**Concise summary of the machine learning methodology**

In this work, we employ a sequential machine learning approach based on Bayesian optimization (BO) to optimize the power conversion efficiency (PCE) of perovskite thin-film photovoltaic devices by varying process conditions. In BO for process optimization, one attempts to find the specific process condition combination, , for which the black-box objective function, , mapping the condition search space (X), to the target property (PCE), is maximal,1

|  |  |
| --- | --- |
|  | (1) |

Such a BO campaign typically starts by selecting an initial sample of process conditions, after which the output property of interest is determined for each of those. Subsequently, a surrogate model – the prior – is trained based on the observations made, and an acquisition function is constructed, which calculates the utility of measuring the PCE at the point *x*. Utility may be measured in a number of ways: the predicted PCE value, the amount of information this new point will provide the surrogate model, the likelihood this point will improve upon the current maximum, etc.2 Once new condition combinations have been acquired, the surrogate model is updated – resulting in the posterior – and acquisition can be repeated. This updating process can be repeated in an iterative manner until significant improvements in PCE are no longer observed.

As a preliminary step in this work, a tentative search space of process conditions for PCE optimization was defined based on expert considerations, with a particular aim to balance experimental feasibility and broadness of the search/parameter space (cf. Table 1). In total, 1.3 billion distinct combinations of process parameters can be distinguished within this space.

|  |  |
| --- | --- |
| Process variable | Total range (interval) |
| Volume NMP (ml) | 0 – 100 (5) |
| Volume DMF (ml) | 0 – 100 (5) |
| Volume DMSO (ml) | 0 – 100 (5) |
| Perovskite concentration (M) | 0.8 – 1.8 (0.1) |
| Annealing temperature (°C) | 100 – 160 (5) |
| Vacuum pressure (Pa) | 20 – 400 (10) |
| Vacuum pressure time (s) | 0 – 55 (2) |

**Table 1**: An overview of the search space considered in this work.

As an initial test to ensure the appropriateness of the selected search space, an exploratory sample of 45 process condition combinations were selected, and PCE values were determined crudely, i.e., a single device was synthesized, and its efficiency measured, for each combination. Significant variations in the measured PCE could be observed across this sample (from 5.6% up to 21.9%), providing confidence that the selected process variables will indeed enable extensive modulation of the efficiency. Consequently, the selected set of process conditions was retained for the actual optimization campaign.

With the search space defined and validated, the BO campaign was initiated. We decided to sample 20 points from the search space at a time, because the most time-consuming step of the experimental setup, i.e., the thermal evaporation, can accommodate up to 20 substrates per batch. As such, this specific batch size can be expected to be the most cost-effective, as it balances the time cost per individual measurement and the feedback/update frequency of the regression model.

The initial sampling of the search space was performed manually, informed by the results from the preliminary screening, and with the aim of covering the ranges of the individual process variables to a reasonable extent, as well as trying to limit the number of solvents added to the precursor (by sampling particularly the regions of the parameter space where 0 ml of either the NMP, DMF or DMSO solvent is used; cf. Fig. 1 for a visual representation of the coverage of every variable).

Afbeelding met schermopname, lijn, Rechthoek, Kleurrijkheid

Automatisch gegenereerde beschrijving

Afbeelding met schermopname, lijn, Rechthoek, plein

Automatisch gegenereerde beschrijving

**Figure 1:** Visual representation of the search space coverage by the initial sample.

During the optimization campaign, we aimed to increase the accuracy and robustness of the registered efficiencies for sampled points – compared to the exploratory sample – by mitigating the influence of the noise inherently associated with individual experimental PCE determinations. As such, four devices were prepared for every combination of process parameters sampled, and only the mean efficiency was registered. To save time and not waste resources, if defects were visually detected on the first device prepared for a specific combination of process conditions, then the device was discarded immediately and no further attempts to synthesize devices were made for this point in the search space.

For the construction of the main surrogate model, Gaussian Process regression with a matern52 kernel was selected.3 Matern52 is a flexible kernel, enabling it to treat potential discontinuities in the data better. Additionally, the kernel was selected in its anisotropic form, so that each of the kernel parameters could be tuned independently for every input variable, facilitating the assignment of a stronger impact/relevance on the predicted output variable, i.e., the measured efficiency, by some variables compared to others. It should be noted that the surrogate model, evaluated at point *x*, yields both a mean prediction for that point, , as well as a standard deviation , which quantifies the uncertainty of the model about its prediction at this point.

To inform the selection of process condition combinations to acquire in subsequent BO batches, a compounded acquisition function was selected. The utility function of upper confidence bound (UCB) was selected as the basis,4

|  |  |
| --- | --- |
|  | (2) |

where is the parameter that adjusts the relative weight to prediction uncertainty over the prediction mean value . In general terms of balancing exploration and exploitation, higher value leads to more exploration, while lower value leads to more exploitation. In this work, we set equal to 1, in line with previous work.5

While it was our goal to focus primarily on the (accurate) efficiencies measured as part of the BO campaign, we also did not want to lose the information from the preliminary sample and the visual defects data altogether during acquisition. As such, two probabilistic constraints were defined, in line with previous work by Buonassisi and co-workers.6 The first of those, , corresponds to the probability that the thin film will be produced without visual defects. This film constraint is defined based on a latent constraint function ,7

|  |  |
| --- | --- |
|  | (3) |

where is a secondary surrogate model in its own right, which is constructed with the help of Gaussian Processes as well, based on the combinations of process conditions evaluated so far. In the training data of this Gaussian process, , is assigned the value +0.5 if the fabricated device did not have visually detectable defects, and -0.5 if the device did contain defects. As a result of this convention, Eq. 3 will evaluate to 0 for all training points which yielded devices with visual defects, and to 1 for all training points resulting in devices without defects. For all other points in the search space, a probability to have a device without defects is interpolated based on the constructed surrogate model.

The second probabilistic constraint, , has been defined with the aim of penalizing regions of the search space around points, evaluated during preliminary sampling, which yielded subpar PCEs. This constraint is defined based on a latent constraint function ,7

|  |  |
| --- | --- |
|  | (4) |

where is another secondary surrogate model, which has again been constructed with the help of Gaussian Processes. The training points for are derived from the process condition combinations and corresponding PCE values for all the points in the preliminary sample, but then with the PCE values rescaled as follows,

|  |  |
| --- | --- |
|  | (5) |

where is the mean of all the PCE values in the preliminary sample. Based on these definitions, it can indeed be inferred that Eq. 4 will evaluate to 0 for all points in the preliminary sample that yielded a PCE below average, whereas it will evaluate to 1 for points that yielded a PCE above average. For all other points in the search space, a probability for the device, produced under the selected conditions, to yield an above average PCE is interpolated based on the constructed surrogate model.

The two constraints are then combined with to yield the final acquisition function as follows,

|  |  |
| --- | --- |
|  | (6) |

Note that the individual constraints have been softened, i.e., rescaled, in Eq. 6 to make the acquisition process more conservative with respect to the film and preliminary screening data. The weighting factors introduced to this end were chosen to make the current data 5 times more important than the preliminary screening data, and the qualitative film quality information is considered half as important as the information of the actual efficiency measurements. By introducing these constraints in the acquisition function, one can expect to sample fewer points in the region where film quality is poor, or where low device efficiency was obtained in the preliminary experiments, throughout the BO campaign.

As a final note, it should be mentioned that since the purpose of the probabilistic constraints is mainly to guide the model away from regions in the search space with a low probability of success, radial basis function kernels were used to model surrogates and . Radial basis functions are less flexible than matern52 kernels, and consequently, their smoother/more rigid behavior intuitively feels more appropriate to crudely define unfavorable regions of the search space.

Surrogate models and acquisition functions were implemented with the help of Emukit8 and GPy.9 All the code associated with this project has been made available through a GitHub repository [ADD LINK HERE].

**First batch**

In Fig. 2, the selected …

**References**

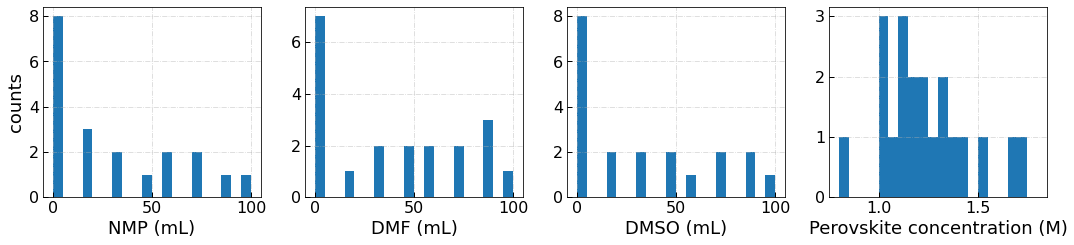
1. Frazier, P. I. (2018). A tutorial on Bayesian optimization. *arXiv preprint* **arXiv:1807.02811**.
2. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, **104**(1), 148-175.
3. Genton, M. G. (2001). Classes of kernels for machine learning: a statistics perspective. *Journal of machine learning research*, **2**, 299-312.
4. Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger (2012). Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Transactions on Information Theory*, **58**, 3250–3265.
5. Rohr, Brian, Helge S. Stein, Dan Guevarra, Yu Wang, Joel A. Haber, Muratahan Aykol, Santosh K. Suram, and John M. Gregoire (2020). Benchmarking the acceleration of materials discovery by sequential learning. *Chemical Science*, **11**(10), 2696-2706.
6. Liu, Z., Rolston, N., Flick, A. C., Colburn, T. W., Ren, Z., Dauskardt, R. H., & Buonassisi, T. (2022). Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing. *Joule*, **6**(4), 834-849.
7. Gelbart, M. A., Snoek, J., & Adams, R. P. (2014). Bayesian optimization with unknown constraints. *arXiv preprint* **arXiv:1403.5607**.
8. (a) Paleyes, A., Pullin, M., Mahsereci, M., McCollum, C., Lawrence, N., & González, J. (2019). Emulation of physical processes with Emukit. In *Second Workshop on Machine Learning and the Physical Sciences, NeurIPS*.
9. GPy: A Gaussian process framework in Python. Available at: \url{[http://github.com/SheffieldML/GPy}](http://github.com/SheffieldML/GPy%7D). (since 2012).

The experiments in this study were performed in two phases. First, an exploratory screening of the tentative search space of process conditions (cf. Table 1) was performed. This initial search space was selected based on expert considerations, intending to balance experimental feasibility and broadness of the search/parameter space. In total, 45 devices were synthesized within this space, and their approximate efficiency was measured crudely. Significant variations in the measured device efficiency were observed, and hence we decided to retain the selected set of process conditions for the eventual optimization campaign, i.e., phase two of this study. In total, 1.3 billion distinct combinations of process variables can be distinguished within this search space.

|  |  |
| --- | --- |
| Process variable | Total range (interval) |
| Volume NMP (ml) | 0 – 100 (5) |
| Volume DMF (ml) | 0 – 100 (5) |
| Volume DMSO (ml) | 0 – 100 (5) |
| Perovskite concentration (M) | 0.8 – 1.8 (0.1) |
| Annealing temperature (°C) | 100 – 160 (5) |
| Vacuum pressure (Pa) | 20 – 400 (10) |
| Vacuum pressure time (s) | 0 – 55 (2) |

During the second phase of the study, the experimental conditions were optimized with the help of Bayesian optimization. To increase the accuracy of the registered efficiencies in this part of the study, measurements were rigorously performed four times on independent samples for every single device, and the average was taken. To not waste resources, films with visually detectable defects were immediately discarded.

The initial sampling of the search space was performed manually, informed by the results from the preliminary screening, and with the aim of covering the ranges of the individual process variables to a reasonable extent, as well as trying to limit the number of solvents added to the precursor (by sampling particularly the regions of the parameter space where 0 ml of either NMP, DMF or DMSO is used; cf. Fig. 1 for a visual representation of the coverage of every variable).



Afbeelding met schermopname, lijn, Rechthoek, plein

Automatisch gegenereerde beschrijving

**Figure 1:** Visual representation of the search space coverage by the initial sample.

For the construction of the surrogate model, Gaussian Process regression with a matern52 kernel was selected. Matern52 is a flexible kernel, enabling it to treat potential discontinuities in the data better. Additionally, the kernel was selected in its anisotropic form, so that each of the kernel parameters could be tuned independently for each input variable, facilitating a stronger impact on the predicted output variable (the efficiency) by some variables compared to others.

To inform the selection of process condition combinations to acquire in the next batch, a compounded acquisition function was selected. In first instance, the utility function of upper confidence bound (UCB) was selected during the first couple of batches. Two probabilistic constraints were included in the acquisition function to incorporate the information about the film quality and the initial search space screening, in line with previous work by XXX. By introducing these constraints, one can expect to sample fewer points in the region where film quality is poor, or where low device efficiency was obtained in the preliminary experiments. To be conservative, we softened the probabilistic constraints for the film quality to a range of 0.5 – 1, and for the initial screening to 0.8 – 1. The weighting factors were chosen to make the current data 5 times more important than the preliminary screening data, and the qualitative film quality information is considered half as important as the information of the actual efficiency measurements. Since the purpose of the probabilistic constraints is mainly to guide the model away from regions in the search space with a low probability of success, radial basis function kernels were used to model these constraints (radial basis functions are less flexible than matern52 kernels, but their more smooth/rigid behavior intuitively feels more appropriate to crudely define inaccessible regions of the search space).