Understanding the Role of Domain Knowledge in Bayesian Optimization under Small-Data Constraints

Bernd Schuscha
Department Simulation
Materials Center Leoben Forschung GmbH
8700 Leoben, Austria
bernd.schuscha@mcl.at

Franz M. Rohrhofer
Department for Methods & Algorithms for AI
Know-Center Research GmbH
8010 Graz, Austria

Daniel Scheiber
Department Simulation
Materials Center Leoben Forschung GmbH
8700 Leoben, Austria

Bernhard C. Geiger
Know-Center Research GmbH
Signal Processing and Speech Communication Laboratory
Graz University of Technology
8010 Graz, Austria

Bayesian optimization (BO) is a sequential model-based optimizing algorithm for expensive black-box functions. The core is a probabilistic surrogate model, typically a Gaussian process, that is updated as new evaluations are collected. At each iteration, an acquisition function balances exploration of uncertain regions against exploitation of promising areas, guiding the selection of the next evaluation point [1]. It is widely used for data-efficient optimization of expensive black-box functions, particularly where evaluations are costly and data is scarce. In materials design and related fields, optimization problems are often accompanied by partial prior knowledge derived from physical models, simulations, or empirical relations. Incorporating such knowledge can in principle improve sample efficiency, but it is not clear when and where in the BO pipeline this knowledge should be injected, nor how sensitive each injection point is to knowledge of varying quality, a critical concern in the small-data regimes typical of experimental materials science, where few evaluations are affordable.

Recent works have proposed multiple strategies to incorporate domain knowledge into BO [2–9]. These include physics-informed surrogate models, expert-informed priors over the search space, and modifications of the acquisition function. While each of these approaches has demonstrated potential benefits, it remains unclear where knowledge should be injected in the BO pipeline to most effectively influence optimization dynamics on a wider range of problems. In this work, we provide a systematic empirical comparison of three conceptually distinct knowledge-injection strategies. Our goal is not to introduce a new optimization algorithm, but to investigate how the placement and flexibility of domain knowledge affect sample efficiency and robustness in small-data regimes.

1 Conceptual Framework

To structure the comparison, we express the different strategies in a unified conceptual form:

$$x_n = \arg\max_{x \in \mathcal{X}} \alpha[g](x)\, \pi(x) + R(x),$$

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

where $\alpha[g](x)$ denotes the acquisition function derived from a surrogate model g, $\pi(x)$ represents a prior over the search space, and $R(x)$ denotes an optional regularization term encoding auxiliary knowledge.

This formulation highlights that domain knowledge can influence Bayesian optimization by shaping the surrogate model (hypothesis space), biasing the search distribution via priors, or modifying the decision criterion through the acquisition function. This formulation serves as a conceptual abstraction to compare different integration strategies and does not define a new optimization method. We consider three placements of domain knowledge:

(i) Surrogate-level injection: Domain knowledge is embedded directly into the surrogate model, for example via physics-inspired mean functions, feature transformations, or domain-informed kernels [3].

(ii) Prior-based injection: Knowledge is encoded as a probabilistic prior over the search space, influencing candidate selection multiplicatively while allowing its impact to decrease over time[7].

(iii) Acquisition-level regularization: An auxiliary domain model contributes additively at the acquisition stage, guiding exploration without altering the surrogate [8, 9].

These strategies differ in how strongly they constrain the hypothesis space and how they interact with uncertainty estimation.

2 Experimental Overview

We evaluate the three strategies across different optimization tasks, including the Branin-Currin multi-objective problem [10], which provides a tractable yet non-trivial testbed for multi-objective BO under controlled conditions.

For the prior- and regularization-level injection strategies, a Gaussian process regression (GPR) is employed as the surrogate model. The tasks vary in dimensionality, objective structure, and noise level. All experiments are conducted under limited evaluation budgets to reflect small-data conditions. Performance is quantified using cumulative normalized hypervolume regret, computed on ground-truth objective values to avoid model-dependent bias. To isolate the effect of knowledge placement, we maintain consistent surrogate architectures and acquisition functions across experiments. Each configuration is repeated five times with different initial points and noise realizations.

All experiments are conducted under strict small-data conditions, with adapting evaluation budgets limited to a maximum of 100 function evaluations, reflecting realistic experimental constraints in materials science where each experiment may require significant time and cost.

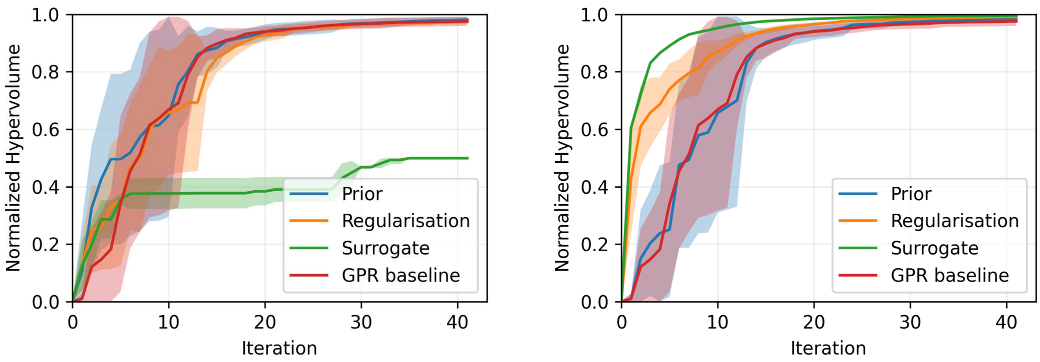
3 Observations

We consider both accurate and partially correct domain knowledge, where partial correctness reflects domain models that capture some but not all structural aspects of the true objective. Across tasks, we observe that the impact of domain knowledge depends strongly on both its placement and its flexibility.

Surrogate-level injection accelerates early-stage learning when the embedded knowledge is adaptable and reasonably aligned with the true objective structure. However, rigid or misspecified knowledge can severely degrade optimization performance. This behavior is illustrated in Fig. 1a.

Prior-based injection provides modest improvements in the early stages but does not consistently yield significant gains compared to vanilla Bayesian optimization. In the present formulation, the injected prior is soft and therefore mainly acts as a term that slightly biases surrogate predictions without fundamentally altering the acquisition landscape. As observational data accumulate, the GPR output becomes more certain over the domain, leading to sampling trajectories that are largely similar to those obtained with a non-informative prior.

Acquisition-level regularization offers flexible guidance. Although it is outperformed by surrogate-level injection when the incorporated knowledge closely matches the ground truth, it provides a more robust mechanism for integrating imperfect domain knowledge. By influencing the ranking of candidate points directly through the acquisition function, this strategy maintains beneficial exploratory behavior even when the injected information is only partially correct. This effect is illustrated in Fig. 1b.



(a) BO with partially correct knowledge

(b) BO with accurate knowledge

Figure 1: Types of knowledge injected in the Branin-Currin multi-objective Bayesian optimization problem

Overall, these results highlight that the effectiveness of knowledge injection in Bayesian optimization depends both on how strongly the injected information constrains the surrogate model and on the level at which it influences the decision-making process. Strong surrogate-level guidance can substantially improve early sample efficiency when well aligned with the objective structure, but increases sensitivity to misspecification. In contrast, more flexible strategies, particularly those acting at the acquisition level, provide more robust performance by preserving the optimizer's ability to adapt to observed data.

This reveals a fundamental trade-off between performance and robustness, governed by the level at which domain knowledge is introduced.

4 Discussion and Outlook

These preliminary findings indicate that, in small-data regimes where only a handful of initial observations are available and the total evaluation budget is tightly constrained, the placement of domain knowledge within the BO pipeline critically determines optimization dynamics. The interaction between knowledge quality, surrogate flexibility, and uncertainty calibration becomes particularly consequential precisely because there is insufficient data to self-correct early misguidance. Rather than asking whether domain-informed BO is beneficial in general, our results suggest that the interaction between knowledge flexibility, uncertainty calibration, and noise level is decisive.

Future work will focus on deeper theoretical analysis of these interactions and on studying fewer real-world materials systems in greater detail to better understand the mechanistic origins of the observed behaviors.

Acknowledgments and Disclosure of Funding

The 'Think Before You Sample: Physics-Informed Active Learning for Green AI' action has received funding from the European Union, via the oc3-2025-TES-01 issued and implemented by the ENFIELD project, under the grant agreement No 101120657.

We acknowledge access to LEONARDO at CINECA, Italy, via an AURELEO (Austrian Users at LEONARDO supercomputer) project.

The authors gratefully acknowledge the financial support under the scope of the COMET program within the K2 Center "Integrated Computational Material, Process and Product Engineering (IC-MPPE)" (Project No 886385). This program is supported by the Austrian Federal Ministries for Economy, Energy and Tourism (BMWET) and for Innovation, Mobility and Infrastructure (BMIMI), represented by the Austrian Research Promotion Agency (FFG), and the federal states of Styria, Upper Austria and Tyrol

References

[1] Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions. *arXiv preprint arXiv:1012.2599*, 2010.

[2] Sean Hooten, Wolfger Peelaers, Thomas Van Vaerenbergh, and Marco Fiorentino. Enhancing scientific bayesian optimization via physics-informed operator priors.

[3] Danial Khatamsaz, Raymond Neuberger, Arunabha Roy, Sina Zadeh, Richard Otis, and Raymundo Arróyave. A physics informed bayesian optimization approach for material design: application to niti shape memory alloys. *npj Computational Materials*, 9(1), 11 2023. ISSN ISSN 2057-3960. doi: 10.1038/s41524-023-01173-7.

[4] Wataru Kobayashi, Takuma Otsuka, Yuki K. Wakabayashi, and Gensai Tei. Physics-informed Bayesian optimization suitable for extrapolation of materials growth. *npj Computational Mathematics*, 11(1):36, February 2025. doi: 10.1038/s41524-025-01522-8.

[5] Maxim A Ziatdinov, Ayana Ghosh, and Sergei V Kalinin. Physics makes the difference: Bayesian optimization and active learning via augmented gaussian process. *Machine Learning: Science and Technology*, 3(1):015003, feb 2022. doi: 10.1088/2632-2153/ac4baa.

[6] Francesco Di Fiore and Laura Mainini. Physics-aware multifidelity bayesian optimization: A generalized formulation. *Computers & Structures*, 296:107302, 2024. ISSN 0045-7949. doi: https://doi.org/10.1016/j.compstruc.2024.107302.

[7] Carl Hvarfner, Danny Stoll, Artur Souza, Marius Lindauer, Frank Hutter, and Luigi Nardi. πbo: Augmenting acquisition functions with user beliefs for bayesian optimization, 04 2022.

[8] Zikai Xie, Xenophon Evangelopoulos, Joseph Thacker, and Andrew Cooper. *Domain Knowledge Injection in Bayesian Search for New Materials*. 09 2023. ISBN 9781643684369. doi: 10.3233/FAIA230587.

[9] Man Luo, Zikai Xie, Huirong Li, Baicheng Zhang, Jiaqi Cao, Yan Huang, Hang Qu, Qing Zhu, Linjiang Chen, Jun Jiang, and Yi Luo. Physics-informed, dual-objective optimization of high-entropy-alloy nanozymes by a robotic ai chemist. *Matter*, 8:102009, 03 2025. doi: 10.1016/j.matt.2025.102009.

[10] Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.