Sequential Learning Methods for the Experimental Optimization of Cell Culture Media for Cellular Agriculture

By

ZACHARY ANTHONY COSENZA
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

CHEMICAL ENGINEERING

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

David E Block (Chair)

Karen McDonald

Keith Baar

Committee in Charge

2022

Contents

Abstract

In this dissertation we focus on the application of several design-of-experiments (DOE) methods to cell culture media development in order to sequentially learn optimal media formulations. These sequential DOE methods use data collected from an experimental system to simultaneously improve the model with additional suggested experiments while at the same time learning the optimal conditions of the experimental system. The purpose of this media is for applications in the cellular agriculture industry, where animal cells are grown for consumption. Starting with a hybrid scheme utilizing radial basis functions with a genetic algorithm and coordinate search method, we discovered that long-term cell growth is not fully correlated with the short-term chemical assays typically used in cell culture. We solved this by successfully deploying a Bayesian model that correlates long and short-term growth assays. We could then predict the information value of new experiments *and* assays jointly, reducing the overall number of experiments needed to solve the optimization problem. This improved Bayesian methodology focuses long-term experiments only on the most promising areas of the design space while allowing simpler short-term growth experiments to fully explore the design space. Using this new approach, we designed a medium with 181% more cell growth than a common commercial formulation with a similar economic cost, while doing so in 38% fewer experiments than an efficient DOE method using a desirability function to parameterize the outcome space. This medium even managed to maintain robust cell growth over four passages. Next, we used a hypervolume function to design experiments to sequentially learn the trade-off between cell growth and media cost in a serum-free system. We found a medium with a 184% improvement in growth over the control at a 71% increase in cost that maintained a high level of cell growth over five passages. Both optimal formulations resulted in robust long-term proliferation of cells, indicating the success of our multi-assay Bayesian approach to optimizing media. Future work could tie imaging software, bio-marker quantification, and techno-economic analysis to improve the accuracy and usefulness of predictions and experimental designs.

Acknowledgments

I'd first like to thank my advisor David E Block and co-advisor Keith Baar. Both are great mentors that imparted their experience while allowing me the freedom to pursue my goals. I'd also like to acknowledge the entire Baar lab, graduates, undergrads, and post-docs, for teaching me everything I know about cell culture and being wonderful lab-mates. I must also thank the members of the Department of Chemical Engineering and Material Science for giving me the resources to thrive, particularly the esprit de corps of the graduate students who have become my close friends and colleagues. The University of California and City of Davis should also be acknowledged for creating such a unique and supportive environment to foster advances in science and engineering, as well as the University of Minnesota Twin Cities for instantiating within me a passion for science. I was generously funded through a fellowship by New Harvest, whose entire organization I am eternally grateful. Finally, I'd like to thank my family and friends for supporting me, especially my mother Laura.

CHAPTER 1

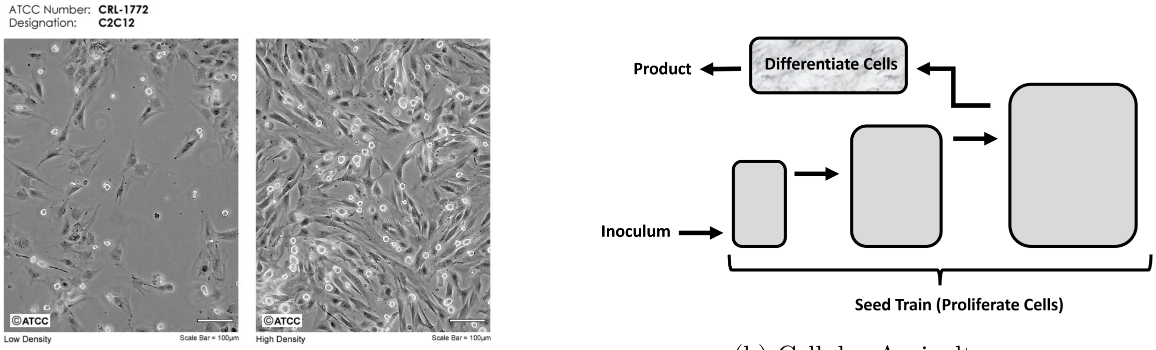
Introduction and Literature Review

This dissertation is focused on exploring experimental optimization methods for the purpose of designing cell culture media for cellular agriculture. This chapter will act as an introduction to the fields of both experimental optimization and cellular agriculture.

1.1. Review of Cellular Agriculture and Cell Culture Media

In cellular agriculture, meat (and other animal products) is derived from populations of cells for consumption [**63**] in an attempt to be more resource efficient and ethical than traditional animal agriculture.

1.1.1. Basic Cellular Agriculture Process. Myoblast, myocytes, and fibroblasts (muscle cells) are cells of greatest interest for the field of cellular agriculture. For texture and taste, adipocytes (fat cells) may be used [**59**] and grown either separately or co-cultured with muscle cells. The choice of animal will also have an effect on the final product and production process because cells from different animals will have different growth characteristics, morphology, and product qualities. The majority of these cell lines are adherent, meaning they require a suitable substrate (surface) to grow. Ideally, cells may be grown in suspension culture (no surface), bringing cellular agriculture in line with typical pharmaceutical practice such as CHO cells [**72**]. Microcarriers (small surfaces in suspension) may also be used to increase the surface area of the total surface [**86**]. Proliferating many cells is not the only consideration in cellular agriculture. Stem cells differentiate into more complex tissue structures depending on time and environmental conditions, which is critical in forming final products that consumers are willing to purchase. For example, C2C12 immortalized murine skeletal muscle cells differentiate into myotubes at high density and when exposed to DMEM + 2% horse serum (Figure 1.1a shows proliferating versus differentiating C2C12 cells). However, because cell differentiation often precludes further proliferation, cells must be periodically passaged (also called sub-cultured) to provide more physical space for growth. This



(b) Cellular Agriculture

(a) C2C12 Cells

FIGURE 1.1. (a) Immortalized cells such as C2C12 (ATCC) have found great use as test-cases for cellular agriculture. Left and right show a low density proliferating culture and high density differentiating culture respectively. (b) This figure was inspired by [**77**] discussing a theoretical cellular agriculture system: Starting with a seed train to proliferate a small population of cells into a large population, the cells then differentiate into their respective final tissue structures, which are harvested and sold.

is typically done by detaching the cells from the substrate using trypsin enzyme and physically placing the cells onto additional surface area. Fundamental techniques in cell culture can be found in [**60**] and a general overview of mammalian cell culture for bio-production uses can be found in ([**8**] pg. 157 - 195). Figure 1.1b shows a high level overview of the cellular agriculture process. Throughout this entire process, media is used to support cells by providing them with nutrients, signal molecules, and an environment for growth. We are focused on reducing the cost of the media while supporting cell proliferation. This is because the media has been identified as the largest contribution to cost (according to [**78**] 55% to over 95% of the marginal cost of the final product). The main considerations for the design of cell culture media in cellular agriculture are (i) the media must be inexpensive, (ii) it must be free of animal products, and (iii) it must support long-term proliferation of relevant cell lines and final differentiation into relevant products.

1.1.2. Media in Cellular Agriculture. The most basic part of a cell culture medium is the basal component, which supplies the amino acids, carbon sources, vitamins, salts, and other fundamental building blocks to cell growth. The optimal pH of cell culture media is around 7.2 - 7.4 which is achieved through buffering with the sodium bicarbonate - (5% - 10%) CO_2 or organic buffers like HEPES. Temperature should be maintained at around 37°C at high humidity to prevent evaporation of media. Osmolarity around 260 - 320 mOsm/kg is maintained by the concentration of inorganic ions salts such as NaCl as well as hormones and other buffers. Inorganic salts also supply potassium, sodium, and calcium to regulate cell membrane potential which is critical for nutrient transport and signalling. Trace metals such as iron, zinc, copper, and selenium are also found in basal media for a variety of tasks like enzyme function [**3**]. Vitamins, particularly B and C, are found in many basal formulations to increase cell growth because they cannot be made by the cells themselves. Nitrogen sources, such as essential and non-essential amino acids, are the building blocks of proteins so are critical to cell growth and survival. Glutamine in particular can be used to form other amino acids [**57**] and is critical for cell growth. It is also unstable in water so is typically supplemented into media as L-alanyl-L-glutamine dipeptide (sold as GlutaMAX). Carbon sources, primarily glucose and pyruvate, are essential as they are linked to metabolism through glycolysis and the pentose-phosphate pathway [**59**]. Fatty acids like lipoic and linoleic acid act as energy storage, precursor molecules, and structural elements of membranes and are sometimes supplied through a basal medium like Ham's F12. Having a sufficient concentration of all of these components is required for proliferating mammalian cells across multiple passages as per (iii) above.

Having a robust basal media is a necessary but not sufficient condition for long-term cell proliferation and differentiation. Serum is a critical aspect of cell culture because it provides a mix of proteins, amino acids, vitamins, minerals, buffers and shear protectors (pg 4 [**3**]). Serum stimulates proliferation and differentiation, transport, attachment to and spreading across substrates, and detoxification [**14**]. Serum (often from horse or cow) has large lot-to-lot variability, zoonotic viruses and contamination ([**6**] pg. 18), as well as the ethical issues associated with collecting serum from animals. Therefore, while it often simplifies cell growth and differentiation, it is critical to remove serum as per point (ii). Supplementation with growth factors like FGF2 [**39**], TGFβ1 [**27**], TNFα [**17**], IGF1, or HGF [**40**] is a common way to induce growth of mammalian muscle cells without the use of serum. Transferrin, another protein found in serum, fulfills a transport role for iron into the cell membrane [**3**]. PDGF and EGF are polypeptide growth factors that initiate cell proliferation [**14**]. Such components enhance cell growth but are expensive and comprise the vast majority of the cost of theoretical cellular agriculture processes [**78**]. Much work has been done on developing serum-free media. The E8 / B8 medium [**52**] for human induced pluripotent stem

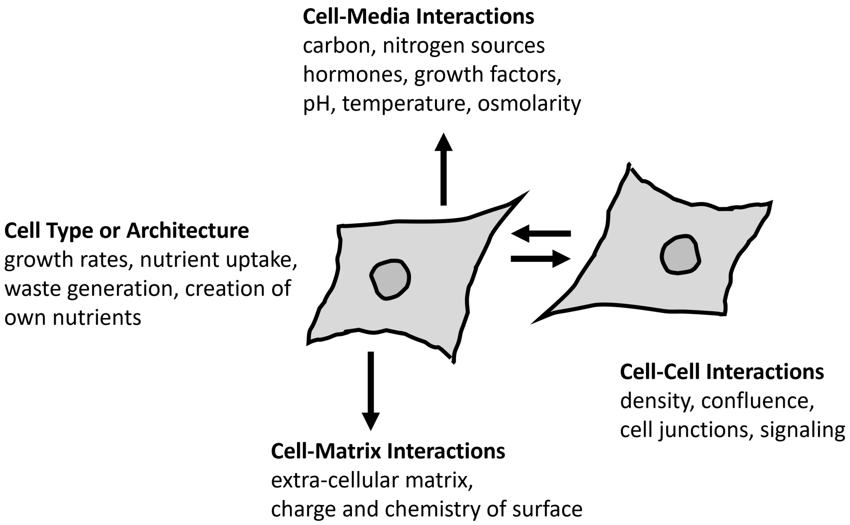

Cell-Media Interactions
carbon, nitrogen sources
hormones, growth factors,
pH, temperature, osmolarity
Cell Type or Architecture
growth rates, nutrient uptake,
waste generation, creation of
own nutrients
Cell-Cell Interactions
density, confluence,
cell junctions, signaling
Cell-Matrix Interactions
extra-cellular matrix,
charge and chemistry of surface

Figure 1.2. Cell Interactions with Environment | figure inspired by [**14**] illustrating different factors, and complexity, in cell proliferation and differentiation. The complexity and number of interactions is an important reason for using black box methods to support research.

cells is based on Dulbecco's Modified Eagle Medium (DMEM) / F12 supplemented with insulin, transferrin, FGF2, TGFβ1, ascorbic acid, and sodium selenite. Beefy-9 by [**80**] is similar to E8 but with additional albumin optimized for primary bovine satellite cells. The approach we will take in this dissertation is to use prior knowledge of biological processes to construct a list of potential media components, and use design-of-experiments (DOE) methods to optimize component concentrations based on cell proliferation. This will be particularly useful for cellular agriculture because by developing and using these statistical tools, as we will see in the next section, DOEs will help develop media quickly and efficiently.

1.1.3. Measuring Cell Growth. One of the most difficult aspects of this work is measuring the quality of media. Viable cells must be counted after a period of time over which the scientist believes the medium will have an effect, which changes depending on cell type, media components, cell density, ECM, pH, temperature, osmolarity, and reactor configuration. If cells grow by adhering to a substrate, then subculturing / passaging may play a role on the health of a cell population, so discounting this effect may have deleterious effects on media design quality [**22**,**23**]. Counting using traditional methods like a hemocytometer or more advanced automatic cell counters using trypan blue exclusion are labor-intensive and prone to error. Cell growth / viability assays are chemical indicators that correlate with viable cell number such as metabolism (AlamarBlue [37, 53], MTT) or DNA / nuclei count (LIVE/DEAD, Hoechst 33342) and can also be used to quantify the effect of media on cells. In chapter 5 we conducted many experiments with different assays and show the inter-assay correlations in Figure 1.3. Notice no assay (top right / blue plots) is perfectly correlated with any other assay because they are collected with different methodologies and fundamentally measure different physical phenomena. For example, AlamarBlue measures the activity of the metabolism in the population of cells, so optimizing a media based on this metric might end up simply increasing the metabolic activity of the cells rather than their overall number. As some of these measurements can be destructive / toxic to the cells (AlamarBlue and LIVE stain for example), continuous measurements to collect data on the change in growth (not just at a single point in time, for an example see [10]) can be tedious. Collecting high-quality growth curves over time may be accomplished using image segmentation and automatic counting techniques. Using fluorescent-stained cells and images, segmentation can be done using algorithms like those discussed [30]. Cells may even be classified based on their morphology dynamically if enough training data is collected to create a generalizable machine learning model. Successfully quantifying the ability of media to grow cells forms the backbone of the novelty of this dissertation.

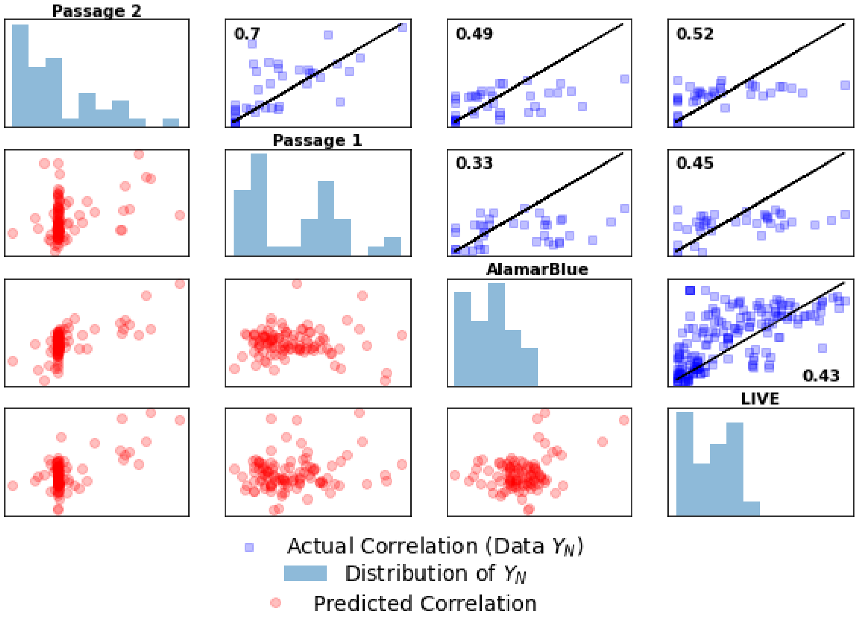


FIGURE 1.3. Correlations Between Cell Growth Metrics | (upper right blue squares) correlation between measurements Y_N of cell growth (with R^2 shown) from experiments conducted in chapter 5 using Passage 2, 1, AlamarBlue, and LIVE stain. x and y axis are the growth measurements of each of the methods located horizontally (to the left) and vertically (below) from the plot respectively. (center histogram) a density plot of N assay outputs Y_N data from Passage 2, 1, AlamarBlue, and LIVE stain. This shows the output distribution of each measurement type. (left bottom) Prediction of random data by model in chapter 5.

1.2. Review of Experimental Optimization Methods

The primary means by which this dissertation will improve cell culture media is through the application of various experimental optimization methods, often called design-of-experiments (DOE). The purpose of DOEs are to determine the best set of conditions x (media concentrations for example) to optimize some output y (cell growth rate or cell density at a specific time for example) by sampling a process for sets of conditions in an optimal manner. If an experiment is time / resource inefficient, then optimizing the conditions of a system may prove tedious. For example, doing experiments at the lower and upper bounds of a 30-dimensional medium like DMEM requires $2^{30} \approx 10^9$ experiments. This militates for methods that can optimize experimental conditions and explore the design space in as few experiments as possible.

1.2.1. Static DOE. The simplest method of exploring the design space and optimizing media is one-factor-at-a-time (OFAT), which was used to create E8 and Beefy-9 serum-free media by varying one component at a time. While good for initial screening, OFAT is inefficient for optimization because it does not take into account interactions between components that is common in biological systems. Full factorial designs are an improvement over OFAT. If there are p factors and l levels (high and low concentration $l = 2$ for example), then the number of experiments needed in a full factorial design is $N = l^p$. Fractional factorial designs reduce the needed experiments to $N = l^{p-k}$ for k generators, where the higher k results in a greater degree of aliasing, or confounding of effects. While the number of experiments is reduced in fractional designs, the contribution of any given component on the cells is unclear. Plackett-Burman designs are $N = p + 1$ economical, typically for systems where only first order effects are expected because they are aliased with second order effects. The 'effects' of a design may be found using the slopes β_i of a linear model (Equation 1.1). If the system is nonlinear, we may model it using a (second order) polynomial model (Equation 1.2), but this often requires a high-fidelity DOE, and the effect of any given component is not trivial to ascertain. Central composite designs are particularly useful for fitting polynomial models but are more experimentally expensive with $N = l^p + n_c + 2p$ experiment using n_c center-points in the middle of the design to evaluate experimental error. Box-Behnken designs, where factors are sampled along the center-points of all pairs of factors in $N = l^p + n_c$ total experiments, are good for training polynomial models where extreme values are not expected to perform well. Figure 1.4 shows an example of a linear versus polynomial model attempting to replicate a nonlinear underlying process with $N = 5$ random experiments. The structure of the data / system should inform the structure of the prediction and complexity of the DOE.

$$\hat{y}(x) = \beta_0 + \sum_{i=1}^{p} x_i \beta_i \tag{1.1}$$

$$\hat{y}(x) = \beta_0 + \sum_{i=1}^{p} x_i \beta_i + \sum_{i=1}^{p} x_i^2 \beta_i + \sum_{i=1}^{p} x_i^2 \beta_{ii} \tag{1.2}$$

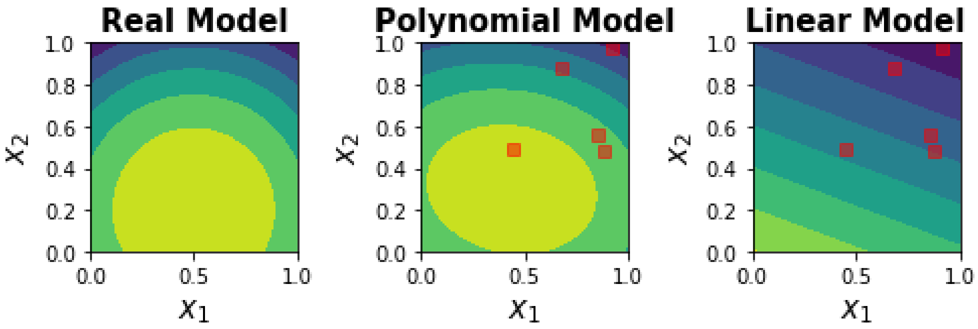


FIGURE 1.4. Examples of Simple Models | (left) plot of $y(x) = -((x_1 - 0.5)^2 + (x_2 - 0.2)^2)$, (center) and (right) are fitted polynomial and linear models trained using ordinary least squares method on $N = 5$ data points shown as red squares.

Other types of designs need to be solved using a computer or a more sophisticated objective function. The "alphabet" designs (A, D, G-optimal for example) are popular. The advantage of them is that (i) we are not limited in the number of experiments to perform at a time, (ii) constraints may be considered easily, and (iii) prior information on coefficients of the model may be considered. One such alphabet design, the D-optimal design, is computed using Equation 1.3, for a the Fisher information matrix (detoted as $\Sigma(x, x)$ for some set of inputs x). Here we attempt to maximally reduce the variance of the coefficients of a statistical model and thus have a model that explains the data well. If the coefficients of the model are multi-variant normal, this becomes $\Sigma = xx^T$, so $\alpha_D(x) = det[xx^T]$ plotted in Figure 1.4. Notice this method is independent of the predicted output. There is a Bayesian interpretation of this design as well. If we try to maximize the distance between the prior and posterior probability distribution of a model we end up with Equation 1.4, which is similar to Equation 1.3 but with an additional matrix R which acts as the prior variance [**49**]. A good rule of thumb is that Bayesian interpretations are similar to standard designs in the limit of large data, where the prior R is "overwhelmed" by the data $\Sigma(x, x)$. This will become important as we try to construct DOEs that conform to our expectations about how the underlying processes work in order to improve model accuracy and experimental efficiency. Other design criteria exists, for example the A-optimal design $\alpha_A(x) = Trace(\Sigma(x, x)^{-1})$ attempts to minimize the sum of diagonal elements of the Fischer information matrix, which may have more desirable statistical properties than D-optimal designs depending on the situation [**44**].

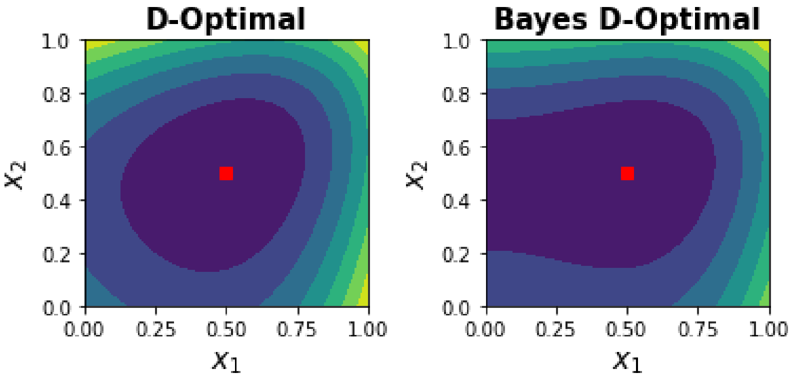


FIGURE 1.5. Example of D-Optimal Contour | using Equation 1.3 (left) and 1.4 (right), D-optimal designs were computed when $x = [0.5, 0.5]$ has already been selected for a polynomial model. For the (left) Bayesian plot $R = \sigma_R^2 * I$ for variance matrix on predictor coefficients $\sigma_R^2 = 10^{-3}$. In this above example, we also set $R_{3,3} = 1$ indicating a low prior variance on x_2, down-weighting the usefulness of sampling x_2 because we believe the variance on that paramater will be low.

$$\alpha_D(x) = det[\Sigma(x, x)] \tag{1.3}$$

$$\alpha_{D,bayes} \approx det[\Sigma(x, x) + R] \tag{1.4}$$

DOEs where samples are located throughout the design space to maximize their spread and diversity according to some distribution are called space-filling designs. The most popular method is the Latin hypercube (Figure 1.6), which are particularly useful for initializing training data for models [**43**] and for sensitivity analysis [**18**]. Maximin designs, where some minimum distance metric is maximized for a set of experiments, can also allow for diversity in samples, with the disadvantage being that in high dimensional systems the designs tend to be pushed to the upper and lower bounds. Thus, we may prefer a Latin hypercube design for culture media optimization because media design spaces may be >30 factors large. Uniform random samples, Sobol sequences, and maximum entropy filling designs [**16**], all with varying degrees of ease-of-implementation and space-filling properties, also may be used. It cannot be known a priori how many sampling points are needed to successfully model and optimize a design space because it is dependent on the number of components in the media system, degree of nonlinearity, and amount of noise expected in the response. Because of these limitations, DOE methods that sequentially sample the design space have gained traction, which will be talked about in the next section.

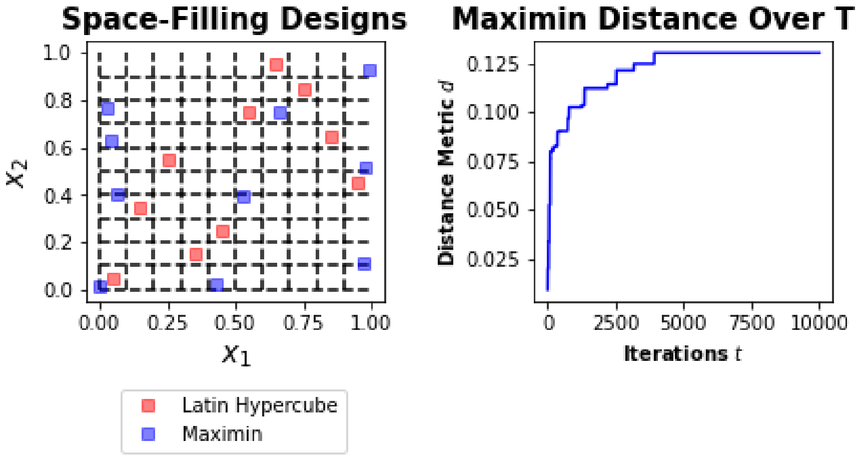


FIGURE 1.6. Space-Filling Designs | (left) Latin hypercube. Notice all rows x_2 and columns x_1 have a single samples without overlap. Maximin designs, determined by maximizing the minimum distance (criteria was $d = \sum_{j=1}^{p}(x_i - x_k)^2)$ between any two points in the group of designs, is also shown by sampling for $t = 10000$ iterations (right plot shows the distance criteria getting better over t with the final distribution in blue on the right).

1.2.2. Sequential DOE. A more data-efficient DOE is to split up individual designs into sequences and use old experiments to inform the new experiments in a campaign. One sequential approach is to use derivative-free optimizers (DFOs) where only function evaluations y are used to sample new designs x. DFOs are popular because they are easy to implement and understand, as they do not require gradients. They are also useful for global optimization problems because they usually have mechanisms to explore the design space to avoid getting stuck in local optima. The genetic algorithm (GA) is a common DFO where a selection and mutation operator is used to find more fit (better y) combinations of genes (combinations of x). In Figure 1.7, notice the GA was able to locate the optimal region of both problems regardless of the degree of multi-modality. [**9**] used a GA to optimize media for rifamycin B fermentation in bacteria where the HPLC titer at the end of 9 days was used to select high performing media combinations from nine metabolites for the next batch of experiments. They allowed for a 1% chance of mutation of each experiment and component to allow for global search. They also discovered that the response space was multi-modal and had interactions between components, which confirmed the need for global optimization of fermentation and bioprocessing problems. [**34**] discusses 17 cases in which GAs have improved media for different organisms for chemical fermentation often by $> 50\%$ yields for problems of > 10 media components. Particle swarm optimization [**28**] is a population-based method that optimizes systems sequentially based on varying x according to a velocity vector v. At the tth iteration of the algorithm a particle x will have the velocity update rule $v_{t+1} = wv_t + c_1 r_1 * (p_{best} - x) + c_2 r_2 (g_{best} - x)$ for random numbers r_1, r_2, coefficients w, c_1, c_2 (note the global and particle optimal points g_{best}, p_{best} respectively). c_1 and c_2 parameterize the exploration-exploitation trade-off, similar to the mutation rate in the GA. w represents the fraction of velocity saved for the next iteration $t + 1$. To implement this one merely computes $x_{t+1} = x_t + v_t$ for a large population of particles over time as the population gradually gravitates to the optimal designs. The Nelder-Mead simplex method, wherein a group of points is moved closer to better values via expansion and contraction steps, is also a popular DFO method. Nelder-Mead is a local optimizer and may be hybridized with other global DFO methods (see section 3.3.4 of [**67**]) to improve convergence. While DFOs don't require gradient calculations and can usually optimize complex multi-modal optimization problems (such as in media design), they require 100's, if not 1000's, of experiments so are limited to fast growing culture systems or computer experiments where experiments are somwhat costless.

The most powerful experimental optimization technique is arguably the model-based sequential DOE, in which a response-surface model (RSM) of the relationship between the input x and output y data is trained, and new samples are constructed based on the predictions of the trained model. The newly collected data is then fed back into the model and used to generate another sequence of samples. [**74**] discusses using combinations of screening DOEs and polynomial RSM to optimize conditions for the fermentation of metabolites such as chitinase, γ-glutamic acid, polysaccharides, chlortetracycline and tetracycline among 20 other metabolites from various organisms. This demonstrates the usefulness of RSMs for fermentation and culture optimization. The primary limitation of polynomial RSMs is their inability to accurately model many factors (usually >5) at a time or systems with significant nonlinearity. Due to their generalizability to modeling different response surfaces, neural networks have been used to optimize bioreactor cultures [**46**] and multi-objective

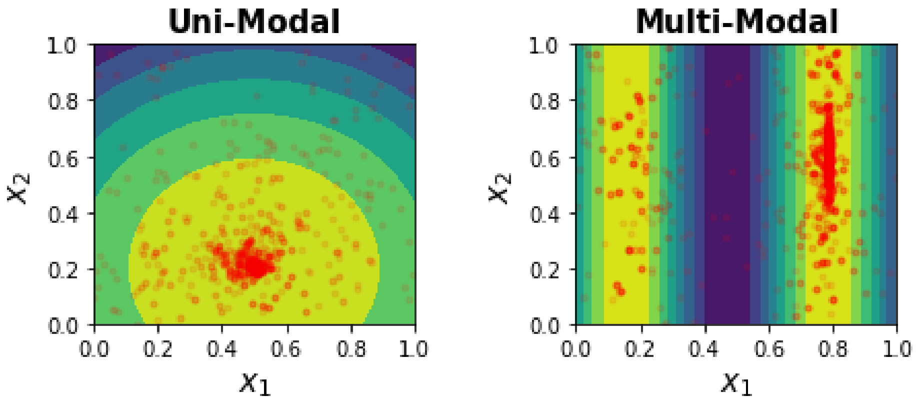


FIGURE 1.7. Single vs Multi-Modal Examples for GA | a GA was used to optimize $y(x) = -((x_1 - 0.5)^2 + (x_2 - 0.2)^2)$ (left) and $y(x) = sin(10x_1)$ (right) in 100 generations of 100 samples per generation. A mutation rate of 1% was used to explore the design space. Crossover between parent genes was done by averaging any two of the 100 parent samples to generate 50 child samples.

protein storage conditions [**68**]. Radial basis function have been used to optimize yeast [**93**] and C2C12 mammalian muscle cell [**22**] culture growth media. Decision trees and neighborhood analysis have been used to optimize media for antibiotics [**9**] and bacteria fermentation [**20**]. An example of an RSM can be seen in Figure 1.8 where a radial basis function maps the input / output relationship in a nonlinear system, then a GA finds new optimal experiments. Over time the predicted contour looks similar to the true function. While these RSMs tend to be more generalizable compared to polynomial and linear models, low-data experimental campaigns common in fermentation and cell culture often obscure the differences between modeling techniques. Additionally, many of these RSM approaches do not take into account prior information about the system to speed up optimization.

Gaussian process (GP) models can also be used for sequential DOE [**70**] in a class of methods known as Bayesian optimization [**71**] or efficient global optimization [**45**]. GPs are distributions over functions $f(x) \sim N(\mu, \Sigma)$, in a similar way that Gaussian distributions are distributions over parameters $x \sim N(\mu, \Sigma)$. The unique thing about GPs is that we may design a mean μ and covariance Σ based on domain knowledge of the fermentation or cell culture system. It is known that, given some input data X and output data Y, a conditional Gaussian distribution $f(x|X, Y) \sim N(\mu, \Sigma)$ has the mean $\mu(x|X, Y)$ and variance matrix $\Sigma(x|X, Y)$ in Equation 1.5 and 1.6 respectively. Therefore, GPs can model uncertainty in the model and experiment, have mathematically well-understood

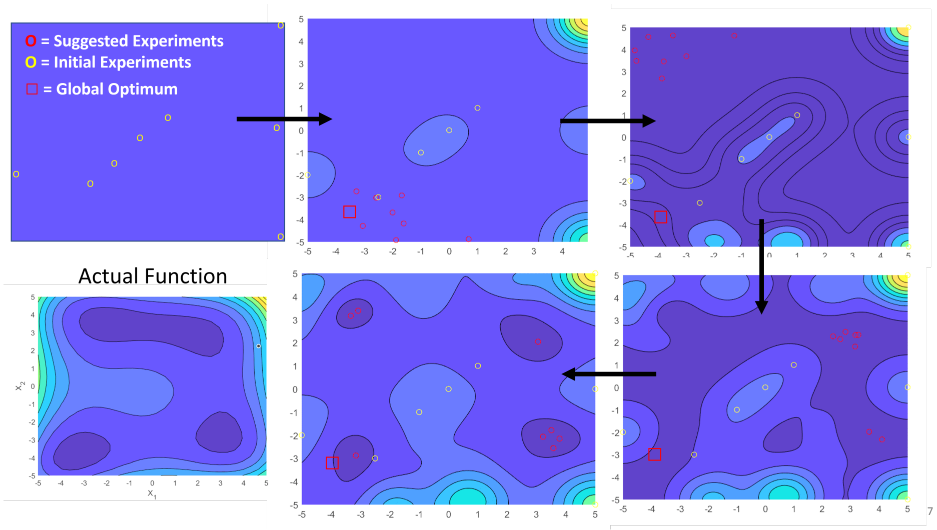


FIGURE 1.8. Example of Sequential DOE | 2-D himmelblau's function $y(x) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$ is optimized using a radial basis function regression model initialized with $N = 6$ random data points. New experiments are suggested using a GA. The red square roughly corresponds to the optimal set of parameters.

properties, and are highly customizable due to our ability to tailor the mean and covariance to the problem of interest. For example, a common covariance that describes the relationship between two given points x and x' is the squared-exponential kernel parameterized by a length-scales λ_k and output-scale σ_f with the functional form $\Sigma(x, x') = \sigma_f^2 * exp(-1/2 \sum_{k=1}^{p} \frac{(x_k - x'_k)^2}{\lambda_k^2})$. This encodes the prior knowledge that culture systems have smooth underlying response surfaces. To further encode smoothness, we may place a normal prior on the length-scales $\lambda_k \sim N(0, \sigma_\lambda^2)$ to induce broader correlations among data points. Another prior might be the knowledge that the underlying function should be linear, so alter Equation 1.5 to have $\mu_0(x) = c^T x$.

$$\mu(x|X, Y) = \mu_0(x) + \Sigma(x, X)(\Sigma(X, X))^{-1}(Y - \mu_0(X)) \tag{1.5}$$

$$\Sigma(x|X, Y) = \Sigma(x, x) + \Sigma_c(x, X)(\Sigma(X, X))^{-1}\Sigma(x, X)^T \tag{1.6}$$

We can compute acquisition functions α to suggest new experiments using the GP predictions. Expected improvement α_{EI} [**41**] pushes new experiments over the best previous value y^* and strikes

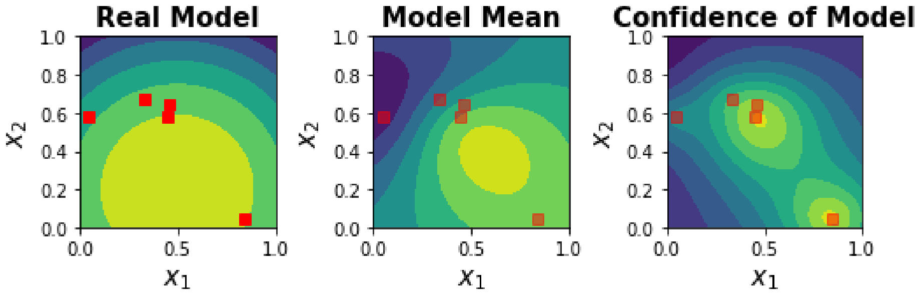


FIGURE 1.9. Gaussian Process | (left) real model with $N = 5$ collected data vs (center) prediction $\mu(x)$ using the data with (right) variance of model. Squared exponential kernel trained using maximum likelihood with L-BFGS-B (a type of optimization algorithm that approximates the Hessian of the function using derivatives) of hyperparameters.

a balance between exploration (high $\sigma(x)$) and exploitation (better $\mu(x)$). It is often the default acquisition function for Bayesian optimization due to this balance. The E and $(f)^+$ operators correspond to the expectation (average) and $max\{f, 0\}$ for some value f. The upper confidence bound α_{UCB} [**79**] allows the user to parameterize the trade-off between exploration and exploitation using the hyperparameter β. This parameter can be dynamically set depending on the results of the experimental campaign or adjusted based on prior knowledge of the scientist. Information-based policies such as max-value entropy search α_{MES} [**82**] have the advantage of quantifying the value of a function evaluation at the unknown minimum $f(x^*)$. In this manner, if it is impossible to define an improvement in the output space, such as collecting data with different units, we can still design experiments. The H and E_{y^*} operators correspond to the entropy of the argument and the expectation (average) value of the argument given you believe y^* is the maximum value in the design space (many more details on the analtical form of this can be found in [**82**]).

$$\alpha_{EI}(x) = E[(\mu(x) - y^*)^+] \tag{1.7}$$

$$\alpha_{UCB}(x) = \mu(x) + \beta\sigma(x) \tag{1.8}$$

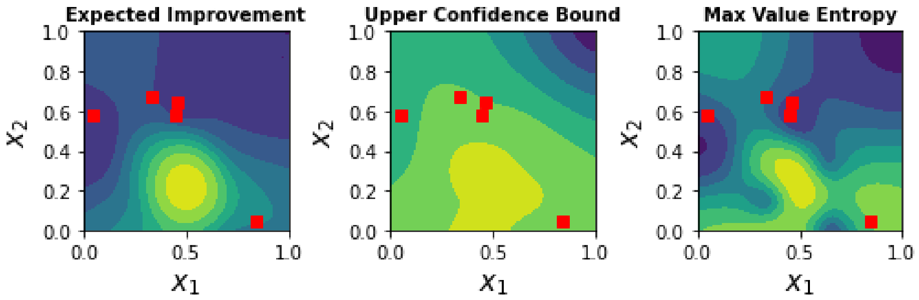


FIGURE 1.10. Bayesian Acquisition Functions | (left) expected improvement, (center) upper confidence bound with $\beta = 0.5$, (right) max-value entropy search.

$$\alpha_{MES}(x) = H(x) - E_{y^*}[H(x|x, x^*)] \tag{1.9}$$

GPs and Bayesian methods are also flexible. For example, [**76**] shows us that priors can be placed on the expected optimal design x^* in a way that allows for faster learning of the true optimal design. This is similar to placing prior beliefs on hyper-parameters θ^* in dynamic systems [**19**] or physics-based experimental / simulation data [**35**] to constraint uncertainty. Measurement noise can be incorporated into a GP by adding a term to the covariance matrix $\Sigma(X, X) + \sigma_\epsilon^2 * I$ where the noise parameter σ_ϵ may be estimated or assumed. This is similar to the Bayesian D-optimal design discussed in the previous section. [**2**] describe a heteroskedastic version of this $\Sigma(X, X) + v(X) * I$ where $v(X)$ is a variance model or set of data, which allows more experimentally uncertain regions of the design space to be modelled as such. [**83**] describe how, if outliers are to be expected, a GP can be modified to be a student-t process. Due to the noisiness of fermentation data it may be useful to consider noise in our process models. Known or unknown constraints can be incorporated into GPs [**54**] as well. For example, a known constraint might be that growth must exceed some minimum value. An unknown constraint might be the existence of excessive foaming in bioreactors, which may be learned from data, but is generally not known ahead of time. Multiple objectives, some of which may compete against one another, can be modeled and optimized using GPs [**11**] and correlations between tasks (potentially with different length-scales or units) may be considered [**81**]. Correlated measurements of the same (or similar) task may be solved using a Bayesian interpretation as well [**61**, **73**]. By correlating measurements, fewer total experiments are often needed. Multi-objective versions of acquisition functions α such as max-value entropy search [**11**] and hypervolume improvement [**26**] exist to turn these GP predictions into a score for a variety of objectives. Fermentation and cell culture systems are often subject to growth vs cost trade-offs so multi-objective Bayesian methods are useful here. Because most bioprocessing experiments can be done using multiple bioreactors or cell culture plates, designing multiple optimal experiments at a time is often necessary. [**89**] shows how, using monte-carlo samples of the GP model, arbitrary numbers of experiments can be designed simultaneously. Knowledge that systems may exhibit separate but interacting local and global responses may militate for additive GPs [**7**]. Experimenters with access to separate computer simulations or algebraic process models may pose their GPs as composites of deterministic or other modeled functions [**5**] and speed up optimization. Bayesian models may even fuse historical data-sets together to estimate optimal model parameters with constrained uncertainty [**35**], and could perhaps be used for optimization as well (transfer learning). More closely related to cell culture media optimization, GPs have been used in a Bayesian optimization scheme to optimize C2C12 growth media for proliferation maximization and cost minimization in chapter 5 of this dissertation.

1.3. Review of Thesis

This dissertation is divided into roughly two equal parts. The first part (chapter 2 and 3) are comprised of the development of a radial basis function genetic algorithm sequential DOE scheme [**21**, **22**]. It drew heavily on the work of [**66**], where a sequential DOE technique was developed on the principle of local random search in areas of high performing media. This algorithm was also dynamic by converging on high performing results and selectively searching the design space when good results were not forthcoming. Additionally, previous work in our lab [**93**, **95**] provided the framework for a sequential DOE based on a truncated GA. This modified GA incorporates uncertainty in the optimal samples found by halting algorithm convergence proportional to the amount of clustering around an optima the GA finds. By hybridizing these two methods, a DOE algorithm called NNGA-DYCORS was developed that solved various computational optimization problems better than either method alone. It was used to optimize a 30-dimensional media for serum-containing C2C12 cell culture with the metric of growth being AlamarBlue reduction after

48 hrs of growth in 96 well plates (in chapter 3 it was renamed HND). Cells were seeded at the same time, concentration, and from the same frozen innoculum so that all experiments were roughly the same. While it was successful at finding media that maximized this metric (as well as minimized a cost metric), the optimal medium did not grow as many cells over additional passages.

To fix this underlying problem, multiple passages needed to be incorporated into the DOE process. This is a very time-consuming process as each passage takes multiple days, many more physical manipulations than simple chemical assays which introduces opportunities for contamination, and difficulty for manual experimentation. To solve this, chemical assays were supplemented with small amounts of manual multi-passage cell counts in a multi-information source Bayesian GP model [**31**] which was used to successfully optimize a 14-dimensional serum-containing media for C2C12 cells [**23**] (chapter 4). Due to the presence of multi-passage data, the final optimal medium grew cells robustly over four passages, provided nearly twice the number of cells at the end of each passage relative to the DMEM + 10% FBS control and traditional DOE method, and did so at nearly the same cost in terms of media components. In the final chapter (chapter 5) the multi-information source GP model was extended to optimize a 26-dimensional serum-free media based on the Essential 8 media [**52**] using a multi-objective metric that improves cell growth while minimizing medium cost. Using this Bayesian metric, a broad set of media samples along the trade-off curve of media quality and cost were found, showing that a designer can be given options in media optimization. In particular, one medium resulted in higher growth over five passages while the control and Essential 8 lagged.

We identify two important future considerations for this work. First, the data collection process, which is the major innovation of this dissertation, needs to be made more robust by actually capturing the long-term growth dynamics of the cells. Fluorescent and brightfield imaging, used to quantify the temporal and spatial changes of the cells, may improve over whole-well AlamarBlue and LIVE/DEAD stains by couting individual cells and collecting more fine-grained growth curves. Additionally, bio-markers of proliferation and cell health such as Pax7, MyoD, and Myogenin may be measured to improve the robustness of predictions and correlations across assays. None of these metrics will aid in optimization if a sufficient model of the relationship between cell growth, media cost, and overall process cost is not considered. Therefore, a techno-economic model of the process is needed to tie together the large-scale production process to bench-top measurements. Secondly, further "white-box" studies that focus on the metabolomics [**58**] of the cell lines would be very useful in defining the upper / lower bounds and important factors of these DOE studies. Developing robust cell lines (that are relevant for cellular agriculture such as bovine, porcine, or avian) adapted to serum-free conditions would open up the design space for use in DOE studies because very poorly-growing cells are difficult to optimize in DOE studies. In general, white-box or traditional studies act to constrain the complexity of future DOE studies, so must be conducted in collaboration with DOE.

CHAPTER 2

A Generalizable Hybrid Surrogate Framework for Expensive Design Optimization Problems

Experimental optimization of physical and biological processes is a difficult task. To address this, sequential surrogate models combined with search algorithms have been employed to solve nonlinear high-dimensional design problems with expensive objective function evaluations. In this article (originally published as [**21**]), a hybrid surrogate framework was built to learn the optimal parameters of a diverse set of simulated design problems meant to represent real-world physical and biological processes in both dimensionality and nonlinearity. The framework uses a hybrid radial basis function/genetic algorithm with dynamic coordinate search response, utilizing the strengths of both algorithms. The new hybrid method performs at least as well as its constituent algorithms in 19 of 20 high-dimensional test functions, making it a very practical surrogate framework for a wide variety of optimization design problems. Experiments also show that the hybrid framework can be improved even more when optimizing processes with simulated noise.

2.1. Introduction

The design and optimization of modern engineering systems often requires the use of high-fidelity simulations and/or field experiments. These black box systems often have nonlinear responses, high dimensionality, and have many local optima. This makes these systems costly and time consuming to model, understand, and optimize when simulations take hours or experiments performed in the lab require extensive time and resources.

The first attempt to improve over experimental optimization methods, such as 'one-factor-at-a-time' and random experiments was through the field of Design of Experiments (DOE). Techniques in DOE have been adapted to many computational [**32**] and experimental fields [**4**, **74**, **90**] in order to reduce the number of samples needed for optimization. These methods often involve performing experiments or simulations at the vertices of the design space hypercube. Full-Factorial Designs are arguably the simplest to implement, where data is collected at all potential combinations of parameters p for all levels l requiring l^p samples in total. Even when $l = 2$ (for 'high' and 'low' levels of the design space) the number of experiments or simulations quickly becomes infeasible so Fractional-Factorial Designs using l^{p-k} experiments for k 'generators' are often used to reduce the burden. While such designs are more efficient, they have lower resolution than full designs and confound potentially important interaction effects. Therefore, DOE techniques are often combined with Response Surface Methodology (RSM) to iteratively move the sampling location, improve model fidelity as more data is collected [**69**], and focus experiments in regions of interest. Stochastic optimization methods such as Genetic Algorithms (GA), Particle Swarm Optimization, and Differential Evolution have also been used to explore design spaces and perform optimization on both simulated [**55**] and experimental data [**9**,**34**,**90**], often requiring fewer experiments than traditional DOE-RSM techniques.

The quickly developing field of surrogate optimization (also called meta-modeling or active learning) attempts to leverage more robust modeling techniques (such as radial basis functions (RBF) [**66**] or Kriging / Gaussian Process models [**88**]) to optimize nonlinear systems. They often employ a stochastic [**95**], uncertainty-based [**45**], or Bayesian [**49**] search algorithm to intelligently select new sample points to query for experimentation or simulation. Due to the variety of modeling techniques and search algorithms available, hybrid algorithms, which attempt to leverage each methods strengths, have proliferated [**33**, **96**]. These hybrid approaches usually involve taking ensembles of surrogate models and asking each surrogate for its best set of predicted query points. New queries are then conducted at these points, often weighted in favor of regions/surrogates with low sample variance or optimal response values. The drawback of many of these algorithms is that they are not always generalizable to design problems of diverse dimensionality and nonlinearity.

A surrogate optimization algorithm is presented here, which uses an evolving RBF model and hybrid search algorithm. This search algorithm selects half of its query points using a Euclidean distance metric truncated to provide diversity in suggested query points. This is based on a neural network genetic algorithm (NNGA) developed for bioprocess optimization [**92**], which has been shown to be more efficient than traditional DOE-RSM methods (note that typically RBF models are not considered neural networks, but we will continue to use this terminology in this chapter for consistency with the original publication). The other half of the query points are selected using a dynamic coordinate search for response surface methods (DYCORS) algorithm based on work developed for computationally expensive simulation [**66**]. DYCORS has been shown to perform better than a variety of popular surrogate optimization techniques. The performance of the NNGA-DYCORS hybrid algorithm is tested against NNGA and DYCORS separately. Further evaluation is performed to probe potentially useful extensions of the hybrid algorithm (1) to address simulated experimental noise, (2) to improve algorithm convergence over time, and (3) to address cases in which certain groups of parameters have a greater influence on the response values than others.

2.2. Methods

2.2.1. RBF Surrogate Model. The surrogate model used is the RBF interpolation model. A cubic RBF $\phi(x) = r^3$ with a linear tail $p(x)$ is used to map input data $x \in R^{n\times p}$ to output data $y \in R^{n\times 1}$ given number of data n with input dimensionality p. The form of the RBF interpolation $s_n(x)$ is shown (Equation 2.1).

$$s_n(x) = \sum_{i=1}^{n} \lambda_i \phi(|(|x - x_i|)|) + p(x) \tag{2.1}$$

Substituting $\phi(x)$ and $p(x)$ gives Equation 2.2,

$$s_n(x) = \sum_{i=1}^{n} \lambda_i (|(|x - x_i|)|)^3 + c_0 + \sum_{j=1}^{p} c_j x_j \tag{2.2}$$

where $||x - x_i||$ is the Euclidean norm of a given point x and all RBF nodes x_i (also called centers). The number of nodes in an RBF model is often tuned to give low bias (many nodes) or low variance (few nodes). For training, the coefficients of the RBF $\lambda \in R^{n\times 1}$ and the linear tail $c \in R^{(d+1)\times 1}$ (for a d parameter design problem) are determined by solving the following system of linear equations shown in Equation 2.3.

$$\begin{pmatrix} \Phi & P \\ P^T & 0_{(p+1)\times(p+1)} \end{pmatrix} \begin{pmatrix} \lambda \\ c \end{pmatrix} = \begin{pmatrix} Y \\ 0_{d+1} \end{pmatrix} \tag{2.3}$$

The matrix $\Phi \in R^{n\times n}$ consists of components $\Phi_{ij} = \phi(|(|x_i - x_j|)|)$. The matrix $P \in R^{n\times(d+1)}$ is comprised of the rows of $[1, x_i^T]$. The output of data i is y_i contained in vector $Y \in R^{n\times 1}$. The coefficient vector can be inverted using singular value decomposition or any linear solver, solving the linear transformation for input data x and output data y. Modifying the equation to exclude the linear terms requires solving $\Phi\lambda = Y$.

2.2.2. NNGA. The NNGA algorithm is based on a RBF-assisted GA. The NNGA uses an RBF model to suggest points that are close to but not directly on top of optima, using a truncated genetic algorithm (TGA). One advantage that GAs have over gradient-based methods is that their randomness allows them to efficiently explore both global and local regions of optimality. This makes them very attractive for an optimization framework attempting to look for global optima while facing uncertainty associated with a sparsely explored parameter space, and thus untrustworthy RBF models. This framework is shown in Figure 2.1(a) and the TGA is illustrated in Figure 2.2.

First, a database of inputs X and outputs Y of N_o total queries is collected (often through a DOE, random queries or Latin Hypercube design). An RBF model is constructed using the training regime discussed in Section 2.2.1. Next, a TGA is run using a randomly initiated population of potential query points with the goal of minimizing the RBF predicted output. In each iteration of the TGA, queries expected to perform the best survive a culling process and have their information propagated into the next iteration by a pairing, crossover and random mutation step. After each iteration, the best predicted query is recorded. When the average normalized Euclidean distance between the TGA's current predicted best query and its next $N-1$ predicted best queries, $d_{av,norm}$, is less than or equal to the critical distance parameter $CD = 0.2$, the TGA is considered to be converged and submits this list of N best points for potential querying (or if the maximum number of iterations has been are clustered down to a final averaged query list of size N using k-means clustering. The final list is reached). This TGA is run a total of $k_{max} = 4$ times, and its query selections from all rounds of TGA queried to give the next set of data for simulation or experiments.

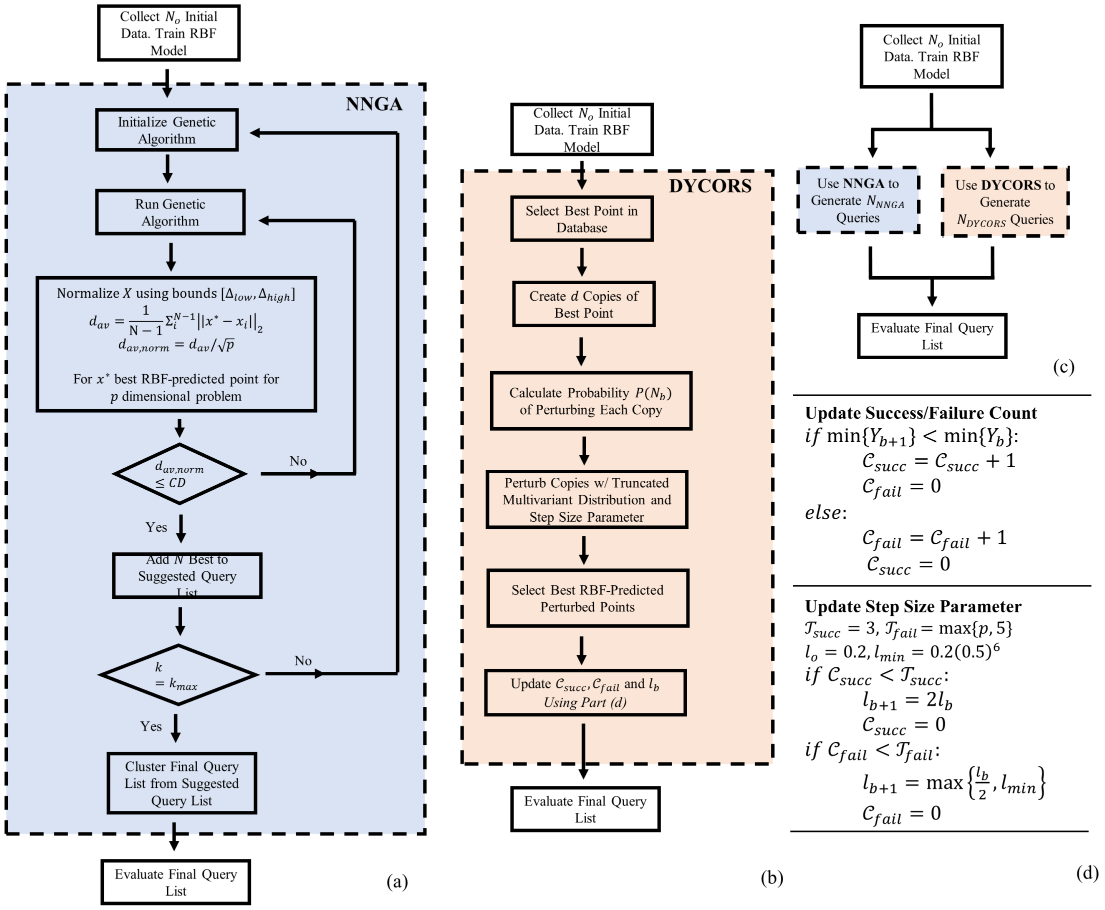

Collect N_o Initial Data. Train RBF Model
NNGA
Initialize Genetic Algorithm
Run Genetic Algorithm
Normalize X using bounds $[\Delta_{low}, \Delta_{high}]$
$d_{av} = \frac{1}{N-1}\Sigma_i^{N-1}||x^* - x_i||_2$
$d_{av,norm} = d_{av}/\sqrt{p}$
For x^* best RBF-predicted point for p dimensional problem
$d_{av,norm} \leq CD$
No
Yes
Add N Best to Suggested Query List
$k = k_{max}$
No
Yes
Cluster Final Query List from Suggested Query List
Evaluate Final Query List
(a)
Collect N_o Initial Data. Train RBF Model
DYCORS
Select Best Point in Database
Create d Copies of Best Point
Calculate Probability $P(N_b)$ of Perturbing Each Copy
Perturb Copies w/ Truncated Multivariant Distribution and Step Size Parameter
Select Best RBF-Predicted Perturbed Points
Update $\mathcal{C}_{succ}$, $\mathcal{C}_{fail}$ and l_b Using Part (d)
Evaluate Final Query List
(b)
Collect N_o Initial Data. Train RBF Model
Use NNGA to Generate N_{NNGA} Queries
Use DYCORS to Generate N_{DYCORS} Queries
Evaluate Final Query List
(c)
Update Success/Failure Count
if $\min\{Y_{b+1}\} < \min\{Y_b\}$:
$\mathcal{C}_{succ} = \mathcal{C}_{succ} + 1$
$\mathcal{C}_{fail} = 0$
$else$:
$\mathcal{C}_{fail} = \mathcal{C}_{fail} + 1$
$\mathcal{C}_{succ} = 0$
Update Step Size Parameter
$\mathcal{T}_{succ} = 3$, $\mathcal{T}_{fail} = \max\{p, 5\}$
$l_o = 0.2$, $l_{min} = 0.2(0.5)^6$
if $\mathcal{C}_{succ} < \mathcal{T}_{succ}$:
$l_{b+1} = 2l_b$
$\mathcal{C}_{succ} = 0$
if $\mathcal{C}_{fail} < \mathcal{T}_{fail}$:
$l_{b+1} = \max\left\{\frac{l_b}{2}, l_{min}\right\}$
$\mathcal{C}_{fail} = 0$
(d)

Figure 2.1. Flow Charts of Optimization Algorithms. (a) NNGA (b) DYCORS (c) hybrid NNGA-DYCORS; (d) the step size adjustment and success / failure count method used in (b) is displayed.

The number of queries per batch N, total number of batches b_{max} and critical distance parameter CD, which controls the degree of truncation, are set by the user.

2.2.3. DYCORS. The DYCORS generates a large list of potential query points based on Gaussian perturbations of the current best point in the training data set. It is dynamic because, as the training data set increases in size, the number of parameters perturbed decreases. In this manner, DYCORS narrows its search of the parameter space overtime. The DYCORS process is shown in Figure 2.1(b) and the parameters used in the algorithm are presented in the discussion below.

First a database with inputs X and outputs Y is collected, and an RBF model constructed. Next, the best point in the current data set x^* is selected and perturbed by a truncated multivariate normal distribution [**13**], bounded by the parameter's bounds $[\Delta_{low}, \Delta_{high}]$ and using a standard deviation $l_b \times \Delta_j$ for each parameter j and current batch step size l_b. This is repeated on $d = min\{100p, 5000\}$ copies of x^* and is to taking the best solution and looking in the general $l_b \times \Delta_j$ region around them for the next points to query. The perturbation appears in the form of Equation 2.4 for a parameter j to be perturbed for a given i copy of x^*:

$$x_{ij} = x_{ij} + N(0, l_b \times \Delta_j) \tag{2.4}$$

DYCORS is modulated by the step size selection algorithm shown in Figure 2.1(d), which counts consecutive successful C_{succ} and failed C_{fail} batches of queries and either doubles (if $C_{succ} \geq T_{succ} = 3$) or halves (if $C_{fail} \geq T_{fail} = max\{p, 5\}$ the step size l_b for the next batch based on thresholds T.This heuristic is employed based on the logic that, if numerous consecutive failures to improve are seen, a minimum parameter set has likely been reached. Thus, the search space is narrowed. In addition to altering l_b over time (with an initial $l_o = 0.2$ and minimum $l_min = 0.2(0.5)^6$), DYCORS also reduces the probability that a point x_{ij} will be perturbed by Equation 2.5 which is dependent on the current number of queries in the database N_b. This has the effect of narrowing down the amount of perturbations per batch as time goes on. After the perturbations are made and the step size l_b is updated, the N best perturbations of x^* are selected to be queried. The process is shown in detail in Figure 2.1(b). The primary way that this implementation of DYCORS differs from the original work is that the N best perturbations of x^* are selected for querying, rather than the single

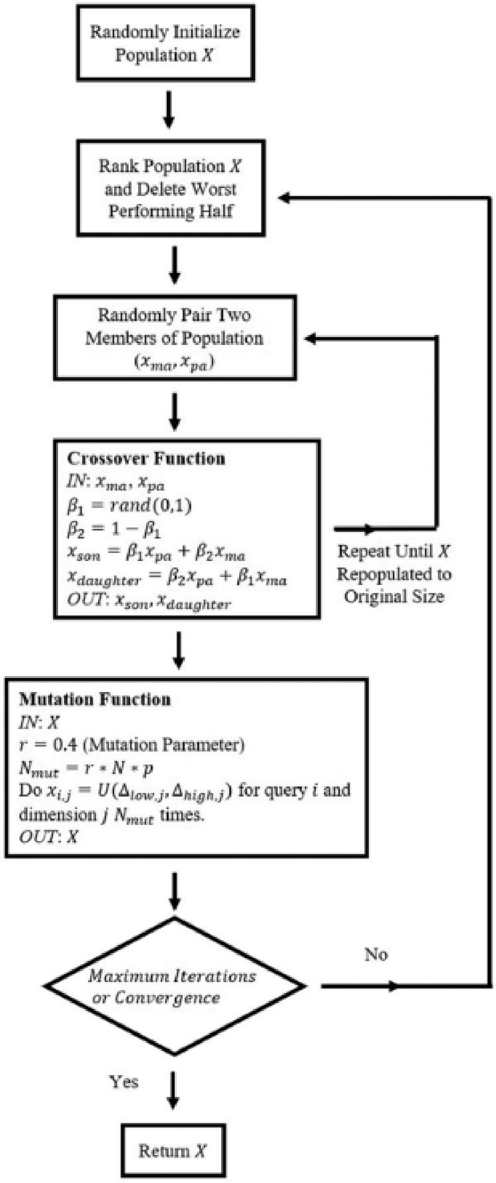

Randomly Initialize Population X
Rank Population X and Delete Worst Performing Half
Randomly Pair Two Members of Population (x_{ma}, x_{pa})
Crossover Function
IN: x_{ma}, x_{pa}
$\beta_1 = rand(0,1)$
$\beta_2 = 1 - \beta_1$
$x_{son} = \beta_1 x_{pa} + \beta_2 x_{ma}$
$x_{daughter} = \beta_2 x_{pa} + \beta_1 x_{ma}$
OUT: $x_{son}, x_{daughter}$
Repeat Until X Repopulated to Original Size
Mutation Function
IN: X
r = 0.4 (Mutation Parameter)
$N_{mut} = r * N * p$
Do $x_{i,j} = U(\Delta_{low,j}, \Delta_{high,j})$ for query i and dimension j N_{mut} times.
OUT: X
Maximum Iterations or Convergence
No
Yes
Return X

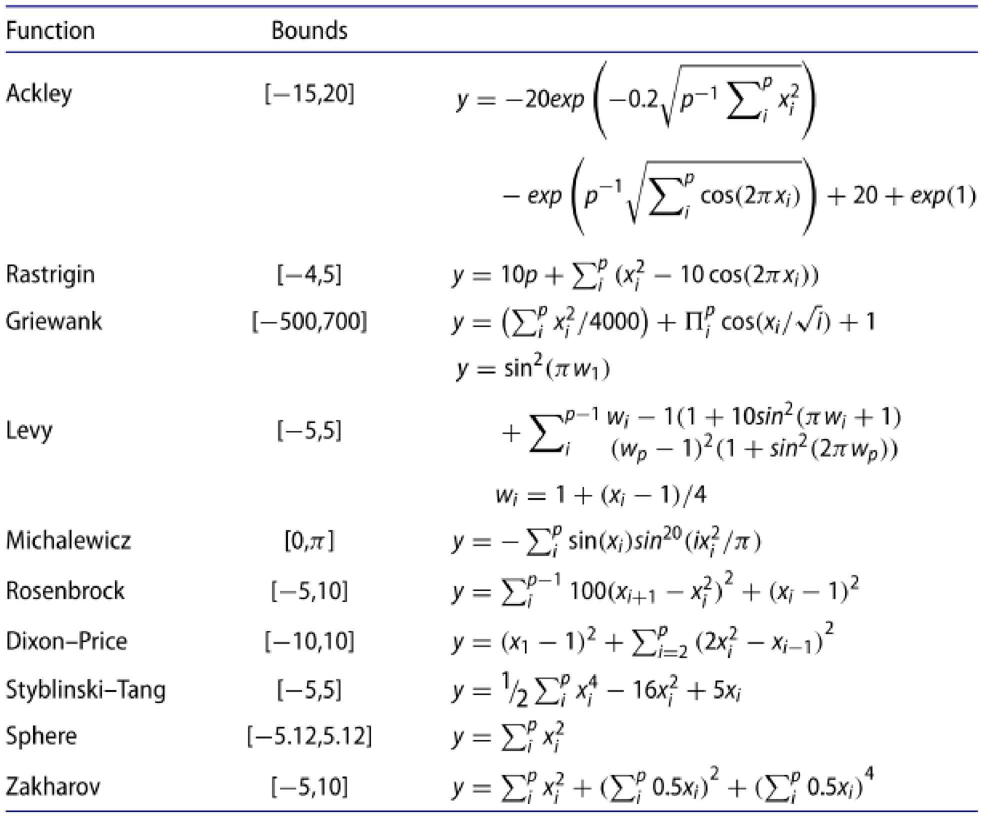
FIGURE 2.2. Truncated Genetic Algorithm. Used as stochastic optimizer for NNGA based on ranking, pairing, crossover, and mutation steps to generate optimal parameter combinations. Maximum iterations set at 100, CD and r (mutation rate) set by user.

best perturbation. For a given N_{max} (total amount of queries) this makes this implementation of DYCORS less efficient, but allows for multiple queries to be generated at once, and thus parallel computations/experiments to be carried out.

$$P(x_{ij} = x_{ij} + N(0, l_b \times \Delta_j)|N_b) = min(20/p, 1) \times (1 - \frac{ln(N_b - N_o + 1)}{ln(N_{max} - N_o)}) \tag{2.5}$$

2.2.4. NNGA-DYCORS. To combine the NNGA and DYCORS surrogate optimization algorithms, the algorithms are run in parallel with a shared data set $\{X, Y\}$ and the RBF model. This is shown by a flowchart in Figure 2.1(c). By having access to the same data, the two algorithms can make different conclusions about new optimal queries. To form the new data set, the suggested queries are combined and a new query conducted. The user can determine how many queries each algorithm suggests each batch. In this article, the NNGA and DYCORS arms of the hybrid find the same number of optimal queries.

2.2.5. Test Functions and Algorithm Assessment. To test the ability of these algorithms to learn arbitrary complex relationships between X and Y and find the minima of the resulting surfaces, optimization is performed on several test functions as shown in Table 2.1.



Function	Bounds	
Ackley	[−15,20]	$y = -20exp\left(-0.2\sqrt{p^{-1}\sum_i^p x_i^2}\right) - exp\left(p^{-1}\sqrt{\sum_i^p \cos(2\pi x_i)}\right) + 20 + exp(1)$
Rastrigin	[−4,5]	$y = 10p + \sum_i^p (x_i^2 - 10\cos(2\pi x_i))$
Griewank	[−500,700]	$y = \left(\sum_i^p x_i^2/4000\right) + \Pi_i^p \cos(x_i/\sqrt{i}) + 1$
Levy	[−5,5]	$y = \sin^2(\pi w_1) + \sum_i^{p-1} w_i - 1(1 + 10sin^2(\pi w_i + 1) (w_p - 1)^2(1 + sin^2(2\pi w_p))$ $w_i = 1 + (x_i - 1)/4$
Michalewicz	[0,π]	$y = -\sum_i^p \sin(x_i)sin^{20}(ix_i^2/\pi)$
Rosenbrock	[−5,10]	$y = \sum_i^{p-1} 100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2$
Dixon–Price	[−10,10]	$y = (x_1 - 1)^2 + \sum_{i=2}^p (2x_i^2 - x_{i-1})^2$
Styblinski–Tang	[−5,5]	$y = {}^1\!/_2 \sum_i^p x_i^4 - 16x_i^2 + 5x_i$
Sphere	[−5.12,5.12]	$y = \sum_i^p x_i^2$
Zakharov	[−5,10]	$y = \sum_i^p x_i^2 + (\sum_i^p 0.5x_i)^2 + (\sum_i^p 0.5x_i)^4$

Table 2.1. Test Functions for the NNGA-DYCORS.

To test the ability of these algorithms to learn arbitrary complex relationships between X and Y and find the minima of the resulting surfaces, optimization is performed on several test functions, as shown in Table 2.1. Simulations were performed on 10-D and 50-D dimensional variants of each test function to simulate low and high-dimensional optimization problems. For each evaluation, all algorithms were run $N_{NNGA} = 5$ and $N_{DYCORS} = 5$ queries per batch from NNGA and DYCORS respectively, in the case 15 times with a randomly selected initial database of size $N_o = 50$ and $N = 10$ queries per batch (with of the hybrid NNGA-DYCORS algorithm). The total number of batches was $b_{max} = 15$, making a total of $N_{max} = 200$ simulated experimental data points as the size of the final data set. To evaluate the optimization algorithms, learning curves were plotted to demonstrate the average optimal (minimum) output of each batch of queries, including error bars that indicate the standard deviation in the minimum output of each batch for the 15 runs. The mean, median, minimum, and standard deviations of the final batch of queries are shown in supplementary Tables A.1 and A.2.

2.2.6. Software and Hardware. Hardware used: Dell Precision 5820 Tower, Intel Xeon W-2145 DDR4-2666 Processor (3.7 GHz), 32 GB Memory. Software used: MATLAB R2019a with Bioinformatics Package.

2.3. Results

2.3.1. The Hybrid Framework versus Constituent Algorithms. The NNGA-DYCORS algorithm was tested against its constituent algorithms, NNGA, and DYCORS individually. Examining the performance of the constituent algorithms (Figure 2.3), the NNGA algorithm consistently works well in high dimensions (50-D), while the DYCORS algorithm performs better in low dimensions (10-D). This was the case both over time (Figure 2.3) and at the final optimal query points (Tables A.1 and A.2). Given these differences in performance, it stands to reason that a hybrid approach would provide a sensible route to a more robust algorithm that could be used on a wider variety of dimensions. As seen in Figure 2.3, the hybrid NNGA-DYCORS often outperforms or performs similarly to the next best constituent algorithm in each experiment. This is reinforced by the data in Tables A.1 and A.2, where the final optimum of the hybrid NNGA-DYCORS is less than or equal to the final optimum of the next best constituent algorithm in 19 of 20 experiments (all but the Michalewicz 50-D). An optimum may be considered better if its upper bound (mean plus standard deviation) is less than the mean of another algorithm's optimum. While this is a rough approximation of the comparative performance of the algorithm, it strongly indicates that the NNGA-DYCORS is robust on a wide variety of problem sets and dimensions. In intermediate cases (those between 10-D and 50-D), the NNGA-DYCORS continued to outperform or perform as well as its most competitive constituent algorithm(data not shown), showing its usefulness in design optimization problems where it is not obvious a priori what dimensionality counts as 'high' and 'low'.

2.3.2. Algorithm Performance in the Presence of Simulated Experimental Noise. To test the effect of random noise on the ability of the surrogate optimization algorithms to find optimal parameters, a random noise e (percent of the deterministic output) was added to the output of the simulation. It is common practice, especially in noisy, low-data, and data-sparse models, to improve the out-of-sample generalizability by model selection procedures such as cross-validation to avoid overfitting. To address the issues with stochasticity in these experiments this, a hyperparameter optimization loop for the number of nodes n_{nodes} in the RBF model was added to the NNGA-DYCORS algorithm, where cross-validation over the database was used to select the optimal n_{nodes}. In this case we deliberately trade higher bias for lower variance to reduce overfitting. As can be seen in Figure 2.4, application of a node optimization scheme improved the learner's performance over the regular scheme (where $n_{nodes} = N_b$) in nearly all cases. It should be noted that in these experiments, the linear tail of the RBF was excluded, so Equation 2.3 was modified to be $\Phi\lambda = Y$ and was solved.

$$y = y + N(0, e \times y) \tag{2.6}$$

2.3.3. Evaluating the Effect of Convergence Parameters on Algorithm Performance. Both NNGA and DYCORS have adjustable convergence parameters that control their design space

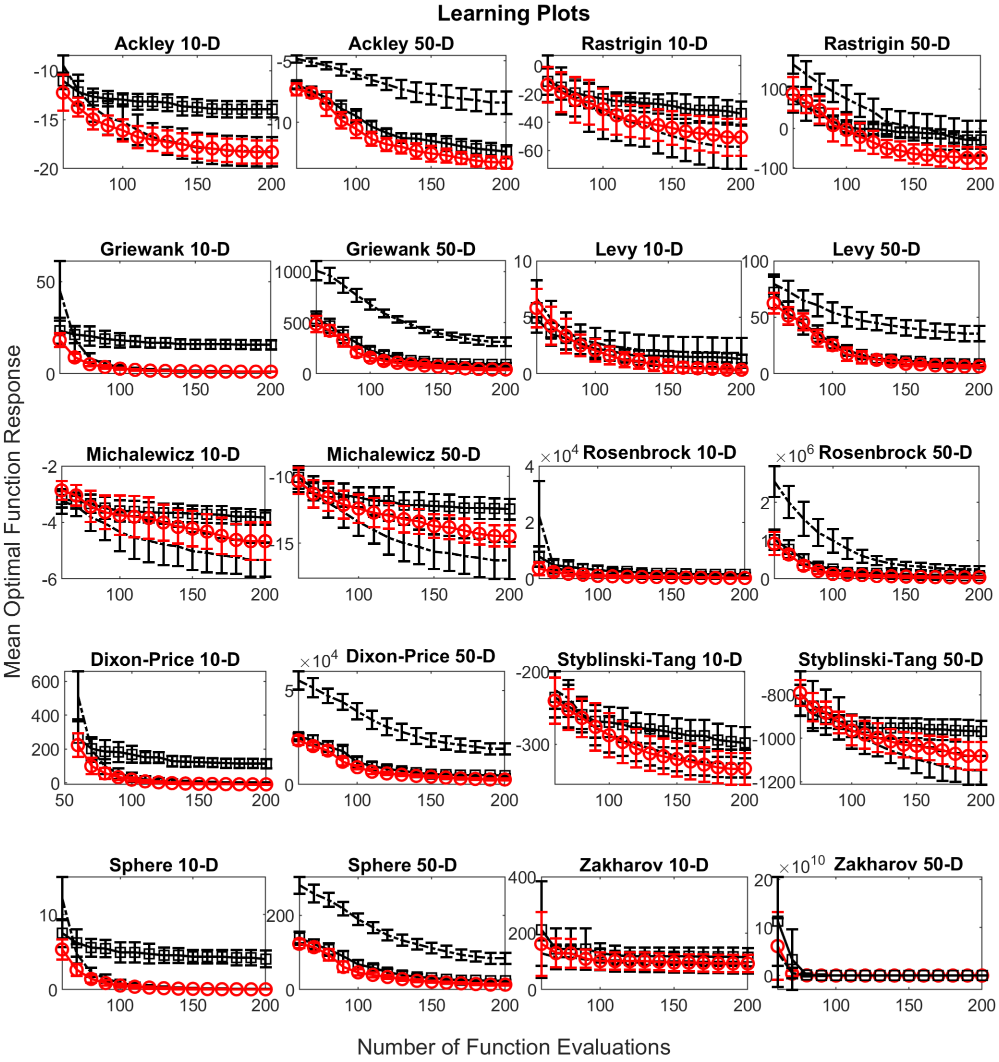


FIGURE 2.3. Hybrid algorithm performance. Squares = NNGA; dotted lines = DYCORS; circles = NNGA-DYCORS. The average minimum of response for each of the test functions in Table 2.1 is plotted against cumulative queries. The hybrid NNGA-DYCORS performs as well as the best NNGA and DYCORS results.

exploration/exploitation trade-off. In other words, both algorithms have a means of avoiding premature convergence to local minima, as predicted by an early (i.e. less accurate and general) surrogate approximation. Here we test the effect of changing these internal search parameters l_b (DYCORS)

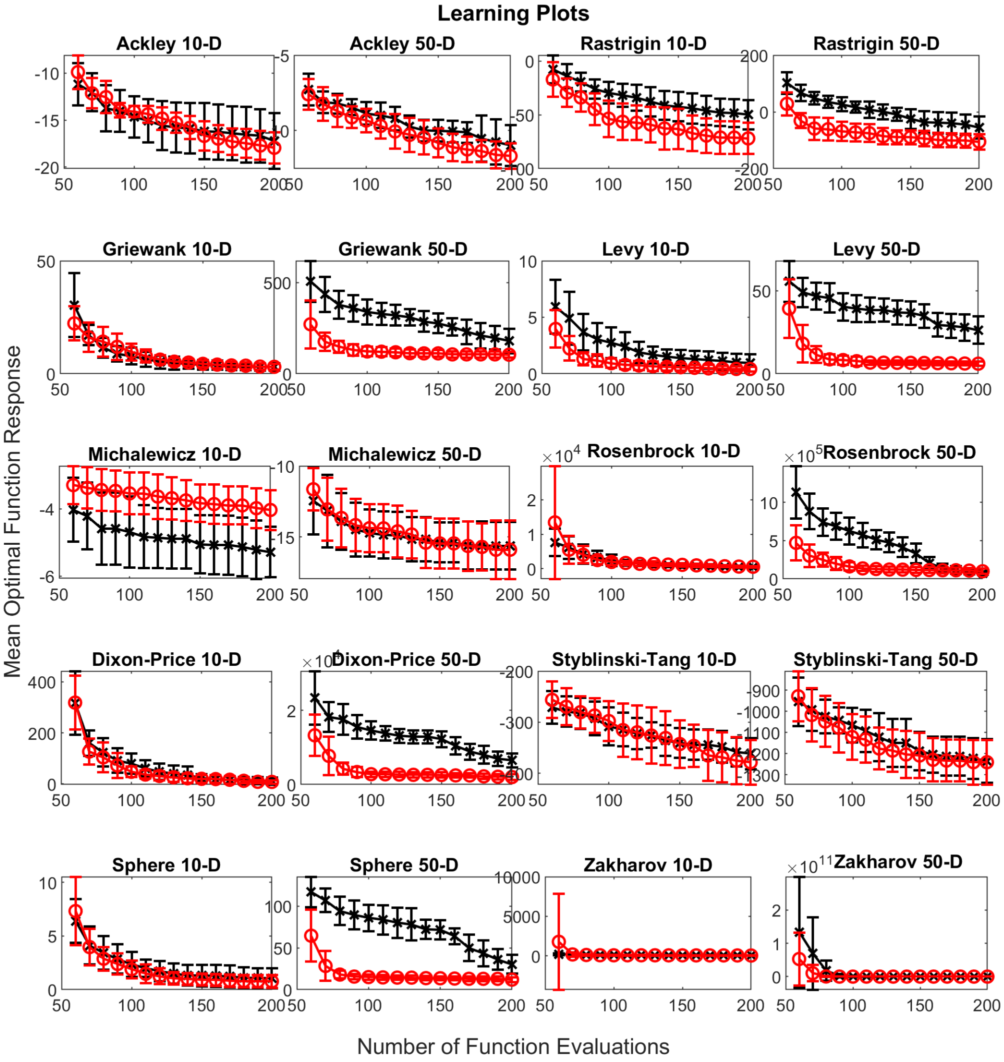


FIGURE 2.4. Algorithm Performance in the Presence of Noise. Circles = NNGA-DYCORS with node optimization; crosses = NNGA-DYCORS without scheme. The average minimum of response for each of the test functions in Table 2.1 is plotted against cumulative queries for noise level $e = 0.2$ (20% of response). Node optimization generally improves learner performance in the presence of simulated experimental noise.

and CD (NNGA). For DYCORS, this has already been suggested using a time-varying strategy **[42]** for current database size N_b. In this method, the step size is recalculated as $l_(b + 1) = C_b l_b$.

$$\theta(N_b) = 2(1 - \frac{ln(N_b - N_o + 1)}{ln(N_{max} + N_o)}) \tag{2.7}$$

$$C_b = \begin{pmatrix} 1 & \theta \geq 1 \\ \theta & 0.5 \leq \theta < 1 \\ 0.5 & \theta \leq 0.5 \end{pmatrix} \tag{2.8}$$

For NNGA, if one defines a maximum and minimum critical distance parameter $CD_1 = 0.2$ and $CD_2 = 0.05$ respectively, then CD can be changed linearly over time using the following formula:

$$CD(N_b) = ((CD_2 - CD_1)/(N_{max} - N_o)) * (N_b - N_o) + CD_1 \tag{2.9}$$

Where N_{max} and N_o are the maximum and initial database sizes respectively. The result of implementing this dynamic parameter approach (Figure 2.5) was that the hybrid learner did not have substantially better performance over the regular hybrid learner. Additionally, these results show that the internal search parameters l_b and CD do not need to be substantially altered.

2.3.4. Evaluating the Effects of a Parameter Subset Selection Algorithm. In engineering systems, certain parameters influence outputs more significantly than others, and often few parameters even matter at all. To simulate this variable response sensitivity while maintaining the nonlinearity and dimensionality of the test problems, an sensitivity vector γ was used to scale the test problems. This vector scales each problem as $x_{scaled,j} = \gamma_j \times x_j$ for parameter j so that 20% of the parameters are scaled up by $\gamma = 2$, 30% are un-scaled, 20% are scaled down by $\gamma = 0.5$, and 30% are neglected in the deterministic function. An example of γ and a scaled 2-D Ackley Function is shown in Figure 2.6 and below.

$\gamma = [2, \ldots 2, 1, \ldots 1, 0.5, \ldots 0.5, 0, \ldots 0]$

Previous work in applying a decision tree-based subset selection strategy to the NNGA algorithm **[94]** reduced the number of queries needed in optimization. To explore this further, an RBF-based

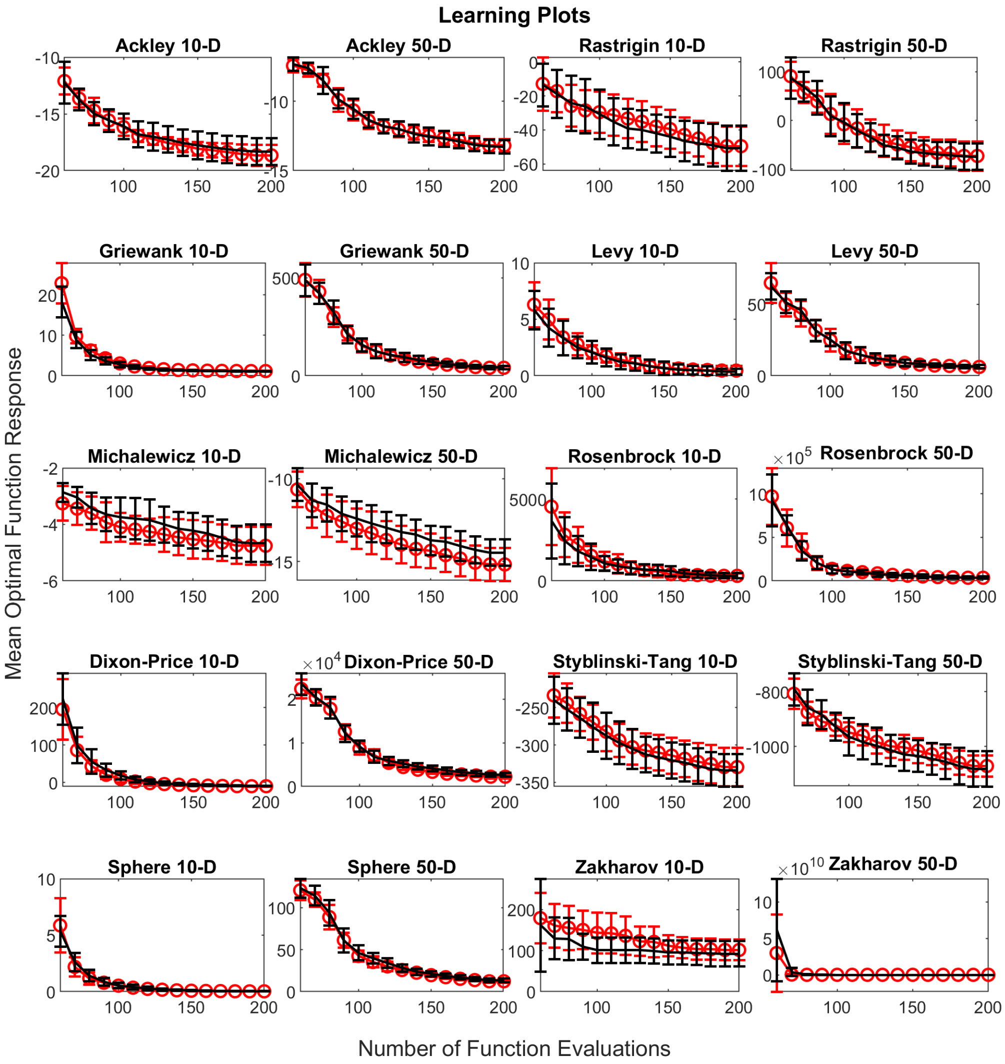


FIGURE 2.5. Algorithm Performance using Dynamic Convergence Parameter Strategy. Circles = NNGA-DYCORS with dynamic convergence parameter strategy; solid lines = NNGA-DYCORS without strategy. The average minimum of response for each of the test functions in Table 2.1 is plotted against cumulative queries. Performance is not much improved by using the dynamic strategy.

subset selection strategy was developed. After $b = 7$ batches of queries (roughly halfway through the entire set of queries), p RBF models are trained with $q = 1 \ldots p$ parameters dropped out. For each neglected parameter q, a cross-validated and averaged out-of-sample correlation coefficient R_{av}^2

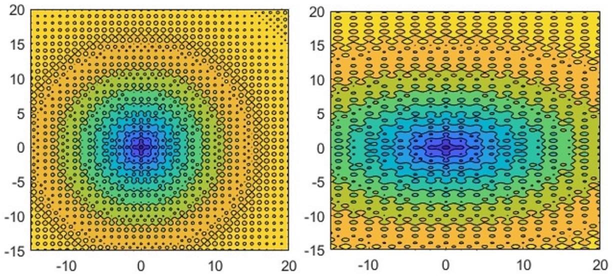


Figure 2.6. Effect of Scaling on Functions. Left figure Ackley Function, right is Ackley Function with ordinate axis modified by $\gamma = 0.5$.

is found using a separate hold-out-set of data. The most important parameters should have the lowest R^2_{av} assuming the RBF model is robust for the database. In experiments using this technique, the DYCORS algorithm selects the most important parameters and only uses that subset in the coordinate-wise perturbation, and the NNGA operates normally. The result was that, while this subset selection method was able to speed up learning in some cases (see Figure 2.7 10-D Ackley Function), it was not able to do so consistently (see Figure 2.7 10-D Michalewicz Function).

2.4. Discussion

There is a seemingly infinite number of modeling techniques, search optimization algorithms, and initialization/infill strategies in the literature to facilitate optimizing expensive objective functions. However, the characteristics of the experimental system and design space are never really known a priori, so having an algorithm that is more efficient than traditional methods and able to work with a wide variety of problems is advantageous. Therefore, the goal of this article was to develop a surrogate optimization framework that could be successfully applied to test problems with a wide range of dimensionality and degrees of nonlinearity. The NNGA-DYCORS algorithm runs two surrogate optimization algorithms in parallel. The NNGA uses a Euclidean distance-based metric to truncate a genetic algorithm, whose best members are k-means cluster distilled into a final query list. This acts as a global optimization process because the internal genetic algorithm searches over the entire design space. The DYCORS algorithm perturbs the best previous queries using a dynamic Gaussian distribution, where the perturbations are adjusted based on cumulative success and the total number of queries in the database. Thus, DYCORS acts as a local search method in

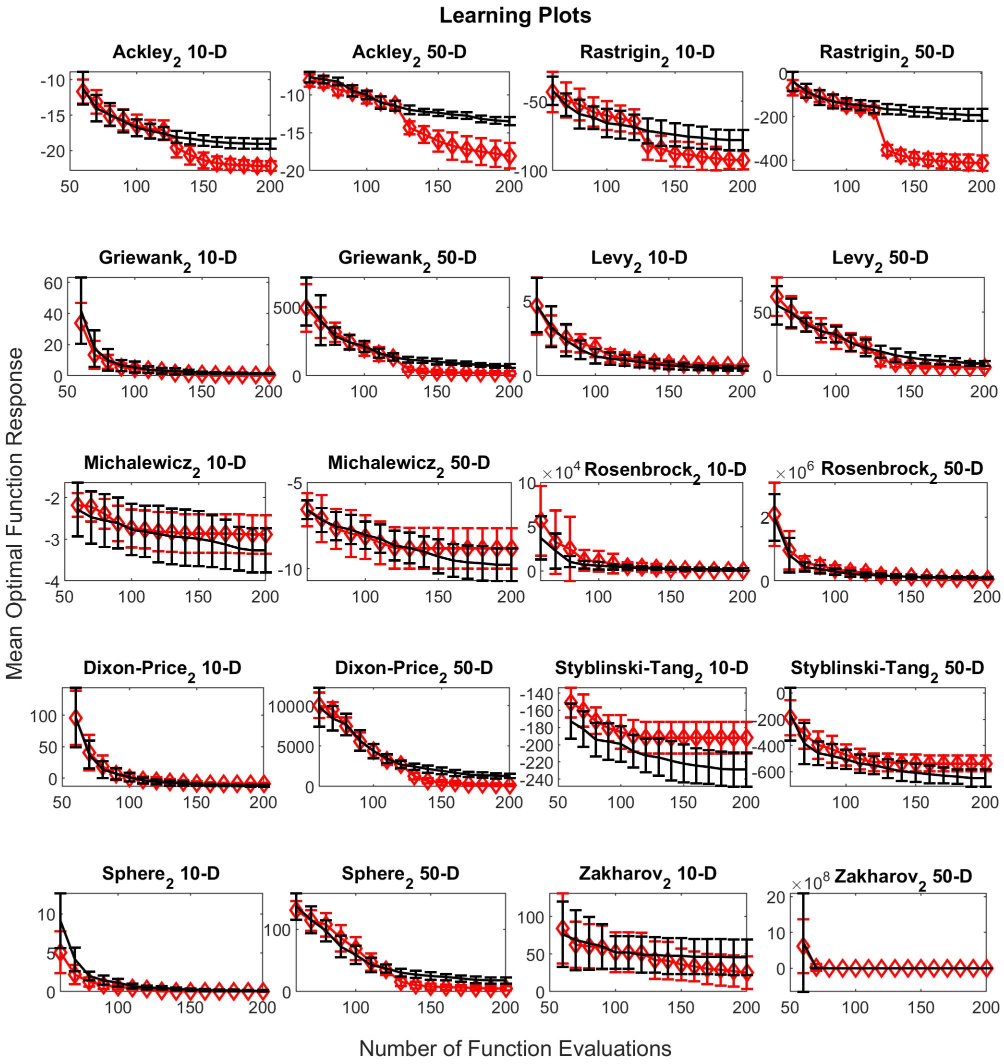


Figure 2.7. Algorithm Performance using Subset Selection. Diamonds = NNGA-DYCORS with subset selection; solid lines = regular NNGA-DYCORs. The average minimum of response for each of the test functions in Table 2.1 is plotted against cumulative queries. Subset selection does not have a consistently positive effect on algorithm performance. The subscript (2) indicates that the sensitivity vector has been applied to the test problem.

the region defined by a Gaussian centred at its best queries. Both arms of the hybrid algorithm use an RBF for prediction.

The result was that the NNGA-DYCORS hybrid algorithm was statistically equal to or outperformed its constituent algorithms in the 19 of 20 test problems. This demonstrates the robustness of the NNGA-DYCORS, as it performs as a best case scenario on a variety of test problem dimensions and shapes. This is important because, in real experimental problems, one does not know the shape of the surface a priori, highlighting the utility of a generalizable optimization framework such as the NNGA-DYCORS. In addition, it is never clear what constitutes a 'high' and 'low' - dimensionality design problem, so an algorithm that performs well in arbitrary dimensions should have large practical value. The DYCORS algorithm was already shown to be competitive compared to other heuristics [**66**], and the NNGA was demonstrated to be significantly more efficient than traditional experimental optimization methods [**92**]. It stands to reason that this hybrid framework should extend the usefulness of both algorithms to test problems of arbitrary dimensionality and degree of nonlinearity.

Using a node optimization scheme to reduce model variance during query selection improves hybrid algorithm performance, especially for noisy surfaces (as could be the case in experimental situations). Practitioners should therefore consider built-in regularization to avoid overfitting of the data when dealing with expensive, data-sparse and noisy systems. Optimizing the number of nodes was specific to this RBF variant, but the optimization loop in Section 3.2 could be applied to any model hyperparameter. In the next set of experiments, the method of making the NNGA-DYCORS convergence parameters dynamic during query selection did not improve performance. This indicates that (1) it may not be fruitful to pursue extensive algorithm parameter adjustments/heuristics for this algorithm, and (2) there is little sensitivity in the selection of algorithm convergence parameters on the outcome, unlike the results in previous articles on the subject [**42, 94**]. Finally, to mimic typical engineering scenarios where response sensitivity varies with the inputs, the test functions were scaled with a sensitivity vector. A subset selection strategy was unable to consistently improve on the regular NNGA-DYCORS performance by focusing the coordinate search on the most sensitive sets of parameters. This may be because the RBF does not adequately model a given test function, so it does not correctly identify the most important parameters in the database, or the coordinate search method does not properly exploit the narrowed parameter space. Generically, it may be useful to reduce the dimensionality of the parameter space, but the strategy of doing so using model adherence ‘drop-out’ experiments was not uniformly successful.

This article demonstrates that the NNGA-DYCORS hybrid learning algorithm outperforms its constituent algorithms in the important criteria of robustness and generalizability to different kinds of problems. Thus, this algorithm can be applied to a wide variety of physical and biological design optimization problems with a degree of assurance that parameter estimates will be optimal while minimizing necessary resources. In addition, as this hybrid is both robust and highly generalizable to many types of design problems, it should be useful for practitioners who are not experts in surrogate optimization methods, and work on a variety of problems of diverse complexity.

CHAPTER 3

Optimization of Muscle Cell Culture Media using Nonlinear Design of Experiments

Optimizing media for biological processes, such as those used in tissue engineering and cultivated meat production, is difficult due to the extensive experimentation required, number of media components, nonlinear and interactive responses, and the number of conflicting design objectives. Here (originally published as [**22**]) we demonstrate the capacity of a nonlinear design- of-experiments (DOE) method to predict optimal media conditions in fewer experiments than a traditional DOE. The approach is based on a hybridization of a coordinate search for local optimization with dynamically adjusted search spaces and a global search method utilizing a truncated genetic algorithm using radial basis functions to store and model prior knowledge. Using this method, we were able to reduce the cost of muscle cell proliferation media while maintaining cell growth 48 h after seeding using 30 common components of typical commercial growth medium in fewer experiments than a traditional DOE (70 vs. 103). While we clearly demonstrated that the experimental optimization algorithm significantly outperforms conventional DOE, due to the choice of a 48 h growth assay weighted by medium cost as an objective function, these findings were limited to performance at a single passage, and did not generalize to growth over multiple passages. This underscores the importance of choosing objective functions that align well with process goals.

3.1. Introduction

Cell culture media is a critical component of bioprocesses such as pharmaceutical manufacturing and the emerging field of cultivated meat products. Optimizing culture media is a difficult task due to the extensive experiments required, number of media components, nonlinear and interactive responses from each component, and conflicting design objectives. Additionally, for cultured meat products, media needs to be less expensive than those currently deployed for other cell culture processes (e.g. biopharmaceutical production), food-grade, consider safety, component stability, and effects on sensory characteristics of final products. Without much in the way of first principles models for these objectives, especially for adherent mammalian muscle cells used for cultivated meat production (as well as fat and connective tissues), media optimization must be done experimentally with constraints on inputs, outputs, and number of experiments.

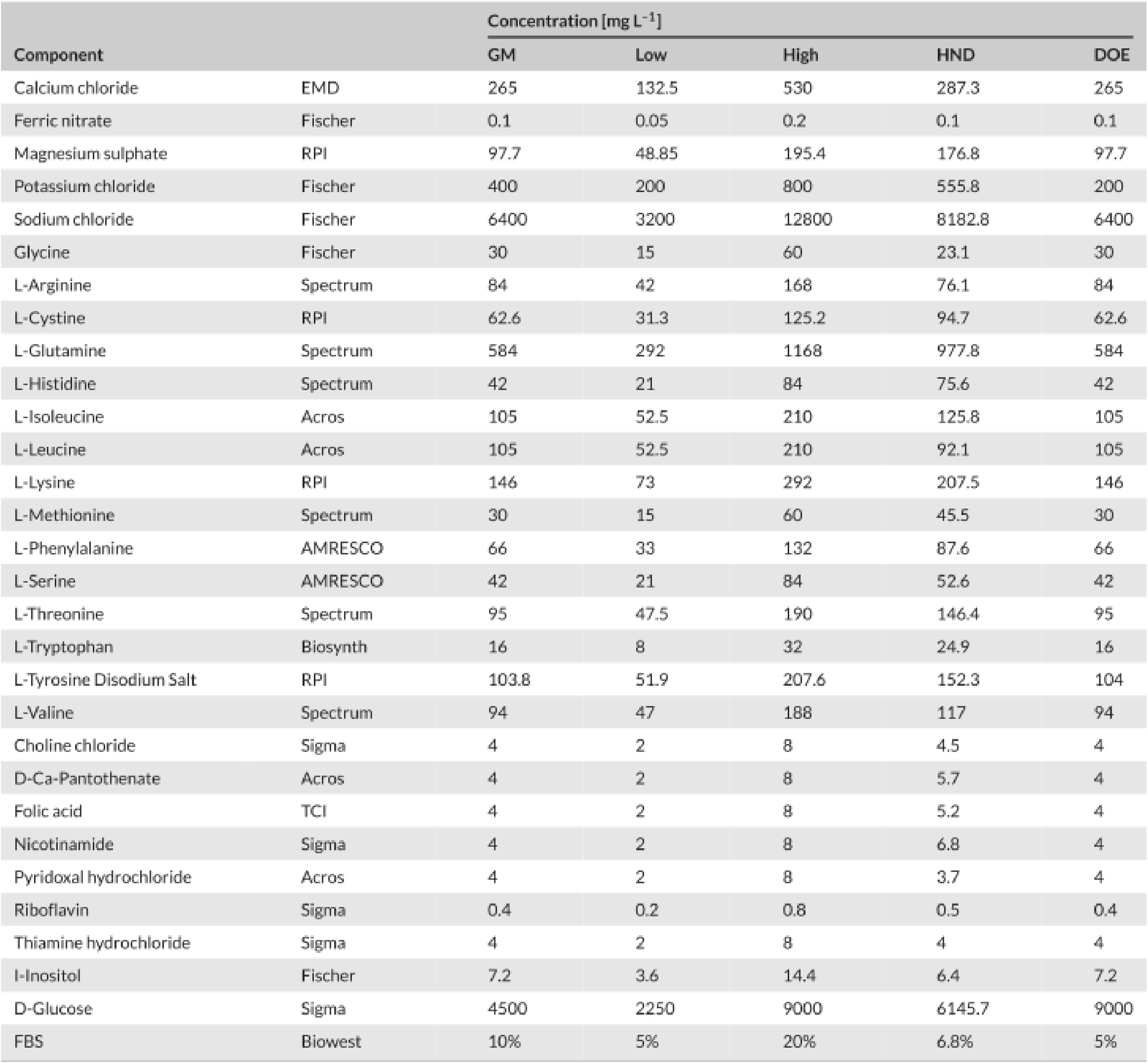
Optimizing one factor at a time or with random experiments is still the most common way of exploring design space. This strategy is very inefficient for large systems (culture media such as DMEM may have up to 30 components [**3**]) and is unable to consider interactions among media components. Design-of-Experiments (DOE) methods are better able to manage large numbers of components in fewer experiments using Factorial, Fractional Factorial, Plackett-Burman, and Central Composite Designs where linear and polynomial models can correlate first order and interactive effects of media components. In general, DOE methods are able to optimize < 10 variables [**74**] and with the help of screening designs can solve problems > 25 variables [**90**], though at the expense of ignoring interactions, screened variables, and easily costing > 100 experiments (when combining typical screening and factorial experiments, although this number can be quite lower if < 5 variables are explored). Experimental optimization of media has also been done using stochastic methods such as genetic algorithms [**34**] and this approach is generally suited to optimizing systems of dimensionality > 15 where DOE methods can become experimentally cumbersome, but also take $\sim$ 200 experiments.

Because the size of the design space increases exponentially with the number of design variables, a natural advance was to use response surface models to capture information about interactions and nonlinearity. These techniques can then be used to sequentially identify optimal culture conditions while simultaneously improving modeling accuracy. Oftentimes experimenters will employ polynomial models to find optimal culture conditions [**69**] but only after extensive DOE to reduce the dimensionality of the problem space to < 5. More advanced modeling techniques are neural networks, decision trees [**92**] and Gaussian processes [**45**] which are often better at generalizing noisy, nonlinear, and multi-modal data. When combined with global optimization methods. Zhang and Block demonstrated that these response surface methods can optimize problems with > 20 variables in less than half the number of experiments as traditional DOE [**93**].

In the previous chapter, this author further improved the robustness of this algorithm by using a hybrid optimization scheme validated on simulated design problems (NNGA-DYCORS). Here we employ this novel nonlinear experimental design algorithm (called HND in this chapter in order to use the same nomenclature as the publication [**22**]) to optimize the proliferation of C2C12 cells while simultaneously reducing media cost by modeling the response surface of culture conditions using an RBF with a hybridized global/local optimization scheme. We then compare this approach to a more traditional DOE method. The organization of this article is as follows: Section 3.2 includes an outline of the experimental and computational methods use in media optimization, Section 3.3 goes over the results and Section 3.4 details a discussion of the results and current challenges.

3.2. Materials and Methods

3.2.1. Media Components and Cell Line. Table 3.1 lists the 30 components of the media system, concentration ranges, and the concentration of the control growth media (GM) used in this work. GM is based on a formulation of DMEM + 10% FBS from HiMedia Cell Culture with 4.5 g/L Glucose and L-Glutamine where FBS is fetal bovine serum (Biowest). All components were stored as aqueous stock solutuions in 2-6 °C sterilized using 0.2 μm pore size micro-filtration (Pall Corporation Acrodisc). The pH was adjusted to 7.2 using 1 M HCl or NaOH solution, and Sodium Bicarbonate (Sigma) buffer at 1850 mg/L was added. C2C12 muscle cells were used for all experiments (ATCC). The cells were stored in liquid N_2 in 10% DMSO (Sigma), 20% FBS, 70% GM at passage 15.



		Concentration [mg L^{-1}]				
Component		GM	Low	High	HND	DOE
Calcium chloride	EMD	265	132.5	530	287.3	265
Ferric nitrate	Fischer	0.1	0.05	0.2	0.1	0.1
Magnesium sulphate	RPI	97.7	48.85	195.4	176.8	97.7
Potassium chloride	Fischer	400	200	800	555.8	200
Sodium chloride	Fischer	6400	3200	12800	8182.8	6400
Glycine	Fischer	30	15	60	23.1	30
L-Arginine	Spectrum	84	42	168	76.1	84
L-Cystine	RPI	62.6	31.3	125.2	94.7	62.6
L-Glutamine	Spectrum	584	292	1168	977.8	584
L-Histidine	Spectrum	42	21	84	75.6	42
L-Isoleucine	Acros	105	52.5	210	125.8	105
L-Leucine	Acros	105	52.5	210	92.1	105
L-Lysine	RPI	146	73	292	207.5	146
L-Methionine	Spectrum	30	15	60	45.5	30
L-Phenylalanine	AMRESCO	66	33	132	87.6	66
L-Serine	AMRESCO	42	21	84	52.6	42
L-Threonine	Spectrum	95	47.5	190	146.4	95
L-Tryptophan	Biosynth	16	8	32	24.9	16
L-Tyrosine Disodium Salt	RPI	103.8	51.9	207.6	152.3	104
L-Valine	Spectrum	94	47	188	117	94
Choline chloride	Sigma	4	2	8	4.5	4
D-Ca-Pantothenate	Acros	4	2	8	5.7	4
Folic acid	TCI	4	2	8	5.2	4
Nicotinamide	Sigma	4	2	8	6.8	4
Pyridoxal hydrochloride	Acros	4	2	8	3.7	4
Riboflavin	Sigma	0.4	0.2	0.8	0.5	0.4
Thiamine hydrochloride	Sigma	4	2	8	4	4
I-Inositol	Fischer	7.2	3.6	14.4	6.4	7.2
D-Glucose	Sigma	4500	2250	9000	6145.7	9000
FBS	Biowest	10%	5%	20%	6.8%	5%

Table 3.1 Details of media design space | components and bounds used in media optimization for proposed method (HND), control optimization method (DOE), and commercial (GM) indicated.

To generate enough cells for these experiments, cells were taken out of storage, thawed, centrifuged at 1500 × g for 5 min and re-suspended in DMEM (Glibco) + 10% FBS in 15 cm cell culture plates (Cellstar, Greiner Bio-One). Cells were then trypsinized (Gen-Clone) in their log phase of growth (∼ 50% confluence, or two days of growth) and plated on 96 well plates (Cellstar, Greiner Bio-One). To plate the cells, trypsinized cells are suspended in phosphate buffered solution (PBS Glibco) and counted using a hemocytometer. The PBS volume was then adjusted so that

5000 cells per well ($\sim$ 15,625 cells/cm^2) could be seeded using 50 μL of PBS into 150 μL of the media being tested (total well volume of 200 μL). The cells were incubated at 37° and 5% CO2 for 48 h post-seeding before measurements of proliferation were made with replicates. For six well plate experiments (Cellstar, Greiner Bio-One) a total volume of 3 mL was used with the same ratios of PBS to media and seeding density (150,000 cells per well), with all other steps being the same.

3.2.2. Assays and Objective Function. After 48 h of incubation, the performance of the media was measured using AlamarBlue [**37**] metabolic colorimetric assay (AB). After pipetting in 10% volume of AB assay (20 μL) for each well, all wells were left to incubate for 3 h at 25° and 5% CO2. The %AB reduction was measured using a microplate reader at 600 and 570 μL using Equation 3.1 with six replicates of each experimental and control well.

$$\%AB = \frac{117216\lambda_{570,media} - 80586\lambda_{600,media}}{155677\lambda_{600,control} - 14625\lambda_{570,control}} \tag{3.1}$$

To quantify the relative proliferation of cells after 48 h of growth, the ratio of %AB for a given medium to %AB for basic GM was used as a metric of the success. The economic cost of a medium was considered by normalizing the %AB ratio by the volume of FBS, which constitutes the vast majority of the media cost [**77**]. Therefore, the objective function α and the optimization problem used in this work (finding the best media components x^*) are as follows, where x_{FBS} is the normalized volume of FBS ranging from $[0, 1]$.

$$\begin{gathered} x^* = argmax_x \alpha(x) \\ \alpha(x) = \frac{\%AB/\%AB_{GM}}{1+x_{FBS}} \\ \bar{x} = \frac{x_i - x_{i,low}}{x_{i,high} - x_{i,low}} \end{gathered}$$

This objective function strikes a balance between a proportionality to cell proliferation and cost, and ease of use. A more elaborate objective function that describes multi-passage dynamics or further economic costs could be employed, but at the expense of significantly more time and labor.

3.2.3. Experimental Design Algorithm. A novel hybrid nonlinear experimental design algorithm (HND) was developed in the previous chapter to optimize high dimensional experimental design systems such as the one outlined above. It is based on a truncated genetic algorithm (TGA) method [**93**] hybridized with a dynamic coordinate search framework (DYCORS) [**66**]. This method starts by constructing an RBF approximation $\hat{y}$ of the system from an initial set of experiments with inputs and outputs $\{X_0, \alpha_0\}$. The RBF takes the form of a sum of n_c cluster λ_i-weighted radial functions $\phi(x, x')$ in Equation 3.2.

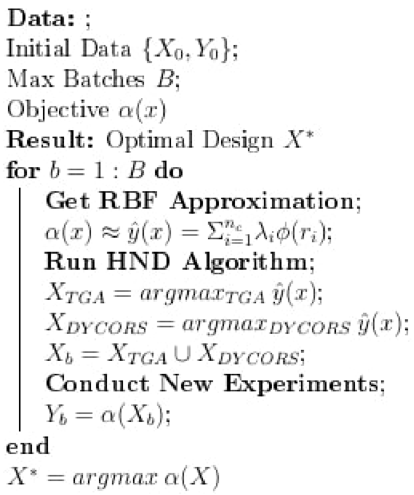
$$\hat{y} = \sum_{i=1}^{n_c} \lambda_i \phi(r_i) \tag{3.2}$$

The radial functions project a set of $[0, 1]$ normalized inputs x and x' (in this case two media concentrations) into a single output space using the Euclidean distance $r = ||x - x'||_2$. This quantifies the difference between two media combinations. Two media that are more similar have smaller r values, so are going to have similar predictions of $\hat{y}$. The radial function used in this work was the cubic function $\phi(x, x') = r^3$. The weights are determined by solving the linear equation for $\Phi(X, X)$ for a training set of data that has been collected $\{X, \alpha\}$.

$$\lambda = (\Phi\Phi)^{-1}\Phi^T\alpha \tag{3.3}$$

To find the optimal locations of the RBF nodes n_c we used the K-means clustering algorithm. This algorithm was repeated for $K = 4$ cross-validated data splits for each batch of experiments, where the n_c with the lowest cross-validated error for the given training set was chosen as the optimal number of clusters. Cross-validation is critical for making sure models generalize well for small amounts of noisy data. In general, higher n_c makes the model more complex (wiggly), so here we balance accuracy with model simplicity / generalizability.

Using the trained RBF model, the two arms of our algorithm, TGA and DYCORS, each suggest five experimental conditions for a total of 10 experiments per batch within the design space $[\times 1/2, \times 2]$ of the GM (see Table 3.1) that optimize α. The TGA arm runs a genetic algorithm (a stochastic global optimization method) over the RBF model to predict the best designs. Because the model is based on a small amount of noisy data, the genetic algorithm is stopped before it can converge to implicitly consider model and experimental uncertainty. The DYCORS arm of the algorithm searches in the region around the best design and picks the best predicted set of designs



```
Data: ;
Initial Data $\{X_0, Y_0\}$;
Max Batches $B$;
Objective $\alpha(x)$
Result: Optimal Design $X^*$
for $b = 1 : B$ do
    Get RBF Approximation;
    $\alpha(x) \approx \hat{y}(x) = \Sigma_{i=1}^{n_c} \lambda_i \phi(r_i)$;
    Run HND Algorithm;
    $X_{TGA} = argmax_{TGA}\, \hat{y}(x)$;
    $X_{DYCORS} = argmax_{DYCORS}\, \hat{y}(x)$;
    $X_b = X_{TGA} \cup X_{DYCORS}$;
    Conduct New Experiments;
    $Y_b = \alpha(X_b)$;
end
$X^* = argmax\, \alpha(X)$
```

Figure 3.1. HND Algorithm

in that region, which expands and contracts based on the quality of previous experiments. The new experiments are conducted and the resulting data is used to correct and retrain the RBF model. To allow the RBF model to generalize better during early periods of optimization, 30 randomly selected experimental conditions were taken initially. The optimization loop was stopped when the α quality of the media showed a lack of improvement. The general framework for the HND is shown in Figure 3.1.

As a control method, a traditional DOE was used to optimize the same media design problem in three steps. (i) A 'Leave-One-Out' (LOO) experiment was conducted where a media composed of all components at their GM concentrations, excluding each individual component,were tested for their proliferation capacity using the %AB metric (α was not used because all media had the same amount of FBS), similar to what was done in previous work [**94**]. The lowest performing components had their concentrations fixed at their respective GM concentrations. Next (ii) a Folded/Un-Folded Plackett-Burman design was implemented with the remaining components at the upper and lower bounds of the design problem. This was done to determine the first order linear effects of each component on the objective function α. A linear model to predict α was used in conjunction with a LASSO algorithm to rank the most important first order effects, and all but the highest impact components were kept at their GM concentrations. Finally, (iii) the remaining components were used to design a Central Composite Design (CCD) where experiments are spread out across the design space to more thoroughly explore potential optimal designs.The best α design from this DOE method was considered the optimal DOE design.

3.2.4. Software and Hardware. Hardware used: Dell Precision 5820 Tower, Intel Xeon W-2145 DDR4-2666 Processor (3.7 GHz), 32 GB Memory. Software used: `MATLAB R2019a` with `Bioinformatics` Package.

3.3. Results

3.3.1. Performance of Traditional DOE for Media Optimization. The DOE-LOO step identified Ferric Nitrate, $MgSO_4$, Glycine, L-Isoleucine, Choline Chloride, Riboflavin, and Thiamine HCl as components that, when left out of GM, had no (or positive) statistical effect on %AB after 48 hr post-seeding (30 experiments needed). These components were set to their respective GM concentration for all subsequent DOE experiments. Next, the DOE-PB with LASSO identified the six most α-important components of the remaining 23 components (KCl, L-Glutamine, Glucose, FBS, L-Cystine, L-Serine). To reduce the number of experiments for the DOE-CCD design, L-Cystine and L-Serine were kept constant at × 1/2 normalized units above and below their GM midpoint concentrations respectively (10.4 and 28 mg/mL)based on the sign of their coefficients (48 experiments required). The remaining four components in the CCD had their upper/lower bounds changed to × 1/2 normalized units above (KCl, L-Glutamine, FBS) and below (Glucose) their GM midpoints. The remaining components were varied in a CCD design, with the best medium being 200 mg/L KCl, 388 mg/L L-Glutamine, 9000 mg/L Glucose, 5% FBS (25 experiments) shown in detail in Table 3.1. An 80% increase in α at 48 hr post-seeding over GM was measured (Figure 3.2 left) using 50% less FBS than GM.

3.3.2. Performance of Novel HND for Media Optimization. For the HND optimization loop, α was used as the objective function and calculated using %AB measured at 48 hr post-seeding at 96 well plate scale (the exact same as the DOE method). The RBF was initially trained with 30 randomly selected experiments. Figure 3.2 shows that the average HND designs improved in both α and %AB metric over time (both cost and proliferation) quickly overcoming standard GM and achieving similar results to the best DOE design (an α difference of 13.3%) with 70 experiments. We have included the proliferation metric (%AB / %AB GM) in Figure 3.2 for completeness even though

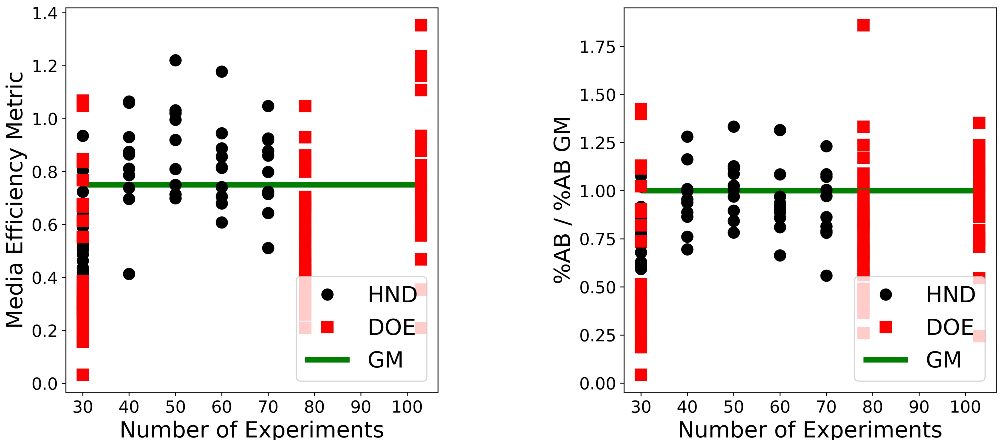


FIGURE 3.2. Iterative improvement of media using HND and DOE | (left) media efficiency metric (right) %AB Proliferation. Both HND and DOE improve over GM.

it was not used as the objective function α in this work. The HND was stopped at 70 experiments because both %AB and α stopped improving. The best medium found had an α measured to be 56% better than GM during the optimization loop using 32.5% less FBS than GM.

3.3.3. Comparison of Media Resulting from Novel HND and Traditional DOE. Figure 3.3 shows the differences between the optimal media. For the most part the HND identified optimal concentrations that were slightly elevated compared to DOE, except for KCl, FBS, and Glucose. It is also notable that both HND and DOE determined that Glucose and FBS should be elevated and reduced in relative to GM. Figure 3.4 shows the media efficiency metric α plotted against the component concentrations for all experiments, demonstrating the nonlinear, interactive, and ultimately non-trivial nature of this experimental design optimization problem. These α-optimal HND and DOE designs were then tested against GM using %AB at 24, 48, and 72 h post-seeding (Figure 3.5), where the designed media have high %AB relative to GM but that advantage is reduced over time. As a further check, α was calculated using raw cell number normalized by the volume of FBS in each experiment (at six well plate scale) where it was found HND and DOE again outperformed GM (Figure 3.5) in terms of the objective function α due to their lower cost. However, both HND and DOE produced 8% and 9% fewer cells respectively, using 70 and 103 total experiments respectively.

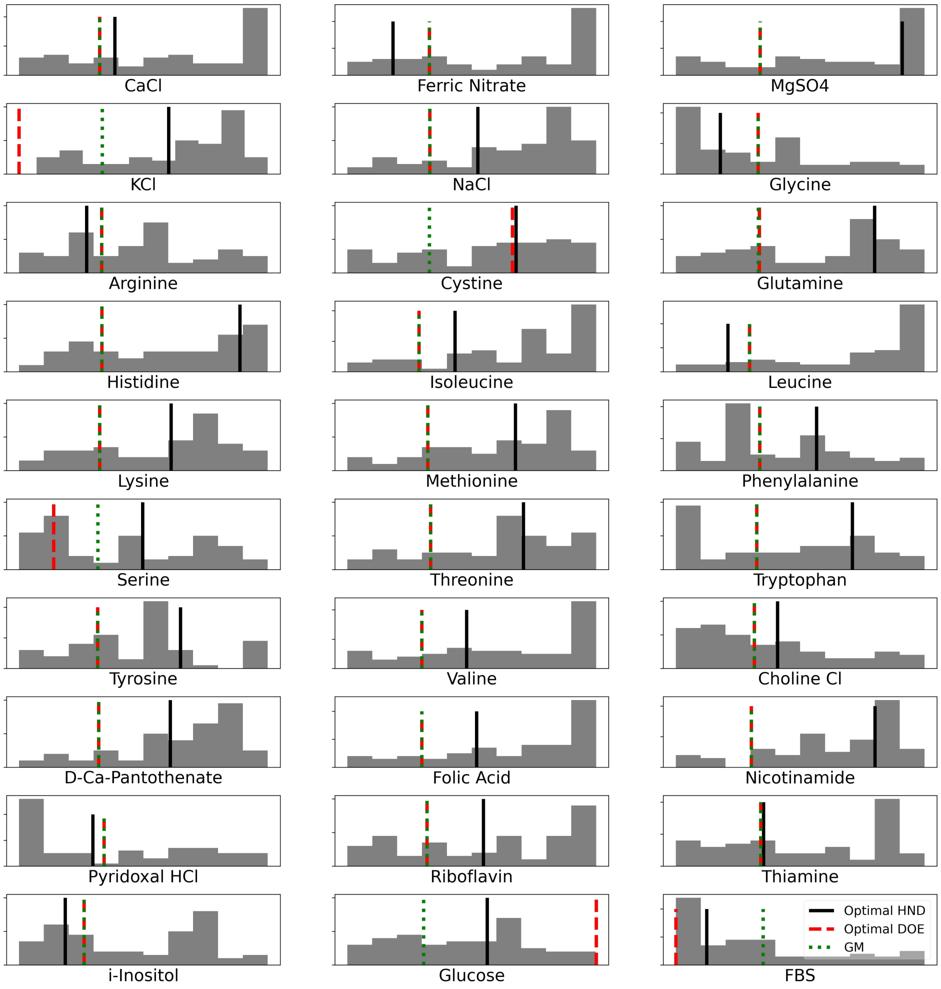


FIGURE 3.3. Distribution of components generated by HND | histogram of HND chosen component concentrations from low to high bound, best DOE and HND results also compared to GM (as horizontal lines and in Table 3.1.

3.3.4. Evaluation of Optimized Media in Multi-Passage Proliferation. Finally, the C2C12 cells were grown in optimal HND, DOE, and GM across five passages to mimic an industrial process where multi-passage dynamics could have a large effect on media design. Figure 3.6 indicates GM cumulatively grew more cells than HND and DOE optimal media by the second passage, and

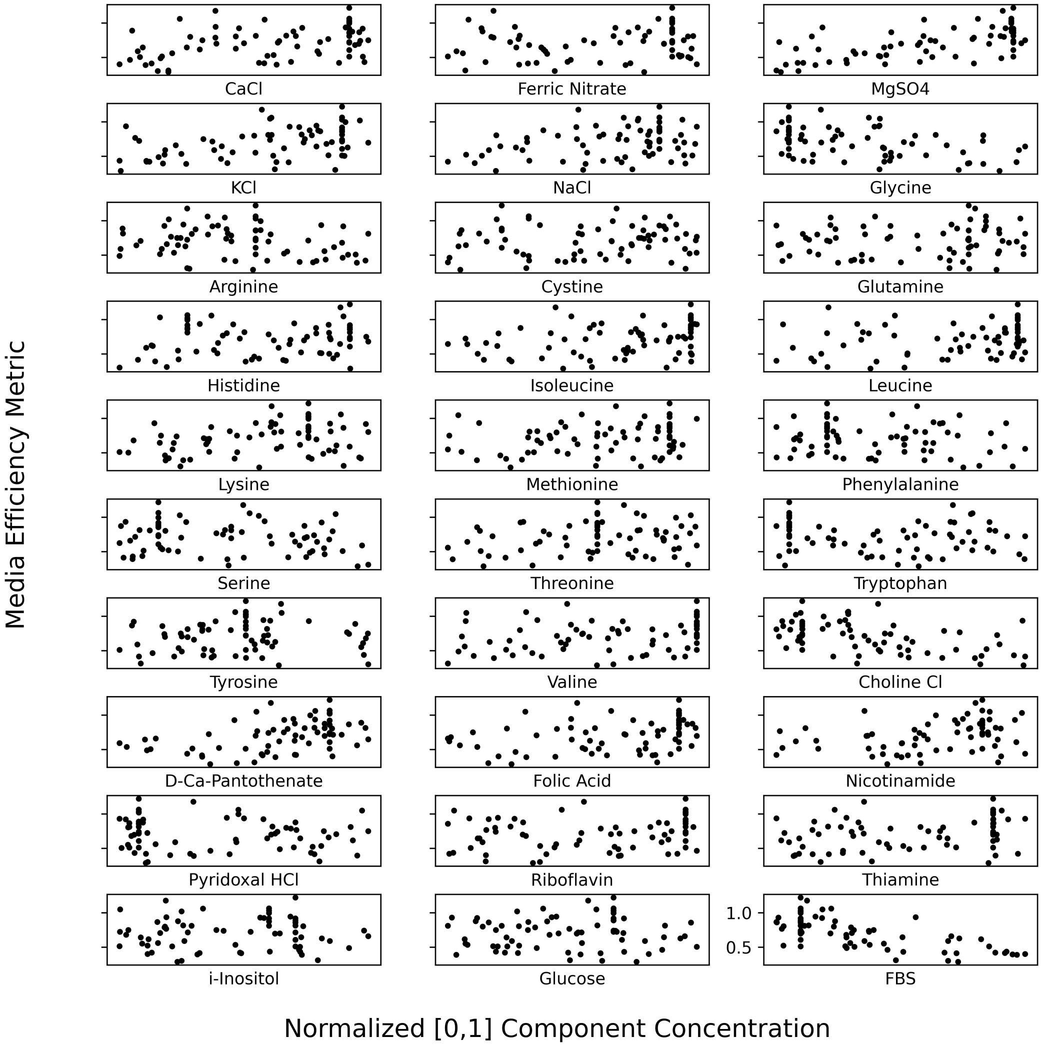


FIGURE 3.4. Input and output of media generated by HND | each dot represents an experiment designed by HND at a chosen component concentrations (normalized to be 0 to 1) and the respective media efficiency metric α

by the third passage had done so at a higher α (again, approximated by number of cells normalized by volume of FBS). Both the optimal HND and DOE media performed roughly the same in terms of cumulative number of cells and media efficiency, but with ×9 and ×11 fewer cells than GM respectively and without a proportional decrease in cost per cell.

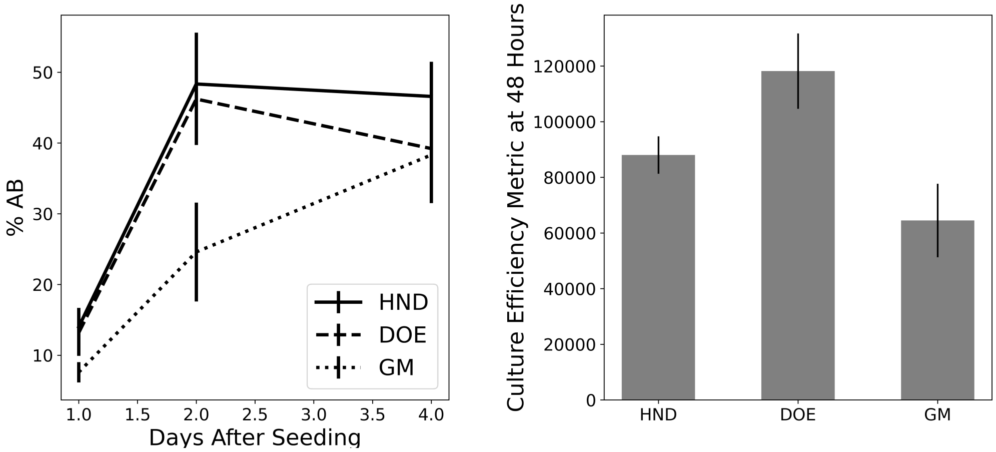


FIGURE 3.5. Result of optimal HND and DOE experiments | (left) %AB Proliferation over time in 96 well plates, error bars are standard deviation of six replicates, seeded at 5000 cells per well (right) cell efficiency metric at 48 h post-seeding in six well plates, error bars are standard deviations of three replicates, seeded at 150,000 cells per well. The media efficiency metric was approximated here by dividing number of cells by concentration of FBS. Raw cell number for HND, DOE, and GM were 594,000, 590,000, and 640,000 cells per well respectively.

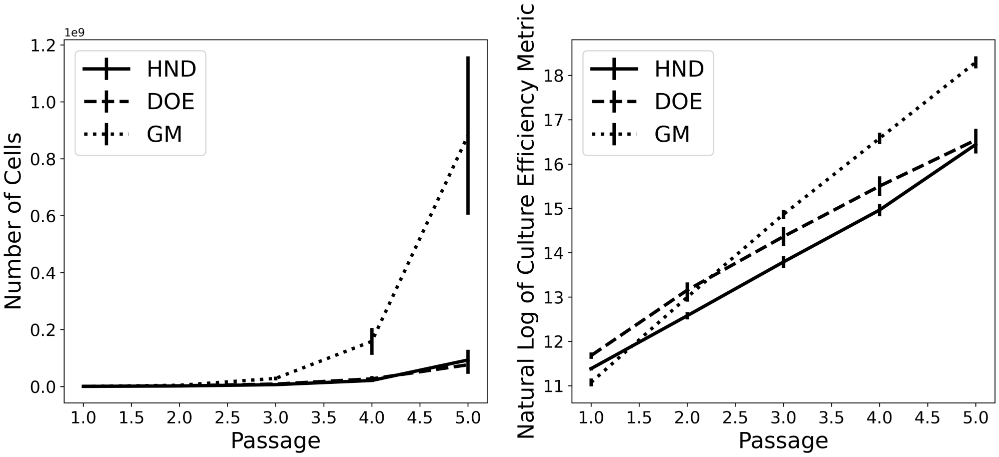


FIGURE 3.6. Optimal media over multiple passages | these media were the best found in optimization experiments. All cell numbers were taken at 48 h post-seeding using a hemocytometer in six-well plates, error bars are standard deviations of three replicates, seeded at 150,000 cells per well (left) (right) natural log of approximate efficiency of media. The media efficiency metric was approximated here by dividing number of cells by concentration of FBS.

3.4. Discussion

It is notable that, despite 30 components used, the HND was able to design a similar media to DOE with a similar degree of proliferation %AB and α in fewer experiments. Additionally, this DOE was more efficient than any single DOE, suggesting that the HND is much more efficient and simpler to use than the typical approach to high dimensional optimization. This is valuable in optimizing media due to the difficulty in collecting large amounts of data with many components. The reasons for the success of this method are likely (i) the balance between global and local optimization, and (ii) the ability of the HBD to accumulate information using the RBF, which can regress on nonlinear, noisy, and interaction-heavy problems, reducing the need for cumbersome dimensionality-reduction experiments used in the traditional DOE.

For the most part HND suggested higher concentrations of most media components than GM or DOE, except for KCl, FBS, and Glucose. This is likely because the DOE method utilized dimensionality reduction. That is, factors that demonstrated insignificant effects were fixed at their GM level and no longer included in the optimization. On the other hand, HND could vary components throughout the optimization process, including increasing component concentrations when they had even a small positive effect. Inclusion of a per component cost (rather than just the cost of FBS) might dampen this effect.

While the RBF can model nonlinear and interactive processes, the effect of each component on α is unclear without further experiments or model validation, a disadvantage of the HND approach. Nonetheless, sensitivity analysis using VARS [**64**, **65**] was conducted and indicates FBS, Glucose, and $MgSO_4$ likely have a significant effect on α, while other effects are more difficult to determine with the limited data available. Sobal sensitivity analysis utilizing polynomial regression like- wise determined FBS, MgSO4, and L-Phenylalanine were the most explanatory components when taking component-component interactions into account. Focusing on optimizing only those components might bring further improvements, which is now feasible because fewer experiments were needed to arrive at this conclusion. Another issue was that the HND algorithm often did not change experimental conditions enough, leading to heavy clustering around early high performing local optima (as seen in Figure 3.3 and 3.4). Myopia (short-termism) should be encoded into the DYCORS arm of the HND to allow for more exploration of the design space, while balancing the need for exploitation of regions of the design space that show promise. It is also possible that initializing the optimization with a more dispersed design would yield a more successful optimization. However, results from [**95**] indicate that the initialization strategy used may not have a large effect. In reality, the impact of initialization is likely to be a strong function of the design surface and how close initial points are to the true optimum, neither of which are know a priori.

Using α as a metric, HND performs similar to DOE, and both better than GM (Figure 3.2). This is true over multiple days after cell seeding and is true when using cell number to calculate α (Figure 3.5), seemingly validating the use of %AB at 48 hr post-seeding in approximating proliferation more generally. However, when measuring cell number at multiple passages (Figure 3.6) both designed media perform worse than GM. This is because the objective function α relied on measurements without multiple passages, so does not account for the dynamics of long-term cellular growth. This was a major shortcoming of the objective function picked, but not the HND or DOE itself. Future work in media design should incorporate more relevant metrics for optimization, such as a multi-passage objective function. Additionally, the %AB metric was not a perfect measure of cell number. Figure 3.5 (left) and Figure 3.2 appears to indicate HND and DOE media outperform GM, but when cell number is measured both optimal media have 8–9% fewer cells. Because AlamarBlue is a metabolic indicator, using it in the objective function for both methods may have biased the process towards higher metabolic activity rather than more proliferation.

Despite these shortcomings, the HND has been demonstrated to be able to optimize high dimensional experimental systems. In our previous work in media optimization, fewer variables (21 components) required more experiments (73–94 data points) to complete. In this work, we demonstrate optimization of 30 components with 70 experiments with no dimensionality reduction or screening designs, to our knowledge, a unique accomplishment in experimental optimization efficiency. Therefore, this represents a valuable proof of concept in the field of experimental optimization. While not able to fully replace first principles understanding of systems often based on the DOE approach (which is ill-advisable in any case), we show that the HND could aid in the optimization of the hardest design problems, including those found in the bioprocessing and larger cultivated meat industry, reducing the cost of experimentation and time-to-market for a new product.

CHAPTER 4

Multi-Information Source Bayesian Optimization of Culture Media for Cellular Agriculture

Culture media used in industrial bioprocessing and the emerging field of cellular agriculture is difficult to optimize due to the lack of rigorous mathematical models of cell growth and culture conditions, as well as the complexity of the design space. Rapid growth assays are inaccurate yet convenient, while robust measures of cell number can be time-consuming to the point of limiting experimentation. In this study (originally published as [**23**]), we optimized a cell culture media with 14 components using a multi-information source Bayesian optimization algorithm that locates optimal media conditions based on an iterative refinement of an uncertainty-weighted desirability function. As a model system, we utilized murine C2C12 cells, using AlamarBlue, LIVE stain, and trypan blue exclusion cell counting assays to determine cell number. Using this experimental optimization algorithm, we were able to design media with 181% more cells than a common commercial variant with a similar economic cost, while doing so in 38% fewer experiments than an efficient design-of-experiments method. The optimal medium generalized well to long-term growth up to four passages of C2C12 cells, indicating the multi-information source assay improved measurement robustness relative to rapid growth assays alone.

4.1. Introduction

Every bioprocess in which cells are the final product or used in the production process requires suitable culture conditions for cell growth and product quality. In the rapidly growing cellular agriculture / cultivated meat industry, where cells are grown for consumption to replace carbon-intensive and often unethical animal agriculture, cost-effective media has been identified as the most critical aspect in scale-up and commercialization [**59**]. Optimizing these conditions is difficult due to a large number of media components with nonlinear and interacting effects between cells, medium, matrix material, and reactor environment [**14**]. Typically, culture media used for processes in cellular agriculture consist of a basal medium of glucose, amino acids, vitamins, and salts (such as the common Dulbecco's Modified Eagle Medium [DMEM]) supplemented with fetal bovine serum (FBS) for improved cell survival. FBS is an undefined, animal-derived serum consisting of proteins, hormones, and other large molecular weight components, and contributes substantially to the cost of media [**84**]. Even when enriched with additional growth factors or FBS, media is often far from optimal for all cell types and requires adaptation and/or optimization [**50**], which is difficult for media mixtures with >30 components, as is common in cell culture.

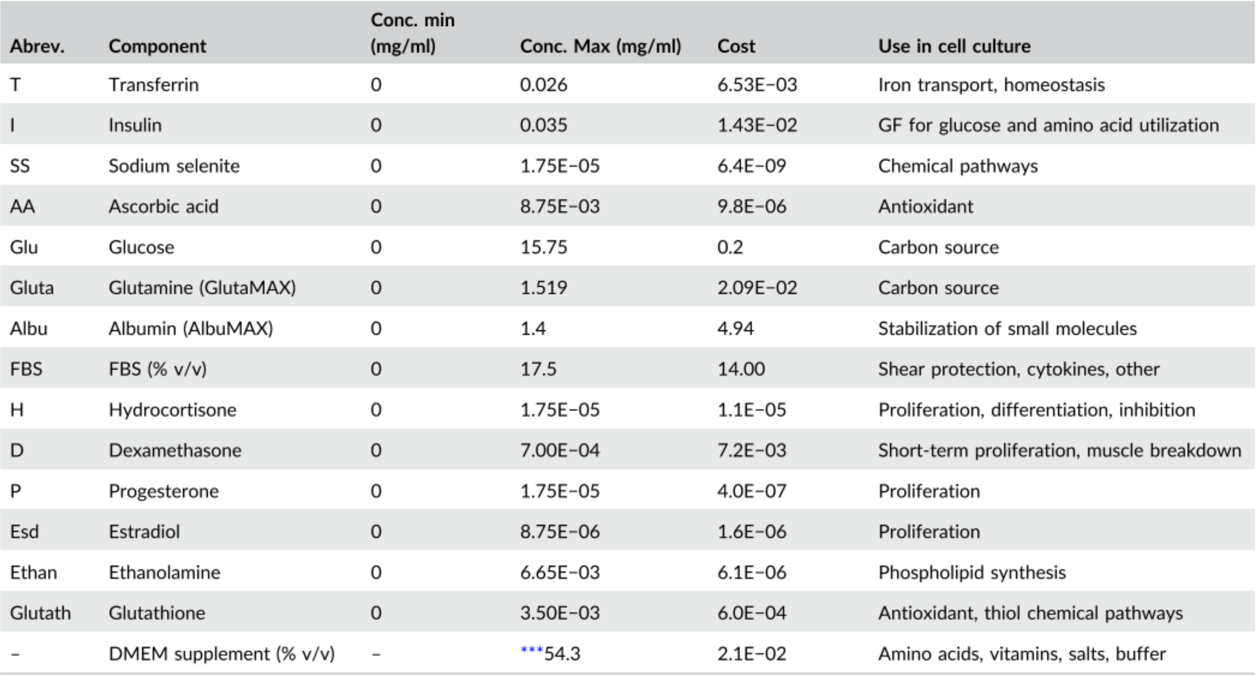
To manage this complexity, design-of-experiments (DOE) meth- ods are often employed in which factors (concentrations or environmental conditions) are set to a user-specified value (usually "high" or "low") and outputs are measured [**74**, **94**]. These DOE designs are arranged in such a way that statistically meaningful correlations can be found in fewer experiments than techniques like intuition, "one-factor-at-a-time" sequences, or random designs. A more advanced form of this is to use sequential, model-based DOEs such as a radial basis function [**21**, **93**, **95**] or Gaussian Process (GP) [**51**], combined with an optimizer/sampling policy, to automatically select sequences of optimal designs. These approaches are often more efficient than traditional DOE at optimizing systems using fewer experiments [**22**] and allow for more natural incorporation of process priors [**19**], measurement noise [**2**,**83**], probabilistic output constraints and constraint learning [**54**], multiobjective [**11**], multipoint [**89**], and multi-information source designs [**42**, **73**, **82**].

Even with these methods available, limitations still exist. In previous work, we applied a machine learning approach to optimize complex media design spaces but had limited success due to the difficulty in measuring cell number for multi-passage growth [**22**]. Therefore, in this study, we utilized a multi-information source (IS) Bayesian model to fuse "cheap" measures of cell biomass (rapid chemical assays which can be done at scale) with more "expensive" but higher quality measurements (cell numbers over time which represents a high-quality metric of growth media quality) to predict long-term medium performance. We refer to the simpler and cheap assays as "low-fidelity" IS, and more complex and expensive assays as "high-fidelity" IS. While not always predictive of long-term growth, these lower fidelity assays are at least correlated with cell health and can help in identifying interesting regions of the design space for further study with the high-fidelity IS. We used this model, with Bayesian optimization (BO) tools, to optimize a cell culture medium with 14 components while minimizing the number of experiments, optimally allocating laboratory resources, and building process knowledge to improve our optimization scheme and model. In Section 4.2 we discuss the computational and experimental components of this BO method. In Section 4.3 we present the results of the BO method in comparison to a traditional DOE method, followed by Section 4.4 where we demonstrate the importance off using multiple sources of information to obtain relevant process knowledge and/or optimization results.

4.2. Methods

4.2.1. Cells and Media Components. The system under consideration was the proliferation of C2C12 (ATCC) cells. These cells are immortalized muscle cells with similar metabolism and growth characteristics as other adherent cell lines useful in the cellular agriculture industry. Cells were stored in 70% DMEM (Gibco), 20% FBS (BioWest), 10% dimethylsulfoxide (Thermo Fischer) freeze medium at -196°C until thawed. Vials were thawed to 25°C and the freezing medium was removed by centrifugation at 1500 × g for 5 min. The centrifuged cell pellet was resuspended in 17 mL of DMEM with 10% FBS and placed on 15 cm sterile plastic tissue culture dishes (at about 10^6 cells/plate) (Cellstar, Greiner Bio-One). Cells were incubated in a 37°C and 5% CO_2 environment. After 24 h the medium was removed, the culture dish washed with Phosphate Buffer Solution (PBS) (Gibco), and fresh DMEM with 10% FBS was introduced. After an additional 24 h, cells were harvested using tripLE solution (Gibco), diluted in PBS, and counted using Countess II with trypan blue exclusion and disposable slides (Invitrogen). The process of removing cells from a plate, counting, and re-plating them with fresh medium is called subculturing or passaging. How well the C2C12 cells survive and grow after passaging is indicative of their long-term potential in a large cellular agriculture process.

The design space was comprised of the components and minimum/maximum concentrations listed in Table 4.1. These components were chosen because they are often used to supplement standard DMEM to improve cell growth; this represents a reasonable test case for the industrial application of these multi-IS BO methods to the cellular agricultural industry. The composition of standard DMEM (such as the medium used above), is shown in Table 4.3, and should not be confused with the base DMEM "supplement" (Gibco), which contains only amino acids, trace metals, salts, and vitamins and none of the other 14 components. pH and osmolarity are not controlled in this study, so act as latent variables.

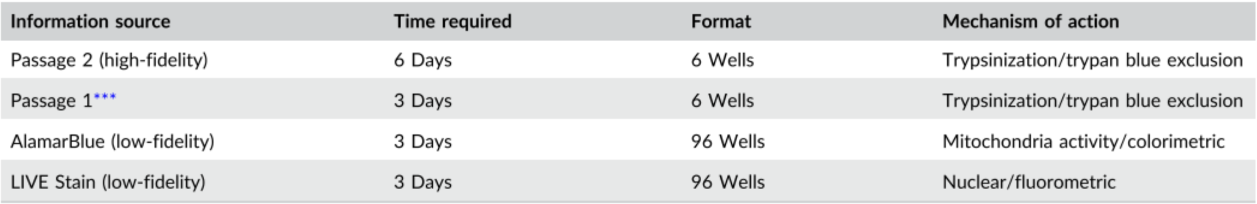


Abrev.	Component	Conc. min (mg/ml)	Conc. Max (mg/ml)	Cost	Use in cell culture
T	Transferrin	0	0.026	6.53E−03	Iron transport, homeostasis
I	Insulin	0	0.035	1.43E−02	GF for glucose and amino acid utilization
SS	Sodium selenite	0	1.75E−05	6.4E−09	Chemical pathways
AA	Ascorbic acid	0	8.75E−03	9.8E−06	Antioxidant
Glu	Glucose	0	15.75	0.2	Carbon source
Gluta	Glutamine (GlutaMAX)	0	1.519	2.09E−02	Carbon source
Albu	Albumin (AlbuMAX)	0	1.4	4.94	Stabilization of small molecules
FBS	FBS (% v/v)	0	17.5	14.00	Shear protection, cytokines, other
H	Hydrocortisone	0	1.75E−05	1.1E−05	Proliferation, differentiation, inhibition
D	Dexamethasone	0	7.00E−04	7.2E−03	Short-term proliferation, muscle breakdown
P	Progesterone	0	1.75E−05	4.0E−07	Proliferation
Esd	Estradiol	0	8.75E−06	1.6E−06	Proliferation
Ethan	Ethanolamine	0	6.65E−03	6.1E−06	Phospholipid synthesis
Glutath	Glutathione	0	3.50E−03	6.0E−04	Antioxidant, thiol chemical pathways
–	DMEM supplement (% v/v)	–	***54.3	2.1E−02	Amino acids, vitamins, salts, buffer

Table 4.1 Note: All components are shown were stored as per manufacturers (PeproTech unless specified) instructions in stock solutions. The concentration (mg/mL) of all media was between the minimum and maximum listed. The cost shown is a unitless scalarization of the relative economic cost of each component. Abbreviations: DMEM, Dulbecco's modified Eagle's medium; FBS, fetal bovine serum. ***All media have a 54.3% v/v (volume percent) base of DMEM supplement (liquid form, no glucose, glutamine, or FBS). Remaining volume (minus component volumes) was made up in water.

4.2.2. Cell Growth Experiments and Assays. For the high-fidelity IS, 750 μL of cell suspension containing 60,000 cells were placed in a six well plate (three replicates) with 2.25 mL of the test medium. For low-fidelity IS, 25 μL of cell suspension containing 2000 cells were placed in 96 well plates (four replicates and two control wells without cells) with 75 μL of the test medium.

All experiments thus had 6250 cells/cm^2 and 312.5 ml/cm^2 of media. After 72 h, all wells were measured using the IS methods shown in Table 4.2.



Information source	Time required	Format	Mechanism of action
Passage 2 (high-fidelity)	6 Days	6 Wells	Trypsinization/trypan blue exclusion
Passage 1***	3 Days	6 Wells	Trypsinization/trypan blue exclusion
AlamarBlue (low-fidelity)	3 Days	96 Wells	Mitochondria activity/colorimetric
LIVE Stain (low-fidelity)	3 Days	96 Wells	Nuclear/fluorometric

Table 4.2 Note: These sources of information (IS) were used to approximate and model C2C12 cell number. In this study, Passage 2 cell numbers were considered the highest-fidelity IS, while AlamarBlue and the LIVE stains were the lowest. ***Passage 1 cell number measurements were necessary to get Passage 2, so were included as a separate IS. Every high-fidelity IS measurement of a medium was also made in parallel with a low-fidelity measurement. Their inter-IS correlations are shown in Figure 4.8c.

The AlamarBlue assay required staining wells with 10% v/v (10 μL) AlamarBlue stock solution (Invitrogen), 4 h of incubation in a 37°C and 5% CO$_2$ incubator, and measurement of 570 nm λ_{570} and 600 nm λ_{600} absorbance wavelengths (Molecular Devices, ImageXpress Pico) as well as the control wells $\lambda_{570,c}$, and $\lambda_{600,c}$, (no cells) to get AlamarBlue reduction metric $AB\%$.

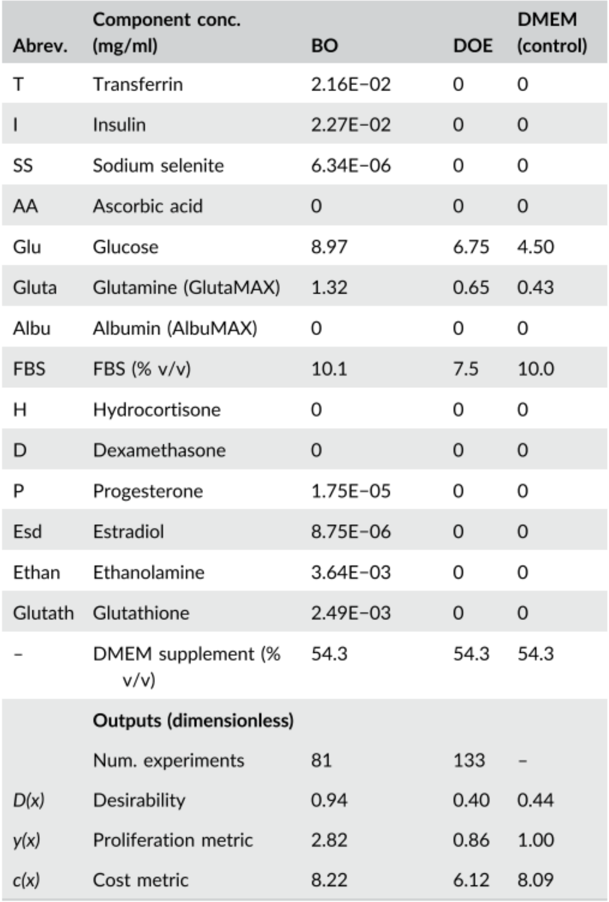
$$\%AB = \frac{117216\lambda_{570} - 80586\lambda_{600}}{155677\lambda_{600,c} - 14625\lambda_{570,c}}$$

The LIVE assay required that the test wells be washed with PBS, and 100 μL of 1 μL LIVE stain Calcein AM (Biotium) be introduced into the test wells and incubated for 1.5 h at 37°C and 5% CO$_2$. The biomass/cell number correlates was then measured using a fluorometer (Molecular Devices, ImageXpress Pico) at Ex/Em 494/530 fluorescein filters and calculated using the emission F_{530} . Both LIVE and AB% metrics are correlated with cell number and thus were the low-fidelity IS metric of cell number.

$$LIVE = F_{530}$$

We also measured the cell number using an automatic cell counter (Countess II) with trypan blue exclusion. This required trypsinization outlined in the previous section. Because we wished to measure long-term cell viability, after the first cell count (Passage 1), we re-seeded the cells under the same conditions and measured the cell count after an additional 72 h (Passage 2). The Passage

2 metric incorporated long-term viability and the effect of trypsinization, and thus was the most robust measurement of cell number. All measurements/correlates of cell number (AB%, LIVE, cell number) were reported as ratios, normalized to the DMEM control (whose concentration is shown in Table 4.3).



Abrev.	Component conc. (mg/ml)	BO	DOE	DMEM (control)
T	Transferrin	2.16E−02	0	0
I	Insulin	2.27E−02	0	0
SS	Sodium selenite	6.34E−06	0	0
AA	Ascorbic acid	0	0	0
Glu	Glucose	8.97	6.75	4.50
Gluta	Glutamine (GlutaMAX)	1.32	0.65	0.43
Albu	Albumin (AlbuMAX)	0	0	0
FBS	FBS (% v/v)	10.1	7.5	10.0
H	Hydrocortisone	0	0	0
D	Dexamethasone	0	0	0
P	Progesterone	1.75E−05	0	0
Esd	Estradiol	8.75E−06	0	0
Ethan	Ethanolamine	3.64E−03	0	0
Glutath	Glutathione	2.49E−03	0	0
–	DMEM supplement (% v/v)	54.3	54.3	54.3
	Outputs (dimensionless)			
	Num. experiments	81	133	–
D(x)	Desirability	0.94	0.40	0.44
y(x)	Proliferation metric	2.82	0.86	1.00
c(x)	Cost metric	8.22	6.12	8.09

Table 4.3 Note: Concentrations (mg/mL) of best BO-designed medium alongside that found by DOE and the DMEM control used throughout this study. The resulting objective function $D(x)$, cell number $y(x)$, and cost $c(x)$ of each medium are shown with the required number of experiments to get the optimal result. Abbreviations: BO, Bayesian optimization; DMEM, Dulbecco's modified Eagle's medium; DOE, design-of-experiments; FBS, fetal bovine serum.

4.2.3. DOE Method. To compare our BO method to a typical method used in the optimization of fermentation / bioprocess systems [**74**, **93**], we used a DOE. We first screened all 14 components using a folded Plackett-Burman (PB) design (which is a normal PB combined with an additional PB design with “high” factors set to “low” and vice versa) using the AlamarBlue IS. A linear model $D(x) = \beta_0 + \Sigma_{i=1}^{p} x_i \beta_i$ let us quantify the desirability of each component using the slope β_i (desirability $D(x)$ will be talked about in Section 4.2.4.2). The lower bound of the PB was set to $x = 0$ (0 mg/mL) and the upper bound $x = 0.28$ (28% of maximum concentration shown in Table 4.1, for example, glucose would be set to 0.28×15.75 mg/mL $= 4.41$ mg/mL) so that component quality could first be judged at modest concentration where nonlinear effects would be minimal. We then set unimportant or harmful ($\beta_i \leq 0$) components to $x = 0$ for the rest of the DOE study, then ran a Box-Behnken (BB) design over the remaining useful components using AlamarBlue IS. The BB was used to estimate an interaction- polynomial model $D(x) = \beta_0 + \Sigma_{i=1}^{p} x_i \beta_i + \Sigma_{1<i<j} x_i x_j \beta_{i,j}$, and a multi-start Newton's Method was used to find the $D(x)$-optimal concentrations inside the design space. If the optimal concentrations were found to be on any edge of the current BB, then the bounds of the design were shifted $\Delta x = 0.145$ dimensionless units in that direction (steepest accent) and another BB was run using these new bounds. This was done because the optimal boundary of the design space is uncertain and needed to be found. The sequential BB was run until the optimal bounds were found or resources exhausted. The best medium was then reported as the optimal point found using multi-start Newton's Method within the final optimal bounds.

4.2.4. BO Method. In standard BO, a function g is modeled using a Gaussian Process (GP) [**61**], characterized by a prior mean μ_0 and covariance Σ, with the property that for any X finite collection of N points with dimensionality p, the prior distribution of the output $g(X)$ is normal $g(x) \sim N(\mu_0, \Sigma)$. The prior determines the directionality and "wigglyness" of the function through the covariance kernel function Σ, which models the relationship between any two points x and x'. We chose the squared exponential function for the kernel to encode the belief that (i) similar experiments are more alike than dissimilar experiments governed by hyperparameters σ_f^2 and $\lambda_{1...p}^2$ and (ii) that the overall biological process underlying the response surface is smooth with (iii) each component response governed by λ_k^2, allowing each component k to have different degrees of "wigglyness".

$$\Sigma(x, x') = \sigma_f^2 \times exp(-1/2 \sum_{k=1}^{p} \frac{(x_k - x'_k)^2}{\lambda_k^2}) \tag{4.1}$$

If we collect N observations of inputs $X_N = [x_1 \dots x_N]$ and outputs $Y_N = [y_1 \dots y_N]$ from the generative process $y(x) = g(x) + \epsilon$ we can get the posterior distribution $g(x)|X_N, Y_N \sim N(\mu(X_N), \Sigma(X_N, X_N))$ where the mean and variance of $g(x)$ are given by Equations 4.2 and 4.3 respectively for homoscedastic noise $\Sigma_\epsilon = \sigma_\epsilon^2 \times I$ with process noise variance σ_ϵ^2.

$$\mu(x) = \mu_0 + \Sigma(X_N, x)(\Sigma(X_N, X_N))^{-1}(Y_N - \mu_0) \tag{4.2}$$

$$\sigma^2(x) = \Sigma(x, x) - \Sigma(X_N, x)(\Sigma(X_N, X_N))^{-1}\Sigma(X_N, x)^T \tag{4.3}$$

A more detailed discussion of GP models can be found in [**70**]. With a predictive model of the mean $\mu(x)$ and variance $\sigma^2(x)$ of cell number, we can use past experimental data inputs X_N and outputs Y_N to inform future process optimization.

The key objective of this study was to maximize C2C12 cell growth/accumulation while minimizing media costs. To do this, measuring cell growth was critical but experimentally expensive. Less expensive assays can approximate cell growth, yet with reduced accuracy. Therefore, it was beneficial to use combinations of cell growth assays to facilitate experimentation while decreasing the overall experimental burden. This provides a balance between quality of information and experimental cost. To this end we adopted the multi-information source GP model introduced by [**61**], which utilizes auxiliary information sources to model an underlying “true” function. We chose this model over the more typical multi-task GP to encode the prior belief that the generative model includes an underlying “true” function and several biased/ variable but correlated auxiliary functions, and to provide the flexibility of allowing different length-scale hyperparameters λ_k for each IS to be learned from the data.

Let us assume a generative model $y = g(x) + \delta(x) + \epsilon$ for a given media combination x at an IS indexed by m. We, therefore, have one independent GP for the underlying function $g(x)$ and one for each auxiliary IS deviation function $\delta(x, m)$ for the mth auxiliary IS (where $m = 0$ references the underlying IS). To implement this, Equation 4.1 is modified by adding an additional kernel (squared exponential) to the original kernel anytime an auxiliary datapoint $m \neq 0$ is referenced and x and x' have the same IS index using an indicator function $1_{m\neq 0}1_{m=l}$.

$$\Sigma(x_m, x'_l) = \Sigma_0(x, x') + 1_{m\neq 0}1_{m=l}\Sigma_m(x, x') \tag{4.4}$$

Further details about the noise model of the GP, training, and using prior information can be found in Appendix B.1. In addition to information on cell numbers, however, we wish to incorporate information about the process cost of x. Therefore, we formulate a cost function $c(x) = c_{min} + \Sigma_{j=1}^{p} c_j x_j$ where c_j is a scaled marginal cost of each media component whose coefficients can be found in Table 4.1.

4.2.4.1. *BO Acquisition Function.* To maximize media utility, we wish to maximize $y(x)$ while minimizing cost $c(x)$ for the highest-fidelity IS. Therefore, we posed this multi-objective optimization problem as a single-objective in the form of a desirability function $D(x)$ [**1**] where cell number and cost are scaled as $\bar{y} = \frac{y(x)-y_L}{y_H-y_L}$ and $\bar{c} = \frac{c(x)-c_H}{c_L-c_H}$ respectively.

$$D(x) = \phi(x)\sqrt{\bar{y}(x)\bar{c}(x)} \tag{4.5}$$

where $\phi(x) = 1_{y(x)\geq y_L}$ is a feasibility indicator function that is non-zero when the predicted $y(x)$ is greater than or equal to some minimum cell number metric y_L. We set $y_L = 0.5$ and $y_H = 2.0$ to exclude media that fail to be 50% as proliferative as the control media to preferentially select high-performance media. We scale $c(x)$ as a "smaller-the-better" metric where $c_L = c_{min} + \Sigma_{j=1}^{p} c_j$ so that we may solve our new cost-aware objective function as a single- objective problem $x^* = argmaxD(x)$.

With a predictive multi-IS GP modeling $\mu(x) \approx y(x)$ and computing $D(x)$ from Equation 4.5, we can use it to suggest optimal media conditions x^*. However, because we would like to solve for some optimal group of $q > 1$ experiments X^* rather than a single $q = 1$ experiment x^* (it is much more efficient to run multiple experiments at a time), we pose the optimization problem as a $p \times q$-dimensional multi-point optimization problem $X^* = argmaxD(X)$ for multiple optimal media conditions at once. This formulation (i) does not consider uncertainty when quantifying the value of a particular set of media components and (ii) does not have an analytical form. We solved both problems by using the multi-point expected improvement function $\alpha(X)$ [**89**].

$$\alpha(X) = E\{(max\{D(X)\} - D^*(X_N))^+)\}$$

where $D^*(X_N)$ is the $D(X)$-optimal desirability of the N points collected and $max\{D(X)\}$ is the $D(x)$-optimal desirability of the q points X evaluated by $\alpha(X)/$ If $max\{D(X)\} - D^*(X_N) \leq 0$ (no improvement from evaluating X) then the "+" operator sets the improvement of the design to $\alpha(X) = 0$. Thus, with $\alpha(X)$ we can quantify the value of multiple points X rather than just a single point x. Evaluating α for any group of experiments X requires further mathematical treatment, which can be found in Appendix B.2.

4.2.4.2. *BO Algorithm.* The BO algorithm that designs optimal experiments is shown in Figure 4.1. After collecting some initial data, the multi-IS GP is trained and X^* found using multi-start L-BFGS-B for some q maximum allowable number of experiments (based on laboratory constraints). The L-BFGS-B optimizer was chosen because it performs well on high dimensional problems, can be ran with multiple restarts thus improving its global optimization capabilities, and has access to gradients and Hessian approximations thus reducing computational time. Because we want to optimize the high-fidelity IS (long-term growth as Passage 2) all calculations in the BO algorithm are done using the high-fidelity IS prediction. With X^* in hand, we now must find the optimal IS to sample. We start by defining the number of high-fidelity samples we are willing to measure $q_0 < q$, with the remaining $q - q_0$ being low-fidelity IS (Figure 4.2).

We can pose the IS-allocation problem as "which q_0 designs in X^* has the highest $\alpha(X)$ in combination"? This requires calculating $\alpha(X)$ for all combinations $\binom{q}{q_0}$ in X^*, and allocating the highest-fidelity budget to the dominant combination. The remaining $q-q_0$ experiments can be allocated to low-fidelity IS. New experiments are collected using the IS and component concentrations found, and the procedure looped until the process was optimized to satisfaction or resources are exhausted.

We started our BO method by initialization with 10 experiments according to Latin Hypercube designs similar to [**61**, **62**] (10 experiments being the approximate capacity of our laboratory at any given time). The algorithm allocates $q = 10$ experiments with $q_0 = 3$ high-fidelity IS and $q - q_0 = 7$ low-fidelity Is using the combinatorial heuristic described above for the optimal group X^*. This was

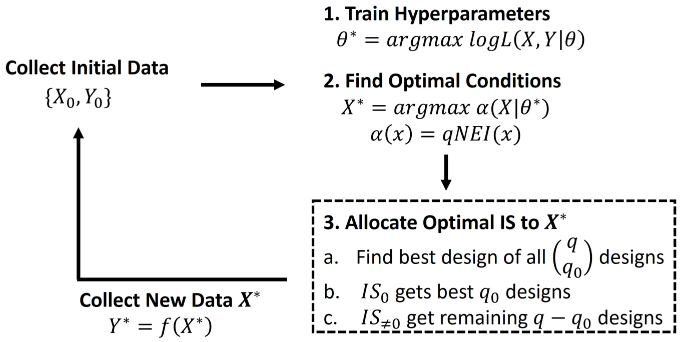

1. Train Hyperparameters
$\theta^* = argmax\ logL(X, Y|\theta)$
Collect Initial Data
$\{X_0, Y_0\}$
2. Find Optimal Conditions
$X^* = argmax\ \alpha(X|\theta^*)$
$\alpha(x) = qNEI(x)$
3. Allocate Optimal IS to X^*
a. Find best design of all $\binom{q}{q_0}$ designs
b. IS_0 gets best q_0 designs
c. $IS_{\neq 0}$ get remaining $q - q_0$ designs
Collect New Data X^*
$Y^* = f(X^*)$

Figure 4.1. BO Algorithm | This loop describes the Bayesian optimization algorithm to maximize some acquisition function $\alpha(X)$ for a process $Y = f(x)$ give q_0 high-fidelity and $q - q_0$ low-fidelity IS samples per batch of experiments. After each batch, the process is repeated until the process is optimized or resources are exhausted. BO = Bayesian optimization.

repeated seven times, with iterative training and optimization stages to improve our model while simultaneously finding optimal media. After 80 experiments (we stopped after exhausting our cell bank) a final high-fidelity IS experiment was performed at the theoretical optima $argmaxD(x)$ for 81 experiments total.

4.2.5. Computational Environment and Packages. Hardware used: Dell Precision 5820 Tower, Intel Xeon W-2145 DDR4-2666 Processor (3.7 GHz), 32 GB Memory. Software used: `python 3.9.7` (for all programming), `gpytorch 1.3.0`, `pytorch 1.8.1`, and `botorch 0.4.0` (for modeling and Bayesian optimization), `pydoe 0.3.8` (for initialization using Latin Hypercube experiments).

4.3. Results

4.3.1. Computational Validation of BO Method. Before optimizing our experimental system, we tested the BO algorithm on various multi-information source mathematical test functions $\{f_1, f_2, f_3, f_4\}$ (Appendix B.3) solving $argmaxf(x)$ using the noisy expected multi-point improvement acquisition function on a 10-dimensional problem. Each f had two low-fidelity test functions ($f_{bias,1}$ and $f_{bias,2}$) which differed substantially from the true test function. Given an extremely limited high-fidelity budget (10 simulations at two per iteration of the optimization loop), the multi-IS GP saw better average performance (higher outputs) compared to a regular GP with otherwise the same model architecture (hyperparameters, training method, priors, etc.). The major limitation of

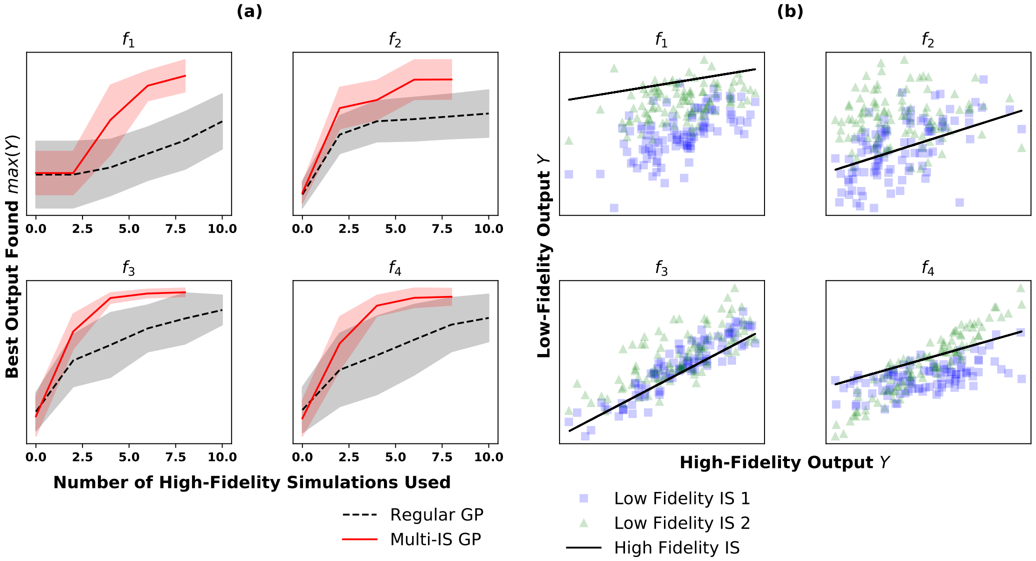


FIGURE 4.2. Simulation Results | Number of cumulative high-fidelity simulations used plotted against average (with standard deviation for five runs of the entire optimization loop) optimal output from f across five sequential iterations of the optimization framework. The multi-IS GP (solid) had access to $q = 15$ total simulations with $q = 20$ high-fidelity and $q - q0 = 13$ low-fidelity simulations per iteration (multi-IS GP has stopped one iteration early to reduce computational burden). The regular GP (dotted) only had access to the $q = 2$ high-fidelity simulations per iteration. Each test function $\{f_1, f_2, f_3, f_4\}$ had two biased low-fidelity versions whose correlations are described by plots (b). Squares and triangles represent a given $f_{bias,1}$ and f_{bias2} respectively. The solid line represents the underlying high-fidelity IS f. Hyperparameter and acquisition function optimization was done using multi-start L-BFGS-B implemented in botorch/scipy.

this experiment is that these test functions do not represent the true biological process. However, as the test functions were created to mimic noisy biological processes, we should be able to differentiate the performance of optimization methods using these results.

4.3.2. Experimental Validation of BO Method. We then applied our BO method to C2C12 media optimization design problem. The BO method achieved a maximum desirability of $D(x) = 0.94$ in 81 total experiments while the DOE only achieved a maximum at $D(x) = 0.40$ requiring 132 experiments. This represented a 132% improvement over the DOE and a 113% improvement over the control DMEM with 38% fewer experiments. The optimal BO medium corresponds to $y(x) = 2.82$ cell number with cost $c(x) = 8.22$, or a 227% improvement in cell number over DOE at a 34% increase in cost, and a 181% improvement in cell number over the DMEM control

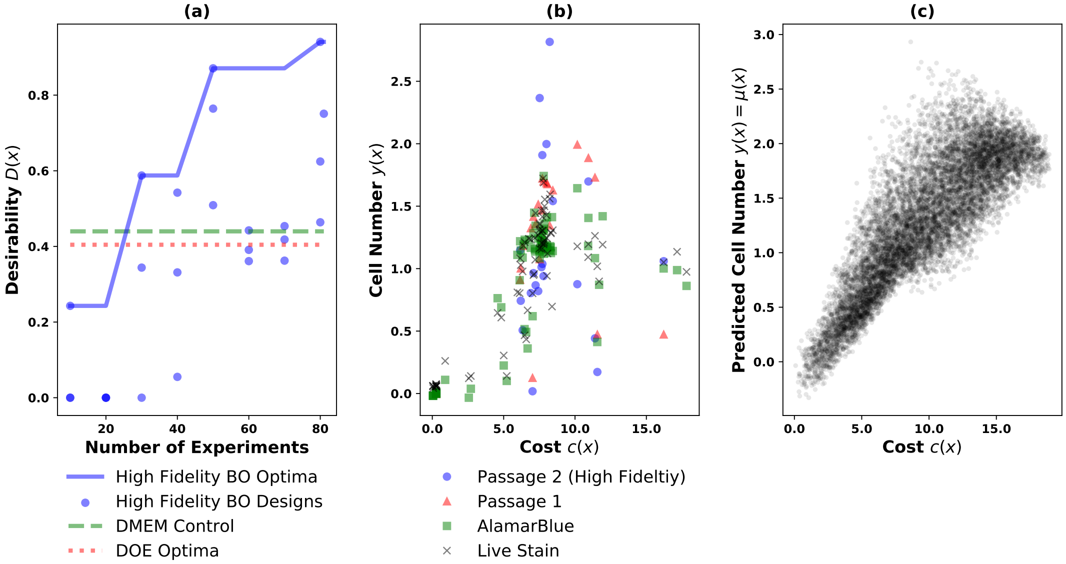


FIGURE 4.3. Learning Curve and Trade-Off Curve of BO Method | (a) Learning curve of $D(x)$ shows BO and DOE method designing experiments over the course of the optimization study. The line and dots represent the high-fidelity IS optima and designs, the dashed and dotted lines represent the DMEM control and DOE optima values for $D(x)$ respectively. Each IS experiment is shown in (b) the trade-off curve indicating a clear trade-off between cost $c(x)$ and cell number $y(x)$, where the dots, triangles, squares, and × represent Passage 2, Passage 1, AlamarBlue, and Live Stain respectively. (c) Simulated trade-off curve also shown for high-fidelity IS also showing a predicted parabolic relationship between competing objectives $y(x)$ and $c(x)$. BO, Bayesian optimization; DMEM, Dulbecco's modified Eagle's medium; DOE, design-of-experiments

at a mere 1.6% increase in cost. As seen in Figure 4.3a the BO method also found a sub-optimal medium, with higher $D(x)$ than DOE and the DMEM control, within 30 experiments, or a 77% reduction in experimental effort.

Table 4.3 shows the media concentrations resulting from the BO and DOE methods along with the DMEM control used throughout this study. The BO method found that transferrin, glutamine, progesterone, and estradiol should be at a high relative concentration. Ascorbic acid, hydrocortisone, and dexamethasone should be at a low / zero concentration. The remaining components should be somewhere in between the two extremes. The DOE method, using only AlamarBlue, used a PB screening design (32 experiments) to reduce the problem size from 14 components to four, finding that glucose, glutamine, albumin, and FBS had the highest positive effect on $D(x)$. Next, four sequential BB designs (25 experiments each), with bounds shifting in the direction of $D(x)$-steepest

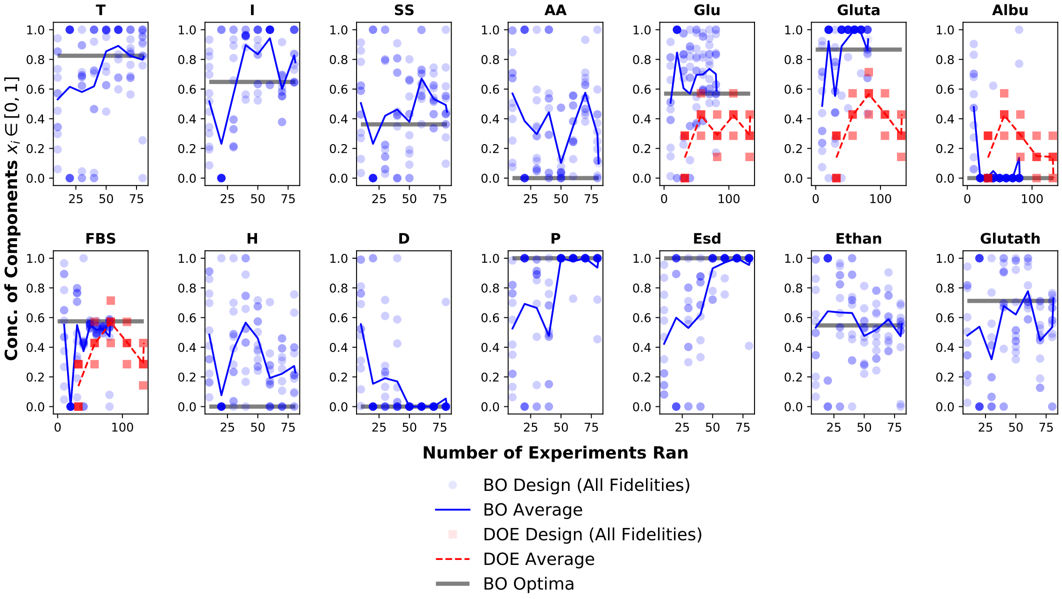


FIGURE 4.4. Learned Optimal Concentration | The conditions of each experiment (concentration ranges in Table 4.1) are shown plotted as a function of the cumulative number of experiments in the BO (circle) and DOE (box) study. The moving average (solid and dashed line for BO and DOE respectively) shows how each method searches for optimal concentrations. The horizontal line represents the final BO optimal concentration. BO, Bayesian optimization; DOE, design-of-experiments.

accent after each BB, used 100 experiments to find the optimal bounds of the four-dimensional factor space. Optimal factors were predicted to be nearly identical to the DMEM control, resulting in nearly identical desirability ($D(x) = 0.40$ vs 0.44 for DOE and DMEM control respectively).

As expected there was a trade-off between a number of cells $y(x)$ and medium cost $c(x)$ captured in Figure 4.3b and 4.3c. More nutrients, especially FBS, improved cell number at the expense of higher cost; this trend then breaks down as more FBS and Albumin have a deleterious effect on growth. We also note from Figure 4.4 that the BO algorithm found the optimal concentration of some components faster than others, as indicated by heavier clustering of data. This is a function of how confident the multi-IS GP was in certain regions of the design space, with denser sampling being indicative of higher confidence in improvement.

4.3.3. Experimental Validation of Long-Term Cell Number Objective Function. The robustness of the multi-IS GP model was evaluated by re-sampling the optimal BO medium which had a cell number metric of $y^* = 2.8 \pm 0.29$. When measured again the cell number metric was

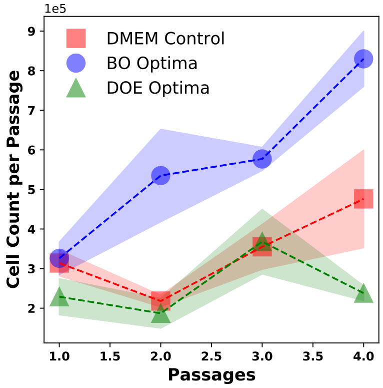


FIGURE 4.5. Long-Term Validation of Optima Media | The optimal BO-designed (dots), DOE (triangles), and DMEM control (squares) media performance up to Passage 4 Each passage was 72 h of growth at 37°C and 5% CO2. Trypsinization took place after each 72 h period to count cells and re-plate them to allow for further growth (standard deviations indicated). The BO method designed an optimal media with substantially improved long-term growth capacity than the DOE or DMEM control. BO, Bayesian optimization; DMEM, Dulbecco's modified Eagle's medium; DOE, design-of-experiments

$y^* = 2.7 \pm 0.93$, indicating measurement and overall system reproducibility. Next, all four optimal media were cultured for 288 total hours (to Passage 4 with 72 h/passage), to determine how well our high-fidelity IS generalized to longer-term growth. The optimal medium designed by the BO method outperformed the DOE and DMEM control substantially in a number of cells grown at Passage 4, with results summarized in Figure 4.5.

4.3.4. Sensitivity Analysis. We then examined the first and second-order effects of each component as predicted by the multi-IS GP (training on all N =81 datapoints). Most components show a parabolic effect in both $y(x)$ and $D(x)$ (Figure 4.6), where the optimal medium is in the middle of the factor space, often in sample dense regions.

To quantify the magnitude of the predicted global effect of each component, we employ the VARS method [**64**,**65**] of sensitivity analysis because standard methods of sensitivity analysis cannot capture the "importance" of a given factor in the presence of nonlinear effects. In VARS we defined $|N(h)|$ as the number of pairs in a set such that all possible pairs of points x^A and x^B are separated by a normalized factor distance h. We then integrated the variance $r = (y(X^A) - y(x^B))^2)$ of all pairs separated by h to get the variogram $\gamma_i = 1/2|N(h)| \sum_{(i,j) \in N(h_i)} r_{i,j}$. If we set $h = 0.1$ (10%

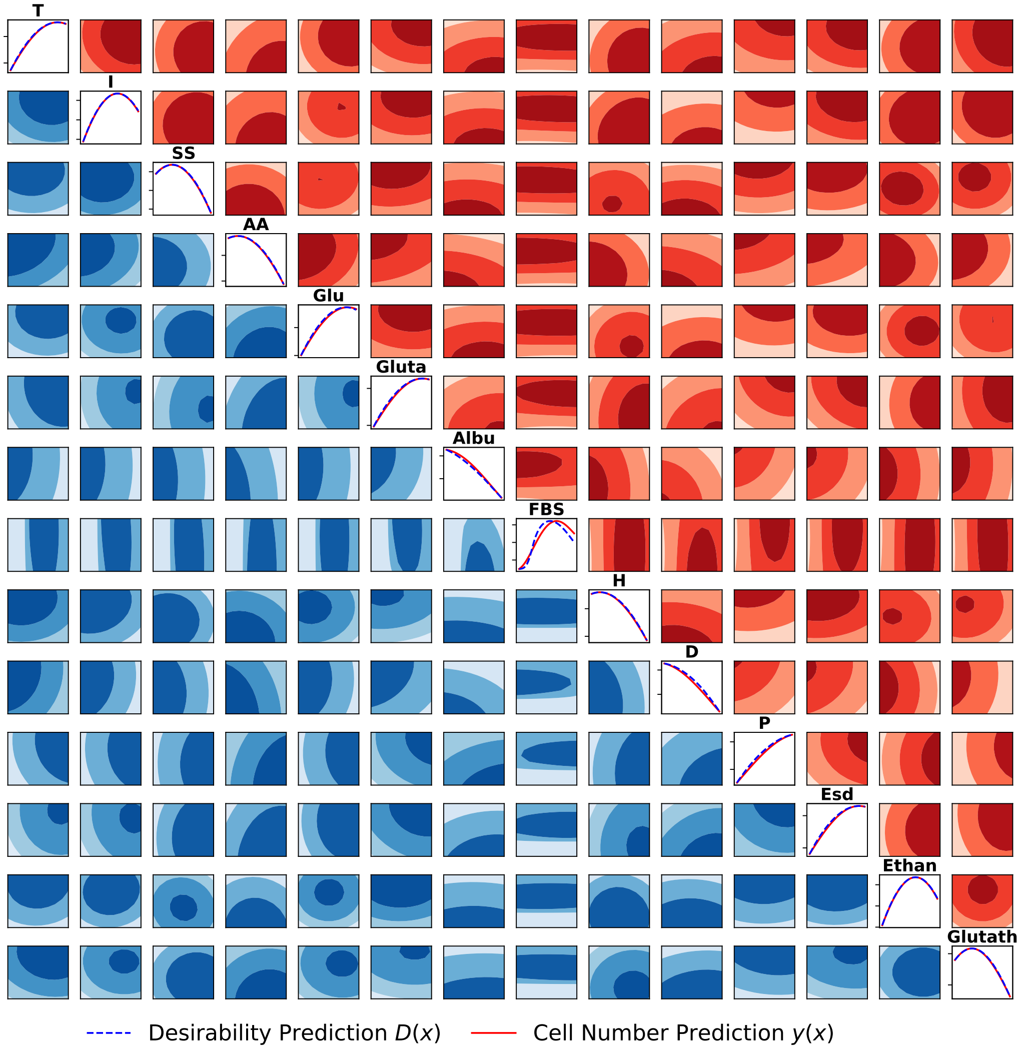


FIGURE 4.6. Predicted First and Second-Order Effects | First-order predicted effects of each component of the high-fidelity IS are shown on diagonal plots (y-axis is not to scale) with solid and dashed lines representing predicted cell number $y(x)$ and desirability $D(x)$ respectively. The “above” diagonal plots are second-order plots for cell number $y(x)$ and “below” are those for desirability $D(x)$. The range of all components as described in Table 4.1. Labels are left off for clarity; to find the axis labels read the x-axis labels horizontal from the diagonal label and read the y-axis labels vertical from the label.

of total normalized factor space for a given component) we are estimating a "local" variability in the output y whereas $h = 0.9$ would be an estimate of the "long-range" effect. Figure 4.7 shows these variograms γ_i for each component integrated to their "local", "medium", and "global" ranges,

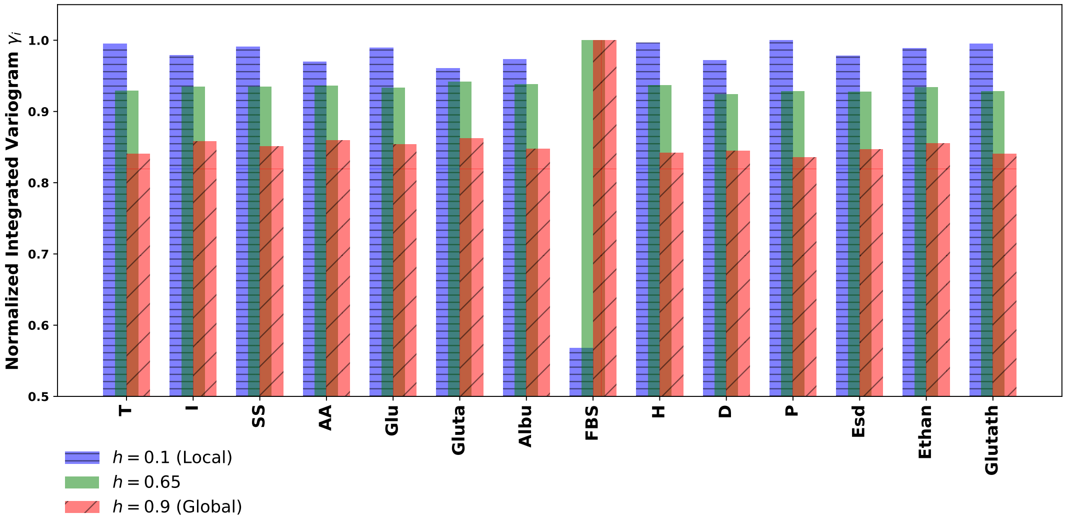


FIGURE 4.7. Variogram Sensitivity Analysis | The local (horizontal hatching), global (diagonal hatching), and mid-range sensitivity of each component on $D(x)$ is indicated by the height of the bars. Albumin, FBS, dexamethasone, and glutamine have the largest effect on $D(x)$, with FBS being by far the most critical component with respect to global sensitivity. Predicted variogram γ_i for each component was formed from $R = 300$ random samples from domain $[0, 1]$.

showing albumin, FBS, dexamethasone, and glutamine have the largest effect on $D(x)$, with FBS being by far the most critical component.

It was also useful to examine the correlations between different IS. The model predicts all IS to have very linear correlations (Figure 4.8c), while Passage 1, having the most experimental noise, had the weakest inter-IS correlations. Biases are predicted at the upper end of the output range as indicated by the deviation from the 45° line in Figure 4.8c. This fact is also evident in the predicted kernel matrix in Figure 4.8c, where the more error-prone Passage 1 data displays high off-diagonal intra-IS correlation, and the other IS shows nearly identical inter and intra-IS correlations.

4.4. Discussion

Production scale cellular agricultural processes will require > 10 passages of cell growth [**59**] so optimizing growth based on single-passage information is not adequate [**22**]. However, multi-passage growth assays are difficult / expensive to measure, and even more difficult to optimize when given many components. We managed this complexity by coupling long-term (i.e., > 1 passages) cell number measurements with simpler but less valuable rapid growth chemical assays

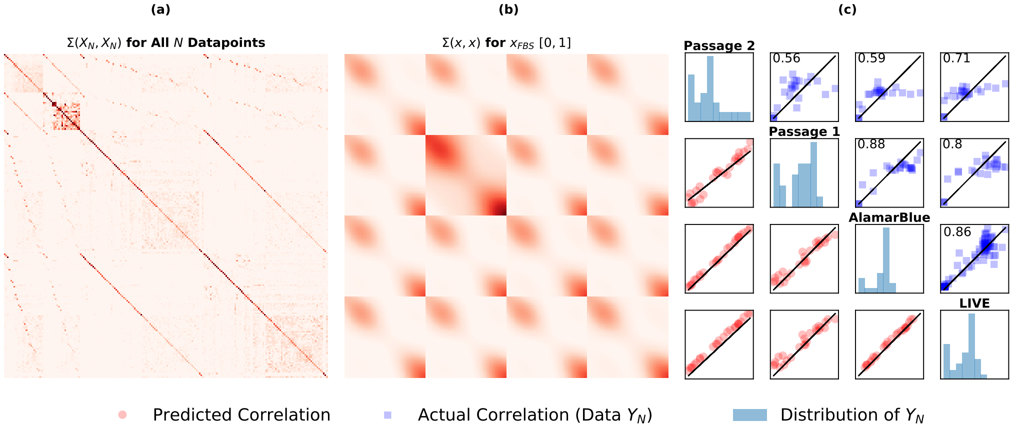


FIGURE 4.8. Kernel Plots and IS Distributions | (a) and (b) Show the output of the kernel $\Sigma(x_m, x_m')$ for all data collected $\{X_N, Y_N\}$ and a simulated data set where only x_{FBS} is varied from $[0, 1]$, respectively. Darker regions indicate large values of Σ, and thus a correlation. Also (c) the various IS cell number / correlate distributions (diagonal histograms) are shown. Above the diagonal (squares) are the actual inter-IS correlations for each IS with their respective R^2 values, and below the diagonal (circles) are the predicted inter-IS correlations for a random data set

(single passage) in murine C2C12 cultures as a model system for cellular agricultural applications, capturing a more holistic model of the process. We combined this with an optimization algorithm that efficiently allocates laboratory resources toward solving $argmax D(x)$ for desirability function $D(x)$, a function that incorporates both cell growth and medium cost. This resulted in a 38% reduction in experimental effort, relative to a comparable DOE method, to find a media 227% more proliferative than the DMEM control at nearly the same cost. As the longer-term passaging study suggests, our Passage 2 objective function and IS were well calibrated to mimicking the complex industrial process of growing large batches of cells over many passages, with Passage 4 cell numbers well predicted by this objective function.

The reasons for the success of the BO are myriad. The BO method iteratively refines a single process model to improve certainty in $D(x)$-optimal regions, whereas the DOE relies on a series of BB designs where the older data sets are ignored because they were outside of the optimal factor space. The BO also used a variety of IS, whereas the DOE only used a single low-fidelity AlamarBlue metric (as is common in analysis of growth media). Looking at Figure 4.8c, the AlamarBlue and

LIVE tended to cluster around the point y =1, making it difficult to distinguish between high-quality and low-quality media. This may be due to the deviation of linearity of the %AB and F_{530} metric at high biomass. The BO method also refined its multi-IS model over the entire feasible design space, allowing it to take advantage of optimal combinations and concentrations of all 14 components over the entire domain, whereas the DOE needed to reduce the design and factor spaces to reduce the number of experiments needed, and may have identified the wrong optimal boundary locations resulting in sub-optimal experimental designs. The BO method was also able to leverage information about process uncertainty to improve the model is poorly understood regions of the design space, whereas the steepest accent method used by the DOE chased after improved $D(x)$ with little regard for overall noise or experimental errors. This was worsened by the sensitivity of the polynomial model to random inter-batch fluctuations in AB%, which may have driven the DOE to sub-optimal media. Note that the success of our BO method should not be taken as generic superiority over all potential instantiations of DOE or commercial media used for C2C12 growth.

While the BO method worked well at solving the experimental optimization problem, the multi-IS GP accuracy was limited to highly sampled regions of the design space, thus limiting the efficacy of sensitivity analysis. This was a conscious decision made to trade off post-facto analysis for sampling media with high desirability $D(x)$. Accuracy was also limited by the low amount of data N available relative to the large dimensionality p, which is inherently the case in complex biological experiments where each batch of q experiments takes > 1 week to evaluate. Finally, the hyperparameters θ^* used in the multi-IS squared exponential kernel were deliberately regularized with prior distributions to smooth the posterior of the prediction $\mu(x)$. Regularization may have diminished the quality of the inter-IS correlations; the model hyperparameters ignored features where IS differed in favor of a simpler correlative structure to explain the data. This is seen in Figure 4.8b and 4.8c, where the kernel evaluations show nearly equal inter-IS correlative strength for most IS used. This may have "squished" / ignored features that could have provided additional information, but at the cost of sampling the design space too widely, again a deliberate choice of model skepticism towards outliers.

Even with these limitations, the BO method clearly performs well on media optimization systems relevant to cellular agriculture, that is, those with multiple and potentially conflicting information sources with varying levels of difficulty in measuring. The media resulting from the BO algorithm supported significantly more C2C12 cell growth with only a small increase in cost. This algorithm performs better than traditional DOE in this case, especially in incorporating critical data from growth after the multiple passages in an affordable manner. With these results, it should be possible to implement this type of experimental optimization algorithm in other systems of importance to cellular agriculture and cell culture production processes with difficult-to-measure output spaces, including for optimization of serum-free media for cell growth and for differentiation.

CHAPTER 5

Multi-Objective Bayesian Optimization of Serum-Free Culture Media for Cellular Agriculture

In this chapter we have extended the multi-information source Gaussian process modeling technique to solve a multi-objective Bayesian optimization problem involving the simultaneous minimization of cost and maximization of growth for serum-free C2C12 cells using a hypervolume improvement acquisition function. In 12 batches of experiments, collected using multiple assays targeting different cellular growth dynamics, we found a medium with a 184% improvement in growth over the control at a 71% increase in cost that maintained a high level of cell growth over five passages. In addition, the algorithm was able to design a variety of high-growth or low-cost alternatives, providing further evidence that sequential DOE techniques can quickly optimize difficult media design problems and provide options to researchers in cellular agriculture.

5.1. Introduction

In this work we applied the approach in the previous chapter to design a serum-free media, which is a necessary precondition to the development of cellular agriculture, for C2C12 cells. The work by [**52**] on Essential 8 (E8 or B8) media has been a good framework for serum-free formulations. They developed their medium for human induced pluripotent stem cell proliferation and stability based on the combination of the DMEM/F12 basal medium and supplementation with insulin, transferrin, FGF2, TGFβ1, ascorbic acid, and sodium selenite. [**80**] took this approach and, by screening multiple growth factors and hormones using a one-factor-at-a-time approach, developed an albumin-enriched B8 formula for the proliferation of bovine satellite cells. Recent work by [**50**] shows that merely seeding cells in serum-free media without additional preparation will be unsuccessful in optimizing serum-free media. A more robust approach is to slowly adapt a cell line to serum-free conditions over multiple passages [**85**]. Sometimes this requires attachment factors or extra-cellular matrix (ECM) material to allow adherent cells to affix themselves to the surface of the culture dish. For a fully animal component-free medium, ECM substitutes like Matrigel may be replaced by dilution cloning or other genetic techniques. The serum-free medium itself must contain the standard vitamins, trace elements, carbohydrates, amino acids, and salts discussed in the previous two chapters, but with additional proteins, enzymes, and growth factors that replace serum [**14**]. These components are particularly expensive and militate for a multi-objective approach to optimizing cell culture media.

The field of multi-objective optimization (MOO) is an extensive and valuable area of research that attempts to solve optimization problems with multiple, and often conflicting, objectives. The region of the design space where one cannot improve one objective without degrading another is the Pareto curve. Usually, there is no single point that dominates the entire design space and lies beyond the Pareto curve, so the MOO problem becomes a matter of finding sets of points X that fall on the curve, or designs that sufficiently represent the preferences of the designer. Cell culture media design, particularly for cellular agriculture, is inherently a MOO problem because improved growth is often found with expensive components. [**48**] used central composite designs to evaluate the effect of several components on a desirability function parameterization of lipid content, carbohydrate consumption and biomass accumulation. In work done to optimize cytokine dosing, [**29**] trained a regularized polynomial model and used a derivative-free optimizer to find the conditions that maximized a desirability function of cell populations. In work by [**34**], genetic algorithms VEGA and SPEA were used to maximize chemical conversion while maintaining biomass of the cyanobacteria organism. In a conference paper, [**87**] used a genetic algorithm MOGA to maximize plant culture biomass and minimize system cost.

There are many ways to solve MOO problems [**56**]. m Objective functions may be "scalarized" into a single objective function $\alpha(x)$ with the most common being a weighted-sum $\alpha(x) = \sum_i^m w_i f_i(x)$ where $\sum_i^m w_i = 1$ and $w_i > 0$ and preference is given to outcomes with higher w_i weights. The weighted exponential sum $\alpha(x) = \sum_i^m w_i f_i(x)_i^\gamma$ also exists where γ_i scales the ith objective. For example, if cell growth is more important than cell morphology then we'd set $w_{growth} > w_{morph}$. If we want the relative importance of cell growth to increase at higher responses, flattening the differences between low and modest growth, then we'd increase γ_{growth}. In lexicographic methods, the objectives are solved sequentially from most to least important objective, where the $i + 1$th objective must be optimized such that the ith objective is not degraded, thus turning the problem into a constrained optimization problem. Another method maximizes the function $\alpha(x) = max_i\{w_i(f_i(x) - f_{i,o})\}$, essentially allowing whichever objective's weighted improvement to dominate the design space when it is large. Genetic algorithms, such as the popular NSGA-II [**47**] have also been used to solve MOO problems. This algorithm works by balancing points that dominate others in the evaluated set (are better in one or more objectives y) with a crowding metric (averaged nearest-neighbor distance) to preserve diversity of solutions across the objective space. NSGA-II then ranks solutions based on their objective value, and break ties using the crowding metric. This can be particularly good for media optimization because we often want a variety of designs along the Pareto curve to consider additional factors like manufacturability and stability. NSGA-II's efficiency has allowed it to act as Pareto samplers for Bayesian approaches to multiple competing objectives [**12**], uncertainty and sensitivity analysis [**36**], and engineering optimization. The challenge with MOO is often in assigning weights such that the optimal designs are optimal from the point of view of the designer. Additionally, designs meant to optimize $\alpha(x)$ may not be distributed evenly across the objective space, giving the designer a distorted view of the objective space. There are also a variety of **multi-objective Bayesian optimization** (MOBO) approaches in the literature. The ParEGO [**38**] approach models each m objective as an independent Gaussian process (GP) with an objective function $\alpha(x) = max_j\{\lambda_j, f_j(x)\} + \rho \sum_i^m \lambda_i f_i$ with a uniformly drawn weight vector $\Lambda = [\lambda_1 \cdots \lambda_m]$. In this manner, the Pareto curve is gradually learned as Λ changes over batches of experiments. [**11**] propose a multi-objective max-value entropy search method. The information value of a given point x for finding the best value y^* is $\alpha(x) = H(y|X) - E_{Y^*}[H(y|X, Y^*]$. [**12**] then went on to generalize this to a multi-fidelity setting in which multiple sources of information could be fused. This is valuable in biological experiments where different assays may be used to measure the same outcome using different chemistry.

Because we have previously worked with GPs to successfully solve Bayesian optimization (BO) problems, we adopted the MOBO approach to solve the media design problem, specifically the noisy expected hypervolume improvement function described in [**26**]. We also used the multi-information source (IS) GP model described in [**61**] to successfully optimize cell culture media with multiple assays to robustly describe long-term cell proliferation [**23**]. We will use this multi-IS GP model again to model long-term cell growth in our serum-free system. In Section 5.2 we will discuss the laboratory materials needed to solve our media design problem, including the cells and chemicals needed, as well as the mathematical derivation of the acquisition function used to solve the MOBO problem. Then in Section 5.3 the results will be presented, followed by Section 5.4 with the discussion of the implications of the results.

5.2. Materials and Methods

5.2.1. Serum-Free Cells. To get the C2C12 cells (ATCC) to proliferate in serum-free conditions, they were first adapted to survive in the commercial Essential 8 (Gibco) (E8) medium by passaging the cells, starting in DMEM (Gibco) and 10% FBS (BioWest), in increasing amounts of E8. Once E8 comprised $> 90\%$ v/v of the medium, cell growth slowed and Matrigel (Corning) was needed to provide ECM. With the Matrigel, the new C2C12 line survived fully in E8. Next we used a dilution cloning technique to select a subset cell line from these cells that could survive without Matrigel. The surviving cells were frozen in Synth-a-Freeze (Gibco) at their fourth passage in -196°C liquid N_2 and are the cells used in the remainder of this chapter. Bovine satellite cells (BSC) were used for verification experiments after the optimization campaign was finished to determine the generalizability of the designed media. BSCs more closely resemble the phenotypes of cells desired in the cellular agriculture industry.

5.2.2. Media Components. The media design space was based on the E8 / B8 formulation [**52**] comprised of basal medium, FGF2, TGFβ1, insulin, transferrin, ascorbic acid, and sodium selenite. We chose to supplement this with nine growth factors which have either been found to improve cell proliferation in [**23**] or by expert opinion. Because the basal component is comprised of >30 individual components it was broken down into groups based on function in cell culture. These component groups (essential and non-essential amino acids, vitamins, salts, trace metals, DNA precursors, fatty acids) were varied during the optimization campaign by the algorithm which we discuss in later sections. Components believed to have significant effects on growth (carbohydrates, ascorbic acid, sodium selenite) were individually varied. NaCl was separated from the general salts group because it had a large effect osmolarity.

Abrev.	**Component**	**Conc. Min**	**Conc. Max**	**Cost (Unitless)**
NEAA	***NEAA	0.5x	5x	1.00E-08
EAA	***EAA	0.5x	5x	1.00E-08
V	***Vitamins	0.5x	5x	1.00E-08
Salt	***Salts	0.5x	5x	1.00E-08
Metal	***Trace Metal	0.5x	5x	1.00E-08
DNA	***DNA Precursor	0.5x	5x	1.00E-08
Fat	***Fatty Acid	0.5x	5x	1.00E-08
SS	Sodium Selenite	7.00E-06	7.00E-05	1.00E-08
AA	Ascorbic Acid	0.03	0.3	1.00E-08
Gluc	Glucose	1.35	13.5	1.00E-08
Gluta	Glutamine	0.22	2.2	1.00E-08
Pyruv	Sodium Pyruvate	0.03	0.3	1.00E-08
NaCl	Sodium Chloride	1.40	14.0	1.00E-08
I	Insulin	0.01	0.1	0.03
T	Transferrin	5.00E-03	0.05	0.004
FGF2	FGF2	3.00E-05	3.00E-04	0.63
TGFb1	TGFβ1	1.00E-06	1.00E-05	0.09
EGF	EGF	0	2.50E-05	0.003
P	Progesterone	0	2.50E-05	1.00E-08
Estra	Estradiol	0	1.25E-05	1.00E-08
IL-6	IL-6	0	6.25E-05	0.08
LIF	LIF	0	1.25E-05	0.02
TGFb3	TGFβ3	0	1.60E-05	0.04
HGF	HGF	0	2.50E-05	0.03
PDGF	PDGF	0	2.50E-05	0.03
PEDF	PEDF	0	2.50E-05	0.04

TABLE 5.1. Serum-Free Medium Design Space | all components were stored as per manufacturers instructions in stock solutions (PreproTech in the case of growth factors or Gibco for glutamine, EAA, and sodium pyruvate). The concentration (mg/mL) of all media was between the minimum and maximum listed. The cost is a unitless coefficient that corresponds to the marginal dollar-denominated cost on the $[0, 1]$ scale. Cell culture sterile water was used to make up the remaining volume not taken up by the components. Notes: ***The max/min concentration is relative to stock concentrations in Appendix C.1. All media have a sodium bicarbonate concentration of 2.44 mg/mL and were stored at 5°C for no longer than 8 days.

5.2.3. Cell Growth Experiments and Assays. We utilized a multi-information source (IS) Bayesian model to combine "cheap" measures of cell biomass (AlamarBlue and LIVE stain) with more “expensive” but higher quality measurements (cell count after 1 and 2 passages) to predict long-term medium performance. We refer to the simpler and cheap assays as “low-fidelity” IS, and more complex and expensive assays as “high-fidelity” IS. To start an experiment for all IS, vials were thawed to 25°C and the freezing medium was removed by centrifugation at 1500 × g for 4 min. The centrifuged cell pellet was resuspended in 17 mL of store-bought E8 (Gibco) and placed on 15 cm sterile plastic tissue culture dishes (Cellstar, Greiner Bio-One). Cells were incubated at 37°C and 5% CO_2 for 48 hrs. Cells were harvested using tripLE solution (Gibco), diluted in PBS, and counted using a Countess II with trypan blue exclusion and disposable slides (Invitrogen). With the known concentration of cells, 96 well plates (for the low-fidelity IS) were seeded at 2,000 cells / well (25 μL of PBS / cell inoculum and 75 μL of test medium) and 6 well plates (for the high-fidelity IS) were seeded at 60,000 cells / well (750 μL of PBS / cell inoculum and 2,250 μL of test medium). The final density of both formats should be roughly 20,000 cells / mL of PBS and medium. After 72 hrs, all wells were measured using the IS methods shown in Table 4.2 and described in Section 4.2.2. Very briefly, the low-fidelity IS AlamarBlue (Invitrogen) and LIVE (Biotium) assay required staining wells with a stock chemical and reading with absorbance and fluorescence on a plate reader (Molecular Devices, ImageXpress Pico). Both signals correlate with cell number. The other low-fidelity IS was the Passage 1 cell count using a Countess II automatic cell counter. The high-fidelity IS, which correlates much better with long-term cell proliferation, the Passage 2 metric, was also measured using the Countess II (but requires an additional 72 hrs of growth and another trypsinization step). This additional 72 hr period is why it is considered a long-term cell growth metrics, but also why it is more tedious to use to optimize a complex media.

5.2.4. MOBO Acquisition Function. We have chosen the hypervolume metric $HV(x)$ (Equation 5.1) to rank the quality of p media combinations based on their growth and cost. If the mth output to maximize is $f_m(x)$ relative to a minimum reference point l_m then $HV(x)$ is the product of $f_m(x)-l_m$ for each output [**25**]. The "+" operator in Equation 5.1 sets $HV(x)=0$ if $f_m(x)-l_m \leq 0$ (this acts as a threshold). To clarify the connection between the GP and the acquisition function, the mean $\mu(x)$ (Equation 5.2) and variance $\sigma^2(x)$ are shown below, where $f_m(x)=\mu(x)$.

$$HV(x) = \prod_{m=1}^{p} [f_m(x) - l_m]^+ \tag{5.1}$$

$$\mu(x) = \mu_0 + \Sigma(X_N, x)(\Sigma(X_N, X_N))^{-1}(Y_N - \mu_0) \tag{5.2}$$

$$\sigma^2(x) = \Sigma(x, x) - \Sigma(X_N, x)(\Sigma(X_N, X_N))^{-1}\Sigma(X_N, x)^T$$

To compute the "improvement" in Equation 5.1 we reformulate the above expression into the product of the minimum between a max-value called u_m, and $f_m(x)$ **[26]** where $z_m(x) = min\{u_m, f_m(x)\}$. As discussed in the cited Daulton papers, a box decomposition algorithm can be used to quickly compute $HVI(x)$ by breaking down the above computation into a piece-wise integration across K rectangles defined by vertices u_k and l_k. We numerically integrate over the rectangles to get the approximation of the hypervolume improvement function $HVI(x)$.

$$HVI(x) = \prod_{m=1}^{p}[z_m(x) - l_m]^+$$

$$HVI(x) \approx \sum_{k=1}^{K}\prod_{m=1}^{p}[z_{m,k}(x) - l_{m,k}]^+$$

Because we can run multiple experiments in a single batch, we can again reformulate $HVI(x)$ into the "multi-point" $qHVI(X)$ where we wish to predict the best q set of experiments X. This can be done using the *inclusion-exclusion* principle for overlapping sets. In practice, this means summing across q points $\sum_{j=1}^{q}(-1)^{j+1}$ and modifying the improvement calculation to incorporate all subsets of the proposed candidate pool X of size j for $j = 1 \cdots q$. This additional calculation prevents double counting of any q overlapping hypervolume sets. Note that $z_{m,k,X_j} = min\{u_k, f_m(X_{i,1}) \cdots f_m(X_{i,j})\}$. Finally, because we have a statistical model in the form of the GP, we formulate an "expected" $qHVI(X)$ as the integral over the posterior distribution over the previous formulation, or $qEHVI(X) = \frac{1}{N}\sum_{t=1}^{q} HVI(X)$ in the case of monte-carlo (MC) sampling of N points (MC is needed because there is no analytical solution to $qHVI(X)$). With sufficiently large N, the MC approximation of $\mu(x)$ should approach Equation 5.2.

$$qHVI(X) = \sum_{X_j \in \Omega}\sum_{j=1}^{q}\sum_{k=1}^{K}\prod_{m=1}^{p}(-1)^{j+1}[z_{m,k,X_j}(x) - l_{m,k}]^+$$

$$qEHVI(X) = \frac{1}{N}\sum_{t=1}^{N}\sum_{X_j \in \Omega}\sum_{j=1}^{q}\sum_{k=1}^{K}\prod_{m=1}^{p}(-1)^{j+1}[z_{m,k,X_j,t}(x) - l_{m,k}]^+$$

MC involves generating a fixed set of normal random numbers $Z \sim N(0, I_N)$ and sampling the GP using the "reparameterization trick" **[91]** where the prediction is sampled as $Y = \mu(X) + L(X)Z$ with Cholesky Decomposition of the covariance matrix $\Sigma(X, X) = L(X)L(X)^T$. Pushing these samples through $qEHVI(X)$ and $\nabla qEHVI(X)$ allows us to solve $X^* = argmax\ qEHVI(X)$ using the multi-start L-BFGS-B optimization algorithm. This is an optimizer that uses function evaluations $f(x)$ and gradients $\nabla f(x)$ to approximate the Hessian matrix of double derivatives $\nabla^2 f(x)$. This approximation speeds up solving $X^* = argmax\ f(x)$ and is commonly used in MOBO and machine learning methods. As we wish to constrain our experiments to achieve some minimum level of growth y_{min} so as not to waste experimental effort in regions of the design space that cannot support cells, we modify $qEHVI(X)$ by multiplying it by an indicator function $\phi(x) = \mathbf{1}\{\mu(x) \geq y_{min}\}$. Because each point q should contribute to the hypervolume proportional to the extent to which it satisfies the constraint, we arrive at the multi-point version of the constrained hypervolume function $\alpha(X)$ by averaging out $\phi(x)$ using the same MC samples. Note that $\phi(x)$ is not differentiable so we replaced it with a sigmoid function $\phi(x) \approx \frac{1}{1+exp(-v(x)/\epsilon)}$ with temperature parameter $\epsilon = 10^{-3}$. We finally arrive at Equation 5.3 which will be the acquisition function to be optimized throughout this work.

$$\alpha(X) = \frac{1}{N}\sum_{t=1}^{N}\sum_{X_j\in\Omega}\sum_{j=1}^{q}\sum_{k=1}^{K}\prod_{m=1}^{p}(-1)^{j+1}[([z_{m,k,X_j,t}(x) - l_{m,k}]^{+})\prod_{x'\in X_j}\phi(x)] \tag{5.3}$$

An example of $\alpha(x)$ and $\phi(x)$ is plotted in Figure 5.1 for glutamine and pyruvate (data collected in this work). The GP model was used to predict the cell growth of the glutamine-pyruvate design space, which was then used to predict $\alpha(x)$ and the feasibility score $\phi(x)$. Notice optimizing $\alpha(x)$ may not correspond to maximising the feasibility of the experiment. This is because (i) cost and (ii) uncertainty reduction is considered in $\alpha(x)$ and not $\phi(x)$.

5.2.5. MOBO Algorithm. The MOBO algorithm that designs optimal experiments is shown in Figure 5.2. After collecting some initial data from a variety of IS, the model was trained and X^* found using multi-start L-BFGS-B for some q maximum allowable number of experiments. Because we want to optimize the high-fidelity IS (Passage 2) all calculations in the MOBO algorithm are done using the high-fidelity IS prediction. With X^*, we now find the optimal IS to sample. We started by defining the number of high-fidelity samples we are willing to measure $q_0 < q$. $\alpha(X)$ was calculated using Equation 5.3 for all combinations $\binom{q}{q_0}$ in X^*, and the dominant combination of

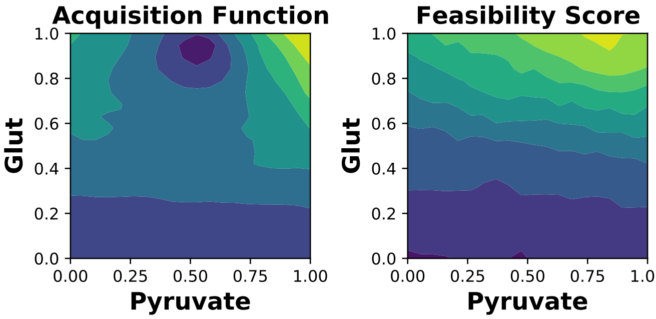


FIGURE 5.1. Plot of Acquisition Function and Feasibility Score for $q = 1$ Experiments | (left) expected hypervolume improvement $\alpha(x)$ and (right) mean feasibility score $\phi(x)$ for glutamine and pryuvate concentration (normalized to $[0, 1]$, which has no effect on the results because the models are trained using the normalized data) calculated using 1000 MC samples using the reparameterization trick. Light / yellow represents higher values. The final pytorch model (which was needed to generate these plots) and data availability is discussed in Appendix C.4. For all experiments, the lower bound for predicted growth $\mu(x)$ was $l_\mu = \bar{\mu} - 4\sigma$ (four standard deviations below the current mean cell growth metric across all assays. The lower bound for cost $l_c = -1.1$, or 10% above the highest possible value of cost (which, because of our unitless scalarization, is always $c_max = 1$.

experiments was allocated to the high-fidelity IS. The remaining $q - q_0$ experiments were allocated to low-fidelity IS. We started our MOBO algorithm in the serum-free experiments by initialization with 10 Latin Hypercube designs [**18**]. The algorithm then allocated $q = 15$ experiments with $q_0 = 3$ high-fidelity IS and $q - q_0 = 12$ low-fidelity IS using the combinatorial heuristic. This was repeated for 12 batches of experiments, where a batch is defined as a single group of q experiments designed by the MOBO algorithm. Because of the enormous time-cost of measuring biological replicates of q_0 cell counts for two passages individually, it was assumed that an averaged technical replicate would capture the underlying trends of the system. As the results will show, this did not appreciably detract from the quality optimal media found even over multiple passages. To further bias our experiments towards high growth regions of the design space, after 9 batches of experiments we ran an additional high-fidelity experiment solving $x^*_G = argmax\, NEI(x)$ as outlined in [**54**] where $NEI(x) = \frac{1}{N}\sum_{t=1}^{N}[max\{f_t(x)\} - max\{f(X)\}]^+$. This is equivalent to maximizing the expected improvement of a single experiment of a noisy function without consideration of cost.

5.2.6. Computational Environment and Packages. Hardware used: Dell Precision 5820 Tower, Intel Xeon W-2145 DDR4-2666 Processor (3.7 GHz), 32 GB Memory. Software used: `python`

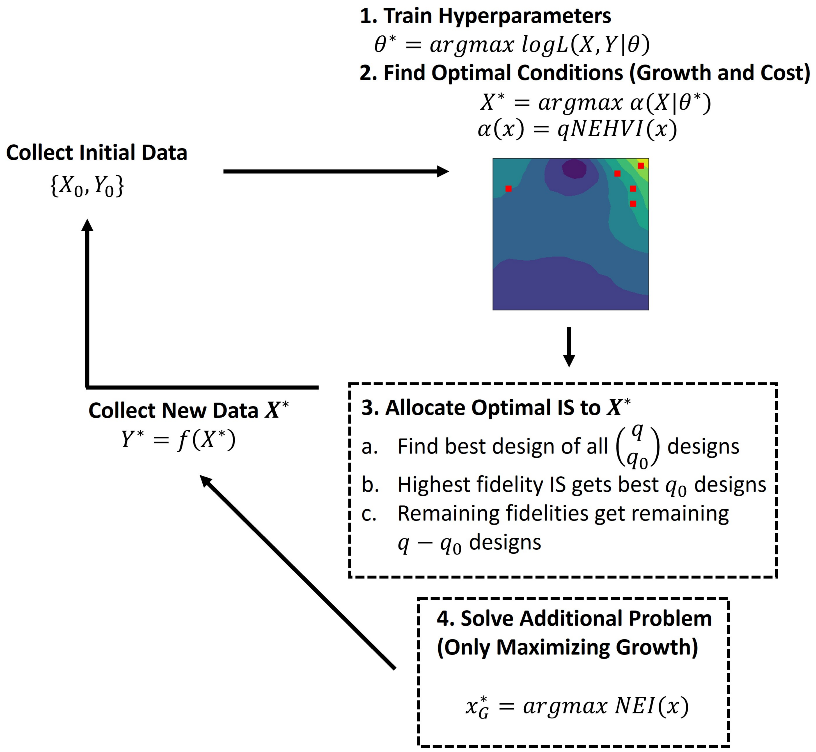
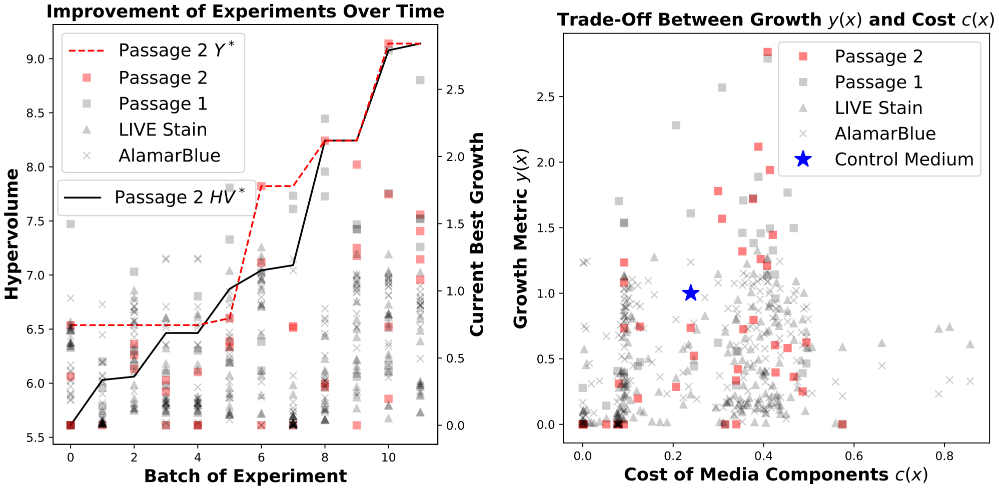

1. Train Hyperparameters
$\theta^* = argmax\ logL(X,Y|\theta)$
2. Find Optimal Conditions (Growth and Cost)
$X^* = argmax\ \alpha(X|\theta^*)$
$\alpha(x) = qNEHVI(x)$
Collect Initial Data
$\{X_0, Y_0\}$
Collect New Data X^*
$Y^* = f(X^*)$
3. Allocate Optimal IS to X^*
a. Find best design of all $\binom{q}{q_0}$ designs
b. Highest fidelity IS gets best q_0 designs
c. Remaining fidelities get remaining $q - q_0$ designs
4. Solve Additional Problem (Only Maximizing Growth)
$x_G^* = argmax\ NEI(x)$

FIGURE 5.2. MOBO Algorithm | this loop describes the MOBO algorithm to maximize $\alpha(X)$ that describes the value of a given set of experiments given q_0 high-fidelity and $q - q_0$ low-fidelity IS samples per batch of experiments. After each batch, the process is repeated until the process is optimized or resources are exhausted. Notes: To increase the presence of high growth conditions, after batch four and nine y_{min} was increased from 0.5 to 0.75 and 1.0 respectively. The minimum standardized variance was thresholded at $\sigma^2_{min} = 0.02$ but this needed to be changed to 0.05 to reduce numerical stability issues with optimizing $\alpha(X)$.

`3.9.7` (for all programming), `gpytorch 1.3.0`, `pytorch 1.8.1`, and `botorch 0.4.0` (for modeling and Bayesian optimization), `pydoe 0.3.8` (for initialization using Latin Hypercube experiments). For neural network test problem `scikit-learn 0.24.1` was used.

5.3. Results

5.3.1. Experimental Validation of MOBO Method. The MOBO algorithm was tested on the computational test problems introduced in the previous chapter with an additional linear cost function that turned the single-objective optimization problem into a MOO problem. The results are in Appendix C.2 and C.3. The new hypervolume acquisition function $\alpha(X)$ (Equation 5.3) performed similarly to the desirability function discussed in the previous chapter. Therefore, because the hypervolume function makes fewer assumptions about the MOO problem, it was chosen to help design experiments for this novel system. Furthermore, empirical studies of the hypervolume [**25**] and noisy-hypervolume [**26**] acquisition function indicate that it is superior to a wide variety of MOO and MOBO solvers on synthetic and data-based optimization problems. The most prominent result from the application of the MOBO algorithm to the serum-free experimental system was the steady improvement in both hypervolume and the Passage 2 high-fidelity IS growth metric in Figure 5.3b. Some of the interesting media designs are highlighted in Table 5.2. Only one medium (OM0) dominated the control medium in both growth and cost, resulting in 23% more growth at 62.5% of the cost of the control. OM0 had notably lower concentrations of major growth factors like insulin, transferrin, FGF2, and TGFβ1 and higher concentrations of progesterone, estradiol, IL6, and LIF. OM1 was another interesting medium that had 78% more growth at only and additional 25% cost. This was due to higher concentrations of the growth factors that OM0 lacked. Finally, OM2 and OM3 had a 112% and 184% improvement in growth at an increase in cost of 62% and 71% respectively. OM2 and OM3 had even higher concentrations of both the insulin, transferrin, FGF2, and TGFβ1 growth factors, while also elevating the concentration of all factors from progesterone to PEDF.



(a)

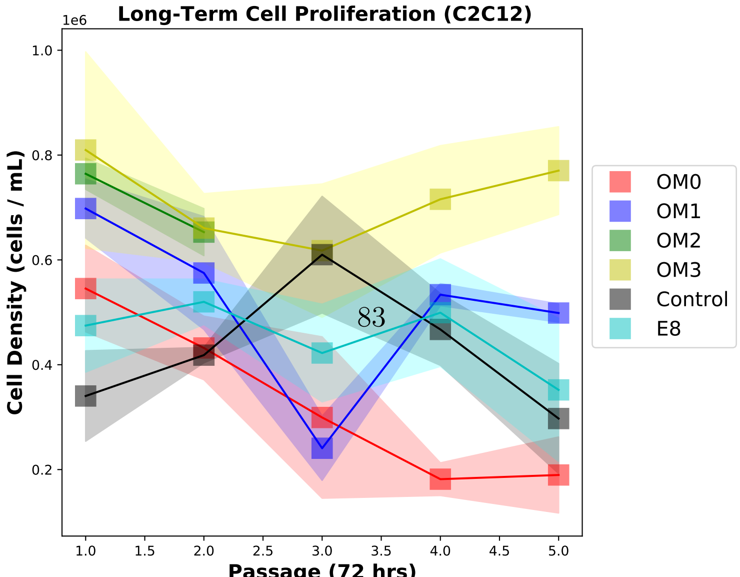
(b)

FIGURE 5.3. MOBO Application to Serum-Free System | plot (a) shows improvement in both Passage 2 growth and hypervolume over time (batches of experiments on the x-axis). The dotted line shows the best performing growth experiment per batch and units are on the right-hand axis. Plot (b) shows the trade-off between growth and cost from all data and IS types. The final pytorch model and data availability is discussed in Appendix C.4.

5.3.2. Long-Term Proliferation. We then tested OM0-OM3, the control medium, and an in-house E8 for five passages to assess the ability of our Passage 2 high-fidelity IS to mimic longer-term effects of the media on C2C12 cells. The cell density after each passage is shown in Figure 5.4a. OM3 was the only media that maintained growth up to five passages. Next, BSC cells were cultured on collagen matrix for three passages with OM0-OM2 media and the control medium. Only OM2 showed significant cell densities throughout the study.

	OM0	OM1	OM2	OM3	Control
NEAA* **	0.75x	1.50x	1.55x	1.20x	1.00x
EAA* **	0.90x	1.35x	1.40x	0.95x	1.00x
V* **	3.60x	4.30x	2.25x	1.60x	1.00x
Salt* **	0.50x	2.50x	1.75x	0.90x	1.00x
Metals* **	5.00x	4.80x	3.40x	3.55x	1.00x
DNA* **	1.95x	3.05x	2.95x	2.00x	1.00x
Fat* **	2.75x	1.25x	2.20x	2.65x	1.00x
SS	4.41E-05	4.83E-05	3.99E-05	4.41E-05	1.40E-05
AA	0.18	0.21	0.23	0.23	0.06
Gluc	1.62	9.59	3.51	6.89	4.05
Glut	1.35	1.95	1.58	1.54	0.43
Pyruvate	0.20	0.25	0.13	0.11	0.06
NaCl	4.76	4.76	5.60	3.64	7.00
I	0.01	0.01	0.02	0.02	0.10
T	5.00E-03	3.35E-02	1.95E-02	2.20E-02	1.00E-02
FGF2	3.00E-05	1.29E-04	1.32E-04	1.35E-04	9.00E-05
TGFb1	1.00E-06	1.00E-06	1.70E-06	2.80E-06	2.00E-06
EGF	4.25E-06	2.25E-05	9.25E-06	1.10E-05	0.00
P	1.73E-05	5.25E-06	1.93E-05	1.50E-05	0.00
Estra	5.75E-06	5.00E-07	4.75E-06	2.38E-06	0.00
IL6	1.13E-05	5.00E-06	3.94E-05	3.94E-05	0.00
LIF	3.75E-07	8.75E-07	4.00E-06	1.63E-06	0.00
TGFb3	0.00	0.00	4.48E-06	7.36E-06	0.00
HGF	0.00	0.00	1.00E-06	2.00E-06	0.00
PDGF	0.00	0.00	9.25E-06	9.50E-06	0.00
PEDF	0.00	0.00	2.50E-06	3.75E-06	0.00
Growth	1.23	1.78	2.12	2.84	1.00
Cost	0.09	0.30	0.39	0.41	0.24

TABLE 5.2. Optimal Media | groups of media that lie on or near the Pareto curve. Only OM2 was found by maximizing $NEI(x)$ rather than $\alpha(X)$. The concentration (mg/mL) of all media was between the minimum and maximum listed in Table 5.1. The cost is a unitless metric of relative economic cost of each component or group. Cell culture sterile water was used to make up the remaining volume not taken up by the components. Notes: ***The max/min concentration is relative to stock concentrations in Appendix C.1. All media have a sodium bicarbonate concentration of 2.44 mg/mL and were stored at 5°C.



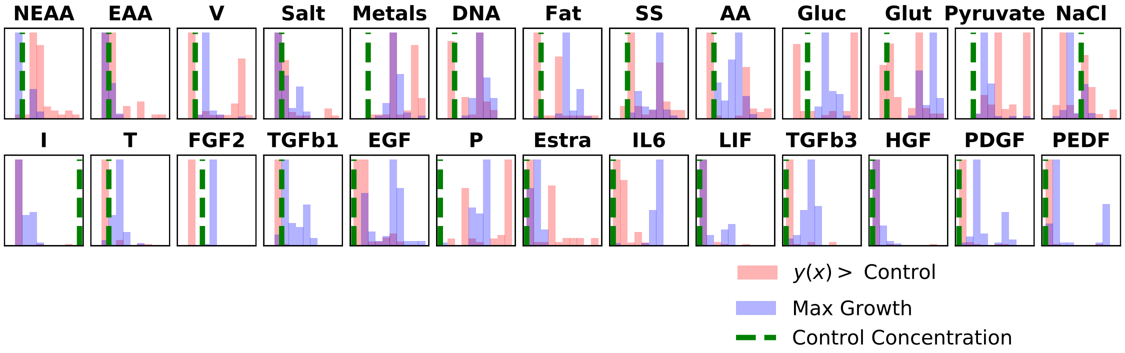


FIGURE 5.5. Distribution of Optimized Samples | samples were taken from 50 restarts of L-BFGS-B optimization algorithm. The "max growth" condition is solved as $argmax\,NEI(x)$. The $y(x) >$ Control condition solve $argmax\,\alpha(x)$ for $y_{min} = 1$, or that the solution must result in higher predicted growth than the control. The x-axis is the $[0, 1$ normalized concentration of each component and the y-axis is the density of sampled points.

5.3.3. Sensitivity Analysis. We then used the GP model to understand the correlations found in the experimental campaign. This was done by optimizing over the final model multiple times using L-BFGS-B at random starting locations for a single $q = 1$ experiment. This will show what the model believes to be the best distribution of experiments. In Figure 5.5 we show the result of samples for two conditions: (i) optimizing $\alpha(x)$ with $y_{min} = 1$ constraint and (ii) only maximizing growth using $NEI(x)$. Simply put, $\alpha(x)$ considers reducing cost while maintaining growth above y_{min} while $NEI(x)$ only maximizes growth. The max growth condition had generally higher concentrations of most growth factors but not necessarily basal components. This confirms the previous section where higher growth was achieved through higher growth factor concentrations, particularly transferrin, FGF2, TGFβ1, EGF, TGFβ3, and PDGF.

Figure 5.5 only tells us what the model thinks are the best conditions and not the relative magnitude of each factor on growth. As a further means of quantifying this, we computed the integrated variogram using the VARS technique [**64**, **65**] for each factor and show it in figure 5.6. VARS values suggest that FGF2, IL6, TGFβ1, and several basal components had significant effects on growth. This mostly confirms the previous section that FGF2, TGFβ1, and several other growth factors had a large effect on growth, but it is impossible to say anything more suggestive than that.

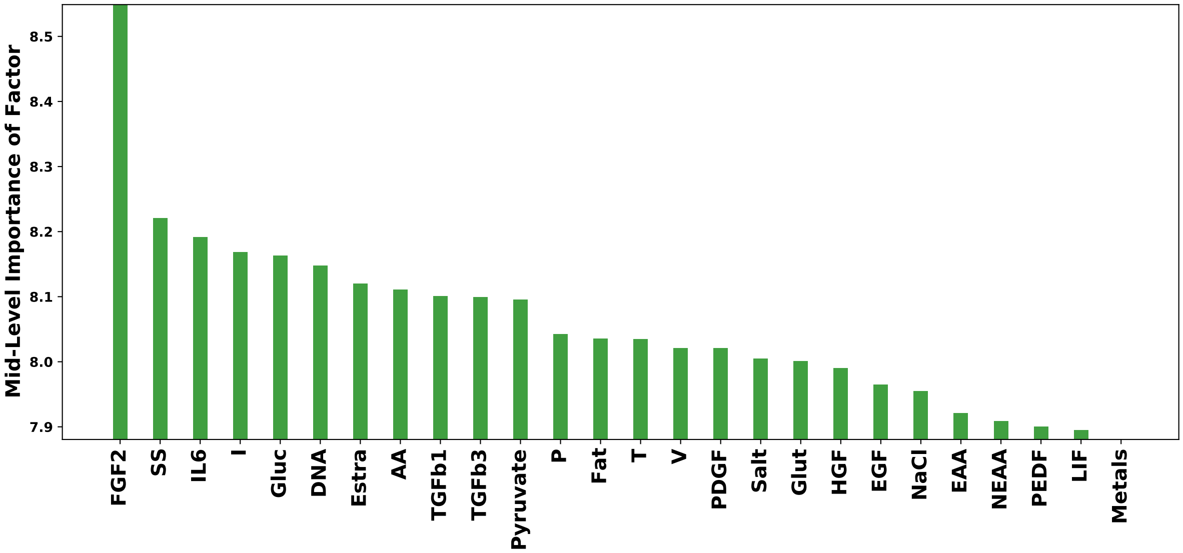


FIGURE 5.6. Mid-Level VARS Variogram | ordered by highest to lowest for the $h = 0.3$ integrated variogram. The height of the bar indicates the relative importance of that component in sensitivity of 30% differences in concentration on growth throughout the design space. Plot was generated using 1000 random samples of the design space.

5.4. Discussion

The MOBO algorithm was successful because a robust, long-term data set was built over time, improving the model as more data were collected. Additionally, the acquisition function (Equation 5.3) was tailored to generate high-value experiments near the Pareto trade-off curve between cell growth and media cost. A separate constraint function translated our need to primarily search for high-growth designs into a mathematical function, as we expected most of the design space to not support cell growth. Some shortcomings of this work are that (i) we didn't compare our MOBO method to an equivalent DOE method, though we have previously shown similar methods are significantly more efficient than traditional DOEs [**23**, **93**]. Additionally, (ii) Figure 5.4 indicates media performance tended to decrease over time. This could be due to morphological changes wrought by the media, physical damage due to passaging, or accumulation of toxins in solution. Clearly, our Passage 2 metric was not enough to fully predict the rapidly changing dynamics of cell growth over multiple passages, though it did so reasonably well given the significant savings in experimental time and resources. Finally, (iii) our media did not generalize well to other cell types (the BSCs) which limits the applicability of the designed media OM0-OM3 to C2C12 cells. However, such a result does indicate the need to re-optimize media and environmental conditions when studying new cell types or cells with significant genetic or metabolic changes, as such our methods could prove even more useful.

In general, the MOBO algorithm was able to design media according to the objective function we picked for this system. Several high-quality serum-free media formulations were discovered which allows for further, more principled, experiments to be made to accompany and expand on the discoveries made in this study. Further work should be performed on correlating biomarkers and morphological attributes to cell differentiation and proliferation, both to improve the robustness of predictions and to simultaneously optimize proliferation and differentiation. Even without these improvements, this work is still relevant to those interested in quickly optimizing their media formulations, generally in the serum-free case, and particularly in the case of difficult-to-measure objectives such as long-term cell growth.

CHAPTER 6

Review of Thesis and Future Work

The first part of this thesis (chapter 2 and 3) was comprised of the development of a radial basis function genetic algorithm sequential DOE scheme [**21**, **22**]. It drew heavily on the work of [**66**], where a sequential DOE technique was developed on the principle of local random search in areas of high performing media. This algorithm was also dynamic by converging on high performing results and selectively searching the design space when good results were not forthcoming. Additionally, previous work in our lab [**93**,**95**] provided the framework for a sequential DOE based on a truncated GA. This modified GA incorporates uncertainty in the optimal samples found by halting algorithm convergence proportional to the amount of clustering around an optima the GA finds. By hybridizing these two methods, a DOE algorithm called NNGA-DYCORS was developed that solved various computational optimization problems better than either method alone. It was used to optimize a 30-dimensional media for serum-containing C2C12 cell culture with the metric of growth being AlamarBlue reduction after 48 hrs of growth in 96 well plates (in chapter 3 it was renamed HND). While it was successful at finding media that maximized this metric (as well as minimized a cost metric), the 48 hr growth metric did not generalize well to multiple passages, and the best medium found degraded over time relative to the control.

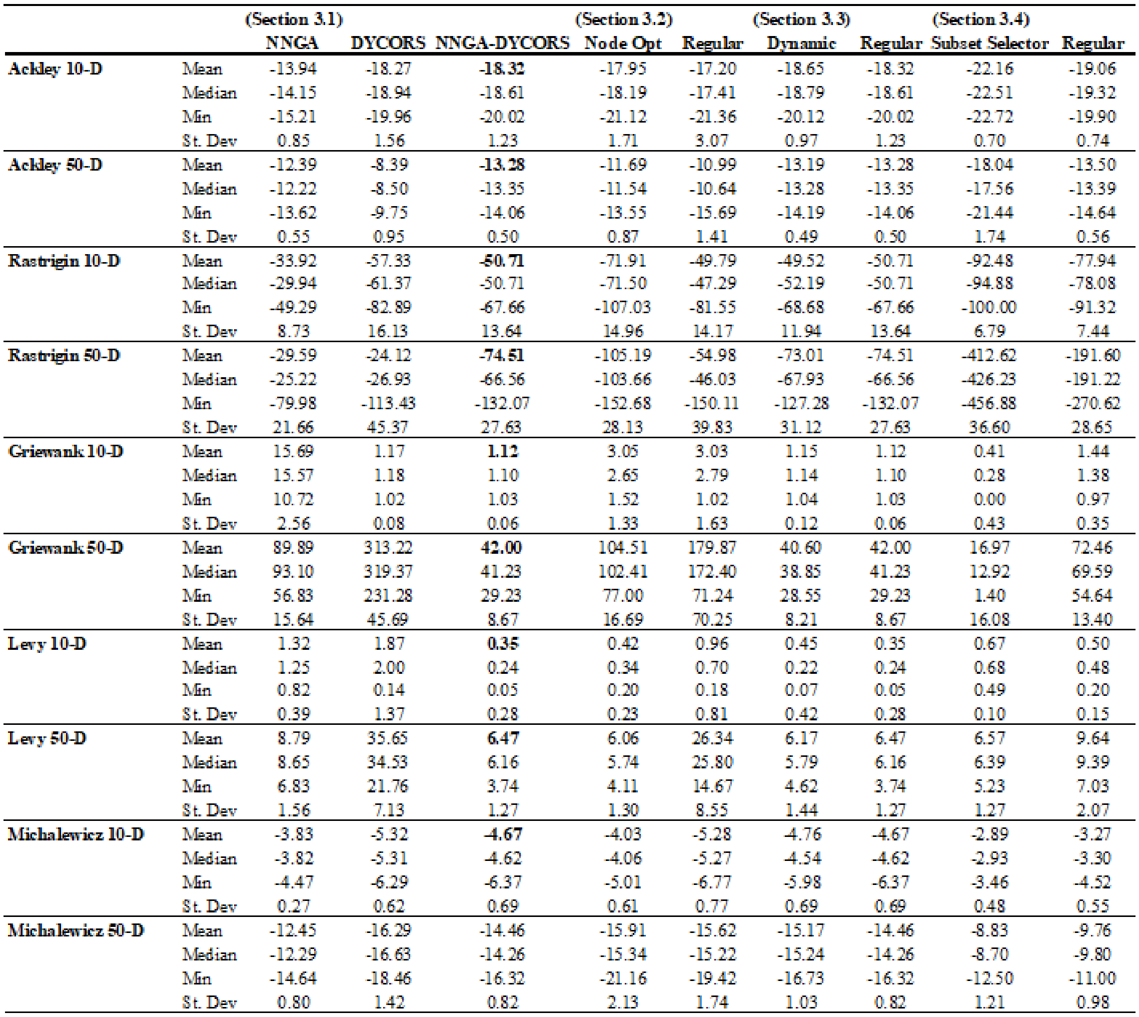
To fix this underlying problem, multiple passages needed to be incorporated into the DOE process. This is a very time-consuming process as each passage takes multiple days, many more physical manipulations than simple chemical assays which introduces opportunities for contamination, and difficulty for manual experimentation. To solve this, chemical assays were supplemented with small amounts of manual multi-passage cell counts in a multi-information source Bayesian GP model [**31**] which was used to successfully optimize a 14-dimensional serum-containing media for C2C12 cells [**23**] (chapter 4). Due to the presence of multi-passage data, the final optimal medium grew cells robustly over four passages, provided nearly twice the number of cells at the end of each passage relative to the DMEM + 10% FBS control and traditional DOE method, and did so at nearly the same cost in terms of media components. In the final chapter (chapter 5) the multi-information source GP model was extended to optimize a 26-dimensional serum-free media based on the Essential 8 media [**52**] using a multi-objective metric that improves cell growth while minimizing medium cost. Using this Bayesian metric, a broad set of media samples along the trade-off curve of media quality and cost were found, showing that a designer can be given options in media optimization. In particular, one medium resulted in higher growth over five passages while the control and Essential 8 lagged.

There are several avenues for future work. First, improving the quality and robustness of the data collected. While the Passage 2 metric did generalize to additional passages (chapter 4), it did not do so for all media (chapter 5). Rather than collect additional long-term cell counts, future researchers should attempt to find biomarker correlates with long-term growth such as Pax7, MyoD, and Myogenin. Due to the lack of expertise and resources, in-depth knowledge of the cascade of signals and molecular interactions in cells were not used to their fullest extent in this thesis, and should be considered in future. Model accuracy could also be improved by incorporating additional data such as brightfield image counts of cell number, fluorescent image counts of nuclei using Hoecht stains, and growth curve data (which is easier to collect with non-destructive techniques like brightfield). As long as the image segmentation parameters are properly tuned, additional data points like the rate of change in cell number, final cell number, cell count at each time-point, and initial cell number could be used to detect high-quality media and elucidate intra-assay correlations. While metabolomic, genomic, and lipidomic analysis [**58**] would be time-consuming to conduct given the amount of experiments we have done here, creating multi-domain models of cellular systems may be yet another route to optimization. Techno-economic analysis (whole plant analysis, capital costs, storage and preparations cost of materials etc) of growth and cost conditions may also allow future DOE studies to translate information collected at the lab-scale closer to the trade-offs considered in industry and large volume bioreactors. Rather than using multi-objective acquisition functions such as hypervolume or desirability maximization, future DOEs could feed raw data into a computer or algebraic techno-economic model and use composite Bayesian optimization [**5**] to solve the experimental optimization problem.

Secondly, fundamental "white-box" studies that focus on non-DOE aspects of the system must be completed in order to constrain the complexity of future DOE studies and set up more interesting or profitable design spaces. Studying the metabolomics of the cell lines would be very useful in defining the upper / lower bounds and important factors of the system. By knowing the limiting factors of the media using spend-media analysis, upper ranges may be adjusted and by knowing waste-product profile, potential synthetic routes may be closed. By finding that, for example, glutamine, is a limiting amino acid in the cell culture system, a higher upper bound may be set which unlocks entirely new designs (as was considered in chapter 5 after considering the work of [**58**], but was not done robustly with every component in chapter 4 or 5). In this thesis, the genetic and morphological profile of our cell lines was not considered. In fact, the adaptation to serum-free and Matrigel-free conditions may have appreciably changed these factors to the point that the optimal media found may not generalize to most C2C12 cells. Therefore, tracking the prevalence of myotubes during differentiation or staining with α-actin or myosin heavy chain antibody [**24**] would both act as a double-check for the way media optimization changes the cell line and as additional data points to consider during DOE campaigns. For example, one may optimize cell growth but with a probabilistic constraint that learns which regions of the design space result in low levels of α-actin or myosin heavy chain response and chooses experiments unlikely to violate that constraint. Future work must also go beyond C2C12s and consider cells that are relevant for cellular agriculture such as bovine, porcine, or avian. As we have seen in chapter 5, media designed for C2C12 cells does not always extend to muscle cells of different animal lineages. These new lines must be adapted to serum-free conditions which would open up the design space to more industrially relevant cell lines. All of these ideas would constitute entire projects and require their own feasibility studies, but would build upon the advances made in this dissertation and body of work.

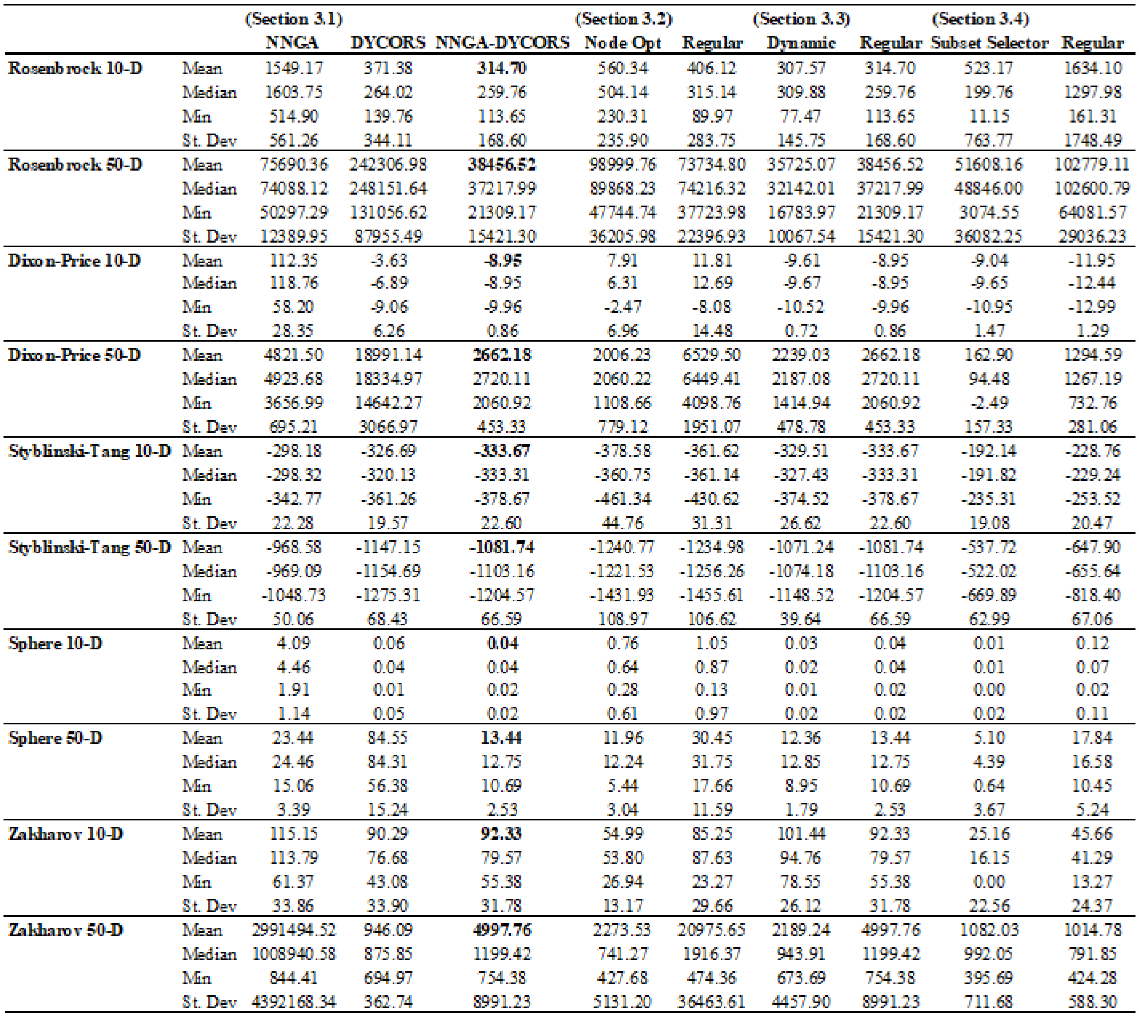
APPENDIX A

Optimization of Muscle Cell Culture Media using Nonlinear Design of Experiments



		(Section 3.1)			(Section 3.2)		(Section 3.3)		(Section 3.4)	
		NNGA	DYCORS	NNGA-DYCORS	Node Opt	Regular	Dynamic	Regular	Subset Selector	Regular
Ackley 10-D	Mean	-13.94	-18.27	**-18.32**	-17.95	-17.20	-18.65	-18.32	-22.16	-19.06
	Median	-14.15	-18.94	-18.61	-18.19	-17.41	-18.79	-18.61	-22.51	-19.32
	Min	-15.21	-19.96	-20.02	-21.12	-21.36	-20.12	-20.02	-22.72	-19.90
	St. Dev	0.85	1.56	1.23	1.71	3.07	0.97	1.23	0.70	0.74
Ackley 50-D	Mean	-12.39	-8.39	**-13.28**	-11.69	-10.99	-13.19	-13.28	-18.04	-13.50
	Median	-12.22	-8.50	-13.35	-11.54	-10.64	-13.28	-13.35	-17.56	-13.39
	Min	-13.62	-9.75	-14.06	-13.55	-15.69	-14.19	-14.06	-21.44	-14.64
	St. Dev	0.55	0.95	0.50	0.87	1.41	0.49	0.50	1.74	0.56
Rastrigin 10-D	Mean	-33.92	-57.33	**-50.71**	-71.91	-49.79	-49.52	-50.71	-92.48	-77.94
	Median	-29.94	-61.37	-50.71	-71.50	-47.29	-52.19	-50.71	-94.88	-78.08
	Min	-49.29	-82.89	-67.66	-107.03	-81.55	-68.68	-67.66	-100.00	-91.32
	St. Dev	8.73	16.13	13.64	14.96	14.17	11.94	13.64	6.79	7.44
Rastrigin 50-D	Mean	-29.59	-24.12	**-74.51**	-105.19	-54.98	-73.01	-74.51	-412.62	-191.60
	Median	-25.22	-26.93	-66.56	-103.66	-46.03	-67.93	-66.56	-426.23	-191.22
	Min	-79.98	-113.43	-132.07	-152.68	-150.11	-127.28	-132.07	-456.88	-270.62
	St. Dev	21.66	45.37	27.63	28.13	39.83	31.12	27.63	36.60	28.65
Griewank 10-D	Mean	15.69	1.17	**1.12**	3.05	3.03	1.15	1.12	0.41	1.44
	Median	15.57	1.18	1.10	2.65	2.79	1.14	1.10	0.28	1.38
	Min	10.72	1.02	1.03	1.52	1.02	1.04	1.03	0.00	0.97
	St. Dev	2.56	0.08	0.06	1.33	1.63	0.12	0.06	0.43	0.35
Griewank 50-D	Mean	89.89	313.22	**42.00**	104.51	179.87	40.60	42.00	16.97	72.46
	Median	93.10	319.37	41.23	102.41	172.40	38.85	41.23	12.92	69.59
	Min	56.83	231.28	29.23	77.00	71.24	28.55	29.23	1.40	54.64
	St. Dev	15.64	45.69	8.67	16.69	70.25	8.21	8.67	16.08	13.40
Levy 10-D	Mean	1.32	1.87	**0.35**	0.42	0.96	0.45	0.35	0.67	0.50
	Median	1.25	2.00	0.24	0.34	0.70	0.22	0.24	0.68	0.48
	Min	0.82	0.14	0.05	0.20	0.18	0.07	0.05	0.49	0.20
	St. Dev	0.39	1.37	0.28	0.23	0.81	0.42	0.28	0.10	0.15
Levy 50-D	Mean	8.79	35.65	**6.47**	6.06	26.34	6.17	6.47	6.57	9.64
	Median	8.65	34.53	6.16	5.74	25.80	5.79	6.16	6.39	9.39
	Min	6.83	21.76	3.74	4.11	14.67	4.62	3.74	5.23	7.03
	St. Dev	1.56	7.13	1.27	1.30	8.55	1.44	1.27	1.27	2.07
Michalewicz 10-D	Mean	-3.83	-5.32	**-4.67**	-4.03	-5.28	-4.76	-4.67	-2.89	-3.27
	Median	-3.82	-5.31	-4.62	-4.06	-5.27	-4.54	-4.62	-2.93	-3.30
	Min	-4.47	-6.29	-6.37	-5.01	-6.77	-5.98	-6.37	-3.46	-4.52
	St. Dev	0.27	0.62	0.69	0.61	0.77	0.69	0.69	0.48	0.55
Michalewicz 50-D	Mean	-12.45	-16.29	-14.46	-15.91	-15.62	-15.17	-14.46	-8.83	-9.76
	Median	-12.29	-16.63	-14.26	-15.34	-15.22	-15.24	-14.26	-8.70	-9.80
	Min	-14.64	-18.46	-16.32	-21.16	-19.42	-16.73	-16.32	-12.50	-11.00
	St. Dev	0.80	1.42	0.82	2.13	1.74	1.03	0.82	1.21	0.98

Table A.1. Final Algorithm Performance Part I. Experiments from Section 2.3.1 where NNGA-DYCORS performed better than or as well as the next best constituent algorithm by one standard deviation have their mean bolded. Variants of NNGA-DYCORS are tested in Sections 2.3.2 – 2.3.4. All data shown are from the final batch of experiments for all algorithms.



		(Section 3.1)			(Section 3.2)		(Section 3.3)		(Section 3.4)	
		NNGA	**DYCORS**	**NNGA-DYCORS**	**Node Opt**	**Regular**	**Dynamic**	**Regular**	**Subset Selector**	**Regular**
Rosenbrock 10-D	Mean	1549.17	371.38	**314.70**	560.34	406.12	307.57	314.70	523.17	1634.10
	Median	1603.75	264.02	259.76	504.14	315.14	309.88	259.76	199.76	1297.98
	Min	514.90	139.76	113.65	230.31	89.97	77.47	113.65	11.15	161.31
	St. Dev	561.26	344.11	168.60	235.90	283.75	145.75	168.60	763.77	1748.49
Rosenbrock 50-D	Mean	75690.36	242306.98	**38456.52**	98999.76	73734.80	35725.07	38456.52	51608.16	102779.11
	Median	74088.12	248151.64	37217.99	89868.23	74216.32	32142.01	37217.99	48846.00	102600.79
	Min	50297.29	131056.62	21309.17	47744.74	37723.98	16783.97	21309.17	3074.55	64081.57
	St. Dev	12389.95	87955.49	15421.30	36205.98	22396.93	10067.54	15421.30	36082.25	29036.23
Dixon-Price 10-D	Mean	112.35	-3.63	**-8.95**	7.91	11.81	-9.61	-8.95	-9.04	-11.95
	Median	118.76	-6.89	-8.95	6.31	12.69	-9.67	-8.95	-9.65	-12.44
	Min	58.20	-9.06	-9.96	-2.47	-8.08	-10.52	-9.96	-10.95	-12.99
	St. Dev	28.35	6.26	0.86	6.96	14.48	0.72	0.86	1.47	1.29
Dixon-Price 50-D	Mean	4821.50	18991.14	**2662.18**	2006.23	6529.50	2239.03	2662.18	162.90	1294.59
	Median	4923.68	18334.97	2720.11	2060.22	6449.41	2187.08	2720.11	94.48	1267.19
	Min	3656.99	14642.27	2060.92	1108.66	4098.76	1414.94	2060.92	-2.49	732.76
	St. Dev	695.21	3066.97	453.33	779.12	1951.07	478.78	453.33	157.33	281.06
Styblinski-Tang 10-D	Mean	-298.18	-326.69	**-333.67**	-378.58	-361.62	-329.51	-333.67	-192.14	-228.76
	Median	-298.32	-320.13	-333.31	-360.75	-361.14	-327.43	-333.31	-191.82	-229.24
	Min	-342.77	-361.26	-378.67	-461.34	-430.62	-374.52	-378.67	-235.31	-253.52
	St. Dev	22.28	19.57	22.60	44.76	31.31	26.62	22.60	19.08	20.47
Styblinski-Tang 50-D	Mean	-968.58	-1147.15	**-1081.74**	-1240.77	-1234.98	-1071.24	-1081.74	-537.72	-647.90
	Median	-969.09	-1154.69	-1103.16	-1221.53	-1256.26	-1074.18	-1103.16	-522.02	-655.64
	Min	-1048.73	-1275.31	-1204.57	-1431.93	-1455.61	-1148.52	-1204.57	-669.89	-818.40
	St. Dev	50.06	68.43	66.59	108.97	106.62	39.64	66.59	62.99	67.06
Sphere 10-D	Mean	4.09	0.06	**0.04**	0.76	1.05	0.03	0.04	0.01	0.12
	Median	4.46	0.04	0.04	0.64	0.87	0.02	0.04	0.01	0.07
	Min	1.91	0.01	0.02	0.28	0.13	0.01	0.02	0.00	0.02
	St. Dev	1.14	0.05	0.02	0.61	0.97	0.02	0.02	0.02	0.11
Sphere 50-D	Mean	23.44	84.55	**13.44**	11.96	30.45	12.36	13.44	5.10	17.84
	Median	24.46	84.31	12.75	12.24	31.75	12.85	12.75	4.39	16.58
	Min	15.06	56.38	10.69	5.44	17.66	8.95	10.69	0.64	10.45
	St. Dev	3.39	15.24	2.53	3.04	11.59	1.79	2.53	3.67	5.24
Zakharov 10-D	Mean	115.15	90.29	**92.33**	54.99	85.25	101.44	92.33	25.16	45.66
	Median	113.79	76.68	79.57	53.80	87.63	94.76	79.57	16.15	41.29
	Min	61.37	43.08	55.38	26.94	23.27	78.55	55.38	0.00	13.27
	St. Dev	33.86	33.90	31.78	13.17	29.66	26.12	31.78	22.56	24.37
Zakharov 50-D	Mean	2991494.52	946.09	**4997.76**	2273.53	20975.65	2189.24	4997.76	1082.03	1014.78
	Median	1008940.58	875.85	1199.42	741.27	1916.37	943.91	1199.42	992.05	791.85
	Min	844.41	694.97	754.38	427.68	474.36	673.69	754.38	395.69	424.28
	St. Dev	4392168.34	362.74	8991.23	5131.20	36463.61	4457.90	8991.23	711.68	588.30

Table A.2. Final Algorithm Performance Part II. Experiments from Section 2.3.1 where NNGA-DYCORS performed better than or as well as the next best constituent algorithm by one standard deviation have their mean bolded. Variants of NNGA-DYCORS are tested in Sections 2.3.2 – 2.3.4. All data shown are from the final batch of experiments for all algorithms.

APPENDIX B

Multi-Information Source Bayesian Optimization of Culture Media for Cellular Agriculture

B.1. Bayesian Model Details

In order to express skepticism over datapoints collected with noise, we incorporated experimental variance measurements into the noise model to get an additional heteroskedastic noise term $\Sigma_\epsilon = (\sigma_\epsilon^2 + v)I$ where $v_i = \frac{1}{a_i - 1}\sum_{j=1}^{a_i}(y_i^j - \bar{y}_i)^2$ for a given measurement y_i with a_i replicates (usually 3 or 5 for this study depending on the IS available).

To fit the model, the optimal hyperparameters θ^* were determined through maximization of the likelihood $L'(X, Y|\theta)$ of the given data X_N and Y_N given a set of hyperparameters $\theta = \lambda, \sigma_f, \mu_0, \sigma$. This was posed as $\theta^* = argmin - logL'(X_N, Y_N|\theta)$.

$$logL'(X_N, Y_N|\theta) = -N/2ln(2\pi) - 1/2Y_N^T K_y^{-1} Y_N - 1/2log|K_y^{-1}|$$

$$K_y = \Sigma(X_N, X_N) + \Sigma_\epsilon$$

$$\nabla_{\theta_i} logL'(X_N, Y_N|\theta) = 1/2Y_N^T K_y^{-1} Y_N - 1/2Tr(K_y^{-1}\nabla_{\theta_i} K_y)$$

We solved for θ^* using the limited memory bound-constrained Broyden–Fletcher–Goldfarb–Shanno (L-BFGS-B) algorithm **[15]** (bounds of λ, σ_f and σ set to $[10^{-2}, 10]$). L-BFGS-B was well suited to our hyperparameter problem because, as a quasi-Newton method, it has access to first and second order (derivative and hessian) information about the negative log likelihood curvature while inverting hessian matrices using less computation than ordinary Newton methods.

We know biological systems produce smooth, unimodal responses to environmental conditions rather than steep, multimodal ones. This assumption can be encoded by placing prior distributions over the hyperparameters and solving for θ^* jointly. A normal prior was placed over the length scale and output scale hyperparameters $\lambda, \sigma_f \sim N(1, 0.25)$ and a gamma prior over the underlying noise $\sigma_\epsilon \sim Gamma(1.1, 0.05)$ with gamma function $\Gamma(1.1)$. This has the effect of (i) regularizing λ

and σ_f values such that they remain near to their lower bound, and thus enforce more interaction between datapoints to smooth the response surface while (ii) constraining the underlying noise σ_ϵ so the response does not flatten out in the presence of little data.

$$Pr(\lambda) = \frac{1}{\sqrt{0.25 \times 2\pi}} \times exp(-1/2(\frac{\lambda - 1}{\sqrt{0.25}})^2)$$

$$Pr(\sigma_f) = \frac{1}{\sqrt{0.25 \times 2\pi}} \times exp(-1/2(\frac{\sigma_f - 1}{\sqrt{0.25}})^2)$$

$$Pr(\sigma_\epsilon) = \frac{0.05^1.1}{\Gamma(1.1)} \times \sigma_\epsilon \times exp(-0.05 \times \sigma_\epsilon)$$

If the priors are independent of each other, they can be expressed as a product $Pr(\theta) = \prod_{i=1}^{(p+1)M+2} Pr(\theta_i)$ and incorporated into the solution to θ^* by modifying the log likelihood.

$$logL(X_N, Y_N|\theta) = -N/2ln(2\pi) - 1/2Y_N^T K_y^{-1} Y_N - 1/2log|K_y^{-1}| + \Sigma_{j=1}^{(p+1)M+2} logPr(\theta_j)$$

$$\nabla_{\theta_i} logL(X_N, Y_N|\theta) = 1/2Y_N^T K_y^{-1} Y_N - 1/2Tr(K_y^{-1} \nabla_{\theta_i} K_y) + \nabla_{\theta_i} logPr(\theta_i)$$

If θ^* is solved for then we can model the output $y(x)$ given x for any IS.

B.2. BO Acquisition Function Details

To take uncertainty captured by $\sigma^2(x)$ into consideration, we take the expectation "E" under the distribution described by θ^* and the N datapoints X_N and Y_N. This was done using the "reparameterization trick" [**91**] where the prediction was sampled as $Y = \mu(X) + L(X)Z$ with Cholesky Decomposition of the covariance matrix $\Sigma(X, X) = L(X)L(X)^T$ and multivariant random normal vector $Z \sim N(0, 1)$. If we take R random samples of Z for points Y, we capture the uncertainty modeled by the GP and utilize it in calculations that do not have analytically tractable integrals such as $\alpha(X)$. First, we sample Y using Z, then calculate $D(X)$ for each sample, thus propagating uncertainty through $D(X)$ as a composite of $y(x)$ and $c(x)$. Note that the cost $c(x)$ was not modeled by a GP, so does not contribute to the uncertainty of $D(X)$. For more information on composite Bayesian optimization see [**5**].

$$\alpha(X) \approx \frac{1}{R} \sum_{r=1}^{R} (max\{D(Y_r|X)\} - D^*(X_N))^+$$

We also express uncertainty in the current optimal point $D^*(X_N)$ by using the samples Z to calculate a distribution of predictions of the union of all points $\Omega = \{X \bigcup X_N\}$ and get the noisy modification of the multi-point expected improvement objective function [**54**]. A sampling policy that maximizes $\alpha(X)$ would, therefore, sample uncertain regions multiple times to improve estimates of the posterior.

$$\alpha(X) \approx \frac{1}{R}\sum_{r=1}^{R}(max\{D(Y_r|X)\} - max\{D^*(Y_{r,N}|\Omega\})^+$$

Gradient estimates of the above equation $\nabla\alpha(X)$ were computed by taking an average of the gradient of each R sample.

$$\nabla\alpha(X) \approx \frac{1}{R}\sum_{r=1}^{R}(max\{\nabla D(\mu(X) + L(X) \times z_r|X)\} - max\{\nabla D^*(\mu(X_N) + L(X_N) \times z_r|\Omega\})^+$$

where $\nabla\mu(X)$ can be easily calculated and $\nabla L(X)$ calculated as per [**75**], both propagated to find $\nabla D(X)$ using chain rule to estimate $\nabla\alpha(X)$. In practice, this was done using auto-differentiation of $R - 2000$ samples Z. These gradient were used in L-BFGS-B to maximize $\alpha(X)$ for a group of q experiments of p-dimension on the domain $[0, 1]$.

B.3. Computational Validation of BO Method Test Function Details

The test functions $\{f_1, f_2, f_3, f_4\}$ used in the computational experiments in this work are shown below.

$$f_1(x) = \sum_i^{10}(x_i - 0.5)^2 + U(-0.2, 0.2)$$

$$f_{1bias,1}(x) = 3\sum_i^{10}(x_i - 0.4)^2 + x_2 - x_4 + x_{10} + U(-0.3, 0.3)$$

$$f_{1bias,2}(x) = \sum_i^{10}(x_i - 0.75)^2 + x_1 - x_5 + x_9 - x_{10} + U(-0.2, 0.2)$$

$$f_2(x) = \sum_j^{10} x_j^2 + \sum_{i=2}^{9} x_i x_{i-1} + U(-0.1, 0.1)$$

$$f_{2bias,1}(x) = f_2(x) - 2(x_3 - x_6 + x_{10}) + 1 + U(-0.1, 0.1)$$

$$f_{2bias,2}(x) = f_2(x) + 2(x_2 - x_5 + x_9) + x_1 x_5$$

$$c = \{0, 0.5, 0.3, 0.8, 1.0\}$$

$$b = \{0, -2.5, 0.3, 0.8, -5\}$$

$$f_3(x) = \sum_{i=1}^{5}(x_i - c_i)^2 + \sum_{i=6}^{1} 0x_i b_i + U(-0.35, 0.35)$$

$$f_{3bias,1}(x) = \sum_{i=1}^{5}(x_i - 0.35 - c_i)^2 + \sum_{i=6}^{1} 0x_i b_i - 2(x_3 - x_6 + x_{10}) + 1 + U(-0.7, 0.7)$$

$$f_{3bias,2}(x) = \sum_{i=1}^{5}(x_i + 0.35 - c_i)^2 + \sum_{i=6}^{1} 0x_i b_i + 2(x_1 + x_5 - x_9) + 2 + x_1 x_5 + U(-0.35, 0.35)$$

$$f_4(x) = \sum_{i=1}^{5}(x_i - c_i)^2 + \sum_{i=6}^{1} 0x_ib_i + x_1x_2 + 2x_6x_4 - 3x_8x_9$$

$$f_{4bias,1}(x) =$$

$$\sum_{i=1}^{5}(x_i - 0.2 - c_i)^2 + \sum_{i=6}^{1} 0x_ib_i + x_1x_2 + 2x_6x_4 - 3x_8x_9 - 2(x_2 + 3x_5 - x_{10}) + U(-0.35, 0.35)$$

$$f_{4bias,2}(x) = \sum_{i=1}^{5}(x_i + 0.2 - c_i)^2 + 2\sum_{i=6}^{1} 0x_ib_i + 2x_1x_2 + 4x_6x_4 - 6x_8x_9 + 2(x_1 - 3x_5 + x_9)$$

B.4. Data and Model Availability

Using Github link `https://github.com/ZacharyCosenza/GradStuff_Cosenza` the input and output data should be available under `DBO_Data_BO_data.txt` (X matrix arranged as experiments in rows and concentrations normalized $[0, 1]$ in columns with last column being the IS indicator, $0 =$ Passage 2, $1 =$ Passage 1, $2 =$ AlamarBlue, $3 =$ LIVE Stain) and `BO_outputs.txt` (first column Y second column variance v). The final optimal model parameters may be loaded from `BO_Model8.pnd` into `DBO_Solver.py` using standard pytorch framework.

APPENDIX C

Multi-Objective Bayesian Optimization of Serum-Free Culture Media for Cellular Agriculture

C.1. Serum-Free Media Components

These tables show the composition of the components in the MOBO design space. These chemicals were grouped together to reduce the dimensionality of the design space while still providing the opportunity for the optimization algorithm to vary their concentrations. There are also various non-grouped components with their own preparation and reconstitution schemes shown in bullet points.

- **Sodium Selenite**: store in PBS at -20°C.
- **Ascorbic Acid**: store in PBS at 5°C, remake every 1-2 weeks.
- **Insulin**: store in water at -20°C. Adjust pH for solubility to 2-3.
- **Transferrin**: store in PBS at -20°C.
- **FGF2**: dilute in 5 mM tris buffer to around 0.1-1 mg/mL and store in PBS at 5°C + 0.10% BSA (bovine serum albumin). PeproTech cat. no. 100-18B.
- **TGFβ1**: dilute in 10 mM citric acid (pH of 3) to around 0.1-1 mg/mL and store in PBS at 5°C + 0.10% BSA. PeproTech cat. no. 100-21.

Component	**Conc. (mg/mL)**
Glycine	0.75
Alanine	0.89
Glutamic Acid	1.47
Proline	1.15
Serine	1.05
Aspartic Acid	1.33
Asparagine-H_2O	1.32
Cysteine-HCl-H_2O	1.76

TABLE C.1. Non-Essential Amino Acids (NEAA) | to make this purchase MEM-NEAA (Gibco) 100x and add Cysteine-HCl-H_2O. Store solution in -3°C.

Component	**Conc. (mg/mL)**
Histidine HCl-H2O	2.10
Isoleucine	2.62
Leucine	2.62
Lysine HCl	3.64
Methionine	0.76
Phenylalanine	1.65
Threonine	2.38
Tryptophan	0.51
Valine	2.34
Tyrosine Disodium Salt Dihydrate	1.80
Arginine HCl	6.32
Cystine	1.20

TABLE C.2. Essential Amino Acids (EAA) | to make this purchase MEM-EAA (Gibco) 50x. Store solution in -3°C.

Component	**Conc. (mg/mL)**
Choline Chloride	0.1
D-Calcium Pantothenate	0.1
Folic Acid	0.1
Nicotinamide	0.1
Pyridoxine HCl	0.1
Riboflavin (Vitamin B2)	0.1
Thiamine HCl	0.1
i-Inositol	0.2
Biotin (Vitamin B7)	1.75E-04
Cobalamin (Vitamin B12)	3.40E-02

TABLE C.3. Vitamins (V) | to make this purchase MEM-Vitamin Solution (Gibco) 100x and add Biotin (Vitamin B7) and Cobalamin (Vitamin B12). Store solution in -80°C away from sunlight.

Component	**Conc. (mg/mL)**
Calcium Chloride (anhyd)	5.83
Potassium Chloride	15.59
Magnesium Sulfate (anhyd)	5.00
Magnesium Chloride (anh)	3.06
Sodium Phosphate Dibasic	6.70
Sodium Phosphate Monobasic	3.13

TABLE C.4. Salts Solution | to make this it is best to make a separate solution for Sodium Phosphate Dibasic and Monobasic and add both solutions to the medium to the desired concentrations. Store both solutions in -3°C.

Component	**Conc. (mg/mL)**
Ferric Sulfate	4.17E-02
Ferric Nitrate	5.00E-03
Zinc Sulfate	4.32E-02
Cupric Sulfate Pentahydrate	1.30E-04

Table C.5. Trace Metals Solution. Store in -3°C away from sunlight.

Component	**Conc. (mg/mL)**
Linoleic Acid	8.40E-03
Lipoic Acid	2.10E-02

Table C.6. Fatty Acids Solution | Lipoic Acid only soluble in ethanol. Store in -80°C.

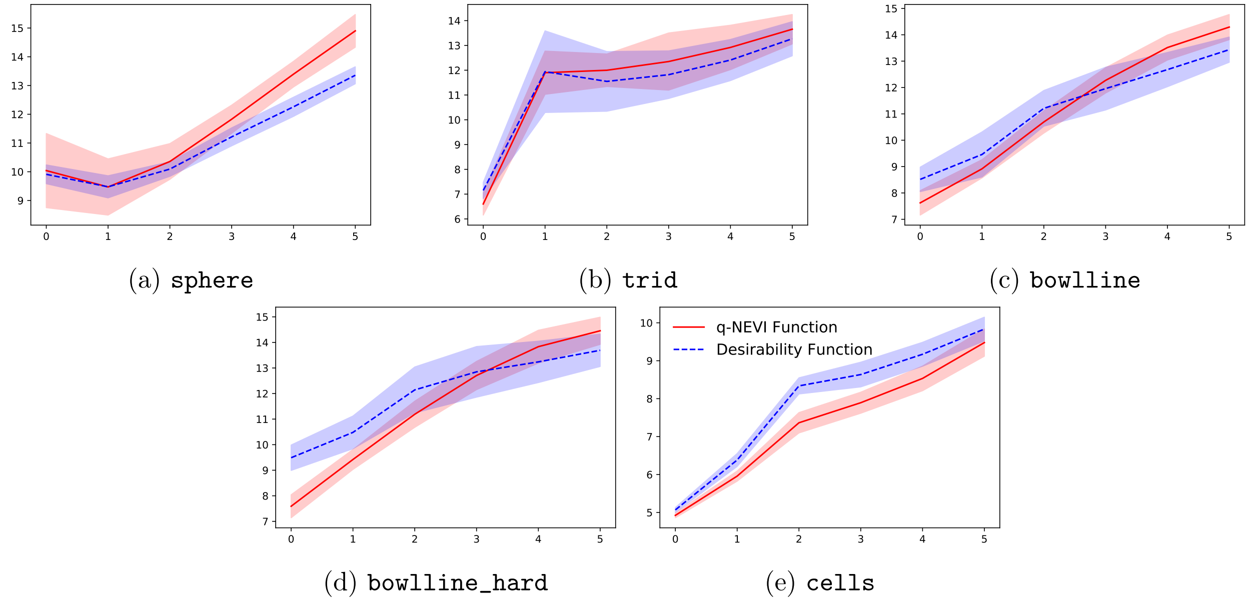
Component	**Conc. (mg/mL)**
Hypoxanthine Na	4.78E-01
Putrescine 2HCl	1.62E-02
Thymidine	7.30E-02

Table C.7. DNA Precursor Solution | Hypoxanthine Na soluble in DMSO, which results in at most 0.2% v/v DMSO in media. Unlikely to by highly toxic to C2C12 cells. Other components are water soluble. Store stock in -3°C away from sunlight. Solution should be remade every 1-2 months.

- **EGF**: reconstitute in sterile water 0.1-1 mg/mL and dilute in PBS + 0.10% BSA, store at -20 - -80°C. PeproTech cat. no. AF-100-15.
- **Progesterone**: store in 100% ethanol at -20°C.
- **Estradiol**: store in 1:4 DMSO:PBS at -20°C.
- **IL-6**: reconstitute in sterile water 0.1-0.5 mg/mL, gently shake for 10 min, and dilute in PBS + 0.10% BSA, store at -20°C.
- **TGFβ3**: reconstitute in 5-10 mM citric acid to 0.1-1 mg/mL, and dilute in PBS + 0.10% BSA, store at -20°C. PeproTech cat. no. 100-36E.
- **HGF**: reconstitute in water to 0.5 mg/mL, dilute in PBS + 0.10% BSA, store at -20 - -80°C. PeproTech cat. no. 100-39H.
- **PDGF**: reconstitute in water to 0.1-1 mg/mL, dilute in PBS + 0.10% BSA, store at -20 - -80°C. PeproTech cat. no. 100-14B.
- **PEDF**: reconstitute in water to 0.1-1 mg/mL, dilute in PBS + 0.10% BSA, store at -20 - -80°C. PeproTech cat. no. 130-13.

C.2. Computational Validation of MOBO Method

We first tested our MOBO algorithm on various multi-objective and multi-information source test functions solving $argmax[f(x), c(x)]$ by maximizing $\alpha(X)$. Four of the test problems (`sphere`, `trid`, `bowlline`, `bowlline_hard`) had two low-fidelity test functions and another had three (`cells`). Each had a cost function $c(x) = \Sigma_{i=1}^{p} c_i x_i$. Looking at $HV(X)$ over time, $\alpha(X)$ performed similar to the desirability function (developed in the previous chapter) for all test problems using the same GP architecture and data availability. We decided the use $\alpha(X)$ for the fact that it did not require choosing parameters to scale the trade-off between growth y and cost c. This is in contrast to the desirability function that requires a specific scalarization through various weights. Particularly when looking at a novel media system, it was useful to remain agnostic to the outcome space and allow for more exploration of the Pareto curve.



(a) `sphere` (b) `trid` (c) `bowlline`

(d) `bowlline_hard` (e) `cells`

FIGURE C.1. Computational Test Results | x-axis and y-axis display batch of experiments and average max value found by the experimental design algorithm. `sphere`, `trid`, `bowlline` and `bowlline_hard` were the same as in Appendix B.3 and C.2. `cells` is discussed in Appendix C.2.

C.3. Computational Test Problem Information

For cells test problem a Latin Hypercube (LH) sample was used to initialize 15 samples for 3 information sources, with 2 random samples being of the highest fidelity. Subsequently 5 samples of 2 of the information sources were made, with 2 additional high-fidelity samples per batch for 5 batches. (The last low-fidelity information source was not used due to a bug in the code). This resulted in a total of $15 \times 3 + 2 + (2 \times 5 + 2) \times 5 = 107$ samples of all information sources. This was repeated 18 times to get a standard deviation. For sphere, trid, bowlline, and bowlline_hard, only 2 low-fidelity information sources were available so 92 total samples where used, again for 18 repeats. Hyperparameter and acquisition function optimization was done using multi-start L-BFGS-B implemented in botorch/scipy. These test problems are described in Appendix B.3. For cells the data-set found in the previous chapter ($N = 248$ for a 14-dimensional design problem with four different information sources) was used to train a neural network model for each information source. The optimal architecture was found using K-fold cross validation error over the entire data-set for MSE minimization for num_layers $= [100, 10]$, $\alpha = [0.00001, 0.0001, 0.001, 0.01, 0.1, 1]$,max_iters $= [100, 1000, 5000]$,tol $= [1^{-4}, 1^{-6}, 1^{-8}, 1^{-10}]$. The final architecture was $\alpha = 1^{-5}$, num_layers $= 10$, max_iters $= 100$, tol $= 1^{-6}$. An example of the different information sources being used with the same optimal architecture is shown below. For neural network test problem scikit-learn 0.24.1 was used for training and testing.

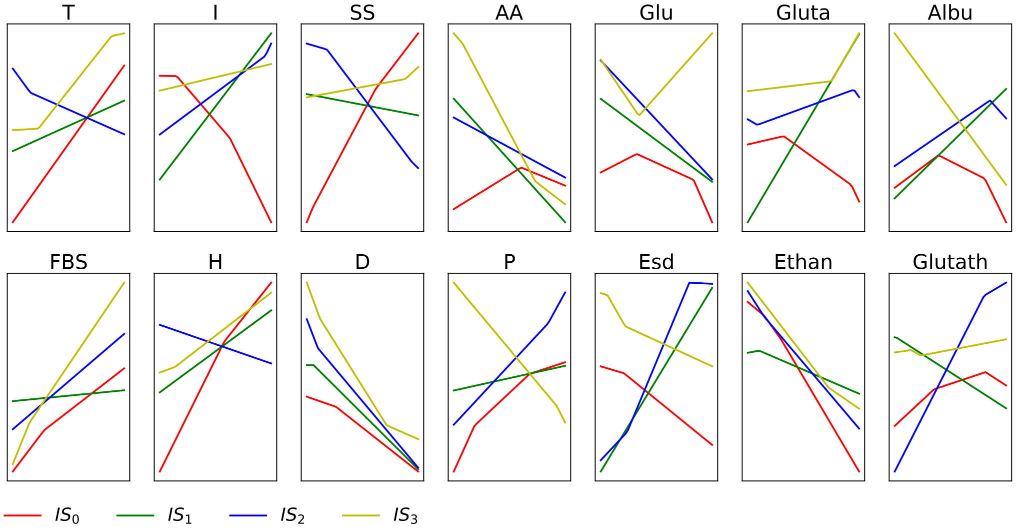


FIGURE C.2. Prediction for Neural Network on `cells` | for the midpoint $X = 0.5$ this figure shows the predictions for four different information sources for $x = [0, 1]$.

C.4. Data and Model Availability

Using Github link `https://github.com/ZacharyCosenza/GradStuff_Cosenza` the input and output data should be available under `X.txt` and `Y.txt` respectively. Each row in `X.txt` is an experiment with conditions along the columns in $[0, 1]$ normalized format. The last column is the information source where 0 = Passage 2, 1 = Passage 1, 2 = AlamarBlue, 3 = LIVE Stain. For `Y.txt` the columns are replicates. The final optimal model parameters may be loaded from `MOBO_Model13.pth` into `MOBO_Solver.py` using standard pytorch framework.

Bibliography

[1] B. Akteke-Ozturk, G. Koksal, and G. W. Weber, *Nonconvex optimization of desirability functions*, Quality Engineering, 30 (2018), pp. 293–310.

[2] B. Ankenman, B. L. Nelson, and J. Staum, *Stochastic Kriging for Simulation Metamodeling*, Operations Research, 58 (2010), pp. 371–382.

[3] M. Arora, *Cell Culture Media: A Review*, 2013.

[4] G. E. Arteaga, E. Li-Chan, M. C. Vazquez-Arteaga, and S. Nakai, *Systematic experimental designs for product formula optimization*, Trends in Food Science and Technology, 5 (1994), pp. 243–254.

[5] R. Astudillo and P. I. Frazier, *Bayesian optimization of composite functions*, 36th International Conference on Machine Learning, ICML 2019, 2019-June (2019), pp. 547–556.

[6] ATCC, *ATCC Animal Cell Culture Guide*.

[7] S. Ba and V. R. Joseph, *Composite Gaussian process models for emulating expensive functions*, Annals of Applied Statistics, 6 (2012), pp. 1838–1860.

[8] R. H. Baltz, A. L. Demain, and J. E. Davies, *Manual of Industrial Microbiology and Biotechnology*, vol. 15, 2010.

[9] P. M. Bapat and P. P. Wangikar, *Optimization of Rifamycin B Fermentation in Shake Flasks Via a Machine-Learning-Based Approach*, Biotechnology and Bioengineering, 86 (2004), pp. 201–208.

[10] A. Barbero, V. Palumberi, B. Wagner, R. Sader, M. J. Grote, and I. Martin, *Experimental and mathematical study of the influence of growth factors on the growth kinetics of adult human articular chondrocytes*, Journal of Cellular Physiology, 204 (2005), pp. 830–838.

[11] S. Belakaria, A. Deshwal, and J. R. Doppa, *Max-value entropy search for multi-objective Bayesian optimization*, Advances in Neural Information Processing Systems, 32 (2019).

[12] ———, *Multi-Fidelity Multi-Objective Bayesian Optimization: An Output Space Entropy Search Approach*, Proceedings of the AAAI Conference on Artificial Intelligence, 34 (2020), pp. 10035–10043.

[13] Z. I. Botev, *The normal law under linear restrictions: simulation and estimation via minimax tilting*, Journal of the Royal Statistical Society. Series B: Statistical Methodology, 79 (2017), pp. 125–148.

[14] D. Brunner, H. Appl, W. Pfaller, and G. Gstraunthaler, *Serum-free Cell Culture : The Serum-free Media Interactive Online Database*, (2010), pp. 53–62.

[15] R. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A Limited Memory Algorithm for Bound Constrained Optimization*, Journal of Scientific Computing, 16 (1995), pp. 1190–1208.

[16] C. C., M. T. J., M. M. D., and Y. D., *Bayesian Prediction of Deterministic Functions, With Applications to the Design and Analysis of Computer Experiments*, J. Am. Stat. Assoc., (1991), p. 953– 963.

[17] . L. Y. P. Chen S. E., Jin B., *TNF-α regulates myogenesis and muscle regeneration by activating p38 MAPK.*, American Journal of Physiology - Cell Physiology, (2007).

[18] T. M. Cioppa and T. W. Lucas, *Efficient nearly orthogonal and space-filling Latin hypercubes*, Technometrics, 49 (2007), pp. 45–55.

[19] M. C. Coleman and D. E. Block, *Retrospective Optimization of Time-Dependent Fermentation Control Strategies Using Time-Independent Historical Data*, Journal of Chemical Technology and Metallurgy, 51 (2006), pp. 726–734.

[20] M. C. Coleman, K. K. S. Buck, and D. E. Block, *An integrated approach to optimization of Escherichia coli fermentations using historical data*, Biotechnology and Bioengineering, 84 (2003), pp. 274–285.

[21] Z. Cosenza and D. E. Block, *A generalizable hybrid search framework for optimizing expensive design problems using surrogate models*, Engineering Optimization, (2020).

[22] Z. Cosenza, D. E. Block, and K. Baar, *Optimization of muscle cell culture media using nonlinear design of experiments*, Biotechnology Journal, 16 (2021), p. 2100228.

[23] Z. Cosenza, D. E. Block, P. I. Frazier, and K. Baar, *Multi - information source Bayesian optimization of culture media for cellular agriculture*, (2022), pp. 1–12.

[24] M. Das, K. Wilson, and J. J. Hickman, *Differentiation of skeletal muscle and integration of myotubes with silicon microstructures using serum-free medium and a synthetic silane substrate*, Nature Protocols, 2 (2007), pp. 1795–1801.

[25] S. Daulton, M. Balandat, and E. Bakshy, *Differentiable expected hypervolume improvement for parallel multi-objective Bayesian optimization*, Advances in Neural Information Processing Systems, 2020-December (2020), pp. 1–30.

[26] ———, *Parallel Bayesian Optimization of Multiple Noisy Objectives with Expected Hypervolume Improvement*, Advances in Neural Information Processing Systems, 3 (2021), pp. 2187–2200.

[27] C. M. E., C. M. J., H. M., L. Barchas, J. J., A. A., and C. I. M., *Relative roles of TGF-β1 and Wnt in the systemic regulation and aging of satellite cell responses.*, Aging Cell, (2009).

[28] K. J. Eberhart R., *A new optimizer using particle swarm theory*, Proceedings of the 6th International Symposium on Micro Machine and Human Science, (1995).

[29] J. M. Edgar, Y. S. Michaels, and P. W. Zandstra, *Multi-objective optimization reveals time- and dose-dependent inflammatory cytokine-mediated regulation of human stem cell derived T-cell development*, npj Regenerative Medicine, 7 (2022).

[30] E. A. M. M. M. A. M. C. N. B. H. Far, (2013).

[31] P. I. Frazier, *A Tutorial on Bayesian Optimization*, (2018), pp. 1–22.

[32] A. A. Giunta, S. F. Wojtkiewicz, and M. S. Eldred, *Overview of modern design of experiments methods for computational simulations*, in 41st Aerospace Sciences Meeting and Exhibit, 2003, pp. 1–17.

[33] J. Gu, G. Y. Li, and Z. Dong, *Hybrid and adaptive meta-model-based global optimization*, Engineering Optimization, 44 (2012), pp. 87–104.

[34] J. Havel, H. Link, M. Hofinger, E. Franco-Lara, and D. Weuster-Botz, *Comparison of genetic algorithms for experimental multi-objective optimization on the example of medium design for cyanobacteria*, Biotechnology Journal, 1 (2006), pp. 549–555.

[35] D. Higdon, J. D. Mcdonnell, N. Schunck, and S. M. Wild, *A Bayesian Approach for Parameter Estimation and Prediction Using a Computationally Intensive Model*, J. Phys. G: Nucl. Phys, (2014).

[36] C. J. Hopfe, *Uncertainty and sensitivity analysis in building performance simulation for decision support and design optimization*, PhD thesis, Technische Universiteit Eindhoven, 2009.

[37] Invitrogen, *AlamarBlue Assay Manual*, tech. rep.

[38] K. J., *Parego: a hybrid algo- rithm with on-line landscape approximation for expensive multiobjective optimization problems*, IEEE Transactions on Evolutionary Computation, (2006), p. 50–66.

[39] M. K. J., T. D., M. S., and P. G. K., *Hepatocyte growth factor affects satellite cell activation and differentiation in regenerating skeletal muscle.*, American Journal of Physiology - Cell Physiology, (2000).

[40] S. J., Y. Y., G. L., K. Y., and K. H., *Involvement of Ras and Ral in chemotactic migration of skeletal myoblasts.*, Molecular and Cellular Biology, (2000).

[41] A. Z. J. Mockus, V. Tiesis, *The application of Bayesian methods for seeking the extremum*, Toward Global Optimization, (1978).

[42] P. Jiang, C. A. Shoemaker, and X. Liu, *Time-varying hyperparameter strategies for radial basis function surrogate-based global optimization algorithm*, IEEE International Conference on Industrial Engineering and Engineering Management, 2017-Decem (2018), pp. 984–988.

[43] S. T. W. Jin R., Chen W., *Comparative Studies of Meta- modeling Techniques Under Multiple Modeling Criteria*, Struct. Multidiscip. Optim, (2001).

[44] B. Jones, K. Allen-Moyer, and P. Goos, *A-optimal versus D-optimal design of screening experiments*, Journal of Quality Technology, 53 (2021), pp. 369–382.

[45] D. R. Jones, M. Schonlau, and W. J. Welch, *Efficient Global Optimization of Expensive Black-Box Functions*, Journal of Global Optimization, 13 (1998), pp. 455–492.

[46] O. K., S. P., P. A. K., and M. B. N., *Modelling of nutrient mist reactor for hairy root growth using Artificial neural network*, Eur. J. Sci. Res., (2013), p. 516–526.

[47] A. P. Kalyanomy DEB, *A fast and elitist multi-objective genetic algoritm:NSGA -II*, 6 (2001), pp. 182–197.

[48] K. Kanaga, A. Pandey, S. Kumar, and Geetanjali, *Multi-objective optimization of media nutrients for enhanced production of algae biomass and fatty acid biosynthesis from Chlorella pyrenoidosa NCIM 2738*, Bioresource Technology, 200 (2015), pp. 940–950.

[49] I. V. Kathryn Chaloner, *Bayesian Experimental Design Review*, Statistical Science, 10 (1996), pp. 273–304.

[50] A. M. Kolkmann, M. J. Post, M. A. Rutjens, A. L. van Essen, and P. Moutsatsou, *Serum-free media for the growth of primary bovine myoblasts*, Cytotechnology, 72 (2020), pp. 111–120.

[51] L. Kotthoff, H. Wahab, and P. Johnson, *Bayesian Optimization in Materials Science: A Survey*, (2021), pp. 1–15.

[52] H.-h. Kuo, X. Gao, J.-m. Dekeyser, K. A. Fetterman, E. A. Pinheiro, C. J. Weddle, M. V. Orman, M. Romero-tejeda, M. Jouni, M. Blancard, T. Magdy, C. Epting, A. L. George, and P. W. Burridge, *Negligible-Cost and Weekend-Free Chemically Defined Human iPSC Culture Hui-Hsuan*, (2019).

[53] K. Kwack, *A New Non-radioactive Method for IL-2 Bioassay*, (2000), pp. 575–578.

[54] B. Letham, B. Karrer, G. Ottoni, and E. Bakshy, *Constrained Bayesian optimization with noisy experiments*, Bayesian Analysis, (2019).

[55] Malinowski, *Comparative Study of Derivative Free Optimization Algorithms*, Industrial Informatics, 7 (2011), pp. 592–600.

[56] R. T. Marler and J. S. Arora, *Survey of multi-objective optimization methods for engineering*, Structural and Multidisciplinary Optimization, 26 (2004), pp. 369–395.

[57] Meng, S. Chen, T. Lao, D. Liang, and N. Sang, *Nitrogen anabolism underlies the importance of glutaminolysis in proliferating cells*, Cell Cycle, 9 (2010), pp. 3921–3932.

[58] E. N. O. Neill, J. C. Ansel, G. A. Kwong, M. E. Plastino, J. Nelson, K. Baar, and D. E. Block, *Spent media analysis suggests cultivated meat media will require species and cell type optimization.*

[59] E. N. O'Neill, Z. A. Cosenza, K. Baar, and D. E. Block, *Considerations for the development of cost-effective cell culture media for cultivated meat production*, Comprehensive Reviews in Food Science and Food Safety, 20 (2021), pp. 686–709.

[60] K. Phelan and K. M. May, *Basic techniques in mammalian cell tissue culture*, Current Protocols in Toxicology, 2016 (2016), pp. A.3B.1–A.3B.22.

[61] M. Poloczek, J. Wang, and P. I. Frazier, *Multi-information source optimization*, in Advances in Neural Information Processing Systems, 2017.

[62] ———, *Multi-information source optimization, Supplementary Material*, Advances in Neural Information Processing Systems, (2017).

[63] M. Post and J. F. Hocquette, *New Sources of Animal Proteins; In Vitro Meat*, Elsevier Ltd, 2017.

[64] S. Razavi and H. V. Gupta, *A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory*, Water Resources Research, 52 (2016), pp. 423–439.

[65] ———, *A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. Application*, Water Resources Research, 52 (2016), pp. 440–455.

[66] R. G. Regis and C. A. Shoemaker, *Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization*, Engineering Optimization, 45 (2012), pp. 529–555.

[67] L. M. Rios and N. V. Sahinidis, *Derivative-free optimization: A review of algorithms and comparison of software implementations*, Journal of Global Optimization, 56 (2012), pp. 1247–1293.

[68] K. S., B. V., P. V., T. R., T. C. K. M., and G. V. D., *Maximizing the native concentration and shelf life of protein: a multiobjective optimization to reduce aggregation.*, Appl. Microbiol. Biotechnol., (2011), p. 99–108.

[69] S. Saval, L. Pablos, and S. Sanchez, *Optimization of a culture medium for streptomycin production using response-surface methodology*, Bioresource Technology, 43 (1993), pp. 19–25.

[70] E. Schulz, M. Speekenbrink, and A. Krause, *A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions*, Journal of Mathematical Psychology, 85 (2018), pp. 1–16.

[71] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, *Taking the human out of the loop: A review of Bayesian optimization*, Proceedings of the IEEE, 104 (2016), pp. 148–175.

[72] A. R. Shazid Md Sharker, *A Review on the Current Methods of Chinese Hamster Ovary (CHO) Cells Cultivation for the Production of Therapeutic Protein*, Curr Drug Discov Technol, (2021), pp. 354–364.

[73] L. Shu, P. Jiang, and Y. Wang, *A multi-fidelity Bayesian optimization approach based on the expected further improvement*, Structural and Multidisciplinary Optimization, 63 (2021), pp. 1709–1719.

[74] V. Singh, S. Haque, R. Niwas, A. Srivastava, M. Pasupuleti, and C. K. M. Tripathi, *Strategies for Fermentation Medium Optimization: An In-Depth Review*, Frontiers in Microbiology, 7 (2017), pp. 1–16.

[75] S. P. Smith, *Differentiation of the cholesky algorithm*, Journal of Computational and Graphical Statistics, 4 (1995), pp. 134–147.

[76] A. Souza, L. Nardi, L. B. Oliveira, K. Olukotun, M. Lindauer, and F. Hutter, *Bayesian Optimization with a Prior for the Optimum*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12977 LNAI (2021), pp. 265–296.

[77] E. A. Specht, D. R. Welch, E. M. Rees Clayton, and C. D. Lagally, *Opportunities for applying biomedical production and manufacturing methods to the development of the clean meat industry*, Biochemical Engineering Journal, 132 (2018), pp. 161–168.

[78] L. Specht, *An analysis of culture medium costs and production volumes for cell-based meat*, tech. rep., 2018.

[79] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, *Gaussian process optimization in the bandit setting: No regret and experimental design*, Proc. Int. Conf. Mach. Learn., (2010), p. 1015–1022.

[80] A. J. Stout, *Simple and effective serum-free medium for sustained expansion of bovine satellite cells for cell cultured meat*, (2021), pp. 1–18.

[81] K. Swersky and R. P. Adams, *Multi-Task Bayesian Optimization*, pp. 1–9.

[82] S. Takeno, H. Fukuoka, Y. Tsukada, T. Koyama, M. Shiga, I. Takeuchi, and M. Karasuyama, *Multi-fidelity Bayesian optimization with max-value entropy search*, arXiv, (2019).

[83] B. D. Tracey and D. H. Wolpert, *Upgrading from gaussian processes to student's-T processes*, AIAA Non-Deterministic Approaches Conference, 2018, 0 (2018).

[84] J. van der Valk, K. Bieback, C. Buta, B. Cochrane, W. G. Dirks, J. Fu, J. J. Hickman, C. Hohensee, R. Kolar, M. Liebsch, F. Pistollato, M. Schulz, D. Thieme, T. Weber, J. Wiest, S. Winkler, and G. Gstraunthaler, *Fetal Bovine Serum (FBS): Past - Present - Future*, Altex, 35 (2018), pp. 99–118.

[85] J. van der Valk, D. Brunner, K. De Smet, Å. Fex Svenningsen, P. Honegger, L. E. Knudsen, T. Lindl, J. Noraberg, A. Price, M. L. Scarino, and G. Gstraunthaler, *Optimization of chemically defined cell culture media - Replacing fetal bovine serum in mammalian in vitro methods*, Toxicology in Vitro, 24 (2010), pp. 1053–1063.

[86] S. Verbruggen, D. Luining, A. van Essen, and M. J. Post, *Bovine myoblast cell production in a microcarriers-based system*, Cytotechnology, 70 (2018), pp. 503–512.

[87] A. Villegas, J. Pablo, D. Aragón, and M. Arias, *Determination of the optimal operation conditions to maximize the biomass production in plant cell cultures of thevetia peruviana using multi-objective optimization*, 2014.

[88] G. G. Wang and S. Shan, *Review of metamodeling techniques in support of engineering design optimization*, Journal of Mechanical Design, Transactions of the ASME, 129 (2007), pp. 370–380.

[89] J. Wang, S. C. Clark, E. Liu, and P. I. Frazier, *Parallel Bayesian Global Optimization of Expensive Functions*, Operations Research, 68 (2020), pp. 1850–1865.

[90] D. Weuster-Botz, *Experimental Design for Fermentation Media Development: Statistical Design or Global Random Search?*, Journal of Bioscience and Bioengineering, 90 (2000), pp. 473–483.

[91] J. T. Wilson, R. Moriconi, F. Hutter, and M. P. Deisenroth, *The reparameterization trick for acquisition functions*, (2017), pp. 1–7.

[92] G. Zhang and D. E. Block, *Integration of Data Mining Into a Nonlinear Experimental Design Approach for Improved Performance*, AIChE Journal, 55 (2009), pp. 3017–3021.

[93] ———, *Using highly efficient nonlinear experimental design methods for optimization of Lactococcus lactis fermentation in chemically defined media*, Biotechnology Progress, 25 (2009), pp. NA–NA.

[94] G. Zhang, D. A. Mills, and D. E. Block, *Development of chemically defined media supporting high-cell-density growth of lactococci, enterococci, and streptococci*, Applied and Environmental Microbiology, 75 (2009), pp. 1080–1087.

[95] G. Zhang, M. M. Olsen, and D. E. Block, *New experimental design method for highly nonlinear and dimensional processes*, AIChE Journal, 53 (2007), pp. 2013–2025.

[96] J. Zhang, S. Chowdhury, and A. Messac, *An adaptive hybrid surrogate model*, Structural and Multidisciplinary Optimization, 46 (2012), pp. 223–238.