Automated navigation of condensate phase behavior with active machine learning

Y.H.A. Leurs[+], W. van den Hout[+], A. Gardin[+], J.L.J. van Dongen, J.C.M. van Hest*, F. Grisoni*, L. Brunsveld*

Institute for Complex Molecular Systems (ICMS), Department of Biomedical Engineering, Eindhoven University of technology, PO Box 513, 5600 MB Eindhoven, The Netherlands.

[+]These authors contributed equally.

*j.c.m.v.hest@tue.nl, f.grisoni@tue.nl, l.brunsveld@tue.nl

Abstract

Biomolecular condensates are essential functional cellular structures that form through phase separation of macromolecules such as proteins and RNA. Synthetic condensates have recently gathered great interest as they can be engineered to better understand the formation mechanism of these cellular condensates and serve as cell-mimetic platforms to develop novel therapeutic strategies. The complexity of the biomolecular components and their reciprocal interactions, however, makes precise engineering and systematic characterization of condensate formation a challenging endeavor. While constructing phase diagrams is a systematic approach to gain comprehensive insight into phase separation behavior, it is a time-consuming and labor-intensive process. Here, we present an automated platform for efficiently mapping multi-dimensional phase diagrams of condensates. The automated platform incorporates a pipetting system for sample formulation, and an autonomous confocal microscope for particle property analysis and characterization. Active machine learning – which allows iterative model improvement – is used to learn from previous experiments and steer future experiments towards an efficient exploration of phase boundaries. The versatility of the pipeline is demonstrated by showcasing its ability to rapidly explore the phase behavior of various polypeptides of opposite charge across formulations, producing detailed and reproducible multidimensional phase diagrams. Beyond identifying phase boundaries, the platform also provides information-rich data, enabling quantification of key condensate properties such as particle size, count, and volume fraction – adding functional insights to phase diagrams. This self-driven platform is robust and generalizable, allowing easy extension to any given combination of condensate-forming materials, ultimately providing key insights into their formation and characteristics.

Introduction

Organization and compartmentalization are fundamental aspects of nature.[1] The spatial arrangement of biomolecules is essential for maintaining cellular function and facilitating metabolic processes, such as molecular transport, energy production, and structural support.[2,3] In this respect, biomolecular condensates have gained significant interest in recent years, as these membrane-less organelles play essential roles in compartmentalization and may contribute to the emergence of cellular complexity.[4,5] Condensates are phase-separated, micron-sized subcellular droplets that are formed through multivalent interactions between (macro)molecules, such as proteins and nucleic acids.[6] Their dynamic formation mechanism and complex biochemistry have become topic of intensive investigation, in particular in the field of molecular and cell biology.[7,8] An alternative approach to provide valuable insight into these structures is to engineer synthetic condensates *in vitro* – outside of the cellular environment.[9–14] This allows a more systematic tuning and study of the physicochemical properties of condensates[15] and also enables the development of self-assembled and/or cell-mimetic platforms that can be used for the exploration of novel therapeutic strategies.[16–22] Although synthetic condensates circumvent the need to take the cell's complexity into account, still significant challenges remain in terms of predicting condensate formation and properties based on the molecular structures and elucidating the effects of environmental factors, such as pH and ionic strength, on condensate formation and properties (as well as the underlying molecular mechanisms).[23,24]

Manually navigating the vast combinatorial space, spanning diverse molecular structures and environmental factors, to investigate coacervate formation (*e.g.*, via phase diagrams) quickly becomes unfeasible as the number of variables to be considered increases. This process involves preparing hundreds to thousands of samples, each with precisely controlled conditions (*e.g.,* concentration, pH, and ionic strength), followed by detailed and consistent analysis of the phase separation parameters.[25–30] Furthermore, identifying phase boundaries by collecting data across a broad range of conditions without specific guidance (*e.g.,* based on intuition) is not only time-consuming and labor-intensive but also prone to human error, highlighting the need for automated, machine learning-driven, high-throughput methods.[31,32]

To address these challenges, recent innovations in high-throughput biochemical assays, microfluidics, and automated microscopy and analysis have enabled new methods to study biomolecular condensates under varied conditions.[33–36] Notwithstanding these advances, fully leveraging the vast datasets these techniques produce opens opportunities to explore condensate behavior more efficiently. Integrating machine learning, particularly active learning[37,38], into this field presents a valuable opportunity to enhance data-driven parameter

exploration, refine predictive models, and reduce the need for extensive experimental input. Active machine learning iteratively selects the most informative data points to analyze and to steer the next iteration of experiments,[39–42] which makes it particularly useful for automation by reducing the amount of data and experimentation needed to achieve accurate results.[39,43]

In this work, we introduce an automated, high-throughput platform designed to map multi-dimensional phase diagrams of biomolecular condensates. Our platform integrates active machine learning for phase mapping optimization, an automated pipetting system for sample formulation, and an autonomous confocal microscope for high-content imaging and detailed sample characterization. Using this platform, we extensively examine the phase behavior of two well-studied polypeptides across a range of formulations. Beyond reproducibly identifying phase boundaries, the platform also produces information-rich data on condensate properties such as particle size, particle count, and volume fraction – offering deep insight into condensate characteristics. To demonstrate the robustness of the approach, we construct higher-dimensional phase diagrams, allowing to uncover how multiple factors influence condensate formation. The automated platform not only accelerates and standardizes phase separation behavior mapping but also enhances our understanding of environmental parameter effects on condensate properties. We expect this approach to increase the application potential of synthetic condensates as a platform for the study of their natural analogues and to engineer self-assembled cell-mimetic platforms.

Results

Closed loop navigation of coacervate formation

Condensate formulation typically involves mixing complementary components, for example, of anionic and cationic nature[44], at specific speeds and durations, in a pH-controlled aqueous solution to form condensate, or more specifically, complex coacervate microdroplets (Figure 1A). This process can be laborious, error-prone, and time-consuming, limiting current capabilities to determine detailed phase diagrams and, correspondingly, the optimal conditions for condensate formation. To date, there is no standardized protocol for producing condensates, and scientists often adhere strictly to formulation techniques that work for their specific applications.[45] Here, we present a generalizable, closed-loop workflow that combines automation and machine learning to (a) standardize and speed up condensate preparation and reduce handling errors, (b) provide an automated characterization approach, and (c) navigate complex coacervate phase diagrams more efficiently, thanks to machine learning predictions. The workflow is based on the following mutually interacting components:

I. *Robotic sample production.* Efficient, accurate, and contamination-free sample preparation is critical for exploring vast experimental spaces with diverse conditions. Our platform addresses these needs with a cost-effective and versatile robotic pipetting platform (Figure 1B) that combines adaptable deck space, scalable reservoir options, and an open-source programming interface. These features enable high-throughput automation of condensate formulations in any pre-defined, multi-dimensional experimental space. Custom features (Figure 1C) allow increasing production rates through optimized liquid handling and prevent cross-contamination by using adaptable dispensing heights for contactless dispensing and different contact points for each liquid via a custom touch-tip functionality. Together, this also reduces plastic consumption, by allowing tip re-use for each distinct liquid.

II. *Automated particle characterization***.** High-throughput condensate analysis requires high-throughput imaging with sufficient spatial resolution and consistent focus, which can be challenging due to heterogeneous and varying sizes of condensate sizes. In our platform, samples are transferred to a 96-well microscopy plate and imaged using an automated confocal microscope (Figure 1D). This setup enables high-speed imaging and precise focus tracking through hardware autofocus. After formulation, condensates naturally settle over time on the glass surface, allowing for 3D reconstruction through dynamically acquired *Z*-stacks at four positions within each sample (Figure 1D). This approach generates technical replicates and accounts for potential inconsistencies. The automated image analysis pipeline involves (a) applying binary thresholding to detect particles (Figure 1E), (b) identifying the optimal *Z*-plane where each particle is in the best focal plane (Figure 1F),

and (c) classifying the sample as phase-separated when a threshold number of particles is observed (Figure 1G), or as non-phase-separated otherwise. Additionally, condensate properties, such as morphology and volume fraction, are extracted for follow-up analysis and characterization.

III. *Active machine learning***.** In our platform, the collected experimental data (Figure 1H) are used to train a Gaussian Process Classifier (GPC), a machine learning model that leverages Bayesian probability to make predictions, while accounting for uncertainty in classification decisions.[46] The model is trained to predict whether a pair of polypeptides at specific concentrations (optionally along with other experimental parameters) will phase-separate. The trained model is then used to predict the phase-separation behavior of the pre-defined experimental space (Figure 1I). Based on the predictions, new experimental points are requested for the next experimental iteration (Figure 1J). This is achieved via the exploration of areas in the phase diagram with high prediction uncertainty (in the form of information entropy, see Methods, Eq. 5), and via diversity-based sampling (via so-called farthest point sampling[47]). The selected points are then produced and characterized (via steps **I** and **II**) and contribute to the next phase of model training.

Thanks to this closed-loop make-analyze-predict cycle, the sample production (step **I**) and characterization (step **II**) produce data for the machine-learning-driven choice of the next experiments (step **III**) – this procedure is repeated until convergence. According to self-driving lab autonomy criteria, our pipeline qualifies as a Level 4 platform, since it integrates multiple hardware operations (*e.g.*, liquid handling and imaging) with iterative, software-driven decision-making.[48] In this framework, the machine learning algorithm autonomously selects future experiments, and the system automatically evolves based on the newly acquired experimental data, while humans are only tasked with defining the initial search space.49 This setup goes beyond traditional, trial-and-error based approaches, and it can generalize to virtually any system: once the initial search space is defined, the condensate phase behavior can be automatically explored in a self-driving manner.

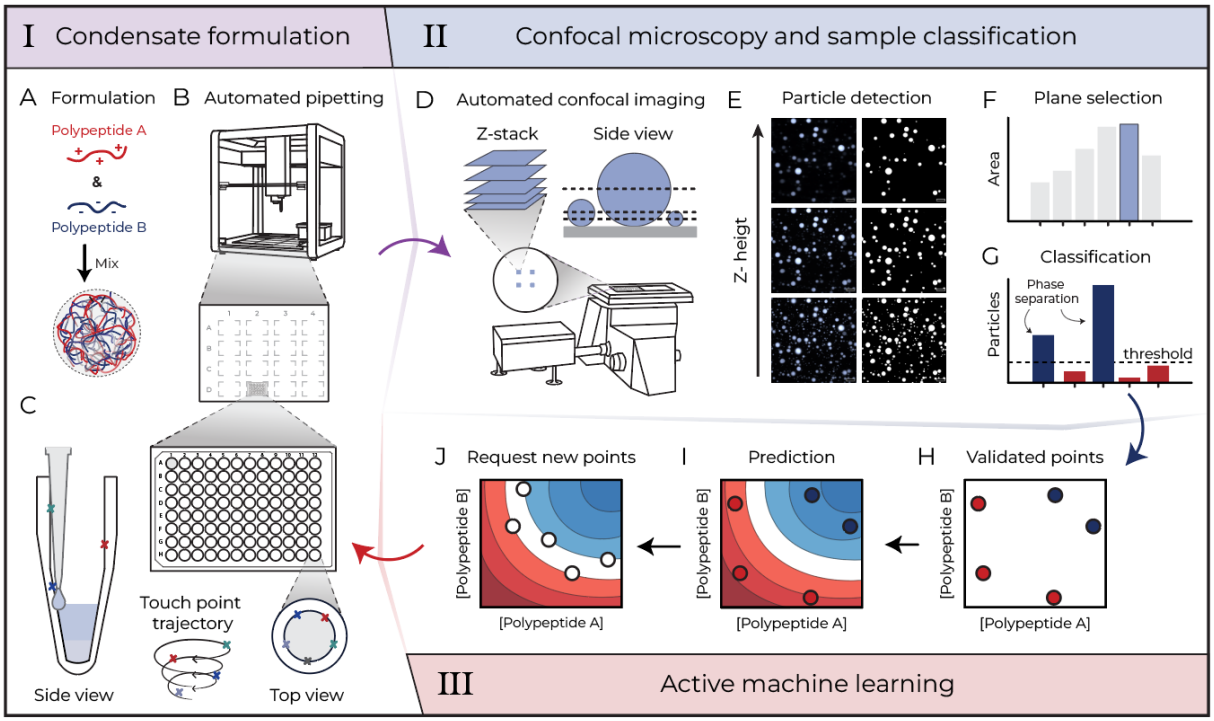


Figure 1: Closed loop navigation of condensate phase diagrams. The workflow is constituted by three parts: (I) condensate formulation, where samples are automatically prepared, (II) confocal microscopy and sample classification, for characterization, and (III) active machine learning, that learns from the collected data and suggests the next experiments. (**A**) Condensate microparticles are formed by mixing cationic and anionic polypeptides, resulting in phase-separated micron-sized droplets. (**B**) Schematic representation of the robotic pipetting platform with 16 flexible deck slots. (**C**) Formulations are prepared in a conical PCR plate, using contactless dispensing with volume tracking. A custom touch-tip functionality follows a touch-point trajectory to ensure accurate dispensing. (**D**) Confocal imaging is performed using dynamic Z-stack acquisition. (**E**) Automated binary Yen-thresholding is applied to each Z-plane for particle detection. (**F**) The optimal Z-plane is selected based on the largest detected area, corresponding to the slice that is best in focus. (**G**) Samples with 12 or more particles are labeled as phase-separated (condensates), while those below the threshold are labeled as non-condensates. (**H**) Experimentally validated data points are incorporated into the machine learning algorithm for training. (**I**) The model predicts a phase diagram based on the acquired experimental data. (**J**) The model then guides the selection of new formulations, restarting the automation cycle at (**A**).

Proof-of-concept: Automated construction of phase diagrams

To showcase the potential of our self-driving platform, we applied it to navigate the phase behavior of poly-L-(lysine) and poly-L-(aspartic acid) (Figure 2), two well-investigated polypeptides in phase separation research.[30,50–53] Even in this case, despite their widespread use, the detailed phase diagrams that capture their binodal boundary remain underexplored, possibly owing, to the need of labor-intensive experiments.[29,30,51] In this context, this condensate system was a useful case-study to investigate how well our automated workflow was suited to effectively determine its phase behavior.

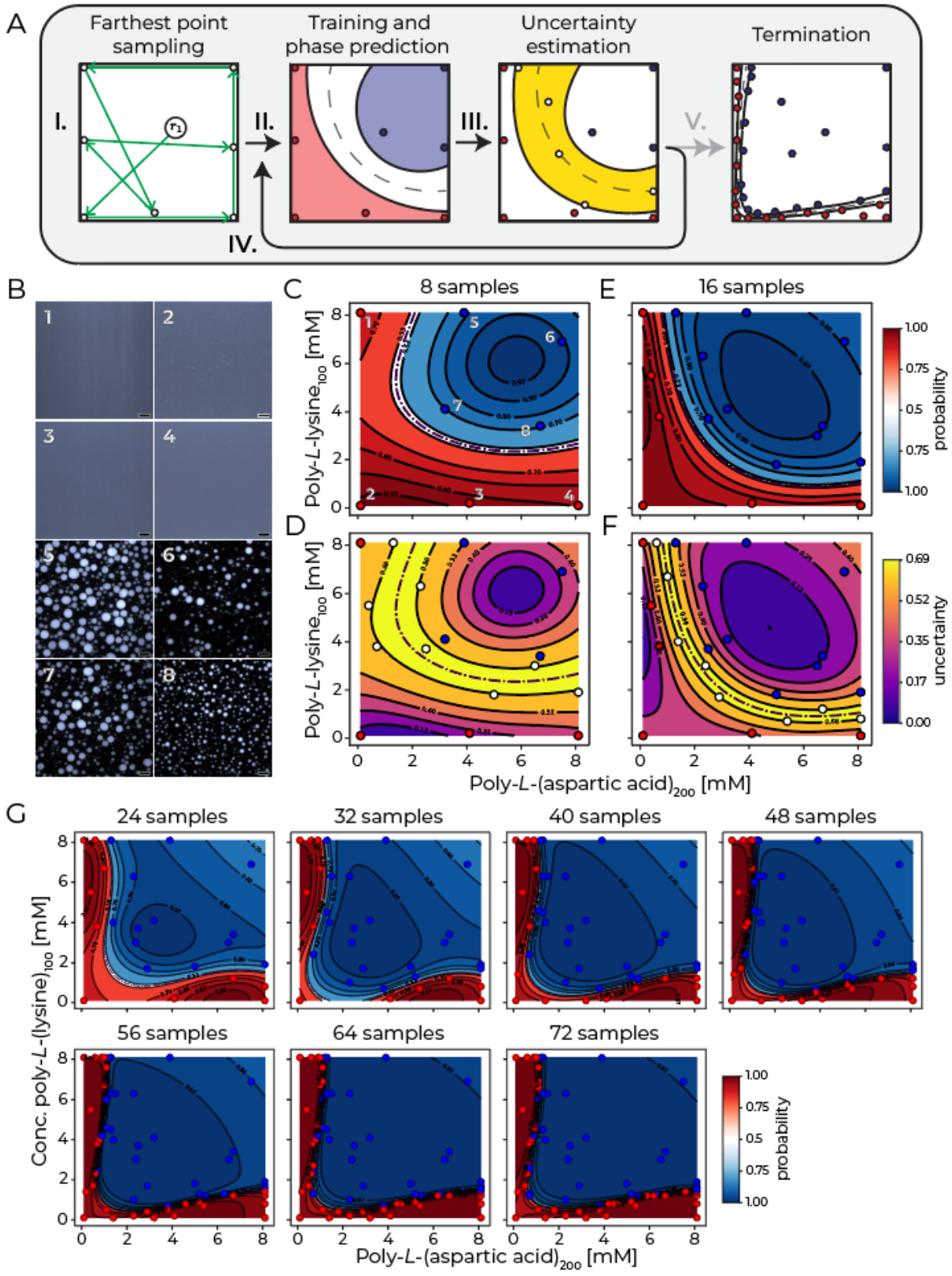
Our experiments followed make-analyze-predict workflow, as follows:

0. *Initialization step* (Figure 2A). We constructed an experimental design space, ranging from 0.1-8.1mM monomer concentration for each polypeptide. As a starting point we used poly-L-(lysine)$_{100}$ and poly-L-(aspartic acid)$_{200}$. Eight points for the experimental formulation and characterization were selected by the farthest point sampling algorithm[47], which starts from a randomly selected point, and then chooses maximally dispersed samples across the design space.
1. *Automated sample production and characterization.* The chosen samples were then formulated and characterized experimentally for their phase separation (Figure 2B). Based on their phase separation behavior, they were labeled as either ‘condensate’, or ‘non-condensate’ for training the machine learning model.
2. *Model training and experiments selection.* The experimentally determined labels were used to train the model and predict the coacervate behavior across the design space (Figure 2C). In particular, the GPC algorithm generates a new phase diagram prediction across the design space. The probabilistic nature of GPC prediction allows to compute an uncertainty measure per class, which we leverage in the form of entropy of the class probabilities (the higher the entropy, the higher the uncertainty across the classes, see Methods, Eq. 4). Once the points within the highest uncertainty regions are identified, farthest point sampling again selects the next batch points for production and characterization (Figure 2D).

After the initialization (step 0), steps 1-2 were iteratively repeated, by adding the new experimental labels to the training dataset and subsequently updating both the phase (Figure 2E) and uncertainty (Figure 2F) landscapes for the next cycle. This active learning process continued until a total of 72 samples were measured across nine cycles (Figure 2A,2G).

After approximately 40 samples (five iterations), only minor changes were observed in the predicted phases, suggesting that the model started to stabilize. Collecting a total of 72 samples further reduced the uncertainty of the predicted phase boundaries (Supplementary

Figures S1-S3). The automated exploration of the phase diagram was carried out in approximately four hours, whereas conducting these experiments manually would have required more than one week. Additionally, the active learning approach generated a detailed phase diagram, a result that would have otherwise required the intuition of an experienced scientist to achieve manually, and potentially many more datapoints. This underlines the platform's effectiveness not only in reducing time but also in directing experimental efforts toward relevant regions for investigation.



* Figure caption on the next page

Figure 2: Machine-learning guided condensate 2D phase diagram mapping. (**A**) Schematic of the active machine learning pipeline used for phase diagram mapping. (I) The initial sample points are selected using farthest point sampling to ensure broad coverage of the design space. (II) A Gaussian Process Classifier is trained on the data, generating a preliminary phase diagram. (III) An uncertainty landscape is computed, highlighting regions with the highest uncertainty. From these regions, new points are sampled using farthest point sampling. (IV) The selected samples are experimentally validated and added to the dataset, refining the phase diagram prediction. (V) Steps I–IV are repeated until convergence is achieved. (**B**) Representative confocal micrographs for the first eight experimentally validated samples. (**C**) The predicted phase diagram for poly-L-(aspartic acid)$_{200}$ and poly-L-(lysine)$_{100}$ based on the validated samples in (**B**). Phase separation is represented by blue points and no separation by red points, with the surface depicting the model's predictions. (**D**) The entropy landscape is constructed based on the prediction in (**C**), and new samples (white points) are selected using farthest point sampling in the high entropy region of the landscape. The requested points are experimentally classified, and a new phase diagram is predicted from the combined data (**E**), along with its associated entropy landscape (**F**). (**G**) Subsequent iterations continue until 72 data points are acquired.

Convergence of condensate phase mapping

A desirable feature when automatically mapping phase diagrams is the unified convergence and reproducibility of the final results regardless of the starting points selection. In fact, while the designed space available for selecting experimental conditions is vast (in this study, a grid of 6,561 points), the models are trained in a low-data regime (up to 72 datapoints), which opens questions about how the underlying patterns and trends are captured.[54] Moreover, given the iterative nature of the approach, initial decisions (*e.g.*, starting points for training) and automation-related challenges (*e.g.*, equipment inconsistencies) might affect decisions in later cycles. To shed light on this key question, we performed three independent replicates using the poly-L-(lysine)$_{100}$ and poly-L-(aspartic acid)$_{200}$ system, so that each replicate was carried out identically (as explained above), but starting from a unique and non-overlapping initial set (step 0) for model training (Figure 3A).

Since the starting sets highly differed across replicates, they resulted in different phase and uncertainty landscapes in early cycles (Figure 3B, Supplementary Figures S4-S7). While each run followed its 'prediction route' across cycles, after approximately 40 samples (cycle number 5), the phase diagrams appeared to converge across the replicates. After collecting 72 samples (cycle number 9), the replicate phase diagrams displayed remarkable similarity and low uncertainty levels.

To further assess the reproducibility of our experiments, we constructed a "ground truth" phase diagram (Supplementary Figure S8) using all data collected across replicates (Supplementary Figures S9-S14). We quantified the prediction agreement between each replicate's predictions (at each cycle) and the ground truth via balanced accuracy (the higher, the more similar the predictions, see Methods Eq. 7).[55] Across replicates, the balanced accuracy steadily increased over successive cycles (Figure 3C), which is especially visible from the fifth cycle onwards, where balanced accuracy reached values consistently above 95% across all replicates. This indicates that, no matter the starting point, all replicates converge to a similar phase diagram in a data-efficient way (*i.e.,* by using substantially less data than the "ground truth" diagram). These results agree with existing active learning literature[39,54,56], showing the potential of this approach to progressively mitigate the effect of the starting data.

The Jensen-Shannon divergence[57] (see Methods, Eq. 8-9) was computed to directly compare phase diagrams (the lower the divergence, the more similar). A "within-replicate" divergence was calculated, by comparing the predicted probabilities of each replicate across consecutive cycles (Figure 3D). The results showed an exponential decrease in divergence values, with substantial changes in the predicted phase diagrams within the first 32 samples (cycle 4) and minimal changes after 56 samples (cycle 7), suggesting that each phase diagram

reached a ‘stable’ state, where additional experiments did not significantly alter predictions. Moreover, we calculated a “between-replicate” divergence (Figure 3E), by comparing the predictions of each cycle across different replicates. The divergence values decreased sharply during the first three cycles, after which they stabilized. These results indicate that only three cycles were necessary to mitigate the stochastic differences by the different starting points, after which the replicates progressively aligned along a common trajectory.

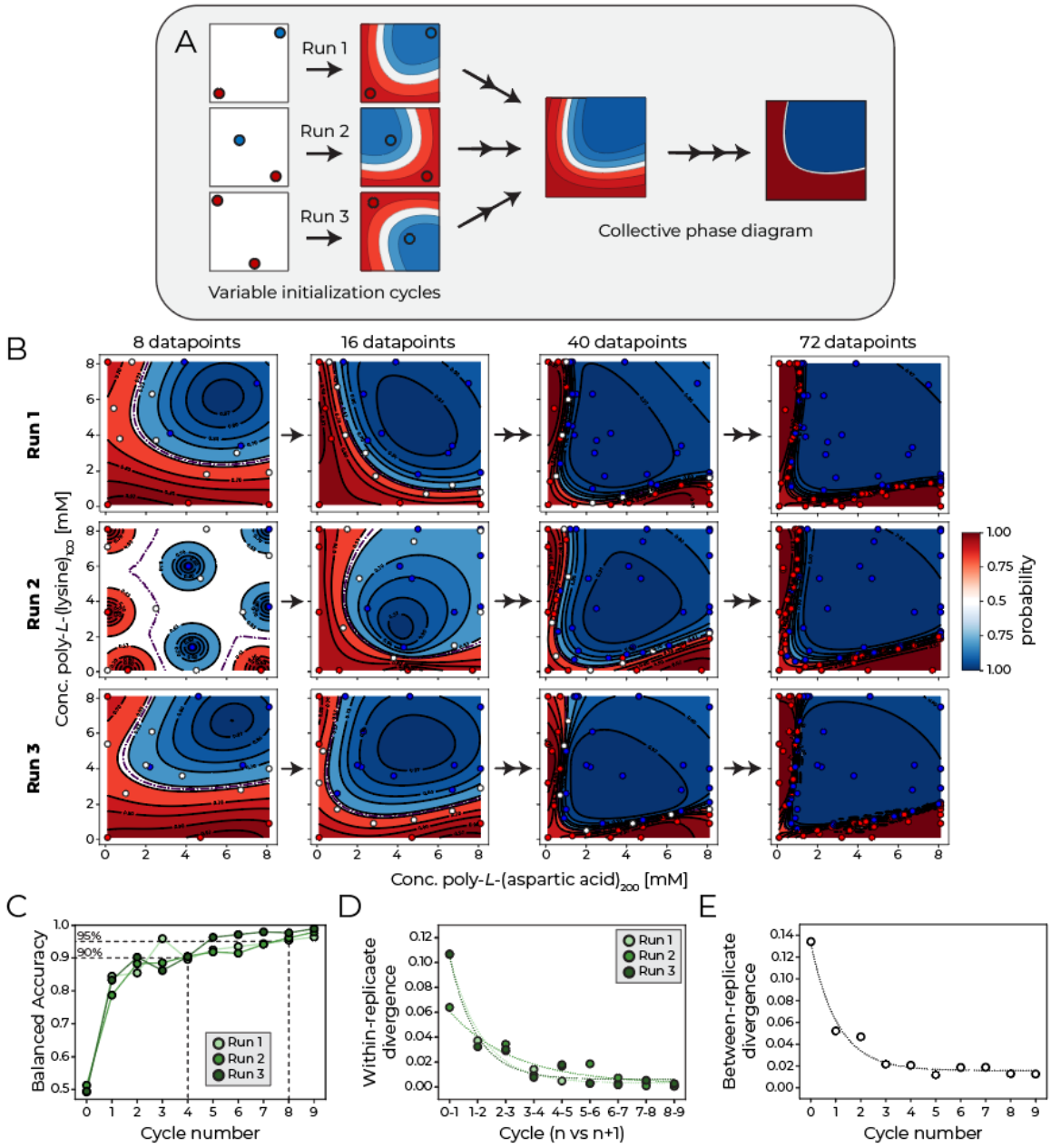


Figure 3: Consistent convergence of phase mapping across replicates. (**A**) Schematic illustration of the experimental workflow used to produce replicated phase diagrams. The initial set of experimental samples is selected by farthest point sampling, resulting in different starting points for each replicate. Each subsequent cycle then follows a unique path to reach the same phase diagram. Reproducibility across replicates is expected only if the machine learning, formulation, and analysis steps are consistent. (**B**) Phase diagrams, showing phase separation in blue and no separation in red, with probability prediction fits (background), validated points (colored dots), and new sample selections (white dots) for three experimental replicates of poly-L-(lysine)$_{100}$ and poly-L-(aspartic acid)$_{200}$ condensates. A total of 72 data points is experimentally validated across 9 cycles. Representative cycles are shown, remaining cycles and entropy maps are reported in Supplementary Figures S2-S7. (**C**) Balanced accuracy plot showing the accuracy on the prediction for each successive cycle with respect to a "ground truth" phase diagram. Cycle 0 represents the balanced accuracy computed with respect of a randomly generated phase diagram as a baseline comparison. (**D**) Average Jensen-Shannon Divergence plot illustrating within-experiment divergence by comparing consecutive cycles for each replicate. This reflects the progressive convergence toward the final phase diagram for each replicate. Cycle 0 is a random phase diagram included as a reference for low similarity. (**E**) Average Jensen-Shannon Divergence plot comparing divergence across replicates at each cycle, highlighting inter-experiment variation. Cycle 0 compares three random phase diagrams and is included as a reference for low similarity.

Mapping condensate properties via phase diagram exploration

Traditional studies on phase separation behavior have primarily focused on determining whether condensates form under specific conditions.[27,28,30,34,58] Our automated data production and characterization pipeline collects additional information beyond phase separation. In particular, depth-resolved imaging from confocal microscopy allows to derive several properties of condensates, including particle count, morphology, and volume fraction, within the phase diagram. Here we compounded the data from the previously described replicates, along with data obtained from optimization experiments, totaling 480 experimentally determined samples (Figure 4A, Supplementary Figure S8). The collected samples show a wide range in particle morphologies, ranging from densely packed condensate clusters to tiny, barely visible particles (Figure 4B). This diversity underscores the variability in condensate formation even within a single “simple” phase, underscoring the necessity of collecting a broader set of properties to gain a deeper understanding of phase behavior.

Here, we focused on the following condensate properties: (a) number of detected condensates, (b) average particle area, and (c) total volume fraction, extrapolated by combining particle counts and area. These properties were mapped onto the compounded phase diagram, and all of them showed evident trends across the experimentally determined space. Low particle counts were for example observable near phase boundaries, while the count increased when both protein concentrations increased (Figure 4C). Particle size showed a similar trend, with larger condensates forming at higher concentrations (Figure 4D). Some regions showed fewer but larger particles, suggesting potential fusion (coalescence) of condensates due to surface saturation. Volume fractions were lower near the phase-separation boundaries and higher toward the inner part of the phase separated region (Figure 4E). This additional information ‘augments’ the insights on condensate systems, by allowing to map phenotypic variations onto the predicted phase separation space. These insights might help guide coacervate formulations in those regions of the experimental space where specific properties are desirable (in addition to the phase-separating behavior).

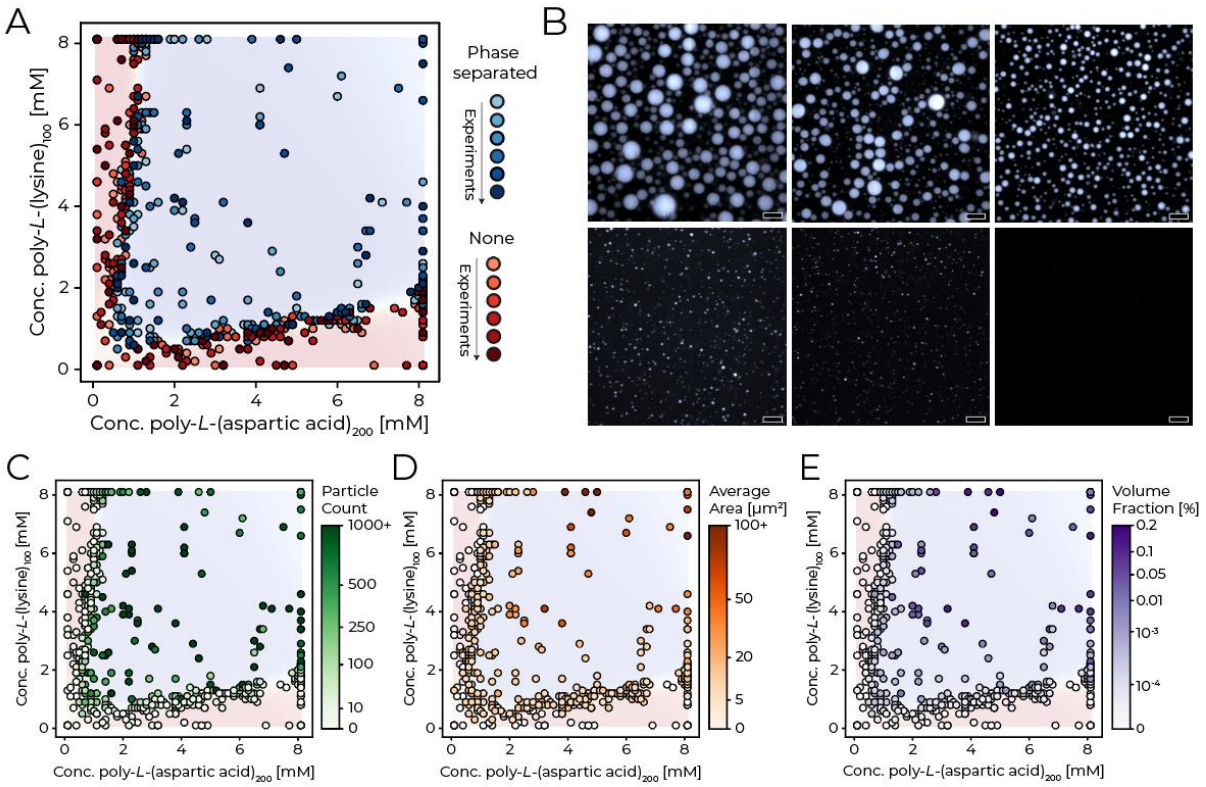


Figure 4: Condensate properties beyond phase boundaries. (**A**) Combined 2D phase diagram of poly-L-(lysine)$_{100}$ and poly-L-(aspartic acid)$_{200}$, based on 480 validated data points. The data is compiled from six independent experiments, represented by varying shades of blue (indicating phase separation) and red (indicating no phase separation). (**B**) Representative confocal micrographs that illustrate the wide range of observed condensate phenotypes (scale bar = 20µm). (**C**) Mapping of the number of detected condensates (represented by dots) overlaid on the phase predictions of the combined dataset extracted from (**A**). (**D**) Mapping of the particle area [µm²] onto the phase predictions of the combined dataset extracted from (**A**). Each dot represents the average particle size. White data points indicate conditions where no particles were detected above the size threshold (500 pixels). The average size is colored using a non-linear gradient. (**E**) Mapping of the condensate volume fraction onto the phase predictions of the combined dataset extracted from (**A**). Each dot represents the apparent volume fraction of the dense/dilute phase, extrapolated from the particle counts (**C**) and average particle size (**D**). The apparent volume fraction is colored using a non-linear gradient.

Identifying structure-separation relationships with automation

To further extend the applicability of our workflow, we applied it to elucidating how polypeptide chain length affects phase separation. We constructed phase diagrams for nine combinations of poly-L-(lysine) and poly-L-(aspartic acid) polypeptides, each differing in chain length (poly-L-(lysine)$_n$: n = 20, 100, 250; poly-L-(aspartic acid)$_n$: n = 30, 100, 200) but with constant overall monomer concentrations. All combinations exhibited phase separation within the tested experimental space (Figure 5, Supplementary Figures S15-S30). However, although these polypeptides share the same structural monomeric unit, their phase behavior, as well as their properties (Supplementary Figures S31-S33) varied considerably. The machine-learning-guided exploration of these phase diagrams was carried out in approximately one week, whereas conducting these experiments manually and based on intuition would have been seriously challenging and labor-intensive.

Notably, even with the more complex and curved diagrams of some of the combinations, we successfully identified well-defined phase boundaries within 72 samples (9 cycles) for all tested conditions. Generally, increasing the length of one polypeptide while keeping the length of the other polypeptide constant enabled phase separation at lower concentrations for the elongated polypeptide, but it required higher concentrations of the fixed-length polypeptide, as visible, for instance, in the case of poly-L-(lysine)$_{20}$ (Figures 5A-C). Similarly, when the poly-L-(lysine) length increased from 20 to 100 or 250 repeats (Figures 5D-I), while maintaining a constant poly-L-(aspartic acid) length, phase separation occurred at lower lysine concentrations, but required higher concentrations of poly-L-(aspartic acid).

These results highlight the delicate balance required in designing polypeptide systems for phase separation. Simply increasing the concentration or length of one polypeptide does not necessarily lead to enhanced phase separation; instead, the process is highly sensitive to the interplay between both polypeptides. Our findings indicate that an optimal balance exists at equal chain lengths of 100 repeats (Figure 5E), where phase separation occurs extensively across most of the investigated chemical space. In some cases, particularly with poly-L-(lysine)$_{250}$, phase boundaries showed slight bends, suggesting complex, non-linear dynamics. These complexities highlight the challenges in controlling and predicting condensate formation, as even minor adjustments at the molecular level can lead to pronounced changes in phase behavior.

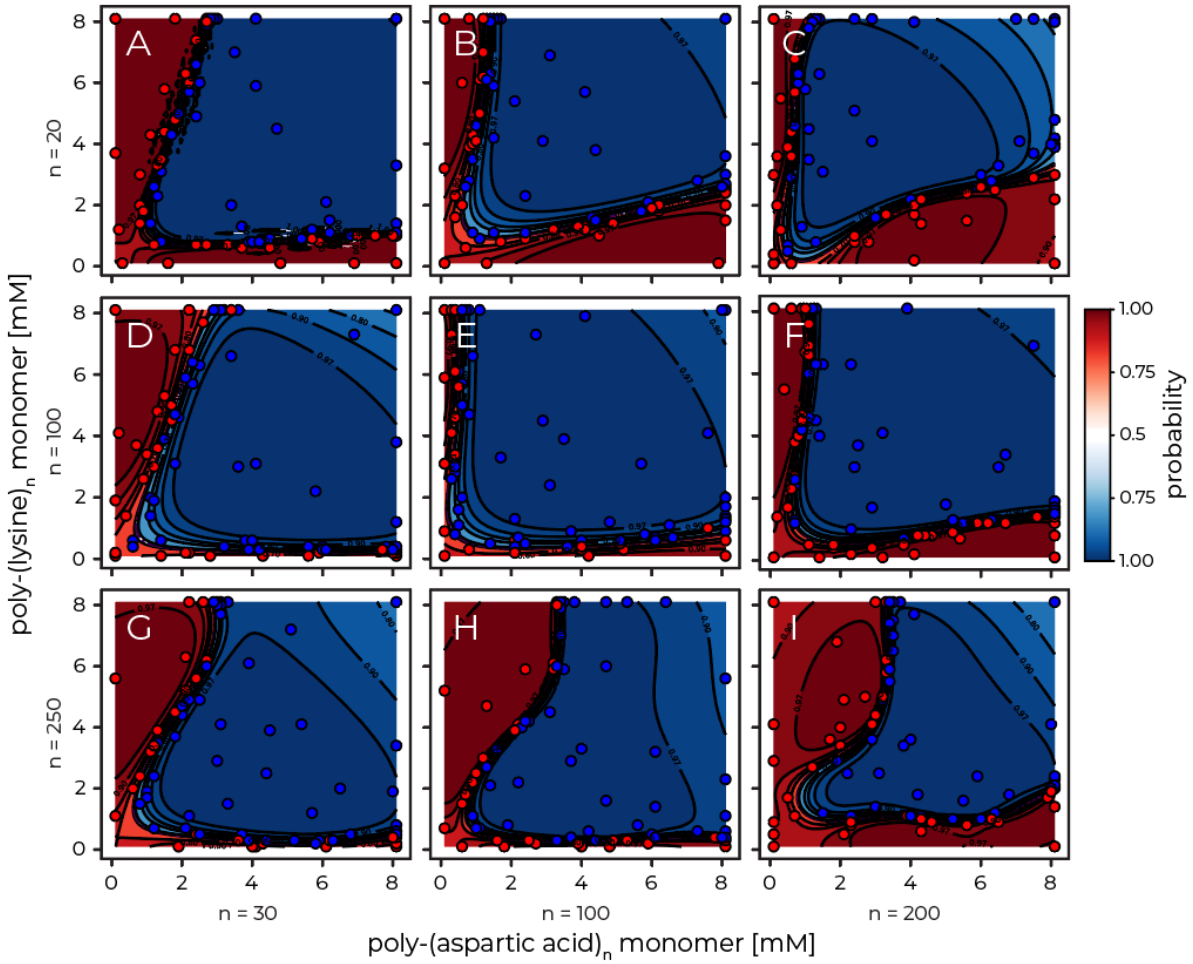


Figure 5: Impact of polypeptide length on phase separation behavior. This figure displays nine phase diagrams illustrating the automated mapping of phase separation for combinations of poly-L-(lysine) and poly-L-(aspartic acid) with varying chain lengths. Panels (**A-C**) represent phase diagrams for poly-L-(lysine) with a chain length of 20, combined with poly-L-(aspartic acid) of lengths 30 (**A**), 100 (**B**), and 200 (**C**). Panels (**D-F**) show poly-L-(lysine) with a chain length of 100, paired with poly-L-(aspartic acid) lengths of 30 (**D**), 100 (**E**), and 200 (**F**). Panels (**G-I**) depict poly-L-(lysine) with a chain length of 250, combined with poly-L-(aspartic acid) lengths of 30 (**G**), 100 (**H**), and 200 (**I**). Datapoints are marked as dots, with blue indicating phase separation and red indicating no phase separation. Each phase map includes a background color gradient derived from predictions based on 72 datapoints per combination, acquired over nine cycles of eight datapoints. Remaining cycles and entropy maps are reported in Supplementary Figures S31-S33.

Navigating phase behavior in complex environments

Building upon these results, we increased experimental complexity by introducing salt (NaCl) as an additional dimension to our system. Salts modulate electrostatic interactions between charged polypeptides and thereby significantly influence condensate phase behavior and properties.[24,59] This expansion increased the potential experimental space from 6,561 points (two dimensions) to 531,441 points (three dimensions). To evaluate the platform's performance, we performed two independent replicates using the poly-L-(lysine)$_{100}$ and poly-L-(aspartic acid)$_{200}$ system for 20 active learning cycles with 32 samples each (640 measured points per replicate; Supplementary Figures S34-S35). These newly acquired points were compounded with previous data to construct a comprehensive 3D "ground truth" phase diagram (Figure 6A-B). As anticipated, salt greatly influenced condensate formation, promoting phase separation at moderate concentrations (150–700 mM), while disrupting it at higher concentrations (1200–1300 mM). Interestingly, some phase-separated regions at higher salt concentrations were identified (Figure 6D, 270° rotation), resulting from salt-induced aggregate phases.

To assess the pipeline's reproducibility and performance, we again calculated the balanced accuracy[55] (See Methods, Eq. 7, Fig. 6C) and within- and between-replicate Jensen-Shannon divergence[57] (see Methods, Eq. 8-9, Fig. 6D-E). As anticipated, all metrics showed consistent improvements across cycles and rapid convergence toward the global phase diagram, with stabilization occurring after approximately eight cycles (256 samples). Notably, these metrics effectively captured the overall progression in identifying phase behavior but may be less sensitive to minor changes in the large design space during the later stages of optimization. Nonetheless, the balanced accuracy continued to improve slightly in subsequent cycles, primarily enhancing the resolution around the phase boundaries (white areas in Fig. 6A, B; Supplementary Figures S34-S35).

Increasing dimensionality introduces challenges, both for machine learning algorithms and due to the formation of distinct aggregate phases. Despite these challenges, we successfully mapped these 3D phase diagrams in just three days. These results not only demonstrate the platform's capability to rapidly explore vast and complex design spaces but also highlight the essential role of machine learning in effectively navigating and elucidating such high-dimensional complex assemblies.

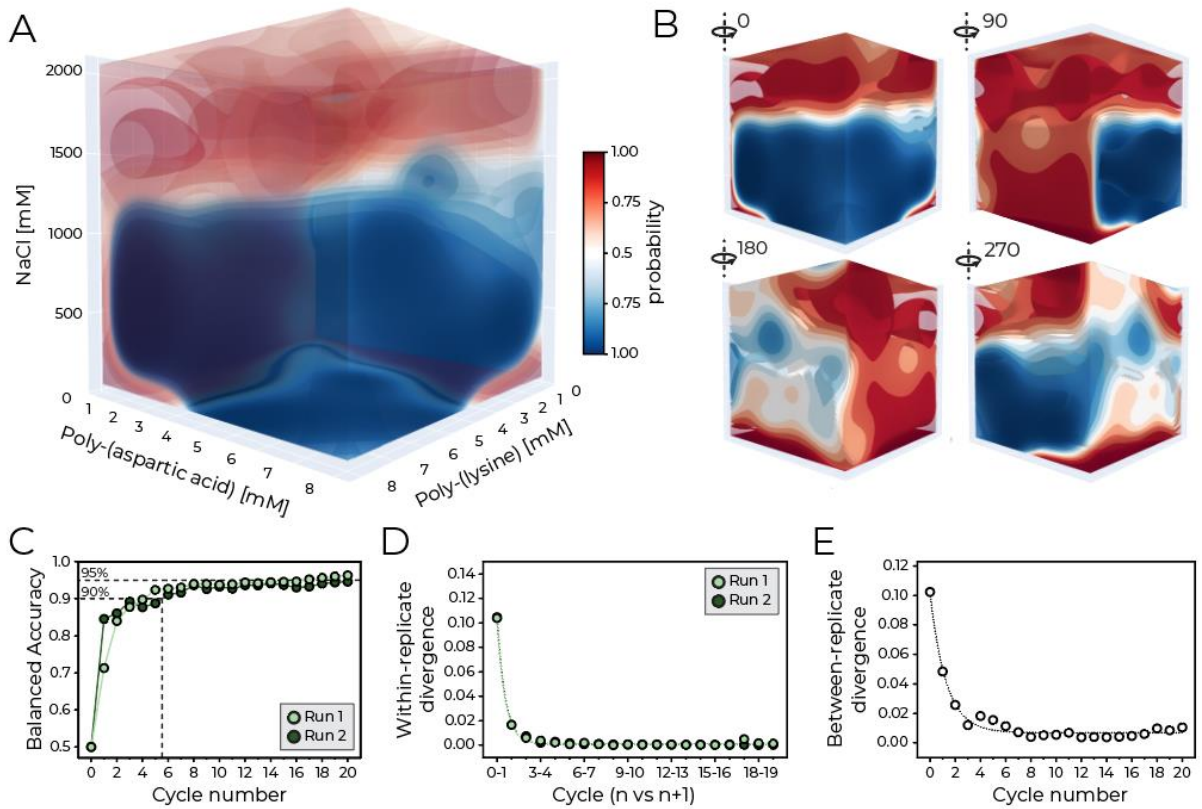


Figure 6: Automated mapping of multi-dimensional phase diagrams. (**A**) Two independent experiments (Supplementary Figure S34-S35) were conducted to explore the effect of salt (NaCl) on the phase behavior of poly-L-(lysine)$_{100}$ and poly-L-(aspartic acid)$_{200}$. The combined dataset, made from 1760 datapoints, was used to construct the "ground truth" three-dimensional phase diagram, here reported. Iso-probability surfaces indicate phase separation (blue, higher opacity) and no phase separation (red, lower opacity). (**B**) Four distinct orientations of the phase diagram with non-transparent surfaces are shown to emphasize phase behavior from different perspectives. (**C**) Balanced accuracy plot showing the accuracy on the prediction for each successive cycle with respect to the "ground truth" phase diagram in panel A. Cycle 0 represents the balanced accuracy computed with respect of a randomly generated phase diagram as a baseline comparison. (**D**) Within-experiment Jensen-Shannon Divergence (JSD) plotted across cycles. This metric tracks convergence by comparing consecutive cycles, illustrating how each replicate approaches the final phase diagram. Cycle 0 reflects divergence from a randomly generated phase diagram. (**E**) Between-experiment Jensen-Shannon Divergence (JSD) across replicates at each cycle. Similar to panel D, Cycle 0 serves as a baseline, representing divergence from a randomly generated phase diagram.

Discussion

In this work, we presented a versatile, machine learning-driven automated platform that rapidly navigates multi-dimensional phase diagrams of condensates. By integrating (a) active machine learning to optimize sample selection and phase diagram navigation, (b) automated pipetting for precise sample formulation, and (c) advanced and automated confocal microscopy for high-content particle characterization, we examined the phase behavior of polypeptides across various formulations and concentration profiles. Our platform reliably and rapidly identified phase boundaries with high accuracy and reproducibility, demonstrating the robustness of our approach. Additionally, it quantified key condensate properties, such as morphology, particle count, and volume fraction, providing insights beyond the traditional binary classifications of phase separation. Moreover, the platform's flexibility enabled rapid exploration of complex phase spaces, allowing to reveal the influence of polypeptide chain length and salt on phase behavior.

Looking forward, numerous opportunities exist to further enhance our platform's capabilities and broaden its applications. By refining the sampling strategies (*e.g.*, by balancing exploration of uncertain regions with exploitation of high-certainty points) the efficiency of phase diagram navigation could be further improved.60 Furthermore, integrating robotics to enhance platform autonomy[61,62] and leveraging machine learning for advanced image analysis can significantly improve condensate classification.[63] Moreover, integrating condensate properties into active machine learning algorithms will allow us to incorporate desirable particle properties in the decision-making process, supporting the design of biomaterials for applications such as drug delivery and tissue engineering.[64] In the future, incorporating molecular information into machine learning models (*e.g.*, via deep learning[65,66]) will enable linking molecular structure with phase behavior, extending beyond the training sets.[67] Finally, the platform's modularity and adaptability make it generalizable to other complex micron-sized assemblies, such as tactoids[68] and microgels.[69] This versatility also opens up opportunities to explore how minor structural and compositional changes in natural proteins, resulting from processes like splicing, mutations, and post-translational modifications, influence condensate behavior, offering valuable insights into phase separation principles under diverse conditions.[70]

Methods

Preparation and Dye Labeling of Polypeptides

All polypeptides used in this study were purchased from Alamanda Polymers. They were dissolved in fresh Milli-Q water (MQ) at 25 mg/mL, then sterile-filtered through a 0.2µm filter, and stored in aliquots at -20°C. Further dilutions were prepared in MQ, with stock solutions maintained at 4°C.

A portion of the poly-L-(lysine) polypeptides was labeled with NHS-Sulfo-Cy5 dye (Lumiprobe) for confocal imaging. The dye was dissolved in DMSO at a concentration of 10 mg/mL and stored at -20°C. Poly-L-(lysine) was labeled in a reaction buffer consisting of 100 mM HEPES (pH 8.0) and 150 mM NaCl in MQ. The polymer-to-dye ratios were 1:3 for poly-L-(lysine) with a chain length of 100 and 1:6 for chain lengths of 20 and 250. The reaction was carried out for two hours at room temperature while shaking at 550 rpm using an Eppendorf MixMate.

Unbound dye was removed using a PD Minitrap G-25 size exclusion column (Cytiva), which was pre-equilibrated with a storage buffer of 25 mM HEPES (pH 7.4) and 100 mM NaCl. Labeling was, if possible, further confirmed by analyzing the flow-through of the dye-labeled polymer after centrifugation with a 3 kDa spin filter (Amicon). All polypeptides were freeze-dried, weighed, and dissolved in MQ. The dye concentration was determined using a nanodrop spectrophotometer (Thermo Scientific NanoDrop 1000). This measurement, combined with the dry weight of the polypeptide, allowed for the calculation of the Degree of Labeling (DoL). The final dye-labeled polypeptides were sterile-filtered (0.2 µm) and stored at -20°C, with additional dilutions prepared in MQ and maintained at 4°C.

Preparation Microscopy Plates

For confocal imaging, black 96-well glass-bottom microscopy plates (Cellvis, 1.5, P96-1.5H-N) were used and the glass surface was passivated to prevent wetting of the condensates. To prepare the surface coating, bovine serum albumin (BSA) was dissolved in MQ at 30 mg/mL and then sterile-filtered through a 0.2 µm filter. A volume of 100 µL of this BSA solution was added to each well. The plates were placed on a MixMate shaker (Eppendorf) and incubated at 500 rpm for 60 minutes at room temperature. After incubation, the BSA solution was discarded, and each well was rinsed three times with 100 µL of MQ water. The plates were then dried overnight, covered with a Kimwipe, and stored at room temperature under a protective cover until use.

Data Architecture and General Automation Workflow

All devices were integrated within a local network and regulated through a central orchestrator workstation, which served as the control hub for the entire platform. The orchestrator contains all necessary protocols and information, and coordinates all device actions and data exchange. Communication with platform components was achieved through USB connections and a local Ethernet network, using TCP-based network communication protocols such as SSH and HTTP.

A centralized data architecture was implemented to manage knowledge transfer between instruments. This architecture included a structured folder system on the orchestrator workstation for organizing Python protocols, instrument logs, raw data storage, and dedicated information transfer files. These information transfer files, detailed below, contained specific instructions for each device—often generated through machine-learning algorithms—and were sent from the orchestrator workstation to individual components. Each device passively listened to the orchestrator workstation, which assigned tasks and actions directly. Devices executed only the actions directed by the orchestrator, forming a streamlined, centralized data workflow across the platform.

- Master File: Central, continuously updated database for all sample details, conditions, and results. It logs sample locations and barcoded plates, directing sample creation, handling, and analysis. A versioned copy was made before each update to maintain data integrity.
- Barcode File: Output from machine learning, which is cross-referenced with the Master File to identify samples to be processed.

- Batch File: Contains a detailed description of polypeptide stocks, including date, version, and degree of labeling for dye-labeled polypeptides. It is essential for calculating component volumes in sample preparation.
- Source File: Tracks materials that are stored in a 96-well plate, including their concentrations and volumes. This file was updated after each pipetting step and versioned once per automation cycle to support accurate records.

This system operated in a closed-loop workflow, where each action depended on information from previous steps, all coordinated by the central orchestrator workstation. The workflow began with the machine learning model, which assessed the chemical space and determined the next set of samples to be measured. It appended these new sample conditions to the Master File and created a matching Barcode File. Next, the pipetting platform used information from the Master, Source, and Batch Files to calculate the required volumes and assign target locations for each sample. During sample preparation, the Source File was updated after each pipetting step to keep track of remaining volumes. Once the samples were prepared, their locations were added to the Master File. The microscope then cross-referenced the Barcode File with the updated Master File to find sample locations and imaging coordinates. It automatically acquired and processed confocal micrographs and added the classification results to the Master File. Finally, the machine learning model retrieved these updated classifications, incorporated them into the chemical space, and initiated the next cycle of experiments.

Automated Sample Preparation

Instrument Setup and Configuration. Samples were prepared automatically using an Opentrons Flex pipetting robot equipped with both single- and 8-channel pipettes (5-1000 µL) and 200 µL tips. The deck was configured as follows: 200uL tip rack in slot B1; 195 mL NEST reservoir filled with MQ in slot C2; Heater-shaker module (Gen 1) with a PCR adapter plate and either a NEST 96-well PCR plate for 2D phase diagrams or an Opentrons Tough 96-well PCR plate for 3D phase diagrams in slot D1; 2mL 96-well deep-well plate (NEST) containing stock solutions in slot D2; waste chute in slot D3; and a 96-well microscopy plate (Cellvis) in slot C3.

Pipette Offset Calibration. The Flex platform was calibrated for height and x/y offsets, following the manufacturer's guidelines.

Source Plate Setup. Stock solutions of HEPES, NaCl, and polypeptides (labeled and unlabeled) were preloaded in the source plate (D2). The robot tracked and updated each well's volume (see Data Architecture and Workflow), prompting refills to bring wells up to 1800 µL when volumes dropped below 200 µL.

Liquid handling. Reagents were dispensed sequentially to achieve a final volume of 150 µL per PCR well: MQ water, HEPES buffer (50 mM, pH 7.4), NaCl (150 mM for 2D or 25-2050 mM for 3D diagrams), dye-labeled poly-L-(lysine) (96-250 nM), unlabeled poly-L-(lysine), and poly-L-(aspartic acid) (0.1-8.1 mM monomer concentration). Final calculations accounted for any additional monomers introduced by the dye-labeled poly-L-(lysine) to ensure accurate concentrations. The same tip was used for multi-dispensing reagents, with new tips used for each aspiration step (except MQ).

Mixing. From NaCl addition onward, samples were mixed (1500 rpm) for 10 seconds. After the final component (poly-L-aspartic acid), samples were mixed (1500rpm) for 5 minutes to promote phase separation.

Custom Dispensing Technique. Transfers used a minimum of 10 µL, leaving 5 µL residual volume in the tip to improve precision. After dispensing, a custom touch-tip function directed the pipette to contact specific points along the well walls for droplet removal (see side view in Figure 1C). Dynamic volume tracking adjusted pipette height and radial position for each touch point, with unique points assigned per liquid to prevent contamination (see top view and trajectory in Figure 1C)

Final Transfer for Imaging. After preparation, 100 µL of each sample was transferred to the imaging plate (C3), which was then sealed with an adhesive aluminum foil seal (ThermoFisher) for confocal imaging. A new tip was used for each well, with samples mixed three times by aspiration/dispensing before transfer.

Automated Confocal Microscopy

Confocal Microscopy Setup and Hardware Configuration. Imaging was conducted on a custom confocal setup integrated by Confocal NL. The microscope consisted of an open-frame inverted microscope (Zaber), with a confocal NL line re-scan system (NL5+) mounted on the left-side camera port. Additionally, the microscope was equipped with a motorized filter wheel (Confocal NL), and a laser autofocus module (Zaber). The NL5+ unit was equipped with an sCMOS camera (Teledyne Photometrix BSI express), providing a large Field of View of 18.8 mm (diagonal). Laser excitation from an Oxxius L4Cc laser diode combiner (containing a 638 nm laser) was coupled to the NL5+ module via an optical fiber. All experiments were conducted using laser power 7%, and a 60x air objective (Nikon, NA 0.95).

Software and connections. All components were controlled via Python. Specifically, pycromanager interacted with Micro-Manager (version 2.0.3) to control laser powers, Z-stacks, and XY positioning. Additionally, the zaber_motion library was connected to the Zaber Launcher (version 2024.11.14) to control the autofocus device.

Automated autofocus adjustments. The autofocus loop involved several steps. To start, the objective was initially directed to a preset Z-position, aligning the autofocus laser within range for the first autofocus attempt. The autofocus was then triggered, aligning the objective with the bottom of the imaging plate. This in-focus focal height was recorded and serves as a reference for the next autofocus loop. After acquisition (see below), the autofocus routine subsequently started each new loop 10 µm below the previously recorded focal plane, searching upward to locate the plate bottom.

XY Positioning and Image Acquisition. The 96-well microscopy plate was mapped into 2x2 grids (550 µm spacing), creating technical replicates within each well. A well-specific event list was created, associating each well with the correct sample barcodes, coordinates, grid locations, and channel information. The scanning algorithm employed a snake pattern, optimizing acquisition time by minimizing travel distance and positional drift across the microscopy plate. Following autofocus, Z-stacks were captured as height additions on top of the recorded autofocus height, using dynamic spacing: a fine 0.5 µm step for the first 5 µm, increasing to 1.0 µm for the next 5 µm, then 2.5 µm for the following 5 µm, and finally 5.0 µm for deeper layers, spanning a total of 50 µm of Z-depth per position. Acquisitions were performed at 5 frames per second.

Verification of Imaging Completion. To monitor imaging progress, a continuous background process compared the number of saved image slices to the expected slice count based on the number of imaging events (i.e., focal planes across wells). Once the saved slice count matched the target, the acquisition was deemed complete, and MM and associated processes were automatically closed.

Automated Image Analysis and Classification. Each acquired micrograph underwent automated analysis to extract sample classifications and particle features. Particles were detected using the scikit-image Python module. Yen thresholding was applied to create a binary mask, which was used to detect particles above 500 pixels (5.87um^2). The extracted particle properties (e.g., X/Y position, area, mean intensity) were saved for each micrograph. Results were then grouped by grid position and sorted by Z-index. For each particle, the slice with the largest detected area was selected as the representative view, which was used for the property mappings performed in this study. Wells were classified based on particle count and distribution, with 12 or more particles across at least three grid positions indicating “Phase Separation” and fewer particles marking “No Phase Separation”.

Machine Learning and Computation

Design of the parameters space. The initial dataset (*i.e.,* cycle 0) for any given system formulation was created by computing a regular D-dimensional grid of points (with D being the number of variables to be considered), where each independent component of the formulation accounts for a dimension. Two of the dimensions were always assigned to the concentration of the two oppositely charged polymers, poly-L-(lysine) and poly-L-(aspartic acid) respectively. Additional dimensions could be added to account for other behaviors. The response variable was represented by an integer that mapped the recorded phase to either coacervate or not. In all our experiments we restrained our formulations to study the coacervation phenomena of two oppositely charged polymers as a function of the two polymer concentrations and the salt concentration. Additionally in the current work, we only focused on 2-D and 3-Dimensional datasets. This means that in the former case (2-D) the salt concentration is fixed and

kept constant, while in the latter case (3-D) it is allowed to change. The range for the polymer concentrations was constrained to be the same for all the experiments, regardless the polymer identity, and it was chosen to be a regularly spaced interval starting from concentration of 0.1 mM to 8.1 mM, with steps of 0.1 mM, giving a total of 81 concentrations values (end points included). Similarly, the range for the variation of the salt concentration was chosen to vary from 50 mM to 2075 mM with steps of 25 mM, giving a total of 81 values. All the ranges were chosen accordingly to the accuracy of the machines used to formulate the solutions. Finally, the dataset was created by filling a 2-D or 3-D regular grid with the values of the variable under investigation, creating a total of 6561 (81x81) points for the 2-D case, and 531441 (81x81x81) for the 3-D case. In all the experiments the response variable was set to -1, the undefined default value, for all the points of the grid.

Selection of new points. Starting from cycle 0, and for each cycle, a subset of points n (*i.e.,* new formulations) was sampled from the available pool of points N. The chosen sampling techniques followed the rules of Farthest Point Sampling (FPS).[47] FPS is a sampling technique used to select a subset of points that are maximally spread out from each other within a given dataset. The goal is to retain points that represent the diversity of the data distribution by maximizing the minimum distance between selected points. Given a starting dataset $X = \{x_1, x_2, \dots, x_N\}$ of N points, a first random point $r_1 \in X$ was selected and added to the set of sampled points $S = \{r_1\}$. For each remaining point $x \in X \backslash S$ the minimum distance to any point in S was computed:

$$d(x) = \min_{r \in S} \|x - r\|. \tag{1}$$

Then, the point x_i with the largest $d(x_i)$ was selected and added to the set of sampled points (*i.e.,* the point farthest from the currently sampled points). This selection was repeated until the desired number n of points was reached. The result is a subset $S \subset X$ of n points that were distributed in such a way that they maintain maximal separation, thereby capturing the structure of the original dataset more effectively than random sampling in cases where spread was important.

Phase diagram (PD) prediction. At each cycle N, a phase diagram was predicted using the data that has been experimentally tested in cycle $N-1$. In the case of $N = 0$, no previous tested data was available, the prediction was skipped, and the FPS selected points were fed to the experimental validation pipeline, where their phase is recorded. For all $N \geq 1$, all the points assigned to the sampled set S, after experimental validation, would be used as the ground truth for a Gaussian Process Classifier (GPC)[46] model, that is going to predict the phase distributions over the entire input space. The GPC models the probability distribution over classes (*e.g.,* the phases) by defining a latent function $f: \mathbb{R}^d \rightarrow \mathbb{R}^K$ that associates each input $x \in \mathbb{R}^d$, contained in the input space, with a set of probabilities $p(y = c|x, S)$, where $c \in \{1, \dots, K\}$ represents the class labels. The training step involved using the subset S to learn the posterior distribution of f, which, in turn, yielded a probabilistic model capable of assigning any points $x_i \in X$ to the probability of belonging to a specific class.

In our case, for each point in the input space the GPC would output a probability vector defined as follows:

$$\boldsymbol{p}_i = [p(y = 1|x_i, S), p(y = 2|x_i, S)], \tag{2}$$

where each component represents the probability of x_i belonging to either the “non-aggregate” ($y = 1$) or “coacervate” ($y = 2$) class. Obviously, given that $\boldsymbol{p}_i$ is a probability vector, it holds that the value for the sum of the individual contribution in Eq. 2 needed to sum up to 1. Thus, the GPC trained on the set of all the sampled and tested points could be used to provide a probabilistic prediction over the entire dataset, simply defined concatenating the individual vectors (Eq. 2) for all the points contained in X,

$$\boldsymbol{P} = [\boldsymbol{p}_1, \dots, \boldsymbol{p}_i, \dots, \boldsymbol{p}_N]. \quad (3)$$

Equation 3 enabled inference about phase membership across all points, essentially representing the phase diagram.

The GPC algorithm used in our work was defined using a Radial Basis Function (RBF) kernel with length scale 1.0, multiplied by a constant kernel with default value of 1.0. In each application of the prediction algorithm, we allowed for an automatic internal optimization step by setting the parameters `n_restarts_optimizer` to 5, and the `max_iter_predict` to 150 (more information can be found on the original GPC Scikit-Learn documentation page[71]).

Uncertainty estimation. At each cycle, to estimate the uncertainty in the phase diagram predictions, we computed the information entropy for each point's probability vector $\boldsymbol{p}_i$ (Eq. 2). The uncertainty was computed as the (information) Entropy $H(x_i)$, and for x_i could be computed as follow:

$$H(x_i) = -\sum_{c=1}^{K} p(y = c|x_i, S) \log p(y = c|x_i, S). \quad (4)$$

Higher entropy values indicate greater uncertainty, providing an uncertainty measure for each point in the phase diagram that is representative of the prediction's confidence level. The entropy range of values is bounded, and it depends on the number of independent classes K. In all our cases, $K = 2$, leading to a range of values that goes from $H = 0$, if either of the two classes was known for certain, *i.e.* $\boldsymbol{p}_i = [1.0,0.0]$, to $H = 0.69$, if both of the two classes were most uncertain, *i.e.* $\boldsymbol{p}_i = [0.5,0.5]$.

Highest uncertainty landscape and exploration. The values of $H(x_i)$ gave direct access to the so-called *uncertainty (phase) landscape* which represented, per cycle, which areas of the design space were most (un)certain. This information was then exploited to select a subset of points $X' \subset X$ that exhibited maximal entropy, within a set range of entropy values:

$$X' = \{x_i \in X | h \leq H(x_i) \leq H_{max}\}. \quad (5)$$

In Eq. 4 the upper-bound limit, H_{max}, represented the maximum value of entropy, defined as:

$$H_{max} = -\sum_{c=1}^{K} \frac{1}{K} \log \frac{1}{K} = \log K, \quad (6)$$

which for $K = 2$ it takes the value of $H_{max} = \log 2 \approx 0.69$. The lower-bound limit can be freely chosen, and in our cases was set it to $h = 0.60$, effectively selecting only the highest uncertainty regions.

The points contained in X' would then be used as the new search space for the FPS algorithm, sampling new suitable points for refining the prediction of the phase diagram. In the context of active learning, this was often referred to as the *exploration* phase of the cycle, where new points were selected trying to maximize the exploration, lowering the overall uncertainty of the predictive algorithm.

Accuracy measurement. To assess the accuracy of our classification model in a way that accounts for class imbalance, we used a balanced accuracy metric.[55] At each cycle a set of labels $Y^{(t)} = \left\{y_i^{(t)}\right\}$ was computed for each point in our dataset from the global vector of probabilities (Eq. 3). Balanced accuracy was defined as the average of the sensitivity for each class. Given the fact that we are dealing with a

binary classification problem we can consider the 'coacervate' class as the 'positive' outcome and the 'non-aggregate' class as the 'negative' outcome.

Then, the balanced accuracy was defined as:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right). \tag{7}$$

In Equation 7, T_P and T_N refer to the true-positive and true-negative predicted labels, while F_P and F_N refer to the false-positive and false-negative predicted labels.

Convergence measurements. To monitor the convergence of the model across cycles and/or experiment replicas, we tracked changes in the phase diagram, represented by the concatenated probability vector $\boldsymbol{P}^{(t)}$ that is outputted from the GPC prediction (Eq. 3). The superscript (t) indicates a cycle specific output of the probability vector. The convergence, in terms of Jensen-Shannon divergence (JSD)[57], could be computed in two main directions: across cycles and across replicas of experiments. The former required two different probability vectors, which belonged to two consecutive cycles of the same experiment, $\boldsymbol{P}^{(t)}$ and $\boldsymbol{Q}^{(t+1)}$, and it was defined as:

$$\text{JSD}\left(\boldsymbol{P}^{(t)}||\boldsymbol{Q}^{(t+1)}\right) = \frac{1}{2}\text{KL}\left(\boldsymbol{P}^{(t)}||\boldsymbol{M}^{(t,t+1)}\right) + \frac{1}{2}\text{KL}\left(\boldsymbol{Q}^{(t+1)}||\boldsymbol{M}^{(t,t+1)}\right), \tag{8}$$

where $\boldsymbol{M}^{(t,t+1)} = \frac{1}{2}\left(\boldsymbol{P}^{(t)} + \boldsymbol{Q}^{(t+1)}\right)$ represents the midpoint distribution. Each term on the right-hand side of Eq. 8 represents the Kullback-Leibler divergence between one of the distributions and the midpoint. By computing the JSD between $\boldsymbol{P}^{(t)}$ and $\boldsymbol{Q}^{(t+1)}$ over successive iterations of the AL algorithm, we obtained a measure of convergence, with decreasing JSD values indicating stabilization of the model prediction across cycles. To compute the JSD across experiment replicas we could average the individual JSD measurements (Eq. 8) as follow,

$$\text{JSD}^{(t,t+1)}_{\text{replica } e} = \text{JSD}\left(\boldsymbol{P}^{(t)}_{\text{replica } e}||\boldsymbol{Q}^{(t+1)}_{\text{replica } e}\right), \tag{9a}$$

$$\widehat{\text{JSD}}^{(t,t+1)} = \frac{1}{E}\sum_{e=1}^{E}\text{JSD}^{(t,t+1)}_{\text{replica } e}. \tag{9b}$$

The average JSD value represented the convergence trend across multiple experimental replicas, which allowed to qualitatively account for the experimental variability.

Software and implementation. All code regarding active machine learning was written in Python 3.12. The Python packages scikit-learn (v.1.5.0) was used for the implementation of the Gaussian Process Classifier and the calculation of the balanced accuracy. SciPy (v.1.13.1) was used for the computation of the information entropy. Pandas (v.2.2.1) was used to handle the datasets. All the other operations (*e.g.*, design space creation, farthest point sampling, and convergence calculation) were carried out with custom scripts using NumPy (v.<2.0.0). For data visualization, matplotlib (v.3.8.4) and plotly (v.5.9.0) were used in combination with Adobe Illustrator.

Data availability

The raw data generated during the active machine learning cycles, as well as the processed data used to create the manuscript's figures, are available on GitHub at https://github.com/molML/activeML-navigation-of-condensate-phases. The data at the time of publishing will be available at: XXX/zenodo.org/XXX. The condensate images acquired using confocal microscopy comprise a substantial dataset and can be shared upon reasonable request to the corresponding authors.

Code availability

The Python code to replicate and extend our active machine learning framework is openly accessible on GitHub at https://github.com/molML/activeML-navigation-of-condensate-phases. The code at the time of publishing is available at: XXX/zenodo.XXX.

References

1. Diekmann, Y. & Pereira-Leal, J. B. Evolution of intracellular compartmentalization. *Biochemical Journal* **449**, 319–331 (2013).
2. Bar-Peled, L. & Kory, N. Principles and functions of metabolic compartmentalization. *Nat Metab* **4**, 1232 (2022).
3. Alberts, B. *et al.* Molecular Biology of the Cell. *Biochem Educ* **22**, 641–695 (1994).
4. Boeynaems, S. *et al.* Protein Phase Separation: A New Phase in Cell Biology. *Trends Cell Biol* **28**, 420–435 (2018).
5. Hyman, A. A., Weber, C. A. & Jülicher, F. Liquid-liquid phase separation in biology. *Annu Rev Cell Dev Biol* **30**, 39–58 (2014).
6. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology 2017 18:5* **18**, 285–298 (2017).
7. Lyon, A. S., Peeples, W. B. & Rosen, M. K. A framework for understanding the functions of biomolecular condensates across scales. *Nature Reviews Molecular Cell Biology 2020 22:3* **22**, 215–235 (2020).
8. Aguzzi, A. & Altmeyer, M. Phase Separation: Linking Cellular Compartmentalization to Disease. *Trends Cell Biol* **26**, 547–558 (2016).
9. Wan, L., Ke, J., Zhu, Y., Zhang, W. & Mu, W. Recent advances in engineering synthetic biomolecular condensates. *Biotechnol Adv* **77**, 108452 (2024).
10. Ramm, B. *et al.* Biomolecular condensate drives polymerization and bundling of the bacterial tubulin FtsZ to regulate cell division. *Nature Communications 2023 14:1* **14**, 1–24 (2023).
11. Welles, R. M. *et al.* Determinants that enable disordered protein assembly into discrete condensed phases. *Nature Chemistry 2024 16:7* **16**, 1062–1072 (2024).
12. Visser, B. S., Lipiński, W. P. & Spruijt, E. The role of biomolecular condensates in protein aggregation. *Nature Reviews Chemistry 2024 8:9* **8**, 686–700 (2024).
13. Buddingh', B. C. & Van Hest, J. C. M. Artificial Cells: Synthetic Compartments with Life-like Functionality and Adaptivity. *Acc Chem Res* **50**, 769–777 (2017).
14. Dai, Y., You, L. & Chilkoti, A. Engineering synthetic biomolecular condensates. *Nature Reviews Bioengineering 2023 1:7* **1**, 466–480 (2023).
15. Erkamp, N. A. *et al.* Biomolecular condensates with complex architectures via controlled nucleation. *Nature Chemical Engineering 2024 1:6* **1**, 430–439 (2024).

16. Song, S. *et al.* Peptide-Based Biomimetic Condensates via Liquid-Liquid Phase Separation as Biomedical Delivery Vehicles. *Biomacromolecules* **25**, 5468–5488 (2024).
17. Mitrea, D. M., Mittasch, M., Gomes, B. F., Klein, I. A. & Murcko, M. A. Modulating biomolecular condensates: a novel approach to drug discovery. *Nature Reviews Drug Discovery 2022 21:11* **21**, 841–862 (2022).
18. Ambadi Thody, S. *et al.* Small-molecule properties define partitioning into biomolecular condensates. *Nature Chemistry 2024 16:11* **16**, 1794–1802 (2024).
19. Dai, Y. *et al.* Programmable synthetic biomolecular condensates for cellular control. *Nature Chemical Biology 2023 19:4* **19**, 518–528 (2023).
20. Duro-Castano, A. *et al.* Capturing "Extraordinary" Soft-Assembled Charge-Like Polypeptides as a Strategy for Nanocarrier Design. *Advanced Materials* **29**, 1702888 (2017).
21. Liu, S. *et al.* Enzyme-mediated nitric oxide production in vasoactive erythrocyte membrane-enclosed coacervate protocells. *Nature Chemistry 2020 12:12* **12**, 1165–1173 (2020).
22. Dzuricky, M., Rogers, B. A., Shahid, A., Cremer, P. S. & Chilkoti, A. De novo engineering of intracellular condensates using artificial disordered proteins. *Nature Chemistry 2020 12:9* **12**, 814–825 (2020).
23. Chin, K. Y., Ishida, S., Sasaki, Y. & Terayama, K. Predicting condensate formation of protein and RNA under various environmental conditions. *BMC Bioinformatics* **25**, 1–14 (2024).
24. Patel, A. *et al.* Biochemistry: ATP as a biological hydrotrope. *Science (1979)* **356**, 753–756 (2017).
25. Castelletto, V., Seitsonen, J., Pollitt, A. & Hamley, I. W. Minimal Peptide Sequences That Undergo Liquid-Liquid Phase Separation via Self-Coacervation or Complex Coacervation with ATP. *Biomacromolecules* **25**, 5321–5331 (2024).
26. Nobeyama, T., Furuki, T. & Shiraki, K. Phase-Diagram Observation of Liquid-Liquid Phase Separation in the Poly(l-lysine)/ATP System and a Proposal for Diagram-Based Application Strategy. *Langmuir* **39**, 17043–17049 (2023).
27. Banani, S. F. *et al.* Compositional Control of Phase-Separated Cellular Bodies. *Cell* **166**, 651 (2016).
28. Boeynaems, S. *et al.* Phase Separation of C9orf72 Dipeptide Repeats Perturbs Stress Granule Dynamics. *Mol Cell* **65**, 1044-1055.e5 (2017).
29. Poudyal, M. *et al.* Intermolecular interactions underlie protein/peptide phase separation irrespective of sequence and structure at crowded milieu. *Nature Communications 2023 14:1* **14**, 1–21 (2023).
30. Cakmak, F. P., Choi, S., Meyer, M. C. O., Bevilacqua, P. C. & Keating, C. D. Prebiotically-relevant low polyion multivalency can improve functionality of membraneless compartments. *Nature Communications 2020 11:1* **11**, 1–11 (2020).
31. Erkamp, N. A., Qi, R., Welsh, T. J. & Knowles, T. P. J. Microfluidics for multiscale studies of biomolecular condensates. *Lab Chip* **23**, 9–24 (2022).
32. Chen, T., Lei, Q., Shi, M. & Li, T. High-throughput experimental methods for investigating biomolecular condensates. *Quantitative Biology* **9**, 255–266 (2021).
33. Nakashima, K. K., André, A. A. M. & Spruijt, E. Enzymatic control over coacervation. *Methods Enzymol* **646**, 353–389 (2021).
34. Arter, W. E. *et al.* Biomolecular condensate phase diagrams with a combinatorial microdroplet platform. *Nature Communications 2022 13:1* **13**, 1–10 (2022).
35. Bray, M. A. *et al.* Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols 2016 11:9* **11**, 1757–1774 (2016).
36. Bremer, A., Mittag, T. & Heymann, M. Microfluidic characterization of macromolecular liquid–liquid phase separation. *Lab Chip* **20**, 4225–4234 (2020).

37. Di Fiore, F., Nardelli, M. & Mainini, L. Active Learning and Bayesian Optimization: A Unified Perspective to Learn with a Goal. *Archives of Computational Methods in Engineering* **31**, 2985–3013 (2024).
38. Reker, D. & Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discov Today* **20**, 458–465 (2015).
39. van Tilborg, D. & Grisoni, F. Traversing chemical space with active deep learning for low-data drug discovery. *Nature Computational Science 2024 4:10* **4**, 786–796 (2024).
40. Khalak, Y., Tresadern, G., Hahn, D. F., De Groot, B. L. & Gapsys, V. Chemical Space Exploration with Active Learning and Alchemical Free Energies. *J Chem Theory Comput* **18**, 6259–6270 (2022).
41. Seegobin, N. *et al.* Optimising the production of PLGA nanoparticles by combining design of experiment and machine learning. *Int J Pharm* **667**, 124905 (2024).
42. Ortiz-Perez, A., van Tilborg, D., van der Meel, R., Grisoni, F. & Albertazzi, L. Machine learning-guided high throughput nanoparticle design. *Digital Discovery* **3**, 1280–1291 (2024).
43. Tamasi, M. J. & Gormley, A. J. Biologic formulation in a self-driving biomaterials lab. *Cell Rep Phys Sci* **3**, (2022).
44. Mason, A. F., Buddingh, B. C., Williams, D. S. & Van Hest, J. C. M. Hierarchical Self-Assembly of a Copolymer-Stabilized Coacervate Protocell. *J Am Chem Soc* **139**, 17309–17312 (2017).
45. Alberti, S., Gladfelter, A. & Mittag, T. Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell* **176**, 419–434 (2019).
46. Rasmussen, C. E. & Williams, C. K. I. Gaussian Processes for Machine Learning. *Gaussian Processes for Machine Learning* (2005) doi:10.7551/MITPRESS/3206.001.0001.
47. Eldar, Y., Lindenbaum, M., Porat, M. & Zeevi, Y. Y. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing* **6**, 1305–1315 (1997).
48. Tom, G. *et al.* Self-Driving Laboratories for Chemistry and Materials Science. *Chem Rev* **124**, 9633–9732 (2024).
49. Canty, R. B., Koscher, B. A., McDonald, M. A. & Jensen, K. F. Integrating autonomy into automated research platforms. *Digital Discovery* **2**, 1259–1268 (2023).
50. van Haren, M. H. I., Visser, B. S. & Spruijt, E. Probing the surface charge of condensates using microelectrophoresis. *Nature Communications 2024 15:1* **15**, 1–10 (2024).
51. Sathyavageeswaran, A., Bonesso Sabadini, J. & Perry, S. L. Self-Assembling Polypeptides in Complex Coacervation. *Acc Chem Res* **57**, 386–398 (2024).
52. Fisher, R. S. & Elbaum-Garfinkle, S. Tunable multiphase dynamics of arginine and lysine liquid condensates. *Nature Communications 2020 11:1* **11**, 1–10 (2020).
53. Ukmar-Godec, T. *et al.* Lysine/RNA-interactions drive and regulate biomolecular condensation. *Nature Communications 2019 10:1* **10**, 1–15 (2019).
54. van Tilborg, D. *et al.* Deep learning for low-data drug discovery: Hurdles and opportunities. *Curr Opin Struct Biol* **86**, 102818 (2024).
55. Thölke, P. *et al.* Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *Neuroimage* **277**, 120253 (2023).
56. Gangwal, A., Ansari, A., Ahmad, I., Azad, A. K. & Wan Sulaiman, W. M. A. Current strategies to address data scarcity in artificial intelligence-based drug discovery: A comprehensive review. *Comput Biol Med* **179**, 108734 (2024).
57. Nielsen, F. On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means. *Entropy 2019, Vol. 21, Page 485* **21**, 485 (2019).
58. Erkamp, N. A. *et al.* Multidimensional Protein Solubility Optimization with an Ultrahigh-Throughput Microfluidic Platform. *Anal Chem* **95**, 5362–5368 (2023).
59. Posey, A. E. *et al.* Biomolecular Condensates are Characterized by Interphase Electric Potentials. *J Am Chem Soc* (2024) doi:10.1021/JACS.4C08946.

60. Maruyama, B. *et al.* Artificial intelligence for materials research at extremes. *MRS Bull* **47**, 1154–1164 (2022).
61. Burger, B. *et al.* A mobile robotic chemist. *Nature 2020 583:7815* **583**, 237–241 (2020).
62. Vescovi, R. *et al.* Towards a modular architecture for science factories. *Digital Discovery* **2**, 1980–1998 (2023).
63. Chen, L., Shi, D., Kang, X., Ma, C. & Zheng, Q. Deep Learning Enabled Comprehensive Evaluation of Jumping-Droplet Condensation and Frosting. *ACS Appl Mater Interfaces* **16**, 25473–25482 (2024).
64. Hickman, R. J., Bannigan, P., Bao, Z., Aspuru-Guzik, A. & Allen, C. Self-driving laboratories: A paradigm shift in nanomedicine development. *Matter* **6**, 1071–1081 (2023).
65. Birolo, R. *et al.* Deep Supramolecular Language Processing for Co-crystal Prediction. (2024) doi:10.26434/CHEMRXIV-2024-VGVHK-V2.
66. Njirjak, M. *et al.* Reshaping the discovery of self-assembling peptides with generative AI guided by hybrid deep learning. *Nature Machine Intelligence 2024* 1–14 (2024) doi:10.1038/s42256-024-00928-1.
67. van Mierlo, G. *et al.* Predicting protein condensate formation using machine learning. *Cell Rep* **34**, (2021).
68. Fu, H. *et al.* Supramolecular polymers form tactoids through liquid–liquid phase separation. *Nature 2024 626:8001* **626**, 1011–1018 (2024).
69. Rovers, M. M. *et al.* Using a Supramolecular Monomer Formulation Approach to Engineer Modular, Dynamic Microgels, and Composite Macrogels. *Advanced Materials* 2405868 (2024) doi:10.1002/ADMA.202405868.
70. Tsang, B., Pritišanac, I., Scherer, S. W., Moses, A. M. & Forman-Kay, J. D. Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations. *Cell* **183**, 1742–1756 (2020).
71. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

Acknowledgments

This work was supported by the National Growth Fund Big Chemistry funded by the Dutch Ministry of Education, Culture and Science. We gratefully acknowledge the Institute for Complex Molecular Systems (ICMS) for providing laboratory facilities. Special thanks to the Chemical Technology IT department, particularly Tom van Teeffelen and Frank Malipaard, for their expert advice and assistance in communication networks. We also extend our gratitude to Cristina Izquierdo Lozano for her support in data management and for the insightful discussions that enriched this work.

Author information

Authors and Affiliations

Institute for Complex Molecular Systems (ICMS), Department of Biomedical Engineering, Eindhoven University of technology, PO Box 513, 5600 MB Eindhoven, The Netherlands.

Contributions

Y.H.A.L., W.H., A.G., J.L.J.D., J.C.M.H., F.G., L.B. designed the automation pipeline. Y.H.A.L., W.H., A.G., J.L.J.D. developed the automation pipeline. Y.H.A.L., W.H., A.G., J.C.M.H., F.G., L.B. designed the experiments. Y.H.A.L., W.H., A.G. performed the experiments. Y.H.A.L., W.H., A.G., J.C.M.H., F.G., L.B. analyzed the data. Y.H.A.L., W.H., A.G., J.C.M.H., F.G., L.B.

wrote the manuscript. J.C.M.H., F.G., L.B. supervised the study. All authors reviewed the manuscript.

Corresponding authors

Correspondence to J.C.M. van Hest (j.c.m.v.hest@tue.nl), F. Grisoni (f.grisoni@tue.nl), L. Brunsveld (l.brunsveld@tue.nl)

Ethics declarations

Competing interests

The authors declare no competing interests.