Chemically-informed data-driven optimization (ChIDDO): Leveraging physical models and Bayesian learning to accelerate chemical research

Journal:	*Reaction Chemistry & Engineering*
Manuscript ID	RE-ART-01-2022-000005.R1
Article Type:	Paper
Date Submitted by the Author:	17-Feb-2022
Complete List of Authors:	Frey, Daniel; New York University Tandon School of Engineering, Chemical and Biological Engineering Shin, Juhee; New York University Tandon School of Engineering, Chemical and Biological Engineering Musco, Christopher; New York University Tandon School of Engineering, Computer Science and Engineering Modestino, Miguel; New York University Tandon School of Engineering, Chemical and Biological Engineering

ARTICLE

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Chemically-informed data-driven optimization (ChIDDO): Leveraging physical models and Bayesian learning to accelerate chemical research†

Daniel Frey[a], Juhee Shin[a], Christopher Musco[b], Miguel A. Modestino*[a]

Current methods of finding optimal experimental conditions, Edisonian systematic searches, often inefficiently evaluate suboptimal design points and require fine resolution to identify near optimal conditions. For expensive experimental campaigns or those with large design spaces, the shortcomings of the status quo approaches are more significant. Here, we extend Bayesian optimization (BO) and introduce a chemically-informed data-driven optimization (ChIDDO) approach. This approach uses inexpensive and low-fidelity information obtained from physical models of chemical processes and are subsequently combined with expensive and high-fidelity experimental data to optimize a common objective function. Using common optimization benchmark objective functions, we describe scenarios in which the ChIDDO algorithm outperforms traditional BO approach, and then implement the algorithm on a simulated electrochemical engineering optimization problem.

Introduction

Edisonian search approaches are widely used in the chemical sciences to discover reactions, process conditions, material compositions, or product formulations with optimal performance for their intended application. These experimental design methods rely on the generation of grids of variables where experimentally accessible conditions are systematically and/or combinatorically explored. While these methods are simple to implement, they often evaluate a suboptimal parameter space where the quality of information derived depends on the numbers of combinations of variables explored, slowing and sometimes preventing the identification of optimal conditions.[1] These shortcomings represent significant impediments for expensive experimental campaigns (*e.g.*, during process scale-up, in fine chemicals or pharmaceuticals) or those with large design spaces that can only afford the implementation of coarse experimental grids, underscoring the need for more efficient experimental optimization methods[2].

Bayesian optimization (BO) has been widely implemented in different fields of research to accelerate experimental optimization.[3-9] BO methods use a surrogate model (SM) that describes an objective function and its probability distribution in the design space to guide the optimization campaign. Each time new experimental data is obtained, the SM is updated to increase its accuracy. In this way, Bayesian statistics and reasoning can be used to select the most informative sequence of experiments and accelerate optimization campaigns.[10] In recent years, BO has been implemented for various applications in the chemical sciences including materials discovery and prediction of their properties,[11-21] design of reactors and chemical processes,[22-33] and the optimization of energy storage materials and devices.[34-38] Data-driven optimization methods such as BO learn and evolve with new experimental data, but they lack *a priori* knowledge of the physical laws that dictate the behavior of the chemical system under study. This can result in the need for large experimental campaigns to accurately model and find the optimal combination of parameters for a given objective function. On the other hand, physical models (*e.g.*, density functional theory, molecular dynamics, continuum models, etc.) could be used to identify optima without the need to perform experimental searches, but they often lack the accuracy to effectively capture the complexity of real systems or require inaccessibly-large computational power. Given the advantages and shortcomings of both optimization approaches, there is an opportunity to leverage *a priori* chemical knowledge in data-driven optimization to reduce the data needs and allow for faster identification of optima.

Herein, we introduce a chemically-informed data-driven optimization (ChIDDO) approach, which is a type of multi-information source optimization (MISO), where inexpensive and low-fidelity information obtained from physical models of chemical processes are combined with high-fidelity experimental data to optimize a common objective function. In

[a] *New York University, Tandon School of Engineering, Department of Chemical and Biomolecular Engineering, 6 Metrotech Center, Brooklyn, NY, E-mail: modestino@nyu.edu; Tel: +1 646-997-3594*
[b] *New York University, Tandon School of Engineering, Computer Science and Engineering, 370 Jay Street, Brooklyn, NY*

† Electronic Supplementary Information (ESI) available: Acquisition function descriptions, objective function descriptions, and supplementary results. See DOI: 10.1039/x0xx00000x

this study, we leverage simulated data to develop a ChIDDO approach that can be implemented broadly in experimental campaigns. While MISO algorithms have been previously implemented to improve BO in computational problems[39-41], the implementation of ChIDDO can extend these advantages to chemical experimentation. In addition, we introduce a new acquisition function, modified ranked batch (MRB) that could improve the selection of a batch of experiments[42].

Experimental

BO Algorithm Description

BO algorithms consist of two main components: a SM and an acquisition function. The SM is used to predict the value of the experimental objective function, y^{pred}, for any set of conditions, $\boldsymbol{x}$. $\boldsymbol{x}$ is a vector of length d, the number of dimensions in the design space. $\boldsymbol{x}$ is bounded by lower and upper bounds for each dimension, $\boldsymbol{x}_{LB}$ and $\boldsymbol{x}_{UB}$, which are arrays of the same dimensionality of $\boldsymbol{x}$. The SM is trained using N^{exp} experimental evaluations of the experimental objective function, which results in vector $\boldsymbol{y}^{exp}$ corresponding to $\boldsymbol{X}^{exp}$. $\boldsymbol{X}^{exp}$ is a matrix with N^{exp} rows and d columns. The ith row of $\boldsymbol{X}^{exp}$, which we denote x_i^{exp}, corresponds to a d dimensional parameter vector to be evaluated. $\boldsymbol{y}^{exp}$ is an array of N^{exp} evaluations of the experimental objective function at each condition, x_i^{exp}, in $\boldsymbol{X}^{exp}$. In this study we use a Gaussian process regressor (GPR) with the radial basis function kernel as the SM.

An acquisition function is used to select the next design condition(s) to evaluate, $\boldsymbol{x}^{next}$, based on how informative the design conditions will be in the goal of optimizing the cost function. Here, we can choose to select a single design condition or a batch of conditions. In the chemical sciences, it is often convenient to run multiple experiments in parallel based on equipment capabilities, so we chose to focus on selecting batches of design conditions. Many different acquisition functions for BO have been developed, and three of the most common are expected improvement (EI),[43] probability of improvement (PI),[44] and upper confidence bound (UCB).[45] In addition to these, we have developed a modified ranked-batch (MRB) mode sampling function inspired by the work of Cardoso et al.[42] The equations for each of acquisition functions are provided in the Electronic Supplemental Information (ESI). An acquisition function uses the current information, $\boldsymbol{X}^{exp}$ and $\boldsymbol{y}^{exp}$, and the SM predictions to calculate how informative a possible design condition is expected to be based on the criteria for the respective acquisition function. To determine the most informative design point to sample next, a maximization method was used to find a local maximum of the acquisition function score. This process was repeated 25 times at different initiation points to get closer to the global maximum solution. The design point with the maximum score was subsequently added to $\boldsymbol{x}^{next}$. For this study a minimization method was used and the negative of the acquisition function score was minimized. The minimization method was the L-BFGS-B method from the scipy.optimize.minimize package. Depending on the batch size used in the optimization campaign, n_b, multiple design conditions can be added to $\boldsymbol{x}^{next}$ by repeating this acquisition function maximization step. After $\boldsymbol{x}^{next}$ is selected, the experimental objective function value(s) are determined to obtain $\boldsymbol{y}^{next}$. Subsequently, $\boldsymbol{x}^{next}$ and $\boldsymbol{y}^{next}$ are appended to $\boldsymbol{x}^{exp}$ and $\boldsymbol{y}^{exp}$.

The EI, PI, and UCB algorithms were run based on their implementation in the modAL active learning framework,[46] which is described in the ESI. The general framework for the BO algorithms presented was also based on the modAL framework. The MRB acquisition function calculated a score consisting of three normalized parameters: a distance score, Δ, an uncertainty score, Γ, and the objective function prediction, Ω. The distance score was calculated as:

$$\Delta = 1 - 1/\left(1 + \min\sqrt{\sum_{i=1}^{d}\left(\boldsymbol{x}_i - \boldsymbol{x}_i^{exp}\right)^2}\right) \quad (1)$$

where $\min\sqrt{\sum_{i=1}^{d}\left(\boldsymbol{x}_i - \boldsymbol{x}_i^{exp}\right)^2}$ is the minimum distance between the proposed set of conditions, $\boldsymbol{x}$, and each of the known sets of conditions, $\boldsymbol{x}^{exp}$. The uncertainty score, Γ, is the standard deviation of the GPR prediction at $\boldsymbol{x}$ normalized compared to the maximum and minimum observed standard deviation. The objective function prediction, Ω, is y^{pred} at $\boldsymbol{x}$ normalized compared to the maximum and minimum

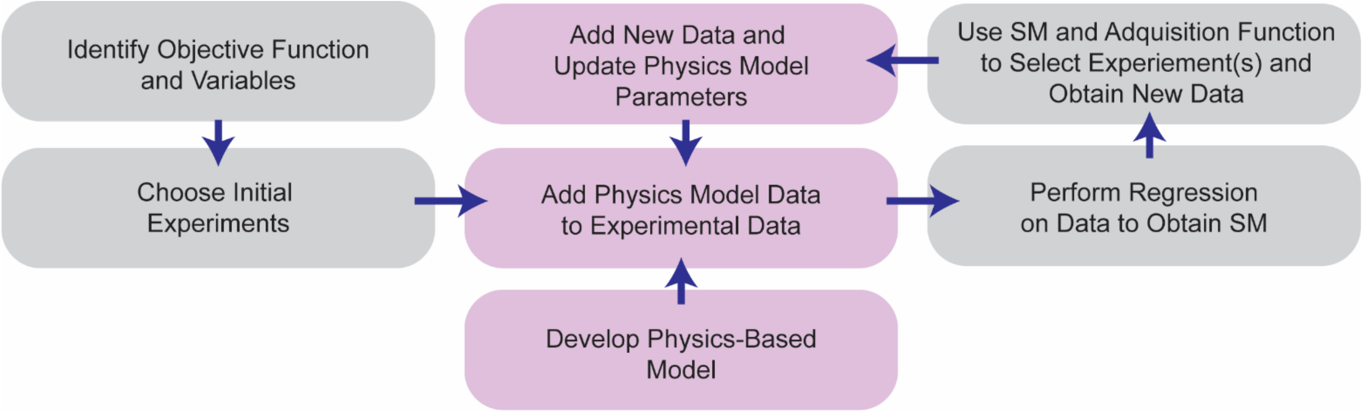
Identify Objective Function and Variables
Choose Initial Experiments
Add New Data and Update Physics Model Parameters
Add Physics Model Data to Experimental Data
Develop Physics-Based Model
Use SM and Adquisition Function to Select Experiement(s) and Obtain New Data
Perform Regression on Data to Obtain SM

Figure 1. Process diagram of the ChIDDO algorithm. The purple blocks correspond to the algorithm steps required to incorporate the physical model. The gray blocks correspond to steps related to experimental data acquisition.

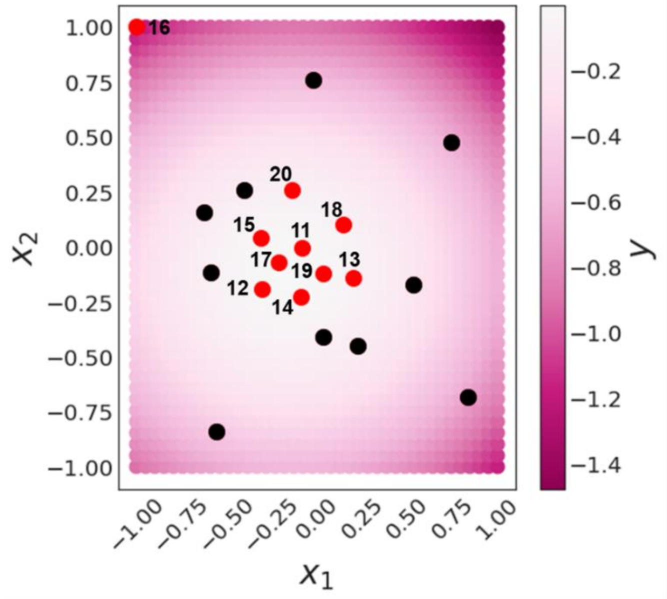

Figure 2. Example of the decision-making process of the BO algorithm using the Sphere objective function. MRB was used as the acquisition function, there were 10 initial random points (black dots), and subsequent points (in red and labelled in order) were selected in batches of 3.

observed prediction. The score that is calculated at each step in the minimization process for the respective $\boldsymbol{x}$ is:

$$\text{Score} = \beta\Delta + \beta\Gamma + \Omega \qquad (2)$$

where β is a tradeoff value. A high value of β encourages more exploration — i.e., encourages searching unknown areas of the design space. A lower value of β encourages exploitation — i.e., searching locally near the current maximum prediction. All of the acquisition functions include a tradeoff value that decreases as more experiments are run, moving from exploration to exploitation. For MRB, β changes linearly from 1 to 0. For UCB, β changes linearly from 4 to 0. For PI and EI, β changes logarithmically from around 0.05 to 1x10^{-7}.

To initiate the algorithm, N_{init} evenly distributed random points were chosen as the initial set of experimental conditions. Our results show robust performance when a random initialization approach is implemented, but other methods that ensure good spatial coverage over the design space and incorporate a degree of randomness could be implemented. The random initialization approach was sone by choosing random experiments to perform without considering the positions of the other initial experiments. In other words, there was no space-filling model for this initialization approach. For the BO algorithm without use of a physics model (referred to as BO from this point on), only these initial points, $\boldsymbol{x}_{init}^{exp}$, were fit by the GPR to generate the SM. After each batch of BO, ($\boldsymbol{x}^{exp}$, $\boldsymbol{y}^{exp}$) increases in size by the batch size, n_b. For the ChIDDO algorithm, before ($\boldsymbol{x}^{exp}$, $\boldsymbol{y}^{exp}$) are passed to the GPR, a certain number of design points from the *a priori* physics model ($\boldsymbol{x}^{phys}$,$\boldsymbol{y}^{phys}$) are appended to ($\boldsymbol{x}^{exp}$, $\boldsymbol{y}^{exp}$). The size of ($\boldsymbol{x}^{phys}$, $\boldsymbol{y}^{phys}$) decreases as the number of experiments that are run increases. For example, if it was decided that a total of 50 experiments would be run before stopping the ChIDDO optimization campaign, N_{total}, and it was chosen to start with 10 experimental points, the ChIDDO algorithm would add $\left(N_{total} - \text{size}(\boldsymbol{x}^{\boldsymbol{exp}})\right)$ data points (40 in this case) calculated from the physics model. These added points were uniformly distributed random design points between the upper and lower bounds. This method allowed for the incorporation of knowledge of the chemical system under study to help guide the initial choice of experiments when less experimental data is available, and progressively increases the amount of experimental data used to generate SM as more empirical evidence becomes available. A general algorithm flowchart is shown in Figure 1 and an example of the decision process in action is shown in Figure 2.

Benchmark Objective Functions

Common objective functions for optimization benchmarking were selected, filling in as a representation of a chemical sciences objective function, and they are described further in the ESI. Each objective function has its own set of parameters that affect the specific shape of the objective function. For example, for the Sphere objective function (an ellipse):

$$f(x) = \sum_{i=1}^{d} P_i(x_i + P_{i+d})^2 \qquad (3)$$

the variable, P, is an array of the $2d$ parameters. For each objective function there is a base set of parameters that results in a base-case objective function shape. To obtain alternate models of the objective functions, P can be randomly perturbed around the base parameters. For all of the studies, 20 alternate models were used as the experimental objective functions.

Depending on the specific objective function, we studied 2-, 3-, 4- and 6-dimensional spaces. Unless otherwise specified, the experimental objective function values, y^{exp}, were exactly equal to the objective function calculation, given the set of parameter values.

Under conditions when noise was added, the objective function values were calculated as:

$$y_i^{noise} = y_i^{exp} + [(y_{max} - y_{min})(2\,\text{rand}(0,1) - 1)]\eta \qquad (4)$$

where y_{max} is the maximum value of the objective function, y_{min} is the minimum value of the objective function, and η is the noise level, defined as the maximum allowable value that could be added or subtracted from y_i^{exp}, which can be viewed as a percentage of the range of y. rand(0,1) is a random variable drawn uniformly from 0 to 1. The η values that were tested were 0.025, 0.05, and 0.1.

Updating the Physics Model Parameters in ChIDDO

Each physics model was initially defined by a set of base-case parameters that could be updated during the optimization process. These base-case models were used as the *a priori* knowledge in the ChIDDO algorithm. The base-case model parameters used are provided in the ESI. Since the initial model parameters are only an estimate, the parameters were updated after each batch of experiments based on the

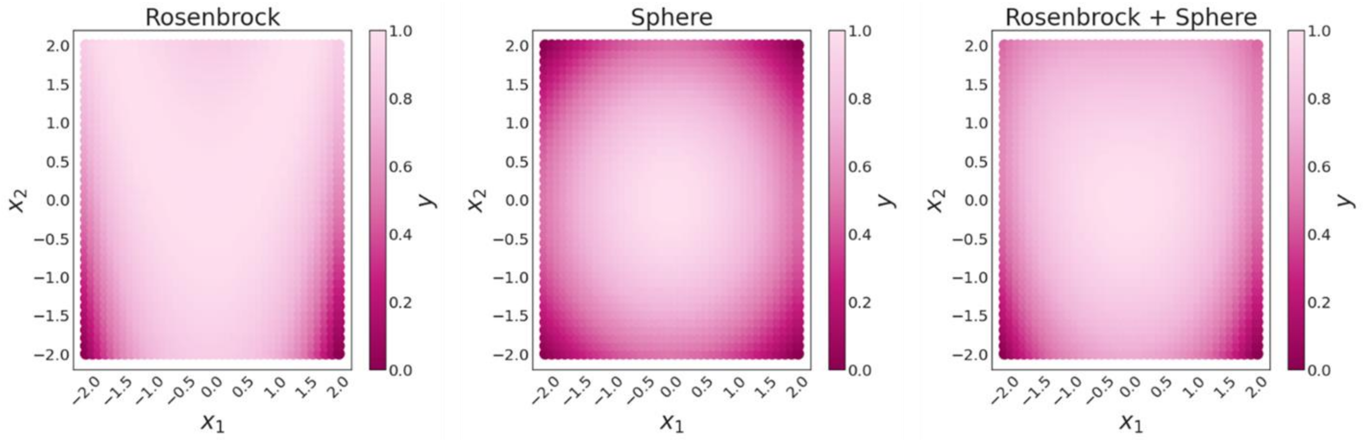

Figure 3. Example of the addition of two dissimilar objective functions. The Rosenbrock function and Sphere function are shown using their respective base case parameters.

new experimental observations. The parameters were updated by using a non-linear least square error regressor to minimize the error between the experimental data and the physics model. The updated model parameters were then used to calculate ($\boldsymbol{x}^{phys}$, $\boldsymbol{y}^{phys}$) for the following batch. With the relatively simple experimental objective functions used in this study, this method of updating the parameters is appropriate. However, for complex objective functions with many different unknown parameters (e.g., continuum models, molecular dynamics), other methods for updating parameters may be needed.

Baseline Search Algorithms

Two different baseline search methods were tested: a grid search and a random search. 100 trials were done for each search method and *N* experimental points were selected for each trial. For the grid search, *N* equally spaced experimental points were selected sequentially between the lower and upper bounds of each variable. For the random search, conditions were chosen at random from the uniform design space defined by the upper and lower bounds. This random search did not consider the locations of the previous selections, so it was possible to have poor representation of the design space (*i.e.,* clustering of points).

Simplified Physics Model

To study how the algorithm performs when a physical model does not accurately represent the chemical process of interest, an objective function was built as a linear combination of two physics models, while the physics model used in ChIDDO was based on only one of them. The values of the combined objective function were calculated as:

$$y^{mixed} = \mathrm{r} * \mathrm{y}_1 + (1-r) * \mathrm{y}_2 \tag{5}$$

where *r* is the mixing ratio, y_1 is the value of the first objective function, and y_2 is the value of the second objective function. For example, the Rosenbrock function could be added to the Sphere function with an *r* of 0.9. In this case, the objective function would more closely, but not perfectly, resemble the Rosenbrock function, as seen in Figure 3.

Electrochemical Model Description

To simulate testing the BO/ChIDDO algorithms on an experimental chemical system, a hypothetical electrochemical system of reactions was considered:

$$\mathrm{A} + \mathrm{B} + \mathrm{e}^- \rightarrow C \tag{6}$$

$$2\mathrm{A} + \mathrm{B}\ + e^- \rightarrow D \tag{7}$$

$$3\mathrm{A} + \mathrm{B} + \mathrm{e}^-\ \rightarrow E \tag{8}$$

$$2\mathrm{B} + \mathrm{e}^-\ \rightarrow F \tag{9}$$

The chosen reaction resembles the electrohydrodymerization of acrylonitrile to adiponitrile, the largest organic electrosynthetic process practices in industry.[1, 47, 48]

The rates of these reactions were modelled by Butler-Volmer kinetics in the form:

$$J_i = J_i^0 \prod\nolimits_{j=A,B} (c_j^{surf})^{\gamma_{ij}} \exp\big((\alpha_i F \eta_i)/RT\big) \tag{10}$$

where J_i is the current density of the respective reaction, *i*, j_i^0 is the exchange current density of the respective reaction, c_j^{surf} is the electrode surface concentration of the respective reactant, *j* (A or B), γ_{ij} is the order of reaction for the respective reactant and reaction, α_i is the average charge transfer coefficient between the two reactants for the respective reaction, *F* is Faraday's constant, η_i is the overpotential for the respective reaction, *R* is the gas constant, and *T* is the temperature in K.

The reactions are simulated in a 1-D domain, representing the diffusion boundary layer, on one end bounded by the bulk electrolyte solution and the other end the electrode surface. The Nernst-Planck equation was used to model the concentration change of each species using diffusion, migration, and generation terms:

$$\frac{\partial \mathrm{c_j}}{\partial \mathrm{t}} = \frac{\sum J_{ij}}{F * \Delta x} + D\frac{\partial^2 c_j}{\partial x^2} + \frac{Dz}{RT}\frac{\partial c_j}{\partial x}\frac{\partial \Phi}{\partial x} \tag{11}$$

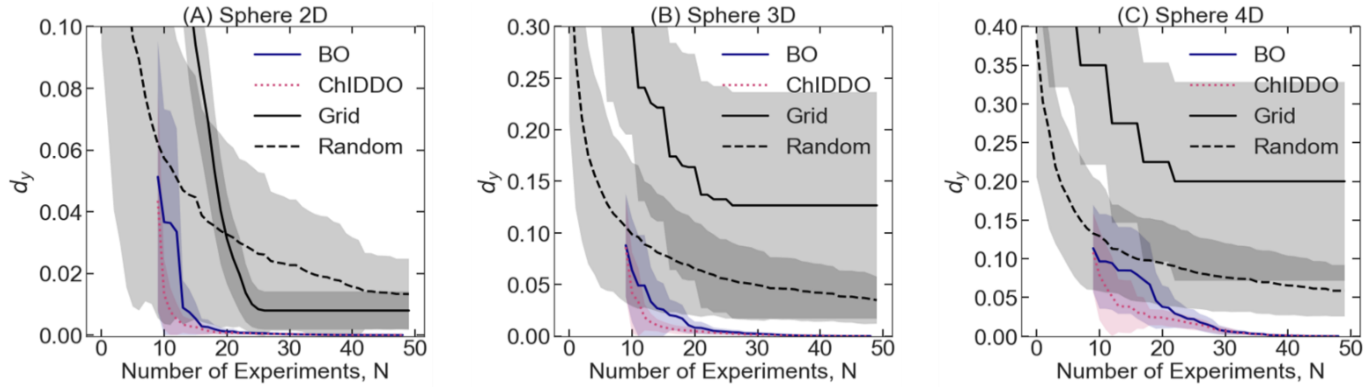

Figure 4. d_y versus number of experiments, *N*, comparing BO and ChIDDO with the Edisonian random and grid search. (A) 2D Sphere, (B) 3D Sphere, (C) 4D Sphere. For each curve, 20 separate searches, *S*, were performed, and the average of the results are the lines shown. The shadow around each of the lines represents the standard deviation. For each of the BO/ChIDDO experiments, the MRB acquisition function was used. The objective function parameter information is provided in the Supplemental Information.

Where c_j is the concentration of the respective reactant or product, *j*, $\sum j_{ij}$ is the sum of the production/consumption rates for the respective species over each reaction, *i*, that the species participates in, Δx is the spacing between each point in the model, *z* is the charge of the species (chosen to be 1), and $\partial \Phi / \partial x$ is the potential gradient.

The Faradaic efficiency (FE) of product D was the value to be optimized. FE is a metric that measures how much of the current participates in the desired reaction. In this case, FE is calculated by dividing the amount of D produced by the total of all produced species (including D). In this system, the concentrations of the reactants could have a large effect on the FE. Therefore, the optimization variables in the 2D design space for this reaction were the bulk concentrations of reactants A and B. Due to the reaction rates and reaction orders of the different reactions, an optimal set of reactant concentrations could be located in the design space.

Data Availability

The code used for all the experiments can be found in a public repository[49].

Results

BO Improvements over Edisonian Approach

To demonstrate the advantages of implementing a BO strategy over an Edisonian approach, we studied the performance of the different optimization approaches on common benchmark functions. Each of these benchmark functions has a different shape and optimization complexity, and by running the algorithms on these different objective functions, we attempted to gain insight into the behavior of the different algorithms. For conciseness, here we present the results for various optimization runs using the Sphere function (Figure 4). A full list of the objective functions, their equations, the base parameters, and optimization results can be found in the ESI.

In our framework, we consider experimental sets, ***S***, which consist of N^{exp} number of experiments with conditions $\boldsymbol{x}^{exp}$ resulting in output performance, $\boldsymbol{y}^{exp}$. The purpose of the BO algorithm is to maximize $\boldsymbol{y}^{exp}$ in the fewest number of experiments. The output of the algorithm generates a set of ($\boldsymbol{x}^{exp}$, $\boldsymbol{y}^{exp}$) results that can be plotted and compared to Edisonian experimental sets that follow either a grid or a random search approach. We evaluate two performance metrics: the normalized deviation from the optimum value, d_y, and the minimum distance from the optimum, d_x, identified by each set of experiments. These two quantities are calculated as,

$$\mathrm{d_y} = \min \frac{y_{max} - y^{exp}}{y_{max} - y_{min}} \qquad (12)$$

$$d_x = \min \sqrt{\sum_{i=1}^{d} (\boldsymbol{x}^{exp} - \boldsymbol{x}^{opt})^2} \qquad (13)$$

where y_{max} is the maximum possible value of the cost function within the constraints of the experimental parameters and y_{min} is the minimum possible value of the cost function within the constraints of the experimental parameters.

In the following studies, the different algorithms (BO and ChIDDO) are run on 20 different Sphere functions which serve as simulated experimental objective functions. Since the experimental objective functions are different, there was some variance in the results between the 20 runs. Therefore, the graphs shown in the following figures show the average of the 20 runs as a solid line, and a shadow around the solid line representing the standard deviation of the 20 runs. In Figure 4, d_y is plotted against the number of experiments, *N*, comparing the Edisonian methods with BO and ChIDDO. The plots for d_x can be found in the ESI. Figure 4A shows how the different search algorithms compare using the 2D Sphere objective function. Even for this simple, parabolic function, the systematic grid search and random search underperform comparatively to BO or ChIDDO. d_y after 30 experiments, d_{y30}, were 0.008 and 0.023 for the grid and random search algorithms, respectively. In comparison, d_{y30} for BO and ChIDDO were both on the order of 10^{-3}. As the design space moves to higher dimensions, Figures 4B and 4C show that the differences between the algorithms increase with dimension size. For the 3D Sphere objective function the enhancements

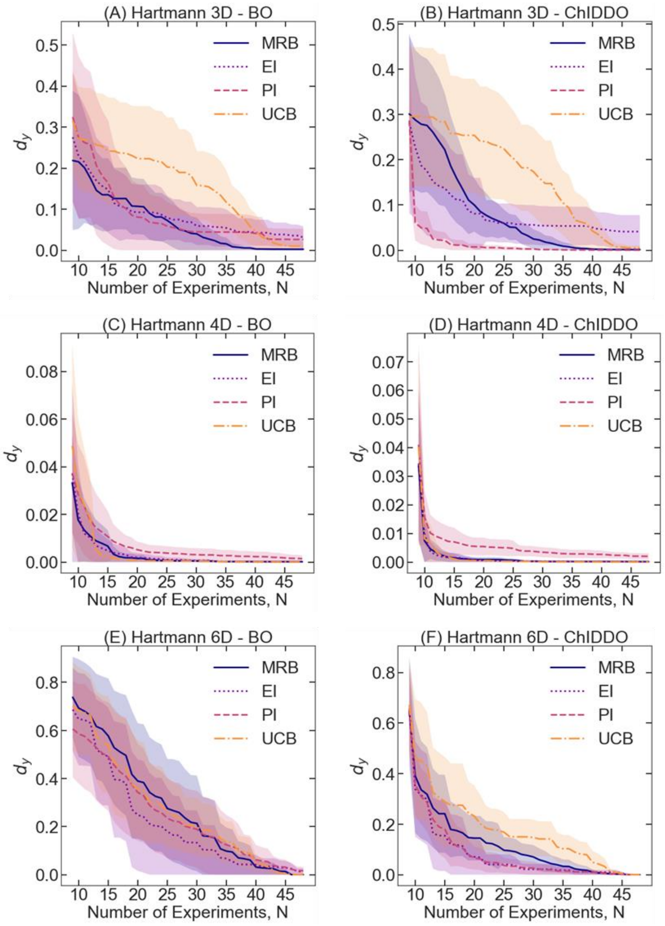

Figure 5. d_y versus number of experiments, *N*, comparing the MRB, EI, PI, and UCB acquisition functions using BO and ChIDDO. (A) 3D Hartmann - BO, (B) 3D Hartmann – ChIDDO, (C) 4D Hartmann – BO, (D) 4D Hartmann – ChIDDO, (E) 6D Hartmann – BO, (F) 6D Hartmann - ChIDDO. For each curve, 25 separate searches, *S*, were performed, and the average of the results are the lines shown. The shadow around each of the lines represents the standard deviation.

are more drastic with d_{y30} being two orders of magnitude smaller for BO compared to the Edisonian algorithms. Because of the larger design space to sample, the grid and random search methods are not capable of searching a fine enough space to find values close to the optimal. It is of interest that the d_{y30} for BO and ChIDDO were very similar, possibly because the Sphere objective function has a well-defined optimum and is therefore easy to identify. As the number of dimensions increases, ChIDDO tends to find near optimal values with fewer experiments than BO. This enhancement is likely because ChIDDO relies on the physics model initially to help more rapidly locate optimal conditions.

Comparison of Different Acquisition Functions

In order to compare how the different acquisition functions behave at identifying optima in objective functions of different dimensionality, Hartmann functions with 3, 4, and 6 dimensions were analyzed and the results are shown in Figure 5. This

objective function was chosen due to its more complex structure (i.e., multiple local optima). Results from other objective functions are provided in the ESI. Figures 5A,B show a comparison of performance when different acquisition functions are used on a 3D Hartmann function. The d_{y30} values for MRB, EI, PI, and UCB using BO were 0.041, 0.069, 0.045, and 0.183, respectively. It appears that all the acquisition functions behaved similarly except for UCB, which shows an order of magnitude worse performance. By the end of the run, it appears that all the acquisition functions reach a similar value of d_y. For the comparison on the 3D Hartmann function using ChIDDO shown in Figure 5B, PI and MRB appeared to perform the best with d_{y30} values of 0.003 and 0.032, respectively, compared with EI (0.056) and UCB (0.182).

Figures 5C,D show the comparison on the 4D Hartmann function using BO and ChIDDO, respectively. Interestingly, all the acquisition functions perform similarly with $d_{y,30}$ values on the order of 10^{-3} or lower and they all reach low values very quickly. The comparison on the 6D Hartmann function is shown in Figures 5E,F. When using BO, all the acquisition functions appear to perform similarly with $d_{y,30}$ values of 0.218 (MRB), 0.134 (EI), 0.195 (PI), and 0.205 (UCB). It is important to note that the standard deviations for the 6D graphs are much larger than for the smaller dimensions. This indicates that the different random starting conditions affected the d_y values more for the 6D space compared with the 3D and 4D spaces, due to the larger complexity of the optimization process with increased dimensionality. Figure 6F shows the comparison using ChIDDO. The $d_{y,30}$ values for MRB, EI, PI, and UCB were 0.079, 0.021, 0.027, and 0.149, respectively. In addition, the standard deviation of d_y is much smaller for ChIDDO than for BO, indicating a more consistent optimization.

When comparing the performance of BO to ChIDDO, it appears that the ChIDDO algorithm performs similarly or better for all of the objective functions. These results show that the ChIDDO algorithm does improve the performance initially, since the physics model information has a larger impact when fewer experiments are available.

Quantifying the Effect of Experimental Noise

So far, we have assumed that experiments run under conditions, x_i, result in exact values of the objective function of interest, $\mathrm{f}(x_i) = y_i$. However, experimental measurements often possess a significant degree of noise. To quantify the effect of the experimental noise and to determine the robustness of the BO and ChIDDO algorithms to noisy experiments, different levels of random noise were added to the objective functions.

Figure 6 compares d_y for different levels of noise using the BO and ChIDDO algorithms with the MRB acquisition function. For the case of the 3D Hartmann using BO (Figure 6A), the d_y for the highest noise level studied (i.e. η=0.1) appears to be slightly higher than the other noise values until about the 37th experiment when d_y approaches the same value for all noise levels. For the 3D Hartmann function using ChIDDO in Figure 6B, the observations are similar to that of BO as the noise had only a small impact on the optimization. Interestingly, the d_y values for the experiments with noise are not substantially different to the experiments without noise, demonstrating the robustness of BO and ChIDDO.

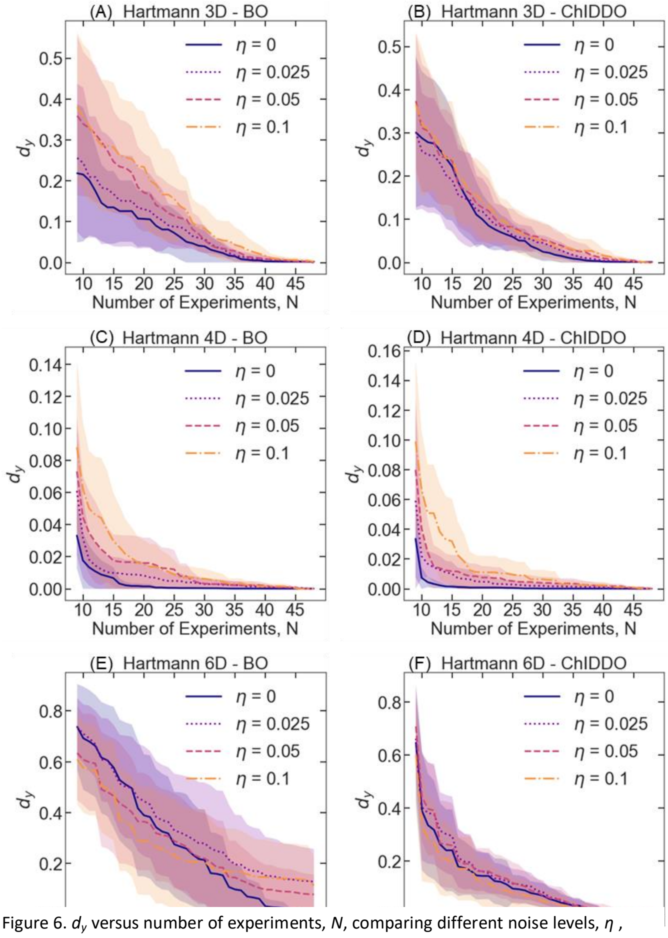

Figure 6. d_y versus number of experiments, *N*, comparing different noise levels, *η* , represented by lines of different colors. (A) 3D Hartmann - BO, (B) 3D Hartmann - ChIDDO, (C) 4D Hartmann - BO, (D) 4D Hartmann – ChIDDO, (E) 6D Hartmann – BO, (F) 6D Hartmann - ChIDDO. For each curve, 25 separate searches, *S*, were performed, and the average of the results are the lines shown. The shadow around each of the lines represents the standard deviation. For each of these studies, the MRB acquisition function was used.

This behavior is also observed for the case of 4D Hartmann function in Figures 6C,D. When using BO, the d_y for η=0.1 remains higher than for the other noise levels until approximately the 32nd experiment when the d_y values start to converge for other noise levels. In the case of the 4D Hartmann function using ChIDDO, the d_y values for each noise level are similar after the 25th experiment. Prior to this, the d_y values for η=0.1 are higher than that of the other noise levels. Contrary to the 3D Hartmann function, the experiments with no noise for both ChIDDO and BO have lower d_y values than the experiments with noise.

Figures 6E,F show the noise comparisons for the 6D Hartmann function. When using BO, all noise levels present similar values for d_y until experiment 30th, and a slightly higher values for η = 0.1 beyond that point. These results indicate that the BO algorithm may be more resistant to noise effects in low-dimensionality design spaces and that overall noise effects are weak within the levels studied. Figure 6F shows that the ChIDDO algorithm performs much better overall for the 6D Hartmann function compared to BO. When ChIDDO is implemented on 3, 4 and 6D Hartmann functions, our observations suggest that noise has only a small impact on the optimization but the values of d_y are lower than those found with BO for a given number of experiments.

Quantifying Effects of Physical Model Accuracy

Experiments in the chemical sciences are often performed under complex geometries, involve multiple kinetic and transport processes, and require molecular-level descriptions of the species involved for an accurate representation. Detailed multidimensional models of these complex chemical systems are often intractable, requiring simplified semi-empirical models that capture with an acceptable level of accuracy the experimental observations. These simplified physical models can still be used in ChIDDO algorithms as they serve as a guide to the optimization and can be complemented and improved by experimental data. To understand how less accurate models affect the performance of our method, we attempted to optimize a mixed objective function that consisted of a linear combination of two functions (Equation 5), while ChIDDO used a physics model that described only one of the functions. Figures 7A and B show d_y as a function of number of experiments for the combination of the 3D Sphere and 3D Hartmann objective functions, while Figures 7C and D present similar results for 6D objective functions. For the example where the Sphere function is used as the physics model, an *r* of 0.1 indicates that the output value for each set of conditions is 10% of the 3D Sphere output value plus 90% of the 3D Hartmann output value. Therefore, a low *r* indicates

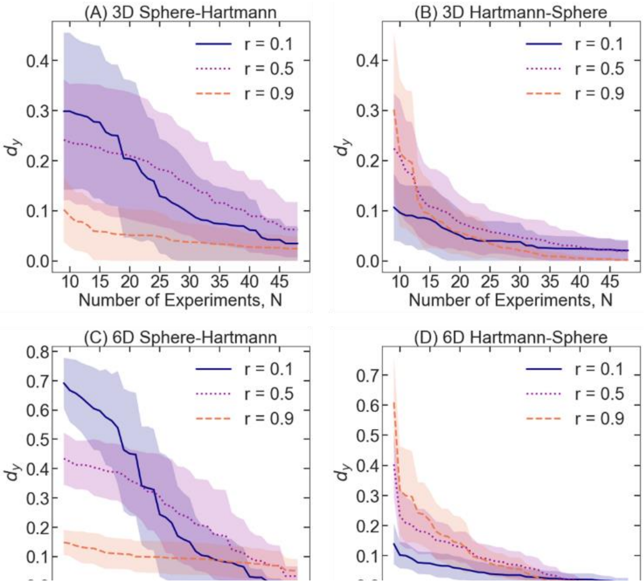

Figure 7. d_y versus number of experiments, *N*, for different objective function mixing ratios, *r*. Larger *r* means more similarity between physics model and experimental objective function. (A) 3D Sphere mixed with 3D Hartmann using Sphere as the simplified physics model. (B) 3D Sphere mixed with 3D Hartmann using Hartmann as the simplified physics model. (C) 6D Sphere mixed with 6D Hartmann using Sphere as the simplified physics model. (D) 6D Sphere mixed with 6D Hartmann using Hartmann as the simplified physics model. For each curve, 25 separate searches, S, were performed, and the average of the results are the lines shown. The shadow around each of the lines represents the standard deviation. For all of these graphs, ChIDDO was used as the AL algorithm and MRB was used as the acquisition function.

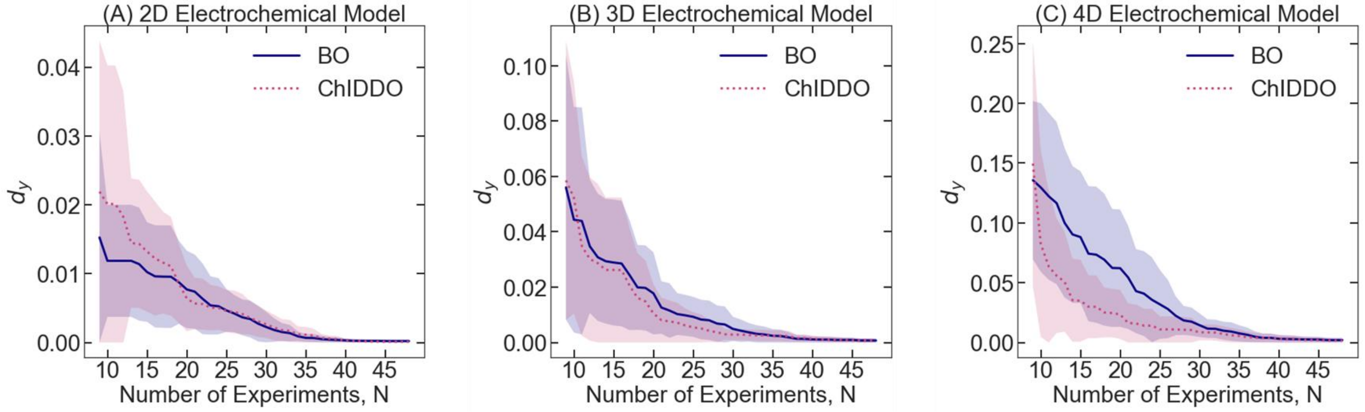

Figure 8. d_y versus number of experiments, *N*, comparing different noise levels, *η* , represented by lines of different colors. (A) 6D Hartmann – BO - PI, (B) 6D Hartmann – BO - EI, (C) 6D Hartmann – BO – MRB. For each curve, 25 separate searches, *S*, were performed, and the average of the results are the lines shown. The shadow around each of the lines represents the standard deviation. For each of these studies, the MRB acquisition function was used.

that the physics model and the experimental points are dissimilar. Conversely, when *r* is high, the physics model and the objective function are close to each other, and one would expect d_y to decrease more rapidly with the number of experiments than in the case of low values of *r*. Interestingly, it appears that the similarity between the physics model and the objective functions only makes a small difference in performance. In these figures it is important to note that since there is a combination of objective functions, the optimum values change for each *r*, resulting in different d_y values after the initial points are run. However, the curves in Figure 7A for *r* = 0.1 and Figure 7B for *r* = 0.9 are based on the same objective function values. From these results, it can be observed that using the Hartmann function as the physics model allowed for improved performance. This could be due to the fact that the Hartmann function incorporates a higher degree of complexity than the Sphere function. For these objective functions with parameters that are easy to regress, the simplified physics model was able to be modified enough to predict values close to the combined objective function. Even with little similarity between the simplified physics model and the combined objective function (*r* = 0.1), the algorithm had adequate performance. However, when implementing more complex physics models that cannot be regressed as easily to match the experimental values, a simplified physics model may show inadequate performance.

Simulation of an Electrochemical Optimization

In the previous sub-sections, we demonstrated the development of ChIDDO using model functions that are difficult to optimize but that are not based on chemical processes. To illustrate the implementation of ChIDDO in a chemical process, we attempted to optimize the Faradaic Efficiency (FE) of product D in the simulated set of electrochemical reactions described in Equations 6 – 9. This is a common objective function in electrochemical processes, where it is often desirable to selectively generate a single product. We studied how BO and ChIDDO performed on electrochemical models with two, three, and four dimensions. For two dimensions, the bulk concentrations of two reactants were the two variables (0.1 – 1 mol dm^{-3}). Voltage (2V – 4V) was used as the third variable and temperature (25C – 80C) was used as the fourth variable. Figure 8 shows the performance of the BO and ChIDDO algorithms on the different dimension electrochemical models. For the 2D and 3D optimizations, the performance of BO and ChIDDO was similar. However, when the fourth dimension was added, ChIDDO outperformed BO, especially at a low number of

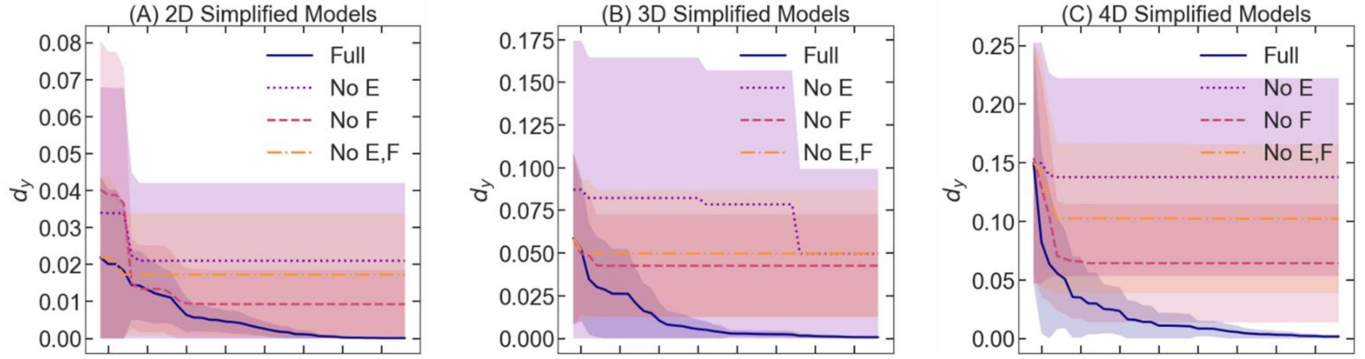

Figure 9. d_y versus number of experiments, *N*, for different electrochemical physics model information. “Full” indicates the model is predicting the same information as the objective function. “No E”, “No F”, and “No EF” indicate the removal of Equations 8 and/or 9 from the physics model information, resulting in a less informative model. (A) 2D electrochemical model. (B) 3D electrochemical model. (C) 4D electrochemical model. For each curve, 25 separate searches, S, were performed, and the average of the results are the lines shown. The shadow around each of the lines represents the standard deviation. For all of these graphs, ChIDDO was used as the AL algorithm and MRB was used as the acquisition function.

experiments. This shows that the physics model allowed the algorithm to identify areas close to maximum without having to search the entire space.

The electrochemical model used in this study has 4 different parallel reactions (Equations 6-9). It is common when simulating a complex reaction network that not all the intermediates or products are known. To test the robustness of the ChIDDO algorithm to incomplete physics models, Equation 8 and/or 9 was removed from the set of physics model reactions. When the full model was used as the physics model, a continuous improvement can be seen as more experiments are incorporated, as seen in Figure 9. However, when one or two reactions are not included in the physics model, the algorithm is not able to improve the optimal value after the first few experiments. This could be the case if the simplified physics model does not agree with the values of the true objective function, leading to experimental selections that are far from the optimal values. After observing the model data from the simplified models, the objective values for the design space have different shapes and magnitudes than the experimental objective function. Examples of the simplified physics model data are shown in the Supplemental Information. After a large number of experiments, the GPR prediction starts to become dominated by the experimental results and the exploration rate decreases, ultimately prompting the algorithm to select suboptimal experiments in close proximity to regions with low d_y values found during the early stage of the optimization. This indicates that it is important to have high accuracy in the physics model, or to extend the exploration phase of the algorithm if the information used in the physics model has large uncertainty.

Conclusions

This work introduced the ChIDDO approach, an optimization methodology where information from physical models of chemical processes is used synergistically with experimental data to potentially improve BO performance. Our results show that both BO and ChIDDO outperform systematic grid or random searches. The ChIDDO algorithm improves the initial performance of the optimization of various types of objective functions, but as more experimental results become available, the performance of BO and ChIDDO tend to converge. The advantages of the inclusion of physical models are more pronounced in optimization problems of high dimensions. This is evident in the case of a 6D Hartmann function, where d_y values for ChIDDO were substantially lower than the BO d_y values, while in the case of 3D and 4D Hartmann optimizations the difference is minimal. Similar results were observed when using data with and without noise. Interestingly, the standard deviation between different experiment was smaller when using ChIDDO, indicating a more consistent optimization regardless of the experimental observations. We also explored scenarios when the physics model may not accurately describe the experimental objective function. In these scenarios, the effect of the inaccuracy of the physics model depends on how easy the physics model can be regressed and modified to resemble take into account the experimental points. For the more constrained physics model used in the electrochemical models, the effect of an inaccurate physics model was drastic. Overall, the importance and potential performance improvements afforded by the physics model information progressively decreases as experimental information increases, and ChIDDO approaches become increasingly similar to BO. Our findings suggest that while the inclusion of physical models of chemical processes may aid the optimization of processes with a large number of optimization parameters, the improvements provided in low-dimensionality optimization problems, such as the 2-D electrochemical reaction optimization example presented, are not significant and data-only approaches are appropriate to rapidly identify optima.

Author Contributions

Daniel Frey: Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft **Juhee Shin**: Investigation, Formal analysis, Software **Christopher Musco**: Writing – review & editing, Supervision **Miguel A. Modestino**: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing

Conflicts of interest

MAM is a director and has a financial interest in Sunthetics, Inc., a startup company in the machine learning optimization space.

Acknowledgements

We thank NYU Tandon School of Engineering and the National Science Foundation for their generous financial support. This material is based upon work supported by the National Science Foundation under grant no. 1943972.

References

1. Blanco DE, Lee B, Modestino MA. Optimizing organic electrosynthesis through controlled voltage dosing and artificial intelligence. Proceedings of the National Academy of Sciences. 2019;116(36):17683-9.
2. Grover A, Markov T, Attia P, Jin N, Perkins N, Cheong B, et al. Best arm identification in multi-armed bandits with delayed feedback. arXiv preprint arXiv:180310937. 2018.
3. Xue D, Balachandran PV, Hogden J, Theiler J, Xue D, Lookman T. Accelerated search for materials with targeted properties by adaptive design. Nat Commun. 2016;7(1):1-9.
4. Vahid A, Rana S, Gupta S, Vellanki P, Venkatesh S, Dorin T. New bayesian-optimization-based design of high-strength 7xxx-series alloys from recycled aluminum. JOM. 2018;70(11):2704-9.
5. Li C, de Celis Leal DR, Rana S, Gupta S, Sutti A, Greenhill S, et al. Rapid Bayesian optimisation for synthesis of short polymer fiber materials. Scientific reports. 2017;7(1):1-10.
6. Abdelrahman H, Berkenkamp F, Poland J, Krause A. Bayesian optimization for maximum power point tracking in photovoltaic power plants. 2016 European Control Conference (ECC). 2016:2078-83.

7. Kikuchi S, Oda H, Kiyohara S, Mizoguchi T. Bayesian optimization for efficient determination of metal oxide grain boundary structures. Physica B: Condensed Matter. 2018;532:24-8.

8. Khajah MM, Roads BD, Lindsey RV, Liu Y-E, Mozer MC. Designing engaging games using Bayesian optimization. Proceedings of the 2016 CHI conference on human factors in computing systems. 2016:5571-82.

9. Lorenz R, Violante IR, Monti RP, Montana G, Hampshire A, Leech R. Dissociating frontoparietal brain networks with neuroadaptive Bayesian optimization. Nat Commun. 2018;9(1):1-14.

10. Frazier P. A tutorial on Bayesian optimization. arXiv: 180702811. 2018.

11. Herbol HC, Hu W, Frazier P, Clancy P, Poloczek M. Efficient search of compositional space for hybrid organic–inorganic perovskites via Bayesian optimization. Computational Materials. 2018;4(1):1-7.

12. Herbol HC, Poloczek M, Clancy P. Cost-effective materials discovery: Bayesian optimization across multiple information sources. Materials Horizons. 2020;7(8):2113-23.

13. Yamashita T, Sato N, Kino H, Miyake T, Tsuda K, Oguchi T. Crystal structure prediction accelerated by Bayesian optimization. Physical Review Materials. 2018;2(1):013803.

14. Ju S, Shiga T, Feng L, Hou Z, Tsuda K, Shiomi J. Designing nanostructures for phonon transport via Bayesian optimization. Physical Review X. 2017;7(2):021024.

15. Ueno T, Rhone TD, Hou Z, Mizoguchi T, Tsuda K. COMBO: an efficient Bayesian optimization library for materials science. Materials discovery. 2016;4:18-21.

16. Hashimoto W, Tsuji Y, Yoshizawa K. Optimization of Work Function via Bayesian Machine Learning Combined with First-Principles Calculation. The Journal of Physical Chemistry C. 2020;124(18):9958-70.

17. Balachandran PV, Kowalski B, Sehirlioglu A, Lookman T. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. Nat Commun. 2018;9(1):1-9.

18. Higgins K, Valleti SM, Ziatdinov M, Kalinin SV, Ahmadi M. Chemical Robotics Enabled Exploration of Stability in Multicomponent Lead Halide Perovskites via Machine Learning. ACS Energy Lett. 2020;5(11):3426-36.

19. Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. Integrating Materials and Manufacturing Innovation. 2017;6(3):207-17.

20. MacLeod BP, Parlane FG, Morrissey TD, Häse F, Roch LM, Dettelbach KE, et al. Self-driving laboratory for accelerated discovery of thin-film materials. Science Advances. 2020;6(20):eaaz8867.

21. Min K, Cho E. Accelerated discovery of potential ferroelectric perovskite via active learning. Journal of Materials Chemistry C. 2020;8(23):7866-72.

22. Rezaeianjouybari B, Sheikholeslami M, Shafee A, Babazadeh H. A novel Bayesian optimization for flow condensation enhancement using nanorefrigerant: a combined analytical and experimental study. Chem Eng Sci. 2020;215:115465.

23. Park S, Na J, Kim M, Lee JM. Multi-objective Bayesian optimization of chemical reactor design using computational fluid dynamics. Comput Chem Eng. 2018;119:25-37.

24. Schweidtmann AM, Clayton AD, Holmes N, Bradford E, Bourne RA, Lapkin AA. Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives. Chem Eng J. 2018;352:277-82.

25. Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, et al. A mobile robotic chemist. Nature. 2020;583(7815):237-41.

26. Granda JM, Donina L, Dragone V, Long D-L, Cronin L. Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature. 2018;559(7714):377-81.

27. Guo Z, Wu S, Ohno M, Yoshida R. Bayesian Algorithm for Retrosynthesis. Journal of Chemical Information and Modeling. 2020;60(10):4474-86.

28. Häse F, Roch LM, Aspuru-Guzik A. Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories. Chemical science. 2018;9(39):7642-55.

29. Häse F, Roch LM, Aspuru-Guzik A. Next-generation experimentation with self-driving laboratories. Trends in Chemistry. 2019;1(3):282-91.

30. Kondo M, Wathsala H, Sako M, Hanatani Y, Ishikawa K, Hara S, et al. Exploration of flow reaction conditions using machine-learning for enantioselective organocatalyzed Rauhut–Currier and [3+ 2] annulation sequence. Chem Commun. 2020;56(8):1259-62.

31. Shields BJ, Stevens J, Li J, Parasram M, Damani F, Alvarado JIM, et al. Bayesian reaction optimization as a tool for chemical synthesis. Nature. 2021;590(7844):89-96.

32. Reker D, Hoyt EA, Bernardes GJ, Rodrigues T. Adaptive Optimization of Chemical Reactions with Minimal Experimental Information. Cell Reports Physical Science. 2020;1(11):100247.

33. Kim K, Lee WH, Na J, Hwang Y, Oh H-S, Lee U. Data-driven pilot optimization for electrochemical CO mass production. Journal of Materials Chemistry A. 2020;8(33):16943-50.

34. Wang Y, Xie T, France-Lanord A, Berkley A, Johnson JA, Shao-Horn Y, et al. Toward Designing Highly Conductive Polymer Electrolytes by Machine Learning Assisted Coarse-Grained Molecular Dynamics. Chem Mater. 2020;32(10):4144-51.

35. Attia PM, Grover A, Jin N, Severson KA, Markov TM, Liao Y-H, et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning. Nature. 2020;578(7795):397-402.

36. Doan HA, Agarwal G, Qian H, Counihan MJ, Rodríguez-López J, Moore JS, et al. Quantum chemistry-informed active learning to accelerate the design and discovery of sustainable energy storage materials. Chem Mater. 2020;32(15):6338-46.

37. Dave A, Mitchell J, Kandasamy K, Wang H, Burke S, Paria B, et al. Autonomous Discovery of Battery Electrolytes with Robotic Experimentation and Machine Learning. Cell Reports Physical Science. 2020:100264.

38. Ebrahimi M, Ting DSK, Carriveau R, McGillis A, Young D. Optimization of a cavern-based compressed air energy storage facility with an efficient adaptive genetic algorithm. Energy Storage. 2020;2(6):e205.

39. Poloczek M, Wang J, Frazier P. Multi-information source optimization. Advances in Neural Information Processing Systems. 2017:4288-98.

40. Swersky K, Snoek J, Adams RP. Multi-task bayesian optimization. Advances in neural information processing systems. 2013:2004-12.

41. Lam R, Allaire DL, Willcox KE, editors. Multifidelity optimization using statistical surrogate modeling for non-hierarchical information sources. 56th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference; 2015.

42. Cardoso TN, Silva RM, Canuto S, Moro MM, Gonçalves MA. Ranked batch-mode active learning. Information Sciences. 2017;379:313-37.

43. Jones DR, Schonlau M, Welch WJ. Efficient global optimization of expensive black-box functions. Journal of Global optimization. 1998;13(4):455-92.
44. Kushner HJ. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. 1964:97-106.
45. Srinivas N, Krause A, Kakade SM, Seeger M. Gaussian process optimization in the bandit setting: No regret and experimental design. arXiv preprint arXiv:09123995. 2009.
46. Acquisition funcitons modAL [cited 2021. Available from: https://modal-python.readthedocs.io/en/latest/content/query_strategies/Acquisition-functions.html.
47. Baizer MM. Electrolytic Reductive Coupling: I. Acrylonitrile. J Electrochem Soc. 1964;111(2):215.
48. Botte GG. Electrochemical manufacturing in the chemical industry. The Electrochemical Society Interface. 2014;23(3):49.
49. Frey D. Chemically-informed-data-driven-optimization-ChIDDO. GitHub repository. 2021.