DESIGN OF PHASE-SEPARATING BIOSYSTEM VIA JOINT DIFFUSION AND POSITIVE-UNLABELED GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Liquid-liquid phase separation (LLPS) is a widespread mechanism by which cells organize their internal environment, leading to the formation of membraneless organelles and biomolecular condensates. The ability to design to generate synthetic condensates properties has significant implications for biotechnology and medicine. While current studies focused on protein-driven phase separation, it is now clear that many condensates are multi-component systems, including proteins, DNA, and RNA. Hereby, We propose a joint diffusion framework that leverages the compositional generative process in and guide a joint generation with a positive-unlabeled (PU) assumption. Experiments on our newly curated multi-component phase separation dataset demonstrate the efficacy of generating unseen biosystems with desired phase behaviors. Our approach demonstrates the feasibility of *in silico* designing and engineering multi-component phase-separating biosystems.

1 INTRODUCTION

Liquid-liquid phase separation (LLPS) enables cells to compartmentalize their internal spaces through the assembly of membraneless organelles and biomolecular condensates Banani et al. (2017); Alberti et al. (2019); Xu et al. (2024). Recent discoveries reveal that many condensates comprise multiple components beyond proteins, including DNA and RNA molecules, and multiple interactions Laflamme & Mekhail (2020); Gu et al. (2022); Welles et al. (2024). Such interplay between these biomolecules, together with their sequence patterns and physicochemical properties, gives rise to complex phase behaviors that are essential for diverse cellular functions and have been implicated in diseases Ding et al. (2024). Despite this importance, designing phase-separating biosystems with specific properties remains exceptionally challenging due to the intricate interplay of multiple factors and the lack of predictive models. The complexity of these phase separating system has motivated the development of novel approaches for generation and design.

Despite rapid advances in protein design methods, phase separation remains an understudied target property, with most design efforts focusing on structure design, property optimization, and function rather than phase behavior modulation Kortemme (2024); Koh et al. (2025). In addition, existing design methods primarily focus on homogeneous systems with single proteins, which limits their applicability to more complex, multi-component systems Dallago et al. (2021); Koh et al. (2025). However, proteins rarely act alone in biological contexts, and many phase-separating condensates comprise multiple biomolecules with very different properties and interactions Laflamme & Mekhail (2020). The ability of such a biocondensate system to undergo phase separation is believed to be encoded in the components themselves, while extrinsic conditions can modulate the phase outcome to a large extent Hardenberg et al. (2020); Singh (2024). Therefore, there is a critical need for design methods that can account for multiple components and their interactions to enable the rational design of phase-separating biosystems with controllable phase behaviors.

While the design and generation of individual proteins is now mature Chronowska et al. (2025), the challenge lies in rationally orchestrating multiple distinct biomolecules to achieve desired collective phase behavior Alberti et al. (2019). There are a few key challenges to overcome in such design tasks. First, it is extremely hard to acquire well-annotated datasets with multi-component phase separation behaviors under diverse experimental conditions, which are essential for training and evaluation at system-level. Jiang et al. (2025); Raymond et al. (2025). Second, the design is

actually performed at a positive-unlabeled space, in which only a very small subset of biomolecular combinations are known to phase separate at given experiments, while the vast majority of combinations lack experimental characterization Jiang et al. (2023); Raymond et al. (2025). This positive-unlabeled setting makes it difficult to distinguish true phase-separating combinations from non-phase-separating ones. Third, experimental conditions are critical factors that modulate phase behavior, as the ability of a system to undergo phase separation depends not only on component properties but also on extrinsic factors like salt concentration, temperature, and pH. Accounting for these conditions in the design process is essential for achieving desired phase behaviors Wang et al. (2022); Arter et al. (2022); Raimondi et al. (2021). Therefore, we must consider these conditions when conducting designing.

To address the aforementioned challenges, we propose to leverage existing mature single-component (protein, RNA, DNA) design methods, such as diffusion models or flow match models Meshchaninov et al. (2024); Geffner et al. (2025) and putting our efforts on how to guide the joint samplings and generation of multi-component phase-separating systems in the positive-unlabeled space Raymond et al. (2025). More specifically, components from compositional generative models serve as building blocks that can be combined to form complex biosystems with desired phase behaviors. PU guidance based on the recorded pair-wise experimental outcomes steers the joint generation process towards optimal direction. Our approach paves the way for rational design of biocondensates.

2 METHOD

2.1 PROBLEM FORMULATION

We formulate the exploration of phase-separating biosystems within a unified context so that we can model the interplay among three key variables:

- **System Components ($\boldsymbol{x}$):** A set of representations of the components. Common representations include, but are not limited to, sequence one-hot vector, intrinsic biophysical characteristics, and language model embeddings derived from pretrained models. The feature space of component i is denoted as $\mathcal{X}_i$.
- **Experimental Conditions ($\boldsymbol{c}$):** A set of representations of extrinsic environmental factors like salt concentration, temperature, solute concentration, crowding agent, pH and etc. These conditions are well curated and normalized to ensure consistency across different experiments. The feature space of experimental conditions is denoted as $\mathcal{C}$.
- **Phase Outcome (y):** A binary variable $y \in \mathcal{Y} = \{0, 1\}$ (or a quantity $y \in \mathcal{Y} = [0, 1]$ from a continuous measure from phase experiments) indicating whether (or how intense) the system undergoes phase separation (dual-phase, $y = 1$) or remains in a single phase ($y = 0$) under given experiment conditions.

A two-component biosystem design task can be formulated as generating a pair of components $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X}_1 \times \mathcal{X}_2$ (product space of two component feature spaces) that can undergo phase separation ($y = 1$) or can not be phase-separating ($y = 0$).

2.2 JOINT DIFFUSION MODEL

Diffusion models are generative models that learn to reverse a gradual noise corruption process, enabling the generation of complex data by iteratively denoising random noise guided by a learned score function. They have demonstrated remarkable success across various domains including image generation, molecular design, and protein structure prediction Yang et al. (2023). We adopt a score-based diffusion framework to model the joint generation of multi-component biosystems. The score function $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$ characterizes the gradient of the log probability density, guiding the generative process towards high-probability regions of the component space. For a two-component system, we define the joint score function as $\mathbf{s}_\theta(\boldsymbol{x}_1, \boldsymbol{x}_2, t) = \nabla_{(\boldsymbol{x}_1, \boldsymbol{x}_2)} \log p(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})$ at diffusion timestep t. During the reverse process, we iteratively denoise the randomly initialized components by following the score function:

$$\mathrm{d}\boldsymbol{x} = -\frac{\beta(t)}{2} \underbrace{\mathbf{s}_\theta(\boldsymbol{x}_1, \boldsymbol{x}_2, t)}_{\text{joint score function}} \mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\mathbf{w}_t \tag{1}$$

where $\beta(t)$ is the noise schedule and $\mathbf{w}_t$ is a Wiener process. The score function is trained by minimizing the score matching objective, enabling efficient sampling Song et al. (2020).

2.3 POSITIVE-UNLABELED GUIDANCE

While diffusion models can generate diverse multi-component biosystems by sampling from the learned data distribution, they lack the ability to steer the generation process towards systems with desired phase separation properties. For targeted (or conditional) we need to guide the generation towards specific outcomes. A guided diffusion generation leverages classifier feedback to preferentially sample from regions of the component space that are more likely to exhibit target phase behaviors. The classifier guidance technique leverages Bayes' rule to decompose the conditional score function:

$$p(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t}|y) = \frac{p(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})p(y|\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})}{p(y)} \tag{2}$$

Taking the gradient of the log probability, the conditional score can be expressed as:

$$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t}|y) = \underbrace{\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})}_{\text{unconditional score, } \mathbf{s}_\theta(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})} + \underbrace{\nabla_{\boldsymbol{x}} \log p(y|\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})}_{\text{classifier guidance}} \tag{3}$$

The first term $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})$ is the unconditional score from the diffusion model, while the second term $\nabla_{\boldsymbol{x}} \log p(y|\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})$ is the adversarial gradient from a classifier trained to predict the phase outcome. In practice, we usually guide the generation towards positive samples ($y = 1$) that exhibit phase separation.

However, in many real-world scenarios, especially in biological experiments, we often encounter positive-unlabeled (PU) data, where only a subset of positive samples are labeled, while the rest of the data remains unlabeled. In detail, let $p_1(\boldsymbol{x}_1, \boldsymbol{x}_2)$ denote the distribution of positive pairs (phase-separating systems) in corresponding feature space, and $p_0(\boldsymbol{x}_1, \boldsymbol{x}_2)$ denote the equivalent distribution of negative pairs (non-phase-separating systems). The overall data distribution can be expressed as a mixture:

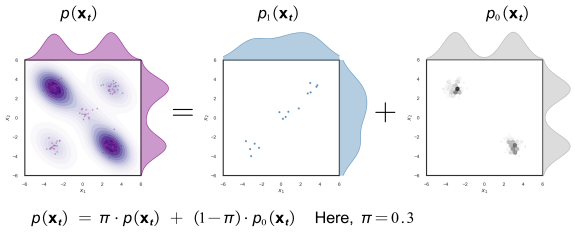

Figure 1: Illustration of the decomposition of joint distribution, positive component and negative component. The unlabeled data by production is independently drawn from the mixture distribution of positive and negative components.

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2) = \pi p_1(\boldsymbol{x}_1, \boldsymbol{x}_2) + (1 - \pi)p_0(\boldsymbol{x}_1, \boldsymbol{x}_2) \tag{4}$$

where π is the prior probability of positive samples in the dataset. Based on the insights from Ivanov (2020); Zeiberg et al. (2020); Raymond et al. (2025), Eq. 3 can be adapted to the PU setting as:

$$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t}|y = 1) = \mathbf{s}_\theta(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t}) + \nabla_{\boldsymbol{x}} \log \frac{p_1(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})}{p(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})} \tag{5}$$

Based on the assumption that the if without phase label, the product distribution should be a mixture of positive and negative distributions and also , i.e., $p(\boldsymbol{x}_{1,t}) \cdot p(\boldsymbol{x}_{2,t}) = \pi p_1(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t}) + (1 - \pi)p_0(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})$. Then training a non-traditional classifier to (PU classifier) discriminate P and U $h(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})$ could avoid the class prior estimation problem in the joint space Raymond et al. (2025) and further simplify the PU-guided score function as,

$$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t}|y = 1) = \underbrace{[\mathbf{s}_\theta(\boldsymbol{x}_{1,t}) \;\; \mathbf{s}_\theta(\boldsymbol{x}_{2,t})]^\top}_{\text{independent score function}} + \underbrace{\nabla_{\boldsymbol{x}} \log \frac{h(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})}{1 - h(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})}}_{\text{positive-unlabeled guidance}} \tag{6}$$

To implement the PU-guided diffusion, we retrain diffusion models to match component scores $\mathbf{s}_\theta(\boldsymbol{x}_{\cdot,t})$ from the corresponding dataset, and train a nested PU classifier $\hat{h}(\boldsymbol{x}_t) = \hat{h}(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})$ by minimizing the empirical cross entropy loss to provide the guidance term in Eq. 6.

3 EXPERIMENTS

We conduct experiments on the newly curated multi-component phase separation dataset to evaluate the effectiveness of the joint diffusion model with PU guidance in designing phase-separating biosystems. Taking "protein(1)-RNA" systems as the system of interest. We examine the single component generation quality, the ability to generate multi-component systems that exhibit desired phase behaviors, and the impact of PU guidance on the generation process.

3.1 DATASET CURATION

Table 1: General component statistics in *db1*, *db2* and *db3*

Dataset	# Entry	# Protein	# DNA	# RNA
db1	6,005	546	102	279
db2	1,188	14	0	3
db3	1,514	37	0	147

We curate a standardized dataset of multi-component phase-separating systems with diverse experimental conditions, integrating annotations of components, conditions, and phase outcomes from three sources. The *db1* dataset from LLPSDB v2.0 Wang et al. (2022) contains 2,917 entries with 586 unique proteins and 6,678 experimental conditions, filtered into protein-only and protein-nucleic acid categories. The *db2* dataset includes 1,188 manually curated entries across *in vitro* and *in vivo* systems, featuring quantitative phase outcomes for high-resolution measurements. The *db3* dataset from RNAPhaSep Zhu et al. (2022) offers numeric records for condition-dependent protein-RNA interactions. This combined resource enables systematic benchmarking and development of computational tools for phase separation studies, with statistics summarized in Table 1.

3.2 COMPONENT GENERATION AND EVALUATION

Table 2: Comparison between source and compositional generation

Component	protein		RNA	
Property	original	generated	original	generated
length[1]	154±54	195±41	110±65	111±48
compositional dist.[2]	0.3148		0.1830	

[1]Length: mean ± std of sequence length. [2]Composition distribution distance:Jensen Shannon divergence of amino acid or nucleotide composition vectors.

We train two separate models for each component type (protein, RNA) in our *db3* dataset (Table 1). For all protein(1) + RNA systems, a phase system with single specie of protein and single specie of RNA, we train a protein diffusion model and an RNA diffusion model in the sequence space respectively. Then, a sampling procedure is performed to generate novel protein and RNA sequences that can participate phase-separating systems when combined. We randomly sample 100 protein and 100 RNA sequences from the trained diffusion models to form the marginal density $p(\boldsymbol{x}_1), p(\boldsymbol{x}_2)$ for joint generation. The diffusion models are trained with diffuser library von Platen et al. with proper parameter settings. We analyze the quality of generated components by comparing their sequence properties with those of real components from *db3*, as it is summarized in Table 2.

3.3 JOINT SAMPLING PROCEDURE

During the joint sampling procedure, we employ Langevin Dynamics to iteratively refine the components while following the PU-guided score function (Eq. 6 and Fig. 2b). Correspondingly, we also joint sample systems without PU guidance as a baseline (Eq. 6 and Fig. 2a). Regardless of the forms of guidance, starting from any initial state, we perform reverse diffusion steps steering samples toward the given joint score field (conditional $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t}|y=1)$ or unconditional score function $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})$) This sampling strategy enables efficient generation of multi-component biosystems with or without guidance. The comparison of two sampling methods are conducted to evaluate the impact of PU guidance on the generation of phase behaviors.

3.4 MULTI-COMPONENT SYSTEMS GENERATION AND EVALUATION

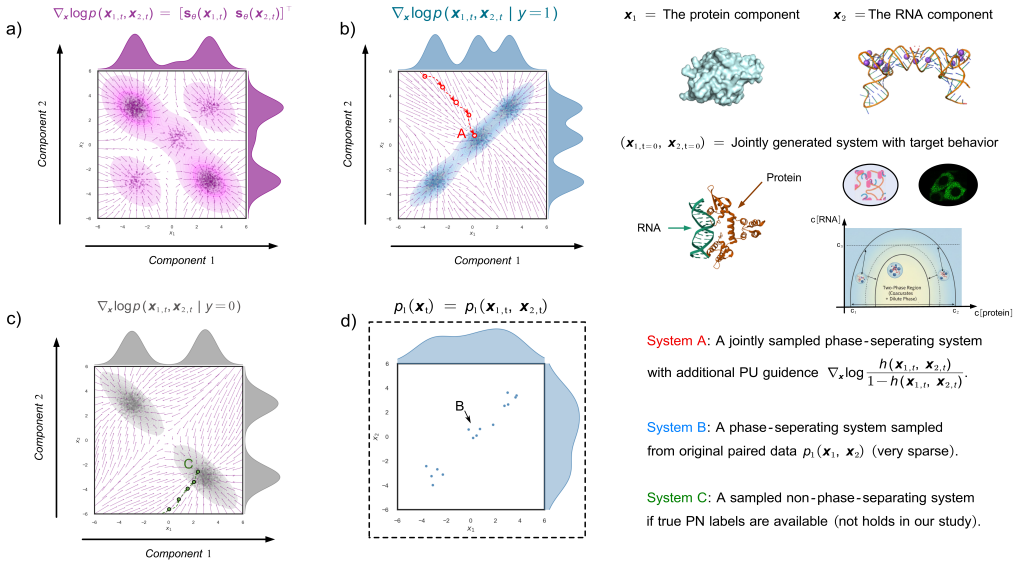

Figure 2: Process of sampling (a) with no guidance; (b) with PU guidance; (c) with negative guidance (not available in real data); (d) from observed positive pairs. Right: cartoon illustration of protein, RNA, and phase behavior. Sampled system A, B and C from the left are explained.

We evaluate the joint generation of multi-component biosystems using the trained nested PU classifier and compositional diffusion models. The aim is to generate protein-RNA pairs that exhibit phase-separating properties. From *db3*, 797 known "protein(1)-RNA" systems that are able to undergo phase separation are treated as positive samples, while all other possible $(100 \times 100 = 1 \times 10^4)$ protein-RNA pairs formed by single-component diffusions are used as unlabeled samples, resulting in a 8% labeled ratio. We trained a nested PU classifier to discriminate P and U systems based on the phase outcome predictor in Jiang et al. (2025).

An *in silico* evaluation is performed on the generated systems to assess their phase-separating propensities. A pretrained phase outcome predictor Jiang et al. (2025) predicts the propensity of phase separation for each generated protein-RNA pair. Results demonstrate that systems generated with PU guidance exhibit significantly higher mean phase separation propensity (0.94 ± 0.03) compared to those generated without guidance $(0.48 \pm 0.02$, p-value $< 0.001)$, shown in Fig. 3a. This showcases the effectiveness of PU guidance in steering the generation process towards systems with desired phase-separating behavior. Phase diagrams of a selected generated systems under varying component concentrations are shown in Fig. 3b, validating the model's ability to generate systems with controllable phase behavior.

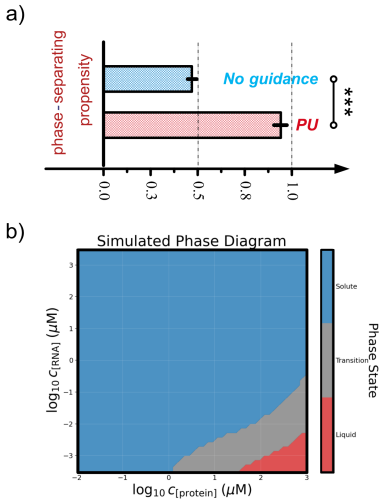

Figure 3: Evaluation of generated systems and their phase behaviors. (a) Comparison of predicted phase-separating propensities with or without guidance. (b) Simulated 3-region phase diagram of a selected system from guided generation in (a).

4 Discussion

Generating phase-separating biosystems with desired properties is a challenging task due to the intricate biophysical interactions. In this work, we curate a comprehensive dataset of phase separation behaviors for multi-component biosystems and proposed a novel approach that combines joint diffusion modeling with positive-unlabeled guidance to control the generation of biomolecular condensates. Future work will refine the guidance strategy, explore more complex multi-component systems, more precisely validate the generated systems and their behaviors in a wet-lab.

REFERENCES

Simon Alberti, Amy Gladfelter, and Tanja Mittag. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell*, 176(3):419–434, 2019.

William E Arter, Runzhang Qi, Nadia A Erkamp, Georg Krainer, Kieran Didi, Timothy J Welsh, Julia Acker, Jonathan Nixon-Abell, Seema Qamar, Jordina Guillén-Boixet, et al. Biomolecular condensate phase diagrams with a combinatorial microdroplet platform. *Nature Communications*, 13(1):7845, 2022.

Salman F Banani, Hyun O Lee, Anthony A Hyman, and Michael K Rosen. Biomolecular condensates: organizers of cellular biochemistry. *Nature reviews Molecular cell biology*, 18(5):285–298, 2017.

Ka Yin Chin, Shoichi Ishida, Yukio Sasaki, and Kei Terayama. Predicting condensate formation of protein and rna under various environmental conditions. *BMC bioinformatics*, 25(1):143, 2024.

Marta Chronowska, Michael J Stam, Derek N Woolfson, Luigi F Di Costanzo, and Christopher W Wood. The protein design archive (pda): insights from 40 years of protein design. *Nature Biotechnology*, pp. 1–3, 2025.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, pp. 2021–11, 2021.

Mingrui Ding, Weifan Xu, Gaofeng Pei, and Pilong Li. Long way up: rethink diseases in light of phase separation and phase transition. *Protein & Cell*, 15(7):475–492, 2024.

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, et al. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025.

Xiang Gu, Ai Zhuang, Jie Yu, Peiwei Chai, Renbing Jia, and Jing Ruan. Phase separation drives tumor pathogenesis and evolution: all roads lead to rome. *Oncogene*, 41(11):1527–1535, 2022.

Maarten Hardenberg, Attila Horvath, Viktor Ambrus, Monika Fuxreiter, and Michele Vendruscolo. Widespread occurrence of the droplet state of proteins in the human proteome. *Proceedings of the National Academy of Sciences*, 117(52):33254–33262, 2020.

Dmitry Ivanov. Dedpul: Difference-of-estimated-densities-based positive-unlabeled learning. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 782–790. IEEE, 2020.

Peiran Jiang, Ruoxi Cai, Jose Lugo-Martinez, and Yaping Guo. A hybrid positive unlabeled learning framework for uncovering scaffolds across human proteome by measuring the propensity to drive phase separation. *Briefings in Bioinformatics*, 24(2):bbad009, 2023.

Peiran Jiang, Adrita Das, Weifeng Wu, Dantong Zhu, Huaiying Zhang, and Jose Lugo-Martinez. Openphase: Condition-aware exploration of multicompo-nent biosystem phase-separating behavior. 2025.

Huan Yee Koh, Yizhen Zheng, Madeleine Yang, Rohit Arora, Geoffrey I Webb, Shirui Pan, Li Li, and George M Church. Ai-driven protein design. *Nature Reviews Bioengineering*, 3(12):1034–1056, 2025.

Tanja Kortemme. De novo protein design—from new structures to programmable functions. *Cell*, 187(3):526–544, 2024.

Guillaume Laflamme and Karim Mekhail. Biomolecular condensates as arbiters of biochemical reactions inside the nucleus. *Communications Biology*, 3(1):773, 2020.

Viacheslav Meshchaninov, Pavel Strashnov, Andrey Shevtsov, Fedor Nikolaev, Nikita Ivanisenko, Olga Kardymon, and Dmitry Vetrov. Diffusion on language model encodings for protein sequence generation. *arXiv preprint arXiv:2403.03726*, 2024.

Daniele Raimondi, Gabriele Orlando, Emiel Michiels, Donya Pakravan, Anna Bratek-Skicki, Ludo Van Den Bosch, Yves Moreau, Frederic Rousseau, and Joost Schymkowitz. In silico prediction of in vitro protein liquid–liquid phase separation experiments outcomes with multi-head neural attention. *Bioinformatics*, 37(20):3473–3479, 2021.

Matt Raymond, Yilun Zhu, Jianxin Zhang, Angela Violi, and Clayton Scott. Joint diffusion sampling via positive-unlabeled guidance for multi-modal data. 2025.

Arunima Singh. Llms predict protein phases. *Nature Methods*, 21(9):1579–1579, 2024.

Marta Skreta, Lazar Atanackovic, Avishek Joey Bose, Alexander Tong, and Kirill Neklyudov. The superposition of diffusion models using the it\^ o density estimator. *arXiv preprint arXiv:2412.17762*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Hiroshi Takahashi, Tomoharu Iwata, Atsutoshi Kumagai, Yuuki Yamanaka, and Tomoya Yamashita. Positive-unlabeled diffusion models for preventing sensitive data generation. *arXiv preprint arXiv:2503.03789*, 2025.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, Steven Liu, William Berman, Yiyi Xu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. URL `https://github.com/huggingface/diffusers`.

Xi Wang, Xiang Zhou, Qinglin Yan, Shaofeng Liao, Wenqin Tang, Peiyu Xu, Yangzhenyu Gao, Qian Li, Zhihui Dou, Weishan Yang, et al. Llpsdb v2. 0: an updated database of proteins undergoing liquid–liquid phase separation in vitro. *Bioinformatics*, 38(7):2010–2014, 2022.

Rachel M Welles, Kandarp A Sojitra, Mikael V Garabedian, Boao Xia, Wentao Wang, Muyang Guan, Roshan M Regy, Elizabeth R Gallagher, Daniel A Hammer, Jeetain Mittal, et al. Determinants that enable disordered protein assembly into discrete condensed phases. *Nature Chemistry*, 16(7):1062–1072, 2024.

Wei-Xin Xu, Qiang Qu, Hai-Hui Zhuang, Xin-Qi Teng, Yi-Wen Wei, Jian Luo, Ying-Huan Dai, and Jian Qu. The burgeoning significance of liquid-liquid phase separation in the pathogenesis and therapeutics of cancers. *International Journal of Biological Sciences*, 20(5):1652, 2024.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4):1–39, 2023.

Daniel Zeiberg, Shantanu Jain, and Predrag Radivojac. Fast nonparametric estimation of class proportions in the positive-unlabeled classification setting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6729–6736, 2020.

Sitao Zhang, Zixuan Jiang, Rundong Huang, Shaoxun Mo, Letao Zhu, Peiheng Li, Ziyi Zhang, Emily Pan, Xi Chen, Yunfei Long, et al. Pro-ldm: Protein sequence generation with a conditional latent diffusion model. *bioRxiv*, pp. 2023–08, 2023.

Haibo Zhu, Hao Fu, Tianyu Cui, Lin Ning, Huaguo Shao, Yehan Guo, Yanting Ke, Jiayi Zheng, Hongyan Lin, Xin Wu, et al. Rnaphasep: a resource of rnas undergoing phase separation. *Nucleic Acids Research*, 50(D1):D340–D346, 2022.

A Appendix

We attached related work, additional datasets curation details, model architecture and training hyperparameter, and experimental details in the appendix.

A.1 Related work and contributions

- **RNAPSEC** is a predictor of protein and RNA under various environmental conditions. It also provides a few embedding methods for RNA, protein and experimental conditions Chin et al. (2024).
- **PRO-LDM** is a deep generative model that can design protein sequences with properties from a learnable conditional latent space Zhang et al. (2023).
- **PU diffusion** is a diffusion model that can generate images with positive-unlabeled guidance to avoid certain types of classes Takahashi et al. (2025).
- **Superposition of Diffusion Models** is a method that can combine multiple pretrained diffusion models at the generation stage under a novel framework termed superposition. Skreta et al. (2024).

Building upon the existing methods and resources summarized above, this work makes the following key contributions:

1. We firstly curate an dataset of phase separation behaviors of multi-component biosystems from existing literature and databases.
2. We propose a joint diffusion framework for the generation of biosystems that undergo phase separation via positive-unlabeled guidance.
3. We showcase that the utility of using PU-guided joint diffusion for the design of multi-component biosystems with desired phase separating behaviors.

A.2 Dataset curation details

We list the preprocessing methods for all three datasets (*db1*, *db2*, *db3*) below.

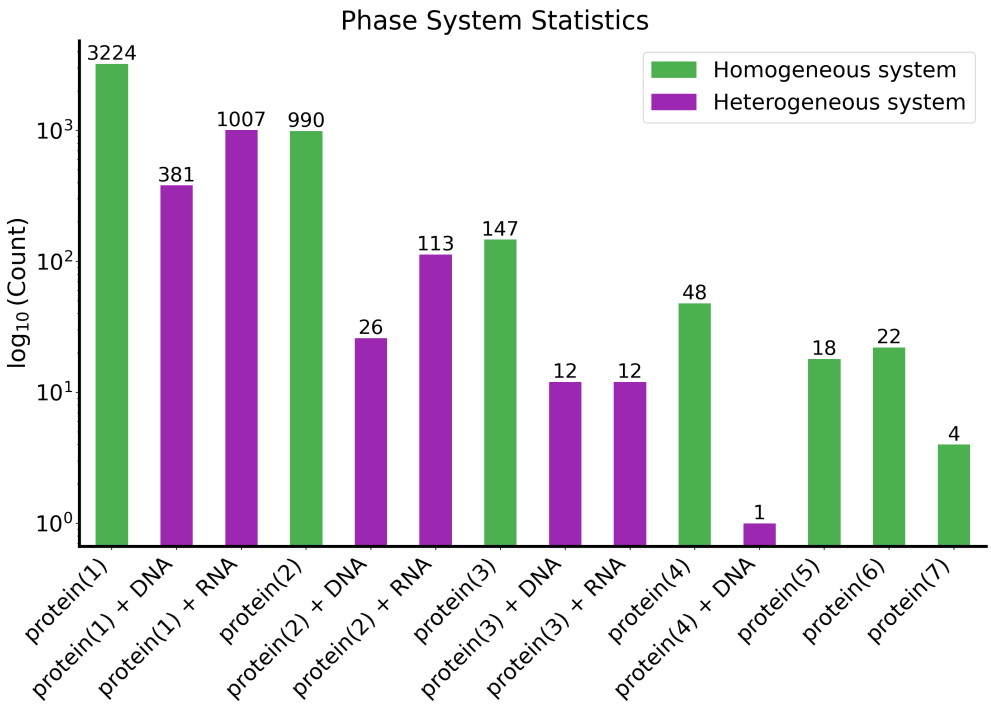

Figure 4: Statistics of system types in *db1* before processing, including homogeneous (protein-only) and heterogeneous (protein-DNA/RNA) systems.

The *db1* Dataset: For the *db1* dataset, we performed a systematic data processing workflow to maximize data quality and kept as many entries as we can for downstream modeling. The main steps were as follows:

1. **System selection:** We retained only representative systems with quantifiable numbers for modeling, including protein(1), protein(2), protein(3), protein(1) + DNA, protein(2) + DNA, protein(1) + RNA, and protein(2) + RNA. Other systems such as protein(4) to protein(7) have fewer numbers and were excluded from the dataset.
2. **Label correction:** All system labels were reviewed and corrected to ensure consistency with the original experimental records. There were some entries with incorrect system labels (e.g. "protein(2)" mislabeled as "protein(1)", "protein(2) + RNA" mislabeled as "protein(1) + RNA").
3. **Data filtering and Deduplication:** Entries with missing, ambiguous, or unparsable experimental conditions were excluded. Duplicate or redundant entries were removed to avoid data leakage between training and testing splits.

The *db2* Dataset: The *db2* dataset is a collection of experimentally validated phase separation outcomes, including sequence information, experimental conditions, and results. It contains 1,188 entries, each representing a single experimental measurement—either one cell (for in vivo data) or one well (for in vitro data). Each entry includes the sequence(s), experimental conditions, and the observed phase separation result. The dataset covers both in vivo and in vitro experiments, and includes both natural and synthetic phase separation constructs.

Table 3: Overview of db2 subsets, phase outcome types, system composition, main variable, and notes

Subset	Phase outcome	System type	Main variable	Notes
LSD1	Binary ($y \in \{0, 1\}$)	"protein(2) + RNA"	Protein and RNA conc.	Ternary exploration
Synthetic IDRs	Intensity fraction ($y \in [0, 1]$)	"protein(1)"	Protein conc.	Synthetic protein
LAF1 RGG	Intensity fraction ($y \in [0, 1]$)	"protein(1)"	Protein conc.	Cell cycle study

1. **Subset 1 (LSD1):** This subset contains 191 entries, each involving a three-component system: two distinct proteins and one RNA species. All data are from in vitro experiments using natural proteins (and mutants) and natural RNA. Main variables include four sequence variants of protein 1, three RNA sequence variants, and wide-ranging concentrations for both protein 1 and RNA. Phase separation outcomes are recorded as TRUE (droplet formation) or FALSE (no droplet formation), enabling exploration of how combinations of protein and RNA sequences at varying concentrations influence phase separation. Other variables (salt, buffer, pressure, temperature, crowding) are fixed and standardized, with no crowding agent used.
2. **Subset 2 (Synthetic IDRs):** This subset contains 278 entries, each involving a single synthetic protein in in vivo experiments conducted in HeLa RMCE cells. All proteins are synthetic, with four different sequences (distinct IDRs, same oligomerization domain). Variables include solute concentration and cell cycle state (interphase or mitosis). All entries display droplet formation (phase separation outcome = TRUE), but the intensity fraction is measured for each entry, representing the fraction of protein undergoing phase separation. This subset focuses on how sequence, concentration, and cell cycle stage impact the extent of phase separation. Other conditions (salt, buffer, pressure, temperature, crowding) are inferred from literature to reflect physiological conditions.
3. **Subset 3 (LAF1 RGG):** This subset contains 719 entries, each involving a single protein in in vivo experiments in HeLa RMCE cells. It includes natural proteins and IDR-mutant variants. Variables include six protein sequences (modified IDRs, same oligomerization domain), solute concentration, and cell cycle state. Phase separation outcomes are recorded as TRUE or FALSE. This subset investigates how sequence variation, concentration, and cell cycle status affect binary phase separation outcomes. Other conditions are assumed to match physiological values. Protein concentration data were derived from fluorescence intensity using a preliminary calibration curve (one replicate); future refinements may slightly adjust reported concentrations.

The *db3* Dataset: The *db3* provides a comprehensive collection of protein-RNA phase separation experiments with detailed annotations.It is derived from RNAPSEC Chin et al. (2024), a manually curated resource that aggregates experimental data Zhu et al. (2022) from the literature referenced in

RNAPhaSep. The construction of *db3* involved systematic collection and annotation of protein-RNA phase separation experiments, including detailed records of protein and RNA sequences, experimental conditions, and observed outcomes.

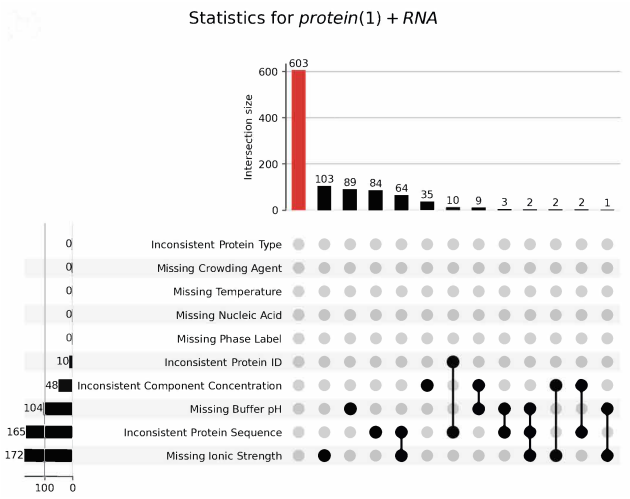

Figure 5: Upset plot showing the distribution of experimental conditions for "protein(1)+ RNA" systems.

Containing 1,514 curated entries, the dataset covers a diverse set of morphologies, including liquid-like, gel-like, and solid-like condensates, as well as negative cases where no phase separation occurs. Specifically, it includes 984 liquid, 92 gel, 53 solid, and 385 non-phase-separating solute entries. The experiments span 37 distinct proteins—such as SARS-CoV-2 nucleoproteins, FUS, and TDP-43—with 96 unique protein sequences and 147 RNA species. The protein UniProt IDs and RNA species are provided in Fig. 6.

A.3 Experimental details

We implement component diffusion by training separate sequence diffusion models for proteins and RNA on the *db3* non-redundant sequences. Sequences are tokenized with 1-indexed vocabularies (20 amino acids for proteins, 4 bases for RNA) and padded or truncated to length 512. The denoising network is a lightweight conditional sequence model with a 64-dim embedding, a 2-layer bidirectional LSTM (hidden size 256, dropout 0.2), and a timestep embedding that is concatenated to the LSTM output and projected to token logits. We use $T = 100$ diffusion steps with a linear beta schedule ($\beta_0 = 10^{-4}$, $\beta_T = 0.02$), optimize with Adam (learning rate 5×10^{-4}, batch size 16) for 500 epochs, and apply gradient clipping at 1.0. During sampling, we run the reverse process from a random token sequence and select the most likely token at each step, injecting mild stochasticity in early steps to promote diversity; we generate 100 protein and 100 RNA sequences for downstream analysis.

The classifier $\hat{h}(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})$ is trained to discriminate between positive and negative samples. We employ a binary classifier architecture consisting of a fully connected neural network with two hidden layers (256 and 128 units respectively, ReLU activation, dropout 0.3). Input features are concatenated embeddings from the diffusion model latent space for both components at timestep t. The classifier is optimized using standard binary cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{N} [y_i \log \hat{h}(\boldsymbol{x}^i_{1,t}, \boldsymbol{x}^i_{2,t}) + (1 - y_i) \log(1 - \hat{h}(\boldsymbol{x}^i_{1,t}, \boldsymbol{x}^i_{2,t}))] \tag{7}$$

where $y_i \in \{0, 1\}$ denotes the PU binary label for sample i. The classifier is trained with Adam optimizer (learning rate 1×10^{-3}, batch size 512) for 200 epochs with early stopping based on validation performance. During guidance, the log-odds output $\log \frac{\hat{h}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{1-\hat{h}(\boldsymbol{x}_1, \boldsymbol{x}_2)}$ is computed and averaged per batch via automatic differentiation to obtain the gradient term in Eq. 6.

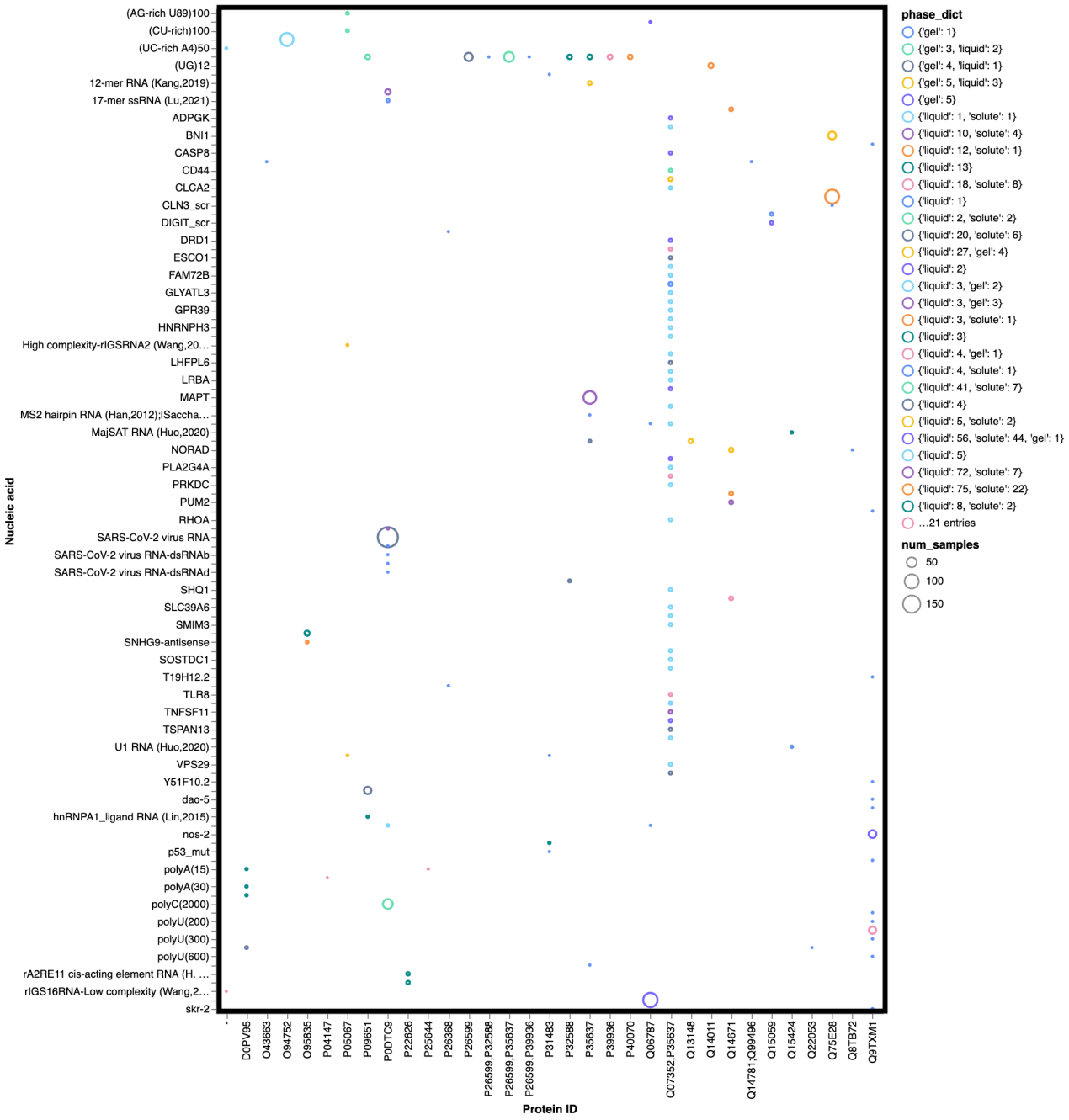

Figure 6: Statistics of protein and RNA species in *db3*. Different phase outcomes are reported under various experimental conditions for each protein-RNA pair.