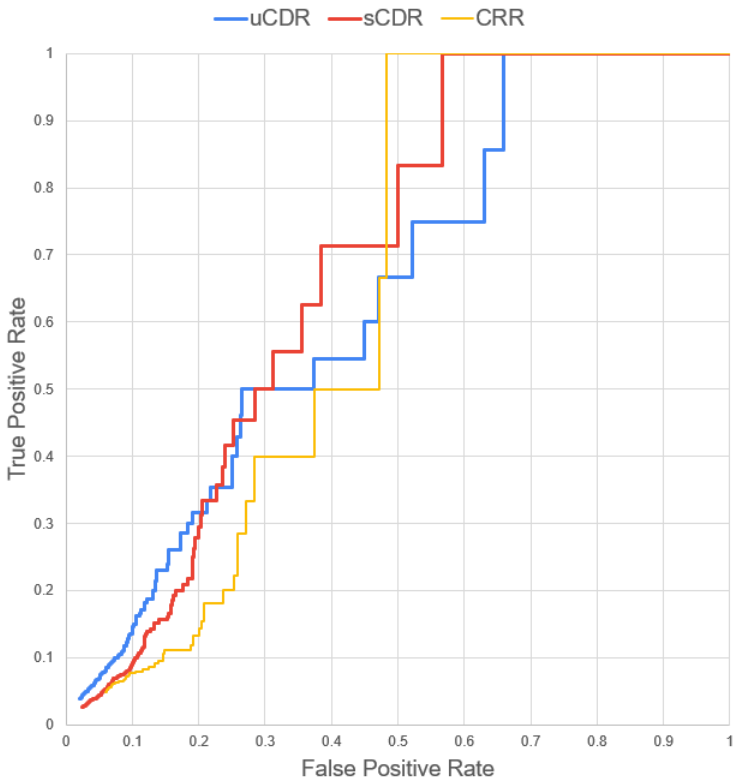

Fig. 3. Receiver operating characteristic (ROC) analysis of the predictive capabilities of different sequence motifs towards phase separation. [1, Fig. 6.]

While the presence of charged sequence motifs alone was proven to be a weak predictor for a protein's participation in phase separation (Fig. 3.), 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.

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. [2]

Method development: I determined that fluorescein is a fluorescent dye that suitably labels constructs of GKAP, LC8, and Drebrin without disrupting the complex formation of the former two.

I participated in designing a microfluidic device that focuses labelled protein samples into the middle of a channel where laminar flow is maintained thereafter. I optimized the recording of signals and developed an analytic software capable of converting measured data into approximated hydrodynamic radii of solute particles.

I discerned that three design units is the largest size a device can be ands still fit onto a single glass slide. Therefore, the main channel of the finalized design was 54 mm long, since turns would disrupt the laminar flow (Fig. 4.).

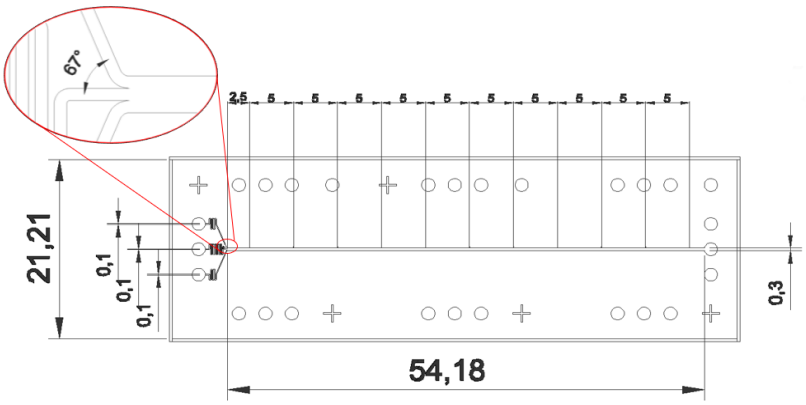

Fig. 4. The final layout of the microfluidic device features equidistant markers, which guided the recording of fluorescent and brightfield signals along the channel at predetermined measurement points. [2, Fig. 7.]

I also determined that the internal surface of the device must be treated with *bovine serum albumin* (BSA) to delay the accumulation of fluorescent particles. Without BSA treatment enough fluorescent particles would accumulate on the surface of the device in 10 minutes to elevate the signal level beyond the upper limit of the microscope (see Fig. 5.). At least 10 minutes must pass after the analyte reached the device before the system reaches its steady state.

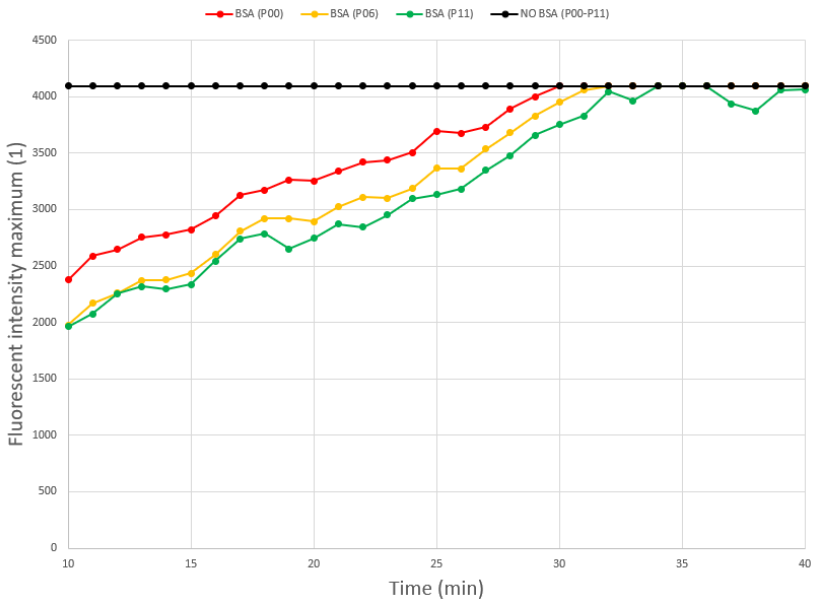

Fig. 5. Peak fluorescent intensity at the intersection of the microfluidic device (P00), about halfway down its main channel (P06), and just before its outlet (P11). Intensity was measured in every minute for half an hour in a microfluidic device treated with 1% BSA, and in another one without treatment. 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. [2, Fig. 8.]

Additionally, the analytic software was set to approximate measured fluorescent intensity profiles with the linear combination of two Gaussian functions, since all “goodness of fit” metrics indicated that the resulting curves are significantly better than those consisting of a single Gaussian function (Fig. 6.).

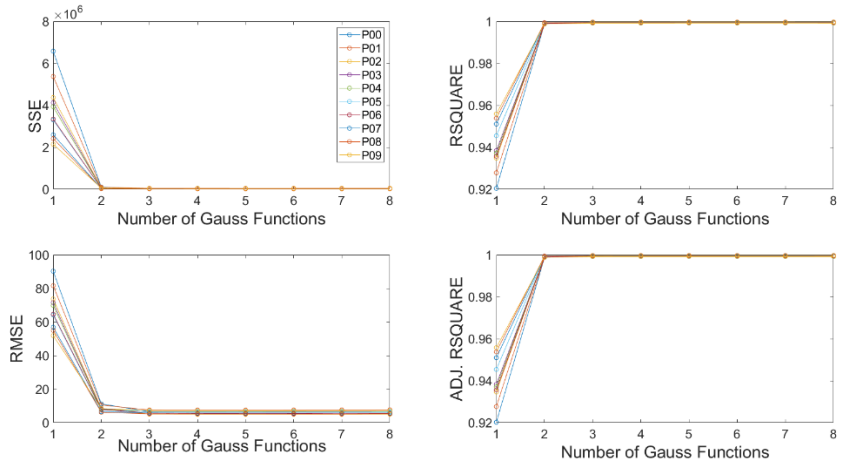

Fig. 6. Goodness of fit (GoF) metrics, calculated for different measurement points (P00-P09) and eight levels of complexity. This measurement used EGFP, but other analytes also exhibited the same trend in all four GoF metrics. [2, Fig. 9.]

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. [2]

a) **Methods:** I carried out measurements with the finalized approach on fluorescein-labelled Drebrin (D233) and GKAP constructs (GKAP-PBM and GKAP-DLC2), GKAP-DLC2 + LC8 complexes where GKAP-DLC2 monomers were labelled beforehand, EGFPs, and samples of fluorescent microspheres 50 nm, 200 nm, and 1100 nm in diameter, respectively.

b) **Results:** The mean values of approximated radii remain consistent between approaches for particles under 10 nm in diameter. At this scale, the standard deviation (STD) is lower in the case of the diffusion-based approach (Table 2.). For larger particles, the diffusion-based approach falls off both in accuracy and precision.

Analyte	Expected radius (nm)	Approx. radius (nm)
GKAP-PBM	1.39	1.16 ± 0.10
GKAP-DLC2	1.27 ± 0.72 1.90 ± 1.00	1.36 ± 0.30
GKAP-DLC2 + LC8	2.17 ± 0.44 3.11 ± 0.76	1.64 ± 0.36
D233	1.81	1.45 ± 0.39
EGFP	2.72 ± 1.01	1.72 ± 0.42
50 nm MS	22.68 ± 2.11	9.28 ± 11.09
200 nm MS	97.09 ± 9.51	15.15 ± 13.36
1100 nm MS	461.61 ± 35.59	6.62 ± 5.78

Table 2. 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 the GKAP-DLC2 + LC8 complex, which are included separately. The mean values and STDs for approximate radii were calculated from multiple diffusion measurements.

My results advance our knowledge of protein phase separation and postsynaptic densities, while providing additional *in silico* and *in vitro* tools

and resources for researchers of these scientific fields. Several robust associations have been identified between charged sequence motifs and protein phase separation. Additionally, a reliable experimental approach has been developed that can monitor PSD proteins and their complexes, providing some insight into the interaction network that possibly initiates phase separation within these cellular components.

Publications

1. 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
2. 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
3. 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

Conference presentations

4. 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)
5. 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)

6. 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].
7. 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)

Contributions

The research condensed into the articles above involved the hard work and invaluable insight of multiple researchers, for which I am immensely grateful. 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.