Draft version December 20, 2021
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 spectroscopic and photometric observations. These methods, however, require sampling a high dimensional space of SED model parameters and, thus, take >100 CPU hours per galaxy. They are not computationally scalable and cannot be feasily used to analyze *millions* of galaxy photometry and spectra from the next-generation galaxy surveys (*e.g.* Rubin, James Webb, and Roman). short description of our SBI method. TODO We combine this SBI-based method with the PROVABGS SED model, which uses non-parameteric star formation and chemical enrichment histories and a flexible dust attenuation model. We then analyze the multi-wavelength photometry of the NASA-Sloan Atlas with our method as well as with the origin PROVABGS SED model using standard Markov Chain Monte Carlo (MCMC) sampling. Our method accurately recovers the posteriors for the SED model parameters and inferred galaxy properties. More importantly, while the standard MCMC sampling took 100 CPU hours, our method took 0.2 secs to infer the full posterior.

1. INTRODUCTION

paragraph on SED modeling TODO

paragraph on SED Modeling with a full Bayesian approach TODO - marginalize over nuisance parameters - accurately capture degeneracies among model parameters - allows informative priors based on prior observations

limitation of Bayesian SED modeling TODO - These methods involve sampling a high dimensional space of SED model parameters. - The dimensionality only increases as SED models become more sophisticated both in terms of stellar population synthesis modeling and accounting for observational effects. - State-of-the-art methods use MCMC or HMC to efficiently sample a high dimensional parameter space. Go through all the latest works and detail their dimensionality and setups. - Despite the use of modern sampling techniques, SED modeling take hundreds of hours per galaxy. Even for ~ 4000 galaxies in the LEGA-C ESO Public Spectroscopic Survey , this required 3.5 million CPU hours Bagpipes.

simulation based inference TODO provide an overview of simulation based inference

2. SIMULATION-BASED INFERENCE

[*] changhoon.hahn@princeton.edu.com

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.* **???**). 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 ~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 ($p(\theta\,|\,\mathbf{x}_{\rm obs})$), 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.* **????????**), and more broadly in physics (*e.g.* **??**)

One simple example of SBI is Approximate Bayesian Computation (ABC; **???**), 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 $\epsilon \rightarrow 0$, 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: $p(\theta\,|\,\rho(F(\theta), \mathbf{x}_{\rm obs}) < \epsilon \rightarrow 0) \equiv p(\theta\,|\,\mathbf{x}_{\rm obs})$.

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. However, despite its advantages, ABC has one major limitation — it is computational inefficient. Rejection sampling means it

New SBI methods, such as density estimation SBI (*e.g.* **?????**),

2.1. *Amortized Neural Posterior Estimation*

sbi (**??**)
? ? ?

- example normalizing flows and MAF

3. SEDFLOW

In this work we apply ANPE to SED modeling of galaxy spectra. tl;dr of intro

3.1. *SED Modeling: PROVABGS*

provabgs set up

3.2. *Training*

paragraph describing the training data

- noise model

description of the ANPE training

- architecture
- validation

4. NASA-SLOAN ATLAS

As a demonstration of its speed and accuracy, we apply SEDflow to optical photometry from the NASA-Sloan Atlas[1] (NSA) with some additional quality cuts. The NSA catalog is a re-reduction of SDSS DR8 (**?**) that includes an improved background subtraction (**?**) and near and far UV photometry from GALEX (). For optical photometry, we use SDSS photometry in the u, g, r, i, and z bands, which are corrected for galactic extinction using **?**. For UV photometry, we use GALEX photometry in the $W1$ and $W2$ bands based on DR6[2]. details about the GALEX force photometry TODO

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

[1] http://nsatlas.org/
[2] http://galex.stsci.edu/GR6/

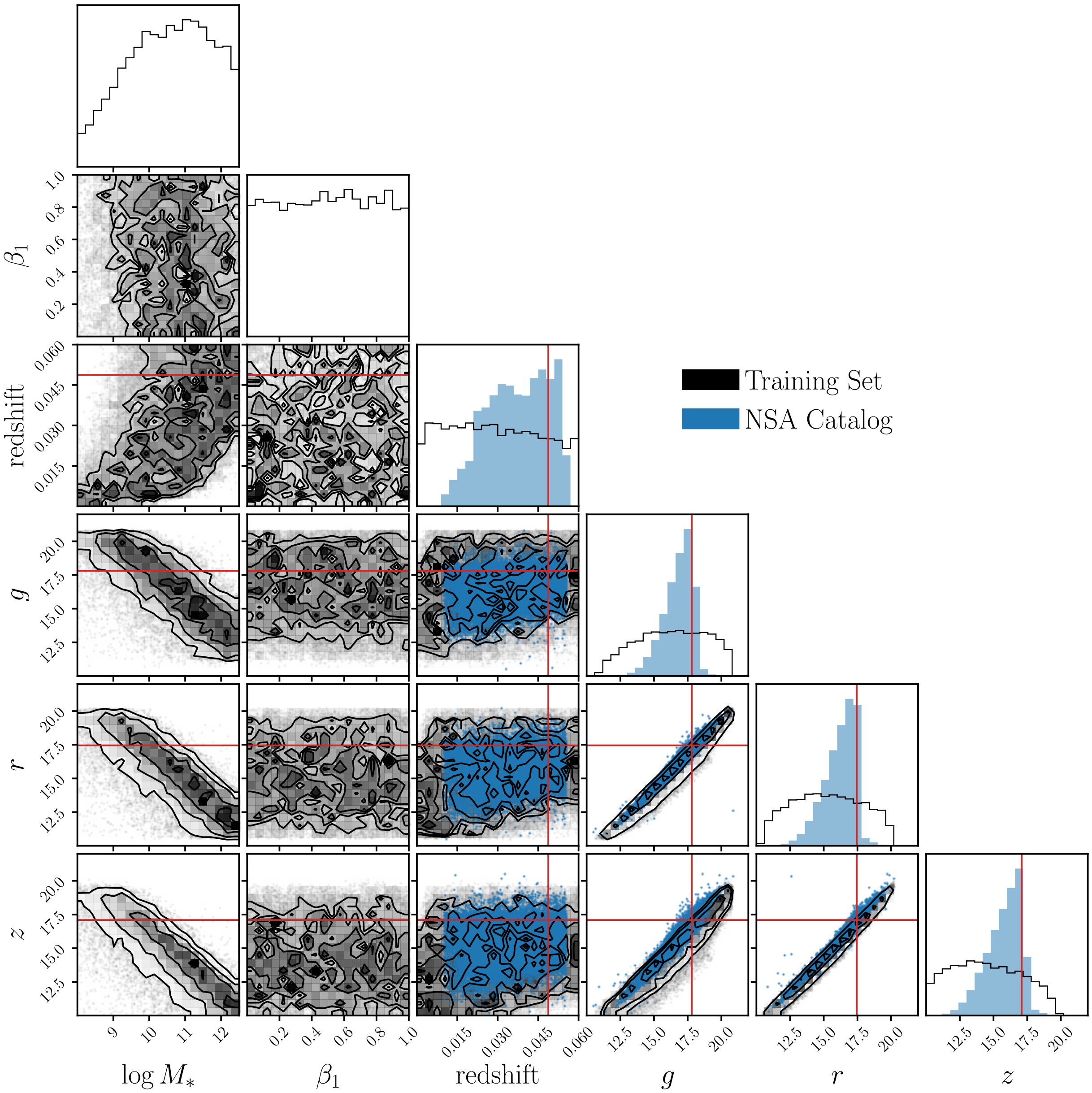

Figure 1. Joint distribution of SED model parameters ($\log M_*$, β_1, redshift) and photometric magnitudes (g, r, z) for our training set. The training set was constructed by sampling parameter values from the prior (Table **??**), constructing SEDs using a theoretical SPS model, and applying our noise model. For details, we refer readers to Section **??**. 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.*

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.

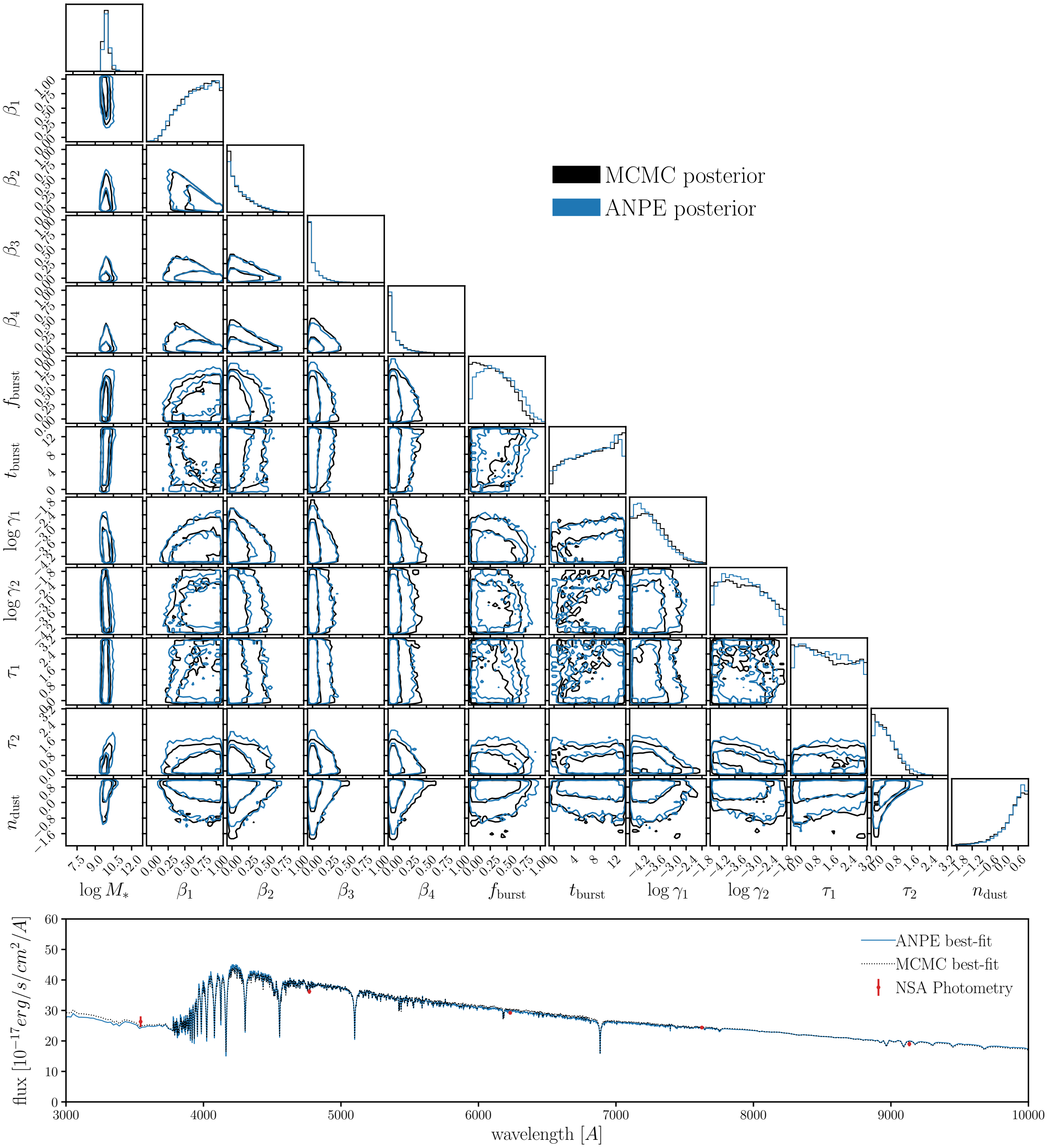

Figure 2. A comparison of the posteriors of the 12 SED model parameters derived from standard MCMC sampling (black) and ANPE (orange) for a randomly selected NSA galaxy. The posteriors are in excellent agreement for all of the parameters. Estimating the posterior using MCMC sampling requires X hours.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

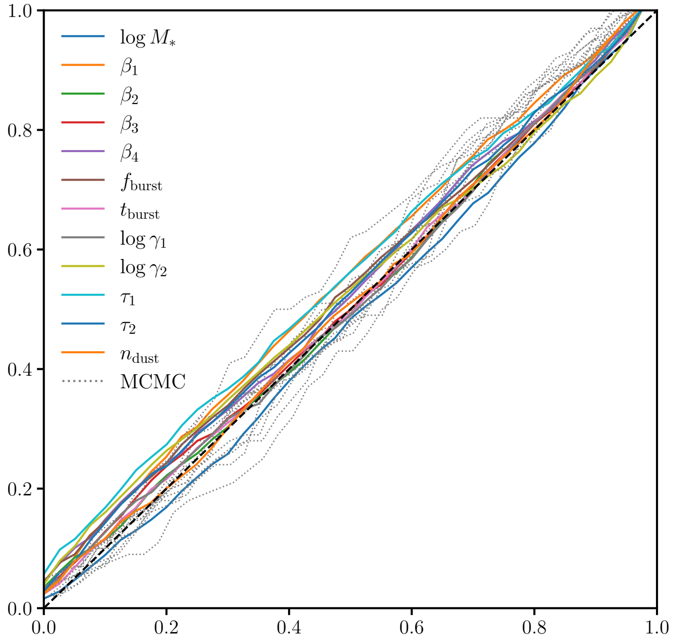

Figure 3. Probability-probability (p-p) plot of the ANPE for 1000 simulated test data. For each SPS parameter, we plot the cumulative distribution function (CDF) of the percentile score of the true value within the ANPE marginalized posterior. For the true posteriors, the percentile score is uniformly distributed so the CDF is diagonal (black dashed). The test data is constructed in the same way as the training data (Section **??**). 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.*

For additional details on the quality flags, we refer readers to the SDSS documentation[3]. After the quality cuts, we have 33,887 galaxies in our NSA sample.

In Figure **??**, we present the distribution of optical and UV magnitudes of the NSA catalog (color).

5. RESULTS

validate the normaling flow SBI posteriors for a single case compare the corner plot of a posterior derived from MCMC with the SBI TODO

validate the derived properties of galaxies for a handful of galaxies TODO

paragraph on the computational advantage TODO

6. SUMMARY

ACKNOWLEDGEMENTS

It's a pleasure to thank Adam Carnall, Kartheik Iyer, Jenny E. Green, Uroš Seljak, Michael A. Strauss, ... for valuable discussions and comments.

APPENDIX

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

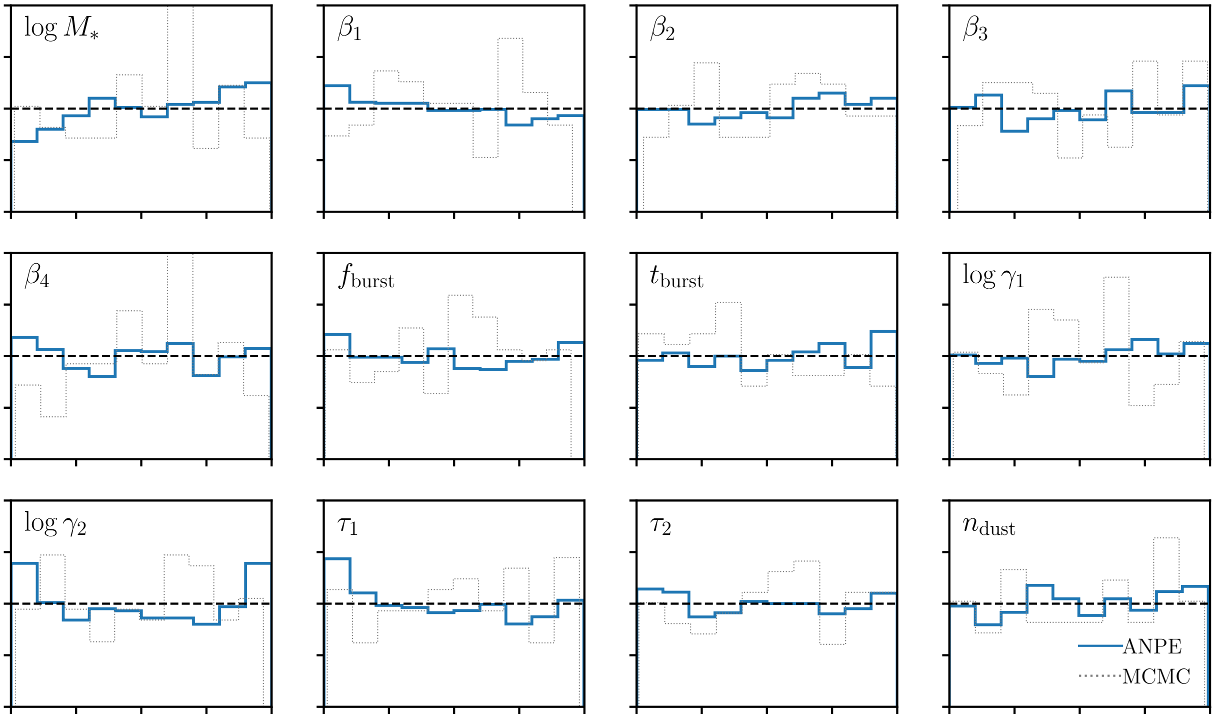

Figure 4. Simulation-based calibration plot of the ANPE for 1000 simulated test data. For each SPS parameter, we plot the cumulative distribution function (CDF) of the percentile score of the true value within the ANPE marginalized posterior. For the true posteriors, the percentile score is uniformly distributed so the CDF is diagonal (black dashed). The test data is constructed in the same way as the training data (Section **??**). 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.*