AI for Science Summit

Flash talks

What makes a wave break? How machine learning can shed light on the underlying physics of breaking waves

Tianning Tang, Department of Engineering Science, University of Oxford

A wide class of supervised machine learning methods are known to be excellent at modelling complex systems empirically. These empirical models, however, usually provide only limited physical explanations about the underlying systems. Instead, so-called “knowledge discovery” methods can be used to explore the governing equations that describe observed phenomena. In this talk, I will focus on how we can use such methods to explore the underlying physics and also model a commonly observed yet not fully understood phenomenon — the breaking of water waves.

In our work, we use symbolic regression to explore the equation that describes wave breaking evolution from a dataset of in silico waves generated using extremely expensive methods (a typical 2D wave with a few breaking periods takes the university’s cluster around two days to simulate). Our work discovers a new boundary equation which provides a reduced order description on how the surface elevation (i.e. the water-air interface) evolves forward in time, including the time period when the wave breaks, which has defied traditional approaches to this problem. Unlike empirical models where the underlying dynamics are hidden in huge numbers of highly tuned parameter values, the physical meaning of each term of our discovered equation can be revealed successfully through mathematical derivation or simulation.

Flash talks

Our equation suggests a new characteristic of breaking waves in deep water – a decoupling between the water-air interface and the fluid velocities. It also hints at much cheaper ways to computationally simulate breaking waves, which we are currently working on.

AI for Ocean Science

Emma Boland, British Antarctic Survey

A brief look at how machine learning techniques are being used to study ocean observations and improve ocean models.

Mathematical conjecture generation using Machine Intelligence

Challenger Mishra, Department of Computer Science and Technology, University of Cambridge

Conjectures hold a special status in mathematics. Good conjectures epitomise milestones in mathematical discovery, and have historically inspired new mathematics and shaped progress in theoretical physics. Hilbert's list of 23 problems and André Weil's conjectures oversaw major developments in mathematics for decades. Crafting conjectures can often be understood as a problem in pattern recognition, for which Machine Learning (ML) is tailor-made. . In this talk, I will propose a framework that allows a principled study of a space of mathematical conjectures. Using this framework and exploiting domain knowledge and machine learning, we generate a number of nontrivial conjectures in number theory and group theory. I will present evidence in support of resulting conjectures and present some new theorems. I will lay out a vision for this endeavour, and conclude by posing some general questions about the pipeline.

Flash talks

AI-enhanced synthesis to save biodiversity

Alec Christie, Department of Zoology, University of Cambridge

Biodiversity is declining rapidly and we need to ensure that decisions to save our planet are informed by the most up-to-date and relevant scientific evidence from the literature. However, current large-scale synthesis methods are bogged down in using slow, manual methods using teams of postdocs to find relevant papers, extract data, critically appraise studies, and synthesise their findings into recommendations. AI could vastly increase the speed of the synthesis pipeline, accelerating the communication of important scientific evidence to key decision-makers, thus improving the effectiveness of conservation actions and solutions. I will outline our work on tackling the start of this pipeline and our future plans to use Large Language Models (LLMs) and the Conservation Evidence database to tackle the trickier problems that lie further down the pipeline.

Customising 3D Cell Segmentation to Study Preimplantation Mouse Embryos

Anita Karsa, Department of Physiology, Development and Neuroscience, University of Cambridge

Biologists hope to obtain critical information on fertility by studying the preimplantation development of mouse embryos. Light sheet microscopy enables them to acquire high-resolution 3D+time images to investigate cell division and cellular arrangement across time. 3D cell segmentation is a crucial first step in the quantitative analysis of these images. However, given the enormous size of these rich datasets, manual segmentation is extremely time-consuming.

Flash talks

Deep-learning-based cell segmentation has been shown to be very promising in 2D, but most of the pre-trained 3D neural networks generalise very poorly due to the lack of annotated 3D data. We trained StarDist3D, a 3D U-Net-based neural network designed for cell segmentation, on representative mouse embryo data acquired by our biologist collaborators to provide them with a fast and accurate image segmentation pipeline for their upcoming studies. The training data was annotated using: a) the well-established StarDist2D in each slice, followed by b) through-slice linear assignment particle “tracking” (based on in-plane distance) to generate 3D segmentations. Both after a) and b), the results were manually checked to guarantee accurate training labels. The number of layers, kernel sizes, and pooling sizes in StarDist3D were carefully selected to ensure an adequate network field-of-view for the cells. The network was trained on the CSD3 cluster for 30 hours. Accurate 3D cell segmentation is now achievable using our trained network with minimal, supervised post processing (15-20 minutes) which is a huge improvement over manual segmentation (several days).

HyperTrack: Neural Combinatorics for High Energy Physics

Mikael Mieskolainen, Department of Physics, Imperial College London

Combinatorial inverse problems in high energy physics span enormous algorithmic challenges. This work presents a new deep learning driven clustering algorithm that utilizes a space-time non-local trainable graph constructor, a graph neural network, and a set transformer. The model is trained with loss functions at the graph node, edge and object level, including contrastive learning and meta-supervision. The algorithm can be applied to problems such as charged particle tracking, calorimetry, pile-up discrimination, jet physics, and beyond.

Flash talks

We showcase the effectiveness of this cutting-edge AI approach through particle tracking simulations. The code is available at github.com/mieskolainen/hypertrack [arXiv:2309.14113, CHEP 2023].

Employing AI to identify the complex interactions of environmental stressors on pollinator health

Rachel Parkinson, Department of Biology, University of Oxford

There is a critical need to investigate how environmental change affects pollinator behaviour so that steps can be taken to mitigate economic and ecological risk. Sound is a typically overlooked component of behaviour despite most animals producing sounds, and it is not known whether sound alone can be used to classify behaviours. I have constructed a raspberry pi-based recording arena for high-throughput behavioural data acquisition from insects, and am developing a multimodal machine learning algorithm that uses sound and video to automatically track behaviour. I am using this system to quantify the effects of environmental stressors, including pesticides and climatic temperature extremes.

A multiscale generative model unveils disorder in domain boundaries

Jiadong Dan, NUS Centre for Bioimaging Sciences, National University of Singapore

In the realm of atomic resolution microscopy, a critical challenge is the identification of significant structural motifs in synthesized materials, especially with limited observational data.

Flash talks

Addressing this, we introduce a novel hybrid generative model, adept at predicting unseen domain boundaries in potassium sodium niobate (KNN) thin films. This model requires minimal observations and circumvents the need for costly first-principles calculations or extensive atomistic simulations. Our findings are twofold: firstly, we demonstrate that domain boundary structures ranging from 1 to 100 nanometers emerge from straightforward, yet probabilistic, local rules. Secondly, we uncover tileable boundary motifs previously unseen, shedding light on their potential influence on the piezoelectric properties of these materials. Furthermore, our model indicates a propensity for creating domain boundaries with maximal configurational entropy. Significantly, our research illustrates that straightforward and interpretable machine learning models can be instrumental in deciphering disorder in complex materials, thereby laying a foundation for future advancements in functional materials design.

Image-based AI diagnosis platform for early drought stress detection in plant leaves

Alice Malivert, Department of Bioengineering, Imperial College London

Drought stress, a lack of accessible water caused by reduced precipitation, salinity, wind and extreme temperatures, is the primary cause of crop loss worldwide and is expected to become more frequent and severe with climate change: currently, drought stress is responsible for over 34% of agricultural production loss in least developed and developing countries. To reduce the drought stress-related issues in agricultural production by selecting appropriate varieties and management techniques, we need to be able to detect the first signs of drought stress in plants.

Flash talks

For that, researchers and farmers alike need a tool that would be fast, simple to use and cost-efficient while providing accurate quantitative measures. I propose to develop an AI assisted tool to detect early signs of drought stress in plant leaf pictures. Ultimately, this project will result in an open online platform to diagnose drought stress in new plant leaf pictures, as an affordable and accessible tool shared with the research and agricultural community.

Towards accelerated, experimental-theoretical closed loop chemical discovery

Austin Mroz, Department of Chemistry, Imperial College London

Novel chemical systems are necessary to fully address the major global challenges facing humanity, including the climate emergency, resource scarcity, and energy consumption needs. Traditional chemical discovery initiatives are founded on intuition-guided, “trial-and-error" processes. Here, small, iterative changes to chemical structure and experimental conditions are made by the researcher.1 This is significantly resource and time intensive; after each small modification, the molecule must be synthesized (often a trial-and-error process in itself) and properties measured. As a result, these workflows are associated with long timescales (~20 years) and high costs.2,3 Recently, computation has accelerated this process via atomistic simulations and data-driven approaches.4,5 The full utility of computation in the chemical discovery pipeline is only realized with the close integration of experiment and theory.6 This could be achieved via closed-loop, experimental-theoretical workflows, which are poised to efficiently identify high-performing candidate compounds for target applications. Here, we use data-driven optimization to guide high-throughout, automated experiments.

Flash talks

Despite the initial success of AI in closed-loop chemical discovery workflows, there still exist major challenges in their scalable implementation – the major bottlenecks being i) the degree of human intervention needed and ii) lack of methods to manage the number and quality of initial data points. This is further complicated by the necessary integration and accommodation of i) varying metadata formats, ii) non-compatible, proprietary characterization software, and iii) hands-on robotic platform calibration and manipulation, among others. Each of which needs to be explicitly addressed to realize coherent, automated closed-loop chemical discovery. Thus, data-driven solutions and supporting software are imperative to seamlessly close-the-loop in experimental-theoretical discovery workflows.

We address these bottlenecks and present an integrated, experimental-theoretical workflow that leverages high-throughput, automated experiments, and abstract computational models with data-driven optimization strategies to drive towards viable supramolecular materials for gas storage and separation applications.

References

[1] J. Am. Chem. Soc., 2022, 144, 18730; [2] Acc. Chem. Res., 2020, 53, 599; [3] Energy Environ. Sci., 2022, 15, 579; [4] Chem. Rev., 2020, 120, 8066; [5] Chem. Rev., 2021, 121, 9816; [6] Adv. Mater, 2021, 33, 2004831.

Posters

Using Supervised Machine Learning Algorithms to Explore Relationships of Gut Hormone Levels to BMI in Healthy and Obese Volunteers

Chris Bannon, Wellcome-MRC Institute of Metabolic Science, University of Cambridge

The gut hormones GLP-1, PYY and GIP have received extensive research, owing to their involvement in appetite regulation and glucose metabolism. Previous studies have shown PYY to have an inverse relationship to BMI, and insulin to be correlated to BMI due to the effects of increased adiposity on insulin resistance. For this study a database was made available with fasting and post meal gut hormones (GLP-1, PYY, GIP) and pancreatic hormones (glucagon and insulin) levels on 205 volunteers with obesity (>30), 29 volunteers with a BMI between 25 and 29, and 15 volunteers with a BMI below 25. Supervised learning algorithms including linear discriminant analysis and random forest classification were used to explore possible predictors of classification of BMI, and different regression techniques used to try predict BMI based on hormone and demographic information. Feature importance results supported previously reported data of PYY and BMI being key predictors of BMI and obesity.

Posters

AI processing of light-sheet microscopy images at exascale

Matthew Archer, University Information Services, University of Cambridge

Light-sheet microscopy is an attractive imaging technique as it provides non-destructive, high-resolution imaging of biological samples, enabling, for instance, cell development to be pictured in time. The high resolution 3D+time data presents a data challenge as resulting images can exceed the petabyte scale. Therefore data is often downsampled to make it more amenable to image processing tasks. In this project, we explore how to engineer a scalable AI and image processing pipeline that can process high-resolution light-sheet microscopy data by utilising the next generation computing hardware, such as the Exascale Test Bed, hosted by CSD3.

Assessing the performance of ChatGPT and Bard against junior uroradiologist expertise on PI-RADS classification

Kang-Lung Lee, Dimitri Kesseler, Tristan Barrett
Department of Radiology, University of Cambridge

Purpose or Learning Objective: Large language models (LLM) have sparked a wave of enthusiasm recently. ChatGPT and Bard are commonly used LLMs and may have some potential in the clinical radiology pipeline, including formulating conclusion remarks. This study aims to compare the classification abilities of ChatGPT, Bard, and a junior uroradiologist in assigning PI-RADS categories based on clinical text reports.

Methods or Background: Clinical prostate MRI text reports from 50 consecutive treatment-naïve patients who underwent mpMRI between 25/11/2022 to 28/12/2022 were included in this study. Clinical history and conclusion remarks were removed from the text reports. Two uroradiologists with 14 and 3 years of prostate MRI reporting experience independently classified PI-RADS categories on the edited text reports. The same reports were inputted manually into online ChatGPT-3.5 and Bard platforms to generate PI-RADS classifications. The classifications of the more experienced reader were deemed definitive. Comparisons were conducted among the classifications of the senior reader to the junior reader, ChatGPT, and Bard. Agreement rates were analysed.

Results or Findings: The senior radiologist assigned 30 reports (60%) as PI-RADS 2, and 3 (6%) as PI-RADS 3, 9 (18%) as PI-RADS 4, 8 (16%) as PI-RADS 5, respectively. The senior and junior radiologists concurred on classifications for 47 reports (94%). Compared to the senior radiologist's classifications, ChatGPT and Bard aligned on 35 (70%) and 32 (64%) reports, respectively. Notably, Bard assigned a PI-RADS 6 classification, not existing in the PI-RADS 2.1, to two patients (4%).

Conclusion: Compared to junior uroradiologist’s PI-RADS classifications, the performance of ChatGPT and Bard demonstrate unsatisfactory performance in the PI-RADS classification task.

Limitations: The limitations of the study are that this is a retrospective study, and only 50 reports were included.

Posters

Bayesian optimization to Accelerate the discovery of molecules for OPV application: Implementation of a fragment-based approach

Mohammed Azzouzi, Department of Chemistry, Imperial College London

We need more efficient, and application tailored organic photovoltaic devices. Exploring the vast chemical space of potential candidates is challenging and impossible with traditional methods. AI methods can help explore this large chemical space. Combining a fragment based approach, Bayesian optimization methods, Expert guidance and deep learning of molecular representation we present a novel way of exploring the chemical space.

Causality between built environment and subjective wellbeing: Integrated application of statistical and machine learning methods

Jerry Chen, Department of Land Economy, University of Cambridge

We apply an integrated framework combining difference-in-differences and synthetic controls ensemble with causal forest to the UK Household Longitudinal Survey. Household relocations are leveraged as natural experiments to quantify the causal relationship between the built environment and subjective wellbeing. We demonstrate the complementarity and interoperability of canonical statistics and novel machine learning methods.

Posters

Understanding and predicting past, present, and future coral reef distribution via multimodal machine learning

Orlando Timmerman, Department of Earth Sciences, University of Cambridge

Coral reefs are complex systems of animal-plant symbiosis on which millions of people rely for food, protection from coastal storms, and income from tourism. Tropical coral species – and the biodiversity they support – are threatened with functional extinction over the coming decades due to coastal pollution, mechanical damage, and sustained ocean temperature rises and ocean acidification driven by anthropogenic greenhouse gas emissions.

Robust, quantitative methods are necessary to direct resource-intensive conservation efforts to areas where future environmental conditions will be most conducive to long-term coral growth, and to inform more radical conservation methods such as assisted migration. Prediction of future environmental suitability first requires an understanding of how historic environmental conditions have resulted in the present-day distribution of coral reef systems.

This work explores the application of several multimodal machine learning methods to predict the present-day distribution of coral using historic environmental data on the Great Barrier Reef. It builds on comparable literature by implementing more sophisticated machine learning models, and by increasing the spatial and temporal resolutions of input data to scales more relevant to ongoing conservation initiatives. This results in greater predictive performance than previous methods.

Posters

Future work will apply these machine learning methods to the output of global climate models to determine forecasted environmental suitability. Emphasising the interpretability and explainability of these methods, the project aims to assist in the optimisation of current coral conservation initiatives, offering a vital contribution to the preservation of the biodiversity and ecosystem services of tropical coral reefs.