Article
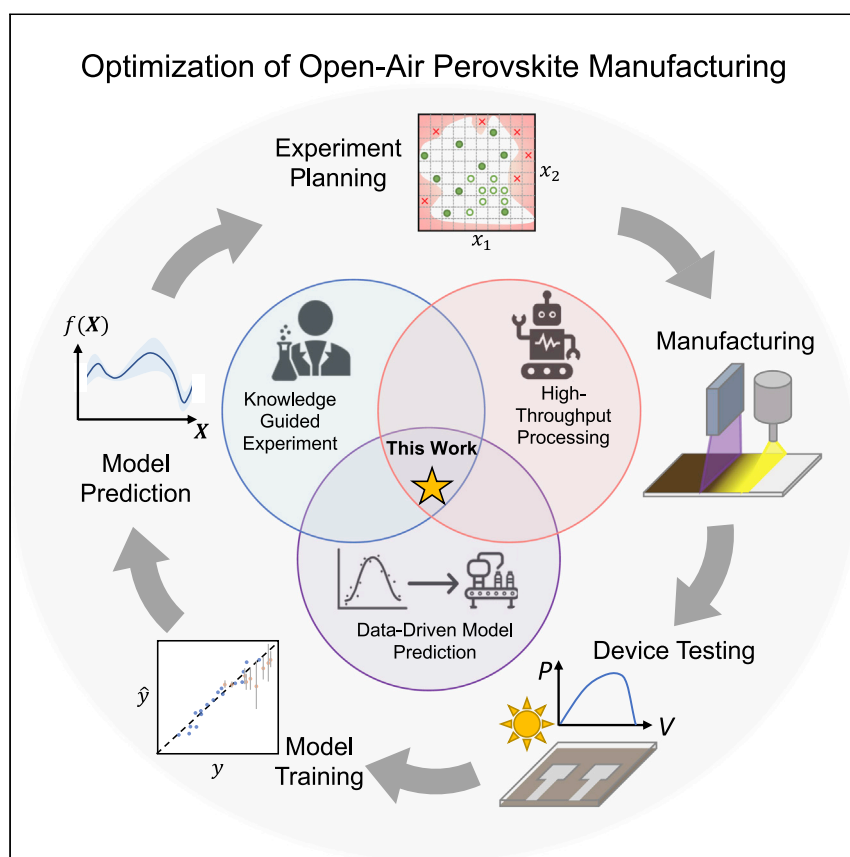
Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing

Optimization of Open-Air Perovskite Manufacturing

Zhe Liu, Nicholas Rolston, Austin C. Flick, Thomas W. Colburn, Zekun Ren, Reinhold H. Dauskardt, Tonio Buonassisi

rhd@stanford.edu (R.H.D.)
buonassi@mit.edu (T.B.)

Highlights

ML accelerates open-air process optimization for perovskite solar cells

With a budget of 100 process conditions, 18.5% device efficiency is achieved

Researchers' domain knowledge can be incorporated into ML process optimization

Benchmarking results show an advantage of ML over traditional methods

Commercialization of perovskite solar cells requires significant efforts to develop scalable manufacturing techniques. Herein, we present a machine learning (ML)-guided framework for the optimization of perovskite devices fabricated by an open-air processing technique. We demonstrate a rapid process optimization route to achieve power conversion efficiency of 18.5%, which is faster than any other conventional researcher-driven design-of-experiment methods. This ML framework can be widely adopted to accelerate perovskite manufacturing scale-up, regardless of processing techniques or device architectures.

Article

Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing

Zhe Liu,[1,4,5] Nicholas Rolston,[2,4] Austin C. Flick,[2] Thomas W. Colburn,[2] Zekun Ren,[3] Reinhold H. Dauskardt,[2,*] and Tonio Buonassisi[1,6,*]

SUMMARY

Developing a scalable manufacturing technique for perovskite solar cells requires process optimization in high-dimensional parameter space. Herein, we present a machine learning (ML)-guided framework of sequential learning for manufacturing the process optimization of perovskite solar cells. We apply our methodology to the rapid spray plasma processing (RSPP) technique for open-air perovskite device fabrication. With a limited experimental budget of screening 100 process conditions, we demonstrated an efficiency improvement to 18.5% as the best result from a device fabricated by RSPP. Our model is enabled by three innovations: flexible knowledge transfer between experimental processes by incorporating data from prior experimental data as a probabilistic constraint, incorporation of both subjective human observations and ML insights when selecting next experiments, and adaptive strategy of locating the region of interest using Bayesian optimization before conducting local exploration for high-efficiency devices. Furthermore, in virtual benchmarking, our framework achieves faster improvements with limited experimental budgets than traditional design-of-experiments methods.

INTRODUCTION

Metal halide perovskites are efficient solar absorbers compatible with low-cost solution processing methods that show promise as an emerging thin-film photovoltaic (PV) technology. Scaling up the fabrication process is currently one of the critical research areas of perovskite technology on the path to commercialization.[1–3] Despite the success of >25% efficient perovskite solar cells in academic labs using spin coating,[4,5] this processing method is not scalable to a manufacturing line. Recently, Rolston et al.[6] demonstrated a high-throughput rapid spray plasma processing (RSPP) method as a scalable open-air fabrication process because it has the potential of achieving low-cost perovskite PV modules at a manufacturing cost of ~$0.2 /W. In addition to cost, the main advantage of this spray-deposition-based technique is its ultrahigh throughput and improved mechanical properties of the thin film in comparison with other scalable processing methods, such as blade coating, slot-die coating, and roll-to-roll printing.

For new scalable perovskite manufacturing processes (including RSPP), it typically takes months to years to achieve process control and reproducibility on the module scale, and several years to estimate the upper potential of the technology. One of the key challenges is that there are many processing parameters to

Context & scale

Perovskite photovoltaics (PVs) have achieved rapid improvement in the past decade for the power conversion efficiency of small-area lab-scale devices. However, successful commercialization still requires the development of low-cost, scalable, and high-throughput manufacturing techniques. Machine learning (ML) for materials science and engineering has been developed in recent years, and it can readily be used to accelerate perovskite manufacturing scale-up.

We demonstrate a Bayesian optimization framework that allows the incorporation of researchers' domain knowledge into the ML-guided loop. In the case of optimizing perovskite solar cells by the open-air rapid spray plasma processing (RSPP) technique, the proposed framework enables a faster optimization in comparison with other conventional researcher-driven design-of-experiment methods. Although it has been shown for RSPP, the ML framework can be broadly used for accelerated development of manufacturing technologies for perovskite PVs.

co-optimize[1,6]—e.g., precursor composition, speed, temperature, head/nozzle height, and curing methods. High-throughput experimentation has been introduced to explore the regions in the high-dimensional parameter space.[7–11] However, in some cases, this high-dimensional optimization problem can be nearly impossible to solve by brute force or even by the sophisticated design of experiments (DoEs).

Sequential machine learning (ML) methods, e.g., Bayesian optimization (BO), have emerged as effective optimization strategies to explore a wide range of chemical reaction synthesis[12–14] and material optimization.[15–25] BO has been shown to work well for optimization problems with under 20 variables[26] or up to 30 variables with some algorithm modifications.[27,28] Therefore, we have chosen to study sequential-learning-based optimization strategies in our current study of RSPP perovskite PV devices.

Current reports of sequential learning studies with classical BO have two common drawbacks: (1) no direct channel to incorporate information from previous relevant studies and (2) inflexibility to adapt researchers' qualitative feedbacks into the iterative loop. On one hand, the ML model sometimes requires a significant amount of data to learn what has already become apparent to the researchers.[29] On the other hand, these drawbacks could discourage material science researchers to adopt ML tools because their domain knowledge could not be utilized in the classical BO iterations.

Previous data and researchers' knowledge are useful information sources for experimental planning. An acquisition function in a BO framework is the "decision-maker" to produce an experimental plan in the next round. Therefore, we can incorporate domain knowledge as probabilistic constraints for the acquisition function (introduced by Gelbart et al.[30]). With the aim of rapid optimization, Sun et al. incorporated density functional theory (DFT) calculations of phase stability as a probabilistic constraint for the experimental acceleration of composition optimization to improve perovskite stability, avoiding compositions susceptible to phase segregation.[31]

Simple information such as a visual assessment of film thickness, color, and structural defects can be a powerful tool that provides additional guidance for intelligent and efficient optimization. In other words, if an experienced researcher identifies a low-quality perovskite film, the subsequent device fabrication is no longer necessary. Defining a probabilistic constraint is an effective and flexible way to incorporate this information into a sequential learning framework.

In this work, we develop a sequential learning framework with probabilistic constraints for rapid process optimization with power conversion efficiency (PCE) as the target variable. As illustrated in Figure 1, the sequential learning framework iteratively learns the process-efficiency relation and suggests new experiments to achieve the optimal target. For a general framework, we start with experimental planning of process conditions with a model-free sampling method for the initial round. Then, perovskite solar cells are fabricated by the RSPP method, and PCE is measured with a solar simulator under standard testing conditions (STCs). With the experimental data of the process parameters and the device PCEs, we train the regression model to learn the process-efficiency relation, and the regression model is subsequently used to predict the PCE (and its prediction uncertainty) for the unsampled regions. Finally, the prediction results are evaluated by an acquisition function together with the constraint information, and therefore, a new round of experiments is planned. By including two additional knowledge constraints (i.e., visual

[1]Deparment of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Department of Materials Science and Engineering, Stanford University, Stanford, CA 94305, USA

[3]Singapore-MIT Alliance for Research and Technology, Singapore 138602, Singapore

[4]These authors contributed equally

[5]Present address: School of Materials Science and Engineering, Northwestern Polytechnical University (NPU), Xi'an, Shaanxi 710072, People's Republic of China

[6]Lead contact

*Correspondence: rhd@stanford.edu (R.H.D.), buonassi@mit.edu (T.B.)
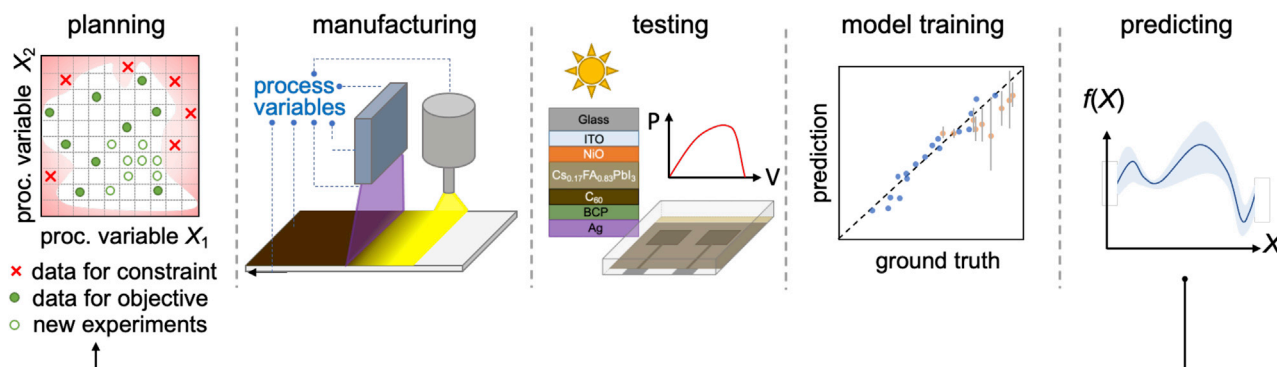
Figure 1. Schematic of sequential learning optimization of perovskite solar cells with probabilistic constraints
The five-step workflow is planning, manufacturing, testing, model training, and prediction. This workflow iterates until the target efficiency is achieved or the maximum experimental budget is reached.

assessment of film quality and previous experimental data from a related study), the optimization framework maps out the parameter space and tends to avoid the less promising regions based on observational film-quality information and prior device data. Hence, sampling suggestions focus on the most promising regions within the parameter space.

To demonstrate the capability of our optimization approach, we consider six key RSPP input variables for the perovskite absorber layer in a device fabrication process. We aimed to exceed the best PCE of solar cells produced with our RSPP. First, we show that efficiency improvements can be obtained with our sequential learning framework, reaching 18.5% efficiency in five experimental iterations. Second, we describe how multiple sources of information were fused into our sequential learning framework as a probabilistic constraint. Third, we analyze the learned relationships between input variables and efficiency, extracting some generalizable insights. Fourth, we benchmark the acceleration factor and enhancement factor of our optimization process against conventional model-free sampling methods for the DoEs sampling methods with optimization simulations, demonstrating an excellent acceleration within the limited experimental budget of fewer than 100 conditions.

RESULTS AND DISCUSSION

Sequential-learning process optimization with improved efficiency
Figure 2A plots the process optimization guided by the sequential learning framework and probabilistic constraints of visual inspection and previous experimental data. PCEs were measured in batches consisting of 20 conditions at a time to enable iteration and model feedback to suggest the subsequent round of process parameters. The highest PCE device of each process condition (the dark-green dots in Figure 2A) was used in the optimization algorithm.

In addition, based on the previous dataset of 45 process conditions (Figure 2C), we defined a top performer with PCE to be a device with PCE above 17% and a good performer to be a device with PCE exceeding 15%. Thus, we had 1 top performer and 6 good performers. Among the 85 of 100 conditions in BO-guided experiments in this work (excluding the low-quality films that did not pass the visual inspection), 45 process conditions achieved >15% PCE (good performers), and 10 process conditions achieved >17% PCE (top performers). The success rate was therefore 47% for the good performers and 12% for the top performers. In contrast, among the Latin
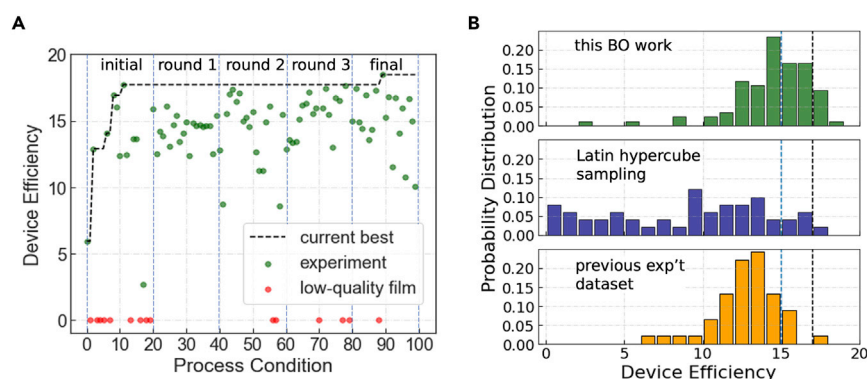
Figure 2. Visualization of the experimental PCE data of solar cells with BO and other methods

(A) The PCEs of the solar cell devices versus experimental process conditions. The solid dark-green dots are the experimental measurements of the highest efficiency among multiple cells at a given condition, and the solid red dots are low-quality films that did not pass the visual inspection. The black line is cumulatively the highest efficiency of all the devices until that specific process condition.

(B) The distribution of power conversion efficiency for three different optimization workflows: BO-guided optimization in this work (85 conditions), Latin hypercube sampling (50 conditions), and a previous experimental dataset by researcher-led optimization (45 conditions). The dashed lines are marked for good performers at 15% PCE (blue) and top performers at 17% PCE (black).

hypercube sampling (LHS)-guided experiments of 50 process conditions (Figure 2B), we found only 6 good performers (12% success rate) and 1 top performer (2% success rate).

Furthermore, the champion process condition produced a best-in-our-lab RSPP device efficiency of 18.5% in fewer than 100 conditions. The champion device in both LHS and previous experiments had never reached 18% PCE. According to a recent review paper on spray-deposited perovskite solar cells,[32] this PCE is comparable with the highest-efficiency devices fabricated by spray deposition in the open air (18.5%)[33] and in the N_2 glove box (19%).[34]

Visual inspection of the perovskite films was done after depositing the perovskite layer with RSPP. The film quality was rated by evaluating color, uniformity, and pinholes. The ratings were intended to be very conservative so that only the lowest-quality films were "tossed." This evaluation step is typical when optimizing with conventional researcher-driven device processing. Example low-quality and high-quality films are shown in Figure S1.

To confirm the validity of our visual criteria, we fabricated additional devices from those low-quality films. We confirmed low PCE values for these conditions (i.e., all below 13.5% with an average PCE of 7.8%), and the corresponding PCE distribution is shown in Figure S2. This validation step confirms that the device fabrication for low-quality films can be skipped. When training our regression model with sequential learning, the film quality ratings were incorporated as a constraint function. Note that only 6 conditions after the first batch produced low-quality films, which further demonstrates that the model sampled the parameter space more effectively over time.

Incorporating knowledge constraints into Bayesian optimization

Figure 3 shows the ML method used in this study. Because a six-dimensional parameter space is difficult to visualize, two parameters of processing speed and substrate
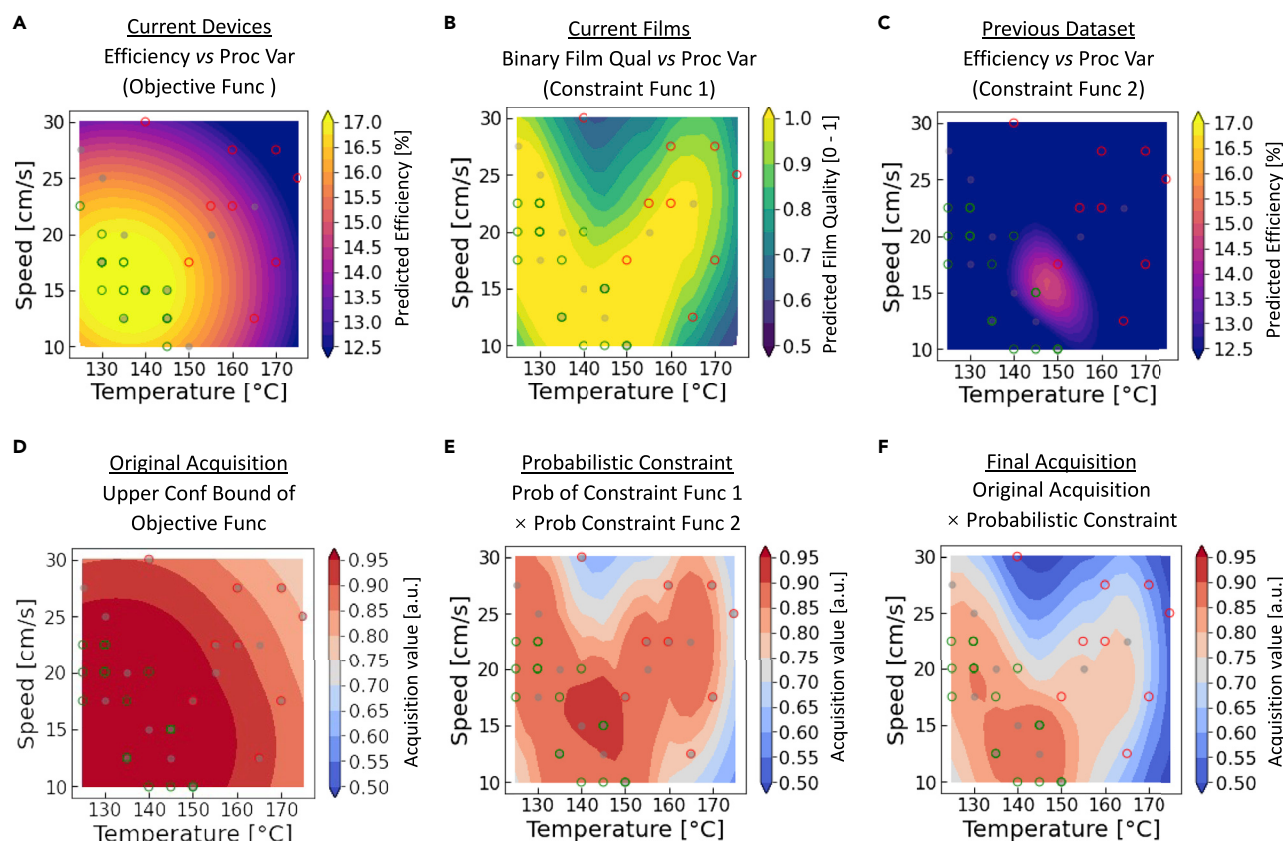
Figure 3. A schematic illustration of probabilistic constraints for the acquisition function in the Bayesian optimization framework

(A–F) The proposed BO framework comprises the process-efficiency relation obtained from the current experimental data (objective function), the binary film quality evaluation (constraint function 1), and the process-efficiency relation from a previous study (constraint function 2). (A)–(C) are the outputs of the three regression models fitted to the respective data. (D)–(F) are the converted acquisition functions based on the trained regression models of (A)–(C).

temperature were chosen as illustrations. A similar plot was obtained for each pair of parameters used in the regression analysis. A regression model was generated for the objective function and two constraint functions. The objective function contains the primary information measured from the experiments by plotting the process-efficiency relationship (Figure 3A). The combination of probabilistic constraints are additional layers included based on visual inspection of film quality (constraint function 1; Figure 3B) and previous experimental results (constraint function 2; Figure 3C).

The resulting acquisition value was determined based on the selection criteria used (i.e., PCE or film quality). These are plotted in Figures 3D–3F and provide the framework for the sequential learning approach, namely how the model learned from each batch to suggest new process conditions with the goal of higher performance. The raw acquisition function in Figure 3D was the upper confidence bound (UCB) of the objective function. Probabilistic constraint 1 is calculated based on the probability of a good-quality film, and probabilistic constraint 2 is calculated based on the probability of a device achieving the above-average PCE in a previous experimental dataset. Both probabilistic constraint functions were scaled to reduce the impact of these constraints and prevent the modification of the acquisition function from being too harsh.
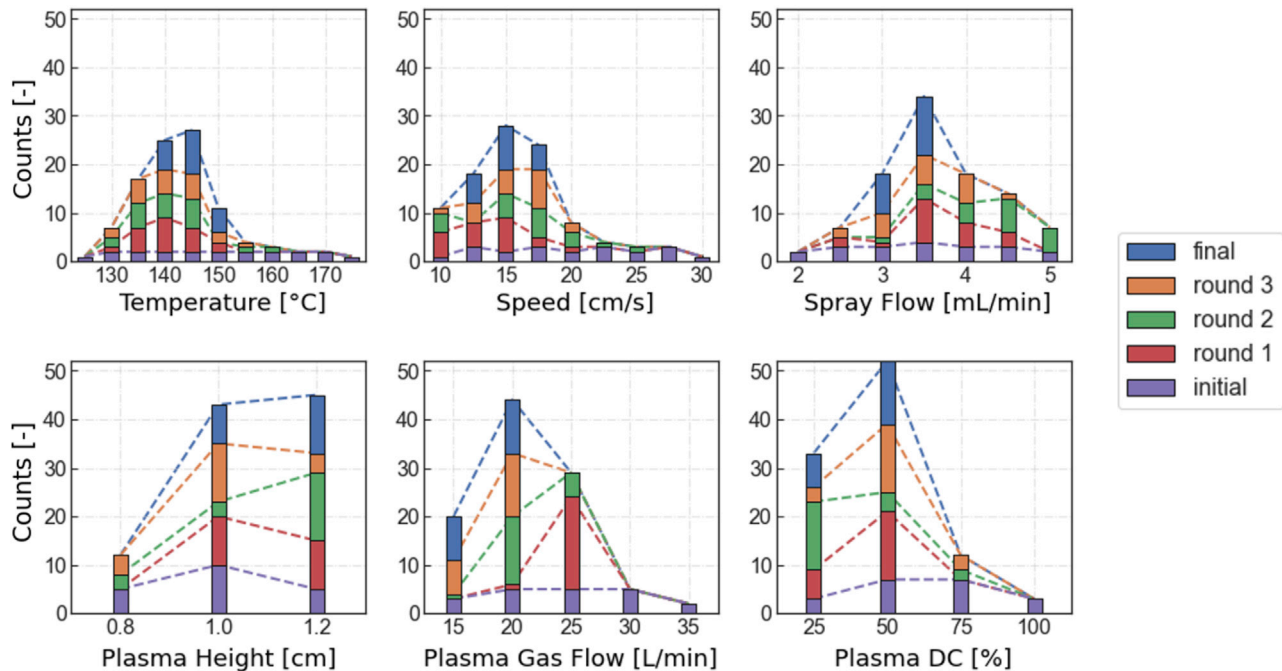
Figure 4. The histogram distribution of process conditions used in each round of the optimization

The initial samples used the LHS method. Rounds 1–3 were the process conditions suggested by Bayesian optimization and subsequently evaluated by experiments. The process converged to a small window after round 3; hence, a subspace with reduced parameter intervals was defined in the final-round optimization.

More specifically, for constraint function 1, the film quality was rated in a binary outcome of either 0 (fail) or 1 (pass). The regression model of film quality was built to interpolate the unsampled regions (see Figure 3E) and sequentially converted to the probability function of passing the film quality assessment. The probabilistic constraint function (i.e., constraint function 1) was then softened to a range of 0.5–1 and multiplied by the raw acquisition function (Figure 3B). See supplemental information section 1.2 for the mathematical definition of probabilistic constraint function. The scaling means that we weigh the device data 2 times more important than the film quality data to avoid any potential bias from our qualitative visual inspection of film quality.

The previous data (that can be found in Figure 1B) for constraint function 2 were acquired during a preliminary optimization of the RSPP setup. We also use these data as a probabilistic constraint instead of a model prior because other variables beyond the six considered herein were slightly modified. For example, the spray and plasma nozzles used in previous experiments were different than those used in this work. Although we believe there is essential knowledge to be transferred (that is not affected too strongly by these modifications), the previous data are not equivalent to the current dataset. This required a more subtle way of representing the information as a probabilistic constraint to the BO framework, instead of directly adding the previous data in the model training.

Iterative visualization of the sample distributions and parameter correlations

Figure 4 visualizes the model learning through iteration as the acquisition function evolved and the process conditions began to converge round by round. Including the initial sampling by LHS, five experimental rounds were conducted for the device
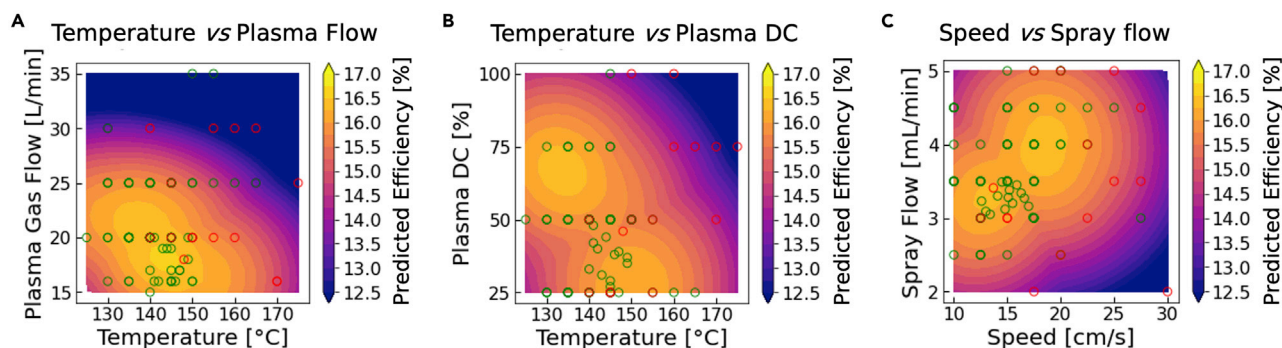
Figure 5. Visualization of the process-efficiency relation based on the trained regression models

The six-dimensional parameter space reduces into two-dimensional contour plots, and the highest efficiency for each condition pair is used for the contour plots. The green open circles are the process conditions made into devices. The red open circles are the process conditions made only into films that failed visual inspection. The color bars indicate the predicted mean efficiency from the trained regression model.

optimization. The experiments in rounds 1–3 followed the suggestions from the BO acquisition function. Figure 4 also shows that the probabilistic constraint began to affect the acquisition from round 1. For example, the higher-temperature and higher-speed regions were sampled less after the initial sampling round because of the probabilistic constraint. This observation is consistent with the probabilistic constraint and acquisition function shown in Figures 3E and 3F.

Due to the limited budget of 100 experimental conditions (or a total of 5 experimental rounds), we opted for a different acquisition method in the final round. Aiming at a further improvement of PCE, we conducted a local optimization in a smaller window of process conditions around the best condition predicted by the regression model. The best condition was found by the particle swarm optimization method,[26,35] which is a common method for the global optimum for any given regression model. The parameter space of process conditions for the final round is shown in Table S2.

The reason for this change is that our Gaussian process (GP) model has the tendency to "smooth out" features in the response surface as shown in Figure 5. A similar finding was found in a previous study in materials science,[17] where the GP fit was also very smooth to avoid potential overfitting to a small number of data points. To increase the probability of finding the optimum value, we use GP as a ringfencing technique to identify a region of the highest probability (i.e., a window of process conditions), rather than a specific point or condition of highest probability. Within this region of highest probability, we combined a few techniques to achieve a balance of exploitation and exploration. Therefore, the final 20 conditions consisted of a best model-predicted condition, nearest neighbors of the best condition for exploitation, and LHS conditions in this region for more exploration.

We visualized the trained regression model of process-efficiency relations by projecting the six-dimensional parameter space into 2D pair-wise contour plots. For each contour plot, we sampled two process variables of interest in a 20 × 20 grid. For every point in the contour plot (i.e., a combination of the two variables), we sampled the remaining four variables 200 times, and we predicted the PCEs using the regression model for the 200 process conditions. Because we were interested in maximizing efficiency, we took the maximum PCE of the 200 process conditions. Therefore, these contour plots are the manifolds of the maximum PCEs in a 2D reduced space. For six process variables, there are a total number of 15 possible
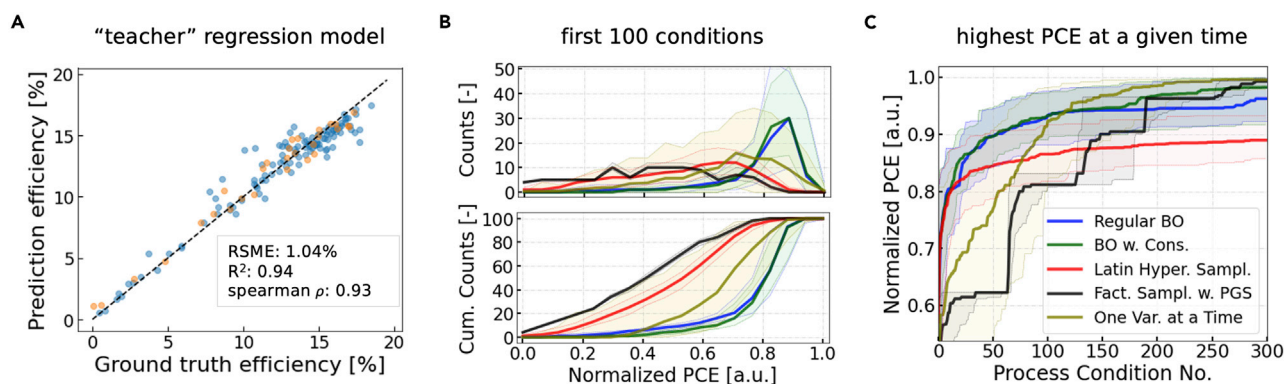
Figure 6. Benchmarking of the newly developed sequential learning framework with the other design-of-experiment methods in the simulated process optimization

(A) Prediction versus ground truth of the trained regression model of gradient boosting with decision trees.

(B and C) (B) The histogram distribution and the cumulative distribution of simulated PCEs of the first 100 process conditions plotted for different sampling methods. (C) The best PCEs at the given number of process conditions for different sampling methods, in the simulated optimizations until the 300th process condition. In (B) and (C), the sampling methods in the comparisons are regular BO (blue), BO with probabilistic constraints (dark green), Latin hypercube sampling (red), factorial sampling with progressive grid subdivision (black), and one-variable-at-at-time sampling (olive); the shaded red area represents the 5%–95% confidence interval of an ensemble of 100 runs with random seeding; all the simulated PCEs are normalized by the highest simulated PCE.

contour plots. Figure 5 shows some examples of the dimension-reduced manifolds. For example, a data point in the contour plot indicates the combination of temperature = 140°C and plasma duty cycle = 20% can achieve a PCE of >16.5% when the other four process variables are fully optimized. Note that the predicted best efficiency by the trained regressor shown in the contour plots is lower than the experimentally measured PCE (i.e., the ground truth) due to the training error of the regression model. This difference is common in BO, but it is still effective because an acquisition function balancing exploitation and exploration compensates for the model error. The training errors after each experimental iteration are found in Figure S4.

The manifold maps also inform correlations between different process variables and their impact on device efficiency. We also projected the suggested experimental conditions onto the contour plots, which helps interpret the decision-making of the algorithm and prevent mistakes at every round. Some correlations in Figure 5 led to new insights of the RSPP method, including a negative correlation between both plasma gas flow/duty cycle and temperature (Figures 5A and 5B). Since a certain curing dose is required to convert the precursor solution to a crystalline perovskite, this result indicated that the higher curing properties of increased plasma gas and duty cycle offset the lower temperature to balance the energy delivered to the perovskite during processing. Additionally, some correlative trends (consistent with previous experimentation) were also observed. For example, higher spray flow correlates with faster speed (see Figure 5C), since a constant precursor dose should be delivered to the substrate to form the desired film thickness.

Benchmarking with design-of-experiment methods using virtual optimization

We conducted a simulated "virtual" optimization with the fully trained "teacher" regression model for a benchmarking comparison between the sequential learning and conventional DoE methods. The "teacher" model is a gradient boosting regression model with decision trees, and it was trained with all the experimental data acquired for this work (i.e., the data from the BO framework and the LHS shown in Figure 2B). The training outcome is shown in Figure 6A. For the "virtual"

Table 1. Optimization ranges for the six process variables (model inputs)

Process variable	Total range (interval)	Process variable	Total range (interval)
Temperature	125°C–175°C (5°C)	plasma height	0.8–1.2 cm (0.2 cm)
Speed	10–30 cm/s (2.5 cm/s)	plasma gas flow	15–35 L/min (5 L/min)
Spray flow	2.0–5.0 mL/s (0.5 mL/min)	plasma duty cycle	25%–100% (25%)

optimization, a good "teacher" model only needs to ensure the predictions follow a monotonic correlation with the ground truth (i.e., the actual efficiency measured from experimentation). Therefore, following the study by Häse et al.,[36] the spearman coefficient was used as an evaluation metric of the model training to assess the strength of the monotonic correlation between predicted and ground-truth efficiencies. This "teacher" regression model approximates the ground-truth manifold with a spearman coefficient of 0.93, and this model was used as a "teacher" model to run simulated optimization. The use of a simulation tool in benchmarking analysis avoids the time-consuming experimental work and has the advantage of generating better statistics for random or pseudorandom processes. However, because of the discrepancy of the absolute values between predictions and ground truths, we normalized the PCE values in the simulated optimization.

As mentioned in the "ML methods" section, the full factorial sampling of all the process conditions in Table 1 is more than 40 k samples, which is not feasible to be explored experimentally. The use of partial factorial DoE methods (e.g., generalized subset design,[37] or D-optimal design[38]) could reduce the sample size to 1/2 or 1/3, but the number of samples is still too large (i.e., greater than 10 k). These methods are not fair comparisons with the BO framework. Alternatively, we used three model-free sampling methods: LHS, one-variable-at-a-time sampling (OVATS), and factorial sampling with progressive grid subdivision (FS-PGS). The LHS, OVATS, and FS-PGS methods are illustrated in detail in the sections 3.2–3.4 of supplemental information. The results of different sampling methods are compared in Figures 6B and 6C, together with the BO framework. Figure 6B shows the PCE distributions for a limited experimental budget of 100 process conditions, whereas Figure 6C shows the evaluation of the best PCE at the given number of process conditions.

First, LHS is a popular model-free sampling method and produces a low-discrepancy pseudorandom sampling sequence (in contrast to the random sampling method). In this case, the outcome of the experiments does not inform the choice of the next round of samples. We ran 100 times and obtained the 5%–95% confidence interval of the statistical distributions. LHS is one of the most commonly used baseline comparison in BO studies,[36] but it is not the best experimental practice in a research lab. Instead, we also implemented more realistic methods used in an experimental optimization, such as, OVATS, and FS-PGS.

Second, OVATS is one of the most popular methods for experimental planning by academic researchers.[39] Therefore, we believe that OVATS is suitable to benchmark ML methods against status quo. In this simulation, we started with one variable to screen and the other variables being constant, and we then fixed the optimal condition of the first variable and screened the second variable orthogonally. We kept iteratively screening the variables until the device efficiency did not further improve. When the efficiency stopped improving, we subdivided the grid and continued to screen more conditions per variables. The batch size varies with the number of levels for each variable. We randomized the choice of initial variable and the sampling sequence, and we ran the methods for 100 times to obtain the 5%–95% confidence interval.

Third, FS-PGS is another method of a sequential DoE using factorial sampling, which is similar to the sampling concept proposed in the studies conducted by Kurchin et al.[40] and Eriksson et al.[41] Two-level grids of process variables were divided and sampled sequentially at every round in FS-PGS, and the parameter space was refined based on the process condition of the highest PCE found in the prior samples. Through this method, the process conditions were exploited locally around the highest efficiency with only a small fraction of the full factorial combinations. The batch size is 64 samples for the initial round (i.e., 2^6) and 63 samples subsequently (i.e., $2^6 - 1$). We randomized the sample sequence for 100 times in each batch and produced the 5%–95% confidence interval.

To understand the manifold of the parameter space, we ran the full-grid sampling methods and 100 k random sampling with the "teacher" model. The PCE distributions are shown in Figure S9. We found that we were looking at a "needle-in-a-haystack" problem, where the top 5 percentile mark is at 0.8 in the normalized PCE, the top 1 percentile mark is at 0.85 in the normalized PCE, and the top 0.1 percentile mark is at 0.9 in the normalized PCE (as shown in Figure S9C). We will use these marks to evaluate the performance of different experimental planning methods.

Figure 6B compares the PCE distributions for different sampling methods, given a limited experimental budget of 100 process conditions. We observe that both the regular BO and the BO with knowledge constraints achieved more high-efficiency devices in the early runs, and both were able to find 50 conditions at the top 1 percentile mark (i.e., 0.8 in the normalized PCE). OVATS found 19 conditions exceeding 0.8 in the normalized PCE. LHS and FS-PGS has less than 5 conditions exceeding 0.8 in the normalized PCE. Therefore, in the task of finding the conditions of the top-one-percentile PCE, the BO methods have much higher probability of success, with an enhancement factor of 2.6× to OVATS, and more than 10× to LHS and FS-PGS.

The highest PCE versus the number of conditions are then compared in Figure 6C. We see that, within the first 100 conditions, the BO methods performed significantly better in improving the PCE than OVATS, LHS, and FS-PGS. Because a few conditions were screened out by film quality as a probabilistic constraint at the first 20 conditions, BO with knowledge constraints was found to be slightly advantageous over the regular BO (and pure LHS).

In the comparison with OVATS and FS-PGS, we can see LHS is an efficient method for sampling, exceeding 0.8 in the normalized PCE (although OVATS and FS-PGS are below 0.7). However, the slope of improvement of LHS method flattened after the first 20 conditions. The BO methods reached 0.9 in the normalized PCE after 50 conditions, whereas LHS reached 0.85, and OVATS and FS-PGS were still below 0.8.

These observations in Figures 6B and 6C indicate that the BO methods can help us to get closed to optimum very quickly during process optimization because it exhibits an impressive acceleration at the small numbers of process conditions explored. However, the acceleration factor begins to decline at higher process conditions, indicating a decreasing rate of improvement of discovery of the highest PCE conditions. For example, at the 0.8 mark in the normalized PCE, the acceleration factor of the BO methods with LHS initial sampling is more than 5.5× over the OVATS method (i.e., 55/10); at the 0.9 mark in the normalized PCE, the acceleration factor of the BO methods is approximately 1.9× over the OVATS method (i.e., 95/50).

We also simulated the scenarios beyond the 100 conditions for a deeper analysis of different sampling methods. We observed the following points in Figure 6C: (1) after 120 conditions, OVATS started overtaking the BO methods. This echoes the necessity of switching to a local search or an exploitative strategy in a later stage of the optimization. (2) The BO with the knowledge constraint method started performing better than the regular BO method after the 160 conditions. (3) The FS-PGS method caught up to the BO methods after the 190 conditions. (4) OVATS and FS-PGS were able to get very closed the optimal PCE toward the end of the 300 conditions, whereas the BO with knowledge constraint method exceeds 0.98 in the normalized PCE, and the regular BO method exceeds 0.96 in the normalized PCE. (5) The OVATS has a significantly larger spread in the 5%–95% confidence interval than the other methods, especially the lower bound of 5 percentile is much lower than the median for the first 100 conditions.

The decreasing rate of improvement in BO suggests that the future improvement of sequential learning may involve adaptive acquisition strategies, which switch between different acquisition methods based on what the algorithm deems to be advantageous. This adaptive acquisition may be implemented through combined acquisition functions or switching to another sampling method. The finding is consistent with our experimental optimization in this work, where the acquisition strategy was switched the final round to increase the chance of the PCE improvement. The first attempt toward this has been implemented in the optimization study of Ag nanoparticle fabrication,[17] where a neural network (NN) model was run in parallel with the GP, to improve regression resolution near the region of maximum interest. In that study, however, the NN was not used to inform the acquisition function, as would be the case here.

Conclusions

In this work, a new sequential learning framework—BO with knowledge constraints—was used for the process optimization of perovskite solar cells, which intelligently incorporated the previous data from preliminary optimization experiments and researchers' visual evaluation of perovskite film quality. The workflow of the BO framework with knowledge constraints effectively mimics the "conventional" iterative optimization method and allows for the flexible incorporation of multiple information sources during process optimization. Compared with regular BO, the concept of knowledge constraints shows two key advantages: (1) it provides a flexible way to incorporate the knowledge learned in the past (or qualitative inputs from an experienced researcher) and (2) it reduces the extra experimental efforts of converting the film into devices when the film quality is visibly poor.

With the experimental planning guided by the BO framework with knowledge constraints, we achieved an 18.5%-efficient solar cell in open-air after optimizing six process variables for perovskite deposition, conducting 5 experimental rounds and screening 100 process conditions. This is the highest PCE achieved to date with the RSPP method and done with only a very small fraction of the more than 40 k possible combinations in a grid search (also known as full factorial sampling). In addition to the best-in-our-lab PCE values, we also discovered 10 process conditions that produced ≥17% PCE devices with BO, significantly outperforming the pseudo-random sampling method of LHS in a parallel run where only 1 of 50 process conditions was above 17% PCE. This indicates a >5× enhancement in finding the process conditions producing the top-performing solar cells.

To benchmark the performance of the BO frameworks (i.e., both regular BO and BO with knowledge constraints), we trained a teacher regression model to produce the

"ground truth" (i.e., simulated experiment results), and we ran virtual optimizations with three common sampling methods for experimental planning, namely LHS, OVATS method, and FS-PGS. We also found that both BO methods significantly overperformed the OVATS method within the first 100 process conditions, but the OVATS method overtook the BO methods afterward. This shows that the true advantage of the BO framework is getting close to the optimum quickly within a small budget of experiments, and further development of adaptive acquisition strategies (e.g., time-evolving acquisition function) might be needed.

In summary, this framework of BO with knowledge constraints could be widely applied in process optimization and materials screening for perovskite PV devices, regardless of deposition methods (e.g., spin coating, blade coating, slot-die coating, or thermal evaporation). Moreover, the knowledge constraint is also flexible to incorporate quantitative film quality measurements, such as light-beam-induced current mapping, photoluminescence imaging, and quasifermi level splitting mapping.[42] This will be particularly helpful to tackle the challenge of perovskite PV scale-up by incorporating the learning from lab-scale devices into the optimization of large-area modules.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact
Further information and requests for resources and materials should be directed to, and will be fulfilled by the lead contact, Tonio Buonassisi (bounassisi@mit.edu)

Materials availability
This study did not generate any unique materials.

Data and code availability
The codes and the datasets used for conducting BO with probabilistic constraints are available in GitHub repository: https://github.com/PV-Lab/SL-PerovskiteOpt.

Solar cell fabrication
The details of solar cell fabrication with an RSPP tool have been discussed elsewhere.[6,42,43,44] Here is a brief overview of the fabrication process. A nickel oxide (NiO_x) layer with a thickness of ~20 nm was deposited by spin coating onto an indium tin oxide (ITO) coated glass substrate (Xin Yan Technologies, 10 Ω/□). The precursor used for NiO_x deposition was $Ni(NO_3)_2$ in 94% vol. ethylene glycol with 6% vol. of ethylenediamine. A perovskite absorber layer of $Cs_{0.17}FA_{0.83}PbI_3$ was then deposited with RSPP using an impact ultrasonic spray nozzle (Sonotek) connected in-line to a slot-type atmospheric plasma jet (Plasmatreat). The basic working principle is to uniformly deposit the precursor solution with the spray nozzle onto a heated glass substrate followed by brief exposure to the afterglow of a nitrogen plasma to achieve rapid solidification and form a polycrystalline thin film. The precursors of cesium iodide (CsI, Sigma-Aldrich), formamidinium iodide (FAI, Sigma-Aldrich), and lead iodide (PbI_2, TCI) were dissolved in the mixed solvent of anhydrous N,N-dimethylformamide (DMF, Acros), and dimethyl sulfoxide (DMSO, Acros) with a volume ratio of 1:2. The molar concentration of the precursors is 0.15 M. Subsequently, a fullerene (C60, MER Corp) layer with a thickness of 45 nm, bathocuproine (BCP, Sigma-Aldrich) layer with a thickness of 7.5 nm, and silver (Ag) electrode with a thickness of 150 nm was deposited by thermal evaporation (Angstrom Amod evaporator) in sequence. The area of each perovskite solar cell is 0.21 cm^2, which is defined by the stencil mask during Ag deposition. This process flow results in a

p-i-n solar cell structure, with seven-layer stacks illuminated through the substrate: glass/ITO/NiO$_x$/Cs$_{0.17}$FA$_{0.83}$PbI$_3$/C$_{60}$/BCP/Ag. The current-voltage (I-V) characteristics of the solar cell devices were characterized under the STC, i.e., at 25°C under the illumination of an AM1.5G solar spectrum. The I-V curves for each device were measured from 1.2 to −0.1 V (a reverse scan) at 0.05 V intervals with a 0.1 s delay at each voltage step. For measurement consistency, there was no preconditioning (e.g., light illumination, voltage bias) before the I-V scans were taken.

Machine learning method
The BO method was used for process optimization. As inputs, we selected six RSPP process variables that affect perovskite absorber layer deposition: substrate temperature (°C), linear speed of the spray and plasma nozzles (cm/s), spray flow rate of precursor liquid (mL/min), gas flow rate into plasma nozzle (L/min), height of plasma nozzle (cm), and plasma duty cycle (%, corresponding to the ratio of time that the plasma receives DC power). The target variable (i.e., output) was solar cell PCE. The optimization bounds were chosen as shown in Table 1. Given the sampling intervals, a full-grid sampling can result in 41,580 unique process conditions.

These six process variables and their optimization bounds were chosen based on historical data of more than 300 process conditions (in a similar setup with different spray and plasma nozzles). The less explored variable in the past was set to a large range, e.g., plasma duty cycle. We have given careful expert consideration when defining this parameter space for optimization, intending to balance experimental feasibility and broadness of the parameter space. The researcher-defined parameter space typically works very well in Bayesian-optimization-guided experimental work.[22]

Sequential learning for process optimization has five steps: initial sampling, regression, prediction, evaluation, and acquisition. The initial round was done by LHS, which is a pseudorandom method that samples the parameter space relatively uniformly in high dimensions. The surrogate model was GP regression with anisotropic Matern52. The choice of GP affords slightly better extrapolation than step-wise regressors like random forest regression;[45] the anisotropic kernel allows the hyperparameters of each input variable to be independently tuned, so that some variables can have a stronger impact on the predicted output variable (i.e., PCE) than others.[46]

The acquisition function consists of two parts. The first part is the objective function (i.e., process-efficiency relation). We used the utility function of UCB[47] to inform which process conditions to acquire in the next batch. The second part is the probabilistic constraint for the acquisition function. The local penalization method[48] was used to select the batch of new data acquisitions. The idea is to generate an envelope function that reduces the acquisition value around the location of sample n in the batch when evaluating sample n + 1. See supplemental information section 2.2 for the experimental cost analysis to determine the optimal batch size (i.e., number of samples in each batch). We used the package of BO in Python called *Emukit*,[49] and the GP package called *GPy*.[50]

To incorporate the film quality information and the dataset from the relevant previous experiments, we defined two probabilistic constraints for the acquisition. By this method, we will sample fewer in the region where film quality is poor, or low PCE conditions in previous experiments. To be conservative, we further softened the probabilistic constraint (i.e., constraint function 2) to a range of 0.8–1. The weighting factors were chosen to make the current data 5 times more important than the

previous data. The knowledge from the previous data was transferred by multiplying the acquisition function with constraint function 2. Note that, because of the difference in scaling factor, constraint function 2 is much weaker than constraint function 1 (i.e., 2.5 times difference). This difference reflects our perspective on how much the film quality and the previous data should influence the acquisition. The film quality should be much more relevant in this optimization than the previous device data.

AUTHOR CONTRIBUTIONS

Z.L., N.R., R.H.D., and T.B. contributed to the project ideation, the initial planning, the data analysis, and the paper revision. Z.L. led the development and execution of the sequential learning framework with probabilistic constraints with Z.R.'s contributions. N.R. led the experimental fabrication and device characterization with contributions from A.C.F. and T.W.C. Z.L. and N.R. coled the project coordination and the manuscript writing (with the inputs from all the co-authors). R.H.D. and T.B. provided guidance during the project execution and the manuscript writing.

DECLARATION OF INTERESTS

Some of the authors (Z.L., Z.R., and T.B.) hold equity in a start-up company, Xinterra, focused on commercializing ML technologies for accelerated materials development. To ensure equal access for all, materials from this study (e.g., code, datasets) are open sourced. One of the authors (T.B.) is a member of the journal's advisory board.

REFERENCES

1. Li, Z.C., Klein, T.R., Kim, D.H., Yang, M., Berry, J.J., van Hest, M.F.A.M., and Zhu, K. (2018). Scalable fabrication of perovskite solar cells. Nat. Rev. Mater. *3*, 1–20.

2. Perini, C.A.R., Doherty, T.A.S., Stranks, S.D., Correa-Baena, J.-P., and Hoye, R.L.Z. (2021). Pressing challenges in halide perovskite photovoltaics—from the atomic to module level. Joule *5*, 1024–1030.

3. Li, D., Zhang, D., Lim, K.-S., Hu, Y., Rong, Y., Mei, A., Park, N.-G., and Han, H. (2021). A review on scaling up perovskite solar cells. Adv. Funct. Mater. *31*, 2008621.

4. Almora, O., Baran, D., Bazan, G.C., Berger, C., Cabrera, C.I., Catchpole, K.R., Erten-Ela, S., Guo, F., Hauch, J., Ho-Baillie, A.W.Y., et al. (2021). Device performance of emerging photovoltaic materials. Adv. Energy Mater. *11*, 2002774.

5. Yoo, J.J., Seo, G., Chua, M.R., Park, T.G., Lu, Y., Rotermund, F., Kim, Y.-K., Moon, C.S., Jeon, N.J., Correa-Baena, J.-P., et al. (2021). Efficient perovskite solar cells via improved carrier management. Nature *590*, 587–593.

6. Rolston, N., Scheideler, W.J., Flick, A.C., Chen, J.P., Elmaraghi, H., Sleugh, A., Zhao, O., Woodhouse, M., and Dauskardt, R.H. (2020). Rapid open-air fabrication of perovskite solar modules. Joule *4*, 2675–2692.

7. Ahmadi, M., Ziatdinov, M., Zhou, Y., Lass, E.A., and Kalinin, S.V. (2021). Machine learning for high-throughput experimental exploration of metal halide perovskites. Joule *5*, 2797–2822.

8. Zhao, Y., Zhang, J., Xu, Z., Sun, S., Langner, S., Hartono, N.T.P., Heumueller, T., Hou, Y., Elia, J., Li, N., et al. (2021). Discovery of temperature-induced stability reversal in perovskites using high-throughput robotic learning. Nat. Commun. *12*, 2191.

9. Du, X., Lüer, L., Heumueller, T., Wagner, J., Berger, C., Osterrieder, T., Wortmann, J., Langner, S., Vongsaysy, U., Bertrand, M., et al. (2021). Elucidating the full potential of OPV materials utilizing a high-throughput robot-based platform and machine learning. Joule *5*, 495–506.

10. Saliba, M. (2019). Polyelemental, multicomponent perovskite semiconductor libraries through combinatorial screening. Adv. Energy Mater. *9*, 1803754.

11. Zhao, Y., Heumueller, T., Zhang, J., Luo, J., Kasian, O., Langner, S., Kupfer, C., Liu, B., Zhong, Y., Elia, J., et al. (2022). A bilayer conducting polymer structure for planar perovskite solar cells with over 1,400 hours operational stability at elevated temperatures. Nat. Energy *7*, 144–152.

12. Burger, B., Maffettone, P.M., Gusev, V.V., Aitchison, C.M., Bai, Y., Wang, X., Li, X., Alston, B.M., Li, B., Clowes, R., et al. (2020). A mobile robotic chemist. Nature *583*, 237–241.

13. Granda, J.M., Donina, L., Dragone, V., Long, D.L., and Cronin, L. (2018). Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature *559*, 377–381.

14. Shields, B.J., Stevens, J., Li, J., Parasram, M., Damani, F., Alvarado, J.I.M., Janey, J.M., Adams, R.P., and Doyle, A.G. (2021). Bayesian reaction optimization as a tool for chemical synthesis. Nature *590*, 89–96.

15. Gongora, A.E., Xu, B., Perry, W., Okoye, C., Riley, P., Reyes, K.G., Morgan, E.F., and Brown, K.A. (2020). A Bayesian experimental autonomous researcher for mechanical design. Sci. Adv. *6*, eaaz1708.

16. MacLeod, B.P., Parlane, F.G.L., Morrissey, T.D., Häse, F., Roch, L.M., Dettelbach, K.E., Moreira, R., Yunker, L.P.E., Rooney, M.B., Deeth, J.R., et al. (2020). Self-driving laboratory for accelerated discovery of thin-film materials. Sci. Adv. *6*, eaaz8867.

17. Mekki-Berrada, F., Ren, Z., Huang, T., Wong, W.K., Zheng, F., Xie, J., Tian, I.P.S., Jayavelu, S., Mahfoud, Z., Bash, D., et al. (2021). Two-step machine learning enables optimized nanoparticle synthesis. npj Comput. Mater. *7*, 55.

18. Attia, P.M., Grover, A., Jin, N., Severson, K.A., Markov, T.M., Liao, Y.H., Chen, M.H., Cheong, B., Perkins, N., Yang, Z., et al. (2020). Closed-loop optimization of fast-charging protocols for batteries with machine learning. Nature *578*, 397–402.

19. Lookman, T., Balachandran, P.V., Xue, D., and Yuan, R. (2019). Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. npj Comput. Mater. *5*, 21.

20. Nikolaev, P., Hooper, D., Webber, F., Rao, R., Decker, K., Krein, M., Poleski, J., Barto, R., and Maruyama, B. (2016). Autonomy in materials research: a case study in carbon nanotube growth. npj Comput. Mater. *2*, 1–6.

21. Balachandran, P.V., Kowalski, B., Sehirlioglu, A., and Lookman, T. (2018). Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. Nat. Commun. *1*, 1668.

22. Ling, J., Hutchinson, M., Antono, E., Paradiso, S., and Meredig, B. (2017). High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. Integr. Mater. Manuf. Innov. *6*, 207–217.

23. Rohr, B., Stein, H.S., Guevarra, D., Wang, Y., Haber, J.A., Aykol, M., Suram, S.K., and Gregoire, J.M. (2020). Benchmarking the acceleration of materials discovery by sequential learning. Chem. Sci. *11*, 2696–2706.

24. Erps, T., Foshey, M., Luković, M.K., Shou, W., Goetzke, H.H., Dietsch, H., Stoll, K., von Vacano, B., and Matusik, W. (2021). Accelerated discovery of 3D printing materials using data-driven multiobjective optimization. Sci. Adv. *7*, eabf7435.

25. Tran, A., Tranchida, J., Wildey, T., and Thompson, A.P. (2020). Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: application to ternary random alloys. J. Chem. Phys. *153*, 074705.

26. James, V., and Miranda, L. (2018). PySwarms: a research toolkit for particle swarm optimization in Python. J. Open Source Softw. *3*, 433.

27. Wang, Z., Gehring, C., Kohli, P., and Jegelka, S. (2018). Batched large-scale Bayesian optimization in high-dimensional spaces. In Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 745–754.

28. Harris, S.J., Harris, D.J., and Li, C. (2017). Failure statistics for commercial lithium ion batteries: a study of 24 pouch cells. J. Power Sources *342*, 589–597.

29. Ziatdinov, M.A., Ghosh, A., and Kalinin, S.V. (2022). Physics makes the difference: Bayesian optimization and active learning via augmented Gaussian process. Mach. Learn. Sci. Technol. *3*, 015022.

30. Gelbart, M.A., Snoek, J., and Adams, R.P. (2014). Bayesian optimization with unknown constraints. Preprint at arXiv. 1403.5607.

31. Sun, S., Tiihonen, A., Oviedo, F., Liu, Z., Thapa, J., Zhao, Y., Hartono, N.T.P., Goyal, A., Heumueller, T., Batali, C., et al. (2021). A data fusion approach to optimize compositional stability of halide perovskites. Matter *4*, 1305–1322.

32. Bishop, J.E., Smith, J.A., and Lidzey, D.G. (2020). Development of spray-coated perovskite solar cells. ACS Appl. Mater. Interfaces *12*, 48237–48245.

33. Su, J., Cai, H., Yang, J., Ye, X., Han, R., Ni, J., Li, J., and Zhang, J. (2020). Perovskite ink with an ultrawide processing window for efficient and scalable perovskite solar cells in ambient air. ACS Appl. Mater. Interfaces *12*, 3531–3538.

34. Ding, J., Han, Q., Ge, Q.Q., Xue, D.J., Ma, J.Y., Zhao, B.Y., Chen, Y.X., Liu, J., Mitzi, D.B., and Hu, J.S. (2019). Fully air-bladed high-efficiency perovskite photovoltaics. Joule *3*, 402–416.

35. Kennedy, J., and Eberhart, R. (1995). Particle swarm optimization. Proceedings of ICNN'95—International Conference on Neural Networks, *4*, pp. 1942–1948.

36. Häse, F., Aldeghi, M., Hickman, R.J., Roch, L.M., Christensen, M., Liles, E., Hein, J.E., and Aspuru-Guzik, A. (2021). Olympus: a benchmarking framework for noisy optimization and experiment planning. Mach. Learn. Sci. Technol. *2*, 035021.

37. Surowiec, I., Vikström, L., Hector, G., Johansson, E., Vikström, C., and Trygg, J. (2017). Generalized subset designs in analytical chemistry. Anal. Chem. *89*, 6491–6497.

38. *NIST/SEMATECH e-Handbook of Statistical Methods*, Section 5.5.2.1 D-Optimal Desings. https://doi.org/10.18434/M32189, Accessed on Mar 30, 2022.

39. Cao, B., Adutwum, L.A., Oliynyk, A.O., Luber, E.J., Olsen, B.C., Mar, A., and Buriak, J.M. (2018). How to optimize materials and devices via design of experiments and machine learning: demonstration using organic photovoltaics. ACS Nano *12*, 7434–7444.

40. Kurchin, R., Romano, G., and Buonassisi, T. (2019). Bayesim: a tool for adaptive grid model

CellPress

fitting with Bayesian inference. Comput. Phys. Commun. *239*, 161–165.

41. Eriksson, D., Ai, U., Pearce, M., Gardner, J.R., Turner, R., and Poloczek, M. (2019). Scalable global optimization via local Bayesian optimization. In 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), pp. 5496–5507.

42. Stolterfoht, M., Wolff, C.M., Márquez, J.A., Zhang, S., Hages, C.J., Rothhardt, D., Albrecht, S., Burn, P.L., Meredith, P., Unold, T., and Neher, D. (2018). Visualization and suppression of interfacial recombination for high-efficiency large-area pin perovskite solar cells. Nat. Energy *3*, 847–854.

43. Hilt, F., Hovish, M.Q., Rolston, N., Brüning, K., Tassone, C.J., and Dauskardt, R.H. (2018). Rapid route to efficient, scalable, and robust

perovskite photovoltaics in air. Energy Environ. Sci. *11*, 2102–2113.

44. Scheideler, W.J., Rolston, N., Zhao, O., Zhang, J., and Dauskardt, R.H. (2019). Rapid aqueous spray fabrication of robust NiO$_x$: a simple and scalable platform for efficient perovskite solar cells. Adv. Energy Mater. *9*, 1803600.

45. Lim, Y.-F., Ng, C.K., Vaitesswar, U.S., and Hippalgaonkar, K. (2021). Extrapolative Bayesian optimization with Gaussian process and neural network ensemble surrogate models. Adv. J. Intell. Syst. *3*, 2100101.

46. Liang, Q., Gongora, A.E., Ren, Z., Tiihonen, A., Liu, Z., Sun, S., Deneault, J.R., Bash, D., Mekki-Berrada, F., Khan, S.A., et al. (2021). Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. npj Comput. Mater. *7*, 118.

47. Srinivas, N., Krause, A., Kakade, S.M., and Seeger, M.W. (2012). Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. IEEE Trans. Inf. Theor. *58*, 3250–3265.

48. González, J., Dai, Z., Hennig, P., and Lawrence, N. (2016). Batch Bayesian optimization via local penalization. Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 648–657.

49. Paleyes, A., Pullin, M., Mahsereci, M., Lawrence, N., and González, J. (2021). Emulation of physical processes with Emukit. Preprint at arXiv. arXiv:2110.13293.

50. GPy. (2012). GPy: A Gaussian Process Framework in Python. https://github.com/SheffieldML/GPy.