

Multi-fidelity Sequential Learning for Accelerated Materials Discovery

Aini Palizhati^{1,2}, Muratahan Aykol², Santosh Suram², Jens Hummelshøj², and Joseph Montoya²

¹ Carnegie Mellon University

² Toyota Research Institute

E-mail: joseph.montoya@tri.global

Abstract. We introduce a new agent-based framework for materials discovery that combines multi-fidelity modeling and sequential learning to lower the number of expensive data acquisitions while maximizing discovery. We demonstrate the framework’s capability by simulating a materials discovery campaign using experimental and DFT band gap data. Using these simulations, we determine how different machine learning models and acquisition strategies influence the overall rate of discovery of materials per experiment. The framework demonstrates that including lower fidelity (DFT) data, whether as a-priori knowledge or using in-tandem acquisition, increases the discovery rate of materials suitable for solar photoabsorption. We also show that the performance of a given agent depends on data size, model selection, and acquisition strategy. As such, our framework provides a tool that enables materials scientists to test various acquisition and model hyperparameters to maximize the discovery rate of their own multi-fidelity sequential learning campaigns for materials discovery.

materials, machine learning, density functional theory

1. Introduction

Materials discovery and optimization is a key process in the development of new technology and is particularly critical to emerging energy technologies for mitigating fossil fuel use. Current materials for fuel cell catalysts, batteries, thermoelectrics, and other such technologies are frequently costly, unsafe, unstable, or otherwise limited in their efficacy. This motivates the discovery of improved materials and the improvement of the materials discovery process itself. The current material discovery process remains time-intensive and ad-hoc, frequently requiring multiple researcher-years of effort. Computational simulations of materials, particularly from density functional theory (DFT)[1, 2], have become an important complement to the discovery process and have even guided the experimental discovery of materials in a number of cases.[3, 4, 5, 1, 2, 6] As both these simulations and high-throughput experiments have been automated for materials property prediction and measurement, large datasets[7, 8, 9, 10] have become

available to researchers. This in turn has enabled the development of accurate and interpretable surrogate models from machine learning designed to quickly predict the desired property of material before either experiment or simulation[3].

While materials discovery and property prediction have been partially addressed by several machine learning strategies and frameworks [11, 12, 13, 14, 15, 16, 17, 18, 19], the diversity of materials measurement and simulation methods, along with the requirement to carry out experiments under budget constraints, require building systems that acquire new data efficiently according to their most recent results. Thus, using sequential learning (SL) and Bayesian optimization (BO) strategies[20, 21, 22, 23, 24, 25, 26, 27] to obtain a property with experiments and/or simulations at varying fidelities and acquisition costs[28, 29, 30, 31, 32, 33, 34, 35] are particularly promising. In SL or BO for materials design, decisions to synthesize, characterize or simulate candidate materials are made with the aid of machine learning in an iterative process. The results of new experiments or computations are fed back to the model and used in subsequent iterations. Recently demonstrated applications of SL in materials science include systems that learn how to perform only the most valuable or relevant DFT simulations using previous iterations[5, 36, 23], improve force fields more rapidly for molecular dynamics simulations[22, 37, 38], or even synthesize carbon nanotubes at new conditions that promote higher yields and higher qualities of product [25]. Multi-fidelity models, on the other hand, combine measurements or experiments performed at varying degrees of accuracy in their fitting and prediction procedures. The advantage of such methods is that they can combine large quantities of cheaply acquired, less accurate data with data acquired via more expensive but accurate methods. Multi-fidelity models can be significantly more accurate in their predictions than similar models trained on single-fidelity datasets.[16, 29, 28] They may also help mitigate resource limitations that make sampling at higher fidelities impractical or impossible in high-throughput.[28, 30, 31] SL and multi-fidelity modeling methods are typically used and investigated independently in the scientific literature, i.e. the relationship between the acquisition strategy and the modeling strategy is chosen on an ad-hoc basis. Therefore, there remains an opportunity to investigate how scientists can optimally use SL and multi-fidelity methods together and develop a framework to experiment with different strategies for their simultaneous use.

In the present work, we introduce a multi-fidelity SL framework for materials discovery with an objective to lower the number of high fidelity acquisitions while maximizing discovery. We demonstrate the framework’s capability using band gap data of inorganic compounds at two fidelities; namely, experimental measurements as the high fidelity source and GGA-level DFT simulations as the low fidelity alternative. Band gap is a fundamental electronic property relevant for a wide range of technological applications[39, 40, 41], and hence been the subject of various high-throughput efforts which have provided large amount of data available at multiple fidelities.[28, 29, 42] Using the framework, we performed several simulations of iterative SL campaigns for the discovery of inorganic materials with a target electronic band gap window under two

different settings. In the first setting, DFT data was considered as prior knowledge and made available for model fitting before any experimental measurements were acquired. In the second setting, DFT and experimental data were acquired together in sequential iterations. These SL procedures determined how different ML models and acquisition strategies influence the overall rate of discovery of materials per experiment. Our results suggest that, even in a simple, automated process, consideration of lower fidelity DFT data in conjunction with experimental data can increase the rate of discovery of materials suitable for solar photoabsorption. We also demonstrate that the type of machine learning model used in the SL procedure controls the extent of an increase in the number of discoveries in the multi-fidelity setting compared to the single-fidelity baseline.

2. Methods

2.1. Dataset Collection and Representation

The band gap dataset was collected from two sources: (1) experimentally reported band gaps aggregated by Zhuo et al.[16] and disseminated via Matminer[43] and (2) GGA-level DFT-computed band gaps disseminated via the Materials Project.[7, 44] The experimental band gaps include 3906 unique compositions. Compositions with multiple measurements were averaged to obtain a mean band gap value. For each composition, the band gap corresponding to the most phase-stable (i.e. lowest computed energy per atom) crystal structure was obtained from the Materials Project. In our discovery campaign simulations, these band gaps from DFT using the Perdew-Burke-Ernzerhof (PBE) functional[45] were considered low fidelity data, as GGA has well-known systematic errors that underestimate experimentally measured band gaps by ~ 0.9 eV.[46]

We processed the data by using a fixed-length vector to encode both the compositions of each material and the level of fidelity. More specifically, stoichiometric compositions were featurized with the matminer ElementProperty featurizer[43]. The levels of fidelity were captured with one-hot encoding, where a binary variable was added for each fidelity level. For example, for experimental data, a “1” was placed under the “experiment” feature and a “0” under the “theory” feature. The final overall features were scaled such that their distribution had a mean of 0 and a standard deviation of 1.

2.2. Multi-fidelity Sequential Learning Procedure

The multi-fidelity SL framework is built on the recently introduced system for Computational Autonomy for Materials Discovery (CAMD)[23]. CAMD follows an agent-based abstraction for convenient design and testing of acquisition strategies from candidate pools in SL-based optimization. Figure 1 outlines the CAMD framework and highlights the newly constructed multi-fidelity acquisition feature. In a given campaign, the (multi-fidelity) seed data and candidate data (search space) go into an agent. A

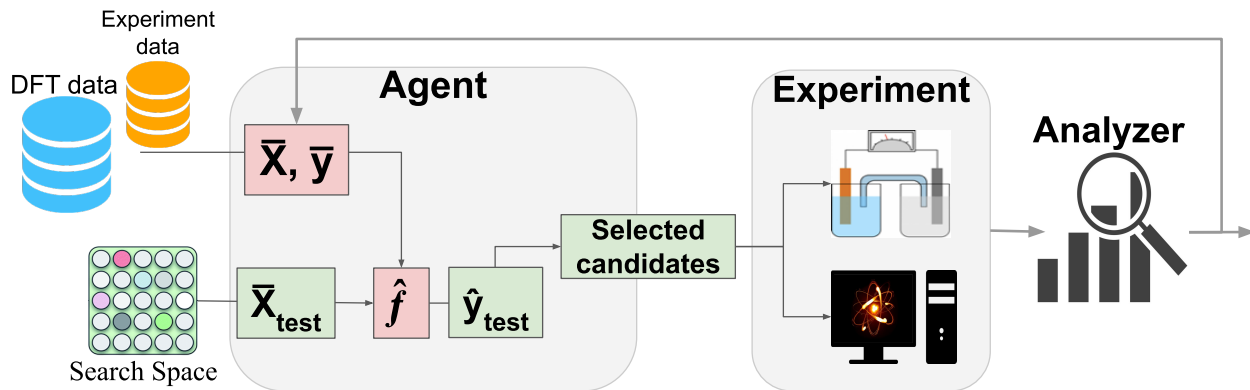


Figure 1. A schematic outlining the multi-fidelity sequential learning framework. The major components of the framework include an agent, experiment, and analyzer. The agent is responsible for the selection of new experiments based on the current state of known data. The experiment acquires new data based on candidates the agent has selected. The analyzer post-processes any experimental data acquired and provides them to the agent for the next round of experiment selections.

preprocessing step in the agent featurizes each member of the seed and candidate data’s composition and fidelity as described in the previous section. The featurized seed data is used to train a machine learning model, which makes predictions on the candidate data for the target property. Using the predictions, the agent then selects candidates at different fidelities. The selected candidates (i.e. those hypothesized as promising by the agent) are sent to an Experiment API, which in our case is an “after-the-fact” (ATF) API that emulates DFT simulation and experimental measurement, respectively, and returns the results from the known dataset. We note that ATF experiment API can later be exchanged for a “true” experiment API to run the actual experiments or simulations with an agent of choice [23]. The analyzer monitors the campaign results and provides an analysis of the experiments in the context of the previously collected data (seed data) and the progress of the campaign. New experimental results are then appended to the seed data and removed from the candidate data, after which the process begins again in the agent phase.

2.3. Agent Design for Materials Discovery

Designing the agent for a multi-fidelity SL procedure required two steps: (1) Selecting an appropriate machine learning model and (2) developing a CAMD-compatible[23] multi-fidelity procedure for acquisition. For model selection, we compared several well-known regression methods, including support vector regression (SVR), k-nearest neighbors (KNN), random forest regression (RFR), and Gaussian process regression (GPR). The selection process included hyperparameter tuning and comparative performance analysis, whose details were provided in Supplementary document section S1. Based

on the results, SVR, RFR, and GPR with noted hyperparameters had qualitatively similar performances and were used consistently for the framework construction and demonstration. However, note that the developed framework (detailed next) allows users to input their ML model and model hyperparameters of choice for the agent.

For the multi-fidelity framework itself, We designed two agents, an epsilon-greedy multi-fidelity agent (ϵ -greedy-MF) and a Gaussian process lower confidence bound[47] derived multi-fidelity agent (GPR_{LCB} -MF). The ϵ -greedy-MF agent takes any supervised machine learning regressor from scikit learn[48] as input and allocates the desired budget for experiment versus theory data, and acquires candidates in each fidelity via exploitation (Algorithm S1). This algorithm prioritizes doing experiments on predicted ideal compositions if they are structurally similar to some known compositions (in seed data). Otherwise, lower-fidelity computations are performed first. The GPR_{LCB} -MF agent (uses GPy[49] package) operates under a total acquisition budget. It acquires candidates factoring in GPR predicted uncertainties in the LCB setting and *hallucination* of information gain from DFT acquisitions analogous to work of Desautels et al. in batch mode LCB[50] (Algorithm S2). The agent aims for prioritizing *exploitation* primarily with high-fidelity experimental measurements, offloading exploratory (higher risk) acquisitions first to lower-fidelity computations. As noted in Algorithm S2, three critical acquisition hyperparameters require optimization - α , β , and γ . α is the uncertainty mixing parameter, which is the weight of uncertainty in the LCB setting. β occur as a threshold for uncertainty percentile. A small beta means the agent wants to be very confident about the prediction on experiments. Unless the uncertainty on a given prediction is very low, it will acquire lower fidelity data first. If β is large, the agent is tolerant to high uncertainty for experiments and will add prospective experimental. γ estimates the influence of lower fidelity data. If γ is 0, then the prospect of the lower fidelity data has to increase the chances of the experiment being successful. Otherwise, lower fidelity data will be acquired first. If γ is very large, then the agent does not care about how much DFT result affects the potential experiment. We simulated various scenarios of these two hyperparameters to optimize the agents. The details and results are in Supplementary document section S4. Details of agents are also made explicit in the code available via the open-source CAMD repository at <https://github.com/TRI-AMDD/CAMD>.

2.4. Performance Metrics for Sequential Learning

In this work, we used previously established active learning metrics (ALM)[37] to quantify the performance of SL agents, and benchmarked the multi-fidelity SL against single-fidelity SL. These metrics are ^{all}ALM, acceleration factor (AF), and enhancement factor (EF), as shown in Figure 2. ^{all}ALM ranges from 0 to 1. It is the fraction of ideal materials identified at an iteration given an SL run. AF is the reduction of required budget (e.g. in time, iterations, or some other consumed resource) between an agent and a benchmark case (e.g. random selection, single-fidelity, or manual human selection) to

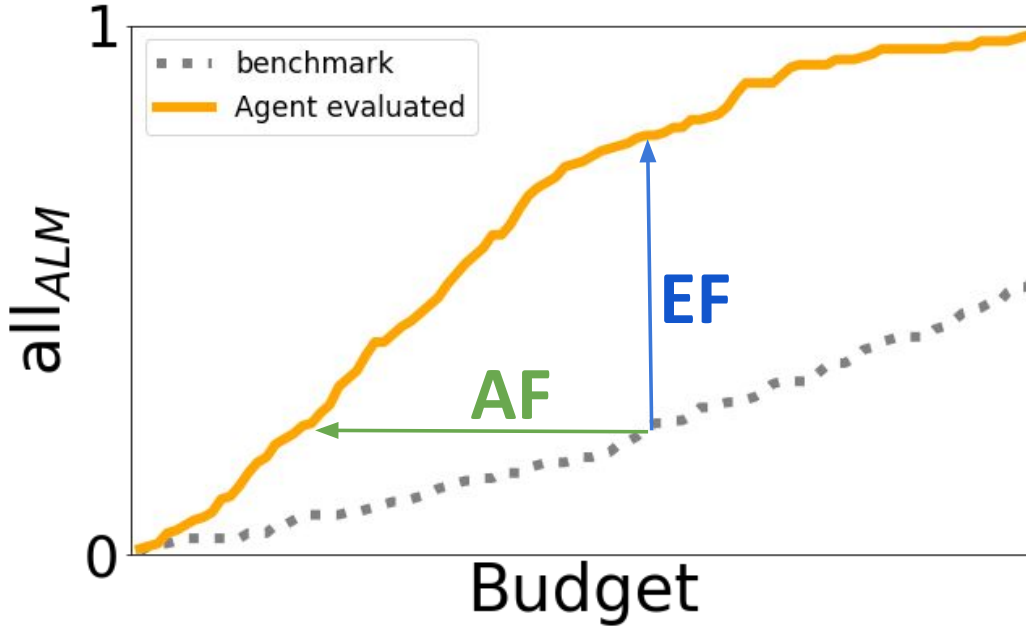


Figure 2. The performance metrics[37] used to evaluate sequential learning procedures. all_ALM (on the y-axis) is the fraction of ideal materials discovered at a given budget. Acceleration factor (AF) is the reduction of the required budget between an agent and a benchmark case to achieve a certain amount of discovery. Enhancement factor (EF) is the performance improvement of an agent compared to the benchmark at a given budget.

reach a particular fraction of ideal candidates ($AF = N_{budget,benchmark} - N_{budget,agent}$). Here, a positive value of AF indicates multi-fidelity campaigns outperform their corresponding single fidelity campaigns because they reduce the required budget needed to achieve a certain amount of discovery. EF shows the performance enhancement at the same consumed experiment budget. More specifically, at the same number of iterations, amount of elapsed time, or some other metric of expended resources, EF quantifies the improvement of materials discovery by a given SL method versus a benchmark method ($EF = \frac{N_{discovery,agent}}{N_{discovery,benchmark}}$). Here, EF greater than one indicates multi-fidelity campaigns outperform their corresponding single fidelity campaigns at a given budget.

2.5. Sequential Learning Objective

For multi-fidelity SL campaign simulations and subsequent performance evaluations of the agents, materials with experimentally measured band gap $\subseteq [1.6, 2.0]$ eV [51], i.e. those with reasonable solar photoabsorption, were considered ideal and set as the “target”.

3. Results

3.0.1. Boundary condition acquisitions: when all DFT data were available. We first tested the acquisition performance of multi-fidelity agents where DFT was considered as a-priori knowledge. The objective here was to determine how an automated experimental SL procedure would be enhanced by a-priori knowledge of a large theoretical dataset. The enhancement is primarily through the *transfer* of knowledge from the large, low-fidelity dataset in making predictions and subsequent acquisitions under the high-fidelity, expensive setting. We simulated runs with three agents: ϵ -greedy agents that used SVR and RFR, and GPR_{LCB} agent. For each agent, we supplied: (1) no seed data and (2) all the DFT data as seed data. The candidate data were the experimental data for both cases, where 202 of 3906 were considered ideal based on the target band gap window defined above. Therefore, (1) represents acquisition based on single (high) fidelity, and (2) represents acquisition based on multi-fidelity (low fidelity present in the seed, high fidelity as target). For convenience, we name them SVR-SF_{boundary}, RFR-SF_{boundary}, and GPR_{LCB}-SF_{boundary}, SVR-MF_{boundary}, RFR-MF_{boundary}, and GPR_{LCB}-MF_{boundary} (SF denotes single fidelity, MF denotes multi-fidelity). We gave all the agents a budget of 20 experiment requests in each iteration and simulated each campaign for 100 iterations. We also provided two limiting cases: (1) a simple agent that randomly chooses data from candidate space and (2) the maximum achievable performance where all the acquisitions are ideal candidates. Put differently, (1) defined a lower bound to ensure that a given agent is delivering some value relative to a trivial selection algorithm, whereas (2) defined the upper bound on agent performance that may be considered a target for when engineering the selection algorithm.

Figure 3 shows the results of the simulated discovery campaigns. For our initial benchmark, all of our SL agents consistently outperformed random acquisitions, demonstrating their ability to deliver value in an automated system. However, the underlying regressors used in each agent influenced their performance. In the single-fidelity case, where there was no access to low-fidelity DFT data, RFR-SF_{boundary} outperformed SVR-SF_{boundary} until ~ 1100 experiment requests, at which roughly 70% of the ideal candidates had been discovered. It also outperformed GPR_{LCB}-SF_{boundary} starting ~ 750 experiment requests. In the multi-fidelity case where all low-fidelity (DFT) data was made available, the SVR-MF_{boundary} agent consistently outperformed RFR-MF_{boundary} and GPR_{LCB}-MF_{boundary} until ~ 1500 experiment requests, at which $\geq 90\%$ of ideal candidates had been discovered. More importantly, we observed that multi-fidelity agents that used SVR-MF_{boundary} (Figure 3b) and GPR_{LCB}-MF_{boundary} (Figure 3d) outperformed their single fidelity counterparts, demonstrating that these regression algorithms can transfer the knowledge available from the lower-fidelity dataset in making predictions for the high-fidelity target, whereas RFR-MF_{boundary} (Figure 3c) cannot.

To illustrate the comparative performance of single and multi-fidelity agents in more detail, we tabulated acceleration factors at 50% and 90% of the total discovery of ideal

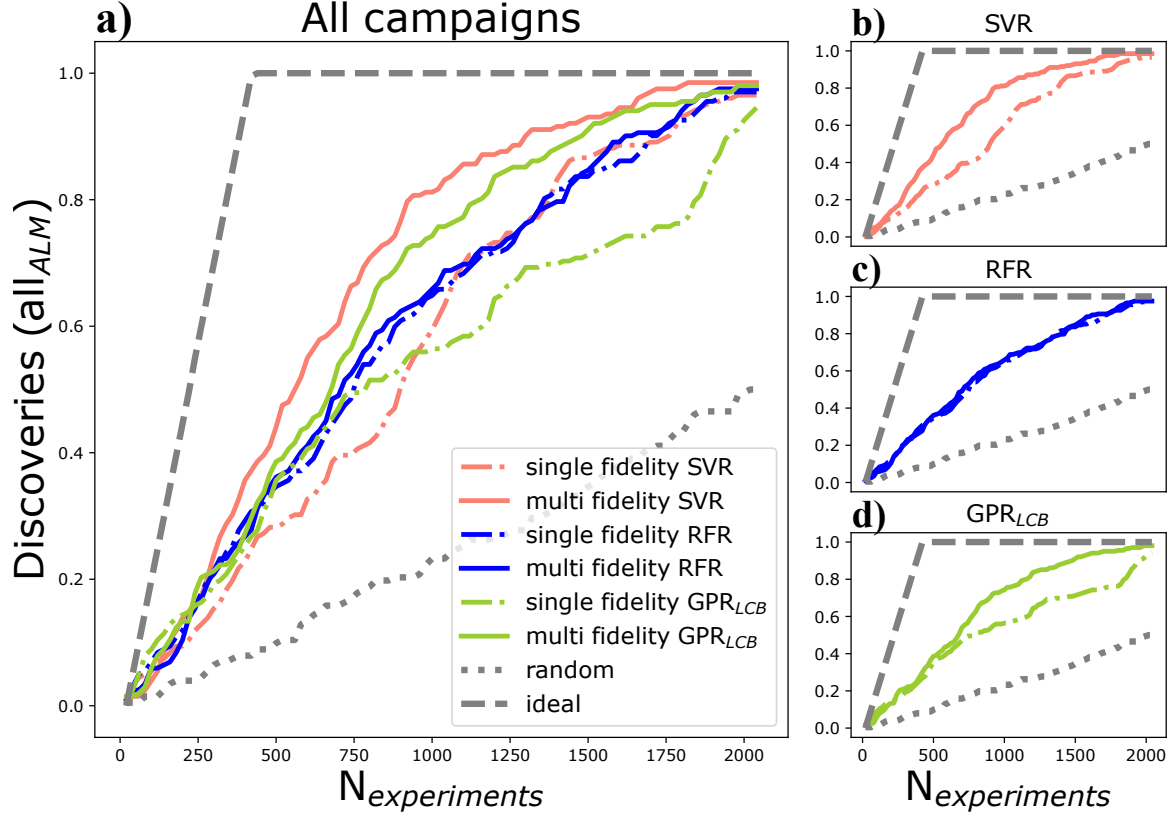


Figure 3. Simulated performance of various campaigns with and without supplying the lower fidelity (DFT) data in the seed. a) shows performance of all campaigns. c-d) shows performance of campaigns separately based on the underlying ML model. For each figure, $N_{\text{experiment}}$ on the x-axis corresponds to the high-fidelity (experiments) acquisition budget. Discoveries (all_{ALM}) on the y-axis corresponds to the fraction of ideal materials discovered from the search space. SVR, RFR, GPR_{LCB} correspond to agents using support vector regressor, random forest regressor, and gaussian process regressor lower confidence bound, respectively. The random acquisition and ideal acquisition baselines are also labeled in the figure, representing the lower and upper bounds of agent performance.

candidates in Table 1. For a multi-fidelity agent with either of the SVR or GPR_{LCB}, 320 or 120 fewer experiments were required to achieve the discovery of 50% of the candidates designed as ideal. Similarly, 460 or 480 fewer experiments were needed to achieve a 90% discovery. In comparison, the RFR-MF_{boundary} agent displayed a much smaller enhancement (40 at 50% and 80 at 90% discovery) over its single-fidelity counterpart. The enhancement factors shown in Figure 4 provided a clearer picture of the comparative performance throughout the campaign. We observe that SVR and GPR_{LCB} multi-fidelity agents briefly underperformed their single-fidelity counterparts in the early stages of campaigns (until ~ 100 and ~ 250 experiments, respectively). After this point, SVR-MF_{boundary} outperformed SVR-SF_{boundary} by a notable margin to achieve enhancement of a factor of ~ 1.5 to 1.6 . This factor diminished slowly as candidates were exhausted for the remainder of the campaign. GPR_{LCB}-MF_{boundary} outperformed GPR_{LCB}-SF_{boundary}

by a smaller amount at the beginning, but the performance further enhanced at ~ 750 experiments, which corresponds to the results in Figure 3d. Lastly, EF of RFR-MF_{boundary} stayed around 1, which means there was not any difference in RFR MF and SF agents’ performance. As an additional analysis, for each set of agents (multi-fidelity and its corresponding single fidelity) during their various stages of the campaign, We computed the reduction in seed data MAE from single fidelity to multi-fidelity. The results again showed having DFT in seed data improved the agent’s prediction ability. The details of the analyses are included in Supplementary document section S2.

Agents	Single fidelity experiments performed for 50% discovery	multi-fidelity AF _{50%}	Single fidelity experiments performed for 90% discovery	multi-fidelity AF _{90%}
SVR	880	320	1740	460
RFR	740	40	1680	80
GPR _{LCB}	780	120	1940	480

Table 1. Acceleration factor of multi-fidelity agents in simple acquisitions. The AF’s are the reduction in number of experiments performed by multi-fidelity agents to achieve a certain amount of discoveries. For each row, we highlighted the agents used, the experiments performed by single fidelity agents to achieve 50% and 90% discovery, and the acceleration factor (AF) of the multi-fidelity agents at those discoveries.

3.0.2. In-tandem acquisitions: when both DFT and experiment data were acquired
 Having looked into two limiting scenarios of with or without low fidelity data prior to campaigns, we now turn to our next main question: when and how should we decide to acquire DFT calculations to support and minimize the number of experimental measurements *during* a sequential, closed-loop data acquisition procedure? To answer this, we simulated another set of campaigns. First, to mimic a more true-to-life discovery process, we split the compositions into seed data and candidate data based on the year of discovery according to the ICSD[52] timeline of their first publication[53], which also mitigated the effect of a random initial seed (Figure 5). For single fidelity campaigns, the data of the first 500 experimentally discovered compositions, up to the discovery year of 1965, were included in the seed data, the remaining were included in the seed data. For multi-fidelity campaigns, the data split was identical, with the addition of corresponding DFT data in each set. Next, we set up the campaigns with a ϵ -greedy agent that used SVR and a GPR_{LCB} agent. We excluded ϵ -greedy agent that used RFR based on its weak a-priori performance. Therefore, a total of four SL campaigns were set up: SVR-SF_{tandem}, GPR_{LCB}-SF_{tandem}, SVR-MF_{tandem}, and GPR_{LCB}-MF_{tandem} (SF denotes single fidelity, MF denotes multi-fidelity). We also provided the two limiting cases of (1) random acquisition and (2) maximum achievable acquisition. For acquisition budget, both SVR-SF_{tandem}, GPR_{LCB}-SF_{tandem}, along with the two limiting case, had a budget of 5 experiment requests. SVR-MF_{tandem} had a fix-ratio budget of 5 experiments and 5 DFT. GPR_{LCB}-MF_{tandem} had a budget of 10 acquisitions, each acquisition can

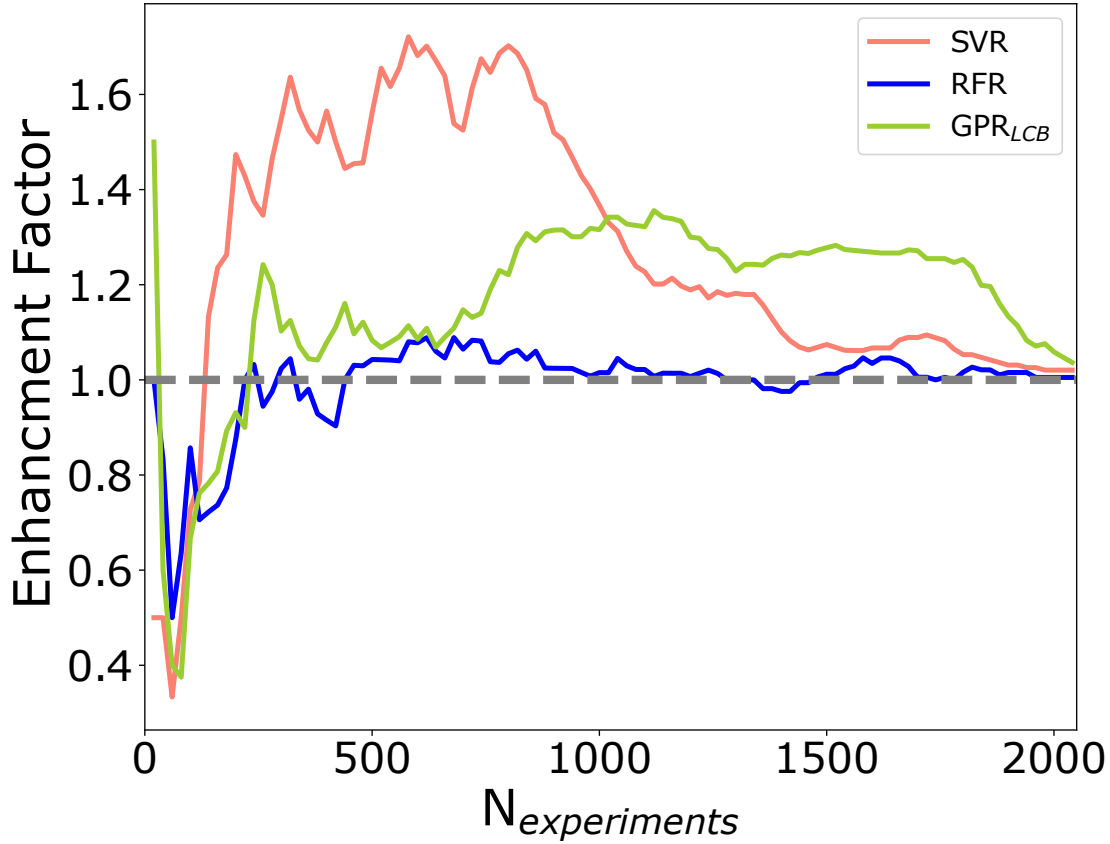


Figure 4. The enhancement factor for quantifying the added value of DFT in the campaigns. The enhancement factors were $\frac{N_{\text{discovery, multi-fidelity}}}{N_{\text{discovery, single-fidelity}}}$ at any given budget.

be either experiments or DFT, depending on the uncertainties and *hallucination* of information gain from DFT. Based on optimization results in Supplementary document section S4, **alpha=0.1**, **beta=10**, and **gamma=0** were used for $\text{GPR}_{\text{LCB}}\text{-MF}_{\text{tandem}}$ to compare against the other SL campaigns.

Figure 6 shows the qualitative results of the simulated campaigns using in-tandem acquisition. Figure 6b shows the $\text{SVR}\text{-MF}_{\text{tandem}}$ agent outperformed its single-fidelity counterpart starting when $N_{\text{experiments}}$ reached around 400. It then stayed slightly ahead until 1000 experiments were acquired, at which point 90% of the ideal materials have been discovered. Figure 6c shows the $\text{GPR}_{\text{LCB}}\text{-MF}_{\text{tandem}}$ agent also outperformed its single-fidelity counterpart until 90% of the ideal materials have been discovered (at ~ 1200 experiments). Compare among all four agents (Figure 6a), $\text{SVR}\text{-MF}_{\text{tandem}}$ agent's performance was similar to $\text{GPR}_{\text{LCB}}\text{-MF}_{\text{tandem}}$ until 750 experiments and became better after that. Furthermore, $\text{GPR}_{\text{LCB}}\text{-MF}_{\text{tandem}}$ agent's performance fell behind $\text{SVR}\text{-SF}_{\text{tandem}}$'s once ~ 1100 experiments had been conducted.

The AF (Table 2) of multi-fidelity acquisitions for $\text{SVR}\text{-MF}_{\text{tandem}}$ and $\text{GPR}_{\text{LCB}}\text{-MF}_{\text{tandem}}$ at 50% discovery were 100 and 226, respectively. At 80% discovery, they were 145 and 111, respectively. The enhancement factors (Figure 7) of $\text{SVR}\text{-MF}_{\text{tandem}}$

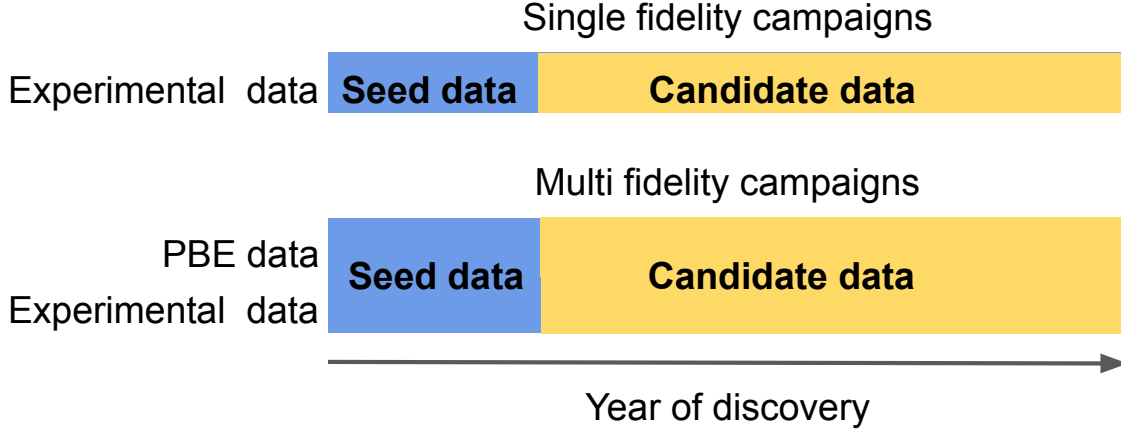


Figure 5. The seed data and candidate data selection for single fidelity campaigns and multi-fidelity campaigns in in-tandem acquisitions. They were decided based on the year of discovery for compositions in the band gap dataset.

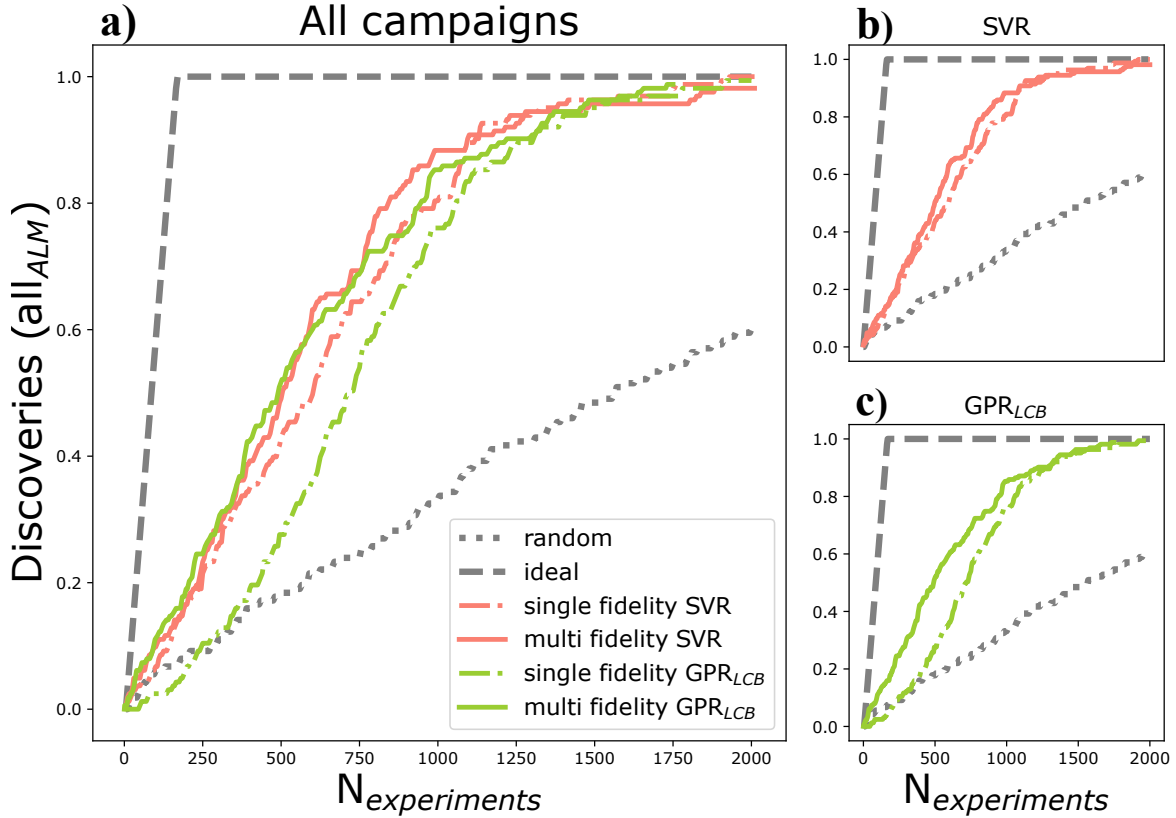


Figure 6. Simulated performance of various campaigns when both DFT and experiment data were acquired during runs. a) were the results of all campaigns combined in one figure. b-c) were the results based on underlying ML used in the agents. $N_{experiment}$ on the x-axis corresponds to the experimental acquisition budget. Discoveries (all_{ALM}) on the y-axis corresponds to the fraction of ideal materials discovered from the search space. The random acquisition and ideal acquisition baselines are also labeled in the figures.

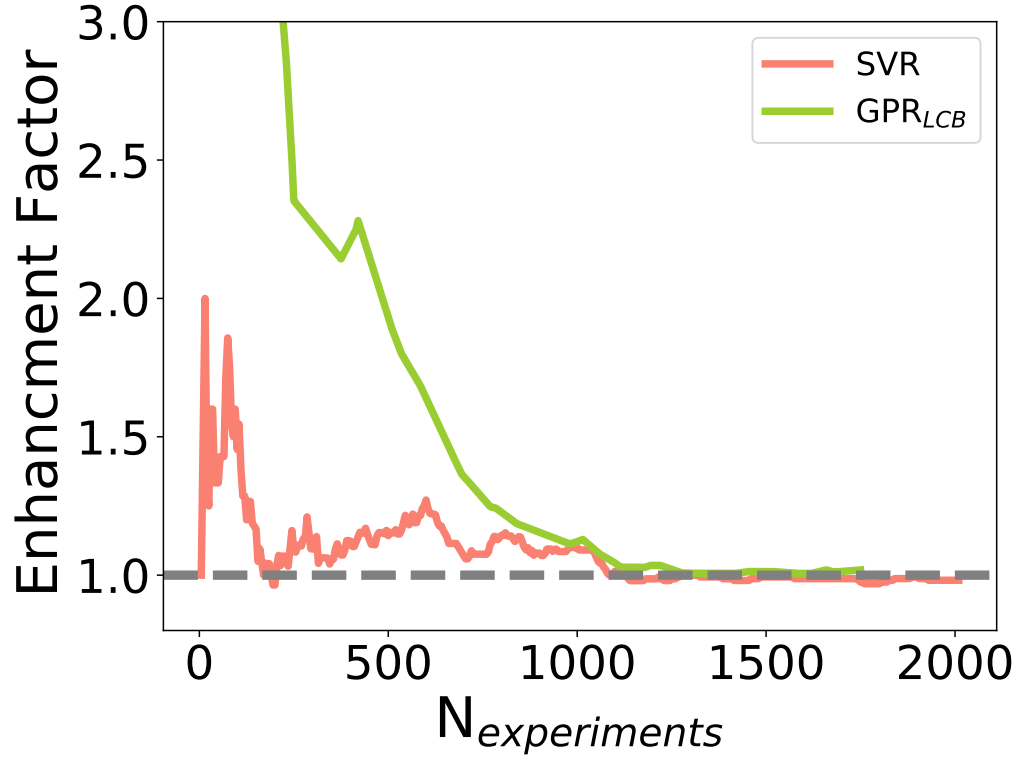


Figure 7. The enhancement factors of in-tandem acquisitions as the campaign progresses.

increased then decreased drastically at first, then stayed slightly above 1 until $N_{\text{experiments}}$ reaches ~ 1100 . The enhancement of $\text{GPR}_{LCB}\text{-MF}_{\text{tandem}}$ was high at the beginning but decreased as the acquisition continued and converged to 1 at ~ 1200 experiments. Combining the results, we noticed that AF and EF were not as large as the previous runs, where all the DFT data was available in the seed space. This was because DFT data were now part of the acquisition, so there was less DFT data in the seed data provided to ML models for the transfer of information from lower fidelity.

Agents	Single fidelity experiments performed for 50% discovery	multi-fidelity $\text{AF}_{50\%}$	Single fidelity experiments performed for 80% discovery	multi-fidelity $\text{AF}_{80\%}$
SVR	605	100	990	145
GPR_{LCB}	720	226	1060	111

Table 2. Acceleration factor of multi-fidelity agents in in-tandem acquisitions. For each row, we highlighted the agents used, the experiments performed by single fidelity agents to achieve 50% and 80% discovery, and the acceleration factor (AF) of the multi-fidelity agents. The AF’s are the reduction in number of experiments performed.

4. Conclusion

In this work, we demonstrate a framework that combines multi-fidelity machine learning models and sequential learning procedures. Using this framework, we simulated a materials discovery process that can acquire band gap data at two fidelities. When all low-fidelity data were provided as a-priori knowledge, both SVR and GPR multi-fidelity agents outperformed their single-fidelity counterparts within 500 experiments and sustained a materials discovery acceleration of 20-60% until over 90% of the target materials had been discovered. When acquiring low and high-fidelity data in-tandem, SVR and GPR multi-fidelity agents still outperformed their single fidelity counterparts. However, the enhancement was lower, based on both the acceleration factor and effectiveness factor.

In summary, we observed a clear trend of multi-fidelity sequential learning agents outperforming those which may only sample at a single fidelity. The results demonstrate that the inclusion of one-hot labeled DFT data either a-priori or acquired in-tandem with experiments can increase the rate at which valuable experiments are acquired. However, the relative performance of multi-fidelity acquisition strongly depends on the dataset size, ML model selection, and acquisition strategy.

Given these dependencies, our framework offers a critical capability that enables the scientific method on the automated discovery process itself. In CAMD, we examine the effects of these hyperparameters not only on ML model performance but the rate at which autonomous agents with ML capabilities can select valuable experiments to perform. Thus, CAMD’s multi-fidelity sequential learning campaigns lay a foundation for future research in which both simulations and experiments can be conducted in tandem with strategies optimized for their relative cost and accuracy.

5. Supporting Information

The supporting information contains details on the preliminary model selection before constructing the multi-fidelity framework. Error rate analyses during boundary condition acquisition, the algorithms related to the two multi-fidelity agents mentioned in the paper, and hyperparameter tuning for GPR_{LCB} multi-fidelity agent. The full details of the code are provided in an open-source repository at <https://github.com/TRI-AMDD/CAMD>.

Acknowledgments

This work was supported by Toyota Research Institute through the Accelerated Materials Design and Discovery program. The authors gratefully acknowledge helpful discussions with Jens Hummelshøj, Chirranjeevi Gopal, Abraham Anapolsky, and Linda Hung.

References

- [1] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** B864
- [2] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133
- [3] Schmidt J, Marques M R, Botti S and Marques M A 2019 *npj Computational Materials* **5** 1
- [4] Cai J, Chu X, Xu K, Li H and Wei J 2019 *Nanoscale Advances* **5** 1
- [5] Tran K and Ulissi Z W 2018 *Nature Catalysis* **1** 696
- [6] Jain A, Shin Y and Persson K A 2016 *Nature Reviews Materials* **1** 1
- [7] Jain A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G and Persson K A 2013 *APL Materials* **1** 011002
- [8] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 *npj Computational Materials* **1** 1
- [9] Curtarolo S, Setyawan W, Hart G L, Jahnatek M, Chepulskii R V, Taylor R H, Wang S, Xue J, Yang K, Levy O, Mehl M J, Stokes H T, Demchenko D O and Morgan D 2012 *Computational Materials Science* **58** 218
- [10] Ong S P 2019 *Computational Materials Science* **161** 143
- [11] Ryan K, Lengyel J and Shatruk M 2018 *Journal of the American Chemical Society* **140** 10158
- [12] Dan Y, Zhao Y, Li X, Li S, Hu M and Hu J 2020 *npj Computational Materials* **6** 1
- [13] Erdem Gunay M and Yildirim R 2020 *Catalysis Reviews - Science and Engineering* **63** 120
- [14] Jennings P C, Lysgaard S, Hummelshøj J S, Vegge T and Bligaard T 2019 *npj Computational Materials* **5** 46
- [15] Ward L, Liu R, Krishna A, Hegde V I, Agrawal A, Choudhary A and Wolverton C 2017 *Physical Review B* **96** 024104
- [16] Zhuo Y, Mansouri Tehrani A and Brgoch J 2018 *Journal of Physical Chemistry Letters* **9** 1668
- [17] Xie T and Grossman J C 2018 *Physical Review Letters* **120** 145301
- [18] Palizhati A, Zhong W, Tran K, Back S and Ulissi Z W 2019 *Journal of Chemical Information and Modeling* **69** 4742
- [19] Balachandran P V, Kowalski B, Sehrioglu A and Lookman T 2018 *Nature Communications* **9** 1
- [20] del Rosario Z, Rupp M, Kim Y, Antono E and Ling J 2020 *The Journal of Chemical Physics* **153** 024112
- [21] Noh J, Back S, Kim J and Jung Y 2018 *Chemical Science* **9** 5152
- [22] Coley C W, Eyke N S and Jensen K F 2020 *Angewandte Chemie International Edition* **59**
- [23] Montoya J H, Winther K T, Flores R A, Bligaard T, Hummelshøj J S and Aykol M 2020 *Chemical Science* **11** 8517
- [24] Bassman L, Rajak P, Kalia R K, Nakano A, Sha F, Sun J, Singh D J, Aykol M, Huck P, Persson K and Vashishta P 2018 *npj Computational Materials* **4** 74
- [25] Nikolaev P, Hooper D, Webber F, Rao R, Decker K, Krein M, Poleski J, Barto R and Maruyama B 2016 *npj Computational Materials* **2** 1
- [26] Häse F, Roch L M, Kreisbeck C and Aspuru-Guzik A 2018 *ACS Central Scienc* **4** 1134
- [27] Ling J, Hutchinson M, Antono E, Paradiso S and Meredig B 2017 *Integrating Materials and Manufacturing Innovation* **6** 207
- [28] Chen C, Zuo Y, Ye W, Li X and Ong S P 2020 *arXiv*
- [29] Pilania G, Gubernatis J E and Lookman T 2017 *Computational Materials Science* **129** 156
- [30] Kandasamy K, Dasarathy G, Schneider J and Póczos B 2017 *ICML'17* **70** 1799
- [31] Tian H and Rangarajan S 2019 *J. Chem. Theory Comput.* **15** 5588
- [32] Greenhill S, Rana S, Gupta S, Vellanki P and Venkatesh S 2020 *IEEE Access* **8** 13937
- [33] Tran A, Tranchida J, Wildey T and Thompson A P 2020 *Journal of Chemical Physics* **153** 074705
- [34] Batra R and Sankaranarayanan S 2020 *JPhys Materials* **3** 031002
- [35] Chen C, Zuo Y, Ye W, Li X and Ong S P 2021 *Nature Computational Science* **1** 46
- [36] Seko A, Togo A, Hayashi H, Tsuda K, Chaput L and Tanaka I 2015 *Phys. Rev. Lett.* **115** 205901
- [37] Rohr B, Stein H S, Guevarra D, Wang Y, Haber J A, Aykol M, Suram S K and Gregoire J M 2020

- Chemical Science* **11** 2696
- [38] Vandermause J, Torrisi S B, Batzner S, Kolpak A M and Kozinsky B 2019 *Preprint at <https://arxiv.org/abs/1904.02042>*
- [39] Jain A, Shin Y and Persson K A 2016 *Computational predictions of energy materials using density functional theory* vol 1 (Nature Publishing Group)
- [40] Chaudhry A, Boutchko R, Chourou S, Zhang G, Grønbech-Jensen N and Canning A 2014 *Phys. Rev. B* **89** 155105
- [41] Polman A, Knight M, Garnett E C, Ehrler B and Sinke W C 2016 *Science* **352**
- [42] Kiselyova N N, Dudarev V A and Korzhuyev M A 2016 *Inorganic Materials: Applied Research* **7** 34
- [43] Ward L, Dunn A, Faghaninia A, Zimmermann N E R, Bajaj S, Wang Q, Montoya J, Chen J, Bystrom K, Dylla M, Chard K, Asta M, Persson K A, Snyder G J, Foster I and Jain A 2018 *Computational Materials Science* **152** 60
- [44] Ong S P, Cholia S, Jain A, Brafman M, Gunter D, Ceder G and Persson K A 2015 *Computational Materials Science* **97** 209
- [45] Perdew J P, Burke K and Ernzerhof M 1996 *Physical review letters* **77** 3865
- [46] Morales-García Á, Valero R and Illas F 2017 *The Journal of Physical Chemistry C* **121** 18862
- [47] Srinivas N, Krause A, Kakade S M and Seeger M 2009 *arXiv preprint arXiv:0912.3995*
- [48] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E 2011 *Journal of Machine Learning Research* **12** 2825
- [49] GPy 2012 a Gaussian process framework in python
- [50] Desautels T, Krause A and Burdick J W 2014 *Journal of Machine Learning Research* **15** 3873
- [51] Hu S, Xiang C, Haussener S, Berger A D and Lewis N S 2013 *Energy and Environmental Science* **6** 2984
- [52] Belsky A, Hellenbrandt M, Karen V L and Luksch P 2002 *Acta Crystallographica Section B Structural Science* **58** 364
- [53] Choudhury R, Aykol M, Gratzl S, Montoya J and Hummelshøj J 2020 *Journal of Open Source Software* **5** 2105