Bayesian Optimization Framework for Data-driven Materials Design

Giovanna Tosato[1], Arnd Koeppe[1,2], Bai-Xiang Xu[4], Michael Selzer[1,2,3], Britta Nestler[1,2,3]

[1]Institute for Applied Materials – Microstructure Modelling and Simulation (IAM-MMS), Karlsruhe Institute of Technology (KIT);
[2]Institute of Nanotechnology – MicroStructure Simulation (INT-MSS), Karlsruhe Institute of Technology (KIT);
[3]Institute for Digital Materials Science (IDM), Karlsruhe University of Applied Sciences;
[4]Institute of Materials Science, Mechanics of Functional Materials, Technische Universität Darmstadt

Research data management systems enable efficient data-driven design and optimization across research disciplines. In materials science, advancing experimental design and optimizing materials with complex properties pushes the boundaries of research. For aqueous foams, the microstructure significantly influences the macroscopic properties, which can be characterized through experimental techniques or simulated using physics-based models. To address the complexity of tuning interrelated parameters in a vast design space, we propose an active learning-based data-driven framework.

This framework leverages Bayesian Optimization to guide the exploration of the search space, prioritizing informative experiments or simulations and minimizing the number of required evaluations. Implemented within the Kadi ecosystem, our workflow promotes data reuse and facilitates seamless collaboration.

1 Introduction

In materials design, optimizing material properties for specific applications is a challenging task, as these properties often exhibit complex behaviors. Effects at multiple scales influence macroscopic material behavior and function, e.g., microscopic pores affect stiffness and thermal insulation. The characterizing parameters of microstructures are often numerous and interconnected through relationships that are only partially understood. This high dimensionality and complex dependencies yield a vast design space, which poses the researchers the challenge of tuning several parameters simultaneously to gain the desired performance.

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: https://doi.org/10.11588/heibooks.1288.c18093

This multifaceted challenge is prominently evident in the study use case of liquid foam evolution, where additionally, the non-linear, time-series nature of the data has to be taken into account. To that end, machine learning can handle unknown, nonlinear, complex, and high-dimensional relationships, but often requires large amounts of data to recapture information.

The acquisition of this data frequently involves performing high numbers of resource-intensive experiments or simulations. Therefore, it is desirable to generate information-rich datasets while minimizing the number of required evaluations. Our strategy to generate small but comprehensive datasets is to exploit Bayesian Optimization and its memory of previously acquired data to guide the selection of the next most informative data point to be evaluated.

Advancing research within the context of e-science and open data, the need for research data management support becomes self-evident. Therefore, we operate within the Kadi ecosystem, our open-source research data infrastructure for materials science[1], which empowers our research following the FAIR principles. The Kadi4Mat repository (Brandt et al. 2021), the workflow engine KadiStudio (Griem et al. 2022), and the interface for AI, KadiAI (Koeppe and CIDS Team 2023) enable structured data storage and reproducible workflows for data analysis and visualization tasks, facilitating sharing and reproducibility.We enclose the Bayesian oracle into a Python framework and make it accessible from the Kadi ecosystem through a node-based graphical interface. This enables the development of "intelligent" study workflows that actively learn to investigate the parameter space.

2 Methods

In materials design, data-driven approaches often outperform traditional trial-and-error and even systematic a-priori planned studies (Himanen et al. 2019). However, the efficiency of data-driven approaches hinges on the ability to guide data collection effectively. Our solution to this problem consists of a design-of-experiments methodology leveraging Bayesian Optimization to efficiently explore and exploit the search space. We iteratively update our model and enrich our dataset by selecting the next most informative data point, guided by previous evaluations. This approach falls in the category of *active learning* (Settles 2009), a subfield of supervised machine learning where a component of the algorithm (an oracle) is responsible for choosing the training data. Motivated by the knowledge that not all samples are equally meaningful, we aim to obtain an expressive dataset that meets the requisites of quality, quantity, and variability while reducing resource usage.

In this work, we implement a query strategy based on a Bayesian oracle. Bayesian Optimization (BO) is known in statistics as a method for global optimization of black-box functions, usually employed in the case of expensive-to-evaluate functions.

1 `https://kadi.iam.kit.edu`

In resource-intensive materials simulations and experiments, microstructure-property relations represent an attractive use-case for BO. This aspect significantly contributes to the growing attention given to BO as an efficient tool for multidimensional optimization in materials design (Kotthoff, Wahab, and Johnson 2021; Kuhn et al. 2021; Zhao et al. 2023; Kulagin et al. 2023).

The main strength of the method lies in the ability to define, work with, and continuously update a cheap *surrogate function* which estimates the behavior of the real (expensive) objective function. The surrogate is a probabilistic model over functions mirroring our objective. A common choice is to use *Gaussian Processes*, which approximate the behavior of the datasets and provide uncertainty estimates. To cover the search space balancing exploration and exploitation, the most promising candidate is selected maximizing a customizable *acquisition function*, see Figure 1.

A customizable active learning framework is under development in the Computational Intelligence and Data Science framework CIDS (Koeppe et al. 2022) within the Kadi ecosystem.

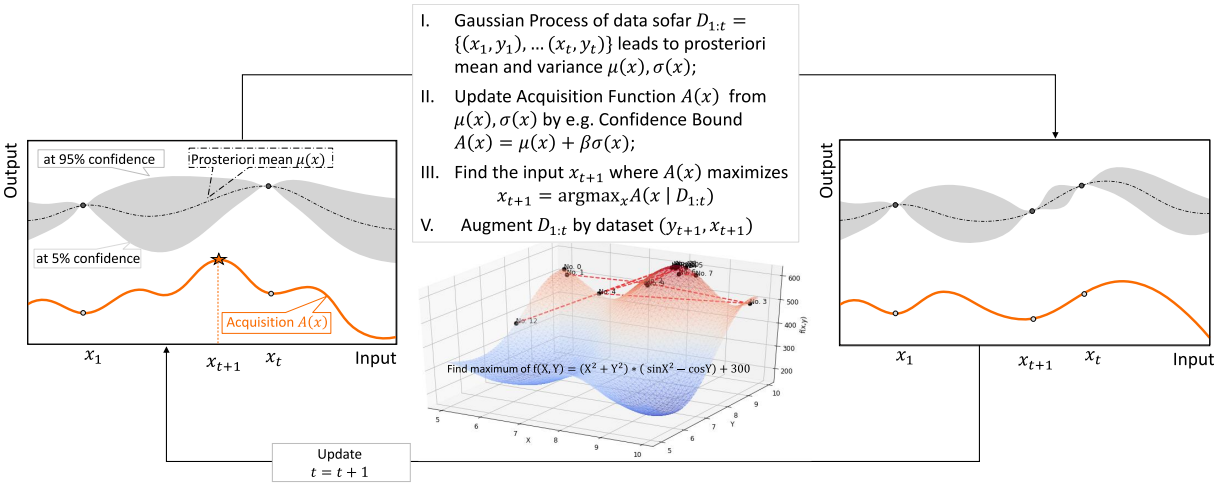

Figure 1: Example iteration of Bayesian Optimization.

3 Workflow

For the use-case of liquid foam evolution, Figure 2 shows how the whole data life cycle exploits and is embedded in Kadi4Mat, which manages every step of the research process, including generation, storage, usage, and sharing. For the generation of simulation data, a KadiStudio workflow automatizes the creation of initial parameters' settings and performs the designed simulation(s) employing the in-house framework Pace3D, Parallel Algorithms for Crystal Evolution in 3D (Hötzer et al. 2018).

A representative temporal evolution analysis of 2D liquid foams was performed on the BwUniCluster2.0 platform, utilizing 40 cores over a computational time of 35 hours. A single timestamp for such simulations occupies around 500MB of storage. To adequately

capture temporal evolution patterns, the study necessitates the examination of 50 to 200 frames for each configuration.

In contrast, the experimental data are generated with a Dynamic Foam Analyzer (DFA100, Krüss GmbH, Hamburg, Germany) from research partners (Jung et al. 2023). Significant observations can span a duration of up to 10 hours, while the equipment generates a complete status dump every 0.2 seconds, yielding a substantial volume of unprocessed data.

For storage, the results are uploaded to Kadi4Mat, imparting structure and enabling access for all the users working on the project.

Kadi allows heterogeneous data formats and corresponding metadata, to answer the needs of an interdisciplinary research field such as materials science. The Bayesian oracle framework has been developed with the same vision in mind and therefore adapts to different (meta)data types. To support fast metadata collection, Kadi offers the possibility to define templates, in combination with validation functionality and, for some type of metadata (e.g. creation time), automatic recording.

For data usage, KadiAI integrates and standardizes AI projects, work packages, and workflows (Koeppe et al. 2021) into the Kadi ecosystem, serving as an interface between RDM and machine learning applications. KadiAI is combined with the Python-based framework CIDS, designed to facilitate the development and implementation of learning algorithms in AI workflows. Thus, the Kadi4Mat ecosystem can leverage data-integrated AI to enable quantitative and qualitative data analysis.

For sharing, Kadi4Mat allows management of access and editing rights for each project, as well as for each specific record. Different roles can be established for working groups or specific registered users. This feature is designed to enhance collaborations among different groups and institutions, while upholding the imperative of controlled data access.

In the vision of a partially-automated laboratory, all data are to be processed in an automated and reproducible manner. Therefore, we develop workflows in KadiStudio that take care of the consequential execution of each research step within a graphical user interface (GUI), using an intuitive point-and-click mechanism. The Bayesian oracle has been implemented as (a set of) nodes, basic building blocks of our workflow system (Figure 3). To monitor the optimization progress, a visualization of the Bayesian oracle current and global state is available in Kadi4Mat.

4 Conclusions

Establishing a common workspace and the possibility of sharing research data on a routine basis has proven to be crucial, especially for collaboration across multiple research groups. Workflows have streamlined the research process, making it faster and providing a simple and intuitive way for researchers, including non-experts, to reproduce their work.

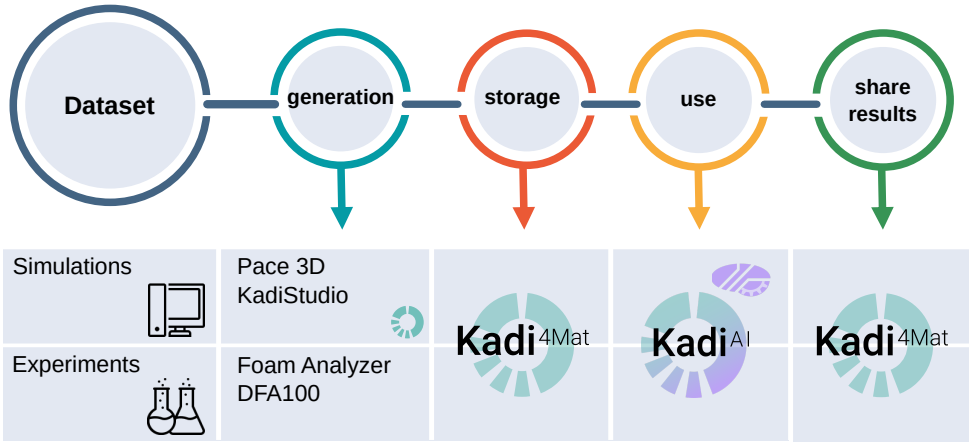

Figure 2: Kadi4Mat integration through the whole research process for the foam evolution use case.

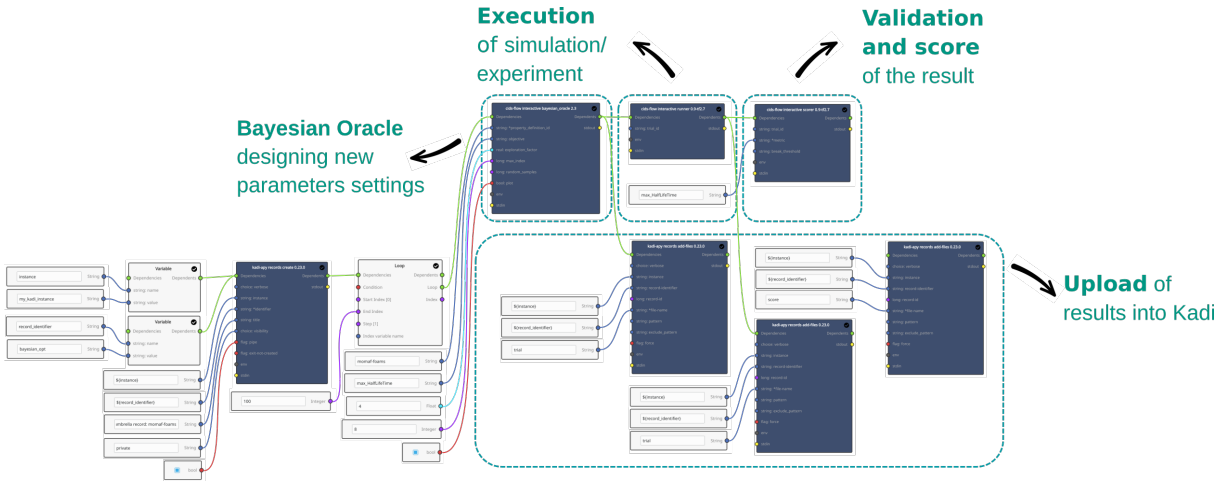

Figure 3: Example workflow implementation of the Bayesian oracle framework.

The possibility to store, share, and work with data in an intuitive manner plays a more and more important role in research. A structured working environment is provided for the day-to-day work of researchers from different groups, places, and disciplines.

Incorporating the Bayesian oracle in this workflow context, in the form of user-friendly nodes, provides a cost-effective solution accessible and utilizable for a broader range of users. Furthermore, this iterative approach presents a valuable opportunity to accelerate the process of materials' development.

Acknowledgements

This work is funded by the Ministry of Science, Research and Art Baden-Württemberg (MWK-BW) in the SDC MoMaF, with funds from the state digitization strategy digital@bw (project No. 57), the BMBF and MWK-BW as part of the Excellence Strategy

of the German Federal and State Governments in the project Kadi4X and the support of the Karlsruhe Nano Micro Facility (KNMFi), a Helmholtz Research Infrastructure at Karlsruhe Institute of Technology within the program MSE No. 43.31.01.

References

Brandt, Nico, Lars Griem, Christoph Herrmann, Ephraim Schoof, Giovanna Tosato, Yinghan Zhao, Philipp Zschumme, and Michael Selzer. 2021. *Kadi4Mat: A Research Data Infrastructure for Materials Science.* 20:8. 1. Ubiquity Press. DOI: https://doi.org/10.5334/dsj-2021-008.

Griem, Lars, Philipp Zschumme, Matthieu Laqua, Nico Brandt, Ephraim Schoof, Patrick Altschuh, and Michael Selzer. 2022. "KadiStudio: FAIR Modelling of Scientific Research Processes". *Data Science Journal* 21 (1): 16. ISSN: 1683-1470. DOI: https://doi.org/10.5334/dsj-2022-016.

Himanen, Lauri, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. 2019. "Data-Driven Materials Science: Status, Challenges, and Perspectives". *Advanced Science* 6 (21). DOI: https://doi.org/10.1002/advs.201900808.

Hötzer, Johannes, Andreas Reiter, Henrik Hierl, Philipp Steinmetz, Michael Selzer, and Britta Nestler. 2018. "The parallel multi-physics phase-field framework Pace3D". *Journal of Computational Science* 26:1–12. DOI: https://doi.org/10.1016/j.jocs.2018.02.011.

Jung, Ole, Mike Barbeck, L U Fan, Fabian Korte, Cuifeng Zhao, Rumen Krastev, Sven Pantermehl, and Xin Xiong. 2023. "Republication: In Vitro and Ex Vivo Analysis of Collagen Foams for Soft and Hard Tissue Regeneration". *In Vivo* 37 (1): 320–328. DOI: https://doi.org/10.21873/invivo.13082.

Koeppe, Arnd, Franz Bamer, Michael Selzer, Britta Nestler, and Bernd Markert. 2021. "Workflow concepts to model nonlinear mechanics with computational intelligence". *PAMM* 21 (1). ISSN: 1617-7061. DOI: https://doi.org/10.1002/pamm.202100238.

———. 2022. "Explainable artificial intelligence for mechanics: physics-explaining neural networks for constitutive models". *Frontiers in Materials* 8:636. DOI: https://doi.org/10.48550/arXiv.2104.10683.

Koeppe, Arnd, and CIDS Team. 2023. *CIDS: 3.1.* Zenodo. Visited on January 11, 2023. DOI: https://doi.org/10.5281/zenodo.7524476.

Kotthoff, Lars, Hud Wahab, and Patrick Johnson. 2021. *Bayesian Optimization in Materials Science: A Survey.* Technical report. Arxiv. DOI: https://doi.org/10.48550/arXiv.2108.00002.

Kuhn, Jannick, Jonathan Spitz, Petra Sonnweber-Ribic, Matti Schneider, and Thomas Böhlke. 2021. "Identifying material parameters in crystal plasticity by Bayesian optimization". *Optimization and Engineering* 23:1489–1523. ISSN: 15732924. DOI: https://doi.org/10.1007/s11081-021-09663-7.

Kulagin, Roman, Patrick Reiser, Kyryl Truskovskyi, Arnd Koeppe, Yan Beygelzimer, Yuri Estrin, Pascal Friederich, and Peter Gumbsch. 2023. "Lattice Metamaterials with Mesoscale Motifs: Exploration of Property Charts by Bayesian Optimization". *Advanced Engineering Materials* 25 (13). ISSN: 1438-1656. DOI: https://doi.org/10.1002/adem.202300048.

Settles, Burr. 2009. *Active Learning Literature Survey.* Technical report. https://burrsettles.com/pub/settles.activelearning.pdf.

Zhao, Yinghan, Patrick Altschuh, Jay Santoki, Lars Griem, Giovanna Tosato, Michael Selzer, Arnd Koeppe, and Britta Nestler. 2023. "Characterization of porous membranes using artificial neural networks". *Acta Materialia* 253:118922. ISSN: 1359-6454. DOI: https://doi.org/10.1016/j.actamat.2023.118922.