



**HAL**  
open science

## Machine-Learning-Based phase diagram construction for high-throughput batch experiments

Ryo Tamura, Guillaume Deffrennes, Kwangsik Han, Taichi Abe, Haruhiko Morito, Yasuyuki Nakamura, Masanobu Naito, Ryoji Katsube, Yoshitaro Nose, Kei Terayama

### ► To cite this version:

Ryo Tamura, Guillaume Deffrennes, Kwangsik Han, Taichi Abe, Haruhiko Morito, et al.. Machine-Learning-Based phase diagram construction for high-throughput batch experiments. *Science and Technology of Advanced Materials: Methods*, 2022, 2 (1), pp.153 - 161. 10.1080/27660400.2022.2076548 . hal-04728516

HAL Id: hal-04728516

<https://hal.science/hal-04728516v1>

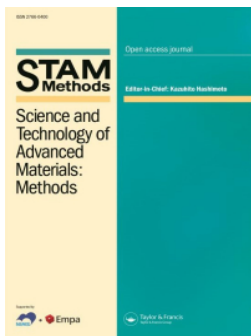
Submitted on 9 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



# Machine-Learning-Based phase diagram construction for high-throughput batch experiments

Ryo Tamura, Guillaume Deffrennes, Kwangsik Han, Taichi Abe, Haruhiko Morito, Yasuyuki Nakamura, Masanobu Naito, Ryoji Katsube, Yoshitaro Nose & Kei Terayama

To cite this article: Ryo Tamura, Guillaume Deffrennes, Kwangsik Han, Taichi Abe, Haruhiko Morito, Yasuyuki Nakamura, Masanobu Naito, Ryoji Katsube, Yoshitaro Nose & Kei Terayama (2022) Machine-Learning-Based phase diagram construction for high-throughput batch experiments, Science and Technology of Advanced Materials: Methods, 2:1, 153-161, DOI: [10.1080/27660400.2022.2076548](https://doi.org/10.1080/27660400.2022.2076548)

To link to this article: <https://doi.org/10.1080/27660400.2022.2076548>



© 2022 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group



[View supplementary material](#)



Published online: 06 Jun 2022.



[Submit your article to this journal](#)



Article views: 2472



[View related articles](#)







[View Crossmark data](#)



Citing articles: 4 [View citing articles](#)

# Machine-Learning-Based phase diagram construction for high-throughput batch experiments

Ryo Tamura <sup>a,b,c</sup>, Guillaume Deffrennes <sup>a</sup>, Kwangsik Han<sup>d</sup>, Taichi Abe <sup>b,d</sup>, Haruhiko Morito<sup>e</sup>, Yasuyuki Nakamura<sup>b</sup>, Masanobu Naito <sup>b</sup>, Ryoji Katsube<sup>f</sup>, Yoshitaro Nose<sup>f</sup> and Kei Terayama<sup>g</sup>

<sup>a</sup>International Center for Materials Nanoarchitectonics (WPI-MANA), National Institute for Materials Science, Tsukuba, Japan; <sup>b</sup>Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba, Japan; <sup>c</sup>Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan; <sup>d</sup>Research Center for Structural Materials, National Institute for Materials Science, Tsukuba, Japan; <sup>e</sup>Institute for Materials Research, Tohoku University, Sendai, Japan; <sup>f</sup>Department of Materials Science and Engineering, Kyoto University, Kyoto, Japan; <sup>g</sup>Graduate School of Medical Life Science, Yokohama City University, Kanagawa, Japan

## ABSTRACT

To know phase diagrams is a time saving approach for developing novel materials. To efficiently construct phase diagrams, a machine learning technique was developed using uncertainty sampling, which is called as PDC (Phase Diagram Construction) package [K. Terayama et al. Phys. Rev. Mater. 3, 033802 (2019)]. In this method, the most uncertain point in the phase diagram was suggested as the next experimental condition. However, owing to recent progress in lab automation techniques and robotics, high-throughput batch experiments can be performed. To benefit from such a high-throughput nature, multiple conditions must be selected simultaneously to effectively construct a phase diagram using a machine learning technique. In this study, we consider some strategies to do so, and their performances were compared when exploring ternary isothermal sections (two-dimensional) and temperature-dependent ternary phase diagrams (three-dimensional). We show that even if the suggestions are explored several instead of one at a time, the performance did not change drastically. Thus, we conclude that PDC with multiple suggestions is suitable for high-throughput batch experiments and can be expected to play an active role in next-generation automated material development.

## ARTICLE HISTORY

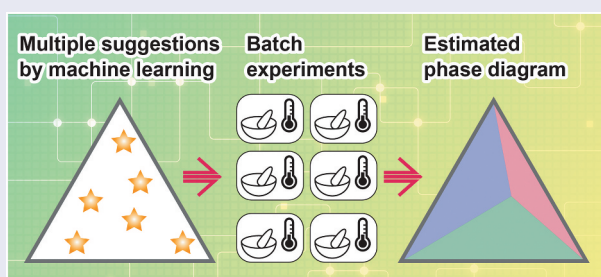
Received 7 March 2022

Revised 23 April 2022

Accepted 8 May 2022

## KEYWORDS





Phase diagram; machine learning; high-throughput batch experiments; lab automation




## 1. Introduction

Phase diagrams provide valuable guidelines for developing novel materials. Through experiments, phase diagrams have been determined in various spaces, such as between processes, external fields, and compositions. In the materials informatics field [1–5], some tools focusing on phase diagrams have been developed using machine learning (ML) [6–9]. Among these tools, we developed and released an ML method based on uncertainty sampling (US) [10] to efficiently construct phase diagrams, which is called PDC (Phase Diagram Construction) package [11,12]. The method evaluates the uncertainty based on an ML model trained under

some conditions in which the phase domains have already been identified; it then suggests the most uncertain point in the phase diagram as the next experimental condition. Based on this suggestion, an experiment is performed, and the corresponding phase domain is identified. Subsequently, the number of known points is increased, and the next condition is suggested using the updated ML model. By applying PDC to well-known phase diagrams, we demonstrated that this iteration process enables us to rapidly construct accurate phase diagrams and find new phase domains [11]. For further verification, PDC was applied to construct a new diagram to obtain Zn-Sn-P films by molecular

**CONTACT** Ryo Tamura  [tamura.ryo@nims.go.jp](mailto:tamura.ryo@nims.go.jp)  International Center for Materials Nanoarchitectonics (WPI-MANA), National Institute for Materials Science, 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan; Kei Terayama  [terayama@yokohama-cu.ac.jp](mailto:terayama@yokohama-cu.ac.jp)  Graduate School of Medical Life Science, Yokohama City University, 1-7-29, Suehiro-cho, Tsurumi-ku, Kanagawa 230-0045, Japan

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/27660400.2022.2076548>

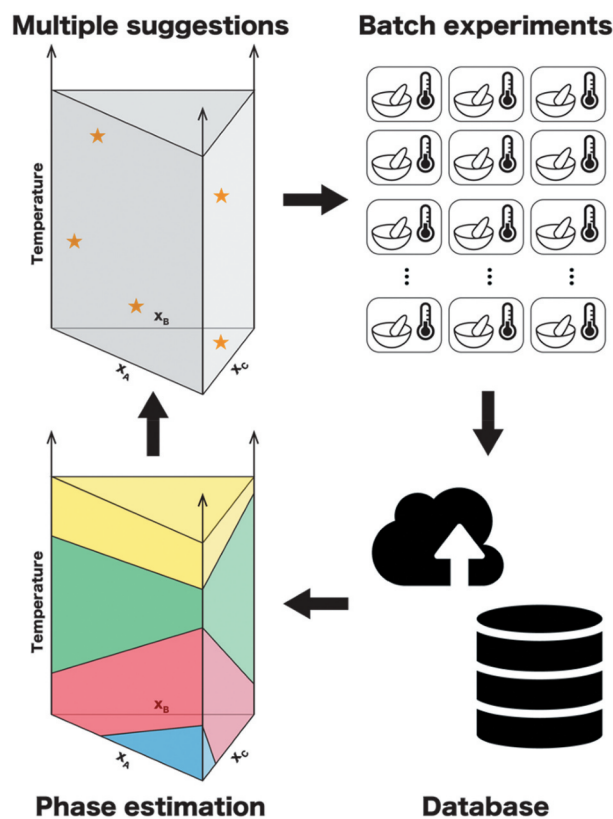
© 2022 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

beam epitaxy [13]. In this study, PDC-assisted experiments detected a new phase that was not observed in the initial stage, and detailed phase boundaries were rapidly determined. Thus, the usefulness of PDC was proven in the case in which one condition is suggested by ML and one experiment is performed in each iteration. In addition, by incorporating Gibbs' phase rule, further acceleration can be realized [14].

Several systems for high-throughput batch experiments, where many experiments can be simultaneously performed, have been developed and many successful results have been reported [15–17]. ML has also been utilized to efficiently characterize the results obtained from high-throughput batch experiments [18–20]. Furthermore, high-throughput batch experiments based on laboratory automation techniques and robotics have recently gained attention as a method for developing innovative materials. Therefore, a combination of high-throughput batch experiments and ML-based methods to suggest the next experimental conditions is required for next-generation automated materials development. To achieve this, multiple conditions must be simultaneously proposed using the ML model [21,22]. For example, to optimize material properties, batch suggestions are obtained using Bayesian optimization tools, such as GPyOpt, COMBO, and PHYSBO packages [23–25]. The use of multiple suggestions can reduce the total number of iterations, reducing the time required for material development. Conversely, the total number of experiments should be increased in general, which corresponds to the cost of material development [25,26]. Thus, multiple suggestions by the ML model do not always yield better performance. In particular, if the cost of a single experiment is high, the total cost would become unrealistic when multiple conditions are suggested by Bayesian optimization.

In this study, we investigated the efficiency of the PDC package when multiple suggestions were explored at a time (see Figure 1). Based on practical considerations, strategies to select multiple conditions were considered, and their performances were compared. Here, the target was a Cu-Mg-Zn ternary alloy. Both isothermal sections of this system (two-dimensional) and its temperature-dependent phase diagram as a whole (three-dimensional) were considered. We show that PDC is suitable for high-throughput batch experiments to efficiently construct phase diagrams. The development of autonomous experimental systems for alloy systems, which are the subject of this study, is in progress [27,28]. To accomplish this objective, in addition to the development of ML-based suggestion methods, synthesis and characterization must be automated, and these attempts are being actively investigated [29–33]. Our proposed method can be commonly implemented in such autonomous systems as 'brain'.



**Figure 1.** Iteration cycles to construct a phase diagram by high-throughput batch experiments with ML. The key to effective construction is an appropriate suggestion of multiple experiments in each cycle.

## 2. PDC with multiple suggestions

Before considering multiple suggestions at each iteration, the conventional PDC, in which one condition is suggested at a time, is introduced in Sec. 2.1. In Sec. 2.2, four strategies for multiple suggestions are considered.

### 2.1. Iterations to consider one suggestion using conventional PDC

In PDC, the following iterations are performed to construct a phase diagram.

- (i) We discretize the phase diagram to be determined, and the position vectors of  $N$  discretized points are prepared as  $\{\mathbf{x}_i\}_{i=1,\dots,N}$ . Here, vector  $\mathbf{x}_i$  can have any dimension corresponding to the dimension of the phase diagram.
- (ii) The initial data points are randomly selected from  $\{\mathbf{x}_i\}_{i=1,\dots,N}$ , and for each point, the phase domain is identified through experiments. The label of the detected phase domain is  $p = 1, \dots, P$  if  $P$  types of phase domains are identified in the initial step. Thus, the initial dataset  $D = \{\mathbf{x}_j, p_j\}_{j=1,\dots,M}$  is prepared, in which the number of initial data points is  $M$ .

- (iii) The probability distribution  $P(p|\mathbf{x})$  of the phase domain labeled by  $p$  at each point  $\mathbf{x}$  estimated by a semi-supervised learning method using training data  $D$ . Here, label propagation (LP) [34] or label spreading (LS) [35] methods are useful for evaluating  $P(p|\mathbf{x})$ .
- (iv) The uncertainty score  $u(\mathbf{x})$  is calculated using  $P(p|\mathbf{x})$ . Three types of scores [11] are implemented in the PDC package:

Least Confident(LC) :

$$u(\mathbf{x}) = 1 - \max_p P(p|\mathbf{x}) \tag{1}$$

Margin Sampling(MS) :

$$u(\mathbf{x}) = 1 - [P(p_1|\mathbf{x}) - P(p_2|\mathbf{x})] \tag{2}$$

Entropy – based Approach(EA) :

$$u(\mathbf{x}) = 1 - \sum_p P(p|\mathbf{x}) \log P(p|\mathbf{x}) \tag{3}$$

where  $P(p_1|\mathbf{x})$  and  $P(p_2|\mathbf{x})$  in Equation (2) are the highest and second-highest probabilities, respectively, at  $\mathbf{x}$ .

- (v) The most uncertain point where  $\mathbf{x}^* = \text{argmax}_{\mathbf{x}} u(\mathbf{x})$  is selected as the next experimental condition.
- (vi) The experiment is performed according to the selected conditions, and the phases are identified. If a new phase is detected, a new label,  $P + 1$ , is attached to  $\mathbf{x}^*$ . Then, the training data are increased as  $D = \{\mathbf{x}_j, p_j\}_{j=1, \dots, M+1}$ .
- (vii) Steps (iii)-(vi) where one condition is suggested by ML are iterated.

## 2.2. Multiple suggestions

We consider four types of strategies to select multiple conditions instead of step (v) in the conventional PDC, referring to the senses of experimentalists as follows. Here, we select  $L$  experimental conditions in one cycle, and the suggested points are  $\{\mathbf{x}_l^*\}_{l=1, \dots, L}$ . The features and disadvantages of each strategy are summarized in Table 1.

### 2.2.1. Only US ranking

The most straightforward approach involves selecting  $L$  experimental conditions in decreasing the order of the uncertainty score  $u(\mathbf{x})$ . We refer to this strategy as the *only US ranking*.

### 2.2.2. Neighbor exclusion

If the proposed points are too close to each other, the process would be inefficient because experiments for neighboring conditions are likely to be irrelevant. Thus, we consider the selection of the next set of experimental conditions such that the proposed points are not too close to each other. In other words, the experimental conditions are chosen in decreasing order of the uncertainty score, but any neighboring points are excluded. In this strategy, a hyperparameter exists that determines the extent to which two points are considered neighboring. Here, we use the distance between the  $k$ th nearest neighbor points in  $\{\mathbf{x}_i\}_{i=1, \dots, N}$ , which is defined as  $\Delta_k = \|\mathbf{x}_i - \mathbf{x}_j\|_{j=k\text{th neighbor of } i}$ . Thus, depending on the value of  $k$ , the suggested points are  $\{\mathbf{x}_l^*\}_{l=1, \dots, L}$  with  $\mathbf{x}_l^* - \mathbf{x}_{l'}^* > \Delta_k$  for  $l, l' = 1, \dots, L$ . We refer to this strategy as *neighbor exclusion*.

### 2.2.3. Same combination exclusion

Selecting numerous experimental conditions near the same predicted phase boundary or invariant point may be inefficient. To avoid such selection, the probability distribution  $P(p|\mathbf{x})$  at each point is useful. If several suggestions have the same combination of phase domains with a large probability,  $P(p|\mathbf{x})$ , it is sufficient to select the experimental condition with the highest uncertainty score from these points. For example, let us consider the case in which only one experiment is selected from all suggestions associated with the same three most likely phase domains. At each point, the three phase domains with the highest probability are denoted as  $p_1, p_2$ , and  $p_3$ . Then, the experimental conditions are selected in decreasing order of uncertainty score  $u(\mathbf{x})$  such that each suggestion has a unique combination of  $[p_1, p_2, p_3]$ . This strategy helps realize the selection in which we do not chose more than one experimental condition near the same predicted invariant point. In this study,

**Table 1.** Features and disadvantages of four types of strategies to select multiple conditions.

	Only US ranking	Neighbor exclusion	Same combination exclusion	Proactive selection
Selection	Points with higher $u(\mathbf{x})$	Points with higher $u(\mathbf{x})$ except for neighbor points	Points with higher $u(\mathbf{x})$ except for around the same predicted boundaries	Predicted proactive uncertain points
Calculation time	Short	Short	Long	Long
Hyperparameters	None	Distance determining neighbor points	Number of phases determining boundaries	Number of proactive steps
Disadvantage	Neighbor points may be suggested.	None	The number of proposals may not be sufficient.	The number of proposals may not be sufficient.

**Table 2.** Numbers of phase domains and discretized points for each target of the Cu-Mg-Zn ternary system.

	Number of domains	Number of discretized points
Isothermal section at 500 K	59	231
Isothermal section at 900 K	30	231
Temperature dependent ternary phase diagram	138	4851

since the target phase diagrams are two- or three-dimensional, we consider the top three phase domains for combination to exclude the suggestions, although this number of phases is arbitrary. Note that the number of combinations is limited; therefore, it may not be possible to select  $L$  points using this strategy. We call this strategy as the *same combination exclusion*.

#### 2.2.4. Proactive selection

Using a possible phase domain predicted by ML as a label for the selected condition with the highest uncertainty score, we can proceed to the next step in the PDC iterations. In the next step, we obtain another experimental condition suggested by ML, depending on the phase domain label used in the previous step. By repeating this procedure, the number of suggested conditions can be increased. However, the points that are useful among these conditions should be considered. These can be selected using the probability distribution. These can be selected using the probability distribution  $P(P|\mathbf{x})$ . In the first step, the value of  $P(P|\mathbf{x})$  can be regarded as the probability that label  $P$  is chosen under the selected condition. In the second step, we can estimate the next probability distribution by ML; however, since we need to consider the probability of the first step, the product of the probabilities of the first and second steps will be useful for the next selected condition in the second step. In further steps, the product of probabilities is evaluated for each suggested condition. The experimental conditions are then selected in decreasing order of products of the probabilities. However, if the number of proactive steps is large, the computation time will drastically increase; therefore, we decided to consider up to three steps ahead. In addition, only the top three phase domains were considered as the possible labels for selected point. Therefore, notably, cases in which  $L$  points cannot be selected can exist. We refer to this strategy as the *proactive selection*.

### 3. Target dataset

To investigate the performance of PDC when multiple suggestions are explored simultaneously, CALPHAD calculations of the Cu-Mg-Zn phase diagram were performed based on the description in Ref. [36]. Phase equilibria data were retrieved between 500

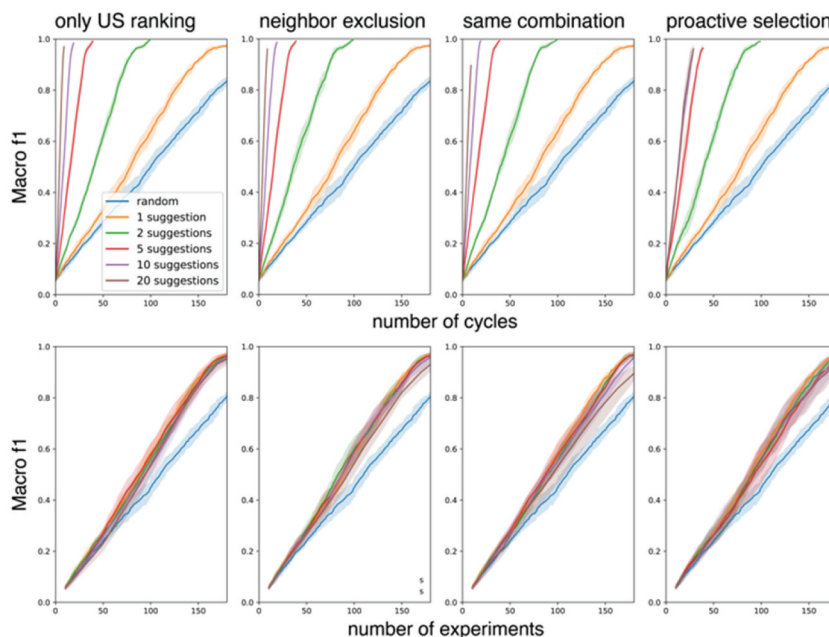
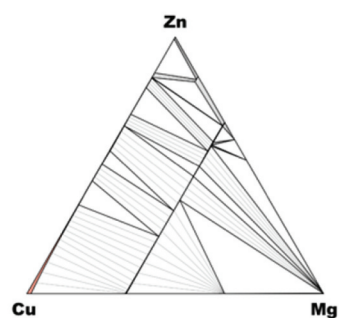
K and 1500 K with steps of 5 at.% and 50 K using the high-throughput calculation function of the Pandat software [37]. Regardless of the number of coexisting phases, a label was created for each unique phase domain. We focused on three different targets: the 500 K and 900 K isothermal sections (two-dimensional) and a temperature-dependent phase diagram as a whole (three-dimensional). The numbers of phase domains and discretized points for each target are summarized in the Table 2.

### 4. Results

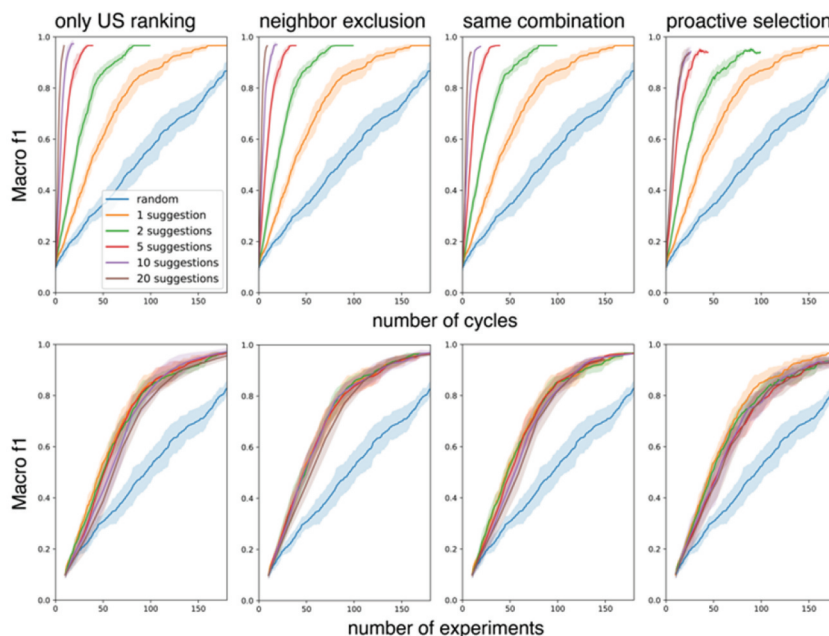
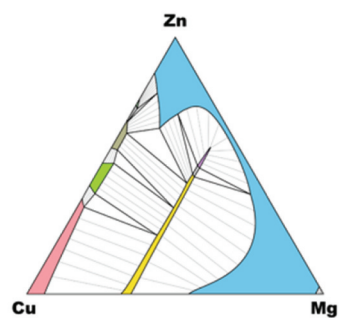
In this study, we used the Macro f1 score [38] to evaluate the efficiency of the proposed method. This score evaluates the agreement between the true phase diagram and the phase diagram predicted by ML when the training data available at that time are used. When the Macro f1 score becomes 1, both phase diagrams are perfectly equivalent. Thus, if we can obtain a large Macro f1 value with a small number of cycles and experiments, it implies that PDC can be efficiently used to construct a phase diagram. As the results depend on the data selected initially, 10 independent iterations with different randomly prepared initial data were performed, and the Macro f1 scores for each iteration were averaged. The number of randomly selected initial data was fixed at 10. As a typical example, LP was used for the ML method to calculate the probability distributions, and LC was used to evaluate the uncertainty score. In the Supplementary Material, some results from other ML models and uncertainty scores are summarized.

In Figure 2, we show the results for ternary isothermal sections, depending on the strategies for suggestions explained in Sec. 2.2. Different numbers of suggestions to be explored in each cycle were considered, and a case in which a single suggestion is randomly selected was presented for comparison. For *neighbor exclusion*, the value of  $k$  was fixed at one. From the viewpoint of the number of cycles (upper panels in Figure 2), this can be reduced by increasing the number of suggestions, regardless of which of the four methods is used. In the case of *proactive selection*, even if the number of suggestions was set to 20, there were cases in which 20 suggestions were not obtained, and the result was almost the same as that in the case of 10 suggestions. When several PDC suggestions were explored simultaneously instead of one at a time, the number of cycles (i.e. the time) required to construct an accurate phase diagram could be reduced.

(a) 500 K



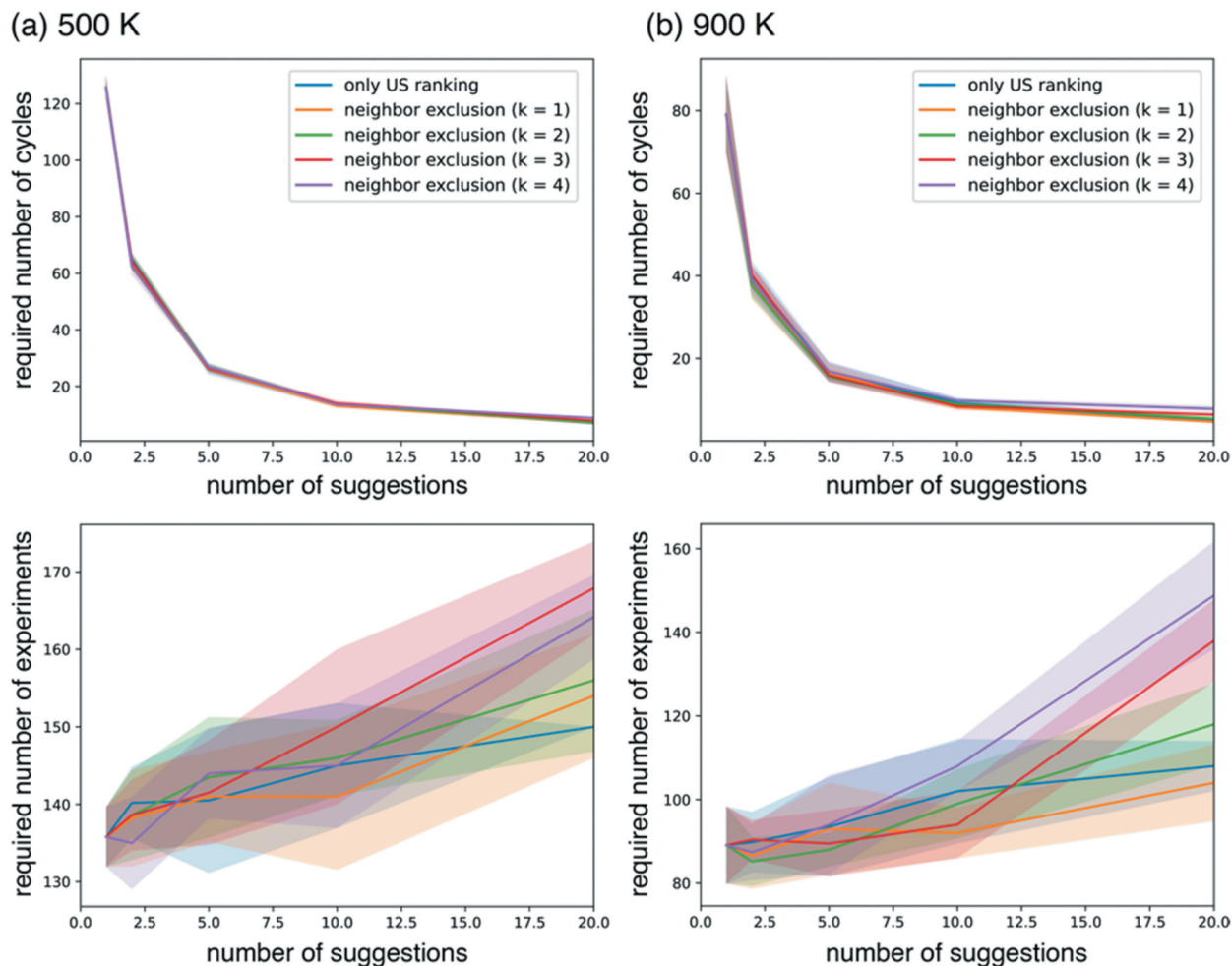
(b) 900 K



**Figure 2.** Target isothermal sections of the Cu-Mg-Zn ternary system and Macro f1 scores depending on the numbers of cycles and experiments for (a) 500 K and (b) 900 K. The four types of multiple-suggestion techniques (*only US ranking*, *neighbor exclusion* with  $k = 1$ , *same combination*, and *proactive selection*) with different numbers of suggestions at each cycle are compared with random selection. Training was performed by LP, and the uncertainty score was evaluated by LC. Ten independent runs were performed, and the line and shaded region represent the mean and error, respectively.

However, the total number of experiments required to obtain an accurate phase diagram is an important factor when multiple conditions are suggested. The results are shown in the lower panels of Figure 2. Interestingly, the number of experiments did not significantly increase, even when the number of suggestions considered in each cycle increased. In particular, when using *neighbor exclusion*, the results with one suggestion and two, five, and ten suggestions were within the error bars for both the 500 K and 900 K sections. Furthermore, the results of the *same combination*

*exclusion* and *proactive selection* were less satisfying than those of the other two methods (see Figure S1), despite the long calculation time required for selection. The Macro f1 scores when different ML models and uncertainty scores were used, depending on the number of cycles and experiments, are summarized in Figures S2-S6. These results imply that LP+LC is useful for both multiple-suggestions and single-suggestion cases [11]. However, the results of PDC with the four types of multiple-suggestion strategies remain better than those of random selection.



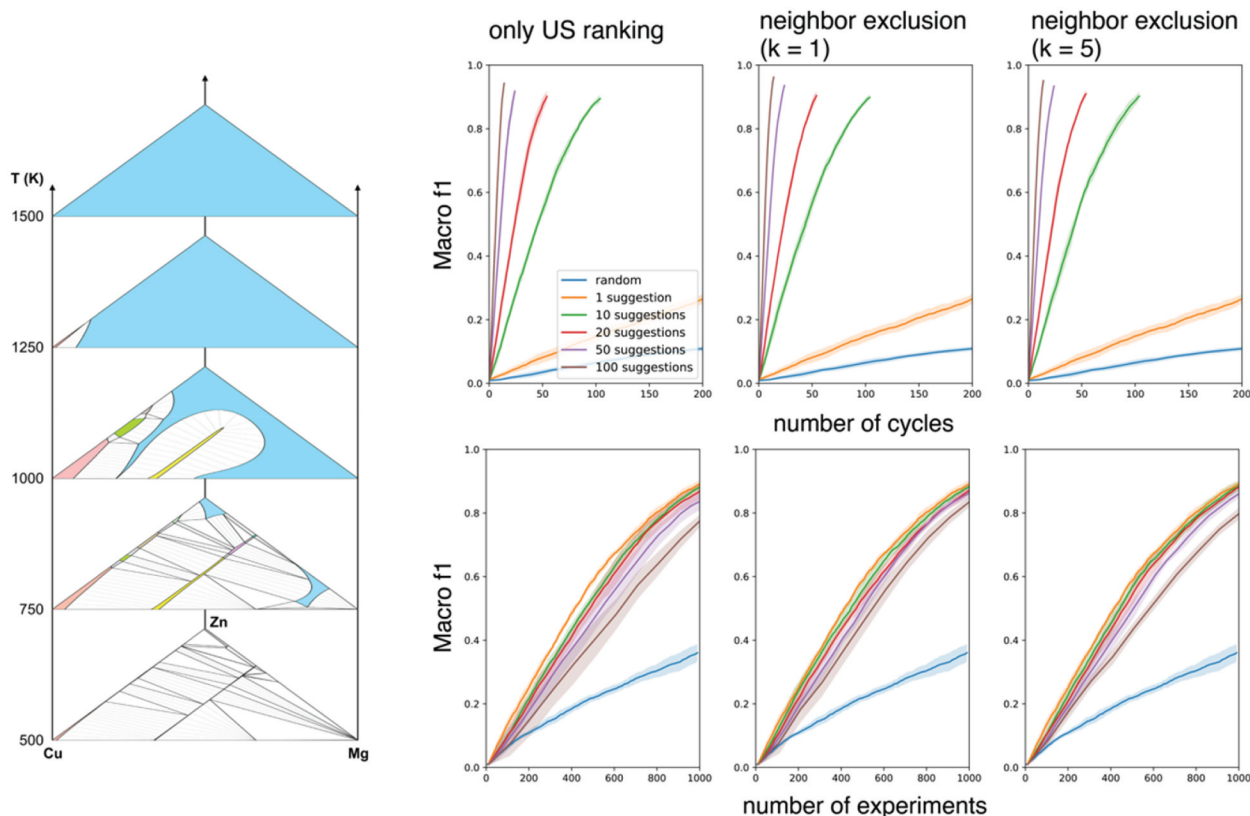
**Figure 3.** Number of cycles and number of experiments required to obtain a value of 0.8 for the Macro f1 score depending on the number of suggestions for isothermal ternary phase diagrams at (a) 500 K and (b) 900 K. The selection strategies of *only US ranking* and *neighbor exclusion* with  $k = 1, 2, 3,$  and  $4$  are used.

Next, we considered the dependence of hyperparameter  $k$  in *neighbor exclusion*. Figure 3 summarizes the number of cycles and the number of experiments required to obtain the value of 0.8 for a Macro f1 score depending on the value of  $k$  in *neighbor exclusion*. As a reference, the result obtained using the *only US ranking* strategy is shown. The number of cycles steadily decreased as the number of suggestions increased in all cases, whereas the required number of experiments slightly increased with the number of suggestions. However, up to approximately ten suggestions, the results with  $k = 1$  at 500 K and  $k = 1, 2, 3$  at 900 K were within the error bar when the number of suggestions was only one (the left-most point). This result supports the idea that multiple suggestions by *neighbor exclusion* are effective in high-throughput batch experiments. The reason why the performance of *neighbor exclusion* is better at 900 K than at 500 K is that the areas of each phase domain are relatively large.

For the 500 K section, nearby candidate points belong to different phase domains because of the existence of many small phase domains, and large  $k$  cases are not inefficient.

To address the case of a three-dimensional phase diagram, the next target is the temperature-dependent phase diagram of the Cu-Mg-Zn ternary system. In Figure 4, we show the Macro f1 score depending on the number of suggestions explored at each cycle based on *only US ranking* and *neighbor exclusion* with  $k = 1$  and  $5$ . Even in the three-dimensional space, PDC reduces the number of cycles as the number of suggestions increases. In addition, regarding the number of experiments, the results with one, ten, and twenty suggestions showed better agreement when  $k = 5$ . Thus, PDC with multiple suggestions is useful in constructing a three-dimensional phase diagram. Furthermore, Figure 5 summarizes the number of cycles and experiments required to obtain a Macro f1 score of 0.8. In this case, as in the case of the





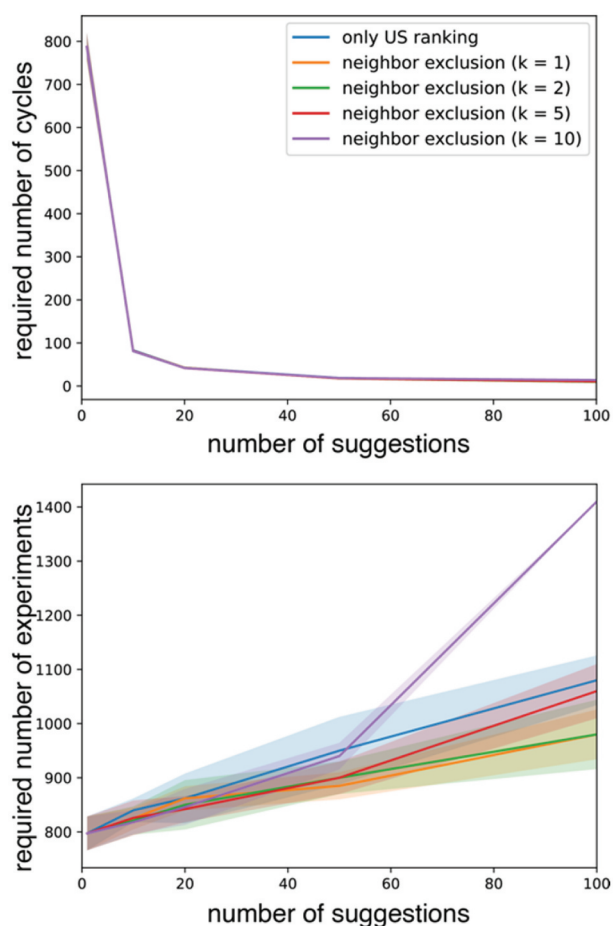
**Figure 4.** Target temperature-dependent Cu-Mg-Zn phase diagram and Macro f1 scores depending on the numbers of cycles and experiments. The two types of multiple-suggestion techniques (*only US ranking* and *neighbor exclusion* with  $k = 1$  and 5) with different numbers of suggestions are compared with random selection. Training was performed by LP, and the uncertainty score was evaluated by LC. Ten independent runs were performed, and the line and shaded region show the mean and error, respectively.

two-dimensional sections, the results for up to ten suggestions are included in the error bars of the single-suggestion case, and twenty suggestions show better efficiency. From another viewpoint, *neighbor exclusion* with  $k = 5$  shows better results up to approximately fifty suggestions, but when the number of suggestions reaches 100, the results become worse. This indicates that if the distance for exclusion is too large, exclusion itself is prioritized over the ranking of the US score, and we cannot obtain effective suggestions. For the same reason,  $k = 10$  was sufficiently large for this problem setting. However, *neighbor exclusion* shows better results than *only US ranking*, and thus *neighbor exclusion* is an effective multiple-suggestion strategy if  $k$  is determined appropriately. Note that a thousand experiments would be impractical for the construction of phase diagrams. To avoid this, discretization must be devised to reduce the number of discretized candidate points. Furthermore, by incorporating Gibbs' phase rule, we can further reduce the number of experiments [14].

An appropriate value of  $k$  strongly depends on the target phase diagram (i.e, the dimension and number of phase domains) and the discretization method. This depends on the number of suggestions. However, based on the results of this study,  $k = 1$  or 2 would be sufficient when we discretize one parameter into approximately twenty parts and the number of suggestions is smaller than twenty. If finer discretization is to be considered, it would be better to increase the value of  $k$  (see Figure S7).

### 5. Discussion and summary

For high-throughput batch experiments, to construct a phase diagram with multiple suggestions using ML, we addressed the efficiency of the US method called PDC. Based on practical considerations, four strategies for selecting multiple conditions were considered. We showed that the number of cycles corresponding to the time needed to accurately determine a phase diagram can be reduced by exploring multiple suggestions for



**Figure 5.** Number of cycles and the number of experiments required to obtain a Macro f1 score of 0.8 depending on the number of suggestions for the temperature-dependent phase diagram. The selection strategies of *only US ranking* and *neighbor exclusion* with  $k = 1, 2, 5,$  and  $10$  are used.

each cycle for all the considered strategies. Furthermore, even if the number of suggestions for each cycle was increased, the total number of experiments did not drastically change. This result implies that PDC with multiple suggestions is suitable for high-throughput batch experiments to efficiently construct phase diagrams. In particular, among the four methods considered, we conclude that the strategy in which the proposed points do not include the neighboring points is useful. This neighbor exclusion strategy was used to create a temperature-composition phase diagram of a crosslinked polymer, and its usefulness was confirmed [39]. This work on polymers highlights that PDC can be applied to a variety of materials and is not limited to metals that are the focus of the present study. Furthermore, PDC can also be utilized to determine the boundaries of a region within a material search space or a process space, even when it is not a ‘phase’. Thus, we updated the PDC package to suggest multiple conditions using two strategies called *only US ranking* and *neighbor exclusion* [40].

By combining PDC with lab automation techniques and a robotics system, accurate phase diagrams can be automatically determined at a high speed and low cost. Thus, we believe that PDC can be expected to play an active role in next-generation automated material development using high-throughput batch experiments.

## Disclosure statement

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

## Funding

This study was supported by a project subsidized by the Core Research for Evolutional Science and Technology (CREST) (Grant Numbers JPMJCR17J2, JPMJCR19J1, JPMJCR19J3) from the Japan Science and Technology Agency (JST).

## ORCID

Ryo Tamura  <http://orcid.org/0000-0002-0349-358X>

Guillaume Deffrennes  <http://orcid.org/0000-0002-3752-2537>

Taichi Abe  <http://orcid.org/0000-0002-5065-0939>

Masanobu Naito  <http://orcid.org/0000-0001-7198-819X>

## References

- [1] Rajan K. Materials informatics. *Mater Today*. 2005;8(10):38–45.
- [2] Pilia G, Wang C, Jiang X, et al. Accelerating materials property predictions using machine learning. *Sci Rep*. 2013;3:1–6.
- [3] 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.
- [4] Terayama K, Sumita M, Tamura R, et al. Black-box optimization for automated discovery. *Acc Chem Res*. 2021;54:1334–1346.
- [5] Tamura R, Takei Y, Imai S, et al. Experimental design for the highly accurate prediction of material properties using descriptors obtained by measurement. *Sci Tech Adv Mater Methods*. 2021;1:152–161.
- [6] Nguyen P, Tran T, Gupta S, et al. Incomplete conditional density estimation for fast materials discovery. In: Proceedings of the 2019 SIAM International Conference on Data Mining (SDM). Society for Industrial and Applied Mathematics; 2019; p. 549–557.
- [7] Dai C, Glotzer SC. Efficient phase diagram sampling by active learning. *J Phys Chem B*. 2020;124:1275–1284.
- [8] Aghaaminiha M, Ghanadian SA, Ahmadi E, et al. A machine learning approach to estimation of phase diagrams for three-component lipid mixtures. *Biochim Biophys Acta - Biomembr*. 2020;1862:183350.
- [9] Tian Y, Yuan R, Xue D, et al. Determining multi-component phase diagrams with desired characteristics using active learning. *Adv Sci*. 2021;8:2003165.

- [10] Settles B. Active learning. *Synth Lect Artif Intell Mach Learn*. 2012;6:1–114.
- [11] Terayama K, Tamura R, Nose Y, et al. Efficient construction method for phase diagrams using uncertainty sampling. *Phys Rev Mater*. 2019;3:033802.
- [12] 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.
- [13] 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.
- [14] Terayama K, Han K, Katsube R, et al. Acceleration of phase diagram construction by machine learning incorporating Gibbs' phase rule. *Scr Mater*. 2022;208:114335.
- [15] Potyrailo RA, Takeuchi I. Combinatorial and high-throughput materials research. *Meas Sci Technol*. 2004;16:E01.
- [16] Shevlin M. Practical high-throughput experimentation for chemists. *ACS Med Chem Lett*. 2017;8:601–607.
- [17] Liu Y, Hu Z, Suo Z, et al. High-throughput experiments facilitate materials innovation: a review. *Sci China Technol Sci*. 2019;62:521–545.
- [18] Kusne AG, Gao T, Mehta A, et al. On-The-Fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci Rep*. 2014;4:6367.
- [19] Ludwig A. Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. *npj Comput Mater*. 2019;5:1–7.
- [20] Li X, Maffettone MP, Che Y, et al. Combining machine learning and high-throughput experimentation to discover photocatalytically active organic molecules. *Chem Sci*. 2021;12:10742–10754.
- [21] Couperthwaite R, Molkeri A, Khatamsaz D, et al. Materials design through batch Bayesian optimization with multisource information fusion. *JOM*. 2020;72:4431–4443.
- [22] Liang Q, Gongora AE, Ren Z, et al. Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. *npj Comput Mater*. 2021;7:1–10.
- [23] <http://sheffieldml.github.io/GPyOpt/>
- [24] Ueno T, Rhone TD, Hou Z, et al. COMBO: an efficient Bayesian optimization library for materials science. *Mater Discovery*. 2016;4:18–21.
- [25] Motoyama Y, Tamura R, Yoshimi K, et al. Bayesian optimization package: PHYSBO. arXiv:211007900.
- [26] Wang Z, Li C, Jegelka S, et al. Batched high-dimensional Bayesian optimization via structural kernel learning. arXiv:170301973.
- [27] Boyce BL, Uchic MD. Progress toward autonomous experimental systems for alloy development. *MRS Bull*. 2019;44:273–280.
- [28] Hart GLW, Mueller T, Toher C, et al. Machine learning for alloys. *Nat Rev Mater*. 2021;6:730–755.
- [29] Spowart JE, Mullens HE, Puchala BT. Collecting and analyzing microstructures in three dimensions: a fully automated approach. *JOM*. 2003;55:35–37.
- [30] Springer H, Raabe D. Rapid alloy prototyping: compositional and thermo-mechanical high throughput bulk combinatorial design of structural materials based on the example of 30mn–1.2c–xal triplex steels. *Acta Materialia*. 2012;60:4950–4959.
- [31] Flynn JM, Shokrani A, Newman ST, et al. Hybrid additive and subtractive machine tools – research and industrial developments. *Int J Mach Tools Manuf*. 2016;101:79–101.
- [32] Azimi SM, Britz D, Engstler M, et al. Advanced steel microstructural classification by deep learning methods. *Sci Rep*. 2018;8:2128.
- [33] Yang L, Haber JA, Armstrong Z, et al. Discovery of complex oxides via automated experiments and data science. *Proc Nat Acad Sci*. 2021;118:e2106042118.
- [34] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. *ICML*. 2003.
- [35] Zhou D, Bousquet O, Lal T, et al. Learning with local and global consistency. *NIPS*. 2003.
- [36] Dreval L, Zeng Y, Dovbenko O, et al. Thermodynamic description and simulation of solidification microstructures in the Cu–Mg–Zn system. *J Mater Sci*. 2021;56:10614–10639.
- [37] Cao W, Chen S-L, Zhang F, et al. PANDAT software with PanEngine, PanOptimizer and PanPrecipitation for multi-component phase diagram calculation and materials property simulation. *Calphad*. 2009;33:328–342.
- [38] [https://scikit-learn/stable/modules/generated/sklearn.metrics.f1\\_score.html](https://scikit-learn/stable/modules/generated/sklearn.metrics.f1_score.html)
- [39] Hu WH, Chen T-T, Tamura R, et al. Topological alternation from structurally adaptable to mechanically stable crosslinked polymer. *Sci Tech Adv Mater*. 2022;23:66–75.
- [40] <https://github.com/tsudalab/PDC>.