Supplementary Information

Sparse Modeling based Bayesian Optimization for Experimental Design

Ryuji Masui[1*], Unseo Lee[1], Ryo Nakayama[2**], Taro Hitosugi[2]

[1]*HACARUS Inc, Kyoto, 604-0835, Japan*

[2]*Department of Chemistry, School of Science, The University of Tokyo, Tokyo, 113-0033, Japan*

* ryuji@hacarus.com

** ryo-nakayama@g.ecc.u-tokyo.ac.jp

[Bayesian optimization]

When using a positive integer $M > 0$ to define an N-dimensional search space as $\mathcal{X} \in [0, M]^N$, the objective involves find $x^* = \mathrm{argmax}_{\mathrm{x}\in\mathcal{X}} f(x)$, where the objective function $f: \mathcal{X} \mapsto \mathbb{R}$ takes the maximum value within the search space. The objective function assumes a black-box nature, where obtaining function values for any $x \in \mathcal{X}$ is possible; however, the evaluation cost is high. In addition, computing the gradients of this objective function is challenging and not necessarily convex.

For such an objective function, f, Bayesian optimization (BO) [1], [2] has been proposed as a technique to determine the optimal solution with minimal trial iterations. BO is characterized by a surrogate model based on Gaussian processes (GPs) and an acquisition function. Gaussian process regression is a nonparametric method that can express nonlinearity and represent estimation uncertainties. A GP is a distribution of functions described by the mean $m(\cdot)$ and covariance $k(\cdot,\cdot)$. It satisfies the following conditions when considering a set of n data points, $x_{1:n}, (x_i \in \mathcal{X})$,

$$f(x_{1:n}) \sim \mathcal{N}\big(m(x_{1:n}), K(x_{1:n}, x_{1:n})\big) \tag{S1}$$

where $K(x_{1:n}, x_{1:n})$ represents the variance-covariance matrix, denoted by a kernel function k such that $K(x_{1:n}, x_{1:n})_{i,j} = k(x_i, x_j)$. The selection of the kernel function k is arbitrary and should be chosen appropriately based on prior information regarding the smoothness of the function, among other factors. Commonly chosen kernel functions include the radial basis function (RBF) and Matern kernels.

Furthermore, for particular data points $x_i \in \mathcal{X}$ and corresponding objective function values $y_i = f(x_i)$, the joint distribution at any arbitrary point $x' \in \mathcal{X}$ can be expressed as follows:

$$\begin{bmatrix} y_{1:n} \\ x' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_{1:n}) \\ m' \end{bmatrix}, \begin{bmatrix} K(x_{1:n}, x_{1:n}) & k(x_{1:n}, x') \\ k(x', x_{1:n}) & k(x', x'] \end{bmatrix}\right)$$

For simplicity, we assume $m(x_{1:n}) = 0$ and $m' = 0$. When given the dataset $(x_{1:n}, y_{1:n})$, the mean μ and variance σ^2 at any arbitrary point $x' \in \mathcal{X}$ can be expressed as follows:

$$\mu(x' \mid D_t) = k(x', x_{1:n}) K(x_{1:n}, x_{1:n})^{-1} y_{1:n}$$
$$\sigma^2(x' \mid D_t) = k(x', x') - k(x', x_{1:n}) K(x_{1:n}, x_{1:n})^{-1} k(x_{1:n}, x')$$

As described above, based on the obtained data, the expected value of prediction $\mu(x')$ and the uncertainty of the prediction $\sigma^2(x')$ for any point x' can be quantified.

Algorithm S1. Bayesian Optimization

Require: An objective function f, a total evaluation budget T, an initial dataset $D_0 = \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i = f(x_i), i = 1, \dots, n\}$.

Ensure: Approximate maximum $x^* = \arg\max_{x \in \mathcal{X}} f(x)$

1: Construct a GP model $\hat{f}_0$ with D_0.

2: f**or** $t = 1, 2, \dots, T$ **do**

3: Find x_t by optimizing the acquisition function q: $x_t = \arg\max_{x \in \mathcal{X}} q\left(x \mid \hat{f}_{t-1}\right)$.

4: Augment the data $D_t = D_{t-1} \cup \{(x_t, f(x_t))\}$.

5: Reconstruct the GP model $\hat{f}_t$ by updating the kernel hyper-parameters with D_t.

6: **end for**

7: **return** the maximum data point x^* in D_t.

In BO, the next set of experimental points is determined to maximize the acquisition function q based on the mean μ and variance σ^2 obtained from this GP. Subsequently, new data points are incorporated to re-estimate the mean and variance, and the process is iterated to determine the next set of experimental points. The choice of the acquisition function in this process is arbitrary, and several are well-known, such as the expected improvement (EI) and upper confidence bound (UCB) [2], [3], [4]. In this study, the EI was used as an acquisition function. The EI function quantifies the improvement expected at point x compared to the maximum objective function value from past data points. Considering the dataset obtained until the t-th experiment $D_t = \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i = f(x_i\}, i = 1, \dots, n_t\}$, and the GP model $\hat{f}_t$ derived from these, the EI function is defined as follows:

$$q(x \mid D_t) = \mathbb{E}\left(\max\left\{0, \hat{f}_t(x) - f(x^+)\right\} \mid D_t\right)$$

where $f(x^+) = \max_{x \in x_{1:n}} f(x)$ represents the maximum objective function value obtained in the past n experiments. The following experimental points are selected to maximize the acquisition function:

$$x_{t+1} = \arg\max_{x \in \mathcal{X}} q(x \mid D_t)$$

The dataset is updated by evaluating the objective function at the selected data points. By repeating this process, $x \in \mathcal{X}$ that maximizes the objective function f can be obtained with a small number of trials. Theoretically, it is possible to attain the optimal solution by conducting a sufficient number of trials. However, owing to the high cost associated with evaluating the objective function, achieving a near-optimal solution within a

minimal number of trials is desirable. Considering the maximum limit of T trials, Algorithm S1 illustrates the BO algorithm.

[Bayesian optimization based on sparse estimation using an ARD kernel]

In BO, as the dimensionality of the search space increases, the number of iterations required to determine the global optimum also increases. Not all explanatory variables are equally important in high-dimensional search spaces. For instance, considering a search space $\mathcal{X}$, where a subspace $\mathcal{X}^\top$ is a subset of $\mathcal{X}$, and its orthogonal complement is denoted as $\mathcal{X}^\perp$ $(i.e. \mathcal{X} = \mathcal{X}^\perp \oplus \mathcal{X}^\perp)$, it is assumed that there exist two functions $f_d: \mathcal{X}^\top \mapsto \mathbb{R}$ and $f_s: \mathcal{X}^\perp \mapsto \mathbb{R}$ along with a sufficiently small constant $\epsilon > 0$, such that:

$$f(x) = f_d(x^\top) + \epsilon f_s(x^\perp) \text{ s.t. } x^\top \in \mathcal{X}^\top, x^\perp \in \mathcal{X}^\perp$$

where the objective function f exhibits only a negligible dependence on $x^\perp$. Therefore, iterating the optimization procedure, while focusing on $\mathcal{X}^\top$, can yield a better approximate solution with fewer iterations as $\dim(\mathcal{X})$ becomes smaller. Consequently, the kernel of the GP regression must express that each component of the search space has a different impact on the objective function.

When the objective function is smooth, the RBF kernel is often chosen as the kernel for the GP regression. The RBF kernel is defined as follows for any $x, x' \in \mathcal{X}$;

$$k_{\mathrm{RBF}}(x, x' \mid \sigma_{\mathrm{RBF}}, \ell) = \sigma_{\mathrm{RBF}}^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$

The RBF kernel has two hyperparameters, σ_{RBF}^2 and ℓ, which denote the covariance and lengthscale, respectively. In the GP regression, it is possible to estimate the most probable hyperparameters from the data. However, when choosing the RBF kernel, the hyperparameters act on the L^2 norm within search spaces of x and x'. Therefore, it is impossible to evaluate the importance of each component of the search space separately.

Applying BO to an RBF kernel may result in experiments that predominantly alter unimportant explanatory variables when they are present. When the objective function has multiple local solutions, changing the unimportant explanatory variables is unlikely to yield an optimal solution because it essentially remains at the same local solution. Therefore, it is crucial to quantify the importance of each explanatory variable to obtain superior approximate solutions with fewer trial iterations. Focusing solely on important explanatory variables can help mitigate unnecessary trials and lead to more efficient optimization.

An automatic relevance-determination (ARD) kernel is proposed as a method to quantify the importance of each explanatory variable [5].

$$k_{\mathrm{ARD}}(x, x' \mid \sigma_{\mathrm{ARD}}, \ell_{1:N}) = \sigma_{\mathrm{ARD}}^2 \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\frac{(x_i - x_i')^2}{\ell_i^2}\right)$$

The ARD kernel encompasses hyperparameters such as the covariance σ_{ARD}^2 and the lengthscale for each component of the search space, denoted as $\ell_{1:N}$ $(i = 1, \ldots, N)$. The relevance of each component to the output is expressed by ℓ_i, where a larger value of ℓ_i indicates a smaller influence of the i-th component on the objective function. Hence, estimating this hyperparameter to fit the data facilitates the quantification of the degree of influence of each explanatory variable on the objective function.

Moreover, judging whether the estimated ℓ_i exceeds the threshold value allows us to separate the components into high- and low-influence components. Focusing only on important explanatory variables with high influence enables the determination of a better approximate solution with fewer trials. Values that maximize EI are utilized as important explanatory variables. By contrast, for unimportant explanatory variables, the values are determined through random sampling from the search space to reduce uncertainty in that direction, thereby enhancing the accuracy of the relevance estimation. Algorithm S2 presents the BO algorithm based on sparse estimation using the ARD kernel.

Algorithm S2. Sparse Bayesian Optimization

Require: An objective function f, a total evaluation budget T, an initial dataset $D_0 = \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i = f(x_i), i = 1, \ldots, n\}$, a threshold ϵ_ℓ.

Ensure: Approximate maximum $x^* = \arg\max_{x \in \mathcal{X}} f(x)$

1: Construct a ARD kernel based GP model $\hat{f}_0$ with D_0.

2: f**or** $t = 1, 2, \ldots, T$ **do**

3: Let ℓ_i be the lengthscale of ARD kernel for each element (each synthesis parameter) i.

4: Partition $(\mathcal{X}^\top, \mathcal{X}^\perp)$ by thresholding $(\ell_i < \epsilon_\ell)$
$\mathcal{X}^\top$: dense subspace, $\mathcal{X}^\perp$: sparse subspace, $\mathcal{X} = \mathcal{X}^\top \oplus \mathcal{X}^\perp$

5: Find $x_t^\top$ by optimizing the acquisition function q: $x_t^\top = \arg\max_{x \in \mathcal{X}^\top} q\left(x \mid \hat{f}_{t-1}\right)$.

6: Choose $x_t^\perp$ by random sampling in $\mathcal{X}^\perp$.

7: $x_t = x_t^\top + x_t^\perp$

8: Augment the data $D_t = D_{t-1} \cup \{(x_t, f(x_t))\}$.

9: Reconstruct the GP model $\hat{f}_t$ by updating the kernel hyper-parameters with D_t.

10: **end for**

11: **return** the maximum data point x^* in D_t.

[Quantifying the importance of explanatory variables using a partial dependence plot]

For a black-box multivariate function $f: \mathbb{R}^p \mapsto \mathbb{R}$, Friedman's partial dependence plot (PDP) is a method to quantify the average change in function values when altering the i-th component [6]. Given a set $S \subset \{1, \dots, p\}$ and its complement denoted as $C = \bar{S}$, the partial dependence function f_S is defined as the mean value when x_C is varied over the neighborhood distribution $dP(x_C)$ with the components of S fixed at x_s, and expressed as:

$$f_S = \mathbb{E}_{x_C}[f(x_S, x_C)] = \int f(x_S, x_C) dP(x_C)$$

However, assuming that evaluating the objective function f is expensive, instead of directly computing f_S, the PDP is estimated using a surrogate model $\hat{f}$ generated by the GP. This estimation is based on the data points x_C^i $(i = 1, \dots, n)$ used to construct the surrogate model. The estimation is as follows:

$$\hat{f}_S = \frac{1}{n} \sum_{i=0}^{n} \hat{f}(x_S, x_C^i)$$

For simplicity, we assume that the cardinality of set S is 1. Consequently, $\hat{f}_S$ can be regarded as a univariate function. By varying the components under consideration, it is possible to understand how the objective function changes, on average. Hence, by calculating the difference between the minimum and maximum values of $\hat{f}_S$ when the components of S, the average effect $\hat{e}_S$ on the objective function can be quantified as follows:

$$\hat{e}_S = \max_{x_S} \hat{f}_S(x_S) - \min_{x_S} \hat{f}_S(x_S)$$

where $\hat{e}_S$ is the average partial dependence effect (APDE).

When the objective function lacks interactions, explanatory variables can be divided into important and unimportant components based on the APDE threshold. The threshold corresponds to the extent of change in the objective function when a specific component is varied, thus facilitating an intuitive determination of this parameter. However, determining the important explanatory variables based solely on the APDE values becomes challenging when the objective function involves interactions. For example, when the objective function value is large only for a particular subregion x_C and small for other x_C, the APDE for x_C is underestimated because of the effect of averaging.

Therefore, when the objective functions interact, a small APDE does not indicate that x_C is an important component.

[Quantifying the importance of explanatory variables using individual conditional expectations]

When the objective function involves interactions, relying solely on the average effect obtained through PDP may result in a misunderstanding. Focusing on individual conditional expectations (ICE) is recommended in such cases, as is well known in the literature [7]. ICE is denoted as the function $\hat{f}(x_S, x_C^i)$ when the i-th instance's component x_C^i, belonging to $C \subset \{1, \dots, p\}$ are fixed. Let x_C^i $(i = 1, \dots, n)$ represent the experimental data obtained in the nth trial. Further, ICE indicates the effect of properties and yields for x_C^i by changing the specific synthesis parameters while the remaining parameters are constant. The average ICE computed for all instances corresponds to the estimated value of PDP $\hat{f}_S$. For each $\hat{f}(x_S, x_C^i)$, the effect of the component S (denoted as $\hat{e}_S^i$) can be computed as follows:

$$\hat{e}_S^i = \max_{x_S} \hat{f}(x_S, x_C^i) - \min_{x_S} \hat{f}(x_S, x_C^i)$$

Furthermore, the effect on the component S can be expressed as:

$$\hat{e}_S^* = \max_{i \in [1,n]} \hat{e}_S^i$$

The effects of the S components on the objective function with variation in the components of S can be quantified. From another perspective, $\hat{e}_S^*$ corresponds to the approximation of $\max_{x_C} \left\{ \max_{x_S} f(x_S, x_C) - \min_{x_S} f(x_S, x_C) \right\}$. Thus, $\hat{e}_S^*$ represents the maximum change in the objective function value when only the component S is changed. Therefore, $\hat{e}_S^*$ is denoted as the maximum partial dependence effect (MPDE).

As an example, Fig. S1 illustrates the PDP and ICE for x_1 for two functions: (a) without interactions, $y_1 = \sin(x_1) + \cos(x_2)$, and (c) with interactions, $y_2 = \sin(x_1)\cos(x_2)$. Figure S1(a) shows the behavior of y_1 within the range of $[0, 2\pi]$ for both x_1 and x_2, and Fig. S1(b) shows the PDP (red solid line) and ICE (black solid line) of y_1. Among the ICE curves, the blue solid line represents max-ICE, which corresponds to the maximum difference between the maximum and minimum values of ICE within the range of x_1. This difference (the gap in the blue dashed line) represents the $\hat{e}_S^*$ (MPDE). In the function without interactions, all the ICE curves exhibit the same magnitude of change. Consequently, the difference between the maximum and minimum PDP (APDE) values

matched that of MPDE. Therefore, the effect on the properties and yields could be quantified using the APDE obtained at a low calculation cost. Similarly, Fig. S1(c) displays the behavior of y_2 within the range of $[0, 2\pi]$ for both x_1 and x_2, and Fig. S1(d) illustrates the PDP (red solid line) and ICE (black solid line) for y_2. Because y_2 involves interactions, the APDE may be very small because of the effects of averaging. In contrast, the MPDE indicates that x_1 has a significant impact on y_2.

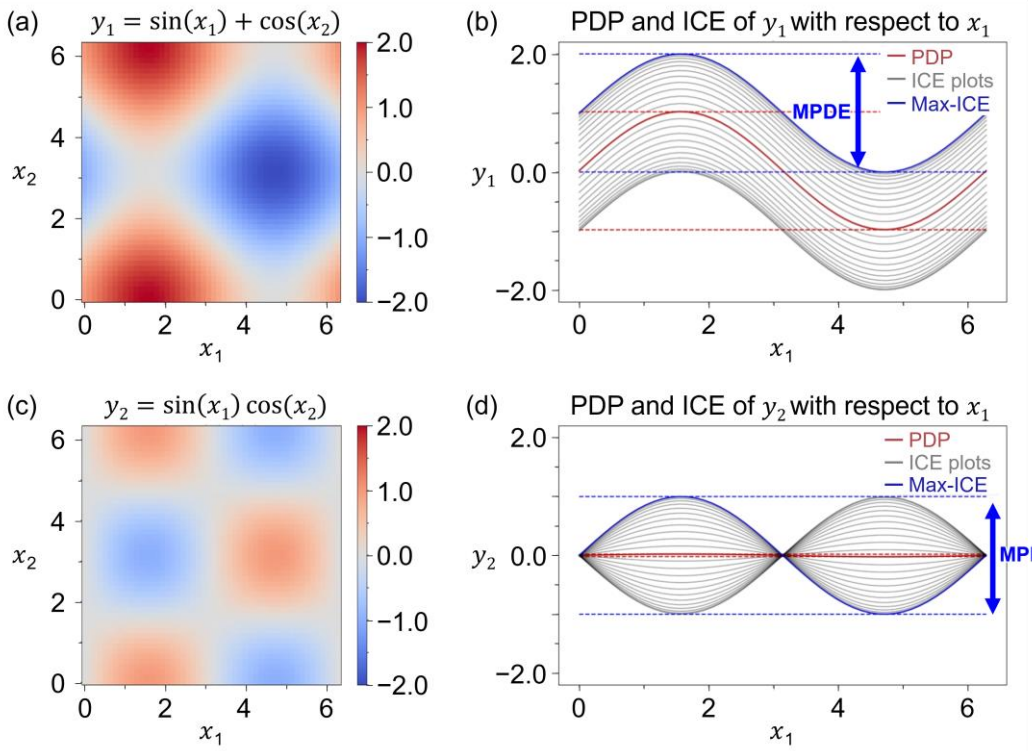

Fig. S1. (a) Function without interaction $y_1 = \sin(x_1) + \cos(x_2)$, and (b) partial dependence plot (PDP) and individual conditional expectation (ICE) for y_1 with respect to x_1. (c) Function with interaction $y_2 = \sin(x_1)\cos(x_2)$, and (d) PDP and ICE for y_2 with respect to x_1. The red, gray, and blue lines represent PDP, ICE, and Max-ICE, which corresponds to the maximum difference between the maximum and minimum values of ICE within the range of x_1.

[Modeling of materials synthesis using isotropic Gaussian functions.

Materials synthesis encompasses several fields, including the synthesis of inorganic and organic compounds. Furthermore, the forms of the materials can vary from powders (bulk) to nanoparticles, thin films, and composites of different materials. Synthesis parameters, such as temperature, pressure, and composition, change the material properties. When a change in synthesis parameters causes a phase transition, materials properties can discontinuously change in response to variations in synthesis parameters.

In contrast, the material properties can continuously change with variations in the synthesis parameters, and the material properties exhibits their maximum/minimum values at specific parameter values. [8], [9], [10], [11], [12], [13], [14] In the case of crystalline materials, a moderately elevated synthesis temperature can improve crystallinity, whereas extremely high temperatures can reduce crystallinity owing to thermal decomposition. [10] In the synthesis of transparent conductive films such as indium tin oxide (ITO), increasing the oxygen partial pressure can enhance the crystallinity and improve the carrier mobility. However, it also reduces the number of oxygen vacancies, resulting in decreased carrier density. [11] Consequently, there is a trade-off between the carrier mobility and carrier density, and the electron conductivity reaches its maximum at a specific oxygen partial pressure. In the synthesis of organic hole-transport materials for perovskite solar cells, the hole mobility changes with the annealing time and dopant concentration, exhibiting peaks in the two-dimensional exploration space, including global and local maximum peaks. [8]

We have already performed simulations of material synthesis with one-, two-, and three-dimensional synthesis parameters, assuming that the global and local optimal peaks are isotropic.[15],[16] We modeled materials synthesis by functions with local optimal peaks generated by combining isotropic Gaussian functions. These studies simulated thin-film synthesis by sputtering with different process window (P_w) widths, where the synthesis parameters are synthesis temperature, oxygen partial pressure, and sputtering power, which are *important* parameters influencing material properties. The process window (P_w) represents the range of synthesis conditions that provide desired material properties. When P_w is large, determining the optimal synthesis conditions is easy. However, in case of small P_w, optimization becomes challenging. In our previous research, based on literature values, we set P_w for temperature, oxygen partial pressure, and sputtering power at 100 °C, 1.0×10^{-4} Pa, and 10 W, respectively. [10], [14], [17]

References

[1] Rasmussen, C. E., & Williams, K. I. Gaussian process for machine learning (MIT Press, 2006).

[2] Jones, D. R., Schonlau, M., & Welch, W. J. Efficient global optimization of expensive black-box functions. *J. Glob. Optim*. **13**, 455-492 (1998).

[3] Brochu, E., Cora, V. M., & de Freitas N. A. Tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599 (2010).

[4] De Palma, A., Mendler-Dünner, C., Parnell T, Anghell, A., & Pozidis H. Sampling acquisition functions for batch Bayesian optimization. arXiv:1903.09434 (2019).

[5] Chen, B., Castro, R., & Krause. A. Joint optimization and variable selection of highdimensional gaussian processes. arXiv:1206.6396 (2012).

[6] Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat*. **29**, 1189-1232 (2001).

[7] Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* **24**, 44-65 (2015).

[8] MacLeod, B. P., *et al*. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci Adv.* **6**, eaaz88 (2020).

[9] Kusada, K., *et al*. Solid solution alloy nanoparticles of immiscible Pd and Ru elements neighboring on Rh: Changeover of the thermodynamic behavior for hydrogen storage and enhanced CO-oxidizing ability. *J. Am. Chem. Soc.* **136**, 1864-1871 (2014).

[10] Tsuruhama, T., Hitosugi, T., Oki, H., Hirose T., & Hasegawa, T. Preparation of layered-rhombohedral $LiCoO_2$ epitaxial thin films using pulsed laser deposition. *Appl. Phys. Express*. **2**, 085502 (2009).

[11] Kim, H., *et. al*. Electrical, optical, and structural properties of indium–tin–oxide thin films for organic light-emitting devices. *J. Appl. Phys*. **86**, 6451 (1999).

[12] Bradley, K., Giagloglou, K., Hayden, B. E., Jungias, H., & Vian C. Reversible perovskite electrocatalysts for oxygen reduction/oxygen evolution. *Chem. Sci.* **10,** 4609-4617 (2019).

[13] Chargui, A., *et al*. Influence of thickness and sputtering pressure on electrical resistivity and elastic wave propagation in oriented columnar tungsten thin films. *Nanomaterials*. **10**, 81 (2020).

[14] Chaoumead, A., Sung, Y., & Kwak, D. The effects of RF sputtering power and gas pressure on structural and electrical properties of ITiO thin film. *Adv. Condens. Matter Phys.* **2012**, 651587 (2012).

[15] Nakayama, R., *et al*. Tuning of Bayesian optimization for materials synthesis: simulation of the one-dimensional case. *STAM Methods*. **2**, 119 (2022).

[16] Xu, H., *et al*. Tuning Bayesian optimization for materials synthesis: simulating two- and three-dimensional cases. *STAM Methods*. **3**, 2210251 (2023).

[17] Shimizu, R., Kobayashi, S., Watanabe, Y., Ando, Y., & Hitosugi, T. Autonomous materials synthesis by machine learning and robots. *APL Mater*. **8**, 1111 (2020).