DIGS: Deep Inference of Galaxy Spectra with Neural Posterior Estimation

Gourav Khullar
Department of Physics and Astronomy;
PITT PACC,
University of Pittsburgh
gourav.khullar@pitt.edu

Brian Nord
Fermi National Accelerator Laboratory;
Kavli Institute for Cosmological Physics &
Department of Astronomy and Astrophysics,
The University of Chicago;
Laboratory for Nuclear Physics, MIT
nord@fnal.gov

Aleksandra Ćiprijanović
Fermi National Accelerator Laboratory
aleksand@fnal.gov

Jason Poh
Department of Astronomy and Astrophysics,
The University of Chicago
jasonpoh@uchicago.edu

Fei Xu
Department of Astronomy and Astrophysics,
The University of Chicago
feixu@uchicago.edu

Ashwin Samudre
School of Computing Science,
Simon Fraser University
ashwin_samudre@sfu.ca

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.

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 show that SBI is capable of inferring very accurate galaxy stellar masses and metallicities. Our methodology also a) produces uncertainty constraints that are comparable to or moderately weaker than traditional inverse-modeling with Bayesian MCMC methods (e.g., 0.17 and 0.26 dex in stellar mass and metallicity for a given galaxy, respectively), and b) conducts rapid SED inference ($\sim 10^5$ galaxy spectra via SBI/SNPE at the cost of 1 MCMC-based fit); this efficiency is needed in the era of JWST and Roman Telescopes.

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, and the star formation history. These parameters can be accurately inferred from a galaxy's spectral energy distribution (SED).

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

Within the last two decades, photometry-based SED fitting has become a pivotal method to measure the aforementioned properties. Ground-based telescopes have been used extensively for large multi-wavelength galaxy surveys – e.g., Sloan Digital Sky Survey (SDSS, [2]), Dark Energy Survey (DES, [1]), and DESI Legacy Imaging Surveys [9] – producing complex high-quality datasets. However, SED studies relying on photometry alone are subject to challenges, such as the age-metallicity-dust degeneracy [28, 11]. SED fitting using spectra mitigate this challenge significantly with measurement of absorption line indices and emission line strengths (e.g., [28, 22] 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 [19, 5, 21, 16]. 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. A 5-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 hours. Moreover, each galaxy spectra requires its own separate inference chains. With the next generation of telescopes, tens of millions of optical and infrared galaxy photometry and spectra will be measured. The need of the hour is to quickly and reliably deduce the physical parameters of galaxies in large surveys as well as from a large number of pixels/spaxels.

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 [24, 23, 14]. These methods allow regression of galaxy parameters, albeit without any uncertainty quantification.

Simulation-based inference [SBI; 8] combined with deep learning can mitigate assumptions (e.g., tractable likelihood) that can plague analytic likelihood/posterior modeling, as well as remove computational bottlenecks. Many astrophysical studies have demonstrated success with SBI in calculating posteriors rapidly and accurately [17, 3, 29, 30, 15]. Recent work has shown that photometric SED data can be used with SBI for fast inference [14]. In this work, we demonstrate a proof-of-concept SBI framework to analyse galaxy spectra and recover posteriors efficiently.

2 Data

We use Prospector [16] to generate simulated SEDs of galaxies. Prospector relies on Markov Chain Monte Carlo (MCMC) sampling for stellar population synthesis (SPS) and parameter inference. It is based on the Python-FSPS framework, with the MILES stellar spectral library and the MIST set of isochrones [7, 20, 12, 10, 6]. We generate a training set of 10000 rest-frame SEDs using a 5-parameter model, with a delayed, exponentially declining (i.e., delayed-tau) star formation history. The SFH – star formation rate vs. time – is:

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

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 [4]). We sample the total stellar mass (M_*), the stellar metallicity $log(Z/Z_{sol})$, a delayed-tau 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 5500Å). Each parameter vector $\boldsymbol{\theta}$ comprises these five parameters. Our SED model assumes a Kroupa IMF [18].

We smooth and resample the simulated SEDs to resemble a medium-resolution spectroscopic survey using Prospector's internal resampling utility. This results in a training set with each galaxy SED sampling rest-frame 3750-9500Å, with 138 flux elements for each SED, which is the data vector $\boldsymbol{x}$ corresponding to each $\boldsymbol{\theta}$. 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 SNR$\sim$20. We conduct data augmentation to scale 10000 noiseless 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 set of 1000 test spectra to conduct posterior diagnostics.

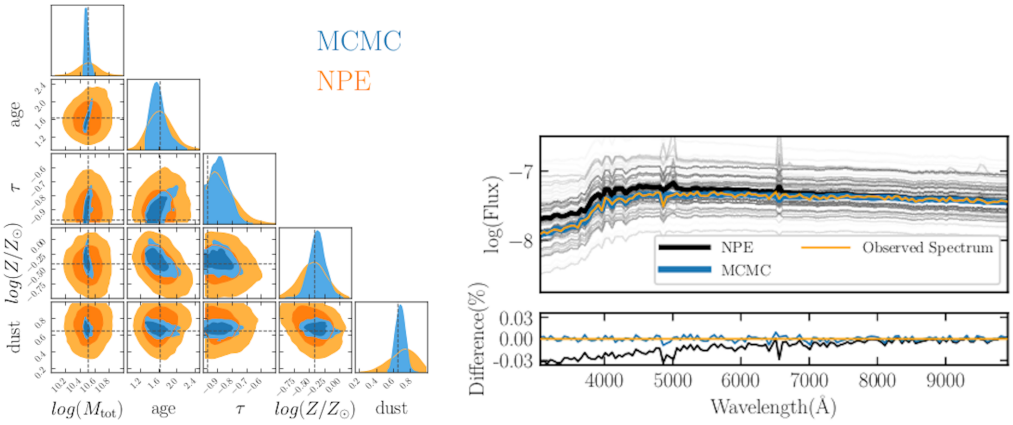

Figure 1: (Left) Corner plot for an example galaxy in our test set for a 5-parameter SED model. This plot shows pairwise posteriors inferred from the MCMC (blue) and SBI neural posterior estimation (orange) analyses. We find that SBI and MCMC recover SED parameters $\boldsymbol{\theta}$ accurately, while SBI constraints for each parameter (for marginalized posteriors) is similar or moderately weaker than MCMC constraints. (Right) Best-fit SEDs from SBI/NPE (black) and MCMC (blue) analyses, and percentage residuals. Note that the difference in observed spectrum (orange) and NPE best-fit spectrum (black) is $< 3\%$, smaller than the Gaussian noise applied to each simulated spectrum.

3 Inference Methodology : SBI

Our objective is to calculate posteriors $p(\boldsymbol{\theta}|\boldsymbol{x})$ of the galaxy parameters derived from a typical SED analysis, where $\boldsymbol{\theta}$ is the set of galaxy properties, and $\boldsymbol{x}$ 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) [25, 13]) 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 observations without having to re-run inference. This "amortized" calculation of posteriors then allows us to infer the posteriors of a "real" galaxy with computational time $<$ 1s. For more details and examples of amortized neural network-based posterior estimation, see [13] or Section 2 of [14].

This work: We use a supervised learning pipeline within an SBI framework via the Macke Lab **sbi** toolbox [27]. 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 $\sim$ 14 hours to train. This analysis generates the set of approximate posterior distributions for our 5 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.

We evaluate the results using a variety of diagnostic tests. First, we compare the recovered SED parameter values with the true values $\boldsymbol{\theta}$ 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 [26]. For a healthy posterior, the SBC ranks of ground truth parameters under the inferred posteriors should follow a uniform distribution. Finally, we compare our results to an MCMC analysis of representative SEDs from the test set.

We fit the same 5-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 use the same $\boldsymbol{\theta}$ prior range and shape as the SBI/NPE analysis, in order to calculate the posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$. We use **emcee** [12] 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.

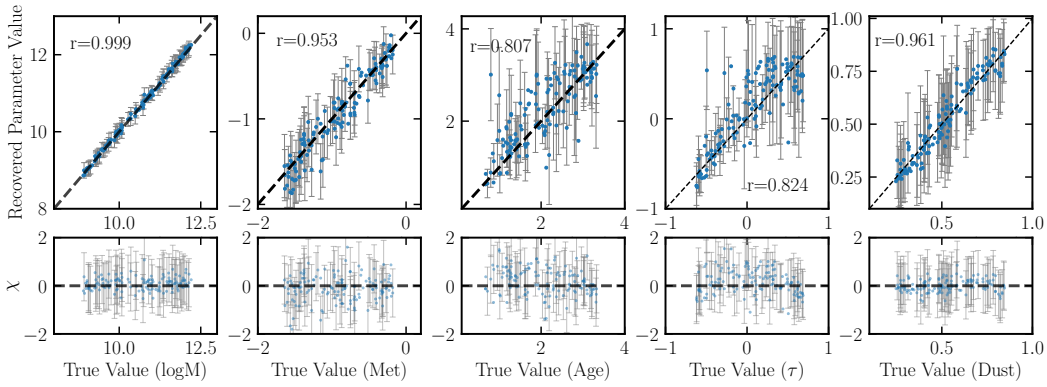

Figure 2: (Top) True vs. recovered values for SED parameters in the test set of 1000 spectra, sampled from the 16-84th percentile range of priors in this study. This demonstrates the accuracy of the predicted models across the entire range of priors. (Bottom) Goodness-of-fit (χ) plots for each parameter, with errorbars corresponding to a value of 1. Please note that the uncertainties plotted here are symmetric (even though the confidence intervals derived from posteriors are not always Gaussian; this plot is used as a visual indicator of the average confidence interval obtained in inference for our test sample for each parameter.

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 $\sim$14 CPU hours to train 2×10^6 simulated spectra with noise. For every subsequent posterior estimation, this setup takes $\sim$0.3s. For our serialized MCMC calculations, we utilise **Prospector** runtime on a 2.7 GHz Quad-Core Intel Core i7 processor, which takes $\sim$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 SEDs randomly sampled from a physically motivated prior(s) range. One such result is shown in the left panel of Fig. 1, for a galaxy with $\log M_{tot}$= 10.51 ($M_\odot$), $log(Z/Z_\odot)$ = -0.41, age = 1.63 Gyr, τ = 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 neural posterior estimation; NPE, in orange) in order to compare constraints across the 5-parameter SED model. In the right panel of Fig. 1, we show the maximum a posteriori (MAP) SED models and model residuals from the SBI/NPE (black) and MCMC (blue) analyses.

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 $\boldsymbol{\theta}$. 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 Fig. 2 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. 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 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 (following

[26]); see Appendix for more details. Moreover, 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. The SBI/NPE model uses 25 hidden units and 10 transform layers: this computation takes $\sim$14 hours to train on a CPU, and $\sim$0.3s per posterior estimation thereafter (MCMC calculation for a single galaxy takes $\sim$24 hours in our setup). Accounting for the cost of training, we can infer accurate posteriors of $\sim 10^5$ galaxy spectra via SBI/SNPE 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 analysis.

5 Conclusion

We demonstrate a proof-of-concept for amortized neural posterior estimation with an SBI framework, that utilizes simulated low-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. In future work, we aim to solve limitations presented here, including a more complex and robust model (e.g., non-parametric star formation histories), better modeling of the age-metallicity degeneracy, higher resolution spectra, more complex and realistic noise models, and training the NPE on summary statistics of the simulated dataset.

6 Broader Impacts Statement

This work contains analyses and methodology that is relevant to all STEM fields – astrophysics, physics, biology, etc. – that rely on experiments that require the assessment of physically motivated uncertainties in an efficient manner. Moreover, the impact of this work on any potential application to the real-world needs to be taken into real consideration. We acknowledge that since real-world data is noisy, any framework claiming to build generative models based on data with complex noise models is attractive for real world applications. We encourage our readers to study the many negative impacts of generative models along the principles and values of ethics and social justice.

Acknowledgements and Disclosure of Funding

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

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, Acquisition of financial support for this publication, Supervision, Writing (review & editing)*; 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*. A. Samudre: *Resources, Writing (review & editing)*.

References

[1] T. M. C. Abbott, F. B. Abdalla, S. Allam, A. Amara, J. Annis, J. Asorey, S. Avila, O. Ballester, M. Banerji, W. Barkhouse, L. Baruah, M. Baumer, K. Bechtol, and et al. The Dark Energy Survey: Data Release 1. , 239(2):18, December 2018.

[2] Romina Ahumada, Carlos Allende Prieto, Andrés Almeida, Friedrich Anders, Scott F. Anderson, Brett H. Andrews, Borja Anguiano, Riccardo Arcodia, Eric Armengaud, Marie Aubert, Santiago Avila, Vladimir Avila-Reese, Carles Badenes, and et al. The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. , 249(1):3, July 2020.

[3] Justin Alsing, Tom Charnock, Stephen Feeney, and Benjamin Wandelt. Fast likelihood-free cosmology with neural density estimators and active learning. , 488(3):4440–4458, September 2019.

[4] Sirio Belli, Andrew B. Newman, and Richard S. Ellis. MOSFIRE Spectroscopy of Quiescent Galaxies at $1.5 < z < 2.5$. II. Star Formation Histories and Galaxy Quenching. , 874(1):17, March 2019.

[5] Adam C. Carnall, Joel Leja, Benjamin D. Johnson, Ross J. McLure, James S. Dunlop, and Charlie Conroy. How to Measure Galaxy Star Formation Histories. I. Parametric Models. , 873(1):44, March 2019.

[6] Jieun Choi, Aaron Dotter, Charlie Conroy, Matteo Cantiello, Bill Paxton, and Benjamin D. Johnson. Mesa Isochrones and Stellar Tracks (MIST). I. Solar-scaled Models. , 823(2):102, June 2016.

[7] C. Conroy and J. E. Gunn. The Propagation of Uncertainties in Stellar Population Synthesis Modeling. III. Model Calibration, Comparison, and Evaluation. , 712:833–857, April 2010.

[8] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *arXiv e-prints*, page arXiv:1911.01429, November 2019.

[9] Arjun Dey, David J. Schlegel, Dustin Lang, Robert Blum, Kaylan Burleigh, Xiaohui Fan, Joseph R. Findlay, Doug Finkbeiner, David Herrera, Stéphanie Juneau, Martin Landriau, Michael Levi, Ian McGreer, Aaron Meisner, Adam D. Myers, John Moustakas, Peter Nugent, Anna Patej, Edward F. Schlafly, Alistair R. Walker, and et al. Overview of the DESI Legacy Imaging Surveys. , 157(5):168, May 2019.

[10] J. Falcón-Barroso, P. Sánchez-Blázquez, A. Vazdekis, E. Ricciardelli, N. Cardiel, A. J. Cenarro, J. Gorgas, and R. F. Peletier. An updated MILES stellar library and stellar population models. , 532:A95, August 2011.

[11] Ignacio Ferreras, Stéphane Charlot, and Joseph Silk. The Age and Metallicity Range of Early-Type Galaxies in Clusters. , 521(1):81–89, August 1999.

[12] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman. emcee: The MCMC Hammer. , 125:306, March 2013.

[13] David S. Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic Posterior Transformation for Likelihood-Free Inference. *arXiv e-prints*, page arXiv:1905.07488, May 2019.

[14] ChangHoon Hahn and Peter Melchior. Accelerated Bayesian SED Modeling using Amortized Neural Posterior Estimation. *arXiv e-prints*, page arXiv:2203.07391, March 2022.

[15] Daniela Huppenkothen and Matteo Bachetti. Accurate X-ray timing in the presence of systematic biases with simulation-based inference. , 511(4):5689–5708, April 2022.

[16] Benjamin D. Johnson, Joel Leja, Charlie Conroy, and Joshua S. Speagle. Stellar Population Inference with Prospector. , 254(2):22, June 2021.

[17] T. Kacprzak, J. Herbel, A. Amara, and A. Ré frégier. Accelerating approximate bayesian computation with quantile regression: application to cosmological redshift distributions. *Journal of Cosmology and Astroparticle Physics*, 2018(02):042–042, feb 2018.

[18] Pavel Kroupa. On the variation of the initial mass function. , 322(2):231–246, April 2001.

[19] J. Leja, B. D. Johnson, C. Conroy, P. G. van Dokkum, and N. Byler. 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. , 837:170, 2017.

[20] J. Leja, B. D. Johnson, C. Conroy, P. G. van Dokkum, and N. Byler. 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. , 837:170, 2017.

[21] Joel Leja, Adam C. Carnall, Benjamin D. Johnson, Charlie Conroy, and Joshua S. Speagle. How to Measure Galaxy Star Formation Histories. II. Nonparametric Models. , 876(1):3, May 2019.

[22] Joel Leja, Benjamin D. Johnson, Charlie Conroy, Pieter van Dokkum, Joshua S. Speagle, Gabriel Brammer, Ivelina Momcheva, Rosalind Skelton, Katherine E. Whitaker, Marijn Franx, and Erica J. Nelson. An Older, More Quiescent Universe from Panchromatic SED Fitting of the 3D-HST Survey. , 877(2):140, June 2019.

[23] Henry W. Leung and Jo Bovy. Deep learning of multi-element abundances from high-resolution spectroscopic data. , 483(3):3255–3277, March 2019.

[24] Christopher C. Lovell, Viviana Acquaviva, Peter A. Thomas, Kartheik G. Iyer, Eric Gawiser, and Stephen M. Wilkins. Learning the relationship between galaxies spectra and their star formation histories using convolutional neural networks and cosmological simulations. , 490(4):5503–5520, December 2019.

[25] George Papamakarios and Iain Murray. Fast ϵ -free inference of simulation models with bayesian conditional density estimation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[26] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration, 2018.

[27] Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J. Gonçalves, David S. Greenberg, and Jakob H. Macke. sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020.

[28] Guy Worthey. Comprehensive stellar population models and the disentanglement of age and metallicity effects. , 95:107, November 1994.

[29] Keming Zhang, Joshua S. Bloom, B. Scott Gaudi, Franç ois Lanusse, Casey Lam, and Jessica R. Lu. Real-time likelihood-free inference of roman binary microlensing events with amortized neural posterior estimation. *The Astronomical Journal*, 161(6):262, may 2021.

[30] Xiaosheng Zhao, Yi Mao, Cheng Cheng, and Benjamin D. Wandelt. Simulation-based Inference of Reionization Parameters from 3D Tomographic 21 cm Light-cone Images. *Astrophys. J.*, 926(2):151, 2022.

Table 1: Simulated SEDs: Model Description and Prior Range

Parameter	**Description**	**Priors**
$M_{\rm total}(M_\odot)$	Total stellar mass formed	$\rm Log_{10}$ 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)	$\rm Log_{10}$ Uniform: [0.1, 1.0]

A Appendix

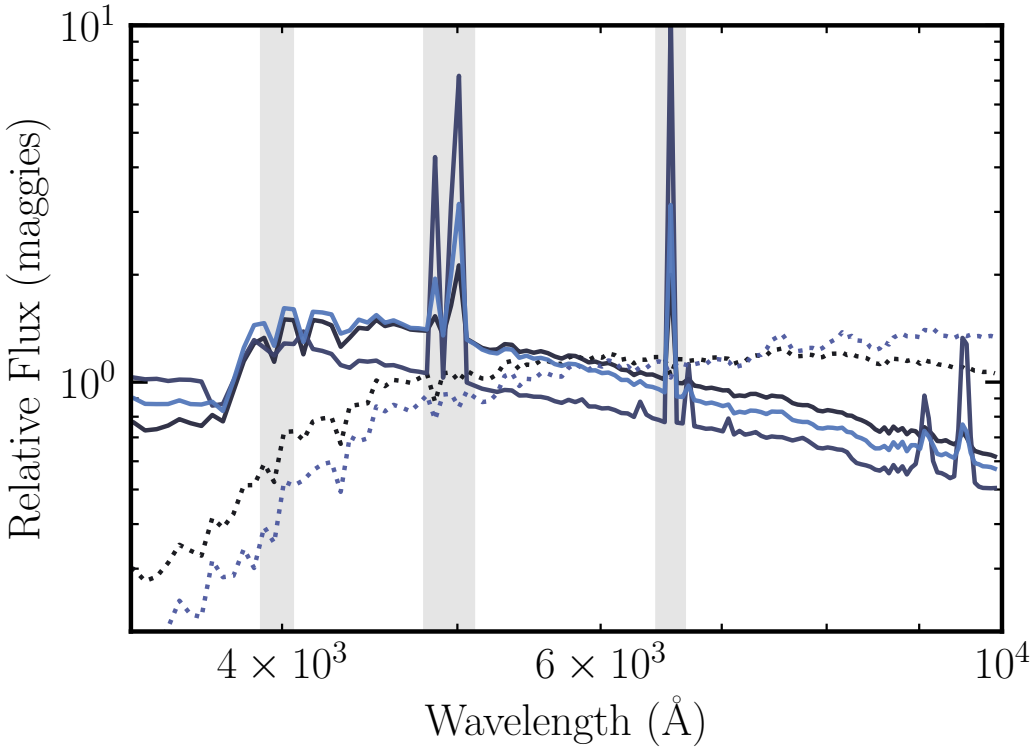

Figure 3: Five Galaxy SEDs randomly sampled from the training set used in this work, with 5 different values of the parameter vector θ. The SEDs are normalized to their median: this respects the fact that the shape of the SED is primarily dictated by the stellar metallicity, age and dust properties, 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 Hβ, [OIII]5007Å and Hα), while dotted lines represent SEDs of galaxies with absorption features and continuum breaks, such as 4000Å .

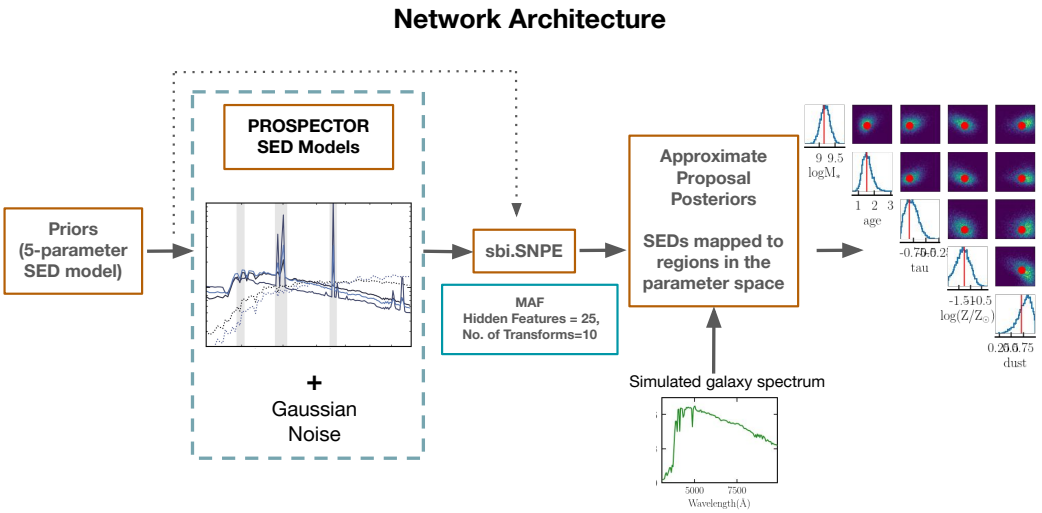

Figure 4: The architecture used in this work to infer galaxy SED properties with spectroscopic data. We use a 5-parameter model and a training set with realistic spectra, that is trained by an NPE to generate approximate posteriors.

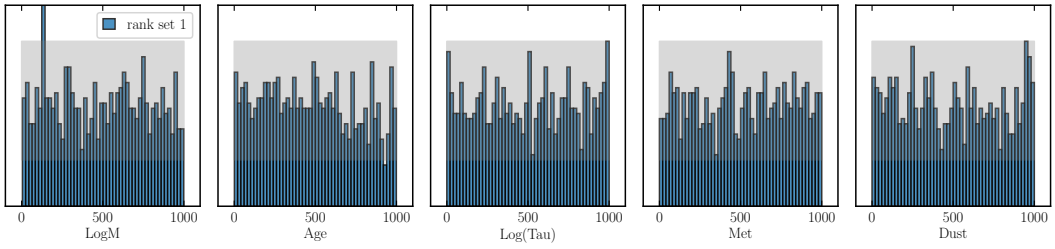

Figure 5: Simulation-Based Calibration rank plots for the SBI/NPE analysis. Each subplot corresponds to a parameter in the SED model. The grey region corresponds to the 95% confidence interval of a uniform distribution, which our parameter rank distributions follow.

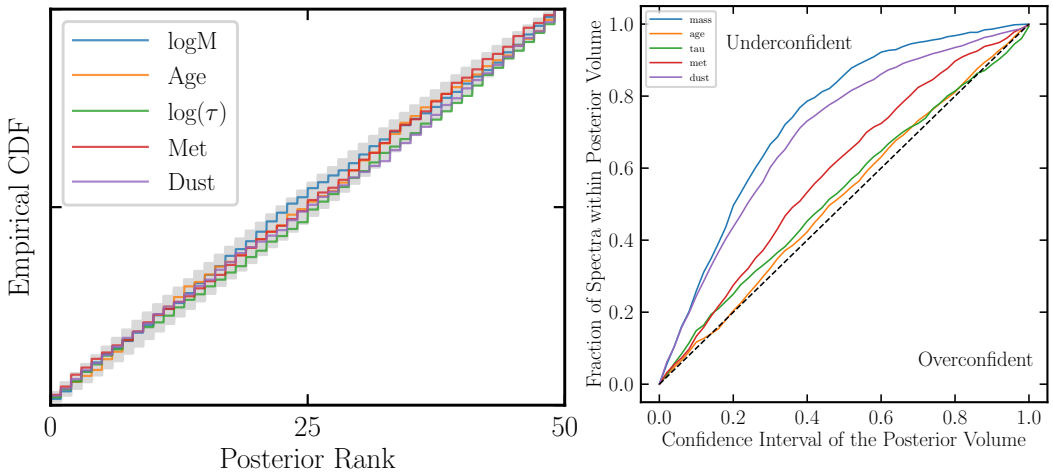

Figure 6: (Left) The cumulative density function (CDF) of posterior ranks for each parameter in our SBI/NPE analysis relative to the 95% confidence interval of a uniform distribution (grey). (Right) Probability coverage plot for each model SED parameter. A well-calibrated posterior estimator will produce curves that closely follow the dashed line (See Section 4 for additional information). For our 5-parameter model, we see that the NPE has fairly well-calibrated uncertainty predictions for the age and τ parameters; it tends to over-predict the posterior uncertainties for three parameters — stellar mass, dust and metallicity. This is consistent with Fig. 2 — 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.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] **See Section 4 and 5.**
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] **We will provide the code, data and instructions in a github repository post-review in the final version.**
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] **See Section 2 and 3 for details.**
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Due to space constraints in the paper, we do not discuss this explicitly, even though our experiments include a comparison between analysis at different random seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] **See Section 3 for details.**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] **Existing open-source codes and packages used have been properly cited in all sections; please see text especially in Section 2 and 3.**
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] **All data used here has been generated using open-source Python code, that that the authors have curated themselves.**
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]