# Integrating AI, automation and multiscale simulations for end-to-end design of phase-separating proteins

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Liquid-liquid phase separation (LLPS) is a fundamental cellular process that is driven by self-assembly of intrinsically disordered proteins (IDPs), protein-RNA complexes, or other bio-molecular systems which can form liquid droplets. Many natural materials including silk, elastin, and gels are a result of LLPS and thus rational design of such phase-separating peptides can have transformative impact, from designing new biologically inspired materials (e.g., clothing) to self-compartmentalized drug-delivery systems for biomedical applications. However, given the intrisinc complexity in the rules governing LLPS, rational design of LLPS undergoing peptides remains challenging. We posit that automation, foundation models integrated with reinforcement learning approaches and multiscale molecular simulations can drive the design of novel peptides that undergo LLPS. We describe our progress towards the goal of end-to-end design of phase separating peptides by summarizing current work at the Argonne National Laboratory's Advanced Photon Source 8ID-I beamline, where a robotic set up in the laboratory is enabled via simulation and extensive testing of such bio-materials. Together, our approach enables the design of novel bio-materials that can undergo phase separation under diverse physiological conditions.

## 1 Introduction

Phase separation in biology is now being widely acknowledged as a fundamental mechanism of cellular control, including cellular compartmentalization as well as in various diseases such as cancer Hyman et al. [2014]. More importantly, several naturally available proteins such as elastin Rodríguez-Cabello et al. [2018], silk Lemetti et al. [2022], Parker et al. [2019], and others are known to undergo phase separation which is likely to influence their overall stability and function inside of cells. Given that phase separation within such proteins is dependent on their (polymer) length, sequence specific linear (amino-acid) motifs, and other factors, a natural question is then in engineering novel constructs of such phase separating peptides/proteins that can possess specific properties Hyman et al. [2014].

Previous studies have examined how sequence composition and polymer length affect phase separation properties in elastin-like polypeptides (ELPs) Christensen et al. [2013]. However, given the diversity of such sequences and the specific linear motifs that they need to phase separate under physiological conditions (e.g., ELPs utilize -(VPXVG)$_n$- motif interspersed with other amino-acid sequences), the combinatorial complexity of the design space entails that an exhaustive evaluation of even a single class of phase separating peptides can be daunting, tedious, and error-prone. Furthermore, the discovery of new phase separating peptides/proteins with diverse mechanisms of self-assembly, there is a need to develop robust experimental and computational workflows that can probe and quantify how phase separation leads to different behaviors under diverse physiological conditions.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

We posit that robotics and automation within the laboratory integrated tightly with artificial intelligence (AI) methods, including generative models and reinforcement learning (RL) can provide an effective platform for not only characterizing phase separation mechanisms, but also in designing novel peptides/proteins that undergo controlled phase separation under diverse conditions. As shown in Fig. 1, our automated design platform at Argonne National Laboratory integrates high-performance computing systems within the Argonne Leadership Computing Facility (ALCF) with the Advanced Photon Source (APS) beamline for characterizing phase separating proteins using x-ray photon correlation spectroscopy (XPCS) and the Advanced Protein Characterization Facility (APCF) to clone, express, and purify protein samples at scale. At the heart of this *self driving lab* is a computational engine that consists of a suite of generative AI models that has been trained on diverse genome-scale data using large language models and fine-tuned on phase-separating protein databases. A RL approach is used to guide the precise modifications to the protein sequence that can predict specific phase separation properties. These are fed into a multiscale simulation framework that uses enhanced sampling techniques guided by AI approaches, namely, DeepDriveMD Brace et al. [2022], Casalino et al. [2021] to characterize molecular interactions that control phase separation. This approach lets us screen over $10^5$ design candidates rapidly, while the APCF can automate the screening of $10^2$-$10^3$ sequence designs. The refined designs (about $10^2$) are then characterized for phase separation at APS-8-ID-I beamline under physiological conditions and the observations are automatically 'piped' through training the AI approaches (so that the design space can be constrained and conditioned appropriately). We provide an overview of progress in developing each of the areas highlighted.

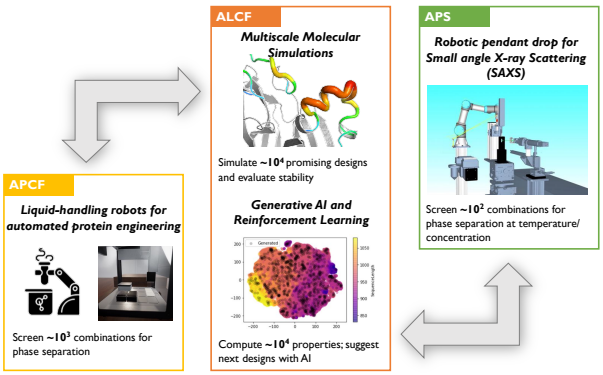


Figure 1: **An overview of our automated platform for designing phase-separating proteins.** Protein engineering is enabled via liquid handling robots enabling screening of $10^3$ peptides. These are fed into simulation and AI workflows that automatically suggest new protein designs for subsequent rounds. Finally, a smaller set of protein designs (about $10^2$) are characterized using X-ray scattering approaches.

## 2 AI-enabled phase-separating protein design and multiscale simulations

### 2.1 Reinforcement learning and generative sequence models

We formulate the design of a single sequence $x$ using reinforcement learning where a policy is trained to optimize a specific objective Sutton and Barto [2018], Silver et al. [2016]. We initialize a policy $\pi = \rho$, where $\rho$ is a pretrained generative language model providing the conditional probability distribution to predict the next tokens in the sequence. The initialized policy $\pi$ is then fine-tuned using RL to perform the protein sequence-specific generation task. Combination of RL and language models have been successful in the past, where RL models were applied to fine-tune pre-trained language models for tasks such as text continuation with positive sentiment or physically descriptive language and summarization Ziegler et al. [2019]. In our work, $\rho$ is obtained using GPT-NeoX Black et al. [2022] trained on diverse protein sequence datasets and fine-tuned on the phase-separating protein databases containing $\sim$6K sequences You et al. [2020].

We employ the proximal policy optimization (PPO) algorithm Schulman et al. [2017] as the RL model. The PPO policy guides the agent's actions which in this case is to insert an amino-acid token from the sequence model vocabulary. The vocabulary consists of 21 amino-acid tokens and other special tokens as part of the tokenization process. In the initial set of experiments for generating novel sequences, the reward structure is simplified such that the reward structure benefits insertion of (valid) amino-acid tokens and penalized for adding special tokens, with a maximum length of 512 tokens. Our experiments suggest (Fig. 2) that the RL training results in generating novel sequences.

## 2.2 Multiscale Simulations

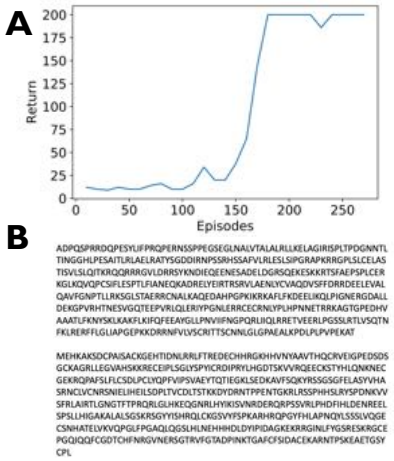


Figure 2: **RL-generated examples of phase-separating proteins.** (A) Returns (expected sum of rewards) are shown to improve with episodes suggesting that the PPO policy in combination with the GPT-NeoX model learns to generate protein sequences by guiding the agent's actions to insert (valid) amino-acid tokens. (B) Examples of novel protein sequences generated after the training with a maximum sequence length of 512 tokens.

We conducted large-scale replica exchange molecular dynamics (REMD) simulations of LLPS phase-separation in the generated peptides undergoing Dignon et al. [2019] to characterize the inter- and intra-molecular interactions that influence LLPS. We separately simulated (i) diffusion of the individual peptides through explicit solvent and (ii) closer-range interactions amongst multiple peptides(i.e., peptide aggregation) in explicit solvent.

The peptides consist of 70-150 amino acids and initially simulated in implicit solvent (see below) for 20ns to reach stable equilibrium conformations; equilibration of the RMSD and radius of gyration occurred within 20ns of simulation. Individual peptides were then simulated in an explicit solvent model (i.e., water and 150mM NaCl in a box providing $\geq$ 2nm padding around the peptide). The peptide and explicit solvent contained roughly $10^5$ atoms. Multiple peptides were similarly simulated in an explicit solvent model,pandas except they were arranged in a 3 x 3 x 3 cuboid configuration with 10nm center-to-center distance between adjacent peptides. The multiple peptide systems contained roughly $10^6$ atoms. These explicit solvent systems used the ff99sb.ILDN force field and TIP3P water model Lindorff-Larsen et al. [2010].

Replica exchange simulations were carried at 64 temperatures between 279.15 and 450 K. Each replica used the Langevin integrator with $1.0ps^-1$ collision rate and 0.004ps time-steps. The replicas were integrated for 2ps between each attempted exchange. The short-range electrostatic interactions and Lennard-Jones interactions were evaluated using a cutoff of 0.9nm. Particle-mesh Ewald summation was used to calculate the long-range electrostatic interactions with an Ewald error tolerance of 0.0005 and Hydrogen mass repartitioning to 1.5amu (to accelerate the integration). Preliminary benchmarking of the REMD give 930ns/day of simulation for systems having $10^5$ atoms on 8 GPU (NVIDIA A100 cards).

# 3 Automated phase-separating protein engineering

The DNA fragments encoding the selected peptide repeats were generated by the overlap-extension rolling circle amplification (OERCA) method Amiram et al. [2011]. The generated clones were sequenced and those with repeats of 20 or more peptides were selected for characterization. The N-terminal $His_6$-tagged proteins were purified using immobilized metal-affinity chromatography and used without the removal of the purification tag. The phase transition of the proteins was measured by either monitoring absorbance at 350nm in a plate reader or monitoring interaction with a fluorescent dye via a real-time PCR detection system (Fig. 3). The phase transition temperature was lower for polypeptides with fewer repeats. Similarly, lower protein concentration, the addition of NaCl or PEG-8000 to the solution also resulted in lower phase transition temperatures. A fully automated system, driven by a Python API, can be used to assemble the various combinations or polypeptides and additives. Current efforts focus on the development of a fully automated closed-loop system capable not only measuring phase transition of a given input

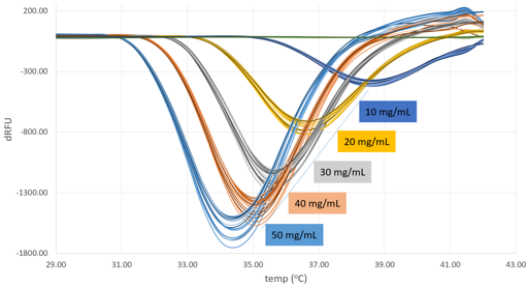


Figure 3: Example phase transition measurement at the Advanced Protein Characterization Facility.

sample without human intervention, but eventually also carry out the synthesis, cloning, and protein purification steps of the workflow.

## 4 Robotic pendant drop enabled small-angle scattering experiments

Robotic pendant drop setup was developed in the adjacent chemistry laboratory of beamline 8-ID-I of Advanced Photon Source and robot programs were implemented on a simulation software RoboDK [2022]. To perform the pendant drop experiments Bera and Antonio [2016], UR3e collaborative robot arm from Universal Robots was utilized as the liquid handling robot. UR3e robot was ideal to operate in tight workspaces such as the beamline, due to its compactness and small footprint. In order to create precise droplets and eliminate the vibration factor, the experimental setup was designed with a pipette docking location. Furthermore, a tool changer (ATI QC-11) was attached to both the end joint of the robot and an Opentrons Single Channel P300 GEN2 electronic pipette. The tool changer lets us lock and unlock its Master and Tool sides with air compression to pick and place the pipette. The pipette was driven to accurately control the volume of the liquid aspirated and dispensed inside the tips and ejecting the tips when needed. An optical microscope was placed by the sample location to provide live video feed of the sample via the reflection of the 45-degree mirror. The mirror has a 1 mm through-hole at its center, allowing x-ray beam to pass through so that optical inspection and x-ray measurements can be performed simultaneously.

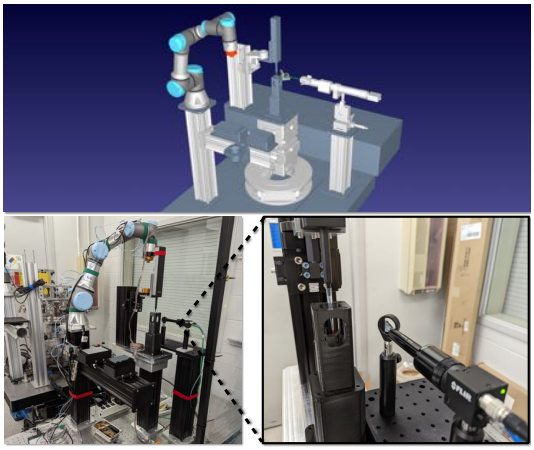

Figure 4: Robotic Setup in The Chemistry Laboratory of Beamline 8-ID-I of APS/Argonne.

RoboDK allows the users to create robot trajectories in the simulation environment as well as execute them on the real robot. The pipette, tool changer and optical microscope were controlled via the DOE light source software (EPICS) Kraimer et al. [2012]. Droplet experiment was simulated in multiple steps by using the 3D CAD drawing of the beamline 8-ID-I of Advanced Photon Source. The experiment starts by locating the robot to the home base, followed by picking up the pipette from the docking location. Subsequently, robot attaches a tip to the pipette from the tip bin and prepares the sample on the 96 well plate by driving the pipette. To obtain the measurements with the prepared sample, the pipette was placed on to the docking location and a droplet is formed by dispensing the sample. Lastly, the pipette is picked up from the docking location to eject the tip to the trash bin and placed back to the docking location, respectively. Fig. 4 shows RoboDK is executing the program on the UR3e, and the robot is performing liquid handing and sample exchange.

## 5 Summary

We have highlighted our progress in developing a self-driving laboratory for designing phase separating proteins. Our approach uses robotics integrated in a functioning beamline (to characterize size and dynamics of phase separation) with AI/ML techniques and high-throughput molecular simulations. The approach also highlighted some important lessons that we learned, including the challenges involved in integrating diverse robotic systems and how such ecosystems of commercial off-the-shelf robotic systems integration can be carried out across at a user facility. Further, it also highlighted the importance of building robust, automated workflow systems that can be used to enable high-throughput bio-materials characterization. Finally, it also helped us reduce the time-to-solution for design cycles of phase separating proteins from several months to about weeks – thus allowing a much rapid exploration of the design space of such materials.

## References

Miriam Amiram, Felipe Garcia Quiroz, Daniel J. Callahan, and Ashutosh Chilkoti. A highly parallel method for synthesizing dna repeats enables the discovery of 'smart'protein polymers. *Nature*

*Materials*, 10(2):141–148, 2011. doi: 10.1038/nmat2942. URL https://doi.org/10.1038/nmat2942.

Mrinal K Bera and Mark R Antonio. Crystallization of keggin heteropolyanions via a Two-Step process in aqueous solutions. *J. Am. Chem. Soc.*, 138(23):7282–7288, June 2016.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.

Alexander Brace, Igor Yakushin, Heng Ma, Anda Trifan, Todd Munson, Ian Foster, Arvind Ramanathan, Hyungro Lee, Matteo Turilli, and Shantenu Jha. Coupling streaming ai and hpc ensembles to achieve 100–1000× faster biomolecular simulations. In *2022 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 806–816, 2022. doi: 10.1109/IPDPS53621.2022.00083.

Lorenzo Casalino, Abigail C Dommer, Zied Gaieb, Emilia P Barros, Terra Sztain, Surl-Hee Ahn, Anda Trifan, Alexander Brace, Anthony T Bogetti, Austin Clyde, Heng Ma, Hyungro Lee, Matteo Turilli, Syma Khalid, Lillian T Chong, Carlos Simmerling, David J Hardy, Julio DC Maia, James C Phillips, Thorsten Kurth, Abraham C Stern, Lei Huang, John D McCalpin, Mahidhar Tatineni, Tom Gibbs, John E Stone, Shantenu Jha, Arvind Ramanathan, and Rommie E Amaro. Ai-driven multiscale simulations illuminate mechanisms of sars-cov-2 spike dynamics. *The International Journal of High Performance Computing Applications*, 35(5):432–451, 2021. doi: 10.1177/10943420211006452. URL https://doi.org/10.1177/10943420211006452.

Trine Christensen, Wafa Hassouneh, Kimberley Trabbic-Carlson, and Ashutosh Chilkoti. Predicting transition temperatures of elastin-like polypeptide fusion proteins. *Biomacromolecules*, 14(5): 1514–1519, 05 2013. doi: 10.1021/bm400167h. URL https://doi.org/10.1021/bm400167h.

Gregory L Dignon, Wenwei Zheng, and Jeetain Mittal. Simulation methods for liquid–liquid phase separation of disordered proteins. *Current Opinion in Chemical Engineering*, 23:92–98, 2019. ISSN 2211-3398. doi: https://doi.org/10.1016/j.coche.2019.03.004. URL https://www.sciencedirect.com/science/article/pii/S2211339818300807. Frontiers of Chemical Engineering: Molecular Modeling.

Anthony A. Hyman, Christoph A. Weber, and Frank Jülicher. Liquid-liquid phase separation in biology. *Annual Review of Cell and Developmental Biology*, 30(1):39–58, 2014. doi: 10.1146/annurev-cellbio-100913-013325. URL https://doi.org/10.1146/annurev-cellbio-100913-013325. PMID: 25288112.

Martin R Kraimer, Janet B Anderson, Andrew N Johnson, W Eric Norum, Jeffrey O Hill, Ralph Lange, Benjamin Franksen, and Peter Denison. Epics application developer's guide. *EPICS Base Release*, 3(11):1–243, 2012.

Laura Lemetti, Alberto Scacchi, Yin Yin, Mengjie Shen, Markus B. Linder, Maria Sammalkorpi, and A. Sesilja Aranko. Liquid–liquid phase separation and assembly of silk-like proteins is dependent on the polymer length. *Biomacromolecules*, 23(8):3142–3153, 08 2022. doi: 10.1021/acs.biomac.2c00179. URL https://doi.org/10.1021/acs.biomac.2c00179.

Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, 2010. doi: https://doi.org/10.1002/prot.22711. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.22711.

Rachael N. Parker, Wenyao A. Wu, Tina B. McKay, Qiaobing Xu, and David L. Kaplan. Design of silk-elastin-like protein nanoparticle systems with mucoadhesive properties. *Journal of Functional Biomaterials*, 10(4), 2019. ISSN 2079-4983. doi: 10.3390/jfb10040049. URL https://www.mdpi.com/2079-4983/10/4/49.

RoboDK. Robodk: Simulate robot applications., 2022. URL https://robodk.com/.

Jose C. Rodríguez-Cabello, Israel González de Torre, Sergio Acosta, Soraya Salinas, and Marcos Herrero. 4 - elastin-like proteins: Molecular design for self-assembling. In Helena S. Azevedo and Ricardo M.P. da Silva, editors, *Self-assembling Biomaterials*, Woodhead Publishing Series in Biomaterials, pages 49–78. Woodhead Publishing, 2018. ISBN 978-0-08-102015-9. doi: https://doi.org/10.1016/B978-0-08-102015-9.00004-6. URL https://www.sciencedirect.com/science/article/pii/B9780081020159000046.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, and Laurent Sifre. George van den driessche et al., mastering the game of go with deep neural networks and tree search. *Nature*, 529: 7587, 2016.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Kaiqiang You, Qi Huang, Chunyu Yu, Boyan Shen, Cristoffer Sevilla, Minglei Shi, Henning Hermjakob, Yang Chen, and Tingting Li. Phasepdb: a database of liquid–liquid phase separation related proteins. *Nucleic acids research*, 48(D1):D354–D359, 2020.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.