More trustworthy Bayesian optimization of materials properties by adding humans into the loop

Armi Tiihonen[*]
Aalto University
Espoo, Finland
armi.tiihonen@gmail.com

Louis Filstroff[†]
ENSAI, CREST
Rennes, France
louis.filstroff@ensai.fr

Petrus Mikkola
Aalto University
Espoo, Finland
petrus.mikkola@aalto.fi

Emma Lehto
Aalto University
Espoo, Finland
emma.lehto@aalto.fi

Samuel Kaski
Aalto University Espoo, Finland, and
The University of Manchester
Manchester, United Kingdom
samuel.kaski@aalto.fi

Milica Todorović
University of Turku
Turku, Finland
milica.todorovic@utu.fi

Patrick Rinke
Aalto University
Espoo, Finland
patrick.rinke@aalto.fi

Abstract

Bayesian optimization (BO) is a popular sequential machine learning optimization strategy for black-box functions. BO has proven to be an effective approach for guiding sample-efficient exploration of materials domains and is increasingly being used in automated materials optimization set-ups. However, when exploring novel materials, sample quality may vary unexpectedly, which, in the worst case, can invalidate the optimization procedure if undetected. This limits the use of highly-automated optimization loops, especially in high-dimensional materials spaces that require more samples. Sample quality may be hard to define unequivocally for a machine but human scientists are usually good at quality assurance, at least on a cursory yet often sufficient level. In this work, we demonstrate that humans can be added into the BO loop as experts to comment on the sample quality, which results in more trustworthy BO results. We implement human-in-the-loop BO via a data fusion approach and simulate BO of experimental perovskite film stability (data from the literature). Our human-in-the-loop approach facilitates automated materials design and characterization by reducing the occurrence of invalid optimization results.

[*]Corresponding author

[†]LF was with the Department of Computer Science, Aalto University, Finland at the time this research was conducted

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

1 Introduction

Bayesian optimization (BO) is a machine learning method designed to globally optimize a black-box function with sequential acquisitions of samples [1-2]. BO is intended specifically for guiding the sample acquisitions – that can be costly – towards the most promising areas of the search space [3]. More precisely, BO utilizes a probabilistic surrogate model (typically, a Gaussian process (GP)) that predicts the target property across a fixed search space of input variables. The surrogate model is improved via active learning by iteratively collecting new samples. The sampling strategy is determined by an acquisition function that has access to the surrogate model. Acquisition functions balance the exploration of new areas against the exploitation of promising, already sampled regions [3].

BO can be combined with automation to form semi- or fully autonomous materials optimization loops in self-driving materials laboratories (the concept illustrated in Figure 1a step (i), in which compositions suggested by BO are synthesized in a laboratory and then measured). In materials science, BO has successfully been used for guiding the experimental or computational exploration of novel materials spaces and for determining the optimal combinations of device settings [4-7].

One of the factors limiting the broader use of BO in experimental materials science is the varying quality of the experimentally prepared samples. The samples should be of sufficient quality to ensure valid characterization results for the target property, which is a necessity in any experimental work. In material composition explorations, for example, the sample preparation conditions (e.g. annealing temperature) might have to be adjusted for specific compositions to maintain sufficient quality. Such adjustments are possible only when the number of required samples is small or when domain knowledge on the nature and location of potential sample quality issues is available.

The increasing level of laboratory automation facilitates a paradigm shift towards higher-dimensional (i.e., more samples needed) and novel search spaces (i.e., more unexpected events during the search). In that scenario, human scientists can no longer manually check all the samples in-situ, else, they would severely slow down the otherwise accelerated optimization loop. Alternatively, automated sample quality tracking via a series of added characterization methods could in principle be implemented, but this is in many applications prohibitively costly and difficult to maintain. Moreover, tracking sample quality fully automatically by a machine would require defining sample quality precisely and unequivocally before having prepared the samples. This is challenging especially in novel search spaces in which unanticipated events frequently occur. An experienced human scientist remains more versatile than a machine at suspecting sample quality issues after a quick inspection and – when needed – confirming them with a suitable characterization method.

We thus propose to add humans into the BO loop to evaluate sample quality, and query them only when the human opinion is needed to prevent the humans from becoming a bottleneck in the loop. Human-in-the-loop approaches do exist in machine learning [8-12] but they are less frequent in materials science [13-14]. In this contribution, we show via a simulated example that is based on real experimental data [4] that adding humans into the BO loop reduces the likelihood of ending the search in erroneous optima. Adding humans into the loop thus increases trust in BO optimization without significantly increasing the overall effort. By showcasing our human-in-the-loop approach, we hope to trigger discussion on the roles of human scientists in the AI-guided design of new materials, and on how sample quality should be addressed in partially for fully self-driving materials laboratories.

2 Results and discussion

2.1 Framework

We investigate the proposed human-in-the-loop BO methodology in an example setting of compositional stability optimization of perovskite film samples (Figure 1a, see the detailed methods description in Supplementary Methods Section A.1). New experimental data has not been generated for this contribution. Instead, we simulate BO loops utilizing experimental BO results from [4] and sample from the environmental stability landscape produced in [4] (Figure 1b) to retrieve emulated experiment data.

The composition of a ternary perovskite $Cs_xMA_yFA_zPbI_3$, where

$$(x, y, z) \in [0, 1]^3, x + y + z = 1, \tag{1}$$

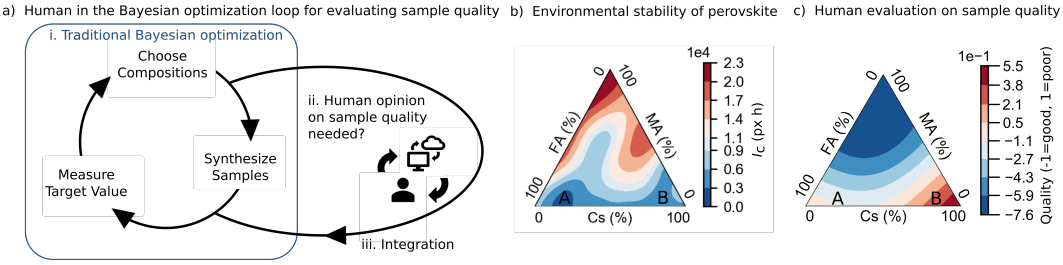

Figure 1: a) Experimental Bayesian optimization (BO) of materials (i) with humans added into the loop for evaluating the quality of synthesized samples (ii-iii). Adding humans requires a query policy (ii) and the integration (iii) of the human response into the loop. The simulated BO loops in this work sample from two Gaussian process models (assumed here in the simulations as the ground truth, a lower value is better for both models): b) the environmental stability of perovskites (BO target) via a proxy variable I_c (the color change of the sample in RGB pixels integrated over the aging test duration in hours), and c) human evaluation on the quality of the perovskite samples fitted based on the pictures of the fresh samples (evaluations on a scale of [-1,1]). The data for fitting both models is from literature [4]. Region A is the pursued optimum that virtual BO runs should find in this work, and Region B is the region to be avoided.

was, in the previous work [4], optimized to survive environmental stress from heat, humidity, and visible light illumination. The BO optimization variables were x, y, and z, the compositional proportions of A site compounds in the perovskite ABX_3 crystal structure. Perovskite thin film samples were prepared with compositions suggested by BO and exposed to aging tests (detailed in Supplementary Section A.1.1). Since the investigated perovskite compositions undergo color changes from dark to yellow when the films degrade [14], the BO minimization target was taken to be the integrated color change of the perovskite film I_c (cf Supplementary Section A.1.1). The loop of preparing samples with compositions suggested by BO (a local penalization batch-mode BO implementation with expected improvement acquisition function and GP surrogate model with Matern kernel) was iterated until the search converged.

Here we adopt the same BO input space x, y, z and target variable I_c to be minimized, sampling from the environmental stability GP model (Figure 1b) that we reproduced from the converged BO search in [4]. It should be mentioned that in [4], the expected improvement acquisition function was modified with a data fusion method to account for a secondary design requirement, thermodynamic phase stability of the compositions. In the current work, we do not use the thermodynamic phase stability data, but we utilize the same data fusion method for adding the humans into the loop. The data fusion method (inspired by the unknown constraint method [16]) uses the estimated probability of filling the secondary design requirement (here, sufficient sample quality evaluated by humans) as a down-weighting cost factor in the BO acquisition function.

An experienced scientist can provide an initial evaluation of perovskite film quality based on its color and uniformity. All the samples produced in the earlier work [4] were photographed. Thus, in this work, we graded the sample quality on a scale of 0 (good quality) - 3 (low quality) post-hoc based on the saved photographs (example pictures shown in Supplementary Figure S2). We subsequently trained a GP model on this data (Figure 1c, see Supplementary Section A.1.2 for details) that serves as the human opinion in the present work. Interestingly, the resulting human opinion model (formed with no knowledge on the individual sample compositions) identifies the same part of the search space as low quality than the thermodynamic phase stability approximation in [4] (see Supplementary Figure S3). This suggests the human evaluation on perovskite film samples indeed links to physically meaningful properties of the samples.

The comparison of the two GP models (Figure 1b-c) reveals a global optimum region (Region A) in the search space that has good quality samples and is environmentally stable. There is also a false optimum region (Region B), which results in seemingly almost equally good environmental stability but many of the samples are actually low quality, leading to invalid stability measurements in this region (the reasons discussed in Supplementary Section A.3). Hence, in this work, we investigate if human input can steer the BO search away from Region B and facilitate convergence towards the optimum in Region A.

2.2 Human-in-the-loop implementation

Human integration requires two steps: deciding when human opinion is required and integrating it into the BO loop (Figure 1a steps ii and iii, the details of our implementation in Supplementary Section A.1.3). For both steps, we have many choices. This work is not a comprehensive evaluation of all the possible choices but instead, it serves as a proof-of-concept human-in-the-loop configuration that provides promising results for materials optimization.

Here, the stability of all the samples is queried and human evaluation is requested on some of these samples at the time when the next samples are suggested. Human evaluation is requested only when there are no previous human evaluations on the compositions nearby (the radius of the compositional exclusion circle, r_{excl}, serving effectively as a hyperparameter) and there are rapid changes in the environmental stability as a function of composition (the limit value of the gradient, g_{excl}, serving as another hyperparameter). The hyperparameter values are chosen based on experimental domain knowledge (see Supplementary Section A.1.3):

$$r_{excl} = 0.1, \tag{2}$$

which is a tenth of the whole compositional range in this work, and

$$g_{excl} = 0.1\sigma/l, \tag{3}$$

where σ and l are the standard deviation and lengthscale of the environmental stability surrogate model, respectively.

The human evaluations are, in this work, integrated into the BO loop via the previously mentioned data fusion method [4]: the point-wise human opinions are approximated with a GP surrogate model that spans the whole search space. The predictive mean from the human GP surrogate model, μ_{human}, are scaled to a probability distribution (the scaling factor β serving effectively as a hyperparameter):

$$P_{\text{human}}(x, y, z) = \frac{1}{1 + e^{\mu_{\text{human}}(x,y,z)/\beta}} \in [0, 1]. \tag{4}$$

Probabilities are integrated into the BO loop as an added cost to the acquisition function α :

$$\alpha_{\text{human}}(x, y, z) = \alpha(x, y, z) * P_{\text{human}}(x, y, z), \tag{5}$$

which suppresses sampling from regions that are likely to have low-quality samples.

2.3 Results

We evaluate the success of the human integration based on three criteria: the optimum environmental stability value found by BO, the compositional distance of this optimum to Region A (evaluated as the simple regret), and the number of queries made on the human (here, number of samples from the human evaluation model). The BO cycle is initiated with two uniformly at random selected samples and iterated 100 rounds. To gather statistics, the BO cycle is repeated 25 times. The human-in-the-loop BO (that queries human only when necessary, Figure 1a steps (i)-(iii)) is compared to vanilla BO (without human integration, Figure 1a step i) and a version that queries human for every sample (Figure 1a steps (i) and (iii)).

Our results in Figure 2a show that vanilla BO frequently converges to other regions than Region A – namely Region B – because the regret remains high. Human integration leads to faster and more reliable convergence of the search to Region A: the human-aided version (orange line) converges in 15 rounds of sampling to a lower regret than vanilla BO in 100 iterations. This is while having shown, on average, 5% of the samples to humans (Figure 2b). A traditional way of evaluating the convergence of BO is to track the optimum value found (here stability). All the compared methods perform similarly by this measure (Supplementary Figure S4) since Regions A and B have almost equal stability values. It is the human evaluation only that reveals that Region B has low quality samples and that the stability measurement values in this region can hence not be trusted.

The ideal rate of querying the human depends on the circumstances for including the human in the first place. For prioritizing increasing trust in the BO results, more samples should be shown to humans, while other factors like the desired acceleration of the optimization cycles sets a limit to how much the human role can be increased. In the low limit of the number of queries, no samples are shown to humans and the performance corresponds to the vanilla BO in Figure 2a. In the high

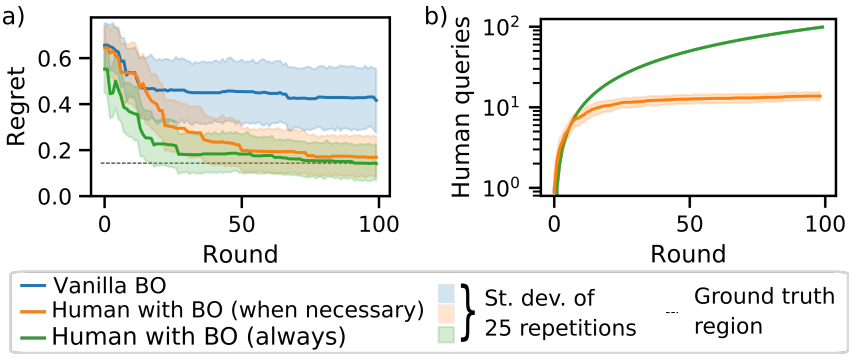

Figure 2: Results of the simulated Bayesian optimization (BO) loops with human or without (vanilla BO). The BO loop is repeated 25 times for statistics, 100 rounds of sampling is performed in each loop, and one sample is sampled in each round. a) Simple regret i.e. the compositional distance of the sample with the best environmental stability found by the BO from the ground truth optimum, Region A. b) The accumulated number of queries made on the humans.

limit, the humans inspect all samples, but the benefits are marginal (Figure 2a), while the human effort is high (Figure 2b). Further work is required to determine an optimal querying frequency that is applicable to new materials science problems. In this context, it should be noted that our implementation does not stop the BO search if the requested human evaluations are not delivered in time (and they can still be provided in the later rounds), which is a choice made to avoid humans becoming a bottleneck.

Adding humans into the BO loop affects the acquisition locations, which indirectly introduces bias from human opinions into the resulting surrogate model of the target property, here environmental stability. Therefore it is particularly important to analyze human-in-the-loop BO results in conjunction with the mean and variance predictions from the surrogate models of both the target property and the human evaluations. The predictions from the surrogate model are less reliable in regions that have been sampled sparsely (i.e., have high entropy). Thus regions avoided due to sample quality issues need to be considered for future searches with refined sample preparation methods. The human-in-the-loop implementation in this work does not remove the low-quality samples from the dataset as they contribute to the variance of the stability surrogate model. This choice also facilitates the BO result analysis – regions with low-quality samples but good proxy variable values can be quickly identified for refined investigation by comparing the two surrogate models. Future work will re-evaluate, if this choice should be refined to optimize the algorithmic performance. Our future work will also compare different human-in-the-loop techniques and implementations.

Our demonstration is on perovskite thin film aging data, for which the visual inspection is a suitable way to estimate sample quality and can be done for few pre-defined samples while transferring the samples to the aging test device. Sample quality issues are present also in other types of materials samples and similar approaches for pre-screening sample quality by an experienced human scientist can be found for many of them. Thus, the presented approach is applicable beyond perovskite optimization tasks.

3 Conclusions

Our proof-of-concept human-in-the-loop Bayesian optimization (BO) process demonstrates that adding humans into experimental materials optimization loops for evaluating sample quality can facilitate the convergence of BO into meaningful optima with samples of sufficient quality (in contrast to wasting sampling in regions with low-quality samples and, consequently, invalid data). Here, we have shown an example case of perovskite composition optimization for environmental stability. Our work illustrates how human-in-the-loop methods could aid AI-guided materials design. Our work aims to facilitate high-dimensional and novel materials searches in partially self-driving laboratories — which are currently hindered by concerns on varying sample quality — and to trigger further discussions on how sample quality should be quantified and ensured in future materials searches.

Acknowledgments and Disclosure of Funding

This work was supported by the Academy of Finland (Flagship program: Finnish Center for Artificial Intelligence FCAI) and the Academy of Finland through projects 334532 and 348180. Authors thank the reviewers for valuable comments during the review process. Authors declare no competing interests.

References

[1] Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. Journal of Global optimization, 13(4), 455-492.

[2] Brochu, E., Cora, V. M., & De Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint arXiv:1012.2599.

[3] Greenhill, S., Rana, S., Gupta, S., Vellanki, P., & Venkatesh, S. (2020). Bayesian optimization for adaptive experimental design: a review. IEEE access, 8, 13937-13948.

[4] Sun, S., Tiihonen, A., Oviedo, F., Liu, Z., Thapa, J., Zhao, Y., ... & Buonassisi, T. (2021). A data fusion approach to optimize compositional stability of halide perovskites. Matter, 4(4), 1305-1322.

[5] Löfgren, J., Tarasov, D., Koitto, T., Rinke, P., Balakshin, M., & Todorović, M. (2022). Machine Learning Optimization of Lignin Properties in Green Biorefineries. ACS Sustainable Chemistry & Engineering, 10(29), 9469-9479.

[6] Wakabayashi, Y. K., Otsuka, T., Krockenberger, Y., Sawada, H., Taniyasu, Y., & Yamamoto, H. (2019). Machine-learning-assisted thin-film growth: Bayesian optimization in molecular beam epitaxy of SrRuO3 thin films. APL Materials, 7(10), 101114.

[7] Langner, S., Häse, F., Perea, J. D., Stubhan, T., Hauch, J., Roch, L. M., ... & Brabec, C. J. (2020). Beyond ternary OPV: high-throughput experimentation and self-driving laboratories optimize multicomponent systems. Advanced Materials, 32(14), 1907801.

[8] Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Pla, O. A., Thomas, O., ... & Klami, A. (2021). Prior knowledge elicitation: The past, present, and future. arXiv preprint arXiv:2112.01380.

[9] Daee, P., Peltola, T., Soare, M., & Kaski, S. (2017). Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. Machine Learning, 106(9), 1599-1620.

[10] Sundin, I., Peltola, T., Micallef, L., Afrabandpey, H., Soare, M., Mamun Majumder, M., ... & Kaski, S. (2018). Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge. Bioinformatics, 34(13), i395-i403.

[11] Bharti, A., Filstroff, L., & Kaski, S. (2022). Approximate Bayesian Computation with Domain Expert in the Loop. arXiv preprint arXiv:2201.12090.

[12] Hvarfner, C., Stoll, D., Souza, A., Lindauer, M., Hutter, F., & Nardi, L. (2022). π BO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization. arXiv preprint arXiv:2204.11051.

[13] Mikkola, P., Todorović, M., Järvi, J., Rinke, P., & Kaski, S. (2020, November). Projective preferential Bayesian optimization. In International Conference on Machine Learning (pp. 6884-6892). PMLR.

[14] 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.

[15] Keesey, R., Tiihonen, A., Siemenn, A. E., Colburn, T. W., Sun, S., Hartono, N. T. P., ... & Buonassisi, T. (2022). An Open-Source Environmental Chamber for Materials-Stability Testing Using an Optical Proxy.

[16] Gelbart, M. A., Snoek, J., & Adams, R. P. (2014). Bayesian optimization with unknown constraints. arXiv preprint arXiv:1403.5607.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Human opinions are in the described setting useful and hence integrated into the optimization. However, this introduces human bias into the optimization cycle which – if left hidden and not considered properly in the optimization result analysis as discussed in this work – may also be harmful for the ultimate target of finding the optimally performing materials.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Described in the appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Cloud provider is not needed in this work.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Data published in a literature article previously.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Methods

A.1.1 Perovskite film data

No perovskite films were prepared as a part of this work but experimental data from literature [4] was used as the basis of simulations of this work. Perovskite film preparation and measurements are detailed in article [4].

As an overview of the experiments in literature [4], 112 $Cs_xMA_yFA_{1-x-y}PbI_3$ perovskite film samples with varying compositions were prepared as a part of an experimental BO loop to find environmentally the most stable perovskite compositions. The films were spin-coated inside a glovebox on top of glass substrates. Solutions over-stoichiometric with PbI_2 and chlorobenzene antisolvent were used in the synthesis. Following the spin-coating, the samples were annealed at 403K for 20 min in a glovebox. Perovskite film samples were prepared in batches of 28 samples with varying compositions, and 4 rounds of aging tests with 28 perovskite films at a time were required for the BO loop to converge. The perovskites were in each aging test degraded for 7,000 minutes under

$85 \pm 2^{\circ}$C temperature, 85±5% air humidity, and 0.15±0.01Sun visible-only illumination, and photographed automatically with 5-minute intervals. A proxy variable for the degradation of the films was defined as [4]:

$$I_c(x, y, z) = \sum_{c=\{R,G,B\}} \int_0^T |c(t, x, y, z) - c(0, x, y, z)|dt, \tag{6}$$

where c is the pixel-averaged color of the films, respectively, t is time, T is aging test duration, and R, G, and B are the red, green, and blue RGB channels in the sample photos, respectively. The smaller I_c value, the more stable the film is.

A.1.2 Implementation of the ground truth models

In the simulated BO loops, the experimental environmental stability measurements and human evaluations of the samples were replaced by querying two models assumed to provide ground truth on environmental stability and human evaluations. The simulated loops were chosen in order to be able to quantify the performance differences between the Bayesian optimization implementations that were compared.

The two ground truth models are Gaussian process (GP) regression models with a radial basis function kernel presented in Figure 1b-c – one for the estimated environmental stability of the perovskite films and another for the human evaluations of the film quality. The fitted ground truth models are shared in the adjoining open code repository HPER (see Section A.1.5). It should be noted that since these two models are fitted based on data resulting from a Bayesian optimization search (documented in literature [4]), they are inherently biased by the unevenly sampled data, and also by measurement variations. They should thus be regarded as the estimates of the ground truth (not the exact ground truth). However, for the purposes of the simulations performed in this work, these models provide a representative case of actual measurement results one could get on perovskite stability and sample quality in a laboratory.

For fitting the environmental stability model, the full dataset of environmental stability measurements of 112 perovskite samples from literature [4] was used. For fitting the human evaluations model, 56 of these samples were evaluated by a human. Only the first half of the samples queried in the BO search documented in [4] were evaluated because in the later stages of the search, the suggestions had already converged into a compositionally limited region. The evaluations were made by the first author of this contribution but sample compositions were hidden from the evaluator at the time of the evaluations.

Mean predictions with Gaussian noise from the environmental stability GP model (Figure 1b) were used instead of environmental stability experiments and noiseless mean predictions from the human evaluation GP model (Figure 1c) were used instead of human opinions during the virtual BO loops in this work.

A.1.3 Implementation of human in the loop for Bayesian optimization

Human-in-the-loop implementation developed for this work is shared in an open code repository HPER (see Section A.1.5). HPER was build on top of a Bayesian optimization Python implementation shared as a part of earlier work [4] in an open code repository SPProC (see Section A.1.5). To clarify the contributions in this work compared to SPProC: SPProC contains a modification to BO Python package GPyOpt for implementing the data fusion approach [4] for integrating additional data sources into the acquisition function of the BO algorithm. In HPER, SPProC is modified by adding the human queries into the main BO loop and fed into the BO algorithm using the data fusion method.

Human is in this implementation queried only when the chances of getting low-quality samples are estimated to be high. The estimation is done at the end of each BO round when the next samples to be suggested are compiled. Only the samples that are suggested as a part of standard BO are subjected to a further evaluation on if the human opinions should be requested for these samples. This choice was made because it may be dismotivating for humans to evaluate samples that had been prepared with a lot of effort but not exposed to the measurement of the actual target value of interest.

The chances of getting low-quality samples are assumed to be high when they are from new regions not having been sampled before, and when rapid changes in the values of the target variable are expected at the composition in question. If these two criteria are met, a human evaluation on the sample is requested at the same time when new samples are suggested. At the beginning of the next BO round, the BO algorithm receives both the target variable data and, separately, the human evaluations (assuming that humans have completed the requested evaluations – the BO will run normally even if the humans would be delayed in providing their opinions). In the integration phase, the accumulated point-wise human evaluations are fed into BO with data fusion [4]: a GPR model is fitted on the human evaluated sample quality data points and the mean predictions of the model, $\mu_{\text{human}}(x, y, z)$, are scaled sigmoidally to a probability distribution of getting good samples:

$$P_{\text{human}}(x, y, z) = \frac{1}{1 + e^{\mu_{\text{human}}(x,y,z)/\beta}}, \tag{7}$$

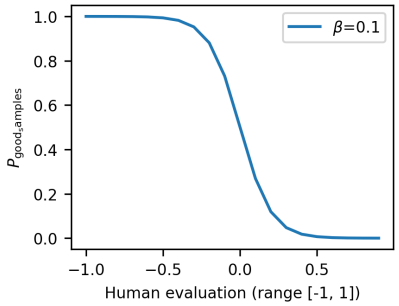

Figure S1: An example of the scaling of probability with scale factor β value of 0.1.

where β is a fixed scale factor. Finally, $P_{\text{human}}(x, y, z)$ is used as an added cost for the BO acquisition function $A(x, y, z)$ to reduce sampling from the regions that are likely to produce low-quality samples:

$$A_{\text{human}}(x, y, z) = A(\theta) * P_{\text{human}}(x, y, z). \tag{8}$$

The benefit of the data fusion approach is that the choice of the scale factor β can be used for describing the certainty of human evaluations on sample quality. When conservative values of β are chosen, no region is fully removed from the BO search but only the tendency to sample from risky regions is reduced. Thus, the likelihood of BO wasting bandwidth to loq-quality samples is reduced but if the risky region seems to contain extremely good samples, BO still samples some samples from there.

The presented approach for integrating human into the loop for BO requires setting three parameters that act effectively as hyperparameters: two parameters affecting when human is queried (the radius of the compositional exclusion zone inside of which human is not queried, r_{excl}, and limit gradient of the environmental stability vs. composition below which human is not queried, g_{excl}) and one affecting the scaling of human observations into a probability distribution (β). In this work, g_{excl} was fixed to lengthscale and standard deviation of the environmental stability surrogate model (GP regression), l and σ, respectively:

$$g_{excl} = 0.1\sqrt{\sigma}/l, \tag{9}$$

thus allowing g_{excl} to adapt to the surrogate model that develops round by round. The maximum possible number of human queries during the BO is limited by the size of the axclusion zone. Here, r_{excl} was fixed to:

$$r_{excl} = 0.1. \tag{10}$$

In practice, the exclusion zone should not be smaller than the experiment resolution (it is estimated in [4] that the resolution of being able to prepare samples with different compositions is 0.01). Additionallt, since the composition range is up to 1, a reasonable r_{excl} value in this application must fulfill $0.01 < r_{excl} > 0.5$. The sigmoid scale factor β determines how conservative the model is towards human observations when translating them to a probability of getting good-quality samples. In this work, we have chosen a conservative estimate $\beta = 0.1$ illustrated in Figure S1. Additionally, because the human evaluations ranged from 0 to 3, they were shifted by 1.5 before the sigmoid transformation for symmetry. Generally, the parameters r_{excl}, g_{excl}, and β may need to be adjusted based on the application. The decisions should be made considering the estimated burden for humans, the speed of the rest of the optimization cycle with respect to the speed of human queries, and the positive and in some cases possibly also negative effects of human bias brought into the optimization cycle. Ways to automatically set the parameters will be investigated in future in order to simplify the use of human integration code.

A.1.4 Evaluating the performance of optimization algorithms

The human-in-the-loop BO described in the previous subsection was compared to two other approaches: vanilla BO that is the traditional BO without human integration, and a human-in-the-loop implementation with a simplistic query policy of asking human opinions on every sample. All the three approaches used as much of identical settings as possible: the acquisition function was expected improvement (this served as the base acquisition function for the human-in-the-loop versions) with jitter of 0.01, the surrogate models were Gaussian process models with Matern kernel for the stability model and radial basis function kernel for the human model, and 100 rounds of sampling one sample on each round were performed.

These BO cycles were repeated 25 times to estimate the performance of the algorithm versions in terms of the current optimum $I_c^*(t)$ found by round t, the number of human queries made, and simple regret:

$$R(x_t^*, y_t^*, z_t^*) = \sqrt{(x_{opt} - x_t^*)^2 + (y_{opt} - y_t^*)^2 + (z_{opt} - z_t^*)^2}, \tag{11}$$

where (x_t^*, y_t^*, z_t^*) is the composition of $I_c^*(t)$ and $(x_{opt}, y_{opt}, z_{opt})$ is the composition of the true optimum.

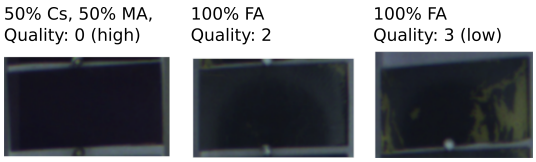

Figure S2: Examples on human evaluation of perovskite film samples.

A.1.5 Data and code availability

The perovskite data used for simulations was produced in work [4] and has been shared by the authors of [4] in a Github repository:

- https://github.com/PV-Lab/SPProC

The raw data produced in [4] has been shared by the corresponding authors of [4] in the following repositories:

- https://doi.org/10.6084/m9.figshare/20506857
- https://doi.org/10.6084/m9.figshare.20523327
- https://doi.org/10.6084/m9.figshare.20521173

The human-in-the-loop code implementation prepared in this work, together with the GP models used as the source of ground truth data in the simulations in this work, are shared in a Github repository:

- https://github.com/TadaSanar/HPER

A.2 Human evaluations as the basis for forming the ground truth model for human-evaluated sample quality in the example perovskite dataset

Human evaluations of perovskite film samples were performed post-hoc from photographs that had been taken from fresh samples before aging them in article [4]. Example evaluations are shown in Figure S1. Human evaluation is mostly linked to the yellowness of the sample (yellow color of lead iodide perovskites indicates either non-photoective yellow phases being present in the film or the perovskite having decomposed even before the aging test, which means yellow films are low quality). Two of the samples in Figure S1 have the same composition ($FAPbI_3$) but they still look different and thus, have received a different evaluation on sample quality. This is an example of sample-to-sample variance that is present in the dataset.

A.3 Competing optima in the example perovskite search space

Figure 1b illustrates that in the example case of optimizing $Cs_xMA_yFA_{1-x-y}PbI_3$ perovskite composition for environmental stability, there are two competing regions with seemingly high environmental stability: Region A and Region B. The model in Figure 1b has been fitted on the measurement results in the source article [4]. These measurements are not direct measurements of the environmental stability of the perovskite samples but measurements of a proxy variable that is easier to probe: the color change of the sample integrated over time during the aging test (more information in Supplementary Section A.1). Both Regions A and B therefore have samples that do not change their color much during an aging test. However, in Region B, the samples tend to suffer from phase de-mixing into minority phases already before the aging test starts [4], which is thus not captured by the integrated color change metric. In extreme cases, the high-Cs films in Region B suffering from de-mixing may even look yellow even as fresh due to the presence of non-photoactive, yellow $CsPbI_3$ phase. The phase de-mixing deteriorates the photovoltaic properties of the films in Region B, resulting in low-quality perovskite samples, and thus, Region B is actually an invalid optimum.

The phase de-mixing issue can be estimated via thermodynamical phase stability. If Gibbs free energy of mixing, dG_{mix}, is positive for a certain composition, it should de-mix into minority phases. The estimated dG_{mix} over the $Cs_xMA_yFA_{1-x-y}PbI_3$ perovskite search space in Figure S3 indeed suggests Region B suffers from phase de-mixing and the rest of the search space is phase-stable.

Interestingly, the comparison of Figure S3 and Figure 1c reveals that human evaluations on sample quality also highlight Region B as a region with low-quality samples. This is likely because the presence of the highest proportions of non-photoactive yellow minority phases is visible even by the eye. Thus, in this search space, human evaluations actually serve as an incomplete, low-fidelity option to more accurate but computationally costly density functional theory calculations.

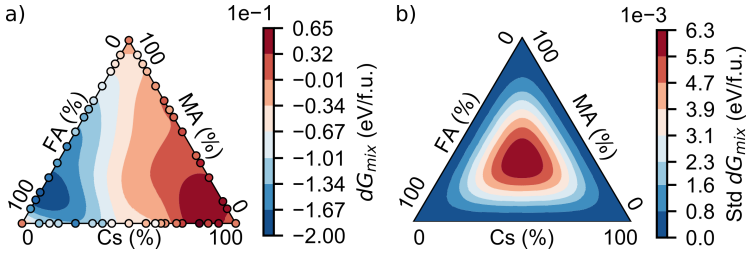

Figure S3: a) Mean and b) standard deviation estimates of Gibbs free energy of mixing modelled using a Gaussian process (GP) regression model and density functional theory data from article [4]. According to the estimate, all the compositions with positive energies should be thermodynamically instable. The data used contains only binary perovskite compositions from the triangle edges (shown in the figure) and the model is thus less reliable with ternary compositions in the interior of the triangle. The GP model was fitted using the same settings than are used in the source article [4].

A.4 Convergence of the optimum stability found during the Bayesian optimization loops

In the test application case presented in this work, Regions A and B produce almost equally low I_c values. Thus, it cannot be detected based on the found optimum value only if the search has converged into Region A or B. The sampling from environmental stability model (Figure 1b) is in the virtual BO implemented with Gaussian noise to mimic real experiments.

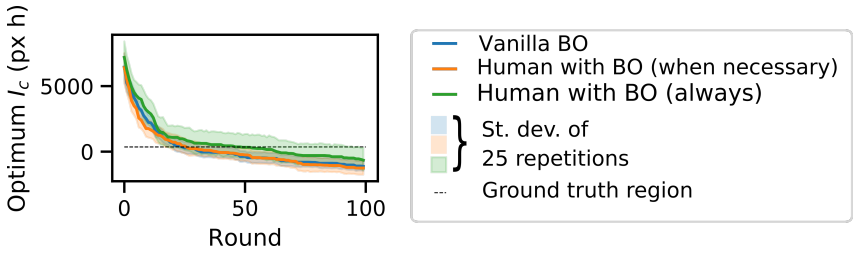

Figure S4: The convergence rate of the Bayesian optimization with (orange and green) and without human in the loop (blue) is similar.