Mathematisches Forschungsinstitut Oberwolfach

Report No. 26/2023

DOI: 10.4171/OWR/2023/26

Machine Learning for Science: Mathematics at the Interface of Data-driven and Mechanistic Modelling

Organized by
Neil Lawrence, Cambridge UK
Jessica Montgomery, Cambridge UK
Bernhard Schölkopf, Tübingen

11 June – 16 June 2023

Abstract. Rapid progress in machine learning is enabling scientific advances across a range of disciplines. However, the utility of machine learning for science remains constrained by its current inability to translate insights from data about the dynamics of a system to new scientific knowledge about why those dynamics emerge, as traditionally represented by physical modelling. Mathematics is the interface that bridges data-driven and physical models of the world and can provide a foundation for delivering such knowledge. This workshop convened researchers working across domains with a shared interest in mathematics, machine learning, and their application in the sciences, to explore how tools of mathematics can help build machine learning tools for scientific discovery.

Mathematics Subject Classification (2020): 60XX, 62XX, 68XX, 85XX, 86A08, 92XX.

Introduction by the Organizers

The workshop *Machine Learning for Science: Mathematics at the Interface of Data-driven and Mechanistic Modelling*, co-organised by Neil Lawrence, Jessica Montgomery, and Bernhard Schölkopf was attended by 40 participants from 11 to 16 June 2023. It set out to consider how mathematical innovations can help produce machine learning tools that can be deployed in support of scientific discovery, creating new interfaces between physical and data-driven modelling approaches. In support of this objective, the workshop convened three discussion themes — Lessons from the application of machine learning in science; Foundational concepts and emerging methods; Machine learning for Earth and climate sciences — which

between them included 19 talks. This workshop report presents a summary of discussion points and insights from the workshop.

The machine learning for science community is a rendezvous point for diverse disciplinary perspectives. Effective research, development, and deployment of machine learning in science requires insights from computer science, domains of application, and software engineering, amongst other areas. Across these domains, mathematics acts as the interface between machine learning and the world, providing a foundation for theories, methods, and tools that enable its safe and effective use for scientific discovery.

Mathematics provides means of formalizing structure. In the context of machine learning for science, it allows researchers to:

- Represent pre-existing domain knowledge, embedding this in machine learning systems to deliver more reliable results;
- Formalize desiderata such as fairness, interpretability, or uncertainty, which are vital to ensure machine learning models align with user needs;
- Model users, and the interactions between the user's model of the machine learning system and machine learning's model of the user.

Opening this workshop, a series of talks focussing on different applications of machine learning in the sciences explored the capabilities of today's machine learning tools [Büttner; Igel; Machuve; Mishra; Müller]. These demonstrated how machine learning can be deployed to: stitch together different data types, allowing researchers to gain a more nuanced view of a system; extract insights from data — and speed up analysis of complex datasets — to gain a more accurate understanding of how the system works, inferring properties of the physical world; and identify areas for experimentation and theorising, showing researchers where they should focus their investigations.

Across scientific domains, today's machine learning systems share a fundamental limitation: the field is not yet at the stage where these systems are directly enabling researchers to generate new causal understandings of physical, biological, or environmental systems. Advanced data analysis has delivered novel insights about the dynamics of these systems, but has not yet allowed researchers to track back from that data to the generation of new scientific knowledge about why those dynamics emerge.

A collection of open mathematical questions — such as how to describe concepts of scientific interest and exploit existing knowledge, how to mathematically represent concepts like fairness, how to make the solutions interpretable — arise from these existing applications. Building on this exploration, the workshop's second discussion theme considered how foundational concepts and methods in the mathematics of machine learning could be applied to overcome the limitations of today's machine learning for science tools [Williamson; Hennig; Ek; Wilkinson; Bah; Kappen; Gregory; Kaski; Rahaman; Macke; Kilbertus].

A combination of established techniques and progress in machine learning is sparking innovative approaches to mathematics in this area. Effective use of established mathematics — statistics and probability; Riemannian geometry; linear

algebra; optimisation; graph theory; numerical analysis; structural causal models; and time series and dynamical systems — is fundamental to further progress in the field. Alongside these mathematical fundamentals, speakers presented ideas for linking established domains in new ways and revisiting ideas proposed before the current wave of progress in machine learning to see how they might apply in today's context. Directions identified as a result include: probabilistic numerics; Imprecise Probabilities; Finslerian geometry; adjoint latent force models; simulation-based inference; Bayesian optimisation; and equivariant models.

The challenge that follows is not only how to use machine learning for science, but how to develop machine learning as a science, connecting progress in methods and applications to theories of the world.

The workshop's third theme considered how these issues collide in a scientific area where progress in research and innovation is vital for human wellbeing: Earth and climate science [Eyring; Reichstein; Benson; Camps-Valls; Cohrs; Diaz; Reiners; Winkler]. Climate change is affecting every part of the globe, manifest in temperature changes, extreme weather events, and knock-on effects for communities and economies. The Earth is a unique system, with highly complex dynamics driven by interactions across land, ocean, ice, and atmospheric sub-systems. Climate models allow researchers to analyse this system, making predictions that are vital for policy development, climate adaptation and climate mitigation. The ability to deploy fast-evolving machine learning techniques to improve the performance of climate models and deliver actionable insights would enhance both scientific knowledge and policy responses. To deliver these, the field needs stable, physically plausible models that generalise well, include the main causal drivers, with reduced uncertainties and increased interpretability.

Both machine learning and physical models are needed to improve understandings of Earth and climate processes and their drivers. Combining these in hybrid models allows researchers to exploit prior knowledge — whether mechanistic, causal, simulation-based, or from invariant relations transferred from other settings — while leveraging insights from data. Such models are helping to: characterise systems for carbon dioxide exchange; analyse how cloud cover affects global temperature change; interrogate how processes like photosynthesis influence carbon dynamics; predict how landscapes and ecosystem services might change under changing climate conditions; forecast extreme weather events; and identify the causal relationships driving these changes in Earth and climate systems.

Progress in machine learning could unlock a shift from today's single-process models — focused on local processes or dynamics — to a digital representation of the Earth that enables the user to interrogate the impact of different policy interventions on the Earth and climate system. Delivering this vision requires advanced application of the methods and tools discussed across the workshop to produce robust, trustworthy models. The research agenda for machine learning and mathematics that follows includes parameter estimation and Bayesian inference; methods to improve generalisation across changing distributions; physics-aware or

hybrid models, and ways of managing these to ensure they capture underlying processes; explainability and interpretability, and their mathematical underpinnings; techniques to stitch together data or models at different levels of granularity; simulation and emulation; and causal machine learning.

Alongside these areas for mathematical progress, the machine learning for science community is grappling with the question of how to foster engagement between different disciplines. Effective deployment of machine learning in science requires both machine learning and domain expertise. There has been a tradition of creating benchmark challenges as a means of engaging the machine learning community: set a challenge, provide data, and have the community develop models in response. Removing the challenge from its scientific context lowers one type of barrier to entry for machine learning researchers interested in the topic — no scientific knowledge is typically required — but this decontextualization of the research question at hand can also result in models that are not well-suited to the scientific context.

Underpinning this question about interdisciplinary dialogue is a mathematics, science, and engineering challenge to create user interfaces that facilitate knowledge exchange between machine learning and science, and between model and user. Methods to encode domain knowledge in the design of machine learning systems, such as embedding physical laws or invariances that constrain the system to deliver physically plausible results, are already being successfully deployed. Innovations in interface design could take this knowledge exchange further.

Tools such as ChatGPT have attracted attention as a means of creating user-friendly interfaces between scientist and machine, but at present the field lacks a sufficiently nuanced understanding of how humans communicate uncertainty through conversation to be able to reliably deploy these in ways that encourage users to interrogate the outputs of these systems and assess their trustworthiness. Progress towards a generalised suite of machine learning tools that can be deployed across fields — AI assistants that can support researchers in their work — will require the development of agents that can help users achieve their goals in conditions where those goals might be unstated, uncertain, or changing. Both new active learning strategies, building on progress in reinforcement learning and Bayesian optimisation, and ways of delivering narratives describing the outputs of a machine learning system can help in delivering this goal.

Tackling the research agenda developed through this workshop will take a community. Time at Oberwolfach has helped create connections, collaborations, and research outputs that are helping strengthen the machine learning for science community. The organisers would like thank the staff of the Mathematisches Forschungsinstitut Oberwolfach for their help before, during, and after the workshop, and participants for their contributions to discussions.

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-2230648, “US Junior Oberwolfach Fellows”.

Workshop: Machine Learning for Science: Mathematics at the Interface of Data-driven and Mechanistic Modelling

Table of Contents

Abstracts

More than the sum of its parts — single-cell multi-omics applications in cancer research

Maren Büttner

The advent of single-cell transcriptomic, epigenomic and proteomic profiling allows to characterize the cellular heterogeneity of healthy and diseased tissue at unprecedented level. The prognosis and treatment of cancer depends on multiple factors such as genetic predisposition, tumor microenvironment, cell-cell communication, and intra-tumor heterogeneity with distinct genetic alterations. These multifaceted events are not reflected by a single molecular modality. Multi-omics sequencing approaches combine several molecular modalities like chromatin accessibility and gene expression profiling of the same cells. Such combined molecular readouts provide essential insights to capture the causes and consequences of cancerous alterations at single-cell and high spatial resolution. At the same time, we require novel machine learning approaches for integrative analysis of multi-omics data.

In my talk, I introduce multimodal single-cell omics data with a particular focus on paired measurements of gene expression and chromatin accessibility in the same cells. The integrative analysis of multiple modalities requires novel computational models. A NeurIPS challenge addressing modality prediction, reconstruction and paired data integration in 2021 and 2022 attracted numerous submissions from the machine learning community outside the genomics domain. While lowering the entry barrier to the field of single-cell genomics, there is the potential of high scoring in the challenge without presenting a productive solution to the field.

My work focuses on understanding the molecular drivers in prostate cancer tumoroids at single-cell resolution, where we quantified gene expression and chromatin accessibility from the same cells. Joint modelling of both gene expression and chromatin accessibility to understand gene regulatory circuitry leverages the concept of "closeness" of regulatory element and regulated gene on the genome. However, cancer cells often exhibit extensive genomic rearrangements (such as copy number variations), which may shuffle regulatory elements next to different genes. Such rearrangements can be reconstructed from high dimensional single-cell data and used to accurately infer the perturbed biological manifold and explore avenues to reprogram cancer cells into less aggressive states.

References

[1] Lance et al, *Multimodal single cell data integration challenge: Results and lessons learned*, Proceedings of Machine Learning Research **176**:162–176, 2022 NeurIPS 2021 Competition and Demonstration Track.

[2] Chan et al, *Lineage plasticity in prostate cancer depends on JAK/STAT inflammatory signaling*, Science, 2022.

[3] Lynch et al, *MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells*, Nature Methods **19** (2022), 1097–1108.

Regression of ecosystem properties: Bias, monotonicity, and uncertainty

Christian Igel

This talk considers the regression of ecosystem properties from remote sensing data, in particular of tree height and biomass from optical satellite imagery and airborne LiDAR with a focus on removing bias, quantifying uncertainty, and imposing monotonicity.

When training deep learning models for least-squares regression, the training error residuals of the final model, selected after a fixed training time or based on performance on a hold-out data set, typically do not sum to zero. This can introduce a systematic error that accumulates if we are interested in the total aggregated performance over many data points. A simple post-processing step after training efficiently solves this problem (Igel and Oehmcke, 2023) and improved the predictions of tree height from satellite imagery (Li et al., 2023) as well as aboveground tree biomass from airborne LiDAR (Oehmcke et al., 2022).

Allometric equations are used to predict tree biomass from remote measurements of crown size (Mugabowindekwe et al., 2022, Tucker et al, 2023) and height estimates (Li et al, 2023). This talk argues for estimating the parameters of standard parametric allometric models by direct gradient-based minimization of the prediction error (Hiernaux et al., 2023) in contrast to the state-of-the-art approach of performing regression after log-log transforming the data. In this context, the question of uncertainty quantification arises and conformal prediction will be discussed as a potential way to derive prediction intervals.

When learning non-parametric allometric models from data, monotonicity of the model is a prerequisite for scientific plausibility. This talk presents a simple modification of the min-max network (MM) architecture for learning smooth (partial) monotonic functions. The network inherits the MM asymptotic approximation properties, can be used within larger deep learning systems trained end-to-end, and performs well compared to more complex state-of-the-art neural networks for monotonic modelling.

References

[1] Hiernaux, P. et al, *Allometric equations to estimate the dry mass of Sahel woody plants from very-high resolution satellite imagery*, Forest Ecology, Management **529**, 2023 **.**

[2] Igel, C. and Oehmcke, S., *Remember to correct the bias when using deep learning for regression!*, Kunstliche Intelligenz, 2023**.**

[3] Li, S. et al, *Deep learning enables image-based tree counting, crown segmentation and height prediction at national scale*, PNAS Nexus **2**(4) (2023)**.**

[4] Mugabowindekwe, M., et al *Nation-wide mapping of tree-level aboveground carbon stocks in Rwanda*, Nature Climate Change, **13** (2022), 91-97**.**

[5] Oehmcke, S., et al *Deep Learning Based 3D Point Cloud Regression for Estimating Forest Biomass*, International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL), ACM (2022)**.**

[6] Tucker, C., et al *Sub-continental scale carbon stocks of individual trees in African drylands*, Nature, **615** (2023), 80-86**.**

Data-driven Solution for Poultry Diseases Diagnostics

Dina Machuve

Coccidiosis, Salmonella, and Newcastle are the common poultry diseases that curtail poultry production if they are not detected early. In Tanzania, these diseases are not detected early due to limited access to agricultural support services by poultry farmers. Deep learning techniques have the potential for early diagnosis of these poultry diseases. In this study, deep learning models were developed to diagnose poultry diseases by classifying healthy and unhealthy fecal images. Unhealthy fecal images may be symptomatic of Coccidiosis, Salmonella, and Newcastle diseases. We collected 1,255 laboratory-labeled fecal images and fecal samples used in Polymerase Chain Reaction diagnostics to annotate the laboratory-labeled fecal images. We took 6,812 poultry fecal photos using an Open Data Kit. Agricultural support experts annotated the farm-labeled fecal images. Then we used a baseline CNN model, VGG16, InceptionV3, MobileNetV2, and Xception models. We trained models using farm and laboratory-labeled fecal images and then fine-tuned them. The test set used farm-labeled images. The test accuracies results without fine-tuning were 83.06 percent for the baseline CNN, 85.85 percent for VGG16, 94.79 percent for InceptionV3, 87.46 percent for MobileNetV2, and 88.27 percent for Xception. Fine Tuning while freezing the batch normalization layer improved model accuracies, resulting in 95.01 percent for VGG16, 95.45 percent for InceptionV3, 98.02 percent for MobileNetV2, and 98.24 percent for Xception, with F1 scores for all classifiers above 75 percent in all four classes. Given the lighter weight of the trained MobileNetV2 and its better ability to generalize, we recommend deploying this model for the early detection of poultry diseases at the farm level. We have ongoing work to train a semantic segmentation model for prediction of the diseases at pixel level using the UNET architecture. The goal of the experiment using semantic segmentation task was to compare the performance to the image classification models. In the last part of the talk, we shared the insights on the deployment challenges of the data-driven solution for poultry diseases diagnostics at the farm level using a mobile application. Smallholder farmers are semi-illiterate with limited access to smartphones. We propose using Conversational IVR technology as digital extension agents to offer targeted information to smallholder farmers who may not have high-end ICT devices. Responsible AI practices on model interpretability and explainability are fundamental.

References

[1] Machuve, D., Nwankwo, E., Mduma, N. and Mbelwa, J. , *Poultry diseases diagnostics models using deep learning*, Front. Artif. Intell **5**:733345, 2022 **.**

[2] Machuve, D., Nwankwo, E., Mduma, N., Mbelwa, H., Maguo, E., and Munisi, C., *Machine Learning Dataset for Poultry Diseases Diagnostics*, https://zenodo.org/record/5801834, 2021**.**

Mathematical conjecture generation using Machine Intelligence

Challenger Mishra

Conjectures hold a special status in mathematics. Good conjectures epitomise milestones in the pipeline of mathematical discovery, and have historically inspired new mathematics and shaped progress in theoretical physics. At the turn of the last century, David Hilbert put forward a list of 23 problems which have since driven development in geometry, number theory, and other domains of pure mathematics.

One such conjecture is the thirteenth problem, which ultimately led to the development of a formidable representation theorem for real continuous multivariate functions in the 1950's (due to Kolmogorov and Arnold). Not only did this posit a new vision of geometry, decades later this helped establish neurocomputing on firmer mathematical footing [13]. Another profound example was a key observation by McKay that led to monstrous moonshine, a phenomenon connecting two seemingly disjointed parts of mathematics, namely, monster groups and modular forms.

Hilbert's lecture at the International Congress of Mathematics in Paris, exemplified the significance of identifying noteworthy problems by making mathematical conjectures and their impact on developing new mathematics. The spirit of proposing open problems and identifying non-trivial conjectures continues to be a rewarding and a common practice endorsed by practitioners and institutions dedicated to furthering mathematics [5]. Other notable lists of problems have been those collected by Landau [8], Weil [23], Thurston [22], and the Clay Mathematics Institute [14].

Conjectures are unproven propositions. Formulating meaningful conjectures can be nontrivial; albeit one can be aided by discovering new patterns and formulating well defined closed form expressions. The Birch and Swinnerton-Dyer conjecture [3, 4], an unsolved Millennium Prize Problem, was one of the early examples of computer assisted conjecture generation, which was proposed based on a numerical tests performed on large data in 1960s, driven by intuitions and expert insights. It is generally known that computers are canonically good at pattern recognition and processing large volumes of data algorithmically. Building on machine intelligence and domain expertise, notable recent strides were made in topics in knot theory, in finding formulae to equate fundamental constants as continued fractions, improving algorithms for matrix multiplication and sorting, among a host of other applications. A common practice in these approaches has been to have interactions with domain experts for a machine guided discovery.

Mathematical inequalities express relations of the form f ¡ g. They are ubiquitously studied across analysis, combinatorics, geometry, and so on, and are essential to bounding quantities of interest across pure and applied natural sciences. Consequently great efforts go into finding new inequalities and proving them [12, 1, 19]. Bounds on prime gaps, ground state energies in physical systems, are merely a couple of examples that highlight the importance of inequalities in mathematics and natural sciences. Further, nontrivial inequalities can have a

range of practical real world applications in a variety of scientific domains, as well as in interdisciplinary endeavours such as machine learning.

Our work highlights the potential for AI-driven conjecture generation. The initial findings in this work indicate that it is possible to generate nontrivial algebraic conjectures in mathematics, which often have relationships to existing long-standing conjectures, such as those given in the case studies inspired by the Hardy-Littlewood conjectures (involving the prime-counting function), and Babai's conjecture (on diameters of Cayley graphs of non-abelian simple groups). We expect to study properties of Hermitian matrices such as Wigner matrices and local Hamiltonians (with applications to machine learning and quantum theory) in the future. One wishful application of such AI-driven inequality-discovery is to find elusive algebraic relations between computationally hard-to-compute quantities, f (such as minimum eigenvalue of some matrix), and computationally efficient-to-compute quantities, g_1, g_2 (such as Trace of functions of the matrix). Being able to closely bound such hard quantities from above and below, $g_1 < f < g_2$, would then lead to discovering new algorithms to efficiently approximate a computationally hard quantity.

Another open direction is to adapt this framework for mathematics education; wherein the oracle we develop outputs non-trivial conjectures which can be used fruitfully to generate novel exercise problems and can be used as a pedagogical tool to offer new practice problems for mathematics education.

These possibilities motivate a principled study of the space of classes of conjectures, understanding their structure, and geometrisability, which we undertook in this work. Giving structure to this space through symmetries, reduces the computational difficulty of the underlying optimization problem, and opens up the space to probe using tools from geometric analysis such as Ricci Flow. Although we take modest theoretical strides in this endeavour, further work remains to understand invariants of free group actions on the conjecture space. Further, the search problems we have devised through case studies are typically low dimensional optimisation problems. Including non-trivial machine architectures (such as neural networks) to capture the latent space would result in formulating more difficult optimization problems. As such, it would be beneficial to realise the *oracle* in the proposed geometric setting in future implementations. This would require us to address some group and invariant theoretic questions we have raised in this work. The machine learning framework of Gaussian processes is a natural way to sample functions from distributions, with the added constraint $f < g$. With further understanding of the conjecture space, and the distribution of conjectures, a machine guided approach facilitated by Gaussian processes could be insightful.

In this work, we have effectively exploited symbolic machine learning through our representations and parameterisations. When considering different representations, further work is required to establish interactions of our conjecture-generation pipeline with formal proof assistants which are built on higher-order logic. In light of the substantial reasoning abilities of natural language, it seems worthwhile to explore natural language representations in the future.

Impactful conjectures have certain commonalities. They are nontrivial with potentially substantial evidence in favour of it. Typically their description is terse, although this is a function of the underlying mathematical representation. Such conjectures can also be a gateway to unlocking new theorems; exemplified by the Riemann Hypothesis. We envision utilising these insights accumulated in [5] to make more meaningful conjectures. In light of the outcomes in this paper, the prospect of exploiting machine learning in exploring conjecture spaces in a geometric fashion, with an aim to generating impactful conjectures is tantalising. This would also benefit from a quantum backend. As such, a truly interdisciplinary approach which also involves domain experts from different branches of mathematics is very much at the backbone of this proposed pipeline.

References

[1] Beckenback, E. and Bellman, R. *Inequalities*, Springer-Verlag, 1965.

[2] Babai, L. and Seress, A. *On the diameter of permutation groups*, European Jounral of Combinatorics, **13**, 4: 231-243, 1992.

[3] Birch, B. and Swinnerton-Dyer, H. *Notes on elliptic curves i*, Journal fur die reine und angewandte Mathematik, **212**: 7-25, 1963.

[4] Birch, B. and Swinnerton-Dyer, H. *Notes on elliptic curves ii*, Journal fur die reine und angewandte Mathematik, **212**: 79-108, 1963 .

[5] Dijkgraaf, R. *The subtle art of the mathematical conjecture*, Quanta Magazine **5**, 2019 .

[6] Davies, A. et al, *Advancing mathematics by guiding human intuition with AI*, Nature **600**, 787:70-74, 2021 .

[7] Fawzi, A. et al, *Discovering faster matrix multiplication algorithms with reinforcement learning*, Nature **610**, 7930:43-45, 2022 .

[8] Guy, R. *Unsolved Problems in Number Theory. Problem Books in Mathematics*, Springer, NY, 3rd Edition, 2004 .

[9] He, Y.H. et al. *Machine learning finite simple groups. work in progress*, 2023 .

[10] Hui, Y.H. and Kim, M., *Learning algebraic structures: preliminary investigations*, International Journal of Data Science in the Mathematical Sciences, 1-20, 2022 .

[11] Hardy, G. and Littlewood, J. *Some problems of 'partitio numerorum'; iii: On the expression of a number as a sum of primes*, Acta Mathematica **44**, 1:1-70, 1923 .

[12] Hardy, G. et al, *Inequalities*, Cambridge University Press, 1952 .

[13] Hecht-Nielsen, R. A. *Kolmogorov's mapping neural network existence theorem*, Proceedings of the international conference on Neural Networks **3**, 11-14, 1987 .

[14] Clay Mathematics Institute, *Millenium Problems*, 2000 .

[15] Lang, S. *Differential and Riemannian manifolds*, Springer Science and Business Media, 2012.

[16] Lee, J.M. *Smooth manifolds*, Springer, 2012 .

[17] Mankowitz, D. et al, *Faster sorting algorithms discovered using deep reinforcement learning*, Nature **618**, 7964:257-263, 2023.

[18] Mincu, G. et al, *Properties of some functions connected to prime numbers*, Journal of Inequalities in Pure and Applied Mathematics [electronic only] **9**, 1, 2008.

[19] Dragoslav, S. et al, *Analytic inequalities*, Springer, 1970.

[20] Raayoni, G. et al, *Generating conjectures on fundamental constants with the Ramanujan Machine*, Nature **590**, 7844:67-73, 2021.

[21] Rosser, J.B. and Shoenfeld, L. *Approximate formulas for some functions of prime numbers*, Illinois Journal of Mathematics, **6**, 1:64-94, 1962.

[22] Thurston, W.P. *Three dimensional manifolds, kleinian groups and hyperbolic geometry*, Bulletin (New Series) of the American Mathematical Society **6**, 3:357-381 1982**.**
[23] Weil, A. *Numbers of solutions of equations in finite fields*, . Bulletin of the American Mathematical Society **55**, 5:497-508, 1949**.**

ML meets quantum chemistry

Klaus-Robert Müller

I will introduce a large body of work on ML for quantum chemistry that I have done with my collaborators and discuss the key (math, ML and Chemistry) challenges (see Chmiela et al 2017, 2023, Unke et al 2021a, Keith et al 2021, Unke et al 2021b). Interestingly ML models have been enabling novel chemical insight (Schütt et al 2017, 2018). The field ML for quantum chemistry emerged from an IPAM program in 2011 (Rupp et al 2012, Snyder et al 2012) and is now a large community with many hundreds of active participants.

References

[1] Rupp, M., Tkatchenko, A., Müller, K. R., and Von Lilienfeld, O. A. *Fast and accurate modeling of molecular atomization energies with machine learning.* Physical review letters, **108**(5), 058301, 2012**..**
[2] Snyder, J. C., Rupp, M., Hansen, K., Müller, K. R., and Burke, K. *Finding density functionals with machine learning.* Physical review letters, **108**(25), 253002, 2012**.**
[3] Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., and Tkatchenko, A. *Quantum-chemical insights from deep tensor neural networks.* Nature communications, **8**(1), 13890, 2017**..**
[4] Schütt, K. T., Sauceda, H. E., Kindermans, P. J., Tkatchenko, A., and Müller, K. R.. *Schnet–a deep learning architecture for molecules and materials.* The Journal of Chemical Physics, **148**(24), 2018**.**
[5] Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., and Müller, K. R. *Machine learning of accurate energy-conserving molecular force fields.* Science advances, **3**(5), e1603015, 2017**.**
[6] Chmiela, S., Vassilev-Galindo, V., Unke, O. T., Kabylda, A., Sauceda, H. E., Tkatchenko, A., and Müller, K. R. *Accurate global machine learning force fields for molecules with hundreds of atoms.* Science Advances, **9**(2), eadf0873, 2023**.**
[7] Keith, J. A., Vassilev-Galindo, V., Cheng, B., Chmiela, S., Gastegger, M., Müller, K. R., and Tkatchenko, A. *Combining machine learning and computational chemistry for predictive insights into chemical systems.* Chemical reviews, **121**(16), 9816-9872, 2021**.**
[8] Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schütt, K. T., ... and Müller, K. R. *Machine learning force fields.* Chemical Reviews, **121**(16), 10142-10186, 2021**.**
[9] Unke, O. T., Chmiela, S., Gastegger, M., Schütt, K. T., Sauceda, H. E., and Müller, K. R. *SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects.* Nature communications, **12**(1), 7273, 2021**.**

Models of Data for Machine Learning

Robert C. Williamson

Machine Learning usually adopts the statistical or actuarial model of data, namely that the data is a "random sample" drawn from some probability distribution. In this talk I will discuss why this might not be appropriate in some cases and offer some interesting alternatives. Specifically, I will illustrate two new models from which one can derive imprecise probabilities (or upper previsions, equivalent to coherent risk measures) augmenting the several that already exist in terms of subjective views of probability:

(1) when the data has relative frequencies which do not converge; and
(2) when not every event has a probability, which turns out to be an intriguing model for the problem of "intersectionality" which has gained much attention in the social sciences.

The above two points are essentially summaries of two papers that were the basis for the talk, details and abstracts of which are appended below:

Rabanus Derr and Robert C. Williamson, *Systems of Precision: Coherent Probabilities on Pre-Dynkin-Systems and Coherent Previsions on Linear Subspaces*, arXiv:2303.0352v3

In literature on imprecise probability little attention is paid to the fact that imprecise probabilities are precise on a set of events. We call these sets systems of precision. We show that, under mild assumptions, the system of precision of a lower and upper probability form a so-called (pre-)Dynkin-system. Interestingly, there are several settings, ranging from machine learning on partial data over frequential probability theory to quantum probability theory and decision making under uncertainty, in which a priori the probabilities are only desired to be precise on a specific underlying set system. Here, (pre-)Dynkin-systems have been adopted as systems of precision, too. We show that, under extendability conditions, those pre-Dynkin-systems equipped with probabilities can be embedded into algebras of sets. Surprisingly, the extendability conditions elaborated in a strand of work in quantum probability are equivalent to coherence from the imprecise probability literature. On this basis, we spell out a lattice duality which relates systems of precision to credal sets of probabilities. We conclude the presentation with a generalization of the framework to expectation-type counterparts of imprecise probabilities. The analogue of pre-Dynkin-systems turn out to be (sets of) linear subspaces in the space of bounded, real-valued functions. We introduce partial expectations, natural generalizations of probabilities defined on pre-Dynkin-systems. Again, coherence and extendability are equivalent. A related, but more general lattice duality preserves the relation between systems of precision and credal sets of probabilities.

Christian Fröhlich, Rabanus Derr, Robert C. Williamson, *Strictly Frequentist Imprecise Probability*, arXiv:2302.03520

Strict frequentism defines probability as the limiting relative frequency in an infinite sequence. What if the limit does not exist? We present a broader theory, which is applicable also to random phenomena that exhibit diverging relative

frequencies. In doing so, we develop a close connection with the theory of imprecise probability: the cluster points of relative frequencies yield a coherent upper prevision. We show that a natural frequentist definition of conditional probability recovers the generalized Bayes rule. This also suggests an independence concept, which is related to epistemic irrelevance in the imprecise probability literature. Finally, we prove constructively that, for a finite set of elementary events, there exists a sequence for which the cluster points of relative frequencies coincide with a prespecified set which demonstrates the naturalness, and arguably completeness, of our theory.

Deep learning only works if its Bayesian, and Bayesian deep learning is easy

Philipp Hennig

Deep learning is perceived as the bane of (not only) Bayesian machine learning. Deep networks learn and predict solely point estimates, and they supposedly "just work" regardless. So is Bayesian learning unnecessary overkill and "scale is all you need"? In this talk, I will review results of Hein et al., showing that commonly used deep architectures in fact exhibit fundamental pathologies in their predictions. I will then show that these pathologies are healed by endowing the network with a probability measure – nearly any probability measure – over the weight space, and that such a "posterior" can help bridge the gap between deep and nonparametric models, turning every trained deep neural network into a Gaussian process. Finally, I will show that such a probability measure can be constructed efficiently using automatic differentiation – at the cost of a couple extra epochs at the end of training, and at a constant cost overhead (one extra backward pass) at inference time, and that it can be calibrated to allow its interpretation as a posterior measure. If time permits, I will show some example uses from scientific applications. I will use these results to argue that all deep learning should be probabilistic, albeit not necessarily Bayesian, deep learning, because it "just works", and "probabilities are all you need". The talk uses results from cooperations with, among others, Matthias Hein, Agustinus Kristiadi, Alexander Immer, Eric Daxberger, Felix Dangel, Frank Schneider and Matthias Bauer.

References

[1] MacKay, D.J.C. *The Evidence Framework applied to Classification Networks*, Neural Computation, 1992.

[2] Kristiadi, A., Hein, M., Hennig, P. *Being Bayesian, even just a bit, fixes overconfidence in ReLU Networks*, ICML 2020.

[3] Daxberger, E. et al. *Laplace Redux – Effortless Bayesian Deep Learning*, Neural Infiormation Processing Systems (NeurIPS) 2021.

[4] Kristiadi, A., Hein, M., Hennig, P. *Learnable Uncertainty under Laplace Approximationsi*, UAI 2021.

[5] Kristiadi, A., Hein, M., Hennig, P. *An infinite-feature extension for Bayesian ReLU Nets that fixes their asymptotic overconfidnce*, NeurIPS 2021.

Machine learning and stochastic metrics

Carl Henrik Ek

Statistical learning is commonly applied to high dimensional but also highly structured data. Therefore a common step is to first learn a latent representation of the data to reduce the dimensionality by exploiting the structures in the data. Given a latent representation we often further analyse the data and relationship between its different instances using simple Euclidean measures. However, for methods that directly specify the generative mapping from the latent to the observed space we can analyse the structure of the manifold using the Riemannian metric defined by pulling back the Euclidean metric from the observed space. This allows us to navigate the manifold in a manner informed by the measure of similarity defined on the observed data and compute quantities such as length and volume.

When the mapping from the latent to the observed space is stochastic this translates to a stochastic Riemannian metric. In this talk we will first show a practical proof of concept on how formulating a robotics navigation using tools from geometry simplifies the problem to that of finding geodesics. We will then show that by propagating the uncertain pull-back metric directly to lengths and volumes leads to a Finslerian structure. We will show that with a stochastic immersion following a Gaussian process leads to closed form expressions of this metric. We will conclude by discussing the benefit of dropping the quadratic constraints and move beyond the Riemannian metrics and how it allows us to embed interesting structures within the metric.

References

[1] Tosi, A., Hauberg, Sören, Vellido, A., and Lawrence, N. D. *Metrics for probabilistic geometries.* In , Uncertainty in Artificial Intelligence, 2014.

[2] Hauberg, S. *Only bayes should learn a manifold (on the estimation of differential geometric structure from data),* https://doi.org/10.48550/arXiv.1806.04994, 2018.

[3] Pouplin, A., Eklund, D., Ek, C. H., and Hauberg, S. *Identifying Latent Distances With Finslerian Geometry.* https://doi.org/10.48550/arXiv.2212.10010, 2022.

Adjoint-aided inference for latent force models

Richard Wilkinson

Linear systems occur throughout engineering and the sciences, most notably as differential equations. In many cases the forcing function for the system is unknown, and interest lies in using noisy observations of the system to infer the forcing, as well as other unknown parameters. In this talk I will show how adjoints of linear systems can be used to efficiently infer forcing functions modelled as Gaussian processes. Adjoints have recently come to prominence in machine learning, but mainly as an approach to compute derivatives of cost functions for differential equation models. Here, we use adjoints in a different way that allows us to analytically compute the least-squares estimator, or the full Bayesian posterior distribution of the unknown forcing. Instead of relying on solves of the original (forward model),

we can recast the problem as n adjoint problems, where n is the number of data points. All that is required is the ability to solve adjoint systems numerically: it does not rely upon additional tractability of the linear system such as the ability to compute Green's functions. Derivation of adjoints can be automated, and is an option in many modern ODE/PDE solvers.

When modelling the unknown forcing function as a Gaussian process, we can convert the process into a sum of basis functions (for example, Mercer eigenfunctions, random Fourier features, Laplacian basis vectors etc). We can then exploit the linearity of the observation operator to write the inference problem as a linear model, thus converting a high dimensional constrained optimization or inference problem, into an unconstrained problem. The approach has a fixed and a priori known computational cost, and results in exact inference for linear problems.

We'll demonstrate the method on both ordinary and partial differential equations, and describe an application to estimating the sources of air pollution in Kampala from a network of low cost air pollution sensors.

Based on Gahungu et al. 2022 (NeurIPS) and Smith et al. 2023 (J. Roy. Stat. Soc. C).

Efficient and Robust Optimization Methods for Training Binarize Deep Neural Networks

Bubacarr Bah

Neural network algorithms have revolutionised Machine Learning in recent years. Successes of these algorithms are spectacular in many supervised learning tasks, especially computer vision and natural language processing tasks.Furthermore, there is a great potential for ML to contribute to knowledge discovery in the form of mechanistic models. However, despite their tremendous successes these algorithms are faced with many challenges including bias, interpretability and robustness to noise. This work could be considered as a step towards improving interpretability and robustness to noise of these algorithms.

Compared to classical deep neural networks its binarized versions are among other things useful for applications on resource-limited devices due to their reduction in memory consumption and computational demands. In this work we study deep neural networks with binary activation functions and continuous or integer weights (BDNN). We show that the BDNN can be reformulated as a mixed-integer linear program with bounded weight space which can be solved to global optimality by classical mixed-integer programming solvers. Additionally a local search heuristic is presented to calculate locally optimal networks. Furthermore to improve efficiency we present an iterative data-splitting heuristic which iteratively splits the training set into smaller subsets by using the k-mean method. Afterwards all data points in a given subset are forced to follow the same activation pattern which leads to a much smaller number of integer variables in the mixed-integer programming formulation and therefore to computational improvements. Finally for the first time a robust model is presented which enforces robustness of

the BDNN during training. All methods are tested on random and real datasets and our results indicate that all models can often compete with or even outperform classical DNNs on small network architectures confirming the viability for applications having restricted memory or computing power.

This talk is based on the following papers.

(1) Kurtz, J. and Bah, B., 2020. *An Integer Programming Approach to Deep Neural Networks with Binary Activation Functions;*ICML 2020 workshop on "Beyond First Order Methods in Machine Learning".
(2) Kurtz, J. and Bah, B., 2021. *Efficient and robust mixed-integer optimization methods for training binarized deep neural networks.* arXiv preprint arXiv:2110.11382.

Why adiabatic quantum annealing is unlikely to yield quantum speed-up

Bert Kappen

Adiabatic quantum annealing (AQA) is a quantum version of simulated annealing and is one of the potentially promising methods to obtain quantum speed up over classical methods for combinatoric optimization problems. AQA maintains a quantum system in the ground state of a Hamiltonian while slowly changing the Hamiltonian through a control parameter z. This procedure allows to change the system from a trivial initial state to a ground state that encodes the solution to the combinatoric optimization problem. The time complexity of the method is determined by the so-called minimal spectral gap, which occurs when the system goes through a phase transition at a critical value of the control parameter $z*$. The smaller the gap, the slower the annealing and the longer the time complexity. We study a class of AQA protocols for which we can analytically compute the minimal spectral gap as $O(1/\sqrt(N))$ with N the total number of configurations of the problem. We also obtain an analytic expression for $z*$ as a partition sum. For some problems, such as Grover search, $z*$ can be efficiently computed and we can design an annealing schedule (how z changes with time) such that a quadratic speed up is obtained: the time complexity of AQA is $O(\sqrt(N)$, while a classical method would require $O(N)$ time. However, in general z* is intractable to compute, making an efficient AQA design unfeasible for most practical combinatoric optimization problems. We conjecture that it is likely that this negative result also applies for any other instance independent transverse Hamiltonians.

References

[1] Grover, L. K. *Quantum mechanics helps in searching for a needle in a haystack.* Physical review letters **79**, 325. 1997**.**
[2] Roland, J. and Cerf, N. J. *Quantum search by local adiabatic evolution.* Physical Review A **65**, 042308, 2002**.**

[3] Farhi, E. et al. *A quantum adiabatic evolution algorithm applied to random instances of an np-complete problem.* Science **292**, 472–475, 2001.
[4] Albash, T. and Lidar, D. A. *Adiabatic quantum computation.* Reviews of Modern Physics **90**, 015002, 2018.

Equivariant Convolutions with Tensor Images

Wilson Gregory

Convolutional neural networks [5] are one of the main tools for doing machine learning on images. These methods assume that the input images are scalar images, but in natural-science domains, image-like data sets can have a geometric object in each pixel. These datasets obey the Geometric Principle: "The laws of physics must all be expressible as geometric (coordinate-independent and reference-frame-independent) relationships between geometric objects (scalars, vectors, tensors, ...) that represent physical entities." [7]. The coordinate-independent relationships define symmetries that should be respected by any function on the dataset, such as a neural network. We will focus on symmetry with respect to translations, rotations of 90 degrees, and reflections.

It is well known that CNNs enforce translational symmetries. In fact, a translationally equivariant linear function can always be written as a convolution with sufficiently large filters [4]. However, rotational and reflection symmetries do not come for free. Symmetries can be enforced by ensuring that the convolution filters are invariant to rotations ([1],[8]). The context of geometric images introduces an additional wrinkle as rotations and reflections of geometric images also rotate and reflect the vectors or tensors in the pixels. A typical CNN on a geometric image will treat the components of the vectors and tensors as channels, but this destroys the structure and makes rotational equivariance impossible. Fortunately, 100 years of tensor analysis dating back to the work of Ricci and Levi-Civita [6] gives us explicit rules for functions on tensors. In particular, any linear function from tensors to tensors can be written with tensor products, contractions, and index permutations [2].

We use these tools from tensor analysis to build our equivariant GeometricImageNet [3]. We use a geometric generalization of convolution with outer products, tensor index contractions, and tensor index permutations to construct vector and tensor image functions that use and benefit from the vector or tensor structure. The framework permits, with a very simple adjustment, restriction to function spaces that are exactly equivariant to translations, discrete rotations, and reflections. In numerical experiments, we find that our model has good generalization for a small simulated physics system, even when trained with a small training set. We expect this tool will be valuable for scientific and engineering machine learning, for example in cosmology or ocean dynamics.

References

[1] Cohen, T. and Welling, M. *Group equivariant convolutional neural networks.* In International Conference on machine learning, pages 2990-2999. PLMR, 2016.

[2] Dimitrienko, Y. *Tensor Analysis and Nonlinear Tensor Functions.* Springer Dordrecht, 1 edition, June 2013.

[3] Gregory, W., Hogg, D.W., Blum-Smith, B., Arias, M.T., Wong, K. and Villar, S. *GeometricImageNet: Extending convolutional neural networks to vector and tensor images* arxiv.org/abs/2305.12585 Preprint, 2023.

[4] Kondor, R. and Trivedi, S. *On the generalization of equivariance and convolution in neural networks to the action of compact groups.* Proceedings of the 35th International Conference on Machine Learning, 2018.

[5] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. and Jackel, L.D. *Backpropagation applied to handwritten zip code recognition.* Neural Computation, **1**(4):541–551, 1989.

[6] Ricci, M.M.G. and Levi-Civita, T. *Méthodes de calcul differentiel absolu et leurs applications.* Mathematische Annalen, **54**(1):125–201, 1900.

[7] Thorne, K.S. and Blandford, R.d. *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics.* Princeton University Press, 2017.

[8] Weiler, M. and Cesa, G.. General e(2)-equivariant steerable, cnns, 2021.

Collaborative ML for Science

Samuel Kaski

Now that research in most empirical fields of science has become computational, in the sense that experiments are designed with simulations or insights derived from existing data, and even replaced with simulations, it is time to ask could we do research more efficiently with new tools, or even do research in new ways. The same questions arise on development work in industry. If we call the constellation of simulation tools a virtual laboratory, an important question is can we have better tools by combining strength across the different fields and developing tools usable across the different virtual laboratories. Many machine learning tools have this aim - they need to have domain-specific elements, such as the specific models which are different in, say, materials science and psychology, but operations such as experimental design given the models are general-purpose operations.

Given common interfaces for the tools in the virtual laboratories in different fields, we can ask could the researchers be helped even more than the current tools are able to. I discussed machine learning based 'sidekick' assistants, able to help other agents research their goals, even when they are not able to yet specify the goal explicitly, or it is evolving. Such assistants can help with tasks ranging from prior knowledge elicitation in modelling, at the simplest, to zero-shot assistance in design and decision making tasks, for instance in drug design. Ultimately they should be helpful for human domain experts in running experiments and solving research problems in simulation-based virtual laboratories. The assistants will be useful tools of domain experts who run such virtual laboratories, and serve as platforms for machine learning researchers to contribute to advancing research across a number of fields, each field running their virtual laboratories which combine field-specific models and domain-agnostic modelling and assistance tools.

References

[1] Klami, A., Damoulas, T., Rinke, P., and Kaski, S. *Virtual Laboratories: Transforming research with AI.* TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.20412540.v1, 2022.

[2] Mikkola, P., Todorovic, M., Järvi, J., Rinke, P., and Kaski, S. *Projective preferential Bayesian optimization.* In Proc. ICML 2020, the 37th International Conference on Machine Learning, vol 119 of PMLR, pp 6884-6892, 2020.

[3] Sundin, I., Voronov, A., Xiao, H., Papadopoulos, K., Bjerrum, E.J., Heinonen, M., Patronov, A., Kaski, S., and Engkvist, O. *Human-in-the-loop assisted de novo molecular design.* Journal of Cheminformatics, **14**:86, 2022.

[4] Mikkola, P., Martinelli, J., Filstroff, L., and Kaski, S. *Multi-fidelity Bayesian optimization with unreliable information sources.* In Proc. AISTATS 2023, the 26th International Conference on Artificial Intelligence and Statistics, vol **206** of PMLR, pp 7425-7454, 2023.

[5] Celikok, M.M., Murena, P-A., and Kaski, S. *Modelling needs user modelling.* Frontiers in Artificial Intelligence, **6**:1097891, 2023.

[6] Lintusaari, J., Vuollekoski, H., Kangasrääsiö, Skytén, K., Järvenpää, Marttinen, P., Gutmann, M., Vehtari, A., Corander, J. and Kaski, S. *ELFI: engine for likelihood-free inference.* Journal of Machine Learning Research, **19**(16):1-7, 2018.

[7] Hämäläinen, A., Celikok, M.M., and Kaski, S. *Amortized surrogates for user models.* In Proc. UAI 2023, in press. .

[8] De Peuter, S., Oulasvirta, A., and Kaski, S. *Toward AI assistants that let designers design.* AI Magazine, **44**(1):85-96, 2023.

[9] De Peuter, S. and Kaski, S. *Zero-shot assistance in sequential decision problems.* In Proc. AAAI-23, in press, 2023.

Beyond Simulation-based inference

Jakob Macke

Many fields of science make extensive use of mechanistic forward models which are implemented through numerical simulators. Simulation-based inference aims to make it possible to perform Bayesian inference on such models by only using model-simulations, but not requiring access to likelihood evaluations. While much progress has recently been made, a lot still remains to be done: spoke about ours, and others, recent work on developing simulation based inference methods using flexible density estimators parameterised with neural networks, on improving their robustness and efficiency, and applications to modelling problems in neuroscience and astrophysics. Finally, I spoke about the prospect of building large-scale models of neural computations in the Drosophila melanogaster by combining connectomic measurements and machine learning.

References

[1] Deistler, M., Macke, J. and Goncalves, P.J. *Energy efficient network activity from disparate circuit parameters*, PNAS 2022.

[2] Greenberg, D., Nonnenmacher, M. and Macke, J. Automatic posterior transformation for likelihood-free inference. International Conference on Machine Learning, 2404-2414, PMLR, 2019.

[3] Gonçalves, P. J., Lueckmann, J. M., Deistler, M., Nonnenmacher, M., Öcal, K., Bassetto, G., ... and Macke, J. H. *Training deep neural density estimators to identify mechanistic models of neural dynamics.* Elife, **9**, e56261, 2020.

[4] Deistler, M., Goncalves, P.J. and Macke, J. *Truncated proposals for scalable and hassle-free simulation-based inference*, Neurips, 2022.
[5] Dax, M., Green, S. R., Gair, J., Macke, J. H., Buonanno, A., and Schölkopf, B. *Real-time gravitational-wave science with neural posterior estimation.* PRL, 2021.
[6] Lappalainen et al. *Connectome-constrained deep mechanistic networks predict neural responses across the fly visual system at single-neuron resolution,* biorxiv 2023.

Experiment design as sequential instrument selection

Niki Kilbertus

Instrumental variable (IV) methods are used to estimate causal effects in settings with unobserved confounding, where we cannot directly experiment on the treatment variable. Instruments are variables which only affect the outcome indirectly via the treatment variable(s). Most IV applications focus on low-dimensional treatments and crucially require at least as many instruments as treatments. This assumption is restrictive: in the natural sciences we often seek to infer causal effects of high-dimensional treatments (e.g., the effect of gene expressions or microbiota on health and disease), but can only run few experiments with a limited number of instruments (e.g., drugs or antibiotics). In such under-specified problems, the full treatment effect is not identifiable in a single experiment even in the linear case. We show that one can still reliably recover the projection of the treatment effect onto the instrumented subspace and develop techniques to consistently combine such partial estimates from different sets of instruments. We then leverage our combined estimators in an algorithm that iteratively proposes the most informative instruments at each round of experimentation to maximize the overall information about the full causal effect.

References

[1] Ailer, E., Hartford, J., and Kilbertus, N. *Sequential Underspecified Instrument Selection for Cause-Effect Estimation.* Proceedings of the 40th International Conference on Machine Learning, 2023.
[2] Janzing, D. *Causal regularization.* Advances in Neural Information Processing Systems, **32**, 2019.
[3] Pfister, N., and Peters, J. *Identifiability of sparse causal effects using instrumental variables.* In Uncertainty in Artificial Intelligence, 1613-1622, PMLR, 2022.

Machine learning for improved understanding and projections of climate change

Veronika Eyring

Climate models are fundamental to understanding and projecting climate change. The models have continued to improve over the years, but considerable biases and uncertainties in their projections remain (Eyring et al., 2021a; Lee et al., 2021). Machine learning provides promising new avenues to improve Earth system models and Earth system understanding (Eyring et al., 2021b; Gentine et al., 2021; Reichstein et al., 2019).

A large contribution to long-standing systematic errors in Earth system models stems from differences in the representation of clouds and convection (i.e., deep clouds) occurring at scales smaller than the resolved model grid resolution that is typically in the order of 40-100 km in the horizontal. This impacts the models' ability to accurately project global and regional climate change, climate variability, and extremes (Gentine et al., 2021). High-resolution, cloud resolving models with horizontal resolution of a few kilometers alleviate many biases of coarse-resolution models for deep clouds and convection (Stevens et al., 2019), but they cannot be run at climate timescales for multiple decades or longer due to high computational costs.

One novel and promising path forward is hybrid climate models that combine ML-based parameterizations of smaller-scale processes with conventional treatments of larger-scale processes (Gentine et al., 2021). Deep learning trained against global storm resolving model output has been successfully substituted for conventional parameterizations of deep convection and cloud cover in climate models, thereby enhancing the fidelity of the host Earth system model to explicitly-resolved clouds and convection (Gentine et al., 2018; Rasp et al., 2018; Grundner et al., 2022). Hybrid models improved in this way have the potential to eliminate systematic biases present in state-of-the-art models participating in the Coupled Model Intercomparison Project Phase 6 (CMIP6, Eyring et al., 2016) and to provide more reliable climate projections.

However, "unconstrained" (standard) deep learning algorithms, such as feedforward neural networks, often learn spurious non-physical relationships which can substantially limit their performance in climate simulations. Combining causal discovery and deep learning can mitigate this problem by identifying direct physical drivers of subgrid-scale processes. Prognostic climate simulations with causally-informed neural network parameterization are stable, accurately represent mean climate and variability of the original climate model, and clearly outperform its non-causal counterpart. Causal discovery can play a key role in improving data-driven parameterizations (informed by causally-consistent physical fields) for both their design and trustworthiness, with implications also in other scientific disciplines (Iglesias-Suarez et al., 2023). Trust and generalizability of the ML models can be further improved by introducing climate invariant variables (Beucler et al., 2021), physical constraints, or equation discovery (Grundner et al., 2023). This approach can drive a paradigm shift in current climate and Earth system modelling towards a new data-driven, yet still physics-aware, ML-based hybrid climate model for improved understanding and projections of climate change.

The work shown in the presentation is performed as part of the European Research Council (ERC) Synergy Grant "Understanding and Modelling the Earth System with Machine Learning (USMILE, https://www.usmile-erc.eu/)".

References

[1] Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., Ahmed, F., O'Gorman, P.A., Neelin, J.D., Lutsko, N.J., Gentine, P., *Climate-Invariant Machine Learning,* https://doi.org/10.48550/arxiv.2112.08440, 2021.

[2] Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J., Taylor, K.E.: *Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization.* Geoscientific Model Development **9**(5), 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

[3] Eyring, V., N.P. Gillett, K.M. Achuta Rao, R. Barimalala, M. Barreiro Parrillo, N. Bellouin, C. Cassou, P.J. Durack, Y. Kosaka, S. McGregor, S. Min, O. Morgenstern, and Y. Sun, *2021: Human Influence on the Climate System. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 423–552, doi:10.1017/9781009157896.005, 2021a.

[4] Eyring, V., Mishra, V., Griffith, G., Chen, L., Keenan, T. F., Turetsky, M. R., Brown, S., Jotzo, F., Moore, F. C. and van der Linden, S. *Reflections and projections on a decade of climate science.* Nat. Clim. Chang. **11**, 279–285. https://doi.org/10.1038/s41558-021-01020-x, 2021b.

[5] Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., and Yacalis, G. *Could machine learning break the convection parameterization deadlock?* Geophysical Research Letters, **45**(11), 5742-5751, 2018.

[6] Gentine, P., Eyring, V. and Beucler, T. *Deep Learning for the Parametrization of Subgrid Processes in Climate Models.* In Deep learning for the Earth Sciences (eds G. Camps-Valls, D. Tuia, X.X. Zhu and M. Reichstein), https://doi.org/10.1002/9781119646181.ch21, 2021.

[7] Grundner, A., Beucler, T., Gentine, P., Iglesias-Suarez, F., Giorgetta, M. A., and Eyring, V. *Deep learning based cloud cover parameterization for ICON.* Journal of Advances in Modeling Earth Systems, 14, e2021MS002959. https://doi.org/10.1029/2021MS002959, 2022.

[8] Grundner, A., T. Beucler, P. Gentine and V. Eyring, *Data-Driven Equation Discovery of a Cloud Cover Parameterization*, arXiv preprint, http://arxiv.org/abs/2304.08063, JAMES, subm., 2023.

[9] Iglesias-Suarez, F., P. Gentine, Solino-Fernandez, B., Beucler, T., Pritchard, M., Runge, J. and Eyring, V. *Causally-informed deep learning to improve climate models and projections,* arXiv preprint, https://arxiv.org/abs/2304.12952, JGR, subm., 2023.

[10] Lee, J.-Y., J. Marotzke, G. Bala, L. Cao, S. Corti, J.P. Dunne, F. Engelbrecht, E. Fischer, J.C. Fyfe, C. Jones, A. Maycock, J. Mutemi, O. Ndiaye, S. Panickal, and T. Zhou, *Future Global Climate: Scenario-Based Projections and NearTerm Information. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 553–672, doi:10.1017/9781009157896.006, 2021.

[11] Rasp, S., Pritchard, M. S., and Gentine, P. *Deep learning to represent subgrid processes in climate models.* Proceedings of the National Academy of Sciences, **115**(39), 9684–9689, 2018.

[12] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, *Deep learning and process understanding for data-driven Earth system science.* Nature **566**(7743), 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

[13] Stevens, B., Satoh, M., Auger, L. et al. *DYAMOND: the DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains.* Prog Earth Planet Sci **6**, 61 (2019). https://doi.org/10.1186/s40645-019-0304-z.

Practical AI challenges for Earth and Sustainability

Markus Reichstein

Humankind is facing environmental challenges well beyond climate, rather a quadruple climate-biodiversity-water-health crises with systemic, interconnected elements. The role of AI can go much beyond improving models only. Integrating different AI systems can be instrumental in tackling global environmental challenges by enhancing monitoring, analysis, and response capabilities. By combining various AI technologies, such as machine learning, computer vision, and natural language processing, it becomes possible to process and analyze vast amounts of data from diverse sources. This integration enables the identification of patterns, early detection of environmental risks, and the development of targeted strategies for mitigation and adaptation. Additionally, AI systems can assist in decision-making processes by providing data-driven insights and recommendations to policymakers and stakeholders, thereby facilitating more informed and effective responses to environmental challenges.

Living in the machine learning and physics interplay for the Earth sciences

Gustau Camps-Valls

Most Earth science problems involve making inferences about the system, where accurate predictions are just a tiny part of the problem. Inferences mean understanding variables relations and deriving physically interpretable models that are simple, parsimonious, and mathematically tractable. Machine learning models alone are excellent approximators but often do not respect the most elementary laws of physics, like mass or energy conservation, so consistency and confidence are compromised [1-4]. I will review the main challenges ahead in the field and introduce several ways to live in the Physics and machine learning interplay that allow us (1) to encode differential equations and learn parameters and processes [1], (2) constrain data-driven models with physics-priors and causal dependence constraints [2], (3) improve parameterizations [1-3], (4) emulate physical models for the sake of speed up simulations, explainability and tractability [1,4], and (5) blend data-driven and process-based models that improve robustness and allow better extrapolations [1,3]. This is a collective long-term AI agenda towards developing and applying algorithms capable of discovering knowledge in the Earth system.

References

[1] Camps-Valls, G., Svendsen, D., Martino, L., Muñoz-Marí, J., Laparra, V., Campos-Taberner, M. and Luengo, D. *Physics-aware Gaussian processes in remote sensing.* Applied Soft Computing **68**:69-82, 2018.

[2] Cortés-Andrés, J., Camps-Valls, G., Sippel, S., Székely, E., Sejdinovic, D., Diaz, E. Pérez-Suay, A., Li, Z., Mahecha, M. and Reichstein, M. *Physics-aware Nonparametric Regression Models for Earth Data Analysis.* Environmental Research Letters **17** (5) 2022.

[3] Camps-Valls, G., Devis Tuia, D., Xiang Zhu, X. and Reichstein, M. (Editors). *Deep learning for the Earth Sciences: A comprehensive approach to remote sensing, climate science and geosciences.* Wiley and Sons 2021.

[4] Reichstein, M., Camps-Valls, G., Stevens, B., Denzler, J., Carvalhais, N., Jung, M. and Prabhat. *Deep learning and process understanding for data-driven Earth System Science.* Nature **566**:195-204, 2019.

Machine Learning for Uncovering Hidden Relationships and Improving Predictions in Coupled Earth System Models

Christian Reiners, Alexander Winkler

Future climatic changes, particularly the warming of the Earth's land surface, are fundamentally influenced by the evolution of biosphere-atmosphere interactions involving carbon, water, and energy in response to increasing greenhouse gases [1]. The understanding and modeling of these biogeochemical and -physical effects are associated with significant uncertainties primarily attributed to limited understanding of underlying processes. However, these effects play a critical role in shaping feedback mechanisms within the interconnected Earth system across different time scales. In this context, our current research focuses on three key processes: On a short time-scale, the response of ecosystems to water and heat stress is of paramount importance, on an intermediate time-scale, phenological shifts in response to climatic changes become prominent and on a long time-scale, the long-term land sink of carbon emissions becomes a crucial factor.

To address the processes occurring at these different time scales, we employ machine learning in two main approaches. Firstly, we utilize machine learning techniques to uncover hidden relationships from observational data to improve our process understanding. An example is the relationship between measurements of plant greenness and meteorological conditions. We applied diverse machine learning techniques to learn the functioning of phenology from observational data and interpret the resulting models [2,3]. Secondly, we use machine learning to incorporate observation-informed process parameterizations and submodels into complex coupled Earth's system models, which are difficult to capture solely through process-based modeling. This approach, called hybrid modeling, allows us, for example, to predict the intricate plant-control over the land-atmosphere flux of water and CO2 inferred from in-situ measurements [4].

Our research in these two directions is primarily focused on recognizing, comprehending, and addressing common challenges. We identify three recurrent key challenges in our studies:

(1) Equifinality
To address the challenge of equifinality, we investigate diverse approaches, including the utilization of regularization techniques (e.g., multi-task learning) and the integration of additional prior knowledge as constraints. More concretely, in [4], we employ a neural network within a hybrid model to predict high-frequent water fluxes from land to the atmosphere by inferring

multiple latent factors that control the flux. We investigate two aforementioned strategies to control for equifinality. Although the solutions in that particular study are tailored to the specific problem at hand, they serve as an initial step towards developing more generalized approaches in the future.

(2) Combining different scales in space and time
Another critical challenge stems from the substantial heterogeneity and non-stationarity inherent in the Earth's system arising from the intricate interplay among processes operating at different temporal and spatial scales. Although these interconnections make it difficult to disentangle and isolate specific processes, such as the relationship between soil moisture and air temperature in controlling carbon uptake by plants [5], they also offer valuable opportunities for insights. In one study, we capitalized on the disparity in time scales between meteorological influences and the influence of elevated CO2 levels of plant productivity. Here, we estimate the climate effect on plant productivity using machine learning and the intermediate time-scales to isolate the CO2 fertilization effect that crucially controls the long-term sink of anthropogenic carbon [6].

(3) Generalization
Further, interpreting the effects isolated by machine learning poses challenges. It is crucial to ensure the neural network accurately captures the intended process to enable correct extrapolation. We aim to develop and use neural networks to simulate factorial experiments and shed light on the functioning on captured relations. In one study, we test extrapolation and estimate the impact of late spring frost events under future global warming as predicted by a neural network trained on contemporary climate data [7]. Furthermore, many existing interpretation methods for neural networks lack the capability to detect patterns arising from multiple interacting inputs that contribute to a specific effect. To this end, we used a specialized method to interpret the prediction of phenology using meteorological data. This method not only identifies the precise timing but also determines the timescale responsible for shifts in phenology, i.e., the seasonality of vegetation [3]. In summary, our use of machine learning involves leveraging it for both uncovering hidden relationships in observational data and improving the predictions using coupled Earth's system models and hybrid modelling.

References

[1] Winkler, A., Myneni, R., Reimers, C., Reichstein, M. et al. *Carbon System State Determines Warming Potential of Emissions*, in review: DOI: 10.21203/rs.3.rs-2865911/v1, 2023.

[2] Liu, G., ..., Reimers, C., Winkler, A. et al. *Data-driven modeling of canopy geenness dynamics accounting for environmental memory effects for vegetation phenology.* in prep, 2023.

[3] Reimers, C. Rachti, D.H. et al. *Understanding Vegetation Phenology using Interpretable Machine Learning based on Wavelet Transform of Meteorological Drivers* in preparation, (www.bgc-jena.mpg.de/ creimers), 2023.

[4] ElGhawi, R., Kraft, B., Reimers, C., Reichstein, M., Körner, M., Gentine, P. and Winkler, A.J. *Hybrid modeling of evapotranspiration: inferring stomatal and aerodynamic resistances using combined physics-based and machine learning.* Environ. Res. Lett. **18**, 034039, 2023 https://doi.org/10.1088/1748-9326/acbbe0**.**

[5] Reimers, C., Winkler, A., Humphrey, V., and Reichstein, M. *Disentangling direct and indirect soil moisture effects on ecosystem carbon uptake with Causal Modeling.* In EGU General Assembly Conference Abstracts (pp. EGU22-11148), 2022 (May)**.**

[6] Zhan, C., Orth, R., Migliavacca, M., Zaehle, S., Reichstein, M., Engel, J., ... and Winkler, A. J. *Emergence of the physiological effects of elevated CO2 on land-atmosphere exchange of carbon and water.* Global Change Biology, **28**(24), 7313-7326, 2022**.**

[7] Liu, G., Migliavacca, M., Reimers, C., Bastos, A., Linscheid, N., Reichstein, M., and Winkler, A. J. *Compound effects of extreme spring temperature fluctuations on vegetation phenology (No. EGU23-14577).* Copernicus Meetings. 2023**.**

Reporter: Jessica Montgomery

Participants

Dr. Bubacarr Bah
MRC Unit The Gambia at LSHTM
Fajara
Atlantic Boulevard
P.O. Box 274
Banjul
GAMBIA

Vitus Benson
Max-Planck-Institut für Biogeochemie
Postfach 10 01 64
07701 Jena
GERMANY

Nathanael Bosch
Universität Tübingen
Maria-von-Linden Straße 6
72076 Tübingen
GERMANY

Dr. Maren Büttner
Calico LLC
1170 Veterans Blvd
South San Francisco 94080
UNITED STATES

Dr. Armando Jose Cabrera Pacheco
Tübingen AI Center
Universität Tübingen
Maria-von-Linden-Straße 6
72076 Tübingen
GERMANY

Prof. Dr. Gustau Camps-Valls
Image Processing Laboratory (IPL)
Parc Científic Universitat de València
C/ Cat. Agustín Escardino Benlloch, 9
46980 Valencia Paterna
SPAIN

Kai-Hendrik Cohrs
Image Processing Laboratory
Universidad de Valencia
C/Catedrático José Beltran, 2
46980 Valencia
SPAIN

Emiliano Díaz
Image and Signal Processing Group
Universitat de Valencia
C/ Cat. Agustín Escardino Benlloch, 9
46980 Valencia
SPAIN

Prof. Dr. Thomas Dietterich
Oregon State University
Institute for Collaborative Robotics and Intelligence Systems
School of Electrical Engineering and Computer Sience
1148 Kelley Engineering Center
Corvallis, OR 97331-5501
UNITED STATES

Dr. Carl Henrik Ek
Department of Computer Science and Technology
University of Cambridge
FE 20, William Gates Building
J.J. Thomson Avenue
Cambridge CB3 0FD
UNITED KINGDOM

Prof. Dr. Veronika Eyring
Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)
Institut für Physik der Atmosphäre (IPA)
Oberpfaffenhofen
82234 Weßling
GERMANY

Dr. Asja Fischer
Fakultät für Informatik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum
GERMANY

Wilson Gregory
Dept. of Appl. Mathematics & Statistics
The Johns Hopkins University
34th and Charles Streets
Baltimore, MD 21218-2682
UNITED STATES

Prof. Dr. Philipp Hennig
Universität Tübingen
Fachbereich Informatik
Lehrstuhl für die Methoden des
Maschinellen Lernens
Maria-von-Linden-Straße 6
72076 Tübingen
GERMANY

Prof. Dr. habil. Christian Igel
Department of Computer Science
University of Copenhagen
Universitetsparken 1
2100 København 2100
DENMARK

Prof. Dr. Bert Kappen
Radboud Universiteit Nijmegen
Postbus 9010
6525 ED Nijmegen
NETHERLANDS

Dr. Samuel Kaski
The University of Manchester
Oxford Road
Manchester M13 9PL
UNITED KINGDOM

Dr. Sammie Katt
Aalto University
P.O. Box 11000
00076 Aalto
FINLAND

Prof. Dr. Niki Kilbertus
Technical University of Munich
School of Computation, Information and
Technology
Department of Computer Science
Boltzmannstraße 3
85748 Garching bei München
GERMANY

Birgit Kühbacher
Institut für Physik der Atmosphäre
Erdsystemmodell-Evaluierung und
-Analyse
Münchner Straße 20
82234 Oberpfaffenhofen-Wessling
GERMANY

Vidhi Lalchand
Department of Computer Science
University of Cambridge
Cambridge CB1 1NP
UNITED KINGDOM

Prof. Dr. Neil Lawrence
Department of Computer Science
and Technology
University of Cambridge
FE 20, William Gates Building
J.J. Thomson Avenue
Cambridge CB3 0FD
UNITED KINGDOM

Prof. Dr. Michael J. Lindsey
Department of Mathematics
University of California, Berkeley
970 Evans Hall
Berkeley, CA 94720-3840
UNITED STATES

Dr. Gilles Louppe
Montefiore Institute
Université de Liège
Allée de la découverte 10
4000 Liège
BELGIUM

Dr. Nan Lu
Department of Computer Science
Universität Tübingen
Maria-von-Linden-Str. 6
72076 Tübingen
GERMANY

Dr. Dina Machuve
DevData Analysts
Third Floor,
MEGA Complex Building
Bondeni Street
P.O. Box 13855
Arusha 23102
TANZANIA, UNITED REPUBLIC OF

Prof. Dr. Jakob Macke
Department of Computer Science
Universität Tübingen
Maria-von-Linden-Strasse 6
72076 Tübingen
GERMANY

Dr. Challenger Mishra
Department of Computer Science
and Technology
University of Cambridge
FE 20, William Gates Building
J.J. Thomson Avenue
Cambridge CB3 0FD
UNITED KINGDOM

Jessica Montgomery
Department of Computer Science and
Technology, University of Cambridge
J.J. Thomson Avenue
Cambridge CB3 0FD
UNITED KINGDOM

Prof. Dr. Klaus-Robert Müller
Computer Science Department
Technical University of Berlin
Sekr. MAR 4-1
Marchstrasse 23
10587 Berlin
GERMANY

Prof. Dr. Jonas Peters
Seminar für Statistik
ETH Zürich
HG G 17
Rämistr. 101
8092 Zürich
SWITZERLAND

Dr. Niklas Pfister
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 København
DENMARK

Nasim Rahaman
Max Planck Institute for Intelligent
Systems
Max-Planck-Ring 4
72076 Tübingen
GERMANY

Prof. Dr. Markus Reichstein
Max-Planck-Institut für Biogeochemie
Hans-Knöll-Str. 10
Postfach 10 01 64
07701 Jena
GERMANY

Dr. Christian Reimers
Max-Planck-Institut für Biogeochemie
Hans-Knöll-Str. 10
07745 Jena
GERMANY

Prof. Dr. Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Max-Planck-Ring 4
72076 Tübingen
GERMANY

Dr. Sebastian Weichwald
Institut for Matematiske Fag
Kobenhavns Universitet
Universitetsparken 5
2100 København
DENMARK

Prof. Dr. Richard Wilkinson
School of Mathematical Sciences
The University of Nottingham
University Park
Nottingham NG7 2RD
UNITED KINGDOM

Prof. Dr. Bob Williamson
Foundations of Machine Learning systems
University of Tübingen
Maria-von-Linden-Str. 6
72076 Tübingen
GERMANY

Dr. Alexander Winkler
Max-Planck-Institut für Biogeochemie
Hans-Knöll-Str. 10
07745 Jena
GERMANY