Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2026, **5**, 1252

Bayesian diversity control for batch-based phase diagram determination

Peiheng Zou,[a] Ryo Tamura [abc] and Koji Tsuda *[abc]

Received 5th November 2025
Accepted 13th February 2026

DOI: 10.1039/d5dd00486a

rsc.li/digitaldiscovery

Machine learning methods are increasingly used in experimental design in phase diagram determination. Some methods perform batch design, where multiple points are sampled from the design space. In this case, it is essential to diversify samples to avoid performing almost identical experiments, and control the diversity level appropriately. Manual diversity control is unintuitive and may require additional trial-and-error in prior to the experiments are started. We propose a Bayesian model called determinantal point process for phase diagram construction (DPP-PDC) that can perform batch design and automatic diversity control simultaneously. Central to this model is the uncertainty-weighted determinantal point process that samples a set of points with high uncertainty under diversity control. Experiments with Cu–Mg–Zn ternary system demonstrate that DPP-PDC can actively control the sample diversity to achieve high efficiency.

1. Introduction

A phase diagram maps various phases of a material depending on theomodynamic variables such as composition, temperature and pressure.[1] Numerous phase diagrams have been drawn for alloys and compounds[2–4] and magnetic strutures.[5–7] Determining the phase diagram of a material is an essential step in materials development, but it requires a considerable number of well-designed experiments by experts with changing experimental parameters, which is cost-intensive in terms of time and human effort.

Some machine learning methods predict the entirety of a phase diagram based on available data, without requiring any experiments.[8–10] Meanwhile, active learning methods aim to guide experiments by recommending experimental parameters for the next attempt.[11–15] In formalization of active learning, a black-box function that maps a design space to an outcome space is given. In phase diagram determination, the design space is specified by the experimental parameters and the outcome space is the set of all possible phase labels. The evaluation of the black-box function at a point in the design space corresponds to phase measurement. The purpose of active learning is to estimate the black-box function with as few evaluations as possible. Given existing data, an active learning algorithm recommends a point in the design space for next evaluation. Intuitively, it is better to suggest the points close to phase boundaries to estimate the function quickly. An active learning method, PDC,[14] based on label propagation[16] and uncertainty sampling,[17] recommends the point of highest phase uncertainty. PDC has been favorably tested both in benchmarks,[14,18] and a real-world study identifying the phase diagram for Zn–Sn–P film deposition using molecular beam epitaxy.[19] In addition, web application AIPHAD[18] offers its active learning service based on PDC.

Active learning algorithms are classified into two types: single-probe and batch-based.[20] PDC is a single-probe method, where the function evaluation is done one-by-one. Batch-based methods assume that evaluations are conducted at multiple points at once. Recently, the importance of batch-based methods is increasing due to the advent of self-driving laboratories.[21] Tamura *et al.*[15] proposed several batch-based methods, but users need to determine a variable to adjust sample diversity manually in advance. Mancias *et al.*[22] proposed to take a percentage of points with maximum uncertainty derived *via* a Gaussian process and apply a *k*-medioid clustering method to sample a batch of points. In this case, the percentage determines the diversity. Fig. 1 shows the importance of diversity control schematically. Misspecification of the diversity level would lead to disastrous results.

To eliminate the need for manual diversity control, we propose a Bayesian approach[23] called DPP-PDC based on the uncertainty-weighted determinantal point process (UwDPP). Given a set of points and the kernel matrix describing closeness among them, the determinant of the kernel matrix represents their diversity. In the *k*-determinantal point process (*k*-DPP),[24] a subset of size *k* is assigned a probablity proportional to the determinant. UwDPP modifies *k*-DPP so that the points with higher phase uncertainty are more likely to be chosen. UwDPP

[a]*Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwa-no-ha, Kashiwa, Chiba 277-8561, Japan. E-mail: tsuda@k.u-tokyo.ac.jp*
[b]*Center for Basic Research on Materials, National Institute for Materials Science, Tsukuba, Ibaraki 305-0047, Japan*
[c]*RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan*

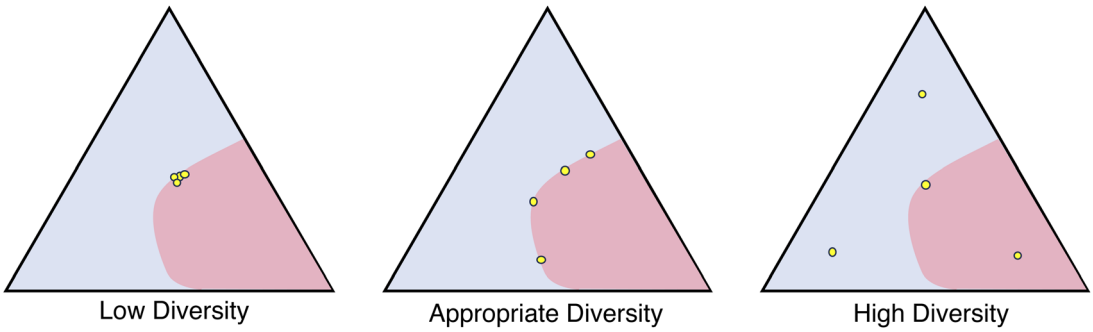

Fig. 1 Diversity control in phase diagram determination. (Left) When diversity is too low, the samples are concentrated to one point. (Middle) When diversity is properly controlled, they distribute close to phase boundaries. (Right) When diversity is too high, the samples are scattered all over the phase diagram.

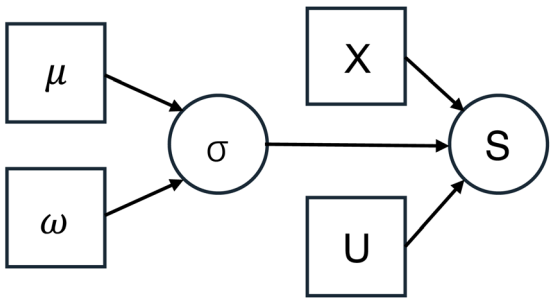

Fig. 2 Graphical model of DPP-PDC. Circular and square nodes represent random and non-random variables, respectively. The arrows from X to Y represent that the generative model of Y depends on X.

has a control variable to adjust how strongly diversity is imposed. A prior distribution is specified to the control variable, making DPP-PDC a Bayesian model. Markov chain Monte Carlo sampling[25] from the posterior distribution allows batch recommendation with automatic diversity adjustment. Note that our method is applicable to any uncertainty measure including the one used by Mancias *et al.*[22]

Using the phase diagram of Cu–Mg–Zn ternary system,[26] we demonstrate that DPP-PDC controls the sample diversity during iterations appropriately to achieve high efficiency in active learning.

2. Method

2.1. PDC

Let us denote the set of candidate points in the design space by $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. Usually, they are designated as grid points in a phase diagram. Let the set of possible phase labels be C. Assume that, for training data points $T \subseteq [1, N]$, the phases $\{y_i\}_{i \in T}$ are known. The task is to recommend a point not in T for our next evaluation. To infer the phases of the remaining points, the scikit-learn[27] implementation of the label propagation algorithm[16] was used. A fully connected graph among X is constructed first, and each edge is weighted with an RBF kernel

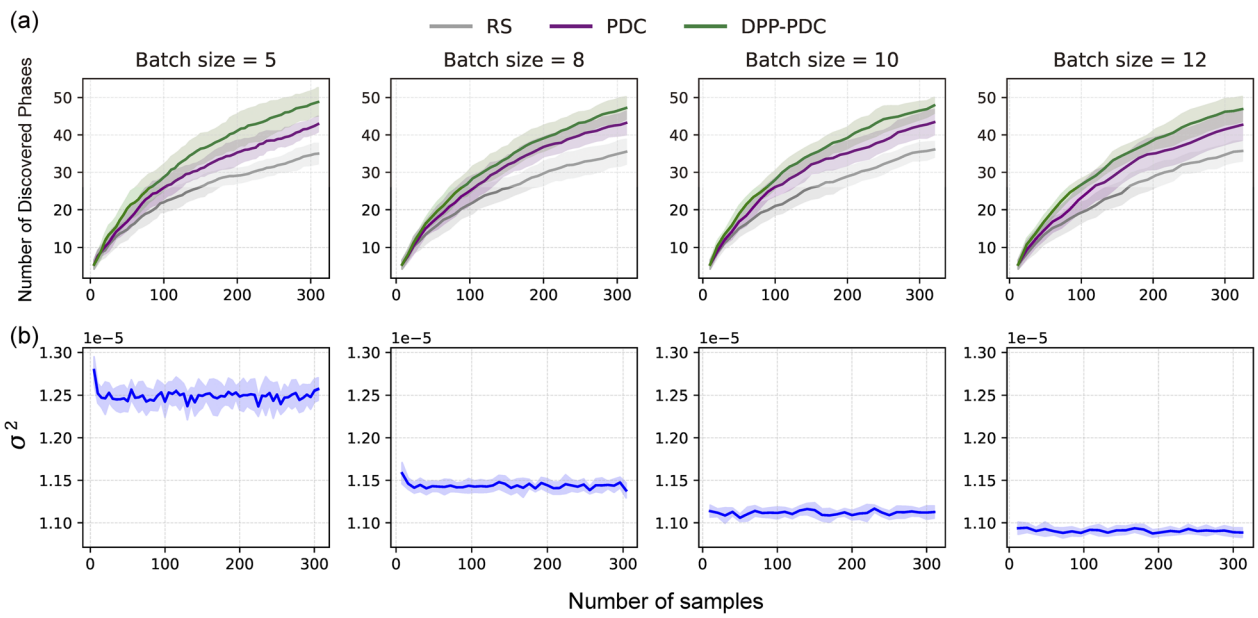

Fig. 3 (a) Phase discovery curves. The x-axis corresponds to the number of samples and the y-axis shows the number of discovered phase labels. RS stands for random sampling. (b) Evolution of control parameter σ^2.

with a lengthscale of 1/20. Based on the weighted graph, the likelihood of point j belonging to phase $c \in C$ is computed as $f_{cj} \in [0, 1]$. The uncertainty of point j is induced as

$$u_j = 1 - \max_{c \in C} f_{cj}, \tag{1}$$

and the point with maximum uncertainty $j_{\max}$ is chosen for recommendation. As a result of the experiment, we obtain a new label $y_{j_{\max}}$. The point is added as $T \leftarrow T \cup \{j_{\max}\}$ and the above procedure is repeated until the budget is met. In early iterations, it is likely that not all phase labels are included in $\{y_i\}_{i \in T}$. In that case, the uncertainty (1) is computed only with the existing labels. As iterations go on, previously unseen phase labels are discovered increasingly.

2.2. Determinantal point process

Let L is a kernel matrix among $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. Commonly, the kernel is determined as a Gaussian kernel, $L(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$, where γ determines its width. Determinantal point processes have been used for sampling a diverse subset S from X in various machine learning applications.[24,28–30] The most fundamental form of DPP is L-ensemble,

$$P^{\mathrm{DPP}}(S) = \frac{\det L_S}{\det(I + L)}, \tag{2}$$

where L_S is the submatrix of L restricted to S. For two points, it is easy to understand that L-ensemble assigns high probability to distant pairs, because

$$\det L_S = L(\boldsymbol{x}_1, \boldsymbol{x}_1)L(\boldsymbol{x}_2, \boldsymbol{x}_2) - L(\boldsymbol{x}_1, \boldsymbol{x}_2)^2.$$

In the two points are close, $L(\boldsymbol{x}_1, \boldsymbol{x}_2)$ gets large, leading to small P^{DPP}. In batch recommendation, it is convenient to have a cardinality constraint. In such a case, one can use k-DPP,[29]

$$P^{k\text{-DPP}}(S) = \frac{\det L_S}{\sum_{S' \in \mathscr{X}_k} \det L_{S'}},$$

where $\mathscr{X}_k$ is the set of all size-k subsets of X. Kathuria *et al.*[28] found it useful to have a parameter to control the diversity and proposed k-DPP with mutual information kernel:

$$P_\sigma^{k\text{-DPPMI}}(S) = \frac{\det(I + L_S/\sigma^2)}{\sum_{S' \in \mathscr{X}_k} \det\left(I + L_{S'}/\sigma^2\right)}, \tag{3}$$

where σ is a control parameter. Let $\lambda_1, \ldots, \lambda_{|S|}$ denote the eigenvalues of L_S. Then, the numerator of (3) is described as

$$\det(I + L_S/\sigma^2) = \prod_{i=1}^{|S|} (1 + \sigma^{-2}\lambda_i).$$

If σ is small and two samples in S are identical, one of the eigenvalue λ_i becomes zero and the probability of S is extremely small. In general, $P_\sigma^{k\text{-DPPMI}}(S)$ is larger when the points are farther from each other, and σ^{-2} determines the strength of the repulsive force. When σ approaches to infinity, $P^{k\text{-DPPMI}}{}_\sigma(S)$ converges to the uniform distribution, completely losing the ability to impose diversity.

2.3. DPP-PDC

Our batch recommendation method, DPP-PDC, uses the same machine learning model as PDC. Given the training data T, it provides the phase uncertainty score $U = \{u_i\}_{i=1}^N$ for all points. Instead of choosing the point of maximum phase uncertainty, we need to pick up several diverse points for batch recommendation. To this aim, k-DPP with mutual information kernel is modified such that those with high uncertainty scores are more likely to be chosen,

$$P_\sigma^{\mathrm{UwDPP}}(S) = \frac{P_\sigma^{k\text{-DPPMI}}(S) \prod_{i \in S} u_i}{\sum_{S' \in \mathscr{X}_k} P_\sigma^{k\text{-DPPMI}}\left(S'\right) \prod_{i \in S'} u_i}. \tag{4}$$

We call this distribution uncertainty-weighted DPP. To automate the choice of parameter σ, the prior distribution of σ as a log-normal distribution,

$$\log(\sigma^2) \sim \mathcal{N}(\mu, \omega),$$

where hyperparameters are fixed as follows: $\mu = -4$ and $\omega = 4$. The graphical model of this Bayesian model is described as Fig. 2.

The posterior distribution of this Bayesian model $P(S|U, \mu, \omega)$ cannot be described analytically. However, sampling S and σ is possible with a Markov chain Monte Carlo (MCMC) method.[23] In MCMC, samples are perturbed randomly and the movement is either accepted or canceled according to a certain rule. Among a plethora of MCMC algorithms, we chose Non-U-Turn sampler[31] implemented in pyMC.[25] This algorithm is particulary useful, because the step size in perturbation is automatically derived. As for S, the perturbation is done by replacing one element randomly. The sampler starts from a random point and the 1000-th sample is adopted as the final solution of S and σ.

3. Results

To test DPP-PDC, we used the phase diagram of Cu–Mg–Zn ternary system.[26] This alloy has applications in automotive and aerospace industries due to low density and exceptional strength. The diagram was calculated by CALPHAD, but confirmed to match available experimental measurements.[26] This phase diagram has three degrees of freedom, that is, fraction of Cu, fraction of Mg and the temperature. The temperature ranges from 500 K to 1500 K. Coexisting phases are given distinct labels (*e.g.*, HCP-ZN + MG2ZN11). This resulted in 71 phase labels in total. Candidate points are defined as the grid points, where 1% and 50 K intervals are adopted for the fractions and the temperature, respectively.

For different batch sizes 5, 8, 10 and 12, DPP-PDC is applied until the number of total samples reach 300. In PDC, top-k samples in terms of the uncertainty are taken as a batch. Fig. 3a shows the number of discovered phase labels against the number of samples (*i.e.*, phase discovery curves). The curve and error bar correspond to the average and standard deviation over 10 runs, respectively. Fig. 4 shows the area under the phase discovery curve for random sampling, PDC and DPP-PDC. Evaluations with respect to other performance measures are

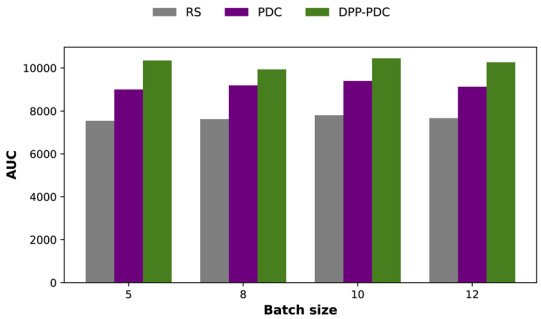

Fig. 4 Area under the phase discovery curve for random sampling (RS), PDC and DPP-PDC.

shown in Fig. S1 in the SI. As expected, DPP-PDC and PDC performed significantly better than random sampling. In all cases, DPP-PDC performed better than PDC, showing the merit of diversity control.

Fig. 3b shows the evolution of control parameter σ^2 for different batch sizes. Although σ^2 has certain variability among iterations, we can observe the clear trend that a smaller value of σ^2, therefore high diversity, is preferred as the number of batches increases. It matches our intuition that diversity must be imposed more strongly in large batch cases to avoid unwanted concentration of recommendations. This result demonstrates that DPP-PDC has an ability to control diversity appropriately and can at least avoid a catastrophic failure caused by extremely small or large diversity.

We tried different hyperparameter settings of DPP-PDC, but the results showed little difference (Fig. S2 in the SI). It is because our prior distribution is set to enforce a minimally weak constraint to allow adaption to data.

We applied the batch sampling method by Mancias *et al.*,[22] where the cut-off parameter is set to 2.5% as instructed in the paper. As shown in Fig. S3 in the SI, its performance was not as good as ours. Batch optimization methods are known to be sensitive to diversity parameter. Bayesian adaptation of the cut-off parameter to our dataset may have improved the performance, but it is out of scope of this paper.

4. Conclusion

In this paper, we have shown how Bayesian modeling can be applied to batch-based phase diagram determination. This approach is more theoretically principled than manual diversity control, and may find various applications in self driving laboratories. Non-Bayesian heuristic methods do not have systematic ways of parameter determination. Their parameters are often determined by prior trial-and-error. When data acquisition is costly as in self-driving laboratories, the cost of trial-and-error would exceed the computational cost of MCMC sampling necessary in Bayesian methods.

Experimental phase measurements may contain noise. At phase boundaries, mislabeling due to noise is likely to occur. We used computational phase diagrams only in our experiments, and did not consider experimental noise. To apply DPP-PDC in experimental phase diagrams, some countermeasures against noise may be necessary.

One drawback of our approach is that the prior distribution of σ^2 is determined in an ad-hoc manner. Given a plenty of phase diagram data, however, it could be determined by cross validation or related model selection methods. Another restriction is that the sampling points are limited to the grid points. It may be problematic when a phase diagram of higher resolution is needed. It would be an interesting future work to develop a Bayesian method for completely continuous settings.

Author contributions

R. T. and K. T. conceived the idea and designed the research. P. Z. implemented the algorithm, and performed computational experiments. All authors wrote the manuscript.

Conflicts of interest

The authors have no conflicts to disclose.

Data availability

Code and datasets used in the paper are available at Github https://github.com/tsudalab/DPP-PDC and Zenodo https://doi.org/10.5281/zenodo.18627155.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d5dd00486a.

Acknowledgements

This work was supported by JSPS Kakenhi 25K01492, JST CREST JPMJCR21O2 and ERATO JPMJER1903.

References

1 Y. A. Chang, S. Chen, F. Zhang, X. Yan, F. Xie, R. Schmid-Fetzer and W. A. Oates, Phase diagram calculation: past, present and future, *Prog. Mater. Sci.*, 2004, **49**, 313–345.
2 K. Kennedy, T. Stefansky, G. Davy, V. F. Zackay and E. R. Parker, Rapid method for determining ternary-alloy phase diagrams, *J. Appl. Phys.*, 1965, **36**, 3808–3810.
3 D. B. Miracle and O. N. Senkov, A critical review of high entropy alloys and related concepts, *Acta Mater.*, 2017, **122**, 448–511.
4 M. Enoki, S. Minamoto, I. Ohnuma, T. Abe and H. Ohtani, Current status and future scope of phase diagram studies, *ISIJ Int.*, 2023, **63**, 407–418.
5 P. Schiffer, A. Ramirez, W. Bao and S.-W. Cheong, Low temperature magnetoresistance and the magnetic phase diagram of $La_{1-x}Ca_xMnO_3$, *Phys. Rev. Lett.*, 1995, **75**, 3336.
6 G. Schmid, S. Todo, M. Troyer and A. Dorneich, Finite-temperature phase diagram of hard-core bosons in two dimensions, *Phys. Rev. Lett.*, 2002, **88**, 167208.
7 J. Reuther, R. Thomale and S. Trebst, Finite-temperature phase diagram of the heisenberg-kitaev model, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2011, **84**, 100406.

8 W. Huang, P. Martin and H. L. Zhuang, Machine-learning phase prediction of high-entropy alloys, *Acta Mater.*, 2019, **169**, 225–236.
9 C. Liu, E. Fujita, Y. Katsura, Y. Inada, A. Ishikawa, R. Tamura, K. Kimura and R. Yoshida, Machine learning to predict quasicrystals from chemical compositions, *Adv. Mater.*, 2021, **33**, 2102507.
10 Y. Oikawa, G. Deffrennes, T. Abe, R. Tamura, and K. Tsuda, "alloym: A large language model for alloy phase diagram prediction", *arXiv*, 2025, preprint, arXiv:2507.22558, DOI: 10.48550/arXiv.2507.22558.
11 C. Dai and S. C. Glotzer, Efficient phase diagram sampling by active learning, *J. Phys. Chem. B*, 2020, **124**, 1275–1284.
12 S. Ament, M. Amsler, D. R. Sutherland, M.-C. Chang, D. Guevarra, A. B. Connolly, J. M. Gregoire, M. O. Thompson, C. P. Gomes and R. B. Van Dover, Autonomous materials synthesis via hierarchical active learning of nonequilibrium phase diagrams, *Sci. Adv.*, 2021, **7**, eabg4930.
13 Y. Tian, R. Yuan, D. Xue, Y. Zhou, Y. Wang, X. Ding, J. Sun and T. Lookman, Determining multi-component phase diagrams with desired characteristics using active learning, *Adv. Sci.*, 2021, **8**, 2003165.
14 K. Terayama, R. Tamura, Y. Nose, H. Hiramatsu, H. Hosono, Y. Okuno and K. Tsuda, Efficient construction method for phase diagrams using uncertainty sampling, *Phys. Rev. Mater.*, 2019, **3**, 033802.
15 R. Tamura, G. Deffrennes, K. Han, T. Abe, H. Morito, Y. Nakamura, M. Naito, R. Katsube, Y. Nose and K. Terayama, Machine-learning-based phase diagram construction for high-throughput batch experiments, *Sci. Technol. Adv. Mater.:Methods*, 2022, **2**, 153–161.
16 X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions", in *Proceedings of the 20th International conference on Machine learning*, ICML-03, 2003, pp. 912–919.
17 D. D. Lewis and W. A. Gale, A sequential algorithm for training text classifiers, *Proceedings of the 17th annual international ACM SIGIR conference (SIGIR 1994)*, Dublin, Ireland, 1994, pp. 3–12.
18 R. Tamura, H. Morito, G. Deffrennes, M. Naito, Y. Nose, T. Abe and K. Terayama, Aiphad, an active learning web application for visual understanding of phase diagrams, *Commun. Mater.*, 2024, **5**, 139.
19 R. Katsube, K. Terayama, R. Tamura and Y. Nose, 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.
20 K. Terayama, M. Sumita, R. Tamura and K. Tsuda, Black-box optimization for automated discovery, *Acc. Chem. Res.*, 2021, **54**, 1334–1346.
21 G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, *et al.*, Self-driving laboratories for chemistry and materials science, *Chem. Rev.*, 2024, **124**, 9633–9732.
22 J. Mancias, B. Vela, J. Flórez-Coronel, R. Tavakoli, D. Allaire, R. Arróyave and D. Tourret, Mapping of microstructure transitions during rapid alloy solidification using bayesian-guided phase-field simulations, *Acta Mater.*, 2025, **297**, 121354.
23 D. Barber, *Bayesian reasoning and machine learning*, Cambridge University Press, 2012.
24 A. Kulesza, B. Taskar, *et al.*, Determinantal point processes for machine learning, *Found. Trends Mach. Learn.*, 2012, **5**, 123–286.
25 O. Abril-Pla, V. Andreani, C. Carroll, L. Dong, C. J. Fonnesbeck, M. Kochurov, R. Kumar, J. Lao, C. C. Luhmann, O. A. Martin, *et al.*, Pymc: a modern, and comprehensive probabilistic programming framework in python, *PeerJ Comput. Sci.*, 2023, **9**, e1516.
26 L. Dreval, Y. Zeng, O. Dovbenko, Y. Du, S. Liu, B. Hu and H. Zhang, Thermodynamic description and simulation of solidification microstructures in the cu–mg–zn system, *J. Mater. Sci.*, 2021, **56**, 10614–10639.
27 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, Scikit-learn: Machine learning in python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
28 T. Kathuria, A. Deshpande and P. Kohli, Batched gaussian process bandit optimization via determinantal point processes, *Adv. Neural Inf. Process. Syst.*, 2016, **29**, 4213–4221.
29 C. Oh, R. Bondesan, E. Gavves and M. Welling, Batch bayesian optimization on permutations using the acquisition weighted kernel, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 6843–6858.
30 E. Nava, M. Mutny, and A. Krause, "Diversified sampling for batched bayesian optimization with determinantal point processes", in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, pp. 7031–7054.
31 M. D. Hoffman, A. Gelman, *et al.*, The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo, *J. Mach. Learn. Res.*, 2014, **15**, 1593–1623.