

## Lens Modeling of STRIDES Strongly Lensed Quasars using Neural Posterior Estimation

SYDNEY ERICKSON,<sup>1,2</sup> SEBASTIAN WAGNER-CARENA,<sup>3,4</sup> PHIL MARSHALL,<sup>1,2</sup> MARTIN MILLON,<sup>1</sup> SIMON BIRRER,<sup>5</sup>  
 AARON ROODMAN,<sup>1,2</sup> THOMAS SCHMIDT,<sup>6</sup> TOMMASO TREU,<sup>6</sup> STEFAN SCHULDT,<sup>7,8</sup> ANOWAR SHAJIB,<sup>9,10</sup>  
 PADMA VENKATRAMAN,<sup>1,11</sup> AND THE LSST DARK ENERGY SCIENCE COLLABORATION

<sup>1</sup>*Kavli Institute for Particle Astrophysics and Cosmology, Department of Physics, Stanford University*

<sup>2</sup>*SLAC National Accelerator Laboratory*

<sup>3</sup>*Center for Data Science, New York University*

<sup>4</sup>*Center for Computational Astrophysics, Flatiron Institute*

<sup>5</sup>*Department of Physics and Astronomy, Stony Brook University*

<sup>6</sup>*Department of Physics and Astronomy, University of California, Los Angeles*

<sup>7</sup>*Dipartimento di Fisica, Università degli Studi di Milano*

<sup>8</sup>*NAF – IASF Milano*

<sup>9</sup>*Kavli Institute for Cosmological Physics, University of Chicago*

<sup>10</sup>*Department of Astronomy and Astrophysics, University of Chicago*

<sup>11</sup>*Department of Astronomy, University of Illinois at Urbana-Champaign*

### ABSTRACT

Strongly lensed quasars can be used to constrain cosmological parameters through time-delay cosmography. Models of the lens masses are a necessary component of this analysis. To enable time-delay cosmography from a sample of  $\mathcal{O}(10^3)$  lenses, which will soon become available from surveys like the Rubin Observatory’s Legacy Survey of Space and Time (LSST) and the Euclid Wide Survey, we require fast and standardizable modeling techniques. To address this need, we apply neural posterior estimation (NPE) for modeling galaxy-scale strongly lensed quasars from the Strong Lensing Insights into the Dark Energy Survey (STRIDES) sample. NPE brings two advantages: speed and the ability to implicitly marginalize over nuisance parameters. We extend this method by employing sequential NPE to increase precision of mass model posteriors. We then fold individual lens models into a hierarchical Bayesian inference to recover the population distribution of lens mass parameters, accounting for out-of-distribution shift. After verifying our method using simulated analogs of the STRIDES lens sample, we apply our method to 14 *Hubble Space Telescope* single-filter observations. We find the population mean of the power-law elliptical mass distribution slope,  $\gamma_{\text{lens}}$ , to be  $\mathcal{M}_{\gamma_{\text{lens}}} = 2.13 \pm 0.06$ . Our result represents the first population-level constraint for these systems. This population-level inference from fully automated modeling is an important stepping stone towards cosmological inference with large samples of strongly lensed quasars.

### 1. INTRODUCTION

The expansion rate of the Universe today, known as the Hubble constant,  $H_0$ , is a central point of debate in modern cosmology. To date, there are unresolved discrepancies in the value of  $H_0$  derived from early and late Universe probes (Verde et al. 2019; Di Valentino et al. 2021; Abdalla et al. 2022). In addition to measuring the current expansion rate, we can measure the expansion rate at past times to build understanding of the nature of dark energy. The expansion history of the Universe is sensitive to any possible evolution of dark energy, often characterized by equation of state parameters  $(w_0, w_a)$ . Recent results from the Dark Energy Spectroscopic Instrument collaboration suggest that  $(w_0, w_a)$  may deviate

from standard  $\Lambda$ CDM values (DESI Collaboration et al. 2024). Considering the ongoing Hubble tension and the possible evidence of the evolution of dark energy, it is important to provide independent measurements of  $(H_0, w_0, w_a)$ .

Strong gravitational lenses can be used to provide independent constraints on both  $H_0$  and  $(w_0, w_a)$  through time-delay cosmography (TDC, Refsdal 1964; Treu et al. 2022; Birrer et al. 2024). This technique uses lenses of time-variable point sources, such as quasars, for independent single-step distance measurements, without requiring any local calibrations. With a sufficiently large sample of lenses, strong lensing can independently constrain the Universe’s expansion at percent-level precision (LSST Science Collaboration et al. 2009; Linder 2011;

The LSST Dark Energy Science Collaboration et al. 2021; Hogg 2023).

The current state-of-the-art in TDC measures  $H_0$  from a sample of 7 strongly lensed quasars (Millon et al. 2020; Birrer et al. 2020). Each lens is modelled extensively using explicit likelihood evaluation, which requires many hours of investigator time and CPU resources per lens. Precision on  $H_0$  from the sample varies (2%, 5%, 8%) depending on the mass modeling assumptions. The most conservative choice (Birrer et al. 2020) includes the mass-sheet degeneracy, and analyzes the lens models within a hierarchical framework, allowing population-level characteristics to be jointly inferred with  $H_0$ . Simultaneous hierarchical inference of cosmology and the lens parameter population is a promising approach. To improve constraining power, we must work towards the analysis of a larger sample of lenses.

The Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST) is slated to observe  $\mathcal{O}(10^3)$  multiply-imaged active galactic nuclei (AGNs) (Oguri & Marshall 2010). Many of these lenses can be used for TDC. To scale TDC analysis for LSST lenses, we need to address the high cost of lens modeling. In the time-delay lens modeling challenge, a team applying the current state-of-the-art forward modeling approach took 500,000 CPU hours and 1700 investigator hours to model 48 lenses (Ding et al. 2021). This translates to roughly 10,400 CPU hours per lens, and 35 investigator hours per lens. This technique will not scale to  $\mathcal{O}(10^3)$  time-delay lenses. We, therefore, must invest time in developing fully automated modeling techniques in order to take advantage of data from upcoming large surveys.

One solution is to further automate the existing modeling process. Several efforts have been made to eliminate investigator intervention and reduce compute time, (e.g., Shajib et al. (2019, 2021); Etherington et al. (2022); Schmidt et al. (2023); Ertl et al. (2023); Schuldt et al. (2023b); Tan et al. (2024)). In the analysis by Schmidt et al. (2023) from the Strong Lensing Insights into the Dark Energy Survey (STRIDES) collaboration (hereafter STRIDES23), CPU time was limited to  $<100$  CPU hours per lens, with two lenses in particular (SDSS J0248+1913 and SDSS J1251+2935) taking 11 and 17 CPU hours each. Investigator time was reduced to roughly 10 hours per lens. These techniques are a great improvement in preparing for large samples, but to handle samples with  $\mathcal{O}(10^3)$  lenses, we need to reduce investigator time and CPU compute time even further.

Another solution is to use auto-differentiable lensing code to allow faster inference through gradient informed sampling (e.g., GIGALENS (Gu et al. 2022), HERCULENS (Galan et al. 2022), GRAVITY.JL (Lombardi

2024))). In one application of HERCULENS to a cluster-scale lens, it took 140 minutes to produce a posterior for  $10^5$  parameters (Galan et al. 2024). Even faster times can be achieved on simpler galaxy-scale lenses. In Lombardi (2024), it took under an hour to infer 9 free parameters describing a lensed quasar from 12 observational constraints. These techniques are much faster than traditional CPU-based techniques, but still require explicit choices for nuisance parameters, such as the source light profile and the point spread function.

We investigate a third option, which is the use of convolutional neural networks (CNNs) for fully autonomous lens modeling pipelines, as first proposed by Hezaveh et al. (2017) and Perreault Levasseur et al. (2017). CNN-based techniques have been applied for strong lens modeling in many works (e.g., Pearson et al. (2019); Madireddy et al. (2019); Schuldt et al. (2021); Poh et al. (2022); Schuldt et al. (2023a); Legin et al. (2023); Gentile et al. (2023); Gawade et al. (2024)). This technique is advantageous for its fast compute time as well as its implicit marginalization over nuisance parameters.

In this work, we employ CNNs for neural posterior estimation (NPE), a technique which approximates the inference of lens model posterior probability density functions (PDFs) (Lueckmann et al. 2019). A network is trained to predict parameters describing a posterior PDF given an image of a strong lens. This is achieved by using training examples to minimize the distance between network-predicted posteriors and the true posterior.

NPE has previously been applied for strong lensing modeling. Particularly interesting for TDC analysis is the application of NPE to model simulated *Hubble Space Telescope* (HST) lensed quasars observations for  $H_0$  inference (Park et al. 2021). We aim to build upon this work by introducing the hierarchical framework from Wagner-Carena et al. (2021) to the analysis of lensed quasars and extending the application to real time-delay lenses.

Our goal is to develop a lens modeling pipeline that will enable TDC constraints from many lenses. Assuming this pipeline will treat lenses within a hierarchical framework, a key metric of interest is the recovery of the lens parameter population model. We adopt the hierarchical framework developed in Birrer et al. (2020) and Wagner-Carena et al. (2021). We require the recovery of an unbiased population model, which will eventually serve as a crucial piece of the TDC analysis. To benchmark our progress towards this goal, we use recovery of the population mean  $\mu(\gamma_{\text{lens}})$  as a metric of interest. We choose  $\gamma_{\text{lens}}$ , because as demonstrated in Suyu (2012), error on  $\gamma_{\text{lens}}$  directly translates to error on  $H_0$ .

To date, NPE has only been applied to real observations of galaxy-galaxy lenses. Particularly interesting is the application of NPE on Hyper Suprime-Cam (HSC) galaxy-galaxy lenses in Schuldt et al. (2023b) and Gawade et al. (2024). In Schuldt et al. (2023b), predicted time-delays from the NPE models are directly compared to time delays computed using mass models from a semi-automated traditional modeling approach. Based on this test, the NPE technique was not recommended for application to HSC time-delay lenses. For the first time, we test the NPE technique on a set of real time-delay lenses, specifically those observed by HST and previously modelled in STRIDES23. We also test an extension of NPE called sequential neural posterior estimation (SNPE), which was first applied to strong lens modeling in Wagner-Carena et al. (2024). When applying NPE on real lenses, training examples are sampled from a large prior volume, since the range of the prior must be wide enough to include any possible lensing configuration. This large volume results in low density of training samples, which weakens the constraining power of NPE. As shown in Kolmus et al. (2024), the density of training samples that are similar to the test set is a performance driver for NPE. This effect can be alleviated with SNPE, which uses sequential generation of training examples to show the network more informative lenses.

We aim to answer the following questions:

- Will the application of NPE for strong lens mass modeling produce reliable models on real time-delay lenses? In verification tests, what is the percent error per lens on the PEMD power-law slope  $\gamma_{\text{lens}}$ ?
- How does the application of SNPE compare to NPE? Does the increased sampling density of SNPE improve precision of lens model posteriors?
- What population constraint can we put on the real data using hierarchical Bayesian inference? In verification tests, what is the percent error on the population mean of the PEMD power-law slope  $\mu(\gamma_{\text{lens}})$ ?

To answer these questions, we design a series of tests that we run on two distinct test sets. First, we verify our performance on a test set deliberately offset from the prior in the “shifted” test set. Then, we verify our performance on realistic lensing configurations by creating a test set that closely mimics the data. We call these simulated lenses “doppelgangers”. Finally, we apply our method to the real HST images of STRIDES lenses.

In Section 2, we introduce our statistical framework for handling many lenses and explain our modeling as-

sumptions. In Section 3, we describe our NPE modeling tool. Then, in Section 4, we describe the datasets we employ our method on. In Section 5 we show the results on our verification test sets and the HST data. We end with a discussion of results in Section 6 and conclusions in Section 7.

## 2. BACKGROUND

We assume a parameterized model for strongly lensed quasars, such that each lens can be described by a set of underlying model parameters,  $\xi$ . We also assume a population-level model that describes the statistical distributions for the properties of the individual lenses.

### 2.1. Individual Lens Models

To model a strong gravitational lens, we divide the system into multiple components. We start with the main deflector galaxy. We assign a mass profile and a light profile to this lensing galaxy. Next, we define the background light that sits behind the main deflector along the line of sight. This component includes a source galaxy light profile and a quasar point source profile. Finally, we include a model for the point spread function (PSF) and additional instrumental effects.

#### 2.1.1. Mass Profile

Lens mass is defined as the 2D projected surface mass density of the lensing galaxy, also known as the convergence,  $\kappa(x, y)$ . We assume the lens galaxy mass is described by a power-law elliptical mass distribution (PEMD) profile (Barkana 1998):

$$\kappa(x, y) = \frac{3 - \gamma_{\text{lens}}}{2} \left( \frac{\theta_E}{\sqrt{q_{\text{lens}}x^2 + y^2/q_{\text{lens}}}} \right)^{\gamma_{\text{lens}} - 1}. \quad (1)$$

The power-law slope,  $\gamma_{\text{lens}}$ , controls the rate at which the density decreases as radius increases. The Einstein radius,  $\theta_E$ , describes the total amount of mass in the lens.  $q_{\text{lens}}$  is the axis ratio, with 1 representing a circle. The coordinates  $x$  and  $y$  are rotated by angle  $\phi_{\text{lens}}$  to be aligned along the major and minor axis, respectively. The coordinates are also shifted to align with the central position of the lens mass,  $(x_{\text{lens}}, y_{\text{lens}})$ . We also include an external shear ( $\gamma_{\text{ext}}$ ,  $\phi_{\text{ext}}$ ) in the model to account for additional distortion from the mass of nearby objects. This shear is described by its strength,  $\gamma_{\text{ext}}$ , and its orientation angle,  $\phi_{\text{ext}}$ .

In total, the mass profile can be described with eight parameters:  $(\theta_E, \gamma_{\text{ext}}, \phi_{\text{ext}}, \gamma_{\text{lens}}, q_{\text{lens}}, \phi_{\text{lens}}, x_{\text{lens}}, y_{\text{lens}})$ . Note that we translate from angular coordinates  $(q_{\text{lens}}, \phi_{\text{lens}})$ ,  $(\gamma_{\text{ext}}, \phi_{\text{ext}})$  to rectangular coordinates  $(e_1,$

$e_2$ ,  $(\gamma_1, \gamma_2)$  to avoid the  $\pi$  periodicity in  $\phi_{\text{lens}}$  and  $\phi_{\text{ext}}$ .

$$\begin{aligned} e_1 &= \frac{1 - q_{\text{lens}}}{1 + q_{\text{lens}}} \cos(2\phi_{\text{lens}}) \\ e_2 &= \frac{1 - q_{\text{lens}}}{1 + q_{\text{lens}}} \sin(2\phi_{\text{lens}}) \end{aligned} \quad (2)$$

$$\begin{aligned} \gamma_1 &= \gamma_{\text{ext}} \cos(2\phi_{\text{ext}}) \\ \gamma_2 &= \gamma_{\text{ext}} \sin(2\phi_{\text{ext}}). \end{aligned} \quad (3)$$

### 2.1.2. Light Profiles

We assume a parametric form for the light profile of a galaxy. This profile is used to define light from the source galaxy and the lensing galaxy. We use the elliptical Sérsic profile (Sérsic 1968):

$$I(x, y) = I_* \exp \left[ -k_* \left\{ \left( \frac{\sqrt{q_* x^2 + y^2/q_*}}{R_*} \right)^{1/n_*} - 1 \right\} \right]. \quad (4)$$

The half-light radius,  $R_*$ , sets the size of the source. The Sérsic index,  $n_*$ , determines how concentrated the light profile is. The profile is scaled by surface brightness amplitude  $I_*$ . Constant  $k_*$  depends on  $n_*$ , and ensures that an ellipse with intermediate-axis length  $R_*$  encloses half of the light. The ellipticity is determined by the axis ratio  $q_*$ . The coordinates  $x$  and  $y$  are rotated by orientation angle  $\phi_*$  to be aligned along the major and minor axis, respectively. The coordinates are also shifted to align with the central position of the galaxy's light ( $x_*$ ,  $y_*$ ).

The quasar point source is simply defined by a magnitude and a position. We assume the position of the quasar point source coincides with the center of the source galaxy ( $x_{\text{src}}$ ,  $y_{\text{src}}$ ). We additionally account for microlensing by including a random fractional change to the apparent magnitude of each point source image in the lens plane. The prior for this microlensing factor, along with all prior choices, is listed in Table 7.

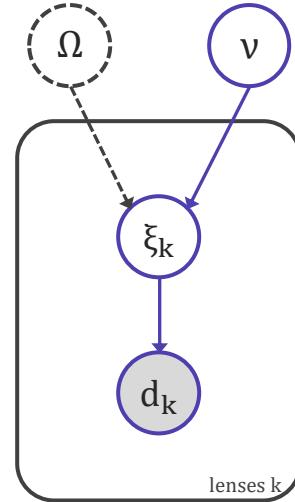
### 2.1.3. Instrumental Effects

During image simulation, we apply a PSF, which smears out light that hits the telescope. This is especially important for point source objects, whose appearance are highly affected by the PSF. We use a library of empirically constructed PSF maps for HST's Wide Field Camera 3 (WFC3) UVIS detector that account for variations in focus and location on the chip (Dauphin et al. 2021)<sup>1</sup>. Our image simulations also include background

noise, using a model that accounts for CCD noise, read noise, sky background, and exposure time.

We also model the dithering effect from HST observations to account for the changes it introduces to the effective PSF and noise profile. When dithering, an observation is split into multiple exposures, where each pointing is slightly offset from the others. Then, images are re-combined using a drizzling algorithm, which accounts for the offsets. To match the dither strategy of the HST observations from STRIDES23, we generate two dither images offset by  $0.5''$  in each direction. We then simulate the drizzling effect using the simulation described in Wagner-Carena et al. (2023), with a final resolution of  $0.04''/\text{pixel}$ , chosen to match the strategy applied in STRIDES23.

## 2.2. Population Model for Many Lenses



**Figure 1.** A PGM showing the hierarchical framework for joint inference of the cosmology ( $\Omega$ ) and lens parameter population model ( $\nu$ ) from  $k$  lenses. The portion of the framework we focus on in this paper is highlighted in solid purple lines. Future inclusion of a cosmological model is shown in dashed grey lines.

We aim to infer a population-level model that describes the statistical distributions for the properties of individual lensed quasars. We are ultimately interested in a joint inference of the cosmological parameters governing the lenses,  $\Omega$ , and the distribution of lens mass properties. Joint recovery of these parameters is important when treating time-delay lenses in a hierarchical framework, as shown in Birrer et al. (2020). In this work, we focus on constraining the hyperparameters,  $\nu$ , that govern the distribution of lens model parameters,  $\xi$ , through the conditional PDF:  $p(\xi|\nu)$  (also referred to as the cPDF). We infer a hyperposterior,  $p(\nu|\{d\})$  from

<sup>1</sup> <https://www.stsci.edu/hst/instrumentation/wfc3/data-analysis/psf>

a dataset of lens observations,  $\{d\}$ , using hierarchical Bayesian inference (HBI). We are especially interested in our ability to recover the hyperparameters that govern the distribution of  $\gamma_{\text{lens}}$ , since recovery of this parameter is especially important for cosmography (Suyu 2012).

To infer  $p(\nu|\{d\})$ , we start from Bayes' theorem:

$$p(\nu|\{d\}) = \frac{p(\{d\}|\nu)p(\nu)}{p(\{d\})}. \quad (5)$$

We infer  $\nu$  from a population of  $k$  lenses. We introduce individual lens parameters  $\xi_k$ , and assume each lens provides an independent constraint. This allows us to break down the problem into constraints on individual lenses.

$$p(\nu|\{d\}) = p(\nu) \prod_k \int \frac{p(d_k|\xi_k)p(\xi_k|\nu)}{p(\{d\})} d\xi_k \quad (6)$$

See Figure 1 for a probabilistic graphical model (PGM) of this inference framework, where the focus of our work is highlighted in purple. Constraining  $p(\nu|\{d\})$  on a set of real time-delay lenses with a fast and automated modeling technique is an important step towards enabling a joint cosmological inference using a large sample of lenses.

### 3. METHOD

We employ NPE to generate approximate lens model posteriors  $q_\phi(\xi_k|d_k, \nu_{\text{int}})$ . For each lens  $k$ , we infer a posterior for PEMD and external shear mass parameters, the lens mass centroid, and the source light centroid:

$$\xi_k = \{\theta_E, \gamma_1, \gamma_2, \gamma_{\text{lens}}, e_1, e_2, x_{\text{lens}}, y_{\text{lens}}, x_{\text{src}}, y_{\text{src}}\}. \quad (7)$$

These parameters are defined in Section 2.1. We additionally attempt to improve the inference of lens model posteriors by applying SNPE. Then, we infer the parameters governing the population-level model  $p(\nu|\{d\})$  from individual lens model posteriors using hierarchical Bayesian inference.

#### 3.1. Neural Posterior Estimation

Given an image of a strong lens, we can infer a posterior for the lens mass parameters using simulation based inference (SBI). We use NPE, which leverages a neural network to approximate the posterior. For a schematic of this method, see the top row of Figure 2. We employ the xResNet-34 architecture (He et al. 2016, 2018). The model learns to approximate inference of lens models from many pairs of simulated images and their underlying lens model parameters  $(d_k, \xi_k)$ .

We generate 500,000 examples  $(d_k, \xi_k)$  for the network to learn from. The data  $d_k$  is an  $80 \times 80$  pixel image, simulated to match HST WFC3 observations, as

described in Section 2.1.3. We define the distribution from which the underlying lensing parameters,  $\xi_k$ , are sampled in Table 7. This distribution can be thought of as an interim prior,  $\nu_{\text{int}}$ . The choice of  $\nu_{\text{int}}$  must be made carefully. The range of  $\nu_{\text{int}}$  is the domain in which the neural network can interpolate. When employing NPE on real data, the range of  $\nu_{\text{int}}$  must be wide enough to ensure support for any possible lensing configuration. Additionally, the distribution  $\nu_{\text{int}}$  will determine the posterior the network attempts to approximate, particularly if the input data is uninformative. Our choice for  $\nu_{\text{int}}$  is described further in Appendix A. We generate training samples using the PALTAS<sup>2</sup> package (Wagner-Carena et al. 2023), which uses LENSTRONOMY<sup>3</sup>, a multi-purpose strong lensing software program (Birrer & Amara 2018; Birrer et al. 2021).

During training, samples from  $\nu_{\text{int}}$  are used to optimize the weights of the network's architecture,  $\phi$ . These weights store a mapping from an input image to a conditional density estimator  $q_\phi(\xi_k|d_k, \nu_{\text{int}})$ . This estimator approximates the lens model posterior  $p(\xi_k|d_k, \nu_{\text{int}})$ . Note that both the conditional density estimator and the posterior are conditioned on the choice of the interim training prior,  $\nu_{\text{int}}$ . In this work,  $q_\phi$  is a 10-dimensional diagonal Gaussian. The assumption of a diagonal covariance matrix is discussed in Section 6.6. To describe  $q_\phi$ , we need a 10-dimensional  $\mu_k$  and 10-dimensional  $\sigma_k$ . This requires the network to have a final fully connected layer with 20 outputs.

To optimize network weights for this task, we minimize the loss function:

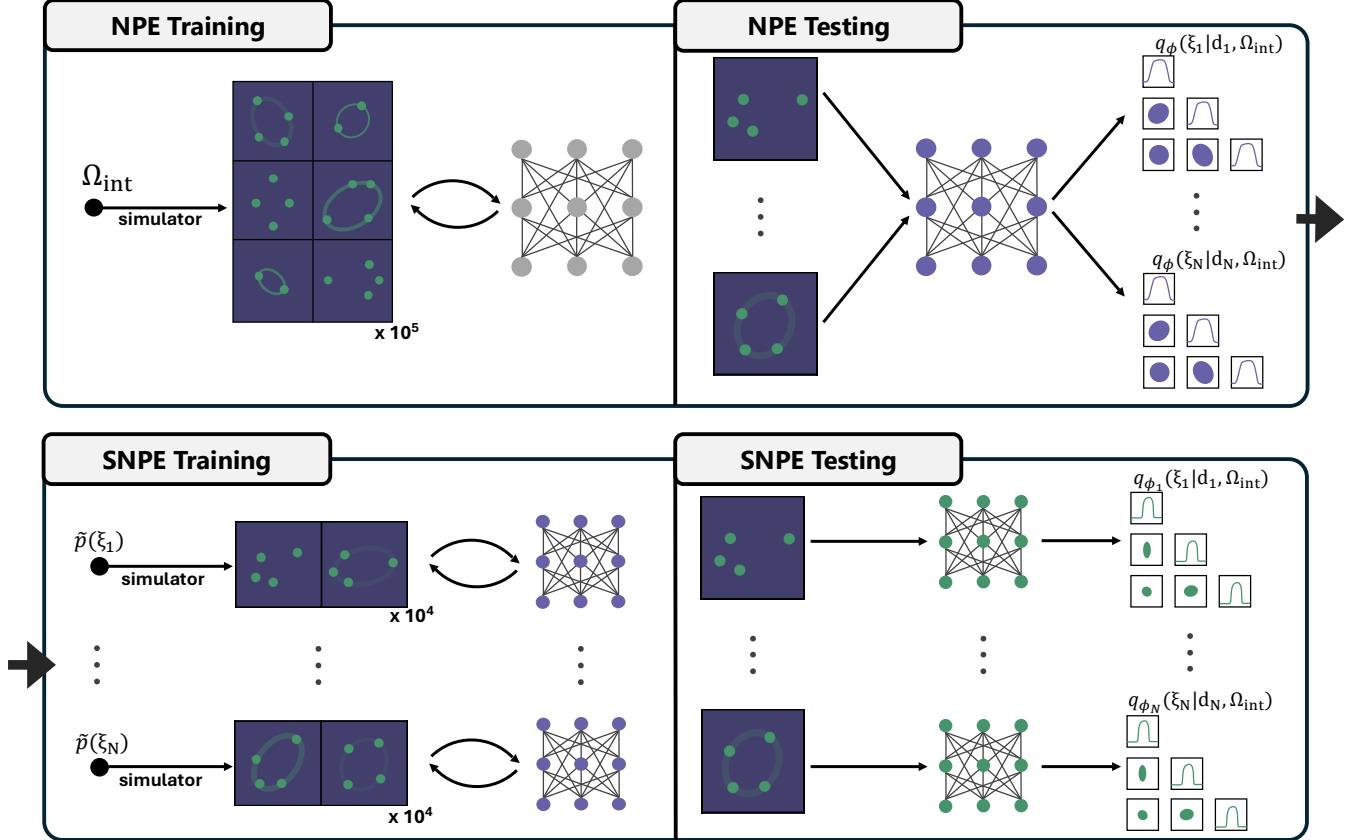
$$L(\phi) = - \sum_{k=1}^N \log [q_\phi(\xi_{k,\text{truth}}|d_k, \nu_{\text{int}})]. \quad (8)$$

With a sufficiently flexible functional form for  $q_\phi$  and a large number of training examples,  $N \rightarrow \infty$ , the approximate posterior converges to the true posterior  $q_\phi(\xi_k|d_k, \nu_{\text{int}}) \rightarrow p(\xi_k|d_k, \nu_{\text{int}})$  when the global minimum of the loss is reached (Papamakarios & Murray 2016). For all aspects of the NPE method, we use the PALTAS package.

During training, we minimize the loss function in Equation 8 using the ADAM optimizer (Kingma & Ba 2014). As an additional augmentation, we randomly rotate the training images each time they are shown to the network. We use 512 images per batch, an initial learning rate of  $5 \times 10^{-4}$ , and an exponential decay schedule for the learning rate. To determine when to stop training

<sup>2</sup> <https://github.com/swagnercarena/paltas>

<sup>3</sup> <https://github.com/lenstronomy/lenstronomy>



**Figure 2.** Diagram of the neural posterior estimation technique. The first round of training is shown in the top row and described in Section 3.1. 5e5 lenses are sampled from the interim prior,  $\nu_{\text{int}}$ . Then, a neural network is trained on those examples. At test time, the network takes in an image of a lens, and outputs the parameters describing an approximate lens model posterior  $q_\phi(\xi_k | d_k, \nu_{\text{int}})$ . The second round of training is sequential, shown in the bottom row and described in Section 3.2. In this step, 5e4 sequential training examples are generated for each lens in the test set. These new lenses are sampled from a proposal distribution  $\tilde{p}(\xi_k)$  that is informed by the NPE posterior  $q_\phi(\xi_k | d_k, \nu_{\text{int}})$ . Then, a copy of the neural network is trained for each lens on the sequential training examples. At test time, each lens is passed through its copy of the neural network to produce parameters describing an approximate lens model posterior  $q_{\phi_k}(\xi_k | d_k, \nu_{\text{int}})$

ing, we evaluate the loss on a held-out validation set at every epoch. The validation set consists of 5,000 lenses sampled from  $\nu_{\text{int}}$ . Validation set loss is used as an early stopping criterion. Training is deemed complete when validation loss has not decreased for 10 epochs. With these settings, we reach convergence after 68 epochs, which takes 5.7 hours on an NVIDIA GeForce RTX 2080 Ti GPU. Using the weights from the last epoch, we pass test images through the network to produce approximate posteriors for each lens,  $q_\phi(\xi_k | d_k, \nu_{\text{int}})$ .

### 3.2. Sequential Neural Posterior Estimation

We investigated multiple ways to further improve the performance of the NPE method. Starting with the simplest investigations, we did not find that longer training or more training examples ( $5 \times 10^6$ ) improved performance. We ultimately hypothesize that the density of training examples near a test example is the most important performance driver, as discussed in Kolmus et al.

(2024). While simply using more training examples does improve the density of samples, our parameter space is high dimensional, which means we may need many orders of magnitude more training examples from  $\nu_{\text{int}}$  to achieve the density of samples required for significant improvement. To achieve higher density of samples in a more efficient way, we explore the application of SNPE (Papamakarios & Murray 2016). Succinctly, during NPE training, a majority of time might be spent on examples that are relatively uninformative for the lenses we are interested in. Rather than increase our training sample size, SNPE improves performance by directly modifying our training distribution.

In SNPE, each object in the test set receives a customized training set that is used for an additional round of training. The sequential training examples are generated by a proposal distribution,  $\tilde{p}(\xi_k)$ , that is informed

by the current NPE posterior. For a schematic of this method, see the bottom row of Figure 2.

We employ one sequential step for our SNPE. A proposal distribution  $\tilde{p}(\xi_k)$  is used to generate new samples in a region of interest. Then, these new samples are used to continue training. We use the proposal:

$$\tilde{p}(\xi_k) \propto (q_\phi(\xi_k|d_k, \nu_{\text{int}})^n p(\xi_k|\nu_{\text{int}})^m)^{\frac{1}{m+n}}. \quad (9)$$

This proposal allows for a trade-off between exploiting the best guess from the current model and retaining the ability to explore the full parameter space. Larger values of  $n$  pull the proposal towards the NPE posterior, while larger values of  $m$  hedge the proposal by favoring the original prior. We investigate the choice of  $m$  and  $n$  in Appendix C. For the remainder of the paper, when we refer to the SNPE technique, we are using a proposal with  $n = 1$  and  $m = 2$ .

With new training examples in hand, we continue optimization of the network weights,  $\phi$ . However, we must modify the loss function to account for the change in training distribution. We employ the loss function derived in Greenberg et al. (2019):

$$L(\phi) = - \sum_k \log q_\phi(\xi_{k,\text{truth}}|d_k, \nu_{\text{int}}) \frac{\tilde{p}(\xi_k)}{p(\xi_k|\nu_{\text{int}})} \frac{1}{Z(d_k, \phi)}. \quad (10)$$

With this modified loss, we maintain guaranteed convergence of the approximate posterior to the true posterior (see Greenberg et al. 2019 for details). Note  $Z(d_k, \phi)$  is a normalization constant:

$$Z(d_k, \phi) = \int q_\phi(\xi_k|d_k, \nu_{\text{int}}) \frac{\tilde{p}(\xi_k)}{p(\xi_k|\nu_{\text{int}})} d\xi_k. \quad (11)$$

When  $q_\phi$ ,  $\tilde{p}$ , and  $p$  are all Gaussian,  $Z(d_k, \phi)$  can be computed analytically.

For each lens in the test set, we generate a copy of the network with randomized weights. Then, 50,000 training examples from proposal  $\tilde{p}(\xi_k)$  are used to optimize the weights of this new network by minimizing Equation 10. All training specifications are kept the same as the NPE run, with the learning rate schedule picking up where it left off. Each SNPE training job is run for 10 epochs, which takes roughly 5 minutes on an NVIDIA GeForce RTX 2080 Ti GPU. Using the weights from the 10th epoch, we pass test images through the network to produce approximate posteriors  $q_{\phi_k}(\xi_k|d_k, \nu_{\text{int}})$ .

### 3.3. Hierarchical Bayesian Inference

After producing mass models for each lens, we are interested in recovering the population-level properties of the lenses. We assume lensing parameters follow a

distribution  $p(\xi|\nu)$ . We constrain  $p(\xi|\nu)$  by inferring the values of  $\nu$ . This results in a hyperposterior,  $p(\nu|\{d\})$ .

We are interested in describing the population-level distribution of 6 lensing parameters:  $(\theta_E, \gamma_1, \gamma_2, \gamma_{\text{lens}}, e_1, e_2)$ . We assume the population-level distributions for these parameters are Gaussian without any covariance. With this assumption we describe the population distribution with 12 hyperparameters: 6 Gaussian means ( $\mathcal{M}$ ) and 6 standard deviations ( $\Sigma$ ). Note we use  $(\mathcal{M}, \Sigma)$  to distinguish the Gaussian hyperparameters from the Gaussian individual posterior parameters.

We assume there is no directional preference of the orientation angles of the ellipticity and shear profiles. This equates to enforcing:  $\mathcal{M}_{e_1}, \mathcal{M}_{e_2}, \mathcal{M}_{\gamma_1}, \mathcal{M}_{\gamma_2} = 0$ . We also assume there is no directional preference in the strength of the ellipticity and shear. This equates to enforcing:  $\Sigma_{e_1, e_1} = \Sigma_{e_2, e_2}$ , and  $\Sigma_{\gamma_1, \gamma_1} = \Sigma_{\gamma_2, \gamma_2}$ . Enforcing these assumptions reduces  $\nu$  to 6 unique parameters:

$$\nu = \{\mathcal{M}_{\theta_E}, \mathcal{M}_{\gamma_{\text{lens}}}, \Sigma_{\theta_E, \theta_E}, \Sigma_{\gamma_{1/2}, \gamma_{1/2}}, \Sigma_{\gamma_{\text{lens}}, \gamma_{\text{lens}}}, \Sigma_{e_{1/2}, e_{1/2}}\}. \quad (12)$$

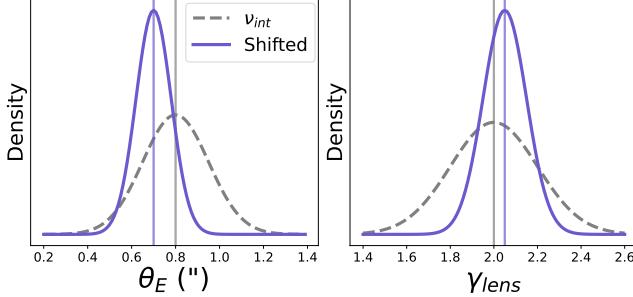
By combining information from individual lens mass models, we can infer the hyperparameters  $\nu$  using hierarchical inference, as described in Section 2.2. Note that we infer 10 mass model parameters, but we only include 6 of those parameters in our hierarchical model. Since the individual posteriors are Gaussian, it is trivial to marginalize over the excluded parameters ( $x_{\text{lens}}, y_{\text{lens}}, x_{\text{src}}, y_{\text{src}}$ ). We start from Equation 6. Then, using Equation C5 from Wagner-Carena et al. (2021), we introduce the neural posterior estimate,  $q_\phi(\xi_k|d_k, \nu_{\text{int}})$ , to the hierarchical inference:

$$p(\nu|\{d\}) \propto p(\nu) \prod_k \int \frac{q_\phi(\xi_k|d_k, \nu_{\text{int}}) p(\xi_k|\nu)}{p(\xi_k|\nu_{\text{int}})} d\xi_k. \quad (13)$$

Note that because we explicitly account for the influence of the interim training prior,  $\nu_{\text{int}}$ , there is a resulting re-weighting term, with  $p(\xi_k|\nu)$  on the numerator, and  $p(\xi_k|\nu_{\text{int}})$  on the denominator. This is how the hierarchical inference can account for distribution shift between the training prior and the test set. Since we have assumed Gaussian forms for the neural posterior estimate  $q_\phi(\xi_k|d_k, \nu_{\text{int}})$ , the training prior  $p(\xi_k|\nu_{\text{int}})$ , and the hyperparameter model  $p(\xi_k|\nu)$ , the integral has an analytic solution.

We use the posterior in Equation 13 along with a flat prior to infer  $\nu$  using the MCMC ensemble sampler in EMCEE<sup>4</sup> (Foreman-Mackey et al. 2013). The up-

<sup>4</sup> <https://github.com/dfm/emcee>



**Figure 3.** Distribution of the shifted test set (solid purple) compared to the training distribution  $\nu_{\text{int}}$  (dashed grey). The shifted test set is designed to test our ability to recover from distribution shift between training set and the test set. The Gaussian distribution of  $\theta_E$  is shifted from  $\mathcal{N}(\mu=0.8, \sigma=0.15)$  to  $\mathcal{N}(\mu=0.7, \sigma=0.08)$ . The Gaussian distribution of  $\gamma_{\text{lens}}$  is shifted from  $\mathcal{N}(\mu=2.0, \sigma=0.2)$  to  $\mathcal{N}(\mu=2.05, \sigma=0.1)$ .

per bounds of the prior for population widths ( $\Sigma_{\theta_E, \theta_E}$ ,  $\Sigma_{\gamma_{1/2}, \gamma_{1/2}}$ ,  $\Sigma_{\gamma_{\text{lens}}, \gamma_{\text{lens}}}$ ,  $\Sigma_{e_{1/2}, e_{1/2}}$ ) are set to the width of the interim training prior for that parameter (see Table 7). This is an assumption that the test set distribution is contained within the range of the training prior.

#### 4. DATA

Our primary goal is to apply our method to real data for the first time. To prepare for this application, we first verify the ability of our method to recover valid posteriors. We construct two simulated test sets, the shifted set (Section 4.1) and the doppelganger set (Section 4.2), for this aim. After verification, our second goal is to assess the robustness of our technique by applying the method to real data. For this aim, we use a set of lensed quasars observed by HST, described in Section 4.3. As we move from the shifted set, to the doppelganger set, to the real HST data, the datasets become increasingly complex and therefore more challenging for our methodology.

##### 4.1. Shifted Test Set

The shifted test set is the first of two simulated verification tests. The goal of the shifted test set is to assess our ability to recover from distribution shift. Distribution shift is the difference between the training distribution and the test distribution. Any mismatch may lead to bias in network predictions on the test set, as discussed in Section 3.1. Thus, it is informative to perform a test on a set of lenses with a distribution deliberately offset from  $\nu_{\text{int}}$ . To create a distribution for the shifted set, we shift and narrow the training distribution in  $\theta_E$  and  $\gamma_{\text{lens}}$ , as shown in Figure 3. We choose these shifts to mimic the shift we see in the doppelganger test set. 20 samples from the shifted distribution are turned

into images using the same simulator as our training set. This test set contains both double and quadruple image configurations. For more details on the creation of the shifted set, see Appendix B.

##### 4.2. Doppelganger Test Set

The doppelganger test set is the second simulated verification test. The goal of this test set is to evaluate our performance on images that are as close as possible to real observations. We imitate the STRIDES data by passing the best fit parameters from STRIDES23 forward modeling to our simulator. This way, we have images of realistic lensing configurations that still have a ground truth we can test against. For more details on the doppelganger simulations, see Appendix B. A comparison of the doppelganger simulations to the HST data is shown in Figure 4.

##### 4.3. HST Data

We model a set of 14 lensed quasars originally presented and analyzed in STRIDES23. We model only a subset of the original sample since we limited this analysis to images that are contained within an  $80 \times 80$  pixel cutout. We plan to model the full set in future work. These lenses were observed by programs HST-GO-15320 and HST-GO-15652 (PI: Treu). We model images taken by HST’s WFC3 in the F814W band, using data products derived in STRIDES23. We refer the reader to STRIDES23 for more details on the individual lenses and data reduction. A gallery of the observations is shown in Figure 4a.

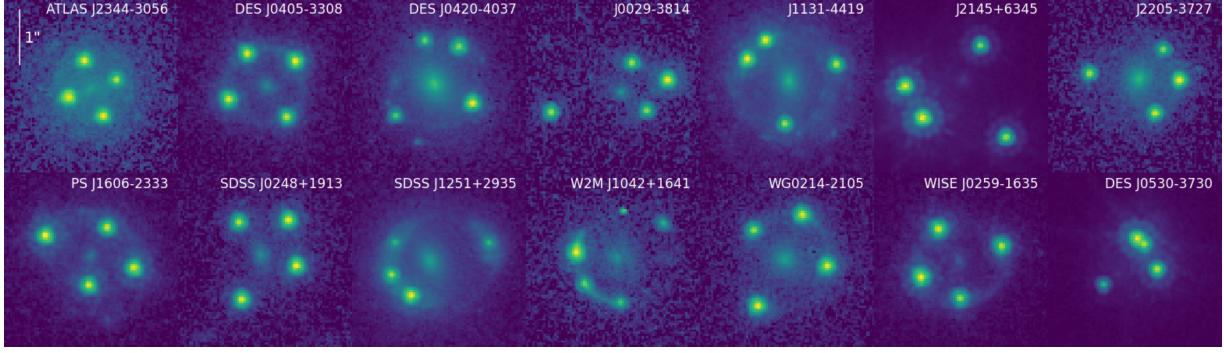
## 5. RESULTS

As introduced in Section 4, we design two verification tests to establish confidence in our method on labelled examples. We present our results on these tests first. Then, we run our method on real data for the first time, giving us an opportunity to assess our robustness.

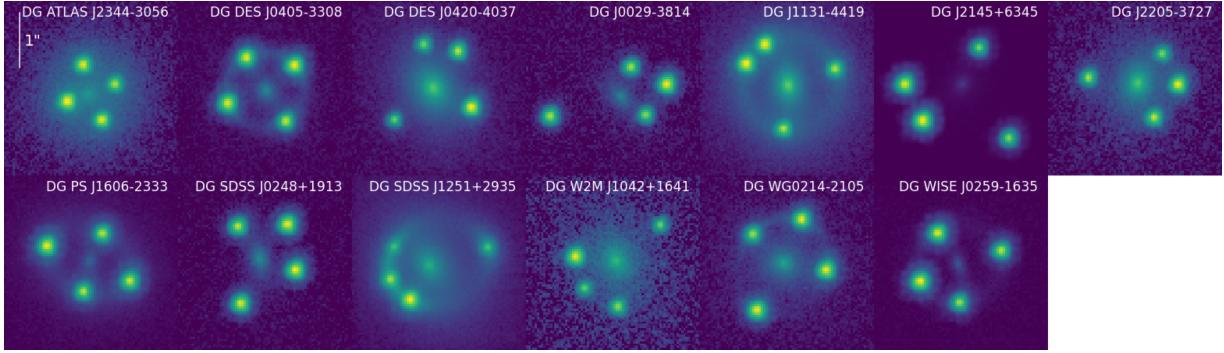
First, we review the outputs of each modeling step and introduce the metrics used to evaluate these outputs in Sections 5.1 and 5.2. Then, we show the performance of our method on the verification test sets in Section 5.3. Finally, we present our results on the real HST lenses in Section 5.4.

##### 5.1. NPE Modeling Metrics

The first step of the modeling process is neural posterior estimation of lens mass models, which produces approximate posteriors  $q_\phi(\xi_k | d_k, \nu_{\text{int}})$  for each lens  $k$ . To evaluate our initial mass modeling step, we want to evaluate how well  $q_\phi(\xi_k | d_k, \nu_{\text{int}})$  captures the ground truth  $\xi_{k,\text{truth}}$  on our verification tests.



(a) HST WFC3 images of STRIDES strongly lensed quasars in the F814W band



(b) Doppelganger simulation counterparts for the STRIDES lenses, created using PALTAS

**Figure 4.** Comparison of real HST data to the simulated doppelganger test set in the F814W band. Images are  $80 \times 80$  pixels with  $0.04''$  resolution. Images are oriented with East to the left, and North to the top. All images are plotted with log-scaled color. Note when using the doppelganger simulation procedure described in Appendix B.2, we were unable to recreate DES J0530–3730.

First we check for introduction of bias. Bias manifests as systematically high or low prediction. To check for bias, we compute the median of the difference between the mean of the posterior,  $\mu_k$  and the ground truth,  $\xi_{k,\text{truth}}$ , across the test set:

$$\text{ME} = \text{median}_k \{ \mu_k - \xi_{k,\text{truth}} \}. \quad (14)$$

Because our test sets contain a small number of lenses – 20 in the shifted set and 13 in the doppelganger set – a deviation from zero is indicative of but does not guarantee bias.

Next, we evaluate the accuracy of the network’s predictions. Accuracy captures how close the prediction is to the ground truth. A modeling technique could achieve zero bias without actually learning information by simply predicting the mean of the test distribution. Evaluating our accuracy ensures that the network is learning information from the data. To benchmark accuracy, we use median absolute error (MAE), which is also evaluated on the mean of the posterior:

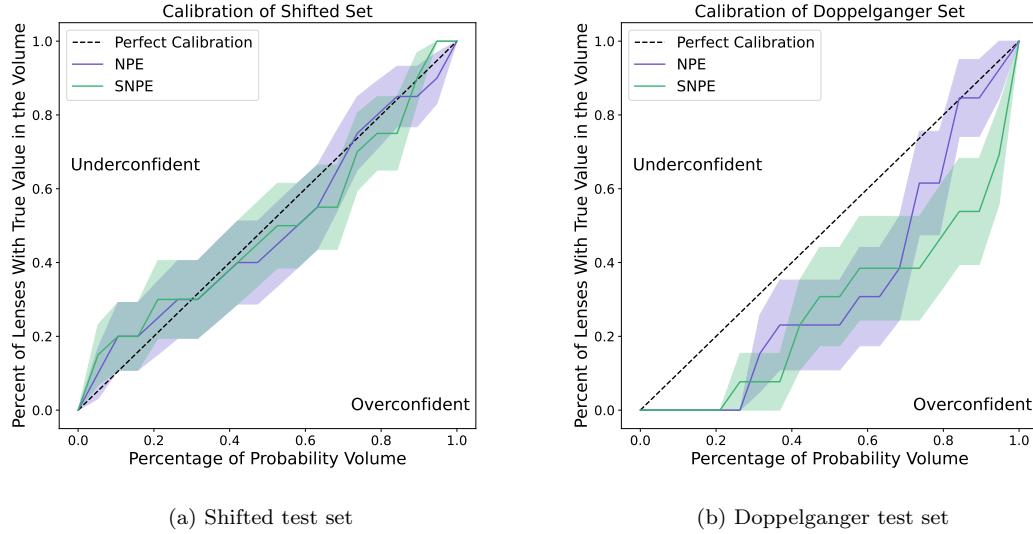
$$\text{MAE} = \text{median}_k \{ |\mu_k - \xi_{k,\text{truth}}| \}. \quad (15)$$

Next, we check for precision, which captures the amount of uncertainty on the predictions. We benchmark precision via the median size of the standard deviation,  $\sigma_k$ :

$$\text{M}(\sigma) = \text{median}_k \{ \sigma_k \}. \quad (16)$$

Note that Median Error, MAE, and Median  $\sigma$  are evaluated separately for each parameter, as shown in Tables 1 and 2.

After evaluating  $\mu_k$  and  $\sigma_k$  separately, we also need to check that the pair are calibrated properly. We assert that the  $(\mu_k, \sigma_k)$  predicted by the network describe Gaussian posteriors for the model parameters. To check whether this is true, we need to evaluate the calibration of the resulting posterior. For different  $x$ , we test whether  $x\%$  of the posterior probability volume contains the ground truth  $x\%$  of the time (see Figure 5). The mathematical formalism for this metric is described in Park et al. (2021) and Wagner-Carena et al. (2021). If  $x\%$  of the probability volume contains the truth *less* than  $x\%$  of the time, the model posteriors are overconfident. If the opposite is true, the model posteriors are underconfident.



**Figure 5.** Calibration curves for the verification test sets. In perfectly calibrated posteriors (dashed line), a given  $x\%$  of the probability volume contains the truth  $x\%$  of the time. Calibration of NPE posteriors is shown in purple. Calibration of SNPE posteriors is shown in green. The shaded region encompasses  $1\sigma$  uncertainty. Doppelganger posteriors are more overconfident than shifted set posteriors for both modeling methods, which is discussed in Section 6.2.

		$\theta_E (\text{''})$	$\gamma_1$	$\gamma_1$	$\gamma_{\text{lens}}$	$e_1$	$e_2$	$x_{\text{lens}} (\text{''})$	$y_{\text{lens}} (\text{''})$	$x_{\text{src}} (\text{''})$	$y_{\text{src}} (\text{''})$
<b>ME</b>	NPE	-0.0	0.0	-0.0	0.01	0.0	-0.02	-0.0	-0.0	0.0	0.0
	SNPE	-0.0	0.0	0.0	0.03	0.01	-0.02	-0.0	0.0	-0.0	0.0
<b>MAE</b>	NPE	0.0	0.01	0.01	0.09	0.03	0.03	0.0	0.0	0.01	0.0
	SNPE	0.01	0.01	0.01	0.09	0.03	0.02	0.0	0.0	0.01	0.01
<b>M(<math>\sigma</math>)</b>	NPE	0.01	0.02	0.02	0.12	0.04	0.04	0.0	0.0	0.01	0.01
	SNPE	0.01	0.02	0.02	0.1	0.04	0.04	0.01	0.01	0.01	0.01

**Table 1.** Metrics for **shifted set** lens models. We report median error (ME, Eqn. 14), median absolute error (MAE, Eqn. 15), and median  $\sigma$  ( $M(\sigma)$ , Eqn. 16). We report performance for both NPE and SNPE modeling.

Finally, we pay special attention to our performance on  $\gamma_{\text{lens}}$ . We compute the average percent error per lens and the average percent precision per lens. This is a useful benchmark, since percent error on  $\gamma_{\text{lens}}$  translates to percent error on  $H_0$  in TDC (Suyu 2012).

### 5.2. Population Inference Metrics

After analyzing each lens separately, individual mass models are combined to infer population-level properties. For each test set, individual posteriors are folded into a hierarchical inference for the hyperposterior  $p(\nu|\{d\})$ .

The output of inference for  $p(\nu|\{d\})$  is a set of 5e3 samples for each hyperparameter in  $\nu$ . We report the median of the samples along with an uncertainty derived from the averaged distance to the lower and upper  $1\sigma$  quantiles. We compare the final values of  $\nu$  to the ground truth, and report the error in units of standard

deviation for all parameters in Tables 4 and 5. For the doppelganger test set, we do not know the true population distribution the lens models are generated from. For these lenses, we take the sample mean and standard deviation, and use those values as a proxy ground truth we aim to recover.

For this modeling stage, we also pay close attention to performance on  $\gamma_{\text{lens}}$ . In the hierarchical inference, our primary parameter of interest is the population mean:  $\mathcal{M}_{\gamma_{\text{lens}}}$ . We evaluate percent error and percent precision on this parameter in Table 3.

### 5.3. Verification Tests

Before application to real data, we verify the performance of our method on two simulated test sets: the shifted set and the doppelganger set. See Section 4 for a full description of these test sets. We compare performance using NPE and SNPE.

		$\theta_E$ ('')	$\gamma_1$	$\gamma_1$	$\gamma_{\text{lens}}$	$e_1$	$e_2$	$x_{\text{lens}}$ ('')	$y_{\text{lens}}$ ('')	$x_{\text{src}}$ ('')	$y_{\text{src}}$ ('')
<b>ME</b>	NPE	-0.0	0.02	0.0	0.09	0.01	0.01	0.0	-0.01	0.0	-0.0
	SNPE	-0.0	0.01	0.0	0.09	0.02	0.02	0.0	-0.0	0.0	-0.0
<b>MAE</b>	NPE	0.01	0.02	0.02	0.12	0.02	0.04	0.01	0.01	0.01	0.01
	SNPE	0.01	0.02	0.01	0.1	0.03	0.02	0.01	0.0	0.0	0.0
<b>M(<math>\sigma</math>)</b>	NPE	0.0	0.02	0.02	0.16	0.03	0.03	0.0	0.0	0.01	0.01
	SNPE	0.0	0.01	0.01	0.12	0.03	0.03	0.0	0.0	0.01	0.0

**Table 2.** Metrics for **doppelganger set** lens models. We report median error (ME, Eqn. 14), median absolute error (MAE, Eqn. 15), and median  $\sigma$  ( $M(\sigma)$ , Eqn. 16). We report performance for both NPE and SNPE modeling.

		Individual Posteriors $p(\gamma_{\text{lens}} d, \nu_{\text{int}})$		Hyperposterior $p(\mathcal{M}_{\gamma_{\text{lens}}} \{d\})$	
		$\gamma_{\text{lens}}$ : % error per lens	$\gamma_{\text{lens}}$ : % precision per lens	$\mathcal{M}_{\gamma_{\text{lens}}}$ : % error	$\mathcal{M}_{\gamma_{\text{lens}}}$ : % precision
<b>Shifted</b>	NPE	4.6	6.1	2.9	1.9
	SNPE	4.2	4.9	2.0	1.4
<b>Doppelganger</b>	NPE	6.7	7.2	6.8	3.6
	SNPE	5.0	5.4	4.4	2.3

**Table 3.** Recovery of  $\gamma_{\text{lens}}$  and  $\mathcal{M}_{\gamma_{\text{lens}}}$  in verification tests. We highlight these results since error on  $\gamma_{\text{lens}}$  directly translates to error on  $H_0$  during TDC. We show results for the shifted test set (top row) and the doppelganger test set (bottom row). We evaluate performance at the individual lens modeling stage ( $\gamma_{\text{lens}}$ , assuming the interim prior) and the subsequent population-level stage ( $\mathcal{M}_{\gamma_{\text{lens}}}$ , assuming and inferring the conditional prior PDF). We also show how results compare for NPE and SNPE lens modeling. We find that the percent error decreases for both  $\gamma_{\text{lens}}$  and  $\mathcal{M}_{\gamma_{\text{lens}}}$  in both verification tests when we switch from NPE to SNPE modeling.

### 5.3.1. Shifted Test Set

We start our verification process with the shifted test set. This set of 20 lenses is designed to test our ability to handle shifts between  $\nu_{\text{int}}$  and the test distribution. First, we evaluate individual lens models. Parameter recovery of these models is summarized in Table 1. The calibration of uncertainties is shown in Figure 5a. The calibration of both NPE and SNPE posteriors scatters about the perfect calibration curve within the  $1\sigma$  uncertainty region, indicating proper recovery of Gaussian  $\mu_k$  and  $\sigma_k$ .

Next, we evaluate recovery of the population-level model. A portion of the posterior  $p(\nu|\{d\})$  is shown in Figure 6a, with all inferred model parameters summarized in Table 4. We see that the error is less than  $2\sigma$  on all hyperparameters.

### 5.3.2. Doppelganger Test Set

After evaluating performance on the shifted set, we move to the doppelganger set. This test set consists of close mocks of the real HST lenses, and aims to test our ability to model realistic lensing configurations. Parameter recovery is summarized in Table 2. We note this performance is slightly worse than the shifted test set, which we will discuss in Section 6.2. The calibration of Gaussian posteriors is summarized by the curves in Figure 5b. Both NPE and SNPE posteriors are overconfident, as evidenced by the departure from the perfect

calibration line. This means that the predicted uncertainties  $\hat{\sigma}_k$  are too small given the distances between  $\hat{\mu}_k$  and  $\xi_{k,\text{truth}}$ . We discuss possible drivers of this effect in Section 6.2. We continue on to the hierarchical modeling step.

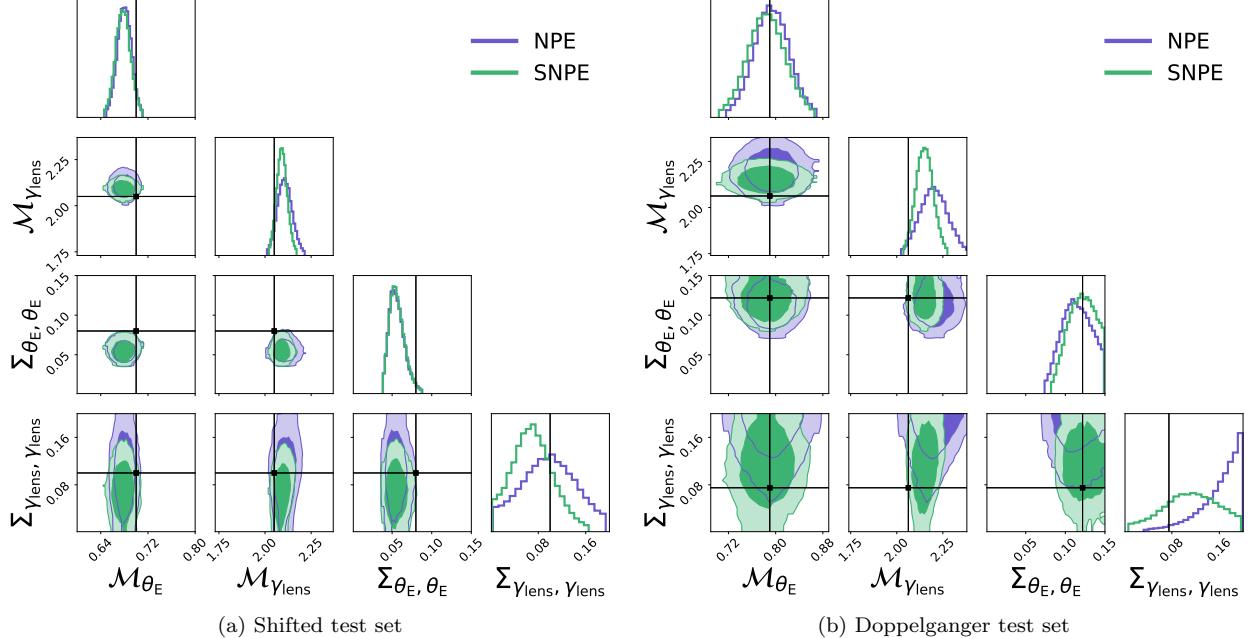
The hyperposterior is shown in Figure 6b and recovery of hyper-parameters is summarized in Table 5. Over-confident calibration of error bars means that during the hierarchical inference, uncertainty that should be attributed to individual posteriors is interpreted as a larger population  $\sigma$ .

Results from this doppelganger test set give us insight into possible systematic bias on  $\gamma_{\text{lens}}$ . We note that percent error on  $\mathcal{M}_{\gamma_{\text{lens}}}$  is higher than the percent precision for both shifted and doppelganger test sets. Performance on these sets is the best indicator of what we can expect on real data. When interpreting  $\mathcal{M}_{\gamma_{\text{lens}}}$  from the real data, we should keep in mind that the result might be affected by a small bias towards higher values.

## 5.4. Application to HST Data

After evaluating performance on verification tests, we apply our modeling technique to HST observations of 14 lensed quasars from the STRIDES sample. See Figure 4a and Section 4.3 for more information on these lenses.

We perform both NPE and SNPE modeling of the data to produce approximate lens model posteriors.



**Figure 6.** 2D contours of the hyperposterior  $p(\nu|d)$  on the verification test sets. Inference from NPE posteriors is shown in purple. Inference from SNPE posteriors is shown in green. The ground truth is shown as a black line. Shaded contours are 68% and 95% intervals. Note for the doppelganger test set, we use the sample mean and standard deviation as a proxy ground truth.

	$\mathcal{M}_{\theta_E}$	$\mathcal{M}_{\gamma_{\text{lens}}}$	$\Sigma_{\theta_E, \theta_E}$	$\Sigma_{\gamma_{1/2}, \gamma_{1/2}}$	$\Sigma_{\gamma_{\text{lens}}, \gamma_{\text{lens}}}$	$\Sigma_{e_{1/2}, e_{1/2}}$
<b>Inferred</b>	$0.68 \pm 0.01$	$2.09 \pm 0.03$	$0.06 \pm 0.01$	$0.12 \pm 0.004$	$0.07 \pm 0.04$	$0.16 \pm 0.02$
<b>Ground Truth</b>	0.7	2.05	0.08	0.12	0.1	0.2
<b>Error (in <math>\sigma</math>)</b>	-2.0	+1.3	-2.0	0.0	-0.8	-2.0

**Table 4.** Inferred hyperparameters  $\nu$  for the **shifted** set using SNPE modeling. Error is reported in units of  $\sigma$  to check for statistical significance.

	$\mathcal{M}_{\theta_E}$	$\mathcal{M}_{\gamma_{\text{lens}}}$	$\Sigma_{\theta_E, \theta_E}$	$\Sigma_{\gamma_{1/2}, \gamma_{1/2}}$	$\Sigma_{\gamma_{\text{lens}}, \gamma_{\text{lens}}}$	$\Sigma_{e_{1/2}, e_{1/2}}$
<b>Inferred</b>	$0.78 \pm 0.03$	$2.15 \pm 0.05$	$0.12 \pm 0.02$	$0.10 \pm 0.01$	$0.11 \pm 0.05$	$0.15 \pm 0.02$
<b>Proxy Ground Truth</b>	0.79	2.06	0.12	0.08	0.08	0.12
<b>Error (in <math>\sigma</math>)</b>	-0.3	+1.8	0.0	+2.0	+0.6	+1.5

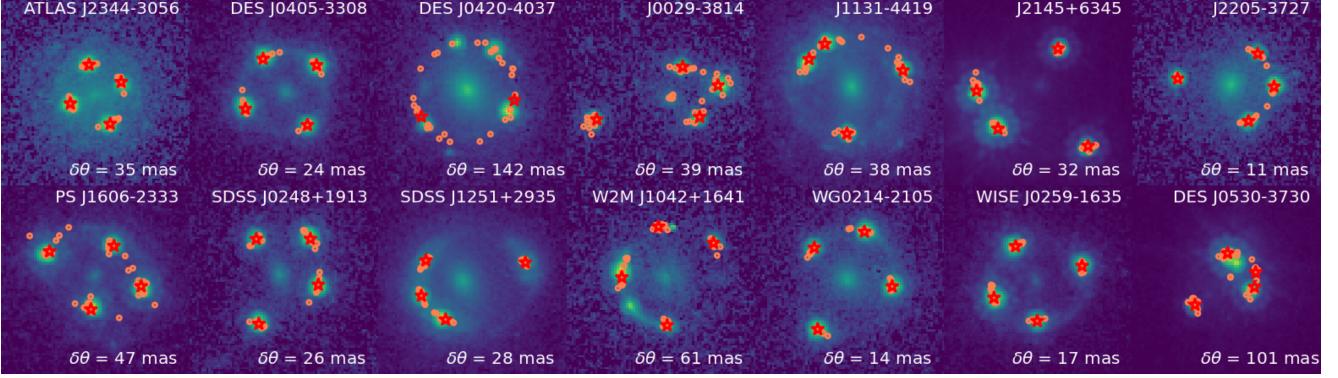
**Table 5.** Inferred hyperparameters  $\nu$  for the **doppelganger set** using SNPE modeling. Error is reported in units of  $\sigma$  to check for statistical significance.

Since we have no ground truth, we are unable to verify NPE and SNPE performance using the metrics previously discussed. Instead, we verify the modeling performance by calculating the image positions predicted by the SNPE lens model. Then, we overlay those image positions on top of the input data. This gives us an opportunity to roughly assess whether the lens model is consistent with the data. See Figure 7. We discuss recovery of point source image positions in Section 6.3.

We then perform a hierarchical inference for the population model using the 14 individual lens models. In

Figure 8, we show the population-level model from both NPE and SNPE posteriors. The hyperparameter values are reported in Table 6 for the inference using SNPE models. We use these values as our final result, since SNPE modeling had better performance on our primary metric of interest,  $\mathcal{M}_{\gamma_{\text{lens}}}$ , in verification tests. We find  $\mathcal{M}_{\gamma_{\text{lens}}} = 2.13 \pm 0.06$  on the real data, which we discuss further in Section 6.4.

## 6. DISCUSSION



**Figure 7.** Image positions predicted by the SNPE lens models overlaid on top of the **real HST data**. HST images are  $80 \times 80$  pixels with  $0.04''$  resolution. Images are oriented with East to the left, and North to the top. All images are plotted with log-scaled color. Image positions computed from the mean of the lens model posterior are shown as red stars, image positions computed from 10 samples from the lens model posterior are shown as orange dots. Image positions computed from the mean of the posterior are compared against image positions from STRIDES23 modeling, and the average difference is quoted as  $\delta\theta$  in mas at the bottom corner of each image. If the SNPE-predicted image positions are farther than  $0.14''$  in either ra or dec from any STRIDES23 image position, that image is discarded from the  $\delta\theta$  calculation.

First, we discuss verification test results. We compare NPE and SNPE performance in Section 6.1 and we discuss discrepancies between shifted set and doppelganger set performance in Section 6.2. Then, we discuss results on real HST data, including image position recovery in Section 6.3, context for our result on  $\mu(\gamma_{\text{lens}})$  in Section 6.4, and a comparison to STRIDES23 modeling in Section 6.5. After understanding results in more detail, we hypothesize possible performance drivers such as the choice of functional form of the individual models in Section 6.6 and the choice of functional form for the cPDF in Section 6.7. We elaborate on our attempt to use hierarchical reweighting in Section 6.8, and pose future directions in Section 6.9. Finally, we report a timing analysis in Section 6.10.

### 6.1. NPE vs SNPE

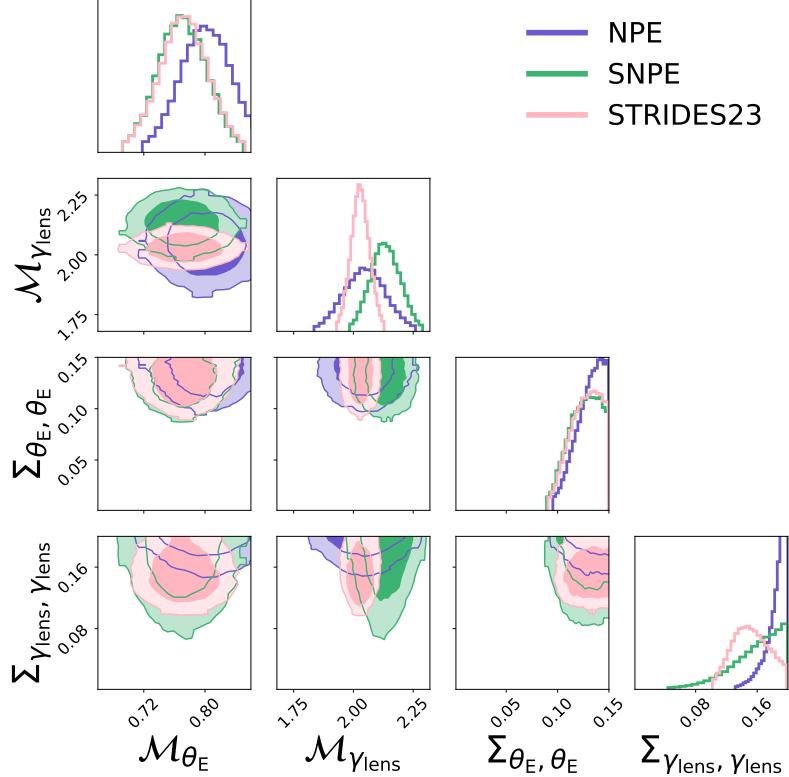
We apply SNPE in an attempt to improve upon the precision of NPE posteriors. In an application of NPE to real data, we do not know the parameter range of the test distribution ahead of time, and thus must choose our training prior  $\nu_{\text{int}}$  to be quite broad in order to fully cover the possible parameter space (see Table 7). We investigate SNPE as a way to cope with this large volume. We especially note that our training prior for the source’s  $R_*$  is much wider than the one chosen in Park et al. (2021). Since the size of the source galaxy is known to have degeneracy with  $\gamma_{\text{lens}}$  (Marshall et al. 2007), enforcing a wider prior for this parameter decreases our ability to break this degeneracy, and learn  $\gamma_{\text{lens}}$ .

We find some evidence for the success of SNPE when assessing performance on  $\gamma_{\text{lens}}$  in Table 3. On both the shifted test set and the doppelganger test set, average

percent error on  $\gamma_{\text{lens}}$  decreased when switching from NPE to SNPE. However, if we look at the median error on  $\gamma_{\text{lens}}$  for the shifted test set in Table 1, a positive bias increases from 0.01 to 0.03 when switching from NPE to SNPE modeling. Here we caution that SNPE may amplify an underlying bias, but we cannot say for certain without further investigation into this effect.

### 6.2. Shifted vs Doppelganger Performance

The only difference between the shifted test set and the doppelganger test set is the way in which the underlying model parameters were sampled. Shifted set lens parameters were sampled from a Gaussian distribution described in Appendix B, while doppelganger set parameters come from forward modeling of real systems. We see some discrepancy between the performance on the two sets, most notably the overconfidence in error bars for the doppelganger set seen in Figure 5. We have three hypotheses for this discrepancy. The first, and simplest, is that the doppelgangers are simply an unlucky statistical draw, since we are dealing with a relatively small sample size. The second hypothesis is that there is a complex selection function on the doppelgangers, putting them in a rarer part of parameter space compared to the shifted set. This selection may be difficult to see when interpreting in lower dimensional space. The third is that the realistic lensing configurations of the doppelgangers are more sensitive to the diagonal covariance assumption. We test the use of full covariance posteriors in Appendix D, and do see some evidence of better calibration on the doppelganger test set. See Section 6.6 for a further discussion of full covariance NPE.



**Figure 8.** 2D contours of the hyperposterior  $p(\nu|d)$  on the **real HST data**. Inference from NPE mass models is shown in purple. Inference from SNPE mass models is shown in green. Inference from STRIDES23 mass models is shown in pink. Shaded contours are 68% and 95% intervals. The NPE constraint on  $\mu(\gamma_{\text{lens}})$  shifts and becomes more precise after applying SNPE, which is accompanied by a more constrained  $\sigma(\gamma_{\text{lens}})$ .

	$M_{\theta_E}$	$M_{\gamma_{\text{lens}}}$	$\Sigma_{\theta_E, \theta_E}$	$\Sigma_{\gamma_{1/2}, \gamma_{1/2}}$	$\Sigma_{\gamma_{\text{lens}}, \gamma_{\text{lens}}}$	$\Sigma_{e_{1/2}, e_{1/2}}$
<b>NPE</b>	$0.80 \pm 0.03$	$2.05 \pm 0.09$	$0.13 \pm 0.01$	$0.11 \pm 0.009$	$0.19 \pm 0.01$	$0.18 \pm 0.02$
<b>SNPE</b>	$0.77 \pm 0.03$	$2.13 \pm 0.06$	$0.13 \pm 0.02$	$0.10 \pm 0.01$	$0.16 \pm 0.03$	$0.18 \pm 0.02$
<b>STRIDES23</b>	$0.77 \pm 0.03$	$2.03 \pm 0.04$	$0.13 \pm 0.02$	$0.08 \pm 0.01$	$0.15 \pm 0.03$	$0.14 \pm 0.02$

**Table 6.** Inferred hyperparameters of the lens population model for the **real HST data**.

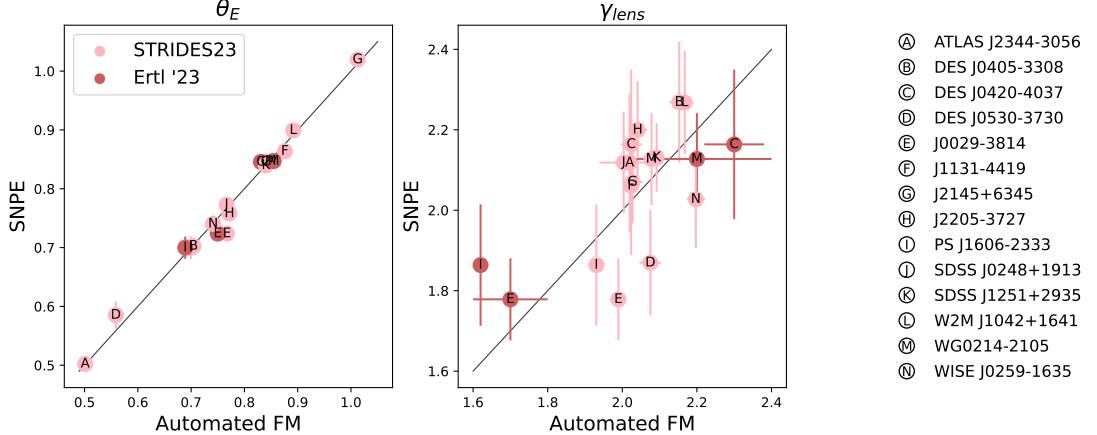
If any of these hypotheses are also true for the HST test set, our error bars on the data may be overconfidently calibrated. Overconfident mis-calibration of error bars has been found in forward modeling (Tan et al. 2024). We caution that mis-calibration of errors in individual posteriors can be absorbed by the population scatter when doing hierarchical inference, so careful error calibration is necessary to provide trustworthy population widths. In this work, we focus on the recovery of the population mean  $M_{\gamma_{\text{lens}}}$ , and acknowledge that better error calibration will be needed in further work for  $\Sigma_{\gamma_{\text{lens}}, \gamma_{\text{lens}}}$ .

### 6.3. Image Position Recovery on Data

We provide a benchmark for lens modeling performance on the real HST data that does not depend on

ground truth lens model parameters. NPE lens models provide lens mass parameters and a source position, which is enough information to compute image positions. We compute the image positions from the predicted lens models, and overlay the positions on top of the input data. We compare the SNPE-predicted image positions to the image positions learned in the STRIDES23 modeling (which we take as a proxy ground truth), and compute the average deviation  $\delta\theta$ . If the SNPE-predicted image positions are farther than  $0.14''$  in either ra or dec from any STRIDES23 image position, that image is discarded from the  $\delta\theta$  calculation. We show this check in Figure 7. We discuss some possible causes of high  $\delta\theta$ .

For lens DES J0420–4017, we hypothesize low performance on this lens is due to the complex source galaxy



**Figure 9.** Comparison of predicted mass model parameters  $\theta_E$  and  $\gamma_{\text{lens}}$  on real HST data. Comparison to STRIDES23 models is shown in light pink, comparison to Ertl et al. (2023) models is shown in red.  $1\sigma$  uncertainty bars are shown for both NPE and automated forward modeling (FM) techniques. Note different priors and different model complexity are assumed for the separate modeling techniques.

structure clearly visible in the lensed arc. The simulated training set did not include complex source light, so this problem could be mitigated with more realistic training simulations in the future. For lens W2M J1042+1641, there is an artifact in the top of the image in the F815W filter (this artifact is not seen in the other filters observed). The network seems to confuse this artifact for a point source image, and ignores the point source image in the bottom left corner. This effect could be mitigated by modeling in multiple filters and simulating similar artifacts in our noise model. For lens DES J0530–3730, the lensing configuration is quite compact. The model produced by STRIDES23 is a swallowtail lens, which is a rare configuration where the caustic overlaps. It is possible that during training, the network never saw a configuration close enough to this one to model the lens correctly. It may be possible to mitigate this problem with larger training sets or more efficient sampling of the parameter space during training.

Despite some lenses having visibly incorrect image positions, we use all 14 lenses for our hierarchical constraint. With our automated modeling technique, we hope to be robust to imperfect modeling on some fraction of the individual lenses when we move to the population level analysis. To check for this robustness, we also perform the population-level inference without DES J0420–4017, W2M J1042+1641, and DES J0530–3730. We find that inclusion or exclusion of these lenses does not significantly change the constraint. For example, without the three lenses, we infer  $\mathcal{M}_{\theta_E} = 0.78 \pm 0.04$  and  $\mathcal{M}_{\gamma_{\text{lens}}} = 2.14 \pm 0.07$ . Comparing to our initial analysis in Table 6, we see these constraints are consistent.

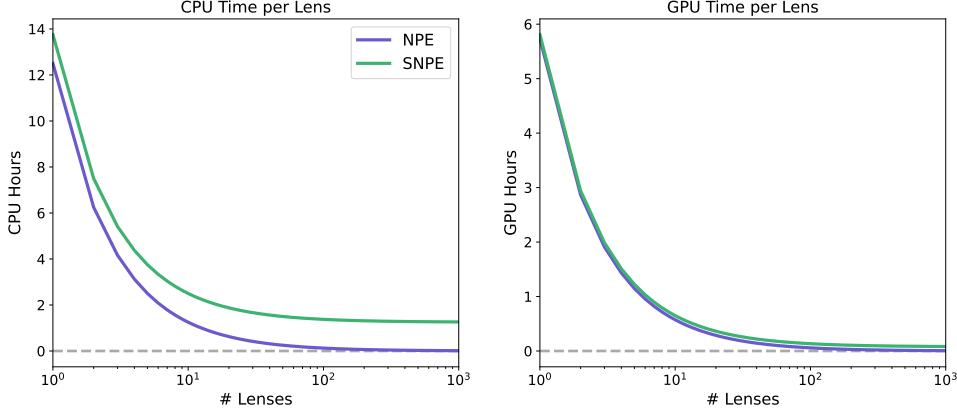
#### 6.4. Population Distribution of the Power-Law Slope

We are interested in our recovery of  $\mathcal{M}_{\gamma_{\text{lens}}}$ , since  $\gamma_{\text{lens}}$  is known to be degenerate with  $H_0$ . Additionally,  $\mathcal{M}_{\gamma_{\text{lens}}}$  has interesting astrophysical implications. On the real HST data, we measure a mean power-law slope of  $\mathcal{M}_{\gamma_{\text{lens}}} = 2.13 \pm 0.06$ . From automated forward modeling, we infer  $\mathcal{M}_{\gamma_{\text{lens}}} = 2.03 \pm 0.04$ .

Both values for the mean power-law slope are consistent with an isothermal profile, and are also consistent with the mean power-law slope at the Einstein radius measured on the SLACS lenses (Auger et al. 2010; Shajib et al. 2021; Etherington et al. 2022; Tan et al. 2024). This indicates a high-degree of similarity between SLACS lenses and this subset of STRIDES deflectors, even though they have been selected differently and span over a different range of redshift. The fact that our measurement is compatible with an isothermal profile can be interpreted in the context of the “bulge-halo” conspiracy (Treu et al. 2006; Buote & Humphrey 2010; Cappellari 2016), which refers to the nearly isothermal profiles of the total matter distribution observed in lensing and kinematic studies, although neither the stellar or dark matter distribution follows a power law with  $\gamma = 2$ . A specific arrangement between the two profiles is required to produce a profile that is nearly isothermal over a wide range of radius. We observe that a similar effect could be at play for this sample of lensed quasars.

#### 6.5. Comparison to Automated Forward Modeling

All of the HST lenses we model were also modeled in STRIDES23 using an automated likelihood-based technique. Four lenses were additionally modelled in Ertl et al. (2023), also using an automated likelihood-based



**Figure 10.** Lens modeling compute time as a function of test set size. We show the CPU and GPU compute time per lens. Note that the curves flatten as the number of lenses increases, since NPE training is an amortized cost. We assume 5e5 lenses are simulated for the NPE training set, and 5e4 lenses are simulated for each SNPE training set. Note that CPU compute time is bounded at 100 CPU hours per lens in STRIDES23.

technique. We compare our results to these automated techniques on the data, but do so with an abundance of caution. There are some significant differences between the techniques, including:

- **Model complexity.** In the automated forward modeling, more complex models are allowed. For example, in [STRIDES23](#), lens light is assumed to have two components. In our NPE modeling, we assume single component lens light.
- **Marginalization over nuisance parameters.** In automated forward modeling, every model component must be modelled exactly. For example, a choice of the PSF model must be made alongside the mass model, and all other model components. In our work, we do not require an explicit choice for the PSF (and other model components), but rather implicitly marginalize over the choice of PSF.
- **Prior choices.** The modeling techniques assume different priors for parameters. This effect is most notable on  $\gamma_{\text{lens}}$ , which is often a prior-dominated parameter.

Despite these differences, we hope to find consistent posteriors between these modeling techniques. We compare the parameter estimates in Figure 9. We find that predictions for  $\theta_E$  are consistent. On the other hand, predictions for  $\gamma_{\text{lens}}$  are not always consistent. In particular, while the discrepancy between modeling methods is often captured by the larger NPE  $1\sigma$  uncertainty, the [STRIDES23](#)  $1\sigma$  uncertainty is not large enough to account for any scatter in modeling results. This is further

evidence for the conclusion presented in [Tan et al. \(2024\)](#) that the uncertainty from forward modeling techniques is often underestimated. The larger  $\sigma$  values from [Ertl et al. \(2023\)](#) may more accurately reflect uncertainty from the automated lens modeling.

We also compare our population model to the population model inferred from [STRIDES23](#) posteriors, as shown in Figure 8 and Table 6. We explicitly account for the informative prior on  $\gamma_{\text{lens}}$  that was assumed for the [STRIDES23](#) modeling by including an interim prior  $\nu_{int}$  in our hierarchical inference with the forward modeling posteriors. The models agree perfectly for  $\mathcal{M}_{\theta_E}$  and  $\Sigma_{\theta_E, \theta_E}$ . There is some discrepancy on  $\mathcal{M}_{\gamma_{\text{lens}}}$ , which is not surprising considering the scatter on individual predictions seen in Figure 9. However, central values are consistent within  $2\sigma$ . Additionally, the precision on  $\mathcal{M}_{\gamma_{\text{lens}}}$  is comparable between the two techniques (3% with SNPE modeling, and 2% with automated forward modeling).

## 6.6. Lens Model Functional Form

Assuming a diagonal Gaussian functional form for the lens model posteriors is a simplification. We do expect covariances between some lens parameters, which our model cannot capture. We suspect that when there is not enough information in an image to break degeneracies between parameters, our diagonal posterior is forced to overconfidently converge to one solution rather than widen to allow both solutions. Ultimately, we suspect that the lack of expressivity of our functional form is one of the main weaknesses of this method.

We experimented with allowing a full covariance in our Gaussian posteriors, thereby enabling the lens models to capture correlations between parameters. This tech-

nique passed both verification tests, with similar performance to the diagonal NPE. However, when we applied the full covariance NPE to the 14 real HST lenses, we saw multiple predictions with a  $\sigma_k(\gamma_{\text{lens}})$  values wider than the  $\gamma_{\text{lens}}$  training distribution. We hypothesize that with the freedom to predict a full covariance matrix, the model is less robust to the domain shift between simulated and real data. We suggest that when increasing the flexibility of the functional form, an accompanying increase in the realistic complexity in the training simulations is necessary.<sup>5</sup> For more details on this test, see Appendix D. We note that there is evidence that a more flexible functional form can improve performance. In one application of neural network modeling to galaxy-galaxy lenses, it was demonstrated that calibration of posteriors was improved when moving from a single Gaussian posterior to a mixture of Gaussians (Legin et al. 2023).

### 6.7. Population Model Functional Form

For this analysis, we assumed the population distribution of lens parameters  $p(\xi|\nu)$  takes a diagonal Gaussian functional form. The true population distribution almost certainly follows a more complex functional form. For example, we know the distribution of  $\theta_E$  has a heavy truncation at 0, and its shape is non-Gaussian. In future analyses, we aim to relax the Gaussian assumption.

### 6.8. Re-Weighting Individual Posteriors

When using lens models within a hierarchical framework, the individual systems' posterior PDFs should be computed by re-weighting the interim posteriors, accounting for the distribution shift between the interim prior and the conditional PDF in the process. This hierarchical re-weighting scheme was derived in Wagner-Carena et al. (2021), and the idea is that each lens model can take advantage of additional information from the population-level model. We applied the same re-weighting scheme for our individual posteriors, with some modifications given the small number of lenses in our sample. For details, see Appendix E.

We found that the resulting re-weighted, “final” posterior PDFs were systematically overconfident in our verification tests. This could stem from miscalibration in the original posteriors that is amplified after re-weighting. Our hypothesis is that this is a likely the result of an insufficiently flexible functional form of the

lens posteriors. We save the investigation of more flexible PDFs for further work.

### 6.9. Timing

NPE and SNPE are advantageous modeling techniques given their speed compared to likelihood-based techniques. We analyze the computing resources required for our method, and project how this requirement will scale for the number of lenses in a given test set. We take stock of the compute used for each step of the modeling process. Then, we summarize the total compute used per lens. Simulation of lenses using PALTAS takes 0.09 CPU seconds per lens. Simulation cost is incurred one time for the NPE training set (12.5 hours for 5e5 lenses), and many times for SNPE training (1.25 hours for 5e4 lenses), which requires a new training set for every lens. Training of the network uses GPU compute. NPE training is only run once, and converges in 5.73 GPU hours. SNPE training is run once for each lens, and uses an additional 0.46 GPU hours per lens. Finally, we use CPU compute to generate predictions. This contribution is negligible, as it only takes 0.7 seconds per lens. We plot the compute time as a function of the number of lenses in the test set in Figure 10. Note if we increase the number of lenses in the test set, this decreases the cost of the initial NPE training per lens, but the SNPE cost remains the same.

### 6.10. Future Directions

We would like to relax the assumption of a Gaussian functional form for the lens model posteriors. An extension to this method would be to use a normalizing flow component to allow freedom of the functional form of the lens posterior, as is done in Poh et al. (2022).

Instead of using SNPE, we could use the NPE approximate posterior as a starting point for likelihood-based forward modeling. This approach was examined in Pearson et al. (2021).

The primary driver of performance with machine learning techniques is often the quality of the training data. To improve the quality of training examples, we should add more realism to our training set simulations. One important factor may be the assumed source galaxy light profile in our training simulation, which is a single Sérsic. More complexity could be included in future applications of this technique. For example, we could introduce postage stamps of real galaxies for the source galaxy light in training simulations. This is done using COSMOS galaxies in Wagner-Carena et al. (2023). Going futher, we could use postage stamps of real galaxies for both the source galaxy and lens light components, as is done in Schuld et al. (2023b). We could also increase the complexity of the mass model, for example,

<sup>5</sup> Note that since the SNPE proposal distribution depends on the NPE posterior, this behavior prevented us from running full-covariance SNPE.

including massive satellites, which we know are present in lenses like PS J1606–2333.

## 7. CONCLUSION

We are working to enable an independent measurement of the Universe’s expansion through time-delay cosmography with lensed quasars discovered by LSST. To enable population-level analyses of a large sample of time-delay lenses, we apply NPE and SNPE for fast and standardizable lens modeling. We use verification tests to establish confidence in our approximate inference technique, ensuring both individual lens models and the population level model are recovered. After assessing verification tests, we apply our technique to real time-delay lenses for the first time. We put the first population-level constraint on a subset of the STRIDES lensed quasars, finding  $\mathcal{M}_{\gamma_{\text{lens}}} = 2.13 \pm 0.06$ .

We address the guiding questions we initially posed:

- Will the application of NPE for strong lens mass modeling produce reliable models on real time-delay lenses? In verification tests, what is the percent error per lens on the PEMD power-law slope  $\gamma_{\text{lens}}$ ?

Answer: NPE and SNPE modeling are successfully used for our first application on HST data. Individual lens constraints recovered during verification tests have roughly 5% error per lens on  $\gamma_{\text{lens}}$ .

- How does the application of SNPE compare to NPE? Does the increased sampling density of SNPE improve precision of lens model posteriors?

Answer: SNPE improves percent error at the individual level and the population level for  $\gamma_{\text{lens}}$ . However, it is unclear that SNPE always improves performance, as evidenced by median absolute error.

- What population constraint can we put on the real data using hierarchical Bayesian inference? In verification tests, what is the percent error on the population mean of the PEMD power-law slope  $\mathcal{M}_{\gamma_{\text{lens}}}$ ?

Answer: We achieve 2% and 4% errors on  $\mathcal{M}_{\gamma_{\text{lens}}}$  in shifted and doppelganger verification tests, respectively, with SNPE modeling. We find a population mean on the real HST data:  $\mathcal{M}_{\gamma_{\text{lens}}} = 2.13 \pm 0.06$ .

With this work, we make necessary steps towards building a fully automated lens modeling pipeline that will be consistently applied to all parts of the LSST strong lens sample.

## 8. ACKNOWLEDGEMENTS

This paper has undergone internal review in the LSST Dark Energy Science Collaboration. We would like to thank internal reviewers Stefan Schuldt and Anowar Shajib for their extensive contributions in this role. We additionally thank Jelle Aalbers, Phil Holloway, Xiangyu Huang, Ralf Kaehler, Narayan Khadka, Tian Li, and Greg Madejski for many useful conversations along the way. We thank Ji Won Park, Tom Collett, Sherry Suyu, and Aymeric Galan for impactful conversations and suggestions.

SE developed all code in LENS-NPE, ran all analysis, produced all figures, and wrote the main body of the text. SWC helped design and implement the method and provided feedback on all aspects. PM helped design the statistical framework and verification tests, and provided feedback on all aspects. MM provided interpretation and context for results and provided feedback on all aspects. SB helped with simulation implementation and provided feedback on all aspects. AR provided feedback on all aspects. TS and TT reduced HST data products, produced forward modeling chains, and provided help handling the data products, in addition to feedback on all results. SS and AS provided feedback and suggestions on all aspects. PV provided feedback on method and results.

HST observations were conducted by programs HST-GO-15320 and HST-GO-15652 (PI: Treu). SE acknowledges funding from the National Science Foundation GRFP, and the Stanford Data Science Scholars program. This work was supported by the U.S. Department of Energy under contract number DE-AC02-76SF00515. MM acknowledges support by the SNSF (Swiss National Science Foundation) through mobility grant P500PT\_203114. SB acknowledge support by the Department of Physics and Astronomy, Stony Brook University. TS and TT acknowledge support by the the National Science Foundation through grant NSF-AST-1906976 and NSF-AST-1907396 “Collaborative Research: Toward a 1% measurement of the Hubble Constant with gravitational time-delays”. SS has received funding from the European Union’s Horizon 2022 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101105167 - FASTIDIoUS. This work was also supported by NASA through the NASA Hubble Fellowship grant HST-HF2-51492 awarded to AS by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555.

The DESC acknowledges ongoing support from the Institut National de Physique Nucléaire et de Physique

des Particules in France; the Science & Technology Facilities Council in the United Kingdom; and the Department of Energy, the National Science Foundation, and the LSST Corporation in the United States. DESC uses resources of the IN2P3 Computing Center (CC-IN2P3-Lyon/Villeurbanne - France) funded by the Centre National de la Recherche Scientifique; the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract

No. DE-AC02-05CH11231; STFC DiRAC HPC Facilities, funded by UK BEIS National E-infrastructure capital grants; and the UK particle physics grid, supported by the GridPP Collaboration. This work was performed in part under DOE Contract DE-AC02-76SF00515.

Source code for this work is publicly available in the repository LENS-NPE<sup>6</sup>. This repository makes use of public software packages PALTAS (Wagner-Carena et al. 2023) and LENSTRONOMY (Birrer & Amara 2018; Birrer et al. 2021).

## REFERENCES

- Abdalla, E., Abellán, G. F., Aboubrahim, A., et al. 2022, Journal of High Energy Astrophysics, 34, 49–211, doi: [10.1016/j.jheap.2022.04.002](https://doi.org/10.1016/j.jheap.2022.04.002)
- Auger, M. W., Treu, T., Bolton, A. S., et al. 2010, The Astrophysical Journal, 724, 511–525, doi: [10.1088/0004-637x/724/1/511](https://doi.org/10.1088/0004-637x/724/1/511)
- Barkana, R. 1998, The Astrophysical Journal, 502, 531–537, doi: [10.1086/305950](https://doi.org/10.1086/305950)
- Birrer, S., & Amara, A. 2018, Lenstronomy: multi-purpose gravitational lens modelling software package. <https://arxiv.org/abs/1803.09746>
- Birrer, S., Shajib, A. J., Galan, A., & et al., M. M. 2020, A&A, doi: [10.1051/0004-6361/202038861](https://doi.org/10.1051/0004-6361/202038861)
- Birrer, S., Shajib, A. J., Gilman, D., et al. 2021, arXiv preprint arXiv:2106.05976
- Birrer, S., Millon, M., Sluse, D., et al. 2024, Space Sci. Rev., 220, 48, doi: [10.1007/s11214-024-01079-w](https://doi.org/10.1007/s11214-024-01079-w)
- Buote, D. A., & Humphrey, P. J. 2010, Dark Matter in Elliptical Galaxies (Springer New York), 235–277, doi: [10.1007/978-1-4614-0580-1\\_8](https://doi.org/10.1007/978-1-4614-0580-1_8)
- Cappellari, M. 2016, Annual Review of Astronomy and Astrophysics, 54, 597–665, doi: [10.1146/annurev-astro-082214-122432](https://doi.org/10.1146/annurev-astro-082214-122432)
- Dauphin, F., Anderson, J., Bajaj, V., et al. 2021, The WFPC2 and WFC3 PSF Database, Instrument Science Report WFC3 2021-12
- DESI Collaboration, Adame, A. G., Aguilar, J., et al. 2024, arXiv e-prints, arXiv:2404.03002, doi: [10.48550/arXiv.2404.03002](https://doi.org/10.48550/arXiv.2404.03002)
- Di Valentino, E., Mena, O., Pan, S., et al. 2021, Classical and Quantum Gravity, 38, 153001, doi: [10.1088/1361-6382/ac086d](https://doi.org/10.1088/1361-6382/ac086d)
- Ding, X., Treu, T., Birrer, S., et al. 2021, Monthly Notices of the Royal Astronomical Society, 503, 1096
- Ertl, S., Schuldt, S., Suyu, S., et al. 2023, Astronomy & Astrophysics, 672, A2
- Etherington, A., Nightingale, J. W., Massey, R., et al. 2022, MNRAS, 517, 3275, doi: [10.1093/mnras/stac2639](https://doi.org/10.1093/mnras/stac2639)
- Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306, doi: [10.1086/670067](https://doi.org/10.1086/670067)
- Galan, A., Caminha, G. B., Knollmüller, J., Roth, J., & Suyu, S. H. 2024, El Gordo needs El Anzuelo: Probing the structure of cluster members with multi-band extended arcs in JWST data. <https://arxiv.org/abs/2402.18636>
- Galan, A., Vernardos, G., Peel, A., Courbin, F., & Starck, J.-L. 2022, Astronomy & Astrophysics, 668, A155, doi: [10.1051/0004-6361/202244464](https://doi.org/10.1051/0004-6361/202244464)
- Gawade, P., More, A., More, S., et al. 2024, Neural network prediction of model parameters for strong lensing samples from Hyper Suprime-Cam Survey. <https://arxiv.org/abs/2404.18897>
- Gentile, F., Tortora, C., Covone, G., et al. 2023, Monthly Notices of the Royal Astronomical Society, 522, 5442–5455, doi: [10.1093/mnras/stad1325](https://doi.org/10.1093/mnras/stad1325)
- Greenberg, D. S., Nonnenmacher, M., & Macke, J. H. 2019, Automatic Posterior Transformation for Likelihood-Free Inference. <https://arxiv.org/abs/1905.07488>
- Gu, A., Huang, X., Sheu, W., et al. 2022, The Astrophysical Journal, 935, 49, doi: [10.3847/1538-4357/ac6de4](https://doi.org/10.3847/1538-4357/ac6de4)
- He, K., Zhang, X., Ren, S., & Sun, J. 2016, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778, doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90)
- He, T., Zhang, Z., Zhang, H., et al. 2018, Bag of Tricks for Image Classification with Convolutional Neural Networks. <https://arxiv.org/abs/1812.01187>
- Hezaveh, Y. D., Levasseur, L. P., & Marshall, P. J. 2017, Nature, 548, 555–557, doi: [10.1038/nature23463](https://doi.org/10.1038/nature23463)
- Hogg, N. B. 2023, Monthly Notices of the Royal Astronomical Society: Letters, 529, L95–L100, doi: [10.1093/mnrasl/slae005](https://doi.org/10.1093/mnrasl/slae005)

<sup>6</sup> <https://github.com/smericks/lens-npe>

- Kingma, D. P., & Ba, J. 2014, Adam: A Method for Stochastic Optimization.  
<https://arxiv.org/abs/1412.6980>
- Kolmus, A., Janquart, J., Baka, T., et al. 2024, Tuning neural posterior estimation for gravitational wave inference. <https://arxiv.org/abs/2403.02443>
- Legin, R., Hezaveh, Y., Perreault-Levasseur, L., & Wandelt, B. 2023, The Astrophysical Journal, 943, 4, doi: [10.3847/1538-4357/aca7c2](https://doi.org/10.3847/1538-4357/aca7c2)
- Linder, E. V. 2011, Physical Review D, 84, doi: [10.1103/physrevd.84.123529](https://doi.org/10.1103/physrevd.84.123529)
- Lombardi, M. 2024, arXiv e-prints, arXiv:2406.15280, doi: [10.48550/arXiv.2406.15280](https://doi.org/10.48550/arXiv.2406.15280)
- LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, LSST Science Book, Version 2.0.  
<https://arxiv.org/abs/0912.0201>
- Lueckmann, J.-M., Bassetto, G., Karaletsos, T., & Macke, J. H. 2019, Likelihood-free inference with emulator networks. <https://arxiv.org/abs/1805.09294>
- Madireddy, S., Ramachandra, N., Li, N., et al. 2019, arXiv preprint arXiv:1911.03867
- Marshall, P. J., Treu, T., Melbourne, J., et al. 2007, ApJ, 671, 1196, doi: [10.1086/523091](https://doi.org/10.1086/523091)
- Millon, M., Galan, A., Courbin, F., et al. 2020, Astronomy & Astrophysics, 639, A101, doi: [10.1051/0004-6361/201937351](https://doi.org/10.1051/0004-6361/201937351)
- Oguri, M., & Marshall, P. J. 2010, MNRAS, 405, 2579, doi: [10.1111/j.1365-2966.2010.16639.x](https://doi.org/10.1111/j.1365-2966.2010.16639.x)
- Papamakarios, G., & Murray, I. 2016, Advances in neural information processing systems, 29
- Park, J. W., Wagner-Carena, S., Birrer, S., et al. 2021, The Astrophysical Journal, 910, 39, doi: [10.3847/1538-4357/abdfc4](https://doi.org/10.3847/1538-4357/abdfc4)
- Pearson, J., Li, N., & Dye, S. 2019, Monthly Notices of the Royal Astronomical Society, 488, 991
- Pearson, J., Maresca, J., Li, N., & Dye, S. 2021, Monthly Notices of the Royal Astronomical Society, 505, 4362–4382, doi: [10.1093/mnras/stab1547](https://doi.org/10.1093/mnras/stab1547)
- Perreault Levasseur, L., Hezaveh, Y. D., & Wechsler, R. H. 2017, The Astrophysical Journal Letters, 850, L7, doi: [10.3847/2041-8213/aa9704](https://doi.org/10.3847/2041-8213/aa9704)
- Poh, J., Samudre, A., Ćiprijanović, A., et al. 2022, Strong Lensing Parameter Estimation on Ground-Based Imaging Data Using Simulation-Based Inference.  
<https://arxiv.org/abs/2211.05836>
- Refsdal, S. 1964, Monthly Notices of the Royal Astronomical Society, doi: [10.1093/mnras/128.4.307](https://doi.org/10.1093/mnras/128.4.307)
- Schmidt, T., Treu, T., Birrer, S., & Shajib, A. J. e. a. 2023, Monthly Notices of the Royal Astronomical Society, 518, 1260–1300, doi: [10.1093/mnras/stac2235](https://doi.org/10.1093/mnras/stac2235)
- Schuldt, S., Cañameras, R., Shu, Y., et al. 2023a, Astronomy & Astrophysics, 671, A147
- Schuldt, S., Suyu, S., Meinhardt, T., et al. 2021, Astronomy & Astrophysics, 646, A126
- Schuldt, S., Suyu, S. H., Cañameras, R., et al. 2023b, Astronomy & Astrophysics, 673, A33, doi: [10.1051/0004-6361/202244534](https://doi.org/10.1051/0004-6361/202244534)
- Sérsic, J. L. 1968, Cordoba
- Shajib, A. J., Treu, T., Birrer, S., & Sonnenfeld, A. 2021, Monthly Notices of the Royal Astronomical Society, 503, 2380–2405, doi: [10.1093/mnras/stab536](https://doi.org/10.1093/mnras/stab536)
- Shajib, A. J., Birrer, S., Treu, T., et al. 2019, Monthly Notices of the Royal Astronomical Society, 483, 5649
- Suyu, S. 2012, Monthly Notices of the Royal Astronomical Society, 426, 868
- Tan, C. Y., Shajib, A. J., Birrer, S., et al. 2024, MNRAS, 530, 1474, doi: [10.1093/mnras/stae884](https://doi.org/10.1093/mnras/stae884)
- The LSST Dark Energy Science Collaboration, Mandelbaum, R., Eifler, T., et al. 2021, The LSST Dark Energy Science Collaboration (DESC) Science Requirements Document.  
<https://arxiv.org/abs/1809.01669>
- Treu, T., Koopmans, L. V., Bolton, A. S., Burles, S., & Moustakas, L. A. 2006, ApJ, 640, 662, doi: [10.1086/500124](https://doi.org/10.1086/500124)
- Treu, T., Suyu, S. H., & Marshall, P. J. 2022, The Astronomy and Astrophysics Review, 30, 8
- Verde, L., Treu, T., & Riess, A. G. 2019, Nature Astronomy, 3, 891, doi: [10.1038/s41550-019-0902-0](https://doi.org/10.1038/s41550-019-0902-0)
- Wagner-Carena, S., Aalbers, J., Birrer, S., et al. 2023, The Astrophysical Journal, 942, 75, doi: [10.3847/1538-4357/aca525](https://doi.org/10.3847/1538-4357/aca525)
- Wagner-Carena, S., Lee, J., Pennington, J., et al. 2024, A Strong Gravitational Lens Is Worth a Thousand Dark Matter Halos: Inference on Small-Scale Structure Using Sequential Methods. <https://arxiv.org/abs/2404.14487>
- Wagner-Carena, S., Park, J. W., Birrer, S., et al. 2021, The Astrophysical Journal, 909, 187, doi: [10.3847/1538-4357/abdf59](https://doi.org/10.3847/1538-4357/abdf59)

## APPENDIX

### A. TRAINING PRIOR

Our choice for the interim training prior is shown in Table 7. When training NPE for scientific application,  $\nu_{\text{int}}$  must be chosen carefully, such that any possible lensing configuration is included in the prior volume. The assumed model for each component is discussed in Section 2.1.

<b>Main Deflector</b>	$\theta_E(\text{''})$	$\mathcal{N}_{\text{trunc}}(\min = 0, \mu = 0.8, \sigma = 0.15)$
	$\gamma_{\text{lens}}$	$\mathcal{N}(\mu = 2.0, \sigma = 0.2)$
	$e_1, e_2$	$\mathcal{N}(\mu = 0.0, \sigma = 0.2)$
	$x_{\text{lens}}, y_{\text{lens}}(\text{''})$	$\mathcal{N}(\mu = 0, \sigma = 0.07)$
	$\gamma_1, \gamma_2$	$\mathcal{N}(\mu = 0, \sigma = 0.12)$
<b>Lens Light</b>	$m_{\text{app}}$	$\mathcal{N}_{\text{trunc}}(\min = 17, \max = 23, \mu = 20, \sigma = 2)$
	$R_*(\text{''})$	$\mathcal{N}_{\text{trunc}}(\min = 0, \mu = 1, \sigma = 0.8)$
	$n_*$	$\mathcal{N}_{\text{trunc}}(\min = 0.5, \mu = 3, \sigma = 2)$
	$e_1, e_2$	$\mathcal{N}_{\text{trunc}}(\min = -0.5, \max = 0.5, \mu = 0.0, \sigma = 0.2)$
	$x_{\text{ll}}(\text{''})$	$\mathcal{N}(\mu = x_{\text{lens}}, \sigma = 0.005)$
	$y_{\text{ll}}(\text{''})$	$\mathcal{N}(\mu = y_{\text{lens}}, \sigma = 0.005)$
<b>Source Light</b>	$m_{\text{app}}$	$\mathcal{N}_{\text{trunc}}(\min = 20, \max = 27, \mu = 23.5, \sigma = 1.7)$
	$R_*(\text{''})$	$\mathcal{N}_{\text{trunc}}(\min = 0, \mu = 0.5, \sigma = 0.8)$
	$n_*$	$\mathcal{N}_{\text{trunc}}(\min = 0.5, \mu = 3, \sigma = 2)$
	$e_1, e_2$	$\mathcal{N}_{\text{trunc}}(\min = -0.5, \max = 0.5, \mu = 0.0, \sigma = 0.2)$
	$x_{\text{src}}(\text{''})$	$\mathcal{N}(\mu = 0, \sigma = 0.1)$
	$y_{\text{src}}(\text{''})$	$\mathcal{N}(\mu = 0, \sigma = 0.1)$
<b>Point Source</b>	$m_{\text{app}}$	$\mathcal{N}_{\text{trunc}}(\min = 19, \max = 25, \mu = 22, \sigma = 2)$
	$x_{\text{ps}}(\text{''})$	$x_{\text{src}}$
	$y_{\text{ps}}(\text{''})$	$y_{\text{src}}$
	$f_{\text{microlensing}}$	$\mathcal{N}_{\text{trunc}}(\min = 0, \mu = 1.0, \sigma = 0.3)$

**Table 7.** Choice of the interim training prior,  $\nu_{\text{int}}$ . Note  $\mathcal{N}$  is a Gaussian, and  $\mathcal{N}_{\text{trunc}}$  is a truncated Gaussian. The main deflector is parameterized by a PEMD and external shear profile (see Section 2.1.1). Lens light and source light are parameterized by a Sérsic profile (see Section 2.1.2). We define the prior on lens light, source light, and point source brightness in apparent magnitude,  $m_{\text{app}}$ . We apply a fractional change to each point source image’s apparent magnitude,  $f_{\text{microlensing}}$ , to account for microlensing.

### B. VERIFICATION TESTS

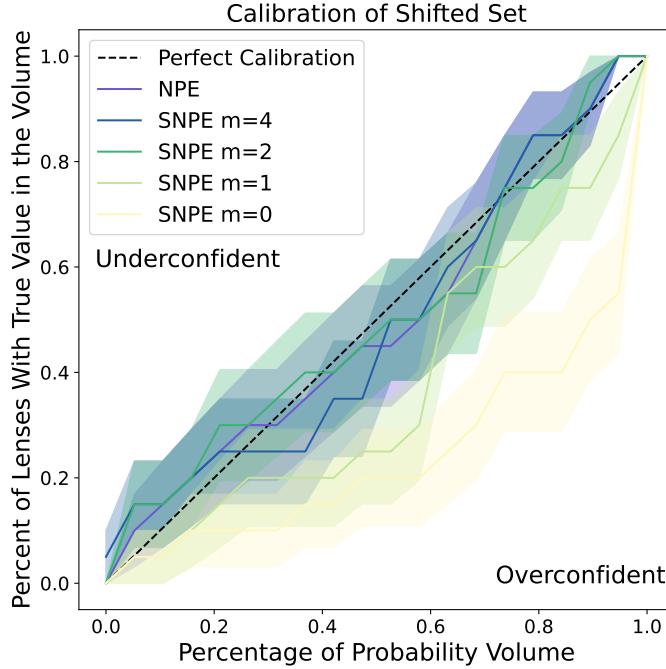
#### B.1. *Shifted Distribution*

To create the shifted set, we start with the distribution of our training prior, detailed in Table 7. Then, we shift and narrow the distribution in two key parameters:  $\theta_E$  and  $\gamma_{\text{lens}}$ . We choose the shifts to mimic the distribution shift we see in the doppelganger test set. We change the distribution for these two parameters as follows:

$$\theta_E \sim \mathcal{N}_{\text{trunc}}(\min = 0, \mu = 0.7, \sigma = 0.08) \quad (\text{B1})$$

$$\gamma_{\text{lens}} \sim \mathcal{N}(\mu = 2.05, \sigma = 0.1). \quad (\text{B2})$$

All other parameters are sampled in the same way as the training set, and the same simulator is used. We draw 20 lenses, both double and quad configurations, from this distribution to make the narrow test set.



**Figure 11.** Calibration curves for the shifted test set. In perfectly calibrated posteriors (dashed line), a given  $x\%$  of the probability volume contains the truth  $x\%$  of the time. Calibration of NPE posteriors is shown in blue-purple. Calibration of SNPE posteriors with different geometrically averaged proposals is shown in blue ( $m=4$ ), green ( $m=2$ ), yellow-green ( $m=1$ ), and yellow ( $m=0$ ). Note  $m=2$  is the choice made for this analysis. The shaded region encompasses  $1\sigma$  uncertainty.

### B.2. Doppelganger Simulations

The [STRIDES23](#) models have more complexity than what is included in our simulations. We simplify the models to match the complexity of our training simulations. For example, the [STRIDES23](#) models have double component lens light, but we only use single component lens light. So, we use only the bulge component for our doppelganger simulations. One key difference is that we ask the network to recover  $(x_{\text{src}}, y_{\text{src}})$ , but the [STRIDES23](#) models fit multiple lens plane image positions  $(x_{\text{img}}, y_{\text{img}})$  rather than a single source position. When making these doppelgangers, we take the image positions and the lens model, and solve for the source position using the numerical solver in [LENSTRONOMY](#). Note that when simulating the doppelganger of DES J0530–3730 with this procedure, we were unable to find a single source position solution that resulted in four point source images, so we exclude this lens from our doppelganger test set. Note that the [STRIDES23](#) lens model for DES J0530–3730 has a rare swallowtail configuration which is very sensitive to the source position.

## C. GEOMETRIC AVERAGING OF SNPE PROPOSALS

Given the training prior  $p(\xi_k)$  and the NPE approximate posterior  $q_\phi(\xi_k | d_k, \nu_{\text{int}})$ , we need to choose a proposal distribution  $\tilde{p}(\xi_k)$  to generate new training examples for SNPE. We investigate a proposal that uses a geometric average of the prior and the NPE approximate posterior:

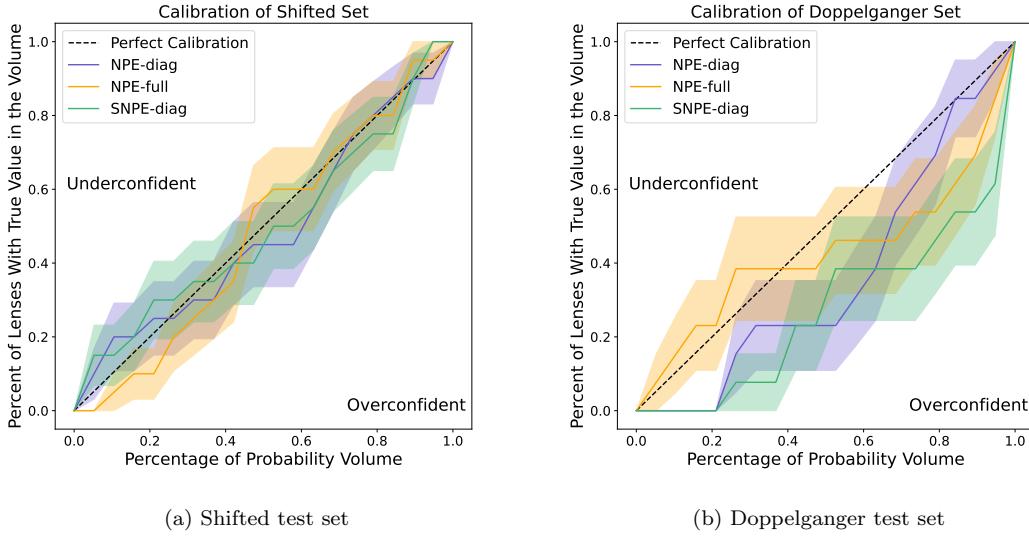
$$\tilde{p}(\xi_k) \propto (q_\phi(\xi_k | d_k, \nu_{\text{int}})^n p(\xi_k | \nu_{\text{int}})^m)^{\frac{1}{m+n}} \quad (\text{C3})$$

Parameter  $n$  is the weight of the posterior in the average, and  $m$  is the weight of the prior in the average. We keep  $n=1$ , but we allow  $m$  to take values  $[0, 1, 2, 4]$ .

We test different values of  $m$  on the shifted test set. We show the calibration of shifted set posteriors for different values of  $m$  in Figure 11. We choose  $m=2$  for our main analysis, since at this value we strike a balance of folding in as much information from the NPE prediction as possible without causing overconfident calibration.

#### D. FULL COVARIANCE NPE TESTS

We investigate NPE with full covariance matrices. The inference technique remains the same as what is described in Section 3.1, except the final layer of the neural network now outputs 10  $\mu_k$  and the 45 elements of the log-cholesky decomposition of the full covariance matrix,  $\Sigma_k$ . We ran all of our analysis metrics on the narrow test set and doppelganger test set with this technique. We summarize performance with calibration plots in Figure 12 and hierarchical inference plots in Figure 13. On verification tests, the full covariance NPE method is well behaved.



**Figure 12.** Calibration curves for the verification test sets with full covariance NPE included as a modeling option. Calibration from different methods is shown, with diagonal NPE (NPE-diag) shown in purple, full covariance NPE (NPE-full) shown in orange, and diagonal SNPE (SNPE-diag) shown in green. In perfectly calibrated posteriors (dashed line), a given x% of the probability volume contains the truth x% of the time. On the narrow test set, NPE-full calibration is consistent with the two other techniques. On the doppelganger test set, NPE-full calibration is closer to the perfect calibration curve, suggesting that inclusion of covariances in posteriors is crucial for the calibration of posteriors.

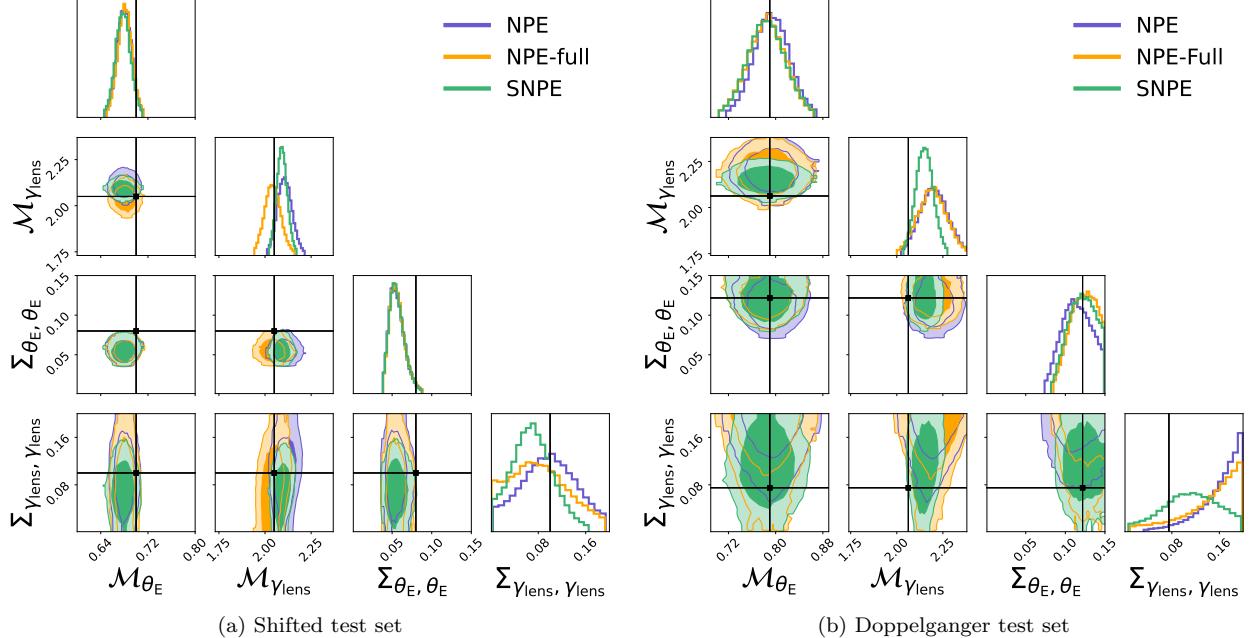
The next step was application of full covariance NPE to the real HST images. In the first step of this analysis, we found indications that this method is not fully robust to complexities in the HST images. For 4 out of the 14 lenses (DES J0420–4037, PS J1606–2333, SDSS J0248+1913, W2M J1042+1641), we found that posterior PDFs had  $\sigma_k(\gamma_{\text{lens}})$  larger than the  $\sigma_k(\gamma_{\text{lens}})$  of the interim prior. For a 5th lens (DES J0530–3730), full covariance NPE results in a posterior wider than the prior in 3 dimensions ( $\gamma_2, e_2, y_{\text{lens}}$ ).

Ultimately, full covariance NPE shows promise given its well behaved performance on the narrow test set and doppelganger test set. However, we hypothesize that full covariance NPE is more sensitive to details in the real HST data than the diagonal covariance NPE. Further work to increase the complexity of training simulations may be necessary before applying full covariance NPE on real data.

#### E. RE-WEIGHTING POSTERIORS

Interim posteriors  $q_\phi(\xi_k | d_k, \nu_{\text{int}})$  are influenced by the choice of hyperparameters  $\nu_{\text{int}}$  that define the interim training prior. Once we have learned the true population hyper-parameters via the hyperposterior  $p(\nu | \{d\})$ , we can use a re-weighting scheme to infer the individual lens posterior PDFs given the cPDF, which replaces the interim prior. This can also be thought of as using information from the population of lenses to refine the posterior of an individual lens. This re-weighting scheme is derived in Wagner-Carena et al. (2021), and ultimately we use the equation:

$$q_\phi(\xi_k | d_k) \propto q_\phi(\xi_k | d_k, \nu_{\text{int}}) \int \frac{p(\xi_k | \nu)}{p(\xi_k | \nu_{\text{int}})} p(\nu | \{d\}_{\neq k}) d\nu \quad (\text{E4})$$

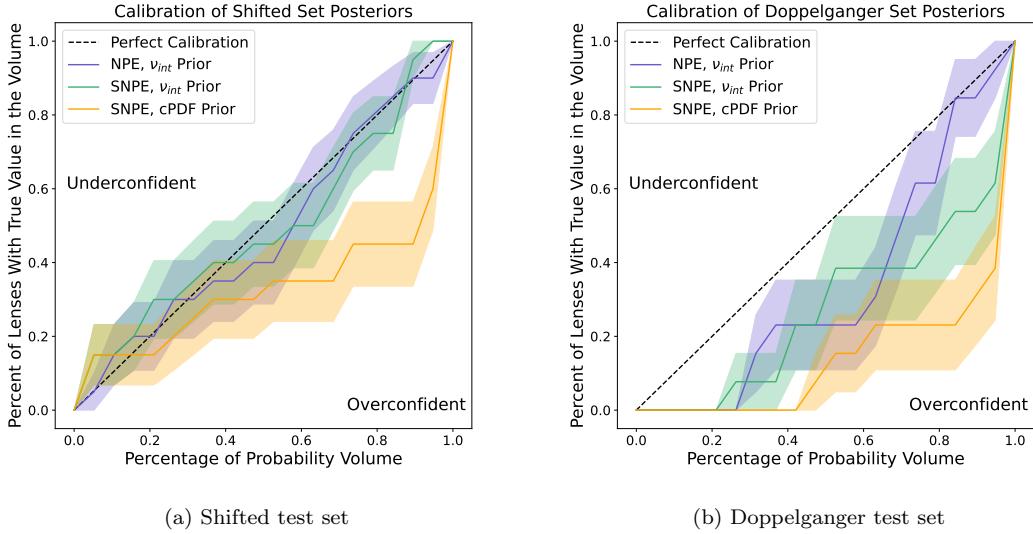


**Figure 13.** 2D contours of the hyperposterior  $p(\nu| \{d\})$  on the verification test sets. Inference from NPE with diagonal covariance posteriors is shown in purple. Inference from NPE with full covariance posteriors is shown in orange. Inference from SNPE with diagonal posteriors is shown in green. The ground truth is shown as a black line. Shaded contours are 68% and 95% intervals. Note for the doppelganger test set, we use the sample mean and standard deviation as a proxy ground truth.

Note that to use this scheme, we need to perform a population-level inference  $k$  times, excluding one lens at a time to generate  $p(\nu| \{d\}_{\neq k})$  for each lens. In practice, we compute the integral in Equation E4 using importance sampling:

$$q_\phi(\xi_k|d_k) \propto q_\phi(\xi_k|d_k, \nu_{\text{int}}) \frac{1}{N} \sum_{\nu \sim p(\nu| \{d\}_{\neq k})} \frac{p(\xi_k|\nu)}{p(\xi_k|\nu_{\text{int}})} \quad (\text{E5})$$

We applied this hierarchical re-weighting of individual posterior PDFs on the verification test sets. To re-weight the posteriors, we take samples from  $q_\phi(\xi_k|d_k, \nu_{\text{int}})$ , and assign each sample a weight by evaluating Equation E5. The resulting calibration curves are shown in Figure 14. We find that the re-weighted posteriors have worse calibration, with the posterior widths becoming too small. We have three hypotheses for the cause of this effect. First, this could be due to imperfect inference of the individual posteriors. Any underlying bias in the initial posteriors influences the population model, and is then folded back in, potentially amplifying any smaller biases. A second hypothesis is that the re-weighting scheme is under-sampled, and probability density cannot move into the tails of a distribution when it needs to. The third hypothesis is that the functional form we asserted for the conditional PDF was insufficiently expressive, leading to artificially strong constraints on the individual lens model parameters. This third hypothesis only holds for the doppelganger test set, since the shifted test set distribution is Gaussian by definition.



**Figure 14.** Calibration curves for the verification test sets with re-weighting. In perfectly calibrated posteriors (dashed line), a given x% of the probability volume contains the truth x% of the time. Calibration of interim NPE posteriors is shown in purple. Calibration of interim SNPE posteriors is shown in green. Interim posteriors are generated under the informative prior  $p(\xi_k | \nu_{\text{int}})$ . Calibration of re-weighted SNPE posteriors is shown in yellow. Re-weighted SNPE posteriors effectively have a different prior assumption, with the cPDF being the new prior. The shaded region encompasses  $1\sigma$  uncertainty. Re-weighting makes error calibration worse, in the direction of increasing overconfidence, on both test sets.