

LETTER • OPEN ACCESS

DIGS: deep inference of galaxy spectra with neural posterior estimation

To cite this article: Gourav Khullar *et al* 2022 *Mach. Learn.: Sci. Technol.* **3** 04LT04

View the [article online](#) for updates and enhancements.

You may also like

- [GALAXY COLOR AND LARGE SCALE STRUCTURE](#)
Nancy Ellman
- [Stellar Populations in Gas-rich Galaxy Mergers. I. Dependence on Star Formation History](#)
Kenji Bekki and Yasuhiro Shioya
- [The Effect of Dense Cluster Environments on Galactic Properties](#)
Ethan Cronk, Matthew B. Bayliss and Keunho Kim



LETTER





DIGS: deep inference of galaxy spectra with neural posterior estimation

OPEN ACCESS

RECEIVED
3 August 2022REVISED
15 September 2022ACCEPTED FOR PUBLICATION
10 October 2022PUBLISHED
28 December 2022

Original Content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#).

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Gourav Khullar^{1,2,3,4,5,*} , Brian Nord^{1,2,3} , Aleksandra Ćiprijanović¹ , Jason Poh^{2,3}  and Fei Xu^{2,3}¹ Fermi National Accelerator Laboratory, Batavia, IL 60510, United States of America² Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, United States of America³ Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, United States of America⁴ Kavli Institute for Astrophysics & Space Research, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, United States of America⁵ Department of Physics and Astronomy and PITT PACC, University of Pittsburgh, Pittsburgh, PA 15260, United States of America

* Author to whom any correspondence should be addressed.

E-mail: gourav.khullar@pitt.edu**Keywords:** simulation-based inference, neural posterior estimation, galaxy evolution, spectroscopy, spectral energy distribution fitting, deep learning, sky surveys**Abstract**

With the advent of billion-galaxy surveys with complex data, the need of the hour is to efficiently model galaxy spectral energy distributions (SEDs) with robust uncertainty quantification. The combination of simulation-based inference (SBI) and amortized neural posterior estimation (NPE) has been successfully used to analyse simulated and real galaxy photometry both precisely and efficiently. In this work, we utilise this combination and build on existing literature to analyse simulated noisy galaxy spectra. Here, we demonstrate a proof-of-concept study of spectra that is (a) an efficient analysis of galaxy SEDs and inference of galaxy parameters with physically interpretable uncertainties; and (b) amortized calculations of posterior distributions of said galaxy parameters at the modest cost of a few galaxy fits with Markov chain Monte Carlo (MCMC) methods. We utilise the SED generator and inference framework Prospector to generate simulated spectra, and train a dataset of 2×10^6 spectra (corresponding to a five-parameter SED model) with NPE. We show that SBI—with its combination of fast and amortized posterior estimations—is capable of inferring accurate galaxy stellar masses and metallicities. Our uncertainty constraints are comparable to or moderately weaker than traditional inverse-modelling with Bayesian MCMC methods (e.g. 0.17 and 0.26 dex in stellar mass and metallicity for a given galaxy, respectively). We also find that our inference framework conducts rapid SED inference ($0.9\text{--}1.2 \times 10^5$ galaxy spectra via SBI/NPE at the cost of 1 MCMC-based fit). With this work, we set the stage for further work that focuses on SED fitting of galaxy spectra with SBI, in the era of JWST galaxy survey programs and the wide-field Roman Space Telescope spectroscopic surveys.

1. Introduction

Understanding the mass assembly of galaxies across cosmic time is a major goal of modern extragalactic astrophysics; solving this question sheds light onto a galaxy's underlying formation and evolution mechanism. Galaxies are well-characterized by features like stellar mass, chemical composition, dust attenuation, current star formation rate (SFR), and the star formation history. These parameters can be accurately inferred from a galaxy's spectral energy distribution (SED).

Within the last two decades, photometry-based SED fitting has become a pivotal method to measure the above properties. Ground-based telescopes have been used extensively for large multi-wavelength galaxy surveys—e.g. Sloan Digital Sky Survey (SDSS, Ahumada *et al* 2020), Dark Energy Survey (Abbott *et al* 2018), and DESI Legacy Imaging Surveys (Dey *et al* 2019)—producing large high-quality complex datasets. However, SED studies relying on photometry alone are subject to many challenges, such as the age-metallicity-dust degeneracy (Worthey 1994, Ferreras *et al* 1999). SED fitting using spectra mitigate this

challenge significantly, especially to constrain galaxy SFRs, formation timescales, and metallicities, with measurement of absorption line indices and emission line strengths (e.g. Worthey 1994, Leja *et al* 2019b and references therein.)

There are several cutting-edge SED-fitting pipelines with Bayesian frameworks that use Markov chain Monte Carlo (MCMC) methods to infer galaxy properties—e.g. CIGALE, MAGPHYS, and Prospector (Leja *et al* 2017a, 2019a, Carnall *et al* 2019, Johnson *et al* 2021). However, the computational time needed by the fitting algorithms in these frameworks—e.g. MCMC or nested sampling has been recently is a major bottleneck. With the next generation of telescopes, like the Vera Rubin Observatory/ Legacy Survey of Space and Time (Ivezić *et al* 2019) and the Dark Energy Spectroscopic Instrument (DESI; Aghamousa *et al* 2016), tens of millions of optical and infrared galaxy photometry and spectra will be measured. Highly-resolved spectra have higher information content and require more complex and flexible models for fitting. Finally, datasets with spaxels (or spatial and spectral pixels) are increasing in number, e.g. data units in integral-field-unit (IFU) spectrograph observations, with JWST IFU spectroscopy of a gravitationally lensed galaxy (Khullar *et al* 2021), or the SDSS-MaNGA survey observations of star forming galaxies.

Fitting a large number of free parameters is computationally expensive. A five-parameter spectral model within a typical SED fitting code—stellar mass, dust attenuation, metallicity, age—converges to a best-fit model solution in 2–10 CPU h. Moreover, each galaxy spectra requires its own separate inference chains. With the advent of spectroscopic surveys that may potentially observe tens of millions of galaxy spectra, the cost of modelling spectra quickly becomes prohibitive. Thus, the need of the hour is to quickly and reliably deduce the physical parameters of galaxies in large surveys, as well as to rapidly analyse thousands of spectra from within one galaxy.

Deep learning applied to galaxy SED fitting allows, in principle, a mapping between an observed SED and the target galaxy's star formation history, with several studies in the last few years alone demonstrating the efficacy of new methodologies (Leung and Bovy 2019, Lovell *et al* 2019, Hahn and Melchior 2022). These methods allow regression of galaxy parameters, albeit without any uncertainty quantification.

Recent developments in deep learning methods have focused on uncertainty quantification. For example, deep ensembles involve retraining a network many times with different initializations to enable uncertainty quantification of the model outputs. (Ganaie *et al* 2021). Furthermore, with Bayesian neural networks (BNNs), deterministic weights of the model are replaced by probability distributions, which allows the model to provide uncertainties of its outputs (Valentin Jospin *et al* 2020) (training a BNN includes assumptions of priors over the network weights and assumes that parameter posteriors can be approximated by well behaved variational distribution e.g. Gaussian distribution or a mixture of multiple Gaussians).

Simulation-based inference (SBI; Cranmer *et al* 2019) combined with deep learning can mitigate assumptions (e.g. tractable likelihood) that can plague analytic likelihood and posterior modelling, as well as remove computational bottlenecks in statistical calculations. Many astrophysical studies have demonstrated success with SBI in calculating posterior estimation in a rapid manner (Kacprzak *et al* 2018, Alsing *et al* 2019, Zhang *et al* 2021, Huppenkothen and Bachetti 2022, Zhao *et al* 2022).

SBI of galaxy spectra is an exciting opportunity because these models (a) do not require the expression of an explicit likelihood, and (b) can calculate approximate posterior distributions of galaxy parameters efficiently, allowing for robust uncertainty quantification. Recent work has shown that photometric SED data can be used with SBI for fast inference, e.g. Hahn and Melchior (2022), and Robeyns *et al* (2022).

In this work, we demonstrate a proof-of-concept SBI framework to analyse galaxy spectra and recover posteriors efficiently for a five-parameter SED model. We expect to scale this work to upcoming galaxy spectroscopic surveys—like the Dark Energy Spectroscopy Instrument Aghamousa *et al* (2016) and the Roman Space Telescope High-Latitude Survey Wang *et al* (2022)—and for SED models with more complex and flexible descriptions of galaxy properties.

In section 2, we describe the simulated data used in this study. In section 3, we describe the SBI network architecture and analysis, and in section 4, we share our results and next steps. The fiducial cosmology model used for all distance measurements as well as other cosmological values assumes a standard flat cold dark matter (or Λ CDM) Universe with a cosmological constant (Λ CDM), corresponding to WMAP9 observations (Hinshaw *et al* 2013).

2. Data

We use Prospector (Johnson *et al* 2021) to generate simulated SEDs of galaxies. Prospector relies on MCMC sampling for stellar population synthesis and parameter inference. It is based on the Python-FSPS framework, with the MILES stellar spectral library and the MESA isochrones & stellar tracks (MIST) set of isochrones (Conroy and Gunn 2010, Falcón-Barroso *et al* 2011, Foreman-Mackey *et al* 2013, Choi *et al* 2016, Leja *et al* 2017a).

Table 1. Simulated SEDs: model description and prior range.

Parameter	Description	Priors
$M_{\text{total}}(M_{\odot})$	Total stellar mass formed	Log_{10} Uniform: $[10^8, 10^{13}]$
$\log(Z/Z_{\odot})$	Stellar metallicity in units of $\log(Z/Z_{\odot})$	Uniform: $[-2.0, 0.2]$
$\tau_{\lambda,2}$	Diffuse dust optical depth	Tophat: $[0.1, 10.00]$
t_{age}	Age of Galaxy (Gyr)	TopHat: $[0, 4]$
τ	e-folding time of SFH (Gyr)	Log_{10} Uniform: $[0.1, 1.0]$

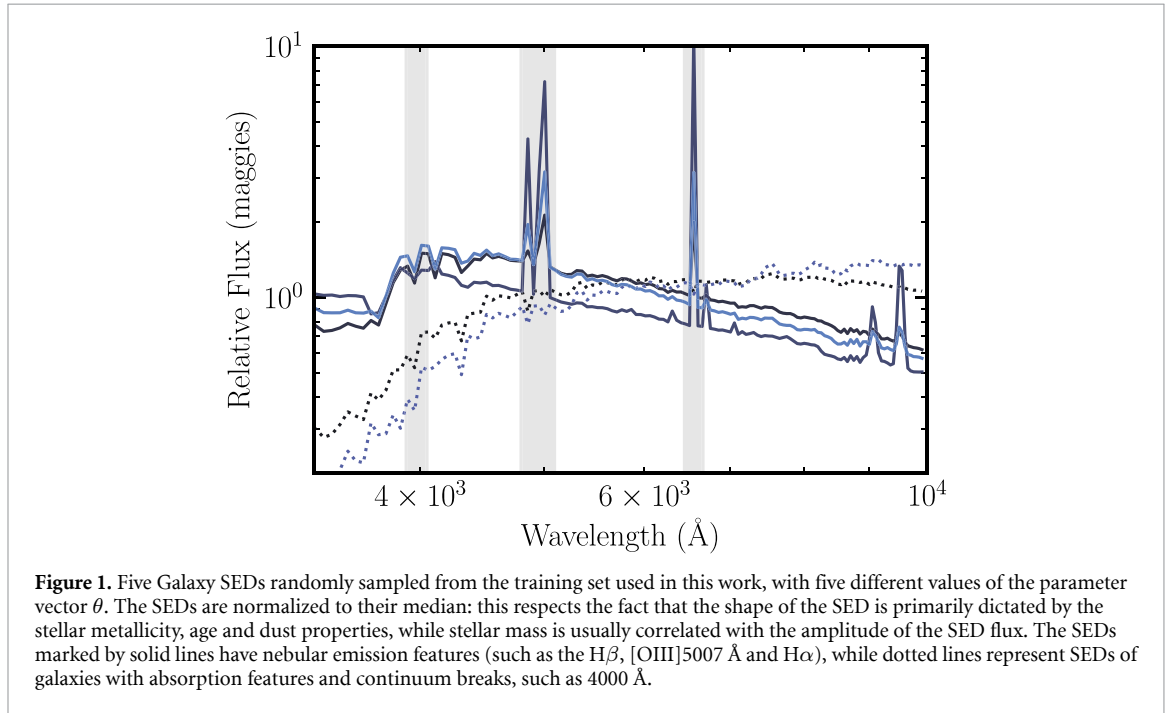


Figure 1. Five Galaxy SEDs randomly sampled from the training set used in this work, with five different values of the parameter vector θ . The SEDs are normalized to their median: this respects the fact that the shape of the SED is primarily dictated by the stellar metallicity, age and dust properties, while stellar mass is usually correlated with the amplitude of the SED flux. The SEDs marked by solid lines have nebular emission features (such as the $H\beta$, $[OIII]5007 \text{ \AA}$ and $H\alpha$), while dotted lines represent SEDs of galaxies with absorption features and continuum breaks, such as 4000 \AA .

We generate a training set of 10 000 rest-frame SEDs using a five-parameter model, with a delayed, exponentially declining (i.e. delayed-tau) star formation history. The SFH—SFR as a function of time—is:

$$\text{SFR}(t, \tau) \propto t/\tau * e^{-t/\tau} \quad (1)$$

where SFR is the star formation rate, t is the epoch at which the star formation history is being evaluated, and τ corresponds to the e-folding time in the delayed- τ SFH model.

This model incorporates physical priors used in survey studies of galaxy mass assembly (e.g. see Belli *et al* 2019). We sample the total stellar mass (M_*), the stellar metallicity $\log(Z/Z_{\text{sol}})$, a delayed and exponentially declining SFH with age t_{age} , the e-folding time τ_{age} (τ from here on), and dust attenuation ($\tau_{\lambda,2}$, corresponding to the optical depth of diffuse dust at 5500 \AA ; from here on referred to as dust.) Each parameter vector θ comprises these five parameters. Our SED model assumes a Kroupa Initial Mass Function (Kroupa 2001). Nebular continuum and line emission are also present.

See table 1 for a description of the prior range for each model parameter. See figure 1 for five example SEDs/spectra from our training set, highlighting the emission and absorption features depending on the type of galaxy (young vs old stellar populations, respectively).

We smooth and resample the simulated SEDs to resemble a medium-resolution spectroscopic survey using Prospector’s internal resampling utility. We use a velocity smoothing parameter (σ_v (km s^{-1}), to account for the contribution of Doppler broadening by stellar velocities, and resolution of the model libraries), and fix it at 350 km s^{-1} ; this smoothing corresponds to $R \sim 100$, similar to a deep galaxy survey conducted with the $R \sim 100$ JWST/NIRSpec prism (Zackrisson *et al* 2017). This results in a training set with each galaxy SED sampling rest-frame $3750\text{--}9500 \text{ \AA}$, with 138 flux elements for each SED, which is the data vector \mathbf{x} corresponding to each θ .

To map our training set to observations, we add stochasticity to the training set in the form of Gaussian noise, to the level of 5% of the flux at a given wavelength, representative of real data at signal-to-noise ratio $\text{SNR} \sim 20$. We conduct data augmentation to scale 10 000 noise-less spectra in our base training set to 2×10^6 spectra with Gaussian noise used in our SBI framework (see section 3).

We also create an additional test of spectra conduct posterior diagnostics. We generate a set of 1000 spectra with noise, from the same parameter prior range.

3. Inference methodology

3.1. SBI

Our objective is to calculate posterior distributions $p(\theta|\mathbf{x})$ of the galaxy parameters derived from a typical SED analysis, where θ is the set of galaxy properties, and \mathbf{x} represents the galaxy spectra. We do this by training our SBI model on the large stochastically-sampled training set of SEDs described in section 2. We utilise Neural Posterior Estimation (NPE) (Papamakarios and Murray 2016, Greenberg *et al* 2019)) which relies on neural networks to train on simulated SEDs with realistic noise, and allow us to estimate ‘amortized’ posterior distributions.

SBI/NPE requires computational time in advance of the actual inference, and evaluates the posterior for different observations without having to re-run inference (this is known as amortization). This ‘amortized’ calculation of posteriors then allows us to infer the posteriors of a ‘real’ galaxy with computational time <1 s. For more details and examples of amortized neural network-based posterior estimation, see Greenberg *et al* (2019) or section 2 of Hahn and Melchior (2022). We provide a short summary below.

NPE uses ‘normalizing flows’ (Tabak and Turner 2013) as a density estimator, which employs an invertible bijective transformation to map a complex distribution (i.e. the true posterior distribution in SED model parameter space) to a simpler and faster-to-calculate distribution (often Gaussian, or a combination of Gaussians). This results in the calculation of approximate posterior distributions, that are assumed to be a good approximation of the underlying posterior distributions of parameters. In particular, we use Masked Autoregressive Flow (MAFs; Uria *et al* 2016, Papamakarios *et al* 2017) incorporated within `sbi` (similar to Hahn and Melchior 2022). MAFs perform well in modelling conditional probability distributions, such as posteriors (see section 2 of Hahn and Melchior 2022 for more details).

3.2. This work

See figure 2 for a depiction of the SBI architecture used in this work. We use a supervised learning pipeline within an SBI framework via the Macke Lab `sbi` toolbox (Tejero-Cantero *et al* 2020). To demonstrate a proof-of-concept, we train on 2×10^6 simulations (where each simulation is a noise-added version of an SED in our training set) in an NPE framework. We use 25 hidden units and 10 transform layers without an embedding network in this framework. Our model trains on features in the raw simulated data; this model converges after 87 epochs and takes ~ 14 h to train. This analysis generates the set of approximate posterior distributions for our five parameter SED model. We also test on other combinations of hidden units and transform layers, and choose the above as the fiducial choice with more robust results.

3.3. Posterior diagnostics

We evaluate the results using a variety of statistical and diagnostic mechanisms. First, to test the precision of our framework, we compare the recovered SED parameter values with the true values θ of the parameters from our test set. Secondly, to test the health of the calculated posteriors, we also perform posterior predictive checks (PPCs) and simulation-based calibration (SBC) checks (Talts *et al* 2018). PPCs validate that the model SEDs corresponding to the distribution of θ values in a given posterior fall within the allowed range. We do this by cross-checking whether our best-fit model-based spectrum looks similar to observed data x (see figure 3).

SBCs provide a quantitative insight into whether the posterior uncertainties are balanced. In this test, we sample θ_i values from the priors, and simulate observations (using our simulator) from these parameters. Following this, we perform inference given each of these observations, which generates SBC posteriors of their own. For a healthy posterior, the SBC ranks of ground truth parameters under the inferred posteriors should follow a uniform distribution (rank plots aid in visually confirming this; see figure 5).

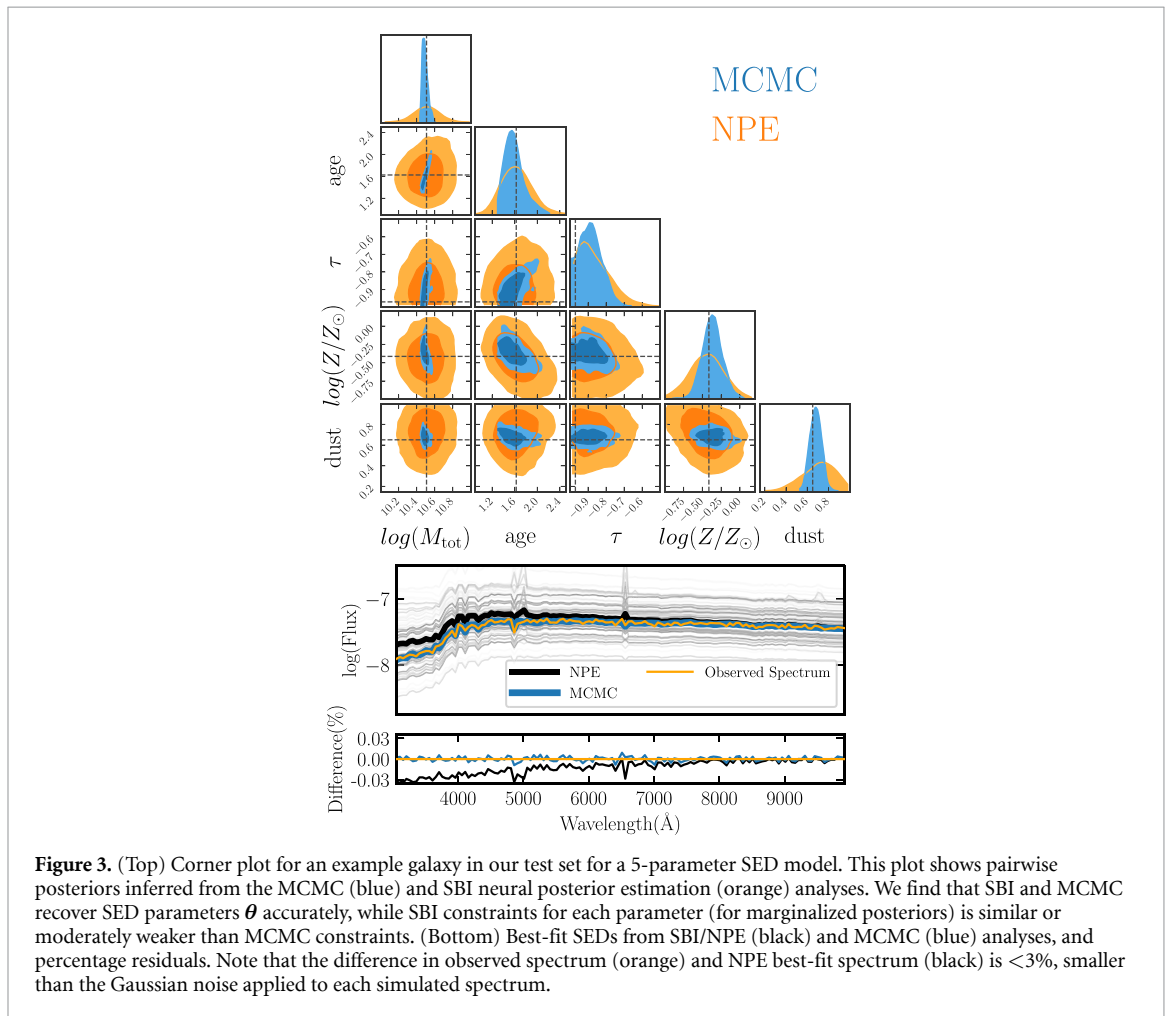
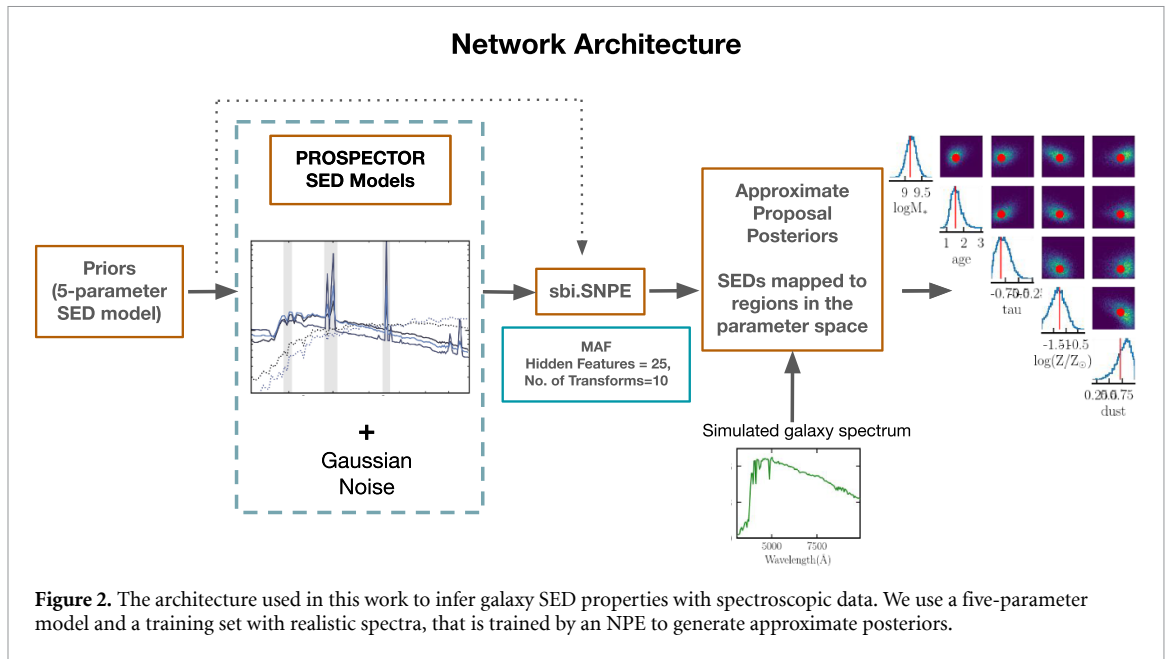
Finally, we compare our results to an MCMC analysis of representative SEDs from the test set.

3.4. MCMC analysis

We fit the same five-parameter SED model to representative galaxy SEDs in the test set using the inference framework Prospector, which calculate Markov-Chain Monte Carlo (MCMC)-based posterior distributions.

We assume the following likelihood function:

$$\ln \mathcal{L}_{\text{spec}}(\mathbf{x}, \theta, \sigma) = \sum_{i=1}^n \ln \left(\frac{1}{\sqrt{2\pi}\sigma_i} \right) - \frac{1}{2} \sum_{i=1}^n \frac{(\mathbf{x}_i - \mathbf{m}(\theta))^2}{\sigma_i^2}$$



where (x, σ_i) are n independent spectral flux elements assumed to be drawn from a Gaussian distribution, and $m(\theta)$ corresponds to the model spectrum for a given parameter set θ .

We use the same θ prior range and shape as the SBI/NPE analysis, in order to calculate the posterior $p(\theta|x)$. We use emcee Foreman-Mackey *et al* (2013) to conduct the MCMC posterior sampling with 128 walkers, 128 iterations and a burn-in with the step set [4096, 4096, 2048, 512].

Note that non-Gaussian or correlated uncertainties are seen in spectral datasets (e.g. magnitude upper limits in the case of non-detections), which are not accurately captured by the above likelihood, making a ‘likelihood-free inference’ like SBI the ideal choice for this analysis. The results from the SBI analysis, posterior diagnostics, and MCMC comparison are shown in section 4.

3.5. Computing resources

For our SBI analysis, we use the Python 3 Google compute engine backend (with the CPU processor AMD EPYC 7B12), which for our network architecture takes ~ 14 CPU hours to train 2×10^6 simulated spectra with noise. For every subsequent posterior estimation, this setup takes ~ 0.3 s.

For our serialized MCMC calculations, we utilise Prospector runtime on a 2.7 GHz Quad-Core Intel Core i7 processor, which takes ~ 14 CPU hours to converge.

4. Results

In this proof-of-concept analysis, we test out to set whether an SBI framework can train on realistic noisy galaxy spectra to estimate amortized posteriors robustly. To test the efficacy of our SBI framework, we use a test set of 1000 randomly sampled SEDs from our prior range (see table 1 and section 2).

One such result is shown in the top panel of figure 3, for a galaxy with $\log M_{\text{tot}} = 10.51 (M_{\odot})$, $\log(Z/Z_{\odot}) = -0.41$, age = 1.63 Gyr, $\tau = 0.11$ Gyr, and dust = 0.65 (a metal-poor dusty galaxy). We plot pairwise posterior distributions estimated from both the MCMC (in blue) and SBI (or NPE, in orange) in order to compare constraints across the five-parameter SED model. The truth values are plotted with black dotted lines. In the bottom panel of figure 3, we show the maximum *a posteriori* SED models and model residuals from the SBI/NPE (black) and MCMC (blue) analyses. Also overplotted are 1000 randomly sampled SEDs from the posteriors (in grey).

We find excellent agreement between the median values of parameters across MCMC and SBI/NPE posteriors (when marginalized over other parameters); these values are also accurate relative to the true parameter values θ . We also observe that the age and metallicity constraints are similar in both analyses for this test galaxy, while MCMC mass and dust estimation is more precise relative to SBI/NPE. This is the first demonstration that the proof-of-concept analysis presented here is effective at recovering galaxy SED parameters with spectroscopic observations.

On running inference on a sample of 1000 test galaxies sampled from the 16th–84th percentile range of priors in this study, we find accurate recovery of SED parameters. See figure 4 for a comparison between true and recovered values of each parameter, as well as χ plots to show goodness-of-fit across the simulated spectroscopic dataset. The recovered values here are the 50th percentile parameter values, and the uncertainties correspond to confidence intervals between 16th–84th percentile in the posteriors. Here, we demonstrate accurate and precise parameter recovery across the majority of the prior range, specifically for stellar mass ($\log M$), while the constraints on metallicity and age are seen to be wider and less precise. We also note that in our analysis, the recovered parameters are the most biased at the edges of the prior ranges, which indicates that the underlying posterior distribution is not being captured in these parameter ranges. For example, the 16th, 50th and 84th percentile of parameter values for a given galaxy are not accurate descriptors of the underlying posteriors near the edges of the prior range. This can be potentially solved by training on a spectroscopic dataset sampled from a prior range marginally wider than the target spectroscopic survey. We also find no substantial difference in the quality of parameter recovery for star forming galaxies (emission line galaxies; median values of the age parameter < 0.75 Gyr), or quiescent systems (galaxies with spectra containing strong absorption line indices; median values of the age parameter > 2.5 Gyr).

We also run extensive PPC and SBC checks to test the accuracy and precision of our SED parameter values, where we find that the posterior distributions in this analysis are well converged. See figure 5 for rank distributions for each parameter—a healthy posterior follows uniform distribution (non-uniform distributions indicates a poorly calibrated posterior; Talts *et al* 2018). This demonstration of well-calibrated uncertainties in our SBI analysis is confirmed in figure 6, where we plot the rank cumulative distribution function on the left panel. In both figures, the grey region corresponds to the 95% confidence interval of a uniform distribution, which our parameter rank distributions follow.

The right panel of figure 6 shows the probability coverage curve of each SED parameter θ . The principle behind a probability coverage plot is as follows: a well-calibrated posterior estimator will produce—for an ensemble of SEDs in the test set—parameter uncertainties that accurately reflect the true underlying uncertainty in the ensemble, e.g. a 68% posterior volume will contain 68% of the true SED parameter values of the test ensemble. Generalizing from this, by plotting the fraction of SED parameters in the test set which fall within the posterior volume as a function of the posterior volume, we obtain curves such as those seen in the right panel of figure 6. A well-calibrated estimator will produce probability coverage curves that closely

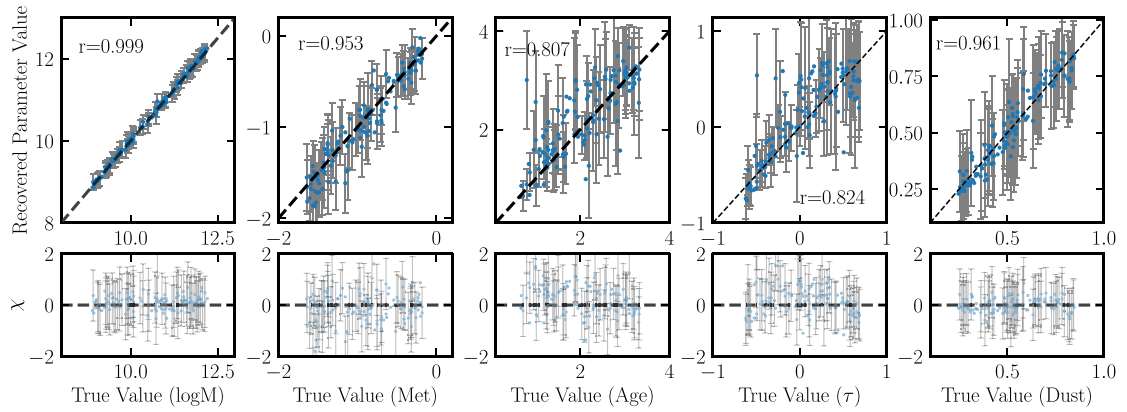


Figure 4. (Top) True vs. recovered values for SED parameters in the test set of 1000 spectra, sampled from the 16–84th percentile range of priors in this study. This demonstrates the accuracy of the predicted models across the entire range of priors. Note that only every alternate errorbar is plotted for visual clarity in the plot. (Bottom) Goodness-of-fit (χ) plots for each parameter, with errorbars corresponding to a value of 1.

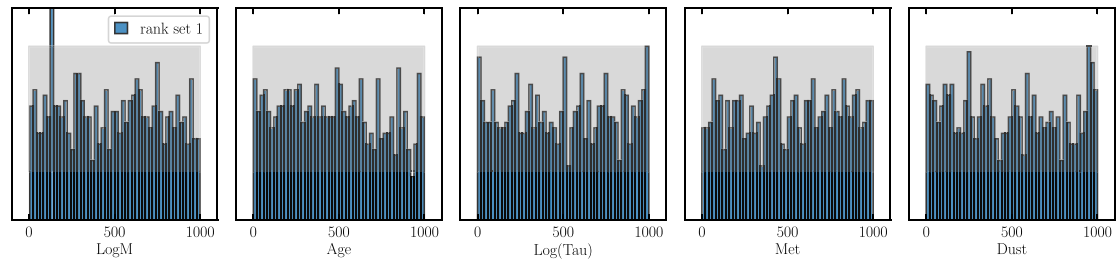


Figure 5. Simulation-Based Calibration rank plots for the SBI/NPE analysis. Each subplot corresponds to a parameter in the SED model. The grey region corresponds to the 95% confidence interval of a uniform distribution, which our parameter rank distributions follow.

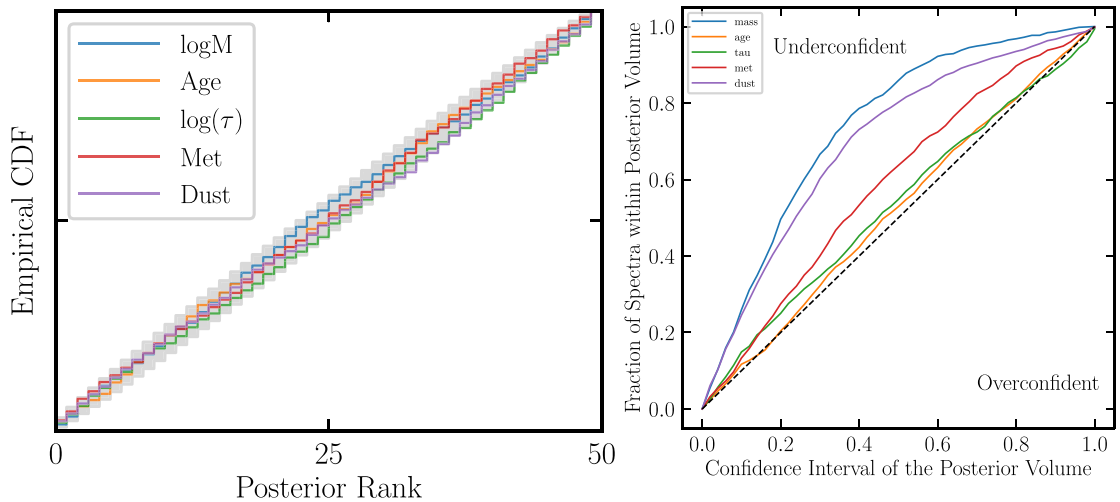


Figure 6. (Left) The cumulative density function (CDF) of posterior ranks for each parameter in our SBI/NPE analysis relative to the 95% confidence interval of a uniform distribution (grey). (Right) Probability coverage plot for each model SED parameter. A well-calibrated posterior estimator will produce curves that closely follow the dashed line (See section 4 for additional information).

trace the diagonal dashed line, while posterior estimates that are under-confident (i.e. over-estimate uncertainties) will produce curves that fall on the upper-left part of the plot.

For our five-parameter model, we see that the NPE has fairly well-calibrated uncertainty predictions for the age and τ parameters. On the other hand, it tends to over-predict the posterior uncertainties for three parameters—stellar mass, dust and metallicity. This is consistent with the results in figure 4—in the goodness-of-fits plots, the scatter in the differences between the predicted and true parameters values across

the test set is smaller than what their error bars would suggest. However, we are encouraged by the fact that the NPE analysis (a) accurately predicts the median best-fit values across the test set for those 3 parameters, and (b) in most scientific applications, over-predicting the uncertainties in an analysis is preferable than the alternative. We aim to continue to further improve the posterior uncertainty calibrations in future work.

We also demonstrate here a significant improvement in inference speeds compared with MCMC inference. As mentioned above, the SBI/NPE model uses 25 hidden units and 10 transform layers: this computation takes ~ 14 h to train on a CPU, and ~ 0.3 s per posterior estimation thereafter (MCMC calculation for a single galaxy takes ~ 24 h in our setup). Accounting for the cost of training, we can effectively infer accurate posteriors of $0.9\text{--}1.2 \times 10^5$ galaxy spectra via SBI/NPE at the cost of 1 MCMC-based fit. This demonstrates that the amortization of posterior estimation in SBI with accurate recovery of SED parameters is the biggest advantage of our proof-of-concept analysis.

4.1. Caveats

This work presents NPE analyses on spectra generated using (a) empirical templates within flexible stellar population synthesis (FSPS) (MILES spectral libraries, and MIST isochrones; Falc3n-Barroso *et al* 2011, Choi *et al* 2016), and (b) with an assumption of uniform (or nearly uniform) SNR and Gaussian noise across wavelength and for each galaxy spectrum. These assumptions have impact on both NPE and MCMC inference of galaxy parameters.

For example, the age-metallicity-dust degeneracy is a systematic that is limited by the information content of the templates, and the wavelength range sampled by the spectra that may or may not contain discriminating information. This affects both MCMC and NPE analyses, and is expected to be mitigated with upcoming spectroscopic surveys of statistical samples of galaxies (DESI, or the Prime Focus Spectrograph; Greene *et al* 2022) as well as simulation studies such as UniverseMachine Behroozi *et al* (2013) and FIRE Ma *et al* (2018).

In addition, SED inference using spectra is only weakly impacted by small ($< 10\%$ – 20%) variations in SNRs across wavelengths (especially in the stellar continuum portions of a given spectrum). Several studies analysing quiescent galaxy spectra Carnall *et al* (2019), Khullar *et al* (2022), Tacchella *et al* (2022) as well as photometric SED fitting and parameter recovery Gladis Magris *et al* (2015), Leja *et al* (2017a) seem to point towards this trend. Star-forming galaxies (with strong emission lines) have wavelength regions with peaked SNRs, that improves constraints on parameters such as instantaneous SFRs and ages. This biases our inferences in favor of young stellar populations with emission line spectra, albeit weakly. In figure 4, we observe that both older and younger stellar populations are recovered with similar precision. Finally, we expect spectroscopic surveys like DESI, PFS and the Roman High-Latitude Survey to improve this systematic uncertainty, as they attempt to reach nearly uniform SNRs across wavelength.

5. Conclusion and next steps

In this work, we demonstrate a proof-of-concept for amortized NPE with an SBI framework, that utilizes simulated low/medium-resolution galaxy spectra. This is the first-of-its-kind demonstration of this technique on spectra, that will enable precise and rapid estimation of galaxy parameter posteriors for billion-galaxy surveys. We also show here a significant improvement in inference speeds, while maintaining accuracy in the recovery of parameters, with precision comparable or moderately weaker than MCMC constraints.

While this work focuses on using an SBI framework to train on galaxy spectra directly (without any summary statistics or embedded nets), we wish to scale this analysis with graphics processing unit (GPU)-processing on suitable summary statistics (Khullar *et al* in prep). Moreover, the combination of highly complex SED models (Leja *et al* 2019a, Suess *et al* 2022b) and high-resolution spectroscopy will enable precise constraints on star formation histories of galaxies. This is especially true in the era of JWST (Labbe *et al* 2022, Leethochawalit *et al* 2022, Nanayakkara *et al* 2022, Suess *et al* 2022a) and Roman Space Telescope (High Latitude Survey; Wang *et al* 2022), where SED analysis of systematic spectroscopic surveys will be bolstered with an SBI framework.

Finally, when using simulation-trained SBI models on future survey data, it is important to consider possible performance issues that will arise from small differences between simulated and real data (due to approximations, unknown physics or computational constraints, imperfect simulation of noise and other observational effects). The drop in performance of simulation-trained models that are applied to real data is a known issue that affects all deep learning models. Mitigation of these problems is an active area of research, which already led to the development of a broad group of methods called *Domain Adaptation* (Csurka 2017, Wang and Deng 2018). These methods allow deep learning model to learn the features shared between simulated and real data and use only these features for inference. This leads to better alignment of the two

data distributions in the latent space of the deep learning model, which leads to improved performance (Ćiprijanović *et al* 2021, 2022). In future work, we will include domain adaptation methods in our SBI frameworks.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

The authors thank Alexander Ji, Egor Danilov, Michael D Gladders for their comments and feedback in the planning and analysis of this work. G K thanks the URA Visiting Scholars Program, 2021, for funding this work through graduate student salary support.

The authors are grateful to the reviewers of the journal MLST for their extremely thoughtful and helpful comments; their efforts and feedback have substantially improved the quality of the manuscript.

This manuscript has been supported by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy (DOE), Office of Science, Office of High Energy Physics.

The authors of this paper have committed themselves to performing this work in an equitable, inclusive, and just environment, and we hold ourselves accountable, believing that the best science is contingent on a good research environment.

We acknowledge the Deep Skies Lab as a community of multi-domain experts and collaborators who have facilitated an environment of open discussion, idea-generation, and collaboration. This community was important for the development of this project.

Author contributions

G Khullar: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Visualization, Writing of original draft; B Nord: Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, Writing (review & editing), Acquisition of the financial support for the project leading to this publication; A Ćiprijanović: Investigation, Methodology, Analysis, Project administration, Resources, Software, Supervision, Writing (review & editing); J Poh: Methodology, Analysis, Resources, Writing (review & editing); F Xu: Methodology, Resources.

Data and code availability statement

The code and dataset used to perform the experiments presented in this paper is openly available in our GitHub repository:


https://github.com/deepskies/digs_sbi

Access to the repository and data are available upon publication.

ORCID iDs

Gourav Khullar  <https://orcid.org/0000-0002-3475-7648>

Brian Nord  <https://orcid.org/0000-0001-6706-8972>

Aleksandra Ćiprijanović  <https://orcid.org/0000-0003-1281-7192>

Jason Poh  <https://orcid.org/0000-0002-5040-093X>

References

- Abbott T M C *et al* 2018 The dark energy survey: data release 1 *Astrophys. J. Suppl. Ser.* **239** 18
- Aghamousa A *et al* (DESI Collaboration) 2016 The DESI experiment part I: science, targeting, and survey design (arXiv:1611.00036)
- Ahumada R *et al* 2020 The 16th data release of the sloan digital sky surveys: first release from the APOGEE-2 southern survey and full release of eBOSS spectra *Astrophys. J. Suppl. Ser.* **249** 3
- Alsing J, Charnock T, Feeney S and Wandelt B 2019 Fast likelihood-free cosmology with neural density estimators and active learning *Mon. Not. R. Astron. Soc.* **488** 4440–58
- Behroozi P S, Wechsler R H and Conroy C 2013 The average star formation histories of galaxies in dark matter halos from $z = 0-8$ *Astrophys. J.* **770** 57
- Belli S, Newman A B and Ellis R S 2019 MOSFIRE spectroscopy of quiescent galaxies at $1.5 < z < 2.5$. II. Star formation histories and galaxy quenching *Astrophys. J.* **874** 17
- Carnall A C, Leja J, Johnson B D, McLure R J, Dunlop J S and Conroy C 2019 How to measure galaxy star formation histories. I. Parametric models *Astrophys. J.* **873** 44

- Choi J, Dotter A, Conroy C, Cantiello M, Paxton B and Johnson B D 2016 Mesa isochrones and stellar tracks (MIST). I. Solar-scaled models *Astrophys. J.* **823** 102
- Ćiprijanović A, Kafkes D, Downey K, Jenkins S, Perdue G N, Madireddy S, Johnston T, Snyder G F and Nord B 2021 DeepMerge—II. Building robust deep learning algorithms for merging galaxy identification across domains *Mon. Not. R. Astron. Soc.* **506** 677–91
- Ćiprijanović A, Kafkes D, Snyder G, Sánchez F J, Perdue G N, Pedro K, Nord B, Madireddy S and Wild S M 2022 DeepAdversaries: examining the robustness of deep learning models for galaxy morphology classification *Mach. Learn.: Sci. Technol.* **3** 035007
- Conroy C and Gunn J E 2010 The propagation of uncertainties in stellar population synthesis modeling. III. Model calibration, comparison and evaluation *Astrophys. J.* **712** 833–57
- Cranmer K, Brehmer J and Louppe G 2019 The frontier of simulation-based inference (arXiv:1911.01429)
- Csurka G 2017 A comprehensive survey on domain adaptation for visual applications *Domain Adaptation in Computer Vision Applications* (Cham: Springer) pp 1–35
- Dey A *et al* 2019 Overview of the DESI legacy imaging surveys *Astron. J.* **157** 168
- Falcón-Barroso J, Sánchez-Blázquez P, Vazdekis A, Ricciardelli E, Cardiel N, Cenarro A J, Gorgas J and Peletier R F 2011 An updated MILES stellar library and stellar population models *Astron. Astrophys.* **532** A95
- Ferreras I, Charlot S and Silk J 1999 The age and metallicity range of early-type galaxies in clusters *Astrophys. J.* **521** 81–89
- Foreman-Mackey D, Hogg D W, Lang D and Goodman J 2013 emcee: the MCMC hammer *Publ. Astron. Soc. Pac.* **125** 306
- Ganaie M A, Hu M, Malik A K, Tanveer M and Suganthan P N 2021 Ensemble deep learning: a review (arXiv:2104.02395)
- Gladis Magris C, Juan Mateu P, Mateu C, Juan Bruzual A, Cabrera-Ziri I and Mejía-Narváez A 2015 On the recovery of galaxy properties from SED fitting solutions *Publ. Astron. Soc. Pac.* **127** 16–30
- Greenberg D S, Nonnenmacher M and Macke J H 2019 Automatic posterior transformation for likelihood-free inference (arXiv:1905.07488)
- Greene J, Bezanson R, Ouchi M and Silverman J (The PFS Galaxy Evolution Working Group) 2022 The prime focus spectrograph galaxy evolution survey (arXiv:2206.14908)
- Hahn C and Melchior P 2022 Accelerated Bayesian SED modeling using amortized neural posterior estimation (arXiv:2203.07391)
- Hinshaw G *et al* 2013 Nine-year Wilkinson microwave anisotropy probe (WMAP) observations: cosmological parameter results *Astrophys. J. Suppl. Ser.* **208** 19
- Huppenkothen D and Bachetti M 2022 Accurate x-ray timing in the presence of systematic biases with simulation-based inference *Mon. Not. R. Astron. Soc.* **511** 5689–708
- Ivezić Ž *et al* 2019 LSST: from science drivers to reference design and anticipated data products *Astrophys. J.* **873** 111
- Johnson B D, Leja J, Conroy C and Speagle J S 2021 Stellar population inference with prospector *Astrophys. J. Suppl. Ser.* **254** 22
- Kacprzak T, Herbel J, Amara A and Réfrégier A 2018 Accelerating approximate Bayesian computation with quantile regression: application to cosmological redshift distributions *J. Cosmol. Astropart. Phys.* **2018** 042
- Khullar G *et al* 2021 Characterizing stellar mass assembly and physical properties in the brightest galaxy in the redshift > 5 Universe JWST Proposal. Cycle 1, ID.#2566 (available at: <https://ui.adsabs.harvard.edu/abs/2021jwst.prop.2566K>)
- Khullar G *et al* 2022 Synthesizing stellar populations in South Pole Telescope galaxy clusters. I. Ages of quiescent member galaxies at $0.3 < z < 1.4$ *Astrophys. J.* **934** 177
- Kroupa P 2001 On the variation of the initial mass function *Mon. Not. R. Astron. Soc.* **322** 231–46
- Labbe I, van Dokkum P, Nelson E, Bezanson R, Suess K, Leja J, Brammer G, Whitaker K, Mathews E and Stefanon M 2022 A very early onset of massive galaxy formation (arXiv:2207.12446)
- Leethochawalit N *et al* 2022 Early results from GLASS-JWST. X: rest-frame UV-optical properties of galaxies at $7 < z < 9$ (arXiv:2207.11135)
- Leja J *et al* 2019b An older, more quiescent universe from panchromatic SED fitting of the 3D-HST survey *Astrophys. J.* **877** 140
- Leja J, Carnall A C, Johnson B D, Conroy C and Speagle J S 2019a How to measure galaxy star formation histories. II. Nonparametric models *Astrophys. J.* **876** 3
- Leja J, Johnson B D, Conroy C, van Dokkum P G van and Byler N 2017a Deriving physical properties from broadband photometry with prospector: description of the model and a demonstration of its accuracy using 129 galaxies in the local universe *Astrophys. J.* **837** 170
- Leung H W and Bovy J 2019 Deep learning of multi-element abundances from high-resolution spectroscopic data *Mon. Not. R. Astron. Soc.* **483** 3255–77
- Lovell C C, Acquaviva V, Thomas P A, Iyer K G, Gawiser E and Wilkins S M 2019 Learning the relationship between galaxies spectra and their star formation histories using convolutional neural networks and cosmological simulations *Mon. Not. R. Astron. Soc.* **490** 5503–20
- Ma X, Hopkins P F, Garrison-Kimmel S, Faucher-Giguère C-A, Quataert E, Boylan-Kolchin M, Hayward C C, Feldmann R and Kereš D 2018 Simulating galaxies in the reionization era with FIRE-2: galaxy scaling relations, stellar mass functions and luminosity functions *Mon. Not. R. Astron. Soc.* **478** 1694–715
- Nanayakkara T *et al* 2022 Early results from GLASS-JWST XVI: discovering a bluer $z \sim 4$ -7 universe through UV slopes (arXiv:2207.13860)
- Papamakarios G and Murray I 2016 Fast ϵ -free inference of simulation models with bayesian conditional density estimation *Advances in Neural Information Processing Systems* ed D Lee, M Sugiyama, U Luxburg, I Guyon and R Garnett (Curran Associates, Inc.) (available at: <https://proceedings.neurips.cc/paper/2016/file/6aca97005c68f1206823815f66102863-Paper.pdf>)
- Papamakarios G, Pavlakou T and Murray I 2017 Masked autoregressive flow for density estimation (arXiv:1705.07057)
- Suess K A, Bezanson R, Nelson E J, Setton D J, Price S H, van Dokkum P, Brammer G, Labbe I, Leja J, Miller T B, Robertson B, Weaver J R and Whitaker K E 2022a Rest-frame near-infrared sizes of galaxies at cosmic noon: objects in JWST’s mirror are smaller than they appeared (arXiv:2207.10655)
- Suess K A, Leja J, Johnson B D, Bezanson R, Greene J E, Kriek M, Lower S, Narayanan D, Setton D J and Spilker J S 2022b Recovering the star formation histories of recently-quenched galaxies: the impact of model and prior choices (arXiv:2207.02883)
- Tabak E G and Turner C V 2013 A family of nonparametric density estimation algorithms *Commun. Pure Appl. Math.* **66** 145–64
- Tacchella S *et al* 2022 Fast, slow, early, late: quenching massive galaxies at $z \sim 0.8$ *Astrophys. J.* **926** 134
- Talts S, Betancourt M, Simpson D, Vehtari A and Gelman A 2018 Validating Bayesian inference algorithms with simulation-based calibration (arXiv:1804.06788)
- Tejero-Cantero A, Boelts J, Deistler M, Lueckmann J-M, Durkan C, Gonçalves P J, Greenberg D S and Macke J H 2020 SBI: a toolkit for simulation-based inference *J. Open Source Softw.* **5** 2505
- Uria B, Côté M A, Gregor K, Murray I and Larochelle H 2016 Neural autoregressive distribution estimation (arXiv:1605.02226)

- Valentin Jospin L, Buntine W, Boussaid F, Laga H and Bennamoun M 2020 Hands-on Bayesian neural networks—a tutorial for deep learning users (arXiv:2007.06823)
- Wang M and Deng W 2018 Deep visual domain adaptation: a survey *Neurocomputing* **312** 135–53
- Wang Y *et al* 2022 The high latitude spectroscopic survey on the Nancy Grace Roman Space Telescope *Astrophys. J.* **928** 1
- Worthey G 1994 Comprehensive stellar population models and the disentanglement of age and metallicity effects *Astrophys. J. Suppl. Ser.* **95** 107
- Zackrisson E *et al* 2017 The spectral evolution of the first galaxies. III. Simulated James Webb Space Telescope spectra of reionization-epoch galaxies with Lyman-continuum leakage *Astrophys. J.* **836** 78
- Zhang K, Bloom J S, Gaudi B S, Lanusse F, Lam C and Lu J R 2021 Real-time likelihood-free inference of roman binary microlensing events with amortized neural posterior estimation *Astron. J.* **161** 262
- Zhao X, Mao Y, Cheng C and Wandelt B D 2022 Simulation-based inference of reionization parameters from 3D tomographic 21 cm light-cone images *Astrophys. J.* **926** 151