

Mercer Features for Efficient Combinatorial Bayesian Optimization

Aryan Deshwal, Syrine Belakaria, Janardhan Rao Doppa

School of EECS, Washington State University
{aryan.deshwal, syrine.belakaria, jana.doppa}@wsu.edu

Abstract

Bayesian optimization (BO) is an efficient framework for solving black-box optimization problems with expensive function evaluations. This paper addresses the BO problem setting for combinatorial spaces (e.g., sequences and graphs) that occurs naturally in science and engineering applications. A prototypical example is molecular optimization guided by expensive experiments. The key challenge is to balance the complexity of statistical models and tractability of search to select combinatorial structures for evaluation. In this paper, we propose an efficient approach referred as *Mercer Features for Combinatorial Bayesian Optimization (MerCBO)*. The key idea behind MerCBO is to provide explicit feature maps for diffusion kernels over discrete objects by exploiting the structure of their combinatorial graph representation. These Mercer features combined with Thompson sampling as the acquisition function allows us to employ tractable solvers to find next structures for evaluation. Experiments on diverse real-world benchmarks demonstrate that MerCBO performs similarly or better than prior methods.

Introduction

The problem of black-box optimization over combinatorial spaces (e.g., sequences and graphs) using expensive evaluations arises in diverse real-world applications. A prototypical example is molecular optimization guided by wet lab experiments. Bayesian optimization (BO) is an efficient framework (Shahriari et al. 2016) for solving such problems and has seen a lot of success for continuous spaces (Snoek, Larochelle, and Adams 2012; Wang and Jegelka 2017; Belakaria et al. 2019; Belakaria et al. 2020a; Belakaria, Deshwal, and Doppa 2020a,b). BO algorithms learn a statistical model, e.g., Gaussian process (GP) (Rasmussen and Williams 2006), and employ it to select the sequence of inputs for evaluation guided by an acquisition function, e.g., expected improvement (EI). The key BO challenge for combinatorial spaces is to appropriately trade-off the complexity of statistical models and tractability of the search process to select structures for evaluation. Prior methods either employ simpler models that are amenable for tractable search (Baptista and Poloczek 2018) or perform heuristic search with complex models (Hutter, Hoos, and Leyton-Brown 2011; Oh

et al. 2019). Additionally, these methods can only work with a limited set of acquisition functions. COMBO (Oh et al. 2019) is the state-of-the-art method that employs a GP-based complex statistical model using discrete diffusion kernels and performs local search to select structures.

In this paper, we ask the following question: *Can we retain the modeling strength of GP-based model using discrete diffusion kernels and still perform tractable search for selecting better structures to improve the overall BO performance?* To answer this question, we propose a novel approach referred as Mercer Features for Combinatorial Bayesian Optimization (MerCBO), pronounced as “Mercè BO”. The main advantage of COMBO comes from the combinatorial graph representation of input space \mathbb{X} that allows using diffusion kernels to define smooth functions over \mathbb{X} . MerCBO computes a closed-form expression of this class of smooth functions in terms of explicit feature maps of diffusion kernels, which we refer as *Mercer features*. The key insight is to exploit the structure of the graph representation in COMBO to extract powerful low-dimensional features.

One main advantage of Mercer features is that they allow us to leverage a large number of acquisition functions and algorithms from the continuous BO literature to improve the BO performance for combinatorial spaces. In MerCBO, we employ Thompson sampling (TS) as the acquisition function by sampling parametric functions from GP posterior via Mercer features. We show that the acquisition function optimization problem with TS is a Binary Quadratic Program (BQP). Inspired by the success of submodular relaxation in the structured prediction literature (Gorelick et al. 2014), we study a fast and scalable submdular relaxation method (Ito and Fujimaki 2016) to solve BQPs for selecting structures. The key idea is to construct a lower bound of the BQP problem in terms of unknown relaxation parameters and iteratively optimize those parameters to improve accuracy.

In many scientific applications (e.g., biological sequence design and chemical design), we have the ability to perform a large number of parallel experiments. Therefore, there is a great need for scalable algorithms for selecting batch of inputs for parallel evaluation to quickly uncover diverse and high-performing structures. MerCBO can be employed to meet this desiderata as Thompson sampling is embarrassingly parallel and its exploration strategy is highly helpful for diversity. Experiments on multiple real-world bench-

marks including biological sequence design show that MerCBO performs significantly better than prior methods.

Key Contributions. We develop and evaluate the MerCBO approach for BO problems over combinatorial spaces.

- Algorithm to compute efficient Mercer features for diffusion kernels to enable improved BO performance.
- Efficient submodular-relaxation approach based on Thompson sampling to select structures for evaluation. Demonstration of its scalability and effectiveness for selecting diverse batch of structures for parallel evaluation.
- Empirical evaluation using diverse real-world benchmarks and comparison with state-of-the-art methods.

Problem Setup and Related Work

A black-box optimization problem over combinatorial spaces is specified by a 2-tuple $\langle \mathbb{X}, O \rangle$, where \mathbb{X} is an input space over discrete structures (e.g., sequences, graphs) and O is the black-box objective function. Without loss of generality, we assume a space of binary structures $\mathbb{X} = \{0, 1\}^n$. Note that we can convert any discrete space into an efficient binary representation by encoding categorical variables. Suppose each structure $\mathbf{x} \in \mathbb{X}$ be represented using n binary variables x_1, x_2, \dots, x_n . We are allowed to evaluate each structure $\mathbf{x} \in \mathbb{X}$ using the objective function $O : \mathbb{X} \rightarrow \mathbb{R}$ to observe the output $y = O(\mathbf{x})$. Each function evaluation is expensive in terms of the consumed resources (e.g., computational or physical). For example, in molecular optimization application, $\mathbf{x} \in \mathbb{X}$ is a candidate molecule represented as a graph and $O(\mathbf{x} \in \mathbb{X})$ corresponds to performing a expensive wet lab experiment to measure the desired property of molecule. The overall goal is to optimize the objective O by minimizing the number of evaluations.

BO methods learn a statistical model M from past function evaluations and select one structure for evaluation in each iteration using an acquisition function α (e.g., expected improvement). To perform this selection, we need to solve the following combinatorial optimization problem:

$$\mathbf{x}_{next} = \arg \max_{\mathbf{x} \in \mathbb{X}} \alpha(M, \mathbf{x}) \quad (0.0.1)$$

The key challenge is to appropriately trade-off the complexity of statistical model M and tractability of solving the above acquisition function optimization (AFO) problem.

Related Work. There is limited work for BO over combinatorial spaces. Prior work addresses the above challenge by using either simple models that are amenable for tractable solvers or perform heuristic search with complex models. BOCS (Baptista and Poloczek 2018) employs a linear second-order model over binary variables that is amenable for semi-definite program based AFO solvers. Discrete walsh functions (Lepřêtre et al. 2019) are also employed to build surrogate models. SMAC (Hutter, Hoos, and Leyton-Brown 2010, 2011) employs random forest as statistical model and a *hand-designed* local search procedure for solving AFO problems. COMBO (Oh et al. 2019) employs GP-based statistical model using discrete diffusion kernels that allows to model higher-order interactions between variables and uses local search with restarts as the AFO solver.

COMBO was shown to perform significantly better than both SMAC and BOCS. COMEX (Dadkhahi et al. 2020) employs multi-linear polynomial as surrogate model with exponential weight updates and executes simulated annealing using the learned model to select inputs for evaluation. In a complementary direction, when AFO problems are not tractable, L2S-DISCO (Deshwal et al. 2020) exploits the relationship between the sequence of AFO problems across BO iterations via learning to improve the accuracy of search.

There is another line of work that assumes the availability of a large dataset of structures. One generic approach is to learn a continuous representation using this dataset and to perform BO in the latent space (Gómez-Bombarelli et al. 2018). One potential drawback of this reduction method is that it can generate invalid structures (Janz et al. 2018; Kusner, Paige, and Hernández-Lobato 2017). There are also methods that employ this big data to learn Bayesian neural networks as statistical models (Hernández-Lobato et al. 2017). Unfortunately, this assumption is violated in applications where small data is the norm.

MerCBO Algorithm

Algorithm 1: MerCBO Algorithm

Input: Input space \mathbb{X} , Black-box objective O , Order of Mercer features MAX, Query budget B

Output: Best input structure found \mathbf{x}_{best}

```

1 MerCBO(input structure  $\mathbf{x}$ , order MAX)
2   Initialize empty list  $\phi = []$ 
3   for  $i=1$  to MAX do
4     Compute the features for the  $i$ th order:
5      $\{\sqrt{e^{-\beta\lambda_i}} \cdot -1^{< r_i, \mathbf{x} >}\}$  and append to  $\phi$ 
6   return Mercer features  $\phi$ 
7
8 Initialize a small-sized training set TRAIN with Mercer
9 features computed for input structures
10 while Query budget does not exceed  $B$  do
11   Learn Gaussian Process model  $M$  using TRAIN
12   Construct Thompson sampling acquisition function by
13   sampling from a parametric approximation of the GP
14   posterior
15   Find  $\mathbf{x}_{next}$  by optimizing Thompson sampling based
16   acquisition function over model  $M$ 
17   begin
18     Step 1: Construct submodular relaxation based
19     lower bound of AFO with parameters
20     while convergence or maximum iterations do
21       Step 2: Solve the relaxed problem using graph
22       cut algorithms
23       Step 3: Update the relaxation parameters to
24       obtain a tighter bound
25     end
26   Evaluate  $O$  at  $\mathbf{x}_{next}$  and add to TRAIN:  $\mathbf{x}_{next}$ ,
27   Mercer_Features( $\mathbf{x}_{next}$ , MAX), and  $O(\mathbf{x}_{next})$ 
28 return Best input  $\mathbf{x}_{best}$  and its objective value  $O(\mathbf{x}_{best})$ 

```

We first provide an overview of MerCBO and its advantages.

Overview of MerCBO algorithm. MerCBO is an instantiation of the generic BO framework. 1) Gaussian Process

with diffusion kernels is the surrogate statistical model. 2) Thompson sampling is used as the acquisition function α . The proposed Mercer features are used to sample from a parametric approximation of the GP posterior to construct the Thompson sampling objective. 3) Acquisition function optimization problem is shown to be a Binary Quadratic Program which is solved using an efficient and scalable submodular-relaxation method. The key idea is to construct a lower bound of the AFO problem in terms of some unknown relaxation parameters and iteratively optimize those parameters to obtain a tighter bound. Algorithm 1 shows the complete pseudo-code of MerCBO. We use a small set of input structures and their function evaluations (denoted TRAIN) to bootstrap the surrogate model. In each iteration, we select the next structure for evaluation \mathbf{x}_{next} by solving the AFO problem; add \mathbf{x}_{next} , its function evaluation $O(x_{next})$, and mercer features of \mathbf{x}_{next} to TRAIN. We repeat these sequence of steps until the query budget is exhausted and then return the best found structure \mathbf{x}_{best} as the output.

Advantages of MerCBO over COMBO and SMAC. **1)** Mercer features allow us to leverage a large number of acquisition functions from the continuous BO literature including Thompson sampling (TS), PES (Hernández-Lobato, Hoffman, and Ghahramani 2014) and MES (Wang and Jegelka 2017) to improve the BO performance for combinatorial spaces. **2)** Retains the modeling strength of complex GP-based model using diffusion kernels and still allows a tractable and scalable AFO procedure with Thompson sampling. **3)** Mercer features combined with TS has many desiderata required for several scientific applications: *diversity* in explored structures and selection of *large batch* of structures for parallel evaluation. Indeed, our experiments on biological sequence design demonstrate the effectiveness of TS (embarrassingly parallel) over COMBO with EI.

Preliminaries

Graph Laplacian. Given a graph $G = (V, E)$, its Laplacian matrix $L(G)$ is defined as $D - A$, where D is the degree matrix and A is the adjacency matrix of G .

Graph Cartesian Product. Given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, their graph Cartesian product ($G_1 \square G_2$) is another graph with vertex set $V(G_1 \square G_2) = V_1 \times V_2$ consisting of the set Cartesian product of V_1 and V_2 . Two vertices (v_1, u_1) and (v_2, u_2) of $G_1 \square G_2$ (where $\{v_1, v_2\} \in V_1$ and $\{u_1, u_2\} \in V_2$) are connected if either, $v_1 = v_2$ and $(u_1, u_2) \in E_2$, or $u_1 = u_2$ and $v_1, v_2 \in E_1$.

Combinatorial Graph Representation of Discrete Space. Recall that we need a graph representation of the discrete space \mathbb{X} to employ diffusion kernels for learning GP models. We consider the combinatorial graph representation (say $G=(V, E)$) proposed in a recent work (Oh et al. 2019). There is one vertex for each candidate assignment of n discrete variables x_1, x_2, \dots, x_n . There is an edge between two vertices if and only if the Hamming distance between the corresponding binary structures is one. This graph representation was shown to be effective in building GP models for BO over discrete spaces (Oh et al. 2019).

Efficient Mercer features for Diffusion Kernel

We are interested in computing explicit feature maps for diffusion kernel over discrete spaces (Kondor and Lafferty 2002; Oh et al. 2019), which is defined using the above combinatorial graph representation as follows:

$$K(V, V) = U \exp(-\beta \Lambda) U^T \quad (0.0.2)$$

where $U = [u_0, u_1, \dots, u_{2^n-1}]$ is the eigenvector and $\Lambda = [\lambda_0, \lambda_1, \dots, \lambda_{2^n-1}]$ is the eigenvalue matrix of the graph Laplacian $L(G)$ and β is a hyper-parameter. For any two given structures $\mathbf{x}_1, \mathbf{x}_2 \in \{0, 1\}^n$, the kernel definition is:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=0}^{2^n-1} e^{-\beta \lambda_i} u_i([\mathbf{x}_1]) u_i([\mathbf{x}_2]) \quad (0.0.3)$$

where $u_i([\mathbf{x}_1])$ denotes the value of the i th eigenvector indexed at the integer value represented by \mathbf{x}_1 in base-2 number system. From the above equation, we can see that the eigenspace of the graph Laplacian $L(G)$ is the central object of study for diffusion kernels. The key insight behind our proposed approach is to exploit the structure of the combinatorial graph G and its Laplacian $L(G)$ for computing its eigenspace in a *closed-form*.

Since G is an exponential-sized graph with 2^n vertices, computing the eigenspace of $L(G)$ seems intractable at first sight. However, G has special structure: G has an equivalent representation in terms of the graph Cartesian product of n sub-graphs G_1, G_2, \dots, G_n :

$$G = (((G_1 \square G_2) \square G_3 \dots) \dots) \square G_n \quad (0.0.4)$$

where each sub-graph G_i represents the i th binary input variable and is defined as a graph with two vertices (labeled 0 and 1) and an edge between them. Therefore, $L(G)$ is equivalently given by:

$$L(G) = L((((G_1 \square G_2) \square G_3 \dots) \dots) \square G_n) \quad (0.0.5)$$

$$L(G) = L(G_1) \oplus L(G_2) \oplus L(G_3) \dots \oplus L(G_n) \quad (0.0.6)$$

where Equation 0.0.6 is due to distributive property of the Laplacian operator over graph Cartesian product via Kronecker sum operator (\oplus) (Hammack, Imrich, and Klavžar 2011) and associative property of the \oplus operator.

Property 1 (Hammack, Imrich, and Klavžar 2011) *Given two graphs G_1 and G_2 with the eigenspace of their Laplacians being $\{\Lambda^1, U^1\}$ and $\{\Lambda^2, U^2\}$ respectively, the eigenspace of $L(G_1 \square G_2)$ is given by $\{\Lambda^1 \boxtimes \Lambda^2, U^1 \otimes U^2\}$ where $\Lambda^1 \boxtimes \Lambda^2 = \{\lambda_i^1 + \lambda_j^2 : \lambda_i^1 \in \Lambda^1, \lambda_j^2 \in \Lambda^2\}$ and \otimes is the Kronecker product operation.*

Property 1 gives a recursive procedure to compute the eigenspace of $L(G)$ based on its decomposition in Equation 0.0.6 provided the eigenspace of each of its constituent $L(G_i)$ is easily computable. Fortunately, in our special case, where each G_i is a simple graph of two vertices (labeled 0 and 1) with an edge between them, it has two eigenvalues $\{0, 2\}$ with corresponding eigenvectors $[1, 1]$ and $[1, -1]$ respectively. The eigenvector matrix $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ is called as a Hadamard matrix of order 2 (2^1 , where 1 in the exponent is

the number of input variables) (Florkowski III 2008; Yackley, Corona, and Lane 2009). Applying the \boxtimes and \otimes operation recursively for n inputs (represented by $\{G_i : i \in \{1, 2, \dots, n\}\}$) as described in property 1, it can be seen that the eigenspace of $L(G)$ has an explicit form given as:

1. $L(G)$ has n distinct eigenvalues $\{0, 2, 4, \dots, 2n\}$ where j th eigenvalue occurs with multiplicity $\binom{n}{j}$, $j \in \{0, 1, 2, \dots, n\}$.
2. The eigenvectors of $L(G)$ are given by the columns of Hadamard matrix of order 2^n .

Hadamard matrix of order 2^n is equivalently defined as:

$$H_{ij} = (-1)^{\langle r_i, r_j \rangle} \quad (0.0.7)$$

where r_i is the n -bit representation of the integer i in base-2 system. Using this definition of Hadamard matrix, each entry of the eigenvectors of $L(G)$ can be computed in a closed form. Therefore, from Equation 0.0.3, the kernel value for any pair of structures \mathbf{x}_1 and \mathbf{x}_2 , $K(\mathbf{x}_1, \mathbf{x}_2)$, can be written in terms of an equivalent sum over binary vectors:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=0}^{2^n-1} e^{-\beta\lambda_i} \cdot -1^{\langle r_i, \mathbf{x}_1 + \mathbf{x}_2 \rangle} \quad (0.0.8)$$

where r_i is the base-2 representation of integer i ranging from 0 to 2^n-1 . We rearrange the RHS of Equation 0.0.8 to delineate the dependency on each input in the pair.

$$K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=0}^{2^n-1} e^{-\beta\lambda_i} \cdot -1^{\langle r_i, \mathbf{x}_1 \rangle} \cdot -1^{\langle r_i, \mathbf{x}_2 \rangle} \quad (0.0.9)$$

$$K(\mathbf{x}_1, \mathbf{x}_2) = \langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle \quad (0.0.10)$$

where $\phi(\mathbf{x})$ corresponds to the proposed (explicit) *Mercer feature maps* of any input structure \mathbf{x} and is given as follows:

$$\phi(\mathbf{x}) = \{i \in [0, 2^n - 1] : \sqrt{e^{-\beta\lambda_i}} \cdot -1^{\langle r_i, \mathbf{x} \rangle}\} \quad (0.0.11)$$

As mentioned earlier, the eigenvalues of $L(G)$ are explicitly given by the set $\{2j : j \in \{0, 1, 2, \dots, n\}\}$, where j th eigenvalue occurs with multiplicity $\binom{n}{j}$. Based on this observation, an elegant way of interpreting the feature maps given in 0.0.11 is based on the number of 1s in the r_i vector (binary expansion of integer i). There are exactly $\binom{n}{j}$ r -vectors with j bits set to 1. Hence, we refer to j as the *order of the Mercer feature maps*. The order variable controls the trade-off between the computational cost and approximation quality of the feature map. We found that the second-order feature maps¹ maintain the right balance as they can be computed efficiently and also allows to perform a tractable and scalable search for acquisition function optimization as described in the next section. Moreover, choosing second order is also prudent from the viewpoint of the definition of diffusion kernels itself, which requires suppressing higher frequency elements of the eigenspace (Oh et al. 2019) to define a smooth function over discrete spaces.

¹Slight abuse of notation. Second-order, from here-on, means concatenation of both first and second-order features.

Tractable Acquisition Function Optimization

In this section, we describe a tractable and scalable acquisition function optimization procedure using the proposed Mercer features and Thompson sampling. Thompson sampling (Chowdhury and Gopalan 2017; Hernández-Lobato et al. 2017; Kandasamy et al. 2018) selects the next point for evaluation by maximizing a sampled function from the GP posterior. We approximate the non-parametric GP model using a parametric Bayesian linear model $f(\mathbf{X}) = \theta^T \phi(\mathbf{x})$ using our proposed Mercer features $\phi(\mathbf{x})$. Given a Gaussian prior, i.e., $\theta \sim \mathcal{N}(0, I)$, the posterior distribution over θ is also a Gaussian with the following mean and covariance:

$$\mu = (\Phi^T \Phi + \sigma^2 \mathbf{I})^{-1} \Phi^T \mathbf{y} \quad (0.0.12)$$

$$\Sigma = (\Phi^T \Phi + \sigma^2 \mathbf{I})^{-1} \sigma^2 \quad (0.0.13)$$

with Φ is the feature matrix with i th row corresponding to $\phi(\mathbf{x}_i)$ and \mathbf{y} is the output vector with $y_i \sim \mathcal{N}(f(x_i), \sigma^2)$.

Acquisition function optimization problem. We sample θ^* from the posterior parametrized by μ and Σ defined in Equation 0.0.13 and minimize the objective $f(\mathbf{x}) = \theta^* \phi(\mathbf{x})$ with respect to $\mathbf{x} \in \{0, 1\}^n$. Suppose $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is a candidate structure with values assigned to n binary variables. We now show how this AFO problem is an instance of Binary quadratic programming (BQP) problem by using second-order feature maps from 0.0.11.

$$\phi(\mathbf{x}) = \left\{ 0 \leq i \leq \binom{n}{2} : \sqrt{e^{-\beta\lambda_i}} \cdot -1^{\langle r_i, \mathbf{x} \rangle} \right\} \quad (0.0.14)$$

The second-order feature maps are composed of two major parts: features constructed by order-1 r -vectors and features constructed by order-2 r -vectors. For order-1 r -vectors, the set $\{0 \leq i \leq n : -1^{\langle r_i, \mathbf{x} \rangle}\}$ is equivalent to $\{0 \leq i \leq n : -1^{x_i}\}$. Similarly, the order-2 feature maps can also be written as $\{1 \leq i \leq n, i+1 \leq j \leq n : -1^{(x_i+x_j)}\}$. Therefore, ignoring the constant corresponding to the 0th index of θ^* , the AFO problem $\min_{\mathbf{x} \in \{0,1\}^n} \theta^* \phi(\mathbf{x})$ becomes:

$$\min_{\mathbf{x} \in \{0,1\}^n} \sum_{i=1}^n \theta_i^* \sqrt{e^{-\beta\lambda_i}} \cdot -1^{x_i} + \sum_{i=1}^n \sum_{j=i+1}^n \theta_{ij}^* \sqrt{e^{-\beta\lambda_{n+i+j}}} \cdot -1^{(x_i+x_j)} \quad (0.0.15)$$

By replacing the -1^{x_i} terms in 0.0.15 with an equivalent term $(1 - 2x_i)$, we get:

$$\min_{\mathbf{x} \in \{0,1\}^n} \sum_{i=1}^n \theta_i^* \sqrt{e^{-\beta\lambda_i}} (1 - 2x_i) + \sum_{i=1}^n \sum_{j=i+1}^n \theta_{ij}^* \sqrt{e^{-\beta\lambda_{n+i+j}}} (1 - 2x_i)(1 - 2x_j) \quad (0.0.16)$$

Rearranging and combining terms with the same degree, we get the following final expression for AFO which is clearly a Binary quadratic programming (BQP) problem:

$$\min_{\mathbf{x} \in \{0,1\}^n} \mathbf{b}^T \mathbf{x} + \mathbf{x}^T \mathbf{A} \mathbf{x} \quad (0.0.17)$$

where \mathbf{b} and \mathbf{A} are defined as given below:

$$b_i = -2 \left(\theta_i^* \sqrt{e^{-\beta \lambda_i}} + \sum_{j=1}^n \theta_{ij}^* \sqrt{e^{-\beta \lambda_{n \cdot i + j}}} \right), 1 \leq i \leq n \quad (0.0.18)$$

$$A_{ij} = \delta_{ij} \cdot 4 \left(\theta_{ij}^* \sqrt{e^{-\beta \lambda_{n \cdot i + j}}} \right), 1 \leq i, j \leq n \quad (0.0.19)$$

where $\delta_{ij} = 1$ if $j > i$ and 0 otherwise.

Efficient submodular-relaxation solver. BQP is a well-studied problem in multiple areas including computer vision (Gorelick et al. 2014). Motivated by the prior success of submodular-relaxation methods in the structured prediction area (Tang, Ayed, and Boykov 2014), we study a fast and scalable approach for solving AFO problems based on recent advances in submodular-relaxation (Ito and Fujimaki 2016). The key idea is to construct an efficient relaxed problem with some unknown parameters and optimize those parameters iteratively to improve the accuracy of solutions. We provide a high-level sketch of the overall algorithm (Deshwal, Belakaria, and Doppa 2020) below.

The objective in 0.0.17 is called as *submodular* if $A_{ij} \leq 0 \forall i, j$. Submodular functions can be exactly minimized by fast strongly polynomial-time graph-cut algorithms (Fujishige 2005). However, in general, the objective might not be submodular. Therefore, a submodular relaxation to the objective in Equation 0.0.17 (Ito and Fujimaki 2016) is constructed by lower bounding the positive terms A^+ of A :

$$\mathbf{x}^T (\mathbf{A}^+ \circ \Gamma) \mathbf{1} + \mathbf{1}^T (\mathbf{A}^+ \circ \Gamma) \mathbf{x} - \mathbf{1}^T (\mathbf{A}^+ \circ \Gamma) \mathbf{1} \leq \mathbf{x}^T \mathbf{A}^+ \mathbf{x} \quad (0.0.20)$$

where $\mathbf{A} = \mathbf{A}^+ + \mathbf{A}^-$,

$$\mathbf{A}^+ = A_{ij} \text{ if } A_{ij} \geq 0 \text{ and } 0 \text{ otherwise} \quad (0.0.21)$$

$$\mathbf{A}^- = A_{ij} \text{ if } A_{ij} \leq 0 \text{ and } 0 \text{ otherwise} \quad (0.0.22)$$

and Γ stands for unknown relaxation parameters. The accuracy of this relaxed problem critically depends on Γ . Therefore, we perform optimization over Γ parameters after initializing them by repeating the following two steps.

1. Solve submodular relaxation lower bound of the BQP objective in 0.0.17 using minimum graph-cut algorithms.
2. Update relaxation parameters Γ via gradient descent

This submodular-relaxation based AFO solver scales gracefully with the increasing input dimension n because of the strongly polynomial complexity $O(n^3)$ of minimum graph cut algorithms (Ahuja, Magnanti, and Orlin 1988).

Experiments and Results

Experimental setting. The source code for state-of-the-art baseline methods was taken from their github links: COMBO (<https://github.com/QUVA-Lab/COMBO>), BOCS (<https://github.com/baptistar/BOCS>), and SMAC (<https://github.com/automl/SMAC3>). For a fair comparison, the priors for all GP hyper-parameters and their

posterior computation were kept the same for both COMBO and MerCBO. COMBO employs a separate hyper-parameter β_i for each dimension to enforce sparsity, which is important in certain applications. For MerCBO also, we include sparsity in sampling θ for Thompson Sampling for all benchmarks other than Ising: $\mathcal{N}((\Phi^T \Phi + \sigma^2 \Upsilon^{-1})^{-1} \Phi^T \mathbf{y}, (\Phi^T \Phi + \sigma^2 \Upsilon^{-1})^{-1} \sigma^2)$, where $\Upsilon = \text{diag}(\beta_1, \beta_2, \dots, \beta_1 \beta_2, \dots, \beta_{n-1} \beta_n)$. It should be noted that we are not introducing any more hyper-parameters than COMBO, but use them in a *strong heirarchical* sense (Bien, Taylor, and Tibshirani 2013).

We ran five iterations of submodular relaxation approach for solving AFO problems and observed convergence. We ran each method on all the benchmarks for 25 random runs and report mean and two times the standard error for results.

Sequential Design Optimization Benchmarks

We employed four diverse benchmarks for sequential design: *function evaluations are performed sequentially*.

Ising sparsification. The probability distribution $p(z)$ (Baptista and Poloczek 2018; Oh et al. 2019) defined by Zero-field Ising model I_p is parametrized by a symmetric interaction matrix J^p whose support is represented as a graph G^p . The goal in this problem is to approximate $p(z)$ with another distribution $q(z)$ such that the number of edges in G^q are minimized. The objective function is defined as:

$$\min_{\mathbf{x} \in \{0,1\}^n} D_{KL}(p||q) + \lambda \|\mathbf{x}\|_1$$

where D_{KL} is the KL-divergence between p and q , and λ is a regularization parameter. The results for this 24-dimensional domain are shown in Figure 1a and 1b. In COMBO (Oh et al. 2019), BOCS was shown to perform better than COMBO on this domain. However, MerCBO shows the best performance among all methods signifying that the proposed approach augments the performance of GP surrogate model.

Low auto-correlation binary sequences (LABS). This problem has diverse applications in multiple fields (Bernasconi 1987; Packebusch and Mertens 2015) including communications where it is used in high-precision interplanetary radar measurements (Shapiro et al. 1968). The goal is to find a binary sequence $\{1, -1\}$ of length n that maximizes the *Merit factor (MF)* defined as follows:

$$\max_{\mathbf{x} \in \{1,-1\}^n} \text{MF}(\mathbf{x}) = \frac{n^2}{E(\mathbf{x})},$$

$$E(\mathbf{x}) = \sum_{k=1}^{n-1} \left(\sum_{i=1}^{n-k} x_i x_{i+k} \right)^2$$

where $E(\mathbf{x})$ is the energy of the sequence. This domain allows us to evaluate all methods on large input dimensions. Results with 40 and 50 dimensions are shown in Figure 1c and 1d respectively. MerCBO shows best performance among all methods showing its effectiveness. COMBO shows poor convergence and matches MerCBO only on 40 dimensional case at the end of allocated budget. SMAC and BOCS show poor performance resulting from their inability to model the underlying structure properly.

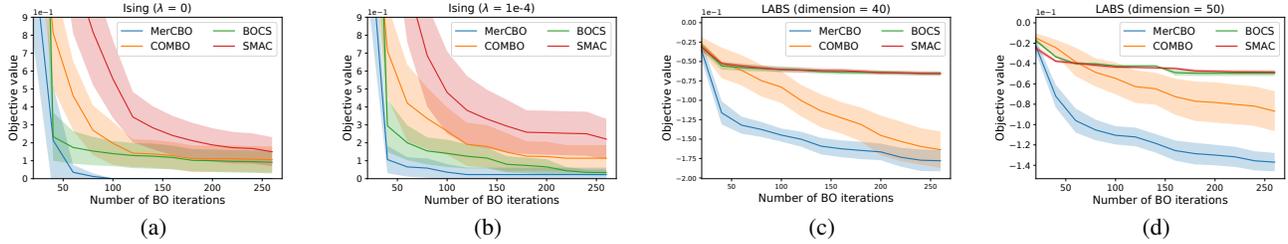


Figure 1: Results for Ising and LABS domains.

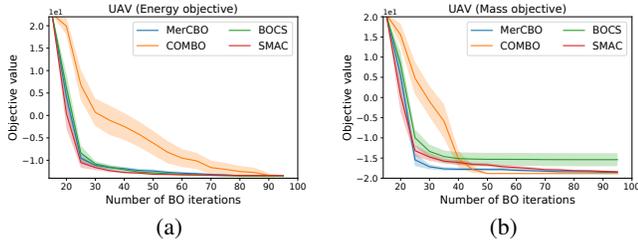


Figure 2: Results for power system design of UAVs.

Electrified aviation power system design. We consider the design of power system for unmanned aerial vehicles (UAVs) via a time-based static simulator (Thurlbeck and Cao 2019; Belakaria et al. 2020b). Each structure is specified using five design variables such as the battery pack configuration and motor size. Evaluation of each design requires performing a computationally-expensive simulation. We consider two variants of this design task: (1) Optimize *total energy*; and (2) Optimize *mass*. We employed a dataset of 250,000 candidate designs for our experiments. The categorical variables are encoded as 16 bit binary variables.

To make the benchmark challenging, we initialized the surrogate models by randomly selecting from worst (in terms of objective) 10% structures. Results for this benchmark are shown in Figure 2. Interestingly, both BOCS and SMAC shows good performance on both mass and total energy objectives of this benchmark. Since EI has the tendency to be relatively more exploitative, COMBO shows poor convergence but reaches the best value at the end of allocated budget. MerCBO converges much faster on both problems, but the performance plateaus out on the mass objective. We attribute this behavior to the naturally exploratory behavior of the Thompson sampling acquisition function.

Parallel Biological Sequence Design

Motivation. Design of optimized biological sequences such as DNA and proteins is a fundamental problem with many important applications (Yang, Wu, and Arnold 2019; Angermueller et al. 2020; Yang et al. 2019). These design problems have the following requirements: uncover a diverse set of structures (*diversity*); select a large batch of structures in each round to perform parallel evaluations (*large-scale par-*

allel experiments); and use parallel experimental resources to accelerate optimization (*real-time accelerated design*).

Benefits of Thompson sampling (TS). TS is a powerful approach that meets these requirements (Hernández-Lobato et al. 2017). Any acquisition function is defined as the expectation of a utility function under the posterior predictive distribution $p(y|\mathbf{x}, D) = \int p(y|\theta)p(\theta|\mathbf{x}, D)$. TS approximates this posterior by a single sample $\theta^* \sim p(\theta|\mathbf{x}, D)$ which inherently enforces exploration due to the variance of this Monte-Carlo approximation. In MerCBO, we can sample as many θ (via Mercer features) as required and can employ scalable AFO solvers for each sample in parallel. However, parallel version of EI (acquisition function employed in COMBO) does not meet most of the requirements as it selects the batch of structures for evaluation sequentially.

Experimental setup. We evaluate TS and EI with GP based models on diverse DNA sequence design problems. The goal is to find sequences that maximize the binding activity between a variety of human transcription factors and every possible length-8 DNA sequence (Barrera et al. 2016; Angermueller et al. 2020). Categorical variables are encoded as 2 bit binary variables. We multiply objective values by -1 to convert the problem into minimization for consistency. We compare the parallel version of EI as proposed in (Snoek, Larochelle, and Adams 2012) and used in multiple earlier works (Hernández-Lobato et al. 2017; Kandasamy et al. 2018) with parallel TS. For a batch of B evaluations in each iteration, parallel-EI works by picking the first input in the same way as in the sequential setting and selects the remaining inputs $j = 2$ to B by maximizing the expected EI acquisition function under all possible outcomes of the first $j - 1$ pending evaluations (aka fantasies (Snoek, Larochelle, and Adams 2012)). On the other hand, TS is easily scalable and parallelizable by sampling B θ 's from the GP posterior and optimizing each of them independently with our MerCBO approach. We run both TS and EI experiments on a 32 core Intel(R) Core(TM) i9-7960X CPU @ 2.80GHz machine. All reported time units are in seconds on the same machine.

Discussion of results. Figure 3 shows the canonical comparison of parallel TS with parallel EI (p-EI) on one transcription factor from the DNA-binding affinity benchmark. The numbers within the plots show the mean objective value for a budget of 500 evaluations. Although parallel-EI is slightly better in terms of optimization performance, TS is extremely fast and is useful for practitioners in time-constrained appli-

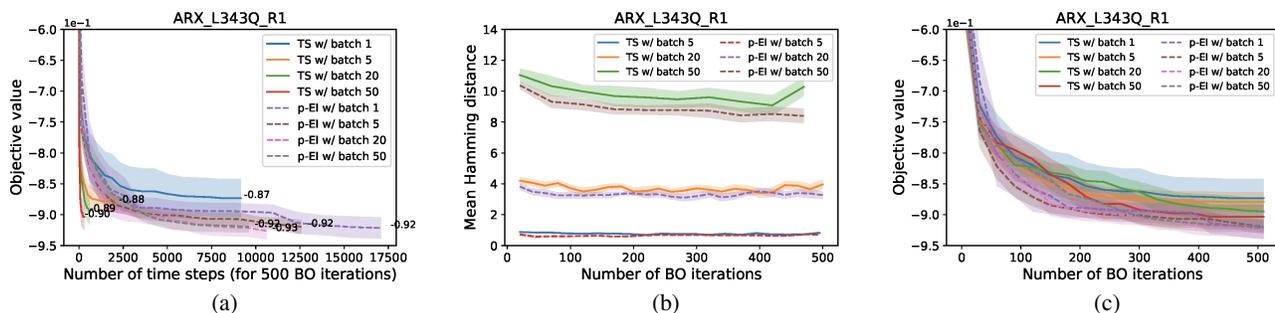


Figure 3: Representative results on biological sequence design problem for one transcription factor.

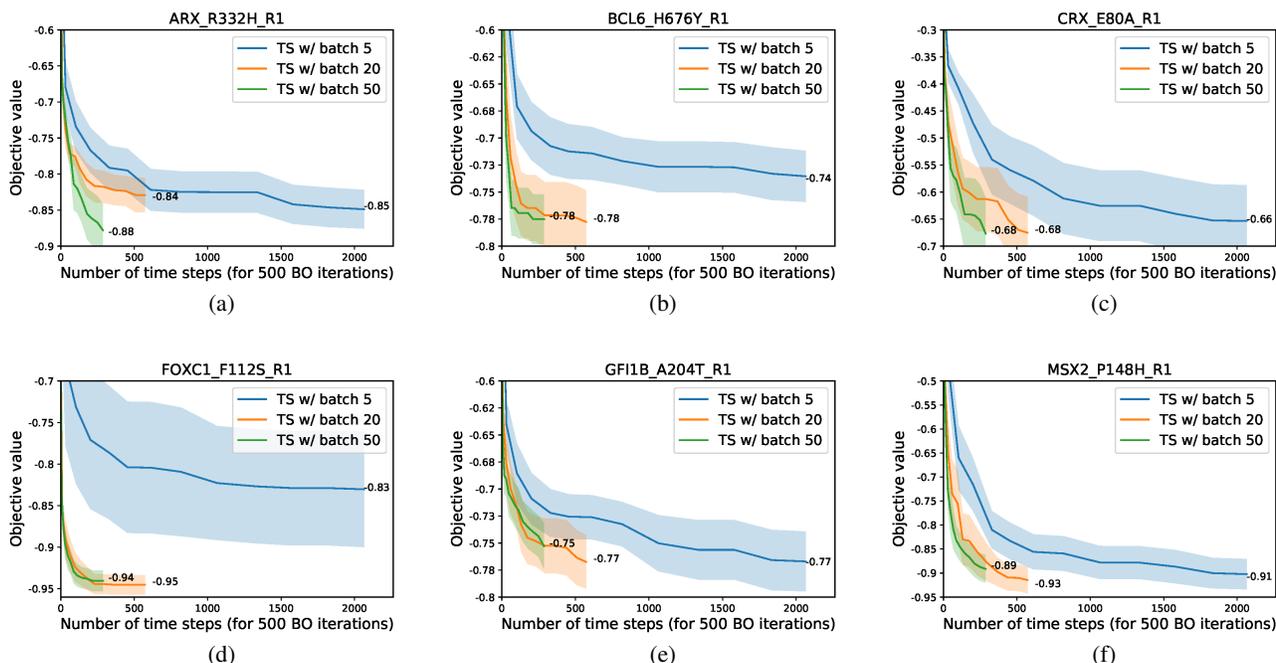


Figure 4: Results on biological sequence design with Thompson sampling for six transcription factors.

cations including drug and vaccine design. The diversity of best batch of sequences is equally important to hedge against the possibility of some chosen candidate designs failing on downstream objectives (Angermueller et al. 2020). Figure 3b shows the results comparing the diversity of sequences (on mean Hamming distance metric) found by TS versus parallel-EI. There are two key observations that can be made from this figure. First, increasing the batch size increases the diversity of sequences. Second, TS finds comparatively more diverse sequences than parallel-EI.

Interestingly, performance of parallel TS improves with increasing batch size. To justify this observation, we further evaluated parallel TS on six other transcription factors as shown in Figure 4. As evident from the figure, performance of parallel TS improves with increasing batch size on five out of six benchmarks. This shows that the exploratory behavior

of TS, which can be bad in some sequential settings, helps in better performance for the parallel setting.

Summary and Future Work

We introduced an efficient approach called MerCBO for optimizing expensive black-box functions over discrete spaces. MerCBO is based on computing Mercer features for diffusion kernels and fast submodular relaxation based acquisition function optimization. We showed that MerCBO produces similar or better performance than state-of-the-art methods on multiple real-world problems. Future work includes using the proposed Mercer features for entropy based acquisition functions such as max-value entropy search.

Acknowledgements

This research is supported in part by National Science Foundation (NSF) grants IIS-1845922, OAC-1910213, and CNS-1955353. The views expressed are those of the authors and do not reflect the official policy or position of the NSF. Jana Doppa would like to thank Tom Dietterich for useful discussions.

References

- Ahuja, R. K.; Magnanti, T. L.; and Orlin, J. B. 1988. *Network Flows: Theory, Algorithms, and Applications*. USA: Prentice-Hall, Inc. ISBN 013617549X.
- Angermueller, C.; Belanger, D.; Gane, A.; Mariet, Z.; Dohan, D.; Murphy, K.; Colwell, L.; and Sculley, D. 2020. Population-Based Black-Box Optimization for Biological Sequence Design. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.
- Baptista, R.; and Poloczek, M. 2018. Bayesian Optimization of Combinatorial Structures. In *Proceedings of the 35th International Conference on Machine Learning*, 462–471.
- Barrera, L. A.; Vedenko, A.; Kurland, J. V.; Rogers, J. M.; Gisselbrecht, S. S.; Rossin, E. J.; Woodard, J.; Mariani, L.; Kock, K. H.; Inukai, S.; et al. 2016. Survey of variation in Human Transcription Factors reveals prevalent DNA Binding changes. *Science* 351(6280): 1450–1454.
- Belakaria, S.; Deshwal, A.; and Doppa, J. R. 2020a. Max-value Entropy Search for Multi-Objective Bayesian Optimization with Constraints. *CoRR* abs/2009.01721. URL <https://arxiv.org/abs/2009.01721>.
- Belakaria, S.; Deshwal, A.; and Doppa, J. R. 2020b. Multi-Fidelity Multi-Objective Bayesian Optimization: An Output Space Entropy Search Approach. In *AAAI conference on Artificial Intelligence (AAAI)*.
- Belakaria, S.; Deshwal, A.; Jayakodi, N. K.; and Doppa, J. R. 2020a. Uncertainty-Aware Search Framework for Multi-Objective Bayesian Optimization. In *AAAI conference on Artificial Intelligence (AAAI)*.
- Belakaria, S.; Jackson, D.; Cao, Y.; Doppa, J. R.; and Lu, X. 2020b. Machine Learning Enabled Fast Multi-Objective Optimization for Electrified Aviation Power System Design. In *IEEE Energy Conversion Congress and Exposition (ECCE)*.
- Belakaria et al. 2019. Max-value Entropy Search for Multi-Objective Bayesian Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 7823–7833.
- Bernasconi, J. 1987. Low Autocorrelation Binary Sequences: Statistical Mechanics and Configuration Space Analysis. *Journal de Physique* 48(4): 559–567.
- Bien, J.; Taylor, J.; and Tibshirani, R. 2013. A Lasso for Hierarchical Interactions. *Annals of statistics* 41(3): 1111.
- Chowdhury, S. R.; and Gopalan, A. 2017. On Kernelized Multi-armed Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, 844–853.
- Dadkhahi, H.; Shanmugam, K.; Rios, J.; Das, P.; Hoffman, S. C.; Loeffler, T. D.; and Sankaranarayanan, S. 2020. Combinatorial Black-Box Optimization with Expert Advice. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 1918–1927.
- Deshwal, A.; Belakaria, S.; and Doppa, J. R. 2020. Scalable Combinatorial Bayesian Optimization with Tractable Statistical models. *CoRR* abs/2008.08177. URL <https://arxiv.org/abs/2008.08177>.
- Deshwal, A.; Belakaria, S.; Doppa, J. R.; and Fern, A. 2020. Optimizing Discrete Spaces via Expensive Evaluations: A Learning to Search Framework. In *AAAI*, 3773–3780.
- Florkowski III, S. F. 2008. Spectral Graph Theory of the Hypercube. Technical report, NAVAL POSTGRADUATE SCHOOL MONTEREY CA. URL <https://apps.dtic.mil/dtic/tr/fulltext/u2/a493796.pdf>.
- Fujishige, S. 2005. *Submodular Functions and Optimization*. ISSN. Elsevier Science.
- Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; and Aspuru-Guzik, A. 2018. Automatic Chemical Design using a Data-driven Continuous Representation of Molecules. *ACS central science* 4(2): 268–276.
- Gorelick, L.; Boykov, Y.; Veksler, O.; Ben Ayed, I.; and DeLong, A. 2014. Submodularization for Binary Pairwise Energies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1154–1161.
- Hammack, R.; Imrich, W.; and Klavžar, S. 2011. *Handbook of product graphs*. CRC press.
- Hernández-Lobato, J. M.; Hoffman, M. W.; and Ghahramani, Z. 2014. Predictive Entropy Search for Efficient Global Optimization of Black-box Functions. In *Advances in neural information processing systems*, 918–926.
- Hernández-Lobato, J. M.; Requeima, J.; Pyzer-Knapp, E. O.; and Aspuru-Guzik, A. 2017. Parallel and Distributed Thompson sampling for Large-scale Accelerated Exploration of Chemical Space. In *Proceedings of the 34th International Conference on Machine Learning*, 1470–1479.
- Hutter, F.; Hoos, H. H.; and Leyton-Brown, K. 2010. Sequential Model-Based Optimization for General Algorithm Configuration (extended version). Technical Report TR-2010-10, University of British Columbia, Department of Computer Science. Available online: <http://www.cs.ubc.ca/~hutter/papers/10-TR-SMAC.pdf>.
- Hutter, F.; Hoos, H. H.; and Leyton-Brown, K. 2011. Sequential Model-based Optimization for General Algorithm Configuration. In *International conference on Learning and Intelligent Optimization*, 507–523.
- Ito, S.; and Fujimaki, R. 2016. Large-Scale Price Optimization via Network Flow. In *Advances in Neural Information Processing Systems*, 3855–3863.
- Janz, D.; van der Westhuizen, J.; Paige, B.; Kusner, M. J.; and Hernández-Lobato, J. M. 2018. Learning a Generative Model for validity in Complex Discrete Structures. *International Conference on Learning Representations*.

- Kandasamy, K.; Krishnamurthy, A.; Schneider, J.; and Poczos, B. 2018. Parallelised Bayesian Optimisation via Thompson Sampling. In *Proceedings of Machine Learning Research*, volume 84, 133–142.
- Kondor, R.; and Lafferty, J. 2002. Diffusion Kernels on Graphs and other Discrete Structures. In *Proceedings of the 19th International Conference on Machine Learning*, volume 2002, 315–322.
- Kusner, M. J.; Paige, B.; and Hernández-Lobato, J. M. 2017. Grammar Variational Autoencoder. In *Proceedings of the 34th International Conference on Machine Learning*, ICML'17. JMLR.org.
- Leprêtre, F.; Verel, S.; Fonlupt, C.; and Marion, V. 2019. Walsh functions as surrogate model for pseudo-boolean optimization problems. In *Proceedings of the Genetic and Evolutionary Computation Conference*, 303–311.
- Oh, C.; Tomczak, J.; Gavves, E.; and Welling, M. 2019. Combinatorial Bayesian Optimization using the Graph Cartesian Product. In *Advances in Neural Information Processing Systems*, 2910–2920.
- Packebusch, T.; and Mertens, S. 2015. Low Autocorrelation Binary Sequences. *Journal of Physics A: Mathematical and Theoretical*, 49 (2016) 165001 doi:10.1088/1751-8113/49/16/165001.
- Rasmussen, C. E.; and Williams, C. K. I. 2006. *Gaussian Processes for Machine Learning*. Adaptive computation and machine learning. MIT Press.
- Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; and de Freitas, N. 2016. Taking the Human out of the loop: A review of Bayesian Optimization. *Proceedings of the IEEE* 104(1): 148–175.
- Shapiro, I. I.; Pettengill, G. H.; Ash, M. E.; Stone, M. L.; Smith, W. B.; Ingalls, R. P.; and Brockelman, R. A. 1968. Fourth Test of General Relativity: preliminary results. *Physical Review Letters* 20(22): 1265.
- Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*, 2960–2968.
- Tang, M.; Ayed, I. B.; and Boykov, Y. 2014. Pseudo-Bound Optimization for Binary Energies. In *European Conference on Computer Vision*, 691–707. Springer.
- Thurlbeck, A. P.; and Cao, Y. 2019. Analysis and Modeling of UAV Power System Architectures. In *2019 IEEE Transportation Electrification Conference and Expo (ITEC)*, 1–8. IEEE.
- Wang, Z.; and Jegelka, S. 2017. Max-value Entropy Search for Efficient Bayesian Optimization. In *Proceedings of the 34th International Conference on Machine Learning*, 3627–3635.
- Yackley, B.; Corona, E.; and Lane, T. 2009. Bayesian Network Score Approximation using a Metagraph Kernel. In *Advances in Neural Information Processing Systems*, 1833–1840.
- Yang, K. K.; Chen, Y.; Lee, A.; and Yue, Y. 2019. Batched Stochastic Bayesian Optimization via Combinatorial Constraints Design. In Chaudhuri, K.; and Sugiyama, M., eds., *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, 3410–3419. PMLR.
- Yang, K. K.; Wu, Z.; and Arnold, F. H. 2019. Machine-learning-guided directed evolution for protein engineering. *Nature methods* 16(8): 687–694.