Autonomous efficient experiment design for materials discovery with Bayesian model averaging

Anjana Talapatra,[1,*] S. Boluki,[2] T. Duong,[1] X. Qian,[2] E. Dougherty,[2] and R. Arróyave[1,3]
[1]*Department of Materials Science & Engineering, Texas A & M University, College Station, Texas 77843, USA*
[2]*Department of Electrical and Computer Engineering, Texas A & M University, College Station, Texas 77843, USA*
[3]*Department of Mechanical Engineering, Texas A & M University, College Station, Texas 77843, USA*

(Received 24 April 2018; revised manuscript received 12 September 2018; published 26 November 2018)

The accelerated exploration of the materials space in order to identify configurations with optimal properties is an ongoing challenge. Current paradigms are typically centered around the idea of performing this exploration through high-throughput experimentation/computation. Such approaches, however, do not account for—the always present—constraints in resources available. Recently this problem has been addressed by framing materials discovery as an optimal experiment design. This work augments earlier efforts by putting forward a framework that efficiently explores the materials design space not only accounting for resource constraints but also incorporating the notion of model uncertainty. The resulting approach combines Bayesian model averaging within Bayesian optimization in order to realize a system capable of autonomously and adaptively learning not only the most promising regions in the materials space but also the models that most efficiently guide such exploration. The framework is demonstrated by efficiently exploring the MAX ternary carbide/nitride space through density functional theory (DFT) calculations.

DOI: 10.1103/PhysRevMaterials.2.113803

I. INTRODUCTION

A. Motivation

The accelerated exploration of the materials design space (MDS) has been recognized for more than a decade as a key enabler for potentially transformative technological developments [1,2]. The development of strategies to integrate simulations and experimental data with expert knowledge is a highly active area of research [3,4]. Over time, different methods have been deployed within conventional, human-centric, materials development frameworks for exploration of the MDS, including high-throughput (HT) experimentation and computation.

Traditional HT experimental [5–7] and computational [8] approaches, while powerful, have important limitations as they (i) employ hardcoded workflows and lack flexibility to iteratively learn and adapt based on the knowledge acquired to assure balanced exploration and exploitation of the MDS and (ii) tend to be suboptimal in resource allocation as these approaches generally rely on highly parallelized exploration of the MDS, even in regions that are of low value relative to the objective, or performance metric, that is sought after.

Resource limitation cannot be overlooked as it is often the case that once a bottleneck in HT workflows has been eliminated (e.g., synthesis of ever more expansive materials libraries), another one suddenly becomes apparent (e.g., need for high-resolution characterization of materials libraries). Regardless of how many bottlenecks are eliminated, the fact that ultimately a human must make decisions about what to do with the acquired information implies that HT frameworks face hard limits that will be extremely difficult to overcome. On the computational front, there exist significant fundamental and technological challenges to the (multiscale) simulation of materials [9] that effectively preclude the HT exploration of MDS beyond the use of (sophisticated) methods—such as DFT-based HT simulations [8]—operating at one scale, with relatively small numbers of degrees of freedom.

B. Experiment design

The goal of any experiment design strategy is to identify an action that results in a desired property, which is usually optimizing an *objective function* of the actions. Without loss of generality, we assume minimization of the objective function $f(\mathbf{x})$:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\chi} f(\mathbf{x}), \tag{1}$$

where χ denotes the action space. In materials discovery, each action is equivalent to an input or design parameter setup or a compound, and χ is the materials design space (MDS).

The objective function can have a closed form as a parametric function, i.e., $f(\mathbf{x}, \theta)$, where θ denotes the parameters. If complete knowledge of the values of the parameters exist, then no experiments are needed. In practice, even if a closed form exists, the true values of the parameters are unknown and they may belong to an uncertainty class Θ, governed by a probability measure. Hence, experiments are desired to gain more knowledge concerning the objective function. It is possible that the parameters of the objective function are directly parameters of an underlying system. For example, in Ref. [11] the underlying system is a gene regulatory network and θ is the set of parameters that govern the network. In this context, the experiment space can be different from the action space, e.g., an experiment determines the true value of a parameter of the underlying system, but an action is a gene

*anjanatalapatra@tamu.edu

perturbation subsequently determined by a medical criterion dependent on the value of the parameter. Typically in the context of materials discovery, each experiment corresponds to applying an action, i.e., setting the input parameters, and observing its true objective value (or a noisy observation of it). Whether or not the experiment and action spaces are identical, the best experiment is determined by optimizing an *acquisition function*.

In materials discovery, f is typically a *blackbox function* without a known closed form, and the cost of querying such a function (through expensive experiments/simulations) at arbitrary query point $\mathbf{x}$ in χ is very high. In these cases a surrogate model can be used to approximate the true objective function. This model can either be parametric or nonparametric. The so-called Bayesian optimization (BO) [12] in the literature corresponds to these cases, where the prior model is sequentially updated after each experiment. Bayesian parametric and nonparametric models are widely used in other fields such as bioinformatics [13–18]. When prior knowledge about the form of the objective function exists and/or many observations of the objective values at different parts of the input space are available, one can use a parametric model as a surrogate model. An example of it for finding the alloy with the least energy dissipation at a specific temperature can be found at [19], where due to the availability of the objective values at many initial input points, the authors have assumed a surrogate parametric function and fixed a subset of its parameters for the experiment design loop.

If, as is often the case, no prior knowledge of the behavior of the objective function is available, and limited initial data points are observed, then one can adopt a nonparametric surrogate model for the objective function. In either case, there is an inherent feature selection step, where different potential feature sets might exist. Moreover, there might be a set of possible parametric families as candidates for the surrogate model. Even when employing nonparametric surrogate models, several choices for the kernel functional form might be available. These translate into different possible surrogate models for the objective function. The common approach is to select a feature set and a single family of models and fix this selection throughout the experiment design loop; however, this is not a reliable approach due to the small initial sample size that is ubiquitous in materials science. In this paper we address this problem by framing experiment design as Bayesian optimization under model uncertainty (BOMU), and incorporating Bayesian model averaging (BMA) within Bayesian optimization. Since in the materials discovery context, the objective function is in most cases a target property of the material; hereafter the surrogate model for the objective function is referred to as the *predictive model*.

In the experiments in this paper no prior knowledge about the functional form of the target properties as functions of the potential features exists, and Gaussian process regression (GPR) [20] is employed as the predictive model for each target property. GPR is a flexible model that imposes only continuity and smoothness prior beliefs and can asymptotically converge to the true objective function. Moreover, in our experiments different predictive models correspond to models based on different potential feature sets. But the approach is by no means limited to this case and can be applied when different predictive models correspond to different parametric families or kernel functional forms of nonparametric models.

A key element in an experiment design strategy is the choice of the acquisition function. The next selected experiment is the one that maximizes the acquisition function, which tries to balance the trade-off between the exploitation of the current belief and the exploration of the unqueried regions of the input space. The acquisition function is itself dependent on the modeling of the objective function. Expected improvement (EI) [10] and knowledge gradient (KG) [21,22] are among the most commonly used acquisition functions, having been originally proposed for experiment design under Gaussian belief over the objective values of input setups and observation noise for an off-line ranking and selection problem. Mean objective cost of uncertainty (MOCU) [11,23,24] is another choice for the acquisition function that is more flexible and quantifies the uncertainty impacting the operational objectives. For the connection of MOCU with KG and EGO, the reader can refer to [24].

In the following sections, we cover the mathematics of our proposed algorithm, but the description in words is as follows:

(i) There is a collection of potential models (e.g., models based on different features sets).

(ii) The models are averaged, based on the (posterior) model probabilities based on initial data set to form a BMA.

(iii) Using the expected acquisition function under the BMA, an experiment is chosen that maximizes the expected acquisition.

(iv) The experiment is run, each model is updated and the (posterior) model probabilities are updated.

(v) The expected acquisition under the updated BMA is computed and an experiment is chosen.

(vi) This iteration is done until some stopping criteria (e.g., while objective not satisfied and budget not exhausted), and the best observation so far is selected as the final suggestion.

In Appendix B we have provided more details about the generalized MOCU for experiment design and how the approach in this paper compares to that.

C. Efficient materials discovery

Resource constraints call for the *efficient* evaluation of materials configurations in order to identify regions in the MDS with the *optimal response*. Bayesian optimization (BO) [12,25] provides a sequential model-based approach to solve the problem: first, a prior belief is prescribed over the objective function and then the model (M) is sequentially refined via Bayesian posterior updating. The domain χ is sampled for a query point $\mathbf{x}_{n+1}$ such that an acquisition function $u(\mathbf{x}|\mathcal{D}_n, M)$—constructed from a model of the observed data $\mathcal{D}_n$—is maximized—see Algorithm 1 and Fig. 1. The stopping criteria can be reaching the desired properties or exhausting the experimental budget.

Having mapped the exploration of the MDS to an expensive *blackbox function*, several groups have already demonstrated the power of Bayesian optimization in the context of *accelerated* materials discovery. Early on, Fujimura *et al.* [26] combined DFT and experimental data to construct a model to

Algorithm 1. Bayesian optimization.

1: Initialize $\mathcal{D}_0$
2: **for** $\mathrm{n} = 0, 1, \ldots$ **do**
3: Update statistical model M
4: Select new $\mathbf{x}_{\mathrm{n}+1}$ by optimizing acquisition function u:

$$\mathbf{x}_{\mathrm{n}+1} = \arg\max_{\mathbf{x}\in\chi} u(\mathbf{x}|\mathcal{D}_{\mathrm{n}}, M)$$

5: Query blackbox function f to obtain y_{n+1}
6: Augment data $\mathcal{D}_{\mathrm{n}+1} = \{\mathcal{D}_{\mathrm{n}}, (\mathbf{x}_{\mathrm{n}+1}, y_{\mathrm{n}+1})\}$
7: **if** stopping criteria reached **then**
8: **break**
9: **end if**
10: **end for**

predict the ionic conductivity of Li-superionic conductors via support vector regression (SVR) [27]. The predicted conductivity σ_{Li} from the SVR model was then used as the *acquisition* function to further explore the Li-superionic conductor space. Seko *et al.* [28] used feature sets derived from DFT calculations and experimentally measured melting points T_m to fit stochastic models based on SVR or Gaussian process regression (GPR) [20] to discover unary and binary crystals with the highest melting point. In that case, the acquisition function used in the sequential exploration of the melting point space χ_{T_m} was the *probability* of improving upon the best value recorded before acquisition $n+1$. These early results introduced the notion of sequential exploration but did not consider the larger implications of framing the materials discovery as the optimization of an expensive *blackbox function*.

Balachandran *et al.* [29] prescribed the need to balance the need to exploit our current knowledge of the MDS χ with the need to explore it. The balance between exploitation and exploration was realized by invoking a proper acquisition function. Balachandran *et al.* proposed using *expected improvement* (EI) [10] in the predicted objective function y by the model $P(y|\mathbf{x}, \mathcal{D})$ over the unexplored regions of χ, given the observed data $\mathcal{D}$. EI can in turn be calculated for unexplored query points $\mathbf{x}$ by the model trained. They demonstrated their design protocol by attempting to predict the MAX phases (ternary layered carbides/nitrides [30]) with maximal/minimal polycrystalline bulk/shear moduli as predicted via DFT calculations. Having demonstrated the power of Bayesian optimization in materials discovery, the same group [31] notably employed the same approach to discover, *via experiments*, NiTi-based shape memory alloys (SMAs) with record-low hysteresis through a minimal experimental effort.

The principled nature of a BO-based materials discovery protocol is amenable to develop full-loop platforms, particularly when attempting to carry out simulation-driven materials development. Indeed, Ju *et al.* [32] recently proposed a framework whereby atomistic transport calculations were combined with a BO framework to identify aperiodic nanostructures with optimal transport properties by examining only an extremely small fraction of the possible configurations. On the experimental front, Nikolaev *et al.* [33] recently demonstrated a fully autonomous closed-loop iterative materials experimentation platform. They demonstrated the system by optimizing the synthesis conditions for carbon nanotubes. In their case, the approach focused on a *greedy* exploitation of the synthesis space by using the predicted rate of growth as the acquisition function—i.e., no exploitation-exploration tradeoff [29,31] was used.

D. Contributions of this paper

While existing computational and experimental deployments of optimal materials discovery constitute significant

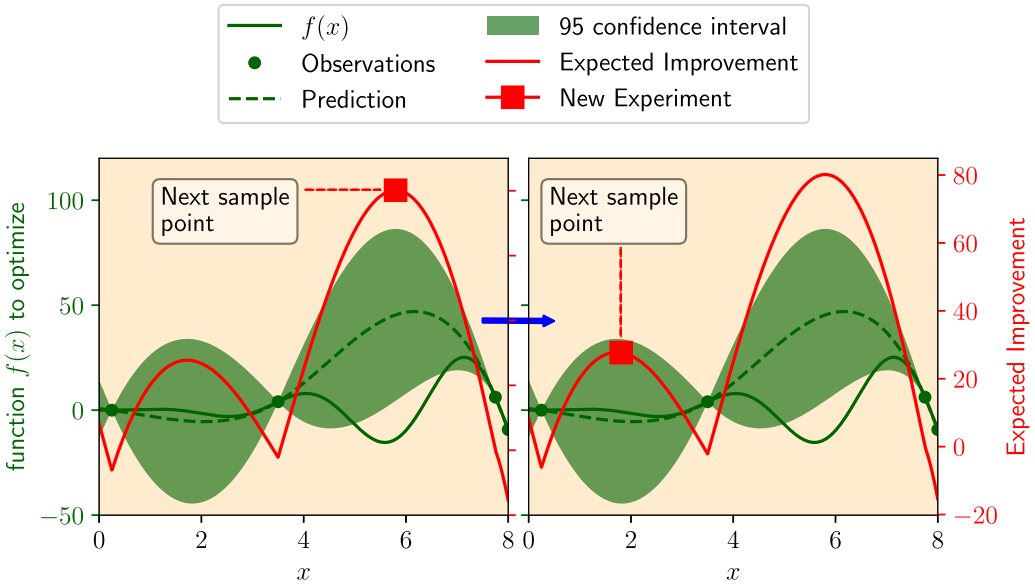

FIG. 1. Schematic illustration of Bayesian optimization (BO): from a limited number of observations on a system (blue solid line) a stochastic model (dashed blue line and shaded area) is built. The next observation is determined by accounting for the tradeoff between the exploitation of the current knowledge and the exploration of the unknown regions of the design domain χ. In this case, expected improvement (EI) is the metric used and thus the policy falls within the efficient global optimization (EGO) framework [10].

advances, there are still significant challenges that remain to be addressed. For example, most BO-based approaches rely on a feature selection step [29,34,35] that necessarily requires a considerable number of feature-property sets to be effective [36]. In other cases, the strength of the approach depends on building sufficient prior knowledge (from informative predictive models [28]) in order for *greedy* approaches to be practical.

Unfortunately, more often than not, the amount of relevant data available before embarking on a materials discovery problem is small. In such situations the nature (and dimensionality) of the design space—χ in the BO formalism—is not known *a priori*. Moreover, it is not even clear which features are best connected to the target performance metric. Finally, the inability of existing approaches to "both build and exploit their internal models, with minimal human hand-engineering" [37] precludes the implementation of truly autonomous materials discovery systems, even in simulation-driven approaches.

In this work we propose a framework that simultaneously (i) accounts for the need to adaptively build increasingly effective models for the accelerated discovery of materials while (ii) accounting for the uncertainty in the models themselves. The framework is then demonstrated by efficiently exploring the MAX ternary carbide space through DFT calculations. Incorporating BMA within Bayesian optimization produces a system capable of autonomously and adaptively learning not only the most promising regions in the materials space but also the models that most efficiently guide such exploration. The framework is also capable of defining optimal experimental sequences in cases where multiple objectives must be met—we note that recent works have begun to address the issue of multiobjective Bayesian optimization [38] in the context of materials discovery. Our approach, however, is different in that the multiobjective optimization is carried out simultaneously with feature selection.

II. BAYESIAN OPTIMIZATION UNDER MODEL UNCERTAINTY

Small sample sizes are ubiquitous in materials science. Experiments—and simulations—are often resource intensive and this imposes significant constraints on any attempt to explore/exploit the MDS. Moreover, in the absence of sufficient information, there are, *a priori*, multiple features that are potentially predictive of the material performance metric of interest. In all the well-known experiment design methods in the literature, one must select the model (the set of predictive features and/or the parametric form or the kernel functional form of the model) before starting the experiment design loop.

Unfortunately, due to small sample size and large number of potential predictive models, the model selection step may not result in the true best predictive model for efficient Bayesian optimization [39,40]. It has been shown that small sample sizes pose a great challenge in model selection due to inherent risk of imprecision and overfitting [39,40], and no feature selection method performs well in all scenarios when sample sizes are small [41]. Thus, by selecting a single model as the predictive model based on small observed sample data, one ignores the model uncertainty [42].

A. Building robust predictive models through Bayesian model averaging

One possible approach to circumvent this problem is to weigh all the possible models by their corresponding probability of being the true model, and use all of these in the experiment design step so that model uncertainty can be taken care of for Bayesian optimization. In other words, the derived predictive model is a marginalized aggregation of all the potential predictive models, weighted by the prior probability and likelihood of the observed data for that model, resulting in the Bayesian model averaging (BMA) method [43,44].

Here we discuss the multioutput case from which the single output can be readily deduced. Let y^j represent the jth output of interest, and $\mathbf{x}$ the corresponding vector of features or materials design parameters, and the observed data be denoted by $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$, where $\mathbf{Y} = [Y^1, \ldots, Y^q]$ is a matrix having the collection of the observed jth output as its jth column, i.e., $Y^j = [y_1^j, \ldots, y_n^j]^T$, where n is the number of observed data points, and $\mathbf{X}$ represent the matrix of the collection of the corresponding observed features. Here, to simplify the notation, we have dropped the subscript denoting the experiment iteration step for $\mathcal{D}$, but note that $\mathcal{D} = \mathcal{D}_n$ at any nth step. The predictive probabilistic model for $\mathbf{y}$ for a new feature vector $\mathbf{x}$ after observing $\mathcal{D}$ is

$$P(\mathbf{y}|\mathbf{x}, \mathcal{D}) = \sum_{i=1}^{L} P(M_i|\mathcal{D}) P(\mathbf{y}|\mathbf{x}, \mathcal{D}, M_i), \tag{2}$$

where $P(\mathbf{y}|\mathbf{x}, \mathcal{D}, M_i)$ represents each potential probabilistic predictive model, and

$$P(M_i|\mathcal{D}) = \frac{P(\mathcal{D}|M_i)P(M_i)}{\sum_{j=1}^{L} P(\mathcal{D}|M_j)P(M_j)}, \tag{3}$$

$$P(\mathcal{D}|M_i) = \int P(\mathcal{D}|\theta_i, M_i) P(\theta_i|M_i) d\theta_i \tag{4}$$

are the (posterior) probability of each model being the true predictive model, and the marginal probability of the observed data under model M_i, respectively. L is the total number of models under consideration, and M_i and θ_i represents the ith model and the vector of ith model parameters, respectively.

If we further assume independence among outputs and let $\mathcal{D}_j$ denote $\{\mathbf{X}, Y^j\}$, we have $P(\mathbf{y}|\mathbf{x}, \mathcal{D}, M_i) = \prod_{j=1}^{q} P(y^j|\mathbf{x}, \mathcal{D}_j, M_i)$ and

$$\begin{aligned} P(\mathcal{D}|M_i) &= \prod_{j=1}^{q} P(\mathcal{D}_j|M_i) \\ &= \prod_{j=1}^{q} \int P\left(\mathcal{D}_j|\theta_i^j, M_i\right) P\left(\theta_i^j|M_i\right) d\theta_i^j. \end{aligned} \tag{5}$$

When each potential probabilistic predictive model M_i is a Gaussian process regression (GPR) model [45], and θ_i^j are the parameters of the covariance function. In fact, each GPR model M_i is defined by a mean (basis) function $[m_i^j(\cdot)]$ and a covariance function $[K_i^j(\cdot, \cdot; \theta_i^j)]$. In this setup, $P(y^j|\mathbf{x}, \mathcal{D}, M_i)$ is a Gaussian distribution, i.e., $P(y^j|\mathbf{x}, \mathcal{D}, M_i) = \mathcal{N}(\mu_i^j(\mathbf{x}), \sigma_i^{2,j}(\mathbf{x}))$, where the predicted

mean and variance of the jth objective function are [45]

$$\mu_i^j(\mathbf{x}) = m_i^j(\mathbf{x}) + K_i^j\left(\mathbf{x},\mathbf{X};\theta_i^j\right)K_i^j\left(\mathbf{X},\mathbf{X};\theta_i^j\right)^{-1}\left[Y^j - m_i^j(\mathbf{X})\right],$$
$$\sigma_i^{2,j}(\mathbf{x}) = K_i^j\left(\mathbf{x},\mathbf{x};\theta_i^j\right) - K_i^j\left(\mathbf{x},\mathbf{X};\theta_i^j\right)K_i^j\left(\mathbf{X},\mathbf{X};\theta_i^j\right)^{-1}K_i^j\left(\mathbf{X},\mathbf{x};\theta_i^j\right). \quad (6)$$

In practice, when using type II maximum likelihood (ML-II) estimation, the covariance function parameters of each model are estimated by maximizing the marginal log likelihood of the observed data under that model, i.e., an estimate $\hat{\theta}_i^j$ is calculated by maximizing

$$\log P\left(\mathcal{D}_j|\theta_i^j, M_i\right) = -\frac{1}{2}\left[Y^j - m_i^j(\mathbf{X})\right]^T K_i^j\left(\mathbf{X},\mathbf{X};\theta_i^j\right)^{-1}\left[Y^j - m_i^j(\mathbf{X})\right] - \frac{1}{2}\left|K_i^j\left(\mathbf{X},\mathbf{X};\theta_i^j\right)\right| - \frac{n}{2}\log 2\pi, \quad (7)$$

where $|\cdot|$ denotes matrix determinant. A quasi-Newton method with multiple random starts can be employed to find the maximum of (7). This estimate $\hat{\theta}_i^j$ is then used in Eq. (6) for prediction purposes under the model assumptions.

For a GPR, $P(\mathcal{D}_j|\theta_i^j, M_i)$ is a multivariate Gaussian probability density function, and $P(\mathcal{D}_j|M_i) = \int P(\mathcal{D}_j|\theta_i^j, M_i)P(\theta_i^j|M_i)d\theta_i^j$ is the marginal probability of the observed data corresponding to jth output under model M_i in Eq. (4), and can be approximated by either first-order expansion of the exponent, or second-order expansion of the exponent known as Laplace approximation method [45]. In the first-order approximation, since $\hat{\theta}_i^j$ is a stationary point of (7), $P(\mathcal{D}_j|M_i)$ can be approximated by $P(\mathcal{D}_j|\hat{\theta}_i^j, M_i)$. In the second-order approximation, $P(\mathcal{D}_j|M_i) \approx P(\mathcal{D}_j|\hat{\theta}_i^j, M_i)\int \exp\{-\frac{1}{2}(\theta_i^j - \hat{\theta}_i^j)^T[-H(\hat{\theta}_i^j)](\theta_i^j - \hat{\theta}_i^j)\}d\theta_i^j$, where $H(\hat{\theta}_i^j)$ is the Hessian matrix of $\log P(\mathcal{D}_j|\theta_i^j, M_i)$ calculated at $\hat{\theta}_i^j$. When all the models are assumed to have the same probability *a priori*, the posterior model probabilities in Eq. (3), i.e., $P(M_i|\mathcal{D}), i = 1,\ldots,L$, are only dependent on the marginal probability of the observed data under each model in Eq. (4), i.e., $P(\mathcal{D}|M_i), i = 1,\ldots,L$.

B. Experiment design by Bayesian optimization

Bayesian experiment design (BED) has the potential to guide efficient search for desired materials by directing sequential search of "optimal" query points to approach the optimal solution [12]. Here we employ the expected improvement (EI) [10] for single objective problems, and an extension of EI to guide the search to approach the Pareto front for multiobjective problems, namely the expected hypervolume improvement (EHVI) [46]. Both EI and EHVI can balance exploration and exploitation up to some extent in guiding the search for optimal solutions.

A *major innovation* in our BED approach is that instead of assuming knowledge of the best predictive model in advance and updating this given predictive model based on the limited number of initial observed data and iterating the experiment design loop based on the updated model—an approach that is taken in the literature—we consider the model uncertainty by including a class of potential predictive models for the task under study. By BMA, the experiment design step is performed based on the weighted average of these potential models. After performing the selected experiment, the new observed data from the experiment is used to update the (posterior) probability of all these potential predictive models. We can see that by taking this approach, as more experiments are done, the true predictive model is selected with a higher probability alongside accelerating the discovery of the material with the desired properties. We note that the proposed BMA also works in cases in which the feature sets are known or fixed but in which different model forms of the GPR—i.e., using different kernels—could potentially have different degrees of fidelity with regards to the available data.

For multiobjective problems, the EHVI under model averaging is

$$EI_{\mathcal{H}}(\mathbf{x}|\mathcal{D}) = \int I_{\mathcal{H}}(\mathbf{y}|\mathbf{x},\mathcal{D})P(\mathbf{y}|\mathbf{x},\mathcal{D})d\mathbf{y} = \int I_{\mathcal{H}}(\mathbf{y}|\mathbf{x},\mathcal{D})\sum_{i=1}^{L}P(M_i|\mathcal{D})P(\mathbf{y}|\mathbf{x},\mathcal{D},M_i)d\mathbf{y} = \sum_{i=1}^{L}P(M_i|\mathcal{D})EI_{\mathcal{H}}(\mathbf{x}|\mathcal{D},M_i), \quad (8)$$

where $I_{\mathcal{H}}(\mathbf{y}|\mathbf{x},\mathcal{D})$ denotes the hyper-volume improvement achieved by observing the outputs at $\mathbf{x}$, E represents expectation, and $EI_{\mathcal{H}}(\mathbf{x}|\mathcal{D},M_i)$ is the ordinary EHVI under model M_i. If the outputs are assumed to be independent $EI_{\mathcal{H}}(\mathbf{x}|\mathcal{D},M_i)$ further simplifies to $\int I_{\mathcal{H}}(\mathbf{y}|\mathbf{x},\mathcal{D})\prod_{j=1}^{q}P(y^j|\mathbf{x},\mathcal{D},M_i)d\mathbf{y}$. The optimal experiment to be performed next is $\mathbf{x}^* = \text{argmax}_{\mathbf{x}\in\mathcal{X}}\,EI_{\mathcal{H}}(\mathbf{x}|\mathcal{D})$, which is the one that maximizes the weighted average EHVI considering all the potential predictive models, based on the iteratively updated (posterior) model probabilities given the observed data. The hypervolume improvement $I_{\mathcal{H}}(\mathbf{y}|\mathbf{x},\mathcal{D})$ is the increase in the hypervolume of the dominated (objective) space achieved by adding the outputs at $\mathbf{x}$ to the observed data, i.e., $I_{\mathcal{H}}(\mathbf{y}|\mathbf{x},\mathcal{D}) = \mathcal{H}(\mathbf{Y}\cup\mathbf{y}) - \mathcal{H}(\mathbf{Y})$. Without loss of generality, if we assume the goal is minimization of all the outputs, the hypervolume dominated by a set of points $\mathbf{A}$ is defined as the volume of the dominated subspace by the points in A, i.e., $\mathcal{H}(\mathbf{A}) = \text{Volume}(\{\boldsymbol{s}\in\mathbb{R}^q | \boldsymbol{s}\prec\boldsymbol{r} \text{ and } \exists\,\mathbf{a}\in\mathbf{A} : \mathbf{a}\prec\boldsymbol{s}\})$, where the domination rule is such that $\mathbf{a}\prec\mathbf{b}$ if and only if $a^j \leqslant b^j$ for all $j = 1,\ldots,q$, and for at least one $j, a^j < b^j$. $\mathbf{r}$ is called a reference or anchor point and is a point dominated by all the possible output values (the whole output space).

For the special case of employing EI-based BED [10], the EI after observing data $\mathcal{D}$ can be computed under model averaging by

$$EI(\mathbf{x}|\mathcal{D}) = \int I(y|\mathbf{x},\mathcal{D})\sum_{i=1}^{L}P(M_i|\mathcal{D})P(y|\mathbf{x},\mathcal{D},M_i)dy = \sum_{i=1}^{L}P(M_i|\mathcal{D})\int I(y|\mathbf{x},\mathcal{D})P(y|\mathbf{x},\mathcal{D},M_i)dy = \sum_{i=1}^{L}P(M_i|\mathcal{D})EI(\mathbf{x}|\mathcal{D},M_i), \quad (9)$$

where $I(y|\mathbf{x}, \mathcal{D})$ denotes the improvement achieved by observing the output of experiment $\mathbf{x}$, and $EI(\mathbf{x}|\mathcal{D}, M_i)$ is the EI under model M_i. In this approach, the optimal experiment to be performed next is $\mathbf{x}^* = \text{argmax}_{\mathbf{x}\in\chi}\ EI(\mathbf{x}|\mathcal{D})$. We can see that the optimal experiment is the one that maximizes the weighted average EI considering all the potential predictive models based on the iteratively updated (posterior) model probabilities given the observed data. In the equations above, the improvement achieved by observing the output of experiment $\mathbf{x}$ is $I(y|\mathbf{x}, \mathcal{D}) = (y^* - y)_+$ when minimization is the goal, and $I(y|\mathbf{x}, \mathcal{D}) = (y - y^*)_+$ when maximization is the goal, where $(a)_+ = a$ if $a > 0$ and is zero otherwise, and y^* denotes the best (lowest/highest for minimization/maximization problems) output observed so far, i.e., the best output in $\mathcal{D}$.

For the GPR model assumptions taken by the experiments in this paper, we have chosen the constant mean function [i.e., $m_i(\mathbf{x}) = c_i$ for single output and $m_i^j(\mathbf{x}) = c_i^j$ for multiple output cases] and the (Gaussian) radial basis function (RBF) kernel, a popular choice, for the covariance function:

$$K_i^j\left(\mathbf{x}, \mathbf{x}'; \theta_i^j\right) = \theta_{i,1}^j \exp\left[-\frac{1}{2}\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\theta_{i,2}^j}\right]. \tag{10}$$

The focus of the experiments in this paper is on showing the power of experiment design considering model uncertainty by BMA in guiding the search towards the optimal compound (with corresponding features design parameters) when the best predictive model is not known in advance, a usual case in practical applications, while also identifying the best predictive model as more data from experiments become available. The algorithm for our proposed Bayesian optimization under model uncertainty (BOMU) framework is shown in Algorithm 2 and the overall framework for autonomous materials discovery is shown in Fig. 2. In Algorithm 2, for the single-objective case, $u(\mathbf{x}|\mathcal{D}_\mathrm{n}, M_i)$ and $u(\mathbf{x}|\mathcal{D}_\mathrm{n})$ correspond to $EI(\mathbf{x}|\mathcal{D}_\mathrm{n}, M_i)$ and $EI(\mathbf{x}|\mathcal{D}_\mathrm{n})$, and for the multiobjective case correspond to $EI_\mathcal{H}(\mathbf{x}|\mathcal{D}_\mathrm{n})$ and $EI_\mathcal{H}(\mathbf{x}|\mathcal{D}_\mathrm{n}, M_i)$, respectively.

In this paper we consider predictive models based on different potential feature sets. The details are provided in the following section.

Algorithm 2. Bayesian optimization under model uncertainty.

1: Initialize $\mathcal{D}_0$
2: **for** $\mathrm{n} = 0, 1, \ldots$ **do**
3: Update statistical model(s), M_i
4: Compute acquisition function u with model averaging:

$$u(\mathbf{x}|\mathcal{D}_\mathrm{n}) = \textstyle\sum_{i=1}^{L} P(M_i|\mathcal{D}_\mathrm{n}) u(\mathbf{x}|\mathcal{D}_\mathrm{n}, M_i)$$

5: Select new $\mathbf{x}_{\mathrm{n}+1}$ by optimizing acquisition function u:

$$\mathbf{x}_{\mathrm{n}+1} = \arg\max_{\mathbf{x}\in\chi} u(\mathbf{x}|\mathcal{D}_\mathrm{n})$$

6: Query blackbox function f to obtain y_{n+1}
7: Augment data $\mathcal{D}_{\mathrm{n}+1} = \{\mathcal{D}_\mathrm{n}, (\mathbf{x}_{\mathrm{n}+1}, y_{\mathrm{n}+1})\}$
8: **if** stopping criteria reached **then**
9: **break**
10: **end if**
11: **end for**

III. DEPLOYMENT OF BOMU: OPTIMAL DISCOVERY OF THE MAX PHASE SPACE

$M_{n+1}AX_n$ phases—M corresponds to a transition metal, A corresponds to group IV and VA elements, and X corresponds to carbon or nitrogen—have a property range within those of ceramics and metals due to the coexistence of both metallic and metallic/covalent bonds within their layered structures [30,47–51], Fig. 3. The bonds between M-A layers tend to be much weaker than those between M-X layers, making them easily deformable while retaining much of the chemical (and thermodynamic) stability of transition MX carbides. While only a very small fraction of the pure ternary MAX phase composition palette has been synthesized to date [52], there is a considerable opportunity to uncover promising chemistries with optimal property sets once different stacking sequences and deviations from stoichiometries in the M, A, and X sites are considered [53,54].

A. Design problem: Optimal mechanical properties in the MAX phase space

Because of their rich chemistry and the wide range of values of their properties [55], MAX phases constitute an adequate material system to test simulation-driven—specifically DFT calculations—materials discovery frameworks. Aryal *et al.* [55], for example, carried out an exhaustive investigation of the structural, electronic, and stability properties of 792 MAX phases with the $M_{n+1}AX_n$ and $n = 1$–4. Balachandran *et al.* [29] used the MAX phases with M_2AX stoichiometry to deploy and test different Bayesian optimization schemes. In this work we use the same system to test the proposed framework.

The MDS for this work is composed of conventional MAX phases with M_2AX and M_3AX_2 stoichiometries. Here M $\in$ {Sc, Ti, V, Cr, Zr, Nb, Mo, Hf, Ti}; A $\in$ {Al, Si, P, S, Ga, Ge, As, Cd, In, Sn, Tl, Pd}; and X $\in$ {C, N}. This results in 216 M_2AX and 216 M_3AX_2 phases. Since we are testing a materials discovery framework, we found it convenient to determine the ground truth of the system beforehand and the mechanical properties of these systems were thus determined before deploying the BOMU framework—our framework has been incorporated into a high-throughput workflow automation tool using the scikit-learn [56] toolbox. The implementation is publicly available at https://gitlab.com/tammal/matpredict. Of the possible MAX phases with the chemistries described above, 29 were found to be elastically unstable and were discarded. The design space thus consists of 403 MAX phases.

The problem was formulated with the goal of identifying the material/materials with (i) the maximum bulk modulus K; (ii) the minimum shear modulus G; and (iii) the maximum bulk modulus and minimum shear modulus. The cases of (i) the maximum bulk modulus K; (ii) the minimum shear modulus G are designed as single-objective optimization problems. The third problem which seeks to identify the materials with the maximum bulk modulus and minimum shear modulus (iii) is designed as a multiobjective problem.

B. Prior knowledge

In this framework it is assumed that some prior knowledge is available before starting the materials discovery task. This

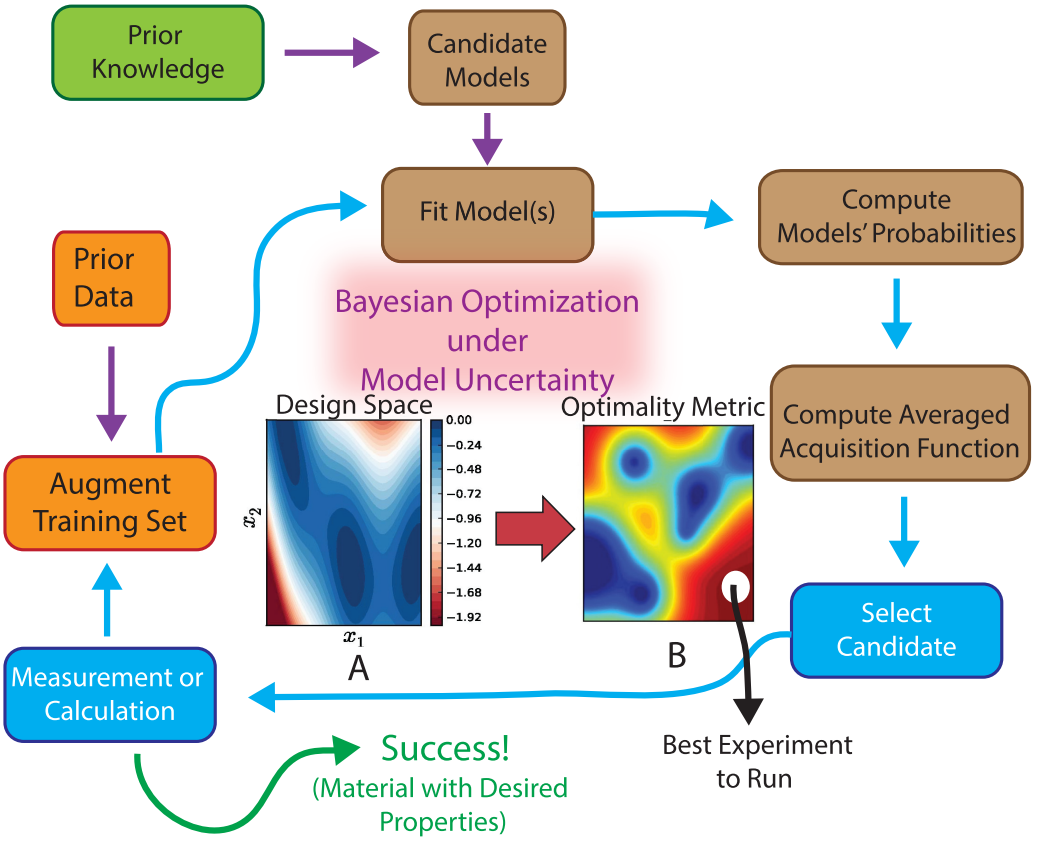

FIG. 2. Schematic of the proposed framework for an autonomous, efficient materials discovery system as a realization of Bayesian Optimization under Model Uncertainty (BOMU). Initial data and a set of candidate models are used to construct a stochastic representation of an experiment/simulation. Each model is evaluated in a Bayesian sense and its probability is determined. Using the model probabilities, an effective acquisition function is computed, which is then used to select the next point in the materials design space that needs to be queried. The process is continued iteratively until target is reached or budget is exhausted. Adapted from the adaptive design strategy framework in [29].

prior knowledge could be as simple as a set of features that are likely to have effects on the materials properties of interest. Here we describe the features that were selected to constitute the MDS:

Each candidate MAX phase c_i of the MDS is determined uniquely by a set of features F_j, which have been distilled from the existing literature and the authors' domain knowledge. In the MDS, the M and A chemical elements comprising the MAX phases sweep along the rows and columns of the periodic table and it is reasonable to expect that features which intrinsically reflect periodic trends would characterize the properties of the MAX phases.

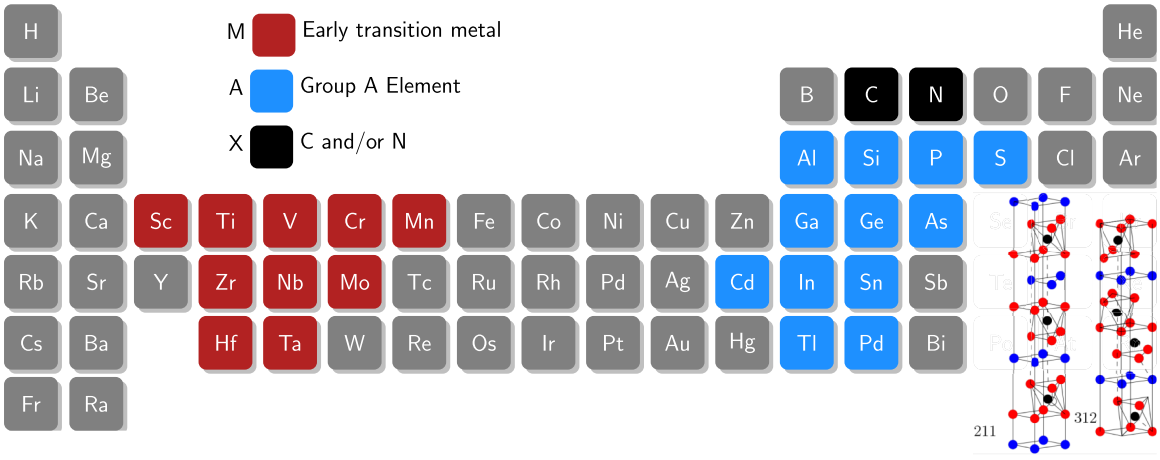

FIG. 3. MAX phases: Ternary (and higher order) layered carbides and nitrides with properties intermediate to those of metals and ceramics [30].

In this work, a total of 15 features were considered: empirical constants C, m which relate the elements comprising the material to its bulk modulus [57]; valence electron concentration C_v; electron to atom ratio $\frac{e}{a}$; lattice parameters a and c; atomic number Z; interatomic distance I_{dist}; the groups according to the periodic table of the M, A, and X elements $\text{Col}_\text{M}, \text{Col}_\text{A}, \text{Col}_\text{X}$, respectively; the order O of MAX phase (whether of order 1 corresponding to M_2AX or order 2 corresponding to M_3AX_2); the atomic packing factor (APF); average atomic radius (rad); and the volume/atom (vol). In relevant cases ($C, m, Cv, \frac{e}{a}, Z, I_{\text{dist}}$, APF, C_v), these features are composition-weighted averages calculated from the elemental values and are assumed to propagate as per the Hume-Rothery rules.

The C, m parameters are related to the bonding character. These are composition-weighted values of the empirical constants reported by Makino *et al.* [57], who proposed that the bulk modulus **K** of elemental substances can be determined by the relation $K = Cr_{\text{ps}}^{-m}$; where r_{ps} is the effective pseudopotential radius. The valence electron concentration C_v is another feature related to the bonding character and is a known marker of the stability of a phase [58,59]. The $\frac{e}{a}$ ratio, which is the average number of itinerant electrons per atom, plays a significant role in the bonding of a solid and is closely related to the valence electron concentration C_v [60].

The lattice parameters c, a for all the domain elements were calculated by DFT by allowing the structures to completely relax. The c lattice parameter is highly correlated to the order of the MAX phase (whether M_2AX or M_3AX_2). The lattice parameters implicitly account for the effect of volume and atomic radius on the elastic properties. Additionally, the c/a ratio characterizes the MAX phases, they being hexagonal close packed (hcp) materials. The relationship between the elastic properties and the c/a ratio for hcp materials has also been extensively studied [61,62]. Here we note that since the determination of the equilibrium structural parameters is approximately an order of magnitude less costly than the full calculation of the elastic constant tensor and thus it is a reasonable proposition to use these DFT-derived quantities to assist in the prediction/discovery of properties that are more costly to acquire.

The atomic number Z, which denotes the number of electrons, is the foremost factor that determines the chemical bonding behavior of a material and defines its chemical properties. The weighted interatomic distance I_{dist} was calculated from the elemental values, which were sourced from the CRC Handbook of Chemistry and Physics [63]. The atomic packing factor (APF) plays an important role in the determination of elastic properties. For example, face centered cubic (fcc) structures tend to be ductile, while hcp structures are brittle. Finally, the structural parameters: average atomic radius (rad) and the volume/atom (vol) were determined from the DFT-determined lattice parameters.

C. Determining candidate models

As discussed, the determination of features comprising the MDS was based off of prior literature and domain knowledge. *A priori*, it is not known which of these features significantly influence the target properties in the materials discovery task.

TABLE I. Feature sets considered in this design.

F_1	$[C, m, C_v, c]$
F_2	$[m, Z, I_{\text{dist}}, \frac{e}{a}]$
F_3	$[\frac{e}{a}, a, c, C_v]$
F_4	$[C, m, C_v, \text{Col}_\text{A}]$
F_5	$[\text{Col}_\text{M}, \text{Col}_\text{A}, \text{Col}_\text{X}, O]$
F_6	$[a, c, \text{APF}, I_{\text{dist}}]$

In the search for new materials with desired properties, such situations are often encountered, where there is a lack of fundamental knowledge relating the intrinsic nature of the material and the desired property. The BOMU approach invoked in this work accounts for uncertainty in the models M_i available to fit the blackbox predictive model to observed data. In our design problem, different models M_i correspond to different subsets $\mathcal{F}_S$ out of the entire feature set $\mathcal{F}$, $\mathcal{F}_S \subseteq \mathcal{F}$.

While one could question the need to define candidate feature subsets $\mathcal{F}_S$ when the entire feature set $\mathcal{F}$ is available, it is important to note that exploring the entire feature set is problematic due to important limitations [64]. First, nonparametric regression is challenging in high-dimensional space, with lower bounds of nearest-neighbor distance between samples depending exponentially on the dimension of the problem [65]. This exponential complexity affects the convergence rate of BO approaches [66]. Second, the computational effort in maximizing the acquisition function also increases in an exponential manner with the number of features.

The general problem of Bayesian optimization in the presence of many potential models (feature sets) is still an outstanding challenge [67] and different approaches have been proposed, including the partitioning of the domain in disjoint subdomains [64] or the use of random embedding [67]. No approach so far provides the means for the BO framework itself to "learn" the optimal model and select the subspace most effectively to reach the target property(ies). Our proposed approach, as will be shown below, *addresses these issues* and thus constitutes a novel approach to effectively reduce the complexity of the BO problem under model uncertainty.

D. Selecting feature sets

Feature selection is an essential component of model construction and learning and is a research area in itself. Application of rigorous feature selection methods can lead to better models with a good understanding of the underlying characteristics of the data. Using the right features reduces the complexity of the model and reduces overfitting. Choosing the right subset of features also improves the accuracy of a model. For the purposes of this work, we elected to see how far one can get by choosing to rely on simpler methods. To reduce the feature space dimensionality of the model, we grouped the features into six sets containing four features each, as shown in Table I. Of the 15 features considered, only 13 were used, with (rad) and (vol) being discarded.

These sets were created *ad hoc*, using a combination of physical insights and an effort to make sets containing features which reflect the effect of electronic structure and chemical bonding character. For example, since C and m are derived

from Makino's empirical model [57], they were grouped together in sets F_1 and F_4. In set F_2, m was used standalone, since the empirical relationship $K = Cr_{ps}^{-m}$ indicates that m is more significant than C, which only introduces the effect of a constant. In set F_5, only the compositional element markers (Col_M, Col_A, Col_X) along with the order O of MAX phase were used, to simulate a feature set which has only the most basic compositional and structural description.

E. Materials discovery/design protocol

GPR models based on six feature subsets in Table I were adopted in our BMA experiment design. For each of the targets (maximizing K, minimizing G, as well as maximizing K/minimizing G) we carried out the sequential experiment design by maximizing the EI or EHVI based on predictive models using single feature sets or BMA using all the feature sets accounting for their probability through first-order (BMA_1) and second-order (BMA_2) Laplace approximation.

The optimization scheme was run for initial data sets (i.e., known data points) of size $N = 2, 5, 10, 15, 20$. The "training set" thus ranges from $\approx 1/200$ to $1/20$ of the MDS. 1500 instances of each initial set N were used to ensure a stable average response. The budget for the optimal design was set at $\approx$20% of the MDS, i.e., 80 materials or calculations. In each iteration, two calculations were done. The selection for the compound(s) to query is based on the optimal policy used: EI or EHVI. Thus the candidates with the maximum and second maximum EI/EHVI are selected for update. This means that for example, for the maximization of the bulk modulus problem for $N = 2$, we initially know the bulk modulus of 2 materials ($N = 2$) and can calculate the bulk modulus of 78 more materials to stay within the budgeted 80 calculations. Since we are calculating the bulk modulus of two materials at a time, this means a total of $78/2 = 39$ iterations for this case. All the input features were normalized, before being fed to the optimization module.

F. DFT calculation parameters

The total energy calculations were performed within the DFT [68] framework, as implemented in the Vienna *ab initio* simulation package (VASP) [69,70]. The generalized gradient approximation (GGA) [71] is used in the form of the parametrization proposed by Perdew, Burke, and Ernzerhof (PBE) [72]. Brillouin zone integrations were performed using a Monkhorst-Pack mesh [73] with at least 5000 k points per reciprocal atom. Full relaxations were realized by using the Methfessel-Paxton smearing method [74] of order one and a final self-consistent static calculation with the tetrahedron smearing method with Blöchl corrections [75]. A cutoff energy of 533 eV was set for all of the calculations and the spin polarizations were taken into account.

To estimate the lattice parameters, the structures were allowed to fully relax to their ground states. The relaxations were carried out in six stages: first stage by allowing change in volume (corresponding to the VASP ISIF = 7 tag), second stage by additionally allowing the relaxation of cell shape (corresponding to the VASP ISIF = 6 tag), third stage by also allowing relaxation of ions (corresponding to the VASP ISIF = 3 tag), fourth stage by allowing only the ions to relax (corresponding to the VASP ISIF =2 tag), fifth stage by allowing full relaxation (VASP ISIF = 3 tag), and a final self-consistent static calculation run. All relaxations were carried out until changes in total energy between relaxation steps were within 1×10^{-6} eV.

The elastic constants were calculated using the stress-strain approach [76,77] where a set of strains ($\epsilon_1; \epsilon_2; \epsilon_3; \epsilon_4; \epsilon_5; \epsilon_6$) were imposed on a crystal, determined using DFT methods as described in [78]. From the n set of strains and the resulting stresses, elastic constants were calculated based on Hooke's law. For these calculations, the ionic positions were relaxed while leaving the lattice shape and volume invariant. These calculations were followed by a static calculation using order-one Methfessel-Paxton smearing method and an auxiliary FFT grid to ensure maximum accuracy in the calculation of interatomic forces. Convergence criteria ensured that calculated elements of elastic constant tensor changed within a few GPa when varying the magnitude of the lattice strain from 1% to 3%. From these elastic constants, various elastic properties have been calculated using the Voigt and Reuss approximations and Voigt-Reuss-Hill averaging [79]. The properties under consideration are: the bulk modulus (K) and the shear modulus (G).

IV. RESULTS

As mentioned earlier, we employ the EI and EHVI acquisition functions in the experiment design loop for single and multiobjective problems, respectively. Hereafter, a single model is a Gaussian process regression (GPR) model based on a single feature set. Also, F_1, F_2, F_3, F_4, F_5, and F_6 denote the six different feature sets considered in our analysis, the GPR models based on those feature sets, and experiment design assuming the underlying model based on those feature sets, interchangeably. In the following "convergence" for each model (feature set) refers to the calculation number in the experiment design iterations based on that model (feature set) when the true optimal design parameters are identified in (nearly) all simulations with 1500 initial data sets with different size N for each setup.

A. Single objective optimization

1. Maximization of bulk modulus (K)

As mentioned earlier, calculations were carried out for different number of initial data instances $N = 2, 5, 10, 15, 20$. The performance trends for all three problems across different values of N are consistent. The technique is found to not significantly depend on quantity of initial data. Figure 4 shows the average number of calculations required to find maximum bulk modulus for $N = 2, 5, 10, 15, 20$ with F_2. Even when we start with very few initial data instances at $N = 2$, the Bayesian experiment design (BED) procedure converges at least as fast as $N = 20$. Using $N = 5$ however, leads to faster convergence than starting with $N = 10, 15, 20$. This shows that it is often more effective to start with a small initial data set. This is advantageous, since in real-world problems, scarcity of data is a common limitation. Consequently, for the sake of brevity, we present results using the representative

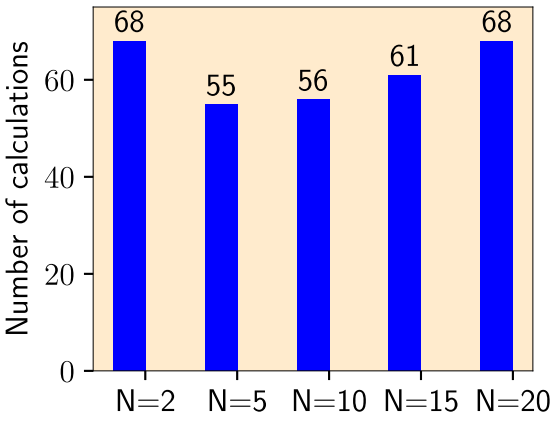

FIG. 4. Average number of calculations required to find maximum bulk modulus for different numbers of initial data instances $N = 2, 5, 10, 15, 20$ using feature set F_2.

case of $N = 10$ only. Results for $N = 2, 5, 15, 20$ may be found in the Supplemental Material [80]. For the first test problem to find the MAX phase with the maximum bulk modulus, the maximum values found in the experiment design iterations based on each model (feature set) averaged over all initial data set instances for $N = 10$ are shown in Fig. 5(a). The dotted line in the figure indicates the maximum bulk modulus = 300 GPa that can be found in the MDS. F_2 is found to be the best performing feature set on average, converging fastest to the maximum bulk modulus. In other words, using the predicted values as well as uncertainty estimation from the GPR model with F_2 in the experiment design loop guides us toward the optimal solution of the problem faster than the other models. F_6 and F_5 on the other hand, are uniformly the worst performing feature sets on average, converging the slowest.

Figure 5(b) shows the swarm plots indicating the number of calculations required to discover the maximum bulk modulus in the MDS using experiment design based on single models for the 1500 initial data instances with $N = 10$. The width of the swarm plot at every vertical axis value indicates the proportion of instances where the optimal design parameters were found at that number of calculations. Bottom heavy, wide bars, with the width decreasing with the number of steps is desirable, since that would indicate that a larger number of instances needed a fewer number of steps to converge.

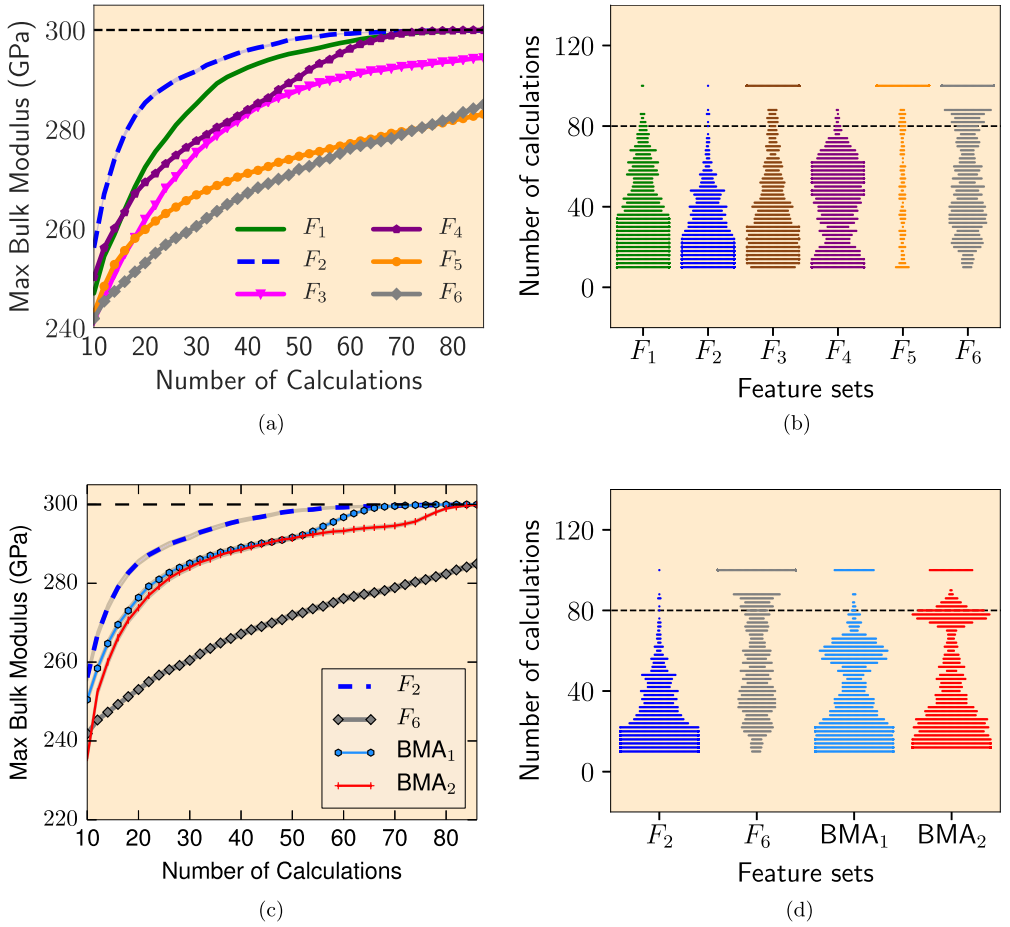

FIG. 5. Representative results for single objective optimization—maximization of bulk modulus for $N = 10$: (a) Average maximum bulk modulus discovered using all described feature sets, (b) swarm plots indicating the distribution of the number of calculations required for convergence using all described feature sets, (c) average maximum bulk modulus discovered using the best feature set F_2, worst feature set F_6, BMA_1 and BMA_2, and (d) swarm plots indicating the distribution of the number of calculations required for convergence using best feature set F_2, worst feature set F_6, BMA_1 and BMA_2.

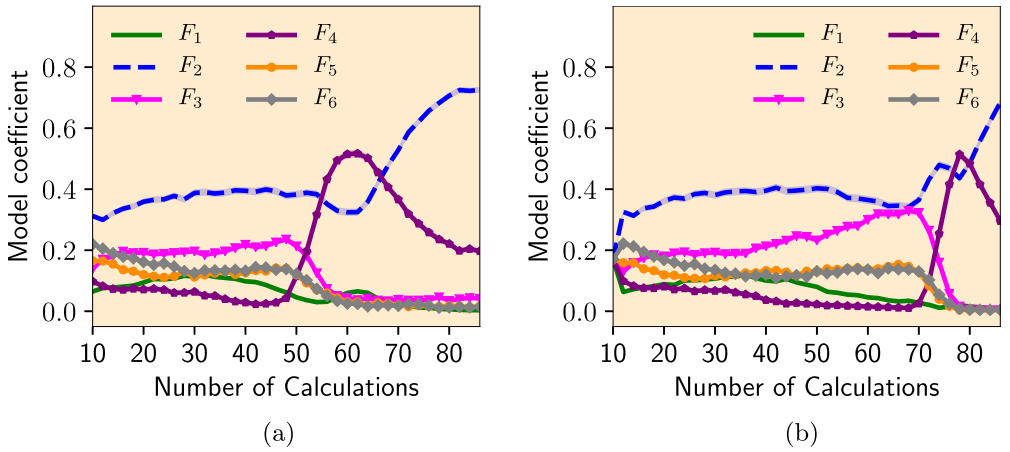

FIG. 6. Average model probabilities for maximizing bulk modulus using (a) BMA_1 and (b) BMA_2.

The dotted line indicates the budget allotted, which was 80 calculations. Instances that did not converge within the budget were allotted a value of 100. Thus the width of the plots at vertical value of 100, corresponds to the proportion of instances which did not discover the maximum bulk modulus in the MDS within the budget. From this figure, it is seen that for F_1, F_2, and F_4 in almost 100% of instances the maximum bulk modulus was identified within the budget, while F_5 is the poorest feature set and the maximum was identified in very few instances.

Figure 5(c) shows the comparison of the average performance of both the first-order and second-order BMA over all initial data set instances with the best performing model (F_2) and worst performing model (F_6). It can be seen that both the first-order and second-order BMA performance in identifying the maximum bulk modulus is consistently close to the best model (F_2). First-order BMA performs as well as if not better than F_2. Figure 5(d) shows the corresponding swarm plots indicating the number of calculations required to discover the maximum bulk modulus in the MDS for the 1500 instances of initial data set for $N = 10$ using first- and second-order BMA, respectively. It can be seen that for a very high percentage of cases the maximum bulk modulus can be found within the designated budget.

In Figs. 6(a) and 6(b) the average model coefficients (posterior model probabilities) of the GPR models based on different feature sets over all instances of initial data set are shown with the increasing number of calculations for BMA_1 and BMA_2, respectively. It can be seen that these model coefficients from BMA may guide automatic selection of the best feature set F_2. For BMA_1 and BMA_2, the average probability of F_2 is (almost) always higher than the other models. Earlier, in Figs. 5(a) and 5(b), F_4 also appears to be a good model and converges at par with F_2 around the 75 calculations. Reflecting this, as the number of available experiments/calculations increases (55 for BMA_1 and 75 for BMA_2), the model probability of F_4 briefly overtakes that of F_2 as indicated in Fig. 6. As more data become available, BMA again considers F_2 as the best model based on the updated model coefficients during the experiment design procedure. Note that such a feature set selection based on BMA is directly determined by the performance of achieving desired operational objectives for experiment design. The actual candidate materials selected during each progressive BED iteration with BMA_1 were analyzed over the 1500 instances, among which the cumulative percentage of choosing candidates with the maximum ($K^1_{\max}$), second maximum ($K^2_{\max}$), and third maximum ($K^3_{\max}$) bulk modulus is indicated in Fig. 7. It is seen that as the BED loop proceeds and the surrogate model improves, the materials with the maximum bulk modulus (top 3 for illustration) are selected more consistently. Specifically, beyond approximately 40 calculations, there is a steep increase in the selection of $K^i_{\max}$ as a candidate, corresponding to the steep increase in the probability of model F_4 and F_2 as illustrated in Fig. 6(a).

2. Maximization of bulk modulus: Noninformative features

To showcase the utility of our BMA approach, we simulate a high-dimensional case by adding 16 noninformative random features, which we compose into subsets F_7, F_8, F_9, and F_{10} of four features each. We carry out two types of calculation using the larger set of 29 (13+16) features. First, we use the BMA_1 approach to find material with maximum K using $F_1, \ldots, F_{10}$; and we use the regular EGO-GP framework to

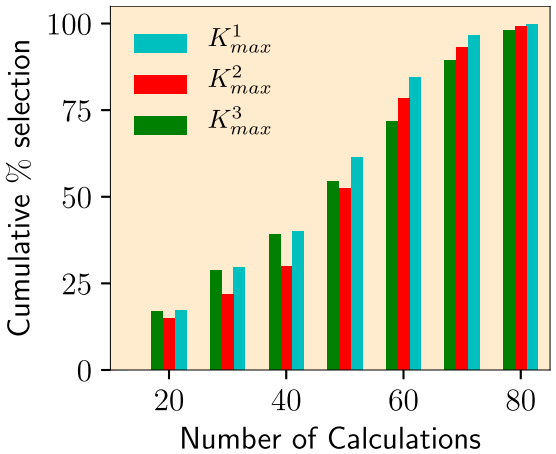

FIG. 7. Percentage of BED selected materials with the maximum ($K^1_{\max}$), second maximum ($K^2_{\max}$), and third maximum ($K^3_{\max}$) bulk modulus with the increasing number of calculations for BMA_1.

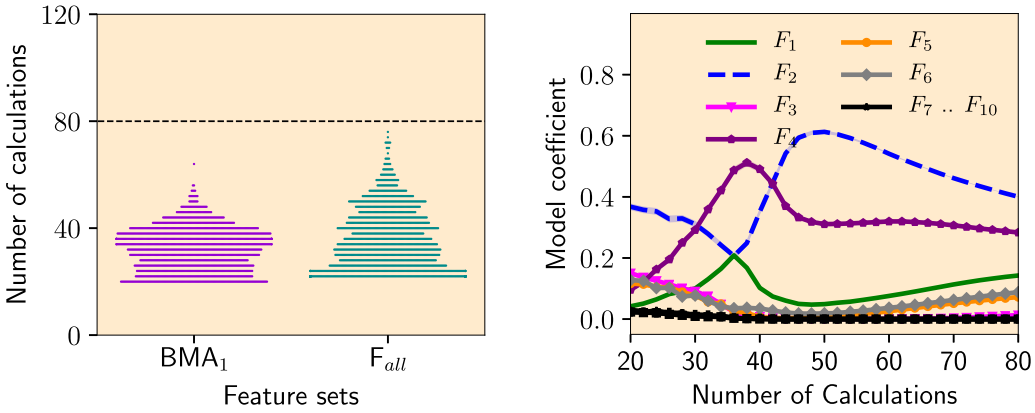

FIG. 8. Representative results for single objective optimization—minimization of shear modulus for $N = 20$ for the case of 29 features: (a) Average model probabilities for maximizing bulk modulus using BMA_1 and F_{all} (b) swarm plots indicating the distribution of the number of calculations required for convergence using BMA_1 and F_{all}.

find the material with maximum K using all 29 features. The results for the same are plotted in Fig. 8. First, we see in Fig. 8(a) that in this case (an actual high dimensional case with a number of noninformative random features) the BMA approach outperforms using all features together. Additionally, tracking the model probabilities as in Fig. 8(b), shows us that the BMA approach effectively picks up the F_2 set as the best feature set, rejects the random feature sets $F_7, \ldots, F_{10}$ (average model probabilities are negligible) and performs better than using F_2 standalone [in Fig. 5(d)].

3. Minimization of shear modulus (G)

Similar to maximization of bulk modulus, the optimization for the minimization problem was carried out for feature sets $F_1, \ldots, F_6$, and then by using BMA_1 and BMA_2. The overall trend in the results was also similar: F_2 is found to be the best performing model on average, converging fastest to the minimum shear modulus. On the other hand, F_6 is uniformly the worst performing feature set on average, converging the slowest. The minimum shear modulus found in the experiment design iterations based on the best model (F_2), worst model (F_6), BMA_1 and BMA_2, averaged over all initial data instances are shown in Fig. 9(a) for $N = 10$. The dotted line in the figure indicates the minimum shear modulus = 10.38 GPa that can be found in the MDS. The performance of both first-order and second-order BMA in identifying the minimum shear modulus lies close to that of the best single model (F_2). Figure 9(b) shows the swarm plots corresponding to the results in Fig. 9(a). It is seen that in almost 100% of the cases the optimal solution (minimum shear modulus) can be found within the designated budget when feature set F_2 is used, while very few instances of convergence are noted for F_6. Using BMA_1 and BMA_2 yields very satisfactory results, as a large majority of the cases converge within budget. Here we see the advantage of using the BMA approach. Without having actually gone through the experiment design loop, one could not know *a priori*, that using F_6 will result in not arriving at the desired material within a reasonable budget with a very high probability. This shows that if one were to just select a

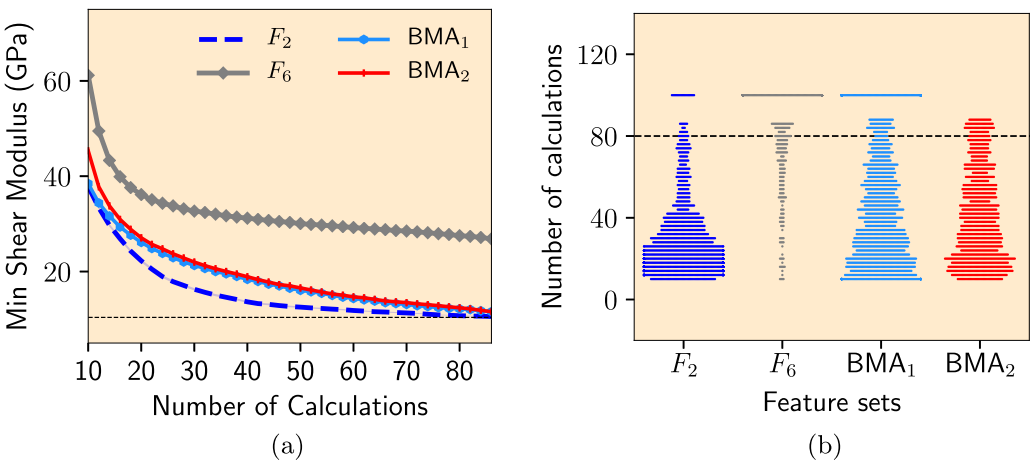

FIG. 9. Representative results for single objective optimization—minimization of shear modulus for $N = 10$: (a) Average minimum shear modulus discovered using the best feature set F_2, worst feature set F_6, BMA_1 and BMA_2, and (d) swarm plots indicating the distribution of the number of calculations required for convergence using best feature set F_2, worst feature set F_6, BMA_1 and BMA_2.

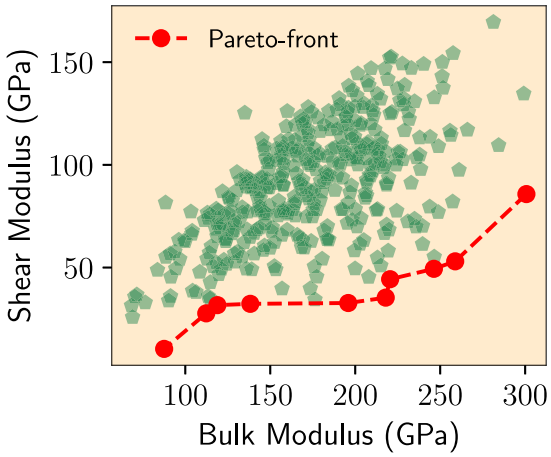

FIG. 10. The Pareto optimal points in the materials property space are marked in red corresponding to the criterion of maximizing bulk modulus and minimizing shear modulus simultaneously. The Pareto set for the MDS consist of ten points as indicated in red.

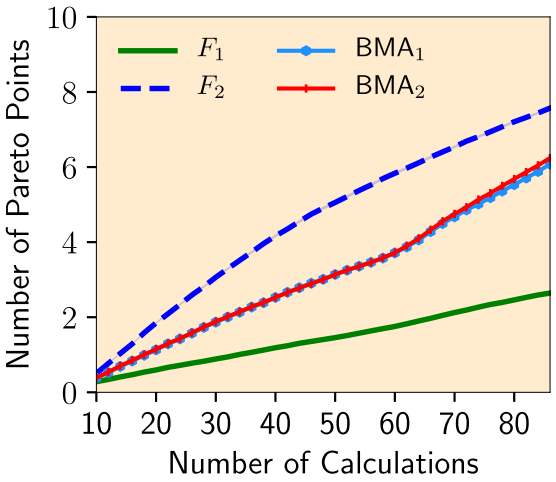

FIG. 11. Average number of true Pareto optimal points found over all initial data set instances for single models, BMA_1, and BMA_2 for $N = 10$.

feature set even using domain knowledge, one may or may not select a good model. However, if one were to use the BMA approach, either BMA_1 or BMA_2, the probability of successfully arriving at the material with desired properties, is very high, since the BMA approach autoselects the best feature set. Results for $N = 2, 5, 15, 20$ as well as the plots for BMA coefficients may be found in the Supplemental Material [80].

B. Multiobjective optimization: Maximize bulk modulus and minimize shear modulus

We now consider multiobjective experiment design to optimize two objectives at the same time: maximizing bulk modulus and minimizing shear modulus. One should note that in our analysis we have already calculated the responses of bulk and shear modulus as materials properties for all the feasible points in the MDS to have the ground truth to compare different models for experiment design. Generally in practice, no knowledge of the responses exists unless one performs all the possible experiments exhaustively. Consequently, none of this information is used in our experiment design procedures. Figure 10 illustrates all the data points in the objective space of materials properties (in green). It can be seen that in this case there does not exist a single optimal solution, and in fact there are ten *Pareto* optimal points comprising the *Pareto front* [81] which is highlighted in red in the figure. Specifically, the Pareto front here is the one-dimensional design curve over which any improvement in one material property (i.e., bulk modulus K) is only achieved through a corresponding sacrifice of another property (here, shear modulus G).

Figure 11 depicts the average performance of the best (F_2) and worst (F_1) models as well as the first- and second-order BMA in finding the true Pareto optimal points versus the number of calculations. Similar to single-objective problems, multiobjective experiment design based on F_2 consistently has the best performance, i.e., it identifies more true Pareto optimal points faster (with smaller budget). Both BMA approaches' performances are consistently in the range of the first best (F_2) single model's performances. Complete results for all cases of N, swarm plots and coefficient plots for the multiobjective scenario may be found in the Supplemental Material [80].

V. DISCUSSION

A. Comparison of first-order and second-order BMA

From the results in the previous sections, we can see that for single-objective experiment design, the performance of the first-order BMA is slightly better than the second-order BMA. On the other hand, the model probabilities in the second-order BMA are more robust, and at any calculation number (sequential experiment iteration), the average posterior probability over all the initial data set instances of the best model in terms of experiment design performance is higher than the other models. The reason is that second-order Laplace approximation, unlike the first-order one, does not rely solely on the fitted values of the parameters of the GPR model to calculate the model probability. In fact, it approximates the model probability by integrating a local expansion of the marginal likelihood over a neighborhood of the fitted parameters values, which may dampen the fluctuations of the fitted values between different sequential experiment iterations. For the multiobjective case, the second-order BMA is slightly better than first-order BMA in terms of both experiment design performance and robustness of identifying the best model in terms of experiment design performance.

B. Remarks on feature sets

The feature sets in our analysis are chosen *a priori* based on domain knowledge. We do not claim that the considered feature sets are among the best possible feature sets for our experiment design problems. We are rather using these to showcase the applicability of the BOMU framework in real-world experiment design problems, where the best model or feature set is often not known, and only a set of possible models might exist based on domain knowledge. The power of

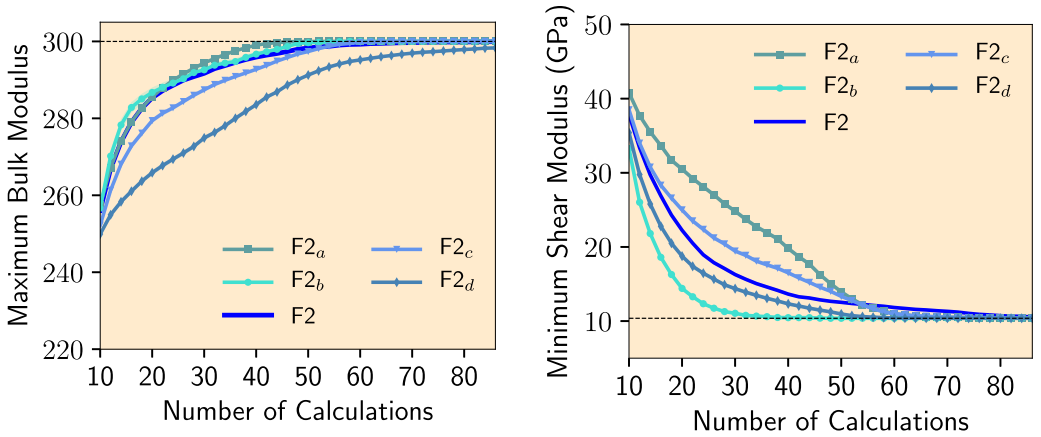

FIG. 12. Average maximum discovered (a) bulk modulus and (b) shear modulus for F_2 and lower-dimensional feature sets (F_{2a}, F_{2b}, F_{2c}, F_{2d}) derived from F_2.

BOMU is that it incorporates the uncertainty over the possible model space, instead of relying on a single model that is selected based on limited initial available data. For instance, we compared experiment design results based on the subsets of F_2 with one feature removed from F_2 (by taking three features at a time): feature set F_{2a}: $[m, Z, I_{\text{dist}}]$, feature set F_{2b}: $[m, Z, e_a]$, feature set F_{2c}: $[m, I_{\text{dist}}, e_a]$, and feature set F_{2d}: $[Z, I_{\text{dist}}, e_a]$.

Figures 12(a) and 12(b) show the corresponding results for maximizing bulk modulus and minimizing shear modulus problems, respectively. From both figures, there are some subsets that can perform better in terms of average optimal objective values discovered over all instances of initial data sets for a fixed initial data set size. Another observation from Figs. 12(a) and 12(b) is that adding noninformative features to a model (feature set) can degrade the experiment design performance, as there are the single models based on some subsets of cardinality three derived from F_2 that can find the optimal compound in the MDS faster than the experiment design based on F_2. One reason is that by adding noninformative features, more dimensions are introduced in the feature space while the information on these dimensions may be irrelevant to their outputs—it does not help better predict the outputs. This has more effect especially when using kernels with a single length-scale parameter, which is the most common practice in the materials literature. This is explicitly indicated in Sec. IV A 2, as the the BOMU approach excels when there are noninformative features, in that it autorejects feature sets $F_6, \ldots, F_{10}$, while converging to the target experiment as fast as the best standalone model F_2. Further discussion is included in Appendix A.

VI. CONCLUSIONS

The Bayesian optimization approach was successfully combined with Bayesian model averaging (BMA) for autonomous and adaptive learning to design a Bayesian experiment design framework under model uncertainty (BOMU) for materials discovery in single- and multiobjective material property space using a test set of MAX phases. It was demonstrated that, while prior knowledge about the fundamental features linking the material to the desired material property is certainly essential to build the materials design space (MDS), the BMA approach may be used to autoselect the *best* features/feature sets in the MDS, thereby eliminating the requirement of knowing the *best* feature set *a priori*. As evident from the extensive results included in the Supplemental Material [80], the BOMU framework is not significantly dependent on the size of the initial data, which enables its use in materials discovery problems where initial data is scant. At the very least, this framework provides a very efficient means of building the initial data set as well, since it may be used to guide experiments or calculations by focusing on gathering data in those sections of the MDS which will result in the most efficient path to achieving the optimal material.

ACKNOWLEDGMENTS

The authors acknowledge the support of NSF through the project DMREF: Accelerating the Development of Phase-Transforming Heterogeneous Materials: Application to High Temperature Shape Memory Alloys, NSF-CMMI-1534534. R.A. and E.D. also acknowledge the support of NSF through Grant No. NSF-DGE-1545403. T.D. acknowledges support of NSF-CMMI-1729335. X.Q. acknowledges the support of NSF through the project CAREER: Knowledge-driven Analytics, Model Uncertainty, and Experiment Design, NSF-CCF-1553281 and NSF-CISE-1835690 (with R.A.). A.T. and R.A. also acknowledge support by the Air Force Office of Scientific Research under AFOSR-FA9550-78816-1-0180 (Program Manager: Dr. Ali Sayir).

A.T. and S.B. contributed equally to this work.

APPENDIX A: IMPLEMENTATION REMARK

The estimation of the (hyper) parameters, including the length-scale parameter, of the GPR model are found by maximizing the marginal likelihood of the data, i.e., ML-II estimation instead of the fully Bayesian treatment. Marginal likelihood might have multiple optima that correspond to different interpretations of the data. When GPR models are trained based on ML-II estimation, depending on the MDS

and selected kernel functions, there is a possibility of overfitting the training data, especially when only a small number of measured data points are available (small-sample training data as initial data points). One thing to note is that experiment design based on GPR models that overfit the training data and assign very low correlation to nearby points in their prediction can yield very poor experiment design performance. One reason being that in this case measuring any point in the MDS will not give much information regarding other points of the MDS, because of the overfitting of the underlying learned surrogate GPR model. Since in our experiments the feature sets were chosen *a priori*, without any knowledge of their suitability for the underlying true model that generates data, in our implementation we have restricted the possible range for the length-scale parameter of the GPR kernel to prevent the models from overfitting the limited number of available data.

APPENDIX B: CONNECTIONS AND DIFFERENCES WITH GENERALIZED MOCU

We would like to close with some remarks concerning the manner in which the experiment design developed in this paper relates to the general theory. In the following we first provide a brief summary of the generalized MOCU introduced in Ref. [24]. Assuming a probability space $\mathcal{M}$ (uncertainty class) with probability measure P, an action space $\mathcal{X}$, and an objective function $f : \mathcal{M} \times \mathcal{X} \to (-\infty, \infty)$, our goal is to find an action $\mathbf{x} \in \mathcal{X}$ that minimizes the unknown true objective function $f(\mathbf{x}; M_t)$ over $\mathcal{X}$, where $M_t \in \mathcal{M}$. A *robust action* is an element $\mathbf{x}^R \in \mathcal{X}$ that minimizes the average of the objective function across all possibilities in the uncertainty class relative to a probability distribution governing the corresponding space. This probability at each time step is the posterior distribution given the observed data points available up to that time step ($\mathcal{D}_n$). Mathematically,

$$\mathbf{x}_n^R = \arg\min_{\mathbf{x}\in\mathcal{X}} E_M[f(\mathbf{x}; M)|\mathcal{D}_n]. \tag{B1}$$

The mean objective cost of uncertainty (MOCU) is the average gain in the attained objective between the robust action and the actual optimal actions across the possibilities:

$$\mathrm{MOCU}_n^{\mathcal{X}}(\mathcal{M}) = E_M\left[f\left(\mathbf{x}_n^R; M\right) - f(\mathbf{x}_M^*; M)|\mathcal{D}_n\right], \tag{B2}$$

where $\mathbf{x}_M^*$ denotes the optimal action for a given M. Note that if we actually knew the true (correct) model, then we would simply use the optimal action for that model and MOCU would be 0. Denoting the set of possible experiments by Ξ, the best experiment ξ_n^* at each time step (in one step look ahead scenario) is the one that maximally reduces the expected MOCU following the experiment, i.e.,

$$\xi_n^* = \arg\min_{\xi\in\Xi} E_\xi\left[E_M\left[f\left(\mathbf{x}_{n+1}^R; M\right)\middle|\xi, \mathcal{D}_n\right]\right] - E_M\left[f\left(\mathbf{x}_n^R; M\right)\middle|\mathcal{D}_n\right]. \tag{B3}$$

In most cases in the context of materials discovery, each experiment is applying an action and observing its cost (or a noisy version of it). Thus, the experiment space is equivalent to the action space.

It is beneficial to recognize that MOCU can be viewed as the minimum expected value of a Bayesian loss function, where the Bayesian loss function maps an action to its differential objective value (for using the given action instead of an optimal action), and its minimum expectation is attained by an optimal robust action that minimizes the average differential objective value. In decision theory, this differential objective value has been referred to as the *regret*.

In Sec. I B we mentioned three possibilities regarding the objective function. In the first case, we have a parametric model where the parameters come from an underlying physical system. An example in medicine is where they characterize a gene regulatory network, the objective function is the likelihood of the cell being in a cancerous state, and the action is to administer a drug [11]. Another example is in imaging where the parameters characterize the image structure, the objective function is an error measure between two images, and the action is to compress the image in order to reduce the number of bits while at the same time maintaining visual fidelity [82]. In this case the action space and experiment space are usually distinct sets.

Another possibility is that the features are known and the parameters come from a surrogate model used in place of the actual physical model, but believed to be appropriately related to the physical model. In the materials example [19] noted in Sec. I B, the surrogate model is based on the time-dependent Ginzburg-Landau (TDGL) theory and simulates the free energy given dopant parameters, the objective function is the energy dissipation, and the action is to find an optimal dopant and concentration. To see how the approach in Ref. [19] fits the above general theory the reader can refer to [24].

A third possibility is that we do not know the physical model and we lack sufficient knowledge to posit a surrogate model with known features/form relating to our objective. This case arises in many scenarios where the objective function is a blackbox function. Nevertheless, we can adopt a model, albeit, one with known predictive properties. This model can be a kernel-based model like a GP. Moreover, this model can consist of a set of possible parametric families, or a kernel-based model with different possible feature sets, or even kernel-based models with different choices for the kernel function. In this paper we have addressed this case when we do not *a priori* have any knowledge about which feature set or model family would be the best, and reliable model selection cannot be performed before starting the experiment design loop. Considering the average prediction from models based on different feature sets or model families weighted by their posterior probability of being the correct model, namely BMA, is one possible approach. In this paper we perform BMA based on possible feature sets that come from domain knowledge.

It is worth mentioning that, in theory, the generalized MOCU can be applied to all these scenarios with a single objective; however, there might be computational issues, especially in the third type of model. For example, when the experiments consist of running expensive simulation models, the computations of MOCU-based experiment design might be extremely heavy, so much so that the experiment design would be more computationally expensive and/or time consuming than the original simulation model.

A last question needs to be addressed. As noted previously, it is known that under certain conditions, MOCU-based experiment design is equivalent to EGO [24]. Could we have used MOCU here, and/or can the procedure proposed in this paper be related to MOCU? In our case, at each time step, after training the GPs based on the current and previous observations (finding the GP hyperparameters that maximize the marginal likelihood of the observed data), each GP provides a Gaussian distribution over the objective values of the actions. Averaging several GPs based on their posterior model probabilities is like mixing weighted Gaussian distributions over the objective value of each action. Based on the sum of weighted Gaussian distributions, the EI or EHVI is calculated for all possible remaining actions for single- or multiobjective scenarios, respectively, and the maximizer is chosen as the next experiment. For the multiobjective case, we cannot employ MOCU. The reason is that the current formulations of MOCU do not contain definitions suitable to multiobjective problems, e.g., no notion of robust action exists in the presence of Pareto optimal solutions. For the single-objective case, assuming the mixture of Gaussian distributions for the objective value of each action given at each time step, and confining the selection of the optimal action in the MOCU framework at each time step to the set of actions whose objective values have been previously observed, the maximizer of EI is equivalent to the solution of (B3). But in practice we have another layer of uncertainty introduced by the model fitting step. If we want to take this uncertainty into account when calculating the expected utility (acquisition value) at each time step, the procedure taken in this paper by employing EI is not equivalent to applying MOCU. To make it so we would have to assume a prior distribution over the hyperparameters of the GPs and when calculating the expected utility (acquisition value) of each potential next experiment at each time step, we would have to consider the corresponding possible updated distributions of the hyperparameters and consequent model probabilities posterior to carrying out the experiment and the possible objective value observation in the next time step. But this would be too computationally costly.

[1] J. P. Holdren *et al.*, *Materials genome initiative for global competitiveness, National Science and Technology Council OSTP* (Washington, DC, 2011).

[2] National Research Council *et al.*, *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security* (National Academies Press, Washington, DC, 2008).

[3] A. Agrawal and A. Choudhary, Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science, APL Mater. **4**, 053208 (2016).

[4] S. R. Kalidindi and M. De Graef, Materials data science: Current status and future outlook, Annu. Rev. Mater. Res. **45**, 171 (2015).

[5] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, and H. Lam, Combinatorial and high-throughput screening of materials libraries: Review of state of the art, ACS Comb. Sci. **13**, 579 (2011).

[6] S. K. Suram, J. A. Haber, J. Jin, and J. M. Gregoire, Generating information-rich high-throughput experimental materials genomes using functional clustering via multitree genetic programming and information theory, ACS Comb. Sci. **17**, 224 (2015).

[7] M. L. Green, C. L. Choi, J. R. Hattrick-Simpers, A. M. Joshi, I. Takeuchi, S. C. Barron, E. Campo, T. Chiang, S. Empedocles, J. M. Gregoire *et al.*, Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies, Appl. Phys. Rev. **4**, 011105 (2017).

[8] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, The high-throughput highway to computational materials design, Nat. Mater. **12**, 191 (2013).

[9] P. Voorhees, G. Spanos *et al.*, Modeling across scales: A roadmapping study for connecting materials models and simulations across length and time scales, Technical Report (The Minerals, Metals & Materials Society, Pittsburgh, PA, 2015).

[10] D. R. Jones, M. Schonlau, and W. J. Welch, Efficient global optimization of expensive black-box functions, J. Global Optim. **13**, 455 (1998).

[11] R. Dehghannasiri, B.-J. Yoon, and E. R. Dougherty, Optimal experimental design for gene regulatory networks in the presence of uncertainty, IEEE/ACM Trans. Comput. Biol. Bioinf. **12**, 938 (2015).

[12] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, Taking the human out of the loop: A review of Bayesian optimization, Proc. IEEE **104**, 148 (2016).

[13] S. Zamani Dadaneh and X. Qian, Bayesian module identification from multiple noisy networks, EURASIP J. Bioinf. Syst. Biol. **2016**, 5 (2016).

[14] S. Boluki, M. Shahrokh Esfahani, X. Qian, and E. R. Dougherty, Constructing pathway-based priors within a Gaussian mixture model for Bayesian regression and classification, IEEE/ACM Trans. Comput. Biol. Bioinf., doi: 10.1109/TCBB.2017.2778715.

[15] S. Boluki, M. S. Esfahani, X. Qian, and E. R. Dougherty, Incorporating biological prior knowledge for Bayesian learning via maximal knowledge-driven information priors, BMC Bioinf. **18**, 552 (2017).

[16] S. Z. Dadaneh, X. Qian, and M. Zhou, Bnp-seq: Bayesian nonparametric differential expression analysis of sequencing count data, J. Am. Stat. Assoc. **113**, 81 (2017).

[17] A. Karbalayghareh, U. Braga-Neto, and E. R. Dougherty, Intrinsically Bayesian robust classifier for single-cell gene expression trajectories in gene regulatory networks, BMC Syst. Biol. **12**, 23 (2018).

[18] A. Karbalayghareh, X. Qian, and E. R. Dougherty, Optimal Bayesian transfer learning, IEEE Trans. Signal Process. **66**, 3724 (2018).

[19] R. Dehghannasiri, D. Xue, P. V. Balachandran, M. R. Yousefi, L. A. Dalton, T. Lookman, and E. R. Dougherty, Optimal experimental design for materials discovery, Comput. Mater. Sci. **129**, 311 (2017).

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

[21] P. I. Frazier, W. B. Powell, and S. Dayanik, A knowledge-gradient policy for sequential information collection, SIAM J. Control Optim. **47**, 2410 (2008).

[22] P. I. Frazier, W. B. Powell, and S. Dayanik, The knowledge-gradient policy for correlated normal beliefs, INFORMS J. Comput. **21**, 599 (2009).

[23] B.-J. Yoon, X. Qian, and E. R. Dougherty, Quantifying the objective cost of uncertainty in complex dynamical systems, IEEE Trans. Signal Process. **61**, 2256 (2013).

[24] S. Boluki, X. Qian, and E. R. Dougherty, Experimental design via generalized mean objective cost of uncertainty, arXiv:1805.01143.

[25] M. Pelikan, D. E. Goldberg, and E. Cantú-Paz, Boa: The Bayesian optimization algorithm, in *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 1* (Morgan Kaufmann, Burlington, MA, 1999), pp. 525–532.

[26] K. Fujimura, A. Seko, Y. Koyama, A. Kuwabara, I. Kishida, K. Shitara, C. A. J. Fisher, H. Moriwake, and I. Tanaka, Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms, Adv. Energy Mater. **3**, 980 (2013).

[27] D. Basak, S. Pal, and D. C. Patranabis, Support vector regression, Neural Inform. Process. Lett. Rev. **11**, 203 (2007).

[28] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single-and binary-component solids, Phys. Rev. B **89**, 054303 (2014).

[29] P. V. Balachandran, D. Xue, J. Theiler, J. Hogden, and T. Lookman, Adaptive strategies for materials design using uncertainties, Sci. Rep. **6**, 19660 (2016).

[30] M. W. Barsoum, *MAX Phases: Properties of Machinable Vernary Carbides and Nitrides* (John Wiley & Sons, New York, 2013).

[31] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, Accelerated search for materials with targeted properties by adaptive design, Nat. Commun. **7**, 11241 (2016).

[32] S. Ju, T. Shiga, L. Feng, Z. Hou, K. Tsuda, and J. Shiomi, Designing Nanostructures for Phonon Transport Via Bayesian Uptimization, Phys. Rev. X **7**, 021024 (2017).

[33] P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto, and B. Maruyama, Autonomy in materials research: A case study in carbon nanotube growth, npj Comput. Mater. **2**, 16031 (2016).

[34] T. Lookman, P. V. Balachandran, D. Xue, J. Hogden, and J. Theiler, Statistical inference and adaptive design for materials discovery, Curr. Opin. Solid State Mater. Sci. **21**, 121 (2016).

[35] R. Jalem, K. Kanamori, I. Takeuchi, M. Nakayama, H. Yamasaki, and T. Saito, Bayesian-driven first-principles calculations for accelerating exploration of fast ion conductors for rechargeable battery application, Sci. Rep. **8**, 5845 (2018).

[36] S. Broderick and K. Rajan, Informatics derived materials databases for multifunctional properties, Sci. Technol. Adv. Mater. **16**, 013501 (2015).

[37] M. Botvinick, D. G. T. Barrett, P. Battaglia, N. de Freitas, D. Kumaran, J. Z. Leibo, T. Lillicrap, J. Modayil, S. Mohamed, N. C. Rabinowitz *et al.*, Building machines that learn and think for themselves, Behavioral Brain Sci. **40**, e255 (2017).

[38] A. M. Gopakumar, P. V. Balachandran, D. Xue, J. E. Gubernatis, and T. Lookman, Multi-objective optimization for materials discovery via adaptive design, Sci. Rep. **8**, 3738 (2018).

[39] C. Sima and E. R. Dougherty, What should be expected from feature selection in small-sample settings, Bioinformatics **22**, 2430 (2006).

[40] U. M. Braga-Neto and E. R. Dougherty, Is cross-validation valid for small-sample microarray classification? Bioinformatics **20**, 374 (2004).

[41] J. Hua, W. D. Tembe, and E. R. Dougherty, Performance of feature-selection methods in the classification of high-dimension data, Pattern Recognition **42**, 409 (2009).

[42] D. Madigan, A. E. Raftery, C. Volinsky, and J. Hoeting, Bayesian model averaging, in *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR* (AAAI, Palo Alto, CA, 1996), pp. 77–83.

[43] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, Bayesian model averaging: A tutorial, Stat. Sci. **14**, 382 (1999).

[44] L. Wasserman, Bayesian model selection and model averaging, J. Math. Psychol. **44**, 92 (2000).

[45] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)* (The MIT Press, Cambridge, 2005).

[46] M. T. M. Emmerich, A. H. Deutz, and J. W. Klinkenberg, Hypervolume-based expected improvement: Monotonicity properties and exact computation, in *Evolutionary Computation (CEC), 2011 IEEE Congress on* (IEEE, Piscataway, NJ, 2011), pp. 2147–2154.

[47] M. Radovic and M. W. Barsoum, MAX phases: Bridging the gap between metals and ceramics, Am. Ceram. Soc. Bull. **92**, 20 (2013).

[48] M. W. Barsoum, The $M_{N+1}AX_N$ phases: A new class of solids: Thermodynamically stable nanolaminates, Prog. Solid State Chem. **28**, 201 (2000).

[49] M. W. Barsoum and M. Radovic, Mechanical properties of the MAX phases, in *Encyclopedia of Materials: Science and Technology*, edited by K. H. J. Buschow, R. Cahn, M. Flemings, B. Ilschner, E. Kramer, S. Mahajan, and P. Veyssiere (Elsevier, Amsterdam, 2004), pp. 1–16.

[50] M. W. Barsoum and M. Radovic, Elastic and mechanical properties of the MAX phases, Annu. Rev. Mater. Res. **41**, 195 (2011).

[51] Z. M. Sun, Progress in research and development on MAX phases: A family of layered ternary compounds, Int. Mater. Rev. **56**, 143 (2011).

[52] M. W. Barsoum, The Mn+ 1 AXn Phases and their Properties, in *Ceramics Science and Technology: Volume 2: Materials and Properties*, edited by R. R. Riedel and I.-W. Chen (Wiley Online Library, 2010), pp. 299-347.

[53] R. Arróyave, A. Talapatra, T. Duong, W. Son, H. Gao, and M. Radovic, Does aluminum play well with others? Intrinsic Al-A alloying behavior in 211/312 MAX phases, Mater. Res. Lett. **5**, 170 (2017).

[54] A. Talapatra, T. Duong, W. Son, H. Gao, M. Radovic, and R. Arróyave, High-throughput combinatorial study of the effect of M site alloying on the solid solution behavior of M_2AlC MAX phases, Phys. Rev. B **94**, 104106 (2016).

[55] S. Aryal, R. Sakidja, M. W. Barsoum, and W.-Y. Ching, A genomic approach to the stability, elastic, and electronic properties of the MAX phases, Phys. Status Solidi B **251**, 1480 (2014).

[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, Scikit-learn: Machine learning in python, J. Machine Learning Res. **12**, 2825 (2011).

[57] Y. Makino and S. Miyake, Estimation of bulk moduli of compounds by empirical relations between bulk modulus and interatomic distance, J. Alloys Compd. **313**, 235 (2000).

[58] J. Karthikeyan, V. Kumar, and P. Murugan, The role of valence electron concentration in tuning the structure, stability, and electronic properties of $Mo_6S_{9-x}I_x$ Nanowires, J. Phys. Chem. C **119**, 13979 (2015).

[59] S. Guo, C. Ng, J. Lu, and C. T. Liu, Effect of valence electron concentration on stability of fcc or bcc phase in high entropy alloys, J. Appl. Phys. **109**, 103505 (2011).

[60] U. Mizutani and H. Sato, Determination of electrons per atom ratio for transition metal compounds studied by FLAPW-Fourier calculations, Philos. Mag. **96**, 3075 (2016).

[61] S. Pronk and D. Frenkel, Large Difference in the Elastic Properties of fcc and hcp Hard-Sphere Crystals, Phys. Rev. Lett. **90**, 255501 (2003).

[62] D. Tromans, Elastic anisotropy of hcp metal crystals and polycrystals, Int. J. Res. Rev. Appl. Sci **6**, 462 (2011).

[63] D. R. Lide *et al.*, *CRC Handbook of Chemistry and Physics* (CRC Boca Raton, 2012).

[64] K. Kandasamy, J. Schneider, and B. Póczos, High dimensional Bayesian optimisation and bandits via additive models, in *International Conference on Machine Learning* (MLR Press, 2015), pp. 295–304.

[65] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression* (Springer Science & Business Media, New York, 2006).

[66] A. D. Bull, Convergence rates of efficient global optimization algorithms, J. Machine Learning Res. **12**, 2879 (2011).

[67] Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. de Feitas, Bayesian optimization in a billion dimensions via random embeddings, J. Artif. Intell. Res. **55**, 361 (2016).

[68] W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev. **140**, A1133 (1965).

[69] G. Kresse and J. Furthmüller, Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set, Phys. Rev. B **54**, 11169 (1996).

[70] G. Kresse and J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. **6**, 15 (1996).

[71] J. P. Perdew and Y. Wang, Accurate and simple analytic representation of the electron-gas correlation energy, Phys. Rev. B **45**, 13244 (1992).

[72] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. **77**, 3865 (1996).

[73] H. J. Monkhorst and J. D. Pack, Special points for Brillouin-zone integrations, Phys. Rev. B **13**, 5188 (1976).

[74] M. P. A. T. Methfessel and A. T. Paxton, High-precision sampling for Brillouin-zone integration in metals, Phys. Rev. B **40**, 3616 (1989).

[75] P. E. Blöchl, O. Jepsen, and O. K. Andersen, Improved tetrahedron method for Brillouin-zone integrations, Phys. Rev. B **49**, 16223 (1994).

[76] Y. Le Page and P. Saxe, Symmetry-general least-squares extraction of elastic data for strained materials from *ab initio* calculations of stress, Phys. Rev. B **65**, 104104 (2002).

[77] N. Singh, A. Talapatra, A. Junkaew, T. Duong, S. Gibbons, S. Li, H. Thawabi, E. Olivos, and R. Arróyave, Effect of ternary additions to structural properties of NiTi alloys, Comput. Mater. Sci. **112**, 347 (2016).

[78] T. Duong, S. Gibbons, R. Kinra, and R. Arróyave, *Ab-initio* approach to the electronic, structural, elastic, and finite-temperature thermodynamic properties of Ti_2AX (A=Al or Ga and X=C or N), J. Appl. Phys. **110**, 093504 (2011).

[79] R. Hill, The elastic behaviour of a crystalline aggregate, Proc. Phys. Soc. Sect. A **65**, 349 (1952).

[80] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevMaterials.2.113803 for details on features used and results for cases other than those presented in the paper and model fidelity discussion.

[81] P. Sirisalee, M. F. Ashby, G. T. Parks, and P. J. Clarkson, Multi-criteria material selection in engineering design, Adv. Eng. Mater. **6**, 84 (2004).

[82] R. Dehghannasiri, X. Qian, and E. R. Dougherty, Intrinsically Bayesian robust Karhunen-Loève compression, Signal Process. **144**, 311 (2018).