Journal Pre-proof

Integration of Bayesian Optimization and Solution Thermodynamics to Optimize Media Design for Mammalian Biomanufacturing

Nelson Ndahiro, Edward Ma, Tom Bertalan, Marc Donohue, Yannis Kevrekidis, Michael Betenbaugh

PII: S2589-0042(25)01205-2

DOI: https://doi.org/10.1016/j.isci.2025.112944

Reference: ISCI 112944

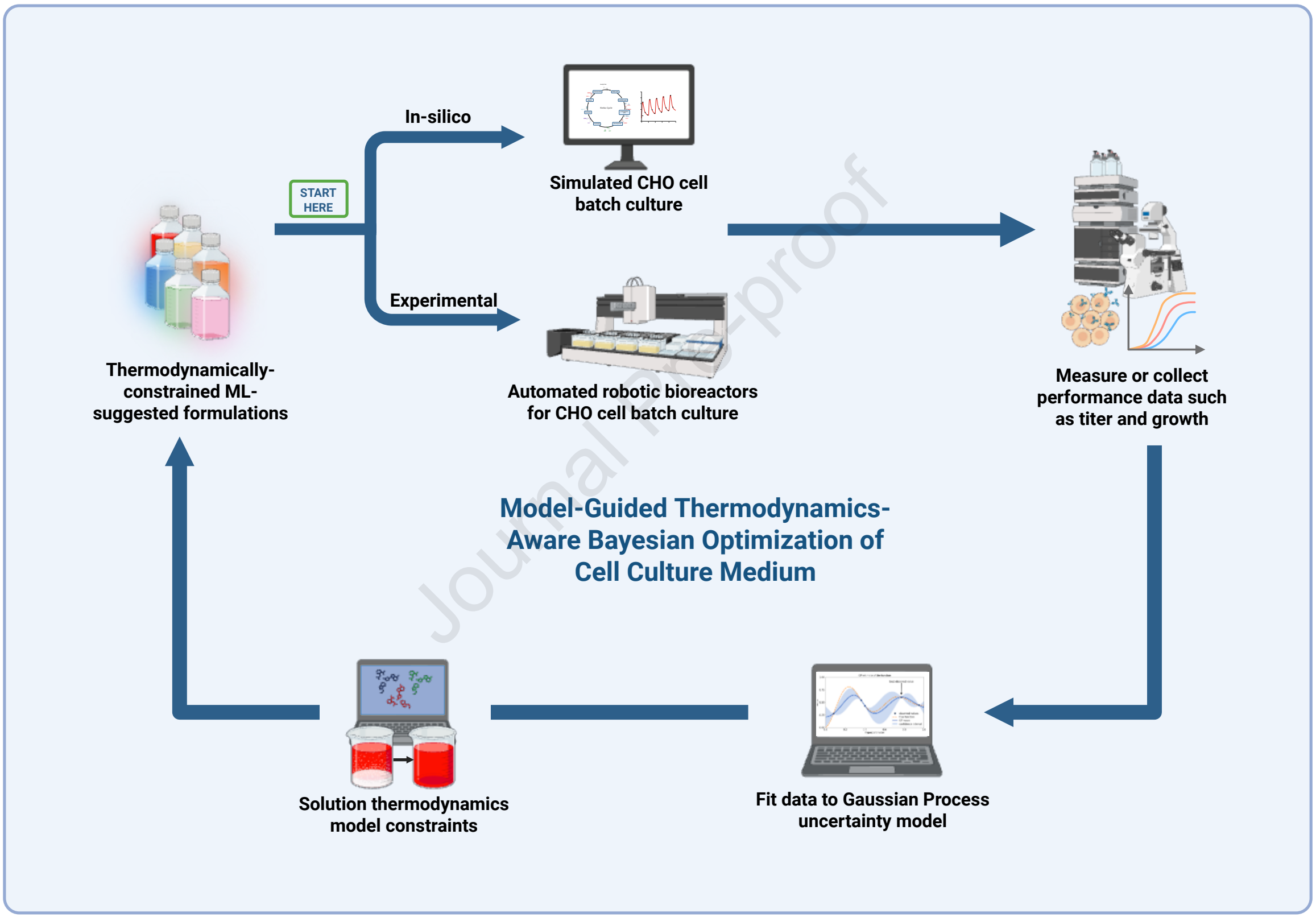
To appear in: *ISCIENCE*

Received Date: 2 January 2025

Revised Date: 11 March 2025

Accepted Date: 17 June 2025

Please cite this article as: Ndahiro, N., Ma, E., Bertalan, T., Donohue, M., Kevrekidis, Y., Betenbaugh, M., Integration of Bayesian Optimization and Solution Thermodynamics to Optimize Media Design for Mammalian Biomanufacturing, *iScience* (2025), doi: https://doi.org/10.1016/j.isci.2025.112944.



Integration of Bayesian Optimization and Solution Thermodynamics to Optimize Media Design for Mammalian Biomanufacturing

Nelson Ndahiro[1], Edward Ma[1], Tom Bertalan[1†], Marc Donohue[1], Yannis Kevrekidis[1,2], Michael Betenbaugh[1,3]

1 Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

2 Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA

3 Lead Contact

† Now at Amgen, Cambridge, MA, USA

*Address all correspondence to:

Michael Betenbaugh, Professor
Department of Chemical and Biomolecular Engineering
Johns Hopkins University
Baltimore, MD 21218
Email: beten@jhu.edu

Key Words: Bayesian, Machine Learning, Solubility, Media Design, Automation, Parallel Design

Summary

Rapid, cost-effective biomanufacturing of products like therapeutics, materials and lab-grown foods depends on optimizing cell culture media, a complex and expensive task due to the combination of components and processing variables. This is especially important for therapeutic production using mammalian systems like Chinese Hamster Ovary (CHO) cells, where long development timelines contribute to high drug costs. Using Bayesian Optimization (BO), adapted for bioprocess applications, our method supports multiple parallel experiments and incorporates thermodynamics-based constraints on media solubility to ensure feasible medium formulations. The approach is validated both in-silico and in experimental bioreactor settings, showing improved product titers compared to classical design of experiments (DOE) methods. This work bridges machine-learning and physical modeling to create a more data-efficient process optimization strategy. The integration of this method into biomanufacturing pipelines together with robotics-assisted bioreactors paves the way for automated bioprocess optimization and more rapidly available and affordable biotherapeutics.

1. Introduction:

Biotherapeutics have emerged as one of the most effective and important pharmaceutical categories of the 21st century. Their ability to treat a variety of diseases, from infectious diseases such as COVID19 to chronic diseases and cancer, have made them versatile tools in the fight against disease. Furthermore, biologics have surpassed small molecules in sales as the dominant treatment modality in disease areas such as solid tumor cancers[1]. Despite this fervent growth there remain challenges related to the cost and development timelines for these products[2]. This is due in part to the complexity of manufacturing these products at scale, where inefficiencies can drive up costs significantly, posing a serious challenge for bioprocess engineers and companies.

The single largest biopharmaceutical product category, monoclonal antibodies, are manufactured primarily by culturing Chinese Hamster Ovary (CHO) cells which generate the recombinant biotherapeutic while growing in industrial bioreactors[3,4]. Implementing an efficient biomanufacturing process requires time consuming and expensive experiments to

identify and develop the right combination of cell line and process variables that produce high product yields. One of the key process factors affecting final product yields is the culture media formulation. Indeed, media development is a cornerstone in the development of optimized mAb manufacturing bioprocesses, as the varied components of the medium have been shown to have strong effects on cell health, growth, mAb titers as well as product quality attributes[5,6,7]. Developing media tailored to a specific mAb product or cell line can be a long, time consuming and expensive trial-and-error endeavor.

One means to reduce media development costs is to apply mathematical models to describe the bioproduction of monoclonal antibodies. Traditional approaches to model-based bioprocess optimization often rely on mechanistic models to predict cell behavior and the role of process parameters[8,9]. While these models can offer valuable insights into the impact of media components, as we and others have shown in previous efforts[10,11,12] their applicability in media design is limited by stringent assumptions about metabolism due to the constraints of the models, whether they are genome scale or kinetic in nature[11,13,14]. These models often offer limited flexibility to describe the variation in cell metabolism that may occur over a bioreactor run. Furthermore, building these models often requires either large or very tailored -omics and time-series datasets in order to parameterize the model to a particular cell line, product, and specific culture conditions[9,14]. As a consequence, researchers have resorted to space-filling designs and extensive rounds of design of experiments (DOE)[15]. Although effective, these methods are resource-intensive, requiring numerous experiments and rigid experimental design frameworks to pinpoint optimal conditions. For example, Box-Behnken design, commonly used in DOE, requires a minimum of $2n(n-1)$ experiments for n variables. This presents a substantial challenge, given the vast number of nutrients and optimization design space created by the unique requirements of different cell lines and bioproducts.

In response to these challenges, the field is increasingly looking to data-driven methodologies. Machine learning, with its ability to handle diverse datasets and uncover complex patterns, presents a promising alternative. Specifically, Bayesian Optimization (BO) has emerged as a powerful tool for navigating high-dimensional design spaces with minimal data[16,17]. BO balances notions of exploration and exploitation, as well as a principled methodology for data-efficient experimental design based on uncertainty quantification. Therefore, BO is usable with limited experimental budgets, making it an ideal candidate for sequential experimental design, especially in contexts where experiments are costly and time-consuming like chemical or biological systems[18,19,20,21]. While data-driven approaches hold promise, these approaches usually are developed in different contexts and for applications such as image processing or text-modeling, often in purely *in-silico* environments[19,22,23]. As a

consequence, relatively little effort has been made to leverage the power of data-driven methods in biomanufacturing.

In this paper we propose an approach to address a bioprocess optimization problem, specifically culture media formulation, using uncertainty-based machine learning models. Our method leverages Bayesian Optimization and we adapt it for bioprocess applications in order to facilitate more efficient experimental design of media formulation. Key features of our approach include the ability to conduct multiple parallel experiments, which leverage a Multi-Scale Multi-Recommendation algorithm (MSMR) and existing parallelized cell culture experimental equipment. The Bayesian Optimization approach is used to evaluate the roles of different process inputs, including specific amino acid media levels and carbon dioxide concentration. Indeed, mammalian cell culture media formulation optimization involves amino acid levels in addition to glucose and other components such as vitamins. Here, machine learning and Bayesian Optimization offers the opportunity to identify the ideal nutrient concentrations through enhanced experimental design more efficiently.

Unfortunately, amino acid levels in media formulation are subject to additional constraints in the form of thermodynamic solubility limits. Indeed, precipitation of amino acids can result in poor biomanufacturing runs and lost batches of media formulation which are even more problematic during scale-up. Furthermore, the presence of specific amino acids can alter, either positively or negatively, the solubility limits of other amino acids. This results in mixtures of amino acids that can either expand or reduce the media design space that is possible. Therefore, for this study, we combined a thermodynamic activity coefficient model, developed by our group, with the machine learning approach. This consideration of thermodynamic solubility limits ensures our machine learning algorithm suggests only feasible media formulations based on precipitation constraints for specific amino acid combinations.

In this study, we first demonstrate our method's efficacy using *in-silico* simulations based on a dynamic flux model of CHO cell metabolism. Next we applied the approach under actual experimental media design conditions for relevant high throughput bioreactor operating conditions. Our findings demonstrate that a machine learning-guided approach, when paired with a thermodynamic-solubility constraining algorithm, can significantly improve the effectiveness of bioprocess media design in terms of a final target protein product titer, outperforming space-filling, a traditional DOE framework.

The integration of machine learning and advanced thermodynamic solubility systems together with robotics-assisted automated bioreactors using Bayesian Optimization techniques represents a transformative approach in media design and development for industrially-relevant CHO bioprocess optimization and mammalian bioproduction platforms in general. Through these example cases, we demonstrate the impact that data-driven optimization can

have on biomanufacturing, offering a glimpse into a future where laboratory automation and advanced machine learning techniques drive rapid data optimization techniques to accelerate development of improved biomanufacturing processes for next generation biotherapeutics needed for future pandemics and complex disease treatments.

2. Model description - Validation of Bayesian Optimization with *in-silico* metabolic model and in-vitro experiments

A hybrid media optimization approach using Bayesian Optimization was developed in three key steps. First, the *in-silico* mechanistic model was employed to simulate the bioprocess and fine-tune the optimization parameters using BO. Next, a thermodynamic model of amino acid solubility was incorporated to constrain the experimental design space, ensuring that only feasible and soluble medium formulations were considered. Finally, the refined models were applied in experimental settings, with adjustments made to enhance the accuracy and performance of the optimization

2.1 Utilizing a mechanistic metabolic model to represent different metabolic conditions

In order to test the feasibility and utility of Bayesian Optimization (BO) for experimental design in bioprocess applications, we employed a CHO model for *in-silico* testing. This model was developed by the authors to simulate the metabolic behavior of CHO cells under different cell media and feeding strategies[24]. This mechanistic model describes CHO cell behavior by employing both linear equations for maintaining mass balance of intracellular metabolites, as well as kinetic equations to model precisely known enzymatic reactions in the central energy metabolism of the cell. The model is designed with a set of parameters specifically fitted for CHO cell lines under various media and feeding conditions. While the simulation does not match perfectly all the experimental data using a single parameter set, it provides a reasonable representation of cell metabolic behavior, including amino acid consumption, cell growth, and titer production. Therefore, this mechanistic model is well-suited for use in *in-silico* experiments to predict a time-series of nutrient profiles and effectively simulate the metabolic processes of CHO cells under different media compositions and with different initial media compositions. The results of these *in-silico* experiments will provide useful/appropriate parameters to be used in subsequent in-vitro high throughput bioreactor validation studies.

The solution spaces of media components of interest were defined for both *in-silico* and in vitro experiments before initiating the Bayesian Optimization. The upper limit of each component is bounded by its solubility as described in section 2.5 below.

2.2 Bayesian Optimization campaign and Prediction Update Policy

A Bayesian Optimization campaign involves a series of iterative experiments aimed at optimizing a specific objective. Each iteration involves selecting a set of input parameters, conducting experiments or simulations, observing the outcomes, and updating the probabilistic model based on the new data. This approach relies on prior input/output distributions and the likelihood of the observed data to calculate the posterior distribution, guiding the prediction process in subsequent iterations. In this use-case for BO, our input variables are the concentration of several basal medium components which will be varied to maximize our measured output: Immunoglobulin G (IgG) protein titer produced by our CHO cells.

By utilizing prior knowledge and continuously refining the probabilistic prediction model based on existing and acquired data, a Bayesian Optimization media campaign is more flexible and should require fewer total conducted experiments than traditional DOE methods which are not dynamically responsive to outcomes. To guide the BO towards accurate predictions, various sets of basal media compositions, well-spaced within the defined solution space, were arbitrarily chosen for an initial round of experiments. The titer of IgG produced at the end of seven days in both in-silico and in-vitro experiments was recorded as one of the optimization objectives for the BO campaign. The model learns a mapping between input media components and output titer by performing regression using Gaussian Processes (GP). By maximizing the conditional probability of the model given the data collected, the mean of the GP model and associated variance can be calculated and used to predict titer of each input formulation along with uncertainty for each prediction:

$$y \sim GP\big(\mu(x), k(x, x'')\big) \qquad (1)$$

$Here\ y\ is\ the\ titer\ with\ mean\ \mu(x), x\ is\ vector\ of\ the\ concentration\ of\ nutrients,$

$k\ is\ the\ covariance\ (or\ kernel)\ function.$

To update the prediction policy with given data and improve the accuracy of the next iteration cycle, the posterior mean and variance distributions at unseen input points of next iteration can be defined by:

$$\mu(x) = k^T * (K + \sigma_n^2 * I)^{-1} * y \qquad (2)$$

$$\sigma^2(x^*) = k(x^*, x^*) - k^T * (K + \sigma_n^2 I)^{-1} * k \qquad (3)$$

$Here\ \mu\ is\ the\ updated\ prediction, k\ is\ the\ covariance\ between\ observed\ and\ unseen$

$points, K\ is\ the\ covariance\ matrix\ of\ observed\ data, \sigma_n^2\ represents\ noise\ variance,$

$and\ I\ is\ the\ identity\ matrix.$

The covariance function, **k**, measures the similarity between pairs of points in the input space, and **K** is the covariance matrix which stores all the pairwise evaluations of **k** for the

whole training data. These two functions allow us to calculate mean and variance given data. The Matern kernel implemented in Python Packages Scikit-Learn[25] and BayesOpt[26] was used due to its flexibility and ability to work on complex real-life systems such as chemical reaction systems[18,20,21].

Posterior means and uncertainties in equation (2) and (3) were leveraged to decide on the best experiments to conduct with respect to an objective (e.g. maximizing titer), given the data collected so far. The choice of the next experiment to conduct was made by an acquisition function (AF) which suggests the next experiment with the most utility.

2.3 Exploratory and Exploitative Modes in Bayesian Optimization

In a BO campaign, the 'explore v.s. exploit' tradeoff plays a crucial role. Exploratory BO involves designing experiments in unexplored areas of the design space, which can lead to the discovery of promising candidates in unexpected regions. This approach emphasizes 'exploring the unknown.' On the other hand, an exploitative algorithm focuses on optimizing the objective (such as maximizing titer), prioritizing regions of the design space that are likely to yield high performance based on existing data. This approach emphasizes 'exploiting the known.'

While there can be multiple rounds of exploration and exploitation, a two-iteration BO campaign was employed in this paper to demonstrate BO's efficiency time-wise. The first iteration prioritized exploration, aiming to investigate new regions of the design space. The second iteration focused on exploitation, concentrating on areas with a high probability of optimizing the titer.

This 'explore-exploit' paradigm is formalized and explained through the expressions of our chosen acquisition function, upper-confidence bound, UCB. An acquisition function acts as a measure of the utility for each possible experimental design given the data. The UCB acquisition function is specifically parametrized by a single parameter **κ**, kappa as denoted in Equation 4:

$$a(x) = \mu(x) + \kappa * \delta(x) \qquad (4)$$

Here a is the measure of the utility of sampling in the next iteration, and δ *is predicted standard deviation.*

The utility measured by the acquisition function a is a function of the predicted mean μ and uncertainty σ (standard deviation) of the Gaussian Processes. **κ** is selected by users to be high when exploration is desired, resulting in high value of a, for areas of high uncertainty (standard deviation term dominates). Conversely, **κ** is then selected to be low when exploitation is desired, prioritizing areas of high prediction of the mean of the GP (mean term dominates). After data is collected and regression with GP is complete, the algorithm uses the acquisition

function to generate the utility of each point. The point of maximum utility is the BO algorithm's choice for the next experimental measurement to be taken. A demonstration of the effect of **κ** on selected experiments is shown in Supplemental Figure 1.

The suggested media formulation then is evaluated, obtaining the associated titer. The data is then added to the model by regression for the second iteration. This paradigm can be repeated in a loop until a desired number of experiments is reached to increase prediction confidence. At the end of a campaign, a retroactive analysis of the experiments selected by the BO algorithm can be conducted to understand the choices made by the algorithm and to compare to alternative experimental design approaches such as different space-filling designs.

2.4 Multiple recommendations

BO is classically formulated as a sequential-sampling optimization problem. That is, only one design is suggested at a time by the acquisition function. This works well for optimizing expensive mathematical functions or scientific models where single model evaluations take relatively short time intervals and can be done repeatedly. In our case, a single cell culture experiment can take from three to four weeks due to the length of a single bioprocess (~7-14 days) and associated preparations (media adaptation, ~9-12 days). Therefore, it is desirable to leverage the parallelized nature of cell culture experimental setups which allow us to run multiple bioprocesses at the same time. In our case, using the Sartorius Automated Micro-Bioreactor (AMBR) system, we can run up to 48 reactors at once. So, to design a more tailored bioprocess-specific BO algorithm, a batched-recommendation approach was selected to make use of parallelized resources available in laboratory setups. MSMR was chosen as it is shown to be superior to other batched-Bayes approaches such as local-penalization and Kriging-believer algorithms[17,27]. It allows us to not make assumptions about the underlying input-output (media-titer) function by sampling multiple length-scales. The length scale is a parameter of the Gaussian Process's covariance function k (Eq. 1, 2, 3). It represents the underlying smoothness of the input-output function, or how much the measured output is expected to vary with small changes in the input. A series of different length scales are chosen by the MSMR to cover a broad range of assumptions about smoothness, and different experiments are suggested from these. The result is a set of diverse experimental recommendations for our parallel bioreactors, informed by previously collected data.

2.5 Thermodynamic model UNIFAC

The feasibility of medium designs suggested by the BO algorithm was further constrained by using thermodynamic models developed by the authors to predict nutrient solubility and precipitation concentrations (Ndahiro, et al, submitted manuscript). The thermodynamic model used was UNIFAC, a group-based contribution model which uses activity coefficient calculations

to predict amino acid solubilities in aqueous mixtures. An extensive database containing both literature and experimentally collected data was used to regress interaction parameters, generating a robust model. The model can predict the effect of adding one amino acid on the solubility of another (or multiple other) amino acid(s) already in solution. By calculating when the activity of each amino acid in a multicomponent solution exceeds its solubility limit at a given temperature, amino acid precipitation can be anticipated computationally; these precipitation concentrations represent upper bounds to the concentrations used in the media design experiments. After each BO recommendation, the nutrient composition was fed into the UNIFAC model to determine if there is a prediction of precipitation. These experiments were then characterized as infeasible.

3. Results:

3.1 Bayesian Optimization applied to bioprocess systems *In-silico* experiments

Before initiating the expensive laboratory experiments, we leveraged the existing *in-silico* CHO metabolic models to build a proof-of-concept of BO in bioprocesses. As they pertain to, a hybrid model featuring enzyme kinetics and flux balances denoting CHO cell behavior was implemented. Incorporating a hybrid approach to model the cell metabolism allows a consideration of known cell metabolic kinetics and steady-state balances, resulting in enhanced predictions than either approach alone. The model performed particularly well in predicting cell metabolism for 7 days in a fed-batch culture[24]. So, this model was employed as a representative simulation environment for our BO campaigns.

To run a virtual cell medium optimization campaign, impactful medium components were selected to vary as a means to test Gaussian Process Modeling and Bayesian Optimization Capabilities. Although a different set of nutrients were ultimately tested in the experiments due to limitation in nutrient components present in the model, it was worthwhile for the selected media nutrients in simulations to have a significant impact on titer so as to clearly showcase the ability of BO to identify promising formulations. First, glucose was chosen as it is the main energy source of the cell and is crucial to cell growth and therefore production of IgG[5,14,28]. Secondly, asparagine is a non-essential amino acid which serves as a carbon and nitrogen source for cell growth and protein production[29,30,31]. Third, lactate is an important molecule in bioprocesses, namely due to its known role in inhibiting cell growth. This often occurs in high glucose media as glucose is metabolized into elevated levels of pyruvate. This causes the lactate dehydrogenase enzyme to convert pyruvate to lactate and accumulate during the cell culture. This leads to slower cell growth and as glucose is depleted, cells shift to consuming lactate[28,32]. Lactate production is linked to glucose consumption, which adds a layer of complexity to the optimization problem, a common occurrence in bioprocess optimization. Glucose, asparagine and lactate were selected as medium components to vary in these simulations due to their

varied roles in cell metabolism, providing for a robust optimization challenge. While adding lactate in cell culture is not common, this was done for our in-silico experiments to explore whether BO could learn this and design the medium composition accordingly.

As mentioned in the model description, the 'explore-exploit' paradigm is crucially important to experimental design in unexplored space, and this can be tuned explicitly using BO. To evaluate whether exploration or exploitation was more suited to our task of optimizing the levels of glucose, asparagine and lactate in the CHO model media, we simulated an array of **κ** values in the UCB acquisition function and monitored the choices made by the BO algorithm. Four **κ** values, namely 0.1, 1, 3 and 10 were selected to span different orders of magnitude and therefore significantly different BO behaviors. A budget of 20 sequential simulated experiments (or iterations) were selected and the same initial starting media formulation was chosen for all **κ** values and the BO experiment was initiated. Titer values on day 7 were recorded for each simulation. A total of 80 simulations (20 iterations [x-axis] with 4 **κ**[A,B,C,D]) were run as seen in Figure 1. The media composition evaluated, and titer obtained by the BO algorithm for each iteration is represented by a datapoint in Figure 1.

The BO campaigns reveal the different learning behaviors of each **κ** value over each of the 20 experiments. Figure 1A through 1D showed the input concentration of glucose, asparagine, and lactate at each simulated experiment, while figure 1E through 1H showed the corresponding titer output on day 7. Strikingly, the very exploitative **κ** value of 0.1 shows very limited learning behavior after 6 experiments as indicated by the plateau in titer measurement, suggesting the presence of a local maximum. The stagnation that is observed after 6 experiments in Figure 1E is preceded by experiments that resulted in lower titers as lactate is increased (particularly at experiment 4), which likely guided the model to remain in the previous areas of higher titers. Conversely, the higher **κ** values of 1 and 3 showed an ability to recover from low titers and continue exploring the design space. Lactate level in both figure 1A and 1B converged and remained at a low level as indicated by the green line in figure 1B and 1C. What's more, the **κ**=1 campaign was able to reach almost 0 mM of lactate, which is in line with a realistic media formulation. Intriguingly, the **κ**=1 campaign also stabilized the media formulation at moderate levels of glucose (except for occasional excursion to explore higher levels) as indicated by the blue line, likely learning the inhibitory effects of high glucose levels led to high lactate production. The **κ**=3 campaign showed similar but more exploratory behavior than **κ**=1. The **κ**=10 campaign showed erratic behavior as it oscillated between low, medium and high values of each metabolite, indicating high levels of exploration but little learning of appropriate values of nutrients.

Another campaign was designed, this time simulating a space-filling experimental design. 20 equally spaced points in the 3-dimensional design space were selected and ran on

the CHO metabolic model. Since this was a classical space-filling approach, there was no sequential experimental design or learning from previous data, and the titers obtained from the campaign are shown in Figure 2. The violin plots in Figure 2 from these campaigns indicate that the best media designs for titer maximization in general came from the BO campaigns with moderate **κ** value . Furthermore, titers obtained from the BO-designs were higher on average, especially at intermediate **κ** values, than those obtained with the space-filling method. Supplementary Figure 2 shows another *in-silico* BO experiment which varied **κ** values but was run with 3 different initial media designs at the start of the BO campaigns. A similar trend was observed, namely that moderate **κ** values performed the best in terms of finding high titer media formulations, showing that BO's ability to optimize a bioprocess objective function is generally independent of the starting point, given sufficient data. The spread of the BO-recommended designs selected with 3 different starting media formulations is visualizable in the form of a Principal Component Analysis (PCA) in Supplementary figure 3. In this case, low to moderate **κ** values result in a reduced space of medium formulations being selected, while at the same time focusing on the most relevant experiments with respect to titer.

3.2 Implementing multiple recommendation for each BO iteration

The in-silico BO campaign results were promising and show that, after tuning the BO parameters, bioprocesses can be effectively optimized using this experimental design framework. However, for this approach to be practical, parallelization of experiments is required. While a budget of 20 experiments is reasonable, if these are all conducted sequentially, the time required to accomplish these experiments would be prohibitively long. Therefore, a framework that recommends batches of experiments would be desirable.

Different approaches to generate multiple parallel experimental recommendations exist with varying levels of complexity. The Kriging-believer algorithm is one such method. In a Kriging-believer framework, like in classical BO, data is collected, and the acquisition function is maximized to suggest a candidate experiment[33]. Now, instead of experimentally generating the objective value of the candidate experiment, the objective value is set to the GP's prediction at that candidate point. In other words, the prediction of the model is taken on faith as the 'real' experimental value, which generates a new maximum for the acquisition function, generating a new design candidate. The new candidate design is then also set to the prediction, and so on until the required number of recommendations per iteration is reached. While this method is simple and intuitive, it has been reported to have limited effectiveness compared to other methods[17,27].

Such a method also may suffer from low diversity of recommendations, which is vitally important for the GP statistical model/approach to learn a complex design space such as bioprocess design. So, other methods such as the Localized Penalty (LP) and Multi-Scale Multi

Recommendation (MSMR) approaches[27,34] were designed to avoid very similar recommendations for each round of recommendations. MSMR was shown by Joy et.al to be superior to LP and other parallelization methods[27]. This method has the added benefit of being applicable when there is little known *a priori* about the shape and smoothness of the underlying design-objective space. This is often the case in bioprocess experiments when different media components, products or even different cell lines have substantially different behavior, yet need to be optimized within a unified framework. MSMR generates multiple candidate response functions and proposes a target number of candidates based on their diversity.

The batch-BO framework with the MSMR algorithm was thus implemented for the *in-silico* testing ground to investigate the effectiveness of the algorithm. The BO campaign was structured to have 10 sequential runs with 4 batches of simulations per run, totalling 40 total simulations. The acquisition function was set to UCB with a $\boldsymbol{\kappa}$=1, which was shown to be relatively exploitative in the non-batched runs. A range of length scale values (smoothness of underlying input-output function) of 1e-6 to 1e6 was given as the MSMR's bounds of search, ensuring a wide range of possible functions are explored. The goal of the campaign was to maximize titer, with a limited experimental budget, while employing a parallelized recommendation approach. The resulting titers from this BO campaign were plotted over each iteration in Figure 3A.

Notably, each candidate media design maximizes an acquisition function associated with a length scale that the algorithm selected. The results of the MSMR approach in Figure 3A show that the method is indeed able to generate candidates with diverse length scale values (y-axis) spanning the range of 1e-6 to 1e6 across the 10 BO iterations (x-axis). While the strength of the MSMR algorithm lies in the fact that we do not need to assume the unknown objective-function's shape, this also means that many length scale values are used, including unrealistic ones. To understand which of the many length scales selected by the MSMR are the most effective, a post-campaign analysis was conducted. Over the course of the campaign, the length scale values of each BO-selected design were recorded and plotted against the resulting titer (Figure 3B). The plot was divided into 4 quarters, separating low and high titers and low and high length scale values. The entire campaign resulted in 25 out of 40 BO recommendations in the high titer quadrants (set to above 4g/L, Q2 and Q3). Strikingly, 80% of the 25 high titer recommendations were from low length scale value (< 1E0) denoted by Q3. Conversely, high length scale values (Q1 and Q2) resulted in 12 out of 15 low titers as visible in the first quadrant Q1. This suggests that under an exploitative $\boldsymbol{\kappa}$=1, 'low' length scale values are able to better optimize this *in-silico* bioprocess, implying that the underlying media-titer function may be relatively smooth. Intuitively, this agrees with the notion that changes in cell IgG production are gradual with respect to changes in media composition with very few big leaps in titer

resulting from a small change in media composition. Informed by this model of our real process, this restricted range was therefore selected for use in the experimental validation of our BO approach.

3.3 Thermodynamics based design constraints with UNIFAC

In many optimization problems, including BO, the solution space is often defined within a hypercubic space where each variable is confined by specified upper and lower bounds, creating a uniform search area. This approach assumes that each variable can be independently varied. However, in bioprocess medium optimization, this assumption does not hold true due to the complex interactions among different components of the system. Different nutrient concentrations, pH, and temperature often influence each other in non-linear and interdependent ways. High concentrations can lead to unintended precipitation of nutrients that exceed their individual solubilities and create infeasible media formulations. Equally, higher concentrations of nutrients may increase the solubility of individual nutrients beyond their individual solubilities, resulting in a feasible medium formulation. These interactions can create constrained or expanded solution spaces without simple cubic boundaries.

In these situations, researchers typically apply heuristics based on experience, or even arbitrary design constraints. In the case of media design, information on feasible media formulations, which provide design constraints, would be advantageous. So, to tailor our BO algorithm to consider feasible formulations, we leveraged a chemical group-based thermodynamic model developed as a means to predict infeasible media formulations. These infeasible operating conditions result from the precipitation of amino acids due to their solubility limits. Precipitation is an important factor in cell media formulation as this phenomenon can impact cell health, product quality and also lead to nutrients becoming unavailable by precipitating out of solution[35,36]. This is especially problematic in bioprocesses, for example, when an amino acid concentration may have positive attributes related with cell growth or titer but cannot be dissolved in media. An additional complication is that amino acid solubilities may be altered due to interactions between amino acids and result in unexpected precipitation and wasted experiments or loss of valuable batches if occurring during scale up.

To address this solubility challenge, UNIFAC, a group-based contribution model which uses activity coefficient calculations, was implemented to predict the solubility levels and precipitation of specific amino acids as described above.. Without using this model, the amount of amino acid that can be incorporated into the media formulation may be either underestimated or overestimated. Taking a two amino acid system as an initial step, current approaches simply define a square solution space based on individual amino acid solubilities without considering thermodynamic interactions as shown by the dashed-line boundaries in Figure 4A and 4B. The dashed edges of this square represent the individual solubilities of amino

acids, the upper bound of our medium design space. Meanwhile, the colored areas in the figures illustrate how estimating solubility of amino acids independently of one another's impact can lead to errors and inappropriate media designs when thermodynamic solubility limits are ignored. Figure 4A illustrates a case of overestimation of the solution space due to interactions between two amino acids, which results in an inappropriately expanded design space. Conversely, Figure 4B illustrates an underestimation of the solution space caused by solubility expansion, where the individual solubility of amino acid A increases due to the presence of amino acid B, highlighting the impact of neglecting amino acid solubility interactions in the design of feasible medium formulation. Hence, a feasibility function was built from UNIFAC to plot soluble regions (green) and insoluble regions (pink) as shown in figure 4C and 4D to demonstrate using different concentration combinations of phenylalanine and glutamine (4C), and glutamine and alanine (4D) respectively. In figure 4C, it is shown that glutamine and phenylalanine would negatively impact each other's maximum solubility through their molecular interactions, meaning that less phenylalanine and glutamine should be added to the media than expected without these interactions. On the other hand, Figure 4D illustrates that glutamine and alanine can interact with each other in a way that enhances each other's maximum solubility, allowing for higher possible concentrations in the media than used in the original design.

In-silico BO campaigns with rigid boundaries based on independent maximum solubility were conducted to demonstrate the precipitation-related failure modes of interactions described above. In Figure 4C, 20% of the BO-designs were infeasible yet were still sampled by the traditional methods. Conversely, in Figure 4D, 13% of feasible design space was ignored through conventional methods due to the synergistic solubility effects between glutamine and alanine. By incorporating thermodynamic constraints, our experimental design framework provides a means for optimizing within the feasible design space, improving the design space scope and value of this Bayesian testing algorithm.

3.4 Experimental validation in AMBR bioreactors

Previous result sections developed an enhanced BO algorithm with parameters compatible with our bioprocesses, the ability to effectively parallelize experimentation though MSMR, as well as implementing design space constraints based on thermodynamic considerations. This initial analysis enabled us to go beyond computational simulations and pursue experimental validation in applicable bioprocess design test cases. For this experimental design, the composition of a Chinese Hamster Ovary (CHO) basal media was considered. In order to maximize production of a model monoclonal antibody, the most widely used biotherapeutic in biotechnology, we started with a custom medium formulation common for commercial applications (see Methods) but with specific critical amino acids that were missing

and available to be varied . The specific amino acids that were missing were Phenylalanine and Glutamine, both amino acids whose levels are known to affect cell growth and productivity[15,37]. In addition to these two nutrient variables, the CO_2 level in the incubator was varied, which has a direct effect on the dissolved CO_2 in media, an important factor in cell production dynamics[38,39].

The Sartorius Automated Micro-Bioreactor (AMBR) was employed to evaluate multiple options for this medium optimization design process. The AMBR system is a powerful high-throughput robotic platform which is widely used to automate bioprocess tasks such as pipetting, passaging, as well as temperature and gas controls. The highly automated reactors make it ideal for consistent and standardized data generation, important to machine learning algorithms such as our BO approach which are dependent on a high degree of data consistency.

CHO-K1 cells were thawed in their native Immediate Advantage media and recovered as described in Materials and Methods. Cells were thawed in a 30mL shake flask culture with regular media for 2 passages and then were transferred into the ABMR bioreactors in BO designed media for 2 more adaptation cycles. Data collection was initiated after adaptation, and each media condition was run as biological duplicates (2 bioreactors per condition). Cell counts were performed daily as well as titer analysis on days 5 and 6 with the data fed back into the BO algorithm to generate experimental designs.

An initial set of 5 media options was selected to serve as seed data for the BO algorithm. The design space of this initial dataset was selected to span from half, base and double (0.5X, 1X, 2X) the base concentration of Phe and Gln in the media. The CO_2 level was set to either 5% or 8%. An AMBR run of 10 bioreactors (5 conditions duplicated) was run in parallel (Table 1) with these initial media conditions. Day 5 was found to be the peak cell density measurement for the culture before cell death and proteolytic activity degraded the product. Cultures were terminated after cell viability dropped below 65%.

This initial seed data was then used to set up the next round of experiments using BO (see Figure 5). The media and CO2 concentrations, along with the measured titers were used to condition the GP model and generate mean predictions and uncertainties. For the next round of experiments, a set of 6 media options (12 reactors) was selected with the goal of maximizing titer. A separate set of 6 media options were chosen to optimize peak VCD, another bioprocess objective often used as a proxy for final titer. To accomplish this goal, a separate GP model was conditioned based on peak VCD from the previous 5 seed experimental data sets. The 12 media alternatives (6 for titer 6 for VCD) were generated by using the MSMR methodology with a range of length values ranging from 10^{-6} to 10^{2}. This selected range was smaller than the *in-silico* tested values (10^{-6} to 10^{6}) since our simulations above ascertained that lower values were likely to be most effective. However, we selected a less restrictive range than the best

performing quadrant Q3 (10^{-6} to 10^{0}, Figure 3B) determined above, as it was expected that the laboratory experimental system will not be reflected identically by the simulation system. Of the 12 BO-designed experiments, 6 of them were generated by conditioning the GP on the training data with titer as the output. The 6 candidates represent 6 media designs to maximize titer. The remaining 6 experiments were generated by conditioning a separate GP model on the peak VCD of the cultures. This was still done with the ultimate final goal of maximizing titer, and to make use of the viability data collected in the initial 5 seed experiments. These two sets of BO 6 media design experiments with biological duplicates were then initiated using the AMBR system.

In addition to titer and viability, pH and osmolarity measurements were collected for the AMBR runs as shown in Supplementary Figure 6 and 7. The pH exhibited consistent patterns across all media conditions and osmolarity falls in the optimal range to support robust cell growth and productivity. In particular, pH initially declined through Day 3 and rose thereafter in all conditions, reflecting an intrinsic link between nutrient composition and cellular metabolic behavior. Osmolarity measurements ranged between 284 and 307 mOsm/kg across all tested formulations, which lies within the physiologically optimal range for CHO cells in batch culture. These findings suggest that the effects of pH were implicitly captured by the algorithm through their coupling to nutrient inputs and cellular responses. Moreover, all cultures were adjusted to a constant volume of 15 mL, minimizing confounding effects from volume-dependent osmolarity changes.

As a comparison to this BO approach, another experimental design approach was run in parallel. Space-filling was selected as a flexible approach that can take a dataset of arbitrary size and recommend a user-specified number of experiments. The experimental design software JMP, was used to generate 12 media designs[40]. The final experimental design consisted of 12 BO-designed media alternatives and 12 media options using space filling experimental design, as described in Table 1 with the overall experimental plan depicted in Figure 5. All media options were conducted as biological replicates, resulting in a total of 48 AMBR mini-bioreactor runs.

The results of this round of experiments are shown in Figure 6. Importantly, the selected formulations show clear differences between the BO and space filling designs. As shown in Figure 6A, the space-filling approach selected a broader range of formulations than the BO. In fact, the range of formulations selected by the BO was relatively restricted compared to that of the space-filling design. Directional dispersion analysis was conducted for BO and space filling approaches as shown in Figure 7[41]. The variance ellipse in Figure 7A indicated a tighter spread along the Phenylalanine axis for the BO approach, reflecting a tendency towards lower phenylalanine values with shifted center. In contrast, figure 7B displayed a more centered

distribution and a wide spread along both axes for the space filling approach, suggesting broader coverage of the solution space. However, the connection plots in figure 7C and 7D revealed that the total distances from the optimal center points were nearly identical, indicating that both approaches provided similar coverage of the solution space. Despite the BO approach showing a bias towards specific media formulation compositions, particularly favoring the lower left quadrant of the design space, both datasets demonstrated comparable diversity. Importantly, the BO approach was still able to identify three higher-performing media formulations in this batch experiment, as illustrated in Figure 6B.The highest producing space filling experiment was still approximately 50% percent of these 3 high performing subsets of media. Low glutamine concentrations and moderate levels of phenylalanine, depending on the CO2 level, tended to produce higher titers in our CHO cells when using the BO algorithm, having learned from the first 5 initial seed data. As a result, the BO was able to examine and design media that are in the 'high producer' region of the design space, illustrating the principal advantage of BO. Interestingly, the BO algorithm selected more low-glutamine and phenylalanine formulations, as indicated by the cluster of points in the lower left quadrant of 7A, and the left downward-shift of the centerpoint in 7C. These nutrient conditions were indeed favored by high producing processes in past studies[15,30,42]. In contrast, the space filling experimental designs did not learn from the results of previous experiments but rather selected designs from an unsampled design space. Furthermore, the poorly performing media formulations from the BO could also be applied in future optimization scenarios in which BO-designed formulations avoid this low titer design space. Overall, this comparison has demonstrated that with the same amount of initial data and resources, a BO algorithm based on simulations to identify parameters, together with preliminary experimental test cases was able to select enhanced media formulation targets superior to those obtained using conventional space filling approaches.

4. **Discussion:**

This study highlights the effectiveness of uncertainty-based machine learning models in optimizing bioprocess experiments. By leveraging *in-silico* mechanistic modeling, we demonstrated the feasibility and effectiveness of Bayesian Optimization (BO) for designing media tailored to generate high-titer mammalian cell cultures. We refined BO to address specific needs in effective bioprocess optimization, investigating the impact of hyperparameters such as 'explore-exploit **κ**' and 'smoothness,' while elucidating optimal parameters for media design.

Additionally, we implemented a multi-recommendation algorithm to facilitate parallel experimental recommendations, crucial for the inherently time-consuming nature of cell culture experiments. Our experimental validation with automated bioreactors confirmed that

BO outperforms traditional space-filling frameworks, achieving superior media designs with an equivalent number of experiments. Furthermore, incorporating thermodynamics-based constraints addressed feasibility issues related to media precipitation.

Given the growing demand for biopharmaceuticals in rapidly compressed time frames for both competitive advantage and addressing emergency deployment requirements such as for Covid-19 or future pandemics, efficient and rapid biomanufacturing design development, optimization, and implementation are paramount. Data-driven, machine learning-guided approaches, like the one developed here, enhance the speed and efficiency of biomanufacturing development by effectively utilizing and building upon existing data sets. These methods, especially when integrated with physics-based knowledge bases, such as the thermodynamics algorithm and kinetic models employed in this study, offer significant opportunities for enhancing the potential for bioproduction capabilities across a range of organisms over significantly shorter development timelines.

This framework can also be extended to more complex bioprocess modes such as fed-batch and perfusion cultures. In these systems, additional process variables—including pH, dissolved oxygen (DO), temperature, feed volume, and feed timing—play a critical role and can be incorporated into the optimization model. However, the response titer and growth in such dynamic systems can be more nuanced. For example, a lower nutrient input might result in a higher titer trajectory early in the process but lead to earlier cell decline, ultimately yielding a similar final titer compared to other strategies. These trade-offs introduce complexity in defining and optimizing objective functions, emphasizing the importance of carefully engineered surrogate models and decision criteria. Nonetheless, the adaptability of Bayesian Optimization and its ability to incorporate complex inputs make it a promising tool for tackling the challenges of fed-batch and perfusion bioprocess optimization.

In turn, this approach can be used synergistically with automation-based approaches that enable high throughput testing to evaluate the validity of these predictive capabilities. This combination of machine-learning based advanced experimental design combined with automation approaches paves the way for an exciting future in automated laboratories with ML based optimization techniques, driving innovation in media formulation and other applications for improving bioprocess efficiencies for next generation biotherapeutics and bioproducts manufacturing.

Resource Availability

Lead contact: Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Professor Michael Betenbaugh (beten@jhu.edu)

Data availability: Raw cell growth curves and corresponding titer measurements generated during media optimization experiments are provided in Supplementary Data table S1.

Code availability: We have uploaded to github with accessibility available upon request. The algorithm was developed under an active Industry/University partnership. As suggested in the cell press' author guide, we will grant access to interested researchers upon request and approval.
Any additional information required to reanalyze the data reported in this article is available from the lead contact upon request.

Materials availability: This study did not generate new unique reagents.

Limitation of Study

The experimental validation involved a batch culture format and a restricted number of amino acids (phenylalanine and glutamine), which may not fully capture the complexity of fed-batch or perfusion processes common in industrial settings. Second, the surrogate Gaussian Process models rely on relatively small training datasets, which may limit its extrapolation to new process variables or cell lines. Future work will expand to different processing platforms and include larger datasets and more cell lines to validate broader scope and applicability.

Acknowledgements

This work was funded and supported by the Advanced Mammalian Biomanufacturing Innovation Center (AMBIC) through the Industry-University Cooperative Research Center Program under U.S. NSF grant number 2100800. We would like to express our gratitude to all AMBIC member companies for their mentorship.

Declaration of interests: The authors declare no competing interests.

Author contributions:

Nelson Ndahiro: Conceptualization, Data Curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing; Edward Ma: Conceptualization, Data Curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing; Tom Bertalan: Formal analysis, Investigation, Methodology; Marc Donohue: Conceptualization, Methodology,Writing - review & editing; Yannis Kevrekidis: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing - review & editing; Michael J. Betenbaugh:

Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Writing - original draft, Writing - review & editing.

Main Figure Titles and Legends

Graphical Abstract: Model-Guided Thermodynamics-Aware Bayesian Optimization of Cell Culture Medium in Bioprocess

Figure 1. Explore and Exploit tradeoff on Bayesian Optimization(BO) campaign in-silico. Moving from left to right (A-D), there is an increasingly explorative BO (higher k) algorithm. The different BO media composition (Glucose (blue), Asparagine(orange), and Lactate(green)) designs are plotted over each iteration in the 20 sequential experiments, campaign. (E-H)The titer measured after each iteration is also plotted (purple). Despite consistently finding high titer designs, the Kappa=0.1 campaign shows very conservative behavior, spending all the time on extremely similar media compositions. Kappa=1 and 3 are more explorative and while they uncover low titer formulations, they also achieve higher titers than k=0.1 during their campaigns. The Kappa=10 produce very exploratory behavior, appearing erratic with no improvement in titer over time.

Figure 2. Titer of the BO campaigns vs Space-fill. The violin plot shows the titers from all the 20 experiment BO campaigns and the space-fill design, which consists of uniformly sampling the design space. The means are shown by the white line, and low kappas (exploitative) have average titers for their campaigns including the highest performing formulations in the k=1 campaign. The space-fill has the lowest average titer and a wide spread of titers shown by the narrow width of the violin plot.

Figure 3. Multi-recommendation in-silico BO campaigns: The Multi Scale Multi Recommendation (MSMR) algorithm was used to generate multiple BO recommendations with an exploitative kappa value. **A.** A 40 experiment BO campaign was initiated with 4 recommendations per each BO iteration. The alpha values selected by the MSMR were recorded and plotted with run number. The spread of the points shows that the set of alpha values selected during the campaign was diverse. **B.** The performance of each experiment (and associated alpha value) was also recorded and plotted. The plot is further divided into 4 quadrants reflecting combinations of alpha and titer values. Strikingly, half of the high titer recommendations (set to above 4), were of alphas of high value. Conversely, low to moderate titers were in the first quadrant. This suggests that 'low' alpha values can accurately optimize this in-silico bioprocess, implying that the underlying media-titer function is relatively smooth (low alpha).

Figure 4: Impact of Thermodynamic constraints on cell culture media experimental design. **A&B** present scenarios that would be missed by assuming normal solubility of amino acids in mixtures. **A.** Example of 2 amino acids that mutually decrease each other's solubility through intermolecular interactions. This means that some mixtures would result in precipitation even though either individually would be soluble at that concentration. **B.** Example of 2 amino acids where the first amino acid (GLY) increases the solubility of amino acid 2 (Serine) but not vice versa. As a result, there is a region where in the presence of serine is more soluble than expected from it's solubility alone in water.

C. 20 points were sampled uniformly in the experimental space where Phenylalanine and Glutamine are modified in medium. The upper bounds of the samples were the solubility of each amino acids (blue box). The UNIFAC model is used to calculate whether precipitation occurs in the mixtures. Green indicates a soluble mixture while purple indicates precipitation. Of the sampled points 5/20 experiments are outside of feasibility. **D.** 20 points were sampled uniformly similarly for Glutamine and Alanine. Again, the upper bounds of the samples were the solubility of each amino acids (blue box). All samples formulations are feasible but ~13% of the feasible space is ignored by the not taking into account the thermodynamic-based nutrient concentration upper bounds.

Figure 5. Experimental setup for Bayesian and space-filling media optimization**.** The experiment began with media lacking glutamine and phenylalanine. After thawing, cells were cultured in basal media for two passages before being transferred to the AMBR15 automated culture system under varying levels of glutamine, phenylalanine, and CO2. Initial conditions (Run 1) were designed to include half, base, and double (0.5X, 1X, 2X) the base concentrations of glutamine and phenylalanine, with CO2 levels set to 5% or 8%. Measured titer and growth data were used to initialize the Bayesian Optimization algorithm. Simultaneously, a classical space-filling design was performed to enhance solution space coverage and construct a surface response curve. Each method generated 12 experimental designs to evaluate and compare their effectiveness.

Figure 6: Experimental comparison of performance of BO and Space-fill designs. A. Only showing the Glutamax and Phenylalanine concentrations, the selected formulations from the Space-filling experiment was uniformly spread, as expected. On the other hand, the BO experiments show a bias for the lower left quadrant, which are likely areas of high expected titer. Additionally , the BO experiments also showed some exploration of the space. **B.** The BO experiments and the Space-filling experiments were run during two different AMBR runs and so running controls (green*) was necessary to ensure data is reliable and reproducible. The titers obtained from the follow on experiments shows that the BO algorithm was able to obtain the highest titer formulations. The BO also identified low performing formulations (some even worse than the space-filling designs). These experiments show that with a similar dataset the BO was able to achieve higher titers with the same number of experiments. Data are represented as mean.

Figure 7. Directional dispersion and coverage analysis of BO and Space-Filling approaches. (A) Variance ellipse for BO shows a tiger spread along the Phenylalanine axis with shifted center toward lower left quadrant, indicating a sampling bias. (B) The Space-Filling approach displays a more centered distribution with broad spread along both axes. (C) and (D) are connection plots for BO and Space-filling methods that visualize the total distances of each sample point from an optimal center to minimize the total distance. The result showed that the total distances were near identical for both methods, and therefore, a comparable diversity in the solution space, despite BO has a bias toward specific media formulations.

Main Table Titles and Legends

Table 1: Table of all media formulations run on the AMBR, and the corresponding titer and peak VCD. The media formulation of all the experiments ran in bioreactors arain the table above. The initial 5 datapoints denoted as Run 1 shows seed data used to prime the BO and space-filling designs. The BO designs generated using the 5 initial datapoints consist of 6 designs maximizing VCD (orange) and 6 designs maximizing Titer (blue). The Space-fill design that followed the initial run is listed along with corresponding titer and VCD.

STAR Methods

Experimental Model and Study Participants Details

Suspension CHO-K1 VRC01 cells (kindly provided by the National Institute of Health) were cultured in Immediate Advantage custom media (Millipore Sigma, Cat 87093C), with 4mM GlutaMax (Thermo Fisher, Cat: 35-050–061) in 125 mL flat-bottom shake flasks. Cell number was counted using a hemocytometer (Electron Microscopy Sciences) and a light microscope (Zeiss) and seeded at a density of 3×10^6 cells/mL. Cells were cultured in an incubator at 37°C, 5% CO_2, 80% humidity.

Method details

Experimental setup and AMBR runs

The cells, which were adapted to Immediate Advantage custom media (Millipore Sigma), with 4mM GlutaMax (Thermo Fisher), were thawed and cultured in 125 mL flat-bottom shake flasks. Before transfer into the AMBR automated cell-culture system, cells were cultured in a 20-30 mL working volume and passaged to $0.3x10^6$ cells/mL every 3-4 days based on counts taken using trypan blue exclusion. Cells were inoculated into the AMBR at a minimum of 2 passages after thawing into shake flasks to ensure appropriate recovery. The CHO-K1 cells were inoculated into the AMBR in their new BO-designed (or space-filling-designed) formulation for media and reactor adaptation. Once in the AMBR, cells passaged every 3-4 days to $0.3x10^6$ cells/mL by action of the AMBR's robotic arm, for at least 3 passages. This ensured that cells were fully acclimated to the change of media as well as to the AMBR. This is an important step and ensures that product titers measured were not coming from a transient state of cell stress associated with a change of environment (and would thus be hard to control and/or reproduce). Then, the cell culture experiment is initiated, where cells are allowed to grow until cell decline in a batch format (no feeding). During the run, cells are counted every day, recording the cell viability (% alive) and viable cell densities (VCD). Supernatants are collected for titer measurement on days 5 and 6 (typical peak titer before cell density decline).

Growth and Viability measurements

Cell culture samples were collected every 24 hours during the cell culture process to measure cell density and viability. Cells were stained and diluted with a 0.2% trypan blue solution (Gibco). Viable cell density (VCD, E6 cells/mL), which was used to assess cell growth, was measured with a hemocytometer (Electron Microscopy Sciences) and a light microscope (Zeiss). Cell culture viability was monitored using the trypan blue dye exclusion method. Both growth and viability were monitored daily.

HPLC titer measurements

An IgG standard curve was constructed with reagent grade IgG from human serum (Millipore Sigma, Cat: I2511). IgG samples at 5 g/L, 2 g/L, 1 g/L, 0.5 g/L, and 0.2 g/L were prepared by performing serial dilution with HPLC grade water.

The binding buffer consists of 1.9 g sodium phosphate monobasic monohydrate, 9.8 g sodium phosphate diabasic heptahydrate, and 5.84 g sodium chloride per liter of HPLC grade water. The elution buffer consists of 6.8 g sodium phosphate monobasic monohydrate and 5.84 g sodium chloride per liter HPLC grade water with pH adjusted to 2.7 using phosphoric acid.

Cell culture samples were collected on days 5 and 6 of the cell culture and centrifuged at 2000 × g. Supernatants were collected for titer quantification.

Titer quantification was performed via an agilent HPLC system (Agilent, Infinity 1200) using a protein A column (Poros 2 µm, 2.1 × 30 mm; Thermofisher). Samples were injected in the column in two technical replicates. Blank samples were run between every sample. Agilent Chemstation was used for data interpretation.

DOE experimental design by JMP

JMP Version 17.2.0 under a Johns Hopkins University student license was used to suggest experiments based on experimental data. The widely used Space-Filling Designs mode was employed to generate experimental setups to try in the AMBR.

Bayesian Optimization experimental design

All Bayesian optimizations were performed using BayesOpt and Scikit-learn libraries in Python. Gaussian process regression with the Matern kernel implemented in BayesOpt and SciKit Learn was used to explore different kernels and acquisition functions. The Matern kernel with default parameters robustly reproduced the cell culture data as suggested by other publications in the space. Either simulation data from the hybrid model or experimental data

from the lab was used to model the effect of cell medium components on titer and suggest new experiments using this implementation of Gaussian Processes and Bayesian Optimization.

Quantification and statistical analysis

Growth and viability measurements were recorded in Microsoft Excel. For each condition (n = 2 biological replicates), data are reported as mean ± SEM.

For HPLC titer quantification, Agilent ChemStation software was used for manual signal integration. Each condition was run in biological duplicate, and each biological replicate was analyzed in technical duplicate. Processed data were interpreted in Microsoft Excel. Titer data from different batches were normalized to the control condition (cells cultured in Immediate Advantage custom media with 4 mM GlutaMax).

References

1. Kimiz-Gebologlu I, Gulce-Iz S, Biray-Avci C. 2018. Monoclonal antibodies in cancer immunotherapy. *Mol. Biol. Rep.* **45**:2935–2940.
2. Chen BK, Yang YT, Bennett CL. 2018. Why biologics and biosimilars remain so expensive: Despite two wins for biosimilars, the supreme court's recent rulings do not solve fundamental barriers to competition. *Drugs* **78**:1777–1781.
3. Dumont J, Euwart D, Mei B, Estes S, Kshirsagar R. 2016. Human cell lines for biopharmaceutical manufacturing: history, status, and future perspectives. *Crit Rev Biotechnol* **36**:1110–1122. https://www.tandfonline.com/doi/full/10.3109/07388551.2015.1084266.
4. Zhu MM, Mollet M, Hubert RS, Kyung YS, Zhang GG. 2017. Industrial Production of Therapeutic Proteins: Cell Lines, Cell Culture, and Purification. In: . *Handbook of Industrial Chemistry and Biotechnology*. Cham: Springer International Publishing, pp. 1639–1669. http://link.springer.com/10.1007/978-3-319-52287-6_29.
5. Combe M, Sokolenko S. 2021. Quantifying the impact of cell culture media on {CHO} cell growth and protein production. *Biotechnol. Adv.* **50**:107761.
6. Coulet M, Kepp O, Kroemer G, Basmaciogullari S. 2022. Metabolic Profiling of CHO Cells during the Production of Biotherapeutics. *Cells* **11**:1929. https://www.mdpi.com/2073-4409/11/12/1929.
7. Reinhart D, Damjanovic L, Kaisermayer C, Kunert R. 2015. Benchmarking of commercially available CHO cell culture media for antibody production. *Appl Microbiol Biotechnol* **99**:4645–4657. http://link.springer.com/10.1007/s00253-015-6514-4.

8. Sha S, Huang Z, Wang Z, Yoon S. 2018. Mechanistic modeling and applications for CHO cell culture development and production. *Curr Opin Chem Eng* **22**:54–61. https://linkinghub.elsevier.com/retrieve/pii/S2211339818300480.
9. Tang P, Xu J, Louey A, Tan Z, Yongky A, Liang S, Li ZJ, Weng Y, Liu S. 2020. Kinetic modeling of Chinese hamster ovary cell culture: factors and principles. *Crit Rev Biotechnol* **40**:265–281. https://www.tandfonline.com/doi/full/10.1080/07388551.2019.1711015.
10. Chen, Y., McConnell, B. O., Gayatri Dhara, V., Mukesh Naik, H., Li, C.-T., Antoniewicz, M. R., & Betenbaugh, M. J. (2019). An unconventional uptake rate objective function approach enhances applicability of genome-scale models for mammalian cells. Npj Systems Biology and Applications, 5(1), 25. https://doi.org/10.1038/s41540-019-0103-6
11. Hefzi H, Ang KS, Hanscho M, Bordbar A, Ruckerbauer D, Lakshmanan M, Orellana CA, Baycin-Hizal D, Huang Y, Ley D, Martinez VS, Kyriakopoulos S, Jiménez NE, Zielinski DC, Quek L-E, Wulff T, Arnsdorf J, Li S, Lee JS, Paglia G, Loira N, Spahn PN, Pedersen LE, Gutierrez JM, King ZA, Lund AM, Nagarajan H, Thomas A, Abdel-Haleem AM, Zanghellini J, Kildegaard HF, Voldborg BG, Gerdtzen ZP, Betenbaugh MJ, Palsson BO, Andersen MR, Nielsen LK, Borth N, Lee D-Y, Lewis NE. 2016. A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. *Cell Syst* **3**:434-443.e8.
12. Li, C.-T., Yelsky, J., Chen, Y., Zuñiga, C., Eng, R., Jiang, L., Shapiro, A., Huang, K.-W., Zengler, K., & Betenbaugh, M. J. (2019). Utilizing genome-scale models to optimize nutrient supply for sustained algal growth and lipid productivity. Npj Systems Biology and Applications, 5(1), 33. https://doi.org/10.1038/s41540-019-0110-7
13. Yeo HC, Hong J, Lakshmanan M, Lee D-Y. 2020. Enzyme capacity-based genome scale modelling of CHO cells. *Metab Eng* **60**:138–147. https://linkinghub.elsevier.com/retrieve/pii/S1096717620300744.
14. Nolan RP, Lee K. 2011. Dynamic model of CHO cell metabolism. *Metab Eng* **13**:108–124. https://linkinghub.elsevier.com/retrieve/pii/S1096717610000856.
15. Ladiwala P, Dhara VG, Jenkins J, Kuang B, Hoang D, Yoon S, Betenbaugh MJ. 2023. Addressing amino acid-derived inhibitory metabolites and enhancing {CHO} cell culture performance through {DOE-guided} media modifications. *Biotechnol. Bioeng.* **120**:2542–2558.
16. Frazier PI. 2018. Bayesian Optimization. In: . *Recent Advances in Optimization and Modeling of Contemporary Problems*. INFORMS, pp. 255–278. http://pubsonline.informs.org/doi/10.1287/educ.2018.0188.
17. Greenhill S, Rana S, Gupta S, Vellanki P, Venkatesh S. 2020. Bayesian optimization for adaptive experimental design: A review. *IEEE Access* **8**:13937–13948.

18. Kumar A, Pant KK, Upadhyayula S, Kodamana H. 2023. Multiobjective Bayesian optimization framework for the synthesis of methanol from syngas using interpretable Gaussian process models. *ACS Omega* **8**:410–421.
19. Merzbacher C, Mac Aodha O, Oyarzún DA. 2023. Bayesian optimization for design of multiscale biological circuits. *ACS Synth. Biol.* **12**:2073–2082.
20. Shields BJ, Stevens J, Li J, Parasram M, Damani F, Alvarado JIM, Janey JM, Adams RP, Doyle AG. 2021. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**:89–96.
21. Wang Y, Chen T-Y, Vlachos DG. 2021. {NEXTorch}: A design and Bayesian optimization toolkit for chemical sciences and engineering. *J. Chem. Inf. Model.* **61**:5312–5319.
22. Borgli RJ, Kvale Stensland H, Riegler MA, Halvorsen P. 2019. Automatic Hyperparameter Optimization for Transfer Learning on Medical Image Datasets Using Bayesian Optimization. In: . *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*. IEEE, pp. 1–6. https://ieeexplore.ieee.org/document/8743779/.
23. Liang Q, Gongora AE, Ren Z, Tiihonen A, Liu Z, Sun S, Deneault JR, Bash D, Mekki-Berrada F, Khan SA, Hippalgaonkar K, Maruyama B, Brown KA, Fisher III J, Buonassisi T. 2021. Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. *NPJ Comput Mater* **7**:188. https://www.nature.com/articles/s41524-021-00656-9.
24. Cui T, Bertalan T, Ndahiro N, Khare P, Betenbaugh M, Maranas C, Kevrekidis IG. 2024. Data-driven and physics-informed modeling of Chinese Hamster Ovary cell bioreactors. Comput. Chem. Eng. 183:108594.
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research* **12**:2825–2830.
26. Fernando Nogueira. 2014. Bayesian Optimization: Open source constrained global optimization tool for Python. *https://github.com/fmfn/BayesianOptimization*. https://github.com/fmfn/BayesianOptimization.
27. Joy TT, Rana S, Gupta S, Venkatesh S. 2020. Batch Bayesian optimization using multi-scale search. *Knowl. Based Syst.* **187**:104818.
28. Fan Y, Jimenez Del Val I, Müller C, Wagtberg Sen J, Rasmussen SK, Kontoravdi C, Weilguny D, Andersen MR. 2015. Amino acid and glucose metabolism in fed-batch {CHO} cell culture affects antibody production and glycosylation. *Biotechnol. Bioeng.* **112**:521–535.
29. Kirsch BJ, Bennun S V., Mendez A, Johnson AS, Wang H, Qiu H, Li N, Lawrence SM, Bak H, Betenbaugh MJ. 2022. Metabolic analysis of the asparagine and glutamine dynamics in

an industrial Chinese hamster ovary fed-batch process. *Biotechnol Bioeng* **119**:807–819. https://onlinelibrary.wiley.com/doi/10.1002/bit.27993.

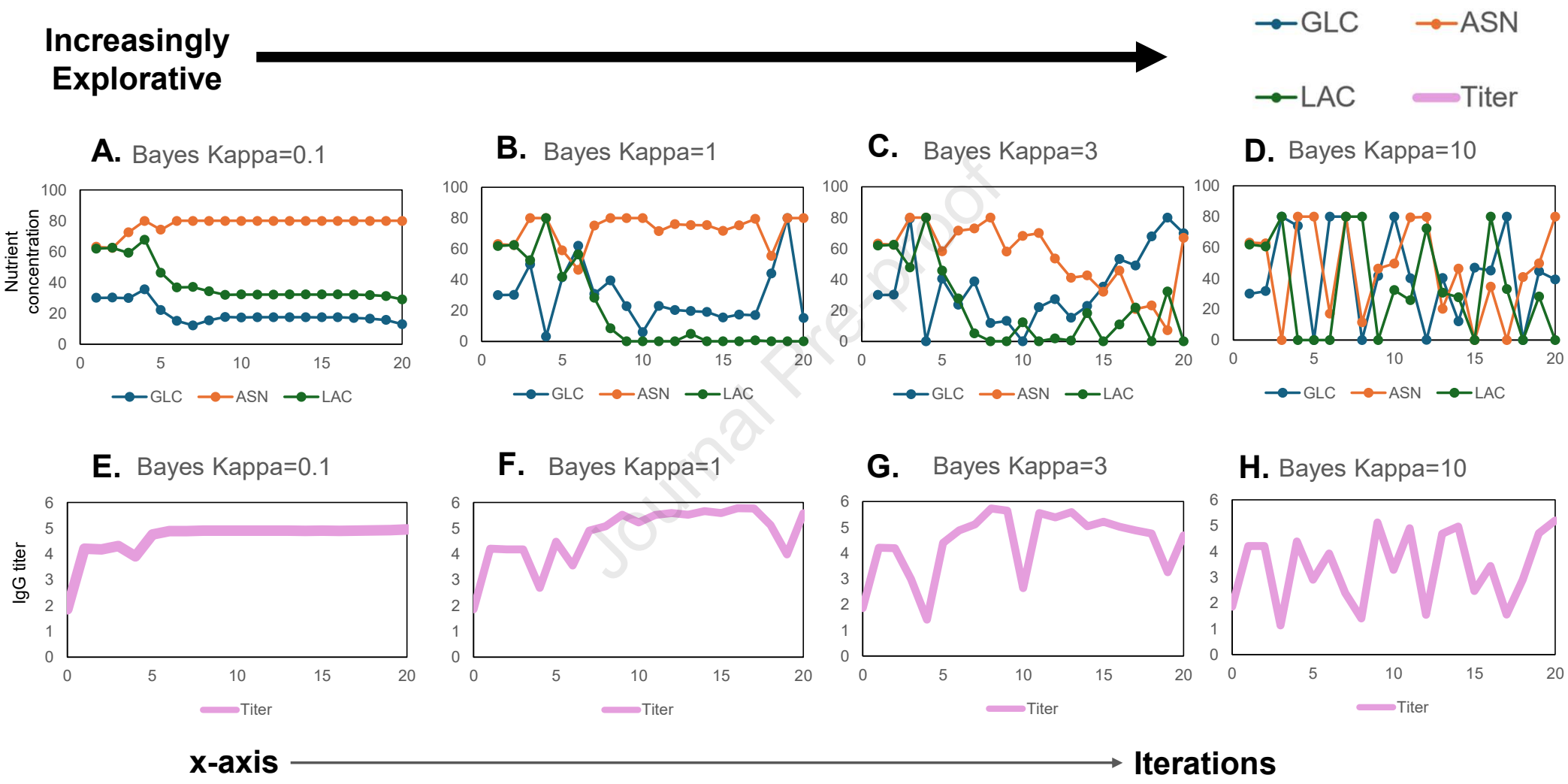
30. Xu P, Dai X-P, Graf E, Martel R, Russell R. 2014. Effects of glutamine and asparagine on recombinant antibody production using {CHO-GS} cell lines. *Biotechnol. Prog.* **30**:1457–1468.
31. Zhang L-X, Zhang W-Y, Wang C, Liu J-T, Deng X-C, Liu X-P, Fan L, Tan W-S. 2016. Responses of {CHO-DHFR} cells to ratio of asparagine to glutamine in feed media: cell growth, antibody production, metabolic waste, glutamate, and energy metabolism. *Bioresour. Bioprocess.* **3**.
32. Quek L-E, Dietmair S, Krömer JO, Nielsen LK. 2010. Metabolic flux analysis in mammalian cell culture. *Metab. Eng.* **12**:161–171.
33. Ginsbourger D, Le Riche R, Carraro L. 2010. Kriging Is Well-Suited to Parallelize Optimization. In: , pp. 131–162. http://link.springer.com/10.1007/978-3-642-10701-6_6.
34. Gonzalez, J., Dai, Z., Hennig, P. & Lawrence N. 2016. Batch Bayesian Optimization via Local Penalization. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research*:648–657. https://proceedings.mlr.press/v51/gonzalez16a.html.
35. Forte T, Grinnell C, Zhang A, Polilli B, Leshinski J, Khattak S. 2023. Methods for identifying precipitates and improving stability of chemically defined highly concentrated cell culture media. *Biotechnol Prog* **39**. https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3345.
36. Stone AT, Dhara VG, Naik HM, Aliyu L, Lai J, Jenkins J, Betenbaugh MJ. 2021. Chemical speciation of trace metals in mammalian cell culture media: looking under the hood to boost cellular performance and product quality. *Curr Opin Biotechnol* **71**:216–224. https://linkinghub.elsevier.com/retrieve/pii/S0958166921001506.
37. Mulukutla BC, Kale J, Kalomeris T, Jacobs M, Hiller GW. 2017. Identification and control of novel growth inhibitors in fed-batch cultures of Chinese hamster ovary cells. *Biotechnol. Bioeng.* **114**:1779–1790.
38. Darja O, Stanislav M, Saša S, Andrej F, Lea B, Branka J. 2016. Responses of CHO cell lines to increased pCO2 at normal (37 °C) and reduced (33 °C) culture temperatures. *J Biotechnol* **219**:98–109. https://linkinghub.elsevier.com/retrieve/pii/S0168165615302133.
39. Lee AP, Kok YJ, Lakshmanan M, Leong D, Zheng L, Lim HL, Chen S, Mak SY, Ang KS, Templeton N, Salim T, Wei X, Gifford E, Tan AH, Bi X, Ng SK, Lee D, Ling WLW, Ho YS. 2021. Multi-omics profiling of a CHO cell culture system unravels the effect of culture pH on cell growth, antibody titer, and product quality. *Biotechnol Bioeng* **118**:4305–4316. https://onlinelibrary.wiley.com/doi/10.1002/bit.27899.
40. JMP®, Version 17.2.0. SAS Institute Inc., Cary, NC, 1989–2024.

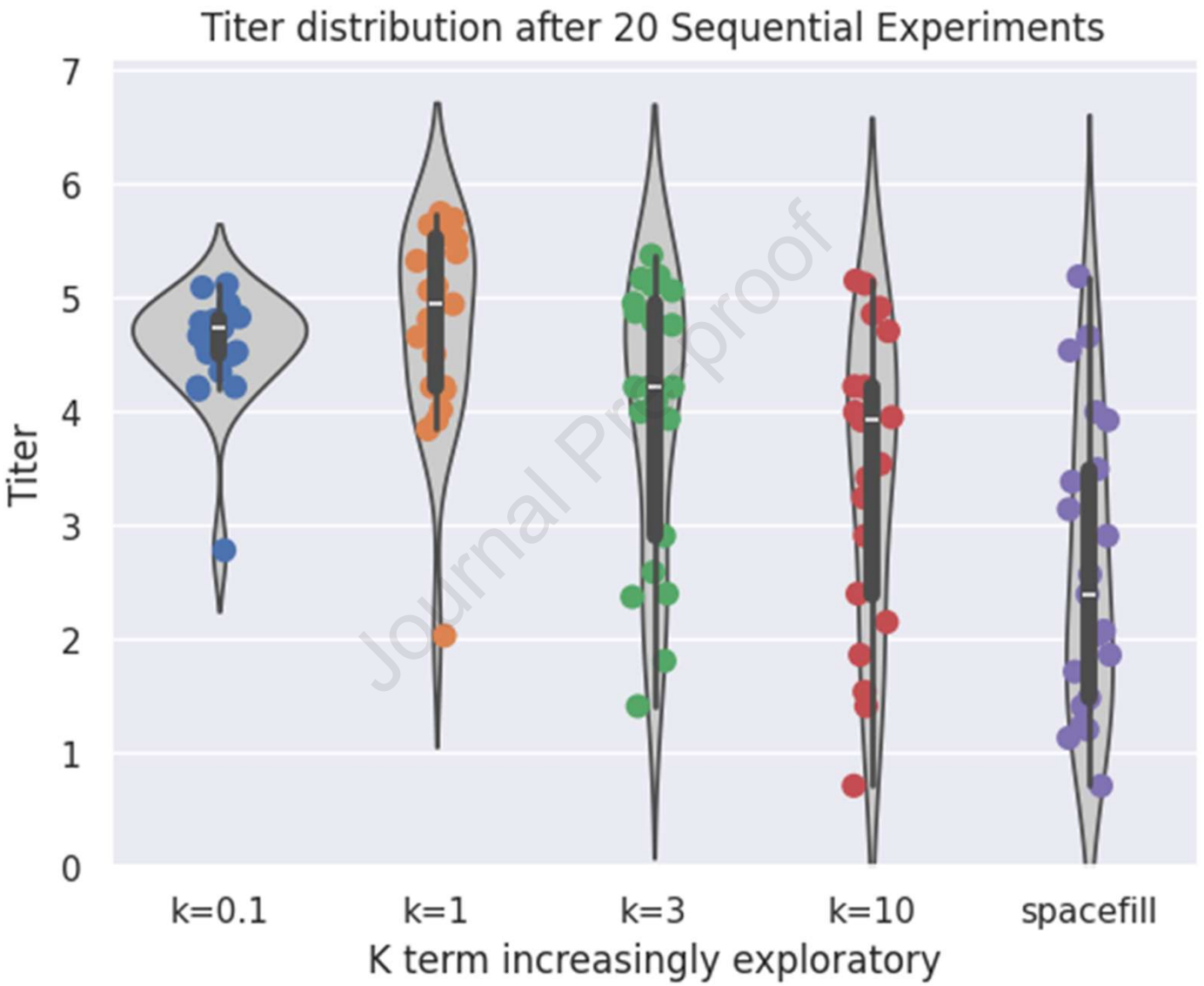
41. Owen, GJ, Chmielewski, AM. 1985. On Canonical Variates Analysis and the Construction of Confidence Ellipses in Systematic Studies. Systematic Zoology, Vol. 34, No. 3.
42. Chen P, Harcum SW. 2005. Effects of amino acid additions on ammonium stressed CHO cells. *J Biotechnol* **117**:277–286.

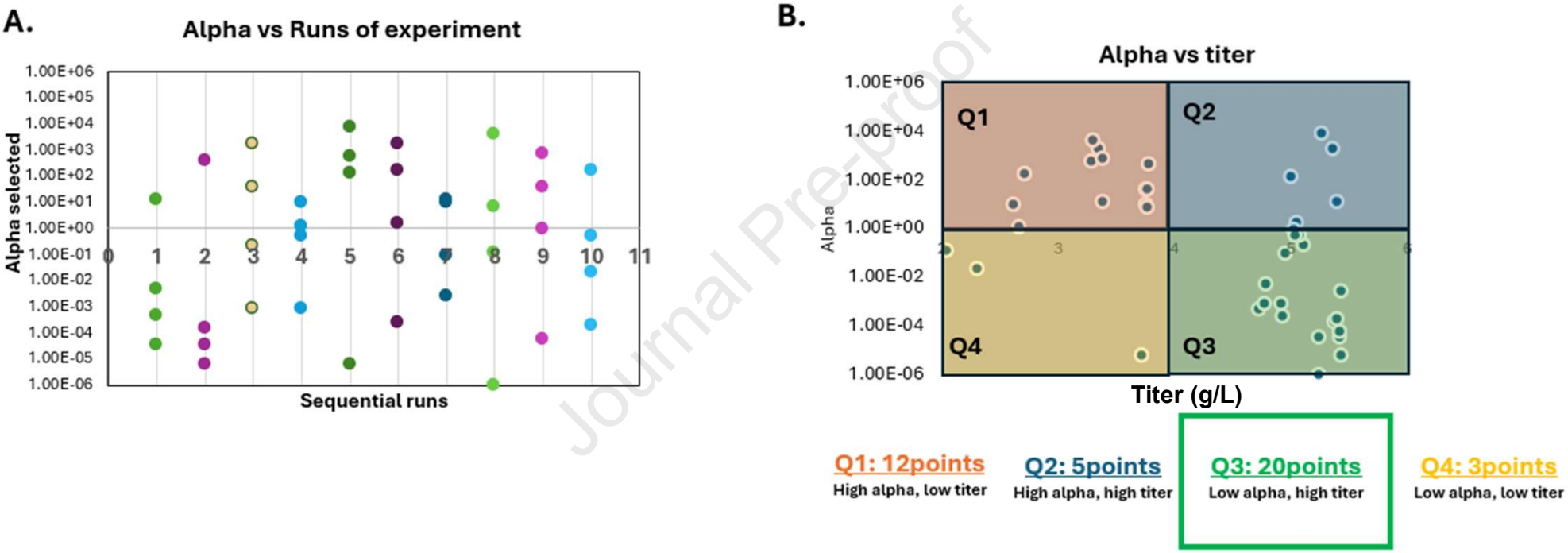
	CO2(%)	Phenylalanine (mM)	Glutamine (mM)	Peak VCD (10^6 cells/mL)	Titer (Day 5) (g/L)
Run 1	5.00	0.77	6.00	13.5	1.36
	8.00	1.56	3.00	15.6	1.39
	5.00	3.08	3.00	15.0	1.60
	8.00	1.56	12.0	9.80	0.816
	8.00	1.56	6.00	11.9	0.763

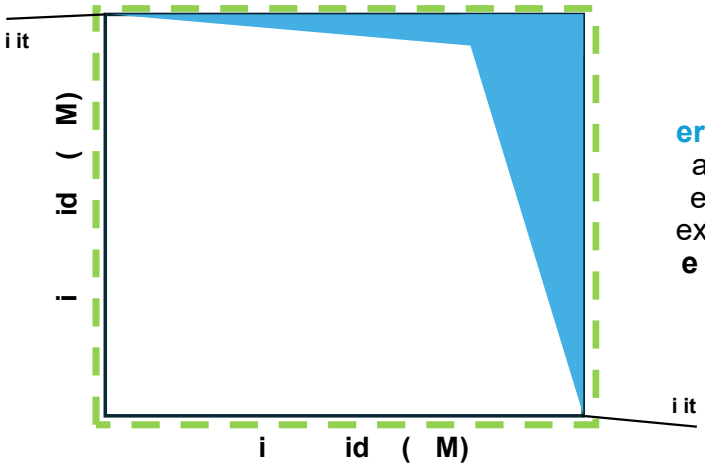
	CO2(%)	Phenylalanine (mM)	Glutamine (mM)	Peak VCD (10^6 cells/mL)	Titer (Day 5) (g/L)
Run 2 (Bayes)	VCD Objective				
	8.00	6.14	11.10	12.60	1.09
	8.00	0.11	6.92	0.26	0.12
	8.00	1.01	2.16	9.74	2.55
	5.00	8.46	2.49	11.23	2.80
	8.00	6.46	15.16	12.33	1.05
	5.00	2.26	18.18	10.13	1.23
	Titer Objective				
	8.00	0.28	19.90	5.90	0.52
	5.00	1.36	0.50	7.16	2.47
	5.00	0.30	7.00	3.29	0.80
	8.00	0.55	5.96	11.50	1.36
	5.00	0.10	10.80	0.11	0.13
	8.00	0.52	19.90	9.18	0.76

	CO2(%)	Phenylalanine (mM)	Glutamine (mM)	Peak VCD (10^6 cells/mL)	Titer (Day 5) (g/L)
Run 2 (Space-fill)	8.00	3.38	19.97	8.62	0.78
	8.00	9.99	18.38	7.14	0.82
	8.00	0.55	5.96	6.42	1.45
	8.00	8.57	15.27	8.83	0.73
	8.00	0.16	0.32	1.26	0.60
	8.00	7.17	8.33	9.90	1.04
	5.00	4.28	10.51	7.82	1.31
	5.00	5.83	1.45	6.30	1.37
	5.00	9.52	5.25	-	-
	5.00	1.56	6.00	10.02	1.37
	5.00	6.34	13.09	7.49	0.96
	5.00	2.12	17.10	8.83	0.90
	5.00	2.12	17.10	10.62	1.23

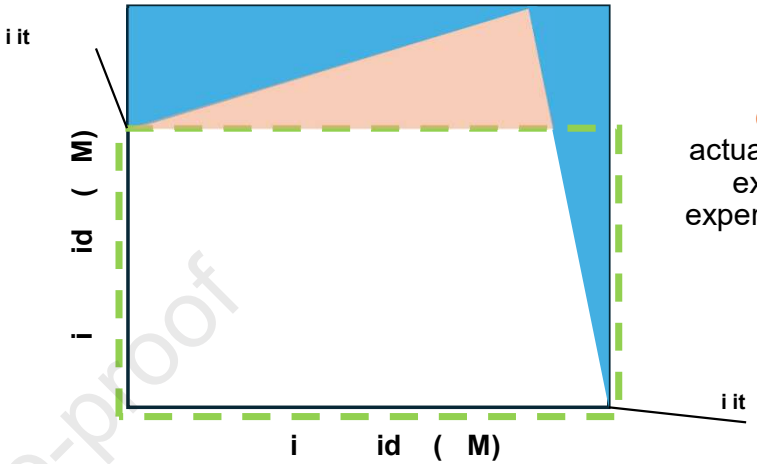






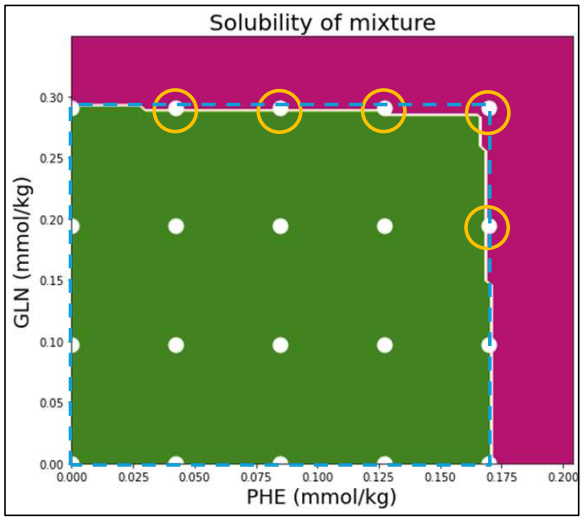


rre t r
ere ti te the actual available space to experiment. Some experiments re t e i e t re ti idered.



rre t r
dere ti te the actual available space to experiment. Some experiments re e i e d re t idered

e i e d G t i e



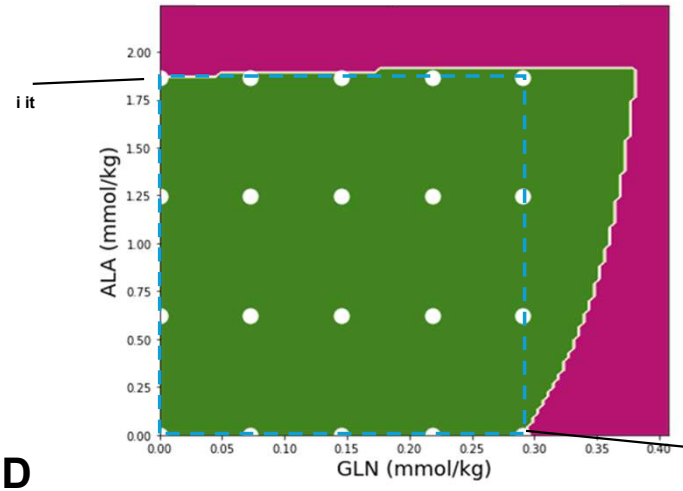
e

l e

re

i e

G t i e i e



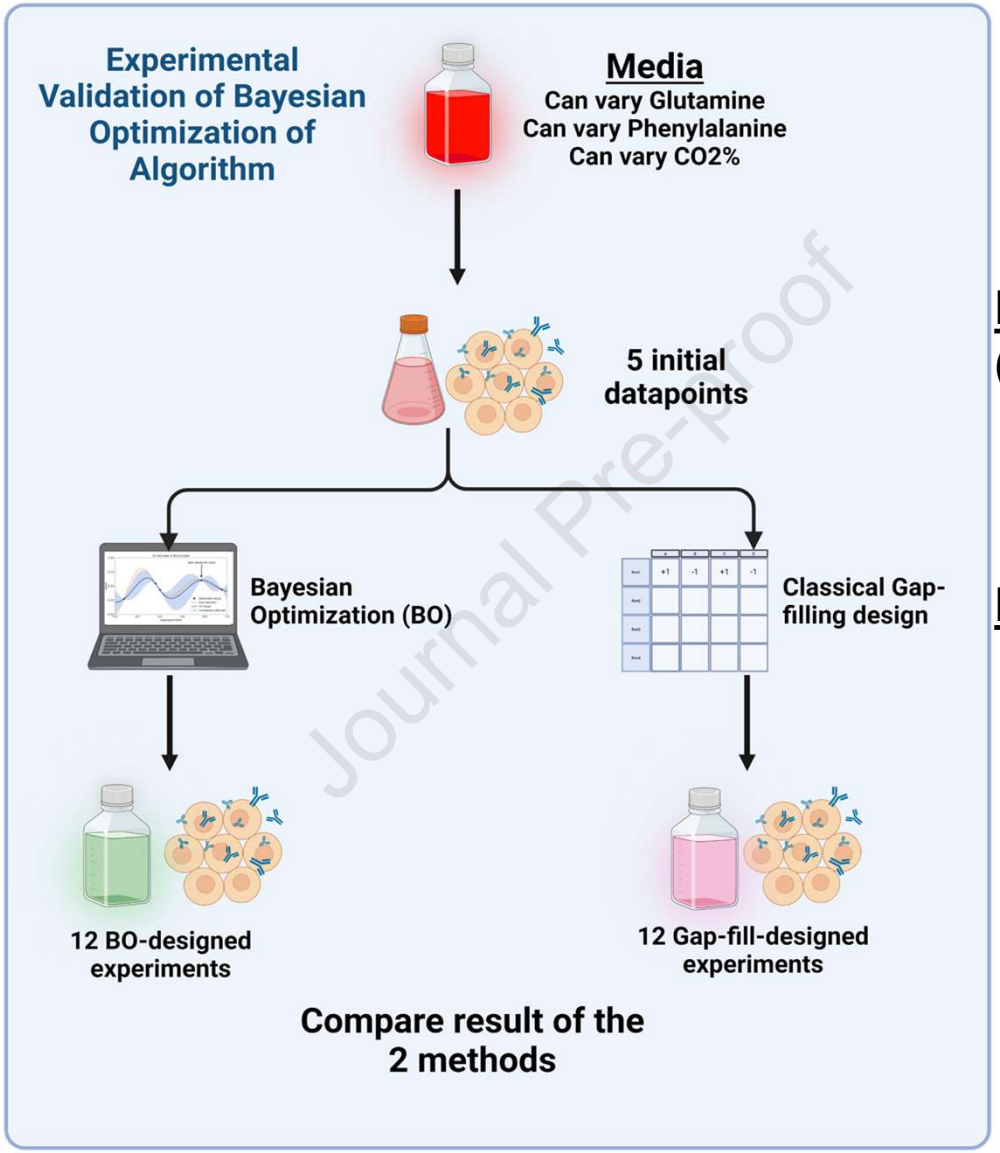
D

e i e

ig red

l e i e

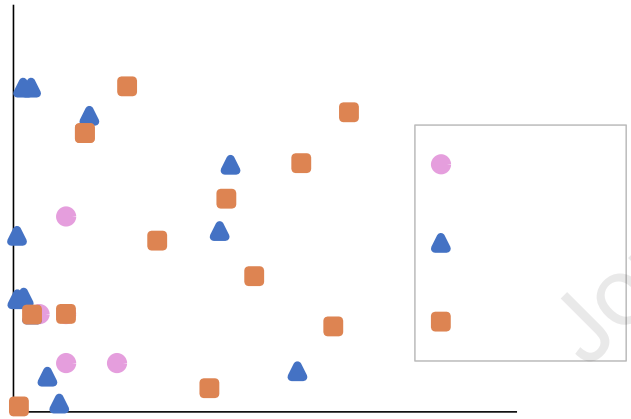
i it
G



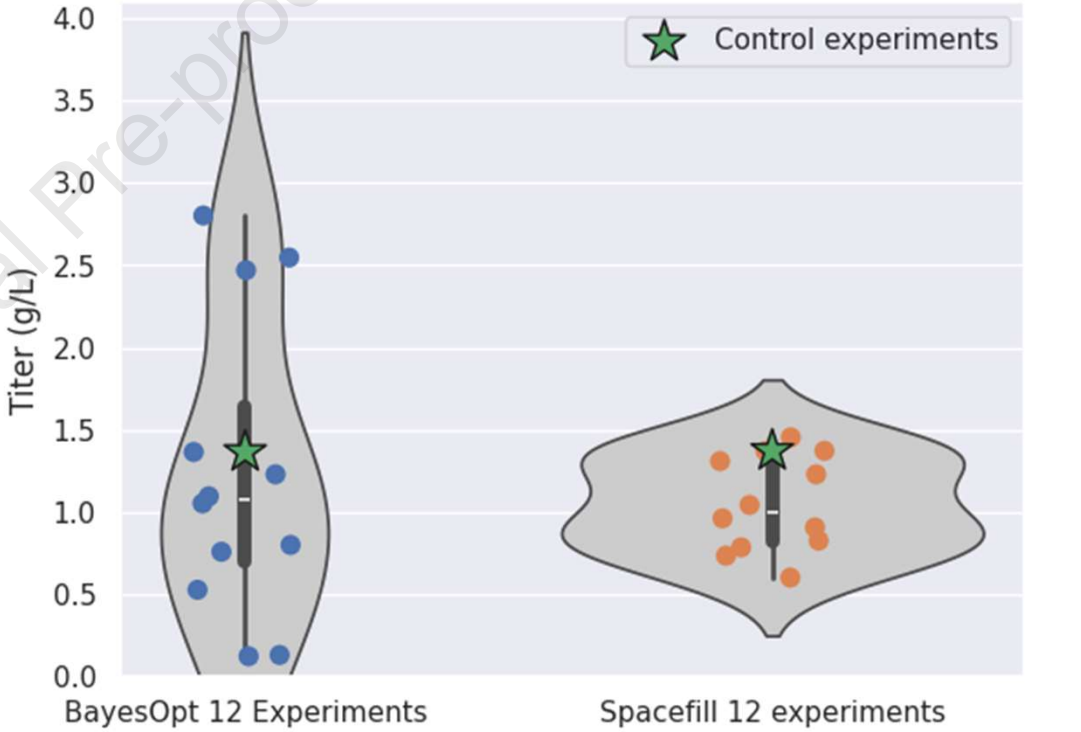
R
(l iti d t et)

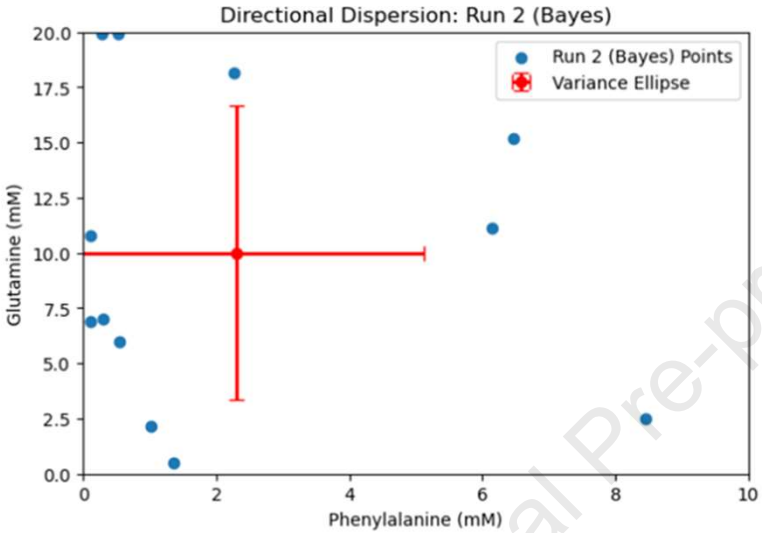
R
G
i

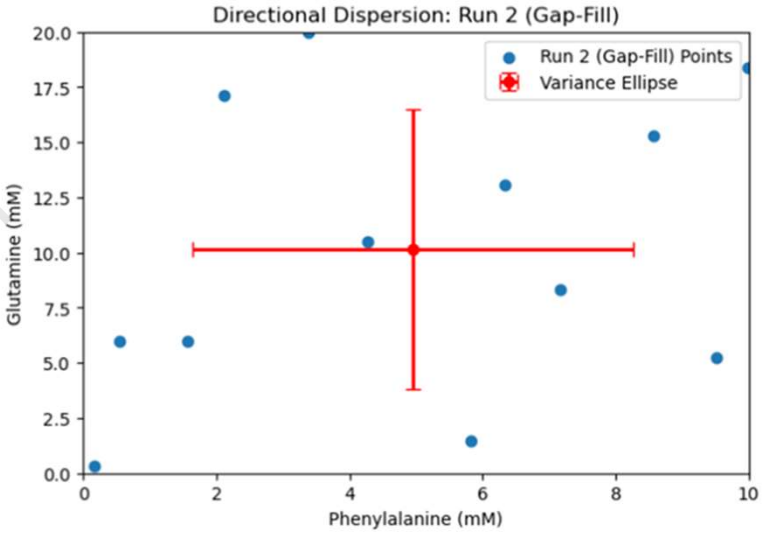
A.

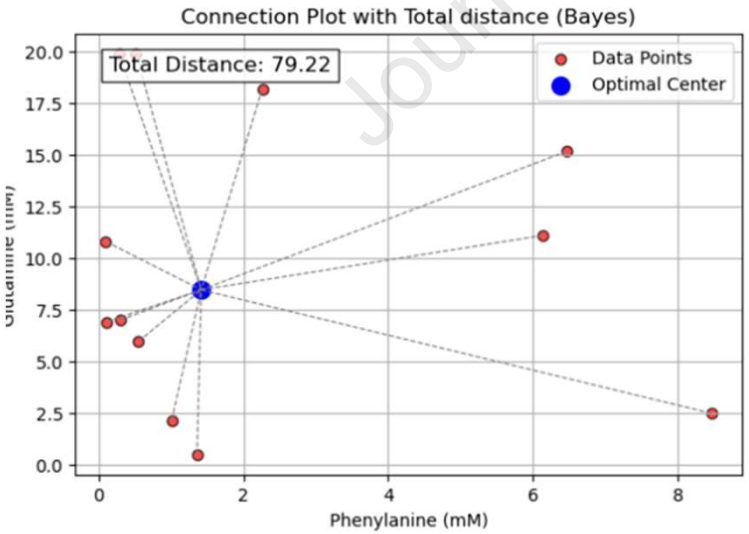


B.

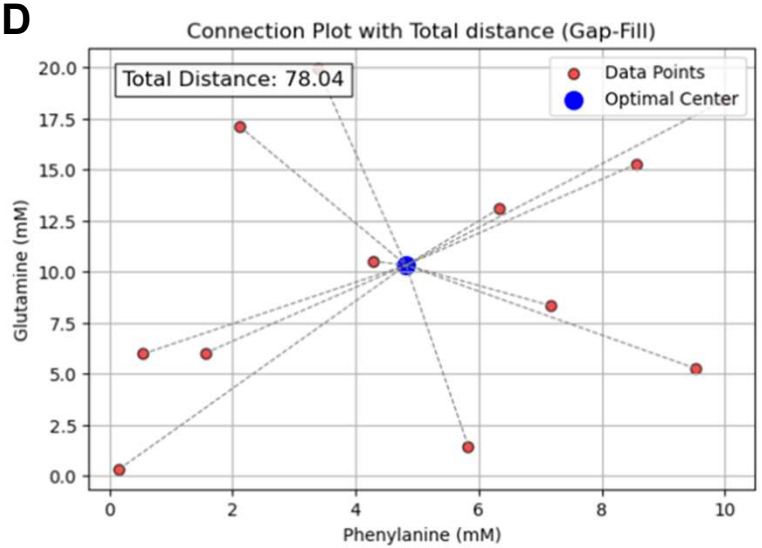








D



Highlights:

- Integrated Bayesian optimization (BO) with thermodynamic constraints for CHO media design
- In-silico fine-tuning and validation of BO algorithm with metabolic hybrid model
- Thermodynamics-awareness avoids amino acid precipitation, ensuring feasible formulations
- Experimental validation with batched-BO in AMBR bioreactors yields higher titers than DOE

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Chemicals, peptides, and recombinant proteins		
Immediate Advantage custom media	Millipore Sigma	Cat#87093C
GlutaMAX	Gibco (Thermo Fisher Scientific)	Cat#35050061
Trypan Blue 0.2%`	Gibco (Thermo Fisher Scientific)	Cat#15250061
L-Phenylalanine	Millipore Sigma	Cat#5202
Critical commercial assays		
Poros Protein A Column (2 µm, 2.1 × 30 mm)	Thermo Fisher Scientific	Cat#1130202E
Deposited data		
Code in a private github repository	jma73@jh.edu	Access upon request
Experimental models: Cell lines		
CHO-K1 cell line	National Institute of Health (NIH)	CRL-9618
Experimental models: Organisms/strains		

Oligonucleotides		
Recombinant DNA		
Software and algorithms		
JMP Pro 17	JMP (SAS Institute)	https://www.jmp.com
Scikit-Learn	(Pedregosa et al., 2011)	https://scikit-learn.org
BayesOpt	(Fernando Nogueira, 2014)	https://github.com/fmfn/BayesianOptimization
Python 3.8	Python Software Foundation	https://python.org
MATLAB (R2022b)	The MathWorks, Inc	https://www.mathworks.com/
Other		
Countess Hemacytometer	life technologies	19350
Countess 3	Invitrogen	AMQAX2000
Infinity 1260 Vialsampler	Agilent Technologies	G7129A
Infinity 1260 Quat Pump	Agilent Technologies	G1311B
Infinity 1260 TCC	Agilent Technologies	G1316A
Infinity 1260 DAD	Agilent Technologies	G4212B
Infinity 1260 FLD	Agilent Technologies	G1321B