npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Bayesian optimization with adaptive surrogate models for automated experimental design

Bowen Lei[1], Tanner Quinn Kirk[2], Anirban Bhattacharya[1], Debdeep Pati[1], Xiaoning Qian[3,4], Raymundo Arroyave[5]✉ and Bani K. Mallick[1]

Bayesian optimization (BO) is an indispensable tool to optimize objective functions that either do not have known functional forms or are expensive to evaluate. Currently, optimal experimental design is always conducted within the workflow of BO leading to more efficient exploration of the design space compared to traditional strategies. This can have a significant impact on modern scientific discovery, in particular autonomous materials discovery, which can be viewed as an optimization problem aimed at looking for the maximum (or minimum) point for the desired materials properties. The performance of BO-based experimental design depends not only on the adopted acquisition function but also on the surrogate models that help to approximate underlying objective functions. In this paper, we propose a fully autonomous experimental design framework that uses more adaptive and flexible Bayesian surrogate models in a BO procedure, namely Bayesian multivariate adaptive regression splines and Bayesian additive regression trees. They can overcome the weaknesses of widely used Gaussian process-based methods when faced with relatively high-dimensional design space or non-smooth patterns of objective functions. Both simulation studies and real-world materials science case studies demonstrate their enhanced search efficiency and robustness.

npj Computational Materials (2021)7:194 ; https://doi.org/10.1038/s41524-021-00662-x

INTRODUCTION

The concept of optimal experimental design, within the overall framework of Bayesian optimization (BO), has been put forward as a design strategy to circumvent the limitations of traditional (costly) exploration of (arbitrary) design spaces. BO utilizes a flexible surrogate model to stochastically approximate the (generally) expensive objective function. This surrogate, in turn, undergoes Bayesian updates as new information about the design space is acquired, according to a predefined acquisition policy. The use of a Bayesian surrogate model does not impose any a priori restrictions (such as concavity or convexity) on the objective function. It was mainly introduced by Mockus[1] and Kushner[2] and pioneered by Jones et al.[3], who developed a framework that balanced the need to exploit available knowledge of the design space with the objective to explore it by using a metric or policy that selects the best next experiment to carry out with the end-goal of accelerating the iterative design process. Multiple extensions have been developed to make the algorithm more efficient[4–7]. This popular tool has been successfully used in a wide range of applications[8,9]. Extensive surveys of this method and its applications can also be found[10–12].

Materials discovery (MD) can be mapped to an optimization problem in which the goal is to maximize or minimize some desired properties of a material by varying certain features/structural motifs that are ultimately controlled by changing the overall chemistry and processing conditions. A typical task in MD is to predict the material properties based on a collection of features and then use such predictions in an inverse manner to identify the specific set of features leading to a desired, optimal performance. The major goal is then to identify how to search the complex material space spanning the elements in the periodic table, arranged in a virtually infinite number of possible configurations and microstructures, as generated by arbitrary synthesis/processing methods, to meet the target properties. Recently, a design paradigm has been proposed—optimal experimental design—built upon the foundation of BO[13–18], which seeks to circumvent the limits of traditional (costly) exploration of the materials design space. Early examples were demonstrated by Frazier and Wang[19], who took into account both the need to harness the knowledge that exists about the design space and the goal of exploring and identifying the best experiment to speed up the iterative design process. The other important task, other than discovering the target position in the space, is the identification of the key factors responsible for most of the variance in the properties of interest during MD[20–22]. This helps us better understand the underlying physical/chemical mechanisms controlling the properties or phenomena of interest, which in turn results in better strategies for MD and design[17,23]. There have been several follow-up papers, mainly extending the algorithm in different applied directions[13–15].

The BO algorithm consists of two major components[10,12]: (i) modeling a (potentially) high-dimensional black-box function, f, as a surrogate of the (expensive-to-query) objective function, and (ii) optimizing the selected criterion considering uncertainty based on the posterior distribution of f to obtain the design points in the feature space Ω. In the procedure, we repeat the two steps until we satisfy the stopping criteria or, as it is often the case in experimental settings, we exhaust the resources available. A critical aspect of BO is the choice of the probabilistic surrogate model used to fit f. A Gaussian process (GP) is the typical choice, as it is a powerful stochastic interpolation method that is distinguished from others by its mathematical explicitness and computational flexibility, and with straightforward uncertainty

[1]Department of Statistics, Texas A&M University, College Station, TX, USA. [2]Department of Mechanical Engineering, Texas A&M University, College Station, TX, USA. [3]Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX, USA. [4]Department of Computer Science & Engineering, Texas A&M University, College Station, TX, USA. [5]Department of Materials Science & Engineering, Texas A&M University, College Station, TX, USA. ✉email: raymundo.arroyave@tamu.edu

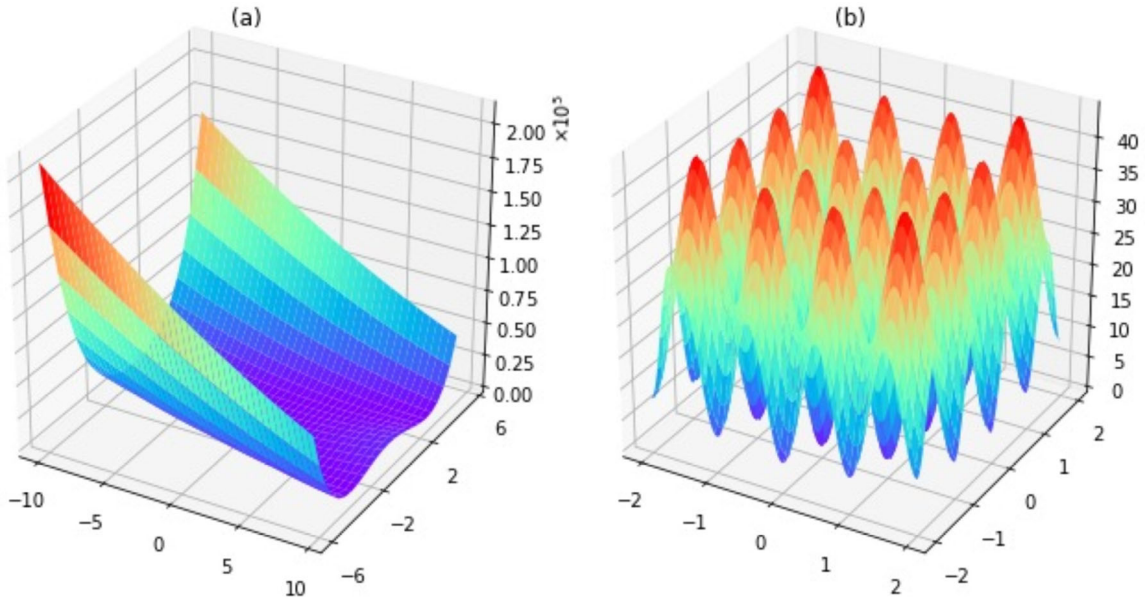

Fig. 1 Plots of black-box functions. a The valley of a two-dimensional Rosenbrock function which has the formula $\mathbf{y} = 100(\mathbf{x}_2 - x_1^2)^2 + (\mathbf{x}_1 - 1)^2$. **b** The frequent and regularly distributed local minima of a two-dimensional Rastrigin function which has the formula $\mathbf{y} = 20 + \sum_{i=1}^{2}[\mathbf{x}_i^2 - 10\cos(2\pi\mathbf{x}_i)]$.

quantification, which makes it broadly applicable to many problems[12,13,24].

Oftentimes, stationary or isotropic GP-based BO may be challenging when faced with (even moderately) high-dimensional design spaces, particularly when very little initial information about the design space is available—in some fields of science and engineering where BO is being used[14,25,26], data sparsity is, in fact, the norm, rather than the exception. In MD problems, data sparsity is exacerbated by the (apparent) high dimensionality of the design space, as a priori, it is possible that many controllable features could be responsible for the materials' behavior/property of interest. In practice, however, the potentially high-dimensional MD space may actually be reduced, as in materials science it is often the case that a small subset of all available degrees of freedom is actually controlling the materials' behavior of interest. Searching over a large dimensional space when only a small subspace is of interest may be highly computationally inefficient. A challenge is then how to discover the dominant degrees of freedom when very little data is available and no proper feature selection can be carried out at the outset of the discovery process. The problem may become more complex due to the existence of interaction effects among the covariates since such interactions are extremely challenging to discover when the available data is very sparse.

We note that there are some more flexible GP-based models, like automatic relevance detection (ARD),[27] which introduces a different scale parameter for each input variable inside the covariance function to facilitate removal of unimportant variables and may alleviate the problem. Recently, Talapatra et al.[14] proposed a robust model for *f*, based on Gaussian mixtures and Bayesian model averaging, as a strategy to deal with the data dimension and sparsity problem. Their framework was capable of detecting subspaces most correlated with optimization objectives by evaluating the Bayesian evidence of competing feature subsets. However, their covariance functions, and in general, most commonly used covariance functions for GP usually induce smoothness property and assume continuity for *f*, which may not necessarily be warranted and limit its performance when *f* is non-smooth or has sudden transitions—this may be a common occurrence in many MD challenges. Also, GP-based methods may still not perform well when the dimension of predictors is relatively high or the choice of the kernel is not suitable for the unknown function[10,12]. Apart from these solutions, there is a broad literature on flexible nonstationary covariance kernels[28]. Deep network kernel is a prominent recent example[29] while its strength may be limited when faced with sparse datasets.

The focus of this paper is to replace GP-based machine learning models with other, potentially more adaptive and flexible, Bayesian models. More specifically, we explore Bayesian spline-based models and Bayesian ensemble-learning methods as surrogate models in a BO setting. Bayesian multivariate adaptive regression splines (BMARS)[30,31] and Bayesian additive regression trees (BART)[32] are used in this paper as they can potentially be superior alternatives to GP-based surrogates, particularly when the objective function, *f*, requires more flexible models. BMARS is a flexible nonparametric approach based on product spline basis functions. BART belongs to Bayesian ensemble-learning-based methods and fits unknown patterns through a sum of small trees. Both of them are well equipped with automatic feature selection techniques.

In this article, we present a fully automated experimental design framework that adopts BART and BMARS as the surrogate models used to predict the outcome(s) of yet-to-be-made observations/queries of/to the expensive "black-box" function. The surrogates are used to evaluate the acquisition policy within the context of BO. Automated algorithm-based experimental design is a growing technology used in many fields such as materials informatics, and biosystems design[22,33,34]. It combines the principles of specific domains with the use of machine learning to accelerate scientific discovery. We compare the performance of this BO approach using non-GP surrogate models against other GP-based BO methods using standard analytic functions, and then present results in which the framework has been applied to realistic materials science discovery problems. We then discuss the possible underlying reasons for the remarkable improvements in performance associated with using more flexible surrogate models, particularly when the (unknown) objective function is very complex and does not follow the underlying assumptions motivating the use of GPs as surrogates.

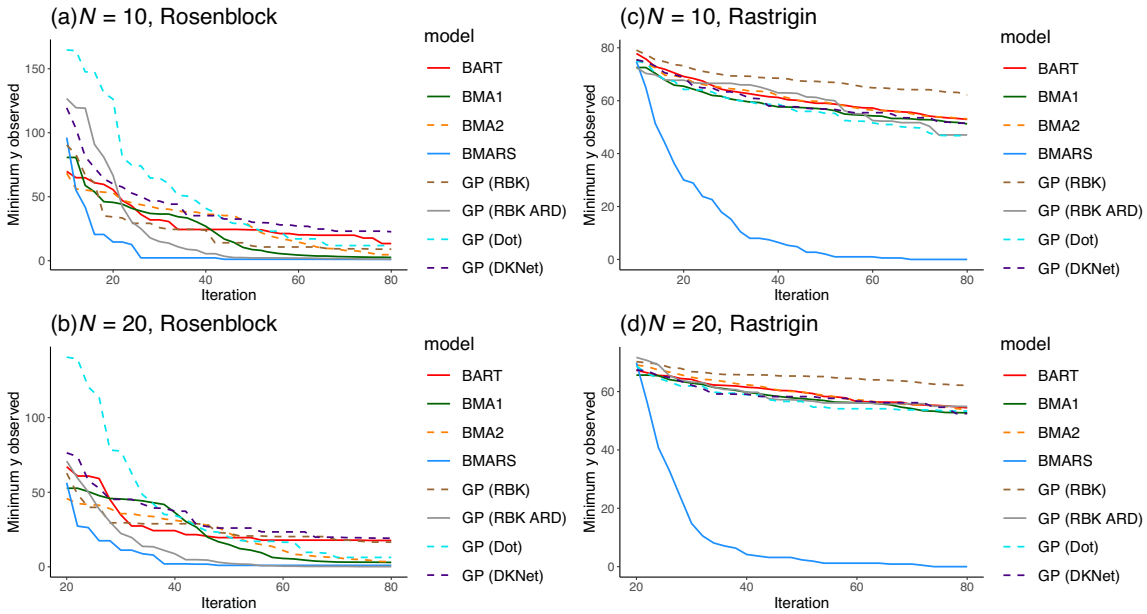

Fig. 2 The average minimum *y* observed based on each model in each iteration. a Rosenbrock function with the initial set of sample size $N = 10$. **b** Rosenbrock function with the initial set of sample size $N = 20$. **c** Rastrigin function with the initial set of sample size $N = 10$. **d** Rastrigin function with the initial set of sample size $N = 20$.

RESULTS

Simulation studies

In this section, we present two simulation studies where we set the Rosenbrock[35] and the Rastrigin[36,37] functions as the black-box function(s) to optimize, respectively. Figure 1 shows a two-dimensional example for each of them. They are two commonly used test functions in optimization benchmark studies. In both optimization tasks, the goal is to find the global minimum point of the unknown function. Therefore, we record the minimum value of observed response **y** in each iteration for each model for comparison. As shown in Fig. 2, a faster decline of the curve indicates a more efficient search for the target point and better performance.

As for the probabilistic models, we compare our proposal, which uses BART[32] and BMARS[30], with the popularly used baseline GP regression with Radial-basis function (GP RBF) kernel[38]. At the same time, RBK kernel with ARD (GP RBK ARD)[27] is considered, as well as nonstationary kernels like the dot-product kernel (GP Dot)[39] and more flexible deep network kernel (GP DKNet)[29]. We also compare them with the Bayesian model average using GP (BMA1 and BMA2)[14], which showed an edge over the benchmark method. BMA1 and BMA2 refer to the use of first- or second-order Laplace approximation to calculate the relevant marginal probabilities of the mixture model. We use a constant mean function for all the GP-based modes above. For the acquisition function, we choose the expected improvement (EI) metric[3] for all the models. To ensure a fair comparison, we also use a random search within the inner optimization problem of the acquisition function.

In order to have a comprehensive performance evaluation, we begin the optimization of the above models with five different sizes of initial datasets ($N = 2, 5, 10, 15, 20$) that are uniformly sampled from the search space[14]. As for each N and each algorithm, the results are based on 100 replicates. To reduce the number of iterations, we choose two samples each time in the workflow. For the stopping criteria, it is regarded as running out of the budget which is set as 80 function evaluations. Relevant results for $N = 10, 20$ are depicted in Fig. 2 and those for other N values can be referred to in Supplementary Note 1.

The Rosenbrock function, also called the Valley or Banana function, is often used as a test case for optimization algorithms[36,40]. The formula of a d-dimensional Rosenbrock function is as follows:

$$f(\mathbf{x}) = \sum_{i=1}^{d-1} [100(\mathbf{x}_{i+1} - \mathbf{x}_i^2)^2 + (\mathbf{x}_i - 1)^2]. \quad (1)$$

This function is unimodal, with the global minimum being at $\mathbf{x}^* = (1, \ldots, 1)$ with $f(\mathbf{x}^*) = 0$, which lays inside a long, narrow, parabolic-shaped flat valley as shown in Fig. 1a. In this function, we have a continuous search space and for each x_i the range is $[-2, 2]$. During the workflow, locating the valley is trivial. However, further convergence to the global optimum is difficult, making this a good test problem.

Here, we set $d = 4$ and simulate the data. Apart from these four important predictors, we also add four uninformative predictors that follow the standard normal distribution. These four new-added features do not affect the response but are designed to augment the dimensionality of the problem, potentially obfuscating the solution and "frustrating" the optimizers. This enables us to check whether these frameworks can reveal true factors properly and lead to an efficient exploration. Otherwise, a lot of unnecessary searches occurring among the insignificant directions can slow down the process of locating the optimum point. Moreover, the quality of the predictions may suffer as a result.

As seen in Fig. 2a, b, the solid blue curves for BMARS show the sharpest decrease, suggesting that it is the most efficient optimizer. BART-based BO (solid red curve) also exhibits competitive performance, relative to popular GP-based techniques like GP (RBK ARD) (solid gray curve), BMA1 (solid dark green curve), and BMA2 (dotted orange curve). Meanwhile, GP (DKNet) (dotted purple curve) cannot show its strength and drops slowly as it requires considerably more data to be trained properly. As seen from (1), the Rosenbrock function overall is a polynomial function with a good smoothness property, which may explain why GP-based surrogates perform competitively despite not being the best. Turning to the final stable stage, it is the blue curve (BMARS) that firstly shows a flat pattern and is closest to the optimum value $f(\mathbf{x}^*) = 0$.

Table 1. The mean value and interquartile range (IQR) of the number of experiments based on each model to find the maximum bulk modulus K in MAX phases with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$.

Model	$N=2$ (Mean)	$N=5$ (Mean)	$N=10$ (Mean)	$N=15$ (Mean)	$N=20$ (Mean)	$N=2$ (IQR)	$N=5$ (IQR)	$N=10$ (IQR)	$N=15$ (IQR)	$N=20$ (IQR)
BART	**36.18**	**39.33**	**32.76**	**38.67**	**43.42**	21	31	22	20.5	18.5
BMA1	55.2	51.66	67.02	59.12	65.78	78	83.5	63	67.5	62
BMA2	57.36	55.48	60.42	66.59	71.7	76.5	77	74	60	52.5
BMARS	**36.82**	**34.94**	41.5	48.94	50.9	24	22	26	24	22.5
GP (RBK)	77.46	76.95	76.86	73.04	75.3	49	59	61	69	62
GP (RBK ARD)	73.3	50.2	**35.5**	**40.05**	**39.8**	67.5	31	13.5	8.5	7.5
GP (Dot)	45.5	45.25	63.1	57.2	65.8	29.5	24.	21.	32.5	17
GP (DKNet)	91.6	78.35	62.9	62.55	81.6	15	43.5	69.5	71	50

The bold values represent the top two models in terms of search performance.

Turning to the Rastrigin function[36,40], it is a nonconvex function used to measure the performance of optimization workflows. The formula for a d-dimensional Rastrigin function reads as follows:

$$f(\mathbf{x}) = 10d + \sum_{i=1}^{d}[\mathbf{x}_i^2 - 10\cos(2\pi\mathbf{x}_i)]. \quad (2)$$

It is based on a quadratic function with an addition of cosine modulation which brings about frequent and regularly distributed local minima as depicted in Fig. 1b. Similar to Rosenbrock's case, the search space is continuous and we focus on $[-2, 2]$ for each direction. Thus, the test function is highly multimodal, making it a challenging task where algorithms easily get stuck in local minima. The global minimum point is $\mathbf{x}^* = (0, \ldots, 0)$ and $f(\mathbf{x}^*) = 0$.

For the simulated data, we set $d = 10$ and again we add five uninformative features following a standard normal distribution. With these five additional variables (or design degrees of freedom), we can assess whether these frameworks are capable of detecting the factors that are truly correlated with the objective function, enabling an efficient exploration of the design space.

As seen in Fig. 2c, d, the solid blue curves for BMARS again exhibit the fastest decline, indicating the best performance. The BART-based BO (solid red curves) follows and presents similar decreased speed with most of the GP-based methods. However, the dotted brown curve seems to be the slowest, which is for the baseline GP (RBK). Considering the convergent stage, the blue curve reaches it between 50 and 60 iterations and the minimum observed y is very close to the global optimum value $f(\mathbf{x}^*) = 0$. The other methods remain in a decreasing pattern with larger values of the minimum observed y. It is no surprise that GP-based methods suffer under this scenario, for which Rastrigin function's quick switch between different local minima may be the reason, especially for GP (RBK). In contrast, with the flexible bases constructed and multiple tree models, BMARS and BART are able to capture this complex trend of f. We note that BART might need a few more training samples to gain more competitive advantages over more flexible GPs like BMA1 and BMA2 due to block patterns of Rastrigin function.

Having established the better overall performance of our proposed non-GP base functions applied to complex BO problems, we will now turn our attention to two materials science-motivated problems.

MD in the MAX phase space

MAX phases (ternary layered carbides/nitrides)[14,41] create an adequate system to investigate the behavior of autonomous materials design frameworks, as a result of both their chemical richness and the wide range of their properties. The pure ternary MAX phase composition palette has so far been explored to a limited degree, so there is also significant potential to reveal promising chemistries with optimal property sets[14,42,43]. For these reasons, we compared different algorithms for searching among the $M_{n+1}AX_n$ phases, where M refers to a transition metal, A refers to group IV and VA elements, and X corresponds to carbon or nitrogen.

Specifically, the materials design space for this work consists of the conventional MAX phases M_2AX and M_3AX_2, where M ∈ {Sc, Ti, V, Cr, Zr, Nb, Mo, Hf, Ti}, A ∈ {Al, Si, P, S, Ga, Ge, As, Cd, In, Sn, Tl, Pd}, and X ∈ {C, N}. The space is discrete which includes 403 stable MAX phases in total, aligned with Talapatra et al.[14]. More discussion about the discrete space in BO can be found in Supplementary Note 4. The goal of the automated algorithm is to provide a fast exploration of the material space, namely to find the most appropriate material design, which is either (i) the maximum bulk modulus K or (ii) the minimum shear modulus G. The results in the following sections are obtained with the aim (i), while those for (ii) can be found in Supplementary Note 2. We point out that while the material design space is small, knowledge of the ground truth can assist significantly in the verification of the solutions arrived at by different optimization algorithms.

For the predictors, we follow the setting in Talapatra et al.[14] and consider 13 possible features in the model: empirical constants C, m, which link the elements of the material to its bulk modulus; valence electron concentration C_v; electron to atom ratio $\frac{e}{a}$; lattice parameters a and c; atomic number Z; interatomic distance l_{dist}; the groups corresponding to the periodic table of the M, A, and X elements Col_M, Col_A, Col_X, respectively; the order O of MAX phase (whether of order 1 according to M_2AX or order 2 according to M_3AX_2); and the atomic packing factor (APF). We note that the features above can potentially be correlated with the intrinsic mechanical properties of MAX phases, although a priori we assume that we have no knowledge as to how such features are correlated. In practice, as was found in ref. [14], only a small subset of the feature space is correlated with the target properties. We note that in ref. [14] the motivation for using Bayesian model averaging was precisely to be able to detect subsets within the larger feature set most effectively correlated with the target properties to optimize.

For the probabilistic model, we align with the simulation study above and compare our suggested framework that uses BART[32] and BMARS[30] to the widely used baselines, including GP (RBK)[38], GP (RBK ARD)[27], GP (Dot)[39], Bayesian model average using GP (BMA1 and BMA2)[14], and GP (DKNet)[29]. For the acquisition function, we choose EI for each of them to ensure a fair comparison. To get a comprehensive picture, we follow the structure in the previous section (where we studied the

Table 2. The top 5 important factors selected by BART for the maximum bulk modulus K in MAX phases with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$.

Setting	Top 1	Top 2	Top 3	Top 4	Top 5
$N = 2$	Col_A	$\frac{e}{a}$	APF	Col_M	c
$N = 5$	Col_A	$\frac{e}{a}$	Col_M	APF	l_{dist}
$N = 10$	Col_A	APF	Col_M	$\frac{e}{a}$	l_{dist}
$N = 15$	Col_A	$\frac{e}{a}$	APF	Col_M	l_{dist}
$N = 20$	Col_A	APF	$\frac{e}{a}$	Col_M	l_{dist}

Table 3. The top 5 important factors selected by BMARS for the maximum bulk modulus K in MAX phases with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$.

Setting	Top 1	Top 2	Top 3	Top 4	Top 5
$N = 2$	$\frac{e}{a}$	Col_A	APF	l_{dist}	c
$N = 5$	Col_A	l_{dist}	APF	$\frac{e}{a}$	c
$N = 10$	Col_A	$\frac{e}{a}$	APF	l_{dist}	a
$N = 15$	Col_A	$\frac{e}{a}$	APF	l_{dist}	c
$N = 20$	Col_A	APF	$\frac{e}{a}$	l_{dist}	a

benchmark Rosenbrock and Rastrigin functions) and start the above models with five different sizes of initial samples ($N = 2, 5, 10, 15, 20$), which are randomly chosen from the design space. For each N, the results are based on 100 replicates. To avoid an excessive number of iterations, we add two materials at a time in the platform. For the stopping criteria, it is set as successfully locating the material with ideal properties or running out of the budget which is set as 80, roughly 20% of the available space. For these replicates not converging within the budget, we follow Talapatra et al.[14] and regard their number of calculations as 100 to avoid an excessive number of evaluations.

Due to the high cost per experiment, the framework has better performance if it needs a fewer number of experiments before finding the candidate with desired properties. Therefore, we use it as a vital criterion for evaluating model capabilities. Table 1 shows the mean value and interquartile range (IQR) of the total number of evaluations searching for the maximum bulk modulus K within the MAX phase design space. The smaller values of the mean and IQR indicate a more efficient and stable platform.

As depicted in Table 1, while GP (Dot), BMA1, and BMA2 are more efficient than GP (RBK) and GP (DKNet) when looking for the maximum bulk modulus K, BART and BMARS can further greatly reduce the number of experiments and maintain a more stable performance compared to GP-based models. For GP (RBK ARD), it achieves good speed when N is larger than 10, but shows poor and unstable performance under small N. Also, considering the interquartile range of each model, BART and BMARS tend to be more robust under each setting and can achieve the goal before 80 iterations, while the other five are more likely to run out of the budget without achieving the objective.

Two possible reasons could explain why BMARS and BART can improve the searching speed much more efficiently than competing strategies. On the one hand, BMARS and BART are known to be more flexible surrogates compared to GP-based methods and are more powerful when faced with unknown and complex mechanisms in real-world data. On the other hand, BMARS and BART usually scale better with the dimension and can be more robust when handling high or even moderately dimensional design spaces.

In MD problems, beyond the identification of optimal regions in the materials design space, it is also desirable to understand the factors/features most correlated with the properties of interest. By taking these predictor and interaction rankings into account, researchers can gain a deeper understanding of the connection between features and material properties. We present the relevant results for the maximum bulk modulus K, and those for the minimum shear modulus G are in Supplementary Note 2. BART and BMARS are endowed with automatic feature selection based on their appearances in the corresponding surrogate models, while baselines GP (RBK) and GP (Dot) cannot identify feature importance relative to the BO objective. Although BMA1 and BMA2 can utilize the coefficient of each component to provide some information about feature importance, they cannot directly tell the exact order of individual variables and interactions.

Under five different scenarios $N \in \{2, 5, 10, 15, 20\}$, Tables 2 and 3 list the top 5 important factors aimed at the maximum value of K using BART and BMARS, respectively. The rankings are based on the median inclusion times of the 100 replicates from the last model when the workflow stops. When using BART, Col_A, $\frac{e}{a}$, Col_M, APF, and l_{dist} are the most useful. While turning to BMARS, Col_A, $\frac{e}{a}$, c, a, APF, and l_{dist} always play a key role. We can see a similar pattern for the top-ranked features between the two models for different N, although some differences exist in their order. Regarding the interactions among features, we measure their importance by counting the coexistence of two of them within each basis function. The more frequently they are used in the same basic function, the greater their influences on material improvement are. The detailed results for the interaction selection can be referred in Supplementary Note 2.

During the material development process, we may not know which features we should add to the model in advance. In light of this, it is usually the case that one considers all possible features during the training *and optimization* to avoid missing important features. This brings an important challenge because it is often not possible to carry out any sort of feature selection ahead of the experimental campaign. Moreover, GP-based BO frameworks tend to become less efficient as the dimension of the design space increases as the required coverage to ensure adequate learning of the response surface is exponential with the number of features[11]. Moreover, the sparse nature of the sampling scheme—BO, after all, is used when there are stringent resource constraints to query the problem space—makes the (learned) response surface very flat over wide regions of the design space, with some interspersed, local highly nonconvex landscapes[44]. These issues make high-dimensional BO very hard. In materials science problems, a key challenge is that many of the potential dimensions of the problem space are uninformative, i.e., they are not correlated with the objective of interest.

It is thus desirable to develop frameworks that are robust against the existence of possibly many uninformative or redundant features. To further check the platform's utility to distill useful information and maintain the speed, we simulate 16 random predictors following the standard normal distribution and mix them with the 13 predictors described above. With these new non-informative features, we use the same automated framework and explore the space for the materials with ideal properties.

As shown from Table 4, BART's performance is not degraded by the newly added unhelpful information and is still the most efficient choice, indicating its robust property. At the same time, although BMARS is slower than the best, it is still competitive compared to other GP-based approaches like BMA1 and BMA2. BART-based BO is clearly capable of detecting non-informative features in a very effective manner.

We also find the top 5 features as well as interaction effects for both BART and BMARS. For the 16 newly added unimportant features, we denote them by $n_1, \ldots, n_{16}$. Tables 5 and 6 summarize the most significant features. We can see that the results do not include $n_1, \ldots, n_{16}$, indicating a good ability to filter out useless information. Compared with Table 2, we can also notice that the

Table 4. The mean value and interquartile range of the number of experiments based on each model to find the maximum bulk modulus K in MAX phases with additional non-informative features with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$.

Model	$N=2$ (Mean)	$N=5$ (Mean)	$N=10$ (Mean)	$N=15$ (Mean)	$N=20$ (Mean)	$N=2$ (IQR)	$N=5$ (IQR)	$N=10$ (IQR)	$N=15$ (IQR)	$N=20$ (IQR)
BART	**32.04**	**28.96**	**34.22**	**34.7**	**39.24**	24.5	22	24	26	22.5
BMA1	61.48	58.48	69.62	70.6	75.06	64	72	55	53.5	46
BMA2	63.02	62.94	68.9	67	74.94	80	75	64	63.5	56
BMARS	63.8	62.35	67.52	66.77	**70.6**	58.5	34.5	56	59	54
GP (RBK)	63.7	67.29	66.7	70.36	72	70.5	65.5	64.5	55.5	48.5
GP (RBK ARD)	73.7	81.25	82.8	74.5	78.3	46.	26.5	41.5	45.5	32
GP (Dot)	**58.1**	**48.45**	**57.85**	**58.75**	72.7	13.	28.5	12.5	14.5	36.5
GP (DKNet)	81.1	81.2	95.	94.65	92.5	36.5	47.5	25.5	18	20.5

The bold values correspond to the top two models in terms of search performance.

Table 5. The top 5 important factors selected by BART for the maximum bulk modulus K in MAX phases with additional non-informative features with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$.

Setting	Top 1	Top 2	Top 3	Top 4	Top 5
$N=2$	APF	Col_A	Col_M	l_{dist}	$\frac{e}{a}$
$N=5$	APF	Col_A	l_{dist}	Col_M	$\frac{e}{a}$
$N=10$	APF	Col_M	Col_A	l_{dist}	$\frac{e}{a}$
$N=15$	APF	Col_A	l_{dist}	Col_M	$\frac{e}{a}$
$N=20$	APF	Col_A	Col_M	l_{dist}	$\frac{e}{a}$

Table 6. The top 5 important factors selected by BART for the maximum bulk modulus K in MAX phases with additional non-informative features with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$ using BMARS.

Setting	Top 1	Top 2	Top 3	Top 4	Top 5
$N=2$	Col_A	$\frac{e}{a}$	APF	Col_M	l_{dist}
$N=5$	Col_A	APF	Col_M	l_{dist}	$\frac{e}{a}$
$N=10$	Col_A	$\frac{e}{a}$	Col_M	AP	l_{dist}
$N=15$	Col_A	$\frac{e}{a}$	APF	l_{dist}	Col_M
$N=20$	Col_A	$\frac{e}{a}$	APF	Col_M	l_{dist}

outputs of BART are very similar to those without additional non-informative data and Col_A, $\frac{e}{a}$, Col_M, APF, and l_{dist} are again frequently chosen in different N showing a robust performance. While compared with Table 3, the selections from BMARS experience more changes and are more influenced by this uncorrelated knowledge.

Moving to the interaction effects, BART successfully neglects unimportant features and maintains its performance. At the same time, BMARS is capable of (almost) filtering out all non-informative features and only leaves a small portion of the interactions between new predictors and the original data. Exact selection results can be found in Supplementary Note 2.

Optimal design for stacking fault energy in high entropy alloy spaces

To further demonstrate our model's advantage, we search among a much larger discrete material design space whose size is 36,273 instead of 403 in the previous section. This dataset represents face-centred cubic (FCC) compositions in the 7-element CoCr-FeMnNiV-Al high entropy alloy (HEA) space. Specifically, we focus on the task of exploring the stacking fault energy (SFE) chemical landscape in this system. SFE is an intrinsic property of crystals that measures their inherent resistance for adjacent crystal plans to shear against each other. Its value can be a good indicator of the (dominant) plastic deformation mechanism of the alloy and is thus a valuable alloy design parameter[45,46]. The SFE in this alloy system has been predicted for each composition using a support vector regressor trained on 498 high-fidelity SFE calculations from density functional theory (DFT) using the axial-next-nearest-neighbor-Ising model[47,48] relating SFE to the lattice energies of (disordered) FCC, hexagonal close-packed (HCP) and double HCP (DHCP) crystals of the same chemical composition. In addition to SFE, stoichiometrically weighted averages and variances were calculated for each composition for 17 pure element properties to generate a total of 34 property-based features.

In this new analysis, we have two goals: namely, to find the global minimum and global maximum in the SFE landscape. Thus, we record the minimum value and the maximum value of the observed response in each iteration for each model for comparison. As we can see, a faster curve increase in Fig. 3a, b and a sharper curve decline in Fig. 3c, d indicate a more efficient search for the target point and better performance.

For the predictors, we choose 41 potential predictors (34 property-based features in addition to the compositions of the seven constituent elements), which provides a larger set of candidate features. This set is a mixture of informative and (potentially) uninformative features and some of the informative features are correlated to each other, which may bring about a more challenging feature selection task. We follow the analysis above and compare our suggested framework that uses BART[32] and BMARS[30] to the GP regression (RBK, RBK ARD, Dot, and DKNet)[27,29,38,39] and Bayesian model average using GP (BMA1 and BMA2)[14]. For the acquisition function, we continue using EI for each of them to maintain a fair comparison. Also, we start the above models with five different sizes of initial samples ($N = 2, 5, 10, 15, 20$), which are randomly chosen from the design space. For each N, the results are based on 100 replicates. Curves for $N = 10$ and 20 are presented here and outputs under other initial sample sizes are summarized in Supplementary Note 3.

As seen in Fig. 3a, b, when looking for the maximum SFE, the solid blue curves for BMARS, solid red curves for BART, and dotted light blue curves for GP (Dot) have the sharpest increase, indicating the best performance. While the other curves representing other GP-based surrogates tend to move slowly.

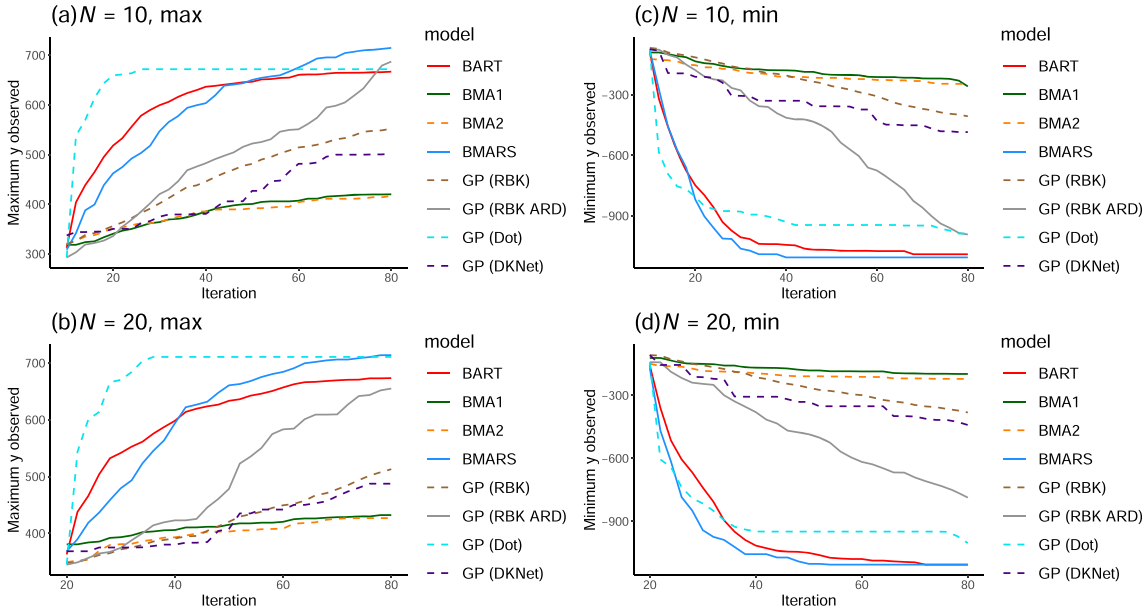

Fig. 3 The average maximum or minimum stacking fault energy (SFE) [mJ·m^{-2}] observed on each model in each iteration. a The maximum SFE with the initial set of sample size $N = 10$. **b** The maximum SFE with the initial set of sample size $N = 20$. **c** The minimum SFE with the initial set of sample size $N = 10$. **d** The minimum SFE with the initial set of sample size $N = 20$.

Table 7. The top five important factors selected by BART for the maximum stacking fault energy (SFE) with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$.

Setting	Top 1	Top 2	Top 3	Top 4	Top 5
$N = 2$	Specific.Heat_Avg	Pauling.EN_Var	Mn	Ni	C_11_Avg
$N = 5$	Mn	Pauling.EN_Var	C._Avg	C_11_Avg	Specific.Heat_Avg
$N = 10$	Specific.Heat_Avg	Pauling.EN_Var	Mn	Ni	Co
$N = 15$	Mn	Specific.Heat_Avg	Co	Pauling.EN_Var	SGTE.LSE_Avg
$N = 20$	Mn	Specific.Heat_Avg	Pauling.EN_Var	Ni	SGTE.LSE_Avg

The figure shows that BART- and BMARS-based BO is capable of finding the materials with SFE values close to the ground-truth maximum in the dataset in ~80 iterations, corresponding to just 0.25% of the total materials design space that could be explored. This is an impressive performance that is eclipsed when considering the performance of BART/BMARS-BO in the minimization problem, as shown in Fig. 3c, d. In Fig. 3c, d, the blue curves for BMARS and red curves for BART drop much faster than other curves, which confirms a more efficient search ability of our methods. In this case, by about ~40 iterations, the optimizer has converged to the points extremely close to the ground-truth minimum in the dataset. This corresponds to about 0.125% of the total materials design space. In this case, the performance of the proposed frameworks is much better than most of the alternatives. Here we note that, although GP (Dot) performs better than BMARS or BART in a few settings, an additional advantage of the latter methods is the automatic detection of important features detailed below.

In this case study, not only the design space has become much larger but also the number of candidate design features has increased. Using other approaches, it would be more difficult to evaluate the significance of the different features (or degrees of freedom) as well as their interactions. Here, we present the corresponding results for finding the maximum SFE, and those for the minimum SFE are in Supplementary Note 3.

Under five different scenarios $N \in \{2, 5, 10, 15, 20\}$, Tables 7 and 8 list the top five factors most correlated with the maximum in the SFE using BART and BMARS, respectively. The rankings are based on the median inclusion times of the 100 replicates from the last model when the workflow stops. For BART, Specific.Heat_Avg, Pauling.EN_Var, Mn, and Ni are the most important features. Meanwhile, turning to BMARS, Specific.Heat_Avg, Pauling.EN_Va, C_11_Avg, and Mn always play vital roles. Comparing top-ranked features for sets of different N, we observe similar patterns, but with a few differences in order. Immediately, one can see that only a few chemical elements are detected to be strongly correlated to the SFE in this HEA system and that, instead, other (atomically averaged) intrinsic properties may be more informative when attempting to predict this important quantity. This implies that focusing exclusively on chemistry as opposed to derived features may not have been an optimal strategy towards BO-based exploration of this space. Notably, Ni figures as the feature highly correlated to SFE in almost all scenarios considered. This is not surprising as Ni is also highly correlated with the stability of FCC over competing phases (such as HCP), and thus, higher Ni content in an alloy should be correlated to higher stability of FCC and higher SFE[49]. Co and Mn also appear as important covariates. In the case of Co, limited experimental studies have shown that increased Co tends to result in lower SFEs in FCC-based HEAs[50]. While trying to understand the underlying reasons for why other covariates (Specific hear, Pauling Electronegativity, etc.) seem to be highly correlated to SFE is beyond the scope of this work, what is notable is that in this framework, such insights can be gleaned at the same time that the materials problem space is

Table 8. The top five important factors selected by BMARS for the maximum stacking fault energy (SFE) with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$.

Setting	Top 1	Top 2	Top 3	Top 4	Top 5
$N = 2$	SGTE.LSE_Avg	Pauling.EN_Var	C_11_Avg	Ni	C._Avg
$N = 5$	Pauling.EN_Var	Specific.Heat_Avg	C_12_Avg	C._Avg	C_11_Avg
$N = 10$	C_11_Avg	Specific.Heat_Avg	Mn	Pauling.EN_Var	C._Var
$N = 15$	Specific.Heat_Avg	Mn	Pauling.EN_Var	Ni	C_11_Avg
$N = 20$	Pauling.EN_Var	Specific.Heat_Avg	SGTE.LSE_Avg	Fe	Mn

being explored. Thus, in an admittedly limited manner, the BART/BMARS-BO framework not only assists in the (very) efficient exploration of materials design spaces but also enhances our understanding of the underpinnings of material behavior. More results about the interactions among features are presented in Supplementary Note 3.

DISCUSSION

In general, there are two major categories of BO: (i) acquisition-based BO (ABO), and (ii) partitioning-based BO (PBO). ABO[10,12,36] is the most traditional and broadly used BO. The key idea is to pick an acquisition function, which is derived from the posterior and then optimized at each iteration to specify the next experiment. On the other hand, PBO[36,51] successfully avoids the optimization of acquisition functions by intelligently partitioning the space based on observed experiments and exploring promising areas, greatly reducing computations. Compared to PBO, ABO usually makes better use of the available knowledge and makes higher quality decisions, leading to a fewer number of needed experiments. In this study, we focused on ABO to construct the autonomous workflow for material discovery.

GP-based BO has been widely used in a number of areas and gradually become a benchmark method[12,13,24] for optimization of expensive "black-box" functions. However, its power can be limited by the intrinsic weaknesses of GP[10,12]. Isotropic covariance functions such as the Matérn and Gaussian kernels commonly employed in the literature have continuous sample paths, which is undesirable in many problems including material discovery as it is well known that the behavior of materials often changes abruptly with minute changes in chemical make-up or (multiscale) microstructural arrangements. Moreover, such isotropic kernels are provably suboptimal[52] in function estimation when there are spurious covariates or anisotropic smoothness. While remedies have been proposed in the literature involving more flexible kernel functions with additional hyperparameters[53] and sparse additive GPs[54,55], tuning and computation of such models can be significantly challenging, especially given a modest amount of data. Thus, in complex material science problems such as ours, Bayesian approaches based on additive regression trees or multivariate splines constitute an attractive alternative to GPs. Attractive theoretical properties of BART, including adaptivity to the underlying anisotropy and roughness, have recently appeared[56].

In this paper, we proposed a fully automated experimental design pipeline where we took advantage of more adaptive and flexible Bayesian models including BMARS[30,31] and BART[32] within an otherwise conventional BO procedure. A wide range of problems in scientific studies can be handled with this algorithm-based workflow, including MD. Both the simulation studies and real data analysis applied to scientifically relevant materials problems demonstrate that using BO with BMARS and BART outperforms GP-based methods in terms of searching speed for the optimal design and automatic feature importance determination. To be more specific, due to its well-designed spline basis, BMARS is able to catch challenging patterns like sudden transitions in the response surface. At the same time, BART also ensembles multiple individual trees and leads to a strong regression algorithm. Resulting from the recursive partitioning structures, they are equipped with a model-free variable selection that is based on feature inclusion frequencies in their basic functions and trees. This enables them to more accurately recognize the trends and correctly reveal the true factors.

We would like to close by briefly discussing potential applications of the framework in the context of autonomous materials research (AMR). Recently, the concept of autonomous experimentation for MD[57] has quickly emerged as an active area of research[58–60]. Going beyond traditional high-throughput approaches to MD[61–63], AMR aims to deploy robotic-assisted platforms capable of the automated exploration of complex materials spaces. Autonomy, in the context of AMR, can be achieved by developing systems capable of automatically selecting the experimental points to explore in a principled manner, with as little human intervention as possible. Our proposed non-GP BO methods seem to have robust performance against a wide range of problems. It is thus conceivable that the experimental design engines of AMR platforms could benefit from algorithms such as those proposed here.

METHODS

Bayesian optimization

BO[10] is a procedure intended to determine the global minimum (or maximum, with the similar procedure) $\mathbf{x}^*$ of an unknown objective function f sequentially and optimally, where $\mathcal{X}$ denotes the search space:

$$\mathbf{x}^* = \underset{\{\mathbf{x} \in \mathcal{X}\}}{\operatorname{argmin}} f(\mathbf{x}). \tag{3}$$

In the common setting of BO, the target function f can be either "black box" or expensive to evaluate, as such a function may represent a resource-intensive experiment or a very complex set of numerical simulations. Thus, we would like to reduce the number of function evaluations as we explore the design space and search for the optimal point. It mainly includes two steps: (i) fitting the hidden pattern of the target function, f, given observed data $\mathbf{D}$ so far based on some surrogate models, and (ii) optimizing selected utility or acquisition functions $u(\mathbf{x}|\mathbf{D})$ based on the posterior distribution of the surrogate estimates of f in order to decide the next sample point to evaluate in the design space $\mathcal{X}$. To be more specific, it generally follows Algorithm 1:

Algorithm 1. **Bayesian optimization (BO)**. **Input:** initial observed dataset $\mathbf{D} = \{(y_i, \mathbf{x}_i), i = 1, \ldots, N\}$. **Output:** candidate with desired properties. 1: Begin with $s = 1$. 2: **while** stopping criteria are not satisfied **do (3 to 7)**. 3: Train the chosen probabilistic model based on data $\mathbf{D}$. 4: Calculate the selected acquisition function $u(\mathbf{x}|\mathbf{D})$. 5: Choose the next experiment point by $\mathbf{x}^{s+1} = \operatorname{argmax}_{\{\mathbf{x}^{s+1} \in \mathcal{X}\}} u(\mathbf{x}|\mathbf{D})$. 6: Get the new point $(y^{s+1}, \mathbf{x}^{s+1})$ and add it into the observed dataset $\mathbf{D}$. 7: $s = s + 1$. 8: **return** the candidate with desired properties

A schematic illustration of BO is shown in Fig. 4—we note that such an algorithm can be implemented in autonomous experimental design platforms. Each of the subplots presents the state after one BO iteration, where they include the true unknown function (blue curve), utility

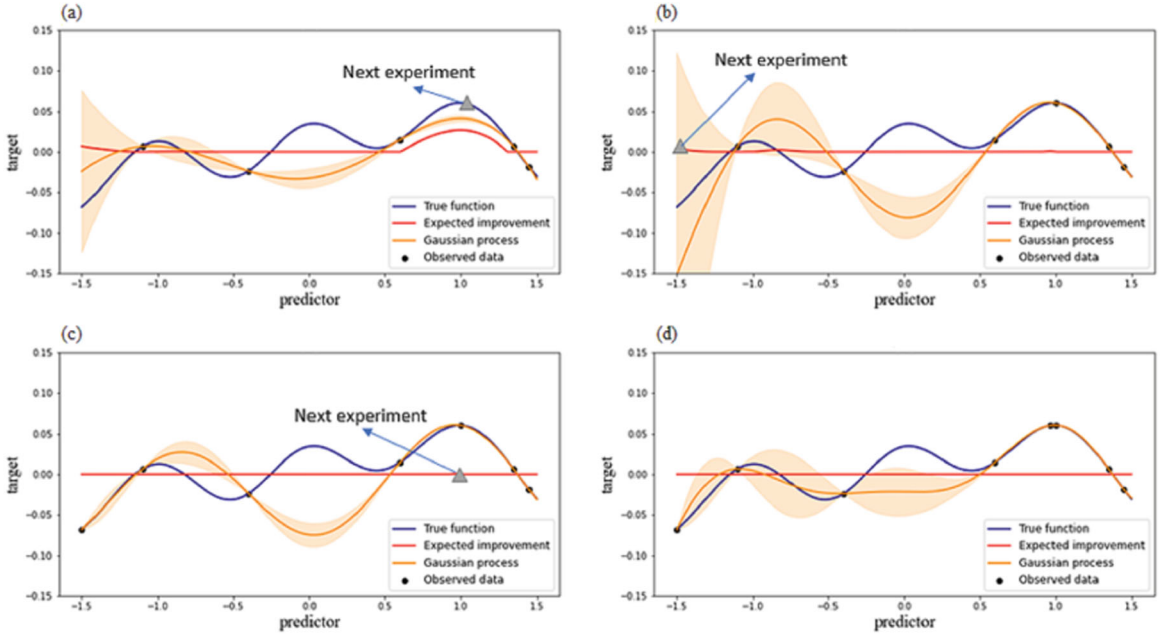

Fig. 4 Schematic illustration of Bayesian optimization (BO). Four subplots (**a**–**d**) give an example of the sequential automated experimental design using BO. They describe the true unknown function (blue curve), expected improvement (red curve), fitted values using GP (orange curve), 95% confidence interval (orange area), observed samples (black points), and the next experiment design (gray triangle) in each iteration.

function—in this case EI—(red curve), fitted values using GP (orange curve), 95% confidence interval (orange shaded area), observed samples (black points), and the next experiment recommended by the utility/acquisition function (gray triangle).

In this sequential optimization strategy, one of the key components is the Bayesian surrogate model for f, which is used to fit the available data[34] and to predict the outcome—with a measure of uncertainty—of experiments yet to be carried out. Another important determinant of BO efficiency is the choice of the acquisition function[34]. It can assist in setting our expectations regarding how much we can learn and gain from a new candidate design. The next design structure to be tested is usually the one that maximizes the acquisition function, balancing the trade-off between exploration and exploitation of the design space. There are many commonly used acquisition functions, such as EI, probability of improvement, upper confidence bound, and Thompson sampling[10,11]. Here, we choose to use EI as the acquisition function, which can find the point that, in expectation, improves on f_n^* the most:

$$u(\mathbf{x}) = \mathbb{E}\mathbb{I}_n(\mathbf{x}) := \mathbb{E}_n[(f(\mathbf{x}) - f_n^*)^+], \tag{4}$$

where f_n^* is the maximum value observed so far, $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}]$ is the expectation taken under the posterior distribution given the observed data, and $b^+ = \max(b, 0)$. We note that we have explored other acquisition functions and the relative performance of the corresponding methods with the same surrogates were not significantly different.

The choice of the surrogate model in BO will have a considerable impact on its performance, including the cost and time involved. As mentioned above, GPs[64] have been widely applied in BO in many applications, including MD[19]. In this work, we utilize BMARS and Bayesian ensemble-learning models, in particular, BART, to help guide the search through the design space more efficiently. We will briefly introduce the potential surrogate models in BO. More detailed technical descriptions of them are included in Supplementary Methods.

GP and model mixing

One of the popular ways in BO is using GP regression as the surrogate model to approximate the unknown f. Given $\mathbf{x}_i \in \mathbb{R}^p$ (design feature vectors) and $y_i (i = 1, \ldots, n)$ (evaluated f values at the corresponding $\mathbf{x}_i$'s, which can be noisy), we aim to fit the pattern of f and predict a new y_* associated with $\mathbf{x}_*$. Usually, we assume that y_i is a function of $\mathbf{x}_i$ with additional noise: $y_i = f(\mathbf{x}_i) + \epsilon_i$, $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. In GP regression, a GP prior is put on the unknown function f and $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^\top$ follows a joint Gaussian distribution:

$$p(\mathbf{f}) \sim \mathcal{N}(\mathbf{f}|m(\mathbf{x}), \mathbf{K}), \quad [K_{ij}] = k(\mathbf{x}_i, \mathbf{x}_j), \tag{5}$$

where $m(\cdot)$ is the mean function and $k(\cdot, \cdot)$ is the kernel function.

A common choice for $m(\cdot)$ is a constant mean function. For $k(\cdot, \cdot)$, there are various candidates and we can decide it based on the corresponding task. The radial-basis function (RBF) kernel is popular to capture stationary and isotropic patterns. RBF kernels with ARD[27] assign different scale parameters for each feature instead of using a common value[27], which can help to identify key covariates determining f. There are also nonstationary kernels, such as dot-product kernels[39] and more flexible deep network kernels[29]. For simplicity, we use $\mathbf{D} = \{\mathbf{x}_{1:n}, \mathbf{y}_{1:n}\}$ to denote the data we have collected. For a new input $\mathbf{x}_*$, the predictive distribution of response y_* is:

$$p(y_*|\mathbf{x}_*, \mathbf{D}) = \mathcal{N}(\mu_*, \sigma_*^2), \tag{6}$$

$$\mu_* = m(\mathbf{x}_*) + k(\mathbf{x}_*, \mathbf{x}_{1:n})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}(\mathbf{y}_{1:n} - m(\mathbf{x}_{1:n})), \tag{7}$$

$$\sigma_*^2 = k(\mathbf{x}_*, \mathbf{x}_*) + \sigma^2 - k(\mathbf{x}_*, \mathbf{x}_{1:n})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}k(\mathbf{x}_{1:n}, \mathbf{x}_*). \tag{8}$$

GP-based nonparametric regression approaches have gained a lot of popularity and have been widely used in various applications[12,13,24]. However, when turning to the sequential experiments in MD, the model may be imprecise and the search may be inefficient if we do not have enough information about the predictive performance of each experimental degree of freedom. Talapatra et al.[14] address this issue by using model mixing to develop multiple GP regression models based on different combinations of the covariates and weigh all the potential models according to their likelihood of being the true model. In this way, they incorporated model uncertainty, leading to a more robust framework capable of adaptively discovering the subset of covariates most predictive of the objective function to optimize.

Bayesian multivariate adaptive regression splines

BMARS[30,65] is a Bayesian version of the classical MARS model[31], which is a flexible nonparametric regression technique. It uses product spline basis functions to model f and it automatically identifies the nonlinear interactions among covariates. The regression develops a relationship between the covariates $\mathbf{x}_i \in \mathbb{R}^p$ and the response $y_i (i = 1, \ldots, n)$ as

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \hat{f}(\mathbf{x}_i) = \sum_{j=1}^{l} a_j B_j(\mathbf{x}_i), \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \tag{9}$$

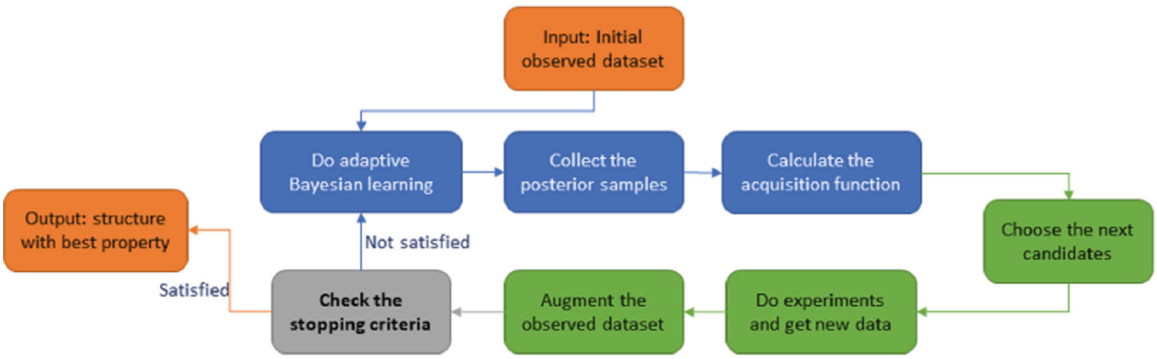

Fig. 5 Workflow of the automated experimental design framework. The overall workflow of the automated experimental design framework is based on Bayesian optimization with adaptive Bayesian learning.

where α_j denotes the relevant coefficient for the basic function B_j taking the form of

$$B_j(\mathbf{x}_i) = \begin{cases} 1, & j = 1, \\ \prod_{q=1}^{Q_j} [s_{qj} \cdot (\mathbf{x}_{i,v(q,j)} - t_{qj})]_+, & j \in \{2, \ldots, I\}, \end{cases} \tag{10}$$

with $s_{qj} \in \{-1, 1\}$, $v(q, j)$ denoting the index of the variables, and the set $\{v(q, j); q = 1, \ldots, Q_j\}$ not repeated. Here t_{qj} tells the partition location, $(\cdot)_+ = \max(0, \cdot)$, and Q_j is the polynomial degree of the basis B_j and also corresponds to the number of predictors involved in B_j. The number of parameters is $O(I)$ and we set the maximum value of I as 500.

To obtain samples from the joint posterior distribution, the computation is mainly based on the reversible jump Metropolis–Hastings algorithms[66]. The sampling scheme only draws the important covariates, hence automatic feature selection is naturally done in this procedure.

Ensemble learning and BART

Apart from model mixing, ensemble learning[67] provides an alternative way of combining models, which is a popular procedure that constructs multiple weak learners and aggregates them into a stronger learner[68–70]. In several circumstances, it is challenging for an individual model to capture the unknown complex mechanism connecting inputs to the output(s) by itself. Therefore, it is a better strategy to use a divide-and-conquer method in the ensemble-learning framework, which allows each of the models to fit a small part of the function. This is the key difference of our adopted Bayesian ensemble learning from the GP-based model mixing strategy in Talapatra et al.[14]. Ensemble learning's robust performance to handle complex data makes it a great candidate for BO[71]. However, it has not been explored to its full potential in the context of optimal experimental design yet. Hence, we choose to combine BO with the Bayesian ensemble learning[72], in particular, BART[32]. As BART is a tree-based model without inherent smoothness assumptions, it is also a more flexible surrogate model when modeling objective functions that are non-smooth, often encountered in MD. This strategy is effective and efficient due to its ability to take advantage of both the ensemble-learning procedure and the Bayesian paradigm.

BART[32] is a nonparametric regression method utilizing the Bayesian ensemble-learning technique. Many simulations and real-world applications confirmed its flexible fitting capabilities[73–75]. Given $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i (i = 1, \ldots, n)$, where it approximates the target function f by aggregating a set of regression trees:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \hat{f}(\mathbf{x}_i) = \sum_{j=1}^{I} g_j(\mathbf{x}_i; T_j, \boldsymbol{M}_j), \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \tag{11}$$

where T_j denotes a binary regression tree, $\boldsymbol{M}_j = (\mu_{j1}, \ldots, \mu_{jb_j})^\top$ denotes a vector of means corresponding to the b_j leaf nodes of T_j, and $g_j(\mathbf{x}_i; T_j, \boldsymbol{M}_j)$ is the function that assigns $\mu_{jt} \in \boldsymbol{M}_j$ to $\mathbf{x}_i$.

Using regularization priors on those trees is critical for the superior performance of this ensemble regression model. That way, each tree will be regularized to explain a small and distinct part of f. This aligns with the essence of ensemble learning, which is about combining weak learners into a stronger model. The number of parameters is correlated with the number of trees I as well as the tree depth d_j and is $O(I \cdot 2^d)$. In our analysis, I is set as 50 and d_j is usually smaller than 6.

At the same time, one can use this regression model for automatic variable selection, which greatly expands its scope of use. The importance of each predictor is based on the average variable inclusion frequency in all splitting rules[32]. Bleich et al.[76] further put forward a permutation-based inferential approach, which is a good alternative for the factor significance determination.

Automated experimental design framework

With BO using BMARS or BART, we propose an autonomous platform for efficient experimental design, aiming at significantly reducing the number of required trials and the total expense to find the best candidate in MD. The framework is depicted in Fig. 5 and the detailed description is as follows.

In this workflow, we begin with an initially observed dataset and the sample size can be as small as two. Then, we train our surrogate Bayesian learning model on the observed dataset and collect the relevant posterior samples. Using these samples, the acquisition function for each potential experiment to perform is calculated. After obtaining the values of the acquisition function, we select the candidates with top scores and do experiments at these points. With the new outcomes, the observed dataset is augmented and the stopping criteria are checked. If the criteria are fulfilled, we stop the workflow and return the candidate with the desired properties. Otherwise, we update the surrogate model by making use of the augmented dataset and use the updated belief to guide the next round of experiments.

Within this fully automated framework, what we need to provide is the initial sample and the stopping criteria. The beginning dataset can be some available data before this project. If we do not have this kind of information, we can randomly conduct a small number of experiments to populate the database and initialize the surrogate models used in the sequential experimental protocol. For the stopping criteria, it can be arriving at the desired properties or running out of the experimental budget[14].

DATA AVAILABILITY

The data files for materials discovery in the MAX phase space and optimal design for stacking fault energy in high entropy alloy space are available upon reasonable request.

Received: 2 July 2021; Accepted: 3 November 2021;
Published online: 03 December 2021

REFERENCES

1. Mockus, J. In *Bayesian Approach to Global Optimization*, 125–156 (Springer, Dordrecht, 1989).
2. Kushner, H. J. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *J. Basic Eng.* **86**, 97–106 (1964).
3. Jones, D. R., Schonlau, M. & Welch, W. J. Efficient global optimization of expensive black-box functions. *J. Glob. Optim.* **13**, 455–492 (1998).
4. Kaufmann, E., Cappé, O. & Garivier, A. On Bayesian upper confidence bounds for bandit problems. In *Proc. 15th International Conference on Artificial Intelligence and Statistics (AISTAT)*, 592–600 (JMLR, 2012).

5. Garivier, A. & Cappé, O. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proc. 24th Annual Conference on Learning Theory*, 359–376 (JMLR Workshop and Conference Proceedings, 2011).
6. Maillard, O.-A., Munos, R. & Stoltz, G. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proc. 24th annual Conference On Learning Theory*, 497–514 (JMLR Workshop and Conference Proceedings, 2011).
7. Auer, P., Cesa-Bianchi, N. & Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47**, 235–256 (2002).
8. Negoescu, D. M., Frazier, P. I. & Powell, W. B. The knowledge-gradient algorithm for sequencing experiments in drug discovery. *INFORMS J. Comput.* **23**, 346–363 (2011).
9. Lizotte, D. J., Wang, T., Bowling, M. H. & Schuurmans, D. Automatic gait optimization with Gaussian process regression. In *Proc. Int. Joint Conf. on Artificial Intelligence,* **7**, 944–949 (2007).
10. Frazier, P. I. Bayesian optimization. In *Recent Advances in Optimization and Modeling of Contemporary Problems*, 255–278 (INFORMS, 2018).
11. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & De Freitas, N. Taking the human out of the loop: a review of Bayesian optimization. *Proc. IEEE* **104**, 148–175 (2015).
12. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. *Adv. Neural Inform. Process. Syst.* **25**, 2960–2968 (2012).
13. Iyer, A. et al. Data-centric mixed-variable Bayesian optimization for materials design. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 59186, V02AT03A066 (American Society of Mechanical Engineers, 2019).
14. Talapatra, A. et al. Autonomous efficient experiment design for materials discovery with Bayesian model averaging. *Phys. Rev. Mater.* **2**, 113803 (2018).
15. Ju, S. et al. Designing nanostructures for phonon transport via Bayesian optimization. *Phys. Rev. X* **7**, 021024 (2017).
16. Ghoreishi, S. F., Molkeri, A., Srivastava, A., Arroyave, R. & Allaire, D. Multi-information source fusion and optimization to realize icme: application to dual-phase materials. *J. Mech. Des.* **140**, 111409 (2018).
17. Khatamsaz, D. et al. Efficiently exploiting process-structure-property relationships in material design by multi-information source fusion. *Acta Mater.* **206**, 116619 (2021).
18. Ghoreishi, S. F., Molkeri, A., Arróyave, R., Allaire, D. & Srivastava, A. Efficient use of multiple information sources in material design. *Acta Mater.* **180**, 260–271 (2019).
19. Frazier, P. I. & Wang, J. Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design*, 45–75 (Springer, 2016).
20. Liu, Y., Wu, J.-M., Avdeev, M. & Shi, S.-Q. Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties. *Adv. Theory Simul.* **3**, 1900215 (2020).
21. Janet, J. P. & Kulik, H. J. Resolving transition metal chemical space: feature selection for machine learning and structure–property relationships. *J. Phys. Chem. A* **121**, 8939–8954 (2017).
22. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 1–13 (2017).
23. Honarmandi, P., Hossain, M., Arroyave, R. & Baxevanis, T. A top-down characterization of NiTi single-crystal inelastic properties within confidence bounds through Bayesian inference. *Shap. Mem. Superelasticity* **7**, 50–64 (2021).
24. Ceylan, Z. Estimation of municipal waste generation of turkey using socio-economic indicators by Bayesian optimization tuned Gaussian process regression. *Waste Manag. Res.* **38**, 840–850 (2020).
25. Moriconi, R., Deisenroth, M. P. & Kumar, K. S. High-dimensional bayesian optimization using low-dimensional feature spaces. *Mach. Learn.* **109**, 1925–1943 (2020).
26. Wang, Z., Hutter, F., Zoghi, M., Matheson, D. & de Feitas, N. Bayesian optimization in a billion dimensions via random embeddings. *J. Artif. Intell. Res.* **55**, 361–387 (2016).
27. Aye, S. A. & Heyns, P. An integrated Gaussian process regression for prediction of remaining useful life of slow speed bearings based on acoustic emission. *Mech. Syst. Signal Process.* **84**, 485–498 (2017).
28. Paciorek, C. J. & Schervish, M. J. Nonstationary covariance functions for gaussian process regression. In *Advances in Neural Information Processing Systems*, 273–280 (Citeseer, 2003).
29. Wilson, A. G., Hu, Z., Salakhutdinov, R. & Xing, E. P. Deep kernel learning. In *Artificial Ontelligence and Statistics*, 370–378 (PMLR, 2016).
30. Denison, D. G., Mallick, B. K. & Smith, A. F. Bayesian MARS. *Stat. Comput.* **8**, 337–346 (1998).
31. Friedman, J. H. Multivariate adaptive regression splines. *Ann. Statist.* 1–67 (1991).
32. Chipman, H. A., George, E. I. & McCulloch, R. E. et al. Bart: Bayesian additive regression trees. *Ann. Appl. Stat.* **4**, 266–298 (2010).
33. HamediRad, M. et al. Towards a fully automated algorithm driven platform for biosystems design. *Nat. Commun.* **10**, 1–10 (2019).
34. Mateos, C., Nieves-Remacha, M. J. & Rincón, J. A. Automated platforms for reaction self-optimization in flow. *React. Chem. Eng.* **4**, 1536–1544 (2019).
35. Bashir, L. Z. & Hasan, R. S. M. Solving banana (rosenbrock) function based on fitness function. *World Sci. News* **12**, 41–56 (2015).
36. Merrill, E., Fern, A., Fern, X. & Dolatnia, N. An empirical study of Bayesian optimization: acquisition versus partition. *J. Mach. Learn. Res.* **22**, 1–25 (2021).
37. Pohlheim, H. GEATbx: Genetic and Evolutionary Algorithm Toolbox for use with MATLAB Documentation. http://www.geatbx.com/docu/algindex-03.html (2008).
38. Vert, J.-P., Tsuda, K. & Schölkopf, B. A primer on kernel methods. *Kernel Methods Comput. Biol.* **47**, 35–70 (2004).
39. Williams, C. K. & Rasmussen, C. E. *Gaussian Processes for Machine Learning*, Vol. 2 (MIT Press, 2006).
40. Molga, M. & Smutnicki, C. Test functions for optimization needs. *Test. Funct. Optim. Needs* **101**, 48 (2005).
41. Barsoum, M. W. *MAX Phases: Properties of Machinable Ternary Carbides and Nitrides* (Wiley, 2013).
42. Aryal, S., Sakidja, R., Barsoum, M. W. & Ching, W.-Y. A genomic approach to the stability, elastic, and electronic properties of the max phases. *Phys. Stat. Sol.* **251**, 1480–1497 (2014).
43. Barsoum, M. W. & Radovic, M. Elastic and mechanical properties of the max phases. *Annu. Rev. Mater. Res.* **41**, 195–227 (2011).
44. Rana, S., Li, C., Gupta, S., Nguyen, V. & Venkatesh, S. High dimensional Bayesian optimization with elastic Gaussian process. In *International Conference on Machine Learning*, 2883–2891 (PMLR, 2017).
45. Chaudhary, N., Abu-Odeh, A., Karaman, I. & Arróyave, R. A data-driven machine learning approach to predicting stacking faulting energy in austenitic steels. *J. Mater. Sci.* **52**, 11048–11076 (2017).
46. Hu, Y.-J., Sundar, A., Ogata, S. & Qi, L. Screening of generalized stacking fault energies, surface energies and intrinsic ductile potency of refractory multicomponent alloys. *Acta Mater.* **210**, 116800 (2021).
47. Denteneer, P. & Soler, J. Energetics of point and planar defects in aluminium from first-principles calculations. *Solid State Commun.* **78**, 857–861 (1991).
48. Denteneer, P. & Van Haeringen, W. Stacking-fault energies in semiconductors from first-principles calculations. *J. Phys. C* **20**, L883 (1987).
49. Cockayne, D., Jenkins, M. & Ray, I. The measurement of stacking-fault energies of pure face-centred cubic metals. *Philos. Mag.* **24**, 1383–1392 (1971).
50. Liu, S. et al. Transformation-reinforced high-entropy alloys with superior mechanical properties via tailoring stacking fault energy. *J. Alloys Compd.* **792**, 444–455 (2019).
51. Wang, S. & Ng, S. H. Partition-based Bayesian optimization for stochastic simulations. In *2020 Winter Simulation Conference (WSC)*, 2832–2843 (IEEE, 2020).
52. Bhattacharya, A., Pati, D. & Dunson, D. Anisotropic function estimation using multi-bandwidth Gaussian processes. *Ann. Stat.* **42**, 352 (2014).
53. Cheng, L. et al. An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. *Nat. Commun.* **10**, 1–11 (2019).
54. Qamar, S. & Tokdar, S. T. Additive Gaussian process regression. Preprint at https://arxiv.org/abs/1411.7009 (2014).
55. Vo, G. & Pati, D. Sparse additive Gaussian process with soft interactions. *Open J. Stat.* **7**, 567 (2017).
56. Ročková, V. & van der Pas, S. et al. Posterior concentration for Bayesian regression trees and forests. *Ann. Stat.* **48**, 2108–2131 (2020).
57. Nikolaev, P. et al. Autonomy in materials research: a case study in carbon nanotube growth. *npj Comput. Mater.* **2**, 1–6 (2016).
58. Kusne, A. G. et al. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat. Commun.* **11**, 5966 (2020).
59. Aldeghi, M., Häse, F., Hickman, R. J., Tamblyn, I. & Aspuru-Guzik, A. Golem: an algorithm for robust experiment and process optimization. *Chem. Sci.* **12**, 14792–14807 (2021).
60. Häse, F. et al. Olympus: a benchmarking framework for noisy optimization and experiment planning. *Mach. Learn. Sci. Technol.* **2**, 035021 (2021).
61. Liu, P. et al. High throughput materials research and development for lithium ion batteries. *High-throughput Exp. Model. Res. Adv. Batter.* **3**, 202–208 (2017).
62. Melia, M. A. et al. High-throughput additive manufacturing and characterization of refractory high entropy alloys. *Appl. Mater. Today* **19**, 100560 (2020).
63. Potyrailo, R. et al. Combinatorial and high-throughput screening of materials libraries: review of state of the art. *ACS Comb. Sci.* **13**, 579–633 (2011).
64. Schulz, E., Speekenbrink, M. & Krause, A. A tutorial on Gaussian process regression: modelling, exploring, and exploiting functions. *J. Math. Psychol.* **85**, 1–16 (2018).
65. Denison, D. G., Holmes, C. C., Mallick, B. K. & Smith, A. F. Bayesian methods for nonlinear classification and regression, Vol. 386 (John Wiley & Sons, 2002).
66. Green, P. J. Reversible jump markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732 (1995).
67. Sagi, O. & Rokach, L. Ensemble learning: a survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **8**, e1249 (2018).

68. Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J. & Woźniak, M. Ensemble learning for data stream analysis: a survey. *Inf. Fusion* **37**, 132–156 (2017).
69. Laradji, I. H., Alshayeb, M. & Ghouti, L. Software defect prediction using ensemble learning on selected features. *Inf. Softw. Technol.* **58**, 388–402 (2015).
70. Chen, X. M., Zahiri, M. & Zhang, S. Understanding ridesplitting behavior of on-demand ride services: an ensemble learning approach. *Transp. Res. Part C* **76**, 51–70 (2017).
71. Zhang, W., Wu, C., Zhong, H., Li, Y. & Wang, L. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geosci. Front.* **12**, 469–477 (2021).
72. Fersini, E., Messina, E. & Pozzi, F. A. Sentiment analysis: Bayesian ensemble learning. *Decis. Support Syst.* **68**, 26–38 (2014).
73. Hill, J., Linero, A. & Murray, J. Bayesian additive regression trees: a review and look forward. *Annu. Rev. Stat. Appl.* **7**, 251–278 (2020).
74. McCord, S. E., Buenemann, M., Karl, J. W., Browning, D. M. & Hadley, B. C. Integrating remotely sensed imagery and existing multiscale field data to derive rangeland indicators: application of Bayesian additive regression trees. *Rangel. Ecol. Manag.* **70**, 644–655 (2017).
75. Sparapani, R. A., Logan, B. R., McCulloch, R. E. & Laud, P. W. Nonparametric survival analysis using Bayesian additive regression trees (bart). *Stat. Med.* **35**, 2741–2753 (2016).
76. Bleich, J., Kapelner, A., George, E. I. & Jensen, S. T. Variable selection for bart: an application to gene regulation. *Ann. Appl. Stat.* **8**, 1750–1781 (2014).

ACKNOWLEDGEMENTS

B.K.M., A.B., and D.P. acknowledge support by NSF through Grant No. NSF CCF-1934904 (TRIPODS). T.Q.K. acknowledges the NSF through Grant No. NSF-DGE-1545403. X.Q. and R.A. acknowledge NSF through Grants Nos. 1835690 and 2119103 (DMREF). The authors also acknowledge Texas A&M's Vice President for Research for partial support through the X-Grants program. Dr. Prashant Singh (Ames Laboratory) is acknowledged for his DFT calculations of SFE in FCC HEAs. Dr. Anjana Talapatra and Dr. Shahin Boluki are acknowledged for facilitating the BMA Code. DFT calculations of the SFEs were conducted with the computing resources provided by Texas A&M High Performance Research Computing.

AUTHOR CONTRIBUTIONS

B.L. and B.K.M. conceived of the concept for non-GP BO. B.L. implemented the algorithms and carried out the experiments. T.Q.K. provided the model for SFE as a function of composition. X.Q., A.B., and D.P. contributed to the discussion on the ML/BO aspects of the work. R.A. and T.Q.K. provided the materials science context and designed the materials science examples. All authors analyzed the results, contributed to the manuscript, and edited it. All authors reviewed the final version of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00662-x.

Correspondence and requests for materials should be addressed to Raymundo Arroyave.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021