Machine Learning: New Perspectives for Science

Tilman Gocht *University of Tübingen, Machine Learning Cluster of Excellence, Germany* tilman.gocht@uni-tuebingen.de

Abstract

The Machine Learning Cluster of Excellence was established in 2019 at the University of Tübingen, Germany. This research cluster aims to advance machine learning to aid scientific understanding in a wide range of disciplines – from medicine and neuroscience to cognitive science, linguistics, and economics, physics and the geosciences – and to better understand and steer the impact of machine learning on scientific practice. In the past years, we have developed the community and workflows to connect machine learning with different scientific disciplines. In the following years, we will continue to harness recent advances in machine learning for the benefit of science, sharpening the machine learning toolset, and tackling the most pressing questions in a wide range of scientific disciplines.

Keywords: Machine Learning

1 Introduction

Over the past decade, Tübingen has emerged as a leading hub for Artificial Intelligence (AI) and Machine Learning (ML) research in Germany and Europe. One of the pillars for this development into an internationally visible research location for AI and ML is the Cluster of Excellence "Machine Learning: New Perspectives for Science" (ML Cluster), which was established at the University of Tübingen, Germany. The ML Cluster receives funding from the German Research Foundation, in its first funding period from 2019 to 2025 in total 37,4 Mio. Euro. In the second funding period from 2026 to 2032, which was just recently approved, the cluster will receive 52,4 Mio. Euro. The aim of the ML Cluster of Excellence is to advance machine learning to aid scientific understanding across a wide range of disciplines – from medicine and neuroscience to cognitive science, linguistics, and economics to physics and the geosciences – and to better understand and steer the impact of machine learning on scientific practices.

2 The first Funding Period: 2019-2025

When our Cluster of Excellence started in 2019 the revolution in machine learning, triggered by the rise of deep learning, had already begun to change industry. However, uptake in science was more limited, even in academic disciplines traditionally close to machine learning, such as neuroscience and bioinformatics.

The ambition when establishing the Cluster of Excellence at the University of Tübingen was to couple the machine learning research community with excellent researchers in a wide range of other scientific disciplines. The central hypothesis driving this Cluster of Excellence is that innovations in machine learning facilitate the discovery process in science and that, vice versa, scientific constraints and requirements can set targets for algorithmic improvements in machine learning.

Approach

Structurally, the first funding period of this Cluster of Excellence was dedicated to two high-level goals: (1) strengthening the field of core machine learning and (2) establishing new connections with carefully selected disciplines at the University of Tübingen that did not yet have machine learning in their toolbox. As a sustainable and lasting investment in the foundations of machine learning in science, we established new professorships for "Machine Learning in Science" and "Explainable Machine Learning". We also hired Independent Research Group leaders who were working at the interface between machine learning and the sciences (from climate science over medical sciences to human and machine cognition and epistemology and ethics of machine learning) to rapidly explore the potential of their work and to build a vibrant community of young researchers.

Furthermore, graduate students and postdoctoral fellows joined the research networks funded through our intramural funding program, where principal investigators from machine learning and the sciences jointly investigated the potential of machine learning for their respective disciplines. To support our research efforts, we successfully established two core facilities: the ML Cloud, which provides computing resources, and the ML Colab to support the adoption of machine learning methods by researchers from other disciplines.

Achievements

Scientific achievements that can be counted as success for our Cluster of Excellence ideally exemplify the close interaction between machine learning and the sciences, either by developing a new machine learning algorithm triggered by a difficult, open question in the sciences, by demonstrating a novel and creative application of machine learning that leads to an important scientific insight or improves a scientific technique, or by closing the circle and doing both. Typically, these achievements do not come as single isolated moonshot breakthroughs, but rather form streams of work that build on each other as described in the following example.

Many scientific disciplines have established mechanistic models of their observed phenomena and measured systems. This is most clearly evident in physics, where there is a long tradition of describing measurements with accurate quantitative models derived from fundamental theories. For example, gravitational waves emerging from the collision of two black holes have become an important way to test gravitational theory. A model for the propagation of the waveform through space can be derived from the solutions to Einstein's

equations for two-body dynamics and gravitational radiation. The parameters of this model can provide insights both into the nature of the merged black holes as well as the underlying theories. Within this field, a significant problem has been that parameter estimation is extremely time-consuming, preventing rapid analysis in the face of ever-increasing data from new detectors such as LIGO (*Aasi et al., 2015*). In the first funding period two principal investigators of the Cluster of Excellence developed a new algorithm for the estimation of gravitational wave parameters using amortized, simulation-based Bayesian inference, which is orders of magnitude faster than previous sampling-based approaches while maintaining accuracy (*Dax et al., 2021*). This joint development built on previous work by Jakob Macke's research group, which had earlier developed and benchmarked new, more versatile, and more accessible algorithms for simulation-based inference (*Tejero-Cantero et al., 2020; Lückmann et al., 2021; Deistler et al., 2022*). The group subsequently extended their algorithmic framework to improve the accuracy (*Dax et al., 2023*), to be able to deal with shifts in the measurement of noise distribution over time (*Wildberger et al., 2023*), and to incorporate geometric symmetries (*Dax et al., 2022*). In sum, this collaboration is a paradigmatic example of how the collaboration of machine learning and physics can lead to new transformative data analysis methods for physics and at the same time inspire the development of new machine learning algorithms.

This is just one example underlining the collaborative approach of the Cluster of Excellence operating at the interface of machine learning and science, more can be found on our blog: https://www.machinelearningforscience.de/en/.

3 The second Funding Period: 2026-2032

New developments in machine learning and the sciences have opened up a range of new questions and opportunities. Based on our existing framework, we continue to structure our research program into the following research areas:

- Developing machine learning algorithms to help scientists discover new theories or models from data is critical in fields where high-throughput experiments generate massive datasets, such as single-cell genomics or physics. While we have made progress in estimating parameters of mechanistic models and uncovering causal relationships, recent advances in machine learning – including gradient-based optimization for complex mechanistic models and the emergence of foundation models – provide new opportunities. Based on our achievements from the first funding period we plan to develop new techniques to discover structured scientific models from data, integrate mechanistic and machine learning models to develop models that obey physical laws, and making causal inference techniques more broadly applicable.

- When combining mechanistic and machine learning models the evaluation and validation of such combined models is key. Hence, new approaches

for testing and validation are needed, as well as techniques for quantifying uncertainty in models and parameters, especially in complex model chains. Given that complex machine learning models are already being increasingly incorporated into experimental pipelines, we need to evaluate the impact of machine learning in such closed-loop settings. Building on our insights from the first funding period, we want address these issues in our next funding period.

- Enabling effective interaction with the entire machine learning workflow in science is essential, especially for scientists navigating complex machine learning workflows consisting of datasets, algorithms, and training procedures without formal machine learning training. They need to tune parameters, understand model outputs, and engage with the models to gain scientific insights, all while dealing with issues of interpretability and transparency. Despite tools to automate model tuning as well as advances in interpretable models and post hoc explanations, challenges remain, particularly due to the rise of large datasets and models, and increasingly complicated model workflows. Building on our work from the previous funding period, we will work on interfaces that allow domain scientists to successfully interact with machine learning workflows.

- The widespread adoption of machine learning models across scientific disciplines is reshaping scientific practices, raising concerns about biases, limited understanding due to reliance on complex models, and the potential marginalization of alternative approaches. Therefore, critical reflection on the impact of machine learning on scientific norms and practices is crucial, necessitating multidisciplinary perspectives. Building on our successful work in philosophy and ethics in the first funding period, we will now widen our scope to integrate insights from philosophy, cultural anthropology, and law to examine how machine learning alters epistemic norms, study how data curation, legal norms, and scientific practices influence 'data journeys' and explore interdisciplinary perspectives on quantitative evaluation metrics such as algorithmic fairness paradigms.

4 Conclusions

The aim of this research cluster is to enable machine learning to play a central role in all aspects of scientific discovery and to understand how such a transformation will impact the scientific approach as a whole. We developed a collaborative approach to address these issues and are open to engaging with other research initiatives to explore how machine learning can advance scientific discoveries.

References

Aasi J., Abbott B., Abbott R., Abbott T., Abernathy M., Ackley K., Adams

C., Adams T., Addesso P., Adhikari R., et al. (2015): Advanced ligo, *Classical and Quantum Gravity*, 32(7): 074001.

Dax M., Green S., Gair J., Macke J., Buonanno A., Schölkopf B. (2021): Real-time gravitational wave science with neural posterior estimation, *Physical Review Letters*, 127(24): 241103.

Dax M., Green S., Gair J., Deistler M., Schölkopf B., Macke J. (2022): Group equivariant neural posterior estimation", *International Conference on Learning Representations (ICLR)*.

Dax M., Green S., Gair J., Pürrer M., Wildberger J., Macke J., Buonanno A., Schölkopf B. (2023): Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference", *Physical Review Letters*, 130(17): 171403.

Deistler M., Goncalves P., Macke J. (2022): Truncated proposals for scalable and hassle-free simulation-based inference", *Neural Information Processing Systems (NeurIPS)*.

Lueckmann J.-M., Boelts J., Greenberg D., Goncalves P., Macke J. (2021): Benchmarking simulation-based inference, *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Tejero-Cantero A., Boelts J., Deistler M., Lueckmann J.-M., Durkan C., Goncalves P., Greenberg D., Macke, J. (2020): SBI - A toolkit for simulation-based inference, *Journal of Open Source Software*, 5(52): 2505.

Wildberger J., Dax M., Green S., Gair J., Pürrer M., Macke J., Buonanno A., Schölkopf B. (2023): Adapting to noise distribution shifts in flow-based gravitational-wave inference, *Physical Review D*, 107(8): 084046.

Acknowledgments

Funded by the German Research Foundation DFG under Germany´s Excellence Strategy (EXC 2064/1, Project 390727645)