DiS

MPHIL

Complete List of Projects for DiS MPhil 2023/24

Contents

1 Deep Convolutional Neural Network for Inverse Problems in Imaging 3

2 Learned Primal-dual Reconstruction 4

3 Relativistic effects in astrometry with Gaia 5

4 The population of merging compact objects observed in gravitational waves 7

5 Application of ring finding techniques to simulated LHCb RICH data 9

6 Visualisation of Materials Energy Landscapes 10

7 Lüscher Zeta function project 11

8 Interpretation of features in chest X-ray images using deep learning 14

9 Application of machine learning to image processing in super-resolution structured illumination microscopy 16

10 Solving quantum lattice many-body systems using convex optimization 19

11 Search for the χc1(3872) meson and measurement of its properties 20

12 Do not underestimate the force: probabilistic machine learning for identifying forcing functions 21

13 Injecting probabilistic machine learning with a healthy dose of physics 23

14 Where has all the physics gone? : Identifying probable physics with probabilistic machine learning 25

15 Classification task, with a twist, for assigning cell-type annotations in single-cell sequencing datasets 27

16 Robustness and reproducibility of non-linear dimensionality reduction, with applications to single-cell sequencing assays 29

17 Methodological aspects of fits of parton distributions 31

18 Discovering the Higgs Boson 33

19 Pathomic Fusion: an interpretable attention-based framework thats integrates genomics and imaging to predict cancer outcomes 34

20 Chaotic amplitude control for the Ising minimisation using optical parametric oscillator system 36

21 Detecting a dominant subnetwork in gene network and protein-protein interaction data 38

22 Improving the resolution of clinical MRI images for detection and diagnosis. 39

23 Fostering gravitational-wave multi-messenger astronomy. 41

24 Learning dominant physical processes with data-driven balance models 43

25 Remapping WMAP 45

26 Baryonic physics - a missing link to small-scale cosmology 47

27 Feature Sensitivity and Model Discrimination in Preclinical Breast Cancer Photoacoustic Imaging 49

28 UNet-based segmentation of kidney cancer in Computed Tomography images 51

29 Forecasting solar and stellar cycles using probabilistic machine learning 52

30 Chemo-dynamical analysis of Milky Way's stellar populations with unsupervised multi-dimensional clustering 54

31 Lagrangian Neural Networks 56

32 Cross-Modal Pre-Training for Astronomical Foundation Models 58

33 Symbolic Distillation of Neural Networks 60

34 How reliably can we measure the mass of an exoplanet just half the mass of Venus? 62

1 Deep Convolutional Neural Network for Inverse Problems in Imaging

Proposer Name	Dr Ander Biguri
Proposer Role and Affiliation	Research Associate (DAMTP)
Proposer Contact Email	ab2860cam.ac.uk
Key Publication	ArXiv Paper 1611.03679

Project Description

Computed Tomography (CT) is an inverse problem that requires high amount of data to be able to produce clinically satisfactory results using standard algorithms, in particular filtered Backprojection (FBP) which is the most commonly used one. Mathematicians have been producing sophisticated algorithms to improve reconstruction quality, in order to be able to reduce dose to the patient, without sacrificing image quality. With the rise of data driven methods, various methods to include machine learning into the CT reconstruction problem have been proposed, and among others, the seminal work of the FBPConvNet algorithm was proposed, where a limited angle CT scan can be mapped to a high quality CT reconstruction This project aims to reproduce the work in the article. The student will learn about inverse problems, mathematics, CT reconstruction, and state of the art methods for machine learning inverse problems solutions.

In this work, we expect the students to reproduce the results, and then further the, either by providing more experimental examples (e.g. different levels of noise, or different level of sparse/limited angle ranges), or by modifying the network and exploring its effects.

Reading List

1. Medical Imaging module on the Mphil
2. This book covers most of the needed theory
3. This is a link to the documentation for the Tomosipo package.
4. This is a link to the code for the Tomosipo package.

Data Access

1. Here is a link to the Data

2 Learned Primal-dual Reconstruction

Proposer Name	Dr Ander Biguri
Proposer Role and Affiliation	Research Associate (DAMTP)
Proposer Contact Email	ab2860cam.ac.uk
Key Publication	ArXiv Paper 1707.06474

Project Description

Computed Tomography (CT) is an inverse problem that requires high amount of data to be able to produce clinically satisfactory results using standard algorithms, in particular filtered Backprojection (FBP) which is the most commonly used one. Mathematicians have been producing sophisticated algorithms to improve reconstruction quality, in order to be able to reduce dose to the patient, without sacrificing image quality. Among others the Primal-Dual algorithm exist. With the rise of data driven methods, various methods to include machine learning into the CT reconstruction problem have been proposed, and among others, the seminal work of the Learned Primal Dual (LPD) algorithm was proposed, where instead of letting the algorithm learn the entire reconstruction, the information of the physics is given to this "iterative unrolled" network. This project aims to reproduce the work in the article. The student will learn about inverse problems, mathematics, CT reconstruction, and state of the art methods for machine learning inverse problems solutions.

Reading List

1. Medical Imaging module on the Mphil
2. This book covers most of the needed theory
3. This is a paper that describes the methods used.
4. This is a link to the code for the Tomosipo package.

Data Access

1. Here is a link to the Data

3 Relativistic effects in astrometry with Gaia

Proposer Name	Dr Christopher Moore
Proposer Role and Affiliation	Associate Professor (IoA + DAMTP)
Proposer Contact Email	c.j.moorebham.ac.uk
Key Publication	Gaia Early Data Release 3: Acceleration of the Solar System from Gaia astrometry

Project Description

With the high quality and quantity of the astrometric data in Gaia EDR3 (and now DR3 as well) it is possible to detect small relativistic effects in the astrometric solutions for distant, extragalactic objects. It is well known that the apparent position of a distant star depends on the observer's velocity; this effect is called aberration. The solar system is moving (at a few hundred km/s) with respect the celestial reference frame. But this velocity is not constant; as the solar system orbits in the potential of the Milky Way, the acceleration towards the galactic centre is 1 cm/s/yr. Therefore, in the few years Gaia has been flying, the velocity of the solar system has changed by a few cm/s and the aberration from this appears as a dipole ("electric" type) in the velocity field of the proper motions of extragalactic objects directed towards the galactic centre. It's remarkable that a relativistic effect associated with such a small change in velocity can be detected! This project will use vector spherical harmonics to flexibly model the proper motion vector field of distant quasars using a sum of low-order vector spherical harmonics. The focus will initially be on the Gaia EDR3 data as we aim to reporoduce the main results from the key publication. - The key result to be reproduced is the measurement of the acceleration (both magnitude and direction) of the solar system barycentre (with respect to the rest frame of the Universe). The project can be extended by

1. looking at other datasets such as the more recent Gaia DR3 or VLBI radio data
2. developing improved, Bayesian methods for the handling of outliers.

More ambitiously, the project might also be extended by looking for (or placing upper limits on) other relativistic effects in the astrometric data. Prominent among these are ultra-low-frequency gravitational waves. These effects exist in the quadrupole sector of the VSH decomposition (as opposed to the dipole sector for the acceleration). The main other way of detecting nanohertz GWs is by pulsar timing arrays. This is very timely; in July 2023 all the major international PTA collaborations (NANOGrav, European PTA, Parkes PTA, and Chinese PTA) announced early evidence for what is probably a stochastic background of GWs generated by inspiralling supermassive black holes. For general interest, see here

Reading List

1. Titov, Lambert & Gontier (2011) "VLBI measurement of the secular aberration drift", A&A Volume 529, A91, link

2. Xu et al. (2012) ”Reconsidering the International Celestial Reference System based on the effect of the secular aberration”, IAU Joint Discussion 7: Space-Time Reference Systems for Future Research at IAU General Assembly-Beijing, link,

3. Titov (2013) ”THE SECULAR ABERRATION DRIFT AND FUTURE CHALLENGES FOR VLBI ASTROMETRY ”, link Truebenbach & Darling (2017) ”The VLBA Extragalactic Proper Motion Catalog and a Measurement of the Secular Aberration Drift”, ApJS 233 3, link

4. Charlot et al. (2020) ”The third realization of the International Celestial Reference Frame by very long baseline interferometry”, A&A, Volume 644, A159, link

Data Access

The main data set for this project will be Gaia data release 3 (DR3). link (You may need to create a free account in order to download large datasets.)

4 The population of merging compact objects observed in gravitational waves

Proposer Name	Dr Isobel Romero-Shaw & Dr Christopher Moore
Proposer Role and Affiliation	Herchel Smith Research Fellow (DAMTP) & Associate Professor (IoA + DAMTP)
Proposer Contact Email	isobel.romeroshaw@gmail.com + c.j.moorebham.ac.uk
Key Publication	The population of merging compact binaries inferred using gravitational waves through GWTC-3

Project Description

This project will study the population properties of compact binary mergers inferred from gravitational-wave (GW) observations from the first three LIGO-Virgo observing runs. This has now reached a total of 76 confidently detected mergers detected below a false alarm rate of 1 per year, reported in the GWTC-3 catalog. This includes binary black holes (BHs), binary neutron stars (NSs), and NS-BH binaries. This project will aim to reproduce some of the most high-profile population-level results from this new field. Some of the key results that could be reproduced are: - the compact object mass distribution (e.g. Figure 2 of the key publication) - the total local merger rate densities (e.g. Table 2 of the key publication) - the cosmological evolution of the population with redshift (e.g. Figure 13 of the key publication) The project will use the publicly-available individual-event posterior samples that are publically available through the GW Open Science Center (GWOSC) [reference]. These are used to define the population-level likelihood function (a.k.a. hyperlikelihood) for the parameters describing the astrophysical population, accounting for the very significant selection effects present in GW astronomy. (For details on this hyperlikelihood, see Ref. Mandel, Gair & Farr (2019) below and for one possible implementation see the BILBY inference software library [https://lscsoft.docs.ligo.org/bilby/].) This likelihood is then be sampled as part of a hierarchical Bayesian analysis. There is scope for significant extension to this project after the initial results are reproduced, for example, through the use of alternative statistical methods for the hierarchical Bayesian inference for the population(possibly including Dirichlet processes or autoregressive processes) or alternative analysis pipelines for the Bayesian parameter estimation at the individual event level.

Reading List

1. LIGO & Virgo collaborations (2021) ”Population properties of compact objects from the second LIGO-Virgo Gravitational-Wave Transient Catalog”, ApJL 913 L7, link

2. LIGO & Virgo collaborations (2019) ”Binary Black Hole Population Properties Inferred from the First and Second Observing Runs of Advanced LIGO and Advanced Virgo”, ApJL 882 L24, https://arxiv.org/abs/1811.12940 Mandel, Farr &

Gair (2019) ”Extracting distribution parameters from multiple uncertain observations with selection biases” MNRAS, 481, 1, 1086-1093, link

3. Ashton et al. (2019) “BILBY: A User-friendly Bayesian Inference Library for Gravitational-wave Astronomy“ ApJS 241 27, link

4. Romero-Shaw et al. (2020) “Bayesian inference for compact binary coalescences with BILBY: validation and application to the first LIGO-Virgo gravitational-wave transient catalogue” MNRAS 499 3, link

Data Access

The data release associated with the Key Paper for this project: link

5 Application of ring finding techniques to simulated LHCb RICH data

Proposer Name	Dr Chris Jones
Proposer Role and Affiliation	Principal Research Associate, High Energy Physics, Cavendish Laboratory
Proposer Contact Email	crj1001@cam.ac.uk
Key Publication	Elastic net for stand-alone RICH ring finding

Project Description

The project will start by first developing a toy simulation of events as recorded by the LHCb RICH detectors. This will use the known radiator gas parameters and detector optics to generate toy Cherenkov rings with size and hit distributions reproducing that in real data as recorded during LHC run2. Once this the framework for generating this toy data has been produced, the student will implement a trackless ring finding approach, the elastic neural network described in the key publication and run this algorithm on the toy data. Performance metrics on how well the algorithm can find and classify the rings will be presented. Once this step is completed the student will then attempt to apply commonly used machine learning techniques to the data and compare the performance to the above elastic neural network.

Reading List

1. Performance of the LHCb RICH detector at the LHC
2. Performance of the LHCb muon system
3. Elastic net for standalone RICH ring finding

Data Access

The project will be based on data simulated by the student so no external data is required.

6 Visualisation of Materials Energy Landscapes

Proposer Name	Prof Chris Pickard
Proposer Role and Affiliation	Sir Alan Cottrell Professor of Materials Science, MSM
Proposer Contact Email	cjp20@cam.ac.uk
Key Publication	Visualizing Energy Landscapes through Manifold Learning

Project Description

Dimensionallity reduction is a key task in the data intensive sciences, and is particularly useful to assist in the visualisation of the increasingly large datasets that high throughput computational methods generate. Materials structure prediction is one of these tasks, and Ab Inito Random Structure Searching generates many putative structures, ranked by quantum mechanical total energy. Stochastic Hyperspace Embedding And Projection (SHEAP) is a dimensionality reduction method designed for visualising these potential energy surfaces. To explore these landscapes efficiently, it is important to understand their topologies. However, they exist in spaces with very large numbers of dimensions, making them difficult to visualise. SHEAP uses dimensionality reduction through manifold learning to effectively visualise the distribution of stable structures across a high-dimensional energy landscape. The aim of this project is to reimplement the SHEAP algorithm, and compare it with the established manifold learning algorithms t-SNE and UMAP for the test datasets described in the key publication. Although designed for materials datasets, SHEAP should be assessed as a technique for general dimensionality reduction in the data intenstive sciences

Reading List

1. Pickard, Chris J., and R. J. Needs. "Ab initio random structure searching." Journal of Physics: Condensed Matter 23, no. 5 (2011): 053201.
2. McInnes, Leland, John Healy, and James Melville. "UMAP: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426 (2018).
3. Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9, no. 11 (2008).

Data Access

Several test datasets are included in the SHEAP bitbucket repository. Additionally: AIRSS data for carbon at 10GPa and the C+N+H+O system at 1GPa Rules of formation of H–C–N–O compounds at high pressure and the fates of planetary ices

7 Lüscher Zeta function project

Proposer Name	Dr David Wilson
Proposer Role and Affiliation	Royal Society Research Fellow, DAMTP
Proposer Contact Email	dw504@cam.ac.uk
Key Publication	Two particle states on a torus and their relation to the scattering matrix and the implementation of its calculation here

Project Description

The Lüscher Zeta function is used to obtain two-hadron scattering information from Lattice Quantum Chromodynamics (QCD). In Lattice QCD studies, correlation functions from the strongly-coupled theory of QCD can be computed in a finite, periodic spatial volume, resulting in a discrete spectrum of energies. In a finite, periodic volume, momentum is quantized such that $\vec{p} = \frac{2\pi}{L}\vec{n}$ for a single hadron, where $\vec{n}$ is a triplet of integers. For two-hadron states, in the absence of interactions, energies are expected at

$$E = (m_1^2 + \vec{p}_1^{\,2})^{1/2} + (m_2^2 + \vec{p}_2^{\,2})^{1/2}\,. \tag{1}$$

This enables a spectrum to be determined by enumerating various low combinations of $\vec{n}$. An example is shown for zero total momentum ($\vec{P} = \vec{p}_1 + \vec{p}_2 = \vec{0}$) in Fig. 1.

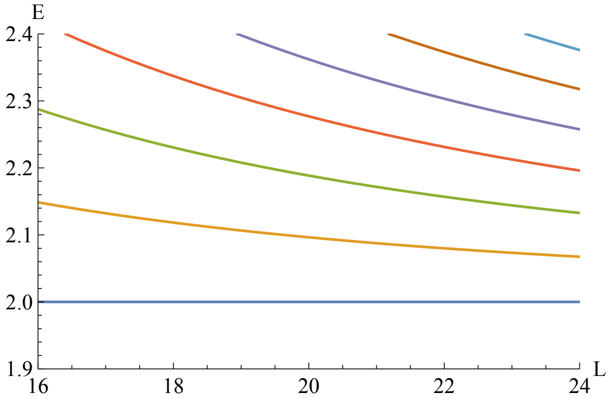

Figure 1: Non-interacting energy levels obtained using Eq. 1 using a few low values of $\vec{n}$.

When we consider an interacting system, the energies obtained are shifted away from the values given by Eq. 1. For weakly attractive or repulsive interactions, the differences are small. Weakly attractive systems produce energies slightly shifted down from Eq. 1; weak repulsion results in small shifts upwards. Resonances can also arise (unstable hadrons that decay to the two-hadron scattering system). These result in an "extra" level and/or an avoided level crossing.

The Lüscher equations considered in this project quantify the correspondence between finite volume energy levels and infinite volume scattering amplitudes. A practical example of this method at work can be found here, for the case of the ρ resonance in $\pi\pi$ scattering.

Project

The aim is to implement various forms of the Lüscher Zeta function and to optimize it. We will first consider the scattering of two hadrons with equal masses $m_1 = m_2$, with zero total momentum, $\vec{P} = 0$. Various extensions exist such as to moving frames (non zero $\vec{P}$) and for unequal hadron-hadron masses. These can be attempted as extensions.

Most sources define the most basic Lüscher Zeta function ($\ell = 0$, $m = 0$) as

$$Z_{00}(s, q^2) = \sum_{\vec{n} \in \mathbb{Z}_3}^{\infty} \frac{1}{(\vec{n}^2 - q^2)^s} \tag{2}$$

where $\vec{q} = \frac{L}{2\pi}\vec{p}$ and $q^2 = \vec{q}.\vec{q}$, where $\vec{p}$ is the corresponding momentum of the finite volume energy state determined from $E = 2(\vec{p}^2 + m^2)^{1/2}$. Only the magnitude of $\vec{p}$ is meaningful. The sum in Eq. 2 is convergent for Re $s > 3/2$. We are interested in the case of $s = 1$, which can be obtained by analytic continuation.

Steps for the project:

1. Code up Eq. 2 in a simple way in the language of your choice (C++/Python preferred).

2. Check that the result is robust with respect to any choices made (for example the number of terms included in the sum or the number of points used in an integral approximation).

3. Extend your code by either generalising it to:

   - $\ell m \neq 00$, $m_1 \neq m_2$ as in Scattering phase shifts for two particles of different mass and non-zero total momentum in lattice QCD
   - $\vec{P} \neq 0$ as in Resonance Scattering Phase Shifts on a Non-Rest Frame Lattice

   or optimise the code to improve the speed of evaluation. For example:

   - How many terms in the sum are necessary to achieve a desired precision?
   - Can the computed precision be quantified efficiently? These things may not be constant as a function of q^2 (and ℓm, m_1, m_2, and $\vec{P}$).
   - Can symmetry be used to reduce the evaluation time?
   - Does parallelism help? Can interpolation help? Think about single evaluations and multiple evaluations at similar q^2 values.
   - Can you find any transformations or alternative representations that are faster in some way?

Reading List

1. If you are interested in the derivation of Eq. 2 from field theory, see Finite-Volume Effects for Two-Hadron States in Moving Frames.

2. For extension to coupled-channel scattering see: Multiple-channel generalization of Lellouch-Luscher formula, Moving Multi-Channel Systems in a Finite Volume with Application to Proton-Proton Fusion, and The coupled-channel scattering on a torus.

3. Extensions to hadrons with spin see Two-particle multichannel systems in a finite volume with arbitrary spin

4. A nice, accessible review of the overall approach and applications can be found in Scattering processes and resonances from lattice QCD

Data Access

For code evaluation and testing, the following data can be used:

- Table A1 of Signatures of unstable particles in finite volume

- Extracting Scattering Phase-Shifts in Higher Partial-Waves from Lattice QCD Calculations contains many figures that can be used for checking and comparison.

8 Interpretation of features in chest X-ray images using deep learning

Proposer Name	Eduardo Gonzalez Solares
Proposer Role and Affiliation	Senior Research Associate, Institute of Astronomy
Proposer Contact Email	eg226@cam.ac.uk
Key Publication	CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison

Project Description

Chest radiography is the most common imaging examination globally, critical for screening, diagnosis, and management of many life threatening diseases. Automated chest radiograph interpretation at the level of practicing radiologists could provide substantial benefit in many medical settings, from improved workflow prioritization and clinical decision support to large-scale screening and global population health initiatives. CheXpert (Chext eXpert) is a large dataset that contains 224,316 chest radiographs of 65,240 patients. The project will focus in using deep learning with the aim to use the CheXpert dataset to train a deep learning model that is able to predict the probability of 14 observations from X-ray images and compare the accuracy with radiologists. It will be useful for the student to have some knowledge of neural networks, Python and either TensorFlow or PyTorch as well as access to a GPU during model traininng.

Key results:

1. Design and train a model on the X-ray images to predict the probability of each of the observations, i.e., given a X-ray chest image determine the most likely pathology.

During the project the student will

1. Become familiar with the particularities associated with medical imaging and X-rays in particular, like biases due to image annotations, data leakage, data augmentation issues or working with limited samples.

2. Learn how the above greatly impact the design of the training, validation and test sets.

3. Experiment with different ways of measuring success as well as loss functions in order to improve the accuracy either for the full dataset or for a subsample

Useful extensions to the project:

1. Investigate the use of techniques like test time augmentation and ensemble models to improve results.

2. Create a web app that allows a user to upload a x-ray image and produce the probability of each observation plus the saliency map.

Reading List

1. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning
2. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Data Access

Data access is available from the key project web page linked above and could be made readily available to the student if necessary somewhere else (HPC or local custer).

9 Application of machine learning to image processing in super-resolution structured illumination microscopy

Proposer Name	Edward Ward
Proposer Role and Affiliation	Research Associate, Dept. Chemical Engineering and Biotechnology
Proposer Contact Email	ew535@cam.ac.uk
Key Publication	Li, X., Wu, Y., Su, Y. et al. Three-dimensional structured illumination microscopy with enhanced axial resolution. Nat Biotechnol (2023)

Project Description

Optical microscopy is a vital tool in biological research, allowing us to view the fundamental processes of life in living systems. Within this field, super-resolution microscopy overcomes the classical diffraction limit of light and enables processes to be probed at the molecular level, revealing the exact mechanisms of disease. Despite their strengths, super-resolution techniques suffer from the drawback that the image reconstruction process can take many hours and the final results are prone to artefacts. This is especially true when imaging live biological samples where the high frame rates needed to capture dynamic events produce large volumes of noisy image data. After reconstruction, this high noise level leads to unreliable experimental results and determining what is a real feature of the sample can become challenging. One of the most popular super-resolution methods is structured illumination microscopy (SIM). Here, a resolution increase is achieved by illuminating samples with patterned light and extracting high resolution information from the interference patterns generated. Recently, Li et al. have developed a method where machine learning (ML) is used to rapidly the process data from a SIM system before the super-resolution reconstruction is performed. They demonstrate that this is preprocessing significantly improves reconstructions and hence the reliability of the images. However, they limit this method to data generated on their own custom system, making it hard to validate the technique on image data from other researchers and limiting the usefulness of the method. This project would first seek to replicate their key results:

1. Demonstrating that the ML processing allows for reliable image reconstructions.
2. Demonstrating that the ML processing improves upon existing methods in the field.
3. Validating the method on data from other research groups.

As well as this validation, there is also huge scope for further development. In the simplest case, this could involve making implementations of the method more accessible to the wider research community. For example, incorporating it into the most commonly used software environments or optimising it to run on lower specification computers. Going forward, there is also significant improvement that can be made on the underlying architecture or training data used. For example, making use of more state-of-the-art networks to encode temporal information into the processing or teaching the networks to compensate for a broader range of imaging artefacts. If successful, this work would find immediate use, both within the host group and for our collaborators, such as the Molecular Neurosciences Group. No background in the biological sciences is expected and this project would be an excellent opportunity for students wishing to gain experience in machine learning for high-throughput image processing. Those interested should feel free to contact Dr Edward Ward for more information or background on the project.

Reading List

The following papers provide a useful source of information on super-resoltuion structured-illumination-microscopy and state-of-the-art applications of ML to structured illumination microscopy:

1. Christian Karras C et al. Successful optimization of reconstruction parameters in structured illumination microscopy – A practical guide, Optics Communications, 436:690-75 (2019)

2. Christensen CN et al. ML-SIM: universal reconstruction of structured illumination microscopy images using transfer learning. Biomed Opt Express. 15:12(5) p2720-2733 (2021)

3. Christensen CN et al. Spatio-temporal Vision Transformer for super-resolution microscopy. arXiv preprint arXiv:2203.00030 (2022).

4. Liang J et al. Swinir: Image restoration using swin transformer. arXiv preprint arXiv:2108.10257, 1833-1844, (2021)

5. Jin, L et al. Deep learning enables structured illumination microscopy with low light levels and enhanced speed. Nat Commun 11:1934 (2020)

Data Access

The dataset used by Li et al. for the ML processing is available through the publication. Additional datasets of SIM data can be found through numerous publications reporting on the applications of SIM. For example:

1. Deep learning enables structured illumination microscopy with low light levels and enhanced speed

2. Machine learning assisted interferometric structured illumination microscopy for dynamic biological imaging

3. A Guide to Structured Illumination TIRF Microscopy at High Speed with Multiple Colors

Additionally, large volumes of data are available from the host lab and our collaborators and artificial data can generated readily on the University's High Performance Cluster (HPC) using in-house software. Implementations of the ML networks are available through the code repositories associated with the source publications.

10 Solving quantum lattice many-body systems using convex optimization

Proposer Name	Prof Hamza Fawzi
Proposer Role and Affiliation	Professor of Applied Mathematics, DAMTP
Proposer Contact Email	hf323@cam.ac.uk
Key Publication	Barthel, Hübener, Solving Condensed-Matter Ground-State Problems by Semidefinite Relaxations, PRL 2012

Project Description

The goal of the project is to use tools from numerical convex optimization, namely semidefinite programming, to solve lattice many-body models from condensed matter physics, such as the Fermi-Hubbard quantum mechanical model. Concretely, the goal of the project is to reproduce the numerical results and figures of the cited papers for the Fermi-Hubbard model, and possibly to extend and improve them (I have some ideas for this). The ideas used in the cited papers have been applied in many different settings in theoretical physics under the name "bootstrap" method (see reading list below). This method has attracted significant attention in the last few years as it provides rigorous results that can be more accurate than Monte Carlo-based methods.

Reading List

For background on semidefinite programming and convex optimization duality, we recommend the following survey:

1. Boyd and Vandenberghe, Semidefinite Programming, SIAM Review, 38(1): 49-95, 1996 (https://web.stanford.edu/ boyd/papers/sdp.html)

The approach based on semidefinite programming relaxations (a.k.a. the bootstrap method, or the sum-of-squares/moment method) is fairly straightforward to grasp and can be understood directly from the main publications cited above. Another short paper which can be useful to read in parallel is:

1. Han, Quantum Many-body Boostrap

More resources can be provided if needed.

Data Access

The obtained numerical results should be compared to the benchmark data in the following paper: Solutions of the 2D Hubbard Model: Benchmarks and Results from a Wide Range of Numerical Algorithms, PRX 2015

11 Search for the χc1(3872) meson and measurement of its properties

Proposer Name	Dr Harry Cliff
Proposer Role and Affiliation	Research Associate, High Energy Physics, Cavendish Laboratory
Proposer Contact Email	cliff@hep.phy.cam.ac.uk
Key Publication	LHCb Collaboration, Determination of the X(3872) Meson Quantum Numbers, PRL 110, 222001 (2013)

Project Description

A search for the χc1(3872) meson (also known as the X(3872)) will be performed using open access data recorded by the LHCb experiment at the Large Hadron Collider in 2011 and 2012. The data sample corresponds to a sample of proton-proton collisions with a total integrated luminosity of 3 fb-1. The χc1(3872) is an exotic hadron whose nature has not yet been conclusively deftermined, and may be either a tetraquark or meson molecule. Determination of its properties, in particular its quantum numbers (total angular momentum J, partity P and charge-conjugation C) can shed light on this mystery. The particle will be searched for in the decay B+→ χc1(3872) $K+$, where the χc1(3872) decays to a J/Ψ meson and two charged pions ($\pi + \pi -$). The higher yield B+ → Ψ(2S) K+ decay will be used as a control channel. Machine learning algorithms will be trained to classify signal and background and optimised to give the maximum possible significance. A search will also be performed for the less frequently produced Ψ(3823) meson. Following observation of the particle, its mass and the ratio of the B+ → χc1(3872) K+ branching fraction with respect to the control channel will be measured using a maximum likelihood fit. Finally (and if time allows), the quantum numbers of of the χc1(3872) will be determined using maximum likelihood fits to the 5D angular distribution of the decay. As no LHCb simulated data are available on an open access basis, all aspects of the analysis will necessarily be carried out using the open access collision data alone.

Reading List

1. Belle Collaboration, Observation of a Narrow Charmoniumlike State in Exclusive $B\pm \rightarrow K \pm \pi + \pi - J/\Psi$ Decays, Phys.Rev.Lett.91:262001,2003.
2. Nicola Mangiafave, Jeremy Dickens, Valerie Gibson, A Study of the Angular Properties of the $X(3872) \rightarrow J/\Psi\pi + \pi-$ Decay, LHCb-PUB-2010-003
3. LHCb Collaboration, Study of the Ψ2(3823) and χc1(3872) states in $B+ \rightarrow (J\Psi\pi + \pi-)K+$ decays, JHEP 08 (2020) 123.

Data Access

12 Do not underestimate the force: probabilistic machine learning for identifying forcing functions

Proposer Name	Dr Henry Moss
Proposer Role and Affiliation	Early Career Advanced Fellow, DAMTP
Proposer Contact Email	hm493@cam.ac.uk
Key Publication	Alvarez, Mauricio, David Luengo, and Neil D. Lawrence. ”Latent force models.” Artificial Intelligence and Statistics. PMLR, 2009.

Project Description

In this project, we will reproduce an important paper on early physics-informed machine learning. The authors propose a way to use Gaussian processes to learn the (unknown) forcing function acting on a physical system from experimental data. This method has been critical for many biological (gene expression) and engineering (stress prediction) problems but has yet to be updated to benefit from recent advances in Gaussian processes. We will start by recreating the experimental results presented in the original paper (i.e. Figures 1,2,3 and Tables 1,2). If time permits, we will investigate novel extensions to this work by:

1. including recent developments from the GP literature:
   (a) investigate more complex physical using GPs suitable for large and high-dimensional datasets
   (b) new kernel approximations to improve the flexibility/applicability of the latent force model.
2. incorporating ideas from the equation discovery literature for scenarios where we want to learn an analytical approximation to the forcing function.
3.

Another ideal outcome from this project would be a robust code implementation that can be contributed to our open-source GP library written in GPJax

Reading List

For a theoretical introduction to GPs:

1. Williams, Christopher KI, and Carl Edward Rasmussen. Gaussian processes for machine learning. Vol. 2. No. 3. Cambridge, MA: MIT press, 2006.

For a practical introduction to GPs (with code):

1. GPJax documentation

For more advanced extensions/applications of the key paper:

1. Guarnizo, C., and Alvarez Lopez. ”Fast kernel approximations for latent force models and convolved multiple-output Gaussian processes.” Uncertainty in Artificial Intelligence: Proceedings of the Thirty-Fourth Conference (2018). AUAI Press, 2018.

2. Ward, Wil, et al. ”Black-box inference for non-linear latent force models.” International Conference on Artificial Intelligence and Statistics. PMLR, 2020.

3. Alvarez, Mauricio A., David Luengo, and Neil D. Lawrence. ”Linear latent force models using Gaussian processes.” IEEE transactions on pattern analysis and machine intelligence 35.11 (2013): 2693-2705.

4. Rogers, T. J., K. Worden, and E. J. Cross. ”On the application of Gaussian process latent force models for joint input-state-parameter estimation: With a view to Bayesian operational identification.” Mechanical Systems and Signal Processing 140 (2020): 106580.

5. Hensman, James, Nicolò Fusi, and Neil D. Lawrence. ”Gaussian processes for Big data.” Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence. 2013

Data Access

We will be reproducing the results from the key publication. The two “toy” problems are based on simulated data that we can easily re-simulate ourselves. The other two “real” problems on motion capture and metal concentration are freely available from Ranked prediction of p53 targets using hidden variable dynamic modeling, CMU Graphics Lab Motion Capture Database and Jura data set, respectively.

13 Injecting probabilistic machine learning with a healthy dose of physics

Proposer Name	Dr Henry Moss
Proposer Role and Affiliation	Early Career Advanced Fellow, DAMTP
Proposer Contact Email	hm493@cam.ac.uk
Key Publication	Long, Da, et al. ”AutoIP: A united framework to integrate physics into Gaussian processes.” International Conference on Machine Learning. PMLR, 2022.

Project Description

In this project, we will reproduce a recent key paper on physics-informed machine learning. The authors propose a clever way to incorporate known underlying physics into previously purely data-driven Gaussian processes regression models, allowing improved predictions and uncertainty quantification when modeling physical systems. We will start by recreating the experimental results presented in Section 5 of this paper (i.e. Figures 1,2,3 and Tables 1,2,3) to redemonstrate the power of including physical knowledge in probabilistic machine learning algorithms. If time permits, we will investigate novel extensions to this work by

1. including recent developments from the GP literature:
   (a) new efficient sampling strategies to allow the inclusion of more flexible physical information
   (b) approximate kernel calculations to improve the flexibility/applicability of this method to more challenging datasets
2. investigate the more realistic scenario where the underlying physics is only partially understood and must be partially identified by ML.

Another ideal outcome from this project would be a robust code implementation that can be contributed to our open-source GP library written in GPJax

Reading List

For a theoretical introduction to GPs:

1. Williams, Christopher KI, and Carl Edward Rasmussen. Gaussian processes for machine learning. Vol. 2. No. 3. Cambridge, MA: MIT press, 2006.

For a practical introduction to GPs (with code):

1. GPJax documentation

For more advanced extensions/applications of the key paper:

1. Hensman, James, Nicolò Fusi, and Neil D. Lawrence. "Gaussian processes for Big data." Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence. 2013.
2. Hensman, James, Alexander Matthews, and Zoubin Ghahramani. "Scalable variational Gaussian process classification." Artificial Intelligence and Statistics. PMLR, 2015.
3. Karniadakis, George Em, et al. "Physics-informed machine learning." Nature Reviews Physics 3.6 (2021): 422-440.
4. Also, see the related works of the key publication named above.

Data Access

We will be reproducing the results from the key publication. There are three "real" problems based on gene expression, on motion capture and metal concentration and are freely available from CMU Graphics Lab Motion Capture Database and Jura data set, respectively.

14 Where has all the physics gone? : Identifying probable physics with probabilistic machine learning

Proposer Name	Dr Henry Moss
Proposer Role and Affiliation	Early Career Advanced Fellow, DAMTP
Proposer Contact Email	hm493@cam.ac.uk
Key Publication	Hidden Physics Models: Machine Learning of Nonlinear Partial Differential Equations or Numerical Gaussian Processes for Time-Dependent and Nonlinear Partial Differential Equations

Project Description

In this project, we will reproduce one of two important paper on physics-informed machine learning for the low-data regime, e.g. as arises from expensive real-world experiments. The authors propose a way to allow Gaussian processes to leverage underlying physical laws to extract patterns from high-dimensional but scarce datasets. The authors consider learning from complex data that is sparse in space (paper 2) or sparse in time (paper 1). The intention of the project is that one of these data regimes is tackled. We will start by recreating the experimental results presented in the original paper (i.e. Figures 1,2,3,4, 5). If time permits, we will investigate novel extensions to this work by:

1. leveraging recent advances in scalable Gaussian processes:
   (a) to allow the learning across many time-steps
   (b) to allow the analysis of higher-dimensional measurements (e.g. climate systems)
2. identify optimal data collection strategies (temporal and spatial)
3. investigate the more realistic scenario where underlying physics are only partially understood

Another ideal outcome from this project would be a robust code implementation that can be contributed to our open-source GP library written in GPJax.

Reading List

For a theoretical introduction to GPs:

1. Williams, Christopher KI, and Carl Edward Rasmussen. Gaussian processes for machine learning. Vol. 2. No. 3. Cambridge, MA: MIT press, 2006.

For a practical introduction to GPs (with code):

1. GPJax documentation

For more advanced extensions/applications of the key paper:

1. Raissi, Maziar, and George Em Karniadakis. "Hidden physics models: Machine learning of nonlinear partial differential equations." Journal of Computational Physics 357 (2018): 125-141.
2. Raissi, Maziar, Paris Perdikaris, and George Em Karniadakis. "Numerical Gaussian processes for time-dependent and nonlinear partial differential equations." SIAM Journal on Scientific Computing 40.1 (2018): A172-A198.
3. Raissi, Maziar. "Deep hidden physics models: Deep learning of nonlinear partial differential equations." The Journal of Machine Learning Research 19.1 (2018): 932-955.
4. Hensman, James, Nicolò Fusi, and Neil D. Lawrence. "Gaussian processes for Big data." Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence. 2013.

Data Access

We will be reproducing the results from the key publication. All the problems are based on simulated data that we can easily re-simulate ourselves, or download from the author's repository here.

15 Classification task, with a twist, for assigning cell-type annotations in single-cell sequencing datasets

Proposer Name	Dr Irina Mohorianu
Proposer Role and Affiliation	Head of Bioinformatics/ Scientific Computing, Cambridge Stem Cell Institute
Proposer Contact Email	iim22@cam.ac.uk
Key Publication(s)	Supervised classification enables rapid annotation of cell atlases, Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling, Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data

Project Description

A crucial step in analysis of single-cell RNAseq data is cell-type annotation. It may be performed manually, in an iterative approach, or using automated or semi-automated methods. The task may occur prior or post clustering of cells (based on expression levels); for the former it may inform the partition of cells; for the latter, it relies on previously identified marker genes, per cell type or condition. A disadvantage of relying exclusively on unsupervised (clustering-derived) partitions as starting point stems from the observed instability of partitioning pipelines e.g. due to random seed, number of selected genes, number of neighbours considered for the graph-based clustering, and clustering method itself. Disadvantages of relying exclusively on markers derive from incomplete information available for cell annotations. In this project, you will [a] evaluate the robustness of inferred annotations using state-of-the-art methods, CellAssign, scType and Garnett and compare and contrast the underlying methods and resulting assignments, [b] combine several supervised approaches discussed previously, in [a], to create a stacked (new) method that combines their advantages. The ensemble/ stack method will be tested with the aim of increasing the sensitivity of boundary detection. The robustness will be assessed using Element Centric Similarity, as described in the ClustAssess pipeline (https://github.com/Core-Bioinformatics/ClustAssess)

Reading List

1. scmap: projection of single-cell RNA-seq data across data sets
2. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing
3. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data

Data Access

The data used for this project comprises two datasets, one for smartSeq sequencing (Cuomo et al 2020,) and one for 10x sequencing (Gribben et al 2023, under revision in Nature; full access to the data and metadata is agreed). The data is publicly available for download on the Gene Expression Omnibus; the expression matrix, which will be the starting point for the project, will be provided alongside all, brief yet necessary, details for understanding the characteristics of the data. The smartSeq data comprises 1800 columns (cells) x 10k rows (genes); the 10x data comprises 20k cells (columns) x 5k expressed genes per cell. The former has 20% missing values (zeros); the latter has 70% missing values (zeros). Both sequencing approaches are current state-of-the-art for both exploratory and in-depth analyses.

16 Robustness and reproducibility of non-linear dimensionality reduction, with applications to single-cell sequencing assays

Proposer Name	Dr Irina Mohorianu
Proposer Role and Affiliation	Head of Bioinformatics/ Scientific Computing, Cambridge Stem Cell Institute
Proposer Contact Email	iim22@cam.ac.uk
Key Publication(s)	Dimensionality reduction for visualizing single-cell data using UMAP, Initialization is critical for preserving global data structure in both t-SNE and UMAP, ClustAssess: tools for assessing the robustness of single-cell clustering, The specious art of single-cell genomics

Project Description

UMAP is a non-linear dimensionality reduction that aims to represent large matrices, with varying proportions of missing values, in a low dimensional space (usually 2 or 3D) by keeping both the global and the local structure and similarities between observations; it was proposed as an alternative to tSNEs (t-distributed stochastic neighbour embedding) which only preserved some local structures, completely distorting conclusions at global, dataset, scale. UMAP representations, usually calculated on a linear dimensionality reduction (PCA) are the state of the art for visualising single-cell dataset. Both tSNE and UMAP are stochastic methods; results differ across runs e.g. when the random seed is changed; acceptable changes include rotations and translations. However more notable changes, that propagate through analyses, are also observed, leading to unstable, not-reproducible outputs. ClustAssess, developed in the Core Bioinformatics group, evaluates the impact of the random seed; the Kobak et al 2021 paper underlines the impact of the initialisation. The Chari et al 2023 shows downstream effects of the intrinsic stochasticity of the methods. The latter also proposes alternative approaches to maximise reproducibility. The project is focused on [a] reproducing the assessement of the UMAP dimensionality reduction with respect to Element Centric Similarity (as described in the ClustAssess manuscript), [b] reproducing the conclusions from the Chari et al manuscript in terms of replacing the L2 norm with the L1 metric for assessing similarity, and evaluating also the generic Lk norm, [c] assessing the modularity of the data, ”hubness”, and integrating it into the embedding. All tasks will be applied on the data extensively presented in the ClustAssess tutorials and documentation. The functions created for reproducing the conclusions from the Chari paper will be included in a new release of the ClustAssess pipeline.

Reading List

1. ClustAssess Git repository

Data Access

The data used for this project comprises two datasets, one for smartSeq sequencing (Cuomo et al 2020,) and one for 10x sequencing (Mende et al 2022). The data is publicly available for download on the Gene Expression Omnibus; the expression matrix, which will be the starting point for the project, will be provided alongside all, brief yet necessary, details for understanding the characteristics of the data. The smartSeq data comprises 1800 columns (cells) x 10k rows (genes); the 10x data comprises 20k cells (columns) x 5k expressed genes per cell. The former has 20% missing values (zeros); the latter has 70% missing values (zeros). Both sequencing approaches are current state-of-the-art for both exploratory and in-depth analyses.

17 Methodological aspects of fits of parton distributions

Proposer Name	Dr James Moore
Proposer Role and Affiliation	Research Associate, DAMTP
Proposer Contact Email	james.moore@damtp.cam.ac.uk
Key Publication	On the determination of uncertainties in parton densities

Project Description

Description of the project and list of key results to be reproduced. Parton distributions are the functions which parametrise the structure of protons in terms of their elementary constituents (quarks and gluons); they are usually determined using precise data from collider experiments like the LHC at CERN. Fits of these functions use a range of methods, including various techniques for uncertainty propagation, and various assumed functional forms for the functions themselves (including neural network parametrisations). This project will study, in a toy scenario, the faithfulness of these different uncertainty propagation methods, together with the effect of using different functional forms. A successful project will first demonstrate a basic understanding of parton distributions and their use in collider physics (a purely intuitive understanding of their use is fine - a detailed understanding of the quantum chromodynamics background is unnecessary). A successful project will then:

1. Investigate, in a toy model, various methods used for uncertainty propagation from experimental data onto parton distribution functions, as described in 2206.10782. This will include a reproduction of Figure 2 from this paper.

2. Investigate, in the same toy model, the effect of using a power law functional form for the parton distributions vs using a neural network functional form for the parton distributions, as described in the same reference. This will include a reproduction of Figures 5 and 8 in the given reference.

3. If time permits, further investigation may be carried out beyond the scope of the paper, including the effect of using toy data modelled by parton distributions entering quadratically in Eq. (28) rather than linearly.

Reading List

1. The primary reference is On the determination of uncertainties in parton densities. This focusses specifically on toy models for parton distribution functions, and the successful candidate need not read beyond this paper.

2. For interested candidates, a recent realistic fit of parton distributions using neural networks is given in The path to proton structure at one-percent accuracy.

Data Access

This project will rely on the candidate generating their own pseudo-data according to Eq.(28) of 2206.10782 (and possibly extending this pseudo-data generation to the quadratic case).

18 Discovering the Higgs Boson

Proposer Name	Matt Kenzie
Proposer Role and Affiliation	Associate Professor, High Energy Physics, Cavendish Laboratory
Proposer Contact Email	mk652@cam.ac.uk
Key Publication	ATLAS Collaboration, Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Phys Lett B 716 (2012) 1, CMS Collaboration, Observation of a New Boson at a Mass of 125 GeV with the CMS experiment at the LHC, Phys Lett B 716 (2012) 30

Project Description

Can you find the Higgs boson with the dataset that was used by the ATLAS and CMS experiments to discover it? You will exploit CERN's Open Access Data platform to study data recorded at the LHC in 2011 and 2012. The Higgs boson will be searched for in data containing Higgs candidate decays to two photons and two Z-boson (which both decay to pairs of leptons). Machine learning techniques will be required to isolate signal candidates from the large amounts of backgrounds. The invariant mass of selected signal candidates will be computed in order to determine if a statistically significant signal can be found by performing a fit to the invariant mass of candidates in the region of 125 GeV. If a signal is found, possible extensions could include studies of the Higgs boson properties. The project is deigned to be carried out by multiple individuals, with each working on a specific decay mode of the Higgs using data from one experiment. If necessary additional data is available from the 2016, 2017 and 2018 runs of the LHC to aid with an unambiguous discovery.

Reading List

1. Status of Higgs Boson Physics
2. Example of physics analysis: the case of SM Higgs boson production in the H → ZZ decay channel in the four-lepton final state
3. Example of physics analysis: the case of SM Higgs boson production in the H → yy decay channel in the two-photon final state
4. About CMS Page

Data Access

Documentation for accessing the data is here: here

19 Pathomic Fusion: an interpretable attention-based framework thats integrates genomics and imaging to predict cancer outcomes

Proposer Name	Dr Mireia Crispin-Ortuzar
Proposer Role and Affiliation	Assistant Professor, Department of Oncology, University of Cambridge
Proposer Contact Email	mc973@cam.ac.uk
Key Publication	Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis

Project Description

This project aims to reproduce the findings presented in the paper ”Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis.” The study provided the foundation for a major 2022 publication (Lipkova et al, Cancer Cell 2022). The paper introduces a novel end-to-end multimodal fusion approach that integrates histopathological image data and genomic markers for the purpose of improving cancer diagnosis and prognosis. Central to the method is a Kronecker product-based mechanism that models pairwise feature interactions across the modalities, coupled with a gating-based attention system to control the expressiveness of each representation. The framework is one of the few examples of deep-learning based multi-modal fusion models including imaging that have been used to answer clinically relevant questions.

1. Preliminary Research [essential] - Conduct a comprehensive review and understanding of the original paper, focusing on the proposed methods, data representation, and fusion mechanisms.

2. Environment Setup [essential] - Establish the computational environment to train the models. Install all required software libraries and dependencies.

3. Data Acquisition [essential] - Obtain the glioma and clear cell renal cell carcinoma datasets from the Cancer Genome Atlas (TCGA) as used in the original study. Obtain the annotations from a previous publication.

4. Model implementation and training [desirable] - Analyse and test the implementation of the proposed deep learning framework. Re-train the models using the training data.

5. Training and Validation [essential] - Validate the trained model’s performance on the test dataset.

6. Interpretability and Documentation [desirable] - Implement the interpretability framework and analyse feature importance and localization across each modality.

Reading List

1. Artificial intelligence for multimodal data integration in oncology
2. Integrating context for superior cancer prognosis
3. Pan-cancer integrative histology-genomic analysis via multimodal deep learning
4. Pan-cancer image-based detection of clinically actionable genetic alterations
5. Predicting cancer outcomes from histology and genomics using convolutional networks

Blog with useful tutorials is here

Data Access

All the details of how to reproduce the results are explained meticulously in a Github repository: https://github.com/mahmoodlab/PathomicFusion

1. The raw data is in the TCGA database: here
2. The annotations can be exported from a previous publication: here
3. Scripts for training and testing models are on Github: here
4. Processed data and trained models are on Google Drive: here

20 Chaotic amplitude control for the Ising minimisation using optical parametric oscillator system

Proposer Name	Natalia Berloff
Proposer Role and Affiliation	Professor of Applied Mathematics, DAMTP
Proposer Contact Email	ngb23@cam.ac.uk
Key Publication	Destabilization of local minima in analog spin systems by correction of amplitude heterogeneity.

Project Description

This paper argues that analogue bistable systems such as the system of optical oscillators, even when simulated on a classical computer, can find low energy states of the Ising Hamiltonian at least as efficiently as current state-of-the-art heuristics. The project consists in implementing the algorithm with amplitude corrections proposed in the key publication and replicating solutions of the combinatorial optimization problems MAX-CUT on the G sets [1] while comparing the performance of this algorithm with that of the BLS [2,3] which itself outperforms other heuristics. See Destabilization of Local Minima in Analog Spin Systems by Correction of Amplitude Heterogeneity for the additional details on the simulation scheme. A potential extention of this project is to use other heuristics for banchmarking of, e.g. Microsoft analog iterative machine [4], simulated annealing or genetic algorithms.

Reading List

1 G sets

2 U. Benlic and J.-K. Hao, Engineering Applications of Artificial Intelligence 26, 1162 (2013).

3 S. Reifenstein, et al "Coherent SAT solvers: a tutorial," Adv. Opt. Photon. 15, 385-441 (2023)

4 K.P.Kalinin et al "Analog Iterative Machine (AIM): using light to solve quadratic optimization problems with mixed variables," arXiv:2304.12594 (2023)

Data Access

[4] The benchmark dataset Gset can be downloaded from G sets. Gset consists of the problems G1 to G81 and, due to some skipped numbers, has a total of 71 problems with a graph size ranging from 800 to 20,000 vertices. There are graphs without weighted edges (all weights are 1) as well as graphs with weighted edges where the weights are either

+1 or -1. Geometrically, the structure of the graphs can be split into three categories (1) Random graphs, (2) Planar graphs and (3) Toroidal graphs. The students should experiment graphs from the representative sets. Each file in the G-set directory has the following structure: the first row lists the number of vertices N and edges E, while following rows list the edges one edge per row. Each such row contains the edge's endpoint vertices i and j as well as its weight w_{ij}.

21 Detecting a dominant subnetwork in gene network and protein-protein interaction data

Proposer Name	Natalia Berloff
Proposer Role and Affiliation	Professor of Applied Mathematics, DAMTP
Proposer Contact Email	ngb23@cam.ac.uk
Key Publication	A Weighted Maximum Clique Method for Identifying Condition-Specific Sub-Network.

Project Description

This key publication reformiulates the sub-network identification problem as a constrained optimization problem for continuous variables. It is an approximation of the general combinatorial problem, based on the theorem posed by Motzkin and Straus [1]. The project aims at implementing the proposed algorithm – the weighted MAXimum clique (WMAXC) method to identify a condition-specific sub-network and replicate WMAXC method on simulated data described in the key publication. The results (original and subnetwork) should be visualised using Cytoscape [2] and compared with the conclusions of the key publication. The possible extention is to use latest state-of-the-art combinatorial optimisation solvers instead of the genetic algorithms used in the key publication and/or use real protein-protein interaction from Spring [4] or other databases.

Reading List

1 Motzkin S, Straus G (1965) Maxima for graphs and a new proof of a theorem of turan. Canadian Journal of Mathematics 17: 533-540.

2 Cytoscape

3 Guo, Zheng, et al. "Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network." Bioinformatics 23.16 (2007): 2121-2128.

4 String Database

Data Access

Use the simulated data described in the section "Simulation studies" of the key publication. The Spring database [4] contains protein network data.

22 Improving the resolution of clinical MRI images for detection and diagnosis.

Proposer Name	Dr Priscilla Canizares
Proposer Role and Affiliation	Senior Research Scientist (Daphne Jackson Fellow) at The Alan Turing Institute and Cambridge University (DAMTP) / Research Associate at Cambridge University (Radiology).
Proposer Contact Email	pc464@cam.ac.uk
Key Publication	Sparse Bayesian pMRI reconstruction with complex Bernoulli-Laplace mixture priors.

Project Description

Magnetic resonance imaging (MRI) is an essential medical imaging tool that uses strong magnetic fields and radio waves to produce detailed images of the inside of the body, such as the brain and spinal cord, the heart and blood vessels, and internal organs. The results of an MRI scan help in the early detection of conditions, such as cancer or heart problems, plan treatments, and assess how effective previous treatment has been. During an MRI scan, the patient must be as still as possible to obtain valuable and informative images. However, MRI scanners take a long time to acquire data, making it challenging for children and persons who cannot breathe due to pre-existing conditions. Scanners with several receiver coils can perform parallel Magnetic Resonance Imaging (pMRI) to accelerate data acquisition and improve MRI images' spatial and temporal resolutions. However, the most challenging task is the full field reconstruction without noise, distortions and artefacts. This project employs a Hierarchical Bayesian model to regularise the pMRI data and exploits sparsity-promoting priors in the image space without any transformation, then, allowing for enhanced images and further speed-up. The key results to be reproduced are (i) the restored MRI scans and (ii) the structural similarity (SSIM) values and the signal-to-noise ratio (SNR) indexes for the different priors proposed. If time allows, this project can be extended by using compressed sensing methods on the MRI image and comparing the resulting images with the ones obtained with the proposed method.

Reading List

1. SENSE: Sensitivity encoding for fast MRI". Magnetic Resonance in Medicine
2. On Bayesian classification with Laplace priors", Pattern Recognition Letters
3. Sparse Signal Recovery with Temporally Correlated Source Vectors Using Sparse Bayesian Learning
4. CS Image reconstruction example: here
5. The Berkeley Advanced Reconstruction Toolbox, BART

Data Access

Dataset: he data set is open-access and available through Kaggle.

23 Fostering gravitational-wave multi-messenger astronomy.

Proposer Name	Dr Priscilla Canizares
Proposer Role and Affiliation	Senior Research Scientist (Daphne Jackson Fellow) at The Alan Turing Institute and Cambridge University (DAMTP) / Research Associate at Cambridge University (Radiology).
Proposer Contact Email	pc464@cam.ac.uk
Key Publication	Detection of gravitational-wave signals from binary neutron star mergers using machine learning

Project Description

In 2017, the Laser Interferometer Gravitational-Wave Observatory (LIGO) and the NASA/ESA Hubble Space Telescope detected a gravitational wave (GW) and the first optical (electromagnetic) counterpart. This historic detection suggested that both signals were emitted by the same event, the merger of two neutron stars, starting the era of GW multi-messenger astronomy. GW multi-messenger astronomy is an exciting new field of research. It relies on the simultaneous detection and joint analysis of signals (messengers) of different nature but emitted by the same GW source, like electromagnetic radiation and/or neutrinos. These messengers, which are produced by different processes within the system, carry complementary physical information, and their joint detection can lead to a better characterisation and understanding of their astrophysical sources. When a GW event occurs, the GW signal is matched (cross-correlated) with a bank of templates to determine whether the system contains a neutron star and, if positive, alert the observational telescopes to search for the (expected) electromagnetic counterpart. Hence, to maximise the observation time, the detection time is crucial. During this project, you will use data from the first GW multi-messenger event and GW models from binary neutron star systems. The data and the models will be matched using classical and deep learning (DL) techniques to determine whether the detected GW signal contains a neutron star and alert the optical telescopes. The key results to be reproduced are the false alarm rate (FAR) and sensitivity of classical vs DL approaches and their corresponding latency time. If time allows, this project can be extended by (i) using a more elaborated GW model than the one employed in the main reference or (ii) exploring and comparing alternative DL architectures.

Reading List

1. The code associated with the key publication listed above can be used to obtain the multirate sampling algorithm employed in the main reference: here
2. The GW templates can be obtained from: Templates
3. GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral

4. First Multimessenger Observations of a Neutron Star Merger

Data Access

Data and data catalogs:

1. Data release for event GW170817
2. Catalog of Observed Gravitational-wave Mergers

24 Learning dominant physical processes with data-driven balance models

Proposer Name	Rich Kerswell
Proposer Role and Affiliation	Professor of Fluid Mechanics
Proposer Contact Email	rrk26@cam.ac.uk
Key Publication	Learning dominant physical processes with data-driven balance models

Project Description

The project revolves around the idea of identifying different spatial regions where specific dynamical balances exist. This is a clustering challenge where the relative magnitudes of all the dynamical terms which appear in the governing equations are considered over space and time. A cluster corresponds to where the same subset of terms are significant and 'balance'. The key publication explores the idea for a number of different examples with the boundary layer situation of probably most interest (e.g. Fig.1). Once this is reproduced/mastered there is the possibility of doing something novel with a polymer-laden flow where the governing equations are much more opaque (there is now a tensor equation describing the polymer configuration augmenting the usual (vector) Navier-Stokes equation). The data source for the later situation can be generated locally using a 2-dimensional code. We have extensive experience simulating what is called Elasto-inertial turbulence (turbulence which needs both inertia and elasticity provided by the polymer to survive and discovered relatively recently in 2013) and more data than we know what to do with. The challenge here would be to identify hitherto unknown balances of subsets of terms which should help our understanding of what is going on especially close to the boundaries where the turbulence seems based.

Reading List

The key paper has now been cited 25 times according to the web of science and that would be a good place to get supplementary reading/ideas.In terms of Elasto-Inertial Turbulence, key refs are

1. Samanta D. et al. 'Elasto-inertial Turbulence" Proc. Nat. Acad. Sciences, 110, 10557-10562, (2013).
2. Sid, S. et al. "Two-dimensional dynamics of elasto-inertial turbulence and its role in polymer drag reduction"
3. Dubief, Y. et al. "First coherent structure in elasto-inertial turbulence" Phys. Rev Fluids 7, 073301 (2022).

Data Access

The turbulent boundary layer data are openly available from the Johns Hopkins Turbulence Database43. Source code for simulating the GNLSE is available at here. Surface current estimates in the Gulf of Mexico are from the HYCOM + NCODA global 1/25°reanalysis (Expt. 50.1) available here. The detonation analog model was simulated with Clawpack44. Further information about the data sets and simulations are included in Supplementary Information. Data for the follow-up work will be generated locally (using Dedalus)

25 Remapping WMAP

Proposer Name	Dr Steven Gratton
Proposer Role and Affiliation	Senior Teaching Associate, DAMTP
Proposer Contact Email	stg20@cam.ac.uk
Key Publication	The WMAP temperature maps and associated papers

Project Description

NASA's WMAP satellite was launched in June 2001 and observed for nine years, producing maps of the microwave sky at frequencies of 23, 33, 41, 61 and 94 GHz. Much of the radiation observed was emitted when the Universe was only about four hundred thousand years old, a tiny fraction of the fourteen billion or so years old it is now. Such "Cosmic Microwave Background" (CMB) radiation is often described as "the afterglow of the Big Bang" and by studying its statistical properties and comparing them to model predictions we can learn about the Universe's composition, evolution and initial conditions. In this project you will reproduce the processing of the "time-ordered data" from the WMAP satellite into sky maps of the radiation. You will learn about the application of iterative data-processing techniques to cosmological datasets and the handling of data on spherical surfaces. Your required target is to reproduce to some level the "individual differencing assembly" first year/DR1 temperature-only maps from the calibrated DR1 time-ordered data using the simple map-making scheme presented in the first set of papers (in particular the data processing methods paper) and first explanatory supplement. You may then go on to make temperature and polarization maps using more data and more advanced algorithms as presented in the later releases. This could include attempting your own calibration, starting say from the DR5 uncalibrated time-ordered data. Extension work might involve investigating the performance of advanced map-making schemes, such as the "bilinear" method presented in arXiv:2210:02243, on WMAP data. You may also consider and/or develop an efficient GPU mapmaking implementation and compare its performance to a CPU-based one.

Reading List

1. See Sec. 2.2 of "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Data Processing Methods and Systematic Error Limits", available from here for a detailed description of iterative mapmaking. Sec. 2.5 gives some refinements it would be nice to incorpate.

2. The explanatory supplement, available from here , in particular Chapter 4, is important for understanding the time-ordered data.

3. The general LAMBDA page for WMAP: here, this contains links to all relevant papers, data products and target results for this project. Following the evolution of the techniques employed for the different releases may be instructive.

4. The full link for the bilinear mapmaking paper mentioned above is here

Data Access

The resouces for this project will come from the ”Legacy Archive for Microwave Background Data” (LAMBDA) website, given above. For example the first-year time-ordered data is found here You can use the ”IDL” code and documentation provided on LAMBDA as a reference for your own work (e.g. for computing the ”pointing” of the satellite’s detectors for each observation). You may find the ”healpy” python package for the handling and visualization of data on the sphere using the ”healpix” pixelization scheme useful. Documentation is here. The astropy.io.fits package might be helpful for handing FITS files in python. Documentation is here.

26 Baryonic physics - a missing link to small-scale cosmology

Proposer Name	Dr Vid Isric
Proposer Role and Affiliation	Senior Kavli Fellow; Kavli Institute for Cosmology, Department of Physics
Proposer Contact Email	vi223@cam.ac.uk
Key Publication	Inferring the impact of feedback on the matter distribution using the Sunyaev Zel‘dovich effect: insights from CAMELS simulations and ACT + DES data

Project Description

Modern observational cosmology has made large strides towards a standard model of cosmology in recent decades. This model predicts that matter is distributed in the Universe in a complex network of knots and filamentary structures called cosmic web. Around 85Recent observations of low-redshift Universe, however, are challenging this paradigm. In particular measurements of the amount of structure in the cosmic web from weak-gravitational lensing, and thermal Sunyaev-Zel'dovich effects shows a deficit of structure compared to the CDM model predictions from the cosmic microwave background. These exciting results could be a hint of new physics in the dark matter sector. However, the picture is complicated by the complexity of the role the ordinary matter plays in galaxy and structure formation. On small scales, the clustering of ordinary matter is expected to deviate from that of the dark matter through effects that galaxies have on the matter distribution in their surroundings. The aim of this project is to explore these effects in a massive set of cosmological simulations of galaxy formation, CAMELS. The simulations were run for a wide range of different cosmologies as well as galaxy formation models. In particular, this project will focus on the results reported in Panday et al. 2023, that suggests that measurements thermal Sunyaev-Zel'dovich (tSZ) effect correlates with the amount of structure on small-scales as induced by galaxy physics. The tSZ effect measure a projected gas pressure profiles in halos, and depends on the density and temperature of the gas around galaxies. Using the simulation data of IllustrisTNG the the student would reproduce certain figures from the Panday et al. 2023 (Fig. 1, Fig.3, top panel of Fig.4 and Fig.5). Using the relations between the tSZ and power spectrum suppression at different scales an interpolation model can be used in order to predict the power spectrum suppression for a given measured tSZ – using either random forest regression or neural network. This can reproduce Fig.6 in the Panday et al. 2023 paper, related to IlustrisTNG (blue points). Project details:

1. Work with IllustrisTNG set of CAMELS simulations

2. Access the power spectra for hydrodynamical simulations and the corresponding N-body simulations (LH-suite), in order to reproduce Fig.1 (blue shaded regions)

3. For each simulation in the hydrodynamical suite, read in halo catalogs in the range of masses from $5x10^12M_sun/h$ to $10^14M_sun/h$ and construct the halo mass function

(halo distribution). For each halo, read in the particle data from the simulations in order to measure tSZ and gas fraction for each halo. Integrating over the mass function of each simulation will result in an average tSZ and gas fraction values per simulation.

4. Reproduce Fig. 3, by colour coding simulations based on the value of cosmological or galaxy formation parameters and using the results you obtained above.

5. Using either gas fractions or tSZ, and integrating the mass function over different mass ranges can reproduce Fig.4 (top panel, related to IllustrisTNG)

6. Fixing the mass range of the halo mass function, and changing the pivot scale where the matter power spectra suppression is measured can reproduce Fig.5. This can be also done for tSZ in addition to the gas fraction reported in the paper.

7. Train a model (either random forest regression or neural network) to predict power spectrum suppression given a tSZ or gas fraction measurement. Reproduce Fig.6 (blue points).

Reading List

1. Villaescusa-Navarro F., et al., 2021, ApJ, 915, 71

2. Chisari N. E., et al., 2019, The Open Journal of Astrophysics, 2, 4

3. Breiman L., 2001, Machine Learning, 45, 5

4. Description of data organisation and data outputs Includes useful tutorials and python packages to navigate the data.

Data Access

The data used in this work will be based on CAMELS simulation project. There are various data access options, either through Binder (recommended) or downloading the data. More information on data access can be found here The data structure and organisation is explained in detail here. Related to this project, the most important data will be Simulations data, SUBFIND catalogs, and Power spectra. The basis of the project is built around IllustrisTNG simulations.

27 Feature Sensitivity and Model Discrimination in Preclinical Breast Cancer Photoacoustic Imaging

Proposer Name	Dr Lorena Escudero Sánchez
Proposer Role and Affiliation	Senior Research Associate, Department of Radiology
Proposer Contact Email	les44@cam.ac.uk
Key Publication	Feasibility and sensitivity study of radiomic features in photoacoustic imaging of patient-derived xenografts

Project Description

This work explores the use of a set of computer vision features used in medical imaging called Radiomics(1), extracted from a novel imaging modality called photoacoustic or optoacoustic (2), which is an increasingly popular method of exploring a tumour's microenvironment. These features account for the textural characteristics of the images, in addition to just the pixel values distributions.

The aim of this project is to test the sensitivity of such features to factors/parameters related to the acquisition and reconstruction of the photoacoustic images obtained from breast cancer patient-derived xenografts from two different models (basal and luminal B), implanted subcutaneously into mice. This projects pays therefore special attention to feature selection (dimensionality reduction) and robustness/sensitivity aspects of features/variables and machine learning methods based on them.

The available data contains the features already extracted from the images (therefore no direct manipulation of images is necessary), and the results to be reproduced are:

1. The sensitivity study presented in the paper with respect to acquisition and reconstruction parameters of each one of the features
2. The study of model discrimination potential of each feature through the results of a machine learning classifier.
3.

The dataset is small and presents a realistic example of real-world healthcare imaging dataset. In particular the second part (model discrimination) can be expanded, and the robustness/reliability of the approach and machine learning algorithm presented in the paper for this part (aiming only at a feasibility study) can be explored.

Reading List

1. Gillies, R., Kinahan, P. & Hricak, H. Radiomics: images are more than pictures, they are data. Radiology 278, 563-577, DOI: 10.1148/radiol.2015151169 (2015).
2. Steinberg, I. et al. Photoacoustic clinical imaging. Photoacoustics 14, 77-98, DOI:10.1016/j.pacs.2019.05.001 (2019)

Data Access

The necessary data (features already extracted from the photoacoustic images) are available here

28 UNet-based segmentation of kidney cancer in Computed Tomography images

Proposer Name	Dr Lorena Escudero Sánchez
Proposer Role and Affiliation	Senior Research Associate, Department of Radiology
Proposer Contact Email	les44@cam.ac.uk
Key Publication	An attempt at beating the 3D U-Net

Project Description

KiTS is a grand challenge to accelerate the development of reliable kidney and kidney tumor semantic segmentation methodologies. There have been three versions so far of this challenge, with increasing number of computed tomography (CT) cases released in 2019, 2021 and 2023. This project will use the released training dataset in the 2019 version,both images and manual segmentations/delienations in NIfTI format (210 cases). The aim of the project is to test three different versions of 3D UNet (with and without residual connections, and with pre-activation) and reproduce similar Dice scores (measuring overlap of the true and predicted segmentations) to those in the paper linked above, which was the winner of the KiTS19 challenge. The project involves handling and manipulating a realistic dataset of medical imaging in a standard format, basic data augmentation, network training (includes a significant amount of GPU hours) and evaluation of appropriate metrics. If time allows, further work can be done by extending the test of such algorithms to a larger dataset with the training sets for KiTS21 and KiTS23 (the latter including different acquisition phases, so expected to produce slightly different results).

Reading List

1. O. Ronneberger, P. Fischer and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, MICCAI 2015, doi: 10.1007/978-3-319-24574-4_28

2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424-432. Springer (2016)

Data Access

The necessary data is available here

29 Forecasting solar and stellar cycles using probabilistic machine learning

Proposer Name	Dr Lalitha Sairam
Proposer Role and Affiliation	Postdoctoral research associate (IoA)
Proposer Contact Email	lalitha.sairam@ast.cam.ac.uk
Key Publication	Forecasting stellar activity

Project Description

Stellar activity produces astrophysical noise signals of different amplitudes and timescales that can hamper both detection and lead to spurious atmospheric species. Despite improved techniques, stellar activity continues impacting exoplanet observations.

To solve this problem we need a robust forecast of optimal observing time based on stellar magnetic activity cycles. A key characteristic of our Sun is its 11 years activity cycle. Recent work shows that by observing the Sun as a star the detection threshold for Earth-like planets is higher during solar activity minima. Hence, the successful discovery of the Earth twin and its atmospheric characterisation would be possible with a more optimal observing strategy based on the activity minima forecast of stars.

The main goal of this project is to develop an ideal machine learning technique that will support a deeper understanding of the mechanisms involved in both solar and stellar activity to predict activity cycles. Firstly, we will use the Gaussian process, a machine learning technique to effectively captures the patterns and irregularities observed in the data, enabling the representation of the inherent variability in activity cycles. The project can be extended by developing Hilbert transformation on solar activity indicators to determine the termination event marking the end of the previous activity cycle and the enhancement in the current cycle. During the second phase of the project, we will apply these models to exoplanet host stars with well-determined activity cycles. These predictions of amplitude and timing of stellar activity cycles will deliver insights into their space weather systems and help us draw parallels to our weather on Earth.

Reading List

1. Forecasting stellar activity
2. Review on methods applied in Sunspot prediction
3. Sunspot cycle prediction using Warped Gaussian process regression
4. Hilbert transformation of solar cycle

Data Access

The data sets for this project will be the following daily sunspot number, Mount Wilson archive data and HARPS-N solar data from Dace

30 Chemo-dynamical analysis of Milky Way's stellar populations with unsupervised multi-dimensional clustering

Proposer Name	Dr GyuChul Myeong, Dr Anke Ardern-Arentsen
Proposer Role and Affiliation	Research Associate(s), Institute of Astronomy
Proposer Contact Email	gm564@cam.ac.uk, aa2437@cam.ac.uk
Key Publication	Milky Way's Eccentric Constitutents with Gaia, APOGEE, and GALAH

Project Description

Large galaxies, such as our Milky Way, go through complex evolutionary phases governed by various different mechanisms. As a result, the present day Galaxy is a composed product of stars from different origins and epochs, although it is not easy to distinguish one star's origin from another individually. Stars formed from the same process and the same source share comparable chemical and dynamical properties which reflects their formation environment. To understand the complex evolutionary history and underlying formation mechanisms of the Galaxy, it is essential to know the chemical and dynamical properties of the stars produced from each process. Using a combination of detailed chemical and dynamical information from Gaia, APOGEE, and GALAH, we attempt to classify the Milky Way stars into an unspecified number of sub-groups with distinguishable chemo-dynamical trends. Gaussian mixture modelling is adopted as an unsupervised clustering method. The chemo-dynamical property of each identified component can help us to trace the Galaxy's evolutionary history. In specific we focus on identifying and studying the Milky Way's known stellar populations, such as GS/E (ex-situ population from a major galactic merger event the Milky Way experienced), Aurora (ancient population from early epoch of Milky Way evolution), Splash (dynamically heated population due to the GS/E merger), and Eos (result of a star-formation as a consequence of GS/E merger).

Reading List

1. From dawn till disc: Milky Way's turbulent youth revealed by the APOGEE+Gaia data

2. Co-formation of the disc and the stellar halo

3. The biggest splash

4. Extreme deconvolution: Inferring complete distribution functions from noisy, hetrogeneous and incomplete observations

Data Access

The main datasets are publically accessable (Gaia DR3, APOGEE DR17, GALAH DR3), but some of the necessary information requires further computations (e.g., orbital integration, energy calculation) which is beyond the scope of the project. The student can contact the PI (GyuChul Myeong; gm564@cam.ac.uk) for the dataset that can be used for the project. The main software required for the project (Extreme Deconvolution) is publically available here. Alternatively, the student can try scikit-learn's Gaussian Mixture Modelling here.

31 Lagrangian Neural Networks

Proposer Name	Dr Miles Cranmer
Proposer Role and Affiliation	Associate Professor, DAMTP/IoA
Proposer Contact Email	mc2473@cam.ac.uk
Key Publication	Lagrangian Neural Networks

Project Description

This project aims to reproduce and further explore the concepts presented in the paper "Lagrangian Neural Networks". The paper presents a new class of neural networks that can learn arbitrary Lagrangians, performing well even in situations where canonical momenta are unknown or difficult to compute. The proposal challenges students to reproduce the experiments presented in the paper, critique the methods used, and explore potential improvements or extensions to the work.

Expected Learning Outcomes:

1. Understand neural differential equations and their potential applications to physics.
2. Understand how physical constraints such as the Euler-Lagrange equations can be imposed on both a learning and inference task.
3. Understand the difference between a "hard" and "soft" physical constraint(or inductive bias) for machine learning.
4. Understand the difficulties of training these models, and strategies for working with non-Gaussian gradients.
5. More generically, skills in critiquing and suggesting improvements to existing scientific works in machine learning.

Reading List

1. Hamiltonian Neural Networks
2. Neural Ordinary Differential Equations
3. David Tong's lecture notes on classical mechanics
4. https://karpathy.github.io/2019/04/25/recipe/

Data Access

The data used in the original paper is a set of very simple mechanics problems integrated with JAX. This dataset is available here. The code for the model is also available at this link, but the student should implement the model themselves (in whatever framework they wish). Though the original model was created in JAX because PyTorch lacked an efficient Jacobian implementation, this is no longer the case, and so PyTorch is a recommended option.

32 Cross-Modal Pre-Training for Astronomical Foundation Models

Proposer Name	Dr Miles Cranmer
Proposer Role and Affiliation	Associate Professor, DAMTP/IoA
Proposer Contact Email	mc2473@cam.ac.uk
Key Publication	AstroCLIP: Cross-Modal Pre-Training for Astronomical Foundation Models

Project Description

This project aims to reproduce some of the results in the paper "AstroCLIP: Cross-Modal Pre- Training for Astronomical Foundation Models" by Lanusse et al., 2023. This paper demonstrates a contrastive learning approach to finding an embedding space shared between galaxy images and spectra. The proposal challenges students to implement the multi-modal embedding in the paper with the contrastive learning framework, as well as the encoders. The student will become comfortable with astronomical datasets, and how to work with them in a machine learning context. The student will also develop intuition for some of the key ideas behind foundation models, and how to use them effectively for downstream tasks.

Expected Learning Outcomes:

1. Hands-on intuition-building experience with a fundamental concept in foundation models: contrastive learning.
2. Intuition for latent spaces, and how they relate to original physical variables.
3. Intuition for how one can use latent representations for tasks other than they were explicitly trained for, which will be increasingly important in the future.
4. Experience with astronomical datasets (which have many advantages to typical ML datasets: they are free, open, massive, and we have fairly good understanding of the underlying relations), and how to work with them in a machine learning context.
5. Skills in critiquing and suggesting improvements to existing scientific works in machine learning at the intersection of machine learning and astrophysics.

Reading List

Related papers:

1. Learning Transferable Visual Models From Natural Language Supervision
2. SimCLR: A Simple Framework for Contrastive Learning of Visual Representations

Supporting Material:

1. DESI Galaxy Survey
2. PROVABGS Catalog
3. Possibly astropy
4. Andrej Karpathy's Neural Network Training Recipe

Data Access

The data used in the paper is from the SDSS "DESI" catalog, crossmatched with the PROVABGS catalog. . The code and specific data links for the paper are given in the GitHub repository, as well as the paper.

33 Symbolic Distillation of Neural Networks

Proposer Name	Dr Miles Cranmer
Proposer Role and Affiliation	Associate Professor, DAMTP/IoA
Proposer Contact Email	mc2473@cam.ac.uk
Key Publication	Discovering Symbolic Models from Deep Learning with Inductive Biases

Project Description

This project aims to reproduce the main results from the paper "Discovering Symbolic Models from Deep Learning with Inductive Biases". The paper demonstrates a technique to translate neural networks into analytic equations, focusing on toy and complex physical systems for test cases with graph neural networks as the architecture, due to their functional similarity to physics. (You can think of this "symbolic distillation" as doing a Taylor expansion on the neural network.) This proposal challenges students to:

1. Implement a graph neural network with a single message passing step, and train and evaluate it on toy n-body systems.
2. Measure the latent representations inside the neural network, specifically the activations of the message functions.
3. Apply symbolic regression using PySR to approximate the message functions with analytic equations.
4. Explore the effect of a bottleneck versus a sparse regularization on a wide message passing space.
5. Evaluate if meaningful physical laws can be recovered, how those laws differ from the ones we are used to, and if those laws generalize better.

Via this process, students will improve their intuition for architectural choices in a neural network, the interpretations of latent variables, and interpretation in a physics-like context. Through this, students will gain intuition for imposing structure on neural networks, interpreting their internal representations, and combining neural networks with symbolic methods. The focus will be reproducing the main results on rediscovering physical laws.

Expected Learning Outcomes:

1. Practical experience working with geometric structures and inductive biases on neural networks.
2. Intuition for how to interpret and extract meaningful representations from neural networks, especially in a physics and model discovery context.
3. An understanding of the tradeoffs between symbolic and neural network-based machine learning approaches, and how they can intersect.

4. Skills in critiquing and suggesting improvements to existing scientific works in machine learning at the intersection of machine learning and astrophysics.

Reading List

Related papers:

1. Relational inductive biases, deep learning, and graph networks
2. Andrej Karpathy's Neural Network Training Recipe
3. PySR paper

Data Access

Simple classical mechanics simulations are provided in the paper code repo. The model exists in this repo as well, but students should re-implement the model from scratch based on the paper.

34 How reliably can we measure the mass of an exoplanet just half the mass of Venus?

Proposer Name	Dr Vinesh Maguire-Rajpaul
Proposer Role and Affiliation	Analyst (Isaac Newton Institute); Visiting Researcher (Cavendish Laboratory)
Proposer Contact Email	vr325@cam.ac.uk
Key Publication	Warm terrestrial planet with half the mass of Venus transiting a nearby star

Project Description

98-59 is a star known to host at least four transiting exoplanets. In 2021, a radial velocity (RV) detection of the innermost planet L 98-59 b was announced, with RV semi-amplitude of just 46 cms. This translated into just half the mass of Venus, making L 98-59 b by far the lowest-mass exoplanet with an apparent RV detection, and thus mass measurement. (Much larger RV signals were also measured for planets c and d.) Yet for more than a decade, astrophysicists have struggled to make convincing detections of low-mass exoplanets with RV signals ≤ 1 ms, due largely to stellar and other nuisance signals that tend to drown out the signals of these small exoplanets that are many trillions of kilometres away. Indeed, the statistical methods used to support the detection of L 98-59 b were recently called into question (relevant pre-print forthcoming by end of 2023). The aim of this project will therefore be to model the publicly-available RV measurements of L 98-59, to reproduce the relatively straightforward detection of planets c and d, and characterise key properties (planetary masses, densities, etc.), while quantifying the statistical evidence supporting (or opposing) the detection of planet b.

Reading List

1. Perryman, Michael. The Exoplanet Handbook. Cambridge University Press, 2018.
2. Characterization of the L 98-59 multi-planetary system with HARPS
3. Statistical Methods for Exoplanet Detection with Radial Velocities
4. A Gaussian process framework for modelling stellar activity signals in radial velocity data
5. The EXPRES Stellar Signals Project II. State of the Field in Disentangling Photospheric Velocities

Data Access

All data to be analysed in this project can be downloaded here