Supplementary Information

Model Descriptions and Additional Results for
Bayesian Optimization with Adaptive Surrogate Models for Automated Experimental Design

---

Supplementary Methods

In GP regression, a GP prior is put on the unknown function f and $\mathbf{f} = (f(\mathbf{x}_1), \cdots, f(\mathbf{x}_n))^\top$ follows a joint Gaussian distribution:

$$p(\mathbf{f}) \sim \mathcal{N}(\mathbf{f}|m(\mathbf{x}), \mathbf{K}), \quad [\mathbf{K}_{ij}] = k(\mathbf{x}_i, \mathbf{x}_j), \tag{1}$$

where m is the mean function and k is the kernel function. For simplicity, we use $\mathbf{D} = \{\mathbf{x}_{1:n}, \mathbf{y}_{1:n}\}$ to denote the data we have collected, which leads to the GP posterior ($i = 1, \cdots, n$):

$$p(\mathbf{f}|\mathbf{x}_{1:n},\mathbf{y}_{1:n}) \sim \mathcal{N}(m_{\text{post}}, k_{\text{post}}), \tag{2}$$
$$m_{\text{post}}(\mathbf{x}_i) = m(\mathbf{x}_i) + k(\mathbf{x}_i, \mathbf{x}_{1:n})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}(\mathbf{y}_{1:n} - m(\mathbf{x}_{1:n})), \tag{3}$$
$$k_{\text{post}}(\mathbf{x}_i) = k(\mathbf{x}_i, \mathbf{x}_{1:n}) - k(\mathbf{x}_i, \mathbf{x}_{1:n})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}k(\mathbf{x}_{1:n}, \mathbf{x}_i). \tag{4}$$

When new input $\mathbf{x}_*$ arrives, the predictive distribution of response y_* is:

$$p(y_*|\mathbf{x}_*, \mathbf{D}) = \mathcal{N}(\mu_*, \sigma_*^2), \tag{5}$$
$$\mu_* = m(\mathbf{x}_*) + k(\mathbf{x}_*, \mathbf{x}_{1:n})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}(\mathbf{y}_{1:n} - m(\mathbf{x}_{1:n})), \tag{6}$$
$$\sigma_*^2 = k(\mathbf{x}_*, \mathbf{x}_*) + \sigma^2 - k(\mathbf{x}_*, \mathbf{x}_{1:n})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}k(\mathbf{x}_{1:n}, \mathbf{x}_*). \tag{7}$$

Under the scheme of BO using GP, the expected improvement can be evaluated analytically as follows:

$$\mathbb{EI}_n(\mathbf{x}_*|\mathbf{D}) = \begin{cases} (\mu_* - f_n^*)\Phi(z) + \sigma_*^2\phi(z) & , \quad \sigma_*^2 > 0 \\ 0 & , \quad \sigma_*^2 = 0, \end{cases} \tag{8}$$

where μ_* and σ_*^2 are the mean and the variance of the predictive distribution of GP at $\mathbf{x}_*$, f_n^* is the maximum value observed up to now, Φ and ϕ are the cumulative distribution function and probability density function of the standard normal distribution, and $z = (\mu_* - f_n^*)/\sigma_*$ if $\sigma_* > 0$, otherwise $z = 0$.

To better account for model uncertainty, Talapatra et al. [5] have used model mixing to develop multiple GP regression models based on different combinations of the available features and weigh all the potential models according to their likelihood of being the true model. Conditioning on the previous experiments, the predictive probabilistic model is

$$p(y_*|\mathbf{x}_*, \mathbf{D}) = \sum_{q=1}^{L} p(M_q|\mathbf{D})p(y_*|\mathbf{x}_*, \mathbf{D}, M_q), \tag{9}$$

where M_q denotes the q-th possible model, $p(y_*|\mathbf{x}_*, \mathbf{D}, M_q)$ is the relevant predictive distribution and L is the number of model candidates. $p(M_q|\mathbf{D})$ represents the relative importance of the q-th model and can be

expressed as

$$p(M_q|\mathbf{D}) = \frac{p(\mathbf{D}|M_q)p(M_q)}{\sum_{s=1}^{L} p(\mathbf{D}|M_s)p(M_s)}, \tag{10}$$

$$p(\mathbf{D}|M_q) = \int p(\mathbf{D}|\boldsymbol{\theta}_q, M_q)p(\boldsymbol{\theta}_q|M_q)d\boldsymbol{\theta}_q. \tag{11}$$

Here $\boldsymbol{\theta}_q$ denotes the parameters of model M_q.

For the q-th GP regression model M_q, the posterior $p(\mathbf{f}|\mathbf{D}, M_q)$ and predictive distribution $p(y_*|\mathbf{x}_*, \mathbf{D}, M_q)$ follow the Gaussian distribution and their corresponding mean and variance have a similar form as in (2) - (7). When calculating the acquisition function and deciding the next design point, the value of the expected improvement is a weighted sum of the values of each potential model:

$$\mathbb{EI}_n(\mathbf{x}_*|\mathbf{D}) = \sum_{q=1}^{L} p(M_q|\mathbf{D})\mathbb{EI}_n(\mathbf{x}_*|\mathbf{D}, M_q), \tag{12}$$

where the weight is the relevant significance of M_q, and $\mathbb{EI}_n(\mathbf{x}_*|\mathbf{D}, M_q)$ can be evaluated using (8).

To compute $p(M_q|\mathbf{D})$, we also need the model evidence $p(\mathbf{D}|M_q)$ and the prior $p(M_q)$. To derive the marginal likelihood, Talapatra et al. [5] proposed the Laplace approximation method where both first-order and quadratic-order expansions of the exponent have been used.

Moving to the Bayesian multivariate adaptive regression splines, following the description of Bayesian multivariate adaptive regression splines (BMARS) [3] in the main text of the manuscript and more specifically about the the Bayesian modeling framework, Denison et al. [3] assumed that the number of basis functions l, as well as their degree of interaction Q_j, their coefficients α_j, their relevant split points t_{qj}, and the sign indicators s_{qj} are model parameters. Assume $\boldsymbol{\theta}^{(l)} = \{\mathcal{B}_1, \cdots, \mathcal{B}_l\}$ where $\mathcal{B}_j$ denotes the relevant parameters $(Q_j, \alpha_j, t_{1j}, \cdots, t_{Q_j,j}, s_{1j}, \cdots, s_{Q_j,j})$ for the basis function B_j. Then, the hierarchical model can be written as

$$p(l, \boldsymbol{\theta}^{(l)}, \mathbf{y}) = p(l)p(\boldsymbol{\theta}^{(l)}|l)p(\mathbf{y}|l, \boldsymbol{\theta}^{(l)}). \tag{13}$$

The joint posterior for l and $\boldsymbol{\theta}^{(l)}$ can be considered in a factorized form

$$p(l, \boldsymbol{\theta}^{(l)}|\mathbf{y}) = p(l|\mathbf{y})p(\boldsymbol{\theta}^{(l)}|l, \mathbf{y}). \tag{14}$$

To obtain samples from the above joint posterior distribution (14), the computation is mainly based on the reversible jump Metropolis-Hastings algorithms [4], where the model is updated by randomly using one of the three moves including (a) a change in a knot location; (b) the creation of a basis function; (c) the removal of a basis function, and then calculate relevant acceptance probability.

As for the factor importance determination, it is based on the number of times when each variable is included in the basis. More specifically, we can see that each basic function B_j depends on Q_j features. Suppose we get S posterior samples for BMARS. If a certain feature is used in one of basic functions in a sample, we increase the inclusion frequency of the corresponding variable by one. After going through all the samples, we can rank factors by comparing the inclusion frequency. The larger the frequency is, the more significant role it plays.

When we utilize BMARS to estimate f in BO, the acquisition function expected improvement cannot be evaluated analytically as in (8) for GP. Hence, we need to estimate it using the MCMC based posterior samples. During the reversible jump Metropolis-Hastings algorithm mentioned above, we get the S posterior samples for the model parameters $\{\alpha_j^{(1)}, B_j(\mathbf{x}_*)^{(1)}\}, \cdots, \{\alpha_j^{(S)}, B_j(\mathbf{x}_*)^{(S)}\}$, $j = 1, \cdots, l_s$, where l_s is the number of basis functions in s-th sample and this number may vary between different samples. With these basic functions and the corresponding coefficients, we can obtain $(f^{(1)}(\mathbf{x}_*), \cdots, f^{(S)}(\mathbf{x}_*))$ which can approximate the posterior predictive distribution of $f(\mathbf{x}_*)$. Next, using the definition of the expected

improvement, we can calculate it using the s MCMC samples as

$$\mathbb{EI}_n(\mathbf{x}_*) \coloneqq \mathbb{E}_n[(f(\mathbf{x}_*) - f_n^*)^+] = \frac{1}{S}\sum_{s=1}^{S}(f^{(s)}(\mathbf{x}_*) - f_n^*)^+. \tag{15}$$

For ensemble Learning and Bayesian additive regression trees, following the description of Bayesian Additive Regression Trees (BART) [2] in the main manuscript, we include more introductions about it here. Use of regularization priors on those trees is critical for the superior performance of this regression model. That way, each tree will be short and explain a small and distinct part of f. This aligns with the idea of ensemble learning which is about combining weak learners into a stronger model. To be more specific, there are three crucial modeling stages, T_j, $\boldsymbol{M}_j$, and σ, respectively. For the part regarding T_j, it is defined in three components: (i) the probability that a node at depth d is not a leaf node:

$$p(T_j) \propto \alpha(1+d)^{-\beta}, \quad \alpha \in (0,1), \beta \in [0,\infty), \tag{16}$$

(ii) the uniform distribution over potential features at each interior node for splitting assignment, and (iii) the possible splitting values for the assigned predictor at each interior node. For the part $\mu_{jt}|T_j$, after scaling the response, people can choose a conjugate normal prior:

$$\mu_{jt} \sim \mathcal{N}(0, \sigma_\mu^2), \quad \text{where } \sigma_\mu = 0.5/k\sqrt{m}. \tag{17}$$

Here k usually is chosen between 1 and 3, and m is the number of trees. As for σ, a good choice is conjugate inverse chi-square distribution $\sigma^2 \sim \upsilon\lambda/\chi_\upsilon^2$, where the hyperparameters can be calibrated by the data.

Turning to the computation of the posterior, Chipman et al. [2] developed a backfitting MCMC algorithm, which mixes very well and is widely used nowadays. Given the observed data, a posterior distribution $p((T_1, M_1), \cdots, (T_l, M_l), \sigma|\mathbf{y})$ can specify this sum-of-trees model and the algorithm generally is a Gibbs sampler. To elaborate, let $T_{(j)}$ be a subset of all trees from the sum other than T_j, and define $M_{(j)}$ the same way. Then the Gibbs sampler first draws σ and $(T_j, M_j),\ j = 1, \cdots, l$ from their full conditionals:

$$\sigma|T_1, \cdots, T_l, M_1, \cdots, M_l, \mathbf{y}, \tag{18}$$

$$(T_j, M_j)|T_{(j)}, M_{(j)}, \sigma, \mathbf{y}. \tag{19}$$

The first part (18) is an inverse gamma distribution, while the second is more challenging to sample. Let $\mathbf{R}_j$ denote partial residuals:

$$\mathbf{R}_j \coloneqq \mathbf{y} - \sum_{k\neq j} g_j(\mathbf{x}; T_k, \boldsymbol{M}_k).$$

With $\mathbf{R}_j$, we can simplify the dependency (19) as below:

$$p((T_j, M_j)|T_{(j)}, M_{(j)}, \sigma, \mathbf{y}) = p((T_j, M_j)|\mathbf{R}_j, \sigma), \tag{20}$$

$$= p(T_j|\mathbf{R}_j, \sigma)p(M_j|T_j, \mathbf{R}_j, \sigma), \tag{21}$$

where the part concerning T_j is in the following form and can be simulated by using the Metropolis–Hastings (MH) algorithm of Chipman et al. [1]

$$p(T_j|\mathbf{R}_j, \sigma) \propto p(T_j)\int p(\mathbf{R}_j|M_j, T_j, \sigma)p(M_j|T_j, \sigma)dM_j. \tag{22}$$

Lastly considering the draws of M_j, we can independently sample from a normal distribution for the terminal node μ_{jt}.

Turning to the variable selection perspective, it is determined in a similar way with BMARS. BART

models the unknown function by a sum of trees. In each binary tree, a feature is chosen when a node is split into two. Thus when a certain feature is used as the splitting variable, the inclusion frequency will increase by one. In the end, a larger frequency indicates a more important role of the relevant feature.

Also similar to BMARS, the expected improvement acquisition function for BART does not have an explicit expression. Therefore, we use the posterior samples to estimate it. Given the observed data $\mathbf{D}$, for input $\mathbf{x}_*$, we can firstly draw samples for the sum-of-trees model $\{T_j^{(1)}, M_j^{(1)}\}, \cdots, \{T_j^{(S)}, M_j^{(S)}\}, j = 1, \cdots, l$. Then by using the tree models, we can get samples $(f^{(1)}(\mathbf{x}_*), \cdots, f^{(S)}(\mathbf{x}_*))$ from the posterior predictive distribution $p(y_*|\mathbf{x}_*, \mathbf{D})$. Next, we calculate the expected improvement function as

$$\mathbb{EI}_n(\mathbf{x}_*) \coloneqq \mathbb{E}_n[(f(\mathbf{x}_*) - f_n^*)^+] = \frac{1}{S}\sum_{s=1}^{S}(f^{(s)}(\mathbf{x}_*) - f_n^*)^+. \tag{23}$$

Supplementary Note 1

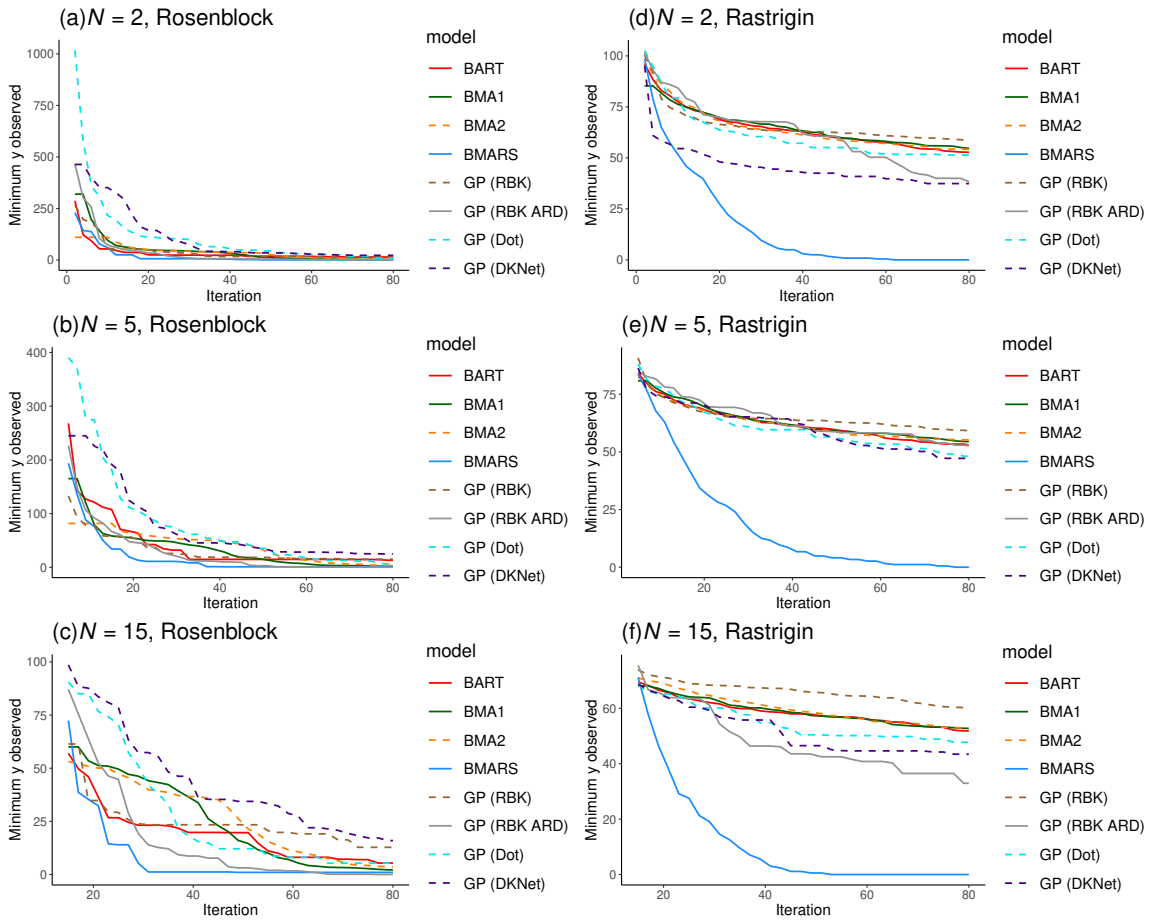
Here we show more results for the Simulation Studies. Supplementary Figure 1 shows the average minimum y observed based on each model in each iteration for Rosenbrock function and Rastrigin function ($N = 2, 5, 15$).

Here we add Gaussian noise to the simulated response and compare whether the models can maintain their performance in the presence of noise. The minimum y observed for Rosenbrock function and Rastrigin function ($N = 2, 5, 10, 15, 20$) are presented in Supplementary Figure 2.

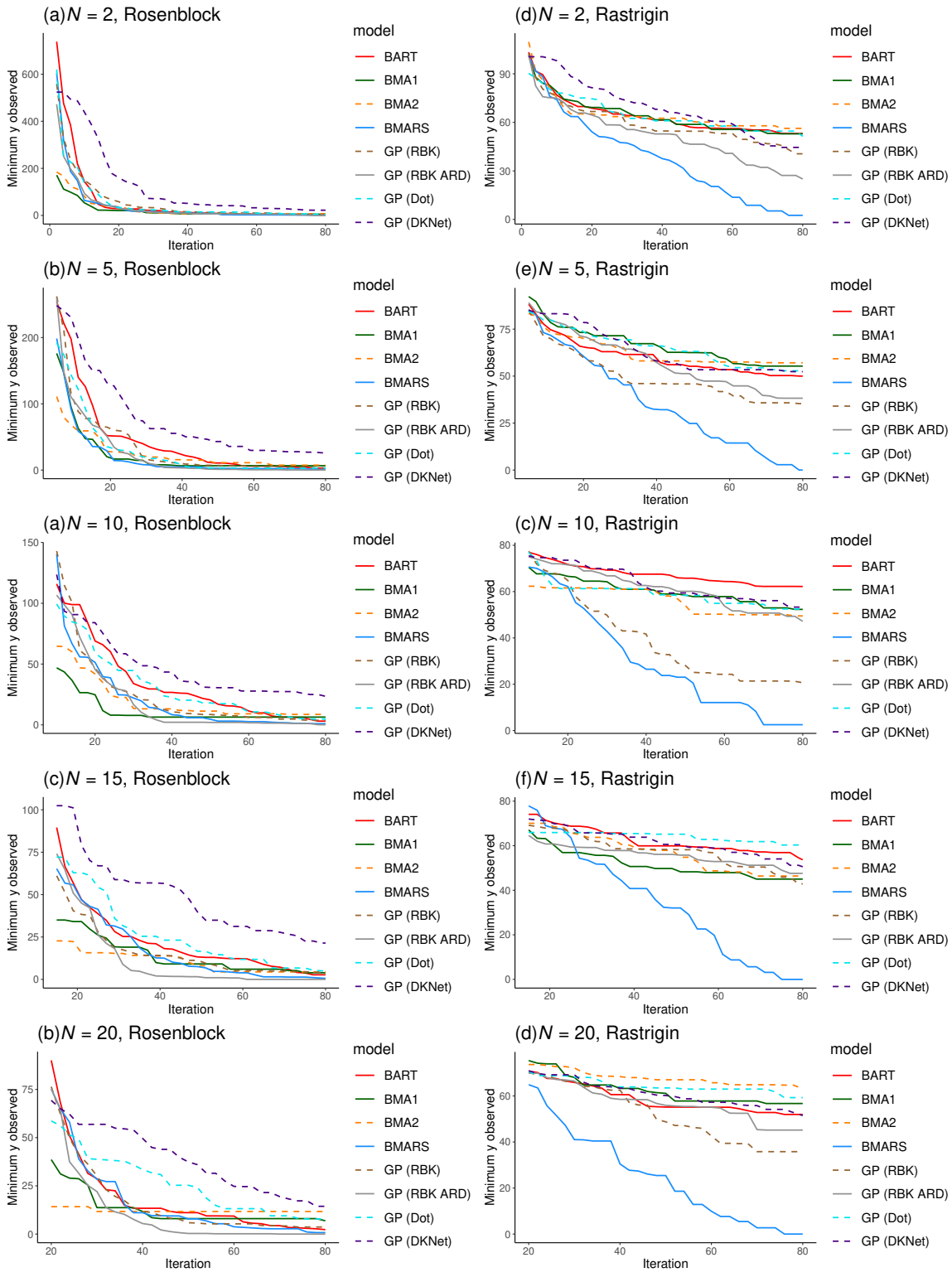
We compare the mean training time of each model under different data sizes. As depicted in Supplementary Table 1, it's true that MCMC is more expensive than standard GPs. However, the time needed for the MCMC algorithm is reasonable and we assume that they are much smaller than the consumed time for experiments. Thus, although we spend more time in training models, we can do fewer experiments and save more time.

Supplementary Table 1: The mean training time (s) based on N experiments from Rastrigin function ($N \in \{20, 30, \cdots, 80\}$).

Model	N=20	N=30	N=40	N=50	N=60	N=70	N=80
BART	0.2766	0.2942	0.3191	0.3588	0.4000	0.4164	0.4652
BMA1	2.601	2.1101	2.6623	2.9554	3.3236	4.524	6.3706
BMA2	2.7629	2.3867	3.1068	3.372	3.8659	4.8576	7.1652
BMARS	1.5395	1.5895	1.787	1.9185	2.0156	2.4459	2.4943
GP (RBK)	0.0065	0.0062	0.0067	0.0066	0.0071	0.0075	0.0074
GP (RBK with ARD)	0.0499	0.0461	0.0607	0.0713	0.075	0.0977	0.0996
GP (Dot)	0.0159	0.0163	0.0187	0.0211	0.0196	0.0233	0.0208
GP (DKNet)	2.3374	2.4776	2.8557	3.1647	3.4218	3.9592	4.3565



Supplementary Figure 1: The average minimum y observed based on each model in each iterations. **a** for Rosenbrock function with the initial set of sample size $N = 2$. **b** for Rosenbrock function with the initial set of sample size $N = 5$. **c** for Rosenbrock function under initial sample size $N = 15$. **d** for Rastrigin function with the initial set of sample size $N = 2$. **e** for Rastrigin function with the initial set of sample size $N = 5$. **f** for Rastrigin function with the initial set of sample size $N = 15$.

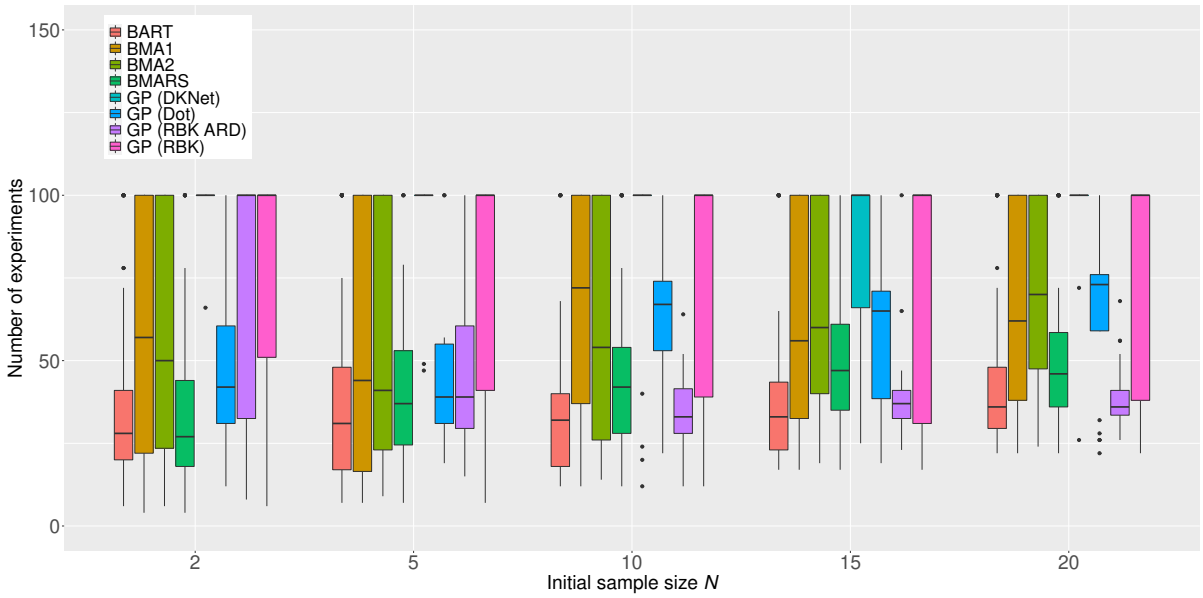


Supplementary Figure 2: The average minimum y observed based on each model in each iterations for Rosenbrock function with error with the initial set of sample size **a** $N = 2$, **b** $N = 5$, **c** $N = 10$, **d** for $N = 15$, **e** for $N = 20$, Rastrigin function with error with the initial set of sample size **f** $N = 2$, **g** $N = 5$, **h** $N = 10$, **i** $N = 15$, **j** $N = 20$.

Supplementary Note 2

Here we show more results for the Materials discovery in the MAX phase space.

To take a closer look at the search speed, as presented in Supplementary Figure 3, we can compare the distribution of the number of experiments under each different model and beginning sample size. It maintains the same pattern with Table 1 of the main manuscript where BART and BMARS are better at guiding the experiments and accelerate the exploration of the design space. GP(RBK ARD) also shows efficient search when N is relevant large. While results of other GP-based techniques are more likely to exceed the budget, especially for GP (DKNet).



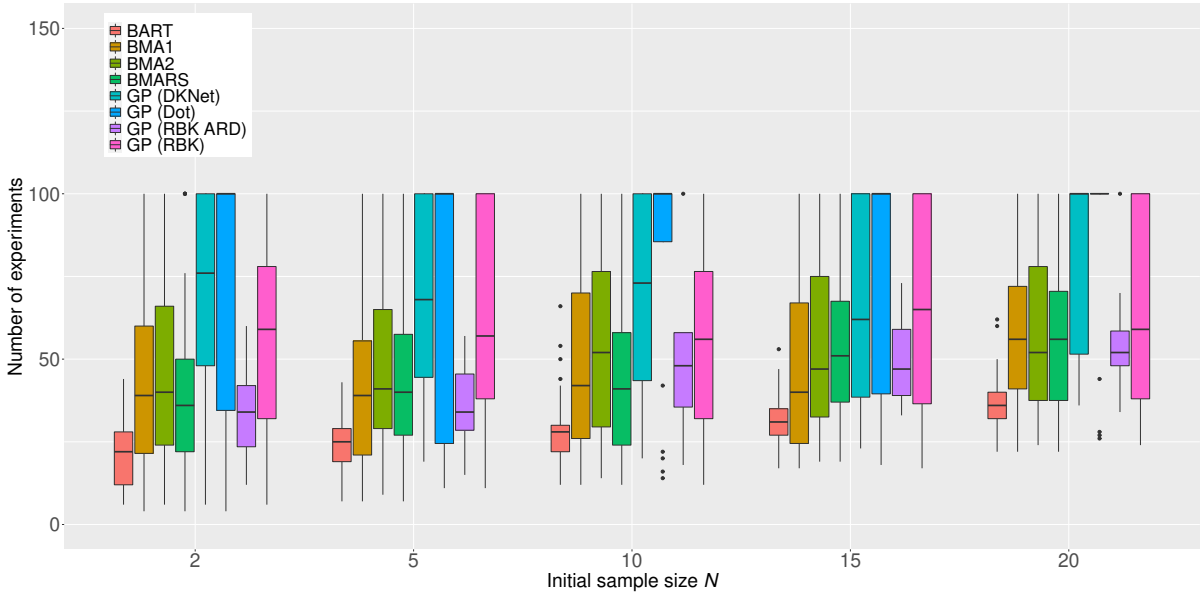
Supplementary Figure 3: Comparison of the number of experiments needed by BART, BMA1, BMA2, BMARS, GP (RBK), GP (RBK ARD), GP (Dot), and GP(DKNet), to find the maximum bulk modulus K with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$.

Supplementary Table 2 show the mean value and interquartile range (IQR) of the total number of evaluations searching for the minimum shear modulus G within the MAX phase design space. The smaller values of the mean and interquartile range mean a more efficient and stable workflow. As presented in Supplementary Table 2, in the exploration for the minimum shear modulus G, BART continues to outperform the other models, while BMARS works as well as GP (RBK ARD), BMA1, and BMA2. However, GP (RBK), GP (Dot), and GP (DKNet) become very slow.

Supplementary Table 2: The mean value and interquartile range (IQR) of the number of experiments based on each model to find the minimum shear modulus G 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	**21.52**	**23.9**	**27.44**	**31.8**	**36.22**	16	10	8	8	8
BMA1	40.72	41.67	47.68	45.08	57.26	38.5	34.5	44	42.5	31
BMA2	46.52	50.23	55.32	54.77	58.76	42	36	47	42.5	40.5
BMARS	38.22	42.81	**42.28**	53.08	**56.6**	28	30.5	34	30.5	33
GP (RBK)	59.12	61.41	57.72	62.89	64.58	46	62	44.5	63.5	62
GP (RBK with ARD)	**33.9**	**36.6**	43.6	**51.3**	60.9	19.5	22	16.5	19	18
GP (Dot)	74.95	72.2	80.7	76.1	86.25	65.5	75.5	14.5	60.5	12.5
GP (DKNet)	70.4	67.6	60.9	62.1	73.2	50.	59.5	33.5	66.	39.5

As depicted in Supplementary Figure 4, we can see the distribution of the number of experiments under each different model and beginning sample size when looking for the minimum shear modulus G. It maintains the same pattern with Supplementary Table 2. BART maintains the best performance in this case and BMARS and GP (RBK ARD) follow. While other methods compared are relatively inefficient.



Supplementary Figure 4: Comparison of the number of experiments needed by BART, BMA1, BMA2, BMARS, GP (RBK), GP (RBK ARD), GP (Dot), and GP(DKNet) to find the minimum shear modulus G with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$.

Supplementary Table 3 and Supplementary Table 4 show the top 5 key interaction effects for the maximum value of K using BART and BMARS, respectively. As summarized in the tables, most of the interactions playing a vital role are between the important features selected by BART and BMARS which are shown in Table 2 and 3 of the main manuscript. Regarding the order in the tables, BART tends to have more stable selection results than BMARS under different N.

Supplementary Table 3: The top 5 important interactions between 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
$N = 2$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : \text{APF}$	$\text{Col}_A : c$	$\text{Col}_A : I_{\text{dist}}$	$\text{Col}_A : \text{Col}_M$
$N = 5$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : c$	$\text{Col}_A : I_{\text{dist}}$	$\text{Col}_A : \text{Col}_M$	$\frac{e}{a} : \text{APF}$
$N = 10$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : \text{Col}_M$	$\text{Col}_A : I_{\text{dist}}$	$\text{Col}_A : \text{APF}$	$\text{Col}_A : c$
$N = 15$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : c$	$\text{Col}_A : \text{Col}_M$	$\text{Col}_A : \text{APF}$	$\text{Col}_A : I_{\text{dist}}$
$N = 20$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : c$	$\text{Col}_A : I_{\text{dist}}$	$\text{Col}_A : \text{Col}_M$	$\text{Col}_A : \text{APF}$

Supplementary Table 4: TThe top 5 important interactions between 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
$N = 2$	$c : \frac{e}{a}$	$a : I_{\text{dist}}$	$a : c$	Col_M : APF	$m : \frac{e}{a}$
$N = 5$	$\text{Col}_A : a$	Z : APF	$C : c$	$C : \frac{e}{a}$	$Z : c$
$N = 10$	$\text{Col}_A : I_{\text{dist}}$	$\text{Col}_A : \text{Col}_M$	$a : m$	Col_M : APF	$m : c$
$N = 15$	$\frac{e}{a}$: APF	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : Z$	$a : C_v$	$C : a$
$N = 20$	$\text{Col}_A : \frac{e}{a}$	C : APF	$c : I_{\text{dist}}$	Col_A : Z	$a : c$

Supplementary Table 5 and Supplementary Table 6 list the top 5 important features with the goal of the minimum point of K using BART and BMARS, respectively.

Supplementary Table 5: The top 5 important factors between factors selected by BART for the minimum shear modulus G 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
$N = 2$	$\frac{e}{a}$	I_{dist}	Col_M	Col_A	c
$N = 5$	$\frac{e}{a}$	Col_M	I_{dist}	Col_A	c
$N = 10$	$\frac{e}{a}$	I_{dist}	Col_M	Col_A	c
$N = 15$	$\frac{e}{a}$	I_{dist}	Col_M	Col_A	APF
$N = 20$	$\frac{e}{a}$	I_{dist}	Col_M	Col_A	APF

Supplementary Table 6: The top 5 important factors between factors selected by BMARS for the minimum shear modulus G 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
$N = 2$	$\frac{e}{a}$	c	a	APF	I_{dist}
$N = 5$	$\frac{e}{a}$	Z	c	Col_A	C_v
$N = 10$	$\frac{e}{a}$	Z	APF	Col_A	c
$N = 15$	$\frac{e}{a}$	Z	Col_A	APF	C_v
$N = 20$	$\frac{e}{a}$	Z	C_v	APF	a

Supplementary Table 7 and Supplementary Table 8 present the top 5 significant interaction effects for the minimum of K using BART and BMARS, respectively.

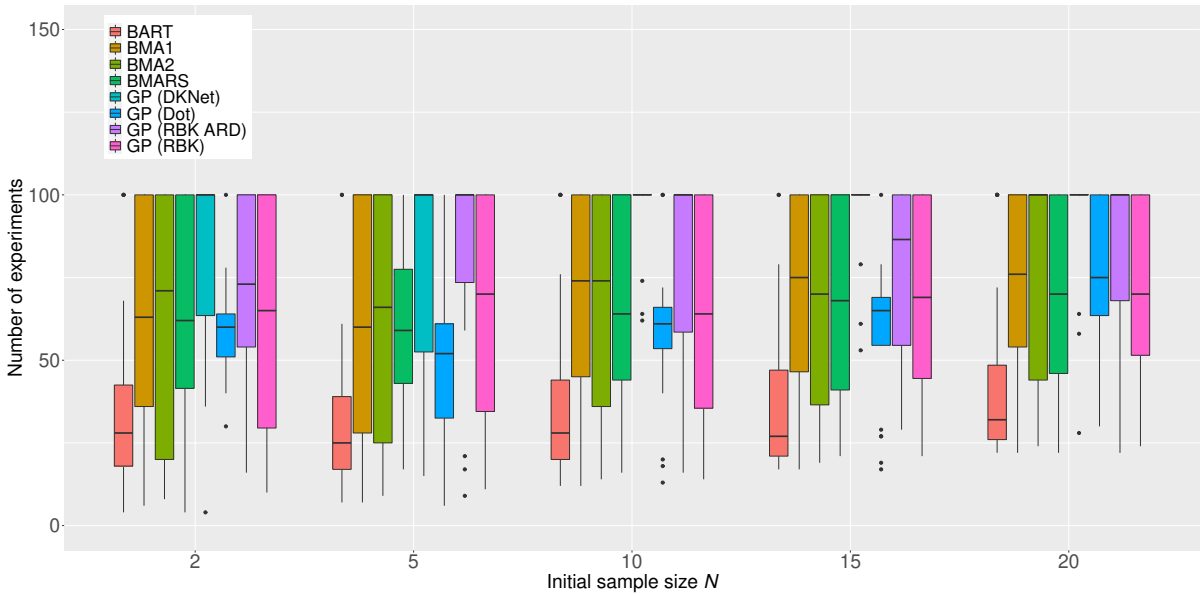
Supplementary Table 7: The top 5 important interactions between factors selected by BART for the minimum shear modulus G in MAX phases with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$.

Setting	Top 1	Top 2	Top 3	Top 4	Top 5
$N = 2$	$\frac{e}{a} : I_{\text{dist}}$	$\frac{e}{a}$: APF	$\text{Col}_A : \frac{e}{a}$	$c : \frac{e}{a}$	$a : \frac{e}{a}$
$N = 5$	$\frac{e}{a} : I_{\text{dist}}$	$\frac{e}{a}$: APF	$\text{Col}_A : \frac{e}{a}$	$c : \frac{e}{a}$	$Z : \frac{e}{a}$
$N = 10$	$\frac{e}{a}$: APF	$\frac{e}{a} : I_{\text{dist}}$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_M : \frac{e}{a}$	$c : \frac{e}{a}$
$N = 15$	$\frac{e}{a}$: APF	$\frac{e}{a} : I_{\text{dist}}$	$\text{Col}_A : \frac{e}{a}$	I_{dist} : APF	$\text{Col}_M : \frac{e}{a}$
$N = 20$	$\frac{e}{a}$: APF	$\text{Col}_A : \frac{e}{a}$	$\frac{e}{a} : I_{\text{dist}}$	I_{dist} : APF	$\text{Col}_M : \frac{e}{a}$

Supplementary Table 8: The top 5 important interactions between factors selected by BMARS for the minimum shear modulus G 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
$N = 2$	$a : \frac{e}{a}$	$Z : \frac{e}{a}$	$\frac{e}{a} : I_{\text{dist}}$	$\text{Col}_A : c$	$\frac{e}{a}$: APF
$N = 5$	$\text{Col}_A : Z$	$C : \text{Col}_A$	$C : c$	C : APF	c : APF
$N = 10$	$Z : \frac{e}{a}$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : c$	$c : \frac{e}{a}$	c : APF
$N = 15$	$\frac{e}{a} : C_v$	$Z : \frac{e}{a}$	$\text{Col}_A : a$	Z : APF	$\frac{e}{a}$: APF
$N = 20$	$a : \frac{e}{a}$	$Z : c$	$a : C_v$	$\text{Col}_A : \frac{e}{a}$	$m : \frac{e}{a}$

As shown from Supplementary Figure 5, BART's performance is not degraded by the newly added unhelpful information and is still the most efficient choice, indicating its robust property. GP (Dot) follows and also shows fast search speed. At the same time, although BMARS is relatively slower than the best, it is still competitive compared to BMA1, BMA2, GP (RBK), and GP (RBK ARD). While GP (DKNet) still doesn't work here.



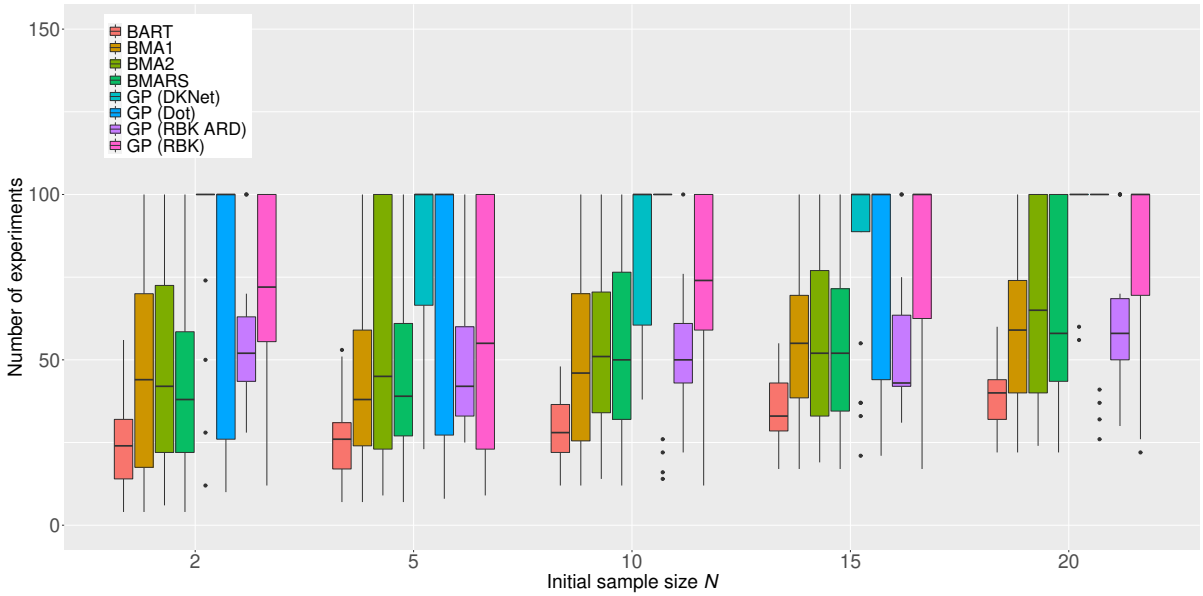
Supplementary Figure 5: Comparison of the number of experiments needed by BART, BMA1, BMA2, BMARS, GP (RBK), GP (RBK ARD), GP (Dot), and GP(DKNet) to find the maximum bulk modulus K with additional non-informative features with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$.

Supplementary Table 9 shows the mean value and interquartile range (IQR) of the total number of evaluations searching for the minimum shear modulus G within the MAX phase design space with additional non-informative features. Similarly, the smaller values of the mean and interquartile range mean a more efficient and stable workflow. As presented in Supplementary Table 9, when searching for the minimum shear modulus G, BART maintains the better performance compared to other models. For BMARS, it works as well as GP (RBK ARD), BMA1 and BMA2. But GP (RBK), GP (Dot), and GP (DKNet) become relatively slow.

Supplementary Table 9: The mean value and interquartile range (IQR) of the number of experiments based on each model to find the minimum shear modulus G 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	**24.12**	**25.48**	**28.78**	**34.28**	**38.8**	18	14	14.5	14.5	12
BMA1	45.42	**42.36**	**48.54**	54.36	**58.6**	52.5	35	44.5	31	34
BMA2	50.26	51.17	55.74	56.69	65.42	50.5	77	36.5	44	60
BMARS	**45.24**	46.6	55.5	**54.28**	64.36	36.5	34	44.5	37	56.5
GP (RBK)	72.02	56.3	74.24	82.1	83.36	77.46	76.95	76.86	73.04	75.3
GP (RBK with ARD)	55.3	49.6	52.1	53.1	62.4	19.5	27.	18.	21.5	18.5
GP (Dot)	70.15	68.95	83.9	78.6	86.8	74.	72.75	45.	56.	42.
GP (DKNet)	88.2	82.65	83.8	84.15	95.8	14.5	33.5	39.5	11.25	20

As presented in Supplementary Figure 6, BART keeps outperforming other models. And BMARS, BMA1, BMA2, GP (RBK ARD) follows and show similar ability. While the rest are more likely to run out of the iteration budget.



Supplementary Figure 6: Comparison of the number of experiments needed by BART, BMA1, BMA2, BMARS, GP (RBK), GP (RBK ARD), GP (Dot), and GP(DKNet) to find the minimum shear modulus G with additional non-informative features with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$.

For the interaction effects when looking for the maximum bulk modulus K in MAX phases with non-informative features as depicted in Supplementary Table 10, BART successfully neglects unimportant features and maintains its performance. At the same time, as in Supplementary Table 11, BMARS is capable of (almost) filtering out all the non-informative features and only leave a small portion of the interactions between new predictors and original data like $\text{Col}_A : n_3$ and $\frac{e}{a} : n_{14}$, indicating a good but relative less robust performance than that of BART.

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

Setting	Top 1	Top 2	Top 3	Top 4	Top 5
$N = 2$	$\text{Col}_A : I_{\text{dist}}$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : \text{Col}_M$	$\text{Col}_A : c$	$\text{Col}_A : \text{APF}$
$N = 5$	$\text{Col}_A : I_{\text{dist}}$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : \text{APF}$	$\text{Col}_A : c$	$\text{Col}_A : \text{Col}_M$
$N = 10$	$\text{Col}_A : I_{\text{dist}}$	$\text{Col}_A : c$	$\text{Col}_A : \text{Col}_M$	$\text{Col}_A : \frac{e}{a}$	$I_{\text{dist}} : \text{APF}$
$N = 15$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : I_{\text{dist}}$	$\frac{e}{a} : \text{APF}$	$\text{Col}_A : \text{APF}$	$\text{Col}_A : I_{\text{dist}}$
$N = 20$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : \text{Col}_M$	$\text{Col}_A : I_{\text{dist}}$	$\text{Col}_A : \text{APF}$	$\text{Col}_A : c$

Supplementary Table 11: The top 5 important interactions between factors for the maximum bulk modulus K 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
$N = 2$	$\text{Col}_A : \frac{e}{a}$	$c : \text{APF}$	$\frac{e}{a} : \text{APF}$	$\text{Col}_A : n_3$	$\text{Col}_A : I_{\text{dist}}$
$N = 5$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_M : \frac{e}{a}$	$I_{\text{dist}} : \text{APF}$	$a : n_8$	$\text{Col}_A : C_v$
$N = 10$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : I_{\text{dist}}$	$\text{Col}_A : c$	$\text{Col}_M : I_{\text{dist}}$	$\frac{e}{a} : n_{14}$
$N = 15$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_M : I_{\text{dist}}$	$\text{Col}_A : c$	$a : c$	$\text{Col}_A : \text{Col}_M$
$N = 20$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_A : c$	$\text{Col}_A : I_{\text{dist}}$	$c : \text{APF}$	$\text{Col}_A : C_v$

Supplementary Table 12 and Supplementary Table 13 list the top 5 important features with the goal of the minimum point of K with additional non-informative features using BART and BMARS, respectively.

Supplementary Table 12: The top 5 important factors for the minimum shear modulus G with additional non-informative features with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$ using BART.

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

Supplementary Table 13: The top 5 important factors for the minimum shear modulus G 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
$N = 2$	$\frac{e}{a}$	Col_A	APF	Z	n_8
$N = 5$	$\frac{e}{a}$	Col_A	APF	c	I_{dist}
$N = 10$	$\frac{e}{a}$	Col_A	APF	Z	n_8
$N = 15$	$\frac{e}{a}$	Col_A	APF	c	a
$N = 20$	$\frac{e}{a}$	Col_A	APF	n_8	Col_M

Supplementary Table 14 and Supplementary Table 15 present the top 5 significant interaction effects for the minimum of K with additional non-informative features using BART and BMARS, respectively.

Supplementary Table 14: The top 5 important interactions between factors for the minimum shear modulus G with additional non-informative features with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$ using BART.

Setting	Top 1	Top 2	Top 3	Top 4	Top 5
$N = 2$	$\text{Col}_A : \frac{e}{a}$	$\frac{e}{a} : \text{APF}$	$\text{Col}_M : \frac{e}{a}$	$\frac{e}{a} : I_{\text{dist}}$	$a : \frac{e}{a}$
$N = 5$	$\text{Col}_A : \frac{e}{a}$	$\frac{e}{a} : I_{\text{dist}}$	$\text{Col}_M : \frac{e}{a}$	$\frac{e}{a} : \text{APF}$	$Z : \frac{e}{a}$
$N = 10$	$\text{Col}_A : \frac{e}{a}$	$\frac{e}{a} : \text{APF}$	$\text{Col}_M : \frac{e}{a}$	$\frac{e}{a} : I_{\text{dist}}$	$\text{Type} : \frac{e}{a}$
$N = 15$	$\frac{e}{a} : \text{APF}$	$\text{Col\_}M : \frac{e}{a}$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_M : I_{\text{dist}}$	$\frac{e}{a} : I_{\text{dist}}$
$N = 20$	$\frac{e}{a} : \text{APF}$	$\frac{e}{a} : I_{\text{dist}}$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_M : \frac{e}{a}$	$I_{\text{dist}} : \text{APF}$

Supplementary Table 15: The top 5 important interactions between factors for the minimum shear modulus G 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
$N = 2$	$a : \frac{e}{a}$	$Z : \frac{e}{a}$	$m : \frac{e}{a}$	$\text{Col}_A : \frac{e}{a}$	$\frac{e}{a} : \text{Cv}$
$N = 5$	$\frac{e}{a} : \text{APF}$	$Z : \frac{e}{a}$	$\text{Col}_M : \frac{e}{a}$	$\frac{e}{a} : I_{\text{dist}}$	$\text{Col}_X : \text{Col}_M$
$N = 10$	$Z : \frac{e}{a}$	$\frac{e}{a} : n_{12}$	$\text{Col}_A : \frac{e}{a}$	$\text{Col}_M : \frac{e}{a}$	$a : Z$
$N = 15$	$\text{Col}_A : \frac{e}{a}$	$m : \frac{e}{a}$	$Z : \frac{e}{a}$	$a : \frac{e}{a}$	$\frac{e}{a} : \text{APF}$
$N = 20$	$Z : \frac{e}{a}$	$\text{Col}_M : \frac{e}{a}$	$\text{Col}_A : \frac{e}{a}$	$\frac{e}{a} : n_{15}$	$\text{Col}_A : c$

Supplementary Note 3

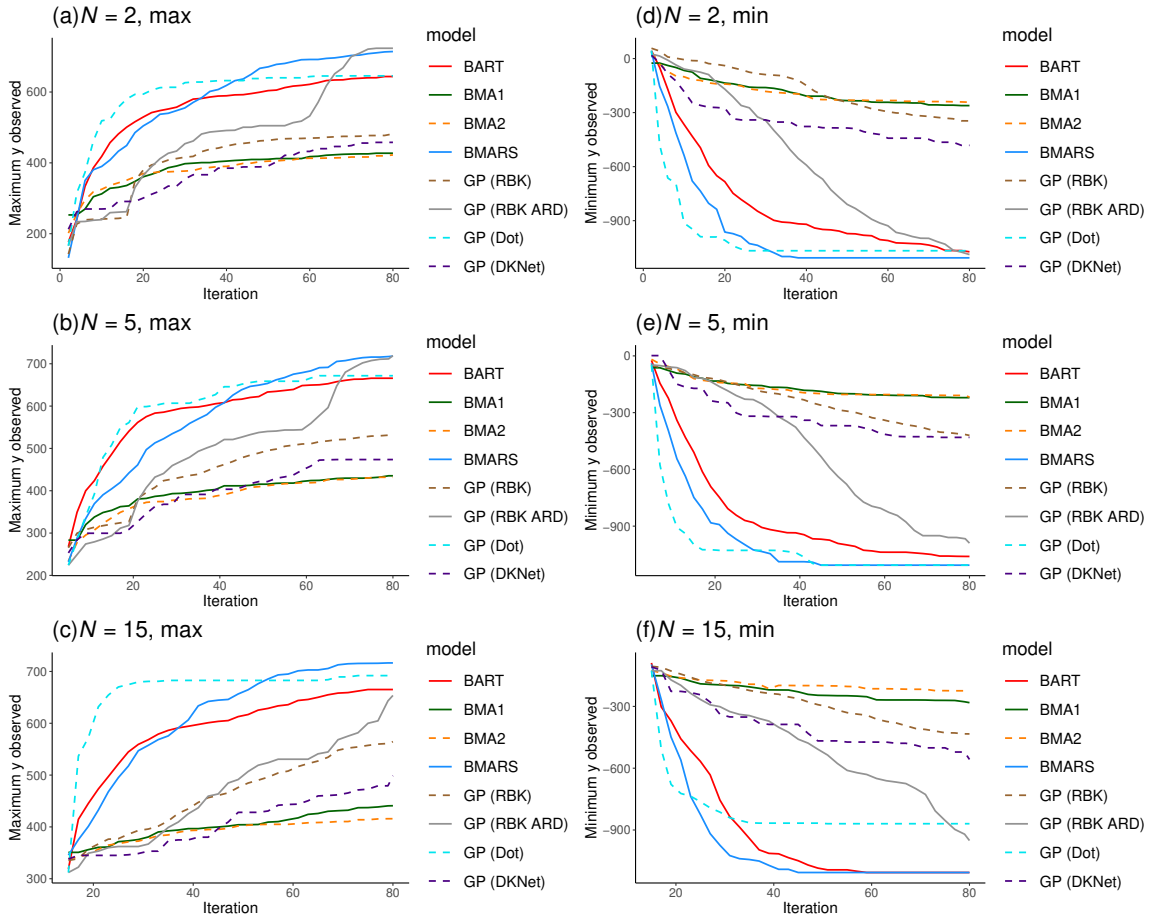
Here we show more results for the Optimal Design for Stacking Fault Energy in High Entropy Alloy Spaces.

As seen in Supplementary Figures 7 (a), (b), and (c), we can see the maximum y observed as 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 most efficient search. In Supplementary Figures 7 (d), (e), and (f), we can find the minimum y observed when searching for the minimum of SFE. The blue curves for BMARS and red curves for BART drop much faster than other curves, which proves a more efficient search ability of our methods.

Considering the interactions among features, we compare their significance by numbering the co-existence of two of them within each basis function. A higher co-appearance frequency suggests a bigger role for certain interactions in determining the interested property. Supplementary Tables 16 and 17 present the top 5 key interaction effects for the maximum value of SFE using BART and BMARS, respectively. As we can see, many of the key interactions are between the selected features in Tables 7 and 8 in the main text of the manuscript.

Supplementary Table 16: The top 5 important interactions selected by BART between factors for the maximum 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
$N = 2$	Pauling.EN_Var : C_11_Avg	V : Specific.Heat_Avg	SGTE.LSE_Avg : Pauling.EN_Var	VASP.PAW.GGA.LSE_Avg : Pauling.EN_Var	Pauling.EN_Var : Specific.Heat_Avg
$N = 5$	Pauling.EN_Var : C._Avg	Pauling.EN_Var : Specific.Heat_Avg	Pauling.EN_Var : C_11_Avg	Mn : Pauling.EN_Var	Melting.Point_Avg : Specific.Heat_Avg
$N = 10$	Pauling.EN_Avg : Specific.Heat_Avg	Co : Specific.Heat_Avg	Melting.Point_Avg : Specific.Heat_Avg	IE_Avg : Specific.Heat_Avg	VASP.PAW.GGA.LSE_Avg : Pauling.EN_Var
$N = 15$	Pauling.EN_Var : Specific.Heat_Avg	Mn : SGTE.LSE_Avg	Mn : Pauling.EN_Var	V : Specific.Heat_Avg	Melting.Point_Avg : Specific.Heat_Avg
$N = 20$	Pauling.EN_Var : Melting.Point_Avg	Pauling.EN_Var : Specific.Heat_Avg	Mn : Pauling.EN_Avg	Mn : Co	Allen.EN_Avg : Specific.Heat_Avg



Supplementary Figure 7: The maximum SFE $[\text{mJ}\cdot\text{m}^{-2}]$ observed in each iteration with the initial set of sample size (a) $N = 2$, (b) $N = 5$, and (c) $N = 15$. The minimum SFE $[\text{mJ}\cdot\text{m}^{-2}]$ observed in each iterations with the initial set of sample size (d) $N = 2$, (e) $N = 5$, and (f) $N = 15$.

Supplementary Table 17: The top 5 important interactions selected by BMARS between factors for the maximum 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
$N = 2$	Metallic.Radius_Var : Mn	Metallic.Radius_Var : Fe	Atomic.Planar.Density_Avg : Fe	Al : Mn	Allen.EN_Avg : Mn
$N = 5$	Pauling.EN_Var : Ni	Melting.Point_Var : V	Melting.Point_Var : Ni	Density_Var : Cr	Metallic.Radius_Var : VASP.PAW.GGA.LSE_Avg
$N = 10$	Melting.Point_Var : Fe	VASP.PAW.GGA.LSE_Var : Mn	Total.Electrons_Var : Mn	Atomic.Weight_Var : Atomic.Weight_Avg	Total.Electrons_Var : IE_Var
$N = 15$	C_12_Avg : Specific.Heat_Var	Specific.Heat_Var : Co	Melting.Point_Avg : Allen.EN_Avg	Total.Electrons_Avg : Atomic.Weight_Var	C._Var : C_12_Avg
$N = 20$	VASP.PAW.GGA.LSE_Avg : Al	Melting.Point_Var : Cr	C_11_Avg : Al	SGTE.LSE_Var : VASP.PAW.GGA.LSE_Var	C._Avg : Density_Var

Under 5 different scenarios $N \in \{2, 5, 10, 15, 20\}$, Supplementary Tables 18 and 19 lists the top 5 factors most correlated to the minimum in the SFE using BART and BMARS, respectively.

Supplementary Table 18: The top 5 important factors for the minimum SFE with the initial set of sample size $N \in \{2, 5, 10, 15, 20\}$ using BART.

Setting	Top 1	Top 2	Top 3	Top 4	Top 5
$N = 2$	C._Avg	SGTE.LSE_Avg	C_11_Var	C_11_Avg	Fe
$N = 5$	C._Avg	C_11_Avg	VASP.PAW.GGA.LSE_Avg	SGTE.LSE_Avg	Pauling.EN_Var
$N = 10$	C._Avg	C_12_Avg	SGTE.LSE_Avg	VASP.PAW.GGA.LSE_Avg	Fe
$N = 15$	C_12_Avg	SGTE.LSE_Avg	C._Avg	C_11_Avg	Valence.e_Var
$N = 20$	C._Avg	C_11_Avg	C_12_Avg	SGTE.LSE_Avg	VASP.PAW.GGA.LSE_Avg

Supplementary Table 19: The top 5 important factors for the minimum SFE 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
$N = 2$	C._Avg	Atomic.Planar.Density_Var	Specific.Heat_Avg	Al	Allen.EN_Var
$N = 5$	SGTE.LSE_Avg	C_11_Var	Allen.EN_Var	Metallic.Radius_Var	Fe
$N = 10$	Specific.Heat_Avg	Fe	C_11_Avg	Valence.e_Var	Pauling.EN_Var
$N = 15$	Specific.Heat_Avg	C._Avg	VASP.PAW.GGA.LSE_Avg	C_12_Avg	C_11_Avg
$N = 20$	Specific.Heat_Avg	C._Avg	Pauling.EN_Var	SGTE.LSE_Avg	Valence.e_Var

For 5 different initial sample sizes $N \in \{2, 5, 10, 15, 20\}$, Supplementary Tables 20 and 21 presents the top 5 significant interactions for the minimum in the SFE using BART and BMARS, respectively.

Supplementary Table 20: The top 5 important interactions between factors for the minimum SFE with initial set of sample size $N \in \{2, 5, 10, 15, 20\}$ using BART.

Setting	Top 1	Top 2	Top 3	Top 4	Top 5
$N = 2$	SGTE.LSE_Avg : C_11_Avg	Specific.Heat_Var : C_12_Avg	VASP.PAW.GGA.LSE_Avg : C_11_Avg	Fe : SGTE.LSE_Avg	IE_Var : C_11_Avg
$N = 5$	C_12_Var : C._Avg	C_12_Avg : C._Avg	Fe : SGTE.LSE_Avg	Atomic.Weight_Var : C_12_Avg	C_11_Var : C_12_Avg
$N = 10$	C_11_Avg : C._Avg	Fe : C._Avg	C_12_Avg : C._Avg	SGTE.LSE_Avg : C._Avg	Fe : C_11_Avg
$N = 15$	SGTE.LSE_Avg : C._Avg	SGTE.LSE_Avg : SGTE.LSE_Var	SGTE.LSE_Avg : C_12_Avg	C_12_Avg : C._Avg	Fe : C._Avg
$N = 20$	C_11_Avg : C._Avg	C_12_Avg : C._Avg	SGTE.LSE_Avg : C_11_Avg	Specific.Heat_Avg : C._Avg	C_11_Avg : C_12_Avg

Supplementary Table 21: The top 5 important interactions between factors for the minimum SFE 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
$N = 2$	Atomic.Planar.Density_Var : Pauling.EN_Avg	Pauling.EN_Var : Ni	Specific.Heat_Avg : Co	Shear.Modulus_Var : Valence.e_Var	Shear.Modulus_Avg : Specific.Heat_Var
$N = 5$	IE_Var : Ni	Total.Electrons_Var : Co	Density_Var : V	Specific.Heat_Avg : V	Allen.EN_Var : VASP.PAW.GGA.LSE_Var
$N = 10$	Specific.Heat_Avg : VASP.PAW.GGA.LSE_Avg	C_11_Avg : IE_Avg	Melting.Point_Var : Valence.e_Avg	Cr : Fe	Valence.e_Avg : Fe
$N = 15$	C_11_Avg : Cr	C._Avg : SGTE.LSE_Avg	Specific.Heat_Avg : Density_Avg	C_12_Avg : Density_Var	IE_Avg : Atomic.Weight_Avg
$N = 20$	Atomic.Weight_Var : SGTE.LSE_Var	Atomic.Planar.Density_Var : Al	IE_Avg : VASP.PAW.GGA.LSE_Avg	C._Var : C_11_Avg	Density_Avg : Fe

Supplementary Note 4

With regards to the practical implementation of Bayesian optimization (BO) on a realistic setting, we note that any BO approach to optimization ultimately operates on a discrete space as the metric indicating utility of a particular observation is evaluated on an actual design point. When the space to explore is a continuous (black box) function, the identification of optimal points is carried out either by sampling the value of the utility function on a fixed grid, random sampling, space-filling sampling or through numerical optimization. Either way, only a finite number of choices is considered at any one time. If the black-box

function to explore is not a function (or model), but rather, a 'real' experimental space, with finite design choices, BO works by evaluating the utility of carrying out experiments at each of the design points yet to be evaluated, given the data that has been acquired thus far, as well as the model being used to predict the outcome of yet-to-be-realized experiments. The next experiments to carry out are chosen such that this utility is maximized.

Supplementary References

[1] Chipman, H. A., George, E. I., McCulloch, R. E. Bayesian CART model search. *Journal of the American Statistical Association* **93**, 935–948 (1998).

[2] Chipman, H. A., George, E. I., McCulloch, R. E. *et al.* Bart: Bayesian additive regression trees. *The Annals of Applied Statistics* **4**, 266–298 (2010).

[3] Denison, D. G., Mallick, B. K. & Smith, A. F. Bayesian MARS. *Statistics and Computing* **8**, 337–346 (1998).

[4] Green, P. J. Reversible jump markov chain monte carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732 (1995).

[5] Talapatra, A. *et al.* Autonomous efficient experiment design for materials discovery with Bayesian model averaging. *Physical Review Materials* **2**, 113803 (2018).