Adjusting the descriptor for a crystal structure search using Bayesian optimization

Nobuya Sato,[1,*] Tomoki Yamashita,[2,3,†] Tamio Oguchi,[2,3] Koji Hukushima,[2,4] and Takashi Miyake[1,2]

[1] *Research Center for Computational Design of Advanced Functional Materials, National Institute of Advanced Industrial Science and Technology, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan*

[2] *Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan*

[3] *Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan*

[4] *Komaba Institute for Science, University of Tokyo, 3-8-1 Komaba, Meguro, Tokyo 153-8902, Japan*

(Received 14 July 2019; revised manuscript received 24 January 2020; accepted 20 February 2020; published 19 March 2020)

We perform a crystal structure search using Bayesian optimization and evaluate its efficiency with a varying parameter value in the descriptor. Applying the crystal structure search to crystalline silicon shows that the efficiency of the search depends heavily on the parameter value. We find that the efficiency is linked to the distribution of the descriptor. Therefore, we introduce an information measure of the distribution to estimate an appropriate parameter value for performing the crystal structure search efficiently. The measure can also be used to predetermine an appropriate parameter value. The validity of the measure is confirmed with its applications to silicon oxide and yttrium-cobalt alloy.

DOI: 10.1103/PhysRevMaterials.4.033801

I. INTRODUCTION

Computational crystal structure searches are a major challenge in materials science [1]. They provide a powerful tool for finding new materials and identifying uncertain crystal structures. Crystal structure searches find the most stable structure or acceptably low-energy structures for a given chemical composition. From a mathematical perspective, this is a global optimization problem, in which the global minimum or local minima of an acceptably small value of a given objective function is identified. In the crystal structure search, evaluating an objective function corresponds to evaluating the potential energy of a crystal structure and descending to a local minimum during global optimization corresponds to relaxing a crystal structure. We often use first-principles methods for the structure relaxation and energy evaluation to compute the potential energy surface accurately. However, the first-principles calculation is quite heavy for some systems, and the calculation must be repeated during the global optimization. Thus, to reduce the computational cost, it is important to minimize the number of calculations performed to find the global minimum.

Many global optimization techniques have been used in crystal structure searches to reduce the number of structure relaxations and energy evaluations, such as the random search algorithm [2,3], evolutionary algorithm [4,5], and particle swarm optimization [6,7]. In addition, a crystal structure search method using Bayesian optimization (BO) [8] has been developed recently [9]. BO is a global optimization machine learning technique for a black-box function that is not explicitly given or requires a large amount of computation to be evaluated. The objective function to be optimized is the potential energy as a function of a crystal structure here. BO suggests a candidate by balancing exploration of a domain far from the obtained data and exploitation of good obtained data.

A crystal structure is represented by a numerical vector called a descriptor that is used to quantify how similar structures are. The search space in BO is spanned by the descriptor. The descriptor is evaluated from atomic coordinates, although it is not a set of atomic coordinates itself, to satisfy the following two requirements. One is to be the same between identical structures to measure the similarity correctly, which is, in other words, to be invariant under translation and rotation of the system and under permutations of atoms of the same chemical element. The other is to be evaluated only from a crystal structure because the energy is not available before executing BO. The mapping of a crystal structure to a descriptor does not need to be injective, which means that descriptors for different structures can be the same. Many descriptors have been proposed, and they often have parameters that must be predetermined [10–13]. Because the parameters change the descriptors or the BO input, as shown in Fig. 1, the parameters would affect the performance of the crystal structure search. The presence of the parameters in the descriptor raises the following two questions. (i) How strongly does the efficiency of the crystal structure search depend on the choice of values of the parameters? (ii) How are the parameter values predetermined to provide an efficient crystal structure search?

In this paper, to answer these two questions, we examine the dependency of the efficiency of the crystal structure search using BO on the parameter in the descriptor by case studies

[*] n.satou@aist.go.jp

[†] Present address: Top Runner Incubation Center for Academia-Industry Fusion, Nagaoka University of Technology, 1603-1 Kamitomioka-machi, Nagaoka, Niigata 940-2188, Japan.

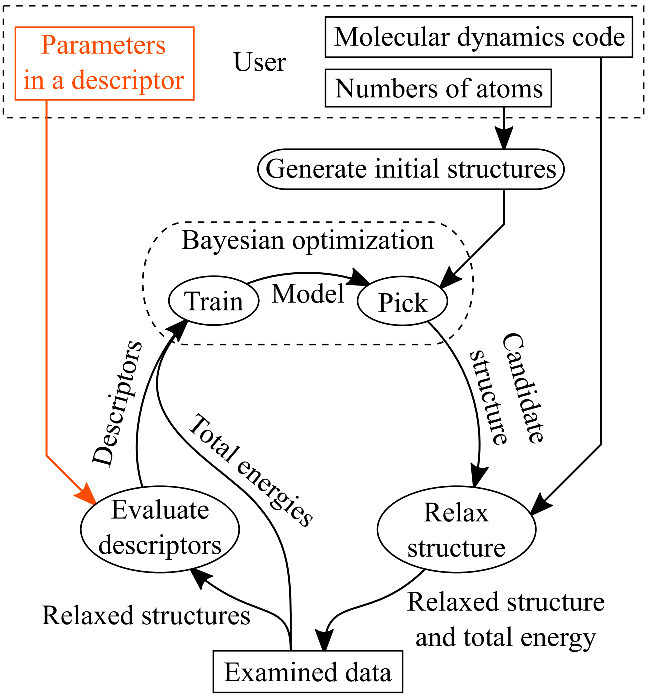

FIG. 1. Flowchart of the crystal structure search using Bayesian optimization (BO). A crystal structure is represented by a descriptor to be treated by BO. Parameters in the descriptor, numbers of atoms, and a molecular dynamics code must be given by a user in advance.

and propose an information measure to determine parameter values. The paper is organized as follows. First, we explain the crystal structure search method and setup for evaluating the efficiency in Sec. II. Next, in Sec. III, we clarify how much a parameter in the descriptor affects the efficiency of the crystal structure search and introduce a related information measure by focusing on crystalline silicon as a model case. We show that the measure can be used to decide a parameter value in the descriptor prior to starting the crystal structure search. The dependency of the efficiency on a parameter and validity of the measure are revealed also for silicon oxide and yttrium-cobalt alloy. Finally, we provide a summary in Sec. IV.

II. METHODS

A. Crystal structure search using Bayesian optimization

We perform the crystal structure search by using the CrySPY code [14], in which an algorithm using BO is implemented [9]. Figure 1 shows the flowchart for the structure search. The code searches for the minimum-energy structure within a pool of crystal structures by using a descriptor and the total energy of a crystal structure as the explanatory and response variables, respectively. Once we give a pool of crystal structures that have identical chemical compositions and the number of atoms per unit cell, the code repeats alternately picking candidates from the pool by BO and relaxing the structures of the candidates. Before performing BO, the descriptors are evaluated for all the structures, and then the descriptors are standardized, namely linearly transformed so that each dimension has a mean of zero and a variance of one.

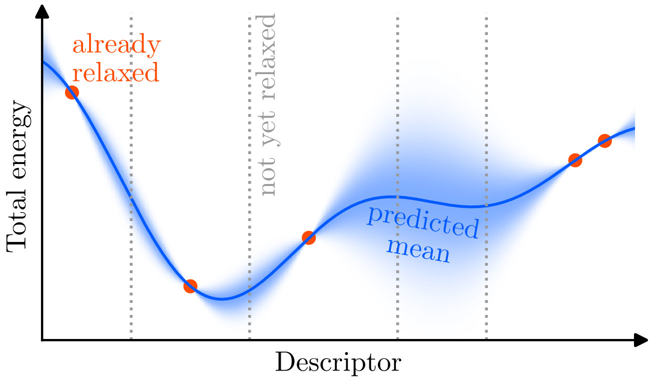

FIG. 2. Schematic view of Bayesian optimization (BO). The descriptor is multidimensional. The circles indicate the training data or descriptors and total energies of already relaxed crystal structures. The opacity of the shading represents the predicted probability of the total energy and the solid line is the mean. Each dotted line indicates a descriptor of a crystal structure that is not yet relaxed. BO predicts total energies for the remaining structures in a probabilistic manner and suggests candidates. A candidate is chosen based on the predicted mean and variance, or expectation value and uncertainty.

We use the F-fingerprint [11] as the descriptor. A schematic of the BO procedures is shown in Fig. 2.

The CrySPY code uses the common Bayesian optimization library (COMBO) [15] to conduct BO. Candidates are selected according to an acquisition function. We use that based on Thompson sampling [16]. The acquisition function is computed from the training data set with the descriptors and total energies already obtained through the Gaussian process. The kernel function in the Gaussian process is the Gaussian kernel with the Euclidean distance, in which hyperparameters are automatically determined by the type II maximum likelihood estimation each time before picking candidates.

The CrySPY code also calls a molecular dynamics code to relax the crystal structures of the selected candidates. The crystal structure relaxation is the most time-consuming part in CrySPY. The CrySPY code can access both *ab initio* and classical molecular dynamics codes. In this study, we use SOIAP [17] and the Vienna Ab initio Simulation Package (VASP) [18]. The ZRL potential [19] is employed in calculations by SOIAP. See Appendix C for details of the ZRL potential. Use of the classical potential, ZRL, reduces computational costs and does not matter to the main aim of the present work of clarifying the dependence of the crystal structure search efficiency on a parameter in the descriptor.

B. Setup for evaluating the efficiency of the crystal structure search

First, we prepare a pool of crystal structures, which consists of 1000 structures generated randomly using the CrySPY code. Next, we repeat alternately performing BO and crystal structure relaxation until we find the minimum-energy structure. BO picks 5 candidates from the pool at a time. Thus, we evaluate the number of structures examined until we find the minimum-energy structure; the fewer the number of structures, the more efficient the crystal structure search. Because the search for the minimum-energy structure is stochastic, we perform 100 minimum searches independently using the same pool of crystal structures and extract summary statistics to evaluate the efficiency.

A component of the descriptor is

$$F_{AB}(r_\alpha) = \sum_{i}^{A,\text{unit cell}} \sum_{j(\neq i)}^{B} \frac{\delta_\sigma(r_\alpha - r_{ij})}{4\pi r_{ij}^2 (N_A N_B / V)\Delta r} - 1, \quad (1)$$

where A and B are chemical elements in a crystal structure, $\sum_i^{A,\text{unit cell}}$ is the summation over A atoms within the unit cell, $\sum_{j(\neq i)}^{B}$ is the summation over B atoms including periodic replicas other than atom i, N_A and N_B are the numbers of A and B atoms in the unit cell, respectively, V is the volume of the unit cell, r_{ij} is the Euclidean distance between atoms i and j, $r_\alpha = r_{\min} + (\alpha - 1)\Delta r$, $\alpha = 1, \ldots, N_{\text{point}}$, $N_{\text{point}} = \lfloor (r_{\max} - r_{\min})/\Delta r \rfloor + 1$, $\lfloor \cdot \rfloor$ is the floor function, and $\delta_\sigma(r)$ is the delta function approximated by the probability density function of the normal distribution with a mean of zero and variance σ^2. If a system is a simple substance, e.g., Si_{16}, the αth component of a descriptor is $F_{\text{SiSi}}(r_\alpha)$. If a system contains multiple chemical elements, a descriptor is a vector constructed by concatenating $F_{AB}(r_\alpha)$'s of all element pairs. For instance, a descriptor of Si_6O_{12} has $3N_{\text{point}}$ components whose first N_{point} components are $F_{\text{SiSi}}(r_\alpha)$'s, middle N_{point} components are $F_{\text{SiO}}(r_\alpha)$'s, and last N_{point} components are $F_{\text{OO}}(r_\alpha)$'s. The four parameters, $r_{\min}$, $r_{\max}$, Δr, and σ, must be given to evaluate the descriptor from a crystal structure. We evaluate the efficiency of the crystal structure search by varying the values of these parameters. The same pool of crystal structures is also used for different parameter values.

III. RESULTS AND DISCUSSION

A. Dependence of efficiency on a parameter in a descriptor

In this and the next subsection, we discuss crystalline silicon, Si_{16}, as a model case. The structure relaxation is performed by SOIAP with the ZRL potential. The diamond-type structure has the lowest energy after relaxing all the structures in the pool. There is only one crystal structure relaxed into the diamond-type structure.

Figure 3 shows the number of structures examined until the minimum-energy structure is found with respect to parameter σ for fixing the remaining parameters, $r_{\min}$, $r_{\max}$, and Δr. The number depends heavily on the value of σ. The minimum search requires a large number of examinations for small and large σ. As a matter of fact, at worst, it is larger than the mean number for the random search of 500.5. The variance of the number is also large in these regions, indicating that efficiency depends strongly on the stochastic process of BO. Such behavior is undesirable for the performance of the crystal structure search. This result demonstrates that the parameters in the descriptor should be adjusted to conduct the crystal structure search efficiently. The mean number is smallest at $\sigma = 16/15$ Å, which is comparable to half of the bond length between Si atoms. This correspondence is probable because σ is the broadening width of the delta function at an atomic position. The broadened delta function can be regarded as an existence probability of a fluctuating atom. In this sense, the descriptor of Eq. (1) represents a crystal structure with fluctuating atoms. The fluctuation would be on the order of a bond length from a physical point of view.

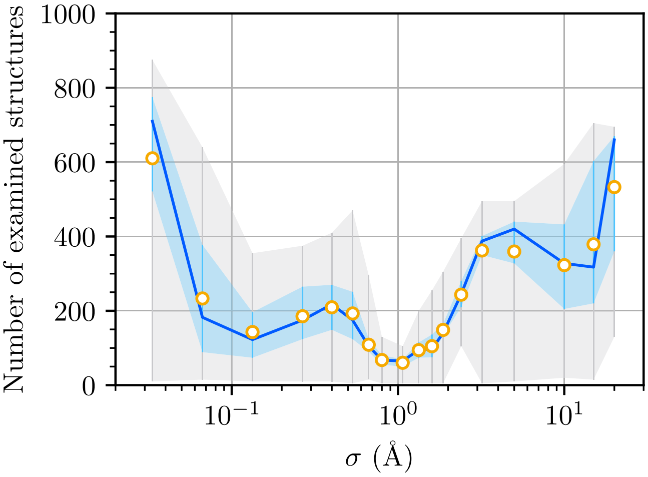

FIG. 3. Number of structures examined until the minimum-energy structure is found with respect to parameter σ in the descriptor. The minimum search is performed for a pool of crystal structures consisting of 1000 randomly generated structures of Si_{16}. The diamond-type structure has the lowest energy after relaxing all the structures in the pool, and there is only one crystal structure relaxed to the diamond-type structure. The minimum search is performed 100 times independently for the same pool. The solid line, inner filled region, and outer filled region indicate the median, lower to upper quartile, and minimum to maximum, respectively. The open circles indicate the means. Parameters other than σ are set as $r_{\min} = 0.5$ Å, $r_{\max} = 20$ Å, and $\Delta r = 0.2$ Å.

When σ is small, function $F(r)$ in Eq. (1), from which the descriptor is sampled, is spiky; therefore, the descriptors are substantially different, even if the crystal structures are similar. In contrast, when σ is large, $F(r)$ is broad, and thus the descriptors are similar even if the crystal structures are substantially different. Thus, parameter σ changes the distribution of the descriptors. These results can also be found in an analysis by agglomerative hierarchical clustering (see Appendix B).

B. Information measure for efficiency estimation

Because the efficiency of the crystal structure search depends strongly on the parameter values, an appropriate value must be determined prior to starting the crystal structure search. The results in Sec. III A suggest that the crystal structure search works efficiently if similarities between pairs of the descriptors are widely distributed from low to high. This implication is also supported by another analysis by agglomerative hierarchical clustering. Therefore, the uniformity of a distribution of similarity between descriptors can be a measure of the efficiency of the crystal structure search, or more specifically, can be used to choose an appropriate parameter value.

We define such a measure in terms of the descriptor $\vec{x}$, which we call the similarity-based information measure (SIM), as

$$S(\{\vec{x}_i\}) = \mathcal{E}[\{s_{ij}\}], \quad (2)$$

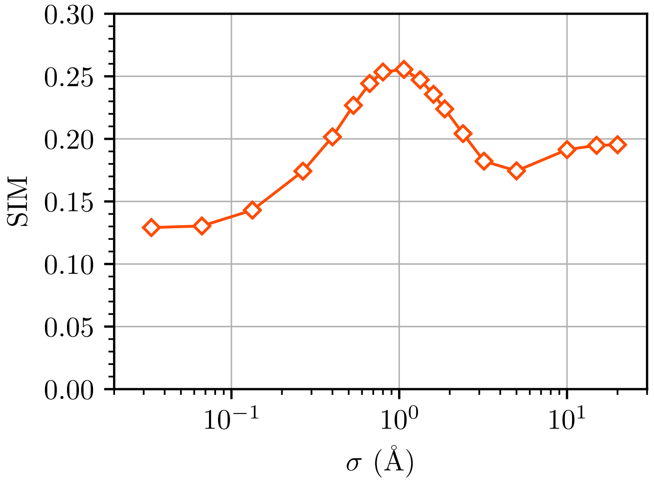

FIG. 4. Similarity-based information measure (SIM) with respect to parameter σ in the descriptor. The set of descriptors is the same as that used in Fig. 3 for each value of σ.

where $\{\vec{x}_i\}$ is a set of standardized descriptors, $\mathcal{E}[\cdot]$ is the cumulative residual entropy (CRE) [20], and $\{s_{ij}\}$ is a set of similarities. The similarity between $\vec{x}_i$ and $\vec{x}_j$ is defined by

$$s_{ij} = \exp\left(-\frac{d_{ij}^2}{2\mathrm{Var}[\{d_{i'j'}\}]}\right), \qquad (3)$$

where $d_{ij} = \|\vec{x}_i - \vec{x}_j\|_2$ and $\mathrm{Var}[\cdot]$ denotes the variance, which is motivated by the Gaussian kernel used in BO. The CRE is one of the information measures extended from the Shannon entropy. If Y follows a non-negative discrete uniform distribution that has a nonzero probability at N points $y_1 < \cdots < y_N$, the CRE is expressed as

$$\mathcal{E}[Y] = -\sum_{k=1}^{N-1} \Delta y_k \left(1 - \frac{k}{N}\right) \ln\left(1 - \frac{k}{N}\right), \qquad (4)$$

where $\Delta y_k = y_{k+1} - y_k$. See Appendix A for details of the CRE. It is practically important that the SIM can be evaluated only from a set of crystal structures without computing their energies. The SIM is one possible definition and there can be other definitions for measuring the efficiency.

Next, we check whether the SIM can be used as a measure of efficiency. Figure 4 shows the SIM with respect to parameter σ. The maximum of the SIM corresponds to the most efficient point, or the minimum of the number of examined structures in Fig. 3. This suggests that we can use the SIM as a measure of efficiency and choose an appropriate parameter value.

C. Application of the information measure for adjusting a parameter

1. Another parameter

Because the SIM measures information about the distribution of descriptors rather than about parameter σ, it is expected that the correspondence of the measure to the efficiency holds also for parameters other than σ. To test this prediction, we use the SIM to adjust Δr under fixed r_{min}, r_{max}, and $\sigma/\Delta r$ as an example. In this case, the number of dimensions of the descriptor is different for each Δr, and a good choice for the value is not as clear as the choice for σ.

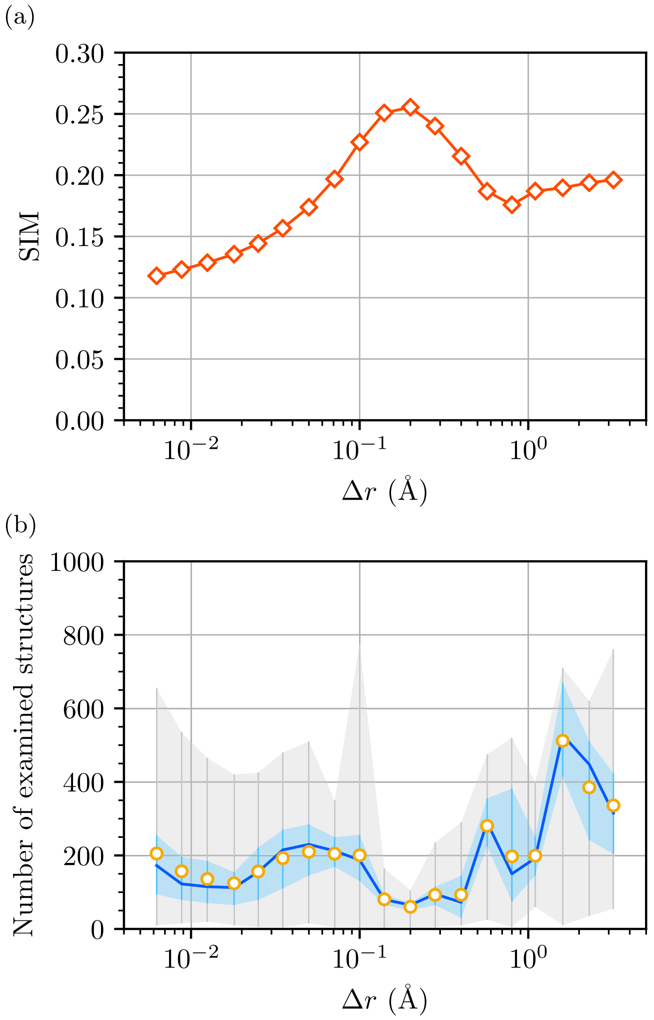

FIG. 5. Adjustment of parameter Δr in the descriptor for Si_{16}. (a) Similarity-based information measure (SIM) with respect to parameter Δr and (b) number of structures examined until the minimum-energy structure is found. The solid line, inner filled region, and outer filled region indicate the median, lower to upper quartile, and minimum to maximum, respectively. The open circles indicate the means. Parameters other than Δr are set as $r_{\mathrm{min}} = 0.5$ Å, $r_{\mathrm{max}} = 20$ Å, and $\sigma = 16\Delta r/3$. Other setups are the same as those of Fig. 3.

Figure 5(a) shows SIMs evaluated along Δr. This result suggests that Δr should be set to 0.2 Å, where the SIM reaches its maximum. We can confirm that the recommended value gives high efficiency [Fig. 5(b)]. This implies that the SIM is generally a good measure for determining parameters in the descriptor.

2. Binary systems

To test whether the measure is applicable to systems other than crystalline silicon, we extend a target of the parameter adjustment to binary systems. Crystal structures of binary systems are basically more complicated than that of crystalline silicon. Therefore, the descriptor needs more dimensions and the crystal structure search is harder. We perform the

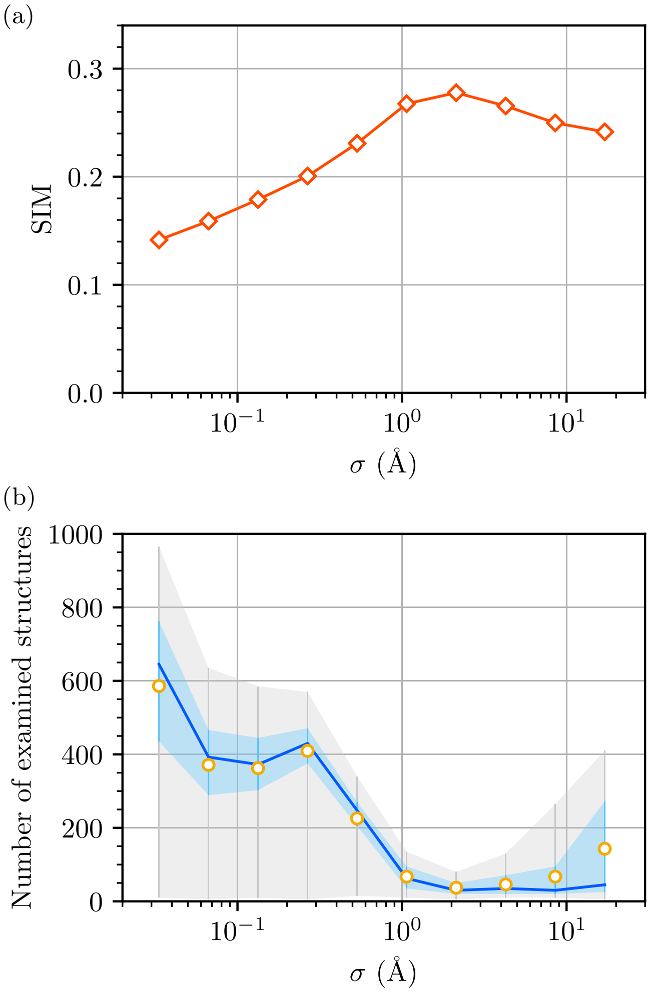

FIG. 6. Adjustment of parameter σ in the descriptor for Si_6O_{12}. (a) Similarity-based information measure (SIM) with respect to parameter σ and (b) number of structures examined until the minimum-energy structure is found. The minimum search is performed for a pool of crystal structures consisting of 1000 randomly generated structures of Si_6O_{12}. There is only one crystal structure relaxed to the minimum-energy structure. The minimum search is performed 100 times independently for the same pool. The solid line, inner filled region, and outer filled region indicate the median, lower to upper quartile, and minimum to maximum, respectively. The open circles indicate the means. Parameters other than σ are set as $r_{\mathrm{min}} = 0.5$ Å, $r_{\mathrm{max}} = 20$ Å, and $\Delta r = 0.2$ Å.

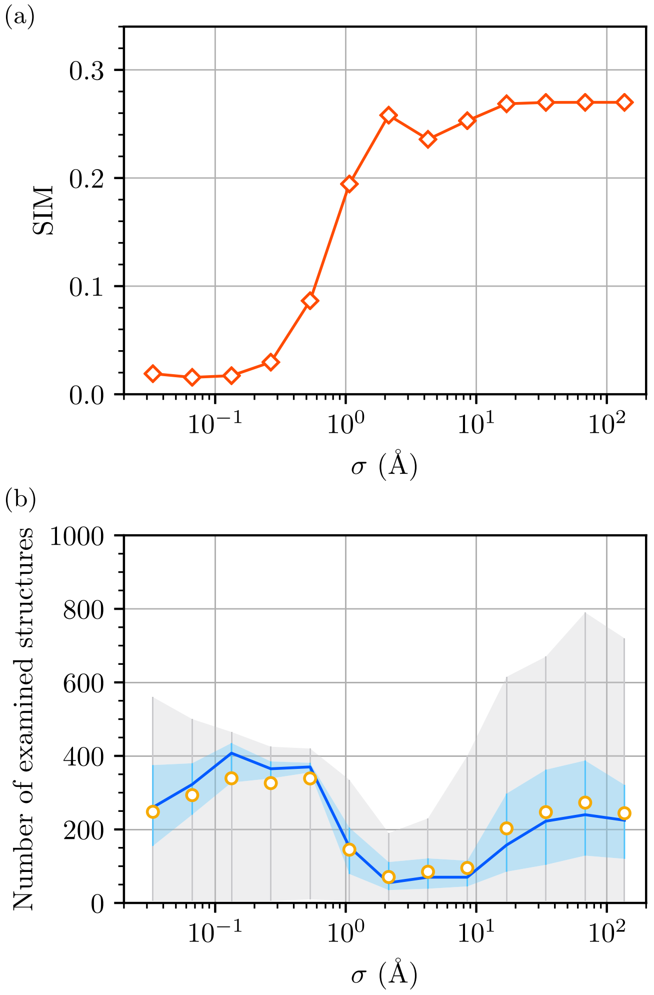

FIG. 7. Adjustment of parameter σ in the descriptor for Y_2Co_{17}. (a) Similarity-based information measure (SIM) with respect to parameter σ and (b) number of structures examined until the minimum-energy structure is found. The minimum search is performed for a pool of crystal structures consisting of 1000 randomly generated structures of Y_2Co_{17}. There are four crystal structures relaxed into the Th_2Zn_{17}-type structure. The minimum search is performed 100 times independently for the same pool. The solid line, inner filled region, and outer filled region indicate the median, lower to upper quartile, and minimum to maximum, respectively. The open circles indicate the means. Parameters other than σ are set as $r_{\mathrm{min}} = 0.5$ Å, $r_{\mathrm{max}} = 20$ Å, and $\Delta r = 0.2$ Å.

parameter adjustment for Si_6O_{12} and Y_2Co_{17}, which have primarily different bonding natures.

Figure 6 shows the results for Si_6O_{12}. The structure relaxation is performed by SOIAP with the ZRL potential. The minimum-energy structure in this pool is not α-quartz which is known to be most stable, although the energy difference between them is less than 1 meV/SiO_2. There is only one crystal structure relaxed to the minimum-energy structure. We can confirm also for this system that the parameter value where the SIM reaches its maximum gives high efficiency.

Figure 7 shows results for Y_2Co_{17}. The structure relaxation is performed by VASP. The minimum-energy structure in this pool is Th_2Zn_{17} type, which is known to be most stable. There are four crystal structures relaxed into the Th_2Zn_{17}-type structure. The SIM has a peak at $\sigma = 32/15$ Å and the crystal structure search is most efficient at that σ. From these two case studies, we found that the SIM could be useful also for a binary system to predetermine a value of a parameter in the descriptor.

As shown in Fig. 7, however, the SIM is highest at $\sigma = 2048/15$ Å and that of the most efficient point is almost equal to but slightly lower than the maximum value. SIMs at these σ's are comparable because distributions of the distance between descriptors, d_{ij} in Eq. (3), are almost the same [Fig. 8(a)]. Meanwhile, d_{ij} itself is different between these σ's as shown in Fig. 8(b). This means that distributions of descriptors are different, hence the different efficiencies.

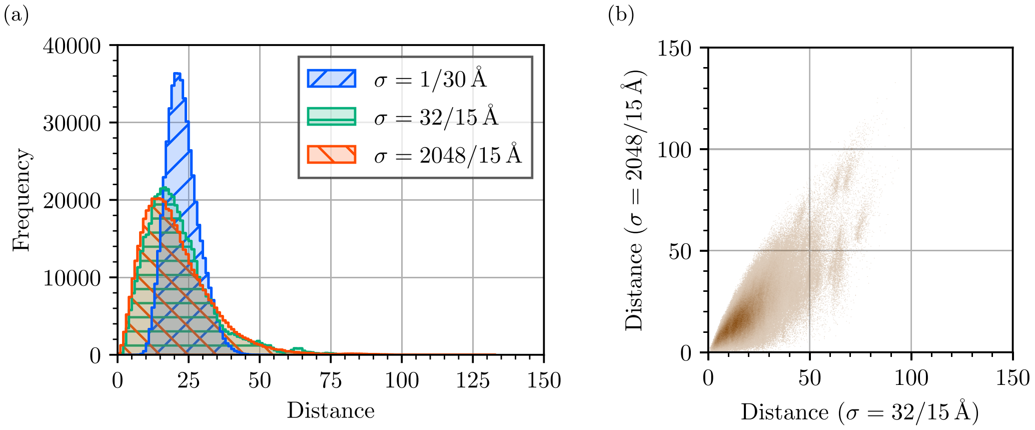

FIG. 8. Distribution of the distance between descriptors for Y_2Co_{17}. (a) Histograms of the distance at typical σ's and (b) correlation of the distance between $\sigma = 32/15$ and $2048/15$ Å. The distance is defined as $d_{ij} = \|\vec{x}_i - \vec{x}_j\|_2$, where $\vec{x}_i$ and $\vec{x}_j$ are the standardized descriptors.

Although the SIM fails to estimate an appropriate parameter value in the case of Y_2Co_{17}, another information measure based on agglomerative hierarchical clustering succeeds (see Appendix B). Agglomerative hierarchical clustering takes account of intercluster distances such as a pair-pair distance other than the pairwise distance. Such a many-body correlation might be important to characterize a distribution of descriptors in the case of Y_2Co_{17}.

We recommend, however, using the SIM rather than the information measure based on clustering. There are several methods for measuring the intercluster distance. The result of clustering depends on the method, and therefore, an estimated parameter value does. In a practical situation, we might be able to estimate an appropriate value by the SIM even in a failed case by finding a local maximum before it is saturated.

IV. CONCLUSION

We performed a crystal structure search using BO for crystalline silicon as a case study. We revealed that the number of structures examined until the minimum-energy structure was found, namely the efficiency of the crystal structure search, depends heavily on a parameter in the descriptor. The efficiency was worse than that of the random search in some cases; therefore, the value of the parameter should be chosen carefully. To predetermine a parameter value in the descriptor, we proposed the SIM as a measure of the efficiency. Because BO uses the energy of a crystal structure as well as the descriptor, no measure defined without the energy can predict the efficiency perfectly. However, we confirmed by case studies of crystalline silicon, silicon oxide, and yttrium-cobalt alloy that the efficiency is high when the SIM reaches its maximum for a parameter, while the most efficient point is located at a local maximum in the case of yttrium-cobalt alloy. Hence, we can adjust a parameter by searching for the maximum of the measure.

In addition to the crystal structure search discussed here, BO has recently been applied to some areas of materials sciences, such as virtual screening of material databases [21]. A common issue in BO is the need to predetermine the descriptor and its parameter. Once a descriptor is chosen, the present work provides a promising way of determining a suitable parameter value for the descriptor.

ACKNOWLEDGMENTS

This work was supported in part by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of Japan as a social and scientific priority issue (Creation of new functional Devices and high-performance Materials to Support next-generation Industries; CDMSI) to be tackled by using a post-K computer and by the Materials research by Information Integration Initiative (MI^2I) project of the Support Program for Starting Up Innovation Hub from the Japan Science and Technology Agency (JST). Part of the computation in this work was performed using the facilities of the Supercomputer Center of the Institute for Solid State Physics at the University of Tokyo.

APPENDIX A: PROPERTIES OF THE CRE

Because we define a measure for a discrete uniform distribution, an information measure used in the definition should be valid for the discrete probability distribution. The CRE is one such information measure, which is defined for a nonnegative random variable X by

$$\mathcal{E}[X] = -\int_0^\infty dx\, \overline{F}(x) \ln \overline{F}(x), \tag{A1}$$

where $\overline{F}(x) = \Pr(X > x)$ is the complementary cumulative distribution function (CCDF). When X follows a discrete uniform distribution, Eq. (A1) results in Eq. (4). The CRE is valid for the discrete probability distribution owing to the CCDF. In contrast, the differential entropy $h[X]$, which is also an extension of the Shannon entropy, cannot be evaluated for the discrete probability distribution because it is defined by using the probability density function $f(x)$ as $h[X] = -\int dx\, f(x) \ln f(x)$. Although the Shannon entropy $H[X]$ can be evaluated for the discrete probability distribution, the value for a discrete uniform distribution depends only on the number of points N as $H[X] = -\sum_{i=1}^{N} \Pr(X = x_i)$

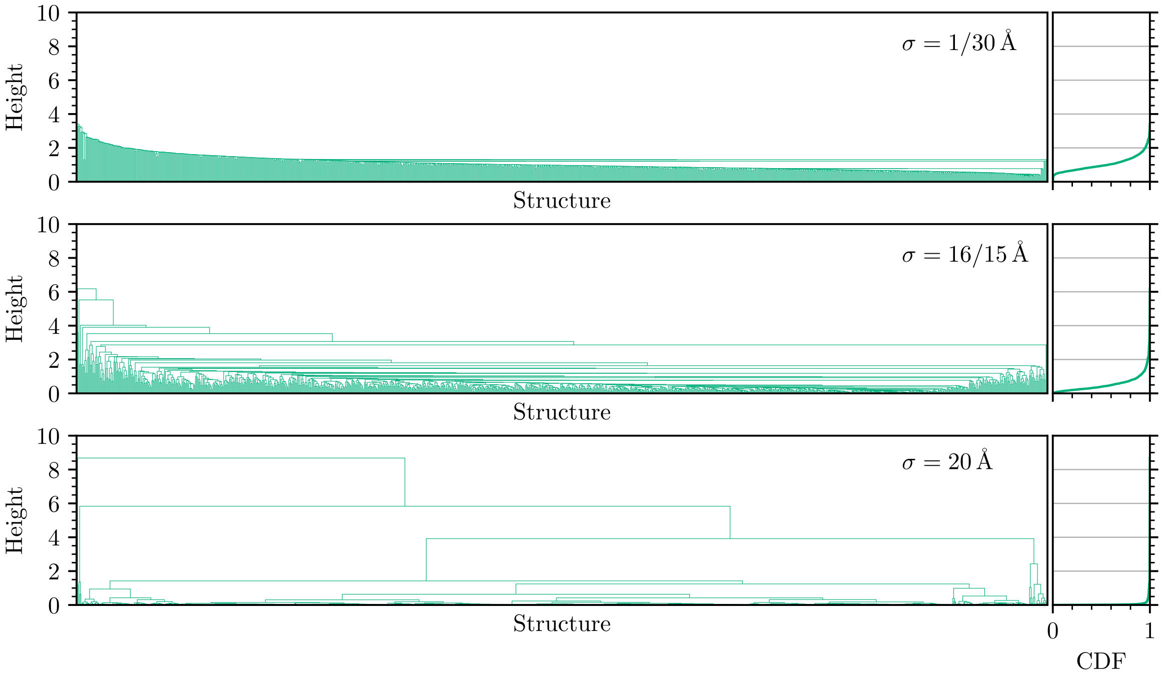

FIG. 9. Dendrograms for Si_{16} at typical σ's. The right-hand panels shows the cumulative distribution function (CDF) as a function of the height of the branch point. Other parameters are set as $r_{\min} = 0.5$ Å, $r_{\max} = 20$ Å, and $\Delta r = 0.2$ Å.

$\ln \Pr(X = x_i) = \ln N$, and thus it is not suitable for use in the definition of a measure.

The CRE is independent from the shift of a random variable [22]. If random variable Y is the linear transformation of random variable X as $Y = aX + b$ with $a > 0$ and $b \geqslant 0$, the CRE of Y is expressed as

$$\mathcal{E}[Y] = a\,\mathcal{E}[X], \tag{A2}$$

which does not depend on shift b.

We show the following two examples of the CRE from Ref. [20].

(1) $\mathcal{E}[X] = a/4$ if random variable X follows the continuous uniform distribution supported on $[0, a]$.

(2) $\mathcal{E}[X] = 1/\lambda$ if random variable X follows the exponential distribution with the mean $1/\lambda$.

Both examples show that the CRE is large when the variable is distributed over a wide range.

APPENDIX B: ANALYSIS BY AGGLOMERATIVE HIERARCHICAL CLUSTERING

We introduce another measure for the efficiency of the crystal structure search based on the dendrogram. Possible crystal structure candidates in the pool used in BO are characterized by a d-dimensional vector of the descriptor. The distance between these crystal structures is evaluated from the descriptors, which depends on the parameters we use in the descriptor. We perform agglomerative hierarchical clustering of set $\{\vec{x}_i/c\}$ by the average linkage method with the Euclidean distance, where $\vec{x}_i$ is the standardized descriptor of the ith candidate, $c = \sqrt{2}\,\Gamma[(d+1)/2]/\Gamma(d/2)$, and $\Gamma(\cdot)$ is the gamma function. The scaling factor, c, is the expectation value of $\|(X_1 \cdots X_d)^{\mathrm{T}}\|_2$ in the case that $X_1, \ldots, X_d$ are independent standard normal random variables. Figure 9 presents typical examples of the dendrogram for three values of parameter σ in the descriptor for Si_{16}. Leaves under the same branch point are a set of similar crystal structures within the distance determined by the height of the branch point in the dendrogram. The tree structure of the dendrogram depends strongly on the value of parameter σ. The branch point decreases on average as the parameter values increases, meaning that the distance between structures is smaller, and the branch point is distributed in a narrower range with small and large parameter values, meaning that the distances between structures are almost identical. These features are also found in the cumulative distribution function in the right-hand panels of Fig. 9.

BO is expected to work efficiently when the distance between a pair of structures in the pool varies from short to long, as mentioned in Sec. III A. As a measure of randomness of the branch point, we define the dendrogram-based information measure (DIM) as

$$D(\{\vec{x}_i\}) = \mathcal{E}[\{h_k\}], \tag{B1}$$

which is the CRE of height h_k of a branch point. As shown in Fig. 10, DIMs reach their maximum around which the efficiencies of crystal structure searches are highest. The DIM, as well as the SIM discussed in Sec. III B, can also be used as a measure of efficiency.

The agglomerative hierarchical clustering relies on the linkage method and distance metric, which affect the resulting dendrogram. Consequently, the value of the DIM and its maximum depend on the details of the linkage method and the metric employed. Furthermore, the DIM also depends on dimension d of the descriptors. Thus, when the parameter in the descriptor changes d, an appropriate scaling of the descriptor is necessary to compare DIMs for different parameter values. Consequently, scaling factor c defined above is introduced.

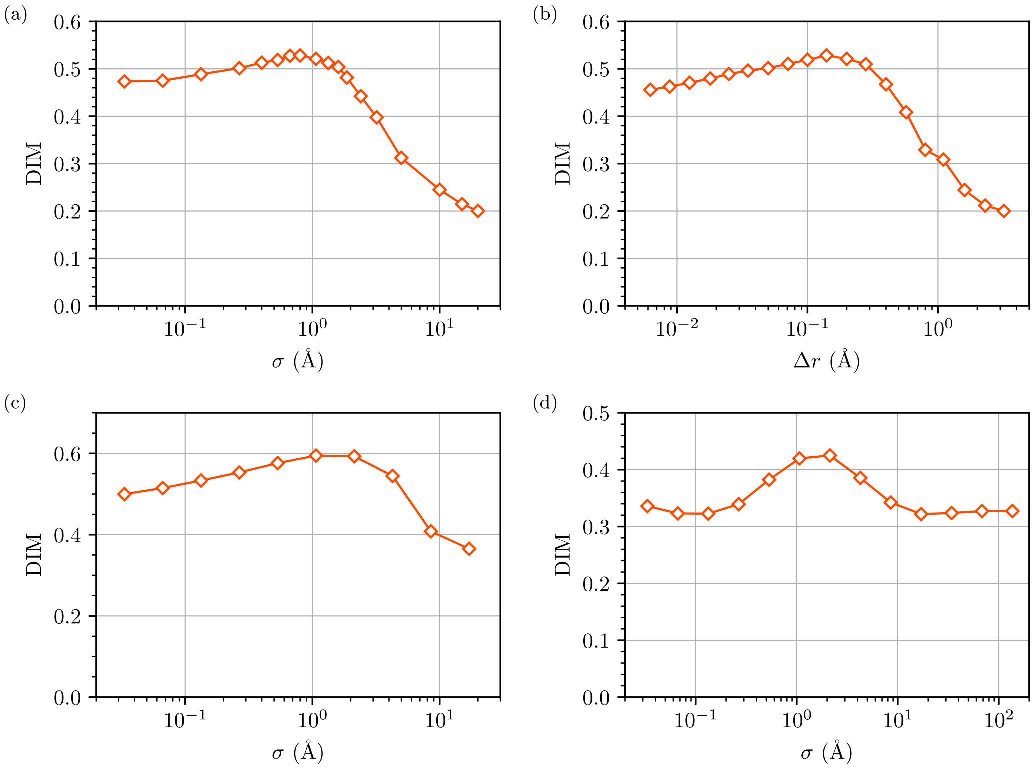

FIG. 10. Dendrogram-based information measure (DIM) for (a), (b) Si_{16}; (c) Si_6O_{12}; and (d) Y_2Co_{17}. Setups are the same as those in the main text.

This is in sharp contrast to the SIM. Therefore, we recommend that the SIM rather than the DIM be used as a measure of efficiency.

APPENDIX C: ZRL POTENTIAL

The ZRL potential [19] is an empirical interatomic potential for the Si-N-O-H system derived from the Tersoff potential. The energy is expressed as

$$E = \frac{1}{2}\sum_{i}^{\text{unit cell}}\sum_{j\,(\neq i)} V_{ij} + \sum_{I} N_I E_I^0 + \sum_{i} E_i^{\mathrm{c}}, \qquad \text{(C1)}$$

where i and j are the indices of atoms, I is the index for the chemical element of the ith atom, V_{ij} is the generalized Morse potential, E_I^0 is the core energy of the Ith element, N_I is the number of atoms of the Ith element, and E_i^{c} is the penalty for under- and overcoordination of the ith atom. E_i^{c} is given by

$$E_i^{\mathrm{c}} = c_I^{(1)}\Delta z_i + c_I^{(2)}(\Delta z_i)^2, \qquad \text{(C2)}$$

where $c_I^{(1)}$ and $c_I^{(2)}$ are parameters,

$$\Delta z_i = \frac{z_i - z_I^0}{|z_i - z_I^0|} f_{\mathrm{s}}\left(\left|z_i - z_I^0\right|\right) \qquad \text{(C3)}$$

is the deviation from the expected coordination number z_I^0, and $f_{\mathrm{s}}(\cdot)$ is the switching function. Definitions not shown here and values of parameters are given in Ref. [19].

We modify the original definition of the switching function given in Eq. (17) of Ref. [19] to remove discontinuity and nonmonotonicity. Our definition is given by

$$f_{\mathrm{s}}(z) = \begin{cases} \lfloor z \rfloor, & \text{if } \{z\} \leqslant z_{\mathrm{T}} - z_{\mathrm{B}}, \\ \lfloor z \rfloor + \frac{1}{2}\left[1 + \sin\left(\frac{\pi}{2}\frac{\{z\} - z_{\mathrm{T}}}{z_{\mathrm{B}}}\right)\right], & \\ & \text{if } z_{\mathrm{T}} - z_{\mathrm{B}} < \{z\} \leqslant z_{\mathrm{T}} + z_{\mathrm{B}}, \\ \lfloor z \rfloor + 1, & \text{if } z_{\mathrm{T}} + z_{\mathrm{B}} < \{z\}, \end{cases} \qquad \text{(C4)}$$

where $z \geqslant 0$, z_{T} and z_{B} are parameters depending on the chemical element, $\lfloor \cdot \rfloor$ is the floor function, and $\{z\} = z - \lfloor z \rfloor$ is the fractional part of z. The original form of $f_{\mathrm{s}}(z)$ is not continuous at $z = z_{\mathrm{T}} \pm z_{\mathrm{B}}, 1, 2, \ldots$ and monotonically decreases over $z \in (z_{\mathrm{T}} - z_{\mathrm{B}}, z_{\mathrm{T}} - z_{\mathrm{B}}/2)$ and $(z_{\mathrm{T}} + z_{\mathrm{B}}/2, z_{\mathrm{T}} + z_{\mathrm{B}})$, which causes discontinuity for the potential energy surface in some situations. In contrast, Eq. (C4) is continuous and monotonically increases for all z, resulting in a continuous potential energy surface even for the case in which the original form produces discontinuity.

[1] J. Maddox, Nature (London) **335**, 201 (1988).

[2] C. J. Pickard and R. J. Needs, Phys. Rev. Lett. **97**, 045504 (2006).

[3] C. J. Pickard and R. J. Needs, J. Phys.: Condens. Matter **23**, 053201 (2011).

[4] A. R. Oganov and C. W. Glass, J. Chem. Phys. **124**, 244704 (2006).

[5] A. O. Lyakhov, A. R. Oganov, H. T. Stokes, and Q. Zhu, Comput. Phys. Commun. **184**, 1172 (2013).

[6] Y. Wang, J. Lv, L. Zhu, and Y. Ma, Phys. Rev. B **82**, 094116 (2010).

[7] H. Wang, Y. Wang, J. Lv, Q. Li, L. Zhang, and Y. Ma, Comput. Mater. Sci. **112**, 406 (2016).

[8] D. R. Jones, M. Schonlau, and W. J. Welch, J. Global Optim. **13**, 455 (1998).

[9] T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, and T. Oguchi, Phys. Rev. Mater. **2**, 013803 (2018).

[10] E. L. Willighagen, R. Wehrens, P. Verwer, R. de Gelder, and L. M. C. Buydens, Acta Crystallogr. B: Struct. Sci. **61**, 29 (2005).

[11] A. R. Oganov and M. Valle, J. Chem. Phys. **130**, 104504 (2009).

[12] J. Behler, J. Chem. Phys. **134**, 074106 (2011).

[13] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, Phys. Rev. B **89**, 205118 (2014).

[14] See https://github.com/Tomoki-YAMASHITA/CrySPY.

[15] T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi, and K. Tsuda, Materials Discovery **4**, 18 (2016).

[16] O. Chapelle and L. Li, in *Advances in Neural Information Processing Systems 24*, edited by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Curran Associates, Inc., 2011), p. 2249.

[17] See https://github.com/nbsato/soiap.

[18] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).

[19] S. R. Billeter, A. Curioni, D. Fischer, and W. Andreoni, Phys. Rev. B **73**, 155329 (2006); **79**, 169904(E) (2009).

[20] M. Rao, Y. Chen, B. C. Vemuri, and F. Wang, IEEE Trans. Inf. Theory **50**, 1220 (2004).

[21] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, Phys. Rev. Lett. **115**, 205901 (2015).

[22] A. Di Crescenzo and M. Longobardi, J. Stat. Plan. Inference **139**, 4072 (2009).