Lessons Learnt from Applying Sequential Learning in High-Dimensional Optimization of Perovskite Solar Cells

Zhe Liu[1,*], Nicholas Rolston[2,*], Zekun Ren[3], Felipe Oviedo[1],Qiaohao Liang[1], Shijing Sun[1]
Reinhold H. Dauskardt[2,†], Tonio Buonassisi[1,†]

[1]Massachusetts Institute of Technology (MIT), Cambridge, MA, United States
[2]Stanford University, Stanford, CA, United States
[3]Singapore-MIT Alliance for Research and Technology (SMART), Singapore

[*]These two authors contributed equally as co-first authors.
[†]Email correspondence: buonassi@mit.edu (T.B.) & rhd@stanford.edu (R.H.D)

Abstract

Introduction

Development of a novel process to fabricate functional materials involves a systematic optimization of high-dimensional process variables (*e.g.*, 10 variables or more), which is a challenging task for human intuition alone to solve the problem efficiently. High-throughput experimentation has recently demonstrated very encouraging outcomes for rapid materials development [1]–[4]. Inspired by the new capability of high-throughput experimentation, preliminary successes were achieved by introducing machine learning (ML) in the loop to guide the experiments of high-dimensional optimization. This iterative machine learning approach with a self-updating regression model and sequential design of the next experiments is known as sequential learning, which is sometimes also referred as active learning.

The most common framework for sequential learning in materials science and engineering is the Bayesian optimization framework. A closed-loop Bayesian optimization was demonstrated for semiconductor thin films. Häse *et al.* developed the chemistry-specific Bayesian optimization *Phoenics* package [5], Roch *et al.* further integrated into a software suite *ChemOS* for materials discovery and optimization [6], and MacLeod *et al.* then deployed the software suite to drive the experimental robot to autonomously optimize semiconductor thin films [7].

Other recent research studies showed the ML-assisted experiment planning can overperform the conventional design of experiments methods based on the domain knowledge from human experts in many specific case studies, *e.g.*, optimizing the reaction yield of chemical synthesis [8], discovering new photocatalysts for CO_2 reduction [9] and searching for the best fast-charging protocol [10]. Besides, to fully assess the acceleration in materials optimization using sequential learning, Ling *et al.* [11] and Rohr *et al.* [12] respectively conducted the benchmarking analysis of sequential learning methods in several cases of different materials databases. A software package *Olympus* was recently created by Häse *et al.* to facilitate a quicker evaluation of sequential learning methods [13].

Although ML is demonstrated as a powerful tool to navigate the optimization task in a high-dimensional space, the black-box nature of machine learning hinders wide adoption in materials science research [14].

Sometimes, materials scientists are not able to fully understand the decisions made by ML algorithms, or to engage intellectually with the ML-driven optimization. To tackle this challenge of understanding ML models during sequential learning, we discuss some generalizable learnings from applying the sequential learning framework to optimize perovskite solar cells: (a) the impact of initialization methods, and acquisition functions on the outcome of sequential learning; (b) the data visualization methods to understand the experimental planning decisions made by sequential learning; (c) the correlation insights from the regression model by applying SHAP analysis on the regression models during sequential learning.

Dataset

Fig. 1 illustrates the open-air spray deposition technique with a post-curing process, which is considered as a scalable fabrication process with ultrahigh throughput. For the perovskite absorber layer $Cs_{0.17}FA_{0.83}Pb(Br_{0.17}I_{0.83})_3$ alone, we selected the 12 key process variables that would have an impact on the perovskite film quality and thereafter the power conversion efficiency (PCE) of the PV devices. There are 4 parameters for precursor solutions, including precursor concentration, DSMO solvent amount, MACl additive amount, and PbI_2 additive amount; 8 other parameters for spray and post-treatment processing, including spray flow rate, linear movement speed, substrate temperature, atmosphere humidity, post-deposition curing method (*i.e.*, gas blade or plasma treatment), height of curing nozzle, curing gas type (*i.e.*, N_2 or dry air), and gas flow rate. By varying these 12 variables, the researchers at Stanford have accumulated an experimental process database of 344 unique conditions (corresponding to XX devices) during the initial process optimization of the solar cells. The optimization of the process variables was done by the one-variable-at-a-time method in a 12-dimensional parameter space, and the data are listed in the database according to the chronological order of the experiments.

We utilized this database for a virtual optimization of the perovskite solar cells using sequential learning. The objective is to find the optimal process conditions that achieves the device with PCE >15% in this database. The advantage of using the existing database over running the actual experiments is to gain statistical understanding of sequential learning (with multiple runs of different initializations and acquisitions) without a huge experimental cost [11], [12].

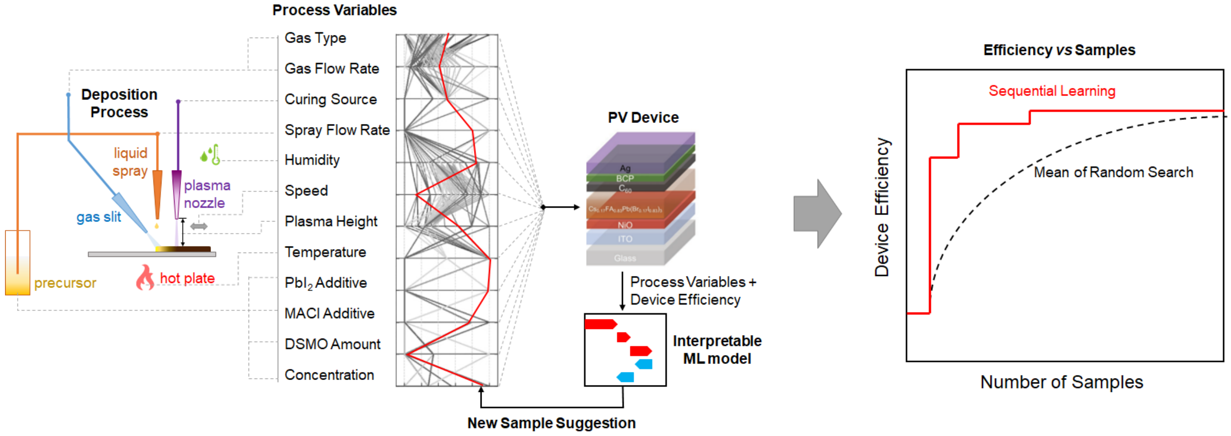

Figure 1: Schematic illustration of the high-dimensional optimization of perovskite solar cells with spray deposition tool, using the interpretable sequential learning framework.

Visualization of the Dataset

To visualize and explore the dataset, the t-student Stochastic Network Embedding (tSNE) method was applied to reduce the 12-dimensional dataset into a two-dimensional space. Figure 2a shows that the dataset has two local maxima for the device efficiency. The global maximum efficiency is 15.x% located in the bottom cluster. The curing sources in Figure 2b reveals the distinction between the bottom cluster and the rest are due to the difference in curing source used. Note that the curing source is a categorical variable, which is encoded in the following optimization. Figure 2c shows another example of a continuous variable, namely temperature, where the temperature gradient is captured in tSNE plot as well.

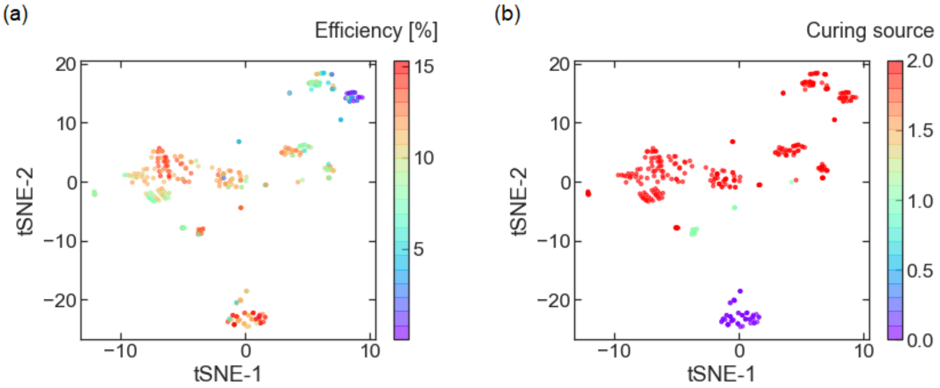

Figure 2: Dataset visualization with the t-student stochastic network embedding tSNE. The tSNE space is a two-dimensional reduced-space representation (i.e., tSNE-1 and tSNE-2) of the original 12-dimensional process variables. (a) Device efficiency overlaid on the tSNE space. (b) The process variable "curing source" is overlaid on the tSNE space, which includes three categories: air, none and plasma. This helps explain gaps in the tSNE.

To better visualize the optimization procedure, we adopted the tSNE dimension reduction to map high-dimension space into 2D space and tracked the sampling sequence in the reduced space.

Initial Sampling Improves the Regression Model from the Start

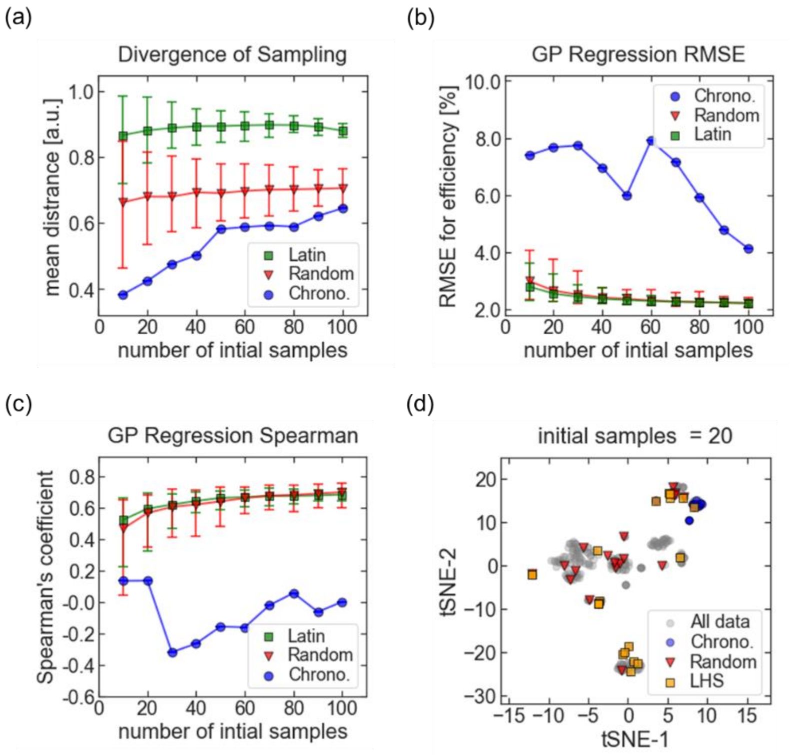

Figure 4: **Comparison of the initial sampling methods at the start of the sequential learning**. (a) Visualization of the initial sampled data using different methods in the tSNE space with a sample size of 20. The tSNE space is two-dimensional reduced space (*i.e.*, tSNE-1 and tSNE-2) of the original 12-dimensional process variables. (b) Divergence metric of the sampled datasets using different sampling methods. The divergence metric is the mean Euclidean distance between each sample (*i.e.*, each sampled process parameter normalized to the 0 – 1 scale) and the centroid of the sampled dataset (*i.e.*, the mean value of all the sampled process parameters). (c) Root mean squared error (RMSE) between the predicted efficiency from GP regression versus ground truth for the unseen data (*i.e.*, the dataset minus the initial samples). (d) Spearman's correlation coefficient between the predicted efficiency from GP regression versus ground truth for the unseen data. The Spearman's coefficient is a metric defined between -1 to 1, examining whether there is a monotonic correlation. In (a), (b), and (c), the symbols indicate the mean of the aggregated results of the 100 separate runs of initial sampling, regression training and prediction. The error bars show the 95% confidence interval. [Both Latin hypercube and random sampling provide more diverse initial sample sets, and LHS also has narrower error bars in the comparison with random sampling.]

Conclusions

In this work, we investigate how we can work cooperatively with an ML algorithm and maximize the acceleration of the optimization process. To do so, we revisit a historical dataset of perovskite devices with open-air spray deposition, and then adopt commonly used sequential learning methods to navigate the process optimization of perovskite solar cells. Based on this case study, we learnt three generalizable points to better utilize machine learning tools for materials research. **First, initial sampling is crucial to ensure rapid optimization.** We compare different initial sampling methods, *e.g.*, chronological, random, and Latin hypercube sequences, and find the initialization method plays an important role to facilitate the subsequent optimization. Without a good initial design of experiments guided by domain knowledge, the acceleration benefits with an ML algorithm is very limited. **Second, the acquisition and objective functions need to be tuned for different research objectives of optimizing device performance or building a predictive**

regression model. By assigning different weights to "exploitation" and "exploration", we evaluate the optimization outcomes for various acquisition functions with different regression models. Echoing previous findings by Rohr *et al.*[7], we also observe that acceleration in building a predictive model is not necessarily easy with sequential learning algorithms, but it could significantly accelerate the process optimization of device efficiency (where up to 2x acceleration could be found). **Third, the algorithm interpretability allows researchers to engage with the ML model and avoid mistakes**. At each step, our framework helps visualize how sampling decisions are made by the ML algorithm: whether an acquisition decision is influenced more by the predicted mean or the predicted uncertainty, whether it favors exploration in the unsampled regions or exploitation in the known regions, or which process variable has the highest impact on the decision. We envision the impact of this work as a human-in-loop machine-learning platform where researchers can understand the framework of ML and utilize ML algorithms to accelerate the research of novel materials.

Acknowledgement

The authors acknowledge the constructive feedbacks from Dr. Armi Tiihonen, Siyu I.P. Tian and other members from the MIT PV research lab. The authors also acknowledge the research funding support for this work from the U.S. Department of Energy (DOE) Solar Energy Technology Office (SETO) Small Innovative Project for Solar (SIPS) program with award number DE-EE0009366. N.R. and R.H.D. acknowledge support from U.S. DOE SETO Photovoltaic Research and Development (PVRD) program with award number DE-EE0008559. Z.R. and T.B. also acknowledge the support from Singapore-MIT Alliance for Research and Technology (SMART), particularly its Independent Research Group of "Low Energy Electronic Systems". SMART is established by the MIT in partnership with the National Research Foundation (NRF) of Singapore

References

[1] H. S. Stein and J. M. Gregoire, "Progress and prospects for accelerating materials science with automated and autonomous workflows," *Chemical Science*, vol. 10, no. 42, pp. 9640–9649, Oct. 2019.

[2] L. T. Schelhas, Z. Li, J. A. Christians, A. Goyal, P. Kairys, S. P. Harvey, D. H. Kim, K. H. Stone, J. M. Luther, K. Zhu, V. Stevanovic, and J. J. Berry, "Insights into operational stability and processing of halide perovskite active layers," *Energy and Environmental Science*, vol. 12, no. 4, pp. 1341–1348, 2019.

[3] J. M. Granda, L. Donina, V. Dragone, D. L. Long, and L. Cronin, "Controlling an organic synthesis robot with machine learning to search for new reactivity," *Nature*, vol. 559, no. 7714, pp. 377–381, 2018.

[4] B. A. Rizkin, A. S. Shkolnik, N. J. Ferraro, and R. L. Hartman, "Combining automated microfluidic experimentation with machine learning for efficient polymerization design," *Nature Machine Intelligence*, vol. 2, no. 4, pp. 200–209, Apr. 2020.

[5] F. Häse, L. M. Roch, C. Kreisbeck, and A. Aspuru-Guzik, "Phoenics: A Bayesian Optimizer for Chemistry," *ACS Central Science*, vol. 4, no. 9, pp. 1134–1145, Sep. 2018.

[6] L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. E. Yunker, J. E. Hein, and A. Aspuru-Guzik, "ChemOS: An orchestration software to democratize autonomous discovery," *PLOS ONE*, vol. 15, no. 4, p. e0229862, Apr. 2020.

[7] B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, and C. P. Berlinguette, "Self-driving laboratory for accelerated discovery of thin-film materials," *Science Advances*, vol. 6, no. 20, p. eaaz8867, May 2020.

[8] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, and A. G. Doyle, "Bayesian reaction optimization as a tool for chemical synthesis," *Nature*, vol. 590, no. 7844, pp. 89–96, Feb. 2021.

[9] M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C. T. Dinh, P. De Luna, Z. Yu, A. S. Rasouli, P. Brodersen, S. Sun, O. Voznyy, C. S. Tan, M. Askerka, F. Che, M. Liu, A. Seifitokaldani, Y. Pang, S. C. Lo, A. Ip, Z. Ulissi, and E. H. Sargent, "Accelerated discovery of CO2 electrocatalysts using active machine learning," *Nature*, vol. 581, no. 7807, pp. 178–183, May 2020.

[10] P. M. Attia, A. Grover, N. Jin, K. A. Severson, T. M. Markov, Y. H. Liao, M. H. Chen, B. Cheong, N. Perkins, Z. Yang, P. K. Herring, M. Aykol, S. J. Harris, R. D. Braatz, S. Ermon, and W. C. Chueh, “Closed-loop optimization of fast-charging protocols for batteries with machine learning,” *Nature*, vol. 578, no. 7795, pp. 397–402, Feb. 2020.

[11] J. Ling, M. Hutchinson, E. Antono, S. Paradiso, and B. Meredig, “High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates,” *Integrating Materials and Manufacturing Innovation*, vol. 6, no. 3, pp. 207–217, Apr. 2017.

[12] B. Rohr, H. S. Stein, D. Guevarra, Y. Wang, J. A. Haber, M. Aykol, S. K. Suram, and J. M. Gregoire, “Benchmarking the acceleration of materials discovery by sequential learning,” *Chemical Science*, vol. 11, no. 10, pp. 2696–2706, Mar. 2020.

[13] F. Häse, M. Aldeghi, R. J. Hickman, L. M. Roch, M. Christensen, E. Liles, J. E. Hein, and A. Aspuru-Guzik, “Olympus: a benchmarking framework for noisy optimization and experiment planning,” *arXiv:2010.04153 [stat.ML]*, Oct. 2020.

[14] “Towards trustable machine learning,” *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 709–710, Oct. 2018.