MACHINE LEARNING
Science and Technology

PURPOSE-LED PUBLISHING™ Part of the MACHINE LEARNING Series

ACCEPTED MANUSCRIPT • OPEN ACCESS

Physics-informed Bayesian optimization of expensive-to-evaluate black-box functions

To cite this article before publication: Ivan Sekulic *et al* 2025 *Mach. Learn.: Sci. Technol.* in press https://doi.org/10.1088/2632-2153/ae1f5f

Physics-informed Bayesian optimization of expensive-to-evaluate black-box functions

Ivan Sekulic,[1,2,*] Jonas Schaible,[3,2] Gabriel Müller,[4] Matthias Plock,[1,2] Sven Burger,[1,2] Víctor J Martínez-Lahuerta,[4] Naceur Gaaloul,[4] and Philipp-Immanuel Schneider[1,2]

[1] *JCMwave GmbH, Bolivarallee 22, 14050 Berlin, Germany*

[2] *Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany*

[3] *Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, 12489 Berlin, Germany*

[4] *Leibniz Universität Hannover, Institut für Quantenoptik, Welfengarten 1, 30167 Hannover, Germany*

Bayesian optimization with Gaussian process surrogates is a popular approach for optimizing expensive-to-evaluate functions in terms of time, energy, or computational resources. Typically, a Gaussian process models a scalar objective derived from observed data. However, in many real-world applications, the objective is a combination of multiple outputs from physical experiments or simulations. Converting these multidimensional observations into a single scalar can lead to information loss, slowing convergence and yielding suboptimal results. To address this, we propose to use multi-output Gaussian processes to learn the full vector of observations directly, before mapping them to the scalar objective via an inexpensive analytical function. This *physics-informed* approach retains more information from the underlying physical processes, improving surrogate model accuracy. As a result, the approach accelerates optimization and produces better final designs compared to standard implementations.

I. INTRODUCTION

Bayesian decision theory with Gaussian processes (GPs) has become a powerful machine learning tool for approximating and optimizing cost functions in physical processes [1–4].

*Electronic address: ivan.sekulic@jcmwave.com

Using Bayesian conditioning on all observed data, GPs construct a probabilistic surrogate model of the objective function. This enables accurate predictions in regions of the parameter space with sufficient data but also allows to quantify uncertainty in regions where data is scarce. Bayesian optimization (BO) strategies use this uncertainty information for strategically balancing exploration of the parameter space with exploitation of high-performance regions. Typically, this improves the optimization efficiency and reduces the number of required evaluations to find the globally optimal solution [3, 4].

The balanced exploration and exploitation of the parameter space works best for *medium-dimensional* problems with typically 4 to 15 parameters. In higher dimensions with abundant local minima, the global search for optimum becomes impractical due to the exponential growth of the parameter space known as *curse of dimensionality*. In these cases, local gradient-based strategies like the quasi-Newton L-BFGS [5] or SLSQP [6] are often more suitable.

BO is particularly valuable for costly objective functions – whether in terms of time, computational resources, or energy consumption – where the gain of reducing the required number of iterations outweighs the computational overhead of training and evaluating GPs. In nanophotonics, for example, optimizing device designs often involves minimizing a cost function derived from the solution of Maxwell's equations. Thus, the optimization requires repeated calls to numerical solvers which are computationally demanding [7–9].

The GP's accuracy has a strong impact on the performance of BO. The next sample to be evaluated is chosen at the maximum of some acquisition function. The acquisition function uses the GP's probability distribution to determine the utility of an evaluation. For the case of a minimization, this utility is large if the mean of the distribution is sufficiently small (exploitation) or if the uncertainty is sufficiently large (exploration). Inaccurate predictions of the mean or uncertainty can thus deteriorate the local convergence or the balance between exploration and exploitation.

In order to increase the fidelity of the predicted distributions, we propose to leverage the knowledge of physical inputs to the objective function. Throughout this work, we denote by $\boldsymbol{f}(\boldsymbol{p}) : \mathcal{X} \to \mathbb{R}^k$ a vector-valued observable coming from a numerical simulation or measurements of a physical experiment, and by $g(\boldsymbol{p}) : \mathcal{X} \to \mathbb{R}$ a scalar-valued objective function constructed from the components of $\boldsymbol{f}(\boldsymbol{p})$. Here, $\boldsymbol{p} \in \mathcal{X} \subseteq \mathbb{R}^d$ is the tunable parameter vector in a d-dimensional search space. To bring the problem into practical light, we consider a

typical inverse design problem in nanophotonics. The scalar objective $g(\boldsymbol{p})$, which is to be minimized over the search space spanned by the device's geometric parameters collected in vector $\boldsymbol{p}$, is commonly defined as the deviation from the desired functionality of the device. For example, in the inverse design of the beam-splitter, which we describe in detail in subsection IV A, the objective function $g(\boldsymbol{p})$ quantifies the deviation from a perfectly uniform distribution of optical power carried by diffraction orders. In this case, the power fluxes associated with each diffraction order are collected into the vector-valued observable $\boldsymbol{f}(\boldsymbol{p})$, and the scalar objective $g(\boldsymbol{p})$ is defined as the variance among the components of $\boldsymbol{f}(\boldsymbol{p})$. In this work we show, that training the GP on values of g while neglecting the individual components of the observable $\boldsymbol{f}$ is associated with an information loss and thus limits the efficiency of classical BO. On the other hand, multi-output GPs trained on observations of $\boldsymbol{f}$ produce multivariate-normal predictions that can be used to determine more faithful estimates of the distribution of possible values of g and thus improve the performance of BO.

The concept of BO for general composite functions has been introduced by Astudillo and Frazier [10]. They propose to train multi-output GPs to learn a multivariate normal random variable $\hat{\boldsymbol{f}}(\boldsymbol{p}) \sim \mathcal{N}\left(\boldsymbol{\mu}(\boldsymbol{p}), \boldsymbol{\Sigma}(\boldsymbol{p})\right)$ of observables. The general mapping $\hat{\boldsymbol{f}} \mapsto \hat{g}$ leads then to a non-Gaussian, non-analytical distribution of $\hat{g}$. However, the acquisition functions such as, for example, expected improvement, can be evaluated numerically through a Monte-Carlo scheme.

For the case of minimizing a Chi-squared deviation $g(\boldsymbol{p}) = \sum_{i=1}^{k} \left(f_i(\boldsymbol{p}) - t_i\right)^2$ to a target vector $\boldsymbol{t} = (t_1, \cdots, t_k)^\top$, Uhrenholt and Jensen [11] proposed to train k independent GP models that predict Gaussian distributions $\hat{f}_i(\boldsymbol{p}) \sim \mathcal{N}\big(\mu_i(\boldsymbol{p}), \sigma_i^2(\boldsymbol{p})\big)$, $1 \leq i \leq k$. The distribution of the random variable $\hat{\chi}^2(\boldsymbol{p}) = \sum_{i=1}^{k} \left(\hat{f}_i(\boldsymbol{p}) - t_i\right)^2$ is approximated by an analytic non-central Chi-squared distribution in order to define analytic acquisition functions. The approach works well for a moderate number of output dimensions k. For inverse measurement applications with $k > 100$ measurement results, the use of k independent GPs leads to large computational costs and seriously underestimates the probability of finding small Chi-squared values. Plock *et al.* have extended the approach to large k values by introducing an effective degree of freedom $\tilde{k} < k$ and using a multi-output GP with a shared covariance function [12].

In this work, we consider the optimization of physical systems based on a minimization of

an objective $g(\boldsymbol{p})$ that depends in an arbitrary nonlinear manner on the vectorial physical response $\boldsymbol{f}(\boldsymbol{p})$ of the system [10]. We benchmark this physics-informed BO method against standard BO using a single-output GP and other heuristic optimization methods on three real-world problems from distinct engineering domains. We provide explanations for the observed superior optimization performance of physics-informed BO by analyzing the prediction accuracy of its surrogate.

This article is structured as follows: In section II, for the sake of completeness, we introduce the notation and the standard theory around single-output BO. Section III details the proposed physics-informed method and highlights its key benefits and possible shortcomings. Section IV presents numerical experiments drawn from real-world engineering applications which demonstrate the significant advantages of physics-informed BO. We discuss the observed results and conclude the work in section V.

II. SINGLE-OUTPUT BAYESIAN OPTIMIZATION

This section provides a concise overview of standard Bayesian optimization utilizing single-output Gaussian processes.

A. Single-output Gaussian process

Bayesian inference starts with a definition of the prior distribution over functions which reflects our assumptions about the objective before any data is observed [1, 3]. In the context of GPs, the prior is specified by a mean function $\mu : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ and a positive definite covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Here we consider the common choice of a constant mean function

$$\mu(\boldsymbol{p}) = \mu_0, \tag{1}$$

and a stationary 5/2-Matérn covariance function [13]

$$k(\boldsymbol{p}, \boldsymbol{p}') = \sigma_0^2 \left(1 + \sqrt{5}\, d + \frac{5}{3}\, d^2\right) \exp\left(-\sqrt{5}\, d\right), \tag{2}$$

that depends only on the variance σ_0 and the relative distance d between the points $\boldsymbol{p}$ and $\boldsymbol{p}'$

$$d(\boldsymbol{p}, \boldsymbol{p}') = \sqrt{\sum_{i=1}^{d} \frac{(p_i - p_i')^2}{l_i^2}} \tag{3}$$

normalized by length scales $l_1, \cdots, l_d$ for each dimension. For any finite set of points $\boldsymbol{P} = [\boldsymbol{p}_1, ..., \boldsymbol{p}_N]$, the random function values $\hat{\boldsymbol{g}}(\boldsymbol{P}) = [\hat{g}(\boldsymbol{p}_1), \cdots, \hat{g}(\boldsymbol{p}_N)]^\top$ follow a multivariate normal (MVN) distribution

$$\hat{\boldsymbol{g}}(\boldsymbol{P}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K}) \text{ with}$$

$$\boldsymbol{\mu} = [\mu(\boldsymbol{p}_1), \cdots, \mu(\boldsymbol{p}_N)]^\top = \mu_0 \mathbf{1}, \tag{4}$$

$$\boldsymbol{K} = [k(\boldsymbol{p}_i, \boldsymbol{p}_j)]_{1 \leq i,j \leq N}, \tag{5}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{K}$ denote prior distribution mean vector and covariance matrix of points in $\boldsymbol{P}$, respectively. We consider partitioning the set $\boldsymbol{P} = \boldsymbol{P}_\mathrm{t} \cup \boldsymbol{P}_\mathrm{q}$ into the subset $\boldsymbol{P}_\mathrm{t}$ of points with known objective values $\boldsymbol{g}(\boldsymbol{P}_\mathrm{t})$ (training points) and the subset $\boldsymbol{P}_\mathrm{q}$ of query points. Using Bayesian inference, the conditional distribution of the function values at the query points $\boldsymbol{g}(\boldsymbol{P}_q)$ is

$$P(\boldsymbol{g}(\boldsymbol{P}_q) \,|\, \boldsymbol{g}(\boldsymbol{P}_t)) = \frac{P(\boldsymbol{g}(\boldsymbol{P}))}{P(\boldsymbol{g}(\boldsymbol{P}_t))}, \tag{6}$$

where $P(\cdot)$ is the probability density function of an observation. This conditional MVN distribution has the closed form [1, 3]

$$\hat{\boldsymbol{g}}(\boldsymbol{P}_\mathrm{q}) \,|\, \boldsymbol{g}(\boldsymbol{P}_\mathrm{t}) \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{K}') \text{ with}$$

$$\boldsymbol{\mu}' = \mu_0 \mathbf{1}_\mathrm{q} + \boldsymbol{K}_\mathrm{tq}^\top \boldsymbol{K}_\mathrm{tt}^{-1} (\boldsymbol{g}(\boldsymbol{P}_\mathrm{t}) - \mu_0 \mathbf{1}_\mathrm{t}), \tag{7}$$

$$\boldsymbol{K}' = \boldsymbol{K}_\mathrm{qq} - \boldsymbol{K}_\mathrm{tq}^\top \boldsymbol{K}_\mathrm{tt}^{-1} \boldsymbol{K}_\mathrm{tq}. \tag{8}$$

Here, $\mathbf{1}_\mathrm{q}$ and $\mathbf{1}_\mathrm{t}$ are vectors with entries 1 of dimension $\|\boldsymbol{P}_\mathrm{q}\|$ and $\|\boldsymbol{P}_\mathrm{t}\|$. $\boldsymbol{K}_\mathrm{qq}$ and $\boldsymbol{K}_\mathrm{tt}$ are the covariance matrices of points in $\boldsymbol{P}_\mathrm{q}$ and $\boldsymbol{P}_\mathrm{t}$, respectively. $\boldsymbol{K}_\mathrm{tq}$ denotes the cross-covariance matrix between the points in $\boldsymbol{P}_\mathrm{t}$ and $\boldsymbol{P}_\mathrm{q}$, while $\boldsymbol{\mu}'$ and $\boldsymbol{K}'$ represent conditional mean vector and covariance matrix, respectively.

The prior distribution, and thus the conditional posterior distribution, depend strongly on the values of the hyperparameters for the mean μ_0, the variance σ_0^2, and the length scales $l_1, \cdots, l_d$. Therefore, these are adapted to the data by choosing values that maximize the probability density $P(\boldsymbol{g}(\boldsymbol{P}_t))$ with $\hat{\boldsymbol{g}}(\boldsymbol{P}_t) \sim \mathcal{N}(\mu_0 \mathbf{1}_\mathrm{t}, \boldsymbol{K}_\mathrm{tt})$.

The inversion of the covariance matrix $\boldsymbol{K}_\mathrm{tt}$ of the training points is not done directly. Instead, a Cholesky factorization is performed for the sake of better numerical stability. This linear algebra operation, required at every iteration of BO, has a computational complexity of $\mathcal{O}(T^3)$, in the number of training points T.

B. Optimization policy and acquisition function

Bayesian optimization is a sequential algorithm; given the T training points in the search space $\mathcal{X}$, the algorithm selects a next training point $\boldsymbol{p}_{T+1}$ that maximizes an *acquisition function* [2]. This function describes the utility of a training point and is designed to guide the search for the optimum by balancing exploration (sampling in areas of parameter space with high uncertainty) and exploitation (sampling in areas where the function is predicted to be optimal). The goal is to reduce the number of expensive objective function evaluations needed to find the global optimum by using information from previous evaluations. In the benchmark examples presented in section IV, each optimization run entails a series of calls to the numerical solver. For the particular problems considered, a single call may take from a few seconds to several minutes. However, as the numerical problem becomes more complex, it is not uncommon for a single evaluation to require several hours, even when executed on high-performance computing architectures. One of the most common acquisition functions is *expected improvement* (EI), defined as:

$$\alpha_{\mathrm{EI}}(\boldsymbol{p}_{\mathrm{q}}) = \mathbb{E}\left[\min\left(0, g_{\min} - \hat{g}(\boldsymbol{p}_{\mathrm{q}})\right)\right], \tag{9}$$

where $g_{\min} = \min\{g(\boldsymbol{p}_1), \cdots, g(\boldsymbol{p}_T)\}$ is the best objective evaluation observed so far. Using Eqs. (7) and (8) to obtain the normal distribution for $\hat{g}(\boldsymbol{p}_q)$ (setting $\boldsymbol{P}_{\mathrm{q}} = [\boldsymbol{p}_{\mathrm{q}}]$), the expression (9) is analytically tractable.

The principle of single-output BO is summarized in Fig. 1.

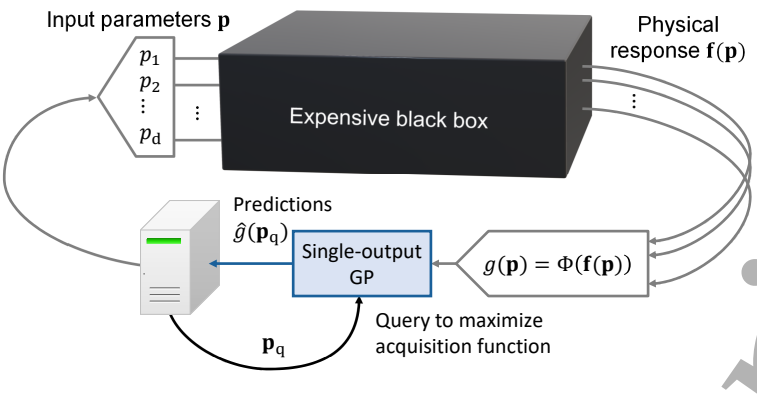

FIG. 1: Principle of single-output Bayesian optimization: The physical response of an expensive black box is mapped to an objective value $g(\boldsymbol{p})$. A single-output GP is trained on each evaluated sample $[g(\boldsymbol{p}_1), \cdots, g(\boldsymbol{p}_T)]$. The acquisition function is maximized by querying the GP obtaining Gaussian predictions $\hat{g}(\boldsymbol{p}_q)$.

III. PHYSICS-INFORMED BAYESIAN OPTIMIZATION

In many optimization scenarios, the physical output of the process under investigation is vectorial. Examples are spectral, temporal or angular responses of a physical system. For the optimization of the system, the response $\boldsymbol{f}(\boldsymbol{p})$ is mapped to a scalar value, i. e.

$$g(\boldsymbol{p}) = \Phi(\boldsymbol{f}(\boldsymbol{p})) \tag{10}$$

to be minimized (or maximized). The mapping $\Phi : \mathbb{R}^k \to \mathbb{R}$ often includes quadratic differences to a target response, but can in general have any form. It formalizes how individual components of $\boldsymbol{f}(\boldsymbol{p})$ are combined into the scalar objective $g(\boldsymbol{p})$. For example, in the beam-splitter case, subsection IV A, Φ computes the variance among the diffraction order power fluxes. In the BEC transport optimization problem, subsection IV B, Φ maps the dynamical properties of the condensate to the final energy objective. Finally, in the color-optimization of the perovskite solar cells, subsection IV C, the learned $\boldsymbol{p}$-dependent variables are mapped into a color-distance measure that serves as the scalar objective. Training a GP only on mapped values $\boldsymbol{g}(\boldsymbol{P}_t) = [g(\boldsymbol{p}_1), \cdots, g(\boldsymbol{p}_T)]$ leads to a loss of information. For example, the Chi-squared deviation $g(\boldsymbol{p}) = \sum_{i=1}^{k} (f_i(\boldsymbol{p}) - t_i)^2$ to a target response $\boldsymbol{t} = (t_1, \cdots, t_k)^\top$ cannot have negative values, while the MVN conditional distribution for $\hat{\boldsymbol{g}}(\boldsymbol{P}_q)$ described by Eqs. (7) and (8) has always a non-vanishing density for negative values [11]. Moreover, a linear approximation of the outputs $f_i(\boldsymbol{p})$ can be used to obtain a higher-order quadratic

approximation of $g(\boldsymbol{p})$.

By training a multi-output GP on the physical response of the system, this information loss can be significantly reduced. This section describes the extension of single-output GPs to multi-output GPs and the mapping of their predictions in order to determine the value of the acquisition function.

A. Multi-output Gaussian process

To define a multi-output GP, we specify a multi-output constant prior mean function $\boldsymbol{\mu}_{\mathrm{mo}} : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}^k$ as

$$\boldsymbol{\mu}_{\mathrm{mo}}(\boldsymbol{p}) = \boldsymbol{\mu}_0. \tag{11}$$

The multi-output prior covariance function $\boldsymbol{k}_{\mathrm{mo}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^k \times \mathbb{R}^k$ is defined as

$$\boldsymbol{k}_{\mathrm{mo}}(\boldsymbol{p}, \boldsymbol{p}') = \boldsymbol{K}_0 \cdot \kappa(\boldsymbol{p}, \boldsymbol{p}'), \tag{12}$$

where $\boldsymbol{K}_0$ is a constant symmetric positive semi-definite $k \times k$ matrix describing the covariance between the outputs, and $\kappa(\boldsymbol{p}, \boldsymbol{p}')$ describes the correlation between points in $\mathcal{X}$ [14, 15]. We use the same definition as in Eqs. (1), (2), only with $\sigma_0^2 = 1$ (the variance of all outputs is explained fully by $\boldsymbol{K}_0$).

In analogy to Eqs. (7) and (8), Bayesian inference leads to a MVN distribution for the vectorial output at different query positions $\boldsymbol{P}_{\mathrm{q}}$. For evaluating the acquisition function, we only need the posterior for a single query point $\boldsymbol{P}_{\mathrm{q}} = [\boldsymbol{p}_{\mathrm{q}}]$. The conditional MVN posterior for that single point evaluates to

$$\hat{\boldsymbol{f}}(\boldsymbol{p}_{\mathrm{q}}) \,|\, \boldsymbol{F}_{\mathrm{t}} \sim \mathcal{N}\left(\boldsymbol{\mu}', \boldsymbol{K}'\right) \text{ with}$$

$$[\boldsymbol{\mu}']_i = [\boldsymbol{\mu}_0]_i + \sum_{m,n=1}^{T} \kappa(\boldsymbol{p}_{\mathrm{q}}, \boldsymbol{p}_m) \left[\boldsymbol{\mathcal{K}}_{\mathrm{tt}}^{-1}\right]_{mn} \left(f_i(\boldsymbol{p}_n) - [\boldsymbol{\mu}_0]_i\right), \tag{13}$$

$$\boldsymbol{K}' = \boldsymbol{K}_0 \left[1 - \sum_{m,n=1}^{T} \kappa(\boldsymbol{p}_{\mathrm{q}}, \boldsymbol{p}_m) \left[\boldsymbol{\mathcal{K}}_{\mathrm{tt}}^{-1}\right]_{mn} \kappa(\boldsymbol{p}_n, \boldsymbol{p}_{\mathrm{q}})\right]. \tag{14}$$

Here, $\boldsymbol{F}_{\mathrm{t}} = \left[f_1(\boldsymbol{p}_1), \cdots, f_1(\boldsymbol{p}_T), f_k(\boldsymbol{p}_1), \cdots, f_k(\boldsymbol{p}_T)\right]^\top$ denotes the vector of training data and $\boldsymbol{\mathcal{K}}_{\mathrm{tt}}$ is the $T \times T$ correlation matrix of the training points. Moreover, $\boldsymbol{\mu}'$ and $\boldsymbol{K}'$ represent the conditional mean vector and the covariance matrix evaluated at a point $\boldsymbol{p}_{\mathrm{q}}$ with the indices n and m running over all training data. We note, that due to the multiplicative definition

of the covariance function in Eq. (12), the conditional mean $[\boldsymbol{\mu}']_i$ for each channel $1 \leq i \leq k$ is independent of observations $f_j(\boldsymbol{p}_m)$ in other channels $j \neq i$.

As for the single-output case, the hyperparameters are chosen to maximize the probability density of the training data $P(\boldsymbol{F}_\mathrm{t})$, where $\hat{\boldsymbol{F}}_\mathrm{t} \sim \mathcal{N}\left(\boldsymbol{\mu}_0 \otimes \mathbf{1}_T, \boldsymbol{K}_0 \otimes \boldsymbol{\mathcal{K}}_\mathrm{tt}\right)$. For optimizing $\boldsymbol{\mu}_0$, we consider terms of the log-probability that depend on $\boldsymbol{\mu}_0$,

$$\log\left[P(\boldsymbol{F}_\mathrm{t})\right] = -\frac{1}{2}\sum_{i,j=1}^{k}\sum_{m,n=1}^{T} X_{im}\left[\boldsymbol{K}_0^{-1}\right]_{ij}\left[\boldsymbol{\mathcal{K}}_\mathrm{tt}^{-1}\right]_{mn} X_{jn} + const., \tag{15}$$

$$\text{with } X_{im} = f_i(\boldsymbol{p}_m) - [\boldsymbol{\mu}_0]_i.$$

At the maximum, we have $\partial \log\left[P(\boldsymbol{F}_\mathrm{t})\right] / \partial[\boldsymbol{\mu}_0]_i = 0$. This yields the optimal value

$$[\boldsymbol{\mu}_0]_i = \frac{\sum_{m,n=1}^{T}\left[\boldsymbol{\mathcal{K}}_\mathrm{tt}^{-1}\right]_{mn} f_i(\boldsymbol{p}_n)}{\sum_{m,n=1}^{T}\left[\boldsymbol{\mathcal{K}}_\mathrm{tt}^{-1}\right]_{mn}}. \tag{16}$$

Likewise, the optimal choice of the $k \times k$ matrix $\boldsymbol{K}_0$ of the covariance hyperparameters can be obtained in closed form as [16]

$$[\boldsymbol{K}_0]_{ij} = \frac{1}{T}\boldsymbol{Y}_i^\top \boldsymbol{\mathcal{K}}_\mathrm{tt}^{-1}\boldsymbol{Y}_j \;\text{ with }\; \boldsymbol{Y}_i = \left[f_i(\boldsymbol{p}_1), \cdots, f_i(\boldsymbol{p}_T)\right]^\top - \boldsymbol{\mu}_0. \tag{17}$$

We note, that for a moderate number of outputs k, the overhead of a multi-output GP for $\boldsymbol{f}(\boldsymbol{p})$ over a single-output GP for $g(\boldsymbol{p})$ is not significant, as discussed by Plock *et al.* [12]. The reason is that the correlation $\kappa(\boldsymbol{p}, \boldsymbol{p}')$ between points $\boldsymbol{p}, \boldsymbol{p}' \in \mathcal{X}$ in the search space is assumed to be identical for all $k \times k$ outputs of the multi-output covariance function (see Eq. (12)). The additional computational overhead arises from the estimation of the posterior mean for each output defined in Eq. (13), due to the different data across the channels coming from components of the vector-valued observable $\boldsymbol{f}(\boldsymbol{p})$. The corresponding computational cost scales as $\mathcal{O}(k \cdot T^2)$ with the number of training points T compared to $\mathcal{O}(T^2)$ for a single-output GP. In some cases, the correlation lengths (i.e. the length scales defined in Eq. (3)) can differ between the outputs. Then, it can be preferable to split the multi-output GP into a set of independent single or multi-output GPs [17, 18].

B. Optimization policy and acquisition function

Generally, the random variable $\hat{g}(\boldsymbol{p}_\mathrm{q}) = \Phi\big(\hat{\boldsymbol{f}}(\boldsymbol{p}_\mathrm{q})\big)$ is non-Gaussian and has no closed form that would allow to evaluate the acquisition function in Eq. (9) analytically. In this

case, one can determine the expectation value by Monte-Carlo sampling. That is, one draws S i.i.d. samples (e. g. $S = 10^4$) from the distribution described by Eqs. (13) and (14) – or from the distributions of the set of independent single or multi-output GPs, and from the drawn samples $\left[\tilde{\boldsymbol{f}}_1(\boldsymbol{p}_\mathrm{q}), \ldots, \tilde{\boldsymbol{f}}_S(\boldsymbol{p}_\mathrm{q})\right]$ one computes the mapped samples $\left[\tilde{g}_1(\boldsymbol{p}_\mathrm{q}), \ldots, \tilde{g}_S(\boldsymbol{p}_\mathrm{q})\right] = \left[\Phi(\tilde{\boldsymbol{f}}_1(\boldsymbol{p}_\mathrm{q})), \ldots, \Phi(\tilde{\boldsymbol{f}}_S(\boldsymbol{p}_\mathrm{q}))\right]$. In most cases, the mapping Φ can be computed efficiently and can be even vectorized, such that it does not lead to a significant computational overhead. With these samples, one can approximate the expected improvement as

$$\alpha_\mathrm{EI}(\boldsymbol{p}_\mathrm{q}) \approx \tilde{\alpha}_\mathrm{EI}(\boldsymbol{p}_\mathrm{q}) = \frac{1}{S}\sum_{i=1}^{S} \min\left(0, g_\mathrm{min} - \tilde{g}_i(\boldsymbol{p}_\mathrm{q})\right). \tag{18}$$

In order to maximize the expected improvement, it is desirable that $\tilde{\alpha}_\mathrm{EI}(\boldsymbol{p}_\mathrm{q})$ is differentiable with respect to $\boldsymbol{p}_\mathrm{q}$. To this end, one can use the *reparametrization trick* [10, 19]. That is, one draws initially S i.i.d. samples $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_S \in \mathbb{R}^k$ from a standard normal distribution and for each query point $\boldsymbol{p}_\mathrm{q}$ one scales the samples to have the right mean $\boldsymbol{\mu}'$ Eq. (13) and covariance $\boldsymbol{K}'$ Eq. (14), i. e.

$$\tilde{\boldsymbol{f}}_i(\boldsymbol{p}_\mathrm{q}) = \boldsymbol{\mu}' + \boldsymbol{L}'\boldsymbol{h}_i, \tag{19}$$

where $\boldsymbol{L}'$ is the Cholesky decomposition of $\boldsymbol{K}'$. The principle of physics-informed BO is summarized in Fig. 2.

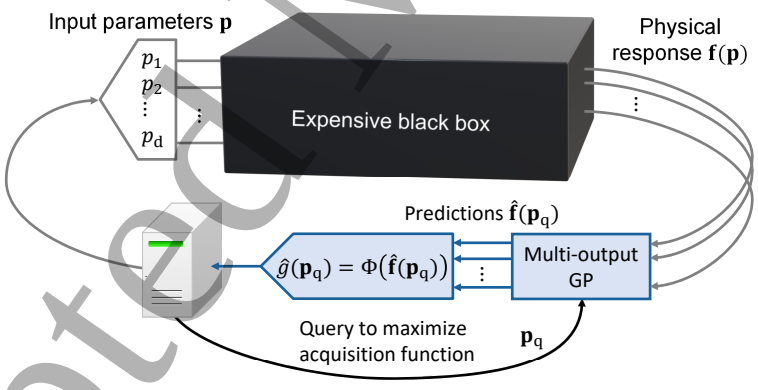

FIG. 2: Principle of physics-informed Bayesian optimization: A multi-output GP is trained on each physical response $[\boldsymbol{f}(\boldsymbol{p}_1), \cdots, \boldsymbol{f}(\boldsymbol{p}_T)]$ of an expensive black box. More generally, also multiple independent single or multi-output GPs can be used. The MVN prediction $\hat{\boldsymbol{f}}(\boldsymbol{p}_\mathrm{q})$ at a query point $\boldsymbol{p}_\mathrm{q}$ is mapped to a scalar posterior $\hat{g}(\boldsymbol{p}_\mathrm{q})$. This generally sample-based posterior is used to maximize a Monte-Carlo approximation of the acquisition function.

C. Illustrative examples

In order to illustrate the approach, we consider three nonlinear vector-valued synthetic examples consisting of three components ($k = 3$) defined in one-dimensional parameter space ($d = 1$). For the first example, we choose $\boldsymbol{f}(x) = [10x^2, \sin(4x), 1.05 + \sin(4x)]^\top$ with the goal to fit the objective $g(x) = 10x^2 + \frac{\sin(4x)}{1.05+\sin(4x)}$, constructed from the combination of its components, on the interval $-\pi \leq x \leq \pi$. This simple example is illustrative because the objective function $f(x)$ has four pronounced sharp minima in the search space, while all components of $\boldsymbol{f}(x)$ are relatively smooth. We define three different GP surrogates,

1. a single-output GP trained with values of the objective $g(x)$ with Gaussian posterior $\hat{g}_{\text{so}}(x)$,

2. a multi-output GP trained with values of the objective $\boldsymbol{f}(x)$ with mapped sample-based posterior $\hat{g}_{\text{mo}}(x)$,

3. a set of three independent single-output GPs trained on each component of $\boldsymbol{f}(x)$ with mapped sample-based posterior $\hat{g}_{\text{mgp}}(x)$.

The last approach determines an optimal length scale for each component individually, while the second approach uses only one length scale for all three components. While the first approach provides a Gaussian posterior defined by its mean and standard deviation, the properties of the other posteriors in terms of a median and uncertainty range between the 16% and 84% quantiles are determined based on 2000 samples.

As shown in Fig. 3, the synthetic example presents a challenging problem for the standard single-output GP with posterior $\hat{g}_{\text{so}}(x)$ (see Fig. 3 panel (a)). The reason is, that $g(x)$ has narrow minima resulting in a small optimal length scale. Hence, at many query points the correlation with the training points is assumed to be small leading to a large uncertainty and a tendency to predict values closer to the mean value. Very short characteristic length scales produce a posterior with low quality of interpolation through the data and high uncertainty, especially in the regions where the objective landscape is smooth (see Fig. 3 panel (a)). The relatively large uncertainty leads to a very explorative search, which deteriorates the convergence of the algorithm. On the contrary, the three-output GP with posterior $\hat{g}_{\text{mo}}(x)$ matches $g(x)$ much better (see Fig. 3 panel (b)). A very similar result is obtained by three independent single-output GPs with posterior $\hat{g}_{\text{mgp}}(x)$ (see Fig. 3 panel

(c)). The similarity between $\hat{g}_{\text{mo}}(x)$ and $\hat{g}_{\text{mgp}}(x)$ arises from the comparable length scales that characterize all the components of $\boldsymbol{f}(x)$. The second example highlights the advantage

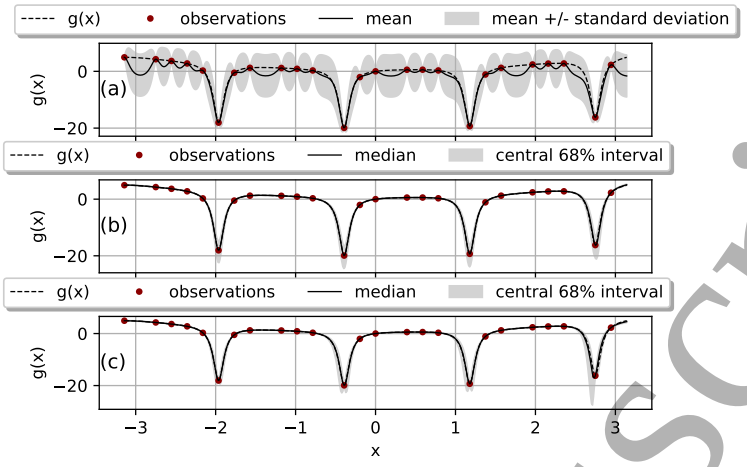

FIG. 3: Fitting the synthetic example defined as $g(x) = 10x^2 + \frac{\sin(4x)}{1.05+\sin(4x)}$ on the domain $-\pi \leq x \leq \pi$ using three different machine learning models on 25 points. Panels correspond to predictive distributions of: single-output GP with posterior $\hat{g}_{\text{so}}(x)$ (panel (a)), multi-output GP with posterior $\hat{g}_{\text{mo}}(x)$ (panel (b)), and three single-output GPs with posterior $\hat{g}_{\text{mgp}}(x)$ (panel (c)).

of incorporating additional information inherent to multi-output models. We consider the synthetic function $\boldsymbol{f}(x) = [3\sin(3x), x, 0.5x^2]^\top$ and aim to fit its squared 2-norm, i.e., $g(x) = \|\boldsymbol{f}(x)\|^2 = 9\sin^2(3x) + x^2 + 0.25x^4$ on the domain $-1 \leq x \leq 1$. The function is sampled at 9 points, deliberately excluding those near the global minimum to assess the impact of missing critical information on the surrogate models' accuracy. Here, all the components of $\boldsymbol{f}(x)$ as well as the objective $g(x)$ are smooth functions in the search space that can be easily interpolated by GPs. Indeed, as we can observe from Fig. 4, the surrogates $\hat{g}_{\text{so}}(x)$, $\hat{g}_{\text{mo}}(x)$ and $\hat{g}_{\text{mgp}}(x)$ appear similar globally, but near the minimum, the models trained on the components of $\boldsymbol{f}(x)$ provide a more accurate estimation of the true objective due to the additional information. Specifically, accurate linear models of the components around the minimum lead to an accurate quadratic model of $g(x)$. Especially in higher-dimensional search spaces where the sample points are scarce, an accurate quadratic model facilitates a quicker convergence to the local minimum as will be demonstrated by numerical examples in section IV. The final synthetic function provides a counter-example where models trained on the components of $\boldsymbol{f}(x)$ provide a less accurate estimation of the objective compared to a model trained on the objective directly. We fit the squared 2-norm of a vector-valued

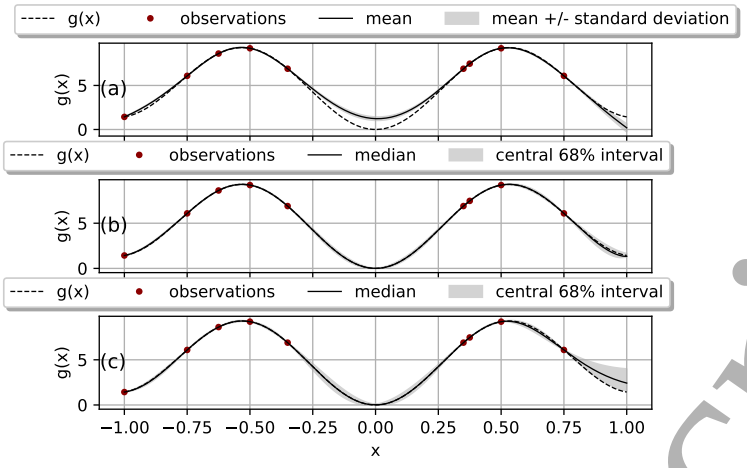

FIG. 4: Fitting the synthetic example defined as $g(x) = 9\sin^2(3x) + x^2 + 0.25x^4$ on the domain $-1 \leq x \leq 1$ using three different GP models on 9 points. Panels correspond to predictive distributions of: single-output GP with posterior $\hat{g}_{\text{so}}(x)$ (panel (a)), multi-output GP with posterior $\hat{g}_{\text{mo}}(x)$ (panel (b)), and three single-output GPs with posterior $\hat{g}_{\text{mgp}}(x)$ (panel (c)).

function $\boldsymbol{f}(x) = [\sin(4x), \cos(4x), x]^\top$ over the domain $-\pi \leq x \leq \pi$ using 12 points. This scenario is unfavorable for multi-output methods due to the oscillatory nature of the two harmonic components of $\boldsymbol{f}(x)$ which are generally more challenging to fit than the final quadratic objective $g(x) = \|\boldsymbol{f}(x)\|^2 = \sin^2(4x) + \cos^2(4x) + x^2 = 1 + x^2$ that does not oscillate at all. We note that this behavior can be common in practice. For example, in photonics the solution of the time-harmonic Maxwell's equations at some observed position contains phase information of the scattered field response $\boldsymbol{E}_p = \|\boldsymbol{E}_p\| e^{i\varphi_p} \hat{\boldsymbol{e}}_p$, $p = x, y, z$. The phases φ_p depend sensitively on the shape of the scatterer. However, the objective is often a function of the intensity of the scattered field $I \propto \|\boldsymbol{E}\|^2$, which is insensitive to the phase information. As evidenced in Fig. 5, the GPs trained on the components of $\boldsymbol{f}(x)$ provide less accurate surrogate models than the GP directly trained on $g(x)$. This effect is particularly pronounced for the posterior $\hat{g}_{\text{mo}}(x)$ that uses a single length scale for all components. If the length scale for the third component is allowed to be larger, the general trend of the objective can be better represented.

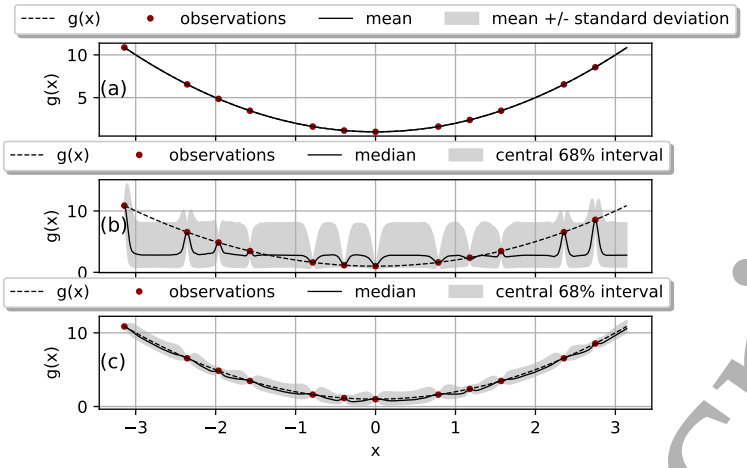

FIG. 5: Fitting the synthetic example defined as $g(x) = \sin^2(4x) + \cos^2(4x) + x^2$ on the domain $-\pi \leq x \leq \pi$ using three different GP models on 12 points. Panels correspond to predictive distributions of: single-output GP with posterior $\hat{g}_{\mathrm{so}}(x)$ (panel (a)), multi-output GP with posterior $\hat{g}_{\mathrm{mo}}(x)$ (panel (b)), and three single-output GPs with posterior $\hat{g}_{\mathrm{mgp}}(x)$ (panel (c)).

IV. NUMERICAL EXPERIMENTS

In this section, we assess the performance of physics-informed BO for three optimization problems from distinct areas of engineering and physics. We benchmark physics-informed BO against classic single-output BO and three heuristic optimization methods, differential evolution (DE), covariance matrix adaptation evolution strategy (CMA-ES) [20], and downhill simplex (Nelder-Mead) search [21]. All methods are implemented in *JCMwave*'s optimization suite *JCMoptimizer* [22] based mainly on the Python packages `scipy.optimize` [23] and `PyTorch`. Each optimization algorithm is executed 10 times from different initial random conditions in order to account for the variability in performance due to different initial samples and the stochastic components in the optimization process. The effectiveness of heuristic algorithms is heavily influenced by a number of hyperparameters which can be adjusted to best fit the specific optimization problem at hand. This process is time-consuming, so we leave them at default values for all the benchmarks.

A. Inverse design of a beam-splitter

Sub-wavelength metasurfaces have received significant attention due to their compactness and flexibility in controlling the phase of the light field [24]. Due to the optical coupling of *meta-atoms* – the metasurface building blocks – optimal designs beyond a library approach are commonly found by their iterative optimization assisted by numerical simulations. As a representative example, we consider a metasurface that acts as a large-angle beam splitter [25] with the aim to scatter light with equal power flux into all diffraction orders.

The considered device consists of periodically aligned supercells constructed from several meta-atoms of different sizes, see Fig. 6. For this study, we focus on the 1D design of the beam-splitter as outlined in [25], assuming geometrical invariance in one direction, allowing for more elaborate benchmark results due to the fast forward problem evaluation. The metasurface is illuminated with a x-polarized plane wave at a wavelength of 1.55 µm propagating along the y-axis, indicated by the wave vector $\boldsymbol{k} = \|\boldsymbol{k}\|\hat{\boldsymbol{a}}_y$ (Fig. 6). The period of the supercell is 7.15 µm allowing for nine diffraction orders $(0, \pm1, \pm2, \pm3, \pm4)$ covering the angular range of 120°. We solve for the electromagnetic fields arising from the interaction of the impinging plane wave and the metasurface by discretizing Maxwell's equations in the frequency domain with higher-order finite element method implemented in the software package *JCMsuite* [26]. We restrict the numerical analysis to one supercell and account for metasurface periodicity by assigning periodic boundary conditions in the x-direction. The infinite domain is truncated by applying perfectly matched layers in the y-direction, and we further reduce the computational domain by half by imposing a mirror-symmetry of the supercell geometry. In the post-processing step, the computed electromagnetic field in the homogeneous half-space above the metasurface is decomposed into a discrete spectrum of diffraction order plane waves via Fourier transform. The positive and negative diffraction orders are identical because of the symmetry present in the supercell geometry, resulting in total of five independent diffracted plane waves.

The design objective in the six-dimensional parameter space $\mathcal{X}$ of meta-atom widths $\boldsymbol{p} = (d_1, \cdots, d_6)^\top \in \mathcal{X} = [200.0\,\text{nm}, 600.0\,\text{nm}]^6$ (see Fig. 6) is to minimize the variance of the power fluxes of the 9 diffracted plane waves

$$g(\boldsymbol{p}) = \frac{1}{9}\sum_{k=1}^{9} f_i^2(\boldsymbol{p}) - \left(\frac{1}{9}\sum_{k=1}^{9} f_i(\boldsymbol{p})\right)^2, \tag{20}$$

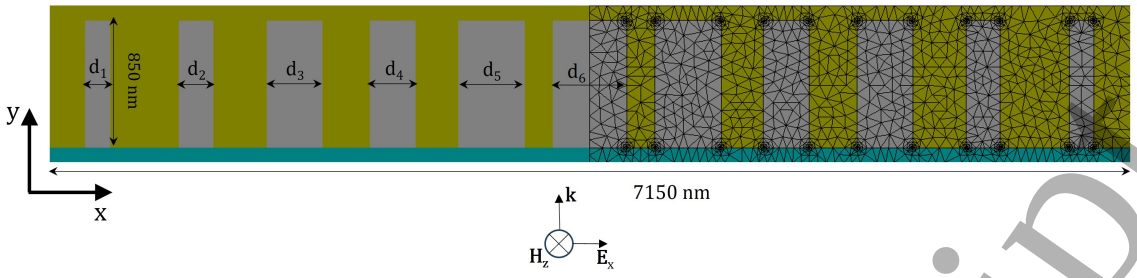

FIG. 6: The beam-splitter supercell setup with a finite element triangulation of half of the mirror-symmetric geometry. Eleven meta-atoms of a height of 850 nm are placed at constant pitch of 7.15 µm. Six independent widths $d_1, ..., d_6$ are variable and subject to optimization.

where $f_i(\boldsymbol{p})$ is the power flux of the ith diffracted plane wave normalized to the power flux of the incoming plane wave. This objective ensures uniformity in the perceived brightness of the projected dots. In order to avoid a solution of very low power flux in all diffraction orders, we set the lower-bound constraint on the total relative transmitted flux as

$$h(\boldsymbol{p}) = \sum_{k=1}^{9} f_i(\boldsymbol{p}) \geq 0.5. \tag{21}$$

To perform a constrained Bayesian optimization, we use *expected constrained improvement*

$$\alpha_{\mathrm{ECI}}(\boldsymbol{p}_{\mathrm{q}}) = \alpha_{\mathrm{EI}}(\boldsymbol{p}_{\mathrm{q}}) \cdot P(\hat{h}(\boldsymbol{p}_{\mathrm{q}}) \geq 0.5) \tag{22}$$

as the acquisition function [27]. For the single-output BO method, both the value of the objective $g(\boldsymbol{p})$ and the constrained total relative flux $h(\boldsymbol{p})$ are learned with two single-output GPs. For the physics-informed BO, a multi-output Gaussian process is used to learn the vector of power fluxes $\boldsymbol{f}(\boldsymbol{p}) = [f_1(\boldsymbol{p}), \cdots, f_9(\boldsymbol{p})]^\top$ and to make predictions about $g(\boldsymbol{p})$ while a single-output GP is employed to learn the constrained relative flux $h(\boldsymbol{p})$.

For the heuristic global optimization methods CMA-ES, differential evolution, and downhill simplex, the constraint has to be encoded into a single scalar objective. Here, we minimize

$$g^*(\boldsymbol{p}) = \Theta\left[h(\boldsymbol{p}) - 0.5\right] \cdot \left[g(\boldsymbol{p}) - g_{\max}\right], \tag{23}$$

where $\Theta[\cdot]$ is the Heaviside step function and $g_{\max} = 0.987$ is an upper bound of the variance of the relative fluxes. For $h(\boldsymbol{p}) < 0.5$ and $g(\boldsymbol{p}) = g_{\max}$, the objective is zero. Otherwise it is negative and takes the minimum value $-g_{\max}$ for a vanishing variance.

In order to asses the average performance of each optimization algorithm, the search process is executed ten times per method with each run starting from a different initial set of points in the parameter space, and continuing for 800 iterations. Fig. 7 shows the cumulative minimum of observed values of $g(\boldsymbol{p})$ that are feasible, i. e. where $h(\boldsymbol{p}) > 0.5$. Clearly, physics-informed BO outperforms all other methods by a large margin. For example, after 100 iterations it finds a feasible beam-splitter configuration with a smaller diffraction order power flux variance than the other approaches after 800 iterations. The additional information about all fluxes enables a much faster convergence to very small values. The second best method is the single-output BO leading to a final flux variance half as large as for the other methods. All heuristic methods perform quite similar with the downhill-simplex method having a slightly worse performance than differential evolution and CMA-ES. To

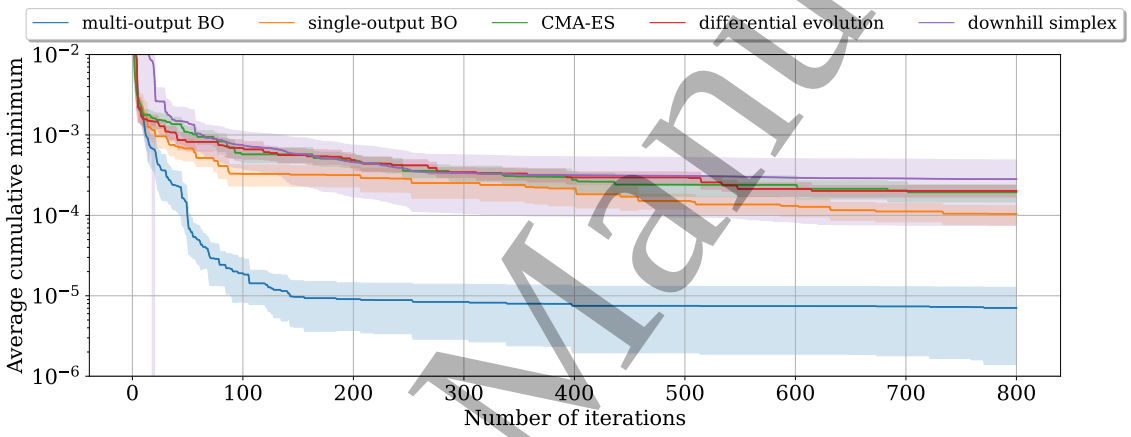

FIG. 7: Benchmark of different optimization approaches on a beam-splitter example. The bold lines represent the mean feasible cumulative minimum of observed variance values defined in Eq. (20) averaged over 10 runs. The shaded areas indicate the respective standard deviations between the runs.

heuristically demonstrate why physics-informed BO outperforms the classical single-output counterpart in this particular example, we conduct an additional numerical study, the results of which are presented in the Appendix, subsection A 1.

B. Transport optimization of quantum gases

Precise state engineering of quantum gases is crucial for advancing quantum technologies including next generation quantum sensors, the preparation of entangled states, and quantum computing. The fast and controlled transport of Bose-Einstein condensates (BECs) has been designed using shortcut-to-adiabaticity (STA) [28, 29] and optimal control theory (OCT) [30, 31] protocols. Both approaches face significant challenges due to the simplifications they require in modeling BEC dynamics. STA primarily focuses on the classical equations of motion of the condensate's center of mass, excluding the quantum nature of the system. On the other hand, OCT is limited by its slow convergence, often requiring millions of executions of the underlying physics simulations to optimize the time-dependent magnetic field ramp $B(t)$ that drives the transport process [30]. As a result, OCT relies on computationally efficient forward problem approximations, such as simplifying trapping potentials to harmonic contributions and describing quantum dynamics within the Thomas-Fermi regime [32].

Initially, in this example case, the BEC resides at the minimum of the magnetic field trap potential at the coordinates $x_0 = y_0 = 0, z_0(t = 0) = 0.45\,\mathrm{mm}$. The goal is to smoothly displace the BEC along the z direction by manipulating the parameters of the trap potential via a carefully designed magnetic field ramp $B(t)$ ending up at the position $z_0(t = t_\mathrm{f}) = 1.65\,\mathrm{mm}$ at the time t_f. To allow for benchmarks with moderate computational effort, we simulate the BEC transport dynamics within the Thomas–Fermi approximation [17]. The BEC dynamics is described with a classical Newtonian equation governing its center of mass z_A motion

$$\ddot{z}_A(t) = -\omega_z^2(t)(z_A(t) - z_0(t)), \tag{24}$$

and the shape dynamics associated with the BEC wave function is described as

$$\ddot{\lambda}_i(t) = \frac{\omega_i^2(0)}{\lambda_i(t)\lambda_x(t)\lambda_y(t)\lambda_z(t)} - \omega_i^2(t)\lambda_i(t), \tag{25}$$

where $\omega_i(t), i = x, y, z$ are the angular frequencies of the confining harmonic potential, and $\lambda_i(t)$ are scaling factors influencing the time evolution of the Thomas-Fermi radii $r_x(t), r_y(t)$ and $r_z(t)$ which define the spatial extent of the BEC paraboloid wave function within the Thomas-Fermi approximation [17, 30]. The position $z_0(t)$ and the frequencies $\omega_i(t)$ of the magnetic field trap are controlled with the driving magnetic field ramp $B(t)$. The differential

equations (24) and (25) can be solved using standard numerical tools such as the Runge-Kutta algorithm. Once the solutions are obtained, they allow for the calculation of the system's *classical* energy

$$E_{\mathrm{cl}}(\boldsymbol{p},t) = \frac{m}{2}\left(\omega_z^2\left[z_A - z_0\right]^2 + \left[\dot{z}_A - \dot{z}_0\right]^2\right) \tag{26}$$

and *quantum* energy

$$E_{\mathrm{qu}}(\boldsymbol{p},t) = \sum_{i=x,y,z}\frac{m}{14}\left(\omega_i^2 r_i^2 + \dot{r}_i^2\right) + \frac{15gN}{28\pi\, r_x r_y r_z}. \tag{27}$$

The average *classical* energy over the transport duration

$$E_{\mathrm{cl}}^{\mathrm{int}}(\boldsymbol{p}) = \frac{1}{t_{\mathrm{f}}}\int_0^{t_{\mathrm{f}}} E_{\mathrm{cl}}(\boldsymbol{p},t)\mathrm{d}t \tag{28}$$

quantifies how far the BEC dynamics deviates from the trap minimum. The larger this measure is, the more anharmonic regions the BEC explores in a realistic trap, and the less accurate the considered approximations become.

Good candidates for the magnetic field ramp are identified at the minimum of the *total* energy cost which is assembled as a weighted sum of the individual energies as [17]:

$$E_{\mathrm{obj}}(\boldsymbol{p}) = \xi_{\mathrm{cl}}E_{\mathrm{cl}}(\boldsymbol{p},t_{\mathrm{f}}) + \xi_{\mathrm{qu}}E_{\mathrm{qu}}(\boldsymbol{p},t_{\mathrm{f}}) + \xi_{\mathrm{cl}}^{\mathrm{int}}E_{\mathrm{cl}}^{\mathrm{int}}(\boldsymbol{p}). \tag{29}$$

Here, $\xi_{\mathrm{cl}}, \xi_{\mathrm{qu}}, \xi_{\mathrm{cl}}^{\mathrm{int}}$ are weighting factors which are selected appropriately to balance the contributions of the respective energies during the optimization process. In order to enforce an optimization to the lower bound of the objective energy $E_{\mathrm{obj}}^{\mathrm{min}}$ dictated by the quantum ground state energy of the trapped BEC, the objective of the optimization is chosen to be

$$g(\boldsymbol{p}) = \mathrm{arsinh}\left(E_{\mathrm{obj}}(\boldsymbol{p}) - E_{\mathrm{obj}}^{\mathrm{min}}\right), \tag{30}$$

where the area hyperbolic sine $\mathrm{arsinh}(\cdot)$ is a bijective transformation with the largest derivative at $E_{\mathrm{obj}}(\boldsymbol{p}) = E_{\mathrm{obj}}^{\mathrm{min}}$. We parametrize the driving magnetic field ramp $B(t)$ with B-splines and 15 degrees of freedom collected in the parameter vector $\boldsymbol{p} \in \mathcal{X} \subset \mathbb{R}^{15}$ [17]. Single-output BO uses a scalar GP to learn directly the objective shown in Eq. (30). On the other hand, the proposed physics-informed BO scheme learns the BEC's dynamical properties at time t_{f}: the center of mass position and velocity, $z_A(t_{\mathrm{f}})$ and $\dot{z}_A(t_{\mathrm{f}})$, as well as the Thomas-Fermi radii $r_x(t_{\mathrm{f}}), r_y(t_{\mathrm{f}}), r_z(t_{\mathrm{f}})$ and their time derivatives $\dot{r}_x(t_{\mathrm{f}}), \dot{r}_y(t_{\mathrm{f}}), \dot{r}_z(t_{\mathrm{f}})$. Finally, the integrated

classical energy $E_{\mathrm{cl}}^{\mathrm{int}}$ is learned. More details of the physics-informed approach are described in [17].

Similarly to the beam-splitter example, we perform a benchmark comparing the convergence of the physics-informed BO to the single-output BO and three heuristic optimization methods for 500 iterations. As shown in Fig. 8, physics-informed BO provides superior convergence compared to other optimization algorithms considered, identifying a magnetic field ramp $B(t)$ in approximately 200 iterations resulting in almost negligible final excess energy. Again, single-output BO is the second best method. Among the heuristic optimization methods, downhill simplex converges significantly better after 200 iterations than CMA-ES and differential evolution.

The fast convergence behavior of physics-informed BO is crucial for the transport optimization of realistic gas systems where the Thomas-Fermi approximation does not hold and numerically intensive, time consuming solutions of the Gross-Pitaevskii equation cannot be avoided.

In subsection A 2 of the Appendix we take a closer look at the predictive capabilities of the single-output and multi-output surrogate models in the vicinity of the minimum of the energy objective $g(\boldsymbol{p})$ (30) and provide possible explanations for the observed superior convergence trends of physics-informed BO.

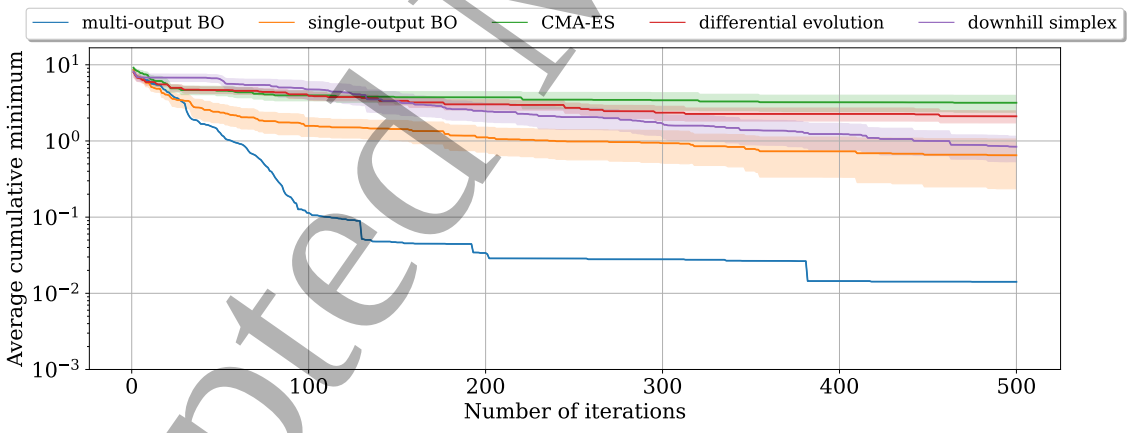

FIG. 8: Benchmark of different optimization algorithms on a BEC transport problem. The bold lines represent the mean convergence of the objective $g(\boldsymbol{p})$ of Eq. (30). The trends are averaged over 10 runs, and the shaded areas indicate respective standard deviations.

C. Optimizing aesthetic appearance of perovskite solar cells

Color aesthetics in photovoltaic modules are essential for building integrated photovoltaics (BIPV) [33]. Multilayer interference structures, such as color filters, can modify the appearance of silicon solar cells by reflecting a narrow part of the visible spectrum [34]. Compared to competing methods, e.g. nanophotonic devices or printed inks, interference structures result in lower efficiency losses, while offering vibrant colors [35]. Emerging perovskite solar cells present an aesthetic challenge with their characteristic gray-purple-brown hues. This optimization targets the aesthetic enhancement of perovskite solar cells by incorporating a bio-inspired MorphoColor filter [34].

The solar cell design is shown schematically in Fig. 9. From top to bottom, it consists of conformal layers of the color filter on structured cover glass, followed by the various layers of a perovskite solar cell, and an absorber layer. The color filter for this optimization study consists of ten layers of alternating high and low refractive index materials. Each layer height can be varied in the range from 0 to 300 nm resulting in the design space $\mathcal{X} = [0, 300\,\mathrm{nm}]^{10}$. The simulation environment uses a net-radiation method and ray-tracing to simulate the reflectance for the 3D-textured multilayer thin-film stack. The model and simulation setup are described in more detail in [18]. The objective is to minimize the cosine-weighted and angle-averaged perception distance ΔE_{00} [18]:

$$g(\boldsymbol{p}) = \sum_{i=1}^{18} \cos(\theta_i) \cdot \Delta E_{00,i}(\boldsymbol{p}), \tag{31}$$

where $\theta_i = 0°, 5°, 10°, 15°, \ldots, 85°$ are the angles of incidence and $\Delta E_{00,i}(\boldsymbol{p})$ is a measure of the color distance of two 3D-sets of color coordinates representing the simulated and the targeted color for angle θ_i. The objective depends on a scaled differences between lightness (ΔL), chroma (ΔC) and hue (ΔH) for the two colors in CIE *Lab* space as:

$$\Delta E^2_{00,i}(\boldsymbol{p}) = \left(\frac{\Delta L_i(\boldsymbol{p})}{S_{\mathrm{L}i}(\boldsymbol{p})}\right)^2 + \left(\frac{\Delta C_i(\boldsymbol{p})}{S_{\mathrm{C}i}(\boldsymbol{p})}\right)^2 + \left(\frac{\Delta H_i(\boldsymbol{p})}{S_{\mathrm{H}i}(\boldsymbol{p})}\right)^2 + R_{\mathrm{T}i}(\boldsymbol{p})\frac{\Delta C_i(\boldsymbol{p})}{S_{\mathrm{C}i}(\boldsymbol{p})}\frac{\Delta H_i(\boldsymbol{p})}{S_{\mathrm{H}i}(\boldsymbol{p})}, \tag{32}$$

where $S_{\mathrm{L}i}$, $S_{\mathrm{C}i}$, and $S_{\mathrm{H}i}$ are correct non-uniformities in CIE *Lab* and $R_{\mathrm{T}i}$ mainly corrects the hue and chroma values in the blue region.

Single-output BO uses a scalar GP to learn the objective value $g(\boldsymbol{p})$ in Eq. (31). For physics-informed BO we use 18 independent multi-output GPs — one for each angle of incidence. Each multi-output GP learns the behavior of the seven $\boldsymbol{p}$-dependent terms appearing in Eq. (32).

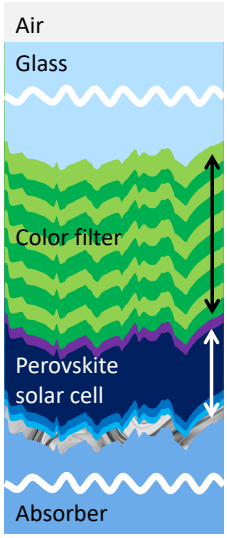

FIG. 9: Schematics of the multilayer thin-film stack of a textured perovskite solar cell with color filter. After an (infinite) air domain and a glass layer, a ten layer color filter (dark and light green) and the layers of a perovskite solar cell follow. The stack is terminated by an (infinite) absorbing layer. Ten independent height parameters of the color filter are the subject of optimization, ranging from 0 to 300 nm.

We perform again a benchmark comparing the convergence of the physics-informed BO to the single-output BO and three heuristic optimization methods for 300 iterations. As shown in Fig. 10, physics-informed BO shows a better convergence to small objective values than single-output BO, CMA-ES, and differential evolution. Within the 300 iterations, only the downhill simplex reaches the comparable minimum. For this example, one observes a large variance between the ten optimization runs. We attribute this to the existence of many rather narrow local minima in the ten-dimensional search space. This can be also observed in a numerical study of the prediction accuracy in the vicinity of the minimum (see Appendix, subsection A 3).

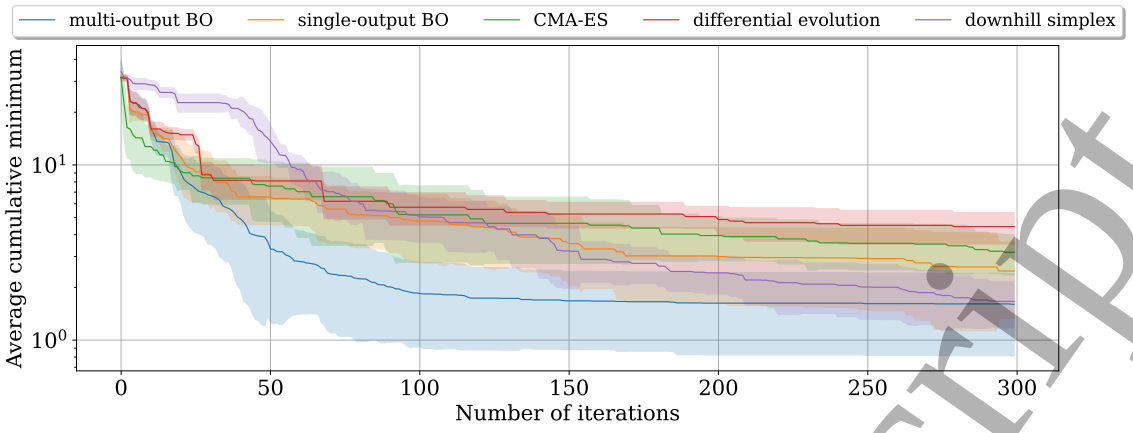

FIG. 10: Benchmark of different optimization algorithms on a color optimization task. The bold lines represent the mean convergence trends averaged over 10 runs, and the shaded areas indicate respective standard deviations.

V. DISCUSSION AND CONCLUSION

In this manuscript, we introduced physics-informed Bayesian optimization based on multi-output Gaussian processes that learn vector-valued functions $\boldsymbol{f}(\boldsymbol{p})$ representing the physical response of a system. This approach allows to gain deeper insight into the behavior of a scalar objective function $g(\boldsymbol{p})$, which results from a generally nonlinear combination of the components of $\boldsymbol{f}(\boldsymbol{p})$. Across three different optimization problems from distinct areas of engineering, we demonstrated that physics-informed BO consistently outperforms the classical single-output BO and standard heuristic approaches reaching better designs in fewer iterations. We attribute this improved performance to the additional information encoded in the physical response of the system, which allows the multi-output surrogate model to approximate the scalar objective landscape more accurately and, consequently, to identify promising regions of the parameter space more effectively. This is supported by extended numerical studies revealing that in all three benchmark experiments the physics-informed approach yields a higher-order approximation of the objective close to the minimum, which is significantly more accurate than the single-channel model (see Appendix A).

There are situations where physics-informed BO may be counterproductive. One illustrative example was provided in subsection III C showing that the approach should not be favored when the learned quantities are highly sensitive to the design parameters, whereas the objective itself is less sensitive. In photonics this is the case for phase information of

the scattered field, which is very sensitive to small changes of the shape of a scatterer. Meanwhile, the objective often only depends on field intensities that are much less sensitive. Another example would be an objective that depends on an averaged temporal response. In this case it is not advantageous to learn the full time-dependent physical behavior since the average is typically a smoother function of the design parameters than each individual time sample, such that it is easier to learn with a single-output GP instead. Moreover, since averaging is a linear transformation, no higher-order information from the multi-output model is obtained. Finally, we note that the estimation of the objective's probability distribution by the prior covariance defined in Eq. (12) neglects local correlations of the outputs. As observed previously in [12] this can lead to a model's underestimation of the probability to observe small objective values if the number of outputs k is large. For large k it could be helpful to transform the output to a lower-dimensional representation. However, this is beyond the scope of this work.

For many practical applications, these limitations are not relevant. Indeed, many objectives describe the performance of a system in terms of some quadratic form of a moderate number k of observables. For example, inverse design tasks in nanophotonics are often formulated as a minimization of the L_2 norm between a target vector $\boldsymbol{f}'$, representing the desired device functionality, and the candidate vector $\boldsymbol{f}(\boldsymbol{p})$, obtained by numerical simulation. Likewise, objectives that contain potential and kinetic energies depend on terms quadratic in observables such as position coordinates and respective velocities (e.g. Eqs. (26), (27)). Our analysis shows that in such cases learning the final scalar objective with a single-output GP inherently leads to a loss of information such that physics-informed BO shows a much faster convergence than classical BO. Therefore, we assume that physics-informed BO will be beneficial in many application fields.

Data availability

The Python scripts and the *JCMsuite* files related to the beam-splitter benchmark example are publicly available in the associated open access data publication [36].

Acknowledgement

This project is funded by the Federal Ministry for Economic Affairs and Energy (BMWi, project number 50WM2253, AI-Quadrat), by the German Federal Ministry of Research, Technology and Space (BMFTR, project number 01IS24005, NanoMaC, and Forschungscampus MODAL, project number 05M20ZBM), and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

Appendix A: Numerical investigations into the convergence of physics-informed Bayesian optimization

For all three discussed benchmarks, physics-informed BO exhibits a better optimization convergence in comparison to single-output BO and global heuristic optimization methods. This means, that the multi-output GPs are able to guide the optimization more effectively. In subsection III C we discussed three mechanisms by which a multi-output GP can provide better or worse predictions for the objective than a single-output GP:

1. The learned outputs $\boldsymbol{f}(\boldsymbol{p}) = [f_1(\boldsymbol{p}), \cdots, f_k(\boldsymbol{p})]^\top$ can have a smoother behavior than the scalar objective $g(\boldsymbol{p}) = \Phi(\boldsymbol{f}(\boldsymbol{p}))$. Due to the larger length scales, the multi-output GP can provide a more accurate interpolation with smaller uncertainty between the training points.

2. The mapping $\Phi : \boldsymbol{f}(\boldsymbol{p}) \mapsto g(\boldsymbol{p})$ can contain quadratic or other higher-order dependencies on the outputs $f_i(\boldsymbol{p})$. Around a local minimum, the multi-output GP can thus provide a higher-order interpolation than the single-output GP with a more exact prediction of the position and depth of the minimum.

3. Conversely to case one, the outputs $f_i(\boldsymbol{p})$ can also exhibit a more oscillatory behavior than the objective $g(\boldsymbol{p})$. It is generally not advisable to learn quantities whose oscillation is not relevant for the objective.

In the following, we will show that especially the second mechanism is relevant for the three benchmark examples. All optimization runs in our benchmarks start with an initial set of randomly distributed training points. Once this random phase is over, Bayesian optimizers

iteratively compute the acquisition function, selecting the point of its maximum as the next training point for evaluation. At this early stage, when training data is extremely scarce, the additional information coming from different channels in multi-output BO proves especially valuable. To validate this claim, we artificially recreate this scenario by drawing a small set of training points around the known objective minimum and fit the single-output and multi-output GP on them. This mimics the optimization phase where both single-output BO and physics-informed BO have sampled around the minimum but have not yet converged.

After training, we generate a set of N random query points $\boldsymbol{p}_{\mathrm{q}1}, \cdots, \boldsymbol{p}_{\mathrm{q}N}$ in the same region and draw for each point S samples $\tilde{g}_{i1}, \cdots, \tilde{g}_{iS}$ from the posterior of the single-output GP $\hat{g}_{\mathrm{so}}(\boldsymbol{p}_{\mathrm{q}i})$ or the posterior of the multi-output GP $\hat{g}_{\mathrm{mo}}(\boldsymbol{p}_{\mathrm{q}i}) = \Phi\big(\hat{\boldsymbol{f}}(\boldsymbol{p}_{\mathrm{q}i})\big)$. We evaluate $N \cdot S$ values for the relative absolute error (RAE) of the predictions of both surrogates as:

$$\varepsilon_{ij} = \left| \frac{\tilde{g}_{ij} - g(\boldsymbol{p}_{\mathrm{q}i})}{g(\boldsymbol{p}_{\mathrm{q}i})} \right|, 1 \leq i \leq N, 1 \leq j \leq S. \tag{A1}$$

1. Inverse design of a beam-splitter

For the case of a beam splitter, we perform an analysis around one of the identified local minima in the six-dimensional search space of meta-atom widths. We select a hypercube with side length 10.0 nm centered at the minimum, and generate 12 training points using a quasi-random Sobol' sequence. On these points, we fit a single-output GP on the variance objective values $g(\boldsymbol{p})$ defined in Eq. (20) and a multi-output GP on the power flux vector $\boldsymbol{f}(\boldsymbol{p}) = [f_1(\boldsymbol{p}), \cdots, f_9(\boldsymbol{p})]^\top$. Next, we sample $N = 200$ query points from a uniform distribution within the considered search hyperspace, run finite element simulations to determine the variance objective at these points, and evaluate the RAE defined in Eq. (A1) of the multi-output and the single-output model predictions based on $S = 2500$ samples. According to the Fig. A1, the RAE of the scalar GP is significantly larger than the one obtained by the multi-output GP indicating better fidelity of the posterior $\hat{g}(\boldsymbol{p}_{\mathrm{q}i}) = \Phi\big(\hat{\boldsymbol{f}}(\boldsymbol{p}_{\mathrm{q}i})\big)$.

To further illustrate this, we examine the behavior of the surrogates by varying the parameter d_4 along one direction of the hypercube while keeping the others fixed at the central minimum position. Comparing panels (a) and (b) in Fig. A2, we observe that multi-output posterior $\hat{g}_{\mathrm{mo}}(\boldsymbol{p}_{\mathrm{q}}) = \Phi\big(\hat{\boldsymbol{f}}(\boldsymbol{p}_{\mathrm{q}})\big)$ aligns more closely with the ground truth variance objective than the single-output posterior $\hat{g}_{\mathrm{so}}(\boldsymbol{p}_{\mathrm{q}})$. This improved fidelity around the minimum enables a BO sampling which is more targeted towards the true minimum leading to a faster

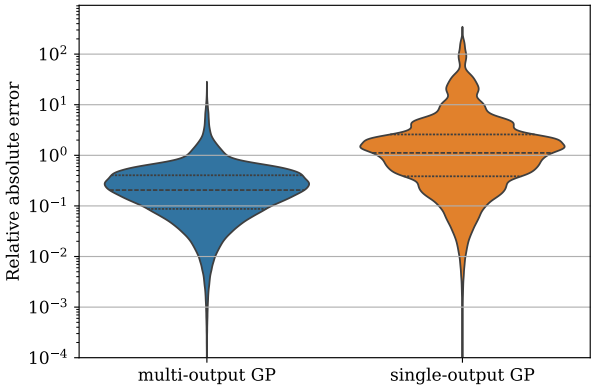

FIG. A1: Relative absolute error distribution of the single-output and the multi-output GP predictions computed on $N = 200$ testing points in the hyperspace around variance objective $v(\boldsymbol{p})$ minimum. The dashed line marks the median and the dotted lines the lower 25% and upper 75% quantiles of the distribution.

convergence, as observed in Fig. 7. It is also noteworthy that, as shown in Fig. A2 (b), the single-output posterior $\hat{g}_{\mathrm{so}}(\boldsymbol{p})$ predicts *negative* values of the objective. This is physically incorrect, since equation (20) ensures that the variance objective is strictly positive.

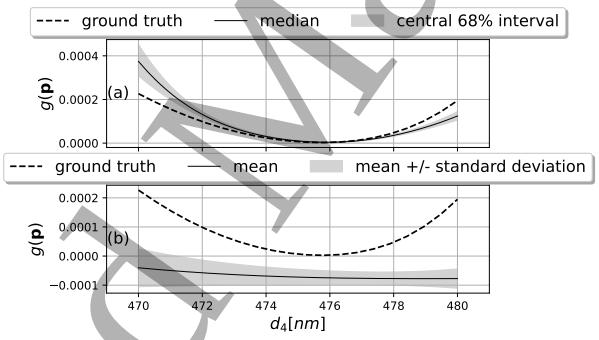

FIG. A2: One-dimensional cross section along the parameter d_4 through a minimum of variance objective (20). Panels (a) and (b) illustrate the behavior of $\hat{g}_{\mathrm{mo}}(\boldsymbol{p})$ and $\hat{g}_{\mathrm{so}}(\boldsymbol{p})$, respectively.

2. Transport optimization of quantum gasses

In analogy to the beam-splitter example, we perform an analysis of the prediction accuracy around the minimum of the objective shown in Eq. (30). A hypercube is constructed in the search space of the B-spline coefficients, centered at the minimum and with side length

of 0.04. In this region we select 16 points sampled from a Sobol' sequence to train a single-output GP on the objective values $g(\boldsymbol{p})$ defined in Eq. (30), and a multi-output GP on the 9 input variables of the objective collected in a vector $\boldsymbol{f}(\boldsymbol{p})$. The surrogates are tested on $N = 200$ uniformly distributed points in the hypercube. The distribution of the relative absolute error RAE of Eq. (A1) is shown in Fig. A3. Again the predictions based on the multi-output GP have generally smaller RAE values.

Fig. A4 shows a one-dimensional slice through the hypercube obtained by varying the control point c_7 while keeping other 14 parameters fixed at the minimum. Very clearly, the multi-output GP provides a good higher order approximation of the minimum, while the predictions of the single-output GP depend only linearly on the control point c_7. This highlights the importance of underlying physical information in such a comparatively high-dimensional search space with $d = 15$ parameters. Here, all $d + 1 = 16$ samples from the training set are required alone for estimating first-order derivatives, while additional $(d^2 + 2)/2 = 120$ training samples would be required to estimate second-order derivatives based on observations of $g(\boldsymbol{p})$ alone. Since the estimated minimum position of the single-output GP largely deviates from the true minimum, the single-output BO is much less sample efficient than physics-informed BO in this case. (see Fig. 8).

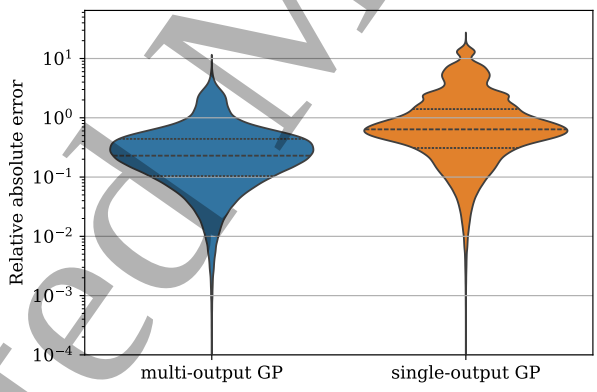

FIG. A3: Relative absolute error distribution of the single-output and the multi-output surrogate model predictions computed on 200 query points in the hyperspace around the energy objective minimum.

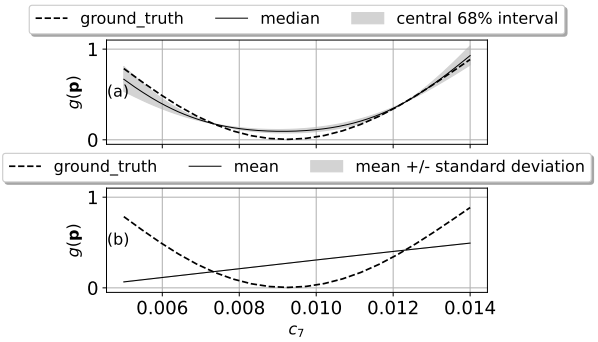

FIG. A4: One-dimensional cross section along the B-spline control point c_7 through the energy objective minimum. Panels (a) and (b) illustrate the behavior of $\hat{g}_{\mathrm{mo}}(\boldsymbol{p})$ and $\hat{g}_{\mathrm{so}}(\boldsymbol{p})$, respectively.

3. Optimizing aesthetic appearance of perovskite solar cells

Similar to the previous two numerical exercises, we assess the prediction accuracy of surrogate models near the minimum of the objective function shown in Eq. (31), defined in a 10-dimensional design space corresponding to the individual layer thicknesses forming a color filter thin-film stack. A test hypercube centered at the optimal color filter design is constructed with a side length of 110 nm, and 10 sample points are selected from a Sobol' sequence within this region to train the GP surrogates. We compare two surrogate models: a single-output GP trained solely on the scalar objective values $g(\boldsymbol{p})$, given by Eq. (31), and a physics-informed model defined as a set of 18 multi-output GPs, each trained on 7 quantities totaling in 126 contributions to the final learned objective. The accuracy of the models in the unexplored regions of the search space is assessed on $N = 200$ uniformly sampled points within the specified hypercube by computing RAE shown in Eq. (A1) between the predictions of the two surrogates and the corresponding ground truth values. Fig. A5 shows that the scalar GP has a noticeably higher RAE compared to the physics-informed model, meaning its predictions are less accurate overall. To better highlight this difference, we examine the prediction accuracy of the surrogate models $\hat{g}_{\mathrm{mgp}}(\boldsymbol{p})$ and $\hat{g}_{\mathrm{so}}(\boldsymbol{p})$ along a one-dimensional slice of the hypercube, generated by varying the thickness of the first layer l_1 of the color filter. Indeed, according to Fig. A6, the physics-informed model $\hat{g}_{\mathrm{mgp}}(\boldsymbol{p})$ provides better approximation of the true objective than the single-output $\hat{g}_{\mathrm{so}}(\boldsymbol{p})$ model. As with the beam-splitter optimization, learning the quantities prior to applying the nonlinear squaring

operation proves to be advantageous, resulting in higher fidelity of the final objective posterior and, ultimately, superior convergence trends linked to physics informed approach, as portrayed on Fig. 10.

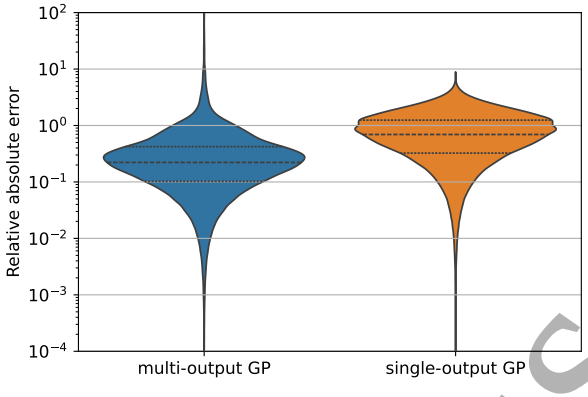

FIG. A5: Relative absolute error distribution of the single-output and the multi-output surrogate model predictions computed on 200 query points in the hyperspace around the color distance objective minimum.

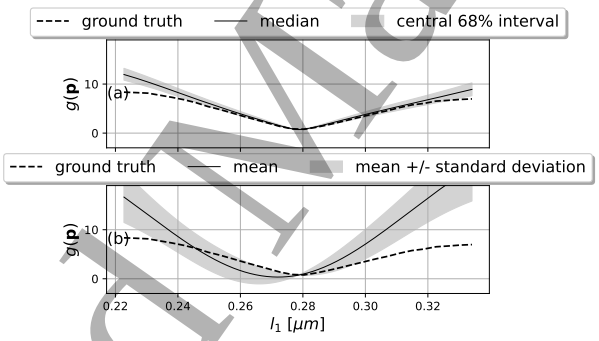

FIG. A6: One-dimensional cross section along the top layer thickness through the color distance objective minimum. Panels (a) and (b) illustrate the behavior of $\hat{g}_{\mathrm{mgp}}(\boldsymbol{p})$ and $\hat{g}_{\mathrm{so}}(\boldsymbol{p})$, respectively.

[1] Rasmussen C. E. and Williams C. K. I., *Gaussian Processes for Machine Learning*, Vol. 2, MIT Press, Cambridge, MA, 2006.

[2] Močkus J., *Bayesian approach to global optimization: theory and applications*, Vol. 37, Springer Science & Bussines Media, 2012.

[3] Garnett R., *Bayesian Optimization*, Cambridge University Press, 2022.

[4] Shahriari B., Swersky K., Wang Z., Adams R. P., and de Freitas N., "Taking the Human Out of the Loop: A Review of Bayesian Optimization," *Proceedings of the IEEE*, Vol. 104, no. 1, pp. 148–175, 2016.

[5] Byrd R. H., Lu P., Nocedal J., and Zhu C., "A limited memory algorithm for bound constrained optimization," *SIAM J. Sci. Comput.*, Vol. 16, pp. 1190–1208, 1995.

[6] Kraft D., "A software package for sequential quadratic programming," Tech. Rep. DFVLR-FB 88-28, 1988.

[7] Rehman S. U. and Langelaar M., "System robust optimization of ring resonator-based optical filters," *J. Light. Technol.*, Vol. 34, pp. 3653—3660, 2016.

[8] Lu J., Boyd S., and Vučković J., "Inverse design of a three-dimensional nanophotonic resonator," *Opt. Express*, Vol. 19, pp. 10563–10570, 2011.

[9] Qin, F., Liu, Z., Zhang, Q., Zhang, H., and Xiao, J., "Mantle cloaks based on the frequency selective metasurfaces designed by Bayesian optimization," *Scientific reports*, Vol. 8, 14033, 2018.

[10] Astudillo R. and Frazier P., "Bayesian optimization of composite functions," *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97, pp. 354–363, 2019.

[11] Uhrenholt A. K. and Jensen B. S., "Efficient Bayesian optimization for target vector estimation," *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, Vol. 89, pp. 2661-2670, 2019.

[12] Plock M., Andrle K., Burger S., and Schneider P. I., "Bayesian target-vector optimization for efficient parameter reconstruction," *Adv. Theory Simul.*, Vol. 5, 2200112, 2022.

[13] Stein M. L., *Interpolation of spatial data: some theory for kriging*, Springer, 1999.

[14] Alvarez M. A., Rosasco L., and Lawrence N. D., "Kernels for vector-valued functions: A review," *Foundations and Trends in Machine Learning*, Vol. 4, pp. 195–266, 2012.

[15] Liu H., Cai J., Ong Y., "Remarks on multi-output Gaussian process regression," *Knowledge-Based Systems*, Vol. 144, pp. 102-121, 2018.

[16] Bonilla E. V., Chai K. M. A., and Williams C. K. I., "Multi-task Gaussian process prediction" *Advances in Neural Information Processing Systems*, Vol. 20, pp. 153-160, 2007.

[17] Müller G., Martínez-Lahuerta V., Sekulic I., Burger S., Schneider P. I., and Gaaloul N., "Bayesian optimization for state engineering of quantum gases," *Quantum Sci. Technol.*, Vol. 10, 015033, 2025.

[18] Schaible J., Winarto H., Škorjanc V., Yoo D., Zimmermann L., Jäger K., Sekulic I, Schneider P. I., Burger S., Wessels A., Bläsi B., and Becker C., "Optimizing aesthetic appearance of perovskite solar cells using color filters," *Solar RRL*, Vol. 9, 2400627, 2025.

[19] Wilson J., Moriconi R., Hutter F., and Deisenroth M., "The reparameterization trick for acquisition functions," *NIPS Workshop on Bayesian Optimization*, 2017.

[20] CMA-ES Covariance matrix adaptation evolution strategy for non-linear numerical optimization in Python (available at: https://pypi.org/project/cma/)

[21] Nelder J. A. and Mead R., "A simplex method for function minimization," *The Computer Journal*, Vol. 7, pp. 308–313, 1965.

[22] https://optimizer.jcmwave.com, 2025.

[23] Virtanen P., Gommers R., Oliphant T. E., Haberland M., Reddy T., Cournapeau D., et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, Vol. 17, pp. 261–72, 2020.

[24] Yu N. and Capasso F., "Flat optics with designer metasurfaces," *Nature Materials*, Vol. 13, pp. 139-150, 2014.

[25] Ni Y., Chen S., Wang Y., Tan Q., Xiao S., and Yang Y., "Metasurface for Structured Light Projection over 120° Field of View," *Nano Lett.*, Vol. 20, pp. 6719–6724, 2020.

[26] Pomplun J., Burger S., Zschiedrich L., and Schmidt F., "Adaptive finite element method for simulation of optical nano structures," *Phys. Status Solidi (b)*, Vol. 244, pp. 3419–3434, 2007.

[27] Gardner J., Kusner M., Xu Z., Weinberger K., and Cunningham J., "Bayesian optimization with inequality constraints," *Proceedings of the 31st International Conference on Machine Learning*, Vol. 32, pp. 937-945, 2014.

[28] Corgier R., Amri S., Herr W., Rudolph J, Guéry-Odelin D., Rasel E. M., Charron E., and Gaaloul N., "Fast manipulation of Bose-Einstein condensates with an atom chip," *New J.*

Phys., Vol. 20, 055002, 2018.

[29] Torrontegui E., Ibánez S., Martínez-Garaot S., Modugno M., del Campo A., Guéry-Odelin D., Ruschhaupt A., Chen X., and Muga J. G., "Shortcuts to adiabaticity," *Adv. At. Mol. Opt. Phys.*, Vol. 62, pp. 117–169, 2013.

[30] Amri S., Corgier R., Sugny D., Rasel E. M., Gaaloul N., and Charron E., "Optimal control of the transport of Bose-Einstein condensates with atom chips," *Sci Rep*, Vol. 9, 5346, 2019.

[31] Pontryagin L. S., Bol'tanskii V. G., Gamkrelidze R. S., and Mischenko E. F., *The mathematical theory of optimal processes*, Pergamon Press, 1964.

[32] Castin Y. and Dum R., "Bose-Einstein condensates in time dependent traps," *Phys. Rev. Lett.*, Vol. 77, pp. 5315-5319, 1996.

[33] Li Y., Li L., Deng W., Zhu D., and Hong L., "Building integrated photovoltaic (BIPV) development knowledge map: A review of visual analysis using CiteSpace," *Buildings*, Vol. 13, 389, 2023.

[34] Bläsi B., Kroyer T., Kuhn T. E., and O. Höhn, "The MorphoColor concept for colored photovoltaic modules," *IEEE Journal of Photovoltaics*, Vol. 11, pp. 1305-1311, 2021.

[35] Macleod H. A., *Thin-Film Optical Filters*, (5th ed.), Taylor & Francis Group, 2017.

[36] Sekulic I., Schaible J., Müller G., Plock M., Burger S., Martínez-Lahuerta V., Gaaloul N., and Schneider P. I., "Data publication for Physics-informed Bayesian optimization of expensive-to-evaluate black-box functions," Zenodo, https://doi.org/10.5281/zenodo.16751508, 2025.