Machine learning-guided design of biomolecular condensates in an automated laboratory

Anonymous author [1]
[1] Paper under double-blind review

Abstract

Biomolecular condensates are phase-separated cellular compartments that regulate signaling, stress responses, and molecular sequestration. Designing synthetic condensates with specified phase behavior and material properties remains difficult due to a complex, context-dependent sequence–property landscape in human cells. Here we present a generative ML–guided design–build–test–learn loop that couples experimental measurements with machine learning to discover condensate-forming sequences. We begin with a domain-expert seed library and perform high-throughput live-cell confocal imaging across multiple cell cycles. An automated image-processing pipeline extracts functionally relevant properties, including saturation concentration, size distribution, and morphology, producing a curated sequence–property dataset. Then, we fit a multi-output Gaussian Process surrogate and use Bayesian Optimization (BO) to propose new candidate sequences, closing the loop between computation and experimentation. Our approach effectively reduces the number of iterations needed to achieve an optimal design over expensive-to-evaluate functions such as the sequence-property landscape for biomolecule condensates The work contributes experimental results, a reusable benchmark dataset, and a practical strategy for generative ML in biomolecular proteomics.

Introduction

Biomolecular condensates, membrane-less organelles formed via liquid–liquid phase separation, play critical roles in cellular such as stress sensing, signal transduction, and reaction compartmentalization [1]. Many condensates are reported to form through various phase separation processes, in which molecules concentrate into a condensed phase when their concentration exceeds the saturation concentration, C_{sat} [2]. The primary drivers of condensate formation are multivalent interactions [3], where molecules interact via multiple sites, particularly among proteins that contain intrinsically disordered regions (IDR), which lack folded structures and behave like flexible polymers (Fig. 2). Synthetic biologists have begun engineering condensates for various applications [4]. For example, the arginine-glycine-glycine (RGG)-rich domain, an IDR, from the C. elegans protein LAF-1 has been used to create synthetic condensates in budding yeast and human cells that can insulate endogenous enzymes CDC24 to control the cell cycle [5].

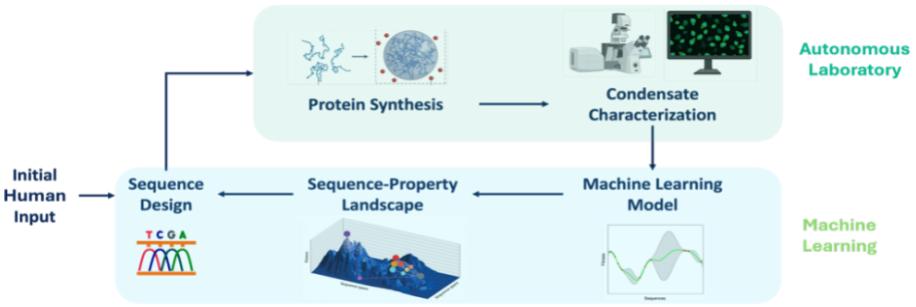

Figure 1. Workflow of ML-guided design of biomolecular condensates in an automated laboratory.

Despite the progress, several hurdles still impede the realization of the full potential of condensates in synthetic biology. For example, the complex sequence-property landscape has not been mapped, as studies on the sequence-property relationship are available only for a limited number of proteins [6]. Such limitations prevent us from rationally designing synthetic condensates with specific properties that facilitate the needed functions. To accelerate condensate engineering, we will leverage machine learning to guide our search in the sequence-property landscape and engineer condensates with targeted properties such as localization, dynamics, and protein compositions in human cells. ML–guided closed-loop experimentation offers a promising solution. A surrogate model (e.g. a Gaussian process) has demonstrated sample-efficient optimization in protein engineering [7], [8], enabling simultaneous optimization of multiple condensate properties, phase boundary, morphology, dynamics, composition, by exploiting shared structure [9], [10]. Coupling the optimization with high-throughput characterization method (i.e. live-cell imaging and lab automation) creates a closed loop for condensate discovery as well as a practical approach in an automated lab.

This abstract presents an integrated wet-lab and machine learning to engineer condensates with target phase separating behavior in human cells. Starting with a human-expert-reviewed library as the design space of interest, we will pilot the experimentation as follows: We will automate molecular cloning, cell culture, and imaging in an autonomous laboratory, deploy a multi-output GP to model saturation concentration, morphology, localization, dynamics, and client composition, and apply hierarchical BO acquisition functions to iteratively select the next batch of sequences. This process continues iteratively, refining the surrogate model until the global optimum, i.e., the desired condensate property, is reached.

Methods

Sequence library construction

Human Domain experts serve the selection of the initial sequences from a search space of interest, and automated experimentation will characterize the condensate property of interest in such a space. The space of interest consists of 10,000-sequence library by combining 1,000 IDRs with 10 coiled-coil domains. IDRs provide conformational flexibility for multivalent interactions, while coiled coils increase interaction valence through oligomerization—a design inspired by natural phase-separating sequences [11].

Wet lab workflow

Each construct comprises an IDR, a fluorescent mEGFP tag, and a coiled-coil domain. All experiments will use HeLa cells, chosen for compatibility with existing cell biology reagents and protocols. Live-cell confocal imaging across multiple cell cycles captures condensate dynamics while avoiding fixation artifacts.

Experimental automation

We propose to conduct large-scale experiments for our project using automated cloning, cell culture, and confocal imaging. After manually establishing the confocal imaging protocol, we will work with the lab to automate the process and enable remote control of the imaging process. Initial manually captured will serve to benchmark the success of the imaging setup and determine if it can generate scientifically meaningful

data. We also plan to establish automated cloning and cell culture. We will develop an automated image processing pipeline to analyze the confocal images for condensates. A machine learning-based cell segmentation model facilitates the identification of cells and the calculation of an average intensity within each cell. This intensity will then be converted into protein concentration using a calibration curve generated with purified mEGFP protein under the same imaging conditions. Automated image analysis will segment the mitotic cells with Cellpose [12], classify them as having condensates or not, and measure the fluorescent intensity within the cells to find the threshold intensity between cells with and without condensates. Plotting the condensate volume against the concentration in each cell will provide the C_{sat}. Other condensate properties such as morphology and size distribution will also be extracted by segmenting the condensates.

Machine learning model

We will use a multi-output GP that will simultaneously model condensate properties that are functionally important (Fig. 3). A few composite GP models will be trained to predict the following : (a) phase boundary – GP_{phase} will model C_{sat} data, (b) morphology – GP_{morph} will model average sphericity of condensate shapes data, (c) localization – GP_{loc} will model average Pearson correlation for co-localization between condensates and cellular structures, (d) dynamics – GP_{dyn} will model the diffusion coefficient, and (e) composition – GP_{comp} will model the number of mitotic proteins in the condensates. Each model will use a Hamming Distance Kernelization to treat sequences as strings and will provide predictions of the unknown sequence-to-property landscape in terms of the mean and covariance. We will optimize each property simultaneously while using the shared information and structure to accelerate the optimization [10]. Additionally, we will explore a hierarchical design [7], where a subset of the GP models is used to filter out sequences (e.g., those with a certain C_{sat}), and only considers the remaining sequences for the next round.

Results

ML-guided design framework

We developed the closed design–build–test–learn loop for biomolecular condensate generation in our autonomous laboratory (Fig. 1). With a human expert reviewed input, the framework explores all possible sequence design by initial selection (empirical representativeness) → sequence selection (BO) → synthesis/imaging (automated lab) → property extraction (image analysis) → model update (GP). The framework has been demonstrated end-to-end in manual mode. Corresponding automation of cloning and imaging is in progress. Preliminary GP models trained on pilot data successfully predict C_{sat}, demonstrating the viability our approach of biomolecular condensate.

Pilot experimentation

We have manually synthesized six IDR–coiled-coil constructs and imaged them in HeLa cells. Five of six formed condensates in mitotic cells, with diverse phenotypes: construct I384-C1 exhibited spindle-pole localization (two condensates per cell), while I271-C1 showed no condensate formation (Fig. 2). This diversity confirms that our sequence library spans a range of phase behaviors and validates the feasibility of live-cell phenotyping.

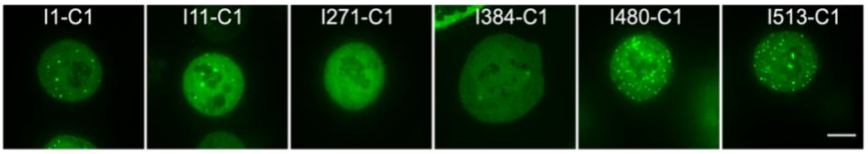

Figure 2. Representative images of pilot constructs. The label Ix-Cy indicates the sequence is a fusion of IDR number x with coiled coil number y. Mitotic HeLa cells. Scale bar, 10 μm.

Property landscape modeling

The proposed multi-output GP captures how IDR and coiled-coil composition influences important condensate properties: phase boundary, morphology, localization, dynamics and composition (Fig. 3). By diving deep into the learnt mapping between the sequence space of interest and property space of interest, it further helps the understanding the inverse correlation with sequence diversity and droplet properties [13]. Such an interpretable landscape enables rational prediction of sequences likely to exhibit desired phenotype combinations. For instance, sequences with intermediate C_{sat}, high sphericity, fast dynamics and stable composition can be prioritized by our inferred landscape for applications requiring mobile, responsive condensates, which is beneficial for drug delivery purpose. This interpretable, multi-property modeling approach accelerates condensate design by reducing reliance on expensive, exhaustive screening and enabling hypothesis-driven sequence optimization grounded in learned biophysical rules.

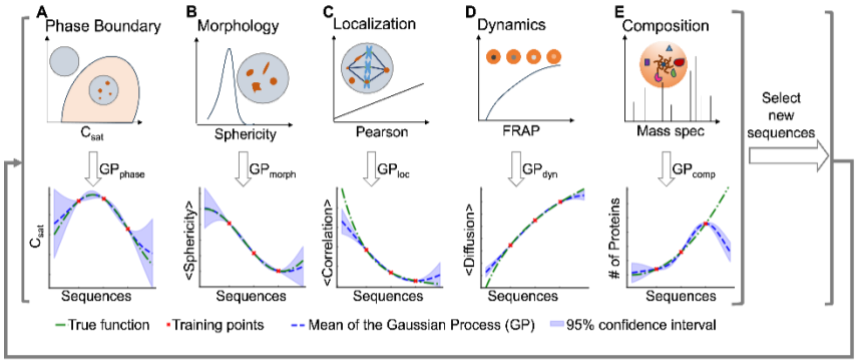

Figure 3. Landscape mapping with iterations between condensate characterization and Bayesian Optimization model.

Conclusion and Discussion

Biomolecular condensates are phase-separated cellular compartments that regulate signaling, stress responses, and molecular sequestration. Design and engineering biomolecular condensates provide novel aspects for therapeutics. In this extended abstract we present a machine learning guided generative framework to discover condensate-forming sequences with target properties. Pilot validation of six constructs of condensates confirms the feasibility of our approach. It reduces expensive experimental iterations while improving predictions across the sequence–property landscape, contributing experimental results, a benchmark dataset, and a practical ML-guided design strategy for biocondensates synthesis. Current limitations include the small pilot dataset and partial automation; scaling to full autonomous operation and expanding the sequence library will be key next steps. Future work will also refine acquisition strategies for multi-objective optimization and evaluate transferability across more diversified cellular contexts.

References

[1] S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen, "Biomolecular condensates: organizers of cellular biochemistry," Nat. Rev. Mol. Cell Biol., vol. 18, no. 5, pp. 285–298, May 2017, doi: 10.1038/nrm.2017.7.

[2] T. Mittag and R. V. Pappu, "A conceptual framework for understanding phase separation and addressing open questions and challenges," Mol. Cell, vol. 82, no. 12, pp. 2201–2214, Jun. 2022, doi: 10.1016/j.molcel.2022.05.018.

[3] P. Li et al., "Phase transitions in the assembly of multivalent signalling proteins," Nature, vol. 483, no. 7389, pp. 336–340, Mar. 2012, doi: 10.1038/nature10879.

[4] Y. Dai, L. You, and A. Chilkoti, "Engineering synthetic biomolecular condensates," Nat. Rev. Bioeng., vol. 1, no. 7, pp. 466–480, Jul. 2023, doi: 10.1038/s44222-023-00052-6.

[5] M. V. Garabedian et al., "Designer membraneless organelles sequester native factors for control of cell behavior," Nat. Chem. Biol., vol. 17, no. 9, pp. 998–1007, Sep. 2021, doi: 10.1038/s41589-021-00840-4.

[6] C. Mayr, T. Mittag, T.-Y. D. Tang, W. Wen, H. Zhang, and H. Zhang, "Frontiers in biomolecular condensate research," Nat. Cell Biol., vol. 25, no. 4, pp. 512–514, Apr. 2023, doi: 10.1038/s41556-023-01102-2.

[7] J. T. Rapp, B. J. Bremer, and P. A. Romero, "Self-driving laboratories to autonomously navigate the protein fitness landscape," Nat. Chem. Eng., vol. 1, no. 1, pp. 97–107, Jan. 2024, doi: 10.1038/s44286-023-00002-4.

[8] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," Aug. 29, 2012, arXiv: arXiv:1206.2944. doi: 10.48550/arXiv.1206.2944.

[9] L.-F. Cheng et al., "Sparse multi-output Gaussian processes for online medical time series prediction," BMC Med. Inform. Decis. Mak., vol. 20, no. 1, p. 152, Jul. 2020, doi: 10.1186/s12911-020-1069-4.

[10] K. Swersky, J. Snoek, and R. P. Adams, "Multi-Task Bayesian Optimization," in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2013. Accessed: Feb. 04, 2026. [Online]. Available: https://papers.nips.cc/paper_files/paper/2013/hash/f33ba15effa5c10e873bf3842afb46a6-Abstract.html

[11] D. W. Sanders et al., "Competing protein-RNA interaction networks control multiphase intracellular organization," Cell, vol. 181, no. 2, pp. 306-324.e28, Apr. 2020, doi: 10.1016/j.cell.2020.03.050.

[12] C. Stringer and M. Pachitariu, "Cellpose3: one-click image restoration for improved cellular segmentation," Nat. Methods, vol. 22, no. 3, pp. 592–599, Mar. 2025, doi: 10.1038/s41592-025-02595-5.

[13] T. Li et al., "Interpretable Active Learning Identifies Iron-Doped Carbon Dots With High Photothermal Conversion Efficiency for Antitumor Synergistic Therapy," Aggregate, vol. 6, no. 7, p. e70060, 2025, doi: 10.1002/agt2.70060.