Beyond Point Sampling: Autonomous Phase Mapping of Biologic Formulation Stability via Hierarchical and Manifold Active Learning

Kiran Vaddi [1 2] **Tyler B. Martin** [1]

[1]*National Institute of Standards and Technology, Gaithersburg, MD, USA* [2]*Theiss Research, La Jolla, CA, USA*. Correspondence to: Kiran Vaddi kiran.vaddi@nist.gov, Tyler B. Martin tyler.martin@nist.gov.

1. Introduction

Biologics, including therapeutic proteins, RNA, and monoclonal antibodies, possess transformative potential for human health; however, their practical application is frequently compromised by physical instability during manufacturing, storage, and distribution [1]. Mechanisms such as cation-induced hydration effects and conformational shifts often drive aggregation or phase separation, necessitating precise formulation design to maintain efficacy [2, 3]. A central bottleneck in this process is the selection of excipients—polymers, surfactants, and small molecules—that stabilize the biologic against these degradation pathways. The resulting design space, defined by high-dimensional excipient compositions and thermodynamic variables like temperature, is sufficiently vast to render conventional trial-and-error screening prohibitively resource-intensive [4, 5].

To address this, we present a self-driving laboratory framework that integrates robotic experimentation with active learning to map biological stability landscapes using autonomous phase mapping [6, 7, 8]. Crucially, our approach respects the inherent physical hierarchy of formulation science: a sample must be synthesized with a target composition before temperature-dependent processing can probe its phase behavior. Consequently, the acquisition problem involves exploring heterogeneous variables: excipient composition as a finite-dimensional vector and temperature as a function-valued continuous profile over time. We propose a novel manifold-based representation to decompose this problem into sequential decision layers. By treating each thermal profile as a point on a Hilbert sphere, we enable efficient resource allocation and robust phase boundary identification. We benchmark this strategy on protein–glycerol systems exhibiting metastable liquid–liquid phase separation, demonstrating that our manifold-based hierarchical acquisition significantly outperforms conventional batch Bayesian active learning in both materials efficiency and detection accuracy.

2. Methods

To establish a baseline for evaluating hierarchical constraints, we implemented a standard Bayesian active learning framework that decouples composition selection from thermal profiling using a discretized Cartesian grid. A Dirichlet Gaussian Process (DGP) classifier [9] was employed to model the probability of categorical phase states (e.g., single-phase vs. two-phase) across the design space. In this baseline approach, the decision process is split into two independent layers. First, the algorithm selects the optimal composition for synthesis by maximizing the marginal entropy of the DGP predictions integrated over the entire temperature dimension. Subsequently, conditioned on this fixed composition, a secondary search is performed along the temperature axis using conditional entropy to identify discrete thermal measurements that maximize information gain.

2.1 Manifold-Based Hierarchical Acquisition

To overcome grid-based resolution limits and repeated DGP fitting for batch temperature selection, our framework models the thermal profile as a continuous function rather than a scalar. Each thermal profile is assigned a score equal to the average marginal entropy of the DGP at different temperature values along the profile for any given composition. We utilize a Square-Root Slope Function (SRSF) representation to project these temperature profiles onto a Hilbert Sphere manifold using the formalism in [10, 11]. We retain the DGP for surrogate modeling but restructure the acquisition into a sequential, mixed-variable optimization problem. First, the optimal excipient composition is selected by maximizing expected entropy in the finite-dimensional composition space, where utility is computed by marginalizing out the temperature variable via Monte Carlo sampling across the manifold. Once the material composition is fixed, the algorithm performs Riemannian optimization [12] directly on the Hilbert Sphere to discover the continuous temperature trajectory that maximizes information gain for that specific sample. Finally, this optimized continuous function is discretized into experimental setpoints (for example, based on the temperature and ramp rate threshold of the hardware), allowing the autonomous system to probe phase boundaries with high resolution exactly where model uncertainty is highest.

3. Results

To rigorously evaluate the proposed manifold-based acquisition strategy, we utilized a "virtual" instrument constructed from literature data [1] characterizing the metastable liquid–liquid phase separation of a protein–glycerol system (Figure 1). We conducted multiple independent active learning campaigns to map this stability landscape, benchmarking the manifold-based framework described in Section 2.1 against the discretized baseline method. All cam-

paigns were initialized identically with a single composition (protein = 40.0 mg/mL, glycerol = 5% vol) subjected to five linearly spaced temperature measurements.

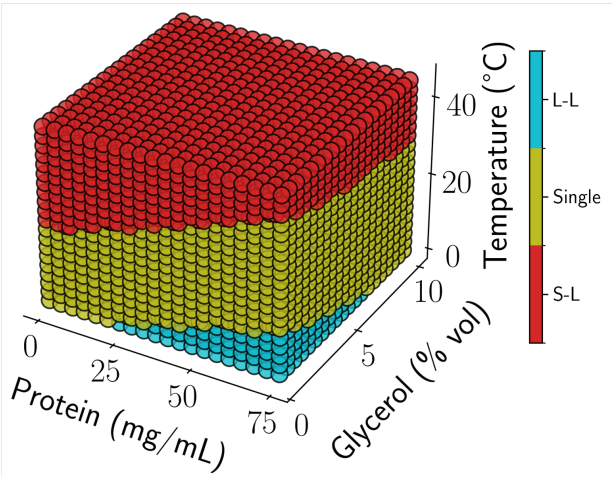

Fig. 1: Ground truth phase diagram for the protein–glycerol system, adapted from Galkin et al. [1]. The plot delineates non-linear phase boundaries distinguishing the homogeneous single-phase region from domains of liquid–liquid (L–L) phase coexistence and solid–liquid (S-L) aggregation.

Figure 2 illustrates the evolution of the posterior marginal entropy for three independent trials. The baseline approach frequently exhibits high variance in exploration pathways, often failing to resolve one of the boundaries (e.g.: boundary between liquid-liquid phase separation and single-phase from 1). In contrast, the manifold-based representation demonstrates remarkable consistency in locating both the phase boundaries across all replicates. This qualitative robustness suggests that treating temperature selection as a function selection on a Hilbert sphere (i.e. going beyond point-sampling in a finite-dimensional space) allows the agent to navigate the mixed-variable space more effectively, avoiding the local minima often encountered in discrete high-dimensional searches.

To quantify this performance, we computed a "boundary score", a scalar metric lower-bounded at zero that measures the similarity between the predicted and ground-truth phase boundaries (Figure 3). The manifold approach consistently achieves higher boundary scores with significantly lower variance compared to the baseline. Furthermore, it converges to the true stability landscape with fewer experimental iterations, thereby reducing the material and temporal overhead of the discovery process. These results indicate that tightly coupling composition selection with optimized, continuous thermal profiling—rather than treating them as decoupled scalar inputs—significantly enhances data efficiency. Ultimately, this establishes the manifold-based framework as a superior resource planning strategy for designing bioformulations.

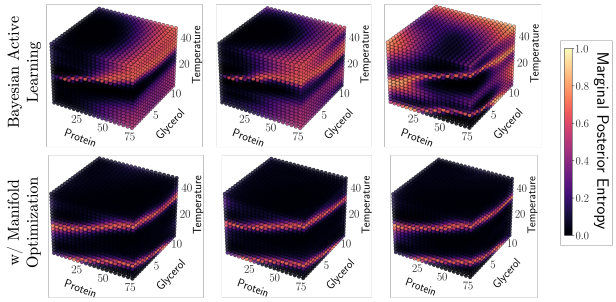

Fig. 2: Comparison of marginal entropies of Gaussian process classifiers at the end of the campaigns (maximum of 50 samples), with each panel corresponding to a single campaign. Across three representative campaigns, the manifold-based approach (bottom row) exhibits a sharper reduction in entropy near phase boundaries compared to the baseline method (top row).

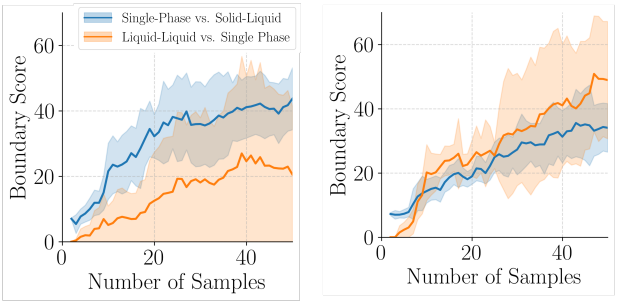

Fig. 3: Quantitative evaluation of phase-boundary detection accuracy as measured by the boundary score. Mean performance (solid lines) and variance (shaded regions) are computed across eight independent campaigns for two distinct phase boundaries: liquid–liquid phase separation versus single phase (orange) and solid–liquid versus single phase (blue). The baseline method is shown in the left panel, and the manifold-based approach introduced in this work is shown in the right panel.

References

[1] Oleg Galkin and Peter G Vekilov. Control of protein crystal nucleation around the metastable liquid–liquid phase boundary. *Proceedings of the National Academy of Sciences*, 97(12):6277–6281, 2000.

[2] Olga Matsarskaia, Michal K Braun, Felix Roosen-Runge, Marcell Wolf, Fajun Zhang, Roland Roth, and Frank Schreiber. Cation-induced hydration effects cause lower critical solution temperature behavior in protein solutions. *The Journal of Physical Chemistry B*, 120(31):7731–7736, 2016.

[3] Stefano Da Vela, Michal K Braun, Andreas Dörr, Alessandro Greco, Johannes Möller, Zhendong Fu, Fajun Zhang, and Frank Schreiber. Kinetics of liquid–liquid phase separation in protein solutions exhibiting lcst phase behavior studied by time-resolved usaxs and vsans. *Soft Matter*, 12(46):9334–9341, 2016.

[4] Matthew J Tamasi and Adam J Gormley. Biologic formulation in a self-driving biomaterials lab. *Cell Reports Physical Science*, 3(9), 2022.

[5] Isabel Waibel, Timo N Schneider, Fiona J Fischer, Poonpat Dumnoenchanvanit, Alina Kulakova, Tin Duy Nguyen, Thomas Egebjerg, Søren Bertelsen, Nikolai Lorenzen, and Paolo Arosio. Bayesian optimization for efficient multi-objective formulation development of biologics. *Molecular Pharmaceutics*, 22(11):6636–6645, 2025.

[6] A Gilad Kusne, Heshan Yu, Changming Wu, Huairuo Zhang, Jason Hattrick-Simpers, Brian DeCost, Suchismita Sarker, Corey Oses, Cormac Toher, Stefano Curtarolo, et al. On-the-fly closed-loop materials discovery via bayesian active learning. *Nature communications*, 11(1):5966, 2020.

[7] Tyler B Martin, Duncan R Sutherland, Austin McDannald, A Gilad Kusne, and Peter A Beaucage. Autonomous small-angle scattering for accelerated soft material formulation optimization. *Chemistry of Materials*, 2025.

[8] Kiran Vaddi, Huat Thart Chiang, Aleksandra Grey, Zachery R Wylie, and Lilo D Pozzo. Autonomous phase mapping of gold nanoparticles synthesis with differentiable models of spectral shape. *npj Computational Materials*, 11(1):335, 2025.

[9] Dimitrios Milios, Raffaello Camoriano, Pietro Michiardi, Lorenzo Rosasco, and Maurizio Filippone. Dirichlet-based gaussian processes for large-scale calibrated classification. *Advances in Neural Information Processing Systems*, 31, 2018.

[10] Anuj Srivastava and Eric P Klassen. *Functional and shape data analysis*, volume 1. Springer, 2016.

[11] Kiran Vaddi, Huat Thart Chiang, and Lilo D Pozzo. Autonomous retrosynthesis of gold nanoparticles via spectral shape matching. *Digital Discovery*, 1(4):502–510, 2022.

[12] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.