Adaptive sampling methods via machine learning for materials screening

Akira Takahashi (a), Yu Kumagai (a), Hirotaka Aoki (a), Ryo Tamura (b,c) and Fumiyasu Oba (a)

(a) Laboratory for Materials and Structures, Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Japan; (b) International Center for Materials Nanarchitectonics (WPI-MANA), National Institute for Materials Science, Ibaraki, Japan; (c) Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan

ABSTRACT
High-throughput virtual screening by using a combination of first-principles calculations and Bayesian optimization (BO) has attracted much attention as a method for efficient material exploration. The purpose of the virtual screening is often to search for the materials whose properties meet a certain target criterion, while the conventional BO aims to find the global extremum. Some recent works use the conventional BO by converting target properties for such motivation. On the other hand, an adaptive sampling method, where the acquisition function is based on the probability that a data point achieves a target property within a specific range, is suggested previously [Kishio et al., Chemom. Intell. Lab. Syst. 127, 70 (2013)]. In this paper, we demonstrate that such adaptive sampling is effective for the exploration of the materials whose properties meet target criteria. We conducted simulations of material exploration using an in-house database constructed by first-principles calculations and compared the performance of the adaptive sampling and conventional BO approaches. Furthermore, we evaluate and discuss the performance of acquisition functions extended to multi-objective problems for material exploration, considering multiple-target properties simultaneously.

1. Introduction
High-throughput virtual screening by using first-principles calculation is a powerful tool for exploring materials with desired properties [1–4]. The recent improvement of computers and algorithms enables us to cover a large search space in the screening and to construct a computational material database possessing more than several hundred thousand or even million entries [5–7]. However, when the number of candidate materials is extremely large and/or computationally demanding approximations such as hybrid functionals [8] and the GW approximation [9] are required for reliable prediction of target material properties, it is still unfeasible to perform brute-force first-principles calculations for all candidate materials because of the high computational cost.

Machine learning techniques are expected to accelerate high-throughput virtual screening by constructing surrogate models and making fast predictions, thereby overcoming this situation [10–17]. Recently, adaptive sampling or black-box optimization, which suggests data points to be added by repeating the construction of a surrogate model and the acquisition of promising data points with demanding evaluation such as first-principles calculations, attracts much attention as

CONTACT
Akira Takahashi takahashi.a.bb@m.titech.ac.jp Laboratory for Materials and Structures, Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8503, Japan

© 2022 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
a method for rapidly discovering optimal data points [18,19]. Bayesian optimization (BO), which is a global optimization method based on adaptive sampling, is one of the most commonly used methods in the material exploration [18–27]. Moreover, a multi-objective method based on BO has been developed for the search of the Pareto-front and applied to the exploration of materials in terms of multiple properties [27–30].

BO is designed for global optimization, the purpose of which is to obtain the data point whose objective value is at the maximum or minimum. However, the purpose of the high-throughput screening is often to collect materials whose target properties meet particular criteria, rather than discovering a few materials whose target properties are extremely large or small. Even if the obtained materials do not have the maximum/minimum target properties, they might still deserve further considerations, such as detailed simulations and synthetic experiments. Examples include the exploration of semiconductors possessing band gaps in a target energy range. In another case, it would be sufficient if a material property is above/below a specific criterion. For example, we often consider materials to be stable if their formation energies are negative. To meet these requirements, some researchers have transformed the target property by a conversion function to search for materials holding band gaps within a desirable range [24,26].

On the other hand, a method for efficiently acquiring a data point holding a target property within a specific range has been developed to determine appropriate experimental conditions quickly [31,32]. Another work has reported a method for searching for data points whose target properties are above/below a specific threshold, specifically, to collect sets of lattice mismatches and interface energies in ‘a low error region’, in which the error in the aspect ratio of the precipitate from experiment and computation is lower than a specific threshold [33]. Recently, such a method has been extended to multi-objective problems, i.e. exploration of data points achieving multiple target properties in specific ranges. It has been applied to the optimization of equipment and operating conditions of ethylene oxide production plants to control multiple properties such as reactor inlet concentration, conversion ratio and product limitations [34]. These methods are expected to be effective in material search under the aforementioned conditions.

In this paper, we report on the development of an adaptive sampling method for high-throughput virtual screening, which enables us to efficiently find materials having a target property within a specific range, based on the previous works [31,33]. This method uses the probability that objective values fall within a specific range as an acquisition function, which determines the priority of data point selection. We assessed the performance of both the conventional BO and our method by conducting simulations of material exploration using an in-house computational database, which contains the formation energies and band gaps of 2,018 nitrides and pnictides obtained by high-throughput first-principles calculations. Moreover, we evaluate the performance of two acquisition functions extended to multiple target properties: one is a simple product of single-objective acquisition functions for respective target properties, and the other is the acquisition function proposed by Iwama and Kaneko [34]. These methods can be applied to the exploration of materials that simultaneously satisfy multiple criteria; for example, the band gap is in a target range and the formation energy is sufficiently low. We also discuss the performance of these adaptive sampling methods for multiple conditions using our computational database.

2. Methodology

2.1. Gaussian process regression

In this work, we employed Gaussian process (GP) regression, as implemented in the PHYSBO code [27] to construct a surrogate model and estimate the mean and the standard deviation of the prediction for adaptive sampling.

To perform the GP regression, we first need to determine the covariance (or kernel) function, which defines the similarity of two points in the descriptor space. In this study, we employed the radial distribution function, which is given by

$$k(x, x') = \exp \left[ -\frac{1}{2\eta^2} \left( x - x' \right)^2 \right],$$

(1)

where $x, x'$ are arbitrary points in the descriptor space and $\eta$ is a hyperparameter. Let us assume that we know data points $D = \{x_i, y_i\}_{i=1,N}$, where $x_i$ and $y_i$ are the descriptor and objective variable of $i$th data, respectively. The mean and standard deviation of the prediction of an arbitrary descriptor $x$, $\mu(x)$, and $\sigma(x)$, respectively, are estimated as

$$\mu(x) = k(x)^T (K + s^2I)^{-1} y,$$

(2)

$$\sigma(x) = k(x, x) + s^2 - k(x)^T (K + s^2I)^{-1} k(x),$$

(3)

where $s$ is a hyperparameter, $I$ is the identity matrix, $y = [y_1, y_2, \ldots, y_N]^T$, $k(x) = [k(x, x_1), k(x, x_2), \ldots, k(x, x_N)]^T$ and $K$ is a kernel matrix denoted as

$$K = \begin{bmatrix}
  k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) \\
  k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) \\
  \vdots & \vdots & \ddots & \vdots \\
  k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N)
\end{bmatrix}.$$  

(4)

The hyperparameters are optimized by maximizing the type-II likelihood [35].
2.2. Conventional Bayesian optimization

As mentioned above, BO is a method for global optimization of a black-box function by repeating the surrogate model construction by machine learning and the data point selection based on the machine learning prediction. In BO, the acquisition function determines the priority of data point selection. In this study, we employed two typical acquisition functions in the conventional BO: probability improvement (PI) and expected improvement (EI).

Based on the GP model, the probability density function of target value $y$ at the data point with descriptor $x$ is represented by the normal distribution function, namely,

$$p(y) = \frac{1}{\sqrt{2\pi} \sigma(x)} \exp \left( -\frac{1}{2 \sigma(x)^2} (y - \mu(x))^2 \right).$$

(5)

PI is the probability of the objective value at descriptor $x$ being higher than any known data points and denoted as

$$PI(x) = \int_{-\infty}^{\infty} p(y)dy = \Phi \left( \frac{\mu(x) - y_{\max}}{\sigma(x)} \right),$$

(6)

where $y_{\max}$ and $\Phi$ are the highest objective value of the known data points and the cumulative distribution function of the normal distribution, respectively. On the other hand, EI is the expected value of the improvement, namely,

$$EI(x) = \int_{y_{\max}}^{\infty} (y - y_{\max}) p(y)dy = \left[ \mu(x) - y_{\max} \right] \Phi \left( \frac{\mu(x) - y_{\max}}{\sigma(x)} \right) + \sigma(x)p(y_{\max}).$$

(7)

Although using $y_{\max}$ for the acquisition function is reasonable and straightforward for global optimization, this would be inappropriate to collect data points that meet a criterion because of the following two reasons. (1) If $\mu(x)$ is smaller than $y_{\max}$, these functions give priority to data points with large $\sigma(x)$ rather than data points with preferable $\mu(x)$. Therefore, if once $y_{\max}$ becomes much higher than the criteria, the conventional BO prioritizes data points with large $\sigma(x)$ values over data points with sufficiently and moderately large $\mu(x)$ and small $\sigma(x)$ values. (2) If the maximum of target property or conversion function is finite and once a data point with the maximum property is found, "probability/expected value of exceeding $y_{\max}$" would be theoretically meaningless because the target values of any other data points do not exceed $y_{\max}$.

2.3. Acquisition function for a target range and its extension to the consideration of multiple objectives

For our purpose, it would be straightforward to use the probability that a value falls within a target range as proposed by Kishio et al. [31], given as

$$p(y_{\text{lower}} \leq y(x) \leq y_{\text{upper}}) = \int_{y_{\text{lower}}}^{y_{\text{upper}}} p(y)dy = \Phi \left( \frac{y_{\text{upper}} - \mu(x)}{\sigma(x)} \right) - \Phi \left( \frac{y_{\text{lower}} - \mu(x)}{\sigma(x)} \right),$$

(8)

where $y_{\text{lower}}$ and $y_{\text{upper}}$ are the lower and upper limits of the target range, respectively. If we consider only one of $y_{\text{lower}}$ or $y_{\text{upper}}$, i.e. $y_{\text{upper}} = +\infty$ or $y_{\text{lower}} = -\infty$, these functions can be denoted as

$$p(y \leq y(x)) = \int_{y_{\text{lower}}}^{+\infty} p(y)dy = 1 - \Phi \left( \frac{y_{\text{upper}} - \mu(x)}{\sigma(x)} \right),$$

(9)

$$p(y(x) \leq y_{\text{upper}}) = \int_{-\infty}^{y_{\text{upper}}} p(y)dy = \Phi \left( \frac{y_{\text{upper}} - \mu(x)}{\sigma(x)} \right),$$

(10)

respectively, as suggested by Tsukada et al. [33]. In this paper, we call these acquisition functions 'Probability that a value falls within a Target Range (PTR)'

The functional form of the PTR is similar to that of the PI. The main difference from the PI is whether thresholds are fixed or updated in each exploration step. The fixed threshold enables us to properly prioritize the promising data points whose target properties are likely to be moderate and sufficiently large/small, even if we acquire data with much larger/smaller target properties than the thresholds. As another advantage of the PTR, we can set the target thresholds without any conversion functions.

Furthermore, such methods can be extended to multiple objectives, aiming to explore materials that meet multiple conditions simultaneously (e.g. materials holding both band gaps within a target range and energies less than a certain criterion). Let us assume that the purpose is to collect the materials whose $n$ target properties $y_1, y_2, \ldots, y_n$ simultaneously satisfy the following inequalities:

...
\[ y_{\text{lower},i} \leq y_i \leq y_{\text{upper},i} \quad (1 \leq i \leq n), \]  

where \( y_{\text{lower},i} \) and \( y_{\text{upper},i} \) are the lower and upper limits of the target range for property \( y_i \). Then, the PTR of each inequality can be calculated by performing GP regression for all of the target properties. Although the covariance of the prediction of each \( y_i \) can be calculated by a multi-output GP [36], we evaluate the probability of satisfying all inequalities by constructing multiple single-output GP models for each \( y_i \) and assume the prediction values are independent. This not only simplifies the method but also enables us to use a different descriptor set for each GP model (e.g., we can select descriptors according to the importance of each target property as discussed in Section 2.3). We denote the descriptor set for the \( i \)th GP model as \( x_i \) and the mean and standard deviation of the posterior distribution of the GP model as \( \mu_i(x_i) \) and \( \sigma_i(x_i) \), respectively. Then, the probability of satisfying Eq. (11) can be expressed as

\[
\prod_{i=1}^{n} p(y_{\text{lower},i} \leq y_i(x_i) \leq y_{\text{upper},i}) = \prod_{i=1}^{n} \Phi\left(\frac{y_{\text{upper},i} - \mu_i(x_i)}{\sigma_i(x_i)}\right) - \Phi\left(\frac{y_{\text{lower},i} - \mu_i(x_i)}{\sigma_i(x_i)}\right),
\]

Furthermore, Iwama and Kaneko have proposed a standardized multi-objective acquisition function in a recent work [34]. The differences between Eq. (12) and their method are twofold: (i) logarithm value is used to prevent the acquisition function from being extremely small and (ii) min-max normalization is applied to balance the scale of each \( p(y_{\text{lower},i} \leq y_i(x_i) \leq y_{\text{upper},i}) \). Their standardized acquisition function is denoted as

\[
\sum_{i=1}^{n} \log\left(\frac{p(y_{\text{lower},i} \leq y_i(x_i) \leq y_{\text{upper},i}) - p_{\text{min},i}}{p_{\text{max},i} - p_{\text{min},i}}\right),
\]

where \( p_{\text{max},i} \) and \( p_{\text{min},i} \) are the maximum and minimum values of \( p(y_{\text{lower},i} \leq y_i(x_i) \leq y_{\text{upper},i}) \) of the known data set, respectively. In this paper, the acquisition functions expressed by Eqs. (12) and (13) are named ‘unstandardized multi-objective probability that values fall within target ranges (PTR)’ and ‘standardized multi-objective PTR’, respectively.

### 2.4. Descriptors

To construct the GP regression model, we considered 185 compositional descriptors implemented in the matminer code [37]. The detail of these descriptors can be found in Table S1 in the Supplemental Materials.

In general, relevant descriptors greatly depend on material properties (e.g., electronic dielectric constants of stable oxides can be accurately predicted by only compositional descriptors, while the prediction of ionic dielectric constants is greatly improved by the addition of structural descriptors [38]). Then, it is important to use different descriptors depending on the material property in order to realize better predictions. The number of descriptors often reaches several tens or one hundred [12,14,15,17,27]. In some cases, especially when crystal structure is available in advance, even several hundreds of descriptors can be considered [38–41]. However, the use of such large numbers of descriptors, thereby increasing the dimension of the search space, may cause the overfitting of the surrogate model and complicate the search. Therefore, in each iteration step of adaptive sampling, we select 10 descriptors based on the permutation importance in the random forest regression [42,43]. This selection is conducted using the scikit-learn library [44]. The effect of the pruning method is evaluated by a dataset as stated below and discussed in Section S2 in the Supplemental Materials. An overview of our adaptive sampling method is illustrated in Figure 1.

### 3. Computational database for performance evaluation

#### 3.1. Data set

To assess the performance of our methods compared with the conventional BO, we used our in-house database. The database is constructed by first-principles calculations of the formation energies (\( E_f \)) and band gaps (\( E_g \)) of nitrides, pnictides and related substances that are selected using the following criteria; hereafter we call these materials nitrides and pnictides for simplicity, including those with uncommon stoichiometries. First, we collected candidate materials from the Materials Project database [5]. We selected the candidate substances satisfying the following conditions: (1) substances include at least one of N, P, As, Sb, and Bi whose oxidation number is not necessarily −3; (2) substances neither include H, O, halogens, noble gases, lanthanoids except La and Ce nor actinides; (3) substances are not simple (elementary) substances; and (4) substances with space group P1 and/or containing more than 40 atoms in the primitive unit cell are excluded due to their high computational costs and/or ambiguities of the crystal structures. The total number of candidates for the first-principles calculations was 5838.

To obtain the formation energies and band gaps of these candidate materials, we use the workflow shown in Figure 2. The computational details of the rough structure optimization can be found in the
Supplemental Materials, while those of the strict optimization of structure and k-mesh density are the same as our previous work on the construction of an oxide database [38] except for the force convergence criterion. We employed the projector augmented-wave (PAW) method [45] as implemented in the VASP code [46,47] for the first-principles calculations. The structure optimizations were performed using 1.3 times the largest default plane-wave cutoff energies for the PAW datasets (ENMAX). The formation energies were estimated for optimized structures with a cutoff energy of 520 eV. The residual atomic forces and stresses of the rough structure optimization are less than 0.2 eV/Å and 6 GPa, while those of the strict optimization 5 meV/Å and 0.1 GPa, respectively. The formation energies and the relaxed structures were obtained by using the Perdew–Burke–Ernzerhof functional tuned for solids (PBEsol) [48] with Hubbard U corrections [49,50], while the band gaps by using the Heyd–Scuseria–Ernzerhof (HSE06) hybrid functional [51,52] since it substantially improves the band gaps of semiconductors including nitrides and pnictides over local and semilocal functionals [3,8,53]. The employed effective U parameters are given in Table S2 in the Supplemental Materials. The employed effective U parameters are given in Table S2 in the Supplemental Materials.

After structure optimization, we removed 183 substances whose crystal structures are duplicates of the others. The duplication was judged in terms of the assessment of the structure matching implemented in

Figure 1. Overview of our adaptive sampling method for (a) single-objective and (b) multi-objective exploration.

Figure 2. Workflow for the construction of the computational database.
3.2. Target materials and conversion functions for the conventional BO

To evaluate the performance of our method for single target property compared with the conventional BO, we investigate the following three cases: (i) the target materials hold $E_g$ within the range of 1.06-1.51 eV, which yields photovoltaic energy conversion efficiencies higher than 32% at the Shockley-Queisser limit [60,61] and, therefore, is relevant to the search for photoabsorber materials; (ii) the target materials have $E_g$ ranging 2.18-2.50 eV, which are desired to solve the 'green gap problem' in the light emitting diode technology [62]; and (iii) the target materials have $E_f$ not higher than $-1$ eV/atom, as a simple criterion of stability. We assume the two cases as multiple conditions: one is the search for substances that satisfy both conditions (i) and (iii), and the other for those satisfying both (ii) and (iii). The number of target materials in our database for each case is shown in Table 1.

To apply the conventional BO to the above purpose, we used several conversion functions following the previous works [23,24,26]. In the conventional BO, the conversion function is used as $y$ value for GP, and conventional PI and EI are adopted for acquisition functions. In the case of $E_g$, we use 'absolute of difference', 'Gaussian' and 'quadratic', namely,

$$\text{Absolute of difference}(E_g) = - \left| E_g - \frac{E_g\text{lower} + E_g\text{upper}}{2} \right|,$$  

$$\text{Gaussian}(E_g) = \exp \left[ -\frac{\left( E_g - \frac{E_g\text{lower} + E_g\text{upper}}{2} \right)^2}{2[eV^2]} \right],$$  

$$\text{Quadratic}(E_g) = - \left( E_g - \frac{E_g\text{lower} + E_g\text{upper}}{2} \right)^2,$$

where $E_g\text{lower}$ and $E_g\text{upper}$ are the lower and upper limits of the target $E_g$ range, respectively. For $E_f$, we used the 'negative' function, namely, $-E_f$ to conform to the maximization problem. Moreover, we used the 'flatten' conversion function denoted as

$$\text{Flatten}(E_f) = \begin{cases} 0, & E_f < E_f\text{upper} \\ E_f\text{upper} - E_f, & E_f \geq E_f\text{upper} \end{cases}$$

where $E_f\text{upper}$ is the upper limit of the target $E_f$ range. These conversion functions are shown in Figure 4.

### Table 1. Number of target materials in our database.

| $E_f$ range | $1.06 \text{ eV} \leq E_f \leq 1.51 \text{ eV}$ | $2.18 \text{ eV} \leq E_f \leq 2.50 \text{ eV}$ | $E_f$ within the other range | Total |
|------------|-----------------|-----------------|-----------------------------|------|
| $E_f \leq -1 \text{ eV/atom}$ | 32 | 22 | 209 | 263 |
| $E_f > -1 \text{ eV/atom}$ | 371 | 127 | 1257 | 1755 |
| Total | 403 | 149 | 1466 | 2018 |

## 4. Results and discussion

### 4.1. Comparison of single-objective acquisition functions

First, we compared the performance of single-objective PTR and conventional acquisition functions, namely, PI and EI with conversion functions. To conduct material exploration simulations, we selected 10
data points randomly as the initial dataset, and ten substances were identified according to acquisition functions in each step. Figure 5 shows the number of identified target substances against the number of observed substances. As can be seen, the PTR acquisition function can more efficiently identify target materials than the conventional BO in all the three cases.

The improvement by using the PTR is the least noticeable for the exploration of data points holding $E_g$ within the range of 1.06-1.51 eV (Figure 5(a)) and the most in the case of $E_f < -1$ eV/atom (Figure 5(c)). As mentioned above, if the maximum value of acquired data greatly exceeds the threshold, the conventional BO prioritizes data points with large prediction uncertainty rather than data points whose target values are likely to be moderate but satisfy the conditions. Thus, such a decrease in performance by the conventional BO can be partly ascribed to the existence of data points whose target values greatly exceed the threshold, i.e. the width of the target range of the conversion function. Compared with the target ranges of the conversion functions for $E_g$ (Figure 4(a, b)), that of the ‘negative’ conversion function for $E_f$ (Figure 4(c)) is wide for the whole range, and the ‘flatten’ conversion function, whose width of the target range is squashed to zero, improves the performance.

However, even the performance of the ‘flatten’ conversion function is still worse than that of the PTR. This would be because the ‘flatten’ function loses information on how lower the formation energy is than the criterion and surrogate models learn less information. As a similar example, for the optimization of a composite function, exploration methods using the information on the composite structure of the objective function improve the performance compared with the conventional BO using only the target value [63,64]. While we could improve the performance of the conventional BO by refining the conversion function, the PTR acquisition function works well without such a design of the conversion function.

Figure 6 shows the $E_f$ ranges of acquired substances by the exploration simulation with $E_{f \text{ upper}} = -1$ eV/atom. Using the PTR, the target substances of each energy ranges (i.e. $E_f \leq -2$ eV/atom, $-2 \leq E_f \leq 1.75$ eV/atom, $-1.75 \leq E_f \leq 1.5$ eV/atom, $-1.5 \leq E_f \leq 1.25$ eV/atom and $-1.25 \leq E_f \leq -1$ eV/atom) are identified at a similar pace (Figure 6(a)). On the contrary, substances with higher formation energies tend to be more slowly identified by the ‘negative’ conversion function (Figure 6(b)). In this case, the substances having $E_f \leq -2$ eV/atom were found at the early stage [see Figure 6(b)], and then the target substances whose $E_f$ is
close to the threshold (e.g. $-1.5 \leq E_f \leq 1.25$ eV/atom and $-1.25 \leq E_f \leq -1$ eV/atom) were acquired less efficiently. This result confirms that the conventional BO is appropriate for global optimization but not for our motivation. The 'flatten' conversion function shows an intermediate trend between the other strategies (Figure 6(c)).

Regarding substances with $E_f$ higher than $E_{f,\text{upper}} = -1$ eV/atom, the substances with lower $E_f$ tend to be obtained early by all strategies, including the PTR.

Figure 7 shows the performance of exploration with various $E_{f,\text{upper}}$. When $E_{f,\text{upper}}$ is $-0.25$ eV/atom, the performance difference between any strategies and the
random search is relatively small, since most substances meet the criterion and many exploration steps are required to complete the exploration (Figure 7(a)). Setting $E_{f,\text{upper}}$ to $-0.5$, $-0.75$ or $-1$ eV/atom, the improvement of the performance by the PTR is relatively noticeable (Figure 7(b-d)). Lowering $E_{f,\text{upper}}$ to $-1.25$ or $-1.5$ eV/atom, the performance improvement becomes smaller (Figure 7(e, f)) and when $E_{f,\text{upper}}$ is $-1.75$ or $-2$ eV/atom, even some strategies of the conventional BO are more efficient than the PTR (Figure 7(g, h)). This would be because $E_{f,\text{upper}}$ is near the minimum $E_{f}$, the problem of the
discrepancy of $E_{\text{upper}}$ and $y_{\text{max}}$ in the conventional BO is alleviated. Additionally, the conventional BO where the value of $y_{\text{max}}$ is gradually changed should be suitable in such cases. Therefore, the advantage of the PTR is especially great when the threshold is not extreme.

4.2. Performance of multi-objective PTR

In this section, we discuss the performance of exploration for multiple target objectives using the unstandardized and standardized multi-objective PTRs. First, we note that even if we use a single-objective PTR, exploring target materials would be more efficient than the random search. Therefore, we also show the number of target materials identified by the single-objective PTR as a baseline. The results of these exploration simulations are shown in Figure 8. There is no significant difference between the performance of the unstandardized and standardized multi-objective PTRs. This would be because (1) since the number of our target values is only two, the acquisition function does not become extremely small even if we do not use logarithm values and (2) after a few target substances are acquired, the min-max normalization does not change the value greatly, since the minimum and maximum values of the single-objective PTRs are nearly 0 and 1 as $\sigma_t(x_i)$ is generally small for the known data set.

At the early stage, both multi-objective methods improve efficiency. The performance of both unstandardized and standardized multi-objective PTRs is worse than that of the single-objective PTR of only $E_t$ in the case of $1.06 \text{ eV} \leq E_g \leq 1.51 \text{ eV}$ and $E_t \leq -1 \text{ eV/atom}$ after the most of target data points are identified, while that for $2.18 \text{ eV} \leq E_g \leq 2.50 \text{ eV}$ and $E_t \leq -1 \text{ eV/atom}$ is still better at the last stage than both single-objective PTR of $E_g$ and $E_t$. The deterioration by using $E_g$ in the former case can be ascribed to the following two reasons. (1) The accurate prediction of $E_g$ would be more difficult than that of $E_t$. Actually, when performing ten-fold cross-validation for the whole data by pruning descriptors and constructing a GP regression model that is the same as the exploration simulation, the coefficients of determination ($R^2$) of $E_g$ and $E_t$ are 0.65 and 0.81, respectively. (2) The ratio of the rest target substances versus the rest of substances satisfying the $E_g$ condition is small at the late stage. This means that the most of substances satisfying $E_g$ condition are not targeted due to $E_t$ and thereby the adaptive sampling focusing on $E_g$ would not improve or even impede the exploration performance. As shown in Figure 9(a), such a ratio of $E_g$ becomes extremely small at the late stage of the exploration compared with that of $E_t$ in the case of $1.06 \text{ eV} \leq E_g \leq 1.51 \text{ eV}$ and $E_t \leq -1 \text{ eV/atom}$. On the other hand, for $2.18 \text{ eV} \leq E_g \leq 2.50 \text{ eV}$ and $E_t \leq -1 \text{ eV/atom}$, the ratio of $E_g$ and $E_t$ Figure 9(b) is about the same at the late stage. These problems could be remedied if (1) the prediction accuracy of $E_g$ can be improved by developing better descriptors and/or increasing the initial dataset size and (2) weighting each single-objective PTR based on the estimated ratio of target substances in the rest of the data.

5. Conclusion

We have developed an exploration method for substances whose target properties meet specific criteria, based on the PTR acquisition function. First, we have compared the performance of the single-objective PTR and conventional BO by conducting exploration simulations using the in-house computational database and found that the PTR identified target materials more efficiently when the criterion of the target property is not extreme. Furthermore, we conducted the multi-objective exploration simulations by using the unstandardized and standardized multi-objective PTRs. The performance is improved by both the multi-objective PTRs unless the ratio of the target substances to the substances satisfying one of the given conditions is extremely small.

Acknowledgements

The computing resources of the Academic Center for Computing and Media Studies at Kyoto University and the Research Institute for Information Technology at Kyushu University were used.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by KAKENHI Grant Numbers JP 19K15275, JP 21K14401, and JP 20H00302 from the Japan Society for the Promotion of Science, CREST Grant Number [PMJCR17J2] from the Japan Science and Technology Agency, and Design & Engineering by Joint Inverse Innovation for Materials Architecture, the Ministry of Education, Culture, Sports, Science, and Technology.

ORCID

Akira Takahashi https://orcid.org/0000-0002-3159-9007
Yu Kumagai https://orcid.org/0000-0003-0489-8148
Ryo Tamura https://orcid.org/0000-0002-0349-358X
Fumiyasu Oba https://orcid.org/0000-0001-7178-5333
References

[1] Curtarolo S, Hart G, Nardelli M, et al. The high-throughput highway to computational materials design. Nat Mater. 2013;12:191.

[2] Hautier G, Miglio A, Ceder G, et al. Identification and design principles of low hole effective mass p-type transparent conducting oxides. Nat Commun. 2013;4:2292.

[3] Hinuma Y, Hatakeyama T, Kumagai Y, et al. Discovery of earth-abundant nitride semiconductors by computational screening and high-pressure synthesis. Nat Commun. 2016;7:11962.

[4] Oba F, Kumagai Y. Design and exploration of semiconductors from first principles: a review of recent advances. Appl Phys Express. 2018;11:060101.

[5] Jain A, Ong SP, Hautier G, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater. 2013;1:011002.

[6] Curtarolo S, Setyawan W, Wang S, et al. AFLOWLIB: ORG: a distributed materials properties repository from high-throughput ab initio calculations. Comput Mater Sci. 2012;58:227–235.

[7] Kirklin S, Saal J, Meredig B, et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. NPJ Comput Mater. 2015;1:15010.

[8] Heyd J, Perdew JE, Scuseria GE. Energy band gaps and lattice parameters evaluated with the Heyd-Scuseria-Ernzerhof screened hybrid functional. J Chem Phys. 2005;123:174101.

[9] Hedin L. New method for calculating the one-particle green’s function with application to the electron-gas problem. Phys Rev. 1965;139:A796.

[10] Curtarolo S, Morgan D, Persson K, et al. Predicting crystal structures with data mining of quantum calculations. Phys Rev Lett. 2003;91:135503.

[11] Balachandran PV, Broderick SR, Rajan K. Identifying the ‘inorganic gene’ for high-temperature piezoelectric perovskites through statistical learning. Proc R Soc A. 2011;467:2271–2290.

[12] Seko A, Maekawa T, Tsuda K, et al. Machine learning with systematic density-functional theory calculations: application to melting temperatures of single- and binary-component solids. Phys Rev B. 2014;89:054303.

[13] Meredig B, Agrawal A, Kirshin S, et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys Rev B. 2014;89:094104.

[14] Ghiringhelli LM, Vybiral J, Levchenko SV, et al. Big data of materials science: critical role of the descriptor. Phys Rev Lett. 2015;114:105503.

[15] Ward L, Agrawal A, Choudhary A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials. NPJ Comput Mater. 2016;2:16028.

[16] Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett. 2018;120:145301.

[17] Bartel CJ, Millican SL, Deml AM, et al. Physical descriptor for the Gibbs energy of inorganic crystal-line solids and temperature-dependent materials chemistry. Nat Commun. 2018;9:4168.

[18] Lookman T, Balachandran PV, Xue D, et al. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. NPJ Comput Mater. 2019;5:21.

[19] Terayama K, Sumita M, Tamura R, et al. Black-box optimization for automated discovery. Acc Chem Res. 2021;54:1334–1346.

[20] Seko A, Togo A, Hayashi H, et al. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization. Phys Rev Lett. 2015;115:205901.

[21] Bassman L, Rajak P, Kalia RK, et al. Active learning for accelerated design of layered materials. Npj Comput Mater. 2018;4:74.

[22] Fukazawa T, Harashima Y, Hou Z, et al. Bayesian optimization of chemical composition: a comprehensive framework and its application to RFe$_2$-type magnet compounds. Phys Rev Mater. 2019;3:053807.

[23] Terayama K, Tsuda K, Tamura R. Efficient recommendation tool of materials by an executable file based on machine learning. Jpn J Appl Phys. 2019;58:098001.

[24] Chen X, Wang C, Li Z, et al. Bayesian optimization based on a unified figure of merit for accelerated materials screening: a case study of halide perovskites. Sci China Mater. 2020;63:1024–1035.

[25] Xin R, Siriwardane EMD, Song Y, et al. Active-learning-based generative design for the discovery of wide-band-gap materials. J Phys Chem C. 2021;125:16118–16128.

[26] Ma XY, Lyu HY, Hao KR, et al. High-efficient ab initio Bayesian active learning method and applications in prediction of two-dimensional functional materials. Nanoscale. 2021;13:14694–14704.

[27] Motoyama Y, Tamura R, Yoshimi K, et al. Bayesian optimization package: PHYSBO. 2021; arXiv:2110.07900, arXiv.org e-Print archive. Cited 2022 Jan 18. https://arxiv.org/abs/2110.07900

[28] Solomou A, Zhao G, Boluski S, et al. Multi-objective Bayesian materials discovery: application on the discovery of precipitation strengthened NiTi shape memory alloys through micromechanical modelling. Mater Des. 2018;160:810–827.

[29] Gopakumar AM, Balachandran PV, Xue D, et al. Multi-objective optimization for materials discovery via adaptive design. Sci Rep. 2018;8:3738.

[30] Suzuki S, Takeno S, and Tamura T, et al. Multi-objective Bayesian optimization using Pareto-frontier entropy. Proceedings of the 37th International Conference on Machine Learning, PMLR (virtual). 2020;119:9279–9288. Available from: http://proceedings.mlr.press/v119/

[31] Kishio T, Kaneko H, Funatsu K. Strategic parameter search method based on prediction errors and data density for efficient product design. Chemom Intell Lab Syst. 2013;127:70–79.

[32] Nakao A, Kaneko H, Funatsu K. Development of an adaptive experimental design method based on probability of achieving a target range through parallel experiments. Ind Eng Chem Res. 2016;55:5726–5735.

[33] Tsukada Y, Takeno S, Karasuyama M, et al. Estimation of material parameters based on precipitation shape: efficient identification of low-error region with Gaussian process modelling. Sci Rep. 2019;9:15794.

[34] Iwama R, Kaneko H. Design of ethylene oxide production process based on adaptive design of experiments and Bayesian optimization. J Adv Manuf Process. 2021;3:e10085.
[35] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2017; arXiv:1412.6980. arXiv.org e-Print archive. Cited 2022 Jan 18. https://arxiv.org/abs/1412.6980

[36] Alvarez MA, Rosasco L, Lawrence ND, et al. Kernels for vector-valued functions: a review. Found Trends Mach Learn. 2012;4(3):195–266.

[37] Ward L, Dunn A, Faghaninia A, et al. Matminer: an open source toolkit for materials data mining. Comput Mater Sci. 2018;152:60–69.

[38] Takahashi A, Kumagai Y, Miyamoto J, et al. Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations. Phys Rev Mater. 2020;4:103801.

[39] Seko A, Hayashi H, Nakayama K, et al. Representation of compounds for machine-learning prediction of physical properties. Phys Rev B. 2017;95:144110.

[40] Ward L, Liu R, Krishna A, et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. Phys Rev B. 2017;96:024104.

[41] Noda Y, Otake M, Nakayama M. Descriptors for dielectric constants of perovskite-type oxides by materials informatics with first-principles density functional theory. Sci Technol Adv Mater. 2020;21(1):92–99.

[42] Breiman L. Random forests. Mach Learn. 2001;45:5–32.

[43] Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Statist Sci. 2001;16(3):199–231.

[44] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine Learning in Python. J Mach Learn Res. 2011;12:2825–2830.

[45] Böähr PE. Projector augmented-wave method. Phys Rev B. 1994;50:17953.

[46] Kresse G, Furthmüller J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Phys Rev B. 1996;54:11169.

[47] Kresse G, Joubert D. From ultrasoft pseudopotentials to the projector augmented-wave method. Phys Rev B. 1999;59:1758.

[48] Perdew JP, Ruzsinszky A, Csonka GI, et al. Restoring the density-gradient expansion for exchange in solids and surfaces. Phys Rev Lett. 2008;100:136406.

[49] Anisimov Z VI, Andersen J. Band theory and Mott insulators: Hubbard U instead of Stoner I. Phys Rev B. 1991;44:943.

[50] Dudarev SL, Bottom GA, Savrasov SY, et al. Electron-energy-loss spectra and the structural stability of nickel oxide: an LSDA+U study. Phys Rev B. 1998;57:1505.

[51] Heyd J, Scuseria GE, Ernzerhof M. Hybrid functionals based on a screened Coulomb potential. J Chem Phys. 2003;118:8207.

[52] Krukau AV, Vydrov OA, Izmaylov AF, et al. Influence of the exchange screening parameter on the performance of screened hybrid functionals. J Chem Phys. 2006;125:224106.

[53] Hinuma Y, Grüneis A, Kresse G, et al. Band alignment of semiconductors from density-functional theory and many-body perturbation theory. Phys Rev B. 2014;90:155405.

[54] Togo A, Tanaka I. Spglib: a software library for crystal symmetry search. 2018, arXiv:1808.01590. arXiv.org e-Print archive. Cited 2022 Jan 18. https://arxiv.org/abs/1808.01590

[55] Hinuma Y, Pizzi G, Kumagai Y, et al. Band structure diagram paths based on crystallography. Comput Mater Sci. 2017;128:140–184.

[56] Kumagai Y, Tsunoda N, Takahashi A, et al. Insights into oxygen vacancies from high-throughput first-principles calculations. Phys Rev Mater. 2021;5:123803.

[57] Ong SP, Richards WD, Jain A, et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci. 2013;68:314–319.

[58] Jain A, Ong SP, Chen W, et al. FireWorks: a dynamic workflow system designed for high-throughput applications. Concurr Comput Pract Exp. 2015;27:5037–5059.

[59] Mathew K, Montoya JH, Faghaninia A, et al. Atoma: a high-level interface to generate, execute, and analyze computational materials science workflows. Comput Mater Sci. 2017;139:140–152.

[60] Shockley W, Queisser HJ. Detailed balance limit of efficiency of p-n junction solar cells. J Appl Phys. 1961;32:510–519.

[61] Rühle S. Tabulated values of the Shockley–Queisser limit for single junction solar cells. Sol Energy. 2016;130:139–147.

[62] de Maur MA, Pecchia A, Penazzi G, et al. Efficiency drop in green InGaN/GaN light emitting diodes: the role of random alloy fluctuations. Phys Rev Lett. 2016;116:027401.

[63] Astudillo R, and Frazier P. Bayesian optimization of composite functions. Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, CA. 2019;97:354–363.

[64] Maddox WJ, Baladat M, Wilson AG, et al. Bayesian optimization with high-dimensional outputs. 2021; arXiv:2106.12997. arXiv.org e-Print archive. Cited 2022 Jan 18. https://arxiv.org/abs/2106.12997