



Volume 15, Issue 11, November 2025

Bayesian Optimisation (Dagstuhl Seminar 25451) <i>Jürgen Branke, Frank Hutter, Giulia Pedrielli, Matthias Poloczek, and Leonard Papenmeier</i>	1
A Roadmap Towards Practical Applications of Neurosymbolic Learning and Reasoning (Dagstuhl Seminar 25452) <i>Thiviyan Thanapalasingam, Annette ten Teije, and Frank van Harmelen</i>	67
Security and Privacy of Large Language Models (Dagstuhl Seminar 25461) <i>Stephan Günnemann, Pavel Laskov, Emil Lupu, Vera Rimmer, Advije Rizvani, and Qianying Liao</i>	87
Online Algorithms beyond Competitive Analysis (Dagstuhl Seminar 25471) <i>Sungjin Im, Nicole Megow, Debmalaya Panigrahi, Sahil Singla, and Golnoosh Shahkarami</i>	114
Approaches and Applications of Inductive Programming (Dagstuhl Seminar 25491) <i>Ute Schmid, Gust Verbruggen, and Sonja Niemann</i>	134
Generalized Voronoi Diagrams and Applications (Dagstuhl Seminar 25492) <i>Andreas Alpers, David P. Bourne, Henning F. Poulsen, and Claudia Redenbach</i> ..	157

ISSN 2192-5283

Published online and open access by

Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at <https://www.dagstuhl.de/dagpub/2192-5283>

Publication date

July, 2026

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <https://dnb.d-nb.de>.

License

This work is licensed under a Creative Commons Attribution 4.0 International license (CC BY 4.0).



In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work under the following conditions, without impairing or restricting the authors' moral rights:

- Attribution: The work must be attributed to its authors.

The copyright is retained by the corresponding authors.

Aims and Scope

The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.

In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

This basic framework can be extended by suitable contributions that are related to the program of the seminar, e. g. summaries from panel discussions or open problem sessions.

Editorial Board

- Elisabeth André
- Franz Baader
- Goetz Graefe
- Reiner Hähnle
- Barbara Hammer
- Lynda Hardman
- Steve Kremer
- Rupak Majumdar
- Heiko Mantel
- Lennart Martens
- Albrecht Schmidt
- Wolfgang Schröder-Preikschat
- Holger Hermanns (*Editor-in-Chief*)
- Heike Wehrheim
- Verena Wolf
- Martina Zitterbart

Editorial Office

Michael Wagner (*Managing Editor*)
Michael Didas (*Managing Editor*)
Jutka Gasiorowski (*Editorial Assistance*)
Dagmar Glaser (*Editorial Assistance*)
Thomas Schillo (*Technical Assistance*)
Christina Schwarz (*Editorial Assistance*)

Contact

Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de

<https://www.dagstuhl.de/dagrep>

Digital Object Identifier: 10.4230/DagRep.15.11.i

Bayesian Optimisation

Jürgen Branke^{*1}, Frank Hutter^{*2}, Giulia Pedrielli^{*3},
Matthias Poloczek^{*4}, and Leonard Papenmeier^{†5}

1 University of Warwick, GB. juergen.branke@wbs.ac.uk

2 Prior Labs – Freiburg, DE & ELLIS Institute Tübingen, DE & Universität
Freiburg, DE. fh@cs.uni-freiburg.de

3 Arizona State University – Tempe, US. giulia.pedrielli@asu.edu

4 Amazon.com – San Francisco, US. matthias.poloczek@gmx.de

5 Universität Münster, DE. Leonard.Papenmeier@uni-muenster.de

Abstract

This report documents the programme and outcomes of Dagstuhl Seminar 25451, “Bayesian Optimisation”, held from November 2–7, 2025. The seminar brought together 39 international experts from machine learning, optimisation, statistics, and engineering to discuss recent advances, open challenges, and emerging research directions in Bayesian optimisation. The programme comprised plenary talks spanning foundational issues, benchmarking, and the growing interaction between Bayesian optimisation and generative AI, alongside focused working groups on key thematic areas. Beyond technical discussions, the seminar placed strong emphasis on community building and worked towards establishing best practices. This report summarises the plenary contributions, the outcomes of the working groups, and additional community-driven activities, including a collection of practical “tricks of the trade” and exploratory benchmarking exercises.

Seminar November 2–7, 2025 – <https://www.dagstuhl.de/25451>

2012 ACM Subject Classification Mathematics of computing → Bayesian computation; Computing methodologies → Gaussian processes; Computing methodologies → Machine learning approaches

Keywords and phrases AutoML, Bayesian optimization, Benchmarking, Gaussian processes

Digital Object Identifier 10.4230/DagRep.15.11.1

1 Executive Summary

Jürgen Branke (University of Warwick, GB)

Frank Hutter (Prior Labs – Freiburg, DE & ELLIS Institute Tübingen, DE & Universität Freiburg, DE)

Giulia Pedrielli (Arizona State University – Tempe, US)

Matthias Poloczek (Amazon.com – San Francisco, US)

License  Creative Commons BY 4.0 International license

© Jürgen Branke, Frank Hutter, Giulia Pedrielli, and Matthias Poloczek

The Dagstuhl Seminar “Bayesian Optimisation” (November 2–7, 2025) brought together 39 leading researchers from 11 countries to discuss the state of the art, emerging challenges, and future directions of Bayesian optimisation (BO). BO is a core methodology for optimizing expensive black-box functions and lies at the intersection of machine learning, optimisation,

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Bayesian Optimisation, *Dagstuhl Reports*, Vol. 15, Issue 11, pp. 1–66

Editors: Jürgen Branke, Frank Hutter, Giulia Pedrielli, and Matthias Poloczek



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

and statistics, with strong contributions from engineering and applied sciences. It combines probabilistic surrogate modelling with principled decision-making to determine which evaluations are most informative under limited budgets.

Although foundational ideas in Bayesian optimisation date back to the 1950s, the field has experienced rapid growth only since around 2010, driven by advances in Gaussian process modelling, computational power, and high-impact applications such as hyperparameter tuning of machine learning methods, simulation-based optimisation, and experimental design. Despite this momentum, BO remains a fragmented research area: there is no dedicated flagship conference, and researchers are dispersed across multiple communities with differing assumptions, evaluation practices, and terminologies. A central aim of the seminar was therefore community building – bringing together researchers from machine learning, operations research, statistics, and engineering to develop a shared understanding of challenges, benchmarks, and research priorities.

The seminar programme featured 12 plenary talks designed to stimulate discussion across subfields, covering topics ranging from the historical development of BO and benchmarking practices to recent advances in high-dimensional optimisation, human-in-the-loop methods, and the integration of large language models and generative AI. In addition, participants worked in five focused working groups on themes identified as critical for the community: human-in-the-loop Bayesian optimisation; multi-fidelity Bayesian optimisation; acquisition strategies; surrogate modelling; and the intersection of Bayesian optimisation with generative AI. The outcomes of these discussions, together with abstracts of the plenary talks, are documented in this report.

Beyond structured discussions, the seminar also facilitated the exchange of practical insights. Participants jointly compiled a collection of “tricks of the trade” reflecting practical wisdom that is often absent from the formal literature. In addition, attendees were invited to apply their preferred BO tools to a small set of benchmark problems, providing a concrete basis for discussions on benchmarking, software libraries, and best practices. Summaries of these activities are also included in the report.

Schloss Dagstuhl once again proved to be an ideal environment for intensive scientific exchange and community formation. Both organisers and participants regarded the seminar as a significant success. Recognising the importance of sustained interaction for this rapidly evolving field, the participants established a Steering Group tasked with coordinating future meetings and related community-building activities. This seminar thus represents not only a snapshot of current research in Bayesian optimisation, but also a foundation for a more cohesive and collaborative BO community going forward.

2 Table of Contents

Executive Summary	
<i>Jürgen Branke, Frank Hutter, Giulia Pedrielli, and Matthias Poloczek</i>	1
Seminar Program	5
Overview of Talks	6
Possible snags in benchmarking noisy or high-dimensional BO	
<i>Mickaël Binois</i>	6
Convergence from any initial design: is it an utopia?	
<i>Antonio Candelieri</i>	6
Bayesian Optimization On and Over Graphs	
<i>Gautam Dasarathy</i>	7
Benchmarking BO – do we really want to learn from neighbouring disciplines?	
<i>Carola Doerr</i>	8
Some things I learned while writing the BayesOpt book	
<i>Roman Garnett</i>	9
Triangulation candidates for Bayesian optimization	
<i>Robert Gramacy</i>	9
Prior-Data Fitted Networks	
<i>Frank Hutter</i>	10
Towards Foundation Models For Black-Box Optimization	
<i>Aaron Klein</i>	10
Bayesian Optimization Without a Clear Objective: Leveraging Interactive Natural Language Feedback	
<i>Maximilian Balandat</i>	10
The Role of Benchmarks in High-Dimensional Bayesian Optimization	
<i>Leonard Papenmeier</i>	11
Automated Generation of Bayesian Optimization Algorithms assisted by LLMs	
<i>Elena Raponi</i>	11
How are transformers and potential AGI affecting Bayesian optimization research?	
<i>Alexander Terenin</i>	12
The Gittins Index: A Design Principle for Decision-Making Under Uncertainty	
<i>Alexander Terenin</i>	12
Working Groups	13
Acquisition Strategies	
<i>Thomas Bartz-Beielstein, Jack Buckingham, Carola Doerr, Matthias Feurer, Roman Garnett, Leonard Papenmeier, and Alexander Terenin</i>	13
Multifidelity	
<i>Mauricio Alvarez, Maximilian Balandat, Cédric Archambeau, Ivo Couckuyt, Frank Hutter, Aaron Klein, Neeratyoy Mallik, Ruth Misener, and Juan Ungredda</i>	19

Human-in-the-loop Bayesian Optimization: Explainability and Interactivity <i>Jürgen Branke, Katharina Eggensperger, Jonathan Fieldsend, Daniel Hernández-Lobato, Mike McCourt, and Sebastian Rojas Gonzalez</i>	25
Surrogate Models for Bayesian Optimization <i>Steven Adriaensen, Nathalie Bartoli, Mickaël Binois, Robert Gramacy, Rodolphe Le Riche, Szu Hui Ng, Inneke Van Nieuwenhuyse, Jixiang Qing, and Antonio Candelieri</i>	36
GenAI and Bayesian Optimization <i>Raul Astudillo, Gautam Dasarathy, Peter Frazier, Daolang Huang, Bryan Kian Hsiang Low, Giulia Pedrielli, Matthias Poloczek, and Elena Raponi</i>	50
Additional Content	59
Benchmarking Bayesian Optimization in Practice: Lessons from a Dagstuhl Seminar Exercise <i>Giulia Pedrielli</i>	59
Bayesian Optimization Tricks Of The Trade <i>Jürgen Branke, Giulia Pedrielli, and Matthias Poloczek</i>	64
Participants	66

3 Seminar Program

Monday 3 November

9:00	Welcome and Introductions
10:45	Formation of Working Groups
12:15	Lunch
13:00	Roman Garnett: History of Bayesian Optimisation
13:30	Max Balandat: Bayesian Optimisation without a clear objective
14:00	Frank Hutter: Prior Fitted Networks
15:00	Coffee Break
16:00	Working Groups
18:00	Dinner

Tuesday 4 November

9:00	Carola Doerr: Benchmarking in the broader black-box optimization context: problems, performance criteria, and data sharing
9:30	Leonard Papenmeier: The Role of Benchmarks in (High-Dimensional) Bayesian Optimization
10:00	Mickaël Binois: Possible snags in benchmarking noisy or high-dimensional BO
10:30	Coffee Break
10:45	Discussion on Pre-Seminar Experimental Results
12:15	Lunch
13:30	Working Groups
15:30	Coffee Break
16:15	Aaron Klein: Towards Foundation Models for Blackbox Optimization
16:45	Elena Raponi: Automatic generation of Bayesian optimization algorithms using LLMs
17:15	Alexander Terenin: How are transformers and potential AGI affecting Bayesian optimization research?
18:00	Dinner

Wednesday 5 November

9:00	Working Group Reports
10:30	Coffee Break
10:45	Working Groups
12:15	Lunch
13:15	Group Photo
13:30	Hiking
15:30	Coffee Break
18:00	Dinner

Thursday 6 November

9:00	Working Groups
10:30	Coffee Break
10:45	Working Groups
12:15	Lunch
13:30	Bobby Gramacy: Triangulation Candidates for Bayesian Optimization
14:10	Antonio Candelieri: Convergence from any initial design: Is it utopia?
14:50	Gautam Dasarathy: Bayesian Optimization on and over graphs
15:30	Coffee Break
16:00	Working Groups
18:00	Dinner

Friday 7 November

9:15	Working Groups Reporting
10:45	Coffee Break
11:00	Tricks of the Trade
11:45	Next meetings
12:15	Lunch

4 Overview of Talks**4.1 Possible snags in benchmarking noisy or high-dimensional BO**

Mickaël Binois (Inria Center at Université Côte d’Azur – Sophia Antipolis, FR)

License  Creative Commons BY 4.0 International license
© Mickaël Binois

In this short presentation, the goal is to highlight possible issues of common practices for assessing BO on noisy or high-dimensional benchmark problems. For noisy problems, it focuses on the challenge of estimating the current best solution. As for high-dimensional test problems, besides general issues for comparing different baselines, the usual practice of adding inactive variables is questioned.

4.2 Convergence from any initial design: is it an utopia?

Antonio Candelieri (University of Milano-Bicocca, IT)

License  Creative Commons BY 4.0 International license
© Antonio Candelieri

Although its well-known sample efficiency and rate convergence proofs for some implementations, Bayesian Optimization’s performance remains strongly affected from the random initial design. A common practice consists into running multiple experiments from different initial designs to mitigate the effect of randomness and analyse robustness of different algorithms. Although this procedure is reasonable for scientific research, it is quite often impractical in real-life applications, due to the high query cost. Moreover, it is not rare to start from

an initial design that has been collected for other – industrial-driven – purposes and that can therefore drastically affect the identification of the optimal solution given a strictly limited query budget. This talk is aimed at quickly revising some of the known issues preventing Bayesian Optimization from convergence to the global optimum, while facilitating the theoretical and methodological discussion about possible solutions to turn an utopia into reality.

References

- 1 A. Candelieri and E. Signori, *MLE-free Gaussian Process Based Bayesian Optimization*, in International Conference on Learning and Intelligent Optimization, pp. 81–95, Cham: Springer Nature Switzerland, June 2024.
- 2 A. Candelieri, *Mastering the Exploration-Exploitation Trade-off in Bayesian Optimization*, arXiv preprint arXiv:2305.08624, 2023.
- 3 C. Benjamins, E. Raponi, A. Jankovic, K. van der Blom, M. L. Santoni, M. Lindauer, and C. Doerr, *Pi Is Back! Switching Acquisition Functions in Bayesian Optimization*, arXiv preprint arXiv:2211.01455, 2022.
- 4 G. De Ath, R. M. Everson, A. A. Rahat, and J. E. Fieldsend, *Greed Is Good: Exploration and Exploitation Trade-offs in Bayesian Optimisation*, ACM Transactions on Evolutionary Learning and Optimization, 1(1), 1–22, 2021.
- 5 J. Berk, S. Gupta, S. Rana, and S. Venkatesh, *Randomised Gaussian Process Upper Confidence Bound for Bayesian Optimisation*, arXiv preprint arXiv:2006.04296, 2020.
- 6 F. Berkenkamp, A. P. Schoellig, and A. Krause, *No-regret Bayesian Optimization with Unknown Hyperparameters*, Journal of Machine Learning Research, 20(50), 1–24, 2019.
- 7 K. P. Wabersich and M. Toussaint, *Advancing Bayesian Optimization: The Mixed Global-Local (MGL) Kernel and Length-scale Cool Down*, arXiv preprint arXiv:1612.03117, 2016.
- 8 Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. de Freitas, *Bayesian Optimization in a Billion Dimensions via Random Embeddings*, Journal of Artificial Intelligence Research, 55, 361–387, 2016.
- 9 Z. Wang and N. de Freitas, *Theoretical Analysis of Bayesian Optimisation with Unknown Gaussian Process Hyper-parameters*, arXiv preprint arXiv:1406.7758, 2014.
- 10 L. Papenmeier, M. Poloczek, and L. Nardi, *Understanding High-dimensional Bayesian Optimization*, arXiv preprint arXiv:2502.09198, 2025.
- 11 C. Hvarfner, E. O. Hellsten, and L. Nardi, *Vanilla Bayesian Optimization Performs Great in High Dimensions*, Proceedings of Machine Learning Research, 235, 20793–20817, 2024.

4.3 Bayesian Optimization On and Over Graphs

Gautam Dasarathy (Arizona State University – Tempe, US)

License © Creative Commons BY 4.0 International license
© Gautam Dasarathy

Main reference Parth Thaker, Mohit Malu, Nikhil Rao, Gautam Dasarathy: “Maximizing and Satisficing in Multi-armed Bandits with Graph Information”, in Proc. of the Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 – December 9, 2022, 2022.

URL http://papers.nips.cc/paper_files/paper/2022/hash/0d561979f0f4bc6127cfcfe9c46ee205-Abstract-Conference.html

Graphs provide a natural language for representing complex, structured systems – from molecules and neural architectures to sensor networks and social systems. Bayesian Optimization (BO) offers a powerful framework for optimizing expensive black-box objectives in such domains, but incorporating graph structure raises new challenges.

This talk explores Bayesian Optimization on and over graphs. I will discuss recent advances that adapt Gaussian-process surrogates and acquisition strategies to graph domains, highlight key theoretical and computational challenges, and outline open questions in structure-aware optimization. A central theme is the role of graph topology – how spectral and combinatorial properties shape learning efficiency. I will also connect these ideas to some of our work on graph bandits, adaptive graph learning, and multifidelity learning, which may hint at a theory of Bayesian optimization that truly understands graphs.

4.4 Benchmarking BO – do we really want to learn from neighbouring disciplines?

Carola Doerr (Sorbonne University – Paris, FR)

License  Creative Commons BY 4.0 International license
© Carola Doerr

Benchmarking is a crucial component in understanding advantages and limitations of Bayesian optimization algorithms. However, to date, researchers in the BO community did not agree on standardized way of benchmarking their algorithms; most papers use ad hoc experimental setups, data sharing and re-using is uncommon.

With the goal to spark a discussion on how to overcome this situation, I will share some benchmarking best practices developed in the broader black-box optimization community. Largely influenced by the COCO platform [2], available at <https://coco-platform.org/>, standardized benchmarking practices and data formats were adopted in the evolutionary computation community. Large data sets are available and explorable through the IO-Hanalyzer component [3] of the modular IOHprofiler environment [1], available at <https://iohanalyzer.liacs.nl/>.

My key take is that it would be much easier for the Bayesian optimization community to adopt similar benchmarking standards than what one might believe.

References

- 1 C. Doerr, H. Wang, F. Ye, S. van Rijn, and T. Bäck, *IOHprofiler: A Benchmarking and Profiling Tool for Iterative Optimization Heuristics*, CoRR, abs/1810.05281, 2018. Available at: <http://arxiv.org/abs/1810.05281>. More up-to-date documentation: <https://iohprofiler.github.io/>.
- 2 N. Hansen, A. Auger, R. Ros, O. Mersmann, T. Tušar, and D. Brockhoff, *COCO: A Platform for Comparing Continuous Optimizers in a Black-box Setting*, Optimization Methods and Software, pp. 1–31, 2020.
- 3 H. Wang, D. Vermetten, F. Ye, C. Doerr, and T. Bäck, *IOHanalyzer: Detailed Performance Analyses for Iterative Optimization Heuristics*, ACM Transactions on Evolutionary Learning and Optimization, 2(1), Article 3, pp. 3:1–3:29, 2022.

4.5 Some things I learned while writing the BayesOpt book

Roman Garnett (Washington University – St. Louis, US)

License © Creative Commons BY 4.0 International license
© Roman Garnett

Main reference Roman Garnett: “Bayesian Optimization”. Cambridge University Press, 2023.

URL <https://doi.org/10.1017/9781108348973>

I spoke about two somewhat niche topics that came up while working on the Bayesian optimization book:

- a brief history of Bayesian optimization, and some commentary on the origin of popular acquisition strategies (UCB, probability of improvement, expected improvement, etc.), which are almost universally misattributed in the literature
- methodology for marginalizing model hyperparameters through an acquisition strategy, and the pitfalls of the common approach of averaging the acquisition function wrt the hyperparameters rather than computing the expected utility with respect to the hyperparameter-marginal model directly

This led to a brief, but engaging discussion.

4.6 Triangulation candidates for Bayesian optimization

Robert Gramacy (Virginia Polytechnic Institute – Blacksburg, US)

License © Creative Commons BY 4.0 International license
© Robert Gramacy

Joint work of Robert Gramacy, Annie Booth, John Smith, Nathan Wycoff

Main reference Robert B. Gramacy, Annie Sauer, Nathan Wycoff: “Triangulation candidates for Bayesian optimization”, in Proc. of the Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 – December 9, 2022, 2022.

URL http://papers.nips.cc/paper_files/paper/2022/hash/e9750610639c3e7a849cff746bf60dbd-Abstract-Conference.html

Bayesian optimization involves “inner optimization” over a new-data acquisition criterion which is non-convex/highly multi-modal, may be non-differentiable, or may otherwise thwart local numerical optimizers. In such cases it is common to replace continuous search with a discrete one over random candidates. Here we propose using candidates based on a Delaunay triangulation of the existing input design. We detail the construction of these “tricands” and demonstrate empirically how they outperform both numerically optimized acquisitions and random candidate-based alternatives, and are well-suited for hybrid schemes, on benchmark synthetic and real simulation experiments.

4.7 Prior-Data Fitted Networks

Frank Hutter (Prior Labs – Freiburg, DE & ELLIS Institute Tübingen, DE & Universität Freiburg, DE)

License  Creative Commons BY 4.0 International license
 Frank Hutter

In this talk, I will discuss a new alternative to the predominant model class used in Bayesian optimization: prior-data fitted networks (PFNs). Like Gaussian processes, PFNs have strong uncertainty estimates, but unlike them they easily support many data complexities, such as mixed discrete/continuous inputs, heterogeneous noise, high dimensions, and good out-of-distribution generalization.

I will first discuss the general method for constructing PFNs and will then deepdive into one particular example: TabPFN, the first foundation model for tabular data. I hope that many colleagues in Bayesian optimization will start to explore PFNs and/or TabPFN for their applications and help develop them further.

4.8 Towards Foundation Models For Black-Box Optimization

Aaron Klein (ScaDS.AI – Leipzig, DE)

License  Creative Commons BY 4.0 International license
 Aaron Klein

Joint work of Aaron Klein, Herilalaina Rakotoarison, Luca Thale-Bombien, David Salinas

Black-box optimization has so far relied on hand-crafted algorithms, carefully tuned for specific problem classes and rarely transferable beyond their original domains. Foundation models hold the promise of fundamentally changing this landscape: rather than manually designing optimizers, we can envision models that learn the process of optimization itself. Unlike representation learning in vision or language, this requires learning algorithms – arguably a substantially more ambitious goal – but one that recent advances in reasoning models suggest may be within reach.

In this talk, we will provide an overview of recent trends in Bayesian optimization for building foundation models and highlight current gaps in the literature.

4.9 Bayesian Optimization Without a Clear Objective: Leveraging Interactive Natural Language Feedback

Maximilian Balandat (Meta – Menlo Park, US)

License  Creative Commons BY 4.0 International license
 Maximilian Balandat

Joint work of Katarzyna Kobalcyk, Zhiyuan Jerry Lin, Benjamin Letham, Zhuokai Zhao, Maximilian Balandat, Eytan Bakshy

Main reference Katarzyna Kobalcyk, Zhiyuan Jerry Lin, Benjamin Letham, Zhuokai Zhao, Maximilian Balandat, Eytan Bakshy: “LILLO: Bayesian Optimization with Interactive Natural Language Feedback”, CoRR, Vol. abs/2510.17671, 2025.

URL <https://doi.org/10.48550/ARXIV.2510.17671>

For many real-world applications, feedback is essential in translating complex, nuanced, or subjective goals into quantifiable optimization objectives. We propose a language-in-the-loop framework that uses a large language model (LLM) to convert unstructured feedback in

the form of natural language into scalar utilities to conduct BO over a numeric search space. Unlike preferential BO, which only accepts restricted feedback formats and requires customized models for each domain-specific problem, our approach leverages LLMs to turn varied types of textual feedback into consistent utility signals and to easily include flexible user priors without manual kernel design. At the same time, our method maintains the sample efficiency and principled uncertainty quantification of BO. We show that this hybrid method not only provides a more natural interface to the decision maker but also outperforms conventional BO baselines and LLM-only optimizers, particularly in feedback-limited regimes.

4.10 The Role of Benchmarks in High-Dimensional Bayesian Optimization

Leonard Papenmeier (Universität Münster, DE)

License © Creative Commons BY 4.0 International license
© Leonard Papenmeier

Joint work of Leonard Papenmeier, Matthias Poloczek, Luigi Nardi

Main reference Leonard Papenmeier, Matthias Poloczek, Luigi Nardi: “Understanding High-Dimensional Bayesian Optimization”, in Proc. of the Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025, Proceedings of Machine Learning Research, Vol. 267, PMLR / OpenReview.net, 2025.

URL <https://proceedings.mlr.press/v267/papenmeier25a.html>

High-dimensional Bayesian optimization (HDBO) is a notoriously difficult challenge. This is due to the curse of dimensionality that, with linearly increasing dimensionality, requires exponentially more observations to maintain the same level of surrogate model accuracy. Recent advances have pushed the envelope to hundreds or even thousands of dimensions on various benchmarks that reflect controlled environments and real-world scenarios. This talk gives insight into specific structural properties of several of these benchmarks, how these properties interplay with design decisions of BO and high-dimensional BO algorithms, and what implicit assumptions several BO algorithms make on the objective function. The goal of this talk is to spark a discussion around best practices for HDBO benchmarking.

4.11 Automated Generation of Bayesian Optimization Algorithms assisted by LLMs

Elena Raponi (Leiden University, NL)

License © Creative Commons BY 4.0 International license
© Elena Raponi

Joint work of Wenhui Li, Niki van Stein, Thomas Bäck, Elena Raponi

Main reference Wenhui Li, Niki van Stein, Thomas Bäck, Elena Raponi: “LLaMEA-BO: A Large Language Model Evolutionary Algorithm for Automatically Generating Bayesian Optimization Algorithms”, CoRR, Vol. abs/2505.21034, 2025.

URL <https://doi.org/10.48550/ARXIV.2505.21034>

Large Language Models (LLMs) are rapidly becoming creative co-pilots for AI, Optimization and AutoML, moving beyond tuning to generating and evolving entire algorithms on demand. This talk introduces LLaMEA-BO, an evolutionary computation framework that uses LLMs to iteratively mutate and improve full codebases, with specific focus on Bayesian Optimization. Our framework uses an evolution strategy to guide an LLM in generating Python code that preserves the key components of BO algorithms: An initial design, a surrogate model, and an

acquisition function. The LLM is prompted to produce multiple candidate algorithms, which are evaluated on the established Black-Box Optimization Benchmarking (BBOB) test suite from the COmparing Continuous Optimizers (COCO) platform. Based on their performance, top candidates are selected, combined, and mutated via controlled prompt variations, enabling iterative refinement. Despite no additional fine-tuning, the LLM-generated algorithms outperform state-of-the-art BO baselines in 19 (out of 24) BBOB functions in dimension 5 and generalize well to higher dimensions, and different tasks (from the Bayesmark framework).

4.12 How are transformers and potential AGI affecting Bayesian optimization research?

Alexander Terenin (Cornell University – Ithaca, US)

License © Creative Commons BY 4.0 International license
© Alexander Terenin

Since the release of ChatGPT, a non-trivial segment of the machine learning community believes that human and/or superhuman artificial general intelligence will arrive in the next half-decade. While one could certainly argue that people with this view are wrong, I would instead like to ask: What if they're right? What would the implications for our research community be? Will Bayesian optimization become a valuable tool used internally as part of every AI company's infrastructure? Or, will it become possible to train transformers to do the same tasks Bayesian optimization does, thereby rendering our methods obsolete? I will not answer these questions: instead, I will introduce concepts that allow them to be asked in a precise form. To do so, I will briefly illustrate recent theory which describes what transformer models can in principle do, before moving to an open-ended discussion and question-answer period focused on showcasing different views from within the community.

4.13 The Gittins Index: A Design Principle for Decision-Making Under Uncertainty

Alexander Terenin (Cornell University – Ithaca, US)

License © Creative Commons BY 4.0 International license
© Alexander Terenin

Joint work of Ziv Scully, Alexander Terenin

Main reference Ziv Scully, Alexander Terenin: "The Gittins Index: A Design Principle for Decision Making Under Uncertainty", pp. 28–70, .

URL <https://doi.org/10.1287/educ.2025.0290>

The Gittins index is a tool that optimally solves a variety of decision-making problems involving uncertainty, including multi-armed bandit problems, minimizing mean latency in queues, and search problems like the Pandora's box model. However, despite the above examples and later extensions thereof, the space of problems that the Gittins index can solve perfectly optimally is limited, and its definition is rather subtle compared to those of other multi-armed bandit algorithms. As a result, the Gittins index is often regarded as being primarily a concept of theoretical importance, rather than a practical tool for solving decision-making problems. The aim of this tutorial is to demonstrate that the Gittins index can be fruitfully applied to practical problems. We start by giving an example-driven introduction to the Gittins index, then walk through several examples of problems it solves

– some optimally, some suboptimally but still with excellent performance. Two practical highlights in the latter category are applying the Gittins index to Bayesian optimization, and applying the Gittins index to minimizing tail latency in queues.

5 Working Groups

5.1 Acquisition Strategies

Thomas Bartz-Beielstein (TH Köln, DE)

Jack Buckingham (University of Warwick – Coventry, GB)

Carola Doerr (Sorbonne University – Paris, FR)

Matthias Feurer (TU Dortmund, DE)

Roman Garnett (Washington University – St. Louis, US)

Leonard Papenmeier (Universität Münster, DE)

Alexander Terenin (Cornell University – Ithaca, US)

License © Creative Commons BY 4.0 International license

© Thomas Bartz-Beielstein, Jack Buckingham, Carola Doerr, Matthias Feurer, Roman Garnett, Leonard Papenmeier, and Alexander Terenin

5.1.1 Introduction

A fundamental question in Bayesian optimization research is how to acquire the next data point. The key challenge comes from the presence of *explore-exploit tradeoffs*: should the algorithm pick a location about which little information is available, in case it might perform better, or a location that one can already expect to perform well? Different acquisition functions resolve these and related tradeoffs in different ways, and the underlying design principles describing these differences are only beginning to be understood.

As a simple example, consider Bayesian optimization under a budget constraint. How should the types of points acquired depend on the remaining budget? In an informal poll of the working group, participants overwhelmingly agreed that different kinds of points should be preferred under large vs. small budgets. If the budget is large, it makes sense to prefer points with high uncertainty more, compared to if the budget is small: intuitively, this is because there is more opportunity left to benefit from taking a risk. In contrast, if the budget is small, a risky acquisition point is unlikely to pay off, so it may be better to stick with points that have a better mean. The question becomes: *how should this intuition be reflected concretely via acquisition function design?*

The default expected improvement acquisition function, in its simplest incarnation, is hyper-parameter-free and therefore cannot be adjusted depending on the budget. On the other hand, UCB can, through its confidence parameter β_T . The exploratory behavior of both probability of improvement and expected improvement can also be tuned by introducing a hyperparameter controlling the improvement target; in the case of probability of improvement, the point maximizing the PI acquisition function for a given target also maximizes the UCB acquisition function for a corresponding confidence parameter and vice versa [12]. Some authors have also considered modifying the expected improvement acquisition strategy to account for the remaining budget via relatively simple, yet effective mechanisms [18]. Recent cost-aware acquisition functions, such as those based on Gittins index theory [25], can be adapted to the budget-constrained setting using Lagrangian relaxation: rather than working with a budget constraint, one can introduce Lagrange multipliers and add a term involving the sum of the costs to the usual simple regret objective, in some cases with appropriate optimality guarantees [30].

The need for different acquisition strategies is also reflected in other differences, beyond behavior with respect to budgets. For example, in high-dimensional Bayesian optimization, it is thought that finding the true global optimum may often be impossible in practice, and therefore be a poor goal for the algorithm to try to achieve. Instead, the algorithm should seek a *local optimum*. The question becomes: in situations with many local optima, how should the algorithm balance the tradeoff between the quality of a potential optimum and its difficulty to find? Comprehensively understanding tradeoffs like these – both from a comparative perspective focused on the behavior, popularity, and performance of existing methods, and from a fundamental theoretical perspective – was a recurring theme among the working group.

In the following, we discuss two main topics that were thoroughly explored throughout the seminar: the initial design and the optimization of the acquisition function. However, there are also other topics in the realm of acquisition strategies that we discussed in less breadth, but that still deserve further discussion, which is why we only mention them briefly. First, the topic can be extended to batch and asynchronous Bayesian optimization. However, recent work suggests that this problem might be easier than initially assumed [24]. Second, instead of using fully global Bayesian optimization, local Bayesian optimization methods could be more beneficial [27]. Third, as an intermediate strategy, it is possible to build kernels that balance the global fit of the function with a local fit around the optimum, potentially alleviating the issue that the global and local lengthscales differ [17].

5.1.2 Initial design

Bayesian optimization algorithms and other surrogate-based search techniques are often initialized with a set of non-adaptively queried points, the so-called *initial design* or *design of experiments (DoE)*. These points are often sampled uniformly at random, following some Latin Hypercube construction, or using quasi-random constructions such as subsets of Sobol’ sequences or similar. Some state-of-the-art surrogate-based optimization algorithms, such as SMAC [10], also use non-adaptive throughout the optimization process, to enforce exploratory behavior. However, later versions of SMAC, which were configured for hyperparameter optimization and blackbox optimization, reduced the fraction of random samples from 50% to 20% and 8%, respectively [16, 26]. We discussed the role of such non-adaptive sampling for Bayesian optimization and whether alternative sampling techniques that take into account the location but not the quality of already sampled points could improve the search behavior. In particular, we discussed the impact of replacing non-adaptive sampling by strategies that (i) maximize some form of “uniformity of distribution” of the collection of already evaluated and newly suggested points or (ii) maximize some minimal distance criterion.

As an example for (ii), we can mention the Morris-Mitchell criterion, a widely used metric for evaluating the space-filling properties of a sampling plan (also known as a design of experiments) [19]. It combines the concept of maximizing the minimum distance between points with minimizing the number of pairs of points separated by that distance. The standard Morris-Mitchell criterion, Φ_q , is defined as:

$$\Phi_q = \left(\sum_{i=1}^m J_i d_i^{-q} \right)^{1/q},$$

where $d_1 < d_2 < \dots < d_m$ are the distinct distances between all pairs of points in the sampling plan. J_i is the number of pairs of points separated by the distance d_i , and q is a positive integer (e.g., $q = 2, 5, 10, \dots$). As $q \rightarrow \infty$, minimizing Φ_q is equivalent to maximizing

the minimum distance d_1 , and then minimizing the number of pairs J_1 at that distance, and so on. One limitation of the standard Φ_q is that its value depends on the number of points in the sampling plan. This makes it difficult to compare the quality of sampling plans with different sample sizes. To address this, [3] proposed a size-invariant version of the criterion, Φ_q^* . Multi-objective optimization can also use Φ_q^* as an additional criterion [3].

Initial investigations such as Bossek et al. [5] suggest a correlation between the so-called L_∞ star discrepancy of the initial design and the quality of the best found solution. However, the size and structure of the best initial design depend to a large extent on the problem's structure, and we have only a very weak understanding of which problem characteristics are most decisive for the most suitable design.

An alternative approach identifies the role of the initial design in learning good model hyperparameters and observes that typical space-filling designs may not be optimal for this purpose. For example, it will not be possible to learn short length scales if all points in the initial design are far apart, and a traditional algorithm will have to discover this “by accident” during the sequential phase. This is not to suggest that the space-filling properties are to be completely discarded, though, since in the presence of large model-mismatch between the black-box function and its GP surrogate, points from the initial design can provide good local corrections to what might otherwise be an overconfident model. One work in this area is the development of a hybrid Latin Hypercube design for which the pairwise distances between points approximately follow a beta distribution [31]. Another recent proposal [11] uses a specific linear combination of two acquisition functions from active learning: BALD [9, 14], which maximizes mutual information between the proposed sample and the GP hyperparameters, and EPIG [4], which minimizes the conditional entropy of the GP given the proposed sample in expectation over the input space.

5.1.3 Optimizing the acquisition function

Most Bayesian optimization strategies require solving an inner optimization problem – the optimization of an *acquisition function* α :

$$\mathbf{x}_{\text{next}} \in \arg \max_{\mathbf{x} \in \mathcal{X}} \alpha(\mathbf{x})$$

Although most acquisition functions have a known form and are amenable to gradient-based optimizers such as L-BFGS-B, the acquisition landscape often becomes highly multimodal as the number of observations grows. While evaluating α on a discrete set of points is effective in lower dimensions, the acquisition function optimization becomes non-trivial in higher dimensions and requires, at a minimum, a multi-start strategy to address the multimodality of the function surface [29]. Another issue that can already arise for problems of moderate dimensionality is vanishing gradients of the acquisition function. Numerically more stable variants of popular acquisition functions help reduce the severity of this issue [1]. Still, for sufficiently high-dimensional problems, even these numerically more stable acquisition functions fall short in providing sufficient gradient signal, and vanishing gradients remain a challenge when maximizing α [21].

To alleviate the difficulties arising in high-dimensional Bayesian optimization, several works focus the search on regions around promising points [23, 6, 28, 22]. Since most acquisition functions are not flat around previously observed points, focusing the search to regions around previously well-performing points reduces the risk of vanishing gradients and is likely to yield points of high acquisition value [1]. The **TuRBO** algorithm [6] implements such a strategy for Thompson sampling (TS). Interestingly, TS does not suffer from vanishing

gradients as it simply searches for the maximum or minimum of a Gaussian process (GP) sample path. However, TS is overly explorative in high dimensions [20], a behavior [6] counterbalance by enforcing locality: Starting from the incumbent observation, **TuRBO** creates a dense grid of candidate points obtained from a GP sample path. To enforce locality, **TuRBO** keeps the values of the incumbent point for $1 - \min(1, \frac{20}{d})$ randomly chosen dimensions, essentially only perturbing a subset of the dimensions for problems with more than 20 dimensions. Here, d is the dimensionality of the search space $\mathcal{X}$. By perturbing only a subset of the dimensions, candidate points remain close to the incumbent observation. Following up on that idea, [22] drop the requirement of the axis-alignment of the perturbation and propose a perturbation strategy that ensures candidate points remain within a certain radius of incumbent points.

The **spotoptim** package [2] implements a multi-objective infill strategy that uses desirability [15] to determine the next point [3]. Multiple criteria can be combined; for example, space-fillingness can be computed using the Morris-Mitchell criterion [19]. However, preliminary results indicate that space-fillingness should be used with caution, as it may push infill points too far from promising regions that have already been evaluated.

The aforementioned strategies focus on subregions of the space, mostly sacrificing global optimization unless a high number of function observations can be afforded. However, for a multi-start acquisition function optimization approach, one may use these strategies to augment the set of *starting points* with local candidates [21]. Combining global and local samples is a seemingly safe strategy, as the starting points are chosen among all samples. In fact, **BoTorch** implements such a strategy with the `sample_around_best` option, but does not enable it by default. Furthermore, **BoTorch** does not even choose the points of maximum initial acquisition value to start the gradient-based optimization, but performs Boltzmann sampling of starting points according to their initial acquisition value to ensure asymptotically space-filling behavior [1]. Arguably, the motivation for this design choice is that *strictly maximizing the acquisition function may not be the best idea*, an idea that has also been stated explicitly within the community [7]. Although, to the best of our knowledge, no works explicitly outline the conditions for this to be fulfilled, various works follow inexact acquisition function maximization schemes – often motivated by reduced computational effort. Delaunay triangulation and Voronoi tessellation leverage empirical observations of the behavior of the acquisition function to select points of high expected acquisition value [8, 8]. Additionally, [13] argue that random grid search may be sufficient for acquisition function optimization.

Despite the numerous strategies to optimize the acquisition function, no clear best strategy has been identified yet, and further research is required.

5.1.4 Outlook

Despite widespread use of Bayesian optimization, the underlying design principles for acquisition strategies are only beginning to be understood, particularly regarding how they should adapt to problem structure, dimensionality, and resource constraints. This is both from an empirical and a theoretical perspective, following other, more basic advances in Bayesian optimization, which have made tackling harder problem classes with structure possible.

References

- 1 Sebastian Ament, Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Unexpected improvements to expected improvement for Bayesian optimization. *Advances in Neural Information Processing Systems*, 36:20577–20612, 2023.

- 2 Thomas Bartz-Beielstein. Sequential parameter optimization cookbook.
- 3 Thomas Bartz-Beielstein. Multi-objective optimization and hyperparameter tuning with desirability functions, 2025.
- 4 Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented Bayesian active learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 7331–7348. PMLR.
- 5 Jakob Bossek, Carola Doerr, and Pascal Kerschke. Initial design strategies and their effects on sequential model-based optimization: an exploratory case study based on BOB. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference, GECCO '20*, page 778–786, New York, NY, USA, 2020. Association for Computing Machinery.
- 6 David Eriksson, Michael Pearce, Jacob Gardner, Ryan D Turner, and Matthias Poloczek. Scalable global optimization via local Bayesian optimization. *Advances in neural information processing systems*, 32, 2019.
- 7 Roman Garnett. *Bayesian optimization*. Cambridge University Press, 2023.
- 8 Robert B Gramacy, Annie Sauer, and Nathan Wycoff. Triangulation candidates for Bayesian optimization. *Advances in Neural Information Processing Systems*, 35:35933–35945, 2022.
- 9 Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning.
- 10 Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In Carlos A. Coello Coello, editor, *Learning and Intelligent Optimization*, pages 507–523, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.
- 11 Carl Hvarfner, David Eriksson, Eytan Bakshy, and Max Balandat. Informed initialization for Bayesian optimization and active learning. In *Advances in Neural Information Processing Systems*.
- 12 Donald R. Jones. A Taxonomy of Global Optimization Methods Based on Response Surfaces. *Journal of Global Optimization*, 21(4):345–383, 2001.
- 13 Hwanwoo Kim, Chong Liu, and Yuxin Chen. Bayesian optimization with inexact acquisition: Is random grid search sufficient? *arXiv preprint arXiv:2506.11831*, 2025.
- 14 Andreas Kirsch, prefix=van useprefix=true family=Amersfoort, given=Joost, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- 15 Max Kuhn. desirability: Function optimization and ranking via desirability functions., 9 2016.
- 16 Marius Lindauer, Matthias Feurer, Katharina Eggensperger, André Biedenkapp, and Frank Hutter. Towards assessing the impact of bayesian optimization’s own hyperparameters, 2019.
- 17 Ruben Martinez-Cantin. Local nonstationarity for efficient bayesian optimization, 2015.
- 18 Jonas Mockus. *Bayesian Approach to Global Optimization: Theory and Applications*. Mathematics and Its Applications. Kluwer Academic Publishers, 1989.
- 19 Max D. Morris and Toby J. Mitchell. Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43(3):381–402, 1995.
- 20 Leonard Papenmeier, Nuojin Cheng, Stephen Becker, and Luigi Nardi. Exploring exploration in Bayesian optimization. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2024.
- 21 Leonard Papenmeier, Matthias Poloczek, and Luigi Nardi. Understanding high-dimensional Bayesian optimization. In *Forty-second International Conference on Machine Learning*, 2024.

- 22 Bahador Rashidi, Kerrick Johnstonbaugh, and Chao Gao. Cylindrical Thompson sampling for high-dimensional Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3502–3510. PMLR, 2024.
- 23 Rommel G Regis and Christine A Shoemaker. Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization. *Engineering Optimization*, 45(5):529–555, 2013.
- 24 Ben Riegler, James A C Odgers, and Vincent Fortuin. Rethinking exploration in asynchronous bayesian optimization: Standard acquisition is all you need. In *The Exploration in AI Today Workshop at ICML 2025*, 2025.
- 25 Ziv Scully and Alexander Terenin. The Gittins index: A design principle for decision making under uncertainty. In *Tutorials in Operations Research: Advances in Analytics and Operations Research: Improving Decisions to Secure the Future*, pages 28–70. INFORMS, 2025.
- 26 Kim Peter Wabersich and Marc Toussaint. Advancing bayesian optimization: The mixed-global-local (mgl) kernel and length-scale cool down, 2016.
- 27 Kaiwen Wu, Kyurae Kim, Roman Garnett, and Jacob Gardner. The behavior and convergence of local bayesian optimization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 73497–73523. Curran Associates, Inc., 2023.
- 28 Kaiwen Wu, Kyurae Kim, Roman Garnett, and Jacob Gardner. The behavior and convergence of local Bayesian optimization. *Advances in neural information processing systems*, 36:73497–73523, 2023.
- 29 Nathan Wycoff, John W. Smith, Annie S. Booth, and Robert B. Gramacy. Voronoi candidates for Bayesian optimization. *CoRR*, abs/2402.04922, 2024.
- 30 Qian Xie, Raul Astudillo, Peter Frazier, Ziv Scully, and Alexander Terenin. Cost-aware Bayesian optimization via the Pandora’s box Gittins index. *Advances in Neural Information Processing Systems*, 37:115523–115562, 2024.
- 31 Boya Zhang, D. Austin Cole, and Robert B. Gramacy. Distance-distributed design for Gaussian process surrogates. 63(1):40–52.

5.2 Multifidelity

Mauricio Alvarez (University of Manchester, GB)

Cédric Archambeau (Helsing – Berlin, DE)

Maximilian Balandat (Meta – Menlo Park, US)

Ivo Couckuyt (Ghent University, BE)

Frank Hutter (Prior Labs – Freiburg, DE & ELLIS Institute Tübingen, DE & Universität Freiburg, DE)

Aaron Klein (ScaDS.AI – Leipzig, DE)

Neeratyoy Mallik (Universität Freiburg, DE)

Ruth Misener (Imperial College London, GB)

Juan Ungredda (ESTECO SpA – Trieste, IT)

License  Creative Commons BY 4.0 International license

© Mauricio Alvarez, Maximilian Balandat, Cédric Archambeau, Ivo Couckuyt, Frank Hutter, Aaron Klein, Neeratyoy Mallik, Ruth Misener, and Juan Ungredda

5.2.1 Introduction

Many scientific, engineering, and machine learning problems require making critical decisions under strict resource constraints. In such settings, directly evaluating the performance metric of interest – through large-scale model training, high-resolution simulations, or physical experiments – is often prohibitively expensive. As a result, practitioners rely on a collection of cheaper approximations, such as simplified models, coarse simulations, or proxy experiments, which differ in accuracy, uncertainty, and cost. These approximations are commonly referred to as fidelities.

More recently, a growing body of work has demonstrated that performance in complex systems often follows regular, predictable patterns as a function of key resource-controlling variables, such as model size, data size, resolution, or experimental scale. These relationships, commonly referred to as scaling laws, provide models that extrapolate performance beyond observed regimes and enable informed decision making about how to allocate resources. While scaling laws have been most prominently studied in the context of large language model pretraining, similar ideas arise naturally in other domains like computational fluid dynamics.

Multi-fidelity methods exploit the availability of multiple information sources with varying querying costs, while scaling laws provide structured models that describe how performance evolves as fidelity or resource levels change. When combined within a Bayesian framework, scaling laws can serve as informative priors or mean functions, while multi-fidelity surrogate models capture deviations, uncertainty, and correlations across fidelities. Crucially, this combination enables the design of more efficient acquisition functions that explicitly trade off cost, uncertainty reduction, and extrapolation to high-fidelity regimes. Arguably, scaling laws are indispensable when high fidelity evaluations are extremely time-consuming and/or costly, and available computing budget can typically support 1 high fidelity evaluation (up to 5 evaluations).

The central goal of this working group is to explore scaling-laws and multi-fidelity Bayesian optimization to predict performance and guide data collection. This includes developing probabilistic models for scaling behavior, quantifying uncertainty in extrapolation, and designing acquisition strategies that actively select which designs, fidelities, and resource levels to evaluate next. Such strategies move beyond naive grid searches or uniform sampling, enabling principled exploration of large design spaces under tight budgets.

We begin by formalizing a general multi-fidelity optimization problem in which design variables and fidelity levels are treated as joint decision variables under a global budget constraint. We then examine several application domains, including scaling laws for large language model pretraining, computational fluid dynamics simulations, and chemical engineering. These case studies illustrate how scaling laws and multi-fidelity structure manifest differently across domains, yet can be unified under a common Bayesian optimization perspective.

5.2.2 Problem Statement

Let $\mathcal{X}$ denote the *design space*, which may be continuous ($\mathcal{X} \subseteq \mathbb{R}^d$) or discrete/finite. Let $\mathcal{W}$ denote the *fidelity space*, which may likewise be continuous ($\mathcal{W} \subseteq \mathbb{R}^p$) or discrete (e.g., $\{1, \dots, M\}$). We consider an unknown response function

$$h : \mathcal{X} \times \mathcal{W} \rightarrow \mathbb{R}.$$

where an evaluation at a design–fidelity pair $(x, w) \in \mathcal{X} \times \mathcal{W}$ produces an observation

$$y = h(x, w),$$

In some settings, such as deterministic simulations, evaluations may be effectively noise-free. In others, including stochastic simulations or physical experiments, it is common to assume additive noise, $y = h(x, w) + \varepsilon_w$, where ε_w is a random noise term whose distribution and variance may depend on the fidelity level w .

Given a *target fidelity* $w^* \in \mathcal{W}$ is specified – corresponding to the most accurate and most expensive evaluation – we aim to identify a design that maximizes the performance,

$$x^* \in \arg \max_{x \in \mathcal{X}} h(x, w^*).$$

We assume access to a total budget $B > 0$, representing a limited resource such as time, monetary cost, computational effort, or experimental capacity. This budget may be allocated sequentially over rounds $t = 1, 2, \dots, T$. At each round, the algorithm selects a design–fidelity pair (x_t, w_t) , observes y_t , and incurs a cost $c(x_t, w_t)$, subject to the constraint

$$\sum_{t=1}^T c(x_t, w_t) \leq B.$$

After consuming the budget, the algorithm returns a recommended design x_r , intended to achieve high performance at the target fidelity, i.e., a large value of $h(x_r, w^*)$.

5.2.3 Application: Scaling Laws for LLM pretraining

Scaling laws are used widely in pretraining generative model, such as LLM, to make high-stakes decisions, for example, which model size, which data size, which hyperparameters, etc.. Current practice is to run large grid searches across model and data scales as well as hyperparameters for each of these, and then fit a parametric law. This is likely highly inefficient since large amounts of the data collected may not be very relevant to improving the fit of the parametric law. Furthermore, the scaling law approaches in the literature (and in practice) typically do not take into account uncertainty. We believe that there is a large opportunity to improve the estimation of scaling laws by using Bayesian techniques (modeling as well as optimization / active learning).

There are multiple aspects to this problem, the primary ones are surrogate modeling for the scaling behavior, and the acquisition strategy for adaptively choosing which models to train on which data under which hyperparameter settings. Both modeling and acquisition strategies will depend on the specific goal we are trying to achieve downstream. In the following, we consider a particular, popular scaling law and propose a GP-based modeling approach for it. We plan to evaluate this proposed model on publicly available data sets.

5.2.3.1 An Example Scaling Law

The well-known Chinchilla Scaling Law [1] models the validation loss of a transformer model by $\mathcal{L}(D, N) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$ where N specifies the number of parameter and D the number of token. To estimate the parameters E, A, B, α, β , it is common to use a predefined grid of different N, D with optimal learning rates LR and batch sizes BS . To estimate the optimal hyperparameters (incl. LR and BS), they need to evaluate another grid search for each N, D pair. This is potentially quite wasteful, and we suggest to instead employ Bayesian modeling together with an active learning strategy to explore the joint space of model size, data size, and hyperparameters.

Specifically, we aim to model the validation loss $\mathcal{L}(\mathbf{x})$ with $\mathbf{x} = (D, N, LR, BS)$ to also include hyperparameters LR and BS in addition to token (D) and parameter (N) count. This has two key advantages:

1. We hypothesize that leveraging additional data points (all parameter settings rather than just the optimal one) improves the predictive quality of the model.
2. The ability to predict the loss not only as a function of the number of tokens D and parameters N but also for the hyperparameters will enable us to predict the optimal hyperparameter settings for large D and N – a key decision one needs to make for training the full model, and often shifts with increasing D and N .

Using a probabilistic model of this behavior will enable uncertainty quantification, which generally is very useful for decision making (esp. given the costs of the training jobs involved) and, also be leveraged in a principled active learning approach to efficiently estimate the parameters of the scaling law.

We assume:

$$\mathcal{L}(\mathbf{x})|f(\mathbf{x}), \theta \sim \mathcal{N}(f(\mathbf{x}), \sigma^2) \quad (1)$$

$$f(\mathbf{x})|\theta \sim \mathcal{GP}(m_{\theta_m}(\mathbf{x}), k_{\theta_k}(\mathbf{x}, \cdot)) \quad (2)$$

where:

$$m_{\theta_m}(\mathbf{x}) = m_{\theta_m}(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta} \quad (3)$$

is the prior mean function with parameters $\theta_m = (E, A, B, \alpha, \beta)$ and k_{θ_k} is the kernel function with θ_k the kernel hyperparameters (incl. lengthscales, etc.).

Given a data set $\mathcal{X}$, we can now fit the parameters $\theta = (\theta_m, \theta_k)$ using maximum likelihood optimization (we could also use other inference approaches such as fully Bayesian / MCMC).

5.2.3.2 Selective selection of hyperparameter configurations

We use Bayesian active learning by disagreement (BALD) [2], which is equivalent to the expected information gain, to sample the hyperparameter configurations to evaluate:

$$U_{\text{BALD}}(\mathbf{x}) = \mathbb{H}[\mathcal{L}(\mathbf{x})|\mathcal{D}] - \mathbb{E}_{p(\theta|\mathcal{D})} \mathbb{H}[\mathcal{L}(\mathbf{x})|\theta] \quad (4)$$

$$= \mathbb{H}[\mathbb{E}_{p(\theta|\mathcal{D})} \mathcal{L}(\mathbf{x})|\theta, \mathcal{D}] - \mathbb{E}_{p(\theta|\mathcal{D})} \mathbb{H}[\mathcal{L}(\mathbf{x})|\theta] \quad (5)$$

where $\mathcal{D} = \{(\mathbf{x}_i, \mathcal{L}_i)\}_{i=1}^n$. Note that $p(\mathcal{L}(\mathbf{x})|\theta, \mathcal{D})$ is the predictive posterior Gaussian process and $p(\mathcal{L}(\mathbf{x})|\theta)$ is the marginal likelihood, both of which are analytically tractable (and have a Gaussian form).

A somewhat simpler approach would be to adopt a Bayesian query by committee approach.

$$U_{\text{BQC}}(\mathbf{x}) = \mathbb{V}_{p(\theta|\mathcal{D})} [\mu_\theta(\mathbf{x}) - \bar{\mu}_\theta(\mathbf{x})|\theta] = \mathbb{E}_{p(\theta|\mathcal{D})} \left[(\mu_\theta(\mathbf{x}) - \bar{\mu}_\theta(\mathbf{x}))^2 | \theta \right] \quad (6)$$

where $\mu_\theta(\mathbf{x})$ is the predictive posterior Gaussian process mean and $\bar{\mu}_\theta(\mathbf{x}) = \mathbb{E}_{p(\theta|\mathcal{D})} [\mu_\theta(\mathbf{x})]$.

In addition, we want to penalize cost. We adopt the approximate cost $c(N, D) = 6ND$ [3], such that

$$U_c(\mathbf{x}) = \frac{U(\mathbf{x})}{c(\mathbf{x})}. \quad (7)$$

How to approximate the marginal posterior $p(\theta|\mathcal{D})$? The challenge with the active selection of the hyperparameter configurations is that the posterior $p(\theta|\mathcal{D})$ is analytically intractable. One strategy is to approximate it via variational inference: $q(\theta) \approx p(\theta|\mathcal{D})$. The expectations appearing in the acquisition functions can then be computed by Monte Carlo integration by sampling from $q(\theta)$.

5.2.3.3 Data

We identified the following published data sources as potential candidates for evaluating our modeling approach:

1. Porian et al. [4] provides the validation loss for various pre-trained language models across an increasing N and D trained with a fixed grid of learning rates and batch sizes.
2. Nezhurina et al. [5] provides data to derive scaling laws for language-vision models.

In addition, ongoing work by some of the authors on training dense transformers on language data using μP parameterization will provide additional data points.

A preliminary investigation of the data sets can be found in the following Colab notebook: <https://colab.research.google.com/drive/1XPEaJ3az8DhwEa6grkKYyOpUjfi0XjqX>

Initial observations from literature and experiments on large language/vision models show consistent scaling behavior with model size, data size, and compute costs. The novelty and challenge of this problem stem from the need for strict extrapolation: at larger scales, both optimal hyperparameters and loss values fall outside the ranges observed in training data. Additionally, the parameter space is sparsely sampled across all dimensions – data (D), model size (N), and hyperparameters (LR, BS, etc.) – due to high computational costs and non-systematic exploration strategies used in data collection.

5.2.4 Applications

5.2.4.1 LLMs

Scaling laws are an important tool for (pre-)training LLMs, as they allow performance to be extrapolated and enable informed decisions about the hyperparameters of the final model. However, accumulating the data needed to fit these scaling laws is a manual, multi-step process that requires human expertise.

The approach presented here would automate this process and reduce the overall computational cost. This would be particularly appealing to smaller or academic labs that do not have access to large amounts of computational resources, as it would enable them to more thoroughly compare different architectures or datasets and ensure sensible scaling.

5.2.4.2 CFD simulations

Computational Fluid Dynamics (CFD) is a domain in which scaling-law multifidelity modeling may be applicable. In CFD applications, the fidelity can often be controlled continuously through physically meaningful parameters such as mesh resolution, time-step size, or solver accuracy. As these fidelity controls are refined, such as refining the mesh or tightening solver tolerances, the simulation fidelity tends to produce more reliable information, but at an increasing computational cost. Because this accuracy–cost trade-off follows regular patterns, coarse simulations can often be used to guide how high-fidelity simulations should be performed. In this setting, scaling laws (whether derived analytically, empirically, or learned probabilistically) can play a central role in guiding multifidelity Bayesian optimization by predicting how the solution error and uncertainty decay with increasing fidelity.

In engineering (e.g., CFD), scaling laws can be constructed for multi-fidelity objective functions (i.e., end-to-end or black-box approaches, or, alternatively, considering multi-fidelity mesh sizes (indirectly defining the objective function, i.e., a grey-box approach).

We propose to study these relationships using the MegaFlow2D dataset and reactor benchmark¹. This dataset provides CFD simulations at multiple resolutions but does not include a specific downstream objective. It is well-suited for exploring super-resolution and multifidelity modeling approaches.

5.2.4.3 Chemistry & Chemical Engineering

Scale-up is a consistent challenge in industrial chemistry and chemical engineering: we want to move from (i) cheap computer simulations to (ii) expensive computer simulations to (iii) cheap experiments on a lab bench to (iv) more expensive experiments on a lab bench to (v) experiments in a pilot plant to (vi) experiments at scale. At each of these levels, we have data on all previous levels and wish to move to the next level. The number of experiments at the coarser level(s) will be large compared to the number of experiments at the finer levels.

Unless there are computer simulations in the chemical engineering examples that obey known and exploitable scaling laws, scaling-law approaches are less applicable to this class of applications. Instead, fidelity levels are best viewed as categorical experimental or computational modalities, such as laboratory-scale experiments versus pilot-scale studies.

Therefore, while scaling laws are less relevant to the chemistry and chemical engineering examples, there are additional considerations that need to be incorporated:

- The input space is highly constrained because not every molecule is synthesizable,
- Each of the fidelities in the output space are subject to uncertainty, these are typically notated in chemistry experiments as the different results for different replicas.

The three examples we have already integrated into our Google Colab document consist on:

- Battery design [6] represents experiments aimed at optimizing electrode materials for lithium-ion batteries by leveraging measurements obtained at different experimental fidelities. The variables dop1–dop4 correspond to controllable electrode design parameters, that represent dopant concentrations in high-nickel cathode materials, which define a specific candidate electrode formulation. Each experiment is conducted at a fidelity level indicated by W, where lower fidelities represent cheaper and faster proxy experiments, most notably coin-cell tests or shortened testing protocols, while higher fidelities correspond to

¹ https://github.com/trsav/reactor_benchmark – The reactor benchmark is provided by Imperial college. Currently, contacts have been established to investigate possible collaboration.

more expensive, time-consuming, and industrially relevant experiments, such as pouch-cell testing that better reflects real battery performance. The output variable Y measures the specific capacity (electrochemically accessible charge per unit mass) at the chosen fidelity and is the variable to be maximized.

- Covalent organic frameworks [7, 8] represents a materials discovery problem over a fixed, finite set of 609 experimentally reported covalent organic frameworks (COFs), where each candidate material is represented by a 14-dimensional continuous feature vector derived from its crystal structure and composition. These 14 features encode physically meaningful properties of each COF that are known to influence gas adsorption behavior. The variable W denotes the simulation fidelity level used to evaluate a given COF: a low-fidelity Monte Carlo integration to compute Xe and Kr Henry coefficients under the dilute approximation (15 min runtime), or a high-fidelity Markov-chain Monte Carlo simulation of the binary grand-canonical ensemble that explicitly models competitive Xe/Kr adsorption (230 min runtime). The response variable Y is the simulated xenon/krypton adsorptive selectivity at room temperature, which quantifies how preferentially a COF adsorbs Xe over Kr.
- Molecule polarizability [8, 9] consists of 1,134 organic molecules, each represented by a vector of molecular descriptors that capture structural and chemical features such as atom types, bonding patterns, and functional groups. Following the authors’ methodology, these high-dimensional descriptors are reduced to a 10-dimensional continuous representation using principal component analysis. The variable W denotes the fidelity level of the polarizability measurement, where the low-fidelity data corresponds to polarizability computed at the Hartree–Fock 6-31G+ level of theory, which is relatively inexpensive, while the high-fidelity data corresponds to experimentally measured polarizability. The response variable Y is the molecular polarizability itself. The dataset is structured to support multi-fidelity Bayesian optimization, with the objective of identifying molecules that maximize experimental polarizability.

References

- 1 Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models. *arXiv e-prints*, page arXiv:2203.15556, March 2022.
- 2 D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 1996.
- 3 J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv:2001.08361 [cs.LG]*, 2020.
- 4 Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving Discrepancies in Compute-Optimal Scaling of Language Models. *arXiv e-prints*, page arXiv:2406.19146, 2024.
- 5 Marianna Nezhurina, Tomer Porian, Giovanni Pucceti, Tommie Keressies, Romain Beaumont, Mehdi Cherti, and Jenia Jitsev. Scaling Laws for Robust Comparison of Open Foundation Language-Vision Models and Datasets. *arXiv e-prints*, page arXiv:2506.04598, June 2025.
- 6 Jose Pablo Folch, Robert M. Lee, Behrang Shafei, David Walz, Calvin Tsay, Mark van der Wilk, and Ruth Misener. Combining multi-fidelity modelling and asynchronous batch bayesian optimization. *Computers & Chemical Engineering*, 172:108194, 2023.

- 7 Nickolas Gantzler, Aryan Deshwal, Janardhan Rao Doppa, and Cory M. Simon. Multi-fidelity bayesian optimization of covalent organic frameworks for xenon/krypton separations. *Digital Discovery*, 2:1937–1956, 2023.
- 8 Víctor Sabanza-Gil, Riccardo Barbano, Daniel Pacheco Gutiérrez, Jeremy S Luterbacher, José Miguel Hernández-Lobato, Philippe Schwaller, and Loïc Roch. Best practices for multi-fidelity bayesian optimization in materials and molecular research. *Nature Computational Science*, pages 1–10, 2025.
- 9 Clyde Fare, Peter Fenner, Matthew Benatan, Alessandro Varsi, and Edward O Pyzer-Knapp. A multi-fidelity machine learning approach to high throughput materials screening. *npj Computational Materials*, 8(1):257, 2022.

5.3 Human-in-the-loop Bayesian Optimization: Explainability and Interactivity

Jürgen Branke (University of Warwick, GB)

Katharina Eggersperger (TU Dortmund, DE)

Jonathan Fieldsend (University of Exeter, GB)

Daniel Hernández-Lobato (Autonomous University of Madrid, ES)

Mike McCourt (Distributional – Toronto, CA)

and Sebastian Rojas Gonzalez (Ghent University, BE)

License  Creative Commons BY 4.0 International license

© Jürgen Branke, Katharina Eggersperger, Jonathan Fieldsend, Daniel Hernández-Lobato, Mike McCourt, and Sebastian Rojas Gonzalez

5.3.1 Introduction

This is the report of the working group on Human-In-the-Loop (HITL) Bayesian Optimization which was part of the Dagstuhl Seminar on Bayesian Optimization. We started by collecting and categorizing different ways in which humans can interact with Bayesian Optimization (BO). These modes of interaction include: providing prior knowledge, by interacting with the algorithm during optimization, or after optimization, when picking or fine-tuning solutions or asking for explanations. Section 5.3.2 summarizes what we found. Later, we focused on two particular aspects: explainability, which is discussed in Section 5.3.3 and bounding future improvements or deciding when to stop – initial results of which are set out in Section 5.3.4.

5.3.2 Human-in-the-loop Optimization

In most optimization scenarios, the HITL provides a priori information about the problem in the form of a proper problem formulation, including objective function(s), decision variables and constraints. However, there can also be additional information that the HITL may provide, such as expert or domain knowledge about possible future scenarios, the location within the search space where good solutions are believed to be, or some already evaluated initial solutions, e.g., from historical observations.

Interactively, the HITL may propose promising solutions (i.e., taking the role of the acquisition function), add or remove constraints, modify the objective, modify the explore/exploit balance, or decide when they are satisfied with the solution and want to stop. For example, in [39], the HITL can provide advice on the next query point through binary accept/reject recommendations.

Preferential BO [15, 31] assumes that it is not possible to formulate a quantifiable objective function; instead, one must rely entirely on the HITL to guide the optimization. Pairwise BO interacts with the HITL by repeatedly asking the HITL which of two solutions is preferred. The goal of the acquisition function is then to pick the most informative pair of solutions to show to the HITL. Preferences can also be stated as rankings when the HITL is capable of simultaneously considering more than two solutions. A very interesting recent development in this area is the use of LLMs to interact with the HITL; this may allow the BO to elicit, in a very natural way, more complex preference information than just which of two solutions is better [24, 23].

After optimization, the HITL may ask for a better understanding of the solution, such as how far the solution might be from the optimum, how robust the solution is, which decision variables are the most influential, or under what circumstances another solution would be better. Section 5.3.3 will look more closely into this latter aspect.

5.3.2.1 Multiobjective Bayesian Optimization

A special case of HITL BO is the case of multi-objective optimization. In the case of multiple objectives, there is usually not a single solution best in all the objectives, but there are several Pareto-optimal solutions with different trade-offs. Therefore, in multi-objective BO, a HITL has to be involved at some point in order to identify the most preferred of the Pareto-optimal solutions. The definition of the *selected* “optimal” solution is subjective and depends on the goals of the HITL. Preference elicitation can be done *a priori*, i.e., before the optimization, interactively during the optimization, or *a posteriori*, after optimization by the HITL picking the most preferred solution from the set of Pareto-optimal solutions.

Progress in this area has evolved from multiple relatively independent research streams. Despite the shared goal of aiding decision-making for the HITL, these research directions have historically evolved in isolation, with contributions spanning multiple fields. For example:

- Analytical methods, including (multi-attribute) utility theory (e.g., [28, 34, 1]);
- Heuristic optimization approaches, especially evolutionary algorithms (e.g., [7, 8, 38]);
- and
- Bayesian optimization frameworks (e.g., [32, 15, 6, 2]).

Recent efforts to bridge the research through hybrid or integrative approaches are promising, but remain scarce and face several unresolved challenges:

Type of Preference Information. Human–algorithm interaction can yield diverse forms of preference data, including:

- weights or priority levels assigned to objectives,
- partial or complete rankings of candidate solutions,
- pairwise comparisons between alternatives,
- specification of aspiration levels or desired improvement directions (which may e.g. relate to the performance of existing designs),
- upper or lower bounds on objective values.

However, the amount and quality of preference information that can be reliably extracted from the Decision Maker (DM) are difficult, especially considering that the ultimate objective is to *select the most preferred solution from a relatively small “menu” of diverse options that are all high-performing*, independently of the characteristics of the problem (e.g., dimensionality, convexity, stochasticity, etc.). Excessive interaction or cognitive demand may induce biases, inconsistencies, and decision fatigue [3]. Moreover, the *informativeness* and *interaction cost* of different preference types may vary across optimization stages, yet

systematic guidance on this aspect is lacking. Most existing preference models are parametric, requiring parameter estimation from limited human feedback [9], which introduces additional sources of computational complexity (and uncertainty).

Frequency and Timing of Interaction. Determining when and how often to engage the HITL remains an open problem. Frequent interactions can lead to inconsistency or transitivity violations (e.g., preferring A over B, B over C, but C over A), whereas infrequent interactions may yield insufficient information and inefficient search behavior [29]. Adaptive interaction strategies that balance informativeness and cognitive load are therefore an important research frontier.

Imperfect and Noisy Feedback. Many existing studies assume that human-provided information is accurate and consistent. In reality, human responses are subject to uncertainty, hesitation, and contextual variability. Modeling and propagating this uncertainty within the optimization process – rather than neglecting it – is crucial for realistic HITL systems [5].

Heterogeneity of Decision Makers. Decision makers differ in their attitudes toward risk, prior knowledge, and cognitive strategies. For instance, risk-averse, risk-neutral, and risk-seeking individuals may exhibit fundamentally different preference structures [16]. Capturing such heterogeneity in a generalizable manner remains challenging, as fully characterizing individual decision models is infeasible without imposing strong assumptions.

Decisions under Uncertainty. The integration of preference learning with stochastic or uncertain optimization remains underexplored relative to deterministic formulations. While utility theory approaches have modeled exogenous uncertainty (arising from environmental variability) [16], Bayesian optimization methods often address endogenous uncertainty (stemming from model or observation noise) [6]. Developing frameworks that jointly account for both uncertainty sources represents a promising but demanding direction for future research [20, 35].

Performance Evaluation and Benchmarking. The evaluation of human-in-the-loop optimization algorithms remains a nascent research area. Traditional performance metrics – typically based on the quantitative assessment of the Pareto front in deterministic problems (e.g., regret or hypervolume) – are inadequate for interactive or stochastic contexts. Emerging studies propose benchmark problems with known solutions [3, 4], but standardized methodologies for evaluating interactive and preference-based optimization remain largely undeveloped.

5.3.3 Problem formulation: Explainability in HITL-BO

A critical component of a successful HITL system is **explainability**. A lack of transparency can cause the user to lose trust in the system. For example: the system solves an optimization problem and finds the best solution x^* based on the current model parameters, θ_0 :

$$x^* = \arg \max_x f(x, \theta_0). \quad (8)$$

The HITL is unsatisfied and asks: “*Why did you not select x' ?*”. Similar to counterfactual explanations in machine learning and following [13], we aim to explain the decision by answering the question: “*In what alternative context θ' would the solution expected by the HITL be optimal, or better than the solution found by BO under θ_0 ?*”.

The proposed solution is to find the smallest change to the preference parameters θ_0 that would make x' a “good” choice. We have explored several mathematical formulations for the problem in the context of BO. We aim to find a new set of parameters θ' that is minimally distant from the original parameters θ_0 , as measured by a distance function $d(\theta', \theta_0)$.

Dividing the decision variables into two sub-groups $x = (a, b)$ where $a \in \mathbb{R}^{m_1}, b \in \mathbb{R}^{m_2}$ and $x \in \mathbb{R}^{m_1+m_2}$, we distinguish the following two cases.

Case 1: x' is a complete alternative solution.

Case 2: The HITL only specifies a partial solution a' , essentially asking the question “*Why does the optimal solution not use a' ?*”.

5.3.3.1 Explaining why the HITL proposal is inferior

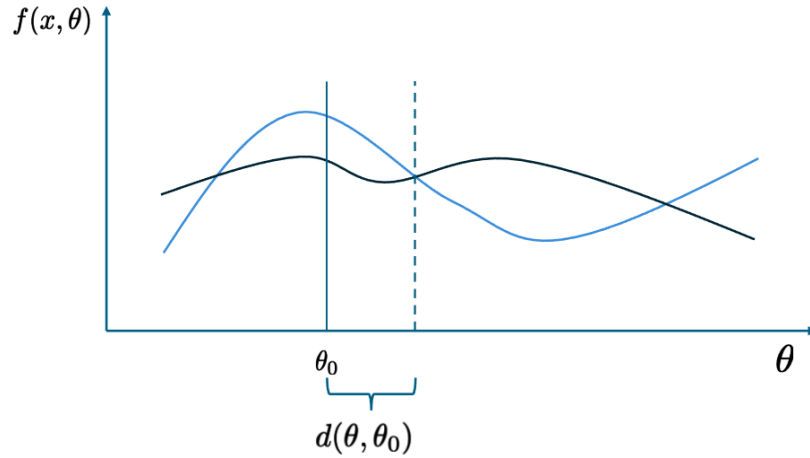
In Case 1, one way to convince the HITL that the identified optimal solution x^* is superior, is to simply compute $f(x', \theta_0)$ and show that $f(x', \theta_0) < f(x^*, \theta_0)$. If the HITL has only specified a partial solution a' as in Case 2, then one would have to demonstrate that even the best $x = (a', b)$ would be worse, i.e., $\max_b f(a', b, \theta_0) < f(x^*, \theta_0)$. This is straightforward to do with BO.

5.3.3.2 Finding a context where the HITL proposal is better

In Case 1, this approach finds the closest possible context (the θ' with the minimum distance $d(\theta', \theta_0)$) where the user’s choice x' would score higher than the system’s original choice x^* .

$$\theta' = \arg \min_{\theta} d(\theta, \theta_0) \text{ s.t. } f(x', \theta) \geq f(x^*, \theta). \quad (9)$$

Figure 1 illustrates the situation. At θ_0 , the function for x^* (blue line) is higher than for x' (black line). The algorithm finds the closest θ' where the blue line is higher than or equal to the black line.



■ **Figure 1** Blue line illustrates $f(x^*, \theta)$, black line illustrates $f(x', \theta)$.

In BO, this can be achieved as follows. First, the goal would be to find a solution where $f(x', \theta) > f(x^*, \theta)$ (if we cannot find one, we can inform the HITL that the proposed x' is worse under any possible context θ). This can simply be achieved by maximizing $f(x', \theta) - f(x^*, \theta)$ over θ , e.g., with expected improvement. Note that every sample for this problem requires the evaluation of two solutions, (x', θ) and (x^*, θ) . Using Knowledge Gradient or Entropy Search, this could also be tackled by allowing only one of the solutions x' or x^* to be evaluated in each iteration.

Once a θ has been found, the goal changes to finding the closest θ , i.e., we search for a reduction in distance subject to still ensuring $f(x', \theta) > f(x^*, \theta)$. A possible acquisition function for this could be an adapted form of constrained EI:

$$\alpha(\theta) = (d(\theta', \theta^*) - d(\theta, \theta^*))\mathbb{P}(f(x', \theta) > f(x^*, \theta)) \quad (10)$$

with θ' being the best θ found so far and $\mathbb{P}(\cdot)$ being the probability that the suggested solution is better under model parameters θ .

Case 2 is a bit more complicated, as it would only fix a partial solution a' , while allowing other variables, namely b , to vary. So, the problem becomes

$$\min d(\theta, \theta_0) \text{ s.t. } \max_b f((a', b), \theta) \geq f(x^*, \theta). \quad (11)$$

This can still be solved using a similar two-phase approach as above, first identifying any θ that satisfies the constraint by optimising over both, θ and b , then aiming to find a better θ . Note, however, that when varying b for a previously evaluated θ , only $f(a', b, \theta)$ needs to be evaluated. This could be taken into account by multiplying the EI value for those points by two.

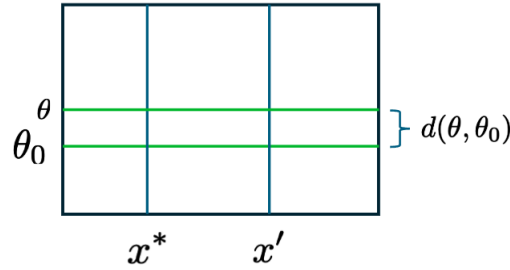
5.3.3.3 Find a context where the HITL proposal is (close to) optimal

For Case 1, this approach seeks the closest context θ' where x' is not just better than x^* , but is “close to optimal”. This means its value is within a small margin ϵ of the *true maximum* value achievable in that new scenario:

$$\theta' = \arg \min_{\theta} d(\theta, \theta_0) \quad | \quad f(x', \theta) \geq (1 - \epsilon) \max_x f(x, \theta). \quad (12)$$

This answers the question, “*What is the smallest change in input parameters that would make x' a genuinely good solution (e.g., within 1% of the best possible function value)?*”, see Figure 2.

This is a bi-level optimisation problem – optimising θ subject to x being optimal, and is thus difficult to solve. For some BO approaches to bi-level problems see [12, 10].



■ Figure 2

Case 2 is even more complicated, involving a maximisation on each side of the constraint:

$$\theta' = \arg \min_{\theta} d(\theta, \theta_0) \text{ s.t. } \max_b f((a', b), \theta) \geq (1 - \epsilon) \max_x f(x, \theta). \quad (13)$$

5.3.3.4 An Entropy Search Approach for Case 1

We observe that if we had access to the blackbox, i.e., $f(\cdot, \cdot)$, we could directly answer all questions regarding the problem of interest. However, our knowledge about $f(\cdot, \cdot)$ is only partial, as a result of solving the optimization problem in (8). Ideally, what we would like

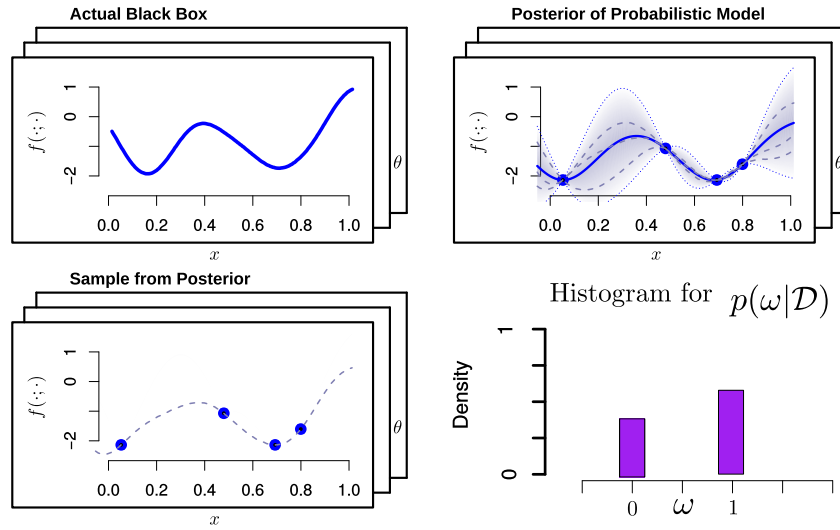
to do is to collect data efficiently, to provide an answer to the question, given Case 1. This involves evaluating the black box at new candidate locations $f(x_{\text{new}}, \theta_{\text{new}})$. However, since evaluating f is expensive, we would like to choose carefully these new evaluations. This is challenging, since providing an answer to the question corresponding to Case 1 is not exactly optimizing f .

In this section we describe how entropy search (ES) [33, 17, 19] can be used to address Case 1, described above. A nice feature of ES is that it can be easily adapted to aim at reducing the uncertainty about some particular feature of the objective, which need not be related to the optimum. This contrasts with most acquisition functions from the BO literature, which are particular designed to find the global optimum and may require complicated tweaking to address other goals (e.g, EI may require to perform two evaluations at each iteration, as described above). To our knowledge, ES has not been used to counterfactual analysis in the context of BO.

To address Case 1, we may define a random variable that represents the answer to the question posed. Namely,

$$\omega \equiv \begin{cases} 1 & \text{if there is a } \theta \text{ for which } x' \text{ is optimal.} \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

Since there is a probabilistic model for f , the black-box function, this induces a probability distribution for ω , as illustrated in Figure 3.

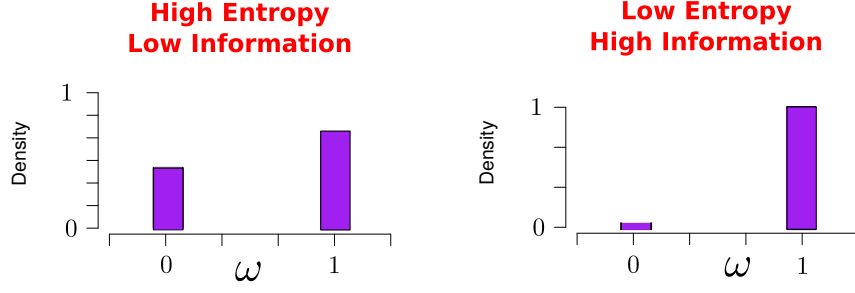


■ **Figure 3** Graphical illustration of the process of generating samples of ω . Note that we have illustrated the dependence of f with respect to θ as an extra axis.

Ideally, we would like to reduce our uncertainty about the potential values of ω , which directly translates into reducing our uncertainty about the potential answer, under Case 1. The idea is that if the entropy of ω is large, we are far from answering the question, and if the entropy of ω is small, we are close to answering the question. This is illustrated in Figure 4.

Therefore, with the goal of reducing the entropy of ω , we can define the following acquisition function:

$$\alpha(x, \theta) = (H[\omega|\mathcal{D}] - \mathbb{E}_y[H[\omega|\mathcal{D} \cup \{x, \theta, y\}]])/d(\theta, \theta_0), \quad (15)$$



■ **Figure 4** Graphical illustration of two distributions for ω . One with high and one with low entropy.

where $H[\cdot]$ denotes differential entropy, $\mathcal{D}$ are the data collected so far, the expectation is with respect to the predictive distribution of the probabilistic model, and $d(\theta, \theta_0)$ is the distance to θ_0 . Thus, we simply favor choosing points that reduce the most the entropy of ω in expectation, while preferring evaluations that are close to θ_0 .

Of course, evaluating (15) is very challenging. However, we can use that such an expression is simply the mutual information between y and ω [19]. Namely, $I(y; \omega)$. Since the mutual information is symmetric, we can swap the variables y and ω , resulting in the following equivalent expression to (15):

$$\alpha(x, \theta) = (H[y|\mathcal{D}] - \mathbb{E}_\omega[H[y|\mathcal{D}, \omega]])/d(\theta, \theta_0), \quad (16)$$

which now involves the entropy of the unconditional predictive distribution, $H[y|\mathcal{D}]$, an expectation with respect to ω , and the entropy of the conditional predictive distribution, $H[y|\mathcal{D}, \omega]$, given ω . The first term is easy to compute, the expectation can be approximated via Monte Carlo, simply by sampling ω , and $H[y|\mathcal{D}, \omega]$ must be approximated.

Computing $H[y|\mathcal{D}]$ is straight-forward. To sample ω , we can generate samples of the black-box function, and check whether ω takes values 1 or 0 under that sample. The conditional predictive distribution whose entropy, $H[y|\mathcal{D}, \omega]$, needs to be computed, requires approximation. Note that this depends on the predictive distribution given the sampled value for ω . Thus, we have to constrain the predictive distribution of the probabilistic model to be compatible with ω . This may be, in principle, done simply by incorporating extra likelihood factors that enforce that:

- there is a θ , for which x' is optimal, if $\omega = 1$,
- there is not a θ for which x' is optimal, if $\omega = 0$,

as it is done in standard predictive entropy search [19]. Depending on the probabilistic model, it can be the case that these factors must be approximated with Gaussians. Similar factors have been used in the context of multi-objective optimization [18].

Of course, there are some challenges to address, like a large number of samples have to be generated to ensure that ω takes values 1. Thus, we may relax that $\omega \in \{0, 1\}$ and simply let $\omega \in [0, 1]$ [11]. In this case, ω will simply give an intuitive idea about how close x' is to being optimal. Ideally, we should also try to use the already observed data $\mathcal{D}$ collected when solving problem (8). Therefore, we may want to impose strong correlations with respect to θ .

5.3.4 A Metric to Quantify the Expected Improvement of the Recommended Solution w.r.t. the Preferences of the HITL

In Bayesian optimization, determining when to stop the algorithm is crucial because the process is inherently expensive, often involving costly evaluations of the objective. Continuing indefinitely can lead to diminishing returns, where additional evaluations provide negligible

improvement in the solution while consuming significant resources. A well-defined stopping criterion ensures efficiency and prevents over-exploration. Moreover, stopping at the right time balances exploration and exploitation, avoiding unnecessary sampling in regions unlikely to yield better results, and thus optimizes resource allocation.

Stopping has been a concern in Bayesian optimization for a considerable time, though mainly in a non-HITL context. As e.g. set out in the recent textbook in the area by [14], from decision theory the time to stop is when the expected gain from the optimal observation exceeds the cost of acquisition (note that practically this requires an accurate utility function representing the decision maker preferences, a well-calibrated predictive model of $f()$ and an accurate model of the mapping from each solution to its evaluation cost). Early works such as by [22] discuss stopping once EI yields a value lower than a certain threshold in terms of the best value seen so far (1%, or 0.01 on the log scale). More recent work has explored this in terms of regret for EI over the best-observed value so far [30]. Using combinations of acquisition functions have also been explored: [26] proposed using EI to determine query points, but applies a threshold on the probability of improvement of the proposed sample to decide when to stop. [27] derives an upper bound for the simple regret (assuming a well-calibrated Gaussian process model), and utilizes this with a threshold for stopping hyperparameter optimization (recommending the threshold is of a similar magnitude as the statistical error of the empirical estimator in order for it to be useful in practice). Recent work by [25] relies on tracking information in a historical window of queried data and optimization is stopped if (i) these data all lie within a convex hull and (ii) local regret in this region falls below some preset threshold. All these approaches do require the user to determine one (or more) termination thresholds *a priori*.² As discussed (along with many other approaches) in [14], stopping may also be prompted by achieving some desired target function value (as distinct from seeking the best possible performing solution under $f()$). Garnett also notes that “*In many cases, the time of termination may in fact not be under the control of the optimization routine at all but instead decided by an external agent.*”.

We intuitively defined a metric that measures what is the expected gain in the objective if we keep evaluating the objective. For this, one only has to consider the knowledge about the objective provided by the probabilistic model. For simplicity, consider the noiseless evaluation setting, and let γ be the best observation so far. We can define a global expected improvement (GlobalEI) metric that estimates the expected improvement in the objective by performing as many evaluations as needed to obtain the objective’s global optimum:

$$\text{GlobalEI} = \mathbb{E}_{f(x^*)}[f(x^*) - \gamma], \quad (17)$$

where the expectation is with respect to the distribution of the global optimum, assuming maximization. Of course, the evaluation of (17) is intractable. However, the expectation can be approximated by Monte Carlo, simply by generating sample functions from the probabilistic model, which can then be optimized very cheaply. Figure 5 and 6 compare this metric with expected improvement and regret, showing that GlobalEI provides more reasonable estimates than EI of how close we are to solving the optimization problem. In Figure 6, the model hyper-parameters (the model is a GP) are fixed to the optimal ones. Thus, there should be no model bias. We speculate that the differences between the GlobalEI and regret are due to sub-optimal optimization when computing GlobalEI using samples from the GP.

² In a number of these works a naive baseline heuristic of stopping once the best found has not improved for a certain (user defined) number of queries is also compared.

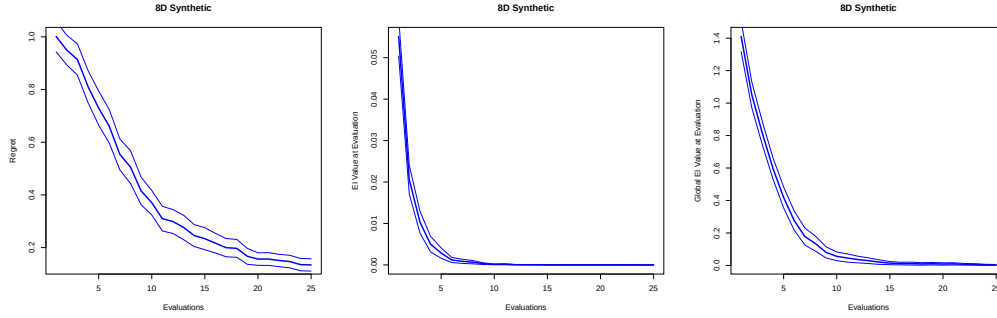


Figure 5 Comparison, on an 8-dimensional synthetic problem, where the objective is sampled from a GP, of (left) regret, expected improvement at the candidate location (center), and the proposed Global EI metric (right). We observe that Global EI does a better job than EI at estimating how far we are from solving the optimization problem. However, both metrics seem to be too optimistic in this case. Model hyper-parameters are estimated from the observed data and not set to optimal values. We report average results over 25 repetitions.

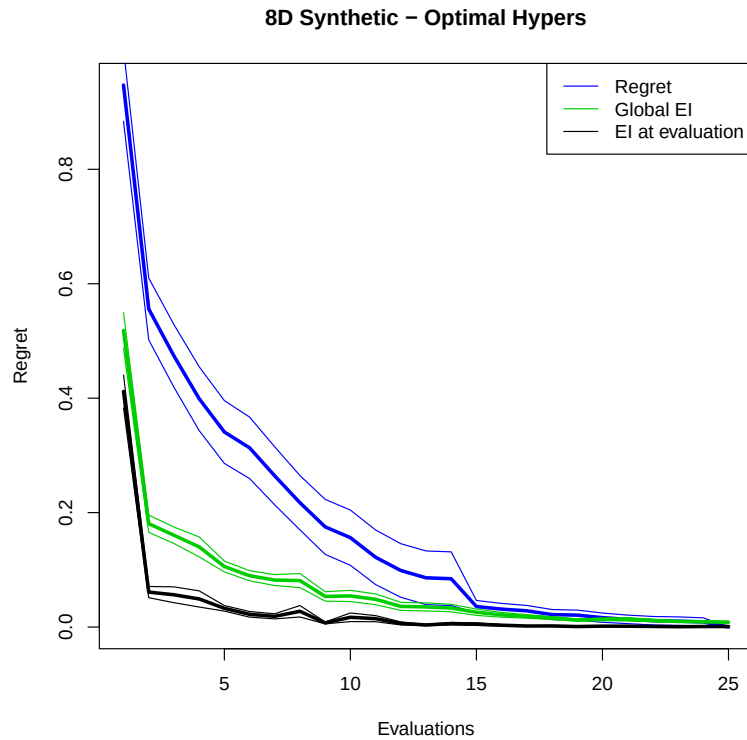


Figure 6 Comparison, on an 8-dimensional synthetic problem, where the objective is sampled from a GP, of (left) regret, expected improvement at the candidate location (center), and the proposed Global EI metric (right). We observe that Global EI does a better job than EI at estimating how far we are from solving the optimization problem. However, both metrics seem to be too optimistic in this case. Model hyper-parameters are fixed to the optimal ones. We report average results over 25 repetitions.

The proposed metric is related to the one described by [36], who, instead of considering the expected value in (17), considers the probability that $f(x^*) - \gamma$ is above a particular threshold value. GlobalEI is equivalent to the criterion described by [21], for minimization instead of maximization. Finally, in the literature, there are also more complicated rules for deciding when to stop when the objective evaluations may have different costs, as described by [37].

References

- 1 A. E. Abbas. *Foundations of Multi-attribute Utility*. Cambridge University Press, Cambridge, UK, 2018.
- 2 M. Adachi, B. Planden, D. A. Howey, M. A. Osborne, S. Orbell, N. Ares, and S. L. . . . Chau. Looping in the Human Collaborative and Explainable Bayesian Optimization. *arXiv preprint arXiv:2310.17273*, 2023.
- 3 B. Afsar, K. Miettinen, and F. Ruiz. Assessing the performance of interactive multiobjective optimization methods: A survey. *ACM Computing Surveys (CSUR)*, 54(4):1–27, 2021.
- 4 B. Afsar, A. B. Ruiz, and K. Miettinen. Comparing interactive evolutionary multiobjective optimization methods with an artificial decision maker. *Complex & Intelligent Systems*, 9(2):1165–1181, 2023.
- 5 B. Armbruster and E. Delage. Decision making under uncertainty when preference information is incomplete. *Management science*, 61(1):111–128, 2015.
- 6 R. Astudillo, Z. J. Lin, E. Bakshy, and P. I. Frazier. qEUBO: A decision-theoretic acquisition function for preferential Bayesian optimization. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS '23)*, pages 1093–1114, 2023.
- 7 J. Branke. Consideration of partial user preferences in evolutionary multiobjective optimization. In J. Branke, K. Deb, K. Miettinen, and R. Słowiński, editors, *Multiobjective Optimization: Interactive and Evolutionary Approaches*, volume 5252 of *Lecture Notes in Computer Science*, pages 157–178. Springer, Berlin / Heidelberg, Germany, 2008. doi:10.1007/978-3-540-88908-3_6.
- 8 J. Branke. MCDA and Multiobjective Evolutionary Algorithms. In S. Greco, M. Ehrgott, and J. R. Figueira, editors, *Multiple Criteria Decision Analysis: State of the Art Surveys*, volume 148 of *International Series in Operations Research and Management Science*, pages 977–1008. Springer, Cham, Switzerland, 2016. doi:10.1007/978-1-4939-3094-4_23.
- 9 J. Branke, S. Corrente, S. Greco, R. Słowiński, and P. Zielniewicz. Using Choquet integral as preference model in interactive evolutionary multiobjective optimization. *European Journal of Operational Research*, 250(3):884–901, 2016.
- 10 R. W. T. Chew, Q. P. Nguyen, and B. K. H. Low. BILBO: BILevel Bayesian Optimization. *arXiv preprint arXiv:2502.02121*, 2025.
- 11 M. De Santis and F. Rinaldi. Continuous reformulations for zero–one programming problems. *Journal of Optimization Theory and Applications*, 153:75–84, 2012.
- 12 O. Ekmekcioglu, N. Aydin, and J. Branke. Bayesian optimization of bilevel problems. *arXiv preprint arXiv:2412.18518*, 2024.
- 13 A. Forel, A. Parmentier, and T. Vidal. Explainable data-driven optimization: from context to decision and back again. In *International Conference on Machine Learning*, pages 10170–10187. PMLR, 2023.
- 14 R. Garnett. *Bayesian Optimization*. Cambridge University Press, 2023.
- 15 J. González, Z. Dai, A. Damianou, and N. D. Lawrence. Preferential Bayesian optimization. In *International Conference on Machine Learning*, pages 1282–1291. PMLR, 2017.
- 16 W. B. Haskell, W. Huang, and H. Xu. Preference elicitation and robust optimization with multi-attribute quasi-concave choice functions. *arXiv preprint arXiv:1805.06632*, 2018.

- 17 P. Hennig and C. J. Schuler. Entropy Search for Information-Efficient Global Optimization. *Journal of Machine Learning Research*, 13(6):1809–1837, 2012.
- 18 D. Hernández-Lobato, J. Hernandez-Lobato, A. Shah, and R. Adams. Predictive entropy search for multi-objective bayesian optimization. In *International conference on machine learning*, pages 1492–1501. PMLR, 2016.
- 19 J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. *Advances in neural information processing systems*, 27, 2014.
- 20 F. Huber, S. Rojas Gonzalez, and R. Astudillo. Bayesian preference elicitation for decision support in multi-objective optimization. *Journal of Multi-Criteria Decision Analysis*, 32(3): e70019, 2025.
- 21 H. Ishibashi, M. Karasuyama, I. Takeuchi, and H. Hino. A stopping criterion for Bayesian optimization by the gap of expected minimum simple regrets. In *International Conference on Artificial Intelligence and Statistics*, pages 6463–6497. PMLR, 2023.
- 22 D. Jones, M. Schonlau, and W. Welch. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13:455–492, 1998.
- 23 A. S. Jovine, T. Ye, F. Bahk, J. Wang, D. B. Shmoys, and P. I. Frazier. LISTEN to Your Preferences: An LLM Framework for Multi-Objective Selection. *arXiv preprint arXiv:2510.25799*, 2025.
- 24 K. Kobalczyk, Z. J. Lin, B. Letham, Z. Zhao, M. Balandat, and E. Bakshy. LILO: Bayesian Optimization with Interactive Natural Language Feedback. *arXiv preprint arXiv:2510.17671*, 2025.
- 25 S. Li, K. Li, and W. Li. “Why Not Looking backward?” A Robust Two-Step Method to Automatically Terminate Bayesian Optimization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 43435–43446. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/870c1e0589822bf37590b84984c345c4-Paper-Conference.pdf.
- 26 R. Lorenz, R. P. Monti, I. R. Violante, A. A. Faisal, C. Anagnostopoulos, R. Leech, and G. Montana. Stopping criteria for boosting automatic experimental design using real-time fMRI with Bayesian optimization. In *5th NIPS Workshop on Machine Learning and Interpretation in Neuroimaging: Beyond the Scanner*. arXiv preprint arXiv:1511.07827, 2015.
- 27 A. Makarova, H. Shen, V. Perrone, A. Klein, J. B. Faddoul, A. Krause, M. Seeger, and C. Archambeau. Automatic Termination for Hyperparameter Optimization. In I. Guyon, M. Lindauer, M. van der Schaar, F. Hutter, and R. Garnett, editors, *Proceedings of the First International Conference on Automated Machine Learning*, volume 188 of *Proceedings of Machine Learning Research*, pages 7/1–21. PMLR, 25–27 Jul 2022. URL <https://proceedings.mlr.press/v188/makarova22a.html>.
- 28 K. Miettinen. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research and Management Science*. Kluwer Academic Publishers, Boston, MA, USA, 1999.
- 29 G. Misitano, B. Afsar, G. Lárraga, and K. Miettinen. Towards explainable interactive multiobjective optimization: R-XIMO. *Autonomous Agents and Multi-Agent Systems*, (2), 2022.
- 30 V. Nguyen, S. Gupta, S. Rana, C. Li, and S. Venkatesh. Regret for Expected Improvement over the Best-Observed Value and Stopping Condition. In M.-L. Zhang and Y.-K. Noh, editors, *Proceedings of the Ninth Asian Conference on Machine Learning*, volume 77 of *Proceedings of Machine Learning Research*, pages 279–294, Yonsei University, Seoul, Republic of Korea, 15–17 Nov 2017. PMLR. URL <https://proceedings.mlr.press/v77/nguyen17a.html>.

- 31 S. Takeno, M. Nomura, and M. Karasuyama. Towards practical preferential Bayesian optimization with skew Gaussian processes. In *International Conference on Machine Learning*, pages 33516–33533. PMLR, 2023.
- 32 P. Viappiani and C. Boutilier. Optimal Bayesian recommendation sets and myopically optimal choice query sets. In *Advances in Neural Information Processing Systems (NIPS)*, volume 23, 2010.
- 33 J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4): 509–534, 2009.
- 34 J. Wallenius, J. S. Dyer, P. C. Fishburn, R. E. Steuer, S. Zionts, and K. Deb. Multiple Criteria Decision Making and Multiattribute Utility Theory: Recent Accomplishments and What Lies Ahead. *Management Science*, 54(7):1336–1349, 2008. doi:10.1287/mnsc.1070.0838.
- 35 H. Wang, J. Branke, and M. Poloczek. Bayesian Optimization with Preference Exploration by Monotonic Neural Network Ensemble. *arXiv preprint arXiv:2501.18792*, 2025.
- 36 J. Wilson. Stopping Bayesian optimization with probabilistic regret bounds. *Advances in Neural Information Processing Systems*, 37:98264–98296, 2024.
- 37 Q. Xie, L. Cai, A. Terenin, P. I. Frazier, and Z. Scully. Cost-aware stopping for Bayesian optimization. *arXiv preprint arXiv:2507.12453*, 2025.
- 38 B. Xin, L. Chen, J. Chen, H. Ishibuchi, K. Hirota, and B. Liu. Interactive Multiobjective Optimization: A Review of the State-of-the-Art. *IEEE Access*, 6:41256–41279, 2018. doi:10.1109/ACCESS.2018.2852976.
- 39 W. Xu, M. Adachi, C. N. Jones, and M. A. Osborne. Principled Bayesian optimization in collaboration with human experts. *Advances in Neural Information Processing Systems*, 37: 104091–104137, 2024.

5.4 Surrogate Models for Bayesian Optimization

Steven Adriaensen (University of Freiburg, DE)

Nathalie Bartoli (DTIS, ONERA, Université de Toulouse, FR)

Mickaël Binois (Inria, Université Côte d’Azur, CNRS, LJAD, Sophia Antipolis, FR)

Robert Gramacy (Virginia Polytechnic Institute – Blacksburg, US)

Rodolphe Le Riche (CNRS LIMOS at Mines Saint-Etienne, FR)

Szu Hui Ng (National University of Singapore, SG)

Inneke Van Nieuwenhuyse (Computational Mathematics, Hasselt University, BE)

Jixiang Qing (Mathematical Sciences, Lancaster University, UK)

Antonio Candelieri (Economics Management and Statistics, University of Milano-Bicocca, IT)

License © Creative Commons BY 4.0 International license

© Steven Adriaensen, Nathalie Bartoli, Mickaël Binois, Robert Gramacy, Rodolphe Le Riche, Szu Hui Ng, Inneke Van Nieuwenhuyse, Jixiang Qing, and Antonio Candelieri

5.4.1 Overview

Surrogates are models of the objective function learned from observed data. This working group focused on surrogate models for Bayesian Optimization (BO), surveying the range of surrogate modeling approaches currently in use and emerging in the literature. We discussed how different surrogate models address key modeling challenges in BO, compared their respective strengths and limitations, and reflected on what fundamentally qualifies as a surrogate model in the BO loop and why/when surrogate-based approaches are preferred over model-free alternatives.

5.4.1.1 What are Surrogate Models?

This report is concerned with **surrogate models** [35, 29] that appear as components of the Bayesian Optimization (BO) algorithm. Recall that the overall goal of such algorithm is to estimate the solutions to the optimization problem

$$\min_{x \in \mathcal{X}} f(x) \quad \text{where } f : \mathcal{X} \rightarrow \mathbb{R} \quad (18)$$

within a limited number of calls to f , a budget denoted as T . A pseudo-code for BO is given in Algorithm 1.

The surrogate, denoted $Y(x)$, is a model of the function f that is learned from past observations $\mathbf{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, where $y_i = f(x_i)$. The surrogate is a key ingredient to make BO cost effective by guiding the search towards high potential parts of the search space $\mathcal{X}$ without calling the true function f . In its most general form, $Y(x)$ is a random

■ **Algorithm 1** Sketch of a Bayesian Optimization algorithm that produces batches of q points at each iteration.

Require: Search space $\mathcal{X}$, Objective function $f : \mathcal{X} \rightarrow \mathbb{R}$

Require: A surrogate model $Y(x)$ (e.g., a Gaussian Process), an acquisition function $\alpha_Y(\cdot)$ (e.g., Expected Improvement, Upper Confidence Bound) that depends on $Y(x)$

Require: Max number of calls to f , T , Initial design points $\mathbf{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, where $y_i = f(x_i)$, $t \leftarrow n$

1: Initialize surrogate model with data $\mathbf{D}$

2: **while** $t \leq T$ **do**

3: Fit the surrogate model to the observed data $\mathcal{D}$

4: Optimize the acquisition function to find the next query points:

$$\{x_{t+1}, \dots, x_{t+q}\} \in \arg \max_{x'_1, \dots, x'_q \in \mathcal{X}^q} \alpha_Y(x'_1, \dots, x'_q)$$

5: Evaluate the objective functions:

$$y_{t+1} = f(x_{t+1}), \dots, y_{t+q} = f(x_{t+q}), \quad t \leftarrow t + q$$

6: Augment the data:

$$\mathbf{D} = \mathbf{D} \cup \{(x_{t+1}, y_{t+1}), \dots, (x_{t+q}, y_{t+q})\}$$

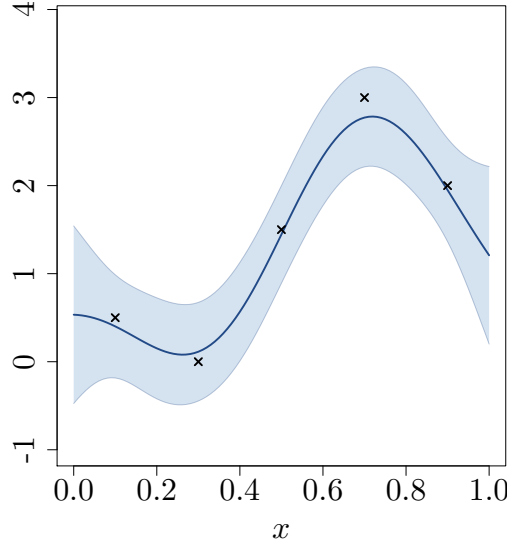
7: **end while**

8: **return**

9: - the best observed point $x^* \in \arg \min_{x_i \in \mathbf{D}} y_i$ (in noiseless context) or the best expected function at the observed points (in a noisy context) $x^* \in \arg \min_{x_i \in \mathbf{D}} \mathbb{Y}(x_i)$

10: - the last surrogate and $\mathbf{D}$

process where each instance $y(x, \omega) : \mathcal{X} \times \Omega \rightarrow \mathbb{R}$, with Ω the sample space, can be seen as a possible f . An illustration of surrogate in 1-D is given in Figure 7. To define a BO algorithm, a complete random process is not needed: the minimal amount of information needed to construct a global optimization algorithm such as a BO is *i*) a prediction of the function f at a new point x and *ii*) a measure of how the value of f is uncertain at this new x . For the illustration in Figure 7, these would be the expectation of $Y(x)$ and its variance, or any monotonous transformations of them. These two pieces of information are required in any acquisition criterion $\alpha_Y(\cdot)$ that trades off intensification of the search in regions of $\mathcal{X}$ that are likely to have high-performance with exploration of unknown regions.



■ **Figure 7** Illustration of a surrogate $Y(x)$. Here, it is a Gaussian process regression in the presence of noise. The solid line is the function prediction, the interval around has size plus/minus 2 standard deviations, the filled tube is $\mathbb{E}Y(x) \pm 2\sqrt{\mathbb{V}Y(x)}$, from [50].

In brief, in this report, surrogates are models of the objective function that provide, at least, both a prediction and a measure of uncertainty. We do not consider models directly outputting activation functions, or distribution models directly proposing points in $\mathcal{X}$ to evaluate (like in evolutionary optimization, e.g., CMA-ES [42, 41]).

5.4.1.2 Why Surrogate Models?

Probabilistic surrogate models enable principled, uncertainty-aware guidance, which is central to achieving high sample efficiency in settings where evaluations are expensive. Beyond guiding where to evaluate next, surrogate models can also inform which solution to return, for example by selecting the model-predicted optimum rather than the best observed evaluation, a distinction that is particularly relevant in settings with observation noise. Moreover, this enables non-myopic strategies that select evaluation points primarily to improve model fidelity rather than for their immediate objective value. Finally, surrogate models naturally support warm-starting by incorporating previously collected data, and provide an additional mechanism for injecting prior knowledge through the model prior. A recent framework for comparing surrogates is proposed by [72].

Different surrogate models are considered here: Gaussian Process in Section 5.4.2, tree-based models in Section 5.4.3, and neural models in Section 5.4.4.

5.4.2 Gaussian Process Models

Gaussian Processes (GPs) constitute the canonical surrogate model in Bayesian Optimization, due to their probabilistic formulation, analytical tractability, and flexibility. A GP defines a distribution over functions such that, for any finite set of inputs, the corresponding outputs follow a multivariate Gaussian distribution,

$$Y(x) \sim \mathcal{N}(m(x), c(x, x')),$$

where the mean function $m(x) \equiv \mathbb{E}Y(x)$ encodes prior beliefs about the function values and the covariance function or kernel $c(x, x') \equiv \text{Cov}(Y(x), Y(x'))$ specifies assumptions about smoothness, correlation structure, and uncertainty.

In practice, most modeling expressiveness arises from the covariance function, which depends only on the inputs and must be positive definite. Its diagonal elements $c(x, x)$ yield uncertainty estimates on the predictions, which are central to acquisition-driven exploration. **Kernel design** [87, 25], therefore plays a fundamental role in GP-based BO, allowing the incorporation of structural assumptions such as stationarity, smoothness, periodicity [24], or compositional structure [23, 32]. Despite the flexibility offered by kernel design, the current practice still largely relies on transformations that reduce modeling to stationary kernels.

A key strength of Gaussian Processes lies in their ability to **accommodate complex input spaces** through kernel construction. This is particularly relevant when optimizing over design spaces where the classical Euclidean distances are not meaningful, e.g., when tuning hyperparameters of Machine Learning algorithms.

Kernels have been specifically defined for **categorical variables** [38], whether qualitative or quantitative [94]. Among the principles used to design these kernels, it seems important to learn the underlying group structure [71, 67]. Embedding in latent (continuous) spaces are, of course, also a key to handling categorical variables [93, 18].

Hierarchical/conditional input spaces [4], which may also be referred to as a variable-size spaces [66], or as hierarchical domains [40] have also their own kernels. This is an important class of search space $\mathcal{X}$ because it describes the numerous, practical situations where specific sub-systems can be added or removed, each coming with its own parameters. The algebraic distance introduced in [75] and the resulting kernel known as the Alg-Kernel, is symmetric positive definite and available in the opensource SMT framework³. This distance is a particular case of the more general hierarchical distance introduced in [40].

In fact, kernels over all types of possible discrete mathematical objects seem to have been defined: **kernels over sets** [31, 80], **over combinatorial variables** [12, 3], **over graphs** [68], over combinatorial variables seen as graphs [63], over object or label preferences [5] ...

Mixed continuous/discrete spaces can be handled by combining continuous/discrete kernels through any kernel combination technique [74]. What is important is that the combinations yields a valid, positive-definite, kernel. Most often, this is achieved by kernel product [18]. Adding kernels is also possible [22].

The performance of GPs with standard kernels deteriorates as **input dimensionality** increases. In high-dimensional spaces, the distances between points become nearly identical (a phenomenon referred to as distance concentration), leading to posterior means and variances that are approximately the same throughout the search space. consequently, the surrogate can no longer meaningfully guide the acquisition function during the search. Moreover, identifiability issues arise, as different kernel parameterizations can produce practically equal likelihoods or posteriors. Recent work addresses this challenge through projection-based kernels [9], active subspaces, and dimension-adaptive modeling strategies [7, 14, 34]. In the context of BO, these approaches are often tightly coupled with trust-region or local modeling assumptions.

Classical GP inference scales cubically with the number of observations, which used to limit the applicability of GPs to **small-data** regimes. In recent research, a wide range of scalable approximations has been developed to accommodate **large data** regimes, including the Vecchia approximation, sparse inducing-point methods, variational formulations, and

³ https://smt.readthedocs.io/en/latest/_src_docs/applications/Mixed_Hier_usage.html

local approximations [45, 85, 13]. Local and partitioned GP models, such as treed or neighborhood-based approaches, further improve scalability while retaining uncertainty estimates [36].

Many real-world optimization problems exhibit **nonstationary output** behavior, both in function values and noise characteristics. Nonstationary kernels, input warping, and locally adaptive GPs provide mechanisms to relax stationarity assumptions [64, 79, 62, 73]. Gaussian processes can also model **heteroscedastic noise** by allowing the observation variance to vary across the input space [33, 82, 6]; yet, this usually requires more data. It also increases inference complexity and often requires a more careful prior specification.

An important advantage of GPs is their ability to incorporate **expert knowledge** through constraints on the output space. Inequality constraints, boundedness, monotonicity, and convexity can be enforced either approximately or exactly through kernel design or posterior conditioning [19, 55]. Physics-informed Gaussian Processes explicitly encode governing equations, such as partial differential equations, as soft or hard constraints on the function space [69, 15, 54].

GPs benefit from their strong theoretical link with Reproducible Kernel Hilbert Spaces (RKHS). Once the GP’s parameters have been estimated, the results (e.g., variance, lengthscales) have a **straightforward interpretation**. In the case of **extrapolation** (away from observed data), predictions will naturally revert to the prior (and are thus strongly influenced by the chosen kernel). Recent work has explored kernels that induce structured extrapolation, such as spectral and Brownian motion-based constructions [88, 92, 52].

GPs also allow the analyst to exploit **derivative information**, by analytically differentiating the posterior. Gradient-enhanced GP models can significantly improve sample efficiency in BO, especially in smooth optimization landscapes [83, 89, 44, 21, 26], and when the function is deterministic. When the output data are noisy, or very sparse, though, the estimated value of the derivative may exhibit high uncertainty. Many BO problems involve **multiple correlated outputs** (e.g., in the case of multi-fidelity evaluations, multi-task objectives, or vector-valued responses). Multi-output Gaussian Processes provide a principled framework for modeling such dependencies through structured covariance functions [76, 16, 57], which can substantially improve the sample efficiency during the optimization. The downside is an increase in modeling/computational complexity. Designing interpretable cross-covariance structures and scaling inference to realistic problem sizes remain open challenges, particularly when combined with other extensions such as nonstationarity or constraints.

5.4.3 Tree-based surrogates and other local surrogates

The scope of these models includes hierarchical treed-models, tree ensembles such as Random Forests (RF) and Bayesian Additive Regression Trees (BART), and other local models [36, 27]. In what follows, we zoom in mainly on tree-based surrogates. These can be split up in two subtypes: **Treed-Models** and **Tree Ensembles**. Treed-Models use trees for recursive search space partitioning. This often leads to simple tree structures, with possibly complex (local) models (e.g., local GPs) fit to the data in each leaf [38, 37]. This approach is highly effective for handling mixed and conditional search spaces. Tree Ensembles, by contrast, represent the objective function as a weighted sum or average of trees with simple leaves. Examples include Random Forests (which are similar to bagging) and BART (similar to boosting) [17].

The primary benefits of tree-based surrogates stem from their “divide and conquer” nature. They are computationally efficient even in the case of **high dimensions** and a **large data**, as complexity adaptively scales with training set size [48]. They are fit for modeling **non-stationary and heteroscedastic output** functions (though, in the case of standard

RFs, heteroscedasticity requires estimating separate trees for the noise variance [17]). They are highly flexible w.r.t. the **input types** and the **structure of the input space**: they do not require any particular assumptions w.r.t. the nature of the input variables, can naturally accommodate hierarchical search spaces, and constraints on the inputs can be easily included. Moreover, ensembles like BART automatically provide **regularization** through an implicit application of Occam’s razor [17]. The practical implementation of the models is supported by various **toolboxes** across Python and R for BART, RF, Treed-GP, and local GP methods (including trust region approaches like TuRBO, [27]).

Nevertheless, tree-based models also present a number of downsides. They are known for their tendency to **overfit**, in the case of **small data** the resulting trees may be overly simple. They also **lack gradient information**, and do not leverage any spatial information across the leaves in the uncertainty estimates and predictions.

The performance of treed-models, when used in the BO loop, fully depends on the type of surrogate used to fit the data in the leaves. Modeling **multiple outputs** may require multiple trees, unless you have MO surrogate models in the leaves. **Acquisition** is computationally expensive, as it is performed per leaf, after which the highest value among the leaves is chosen. For ensembles, inference is often performed via Monte Carlo, and acquisition is usually done using candidate sets. Ensembles do not easily allow for the inclusion of **expert knowledge**.

5.4.4 Neural Models

We discussed a broad class of surrogate models based on neural networks for predicting unseen function values in Bayesian optimization. This category encompasses a wide range of approaches with very different characteristics. Classical methods rely on training Bayesian neural networks [1, 65] (BNNs) directly on the observed data within the BO loop, while more recent approaches leverage pretrained models, such as prior-data fitted networks [59] (PFNs) or large language models (LLMs), that perform Bayesian-style prediction via in-context learning, i.e., observations are an input used to condition the pretrained model.

While diverse in their realizations, neural surrogate models share several common traits. Compared to Gaussian processes, they offer **richer functional priors**, **non-stationary output scales**, and **heteroscedastic noise modeling** without explicit kernel design. Relative to tree-based models, they provide **smooth predictive surfaces** and **efficient input gradients**, which are for example useful for gradient-based acquisition function optimization. However, these methods also share notable drawbacks, including **training instabilities**, **limited interpretability**, and **limited integration in existing toolkits**, compared to more classical surrogates.

Bayesian Neural Networks (BNNs)

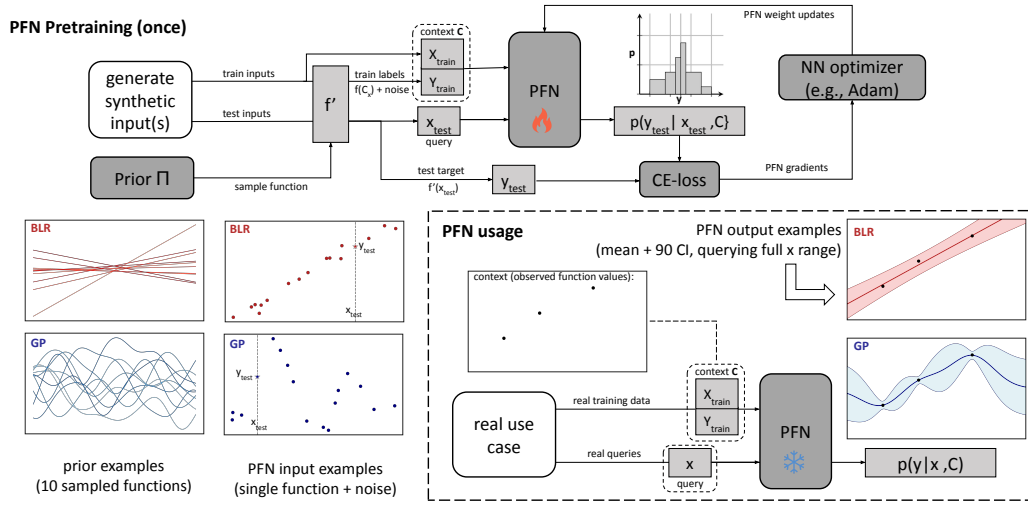
BNNs [1, 65] represent the integration of Bayesian inference with neural networks, where prior distributions are placed over network components (typically the parameters) and Bayesian inference is performed to quantify epistemic uncertainty [60, 56]. Given a prior $p(\mathbf{w})$ over weights and observed data $\mathcal{D}$, the posterior $p(\mathbf{w}|\mathcal{D})$ induces a predictive distribution $p(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x}, \mathbf{w})p(\mathbf{w}|\mathcal{D}) d\mathbf{w}$ that naturally provides uncertainty quantification. Since exact inference is intractable, various approximation schemes have been proposed, including Markov Chain Monte Carlo (MCMC) methods [60], variational inference [39, 8], Laplace approximations [56, 20], and partial Bayesian methods that place priors only on a subset of parameters [77], such as last-layer approaches [78, 43, 11].

Compared to Gaussian Processes, BNNs offer the ability to **model non-stationary functions**, learn **problem-specific similarity metrics**, and **scale to large observation sets** where GP cubic complexity becomes prohibitive. BNNs have been explored as surrogate models in BO [81, 78], with recent work demonstrating competitive performance with GPs on standard benchmarks [53, 11]. However, BNNs present their own caveats typically stemming from **approximate inference**, and empirically their performance depends on architectural choices and prior specification, **requiring careful tuning** to achieve well-calibrated uncertainty quantification, and to avoid **overfitting in low data regimes** [86, 28, 53].

Prior-data Fitted Networks (PFNs)

PFNs [59] are models pretrained to do in-context Bayesian prediction for a specific function prior $\Pi = p(f)$. Note that PFNs are closely related (can be seen as a special instance of) conditional Neural processes [30, 61]. Typically, a PFN is a transformer neural network [84], pretrained on synthetic data by repeatedly sampling a function $f' \sim \Pi(f)$, picking “train” and “test” inputs (X_{train}, X_{test}) (often uniform random noise), generating labels $Y_{train} = f'(X_{train})$ (optionally with observation noise), and minimizing expected cross-entropy of the predictive distribution over queries $x_{test} \in X_{test}$ conditioned on context $C = (X_{train}, Y_{train})$. Note that a PFNs output a bar distribution over finely discretized output range, where the PFN has an output per bar, corresponding to the probability that y_{test} falls in the associated interval. Under risk minimization, the solution matches the Bayesian posterior predictive $p(y_{test} | x_{test}, C)$ induced by Π , so PFNs learn Bayesian one-shot prediction in a single forward pass. For example, a PFN pretrained on samples of a GP prior, provably approximates GP prediction in a single forward pass. The training and usage of PFNs are illustrated in Figure 8. While TabPFN [47, 46], a tabular foundation model, arguably remains the most prominent application of this paradigm, PFNs have already been explored in the context of BO. Müller *et al.* [58] first proposed PFNs as a surrogate for black box BO in 2023. Recent work has explored PFN surrogates in the context of Freeze-thaw BO [70, 51] and high-dimensional BO [91]. Recently, PFNs have been integrated in the popular BoTorch framework [2].

In contrast to BNNs, PFNs are **not retrained on observed function values**, but amortize Bayesian inference through pretraining on synthetic tasks, yielding **strong performance with very few observations**. This makes them attractive for Bayesian optimization, where **posterior updates reduce to a single forward pass**, avoiding retraining, hyperparameter optimization, and associated instabilities. PFNs also naturally support **mixed input spaces** including continuous, discrete, and categorical variables, as demonstrated by TabPFN on heterogeneous tabular data, which is a practical advantage for BO with structured or conditional search spaces. At the same time, current PFN transformer architectures rely on dense attention, leading to **quadratic compute and memory growth** and making **very large contexts computationally demanding** compared to tree-based surrogates. Another trade-off is that PFNs offer **explicit control over the functional prior**, but **changing this prior requires pretraining from scratch**. As with any learned surrogate, **strong out-of-distribution mismatch may lead to instability**, although empirical performance is **often reasonable**. Overall, PFN-based Bayesian optimization remains **an emerging and experimental paradigm**, with several modeling and design choices still under active investigation.



■ **Figure 8** Illustration of the training and usage of prior-data fitted networks (PFNs), with specific in-/output examples fitting a Bayesian Linear Regression (BLR) and Gaussian Process (GP) prior.

Large Language Models (LLMs)

LLMs are general-purpose neural networks trained on massive text corpora to model and generate natural language, and have emerged as a versatile foundation for a wide range of reasoning, prediction, and decision-making tasks [10]. They can be employed for Bayesian optimization in many different ways, ranging from proposing candidate points [90] to guiding search through natural-language feedback [49]; these broader uses are discussed in the GenAI for BO working group (Section 5.5). In our discussions, we focused on a deliberately narrow and directly comparable setting, where the LLM is used *as a surrogate model* via in-context learning: previously observed input–output pairs (x, y) are provided in the prompt, a new candidate x_{test} is queried, and the model predicts the corresponding outcome y_{test} . While this setup is natural, it may not be the most practical way to deploy LLMs in BO, and we are not aware of prior work considering this exact formulation, it enables a clean conceptual contrast with PFNs by isolating in-context prediction with frozen weights.

In this setting, LLMs share some appealing properties with PFNs, most notably **data efficiency via in-context learning**, allowing them to adapt from very few observations without retraining. Beyond this, LLMs offer distinctive advantages: they can leverage **broad world knowledge**, support flexible textual input and output, and are **easy to condition on auxiliary information**, enabling predictions to be augmented with explanations, constraints, or meta-information that go beyond scalar regression targets. At the same time, these strengths come with substantial drawbacks when used as BO surrogates. LLMs are typically **large and computationally expensive**, making repeated surrogate queries costly. Their inductive biases are **implicit and hard to control**, limiting the ability to align model behavior with a desired functional prior. Moreover, predictions can be **sensitive to prompt formulation and syntax**, introducing brittleness and requiring careful prompt tuning. As a result, while LLMs provide a flexible and expressive baseline for in-context surrogate modeling, their use in this direct surrogate role is best viewed as **experimental** and primarily valuable for comparison and exploration rather than as a default practical choice.

5.4.4.1 Conclusions

The discussions in this working group underscored that surrogate modeling remains a central and evolving design choice in Bayesian optimization rather than a settled component. Gaussian processes continue to be the default surrogate model due to their strong theoretical foundations, flexibility, and principled uncertainty estimates, but they also exhibit limitations in scalability, representational capacity, and prior specification, which are often addressed through increasingly complex extensions. This has motivated interest in alternative surrogate models, including neural approaches. While traditional in-weight learning methods such as Bayesian neural networks have so far had limited impact in BO, possibly due to the structured, tabular nature of BO data, the low-data regime, and the need for frequent model updates, settings in which deep learning has historically been less competitive. More recent approaches based on in-context learning, including prior-data fitted networks and large language models, offer a qualitatively different perspective by shifting weight learning to a pretraining phase and avoiding retraining during optimization. These models have shown strong performance in low-data and tabular settings and introduce new capabilities, such as explicit control over prior assumptions (PFNs) or access to broad world knowledge and flexible input–output interfaces (LLMs). Finally, we emphasized that surrogate models should not be considered in isolation, as their effectiveness depends on interactions with other components of the Bayesian optimization pipeline, and that model fit and prior mismatch remain underexplored issues where improved evaluation practices could yield deeper insight.

References

- 1 Julyan Arbel, Konstantinos Pitas, Mariia Vladimirova, and Vincent Fortuin. A primer on bayesian neural networks: Review and debates. *arXiv preprint arXiv:2309.16314*, 2023.
- 2 Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. Botorch: A framework for efficient monte-carlo bayesian optimization. *Advances in neural information processing systems*, 33:21524–21538, 2020.
- 3 Ricardo Baptista and Matthias Poloczek. Bayesian optimization of combinatorial structures. In *International conference on machine learning*, pages 462–471. PMLR, 2018.
- 4 Lucas Baraton, Annafederica Urbano, Loïc Brevault, and Mathieu Balesdent. Bayesian quality-diversity optimization for conditional search-space problems: L. baraton et al. *Optimization and Engineering*, pages 1–46, 2025.
- 5 Alessio Benavoli and Dario Azzimonti. A tutorial on learning from preferences and choices with gaussian processes. *arXiv preprint arXiv:2403.11782*, 2024.
- 6 Mickael Binois, Robert B Gramacy, and Mike Ludkovski. Practical heteroscedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821, 2018.
- 7 Mickael Binois and Nathan Wycoff. A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization. *ACM Transactions on Evolutionary Learning and Optimization*, 2(2):1–26, 2022.
- 8 Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.
- 9 Mohamed Amine Bouhlel, Nathalie Bartoli, Abdelkader Otsmane, and Joseph Morlier. Improving kriging surrogates of high-dimensional design models by partial least squares dimension reduction. *Structural and Multidisciplinary Optimization*, 53(5):935–952, 2016.
- 10 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

- 11 Paul Brunzema, Mikkel Jordahn, John Willes, Sebastian Trimpe, Jasper Snoek, and James Harrison. Bayesian optimization via continual variational last layer training. In *The Thirteenth International Conference on Learning Representations*, 2025.
- 12 Poompol Buathong, David Ginsbourger, and Tupaluck Krityakierne. Kernels over sets of finite sets using rkhs embeddings, with application to bayesian (combinatorial) optimization. In *International conference on artificial intelligence and statistics*, pages 2731–2741. PMLR, 2020.
- 13 Jian Cao, Joseph Guinness, Marc G Genton, and Matthias Katzfuss. Scalable gaussian-process regression and variable selection using vecchia approximations. *The Journal of Machine Learning Research*, 23(1):15799–15828, 2022.
- 14 Jialei Chen, Zhehui Chen, Chuck Zhang, and CF Jeff Wu. Apik: Active physics-informed kriging model with partial differential equations. *SIAM/ASA Journal on Uncertainty Quantification*, 10(1):481–506, 2022.
- 15 Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart. Solving and learning nonlinear pdes with gaussian processes. *Journal of Computational Physics*, 447:110668, 2021.
- 16 Zexun Chen, Bo Wang, and Alexander N Gorban. Multivariate gaussian and student-t process regression for multi-output prediction. *Neural Computing and Applications*, 32(8):3005–3028, 2020.
- 17 Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, pages 266–298, 2010.
- 18 Jhouben Cuesta Ramirez, Rodolphe Le Riche, Olivier Roustant, Guillaume Perrin, Cedric Duranton, and Alain Gliere. A comparison of mixed-variables bayesian optimization approaches. *Advanced Modeling and Simulation in Engineering Sciences*, 9(1):6, 2022.
- 19 Sébastien Da Veiga and Amandine Marrel. Gaussian process modeling with inequality constraints. In *Annales de la faculté des sciences de Toulouse*, volume 21, pages 529–555, 2012.
- 20 Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021.
- 21 Filip de Roos, Alexandra Gessner, and Philipp Hennig. High-dimensional gaussian process inference with derivatives. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2535–2545. PMLR, 18–24 Jul 2021.
- 22 Nicolas Durrande, David Ginsbourger, and Olivier Roustant. Additive covariance kernels for high-dimensional gaussian process modeling. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 21, pages 481–499, 2012.
- 23 Nicolas Durrande, David Ginsbourger, Olivier Roustant, and Laurent Carraro. Anova kernels and rkhs of zero mean functions for model-based sensitivity analysis. *Journal of Multivariate Analysis*, 115:57–67, 2013.
- 24 Nicolas Durrande, James Hensman, Magnus Rattray, and Neil D Lawrence. Detecting periodicities with gaussian processes. *PeerJ Computer Science*, 2:e50, 2016.
- 25 David Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.
- 26 David Eriksson, Kun Dong, Eric Lee, David Bindel, and Andrew G Wilson. Scaling gaussian process regression with derivatives. In *Advances in Neural Information Processing Systems*, pages 6866–6876, 2018.
- 27 David Eriksson, Michael Pearce, Jacob Gardner, Ryan P Turner, and Matthias Poloczek. Scalable global optimization via local bayesian optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

- 28 Jonathan Foldager, Mikkel Jordahn, Lars K Hansen, and Michael R Andersen. On the role of model uncertainties in bayesian optimisation. In *Uncertainty in Artificial Intelligence*, pages 592–601. PMLR, 2023.
- 29 Alexander Forrester, Andras Sobester, and Andy Keane. *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons, 2008.
- 30 Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International conference on machine learning*, pages 1704–1713. PMLR, 2018.
- 31 Roman Garnett, Michael A Osborne, and Stephen J Roberts. Bayesian optimization for sensor set selection. In *Proceedings of the 9th ACM/IEEE international conference on information processing in sensor networks*, pages 209–219, 2010.
- 32 David Ginsbourger, Xavier Bay, Olivier Roustant, and Laurent Carraro. Argumentwise invariant kernels for the approximation of invariant functions. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 21, pages 501–527, 2012.
- 33 Paul W Goldberg, Christopher KI Williams, and Christopher M Bishop. Regression with input-dependent noise: A gaussian process treatment. *Advances in neural information processing systems*, pages 493–499, 1998.
- 34 Miguel González-Duque, Richard Michael, Simon Bartels, Yevgen Zainchkovskyy, Søren Hauberg, and Wouter Boomsma. A survey and benchmark of high-dimensional bayesian optimization of discrete sequences. *Advances in Neural Information Processing Systems*, 37:140478–140508, 2024.
- 35 Robert B Gramacy. *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC, 2020.
- 36 Robert B Gramacy and Daniel W Apley. Local gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578, 2015.
- 37 Robert B Gramacy and Herbert KH Lee. Bayesian treed gaussian process models with optimization applications. *Journal of Computational and Graphical Statistics*, 17(4):964–980, 2008.
- 38 Robert B Gramacy and Matthew Taddy. Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an R package for treed gaussian process models. *Journal of Statistical Software*, 33(1):1–48, 2010.
- 39 Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.
- 40 Edward Hallé-Hannan, Charles Audet, Youssef Diouane, Sébastien Le Digabel, and Paul Saves. A distance for mixed-variable and hierarchical domains with meta variables. *Neuro-computing*, 653:131208, 2025.
- 41 Nikolaus Hansen. The CMA evolution strategy: A tutorial. *arXiv preprint arXiv:1604.00772*, 2016.
- 42 Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evol. Comput.*, 9(2):159–195, June 2001.
- 43 James Harrison, John Willes, and Jasper Snoek. Variational bayesian last layers. In *The Twelfth International Conference on Learning Representations*, 2024.
- 44 Xu He and Peter Chien. On the instability issue of gradient-enhanced gaussian process emulators for computer experiments. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):627–644, 2018.
- 45 James Hensman, Alexander G Matthews, Maurizio Filippone, and Zoubin Ghahramani. Mcmc for variationally sparse gaussian processes. *Advances in neural information processing systems*, 28, 2015.

- 46 Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeyer, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.
- 47 Noah Hollmann, Samuel Müller, Katharina Eggenberger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.
- 48 Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *LION*, 2011.
- 49 Katarzyna Kobalczuk, Zhiyuan Jerry Lin, Benjamin Letham, Zhuokai Zhao, Maximilian Balandat, and Eytan Bakshy. Lilo: Bayesian optimization with interactive natural language feedback. *arXiv preprint arXiv:2510.17671*, 2025.
- 50 Rodolphe Le Riche and Nicolas Durrande. An overview of kriging for researchers. Summer school on Modeling and Numerical Methods for Uncertainty Quantification, slides as HAL course cel-02285439, <https://hal.science/cel-02285439>, September 2019.
- 51 Dong Bok Lee, Aoxuan Silvia Zhang, Byungjoo Kim, Junhyeon Park, Steven Adriaensen, Juho Lee, Sung Ju Hwang, and Hae Beom Lee. Cost-sensitive freeze-thaw bayesian optimization for efficient hyperparameter tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- 52 Minyong R Lee and Art B Owen. Single nugget kriging. *Statistica Sinica*, pages 649–669, 2018.
- 53 Yucen Lily Li, Tim G. J. Rudner, and Andrew Gordon Wilson. A study of bayesian neural network surrogates for bayesian optimization. In *The Twelfth International Conference on Learning Representations*, 2024.
- 54 Da Long, Zheng Wang, Aditi Krishnapriyan, Robert Kirby, Shandian Zhe, and Michael Mahoney. Autoip: A united framework to integrate physics into gaussian processes. In *International Conference on Machine Learning*, pages 14210–14222. PMLR, 2022.
- 55 Andrés F López-Lopera, François Bachoc, Nicolas Durrande, and Olivier Roustant. Finite-dimensional gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1224–1255, 2018.
- 56 David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 05 1992.
- 57 Wesley J Maddox, Maximilian Balandat, Andrew G Wilson, and Eytan Bakshy. Bayesian optimization with high-dimensional outputs. *Advances in Neural Information Processing Systems*, 34:19274–19287, 2021.
- 58 Samuel Müller, Matthias Feurer, Noah Hollmann, and Frank Hutter. PFNs4BO: In-context learning for Bayesian optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25444–25470. PMLR, 23–29 Jul 2023.
- 59 Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022.
- 60 Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, Berlin, Heidelberg, 1996.
- 61 Tung Nguyen and Aditya Grover. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. In *International Conference on Machine Learning*, pages 16569–16594. PMLR, 2022.
- 62 Marcus M Noack and James A Sethian. Advanced stationary and nonstationary kernel designs for domain-aware gaussian processes. *Communications in Applied Mathematics and Computational Science*, 17(1):131–156, 2022.

- 63 Changyong Oh, J Tomczak, Efstratios Gavves, and Max Welling. Combo: Combinatorial bayesian optimization using graph representations. In *ICML Workshop on Learning and Reasoning with Graph-Structured Data*, 2019.
- 64 Christopher J Paciorek and Mark J Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics: The official journal of the International Environmetrics Society*, 17(5):483–506, 2006.
- 65 Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, Laurence Aitchison, Julyan Arbel, David Dunson, Maurizio Filippone, Vincent Fortuin, Philipp Hennig, José Miguel Hernández-Lobato, et al. Position: Bayesian deep learning is needed in the age of large-scale ai. In *International Conference on Machine Learning*, 2024.
- 66 Julien Pelamatti, Loic Brevault, Mathieu Balesdent, El-Ghazali Talbi, and Yannick Guerin. Bayesian optimization of variable-size design space problems. *Optimization and Engineering*, 22:387–447, 2021.
- 67 Raphaël Carpintero Perez, Sébastien Da Veiga, and Josselin Garnier. A reproducible comparative study of categorical kernels for gaussian process regression, with new clustering-based nested kernels. *arXiv preprint arXiv:2510.01840*, 2025.
- 68 Raphaël Carpintero Perez, Sébastien Da Veiga, Josselin Garnier, and Brian Staber. Gaussian process regression with sliced wasserstein weisfeiler-lehman graph kernels. In *International Conference on Artificial Intelligence and Statistics*, pages 1297–1305. PMLR, 2024.
- 69 Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Numerical gaussian processes for time-dependent and nonlinear partial differential equations. *SIAM Journal on Scientific Computing*, 40(1):A172–A198, 2018.
- 70 Herilalaina Rakotoarison, Steven Adriaensen, Neeratyoy Mallik, Samir Garibov, Eddie Bergman, and Frank Hutter. In-context freeze-thaw bayesian optimization for hyperparameter optimization. In *International Conference on Machine Learning*, pages 41982–42008. PMLR, 2024.
- 71 Olivier Roustant, Esperan Padonou, Yves Deville, Aloïs Clément, Guillaume Perrin, Jean Giorla, and Henry Wynn. Group kernels for gaussian process metamodels with categorical inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):775–806, 2020.
- 72 Kellin N Rumsey, Graham C Gibson, Devin Francom, and Reid Morris. All emulators are wrong, many are useful, and some are more useful than others: A reproducible comparison of computer model surrogates. *arXiv preprint arXiv:2512.09060*, 2025.
- 73 Annie Sauer, Andrew Cooper, and Robert B Gramacy. Non-stationary gaussian process surrogates. *arXiv preprint arXiv:2305.19242*, 2023.
- 74 Paul Saves, Youssef Diouane, Nathalie Bartoli, Thierry Lefebvre, and Joseph Morlier. A mixed-categorical correlation kernel for gaussian process. *Neurocomputing*, 550:126472, 2023.
- 75 Paul Saves, Rémi Lafage, Nathalie Bartoli, Youssef Diouane, Jasper H. Bussemaker, Thierry Lefebvre, John. T. Hwang, Joseph Morlier, and Joaquim R. R. A. Martins. SMT 2.0: A surrogate modeling toolbox with a focus on hierarchical and mixed variables Gaussian processes. *Advances in Engineering Software*, 188:1–15, 2024.
- 76 Amar Shah, AC UK, Zoubin Ghahramani, and CAM AC. Pareto frontier learning with expensive correlated objectives. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1919–1927, 2016.
- 77 Mrinank Sharma, Sebastian Farquhar, Eric Nalisnick, and Tom Rainforth. Do bayesian neural networks need to be fully stochastic? In *International Conference on Artificial Intelligence and Statistics*, pages 7694–7722. PMLR, 2023.
- 78 Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable bayesian optimization using deep neural networks. In *International conference on machine learning*, pages 2171–2180. PMLR, 2015.

- 79 Jasper Snoek, Kevin Swersky, Richard S Zemel, and Ryan P Adams. Input warping for bayesian optimization of non-stationary functions. In *International Conference on Machine Learning*, pages 1674–1682, 2014.
- 80 Babacar Sow, Rodolphe Le Riche, Julien Pelamatti, Merlin Keller, and Sanaa Zannane. Optimization and metamodeling of functions defined over clouds of points. In *workshop SAMOURAI*, Paris – FR, December 2024. réseau thématique Uncertainty Quantification (RT UQ). HAL document hal-04830176, <https://hal.science/hal-04830176>.
- 81 Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, and Frank Hutter. Bayesian optimization with robust bayesian neural networks. *Advances in neural information processing systems*, 29, 2016.
- 82 Michalis K Titsias and Miguel Lázaro-Gredilla. Variational heteroscedastic gaussian process regression. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 841–848, 2011.
- 83 Selvakumar Ulaganathan, Ivo Couckuyt, Tom Dhaene, Joris Degroote, and Eric Laermans. High dimensional kriging metamodeling utilising gradient information. *Applied Mathematical Modelling*, 40(9-10):5256–5270, 2016.
- 84 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- 85 Ke Wang, Geoff Pleiss, Jacob Gardner, Stephen Tyree, Kilian Q Weinberger, and Andrew Gordon Wilson. Exact gaussian processes on a million data points. *Advances in neural information processing systems*, 32, 2019.
- 86 Florian Wenzel, Kevin Roth, Bastiaan S. Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? In *Proceedings of the 37th International Conference on Machine Learning, ICML’20*. JMLR.org, 2020.
- 87 Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.
- 88 Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1067–1075, 2013.
- 89 Anqi Wu, Mikio C Aoi, and Jonathan W Pillow. Exploiting gradients and hessians in bayesian optimization and bayesian quadrature. *arXiv preprint arXiv:1704.00060*, 2017.
- 90 Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2023.
- 91 Rosen Ting-Ying Yu, Cyril Picard, and Faez Ahmed. Git-bo: High-dimensional bayesian optimization with tabular foundation models. *arXiv preprint arXiv:2505.20685*, 2025.
- 92 Ning Zhang and Daniel W Apley. Brownian integrated covariance functions for gaussian process modeling: Sigmoidal versus localized basis functions. *Journal of the American Statistical Association*, 111(515):1182–1195, 2016.
- 93 Yichi Zhang, Siyu Tao, Wei Chen, and Daniel W Apley. A latent variable approach to gaussian process modeling with qualitative and quantitative factors. *Technometrics*, 62(3):291–302, 2020.
- 94 Qiang Zhou, Peter ZG Qian, and Shiyu Zhou. A simple approach to emulation for computer models with qualitative and quantitative factors. *Technometrics*, 53(3), 2011.

5.5 GenAI and Bayesian Optimization

Raul Astudillo (California Institute of Technology – Pasadena, US)

Gautam Dasarathy (Arizona State University – Tempe, US)

Peter Frazier (Cornell University – Ithaca, US)

Daolang Huang (Aalto University, FI)

Bryan Kian Hsiang Low (National University of Singapore, SG)

Giulia Pedrielli (Arizona State University – Tempe, US)

Matthias Poloczek (Amazon.com – San Francisco, US)

Elena Raponi (Leiden University, NL)

License © Creative Commons BY 4.0 International license

© Raul Astudillo, Gautam Dasarathy, Peter Frazier, Daolang Huang, Bryan Kian Hsiang Low, Giulia Pedrielli, Matthias Poloczek, and Elena Raponi

5.5.1 Overview

Working Group 5 explored emerging intersections between generative AI (GenAI) and Bayesian optimization (BO). Our discussions converged on four overarching themes:

1. **LLM-based synthetic task generation** to improve meta-BO, mitigate prior misspecification, and broaden the distribution of training tasks (including connections to probabilistic numerics and PFNs).
2. **Bayesian optimization across the LLM training and inference pipeline**, including data mixture optimization, model merging, and inference-time orchestration.
3. **BO guided by generative models** (e.g., diffusion models for protein design and safety verifiers), using learned priors to constrain search to plausible regions while enabling effective exploration.
4. **Rich text-based human feedback for BO**, where LLMs allow algorithms to interpret semantically rich user feedback, greatly increasing information per human interaction.

Together, these directions point toward a long-term vision of a *task-agnostic neural optimization engine*: a foundation model for optimization that can be deployed across scientific and engineering domains, guided by symbolic structure, generative priors, and human intent.

5.5.2 Research Ideas

5.5.2.1 LLM-Based Synthetic Task Generation (for Meta-BO and PFNs)

Contributors

- First-draft writers: Giulia Pedrielli, Daolang Huang
- Other interested parties: Elena Raponi, Raul Astudillo

Motivation. Meta-Bayesian optimization (meta-BO) has shown impressive gains in sample efficiency by learning reusable optimization policies across tasks. However, its success hinges on the quality and breadth of the pre-training task distribution. Most prior work (e.g., NAP, PFN v1, ACE) synthesizes tasks from Gaussian process (GP) priors or structural causal models (SCMs). While convenient, these generators induce two key practical failure modes:

1. **Prior misspecification.** Many real optimization problems differ substantially from GP/SCM assumptions, causing meta-policies to overfit the synthetic pre-training distribution and degrade at deployment.

2. **Real-data scarcity.** In domains such as protein or material design, collecting diverse, well-labeled optimization traces is expensive, making it difficult to correct misspecification with sufficient supervision.

This motivates a principled way to systematically generate better-matched pre-training data. Our idea is to leverage LLMs to generate task priors and simulators, broadening the support of synthetic tasks beyond GP/SCM and targeting high-information training signals aligned with downstream objectives.

Problem Formulation. We aim to define a generative-symbolic pipeline that produces a distribution over optimization tasks

$$T = (\mathcal{X}, f, c, \theta_T),$$

where tasks differ in domain, structure, constraints, and parameters but expose a common query interface. A meta-BO policy

$$\pi_\theta : \mathcal{H}_n \rightarrow \mathcal{X}$$

maps a history of queries and outcomes to the next query point. The training objective is to minimize distributional regret:

$$\min_{\theta} \mathbb{E}_{T \sim \mathcal{T}} [R(\pi_\theta; T)].$$

5.5.2.2 Main Challenges

Symbolic task generation. We identified three macro areas as especially relevant:

1. **Compositional and operator-based modeling.** Symbolic and graph-based equation modeling, code generation, and scientific machine learning (SciML) offer formalisms for composing simulators from reusable operators. Probabilistic programs and category-theoretic approaches provide rigorous semantics for composing open systems as “boxes” with typed ports. These can underpin an operator algebra for task generation. Key challenges include:
 - designing operators with sufficient expressivity while avoiding computational intractability,
 - handling heterogeneous aspects and domains within a common representation.
2. **Equation discovery.** Methods such as SINDy and its variants (with control, boundary-value problems, etc.) discover sparse governing equations from trajectories by selecting terms from a dictionary of candidate operators. These equations can be turned into simulation code directly. Symbolic regression methods (e.g., AI-Feynman, PySR) discover closed-form expressions, with recent work incorporating units and dimensional constraints to maintain physical validity. Such approaches yield human-readable right-hand sides that can be integrated into simulation frameworks like Modelica.
3. **Domain-Specific Language (DSL) synthesis.** DSLs allow LLMs to operate over *operators* (parameterizable objects) rather than raw files, making search and verification easier. Systems such as DreamCoder, which learn and refine a DSL of primitives while solving tasks, exemplify this paradigm. This aligns well with the idea of “teaching the model an operator algebra, then synthesizing programs that compose them.”

Meta-BO. Assuming that a symbolic/LLM pipeline can provide a stream of verified simulators and task specifications, the next challenge is to learn meta-BO policies that systematically exploit this rich task distribution:

- **Robust training paradigms.** Pure behavior cloning from expert BO traces is stable but capped by the expert’s biases. Pure RL on regret or feasibility objectives is well aligned but unstable and expensive. A key challenge is to combine them: pre-train on high-quality trajectories and then fine-tune with regret-, feasibility-, or level-set-aligned objectives.
- **Architectures for heterogeneous histories.** The policy must handle variable input dimensions, constraints, multi-objective settings, and tasks where success includes not only finding the global optimum but also locating feasible points, reaching target level-sets, or covering a Pareto front.
- **Alternative utilities beyond the optimal value.** Practitioners often care about feasibility rate, time-to-feasibility, level-set coverage, or cost-aware regret. The challenge is to define unified, learnable utilities and surrogates that make meta-training stable while reflecting these practical objectives.
- **Robustness to generator imperfections.** Even with verification, the task generator may be biased. The meta-BO model should not exploit artifacts of the generator but instead learn algorithms that transfer to real-world problems.

Pipeline. At a high level, the system consists of three main stages:

1. **Symbolic proposal via LLMs.** LLMs propose task specifications, simulators, and operator compositions within a DSL.
2. **Verification and compilation.** Candidate simulators are checked for consistency (e.g., units, dimensions), tested via automatic unit tests, and compiled into executable form.
3. **Meta-BO training.** A meta-BO policy π_θ is trained on tasks sampled from the verified generator using a combination of imitation learning and utility-aligned optimization to produce decision rules that generalize across tasks.
4. **(Optional) Co-learning.** An additional direction is to close the loop between the task generator and the meta-BO policy. Instead of treating the LLM-based generator as fixed, we can use the performance of the learned policy as feedback to refine the task distribution: identify failure modes, adjust prompts, modify operator choices or parameter ranges, and encourage more diverse and informative tasks.

Outlook: Toward a Task-Agnostic Neural Optimization Engine. Ultimately, the vision underlying this line of work is a task-agnostic neural optimization engine: a foundation model for optimization that can be deployed across domains with minimal adaptation. Given a standardized description of a new task, the model should rapidly propose effective queries without rebuilding a bespoke BO pipeline.

Realizing this vision raises several additional challenges, including:

- defining universal task representations that capture symbolic, statistical, and geometric structure,
- aligning training task distributions with real-world variability,
- transferring operator-based semantics across scientific domains,
- providing safety, interpretability, and robustness guarantees for generated tasks and learned policies.

5.5.2.3 Bayesian Optimization for the LLM Training and Inference Pipeline

Contributors

- First-draft writers: Bryan Low, Gautam Dasarathy
- Other interested parties: Peter Frazier

Motivation. Bayesian optimization (BO) and bandit methods offer a unified lens on several high-impact but expensive decision problems in the large language model (LLM) pipeline:

1. selecting data mixtures for training and adaptation,
2. merging or interpolating pretrained models,
3. directing agentic workflows or routing decisions at inference time.

Each problem involves a costly, noisy black-box objective, a structured decision space, and opportunities for exploiting multi-fidelity or derivative-like information.

Optimizing Data Mixtures. Data composition determines how training samples from different domains are combined under a fixed training protocol. BO constructs a surrogate model of validation performance based on a small number of observed mixtures and selects new candidates via acquisition functions such as expected improvement or knowledge gradient. Multi-fidelity BO integrates cheaper proxies – smaller models, truncated runs, or reduced datasets – with occasional full evaluations.

Recent work suggests that performance as a function of mixture proportions often obeys predictable *data mixing laws* that transfer across scales, enabling nested optimization where low-fidelity surrogates guide higher-fidelity decisions. Derivative-like signals (e.g., sensitivity of validation loss to marginal increases in data buckets) could further accelerate convergence. While full pretraining remains too costly to explore, adaptation and post-training reweighting provide practical arenas for BO-driven mixture optimization.

As an example, [4] introduces DUET, a global-to-local algorithm that combines influence functions for data selection with Bayesian optimization to iteratively refine data mixtures based on feedback from an unseen task. It provides both theoretical convergence guarantees and empirical improvements over competing methods.

Optimizing Model Mixtures. Model mixing or merging seeks to combine the capabilities of separately trained LLMs – whether checkpoints, domain specialists, or fine-tuned variants – into a single unified model. When models share architecture and tokenization, parameters can often be interpolated or aligned directly; when they differ, merging typically proceeds through ensembling or distillation rather than literal weight fusion.

Given a set of models and merge parameters (e.g., interpolation coefficients), the performance of the merged model defines another complex black-box optimization landscape. BO can efficiently explore low-dimensional merge spaces using multi-fidelity evaluations such as limited benchmarks or reduced-context inference. Incorporating similarity features between models, such as representation overlap or calibration distance, into the surrogate may improve sample efficiency. Future directions include BO-guided tuning of structured merges (mixtures-of-experts, layerwise blends) and developing reliable low-cost proxies for alignment or robustness metrics.

Optimizing Agentic Workflows and Routing Policies. As language models are increasingly embedded in multi-step or agentic workflows, a key challenge is deciding which model, prompt, or tool to invoke at each stage. These routing or orchestration policies must balance accuracy, latency, and cost, often under uncertainty about user intent or task difficulty. Evaluating a routing decision typically requires running full interactions or multi-turn rollouts, which are expensive and noisy.

Bayesian optimization can guide the offline design of such policies by exploring high-level configuration choices – such as thresholds, model assignments, or prompt templates – using cheaper approximations like partial rollouts, simulated environments, or reduced-context prompts. Once deployed, online bandit algorithms can adapt these decisions based on real feedback, allocating traffic toward more promising routes while maintaining safety constraints. Together, offline BO for global structure and online bandits for continual adaptation offer a practical path toward efficient and robust agentic control.

As an example of a related perspective, [4] also introduces A-BAD-BO, a BO-based algorithm for optimizing complex AI systems (“AutoAI”) by leveraging local loss signals from system components.

Cross-Cutting Themes. The three settings we discussed – data mixing, model merging, and routing – share a common structure. Each involves a search over a large and expensive design space where direct experimentation is costly. In every case, there exist cheaper or approximate sources of information that can guide this search, such as small-scale models, truncated training runs, or limited-context evaluations. These can be organized into a hierarchy of fidelities that BO is naturally equipped to exploit.

Additional signals can also be incorporated into this process. Examples include estimated sensitivities of validation loss to different data sources, measures of similarity between models before merging, and statistics collected from partial rollouts during inference. These signals provide side information that can shape the surrogate model used by BO, improving both efficiency and robustness.

Although these problems are usually treated independently, they are in fact coupled. The data mixture chosen during training affects the diversity and quality of models available for merging. The merged model influences how routing or orchestration policies should behave at inference time. In the opposite direction, observations from deployment can identify where the model’s behavior is lacking, which in turn informs how future training data or compute should be allocated.

Viewing the entire training and inference pipeline as a single, hierarchical optimization process makes these dependencies explicit. Information collected at one stage can serve as prior knowledge at another, reducing redundant experimentation. Surrogate models built for data selection may inform later stages such as model merging or routing. This integrated perspective does not yet have a practical implementation, but it points toward a coherent optimization framework in which choices about data, model parameters, and inference strategies are treated as parts of one evolving system.

Outlook. BO’s strength lies in sample-efficient search under extreme cost asymmetries. Near-term gains are likely in post-training data reweighting and model merging, where iteration is feasible. Longer-term opportunities include multi-fidelity theory for scaling transfer, gradient-informed surrogates, and unified optimization frameworks that bridge training, adaptation, and inference.

5.5.2.4 BO with Generative Priors for Protein Design, Safety Verifiers, and Beyond

Contributors

- First-draft writers: Raul Astudillo, Elena Raponi
- Other interested parties: Peter Frazier, Gautam Dasarathy

Context. Recent work on steering generative models with experimental data for protein fitness optimization includes, for example, [13] and [14]. These and related papers illustrate how diffusion and other generative models can be used as learned priors in scientific design problems.

Problem Setting. Generative models capture rich, high-dimensional structure from large datasets. In domains such as protein design, they learn distributions over biologically valid sequences, implicitly encoding physical, evolutionary, and functional constraints.

Bayesian optimization (BO), by contrast, is a principled framework for optimizing expensive black-box objectives, balancing exploration and exploitation through probabilistic modeling. Yet, BO lacks a natural mechanism to incorporate the powerful prior knowledge encoded by generative models. This gap motivates a central question: *How can learned generative priors guide Bayesian optimization to search more efficiently while remaining faithful to data-driven structure?*

Let $\mathbb{X}$ denote the input space and $f : \mathbb{X} \rightarrow \mathbb{R}$ the objective function to be optimized. We assume access to a generative model that defines a prior distribution $\pi_0(x)$, highlighting regions of $\mathbb{X}$ corresponding to plausible or desirable configurations. Our goal is to identify evaluation points that maximize f while effectively leveraging the structural information encoded in $\pi_0(x)$.

Relation to Prior Work. Given data $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, let $\alpha_n : \mathbb{X} \rightarrow \mathbb{R}$ denote a user-defined acquisition function. Prior work has explored various approaches to integrate the inductive bias encoded by π_0 with the predictive guidance provided by α_n across different settings.

Hvarfner et al. introduced π -BO [6], scaling the acquisition function by a user-specified prior $\pi(x)$ to encode expert beliefs:

$$\tilde{\alpha}_n(x) = \pi(x) \alpha_n(x).$$

Yang et al. [13] explored a related idea in the context of priors encoded by generative models such as diffusion models, yielding a sampling policy

$$x_{n+1} \sim \pi_{n+1} \propto \pi_0(x) \exp\left(\frac{\alpha_n(x)}{\beta}\right), \quad (19)$$

where β balances exploration and adherence to the generative prior.

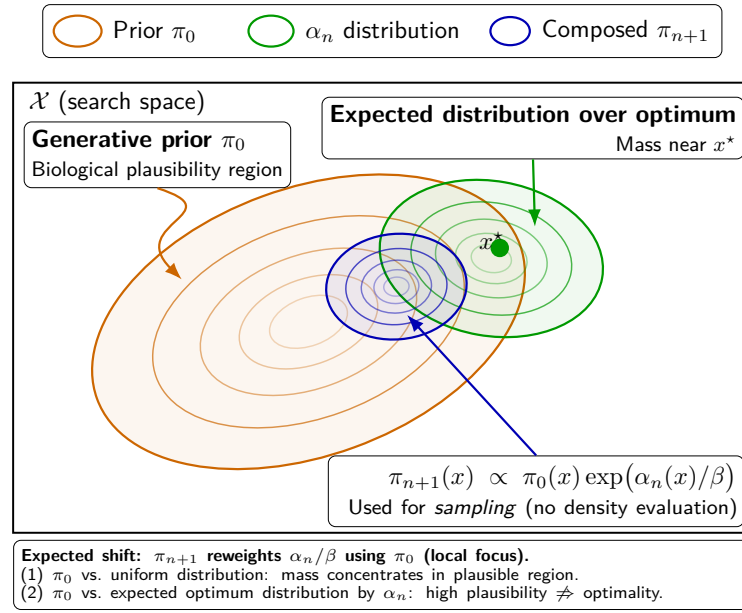
Opportunities for Exploration

Coupled vs. decoupled formulations. Most existing approaches combine the generative prior π_0 and acquisition function α_n post hoc. A key question is whether these components should instead be co-learned or co-adapted – for example, through conditional parameterization.

Prior alignment and bias control. Generative priors capture *plausibility*, not necessarily *optimality*. How can we align or temper π_0 to balance structural validity with performance? Adaptive temperature schedules, mixture priors, or meta-learned weighting schemes may help prevent over-constraining the search while retaining useful bias.

Generative acquisition functions. Rather than using a fixed acquisition function, could a generative model learn to *represent and sample from* the acquisition landscape itself? This raises the possibility of *generative acquisition functions* – models that internalize both uncertainty and prior structure, potentially bridging BO and reinforcement learning.

Theoretical understanding. What are the convergence, regret, or consistency guarantees for sampling-based policies such as Eq. (19)? Can they be formalized as entropy-regularized optimization, energy-based inference, or constrained search in distribution space? Clarifying these connections could ground new hybrid methods in a principled theoretical framework.



■ **Figure 9** Conceptual overview: Bayesian optimization guided by a generative prior π_0 learned from data. The generative model constrains exploration to biologically plausible regions of the search space, while the acquisition function $\alpha_n(x)$ drives optimization toward higher predicted performance.

5.5.2.5 Rich Text-Based Human Feedback

Contributors

- First-draft writers: Matthias Poloczek, Peter Frazier
- Other interested parties: Raul Astudillo, Elena Raponi, Matthias Feurer, Antonio, Gautam Dasarathy, Daolang Huang

Introduction

There is a long history of work on BO and human interaction. This includes:

- **BO for preference learning**, e.g., see [5, 10, 2, 8, 9, 11]. Here, our objective function in a BO problem is a human decision-maker's utility function, which the human cannot specify as a mathematical expression. Instead, we interrogate this utility via preference queries.
- **BO with human-in-the-loop feasibility assessment**, e.g., [1, 12], where a human expert reviews inputs proposed by a BO algorithm and may label them infeasible before they are evaluated.
- **Elicitation of informative priors**, where human expertise is translated into prior distributions through structured questioning (e.g., quantiles, ranges, or plausible values).

Traditionally, human–algorithm interactions in BO have a simple form where the human provides only a few bits of feedback per interaction. For example, in BO for preference learning, the most common interaction is a pairwise comparison: the user is presented with two items and asked which one they prefer. In feasibility assessment, the human often provides a single bit of information (feasible or infeasible). In prior elicitation, humans answer structured questions that are then converted into prior distributions.

Large language models (LLMs) present an opportunity to interpret *richer text-based feedback from the human* and incorporate this into our algorithms. This promises many more bits per interaction, potentially enabling much more sample-efficient optimization.

Novel Challenges

Effective use of rich text-based human feedback raises novel challenges. Many of these can be categorized into questions about inference and questions about experiment and interaction design.

Inference.

- How do we interpret text-based human feedback?
- How do we leverage LLMs in this interpretation while maintaining probabilistic coherence?
- How do we combine information from LLMs with traditional probabilistic models (e.g., Gaussian processes or other surrogates)?
- How do we deal with variability, bias, and lack of reliability in LLM outputs?

Interaction and experiment design.

- **Form of interaction.** What forms of interaction are most effective? For example, in preferential BO, we could ask:
 - “Tell me what you are looking for in a solution.”
 - “Which solution attributes are most important?”
 - “Compare solutions A and B. Tell me which one you prefer and why.”
 - “Please improve solution A by making a small change and tell me why it is better.”
 Effectiveness encompasses useful bits per interaction, ease of interpretation, and cognitive load for the human.
- **Question selection.** What is the best way to choose questions to ask the human? Should we outsource this to an LLM, rely on traditional acquisition functions, or combine both?
- **Resource allocation.** In light of this new way of getting feedback, how do we balance effort spent on querying the expensive black box with effort spent on interacting with the human?
- **Supporting human introspection.** Humans use introspection to explore their own preferences as they consider options in multi-objective problems. An ideal algorithm would clearly identify key tradeoffs and present them in a way that helps the human focus their introspective efforts.
- **Evaluation.** How do we evaluate such systems? Possible directions include:
 - using an LLM as a judge in simulated user studies,
 - online studies with real user traffic,
 - offline evaluation on comprehensive user datasets where multiple forms of feedback are available.

Related Work

- Max: <https://arxiv.org/abs/2510.17671>
- Peter: <https://arxiv.org/abs/2510.25799>
- Papers identified by Matthias: [7, 3]

[7] considers engineering tasks where the feasible space does not have an explicit mathematical form and proposes eliciting human prior knowledge in natural language to learn the feasible set for BO.

[3] proposes a human–AI collaborated experimental workflow, “Bayesian Optimized Active Recommender System” (BOARS), where the user can dynamically change the target objective on the fly.

More Specific Discussion of Preferential BO

Formal problem description. Formally, we have a feasible set $\mathbb{X}$, a computationally expensive black box that computes a collection of metrics $y = f(x)$ for any $x \in \mathbb{X}$, and an unknown user utility function $u(y)$. We seek to optimize $u(y)$ subject to the expense of evaluations and the constraint that u can only be interrogated through human feedback, potentially in rich textual form.

Important problem characteristics.

- **Dimensionality of y .** Is y low-dimensional (e.g., a small set of KPIs) or high-dimensional (e.g., images, videos, or long-form text)? The nature of y profoundly affects the form and cost of human feedback.
- **Need for physical instantiation.** Do we need to physically instantiate x to observe y and thus obtain feedback? For example, in materials design, x might be a composition, and y could involve physical properties that require lab experiments.

Connections to Other Areas. Connections to work on LLM alignment and assurance are natural: in both settings, one seeks to interpret human feedback and use it to guide system behavior. However, in our setting we typically do not modify the LLM itself; rather, we use the LLM as a tool to interpret feedback and support BO.

BO for Prompt Optimization

- Names from shared notes: Elena, Matthias P., Matthias Feurer, Antonio

This topic was identified as an important direction but was not developed during the seminar. We anticipate that many of the ideas above – LLM-based task generation, generative priors, and rich text-based human feedback – will play a role in BO for prompt and prompt-pipeline optimization. We leave a detailed treatment of this topic for future work.

References

- 1 Arun Kumar A V, Santu Rana, Alistair Shilton, and Svetha Venkatesh. Human–AI collaborative Bayesian optimisation. In *Advances in Neural Information Processing Systems*, pages 16233–16245, 2022.
- 2 Raul Astudillo and Peter Frazier. Multi-attribute Bayesian optimization with interactive preference learning. In *International Conference on Artificial Intelligence and Statistics*, pages 4496–4507, 2020.
- 3 Arpan Biswas, Yongtao Liu, Nicole Creange, Yu-Chen Liu, Stephen Jesse, Jan-Chi Yang, Sergei V Kalinin, Maxim A Ziatdinov, and Rama K Vasudevan. A dynamic Bayesian optimized active recommender system for curiosity-driven partially human-in-the-loop automated experiments. *npj Computational Materials*, 10(1):29, 2024.
- 4 Zhiliang Chen, Gregory Kang Ruey Lau, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. Duet: Optimizing training data mixtures via feedback from unseen evaluation tasks, 2025.
- 5 Wei Chu and Zoubin Ghahramani. Preference learning with Gaussian processes. In *International Conference on Machine Learning*, pages 137–144, 2005.

- 6 Carl Hvarfner, Danny Stoll, Artur Souza, Marius Lindauer, Frank Hutter, and Luigi Nardi. π BO: Augmenting acquisition functions with user beliefs for Bayesian optimization, 2022.
- 7 Cole Jetton, Matthew Campbell, and Christopher Hoyle. Constraining the feasible design space in Bayesian optimization with user feedback. *Journal of Mechanical Design*, 146(4):041703, 2023.
- 8 Zhiyuan Jerry Lin, Raul Astudillo, Peter Frazier, and Eytan Bakshy. Preference exploration for efficient Bayesian optimization with multiple outcomes. In *International Conference on Artificial Intelligence and Statistics*, pages 4235–4258, 2022.
- 9 Ryota Ozaki, Kazuki Ishikawa, Youhei Kanzaki, Shion Takeno, Ichiro Takeuchi, and Masayuki Karasuyama. Multi-objective Bayesian optimization with active preference learning. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 14490–14498, 2024.
- 10 Juan Ungredda and Juergen Branke. When to elicit preferences in multi-objective Bayesian optimization. In *Genetic and Evolutionary Computation Conference Companion Proceedings*, pages 1997–2003, 2023.
- 11 Hanyang Wang, Juergen Branke, and Matthias Poloczek. Bayesian optimization with preference exploration by monotonic neural network ensemble. In *Advances in Neural Information Processing Systems*, 2025.
- 12 Wenjie Xu, Masaki Adachi, Colin N Jones, and Michael A Osborne. Principled Bayesian optimization in collaboration with human experts. *Advances in Neural Information Processing Systems*, 37:104091–104137, 2024.
- 13 Jason Yang, Wenda Chu, Daniel Khalil, Raul Astudillo, Bruce James Wittmann, Frances H. Arnold, and Yisong Yue. Steering generative models with experimental data for protein fitness optimization. In *Advances in Neural Information Processing Systems*, 2025.
- 14 Taeyoung Yun, Kiyoung Om, Jaewoo Lee, Sujin Yun, and Jinkyoo Park. Posterior inference with diffusion models for high-dimensional black-box optimization. In *Forty-second International Conference on Machine Learning*, 2025.

6 Additional Content

6.1 Benchmarking Bayesian Optimization in Practice: Lessons from a Dagstuhl Seminar Exercise

Giulia Pedrielli (Arizona State University – Tempe, US)

License © Creative Commons BY 4.0 International license
© Giulia Pedrielli

This report summarizes a hands-on benchmarking exercise conducted during the Dagstuhl Seminar 25451 (November 2025), organized with the explicit goal of fostering discussion around practical aspects of Bayesian Optimization (BO). Rather than comparing abstract algorithmic formulations, the exercise was designed to surface the “tricks of the trade”: implementation details and design choices that materially affect performance but are often under-specified or abstracted away in the literature.

The intent of this report is not to establish winners or propose a standardized benchmark, but to document observations that emerged from systematically varying algorithmic choices under controlled experimental conditions, and to articulate reflections relevant to both researchers and practitioners.

6.1.1 Scope of the Exercise and Experimental Setup

The exercise focused on single-objective, black-box optimization under finite evaluation budgets. Algorithms were evaluated on a small suite of well-known continuous test functions of varying dimensionality and landscape characteristics, including Ackley (100D), Rosenbrock (6D), and Michalewicz (10D). Budgets were deliberately limited to emphasize sample efficiency rather than asymptotic convergence.

6.1.1.1 Algorithm Characterization

A central design principle of the exercise was to characterize BO algorithms not as monolithic entities, but as compositions of interacting design choices, including:

- Surrogate model class (e.g., Gaussian process variants)
- Kernel choice and hyperparameter handling
- Acquisition function family
- Acquisition optimization strategy
- Initialization strategy (number and type of initial design points)
- Budget allocation policy (initial design vs adaptive phase)
- Noise modeling assumptions
- Restart or trust-region mechanisms

This decomposition allowed participants to reason about why performance differed, rather than merely whether it differed.

6.1.2 Performance Metrics and Visualization

Algorithmic choices were observed under multiple complementary views:

1. Sample efficiency curves, showing best-so-far objective value as a function of evaluations.
2. Rank-based comparisons, aggregating relative performance across functions.
3. Budget sensitivity plots, highlighting the impact of how evaluation budgets are allocated between initialization and adaptive optimization.

The use of rank-based summaries was motivated by the observation that raw objective values are often not comparable across problems or scales, while ranks more clearly expose robustness and consistency.

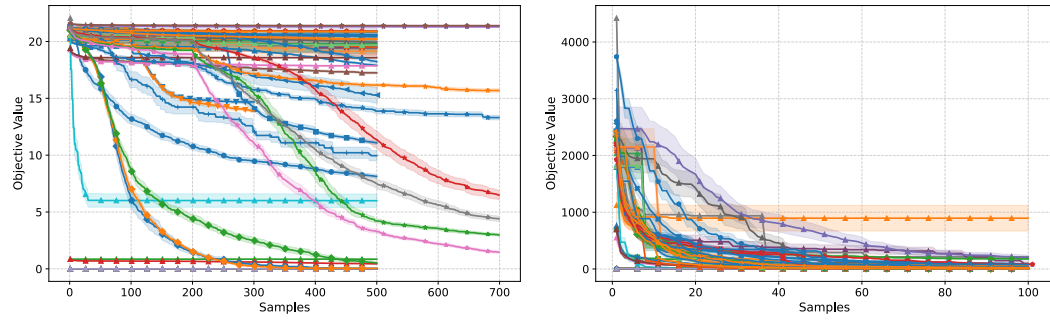
6.1.3 Empirical Results

6.1.3.1 Sample Efficiency

Across all test functions, we observed substantial variation in early-stage performance, even among algorithms that share nominally similar formulations. Figures illustrating convergence and sampled values (e.g., Ackley 100D and Rosenbrock 6D) show that initialization strategy and early hyperparameter behavior often dominate performance in the first 20–40% of the budget.

6.1.3.2 Rank-Based Comparisons

The overall average rank plot reveals that algorithm rankings are often compressed and unstable: small differences in setup can shift relative positions, and no single configuration consistently dominates across all problems. Aggregated rank statistics indicate that several algorithmic configurations perform comparably on average, with performance differences



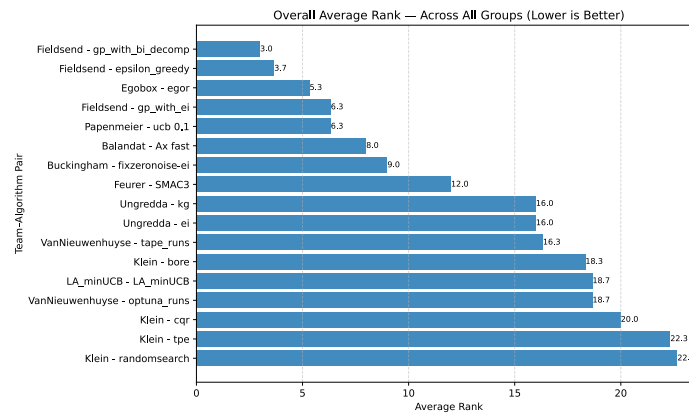
(a) Ackley (100D)

(b) Rosenbrock (6D)



(c) Plots legend.

Figure 10 Sample efficiency of representative Bayesian optimization configurations. Plots report the best objective value observed so far as a function of the evaluation budget. Differences in early-stage behavior persist throughout the optimization horizon, highlighting the impact of initialization and surrogate configuration choices.



■ **Figure 11** Average rank calculated for each approach across test functions.

frequently within statistical noise for the given budgets. This observation reinforces the notion that headline rankings without detailed configuration context can be misleading.

6.1.3.3 Budget Allocation Effects

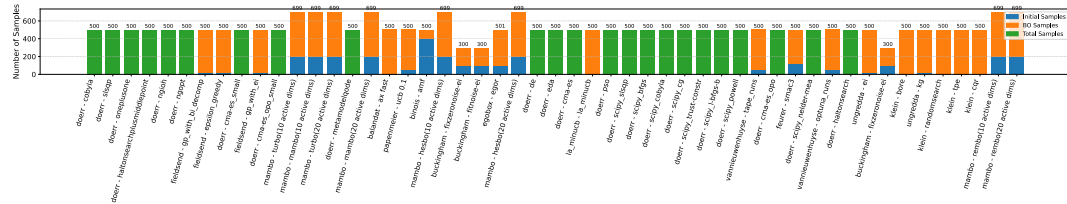
Explicit plots separating initial sampling from adaptive optimization show that how the budget is spent can matter as much as which acquisition function is used. Over-investment in initial design can delay exploitation, while overly aggressive early exploitation can lead to premature convergence. Sample allocation plots illustrate that moderate, problem-aware initialization strategies tend to offer better robustness across test functions.

6.1.4 Reflections on Algorithmic Choices

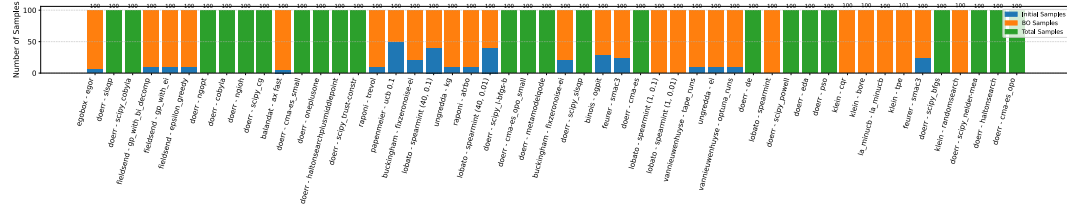
Several consistent themes emerged from the exercise and subsequent discussion:

- *Initialization Is Not a Detail:* The number, placement, and diversity of initial samples strongly influence surrogate fit and acquisition behavior. In high dimensions, naive space-filling designs can consume disproportionate budget with limited benefit.
- *Hyperparameter Handling Drives Early Behavior:* Surrogate hyperparameters, especially kernel length-scales, often dominate acquisition decisions early on. Aggressive re-estimation can destabilize optimization, while overly conservative strategies may delay adaptation.
- *Acquisition Optimization Quality Matters:* Differences in acquisition maximization (e.g., local vs global optimization, restart strategies) can outweigh differences between acquisition functions themselves, particularly under small budgets.
- *Budget Allocation Encodes Implicit Assumptions:* Choices about when to transition from exploration to exploitation reflect implicit beliefs about problem smoothness and noise. Making these assumptions explicit improves both interpretability and reproducibility.
- *More complex algorithm characterization may be needed:* it is a possibility that the lack of separation between algorithms partially lies in the choice of algorithm parameters. In other words, other working mechanisms of the methods may be required to understand the behaviors (e.g., use of a single/multiple surrogates, optimizer for the acquisition function, use of projections/embeddings).

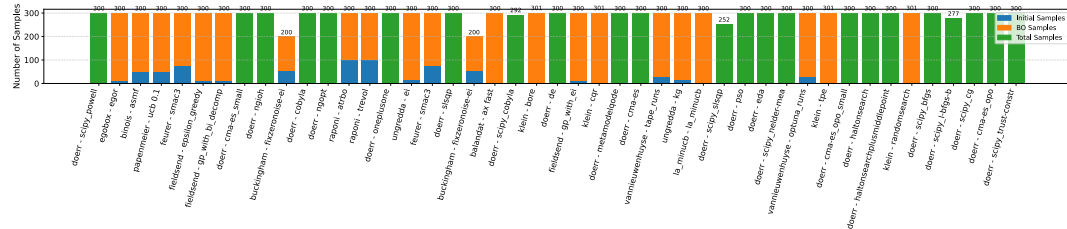
As an example of notable aspects that rose from the discussions during the seminar, the use of multiple surrogates dynamically and adaptively revealed to be potentially fruitful, on the



(a) Ackley (100D)



(b) Rosenbrock (6D)



(c) Michalewicz (10D)

■ **Figure 12** Budget sensitivity.

other hand, strategies such as random projections revealed their lack of robustness in that they showed good performance but only due to the fact that the 100D function chosen was symmetric, and the same performance was not observed across benchmarks.

6.1.5 Concluding Remarks

This exercise highlights that practical BO performance is shaped less by isolated algorithmic components and more by the interaction of design choices under finite budgets. The results reinforce the need for benchmarking practices that foreground configuration metadata and encourage reflective analysis over leaderboard-style comparisons.

Rather than advocating for a particular “best” algorithm, the Dagstuhl exercise underscores the value of transparency, careful experimental design, and explicit acknowledgment of the practical decisions that ultimately govern performance.

6.2 Bayesian Optimization Tricks Of The Trade

Jürgen Branke (*University of Warwick, GB*)

Giulia Pedrielli (*Arizona State University – Tempe, US*)

Matthias Poloczek (*Amazon.com – San Francisco, US*)

License  Creative Commons BY 4.0 International license
© Jürgen Branke, Giulia Pedrielli, and Matthias Poloczek

6.2.1 Introduction

At the Dagstuhl Seminar 25451 on Bayesian Optimization, we asked all participants to fill in a questionnaire and tell us their tricks of the trade. The questions we asked were:

1. What did you wish someone had told you before you started working in BO?
2. What's something you usually do but never see written in papers?
3. What's a mistake you made once and vowed never to repeat?
4. When reviewing others' work, what silent red flags tell you they're beginners?

In addition, we have asked chatGPT 5 to answer the same questions. The following summarizes the responses we received.

6.2.2 What I wish I had known earlier

- Always check and tune the model before doing optimisation.
- The details of implementation matter much more than expected (HP optimisation, AF optimisation).
- Having a solid foundation in GP bandits helps enormously.
- You don't have to be a strict Bayesian to do good BO.
- Try different kernels and trend functions; default choices rarely work best. If you go with one kernel, Matérn 5/2 seems a generally accepted choice.
- Nugget effect (modelling a small constant noise) improves stability in deterministic problems.
- Random axis-aligned subspace perturbations (RAASP) which samples around the current optimal point along which differ only in a random subset of dimensions or cylindrical Thompson sampling are good ways for candidate sampling especially in higher dimensions.
- Thompson sampling works when other AFs fail, probably because it is much easier to optimise than other AFs.
- Always start with smaller tests before scaling up.
- Establish clear baselines, such as random search, CMA-ES or vanilla BO.
- Save **EVERYTHING** you can, acqf values, commits, number of initial points, lengthscales, etc. Saves a lot of debugging time.
- Model mismatch and AF optimization can have a huge impact on results. Therefore one should always also test on a GP generated test problem on a discrete set where the AF can be optimized via complete enumeration.
- Setting the prior mean is application-dependent, often a constant, sometimes parametrized, for LLM even a scaling law.

6.2.3 Unwritten practices

- Scale all inputs to the unit hypercube
- Test on simple 2D problems first that allow visualization.
- Standardise objectives on the fly during multi-objective optimisation.

- Use multi-start when fitting the GP.
- Warm-start the model fit from the previous solution.
- Put lower bounds on kernel length scales to prevent numerical issues.
- Use logEI rather than EI.
- Smart points to optimize the acquisition function help a lot. Best posterior mean location, location of the best sampled location.
- Validate performance on a test set, not just on benchmarks used for deciding BO hyperparameters.

6.2.4 Mistakes to avoid

- Forgetting to update the reference point for expected hypervolume improvement calculation in multi-objective optimisation.
- Using unstable implementations of BFGS or other optimisers.
- Forgetting to check the GP fit.
- Failing to rescale data before fitting models.
- Starting benchmarking before checking that the code is correct.
- Not saving data, forcing repeated and inconsistent re-runs.
- Evaluating on too few problems or only on familiar ones.

6.2.5 Beginner red flags

- No error bars or uncertainty measures in reported results.
- Too many arbitrary “magic” constants in experiments.
- Only one example tested, with no justification.
- Misuse of terminology (e.g., calling BO “EGO”).
- Wrong or missing citations for well-known work.
- Re-inventing already existing methods.
- Use squared exponential kernel without justification
- Irrelevant details swamping main explanations (move details to appendix).
- Experimental setups not clearly described.
- No ablation study on hyperparameters, or no explanation on how they were chosen.
- AF and HP optimization process not explained.
- Mixing up GP and acquisition-function concepts.
- Lack of replication or reproducibility checks.
- No scientific hypothesis that is clearly defined and rejected/not rejected at the end.
- Using an unsuitable method as baseline for the empirical evaluation.

6.2.6 Other good practices

- Report, not just validate, the performance on unseen test tasks.
- Accessible codebase in addition to high-level description of the code in the paper.

Participants

- Steven Adriaensen
Universität Freiburg, DE
- Mauricio A Álvarez
University of Manchester, GB
- Cedric Archambeau
Helsing – Berlin, DE
- Raul Astudillo
California Institute of Technology
– Pasadena, US
- Maximilian Balandat
Meta – Menlo Park, US
- Nathalie Bartoli
ONERA – Toulouse, FR
- Thomas Bartz-Beielstein
TH Köln, DE
- Mickaël Binois
Inria Center at Université Côte
d'Azur – Sophia Antipolis, FR
- Jürgen Branke
University of Warwick, GB
- Jack Buckingham
University of Warwick –
Coventry, GB
- Antonio Candelieri
University of Milano-Bicocca, IT
- Ivo Couckuyt
Ghent University, BE
- Gautam Dasarathy
Arizona State University –
Tempe, US
- Carola Doerr
Sorbonne University – Paris, FR
- Katharina Eggersperger
TU Dortmund, DE
- Matthias Feurer
TU Dortmund, DE
- Jonathan Fieldsend
University of Exeter, GB
- Peter Frazier
Cornell University – Ithaca, US
- Roman Garnett
Washington University –
St. Louis, US
- Robert Gramacy
Virginia Polytechnic Institute –
Blacksburg, US
- Daniel Hernández-Lobato
Autonomous University of
Madrid, ES
- Daolang Huang
Aalto University, FI
- Frank Hutter
Prior Labs – Freiburg, DE &
ELLIS Institute Tübingen, DE &
Universität Freiburg, DE
- Aaron Klein
ScaDS.AI – Leipzig, DE
- Rodolphe Le Riche
CNRS – Aubière, FR
- Bryan Kian Hsiang Low
National University of
Singapore, SG
- Neeratyoy Mallik
Universität Freiburg, DE
- Mike McCourt
Distributional – Toronto, CA
- Ruth Misener
Imperial College London, GB
- Szu Hui Ng
National University of
Singapore, SG
- Leonard Papenmeier
Universität Münster, DE
- Giulia Pedrielli
Arizona State University –
Tempe, US
- Matthias Poloczek
Amazon.com – San Francisco, US
- Jixiang Qing
Imperial College London, GB
- Elena Raponi
Leiden University, NL
- Sebastian Rojas Gonzalez
Ghent University, BE
- Alexander Terenin
Cornell University – Ithaca, US
- Juan Ungredda
ESTECO SpA – Trieste, IT
- Inneke Van Nieuwenhuysse
Hasselt University –
Diepenbeek, BE



A Roadmap Towards Practical Applications of Neurosymbolic Learning and Reasoning

Thiviyan Thanapalasingam^{*1}, Annette ten Teije^{*2}, and Frank van Harmelen³

1 University of Amsterdam, NL. thiviyan.t@gmail.com

2 VU Amsterdam, NL. annette.ten.teije@vu.nl

3 VU Amsterdam, NL. frank.van.harmelen@vu.nl

Abstract

Neurosymbolic Artificial Intelligence (NeSy) promises to combine the scalability and adaptability of neural networks with the precision and interpretability of logic-based reasoning systems. Despite significant theoretical advances and growing interest in the field, the development of NeSy systems that seamlessly integrate symbolic and sub-symbolic elements at scale for real-world practical applications remains an open challenge. This Dagstuhl Seminar brought together leading researchers from diverse fields including machine learning, logic, knowledge representation, knowledge discovery, robotics, healthcare, and natural sciences to chart a roadmap for advancing NeSy systems from synthetic benchmarks to practical deployment. The seminar addressed four key challenges: (I) developing realistic benchmarks and standardized evaluation frameworks for NeSy systems, (II) identifying and advancing applications in robotics, healthcare, and scientific discovery, (III) creating open-source libraries to lower entry barriers and accelerate development, and (IV) creating accessible educational materials to broaden participation in the field. Through structured discussions, breakout sessions, and collaborative planning, participants developed concrete action plans and deliverables, including a planned Communications of the ACM article, open-source library development (ULLER, DeepLog), educational resources, and improvements to the field's Wikipedia entry. This report documents the seminar's presentations, discussions, and findings, and presents strategic recommendations for realizing the promise of Neurosymbolic AI in real-world applications.

Seminar November 2–7, 2025 – <https://www.dagstuhl.de/25452>

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Computing methodologies → Symbolic and algebraic algorithms; Theory of computation → Logic

Keywords and phrases Neurosymbolic Artificial Intelligence, Machine Learning, Logic, Knowledge Representation & Reasoning, Sub-symbolic & Symbolic Integration

Digital Object Identifier 10.4230/DagRep.15.11.67

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

A Roadmap Towards Practical Applications of Neurosymbolic Learning and Reasoning, *Dagstuhl Reports*, Vol. 15, Issue 11, pp. 67–86

Editors: Thiviyan Thanapalasingam, Annette ten Teije, and Frank van Harmelen



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Thiviyan Thanapalasingam (University of Amsterdam, NL, thiviyan.t@gmail.com)

Annette ten Teije (VU Amsterdam, NL, annette.ten.teije@vu.nl)

Frank van Harmelen (VU Amsterdam, NL, frank.van.harmelen@vu.nl)

License  Creative Commons BY 4.0 International license

© Thiviyan Thanapalasingam, Annette ten Teije, and Frank van Harmelen

This summary presents the outcomes of our Dagstuhl Seminar on “A Roadmap Towards Practical Applications of Neurosymbolic Learning and Reasoning” (25452). The seminar brought together 23 researchers from diverse fields to address the critical challenge of transitioning Neurosymbolic AI (NeSy) systems from synthetic benchmarks to real-world applications. The seminar focused on:

- Identifying key challenges that impede the practical deployment of NeSy systems,
- Developing concrete action plans and collaborative structures to address these challenges,
- Establishing deliverables including educational resources, open-source libraries, and an article in a high-impact journal.

As a result of the seminar, the following challenges have been identified and addressed through structured discussions and working groups:

1. **Benchmarks and Evaluation:** The field’s current reliance on synthetic problems (e.g., MNIST Addition, MNIST Sudoku) fails to capture real-world complexity. Participants developed plans for creating realistic benchmark suites that incorporate incomplete rule sets, scalability requirements, and standardized evaluation metrics beyond traditional machine learning measures, including reasoning correctness, rule violation detection, and interpretability assessment.
2. **Domain Applications:** Three priority application areas were identified (robotics, natural sciences, and healthcare) where NeSy systems can provide substantial value. Working groups outlined concrete use cases while identifying specific technical challenges and highlighting existing applications of NeSy systems from other AI subfields.
3. **Open-Source Infrastructure:** Following the success of tools like PyTorch Geometric in accelerating deep learning adoption, participants discussed with the creators of Pylon and ULLER strategies to lower entry barriers for both researchers and practitioners.
4. **Educational Resources:** To broaden participation and accelerate field growth, participants committed to writing a comprehensive handbook/textbook on NeSy foundations, using machine learning as a foundation, and targeting graduate students and researchers from adjacent fields.

The seminar produced several concrete outcomes and commitments for future collaboration. Participants agreed to co-author a CACM article to raise awareness across the broader scientific community, initiate development of open-source Python libraries for NeSy methods, create a comprehensive benchmark suite with standardized evaluation protocols, and update the Wikipedia entry for Neurosymbolic AI to improve public-facing documentation of the field.

Plenary discussions on “*What would be a good textbook for teaching Neurosymbolic Learning and Reasoning?*” and “*What is the role of Neurosymbolic AI in the era of foundational models?*” produced important insights regarding how large language models both motivate NeSy research by highlighting reasoning limitations and provide new opportunities through neuro-symbolic integration with foundation models.

The seminar’s roadmap positions the NeSy community to address pressing challenges in AI by focusing on practical applications and building supporting infrastructure.

2 Table of Contents

Executive Summary

Thiviyan Thanapalasingam, Annette ten Teije, and Frank van Harmelen 68

Overview of Talks

Neuro-Symbolic AI in HealthCare: Integrating Learning and Reasoning for Trustworthy Healthcare AI	
<i>Maria-Esther Vidal</i>	71
Neurosymbolic AI for Trustworthy Automation: From Datacenters to Agentic Systems	
<i>Byron Cook</i>	71
Toward Derivable Scientific Discovery: Integrating Data, Theory, and the Future of Machine-Aided Discovery	
<i>Cristina Cornelio</i>	72
Tooling for Reproducible Neurosymbolic Research: ULLER	
<i>Emile van Krieken</i>	73
DeepLog: A Theoretical and Operational Framework for Neurosymbolic AI	
<i>Robin Manhaeve</i>	73
Benchmarking Neurosymbolic AI	
<i>Maarten C. Stol</i>	73
Neuro-Symbolic AI and Robotics: Bridging Perception and Sensemaking	
<i>Agnese Chiatti</i>	74
Neurosymbolic AI Education: Integrating Learning and Reasoning	
<i>Luís C. Lamb</i>	74

Working Groups

Benchmarking Neurosymbolic AI (Working Group 1)	
<i>Robert Peharz, Antonio Vergari, Mathias Niepert, Guy Van den Broeck, Benjie Wang, Emile van Krieken, and Annette ten Teije</i>	75
Benchmarking Neurosymbolic AI (Working Group 2)	
<i>Alberto Speranzon, Lia Morra, Maria-Esther Vidal, and Michael Cochez</i>	75
Benchmarking Neurosymbolic AI (Working Group 3)	
<i>Luis Lamb, Guy Van den Broeck, Sebastijan Dumancic, Thiviyan Thanapalasingam, Cristina Cornelio, Maarten C. Stol, and Frank van Harmelen</i>	76
Neuro-Symbolic AI in Health Care	
<i>Robert Peharz, Maria-Esther Vidal, Frank van Harmelen, Annette ten Teije, and Cristina Cornelio</i>	78
Architectures for Probabilistic Neurosymbolic AI	
<i>Robert Peharz, Antonio Vergari, Mathias Niepert, Guy Van den Broeck, Benjie Wang, Emile van Krieken, Luis C. Lamb, Lia Morra, Michael Cochez, and Sebastijan Dumancic</i>	79

Neurosymbolic Methods for Safe and Reliable Robotic Systems <i>Guy Van den Broeck, Alberto Speranzon, Agnese Chiatti, Annette ten Teije, Cristina Cornelio, Thiviyan Thanapalasingham, and Sebastijan Dumancic</i>	80
Updating the Wikipedia Entry for Neurosymbolic AI <i>Emile van Krieken, Robin Manhaeve, and Luis C. Lamb</i>	81
Plenary Discussions	
Tooling for Neurosymbolic AI	81
Bridging NeSy and LLMs	82
Education Materials for NeSy AI	83
Communicating Results	
Planning on joint paper	84
Wikipedia Entry for Neurosymbolic AI	84
Educational Materials	84
Participants	86

3 Overview of Talks

3.1 Neuro-Symbolic AI in HealthCare: Integrating Learning and Reasoning for Trustworthy Healthcare AI

Maria-Esther Vidal (TIB Leibniz-Informationszentrum – Hannover, DE & Leibniz Universität Hannover, DE, vidal@l3s.de)

License  Creative Commons BY 4.0 International license
© Maria-Esther Vidal

Artificial Intelligence is rapidly transforming healthcare, yet its integration into clinical practice remains constrained by concerns of bias, opacity, and lack of accountability. The FUTURE-AI framework defines six guiding principles (Fairness, Universality, Traceability, Usability, Robustness, and Explainability) as prerequisites for trustworthy medical AI. However, these principles are still largely conceptual; they require technical mechanisms that can make them actionable and verifiable within AI pipelines.

This talk examines how Neuro-Symbolic AI (NeSy), which combines neural prediction and symbolic reasoning, can address these challenges by integrating semantics, logic, and ethics directly into the computational core of healthcare AI systems. NeSy unites the statistical power of neural models with the interpretability and formal rigor of symbolic representations. Through this integration, fairness can be achieved by tracing and reasoning over bias; universality and robustness can arise from semantic consistency and invariant knowledge; and traceability and explainability can be realised through explicit reasoning chains and formal provenance models.

The presentation illustrates these ideas through two case studies: (1) ontology-based bias documentation and reasoning in machine learning pipelines, and (2) explainable prediction in lung-cancer knowledge graphs. Both examples demonstrate how symbolic reasoning complements data-driven learning to improve interpretability, reliability, and trust.

The talk concludes by identifying open research directions, formalisms for specification and validation, scalable multimodal reasoning, and human-centred explainability, toward the realisation of ethically grounded, transparent, and trustworthy Neuro-Symbolic AI for healthcare.

3.2 Neurosymbolic AI for Trustworthy Automation: From Datacenters to Agentic Systems

Byron Cook (Amazon Web Services – Seattle, US & University College London, GB, byron.cook@amazon.com)

License  Creative Commons BY 4.0 International license
© Byron Cook

We are witnessing a scientific revolution in computing, driven by the shift from microprocessors to datacenters as the dominant computational model. This transformation has enabled not only generative AI, through access to massive computational resources, but also fundamental advances in formal reasoning tools, which now achieve 10x performance improvements through distributed reasoning and lemma sharing.

This evolution presents a unique opportunity for neurosymbolic approaches: language models can discover inductive invariants, ranking functions, and rely/guarantee constraints, while mechanical theorem provers verify these artifacts with proof-producing SAT/SMT

solvers. The past decade has demonstrated that automated reasoning is no longer confined to hardware and safety-critical systems, cloud providers like AWS now apply it broadly across their infrastructure, from customer-facing tools (IAM Access Analyzer, S3 Block Public Access) to internal proof projects in cryptography, identity, and virtualization.

However, as customers adopt generative and agentic AI, they face critical challenges around safety, truthfulness, and trustworthiness. This talk presents Amazon’s approach through several products: Automated Reasoning checks in Bedrock Guardrails, which combines pre-deployment requirements analysis with inference-time verification; Amazon Kiro, which helps build formal specifications for programs; and Strands and AgentCore, which apply compositional reasoning to agent safety and liveness properties using the Cedar permissions framework.

The presentation demonstrates how neurosymbolic techniques address key challenges, including redundant translation with semantic equivalence checking, iterative refinement, synthesis of axiomatic vocabularies, and specification composition. It explores how these ideas extend to a new style of GenAI-native development, where semantic guardrails in the form of proved properties prevent unwanted bugs in code generated by agentic developer tools.

Finally, the talk addresses the critical importance of soundness metrics, proposing that the community measure soundness in terms of “nines” (e.g., “five 9s” = 99.999% sound), similar to durability and availability contracts in IT systems. Metrics like “Valid@N” help track whether neurosymbolic systems can reliably return valid proofs rather than defaulting to “unknown” or “invalid.” The talk concludes by discussing persistent proofs as a foundation for proof search, repair, and distributed lemma sharing in large-scale neurosymbolic systems.

3.3 Toward Derivable Scientific Discovery: Integrating Data, Theory, and the Future of Machine-Aided Discovery

Cristina Cornelio (Samsung AI – Cambridge, GB, c.cornelio@samsung.com)

License  Creative Commons BY 4.0 International license
© Cristina Cornelio

Scientists aim to build models that not only fit observations but also explain them. Traditionally, theories were developed from domain knowledge and then tested against data, whereas modern machine learning instead extracts patterns directly from large datasets. However, discovering meaningful models within data remains an ongoing challenge.

This talk explores two methods that address this challenge by connecting empirical evidence with formal knowledge: AI-Descartes, which uses symbolic regression to propose candidate laws and logical reasoning to select those consistent with known axioms; and AI-Hilbert, which combines polynomial optimization with logical theories ensuring both theoretical consistency and empirical validation at the same time.

These methods point toward a new paradigm in which data and theory jointly shape scientific discovery, opening new questions and inviting further exploration of the emerging field of AI for science.

3.4 Tooling for Reproducible Neurosymbolic Research: ULLER

Emile van Krieken (VU Amsterdam, NL)

License © Creative Commons BY 4.0 International license
© Emile van Krieken

One of the major transformations behind computational research is the emergence of frictionless reproducibility. In Machine Learning research, executing a highly complicated new experiment is easy due to well-developed modern tools like PyTorch, HuggingFace Transformers, and Weights and Biases. What is missing for frictionless reproducibility in neurosymbolic AI? We see opportunities in common formats for sharing knowledge, and for well-developed standard algorithms for (neurosymbolic) reasoning. Therefore, we introduce two Python libraries in development. ULLER aims to represent knowledge in simple Python. DeepLog efficiently implements building blocks for reasoning in neurosymbolic models on the GPU. These libraries are a push towards standard tooling in neurosymbolic research.

3.5 DeepLog: A Theoretical and Operational Framework for Neurosymbolic AI

Robin Manhaeve (KU Leuven, BE)

License © Creative Commons BY 4.0 International license
© Robin Manhaeve

We present a theoretical and operational framework for neurosymbolic AI called DeepLog. DeepLog introduces building blocks and primitives for neurosymbolic AI that make abstraction of commonly used representations and computational mechanisms used in neurosymbolic AI. DeepLog can represent and emulate a wide range of neurosymbolic systems. DeepLog is implemented in software, relies on the latest insights in implementing algebraic circuits on GPUs, and is declarative in that it is easy to obtain different neurosymbolic models by making different choices for the underlying algebraic structures and logics. The generality and efficiency of the DeepLog neurosymbolic machine is demonstrated experimentally.

3.6 Benchmarking Neurosymbolic AI

Maarten C. Stol (BrainCreators – Amsterdam, NL & VU Amsterdam, NL)

License © Creative Commons BY 4.0 International license
© Maarten C. Stol

Neurosymbolic (NeSy) benchmarking lacks systematic desiderata derived from the separate benchmarking cultures of Machine Learning (ML) and Automated Reasoning (AR). A combination of current practices is insufficient, as e.g., ML benchmarking often struggles with “construct validity”: overly general claims from highly specific data and tasks. While the nature of NeSy construct validity is an open question, we claim NeSy is now in a position to inherit the successes, and overcome the flaws of both fields.

We advocate a shift from static benchmarks and incremental SOTA improvements towards understanding how systems work and fail. We propose benchmark generators based on controllable notions of problem instance hardness, inheriting AR notions (phase transitions)

and ML notions (label noise, OOD data). Such generators enable analysis of how tuning hardness can reverse SOTA rankings and explain performance rather than merely measure it. Moreover, utilizing multiple metrics allows us to then analyze performance trade-offs and trade-off gradients. The resulting synthetic data serves not to replace real-world data, but to analyze and categorize its inherent hardness aspects. We demonstrate this approach with two use cases (maritime surveillance, infrastructure maintenance) involving geo-spatial data and object relations. The potential for NeSy in these use cases is high due to an abundance of background knowledge available to support their AR and ML components.

3.7 Neuro-Symbolic AI and Robotics: Bridging Perception and Sensemaking

Agnese Chiatti (Polytechnic University of Milan, IT)

License  Creative Commons BY 4.0 International license
© Agnese Chiatti

Cornerstone objectives and aspirations in NeSy AI and Robotics overlap in many ways. In Robotics, the demand for world models capable of bridging the gap between sub-symbolic perceptual data and symbolic representations has led to the development of semantic maps. Semantic maps are a paradigmatic example of Neuro-Symbolic representation, as they integrate the outputs of both sub-symbolic and symbolic processing modules.

Establishing a deeper synergy between NeSy AI and Robotics holds significant promise. On the one hand, NeSy approaches can mitigate challenges of verifiability and safety that arise in the use of purely sub-symbolic learning and reasoning techniques, thus enabling reliable and trustworthy robot deployment in real-world contexts. On the other hand, robots, conceived as embodied agents, offer a natural platform for advancing NeSy research by supporting a bidirectional feedback process: (i) incorporating top-down knowledge and constraints to guide reasoning and monitoring, and (ii) discovering new symbolic knowledge through active, bottom-up exploration of the environment. However, the divergence in terminology, methodologies, and benchmarking practices between the NeSy AI and Robotics communities continues to hinder progress at their intersection.

In this talk, I will advocate for drawing insights from both fields, from the perspective of robotic perception and sensemaking. I will discuss use cases in health and safety monitoring, assisted living, and general-purpose visual reasoning. Then, I will present an overview of current research trends as a basis for broader discussion.

3.8 Neurosymbolic AI Education: Integrating Learning and Reasoning

Luís C. Lamb (CatholicTech – Cambridge, US & UFRGS – Porto Alegre, BR)

License  Creative Commons BY 4.0 International license
© Luís C. Lamb

In this talk, I outline the contents of a proposed graduate-level course on Neurosymbolic Artificial Intelligence. The structure of the course and the selection of topics are designed in alignment with the educational objectives and the intended learning outcomes. The lectures and accompanying materials provide contextual background, conceptual foundations, and applied perspectives. The proposed syllabus includes the following topics, to be covered

over a semester: Introduction; History and Evolution of Computer Science and Artificial Intelligence; Basic concepts: symbolic reasoning and connectionist learning; Integrating Reasoning and Learning; Neurosymbolic AI approaches, tools, and applications; Perspectives of the field. Keywords: Neurosymbolic AI; Education; Learning; Reasoning.

4 Working Groups

4.1 Benchmarking Neurosymbolic AI (Working Group 1)

Robert Peharz (TU Graz, AT)

Antonio Vergari (University of Edinburgh, GB)

Mathias Niepert (Universität Stuttgart, DE)

Guy Van den Broeck (UCLA, US)

Benjie Wang (UCLA, US)

Emile van Krieken (VU Amsterdam, NL)

Annette ten Teije (VU Amsterdam, NL)

License © Creative Commons BY 4.0 International license

© Robert Peharz, Antonio Vergari, Mathias Niepert, Guy Van den Broeck, Benjie Wang, Emile van Krieken, and Annette ten Teije

In our break-out benchmark session, we began by clarifying what neurosymbolic (NeSy) systems are and why they matter, which naturally led into a deeper discussion of their goals, desiderata, and underlying premises. Several key promises of NeSy systems emerged: (1) data efficiency, both during training and inference; (2) improved generalization and transferability under distribution shifts while maintaining symbolic constraints; (3) controllability, including safety guarantees and reliability; and (4) modularity at the system level, enabling clearer communication and interfacing between components. Building on this discussion, we sketched a small ontology for the space of neurosymbolic approaches. We closed by examining current successes that fit the NeSy paradigm, even when they are not explicitly labeled as such, such as LLMs combined with tool-use and verifiers, and systems like AlphaGeometry. Overall, we concluded that what makes neurosymbolic systems distinctive is the combination of learning, reasoning, interfacing, and guarantees.

4.2 Benchmarking Neurosymbolic AI (Working Group 2)

Alberto Speranzon (Lockheed Martin Advanced Technology Labs – Eagan, US)

Lia Morra (Polytechnic University of Turin, IT)

Maria-Esther Vidal (TIB Leibniz-Informationszentrum – Hannover, DE & Leibniz Universität Hannover, DE)

Michael Cochez (ELLIS Institute – Espoo, FI & Åbo Akademi University – Turku, FI)

License © Creative Commons BY 4.0 International license

© Alberto Speranzon, Lia Morra, Maria-Esther Vidal, and Michael Cochez

Benchmarks in neurosymbolic AI should not be understood merely as tools for comparing different approaches; they also shape the direction of entire research communities. A well-designed benchmark provides clear targets to work toward and helps reveal the boundaries

and limitations of current state-of-the-art systems. This requires identifying the essential properties of a good AI benchmark: properties that hold independently of specific performance metrics and that may be inspired by the standards of other scientific fields [1, 2]. Because neurosymbolic AI systems, especially in domains like healthcare, must satisfy multidimensional objectives, benchmarks must be able to evaluate these different dimensions. For instance, the FUTURE-AI principles (Fair, Universal, Traceable, Usable, Robust, Explainable) [3] illustrate the variety of criteria that a single benchmark may need to assess.

To be genuinely useful, benchmarks should be prescriptive: they must specify not only which aspects of performance to measure, but also how these measurements should be made, interpreted, and reported. They should explicitly surface the trade-offs inherent in different solutions, enabling researchers to understand how improving one property may hinder another.

A major risk in neurosymbolic AI is the development of benchmarks that appear favourable to NeSy approaches while failing to reflect the real complexities and requirements of the target domain. Avoiding such “bad” benchmarks is essential to ensuring that progress is meaningful and aligned with real-world needs.

References

- 1 Guideline for reporting Experimental Results in Life Sciences. <https://www.nature.com/documents/nr-reporting-life-sciences-research.pdf>
- 2 Guideline for reporting Experimental Results in Physical Sciences. https://www.nature.com/documents/Physical_sciences_reproducibility.pdf
- 3 Lekadir K, Frangi A F, Porras A R, Glocker B, Cintas C, Langlotz C P et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare BMJ 2025; 388 :e081554 doi:10.1136/bmj-2024-081554

4.3 Benchmarking Neurosymbolic AI (Working Group 3)

Luis Lamb (CatholicTech – Cambridge, US & UFRGS – Porto Alegre, BR)

Guy Van den Broeck (UCLA, US)

Sebastijan Dumancic (TU Delft, NL)

Thiviyan Thanapalasingam (University of Amsterdam, NL)

Cristina Cornelio (Samsung AI – Cambridge, GB)

Maarten C. Stol (BrainCreators & VU Amsterdam, NL)

Frank van Harmelen (VU Amsterdam, NL)

License © Creative Commons BY 4.0 International license

© Luis Lamb, Guy Van den Broeck, Sebastijan Dumancic, Thiviyan Thanapalasingam, Cristina Cornelio, Maarten C. Stol, and Frank van Harmelen

This working group identified three key lessons for benchmark design, outlined in this section. A central idea in neurosymbolic AI is that a little logic goes a long way. Many real-world problems do not require full-blown symbolic reasoning or heavy domain theories; instead, lightweight logical constraints can already provide meaningful structure and improve learning outcomes. This perspective encourages focusing on simple forms of domain knowledge and considering applications where only limited, easily articulated knowledge is needed. Industry examples point in this direction as well: systems such as DeepMind’s structured reasoning models or OpenAI’s JSON-constrained generation show how modest symbolic scaffolding can guide powerful neural models. Exploring lighter logical systems that can stand in for prompt engineering or complement it, may reveal a productive balance between expressivity and

practicality. Finding this “sweet spot” of expressivity is crucial. Lightweight knowledge can make models more data-efficient, more reliable, and more controllable, without overburdening them with complex logical formalisms. This aligns with the broader idea that intelligent systems must be educable: as Valiant [1] argues, humans’ unique strength lies in our ability to acquire and use small, structured pieces of knowledge. Neurosymbolic systems should strive for a similar balance.

Lesson for benchmarking design. avoid tasks that are overly logic-heavy and instead design evaluation setups that test how effectively a system can use minimal but meaningful knowledge. Potential paths forward include creating new knowledge-centric datasets from scratch such as ARC or augmenting existing datasets with structured knowledge, using as little additional data as possible while still leveraging symbolic insights.

Industry’s main promise in adopting neurosymbolic AI lies in its potential to deliver safety, reliability, and transparency qualities that become essential in sectors where failure tolerance is extremely low and data is often scarce. These needs are not limited to big tech; industries such as aerospace, manufacturing, finance, and cybersecurity all operate in environments where mistakes are costly and explanations matter. Neurosymbolic approaches can add value precisely because they offer structured guarantees, improved data efficiency, and stronger out-of-distribution robustness critical features given that most real-world prediction tasks fall outside the training distribution and most industrial datasets are small. At the same time, documenting how knowledge is obtained and incorporated, as seen in cybersecurity workflows, becomes a practical requirement for trust and auditability. Instead of investing months into refining prompts, organisations want principled, knowledge-driven methods that scale reliably.

Lesson for benchmark design. include small-data scenarios, OOD tasks, and conditions reflecting fault intolerance to meaningfully test systems intended for industrial deployment.

Neurosymbolic AI should broaden its view beyond purely logical formalisms and consider a wider variety of symbolic representations, including state machines, knowledge graphs, JSON schemas, grammars, and domain-specific languages such as PSL or PDDL. Current NeSy research often overemphasizes declarative logic, even though many practical systems like Python-based code generators or structured planners, already use symbolic constraints to reduce hallucinations and improve reliability. Likewise, perception systems can benefit from injecting world knowledge, such as object permanence or contextual relations, to clean or enrich scene graphs, and robotic navigation can be improved by adding structured knowledge on top of bounding-box-based vision. Instead of searching for entirely new NeSy applications, it may be more productive to highlight the neurosymbolic components that already underpin existing ones, such as code generation or multimodal reasoning in VLMs.

Lesson for benchmark design. avoiding constraints that require logic as the only declarative language and instead standardizing input formats for both data and knowledge, independent of their expressivity.

References

- 1 The Importance of Being Educable: A New Theory of Human Uniqueness, Leslie Valiant, ISBN: 9780691230566.

4.4 Neuro-Symbolic AI in Health Care

Robert Peharz (TU Graz, AT)

Maria-Esther Vidal (TIB Leibniz-Informationszentrum – Hannover, DE & Leibniz Universität Hannover, DE)

Frank van Harmelen (VU Amsterdam, NL)

Annette ten Teije (VU Amsterdam, NL)

Cristina Cornelio (Samsung AI – Cambridge, GB)

Agnese Chiatti (Polytechnic University of Milan, IT)

License  Creative Commons BY 4.0 International license

© Robert Peharz, Maria-Esther Vidal, Frank van Harmelen, Annette ten Teije, and Cristina Cornelio

The FUTURE-AI framework (Fairness, Universality, Traceability, Usability, Robustness, and Explainability) [1] provides a conceptual foundation for trustworthy medical AI, but its principles require concrete technical mechanisms to become actionable. Our discussion focused on what these dimensions mean in the context of neurosymbolic (NeSy) systems, and whether such systems may naturally perform better on some aspects of the framework.

We explored each FUTURE dimension through a NeSy lens. For **Fairness**, NeSy systems offer tools such as counterfactual reasoning and causal interventions between features, as well as ontology-informed learning both of which enable more transparent bias detection and mitigation. Under **Universality**, their combination of symbolic structure and neural generalization supports improved transferability and out-of-distribution performance; symbolic rules further allow human inspection and causal validation. In terms of **Traceability**, the picture is mixed: production traceability (e.g., reproducibility of training pipelines) does not automatically improve, but execution traceability, being able to follow symbolic reasoning steps, can benefit substantially from NeSy architectures. For **Usability**, symbolic components introduce shared concepts and interfaces that may make human interaction with models more intuitive. **Robustness** and **Explainability** also stand to gain from explicit reasoning layers, structured constraints, and semantic consistency. Overall, the discussion highlighted that while NeSy systems do not guarantee trustworthiness, they may provide promising pathways for operationalizing several dimensions of the FUTURE-AI framework.

References

- 1 FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare, FUTURE-AI consortium BMJ 2025; 388 doi: <https://doi.org/10.1136/bmj-2024-081554>

4.5 Architectures for Probabilistic Neurosymbolic AI

Robert Peharz (TU Graz, AT)

Antonio Vergari (University of Edinburgh, GB)

Mathias Niepert (Universität Stuttgart, DE)

Guy Van den Broeck (UCLA, US)

Benjie Wang (UCLA, US)

Emile van Krieken (VU Amsterdam, NL)

Luis C. Lamb (CatholicTech – Cambridge, US & UFRGS – Porto Alegre, BR)

Lia Morra (Polytechnic University of Turin, IT)

Michael Cochez (ELLIS Institute – Espoo, FI & Åbo Akademi University – Turku, FI)

Sebastijan Dumancic (TU Delft, NL)

License  Creative Commons BY 4.0 International license

© Robert Peharz, Antonio Vergari, Mathias Niepert, Guy Van den Broeck, Benjie Wang, Emile van Krieken, Luis C. Lamb, Lia Morra, Michael Cochez, and Sebastijan Dumancic

This working group examined fundamental architectural components underlying probabilistic neurosymbolic systems, with a particular focus on the product of experts formulation. In this framework, one expert represents a probability distribution that encodes beliefs over world states, while the other encodes logical constraints that filter invalid states. Computing the weighted model count (or weighted model integral for continuous domains) by summing this product over all possible states forms the computational backbone of many neurosymbolic pipelines.

Knowledge compilation routines that translate logical constraints into circuit representations with specific structural properties enable closed-form computation of weighted model counts. The determinism property, while enabling tractable inference, restricts circuit expressiveness and may cause exponential blow-up in compiled circuit size. Relaxing determinism constraints could yield smaller circuits and more scalable pipelines. One promising direction, currently explored in work on steering LLM generation with automata, involves translating CNF formulas into partial DNF representations. Hybrid strategies combining these approaches represent an important research direction with broad applicability across neurosymbolic systems.

The discussion also addressed integration of circuit representations into large language models for constraint enforcement and generation acceleration. Training circuits to approximate LLM distributions was compared with approaches that embed circuits as final layers, offering greater expressiveness but potentially limiting constraint enforcement to bounded token windows. Finally, analogies to constrained generation in diffusion models were explored, noting that classifier-guided approaches employ product-of-expert formulations for score matching while sidestepping exact partition function computation. Combining these directions with tractable circuit models presents opportunities for novel hybrid neurosymbolic architectures.

4.6 Neurosymbolic Methods for Safe and Reliable Robotic Systems

Guy Van den Broeck (UCLA, US)

Alberto Speranzon (Lockheed Martin Advanced Technology Labs – Eagan, US)

Agnese Chiatti (Polytechnic University of Milan, IT)

Annette ten Teije (VU Amsterdam, NL)

Cristina Cornelio (Samsung AI – Cambridge, GB)

Thiviyan Thanapalasingam (University of Amsterdam, NL)

Sebastijan Dumancic (TU Delft, NL)

License © Creative Commons BY 4.0 International license

© Guy Van den Broeck, Alberto Speranzon, Agnese Chiatti, Annette ten Teije, Cristina Cornelio, Thiviyan Thanapalasingam, and Sebastijan Dumancic

This working group examined how neurosymbolic methods can improve the reliability and safety of robotic systems. A central observation was that LLMs perform impressively for hierarchical planning, until they do not. Current approaches use symbolic verifiers to validate LLM-generated plans, feeding errors back into the prompt for iterative refinement, an approach sometimes characterised as “prompt and pray”. While this works in many scenarios, LLMs fundamentally lack the safety guarantees required for real-world deployment, where even controlled environments can produce unexpected uncertainty.

Neurosymbolic methods offer principled ways to handle such ambiguity. Three aspects of assurance were identified as critical: system design, concerning how to architect systems with built-in safety properties; runtime monitoring, using techniques such as signal temporal logic to detect and respond to violations during execution; and task evaluation, measuring success in completing objectives. Additionally, small-data requirements in robotics applications align well with the data efficiency that neurosymbolic systems naturally provide.

Recent benchmarking efforts such as the Embodied Agent Interface [1] offer standardised evaluation for LLM-based decision making, and work on Bayesian-conditioned controllable generation has shown success rate improvements from 20% to 40% by encoding constraints, demonstrating that neurosymbolic approaches yield measurable benefits. A key insight was that rather than imposing all possible constraints, selectively adding those with high failure probability proves more effective, raising the possibility of learning which constraints matter most.

The group also noted that edge computing demands smaller models, and neurosymbolic constraints can help achieve compactness without sacrificing performance. Finally, connections between vision-language models and world models were discussed as both containing neurosymbolic components, and the division between robotics and AI venues was noted as a barrier to knowledge sharing across communities.

References

- 1 Li, Manling and Zhao, Shiyu and Wang, Qineng and Wang, Kangrui and Zhou, Yu and Srivastava, Sanjana and Gokmen, Cem and Lee, Tony and Li, Li Erran and Zhang, Ruohan and Liu, Weiyu and Liang, Percy and Fei-Fei, Li and Mao, Jiayuan and Wu, Jiajun. “Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making”. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track (Oral)*, 2024.

4.7 Updating the Wikipedia Entry for Neurosymbolic AI

Emile van Krieken (VU Amsterdam, NL)

Robin Manhaeve (KU Leuven, BE)

Luis C. Lamb (CatholicTech – Cambridge, US & UFRGS – Porto Alegre, BR)

License © Creative Commons BY 4.0 International license
© Emile van Krieken, Robin Manhaeve, and Luis C. Lamb

This working group examined the Wikipedia entry for Neurosymbolic AI¹ and found it incomplete, with an unclear definition, limited historical coverage, missing connections to adjacent fields, and minimal discussion of practical applications. The page also lacks adequate treatment of the field’s relationship to ethical AI, explainable AI, safe AI, and trustworthy AI research.

Inspired by the well-organized Machine Learning Wikipedia page, the group developed a comprehensive restructuring plan with the following proposed additions and revisions:

- Rewriting the introduction and definition for clarity and completeness
- Expanding the history section with improved coverage of the field’s development
- Adding a new section on relationships to other fields, including machine learning and deep learning, knowledge representation and reasoning, theorem proving, cognitive science, ethical and responsible AI, explainable AI, safe AI, trustworthy AI, databases and software engineering, and artificial general intelligence
- Creating a new applications section documenting real-world deployments
- Revising the research challenges section to reflect current open problems
- Updating the implementations and software section with recent developments

Following the seminar, participants committed to collaboratively editing the entry, aiming to provide the broader scientific community and general public with an accurate and up-to-date resource on Neurosymbolic AI.

5 Plenary Discussions

5.1 Tooling for Neurosymbolic AI

This plenary discussion focused on the infrastructure needed to achieve “frictionless reproducibility” in neurosymbolic AI research. In machine learning, executing complex experiments has become straightforward thanks to well-developed tools such as PyTorch and PyTorch Geometric [1]. The NeSy field lacks comparable infrastructure, which hinders both newcomers and experienced researchers. Two complementary Python libraries under development were presented as steps toward addressing this gap. ULLER (Unified Language for LEarning and Reasoning) [2] provides a Python domain-specific language for expressing background knowledge, with compilers targeting different NeSy system formats. ULLER decouples the syntax of knowledge representation from the semantics provided by specific NeSy systems, supporting classical first-order logic, fuzzy logic, and probabilistic semantics. Users can define constraints once and then invoke different NeSy systems with minimal code changes. DeepLog [3] operates at a lower level of abstraction, providing a neurosymbolic machine consisting of two components: a language for specifying neurosymbolic models based on

¹ https://en.wikipedia.org/wiki/Neuro-symbolic_AI

annotated grounded first-order logic, and extended algebraic circuits as computational graphs. DeepLog efficiently implements reasoning primitives on the GPU and can emulate a wide range of existing neurosymbolic systems including DeepProbLog, NeurASP, and Semantic Loss. Together, ULLER addresses the first-order knowledge level while DeepLog handles the propositional and inference levels. The discussion also examined Pylon [4], an earlier NeSy framework that struggled to gain community adoption despite its capabilities. This raised concerns about the risk of repeatedly reinventing tools that fail to achieve critical mass. Drawing parallels to the challenges faced by Semantic Web technologies, the group noted that even well-designed tooling requires sustained effort to build community uptake. Standardised input formats for both data and knowledge, independent of their logical expressivity, were identified as a key enabler for broader NeSy adoption and for avoiding fragmentation across competing frameworks.

References

- 1 Fey, Matthias and Lenssen, Jan Eric. “Fast Graph Representation Learning with PyTorch Geometric.” In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
- 2 van Krieken, Emile and Badreddine, Samy and Manhaeve, Robin and Giunchiglia, Eleonora. “ULLER: A Unified Language for Learning and Reasoning.” In *Proceedings of the 18th International Conference on Neural-Symbolic Learning and Reasoning (NeSy 2024)*, pp. 219–239, 2024.
- 3 Derkinderen, Vincent and Manhaeve, Robin and Adriaensen, Rik and Van Praet, Lucas and De Smet, Lennert and Marra, Giuseppe and De Raedt, Luc. “The DeepLog Neurosymbolic Machine.” *arXiv preprint arXiv:2508.13697*, 2025.
- 4 Ahmed, Kareem and Teso, Stefano and Chang, Kai-Wei and Van den Broeck, Guy and Vergari, Antonio. “Pylon: A PyTorch Framework for Learning with Constraints.” In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*, 2022.

5.2 Bridging NeSy and LLMs

In the plenary session on “*What is the role of Neurosymbolic AI in the era of foundational models?*”, the discussion centered on emerging patterns in how neurosymbolic (NeSy) methods can enhance and structure the use of large language models. Several concrete integration patterns were identified. LLMs can assist in knowledge extraction and construction, transforming unstructured text into symbolic or semi-formal representations. They can also be pushed toward formal reasoning, for example through constrained prompting or by delegating certain reasoning steps to external symbolic engines. Another pattern involves feeding symbolic feedback back into the LLM, either dynamically during generation or as a post-processing verification layer, helping to rectify inconsistencies. Finally, symbolic systems can be used to generate high-quality training or fine-tuning data, enabling LLMs to inherit structured reasoning capabilities they cannot easily learn from raw text.

A key takeaway from the session was the mutually reinforcing relationship between LLMs and symbolic methods. LLMs can dramatically reduce the cost of constructing and maintaining formal representations by automating parts of the knowledge engineering pipeline. Conversely, symbolic constraints, logic, and verifiable representations help reduce hallucinations and improve reliability in LLM outputs. Looking ahead, the group envisioned a future in which elements such as logic, causality, and other structured reasoning formalisms become integrated directly into model training regimes, rather than remaining purely external modules. This deeper fusion could lead to more robust, interpretable, and systematically grounded AI systems that retain the flexibility of foundation models while gaining the trustworthiness and rigor of symbolic reasoning.

5.3 Education Materials for NeSy AI

In the plenary discussion on educational materials for neurosymbolic AI, several overarching themes emerged. The consensus was that the primary target audience should be advanced master's students or early-stage PhD students. A key tension discussed was the balance between practical, hands-on learning and deeper research-oriented understanding: should the material focus on using neurosymbolic tools, or on developing the theoretical foundations behind them? The group agreed that a modern resource should avoid the static, paper-collection style of traditional textbooks and instead adopt an agile, update-friendly format. Given the increasing momentum in neurosymbolic AI, many felt that now is an ideal time to develop a comprehensive teaching resource or textbook.

Prerequisites were another central point: students should enter with a working knowledge of logic and formal representation (potentially including exposure to symbolic programming), machine learning fundamentals, and a clear vocabulary that bridges terminology across ML and knowledge representation (KR). This led to multiple possible design directions for structuring the material. One idea is framed as “advanced ML”, is to start from familiar ML concepts and progressively extend them with symbolic reasoning, explicitly motivating each addition through ML's shortcomings. Another approach begins with the high-level “why learning and reasoning?”, incorporating Kautz's taxonomy, examples of existing systems, and industrial case studies. A more foundational option would foreground probability, Boolean circuits, and the expressivity of neural networks and logics, culminating in a taxonomy of neurosymbolic system classes. Yet another design path would contrast ML and KR directly: starting from the weaknesses of each and showing how the two fields complement and strengthen one another. A narrative “product-development” storyline was also proposed, modeling the material as the evolution of increasingly capable NeSy systems.

The discussion also highlighted concrete tools suitable for teaching, such as PyNeuralog, ULLER, and SKI-Lang (to be introduced at ECAI 2025). Additional topics worth integrating include inductive programming, the intellectual history of symbolic vs. connectionist AI (“the two AI religions”), and broader historical reflections inspired by Kuhnian paradigm shifts. Luis Lamb existing course structure on Neurosymbolic AI Education was referenced as a strong template, spanning introductions, historical context, symbolic vs. connectionist basics, motivations for integration, tools and applications, and forward-looking perspectives. Finally, the group explored different strategies for generating the teaching materials. Options ranged from writing an entirely new textbook, to producing a Version 0 using generative AI and iterating, to assembling handouts and slides that later evolve into a book. Alternatives included extending an existing text, collaborating with current authors, or producing lecture videos as the first step. Several existing books were reviewed [1, 2, 3] with the consensus that none fully addresses both the conceptual depth and modern research landscape needed for a master- or PhD-level neurosymbolic AI course. An online course was also considered as a complementary or preliminary step. Overall, the discussion reflected a shared recognition that neurosymbolic AI education is both timely and ripe for a coherent, rigorous, and modern learning resource.

References

- 1 Dingli, A., & Farrugia, D. (2023). *Neuro-Symbolic AI: Design transparent and trustworthy systems that understand the world as you do*. Birmingham & Mumbai: Packt Publishing. ISBN 978-1-80461-762-5. O'Reilly Media
- 2 Shakarian, P., Simari, G. I., Baral, C., Xi, B., & Pokala, L. (2023). *Neuro-Symbolic Reasoning and Learning*. Springer. ISBN 978-3-031-39179-8.
- 3 Bhuyan, B. P., Ramdane-Cherif, A., Singh, T. P., & Tomar, R. (2025). *Neuro-Symbolic Artificial Intelligence: Bridging Logic and Learning (Studies in Computational Intelligence, Vol. 1176)*. Springer Nature Singapore. ISBN 978-981-97-8171-3. Scribd

6 Communicating Results

6.1 Planning on joint paper

As a result of an intensive week of working together around the theme of neurosymbolic approaches and its applications, we aim for a Communications of the ACM paper for a broad audience. Below are the contours of the paper.

Background. Neuro-symbolic AI systems that combine learning with reasoning are rapidly gaining attention in recent years. It is becoming increasingly clear that the data-intensive probabilistic approaches of neural networks and language models must be complemented by symbolic algorithms in order to function in domains with high stakes and low data regimes, which are typical in many real-world applications.

Message. We argue that the combination of learning and reasoning is not just a single niche approach in AI, but is actually a spectrum that encompasses many already successful techniques. This leads to rethinking the design of AI systems in general within this spectrum. We show that many leading AI systems can be “rationally reconstructed” using four neuro-symbolic principles: Reasoning, Assurances, Interfacability and Learning (RAIL for short). Firstly, this gives us a unified view on many seemingly disparate designs of AI systems, ranging from physics-aware machine learning to DeepMind’s Alpha-X suite (AlphaGo, AlphaGeometry, AlphaFold) to Knowledge Completion, Causal Learning, and LLMs augmented with tool usage. Secondly, the neuro-symbolic RAIL principles allow engineers to make informed choices about AI system design, by moving along the axes for each of these 4 principles.

Goal.

- (i) We show that many works in AI share some/all of the NeSy principles.
- (ii) This gives a unified look at many disconnected lines of work and reveals possible connections/opportunities/challenges.
- (iii) The NeSy field will benefit from learning the lessons learned in these other communities.
- (iv) Other communities would benefit from looking at their own work through the NeSy lens.

6.2 Wikipedia Entry for Neurosymbolic AI

A working group has committed to improving the Wikipedia entry for Neurosymbolic AI, which currently lacks comprehensive coverage of the field’s scope, methods, and applications. The group will rewrite the introduction for clarity, expand the historical overview, add sections on relationships to adjacent fields (including explainable AI, safe AI, and trustworthy AI), and document real-world applications. The goal is to establish a reliable, publicly accessible reference that reflects the current state of neurosymbolic research. See Section 4.7 for details on the proposed changes.

6.3 Educational Materials

A working group has committed to developing educational resources for neurosymbolic AI, targeting advanced master’s students and early-stage PhD researchers. The planned deliverable is a course structure for a comprehensive textbook covering both theoretical

foundations and practical tooling. The group aims to bridge the gap between machine learning and knowledge representation communities by establishing shared vocabulary and providing hands-on exercises using tools such as ULLER and DeepLog. See Section 5.3 for the full discussion on content structure and pedagogical approaches.

Participants

- Agnese Chiatti
Polytechnic University of Milan, IT
- Michael Cochez
ELLIS Institute – Espoo, FI & Åbo Akademi University – Turku, FI
- Byron Cook
Amazon Web Services – Seattle, US & University College London, GB
- Cristina Cornelio
Samsung AI – Cambridge, GB
- Artur d’Avila Garcez
City University of London, GB
- Sebastijan Dumancic
TU Delft, NL
- Egor V. Kostylev
University of Oslo, NO
- Luis C. Lamb
CatholicTech – Cambridge, US & UFRGS – Porto Alegre, BR
- Robin Manhaeve
KU Leuven, BE
- Lia Morra
Polytechnic University of Turin, IT
- Mathias Niepert
Universität Stuttgart, DE
- Robert Peharz
TU Graz, AT
- Ute Schmid
Universität Bamberg, DE
- Alberto Speranzon
Lockheed Martin Advanced Technology Labs – Eagan, US
- Maarten C. Stol
Braincreators – Amsterdam, NL & VU Amsterdam, NL
- Annette ten Teije
VU Amsterdam, NL
- Thiviyan Thanapalasingam
University of Amsterdam, NL
- Guy Van den Broeck
UCLA, US
- Frank van Harmelen
VU Amsterdam, NL
- Emile van Krieken
VU Amsterdam, NL
- Antonio Vergari
University of Edinburgh, GB
- Maria-Esther Vidal
TIB Leibniz-Informationszentrum – Hannover, DE & Leibniz Universität Hannover, DE
- Benjie Wang
UCLA, US



Security and Privacy of Large Language Models

Stephan Günnemann^{*1}, Pavel Laskov^{*2}, Emil Lupu^{*3},
Vera Rimmer^{*4}, Advije Rizvani^{†5}, and Qianying Liao^{†6}

- 1 TU München – Garching, DE. s.guennemann@tum.de
- 2 Universität Liechtenstein – Vaduz, LI. pavel.laskov@uni.li
- 3 Imperial College London, GB. e.c.lupu@imperial.ac.uk
- 4 KU Leuven, BE. vera.rimmer@kuleuven.be
- 5 Universität Liechtenstein – Vaduz, LI. advije99@gmail.com
- 6 KU Leuven, BE. qianying.liao@kuleuven.be

Abstract

Large Language Models (LLMs) have rapidly evolved from experimental systems capable of generating simple text to powerful general-purpose tools that solve exam-level problems, assist human experts, summarize complex documents, and write code, leading to their widespread deployment in production environments to improve productivity. Their rapid adoption has outpaced scientific understanding of the associated security and privacy risks, creating a growing gap between real-world use and trust in LLM-based applications. Unlike traditional machine learning models, LLMs are inherently general-purpose, operate at unprecedented scale, are often proprietary, and are accessed through interactive dialog interfaces. This makes their behavior difficult to predict and introduces novel attack surfaces and emerging phenomena such as hallucinations. Existing methods and analyzes for conventional machine learning are insufficient to cope with these aspects. So, this Dagstuhl Seminar “Security and Privacy of Large Language Models” (25461) aimed to initiate a systematic scientific discussion on the security and privacy of language models by bringing together researchers from AI, security, privacy, and natural language processing to address three central questions: how safe and secure are LLMs in adversarial environments; to what extent users’ privacy may be at risk when interacting with LLMs; and what other categories of societal impact may arise from the rapid advancement of LLM technologies. The seminar sought to systematize existing knowledge, identify high-risk applications and prevalent attack vectors, discuss defense strategies and deployment challenges, and outline directions for future research in LLM security and privacy.

Seminar November 9–14, 2025 – <https://www.dagstuhl.de/25461>

2012 ACM Subject Classification Security and privacy → Software and application security;
Security and privacy → Human and societal aspects of security and privacy; Computing
methodologies → Natural language processing

Keywords and phrases Large Language Models, Artificial Intelligence, Security and Privacy

Digital Object Identifier 10.4230/DagRep.15.11.87

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed
under a Creative Commons BY 4.0 International license

Security and Privacy of Large Language Models, *Dagstuhl Reports*, Vol. 15, Issue 11, pp. 87–113

Editors: Stephan Günnemann, Pavel Laskov, Emil Lupu, and Vera Rimmer



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Stephan Günnemann (TU München – Garching, DE)

Pavel Laskov (Universität Liechtenstein – Vaduz, LI)

Emil Lupu (Imperial College London, GB)

Vera Rimmer (KU Leuven, BE)

License  Creative Commons BY 4.0 International license
© Stephan Günnemann, Pavel Laskov, Emil Lupu, and Vera Rimmer

Overview

Large Language Models (LLM) have rapidly evolved from simple toys capable of generating stories about unicorns to powerful tools that can solve exam-level problems, advise human experts, summarize texts, and write code, among other tasks. The remarkable advances in LLM capabilities over the past few years have led to their rapid adoption and deployment in production environments without allowing to the time necessary to understand, analyze, and mitigate the resulting security risks. As a result, the security and privacy aspects are not well understood in the scientific community, causing a concerning lack of trust in the widely deployed LLM applications. This Dagstuhl Seminar “Security and Privacy of Large Language Models” (25461) was an attempt to bring together leading researchers in AI, security, privacy, and NLP to systematize knowledge, assess systemic risks, identify mitigation challenges, and map future directions in this area.

LLMs are qualitatively different from more traditional machine learning models. They are intended to be “general purpose” and handle a broad variety of tasks. While an image classifier trained to distinguish cats from dogs cannot even distinguish cars from airplanes, a language model is expected to be useful on any question expressible natural language. This makes it much more challenging to predict the context in which they will be deployed, the tasks they will be given and how they will perform. There are also many other differences. For example, these models have hundreds of billions to trillions of parameters and are often private to companies that sell their services. This makes it more difficult to understand how they work, a challenging question already even when models are open. Loss functions optimized by LLMs are extremely complex, with local minima being a tangible threat to their generalization capabilities – a threat that also results in the widely recognized “hallucination” phenomena. Last, but not least, the learning components of LLMs are not directly exposed to end users but accessed via a dialog interface (and its own learning mechanisms) that itself can lead to “intriguing properties”.

This Dagstuhl Seminar aimed to bring together researchers from a diverse set of backgrounds to discuss research directions that could lead to the scientific foundation for privacy and security of LLMs.

Goal of the Seminar

The seminar focused on three main themes of discussion, consistently with the issues highlighted above:

- How secure are language models? When an adversary is present and interacts with these models, can they cause the model to behave incorrectly? Which mitigations are effective?
- How private are language models? Can an adversary interact with the model to violate the privacy of users? What are the possible privacy risks?
- What other categories of societal impact may arise from the rapid advancement and deployment of LLMs within information and institutional systems?

In addition to these themes, the seminar also provided an open environment for the discussion of related topics. Emerging research in this area is often exploratory and rapidly evolving. So, organisers deliberately encouraged participants to bring different perspectives from the academic and the industry environments.

Overall Organization and Schedule

The structure and schedule of the seminar combined the advantages of conventional conference formats with the specific traditions of Dagstuhl Seminars in the following schedule:

Schedule	Activities
Day 1	Seminar introduction, short self-introductions by participants, six keynote presentations
Day 2	Three keynote presentations, a brain-storming session on challenges and open problems, the building of working groups, one breakout session
Day 3	Two breakout sessions, one keynote presentation, social event
Day 4	Reporting from breakout sessions, a podium discussion of societal impact of LLM security
Day 5	A discussion of future research on LLM security, summary of results

In what follows, the report presents an overview of the contributed talks, a summary of the main results that emerged from the seminar discussions, and detailed reports from the formed working groups.

2 Table of Contents

Executive Summary

<i>Stephan Günnemann, Pavel Laskov, Emil Lupu, and Vera Rimmer</i>	88
--	----

Overview of Talks

Is AI Security Doomed? <i>Ilia Shumailov</i>	92
Inducing bias and measuring consistency in LLM responses <i>Lujo Bauer</i>	92
Phishing in the LLM Era: Challenges and Opportunities <i>Giovanni Apruzzese</i>	93
LLM Security & Privacy: A Perspective <i>Giovanni Cherubin</i>	93
Capture the Narrative: Social media manipulation wargaming for cyberliteracy and research <i>Hammond Pierce</i>	94
LLM for malware detection and classification <i>Lorenzo Cavallaro</i>	94
Impossibility Results in AI Safety <i>Eric Wong</i>	95
AI-based Traffic Modeling for Network Security and Privacy: Challenges Ahead <i>Dinil Mon Divakaran</i>	95
Safety of Large Language Models Beyond English <i>Wojciech Kusa</i>	96
LLM Security – An industrial and an end-users' perspective <i>Kathrin Grosse</i>	97

Working Groups

AI System Capabilities Certification <i>Tobias Callies, Jonathan Petit, Mathilde Raynal, and Eric Wong</i>	98
On LLM manipulation of humans <i>Lujo Bauer, Rebekka Burkholz, Lorenzo Cavallaro, Andreas Happe, Pavel Laskov, Nils Lukas, Hammond Pearce, and Lea Schoenherr</i>	99
Privacy risks of LLMs <i>Giovanni Cherubin, Ana-Maria Cretu, Stephan Guenemann, Qianying Liao, Vera Rimmer, and Ali Satvaty</i>	101
Responsible Usage of LLMs in Peer Review <i>Giovanni Apruzzese, Amir Houmansadr, Marc Juarez, Florian Kerschbaum, Wojciech Kusa, Fabio Pierazzi, Konrad Rieck, and Christian Wressnegger</i>	104
Modeling the Attack Surface of LLM- and AI-enabled Systems <i>Hyrum Anderson, Javier Carnerero Cano, Dinil Mon Divakaran, Mario Fritz, Emil C. Lupu, Irina Nicolae, and Tim Van hamme</i>	107

S. Günnemann, P. Laskov, E. Lupu, V. Rimmer, A. Rizvani, and Q. Liao **91**

LLM Safety & Security Benchmarks
*Maksym Andriushchenko, Marc Fischer, Kathrin Grosse, Daphne Ippolito, Andrew
Paverd, Advije Rizvani, and Ilia Shumailov* 108

Participants 113

3 Overview of Talks

3.1 Is AI Security Doomed?

Ilia Shumailov (AI Security Company – London, GB)

License  Creative Commons BY 4.0 International license
© Ilia Shumailov

This talk examines emerging security risks in modern large language model (LLM) systems beyond model-level performance, with a focus on system opacity and prompt injection attacks. We first move past the abstraction of simple inference interfaces (e.g., `model.generate()`) to analyze the complex, layered machine learning pipelines underlying contemporary LLM deployments, highlighting how pre-trained models, third-party components, and opaque dependencies introduce untrusted origins and systemic vulnerabilities. We then turn to prompt injection attacks, which have proven both pervasive and practically exploitable, and critically assess the current landscape of proposed mitigations. Despite significant research attention, effective and deployable defenses remain limited. The talk concludes by outlining promising directions for system-level defenses that aim to mitigate prompt injection risks in realistic LLM-based applications.

3.2 Inducing bias and measuring consistency in LLM responses

Lujo Bauer (Carnegie Mellon University – Pittsburgh, US)

License  Creative Commons BY 4.0 International license
© Lujo Bauer

This talk explores a subtle but powerful threat to user autonomy arising from prompt optimization services for large language models (LLMs). While such services aim to reduce the burden of prompt engineering, they also introduce a new and largely overlooked attack surface: untrusted prompt providers can deliberately manipulate prompts to bias model outputs. We demonstrate that minor synonym-level modifications to prompts can significantly influence LLM behavior, increasing the likelihood that models mention targeted concepts – such as brands, political entities, or nations – by up to 78%. Through a user study, we show that these adversarially perturbed prompts are indistinguishable from benign prompts to human users, successfully steer model responses, and increase user attention to the targeted concepts without raising suspicion. The talk highlights the practical feasibility and societal implications of prompt-based manipulation and discusses mitigation strategies, including the need for warnings and safeguards when relying on prompts from untrusted sources.

3.3 Phishing in the LLM Era: Challenges and Opportunities

Giovanni Apruzzese (Reykjavik University, IS)

License © Creative Commons BY 4.0 International license
© Giovanni Apruzzese

Every day, our inboxes are flooded with unsolicited emails, ranging between annoying spam to more subtle phishing scams. Unfortunately, despite abundant prior research efforts proposing solutions that achieve near-perfect accuracy and despite numerous organizations employing phishing education campaigns, the reality is that countering malicious emails still remains an unsolved dilemma – which has been exacerbated by the advent of LLM.

This talk will cover two recent publications (AsiaCCS’25, and AISEC’25) which present novel results featuring the application of LLMs in the phishing-email context. First, I will show the effectiveness that LLM-generated emails, crafted with naive OSINT techniques, can have against 18k employees of three different organizations. Then, I will focus on the research side, and elucidate various “open problems” that affect academic literature on phishing-email detection – such as the excessive reliance on datasets containing emails exchanged up to 20 years ago. Finally, I will present how LLMs can be used defensively, i.e., to generate large-scale datasets of phishing emails which could resemble those sent by real attackers, and which could be used to benchmark existing (and future) detection methods – including, of course, those based on LLMs!

3.4 LLM Security & Privacy: A Perspective

Giovanni Cherubin (Microsoft Research – Cambridge, GB)

License © Creative Commons BY 4.0 International license
© Giovanni Cherubin

When developing security and privacy for “traditional” supervised Machine Learning (ML), the research community has matured a lot of wisdom. Arguably, this is in part attributable to the deep theoretical understanding of ML itself, which was developed by scientists such as T. Cover, P. Hart, C. J. Stone, A. Chervonenkis, V. Vapnik, and L. Valiant in the second half of 1900. Thanks to this, security & privacy for traditional ML flourished, with successes (e.g., privacy-preserving ML), as well as with the uncovering of negative results (e.g., impossibility of preventing adversarial examples in some domains). Conversely, developing a similarly rich understanding of auto-regressive models, such as Large Language Models (LLMs) seems out of reach for the foreseeable future. This makes securing this new learning pattern an extremely complex endeavor.

In this invited talk, we focus on the main security threat to LLMs, prompt injection attacks. We look at the defence paradigms to date, and show how each of them has inherent limitations; in conjunction, we highlight gaps and promising avenues in this field. Finally, we look at how traditional ML and LLMs behave differently also when it comes to privacy; this suggests that applying to LLMs our largely mature understanding of ML privacy is far from immediate, and can lead to mistakes. We hope that this analysis can help the research community prioritize their focus over the coming decade.

3.5 Capture the Narrative: Social media manipulation wargaming for cyberliteracy and research

Hammond Pierce (UNSW – Sydney, AU)

License  Creative Commons BY 4.0 International license
© Hammond Pierce

This talk presents Capture the Narrative, an educational competition designed to expose and study the mechanics of AI-driven information manipulation in online social platforms. The primary intention of the competition is to raise awareness of how coordinated automated agents can shape public discourse, while providing participants with hands-on experience in adversarial thinking, social media dynamics, and cyber-security challenges. Participants formed teams to design and deploy bots that operated on a simulated social media platform during a fictional election campaign, with the goal of amplifying or suppressing narratives to influence public opinion and polling outcomes. The competition unfolded over several weeks, during which millions of posts were generated and interacted with under controlled conditions. The outcome demonstrated that even simple, well-coordinated automated strategies can dominate online narratives and meaningfully affect collective perceptions, ultimately altering the simulated election result. The talk reflects on the lessons learned from the competition, including the effectiveness of narrative manipulation, the vulnerabilities of social platforms, and the value of experiential approaches for understanding AI-enabled influence operations.

3.6 LLM for malware detection and classification

Lorenzo Cavallaro (University College London, GB)

License  Creative Commons BY 4.0 International license
© Lorenzo Cavallaro

Large Language Models (LLMs) are emerging as a powerful paradigm for malware detection and classification, offering a flexible alternative to traditional signature-based and heuristic approaches. By analyzing source code, decompiled binaries, and system call traces, LLMs can reason about program semantics and malicious intent, enabling accurate classification of malware families and even zero-shot detection of previously unseen threats. This talk surveys recent advances in applying LLMs to malware analysis, including code-level reasoning over malicious logic, the use of decompilation tools to bridge binary and source-level analysis, and the classification of system call sequences to distinguish benign from malicious behavior. We also discuss context-driven approaches for domain-specific threats, such as Android malware and script-based attacks, as well as the role of LLMs in explainable malware analysis through natural-language descriptions of observed behavior. While modern transformer-based models have demonstrated significant performance gains on complex and evolving malware families, important challenges remain, including context window limitations, structural complexity of real-world malware, and adversarial evasion techniques. The talk concludes by outlining open research directions for scaling, hardening, and operationalizing LLM-based malware detection systems.

3.7 Impossibility Results in AI Safety

Eric Wong (University of Pennsylvania – Philadelphia, US)

License © Creative Commons BY 4.0 International license
© Eric Wong

What kinds of defenses are possible for safety research? We will discuss two impossibility results, one theoretical and one empirical. Theoretically, we will demonstrate how rule-following can always be attacked with token-level jailbreaks as an architectural inevitability of attention. Empirically, we will demonstrate how a dedicated attacker aiming to misuse a model can use decomposition to hide their intent and make their queries indistinguishable from benign queries. We will end with some optimistic discussion on how these studies inform potential ways forward.

References

- 1 Xue, Anton, Avishree Khare, Rajeev Alur, Surbhi Goel, and Eric Wong. “Logicbreaks: A framework for understanding subversion of rule-based inference.” arXiv preprint arXiv:2407.00075 (2024).
- 2 Brown, Davis, Mahdi Sabbaghi, Luze Sun, Alexander Robey, George J. Pappas, Eric Wong, and Hamed Hassani. “Benchmarking Misuse Mitigation Against Covert Adversaries.” arXiv preprint arXiv:2506.06414 (2025).
- 3 Guardieiro, Vitoria, Adam Stein, Avishree Khare, and Eric Wong. “Instruction Following by Boosting Attention of Large Language Models.” arXiv preprint arXiv:2506.13734 (2025).

3.8 AI-based Traffic Modeling for Network Security and Privacy: Challenges Ahead

*Dinil Mon Divakaran (Institute for Infocomm Research, A*STAR – Singapore, SG)*

License © Creative Commons BY 4.0 International license
© Dinil Mon Divakaran

This talk reflected on the advancements in network traffic analysis using machine learning (ML) and discussed the challenges that remain in effectively securing networks and protecting user privacy.

AI-based traffic models are developed to address various challenging problems in network security and privacy (NetS&P). Deep Learning (DL) models are particularly valuable as network data is enormous (a 1 Gbps link can generate hundreds of GBs per day) and the feature space is vast (representation of a short browsing session may require 2,000 features).

Some of the common and relevant NetS&P tasks where AI models are developed are i) Detection of network anomalies, ii) Classification of attacks into different types, iii) Website fingerprinting attack on user privacy, and iv) IoT device identification. There are also other problems of interest, such as LLM token inference attack, de-anonymization attack on Tor networks, countering of network fingerprinting attacks on user privacy, evading network-based censorship, etc. Research on these problems have helped us identify the features from (encrypted) network traffic that are useful for modeling.

Despite these advancements, there are significant challenges in the domain of traffic modeling and analysis that need to be addressed to effectively secure our networks from evolving threats and attacks.

Data Challenge: Lack of high-quality, labeled real-world network traffic data, primarily due to the privacy risks associated with leaking sensitive user information. Existing open datasets are shown to have artifacts that affect the evaluations, consequently creating hurdles for objective evaluations and generalization of AI models developed.

Practical Deployment: Feature extraction for ML inference is challenging as it demands real-time per-packet processing at network speeds of up to hundreds of Gbps. Programmable data planes presents an opportunity to tackle this.

Explainability: Cybersecurity solutions need to provide human-consumable explanations for the predicted results to build trust between different stakeholders, to assist analyst decide on the next steps, to improve model performance as well as to help build cost-effective solutions.

Model Convergence: Is it time to move beyond multiple task-specific model, to build a foundation model for network traffic analysis? An effective foundation network model must possess key properties, including flexible representation of traffic to deal with different formats and granularity of packet captures at different networks, robustness to missing and noisy labels during fine-tuning, and explainability for various downstream tasks.

References

- 1 Dinil Mon Divakaran, “Traffic Modeling for Network Security and Privacy: Challenges Ahead,” in *Proc. AAAI Workshop on AI for Cyber Security*, 2026.

3.9 Safety of Large Language Models Beyond English

Wojciech Kusa (NASK – Warsaw, PL)

License  Creative Commons BY 4.0 International license
© Wojciech Kusa

As Large Language Models (LLMs) continue to evolve, ensuring their safety across multiple languages has become a critical concern. While LLMs demonstrate impressive capabilities in English, their safety mechanisms may not generalize effectively to other languages, leading to disparities in toxicity detection, bias mitigation, and harm prevention. In this talk, I will present a systematic review examining the multilingual safety of LLMs by synthesizing findings from recent studies that evaluate their robustness across diverse linguistic and cultural contexts beyond the English language. I will discuss the methodologies used to assess multilingual safety and identify challenges such as dataset availability and evaluation biases. Based on this analysis, I will highlight gaps in multilingual safety research and provide recommendations for future work. I will also show recent results in multilingual refusal alignment, illustrating where smaller language models struggle with following instructions in low-resource language settings.

3.10 LLM Security – An industrial and an end-users’ perspective

Kathrin Grosse (IBM Research Europe – Zürich, CH)

License © Creative Commons BY 4.0 International license
© Kathrin Grosse

Joint work of Javier Carnerero Cano, Marc Stoecklin, Ian M. Molloy, Vinayaka Pandit, Mark Purcell

LLMs are widely used in both industry and by end users. Yet, on both ends of these spectrums, there are very different challenges, which we discuss in this talk. On the one hand, companies must ensure the availability, confidentiality, and integrity of systems on a level that supports both governance and auditability. Due to the complexity of LLM-based programs, this requires many aspects when, for example, properly implementing access control. Orthogonally, to ensure governance and safety of these systems, proper testing suits are needed. We review ARES, IBM’s open source library. On the other hand, end-users face different dilemmas when interacting with LLMs. They are often unaware of pitfalls and engage, potentially unaware, in security-relevant behavior. Also, about a third of the end-users admitted to jailbreaking their chatbots. However, this behavior is not malicious but motivated by play and curiosity about the limitations of the new, end-user-friendly technology. Lastly, we briefly discuss AI security incident reporting as a possible solution to the open questions raised in both areas.

References

- 1 Kathrin Grosse, Nico Ebert: Prevalence of Security and Privacy Risk-Inducing Usage of AI-based Conversational Agents. CoRR abs/2510.27275 (2025)

4 Working Groups

During a match-making session on Day 2, the interests expressed by the participants were consolidated into a set of six working groups, each addressing a distinct theme related to the security, privacy, and safety of LLMs and AI-enabled systems.

The topics discussed in the subsequent break-out sessions were tailored to the specific interests of the participants in each working group and were selected through informal topic-selection discussions during the seminar. The following themes and underlying open questions are representative of the discussions that emerged across the groups:

- **AI system capabilities certification.** How can the capabilities and limitations of AI systems be systematically characterized and certified? What forms of assurance or guarantees are meaningful for complex, adaptive systems such as LLMs?
- **LLM manipulation on humans.** To what extent can LLMs influence, persuade, or manipulate users, intentionally or unintentionally? How should such risks be modeled, measured, and mitigated in real-world deployments?
- **Privacy risks of LLMs.** What new privacy threats arise from training, fine-tuning, and interacting with LLMs? How can risks such as memorization, inference, and data leakage be systematically assessed and reduced?
- **Attack surface of LLMs.** How can the attack surface of LLM-based and AI-enabled systems be decomposed and analyzed at the system level, including components such as prompts, tools, agents, external services, and data pipelines?
- **LLM safety and security benchmarks.** What benchmarks are needed to meaningfully evaluate the safety and security of LLMs beyond traditional accuracy metrics? How can benchmarks remain relevant in the face of rapidly evolving models and attack techniques?

- **Practical implications and deployment challenges.** How do all these concerns translate to real-world applications, and what trade-offs arise between security, privacy, safety, usability, and performance?
- **Societal impact of LLMs.** How may the widespread deployment of LLMs affect the functioning of society, including economic structures, legal systems, academic peer review, and democratic processes? Are the existing AI regulatory frameworks able to govern their development and deployment? In parallel, how might LLMs themselves become tools within legal and institutional systems, and what risks arise from such integration?
- **Concentration of AI capabilities.** How does the growing concentration of cutting-edge LLMs and AI infrastructure within a small number of large technology companies affect research, innovation, and societal power dynamics? What mechanisms could help democratize access to models, data, and compute so that academia and broader society can meaningfully participate in the development and governance of advanced AI systems?

While the seminar could not fully address all of the questions outlined above, identifying and articulating them was itself an important outcome of the event. The discussions highlighted these issues as central challenges for the research community and for society at large, helping to frame a broader agenda for future work in the area of LLM safety and security.

Among the posed themes, six were explored in greater depth through dedicated working groups, each of which produced a written report summarizing their findings and perspectives. In the remainder of this document, the working group reports on the following topics are presented:

- AI system capabilities certification;
- LLM manipulation of humans;
- Privacy risks of LLMs;
- Responsible usage of LLMs in peer review;
- Modeling the attack surface of LLM- and AI-enabled systems;
- LLM safety and security benchmarks.

4.1 AI System Capabilities Certification

Tobias Callies (Universität der Bundeswehr München, DE)

Jonathan Petit (Qualcomm – Boxborough, US)

Mathilde Aliénor Raynal (EPFL Lausanne, CH)

Eric Wong (University of Pennsylvania – Philadelphia, US)

License © Creative Commons BY 4.0 International license
© Tobias Callies, Jonathan Petit, Mathilde Raynal, and Eric Wong

Models are being released or deployed at a record pace. Nowadays, users are exposed to chatbots on most websites, while developers must decide which pre-trained model to deploy or fine-tune. This explosion of models makes it difficult to decide which one to use (from a user or developer standpoint), or more importantly, which one to trust. Indeed, how can a user (or an AI agent) be aware of the security, privacy, and safety capabilities of the AI system she interacts with? We propose a solution to increase the transparency and trust in AI systems. More specifically, the proposed solution should aim to:

- Increase transparency for the user about the AI system/model capabilities using state-of-the-art Human-Computer-Interaction techniques.

- Create accountability (for the model deployer).
- Enable “model shopping” and specialized models.

As we noted that existing certification regimes helped improving the reliability and cybersecurity of certified devices (e.g., RED certification of wireless devices, U.S Cyber Trust Mark, ISO 21434 for automotive), we devised a certification process that would output a AI trustworthiness label.

4.1.1 Methodology

This section briefly explains the methodology followed by the group.

Definition. First, we defined the term “capability” of an AI system. As the focus of this seminar is LLM security, we focused on security-related capabilities such as “prompt injection protection”.

Process. Then, we specified a three-step certification process:

- Step 1 (Define): Specify AI system task(s) and harm(s). Tasks define the responsibilities and actions an AI system is expected to perform. Harms define and categorize potential threats to protection-worthy assets.
- Step 2 (Measure): Specify the evaluation process and parameters. For example, use a population of AI agents that is representative of the target population. The objective is to assess the probability of being exposed to the harm(s) defined in Step 1. We proposed a badge system that would enable users to understand the risk of using the system under test.
- Step 3 (Update): Any respectable certification process must be updated to reflect the ecosystem.

4.1.2 Next Steps

After presenting to the entire seminar’s participants, we gathered extensive feedback and interest in writing a concept paper. We will work towards that goal.

4.2 On LLM manipulation of humans

Lujo Bauer (Carnegie Mellon University – Pittsburgh, US)

Rebekka Burkholz (CISPA – St. Ingbert, DE)

Lorenzo Cavallaro (University College London, GB)

Andreas Happe (TU Wien, AT)

Pavel Laskov (Universität Liechtenstein – Vaduz, LI)

Nils Lukas (MBZUAI – Abu Dhabi, AE)

Hammond Pearce (UNSW – Sydney, AU)

Lea Schönherr (CISPA – Saarbrücken, DE)

License © Creative Commons BY 4.0 International license

© Lujo Bauer, Rebekka Burkholz, Lorenzo Cavallaro, Andreas Happe, Pavel Laskov, Nils Lukas, Hammond Pearce, and Lea Schoenherr

Humans have always sought to influence one another, toward both positive and negative ends. Political leaders campaign for the support of their citizens. Suitors emphasise their most attractive qualities. Scammers and criminals attempt to manipulate victims into surrendering valuable assets.

Historically, such influence campaigns were constrained by the inherent capabilities and resources of the influencing party. A single scammer could reach only a small number of targets at a time. Large-scale influence required large organisations, often at the scale of governments (e.g., mass propaganda, nation-level education policy).

With the rise of generative AI (GenAI), these constraints are fundamentally changing. Influencers can now scale their campaigns both in breadth (tailoring highly customised messages to many more targets) and in depth (conducting extensive background research on individuals or groups autonomously). Further, GenAI-based influence needs not be bounded by human limitations such as tiredness, empathy, or guilt. In our discussion, we examine a taxonomy of GenAI-based influence, pose open questions for the research community, and offer our perspective on what the future may hold for autonomous influence campaigns.

In interactions between humans and an LLM-driven influencer, communication can unfold in multiple ways. It may follow an intentional solicitation (e.g., a user engaging with a chatbot) or occur without the human ever realising that an AI system is involved. Influence can also take place over multiple turns or within a single exchange. Depending on the application, the agent may be granted access to additional tools and information that amplify its influence – for example, by retrieving external information or personal data about the human target.

There is therefore a large design space to consider when thinking about the many ways humans may be influenced by LLMs. We organise our thoughts into four questions for future research:

4.2.1 Q1: On Influencing Humans – What are the properties of content that best influence humans in this setting?

We already know some common patterns that shape how people respond. Messages that appear trustworthy, personalised, and context-aware can lower suspicion and increase engagement. Emotional triggers such as fear, sympathy, outrage, or excitement can drive rapid decisions. Creating a sense of urgency or stress can cause people to act before they have time to reflect. Decomposing a questionable request into small, seemingly harmless steps (a “foot-in-the-door” strategy) can make it easier for someone to agree without noticing the cumulative effect. Understanding these patterns in the context of GenAI is crucial so that people, platforms, and policymakers can better recognise and resist manipulation.

4.2.2 Q2: LLM-based Capabilities – What could LLMs enable in this setting that is infeasible/uneconomical for pure human/non-LLM influencers?

LLMs enable unprecedented scalability, allowing attackers and other influencers to dramatically increase the number of people they can target, including in semi-autonomous ways. This efficiency can make small or low-value operations economically viable, because generating customised content and interactions becomes extremely cheap. LLMs can also deepen the quality and speed of reconnaissance, rapidly gathering and analysing large volumes of unstructured data to tailor messaging, timing, and strategy to individual targets or micro-groups. Even partial automation can substantially increase reach and impact. Unlike humans, LLMs do not become tired, demotivated, remorseful, or empathetic; if misused, they can sustain persistent, high-volume activity without emotional or cognitive limits.

4.2.3 Q3: LLM-powered Influencers – How is an LLM-powered “influencer” best designed to enable Q1/Q2?

Designing an LLM-powered “influencer” raises difficult technical, ethical, and legal questions. We must define what counts as “influence” and how to measure it – for example, changes in attitudes, short-term engagement (clicks, shares), or longer-term behaviour. We must also account for who is being influenced, the context of the interaction, the time and effort required to influence them, and realistic limits on how far automated systems can shape beliefs and actions.

“Good” or prosocial influencers should reject mis- and disinformation and favour honest, transparent communication. Their design must incorporate robust legal and ethical safeguards that respect autonomy and rights, avoid covert manipulation, and support informed, voluntary choices. By contrast, “bad” influencers (such as those deployed by scammers or state actors with malicious intent) will not be constrained by such norms, and may optimise solely for desired, potentially harmful outcomes. Understanding these unconstrained designs is essential for anticipating threat models and developing effective mitigations and defenses.

4.2.4 Q4: Mitigation – How can LLM-powered influencers be mitigated?

Mitigating the risks of LLM-powered influencers is challenging because it requires understanding and measuring their impact on human cognition and behaviour. While LLMs sometimes exhibit recognisable patterns – such as repetitive phrasing, overly consistent tone, or specific stylistic signatures – these signals are imperfect and may diminish as models improve.

Addressing these risks will likely require a combination of technical, institutional, and societal responses. On the technical side, stronger methods are needed for detecting and tracing influence, including behavioural signals, metadata standards, platform-level transparency mechanisms, and clearer audit trails. Techniques such as watermarking, provenance tracking, and prompt extraction (e.g., retrieving a model’s system prompt or high-level intentions) may play supporting roles. On the societal side, increasing information diversity, investing in public education, and strengthening digital literacy can all help build resilience to manipulation. Finally, any defensive approach must explicitly account for the dual-use nature of LLM influencers: the same tools can be deployed by both “good” and “bad” actors, which complicates which interventions are feasible, effective, and acceptable.

4.3 Privacy risks of LLMs

Giovanni Cherubin (Microsoft Research – Cambridge, GB)

Ana-Maria Cretu (EPFL – Lausanne, CH)

Stephan Günnemann (TU München – Garching, DE)

Qianying Liao (KU Leuven, BE)

Vera Rimmer (KU Leuven, BE)

Ali Satvaty (University of Groningen, NL)

License © Creative Commons BY 4.0 International license

© Giovanni Cherubin, Ana-Maria Cretu, Stephan Günnemann, Qianying Liao, Vera Rimmer, and Ali Satvaty

Large Language Models (LLMs) introduce numerous privacy challenges, some of which were known in traditional Machine Learning (ML) systems, others that are entirely novel. We discuss ways in which personal data of individuals can be misused in the lifecycle of LLMs,

ranging from data leakage and sensitive inferences to LLMs influencing the behavior of users. We also map out how LLMs differ from traditional ML applications, and how those differences can scale up existing threats or introduce new privacy harms.

4.3.1 Data leakage from LLMs

LLMs risk leaking the data of users on which they were trained. There are four different stages in the LLM lifecycle where they ingest data.

In the **pre-training stage**, LLMs are trained on vast amounts of publicly available personal data. It is debatable whether the leakage of personal data about a user that was published with their consent constitutes a privacy violation. However, in many cases personal data is online without the consent of the user, or the user may decide to remove their online data after it was scraped and used in LLM training; this is a great source for concern. A further concern is the potential increase in the ability to profile the users or infer sensitive information about them, as compared to not having been trained at all on their data.

In the **fine-tuning stage**, LLMs are trained on smaller and domain-specific datasets (e.g., medical notes, transaction data, customer chats). Fine-tuning on private data can amplify privacy exposure by making the model more aware of individuals or groups, and may lead to increased leakage about individuals due to the smaller dataset size.

In the **in-context learning stage**, private data can be embedded directly in prompts or system instructions. This data may leak in subsequent user interactions, and it is particularly hard to prevent this from happening.

To limit the risks of data leakage, the ML privacy community has built a solid understanding of the various attacks and defenses for traditional ML (supervised learning). Yet, this research is not always directly applicable to LLMs, as they differ from classic ML in several respects. For example, LLMs are trained as general purpose models, which increases the attack surface. Further, it is harder to audit their privacy, as this requires to contextually determine the “privacy unity”, e.g., whether to protect individual documents, all the data belonging to a user, or specific secrets that are shared across users.

Beyond training data leakage, there is a risk of data exfiltration attacks in the **AI agentic stage**, where the LLM can access user files and integrate with external systems. The boundaries between stored data, model memory, and ephemeral inference become increasingly blurred. This complicates GDPR-style risk assessment, transparency obligations, and expectations about how long data persists, whether deletion is respected, and what constitutes a privacy unit in LLMs.

4.3.2 Privacy harms from interactions between a user and the LLM

Privacy harms may arise from sharing information with a system that mimics trustworthiness. For example, users increasingly disclose intimate details to LLM-powered chatbots, far more than they would to friends or doctors. Based on these interactions, LLM-powered chatbots can create internal profiles of users on inferred attributes, preferences, vulnerabilities, or political leanings. These inferences can persist across conversations (e.g., ChatGPT’s memory feature), shape future responses, and influence user’s behavior. Even if the model does not exploit this information, profiling raises questions about autonomy, manipulation, and privacy as freedom from interference (e.g., Article 12 GDPR, UN privacy rights).

4.3.3 LLMs as privacy attackers

At scale, LLMs can also function as powerful tools to help attackers violate people's privacy, by exploiting linkages and contextual inferences. LLMs can aggregate data from web browsing tools, API calls, and user history, enabling attacks that combine inference, linkage, and manipulation. LLMs can also synthesize facts from multiple data sources to infer information about individuals from public fragments, e.g., link a user's LinkedIn to anonymous postings or guessing medical, financial, or political attributes. Unlike search engines, LLMs generate new spans of text, making the disclosure harder to track and validate, especially since ground truth is unavailable for many inferred secrets. Even if the inferred information is "public", the act of connecting, aggregating, or contextualizing can scale the privacy harms that previously required a significant human effort.

4.3.4 LLMs as the privacy defender

Paradoxically, LLMs may be a better tool to defend privacy than traditional ML systems, because of their ability to understand natural language and context; crucially, this sets them apart from standard privacy techniques (e.g., DP mechanisms). One example is text anonymization, which typically requires removing personally identifiable information and paraphrasing to remove secrets. LLMs could enforce personal-data minimization, detect exfiltration attempts, refuse risky actions, or provide dynamic privacy explanations to users. However, realizing this defensive potential requires solving major open problems: defining what privacy means in generative contexts, creating measurable risk metrics, distinguishing intended vs. non-intended behavior, and ensuring that safeguards cannot be bypassed.

4.3.5 Conclusions

These discussions highlight that privacy in LLMs goes beyond traditional ML concerns: from data leakage at various training stages, to inference-based profiling and large-scale aggregation. Interestingly, LLMs can also offer novel means to enhance privacy, e.g., via more contextually aware anonymization. Further research is essential to adequately describe privacy threats, clarify acceptable use contexts, prevent misuse, and to harness the potential privacy benefit that LLMs may offer.

4.4 Responsible Usage of LLMs in Peer Review

Giovanni Apruzzese (Universität Liechtenstein – Vaduz, LI)

Amir Houmansadr (University of Massachusetts Amherst, US)

Marc Juarez (University of Edinburgh, GB)

Florian Kerschbaum (University of Waterloo, CA)

Wojciech Kusa (NASK – Warsaw, PL)

Fabio Pierazzi (University College London, GB)

Konrad Rieck (BIFOLD – Berlin, DE & TU Berlin, DE)

Christian Wressnegger (Karlsruher Institut für Technologie, DE)

License © Creative Commons BY 4.0 International license

© Giovanni Apruzzese, Amir Houmansadr, Marc Juarez, Florian Kerschbaum, Wojciech Kusa, Fabio Pierazzi, Konrad Rieck, and Christian Wressnegger

Large language models (LLMs) are increasingly embedded in scientific research practices. Their use for writing assistance, programming, and exploratory analysis is now widespread and well documented [1, 2, 3, 4, 5]. More recently, LLMs have begun to influence peer review, raising concerns regarding accountability, fairness, quality, and policy compliance [6, 7]. Empirical analyses have revealed that, following the release of ChatGPT, the reviews submitted for ICLR’23 included a substantially higher number of words typically associated to LLMs [8]. In response, venues and publishers have adopted policies ranging from explicit prohibitions on LLM use [9] to controlled LLM integration into the review process [10], as well as revising their respective authorship and reviewing guidelines [11, 12].

This working group examined “responsible LLM usage in peer review” from a security-oriented perspective. The guiding assumption was that irresponsible LLM usage (e.g., delegating substantial intellectual tasks without human oversight, or violating explicit venue policies) can compromise the integrity of the scientific process. The intention was to discuss how security mechanisms, whose real-world objective is more often to discourage rather than complete prevention, can help in addressing this issue of “irresponsible” usage of LLM in peer review [13, 14].

4.4.1 Discussed Problems

The discussion mostly concerned the detectability of LLM-generated content, particularly LLM-written reviews. Participants broadly agreed that they had encountered such reviews and that these were often unhelpful or incorrect. The group discussed several detection strategies:

- Canary-based approaches, such as including hidden prompts in papers to reveal whether reviewers submit papers to LLMs, have been proposed in prior work [15]. However, such approaches rely on security through obscurity and are easily circumvented once known.
- Automated detectors and watermarking techniques (both academic and commercial [16, 17, 18, 19]) were also discussed. These methods were viewed as fundamentally unreliable due to false positives (including flagging “benign” uses, such as language polishing) and ease of evasion through minor editing or prompt engineering.
- More elaborate strategies, such as finding the prompt that generates a given review, were deemed impractical, due to the intrinsic difficulty in finding a prompt without falling into subjective choices, as well as raising potential confidentiality issues (since the process involves submitting the prompt to an LLM alongside the reviewed paper – which can be done by the authors of the paper, but not by, e.g., PC members during a confidential peer-review process).

The group therefore converged on the conclusion that robust detection of LLM-generated reviews is not currently feasible.

The group also questioned whether LLM usage in scientific contexts is intrinsically problematic. Opinions varied considerably. In the context of peer review, three use cases were examined:

- First, using LLMs for language polishing or grammar was viewed by some as not necessary, while others argued that improved clarity benefits authors and reviewers alike [5].
- Second, generating reviews with LLMs – either partially or entirely – was considered unacceptable by many participants, as reviewing is meant to be done by “peers” (i.e., humans); moreover, reliance on LLMs is disrespectful towards other reviewers who spend considerable time for their reviewing duties. Others, however, argued that what really matters is the quality of a review, rather than the process by which reviews are produced: if LLMs can lead to better reviews, their usage would be justified.
- Third, the use of LLMs to analyse the submitted reviews (an application evaluated by the organizers of ICLR’25 [10]) was viewed as potentially beneficial. However, some concerns were still raised about the practical utility of LLM feedback and the difficulty of objectively measuring improvements in review quality.

The group also discussed LLM usage in writing scientific articles. Using LLMs for editing and stylistic refinement was generally considered acceptable. However, generation of content from scratch raises concerns such as hallucinated references and factual inaccuracies [20]. This problem is aggravated by the fact that identifying such errors is challenging for reviewers, who are not typically expected to verify every citation and its surrounding context. Participants emphasized the need for automated mechanisms to support reference checking.

The group also discussed LLM usage by students, particularly PhD candidates. Participants agreed that LLM use in graded coursework (e.g., for exams involving use of internet-capable devices, or for homework) was problematic. However, greater concern was expressed about “undeclared” usage of LLMs for writing research documents, given that supervisors and co-authors remain accountable for any errors introduced. The importance of educating early-career researchers about the pros-and-cons of LLMs in scientific work was strongly emphasized.

Disclosure of LLM usage in peer-review contexts generated further disagreement. From the author side, some venues explicitly require authors to disclose LLM usage (e.g., SaTML’26, or S&P’26); however, it is unclear if authors truly comply with such requirements (participants reported accounts of clearly LLM-written content in papers they reviewed which had not been acknowledged). From the reviewing side, disclosure may also undermine trust in reviews, even when LLMs are used responsibly. In either case, while disclosure could increase transparency, it is uncertain how to ensure that researchers comply with disclosure requirements.

4.4.2 Possible Approaches

Despite divergent views, a shared conclusion emerged: the quality, correctness, and accountability of scientific content matter more than the tools used to produce it. Based on this principle, several recommendations were formulated.

- Reviewers should remain engaged beyond submitting their own reviews by critically examining other reviews and correcting factual errors when possible. If LLMs are (responsibly) used, any time savings should be reinvested in enhancing the overall quality of the peer-review, which goes beyond the mere act of submitting a review, since it also encompasses discussions with other reviewers.

- The community should also place greater value on high-quality reviewing by strengthening incentive and recognition mechanisms, as well as pointing out cases of “policy violations” wherein some researchers (whose identity should not be disclosed) received some penalty due to illegitimate usage of LLMs.
- Finally, more education, especially for young researchers, is needed to foster a nuanced understanding of LLM capabilities, limitations, and responsibilities.

4.4.3 Conclusion

Overall, responsible LLM usage in peer review cannot be guaranteed via detection alone. Instead, preserving review quality requires clearer norms, better incentives, targeted education, and accountability mechanisms that discourage irresponsible behavior while acknowledging the practical role LLMs may play.

References

- 1 A. Birhane, A. Kasirzadeh, D. Leslie, and S. Wachter, “Science in the age of large language models,” *Nature Reviews Physics*, vol. 5, no. 5, pp. 277–280, 2023.
- 2 M. Binz, S. Alaniz, A. Roskies, B. Aczel, C. T. Bergstrom, C. Allen, D. Schad, D. Wulff, J. D. West, Q. Zhang *et al.*, “How should the advancement of large language models affect the practice of science?” *Proceedings of the National Academy of Sciences*, vol. 122, no. 5, p. e2401227121, 2025.
- 3 J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J.-N. Eckardt, N. G. Laleh, C. M. L. Löffler, S.-C. Schwarzkopf, M. Unger, G. P. Veldhuizen *et al.*, “The future landscape of large language models in medicine,” *Communications medicine*, vol. 3, no. 1, p. 141, 2023.
- 4 M. Naddaf, “How are researchers using AI? Survey reveals pros and cons for science,” *Nature*, 2025, <https://www.nature.com/articles/d41586-025-00343-5>.
- 5 M. Geng, C. Chen, Y. Wu, Y. Wan, P. Zhou, and D. Chen, “The impact of large language models in academia: from writing to speaking,” in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 19 303–19 319.
- 6 J. Zou, “ChatGPT is transforming peer review – how can we use it responsibly?” *Nature*, vol. 635, no. 8037, pp. 10–10, 2024, <https://www.nature.com/articles/d41586-024-03588-8>.
- 7 G. O’Brien, “How scientists use large language models to program,” in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–16.
- 8 W. Liang, Z. Izzo, Y. Zhang, H. Lepp, H. Cao, X. Zhao, L. Chen, H. Ye, S. Liu, Z. Huang *et al.*, “Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews,” in *International Conference on Machine Learning*. PMLR, 2024, pp. 29 575–29 620.
- 9 CVPR, “CVPR 2025 Reviewer Guidelines,” 2025, <https://cvpr.thecvf.com/Conferences/2025/ReviewerGuidelines>.
- 10 N. Thakkar, M. Yuksekgonul, J. Silberg, A. Garg, N. Peng, F. Sha, R. Yu, C. Vondrick, and J. Zou, “Can LLM feedback enhance review quality? a randomized study of 20k reviews at ICLR 2025,” *arXiv:2504.09737*, 2025.
- 11 ACM, “ACM Policy on Authorship,” 2025, <https://www.acm.org/publications/policies/new-acm-policy-on-authorship>.
- 12 – – , “ACM Peer Review Policy Frequently Asked Questions,” 2025, <https://www.acm.org/publications/policies/peer-review-faq>.
- 13 V. M. Bier, “Choosing what to protect,” *Risk Analysis: An International Journal*, vol. 27, no. 3, pp. 607–620, 2007.
- 14 L. A. Gordon and M. P. Loeb, “The economics of information security investment,” *ACM Transactions on Information and System Security (TISSEC)*, vol. 5, no. 4, pp. 438–457, 2002.

- 15 M. G. Collu, U. Salviati, R. Confalonieri, M. Conti, and G. Apruzzese, “Publish to perish: Prompt injection attacks on LLM-assisted peer review,” *arXiv preprint arXiv:2508.20863*, 2025.
- 16 Pangram, “AI Detection that actually works,” 2025, <https://www.pangram.com/>.
- 17 M. Naddaf, “Major AI conference flooded with peer reviews written fully by AI,” *Nature*, 2025, <https://www.nature.com/articles/d41586-025-03506-6>.
- 18 R. Zhang, S. S. Hussain, P. Neekhara, and F. Koushanfar, “REMARK-LLM: A robust and efficient watermarking framework for generative large language models,” in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 1813–1830.
- 19 T. Lee, S. Hong, J. Ahn, I. Hong, H. Lee, S. Yun, J. Shin, and G. Kim, “Who wrote this code? watermarking for code generation,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 4890–4911.
- 20 P. Esau, A. Cui, and A. Adam, “GPTZero finds over new 50 hallucinations in ICLR 2026 submissions,” 2025, <https://gptzero.me/news/iclr-2026/>.

4.5 Modeling the Attack Surface of LLM- and AI-enabled Systems

Hyrum Anderson (Cisco Systems – San Jose, US)

Javier Carnerero Cano (IBM Research – Dublin, IE)

*Dinil Mon Divakaran (Institute for Infocomm Research, A*STAR – Singapore, SG)*

Mario Fritz (CISPA – Saarbrücken, DE)

Emil C. Lupu (Imperial College London, GB)

Irina Nicolae (Apple – Paris, FR)

Tim Van hamme (KU Leuven, BE)

License © Creative Commons BY 4.0 International license

© Hyrum Anderson, Javier Carnerero Cano, Dinil Mon Divakaran, Mario Fritz, Emil C. Lupu, Irina Nicolae, and Tim Van hamme

The rapid integration of large language models (LLMs) into different applications across various enterprise sectors has led to new and complex security challenges, as LLMs have inherent weaknesses such as non-deterministic output, hallucinations, and privacy leakages. Although recent research has focused largely on model-level issues such as prompt injection, the field lacks a systematic characterization of threats that emerge when LLMs are embedded in real applications, particularly with respect to implications such as business risks, motivation and effort estimates, both from an attacker and defender perspective. In this work, we attempt to fill this gap by (i) proposing a risk assessment framework for modeling the attack surface of LLM systems and (ii) mapping existing defense strategies to their roles and impacts from a system-level viewpoint.

4.5.1 Impact-oriented LLM Application Threat Modeling Framework

We outline an impact-oriented threat-modeling framework for LLM applications that connects business risks to system-level vulnerabilities. The framework integrates four elements: (i) a business-impact analysis that identifies critical assets and evaluates the consequences of their compromise; (ii) an attack-surface modeling process that derives concrete system vulnerabilities from high-level impact scenarios; (iii) a characterization of threat sources, spanning highly capable adversaries (e.g., nation-state actors) to insiders and accidental actors; and (iv) a lightweight risk-scoring method that combines scenario likelihood with impact severity. Together, these components provide a structured approach for assessing and prioritizing risks in LLM-enabled systems.

4.5.2 Mitigating the Risks

Risk-mitigation decisions can be guided by two complementary system-level perspectives. The weakest-link view assesses risk along the easiest attack path, revealing the objectives and vectors an adversary is most likely to exploit. The full attack-surface view aggregates risk across all feasible paths, highlighting overall exposure and capturing cases where multiple comparable or cumulative attack routes create material risk even when no single path dominates.

The proposed framework formalizes how impacts, vulnerabilities, threat sources, and risk metrics interrelate, providing a structured basis for prioritizing defenses and guiding secure system design. In doing so, it offers a path for organizations to assess, communicate, and mitigate the real-world risks introduced by LLM integration. More broadly, the framework is intended to equip practitioners with practical guidance for securing LLM-enabled systems and to help researchers identify and prioritize high-impact directions in LLM security and privacy.

4.6 LLM Safety & Security Benchmarks

Maksym Andriushchenko (ELLIS – Tübingen, DE)

Marc Fischer (Snyk – Zürich, CH)

Kathrin Grosse (IBM Research Europe – Zürich, CH)

Daphne Ippolito (Carnegie Mellon University – Pittsburgh, US)

Andrew Paverd (Microsoft Research – Cambridge, GB)

Advije Rizvani (Universität Liechtenstein – Vaduz, LI)

Ilia Shumailov (AI Security Company – London, GB)

License © Creative Commons BY 4.0 International license

© Maksym Andriushchenko, Marc Fischer, Kathrin Grosse, Daphne Ippolito, Andrew Paverd, Advije Rizvani, and Ilia Shumailov

This working group investigated the challenges in evaluating the trustworthiness (specifically security and safety) of current state-of-the-art large language models (LLMs), AI applications, defenses meant to augment them, and attacks considered to break them. This spans the robustness of models to jailbreaks, prompt injections by users, indirect prompt injections via tools or third party actors as well as ways to defend against these. We outline current issues and summarize our recommendations for further benchmarks and evaluations.

The ubiquitous use of large language models in virtually all areas of life and to dynamically drive the behavior of computer programs (“AI Applications” and “AI Agents”) raises questions about their trustworthiness and how to establish it. Key among these concerns are safety and security, e.g.

Safety Entails, among other things, the alignment of the underlying AI model and the overall AI application, i.e., whether a model will refuse to engage in hate speech or to explain or enact dangerous practices (e.g. “How to build a Bomb?”). While these properties are often established at training time, this initial alignment is the often circumvented by jailbreaks [9] and/or direct prompt injections [8].

Security Assesses whether the AI model or AI application can engage in dangerous or destructive behavior by hijacking it’s behavior, e.g. leaking or deleting information, in some way. This is typically achieved via indirect prompt injections.

Typically, these properties are best measured using benchmarks [10, 11]. In this work, we outline the current challenges and shortcoming of these benchmarks, our suggestions to overcome these along with open challenges.

4.6.1 Benchmarking AI Security & Safety

An enduring cycle of incremental advancements in AI defences characterizes the current landscape of AI security literature [2, 1, 5]. This research typically follows a predictable three-part loop: the specification of a system evaluation, the development of attacks against that system, and the subsequent creation of defenses [3]. There is the cycle of defenses and attacks each getting better as newly devised attacks are tailored to exploit the specific mechanisms of defenses [2, 1, 5], or evaluations that showed the performance of defense were flawed [4]. In either case, new defenses are proposed to remove the uncovered mechanisms. There is also the larger cycle of researchers realizing that the specification of the problem itself is flawed [3], and the attacks and defenses being created are insufficiently relevant to real-world systems [6, 7]. This all perpetuates a reactive and iterative sequence of “Specification → Attack → Defence → ... → Attack → Defence.”

We argue that the non-determinism in LLMs and the fact that advancements in LLMs are typically interoperable with their predecessors lead to a rapid and low-reproducibility development cycle for AI Security & Safety compared to other areas of Cyber Security. LLMs are increasingly being incorporated into the specification of the problem (e.g., the LLM underlying the agent in an agentic system), the attacks and defense methods, and the evaluation of attack success (e.g., LLM-as-judge methods). The behaviour of LLMs is non-deterministic—the same prompt can yield different generations, and small changes to the prompt can result in significant changes to the generations. This means, all the components which rely on LLMs are also non-deterministic in their behaviour. Moreover, it is usually trivial to swap out LLM x with its successor LLM $x + 1$, or with an entirely different LLM. This means that:

- Updated LLMs may be inherently less vulnerable to attacks without the need for post-hoc defenses. They may have also been trained on the evaluation scenario, resulting in inflated confidence in scenarios being “solved.”
- Attack and defense development becomes faster when there are so many “easy” parameters that can be changed that may significantly improve attack/defense success.

This paradigm is further complicated by the “noisy” nature of evaluating the efficacy of these attacks and defenses, which are often assessed in limited and simplistic “Rosetta Stone” environments, for example with attacks that only exceed with low probability. Consequently, many proposed attacks and defenses offer only short-term value, quickly losing their relevance as they are over-fitted to environments that are challenging and time-consuming to build and maintain. At the same time, environment data is easily incorporated into training future models, invalidating the assessment.

We argue that the only way to break this inefficient cycle, the research community must shift its focus towards two more fundamental goals: the creation of attacks that can generally break existing defenses and the development of defenses that are secure by design. We contend that progress in either of these areas is contingent upon incentivizing the development of meaningfully diverse, realistic, and dynamic evaluation environments. Such environments are critical for rigorously assessing the generalizability of defensive measures and the true potency of attack vectors, thereby fostering more robust and lasting advancements in AI Security & Safety.

Some proposed attacks only succeed with a low probability, limiting their practical impact. Consequently, many of these meticulously crafted attacks and defenses offer only short-term value, quickly losing their relevance as they are overfitted to static evaluation environments that are both challenging and resource-intensive to build and maintain. The problem is exacerbated by the fact that data from these evaluation environments can be easily incorporated into the training of future models, thereby invalidating the very assessments they were designed for. This is a critical issue, as companies are incentivized to “train on the evals,” creating a situation where the benchmarks for security become moving targets that are constantly being gamed.

4.6.2 Strong Attacks, Secure by Design

To break this inefficient and unsustainable cycle, the research community must fundamentally shift its focus toward two more foundational goals: the creation of attacks that can generally break all existing defenses and the development of defenses that are secure by design. Progress in either of these ambitious directions is contingent upon a critical prerequisite: the development of diverse, realistic, and dynamic evaluation environments. Such environments are paramount for rigorously assessing the generalizability of defensive measures and the true potency of attack vectors, thereby fostering more robust and lasting advancements in AI security.

Upon deeper investigation into the complexities of this challenge, it becomes apparent that the evaluation of an AI system is not a monolithic task. A fundamental tension exists between evaluating the system as a whole versus assessing its individual components. Each component possesses unique semantics, constraints, and its own input-output domains, making a holistic evaluation incredibly difficult. The problem is compounded by the fact that it is often impossible to decompose a single model inference into its sub-components for individual evaluation. This challenge is further magnified when considering the trade-offs between evaluation time and attack efficiency. A thorough evaluation may require significant computational resources and time, which may not be practical in rapidly evolving threat landscapes.

4.6.3 Dynamic Efficient Benchmarks

Furthermore, the nature of attacks themselves is context-dependent. For instance, the relevance of “imperceptible” attacks from the old adversarial example literatures, which are subtle manipulations to inputs that are not easily detectable by humans, is not a universal constant. For many attacks, imperceptibility is a property of the context in which the model appears, such as within the structure of a markdown document or as invisible text. This highlights the limitations of evaluations conducted in sterile, context-free environments. Similarly, many current evaluations operate with restricted attack budgets, limiting the number of dimensions an attacker can manipulate. In reality, attackers may have unrestricted budgets, allowing for far more sophisticated and multi-faceted attacks. A crucial insight is that many successful attacks are, in fact, functionally exploiting the defenses themselves; they target the interpretive nature of the defenses, turning their own logic against them. This is particularly true for adaptive attacks, which are tailored to the specific defenses in place. The need for more robust evaluation is also underscored by the increasing prevalence of AI in multi-modal domains, including audio and vision, which present new and more continuous domains for potential attacks. At the same time, such expensive attacks, compared with the need to evaluate across various and wide settings, can make the benchmarking and

development of new attacks prohibitively expensive. In an adhoc exploration we found many datapoints in existing benchmarks to be highly correlated (succeeding or failing jointly). We thus call for benchmarks to provide representative core sets for quick evaluation and encourage the exploration of methods to determine these automatically.

The current paradigm also struggles with the comparative evaluation of different defense systems. While it is understood that defenses need to be evaluated using strong, adaptive attacks, these attacks are, by their very nature, specific to a particular defense and environment. This specificity makes it incredibly challenging to compare the relative strengths and weaknesses of different defensive strategies in a generalized manner. Without the ability to conduct meaningful comparative evaluations, the community risks investing in defenses that are only effective against a narrow and perhaps irrelevant set of attacks.

4.6.4 Recommendations

To conclude this working group report, we outline our recommendations based on the above discussion. Above all we encourage the development of secure/safe-by-design approaches and strong adaptive attacks along with efficient, comprehensive, dynamic, interchangeable evaluation environments. In particular, benchmarks need to be dynamic in the sense of (i) multiple adaptive versions of attacks are run to deal with the stochasticity of LLMs, (ii) while the scaffolding of benchmarks is static the concrete content changes dynamically to avoid being trained on by foundation models, (iii) to further mitigate overfitting, new benchmarks are continually developed and published. To facilitate this, we call for the reuse of standardized eval hardnesses in which attacks and defenses can be deployed in a plug-and-play manner. We find datapoints within existing benchmarks to be highly correlated. Thus, we need to be able to efficiently discover the most insightful datapoints and run the evals on them. While it is crucial to test full AI applications and AI agents, often the key vulnerability can be broken down into a single jailbreak of prompt injection. We believe that benchmarks that highlight both these pathological cases as well as the full environments, can help make progress at a rapid scale. We further encourage investigation into quantifying attack budgets with respect to cost, time and imperceptibility constraints and welcome arguments to the required economics of AI safety and security, though we caution to artificially limit benchmarks with these arguments.

Our goal with these measures is to drive the creation of safe and secure AI systems by making it easy to evaluate newly proposed systems against the current state-of-the-art adaptive and generic attacks.

References

- 1 Nicholas Carlini. Cutting through buggy adversarial example defenses: fixing 1 line of code breaks Sabre. arXiv preprint arXiv:2405.03672, 2024.
- 2 Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.
- 3 Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems*, volume 33, pages 1633–1645, 2020.
- 4 Maura Pintor, Luca Demetrio, Angelo Sotgiu, Ambra Demontis, Nicholas Carlini, Battista Biggio, and Fabio Roli. Indicators of attack failure: Debugging and improving optimization of adversarial examples. In *Advances in Neural Information Processing Systems*, volume 35, pages 23063–23076, 2022.

- 5 Reza Shokri et al. Bypassing backdoor detection algorithms in deep learning. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 175–183, 2020.
- 6 Giovanni Apruzzese, Hyrum S. Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. “real attackers don’t compute gradients”: bridging the gap between adversarial ML research and practice. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 339–364, 2023.
- 7 Kathrin Grosse, Lukas Bieringer, Tarek R. Besold, and Alexandre M. Alahi. Towards more practical threat models in artificial intelligence security. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4891–4908, 2024.
- 8 Sahar Abdelnabi, Kai Greshake, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not What You’ve Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *AISec@CCS*, pages 79–90. ACM, 2023.
- 9 Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How Does LLM Safety Training Fail? In *NeurIPS*, 2023.
- 10 Edoardo DeBenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. In *NeurIPS*, 2024.
- 11 Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David A. Forsyth, and Dan Hendrycks. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In *ICML*, 2024.

Participants

- Hyrum Anderson
CISCO Systems – San Jose, US
- Maksym Andriushchenko
ELLIS – Tübingen, DE
- Giovanni Apruzzese
Reykjavik University, IS
- Lujo Bauer
Carnegie Mellon University – Pittsburgh, US
- Rebekka Burkholz
CISPA – St. Ingbert, DE
- Tobias Callies
Universität der Bundeswehr München, DE
- Javier Carnerero Cano
IBM Research – Dublin, IE
- Lorenzo Cavallaro
University College London, GB
- Giovanni Cherubin
Microsoft Research – Cambridge, GB
- Ana-Maria Cretu
EPFL – Lausanne, CH
- Dinil Mon Divakaran
Institute for Infocomm Research, A*STAR – Singapore, SG
- Marc Fischer
Snyk – Zürich, CH
- Mario Fritz
CISPA – Saarbrücken, DE
- Kathrin Grosse
IBM Research Europe – Zürich, CH
- Stephan Günnemann
TU München – Garching, DE
- Andreas Happe
TU Wien, AT
- Amir Houmansadr
University of Massachusetts Amherst, US
- Daphne Ippolito
Carnegie Mellon University – Pittsburgh, US
- Marc Juarez
University of Edinburgh, GB
- Florian Kerschbaum
University of Waterloo, CA
- Wojciech Kusa
NASK – Warsaw, PL
- Pavel Laskov
Universität Liechtenstein – Vaduz, LI
- Nils Lukas
MBZUAI – Abu Dhabi, AE
- Emil C. Lupu
Imperial College London, GB
- Irina Nicolae
Apple – Paris, FR
- Andrew Paverd
Microsoft Research – Cambridge, GB
- Hammond Pearce
UNSW – Sydney, AU
- Jonathan Petit
Qualcomm – Boxborough, US
- Fabio Pierazzi
University College London, GB
- Liao Qianying
KU Leuven, BE
- Mathilde Aliénor Raynal
EPFL – Lausanne, CH
- Konrad Rieck
BIFOLD – Berlin, DE & TU Berlin, DE
- Vera Rimmer
KU Leuven, BE
- Advije Rizvani
Universität Liechtenstein – Vaduz, LI
- Ali Satvaty
University of Groningen, NL
- Lea Schönherr
CISPA – Saarbrücken, DE
- Ilia Shumailov
AI Security Company – London, GB
- Tim Van hamme
KU Leuven, BE
- Eric Wong
University of Pennsylvania – Philadelphia, US
- Christian Wressneger
Karlsruher Institut für Technologie, DE



Online Algorithms beyond Competitive Analysis

Sungjin Im^{*1}, Nicole Megow^{*2}, Debmalya Panigrahi^{*3}, Sahil Singla^{*4},
and Golnoosh Shahkarami^{†5}

1 University of California – Santa Cruz, US. sim9@ucsc.edu

2 Universität Bremen, DE. nicole.megow@uni-bremen.de

3 Duke University – Durham, US. debmalaya.panigrahi@gmail.com

4 Georgia Institute of Technology – Atlanta, US. ssingla@gatech.edu

5 MPI für Informatik – Saarbrücken, DE & Universität Bremen, DE.
golnooshshah@gmail.com

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25471: “Online Algorithms beyond Competitive Analysis”. This seminar brought 45 leading researchers in online algorithms and related areas together for a discussion of the recent progress and challenges in various aspects of online algorithms. The seminar included several talks of varying lengths and an open problem session. In addition, sufficient time was set aside for research discussions and collaborations.

Seminar November 16–21, 2025 – <https://www.dagstuhl.de/25471>

2012 ACM Subject Classification Theory of computation → Online algorithms

Keywords and phrases algorithms under uncertainty, beyond worst-case algorithm design, online algorithms

Digital Object Identifier 10.4230/DagRep.15.11.114

1 Executive Summary

Sungjin Im (University of California – Santa Cruz, US)

Nicole Megow (Universität Bremen, DE)

Debmalya Panigrahi (Duke University – Durham, US)

Sahil Singla (Georgia Institute of Technology – Atlanta, US)

License  Creative Commons BY 4.0 International license

© Sungjin Im, Nicole Megow, Debmalya Panigrahi, and Sahil Singla

Traditionally, algorithms are studied in a setting where the entire input is known in advance and the goal is to compute an optimal solution with respect to a given objective. In many applications, however, this paradigm is inadequate: inputs arrive over time and decisions must be made as information is revealed. This has given rise to the field of online algorithms, in which input revelation and algorithmic decisions are interleaved, creating a distinctive mix of computational and information constraints. A central theoretical tool for analyzing such algorithms is *competitive analysis*, which extends worst-case reasoning to the online setting by comparing the performance of an online algorithm to an offline optimum that has full knowledge of the complete input sequence. Over the past decades, this framework has generated a deep and influential body of results, shaped a set of core problems, and established online algorithms as a central area in algorithm design.

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Online Algorithms beyond Competitive Analysis, *Dagstuhl Reports*, Vol. 15, Issue 11, pp. 114–133

Editors: Sungjin Im, Nicole Megow, Debmalya Panigrahi, Sahil Singla, and Golnoosh Shahkarami



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

At the same time, it has long been recognized that competitive analysis, despite its elegance and conceptual clarity, can be too pessimistic in practice, and can fail to separate algorithms whose empirical performance differs substantially. A canonical illustration is online caching: many deterministic policies share the same worst-case competitive ratio, yet strategies such as least-recently-used can outperform others by large margins on realistic workloads. Motivated by such gaps between worst-case guarantees and observed behavior, the community has increasingly developed and studied beyond-worst-case models for online computation. These approaches aim to preserve rigorous guarantees while incorporating additional structure or flexibility that better reflects practical environments, thereby enabling finer-grained comparisons between algorithms that standard competitive analysis treats as essentially equivalent.

The Dagstuhl Seminar “Online Algorithms beyond Competitive Analysis” (25471) was organized in the context of this broader shift. It brought together researchers working on beyond-worst-case perspectives in online algorithms, to exchange viewpoints across sub-communities, and to identify promising directions and shared challenges. Discussions were structured around three representative themes:

- **Online algorithms with predictions** (learning-augmented algorithms): models in which an online algorithm is given auxiliary information about the future (e.g., forecasts derived from past data) and is designed to benefit from accurate predictions while remaining robust when predictions are noisy or unreliable.
- **Online and dynamic algorithms with recourse**: frameworks that relax strict irrevocability by allowing limited revisions to past decisions, capturing settings where changes are possible but costly, and studying how bounded recourse improves achievable guarantees and enables dynamic variants where requirements may both arrive and depart.
- **Online stochastic input models**: models that assume stochastic structure in arrivals (such as random order or distributional assumptions) to circumvent worst-case barriers, together with emerging efforts to make such models robust to deviations such as outliers or mild correlations.

Organization of the Seminar

The seminar brought together 45 researchers working on online algorithms. The participants embraced both senior and junior researchers, including a number of postdocs and advanced PhD students. During the five days of the seminar, 28 talks of different lengths took place. Six keynote speakers gave an overview of the state-of-the art of the respective area or presented recent highlight results in 60 minutes:

- Christian Coester: A Mirror Descent Approach for Online Algorithms
- Shuchi Chawla: Pandora’s Box
- Jose Correa: Prophet Inequalities
- Anupam Gupta: Online Algorithms with Samples
- Ziv Scully: Survey on Queueing Theoretic Scheduling
- David Wajc: Philosopher inequalities: online Bayesian selection beyond competitive analysis

The remaining slots were filled with shorter talks of 30 minutes and a few 8-minute spotlight talks.

Outcomes

Organizers and participants regard the seminar as a great success. The seminar achieved the goal to bring together researchers from different sub-communities addressing beyond-worst-case analyses, share the state-of-the art research and discuss the current major challenges. The talks were excellent and stimulating; participants connected and actively met in working groups in the afternoon and evenings.

The organizers wish to express their gratitude towards the Scientific Directorate and the administration of the Dagstuhl Center for their great support for this seminar.

2 Table of Contents

Executive Summary

Sungjin Im, Nicole Megow, Debmalaya Panigrahi, and Sahil Singla 114

Overview of Talks

Randomization and Stochasticity in Learning-Augmented Algorithms <i>Spyros Angelopoulos</i>	119
Competitive Bundle Trading <i>Yossi Azar</i>	119
The Power of Greedy for Online Minimum Cost Matching on the Line <i>Eric Balkanski</i>	120
Machine Learning for Online Resource Allocation <i>Benjamin Moseley</i>	120
Competitive Recourse (Beyond Absolute Recourse) <i>Niv Buchbinder</i>	121
Costly Information Combinatorial Selection <i>Shuchi Chawla</i>	121
Competitive Analysis via Mirror Descent <i>Christian Coester</i>	121
Prophet Inequalities: Old and New <i>José R. Correa</i>	122
On the Informativeness of Moments in Optimal Stopping <i>Andrés Cristi</i>	122
The Cost of Consistency: Submodular Maximization with Constant Recourse <i>Paul Dütting</i>	123
$O(1/\epsilon)$ is the answer in online weighted throughput maximization <i>Franziska Eberle</i>	123
Learning-augmented algorithms for MTS with bandit access to multiple predictors <i>Marek Eliáš</i>	124
Algorithmic Contract Design <i>Michal Feldman</i>	124
Online Algorithms with a Sample <i>Anupam Gupta</i>	125
Optimal Stopping with a Predicted Prior <i>Zhiyi Huang</i>	125
Prophet and Secretary at the Same Time <i>Thomas Kesselheim</i>	126
Online Bin Packing under Random-Order and IID model <i>Arindam Khan</i>	126
Optimal Online Algorithm for Robust Flow Time Scheduling <i>Amit Kumar</i>	127

A Model for Data Filtering <i>Ravi Kumar</i>	127
Random Order Online Learning <i>Stefano Leonardi</i>	128
Open problems in Knapsack Prophet Inequality and Contention Resolution <i>Will Ma</i>	128
Single-Sample Prophet Inequalities <i>Rebecca Reiffenhäuser</i>	129
Approximating Matroid Basis Testing for Partition Matroids using Budget-In-Expectation <i>Kevin Schewior</i>	129
Incremental Graph Algorithms with Predictions <i>Shikha Singh</i>	130
Improved Approximation Algorithms Using “Edge Predictions” <i>Sandeep Silwal</i>	130
Learning-Augmented Online Algorithms for Nonclairvoyant Joint Replenishment Problem with Deadlines <i>Seeun William Umboh</i>	131
Philosopher Inequalities: Online Bayesian Selection Beyond Competitive Analysis <i>David Wajc</i>	131
Open problems	
Revisiting the free-order prophet for one-dimensional parameter classes of distributions <i>Vianney Perchet</i>	131
Applications of Correlated Sampling from the Hypersimplex <i>David Wajc</i>	132
Participants	133

3 Overview of Talks

3.1 Randomization and Stochasticity in Learning-Augmented Algorithms

Spyros Angelopoulos (International Lab on Learning Systems – Montreal, CA)

License © Creative Commons BY 4.0 International license
© Spyros Angelopoulos

Randomization and stochasticity introduce new challenges while also enabling meaningful extensions in the design and analysis of learning-augmented algorithms. This presentation will discuss such challenges, with a focus on three main themes: (i) the characterization of Pareto-optimal randomized strategies, illustrated through the framework of search games; (ii) the use of distributional prediction oracles, where the nature of the prediction is probabilistic; and (iii) risk-based analysis, which incorporates notions of uncertainty and risk into the design and evaluation of learning-augmented algorithms.

3.2 Competitive Bundle Trading

Yossi Azar (Tel Aviv University, IL)

License © Creative Commons BY 4.0 International license
© Yossi Azar
Joint work of Yossi Azar, Niv Buchbinder, Roie Levin, Or Vardi
Main reference Yossi Azar, Niv Buchbinder, Roie Levin, Or Vardi: “Competitive Bundle Trading”, CoRR, Vol. abs/2507.23047, 2025.
URL <https://arxiv.org/abs/2507.23047>

Allocating or selling a set of resources to a sequence of arriving customers has been extensively studied under a wide range of models and objectives. However, a natural extension arises when the algorithm is also allowed to purchase additional resources from suppliers. This setting arguably captures the fundamental scenario faced by any trader or retailer. A natural objective in this context is to maximize the cumulative profit over the online sequence. That is, the difference between revenue from sales and expenses from purchases. This problem of maximizing the difference between two competing quantities appears to be significantly more challenging than the sell-only case, which may explain why it has remained largely unexplored.

We initiate the study of such a trading problem and design an algorithm that achieves a logarithmic competitive ratio relative to the optimal offline solution. Our approach employs an exponential-weight-update dynamic pricing scheme, and our analysis uses a dual-fitting argument that relates the algorithm’s profit to a linear programming relaxation that upper bounds the optimal offline profit. We also prove (nearly) matching lower bounds. Finally, we extend our results by designing an incentive-compatible mechanism for the setting in which customers are strategic and may misreport their true valuations.

3.3 The Power of Greedy for Online Minimum Cost Matching on the Line

Eric Balkanski (Columbia University – New York, US)

License © Creative Commons BY 4.0 International license

© Eric Balkanski

Joint work of Eric Balkanski, Yuri Faenza, Noémie Périer

Main reference Eric Balkanski, Yuri Faenza, Noémie Périer: “The Power of Greedy for Online Minimum Cost Matching on the Line”, in Proc. of the 24th ACM Conference on Economics and Computation, EC ’23, p. 185–205, Association for Computing Machinery, 2023.

URL <https://doi.org/10.1145/3580507.3597794>

We consider the online minimum cost matching problem on the line, in which there are n servers and, at each of n time steps, a request arrives and must be irrevocably matched to a server that has not yet been matched to, with the goal of minimizing the sum of the distances between the matched pairs. Despite achieving a worst-case competitive ratio that is exponential in n , the simple greedy algorithm, which matches each request to its nearest available free server, performs very well in practice. A major question is thus to explain greedy’s strong empirical performance.

In this talk, we aim to understand the performance of greedy over instances that are at least partially random. When both the requests and the servers are drawn uniformly and independently from $[0, 1]$, we show that greedy is constant competitive and extend this constant competitive ratio to a setting with a linear excess of servers. Moreover, we show that in the semi-random model where requests are still drawn uniformly and independently but servers are chosen adversarially, greedy achieves an $O(\log n)$ competitive ratio. When the requests arrive in a random order but are chosen adversarially, it was previously known that greedy is $O(n)$ -competitive. Even though this one-sided randomness allows a large improvement in greedy’s competitive ratio compared to the model where requests are adversarial and arrive in a random order, we show that it is not sufficient to obtain a constant competitive ratio by giving a tight $\Omega(\log n)$ lower bound. These results invite further investigation about how much randomness is necessary and sufficient to obtain strong theoretical guarantees for the greedy algorithm for online minimum cost matching, on the line and beyond.

3.4 Machine Learning for Online Resource Allocation

Benjamin Moseley (Carnegie Mellon University – Pittsburgh, US)

License © Creative Commons BY 4.0 International license

© Benjamin Moseley

Joint work of Benjamin Moseley, Andrew Li, Lin An

Main reference Lin An, Andrew A. Li, Benjamin Moseley, Gabriel Visotsky: “Best of Many in Both Worlds: Online Resource Allocation with Predictions under Unknown Arrival Model”, CoRR, Vol. abs/2402.13530, 2024.

URL <https://doi.org/10.48550/ARXIV.2402.13530>

This talk will discuss a model for augmenting discrete optimization algorithms with machine learned predictions to improve algorithm performance. First the talk will discuss online resource allocation using predictions. Predictions will be leveraged to go beyond the worst case lower bounds in both the regret and competitive analysis frameworks.

3.5 Competitive Recourse (Beyond Absolute Recourse)

Niv Buchbinder (Tel Aviv University, IL)

License © Creative Commons BY 4.0 International license
© Niv Buchbinder

Joint work of Niv Buchbinder, Sayan Bhattacharya, Roie Levin, Thatchaphol Saranurak, Seffi Naor, David Wajc

In this talk, I will explore the concepts of absolute and competitive recourse in the analysis of low-recourse algorithms. I will argue that competitive recourse provides a more meaningful framework for evaluating such algorithms. Finally, I will survey several recent results on competitive recourse and highlight a number of intriguing open problems.

3.6 Costly Information Combinatorial Selection

Shuchi Chawla (University of Texas – Austin, US)

License © Creative Commons BY 4.0 International license
© Shuchi Chawla

Joint work of Dimitris Christou, Trung Dang, Amit Harlev, Ziv Scully

Main reference Shuchi Chawla, Dimitrios Christou, Amit Harlev, Ziv Scully: “Combinatorial Selection with Costly Information”, in Proc. of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2026, Vancouver, BC, Canada, January 11-14, 2026, pp. 664–717, SIAM, 2026.
URL <https://doi.org/10.1137/1.9781611978971.26>

Main reference Shuchi Chawla, Dimitris Christou, Trung Dang: “Commitment Gap via Correlation Gap”, CoRR, Vol. abs/2508.20246, 2025.

URL <https://doi.org/10.48550/ARXIV.2508.20246>

Imagine you are hiring a graduate student from a large pool of candidates. How much effort should you spend reading recommendation letters? Should you interview some candidates, or none at all? Many decision-making problems involve this same dilemma: investing in information acquisition upfront can improve outcomes, but at a cost. Balancing this trade-off requires carefully weighing whether the value of additional signals justifies the resources spent to obtain them.

We capture this interplay through a model we call Costly Information Combinatorial Selection (CICS). In CICS, the algorithm is given a ground set of elements whose values are drawn independently from known distributions. The realized values can be revealed incrementally through costly probing actions, and the objective is to select a feasible subset maximizing total value minus information costs. In this talk, I will introduce the model and describe two approaches for designing approximation algorithms. Along the way, we will see how CICS connects to classical stochastic selection problems, including Pandora’s Box and prophet inequalities.

3.7 Competitive Analysis via Mirror Descent

Christian Coester (University of Oxford, GB)

License © Creative Commons BY 4.0 International license
© Christian Coester

Mirror descent is a general technique originating from convex optimization and used extensively in online learning, which in recent years also led to the successful discovery of optimal competitive ratios for various fundamental problems in online algorithms. This talk surveys how the technique can be used to obtain competitive online algorithms.

3.8 Prophet Inequalities: Old and New

José R. Correa (University of Chile – Santiago, CL)

License  Creative Commons BY 4.0 International license
© José R. Correa

In this talk, I will survey some of the most widely used techniques for designing and analyzing online algorithms under stochastic input. The playground is that of the classic prophet inequality. This states that, when faced with a finite sequence of non-negative independent random variables, a gambler who knows their distribution and is allowed to stop the sequence at any time, can obtain, in expectation, at least half as much reward as a prophet who knows the values of each random variable and can choose the largest one.

We begin by reviewing the classic “balanced prices” proof of this result, due to Samuel Cahn. We then introduce online contention resolution schemes of Feldman, Svensson, and Zenklusen, and discuss how to adapt them to the prophet inequality setting. To wrap up, we dive into the data-driven approach recently introduced by Rubinstein, Wang, and Weinberg. In all cases, we provide additional references to the literature and discuss some of the more recent research and open problems. In the final part of the talk, we introduce a new technique to tackle prophet inequalities based on Ordinary Differential Equations and show how this approach provides a unified view of the problem and all known algorithms for it. This last part is based on joint work with Fichtl, Jlibene, Laraki, Livanos, Schewior, and Verdugo.

3.9 On the Informativeness of Moments in Optimal Stopping

Andrés Cristi (EPFL – Lausanne, CH)

License  Creative Commons BY 4.0 International license
© Andrés Cristi

Joint work of Andrés Cristi, José Correa, Vasilis Livanos, Víctor Verdugo, Jiechen Zhang

We study a variant of the prophet inequality with limited information, where the decision maker only has access to the first k moments of each random variable, rather than their full distributions. In this work, it is proved that even with full moment knowledge (i.e., $k = \infty$), the best possible competitive ratio is $\Theta(1/\log n)$, which we show can already be achieved with only knowledge of the first moment. While the lower bound is simple and is attained by a standard bucketing algorithm, the upper bound requires a subtle construction. This involves using Vandermonde matrices first to construct a parametrized family of distributions for which the first k moments coincide, and for which the expected maximum of n such copies varies widely across different parameter choices. Then, we use Prokhorov’s Theorem to prove the existence of limit distributions, which we show have all their moments equal. Finally, we show how an adversary can select *equally looking* instances combining these distributions, making it impossible for the decision maker to obtain a factor better than $O(\log n)$ of the expected maximum. Our result, in particular, implies that to obtain improved prophet inequalities, further assumptions, beyond moment knowledge, are needed. To showcase this direction, we establish improved bounds under additional distributional assumptions, such as MHR, bounded coefficient of variation, or bounded support.

3.10 The Cost of Consistency: Submodular Maximization with Constant Recourse

Paul Dütting (Google – Zürich, CH)

License © Creative Commons BY 4.0 International license

© Paul Dütting

Joint work of Paul Dütting, Federico Fusco, Silvio Lattanzi, Ashkan Norouzi-Fard, Ola Svensson, Morteza Zadimoghaddam

Main reference Paul Dütting, Federico Fusco, Silvio Lattanzi, Ashkan Norouzi-Fard, Ola Svensson, Morteza Zadimoghaddam: “The Cost of Consistency: Submodular Maximization with Constant Recourse”, in Proc. of the 57th Annual ACM Symposium on Theory of Computing, STOC 2025, Prague, Czechia, June 23-27, 2025, pp. 1406–1417, ACM, 2025.

URL <https://doi.org/10.1145/3717823.3718131>

In this work, we study online submodular maximization, and how the requirement of maintaining a stable solution impacts the approximation. In particular, we seek bounds on the best-possible approximation ratio that is attainable when the algorithm is allowed to make at most a constant number of updates per step. We show a tight information-theoretic bound of $2/3$ for general monotone submodular functions, and an improved (also tight) bound of $3/4$ for coverage functions. Since both these bounds are attained by non poly-time algorithms, we also give a poly-time randomized algorithm that achieves a 0.51 -approximation. Combined with an information-theoretic hardness of $1/2$ for deterministic algorithms from prior work, our work thus shows a separation between deterministic and randomized algorithms, both information theoretically and for poly-time algorithms.

3.11 $O(1/\epsilon)$ is the answer in online weighted throughput maximization

Franziska Eberle (TU Berlin, DE)

License © Creative Commons BY 4.0 International license

© Franziska Eberle

Main reference Franziska Eberle: “ $O(1/\epsilon)$ Is the Answer in Online Weighted Throughput Maximization”, in Proc. of the 41st International Symposium on Theoretical Aspects of Computer Science (STACS 2024), Leibniz International Proceedings in Informatics (LIPIcs), Vol. 289, pp. 32:1–32:15, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024.

URL <https://doi.org/10.4230/LIPIcs.STACS.2024.32>

We study a fundamental online scheduling problem where jobs with processing times, weights, and deadlines arrive online over time at their release dates. The task is to preemptively schedule these jobs on a single or multiple (possibly unrelated) machines with the objective to maximize the weighted throughput, the total weight of jobs that complete before their deadline. To overcome known lower bounds for the competitive analysis, we assume that each job arrives with some slack $\epsilon > 0$; that is, the time window for processing job j on any machine i on which it can be executed has length at least $(1 + \epsilon)$ times j ’s processing time on machine i . Our contribution is a best possible online algorithm for weighted throughput maximization on unrelated machines: Our algorithm is $O(1/\epsilon)$ -competitive, which matches the lower bound for unweighted throughput maximization on a single machine. Even for a single machine, it was not known whether the problem with weighted jobs is “harder” than the problem with unweighted jobs. Thus, we answer this question and close weighted throughput maximization on a single machine with a best possible competitive ratio $\Theta(1/\epsilon)$.

3.12 Learning-augmented algorithms for MTS with bandit access to multiple predictors

Marek Eliáš (Bocconi University – Milan, IT)

License © Creative Commons BY 4.0 International license

© Marek Eliáš

Joint work of Marek Eliáš, Matei Coşa

Main reference Matei Gabriel Coşa, Marek Eliáš: “Learning-Augmented Algorithms for MTS with Bandit Access to Multiple Predictors”, CoRR, Vol. abs/2506.05479, 2025.

URL <https://doi.org/10.48550/ARXIV.2506.05479>

We consider the following problem: We are given ℓ heuristics for Metrical Task Systems (MTS), where each might be tailored to a different type of input instances. While processing an input instance received online, we are allowed to query the action of only one of the heuristics at each time step. Our goal is to achieve performance comparable to the best of the given heuristics. The main difficulty of our setting comes from the fact that the cost paid by a heuristic at time t cannot be estimated unless the same heuristic was also queried at time $t - 1$. This is related to Bandit Learning against memory bounded adversaries (Arora et al., 2012 [1]). We show how to achieve regret of $O(OPT^{2/3})$ and prove a tight lower bound based on the construction of Dekel et al. (2013) [2].

References

- 1 Arora, R., Dekel, O., and Tewari, A. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Conference on Machine Learning, ICML’12*, pp. 1747–1754, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.
- 2 Dekel, O., Ding, J., Koren, T., and Peres, Y. Bandits with switching costs: $T^{2/3}$ regret. *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, 2013. URL <https://api.semanticscholar.org/CorpusID:9425790>.

3.13 Algorithmic Contract Design

Michal Feldman (Tel Aviv University, IL)

License © Creative Commons BY 4.0 International license

© Michal Feldman

Joint work of Michal Feldman, Paul Dütting, Tomer Ezra, Yoav Gal-Tzur, Thomas Kesselheim, Maya Schlesinger

Main reference Michal Feldman: “Algorithmic Contract Design (Keynote)”, in Proc. of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, p. 2, Association for Computing Machinery, 2024.

URL <https://doi.org/10.1145/3618260.3664273>

Algorithmic contract design is a new frontier at the interface of economics and computation, studying scenarios where a principal delegates the execution of a costly project to an agent or a team of agents, and incentivizes them through a contract that specifies payments contingent on the project’s success. This domain has gained increasing interest from the theoretical computer science community, particularly in the realm of combinatorial contracts. In this talk, I will survey two distinct models of combinatorial contracts, each illustrating unique sources of complexity encountered in contract design. The first model allows an agent to select from a set of possible actions, examining the intricate dependencies among these actions. The second model involves motivating a team of agents, focusing on the interdependencies within various agent combinations. I will present both (approximation) algorithms and hardness results concerning the optimal contract problem in these settings.

3.14 Online Algorithms with a Sample

Anupam Gupta (New York University, US)

License © Creative Commons BY 4.0 International license
© Anupam Gupta

While analyzing algorithms in the worst-case has long served us well, recent years have seen an exciting surge in analyzing algorithms in models that go beyond the worst case. We consider the classical problem of load-balancing, where jobs arrive online and must be assigned to collections of machines to minimize the maximum load. Can we get results better than the adversarial setting if we have a small sample of the upcoming data? A recent line of work has explored this research direction, and we give a survey of some results and techniques in the area.

3.15 Optimal Stopping with a Predicted Prior

Zhiyi Huang (University of Hong Kong, HK)

License © Creative Commons BY 4.0 International license
© Zhiyi Huang
Joint work of Tian Bai, Zhiyi Huang, Chui Shan Lee, Dongchen Li
Main reference Tian Bai, Zhiyi Huang, Chui Shan Lee, Dongchen Li: “Optimal Stopping with a Predicted Prior”, CoRR, Vol. abs/2511.03289, 2025.
URL <https://doi.org/10.48550/ARXIV.2511.03289>

There are two major models of value uncertainty in the optimal stopping literature: the secretary model, which assumes no prior knowledge, and the prophet inequality model, which assumes full information about value distributions. In practice, decision makers often rely on machine-learned priors that may be erroneous. Motivated by this gap, we formulate the model of optimal stopping with a predicted prior to design algorithms that are both consistent, exploiting the prediction when accurate, and robust, retaining worst-case guarantees when it is not.

Existing secretary and prophet inequality algorithms are either pessimistic in consistency or not robust to misprediction. A randomized combination only interpolates their guarantees linearly. We show that a family of bi-criteria algorithms achieves improved consistency-robustness trade-offs, both for maximizing the expected accepted value and for maximizing the probability of accepting the maximum value. We further prove that for the latter objective, no algorithm can simultaneously match the best prophet inequality algorithm in consistency, and the best secretary algorithm in robustness.

3.16 Prophet and Secretary at the Same Time

Thomas Kesselheim (*Universität Bonn, DE*)

License © Creative Commons BY 4.0 International license
© Thomas Kesselheim

Joint work of Thomas Kesselheim, Gregory Kehne

Main reference Gregory Kehne, Thomas Kesselheim: “Prophet and Secretary at the Same Time”, CoRR, Vol. abs/2511.09531, 2025.

URL <https://doi.org/10.48550/ARXIV.2511.09531>

Many online problems are studied in stochastic settings for which inputs are samples from a known distribution, given in advance, or from an unknown distribution. Such distributions model both beyond-worst-case inputs and, when given, partial foreknowledge for the online algorithm. But how robust can such algorithms be to misspecification of the given distribution? When is this detectable, and when does it matter? When can algorithms give good competitive ratios both when the input distribution is as specified, and when it is not?

We consider these questions in the setting of optimal stopping, where the cases of known and unknown distributions correspond to the well-known prophet inequality and to the secretary problem, respectively. Here we ask: Can a stopping rule be competitive for the i.i.d. prophet inequality problem and the secretary problem at the same time? We constrain the Pareto frontier of simultaneous approximation ratios (α, β) that a stopping rule can attain. We introduce a family of algorithms that give nontrivial joint guarantees and are optimal for the extremal i.i.d. prophet and secretary problems. We also prove impossibilities, identifying (α, β) unattainable by any adaptive stopping rule. Our results hold for both n fixed arrivals and for arrivals from a Poisson process with rate n . We work primarily in the Poisson setting, and provide reductions between the Poisson and n -arrival settings that may be of broader interest.

3.17 Online Bin Packing under Random-Order and IID model

Arindam Khan (*Indian Institute of Science – Bangalore, IN*)

License © Creative Commons BY 4.0 International license
© Arindam Khan

We consider the online bin packing problem under two stochastic settings. First, we study bin packing under the i.i.d. model, where item sizes are sampled independently and identically from a distribution in $[0, 1]$. We provide a linear time $(1 + \epsilon)$ -competitive algorithm for online bin packing under i.i.d. Model. We then study the random-order model, where an adversary chooses the instance, but the order of arrival of items in the instance is drawn uniformly at random from the set of all permutations of the items. Kenyon [SODA’96] conjectured the random-order ratio is 1.15. We make some progress towards the conjecture.

3.18 Optimal Online Algorithm for Robust Flow Time Scheduling

Amit Kumar (Indian Institute of Technology – New Delhi, IN)

License © Creative Commons BY 4.0 International license

© Amit Kumar

Joint work of Amit Kumar, Anupam Gupta, Debmalya Panigrahi, Zhaozi Wang

Main reference Anupam Gupta, Amit Kumar, Debmalya Panigrahi, Zhaozi Wang: “An Optimal Online Algorithm for Robust Flow Time Scheduling”, in Proc. of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2026, Vancouver, BC, Canada, January 11-14, 2026, pp. 1214–1238, SIAM, 2026.

URL <https://doi.org/10.1137/1.9781611978971.47>

The problem of minimizing the total flow time on a single machine is one of the few problems for which we can give an optimal online algorithm: just schedule the job with the shortest remaining processing time (SRPT). However, this requires knowledge of the true running time p_j of each job j . Azar, Leonardi, and Tseitou recently asked: what if we are given estimates p'_j for each job, such that the multiplicative error between p_j and p'_j (called the distortion) is at most μ ? It is easy to construct examples where no algorithm can be $o(\mu)$ competitive; can we get $O(\mu)$ competitiveness?

We show how to achieve this asymptotically optimal result, improving on the previous best result of $O(\mu \log \mu)$. Moreover, we give a very simple algorithm to get this tight bound of $O(\mu)$; the previous ZigZag algorithm was relatively more involved. Our proof is via a dual-fitting argument based on the idea of a reduced instance: we consider an LP relaxation based on knapsack-cover inequalities, and show a solution with a large dual value. Our ideas can also be extended to give alternative dual-fitting arguments for previously analyzed algorithms (like SRPT, ZigZag, and others for minimizing the total flow time, and the Bansal-Dhamdhere algorithm for weighted flow-time minimization).

3.19 A Model for Data Filtering

Ravi Kumar (Google – Mountain View, US)

License © Creative Commons BY 4.0 International license

© Ravi Kumar

Joint work of Ravi Kumar, Aditya Bhaskara, Ashok Cutkosky, Manish Purohit

Selecting high quality data is a crucial prerequisite for training and fine-tuning large language models. We propose a rigorous theoretical framework for data filtering that lets us reason about the alignment of the filtered data to the task of interest while being fully oblivious to the specifics of the downstream learning model. Formally, given a target distribution d and access to a distribution D that is a mixture of d and some other unknown distribution D' , we model the data filtering problem as designing a filter such that the distribution D conditioned on passing the filter is close in total variation distance to the target distribution d . Our main result is a simple filter that achieves a total variation distance of roughly $\tilde{O}\left(\frac{1}{N_d} + \frac{1}{pN_D}\right)$, where N_D is the number of samples from D , N_d is the number of samples from d , and p is the mixing probability. We also show nearly matching information-theoretic lower bounds.

3.20 Random Order Online Learning

Stefano Leonardi (Sapienza University of Rome, IT)

License © Creative Commons BY 4.0 International license
© Stefano Leonardi

Joint work of Stefano Leonardi, Martino Bernasconi, Andrea Celli, Riccardo Colini-Baldeschi, Federico Fusco, Matteo Russo

Main reference Martino Bernasconi, Andrea Celli, Riccardo Colini-Baldeschi, Federico Fusco, Stefano Leonardi, Matteo Russo: “Online Learning in the Random Order Model”, CoRR, Vol. abs/2510.02820, 2025.

URL <https://doi.org/10.48550/ARXIV.2510.02820>

In the random-order model for online learning, the sequence of losses is chosen upfront by an adversary and presented to the learner after a random permutation. While algorithms for adversarial inputs naturally maintain their regret guarantees in random order, simple no-regret algorithms exist for the stochastic model that fail against random-order instances. In this paper, we propose a general template to adapt stochastic learning algorithms to the random-order model without substantially affecting their regret guarantees.

3.21 Open problems in Knapsack Prophet Inequality and Contention Resolution

Will Ma (Columbia University – New York, US)

License © Creative Commons BY 4.0 International license
© Will Ma

Main reference Will Ma, Calum MacRury, Cliff Stein: “Forward-backward Contention Resolution Schemes for Fair Rationing”, in Proc. of the 26th ACM Conference on Economics and Computation, EC 2025, Stanford University, Stanford, CA, USA, July 7-10, 2025, p. 945, ACM, 2025.

URL <https://doi.org/10.1145/3736252.3742654>

Main reference Jiashuo Jiang, Will Ma, Jiawei Zhang: “Tight Guarantees for Multi-unit Prophet Inequalities and Online Stochastic Knapsack”, in Proc. of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 – 12, 2022, pp. 1221–1246, SIAM, 2022.

URL <https://doi.org/10.1137/1.9781611977073.51>

A sequence of items $i = 1, \dots, n$ arrive with random sizes S_i and weights W_i drawn from known distributions, and must be irrevocably accepted or rejected, subject to the constraint that the total size accepted is at most 1. What is the maximum total weight that an online algorithm can accept in expectation, compared to a prophet who sees the realizations in advance? Surprisingly, the best-known algorithmic guarantee for this problem relies on the (loose) knapsack polytope followed by online contention resolution, and better guarantees are difficult to achieve even under random-order or IID arrivals. This talk surveys some of these challenges and puzzling open problems.

3.22 Single-Sample Prophet Inequalities

Rebecca Reiffenhäuser (*University of Amsterdam, NL*)

- License** © Creative Commons BY 4.0 International license
© Rebecca Reiffenhäuser
- Joint work of** Constantine Caramanis, Paul Dütting, Matthew Faw, Federico Fusco, Philip Lazos, Stefano Leonardi, Orestis Papadigenopoulos, Emmanouil Pountourakis, Rebecca Reiffenhäuser, Thomas Kesselheim, Brendan Lucier, Sahil Singla
- Main reference** Constantine Caramanis, Paul Dütting, Matthew Faw, Federico Fusco, Philip Lazos, Stefano Leonardi, Orestis Papadigenopoulos, Emmanouil Pountourakis, Rebecca Reiffenhäuser: “Single-Sample Prophet Inequalities via Greedy-Ordered Selection”, in Proc. of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 – 12, 2022, pp. 1298–1325, SIAM, 2022.
- URL** <https://doi.org/10.1137/1.9781611977073.54>
- Main reference** Paul Dütting, Thomas Kesselheim, Brendan Lucier, Rebecca Reiffenhäuser, Sahil Singla: “Online Combinatorial Allocations and Auctions with Few Samples”, in Proc. of the 65th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2024, Chicago, IL, USA, October 27-30, 2024, pp. 1231–1250, IEEE, 2024.
- URL** <https://doi.org/10.1109/FOCS61266.2024.00081>

The classic prophet inequality uses knowledge of the distributions that produce all online values to circumvent the impossibility of obtaining constant competitive ratios. However, in many problem versions, complete distributional knowledge is unnecessary to achieve this: instead, it often suffices to obtain a single sample (value) from each distribution before start of the algorithm. The talk introduces a general technique to obtain such Single-Sample Prophet inequalities, and demonstrates its generality by applying to different combinatorial versions of the problem.

3.23 Approximating Matroid Basis Testing for Partition Matroids using Budget-In-Expectation

Kevin Schewior (*Universität Köln, DE*)

- License** © Creative Commons BY 4.0 International license
© Kevin Schewior
- Joint work of** Lisa Hellerstein, Benedikt M. Plank, Kevin Schewior
- Main reference** Lisa Hellerstein, Benedikt M. Plank, Kevin Schewior: “Approximating Matroid Basis Testing for Partition Matroids using Budget-In-Expectation”, in Proc. of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2026, Vancouver, BC, Canada, January 11-14, 2026, pp. 2588–2616, SIAM, 2026.
- URL** <https://doi.org/10.1137/1.9781611978971.94>

We consider the following Stochastic Boolean Function Evaluation problem, which is closely related to several problems from the literature. A matroid $\mathcal{M}$ (in compact representation) on ground set E is given, and each element $i \in E$ is active independently with known probability $p_i \in (0, 1)$. The elements can be queried, upon which it is revealed whether the respective element is active or not. The goal is to find an adaptive querying strategy for determining whether there is a basis of $\mathcal{M}$ in which all elements are active, with the objective of minimizing the expected number of queries.

When $\mathcal{M}$ is a uniform matroid, this is the problem of evaluating a k -of- n function, first studied in the 1970s. This problem is well-understood, and has an optimal adaptive strategy that can be computed in polynomial time.

Taking $\mathcal{M}$ to instead be a partition matroid, we show that previous approaches fail to give a constant-factor approximation. Our main result is a polynomial-time constant-factor approximation algorithm producing a randomized strategy for this partition matroid problem. We obtain this result by combining a new technique with several well-established

techniques. Our algorithm adaptively interleaves solutions to several instances of a novel type of stochastic querying problem, with a constraint on the *expected* cost. We believe that this type of problem is of independent interest, will spark follow-up work, and has the potential for additional applications.

3.24 Incremental Graph Algorithms with Predictions

Shikha Singh (Williams College – Williamstown, US)

License © Creative Commons BY 4.0 International license
© Shikha Singh

Joint work of Shikha Singh, Samuel McCauley, Benjamin Moseley, Aidin Niaparast, Helia Niaparast

Main reference Samuel McCauley, Benjamin Moseley, Aidin Niaparast, Helia Niaparast, Shikha Singh: “Incremental Approximate Single-Source Shortest Paths with Predictions”, in Proc. of the ICALP, LIPIcs, Vol. 334, pp. 117:1–117:20, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2025.

URL <https://doi.org/10.4230/LIPICS.ICALP.2025.117>

In an incremental graph problem, the vertices of the graph G are known ahead of time and the edges are revealed over time. The goal is to efficiently maintain some property (such as connectivity or shortest path) of the current graph G_t . Predictions about future inserts in such problems help interpolate smoothly between the fully online and fully offline settings. This talk describes recent results that use a recursive decomposition to leverage predictions and design faster incremental graph algorithms for shortest path and strong connectivity.

3.25 Improved Approximation Algorithms Using “Edge Predictions”

Sandeep Silwal (University of Wisconsin-Madison, US)

License © Creative Commons BY 4.0 International license
© Sandeep Silwal

Joint work of Sandeep Silwal, Anders Aamand, Justin Y. Chen, Siddharth Gollapudi, Hao Wu, Yuhao Liu

Augmenting constraint satisfaction problems (CSPs) with predictions has been a recent popular theme in learning-augmented algorithms. In this talk I will describe improved approximation algorithms for CSPs using “a prediction per constraint.” In the context of NP-hard graph problems, each edge provides two bits of information, corresponding to predictions about whether each of its endpoints belongs to an optimal solution. In other contexts, such as Max3SAT, every clause gives information about the variables participating in the clause. Even with weak predictions where each bit is only ϵ -correlated with the true solution, this information allows us to break approximation barriers in the standard setting. We develop algorithms with improved approximation ratios for MaxCut, Vertex Cover, Set Cover, and Maximum Independent Set problems (among others). Across these problems, our algorithms share a unifying theme, where we separately satisfy constraints related to high-degree vertices (using predictions) and low-degree vertices (without using predictions) and carefully combine the answers. We will also discuss the many open problems related to the “strengths” of different prediction models for CSPs and potential future directions.

3.26 Learning-Augmented Online Algorithms for Nonclairvoyant Joint Replenishment Problem with Deadlines

Seeun William Umboh (*The University of Melbourne, AU*)

License © Creative Commons BY 4.0 International license

© Seeun William Umboh

Joint work of Seeun William Umboh, Jeremy Fineman, Michael Dinitz

Main reference Michael Dinitz, Jeremy T. Fineman, Seeun William Umboh: “Learning-Augmented Online Algorithms for Nonclairvoyant Joint Replenishment Problem with Deadlines”, CoRR, Vol. abs/2511.16094, 2025.

URL <https://doi.org/10.48550/ARXIV.2511.16094>

This talk considers using predictions in the context of the online Joint Replenishment Problem with Deadlines. Prior work includes asymptotically optimal competitive ratios of $O(1)$ for the clairvoyant setting and $O(\sqrt{n})$ for the nonclairvoyant setting, where n is the number of items. Our goal is to significantly reduce the competitive ratio in the nonclairvoyant case by leveraging predictions.

3.27 Philosopher Inequalities: Online Bayesian Selection Beyond Competitive Analysis

David Wajc (*Technion – Haifa, IL*)

License © Creative Commons BY 4.0 International license

© David Wajc

Joint work of David Wajc, Mark Braverman, Mahsa Derakhshan, Seffi Naor, Christos Papadimitriou, Tristan Pollner, Amin Saberi, Aravind Srinivasan

In online Bayesian selection problems, a seller is faced with a stream of buyers arriving sequentially. Each arriving buyer makes, for each item on sale, a take-it-or-leave-it offer drawn from some known distribution, which the seller must immediately either accept or decline. It is common to analyze the seller’s gain via the lens of competitive analysis, comparing the seller’s gain to that of the optimal offline solution, obtained by a “prophet” who knows the future. However, the seller might not even be able to compete with the optimal online algorithm, which may be computationally intractable to run. In this case the optimal online algorithm is obtained by a “philosopher” with sufficient time to think (compute). In this talk I will discuss some developments on online algorithms efficiently approximating this philosopher.

4 Open problems

4.1 Revisiting the free-order prophet for one-dimensional parameter classes of distributions

Vianney Perchet (*ENSAE – Palaiseau, FR*)

License © Creative Commons BY 4.0 International license

© Vianney Perchet

In their 1985 paper, Hill and Hordijk [1] defined, in the now classical prophet inequality problem, as “descending optimal” any class of distribution, parameterized by a one-dimensional parameter, such that the value of the online optimal algorithm (i.e., dynamic programming) is the highest when the distributions are observed in decreasing order with respect to their parameter.

We wonder if we can give a simple characterization of classes that are descending optimal, maybe when those random variables have the same distribution up to some (known) multiplicative factor, the one-dimensional parameter. It is known that Bernoulli, uniform and exponential distributions are indeed descending optimal.

An alternative consists in the same question when the dynamic programming is replaced with the best “single-threshold” algorithm.

References

- 1 Hill, T. P., Hordijk, A. (1985). *Selection of order of observation in optimal stopping problems*. Journal of applied probability, 22(1), 177-184.

4.2 Applications of Correlated Sampling from the Hypersimplex

David Wajc (Technion – Haifa, IL)

License  Creative Commons BY 4.0 International license
© David Wajc

Joint work of Joseph (Seffi) Naor, Nitya Raju, Abhishek Shetty, Aravind Srinivasan, Renata Valieva, David Wajc
Main reference Joseph (Seffi) Naor, Nitya Raju, Abhishek Shetty, Aravind Srinivasan, Renata Valieva, David Wajc: “Dimension-Free Correlated Sampling for the Hypersimplex”, in Proc. of the 17th Innovations in Theoretical Computer Science Conference, ITCS 2026, Bocconi University, Milan, Italy, January 27-30, 2026, LIPIcs, pp. 104:1–104:20, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026.
URL <https://doi.org/10.4230/LIPICS.ITCS.2026.104>

► **Definition 1.** Let $\Delta n, k \triangleq x \in [0, 1]^n : \|x\|_1 \leq k$ be the convex hull of origin and the hypersimplex (the same polytope with the inequality replaced with an equality).

► **Definition 2.** A distribution D over deterministic algorithms $A : \Delta n, k \rightarrow 2^{[n]}$ mapping points in $\Delta n, k$ to subsets of $[n]$ of cardinality at most k is an α -stretch correlated sampling algorithm for $\Delta n, k$ if it satisfies the following properties for all $x, y \in \Delta n, k$:

(P1) (**Cardinality**) $|A(x)| \in [\|x\|_1, \lceil \|x\|_1 \rceil]$ (hence $|A(x)| \leq k$) for all A in the support of D .

(P2) (**Marginals**) $\Pr_{A \sim D}[i \in A(x)] = x_i$ for all $i \in [n]$.

(P3) (**Stretch**) $\mathbb{E}_{A \sim D}[|A(x) \oplus A(y)|] \leq \alpha \cdot \|x - y\|_1$, where $\oplus$ denotes the symmetric difference.

For $k = 1$, it is known that $\alpha = 2$ is best possible. Algorithms for this problem, referred to as correlated sampling, have found applications to approximation algorithms, parallel repetition, replicability and privacy, parallel sampling and parallelizing next-token generation, cryptography, and locality sensitive hashing.

For $k > 1$, a work of Chen et al. ICALP17 shows that $\alpha = O(\log n)$ is attainable, improved to a dimension-free bound of $O(\log k)$ by Naor et al. ITCS26. The latter algorithm is also implementable efficiently in parallel and dynamic settings, and furthermore preserves submodular objectives. The recent work shows some simple applications, but it seems plausible that this machinery has further applications. What applications does this generalization of correlated sampling enable?

Participants

- Anders Aamand
University of Copenhagen, DK
- Spyros Angelopoulos
International Lab on Learning
Systems – Montreal, CA
- Antonios Antoniadis
University of Twente –
Enschede, NL
- Yossi Azar
Tel Aviv University, IL
- Eric Balkanski
Columbia University –
New York, US
- Niv Buchbinder
Tel Aviv University, IL
- Shuchi Chawla
University of Texas – Austin, US
- Christian Coester
University of Oxford, GB
- José R. Correa
University of Chile –
Santiago, CL
- Andrés Cristi
EPFL – Lausanne, CH
- Sami Davies
University of California –
Berkeley, US
- Paul Dütting
Google – Zürich, CH
- Franziska Eberle
TU Berlin, DE
- Marek Eliáš
Bocconi University – Milan, IT
- Michal Feldman
Tel Aviv University, IL
- Rohan Ghuge
University of Texas – Austin, US
- Anupam Gupta
New York University, US
- Zhiyi Huang
University of Hong Kong, HK
- Sungjin Im
University of California –
Santa Cruz, US
- Thomas Kesselheim
Universität Bonn, DE
- Arindam Khan
Indian Institute of Science –
Bangalore, IN
- Amit Kumar
Indian Institute of Technology –
New Delhi, IN
- Ravi Kumar
Google – Mountain View, US
- Stefano Leonardi
Sapienza University of Rome, IT
- Roie Levin
Rutgers University –
New Brunswick, US
- Vasilis Livanos
University of Chile –
Santiago, CL
- Will Ma
Columbia University –
New York, US
- Nicole Megow
Universität Bremen, DE
- Marco S. Molinaro
Microsoft Research – Redmond,
US & PUC – Rio de Janeiro, BR
- Benjamin J. Moseley
Carnegie Mellon University –
Pittsburgh, US
- Seffi Naor
Technion – Haifa, IL
- Debmalya Panigrahi
Duke University – Durham, US
- Vianney Perchet
ENSAE – Palaiseau, FR
- Rebecca Reiffenhäuser
University of Amsterdam, NL
- Adi Rosén
CNRS – Paris, FR &
Université Paris Cité, FR
- Barna Saha
University of California –
San Diego, US
- Kevin Schewior
Universität Köln, DE
- Ziv Scully
Cornell University – Ithaca, US
- Golnoosh Shahkarami
MPI für Informatik –
Saarbrücken, DE
- Sandeep Silwal
University of Wisconsin-
Madison, US
- Shikha Singh
Williams College –
Williamstown, US
- Sahil Singla
Georgia Institute of Technology –
Atlanta, US
- Ola Svensson
EPFL – Lausanne, CH
- Seeun William Umboh
The University of Melbourne, AU
- David Wajc
Technion – Haifa, IL



Approaches and Applications of Inductive Programming

Ute Schmid^{*1}, Gust Verbruggen^{*2}, and Sonja Niemann^{†3}

1 Universität Bamberg, DE. ute.schmid@uni-bamberg.de

2 Microsoft – Keerbergen, BE. gustverbruggen@gmail.com

3 bidt – München, DE. sonja.niemann@bidt.digital

Abstract

The Dagstuhl Seminar “Approaches and Applications of Inductive Programming” (AAIP) took place in 2025 for the seventh time. The focus of this seminar series is methods and applications of learning computer programs from incomplete and informal specifications such as input/output examples or natural language prompts. Researchers come from different areas, mostly from machine learning and other branches of artificial intelligence research, cognitive scientists interested in human learning in complex domains, and researchers with a background in formal methods and programming languages. The focus of the AAIP 2025 seminar was the application of large language models to code generation and the evaluation of quality of generated code in comparison with other, classic inductive programming approaches. Furthermore, neuro-symbolic approaches to inductive programming, were explored. Applications of inductive programming in different domains such as biomedical scientific discovery, control code generation in complex industrial settings, and programming education were discussed.

Seminar November 30 – December 5, 2025 – <https://www.dagstuhl.de/25491>

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Human-centered computing → Human computer interaction (HCI); Theory of computation → Logic; Computing methodologies → Machine learning; Theory of computation → Semantics and reasoning; Software and its engineering → Software creation and management

Keywords and phrases Diffusion Models, Inductive Programming, LLMs, Program Synthesis, Programming by Examples

Digital Object Identifier 10.4230/DagRep.15.11.134

1 Executive Summary

Ute Schmid (Universität Bamberg, DE)

Gust Verbruggen (Microsoft – Keerbergen, BE)

License  Creative Commons BY 4.0 International license
© Ute Schmid and Gust Verbruggen

Inductive programming (IP) research, also called inductive program synthesis, is concerned with approaches to learn computer programs from incomplete and informal specifications, such input/output examples or natural language prompts. IP research incorporates different areas of computer science, especially machine learning, automated reasoning, program verification, programming language theory, and software engineering. Furthermore, IP has strong relations to cognitive science, where IP can help build models of human inductive learning, and to provide intelligent tutoring systems for programming education. IP is also relevant to industry, where it supports tools for end-user programming.

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Approaches and Applications of Inductive Programming, *Dagstuhl Reports*, Vol. 15, Issue 11, pp. 134–156

Editors: Ute Schmid, Gust Verbruggen, and Sonja Niemann



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The Dagstuhl Seminar “Approaches and Applications of Inductive Programming” (AAIP) took place in 2025 for the seventh time. Given the recent developments in generative methods and their applications to code generation, the focus topic of the seminar was the application of large language models to code generation and the evaluation of quality of generated code in comparison with other classic inductive programming approaches. Recent research on large language models as well as diffusion methods for code generation from examples were presented. Suitable benchmarks and evaluation strategies taking into account reliability and quality of generated code were discussed. Furthermore, advances in inductive logic programming, functional program synthesis, and probabilistic programming were presented.

Cognitive science research on IP approaches as models for human concept learning, abstract visual reasoning and problem solving were presented together with empirical results. Furthermore, applications of IP methods for intelligent tutoring in the domain of programming education were discussed. IP methods have been matured over the last decade such that applications in complex real world domains can be tackled. Applications of classic as well as generative IP methods in scientific discovery in biomedicine as well as control code generation in complex industrial settings, were presented.

Core aspects of ongoing IP research were discussed in four working groups: Neuro-symbolic program induction, gradual refinement methods for concept induction and human-like iterative programming, programming education in the age of large language models, inductive programming for complex biological problems, and IP benchmarks and evaluation strategies suitable for comparing performance and quality of programs generated with classic as well as generative IP approaches.

Based on the presentations and in-depth discussions of ongoing research in IP, there was a general agreement that despite the impressive progress of generative methods for code generation, inductive programming is not solved. The main challenges are in combinations of generative approaches with other methods of inductive programming and machine learning to support generation of programs that are reliable and comprehensible. Furthermore, there was a general agreement between participants that the interdisciplinary exchange between core IP researchers with cognitive scientists, and programming language researchers was highly fruitful. Finally, the participation of researchers from industry was perceived as very inspiring, showing advances and challenges for real world applications of inductive programming.

2 Table of Contents

Executive Summary

<i>Ute Schmid and Gust Verbruggen</i>	134
---	-----

Overview of Talks

Concept Learning as Coarse-To-Fine Probabilistic Program Synthesis <i>Maddy Bowers</i>	138
Hypothesis Space Processing for Efficient Rule Learning Through Inductive Logic Programming <i>David M. Cerna</i>	138
Learning cancer programs <i>Jasmin Fisher</i>	140
Computer-Aided Programming and Problem Solving <i>Frank Jäkel</i>	140
Inductive Functional Programming: Igor2 (vs ChatGPT) and Generalized I/O-Examples <i>Emanuel Kitzelmann</i>	140
LLM-based feature generation from text for interpretable machine learning <i>Tomáš Kliegr</i>	141
Generating Industrial Control Logic from Natural Language Requirements <i>Heiko Kozirolek</i>	142
On Solving Visual Analogies <i>Johannes Langer</i>	142
Evaluating Large Language Models in Programming-by-Example <i>Félix Martí Pérez</i>	143
Beyond Performance: Towards Reliable Large Language Models <i>Yael Moros Daval</i>	143
ConDABench: Interactive Evaluation of Data Analysis Assistants <i>Arjun Radhakrishna</i>	143
A Short History of Inductive Programming <i>Ute Schmid</i>	144
Prompt Programming <i>Adish Singla</i>	145
Execution-guided within-prompt search for programming-by-example <i>Gust Verbruggen and Mukul Singh</i>	145
Cyclic Bayesian networks: From probabilistic modelling to probabilistic programming <i>Felix Weitzkämper</i>	145

Working groups

Inductive programming for complex biological problems <i>David Cerna, Jasmin Fisher, Pierre Flener, Tomáš Kliegr, Heiko Kozirolek, Raoul van Doren, and Felix Weitzkämper</i>	146
--	-----

Neuro-Symbolic Program Induction <i>Quinten Dewulf, Wang-Zhou Dai, Cesar Ferri Ramirez, Johannes Fürnkranz, Johannes Langer, and Zachary Siegel</i>	147
Gradual refinement in coarse-to-fine concept induction and human-like iterative programming <i>Alex Lew, Maddy Bowers, Gabriel Grand, Emanuel Kitzelmann, David Nagy, Tejaswini Ojha, Mukul Singh, Tobias Thomas, Dominik Ürüm, and Hanqi Zhou . . .</i>	150
Benchmarks and Evaluation <i>Yael Moros Daval, Félix Martí Pérez, Yewen Pu, Arjun Radhakrishna, Ute Schmid, and Gust Verbruggen</i>	152
Programming Education in the Age of Large Language Models <i>Hila Peleg, Frank Jäkel, Sven Mayer, and Sonja Niemann</i>	153
Participants	156

3 Overview of Talks

3.1 Concept Learning as Coarse-To-Fine Probabilistic Program Synthesis

Maddy Bowers (MIT – Cambridge, US)

License © Creative Commons BY 4.0 International license
© Maddy Bowers

Joint work of Maddy Bowers, Mauricio Barba da Costa, Xiaoyan Wang, Joshua Tenenbaum, Vikash Mansinghka, Armando Solar-Lezama, Alexander Lew

In this talk I'll discuss a family of new concept induction techniques inspired by two key aspects of human concept learning – our ability to judge how promising a vague, partial hypothesis is, and our ability to gradually refine these vague explanations of observations to precise ones. First, I will argue that coarse hypotheses should be represented as probabilistic programs that combine deterministic logic that explains part of the program structure with probabilistic sampling to encode uncertainty about what else might be going on. Second, I'll show how we can measure the quality of one of these coarse hypotheses precisely by calculating likelihoods of input-output examples, and present Pluck, a new tool we have developed computes these likelihoods really efficiently. Third, I'll show how these likelihood scores can be used to guide synthesis, and finally I'll conclude with some preliminary experimental results.

3.2 Hypothesis Space Processing for Efficient Rule Learning Through Inductive Logic Programming

David M. Cerna (Dynatrace Research – Linz, AT)

License © Creative Commons BY 4.0 International license
© David M. Cerna

Joint work of David M. Cerna, Andrew Cropper

Main reference Andrew Cropper, David M. Cerna, Matti Järvisalo: “Symmetry breaking for inductive logic programming”, CoRR, Vol. abs/2508.06263, 2025.

URL <https://doi.org/10.48550/ARXIV.2508.06263>

Main reference Andrew Cropper, Filipe Gouveia, David M. Cerna: “Honey, I shrunk the hypothesis space (through logical preprocessing)”, CoRR, Vol. abs/2506.06739, 2025.

URL <https://doi.org/10.48550/ARXIV.2506.06739>

Main reference Andrew Cropper, David M. Cerna: “Efficient rule induction by ignoring pointless rules”, CoRR, Vol. abs/2502.01232, 2025.

URL <https://doi.org/10.48550/ARXIV.2502.01232>

Inductive Logic programming (ILP) combines knowledge representation and machine learning [7, 3]. The goal is to search a hypothesis space, consisting of logic programs, to find a hypothesis that generalizes the given training examples and background knowledge. Early ILP approaches such as *foil* iteratively specialize rules till only positive examples are accepted and repeat this process until all positive examples are accepted. This process was refined in systems like *Aleph* by restricting the search space of specializations to hypotheses that θ -subsume the so-called *bottom clause*, i.e., the most specific clause which accepts a given positive example.

Modern ILP often follows the *meta-learning* approach, i.e., perform search through an encoding of the hypothesis space. In particular, *Popper* [4] encodes the generation of logic programs modulo the background knowledge as an ASP program (*generate phase*). As *Popper* tests the hypothesis produced by the generate phase, additional constraints are added guiding the search at subsequent steps. This leads to a significant decrease in the number

of tested hypotheses while maintaining optimality guarantees. These constraints are based on propositional subsumption and are thus computationally feasible to check. Popper's approach to ILP has led to significant advancements in the tasks such systems can handle. For example learning with *negative invention* [1], *Higher-order* [8], *very large programs* [5], and competitive performance on ARC [6].

In this abstract, we focus on alternative constraints based on the semantics of the background knowledge and **symmetry breaking**, which can improve performance in all of the above areas. For example, in [2] we introduce the notion of *reducible* literal, i.e., given a rule r containing l , the following holds: $B \models r \leftrightarrow r \setminus \{l\}$ where B is the background knowledge. We use this concept to introduce additional constraints while learning. In some cases, we can constrain the hypothesis space prior to learning, for example, if a predicate in the background knowledge is satisfied by a finite number of instantiations. This limits the number of times it can occur in a rule. For example, consider the following where $r_1 \preceq_\theta r_2$:

$$\begin{aligned} B &= \{ \text{edge}(a, b), \text{edge}(b, c), \text{edge}(c, a) \} \\ r_1 &= h \leftarrow \text{edge}(A, B), \text{edge}(B, C), \text{edge}(C, D), \text{edge}(D, E) \\ r_2 &= h \leftarrow \text{edge}(A, B), \text{edge}(B, C), \text{edge}(C, A) \\ \theta &= \{ D \mapsto A, E \mapsto B \} \end{aligned}$$

We refer to this property as *recall redundant*, similar properties are definable for literal combinations that make a rule unsat, literal combinations that are always true, and implication relations between literals. A final approach to constraining the hypothesis space is to remove rule variants. This is achieved by ordering the literals by their argument tuples and normalizing with respect to certain properties.

$$\begin{aligned} r_1 &: h(A, B) \leftarrow p(A, E), p(B, C), p(C, D). \\ r_2 &: h(A, B) \leftarrow p(A, C), p(B, E), p(C, D). \end{aligned}$$

Observe $r_1 \sigma_2 = r_2$ where $\sigma_2 = \{E \mapsto C, C \mapsto E\} \{E \mapsto D\}$. In the above case, we order the variables and ensure that variables are witnessed by smaller variables (with respect to a variable order). This normalization is sound (does not remove all variants), but incomplete. We will cover these approaches in more detail during the talk.

References

- 1 David M. Cerna and Andrew Cropper. Generalisation through negation and predicate invention. *AAAI*, 38(9):10467–10475, 2024.
- 2 Andrew Cropper and David M. Cerna. Efficient rule induction by ignoring pointless rules, 2025.
- 3 Andrew Cropper and Sebastijan Dumancic. Inductive logic programming at 30: A new introduction. *J. Artif. Intell. Res.*, 74:765–850, 2022.
- 4 Andrew Cropper and Rolf Morel. Learning programs by learning from failures. *Mach. Learn.*, 110(4):801–856, 2021.
- 5 Céline Hocquette, Andreas Niskanen, Rolf Morel, Matti Järvisalo, and Andrew Cropper. Learning big logical rules by joining small rules. In *IJCAI*, pages 3430–3438, 2024.
- 6 Céline Hocquette and Andrew Cropper. Relational decomposition for program synthesis, 2025.
- 7 Stephen Muggleton. Inductive logic programming. *New Generation Computing*, 8(4):295–318, 1991.
- 8 Stanislaw J. Purgal, David M. Cerna, and Cezary Kaliszyk. Learning higher-order logic programs from failures. In Luc De Raedt, editor, *IJCAI*, pages 2726–2733, 2022.

3.3 Learning cancer programs

Jasmin Fisher (University College London, GB)

License  Creative Commons BY 4.0 International license
© Jasmin Fisher

Digital twins of cancer tumours are emerging as transformative tools in oncology, enabling more personalised and adaptive treatment strategies. In this talk, I will present a growing library of cancer digital twins spanning multiple cancer types, including breast, lung, skin, and brain. These interpretable, executable models capture molecular mechanisms within tumour cells and their microenvironment, allowing us to uncover and anticipate treatment-resistance mechanisms and design patient-specific strategies to improve outcomes. Currently, these models are manually constructed by experts, a process that is time-consuming and costly. The next challenge is to harness generative AI to learn from vast cancer datasets and automatically build explainable, executable programs (“glass-box” models). This approach would accelerate model development, enhance scalability and align with the goals of inductive programming, making it a timely topic for the Dagstuhl Seminar.

3.4 Computer-Aided Programming and Problem Solving

Frank Jäkel (TU Darmstadt, DE)

License  Creative Commons BY 4.0 International license
© Frank Jäkel

Computers are tools for thought. Scientists and engineers use computers to solve problems that they cannot solve without them. However, more often than not learning the relevant tools is cumbersome. An important goal of basic research in cognitive science is to build the conceptual foundations for future AI tools that naturally support human problem solving. In this talk I will present the results of two studies on how people solve two important classes of problems: constraint-satisfaction problems and guesstimation problems. I will also discuss how these results might inform the development of human-AI interaction.

3.5 Inductive Functional Programming: Igor2 (vs ChatGPT) and Generalized I/O-Examples

Emanuel Kitzelmann (Technische Hochschule Brandenburg, DE)

License  Creative Commons BY 4.0 International license
© Emanuel Kitzelmann

The inductive functional programming system Igor2 induces recursive functions from i/o-examples over user-defined algebraic data types. In this talk, I review its refinement operators using the task of swapping two elements in a list at specified indices, assuming no library functions are available (only list constructors and integers). I demonstrate that even ChatGPT 5.1 still struggles with this problem. Based on an example from the Lists benchmark for inductive programming, I introduce and motivate the problem of generalizing i/o-examples as a preprocessing step for recursive generalization.

3.6 LLM-based feature generation from text for interpretable machine learning

Tomáš Kliegr (*University of Economics – Prague, CZ*)

License © Creative Commons BY 4.0 International license

© Tomáš Kliegr

Joint work of Vojtech Balek, Lukáš Sýkora, Vilém Sklenák, Tomáš Kliegr

Main reference Vojtech Balek, Lukáš Sýkora, Vilém Sklenák, Tomáš Kliegr: “LLM-based feature generation from text for interpretable machine learning”, *Mach. Learn.*, Vol. 114(11), p. 241, 2025.

URL <https://doi.org/10.1007/S10994-025-06867-1>

Applying symbolic rule learning to unstructured text is difficult: bag-of-words features produce sparse, brittle rules, while neural embeddings deliver accuracy at the cost of transparency. We bridge this gap in two complementary ways. First, we present SMER, a transparency approach for classifiers built from averaged word embeddings + logistic regression, where logit linearity enables an exact (perfect-fidelity) decomposition of document decisions into word-level contributions. Second, we propose a structural approach that uses LLMs to map documents to low-dimensional, human-interpretable concepts (e.g., rigor, novelty, replicability), which can be directly consumed by symbolic learners such as action rules. On scientometric benchmarks, fusing LLM-derived concepts with standard text features improves predictive performance over either representation alone, while remaining interpretable. We also address semantic bias in LLM prompting – where priors can override logic – via the “Meaningless is better” hypothesis: hashing bias-inducing words into internally referenceable strings improves performance on logical and statistical tasks. We conclude with an ecosystem overview, including the FeatureLab application and a GPU-accelerated action-rules package, and illustrate applications in research-quality prediction and microbial media modeling.

References

- 1 V. Balek, L. Sýkora, V. Sklenák, and T. Kliegr, “LLM-based feature generation from text for interpretable machine learning,” *Machine Learning*, vol. 114, no. 11, p. 241, 2025. doi: 10.1007/s10994-025-06867-1.
- 2 L. Dvořáčková, M. P. Joachimiak, M. Černý, A. Kubecová, V. Sklenák, and T. Kliegr, “Explaining word embeddings with perfect fidelity: a case study in predicting research impact,” *Machine Learning*, vol. 114, no. 12, p. 265, 2025. doi: 10.1007/s10994-025-06870-6.
- 3 M. Chadimová, E. Jurásek, and T. Kliegr, “Meaningless is better: Hashing bias-inducing words in LLM prompts improves performance in logical reasoning and statistical learning,” *Information Processing & Management*, vol. 63, no. 3, p. 104493, 2026. doi: 10.1016/j.ipm.2025.104493.
- 4 L. Sýkora and T. Kliegr, “action-rules: GPU-accelerated Python package for counterfactual explanations and recommendations,” *SoftwareX*, vol. 29, p. 102000, 2025. doi: 10.1016/j.softx.2024.102000.
- 5 P. Máša, T. Kliegr, and M. P. Joachimiak, “Explainable rule-based prediction of cultivation media for microbes,” *Computational and Structural Biotechnology Journal*, vol. 27, pp. 5194–5206, 2025. doi: 10.1016/j.csbj.2025.10.014.

3.7 Generating Industrial Control Logic from Natural Language Requirements

Heiko Kozirolek (ABB – Mannheim, DE)

License  Creative Commons BY 4.0 International license

© Heiko Kozirolek

Joint work of Heiko Kozirolek, Thilo Braun, Virendra Ashiwal, Sofia Linsbauer, Marthe Ahlgreen Hansen, Karoline Grotterud

Main reference Heiko Kozirolek, Thilo Braun, Virendra Ashiwal, Sofia Linsbauer, Marthe Ahlgreen Hansen, Karoline Grotterud: “Spec2Control: Automating PLC/DCS Control-Logic Engineering from Natural Language Requirements with LLMs – A Multi-Plant Evaluation”, CoRR, Vol. abs/2510.04519, 2025.

URL <https://doi.org/10.48550/ARXIV.2510.04519>

Distributed control systems (DCS) manage the automation for many industrial production processes (e.g., power plants, chemical refineries, steel mills). Programming the software for such systems remains a largely manual and tedious process, incurring costs of millions of dollars for extensive facilities. Large language models (LLMs) have been found helpful in generating DCS control logic, resulting in commercial copilot tools. Today, these tools are focused on textual notations, they provide limited automation, and have not been tested on large datasets with realistic test cases. We introduce Spec2Control, a highly automated LLM workflow to generate graphical control logic directly from natural language user requirements. Experiments using an open dataset with 10 control narratives and 65 complex test cases demonstrate that Spec2Control can successfully identify control strategies, can generate 98.6% of correct control strategy connections autonomously, and can save between 94-96% of human labor. Spec2Control is being integrated into commercial ABB engineering tools, but is also available as an open-source variant for independent validation.

3.8 On Solving Visual Analogies

Johannes Langer (Universität Bamberg, DE)

License  Creative Commons BY 4.0 International license

© Johannes Langer

Joint work of Johannes Langer, Ute Schmid

Main reference Johannes Langer, Ute Schmid: “Flexible Segmentation for Rule Learning over Object Representations in Abstract Visual Reasoning”, in Proc. of the Twelfth Annual Conference on Advances in Cognitive Systems, 2025.

URL <https://openreview.net/forum?id=PSsAsrCkFx>

The ability to reason in abstract domains is a key component of human intelligence, and is tested in many intelligence tests such as Raven’s Progressive Matrices (RPM). Such tests often require humans to perform analogical reasoning over visually presented problem premises. A more recent and AI-focused example of such a test is the Abstraction and Reasoning Corpus (ARC). To solve such analogies in a human-like manner, we attempted to use Inductive Logic Programming due to its ability to learn relations from little training data. This talk discussed the challenges of that approach and new ideas which arise from it.

3.9 Evaluating Large Language Models in Programming-by-Example

Félix Martí Pérez (Technical University of Valencia, ES)

License © Creative Commons BY 4.0 International license
© Félix Martí Pérez

Traditionally, PBE relied on Inductive Programming (IP) systems based on symbolic logic. However, the rise of Large Language Models (LLMs) offers a powerful alternative. This presentation starts by explaining our past work on evaluating the performance of LLMs against traditional IP solvers (MagicHaskeller) on a curated dataset of both standard algorithmic problems and novel “atypical” tasks. We demonstrate that while LLMs offer superior flexibility, they exhibit significant language bias (favouring Python over Haskell) and struggle with novel logic patterns. We then summarise recent research on evaluating LLMs within the PBE paradigm and conclude by outlining current limitations and promising future directions for building more robust and general PBE benchmarks.

3.10 Beyond Performance: Towards Reliable Large Language Models

Yael Moros Daval (Technical University of Valencia, ES)

License © Creative Commons BY 4.0 International license
© Yael Moros Daval
Joint work of Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, José Hernández-Orallo
Main reference Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, José Hernández-Orallo: “Larger and more instructable language models become less reliable”, *Nat.*, Vol. 634(8032), pp. 61–68, 2024.
URL <https://doi.org/10.1038/S41586-024-07930-Y>

Although scaling and alignment improve LLM capability, they also undermine reliability. Models continue to make errors on simple tasks, and difficulty no longer predicts where failures will occur. Refusal behaviour diminishes as models produce fluent but incorrect answers that users often fail to detect. Confidence-estimation and abstention methods reintroduce calibrated uncertainty signals, helping models recognise low-confidence situations and avoid misleading outputs.

3.11 ConDABench: Interactive Evaluation of Data Analysis Assistants

Arjun Radhakrishna (Microsoft – Redmond, US)

License © Creative Commons BY 4.0 International license
© Arjun Radhakrishna
Joint work of Avik Dutta, Priyanshu Gupta, Hosein Hasanbeig, Rahul Pratap Singh, Harshit Nigam, Sumit Gulwani, Arjun Radhakrishna, Gustavo Soares, Ashish Tiwari
Main reference Avik Dutta, Priyanshu Gupta, Hosein Hasanbeig, Rahul Pratap Singh, Harshit Nigam, Sumit Gulwani, Arjun Radhakrishna, Gustavo Soares, Ashish Tiwari: “ConDABench: Interactive Evaluation of Language Models for Data Analysis”, *CoRR*, Vol. abs/2510.13835, 2025.
URL <https://doi.org/10.48550/ARXIV.2510.13835>

Real-world data analysis tasks often come with under-specified goals and unclear data. In the setting of data analysis assistants, user interaction is necessary to understand and disambiguate a user’s intent, and hence, is essential for solving these complex tasks. Existing benchmarks for evaluating LLM-driven data analysis assistants on these tasks do not capture

these complexities or provide first-class support for interactivity. In this task, I will discuss ConDABench, a benchmarking framework and a collection of tasks for evaluating data analysis assistants in the interactive setting. In ConDABench, a user-proxy agent plays the role of the user and is responsible for steering the assistant and responding to any questions or clarifications from the assistant. ConDABench provides a set of 1420 real-world data analysis tasks, each of which asks the data analysis assistant to reproduce key results from public data journalism articles, starting from the raw unclean data. ConDABench is an avenue for model builders to measure progress towards truly collaborative models that can complete complex interactive data analysis tasks.

3.12 A Short History of Inductive Programming

Ute Schmid (Universität Bamberg, DE)

License © Creative Commons BY 4.0 International license
© Ute Schmid

In the talk, the history of the research field of inductive programming (IP) and the Dagstuhl Seminar series (since 2013) as well as the preceding workshops on Approaches and Applications of Inductive Programming (AAIP, since 2005) has been presented. Inductive programming research (see Flener and Schmid 2008; 2024) started in the 1970ies with a focus on construction of functional (Lisp) programs from input/output examples. First approaches were data-driven, that is, generalisazion over the structure of the examples. These were complemented with generate-and-test approaches for functional as well as logic program synthesis. Currently, focus is on prompt-based generation from LLMs. Besides algorithmic approaches to inductive programming, in the talk relations to cognitive science research on learning complex productive rules from examples (e.g., Schmid and Kitzelmann, 2011; Rule et al., 2024) as well as programming education (e.g., Phung et al., 2023) have been presented.

References

- 1 Flener, P., & Schmid, U. (2008). An introduction to inductive programming. *Artificial Intelligence Review*, 29(1), 45-62.
- 2 Flener, P., & Schmid, U. (2024). Inductive Programming. In *Encyclopedia of Machine Learning and Data Science*. Springer, New York, NY.
- 3 Schmid, U., & Kitzelmann, E. (2011). Inductive rule learning on the knowledge level. *Cognitive Systems Research*, 12(3-4), 237-248.
- 4 Rule, J. S., Piantadosi, S. T., Cropper, A., Ellis, K., Nye, M., & Tenenbaum, J. B. (2024). Symbolic metaprogram search improves learning efficiency and explains rule learning in humans. *Nature Communications*, 15(1), 6847.
- 5 Phung, T., Pădurean, V. A., Cambronero, J., Gulwani, S., Kohn, T., Majumdar, R., ... & Soares, G. (2023, August). Generative AI for programming education: benchmarking ChatGPT, GPT-4, and human tutors. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 2* (pp. 41-42).

3.13 Prompt Programming

Adish Singla (MPI-SWS – Saarbrücken, DE)

License © Creative Commons BY 4.0 International license
 © Adish Singla
URL <https://www.promptprogram.org/>

Computing students increasingly rely on generative AI tools for programming assistance, often without formal instruction or guidance. This change highlights the need to teach students how to interact effectively with AI models, particularly through natural-language prompts, to generate and critically evaluate code when solving computational tasks. To address this, we developed a novel platform for prompt programming. This talk will demonstrate the platform, discuss initial results, and highlight plans for future work.

3.14 Execution-guided within-prompt search for programming-by-example

Gust Verbruggen (Microsoft – Keerbergen, BE) and Mukul Singh (Microsoft – Redmond, US)

License © Creative Commons BY 4.0 International license
 © Gust Verbruggen and Mukul Singh
Joint work of Gust Verbruggen, Sumit Gulwani, Vu Le, Ashish Tiwari, Mukul Singh
Main reference Gust Verbruggen, Ashish Tiwari, Mukul Singh, Vu Le, Sumit Gulwani: “Execution-guided within-prompt search for programming-by-example”, in Proc. of the The Thirteenth International Conference on Learning Representations, 2025.
URL <https://openreview.net/forum?id=PY56Wur7S0>

Large language models (LLMs) can generate code from examples without being limited to a DSL, but they lack search, as sampled programs are independent. We use an LLM as a policy that generates lines of code and then join these lines of code to let the LLM implicitly estimate the value of each of these lines in its next iteration. We further guide the policy and value estimation by executing each line and annotating it with its results on the given examples. This allows us to search for programs within a single (expanding) prompt until a sound program is found, by letting the policy reason in both the syntactic (code) and semantic (execution) space. We evaluate within-prompt search on straight-line Python code generation using five benchmarks across different domains (strings, lists, and arbitrary Python programming problems). We show that the model uses the execution results to guide the search and that within-prompt search performs well at low token budgets. We also analyze how the model behaves as a policy and value, show that it can parallelize the search, and that it can implicitly backtrack over earlier generations.

3.15 Cyclic Bayesian networks: From probabilistic modelling to probabilistic programming

Felix Weitzkämper (German UDS – Potsdam, DE)

License © Creative Commons BY 4.0 International license
 © Felix Weitzkämper
Joint work of Vera Koponen, Felix Weitzkämper

Bayesian networks are a popular target formalism for statistical/probabilistic machine learning as well as knowledge-based model construction. So what distinguishes learning Bayesian networks from inductive probabilistic programming? In particular, what is missing from

Bayesian networks to be counted a full probabilistic programming language? In this talk, I argue that after relational Bayesian networks [1] have addressed a key expressive deficiency, recursion is the main missing ingredient. For fully random Bayesian networks, in which all probabilities are properly between 0 and 1, approaches based on irreducible Markov chains provide a uniquely determined semantics for networks with cycles [2, 3]. I argue that fixed-point methods from traditional logic and discrete mathematics are able to give semantics also to a large class of cyclic networks with deterministic constraints, which suffices for instance to embed probabilistic logic programming.

References

- 1 Manfred Jaeger. Relational Bayesian networks. Proceedings of UAI 1997, pp. 266–273.
- 2 Mieczysław Kłopotek. Cyclic Bayesian network: Markov process approach. *Studia Informatica: systems and information technology* 1(7), 47–55 (2006)
- 3 Christel Baier, Clemens Dubslaff, Holger Hermanns, Nikolai Käfer. On the Foundations of Cycles in Bayesian Networks. Principles of Systems Design. Lecture Notes in Computer Science, vol 13660, pp 343–363, 2022

4 Working groups

4.1 Inductive programming for complex biological problems

David Cerna (Dynatrace Research – Linz, AT), Jasmin Fisher (University College London, GB), Pierre Flener (Uppsala University, SE), Tomáš Kliegr (University of Economics – Prague, CZ), Heiko Koziol (ABB – Mannheim, DE), Raoul van Doren (ETH Zürich, CH), and Felix Weitekämper (German UDS – Potsdam, DE)

License  Creative Commons BY 4.0 International license

© David Cerna, Jasmin Fisher, Pierre Flener, Tomáš Kliegr, Heiko Koziol, Raoul van Doren, and Felix Weitekämper

The motivating problem of the working group is the induction of gene regulatory networks in biology. Those networks are modelled as qualitative networks, a multi-valued generalisation of Boolean networks. Gene regulatory networks are highly relevant for understanding the mechanism of tumor development.

Some of the defining features of this domain, which add intrinsic interest as a target domain for inductive programming approaches, are non-determinism and heavy use of recursion. Furthermore, transparency of both the final output and the methodology used to get there is essential for trusting the discovered networks.

Recent developments in symbolic and sub-symbolic machine learning raise the prospect of inducing such gene regulatory models directly from data. This raises several challenges:

The first challenge is gathering reliable data in a machine-readable format. Despite recent progress in the availability of structured databases, their reliability can be variable. Due to the discrete nature of the systems, single errors can have unforeseen consequences. Hence, tight quality control with human oversight must be integrated into the data curation process.

The second challenge is the induction of the qualitative network from the garnered data. This involves integrating these various forms of examples into a transparent learning framework that provides appropriate guarantees. Existing approaches involve augmenting human-created network structure with a combination logic supported by machine learning.

The third challenge is maintaining the integrity of existing gene regulatory networks in the face of new data. This extends the traditional belief revision problem by the issue of potentially augmenting the vocabulary involved in the network.

4.2 Neuro-Symbolic Program Induction

Quinten Dewulf (KU Leuven, BE), Wang-Zhou Dai (Nanjing University – Suzhou, CN), Cesar Ferri Ramirez (Technical University of Valencia, ES), Johannes Fürnkranz (Johannes Kepler Universität Linz, AT), Johannes Langer (Universität Bamberg, DE), and Zachary Siegel (MIT – Cambridge, US)

License  Creative Commons BY 4.0 International license

© Quinten Dewulf, Wang-Zhou Dai, Cesar Ferri Ramirez, Johannes Fürnkranz, Johannes Langer, and Zachary Siegel

4.2.1 Two Types of Neuro-Symbolic Program Induction

In this discussion group we focus on neuro-symbolic program induction. It quickly became apparent that there was an important distinction between two ways to interpret the topic of neuro-symbolic program induction – as *[Neuro-Symbolic] Program Induction* and *[Neuro-Symbolic Program] Induction*. The former addresses neuro-symbolic approaches for primarily inducing *symbolic* programs. Examples for this line of research include discretizing non-symbolic inputs for symbolic program induction [4], or neurally guiding program search as in DreamCoder [3]. The latter describes the problem of inducing programs which themselves are inherently neuro-symbolic in nature.

4.2.2 What are the benefits of using neuro-symbolic methods to synthesize symbolic programs?

Consider two techniques for synthesizing programs: a symbolic-only system that performs enumerative search over a DSL, and a large language model (LLM) from which we sample programs. The symbolic-only system comes with completeness guarantees and can explore programs that an LLM would not otherwise discover if those types of programs are not within the LLM’s training distribution. In addition, particularly if the domain is known or constrained, heuristics can be used to guide search over programs to make the system quite powerful. On the other hand, such systems will struggle to produce longer programs since the search space increases exponentially with the depth of the program. Although techniques such as library learning and forming abstractions can help, an unguided search will likely struggle to work in a general setting and lacks the type of “common sense” knowledge that people may use when writing code.

The neural-only system can sample programs effectively that are within the training distribution of LLMs. We see the success of such methods due to the prevalence of using coding assistants, which are already in high use by software developers. These systems, however, struggle to generate code out of distribution, and additionally struggle to manage large code bases. Each piece of code sampled from an LLM is relatively expensive compared to a piece of code generated via search, so the LLM’s prior is highly important for generating useful programs.

These benefits naturally raise the question of whether a neuro-symbolic system could obtain the advantages of both methods; namely, combining the “common sense” statistical knowledge of neural networks with the completeness guarantees and other heuristics of symbolic synthesizers. One example is DreamCoder [3], which uses a neural network to guide symbolic search over programs. The neural network can be viewed as a heuristic that accelerates search, allowing the system to discover programs that would not be encountered in a reasonable time frame otherwise. We believe that future promising work might use neural methods such as LLMs to guide symbolic search for programs, potentially in combination with other heuristics to guide program search more effectively.

4.2.3 What is a neuro-symbolic program?

There are already several languages for representing neuro-symbolic programs. Best-known is possibly DeepProbLog [6], which extends probabilistic logic programming with subsymbolic predicates that can, for example, be defined with deep neural networks. Similarly, Neur-ASP [9] extends Answer Set Programming with neural networks for estimating a probability distribution over atomic facts. Scallop [5] is a language similar to Datalog, which features an underlying differentiable reasoning engine. ULLER [8] is an attempt to unify various approaches and provide a unified framework for learning and reasoning.

However, all these approaches aim at extending conventional programming paradigms (such as Python, Prolog, ASP, etc.) with mechanisms for bringing in neurosymbolic predicates or functions. In all cases, the program itself has a clearly identifiable symbolic “spine” that connects neurosymbolic predicates or functions. The question of what constitutes a neuro-symbolic program may go deeper and could require re-thinking the paradigm of programming. *Vibe programming*, where one obtains a conventional program through an iterative neuro-symbolic refinement process, is one example. A related possibility is that the *process* that defines the final program may itself be considered the program, and the final program text be less central – analogous to how compilers made machine and assembly code largely obsolete for many programmers. In the seminar we saw several neuro-symbolic programming paradigms that foreshadow such developments.

4.2.4 What are the benefits of synthesizing neuro-symbolic programs?

The main benefits of neuro-symbolic programs as a representation come from inheriting advantages of both symbolic and neural representations. Symbolic programs are naturally verifiable and interpretable, since we understand how they work, which makes them more usable in domains requiring performance guarantees. Relatedly, they can be more accurate for specific problems compared to neural methods: if we understand the structure of a problem, a symbolic solution can provide guarantees on finding solutions. Symbolic programs can also generalize well by leveraging representations tailored to the problem; by learning appropriate abstractions, solutions can become reusable.

Despite these benefits, compared to neural techniques, symbolic programs can be harder to scale with access to additional data, since there generally is not an analogue of “taking a gradient step” to update a symbolic program in response to data. With access to large datasets, neural methods can be scaled up to be quite general in the sense of covering a broad distribution of inputs. This scalability is shown clearly in the success of large neural models like LLMs, which achieve impressive performance across a wide range of tasks.

There are often situations where a problem can be well specified with some outer program structure, even if certain internal components may be better represented with neural predicates. For instance, consider controlling the behavior of a self-driving car at a traffic stop. There is a well-defined symbolic behavior (when the light is green, proceed; when the light is red, stop). On the other hand, identifying the color of the traffic light may be easier to specify using a neural predicate. By combining such neural predicates with symbolic decision flow, neuro-symbolic programs can provide additional guarantees and interpretability over fully neural systems, and additional flexibility over symbolic-only models.

4.2.5 Why induce neuro-symbolic programs in a joint fashion?

Inducing neuro-symbolic programs by jointly learning neural weights and symbolic knowledge is, in general, not reducible to simpler symbolic program induction or to learning only neural weights. Many methods in neuro-symbolic AI rely on hand-engineered symbolic knowledge

to provide guarantees, data efficiency, and improved accuracy when training their neural components. This makes such models highly dependent on the quality and completeness of the knowledge provided by human experts. Inducing neuro-symbolic programs from data alleviates this problem by allowing missing or unknown symbolic knowledge to be learned during training.

One might argue that there is no need for joint learning of neural weights and symbolic knowledge in neuro-symbolic AI, since neural perception can be outsourced to generalist vision-language models, reducing the problem to symbolic program induction [1]. However, such methods may fail to account for problem instances containing niche subsymbolic input that is out of distribution for any generalist model, and they neglect the considerable cost associated with using such models.

4.2.6 How to approach induction of neuro-symbolic programs

Jointly learning structure and parameters is difficult because the search spaces interact: structural choices redefine the parameter landscape, and parameter updates can bias the search toward suboptimal structures. Neuro-symbolic systems face this acutely, as the symbolic component demands discrete structural commitments while the neural component relies on continuous optimization. Effective approaches typically alternate or blend these two processes, using differentiable relaxations, guided search, or meta-learning priors to keep the joint space navigable.

A practical approach is to employ abduction-guided search, where candidate neuro-symbolic programs are generated, scored against raw data, and refined using differentiable components to prune the hypothesis space. Concretely, one can combine perceptual models with an abductive logic engine in a loop: propose symbolic rules, test them against observations, and iteratively update the neural perception and the symbolic hypothesis until convergence [2]. Another approach (inspired by meta-learning) uses a symbolic program representation for compositional generalization together with neural program synthesis, enabling efficient guess-and-verify in program space [7].

References

- 1 Vít Balek, Lukáš Sýkora, Václav Sklenák, and Tomáš Kliegr. *LLM-based feature generation from text for interpretable machine learning*. *Machine Learning*, 114(11):241, 2025. <https://doi.org/10.1007/s10994-025-06867-1>
- 2 Wang-Zhou Dai and Stephen H. Muggleton. *Abductive knowledge induction from raw data*. *arXiv preprint arXiv:2010.03514*, 2020.
- 3 Kevin Ellis, Lucas Wong, Michael Nye, Mathias Sablé-Meyer, Luc Cary, Laura Anaya Pozo, Luke Hewitt, Armando Solar-Lezama, and Joshua B. Tenenbaum. *DreamCoder: Growing generalizable, interpretable knowledge with wake-sleep Bayesian program learning*. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220050, 2023. <https://doi.org/10.1098/rsta.2022.0050>
- 4 Johanna Langer and Ute Schmid. *Flexible Segmentation for Rule Learning over Object Representations in Abstract Visual Reasoning*. In *Twelfth Annual Conference on Advances in Cognitive Systems*, 2025. <https://openreview.net/forum?id=PSsAsrCkFx>
- 5 Zhuo Li, Junjie Huang, and Mayur Naik. *Scallop: A Language for Neurosymbolic Programming*. *arXiv arXiv:2304.04812*, 2023. <https://doi.org/10.48550/arXiv.2304.04812>
- 6 Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. *DeepProbLog: Neural Probabilistic Logic Programming*. *arXiv arXiv:1805.10872*, 2018. <https://doi.org/10.48550/arXiv.1805.10872>

- 7 Maxwell Nye, Armando Solar-Lezama, Joshua B. Tenenbaum, and Brenden M. Lake. *Learning compositional rules via neural program synthesis*. In *Advances in Neural Information Processing Systems*, volume 33, pages 10832–10842, 2020.
- 8 Emile van Krieken, Selim Badreddine, Robin Manhaeve, and Enrico Giunchiglia. *ULLER: A Unified Language for Learning and Reasoning*. *arXiv* arXiv:2405.00532, 2024. <https://doi.org/10.48550/arXiv.2405.00532>
- 9 Zhun Yang, Adam Ishay, and Joohyung Lee. *NeurASP: Embracing Neural Networks into Answer Set Programming*. *arXiv* arXiv:2307.07700, 2023. <https://doi.org/10.48550/arXiv.2307.07700>

4.3 Gradual refinement in coarse-to-fine concept induction and human-like iterative programming

Alex Lew (Yale University – New Haven, US), Maddy Bowers (MIT – Cambridge, US), Gabriel Grand (MIT – Cambridge, US), Emanuel Kitzelmann (Technische Hochschule Brandenburg, DE), David Nagy (Universität Tübingen, DE), Tejaswini Ojha, Mukul Singh (Microsoft – Redmond, US), Tobias Thomas (TU Darmstadt, DE), Dominik Ürüm (TU Darmstadt, DE), and Hanqi Zhou (Universität Tübingen, DE)

License © Creative Commons BY 4.0 International license

© Alex Lew, Maddy Bowers, Gabriel Grand, Emanuel Kitzelmann, David Nagy, Tejaswini Ojha, Mukul Singh, Tobias Thomas, Dominik Ürüm, and Hanqi Zhou

A key research question for the field of program induction is how to generate programs *iteratively*, refining them gradually in response to various signals. We investigate this question in two contexts. (1) Modeling the process by which engineers iteratively improve their code, on both short time scales (fixing bugs, refactoring) and longer ones (evolving large-scale codebases over years); and (2) modeling the coarse-to-fine aspects of how humans learn concepts from examples.

4.3.1 Automating human-like iterative coding

Human engineers develop programs in stages: A programmer may develop one function in an iterative fashion over the course of minutes or hours, writing tests, sketching a possible approach with comments, filling in the sketch with code, running tests and inspecting results, fixing bugs, and so on [1]. A key question is how to model or automate these iterative processes.

Model architectures for iterative revision. Today, autoregressive (“causal”) transformers are the predominant architecture for LLMs. Such architectures (1) generate tokens left-to-right at inference time, and (2) are trained to maximize the left-to-right generation probability of programs in the training data. One important research direction for modeling iterative coding processes is to explore alternative architectures that may be able to learn to generate text in more human-like ways.

Diffusion models are one such alternative. Current diffusion models for text assume Gaussian noise in latent space, which recovers some but not the whole structure of the iterative process of coding [2, 5]. Thus, a possible extension would be to replace the continuous noise process with an explicit, possibly discrete, noise process in the token or text space. Another one would be to infer the latent generative process of the code, which is in reality never left-to-right, but involves the whole range of the iterative complexity described above. Yet,

the latter approach requires either a dataset including at least some of the intermediate steps of the generative process, or a more sophisticated training algorithm that co-learns a noiser and denoiser (e.g., based on EM or variational inference – see, e.g., the STaR family of algorithms [3, 4], or using amortized sampling of the latent representation over data [6].

A further architectural challenge is that, in human problem solving, there appear to be intermediate memory representations that are less granular than the raw token sequence in an LLM’s context window, yet more granular and rapidly updated than what is stored in the weights. For coding agents, this might include persistent information about how to interact with a given user and project such as preferred coding conventions or the appropriate level of detail for clarifying questions from the coding agent. At present, these forms of task context are implemented through a hand-engineered outer loop that incorporates the LLM, injecting this information into the context window via system prompts, personal custom instructions, and explicit context distillation across sessions. An open question is whether some of this intermediate task memory can be incorporated into the architecture and learning dynamics of the model itself, rather than relying on our current understanding and ability to explicitly specify the human coding process.

4.3.2 Modeling human-like coarse-to-fine concept induction

The second question of our discussion group centered on modeling *concept learning*, rather than human software engineering. When solving concept induction problems, people often begin from coarse, vague hypotheses that they incrementally refine into increasingly precise explanations [9]. The talk “Concept learning as coarse-to-fine probabilistic program induction” [7], presented on the first day of the seminar, offered an approach to modeling this kind of reasoning by representing coarse hypotheses as probabilistic programs and evaluating their quality in terms of the likelihood of the observed examples under the programs. However, an open question with the approach is how to decide what set of refinements to consider. In particular, having a small core set of low-level refinements, such as those given implicitly by the productions of the target language grammar, can be suboptimal since we may require applying multiple low-level refinements in sequence before seeing an improvement in hypothesis quality as measured by the likelihood of the data under the program. This suggests introducing new individual refinements that correspond to introducing a whole chunk of program at once.

We discussed how this kind of refinement learning can be framed as a form of library learning in the style of DreamCoder [8]. Library learning could identify common structure across successful or promising coarse-to-fine search traces, enabling the discovery of useful refinement chunks (e.g., a learned `map`). These chunks can then be added to the library of possible refinements for downstream use in synthesis. This could also help bridge a gap in the refinement process, where the intermediate hypothesis the search would otherwise have to proceed through is judged as a low-quality coarse hypothesis even though the destination is high-quality.

We believe that these insights on the coarse-to-fine aspects of human concept learning are applicable not only to modeling concept induction, but also to scaling synthesis in more general automated coding settings.

References

- 1 J. S. Rule, *The Child as Hacker: Building More Human-Like Models of Learning*, Ph.D. dissertation, Massachusetts Institute of Technology, 2020.
- 2 M. Singh, J. Cambronero, S. Gulwani, V. Le, C. Negreanu, and G. Verbruggen, “CodeFusion: A pre-trained diffusion model for code generation,” in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2023, pp. 11697–11708.
- 3 E. Zelikman, G. R. Harik, Y. Shao, V. Jayasiri, N. Haber, and N. Goodman, “Quiet-STaR: Language models can teach themselves to think before speaking,” in *Proc. First Conf. on Language Modeling*, 2024.
- 4 G. Dong, Y. Chen, X. Li, J. Jin, H. Qian, Y. Zhu, H. Mao, G. Zhou, Z. Dou, and J.-R. Wen, “Tool-Star: Empowering LLM-brained multi-tool reasoner via reinforcement learning,” *arXiv preprint*, 2025.
- 5 M. Singh, G. Verbruggen, V. Le, and S. Gulwani, “Diffusion is a code repair operator and generator,” *arXiv preprint*, 2025.
- 6 S. Venkatraman et al., “Amortizing intractable inference in diffusion models for vision, language, and control,” in *Advances in Neural Information Processing Systems*, vol. 37, A. Globerson et al., Eds., 2024, pp. 76080–76114.
- 7 M. Bowers, A. Lew, W. Qi, J. S. Rule, V. Mansinghka, J. Tenenbaum, and A. Solar-Lezama, “Concept learning as coarse-to-fine probabilistic program induction,” in *Proc. Annual Meeting of the Cognitive Science Society (CogSci)*, vol. 46, 2024.
- 8 K. Ellis, C. Wong, M. Nye, M. Sablé-Meyer, L. Morales, L. Hewitt, and J. B. Tenenbaum, “DreamCoder: Bootstrapping inductive program synthesis with wake-sleep library learning,” in *Proc. 42nd ACM SIGPLAN Int. Conf. on Programming Language Design and Implementation (PLDI)*, 2021, pp. 835–850.
- 9 N. D. Goodman, J. B. Tenenbaum, J. Feldman, and T. L. Griffiths, “A rational analysis of rule-based concept learning,” *Cognitive Science*, vol. 32, no. 1, pp. 108–154, 2008.

4.4 Benchmarks and Evaluation

Yael Moros Daval (Technical University of Valencia, ES), Félix Martí Pérez (Technical University of Valencia, ES), Yewen Pu (Nanyang TU – Singapore, SG), Arjun Radhakrishna (Microsoft – Redmond, US), Ute Schmid (Universität Bamberg, DE), and Gust Verbruggen (Microsoft – Keerbergen, BE)

License © Creative Commons BY 4.0 International license
 © Yael Moros Daval, Félix Martí Pérez, Yewen Pu, Arjun Radhakrishna, Ute Schmid, and Gust Verbruggen

Benchmarking and evaluation of AI systems are currently in a state of rapid transition, largely driven by the generalization capabilities of large language models (LLMs). Classical inductive programming benchmarks, for example lists, strings, and grid-based tasks such as ARC, were designed to probe abstraction, compositionality, and data efficiency. At present, many of these benchmarks are effectively saturated: they are solvable either through scale, by learning strong statistical priors over program spaces, or through inference-time computation that enables iterative reasoning and code refinement, often at substantial cost. As a result, functional correctness on static input–output tasks no longer provides a meaningful signal for progress in symbolic or hybrid AI systems. Recent evaluation efforts have therefore shifted toward more realistic and complex settings, including software engineering benchmarks, interactive program synthesis, and agent-style tasks that embed systems in execution environments.

These benchmarks trade ease of evaluation for realism and complexity, frequently requiring partial automation, human assessment, or LLM-based judges. While such approaches better reflect real-world use, they introduce challenges in reproducibility, standardization, and interpretability of results. Overall, the field is moving from closed-form, ground-truth-driven benchmarks toward in-situ evaluation that models user interaction, iterative refinement, and environment feedback. For inductive and logic-based programming, this shift necessitates new benchmarks that explicitly test collaboration between symbolic methods, learned models, and users, rather than raw problem-solving ability alone.

4.5 Programming Education in the Age of Large Language Models

Hila Peleg (Technion – Haifa, IL), Frank Jäkel (TU Darmstadt, DE), Sven Mayer (TU Dortmund, DE), and Sonja Niemann (bidt – München, DE)

License © Creative Commons BY 4.0 International license
© Hila Peleg, Frank Jäkel, Sven Mayer, and Sonja Niemann

The wide adoption of large language models (LLMs) to assist programmers raises the question: How should programming be taught in the future? It is sometimes claimed that students will not have to learn programming at all. However, there will always be demand for expert programmers. There is currently still disagreement on how many people will have to acquire how much programming expertise. If programming assistants become better still, this will increase access to automation for many people who do not know how to code and will not have to learn it either. Therefore, the number of people who feel that they have to learn programming in order to become proficient computer users will decrease. Until very recently, programming and computational thinking were considered essential skills for the 21st century, and educators tried to integrate them into school curricula. Since LLMs are marketed as being able to replace programmers [5], this consensus is under increasing pressure. However, computational thinking will still be required of users of LLMs, and programming is a good way to learn it. Furthermore, if programming were to be taught less, there would be fewer people who potentially become experts down the road.

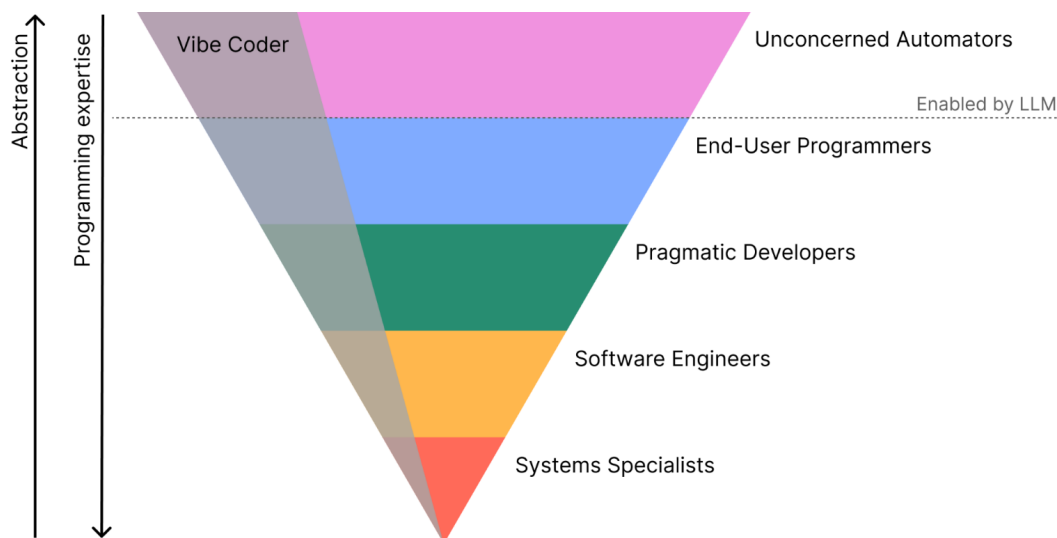


Figure 1 Possible distribution of programmers, abstraction decreases towards the wide base of Unconcerned Automators while programming expertise increases towards the Systems Specialists.

It is useful to think about the people who program today as an inverted pyramid. Most programming is done by users who automate tasks in some application – very often Excel. These end-user programmers are not professional programmers but need some programming skills to get their work done. LLMs currently make this kind of automation accessible to more and more users. These users who don't know how to program will, however, mostly be unconcerned with code. This is in contrast to pragmatic programmers who have more expertise than the end-user programmers. These people have problems that require them to develop custom code, and they often use professional tooling and libraries even though they are not professional software developers. Scientists, engineers, or data analysts fall into this class, but also, e.g., web developers. Their work relies on the work of far fewer software developers who write the libraries and tools that they rely on. Their code, in turn, is built on the work of even fewer system specialists who develop operating systems, compilers, or programming languages. The challenge is to educate each group at the right level while leaving the door open for those on the higher rungs of the pyramid to be able to move to the lower rungs if they need or want to.

To meet this challenge, it is interesting to consider the parallels between programming with an LLM and pair programming, where one person is the “driver,” who only types code, and the other the “navigator,” who decides what code to write. This is like the acceleration mode in Grounded Copilot [1], removing low-level concerns like syntax. How can we ensure that the roles aren't reversed, i.e., that the human user becomes the “driver” and the LLM the “navigator”? Such a reversal would make it difficult for LLM users to acquire the expertise to transition to the lower rungs of the pyramid. One way to avoid this is to design LLM support in a way that it is more constrained for beginners, thus making sure that they have to think about the code generated by the LLM. Such a system only allows for more and more auto-completion as higher levels of programming expertise are “unlocked”. This would be similar to some programming languages specifically created to teach programming at progressively greater degrees of difficulty, like Hedy [2] or Racket [4, 3].

Since expertise is contextual, programmers at all levels might find themselves “vibe coding” at times. Since the definition is cases where the code does not matter, this can be relevant for one-time pieces of code in areas of low expertise, but could also be a very useful tool for things like prototyping an idea, just as a proof-of-concept, generating code that will be thrown out and rewritten with more care and attention later. This makes vibe coding a tool that can be taught, stressing the importance of not keeping any such code around longer than one would keep a prototype, while being aware that this code might find its way into production despite the advice of the tutors.

Programming assistants have the potential to bring the full power of computing to more people, but there is also the danger of skill loss. The design of such tools, therefore, needs to ensure that students can not just solve programming problems faster but also improve their conceptual understanding and programming skills. Otherwise, we might find ourselves in a future without expert programmers at the apex of the inverted pyramid.

References

- 1 Shraddha Barke, Michael B. James, and Nadia Polikarpova. Grounded Copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages*, 7(OOPSLA1), Article 78, 2023.
- 2 Felienne Hermans. Hedy: A gradual language for programming education. In *Proceedings of the 2020 International Computing Education Research Conference (ICER '20)*, pages 1–12, 2020.

- 3 Michael Hanus and Jan Rasmus Tikovsky. A parser generator system for level-based programming languages. In *Proceedings of the Software Engineering Workshops 2016 (SE-WS 2016)*, CEUR Workshop Proceedings, volume 1559, pages 3–24, 2016.
- 4 The Racket Documentation. DrRacket Language Levels. Available at: https://download.racket-lang.org/docs/5.1.2/html/unstable/DrRacket_Language_Levels.html, 2011. Version 5.1.2. Accessed: December 5, 2025.
- 5 Mark Tyson. Jensen Huang says kids shouldn't learn to code – they should leave it up to AI. *Tom's Hardware*, February 25, 2024. Available at: <https://www.tomshardware.com/tech-industry/artificial-intelligence/jensen-huang-advises-against-learning-to-code-leave-it-up-to-ai>.

Participants

- Maddy Bowers
MIT – Cambridge, US
- David Cerna
Dynatrace Research – Linz, AT
- Wang-Zhou Dai
Nanjing University – Suzhou, CN
- Quinten Dewulf
KU Leuven, BE
- Cesar Ferri Ramirez
Technical University of
Valencia, ES
- Jasmin Fisher
University College London, GB
- Pierre Flener
Uppsala University, SE
- Johannes Fürnkranz
Johannes Kepler Universität
Linz, AT
- Gabriel Grand
MIT – Cambridge, US
- Katerina Hrudková
University of Economics –
Prague, CZ
- Frank Jäkel
TU Darmstadt, DE
- Emanuel Kitzelmann
Technische Hochschule
Brandenburg, DE
- Tomáš Kliegr
University of Economics –
Prague, CZ
- Heiko Koziol
ABB – Mannheim, DE
- Johannes Langer
Universität Bamberg, DE
- Alex Lew
Yale University – New Haven, US
- Félix Martí Pérez
Technical University of
Valencia, ES
- Sven Mayer
TU Dortmund, DE
- Yael Moros Daval
Technical University of
Valencia, ES
- David Nagy
Universität Tübingen, DE
- Sonja Niemann
bidt – München, DE
- Stassa Patsantzis
University of Surrey –
Guildford, GB
- Hila Peleg
Technion – Haifa, IL
- Yewen Pu
Nanyang TU – Singapore, SG
- Arjun Radhakrishna
Microsoft – Redmond, US
- Ute Schmid
Universität Bamberg, DE
- Zachary Siegel
MIT – Cambridge, US
- Mukul Singh
Microsoft – Redmond, US
- Adish Singla
MPI-SWS – Saarbrücken, DE
- Tobias Thomas
TU Darmstadt, DE
- Dominik Ürum
TU Darmstadt, DE
- Raoul van Doren
ETH Zürich, CH
- Gust Verbruggen
Microsoft – Keerbergen, BE
- Felix Weitkämper
German UDS – Potsdam, DE
- Hanqi Zhou
Universität Tübingen, DE



Generalized Voronoi Diagrams and Applications

Andreas Alpers^{*1}, David P. Bourne^{*2}, Henning F. Poulsen^{*3}, and
Claudia Redenbach^{*4}

1 University of Liverpool, GB. andreas.alpers@liverpool.ac.uk

2 Heriot-Watt University – Edinburgh, GB. d.bourne@hw.ac.uk

3 Technical University of Denmark – Lyngby, DK. hfpo@dtu.dk

4 RPTU Kaiserslautern-Landau, DE. claudia.redenbach@rptu.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25492 “Generalized Voronoi Diagrams and Applications”, which was held from 30 November to 5 December 2025. The seminar brought together 28 scientists working on various aspects of generalized Voronoi diagrams such as efficient diagram computation, fitting diagrams to observed cell systems or using Voronoi diagrams to model material microstructure. The report contains the abstracts of the tutorials and presentations held during the seminar. It also contains a summary of the discussion sessions that were organized in smaller break-out groups.

Seminar November 30 – December 5, 2025 – <https://www.dagstuhl.de/25492>

2012 ACM Subject Classification Applied computing → Physical sciences and engineering; Theory of computation → Computational geometry; Applied computing → Mathematics and statistics

Keywords and phrases Laguerre tessellation, generalized balanced power diagram, anisotropic Voronoi diagram, cellular materials, grains, polycrystals

Digital Object Identifier 10.4230/DagRep.15.11.157

1 Executive Summary

Andreas Alpers (University of Liverpool, GB, andreas.alpers@liverpool.ac.uk)

David P. Bourne (Heriot-Watt University – Edinburgh, GB, d.bourne@hw.ac.uk)

Henning F. Poulsen (Technical University of Denmark – Lyngby, DK, hfpo@dtu.dk)

Claudia Redenbach (RPTU Kaiserslautern-Landau, DE, claudia.redenbach@rptu.de)

License  Creative Commons BY 4.0 International license

© Andreas Alpers, David P. Bourne, Henning F. Poulsen, and Claudia Redenbach

Motivation

Voronoi diagrams are fundamental geometric structures that divide space into regions based on proximity to points or sites. They play an important role in many fields of science and have become cornerstones of disciplines such as computational geometry, scientific computing, and optimization.

At the heart of this seminar are generalized Voronoi diagrams, with a primary focus on Laguerre diagrams and their rather recent generalizations – anisotropic power diagrams and generalized balanced power diagrams – which introduce piecewise quadratic cell boundaries, enabling more precise modeling of complex structures. Recently these diagrams have drawn significant interest from the materials science community for their remarkable ability to

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Generalized Voronoi Diagrams and Applications, *Dagstuhl Reports*, Vol. 15, Issue 11, pp. 157–178

Editors: Andreas Alpers, David P. Bourne, Henning F. Poulsen, and Claudia Redenbach



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

accurately depict the inner structures of polycrystalline materials, such as metals, ceramics, and rocks, offering valuable insights into their composition and properties. Due to their ability to follow more complex boundary structures, they are also of emerging interest in Voronoi-based image processing applications, e.g., in superpixel generation and image segmentation.

Current methods for computing generalized Voronoi diagrams from imaging data are limited to moderately sized diagrams comprising of about 1,000-10,000 regions. To handle large data sets enabling computations at high resolution, large sample size, and multiple time steps, it is vital to develop efficient algorithms. However, the research communities behind the algorithm development, geometric clustering, optimal transport, stochastic geometry, and computational geometry are largely disjoint, with each field relying on different mathematical foundations and techniques. Applications such as the modelling of materials microstructures by Voronoi diagrams present yet another set of challenges for diagram calculations. Examples are the statistical fitting of models to observed data, fitting for incomplete data where only partial information of the cell system is observed, and handling of material inhomogeneities and edge effects.

The goal of this Dagstuhl Seminar was to bring together leading researchers from the diverse mathematical areas and materials scientists to learn from each other, spark vibrant discussions, identify common grounds, advance their fields, and set the path for successful collaborations towards developing fast and robust generalized Voronoi diagram algorithms addressing the needs of the practitioners working in imaging and microstructure modelling.

Programme

The seminar began with 3-minute/1-slide introductions by all the participants, which was a great way to break the ice since the seminar brought together diverse communities, and a lot of the participants had not met before.

This was followed by five 1-hour tutorials (four on Monday and one on Tuesday) to establish a common language and introduce some basic notions from experimental and computational materials science (Henning F. Poulsen and Orkun Furat), discrete mathematics (Peter Gritzmann), optimal transport (Bruno Lévy), and stochastic geometry (Claudia Redenbach). In each tutorial we saw different applications of generalized Voronoi diagrams and different tools and results.

The tutorials were followed by short research talks (20-minute talks plus 10 minutes of questions and discussion). All participants were invited to speak.

In addition, there were two 2-hour discussion sessions (reports from the breakout groups can be found below) and a software demonstration session, where Romain Quey gave a demo of *Neper* and David Bourne gave a demo of the *SyntheticMic* GUI, which are both software packages for generating synthetic polycrystalline microstructures using Laguerre diagrams. Bruno Lévy gave a tutorial on his geometry-processing library *Geogram*.

In addition to the scientific programme, there was an 8-km hike on Wednesday afternoon and a piano recital by Piere Calka on Thursday evening.

Outcomes

One of the immediate outcomes of the seminar was knowledge exchange between what had previously been quite disjoint communities (materials science, computational and stochastic geometry, discrete mathematics, and optimal transport theory). Another immediate outcome was the identification of several open problems including how to reconstruct a 3D micro-

structure (or diagram) from 2D slices, how to model grain growth using generalized Voronoi diagrams, and how to efficiently compute an analytic representation of anisotropic power diagrams. In addition, initial follow-up activities have already begun, including invitations for research talks and online discussions to develop specific problems identified during the seminar.

Regarding longer-term outcomes, we expect the collaborations initiated at the seminar to result in joint grant applications and research articles, and to stimulate further focused workshops and research visits.

Overall, we feel that this was a very successful seminar. This impression is confirmed by the survey filled in by the participants. We are grateful to Christian Jung for organizing and compiling the abstracts. On behalf of the participants, we would like to thank Dagstuhl for enabling us to hold the seminar at Schloss Dagstuhl, which was the ideal venue.

2 Table of Contents

Executive Summary

Andreas Alpers, David P. Bourne, Henning F. Poulsen, and Claudia Redenbach . . . 157

Overview of Tutorials

Introduction to artificial neural networks and their applications in structural analysis and modeling of polycrystalline materials <i>Orkun Furat, Lukas Fuchs, and Volker Schmidt</i>	162
Diagrams, Clustering, and Coresets – and the representation of polycrystals <i>Peter Gritzmann</i>	163
Optimal Transport: a new asset in the numerical analysis toolbox <i>Bruno Levy</i>	163
Random tessellations <i>Claudia Redenbach</i>	164

Overview of Talks

PyAPD: A Python library for working with minimisation diagrams with GPU acceleration <i>Maciej Buze</i>	164
A few asymptotic results for functionals of Poisson-Voronoi cells <i>Pierre Calka</i>	165
Overlay of Voronoi tessellations for lunar Euclidean MSTs <i>Herbert Edelsbrunner</i>	166
CVT-based data-adaptive family of partitioning algorithms and materials modeling <i>Maria Emelianenko</i>	166
Provably Robust Perspective- n -Point Algorithms for Camera Pose Estimation <i>Dan Feldman</i>	167
An optimal transport model for dynamical shapes, collective motion and cellular aggregates <i>Jean Feydy</i>	168
Data-driven stereology: Integrating generative adversarial networks with tessellation-based models for 3D microstructure reconstruction <i>Orkun Furat, Lukas Fuchs, and Volker Schmidt</i>	168
Geometric properties of random Laguerre tessellations <i>Anna Gusakova</i>	169
The Road to the Closest Point is Paved by Good Neighbors <i>Sariel Har-Peled</i>	169
Stochastic Microstructure Modelling Based on Random Tessellations <i>Christian Jung</i>	170
When grains go wild! Tracking the emergence and persistence of abnormal grain growth using time-resolved 3D x-ray microscopy <i>Carl Krill</i>	170

Semi-discrete convex order and Laguerre tessellation fitting	
<i>Andrea Natale</i>	171
From classical models to Voronoi diagrams with graph-valued seeds – applications to morphological modeling in materials science	
<i>Matthias Neumann</i>	172
Voronoi-like graphs with applications	
<i>Evanthia Papadopoulou</i>	173
Statistics of random marked tessellations	
<i>Zbynek Pawlas</i>	173
A Historical Perspective on the Use of Generalised Voronoi Diagrams in Polycrystal Mechanics	
<i>Romain Quey</i>	174
Statistical Inference for Poisson-Laguerre tessellation: inversion and reconstruction	
<i>Martina Vittorietti</i>	175
Working Groups	
Algorithms and Data Structures	
<i>David P. Bourne</i>	176
Materials Applications	
<i>Andreas Alpers and Claudia Redenbach</i>	176
Software Demos	
Neper (Romain Quey)	177
SynthetMic (David Bourne)	177
Geogram (Bruno Lévy)	177
Participants	178
Remote Participants	178

3 Overview of Tutorials

3.1 Introduction to artificial neural networks and their applications in structural analysis and modeling of polycrystalline materials

Orkun Furat (*University of Southern Denmark – Odense, DK*), Lukas Fuchs (*Universität Ulm, DE*), and Volker Schmidt (*Universität Ulm, DE*)

License  Creative Commons BY 4.0 International license
© Orkun Furat, Lukas Fuchs, and Volker Schmidt

A broad range of engineered as well as natural materials exhibit polycrystalline microstructures, i.e., they consist of a large number of crystals (also called grains) with different atomic lattice orientations arranged in a space-filling manner. The macroscopic (functional) properties of these materials are often strongly influenced by their internal grain architecture, typically observed in experimentally acquired 2D/3D microscopic image data.

Artificial neural networks (ANNs) have become powerful tools for extracting information from this type of image data: convolutional neural networks can segment grains in 2D/3D image data, which in turn facilitates subsequent analysis and modeling workflows [1]. Beyond segmentation, ANNs can be deployed as generative models. For example, generative adversarial networks (GANs) can synthesize image data that exhibits grain architectures that are similar to experimentally observed ones [2]. Such artificially generated microstructures can serve as input for further numerical simulations, enabling virtual materials testing and thereby reducing experimental effort. Moreover, generative approaches can help overcome the limitations of 2D imaging techniques by learning to generate 3D images of grain architectures that are consistent with 2D observations.

Besides representing grain architectures as discretized image data, grain architectures can also be described in a space-continuous manner using tessellation models, in particular, power diagrams. Such representations are advantageous, as they are typically defined by relatively few parameters. Methods from ANNs can contribute to this representational framework by enabling the efficient fitting of power diagrams to segmented image data [3]. In particular, by leveraging programming libraries that enable automatic differentiation, gradient-based optimization can be used to identify parameters of power diagrams that best approximate the grain architectures observed in image data. Beyond deterministic fitting, differentiable implementations of power diagrams enable ANN-guided generative modeling [4]. ANNs can be deployed to learn the probability distribution of power diagram parameters instead of directly learning the distribution of high-dimensional image data, which would be the common approach in conventional GAN-based modeling. The combination of power diagrams with ANNs, induces geometric regularization, as generated samples are constrained to physically reasonable morphologies by the chosen type of power diagram. In particular, this approach facilitates learning the probability distribution of 3D grain architectures from just a few 2D images, thereby reducing the need for costly and time-intensive 3D experimental characterization.

References

- 1 O. Furat, M.Y. Wang, M. Neumann, L. Petrich, M. Weber, C.E. Krill III, and V. Schmidt. *Machine learning techniques for the segmentation of tomographic image data of functional materials*. *Frontiers in Materials* 6, 145, 2019.
- 2 S. Kench and S.J. Cooper. *Generating three-dimensional structures from a two-dimensional slice with generative adversarial network-based dimensionality expansion*. *Nature Machine Intelligence* 3, 299–305, 2021.

- 3 A. Alpers, O. Furat, C. Jung, M. Neumann, C. Redenbach, A. Saken, and V. Schmidt. *Fitting generalized power diagrams to 3D image data: A prerequisite for virtual materials testing*. Preprint (submitted).
- 4 L. Fuchs, O. Furat, D.P. Finegan, J. Allen, F.L.E. Usseglio-Viretta, B. Ozdogru, P.J. Weddle, K. Smith, and V. Schmidt. *Generating multi-scale Li-ion battery cathode particles with radial grain architectures using stereological generative adversarial networks*. Communications Materials 6, 4, 2025.

3.2 Diagrams, Clustering, and Coresets – and the representation of polycrystals

Peter Gritzmann (TU München, DE)

- License** © Creative Commons BY 4.0 International license
© Peter Gritzmann
- Joint work of** Peter Gritzmann, Andreas Alpers, Maximilian Fiedler, Fabian Klemm
- Main reference** Andreas Alpers, Maximilian Fiedler, Peter Gritzmann, Fabian Klemm: “Turning Grain Maps into Diagrams”, SIAM J. Imaging Sci., Vol. 16(1), pp. 223–249, 2023.
URL <https://doi.org/10.1137/22M1491988>
- Main reference** Andreas Alpers, Maximilian Fiedler, Peter Gritzmann, Fabian Klemm: “Dynamic grain models via fast heuristics for diagram representations”, Philosophical Magazine, Vol. 103(10), pp. 948–968, Taylor & Francis, 2023.
URL <https://doi.org/10.1080/14786435.2023.2180679>
- Main reference** Maximilian Fiedler, Peter Gritzmann, Fabian Klemm: “On resolution coresets for constrained clustering”, Journal of Computational Mathematics and Data Science, Vol. 12, p. 100100, 2024.
URL <https://doi.org/10.1016/j.jcmds.2024.100100>
- Main reference** Maximilian Fiedler, Peter Gritzmann: “Coresets for Weight-Constrained Anisotropic Assignment and Clustering”, CoRR, Vol. abs/2203.10864, 2022.
URL <https://doi.org/10.48550/ARXIV.2203.10864>

We give a brief account on diagrams with a view towards representing and analyzing polycrystalline materials. In particular, weight-constrained anisotropic clustering allows to compute diagram representations of polycrystals from data on the volume, center and moments of the grains which are available through tomographic measurements. Also we show the effect of coreset techniques for significantly accelerating the computations.

3.3 Optimal Transport: a new asset in the numerical analysis toolbox

Bruno Levy (INRIA – Villers-lès-Nancy, FR)

- License** © Creative Commons BY 4.0 International license
© Bruno Levy

The Fast Fourier Transform (FFT) had a considerable impact in numerical analysis, by making it very easy to solve the Poisson equation, omnipresent in Physics. According to Yann Brenier’s intuition, the Monge-Ampère equation that characterizes Optimal Transport solutions can be considered as a non-linear version of the Poisson equation. The Monge-Ampère equation plays a role each time a least action principle is constrained by a conservation law, a very general configuration in Physics. This motivates the quest for an efficient Monge-Ampère solver, that could play a role similar to the FFT for this type of problems, as well as for some problems in modern Machine Learning. After a general introduction to the topic, I shall focus on a particular configurations (semi-discrete L2 transport), that involves some generalized Voronoi diagrams (Laguerre diagrams), and for which some specific algorithms (Kitagawa-Merigot-Thibert) can solve problems of very large scale (billions points).

3.4 Random tessellations

Claudia Redenbach (RPTU Kaiserslautern-Landau, DE)

License © Creative Commons BY 4.0 International license
© Claudia Redenbach

Main reference Claudia Redenbach, Christian Jung: “Random Tessellations: An Overview of Models”, pp. 35–80, Springer Nature Switzerland, 2025.

URL https://doi.org/10.1007/978-3-031-87264-8_2

Random tessellations are a prominent class of models in stochastic geometry. In our tutorial, the notion of a random tessellation and basic geometric characteristics of random tessellations are introduced. Several prominent model classes are presented: Voronoi tessellations and their weighted generalizations, hyperplane tessellations, and STIT tessellations. Some statistical methods and applications are also discussed.

4 Overview of Talks

4.1 PyAPD: A Python library for working with minimisation diagrams with GPU acceleration

Maciej Buze (Lancaster University, GB)

License © Creative Commons BY 4.0 International license
© Maciej Buze

Joint work of Maciej Buze, Jean Feydy, Steven M. Roper, Karo Sedighiani, David P. Bourne

Minimisation diagrams, such as Voronoi tessellations, play a pivotal role across science and engineering, with applications ranging from modelling biological and chemical structures to mesh generation for finite element methods [1] and nearest-neighbour searches [2]. They have strong theoretical ties to optimal transport [3] and can be fitted to data using logistic regression, which underpins a lot of modern machine learning methodology [4].

In my talk I will present PyAPD [5], a Python library for working with minimisation diagrams, which leverages a specialised GPU-oriented library for symbolic kernel operations (KeOps [6]) to achieve near instantaneous computation of a generic large minimisation diagram (a 3-4 orders of magnitude speed-up versus baseline GPU (PyTorch) and CPU (NumPy) implementations).

In particular, I will showcase the capabilities of the library in the context of microstructure modelling [7], discuss on-going extensions (unbalanced optimal transport and higher-order minimisation diagrams) and present ideas for collaboration enabled by the ideas underpinning PyAPD.

References

- 1 F. Aurenhammer. *Voronoi diagrams — a survey of a fundamental geometric data structure*. ACM Computing Surveys, 23(3), 345-405 (1991).
- 2 S. Har-Peled and N. Kumar. *Approximating minimization diagrams and generalized proximity search*. SIAM Journal on Computing, 44(4), 944-974 (2015).
- 3 Q. Merigot and B. Thibert. *Optimal transport: discretization and algorithms*. Handbook of Numerical Analysis, Vol. 22, pp. 133-212 (2021), Elsevier.
- 4 K.P. Murphy. *Machine learning: A probabilistic perspective*. MIT press (2012).
- 5 PyAPD: A Python library for computing anisotropic power diagrams using GPU acceleration, URL: <https://github.com/mbuze/PyAPD>

- 6 B. Charlier, J. Feydy, J.A. Glaunès, F.-D. Collin, G. Durif. *Kernel Operations on the GPU, with Autodiff, without Memory Overflows*. Journal of Machine Learning Research 22, 1–6 (2021).
- 7 M. Buze, J. Feydy, S.M. Roper, K. Sedighiani and D.P. Bourne. *Anisotropic power diagrams for polycrystal modelling: Efficient generation of curved grains via optimal transport*. Computational Materials Science, Volume 245 (2024).

4.2 A few asymptotic results for functionals of Poisson-Voronoi cells

Pierre Calka (*University of Rouen, FR*)

License © Creative Commons BY 4.0 International license
© Pierre Calka

Joint work of Pierre Calka, Aurélie Chapron, Nicolas Chenavier, Cecilia D’Errico, Yann Demichel, Nathanaël Enriquez

In this talk, we consider the classical model of a random Voronoi tessellation generated by the points of a Poisson point process and derive in various settings distributional properties or expectation estimates for several functionals of the cells.

When the underlying Poisson point process is homogeneous in the Euclidean space, we can define the notion of typical cell in the Palm meaning, which is equal in distribution to the Voronoi cell generated by a fixed deterministic nucleus added to the point process, see e.g. [7] or [8, Section 10.2].

It is easy to show that the law of the volume of the largest ball centered at the nucleus and included in the typical cell is exponential. The distribution of the maximal distance from a point in the cell to the nucleus is another matter. In a recent work [3], we obtain precise asymptotics for the distribution tail of that distance and deduce from it and from a previous work [2] the fact that the extremal index related to the sequence of such distances for all Voronoi cells included in a large box is equal to $(2d)^{-1}$.

In the setting of the two-dimensional Poisson-Voronoi tessellation conditioned on the event that the typical cell contains a fixed large shape, we derive asymptotics for the expectation of the defect volume, perimeter and number of vertices of the cell [5]. Moreover, in view of studying the neighboring cells with an elongated shape, we study a toy model where the Voronoi nuclei are located in the lower half-space and we obtain a limit shape result for the renormalized typical Poisson-Voronoi cell with fixed height [6].

In the setting of a Poisson-Voronoi tessellation along a Riemannian manifold, we study the Voronoi cell associated with a deterministic nucleus and show that when the intensity goes to infinity, the expectation of the number of vertices of the cell has a two-term expansion which involves the scalar curvature of the manifold at the nucleus point [1].

Finally, we present a model of iterated Voronoi tessellations introduced by S. Zuyev and K. Tchoumatchenko [9]. Inspired by their construction, we propose a planar toy model for a fractal random set which is based on a sequence of regular tessellations. We then provide explicit values for its box and Hausdorff dimensions [4].

In the process, we discuss a few open questions related to these topics.

References

- 1 Pierre Calka, Aurélie Chapron and Nathanaël Enriquez. *Poisson-Voronoi tessellation on a Riemannian manifold*. Int. Math. Res. Not. 2021, 5413–5459, 2021.
- 2 Pierre Calka and Nicolas Chenavier. *Extreme values for characteristic radii of a Poisson-Voronoi tessellation*. Extremes 17, 359–385, 2014.

- 3 Pierre Calka, Cecilia D’Errico and Nathanaël Enriquez. *Sharp asymptotics for the maximal distance from the boundary to the nucleus of a typical Poisson-Voronoi cell*. Preprint available at arXiv:2511.11189, 2025.
- 4 Pierre Calka and Yann Demichel. *Fractal random sets associated with multitype Galton-Watson trees*. To appear in Probab. Theory Related Fields, 2025.
- 5 Pierre Calka, Yann Demichel and Nathanaël Enriquez. *Large planar Poisson-Voronoi cells containing a given convex body*. Ann. H. Lebesgue 4, 711–757, 2021.
- 6 Pierre Calka, Yann Demichel and Nathanaël Enriquez. *Elongated Poisson-Voronoi cells in an empty half-plane*. Adv. Math. 410, Part A, 108721, 2022.
- 7 Jesper Møller. *Lectures on random Voronoi tessellations*. Springer, 1994.
- 8 Rolf Schneider and Wolfgang Weil. *Stochastic and integral geometry*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2008.
- 9 Konstantin Tchoumatchenko and Sergei Zuyev. *Aggregate and Fractal Tessellations*. Prob. Theory and Related Fields 121(2), 198–218, 2001.

4.3 Overlay of Voronoi tessellations for lunar Euclidean MSTs

Herbert Edelsbrunner (IST Austria – Klosterneuburg, AT)

License © Creative Commons BY 4.0 International license
© Herbert Edelsbrunner

Related or higher-order Voronoi tessellations, we consider the overlay of several order-1 Voronoi tessellations. These tessellations are motivated by the study of chromatic persistent homology and, in particular, the equivalence of the relative chromatic diagram and a more conventional diagram that analyzes the persistent growth pattern of lunes (which we define as intersections of equal-size balls).

4.4 CVT-based data-adaptive family of partitioning algorithms and materials modeling

Maria Emelianenko (George Mason University – Fairfax, US)

License © Creative Commons BY 4.0 International license
© Maria Emelianenko
Joint work of Maria Emelianenko, Guy B. Oldaker, Katherine Saleme Ruiz, Dmitry Golovaty
Main reference Maria Emelianenko, Guy B. Oldaker IV: “On the optimality of voronoi-based column selection”, Numerical Methods for Partial Differential Equations, Vol. 40(6), p. e23137, 2024.
URL <https://doi.org/10.1002/num.23137>

In the first part of the talk, we present a unified family of data-adaptive partitioning algorithms that extends several well-known methods (e.g., k-means and k-subspaces) [1, 2]. Indexed by a single parameter and employing a common centroidal Voronoi tessellation-based minimization strategy, the algorithms are easy to use and interpret, and scale well to large, high-dimensional problems. The data-adaptive framework developed in this work: (a) exhibits skill at automatically uncovering data structures and problem parameters without any expert knowledge, (b) can be used to augment other existing methods. Analytical results and numerical experiments presented in this talk highlight the advantages of the new methodology in the context of several disparate application areas, including subspace clustering, model order reduction, and matrix approximation.

The second part highlights recent work on microstructure modeling in polycrystals, using large-scale curvature-driven grain growth simulations and mesoscopic formulations [3]. It is shown that microstructure entropy plays an important role both in the formation of grain boundary statistics as well as in the form of the elastic and plastic response. Stochastic modeling, including random walk descriptions, and applications to additive manufacturing microstructure characterization are presented.

References

- 1 M. Emelianenko and G. Oldaker. *On the optimality of Voronoi-based column selection*. Numer. Methods Partial Differ. Eq., 2024.
- 2 G.B. Oldaker IV and M. Emelianenko. *A Unifying Family of Data-Adaptive Partitioning Algorithms*. arXiv preprint arXiv:2412.16713, 2024.
- 3 A. H. J. Sazo, C. E. Torres, M. Emelianenko, and D. Golovaty, Mater. Res. Express 6 055506, 2019.

4.5 Provably Robust Perspective- n -Point Algorithms for Camera Pose Estimation

Dan Feldman (*University of Haifa, IL*)

License  Creative Commons BY 4.0 International license
© Dan Feldman

In the Perspective- n -Point problem (PnP), we are given a 2D image of a known 3D object that was captured by a pinhole camera, and the goal is to estimate the position of the camera. More formally, the input is a set of n points $p_1, \dots, p_n$ in $\mathbb{R}^3$, and corresponding n lines $\ell_1, \dots, \ell_n$, where each line intersects the origin of $\mathbb{R}^3$ (the pinhole). We wish to compute an alignment, i.e., a rotation matrix $R \in \text{SO}(3)$ and a translation vector $t \in \mathbb{R}^3$, that minimizes some function $f(\text{dist}(Rp_1 + t, \ell_1), \dots, \text{dist}(Rp_n + t, \ell_n))$ of the distances between each aligned point to its paired line. The Least-Mean-Square function $f(\cdot) = \|\cdot\|_2^2$ is easier to compute, but sensitive to noise and outliers.

We suggest the first provably good approximation algorithms for the global optimum of robust versions of the PnP. These include: sum of (non-squared) distances $f = \|\cdot\|_1$, M -estimators and m -outliers (where large distances have a bounded effect or are ignored, respectively). The expected running time is $O(n \log n)$ for a sufficiently large n , and the multiplicative approximation factor is $O(1)$ in the worst-case.

This is by providing a generic algorithm that minimizes any given log-Lipschitz function f of the distances. It forges novel links between objects in geometric topology, such as Clifford torus, optimization techniques from computational geometry, such as core-sets, and commutative algebra that is used for solving the polynomial system. While PnP was selected for demonstration, we expect applications to many other open problems, especially in computer vision.

4.6 An optimal transport model for dynamical shapes, collective motion and cellular aggregates

Jean Feydy (INRIA – Paris, FR)

License  Creative Commons BY 4.0 International license
© Jean Feydy

Many living and physical systems such as cell aggregates, tissues or bacterial colonies behave as unconventional systems of particles that are strongly constrained by volume exclusion and shape interactions. Understanding how these constraints lead to macroscopic self-organized structures is a fundamental question in e.g. developmental biology. To this end, various types of computational models have been developed. Here, we introduce a new framework based on optimal transport theory to model particle systems with arbitrary dynamical shapes and deformability properties. Our method builds upon the pioneering work of Brenier on incompressible fluids and its recent applications to materials science. It lets us specify the shapes and volumes of individual cells and supports a wide range of interaction mechanisms, while automatically taking care of the volume exclusion constraint at an affordable numerical cost. We showcase the versatility of this approach by reproducing several classical systems in computational biology. Our Python code is freely available at <https://iceshot.readthedocs.io/en/latest/>

4.7 Data-driven stereology: Integrating generative adversarial networks with tessellation-based models for 3D microstructure reconstruction

Orkun Furat (University of Southern Denmark – Odense, DK), Lukas Fuchs (Universität Ulm, DE), and Volker Schmidt (Universität Ulm, DE)

License  Creative Commons BY 4.0 International license
© Orkun Furat, Lukas Fuchs, and Volker Schmidt

Many natural materials exhibit a cellular microstructure, where the material is composed of individual cells or domains with distinct properties, despite sharing the same composition. Such structures are observed, for example, in metallic materials, biological systems, synthetic materials like foams and rocks. The spatial arrangement, size distribution, interfacial characteristics and morphological features of these cells critically determine the macroscopic effective properties of the bulk material. Consequently, the investigation of cellular structures is of great importance. Tessellations provide a natural and mathematically rigorous framework for partitioning space into cells, making them particularly well suited for modeling cellular microstructures. Their parameters have clear geometric interpretations, such as cell size, shape and adjacency, which directly correspond to observable microstructural characteristics.

By combining point stochastic processes with distance-based tessellation models such as Poisson–Voronoi and Laguerre tessellations, it becomes possible to generate random tessellations whose realizations are statistically similar to those observed in experimental image data [1]. This allows the efficient generation of large datasets of model realizations, facilitating the systematic investigation of structure–property relationships. However, 3D datasets, that are usually required for calibrating such random 3D tessellation models are often unavailable due to the high cost or complexity of their acquisition.

This talk addresses this issue, by introducing a data-driven framework that leverages parametric tessellation models to statistically calibrate 3D random tessellation models using only 2D image data while preserving morphological interpretability[2, 3]. Specifically, we

employ generative adversarial networks (GANs) to learn the tessellation parameters from 2D images, enabling efficient inference of realistic 3D cellular morphologies without the need for expensive 3D measurements. This approach bridges the gap between physically interpretable stochastic geometry and the generative capabilities of deep learning. It offers a data-efficient pathway for reconstructing and simulating realistic 3D microstructures, supporting applications in materials design, stereological analysis and computational materials science.

References

- 1 O. Furat, L. Petrich, D. Finegan, D. Diercks, F. Usseglio-Viretta, K. Smith and V. Schmidt. *Artificial generation of representative single Li-ion electrode particle architectures from microscopy data*. npj Computational Materials 7, 105, 2021.
- 2 L. Fuchs, O. Furat, D.P. Finegan, J. Allen, F.L.E. Usseglio-Viretta, B. Ozdogru, P.J. Weddle, K. Smith and V. Schmidt. *Generating multi-scale Li-ion battery cathode particles with radial grain architectures using stereological generative adversarial networks*. Communications Materials 6, 4, 2025.
- 3 L. Fuchs, T. Wilhelm, O. Furat and V. Schmidt. *Adversarial learning and Voronoi tessellations: A stereological framework for 3D cellular microstructure generation*. Preprint (submitted).

4.8 Geometric properties of random Laguerre tessellations

Anna Gusakova (Universität Münster, DE)

License © Creative Commons BY 4.0 International license
© Anna Gusakova

Joint work of Anna Gusakova, Mathias in Wolde-Lübke

In this talk we consider stationary random Laguerre tessellations whose generating point set is given by some inhomogeneous Poisson point process in space-time $\mathbb{R}^d \times \mathbb{R}$. We will discuss sectional properties of Poisson-Laguerre tessellations revealing their relation to the classical Poisson-Voronoi tessellation. Furthermore, we will study the class of the corresponding dual Laguerre tessellations which, under some natural assumptions, give rise to simplicial random tessellations. For these models we will be interested in the geometric properties of their typical cell.

4.9 The Road to the Closest Point is Paved by Good Neighbors

Sariel Har-Peled (University of Illinois – Urbana-Champaign, US)

License © Creative Commons BY 4.0 International license
© Sariel Har-Peled

Joint work of Sariel Har-Peled, Benjamin Raichel, Eliot W. Robson

Main reference Sariel Har-Peled, Benjamin Raichel, Eliot W. Robson: “The Road to the Closest Point is Paved by Good Neighbors”, in Proc. of the 2026 Symposium on Simplicity in Algorithms, SOSA 2026, Vancouver, BC, Canada, January 12-14, 2026, pp. 1–11, SIAM, 2026.

URL <https://doi.org/10.1137/1.9781611978964.1>

Given a set P of n points in $\mathbb{R}^d$, and a parameter $\varepsilon \in (0, 1)$, we present a new construction of a directed graph G , of size $O(\frac{n}{\varepsilon^d})$, such that $(1 + \varepsilon)$ -ANN queries can be answered by performing a greedy walk on G , repeatedly moving to a neighbor that is (significantly) better than the current point. To the best of our knowledge, this is the first construction of a linear size with no dependency on the spread of the point set. The resulting query time, is $O(\varepsilon^{-d} \log \Psi)$, where Ψ is the spread of P . The new construction is surprisingly simple and should be practical.

4.10 Stochastic Microstructure Modelling Based on Random Tessellations

Christian Jung (RPTU Kaiserslautern-Landau, DE)

License © Creative Commons BY 4.0 International license

© Christian Jung

Joint work of Christian Jung, Katja Schladitz, Claudia Redenbach

Main reference Christian Jung, Claudia Redenbach: “Crack Modeling via Minimum-Weight Surfaces in 3d Voronoi Diagrams”, in *Journal of Mathematics in Industry*, 13(10), 2023.

URL <https://doi.org/10.1186/s13362-023-00138-1>

Modelling the microstructure of industrial materials allows for the generation of large synthetic data sets that can be used for numerical simulations and the training of machine learning models. In this talk we consider stochastic microstructure models based on random tessellations and several applications. First, we establish a model for crack structures in concrete. Cracks are assumed to take the path of least resistance when propagating through concrete. The model is based on the idea of weighting the facets of a 3d Voronoi tessellation. Bounded by a contour, a connected set of facets of minimal weight is computed by solving a binary integer program. The point process models that generate the Voronoi diagrams are interchangeable such that the model realizations show several levels of regularity. This makes the model highly flexible and allows for the generation of a whole range of different crack structures. Second, germ-grain processes based on the cells of Gibbs-Laguerre tessellations are used for modelling solid particles in active protective coatings. Our goal is to compute realizations of Gibbs-Laguerre tessellations whose distributions of cell shape and size are similar to those of the observed particles. Several techniques are proposed to also account for the correlation between these characteristics.

References

- 1 C. Jung, C. Redenbach, and K. Schladitz. *VoroCrack3d: An annotated data set of 3d CT concrete images with synthetic crack structures*. Data in Brief, 54:110474, 2024.
- 2 C. Jung and C. Redenbach. *An Analytical Representation of the 2d Generalized Balanced Power Diagram*. Computational Geometry, 121:102101, 2024.
- 3 K. Schladitz, C. Jung, S. Flenner, M. Godehardt, B. Grevelhörster, I. Greving, P. Klein, N. Konchakova, C. Redenbach, P. Visser, and J. Zaninović. *Geometric modelling of corrosion inhibitor pigments in active protective coatings based on SR-nano-CT images*. Progress in Organic Coatings, 197:108762, 2024.

4.11 When grains go wild! Tracking the emergence and persistence of abnormal grain growth using time-resolved 3D x-ray microscopy

Carl Krill (Universität Ulm, DE)

License © Creative Commons BY 4.0 International license

© Carl Krill

Joint work of Carl Krill, André Schulz-Harder, Jule M. Dake, Karolína Gutbrod, Madlen Atzen, Markus Ziehmer, Thomas Wilhelm, Orkun Furat, Volker Schmidt, Haixing Fang, Pierre-Olivier Autran, Wolfgang Ludwig

Usually, the coarsening of a polycrystalline material is a rather civilized affair: adjacent grains exchange atoms at moderate rates, only a small fraction of grains disappears over a given interval, and the grain-size distribution remains unimodal. In some systems, however, certain participants in this competition develop a seemingly insatiable appetite for growth! The

resulting microstructure features a subpopulation of oversized, *abnormal* grains embedded in a matrix of much smaller ones, producing the characteristic bimodal grain-size distribution associated with abnormal grain growth (AGG).

Despite the prevalence of AGG, the conditions that trigger the first abnormal grains and the mechanisms that sustain their growth remain poorly understood. To address these questions, we have used time-resolved diffraction-contrast tomography (DCT) to capture the evolution of AGG in three dimensions in the commercial aluminum alloy 5252. These measurements, supplemented by phase-contrast tomography to map second-phase pinning particles, reveal persistent migration of abnormal/matrix interfaces through a dense particle field. However, the *rate* of such migration is observed to be strongly correlated with the number density of particles encountered by each boundary.

Together, the time-resolved observations suggest that a single underlying mechanism may be responsible for both the emergence of abnormal grains and the persistent growth of these grains at the expense of the surrounding matrix.

4.12 Semi-discrete convex order and Laguerre tessellation fitting

Andrea Natale (*Université Paris-Saclay, Inria, FR*)

License © Creative Commons BY 4.0 International license

© Andrea Natale

Joint work of David P. Bourne, Thomas O. Gallouët, Quentin Merigot, Andrea Natale

Laguerre tessellations offer a computationally tractable way to describe a large number of convex partitions of Euclidean space. For this reason, they have become popular in computational geometry, imaging and numerical analysis, both as a modeling and a discretization tool.

In material science, Laguerre tessellations are used to describe the microstructure of polycrystals. In this context, one generally needs to fit various type of data describing polycrystalline structures by means of tessellations. Motivated by these applications, we study the problem of recovering the generators and weights associated with a given Laguerre tessellation, provided the volumes and barycenters of its cells, which is an idealized version of a common scenario encountered in practice [3].

To solve this problem, we make explicit the connection between barycenters obtained from Laguerre tessellations with prescribed cell volumes and semi-discrete Optimal Transport [4], developing a point of view proposed in [2]. Specifically, we find that any vector of barycenters arising from a Laguerre tessellation of cells with prescribed volumes can be understood as an exposed point of a convex set C defined by convex order relations [1]. Moreover, maximizing a linear function over this set is equivalent to solving a semi-discrete Optimal Transport problem. We leverage this characterization to propose a numerical algorithm that approximately solves the reconstruction problem by computing an orthogonal projection onto C . This can be computed efficiently using classical convex optimization algorithms. Importantly, the proposed approach can be used more generally on experimental data which does not necessarily arise from a Laguerre tessellation, which is the typical scenario in the material science applications motivating this work.

References

- 1 D. P. Bourne, T. O. Gallouët, Q. Merigot, and A. Natale. *Semi-discrete convex order and Laguerre tessellation fitting*. HAL preprint: hal-05008630, 2025.

- 2 D. P. Bourne, M. Pearce, and S. M. Roper. *Inverting Laguerre tessellations: Recovering tessellations from the volumes and centroids of their cells using optimal transport*. ESAIM: Mathematical Modelling and Numerical Analysis, 59(5):841–871, 2025.
- 3 L. Petrich, J. Staněk, M. Wang, D. Westhoff, L. Heller, P. Šittner, C. E. Krill III, V. Beneš, and V. Schmidt. *Reconstruction of grains in polycrystalline materials from incomplete data using Laguerre tessellations*. Microscopy and Microanalysis, 25(3):743–752, 2019.
- 4 Q. Mérigot and B. Thibert. *Optimal transport: discretization and algorithms*. Handbook of Numerical Analysis, Vol. 22, pp. 133–212, 2021, Elsevier.

4.13 From classical models to Voronoi diagrams with graph-valued seeds – applications to morphological modeling in materials science

Matthias Neumann (TU Graz, AT)

License  Creative Commons BY 4.0 International license
© Matthias Neumann

Stochastic 3D modeling of the morphology of micro- and nanostructures allows for generating large databases that can be used to quantify relationships between morphology and effective physical properties of functional materials, such as polycrystalline materials, fuel cells or porous silica. The latter is crucial for resource-efficient optimal design of these materials. In this talk, we present three different parametric morphological models based on random tessellations. First, we show a parametric model based on marked tessellations developed for polycrystalline materials, which exhibit so-called twinning effects. This requires a particular dependency structure for the $SO(3)$ -valued marks which model the crystal orientations of the grains [1]. Second, we discuss a model for three-phase structures in solid oxide fuel cell anodes that guarantees that all three phases are almost surely connected. For this purpose, a special kind of Voronoi diagram is used that consists only of three cells, and the seeds are three independent random graphs. The vertices of the latter are modeled by homogeneous Poisson point processes [2, 3]. A generalization of this model leads to the third application, where the morphology of porous silica is modeled. In this case, the vertices of the random graphs are modeled by Matérn cluster processes. The model is then fitted to three differently manufactured samples of porous silica [4]. Finally, an outlook on future research and open problems regarding the considered models are presented.

References

- 1 P. Rieder, M. Neumann, L. Monteiro Fernandes, A. Mulard, H. Proudhon, F. Willot, and V. Schmidt. *Stochastic 3D microstructure modeling of twinned polycrystals for investigating the mechanical behavior of γ -TiAl intermetallics*. Computational Materials Science, 238, 112922, 2024.
- 2 M. Neumann, J. Staněk, O. Pecho, L. Holzer, V. Beneš, and V. Schmidt. *Stochastic 3D modeling of complex three-phase microstructures in SOFC-electrodes with completely connected phases*. Computational Materials Science, 118, 353–364, 2016.
- 3 M. Neumann, B. Abdallah, L. Holzer, F. Willot, and V. Schmidt. *Stochastic 3D modeling of three-phase microstructures for predicting transport properties: a case study*. Transport in Porous Media, 128, 179–200, 2019.
- 4 B. Prifling, M. Neumann, D. Hlushkou, C. Kübel, U. Tallarek, and V. Schmidt. *Generating digital twins of mesoporous silica by graph-based stochastic microstructure modeling*. Computational Materials Science, 187, 109934, 2021.

4.14 Voronoi-like graphs with applications

Evanthia Papadopoulou (USI – Lugano, CH)

License © Creative Commons BY 4.0 International license
© Evanthia Papadopoulou

Main reference Evanthia Papadopoulou: “Abstract Voronoi-Like Graphs: Extending Delaunay’s Theorem and Applications”, in Proc. of the 39th International Symposium on Computational Geometry, SoCG 2023, Dallas, Texas, USA, June 12–15, 2023, LIPIcs, pp. 52:1–52:16, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.

URL <https://doi.org/10.4230/LIPICS.SOCG.2023.52>

Abstract Voronoi diagrams (AVDs) offer a unifying framework for various concrete Voronoi diagrams involving points and generalised sites, such as segments, disks, or polygons, in various metrics. Rather than the geometric notions of sites and circles, AVDs are based on bisecting curves that satisfy some simple combinatorial properties. In this talk I will give an overview of AVDs and define a more relaxed Voronoi-like variant. Voronoi-like diagrams are graphs in the arrangement of bisecting curves that fall under the abstract framework, whose vertices are “locally Voronoi” corresponding to vertices in Voronoi diagrams of three sites. Voronoi-like graphs turn out to be Voronoi diagrams subject to perhaps missing some faces, if Voronoi regions can be disconnected. This result extends Delaunay’s theorem beyond points in the Euclidean plane, to Voronoi diagrams that fall under the AVD framework. We have used Voronoi-like graphs to derive expected linear-time algorithms to compute Voronoi trees and forests. We also expect Voronoi-like graphs to be relevant to Voronoi recognition problems.

4.15 Statistics of random marked tessellations

Zbynek Pawlas (Charles University – Prague, CZ)

License © Creative Commons BY 4.0 International license
© Zbynek Pawlas

Joint work of Zbynek Pawlas, Viktor Beneš, Daniela Flimmel, Luděk Heller, Iva Karafiátová, Jesper Møller, Filip Seitzl, Jakub Staněk, Joseph E. Yukich

We present an overview of several statistical problems arising in the analysis of random marked tessellations. First, we consider non-parametric estimation of the grain-mark distribution. Using the minus-sampling technique, we construct an unbiased estimator of Horvitz–Thompson type. For a stationary weighted Voronoi tessellation generated by a Poisson point process, we establish variance asymptotics, weak consistency, and asymptotic normality under mild conditions on the weights of the cells in [1].

The second problem is motivated by applications in materials science and concerns the fitting of parametric models for marks taking values in $SO(3)$. Conditional on the Laguerre tessellation, we propose in [2] interaction models for the distribution of cubic crystal lattice orientations, where interactions occur between pairs of orientations associated with neighbouring cells of the tessellation.

Finally, we describe a statistical test introduced in [3] for deciding whether orientations are assigned independently to the cells of the underlying tessellation. The distribution of a given orientation may either be identical for all cells or depend on the corresponding cell. For both settings, non-parametric tests are developed.

References

- 1 D. Flimmel, Z. Pawlas, and J. E. Yukich. *Limit theory for unbiased and consistent estimators of statistics of random tessellations*. Journal of Applied Probability, 57, 679–702, 2020.
- 2 I. Karafiátová, J. Møller, Z. Pawlas, J. Staněk, F. Seidl, and V. Beneš. *Fitting the grain orientation distribution of a polycrystalline material conditioned on a Laguerre tessellation*, Spatial Statistics, 55, 100747, 2023.
- 3 Z. Pawlas, I. Karafiátová, and L. Heller. *Random tessellations marked with crystallographic orientations*. Spatial Statistics, 39, 100469, 2020.

4.16 A Historical Perspective on the Use of Generalised Voronoi Diagrams in Polycrystal Mechanics

Romain Quey (CNRS / Mines Saint-Etienne, FR)

License  Creative Commons BY 4.0 International license
© Romain Quey

In materials science, particularly in polycrystal mechanics, simulation is essential for studying the mechanical response and microstructure evolution of materials under external loads. Because polycrystalline materials have random microstructures, generalized Voronoi tessellations are the preferred approach for representing them. However, the appropriate model must be used for each application, which is why models of different levels of refinement have been used in the past. In the 1990s, simulations focused on replicating macroscopic properties, such as strain-strain response and crystallographic texture – i.e., the statistical distribution of the crystal orientations of the grains. In this context, and to facilitate mesh generation, researchers widely used regularly shaped grains, such as cubes, dodecahedra, and truncated octahedra. As experimental methods became available to probe local deformation or microstructural evolution during straining, simulations matured as well. Voronoi tessellations emerged as dominant polycrystal models. However, Poisson Voronoi tessellations contain a large number of small edges, posing a problem for simulations based on the finite element method. A “regularization” technique was developed to solve this problem and enable robust simulations up to 30–40% strain [1]. The next experimental breakthrough involved three-dimensional (3D) measurements of microstructural and mechanical states during deformation using high-energy X-ray diffraction. Around the same time, Laguerre tessellations were first used as polycrystal models. In an optimization context, Laguerre tessellations can be generated from any form of input, whether statistical, incomplete (e.g., grain centroids and volumes), or complete (e.g., polycrystal map) [2]. Despite the resulting convexity of the grains, which is not necessarily present in a real polycrystal, it was shown that the main properties of an experimental deformation field can be accurately reproduced [3]. The latest X-ray diffraction methods allow the elastic strain field in a deformed polycrystal to be mapped on a per-voxel basis. This again requires new developments in polycrystal modeling. Anisotropic power diagrams can reproduce the non-convexity of grains with remarkable agreement with real microstructures. However, meshing such microstructures remains an open problem. Similarly, microstructures obtained using new processing techniques, such as additive manufacturing, differ significantly from those obtained using traditional techniques, which presents an additional challenge for microstructure modeling.

Neper is an open-source software package designed for polycrystal modeling based on generalized Voronoi tessellations, especially for comparative studies between experiments and finite element simulations [4]. When used with its companion program, FEPX, Neper

provides a complete workflow, from tessellation generation to managing, processing, and visualizing simulation results [5]. As Neper is a large open-source program, it is a potential platform for efficiently implementing new methods.

References

- 1 R. Quey, P.R. Dawson, and F. Barbe. *Large-scale 3D random polycrystals for the finite element method: Generation, meshing and remeshing*. Computer Methods in Applied Mechanics and Engineering. 200, 1729-1745, 2011.
- 2 R. Quey and L. Renversade. *Optimal polyhedral description of 3D polycrystals: method and application to statistical and synchrotron X-ray diffraction data*. Computer Methods in Applied Mechanics and Engineering, 330, 308-333, 2018.
- 3 M. Gille, H. Proudhon, J. Oddershede, R. Quey, and T.F. Morgeneyer. *3D strain heterogeneity and fracture studied by X-ray tomography and crystal plasticity in an aluminium alloy*. International Journal of Plasticity, 183, 104146, 2024.
- 4 Neper, <https://neper.info>.
- 5 R. Quey and M. Kasemer. *The Neper/FEPIX project: free / open-source polycrystal generation, deformation simulation, and post-processing*. IOP Conference Series Materials Science and Engineering. 1249, 012021, 2022.

4.17 Statistical Inference for Poisson-Laguerre tessellation: inversion and reconstruction

Martina Vittorietti (TU Delft, NL)

License © Creative Commons BY 4.0 International license
© Martina Vittorietti

Main reference Thomas van der Jagt, Geurt Jongbloed, Martina Vittorietti: “Nonparametric inference for Poisson-Laguerre tessellations”, Scandinavian Journal of Statistics, Vol. 52(4), pp. 1816–1851, 2025.
URL <https://doi.org/10.1111/sjos.70011>

In this talk, we consider statistical inference for Poisson-Laguerre tessellations in $\mathbb{R}^d$. The object of interest is a distribution function F which describes the distribution of the arrival times of the generator points. The function F uniquely determines the intensity measure of the underlying Poisson process. We consider two scenarios: first, we assume to know the points of the Poisson process which generate non-empty cells and the actual cells corresponding to these points and construct two nonparametric estimators. The proposed estimators are proven to be strongly consistent, as the observation window expands unboundedly to the whole space. Then, we assume to do not know the points and to want to retrieve them inverting the tessellation. We provide sufficient conditions for consistently inverting the observed Poisson-Laguerre tessellation, as the observation window expands unboundedly to the whole space. The performance of the two estimators, previously proposed is investigated also in this setting.

5 Working Groups

In the working group sessions, the participants split up into smaller groups. The discussions mainly evolved around two topics: algorithmic problems related to generalized Voronoi diagrams and the use of generalized Voronoi diagrams in materials science applications. The composition of the groups differed from session to session.

5.1 Algorithms and Data Structures

David Bourne (Heriot-Watt University – Edinburgh, GB, d.bourne@hw.ac.uk)

License  Creative Commons BY 4.0 International license
© David P. Bourne

Discussed Problems

This group discussed methods for reconstructing Laguerre tessellations from incomplete data, such as the volumes and centroids of the cells. We compared various methods from different communities including stochastic optimisation (the cross-entropy method), stochastic geometry, and optimal transport theory.

Another topic of the discussion was the lack of efficient data structures for storing diagrams with curved cell boundaries such as generalized balanced power diagrams. We also discussed efficient algorithms for computing such diagrams.

5.2 Materials Applications

Andreas Alpers (University of Liverpool, GB, andreas.alpers@liverpool.ac.uk)

Claudia Redenbach (RPTU Kaiserslautern-Landau, DE, claudia.redenbach@rptu.de)

License  Creative Commons BY 4.0 International license
© Andreas Alpers and Claudia Redenbach

Discussed Problems

The groups discussed various aspects related to the modelling of materials microstructures. A particular focus was on representing the grain system in polycrystalline materials.

In many applications, only fitting of the grain size distribution is considered. Grain sizes are often observed empirically to follow approximately log-normal distributions, although this behavior is not universal and its theoretical justification remains largely phenomenological. For realistic grain systems, also grain shape including anisotropy should be taken into account. In particular, large cells of Laguerre diagrams often turn out to be too spherical when compared to real grains.

For finite element simulations, the cells of the diagram have to be meshed. Short cell edges can result in computationally inefficient meshes. Hence, a regularization step that resolves short edges should be applied prior to the mesh generation.

Differences between the microstructure of additively manufactured materials and materials produced by conventional approaches were discussed. While additive manufacturing gives more control on the large scale structure, it results in very chaotic microstructures that pose new challenges for microstructure modelling. This is a particular problem for parts with complex shapes which induce massive edge effects that must be reproduced in the model. Simply masking a homogeneous tessellation model with the desired shape is not a solution to this problem.

The observation and modelling of grain growth in polycrystalline materials was another topic of a lively discussion. In particular, individual grains disappear during grain growth. The resulting topological changes in the diagram must be correctly reproduced. This, in turn, underscores the importance of analytic boundary representations for materials science applications, since they enable the inference of geometric properties such as dihedral angles.

It was further suggested that, rather than first inferring geometric quantities such as volumes and centres of mass from diffraction data and subsequently constructing a generalized Voronoi diagram, it may be possible to reconstruct such diagrams directly from diffraction data using learning-based methods, potentially reducing modelling bias and improving computational efficiency.

6 Software Demos

6.1 Neper (Romain Quey)

Romain Quey gave a demonstration of how to use the software for polycrystal generation and how to visualize and output its results.

“Neper is a free / open source software package for polycrystal generation and meshing. It can be used to generate polycrystals with a wide variety of morphological properties, from very simple morphologies (simple tessellations, grain-growth microstructures, ...) to complex, multiphase or multiscale microstructures that involve grain subdivisions. The resulting tessellations can be meshed into high-quality meshes suitable for finite-element simulations.” (Source: <https://neper.info/>).

6.2 SynthetMic (David Bourne)

David Bourne gave a demonstration of his software library *SynthetMic* (<https://github.com/synthetic-microstructures/synthetmic>), which was developed in collaboration with R. O. Ibraheem (M-DICE, Heriot-Watt University) and S. M. Roper (University of Glasgow). This is a python library for generating 2D and 3D synthetic polycrystalline microstructures using Laguerre tessellations. It includes a fast algorithm for generating grains of prescribed volumes using optimal transport theory. The demo focused on the graphical user interface (web app) for the library, which can be found here: <https://david-bourne.shinyapps.io/synthetmic-gui/>.

6.3 Geogram (Bruno Lévy)

Bruno Lévy gave a demonstration of his geometry-processing library *Geogram* (<https://github.com/BrunoLevy/geogram>) and a tutorial on how to install it. Geogram includes code for meshing, Delaunay triangulations, Laguerre diagrams, and solving semi-discrete optimal transport problems.

Participants

- Andreas Alpers
University of Liverpool, GB
- David P. Bourne
Heriot-Watt University –
Edinburgh, GB
- Maciej Buze
Lancaster University, GB
- Pierre Calka
University of Rouen, FR
- Herbert Edelsbrunner
IST Austria –
Klosterneuburg, AT
- Maria Emelianenko
George Mason University –
Fairfax, US
- Dan Feldman
University of Haifa, IL
- Jean Feydy
INRIA – Paris, FR
- Lukas Fuchs
Universität Ulm, DE
- Orkun Furat
University of Southern Denmark –
Odense, DK
- Peter Gritzmann
TU München, DE
- Anna Gusakova
Universität Münster, DE
- Sarel Har-Peled
University of Illinois –
Urbana-Champaign, US
- Mathias in Wolde-Lübke
Universität Münster, DE
- Christian Jung
RPTU – Kaiserslautern, DE
- Carl Krill
Universität Ulm, DE
- Bruno Levy
INRIA – Villers-lès-Nancy, FR
- Dawid Lipinski
Lancaster University, GB
- Andrea Natale
Université Paris-Saclay, Inria, FR
- Matthias Neumann
TU Graz, AT
- Evanthia Papadopoulou
USI – Lugano, CH
- Zbynek Pawlas
Charles University – Prague, CZ
- Romain Quey
CNRS / Mines Saint-Etienne, FR
- Claudia Redenbach
RPTU – Kaiserslautern, DE
- Katja Schladitz
Fraunhofer ITWM –
Kaiserslautern, DE
- Volker Schmidt
Universität Ulm, DE
- Raimund Seidel
Universität des Saarlandes –
Saarbrücken, DE
- Martina Vittorietti
TU Delft, NL

Remote Participants

- Henning F. Poulsen
Technical University of Denmark
– Lyngby, DK

