UNIVERSITY OF CAMBRIDGE

ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

Schmidt Sciences

2024

ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

DONOR REPORT

Forward

2024 has brought a new wave of excitement about the potential of AI for science. Thanks to the continuing support of Schmidt Sciences, the University of Cambridge has been positioned to respond to this excitement with a portfolio of research, training, and community-building activities convened by the Accelerate Programme for Scientific Discovery. During year four of this programme, we've delivered a step-change in our engagement with the Cambridge AI for science community. We are pleased to share the following report that introduces the highlights from this work.

Our core research team continues to advance dynamic research agendas, in areas from string theory to sleep science, to human-machine collaboration. We are delighted to have welcomed a new cohort of early career researchers to the team this year: Timo Hromadka (Research Assistant), Haochen Liu (Research Assistant), Stathis Megas (Research Associate), Kevin Monteiro, (Research Assistant), Diana Robinson (Research Assistant), Dinithi Sumanaweera (Research Associate), and Hongyu Zhou (Research Associate).

Our collaborative research schemes have also grown the network of researchers that are engaged in delivering our aspirations for AI in science at the University. We now support 33 collaborative research projects, which are helping establish new research teams in cancer research, paleoecology, legal practice, neonatal monitoring, and more. Alongside these projects, we are supporting the development of a new cross-University, challenge-led programme in AI for education and cultural heritage, which will be launched publicly in 2025.

A highlight of 2024 has been the expansion of our training and networking opportunities for Cambridge researchers. Over 1000 researchers have engaged in our courses, workshops, events and engineering clinics this year. We've introduced new courses in Large Language Models (LLMs) and Diffusion Models for science, and we've grounded these in new hands-on learning sessions where researchers can access coaching in how to deploy AI. We plan to further grow this network in 2025, introducing 'train the trainer' sessions that will create a cohort of Accelerate AI for science ambassadors across departments.

We were pleased this year to be selected as an inaugural member of a new European collaboration in AI for science. The European Leadership in Innovation with AI and Science (ELIAS) Alliance will bring together a network of innovation hubs in Cambridge, Amsterdam, Barcelona, Copenhagen, Munich, Potsdam, Tübingen, and Zurich. Over the next year, we will work together to create opportunities for students to build skills in AI for science and entrepreneurship, and to create an environment of open AI innovation across Europe. Our collaboration with the International Computation and AI Network has also been developing at pace, with pilot projects in AI for weather forecasting, crop disease diagnosis, and humanitarian response all delivering early results. We hope that this report demonstrates the impact of the Accelerate Science donation from Schmidt Sciences. We look forward to working with you to continue to grow our work as we move into year 5 of the Programme.

Best regards,

Jessica Montgomery
Director, Accelerate Programme for Scientific Discovery and ai@cam

Neil Lawrence
DeepMind Professor of Machine Learning

Contents

↗ Research

Section One: **Research**

Accelerate Science pursues research that applies machine learning to the scientific challenges of the 21st century, generating insights that accelerate scientific progress and creating AI tools that can deliver benefits for science and society.

Advancing mathematics with machine intelligence

Dr. Challenger Mishra
Research Fellow

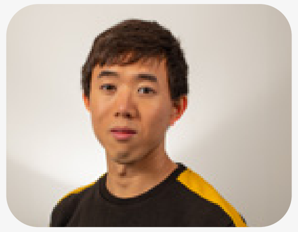
Dr Challenger Mishra leads a team developing new machine learning methodologies to study complex geometries called Calabi-Yau manifolds.

Research published by the group in 2024 has solved a long standing problem using AI to extract physical predictions from string theory. This has led to widespread interest in the group's work and team members have presented at the Harvard University Center of Mathematical Sciences and Applications and the String Data Conference at the Yukawa Institute for Theoretical Physics, Kyoto University, Kyoto, Japan. Work from the group has advanced the use of AI for mathematical conjecture generation. This opens new research directions and could lead to innovative methods for teaching mathematics in future.

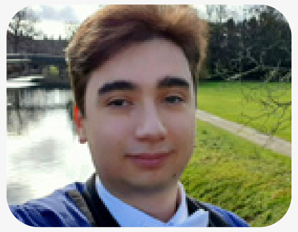
↙ PhD Students

Daattavya Aggarwal
PhD Student

Oisin Kim
PhD Student



Justin Tan
PhD Student



Viktor Mirjanic
PhD Student

AI for Mental Health

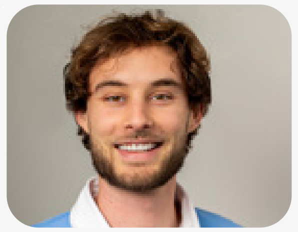
Dr Sarah Morgan
Research Fellow

Dr Sarah Morgan leads a research team using multimodal data to improve diagnosis of mental health conditions.

The team's research has generated new understandings of how speech can be used to detect psychotic disorders, helping develop a roadmap of how brain regions are related.

This year, Sarah has continued to build collaborations across the healthcare sector nationally and internationally. She spearheaded a project to gather speech samples from patients with schizophrenia in Nigeria as part of a feasibility assessment on the use of computational speech analysis in this setting. In May 2024, Sarah took up a new position as Senior Lecturer at King's College London. She continues to collaborate with Accelerate Science and to co-supervise the PhD students below as they complete their studies.

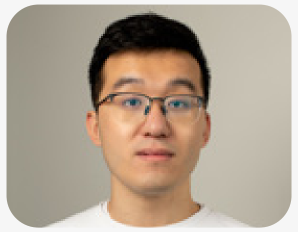
↙ PhD Students



Benjamin Chidiac
PhD student

Shankhla Pandey
PhD Student

Isaac Sebenius
PhD Student



Zhiyuan Song
PhD Student

Complex Systems, Explainable AI and Healthcare

Dr Soumya Banerjee
Senior Research Associate

Dr Soumya Banerjee is building a programme of work focused on explainable AI and its application in the healthcare context.

Soumya's work has led to development of a framework for applying participative data science to healthcare, enabling data scientists, clinicians, and patients to work together.

Highlights in 2024 include new work on the nature of intelligence and presenting initial findings at several international conferences. Soumya has supervised 5 masters students in 2024 and will present outcomes from these projects at NeurIPS workshops in December 2024. Research publications developed from work with masters students in 2023 have also been under development, with a paper published in Nature Scientific Reports on neural networks for abstraction and reasoning.

AI and Sleep

Dr Sam Nallaperuma
Senior Research Associate

Dr Sam Nallaperuma uses AI to solve modelling and optimisation problems in biological networks.

Sam's research in sleep science has shown how AI could help deliver new therapeutic interventions for patients with insomnia.

This year Sam has worked with her team to develop initial versions of the three key components of the proposed digital twin model 'BrainTwin'. New research from her team has quantitatively validated model components relating to sleep quality scoring, therapeutic generation, and brain simulation. In the next phase of work for this project, the model will be validated with human participants (both healthy volunteers and patients with insomnia) in a clinical setting. Sam has brought on board three research assistants to develop work in this area and has supervised two masters students working on AI and stress management.

↙ Research Assistants

Timo Hromadka
Research Assistant

Haochen Liu
Research Assistant

Kevin Monteiro
Research Assistant

Applying Probabilistic Machine Learning to Scientific Challenges

Aditya Ravuri
PhD Student

Working with Neil Lawrence, PhD student Aditya Ravuri has continued to explore applications of probabilistic machine learning in scientific contexts that range from cell biology to animal behaviour. Aditya's research has led to a new paradigm for probabilistic dimensionality reduction. In 2024 Aditya has been working on ProbDR 2, an analysis tool for visualisation, noise reduction and efficient downstream data processing.

Metascience

Hongyu Zhou
Research Associate

Hongyu joined Accelerate Science in September 2024 as a Research Associate. He is studying the production, consumption, and diffusion of scientific knowledge, with the aim of developing frameworks to accelerate R&D. In the coming months, he will be developing a project examining the adoption of AI across the sciences, with the aim of understanding the state of AI for science.

↗ Research

Using AI in Image Processing

Ander Biguri and Moshe Elisaof work in Professor Carola-Bibiane Schönlieb's lab in the Department of Applied Mathematics and Theoretical Physics. Work in Professor Schönlieb's group focuses on image analysis across a range of applications.

Dr Ander Biguri
Senior Research Associate

This year Ander has worked with collaborators to bridge different elements of the literature on machine learning-driven methods for Computed Tomography (CT) image reconstruction. His work has included the first experiments training AI systems for CT image recognition with real data, the first 3D CT reconstruction challenge and the first study on the limitations of synthetic data in this topic. This work is all contained within a new library for the field, LION (https://github.com/CambridgeCIA/LION) which collates AI tools for tomographic reconstruction.

Dr Moshe Elisaof
Research Associate

Dr Moshe Elisaof works on inverse problems. This year, his work has advanced the field of Geometric Deep Learning, specifically in the topic of Graph Neural Networks (GNNs), by promoting adaptivity into GNNs. Moshe's research on adaptive activation functions and normalization layers in GNNs has been accepted for publication at NeurIPS 2024.

Combining Genomics and Machine Learning

Stathis Megas
Research Associate

Dr Stathis Megas joined the team working in Professor Sarah Teichmann's lab.

His research focus is in the mathematical design of deep learning algorithms for disentangled representations and causal inference in single-cell and spatial genomics, with the aim of understanding gene regulation and protein-ligand interactions.

Dr Dinithi Sumanaweera
Research Associate

Dr Dinithi Sumanaweera joined the Accelerate Science team as a Research Associate in Professor Sarah Teichmann's lab in the Wellcome Sanger Institute and Cambridge Stem Cell Institute.

Dinithi has been working on work on single-cell trajectory alignment modelling. She has also been involved in collaborations to apply the group's trajectory alignment framework for various contexts (i.e. skin organoids, gut organoids) and has presented this work at conferences including the International Conference on Intelligent Systems for Molecular Biology in Montreal. Dinithi has recently been appointed to a new role with GlaxoSmithKline.

AI Assistants for Scientists and Engineers

Carl Henrik Ek
Professor of Statistical Learning

Professor Carl Henrik Ek continues to work in collaboration with Boeing to explore the engineering design process from a data-driven perspective. Research Associate Erik Bodin and PhD student Sam Willis have joined his team to drive progress in this area. Carl Henrik also holds a position as visiting Professor at Karolinska Institute in Stockholm and is working with Professor Erik Melén developing research in AI and cardiology. In 2025 Carl Henrik will host a research visit from a clinical doctor from the Karolinska Institute to support development of machine learning skills.

Diana Robinson
Research Assistant

Diana joined the team in May 2024 as a Research Assistant for a project focused on the Interfaces between researchers and AI, led by Professor Neil Lawrence. Diana specialises in Human Computer Interaction, working on developing machine learning tools for and with domain experts. She is particularly focused on supporting reasoning with uncertainty. Recent projects have investigated the development of clinical decision support tools for ICU doctors and research practices of an Assyriologist. Diana has also been convening a collaborative project to set out a research agenda for the development of software tools from user requirements. This work is being carried out in collaboration with Cogna, a company working on AI powered software development.

Radzim Sendyka
Research Assistant

Also working on the Interfaces project, Radzim is exploring best avenues for collaboration with domain experts using machine learning including a continuing project with Dr Jonathan Tenney in the Department of Archaeology to decipher ancient texts. By applying AI methods to these collections, new connections across ancient texts have emerged including insights on locations and date of the creation of stone tablets. Radzim has also been working with Cogna, in the space of code synthesis, data wrangling, and user interfaces.

Accelerate's collaborative research scheme:

Accelerate's collaborative research programme is now in its third year. The scheme provides funding, software engineering, and convening support to researchers across the University of Cambridge working at the interface of AI and the sciences.

Project updates – 2023 awards

In 2023, Accelerate onboarded a cohort of 11 projects through this scheme with researchers based across 12 departments. Throughout 2024, awardees have delivered events and workshops, built new AI tools or applications, and developed connections across fields from veterinary medicine to legal applications of LLMs.

Explore ↗

New AI Applications

- Using data from the Cardiovascular eHospital Research Database, a research team based in the Department of Applied Mathematics and Theoretical Physics have built a data pipeline that will allow the development of machine learning models that can personalise treatment for stroke patients. (Project leads: Dr. Smriti Agarwal, Victor Phillip Dahdaleh Heart & Lung Research Institute, Professor James Rudd, Department of Medicine, Professor Carola Schönlieb, Department of Applied Mathematics and Theoretical Physics)
- Dr Angelica Aviles-Rivero and her team have developed novel techniques for image reconstruction and denoising that can be deployed in medical applications. (Project lead: Dr Angelica Aviles-Rivero, Department of Applied Mathematics and Theoretical Physics)
- Working in collaboration with academics at the University of Warwick, Professor Sergio Bacallado de Lara and his team have carried out pre-clinical testing of molecules that could be developed into new antibiotics, which were identified by a transfer learning approach. (Project lead: Professor Sergio Bacallado de Lara, Department of Pure Mathematics and Mathematical Statistics)
- Using real-world data to optimise AI models for use in monitoring devices, Alex Grafton is integrating AI into clinical observation systems for premature babies. (Project lead: Alex Grafton, PhD student, Department of Engineering)
- A machine learning model developed by Maya Juman has predicted which fruit bat species are likely hosts of zoonotic paramyxoviruses, providing a proof of concept for model-guided viral discovery in museum collections. (Project lead: Maya Juman, PhD student, Department of Veterinary Medicine)
- AI-enabled analysis of court case outcomes has shown that LLMs can be used to extract data from legal documents and predict the outcome of court cases. (Project lead: Professor Felix Steffek, Faculty of Law)

Catalysing collaboration

- Over 100 in-person and a further 400 online participants joined the Machine Learning and AI for Hard-To-Treat Cancers: Datasets, Pipelines and Clinical Implementation Symposium. Discussions were centred around AI in precision oncology, including early prediction and detection, integration of multiple data streams to aid clinical decision making, and collaboration between industry, academia and the NHS. (Project leads: Professor Mireia Crispin-Ortuzar and Eleanor Wolmark, Department of Oncology).
- The Robust Cancer Early Detection Systems Under Distribution Shifts and Uncertainty Workshop, brought together leading experts across machine learning, data science, and clinical research. The workshop led to a collaboration with Springer Nature's British Journal of Cancer, initiating a paper call for a special edition. (Project lead: Dr Samantha Ip, Department of Public Health and Primary Care).
- The Embodied AI and Evolutionary Soft Robotics Workshop, brought together researchers to discuss the application of embodied AI in evolutionary soft robotics and how we can combine ideas and actions, theory and practice to uncover the latent potential of embodied AI, and shape the trajectory of AI-driven robotics (Project lead, Yue Xie, Marie Sklodowska-Curie Future Roads Fellow, Department of Engineering)

Developing communities of practice

↗ Project Updates

- In partnership with the Pulitzer Centre, a project led by Dr Anne Alexander has convened networks of investigative journalists to develop machine learning for journalistic investigations, initially focused on a project on illegal mining in the Amazon basin. (Project lead: Dr Anne Alexander, Director of Learning, Cambridge Digital Humanities)
- 25 researchers from diverse disciplines, including paleoecology, environmental science, statistics, and AI attended the PaleoStats workshop to explore innovations in paleoecological research, creating a new network in Paleostats and AI (Project lead: Dr Marco Aquino-Lopez, Department of Geography).

Translating emerging findings to wider impact

To translate promising approaches or prototypes developed by these projects into real-world impact, in 2024 we trialled an Impact funding scheme. With the support of this scheme:

- Dr Anne Alexander and collaborators at the Pulitzer Centre will work with the Rainforest Investigations Network, non-profit investigative journalism organisation Watershed Investigations, and local journalists from Uganda and Northern Ireland to build capacity in the use of machine learning for public interest journalism.
- Dr Edward Harding will develop a digital platform to increase the accessibility of research into pre-clinical phenotyping and drug discovery.
- Dr Ines Prata Machado will convene patients academics and policy makers to develop resources that communicate the current state of AI research in medical sciences to the public and UK policy makers.
- Professor Felix Steffek and his team will engage legal professionals with their technical work using AI to predict the outcome of court cases.

Projects funded in 2024

In 2024, this collaborative project scheme received 113 applications, with 13 projects selected for funding.

Harnessing AI for metagenomic discovery of cryptic pathogen lineages

Alexandre Almeida, MRC Career Development Fellow, Department of Veterinary Medicine

The human intestinal tract is colonised by a community of microorganisms — the human gut microbiome — with beneficial roles to human health. However, many microbial species inhabiting the human gut have the potential to cause disease including species implicated in severe infections and antibiotic resistance worldwide. This project aims to leverage high-resolution metagenomic strategies powered by AI to pre-emptively identify and track emerging pathogen variants.

Automating Data Analysis with Large Language Model Agents

Boris Bolliet, Assistant Teaching Professor in the MPhil in Data Intensive Science, Department of Physics, Cavendish Astrophysics Group

Many researchers work in large scientific collaborations where they must process, analyse, and interpret vast volumes of data using advanced statistical methods, including machine learning and AI-based techniques. The project team will explore whether it is possible to use LLMs to automate and optimise scientific workflows. The goal of this project is to develop *cmbagent*, a Multi-Agent System powered by pre-trained LLMs and designed to automate complex data analysis tasks.

Multi-modal Foundation Models for the Early Detection of Neurodegenerative Diseases

Zhongying Deng, Post-doctoral Research Associate at the Department of Applied Mathematics and Theoretical Physics

Neurodegenerative diseases like Alzheimer's and Parkinson's are chronic, incurable conditions that gradually impair brain function. Current diagnostic methods often rely on a single type of data, limiting their ability to detect neurodegenerative disease. This project addresses this limitation by using multi-modal data, e.g., MRI, PET, and genomics data, to train large-scale foundation models to improve early detection.

Enhancing the deterrence effect of anti-poaching patrols in protected areas using machine learning approaches

Charles A. Emogor, Schmidt Science Postdoctoral Fellow, Departments of Zoology and Computer Science and Technology
Anil Madhavapeddy, Professor of Planetary Computing Department of Computer Science and Technology, University of Cambridge

This project aims to apply machine learning methods to help reduce illegal hunting plaguing thousands of protected areas globally. Using existing data spanning decades from seven protected areas in three sub-Saharan African countries, the team will build prototype models to help park rangers identify where and when to patrol. This work spans savanna and forest habitats and will be the first multi-site effort using spatial data from rangers, allowing the team to capture the socio-economic variability in the distribution of poaching related threats.

Unlocking Educational Potential: Exploring the Impact of LLM Chatbot Tutors on Student Learning Psychology and Behaviours

Megan Ennion, PhD Student, Faculty of Education
Ros McLellan, Associate Professor in Teacher Education and Development / Pedagogical Innovation, Faculty of Education

Despite a growing body of research suggesting LLM tutors hold considerable potential to enhance student learning experiences and outcomes, technological advances are not yet translating into real world practice, and there are important evidence gaps around the interaction of such tutors with students. This study focuses on the effect of LLM tutors on student learning psychology and key behaviours and offers a preliminary exploration of the methodological potential of LLMs in educational research.

Predictive Modelling of OCD Subtypes and Comorbidities to Enhance Personalized Treatment Strategies

Máiréad Healy, PhD Student, Department of Psychology
Zoe Kourtzi, Professor of Experimental Psychology, Department of Psychology

OCD is a complex and heterogeneous condition, where comorbidities such as Autism Spectrum Disorder and Attention-Deficit/Hyperactivity Disorder complicate diagnosis and treatment, leading to suboptimal outcomes in clinical settings. This project aims to advance the diagnosis and treatment of OCD by developing a sophisticated machine learning model to accurately identify subtypes and associated comorbidities enabling personalized treatment strategies to improve patient care and outcomes.

Reconstructing Patient Journey for Early Detection of Endometriosis-Associated Ovarian Cancer

Dr Golnar Mahani, Research Associate, Department of Oncology, Cancer Research UK, Cambridge Institute and Early Cancer Institute

Ovarian cancer is the most lethal gynaecologic malignancy, with less than half of patients surviving five years post-diagnosis. This poor prognosis reflects lack of early detection strategies as well as rapid emergence of chemoresistance. This project aims to develop new early detection strategies by focusing on endometriosis-associated ovarian cancer (EAOC). By reconstructing patient journeys using LLMs the project aims to identify potential EAOC cohorts and gain critical insights into early signs of malignancy.

AI-powered Demand Forecasting for Enhanced Healthcare Resource Allocation in the NHS

Zidong Liu, PhD Student, Cambridge Judge Business School

This proof-of-concept project explores the potential of AI in predicting and managing healthcare demand across the UK's primary and secondary care sectors. The NHS grapples with escalating service demand due to an ageing population and resource constraints, exacerbated by the COVID-19 pandemic. This project addresses these challenges by developing predictive models that accurately forecast healthcare demand, enabling proactive resource allocation and improved patient outcomes.

Exploring Interdisciplinary Frontiers: Cognitive Science, Computational Modeling, and Artificial Intelligence

Runhao Lu, PhD student (Gates Cambridge Scholar), MRC Cognition and Brain Sciences Unit
Alexandra Woolgar, Programme Leader, MRC Cognition and Brain Sciences Unit

This project aims to foster academic exchange and collaboration among leading scholars and early career researchers in the UK and EU, within the fields of AI, cognitive neuroscience, psychology, philosophy, computer science, and robotics. By facilitating partnerships between institutions, the team aim to advance both AI's role in understanding brain function and how neuroscience can inspire new AI frameworks. Through interdisciplinary collaboration, the project seeks to bridge the gap between these rapidly evolving fields, driving innovative solutions and expanding the impact of cognitive science and related domains.

Voice in the Machine: Harnessing Speech AI for Naturalistic Prosody in Audiological Assessment

Alexis Deighton MacIntyre, Research Fellow (Leverhulme Trust), MRC Cognition and Brian Sciences Unit

Auditory disabilities affect people from all walks of life, and the World Health Organisation estimates that nearly 2.5 billion people will live with some degree of hearing loss by 2050. Clinicians and researchers rely on listening tests to assess speech perception. Breakdown between tests and realistic communicative scenarios may misalign patient expectations with outcomes and contribute to under-diagnosis of hearing disorders. This project will explore the experimental control and generative capabilities of state-of-the-art speech synthesis to produce audiological testing materials and conduct evaluation to ensure fit for use in clinical tests.

AI meets cultural heritage – Non-invasive imaging and machine learning techniques for the reconstruction of degraded historical sheet music

Anna Breger, Senior Postdoctoral Researcher, Department of Applied Mathematics and Theoretical Physics
Carola-Bibiane Schönlieb, Professor of Applied Mathematics, Head of Cambridge Image Analysis group, Department of Applied Mathematics and Theoretical Physics

This project will explore the possibilities, challenges and limitations of imaging and machine learning methods for reconstructing degraded historical sheet music. Such degradations may happen due to chemical or physical damage. The project team will form a novel collaboration network spanning libraries, imaging laboratories and AI imaging researchers. Samples will be selected of musical manuscripts of historical interest and the team will employ advanced imaging systems to scan these manuscripts and apply standard as well as newly developed, advanced machine learning methods to reconstruct the degraded parts.

Mitigating Confounding Variables in MRE Brain Imaging through Domain-Invariant Contrastive Learning

Jakob Träuble, PhD Student, Department of Chemical Engineering and Biotechnology
Gabriele Kaminski Schierle, Professor in Molecular Biotechnology, Department of Chemical Engineering and Biotechnology

Magnetic Resonance Elastography (MRE) is a promising brain imaging technique that measures the mechanical properties of brain tissue and shows potential for detecting neurological changes earlier than traditional MRI. However, its broader application is hindered by significant variability in datasets that distort machine learning models trained on such data. This project tackles this issue by developing a machine learning framework to account for and reduce the influence of these confounding variables with the goal to establish a more reliable framework for analysing MRE data, potentially improving the early diagnosis and treatment of neurological diseases.

Online training of large-scale Fortran-based hybrid computational science models, with applications in climate science

Joe Wallwork, Research Software Engineer, Institute of Computing for Climate Science (ICCS), University Information Services
Jack Atkinson, Senior Research Software Engineer, Institute of Computing for Climate Science (ICCS)

Many fields make use of large-scale Fortran codes that have been developed over decades through international research efforts, for example numerical weather and climate prediction. Recent advances in machine learning have opened up several exciting opportunities to advance these models. However, this presents a significant software challenge as much maching learning is conducted in Python. The team has developed a software solution - FTorch - that allows Python-based ML models written in PyTorch to be easily interfaced with Fortran codebases. This project seeks to extend the functionality of this software to facilitate online training of models.

Section Two: **Teaching & Learning**

Accelerate is equipping early career researchers and future research leaders with the data science and machine learning skills that can help them drive a new wave of scientific progress.

What participants have said:

"The next project in my PhD will require the use of LLMs, but I have little experience with them. The workshop gave me a good background in some of the key concepts of LLMs, and introduced me to things that will help my workflow for this project."

Feedback from participant, AI and Large Language Models workshop

"My goal was to fundamentally understand how LLMs work and learn how to get started working with them, which the workshop has enabled within just a day. I now know enough to navigate the very huge, jargon-filled landscape and specifically look for what I need to get started."

Feedback from participant, AI and Large Language Models workshop

"I was fortunate to take part in the first Accelerate Programme Study Group on LLMs with a highly interdisciplinary cohort. Learning from this group and receiving subsequent tailored support from Accelerate team members has unlocked incredible new realities for my PhD and career. A year later I'm teaching and lecturing in my own field on several of the methods I first learnt from Accelerate, proving the nuclear reaction effect of their work."

Feedback from Jacob Forward, PhD student, Faculty of History

Building Cambridge's Community of AI for Science Practitioners

The Accelerate Programme now offers postgraduate students and researchers at the University access to a suite of training courses that allow participants to progress from learning to code in Python to deploying open source foundation models in their research. During 2024, over 250 participants have engaged with these courses, which now regularly have waiting lists exceeding 100 people.

1 Python for Science

Our Python for Science course offers a self-paced learning module for researchers across disciplines who want to begin building their data skills to help answer their research questions. The course has now received over 500 registrations.

2 Publishing & Packaging

Our Publishing and Packaging workshop equips researchers with knowledge of workflows and tools they can use to package and publish their code.

3 Large Language Models

Our Large Language Models workshop is designed to support researchers to develop knowledge of large language models (LLMs) for use in scientific research and give them the confidence to take the first steps in using LLMs in their own work.

4 Diffusion Models

Our Diffusion Models workshop equips participants with an understanding of how generative AI and diffusion models work and shares an overview of research applications.

5 Hands on AI Workshop

Our Hands on AI workshop has now given over 50 participants the opportunity to work on real-life ML problems and data, and gain confidence in using tools that they can take back to their own research. We have also developed an additional workshop track focussed on LLMs.

6 Machine Learning Academy

Our Machine Learning Academy has now welcomed across 3 cohorts to a year long course in partnership with Cambridge Spark with an opportunity to develop machine learning skills across a range of specialist topics.

7 Machine Learning & Physical World

Our Machine Learning and the Physical World course provides the methodological background to machine learning applied to systems that interact directly with the real world.

In addition to these offers, our research team have delivered teaching across a range of courses:

- Challenger Mishra has taught on the graduate courses Theory of Deep Learning and Physics, Geometry, and Machine Learning, focussing on connections between machine learning and well-understood phenomena in theoretical physics.
- Sam Nallaperuma has taught on the MPhil on Basic and Translational Neuroscience, delivering teaching on "Digital twin brain modelling: Where artificial intelligence and neuroscience meet".
- Ryan Daniels co-supervised two Part 2 students from the Undergraduate Computer Science Tripos in partnership with the Mercedes-AMG Petronas F1 team. The students work with the Mercedes Strategy team to develop machine learning models for making pit stop predictions and imputing missing data.

The Accelerate Science team are now working with the community through a train the trainer programme to build capacity in the AI for science community to deliver further workshops in 2025 and continue to scale our training activities to meet demand.

Programme Management and Machine Learning Engineering

"This year we have expanded our engagement across the AI for science community at the University of Cambridge and have delivered over 1000 engagements through courses, workshops and events. Our programme of new training courses and workshops has proven popular and our series of AI clinic events has reached new audiences across the University in diverse fields from chemistry to the humanities."

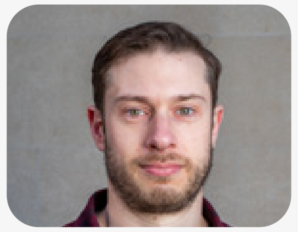
- Katie Light, Programme Manager

The staff of the Accelerate Science Programme work to deliver events and activities that engage researchers across the University.

↙ The Team

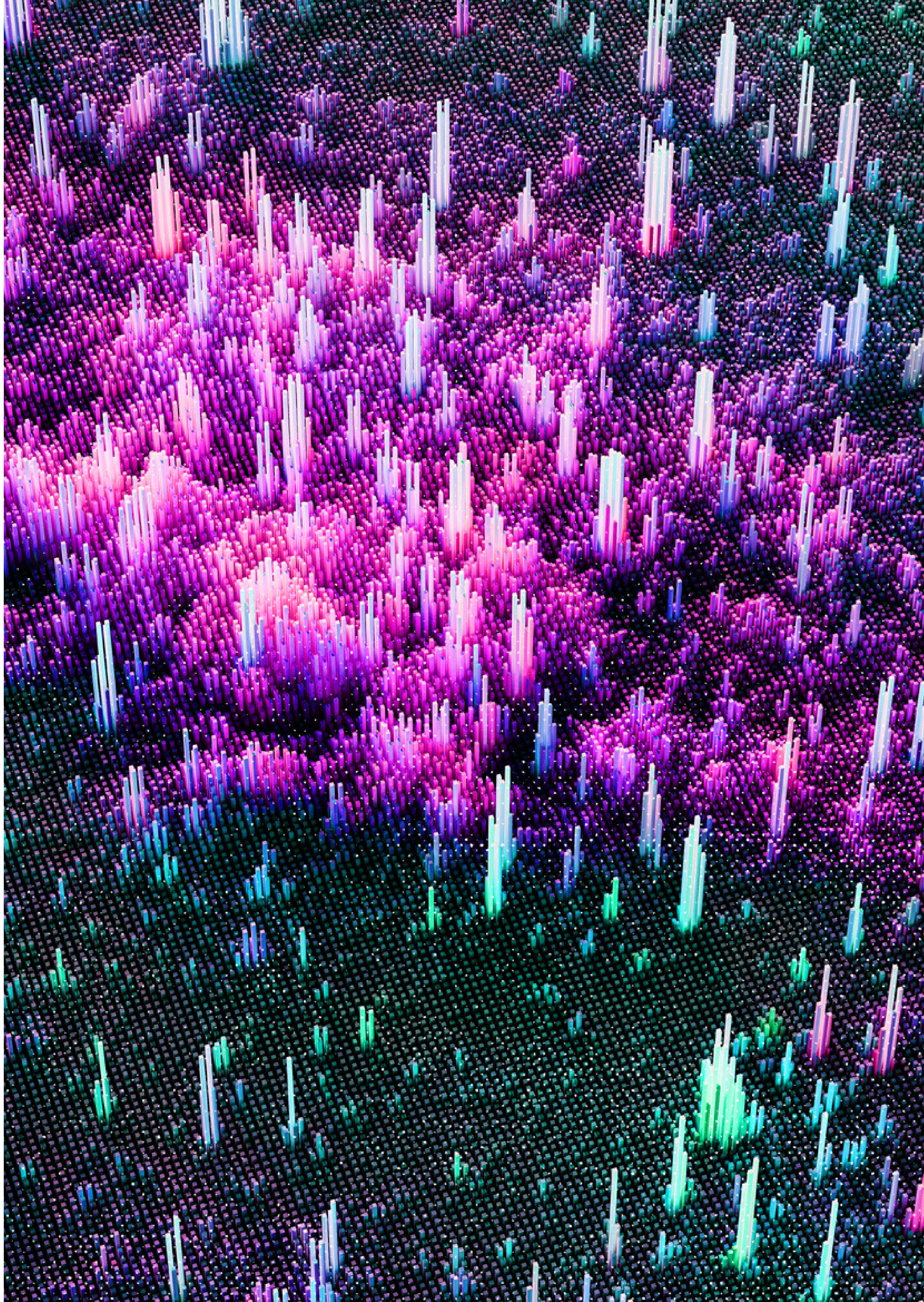
Dr Catherine Breslin, Consultant Machine Learning Engineer

Caroline Chater, Programme Coordinator



Ryan Daniels, Senior Machine Learning Engineer

Katie Light, Programme Manager



Community

Section Three:

Reaching out to the AI for science community

Building connections and developing capacity across research, policy and practice.

AI Clinic

Accelerate's AI Clinic provides researchers across the University with practical software engineering support to deploy AI in their research. This year clinic activity has grown, with 13 AI café events bringing together 128 researchers and experts in machine learning to discuss the latest research and provide support with queries at all stages of the research pipeline.

The Clinic's consultancy service has supported 118 research projects in 2024, providing both short and long-term advice for researchers working with machine learning techniques in their work.

13
AI café events

128
Researchers & experts connected

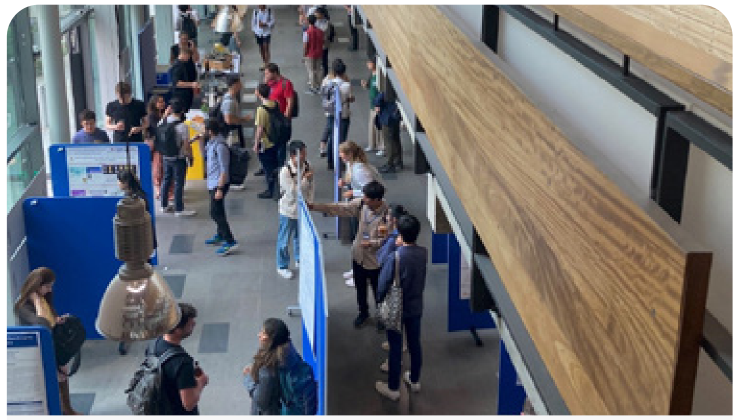
118
Research projects supported

35
Departments

Connecting to national and international dialogues about AI

We have continued to connect Accelerate's work in Cambridge to national and international conversations on AI in Science. Highlights from this work include:

- A programme of talks delivered to over 100 students at the European Laboratory for Learning and Intelligent Systems (ELLIS) Summer School on Probabilistic Machine Learning in July.
- Launch of the European Leadership in Innovation with AI and Science (ELIAS) Alliance, in which Cambridge is a founding node, in September.
- A policy workshop that convened UK and EU policymakers to explore policy frameworks to accelerate AI adoption in science in November.
- An interdisciplinary research workshop on responsible adoption of LLMs in science.
- A workshop on the development of HPC facilities to support wider adoption of AI for science, with Cambridge's Dawn supercomputer team.
- Welcoming over 100 members of the international AI for science community to Cambridge for our AI for Science summit in November, including Schmidt Sciences fellows from London, Oxford, Michigan, Singapore, Toronto and Boston.



Poster session at the ELLIS summer School, Cambridge, July 2024.

AI for Science Summit, Cambridge, November 2024.

Graphic capture of discussions at LLMs for science: best practices for safe and effective deployment workshop, Cambridge, September 2024.

Convening across the University AI community with ai@cam

ai@cam is Cambridge University's mission to develop AI that serves science, citizens, and society. In 2024, as part of ai@cam's AI-deas incubator for challenge-led research projects, Accelerate Science and ai@cam launched a call for AI-deas in education and cultural heritage. The call received 52 expressions of interest with applicants from 28 departments across the University. Two project teams have been awarded funding for interdisciplinary projects working to develop research that connects AI development to our rights to education, cultural life, and access to science.

AI for Cultural Heritage Hub (ArCH)

Led by Tuan Pham, Head of Digital Innovation & Development, University Library and Amelie Roper, Head of Research, University Library

Cambridge's galleries, libraries and museums institutions hold millions of objects spanning the globe and millennia and representing an unparalleled repository of cultural and natural history. Challenges such as analogue formats, handwritten documentation, fragmented and dispersed objects, multi-lingual sources and multi-dimensional surfaces render much of this data inaccessible. The AI for Cultural Heritage Hub (ArCH) will transform these challenges into opportunities for innovation. ArCH will be a collaborative, interdisciplinary endeavour that deploys the convening power of Cambridge's distributed network of collections and expertise for the benefit of all. The interdisciplinary team will build collaborative networks of curators, collections researchers, IT professionals and AI experts locally and globally and will model solutions that support users throughout the research lifecycle, setting standards and providing a proof of concept for future initiatives.

Opportunities and Potential Risks of AI in Supporting Evaluation (OpRaise)

Led by Dr Deborah Talmi, Senior Fellow of the Higher Education Academy, Associate Professor and Director of Curriculum Development, Department of Psychology

Higher Education has been unsettled by the potential for AI to disrupt conventional assessment and marking practices. Frontier AI models already produce passable university essays and sophisticated summative and formative feedback, and their performance improves exponentially. Given the importance of student essays in student assessment, universities must prepare for the challenges that this innovation poses to their core mission. This project focuses on the academic evaluation of written products. The team will address the key question of what value human assessment of essay submission still has in the age of AI. OpRaise will support University teaching by offering relevant quantitative and qualitative data bearing on this question. This information will provide a solid basis from which to develop strategies for effectively harnessing human-AI collaborations in university assessments, so that Higher Education institutions best continue to serve citizens and societies.

Annexes

Annex 1: 2024 Research Outputs

Members of the Accelerate Programme are listed in **bold.**

Banerjee, S., 2024. Generating complex explanations for artificial intelligence models: an application to clinical data on severe mental illness. *Life, 14(7)*, p.807.

This paper presents an explainable artificial intelligence methodology for predicting mortality in patients. The research combines clinical data from an electronic patient healthcare record system with factors relevant for severe mental illness and then applies machine learning. The machine learning model is used to predict mortality in patients with severe mental illness.

Banerjee, S., 2024, November. Building Trustworthy AI: The Role of Patient and Public Involvement in Healthcare AI Development. In *Proceedings of the AAAI Symposium Series* (Vol. 4, No. 1, pp. 329-331).

This paper explores community involvement in AI systems in healthcare projects and argues that AI algorithms for healthcare should be co-designed with patients and healthcare workers, so that they are useful and trustworthy. The paper also suggests a roadmap for this collaborative approach in AI model building.

Berglund, P., Butbaia, G., Hübsch, T., Jejjala, V., **Mishra, C**., Peña, D.M. and **Tan, J.**, 2024. cymyc--Calabi-Yau Metrics, Yukawas, and Curvature. *arXiv preprint arXiv*:2410.19728.

The authors introduce a high-performance Python library written in Jax for numerical investigation of the geometry of a large class of Calabi-Yau manifolds and their associated moduli spaces. The authors develop a well-defined geometric ansatz to numerically model tensor fields of arbitrary degree, considered as sections of the tensor bundle, on this class of Calabi-Yau manifolds. This library includes a machine learning component which incorporates this ansatz to compute tensor fields of interest on these spaces by finding an approximate solution to the system of partial differential equations they should satisfy.

Berglund, P., Butbaia, G., Hübsch, T., Jejjala, V., Peña, D.M., **Mishra, C. and Tan, J.**, 2024. Precision String Phenomenology. *arXiv preprint arXiv:*2407.13836.

In order to make physical predictions, we require certain geometric data on a given Calabi-Yau manifold. This is recovered by considering the system of partial differential equations which a given geometric object should satisfy, subsequently solved using a neural network ansatz. Function optimisation on general manifolds is highly nontrivial - any sensible model must obey a set of geometric constraints imposed by the topology of the manifold. The authors demonstrate, standard neural network solvers have difficulty adhering to these conditions tabula rasa, resulting in nonphysical predictions. The authors generate a very general ansatz, valid for any Calabi-Yau manifold, which respects these conditions by construction, and allows the modelling of tensor fields on Calabi-Yau manifold of arbitrary degree. These novel methods agree with independent numerical computations and leave the authors well-poised to generalise this program to a wider class of models.

Biguri, A. and Mukherjee, S., 2024, April. Advancing the frontiers of deep learning for low-dose 3D cone-beam computed tomography (CT) reconstruction. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)* (pp. 81-82). IEEE.

Image reconstruction from the X-ray attenuation measurement in computed tomography (CT) can be formulated as an inverse problem for recovering a function in the 3D space from its line integrals. In this work the authors propose and evaluate a 3D CBCT reconstruction challenge for machine learning methods.

Bober-Irizar, M. and **Banerjee, S.**, 2024. Neural networks for abstraction and reasoning. *Scientific Reports, 14*(1), p.27823.

This paper looks at several novel approaches for solving the Abstraction & Reasoning Corpus (ARC), a dataset of abstract visual reasoning tasks introduced to test algorithms on broad generalization. The authors revisit adapt the DreamCoder neurosymbolic reasoning solver to ARC and present the Perceptual Abstraction and Reasoning Language (PeARL) language, which allows DreamCoder to solve ARC tasks, and propose a new recognition model that allows us to significantly improve on the previous best implementation. The code is available here:

https://github.com/mxbi/dreamcoder-arc, https://github.com/mxbi/arckit

Breger, A., **Biguri, A.**, Landman, M.S., Selby, I., Amberg, N., Brunner, E., Gröhl, J., Hatamikia, S., Karner, C., Ning, L. and Dittmer, S., 2024. A study of why we need to reassess full reference image quality assessment with medical images. *arXiv preprint arXiv*:2405.19097.

The paper highlights that the commonly used metrics for image quality assessment, SSIM and PSNR, do a poor job at selecting the best medical images in 8 different clinical scenarios. These metrics may be of use for natural images, but optimizing their performance for medical images is undesirable, in general.

Butbaia, G., Peña, D.M., **Tan, J.**, Berglund, P., Hübsch, T., Jejjala, V. and **Mishra, C.**, 2024. Physical Yukawa Couplings in Heterotic String Compactifications. *arXiv preprint arXiv:*2401.15078.

This work builds a machine learning framework to compute fundamental physical parameters (consisting of particle masses and interaction strengths), corresponding to any given Calabi-Yau manifold within a broad class of string models. The authors verify correctness of the approximation by developing the mathematical theory to enable an independent, numerically exact verification, possible for a special subset of string models. The excellent agreement between both approaches opens the door to the application of the machine learning approach to a wider class of string models where no exact computation is available.

Ek, C.H., Kim, O. and Mishra, C., 2024. Calabi-Yau metrics through Grassmannian learning and Donaldson's algorithm. *arXiv preprint arXiv:*2410.11284.

Motivated by recent progress in the problem of numerical Kähler metrics, the authors survey machine learning techniques in this area, discussing both advantages and drawbacks. The authors present a novel approach to obtaining Ricci-flat approximations to Kähler metrics, applying machine learning within a `principled' framework. In particular, they use gradient descent on the Grassmannian manifold to identify an efficient subspace of sections for calculation of the metric. This approach is combined with both Donaldson's algorithm and learning on the h-matrix itself. The methods are implemented on the Dwork family of threefolds, commenting on the behaviour at different points in moduli space.

Eliasof, M., Haber, E. and Treister, E., 2024, March. Feature transportation improves graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 11, pp. 11874-11882).

Graph neural networks (GNNs) have shown remarkable success in learning representations for graph-structured data. However, GNNs still face challenges in modeling complex phenomena that involve feature transportation. In this paper, the authors propose a novel GNN architecture inspired by Advection-Diffusion-Reaction systems, called ADR-GNN and show that it improves or offers competitive performance compared to state-of-the-art networks.

Eliasof, M., Haber, E., Treister, E. and Schönlieb, C.B.B., 2024, April. On The Temporal Domain of Differential Equation Inspired Graph Neural Networks. In *International Conference on Artificial Intelligence and Statistics* (pp. 1792-1800). PMLR.

Graph Neural Networks (GNNs) have demonstrated remarkable success in modeling complex relationships in graph-structured data. A recent innovation in this field is the family of Differential Equation-Inspired Graph Neural Networks (DE-GNNs), which leverage principles from continuous dynamical systems to model information flow on graphs with built-in properties such as feature smoothing or preservation. However, existing DE-GNNs rely on first or second-order temporal dependencies. In this paper, the authors propose a neural extension to those pre-defined temporal dependencies through a new model, TDE-GNN, which can capture a wide range of temporal dynamics that go beyond typical first or second-order methods, and provide use cases where existing temporal models are challenged.

Eliasof, M., Murarib, D., Sherrya, F. and Schönlieb, C.B., 2024. Resilient Graph Neural Networks: A Coupled Dynamical Systems Approach.

Graph Neural Networks (GNNs) have established themselves as a key component in addressing diverse graph-based tasks. Despite their notable successes, GNNs remain susceptible to input perturbations in the form of adversarial attacks. This paper introduces an innovative approach to fortify GNNs against adversarial perturbations through the lens of coupled dynamical systems. A distinctive feature of the proposed approach is the simultaneous learned evolution of both the node features and the adjacency matrix, yielding an intrinsic enhancement of model robustness to perturbations in the input features and the connectivity of the graph.

Eliasof, M., Bevilacqua, B., Schönlieb, C.B. and Maron, H., 2024. GRANOLA: Adaptive Normalization for Graph Neural Networks. *arXiv preprint arXiv:*2404.13344.

Despite their widespread adoption, the incorporation of off-the-shelf normalization layers within a GNN architecture may not effectively capture the unique characteristics of graph-structured data, potentially reducing the expressive power of the overall architecture. The authors propose GRANOLA, a novel graph-adaptive normalization layer. Unlike existing normalization layers, GRANOLA normalizes node features by adapting to the specific characteristics of the graph, particularly by generating expressive representations of its neighborhood structure, obtained by leveraging the propagation of Random Node Features (RNF) in the graph. Extensive empirical evaluation of various graph benchmarks underscores the superior performance of GRANOLA over existing normalization techniques.

Gan, Z., Shevchenko, M., **Nallaperuma Herzberg, S.** and Savory, S.J., 2024. Fast neural network inverse model to maximize throughput in ultra-wideband WDM systems. *Optics Express,* 32(22), pp.38642-38654.

Ultra-wideband systems expand the optical bandwidth in wavelength-division multiplexed (WDM) systems to provide increased capacity using the existing fiber infrastructure. In this paper, the authors propose a fast and accurate data-driven deep neural network-based physical layer in this paper which can achieve 99 – 100% throughput compared to the semi-analytical approach with more than 2 orders of magnitude improvement in computational time.

García-San-Martín, N., Bethlehem, R.A., Mihalik, A., Seidlitz, J., **Sebenius, I.**, Alemán-Morillo, C., Dorfschmidt, L., Shafiei, G., Ortiz-García de la Foz, V., Merritt, K., David, A., **Morgan, S.E** et. al. 2024. Molecular and micro-architectural mapping of gray matter alterations in psychosis. *Molecular Psychiatry*, pp.1-10.

This paper explores structural magnetic resonance imaging data from 38,696 healthy controls and 1256 psychosis-related conditions. Using a normative modeling approach, the authors generated centile scores for cortical gray matter (GM) phenotypes, identifying deviations in regional volumes below the expected trajectory for all conditions, with a greater impact on the clinically diagnosed ones, FEP and chronic. Additionally, 46 neurobiological features from healthy individuals were mapped to these abnormal centiles using a multivariate approach. Results revealed that neurobiological features were highly co-localized with centile deviations, where metabolism, cerebral blood flow and neurotransmitter concentrations showed the most consistent spatial overlap with abnormal GM trajectories. Taken together these findings shed light on the vulnerability factors that may underlie atypical brain maturation during different stages of psychosis.

Haket, N. and **Daniels, R**., 2024. BERT's Conceptual Cartography: Mapping the Landscapes of Meaning. *arXiv preprint arXiv:*2408.07190.

In this paper, the authors bring tools from natural language processing into the domain of conceptual engineering, a branch of philosophy dealing with how to change public usage of words and phrases. The authors use the spoken component of the British National Corpus and BERT to create contextualised word

embeddings, and use Gaussian Mixture Models, a selection of metrics, and qualitative analysis to visualise and numerically represent lexical landscapes.

Holm, H. and **Banerjee, S.**, 2024. Intelligence in animals, humans and machines: a heliocentric view of intelligence? *Ai & Society* pp.1-3.

The authors argue that current conceptions of intelligence are anthropocentric. Looking at intelligence in both biological and artificial (human-engineered) systems can yield a more sobering and nuanced view of intelligence. Taking this into account, the authors suggest a radical reconceptualisation of intelligence.

Kiss, M.B., **Biguri, A.**, Schönlieb, C.B., Batenburg, K.J. and Lucka, F., 2024. Learned denoising with simulated and experimental low-dose CT data. *arXiv preprint arXiv:*2408.08115.

This paper analyses what happens when artificial noise is added to experimental data at the same level of real noise, and cross-compare AI methods trained in each. The authors showcase that even the most sophisticated of the "fast" X-ray noise simulators do not capture the complexity of real noise, and thus any model trained with such simulated noise may not generalize to real data.

Lawrence, N.D. and Montgomery, J., 2024. Accelerating AI for science: open data science for science. *Royal Society Open Science,* 11(8), p.231130.

Aspirations for AI as a catalyst for scientific discovery are growing. High-profile successes deploying AI in domains such as protein folding have highlighted AI's potential to unlock new frontiers of scientific knowledge. However, the pathway from AI innovation to deployment in research is not linear. Those seeking to drive a new wave of scientific progress through the application of AI require a diffusion engine that can enhance AI adoption across disciplines. Lessons from previous waves of technology change, experiences of deploying AI in real-world contexts and an emerging research agenda from the AI for science community suggest a framework for accelerating AI adoption. This framework requires action to build supply chains of ideas between disciplines; rapidly transfer technological capabilities through open research; create AI tools that empower researchers; and embed effective data stewardship. Together, these interventions can cultivate an environment of open data science that deliver the benefits of AI across the sciences.

Mantri, K.S.I., Wang, X., Schönlieb, C.B., Ribeiro, B., Bevilacqua, B. and **Eliasof, M.**, 2024. Digraf: Diffeomorphic graph-adaptive activation function. *arXiv preprint arXiv:*2407.02013.

In this paper, the authors propose a novel activation function tailored specifically for graph data in Graph Neural Networks (GNNs), DiGRAF, leveraging Continuous Piecewise-Affine Based (CPAB) transformation. In addition to its graph-adaptivity and flexibility, DiGRAF also possesses properties that are widely recognized as desirable for activation functions, such as differentiability, boundness within the domain, and computational efficiency. The authors conduct an extensive set of experiments across diverse datasets and tasks, demonstrating a consistent and superior performance of DiGRAF compared to traditional and graph-specific activation functions, highlighting its effectiveness as an activation function for GNNs. The code is available here: https://github.com/ipsitmantri/DiGRAF

Miller, J. and **Banerjee, S.**, The BACON system for equation discovery from scientific data: Transforming classical artificial intelligence with modern machine learning approaches. Preprint.

BACON is a heuristic-based computational scientific discovery system, which aims to find invariants in multivariable systems. In this paper, the authors demonstrate how they have rebuilt BACON in a modern computing language, and improve the noise-resilience of BACON. The framework is applied to a number of exemplar problems in physics and mathematics. The authors suggest that there is potential in these forgotten approaches that modern deep learning systems can learn from.

Montagnese, M., Mehta, M., ffytche, D., Firbank, M., Lawson, R.A., Taylor, J.P., Bullmore, E.T. and **Morgan, S.E.**, 2024. Disrupted functional brain network associated with presence of hallucinations in Parkinsons Disease. *medRxiv*, pp.2024-09.

Background hallucinations negatively impact quality of life in Parkinsons disease, yet their neural mechanisms remain poorly understood, particularly in early disease stages. This study aimed to identify functional connectivity differences associated with visual hallucinations in early Parkinsons Disease and to validate these findings across independent datasets. Resting state fMRI data from two prior independent studies was used. Group differences in functional connectivity were assessed within predefined cytoarchitectonic cortical classes and functional networks, followed by whole-brain analysis using Network-Based Statistics (NBS). Associations with clinical measures, including hallucination severity, motor symptoms, cognition, and attention, were also evaluated. The identified functional subnetwork shows promise as a potential biomarker and therapeutic target for Parkinsons disease psychosis, warranting further investigation and validation in future studies.

Mostowsky, P., Dutordoir, V., Azangulov, I., Jaquier, N., Hutchinson, M.J., **Ravuri, A.**, Rozo, L., Terenin, A. and Borovitskiy, V., 2024. The GeometricKernels Package: Heat and Matérn Kernels for Geometric Learning on Manifolds, Meshes, and Graphs. *arXiv preprint arXiv:*2407.08086.

Kernels are a fundamental technical primitive in machine learning. In recent years, kernel-based methods such as Gaussian processes are becoming increasingly important in applications where quantifying uncertainty is of key interest. In settings that involve structured data defined on graphs, meshes, manifolds, or other related spaces, defining kernels with good uncertainty-quantification behavior, and computing their value numerically, is less straightforward than in the Euclidean setting. To address this difficulty, the authors present GeometricKernels, a software package which implements the geometric analogs of classical Euclidean squared exponential - also known as heat - and Matérn kernels, which are widely-used in settings where uncertainty is of key interest. As a byproduct, the authors obtain the ability to compute Fourier-feature-type expansions, which are widely used in their own right, on a wide set of geometric spaces. Code available at: https://github.com/geometric-kernels/GeometricKernels

Olowu, S., **Lawrence, N. and Banerjee, S.**, 2024. Enhancing Patient Stratification and Interpretability Through Class-Contrastive and Feature Attribution Techniques. *medRxiv,* pp.2024-03. medRxiv 2024.03.25.24304824

A crucial component of the treatment of genetic disorders is identifying and characterising the genes and gene modules that drive disease processes. In this work, the team develop a comprehensive set of explainable machine learning techniques to perform patient stratification for inflammatory bowel disease.

Ravuri, A., Cooper, E. and Yamagishi, J., 2024, April. Uncertainty as a predictor: Leveraging self-supervised learning for zero-shot mos prediction. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)* (pp. 580-584). IEEE.

Predicting audio quality in voice synthesis and conversion systems is a critical yet challenging task, especially when traditional methods like Mean Opinion Scores (MOS) are cumbersome to collect at scale. This paper addresses the gap in efficient audio quality prediction, especially in low-resource settings where extensive MOS data from large-scale listening tests may be unavailable. The authors demonstrate that uncertainty measures derived from out-of-the-box pre-trained self-supervised learning (SSL) models, such as wav2vec, correlate with MOS scores. The paper explores the extent of this correlation across different models and language contexts, revealing insights into how inherent uncertainties in SSL models can serve as effective proxies for audio quality assessment.

Ravuri, A. and Lawrence, N.D., 2024. Towards One Model for Classical Dimensionality Reduction: A Probabilistic Perspective on UMAP and t-SNE. *arXiv preprint arXiv:*2405.17412.

This paper shows that the dimensionality reduction methods, UMAP and t-SNE, can be approximately recast as MAP inference methods corresponding to a generalized Wishart-based model introduced in ProbDR. This interpretation offers deeper theoretical insights into these algorithms, while introducing tools with which similar dimensionality reduction methods can be studied.

Ravuri, A., Muir, J. and **Lawrence, N.D.**, 2024. On Feature Learning for Titi Monkey Activity Detection. *arXiv preprint arXiv:*2407.01452.

This paper introduces a robust machine learning framework for the detection of vocal activities of Coppery titi monkeys. Utilizing a combination of MFCC features and a bidirectional LSTM-based classifier, the authors effectively address the challenges posed by the small amount of expert-annotated vocal data available. This approach significantly reduces false positives and improves the accuracy of call detection in bioacoustic research. Initial results demonstrate an accuracy of 95\% on instance predictions, highlighting the effectiveness of our model in identifying and classifying complex vocal patterns in environmental audio recordings. Moreover, the authors show how call classification can be done downstream, paving the way for real-world monitoring.

Robinson, D., Cabrera, C., Gordon, A.D., **Lawrence, N.D**. and Mennen, L., 2024. Requirements are All You Need: The Final Frontier for End-User Software Engineering. *arXiv preprint arXiv:*2405.13708.

This paper maps out a research agenda for building software just from requirements, discussing challenges and opportunities in three areas: elicitation, testing, and deployment and maintenance. The authors discuss the research needed to bridge the gap between where we are today and these imagined systems of the future.

Runkel, C., **Biguri, A**. and Schönlieb, C.B., 2024. Continuous Learned Primal Dual. *arXiv preprint arXiv:*2405.02478.

Neural ordinary differential equations (Neural ODEs) propose the idea that a sequence of layers in a neural network is just a discretisation of an ODE, and thus can instead be directly modelled by a parameterised ODE. The authors propose an upgrade to the well known primal dual method for reconstruction, using the literature of neural ODEs, to interpret the network as solving an ODE rather than a set of discrete steps.

Sumanaweera, D., Suo, C., Cujba, A.M., Muraro, D., Dann, E., Polanski, K., Steemers, A.S., Lee, W., Oliver, A.J., Park, J.E. and Meyer, K.B., 2024. Gene-level alignment of single-cell trajectories. *Nature Methods*, pp.1-14.

This paper presents Genes2Genes, a new Bayesian information-theoretic alignment framework for single-cell pseudotime trajectories. It demonstrates the utility of trajectory alignment for disease cell-state analysis and in vitro vs. in vivo comparative studies for organoid protocol refinement. Genes2Genes was developed as an open source Python package available at: https://github.com/Teichlab/Genes2Genes.

Zakariaei, N., Rout, S., Haber, E. and **Eliasof, M.**, 2024. Advection Augmented Convolutional Neural Networks. *arXiv preprint arXiv:*2406.19253.

Many problems in physical sciences are characterized by the prediction of space-time sequences. Modern techniques for the solution of these problems typically combine Convolution Neural Networks (CNN) architecture with a time prediction mechanism. However, oftentimes, such approaches underperform in the long-range propagation of information and lack explainability. In this work, the authors introduce a physically inspired architecture for the solution of such problems. Namely, they propose to augment CNNs with advection by designing a novel semi-Lagrangian push operator and show that the proposed operator allows for the non-local transformation of information compared with standard convolutional kernels.

Zhao, S., **Ravuri, A**., Lalchand, V. and **Lawrence, N.D**., 2024. Scalable Amortized GPLVMs for Single Cell Transcriptomics Data. *arXiv preprint arXiv:*2405.03879.

Dimensionality reduction is crucial for analyzing large-scale single-cell RNA-seq data. Gaussian Process Latent Variable Models (GPLVMs) offer an interpretable dimensionality reduction method, but current scalable models lack effectiveness in clustering cell types. The authors introduce an improved model, the amortized stochastic variational Bayesian GPLVM (BGPLVM), tailored for single-cell RNA-seq with specialized encoder, kernel, and likelihood designs. This model matches the performance of the leading single-cell variational inference (scVI) approach

on synthetic and real-world COVID datasets and effectively incorporates cell-cycle and batch information to reveal more interpretable latent structures as demonstrated on an innate immunity dataset.

Annex 2: MPhil projects supervised in 2024

In 2024, Soumya Banerjee and Sam Nallaperuma supervised 7 master's students. Details of their project titles are provided below:

Supervised by Soumya Banerjee

Kiril Bikov, Multi-Agent System with Reflection for the Abstraction and Reasoning Corpus

N'yoma Diamond, On the Ethical Considerations of Generative Agents

Alexandra Herghelegiu, Towards Differentiable Automatic Concept Discovery in Graph Neural Networks

Jonah Miller, The BACON system for equation discovery from scientific data: Transforming classical artificial intelligence with modern machine learning approaches

Jad Sbai, DECoD: A framework for Data-centric and Explainable Imputation of Clinical Data

Supervised by Sam Nallaperuma

Rishab Balse, Generative AI based therapy development for stress management

Sonia Kozsut, Diagnostic framework for stress using AI and EEG

UNIVERSITY OF CAMBRIDGE

APSCI ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

Schmidt Sciences

2024

ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

DONOR REPORT