scientific reports

Check for updates

OPEN

Performance of uncertainty-based active learning for efficient approximation of black-box functions in materials science

Ai Koizumi[1]✉, Guillaume Deffrennes[2], Kei Terayama[3,4,5]✉ & Ryo Tamura[1,5,6]✉

Obtaining a fine approximation of a black-box function is important for understanding and evaluating innovative materials. Active learning aims to improve the approximation of black-box functions with fewer training data. In this study, we investigate whether active learning based on uncertainty sampling enables the efficient approximation of black-box functions in regression tasks using various material databases. In cases where the inputs are provided uniformly and defined in a relatively low-dimensional space, the liquidus surfaces of the ternary systems are the focus. The results show that uncertainty-based active learning can produce a better black-box function with higher prediction accuracy than that by random sampling. Furthermore, in cases in which the inputs are distributed discretely and unbalanced in a high-dimensional feature space, datasets extracted from materials databases for inorganic materials, small molecules, and polymers are addressed, and uncertainty-based active learning is occasionally inefficient. Based on the dependency on the material descriptors, active learning tends to produce a better black-box functions than random sampling when the dimensions of the descriptor are small. The results indicate that active learning is occasionally inefficient in obtaining a better black-box function in materials science.

Keywords Active learning, Alloy, Semiconductor, Polymer, Molecule, PHYSBO

In materials science, materials are synthesized based on the information required to develop these (such as compositions, structures, and processes), and their properties are measured. It can be considered as a function of the information on materials as inputs and material properties as outputs. This function is called a black-box function (BBF) because it cannot be described analytically. To approximate a BBF in materials science, machine learning (ML) models can be used to accurately predict material properties for various inputs. Recently, the concept of black-box optimization (BBO) has attracted attention for the efficient optimization of the inputs of BBF[1–3]. In BBO, the potential materials are selected using an ML model approximating the BBF. Their properties are obtained experimentally or by simulations. Using the newly obtained data, the ML model is updated continually through optimization. Although it is well known that the input of a BBF can be explored efficiently by BBO[4–9], this approach does not ensure the accuracy of the BBF approximation.

The training dataset and the input of a BBF determine the accuracy of the BBF approximation. To obtain a better input, feature selection methods are useful[10–12]. However, in the case where we do not have enough training data, it is difficult to select better input because it may be highly data dependent. To prepare informative training datasets for constructing a fine approximation of a BBF, Active Learning (AL)[13–16] is useful. AL has been studied in informatics primarily for classification tasks. It addresses the problem of selecting unlabeled data to be observed to improve the prediction accuracy of ML models. Studies using AL in materials science have also been conducted. A study using ML classification models was the phase diagram construction method[17–19]. This method uses the uncertainty sampling approach[20] as AL. The most uncertain point in the phase diagram is selected as

[1]Center for Basic Research on Materials, National Institute for Materials Science, 1-1, Namiki, Tsukuba, Ibaraki 305-0044, Japan. [2]Univ. Grenoble Alpes, CNRS, Grenoble INP, SIMaP, F-38000 Grenoble, France. [3]Graduate School of Medical Life Science, Yokohama City University, 1-7-29, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [4]MDX Research Center for Element Strategy, Tokyo Institute of Technology, 4259, Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8501, Japan. [5]RIKEN Center for Advanced Intelligence Project, 1-4-1, Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan. [6]Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5, Kashiwa-no-ha, Kashiwa, Chiba 277-8561, Japan. ✉email: koizumi.ai@nims.go.jp; terayama@yokohama-cu.ac.jp; tamura.ryo@nims.go.jp

the next candidate for the experiments. This method can actively sample points near the phase boundaries. A detailed phase diagram can be obtained with 20% of the experiments that would have been required for random sampling. Another study that targeted ML regression models was on X-ray magnetic circular dichroism spectroscopy[21]. The study considered the extent to which the number of experiments required to obtain detailed spectra can be reduced. In this technique, a one-dimensional Gaussian process regression (GPR) learns the data measured from a small number of experiments. An experiment to determine the X-ray energy with the largest variance evaluated by GPR was performed. This procedure was iterated. It was reported that a detailed spectrum could be obtained with 20% of the total number of experiments. Tian et al. investigated the efficiency of AL for different material datasets[22,23]. They reported that a fine approximation of a BBF can be realized by AL when the dimension of the inputs is small. In addition, Jose et al. applied various AL methods for regression to superconductivity data and compared their accuracies[24]. The AL methods considered are model-free methods (which select data only from the input information) and model-based methods (which select data according to a trained ML model). They reported that the accuracy depends strongly on the method. Many methods are less accurate than random sampling. Therefore, the effectiveness of AL is not ensured for regression problems in materials science.

In this study, we addressed the performance of AL on various materials datasets when material and molecular descriptors are used. In the fields of materials informatics and chemoinformatics, various descriptors generated by compositions and structures are introduced as input to a BBF. For example, the Matminer descriptors and the Morgan fingerprint are commonly used for inorganic and organic materials, respectively. Their dimensions are often large (45 and 2048 dimensions for the former and the latter, respectively). In addition, to provide useful information for the case where the dataset is built from scratch, we considered the case where the AL starts with almost no data. In such a case, it is difficult to select important elements in the descriptors using feature selection methods due to the small size of the training dataset. Therefore, we focused here on the performance of AL without feature selection for material and molecular descriptors. Furthermore, only model-based AL was considered. Uncertainty sampling is the simplest model-based AL. In uncertainty-based approaches, the most uncertain point defined by the ML prediction results is selected as the next candidate for labeling (Fig. 1). Here, the uncertainty value was calculated using GPR as described in a previous study[21]. The Python package PHYSBO[25] was used to learn the GPR model. As materials datasets, the liquidus surfaces of ternary systems were prepared as cases in which the inputs for a BBF were given uniformly and defined in a low-dimensional space. Data were generated by CALPHAD calculations with a constant composition step as shown in Fig. 1. Materials databases focus on cases in which the inputs for a BBF are distributed discretely and unbalanced in a high-dimensional feature space. Specifically, datasets of bandgaps and dielectric constants for electrons and lattices calculated by the density functional theory (DFT) for inorganic materials[26], absorption wavelengths and intensities calculated by DFT for small molecules[27], and glass transition temperatures of polymers obtained from PoLyInfo[28] were used. These data have been converted into material or molecular descriptors, and these descriptors are not uniformly distributed in the feature space, i.e. the distribution is unbalanced. We observed that the accuracy of uncertainty-based AL depended strongly on the dataset. In particular, for materials databases, AL tends to produce a finer approximation of a BBF than random sampling when the dimensions of the material descriptors are small.

Uncertainty-based active learning

This section describes the verification of the performance of the uncertainty-based AL method. Let the dataset $\{\mathbf{x}_i, y_i\}_{i=1,\ldots,N}$ consist of N data points with known labels. Here, $\mathbf{x}_i$ is the input for a BBF, and y_i is the

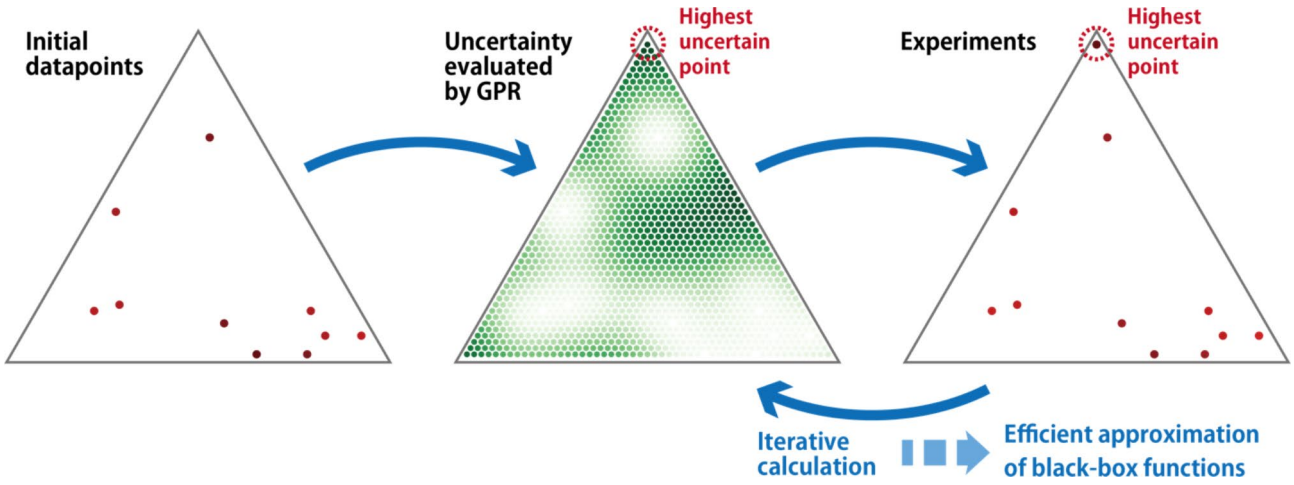

Fig. 1. Flow of uncertainty-based active learning. First, a Gaussian process regression (GPR) model is trained by the initial datapoints. Next, the datapoint with the highest uncertainty calculated by GPR is selected, and the property is measured experimentally. By iterating the training of the GPR model, selection of the most uncertain point, and experiments, an efficient approximation of BBFs can be achieved.

output value of the BBF. Here, we focus on the materials datasets, where $\mathbf{x}_i$ and y_i are the material descriptors and material properties, respectively. Using known datasets, we hypothetically performed AL and verified its performance according to the following procedure:

1. The validation data, $N_{\rm val}$, are selected in advance from N data. To ensure variety in the validation dataset, the interval between the minimum and maximum values of the outputs i.e., $[\min_i y_i,\ \max_i y_i]$, is divided into 100 equal bins. The validation dataset is prepared by randomly selecting a data point from each bin. Therefore, the maximum number of $N_{\rm val}$ is 100.
2. $N_{\rm ini}$ data are selected randomly from the remaining $N_{\rm train} = N - N_{\rm val}$ data as initial data. This dataset is denoted as $D = \{\mathbf{x}_j, y_j\}_{j=1,\cdots,N_{\rm ini}}$.
3. Using D as training data, the GPR model is learned.
4. Using the GPR model, predictions are made for $N_{\rm val}$, and the prediction accuracy is evaluated.
5. Using the GPR model, predictions are made for the data in $N_{\rm train}$ that are not included in D. The data with the largest value of acquisition function was added to D.
6. Steps 3, 4, and 5 are repeated.

To perform AL, we consider the following four types of acquisition functions used in Step 5:

$$f_{\rm US}(\mathbf{x}) = \sigma(\mathbf{x}), \tag{1}$$

$$f_{{\rm TS}-\mu}(\mathbf{x}) = {\rm TS}(\mathbf{x}) - \mu(\mathbf{x}), \tag{2}$$

$$f_{\rm TS}(\mathbf{x}) = {\rm TS}(\mathbf{x}), \tag{3}$$

$$f_{\rm Random}(\mathbf{x}) = \text{Uniform random number in } [0,1], \tag{4}$$

where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are the mean and standard deviation of the prediction values from the GPR model when $\mathbf{x}$ is inputted as a material descriptor. ${\rm TS}(\mathbf{x})$ is the acquisition function by Thompson sampling. Here, the sampling is performed from the normal distribution with mean $\mu(\mathbf{x})$ and standard deviation $\sigma(\mathbf{x})$, which is obtained by PHYSBO. $f_{\rm US}(\mathbf{x})$ is the simplest function for uncertainty-based AL. It selects the point at which the prediction uncertainty is the highest. $f_{{\rm TS}-\mu}(\mathbf{x})$ is also a function for uncertainty-based AL. It is based on Thompson sampling, which is generated from the normal distribution with mean zero and standard deviation $\sigma(\mathbf{x})$. In contrast, $f_{\rm TS}(\mathbf{x})$ is an acquisition function for Bayesian optimization used to search for materials with better properties. $f_{\rm Random}(\mathbf{x})$ is the case wherein data are selected randomly in each iteration. It is the basis for determining whether the uncertainty-based AL is effective. The methods using $f_{\rm US}(\mathbf{x})$, $f_{{\rm TS}-\mu}(\mathbf{x})$, $f_{\rm TS}(\mathbf{x})$, and $f_{\rm Random}(\mathbf{x})$ are called US, TS-μ, TS, and Random, respectively.

To evaluate the prediction accuracy in step 4, two cases were considered. That is, GPR or random forest regression (RFR) models were used to predict on the validation dataset. In the first case, the same ML model was used to evaluate the uncertainty and prediction accuracy. Meanwhile, the GPR and RFR models were used to evaluate the uncertainty and prediction accuracy, respectively, in the second case. The prediction accuracy was calculated as the coefficient of determination R^2between the true and predicted values for the validation data. PHYSBO was used to learn the GPR models, and scikit-learn[29] was used to learn the RFR models. To obtain the statistics, 200 independent trials were performed with different initial data selections in Step 2, and the means and standard deviations were calculated.

Results

In Section "Liquidus surfaces of ternary alloys", the liquidus surfaces of the ternary systems are focused on as cases where the inputs are given uniformly and defined in a relatively low-dimensional space. For cases where the inputs were distributed discretely and unbalanced in a high-dimensional feature space, datasets extracted from materials databases of inorganic materials (Section "Inorganic materials datasets"), small molecules (Section "Small molecule datasets"), and polymers (Section "Polymers dataset") were considered.

Liquidus surfaces of ternary alloys

We investigated the performance of AL in predicting the liquidus surfaces of ternary alloys obtained from CALPHAD calculations. In this study, the ternary alloys Al-Si-Zn[30], Cu-Mg-Zn[31], and Al-Mg-Zn[32] were focused on. Although the ternary liquidus temperatures can be projected into a two-dimensional space, for simplicity, we used the composition of the three elements as an input for a BBF. The compositions were discretized in 2% increments. The total number of data was $N = 1326$. Figure 2 shows the liquidus temperatures and scatter plots when predicting $N_{\rm val}$ using all the remaining $N - N_{\rm val}$ data for training by GPR. These demonstrate that the liquidus temperatures can be predicted accurately. The prediction accuracies depending on the iteration steps when AL with $N_{\rm ini} = 10$ was performed are shown in Fig. 2. Here, GPR was used as the prediction model. In all the cases, US achieved the highest accuracy. That is, better R^2 values were obtained when the number of iteration steps was small. Furthermore, US and TS-μ showed similar and better results. However, in all the cases, TS showed the most inferior results. This indicates that the prediction accuracy occasionally does not improve when BBO is performed. Similar results were obtained when the RFR was used as the prediction model (Fig. S1). Thus, we conclude that for the liquidus surface data, the AL is effective in obtaining better BBFs. This conclusion is consistent with the results of a previous study[22].

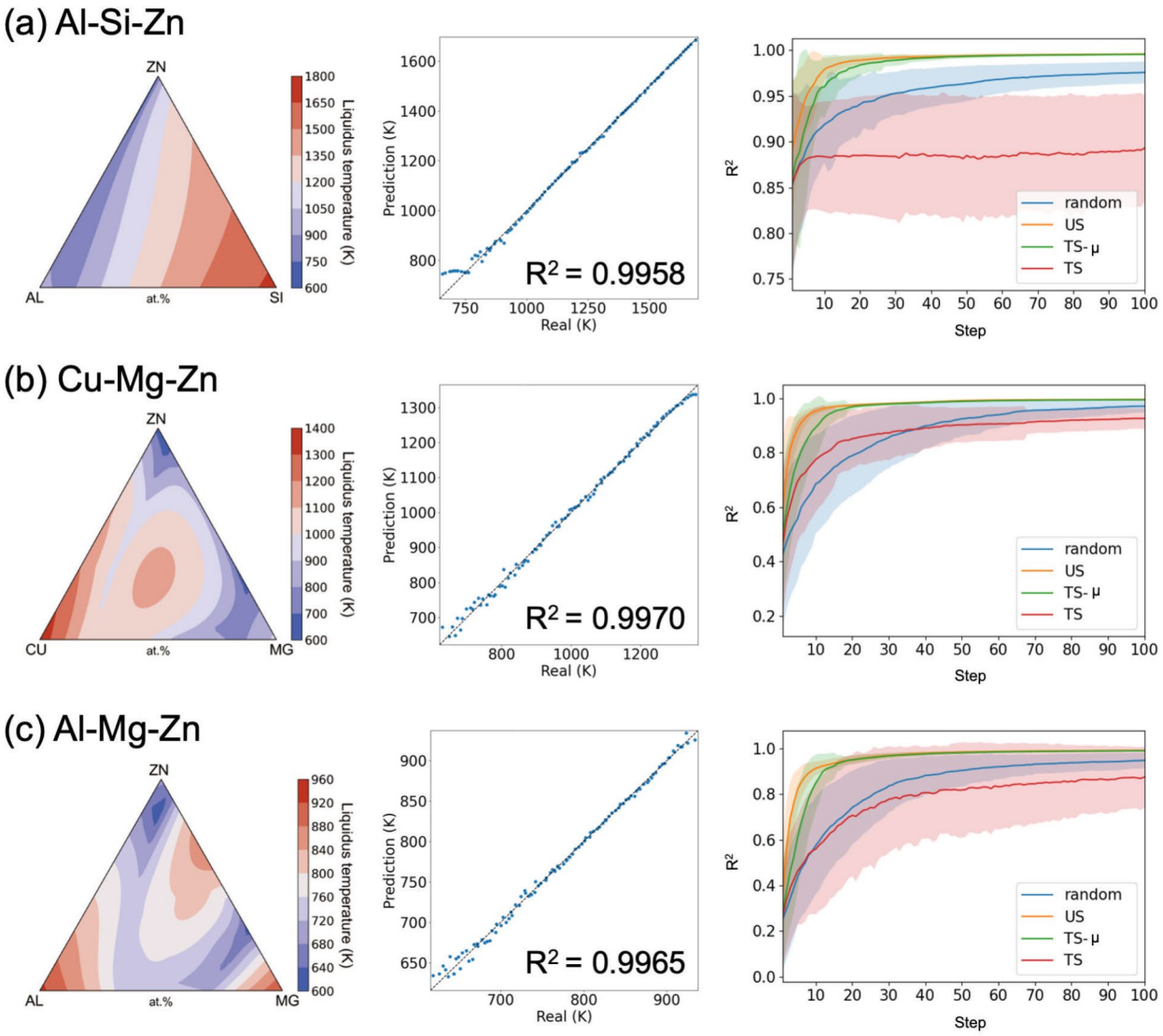

Fig. 2. Liquidus temperature (left panels), scatter plot when predicting N_{val} data using all the remaining $N - N_{\mathrm{val}}$ data for training (center panels), and the prediction accuracy depending on the iteration steps (right panels) for the (**a**) Al-Si-Zn, (**b**) Cu-Mg-Zn, and (**c**) Al-Mg-Zn systems when the prediction model is GPR. The number of initial data is fixed as $N_{\mathrm{ini}} = 10$. The 200 independent runs are performed. The mean and standard deviation are depicted as lines and shaded areas, respectively.

Inorganic materials datasets

In this section, we focus on the physical properties of inorganic materials. We extracted $N = 1255$ semiconductors from the semiconductor dataset reported in Ref.[26] containing bandgap values and dielectric constants for electrons and lattices for oxides calculated by DFT. Six descriptors were used to train the ML prediction model based on these properties. Compositional descriptors were obtained from the composition information and properties of the pure elements. Matminer descriptors[33], magpie descriptors[34], and Deml descriptors[35]were adopted in this study. As the structural descriptor, orbital field matrix[36], JarvisCFID descriptors[37], and radial distribution function were used. These descriptors were generated for $N = 1255$ oxides using the matminer python package[33]. The components in the descriptors where all the oxides had an equal value were removed. The dimensions of the descriptors are summarized in Table S1.

Figure 3 shows the scatter plots when predicting N_{val} using all the remaining $N - N_{\mathrm{val}}$ data for training when the prediction model was GPR. The results by matminer descriptors and orbital field matrix are shown here. The other results are summarized in Figs. S2 and S3. For the dielectric constants, a logarithmic scale was adopted because its prediction accuracy is better than that of the normal scale. The results showed that it is difficult to predict the dielectric constant of a lattice using both compositional and structural descriptors. To validate the AL strategy, the prediction accuracy depending on the number of iteration steps when $N_{\mathrm{ini}} = 10$ is shown in Fig. 3. When the matminer descriptor was used, US and TS-μ performed better than random

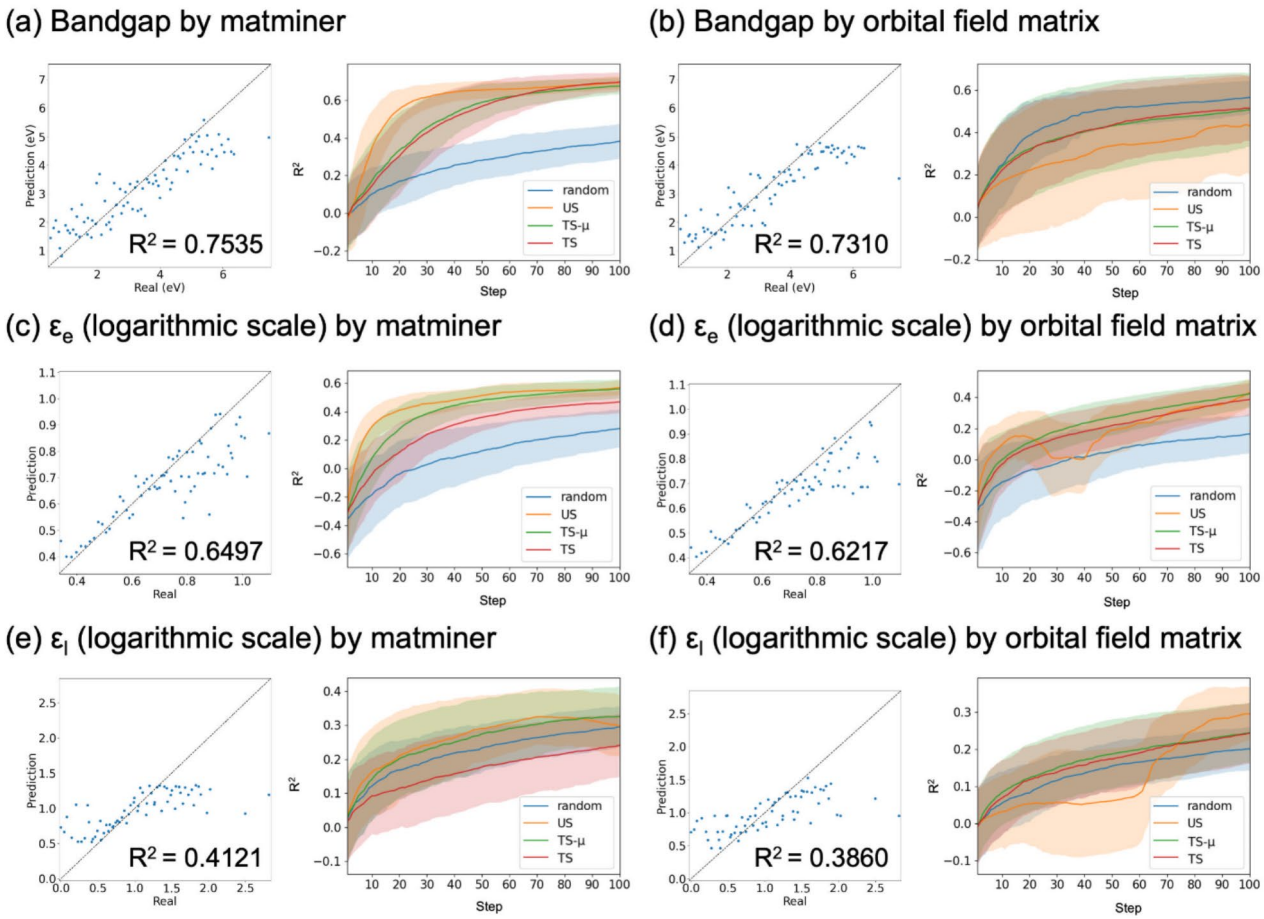

Fig. 3. Scatter plots when predicting N_{val} data using all the remaining $N - N_{\mathrm{val}}$ data for training (left panels) and the prediction accuracy depending on the iteration steps (right panels) for bandgaps, dielectric constants for electron $(\varepsilon_{\mathrm{e}})$ and lattice $(\varepsilon_{\mathrm{l}})$ by matminer and orbital field matrix. The ML model is trained by GPR. The 200 independent runs are performed, and the mean and standard deviation are plotted as lines and shaded areas, respectively.

sampling for the three physical properties. However, when the orbital field matrix was used, the AL strategy was occasionally ineffective. Thus, for an inorganic dataset, the performance of AL depends strongly on the material properties and descriptors. Similar results were obtained when an RFR model was used to evaluate the prediction accuracy (Figs. S4 and S5).

Small molecule datasets

Next, we considered the performance of AL for the calculated absorption wavelengths and intensities of the small molecules. We extracted $N = 1255$ molecules from the ZINC database[38] for our analyses to arrange the amount of data used in Sections "Inorganic materials datasets" and "Polymers dataset". Their properties were calculated using the DFT in Ref.[27]. SMILES notation is used to describe the structure of the molecules. These strings need to be converted into numerical vectors to train an ML prediction model. Here, the Morgan fingerprint[39], MACCS key, and topological fingerprint obtained using RDKit[40] were used. The dimensions of each descriptor are summarized in Table S1. Figure 4(a) and (b) show the scatter plots when predicting the logarithmic scale of each property for N_{val} data using all the remaining $N - N_{\mathrm{val}}$ data by GPR. For these figures, a topological fingerprint was used. The other cases are summarized in Fig. S6. Figure 4(a) and (b) show the prediction accuracy depending on the iteration steps when AL with $N_{\mathrm{ini}} = 10$ was performed. No difference in prediction accuracy between the random sampling and AL methods was observed in any case. When an RFR model was used, all acquisition functions gave similar results (Fig. S7). Thus, for our small-molecule datasets, where wavelength and intensity are focused as material properties, the AL strategy is not effective and random sampling is sufficient to obtain better BBFs.

Polymers dataset

Finally, uncertainty-based AL was validated using a polymer database. Homopolymers recorded in the PoLyInfo database[41] were used. Their properties were set to the glass transition temperature. From the PoLyInfo database, we extracted $N = 1255$ data points for our analyses to arrange the number of data used in Sections "Inorganic

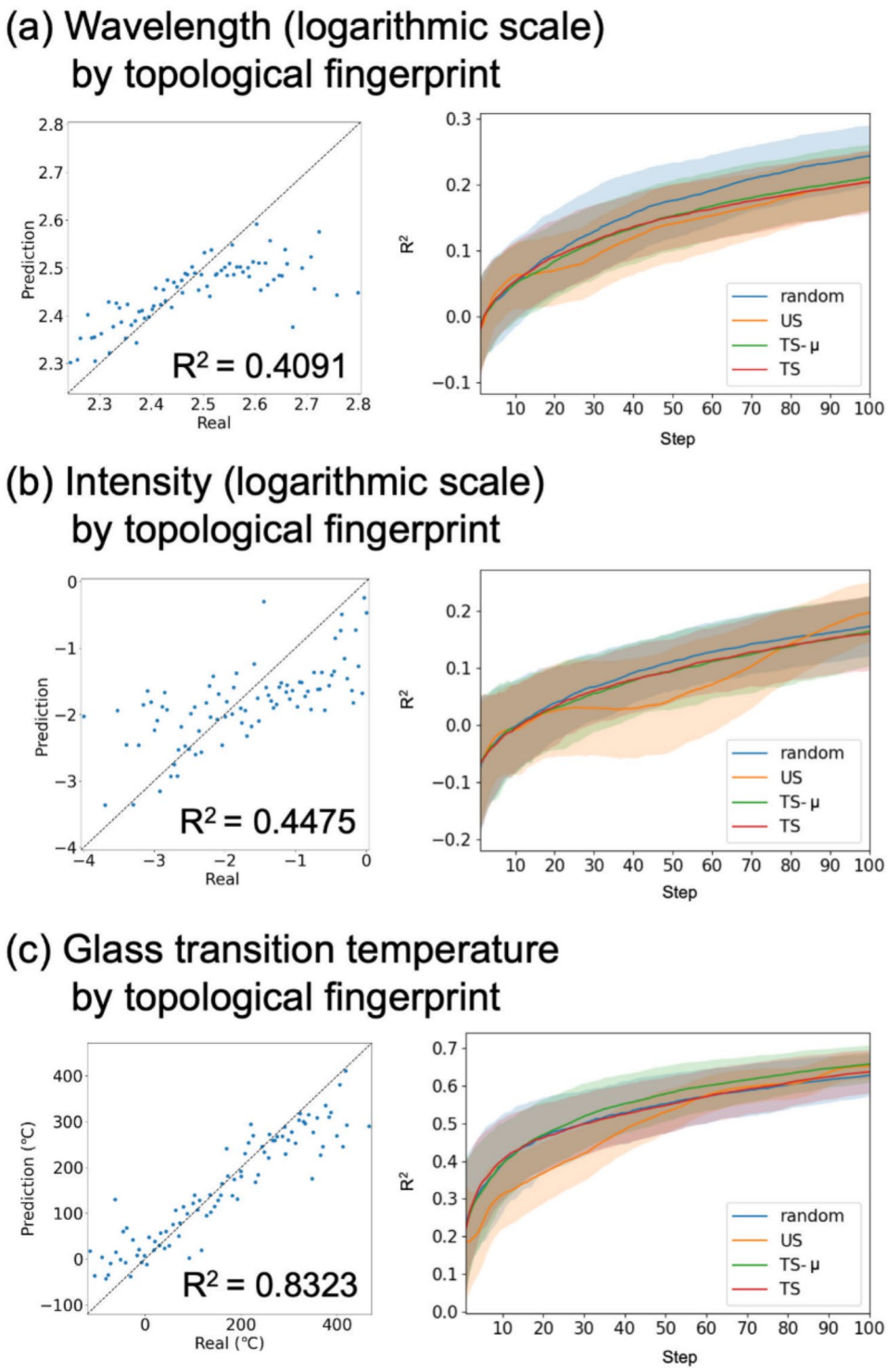

Fig. 4. Scatter plots when predicting N_{val} data using all the remaining $N - N_{\mathrm{val}}$ data for training (left panels) and the prediction accuracy depending on the iteration steps (right panels) for (**a**) the absorption wavelength prediction, (**b**) intensity prediction in small molecule dataset, and (**c**) the glass transition temperature prediction in homopolymer dataset. The ML model is trained by GPR. The 200 independent runs are performed, and the mean and standard deviation are depicted as lines and shaded areas, respectively.

materials datasets" and "Small molecule datasets". In the PoLyInfo database, the polymers are represented by their monomer structures using the SMILES format. The SMILES is converted into Morgan fingerprints, MACCS key, and topological fingerprint by RDKit. The dimensions of each descriptor are summarized in Table S1. The results obtained when topological fingerprints were used are summarized in Fig. 4. The other cases are summarized in Fig. S8. The figure shows that the glass transition temperature of N_{val} data can be predicted accurately using all the remaining $N - N_{\mathrm{val}}$ data. However, there was no difference in prediction accuracy between the random sampling and AL methods. In the first half of iterations, the results obtained by US were inferior to those obtained by random sampling. Regardless of whether we used other descriptors or the RFR, the AL strategy was not more effective than random sampling (Fig. S9). For the material datasets considered in Sections "Inorganic materials datasets", "Small molecule datasets", and "Polymers dataset", even when BBO was performed (i.e., TS was used as an acquisition function), the accuracy of approximating a BBF was similar to that for random sampling.

Discussion

Relationship between data distribution and performance of AL

The following equation introduces an indicator of whether AL performs better than random sampling:

$$\langle \Delta R^2 \rangle = \frac{1}{N_{\mathrm{step}}} \sum_{i=1}^{N_{\mathrm{step}}} \left(R^2_{i,\mathrm{AL}} - R^2_{i,\mathrm{random}} \right), \tag{5}$$

where $R^2_{i,\mathrm{AL}}$ and $R^2_{i,\mathrm{random}}$ are R^2 values between the real and predicted values for N_{val} data at the ith step when AL and random sampling are performed, respectively. Here, N_{step} was set to 100 steps, and US was used as the acquisition function for AL. $\langle \Delta R^2 \rangle > 0$ implies that AL produced a better prediction model than random sampling. $\langle \Delta R^2 \rangle < 0$ implies that AL was not effective. Figure 5(a) shows the results of $\langle \Delta R^2 \rangle$ on a two-dimensional space, where the variance value of the output values $\{y_i\}_{i=1,\ldots,N}$ and the dimension of the descriptors are used as each axis, when $N_{\mathrm{ini}} = 10$. To evaluate the variance, the output values were normalized

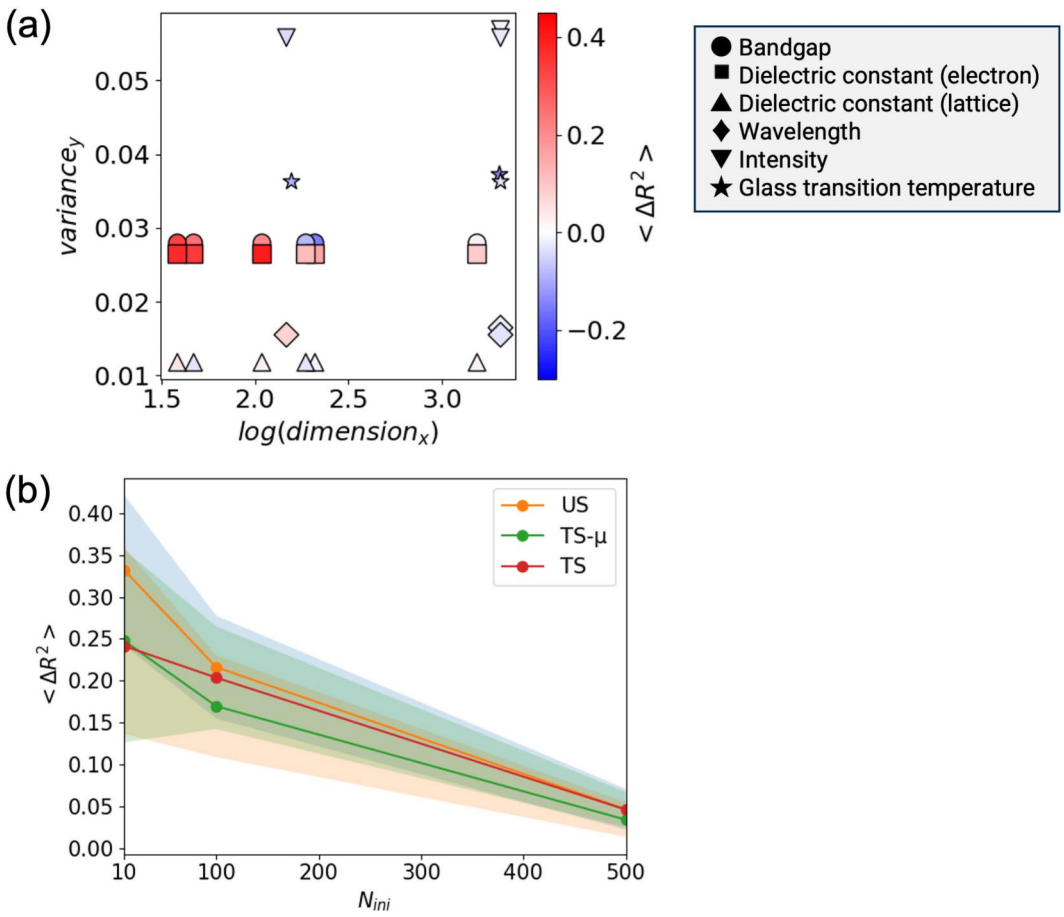

Fig. 5. (**a**) Results of $\langle \Delta R^2 \rangle$ in a two-dimensional space, where the variance of the objective functions $\{y_i\}_{i=1,\ldots,N}$ is vertical axis and the dimension of the inputs for a BBF is horizontal axis, when $N_{\mathrm{ini}} = 10$. (**b**) Results of $\langle \Delta R^2 \rangle$ for these properties after 100 iterations when $N_{\mathrm{ini}} = 10, 100$, and 500 for the bandgap with the matminer descriptor. The ML model is trained by GPR. The depicted dimensions in Panel (**a**) are active dimension which is summarized in Table S1.

such that the minimum and maximum values were zero and one, respectively. The prediction model was GPR. The results for the inorganic materials, small molecules, and polymer datasets were summarized. Figure 5(a) shows that AL tends to produce a better prediction model than random sampling when the descriptor dimension is small. In general, it is well known that the performance of Bayesian optimization degrades for higher dimensions[42]. Various methods such as REMBO[43], LINEBO[44], and SAASBO[45] have been developed to address this problem. In approximating the BBF, the performance may also be improved by considering the process of addressing the high dimensions applied in the above studies. However, no relationship was observed between the variance of the output values and the effectiveness of AL. A similar trend was observed when the RFR was used as the prediction method (see Fig. S10). In addition, we investigated the data distributions using principal component analysis implemented in scikit-learn[29], and results are summarized in Figs. S11 and S12. However, clear relationship between distributions and performance of AL cannot be extracted.

Relationship between the number of initial data and performance of AL

We considered the efficiency of AL and random sampling depending on the number of initial data points. Using a matminer descriptor for the bandgap in inorganic materials, it was verified that AL performed better than random sampling for $N_{\mathrm{ini}} = 10$ (see Fig. 3). Figure 5(b) shows the result of $\langle \Delta R^2 \rangle$ for the bandgap after $N_{\mathrm{step}} = 100$ iterations when $N_{\mathrm{ini}} = 10, 100$, and 500. As N_{ini} increased, $\langle \Delta R^2 \rangle$ approached zero. This indicated that the performance of AL is approximately equal to that of random sampling when N_{ini} is sufficiently large. This implies that if N_{ini} is sufficiently large, a better prediction model can be constructed in the initial stage. From this point onward, adding more data did not significantly alter the prediction accuracy, and the difference between AL and random sampling was insignificant.

Conclusion

In this study, we tested the performance of AL in constructing a fine approximation of the BBF on several material datasets. First, we focused on the BBF for the liquidus temperature in the ternary systems, whose inputs were given as continuous values. This continuous space was discretized uniformly. The next point to be evaluated was selected using an uncertainty-based AL approach. The results showed that uncertainty sampling produced a finer approximation of the BBF more efficiently than random sampling. However, when a material database was used in which the inputs were distributed discretely and unbalanced in a high-dimensional space, the results depended strongly on the dataset. In particular, when the dimensions of the descriptors were high, the difference between AL and random sampling was not verified. Using material descriptors, the inputs for a BBF can straightforwardly be over 100-dimensional. Thus, for many material datasets, AL is ineffective at producing a prediction model with a high accuracy compared with random sampling.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 2 October 2023; Accepted: 16 October 2024
Published online: 06 November 2024

References

1. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput. Mater.* **5**, 1–17 (2019).
2. Terayama, K., Sumita, M., Tamura, R. & Tsuda, K. Black-box optimization for automated discovery. *Acc. Chem. Res.* **54**, 1334–1346 (2021).
3. Jin, Y. & Kumar, P. V. Bayesian optimisation for efficient material discovery: A mini review. *Nanoscale* **15**, 10975–10984 (2023).
4. Sakurai, A. et al. Ultranarrow-band wavelength-selective thermal emission with aperiodic multilayered metamaterials designed by bayesian optimization. *ACS Cent. Sci.* **5**, 319–326 (2019).
5. Ju, S. et al. Designing nanostructures for phonon transport via bayesian optimization. *Phys. Rev. X* **7**, 021024 (2017).
6. Minami, T. et al. Prediction of repeat unit of optimal polymer by Bayesian optimization. *MRS Adv.* **4**, 1125–1130 (2019).
7. Fukazawa, T., Harashima, Y., Hou, Z. & Miyake, T. Bayesian optimization of chemical composition: A comprehensive framework and its application to RFe12-type magnet compounds. *Phys. Rev. Mater.* **3**, 053807 (2019).
8. Tamura, R. et al. Machine learning-driven optimization in powder manufacturing of Ni-Co based superalloy. *Mater. Des.* **198**, 109290 (2021).
9. Tamura, R. et al. Automatic Rietveld refinement by robotic process automation with RIETAN-FP. *Sci. Technol. Adv. Mater. Methods* **2**, 435–444 (2022).
10. Li, J. et al. Feature selection: A data perspective. *ACM Comput. Surv. (CSUR)* **50**(6), 1–45 (2017).
11. Janet, J. P. et al. Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships. *J. Phys. Chem. A* **121**(46), 8939–8954 (2017).
12. Balachandran, P. V. et al. Importance of feature selection in machine learning and adaptive design for materials. in *Materials discovery and design: By means of data science and optimal learning* 59–79 (Springer, New York, 2018).
13. Johnson, N. L. Sequential analysis: A survey. *J. R Stat. Soc. Ser. A (Gen.)* **124**, 372–411 (1961).
14. Ford, I. & Silvey, S. D. A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika* **67**, 381–388 (1980).
15. Hino, H. Active learning: Problem settings and recent developments. Preprint at https://doi.org/10.48550/arXiv.2012.04225 (2020).
16. Settles, B. Active Learning Literature Survey.
17. Terayama, K. et al. Efficient construction method for phase diagrams using uncertainty sampling. *Phys. Rev. Mater.* **3**, 033802 (2019).
18. Terayama, K. et al. Acceleration of phase diagram construction by machine learning incorporating Gibbs' phase rule. *Scr. Mater.* **208**, 114335 (2022).

19. Tamura, R. et al. Machine-learning-based phase diagram construction for high-throughput batch experiments. *Sci. Technol. Adv. Mater. Methods* **2**, 153–161 (2022).
20. Lewis, D. D. & Catlett, J. Heterogeneous uncertainty sampling for supervised learning. in *Machine Learning Proceedings 1994* (eds. Cohen, W. W. & Hirsh, H.) 148–156 (Morgan Kaufmann, San Francisco, CA, 1994). https://doi.org/10.1016/B978-1-55860-335-6.50026-X.
21. Ueno, T. et al. Adaptive design of an X-ray magnetic circular dichroism spectroscopy experiment with Gaussian process modelling. *npj Comput. Mater.* **4**, 1–8 (2018).
22. Tian, Y. et al. Efficient estimation of material property curves and surface via active learning. *Phys. Rev. Mater.* **5**, 013802 (2021).
23. Xian, Y. et al. Compositional design of multicomponent alloys using reinforcement learning. *Acta Mater.* **274**, 120017 (2024).
24. Jose, A. et al. Regression tree-based active learning. *Data Min. Knowl. Disc.* https://doi.org/10.1007/s10618-023-00951-7 (2023).
25. Motoyama, Y. et al. Bayesian optimization package: PHYSBO. *Comput. Phys. Commun.* **278**, 108405 (2022).
26. Takahashi, A., Kumagai, Y., Miyamoto, J., Mochizuki, Y. & Oba, F. Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations. *Phys. Rev. Mater.* **4**, 103801 (2020).
27. Terayama, K. et al. Pushing property limits in materials discovery via boundless objective-free exploration. *Chem. Sci.* **11**, 5959–5968 (2020).
28. Otsuka, S. et al. PoLyInfo: Polymer database for polymeric materials design. in *2011 International Conference on Emerging Intelligent Data and Web Technologies* 22–29. https://doi.org/10.1109/EIDWT.2011.13 (2011).
29. scikit-learn: machine learning in Python — scikit-learn 0.24.1 documentation. https://scikit-learn.org/stable/.
30. Jacobs, M. H. G. & Spencer, P. J. A critical thermodynamic evaluation of the systems Si-Zn and Al-Si-Zn. *Calphad* **20**, 307–320 (1996).
31. Dreval, L. et al. Thermodynamic description and simulation of solidification microstructures in the Cu–Mg–Zn system. *J. Mater. Sci.* **56**, 10614–10639 (2021).
32. Naohiro, H., Kazuki, N., Masanori, E. & Hiroshi, O. Thermodynamic analysis of the Al-Mg-Zn ternary system. *J. Jpn. Inst. Met. Mater.* **84**, 141–150 (2020).
33. Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
34. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 1–7 (2016).
35. Deml, A. M., O'Hayre, R., Wolverton, C. & Stevanović, V. Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression. *Phys. Rev. B* **93**, 085142 (2016).
36. Lam Pham, T. et al. Machine learning reveals orbital interaction in materials. *Sci. Technol. Adv. Mater.* **18**, 756–765 (2017).
37. Choudhary, K., DeCost, B. & Tavazza, F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Phys. Rev. Mater.* **2**, 083801 (2018).
38. Irwin, J. J. et al. ZINC20—A free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).
39. Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. Doc.* **5**, 107–113 (1965).
40. RDKit. https://www.rdkit.org/.
41. PoLyInfo. https://polymer.nims.go.jp/.
42. Binois, M. & Wycoff, N. A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization. *ACM Trans. Evol. Learn. Optim.* **2**(2), 1–26 (2022).
43. Wang, Z. et al. Bayesian optimization in high dimensions via random embeddings. in *23rd International Joint Conference on Artificial Intelligence* (2013).
44. Kirschner, J. et al. Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces. arXiv preprint arXiv:1902.03229 (2019).
45. Eriksson, D. & Jankowiak, M. High-dimensional Bayesian optimization with sparse axis-aligned subspaces. Uncertainty in Artificial Intelligence. *Proc. Mach. Learn. Res.*, (2021).

Acknowledgements

This study was supported by a project subsidized by the Core Research for Evolutional Science and Technology (CREST) (Grant Number JPMJCR2234) and KAKENHI 21H01008. It was also supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) as a "Simulation- and AI-driven next-generation medicine and drug discovery" based on "Fugaku" (JPMXP1020230120), "Feasibility studies for the next-generation computing infrastructure", and Data Creation and Utilization Type Material Research and Development Project Grant Number JPMXP1122683430. The computations in the present study were performed on supercomputers at the Supercomputer Center, Institute for Solid State Physics, University of Tokyo, and the National Institute for Materials Science. We would like to thank Editage (www.editage.jp) for English language editing.

Author contributions

A.K., K.T., and R.T. conceived the idea and designed the research. A.K. conducted the investigations and calculations. G.D., K.T. and R.T. prepared the datasets. All members contributed to the preparation of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-76800-4.

Correspondence and requests for materials should be addressed to A.K., K.T. or R.T.