'Statistical Challenges for Next-Generation Astronomical Surveys (StatsChall)' Abstracts

Talks

Understanding the impact of emulators in our inference pipelines

Authors

Harry Bevins (1)

(1) University of Cambridge

Abstract

Neural network emulators are widely used in astrophysics and cosmology to approximate computationally expensive simulations inside inference loops, and their role will only grow as surveys like Rubin, Euclid, and the SKA generate data at unprecedented scale. However, understanding the uncertainty they introduce is crucial for responsible and reliable science. Researchers often measure emulator accuracy in signal space and use ad hoc rules of thumb to determine what is "good enough" for inference. We present a theoretically motivated approach to translate emulator error directly into an error on the recovered posterior, quantified via a Kullback–Leibler divergence. Under assumptions of linearity, uncorrelated noise, and a Gaussian likelihood, we provide a rigorous bound on the maximum incorrect information inferred when using an emulator. The resulting metric is instrument-aware, depending explicitly on the noise level, and can be used both to build confidence in inference products and as an instrument-aware loss function during emulator training. We demonstrate performance on mock observations from 21-cm instruments and argue that with principled accuracy metrics, emulators are a powerful and trustworthy component of modern inference pipelines.

Evaluating Systematic Biases in the Modelling of Double-source-plane Lenses from Wide-area Surveys

Authors

Duncan Bowden (1), Aprajita Verma (1)

(1) University of Oxford

Abstract

Double-source-plane lenses (DSPLs) are a unique and rare type of strong lens consisting of two source galaxies at different redshifts. They are particularly valuable for many applications due to the additional constraining power the second source provides on the mass of the deflector and the dependence of the ratio of Einstein radii on cosmology. By the end of the LSST and Euclid surveys, we will have found O(10^2) and O(10^3), a factor of over 100 more than those currently known. We must be prepared to model many systems, which can require 100s of parameters each, in a consistent and computationally efficient manner. We do so using fully forward models from Bayesian codes such as Herculens, which are able to perform parameter inference using Hamiltonian Monte Carlo on GPUs through an auto-differentiable framework. By exploring the systematic error introduced by mass model assumptions for the intermediate source using simulated lenses, we seek to determine how well we can recover the underlying truth. This will directly impact our ability to infer cosmological parameters from the unprecedentedly large samples of DSPLs from forthcoming surveys.

SHADO: Operator-Based Signal Analysis and Feature Extraction for Patchy Multi-Instrument Photometry

Authors

Danny Dixon (1), Adam Ingram (1)

(1) Newcastle University

Abstract

SHADO is a framework for constructing operators directly on raw time-series observations so we can (i) separate intrinsic signal morphology from sampling artefacts and noise, (ii) share information across channels (bands/instruments), and (iii) produce features, reconstructions, and uncertainties that remain consistent under irregular, gappy, multi-modal observing strategies such as ZTF x TESS.

Rather than forcing data onto a uniform grid, SHADO composes a stack of complementary operators: Spectrum (repeating structures), Orthogonal transforms (compact multiresolution bases), Diffusion (uncertainty-aware smoothing on irregular grids), Homology (scale-robust morphology summaries and anchor selection via lazy/witness complexes), and Action (physics-informed variational priors expressed as optimisation objectives).

We validate SHADO on simulated targets where TESS provides near-continuous sector coverage of ~27 days to “lock” morphology, while ZTF supplies long-baseline but gappy sampling that stress-tests aliasing and robustness. Finally, we show how SHADO plugs into our BLuE self-lensing model for quiescent MS-BH binaries: a fast, physically grounded model in which the observed flux is the product of gravitational self-lensing (sharp periodic flares), relativistic Doppler boosting, and ellipsoidal variations, with eccentric orbits handled via a phase-time mapping and multi-band predictions enabled through wavelength-dependent Doppler and limb-darkening prescriptions

From Months to Seconds: Accelerating Next-Generation Cosmological Inference with unimpeded for Bayesian Model Comparison and Tension Quantification Across Cosmological Surveys

Authors

Dily Duan Yi Ong (1), Will Handley (1), David Yallup (1)

(1) University of Cambridge

Abstract

As precision astronomical surveys deliver unprecedented data complexity, disparities between multi-probe datasets demand robust statistical methods for model comparison and tension quantification. However, calculating Bayesian evidences remains a computationally prohibitive bottleneck, often requiring thousands of CPU hours, limiting the rapid validation of new datasets and extended models. To address these challenges, I present unimpeded [arXiv:2511.04661, 2511.05470], an open-source, pip-installable Python library designed to democratise access to these expensive calculations. Conceptually analogous to the Planck Legacy Archive (PLA), which provides MCMC chains restricted to parameter estimation without direct access to Bayesian evidences, unimpeded is a reusable library of nested sampling chains and Bayesian evidences, enhanced by machine learning emulators utilising normalising flows. It spans a massive pre-computed grid across 13 cosmological surveys and 8 models, featuring built-in comprehensive tension quantification tools. unimpeded acts as a scalable inference tool, transforming months of supercomputer time into seconds on a laptop, and enabling systematic Bayesian model comparison and tension quantification using just a few lines of code within unimpeded. To demonstrate the framework's ability to untangle complex inter-dataset systematics, I will present a full Bayesian reanalysis of the DESI DR2 BAO data in combination with Planck CMB measurements, multiple supernova catalogues (Pantheon+, Union3, DES-SN5YR, DES-Dovekie), and DES-Y1 weak lensing [arXiv:2511.10631, 2603.05472]. While frequentist approximations previously reported a preference for dynamical dark energy w_0w_aCDM over ΛCDM, our direct Bayesian analysis reveals that the multi-probe combination actually favours the simpler ΛCDM model. I will explain how this systematic discrepancy arises from the Jeffreys–Lindley paradox. Using unimpeded, our tension metrics rapidly identified that the apparent preference for new physics was driven by a significant conflict between DESI DR2 and the original DES-SN5YR supernova sample within ΛCDM. Notably, once the recalibrated DES-Dovekie supernova calibration is applied, this tension is mitigated and the Bayesian evidence for dynamical dark energy vanishes. These findings highlight rigorous Bayesian tension quantification as a crucial safeguard against interpreting dataset systematics as new physics. Ultimately, unimpeded empowers researchers to rapidly extract robust scientific constraints and conduct cross-disciplinary tests in the precision multi-probe era.

Behind the Spectrum: Disentangling Galaxy Physics from Observables with Variational AutoEncoders

Authors

Benedict Van den Bussche (1), Sinan Deger (2), Hiranya Peiris (3), Stephen Thorp (2), Anik Halder (2), Daniel Mortlock (4), Boris Leistedt (5), Joy Gong (2), Madalina Tudorache (2), Gurjeet Jagwani (6)

(1) Institute of Astronomy & Kavli Institute for Cosmology, University of Cambridge
(2) Institute of Astronomy and Kavli Institute for Cosmology, University of Cambridge
(3) Institute of Astronomy, Kavli Institute for Cosmology, and Cavendish Laboratory, University of Cambridge, and The Oskar Klein Centre, Department of Physics, Stockholm University
(4) Astrophysics Group and Department of Mathematics, Imperial College London
(5) Astrophysics Group, Imperial College London, Blackett Laboratory
(6) Institute of Astronomy and Kavli Institute for Cosmology and Research Computing Services, University of Cambridge

Abstract

The mapping between galaxy spectra and the underlying physics is inherently non-linear and complex. With an unprecedented volume of data, high-precision interpretation and diagnostics of these observables are essential to understanding galaxy evolution across cosmic time. In this talk, I present a data-driven framework to decode this relationship using the pop-cosmos generative galaxy population model.
By leveraging a beta-Variational Autoencoder (beta-VAE), we identify the most informative astrophysical features encoded within spectral energy distributions (SEDs), compressing them into a compact, five-dimensional latent space. Our model reconstructs galaxy SEDs with high fidelity (2% fractional error) while successfully disentangling the underlying factors of variation of the spectrum. Using the information-theoretic metric Mutual Information (MI), we demonstrate that these learned latent variables map directly onto key physical processes and parameters: stellar mass, star formation rate, dust attenuation, metallicity, and ionisation. We also relate broadband colours and specific emission lines to the properties and their mapping.
Crucially, the emergent grouping of these properties and features in the latent space highlights the dominant physical drivers - stellar mass, recent star formation, dust, and different ionisation potentials - encoded within the observables. This work provides improved, quick galaxy evolution diagnostics with a robust, interpretable connection between SEDs and the physics within galaxies.

Statistical Challenges for Cosmology in the Era of Stage-IV Surveys (Invited)

Authors

Alessio Spurio Mancini (1)

(1) Royal Holloway, University of London

Abstract

The next generation of cosmological surveys — including Euclid, the Vera Rubin Observatory, and the SKA — will deliver datasets of transformative statistical power, mapping the large-scale structure of the Universe across unprecedented volumes and redshift ranges. Extracting robust cosmological constraints from these data is, however, far from straightforward. The challenges are manifold: forward models are computationally expensive and only approximate; likelihoods are often intractable due to complex survey geometries, systematics, and massive data compression requirements; posterior distributions are extremely high-dimensional; and the interplay between cosmological signal and astrophysical nuisances — baryonic feedback, intrinsic alignments, nonlinear galaxy bias — introduces significant modelling uncertainty.

In this talk, I survey these statistical challenges and discuss how modern machine learning and probabilistic programming methods are beginning to address them. Neural emulators can replace costly forward models while preserving the fidelity required for unbiased inference. Simulation-based inference circumvents the need for explicit likelihoods, enabling analyses at the field level or with optimal data summaries that go beyond two-point statistics. Differentiable pipelines allow gradients to flow through the full analysis chain, opening the door to scalable variational inference and efficient posterior sampling.

I will discuss both the genuine promise and the real limitations of these approaches, with particular attention to validation, prior sensitivity, and the conditions under which ML-accelerated inference can be trusted for precision cosmology.

Posters

Strategies Utilising High-Frequency Interferometric Data to Explore SETI Parameter Space

Authors

Louisa Mason (1)

(1) University of Manchester

Abstract

The search for extraterrestrial intelligence (SETI) spans multiple dimensions of parameter space - such as spatial volume, frequency and time resolution – to hunt for narrowband radio technosignatures. This has traditionally focused on the 'water hole' - 1.42 to 1.66 GHz – with only recent efforts extending into millimetre and submillimetre wavelengths. We explore key considerations to conduct high-frequency SETI surveys using ALMA and present the first technosignature survey using archival ALMA data at Band 3 (90.642 GHz and 93.151 GHz). Placing constraints on the prevalence of extraterrestrial transmitters for an individual survey is essential to understanding the searched parameter space for technosignatures. We propose the use of galactic modelling – using the Besançon Galactic Model – to simulate the stellar population (or 'bycatch') sensitive within the field of view of a radio telescope, overcoming restrictions from astrometric catalogues due to magnitude limits and confusion. This work offers a crucial scalable strategy to utilise interferometric radio data in the search for technosignatures, especially at higher frequencies, as well as determining a truer statistical representation of the extent searched by radio SETI surveys.

Bayesian Technosignature Detection

Authors

Kelvin Wandia (1)

(1) University of Manchester

Abstract

Radio technosignature searches produce millions of candidate signals per survey, yet the classification pipelines determining which candidates warrant a follow-up has remained fundamentally unchanged for decades, with signal-to-noise thresholding, heuristic radio frequency interference (RFI) filters, and manual visual inspection often used. These approaches do not quantify the confidence of individual classifications, neither do they produce probabilistic principled scoring for follow-up prioritisation. Null results, which are commonly encountered, are also converted into fixed effective isotropic radiated power (EIRP) instead of evolving probabilistic constraints. We propose a Bayesian multi-hypothesis inference framework, that addresses all three limitations by drawing directly on methods developed

for gravitational-wave detection, a field that faced structurally identical problems before the first LIGO detections. The designed framework is positioned as a downstream complement of existing machine learning pipelines, not a replacement, that is, machine learning finds the candidates and Bayesian inference determines what they mean.

Generating Missing Abundances in Large Spectroscopic Surveys APOGEExGAIA

Authors

Salma Louhaichy (1)

(1) Liverpool John Moores University

Abstract

Large spectroscopic surveys such as APOGEE provide high-dimensional chemical abundance measurements for hundreds of thousands of stars, but these datasets are intrinsically incomplete as many elemental abundances are missing or unreliable due to observational limitations, spectral degeneracies, and signal-to-noise variations... This structured missingness poses a significant challenge for downstream studies of Galactic chemical evolution.

We present a machine learning framework to infer missing stellar abundances by learning the underlying chemical abundance manifold. We first construct a "high-trust" dataset with explicit per-element reliability masks derived from observational uncertainties. A neural network baseline is trained using a heteroscedastic likelihood to predict abundances and associated uncertainties, achieving typical errors of ~0.09 dex and providing a probabilistic reference model.

We then develop a conditional generative model based on a Wasserstein GAN with gradient penalty, which generates complete abundance vectors conditioned on stellar parameters and observed elements. The model is trained using a combination of adversarial and reconstruction objectives to preserve both pointwise accuracy and the multivariate structure of chemical space.

We validate the method through controlled masking experiments, achieving reconstruction errors comparable to the neural network baseline, and demonstrate robustness across challenging regimes including dwarf stars, low signal-to-noise spectra, and metal-poor populations. The generative model preserves astrophysical structure, including element correlations and the α-enhancement sequence.

Unlike purely predictive neural networks, which optimise element-wise accuracy and tend to regress toward mean trends, the generative approach learns the full joint distribution of chemical abundances. This enables it to produce abundance completions that are not only accurate on average, but also globally consistent across elements, maintaining the intrinsic correlations imposed by nucleosynthesis and Galactic chemical evolution. As a result, the generated abundances better populate the underlying chemical manifold, avoiding unphysical combinations of elements and preserving population-level features such as sequences and scatter. This leads to higher-quality synthetic data that is more suitable for downstream astrophysical analyses than deterministic predictions alone.