Simulation Based Inference for Galaxy Evolution

Bristol, UK

9 – 12 April, 2024

Contents

Rationale

Inferring the physical parameters of galaxy populations from their observed properties is one of the key problems in galaxy evolution, and a necessary ingredient in many cosmological analyses. Typical approaches require an explicit or computable form of the likelihood, and significant computational resources. Simulation Based Inference (SBI) approaches circumvent this requirement, and when combined with machine learning methods, significantly reduce the computational cost of inference. Such approaches are necessary to process the millions to billions of sources from the latest generation of instruments.

Simulation based inference is currently being applied to many astrophysical problems which face similar computing barriers. In this workshop, we aim to bring together both experts in SBI as well as galaxy evolution experts with SBI experience, to discuss problems and solutions faced when applying SBI to problems in galaxy evolution. This will be a small, focused meeting to enable plenty of time for discussion and practical sessions.

Invited Interactive Sessions

Introduction to SBI: parameter inference & model comparison

Niall Jeffrey
University College London

Introduction to the LTU-ili Inference Pipeline

Matthew Ho
Institut d'Astrophysique de Paris

Out-of-distribution

Marc Huertas-Company
Instituto de Astrofisica de Canarias - IAC

Normalising Flows

Ioana Ciuca
Australian National University

Tools

Adam Carnall & Christopher Lovell
University of Edinburgh, University of Portsmouth

Engine Choice

Lucas Makinen
Imperial College London

Contributed Abstracts

CAMELS-SAM: untangling the galaxy-halo connection with machine learning in observational space

Lucia Perez
Center for Computational Astrophysics

While many advanced methods for the statistical analysis of cosmological data have been developed to handle the looming breadth of new astronomical observations, a common constraint is the availability of training cosmological simulation data, especially that which includes realistic galaxy formation physics and the volume and/or resolution necessary to match observations. CAMELS-SAM and its newest updates offer crucial and unique training data sets of realistic galaxies across an enormous range of cosmologies and galaxy physics formulations. As the larger-volume 'hump' of the Cosmology and Astrophysics with MachinE Learning Simulations (CAMELS) project, CAMELS-SAM uses semi-analytic models (SAMs) of galaxy formation to flexibly and quickly generate galaxies over 1000 dark-matter only simulations of L=10 h^{-1} cMpc and $N = 640^3$. We present: the updated galaxy catalogs using the Santa Cruz SAM with 1) up to 9 varied astrophysical parameters, with 2) complete photometry generated for all galaxies; and, 3) new galaxy catalogs with the L-Galaxies SAM. We show initial exploration of how well neural networks are able constrain OmegaM and sigma8 while marginalizing over both varied parameters within a particular SAM, as well as across distinct SAMs; and the many possibilities that the CAMELS-SAM datasets offer for the next generation of simulation-based inference on observed galaxies.

What can we learn from individual galaxies?

Francisco Villaescusa-Navarro
Flatiron Institute

Can we learn something about cosmology and or astrophysics from indivdual galaxies by combining simulations with machine learning?

Inferring stellar population properties with Bayesian Deep Learning

Patricia Iglesias-Navarro
Instituto Astrofisico de Canarias

Spectral energy distributions (SEDs) encode information about the stellar populations within galaxies. By investigating the properties of these stars, such as their ages, masses, and metallicities, we can gain insights into the underlying physical processes that drive the growth and transformation of galaxies over cosmic time, in particular, the triggers and quenching mechanisms of star formation. For this purpose, we explore an amortized implicit inference approach to estimate the posterior distribution of redshift, metallicity and non-parametric star formation histories (SFHs) of galaxies, i.e. star formation rate as a function of cosmic time, using both optical absorption spectra, and photometry from ACS and NIRCam filters. Fed with the MILES stellar population models, we generate a sample of synthetic SEDs to train and test our model. We show that our approach is capable of reliably estimating the mass assembly of an integrated stellar population with, crucially, well-calibrated uncertainties. Once trained, deriving the posteriors takes 1 second per galaxy, 1e5 times faster than classical MCMC sampling, being able to address a large number of galaxies, and also to perform a thick sampling of the posteriors, estimating both the deterministic trends and the inherent uncertainty of this highly degenerated inversion problem, so far inaccessible for more traditional methods. First, we apply our pipeline to real observations of very massive elliptical galaxies from the Sloan Digital Sky Survey and show that it recovers ranges of star formation rate over time consistent with the spectra, as well as the expected relation between age and velocity dispersion. As a preliminary work, we apply this method to resolved galaxies from the JWST Advanced Deep Extragalactic Survey (JADES), fitting every pixel to study gradients in the stellar populations properties, showing a good generalization to data. We believe that our framework, a machine-learning-based implicit inference applied to SED fitting, is remarkably promising to deal with the size and complexity of upcoming galaxy surveys.

Galaxy physical property inference from photometry and spectroscopy with BAGPIPES

Adam Carnall
Edinburgh University

I am the developer of the Bagpipes galaxy spectral fitting code, and have worked extensively on physical parameter recovery, in particular the effects of star-formation history priors and spectrophotometric calibration systematic uncertainties. I'd be happy to talk about these things, give a Bagpipes tutorial, or provide an update on JWST high-redshift science if any of these would be of interest. Also happy to just come along and listen, provided I don't end up with too many other commitments in the spring.

Modernizing SED Fitting with Machine Learning Techniques

Dhruv Zimmerman

University of Florida

Most inference about the formation and evolution of galaxies in the universe relies on the spectral energy distribution (SED) fitting to determine galaxy properties. With more large surveys and JWST results on the horizon, it is especially important that the SED modeling and star formation histories (SFHs) are correct for the community to draw appropriate conclusions about the evolution of the universe from observations. SED fitting relies on carefully constructing models with many moving pieces and assumptions, and as a result often has large uncertainties and even systematic biases. We leverage the ground truth knowledge of the physical properties of galaxies gained from simulations and run radiative transfer to produce synthetic spectra. We construct a machine learning-based model to estimate galaxy properties from our synthetic photometry that provides reasonable uncertainties and is robust to available photometry to try to improve on traditional SED fitting techniques for the use of the community. We also model the star formation histories (SFH) of very early universe galaxies by using simulations that feature bursty star formation histories.

Learning from Cosmological Simulations

Marc Huertas-Company

IAC

I will present efforts from our group for doing inference using the mock outputs of cosmological simulations. I will in particular describe efforts to perform robust inference of the mass accretion histories by using a foundational model trained on multiple simulations. I will also show a probabilistic approach for finding progenitor galaxies in observations, informed by models.

COSMOPOWER: neural cosmological emulators for scalable and efficient simulation-based inference

Alessio Spurio Mancini
Royal Holloway, University of London

COSMOPOWER is a state-of-the-art Machine Learning framework adopted by all major Large-Scale Structure (LSS) and Cosmic Microwave Background (CMB) international collaborations for acceleration of their cosmological inference pipelines. It achieves orders-of-magnitude acceleration by replacing the expensive computation of cosmological power spectra, traditionally performed with a Boltzmann code, with neural emulators. I will present recent additions to COSMOPOWER which render it into a fully-differentiable library for cosmological inference. I will demonstrate how it is possible to use its differentiable emulators to enable scalable and efficient simulation-based inference with hierarchical modelling. Leveraging the benefits of automatic differentiation, XLA optimisation, and the ability to run on GPUs and TPUs, COSMOPOWER significantly enhances the performance of neural density estimation for simulation-based inference, augmenting it with the simulator gradients. I will show how COSMOPOWER allows the user to create end-to-end pipelines that achieve unparalleled accuracy in the final cosmological constraints, due to their ability to efficiently sample an unprecedentedly large number of nuisance parameters for the modelling of systematic effects.

Impact of galaxy evolution modelling on simulation-based inference of large-scale structure

Maximilian von Wietersheim-Kramsta
University College London

Measuring large-scale structure through weak gravitational lensing and galaxy clustering is a powerful probe to infer the cosmological model of the Universe. As the precision and resolution of galaxy surveys improves, non-linear scales become more important, and standard assumptions, such as a linear galaxy bias may break down. To address this, I will show how fast lognormal random field simulations can be calibrated with measurements of the galaxy bias, and the baryon density from the FLAMINGO hydrodynamical simulations which incorporate realistic halo occupation distributions, baryonic physics and galaxy evolution models. I will also present how these can be applied to a simulation-based inference analysis of weak lensing surveys by showing the results of our cosmic shear analysis with SBI of the Kilo-Degree Survey's KiDS-1000 data release. Lastly, I will highlight how such models and SBI will help address the modelling/inference challenges facing upcoming stage IV galaxy surveys, such as Euclid.

Practical considerations for SBI

Peter Melchior

Princeton University

My talk will deal with questions that arise in practical application of SBI to larger galaxy samples, based on experience from the development of the methods SEDFlow and PopSED. In particular, I will discuss 1) the noise model, its validity and how it can best be used in SBI applications; 2) the effects of a precomputed simulation suite. This talk is meant to raise awareness for problems that can profoundly challenge the reliability of SBI and also draw attention to the lack of diagnostics to determine if such problems have actually occurred in a given analysis.

A flexible model of galaxy formation for SBI

Suchetha Cooray

National Astronomical Observatory of Japan

The critical requirement for inferring the physical parameters of galaxy populations from their observed properties with SBI will be a flexible galaxy formation simulator that potentially covers the real parameter space of galaxy formation. I am developing a flexible galaxy-halo connection model that simulates galaxy SED populations and field images by modeling consistent galaxy growth histories within cosmological structure formation. This model builds on the UniveseMachine (Behroozi et al. 2019) where I model dark matter halo property-dependent star formation rates, dust, metallicity and morphology to derive galaxy field images from dark matter-only simulations. SEDs are calculated using a stellar population synthesis, and SED-conditioned galaxy images are produced by a conditional diffusion model. Key outcomes of the model include a fully physical, self-consistent picture of galaxy stellar masses, star formation histories, dust, and metallicity from $z = 0$ to 15; significantly reduced uncertainties on the evolution of galaxies in dark matter halos; and mock catalogs and images for arbitrary current and future wide-field surveys that match the latest observations.

A physically motivated simulated sample of galaxies for machine learning-based identification of main sequence galaxies in the local Universe

Charalampos Daoutis
University of Crete

The analysis of the Spectral Energy Distributions (SEDs) of galaxies has become a powerful tool for deriving the properties of galaxies and investigating their evolution. Similarly, SEDs have been a powerful tool for exploring the observable properties of galaxies under different assumptions and evolutionary models. In this project, we employ a stellar population synthesis code, Flexible Stellar Population Synthesis (FSPS), to generate a simulated sample of galaxy SEDs representative of star-forming galaxies on the main sequence (MS) in the local universe. In order to ensure that the simulated SEDs are representative of the general galaxy populations the distributions of key SED parameters (e.g. star-formation history, metallicity, extinction, dust-related parameters) are based on the measured distributions of large galaxy samples. Following this process, a library of 797,000 SEDs from FUV to FIR is produced. Our results show that the simulated sample of normal star-forming galaxies exhibits spectroscopic and photometric properties that closely align with those observed in the local universe. This agreement validates the use of simulated data for training machine learning models. Therefore, using our simulated sample of galaxies and a machine learning algorithm we define a diagnostic tool for identifying MS star-forming galaxies based on custom bands applied to mid-IR spectra in the 5μ m - 30μ m range. These models, which are applicable to JWST MIRI spectra, achieve 90% completeness in the identification of MS galaxies. Finally, we aim to extend this methodology in order to infer dust parameters for large samples of star-forming galaxies from JWST archival data.

Simulation-based inference of cosmic shear with the Kilo-Degree Survey

Kiyam Lin
University College London

The standard approach to inference from cosmic large-scale structure data employs summary statistics that are compared to analytic models in a Gaussian likelihood with pre-computed covariance. To overcome the idealising assumptions about the form of the likelihood and the complexity of the data inherent to the standard approach, we employed simulation-based inference (SBI), which learns the sampling distribution as a probability density parameterised by a neural network. We first validated that SBI is a viable methodology by testing its application towards a suite of exactly Gaussian-distributed data vectors for the most recent Kilo-Degree Survey (KiDS) weak gravitational lensing analysis (Lin et al., 2023). New simulations were then constructed and used to train our neural density estimators. We compare the differences between our version of the fiducial KiDS analysis done using SBI vs.

our new analysis that include previously not modelled systematics such as variable depth and also compare the effects of different levels of Gaussianity imposed on the inference. We present the newest KiDS-1000 results analysed using SBI (von Wietersheim-Kramsta in prep.)

The Flagship simulation to infer galaxy physical parameters

Francisco Castander

ICE-CSIC, IEEC, Barcelona

The Flagship simulation is a large scale simulation that was produced to help the Euclid mission data processing and science analysis. We use the Flagship simulation galaxy mock catalogue to calibrate methods to infer the physical parameters of galaxies. We will present current work on inferring the stellar mass of DESI galaxies from the legacy survey photometry used to select the galaxy targets. We validate the performance against the spectroscopic measurements.

Get the Most out of Your Data: Simulation Based Inference and How to Think About Your Data Distribution

T. Lucas Makinen

Imperial College London

Simulation-based inference makes physical parameter inference possible from arbitrarily-shaped data which otherwise might be intractable. Data compression to informative (neural) summaries is an essential step in this process. I will give an overview of SBI and discuss approaches to choosing the best network for your data, with an emphasis on set-like data. I will present examples of graph- and set-based analyses for cosmology and galaxy evolution.

Simulation Based Inference in Galaxy Evolution with AstroFM Self-Supervised Pre-trained Models

Liam Parker
Flatiron Institute/UC Berkeley

We present AstroFM, a cross-modal foundation model for galaxies. Our approach takes information about galaxies from two separate modalities - images and optical spectra - and embeds and aligns both into a shared, physically meaningful embedding space using new-to-astronomy models and training regimes. Specifically, our training procedure consists of two distinct steps: first, we embed galaxy images and spectra separately by pre-training the first-ever transformer-based architectures for galaxy images and spectra in a self-supervised context, using self-distillation from DINO_v2 and a mask-filling strategy respectively. We then align these embeddings using a CLIP-inspired contrastive objective. We apply our method to multi-band images and spectrograms from the Dark Energy Spectroscopic Instrument (DESI) and demonstrate that our embeddings contain valuable physical information across a variety of downstream tasks. In the context of galaxy evolution simulation-based inference, we propose that our self-supervised, cross-modal embeddings can be used as effective summary statistics in a traditional SBI pipeline. Indeed, when starting from the pre-trained embeddings, downstream property estimation tasks can be performed by simply training a normalizing flow on a small set of known galaxies or high-fidelity simulations.

Modeling the Kinematics of Central and Satellite Galaxies Using Normalizing Flows

James Kwon
University of California, Santa Barbara

Galaxy clustering contains information on cosmology, galaxy evolution, and the relationship between galaxies and their dark matter hosts. On small scales, the detailed kinematics of galaxies within their host halos determines the galaxy clustering. In this paper, we investigate the dependence of the central and satellite galaxy kinematics on $\boldsymbol{\theta}$, the intrinsic host halo properties (mass, spin, concentration), cosmology ($\Omega_{\rm m}$, σ_8), and baryonic feedback from active galactic nuclei and supernovae ($A_{\rm AGN1}$, $A_{\rm AGN2}$, $A_{\rm SN1}$, $A_{\rm SN2}$). We utilize 2,000 hydrodynamic simulations in CAMELS run using IllustrisTNG and SIMBA galaxy formation models. Focusing on central and satellite galaxies with $M > 10^9 M_*$, we apply neural density estimation (NDE) with normalizing flows to estimate their $p(\Delta r \,|\, \boldsymbol{\theta})$ and $p(\Delta v \,|\, \boldsymbol{\theta})$, where Δr and Δv are the magnitudes of the halo-centric spatial and velocity offsets. With NDE, we accurately capture the dependence of galaxy kinematics on each component of $\boldsymbol{\theta}$. For central galaxies, we identify significant spatial and velocity biases dependent on halo mass, concentration, and spin. For satellite distributions, we find significant deviations from an NFW profile and evidence that they consist of distinct orbiting and infalling populations. However, we find no significant dependence on $\boldsymbol{\theta}$ besides a weak dependence on host halo spin. For both central and

satellite galaxies, there is no significant dependence on cosmological parameters and baryonic feedback. These results provide key insights for improving the current halo occupation distribution (HOD) models. This work is the first in a series that will re-examine and develop HOD frameworks for improved modeling of galaxy clustering at smaller scales.

Simulation based inference for galactic outflows using NIHAO and GECKOS

Anna Lena Schaible
Heidelberg University

The formation and evolution of galaxies is a complex and multiscale problem, where gas flows play an important role. Cosmological simulations have illustrated that feedback mechanism from stellar winds, radiation fields, supernovae and active galactic nuclei (AGNs) can effectively drive galactic winds. However, the details of the driving mechanisms behind gas in- and outflows are still unclear. Also, many cosmological simulations overlook Cosmic Rays (CRs, feedback from relativistic particles), which are another source of non-thermal feedback. We use the state-of-the-art cosmological hydrodynamical simulation of Milky Way like galaxies simulations to study the effect of different feedback processes on gas in- and outflows at different gas temperatures to investigate in the underlying mechanisms driving gas flows. To validate and extend our findings, we study the edge-on spiral galaxies from the GECKOS (Generalising Edge-on galaxies and their Chemical bimodalities, Kinematics, and Outflows out to Solar environments) survey, a large MUSE program recently granted. We plan to use simulation-based inference to test the different feedback processes in our simulations on the observational data.

Self-consistent population synthesis of AGN from observational constraints in the X-rays

Dimitra Gerolymatou
University of Geneva

The cosmic X-ray background (CXB) is produced by the collective emission of the whole population of active galactic nuclei (AGN), thus providing key information about the emission properties of the AGN population. Equally important, X-ray surveys of AGN provide direct constraints on the properties of individual AGN, such as their luminosity and obscuration. Previous AGN population synthesis models have not addressed such constraints self-consistently, i.e. by intrinsically linking obscuration and reflection. Here we use a simulation-based inference (SBI) tool to determine the geometrical and physical properties of the AGN population. We perform numerical simulations with our ray-tracing code, RefleX, which allows the self-consistent modelling of the X-ray emission of AGN with flexible circumnuclear geometries. We create our synthetic population by sampling an X-ray luminosity function and we use the RefleX-simulated emission of the AGN population to construct gradually more complex physically motivated geometrical models and simultaneously reproduce the CXB and obscuration properties obtained in different X-ray surveys, i.e. the observed NH distribution and fraction of absorbed AGN. Finally, we explore the relationship between reflection and obscuration and derive the intrinsic fraction of Compton-thick AGN.

A Universe To Be Decided: Towards Specialized Foundation Models for Advancing Astronomy

Ioana Ciuca
Australian National University

I discuss the application of Foundation Models in Astronomy through the collaborative efforts of the UniverseTBD consortium with a mission to democratize Science for everyone. One of our key objectives is to overcome the limitations of general-purpose Foundation Models, such as producing limited information in specialized fields. To this end, we have developed the first specialized large language model for Astronomy, AstroLLaMa-1. This model, enhanced by exposure to domain-specific literature from the NASA Astrophysics Data System and ArXiv, demonstrates improved text completion and embedding capabilities over existent GPT models. I further discuss the potential of LLMs in generating complex scientific hypotheses and extracting meaningful insights from astronomy literature. Our findings, validated by human experts, demonstrate the LLM capability in informed scientific critique and uncover intriguing patterns in the embedding space, highlighting the potential of LLMs to augment scientific inquiry. I will also discuss preliminary work with the multi-modal model AstroLLaVA, which allows us to interact with astronomical images via natural language. Through the work of UniverseTBD, we aim to explore how artificial intelligence can assist human intelligence in Astronomy and, more broadly, Science.

A Hierarchy of Normalising Flows for Modelling the Galaxy-Halo Relationship

Christopher Lovell
University of Portsmouth

We will present a unified model of the conditional dependence of galaxy properties on cosmological, astrophysical and host halo parameters. Using a large sample of galaxies taken from the Cosmology and Astrophysics with MachinE Learning Simulations (CAMELS) project, a suite of hydrodynamic simulations varying both cosmological and astrophysical parameters, we train a normalizing flow (NF) to map the probability density of various galaxy and halo properties. We simultaneously condition on the host subhalo mass of each galaxy, as well as the values of Ω_m, σ_8, two parameters describing the strength of supernovae feedback, and two describing the strength of AGN feedback. The model successfully reproduces the joint and conditional dependence of the various parameters, and by leveraging the learnt conditional relationships we can explore a wide range of interesting questions, whilst enabling simple marginalisation over nuisance parameters. The NF architecture also allows for rapid sampling and scoring. The former allows us to treat the model as a generative model; we demonstrate this by using the conditional dependence on subhalo mass to map galaxies to dark matter halos in N-body simulations, reproducing the galaxy stellar mass function for arbitrary values of our conditional parameters. We also use the scoring capabilities of the NF to treat the model as an inference engine, allowing us to carry out simulation based inference provided the properties of a single galaxy. Finally, we combine the generative model within a simulation based inference framework to infer cosmological and astrophysical parameters from key galaxy distribution function in a self-consistent way. The model represents a unique and flexible approach to modelling the galaxy–halo relationship.

Sequential Neural Score Estimation: Likelihood-Free Inference with Conditional Score Based Diffusion Models

Jack Simons

University of Bristol

We introduce Sequential Neural Posterior Score Estimation (SNPSE), a score-based method for Bayesian inference in simulator-based models. Our method, inspired by the remarkable success of score-based methods in generative modelling, leverages conditional score-based diffusion models to generate samples from the posterior distribution of interest. The model is trained using an objective function which directly estimates the score of the posterior. We embed the model into a sequential training procedure, which guides simulations using the current approximation of the posterior at the observation of interest, thereby reducing the simulation cost. Importantly, there are no restrictions on our neural-network architecture, thus avoiding the paradigm of normalising flows.

Author Index