**Submitted to *INFORMS Journal on Data Science***
**manuscript**

# Adaptive Exploration and Optimization of Materials Crystal Structures

Arvind Krishna*, Huan Tran†, Chaofan Huang‡, Rampi Ramprasad†, V. Roshan Joseph‡
*Department of Statistics and Data Science, Northwestern University, Evanston, IL 60208
†School of Material Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332,
‡H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332
krish@northwestern.edu, huan.tran@mse.gatech.edu, chaofan.huang@gatech.edu, rampi.ramprasad@mse.gatech.edu,
roshan@gatech.edu

A central problem of materials science is to determine whether a hypothetical material is stable without being synthesized, which is mathematically equivalent to a global optimization problem on a highly non-linear and multi-modal potential energy surface (PES). This optimization problem poses multiple outstanding challenges, including the exceedingly high dimensionality of the PES and that PES must be constructed from a reliable, sophisticated, parameters-free, and thus, very expensive computational method, for which density functional theory (DFT) is an example. DFT is a quantum mechanics based method that can predict, among other things, the total potential energy of a given configuration of atoms. DFT, while accurate, is computationally expensive. In this work, we propose a novel expansion-exploration-exploitation framework to find the global minimum of the PES. Starting from a few atomic configurations, this "known" space is expanded to construct a big candidate set. The expansion begins in a non-adaptive manner, where new configurations are added without considering their potential energy. A novel feature of this step is that it tends to generate a space-filling design without the knowledge of the boundaries of the domain space. If needed, the non-adaptive expansion of the space of configurations is followed by adaptive expansion, where "promising regions" of the domain space (those with low energy configurations) are further expanded. Once a candidate set of configurations is obtained, it is simultaneously explored and exploited using Bayesian optimization to find the global minimum. The methodology is demonstrated using a problem of finding the most stable crystal structure of Aluminum.

*Key words*: Active learning, Adaptive design, Bayesian optimization, Computer experiments, Crystal structure prediction, Gaussian process model, Space-filling design.

## 1. Introduction

One of the most ambitious goals of material scientists is to discover and design new materials with desirable properties and applications (Franceschetti & Zunger 1999, Weymuth & Reiher 2014, d'Avezac et al. 2012, Xiang et al. 2013, Huan et al. 2015, Mannodi-Kanakkithodi et al. 2016). Until the present time, material discoveries are largely driven by expensive and time-consuming trial-and-error approaches, i.e., they must be physically synthesized and tested in a laboratory with limited guidance beyond empirical rules and experience. However, under some scenarios, some properties of a material can be computed without synthesizing it, if its atomic structure is known.

Predicting the stable atomic configurations of a given set of atoms can be mathematically formulated as an optimization problem. The most stable configuration is the global minimum of the potential energy surface corresponding to all possible atomic configurations. This is a very active research area in the emerging era of materials discovery and design, when a large number of hypothetical materials should be examined by computational methods before some of them can be advanced to the synthesizing and testing steps. The main objective of materials structure prediction (Oganov et al. 2019) is searching for low-energy atomic configurations of a given set of atoms.

Although we are primarily interested in the global minimum of the PES, certain local minima may also be useful (Vu et al. 2021, Therrien et al. 2021). External perturbations such as temperature, pressure, and other kinetic-related factors may bring a local minimum down to be the global minimum at a specific condition (Huan 2018, Kobayashi et al. 2019, Gaida et al. 2021), or drive the atomic configuration to land at some nearby (accessible) local minima. Therefore, configurations that are very far from (and/or very well-separated by a high potential energy barrier with) the global minimum may also be reliable.

The specific class of materials addressed in this work is crystal. A crystal can be imagined as an infinitely repeated array of a unit cell along three Cartesian dimensions. Crystal materials are dominant in material science because of two main reasons. First, a majority of materials are crystals and/or can be modeled very well by crystal models. Second, because of its periodicity, crystal models are small enough so that physics-based computational methods such as the Density Functional Theory (DFT) (Hohenberg & Kohn 1964, Kohn & Sham 1965), the most reliable (but expensive) parameter-free computational method, may be used at an acceptable cost.

The least biased and non-empirical approaches to crystal structure prediction involve computational optimization (Oganov 2011). These approaches involve explicit computation of the potential energy of the crystal structure, followed by solving an optimization problem to find the crystal structure corresponding to the least energy, or the thermodynamically most stable configuration. Pickard & Needs (2006, 2011) developed a random-search based method to find the stable crystal structure configuration. The underlying idea in this method is to use the DFT to optimize a randomly generated set of crystal structures, driving each of them to the nearest local minimum. With a large number of randomly generated samples, these approaches can successfully identify the most stable crystal structure configuration in many cases. Some other recently developed methods are simulated annealing (Pannetier et al. 1990, Schön & Jansen 1996, Tekin et al. 2010), basin hopping (Wales & Doye 1997), minima hopping (Goedecker 2004), metadynamics (Martoňák et al. 2006), evolutionary algorithms (Trimarchi et al. 2009), and USPEX (Universal Structure Predictor: Evolutionary Xtallography) approach (Oganov & Glass 2006, Glass et al. 2006, Oganov et al. 2011), which is based on evolutionary algorithms.

While these methods are different in many aspects, most notably the employed (global) optimization algorithms, they do share two common fundamental problems. First, given a set of atoms, how to thoroughly explore the configurational space, and second, within this accessible domain, how to efficiently identify the global minimum? Since the number of local minima of the potential energy scales up exponentially with the number of atoms in the system (Berry 1993, Stillinger 1999), both problems are enormously challenging. In most of the practical cases, there is essentially no way to guarantee that the entire configurational space can be explored, and for this reason, new developments in this active research area are still in progress.

We have developed an expansion-exploration-exploitation framework to address these problems, i.e., (i) enlarging the accessible domain of the search space, and (ii) finding the global minimum of the PES within this domain. For the first problem, we expand the space spanned by a few possible configurations by perturbing them and generating more configurations in their neighborhood. The configurations are generated such that they continuously expand the spanned domain space of configurations, especially towards the low-energy regions of the domain space. Once a representative candidate set of configurations is obtained, a Bayesian optimization procedure (Jones et al. 1998) is used for exploring the domain space regions with high uncertainty in the potential energy estimate while simultaneously exploiting the low-energy regions to find the global minimum and reliable local minima among the candidate set of configurations.

This article is organized as follows. In Section 2, we discuss the characteristics of the crystal structure prediction problem, and the associated constraints and challenges. In Section 3, we describe the developed methodology that addresses these constraints and challenges. In Section 4, we illustrate the effectiveness of our methodology on the problem of finding the crystal structure of $Al_8$ (Aluminum), where the true structure is already known (see Figure 1). We conclude the article with some remarks in Section 5.

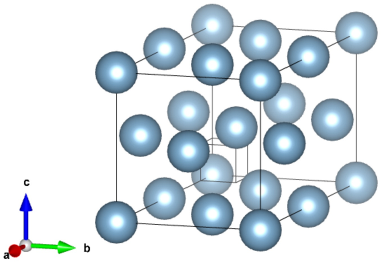

**Figure 1** A supercell of the body-centered cubic (ground state) crystal structure of Aluminium. The ideal crystal is obtained by infinitely repeating this cell in three dimensions.

## 2. Problem characteristics, constraints and challenges

This section describes the characteristics of the crystal structure prediction problem, and the associated constraints and challenges, as summarised in Figure 2.

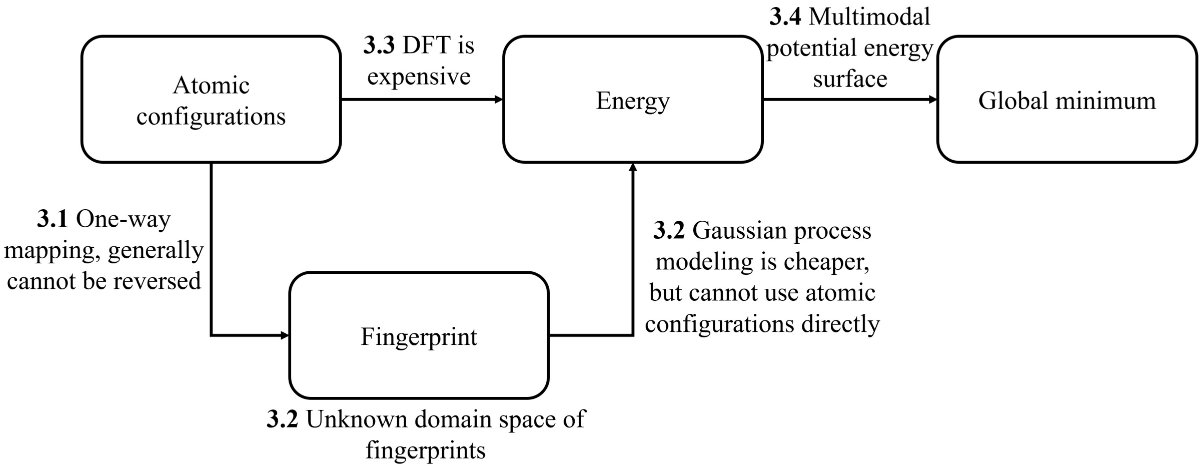
Atomic configurations
**3.3** DFT is expensive
Energy
**3.4** Multimodal potential energy surface
Global minimum
**3.1** One-way mapping, generally cannot be reversed
**3.2** Gaussian process modeling is cheaper, but cannot use atomic configurations directly
Fingerprint
**3.2** Unknown domain space of fingerprints

**Figure 2** Problem characteristics and the associated constraints and challenges.

### 2.1. Expensive energy computation: Density functional theory (DFT)

We use single-point DFT computations to determine the potential energy of the atomic configuration under consideration. Such calculations are performed using the ABINIT package (Gonze et al. 2016), employing the Perdew-Burke-Ernzerhof functional (Perdew et al. 1996) for the quantum mechanical exchange-correlation energies. The electron-nuclear interactions are computed with help from the norm-conserving Hartwigsen-Goedecker-Hutter pseudopotentials (Hartwigsen et al. 1998). For our calculations, the Brilouin zone is sampled by a dense Monkhorst-Pack k-point mesh (Monkhorst & Pack 1976), and a basis set of plane waves with kinetic energy up to 550 eV.

To find the most stable crystal structure configuration, we need to find the one with the least potential energy. Any reliable predictions of materials structure must be done with accurate-enough methods to compute the potential energy, and DFT is possibly the least expensive method of this kind. However, the time needed for a DFT computation scales as $N_A^3$ where $N_A$ is the number of atoms. It also becomes quickly expensive if numerous other parameters are adjusted to reach higher levels of accuracy. Each evaluation of the potential energy on supercomputers requires hours or even days, depending on the size of the atomic system under investigations. This is why predicting a simple crystal structure by computations was regarded as "*one of the continuing scandals in the physical sciences*" in 1988 by a Nature's editor, Sir John Maddox (Maddox 1988). Although structure prediction methods have evolved dramatically since then and have led to numerous new materials predicted computationally and realized experimentally (Oganov et al. 2019), this remains a major bottleneck of contemporary materials discoveries. The expensive DFT computations constrain us to evaluate the energy for only a few configurations, which gives rise to the challenge of optimizing a huge potential energy surface, while observing it at only a few points.

### 2.2. Crystal structure representation: One-sided mapping

A crystal model of a material includes a parallelepiped unit cell defined by three basis vectors $\vec{a}$, $\vec{b}$, and $\vec{c}$, a given set of $N_A$ atoms arranged in the unit cell, and an assumption that the unit cell is infinitely repeated along $\vec{a}$, $\vec{b}$, and $\vec{c}$. Figure 1 shows an example of a unit cell. Because a material does not change under rigid translations and rotations, three vectors $\vec{a}$, $\vec{b}$, and $\vec{c}$ can be uniquely determined by six independent numbers. Therefore, the crystal structure prediction is mathematically equivalent to a global optimization problem on the PES defined in a $3N_A + 6$ dimensional space ($N_A$ has no upper limit, and its typical values can be as high as 100).

As mentioned earlier, we use the DFT to compute the potential energy of a crystal structure. Potential energy of a crystal structure depends only on the relative distance between its atoms, and not on the absolute positions of atoms. This implies that the energy is invariant to translational, rotational and permutational operations of alike atoms, in the crystal structure configuration. Such transformations change the Cartesian coordinates of the atom, but do not change the material in any physical and chemical way. For this reason, we used the Cartesian coordinate system for energy computations with DFT, while using the *AGNI* (Adaptive, Generalizable and Neighborhood Informed) fingerprint (Batra et al. 2019) to map the atomic configurations onto their energy. All the redundant Cartesian coordinate system atomic configurations correspond to a unique *AGNI* fingerprint. An *AGNI* fingerprint captures the atomic-level information of the structure pretty well while preserving the material presentation under such "identity" transformations in the materials space.

The *AGNI* fingerprint used in this work is defined as $f := \{S_k; V_k\}_{k=1}^{n}$, where the scalar components $S_k$ and the vectorial components $V_k$ are given by

$$S_k = \sum_{i \neq j} \mathcal{G}(r_{ij}, \sigma_k) f_c(r_{ij}), \tag{1}$$

and

$$V_k = \sqrt{\sum_{\alpha=x,y,z} \left[ \sum_{i \neq j} \frac{r_{ij}^{\alpha}}{r_{ij}} \mathcal{G}(r_{ij}, \sigma_k) f_c(r_{ij}) \right]^2}, \tag{2}$$

respectively. Here, $r_{ij}$ is the distance between atoms $i$ and $j$, $r_{ij}^{\alpha}$ is the projection of $r_{ij}$ onto the Cartesian axis $\alpha$, $\mathcal{G}(r, \sigma_k)$ is the Gaussian function centered at 0 with varying width $\sigma_k$:

$$\mathcal{G}(r, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left( \frac{-r^2}{2\sigma_k^2} \right), \tag{3}$$

and $f_c(.)$ is a cutoff function that is used for disregarding interaction among atoms that are further than a distance $R_c$ from each other:

$$f_c(r) \equiv \frac{1}{2} \left[ \cos(\pi r / R_c) + 1 \right]. \tag{4}$$

While summarizing over the atoms $i$ and $j$, the periodicity of the unit cell is considered, i.e., for an atom, its neighbors in all the repeated images of the unit cell are also taken into account. The cutoff function, $f_c(r)$, defined in (4) is used for restricting the neighborhood to a radius of $R_c$. We used $R_c = 8$ Å in this work, because the interaction between two atoms at this distance is negligible. From the mathematical point of view, *AGNI* fingerprint is a way of projecting the atomic positions onto a set of predefined basis functions. Here, we used $n = 16$ functions $\mathcal{G}(r, \sigma_k)$, thus our fingerprint $f$ has $2n = 32$ components or dimensions. Note that the accuracy of the model using a fingerprint increases as the dimension increases and then saturates. Our tests indicate that after 32 dimensions, the increase in model accuracy with increasing dimensions is negligible. The *AGNI* fingerprint is one of the numerous material fingerprints (Behler & Parrinello 2007, Bartók et al. 2010) developed during the last decade.

Although the *AGNI* fingerprint eliminates redundancy in the Cartesian coordinate representation of the crystal structure configuration, and reduces the PES dimensionality from $3N_A + 6$ (as described in Section 2) to 32, it introduces a constraint. The constraint is that the *AGNI* fingerprint function, as defined in (1) and (2), is non-invertible, i.e., we can only map a configuration in the Cartesian coordinate system to the *AGNI* system, but not vice-versa.

In the absence of the above constraint, we could have used a continuous optimization procedure in the *AGNI* fingerprint space to find a solution and transform it to the physically interpretable Cartesian coordinate system. However, in the presence of this constraint, we will need to use a discrete optimization approach of considering a candidate set of configurations, finding their corresponding fingerprints, and then finding the one with the least energy. This gives rise to the challenge that the candidate set of fingerprints must contain the solution(s) or fingerprints "close enough" to the solution(s).

Tripathy et al. (2016), Siivola et al. (2021), Wang et al. (2022) develop the Gaussian process model and optimize it in the latent space. However, we cannot use these methods as there is no reverse mapping from the *AGNI* fingerprint space (latent space) back to the original input space (Cartesian coordinate system). Chen et al. (2020) proposed formulating a least square problem to map the solution back to the input space. However, this becomes a discrete optimization problem in our case, which is computationally very expensive to solve.

### 2.3. Crystal structure representation: Unknown domain space

The domain space of the crystal structure configuration is more intuitive than that of the *AGNI* fingerprint. This gives rise to the challenge of obtaining a candidate set of fingerprints that are representative of all fingerprints, or a candidate set of crystal structure configurations that are representative of all possible configurations. If we knew the domain space, we could have used a space-filling design (Joseph 2016) to obtain a representative candidate set of fingerprints. However, the challenge is to find a representative set of fingerprints without the knowledge of their domain space.

There has been some work done for performing Bayesian optimization in an unknown input space (Shahriari et al. 2016, Nguyen et al. 2017, Ha et al. 2019). However, these methods are not applicable to our problem due to several reasons. First, these methods systematically expand the input space to search for the optimum. But, in our problem, we do not expand the input space (i.e., the Cartesian coordinate system space) directly. We map the input space to a feature space (the *AGNI* fingerprint space), which is expanded systematically to search for the optimum. Second, we cannot avoid the input space, and apply these methods directly on the feature space because a solution in the feature space cannot be mapped back to the input space, and the solution is to be found in the input space. Third, we cannot avoid the feature space and apply these methods on the input space. With the large amount of redundancy in the input space, the number of minima will be too large making the problem unnecessarily complex. There has been some work done to identify the feasible domain space when the search space is unknown. Basudhar & Missoum (2008) used Support Vector Machines, while Chen & Fuge (2017) used active learning to identify the feasible domain in an unbounded space. However, in our problem, it is not useful to identify the feasible feature space because the points in this space cannot be mapped back to the input space. Even though we expand the feature space, the expansion must be driven from the input space so that we can trace the solution in the feature space back to the input space.

### 2.4. Multi-modal potential energy surface (PES)

The potential energy surface is highly nonlinear and multi-modal. Given that we can observe it at only a few points, due to the expensive DFT computations, it becomes challenging to accurately model all the modalities. As the number of local minima scale up exponentially with the number of atoms in a unit cell, $N_A$, the challenge is even bigger for crystal structures with large $N_A$. However, modeling the multi-modalities is necessary to find the global minimum as well as other reliable local minima.

## 3. Methodology

We have developed an expansion-exploration-exploitation framework for crystal structure prediction that addresses all the challenges presented in Section 3. This framework is implemented in two steps. The first step is *domain space expansion*, where we expand the space spanned by a few possible configurations by iteratively adding more configurations. This leads to a candidate set of configurations that will ideally either span the entire domain space of possible configurations or at least span the space of stable configurations. The expansion step consists of a sequence of two sub-steps: *non-adaptive expansion* and *adaptive expansion.* Non-adaptive expansion refers to adding configurations without considering their potential energy. This tends to include unexpected configurations in our candidate set. If needed, this step is followed by adaptive expansion, which tends to add configurations that further expand the low-energy regions of the domain space. The expansion step is followed by simultaneous exploration and exploitation of the domain space spanned by the candidate set to find the configuration that corresponds to the minimum potential energy. We will explain these steps in the three sub-sections below.

In the sub-sections below, the number of initial configurations are denoted as $n_0$, the number of configurations added in the non-adaptive and adaptive expansion steps are denoted as $n_1$ and $n_2$ respectively. The cumulative number of configurations in the candidate set at the end of the non-adaptive and adaptive expansion steps are denoted as $N_1$ and $N_2$ respectively. The number of iterations in the Bayesian optimization procedure is denoted as $n_3$. As no new configurations are added in the Bayesian optimization procedure, the number of configurations at the end of our methodology remain $N_2$.

### 3.1. Non-adaptive domain space expansion

The purpose of this step is to obtain a candidate set of configurations that span as much domain space as possible. We start from a set of few possible configurations, and iteratively add those configurations to the set that expand their spanned domain space. The potential energy of the configurations is ignored, while developing the candidate set, to serve two purposes. First, it may lead us to regions of the domain space where a low-energy configuration is unexpected. Second, it saves the computational resources for calculating the energy and helps us obtain a larger candidate set within a given time period.

We will explain the algorithm with a toy example. Let the fingerprint domain space, be $[-3,3] \times [-3,3]$. However, in practice, we are not aware of the fingerprint domain space. So, we will not feed this domain space to our algorithm. Nevertheless, the objective of our algorithm will be to find a candidate set of fingerprints that fill this space.

The algorithm begins by considering the initial candidate set of few possible atomic configurations, say $\mathcal{C} = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_{n_0}\}$, where $\boldsymbol{c}_1, \cdots, \boldsymbol{c}_{n_0}$ are the $n_0$ initial atomic configurations. Let their corresponding fingerprints be $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n_0}\}$. Let us assume that there are a set of $n_0 = 5$ possible fingerprints for our toy example, as shown in Figure 3 (left), and we have a budget of expanding it to $N_1$ fingerprints.

To expand the domain space spanned by the fingerprints, we will identify the most sparsely populated region of the domain space, and generate fingerprints around it. We define the most sparsely located fingerprint as the one that has the farthest nearest neighbor in any of its neighborhoods. A fingerprint-neighborhood is defined as the space on either side of the fingerprint along each dimension. Thus, for a $p$-dimensional fingerprint, there are $2p$ neighborhoods - two on either side of it along each dimension. We intend to identify the fingerprint that has the farthest nearest neighbor in any of its $2p$ neighborhoods.

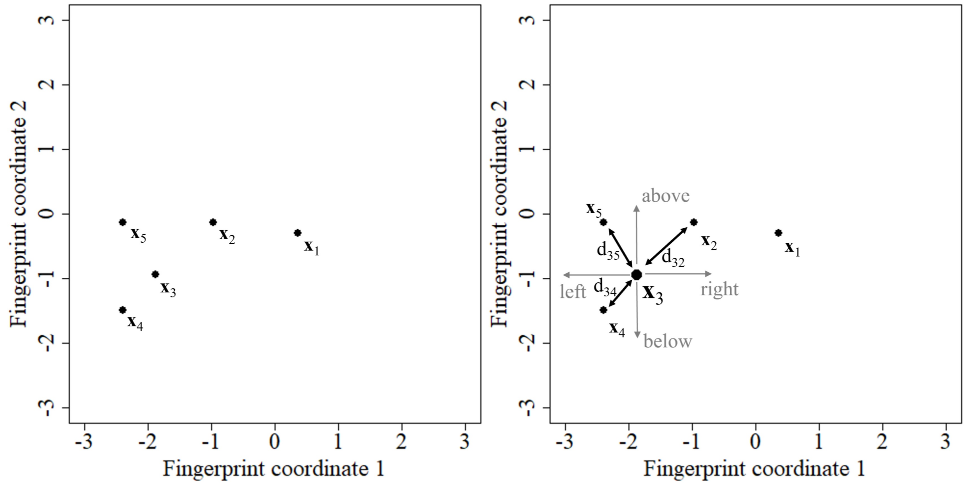

**Figure 3 (left): Initial $n_0 = 5$ fingerprints; (right): Distances to the closest neighbors of fingerprint $\boldsymbol{x}_3$ in all of its $2p = 4$ neighborhoods.**

Consider the two-dimensional fingerprints in Figure 3. Let us find the distance to the farthest nearest neighbor of fingerprint $\boldsymbol{x}_3$ in Figure 3 (right). For that we will find the distance to the nearest neighbor in $2p = 2 \times 2 = 4$ neighborhoods - above and below $\boldsymbol{x}_3$, and right and left of $\boldsymbol{x}_3$. The nearest neighbors above and below are $\boldsymbol{x}_3$ are at a distance of $d_{35} = 0.96$ and $d_{34} = 0.75$ respectively, while those on the left and right are at distance of $d_{34} = 0.75$ and $d_{32} = 1.22$ respectively. Thus, the farthest nearest neighbor of $\boldsymbol{x}_3$ is at a distance of $\max(d_{32}, d_{34}, d_{35}) = 1.22$.

Figure 4 (left) visualizes the distance to the farthest nearest neighbor with a circle having radius half of that distance, around each fingerprint. Clearly, the fingerprint $\boldsymbol{x}_1$ is the most sparsely located fingerprint. Let us label the fingerprint $\boldsymbol{x}_1$ as $\boldsymbol{x}_{sparse}$. So, we will find a fingerprint in the space around $\boldsymbol{x}_{sparse}$, and add it to the candidate set to expand the spanned domain space.

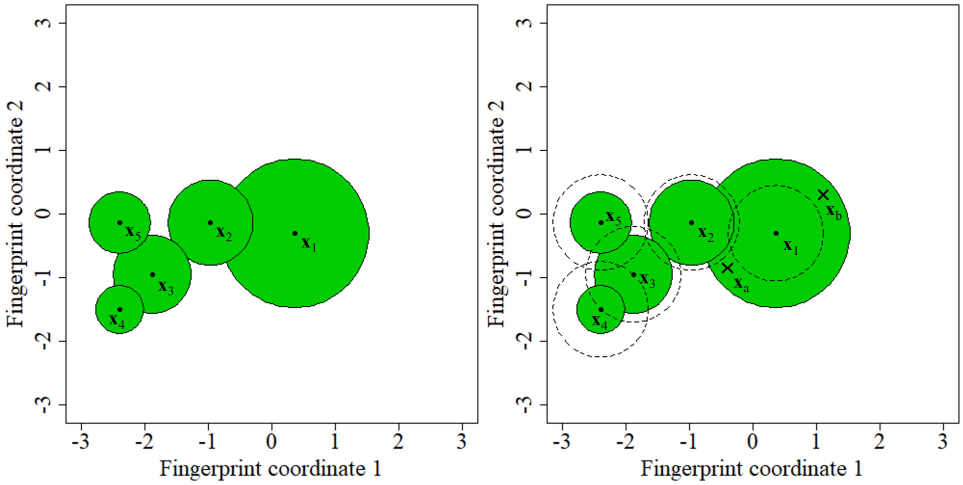

**Figure 4 (left): Space spanned by the initial $n_0 = 5$ fingerprints; (right): Dotted circle around each fingerprint with radius $t$, showing the minimum distance necessary between them and an acceptable newly generated fingerprint; Two examples of acceptable new fingerprints for inclusion in the candidate set - $\boldsymbol{x}_a, \boldsymbol{x}_b$.**

To generate a fingerprint around $\boldsymbol{x}_{sparse}$, we randomly perturb the atomic configuration corresponding to $\boldsymbol{x}_{sparse}$ to generate another configuration, and fingerprint it. This new fingerprint is added to the candidate set if it is “far enough” from its nearest neighboring fingerprint. We use a threshold distance $t_i$ (for the *ith* iteration), which will be defined later, to check if the randomly generated fingerprint is “far enough”. If it is not “far enough”, then we discard it, and again perturb the same atomic configuration.

The purpose of the threshold distance $t_i$ is twofold. First, it is used for avoiding redundancy of fingerprints in the candidate set. Second, it ensures that the fingerprints are evenly spaced-out in their domain space. If $t_i$ is the threshold distance in the *ith* iteration, and $d_{min,i}$ is the distance of the new fingerprint to its nearest neighbor in this iteration, then the threshold distance for the next iteration is given by:

$$t_{i+1} = \frac{t_i + d_{min,i}}{2}, \ \forall i > 1. \tag{5}$$

The term $d_{min,i}$ ensures that the threshold distance is large when large parts of the domain space are unexplored, and small if the domain space is already well-explored. This makes the fingerprints spread farther apart until the entire domain space has been explored. Once the domain space has been explored, the threshold distance decreases so that new fingerprints may be added to the candidate set, until the budget of $N_1$ fingerprints is exhausted. The term $t_i$ ensures that the threshold distance does not change abruptly for an abrupt change in $d_{min,i}$. For the first iteration, $i = 1$, $t_1$ is taken as the mean of the distances to the nearest neighboring fingerprint for each fingerprint.

In our toy example, $\boldsymbol{x}_1$ is the fingerprint identified with the farthest nearest neighbor. So, we perturb the atomic configuration corresponding to it to generate another one, and fingerprint it. In all the examples of this Section, we perturb a fingerprint using a multivariate normal distribution. The mean of the distribution is the coordinates of the fingerprint being perturbed, and the covariance is a diagonal matrix, where the diagonal elements are the mean distance of a fingerprint to its nearest neighbor in the initial candidate set. Figure 4 (right) shows the five fingerprints in the candidate set with a dotted circle around them whose radius is equal to the threshold distance $t_1 = 0.75$. If the new fingerprint falls inside any of the dotted circles, then it will be rejected on account of being redundant with the fingerprints in the candidate set. Figure 4 (right) shows two examples of an acceptable new fingerprint - $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$. Both of them expand the space spanned by the candidate set of fingerprints.

We repeat the exercise of identifying and adding a fingerprint around the most sparsely located one, until we have added the desired number of fingerprints in the candidate set. Figure 5 shows the results obtained when we have a candidate set of $N_1 = 200$ fingerprints, and $N_2 = 400$ fingerprints. Our algorithm performs well in (a) providing a candidate set of fingerprints that spans the entire domain space, (b) spacing-out fingerprints such that they evenly span the domain space within a given budget of $N_1$ fingerprints. The non-adaptive domain space expansion algorithm is included as Algorithm 1.

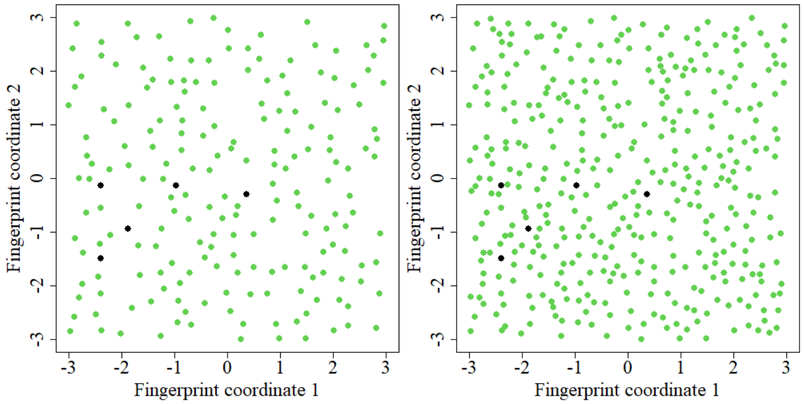

**Figure 5** **Candidate set of $N_1 = 200$ fingerprints (left) and $N_1 = 400$ fingerprints (right) in the toy example.**

Regarding the choice of $N_1$, ideally, it should be the number of fingerprints that are sufficient to capture the global minimum. However, since it is impossible to know this beforehand, we suggest

**Algorithm 1** : Non-adaptive domain space expansion

```
Input $N_1, \mathcal{C}, \mathcal{X}$
$i \leftarrow n_0$
Compute $\boldsymbol{R} = \{r_1, \cdots, r_{n_0}$ {Distance to farthest nearest neighbor among $2p$ neighborhoods, for each fingerprint }
Compute $\boldsymbol{D} = \{d_1, \cdots, d_{n_0}$ {Distance to the nearest neighbor of each fingerprint}
$t \leftarrow \text{mean}(\boldsymbol{D})$
while $i \leq N_1$ do
    per $\leftarrow \arg\max_i r_i$
    Randomly perturb $\boldsymbol{c}_{per}$ to generate $\boldsymbol{c}_{new}$
    $\boldsymbol{x}_{new} \leftarrow \text{fingerprint}(\boldsymbol{c}_{new})$
    Compute $d_{min}$ {Distance to the nearest neighbor of $\boldsymbol{x}_{new}$}
    if $d_{min} > t$ then
        $\mathcal{C} \leftarrow \text{append}(\mathcal{C}, \boldsymbol{c}_{new})$
        $\mathcal{X} \leftarrow \text{append}(\mathcal{X}, \boldsymbol{x}_{new})$
        $i \leftarrow i + 1$
        Update $\boldsymbol{R}, \boldsymbol{D}$
    end if
    $t \leftarrow 0.5(t + d_{min})$
end while
Output $\mathcal{C}, \mathcal{X}, \boldsymbol{R}, \boldsymbol{D}, t$
```

a candidate set of size $N_1 = 100p$, as a thumb-rule, for a $p$-dimensional fingerprint. A larger $N_1$ could be used, but it will increase the computational and storage costs as discussed at the end of Section 4.3.

### 3.2. Adaptive domain space expansion

Adaptive domain space expansion is an optional step. It is not needed if the candidate set, obtained with the non-adaptive expansion algorithm, captures the global minimum. Although it is impossible to determine if the global minimum has been captured, it is possible to make an educated guess if further expansion of a particular region of the explored domain space may help capture more minima, one of which may potentially be the global minimum.

To illustrate the need for adaptive expansion, we will consider an example, where it is assumed that the fingerprints are two-dimensional and their potential energy is given by the Branin function (Surjanovic & Bingham 2013). Assume that there are $n_0 = 10p = 20$ initial fingerprints as shown in Figure 6 (left). We use the non-adaptive expansion algorithm to expand them to a set of $N_1 =$

$100p = 200$ fingerprints as shown in Figure 6 (right). As the budget uptil the non-adaptive expansion ($N_1 = 200$) step is exhausted, we will shift the focus to only the "promising regions" or the low-energy regions of the domain space, instead of the sparsely populated regions. We use DFT to compute the energy of the initial $10p = 20$ fingerprints. Then, a model-based approach will be used for identifying the low-energy "promising regions".

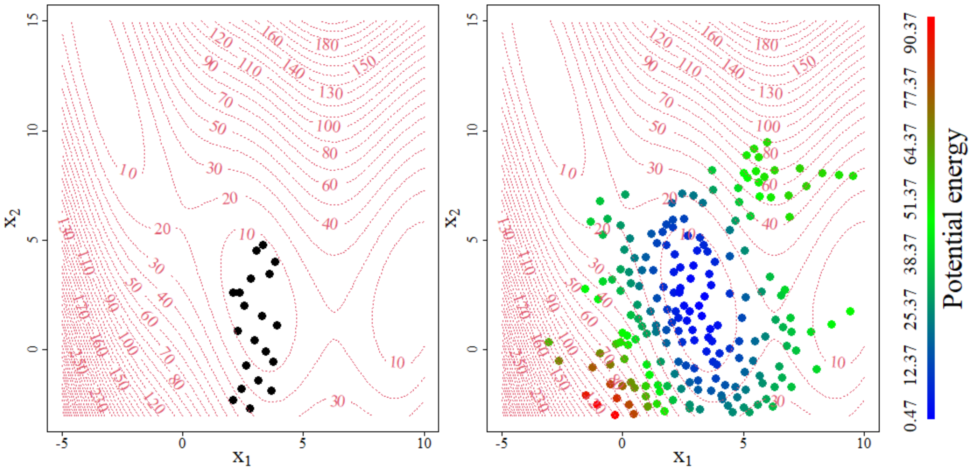

**Figure 6** **Candidate set of $n_0 = 20$ initial fingerprints (left) and $N_1 = 200$ fingerprints obtained from the non-adaptive expansion algorithm (right), shown over the contour plot of the Branin function.**

As the $n_0$ fingerprints corresponding to the initially known configurations are likely to be stable, we compute their potential energy. If $n_0 < 10p$, we suggest to choose the remaining fingerprints from the set of $N_1$ fingerprints, by augmenting the initial set with a space-filling design such as the maximum projection (MaxPro) design (Joseph et al. 2015). The augmented design can be obtained sequentially by adding one fingerprint at a time using the `MaxProAugment` function in the R package `MaxPro` (Ba & Joseph 2018).

We use Gaussian Process (GP) for modeling the energy data. Assume that $e(\cdot)$ is a realization of a GP:

$$e(\boldsymbol{x}) \sim GP(\mu, C(\boldsymbol{x};\cdot)), \tag{6}$$

where $\mu$ is the mean and $C(\boldsymbol{x}_u;\boldsymbol{x}_v) = Cov\{f(\boldsymbol{x}_u), f(\boldsymbol{x}_v)\}$ is the covariance function. See Santner et al. (2018) for details on GP modeling. Given the energy-data, the posterior distribution of $e(\boldsymbol{x})$ is given by

$$e(\boldsymbol{x})|\boldsymbol{e} \sim \mathcal{N}(\hat{e}(\boldsymbol{x}), s^2(\boldsymbol{x})), \tag{7}$$

where

$$\hat{e}(\boldsymbol{x}) = \mu + C(\boldsymbol{x};\boldsymbol{S})C^{-1}(\boldsymbol{S};\boldsymbol{S})(\boldsymbol{e} - \mu\boldsymbol{1}) \tag{8}$$

is the surrogate model,

$$s^2(\boldsymbol{x}) = C(\boldsymbol{x};\cdot) - C(\boldsymbol{x};\boldsymbol{S})C^{-1}(\boldsymbol{S};\boldsymbol{S})C(\boldsymbol{S};\cdot) \tag{9}$$

is the variance, $\boldsymbol{S} = [\boldsymbol{x}_1^T, \cdots, \boldsymbol{x}_n^T]^T$, $C(\boldsymbol{x};\boldsymbol{S})$ is the covariance vector with $i$th element $C(\boldsymbol{x};\boldsymbol{S}_i)$, $C(\boldsymbol{S};\boldsymbol{S})$ is the covariance matrix, and $\mathbf{1}$ is a vector of 1's. We use the R package `DiceKriging` (Roustant et al. 2012) to fit the GP model using a Gaussian covariance function. Figure 6 (right) shows the potential energy predictions (shown in color) based on the fitted GP model. We see that the estimated global minimum seems to lie in the "interior" of the candidate set of configurations. In this case, there is no need to further expand the "low-energy" region, as it is already surrounded by the candidate set of configurations.

Now, let us consider another scenario, where the the initial set of $10p$ fingerprints are as shown over the contour plot of the Branin function in Figure 7 (left). Figure 7 (center) shows the non-adaptive expansion along with the potential energy (shown in color) based on the fitted GP model. The estimated global minimum seems to lie at the "boundary" of the candidate set of configurations. In this case the "low-energy" region is not well explored on all sides. Thus, in this case, we need to further expand the "low-energy" region to ensure that the minimum of the "low-energy" region, which may potentially be the global minimum or a reliable local minimum, is included in the candidate set of configurations.

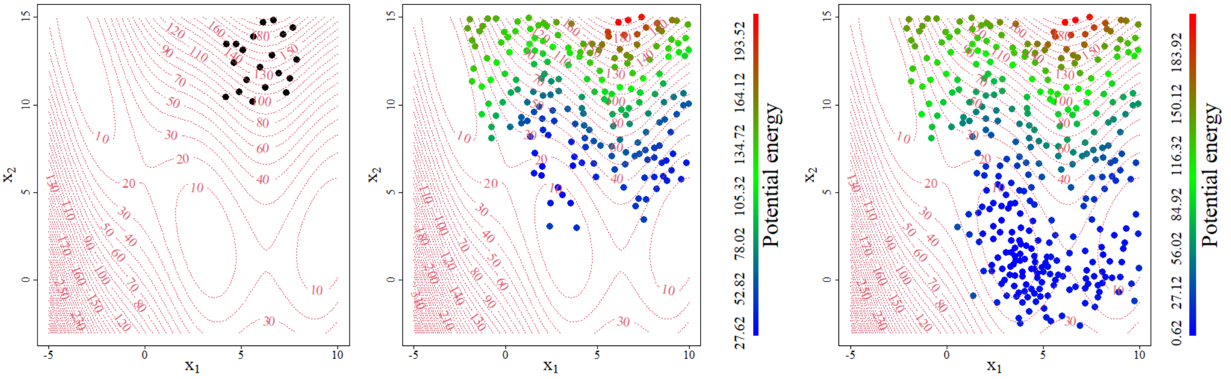

**Figure 7 Candidate set of $n_0 = 20$ initial fingerprints (left); $N_1 = 200$ fingerprints obtained from the non-adaptive expansion algorithm (center); and $N_2 = 341$ fingerprints obtained after the adaptive expansion algorithm (right), shown over the contour plot of the Branin function.**

The definition of the "boundary" and the "interior" of the candidate set of configurations is based on the dimension $p$ of the fingerprint. A two-dimensional fingerprint, assumed to be at the origin, lies in the interior of the domain space if it has neighboring fingerprints in each of the four quadrants, within a distance $r$ around it. Otherwise, it lies on the boundary of the domain space spanned by the candidate set of fingerprints. Here $r$ is taken to be the maximum distance to the nearest neighbor for the candidate set of $N_1$ fingerprints obtained at the end of the non-adaptive expansion algorithm. Figure S1 in the Supplementary material shows examples of two-dimensional fingerprints that lie on the boundary or in the interior of the domain space spanned by the candidate set.

For the case shown in Figure 7 (center), we push the boundary of the low-energy region by perturbing the lowest energy fingerprint that lies on the boundary of the spanned domain space. With the addition of every 10 fingerprints (thumb-rule) to the candidate set, DFT is used for computing the potential energy of the fingerprint with the least energy estimate. The GP model is then updated to better estimate the energy in the newly explored lower-energy domain space. A periodic model-update helps navigate the expansion of the lower-energy region. If the fingerprint having the minimum estimated potential energy does not change within 10 successive DFT computations (thumb-rule), we stop the algorithm. The adaptive domain space exploration algorithm is included as Algorithm 2.

Figure 7 (right) shows the result of applying the adaptive expansion algorithm to the scenario presented in Figure 7 (center). The algorithm adaptively expands the set of $N_1 = 200$ fingerprints obtained at the end of non-adaptive expansion to $N_2 = 341$ fingerprints. Note that the algorithm continues to expand until the low-energy region is fully explored, and the minimum is well surrounded by the candidate set of fingerprints.

We tested our algorithms (non-adaptive + adaptive expansion) to find the minimum of two different kinds of functions - sphere function and Schwefel function (Surjanovic & Bingham 2013). The sphere function in $p$ dimensions is given by:

$$f(x) = \sum_{i=1}^{p} x_i^2; x_i \in [-5.12, 5.12], \tag{10}$$

while the Schewefel function in $p$ dimensions is given by:

$$f(x) = 418.9829p - \sum_{i=1}^{p} x_i \sin(\sqrt{|x_i|}); x_i \in [-500, 500] \tag{11}$$

The sphere function is simple and smooth with only one minimum, while the Schewefel function is complex with several local minima. These functions are generalizable to any dimension $p$. We consider two distinct values of dimension: $p = 2$ and $p = 10$ for both the functions. We also consider two distinct values of budget (until the non-adaptive expansion step): $N_1 = 50p$ and $N_1 = 100p$.

The feasible domain space of the sphere function is taken as $[-5.12, 5.12]^p$ and that of the Schwefel function to be $[-500, 500]^p$. However, as the feasible domain space is assumed to be unknown, we generate the initial set of $10p$ fingerprints as a maximin Latin hypercube design (Morris & Mitchell 1995) in a smaller sub-space: $[1.5, 4]^p$ for the sphere function and $[250, 400]^p$

**Algorithm 2** : Adaptive domain space expansion

```
Import $\mathcal{C}, \mathcal{X}, \boldsymbol{R}, \boldsymbol{D}, t$ {Obtained at the end of the non-adaptive expansion procedure}
$\mathcal{X} \leftarrow \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n_0}\}$
Augment $\mathcal{X}_{n_0}$ by $10p - n_0$ space-filling fingerprints to obtain $\mathcal{X}_{DFT}$ {R package:MaxPro }
$\boldsymbol{e} \leftarrow DFT(\mathcal{X}_{DFT})$; model $\leftarrow GP(\mathcal{X}_{DFT}, \boldsymbol{e})$
$\hat{\boldsymbol{e}} \leftarrow GP.predict(\mathcal{X})$
flag $\leftarrow 1$; DFT_period $\leftarrow 0$; iter $\leftarrow$ 0;
while flag $= 1$ do
  Find $\boldsymbol{c}_{per}$, the configuration with minimum estimated energy lying on the boundary
  Lines $8 - 10$ from the non-adaptive domain space expansion algorithm
  if $d_{min} > t$ then
    Lines $12 - 15$ from the non-adaptive domain space expansion algorithm
    DFT_period $\leftarrow$ DFT_period+1
    $e_{new} \leftarrow GP.predict(\mathcal{X}_{new})$
    $\hat{\boldsymbol{e}} \leftarrow$ append$(\hat{\boldsymbol{e}}, e_{new})$
    if DFT_period $= 10$ then
      iter $\leftarrow$ iter+1
      Find $\boldsymbol{c}_{min}$, the configuration with the minimum estimated potential energy
      $e_{min} \leftarrow DFT(\boldsymbol{c}_{min})$; $\boldsymbol{e} \leftarrow$ append$(\boldsymbol{e}, e_{min})$
      $\mathcal{X}_{DFT} \leftarrow$ append$(\mathcal{X}_{DFT}, \boldsymbol{c}_{min})$
      model $\leftarrow GP(\mathcal{X}_{DFT}, \boldsymbol{e})$
      $\hat{\boldsymbol{e}} \leftarrow GP.predict(\mathcal{X})$
      if iter $= 10$ then
        iter $= 0$
        if $\boldsymbol{c}_{min}$ has not changed since the last 10 DFT computations then
          flag $= 0$
        end if
      end if
    end if
  end if
  $t \leftarrow 0.5(t + d_{min})$
end while
Output $\mathcal{C}, \mathcal{X}, \mathcal{X}_{DFT}, \boldsymbol{e}$
```

for the Schwefel function. Note that the global minimum of the sphere function, $(0, ..., 0)$, and that of the Schewefel function, $(420.9687, ..., 420.9687)$, are outside these sub-spaces. We compare our results to a baseline maximin Latin hypercube sample of size $N_2$ generated using the R package `lhs` (Carnell 2016) in the sub-space of the initial set of fingerprints. We perform 30 simulations for each unique combination of function, dimension ($p$), and budget until non-adaptive expansion ($N_1$).

Figure 8 shows the distribution the minimum energy in the candidate set of fingerprints obtained at each step of our method, and in the baseline sample. There are several important points to note. First, the candidate set generated by our method (at the end of the adaptive expansion step) has a fingerprint with a lower energy, on average, as compared to the maximin Latin hypercube design (baseline sample) of the same size. Second, the adaptive expansion step of our algorithm helps navigate the search towards lower energy regions of the feasible space. Third, a larger budget in the non-adaptive expansion step leads to more exploration of the feasible domain space, which results in lower energy fingerprints in the candidate set, as expected. Fourth, our method scales up well, performing better than the state-of-the-art method even for higher dimensions. Fifth, our method works well even for very complex and highly multimodal functions.

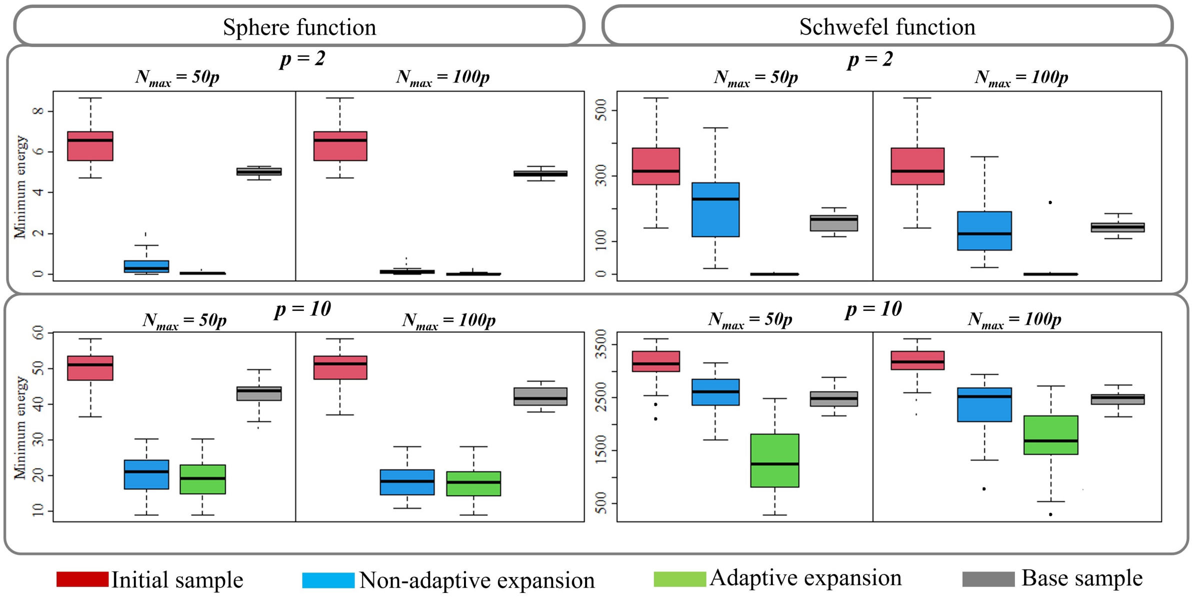

**Figure 8** Distribution of the minimum energy in the candidate set of fingerprints obtained at each step of our method (initial sample, non-adaptive expansion, and adaptive expansion) and in the baseline sample.

### 3.3. Exploration and exploitation of the domain space: Bayesian optimization

The purpose of this step is to identify the crystal structure configuration with the least potential energy in this candidate set of size $N_2$ obtained from the expansion steps. As mentioned in Section

2.1, energy computation using the DFT is too expensive, which makes it impractical to compute the energy for all the $N_2$ configurations. So, we will use the GP model developed during the adaptive expansion procedure to estimate the energy of all the $N_2$ fingerprints. Bayesian optimization (Jones et al. 1998, Frazier 2018) will then be used for iteratively optimizing and updating the model. The method will let us identify the global minimum with DFT computations over only a small fraction of the $N_2$ fingerprints in the candidate set.

Let $\boldsymbol{e}$ be the vector of potential energy, computed using DFT, and $s(\cdot)$ be the standard error of their energy estimate. Then, the expected improvement criterion can be expressed as the following closed form (Jones et al. 1998):

$$EI(\boldsymbol{x}) = [\min(\boldsymbol{e}) - \hat{e}(\boldsymbol{x})]\Phi\left(\frac{\min(\boldsymbol{e}) - \hat{e}(\boldsymbol{x})}{s(\boldsymbol{x})}\right) + s(\boldsymbol{x})\phi\left(\frac{\min(\boldsymbol{e}) - \hat{e}(\boldsymbol{x})}{s(\boldsymbol{x})}\right), \tag{12}$$

where $\Phi$ and $\phi$ are respectively the cumulative distribution function and the probability density function of the standard normal distribution, $\hat{e}(.)$ is the estimated energy as defined in (8), and $s(\boldsymbol{x})$ is as defined in (9). As the surrogate model is cheap, we evaluate (12) on all the $N_2$ fingerprints in the candidate set $\mathcal{X}$, and find the one that maximizes it:

$$\boldsymbol{x}_{new} = \underset{\boldsymbol{x}\in\mathcal{X}}{\arg\max}\, EI(\boldsymbol{x}). \tag{13}$$

DFT is used for computing the potential energy at $\boldsymbol{x}_{new}$, and the energy-data set is updated to include $[\boldsymbol{x}_{new}, e(\boldsymbol{x}_{new})]$. Note that we are able to compute the energy because $\boldsymbol{x}_{new} \in \mathcal{X}$ and therefore, we know the corresponding Cartesian coordinate configuration. Now we use (7) to update our surrogate model based on the updated energy-data. The expected improvement (EI) criterion balances exploration of the PES with exploitation, thereby simultaneously addressing both the objectives of crystal structure prediction - exploiting low-energy regions to search for the minimum and exploring new and unusual domains of the PES. We stop the Bayesian optimization algorithm, when the expected improvement becomes negligible as compared to the current estimate of the minimum potential energy. The Bayesian optimization algorithm is included as Algorithm 3.

The computational complexities of the three Algorithms 1, 2, and 3 are $O((n_0+n_1)^2p^2)$, $O(n_0+n_2/10)^3n_2)$, and $O((n_0+n_2/10+n_3)^3n_3)$, respectively. See Section B of the Appendix for details. In addition to computational costs, there are storage costs as it is required to store $N_2 = n_0+n_1+n_2$, $p$-dimensional fingerprints, their corresponding Cartesian coordinate configurations, and the vectors consisting of the estimated energy and distance to the nearest neighbor of each fingerprint. However, the main cost of the methodology is the cost of running $(n_0+n_2/10+n_3)$ DFT computations.

**Algorithm 3** : Bayesian optimization

```
Import $\mathcal{C}, \mathcal{X}$ {Obtained from the expansion algorithms}
Input $t_{EI}$
**if** Algorithm 2 : Adaptive domain space expansion is used **then**
    Import $\mathcal{X}_{DFT}, \boldsymbol{e}$
**else**
    Lines $2-5$ from the adaptive domain space expansion algorithm
**end if**
$e_{min} \leftarrow \min(\boldsymbol{e})$
$EI_i \leftarrow EI(\boldsymbol{x}_i); i \in \{1, \cdots, N_2\}$ {Use (12)}
percent_improve $\leftarrow \max(\boldsymbol{EI})/e_{min}$
**while** percent_improve $\leq t_{EI}$ **do**
    $i\_maxEI \leftarrow \arg\max_i(\boldsymbol{EI}); i \in \{1, \cdots, N_2\}$
    $e_{new} \leftarrow DFT(\boldsymbol{x}_{i\_maxEI})$
    $\mathcal{X}_{DFT} \leftarrow$ append$(\mathcal{X}_{DFT}, \boldsymbol{x}_{i\_maxEI})$
    $\boldsymbol{e} \leftarrow$ append$(\boldsymbol{e}, e_{new})$
    $model \leftarrow GP(\mathcal{X}_{DFT}, \boldsymbol{e})$
    $EI_i \leftarrow EI(\boldsymbol{x}_i); i \in \{1, \cdots, N_2\}$ {Use (12)}
    percent_improve $\leftarrow \max(\boldsymbol{EI})/e_{min}$
**end while**
$i\_stable \leftarrow \arg\min_i \boldsymbol{e}; i \in \{1, \cdots, nrows(\mathcal{X}_{DFT})\}$
Output $\boldsymbol{x}_{i\_stable}, \boldsymbol{c}_{i\_stable}$
```

## 4. Example: Crystal structure of $Al_8$

The objective is to find the most stable crystal structure configuration of $Al_8$, or the configuration of eight Aluminum (Al) atoms arranged in a parallelepiped unit cell. Note that the most stable configuration of $Al_8$ is the face-centered cubic (fcc), which is already known and is shown in Figure 1. Because the primitive cell of fcc $Al$ has one atom, it must be included in the configuration space of $Al_8$. For our candidate set, eight-atoms structures were created by (1) randomly choosing a specific value of volume $v$ falling within ±5% of the known specific volume of the Al fcc structure, (2) randomly selecting three vectors $\vec{a}$, $\vec{b}$, and $\vec{c}$ of the unit cell so that its volume, given by $V \equiv \vec{a} \cdot (\vec{b} \times \vec{c}) = 8v$, and (3) randomly arranging the eight Al atoms in the cell so that the distance between any pairs is larger than 2.0 Å. Two constraints (1) and (3) of this procedure, which were formulated from the known facts of the fcc Al structure, clearly limit the examined configuration space but the search domain remains staggering and certainly contains the global minimum.

Within our expansion-exploration-exploitation framework, we start from a set of 270 initial configurations for which the energy is computed at the DFT level. Note that the DFT calculations performed herein involve *only single-point energy calculations* but not any local optimizations, which may be $10^3 - 10^4$ times more expensive. Therefore, in general, none of the examined structures is a local minimum of the PES. However, the main objectives of this work, i.e., *(diversely) filling the configuration space and searching for the global minimum of a big and diverse structure dataset*, can be demonstrated and is very useful for material structure prediction.

We used the ABINIT package for DFT computations, and selected the parameters that ensure a typical level of accuracy needed in Materials Science. During the process of exploring the configuration space using our active learning algorithm, a DFT computation took about 122 seconds on average. Clearly, this is a major bottleneck compared to ML predictions, which need a fraction of a second.

To obtain a candidate set of fingerprints, we start with the non-adaptive domain space expansion algorithm (Algorithm 1). We have a set of $n_0 = 270$ initially known atomic configurations of $Al_8$. These configurations become the input $\mathcal{C}$, and their corresponding fingerprints become the input $\mathcal{X}$. The fingerprints have a dimension of $p = 32$. As per the thumb-rule mentioned earlier, we take $N_1 = 100p = 3,200$. For perturbing the most sparsely located atomic configuration, we randomly change the position of each atom by a maximum of 0.1 Angstrom.

As the *AGNI* fingerprint is 32-dimensional, the candidate set of fingerprints cannot be visualized directly. We use Principal Component Analysis (Jackson 2005), or PCA to visualize the first three PCs, which capture 97% of their variance. Figure 9 (left) shows the PCs of the initial set of fingerprint (grey circles) that are input to the non-adaptive space expansion algorithm (Algorithm 1). The solid black circle (in this and all the subsequent figures) corresponds to the most stable fingerprint, or the fingerprint that has the minimum potential energy. This is not a part of the initial candidate set. However, this is the solution that we hope to achieve. Note the relatively large gap between it and the initial set of fingerprints. Ideally, our expansion algorithms will expand the initial candidate set to include the most stable fingerprint, and then the Bayesian optimization algorithm should identify this fingerprint as the global minimum.

Figure 9 (center) shows the PCs for the candidate set of $N_1 = 3,200$ fingerprints obtained using our non-adaptive space expansion algorithm. Note that the algorithm expands the volume of the domain space spanned by the initial candidate set of fingerprints, which reduces the gap between different clusters of fingerprints in the initial candidate set.

Figure 9 (right) shows the PCs for the candidate set of $N_1 = 3,200$ fingerprints obtained using the Ab initio random structure searching approach (AIRSS) (Pickard & Needs 2011). Figure 10 shows the median distance to the nearest neighbor of a fingerprint as the candidate set size increases. As

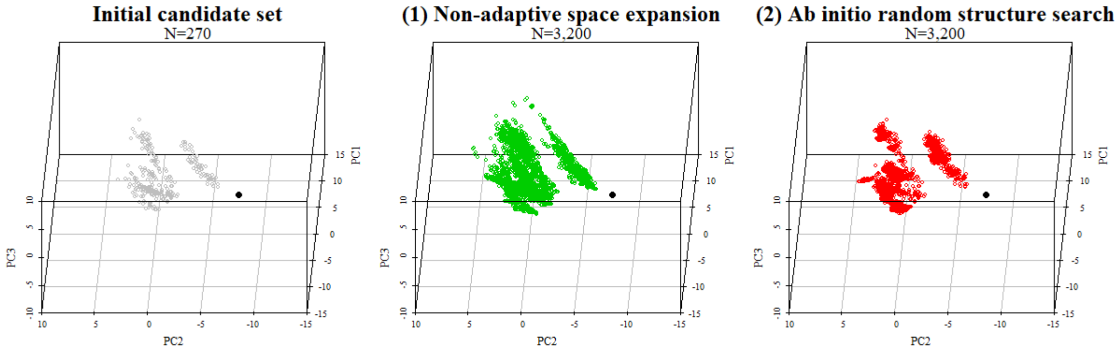

**Figure 9** **Initial set of fingerprints (left), and candidate set of $N_1 = 3,200$ fingerprints obtained using non-adaptive space expansion algorithm (center) and the Ab initio random structure searching approach (right). The solid black circle is the true global optimum, which is not included in the candidate set.**

expected, our proposed candidate set keeps expanding the spanned domain space leading to more spaced-out fingerprints. On the other hand, the Ab initio random structure searching approach results in a set of closely packed fingerprints leading to wastage of resources, and failure to consider more distinct configurations.

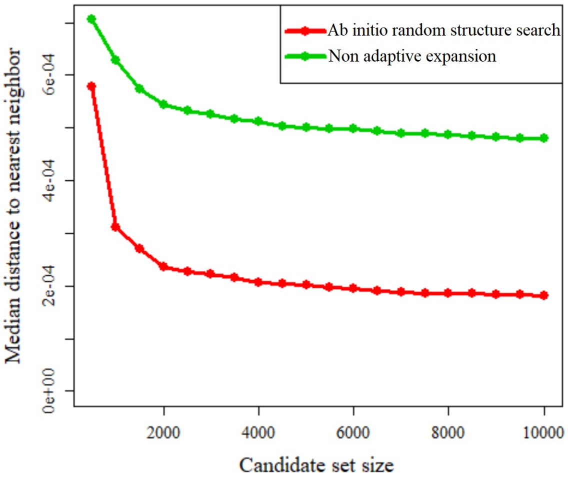

**Figure 10** **Distance to the nearest neighbor vs candidate set size for fingerprints.**

Figure 11 (top) shows examples of a couple of configurations corresponding to the fingerprints obtained from the non-adaptive domain space expansion algorithm. Figure 11 (bottom) shows the X-ray diffraction patterns corresponding to these configurations. The stark differences in these configurations, as evident by their X-ray diffraction pattern, shows that the algorithm spans through quite distinct regions of the domain space.

After obtaining a candidate set of 3, 200 fingerprints from the non-adaptive expansion algorithm, we compute the energy for $10p = 320$ fingerprints, using DFT, to develop a GP model. These 320

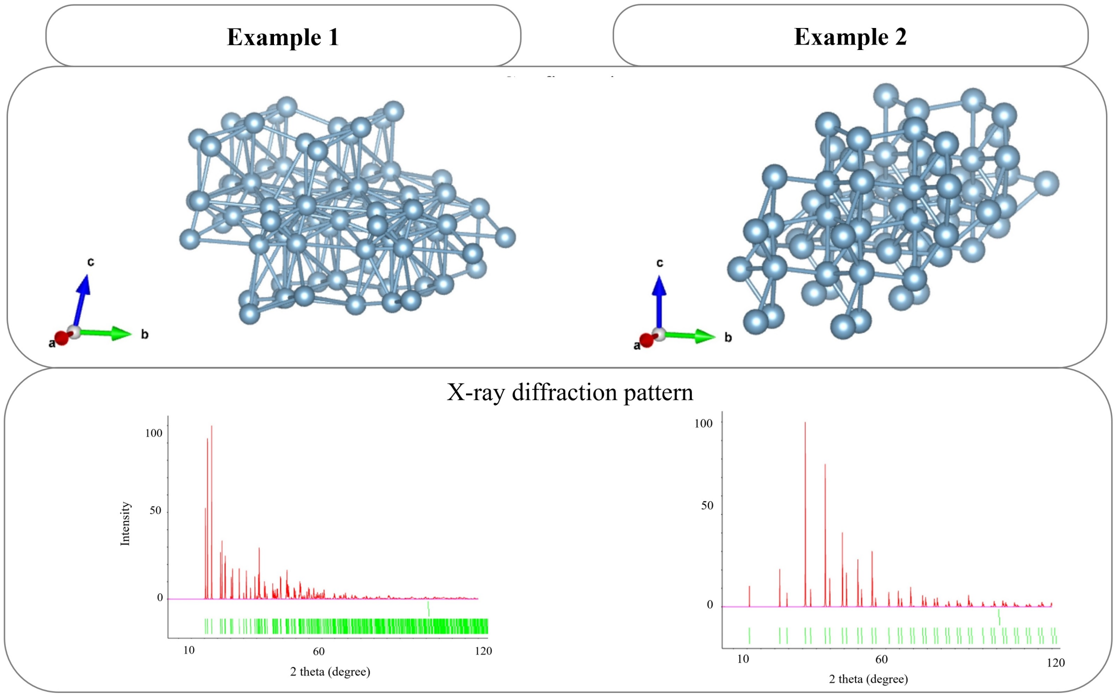

**Figure 11** **Configuration (top) and X-ray diffraction pattern (bottom) of a couple of example structures.**

fingerprints include the initial candidate set of 270 fingerprints corresponding to known configurations, which are augmented by another 50 fingerprints from the candidate with the space-filling MaxPro design. Figure S2 in the Supplementary material shows the initial fingerprints and the ones augmented using the MaxPro design.

Figure 12 (left) shows the predicted potential energy of the candidate set using the GP model. The low-energy regions seem to be around the boundary of the spanned domain space. Using the boundary definition mentioned earlier and illustrated in Figure S1 (of the Supplementary material), we find that the estimated minimum of the candidate set is actually at the boundary of the spanned domain space. So, it is necessary to further expand this low energy region as it may lead to further minimization of the current estimate of the energy-minimum. Note that we work in the space of the first three principal components for determining the boundary because it is very expensive to work in the 32-dimensional fingerprint space.

The adaptive expansion algorithm (Algorithm 2) is used for further expanding the identified low-energy region of the spanned domain space. Figure 12 (right) shows the updated candidate set of fingerprints and the predicted potential energy after the adaptive expansion procedure. The algorithm adds 530 fingerprints to the candidate set, and DFT computations are done for every $10th$ fingerprint added to the set, i.e., for a total of 53 fingerprints. The algorithm stops when the estimated minimum does not change with 10 successive DFT computations. We observe that the algorithm succeeds in expanding the candidate set towards the unknown true global minimum!

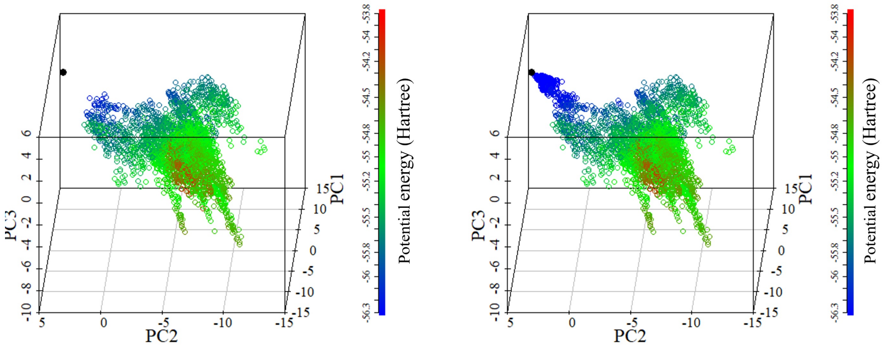

**Figure 12 Potential energy of the candidate set of fingerprints obtained after non-adaptive expansion (left) and adaptive expansion (right).**

Once a candidate set of fingerprints is obtained by the expansion algorithms, we explore and exploit it for the global minimum of potential energy, using the Bayesian optimization algorithm (Algorithm 3). The input $\mathcal{X}$ in this example is the candidate set of $3,200+530=3,730$ fingerprints that we obtained using the expansion algorithms.

The circles in Figure 13 are the candidate set of $N_1 = 3,200$ fingerprints. The potential energy was computed for a set of $320+53=373$ fingerprints (in the adaptive domain space expansion algorithm), shown as squares in Figure 13 (left). A GP model was fitted using the potential energy data of the 373 fingerprints. Then, iterative Bayesian optimization procedure of updating the GP model, and adding a point in the known-data based on the EI criterion is continued until the expected improvement becomes lesser than a threshold value $t_{EI}$. We chose $t_{EI}$ to be 0.001% of the current minimum value of potential energy in the energy-data. Thus, for the $ith$ iteration, the threshold value is:

$$t_{EI} = 10^{-5} \times \min(e_1, \cdots, e_{n+i}). \tag{14}$$

This is a reasonably low value as the potential energy in the known-data varies by 5% around its mean. Jones et al. (1998) stop the Bayesian optimization algorithm when the expected improvement is 1% of the current best function value. We have used a much more conservative stopping criterion. For a different structure, a reasonably low cutoff must be chosen depending on the variance in the potential energy of the candidate set of configurations. With the above threshold, the algorithm stopped after the $95th$ iteration.

The fingerprints iteratively added in the known-data are shown as blue triangles in Figure 13 (right). There are three points to note about the iteratively added fingerprints. First, 93 of the 95 fingerprints are selected in the region around the global minimum, which shows that the algorithm does well in exploiting the "promising" region of the domain space. Second, two fingerprints are selected in regions far away from the global minimum. These are regions at the boundary of the domain space, where probably there is high uncertainty in the potential energy estimate of the GP surrogate model. This shows the exploratory nature of the algorithm, where it tries to find the global minimum in regions other than the “promising region”. This exploratory feature of the algorithm makes it better than the state-of-the-art approaches such as basin hopping (Wales & Doye 1997) and minima hopping (Goedecker 2004), which focus only on exploiting the “promising region” of the domain space for the global minimum. Third, though the true global minimum was not a part of the candidate set, we identified the fingerprint closest to it as the global minimum! Thus, the algorithm provided a solution that is potentially very similar to the true global minimum. The expected improvement criterion, and the energy of configurations selected for DFT computations are visualized in Figure S3 in the Supplementary material.

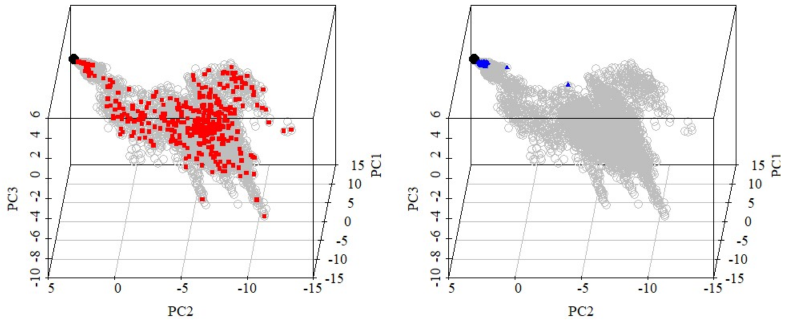

**Figure 13 (left): initial set of** $373$ **fingerprints for which the potential energy is computed using DFT (red squares) and (right):** $95$ **fingerprints selected during the Bayesian optimization (blue triangles), which are shown over the candidate set of** $3,200$ **fingerprints (circles).**

The algorithm’s output and our solution is the configuration corresponding to the fingerprint selected in the $368th$ DFT computation, as it has the minimum potential energy in the energy-data. Figure 14 (top) shows that the estimated structure looks quite similar to the true structure. The similarity in the estimated and true structures is more evident in the X-ray diffraction pattern of the structures as shown in Figure 14 (bottom).

## 5. Conclusion

We have developed an active learning method to (1) obtain a candidate set of crystal structure configurations that expands the space of a few initially known configurations, and (2) find the configuration with the lowest potential energy in the set. The novelty of our approach is the expansion-exploration-exploitation framework that extends the traditionally used exploration-exploitation Bayesian optimization framework to better achieve the objectives of crystal structure prediction. The expansion algorithms ensure that the candidate set continues to expand with the addition of each fingerprint - first in arbitrary directions (with respect to potential energy) to explore new and possible unusual domains of the PES, and then towards lower-energy atomic configurations. Our

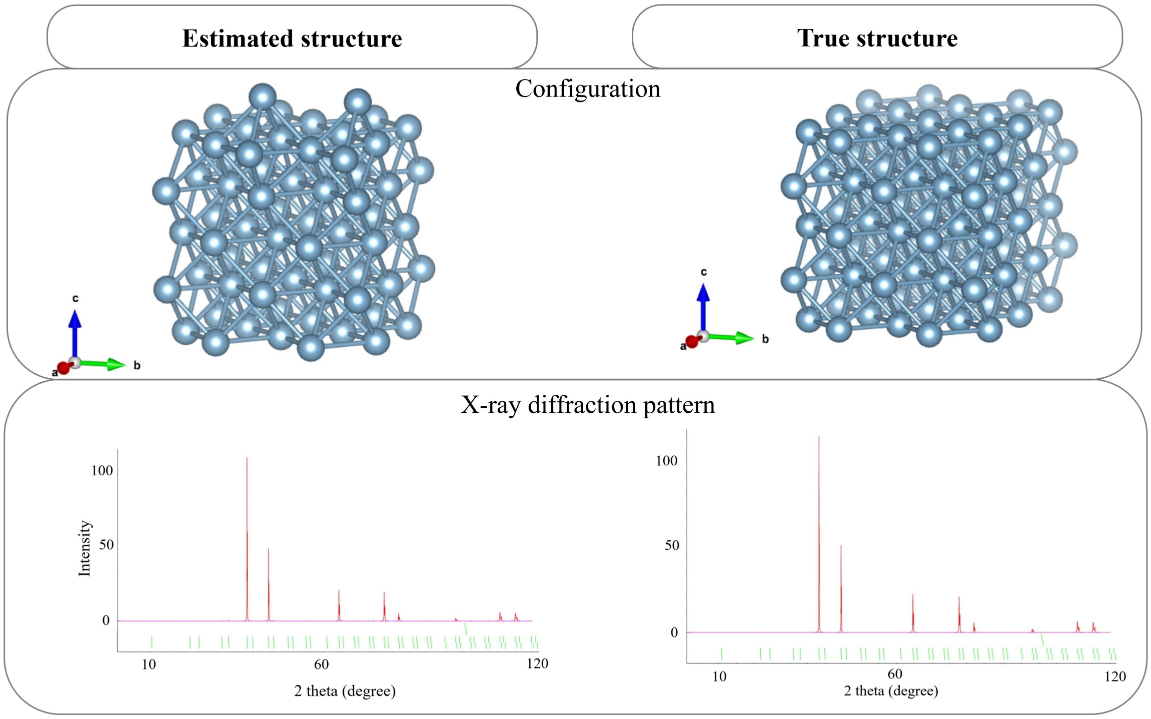

**Figure 14** **Configuration (top) and X-ray diffraction pattern (bottom) of the estimated and true structures.**

algorithm provides a space-filling design without the knowledge of the boundaries of the design space. This is a novel contribution to the field of experimental design, where most of the work on space-filling design is focused on cases of known design space.

Although we demonstrated our approach on a simple problem of finding the stable configuration of $Al_8$, the new concepts are powerful and can easily be generalized to more realistic problems. A recent and interesting application of the proposed method in inverse designs can be found in Krishna et al. (2022).

## Acknowledgments

The authors thank a U.S. National Science Foundation grant DMREF-1921873 and XSEDE through grant DMR170031.

## References

Ba, S. & Joseph, V. R. (2018), 'Maxpro: Maximum projection designs. R package version 4.1-2', *URL: https://cran.r-project.org/web/packages/MaxPro* .

Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. (2010), 'Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons', *Phys. Rev. Lett.* **104**(13), 136403.

Basudhar, A. & Missoum, S. (2008), 'Adaptive explicit decision functions for probabilistic design and optimization using support vector machines', *Computers & Structures* **86**(19-20), 1904–1917.

Batra, R., Tran, H. D., Kim, C., Chapman, J., Chen, L., Chandrasekaran, A. & Ramprasad, R. (2019), 'General atomic neighborhood fingerprint for machine learning-based methods', *J. Phys. Chem. C* **123**(25), 15859–15866.

Behler, J. & Parrinello, M. (2007), 'Generalized neural-network representation of high-dimensional potential-energy surfaces', *Phys. Rev. Lett.* **98**(14), 146401.

Berry, R. S. (1993), 'Potential surfaces and dynamics: What clusters tell us', *Chemical reviews* **93**(7), 2379–2394.

Carnell, R. (2016), 'Package 'lhs", *CRAN. https://cran. rproject. org/web/packages/lhs/lhs. pdf* .

Chen, J., Zhu, G., Yuan, C. & Huang, Y. (2020), 'Semi-supervised embedding learning for high-dimensional bayesian optimization', *arXiv preprint arXiv:2005.14601* .

Chen, W. & Fuge, M. (2017), 'Beyond the known: Detecting novel feasible domains over an unbounded design space', *Journal of Mechanical Design* **139**(11).

d'Avezac, M., Luo, J.-W., Chanier, T. & Zunger, A. (2012), 'Genetic-algorithm discovery of a direct-gap and optically allowed superstructure from indirect-gap si and ge semiconductors', *Phys. Rev. Lett.* **108**(2), 027401.

Franceschetti, A. & Zunger, A. (1999), 'The inverse band-structure problem of finding an atomic configuration with given electronic properties', *Nature* **402**(6757), 60–63.

Frazier, P. I. (2018), 'Bayesian optimization', *INFORMS Tutorials* pp. 255–278.

Gaida, N. A., Niwa, K., Sasaki, T. & Hasegawa, M. (2021), 'Phase relations and thermoelasticity of magnesium silicide at high pressure and temperature', *J. Chem. Phys.* **154**(14), 144701.

Glass, C. W., Oganov, A. R. & Hansen, N. (2006), 'Uspex—evolutionary crystal structure prediction', *Computer physics communications* **175**(11-12), 713–720.

Goedecker, S. (2004), 'Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems', *J. Chem. Phys.* **120**(21), 9911–9917.

Gonze, X., Jollet, F., Araujo, F. A., Adams, D., Amadon, B., Applencourt, T., Audouze, C., Beuken, J.-M., Bieder, J., Bokhanchuk, A., Bousquet, E., Bruneval, F., Caliste, D., Côté, M., Dahm, F., Pieve, F. D., Delaveau, M., Gennaro, M. D., Dorado, B., Espejo, C., Geneste, G., Genovese, L., Gerossier, A., Giantomassi, M., Gillet, Y., Hamann, D., He, L., Jomard, G., Janssen, J. L., Roux, S. L., Levitt, A., Lherbier, A., Liu, F., Lukačević, I., Martin, A., Martins, C., Oliveira, M., Poncé, S., Pouillon, Y., Rangel, T., Rignanese, G.-M., Romero, A., Rousseau, B., Rubel, O., Shukri, A., Stankovski, M., Torrent, M., Setten, M. V., Troeye, B. V., Verstraete, M., Waroquiers, D., Wiktor, J., Xu, B., Zhou, A. & Zwanziger, J. (2016), 'Recent developments in the abinit software package', *Comput. Phys. Commun.* **205**, 106 – 131.

Ha, H., Rana, S., Gupta, S., Nguyen, T., Venkatesh, S. et al. (2019), 'Bayesian optimization with unknown search space', *Advances in Neural Information Processing Systems* **32**.

Hartwigsen, C., Goedecker, S. & Hutter, J. (1998), 'Relativistic separable dual-space gaussian pseudopotentials from H to Rn', *Phys. Rev. B* **58**, 3641.

Hohenberg, P. & Kohn, W. (1964), 'Inhomogeneous electron gas', *Phys. Rev.* **136**, B864–B871.

Huan, T. D. (2018), 'Pressure-stabilized binary compounds of magnesium and silicon', *Phys. Rev. Materials* **2**(2), 023803.

Huan, T. D., Mannodi-Kanakkithodi, A. & Ramprasad, R. (2015), 'Accelerated materials property predictions and design using motif-based fingerprints', *Phys. Rev. B* **92**(1), 014106.

Jackson, J. E. (2005), *A user's guide to principal components*, Vol. 587, John Wiley & Sons.

Jones, D. R., Schonlau, M. & Welch, W. J. (1998), 'Efficient global optimization of expensive black-box functions', *Journal of Global optimization* **13**(4), 455–492.

Joseph, V. R. (2016), 'Space-filling designs for computer experiments: A review', *Quality Engineering* **28**(1), 28–35.

Joseph, V. R., Gul, E. & Ba, S. (2015), 'Maximum projection designs for computer experiments', *Biometrika* **102**(2), 371–380.

Kobayashi, Y., Naito, M., Sudoh, K., Gentils, A., Bachelet, C. & Bourcois, J. (2019), 'Formation of crystallographically oriented metastable $Mg_{1.8}Si$ in Mg ion-implanted Si', *Crystal Growth & Design* **19**(12), 7138–7142.

Kohn, W. & Sham, L. (1965), 'Self-consistent equations including exchange and correlation effects', *Phys. Rev.* **140**, A1133–A1138.

Krishna, A., Craig, S. R., Shi, C. & Joseph, V. R. (2022), 'Inverse design of acoustic metasurfaces using space-filling points', *Applied Physics Letters* **121**(7), 071701.

Maddox, J. (1988), 'Crystals from first principles', *Nature* **335**(6187), 201–201.

Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. (2016), 'Machine learning strategy for the accelerated design of polymer dielectrics', *Sci. Rep.* **6**, 20952.

Martoňák, R., Donadio, D., Oganov, A. R. & Parrinello, M. (2006), 'Crystal structure transformations in sio 2 from classical and ab initio metadynamics', *Nature materials* **5**(8), 623–626.

Monkhorst, H. J. & Pack, J. D. (1976), 'Special points for brillouin-zone integrations', *Phys. Rev. B* **13**, 5188.

Morris, M. D. & Mitchell, T. J. (1995), 'Exploratory designs for computational experiments', *Journal of statistical planning and inference* **43**(3), 381–402.

Nguyen, V., Gupta, S., Rane, S., Li, C. & Venkatesh, S. (2017), Bayesian optimization in weakly specified search space, *in* '2017 IEEE International Conference on Data Mining (ICDM)', IEEE, pp. 347–356.

Oganov, A. R. (2011), *Modern methods of crystal structure prediction*, John Wiley & Sons.

Oganov, A. R. & Glass, C. W. (2006), 'Crystal structure prediction using ab initio evolutionary techniques: Principles and applications', *The Journal of chemical physics* **124**(24), 244704.

Oganov, A. R., Lyakhov, A. O. & Valle, M. (2011), 'How evolutionary crystal structure prediction works and why', *Accounts of chemical research* **44**(3), 227–237.

Oganov, A. R., Pickard, C. J., Zhu, Q. & Needs, R. J. (2019), 'Structure prediction drives materials discovery', *Nat. Rev. Mater.* **4**(5), 331–348.

Pannetier, J., Bassas-Alsina, J., Rodriguez-Carvajal, J. & Caignaert, V. (1990), 'Prediction of crystal structures from crystal chemistry rules by simulated annealing', *Nature* **346**(6282), 343–345.

Perdew, J. P., Burke, K. & Ernzerhof, M. (1996), 'Generalized gradient approximation made simple', *Phys. Rev. Lett.* **77**, 3865–3868.

Pickard, C. J. & Needs, R. (2006), 'High-pressure phases of silane', *Physical review letters* **97**(4), 045504.

Pickard, C. J. & Needs, R. (2011), 'Ab initio random structure searching', *Journal of Physics: Condensed Matter* **23**(5), 053201.

Roustant, O., Ginsbourger, D. & Deville, Y. (2012), 'Dicekriging, diceoptim: Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization'.

Santner, T. J., Williams, B. J. & Notz, W. I. (2018), *The design and analysis of computer experiments*, Springer.

Schön, J. C. & Jansen, M. (1996), 'First step towards planning of syntheses in solid-state chemistry: determination of promising structure candidates by global optimization', *Angewandte Chemie International Edition in English* **35**(12), 1286–1304.

Shahriari, B., Bouchard-Côté, A. & Freitas, N. (2016), Unbounded bayesian optimization via regularization, *in* 'Artificial intelligence and statistics', PMLR, pp. 1168–1176.

Siivola, E., Paleyes, A., González, J. & Vehtari, A. (2021), 'Good practices for bayesian optimization of high dimensional structured spaces', *Applied AI Letters* **2**(2), e24.

Stillinger, F. H. (1999), 'Exponential multiplicity of inherent structures', *Physical Review E* **59**(1), 48.

Surjanovic, S. & Bingham, D. (2013), 'Virtual library of simulation experiments: Test functions and datasets', Retrieved August 21, 2022, from `http://www.sfu.ca/~ssurjano`.

Tekin, A., Caputo, R. & Züttel, A. (2010), 'First-principles determination of the ground-state structure of libh 4', *Phys. Rev. Lett.* **104**(21), 215501.

Therrien, F., Jones, E. B. & Stevanović, V. (2021), 'Metastable materials discovery in the age of large-scale computation', *Applied Physics Reviews* **8**(3), 031310.

Trimarchi, G., Freeman, A. J. & Zunger, A. (2009), 'Predicting stable stoichiometries of compounds via evolutionary global space-group optimization', *Phys. Rev. B* **80**(9), 092101.

Tripathy, R., Bilionis, I. & Gonzalez, M. (2016), 'Gaussian processes with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation', *Journal of Computational Physics* **321**, 191–223.

Vu, T. N., Nayak, S. K., Nguyen, N. T. T., Alpay, S. P. & Tran, H. (2021), 'Atomic configurations for materials research: A case study of some simple binary compounds', *AIP Adv.* **11**(4), 045120.

Wales, D. J. & Doye, J. P. (1997), 'Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms', *The Journal of Physical Chemistry A* **101**(28), 5111–5116.

Wang, L., Yerramilli, S., Iyer, A., Apley, D., Zhu, P. & Chen, W. (2022), 'Scalable gaussian processes for data-driven design using big data with categorical factors', *Journal of Mechanical Design* **144**(2).

Weymuth, T. & Reiher, M. (2014), 'Inverse quantum chemistry: Concepts and strategies for rational compound design', *International Journal of Quantum Chemistry* **114**(13), 823–837.

Xiang, H., Huang, B., Kan, E., Wei, S.-H. & Gong, X. (2013), 'Towards direct-gap silicon phases by the inverse band structure design approach', *Phys. Rev. Lett.* **110**(11), 118702.