Draft version March 6, 2022
Typeset using LaTeX **preprint** style in AASTeX63

Accelerated Bayesian SED Modeling using Amortized Neural Posterior Estimation

ChangHoon Hahn[1,*] and Peter Melchior[1,2]

[1]*Department of Astrophysical Sciences, Princeton University, Princeton NJ 08544, USA*
[2]*Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08544, USA*

ABSTRACT

State-of-the-art spectral energy distribution (SED) analyses use a Bayesian framework to infer the physical properties of galaxies from observed photometry or spectra. They require sampling from a high-dimensional space of SED model parameters and take $> 10-100$ CPU hours per galaxy, which renders them practically infeasible for analyzing the *billions* of galaxies that will be observed by upcoming galaxy surveys (*e.g.* DESI, PFS, Rubin, Webb, and Roman). In this work, we present an alternative scalable approach for rigorous Bayesian inference using Amortized Neural Posterior Estimation (ANPE). ANPE is a simulation-based inference method that employs neural networks to estimate the posterior probability distribution over the full range of observations. Once trained, it requires no additional model evaluations to estimate the posterior. We present, and publicly release, SEDflow, an ANPE method to produce posteriors of the recent Hahn et al. (2022) SED model from optical photometry. SEDflow takes $\sim$1 *second per galaxy* to obtain the posterior distributions of 12 model parameters, all of which are in excellent agreement with traditional Markov Chain Monte Carlo sampling results. We also apply SEDflow to 33,884 galaxies in the NASA-Sloan Atlas and publicly release their posteriors.

Keywords: galaxies: evolution – galaxies: statistics

1. INTRODUCTION

Physical properties of galaxies are the building blocks of our understanding of galaxies and their evolution. We can determine properties such as stellar mass (M_*), star formation rate (SFR), metallicity (Z), and age ($t_{\rm age}$) of a galaxy by analyzing its spectral energy distribution (SED). Theoretical modeling of galaxy SEDs is currently based on stellar population synthesis (SPS) and describes the SED as a composite stellar population constructed from isochrones, stellar spectra, an initial mass function (IMF), a star formation and chemical evolution history, and dust attenuation (*e.g.* Bruzual & Charlot 2003; Maraston 2005; Conroy et al. 2009, see Walcher et al. 2011; Conroy 2013 for a comprehensive review). Some models also include dust and nebular emissions as well as emissions from active galactic nuclei (*e.g.* Johnson et al. 2021). In state-of-the-art SED modeling, theoretical

* changhoon.hahn@princeton.edu.com

SPS models are compared to observed SEDs using Bayesian inference, which accurately quantifies parameter uncertainties and degeneracies among them (Acquaviva et al. 2011; Chevallard & Charlot 2016; Leja et al. 2017; Carnall et al. 2018; Johnson et al. 2021; Hahn et al. 2022). The Bayesian approach also enables marginalization over nuisance parameters, which are necessary to model the effects of observational systematics (*e.g.* flux calibration).

However, current Bayesian SED modeling methods, which use Markov Chain Monte Carlo (MCMC) sampling techniques, take $10-100$ CPU hours per galaxy (*e.g.* Carnall et al. 2019; Tacchella et al. 2021). While this is merely very expensive with current data sets of hundreds of thousands of galaxy SEDs, observed by the Sloan Digital Sky Survey (SDSS; York et al. 2000), DEEP2 (Davis et al. 2003), zCOSMOS (Scoville et al. 2007; Lilly et al. 2007), and GAMA (Baldry et al. 2018), it is prohibitive for the next generation of surveys. Over the next decade, surveys with the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2016), the Prime Focus Spectrograph (PFS; Takada et al. 2014), the Vera C. Rubin Observatory (Ivezic et al. 2019), the James Webb Space Telescope (Gardner et al. 2006), and the Roman Space Telescope (Spergel et al. 2015), will observe *billions* of galaxy SEDs. The task of SED modeling alone for these surveys would amount to tens or hundreds of billions of CPU hours, exceeding *e.g.* the compute allocation of the Legacy Survey of Space and Time (LSST) data release production over its entire lifetime[1] by at least two orders of magnitude. Recently, Alsing et al. (2020) adopted neural emulators to accelerate SED model evaluations by three to four orders of magnitude — posterior inference takes minutes per galaxy. While this renders current data sets within reach, the next generation data sets will still require tens or hundreds of millions of CPU hours whenever any aspect of the SED model is altered. Furthermore, this still practically precludes rapid analyses of upcoming transient surveys, especially LSST, which will report $\sim$10,000 alerts per minute[2].

But Bayesian inference does not require MCMC sampling. Simulation-based inference (SBI) is a rapidly developing class of inference methods that offers alternatives for many applications (see Cranmer et al. 2020, and references therein). Many SBI methods leverage the latest developments in statistics and Machine Learning for more efficient posterior estimation (Papamakarios et al. 2017; Alsing et al. 2019; Hahn et al. 2019; Dax et al. 2021; Huppenkothen & Bachetti 2021; Zhang et al. 2021). Of particular interest for SED modeling is a technique called Amortized Neural Posterior Estimation (ANPE). Instead of using MCMC to sample the posterior for every single galaxy separately, ANPE uses neural density estimators (NDE) to build a model of the posterior for *all* observable galaxies. Once the NDE is trained, generating the posterior requires only the observed SED and no additional model evaluations.

In this work, we present SEDflow, a method that applies ANPE to Bayesian galaxy SED modeling using the recent Hahn et al. (2022) SED model. We demonstrate that we can derive accurate posteriors with SEDflow and make Bayesian SED modeling fully scalable for the billions of galaxies that will be observed by upcoming surveys. As further demonstration, we apply SEDflow to the optical photometry of $\sim 33,000$ galaxies in the NASA-Sloan Atlas (NSA). We begin in Section 2

[1] $\approx$2 billion core hours; https://dmtn-135.lsst.io/
[2] https://dmtn-102.lsst.io/

by describing SBI using ANPE. We then present how we design and train SEDFLOW in Section 3 and describe the NSA observations in Section 4. We validate the accuracy of the posteriors from SEDFLOW in Section 5 and discuss the implications of our results and future steps in Section 6.

2. SIMULATION-BASED INFERENCE

The goal of Bayesian SED modeling, and probabilistic inference more broadly, is to infer the posterior probability distributions $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ of galaxy properties, $\boldsymbol{\theta}$, given observations, $\boldsymbol{x}$. For a specific $\boldsymbol{\theta}$ and $\boldsymbol{x}$, we can evaluate the posterior using Bayes' rule, $p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{\theta})\; p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ denotes the prior distribution and $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ the likelihood, which is typically assumed to have a Gaussian functional form:

$$\ln p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = -\frac{1}{2} \left(\boldsymbol{x} - m(\boldsymbol{\theta})\right)^T \mathbf{C}^{-1} \left(\boldsymbol{x} - m(\boldsymbol{\theta})\right). \tag{1}$$

$m(\boldsymbol{\theta})$ is the theoretical model, in our case a galaxy SED model from SPS. $\mathbf{C}$ is the covariance matrix of the observations. In practice, off-diagonal terms are often ignored and measured uncertainties are used as estimates of the diagonal terms.

Simulation-based inference (SBI; also known as "likelihood-free" inference) offers an alternative that requires no assumptions about the form of the likelihood. Instead, SBI uses a generative model, *i.e.* a simulation F, to generate mock data $\boldsymbol{x}'$ given parameters $\boldsymbol{\theta}'$: $F(\boldsymbol{\theta}') = \boldsymbol{x}'$. It uses a large number of simulated pairs $(\boldsymbol{\theta}', \boldsymbol{x}')$ to directly estimate either the posterior $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$, the likelihood $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, or the joint distribution of the parameters and data $p(\boldsymbol{\theta}, \boldsymbol{x})$. SBI has already been successfully applied to a number of Bayesian parameter inference problems in astronomy (*e.g.* Cameron & Pettitt 2012; Weyant et al. 2013; Hahn et al. 2017; Kacprzak et al. 2018; Alsing et al. 2018; Wong et al. 2020; Huppenkothen & Bachetti 2021; Zhang et al. 2021) and in physics (*e.g.* Brehmer et al. 2019; Cranmer et al. 2020).

2.1. *Amortized Neural Posterior Estimation*

SBI provides another a critical advantage over MCMC inference methods — it enables *amortized inference*. For SED modeling using MCMC, each galaxy requires $>10^5$ model evaluations to accurately estimate $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ (Hahn et al. 2022, Kwon *et al.* in prep.). Moreover, model evaluations for calculating the posterior of one galaxy cannot be used for another. This makes MCMC approaches for SED modeling of upcoming surveys computationally infeasible.

With density estimation SBI, we require a large number ($\sim 10^6$) of model evaluations only initially to train a neural density estimator (NDE), *i.e.* a neural network with parameters $\boldsymbol{\phi}$ that is trained to estimate the density $p_\phi(\boldsymbol{\theta} \,|\, \boldsymbol{x}')$. If the training covers the entire or the practically relevant portions of the $\boldsymbol{\theta}$ and $\boldsymbol{x}$ spaces, we can evaluate $p_\phi(\boldsymbol{\theta} \,|\, \boldsymbol{x}_i)$ for each galaxy i with minimal computational cost. The inference is therefore amortized and no additional model evaluations are needed to generate the posterior for each galaxy. This technique is called Amortized Neural Posterior Estimation (ANPE) and has recently been applied to a broad range of astronomical applications from analyzing gravitational waves (*e.g.* Wong et al. 2020; Dax et al. 2021) to binary microlensing lensing (Zhang et al. 2021). For SED modeling, the choice in favor of using ANPE is easy: the entire upfront cost for ANPE model evaluations would only yield posteriors of tens of galaxies with MCMC.

ANPE makes two important assumptions. First, the simulator F is capable of generating mock data $\boldsymbol{x}'$ that is practically indistinguishable from the observations. In terms of the expected signal, m in Eq. 1, this is the same requirement as any probabilistic modeling approach. But unlike likelihood-based evaluations, such as conventional MCMC, data generated for SBI need to include all relevant noise terms as well. We address both aspects in Sections 3.2 and 6.1. Second, ANPE assumes that the NDE is trained well: $p_\phi(\boldsymbol{\theta} \,|\, \boldsymbol{x}')$ is a good approximation of $p(\boldsymbol{\theta} \,|\, \boldsymbol{x}')$, and therefore of $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$. We assess this in Section 5.

ANPE commonly employs so-called "normalizing flows" (Tabak & Vanden-Eijnden 2010; Tabak & Turner 2013) as density estimators. Normalizing flow models use an invertible bijective transformation, f, to map a complex target distribution to a simple base distribution, $\pi(\boldsymbol{z})$, that is fast to evaluate. For ANPE, the target distribution is $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ and the $\pi(\boldsymbol{z})$ is typically a simple multivariate Gaussian, or mixture of Gaussians. The transformation $f : \boldsymbol{z} \to \boldsymbol{\theta}$ must be invertible and have a tractable Jacobian. This is so that we can evaluate the target distribution from $\pi(\boldsymbol{z})$ by a change of variable:

$$p(\boldsymbol{\theta} \,|\, \mathbf{x}) = \pi(\boldsymbol{z}) \left| \det \left(\frac{\partial f^{-1}}{\partial \boldsymbol{\theta}} \right) \right|. \tag{2}$$

Since the base distribution is easy to evaluate, we can also easily evaluate the target distribution. A neural network is trained to obtain f, the collection of its parameters form $\boldsymbol{\phi}$. The network typically consists of a series of simple transforms (*e.g.* shift and scale transforms) that are each invertible and whose Jacobians are easily calculated. By stringing together many such transforms, f provides an extremely flexible mapping from the base distribution.

Many different normalizing flow models are now available in the literature (*e.g.* Germain et al. 2015; Durkan et al. 2019). In this work, we use Masked Autoregressive Flow (MAF; Papamakarios et al. 2017). The autoregressive design (Uria et al. 2016) of MAF is particularly well-suited for modeling conditional probability distributions such as the posterior. Autoregressive models exploit chain rule to expand a joint probability of a set of random variables as products of one-dimensional conditional probabilities: $p(\boldsymbol{x}) = \prod_i p(x_i \,|\, x_{1:i-1})$. They then use neural networks to describe each conditional probability, $p(x_i \,|\, x_{1:i-1})$. In this context, we can add a conditional variable y on both sides of the equation, $p(\boldsymbol{x} \,|\, \boldsymbol{y}) = \prod_i p(x_i \,|\, x_{1:i-1}, \boldsymbol{y})$, so that the autoregressive model describes a conditional probability $p(\boldsymbol{x} \,|\, \boldsymbol{y})$. One drawback of autoregressive models is their sensitivity to the ordering of the variables. Masked Autoencoder for Distribution Estimation (MADE; Germain et al. 2015) models address this limitation using binary masks to impose the autoregressive dependence and by permutating the order of the conditioning variables. A MAF model is built by stacking multiple MADE models. Hence, it has the autoregressive structure of MADE but with more flexibility to describe complex probability distributions. In practice, we use the MAF implementation in the `sbi` Python package[3] (Greenberg et al. 2019; Tejero-Cantero et al. 2020).

3. SEDFLOW

[3] https://github.com/mackelab/sbi/

In this section, we present SEDFLOW, which applies ANPE to galaxy SED modeling for a scalable and accelerated approach. For our SED model, we use the state-of-the-art PROVABGS model from Hahn et al. (2022). Although many SED models have been recently used in the literature (*e.g.* BAGPIPES, Carnall et al. 2018; PROSPECTOR, Leja et al. 2017; Johnson et al. 2021), we choose PROVABGS because it will be used to analyze >10 million galaxy spectrophotometry measured by the DESI Bright Galaxy Survey (Hahn *et al.* in prep.). Below, we describe the PROVABGS model, the construction of the SEDFLOW training data using PROVABGS, and the training procedure for SEDFLOW.

3.1. *SED Modeling: PROVABGS*

We use the state-of-the-art SPS model of the PROVABGS (Hahn et al. 2022). The SED of a galaxy is modeled as a composite of stellar populations defined by stellar evolution theory (in the form of isochrones, stellar spectral libraries, and an initial mass function) and its star formation and chemical enrichment histories (SFH and ZH), attenuated by dust (see Walcher et al. 2011; Conroy 2013, for a review). The PROVABGS model, in particular, utilizes a non-parametric SFH with a starburst, a non-parametric ZH that varies with time, and a flexible dust attenuation prescription.

The SFH has two components: one based on non-negative matrix factorization (NMF) bases and the other, a starburst component. The SFH contribution from the NMF component is a linear combination of four NMF SFH basis functions, derived from performing NMF (Lee & Seung 1999; Cichocki & Phan 2009; Févotte & Idier 2011) on SFHs of galaxies in the Illustris cosmological hydrodynamical simulation (Vogelsberger et al. 2014; Genel et al. 2014; Nelson et al. 2015). The NMF SFH prescription provides a compact and flexible representation of the SFH. The second starburst component consists of a single stellar population (SSP) and adds stochasticity to the SFH.

The ZH is similar defined using two NMF bases dervied from Illustris. This ZH prescription enables us to flexibly model a wide range of ZHs and, unlike most SED models, it does not assume constant metallicity over time, which can significantly bias inferred galaxy properties (Thorne et al. 2021). The stellar evolution theory is based on Flexible Stellar Population Synthesis (FSPS; Conroy et al. 2009; Conroy & Gunn 2010) with the MIST isochrones (Paxton et al. 2011, 2013, 2015; Choi et al. 2016; Dotter 2016), the Chabrier (2003) initial mass function (IMF), and a combination of the MILES (Sánchez-Blázquez et al. 2006) and BaSeL (Lejeune et al. 1997, 1998; Westera et al. 2002) libraries. The SFH and ZH are binned into 43 logarithmically-space time and SSPs are evalulated at each time bin using FSPS. The SSPs are summed up to get the unattenuated rest-frame galaxy SED.

Lastly, PROVABGS attenuates the light from the composite stellar population using the two component Charlot & Fall (2000) dust attenuation model with diffuse-dust (ISM) and birth cloud (BC) components. All SSPs are attenuated by the diffuse dust using the Kriek & Conroy (2013) attenuation curve. Then, the BC component provides extra dust attenuation on SSPs younger than 100 Myr with young stars that are embedded in modecular clouds and HII regions. In total the PROVABGS SED model has 12 parameters: stellar mass (M_*), six SFH parameters ($\beta_1, \beta_2, \beta_3, \beta_4, t_{\rm burst}, f_{\rm burst}$), two ZH parameters ($\gamma_1, \gamma 2$), and three dust attenuation parameters ($\tau_{\rm BC}, \tau_{\rm ISM}, n_{\rm dust}$). Each PROVABGS model evaluation takes ∼340 ms.

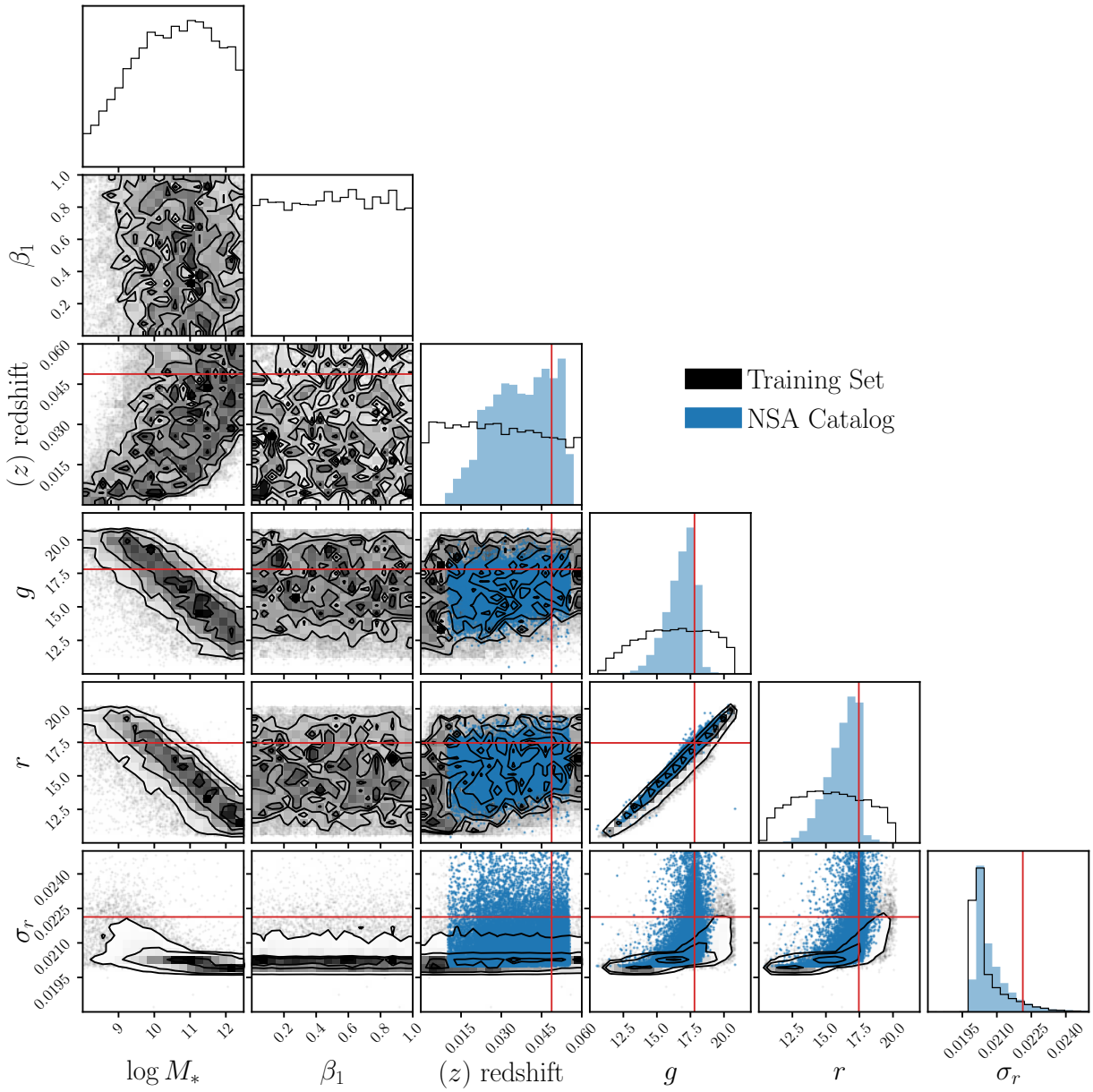

Figure 1. The distribution of SED model parameters, redshift, photometric magnitudes, and uncertainties of the data used to train SEDFLOW. We plot only a subset of the parameters ($\log M_*$, β_1) and photometric bands for clarity. The training data was constructed by sampling SED model parameters from the prior and forward modeling optical photometry using the PROVABGS and noise models (Section 3). For comparison, we present the distribution of redshift, magnitudes, and uncertainties for galaxies in the NSA catalog (blue). *The training set encompasses the observations; thus,* SEDFLOW *can be used to infer the posterior for the NSA galaxies.*

3.2. *Training Data*

In this section, we describe how we construct the training data for SEDFLOW using the PROVABGS SED model. First, we sample $N_{\rm train}$ SED model parameters from a prior: $\theta' \sim p(\theta)$. We use

the same priors as Hahn et al. (2022): uniform priors over $M_*, t_{\rm burst}, f_{\rm burst}, \gamma_1, \gamma_2, \tau_{\rm BC}, \tau_{\rm ISM}, n_{\rm dust}$ with broad conservative ranges and Dirichlet prior over $\beta_1, \beta_2, \beta_3, \beta_4$, chosen to normalize the NMF SFH. For each θ', we also uniformly sample a redshift within the range of the NSA: $z' \sim \mathcal{U}(0., 0.2)$. Next, we forward model mock observables. We calculate the rest-frame galaxy SED from PROVABGS and redshift it: $F(\lambda; \theta', z)$. Afterwards, we convolve F with optical broadband filters, R_X, to generate noiseless photometric fluxes:

$$f_X(\theta', z') = \int F(\lambda; \theta', z')\, R_X(\lambda)\, {\rm d}\lambda \tag{3}$$

The next step of the forward model is to apply noise. We assign photometric uncertainties, σ'_X, by sampling an estimate of the observed $p(\sigma_X|f_X)$ of NSA galaxies. Then, we apply Gaussian noise

$$\hat{f}_X(\theta', z', \sigma'_x) = f_X(\theta', z') + n_X \quad \text{where } n_X \sim \mathcal{N}(0, \sigma'_X) \tag{4}$$

to derive the forward modeled photometric flux.

For our estimate of $p(\sigma_X|f_X)$, we use an empirical estimate based on NSA photometry and measured uncertainties. For each of the five optical bands, we separately estimate

$$\hat{p}(\sigma_X|f_X) = \mathcal{N}\left(\mu_{\sigma_X}(f_X), \sigma_{\sigma_X}(f_X)\right) \tag{5}$$

as a Gaussian in magnitude-space. μ_{σ_X} and σ_{σ_X} are the median and standard deviation of σ_X as a function of f_X that we estimate by evaluating them in f_X bins and interpolating over the bins. Any θ' that is assigned a negative σ'_X is removed from our training data. We also remove any training data with $f_X(\theta')$ outside the range of NSA photometry.

As SBI requires an accurate noise model, one might be concerned about the simplicity of our model. If the noise model is incorrect, estimates of the parameters θ would be biased by an amount that is impossible to predict. We therefore add the noise variances σ_X to the conditioning variables of the posterior model, *i.e.* we train and evaluate $p_\phi(\boldsymbol{\theta} \,|\, \boldsymbol{x}, \{\sigma_X\})$. This means we merely have to ensure that σ'_X spans the observed σ_X values in order to have a posterior that is robust to our choice of noise model. We discuss this further in Section 6.

In total, we constuct $N_{\rm train} = 1,131,561$ sets of SED parameters, redshift, photometric uncertainties, and mock NSA photometry. In Figure 1, we present the distribution of the training data $\{(\theta', z, \sigma'_X, \hat{f}_X)\}$ (black). We include select SED model parameters ($\log M_*$, β_1), redshift, photometry in the g and r bands, and photometric uncertainty in the r band. We also include the (z, σ_X, f_X) distribution of NSA galaxies (blue), for comparison. The photometry and uncertainties are in magnitude-space. The distribution of the training data spans the distribution of NSA galaxies.

3.3. *Training* SEDflow

For SEDflow, we use a MAF normalizing flow model (Section 2.1) with 15 MADE blocks, each with 2 hidden layers and 500 hidden units. In total, the model has 7,890,330 parameters, $\boldsymbol{\phi}$. We determine this architecture through experimentation. Our goal is to determine $\boldsymbol{\phi}$ of the MAF model, $p_\phi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$, so that it accurately estimates the posterior probability distribution $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$. $\boldsymbol{\theta}$ represent

the SED parameters and $\boldsymbol{x} = (f_X, \sigma_X, z)$. We do this by minimizing the KL divergence between $p_\phi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ and $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$: $D_{\rm KL}(p \,||\, p_\phi)$.

In practice, we split the training data into a training and validation set with a 90/10 split. Afterwards, we maximize the total log likelihood $\sum_i \log p_\phi(\boldsymbol{\theta}_i \,|\, \boldsymbol{x}_i)$ over training set, which is equivalent to minimizing $D_{\rm KL}(p \,||\, p_\phi)$. We use the ADAM optimizer (Kingma & Ba 2017) with a learning rate of 5×10^{-4}. To prevent overfitting, we evaluate the total log likelihood on the validation data at every training epoch and stop the training when the validation log likelihood fails to increase after 20 epochs. Training our model with a batch size of 50 takes roughly a day on a single 2.6 GHz Intel Skylake CPU. Given our small batch size, we find similar training times when using CPUs or GPUs.

4. NASA-SLOAN ATLAS

As a demonstration of its speed and accuracy, we apply SEDFLOW to optical photometry from the NASA-Sloan Atlas[4] (NSA). The NSA catalog is a re-reduction of SDSS DR8 (Aihara et al. 2011) that includes an improved background subtraction (Blanton et al. 2011). We use SDSS photometry in the u, g, r, i, and z bands, which are corrected for galactic extinction using Schlegel et al. (1998). To ensure that the galaxy sample is not contaminated, we impose a number of additional quality cuts by excluding:

- objects where the centroiding algorithm reports the position of the peak pixel in a given band as the centroid. The SDSS photometric pipeline can struggle to accurately define the center of objects near the edge or at low signal-to-noise, so these cases are often associated with spurious objects.
- objects that have pixels that were not checked for peaks by the deblender.
- objects where more than 20% of point-spread function (PSF) flux is interpolated over as well as objects where the interpolation affected many pixels and the PSF flux error is inaccurate. The SDSS pipeline interpolates over pixels classified as bad (*e.g.* cosmic ray).
- objects where the interpolated pixels fall within 3 pixels of their center and they contain a cosmic ray that was interpolated over.
- objects that were not detected at $\geq 5\sigma$ in the original frame, that contain saturated pixels, or where their radial profile could not be extracted.

By excluding these objects, we avoid complications from artificats in the photometry that we do not model. For additional details on the quality flags, we refer readers to the SDSS documentation[5]. After the quality cuts, we have a total of 33,883 NSA galaxies in our sample.

5. RESULTS

Now that we have trained SEDFLOW, we can estimate the posterior, $p(\boldsymbol{\theta} \,|\, \boldsymbol{x}_i)$, for any $\boldsymbol{x}_i = \{f_{X,i}, \sigma_{X,i}, z_i\}$. In practice, we do this by drawing samples from the SEDFLOW NDE model. Since

[4] http://nsatlas.org/
[5] https://www.sdss.org/dr16/algorithms/flags_detail

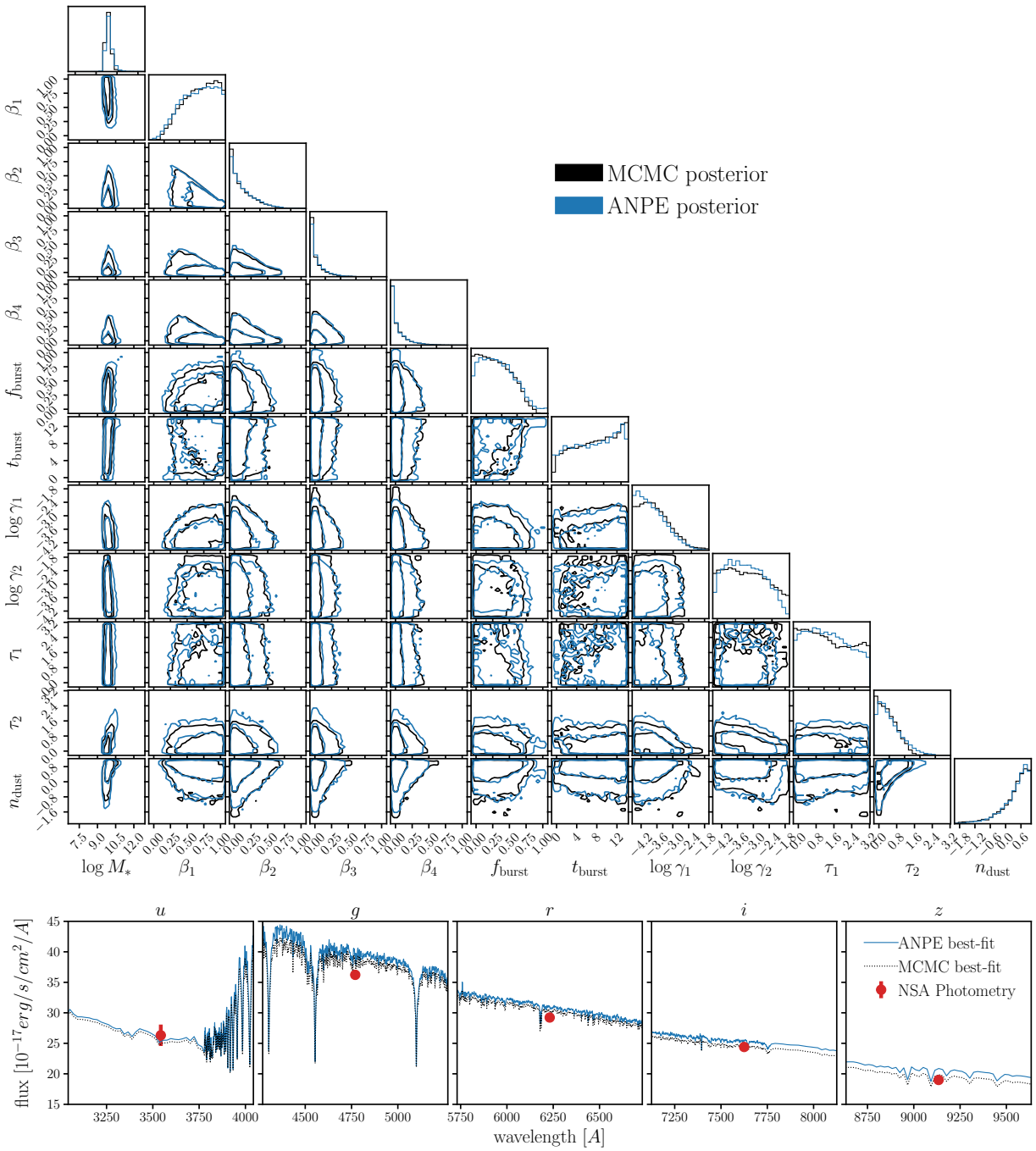

Figure 2. A comparison of the posteriors of the 12 SED model parameters derived from standard MCMC sampling (black) and our SEDFLOW (blue) for an arbitrarily selected NSA galaxy (`NSAID` = 72). The posteriors are in excellent agreement for all of the SED parameters. In the bottom panel, we present the SEDs of the the best-fit parameter values from the SEDFLOW (blue) and MCMC posteriors (black dotted), which are all in excellent agreement with the observed NSA photometric flux (red). Estimating the posterior using MCMC sampling requires 10 CPU hours. Even using neural emulators to accelerate likelihood evaluations, MCMC sampling requires 10 CPU minutes. *With* SEDFLOW*, inferring the full posterior takes 1 second per galaxy.*

we use a normalizing flow, this is trivial: we generate samples from the target distribution of the normalizing flow, a multivariate Gaussian distribution in our case, then we transform the samples using the bijective transformation in Eq. 2 that we trained. The transformed samples follow $p_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}_i)$ and estimate the posterior, $p(\boldsymbol{\theta} \mid \boldsymbol{x}_i)$.

Next, we validate the accuracy of the SEDFLOW posterior estimates. As a first test, we compare the posterior from SEDFLOW to the posterior derived from MCMC sampling for a single arbitrarily chosen NSA galaxy in Figure 2 (`NSAID` = 72). In the top, we present the the posterior distribution of the 12 SED model parameters for the SEDFLOW posterior (blue) and MCMC posterior (black). *The* SEDFLOW *posterior is in excellent agreement with the MCMC posterior for all of the SED parameters.*

In the bottom of Figure 2, we compare the SEDs of the best-fit parameter values from the SEDFLOW (blue) and MCMC posteriors (black dotted). We also include the NSA photometric flux of the selected galaxy (red). The best-fit SED from the SEDFLOW posterior is also in good agreement with both the MCMC best-fit SED and the NSA photometry.

The key advantage of ANPE is that it enables accurate Bayesian inference orders of magnitude faster than conventional methods. We derive the MCMC posterior using the ZEUS ensemble slice-sampler (Karamanis & Beutler 2020) with 30 walkers and 10,000 iterations. 2,000 of the iterations are discarded for burn-in. In total, the MCMC posterior requires >100,000 SED model evaluations. Since each evaluation takes ∼340 ms, it takes ∼10 CPU hours for a single MCMC posterior. Recently, SED modeling has adopted neural emulators to accelerate SED model evaluations (Alsing et al. 2020). In Hahn et al. (2022), for instance, the PROVABGS emulator takes only ∼2.9 ms to evaluate, >100× faster than the original model. Yet, even with emulators, due to the number of evaluations necessary for convergence, an MCMC posterior takes ∼10 CPU minutes. Meanwhile, after training, *the* SEDFLOW *posterior takes* 1 *second* — $>10^4\times$ *faster than MCMC.*

The posteriors from SEDFLOW and MCMC are overall in excellent for NSA galaxies, besides the one in Figure 2. However, we do not know the true SED parameters for these galaxies so to further validate SEDFLOW, we use test synthetic photometry, where we know the truth. We sample 1000 SED parameters from the prior, $\{\boldsymbol{\theta}_i^{\rm test}\} \sim p(\boldsymbol{\theta})$, and forward model synthetic NSA observations, $\{\boldsymbol{x}_i^{\rm test}\}$, for them in the same way as the training data (Section 3.2). Afterwards, we generate posteriors for each of $\boldsymbol{x}_i^{\rm test}$ using SEDFLOW: $\{p(\boldsymbol{\theta} \mid \boldsymbol{x}_i^{\rm test})\}$.

In Figure 3, we present the probability-probability (p-p) plot of the SEDFLOW posteriors for the test data. The p-p plot presents the cumulative distribution function (CDF) of the percentile score of the true value within the marginalized posterior for each parameter. For true posteriors, the percentiles are uniformly distributed so the CDF is a diagonal (black dashed). Overall, the CDFs for SEDFLOW lie close to the diagonal for each of the SED parameters. *Hence, the* SEDFLOW *posteriors are in excellent agreement with the true posteriors.*

In Figure 3, we also include the CDFs of the SED parameters for the MCMC posteriors derived for a subset of 100 test observations (gray dotted). Comparing the CDFs from the MCMC posteriors to those of SEDFLOW, we find that the SEDFLOW posteriors are actually in better agreement with the true posteriors. This is due to the fact that MCMC posteriors are also only estimates of the true

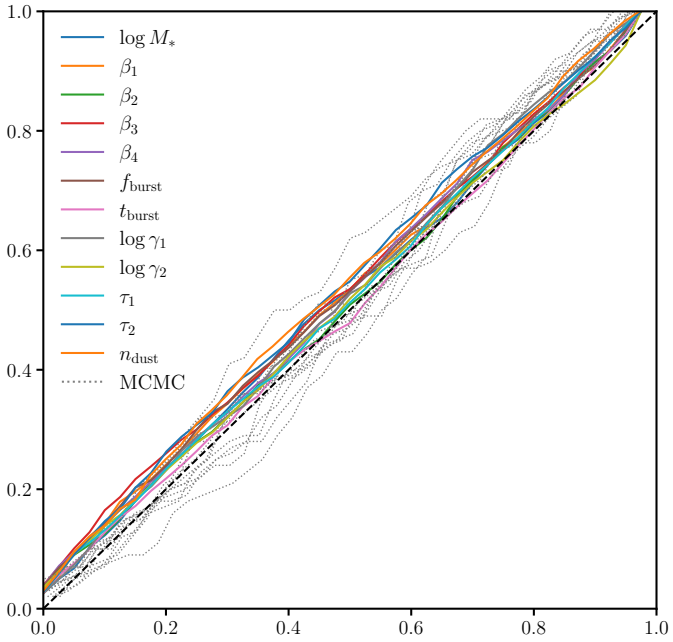

Figure 3. Probability-probability (p-p) plot of the SEDflow posteriors for 1000 synthetic test observations, with known true parameter values. We plot the CDFs of the percentile score of the true values within the marginalized SEDflow posteriors for each SED parameter. For the true posteriors, the percentile score is uniformly distributed so the CDF is diagonal (black dashed). For reference, we include the p-p plot of the posterior estimated from MCMC sampling (gray). *The* SEDflow *posteriors are in excellent agreement the true posteriors.*

posterior and are subject to limitations in initialization, sampling, and convergence. Posteriors from SEDflow are not impacted by these limations, so the comparison highlights additional advantages of an ANPE approach besides the $>10^4\times$ performance improvement.

We examine another validation of the SEDflow posteriors using simulation-based calibration (SBC; Talts et al. 2020). Rather than using percentile scores, SBC examines the distribution of the rank statistics of the true parameter values within the marginalized posteriors. It addresses the limitation that the CDFs only asymptotically approach the true values and that the discrete sampling of the posterior can cause artifacts in the CDFs. In Figure 4, we present SBC of each SED parameter for the SEDflow posteriors (blue) using the 1000 test observations. For comparison, we include the SBC for the MCMC posteriors (gray dotted). Similar to the percentile score, the distribution of the rank statistic is uniform for the true posterior (black dashed). The rank statistic distribution for the SEDflow posteriors are nearly uniform for all SED parameters. Hence, *we confirm that the* SEDflow *posteriors are in excellent agreement with the true posterior.*

An advantage of SBC is that by examining the deviation of rank statistics distribution from uniformity, we can determine how the posterior estimates deviate from the true posteriors. For instance, if the distribution has a U-shape where the true parameter values are more often at the lowest and highest ranks, then the posterior estimates are narrower than the true posteriors. If the distribution has a $\cap$-shape, then the posterior estimates are broader than the true posteriors. Any asymmetry

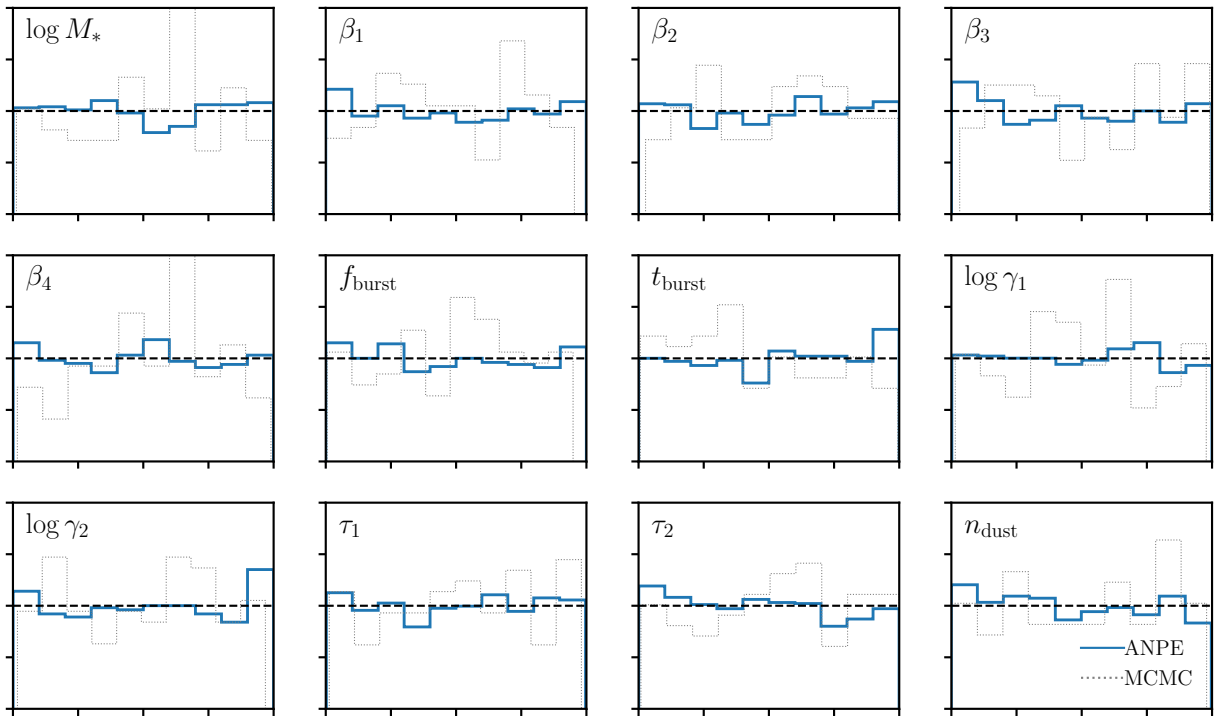

Figure 4. Simulation-based calibration plot of the SEDflow posteriors for 1000 synthetic test observations. The histogram in each panel represents the distribution of the rank statistic of the true value within the marginalized SEDflow posterior (blue) for each SED parameter. For the true posterior, the rank statistics will have a uniform distribution (black dashed). For reference, we include the rank distribution of the MCMC posteriors for a subset of 100 test data (gray dotted). The rank statistic distribution of SEDflow is nearly uniform for all of the SED parameters. Therefore, SEDflow *provides unbiased and accurate estimates of the true posteriors.*

in the distribution implies that the posterior estimates are biased. For the SEDflow posteriors, we find none of these features for any of the SED parameters. Hence, SEDflow provides unbiased and accurate estimates of the true posteriors for all SED parameters.

With the accuracy SEDflow validated, we apply it to derive posteriors for all of our NSA galaxies (Section 4). Analyzing all 33,883 NSA galaxies takes <12 CPU hours. For each galaxy, we generate 10,000 samples of the 12-dimensional posterior, $p(\boldsymbol{\theta} \mid \boldsymbol{x})$. These posteriors on SED parameters represent posteriors on stellar mass, SFH, ZH, and dust content of the NSA galaxies. To maximize the utility of the posteriors further, we use them to derive posteriors on the following additional galaxy properties: SFR averaged over 1Gyr, mass-weighted metallicity, and mass-weighted stellar age (see Eq. 17 in Hahn et al. 2022 for the exact calculation). We publicly release all of the posteriors for our NSA sample at some-url-here. Furthermore, SEDflow model and all of the software used to train it are publicly available at https://github.com/changhoonhahn/SEDflow/.

6. DISCUSSION

6.1. *Forward Model*

In the previous section, we demonstrated the accuracy of SEDFLOW posteriors. Nevertheless, a primary determining factor the fidelity of SEDFLOW, or any ML model, is the quality of the training data set and, thus, the forward model used to construct it. Below, we discuss the caveats and limitations of our forward model, which has two components: the PROVABGS SED model and a noise model (Section 3.2).

First, for our noise model, we assign noiseless photometric fluxes uncertainties based on an empirical estimate of $p(\sigma_X \mid f_X)$ for each band independently. Afterwards, the assigned σ_X is used to apply Gaussian noise to the photometric flux (Eq. 4). This is a simplicistic noise model and, as the bottom right ($g - \sigma_r$ and $r - \sigma_r$) panels of Figure 1 reveal, there are discrepancies in the magnitude versus uncertainty distributions of the training data and observations. Despite these discrepancies, SEDFLOW provides excellent estimates of the true posterior. This is because we design our ANPE to include σ_X as a conditional variable (Section 3.3). The $f_X - \sigma_X$ distribution of our training data does not impact the accuracy of the posteriors as long as there are sufficient training data near $\boldsymbol{x}$ to train the NDE in that region.

A more accurate noise model will, in theory, improve the performance of SEDFLOW because the $\boldsymbol{x}$-space of the training data will more effectively span the observations. In other words, there will be fewer training data expended in regions of $\boldsymbol{x}$-space that are devoid of observations. However, for our application to SED modeling, we do not find significantly improved performance when we alter the noise model. This suggests that even with our simplistic noise model, the $\boldsymbol{x}$-space of observations is covered sufficiently well by the training data. We note that when we decrease $N_{\rm train}$ to below 500,000, SEDFLOW posteriors are significantly less accurate. A more realistic forward model may reduce this $N_{\rm train}$ threshold for accurate posteriors. However, generating $N_{\rm train} \sim 1,000,000$ training has a negligible computational cost compared to MCMC SED modeling, so we do not consider it necessary to explore this further.

Next, we consider limitations in the PROVABGS SED model used in our forward model. Our SED model uses a compact and flexible prescription for SFH and ZH that can describe a broad range of SFHs and ZHs. However, the prescription is derived from simulated Illustris galaxies, whose SFHs and ZHs may be not reflect the full range of SFHs and ZHs of real galaxies. If certain subpopulations of observed galaxies have SFHs and ZHs that cannot be described by the PROVABGS prescriptions, they cannot be accurately modeled. Furthermore, even if the PROVABGS SFH and ZH prescriptions are sufficient, there are limitations in our understanding of stellar evolution.

There is currently no consensus in the stellar evolution, stellar spectral libraries, or IMF of galaxies (*e.g.* Treu et al. 2010; van Dokkum & Conroy 2010; Rosani et al. 2018; Ge et al. 2019; Sonnenfeld et al. 2019). The PROVABGS model uses MIST isochrones, Chabrier (2003) IMF, and the MILES + BaSeL spectral libraries. These choices limit the range of SEDs that can be produced by the training data. For instance, if galaxies have significant variations in their IMF, assuming a fixed IMF would falsely limit the range of our training data. A more flexible SED model that includes uncertainties in SPS would broaden the range of galaxy SEDs that can be modeled. Data-driven approaches may also enable SED models to be more descriptive (*e.g.* Hogg et al. 2016; Portillo et al. 2020). Improving SED models, however, is beyond the scope of this work. Our focus is on improving the Bayesian

inference framework. In that regard, the limitations of the SED model equally impacts conventional approaches with MCMC.

We encounter the caveats above when we apply SEDFLOW to the NSA catalog. For a small fraction of NSA galaxy SEDs (591 out of 33,883), SEDFLOW generates posteriors that are outside of the prior volume. We encounter these failures because the photometry or photometric uncertainties of these galaxies lie outside of the support of the training data where SEDFLOW is not trained. They either have higher photometric uncertainties, for a given magnitude, or bluer photometric colors than the training data. Some of these may be observational artifacts or problematic photometry. Nevertheless, SEDFLOW fails because the limitations in our noise and SPS models, mentioned above, prevent us from construct training data near them. Since this issue only affects a small fraction of the NSA galaxies, we flag them in our catalog and infer their galaxy properties by applying PROVABGS with MCMC sampling for completeness. For more details on these failures, we refer readers to Appendix A.

To test for limitations of the forward model, we can construct additional tests of posteriors derived from ANPE. For instance, the χ^2 of the best-fit parameter value from the estimated posterior can be used to assess whether the best-fit model accurately reproduces observations. This would only require one additional model evaluation per galaxy. As another test of the posteriors, one can construct an Amortized Neural Likelihood Estimator (ANLE) using the same training data. Unlike the ANPE, which estimates $p(\boldsymbol{\theta} \,|\, f_X, \sigma_X, z)$, the ANLE would estimate $p(f_X \,|\, \boldsymbol{\theta}, \sigma_X, z)$. We can then further validate the posteriors by assessing whether the observed photometry lies within the ANLE. Based on the overall high level of accuracy of SEDFLOW posteriors, we do not explore these additional tests; however, they can be used to further validate any ANPE posterior.

6.2. *Advantages of* SEDFLOW

The primary advantage of SEDFLOW is its computational speed. This becomes even more relevant if one wants to add further parameters to address the concerns about the choices in current SPS models we described in Section 6.1. To relax these assumptions, SPS models would need to introduce additional parameters that flexibly model these uncertainties (Conroy et al. 2009; Conroy & Gunn 2010). While the dimensionality of current SPS models already makes MCMC methods computational infeasible, ANPE has already been applied to higher dimensional applications. For instance, Dax et al. (2021) constructed an accurate ANPE for a 15-dimensional model parameter space and 128-dimensional conditional variable space. NDE is an actively developing field in ML and new methods are constantly emerging (*e.g.* Wu et al. 2020; Dhariwal & Nichol 2021). Since ANPE can handle higher dimensionality, we can in the future include additional parameters that model uncertainties in SPS. This will not only improve our SED modeling, but also improve our understanding of stellar evolution and the IMF.

In addition to enabling scalable SED modeling for the next generation galaxy surveys, SEDFLOW will enable us to tackle other key challenges in SED modeling. For example, recent works have demonstrated that priors of SED models can significantly impact the inferred galaxy properties (Carnall et al. 2018; Leja et al. 2019; Hahn et al. 2022). Even "uniformative" uniform priors on SED model parameters can impose undesirable priors on derived galaxy properties such as M_*, SFR, SFH, or ZH. To avoid significant biases, galaxy studies must carefully select priors and validate their

results using multiple different choices. With an MCMC approach, selecting a different prior means reevaluating every posterior and repeating all the SED model evaluations in the MCMC sampling.

For an ANPE approach, the prior is set by the distribution of parameters in the training data. For a new prior, instead of reconstructing the training data, we can resample it in such a way that the parameters follow the new prior. Then, the ANPE model can be re-trained, re-validated on the test data, and re-deployed on observations. Each of these steps require substantially less computational resources than generating a new set of training data or using MCMC methods. Hence, the ANPE approach provides a way to efficiently vary the prior without multiplying computational costs.

7. SUMMARY AND OUTLOOK

By analyzing the SED of a galaxy, we can infer detailed physical properties such as its stellar mass, star formation rate, metallicity, and dust content. These properties serve as the building blocks of our understanding of how galaxies form and evolve. State-of-the-art SED modeling methods use MCMC sampling to perform Bayesian statistical inference. They derive posterior probability distributions of galaxy properties given observation that accurately estimate uncertainties and parameter degeneracies to enable more rigorous statistical analyses. For the dimensionality of current SED models, deriving a posterior requires $\gtrsim$100,000 model evaluations and take $\gtrsim 10-100$ CPU hours per galaxy. Upcoming galaxy surveys, however, will observe *billions* of galaxies using *e.g.* DESI, PFS, Rubin observatory, James Webb Space Telescope, and the Roman Space Telescope. Analyzing all of these galaxies with current Bayesian SED models is infeasible and would require hundreds of billions of CPU hours. Even with recently a proposed emulator, which accelerates model evaluations by three to four orders to magnitude, the computation cost of SED modeling would remain a major bottleneck for galaxy studies.

We demonstrate in this work that Amortized Neural Posterior Estimation (ANPE) provides an alternative *scalable* approach for Bayesian inference in SED modeling. ANPE is a simulation-based inference method that formulates Bayesian inference as a density estimation problem and uses neural density estimators (NDE) to approximate the posterior over the full space of observations. The NDE is trained using parameter values drawn from their respective prior and mock observations simulated with these parameters. Once trained, a posterior can be obtained from the NDE by providing the observations as the conditional variables without any additional model evaluations.

In this work, we present SEDflow, a galaxy SED modeling method using ANPE and PROVABGS, a flexible SED model that uses a compact non-parameteric SFH and ZH prescriptions and was recently validated in Hahn et al. (2022). Furthermore, we apply SEDflow to optical photometry from the NASA-Sloan Atlas as demonstration and validation of our ANPE approach. We present the following key results from our analysis.

- We train SEDflow using a data set of $\sim$1 million SED model parameters and forward model synthetic SEDs. The parameters are drawn from a prior and the forward model is based on the PROVABGS and noise models. We design the ANPE to estimate $p(\theta|f_X, \sigma_X, z)$, where f_X, σ_X, and z are the photometry, photometric uncertainty, and redshift, respectively. For its architecture, we use a MAF normalizing flow with 15 MADE blocks each with 2 hidden layers and 500 hidden units. Training SEDflow requires roughly 1 day on a single CPU. Once

trained, deriving posteriors of galaxy properties for a galaxy takes $\sim$1 second, $10^5\times$ faster than traditional MCMC sampling.

- Posteriors derived using SEDFLOW show excellent agreement with posteriors derived from MCMC sampling. We further validate the accuracy of the posteriors by applying SEDFLOW to synthetic observations with known true parameter values. Based on statistical metrics used in the literature (p-p plot and SBC), we find excellent agreement between the SEDFLOW and the true posterior.
- Lastly, we demonstrate the advantages of SEDFLOW by applying it to the NASA-Sloan Atlas. Estimating the posterior of $\sim$34,000 galaxies takes $<$12 CPU hours. We make the catalog of posteriors publicly available at URL. For each galaxy, the catalog contains posteriors on all 12 PROVABGS SED model parameters. In terms of galaxy properties, the catalog includes posteriors of M_*, average SFR over 1Gyr, mass-weighted metallicity, mass-weighted age, and dust optical depth.

This work highlights the advantages of using an ANPE approach to Bayesian SED modeling. Our approach can easily be extended beyond this application. For instance, we can include multi-wavelength photometry at ultra-violet (UV) or infrared (IR) wavelengths. We can also modify SEDFLOW to infer redshift from photometry. In SEDFLOW, we include redshift as a conditional variable, since NSA provides spectroscopic redshifts. However, redshift can be included as an inferred variable rather than a conditional one. Then, we can apply SEDFLOW to infer galaxy properties from photometric data sets without redshift measurements while marginalizing over the redshift prior. If we do not require spectroscopic redshifts, SEDFLOW can be extended to much larger data sets that span fainter and broader galaxy samples. Conversely, we can use SEDFLOW to infer more physically motivated photometric redshifts, where we marginalize over our understanding of galaxies rather than using templates.

The ANPE approach to SED modeling can also be extended to galaxy spectra. Constructing an ANPE for the full data space of spectra, however, requires estimating a dramatically higher dimensional probability distribution. SDSS spectra, for instance, have $\sim$3,600 spectral elements. Furthermore, in our approach we include the uncertainties of observables as conditional variables, which double the curse of dimensionality. Recent works, however, have demonstrated that galaxy spectra can be represented in a compact low-dimensional space using autoencoders (Portillo et al. 2020, Melchior & Hahn, in prep.). In Portillo et al. (2020), they demonstrate that SDSS galaxy spectra can be compressed into 7-dimensional latent variable space with little loss of information. Such spectral compression dramatically reduces the dimensionality of the conditional variable space to dimensions that can be tackled by current ANPE methods. We will explore SED modeling of galaxy spectrophotometry using ANPE and spectral compression in a following work.

ACKNOWLEDGEMENTS

It's a pleasure to thank Adam Carnall, Miles Cranmer, Kartheik Iyer, Andy Goulding, Jenny E. Green, Uroš Seljak, Michael A. Strauss, ... for valuable discussions and comments. This work was supported by the AI Accelerator program of the Schmidt Futures Foundation.

APPENDIX

A. TESTING OUTSIDE SEDFLOW TRAINING RANGE

For a small number of NSA galaxies, 591 out of 33,883, SEDFLOW does not produce valid posteriors. The normalizing flow of SEDFLOW generates posteriors that are entirely outside of the prior volume because the observables of the NSA galaxies lie outside of the support of the SEDFLOW training data (Section 3.2). These galaxies lie outside of the support for two reasons: (1) they have unusually high photometric uncertainties that are not accounted for in our noise model or (2) they have photometric colors that cannot be modeled by our SED model. In Figure 5, we present the distribution of photometric magnitudes, uncertainties, and redshift ($\boldsymbol{x} = \{f_X, \sigma_X, z\}$) of these NSA galaxies. We mark galaxies that are outside the SEDFLOW training data support (black) due to (1) in orange and (2) in blue.

We classify NSA galaxies as (1) in Figure 5, if they have σ_X that is unusually high for a given f_X in at least one photometric band: $\sigma_X > \mu_{\sigma_X}(f_X) + 3\sigma_{\sigma_X}(f_X)$ (see Eq. 4). There are 490 NSA galaxies without valid SEDFLOW posteriors due to (1). These galaxies have a $\boldsymbol{x}$ distribution that is significant discrepant from the distribution of the training data, with many galaxies that lie well beyond the locus of training data points. The SEDFLOW estimate of $p(\boldsymbol{\theta} \mid f_X, \sigma_X, z)$ is only accurate in regions of $\boldsymbol{x}$-space where there is sufficient training data. This requirement is not met for these NSA galaxies. In principle, if we use a more conservative noise model than Eq. 4 and construct noisier training data, we can expand the support of SEDFLOW. SEDFLOW would then produce sensible posteriors for more NSA galaxies. However, there is an inherent trade-off. For a training data set of fixed size, a more conservative noise model would reduce the amount of training data in $\boldsymbol{x}$-space where the vast majority of NSA galaxies lie and can reduce the accuracy of the posteriors in these regions. Since, SEDFLOW fails for only a small fraction of NSA galaxies, we do not explore more conserative noise models in this work.

Next, we examine the NSA galaxies that lie outside of the SEDFLOW support because they have colors that cannot be modeled by our SED model. In Figure 5, we classify NSA galaxies as (2) if any of their colors (*e.g.* $u-r$, $u-g$, $r-z$) is bluer than the 99.9% percentile of the training data color. There are 98 NSA galaxies without valid SEDFLOW posteriors due to (2). The i versus z magnitude panel in particular highlights how a significant number of NSA galaxies are bluer than the training data. The fact that the training data do not span these colors suggests that the PROVABGS SED model may not fully describe all types of galaxies in observations. As we discuss in Section 6, this may be due to limitations in the SFH and ZH prescription in the PROVABGS model or even in our understanding of stellar evolution. Limitations of the SED model equally impacts conventional MCMC sampling approaches and is beyond the scope of this work. We, therefore, do not examine the issue further. For completeness, we derive posteriors for NSA galaxies for which SEDFLOW fails to estimate valid posteriors using PROVABGS with MCMC sampling with the same configuration as Hahn et al. (2022).

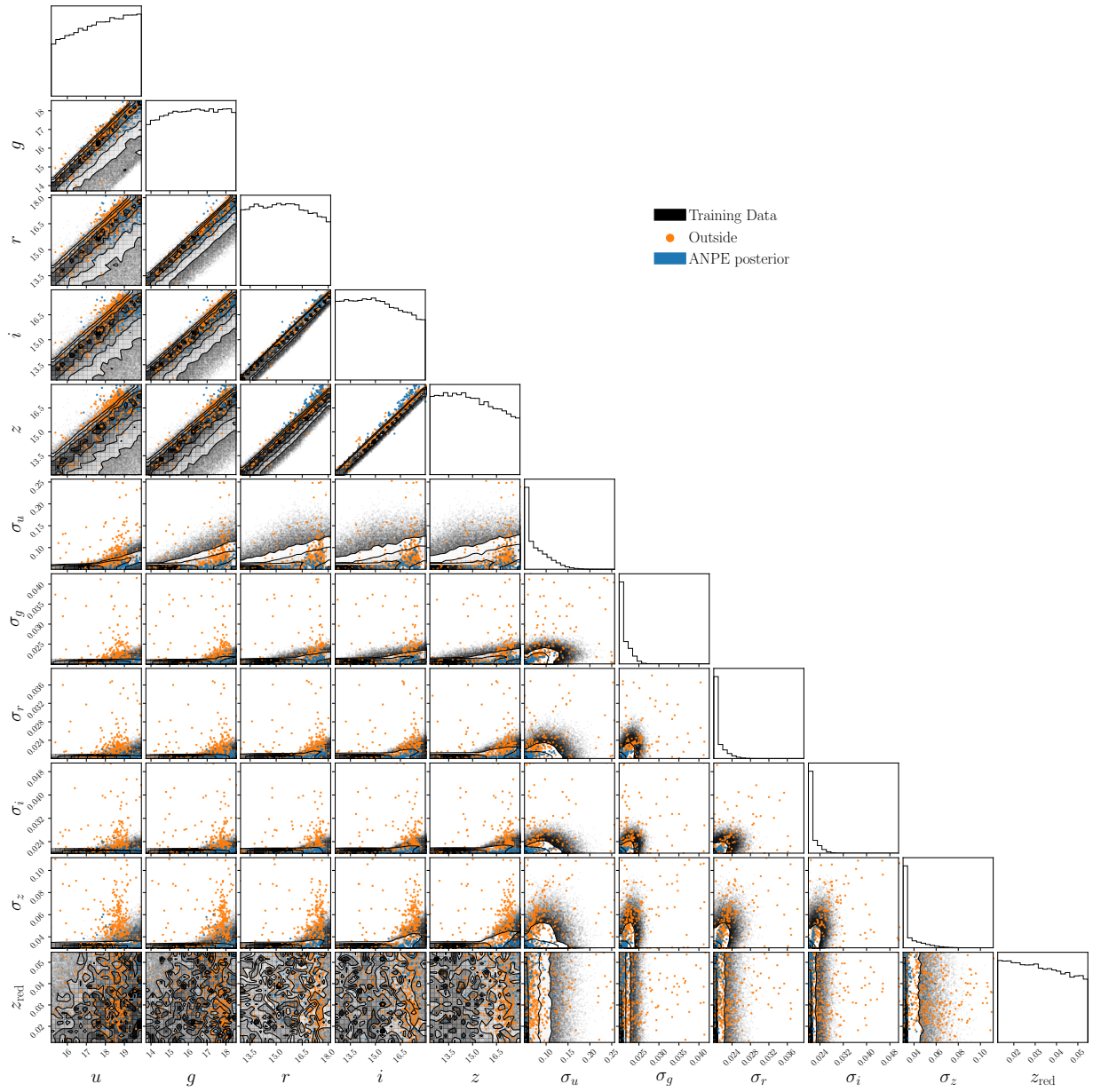

Figure 5. The distribution of photometric magnitudes, uncertainities, and redshift of NSA galaxies for which SEDFLOW does not produce valid posteriors. These galaxies lie outside of the suport of the SEDFLOW training data (black) so SEDFLOW cannot accurately estimate their posteriors. We mark the NSA galaxies that have unusually high photometric uncertainties that are not accounted for in our noise model in orange and galaxies that have photometric colors that cannot be modeled by our SED model in blue.

REFERENCES

Acquaviva, V., Gawiser, E., & Guaita, L. 2011, The Astrophysical Journal, 737, 47, doi: 10.1088/0004-637X/737/2/47

Aihara, H., Allende Prieto, C., An, D., et al. 2011, The Astrophysical Journal Supplement Series, 193, 29, doi: 10.1088/0067-0049/193/2/29

Alsing, J., Charnock, T., Feeney, S., & Wandelt, B. 2019, Monthly Notices of the Royal Astronomical Society, 488, 4440, doi: 10.1093/mnras/stz1960

Alsing, J., Wandelt, B., & Feeney, S. 2018, arXiv:1801.01497 [astro-ph]. http://arxiv.org/abs/1801.01497

Alsing, J., Peiris, H., Leja, J., et al. 2020, The Astrophysical Journal Supplement Series, 249, 5, doi: 10.3847/1538-4365/ab917f

Baldry, I. K., Liske, J., Brown, M. J. I., et al. 2018, Monthly Notices of the Royal Astronomical Society, 474, 3875, doi: 10.1093/mnras/stx3042

Blanton, M. R., Kazin, E., Muna, D., Weaver, B. A., & Price-Whelan, A. 2011, The Astronomical Journal, 142, 31, doi: 10.1088/0004-6256/142/1/31

Brehmer, J., Louppe, G., Pavez, J., & Cranmer, K. 2019, arXiv:1805.12244 [hep-ph, physics:physics, stat]. http://arxiv.org/abs/1805.12244

Bruzual, G., & Charlot, S. 2003, Monthly Notices of the Royal Astronomical Society, 344, 1000, doi: 10.1046/j.1365-8711.2003.06897.x

Cameron, E., & Pettitt, A. N. 2012, Monthly Notices of the Royal Astronomical Society, 425, 44, doi: 10.1111/j.1365-2966.2012.21371.x

Carnall, A. C., Leja, J., Johnson, B. D., et al. 2019, The Astrophysical Journal, 873, 44, doi: 10.3847/1538-4357/ab04a2

Carnall, A. C., McLure, R. J., Dunlop, J. S., & Davé, R. 2018, Monthly Notices of the Royal Astronomical Society, 480, 4379, doi: 10.1093/mnras/sty2169

Chabrier, G. 2003, Publications of the Astronomical Society of the Pacific, 115, 763, doi: 10.1086/376392

Charlot, S., & Fall, S. M. 2000, The Astrophysical Journal, 539, 718, doi: 10.1086/309250

Chevallard, J., & Charlot, S. 2016, Monthly Notices of the Royal Astronomical Society, 462, 1415, doi: 10.1093/mnras/stw1756

Choi, J., Dotter, A., Conroy, C., et al. 2016, The Astrophysical Journal, 823, 102, doi: 10.3847/0004-637X/823/2/102

Cichocki, A., & Phan, A.-H. 2009, IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences, 92, 708, doi: 10.1587/transfun.E92.A.708

Conroy, C. 2013, Annual Review of Astronomy and Astrophysics, 51, 393, doi: 10.1146/annurev-astro-082812-141017

Conroy, C., & Gunn, J. E. 2010, The Astrophysical Journal, 712, 833, doi: 10.1088/0004-637X/712/2/833

Conroy, C., Gunn, J. E., & White, M. 2009, The Astrophysical Journal, 699, 486, doi: 10.1088/0004-637X/699/1/486

Cranmer, K., Brehmer, J., & Louppe, G. 2020, Proceedings of the National Academy of Sciences, 117, 30055, doi: 10.1073/pnas.1912789117

Davis, M., Faber, S. M., Newman, J., et al. 2003, 4834, 161, doi: 10.1117/12.457897

Dax, M., Green, S. R., Gair, J., et al. 2021, arXiv:2106.12594 [astro-ph, physics:gr-qc]. http://arxiv.org/abs/2106.12594

DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv:1611.00036 [astro-ph]. http://arxiv.org/abs/1611.00036

Dhariwal, P., & Nichol, A. 2021, arXiv:2105.05233 [cs, stat]. http://arxiv.org/abs/2105.05233

Dotter, A. 2016, The Astrophysical Journal Supplement Series, 222, 8, doi: 10.3847/0067-0049/222/1/8

Durkan, C., Bekasov, A., Murray, I., & Papamakarios, G. 2019, arXiv:1906.04032 [cs, stat]. http://arxiv.org/abs/1906.04032

Févotte, C., & Idier, J. 2011, arXiv:1010.1763 [cs]. http://arxiv.org/abs/1010.1763

Gardner, J. P., Mather, J. C., Clampin, M., et al. 2006, Space Science Reviews, 123, 485, doi: 10.1007/s11214-006-8315-7

Ge, J., Mao, S., Lu, Y., Cappellari, M., & Yan, R. 2019, Monthly Notices of the Royal Astronomical Society, 485, 1675, doi: 10.1093/mnras/stz418

Genel, S., Vogelsberger, M., Springel, V., et al. 2014, Monthly Notices of the Royal Astronomical Society, 445, 175, doi: 10.1093/mnras/stu1654

Germain, M., Gregor, K., Murray, I., & Larochelle, H. 2015, arXiv:1502.03509 [cs, stat]. http://arxiv.org/abs/1502.03509

Greenberg, D. S., Nonnenmacher, M., & Macke, J. H. 2019, Automatic Posterior Transformation for Likelihood-Free Inference, Tech. rep. https://ui.adsabs.harvard.edu/abs/2019arXiv190507488G

Hahn, C., Beutler, F., Sinha, M., et al. 2019, Monthly Notices of the Royal Astronomical Society, 485, 2956, doi: 10.1093/mnras/stz558

Hahn, C., Vakili, M., Walsh, K., et al. 2017, Monthly Notices of the Royal Astronomical Society, 469, 2791, doi: 10.1093/mnras/stx894

Hahn, C., Kwon, K. J., Tojeiro, R., et al. 2022, The DESI PRObabilistic Value-Added Bright Galaxy Survey (PROVABGS) Mock Challenge, Tech. rep. https://ui.adsabs.harvard.edu/abs/2022arXiv220201809H

Hogg, D. W., Casey, A. R., Ness, M., et al. 2016, arXiv:1601.05413 [astro-ph], doi: 10.3847/1538-4357/833/2/262

Huppenkothen, D., & Bachetti, M. 2021, Accurate X-ray Timing in the Presence of Systematic Biases With Simulation-Based Inference, Tech. rep. https://ui.adsabs.harvard.edu/abs/2021arXiv210403278H

Ivezic, Z., Kahn, S. M., Tyson, J. A., et al. 2019, The Astrophysical Journal, 873, 111, doi: 10.3847/1538-4357/ab042c

Johnson, B. D., Leja, J., Conroy, C., & Speagle, J. S. 2021, The Astrophysical Journal Supplement Series, 254, 22, doi: 10.3847/1538-4365/abef67

Kacprzak, T., Herbel, J., Amara, A., & Réfrégier, A. 2018, Journal of Cosmology and Astro-Particle Physics, 2018, 042, doi: 10.1088/1475-7516/2018/02/042

Karamanis, M., & Beutler, F. 2020, arXiv e-prints, arXiv:2002.06212. https://ui.adsabs.harvard.edu/abs/2020arXiv200206212K

Kingma, D. P., & Ba, J. 2017, arXiv:1412.6980 [cs]. http://arxiv.org/abs/1412.6980

Kriek, M., & Conroy, C. 2013, The Astrophysical Journal Letters, 775, L16, doi: 10.1088/2041-8205/775/1/L16

Lee, D. D., & Seung, H. S. 1999, Nature, 401, 788, doi: 10.1038/44565

Leja, J., Carnall, A. C., Johnson, B. D., Conroy, C., & Speagle, J. S. 2019, ApJ, 876, 3, doi: 10.3847/1538-4357/ab133c

Leja, J., Johnson, B. D., Conroy, C., van Dokkum, P. G., & Byler, N. 2017, The Astrophysical Journal, 837, 170, doi: 10.3847/1538-4357/aa5ffe

Lejeune, T., Cuisinier, F., & Buser, R. 1997, A & A Supplement series, Vol. 125, October II 1997, p.229-246., 125, 229, doi: 10.1051/aas:1997373

—. 1998, Astronomy and Astrophysics Supplement, v.130, p.65-75, 130, 65, doi: 10.1051/aas:1998405

Lilly, S. J., Le Fèvre, O., Renzini, A., et al. 2007, The Astrophysical Journal Supplement Series, 172, 70, doi: 10.1086/516589

Maraston, C. 2005, Monthly Notices of the Royal Astronomical Society, 362, 799, doi: 10.1111/j.1365-2966.2005.09270.x

Nelson, D., Pillepich, A., Genel, S., et al. 2015, Astronomy and Computing, 13, 12, doi: 10.1016/j.ascom.2015.09.003

Papamakarios, G., Pavlakou, T., & Murray, I. 2017, arXiv e-prints, 1705, arXiv:1705.07057. http://adsabs.harvard.edu/abs/2017arXiv170507057P

Paxton, B., Bildsten, L., Dotter, A., et al. 2011, The Astrophysical Journal Supplement Series, 192, 3, doi: 10.1088/0067-0049/192/1/3

Paxton, B., Cantiello, M., Arras, P., et al. 2013, The Astrophysical Journal Supplement Series, 208, 4, doi: 10.1088/0067-0049/208/1/4

Paxton, B., Marchant, P., Schwab, J., et al. 2015, The Astrophysical Journal Supplement Series, 220, 15, doi: 10.1088/0067-0049/220/1/15

Portillo, S. K. N., Parejko, J. K., Vergara, J. R., & Connolly, A. J. 2020, The Astronomical Journal, 160, 45, doi: 10.3847/1538-3881/ab9644

Rosani, G., Pasquali, A., La Barbera, F., Ferreras, I., & Vazdekis, A. 2018, Monthly Notices of the Royal Astronomical Society, 476, 5233, doi: 10.1093/mnras/sty528

Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, The Astrophysical Journal, 500, 525, doi: 10.1086/305772

Scoville, N., Aussel, H., Brusa, M., et al. 2007, The Astrophysical Journal Supplement Series, 172, 1, doi: 10.1086/516585

Sonnenfeld, A., Jaelani, A. T., Chan, J., et al. 2019, Astronomy & Astrophysics, Volume 630, id.A71, <NUMPAGES>19</NUMPAGES> pp., 630, A71, doi: 10.1051/0004-6361/201935743

Spergel, D., Gehrels, N., Baltay, C., et al. 2015, Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report, Tech. rep. https://ui.adsabs.harvard.edu/abs/2015arXiv150303757S

Sánchez-Blázquez, P., Peletier, R. F., Jiménez-Vicente, J., et al. 2006, Monthly Notices of the Royal Astronomical Society, 371, 703, doi: 10.1111/j.1365-2966.2006.10699.x

Tabak, E. G., & Turner, C. V. 2013, Communications on Pure and Applied Mathematics, 66, 145, doi: 10.1002/cpa.21423

Tabak, E. G., & Vanden-Eijnden, E. 2010, Communications in Mathematical Sciences, 8, 217, doi: 10.4310/CMS.2010.v8.n1.a11

Tacchella, S., Conroy, C., Faber, S. M., et al. 2021, arXiv e-prints, 2102, arXiv:2102.12494. http://adsabs.harvard.edu/abs/2021arXiv210212494T

Takada, M., Ellis, R. S., Chiba, M., et al. 2014, Publications of the Astronomical Society of Japan, 66, R1, doi: 10.1093/pasj/pst019

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. 2020, arXiv:1804.06788 [stat]. http://arxiv.org/abs/1804.06788

Tejero-Cantero, A., Boelts, J., Deistler, M., et al. 2020, Journal of Open Source Software, 5, 2505, doi: 10.21105/joss.02505

Thorne, J. E., Robotham, A. S. G., Davies, L. J. M., et al. 2021, Monthly Notices of the Royal Astronomical Society, 505, 540, doi: 10.1093/mnras/stab1294

Treu, T., Auger, M. W., Koopmans, L. V. E., et al. 2010, The Astrophysical Journal, 709, 1195, doi: 10.1088/0004-637X/709/2/1195

Uria, B., Côté, M.-A., Gregor, K., Murray, I., & Larochelle, H. 2016, arXiv:1605.02226 [cs]. http://arxiv.org/abs/1605.02226

van Dokkum, P. G., & Conroy, C. 2010, Nature, 468, 940, doi: 10.1038/nature09578

Vogelsberger, M., Genel, S., Springel, V., et al. 2014, Monthly Notices of the Royal Astronomical Society, 444, 1518, doi: 10.1093/mnras/stu1536

Walcher, J., Groves, B., Budavári, T., & Dale, D. 2011, Astrophysics and Space Science, 331, 1, doi: 10.1007/s10509-010-0458-z

Westera, P., Lejeune, T., Buser, R., Cuisinier, F., & Bruzual, G. 2002, Astronomy and Astrophysics, 381, 524, doi: 10.1051/0004-6361:20011493

Weyant, A., Schafer, C., & Wood-Vasey, W. M. 2013, The Astrophysical Journal, 764, 116, doi: 10.1088/0004-637X/764/2/116

Wong, K. W. K., Contardo, G., & Ho, S. 2020, Physical Review D, 101, 123005, doi: 10.1103/PhysRevD.101.123005

Wu, H., Köhler, J., & Noé, F. 2020, arXiv:2002.06707 [physics, stat]. http://arxiv.org/abs/2002.06707

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, The Astronomical Journal, 120, 1579, doi: 10.1086/301513

Zhang, K., Bloom, J. S., Gaudi, B. S., et al. 2021, doi: 10.3847/1538-3881/abf42e