ML-IAP/CCA-2023

Monday, November 27, 2023 - Friday, December 1, 2023

Dual node

Book of Abstracts

Contents

Contributed talks / 2

Toward automated discovery of analytical physical laws from data using deep reinforcement learning

Author: Wassim TENACHI[1]

Co-authors: Rodrigo Ibata [1]; Foivos Diakogiannis [2]

[1] *Observatoire Astronomique de Strasbourg*

[2] *Data61, CSIRO, Australia*

Corresponding Authors: rodrigo.ibata@astro.unistra.fr, foivos.diakogiannis@data61.csiro.au, wassim.tenachi@astro.unistra.fr

Symbolic Regression is the study of algorithms that automate the search for analytic expressions that fit data. With new advances in deep learning there has been much renewed interest in such approaches, yet efforts have not been focused on physics, where we have important additional constraints due to the units associated with our data.

I will present Φ-SO, a Physical Symbolic Optimization framework for recovering analytical symbolic expressions from physical data using deep reinforcement learning techniques. Our system is built, from the ground up, to propose solutions where the physical units are consistent by construction, resulting in compact, physical, interpretable and intellegible analytical models. This is useful not only in eliminating physically impossible solutions, but because it restricts enormously the freedom of the equation generator, thus vastly improving performances.

The algorithm can be used to fit noiseless data, which can be useful for instance when attempting to derive an analytical property of a physical model, and it can also be used to obtain analytical approximations to noisy data or even open up the black box that are neural networks. I will showcase our machinery on a panel of astrophysical cases ranging from high energy astrophysics to galactic dynamics, all the way to cosmology. I will then touch on our preliminary results in applying this type of approach to physical differential equations.

Contributed talks / 4

SNAD: enabling discovery in the era of big data

Author: Maria Pruzhinskaya[None]

Co-author: +SNAD team

Corresponding Author: pruzhinskaya@gmail.com

In the era of wide-field surveys and big data in astronomy, the SNAD team (https://snad.space) is exploiting the potential of modern datasets for discovery new, unforeseen, or rare astrophysical phenomena. The SNAD pipeline was built under the hypothesis that, although automatic learning algorithms have a crucial role to play in this task, the scientific discovery is only completely realized when such systems are designed to boost the impact of domain knowledge experts. Our key contributions include the development of the Coniferest Python library, which offers implementations of two adaptive learning algorithms with an "expert in loop", and the creation of the SNAD Transient Miner, facilitating the search for specific types of transients. We have also developed the SNAD Viewer, a web portal that provides a centralized view of individual objects from the Zwicky Transient Facility's (ZTF) data releases, making the analysis of candidates in anomalies more efficient. Finally, when applied to ZTF data, our approach has yielded over a hundred new supernova candidates, along with few other non-catalogued objects, such as red dwarf flares, active galactic nuclei, RS CVn type variables, and young stellar objects.

Posters / 6

Fast nested sampling with deep neural network model emulators

Author: Johannes Buchner[1]

[1] *Max Planck Institute for extraterrestrial Physics*

Corresponding Author: jbuchner@mpe.mpg.de

Elaborate simulations of physical systems can be approximated by deep learning model emulators, aka surrogate models, based on training data generated from the full model. Because of powerful deep learning libraries and the enormous speed-up to compute model components or the full likelihood, model emulators becoming more common in astronomy. An interesting computational property of deep neural networks on GPU/CPU/TPUs is that the evaluation cost with one model instance is almost the same as the evaluation cost of hundreds of model instances. JAX-based models are limited to a fixed number of model instances. Tailored to this emulator computational model, I will present three new Bayesian inference algorithms based on nested sampling, implemented in UltraNest. Two enable rapid inference in physical systems with 100 or more parameters, currently powering inference on supernova explosions. The third provides robustness guarantees and is ideal low-dimensional inference of many data sets. This is common in large astronomical surveys, and in heavy use in systematic eROSITA and XMM X-ray data analyses.

Posters / **7**

How to create powerful machine learning projects in astronomy

Authors: Johannes Buchner[1]; Sotiria Fotopoulou[2]

[1] *Max Planck Institute for extraterrestrial Physics*

[2] *University of Bristol*

Corresponding Authors: sotiria.fotopoulou@bristol.ac.uk, jbuchner@mpe.mpg.de

Large, freely available, well-maintained data sets have made astronomy a popular playground for machine learning projects. Nevertheless, robust insights gained into both machine learning and physics could be improved by clarity in problem definition and establishing workflows that critically verify, characterize and calibrate machine learning models. We provide a collection of guidelines for setting up machine learning projects to make them likely useful for science, less frustrating and time-intensive for the scientist and their computers, and more likely to lead to robust insights. We draw examples and experience from astronomy, but the advice is potentially applicable to other areas of science. The recommendations have been influenced by projects with students, and discussions at conferences including ML-IAP2021 in Paris.

Contributed talks / **8**

Deconstructing the galaxy merger sequence with machine vision

Author: Robert Bickley[1]

[1] *University of Victoria*

Corresponding Author: rbickley@uvic.ca

Galaxy mergers are unique in their ability to transform the morphological, kinematic, and intrinsic characteristics of galaxies on short timescales. The redistribution of angular momentum brought on

by a merger can revive, enhance, or truncate star formation, trigger or boost the accretion rate of an AGN, and fundamentally alter the evolutionary trajectory of a galaxy.
These effects are well studied in spectroscopically distinct galaxy pairs, but less so in pre- and post-coalescence merger systems on account of their rarity, and complications surrounding their identification by traditional morphological metrics.
To overcome this obstacle, we use bespoke machine learning morphological classifications to search for merging and merged galaxies in two imaging surveys: the latest data release from the deep and high-resolution Canada France Imaging Survey (CFIS/UNIONS), and the Dark Energy Camera Legacy Survey (DECaLS). I will present the details of our machine learning methodology, and offer our work as a case study on the flexibility and utility of machine vision as a bridge between observations and simulations.
Thanks to new large datasets and methodological advantages ushered in by the popularization of machine learning in astronomy, I will present for the first time an updated, abundant, and pure sample of pre- and post-mergers, and show the results of a temporal study following the star formation and multi-wavelength AGN demographics of galaxy mergers all the way through to coalescence.

Contributed talks / **9**

Machine learning cosmology from void properties

Author: Bonny Y. Wang[1]

Co-authors: Alice Pisani [2]; Benjamin Wandelt [3]; Francisco Villaescusa-Navarro [4]

[1] *Flatiron Institute / Carnegie Mellon University*

[2] *Flatiron Institute / The Cooper Union*

[3] *Institut d'Astrophysique de Paris / The Flatiron Institute*

[4] *Flatiron Institute*

Corresponding Authors: ywang@flatironinstitute.org, apisani@flatironinstitute.org, bwandelt@iap.fr, fvillaescusa@flatironinstitute.org

Cosmic voids are the largest and most underdense structures in the Universe. Their properties have been shown to encode precious information about the laws and constituents of the Universe. We show that machine learning techniques can unlock the information in void features for cosmological parameter inference. Using thousands of void catalogs from the GIGANTES dataset, we explore three properties of voids: ellipticity, density contrast, and radius. Specifically, we train 1) fully connected neural networks on histograms from void properties and 2) deep sets from void catalogs, to perform likelihood-free inference on the value of cosmological parameters. Our results provide an illustration of how machine learning can be a powerful tool for constraining cosmology with voids.

Posters / **10**

Exploring new SHORES

Author: Meriem Behiri[1]

[1] *SISSA*

Corresponding Author: mbehiri@sissa.it

The Serendipitous H-ATLAS fields Observations of Radio Extragalactic Sources (SHORES, PI: Marcella Massardi) is a brand new survey 2.1 GHz performed with the Australia Telescope Compact Array.
It is composed of 30 discontinuous fields covering a total area of 15 sq. deg,. in the Herschel-ATLAS Southern Galactic Pole region (see Eales+2010), centred in candidate lensed galaxies (Negrello+14). With more than 200 hours of observing time, we reached ~ 30μJy sensitivities.

These fields have the perks of being covered by Herschel observations (H-ATLAS sgp) and many other surveys (KIDS, SDSS, GAMA...). With SHORES we aim at:
characterizing the galactic populations in the radio bands up to high redshift
reconstructing the radio luminosity function
understanding the polarization of a wide range of galaxies populations. This survey's wide and panchromatic coverage makes it a perfect target for cosmological studies, both from a structure formation and a foreground point of view, and a perfect playground for the upcoming SKAO observations and testing the latest ML techniques, which will help reach our goals and beyond.

Contributed talks / 11

A neural-network emulator for the Lyman-α flux power spectrum

Author: Laura Cabayol-Garcia[1]

Co-authors: Andreu Font-Ribera [2]; Jonás Chaves-Montero [2]

[1] *IFAE/PIC*

[2] *IFAE*

Corresponding Authors: jchaves@ifae.es, afont@ifae.es, lcabayol@pic.es

The Lyman-α forest presents a unique opportunity to study the distribution of matter in the high-redshift universe and extract precise constraints on the nature of dark matter, neutrino masses, and other extensions to the ΛCDM model. However, accurately interpreting this observable requires precise modeling of the thermal and ionization state of the intergalactic medium, which often relies on computationally intensive hydrodynamical simulations. In this study, we introduce the first neural-network emulator capable of rapidly predicting the one-dimensional Lyman-α flux power spectrum (P_{1D}) as a function of cosmological and IGM parameters.
Traditionally, Gaussian processes have been the preferred choice for emulators due to their ability to make robust predictions with fewer training data points. However, this advantage comes at the cost of runtimes that scale cubically with the number of data points. With the continuous growth of training data sets, the need to transition to algorithms such as neural networks becomes increasingly crucial. Unlike other methods, neural networks provide a linear scaling between time and the number of training points. This scalability is particularly advantageous as it allows for efficient processing even with large datasets. Additionally, the use of GPUs further accelerates neural-network computations, enhancing the speed and efficiency of the training process.
Our emulator has been specifically designed to analyze medium-resolution spectra from the Dark Energy Spectroscopic Instrument (DESI) survey, considering scales ranging from $k_{\parallel} = 0.1$ to $4\ \mathrm{Mpc}^{-1}$ and redshifts from $z = 2$ to $z = 4.5$. DESI employs a sophisticated instrument equipped with thousands of optical fibers that simultaneously collect spectra from millions of galaxies and quasars. Indeed, DESI started 2 years ago, and it has already doubled the amount of quasar spectra previously obtained.
Our approach involves modeling P_{1D} as a function of the slope and amplitude of the linear matter power spectrum, rather than directly as a function of cosmological parameters. We demonstrate that our emulator achieves sub-percent precision across the entire range of scales. Additionally, the emulator maintains this level of accuracy for three ΛCDM extensions: massive neutrinos, running of the spectral index, and curvature. It also performs at the percent level for thermal histories not present in the training set.
To emulate the probability distribution of P_{1D} at any given k scale, we employ a mixture density network. This allows us to estimate the emulator's uncertainty for each prediction, enabling the rejection of measurements associated with high uncertainty. We have observed that the neural network assigns higher uncertainties to inaccurate emulated P_{1D} values and to training points that lie close to the limits of the convex hull. Furthermore, by emulating the probability distribution of P_{1D}, we can estimate the covariance of the emulated values, providing insights into the correlation at different scales. While further investigations are required to enhance our understanding of P_{1D} measurement covariances, we are pleased to note that, to the best of our knowledge, this study represents the first instance in which a complete emulator covariance is provided for P_{1D} emulators, rather than solely focusing on the diagonal elements.
Given the demonstrated sub-percent precision, robustness to ΛCDM extensions, and the ability to

estimate uncertainties, we expect that the developed neural network emulator will play a crucial role in the cosmological analysis of the DESI survey.

Contributed talks / 12

Machine learning for new physics

Author: Agnès Ferté[1]

[1] *SLAC/Stanford U*

Corresponding Author: ferte@slac.stanford.edu

While the benefits of machine learning for data analysis are widely discussed, I will argue that machine learning has also the great potential to inform us on interesting directions in new physics. Indeed, the current approach to solve the big questions of cosmology today is to constrain a wide range of cosmological models (such as cosmic inflation or modified gravity models), which is costly. In our recently published approach https://arxiv.org/abs/2110.13171, we propose to use unsupervised learning to map models according to their impact on cosmological observables. We can thus visualize which models have a different impact and therefore are worth investigating further, using this map as a guide to unlock to information about new physics from the new generation of cosmological surveys. In this talk, I will explain the approach, its use case and its application to the space of modified gravity probed by cosmic shear.

Posters / 14

Convolutional Neural Networks to study Complex Organic Molecules in Radioastronomy

Authors: Nina Kessler[1]; Timea Csengeri[None]; Sylvain Bontemps[None]; David Cornu[None]

[1] *Laboratoire d'Astrophysique de Bordeaux*

Corresponding Author: nina.kessler@u-bordeaux.fr

During the process of star formation, a wide variety of molecules can form. The use of ALMA interferometer has made it possible to detect a richness of complex organic molecules (COMs) towards hot cores and hot corinos by studying their rotational transitions. However, the analysis of such spectra is a tedious work and actual technics are not optimal, especially for analyzing a large sample of spectra in a systematic way. Moreover, the amount of data related to these observations has increased considerably in recent years. Therefore, it becomes necessary to develop new tools based on Artificial Intelligence to automate line detection and identification. In this context, we set ourselves the challenge of building an appropriate Neural Network architecture that is able to catch the fine details of molecular signature. This presentation would be the opportunity to discuss our first results on the building of CNNs to facilitate the analysis of large samples of (sub)millimeter spectra.

Contributed talks / 16

Emulating the Universe: overcoming computational roadblocks with Gaussian processes

Author: Benjamin Giblin[1]

[1] *University of Edinburgh*

Corresponding Author: bengib@roe.ac.uk

Whether it's calibrating our analytical predictions on small scales, or devising all new probes beyond standard two-point functions, the road to precision cosmology is paved with numerical simulations. The breadth of the parameter space we must simulate, and the associated computational cost, however, present a serious challenge. Fortunately, emulators based on Gaussian processes and neural networks provide a way forward, allowing for numerical models to be constructed through training machine learning algorithms on a tractable number of mocks. In this talk, I will present cosmological constraints derived from new statistics made possible by a simulation-based emulator model, and argue that this new approach presents a practical and environmentally-conscious path towards accurate cosmological inference.

Contributed talks / **17**

Machine Learning Powered Inference in Cosmology

Author: Pablo Lemos[1]

[1] *Mila - Université de Montréal*

Corresponding Author: plemos91@gmail.com

The main goal of cosmology is to perform parameter inference and model selection, from astronomical observations. But, uniquely, it is a field that has to do this limited to a single experiment, the Universe we live in. With compelling existing and upcoming cosmological surveys, we need to leverage state-of-the-art inference techniques to extract as much information as possible from our data.

In this talk, I will begin present Machine Learning based methods to perform inference in cosmology, such as simulation-based inference, and stochastic control sampling approaches. I will show how we can use Machine Learning to perform parameter inference of multimodal posterior distributions on high dimensional spaces. I will finish by showing how these methods are being used to improve our knowledge of the Universe, by presenting the results from the SimBIG analysis on simulation-based inference from large-scale structure data.

Posters / **18**

Extracting the Full Cosmological Information of Galaxy Surveys with SimBIG

Author: ChangHoon Hahn[1]

[1] *Princeton University*

Corresponding Author: changhoon.hahn@princeton.edu

The 3D distribution of galaxies encodes key cosmological information that can probe the growth and expansion history of the Universe. In my talk, I will present how we can leverage simulations and machine learning to go beyond current analyses and extract the full cosmological information of the next-generation galaxy surveys. In particular, I will present SimBIG, a forward modeling framework for analyzing galaxy clustering using simulation-based inference based on normalizing flows. I will show the latest results from applying SimBIG to BOSS observations to analyze the bispectrum, wavelet scattering transform, and a field-level summary based on convolutional neural networks—all down to small, non-linear, scales. By robustly extracting additional cosmological information, we constrain ΛCDM parameters, Ω_b, h, n_s, Ω_m, and σ_8, that are 2.4, 1.5, 1.7, 1.2, and 2.7$\times$ tighter than

standard power spectrum analyses. With this increased precision, we derive constraints on the Hubble constant, H_0, and $S_8 = \sigma_8\sqrt{\Omega_m/0.3}$ that are competitive with other cosmological probes and inform cosmic tensions, even with a sample that only spans 10\% of the full BOSS volume. Lastly, I will discuss how SimBIG can be extended to upcoming spectroscopic galaxy surveys (DESI, PFS, Euclid) to produce leading H_0 and S_8 constraints.

Contributed talks / **19**

Cosmology with Galaxy Photometry Alone

Author: ChangHoon Hahn[1]

[1] *Princeton University*

Corresponding Author: changhoon.hahn@princeton.edu

In this talk I will present the first cosmological constraints from only the observed photometry of galaxies. Villaescusa-Navarro *et al.* (2022) recently demonstrated that the internal physical properties of a single galaxy contain a significant amount of cosmological information. These physical properties, however, cannot be directly measured from observations. I will present how we can go beyond theoretical demonstrations to infer cosmological constraints from actual galaxy observables (*e.g.* optical photometry) using neural density estimation and the CAMELS suite of hydrodynamical simulations. We find that the cosmological information in the photometry of a single galaxy is limited. However, we can combine the constraining power of photometry from many galaxies using hierarchical population inference and place significant cosmological constraints. With the observed photometry of ∼15,000 NASA-Sloan Atlas galaxies, we constrain $\Omega_m = 0.310^{+0.080}_{-0.098}$ and $\sigma_8 = 0.792^{+0.099}_{-0.090}$.

Posters / **22**

CosmoPower-JAX: high-dimensional Bayesian inference with differentiable cosmological emulators

Author: Davide Piras[1]

Co-author: Alessio Spurio Mancini [2]

[1] *University of Geneva*

[2] *University College London*

Corresponding Author: davide.piras@unige.ch

We present CosmoPower-JAX, a JAX-based implementation of the CosmoPower framework, which accelerates cosmological inference by building neural emulators of cosmological power spectra. We show how, using the automatic differentiation, batch evaluation and just-in-time compilation features of JAX, and running the inference pipeline on graphics processing units (GPUs), parameter estimation can be accelerated by orders of magnitude with advanced gradient-based sampling techniques. These can be used to efficiently explore high-dimensional parameter spaces, such as those needed for the analysis of next-generation cosmological surveys. We showcase the accuracy and computational efficiency of CosmoPower-JAX on two simulated Stage IV configurations. We first consider a single survey performing a cosmic shear analysis totalling 37 model parameters. We validate the contours derived with CosmoPower-JAX and a Hamiltonian Monte Carlo sampler against those derived with a nested sampler and without emulators, obtaining a speed-up factor of O(10ˆ3). We then consider a combination of three Stage IV surveys, each performing a joint cosmic shear and galaxy clustering (3x2pt) analysis, for a total of 157 model parameters. Even with such a high-dimensional parameter space, CosmoPower-JAX provides converged posterior contours in 3 days, as opposed to the estimated 6 years required by standard methods. CosmoPower-JAX is fully written

in Python, and we make it publicly available to help the cosmological community meet the accuracy requirements set by next-generation surveys (https://github.com/dpiras/cosmopower-jax).

Contributed talks / 23

DE-VAE: a representation learning architecture for a dynamic dark energy model

Author: Davide Piras[1]

Co-author: Lucas Lombriser [1]

[1] *University of Geneva*

Corresponding Author: davide.piras@unige.ch

We present DE-VAE, a variational autoencoder (VAE) architecture to search for a compressed representation of beyond-ΛCDM models. We train DE-VAE on matter power spectra boosts generated at wavenumbers k ∈ (0.01 − 2.5) h/Mpc and at four redshift values z ∈ (0.1, 0.48, 0.78, 1.5) for a dynamic dark energy (DE) model with two extra parameters describing an evolving DE equation of state. The boosts are compressed to a lower-dimensional representation, which is concatenated with standard CDM parameters and then mapped back to reconstructed boosts; both the compression ("encoder") and the reconstruction ("decoder") components are parametrized as neural networks. We demonstrate that a single latent parameter can be used to predict DE power spectra at all k and z within 2σ, where the Gaussian error includes cosmic variance, shot noise and systematic effects for a Stage IV-like survey. This single parameter shows a high mutual information (MI) with the two DE parameters, and we obtain an explicit equation linking these variables through symbolic regression. We further show that considering a model with two latent variables only marginally improves the accuracy predictions, and that a third latent variable has no significant impact on the model's performance. We discuss how the DE-VAE framework could be extended to search for a common lower-dimensional parametrization of different beyond-ΛCDM models, including modified gravity and braneworld models. Such a framework could then both potentially serve as an indicator of the existence of new physics in cosmological datasets, and provide theoretical insight into the common aspects of beyond-ΛCDM models.

Contributed talks / 24

Investigations for LSST with Machine Learning: Photometric redshift predictions, strong lens detection and mass modeling

Author: Stefan Schuldt[None]

Corresponding Author: stefan.schuldt@unimi.it

Photometric redshifts and strong lensing are both integral for stellar physics and cosmological studies with the Rubin Observatory Legacy Survey of Space and Time (LSST), which will provide billions of galaxy images in six filters, including on the order of 100,000 galaxy-scale lenses. To efficiently exploit this huge amount of data, machine learning is a promising technique that leads to an extreme reduction of the computational time per object.
Since accurate redshifts are a necessity for nearly any astrophysical study, precise and efficient techniques to predict photometric redshifts are crucial to allow for the full exploitation of the LSST data. To this end, I will highlight in the first part of my talk the novel ability of using convolutional neural networks (CNNs) to estimate the photometric redshifts of galaxies. Since the image quality from LSST is expected to be very similar to that of the Hyper Suprime-Cam (HSC), and training a network on realistic data is crucial to achieve a good performance on real data, the network is trained on real HSC cutouts in five different filters. The good performance will be highlighted with a detailed comparison to the Direct Empirical Photometric (DEmP) method, a hybrid technique with one of the best performances on HSC images.

To address further challenges in efficiently analyzing the huge amount of data provided by LSST, I will present in the second part of my talk some recent machine learning techniques developed within the HOLISMOKES collaboration, which focus on the exploitation of strongly lensed supernovae (SNe). These very rare events offer promising avenues to probe stellar physics and cosmology. For instance, the time-delays between the multiple images of a lensed SN allow for a direct measurement of the Hubble constant (H0) independently from other probes. This allows one to assess the current tension on the H0 value, and the possible need for new physics. Furthermore, these lensed SNe also help constrain the SN progenitor scenarios by facilitating follow-up observations in the first hours after the explosion. In particular, I will summarize our deep learning methods to search for lensed SNe in current and future wide-field time-domain surveys, and focus on our new achievements in the automation of strong-lens modeling with a residual neural network. To train, validate, and test these networks, we mock up images based on real observed galaxies from HSC and the Hubble Ultra Deep Field. These networks are further tested on known real systems to estimate the true performance on real data.
For all the networks, the main advantage is the opportunity to apply these easily and fully automated to millions of galaxies with a huge gain in speed. Both regression networks are able to estimate the parameter values in fractions of a second on a single CPU while the lens modeling with traditional techniques typically takes weeks. With these networks, we will be able to efficiently process the huge amount of expected detections in the near future by LSST.

Contributed talks / 25

Efficient and fast deep learning approaches to denoise large radioastronomy line cubes and to emulate sophisticated astrophysical models

Author: Lucas Einig[1]

Co-authors: Jocelyn Chanussot [2]; Jérôme Pety [1]; Maryvonne Gerin [3]; Pierre Palud [4]

[1] *Institut de Radioastronomie Millimétrique*

[2] *GIPSA-Lab*

[3] *Observatoire de Paris*

[4] *CRIStAL*

Corresponding Author: einig@iram.fr

The interstellar medium (ISM) is an important actor in the evolution of galaxies and provides key diagnostics of their activity, masses and evolutionary state. However, surveys of the atomic and molecular gas, both in the Milky Way and in external galaxies, produce huge position-position-velocity data cubes over wide fields of view with varying signal-to-noise ratios. Besides, inferring the physicals conditions of the ISM from these data requires complex and often slow astrophysical codes.

The overall challenge is to reduce the amount of human supervision required to analyze and interpret these data. I will describe two applications of deep learning to tackle this challenge.

1/ I will first introduce a self-supervised denoising method adapted to molecular line data cubes (Einig et al. 2023). The proposed autoencoder architecture compensates for the lack of redundancy between channels in line data cubes compared to hyperspectral Earth remote sensing data. When applied to a typical data cube of about 10ˆ7 voxels, this method allows to recover the low SNR emission without affecting the signals with high SNR. The proposed method surpasses current state of the art denoising tools, such as ROHSA and GaussPY+, which are based on multiple Gaussian fitting of line profiles.

2/ Numerical simulations are usually too slow to be used in Bayesian inference framework, as it requires numerous model evaluations. Here, I will present a supervised method to derive fast and light neural-network based emulations of a model from a grid of precomputed outputs (Palud et al. 2023). This emulator is compared with four standard classes of interpolation methods used to emulate the Meudon PDR code, a characteristic ISM numerical model. The proposed strategies yield

networks that outperform all interpolation methods in terms of accuracy on outputs that have not been used during training. Moreover, these networks are 1,000 times faster than accurate interpolation methods, and require at most 10 times less memory. This paves the way to efficient inferences using wide-field multi-line observations of the ISM. The proposed strategies can easily be adapted to other astrophysical models.

References:
Einig et al. 2023, A&A, in press
Palud et al. 2023, subm. to A&A

Contributed talks / **26**

SBI meets reality: simulation-based inference in practical cosmology applications

Author: Benjamin Joachimi[1]

[1] *University College London*

Corresponding Author: b.joachimi@ucl.ac.uk

Simulation-based inference (SBI) building on machine-learnt density estimation and massive data compression has the potential to become the method of choice for analysing large, complex datasets in survey cosmology. I will present recent work that implements every ingredient of the current Kilo-Degree Survey weak lensing analysis into an SBI framework which runs on similar timescales as a traditional analysis relying on analytic models and a Gaussian likelihood. We show how the SBI analysis recovers and, in several key aspects, goes beyond the traditional approach. I will also discuss challenges and their solutions to SBI-related data compression and goodness-of-fit in several real-world cosmology applications.

Posters / **27**

The halo-galaxy connection from a machine learning perspective

Authors: Antonio Montero-Dorta[1]; Natalí de Santi[2]; Natália Rodrigues[2]; Raul Abramo[2]

[1] *Universidad Técnica Federico Santa María*

[2] *Universidade de São Paulo*

Corresponding Authors: natalidesanti@gmail.com, natalia.villa.rodrigues@usp.br

The relationship between galaxies and halos is central to describing galaxy formation and a fundamental step toward extracting precise cosmological information from galaxy maps. However, this connection involves several complex processes that are interconnected. Machine learning methods are flexible tools that can learn complex correlations between a large number of features but are traditionally designed as deterministic estimators.
In this work, we use the IllustrisTNG300-1 simulation and investigate how machine learning methods capable of predicting distributions can accurately reproduce features of different galaxy populations based on their host halo properties. In particular, we study how the models can quantify the uncertainty related to the intrinsic scatter in the halo-galaxy connection.

Contributed talks / **28**

Finding Observable Environmental Measures of Halo Properties using Neural Networks

Author: Haley Bowden[1]

Co-authors: Andrew Hearin [2]; Peter Behroozi [1]

[1] *University of Arizona*

[2] *Argonne National Laboratory*

Corresponding Authors: behroozi@arizona.edu, hbowden@arizona.edu

Simulations have revealed correlations between the properties of dark matter halos and their environment, made visible by the galaxies which inherit these connections through their host halos. We define a measure of the environment based on the location and observable properties of a galaxy's nearest neighbors in order to capture the broad information content available in the environment. We then use a neural network to learn the connection between the multi-dimensional space defined by the observable properties of galaxies and the properties of their host halos using mock galaxy-catalogs from UNIVERSEMACHINE. The trained networks will: 1) reveal new connections between galaxy, halo, and environment; 2) serve as a powerful tool for placing galaxies into halos in future cosmological simulations; and 3) be a framework for inferring the properties of real halos from next-generation survey data, allowing for direct comparison between observational statistics and theory. We will first show the results of estimating the masses of halos and sub-halos. This will be followed by preliminary results on halo properties beyond mass, including satellite membership and concentration.

Posters / 29

Calculating enclosed mass with machine learning and line-of-sight data

Author: Jorge Sarrato-Alós[1]

Co-authors: Arianna Di Cintio [1]; Christopher Brook [1]

[1] *Institute of Astrophysics of the Canary Islands*

Corresponding Authors: cbrook@iac.es, adicintio@iac.es, jorgesarrato@gmail.com

Accurately determining the mass distribution within galaxies is crucial for understanding their formation and evolution. Previous research has traditionally relied on analytical equations based on the Jeans equation to estimate the enclosed mass with minimum projection effect. In this study, we present a novel approach to predict the enclosed mass within a given radius using a machine learning model trained on line of sight data of high-resolution cosmological hydrodynamical simulations. Our dataset comprises a diverse sample of galaxies spanning a wide range of masses.

To train the model, we utilize projected positions and velocities of stars within the galaxies. Multiple training iterations are performed, each with the mass enclosed within a different radius as the target variable. By systematically varying the radius, we identify the optimal value at which the neural network exhibits the highest precision in predicting the enclosed mass.

Our results demonstrate the effectiveness of the machine learning-based approach in predicting galaxy mass within a specific radius. The trained model offers a valuable tool for studying galaxy properties, such as mass distribution and gravitational potential, providing insights into the formation and dynamics of galaxies. This work also highlights the utility of machine learning techniques for studying galaxies through line of sight data.

Contributed talks / 30

Neutrino mass constraint from an Implicit Likelihood Analysis of BOSS voids

Author: Leander Thiele[1]

Co-authors: Alice Pisani ; Benjamin Wandelt ; ChangHoon Hahn ; David Spergel ; Elena Massara ; Shirley Ho

[1] *Princeton University*

Corresponding Author: lthiele@princeton.edu

Cosmic voids identified in the spatial distribution of galaxies provide complementary information to two-point statistics. In particular, constraints on the neutrino mass sum, $\sum m_\nu$, promise to benefit from the inclusion of void statistics. We perform inference on the CMASS NGC sample of SDSS-III/BOSS with the aim of constraining $\sum m_\nu$. We utilize the void size function, the void-galaxy cross power spectrum, and the galaxy auto power spectrum. To extract constraints from these summary statistics we use a simulation-based approach, specifically implicit likelihood inference. We populate approximate gravity-only, particle neutrino cosmological simulations with an expressive halo occupation distribution model. With a conservative scale cut of $k_{\rm max} = 0.15\,h{\rm Mpc}^{-1}$ and a Planck-inspired ΛCDM prior, we find upper bounds on $\sum m_\nu$ of 0.43 and $0.35\,{\rm eV}$ from the galaxy auto power spectrum and the full data vector, respectively ($95\,\%$ credible interval). We observe hints that the void statistics may be most effective at constraining $\sum m_\nu$ from below. We also substantiate the usual assumption that the void size function is Poisson distributed.

Contributed talks / 31

Spatially Variant Point Spread Functions for Bayesian Imaging

Author: Vincent Eberle[1]

Co-authors: Margret Westerkamp [1]; Matteo Guardiani [1]; Philipp Frank [2]; Julia Stadler [3]; Philipp Arras [2]; Torsten Enßlin [4]

[1] *Max Planck Institute for Astrophysics / Faculty of Physics, Ludwig-Maximilians-Universität München (LMU)*

[2] *Max Planck Institute for Astrophysics*

[3] *Max Planck Institute for Astrophysics / Excellence Cluster ORIGINS*

[4] *MPI for Astrophysics*

Corresponding Authors: ensslin@mpa-garching.mpg.de, philipp@mpa-garching.mpg.de, matteani@mpa-garching.mpg.de, margret@mpa-garching.mpg.de, veberle@mpa-garching.mpg.de, jstadler@mpa-garching.mpg.de, xray@philipp-arras.de

When measuring photon counts from incoming sky fluxes, observatories imprint nuisance effects on the data that must be accurately removed. Some detector effects can be easily inverted, while others are not trivially invertible such as the point spread function and shot noise. Using information field theory and Bayes' theorem, we infer the posterior mean and uncertainty for the sky flux. This involves the use of prior knowledge encoded in a generative model and a precise and differentiable model of the instrument.
The spatial variability of the point spread functions as part of the instrument description degrades the resolution of the data as the off-axis angle increases. The approximation of the true instrument point spread function by an interpolated and patched convolution provides a fast and accurate representation as part of a numerical instrument model. By incorporating the spatial variability of the point spread function, far off-axis events can be reliably accounted for, thereby increasing the signal-to-noise ratio.
The developed reconstruction method is demonstrated on a series of Chandra X-ray observations of the Perseus galaxy cluster.

Keywords: spatially variant point spread functions; deconvolution; deblurring; X-ray imaging; information field theory; Perseus galaxy cluster; Bayesian imaging

Posters / 32

Bayesian Spatio-spectral Imaging of SN1006 in X-ray

Author: Margret Westerkamp[1]

Co-authors: Vincent Eberle [1]; Matteo Guardiani [1]; Philipp Frank [2]; Lukas Platz [3]; Philipp Arras [2]; Jakob Knollmüller [4]; Julia Stadler [5]; Torsten Enßlin [6]

[1] *Max Planck Institute for Astrophysics; Ludwig-Maximilians-Universität München*

[2] *Max Planck Institute for Astrophysics*

[3] *Max Planck Institute for Astrophysics; Ludwig-Maximilians-Universität München; Institute for Biological and Medical Imaging, Helmholtz Zentrum München; Institute of Computational Biology, Helmholtz Zentrum München; Technical University Munich, School of Medicine*

[4] *Technical University Munich, TUM School of Natural Sciences; Excellence Cluster ORIGINS*

[5] *Max Planck Institute for Astrophysics; Excellence Cluster ORIGINS*

[6] *Max Planck Institute for Astrophysics; Ludwig-Maximilians-Universität München; Excellence Cluster ORIGINS*

Corresponding Authors: philipp@mpa-garching.mpg.de, matteani@mpa-garching.mpg.de, margret@mpa-garching.mpg.de, veberle@mpa-garching.mpg.de, lplatz@mpa-garching.mpg.de, jakob@knollmueller.de, jstadler@mpa-garching.mpg.de, xray@philipp-arras.de

The supernova remnant SN1006 has been studied extensively by various X-ray instruments and telescopes due to its historical record, its proximity, and its brightness. In order to accurately study the properties of this remnant itself, it is essential to obtain a detailed and denoised view of its small-scale structures, given the existing observations. Here, we present a Bayesian spatio-spectral image reconstruction method, based on information field theory, that aims to separate the emission of the remnant from that of other sources in the field.
We describe our priors using generative models that incorporate knowledge of the spatial and spectral correlation structure of the remnant and of other sources, such as point sources and background radiation. Combined with a likelihood model that allows the fusion of multiple data sets and instrument descriptions, we obtain the posterior distribution of the remnant's emission at each point in space and frequency. Furthermore, we introduce a multi-step approach where the spatial reconstruction obtained for a single energy range is used to derive an informed starting point for the full spatio-spectral reconstruction in order to speed up the imaging process.
The developed method is applied to the latest merged Chandra data available to date on SN1006, providing a high quality visualisation of its complex features.
Keywords: SN1006, information field theory, X-ray imaging, Bayesian imaging, spatio-spectral reconstruction, component separation, generative models

Posters / 33

Estimation of Galaxy properties in 3D MUSE archival data with convolutional neural networks

Author: Alejandra Fresco[1]

[1] *University of Milan-Bicocca*

Corresponding Author: alejandra.frescoarrom@unimib.it

Next generation instruments are focused on producing massive amounts of spectroscopic data that require new approaches that are computationally efficient and more accurate. While traditional processes such as the convolution-based template matching have been proven successful, they are computationally demanding. Machine learning methods have proven to be orders of magnitude faster and showing promising results. Here we built on the previous efforts and explore these techniques further. Using simulated 3D MUSE cubes, we train a Convolutional Neural Network (CNN) to detect and measure the Lyman-alpha, C IV, and He II emission lines, in order to trace the overdensities and characterise the large scale structure environment. We then test the accuracy of the

CNN against real data using ˜300 deep field MUSE cubes of archival data. With this work we will have a new tool for processing and characterising large amounts of data, which will be faster and less computationally demanding. We will also be able to set tighter constraints on this new method of quasar spectra analysis, aiming to extrapolate the results to the new upcoming massive spectroscopic surveys.

Posters / **34**

Cosmological Parameter Inference Machine Learning Algorithms with Constrained Cosmological Simulations

Author: Elena Hernandez Martinez[1]

[1] *Ludwig-Maximilian University Munich (LMU, Universitäts Sternwarte)*

Corresponding Author: elenahmd@gmail.com

The ΛCDM model stands as the prevailing framework in cosmology, yet discrepancies between Cosmic Microwave Background (CMB) and late universe probes underscore incomplete understanding of essential cosmological parameters, like Ωm and σ8, which govern matter density and density fluctuations in the Universe. To address the limitations of traditional statistical methods, we have developed a novel set of constrained cosmological simulations known as SLOW. These simulations have demonstrated exceptional precision in replicating observed structures within the Local Universe within a cosmological box of size 500 Mpc/h, rendering them an exemplary testbed for diverse cosmological investigations, including the application of Machine Learning techniques for precise cosmological parameter inference within our Local Universe.

Contributed talks / **35**

Who threw that rock? Tracing the path of martian meteorites back to the crater of origin using ML

Author: Konstantinos Servis-Nussbaum[1]

Co-authors: Anthony Lagain [2]; Gretchen Benedix [2]; John Fairweather [2]

[1] *Pawsey/CSIRO*

[2] *Curtin SSTC*

Corresponding Authors: anthony.lagain@curtin.edu.au, g.benedix@curtin.edu.au, john.fairweather@postgrad.curtin.edu.au, knservis@gmail.com

We created an ML pipeline able to efficiently detect craters in a large dataset of georeferenced images. We used it to create a detailed database of craters on rocky bodies in the solar system including Mars. The Mars crater database was of sufficient detail to enable us to determine the likely origin of a number of meteorites that we have collected on Earth. As a consequence, it is possible to get a better picture of the early formation processes of Mars using a sample from Mars, before the first sample-return mission has been organized. In this presentation, we will see how we have structured our pipeline and the technologies used to produce that data product.

Contributed talks / **36**

Extending the Reach of Gaia DR3 with Self-Supervision

Author: Aydan McKay[1]

Co-author: Sébastien Fabbro [2]

[1] *University of Victoria*

[2] *NRC Herzberg Astronomy and Astrophysics*

Corresponding Authors: aydanmckay@uvic.ca, sebastien.fabbro@nrc-cnrc.gc.ca

The Gaia Collaboration's 3rd data release (DR3) provides comprehensive information including photometry and kinematics on more than a billion stars across the entire sky up to $G \approx 21$, encompassing approximately 220 million stars with supplementary low-resolution spectra ($G < 17.6$). These spectra offer derived valuable stellar properties like [Fe/H], $\log g$, and T_{eff}, serving as proxies to identify and characterize significant stellar structures, such as stellar streams formed from past minor galaxy mergers with the Milky Way.

In pursuit of constraining the chemo-dynamical history of the Galaxy with data-driven algorithms, we propose a novel self-supervised approach implementing masked stellar modelling (MSM) exploiting multiple spectroscopic and photometric surveys to extend beyond the limitations of DR3's low-resolution spectra. We incorporate diverse imaging surveys that span ultraviolet to near-infrared wavelengths across the celestial sphere. The MSM employs a powerful encoder to generate informative embeddings, containing crucial information for downstream tasks, facilitated by an extensive training sample. By leveraging these embeddings, similarity searches on the complete database of embeddings can be conducted instantly. Moreover, spectroscopic surveys often exhibit inconsistencies due to varying assumptions in their respective derivations of stellar characteristics. The MSM method offers the ability to fine-tune any survey on specific stellar astrophysics tasks with much fewer labels, and thanks to its extensive training set, is more robust to misrepresentativity. The stellar embeddings result in a self-consistent dataset, effectively establishing a comprehensive stellar model.

Overall, this research showcases an innovative data-driven approach to utilize various surveys and spectral products, empowering researchers to make significant strides in understanding the Milky Way's history and dynamics. The methodology's effectiveness in regression tasks and its scalability will be highlighted, shedding light on its broader applicability.

Contributed talks / **37**

Prioritising Follow-up for Transient Surveys in the New Era of Time-Domain Astronomy

Author: Daniel Muthukrishna[1]

[1] *Massachusetts Institute of Technology*

Corresponding Author: daniel.muthukrishna@gmail.com

New large-scale astronomical surveys such as the Vera Rubin Observatory's Legacy Survey of Space and Time (LSST) have the potential to revolutionize transient astronomy, providing opportunities to discover entirely new classes of transients while also enabling a deeper understanding of known supernovae. LSST is expected to observe over 10 million transient alerts every night, over an order of magnitude more than any preceding survey. In this talk, I'll discuss the issue that with such large data volumes, the astronomical community will struggle to prioritize which transients - rare, interesting, or young - should be followed up. I address three major challenges: (1) automating real-time classification of transients, (2) automating serendipity by identifying the likelihood of a transient being interesting and anomalous, and (3) identifying the epoch time in order to observe transients early to understand their central engine and progenitor systems. I present machine learning and Bayesian methods of automating real-time classification, anomaly detection, and predicting epoch times of transients. Our ability to classify events and identify anomalies improves over the lifetime of the light curves.

Contributed talks / 38

Current progress and challenges from the Cosmology and Astrophysics with MachinE Learning Simulations (CAMELS) project

Author: Daniel Angles-Alcazar[1]

[1] *University of Connecticut*

Corresponding Author: angles-alcazar@uconn.edu

Large-volume cosmological hydrodynamic simulations have become a primary tool to understand supermassive black holes (SMBHs), galaxies, and the large-scale structure of the Universe. However, current uncertainties in sub-grid models for core physical processes such as feedback from massive stars and SMBHs limit their predictive power and plausible use to extract information from extragalactic surveys. In this talk, I will present an overview of the Cosmology and Astrophysics with MachinE Learning Simulations (CAMELS) project, containing thousands of simulations implementing different cosmological and astrophysical parameters, sub-grid galaxy formation implementation, and hydrodynamics solver, and designed to train machine learning algorithms to maximize the extraction of information from cosmological surveys while marginalizing over uncertainties in sub-grid physics. I will show illustrative examples of the broad range of possible applications of CAMELS, discuss recent progress and challenges building robust simulation-based inference models for cosmology, and advertise the latest additions to the ever-growing CAMELS public data repository.

Contributed talks / 39

Machine-directed gravitational-wave counterpart discovery

Author: Niharika Sravan[1]

[1] *Drexel University*

Corresponding Author: niharika.sravan@gmail.com

Joint observations in electromagnetic and gravitational waves shed light on the physics of objects and surrounding environments with extreme gravity that are otherwise unreachable via siloed observations in each messenger. However, such detections remain challenging due to the rapid and faint nature of counterparts. Protocols for discovery and inference still rely on human experts manually inspecting survey alert streams and intuiting optimal usage of limited follow-up resources. Strategizing an optimal follow-up program requires adaptive sequential decision-making given evolving light curve data that (i) maximizes a global objective despite incomplete information and (ii) is robust to stochasticity introduced by detectors/observing conditions. Reinforcement learning (RL) approaches allow agents to implicitly learn the physics/detector dynamics and the behavior policy that maximize a designated objective through experience.
To demonstrate the utility of such an approach for the kilonova follow-up problem, we train a toy RL agent for the goal of maximizing follow-up photometry for the true kilonova among several contaminant transient light curves. In a simulated environment where the agent learns online, it achieves 3x higher accuracy compared to a random strategy. However, it is surpassed by human agents by up to a factor of 2. This is likely because our hypothesis function (Q that is linear in state-action features) is an insufficient representation of the optimal behavior policy. More complex agents could perform at par or surpass human experts. Agents like these could pave the way for machine-directed software infrastructure to efficiently respond to next generation detectors, for conducting science inference and optimally planning expensive follow-up observations, scalably and with demonstrable performance guarantees.

Contributed talks / 40

Data Compression and Inference in Cosmology with Self-Supervised Machine Learning

Authors: Aizhan Akhmetzhanova[1]; Cora Dvorkin[None]; Siddharth Mishra-Sharma[None]

[1] *Harvard University*

Corresponding Authors: cdvorkin@g.harvard.edu, smsharma@mit.edu, aakhmetzhanova@g.harvard.edu

The influx of massive amounts of data from current and upcoming cosmological surveys necessitates compression schemes that can efficiently summarize the data with minimal loss of information. We introduce a method that leverages the paradigm of self-supervised machine learning in a novel manner to construct representative summaries of massive datasets using simulation-based augmentations. Deploying the method on hydrodynamical cosmological simulations, we show that it can deliver highly informative summaries, which can be used for a variety of downstream tasks, including precise and accurate parameter inference. We demonstrate how this paradigm can be used to construct summary representations that are insensitive to prescribed systematic effects, such as the influence of baryonic physics. Our results indicate that self-supervised machine learning techniques offer a promising new approach for compression of cosmological data as well its analysis.

Contributed talks / 41

Gravitational Wave Paleontology and the Progenitor Uncertainty Challenge

Author: Floor Broekgaarden[1]

[1] *Columbia University, Simons Foundation (Junior Fellow)*

Corresponding Author: fsbroekgaarden@gmail.com

We are on the precipice of the Big Data gravitational wave (GW) era. Pairs of stellar-mass black holes (BHs) or neutron stars (NSs) across our vast Universe occasionally merge, unleashing bursts of gravitational waves that we can observe here on Earth since their first detection in 2015. Over the next few years, the population of detected mergers will rapidly increase from a few hundred to many million detections per year as new GW observing runs (LIGO/Virgo O4; 2023, LIGO/Virgo O5; 2025) and next-generation detectors (Cosmic Explorer; Einstein Telescope, LISA; 2035) provide data with ever-increasing precision and to larger distances, pushing the reach of gravitational-wave astronomy to the edge of the observable Universe; revolutionizing our view of the cosmos. Making the most of these observations and the rapidly increasing landscape of gravitational-wave detections requires comparing the observed properties, such as their rates, BH and NS masses, and BH spins, to theoretical "population synthesis models"simulating their formation pathways. However, at present, this endeavor is limited by the so-called progenitor "Uncertainty Challenge": uncertainties within the theoretical models are so large, and the models so computationally expensive, that learning about the underlying fundamental physical processes in the lives and deaths of massive stars from observations is completely out of reach, especially for rare events. All present-day simulations thus pay a high price by using highly approximate algorithms that treat the physical processes in a simplified way, or by limiting the total number of simulations, restricting the exploration of the impact of the uncertain physical input assumptions beyond a few variations. In this talk I will introduce the problem and lead an interactive discussion with the participants to investigate statistical techniques to tackle the key Uncertainty Challenge bottleneck in two key areas: (i) improving the sampling of rare events (such as GW sources) in simulations by improving techniques such as adaptive importance sampling, Markov Chain Monte Carlo, and nested sampling and (ii) developing effective emulators predicting model outcomes from small parameter explorations by improving upon techniques from deep learning, normalizing flows, uncertainty quantification, and Gaussian process regression.

Posters / 43

Systematic biases in machine learning and their impact on astronomy research

Author: Lior Shamir[1]

[1] *Kansas State University*

Corresponding Author: lshamir@mtu.edu

Machine learning, and in particular deep neural networks (DNNs), have become primary tools for automatic annotation and analysis of astronomical data. Given that astronomy have been becoming increasingly more dependent on Earth-based and space-based digital sky surveys generating vast pipelines of astronomical data, a large number of DNN-based solutions have already been proposed and applied. But although DNNs are accurate and effective, they also introduce biases that are difficult to notice, profile, and control. Here I describe simple experiments that show that even properly trained DNNs with no apparent flaws in the design process can lead to small but consistent biases that are very difficult to notice, and can therefore be viewed incorrectly as new discoveries. The experiments show that these biases exist in image data, as well as photometry and spectroscopy data when the data are analyzed by machine learning algorithms. Such biases can lead to unusual patterns that can be observed in catalogs and data products prepared with the involvement of machine learning. These biases are often difficult to notice, and their presence is not necessarily expected by unsuspecting data users. Therefore, such biases might lead to incorrect conclusions about astronomy, while they are in fact properties of the data annotation algorithms. Therefore, catalogs and data products generated with the involvement of DNNs should be used with caution, and consumers of such catalogs must be fully aware of the vulnerability of DNNs to complex biases.

Contributed talks / 45

Dealing with systematic effects: the issue of robustness to model misspecification

Authors: Florent Leclercq[1]; Tristan Hoellinger[1]

[1] *Institut d'Astrophysique de Paris*

Corresponding Author: florent.leclercq@iap.fr

Model misspecification is a long-standing problem for Bayesian inference: when the model differs from the actual data-generating process, posteriors tend to be biased and/or overly concentrated. This issue is particularly critical for cosmological data analysis in the presence of systematic effects. I will briefly review state-of-the-art approaches based on an explicit field-level likelihood, which sample known foregrounds and automatically report unknown data contaminations. I will then present recent methodological advances in the implicit likelihood approach, with arbitrarily complex forward models of galaxy surveys where all relevant statistics can be determined from numerical simulations. The method (Simulator Expansion for Likelihood-Free Inference, SELFI) allows to push analyses further into the non-linear regime than state-of-the-art backward modelling techniques. Importantly, it allows a check for model misspecification at the level of the initial matter power spectrum before final inference of cosmological parameters. I will present an application to a Euclid-like configuration.

Contributed talks / 46

A Bayesian Neural Network based ILC method to estimate accurate CMB polarization power spectrum over large angular scales

Author: Sarvesh Kumar Yadav[1]

[1] *Raman Research Institute, Banglore, India*

Corresponding Author: sarveshkesav@gmail.com

Observations of the Cosmic Microwave Background (CMB) radiation have made significant contributions to our understanding of cosmology. While temperature observations of the CMB have greatly advanced our knowledge, the next frontier lies in detecting the elusive B-modes and obtaining precise reconstructions of the CMB's polarized signal in general. In anticipation of proposed and upcoming CMB polarization missions, this study introduces a novel method for accurately determining the angular power spectrum of CMB E-modes and B-modes. We have developed a Bayesian Neural Network (BNN)-based approach to enhance the performance of the Internal Linear Combination (ILC) technique. Our method is applied separately to the frequency channels of both the LiteBird and ECHO (also known as CMB-Bharat) missions and its performance is rigorously assessed for both missions. Our findings demonstrate the method's efficiency in achieving precise reconstructions of both CMB E-modes and CMB B-mode angular power spectra, with errors constrained primarily by cosmic variance.

Posters / **47**

Towards Automatic Point Source Detection

Authors: Matteo Guardiani[1]; Vincent Eberle[1]; Margret Westerkamp[1]; Philipp Frank[1]; Torsten Enßlin[2]

[1] *Max Planck Institute for Astrophysics*

[2] *MPI for Astrophysics*

Corresponding Authors: ensslin@mpa-garching.mpg.de, philipp@mpa-garching.mpg.de, matteani@mpa-garching.mpg.de, margret@mpa-garching.mpg.de, veberle@mpa-garching.mpg.de

For a deep understanding of the Universe, it is crucial to rely on complete and accurate information on its primary constituents. These constituents, such as galaxies, black holes, supernovae, and other compact objects, show distinct features in the sky and therefore imprint differently on astronomical data. In this work, we leverage these differences to construct statistical models for their a priori independent distributions in the sky. This not only enhances the overall observation reconstruction, but also allows to segregate the flux of the various components that populate the sky and more accurately study their individual features. Specifically, we introduce a new technique that uses a notion of latent-space model stress to automatically separate point-like sources from diffuse, correlated structures. We showcase our results on publicly available data.

Contributed talks / **48**

Field-Level Inference with Microcanonical Langevin Monte Carlo

Author: Adrian Bayer[1]

[1] *Princeton University / Simons Foundation*

Corresponding Author: abayer@princeton.edu

Extracting optimal information from upcoming cosmological surveys is a pressing task, for which a promising path to success is performing field-level inference with differentiable forward modeling. A key computational challenge in this approach is that it requires sampling a high-dimensional

parameter space. In this talk I will present a new promising method to sample such large parameter spaces, which improves upon the traditional Hamiltonian Monte Carlo, to both reconstruct the initial conditions of the Universe and obtain cosmological constraints.

Contributed talks / 50

Opportunities and challenges of machine learning for astrophysics

Author: Jason McEwen[1]

[1] *UCL*

Corresponding Author: jason.mcewen@gmail.com

Machine learning (ML) is having a transformative impact on astrophysics. The field is starting to mature, where we are moving beyond the naive application of off-the-shelf, black-box ML models towards approaches where ML is an integral component in a larger, principled analysis methodology. Furthermore, not only are astrophysical analyses benefiting from the use of ML, but ML models themselves can be greatly enhanced by integrating knowledge of relevant physics. I will review three maturing areas where ML and astrophysics have already demonstrated some success, while still providing many further opportunities and challenges. (1) Physics-enhanced learning integrates knowledge of relevant physics into ML models, either through augmentation, encoding symmetries and invariances, encoding dynamics, or directly through physical models that are integrated into the ML model. (2) In statistical learning, ML and statistical models are tightly coupled to provide probabilistic frameworks, often in a Bayesian setting, that offer uncertainty quantification, generative models, accelerated inference, and data-driven priors. (3) For scientific analyses in particular, it is important that ML models are not opaque, black-boxes but are intelligible, ensuring truthfulness, explainability and interpretability. Throughout I will provide numerous examples of astrophysical studies where such approaches have or are being developed and applied, in the context of upcoming observations from the Euclid satellite, the Rubin Observatory Legacy Survey of Space and Time (LSST), and the Square Kilometre Array (SKA). Finally, I will highlight outstanding challenges and some thoughts on how these may be overcome.

Posters / 51

Self-supervised learning applied to outlier detection: searching for jellyfish in the ocean of data from upcoming surveys

Author: Yash Gondhalekar[None]

Co-authors: Ana Chies Santos [1]; Carolina Queiroz ; Rafael de Souza [2]

[1] *Universidade Federal do Rio Grande do Sul*

[2] *University of Hertfordshire*

Corresponding Authors: yashgondhalekar567@gmail.com, ana.chies@ufrgs.br, r.da-silva-de-souza@herts.ac.uk, c.queirozabs@gmail.com

Human visual classification has been the traditional approach to identifying galaxies possessing extreme ram-pressure stripping, the so-called Jellyfish galaxies. However, this approach can lead to misclassifications due to human biases and is unsuitable for large-scale galaxy surveys. In this study, we employ self-supervised learning on a dataset of $\sim$200 images to extract semantically meaningful representations of galaxies. Despite the small dataset size, a similarity search using these representations demonstrates the robustness of the approach and slightly better performance than traditional supervised learning. Using self-supervised learning, we propose a straightforward framework

for assigning JClass, a categorical stripping measure, using a k-nearest neighbor search in the self-supervised representation space. Our method can assist human visual classifiers and help improve the quality of JClass by significantly eliminating biases due to visual subjectiveness or supervised learning. Our framework is versatile and can be applied to various astronomical scenarios requiring the identification of rare objects within massive datasets.

Posters / **52**

Reinventing Astronomical Survey Scheduling with Reinforcement Learning: Unveiling the Potential of Self-Driving Telescopes

Author: Franco Terranova[1]

Co-authors: Brian Nord [2]; Eric Neilsen [2]; M Voetberg [2]

[1] *University of Pisa, Fermilab*

[2] *Fermilab*

Corresponding Authors: neilsen@fnal.gov, f.terranova2@studenti.unipi.it, maggiev@fnal.gov, nord@fnal.gov

The rise of cutting-edge telescopes such as JWST, the Large Synoptic Survey Telescope (LSST), and the Nancy Grace Roman telescope (NGRT) has introduced a new era of complexity in the realm of planning and conducting observational cosmology campaigns.
Astronomical observatories have traditionally relied on manual planning of observations, e.g., human-run and human-evaluated simulations for every observing scenario, which can potentially result in suboptimal observations in terms of scheduling optimization.
Reinforcement learning (RL) has been well-demonstrated as a valuable approach for training autonomous systems, and it may provide the basis for self-driving telescopes capable of scanning the sky and collecting valuable data.
We have developed a framework for statistical learning-based optimization of telescope scheduling that can enhance data acquisition given a predefined scientific reward, e.g., optimizing for a volume-based survey.
The observational campaign is framed as a Markov Decision Process (MDP), a mathematical framework that effectively captures the essence of sequential decision-making.
We compared several RL algorithms applied to a simulated offline dataset, i.e., pre-recorded interactions between the telescope and the sky, considering a discrete set of sky locations the telescope is allowed to visit.
In our study, we conducted comparisons between policy-based, value-based methods, and evolutionary computation strategies.
Value-based methods, and in particular Deep Q-Networks (DQNs), have shown remarkable success in the optimization of astronomical observations for our dataset.
Our experimental results on the test set, demonstrate that the combination of dataset preprocessing techniques, along with the combination of well-known improvements from the literature, such as Dueling DQN, n-steps Bellman unrolling, and noisy networks, yield high performances and capabilities to generalize on unseen data for our task.
In the full environment, the average reward value in each state was found to be 92%±5% of the maximum possible reward. The results from the test set showed an average value of 87%, with a standard deviation of 9%.

Contributed talks / **53**

Probing primordial non-Gaussianity by reconstructing the initial conditions with convolutional neural networks

Author: Xinyi Chen[1]

Co-authors: Nikhil Padmanabhan [1]; Daniel Eisenstein [2]; Fangzhou Zhu [3]; Sasha Gaines [1]

[1] *Yale University*

[2] *Harvard University*

[3] *Google LLC*

Corresponding Author: xinyi.chen@yale.edu

Inflation remains one of the enigmas in fundamental physics. While it is difficult to distinguish different inflation models, information contained in primordial non-Gaussianity (PNG) offers a route to break the degeneracy. In galaxy surveys, the local type PNG is usually probed by measuring the scale-dependent bias in the power spectrum. We introduce a new approach to measure the local type PNG by computing a three-point estimator using reconstructed density field, a density field reversed to the initial conditions from late time. This approach offers an alternative way to the existing method with different systematics and also organically follows the procedure of BAO analysis in large galaxy surveys. We introduce a reconstruction method using convolutional neural networks that significantly improves the performance of traditional reconstruction algorithms in matter density field, which is crucial for more effectively probing PNG. This pipeline can be applied to the ongoing Dark Energy Spectroscopical Instrument (DESI) and *Euclid* surveys, as well as upcoming projects, such as the *Nancy Roman Space Telescope*.

Contributed talks / **54**

Multiview Symbolic Regression in astronomy

Author: Etienne Russeil[1]

Co-authors: Emille Ishida [1]; Emmanuel Gangler [1]; Fabricio Olivetti de França [2]; Konstantin Malanchev [3]

[1] *Laboratoire de Physique de Clermont (LPC)*

[2] *Heuristics, Analysis and Learning Laboratory (HAL), Federal University of ABC*

[3] *Department of Astronomy, University of Illinois*

Corresponding Authors: emille.ishida@clermont.in2p3.fr, etienne.russeil@clermont.in2p3.fr

Symbolic Regression is a data-driven method that searches the space of mathematical equations with the goal of finding the best analytical representation of a given dataset. It is a very powerful tool, which enables the emergence of underlying behavior governing the data generation process. Furthermore, in the case of physical equations, obtaining an analytical form adds a layer of interpretability to the answer which might highlight interesting physical properties.
However equations built with traditional symbolic regression approaches are limited to describing one particular event at a time. That is, if a given parametric equation was at the origin of two datasets produced using two sets of parameters, the method would output two particular solutions, with specific parameter values for each event, instead of finding a common parametric equation. In fact there are many real world applications –in particular astrophysics – where we want to propose a formula for a family of events which may share the same functional shape, but with different numerical parameters
In this work we propose an adaptation of the Symbolic Regression method that is capable of recovering a common parametric equation hidden behind multiple examples generated using different parameter values. We call this approach Multiview Symbolic Regression and we demonstrate how it can reconstruct well known physical equations. Additionally we explore possible applications in the domain of astronomy for light curves modeling. Building equations to describe astrophysical object behaviors can lead to better flux prediction as well as new feature extraction for future machine learning applications.

Contributed talks / **55**

Towards an Astronomical Foundation Model for Stars with a Transformer-based Model

Authors: Henry Leung[1]; Jo Bovy[1]

[1] *University of Toronto*

Corresponding Authors: bovy@astro.utoronto.ca, henrysky.leung@utoronto.ca

Rapid strides are currently being made in the field of artificial intelligence using Transformer-based models like Large Language Models (LLMs). The potential of these methods for creating a single, large, versatile model in astronomy has not yet been explored except for some uses of the basic component of Transformer –the attention mechanism. In this talk, we will talk about a framework for data-driven astronomy that uses the same core techniques and architecture as used by LLMs without involving natural language but floating point data directly. Using a variety of observations and labels of stars as an example, we have built a Transformer-based model and trained it in a self-supervised manner with cross-survey data sets to perform a variety of inference tasks. In particular, we have demonstrated that a single model can perform both discriminative and generative tasks even if the model was not trained or fine-tuned to do any specific task. For example, on the discriminative task of deriving stellar parameters from Gaia XP spectra, our model slightly outperforms an expertly trained XGBoost model in the same setting of inputs and outputs combination. But the same model can also generate Gaia XP spectra from stellar parameters, inpaint unobserved spectral regions, extract empirical stellar loci, and even determine the interstellar extinction curve. The framework allows us to train such foundation models on large cross-survey, multidomain astronomical data sets with a huge amount of missing data due to the different footprints of the surveys. This demonstrates that building and training a single foundation model without fine-tuning using data and parameters from multiple surveys to predict unmeasured observations and parameters is well within reach. Such 'Large Astronomy Models' trained on large quantities of observational data will play a large role in the analysis of current and future large surveys.

Posters / 56

Classifying X-ray sources with Supervised Machine Learning: Challenges and Solutions

Author: Hui Yang[1]

Co-authors: Oleg Kargaltsev [1]; Jeremy Hare [2]; Steven Chen [1]; Igor Volkov [3]; Blagoy Rangelov [4]; Yichao Lin [1]

[1] *The George Washington University*

[2] *NASA GSFC*

[3] *Whitespace*

[4] *Texas State University*

Corresponding Author: huiyang@gwmail.gwu.edu

Millions of serendipitous X-ray sources have been discovered by modern X-ray observatories like Chandra, XMM-Newton, and recently eROSITA. For the vast majority of Galactic X-ray sources the nature is unknown. We have developed a multiwavelength machine-learning (ML) classification pipeline (MUWCLASS) that uses the random forest algorithm to quickly perform classifications of a large number of sources to learn about their astrophysical nature. This approach enables quick follow-up observations of interesting sources and population studies of various kinds. MUWCLASS has been applied to Chandra Source Catalog and XMM-DR13 catalog, augmented with multiwavelength properties obtained by cross-matching to surveys performed at other wavelengths. In this talk, I will demonstrate and discuss some common obstacles encountered in supervised ML (e.g., biases between training data and unclassified data, imbalanced training data, missing values, high-dimensionality) in the context of X-ray source classification. I will also present recent developments we have implemented to address some of those issues (e.g., astrophysically-motivated oversampling, accounting for feature uncertainties and absorption/extinction biases, probabilistic cross-matching and probabilistic class inference).

Posters / **57**

Artificial Intelligence at the Service of Space Astrometry: A New Way to Explore the Solar System

Author: giulio quaglia[1]

Co-authors: Guillaume Tochon [2]; Valery Lainey [1]

[1] *Observatoire de Paris, Paris*

[2] *EPITA, Paris*

Corresponding Authors: giulioquaglia1@gmail.com, guillaume.tochon@lrde.epita.fr, lainey@imcce.fr

Advancements in artificial intelligence (AI) have opened new horizons for space exploration, particularly in the domain of astrometry. This research investigates the integration of AI techniques, specifically deep neural networks, with space astrometry using the Cassini-Huygens images database. The primary objective is to establish a robust algorithm for the detection and classification of astronomical sources, in order to process them for a better understanding of the solar system and the possible discovery of new satellites around Saturn.
The methodology employed involves a multi-step process. Firstly known stars and satellites' positions are located in the images using ARAGO, a software package designed for the astrometric measurement of natural satellite positions in images taken using the Imaging Science Subsystem (ISS) of the Cassini spacecraft. A personalised detection system, using classical image processing techniques such as mathematical morphology, is then applied to identify all the bright sources, subsequently forming a labeled database for every image including source positions, bounding boxes and corresponding classes—divided in stars, satellites, cosmic rays, and suspicious objects (which could be uncatalogued satellites or stars). The database is used to train a YOLOv5 architecture, customized for small object detection, enabling the accurate identification and classification of sources within Cassini images.
Initial analysis shows promising results, with some room for improvement due to the challenging task of detecting and classifying sources of a few pixels only.
The implications of this research are far-reaching. Spatial and temporal characterizations of cosmic rays around Saturn's magnetosphere could lead to new insights into the behavior of high-energy particles. Moreover, the detection of new small satellites orbiting Saturn holds the potential to better understand the formation and evolution of the solar system.
Beyond Saturn, the proposed methodology can be adapted to the Jupiter system using data from the upcoming JUICE mission, broadening its applicability to a wider astronomical context.
In conclusion, the fusion of artificial intelligence and space astrometry, as demonstrated in this study, introduces a promising paradigm for the exploration of the universe and in particular our solar system.

Contributed talks / **58**

Subhalo effective density slope measurements from HST strong lensing data with neural likelihood-ratio estimation

Authors: Atinç Çagan Şengül[1]; Cora Dvorkin[1]; Gemma Zhang[1]

[1] *Harvard University*

Corresponding Authors: yzhang7@g.harvard.edu, sengul@g.harvard.edu, cdvorkin@g.harvard.edu

The CDM model is in remarkable agreement with large-scale observations but small-scale evidence remains scarce. Studying substructure through strong gravitational lensing can fill in the gap on small scales. In the upcoming years, we expect the number of observed strong lenses to increase by several orders of magnitude from ongoing and future surveys. Machine learning has the potential to optimally analyze these images, but its application to real observations remains limited. I will present the first application of machine learning to the analysis of subhalo properties in real strong lensing observations. Our work leverages a neural simulation-based inference technique in order to

infer the density slopes of subhalos. I will compare our method's prediction on HST images to the expected CDM measurements and discuss the implication of our work.

Contributed talks / **59**

DeepSZSim: Fast Simulations of the Thermal Sunyaev–Zel'dovich Effect in Galaxy Clusters for Simulation-based Inference

Author: Eve Vavagiakis[1]

Co-authors: Brian Nord [2]; Brian Zhang [3]; Camille Avestruz [4]; Elaine Ran [1]; Hanzhi Tan [3]; Ioana Cristescu [5]; Kush Banker [3]; Samuel McDermott [3]

[1] *Cornell University*

[2] *Fermilab*

[3] *University of Chicago*

[4] *University of Michigan*

[5] *University of Richmond*

Corresponding Authors: ev66@cornell.edu, samueldmcdermott@gmail.com, nord@fnal.gov

Simulations of galaxy clusters that are well-matched to upcoming data sets are a key tool for addressing systematics (e.g., cluster mass inference) that limit current and future cluster-based cosmology constraints. However, most state-of-the-art simulations are too computationally intensive to produce multiple versions of relevant physics systematics. We present DeepSZSim, a lightweight framework for generating simulations of Sunyaev–Zel'dovich (SZ) effect clusters based on average thermal pressure profile models. These simulations offer a fast and flexible method for generating large datasets for testing mass inference methods like machine learning and simulation-based inference. We present these simulations and their place within the larger Deep Skies nexus of versatile, multi-wavelength galaxy cluster and cosmic microwave background simulators. We discuss progress and prospects for using these SZ simulations for machine learning, including simulation-based inference of cluster mass.

Posters / **60**

Exploring the Link Between the Star Formation History and the Morphology of Galaxies Using CNNs

Author: Juan Pablo Alfonzo[1]

Co-authors: Kartheik Iyer [2]; Masayuki Akiyama [1]

[1] *Tohoku University*

[2] *Columbia University*

Corresponding Author: juanpabloalfonzo@astr.tohoku.ac.jp

We study the connection between the factors regulating star formation in galaxies on different spatial and temporal scales and connect morphological features (such as bars, bulges and spiral arms) with their integrated star formation on different timescales. This is being done using machine learning methods, specifically using convolutional neural networks (CNNs). The network is trained on a subset of galaxies in the SDSS-IV MaNGA DR17 (0<z<0.1) (N˜10,010) .The CNN network is trained to predict SFR, stellar mass and the t50 of galaxies. Furthermore, we use the network prediction to construct the galaxies'star formation histories using the dense_basis SED fitting algorithm. The target values are taken from the schema data of the SDSS surveys which use more traditional methods (i.e spectral fitting) to acquire the values of the parameters for each galaxy. Additionally, we

explore the use of transfer learning on the ResNet50 architecture with pretrained weights from ImageNet. We focus on interpretability of the trained network using various XAI methods such as SHAP (SHapley Additive exPlanations), to see what parts of galaxy images the network is focusing on to make its predictions. With this, we can explore which morphological features of galaxies have the greatest impact on predicted star formation history parameters, and use it to gain insights on the links between the underlying physical processes regulating star formation in galaxies.

Contributed talks / 61

CNNs reveal crucial degeneracies in strong lensing subhalo detection

Author: Conor O'Riordan[1]

Co-author: Simona Vegetti [1]

[1] *Max Planck Institute for Astrophysics*

Corresponding Author: conor@mpa-garching.mpg.de

Strong gravitational lensing has become one of the most important tools for investigating the nature of dark matter (DM). This is because it can be used to detect dark matter subhaloes in the environments of galaxies. The existence of a large number of these subhaloes is a key prediction of the most popular DM model, cold dark matter (CDM). With a technique called *gravitational imaging*, the number and mass of these subhaloes can be measured in strong lenses, constraining the underlying DM model.

Gravitational imaging however is an expensive method. This is mostly due to the final stage of the analysis: so-called *sensitivity mapping*. Here, the observation is analysed to find the smallest detectable subhalo in each pixel. This information can be used to turn a set of subhalo detections and non-detections into an inference on the dark matter model. We have previously introduced a machine learning technique that uses a set of large convolutional neural networks (CNNs) to replace the expensive sensitivity mapping stage [1]. We exploited this new technique to test the sensitivity of *Euclid* strong lenses to dark matter subhaloes. Analysing 16,000 simulated *Euclid* strong lens observations we found that subhaloes with mass larger than $M > 10^{8.8\pm0.2} M_{\odot}$ could be detected at 3σ in that data, and that the entire survey should yield ~ 2500 new detections.

In the current work, we take our method much further to understand a crucial systematic uncertainty in subhalo detection: the angular structure of the lens mass model. We train an ensemble of CNNs to detect subhaloes in highly realistic HST images. The models use an increasing amount of angular complexity in the lensing galaxy mass model, parametrised as an elliptical power-law plus multipole perturbations up to order 4, and external shear. Multipole perturbations allow for boxy/discy structure in the lens galaxy. This is commonly found in the light profiles of elliptical galaxies but is almost always missing from the mass profile in strong lensing studies.

We find that multipole perturbations up to 1 per cent are large enough to cause false positive subhalo detections at a rate of 20 per cent, with order 3 perturbations having the strongest effect. We find that the area in an observation where a subhalo can be detected drops by a factor of 10 when multipoles up to an amplitude of 3 per cent are allowed in the mass model. However, the mass of the smallest subhalo that can be detected does not change, with a detection limit of $M > 10^{8.2} M_{\odot}$ found at 5σ regardless of model choice. Assuming CDM, we find that HST observations modelled without multipoles should yield a detectable subhalo in 4.8 per cent of cases. This drops to 0.47 per cent when the lenses are modelled with multipoles up to 3 per cent amplitude. The loss of expected detections is due to the effect of the previously detectable objects being consistent with multipoles of that strength. To remain reliable, **strong lensing analyses for dark matter subhaloes must therefore include angular complexity beyond the elliptical power-law**.

[1] O'Riordan C. M., Despali, G., Vegetti, S., Moliné, Á., Lovell, M., MNRAS **521**, 2342 (2023)

Contributed talks / 62

Selection functions of strong lens finding neural networks

Author: Aniruddh Herle[1]

Co-authors: Conor O'Riordan [2]; Simona Vegetti [2]

[1] *Max-Planck Institute for Astrophysics, + LMU*

[2] *Max Planck Institute for Astrophysics*

Corresponding Authors: conor@mpa-garching.mpg.de, aniruddh.herle@gmail.com

Convolutional neural networks (CNNs) are now the standard tool for finding strong gravitational lenses in imaging surveys. Upcoming surveys like Euclid will rely completely on CNNs for strong lens finding but the selection function of these CNNs has not yet been studied. This is representative of the large gap in the literature in the field of machine learning applied to astronomy. Biases in CNN lens finders have the potential to influence the next generation of strong lens science unless properly accounted for. In our work we have quantified, for the first time, this selection function. We also explore the implications of these biases for various strong lens science goals.

We find that CNNs with similar architecture and training data as is commonly found in the lens finding literature are biased classifiers. We use three training datasets, representative of those used to train galaxy-galaxy and galaxy-quasar lens finding neural networks. The networks preferentially select systems with larger Einstein radii, as in this case the source and lens light is most easily disentangled. Similarly, the CNNs prefer large sources with more concentrated source-light distributions, as they are more distinct from the extended lens light.

The model trained to find lensed quasars shows a stronger preference for higher lens ellipticities than those trained to find lensed galaxies. The selection function is independent of the slope of the power-law of the mass profiles, hence measurements of this quantity will be unaffected. We find that the lens finder selection function reinforces the lensing cross-section. In general, we expect our findings to be a universal result for all galaxy-galaxy and galaxy-quasar lens finding neural networks.

Based on work in Herle A., O'Riordan C. M., Vegetti, S., arXiv:2307.10355, submitted to MNRAS. arXiv submission

Contributed talks / 63

Machine learning as a key component in the science processing pipelines of space- and ground-based surveys?

Author: Jeroen Audenaert[1]

[1] *MIT*

Corresponding Author: jeroena@mit.edu

Machine learning is becoming an essential component of the science operations processing pipelines of modern astronomical surveys. Space missions such as NASA's Transiting Exoplanet Survey Satellite (TESS) are observing millions of stars each month. In order to select the relevant targets for our science cases from these large numbers of observations, we need highly automated and efficient classification methods. Only afterwards, more detailed astrophysical studies can be done to derive the physical parameters of the selected stars. Given the increasing data volumes, machine learning techniques, and in particular physically interpretable machine learning models, prove to be the ideal instruments to achieve this. In this talk, I will draw from our experiences in developing the TESS Data for Asteroseismology (T'DA) machine learning classification pipeline to (i) discuss the challenges and opportunities associated to the development of such pipelines, (ii) share our insights with regard to the used machine learning techniques and identify where they could be improved,

and (iii) give an outlook of how machine learning could be incorporated into the science processing pipelines of ground- and space-based surveys.

Posters / 64

Learning the Reionization History from High-z Quasar Damping Wings with Simulation-based Inference

Author: Timo Kist[1]

Co-author: Joseph F. Hennawi [2]

[1] *Leiden Observatory*

[2] *Leiden Observatory, UC Santa Barbara*

Corresponding Authors: hennawi@strw.leidenuniv.nl, kist@strw.leidenuniv.nl

The damping wing signature of high-redshift quasars in the intergalactic medium (IGM) provides a unique way of probing the history of reionization. Next-generation surveys will collect a multitude of spectra that call for powerful statistical methods to constrain the underlying astrophysical parameters such as the global IGM neutral fraction as tightly as possible. Inferring these parameters from the observed spectra is challenging because non-Gaussian processes such as IGM transmission causing the damping wing imprint make it impossible to write down the correct likelihood of the spectra.
We will present a tractable Gaussian approximation of the likelihood that forms the basis of a fully differentiable Hamiltonian Monte-Carlo inference scheme implemented in JAX. Our scheme can be readily applied to real observational data and is based on realistic forward-modelling of high-redshift quasar spectra including IGM transmission and heteroscedastic observational noise. In contrast to most previous approaches, we do not only use the smooth part of the spectrum redward of the Lyman-alpha line to infer the quasar continuum but also the information encoded in the Lyman-alpha forest, taking into account the full covariance between the red and the blue part of the spectrum.
We improve upon our Gaussian likelihood approximation by learning the true likelihood with a simulation-based version of the inference scheme. To this end, we train a normalizing flow as neural likelihood estimator as well as a binary classifier as likelihood ratio estimator and incorporate them into our inference pipeline.
We provide a full reionization forecast for Euclid by applying our procedure to a set of realistic mock observational spectra resembling the distribution of Euclid quasars and realistic spectral noise. By inferring the IGM neutral fraction as a function of redshift, we show that our method applied to upcoming observational data can robustly constrain its evolution up to ~5% at all redshifts between $6 < z < 11$.

Posters / 65

Domain Adaptation in Gravitational Lens Analysis

Authors: Paxson Swierc[None]; Yifan(Megan) Zhao[None]

Co-authors: Aleksandra Ciprijanovic [1]; Brian Nord [2]

[1] *Fermi National Accelerator Laboratory*

[2] *Fermilab*

Corresponding Authors: yifanzyf@uchicago.edu, pswierc@uchicago.edu, aleksand@fnal.gov, nord@fnal.gov

Upcoming surveys are predicted to discover galaxy-scale strong lenses on the magnitude of 10^5, making deep learning methods necessary in lensing data analysis. Currently, there is insufficient

real lensing data to train deep learning algorithms, but training only on simulated data results in poor performance on real data. Domain adaptation can bridge the gap between simulated and real datasets. We adopt domain adaptation on the estimation of Einstein radius in simulated galaxy-scale gravitational lensing images. We evaluate two domain adaptation techniques - domain adversarial neural networks (DANN) and maximum mean discrepancy (MMD). We train on a source domain of simulated lenses and apply it to a target domain with emulation of DES survey conditions. We show that both domain adaptation techniques can significantly improve the model performance on the more complex target domain datasets. Our results show the potential of using domain adaptation to perform analysis on future survey data with a deep neural network trained on simulated data.

Contributed talks / 66

Perturbation theory emulator for cosmological analysis

Author: Svyatoslav Trusov[1]

[1] *LPNHE*

Corresponding Author: strusov@lpnhe.in2p3.fr

The data from the new generation of cosmological surveys, such as DESI (DESI Collaboration et al. 2022), have already started taking data, and even more will arrive with Euclid (Laureijs et al. 2011) and the LSST of Vera Rubin Observatory (Ivezić et al. 2019) starting soon. At the same time, the classical methods of analysing RSD and BAO with 2-point statistics provide less strenuous constraints than for example a full-modelling analysis (Ivanov et al. 2020). Such an analysis does however require much more computational power.

We present an emulator based on the feedforward neural network which allows us to significantly speed up analytical computations of the 2-point statistics in both Fourier and configuration space (Trusov et al. in prep). Our approach is based on emulating the perturbation theory (PT) quantities, which are later combined with bias terms to produce the non-linear prediction of the 2-point statistics for any galaxy sample. We compare the performance of our approach against publicly available PT codes using mocks based on the AbacusSummit simulations (Maksimova et al. 2021, Garrison et al. 2021), where our tool performs significantly faster without any noticeable loss of precision.

Contributed talks / 67

Field-level Emulator within Bayesian Origin Reconstruction from Galaxies (BORG)

Author: Ludvig Doeser[1]

[1] *Stockholm University*

Corresponding Author: ludvig.doeser@fysik.su.se

Unlocking the full potential of next-generation cosmological data requires navigating the balance between sophisticated physics models and computational demands. We propose a solution by introducing a machine learning-based field-level emulator within the HMC-based Bayesian Origin Reconstruction from Galaxies (BORG) inference algorithm. The emulator, an extension of the first-order Lagrangian Perturbation Theory (LPT), achieves remarkable accuracy compared to N-body simulations while significantly reducing evaluation time. Leveraging its differentiable neural network architecture, the emulator enables efficient sampling of the high-dimensional space of initial conditions. To demonstrate its efficacy, we use the inferred posterior samples of initial conditions to run constrained N-body simulations, yielding highly accurate present-day non-linear dark matter fields compared to the underlying truth used during inference.

Contributed talks / 68

Fast realistic, differentiable, mock halo generation for wide-field galaxy surveys

Author: Simon Ding[1]

[1] *Institut d'Astrophysique de Paris (IAP)*

Corresponding Author: simon.ding@iap.fr

Accurately describing the relation between the dark matter over-density and the observable galaxy field is one of the significant challenges to analyzing cosmic structures with next-generation galaxy surveys. Current galaxy bias models are either inaccurate or computationally too expensive to be used for efficient inference of small-scale information.
In this talk, I will present a hybrid machine learning approach called the Neural Physical Engine (NPE) that addresses this problem. The network architecture, first developed and tested by Charnock et al. (2020), exploits physical information of the galaxy bias problem and is suitable for zero-shot learning within field-level inference approaches.
Furthermore, the model can efficiently generate mock halo catalogues on the scales of wide-field surveys such as Euclid. Finally, I will also show that those generated mocks are consistent with full phase-space halo finders, including the 2-point correlation function.

Contributed talks / 69

Deep Learning Generative Models to Infer Mass Density Maps from SZ, X-ray and Galaxy Members Observations in Galaxy Clusters

Author: Daniel de Andrés[1]

Co-authors: Gustavo Yepes [2]; Marco De Petris [3]; Weiguang Cui [4]

[1] *Universidad Autónoma de Madrid*

[2] *UAM*

[3] *Sapienza Universita di Roma*

[4] *UAM and University of Edinburgh*

Corresponding Author: daniel.deandres@uam.es

In our previews works, e.g., arXiv:2209.10333, deep learning techniques have succeeded in estimating galaxy cluster masses in observations of Sunyaev Zel'dovich maps, e.g. in the Planck PSZ2 catalog and mass radial profiles from SZ mock maps. In the next step, we explore inferring 2D mass density maps from mock observations of SZ, X-ray and stars using THE THREE HUNDRED (The300) cosmological simulation. In order to do that, we investigate state-of-the-art deep learning models that have been proven to be successful for image generation in multiple areas of research including astrophysics and medical imaging. These models are conditioned to observations, e.g. SZ maps, to generate the most likely matter 2D distribution given our dataset, composed of around 140 thousand mock maps from The300. We show that these models can successfully infer the 2D matter distribution with a scatter of around 14% in their pixel distribution and reproduce the matter power spectrum when comparing the generated maps with the ground-truth from the simulations. One of the main advantages of these generative models, is that they can effectively combine several inputs views and extract the useful features of each of them to infer mass density maps. By combining SZ, X-ray and stars in a multichannel approach, the scatter is reduced by a factor of 2 in comparison with the scatter that is computed when considering only the single-view models.

The next natural step of this project is to apply DL models on high resolution SZ observation, such as NIKA2, SPT and ACT. However, mock images needed for training deep learning models must

fully take into consideration the observational impact of the telescopes in order to mimic real observations.

Posters / **70**

Galaxy cluster detection on SDSS images using deep machine learning

Authors: Kirill Grishin[1]; Simona Mei[2]; Stephane Ilic[3]

[1] *Astroparticule et Cosmologie (APC)*

[2] *AstroParticule et Cosmologie (APC)*

[3] *IJCLab, Université Paris-Saclay*

Corresponding Authors: ilic@ijclab.in2p3.fr, mei@apc.in2p3.fr, kirillg6@gmail.com

Galaxy clusters are a powerful probe of cosmological models. Next generation large-scale optical and infrared surveys will reach unprecedented depths over large areas and require highly complete and pure cluster catalogs, with a well defined selection function. We have developed a new cluster detection algorithm YOLO-CL, which is a modified version of the state-of-the-art object detection deep convolutional network YOLO, optimized for the detection of galaxy clusters (Grishin, Mei, Ilic 2023). We trained YOLO-CL on color images of the redMaPPer cluster detections in the SDSS. We find that YOLO-CL detects 95–98% of the redMaPPer clusters, with a purity of 95–98% calculated by applying the network to SDSS blank fields. When compared to the MCXC2021 X-ray catalog in the SDSS footprint, YOLO-CL is more complete then redMaPPer, which means that the neural network improved the cluster detection efficiency of its training sample: it detects 98% of clusters with mean X-ray surface brightness of 20×10⁻¹⁵ erg/s/cm²/arcmin² while redMaPPer is 98% complete above 55×10⁻¹⁵ erg/s/cm²/arcmin². The YOLO-CL selection function is approximately constant with redshift, with respect to the MCXC2021 cluster mean X-ray surface brightness. YOLO-CL shows high performance when compared to traditional detection algorithms applied to SDSS. Deep learning networks benefit from a strong advantage over traditional galaxy cluster detection techniques because they do not need galaxy photometric and photometric redshift catalogs. This eliminates systematic uncertainties that can be introduced during source detection, and photometry and photometric redshift measurements. Our results show that YOLO-CL is an efficient alternative to traditional cluster detection methods. In general, this work shows that it is worth exploring the performance of deep convolutional networks for future cosmological cluster surveys, such as the Rubin/LSST, Euclid or the Roman Space Telescope surveys.

Contributed talks / **71**

Latent space out-of-distribution detection of galaxies for deblending in weak lensing surveys

Author: Jelle Mes[1]

[1] *Leiden Observatory, Leiden University, The Netherlands*

Corresponding Author: mes@strw.leidenuniv.nl

Upcoming photometric surveys such as the Legacy Survey of Space and Time (LSST) will image billions of galaxies, an amount required for extracting the faint weak lensing signal at a large range of cosmological distances. The combination of depth and area coverage of the imagery will be unprecedented ($r \sim 27.5$, $\sim 20\,000\,\mathrm{deg}^2$), and processing it will be fraught with many challenges. One of the most pressing issues is the fact that roughly 50% of the galaxies will be "blended", where its projection on our detectors will overlap with other astronomical objects along the same line of

sight. Without appropriate "deblending"algorithms, the blends introduce an unacceptable error on the weak lensing signal.
Several deblending algorithms have emerged up the past years, of which the most promising are based on deep neural networks (DNNs). DNNs are known to be highly sensitive to a difference in the distributions of the training and validation datasets. As the true deblended image of a blend, needed for supervised learning, is in most cases unobtainable due to the line of sight projection, training data has to be generated algorithmically. This training data will by its nature have a limited coverage of the high dimensional space that spans all galaxies that will be observed with the LSST. In other words, many galaxies and blends observed by the LSST will be out of distribution (OOD) and the DNNs will perform poorly on them. We have developed a method to classify blends on being OOD or in-distribution (IID) based on the distribution of an input blend sample in the latent space of a β-VAE, compared to the latent space distribution of the training sample. We will present the results of the OOD flagging, demonstrating that the latent space is indeed a useful tool for identifying OOD samples. Furthermore, we will discuss the ensuing reduction on the error of shear and photometry measurements when rejecting OOD samples for the weak lensing analysis.
Though core components of our method build on an existing deblending algorithm by Arcelin et al. (2021), the addition of this successful OOD detection technique is essential for its proper functioning on future LSST imagery. The blends flagged as OOD can, in future pipelines, be separated from the IID blends to prevent contamination of the weak lensing signal or be deblended with a method specifically tuned to OOD blends.

Contributed talks / **72**

Likelihood-free Forward Modeling for Cluster Weak Lensing and Cosmology

Author: Sut Ieng Tam[1]

[1] *Institute of Astronomy and Astrophysics, Academia Sinica (ASIAA)*

Corresponding Author: sitam@asiaa.sinica.edu.tw

Likelihood-free inference provides a rigorous way to preform Bayesian analysis using forward simulations only. It allows us to account for complex physical processes and observational effects in forward simulations. In this work, we use Density-Estimation Likelihood-Free Inference (DELFI) to perform a likelihood-free forward modelling for Bayesian cosmological inference, which uses the redshift evolution of the cluster abundance together with weak-lensing mass calibration. The analysis framework developed in this study will be powerful for cosmological inference in relation to ongoing cluster cosmology programs, such as the XMM-XXL survey and the eROSITA all-sky survey, combined with wide-field weak-lensing surveys.
In this talk, I will first present the convergent solutions for the posterior distribution which employ the synthetic cluster catalogue generated from our forward model, and then I will show some preliminary results by applying this method to the HSC data.

Posters / **73**

Leveraging Machine Learning for Retrieving Exoplanet Atmosphere Parameters from the upcoming ARIEL Space Telescope Spectra

Authors: Ethan Tregidga[1]; Mayeul Aubin[2]

[1] *Center for Astrophysics, École polytechnique fédérale de Lausanne*

[2] *AstroAI @ Center for Astrophysics | Harvard & Smithsonian, Ecole Polytechnique*

Corresponding Authors: m.mayeul.aubin@gmail.com, ethan@treggal.com

The study of exoplanet atmospheres plays a vital role in understanding their composition. However, extracting accurate atmospheric parameters from transmission spectra poses significant challenges. Bayesian sampling algorithms, although effective, can be time-consuming and laborious. As an alternative, machine learning techniques offer promising avenues to expedite and enhance this process.

In this poster, I will present a new model we developed in the AstroAI group at the Center for Astrophysics which retrieves the atmospheric parameters of exoplanets for observations with the upcoming ARIEL space telescope. Our model is based on Normalising Flows, a machine-learning technique that allows us to generate probability distributions of the parameters for each observed spectrum, and thus gain valuable insights into the plausible compositions for each specific spectrum. This work won the 2023 ARIEL data challenge organized by the European Space Agency (ESA). To tackle this task, we put together an interdisciplinary team of experts in Machine Learning, Astronomy, Molecular Spectroscopy, and Exoplanetary Research.

Our model is based on Normalising Flows, a machine learning technique that allows us to generate probability distributions of the parameters for each observed spectrum, and thus gain valuable insights into the plausible compositions for each specific spectrum.

Through this interdisciplinary approach that merges astrophysics and machine learning, we aim to advance our understanding of exoplanet atmospheres and the use of machine learning tools in Simulated Based Inference in astrophysics. Our research showcases the capabilities of AI to revolutionize the analysis of exoplanetary data, preparing the ground for more efficient and accurate characterization of exoplanets in the future.

Contributed talks / **74**

Significance Mode Analysis (SigMA) for hierarchical structures

Author: Sebastian Ratzenböck[1]

Co-authors: João Alves [1]; Immanuel Bomze [1]; Torsten Möller [1]

[1] *University of Vienna*

Corresponding Author: sebastian.ratzenboeck@univie.ac.at

We present an innovative clustering method, Significance Mode Analysis (SigMA), to extract co-spatial and co-moving stellar populations from large-scale surveys such as ESA *Gaia*. The method studies the topological properties of the density field in the multidimensional phase space. The set of critical points in the density field gives rise to the cluster tree, a hierarchical structure in which leaves correspond to modes of the density function. Typically, however, non-parametric density estimation methods lead to an over-clustering of the input data. We propose an interpretable cluster tree pruning strategy by determining minimum energy paths between pairs of neighboring modes directly in the input space. We present a statistical hypothesis test that examines deviations from unimodality along these paths, which provides a measure of significance for each pair of clusters.
We apply SigMA to *Gaia* EDR3 data of the closest OB association to Earth, Scorpio-Centaurus (Sco-Cen), and find 37 co-moving clusters in Sco-Cen. These clusters are independently validated using astrophysical knowledge and, and to a certain extent, by their association with massive stars too bright for Gaia, both unknown to SigMA. Our findings suggest that the OB association is more actively star-forming and dynamically richer than previously thought. This application demonstrates that SigMA allows for an accurate census of young populations, quantify their dynamics, and reconstruct the recent star formation history of the local Milky Way.

Contributed talks / **75**

Modeling galaxy orientations on the SO(3) manifold with score-based generative models

Authors: Yesukhei Jagvaral[1]; Rachel Mandelbaum[1]; Francois Lanusse[2]

[1] *Carnegie Mellon University*

[2] *CNRS*

Corresponding Authors: rmandelb@andrew.cmu.edu, eeseight@gmail.com, francois.lanusse@cnrs.fr

Upcoming cosmological weak lensing surveys are expected to constrain cosmological parameters with unprecedented precision. In preparation for these surveys, large simulations with realistic galaxy populations are required to test and validate analysis pipelines. However, these simulations are computationally very costly – and at the volumes and resolutions demanded by upcoming cosmological surveys, they are computationally infeasible.
Here, we propose a Deep Generative Modeling approach to address the specific problem of emulating realistic 3D galaxy orientations in synthetic catalogs. For this purpose, we develop a novel Score-Based Diffusion Model specifically for the SO(3) manifold. The model accurately learns and reproduces correlated orientations of galaxies and dark matter halos that are statistically consistent with those of a reference high-resolution hydrodynamical simulation.

Posters / **76**

Galaxy Merger identification using the effect of low-surface-brightness features on the sky background measurement

Author: Luis Suelves[1]

[1] *NCBJ, Poland*

Corresponding Author: luis.suelves@ncbj.gov.pl

One of the main ongoing intersections of machine learning and astronomy is the classification of galaxy types such as merging galaxies in large-scale surveys. In this work, we built a class-balanced training dataset using SDSS DR6 galaxies classified in Galaxy Zoo DR1, where the mergers were visually confirmed galaxy pairs from Darg et al (2010). We wanted to test the potential of training a Neural Network (NN) using only photometric information, which led us to discover that the SDSS DR6 sky background error could provide a NN model with an accuracy of 92.64 ± 0.15 % in training and 92.36 ± 0.21 % in test. Moreover, we found out that this parameter could suffice to separate mergers by simply drawing a decision boundary in the g - r bands plane, obtaining a 91.59 % using all our data. We consider that this sky background error is sensitive to the low-surface-birghtness tidally-stripped material surrounding the merging sources. Present work is focused on applying this knowledge that was found with the aid of the NN, both to data outside of the training sample but also in SDSS DR6 + GZ DR1, and to a sample in the Subaru/HSC North Ecliptic Pole.

Posters / **77**

Neural Networks for Super Resolution of X-Ray Line Emission Mapper Images

Authors: Cecilia Garraffo[1]; Ethan Tregidga[2]; James Steiner[1]

Co-authors: AstroAI [1]; Tao Tsui [3]

[1] *Center for Astrophysics*

[2] *Center for Astrophysics, École polytechnique fédérale de Lausanne*

[3] *Harvard*

Corresponding Authors: jiataocui@gmail.com, cgarraffo@cfa.harvard.edu, astroai@cfa.harvard.edu, james.steiner@cfa.harvard.edu, ethan@treggal.com

The line emission mapper (LEM) is a proposed X-ray probe for high spectral resolution survey observations targeting galaxies and clusters of galaxies to characterise the circumgalactic and intergalactic medium better. The mission will use a microcalorimeter array with 1-2 eV resolution, capturing individual emission lines and offering the ability to spatially map elemental emission within galaxies, supernovae remnants, and more. We propose two methods of machine learning super-resolution to enhance LEM's capabilities and bring the advantages of LEM to archival data. The first project will improve the spatial resolution of LEM by leveraging the high spatial resolution of Chandra and training a network to interpolate features from low-dimensional to high-dimensional space. The second project will apply machine learning to LEM observations to infer high-spectral resolution results from the vast archive of low-spectral resolution Chandra data. As LEM is still in development, we are presently working with simulated data sets for galaxies from which we generate mock observations with Chandra, LEM and a theoretical high spatial resolution version of LEM for training the network.

Posters / **78**

Determining Physical Parameters of Serendipitous Sources using AI

Authors: Cecilia Garraffo[1]; Ethan Tregidga[2]; Gerrit Schellenberger[1]; Joshua Wing[1]

Co-author: AstroAI [1]

[1] *Center for Astrophysics*

[2] *Center for Astrophysics, École polytechnique fédérale de Lausanne*

Corresponding Authors: jwing@cfa.harvard.edu, gerrit.schellenberger@cfa.harvard.edu, cgarraffo@cfa.harvard.edu, astroai@cfa.harvard.edu, ethan@treggal.com

Galaxy groups are gravitationally bound structures composed of galaxies and a hot X-ray-emitting gas that envelops the entire group. These systems are balanced with gravitational potential pulling inwards and thermal pressure from the hot gas pushing outwards. Questions remain about how this balance is altered when galaxies within the group undergo periods of star formation or when supermassive black holes at the centers of some group galaxies become active and outburst. Identifying galaxy groups at very large distances with the next generation of X-ray observatories can help to answer these questions. Doing so with conventional methods is computationally expensive. We are using simulated data from the line emission mapper (LEM), a next-generation X-ray probe for high spectral resolution survey observations targeting galaxies and clusters of galaxies to characterise the circumgalactic and intergalactic medium better. Our initial results from training a CNN accurately and quickly identify galaxy groups' distance, age, mass, and chemical composition based on simulated high-resolution X-ray spectra. This allows us to serendipitously identify galaxy groups in the background of observations of other astronomical sources.

Contributed talks / **79**

Cosmological constraints from HSC survey first-year data using deep learning

Author: Zoltan Haiman[1]

[1] *Columbia University*

Corresponding Author: zoltan@astro.columbia.edu

We present cosmological constraints from the Subaru Hyper Suprime-Cam (HSC) first-year weak lensing shear catalogue using convolutional neural networks (CNNs) and conventional summary statistics. We crop 19 $3 \times 3\mathrm{deg}^2$ sub-fields from the first-year area, divide the galaxies with redshift $0.3 < z < 1.5$ into four equally-spaced redshift bins, and perform tomographic analyses. We develop a pipeline to generate simulated convergence maps from cosmological N-body simulations, where we account for effects such as intrinsic alignments (IAs), baryons, photometric redshift errors, and point spread function errors, to match characteristics of the real catalogue. We train CNNs that can predict the underlying parameters from the simulated maps, and we use them to construct likelihood functions for Bayesian analyses. In the Λ cold dark matter model with two free cosmological parameters Ω and σ_8, we find $\Omega = 0.278^{+0.037}_{-0.035}$, $S_8 \equiv (\Omega/0.3)^{0.5}\sigma_8 = 0.793^{+0.017}_{-0.018}$, and the IA amplitude $A_{\rm IA} = 0.20^{+0.55}_{-0.58}$. In a model with four additional free baryonic parameters, we find $\Omega = 0.268^{+0.040}_{-0.036}$, $S_8 = 0.819^{+0.034}_{-0.024}$, and $A_{\rm IA} = -0.16^{+0.59}_{-0.58}$, with the baryonic parameters not being well-constrained. We also find that statistical uncertainties of the parameters by the CNNs are smaller than those from the power spectrum (5-24 percent smaller for S_8 and a factor of 2.5-3.0 smaller for Ω), showing the effectiveness of CNNs for uncovering additional cosmological information from the HSC data. With baryons, the S_8 discrepancy between HSC first-year data and Planck 2018 is reduced from $\sim 2.2\,\sigma$ to $0.3 - 0.5\,\sigma$.

Posters / 80

Comparing Automated Posterior Estimation Techniques for Modeling Strong Lenses In Ground-based Survey Data

Authors: Jason Poh[1]; Ashwin Samudre[2]; Aleksandra Ćiprijanović[3]; Brian Nord[3]

Co-authors: Gourav Khullar [4]; Dimitrios Tanoglidis [5]; Joshua Frieman [1]

[1] *University of Chicago*

[2] *Simon Fraser*

[3] *Fermilab*

[4] *University of Pittsburgh*

[5] *University of Pennsylvania*

Corresponding Authors: jasonpoh@uchicago.edu, gourav.khullar@pitt.edu, ashwin.samudre@gmail.com, dtanogli@sas.upenn.edu, jfrieman@uchicago.edu, aleksand@fnal.gov, nord@fnal.gov

Current and future ground-based cosmological surveys, such as the Dark Energy Survey (DES), and the Vera Rubin Observatory Legacy Survey of Space and Time (LSST), are predicted to discover thousands to tens of thousands of strong gravitational lenses. The large number of strong lenses discoverable in future surveys will make strong lensing a highly competitive and complementary cosmic probe. However, conventional lens modeling techniques are unable to scale up to the sheer number of lenses that will be discovered through upcoming surveys. Therefore, the use of automated lens analysis techniques is necessary. We demonstrate that machine learning methods can be used to automate the inference of informative model posteriors of strong lensing systems in ground-based surveys with credible uncertainty estimation. We present two Simulation-Based Inference (SBI) approaches for lens parameter estimation of galaxy-galaxy lenses. We demonstrate applications of Neural Posteriors Estimators (NPEs) and Bayesian Neural Network (BNNs) to automate the inference of a 12-parameter lensing system for DES-like ground-based imaging data. We apply a suite of diagnostics (e.g., posterior coverage and SBC) to validate the performance of our methods. We find that NPEs outperform the BNN, producing posterior distributions that are for the most part both more accurate and more precise; in particular, several source-light model parameters are systematically biased in the BNN implementation.

Posters / 81

Probing Supermassive Black Hole-Host Galaxy Scaling Relations in Cosmological Simulations with Machine Learning

Author: Yuan Li[1]

[1] *University of North Texas*

Corresponding Author: yuan.astro@gmail.com

Observations have established intriguing correlations between supermassive black holes (SMBHs) and their host galaxies. However, state-of-the-art cosmological simulations have revealed discrepancies in the slope, amplitude, and scatter of the scaling relations when compared to both observational data and among different simulations. Understanding the underlying physical mechanisms responsible for these scaling relations remains a challenging task, although it has been demonstrated that SMBH feedback plays a crucial role in shaping them within simulations. In this study, we conduct a comprehensive analysis of SMBH-host scaling relations in three cosmological simulations, namely Illustris, TNG, and EAGLE. Leveraging the power of machine learning techniques, we quantify the tightness of each scaling relation across the different simulations. We find the M-sigma relation to be the tightest scaling relation across simulations except for TNG. EAGLE exhibits more scattered scaling relations overall. Additionally, we explore the dependence of scatter on SMBH mass and investigate SMBH-host "fundamental planes." Our analysis sheds light on the coevolution of SMBHs and galaxies in cosmological simulations with different SMBH feeding and feedback implementations. Our work also paves way for future studies to connect observations and simulations, and provide constraints for theoretical models.

Posters / 82

Data-Driven Discovery: Machine Learning for the Detection and Characterization of X-ray Transients

Author: Steven Dillmann[1]

Co-authors: Rafael Martínez-Galarza [2]; Rosanne Di Stefano [2]; Vinay Kashyap [2]

[1] *University of Cambridge*

[2] *Smithsonian Astrophysical Observatory*

Corresponding Authors: distefano.rosanne@gmail.com, jmartine@cfa.harvard.edu, steven.dillmann18@imperial.ac.uk, vkashyap@cfa.harvard.edu

Recent serendipitous discoveries in X-ray astronomy such as extragalactic fast X-ray transients, Quasi-periodic eruptions, extroplanetary transits, and other rare short-duration phenomena in the X-ray sky highlight the importance of a systematic search for such events in X-ray archives. Variable-length time series data in form of X-ray eventfiles present a challenge for the identification of characteristic features of these time-domain anomalies with machine learning applications. Novel equal-length data representations of X-ray eventfiles capturing both time and energy information are introduced. We use these eventfile representations as features for an unsupervised X-ray transient detection pipeline involving principal component analysis or autoencoder feature learning followed by dimensionality reduction and clustering. The association of these clusters with previously identified transients produces a new set of X-ray transient candidates. Supervised regression and classification models are trained to characterize and predict the time-domain and spectral properties of X-ray eventfiles. We find 8956 X-ray transient candidates in the Chandra archive including a confirmed eclipsing low-mass X-ray binary system and a potential accretion-powered X-ray pulsar. The developed data science tools and catalog of X-ray transient candidates are made publicly available for the advancement of data-driven discoveries in the X-ray astronomy community.

Contributed talks / 83

Reconstruction of cosmological initial conditions with sequential simulation-based inference

Author: Oleg Savchenko[1]

[1] *University of Amsterdam*

Corresponding Author: savchenkooleg42@gmail.com

Knowledge of the primordial matter density field from which the present non-linear observations formed is of fundamental importance for cosmology, as it contains an immense wealth of information about the physics, evolution, and initial conditions of the universe. Reconstructing this density field from the galaxy survey data is a notoriously difficult task, requiring sophisticated statistical methods, advanced cosmological simulators, and exploration of a multi-million-dimensional parameter space. In this talk, I will discuss how Gaussian Autoregressive Neural Ratio Estimation (a recent approach in simulation-based inference) allows us to tackle this problem and sequentially obtain data-constrained realisations of the primordial dark matter density field in a simulation-efficient way for general non-differentiable simulators. In addition, I will describe how graph neural networks can be used to get optimal data summaries for galaxy maps, and how our results compare to those obtained with classical likelihood-based methods such as Hamiltonian Monte Carlo.

Posters / **84**

Unlocking fast cosmological parameter inference from Euclid with Marginal Neural Ratio Estimation

Author: Guillermo Franco Abellán[1]

[1] *GRAPPA Institute, University of Amsterdam*

Corresponding Author: g.francoabellan@uva.nl

The Euclid space telescope will measure the shapes and redshifts of billions of galaxies, probing the growth of cosmic structures with an unprecedented precision. However, the increased quality of these data also means a significant increase in the number of nuisance parameters, making the cosmological inference a very challenging task. In this talk, I discuss the first application of Marginal Neural Ration Estimation (MNRE) (a recent approach in so-called simulation-based inference) to Euclid primary observables, like cosmic shear and galaxy-clustering spectra. Using expected Euclid experimental noise, I show how it's possible to recover the posterior distribution for the cosmological parameters using an order of magnitude fewer simulations than conventional likelihood-based methods. This result supports that MNRE is a powerful framework to analyse Euclid data, allowing to extend the model complexity beyond what its currently achievable with standard MCMC.

Contributed talks / **85**

Domain Adaptive Graph Neural Networks for Constraining Cosmological Parameters Across Multiple Data Sets

Author: Andrea Roncoli[1]

Co-authors: Aleksandra Ciprijanovic [1]; Brian Nord [2]; Francisco Villaescusa-Navarro [3]; M Voetberg [2]

[1] *Fermi National Accelerator Laboratory*

[2] *Fermilab*

[3] *Flatiron Institute*

Corresponding Authors: andrea.roncoli.cs@gmail.com, aleksand@fnal.gov, maggiev@fnal.gov, nord@fnal.gov, fvillaescusa@flatironinstitute.org

State of the art astronomical simulations have provided datasets which enabled the training of novel deep learning techniques for constraining cosmological parameters. However, differences in subgrid physics implementation and numerical approximations among simulation suites lead to differences in simulated datasets, which pose a hard challenge when trying to generalize across diverse data domains and ultimately when applying models to observational data.
Recent work reveals deep learning algorithms are able to extract more information from complex cosmological simulations than summary statistics like power spectra. We introduce Domain Adaptive Graph Neural Networks (DA-GNNs), trained on CAMELS data, inspired by CosmoGraphNet (Villanueva-Domingo et al 2023). By utilizing GNNs, we can capitalize on their capacity to capture both astrophysical and topological features of galaxy distributions. Mixing these capabilities with domain adaptation techniques such as Maximum Mean Discrepancy (MMD), which enable extraction of domain-invariant features, our framework demonstrates enhanced accuracy and robustness. We present experimental results, including the alignment of distributions across domains through data visualization.
These findings suggest that DA-GNNs are an efficient way of extracting domain independent cosmological information, a vital step toward robust deep learning for real cosmic survey data.

Contributed talks / 86

Simulation-based inference with non Gaussian statistics in the Dark Energy Survey

Author: Marco Gatti[1]

[1] *UPenn*

Corresponding Author: mgatti29@sas.upenn.edu

In recent years, non-Gaussian statistics have been growing in popularity as powerful tools for efficiently extracting cosmological information from current weak lensing data. Their use can improve constraints on cosmological parameters over standard two-point statistics, can additionally help discriminate between general relativity and modified gravity theories, and can help to self-calibrate astrophysical and observational nuisance parameters. During this talk, I will present an end-to-end simulation-based inference (SBI) framework that allows us to use common non-Gaussian statistics (e.g., higher order moments, peaks, scattering transform, phase wavelet harmonics) to constraints cosmological parameters. The pipeline relies on a neural network compression of the summary statistics and estimates the parameter posteriors using a mixture of Neural Density Estimators (NDEs). I will use the pipeline to compare the performance of different summary statistics in terms of cosmological parameters constraining power. I will then show constraints on data using the Dark Energy Survey year 3 weak lensing data. I will also be discussing the impact of observational systematics, and the main challenges ahead in view of stage IV surveys.

Contributed talks / 87

A Reanalysis of BOSS Galaxy Clustering Data with a Simulation-Based Emulator of the Wavelet Scattering Transform

Author: Georgios Valogiannis[None]

Corresponding Author: gv89@cornell.edu

Optimal extraction of the non-Gaussian information encoded in the Large-Scale Structure (LSS) of the universe lies at the forefront of modern precision cosmology. We propose achieving this task through the use of the Wavelet Scattering Transform (WST), which subjects an input field to a layer of non-linear transformations that are sensitive to non-Gaussianities through a generated set of WST

coefficients. In order to assess its applicability in the context of LSS surveys, we perform the first WST application on actual galaxy observations, through a WST analysis of the BOSS DR12 CMASS dataset. We lay out the detailed procedure on how to capture all necessary layers of realism for an application on data obtained from a spectroscopic survey, including the effects of redshift-space anisotropy, non-trivial survey geometry, the shortcomings of the dataset through a set of systematic weights and the Alcock-Paczynski distortion effect. Using the suite of Abacus summit simulations, we construct an emulator for the cosmological dependence of the WST coefficients and perform a likelihood analysis of the CMASS data to obtain the marginalized errors on cosmological parameters. The WST is found to deliver a substantial improvement in the values of the predicted 1σ errors compared to the regular galaxy power spectrum. Lastly, we discuss recent progress towards applying these techniques in order to fully harness the constraining power of upcoming spectroscopic observations by Stage-IV surveys such as DESI and Euclid.

Posters / **88**

Imaging hidden worlds? Exploring the SpHere INfrared survey for Exoplanets (SHINE) through deep learning

Author: Carles Cantero Mitjans[1]

Co-author: Mariam Sabalbal [1]

[1] *Université de Liège*

Corresponding Authors: ccantero@uliege.be, mariam.sabalbal@uliege.be

The detection of exoplanets has become one of the most active fields in astrophysics. Despite the fact that most of these discoveries have been made possible through indirect detection techniques, the direct imaging of exoplanets using 10-meter-class ground-based telescopes is now a reality. Achieving this milestone is the result of significant advances in the field of high-contrast imaging (HCI), including extreme adaptive optics systems and cutting-edge coronagraphs on the telescopes, along with dedicated post-processing techniques for image detection.

The detection capabilities of traditional post-processing techniques have been enhanced by a new family of machine learning-based methods in HCI. In the realm of deep learning, our group developed SODINN (Gomez Gonzalez et al., 2018), a binary classifier that employs a convolutional neural network to distinguish between companion signatures and residual noise in long exposures. Additionally, we have recently introduced a novel version of SODINN, known as NA-SODINN (Cantero et al., 2023), that better captures correlations in image noise, thus pushing the detection limits for fainter companions. These two algorithms have undergone testing with synthetic companion signatures, and we are now well-positioned to apply them to real HCI surveys in the search for potential companions.

In this poster, we present preliminary results from applying the SODINN and NA-SODINN algorithms to the F150 sample of the Sphere INfrared survey for Exoplanets (SHINE) survey (Langlois et al., 2021), which gathers observations of 150 stars using the SPHERE high-contrast imager at the VLT. The project aims to re-analyze this survey in an attempt to reveal yet undetected companions.

Contributed talks / **89**

Reionisation time fields reconstruction from 21 cm signal maps

Author: Julien Hiegel[1]

[1] *Observatoire Astronomique de Strasbourg - CNRS*

Corresponding Author: julien.hiegel@astro.unistra.fr

During the Epoch of reionisation, the intergalactic medium is reionised by the UV radiation from the first generation of stars and galaxies. One tracer of the process is the 21 cm line of hydrogen that will be observed by the Square Kilometre Array (SKA) at low frequencies, thus imaging the distribution of ionised and neutral regions and their evolution.

To prepare for these upcoming observations, we investigate a deep learning method to predict from 21 cm maps the reionisation time field treion, i.e. the time at which each location has been reionised. treion encodes the propagation of ionisation fronts in a single field, gives access to times of local reionisation or to the extent of the radiative reach of early sources. Moreover it gives access to the time evolution of ionisation on the plane of sky, when such evolution is usually probed along the line-of-sight direction.

We trained a convolutional neural network (CNN) using simulated 21 cm maps and reionisation times fields produced by the simulation code 21cmFAST . We also investigate the performance of the CNN when adding instrumental effects.

Globally, we find that without instrumental effects the 21 cm maps can be used to reconstruct the associated reionisation times field in a satisfying manner: the quality of the reconstruction is dependent on the redshift at which the 21 cm observation is being made and in general it is found that small scale features are smoothed in the reconstructed field, while larger
scale features are well recovered. When instrumental effects are included, the scale dependance of reconstruction is even further pronounced, with significant smoothing on small and intermediate scales.

The reionisation time field can be reconstructed, at least partially, from 21 cm maps of IGM during the Epoch of reionisation. This quantity can thus be derived in principle from observations and should then provide a mean to investigate the effect of local histories of reionisation on the first structures that appear in a given region.

Contributed talks / 90

Generative Topographic Mapping for tomographic redshift estimates

Authors: Grégoire Aufort[1]; Olivier Ilbert[2]

[1] *IAP*

[2] *LAM*

Corresponding Authors: gregoire.aufort@iap.fr, olivier.ilbert@lam.fr

In this talk, we explore the use of Generative Topographic Mapping (GTM) as an alternative to self-organizing maps (SOM) for deriving accurate mean redshift estimates for cosmic shear surveys. We delve into the advantages of the GTM probabilistic modeling of the complex relationships within the data, enabling robust estimation of redshifts. Through comparative analysis, we showcase the effectiveness of GTM in producing tomographic redshift estimates, thereby contributing to the advancement of cosmological studies. Our findings underscore GTM's potential as a powerful tool for redshift inference in large-scale astronomical surveys.

Contributed talks / 91

Before real data: pressing graph neural networks to do field-level simulation-based inference with galaxies

Author: Natalí Soler Matubaro de Santi[1]

Co-authors: Francisco Villaescusa-Navarro [2]; Raul Abramo [3]

[1] *University of São Paulo*

[2] *Flatiron Institute*

[3] *Universidade de São Paulo*

Corresponding Authors: natalidesanti@gmail.com, fvillaescusa@flatironinstitute.org

Field level likelihood-free inference is one of the brand new methods to extract cosmological information, over passing inferences of the usual and time-demanding traditional methods. In this work we train different machine learning models, without any cut on scale, considering a sequence of distinct selections on galaxy catalogs from the CAMELS suite in order to recover the main challenges of real data observations. We consider mask effects, peculiar velocity uncertainties, and galaxy selection effects. Also, we are able to show that we obtain a robust model across different sub-grid physical models such as Astrid, SIMBA, IllustrisTNG, Magneticum, and SWIFT-EAGLE using only galaxy phase-space information (3D positions and 1D velocity).
Moreover, we are able to show that the model can still track the matter content of the simulations keeping only the 2D positions and 1D velocity. The main purpose is to provide a proof of concept that graph neural networks, together with moment neural networks, can be used as a useful and powerful machinery to constrain cosmology for the next generation of surveys.

Posters / **92**

Cosmology Constraints from Strong Gravitational Lensing using Hierarchical Simulation Based Inference

Author: Sreevani Jarugula[1]

Co-authors: Aleksandra Ciprijanovic [2]; Becky Nevin [1]; Brian Nord [1]; Jason Poh [3]

[1] *Fermilab*

[2] *Fermi National Accelerator Laboratory*

[3] *University of Chicago*

Corresponding Authors: jasonpoh@uchicago.edu, rnevin@fnal.gov, jarugula@fnal.gov, aleksand@fnal.gov, nord@fnal.gov

Strong lenses are valuable probes of both astrophysics and cosmology, but traditional modeling methods for each system are computationally expensive. In addition, these methods won't be able to cope with the millions of lenses that will be discovered in the next generation of cosmic telescopes and surveys. New tools for inference, like Simulation-Based Inference (SBI) using Neural Posterior Estimation, present an opportunity for addressing this challenge. We perform SBI to simultaneously infer lensing system parameters and the dark energy equation of state from a population of galaxy-galaxy strong lenses. We compare the constraining power of a population level inference with the inference from individual lenses.
This analysis is important for cosmology inference from the upcoming strong lens follow-up surveys such as The 4MOST Strong Lensing Spectroscopic Legacy Survey (4SLSLS).

Contributed talks / **93**

EFTofLSS meets simulation-based inference: σ_8 from biased tracers

Author: Beatriz Tucci[1]

[1] *Max Planck Institute for Astrophysics*

Corresponding Author: beatriz.tucci@gmail.com

Modern cosmological inference typically relies on likelihood expressions and covariance estimations, which can become inaccurate and cumbersome depending on the scales and summary statistics under consideration. Simulation-based inference, in contrast, does not require an analytical form for the likelihood but only a prior distribution and a simulator, thereby naturally circumventing these issues. In this talk, we will explore how this technique can be used to infer σ_8 from a forward model based on Lagrangian Perturbation Theory and the bias expansion. The power spectrum and the bispectrum are used as summary statistics to obtain the posterior of the cosmological, bias and noise parameters via neural density estimation.

Posters / 94

Field-level BAO inference

Author: Ivana Babic[1]

[1] *Max Planck for Astrophysics*

Corresponding Author: ibabic@mpa-garching.mpg.de

The BAO feature is damped by non-linear structure formation, which reduces the precision with which we can infer the BAO scale from standard galaxy clustering analysis methods. A variety of techniques, known as BAO reconstruction, have been proposed to mitigate this damping effect; however, in order to work, these methods need to make assumptions abut bias and cosmology as well as to rely on the compression functions. In our study, we combine forward modeling with field-level inference in the goal of extracting the size of BAO scale using HMC sampling. Unlike traditional methods, field-level approach does not require reconstruction and permits full information extraction without relying on n-point functions. To fully gauge the gain of this approach, we are conducting a thorough comparison with n-point functions analysis, employing both standard likelihood-based and simulation-based inference methods.

Posters / 95

Generating multi-component Cosmological fields with Normalizing Flows

Authors: Kimmy Wu[1]; Matiwos Mebratu[2]

[1] *SLAC National Laboratory*

[2] *Stanford University*

Corresponding Authors: mmebrat1@stanford.edu, wlwu@stanford.edu

We present a technique to improve the accuracy and training efficiency of normalizing flows for multiple images in the context of cosmology. Normalizing flows are powerful deep generative models that can learn complex probability distributions through invertible transformations applied to a simple distribution. They are well-suited for both image generation and density estimation, enabling precise likelihood evaluation and efficient sampling. However, due to the inherent constraint of invertibility, normalized flows exhibit limited expressiveness when compared to other well-known models like GANs. Yet, past research has demonstrated that this limitation can be addressed by employing more expressive priors, such as resampled normal Gaussian or correlated priors, effectively enhancing the capabilities of normalizing flows. Our ongoing work focuses on developing a technique to enable normalizing flows to capture correlated non-Gaussian fluctuations in realistic mm-wave extragalactic foreground simulations, which consist of multiple components, based on N-body simulations. As part of this endeavor, we investigate and compare the efficiency of normalizing flows in generating two-component correlated non-Gaussian maps using various priors. The

three priors considered are as follows: the normal distribution (the simplest setup), the correlated prior (which incorporates the auto-spectrum of each target maps but not their correlations), and the component-correlated prior (which incorporates both the auto- and the cross-spectra of the target maps). We find that, when dealing with f_{nl} type local non-Gaussianity the application of correlated and component-correlated priors results in more accurate representations of the target distribution compared to other approaches. We will next apply this on realistic simulations of the CMB lensing convergence and extragalactic foreground fields, with the goal of field-level inference with mm-wave data.

Contributed talks / 96

The terms Eisenstein and Hu missed

Author: Deaglan Bartlett[1]

Co-authors: Benjamin Wandelt [2]; Miles Cranmer [3]

[1] *Institut d'Astrophysique de Paris*

[2] *Institut d'Astrophysique de Paris / The Flatiron Institute*

[3] *Cambridge University*

Corresponding Authors: miles.cranmer@gmail.com, deaglan.bartlett@iap.fr, bwandelt@iap.fr

The matter power spectrum of cosmology, P(k), is of fundamental importance in cosmological analyses, yet solving the Boltzmann equations can be computationally prohibitive if required several thousand times, e.g. in a MCMC. Emulators for P(k) as a function of cosmology have therefore become popular, whether they be neural network or Gaussian process based. Yet one of the oldest emulators we have is an analytic, physics-informed fit proposed by Eisenstein and Hu (E&H). Given this is already accurate to within a few percent, does one really need a large, black-box, numerical method for calculating P(k), or can one simply add a few terms to E&H? In this talk I demonstrate that Symbolic Regression can obtain such a correction, yielding sub-percent level predictions for P(k).

Contributed talks / 97

Data-driven galaxy morphology at $z > 3$ with contrastive learning and cosmological simulations

Author: Jesús Vega Ferrero[1]

Co-authors: Marc Huertas-Company [2]; Luca Costantin [3]; Pablo G. Pérez González [3]

[1] *Universidad de Valladolid (UVa)*

[2] *Instituto de Astrofísica de Canarias*

[3] *Centro de Astrobiología (CAB), CSIC-INTA*

Corresponding Authors: astrovega@gmail.com, marc.huertas.company@gmail.com

Visual inspections of the first optical rest-frame images from JWST have indicated a surprisingly high fraction of disk galaxies at high redshifts. Here, we alternatively apply self-supervised machine learning to explore the morphological diversity at $z \geq 3$.

Our proposed data-driven representation scheme of galaxy morphologies, calibrated on mock images from the TNG50 simulation, is shown to be robust to noise and to correlate well with physical properties of the simulated galaxies, including their 3D structure. We apply the method simultaneously to F200W and F356W galaxy images of a mass-complete sample ($M_*/M_\odot > 10^9$) at $3 \leq z \leq 6$

from the first JWST/NIRCam CEERS data release. We find that the simulated and observed galaxies do not exactly populate the same manifold in the representation space from contrastive learning. We also find that half the galaxies classified as disks (either CNN-based or visually) populate a similar region of the representation space as TNG50 galaxies with low stellar specific angular momentum and non-oblate structure.

Although our data-driven study does not allow us to firmly conclude on the true nature of these galaxies, it suggests that the disk fraction at $z \geq 3$ remains uncertain and possibly overestimated by traditional supervised classifications.

Posters / **98**

Identifying stellar disk truncations in Euclid galaxy images using Segment Anything Model (SAM)

Author: Jesús Vega Ferrero[1]

Co-authors: Fernando Buitrago [1]; Jesús Fernández [1]; Benjamín Sahelices [1]

[1] *Universidad de Valladolid (UVa)*

Corresponding Author: astrovega@gmail.com

Stellar disk truncations are a long-sought galactic size indicator based on the radial location of the gas density threshold for star formation, i.e., the edge/limit of the luminous matter in a galaxy. The study of galaxy sizes is crucial for understanding the physical processes that shape galaxy evolution across cosmic time. Current and future ultradeep and large-area imaging surveys, such as the JWST and the ESA's Euclid mission, will allow us to explore the growth of galaxies and trace the limits of star formation in their outskirts.

The task of identifying the disk truncations in galaxy images is, therefore, equivalent to what is called (informed) image segmentation in computer vision. Recently, the Meat AI research team has published the Segment Anything Model (SAM, *Kirillov et al. 2023*). SAM is a deep learning model that is capable of segmenting any type of data (including text, images, and audio) into smaller components or segments. The model is designed to be highly adaptable and versatile making it suitable for a wide range of applications.

In preparation for automatically identifying disk truncations in the galaxy images that will be soon released by Euclid, we run SAM over a dataset of 1048 disc galaxies with $M_* > 10^{10} M_\odot$ and $z < 1$ within the HST CANDELS fields presented in *Buitrago et al. 2023 (A&A, in press)*. We 'euclidize' the HST galaxy images by making composite RGB images using the H, J and I+V HST filters, respectively. Using these images as input for the SAM, we retrieve various truncation masks for each galaxy image given different configurations of the input dataset (i.e. varying the stretch and normalization of the input images) and of the SAM pipeline. Finally, we present a comparison of the truncations obtained with the SAM on the whole 'euclidized' dataset with the results presented in: a) *Buitrago et al. 2023 (A&A, in press)*, in which truncations are evaluated using the radial positions of the edge in the light profiles of galaxies —inferred in a non-automated way; b) *Fernández-Iglesias et al. 2023 (A&A, submitted)*, in which segmented images of truncations are automatically obtained using a U-Net.

Contributed talks / **99**

TheLastMetric: ML for statistically rigorous observing strategy optimization

Author: Alex Malz[1]

Co-authors: François Lanusse [2]; John Franklin Crenshaw [3]; Bryan Scott [4]; Melissa Graham [3]

[1] *Carnegie Mellon University*

[2] *CNRS*

[3] *University of Washington*

[4] *Northwestern University*

Corresponding Author: aimalz@nyu.edu

Most applications of ML in astronomy pertain to classification, regression, or emulation, however, ML has the potential to address whole new categories of problems in astronomical big data. This presentation uses ML in a statistically principled approach to observing strategy selection, which encompasses the frequency and duration of visits to each portion of the sky and impacts the degree to which the resulting data can be employed toward any scientific objective, let alone the net effect on many diverse science goals. Aiming to homogenize the units of observing strategy metrics across different science cases and minimize analysis model-dependence, we introduce TheLastMetric, a variational approximation to the lower bound of mutual information between a physical parameter of interest and anticipated data, a measure of the potentially recoverable information, under a given observing strategy. We demonstrate TheLastMetric in the context of photometric redshifts (photo-zs) from the upcoming Legacy Survey of Space and Time (LSST) on the Vera C. Rubin Observatory, showing qualitative agreement with traditional photo-z metrics and improved discriminatory power without assuming a photo-z estimation model. In combination with evaluations on other physical parameters of interest, TheLastMetric isolates the subjective assessment of relative priority of a science goal from the units-dependent sensitivity of its metric, enhancing the transparency and objectivity of the decisionmaking process. We thus recommend the broad adoption of TheLastMetric as an appropriate and effective paradigm for community-wide observing strategy optimization.

Contributed talks / 100

Extracting physical rules from ensemble machine learning for the selection of radio AGN.

Author: Rodrigo Carvajal[None]

Corresponding Author: racarvajal@ciencias.ulisboa.pt

Studying Active Galactic Nuclei (AGN) is crucial to understand processes regarding birth and evolution of Super-Massive Black Holes and their connection with star formation and galaxy evolution. However, few AGN have been identified in the EoR (z > 6) making it difficult to study their properties. In particular, a very small fraction of these AGN have been radio detected. Simulations and models predict that future observatories might increase these numbers drastically.

It becomes fundamental, then, to establish connections between radio emission and other multi-wavelength properties at high z. Recent wide-area multi-survey data have opened a window into obtaining these connections and rules.

At the same time, the development and operation of large-scale radio observatories, renders the use of regular AGN detection and redshift determination techniques inefficient. Machine Learning (ML) methods can help to predict the detection of AGN and some of their properties. We have developed, then, a series of ML models that, using multi-wavelength photometry, can produce a list of Radio Galaxy candidates, with their predicted redshift values.

More importantly, we have also applied some state-of-the-art feature importance techniques to understand which physical properties drive the predictions made by our models. From these techniques, it is possible to derive indicators for the selection of studied sources.

We will present the results of applying these models and techniques on near-infrared (NIR)-selected sources from the HETDEX Spring Field and the Stripe 82 Field. Furthermore, using feature importances, we will describe which properties hold the highest predicting power and the derivation of a efficient colour-colour criterion for the identification of AGN candidates. Moreover, we will introduce our efforts to apply said models and procedures to data in the area of the Evolutionary Map of the Universe (EMU, a precursor of the SKA Observatory) Pilot Survey.

Posters / **102**

Detecting the edges of galaxies with Deep Learning

Author: Jesús Fernández Iglesias[1]

Co-authors: Benjamín Sahelices [2]; Fernando Buitrago [2]

[1] *University of Valladolid (School of Informatic Engineering)*

[2] *Universidad de Valladolid (UVa)*

Corresponding Author: jssfernandez0@gmail.com

Galaxy edges/truncations are Low Surface Brightness (LSB) features located in the galaxy outskirts that delimit the distance up to where the gas density enabled efficient star formation. Therefore, they constitute true galaxy edges. As such, they could be interpreted as a non-arbitrary means to determine the galaxy size, and this is also reinforced by the smaller scatter in the galaxy mass-size relation when comparing them with other size proxies. However there are several problems attached to this novel metric, namely the access to deep imaging and the need to contrast surface brightness, color and mass profiles to derive the edge position. While the first hurdle is already overcome by new ultra-deep galaxy observations, we hereby propose the use of Machine Learning algorithms in order to determine the position of these features for very large datasets. We compare the semantic segmentation by our Deep Learning models with the results obtained by humans for HST observations of a sample of massive disk galaxies at $z < 1$. In addition, the concept of astronomic augmentations is introduced to endow the inputs of the networks with physical meaning. Our findings (to appear in Fernández-Iglesias et al. 2023 in press) suggest that similar performances than humans could be routinely achieved, although in the majority of cases the best results are obtained by combining (with a pixel-by-pixel democratic vote) the output of several neural networks using ensemble learning. Specifically, the experiments show a great similarity between the semantic segmentation performed by the AI compared to the human model, with an average Dice of 0.8969 for the best model and an average Dice of 0.9104 for the best ensemble. This methodology will be profusely used in future datasets such as Euclid where our team has the expertise to create a LSB-compliant data reduction. We also offer to the community our Machine learning algorithms in the repository `https://github.com/jesusferigl`

Contributed talks / **103**

Field-level inference of primordial non-Gaussianity, using next-generation galaxy surveys

Author: Adam Andrews[1]

[1] *INAF OAS Bologna*

Corresponding Author: adam.andrews@inaf.it

A significant statement regarding the existence of primordial non-Gaussianity stands as one of the key objectives of next-generation galaxy surveys. However, traditional methods are burdened by a variety of issues, such as the handling of unknown systematic effects, the combination of multiple probes of primordial non-Gaussianity, and the capturing of information beyond the largest scales in the data. In my presentation, I will introduce my pioneering work of applying field-level inference to constrain primordial non-Gaussianity galaxy surveys. I will discuss how my method can resolve the challenges faced by other approaches and how I can capture more information from the data compared to traditional methods. Additionally, I will explore the additional data products that my method enables and delve into other potential applications. Finally, I will briefly touch upon the future use of field-level inference to study the primordial universe, along with the promises and challenges inherent in this approach.

Contributed talks / 104

Vision Transformers for Cosmological Inference from Weak Lensing

Author: Shubh Agrawal[1]

Co-authors: Marco Gatti [2]; Bhuvnesh Jain [1]

[1] *University of Pennsylvania*

[2] *UPenn*

Corresponding Authors: bjain@physics.upenn.edu, shubh@sas.upenn.edu, mgatti29@sas.upenn.edu

Weak gravitational lensing is an excellent quantifier of the growth of structure in our universe, as the distortion of galaxy ellipticities measures the spatial fluctuations in the matter field density along a line of sight. Traditional two-point statistical analyses of weak lensing only capture Gaussian features of the observable field, hence leaking information from smaller scales where non-linear gravitational interactions yield non-Gaussian features in the matter distribution. Higher-order statistics such as peak counts, Minkowski-functionals, three-point correlation functions, and convolutional neural networks, have been introduced to capture this additional non-Gaussian information and improve constraints on key cosmological parameters such as Ω_m and σ_8.

We demonstrate the potential of applying a self-attention-based deep learning method, specifically a Vision Transformer, to predict cosmological parameters from weak lensing observables, particularly convergence κ maps. Transformers, which were first developed for natural language processing and are now at the core of generative large language models, can be used for computer vision tasks with patches from an input image serving as sequential tokens analogous to words in a sentence. In the context of weak lensing, Vision Transformers are worth exploring for their different approach to capturing long-scale and inter-channel information, improved parallelization, and lack of strong inductive bias and locality of operations.

Using transfer learning, we compare the performance of Vision Transformers to that of benchmark residual convolutional networks (ResNets) on simulated wCDM theory predictions for κ, with noise properties and sky coverage similar to DESY3, LSSTY1, and LSSTY10. We further use neural density estimators to investigate the differences in the cosmological parameters' posteriors recovered by either deep learning method. These results showcase a potential astronomical application derived from the advent of powerful large language models, as well as machine learning tools relevant to the next generation of large-scale surveys.

Contributed talks / 105

Galaxy modeling with physical forward models and generative neural networks

Author: Peter Melchior[1]

[1] *Princeton University*

Corresponding Author: peter.m.melchior@gmail.com

Detection, deblending, and parameter inference for large galaxy surveys have been and still are performed with simplified parametric models, such as bulge-disk or single Sersic profiles. The complex structure of galaxies, revealed by higher resolution imaging data, such as those gathered by HST or, in the future, by Euclid and Roman, makes these simplifying assumptions problematic. Biases arise in photometry and shape measurements, and I will discuss examples for both.

On the other hand, non-parametric modeling also has a long history in many fields of image processing. But it is limited to signal-to-noise regimes that are high by the standards of most astrophysical

surveys. This weakness can be overcome by specifying priors over the space of galaxy images. I will present a new codebase, scarlet2, written entirely in jax, for modeling complex extragalactic scenes. I will also discuss how to integrate data-driven priors in the form of score models, and show examples of sampling from posteriors to assess the uncertainties in heavily blended configurations. I will conclude with an outlook of how these tools need to be extended to fully exploit the data from the combination of optical surveys that will shape astrophysics in the 2020s.

Contributed talks / **106**

Explaining dark matter halo abundance with interpretable deep learning

Author: Ningyuan (Lillian) Guo[1]

Co-authors: Andrew Pontzen [1]; Hiranya Peiris [2]; Luisa Lucie-Smith [3]

[1] *University College London*

[2] *University College London/Stockholm University*

[3] *MPA*

Corresponding Author: ningyuan.guo.20@ucl.ac.uk

The halo mass function describes the abundance of dark matter halos as a function of halo mass and depends sensitively on the cosmological model. Accurately modelling the halo mass function for a range of cosmological models will enable forthcoming surveys such as Vera C. Rubin Observatory' s Legacy Survey of Space and Time (LSST) and Euclid to place tight constraints on cosmological parameters. Due to the highly non-linear nature of halo formation, understanding which quantities determine the halo mass function for different cosmological models is difficult. We present an interpretable deep learning framework that allows us to find, with minimal prior assumptions, a compressed representation of the information required to accurately predict the halo mass function. We use neural network models that consist of an encoder-decoder architecture: the encoder compresses the input linear matter power spectrum and growth function into a low-dimensional representation, and the decoder uses this representation to predict halo abundance given a halo mass. We train the network to predict the halo mass function at redshift z=0 to better than 1% precision for a range of cosmological parameters. We then interpret the representation found by the network via measuring mutual information between the representation and quantities such as the ground truth halo number densities, the power spectrum, and cosmological parameters. This can enable us to gain new insights on what physics is involved in the process of halo formation, and a better understanding of how to accurately model the halo mass function for different cosmological models. The framework can also be extended to model the halo mass function over a range of redshifts.

Contributed talks / **107**

Convolutional Neural Networks for Exoplanet Detection in Photometric Light Curves From Massive Data Surveys

Author: Stela Ishitani Silva[1]

[1] *NASA GSFC*

Corresponding Author: stela.ishitani@gmail.com

Amidst the era of astronomical surveys that collect massive datasets, neural networks have emerged as powerful tools to address the challenge of exploring and mining these enormous volumes of information from our sky. Among the obstacles in the study of these surveys is the identification of exoplanetary signatures in the photometric light curves. In this presentation, we will discuss how convolutional neural networks can significantly facilitate the detection of exoplanets, focusing on

two exoplanetary detection methods: (1) planetary transits and (2) gravitational microlensing. We will elaborate on (1) their proven success in detecting planetary transit signals within the Transiting Exoplanet Survey Satellite data and (2) our ongoing project to identify gravitational microlensing events using the nine-year Microlensing Observations in Astrophysics dataset. Our strategy proposes using only raw photometric light curves as input for our neural network pipeline, which, after training, can detect the desired signal in a light curve in milliseconds. Looking towards future space missions, we will discuss the role of neural networks as an alternative pipeline to accelerate the identification of potential exoplanet candidates in the Nancy Grace Roman Space Telescope data.

Contributed talks / **108**

Assessing and Benchmarking the Fidelity of Posterior Inference Methods for Astrophysics Data Analysis

Author: Becky Nevin[1]

Co-authors: Aleksandra Ciprijanovic [2]; Brian Nord [1]; Jason Poh [3]; Samuel McDermott [3]; Sreevani Jarugula [1]

[1] *Fermilab*

[2] *Fermi National Accelerator Laboratory*

[3] *University of Chicago*

Corresponding Authors: jasonpoh@uchicago.edu, rnevin@fnal.gov, jarugula@fnal.gov, aleksand@fnal.gov, samueldmcdermott@gmail.com, nord@fnal.gov

In this era of large and complex astronomical survey data, interpreting, validating, and comparing inference techniques becomes increasingly difficult. This is particularly critical for emerging inference methods like Simulation-Based Inference (SBI), which offer significant speedup potential and posterior modeling flexibility, especially when deep learning is incorporated. We present a study to assess and compare the performance and uncertainty prediction capability of Bayesian inference algorithms –from traditional MCMC sampling of analytic functions to deep learning-enabled SBI. We focus on testing the capacity of hierarchical inference modeling in those scenarios. Before we extend this study to cosmology, we first use astrophysical simulation data to ensure interpretability. We demonstrate a probabilistic programming implementation of hierarchical and non-hierarchical Bayesian inference using simulations derived from the DeepBench software library, a benchmarking tool developed by our group that generates simple and controllable astrophysical objects from first principles. This study will enable astronomers and physicists to harness the inference potential of these methods with confidence.

Contributed talks / **110**

Harnessing Differentiable and Probabilistic Programming for Scalable and Robust Statistical Analysis of Astronomical Surveys

Author: Alessio Spurio Mancini[1]

[1] *University College London*

Corresponding Author: a.spuriomancini@ucl.ac.uk

I present a novel, general-purpose Python-based framework for scalable and efficient statistical inference by means of hierarchical modelling and simulation-based inference.

The framework is built combining the JAX and NumPyro libraries. The combination of differentiable and probabilistic programming offers the benefits of automatic differentiation, XLA optimization, and the ability to further improve the computational performance by running on GPUs and TPUs as well. These properties allow for efficient sampling through gradient-based methods, and for significantly enhanced performance of neural density estimation for simulation-based inference, augmented by the simulator gradients.

The framework seamlessly integrates with the recently developed COSMOPOWER-JAX and JAX-COSMO libraries, making it an ideal platform to solve Bayesian inverse problems in cosmology. Beyond cosmology, the framework is designed to be a versatile, robust tool for cutting-edge analysis of astronomical surveys. I demonstrate its practical utility through applications to various domains, including but not limited to weak lensing, supernovae, and galaxy clusters.

Contributed talks / **111**

Causal graphical models for galaxy surveys

Author: Serafina Di Gioia[1]

[1] *ICTP*

Corresponding Author: sdigioia@sissa.it

A fundamental task of data analysis in many scientific fields is to determine the underlying causal relations between physical properties as well as the quantitative nature of these relations/laws. These laws are the fundamental building blocks of scientific models describing observable phenomena. Historically, causal methods were applied in the field of social sciences and economics (Pearl, 2000), where causal relations were investigated by means of interventions (manipulating and varying features of systems to see how systems react). However, since we can observe one single world and one single Universe we cannot use interventions for recovering causal models describing our data in disciplines such as astrophysics or climate sciences. It is therefore necessary to discover causal relations by analyzing statistical properties of purely observational data, a task known as causal discovery or causal structure learning.

In S. Di Gioia et al, 2023 (in preparation), in collaboration with R. Trotta, V. Acquaviva, F. Bucinca and A. Maller, we perform causal model discovery on simulated galaxy data, to better understand which galaxy and halo properties are the drivers of galaxy size, initially at redshift z = 0. In particular, we used a constraint-based structure learning algorithm, called kernel-PC, based on a Python parallel code developed by the author, and, as input data, the simulated galaxy catalog generated with the Santa Cruz semi-analytic model (SC-SAM). The SC-SAM was built on the merger trees extracted from the dark matter-only version of the TNG-100 hydro-dynamical simulation, which showed to describe successfully the full spectra of observed galaxy properties, from z=0 to z=3-4 (Gabrielpillai et al., 2022).

In my talk I will present the main results of this work, together with an overview of the most common algorithms to perform causal discovery, in the framework of Causal Graphical Models, focusing on their potential applicability to upcoming astronomical surveys. Future applications of this method include dimensionality reduction and Bayesian model discovery.

Contributed talks / **112**

Doing More With Less; Label-Efficient Learning for Euclid and Rubin

Author: Mike Walmsley[1]

[1] *University of Toronto*

Corresponding Author: walmsleymk1@gmail.com

Deep learning is data-hungry; we typically need thousands to millions of labelled examples to train effective supervised models. Gathering these labels in citizen science projects like Galaxy Zoo can take years, delaying the science return of new surveys. In this talk, I'll describe how we're combining simple techniques to build better galaxy morphology models with fewer labels.

First 1, we're using large-scale pretraining with supervised and self-supervised learning to reduce the number of labelled galaxy images needed to train effective models. For example, using self-supervised learning to pretrain on unlabelled Radio Galaxy Zoo images halves our error rate at distinguishing FRI and FRII radio galaxies in a separate dataset.

Second [2], we're continually retraining our models to prioritise the most helpful galaxies for volunteers to label. Our probabilistic models filter out galaxies they can confidently classify, leaving volunteers able to focus on challenging and interesting galaxies. We used this to measure the morphology of every bright extended galaxy in HSC-Wide in weeks rather than years.

Third [3], we're using natural language processing to capture radio astronomy classes (like "FRI"or "NAT") through plain English words (like "hourglass") that volunteers use to discuss galaxies. These words reveal which visual features are shared between astronomical classes, and, when presented as classification options, let volunteers classify complex astronomical classes in an intuitive way.

We are now preparing to apply these three techniques - pretraining, active learning, and natural language labels - to provide day-one galaxy morphology measurements for Euclid DR1.

Contributed talks / **113**

Embedding Neural Networks in ODEs to Learn Linear Cosmological Physics

Author: James Sullivan[1]

[1] *UC Berkeley*

Corresponding Author: jmsullivan@berkeley.edu

The ΛCDM cosmological model has been very successful, but cosmological data indicate that extensions are still highly motivated. Past explorations of extensions have largely been restricted to adding a small number of parameters to models of fixed mathematical form. Neural networks can account for more flexible model extensions and can capture unknown physics at the level of differential equation models. I will present evidence that it is possible to learn missing physics in this way at the level of linear cosmological perturbation theory as well as quantify uncertainty on these neural network predictions. This is accomplished through Bolt, the first differentiable Boltzmann solver code - the gradients provided by Bolt allow for efficient inference of neural network and cosmological parameters. Time permitting, I will also present other aspects of Bolt, such as the use of iterative methods of solution, choice of automatic differentiation algorithm, and stiff ODE solver performance.

Contributed talks / **114**

Optimizing Galaxy Sample Selections for Weak Lensing Cluster Cosmology

Author: Markus Rau[1]

[1] *Argonne National Laboratory*

Corresponding Author: markusmichael.rau@gmail.com

Weak Lensing Galaxy Cluster Masses are an important observable to test the cosmological standard model and modified gravity models. However cluster cosmology in optical surveys is challenged by sources of systematics like photometric redshift error. We use combinatorial optimization schemes and fast Machine Learning assisted model evaluation to select galaxy source samples that minimize the expected systematic error budget while maintaining sufficient signal-to-noise in the measurement to meet the stringent science requirements of surveys like LSST.

Contributed talks / 115

Anomaly detection using local measures of uncertainty in latent representations

Author: Fiona Porter[1]

Co-author: Anna Scaife [1]

[1] *Jodrell Bank Centre for Astrophysics*

Corresponding Author: fiona.porter-2@manchester.ac.uk

As upcoming SKA-scale surveys open new regimes of observation, it is expected that some of the objects they detect will be "unknown unknowns": entirely novel classes of sources which are outside of our current understanding of astrophysics. The discovery of these sources has the potential to introduce new fields of study, as it did for e.g. pulsars, but relies upon us being able to identify them within peta- or exascale data volumes. Automated anomaly detection using machine learning offers a promising method to ensure that atypical sources are not lost within the data, but these methods are typically incapable of simultaneously classifying non-anomalous sources and identifying anomalies, resulting in separate models being needed to complete both necessary tasks. In this talk, we discuss the possibility of using uncertainty metrics derived from an image classification model to provide anomaly detection within a classification pipeline.

Fanaroff-Riley (FR) galaxies, a type of radio-loud AGN, are among the sources that are expected to see a drastic increase in known population with upcoming large-scale radio surveys, and provide a useful test population for outlier detection because in addition to two "standard" morphologies (FRI and FRII) there are numerous rare morphological subclasses which are particularly useful for the study of AGN environments and engines. Using the MiraBest dataset of Fanaroff-Riley galaxies, we trained supervised deep learning model on binary Fanaroff-Riley sources, reserving hybrid FR galaxies to serve as a sample of "anomalous" objects that might be mistaken for in-distribution sources. Our model architecture used dropout at test time to approximate a Bayesian posterior on predictions, allowing for uncertainty in a class label to be expressed by calculating predictive entropy.

Highly anomalous out-of-distribution sources were found to be located in sparse regions of latent space and hence were easily identifiable, but hybrid sources could not easily be isolated from binary FR galaxies in either latent space or by entropy value alone. Instead, we created a measure of typical local entropy by calculating the average entropy of the nearest ten training set sources to any given point in latent space; this allowed for objects with atypically high or low entropy relative to nearby sources to be identified regardless of the absolute value of their entropy.

Using a test set of both in-distribution binary FRs and "anomalous" hybrid sources, we find that the in-distribution sources show no significant departure from the training set entropy, but hybrid sources have significantly higher entropy than their surroundings in all regions of latent space except where the local entropy is itself maximal. All sources more than 3σ from the local entropy were found to be hybrids, and flagging using this method alone detected 30% of the hybrid sample; the majority of the remaining hybrid sources were found to have near-maximal entropy, meaning that additionally flagging high-entropy sources would allow for both these and the most uncertainly-labelled in-distribution FR galaxies to be inspected while avoiding unnecessary flagging of low-uncertainty sources.

Posters / **116**

Deciphering Black-Hole Physics with Modern Machine-Learning Methods

Author: Thaddaeus Kiker[1]

Co-authors: James Steiner [2]; Joanna Kuraszkiewicz [2]; Markus Rau [3]

[1] *AstroAI CfA Harvard, Columbia University*

[2] *AstroAI CfA Harvard*

[3] *Argonne National Laboratory*

Corresponding Authors: jkuraszkiewicz@cfa.harvard.edu, tjk2147@columbia.edu, markusmichael.rau@gmail.com, james.steiner@cfa.harvard.edu

Supermassive black holes reside in the center of almost every galaxy. Today's supermassive black holes are mostly dormant (like the one at the center of our Milky Way), but in the past, they were actively accreting large amounts of matter and releasing vast amounts of energy. Galaxies with the brightest, most active supermassive black holes, called active galactic nuclei (AGN), are the most luminous objects in the universe. AGNs show many visible and ultraviolet emission lines, which probe the accreting material's physical conditions and the black hole's properties.

We discuss our work towards building a generative model of AGN spectra that will help us to study correlations between emission lines to derive insight into the accretion process, starting from a model to cluster AGN spectra directly using spectral input.

Posters / **117**

Improving astrophysical scaling relations with machine learning

Authors: Digvijay Wadekar[1]; Francisco Villaescusa-Navarro[2]; Leander Thiele[3]; Miles Cranmer[4]; Shirley Ho[None]

Co-author: David Spergel

[1] *Institute for Advanced Study (IAS)*

[2] *Flatiron Institute*

[3] *Princeton University*

[4] *Cambridge University*

Corresponding Authors: miles.cranmer@gmail.com, jayw@ias.edu, lthiele@princeton.edu, fvillaescusa@flatironinstitute.org

Finding low-scatter relationships in properties of astrophysical systems is important to estimate their masses/distances. I will show how interpretable ML tools like symbolic regression can be used to expeditiously search for these low-scatter relations in abstract high-dimensional astrophysical datasets. I will present new scaling relations between properties of galaxy clusters that we obtained using ML. I will also highlight advantages of using interpretable ML tools instead of deep neural networks for particular astrophysical problems.

Contributed talks / **118**

Fishnets: Mapping Information Geometry with Robust, Scalable Neural Compression

Author: Lucas Makinen[1]

Co-authors: Benjamin Wandelt [2]; Justin Alsing [3]

[1] *Imperial College London*

[2] *Institut d'Astrophysique de Paris / The Flatiron Institute*

[3] *Stockholm University*

Corresponding Authors: justin.alsing@fysik.su.se, l.makinen21@imperial.ac.uk, bwandelt@iap.fr

Data compression to informative summaries is essential for modern data analysis. Neural regression is a popular simulation-based technique for mapping data to parameters as summaries over a prior, but is usually agnostic to how uncertainties in information geometry, or data-summary relationship, changes over parameter space. We present Fishnets, a general simulation-based, neural compression approach to calculating the Fisher information and score for arbitrary data structures *as functions of parameters*. These compression networks can be scaled information-optimally to arbitrary data structures, and are robust to changes in data distribution, making them ideal tools for cosmological and graph dataset analyses.

Contributed talks / **119**

The Cosmic Graph: Optimal Information Extraction from Large-Scale Structure using Catalogues

Author: Lucas Makinen[1]

Co-authors: Alan Heavens [1]; Benjamin Wandelt [2]; Natalia Porqueres [3]; Pablo Lemos [4]; Tom Charnock

[1] *Imperial College London*

[2] *Institut d'Astrophysique de Paris / The Flatiron Institute*

[3] *University of Oxford*

[4] *Mila - Université de Montréal*

Corresponding Authors: tom@charnock.fr, phys2398@ox.ac.uk, a.heavens@imperial.ac.uk, l.makinen21@imperial.ac.uk, plemos91@gmail.com, bwandelt@iap.fr

The cosmic web, or Large-Scale Structure (LSS) is the massive spiderweb- like arrangement of galaxy clusters and the dark matter holding them together under gravity. The lumpy, spindly universe we see today evolved from a much smoother, infant universe. How this structure formed and the information embedded within is considered one of the "Holy Grails"of modern cosmology, and might hold the key to resolving existing "tensions"in cosmological theory. But how do we go about linking this data to theory? Cosmological surveys are comprised of millions of pixels, which can be difficult for samplers and analytic likelihood analysis. This also poses a problem for simulation- based inference: how can we best compare simulations to observed data? Information Maximising Neural Networks (IMNNs) offer a way to compress massive datasets down to (asymptotically) lossless summaries that contain the same cosmological information as a full sky survey, as well as quantify the information content of an unknown distribution. We will look at LSS assembled as a graph (or network) from discrete catalogue data, and use graph neural networks in the IMNN framework to optimally extract information about cosmological parameters (theory) from this representation. We will make use of the modular graph structure as a way to open the "black box"of simulation-based inference and neural network compression to show where cosmological information is stored.

Posters / **120**

Large scale structure: information content, scalable neural summaries and scaling laws for the neural network

Authors: Anirban Bairagi[1]; Benjamin Wandelt[2]

[1] *Institut d'Astrophysique de Paris*

[2] *Institut d'Astrophysique de Paris / The Flatiron Institute*

Corresponding Authors: anirban.bairagi@iap.fr, bwandelt@iap.fr

How much cosmological information does a cube of dark matter contain? Are we utilising the full potential of information available within a density field? Neural summaries aim to extract all these informations; but success depends on the availability of simulations, network architecture and hyperparameters, and the ability to train the networks. Even for the simplest summary statistics power spectrum we need 7 layers to get best possible optimal result from Quijote dark matter simulations. Hyperparameter tuning for every additional layer is not possible every time while trying different architecture. So an extensive hyperparameter search for a single layer perceptron on P(k) has been done to match the results with linear regression prediction and these hyperparameters are expected to be scalable for larger networks from the current studies on Large Neural Networks. Our study on loss vs number of training simulation suggests currently available 2000 latin hypercube simulations are not enough to reach the optimal regime. On the other hand fitting high resolution, large volume of cosmological data from next generation surveys like DESI or EUCLID into a GPU with the best performance built till date and infer parameter constraints from it is almost impossible without losing some amount information available within the data. Current methods tend to use the low resolution and smaller volume data because of which we are throwing away a larger part of the information. We have come up with a method that can solve this issue by combining sub volumes of density field with the power spectrum.

Contributed talks / 121

Debating the Benefits of Differentiable Cosmological Simulators for Weak Lensing Full-Field Inference (LSST Y10 case study)

Author: Justine Zeghal[1]

Co-authors: Alexandre Boucaud [2]; Denise Lanzieri [3]; François Lanusse [2]

[1] *APC, CNRS*

[2] *CNRS*

[3] *CEA Saclay*

Corresponding Authors: denise.lanzieri@cea.fr, alexandre.boucaud@apc.in2p3.fr, zeghal@apc.in2p3.fr, francois.lanusse@cea.fr

Conventional cosmic shear analyses, relying on two-point functions, do not have access to the non-Gaussian information present at the full field level, thus limiting our ability to constrain with precision cosmological parameters. Performing Full-Field inference is in contrast an optimal way to extract all available cosmological information, and it can be achieved with two widely different methodologies:
Explicit high-dimensional inference through the use of Bayesian Hierarchical Model (BHM)
Implicit Inference (also known as Simulation-Based Inference or Likelihood-Free Inference)
It is evident that differentiability of the forward model is essential for explicit inference, as this approach requires exploring a very high dimensional space, which is only practical with gradient-based inference techniques (HMC, Variational Inference, etc). In this work, we consider the question of whether implicit inference approaches can similarly benefit from having access to a differentiable simulator in a cosmological full-field inference scenario. Indeed, several methods (including ours) have been developed in recent years to leverage the gradients of the simulator to help constrain the inference problem, but the benefits of these gradients are problem dependent, raising the question of their benefit for cosmological inference. To answer this question, we consider a simplified full-field weak lensing analysis, emulating an LSST Y10 setting, and benchmark state-of-the art implicit inference methods making use or not of gradients.

This setting allows us to ask a first question: "What is the best method to optimally recover cosmological parameters for an LSST full-field weak lensing analysis with the minimum number of forward model evaluations?"There, our results suggest that gradient-free SBI methods are the most effective for this particular problem, and we develop some insights explaining why.

Posters / **122**

An Observationally Driven Multifield Approach for Probing the Circum-Galactic Medium with Convolutional Neural Networks

Author: Naomi Gluck[None]

Corresponding Author: naomi.gluck@yale.edu

The circum-galactic medium (CGM) can feasibly be mapped by multiwavelength surveys covering broad swaths of the sky. With multiple large datasets becoming available in the near future, we develop a likelihood-free Deep Learning technique using convolutional neural networks (CNNs) to infer broad-scale properties of a galaxy's CGM and its halo mass for the first time. Using CAMELS (Cosmology and Astrophysics with MachinE Learning Simulations) IllustrisTNG, SIMBA, and Astrid, we train CNNs on 2D maps of Soft X-ray and 21-cm (HI) radio to trace hot and cool gas, respectively, around galaxies, groups, and clusters. The tested CNN provides inferences on halo mass, CGM mass, metallicity, temperature, and cool gas fraction. Our CNN creates the unique opportunity to simultaneously train and test with HI and X-ray maps as a "multifield" dataset, inferring CGM properties significantly better than either alone. Applying multiwavelength survey limits to the CNN shows that X-ray is not powerful enough to infer low halo masses. Creating a multifield with HI is essential for inference refinement. Generally, a CNN trained and tested on Astrid (SIMBA) can most (least) accurately infer CGM properties. Cross-simulation analysis – raining on one simulation and testing on another – is then performed to quantify model robustness. Models that are not robust within cross-simulation analysis will not produce robust inferences when the simulation-based testing set is exchanged with multiwavelength observational data. Future efforts including saliency analysis, higher resolution maps, and a more extensive multifield are underway to overcome this challenge.

Contributed talks / **123**

Scientific Discovery from Ordered Information Decomposition

Author: Matthew Ho[1]

[1] *Institut d'Astrophysique de Paris*

Corresponding Author: matthew.ho@iap.fr

How can we gain physical intuition in real-world datasets using 'black-box' machine learning? In this talk, I will discuss how ordered component analyses can be used to seperate, identify, and understand physical signals in astronomical datasets. We introduce Information Ordered Bottlenecks (IOBs), a neural layer designed to adaptively compress data into latent variables organized by likelihood maximization. As an nonlinear extension of Principal Component Analysis, IOB autoencoders are designed to be truncated at any bottleneck width, controlling information flow through only the most crucial latent variables. With this architecture, we show how classical neural networks can be easily extended to dynamically order latent information, revealing learned structure in multi-signal datasets. We demonstrate how this methodology can be extended to structure and classify physical phenomena, discover low-dimensional symbolic expressions in high-dimensional data, and regularize implicit inference. Along the way, we present several astronomical applications including emulation of CMB power spectrum, analysis of binary black hole systems, and dimensionality reduction of galaxy properties in large cosmological simulations.

Posters / **125**

Deep Learning and Hierarchical Inference to infer H0 from Strong Gravitational Lenses

Author: Sydney Erickson[1]

[1] *Stanford*

Corresponding Author: sydney3@stanford.edu

To achieve a high precision measurement of the Hubble constant from strongly lensed AGN, we need to take advantage of the 1,000s of new strong lens observations that will come from the next generation of survey telescopes. In preparation for modeling roughly 10 times more lenses than are currently known, we have been developing a machine learning lens modeling technique. We use a deep convolutional neural network for neural posterior estimation of lens model posterior PDFs. These posteriors are then combined in a hierarchical inference to recover population hyperparameters and correct individual posteriors for bias from the choice of interim training prior. As a first step in preparing our pipeline, we validate our technique by testing it on real lens images for the first time, using a sample of 14 quadruply-lensed quasars imaged by HST for the STRIDES collaboration. Given our flexible and fast training methodology, we test how key potential sources of systematic error, such as image rizzling, PSF modeling, and source morphology, affect the network predictions. We also test the robustness of our methods, including a direct comparison to lensing constraints from traditional modeling. Our
results show that deep learning lens modeling is a powerful probe of systematics, providing insights that are not possible with traditional modeling. We find improved simulation realism and further algorithmic development are required before further scientific application of the pipeline to real data.

Contributed talks / **126**

HySBI - Hybrid Simulation-Based Inference

Author: Chirag Modi[1]

Co-author: Oliver Philcox [2]

[1] *Flatiron Institute*

[2] *Columbia University*

Corresponding Author: cmodi@flatironinstitute.org

We present a novel methodology for hybrid simulation-based inference (HySBI) for large scale structure analysis in cosmology. Our approach combines perturbative analysis on the large scales which can be modeled analytically from first principles, with simulation based implicit inference (SBI) on small, non-linear scales that cannot be modeled analytically. As a proof-of-principle, we apply our method to dark matter density fields to constrain cosmology parameters using power spectrum on the large scales, and power spectrum and wavelet coefficients on small scales. We highlight how this hybrid approach can mitigate the computational challenges in applying SBI to the future cosmological surveys, and discuss the roadmap to extend this approach for analyzing survey data.

Contributed talks / **128**

Sampling with Hamiltonian Neural Networks

Author: Vincent Souveton[1]

[1] *LMBP - Université Clermont Auvergne*

Corresponding Author: vincent.souveton@ext.uca.fr

Normalizing Flows (NF) are Generative models which transform a simple prior distribution into the desired target. They however require the design of an invertible mapping whose Jacobian determinant has to be computable. Recently introduced, Neural Hamiltonian Flows (NHF) are Hamiltonian dynamics-based flows, which are continuous, volume-preserving and invertible and thus make for natural candidates for robust NF architectures. In particular, their similarity to classical Mechanics could lead to easier interpretability of the learned mapping. In this presentation, I will detail the NHF architecture and show that they may still pose a challenge to interpretability. For this reason, I will introduce a fixed kinetic energy version of the model. Inspired by physics, this approach improves interpretability and requires less parameters than the original model. I will talk about the robustness of the NHF architecture, especially its fixed-kinetic version, on a simple 2D problem and present first results in higher dimension. Finally, I will show how to adapt NHF to the context of Bayesian inference and illustrate the method on an example from cosmology.

Review presentations / **130**

TBD: ML and Bayesian inference in cosmology

Author: Jens Jasche[1]

[1] *Stockholm University*

Corresponding Author: jens.jasche@fysik.su.se

Review presentations / **131**

Capitalizing on Artificial Intelligence for LSS Cosmology

Corresponding Author: tomasz.kacprzak@phys.ethz.ch

In this review talk, I will show how artificial intelligence can bring tangible benefits to cosmological analysis of large-scale structure.
I will focus on how the use of AI in the framework of Simulations-Based Inference to achieve scientific objectives that would not be attainable with classical 2-pt function analyses. I will show three avenues where, in my opinion, AI can bring the most benefits: reaching the information floor of limited survey data via SBI analysis, accelerating simulations for SBI, and breaking degeneracies between cosmological probes and astrophysical nuissance fields. I will discuss new challenges that come with AI-based analyses. Finally, I will present outlook for exciting future applications for AI analysis of LSS.

Review presentations / **132**

Generative models to assist sampling

Corresponding Author: marylou.gabrie@polytechnique.edu

Deep generative models parametrize very flexible families of distributions able to fit complicated datasets of images or text. These models provide independent samples from complex high-distributions

at negligible costs. On the other hand, sampling exactly a target distribution, such a Bayesian posterior or the Boltzmann distribution of a physical system, is typically challenging: either because of dimensionality, multi-modality, ill-conditioning or a combination of the previous. In this talk, I will discuss opportunities and challenges in enhancing traditional inference and sampling algorithms with learning.

Review presentations / **133**

TBD: Symbolic regression

Corresponding Author: miles.cranmer@gmail.com

Review presentations / **134**

TBD: Symmetries in deep learning

Corresponding Author: svillar3@jhu.edu

Review presentations / **135**

TBD: Deep learning and numerical simulations

Corresponding Author: rcroft@cmu.edu

Review presentations / **136**

TBD: Domain adaptation

Author: David Shih[1]

[1] *Rutgers University*

Corresponding Author: shih@physics.rutgers.edu

Review presentations / **137**

TBD: ML and Bayesian inference in cosmology (Replay)

Author: Jens Jasche[1]

[1] *Stockholm University*

Corresponding Author: jens.jasche@fysik.su.se

Review presentations / **138**

TBD: Symmetries in deep learning (Replay)

Corresponding Author: svillar3@jhu.edu

Review presentations / **139**

Generative models to assist sampling

Corresponding Author: marylou.gabrie@polytechnique.edu

Deep generative models parametrize very flexible families of distributions able to fit complicated datasets of images or text. These models provide independent samples from complex high-distributions at negligible costs. On the other hand, sampling exactly a target distribution, such a Bayesian posterior or the Boltzmann distribution of a physical system, is typically challenging: either because of dimensionality, multi-modality, ill-conditioning or a combination of the previous. In this talk, I will discuss opportunities and challenges in enhancing traditional inference and sampling algorithms with learning.

Review presentations / **140**

TBD: Domain adaptation (Replay)

Corresponding Author: shih@physics.rutgers.edu

Review presentations / **141**

TBD: Deep learning and numerical simulations (Replay)

Corresponding Author: rcroft@cmu.edu

Review presentations / **142**

Capitalizing on Artificial Intelligence for LSS Cosmology (replay)

Corresponding Author: tomasz.kacprzak@phys.ethz.ch

In this review talk, I will show how artificial intelligence can bring tangible benefits to cosmological analysis of large-scale structure.
I will focus on how the use of AI in the framework of Simulations-Based Inference to achieve scientific objectives that would not be attainable with classical 2-pt function analyses. I will show three avenues where, in my opinion, AI can bring the most benefits: reaching the information floor of limited survey data via SBI analysis, accelerating simulations for SBI, and breaking degeneracies between cosmological probes and astrophysical nuissance fields. I will discuss new challenges that

come with AI-based analyses. Finally, I will present outlook for exciting future applications for AI analysis of LSS.

Review presentations / **143**

TBD: Symbolic regression (Replay)

Corresponding Author: miles.cranmer@gmail.com

Posters / **144**

Beyond Summary Statistics: Leveraging Generative Models for Robust and Optimal Field-Level Weak Lensing Analysis

Author: Biwei Dai[1]

[1] *UC Berkeley*

Corresponding Author: biwei@berkeley.edu

Deep learning (DL) methods have demonstrated great potential for extracting rich non-linear information from cosmological fields, a challenge that traditional summary statistics struggle to address. Most of these DL methods are discriminative models, i.e., they directly learn the posterior constraints of cosmological parameters. In this presentation, I will make the argument that learning the field-level likelihood function using generative modeling approaches such as Normalizing Flows usually leads to more effective extraction of cosmological information. This approach also enables anomaly detection to improve the robustness of the analysis. To scale the modeling to high dimensional data and improve its generalization capabilities, we further incorporate physical inductive biases, such as symmetries and multiscale structure, into the architecture of the normalizing flow models. On mock weak lensing maps, I will show that the model leads to significant improvement in constraining power compared to power spectrum and alternative DL models. I will also show that it is able to detect domain shifts between training simulations and test data, such as noise miscalibration and baryonic effect, which, if left unaddressed, could introduce systematic biases in parameter constraints. Finally, I will also show some preliminary results of our ongoing work on applying this model to the field-level cosmic shear analysis for HSC.

Contributed talks / **145**

ChatGaia

Corresponding Author: nolan.koblischke@mail.utoronto.ca

Review presentations / **146**

Deep learning algorithms for morphological classification of galaxies

Galaxies exhibit a wide variety of morphologies which are strongly related to their star formation histories and formation channels. Having large samples of morphologically classified galaxies is

fundamental to understand their evolution. In this talk, I will review my research related to the application of deep learning algorithms for morphological classification of galaxies. This technique is extremely successful and has resulted in the release of morphological catalogues for important surveys such as SDSS, MaNGA or Dark Energy Survey. I will describe the methodology, based on supervised learning and convolutional neural networks (CNN). The main disadvantage of such approach is the need of large labelled training samples, which we overcome by applying transfer learning or by 'emulating'the faint galaxy population. I will also show current challenges for the classification of galaxy images with CNNs, such as the detection, classification and segmentation of low surface brightness features, which will be of great relevance for surveys such as HSC-SSP or ARRAKIHS, and our current plans for addressing them properly.

Debates / **147**

Concluding remarks

Corresponding Author: liciaverde@gmail.com