Draft version January 26, 2022
Typeset using LaTeX **preprint** style in AASTeX63

Accelerated Bayesian SED Modeling using Amortized Neural Posterior Estimation

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

[1]*Department of Astrophysical Sciences, Princeton University, Peyton Hall, 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 and spectra. They require sampling a high dimensional space of SED model parameters and, thus, take $> 10-100$ CPU hours per galaxy. Current models are not computationally feasible for analyzing the *millions* of galaxies that will be observed by upcoming galaxy surveys (*e.g.* DESI, PFS, Rubin, James 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 exploits neural network models to estimate the posterior probability distribution over the full range of observations. It requires no additional model evaluations to estimate the posterior after training. To demonstrate the advantages of an ANPE approach, we present SEDflow, an SED modeling method that uses ANPE with the recent **?** SED model, and apply it to optical photometry. Once trained, deriving the posterior of galaxy properties with SEDflow takes $\sim$1 *second per galaxy*. Futhermore, we validate SEDflow using posteriors derived from standard Markov Chain Monte Carlo (MCMC) sampling and synthetic observations, with known true parameter values. The SEDflow posteriors are in excellent agreement with both the MCMC estimates and the true posteriors. We therefore conclude that using ANPE we derive accurate posteriors $> 10,000\times$ faster than standard SED modeling methods. Lastly, we apply SEDflow to 33,887 galaxies in the NASA-Sloan Atlas and publicly release the posteriors of galaxy properties for every galaxy.

Keywords: galaxies: evolution – galaxies: statistics

1. INTRODUCTION

Over the past few decades, observations from large galaxy surveys such as the Sloan Digital Sky Survey (SDSS; York et al. 2000), DEEP2 (Davis et al. 2003), COSMOS and zCOSMOS (Scoville et al. 2007; Lilly et al. 2007), and GAMA (Baldry et al. 2018) have transformed our understanding of how galaxies form and evolve. The building blocks of our physical insight into galaxies are their physical properties, *e.g.* stellar mass (M_*), star formation rate (SFR), and metallicity (Z), measured

* changhoon.hahn@princeton.edu.com

from these observations. The primary way for measuring these galaxy properties is by analyzing their spectral energy distribution (SED).

All of the physical processes in a galaxy leave an imprint on its SED: *e.g.* its star formation history, chemical enrichment history, dust content. For instance, the SED over the ultraviolet to infrared wavelengths is primarily composed of light from the galaxy's stellar populations. Some of this stellar light is reprocessed by the gas and dust in its interstellar medium. The goal of SED modeling is to extract these detailed physical properties of galaxies that are encoded in observed SEDs. SED modeling involves three key components: the observations, a physical SED model, and a statistical inference framework for deriving physical properties from the comparisons between the observations and SED models.

Current SED models are based on stellar population synthesis (SPS). Broadly speaking, they model the SED of a galaxy as a compsite stellar population constructed based on isocrhones, 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; **?**). Some models also include dust and nebular emissions as well as emissions from active galactic nuclei (AGN) (*e.g.* Johnson et al. 2021). For a comprehensive review on SPS and SED modeling, we refer readers to Walcher et al. (2011) and Conroy (2013).

In this work, we focus on the third component of SED modeling: the statistical framework for comparing SED models to observations and inferring galaxy properties. State-of-the-art SED modeling use a Bayesian parameter inference framework. This approach has a number of key advantages over maximum-likelihood approaches (*e.g.* Cid Fernandes et al. 2005; Tojeiro et al. 2007; Koleva et al. 2008) that were often used in the past. In Bayesian inference, the goal is to estimate $p(\theta \,|\, \boldsymbol{x})$, the probability distribution of parameters θ, in our case galaxy properties, given observation $\boldsymbol{x}$. Inferring $p(\theta \,|\, \boldsymbol{x})$ provides an accurate estimate of parameter uncertainties and any degerancies among them. These uncertainties can then be propagated to for more accurate statistical analyses. The Bayesian framework also enables marginalization over nuisance parameters, which are necessary to model the effects of observational systematics (*e.g.* flux calibration). In principle, informative priors based on previous observations can also be exploited in Bayesian inference to derive more precise constraints on galaxy properties

In practice, current methods use Markov Chain Monte Carlo (MCMC) based sampling techniques to explore and estimate the posterior (Acquaviva et al. 2011; Chevallard & Charlot 2016; Carnall et al. 2017; Leja et al. 2017). MCMC sampling enables more accurate posterior estimation and addresses the curse of dimensionality that restricts grid-based techniques used in the past (Kauffmann et al. 2003; Burgarella et al. 2005; Salim et al. 2007; da Cunha et al. 2008). Grid-based techniques require SED models to be pre-computed over a grid in parameter space. Hence, as the dimensionality of parameter space increases for more sophisticated models, these techniques require exponentially large number of model evaluations. For MCMC, the number of evaluations scales roughly linearly with the number of parameters so they can be used to efficiently sample high dimensional parameter spaces. The Johnson et al. (2021) SED modeling, for instance, uses MCMC to sample a 16-dimensional parameter space.

Despite these advantages, current Bayesian SED modeling methods are *not scalable* as upcoming galaxy surveys will observe an unprecedented number of galaxies. The Dark Energy Spectroscopic Survey (DESI; Collaboration et al. 2016), the Prime Focus Spectrograph (PFS; Takada et al. 2014), Rubin observatory (Ivezić et al. 2019), and Roman Space Telescope (Spergel et al. 2015) will all observe *millions* of galaxy SEDs. Meanwhile, current SED modeling takes 10-100 CPU hours per galaxy (*e.g.* Carnall et al. 2019; Tacchella et al. 2021). With current methods, inferring galaxy properties from rigorous Bayesian SED modeling for upcoming surveys would require *billions* of CPU hours.

The computational challenge is further exacerbated by additional factors. First, more accurate and sophisticated SED models require additional parameters. For instance, modeling any additional physical processes (*e.g.* nebular emission) requires additional parameters. More accurate modeling of observational effects also requires additional parameters. As the parameter space expands, conventional sampling techniques become less efficient. Second, it has recently come to light that the prior, used to evaluate the posterior, can significantly impact the inferred galaxy properties (Carnall et al. 2017; Leja et al. 2017; **?**). This is a consequence of the fact that galaxy properties are not directly SED model parameters but rather derived properties. Hence, deriving robust galaxy properties requires validating the properties using multiple different priors. For MCMC-based SED modeling, posteriors would need to be entirely re-evaluated for each prior. Lastly, current SED modeling make explicit choices for its stellar spectral library and initial mass function, even though there is no consensus. Incorporating the uncertainties and variations in these choices would either require including them in the SED model or exploring multiple different choices. Both dramatically increases the computational costs of current SED modeling.

Rigorous Bayesian inference does not require MCMC sampling of high-dimensional posteriors. Simulation-based inference (SBI; also known as "likelihood-free inference") is a rapidly developing class of inference methods that offers better alternatives for many applications (see Cranmer et al. 2020, and reference 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). One such SBI technique optimal for scalable Bayesian inference is Amortized Neural Posterior Estimation (ANPE).

ANPE utilizes a density estimation approach to SBI. Instead of MCMC sampling the posterior, $p(\theta \,|\, \mathbf{x_i})$, of a single galaxy, i, ANPE uses neural density estimators (NDE) and training data to estimate the $p(\theta \,|\, \mathbf{x})$ distribution over the full space of observables $\mathbf{x}$. NDEs are density models parameterized by neural networks that estimate a density/probability distributions. The training data are precomputed $\{(\theta_i, \mathbf{x'_i}\}$ pairs. For our application, θ is the SED model parameters, sampled from a prior, and $\mathbf{x'_i}$ are mock observables, such as photometry, constructed from the SED model run on θ_i. Once the NDE is trained using this data, generating the posterior for a galaxy only requires plugging in the observation $\mathbf{x_i}$ — this takes $\sim$1 second.

ANPE addresses one of the major drawbacks of MCMC: SED model evaluations used for one galaxy cannot be used for another. This is why MCMC sampling has to be repeated for every galaxy.

In ANPE, models evaluations are only used to construct the training data of the NDE. Although the training set typically requires more evaluations than for a single MCMC posterior, no additional evaluations are necessary after training. Hence, the computational cost is *amoritzed* and only a minuscule fraction of the cost of performing MCMC to analyze a large set of observations. ANPE has recently been applied to a wide range of applications in physics and astronomy with remarkable success (Stein et al. 2020; Wong et al. 2020; Dax et al. 2021; Zhang et al. 2021).

In this work, we apply SBI using ANPE to Bayesian galaxy SED modeling. We demonstrate that with ANPE, we can make Bayesian SED modeling fully scalable for the millions of galaxies that will be observed by upcoming surveys. Furthermore, as further demonstration, we apply ANPE to analyze optical photometry of $\sim$33,000 galaxies in the NASA-Sloan Atlas (NSA; http://www.nsatlas.org/). We begin in Section 2 by describing SBI method using ANPE. We describe the NSA observations in Section 3. Then, we present SEDflow, our SED modeling using ANPE, in Section 4. We validate the accuracy of the posteriors from SEDflow in Section 5 and discuss the impliciation of our results and future steps in Section 5.1. Finally, we summarize and conclude in Section 6.

2. SIMULATION-BASED INFERENCE

The ultimate goal of Bayesian SED modeling, and probabilistic inference more broadly, is to infer the posterior probability distributions of galaxy properties, θ, given observations, $\mathbf{x}_{\rm obs}$ — $p(\theta \,|\, \mathbf{x}_{\rm obs})$. We can evaluate the posterior at a specific θ and $\mathbf{x}$ using Bayes' rule, $p(\theta \,|\, \mathbf{x}_{\rm obs}) \propto p(\theta)\ p(\mathbf{x}_{\rm obs} \,|\, \theta)$. $p(\theta)$ is the prior distribution, which we specify. And $p(\mathbf{x}_{\rm obs} \,|\, \theta)$ is the likelihood, which is *typically* evaluated using a surrogate Gaussian functional form:

$$\ln p(\mathbf{x}_{\rm obs} \,|\, \theta) = -\frac{1}{2}(\mathbf{x}_{\rm obs} - m(\theta))^t \mathbf{C}^{-1} (\mathbf{x}_{\rm obs} - m(\theta)). \tag{1}$$

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

In the standard approach, the full posterior distribution is derived by evaluating the posterior with a sampling technique such as Markov Chain Monte Carlo (MCMC) or nested sampling (*e.g.* Carnall et al. 2017; Leja et al. 2019b; Tacchella et al. 2021). These sampling techniques are essential for the efficient exploration of the relatively higher dimensionality of SED model parameter space. Even advanced techniques, however, are subject to major limitations. For instance, MCMC sampling techniques can struggle to accurately estimate multimodal and degenerate posteriors. Many also require significant hand-tuning by the user. More importantly, despite their efficiency, these techniques require on the order of a *million* SED model evaluations to derive a posterior — this can take $\sim$100 of CPU hours per galaxy. Analyzing the tens of millions of spectra or billions of photometry from upcoming surveys (*e.g.* DESI, Rubin, Roman) with these approaches would thus require *billions of CPU hours*.

Simulation-based inference (SBI; also known as "likelihood-free" inference) offers a more scalable approach to Bayesian SED modeling. At its core, SBI involves any method that uses a forward model of the observed data to directly estimate the posterior — $p(\theta \,|\, \mathbf{x})$, the likelihood — $p(\mathbf{x} \,|\, \theta)$, or the

joint distribution of the parameters and data — $p(\theta, \mathbf{x})$. SBI methods have 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 more broadly in physics (*e.g.* Brehmer et al. 2019; Cranmer et al. 2020)

One simple and pedagogical example of SBI is Approximate Bayesian Computation (ABC; Rubin 1984; Pritchard et al. 1999; Beaumont et al. 2002), which uses a rejection sampling framework to estimate the posterior. First, parameter values are sampled from the prior: $\theta' \sim p(\theta)$. The forward model, F is then run on the sampled θ' to generate simulated data $F(\theta') = \mathbf{x}'$. If the simulated $\mathbf{x}'$ is 'close' to the observed $\mathbf{x}_{\rm obs}$, usually based on a threshold on some distance metric $\rho(\mathbf{x}', \mathbf{x}_{\rm obs}) < \epsilon$, θ' is kept. Otherwise, θ' is rejected. This process is repeated until there are enough samples to estimate the posterior. The estimated posterior from ABC can be written as $p(\theta \,|\, \rho(F(\theta), \mathbf{x}_{\rm obs}) < \epsilon)$. In the case where $\epsilon \to 0$, the conditional statement is equivalent to the condition $F(\theta) = \mathbf{x}_{\rm obs}$; thus, the estimated ABC posterior is *equivalent* to the true posterior: $p(\theta \,|\, \rho(F(\theta), \mathbf{x}_{\rm obs}) < \epsilon \to 0) \equiv p(\theta \,|\, \mathbf{x}_{\rm obs})$.

ABC produces unbiased estimates of the posterior and only requires a forward model of the observed data. It makes no assumptions on the likelihood and, therefore, relaxes the assumptions that go into surrogate likelihood methods. Nevertheless, ABC is based on rejection sampling and thus requires comparable number of model evaluations as standard MCMC sampling based techniques. ABC is only the simplest SBI method; new SBI methods can infer posteriors with much fewer model evaluations. Density estimaiton-based SBI methods (*e.g.* Papamakarios et al. 2017; Alsing et al. 2018; Hahn et al. 2019; Greenberg et al. 2019; Tejero-Cantero et al. 2020), for instance, use model evaluations to fit estimates of $p(\theta \,|\, \mathbf{x}_{\rm obs})$, $p(\mathbf{x} \,|\, \theta$, or $p(\theta, \mathbf{x})$ probability distributions. They can exploit recent advances in neural density estimation (NDE) that increasingly enable high-fidelity density estimation with fewer samples of the distribution. For instance, the NDE in Papamakarios et al. (2017) accurately estimates the $28 \times 28 = 784$-dimensional distribution of the MNIST dataset[1] with only tens of thousands of samples.

2.1. *Amortized Neural Posterior Estimation*

Density estimation SBI provides a critical advantage over MCMC sampling-based inference methods — it enables *amortized inference*. With SED modeling using MCMC sampling, each galaxy requires $>10^5$ model evaluations to accurately estimate $p(\theta \,|\, \mathbf{x}_{\rm obs})$. Moreover, model evaluations for calculating the posterior of one galaxy cannot be used for another, so for large galaxy surveys with $>10^6$ galaxies (*e.g.* Ahumada et al. 2020) this would require >100 billion model evalulations. Upcoming galaxy surveys will observe orders of magnitude more galaxies (*e.g.* DESI, PFS, Rubin, Roman). Bayesian SED modeling with MCMC sampling for these large galaxy surveys is utterly *computationally infeasible*.

On the other hand, if we use density estimation SBI for SED modeling, we do not require a large number of model evaluations for each galaxy. An initial set of model evaluations ($\sim 10^6$) is necessary to train an NDE to accurately estimate $\hat{p}(\theta \,|\, \mathbf{x}_{\rm obs})$ — *i.e.* Neural Posterior Estimation (NPE). Once

[1] http://yann.lecun.com/exdb/mnist/

trained, $\hat{p}(\theta \,|\, \mathbf{x}_{\rm obs})$ is over the full θ and $\mathbf{x}_{\rm obs}$-space so we can sample $\hat{p}(\theta \,|\, \mathbf{x}_{{\rm obs},i})$ for each galaxy with minimal computational cost. Hence, the inference is now amortized and no additional model evaluations are needed to generate the posterior for each galaxy. In total, SED modeling with SBI for $>10^6$ galaxies requires the same number of model evaluations as analyzing tens of galaxies using the MCMC sampling.

Amoritzed inference using density esimation SBI 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). They primarily use a class of NDE called normalizing flows (Tabak & Vanden-Eijnden 2010; Tabak & Turner 2013). Normalizing flow models use an invertible bijective transformation, f, to map a complex target distribution to a simple base distribution, $\pi(z)$, that is fast to evaluate. In the case of NPE, the target distribution is $p(\theta \,|\, \mathbf{x})$ and the $\pi(z)$ is typically a simple multivariate Gaussian, or mixutre of Gaussians.

The transformation $f : z \rightarrow \theta$ must be invertible and have a tractable Jacobian. This is so that we can evaluate the target distribution from $\pi(z)$ using change of variable:

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

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 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 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 *i.e.* 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: $p(x) = \prod_i p(x_i \,|\, x_{1:i-1})$. 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, $p(x \,|\, y) = \prod_i p(x_i \,|\, x_{1:i-1}, y)$, so that the autoregressive model describes a conditional probability $p(x \,|\, y)$. 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 by dropping out connections of a fully-connected autoencoder using weight matricies with binary masks. This ensures that the MADE model is autoregressive and can be efficiently calculated on a GPU. A MAF model is built by stacking multiple MADE models, where each MADE models the random numbers of the next MADE in the stack. MAF 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 (Greenberg et al. 2019; Tejero-Cantero et al. 2020).

3. NASA-SLOAN ATLAS

As a demonstration of its speed and accuracy, we apply SEDFLOW to optical photometry from the NASA-Sloan Atlas[2] (NSA) with some additional quality cuts. 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).

We impose a number of additional quality cuts to the NSA photometry. The SDSS photometric pipeline can struggle to accurately define the center of objects near the edge or at low signal-to-noise. In some cases, the centroiding algorithm will report the position of the peak pixel in a given band as the centroid. These cases are often associated with spurious objects, so we exclude them from our sample. We also exclude objects that have pixels, which were not checked for peaks by the deblender. The SDSS pipeline interpolates over pixels classified as bad (*e.g.* cosmic ray). We exclude 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. We also exclude objects where the interpolated pixels fall within 3 pixels of their center and they contain a cosmic ray that was interpolated over. Lastly, we exclude any 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 imposing these quality cuts, we avoid complications from artificats in the photometry that we do not model. In principle, we can relax the cuts if we were to include observational effects in our model. For additional details on the quality flags, we refer readers to the SDSS documentation[3]. After the quality cuts, we have a total of 33,887 galaxies in our NSA sample.

4. SEDFLOW

The goal of this work is to demonstrate that we can exploit ANPE to to accelerate Bayesian galaxy SED modeling. For our SED model, we use the state-of-the-art PROVABGS SED model from **?**. Although many SED models have been recently used in the literature (*e.g.* BAGPIPES, Carnall et al. 2017; PROSPECTOR, Leja et al. 2017; Johnson et al. 2021), we choose PROVABGS because it will used to analyze >10 million galaxy spectrophotometry that will be measured by the DESI Bright Galaxy Survey (Ruiz-Macias et al. 2021; **?**). Below, we describe the PROVABGS model in further detail. We also present how we construct the training data for the ANPE using PROVABGS and any additional details for training the ANPE.

4.1. *SED Modeling: PROVABGS*

To model SEDs, we use the state-of-the-art stellar population synthesis (SPS) model of the PROVABGS (Hahn *et al.* 2022). With SPS modeling, we model the SED of a galaxy a 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 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.

[2] http://nsatlas.org/
[3] https://www.sdss.org/dr16/algorithms/flags_detail

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, that are derived from performing NMF (Lee & Seung 1999; Cichocki & Phan 2009; Févotte & Idier 2011) on smoothed SFHs of simulated galaxiese of the Illustris cosmological hydrodyanmic simulations (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, assuming that the SFHs of Illustris galaxies resemble the SFHs of real galaxies. To add stochasticity to the SFH, we include a second star burst component that consists of a single stellar population (SSP).

The ZH is similar defined using two NMF bases dervied from Illustris ZHs. Most SPS models assume constant metallicity over time (*e.g.* Carnall et al. 2017; Leja et al. 2019a); however, this assumption can significantly bias inferred galaxy properties (Thorne et al. 2021). Instead by using the NMF prescription, we can flexibly model a range of different ZHs with only two extra parameters. 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 bin and SSPs are evalulated at each time bin using FSPS. The SSPs are summed up to get the unattenuated rest-frame galaxy SED.

Finally, 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 free 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.

4.2. *Training Data*

In this section, we describe how to we construct the training data using our 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 $M_*, t_{\rm burst}, f_{\rm burst}, \gamma_1, \gamma 2, \tau_{\rm BC}, \tau_{\rm ISM}, n_{\rm dust}$ and Dirichlet prior over $\beta_1, \beta_2, \beta_3, \beta_4$. The Dirichlet prior is chosen for the normalization of the NMF SFH while the rest are chosen to span an extensive range of galaxy SEDs.

For each sampled SED parameters, we construct mock observables — *i.e.* NSA optical photometry (Section 3) — We first construct the rest-frame galaxy SED using the SED model, $F(\lambda; \theta')$. Then, for a given redshift, z, we redshift the SED and convolve it with optical broadband filters, R_X to generate noiseless photometric fluxes:

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

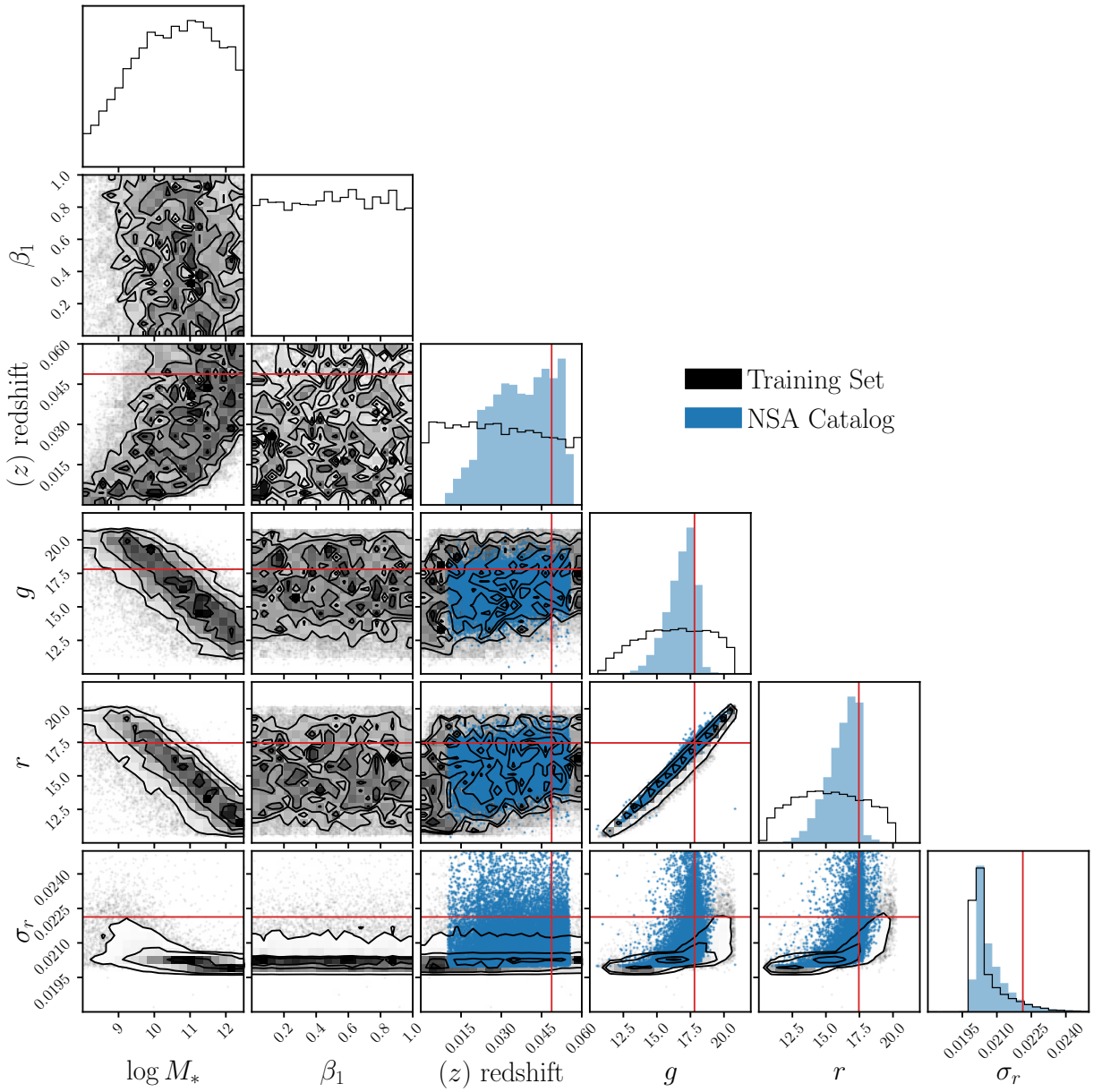

Figure 1. Joint distribution of SED model parameters ($\log M_*$, β_1, redshift) and photometric magnitudes (g, r, z) for our training set. The training set was constructed by sampling parameter values from the prior, constructing SEDs using a theoretical SPS model, and applying our noise model. For details, we refer readers to Section 4. For comparison, we present the distribution of magnitudes for galaxies in the NSA catalog (blue). *The training set fully encompasses the observations, thus, our* SEDFLOW *method can be used to infer the posterior for all NSA galaxies.*

Next, we apply a noise model.

In our noise model, we first assign photometric uncertainties, σ'_X, to the training data by sampling an estimate of $p(\sigma_X|f_X)$. Then, we apply Gaussian noise

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

to derive photometric flux. We emphasize that the accuracy of the posteriors from our ANPE is not particularly sensitive to the noise model. This is because σ_X is included in the conditional statement of our posterior. Hence, we only require that σ'_X of the training data spans the observed σ_X values. We discuss this further in Section 5.1.

Since we do not require a high fidelity estimate of $p(\sigma_X|f_X)$, we use a simple empirical estimate of $p(\sigma_X|f_X)$ based on NSA photometry and uncertainties. For each band, we separately estimate

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

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 bin 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.

In total, we constuct $N_{\rm train} = 1,131,561$ sets of SED parameters, redshift, photometric uncertainty, and mock NSA photometry: $\{(\theta', z, \sigma'_X, \hat{f}_X(\theta')\}$. There are 12 SED parameters and photometry and uncertainties in each of the 5 NSA bands. In Figure 1, we present the joint distribution of select SED model parameters ($\log M_*$, β_1), redshift, photometry (g and r bands), and photometric uncertainty (r band) of our training data (black). We also include the distribution of redshift, photometry, and uncertainty of NSA galaxies (blue). The photometry and uncertainties are in magnitude-space. The distribution of the training data fully spans the distribution of NSA galaxies. Hence, the training data provides support over the full (f_X, σ_X, z) space of the NSA observations and the trained ANPE can accurately estimate posteriors for all NSA galaxies.

4.3. *Training ANPE*

For our ANPE, 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 X free hyperparameters, $\mathbf{h}$. Our goal is to determine $\mathbf{h}$ of the MAF model, $q(\theta\,|\,\mathbf{x}, \mathbf{h})$, so that it accurately estimates the posterior probability distribution $p(\theta\,|\,\mathbf{x})$. Here, θ represent the SED parameters and $\mathbf{x} = (f_X, \sigma_X, z)$. We do this by minimizing the KL divergence between $q(\theta\,|\,\mathbf{x}, \mathbf{h})$ and $p(\theta\,|\,\mathbf{x})$: $D_{\rm KL}(p\,||\,q)$.

In practice, we first split the simulated data from Section 4.2 into a training and validation data set with a 90/10 split. Afterwards, we maximize the total log likelihood $\sum_i \log q(\theta_i\,|\,\mathbf{x}_i)$ on training data $\{(\theta_i, \mathbf{x}_i)\}$. Maximizing the total log likelidhood is equivalent to minimizing $D_{\rm KL}(p\,||\,q)$, which is difficult to directly compute. 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 training epochs. Training our model with a training 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. mention that we're estimating $p(\theta\,|\,f_X, \sigma_X, z)$. TODO

5. RESULTS

Now that we have trained our ANPE, we validate whether the posteriors we infer using it are accurate. As a first test, we compare the posterior from ANPE to the posterior derived from MCMC

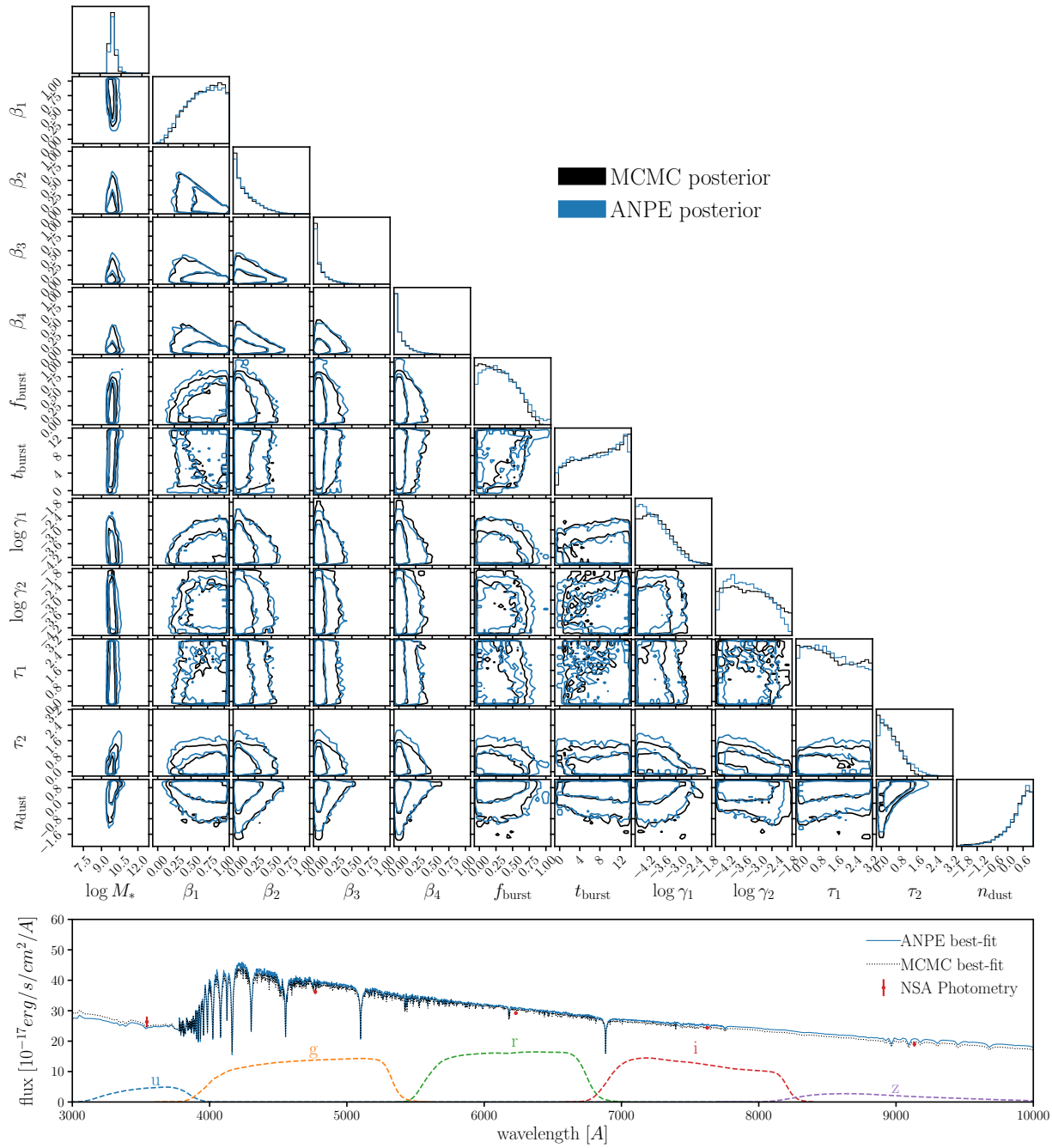

Figure 2. A comparison of the posteriors of the 12 SED model parameters derived from standard MCMC sampling (black) and our ANPE (blue) for a single arbitrarily selected NSA galaxy. The posteriors are in excellent agreement for all of the parameters. Estimating the posterior using MCMC sampling requires X hours. Even using neural emulators to accelerate likelihood evaluations, MCMC sampling requires Y hours. *With ANPE, inferring the full posterior requires 1 second per galaxy.*

sampling for a single arbitrarily chosen NSA galaxy in Figure 2. In the top, we present the the posterior distribution of the 12 SED model parameters (Section 4.1) for the ANPE posterior (blue) and MCMC posterior (black). *The ANPE posterior is in excellent agreement with the MCMC posterior for all of the SED parameters.*

In the bottom of Figure 2, we compare the SED distributions of the best-fit parameter values from the ANPE (blue) and MCMC posteriors (black dotted). We also include the NSA photometric flux of the selected galaxy (red) and mark the optical broadband response curves (dashed). The best-fit SED from the ANPE posterior is also in good agreement with both the MCMC best-fit 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 ∼10 CPU hours for a single MCMC posterior. SED modeling has recently adopted neural emulators to accelerate SED model evaluations. In Hahn *et al.* (2022), for instance, the PROVABGS emulator takes ∼2.9 ms to evaluate, $> 100\times$ 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 ANPE posterior takes* 1 *second — $>10^5\times$ faster than MCMC.*

Besides the selected NSA galaxy in Figure 2, the posteriors from ANPE and MCMC are overall in excellent for NSA galaxies. However, we do not know the true SED parameters for these galaxies so to further validate the ANPE posteriors, we use test synthetic photometry, where we know the truth. We sample 1000 SED parameters from the prior, $\{\theta_i^{\rm test}\} \sim p(\theta)$, and construct synthetic NSA galaxy photometry, $\{\mathbf{x}_i^{\rm test}\}$, for them in the same way as the training data in Section 4.2. Afterwards, we generate posteriors for each of test data using our ANPE: $\{p(\theta \,|\, \mathbf{x}_i^{\rm test})\}$.

In Figure 3, we present the probability-probability (p-p) plot of the ANPE 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 SED parameter. For the true posterior, the percentiles are uniformly distributed so the CDF is a diagonal (black dashed). *Overall, the ANPE posteriors are in good agreement with the true posteriors for each of the SED parameters.*

We also include in Figure 3 the CDFs of the SED parameters for the MCMC posteriors derived for a subset of 100 test galaxy photometry (gray dotted). Based on the CDFs, the ANPE posteriors are actually in better agreement with the true posteriors than those derived from MCMC. This is due to the fact that MCMC posteriors are only estimates of the true posterior and are subject to limitations in initialization, sampling, and convergence. Meanwhile, these limitations do not impact posteriors from ANPE and so the comparison highlights additional advantages of ANPE besides the $10^5\times$ speed up.

We examine another validation of the ANPE 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 fact that the

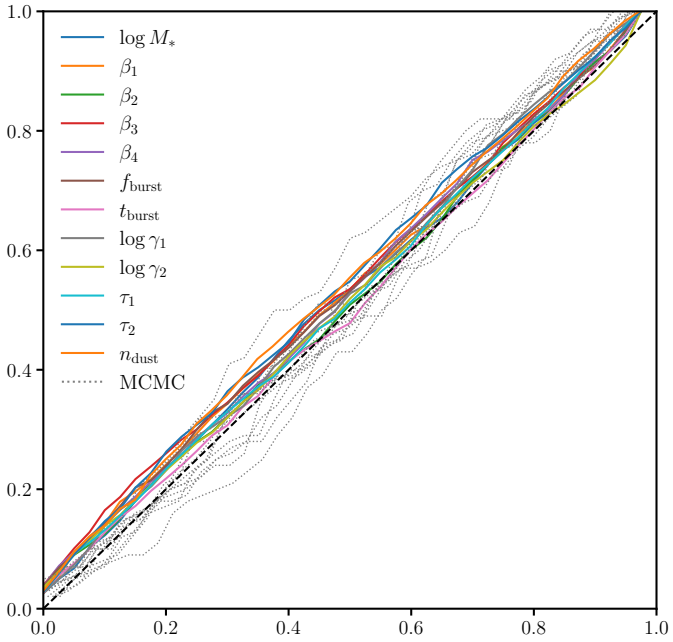

Figure 3. Probability-probability (p-p) plot of the ANPE for 1000 simulated test data. For each SED parameter, we plot the cumulative distribution function (CDF) of the percentile score of the true value within the ANPE marginalized posterior. For the true posteriors, the percentile score is uniformly distributed so the CDF is diagonal (black dashed). The test data is constructed in the same way as the training data (Section 4.2). For reference, we include the p-p plot of the posterior estimated from MCMC sampling (gray). *The ANPE is in good agreement the true posterior.*

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 model parameter for the ANPE posteriors (blue) using the 1000 test data. 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 distribution of rank statistic for the ANPE posteriors are overall in good agreement with a uniform distribution. Hence, the ANPE posteriors are in good 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 ANPE posteriors deviate from the true posteriors. For most of the SED parameters, we find little deviation from uniformity so the ANPE posterior is accurate. However, we find a noticeable deviation for $\log \gamma_2$, the coefficient of the second ZH NMF basis function. The rank statistic distribution of $\log \gamma_2$ has a symmetric U shape where the true parameter values are more often at the lowest and highest ranks. This means that on average the ANPE posterior for $\log \gamma_2$ is narrower than the true posterior and underestimates the uncertainty. To better gauge the actual impact of such a discrepancy, we examine the distributions of the MCMC posterior. Since we only use 100 test data for MCMC the distributions are noisy; however, we find overall significantly large discrepancies from uniformity (see *e.g.* $\log M_*$ and β_4 panels). Based on this comparison, we

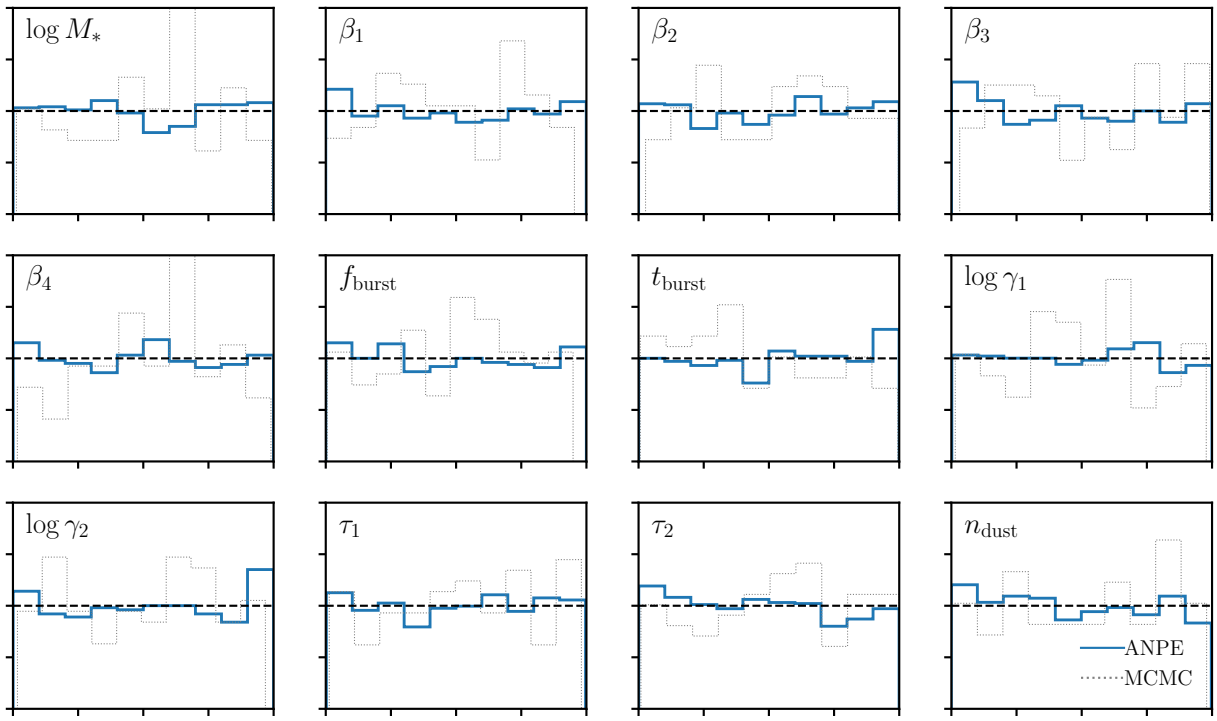

Figure 4. Simulation-based calibration plot of the ANPE for 1000 simulated test data. For each SED parameter, we plot the histogram of the rank statistic of the true value within the marginalized ANPE posterior (blue). For the true posterior, the rank statistics will have a uniform distribution (black dashed). For reference, we include the rank dstirubtion of the MCMC posterior for a subset of 100 test data (gray dotted). *The ANPE is in good agreement the true posterior.*

conclude that our ANPE produces posteriors that are in good agreement with the true posteriors. revisit once we have the final ANPE

With the accuracy of our ANPE validated, we now apply it to derive posteriors for all of our NSA galaxies (Section 3). elaborate on this We make the

TODO

5.1. *Discussion*

With ANPE, we can generate posteriors for SED modeling $>10^5$ faster than conventional MCMC-based methods. The primarily concern for ANPE is the accuracy of the posteriors. One of the key factors that determines the accuracy of the ANPE is the training data. To construct our training data, we use simple noise model with a Gaussian estimate of $p(\sigma_X \mid f_X)$ and treat each photometric band separately, ignoring any covariance (Section 4.2). The actual $p(\sigma_X \mid f_x)$ distrubiton for NSA, as Figure 1 illustrates, is not Gaussian. There are also significant covariances among the photometry in different bands. Despite the shortcomings of our training data, our ANPE is *not* sensitive to the accuracy of our noise model. This is because we include σ_X as a conditional variable in our ANPE (Section 4.3). Hence, as long as the observed $\mathbf{x}_{\rm obs} = (f_X, \sigma_X, z)$ is sufficiently within the $\mathbf{x}$-space support of the training data, the ANPE produces accurate posteriors.

A more accurate noise model would in principle improve the performance of ANPEs since the $\mathbf{x}$-space of the training data will more effectively span the observations. In other words, there will be fewer training data expended in regions of $\mathbf{x}$-space that are not occupied by observations. However, for our application to SED modeling, we do not find significantly better performance when we replace the noise model to a more sophisticated one. Instead, we find that having a sufficient number of training data is a more important factor for the accuracy of the ANPE. For instance, we find significantly lower accuracy in the posteriors when we use less than $<$500,000 training data points.

The ANPE produces accurate posteriors for $\mathbf{x}_{\rm obs}$ within the $\mathbf{x}$ support of the training data. However, some objects in the NSA catalog are outside of the support. For instance, the g, r, σ_r panels of Figure 1 reveal a number of NSA objects (blue) that lie outside of the $g-r$ color distribution of the training data (black). This is not due to deficiencies in the noise model since the $g-\sigma_r$ and $r-\sigma_r$ relations of these NSA objects are within that of the training data. Some of these objects are likely observational artifacts. Even with the quality cuts in Section 3, there are likely objects in our samples with problematic photometry. list a few typical examples of problematic photometry. The SEDs of such artifcats cannot be modeled using an SPS model combined with noise, so they can lie outside of the training $\mathbf{x}$ support.

Alternatively, the training data may provide insufficient support because the SED model we use to generate them does not sufficiently describe the full range of true galaxy SEDs. In this work we use the PROVABGS SED model (Section 4.1 and **?**). This model provides compact and flexible prescriptions for SFH and ZH that can describe a broad range of SFHs and ZHs (**?**). However, the prescriptions were derived from the SFHs and ZHs of simulated galaxies in Illustris, which may not reflect the full range of SFHs and ZHs of ZHs of actual galaxies. Even if the PROVABGS SFH and ZH prescriptions can accurately describe the true SFH and ZH, there are limitations in our understanding of stellar evolution.

There is currently no firm consensus in SPS: *e.g.* stellar evolution, stellar spectral libraries, or the IMF (*e.g.* **?????**). The PROVABGS model uses MIST isochrones, Chabrier (2003) IMF, and the MILES + BaSeL spectral libraries. These choices limit the range of SED that can be produced by the training data. For instance, if galaxies have significant variations in their IMF, assuming a fixed IMF would reduce the $\mathbf{x}$ support of our training data. A more flexible SED model that includes uncertainties in SPS would widen 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). Extending SED models, however, is beyond the scope of this work, which focuses on improving the Bayesian framework for SED modeling. We emphasize that conventional approaches with MCMC are *equally* impacted by the limitations of the SED model.

In Section 5, we assessed the accuracy of the ANPE posteriors using posteriors derived using MCMC and test SEDs, where we know the true parameter values. Both cases demonstrate the sufficient accuracy of ANPE posteriors for this work. However, for applications that require even higher fidelity posteriors, there are further tests. For instance, the χ^2 of the best-fit parameter value from the ANPE posterior can be used to test whether the best-fit SED model whether accurately reproduces the observed SED. This only requires one additional SED model evaluation per galaxy.

For an even more rigorous test of the posteriors, one can construct an Amortized Neural Likelihood Estimator (ANLE) using the same training data and neural density estimator setup. Unlike the ANPE, which estimates $p(\theta \,|\, f_X, \sigma_X, z)$, the ANLE estimates $p(f_X \,|\, \theta, \sigma_X, z)$. We can then further validate the posteriors by assessing whether the observed photometry lies within the ANLE.

In this work, the ANPE estimates a 12-dimensional probability distribution with an 11-dimensional conditional variable space. However, future SED modeling will likely require higher dimensionality. As mention above, SED models currently do not account for uncertainties in stellar evolution, spectral libraries, or the IMF. Including these uncertainties as well as any additional parameters to account for observational systematics (*e.g.* zero-point calibration, see also http://www.nsatlas.org/caveats) will significantly increase the dimensionality of the posterior. Higher dimensionality significantly increases the computational cost of current methods: the number of evaluation scales roughly linearly with the number of dimensions for MCMC. Fortunately, ANPE has already been applied to higher dimensional applications. Dax et al. (2021), for instance, constructed an accurate ANPE for a 15-dimensional model parameter space and 128-dimensional conditional variable space.

We demonstrate that we can exploit ANPE to accelerate SED modeling of galaxy photometry. The next step is to apply ANPE to SED modeling of 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 X spectral elements. Furthermore, in our approach we include the uncertainties as conditional variables and double the curse of dimensionality. TODO 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 10-dimensional latent variable space with little loss of information. With this spectral compression, we can dramatically reduce the dimensionality of the conditional variable space. We will explore SED modeling of galaxy spectrophotometry using ANPE and spectral compression in a following work.

One of the key ingredients in Bayesian inference is the prior. For SED modeling, recent works have demonstrated that model priors play a crucial role in the derived galaxy properties (Carnall et al. 2018; Leja et al. 2019a; **?**). Even "uniformative" uniform priors on SED model parameters can impose undesirable priors on derived galaxy properties such as M_*, SFR, SFH, or ZH. This underscores the importance of carefully selecting priors and validating results using multiple different priors. For SED modeling with MCMC, selecting a different prior requries reevaluating posteriors for every galaxy. This means repeating all the SED model evaluations in the MCMC sampling. For ANPE, the prior is set by how the parameters of the training data are sampled. For a new prior, we can subsample the training data so that its parameters follow a newly selected prior. Afterwards, the ANPE can be re-trained, re-validated on the test data, and re-deployed on the entire dataset. Re-training the ANPE has significant computational cost, however, it requires no where near the Hence, ANPE has the additional advantage that we can vary the prior without additional model evalulations. TODO

extra advantages of faster posteriors reemphasize that we can meet the needs of DESI, PFS, Rubin, JWST, and Roman.

6. SUMMARY

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. Posteriors also enable marginalization over any nuisance parameters. 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 *millions* 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 *billions* of CPU hours. Rigorous inference will soon be the major bottleneck for galaxy studies.

We demonstrate in this work that Amortized Neural Posterior Estimation (ANPE) provides an altnerative *scalable* approach for Bayesian inference in SED modeling. ANPE is a simulation-based inference method that leverages the latest developments in ML. It formulates Bayesian inference into a density estimation problem and uses neural density estimators (NDE) to estimate the posterior over the full space of observations. The NDE is trained using parameter values drawn from the prior and mock observations constructed using them and a forward model. Once trained, a posterior can be derived using the NDE by plugging in 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 **?**. Furthermore, we apply SEDflow to optical photometry from the NASA-Sloan Atlas as demonstration and validation of our ANPE approach. In our analysis we present the key results below.

- We train SEDflow using a data set of $\sim$1 million SED model parameters and synthetic SEDs construted using them and a forward model. The parameters are drawn from a prior and the forward model is based on PROVABGS and a Gaussian photometric noise model. 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 — $5 \times 10^4\times$ faster than current methods.

- Posteriors derived using SEDflow show excellenet agreement with posteriors derived from MCMC sampling. We further validate the accuracy of its 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 X CPU hours. We make the catalog of posteriors publicly available at urlhere. 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. TODO

Our work clearly highlights the advantages of using an ANPE approach to SED modeling. Such scalable methods for Bayesian SED modeling will enable us to analyze the millions of galaxies that will be observed by upcoming experiments. As we discuss in Section 5.1, however, an ANPE approach requires careful construction of the training data. Its accuracy depends on quality of the forward model and limitations in, say, the SED model can impact the accuracy of the ANPE posteriors outside of the training data support. These are equally important considerations and limitations in current SED modeilng methods. In this work, we focus on SED modeling of optical photometry; however, SEDflow can easily be extended to multi-wavelength photometry. In fact, with the latest developments in efficiently dimensionality reduction of spectra (Portillo et al. 2020, , Melchior & Hahn 2022), SEDflow can even be extended to galaxy spectra. We wil explore this in subsequent 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.

APPENDIX

REFERENCES

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

Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, The Astrophysical Journal Supplement Series, 249, 3, doi: 10.3847/1538-4365/ab929e

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]. https://arxiv.org/abs/1801.01497

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

Beaumont, M. A., Zhang, W., & Balding, D. J. 2002, Genetics, 162, 2025

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]. https://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

Burgarella, D., Buat, V., & Iglesias-Páramo, J. 2005, Monthly Notices of the Royal Astronomical Society, 360, 1413, doi: 10.1111/j.1365-2966.2005.09131.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. 2018, arXiv:1811.03635 [astro-ph]. https://arxiv.org/abs/1811.03635

Carnall, A. C., McLure, R. J., Dunlop, J. S., & Davé, R. 2017, arXiv:1712.04452 [astro-ph]. https://arxiv.org/abs/1712.04452

Carnall, A. C., McLure, R. J., Dunlop, J. S., et al. 2019, Monthly Notices of the Royal Astronomical Society, 490, 417, doi: 10.1093/mnras/stz2544

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

Cid Fernandes, R., Mateus, A., Sodré, L., Stasińska, G., & Gomes, J. M. 2005, Monthly Notices of the Royal Astronomical Society, 358, 363, doi: 10.1111/j.1365-2966.2005.08752.x

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

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

da Cunha, E., Charlot, S., & Elbaz, D. 2008, Monthly Notices of the Royal Astronomical Society, 388, 1595, doi: 10.1111/j.1365-2966.2008.13535.x

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]. https://arxiv.org/abs/2106.12594

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]. https://arxiv.org/abs/1906.04032

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

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]. https://arxiv.org/abs/1502.03509

Greenberg, D. S., Nonnenmacher, M., & Macke, J. H. 2019, Automatic Posterior Transformation for Likelihood-Free Inference

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

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

Ivezić, Ž., 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
Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, Monthly Notices of the Royal Astronomical Society, 341, 33, doi: 10.1046/j.1365-8711.2003.06291.x
Kingma, D. P., & Ba, J. 2017, arXiv:1412.6980 [cs]. https://arxiv.org/abs/1412.6980
Koleva, M., Prugniel, P., Ocvirk, P., Le Borgne, D., & Soubiran, C. 2008, Monthly Notices of the Royal Astronomical Society, 385, 1998, doi: 10.1111/j.1365-2966.2008.12908.x
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. 2019a, 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
Leja, J., Speagle, J. S., Johnson, B. D., et al. 2019b, arXiv, arXiv:1910.04168
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
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
Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. 1999, Molecular Biology and Evolution, 16, 1791, doi: 10.1093/oxfordjournals.molbev.a026091
Rubin, D. B. 1984, The Annals of Statistics, 12, 1151
Ruiz-Macias, O., Zarrouk, P., Cole, S., et al. 2021, Monthly Notices of the Royal Astronomical Society, 502, 4328, doi: 10.1093/mnras/stab292
Salim, S., Rich, R. M., Charlot, S., et al. 2007, The Astrophysical Journal Supplement Series, 173, 267, doi: 10.1086/519218
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
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
Spergel, D., Gehrels, N., Baltay, C., et al. 2015, Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report
Stein, G., Seljak, U., & Dai, B. 2020, Unsupervised In-Distribution Anomaly Detection of New Physics through Conditional Density Estimation
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
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]. https://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

Tojeiro, R., Heavens, A. F., Jimenez, R., & Panter, B. 2007, Monthly Notices of the Royal Astronomical Society, 381, 1252, doi: 10.1111/j.1365-2966.2007.12323.x

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

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

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