Reliable AGN Detection and Flux Estimation from Photometry

Initial Progress at KSPA 2019

Mike Walmsley[1], Sotiria Fotopoulou[2]*, Ivana Damjanov[3],
François Lanusse[4], Chris Lintott[1], Nesar Ramachandra[5],
Nic Ross[6], Yuan-Sen Ting[7,8,9]†

[1]Oxford Astrophysics, Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH, UK
[2]Centre for Extragalactic Astronomy, Department of Physics, Durham University, Durham DH1 3LE, UK
[3]Department of Astronomy and Physics, Saint Mary's University, 923 Robie Street, Halifax, NS B3H 3C3, Canada
[4]Berkeley Center for Cosmological Physics and Department of Physics, University of California, Berkeley, CA 94720
[5]High Energy Physics Division, Argonne National Laboratory, Lemont, IL 60439
[6]Institute for Astronomy, University of Edinburgh, Royal Observatory, Edinburgh, EH9 3HJ, United Kingdom
[7]Institute for Advanced Study, Princeton, NJ 08540, USA
[8]Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA
[9]Observatories of the Carnegie Institution of Washington, 813 Santa Barbara Street, Pasadena, CA 91101, USA

Abstract

The observed colours of galaxies are a combination of stellar emission, dust emission and absorption, as well as the contribution from the active galactic nucleus (AGN). Fortunately, the AGN emission has a very distinct shape, compared to the stellar emission, visible even in broad-band photometric measurements. We exploit the impact of AGN in the observed broad-band colors to estimate posteriors for plausible AGN flux and disk inclination.

Spectral energy distribution (SED) fitting codes use a predefined model library to estimate physical parameters such as stellar mass and star-formation rate. In this work, we extended such a software, Prospector, to include also an AGN component in the blue part of the SED originating from the accretion disk. In the era of large datasets, a severe disadvantage of all SED fitting codes is the computing resources required. To sample fast enough for practical use on million galaxy surveys, we emulate a complete galaxy and AGN model using neural networks. Preliminary results show that we recover correctly the expect amount of AGN contribution using the CPz sample of Fotopoulou & Paltani (2018). In our final publication, we aim to show that we can recover AGN flux, inclination and extinction from simulated observations and observations from the XXL Survey and OSSY Database.

Our photometric AGN identification method is crucial to avoid Euclid/LSST weak lensing systematics introduced by poor photo-z estimates. Our method could also measure the fraction of AGN flux per galaxy over cosmic time, helping investigate the impact of AGN feedback on star formation and the galaxy luminosity distribution.

1 Introduction

Photometric estimation of AGN flux is important for understanding how AGN affect galaxy evolution. Unpicking the physics behind how AGN interact with their hosts often relies on measuring correlations between AGN and other galaxy properties (mass, star formation, quenching, merger history, morphology, etc). The task of estimating AGN flux is often replaced with the simpler task of determining whether or not the source flux is stellar-dominated, AGN dominated, or composite (Padovani et al., 2017). These three physical regimes are an approximation of the inherently continuous balance between stellar and AGN flux.

Estimating AGN flux is also crucial for Euclid to accurately measure cosmological parameters with weak lensing. AGN introduce systematics in such measurements because they alter galaxy colours and therefore bias the photo-z distance estimates required for weak lensing (Salvato et al., 2019). At the billion-source scale

*Swiss National Science Foundation Fellow
†NASA Hubble Fellow

of Euclid, systematics are the limiting factor for precision cosmology (Laureijs et al., 2011). Our method identifies galaxies likely to be 'contaminated' with AGN flux, allowing them to be cleaned from the weak lensing sample.

Previous work on estimating AGN flux covers a wide range of methods. One particularly relevant method is the application of MCMC methods to perform Bayesian inference of AGN properties on known hosts. These often use sophisticated high-dimensional models of AGN (Rivera et al., 2016). Sample sizes are typically in the tens to thousands, perhaps due to computational cost; each inference may take several CPU-days. Such methods are popular where plentiful multi-wavelength data is available, often including optical or IR spectra (i.e. not on photometric-only data).

At KSPA 2019, we aimed to apply Bayesian inference to AGN flux measurement (which reduces to AGN detection if desired) using only photometric data. The key advantage of using Bayesian inference over e.g. color cuts (Assef et al., 2017) or supervised learning (Fotopoulou & Paltani, 2018; Nakoneczny et al., 2019) is that we explicitly marginalise over all possibilities. This allows us to make wide but well-calibrated predictions where the data is genuinely inconclusive - which we expect from photometry. We also require our method to be sufficiently fast (i.e. have a low computational cost) to be applied to the largest surveys. To dramatically speed up inference, we emulate our AGN/galaxy model with a neural network.

2 Forward Model of Galaxy + AGN

To perform inference, we need to know what photometric observations we would expect given a set of galaxy parameters. In the literature, the description of galaxy SEDs is performed with two strategies 1) empirical models derived from observed galaxies (see Brown et al. 2014 for a recent example) and 2) stellar populations synthesis models which build realistic galaxy SEDs by assuming an initial stellar mass function, star formation history, and gas/dust component e.g. (Bruzual & Charlot, 2003; Maraston, 2005). The former approach allows for realistic representation of galaxies, while the latter allows for the estimation of physical parameters such as mass and star formation rate. In this work, we use synthetic models as we are interested in infering physical parameters.

2.1 Prospector/FSPS galaxy model

We use the Python package `Prospector` (Leja et al., 2017) as a framework for our method. `Prospector` is a galaxy SED fitting tool designed to create realistic galaxy SEDs, and to sample these SEDs given some observation. The source frame galaxy SEDs are created using `FSPS` (Conroy et al., 2009) (called via `pyFSPS`). Mock observations are made by redshifting the SED and then applying bandpasses for each desired filter (via `sedpy`). Given a user-provided observation and associated uncertainties, Prospector calculates the log-likelihood of each galaxy parameter vector θ as (by default) the log-likelihood of the observation from the mock observation observed under Gaussian noise.

Prospector can estimate posteriors from this log-likelihood and user-defined priors by sampling; either through affine-invariant ensemble MCMC (via `emcee`) or nested sampling (via `dynesty`). Prospector has previously been used to estimate galaxy age and SFR from photometry for the 3D-HST survey (Leja et al., 2019).

We require a model flexible enough to explain real observations, but with few enough free parameters that those parameters could be reasonably constrained by our limited observations. Building on (Leja et al., 2017), we assume the following galaxy model (each entry corresponds to an FSPS argument):

Free Parameters

1. Stellar mass: log-uniform prior $[10^9, 10^{12}]$ $M_\odot$
2. Star formation: delay-τ model, with log-uniform τ prior [0.1, 30]
3. Dust optical depth at 5500Å: 0.6, with uniform prior [0, 2]

Fixed Parameters

1. Initial mass function: Kroupa, following Kroupa (2001)
2. Dust law: Calzetti (Calzetti et al., 2000)
3. Dust emission: using Draine & Li (2007), reference intensity $U_{min} = 1.$ (i.e. MW-like), PAH fraction by mass $\gamma = 4\%$, and $q_{pah} = 0.1\%$ dust in high radiation intensity
4. IGM absorbtion: Madau attention at 1.
5. Metallicity: solar (i.e. `logzsol`=0.)

To make progress in the limited time available during the Kavli Program, we fix the redshift of each galaxy to the spectroscopic redshift. We will extend our model to preserve the redshift as a free parameter and update our results.

2.2 AGN Components

Accurate physical models of AGN SEDs remain an ongoing research challenge - in part due to the extensive variation of observed SEDs. Fortunately, our limited (photometry) data ensures that we need only be sensitive to the most crucial of these variations.

We create an SED model composed of accretion disk and dusty torus components. The normalisation of disk and torus are allowed to vary independently. This allows us to model AGN where the disk dominates the SED, and AGN with a heavily obscured disk but bright torus. Both models are described in the following sections.

2.2.1 Accretion Disk

We use SDSS quasar observations to model the accretion disk. For quasars, the accretion disk is expected to dominate in wavelengths short of 1 micron. We use the median composite radio-quiet quasar reported in Shang et al. (2011) as a template and vary the normalisation factor. As we model the torus independently, we apply an arbitrary power-law damping to the template at wavelengths above 1 micron to exclude any contribution at longer wavengths.

We apply an independent Calzetti extinction law to the disk component to allow for different typical dust optical depths for the (galaxy) stellar and (AGN) disk environments. Following Calzetti et al. (2000), this extinction is given by:

$$f_{reddened} = f_0 \cdot 10^{-0.4 \cdot k \cdot \text{EB-V}} \tag{1}$$

where k is the wavelength.

2.2.2 Dusty Torus

To create our torus SED model[1], we use the simulation `CLUMPY` by Nenkova et al. (2008). AGN tori were previously thought to be composed of homogeneously distributed dust, but failures to model observations have led researchers to view tori as composed of dusty clumps. CLUMPY simulates the rest-frame SED that would be observed from the (re-)emission of such a clumpy torus under a pre-defined geometrical configuration. The main parameters are, the inner radius R_d, set by the dust sublimation temperature T_d, the outer radius R_0, the total number of clouds N_0, the opening angle, σ, and the inclination with respect to the observer i. The clumps are of equal optical depth τ_V and distributed with radial density profile r^{-q} out to $Y = \frac{R_0}{R_d}$ and various possible angular distributions.

Nenkova et al. (2008) provides a grid of SEDs calculated at each possible combination of these parameters. Our 12 photometric bands do not provide sufficient information to constrain all 6 torus parameters (in addition to the galaxy and accretion disk components). To restrict our free parameter space, we assume the physically reasonable values of:

[1]Even though Prospector offers the same AGN torus templates, we decided to introduce them as a separate component, because in the Prospector implementation they are linked to the galaxy flux. Given that we want to have solutions that are 100% dominated by the quasar, we must allow the torus model to contribute independently from the galaxy emission.

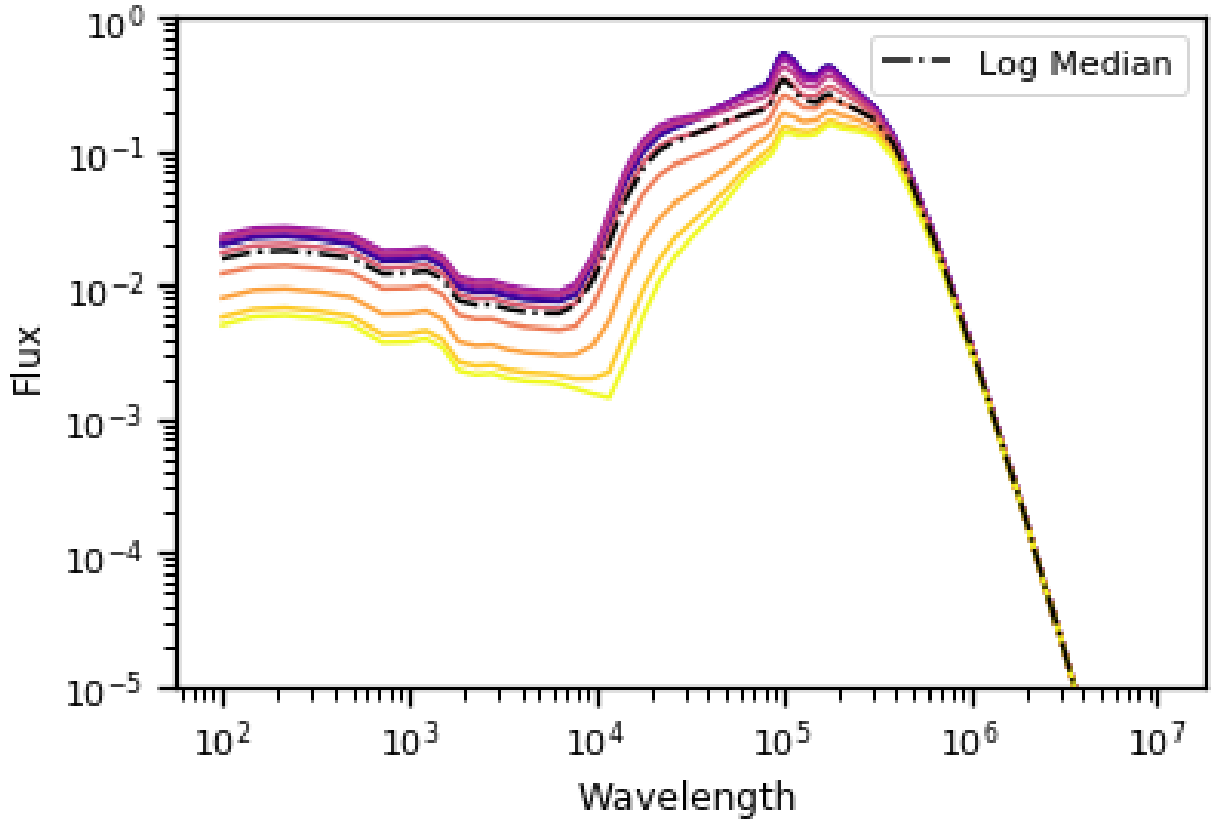

Figure 1: Dusty torus simulations by inclination. Based on simulations by Nenkova et al (2008).

- Opening angle: $\sigma = 30\,\mathrm{deg}$
- Cloud radial distribution: $q = 3$
- Disk size: $Y = 30$
- Number of clouds: $N_0 = 5$

Inclination has the most significant effect on the resulting photometry and so we allow the inclination to vary. Given our fixed parameters above, we interpolate between the varied-inclination SEDs to create an SED as a function of arbitrary inclination $f_{torus}(i)$. For the prototype developed at the Kavli Program, inclination was held fixed at 30 deg. This model was used for the results currently presented in this report.

Mirroring the AGN disk model, we apply an arbitrary power-law damping to the torus SED below 1 micron, to account for any intervening extinction.

3 Sampling Approaches

We have defined a model for the photometry we expect given a galaxy with (potential) AGN disk and torus SED contributions. We would like to be able to calculate posteriors for some observation by sampling this model. We also require this sampling to be fast enough to scale to modern photometric datasets on the order of 10 million galaxies, such as the XXL Survey (Fotopoulou in prep.).

First, we demonstrate that our model is flexible enough to provide plausible fits to real observations. To do this, we use standard MCMC and nested sampling approaches. However, these approaches are too slow

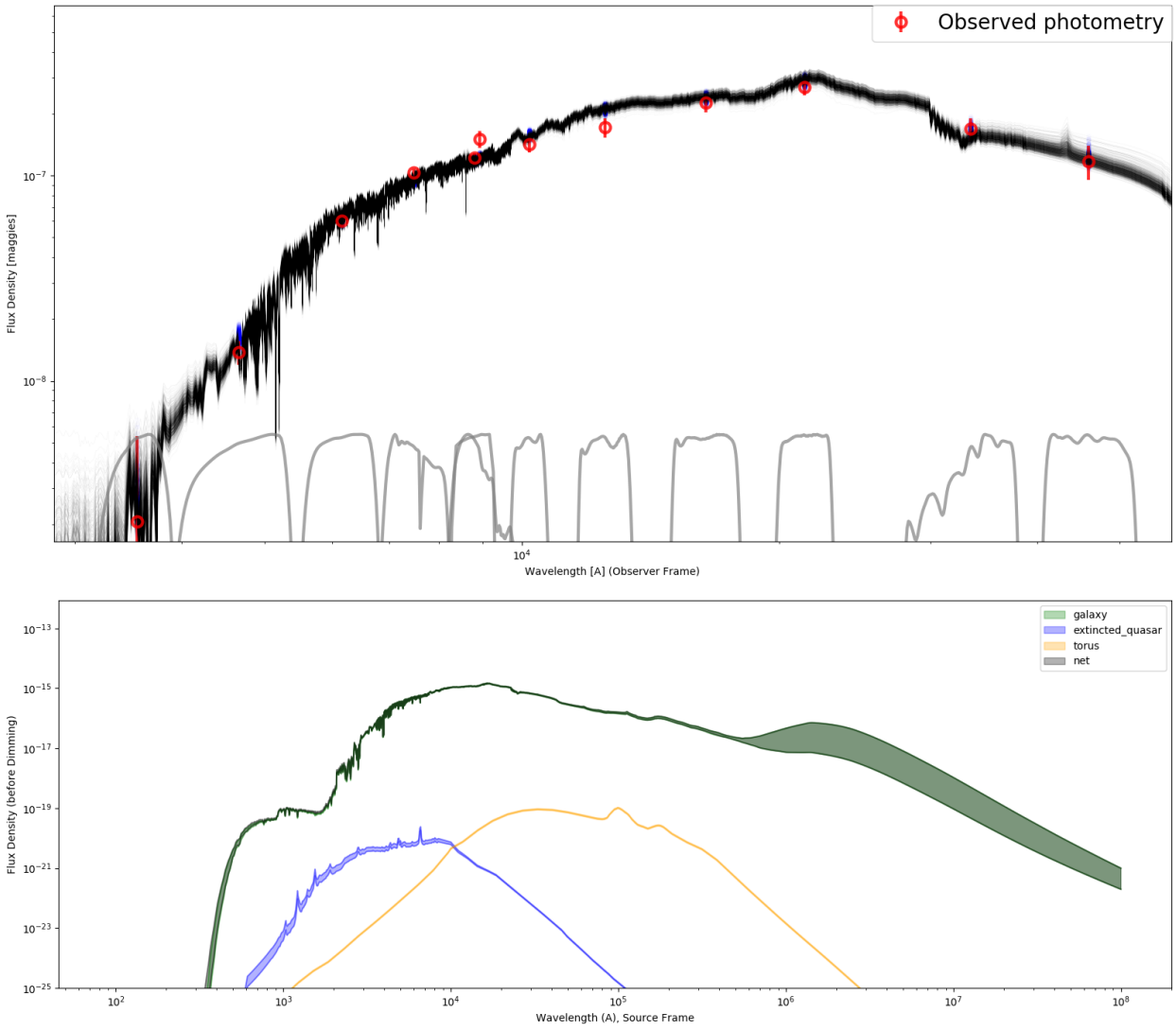

Figure 2: Model fit via nested sampling to spectroscopically-identified starforming galaxy.

for million-galaxy-scale use. To resolve this, we present a neural network emulation method to provide a 100x speedup, enabling us to calculate reliable posteriors at only modest computational cost.

3.1 Baseline Fits Using Nested Sampling or Gradient-Free MCMC

Here, we use `dynesty` (nested sampling) and `emcee` (MCMC) to sample our model. `emcee` implements an affine-transform-invariant version of Metropolis-Hastings MCMC. This MCMC approach is gradient-free; it neither requires nor benefits from any knowledge of the forward model gradients. We make a qualitative inspection of the fits and find empirically that 1) our model successfully reproduces the observations and 2) nested sampling is sometimes able to find parameters which reproduce the observations where MCMC cannot, suggesting that MCMC is not always able to fully explore the posterior in the time available.

We first apply our model to a galaxy spectroscopically identified as starforming without a significant AGN component (Figure 2). Our model is able to produce a fit which reproduces the observations. The posteriors correctly identify the SED contribution of both AGN disk and AGN torus to be negligible (Figure 3).

We next apply our model to a galaxy spectroscopically identified as a quasar, shown in Figure 4. Our

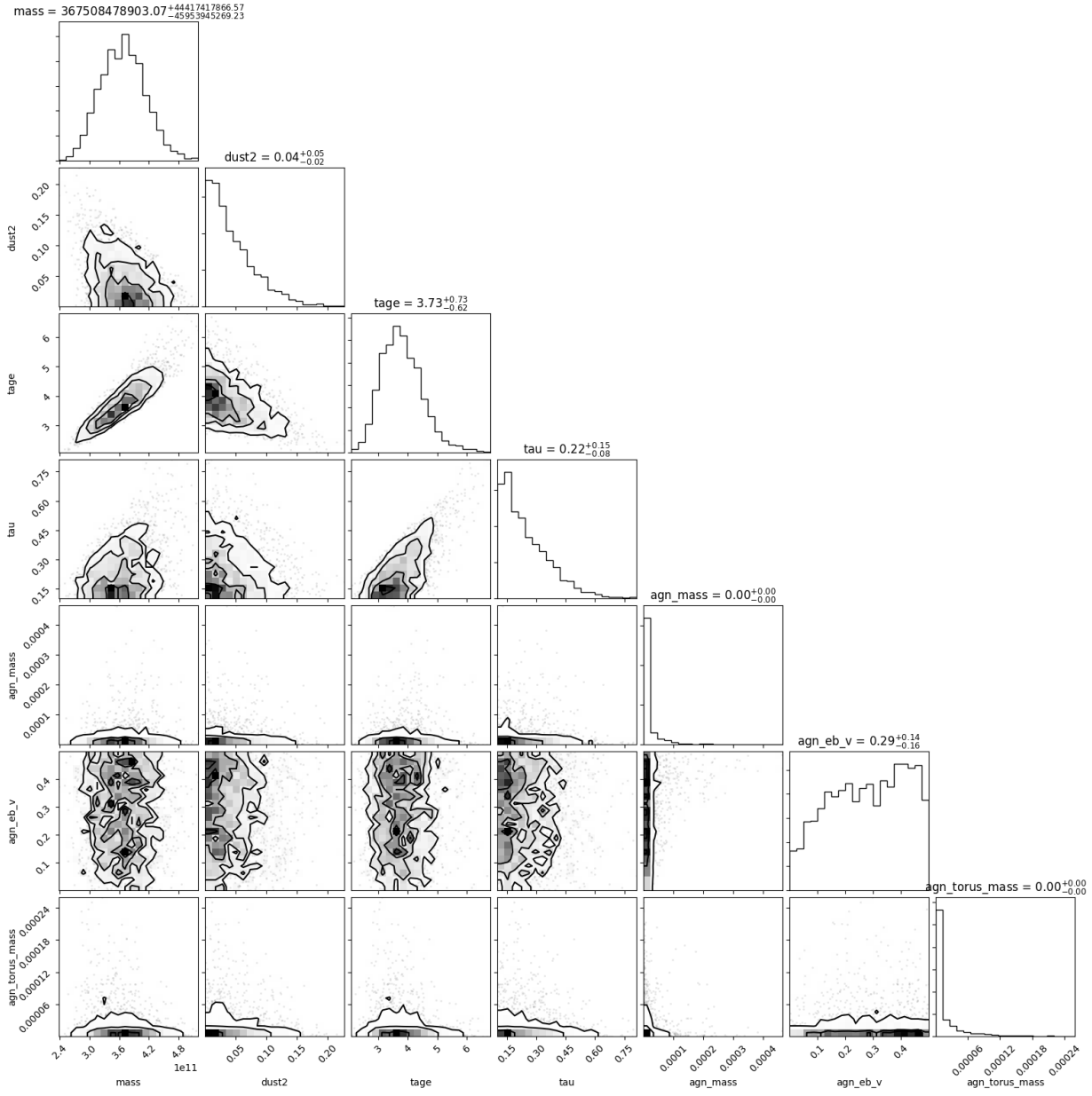

Figure 3: Posteriors of galaxy model parameters of the starforming galaxy of Fig. 2.

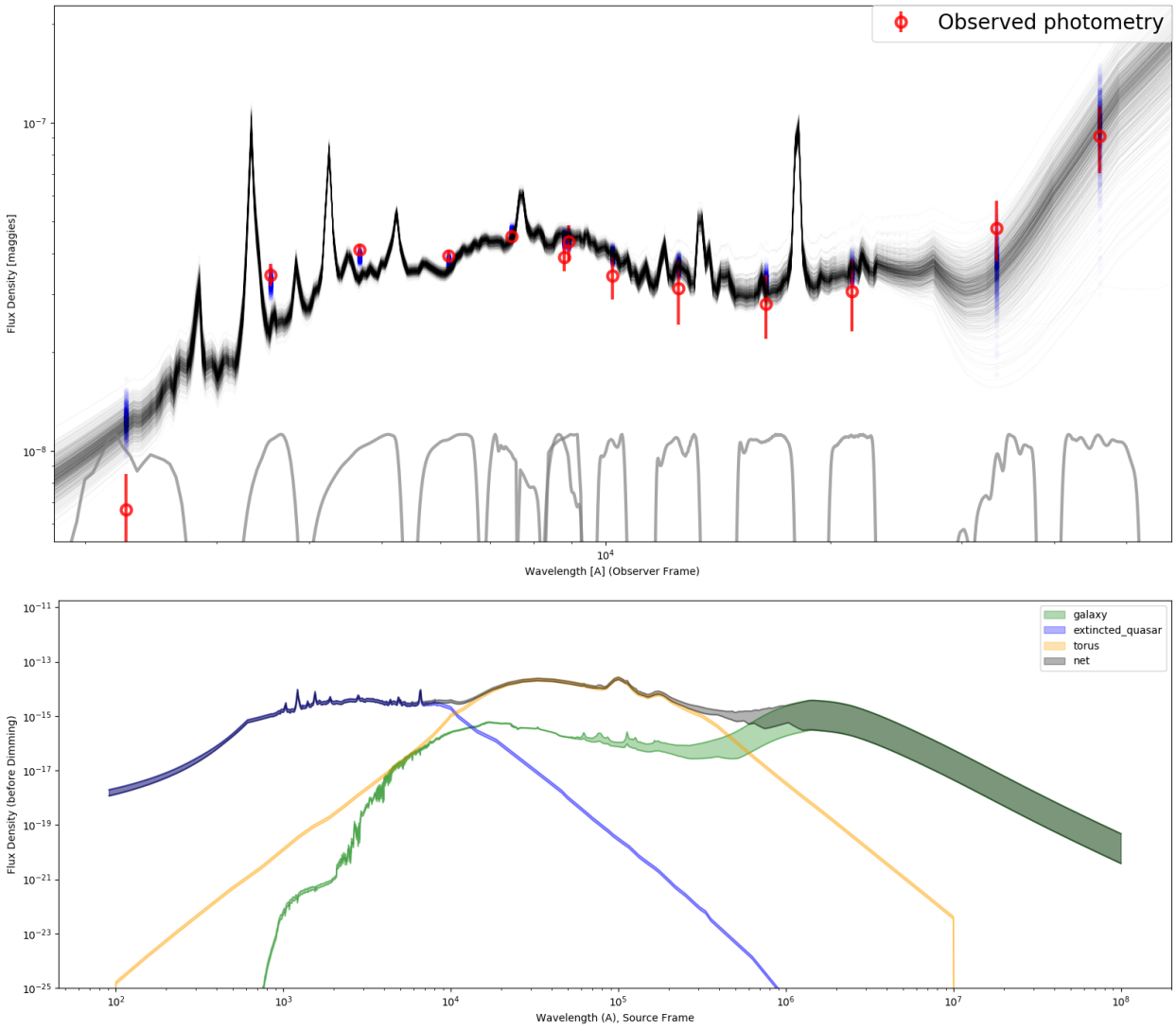

Figure 4: Model fit via nested sampling to spectroscopically-identified quasar.

model is again able to produce a plausible fit. The posteriors correctly identify the SED being AGN-dominated, with significant contributions from both the disk and torus. The galaxy contribution is only non-negligible in the long-wavelength regime, as expected from cold gas emission.

Unfortunately, sampling our model is prohibitively slow for large surveys. Each sample takes 27ms, and hence calculating posteriors for each galaxy takes approximately 20 minutes. 100,000 galaxies would require approx. 14,000 CPU-hours (3.8 CPU-years). In the next section, we use neural network emulation to drastically reduce this.

3.2 Neural Network Emulation and Hamiltonian Monte Carlo

Our model is effectively a mapping between galaxy (and AGN) parameters θ and photometric observations x. Calculating x from θ is the time-consuming part of each forward pass, as the log-likelihood is quick to calculate analytically. We aim to speed up this calculation.

To do so, we will learn the mapping with a neural network, $\tilde{x} = f_w(\theta)$. Because the dimensionality of input and output are relatively low (12 photometric bands and 7 galaxy parameters), a simple dense neural

network is sufficient. To train our network, we create a Latin hypercube of 1 million possible galaxy/AGN model parameters $/theta$. We calculate synthetic observations x for each $/theta$ to construct a training set of 1 million $(/theta, x)$ pairs.

By leveraging the compiled graph approach of TensorFlow, we can evaluate our trained network extremely efficiently. For our simple network, most of the evaluation time for a single call is overhead. We can evaluate multiple chains in parallel by providing the current θ of each chain along the batch dimension. This allows us to generate samples across several hundred chains at almost identical computational cost as for a single chain. Further, because our network is differentiable (unlike the complex original model), we can apply Hamiltonian Monte Carlo to efficiently explore the parameter space.

Figure 5 shows an example of our NN/HMC method recovering the correct parameter values for a simulated galaxy observation, where the 'real' photometry was calculated with the full galaxy model. We achieve a sampling time of 0.17ms per sample, a factor of 100 faster than with the full model.

4 Conclusion

During the Kavli Summer Program in Astrophysics 2019, we explored applying Bayesian inference and neural network emulation to detect and estimate AGN flux from photometry. We extended the galaxy SED fitting package `Prospector` to include independent AGN disk and dusty torus components. In a qualitative investigation using standard sampling approaches (nested sampling and gradient-free MCMC), we find that our new forward model correctly identifies SDSS starforming galaxies as likely AGN-free and SDSS quasars as AGN-dominated. We then train a neural network to emulate our forward model, which both dramatically increases sampling speed (0.17ms vs 27ms per sample) and provides forward model gradients, allowing the use of Hamiltonian MCMC sampling.

We would like to emphasise that *these results are preliminary and not yet suitable for citation.* Our goal has been to establish if our Bayesian approach is possible, and if it can be made sufficiently fast for practical use. Next, we need to quantitatively confirm that our method is reliable. We will do this by verifying that the parameters of synthetic galaxies are correctly recovered, particularly for synthetic galaxies with observations similar to real SDSS galaxies. We also hope to show that we correctly report significant AGN flux for galaxies with broad-line-identified AGN in the XXL or OSSY catalogs.

We are grateful to the organisers and funders of KSPA 2019 for providing the opportunity and encouragement to collaborate on this work. We hope that this method will ultimately allow researchers to remove the AGN systematic from weak lensing cosmological parameter estimation, and to investigate how AGN impact galaxy evolution through a new dimension - the AGN flux fraction.

References

Assef R. J., Stern D., Noirot G., Jun H. D., Cutri R. M., Eisenhardt P. R. M., 2017, http://dx.doi.org/10.3847/1538-4365/aaa00a The Astrophysical Journal Supplement Series, 234, 23

Brown M. J. I., et al., 2014, http://dx.doi.org/10.1088/0067-0049/212/2/18 , https://ui.adsabs.harvard.edu/abs/2014ApJS..212...18B 212, 18

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

Calzetti D., Armus L., Bohlin R. C., Kinney A. L., Koornneef J., StorchiBergmann T., 2000, http://dx.doi.org/10.1086/308692 The Astrophysical Journal, 533, 682

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

Draine B. T., Li A., 2007, http://dx.doi.org/10.1086/511055 The Astrophysical Journal, 657, 810

Fotopoulou S., Paltani S., 2018, http://dx.doi.org/10.1051/0004-6361/201730763 Astronomy & Astrophysics, 619, A14

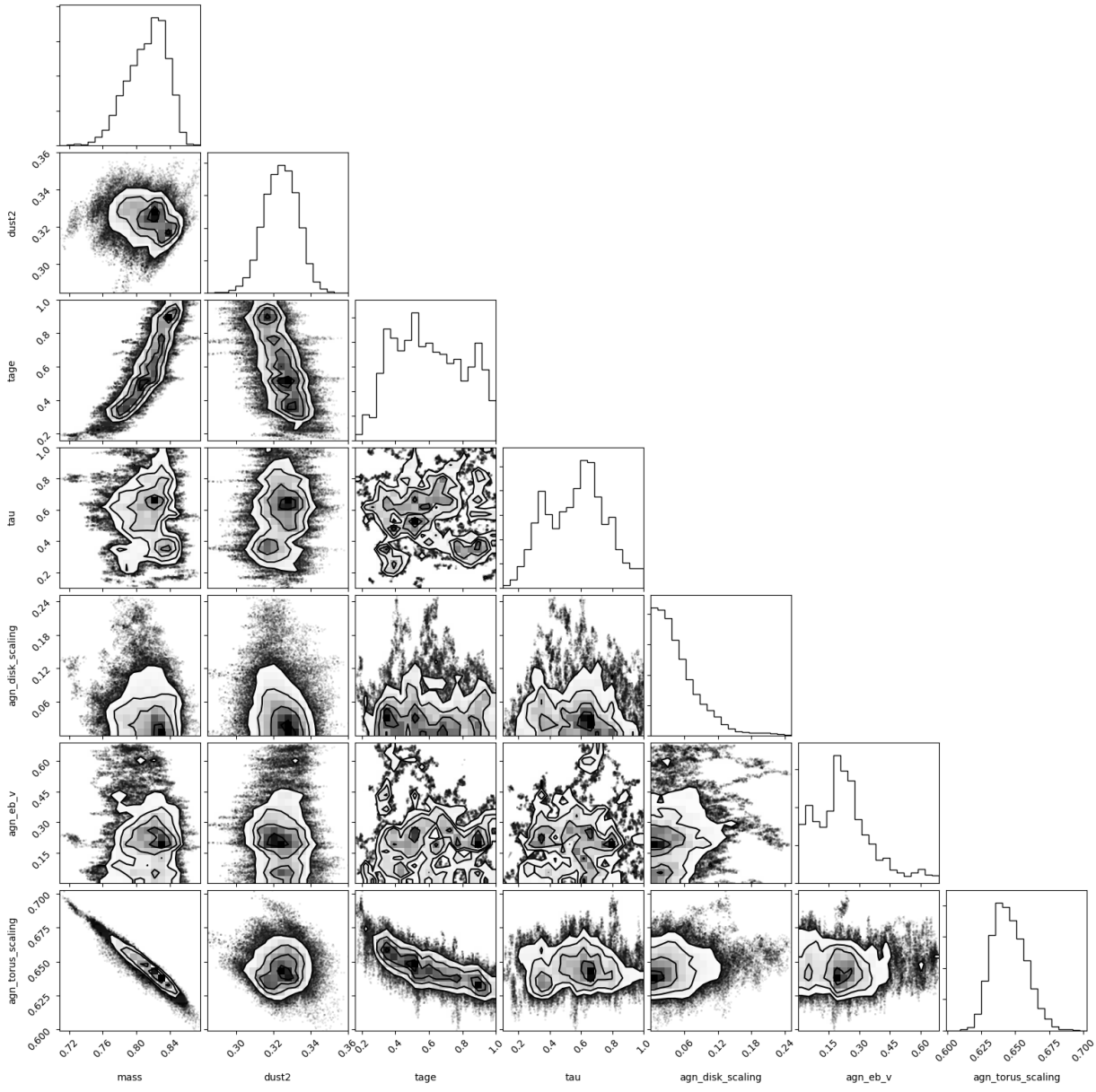

Figure 5: Posteriors of galaxy model parameters recovered from mock photometric observation. Posteriors calculated via HMC sampling, enabled by neural network emulation.

Kroupa P., 2001, http://dx.doi.org/10.1046/j.1365-8711.2001.04022.x Monthly Notices of the Royal Astronomical Society, 322, 231

Laureijs R., et al., 2011, http://dx.doi.org/10.1088/0264-9381/18/14/306 Arxiv preprint

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

Leja J., et al., 2019, http://dx.doi.org/10.3847/1538-4357/ab1d5a The Astrophysical Journal, 877, 140

Maraston C., 2005, http://dx.doi.org/10.1111/j.1365-2966.2005.09270.x , https://ui.adsabs.harvard.edu/abs/2005MNRAS.362..799M 362, 799

Nakoneczny S., Bilicki M., Solarz A., Pollo A., Maddox N., Spiniello C., Brescia M., Napolitano N. R., 2019, http://dx.doi.org/10.1051/0004-6361/201834794 Astronomy & Astrophysics, 624, A13

Nenkova M., Sirocky M. M., Ivezic Z., Elitzur M., 2008, http://dx.doi.org/10.1086/590482 The Astrophysical Journal, Volume 685, Issue 1, pp. 147-159 (2008)., 685, 147

Padovani P., et al., 2017,] 10.1007/s00159-017-0102-9

Rivera G. C., Lusso E., Hennawi J. F., Hogg D. W., 2016,] 10.3847/1538-4357/833/1/98

Salvato M., Ilbert O., Hoyle B., 2019, http://dx.doi.org/10.1038/s41550-018-0478-0 Nature Astronomy, 3, 212

Shang Z., et al., 2011, http://dx.doi.org/10.1088/0067-0049/196/1/2 The Astrophysical Journal Supplement Series, 196, 2