Experimental design for the highly accurate prediction of material properties using descriptors obtained by measurement

Ryo Tamura¹,²,³, Yuki Takei¹,², Shinichiro Imai¹,², Maki Nakahara¹,², Satoshi Shibata²,³, Takashi Nakanishi¹,² and Masahiko Demura¹,²,³

¹Materials Open Platform for Chemistry, National Institute for Materials Science (NIMS), Ibaraki, Japan; ²International Center for Materials Nanoarchitectonics (WPI-MANA), National Institute for Materials Science (NIMS), Ibaraki, Japan; ³Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science (NIMS), Ibaraki, Japan; ⁴Informatics Initiative, Digital Value Co-Creation, Asahi KASEI Corporation, Tokyo, Japan; ⁵Science & Innovation Center, Mitsubishi Chemical Corporation, Kanagawa, Japan; ⁶Mitsui Chemicals, Inc, Chiba, Japan; ⁷Advanced Materials Development Laboratory, Sumitomo Chemical Co, Chiba, Japan

ABSTRACT
In materials science, both controllable and uncontrollable descriptors can be used to characterize materials. Examples of controllable descriptors include the composition of elements and fabrication processes; in contrast, uncontrollable descriptors are generated by experimental data characterizing particular samples, such as raw spectral data or specific gravity. In this study, we consider an experimental design to obtain a highly accurate prediction model where the uncontrollable descriptors of materials are features and its material properties are labels. In general, as uncontrollable descriptors are more closely related to material properties, predictions based on them will be more accurate. The goal of the experimental design in the present study is not the improvement of the material properties as such but the prediction of their properties. To realize this design, we select appropriate controllable descriptors for the synthesis of the material to improve the accuracy of the model when the corresponding uncontrollable descriptors and material properties are added to the training data. We propose two experimental design methods, one based on Bayesian optimization and the other on uncertainty sampling. Using a polymer database in which controllable and uncontrollable descriptors, and mechanical properties are recorded, we confirm that our method can select an appropriate candidate material to train a highly accurate prediction model in which the material properties are predicted by uncontrollable descriptors. Our proposed method can be applied to materials developments where uncontrollable descriptors are more easily obtained by experiments than obtaining target material properties; it will also be useful for extracting the relationship between structure and properties of a material.

ARTICLE HISTORY
Received 5 June 2021
Revised 13 July 2021
Accepted 30 July 2021

KEYWORDS
Uncontrollable descriptors; experimental design; polymer

CLASSIFICATIONS
Data analysis (AI, machine learning, data-driven analysis, descriptor development, structure search/identification)

CONTACT
Ryo Tamura (tamura.ryo@nims.go.jp); Takashi Nakanishi (nakanishi.takashi@nims.go.jp); Masahiko Demura (demura.masahiko@nims.go.jp)
Materials Open Platform for Chemistry, National Institute for Materials Science (NIMS), 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan

Article classification: Data analysis (AI, machine learning, data-driven analysis, descriptor development, structure search/identification)

Supplemental data for this article can be accessed here.

© 2021 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.
1. Introduction

Various successful applications of machine learning have been reported in materials science [1–18]. In particular, predictions using machine learning regression models is powerful for materials science because properties of materials that have not yet been synthesized can be predicted without their actual synthesis and measurements. To successfully conduct research using machine learning predictions, it is important to construct a highly accurate regression model with a training dataset. In general, the accuracy of the regression model is strongly dependent on the amount of training data. However, in materials science, obtaining data requires a significant amount of time and high costs. Thus, it is challenging to create accurate prediction models with as little training data as possible.

A reliable strategy for obtaining an accurate prediction model is feature selection, which searches for important features that increase prediction performance [19–22]. In this study, however, we focus on another strategy: active learning, which selects new data points to be added to the training dataset to improve the prediction accuracy. Here, new data is selected from the candidate materials, which have not been synthesized, prepared in advance. Uncertainty sampling in active learning is useful: the datapoint with the highest uncertainties is selected to improve the prediction accuracy [23–27]. For example, these uncertainties correspond to the deviations of the predicted values, evaluated by such as Gaussian process regression.

However, uncertainty sampling is not always possible for problems in the materials science. As a characteristic of materials science, two types of material descriptors can be used as features when training the regression model (see Figure 1). The first type is controllable descriptors, which can be chosen in advance without synthesizing the material, such as compositions of elements and fabrication processes. The second type is uncontrollable descriptors, which are those generated from experimental data such as raw spectral data and specific gravity of samples and are thus unknown without synthesis. If the features used for regression models are limited to controllable descriptors, a new point to improve the prediction accuracy can be selected by conventional uncertainty sampling. This is because, when the features are controllable descriptors, their ground-truth values are known in advance, and thus the uncertainty (as evaluated by a specific machine-learning prediction) is exactly defined, even if the material has not actually been synthesized. On the other hand, when uncontrollable descriptors are used as features, the situation changes completely: as the candidate dataset cannot be prepared for uncontrollable descriptors without synthesis and measurements, the selection of uncontrollable descriptors by conventional uncertainty sampling is impossible.

This study seeks to obtain a highly accurate prediction model of material properties from uncontrollable descriptors with as few experiments as possible. In general, the dimensions of uncontrollable descriptors generated by experimental data can be increased by performing various measurements, and uncontrollable descriptors are more closely related to material properties in many cases. Therefore, using uncontrollable descriptors should improve prediction accuracy. In addition, if uncontrollable descriptors are obtained...

Figure 1. Examples of controllable and uncontrollable descriptors, and material properties. In this study, we develop an experimental design for obtaining an accurate prediction model where material properties are predicted by uncontrollable descriptors; i.e. we design a method to suggest controllable descriptors for synthesis and measurement to improve the prediction accuracy.
more easily than the target material properties, difficult-to-measure experimental data can be predicted from easy-to-measure ones. This model can be used in various ways (e.g., materials optimization and repurposing of materials), as will be discussed in Sec. 5. Furthermore, using the prediction model, the relationship between structure expressed by uncontrollable descriptors and properties of a material can be understood.

The methodology of the experimental design proposed in this study is as follows. First, we prepare a candidate dataset that is constructed using only controllable descriptors. When a controllable descriptor is selected, the uncontrollable descriptors and target material properties are obtained by synthesizing the material and conducting measurements. Then, we train a regression model to predict the target property from the uncontrollable descriptors without using the controllable ones. Next, we select appropriate controllable descriptors from the candidate dataset such that the accuracy of the prediction model using the uncontrollable descriptors will be high. Unfortunately, the conventional uncertainty sampling method cannot be used directly in this step and thus development of a new method is necessary. We propose two methods for using uncontrollable descriptors in active learning, one based on Bayesian optimization and one based on uncertainty sampling. To demonstrate the performance of these methods, a polypropylene dataset is used in this study.

2. Experimental design method

2.1. Problem establishment

The controllable and uncontrollable descriptors are represented by \( \mathbf{x}_i \in \mathbb{R}^d \) and \( \mathbf{x}_i' \in \mathbb{R}^d \), respectively; \( y_i \in \mathbb{R} \) is the target material property. First, we prepare a dataset constructed using only the controllable descriptors of the candidate materials; this is expressed as \( D_c = \{ \mathbf{x}_i \}_{i=1,...,N} \) where the number of candidate materials is \( N \). We assume that, for only \( M \) controllable descriptors in \( D_c \), expressed as \( D_{c,M} = \{ \mathbf{x}_i \}_{i=1,...,M} \) the uncontrollable descriptors \( \{ \mathbf{x}_i \}_{i=k=1,...,M} \) and target properties \( \{ y_k \}_{k=1,...,M} \) are experimentally obtained. The training data for the regression model is expressed as \( D_{t,M} = \{ \mathbf{x}_k, y_k \}_{k=1,...,M} \). The prediction performance of this regression model, where the uncontrollable descriptors are features and the target properties are labels, is evaluated by the leave-one-out cross-validation method. To improve the prediction performance, we select the next appropriate candidate material characterized by the controllable descriptors from the remaining dataset in \( D_c \setminus D_{c,M} \), where the number of data points is given by \( N - M \). The dataset for the controllable descriptors and target properties, when the number of data points is \( M \), is denoted as \( D_{t,M} = \{ \mathbf{x}_k, y_k \}_{k=1,...,M} \).

2.2. Bayesian optimization-based method

In this section, we introduce a method based on Bayesian optimization. The following steps are taken to select the \( M + 1 \)st data using this scheme. First, the training dataset, \( D'_{t,M} \), is divided randomly into a set of \( M - L \) data and a set of \( L \) data, which are denoted as \( D'_{t,M-L} \) and \( D'_{t,L} \) respectively, where \( L \) is a hyperparameter. The corresponding datasets for the controllable descriptors are denoted as \( D_{c,M-L} \) and \( D_{c,L} \), which are constructed by the division of \( D_{c,M} \). Second, one datapoint is extracted from \( D'_{t,L} \).

We use this extracted datapoint and \( D'_{t,M-L} \) as the training data to create the regression model. The prediction accuracy of the trained regression model is evaluated using such as the leave-one-out error calculated by the mean squared error (MSE). Third, for each datapoint in \( D'_{t,L} \), the second step is performed. Fourth, the Gaussian process regression is trained using the controllable descriptors in \( D_{c,L} \) as features and the previously evaluated prediction accuracies as labels. The amount of training data is denoted by \( L \). Fifth, from \( D_{c} \setminus D_{c,M} \), in terms of Bayesian optimization using the trained Gaussian process regression from the previous step, a single datapoint is selected to minimize the prediction accuracy. Here, the expected improvement (EI) is used as the acquisition function. Finally, for the selected data, the uncontrollable descriptors and material properties are obtained by experiments. The number of known data becomes \( M + 1 \), and the datasets are updated to \( D_{c,M+1} \) and \( D'_{t,M+1} \). Using the dataset \( D'_{t,M+1} \) as the training data, we construct a regression model where the material properties are predicted by uncontrollable descriptors. The prediction accuracy of this regression model is evaluated. The regression model should be the same as that used in step two.

This method selects an \( M + 1 \)st datapoint that improves the prediction accuracy when the selected datapoint is added to the dataset \( D'_{t,M-L} \), not to \( D'_{t,M} \). The original goal is to select the next datapoint to be added to the dataset \( D'_{t,M} \). Thus, our proposed algorithm is regarded as an approximation of this original goal. Decreasing \( L \) to bring \( D'_{t,M-L} \) close to \( D'_{t,M} \) would improve this approximation. However, the Gaussian process regression has \( L \) training data, so decreasing \( L \) will worsen its prediction performance. Hence, \( L \) should be adjusted in a trade-off to find the best value. The
Algorithm 1 Bayesian optimization-based experimental design (BOED) method

Prepare candidate dataset: \(D_c = \{x_i\}_{i=1,...,N}\)
for \(m = 1, 2, \ldots\) do
  Prepare datasets including \(M\) known points: \(D_{c,M} = \{x_k\}_{k=1,...,M}\) and \(D_{L,M} = \{x'_k, y_k\}_{k=1,...,M}\)
  Divide \(D_{L,M}\) randomly into a set of \(M - L\) data \(D_{L,M-L}\) and a set of \(L\) data \(D_{L,L}\)
  for \(l = 1, 2, \ldots, L\) do
    Prepare training dataset \(D'_{L,M-L+1} = \{D'_{L,M-L}, (X'_l, y_l)\}\).
    Train regression model using \(D_{L,M-L+1}\).
    Evaluate the prediction accuracy \(\Delta_l\) by leave-one-out cross-validation of the trained regression model.
    Augment the dataset: \(D_{L,L+1} = \{D_{L,L-1}, (x_l, \Delta_l)\}\).
  end for
Train the Gaussian process regression using \(D_{L,L}\).
Select \(x_m\) by optimizing the acquisition function defined by the trained Gaussian process regression from the remaining dataset \(D_c \setminus D_{c,M}\).
Measure uncontrollable descriptors \(y_m\) and material property \(y_m\) by performing experiments.
Augment the dataset: \(D_{c,M+1} = \{D_{c,M}, x_m\}\) and \(D'_{L,M+1} = \{D'_{L,M}, (x_m, y_m)\}\). Train regression model using \(D'_{L,M+1}\). Evaluate the prediction accuracy when the number of training data is \(M + 1\). End for

Figure 2. Schematic of Bayesian optimization-based experimental design when there are \(M\) known datapoints. The Gaussian process regression model is trained to predict the prediction error of \(D_c \setminus D_{c,M}\). Notation for datasets is explained in Section 2.2.

2.3. Uncertainty sampling-based method

In the alternative method based on uncertainty sampling, the \(M + 1\) st data is selected using the following steps. First, the Gaussian process regression is trained using dataset \(D_{c,M}\) which contains \(M\) training data. Second, a single datapoint is selected from \(D_c \setminus D_{c,M}\) with the largest deviation evaluated by the trained Gaussian process; this step is inspired by conventional uncertainty sampling. Third, by experimentally obtaining the uncontrollable descriptors and material properties for the selected data, the number of known data becomes \(M + 1\). Thus, the datasets are updated to \(D_{c,M+1}\), \(D_{L,M+1}\), and \(D'_{L,M+1}\). We train a regression model using the training dataset \(D'_{L,M+1}\), where the uncontrollable descriptors are features and the material properties are labels. Then, we evaluate the prediction accuracy of this regression model.

This method selects the \(M + 1\) st data by choosing which datapoint has the highest uncertainty for the trained Gaussian process regression where the material properties are predicted by uncontrollable descriptors. The original goal is to select the most uncertain data for a regression model in which the material properties are predicted by uncontrollable descriptors. If there is any correlation between the controllable and uncontrollable descriptors, this algorithm will be successful. The pseudo-code of this uncertainty sampling-based experimental design method (USED) is shown in Algorithm 2.

Algorithm 2 Uncertainty sampling-based experimental design (USED) method
3. Target dataset

To demonstrate the efficiency of our proposed methods in Section 2, 75 types of polypropylene data are used. The material properties outlined here are the mechanical properties of the Charpy impact test and the tensile modulus. The three controllable descriptors for the polymers are set to the molecular weight, mmm in tacticity, and the injection-cooling temperature. The uncontrollable descriptors are prepared from differential scanning calorimetry (DSC), wide-angle X-ray diffraction (WAXD), pulse nuclear magnetic resonance (puls-NMR), the skin layer observed by polarizing optical microscopy, and specific gravity. In these measured data, information about the higher-order structure of polymers is included.

In this study, we verify the efficiency for the case where the prediction of mechanical properties is achieved from uncontrollable descriptors. This is because, if this prediction model is poor, appropriate selection is impossible. Thus, some uncontrollable descriptors are selected in advance. Specifically, the Pearson correlation coefficients between each experimental data and the mechanical property are evaluated, and the top 10 descriptors are extracted for each mechanical property. Two investigations are executed, using the 1st to 5th and 6th to 10th uncontrollable descriptors in descending order, respectively. In Section 4, the results of the experimental design for the case of the three-dimensional controllable descriptors and five-dimensional uncontrollable descriptors are shown.

4. Results

4.1. Comparison of our algorithms and random sampling

We execute the experimental designs described above using the proposed methods and the polymer database. In our implementation, we use the Python library sci-kit-learn [28]. Elastic net regression is used when the mechanical property is predicted by uncontrollable descriptors. In the Supplementary Information, the results of the Gaussian process regression and partial least-squares (PLS) regression are shown. First, we address the performance of regression models when all 75 data points are used as the training data. Figure 3 shows scatter plots of the real and predicted mechanical properties of the test data depending on the features used, i.e. the controllable, 1st to 5th uncontrollable, and 6th to 10th uncontrollable descriptors. The test data is prepared using leave-one-out cross validation, and the prediction results are plotted. The MSE is evaluated for the test data corresponding to the leave-one-out error (LOOE), which is indicated in Figure 3. In addition, the coefficient of determination, $R^2$, for the whole dataset is evaluated and is also denoted in Figure 3. The prediction accuracy for both metrics is better with the 1st to 5th uncontrollable descriptors than with the 6th to 10th descriptors. For the Charpy impact test, the uncontrollable descriptors are better suited as features than uncontrollable descriptors; on the other hand, the uncontrollable descriptors exhibit a higher prediction accuracy for the tensile modulus. These prediction accuracies are the higher bounds of the prediction performance in our experimental design.

Figure 4 shows the prediction accuracies, which are defined by the LOOE calculated using the MSE and $R^2$ for predictions of the whole dataset (i.e. all 75 data points), depending on the iteration step, $m$, of the experimental design when the training data is $D_{1:M} = \{x_k, y_k\}_{k=1,\ldots,M}$. Here, $M$ is the number of known points at each step. When the training data is $D_{1:M}$, predictions for the truly unknown data, $D_c \setminus D_{1:M}$, at each step are included in $R^2$. On the other hand, the LOE is evaluated by only $D_{1:M}$. Here, the number of initial data points is 10, where these initial data are randomly selected from $D_c$. Then, in each iteration step, $m$, the number of training datapoints is $M = m + 10$. In addition, 100 runs of the experimental design with different initial choices are performed and the prediction accuracies are averaged. The results using BOED and USED are compared with those obtained when the next candidate material is randomly selected from the remaining candidates. For BOED, we consider the case where the hyperparameter $L$ is set to an integer of $M/3$, so that the value of $L$ increases with the number of steps.
4.2. Hyperparameter dependency of the BOED

In this section, we address the dependence of the hyperparameter L on the prediction accuracies in the experimental design. As explained in Section 2.2, finding the optimal value of L involves a trade-off between approximation quality (i.e. distance of $D_{\text{LM}}$ from $D_{\text{LM}}$) and number of training data in the Gaussian process. In Figure 5, we show the prediction accuracies depending on the iteration step when L is changed to integers of $M/2$, $M/3$, $M/4$, and $M/5$. In almost all cases, for a smaller value of L, a better accuracy is realized. For $M/2$, which has the largest number of training data in the Gaussian process, the worst accuracy is obtained. Thus, in the trade-off, the improvement of the approximation is more important than the increase in the amount of training data. The number of initial data is set at 10 and thus L cannot be reduced further. If the amount of data is increased, it is expected that better results will be obtained by controlling L.

5. Discussion and summary

We proposed methods for conducting experimental designs to improve the prediction accuracy of a regression model where the properties of material are predicted by descriptors obtained by measurement. In other words, this study considered the treatment of uncontrollable descriptors, which are those generated from experimental data such as raw spectra and specific gravity and are thus unknown without synthesis, as features for prediction models in active learning. The methodology of the study is as follows. We focus on the situation in which if controllable descriptors, such as composition and processes, are
determined, corresponding uncontrollable descriptors and target material properties are obtained by synthesis and measurements. Then, the material expressed by controllable descriptors to be synthesized is appropriately selected to improve the accuracy of the regression model for predicting material property by uncontrollable descriptors.

Two methods (based on Bayesian optimization and uncertainty sampling, respectively) were proposed, and their efficiencies were addressed. We used the polypropylene database and the Charpy impact test and tensile modulus as target properties. Our proposed experimental designs could appropriately select the next candidate material to train a highly accurate prediction model where the material properties are predicted by the uncontrollable descriptors. With a small number of training data, better predictions were realized than by random sampling. We found that the proposed method works better when regression models with better prediction performances will be achieved. In the main text, the results of the elastic net regression were introduced; in Figures. S1-6, in the

![Figure 4](image_url)

*Figure 4. Method dependence of the prediction accuracies, which are defined as the LOOE calculated by MSE (upper panels) and $R^2$ for predictions of whole dataset (lower panels), against the iteration steps when (a) Charpy impact test or (b) tensile modulus is targeted. Elastic net regression is chosen as the regression model. The results of Bayesian optimization-based experimental design (BOED), uncertainty-sampling-based experimental design (USED), and random sampling are compared. The left panels are the results when the 1st to 5th uncontrollable descriptors are used, and the right panels are the results for the 6th to 10th uncontrollable descriptors.*
Supplementary Information, the results of the Gaussian process and PLS regression are shown. We confirm that our experimental design works well when these regression methods are used. Note that although we used only uncontrollable descriptors as features for the prediction model in this study, our method can be also applied to create a better prediction model using both controllable and uncontrollable descriptors.

If the uncontrollable descriptors are obtained more easily than the target material properties, the prediction of difficult-to-measure experimental data from easy-to-measure ones can be achieved by our experimental design, accelerating materials design and reducing the cost of materials development. For example, materials optimization can be conducted by using our prediction model in the determination of target material properties. Note that such materials optimization can also be performed using only controllable descriptors. However, there are many more uncontrollable than controllable descriptors, and

Figure 5. Hyperparameter L dependence for the prediction accuracies in Bayesian optimization-based experimental design, which are defined as the LOOE calculated by the MSE (upper panels) and R² for predictions of the whole dataset (lower panels), against the iteration steps when (a) Charpy impact test or (b) tensile modulus is targeted. Elastic net regression is chosen as the regression model. The calculated L takes integer values of M/2, M/3, M/4, and M/5, where M is the number of known points.
uncontrollable descriptors are more closely related to material properties in many cases. Therefore, using uncontrollable descriptors allows better prediction of material properties and more effective optimizations. Also, if an accurate prediction model can be obtained, materials can be repurposed, and databases including uncontrollable descriptors (such as existing XRD databases) can be screened using the prediction model to discover materials having desired target properties.

From another perspective, using the prediction model and collecting data by our experimental design can lead to deeper understanding of materials. This is true even in the case that both uncontrollable descriptors and material properties are difficult to measure. For example, the origin of materials’ properties can be clarified by extracting relations between them and the materials’ structure expressed by uncontrollable descriptors.

In addition, because our method can reduce the number of experiments required for a good prediction model, it is well-suited for use with laboratory automation technologies [29–31], which have recently attracted much attention. Furthermore, even when target material properties themselves cannot be measured robotically, materials optimization can be conducted automatically by replacing the determination of the target material property with the prediction obtained by our experimental design.

This study had a preliminary character, proposing the method and demonstrating its usefulness using a specific polymer database. Future aims include the execution of the real experimental design without the use of the database and the actual experimental generation of a dataset with the help of our proposed methods. To consider a case in which a highly accurate prediction model could be obtained, we selected and used uncontrollable descriptors that were known in advance to have strong correlations with the target property. However, when designing actual experiments, useful uncontrollable descriptors might not be known ahead of time, and thus further improvement of the method is necessary for practical use.

There are both controllable and uncontrollable descriptors; the need to predict properties with high accuracy from uncontrollable ones occurs frequently throughout materials science. We believe that our proposed method, even if it is not yet optimal, points to one way that this unique problem can be solved. We expect our study to stimulate the development of a method suitable for this unique problem.

**Acknowledgements**

We thank Miharu Shimizu for conducting the simulations. We appreciate all members in the Materials Open Platform for Chemistry, composed of NIMS and four companies (Asahi Kasei Corporation, Mitsubishi Chemical Corporation, Mitsui Chemicals, Inc., and Sumitomo Chemical Co., Ltd.), for valuable discussions. In particular, we thank Fumisato Goto, Hiroshi Hamamatsu, Takeaki Kakigano, Kazunori Kamio, Shuichi Kimata, Shinichi Kitade, Teiichiro Kohno, Kenji Nagata, Toshinao Shioya, Keitaro Sodeyama, and Hisao Takesuchi for useful discussions. We would like to thank Editage (www.editage.com) for English language editing.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**ORCID**

Ryo Tamura [http://orcid.org/0000-0002-0349-358X](http://orcid.org/0000-0002-0349-358X)

Takashi Nakanishi [http://orcid.org/0000-0002-8744-782X](http://orcid.org/0000-0002-8744-782X)

Masahiko Demura [http://orcid.org/0000-0002-7308-3041](http://orcid.org/0000-0002-7308-3041)

**References**

[1] Pilania G, Wang C, Jiang X, et al Accelerating materials property predictions using machine learning. Sci Rep. 2013;3:1–6.

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

[3] 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.

[4] Ikebata H, Hongo K, Isomura T, et al Bayesian molecular design with a chemical language model. J Comput Aided Mol Des. 2017;31:379–391.

[5] Ju S, Shiga T, Feng L, et al Designing nanostructures for phonon transport via Bayesian optimization. Phys Rev X. 2017;7:021024.

[6] Kim E, Huang K, Jegelka S, et al Virtual screening of inorganic materials synthesis parameters with deep learning. Npj Comput Mater. 2017;3:1–9.

[7] Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci. 2017;129:156–163.

[8] Gómez-Bombarelli R, Wei JN, Duvenaud D, et al Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4:268–276.

[9] Sawada R, Iwasaki Y, Ishida M. Boosting material modeling using game tree search. Phys Rev Materials. 2018;2:103802.

[10] Sumita M, Yang X, Ishihara S, et al Hunting for organic molecules with artificial intelligence: molecules optimized for desired excitation energies. ACS Cent Sci. 2018;4:1126–1133.

[11] Sakurai A, Yada K, Simomura T, et al Ultranarrow-band wavelength-selective thermal emission with aperiodic multilayered metamaterials designed by Bayesian optimization. ACS Cent Sci. 2019;5:319–326.

[12] Wu S, Kondo Y, Kakimoto M, et al Machine-learning-assisted discovery of polymers with high thermal
conductivity using a molecular design algorithm. Npj Comput Mater. 2019;5:1–11.

[13] Minami T, Kawata M, Fujita T, et al Prediction of repeat unit of optimal polymer by Bayesian optimization. MRS Adv. 2019;4:1125–1130.

[14] Homma K, Liu Y, Sumita M, et al Optimization of a heterogeneous ternary Li$_3$PO$_4$–Li$_2$BO$_3$–Li$_2$SO$_4$ mixture for Li-ion conductivity by machine learning. J Phys Chem C. 2020;124:12865–12870.

[15] Kitai K, Guo J, Ju S, et al Designing metamaterials with quantum annealing and factorization machines. Phys Rev Res. 2020;2:013319.

[16] Nakajin K, Minami T, Kawata M, et al Prediction of physical properties of thermosetting resin by using machine learning and structural formulas of raw materials. MRS Adv. 2020;5:1567–1575.

[17] Ohkubo I, Hou Z, Lee JN, et al Realization of closed-loop optimization of epitaxial titanium nitride thin-film growth via machine learning. Mater Today Phys. 2021;16:100296.

[18] Tamura R, Osada T, Minagawa K, et al Machine learning-driven optimization in powder manufacturing of Ni-Co based superalloy. Mater Des. 2021;198:109290.

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

[20] Stanev V, Oses C, Kusne AG, et al Machine learning modeling of superconducting critical temperature. Npj Comput Mater. 2018;4:1–14.

[21] Sodeyama K, Igarashi Y, Nakayama T, et al Liquid electrolyte informatics using an exhaustive search with linear regression. Phys Chem Chem Phys. 2018;20:22585–22591.

[22] Tamura R, Watanabe M, Mamiya H, et al Materials informatics approach to understand aluminum alloys. Sci Technol Adv Mater. 2020;21:540–551.

[23] 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;5:1–17.

[24] Pruksawan S, Lombard G, Samitsu S, et al Prediction and optimization of epoxy adhesive strength from a small dataset through active learning. Sci Technol Adv Mater. 2019;20:1010–1021.

[25] Terayama K, Tamura R, Nose Y, et al Efficient construction method for phase diagrams using uncertainty sampling. Phys Rev Mater. 2019;3:033802.

[26] 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.

[27] Katsube R, Terayama K, Tamura R, et al Experimental establishment of phase diagrams guided by uncertainty sampling: an application to the deposition of Zn–Sn–P films by molecular beam epitaxy. ACS Mater. Lett. 2020;2:571–575.

[28] scikit-learn: machine learning in Python [Internet]. [cited 2021 Jan 29]. Available from: https://scikit-learn.org/stable/

[29] Matsuda S, Nishioka K, Nakanishi S. High-throughput combinatorial screening of multi-component electrolyte additives to improve the performance of Li metal secondary batteries. Sci Rep. 2019;9:6211.

[30] Burger B, Maffettone PM, Gusev VV, et al A mobile robotic chemist. Nature. 2020;583:237–241.

[31] Shimizu R, Kobayashi S, Watanabe Y, et al Autonomous materials synthesis by machine learning and robotics. APL Mater. 2020;8:111110.