Draft version October 24, 2025
Typeset using LaTeX **twocolumn** style in AASTeX63

Lens Model Accuracy in the Expected LSST Lensed AGN Sample

Padmavathi Venkatraman [1,2] Sydney Erickson [2,3] Phil Marshall [2,3] Martin Millon [4,2] Philip Holloway [5,6] Simon Birrer [7] Steven Dillmann [2,8,9] Xiangyu Huang [7] Sreevani Jaragula [10] Ralf Kaehler,[2,3] Narayan Khadka [7,2,3] Grzegorz Madejski,[2,3] Ayan Mitra [11,1] Kevil Reil [3] Aaron Roodman [2,3] and the LSST Dark Energy Science Collaboration

[1] *Department of Astronomy, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA*
[2] *Kavli Institute for Particle Astrophysics and Cosmology, Department of Physics, Stanford University, Stanford, CA 94309, USA*
[3] *SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA*
[4] *Institute for Particle Physics and Astrophysics, ETH Zurich, Wolfgang-Pauli-Strasse 27, CH-8093 Zurich, Switzerland*
[5] *Department of Physics, University of Oxford, Keble Road, Oxford, UK*
[6] *Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth, UK*
[7] *Department of Physics, Stony Brook University, NY, U.S.A.*
[8] *Stanford Artificial Intelligence Laboratory, Department of Computer Science, Stanford University, Stanford, CA 94309, USA*
[9] *Stanford Institute of Computational and Mathematical Engineering, Stanford University, Stanford, CA 94309, USA*
[10] *Fermi National Accelerator Laboratory, Kirk Road and Pine Street, Batavia IL 60510-5011, USA*
[11] *Center for AstroPhysical Surveys, National Center for Supercomputing Applications, University of Illinois Urbana-Champaign, Urbana, IL, 61801, USA*

(Received; Revised; Accepted)

ABSTRACT

Strong gravitational lensing of active galactic nuclei (AGN) enables measurements of cosmological parameters through time-delay cosmography (TDC). With data from the upcoming LSST survey, we anticipate using a sample of $\mathcal{O}(1000)$ lensed AGN for TDC. To prepare for this dataset and enable this measurement, we construct and analyze a realistic mock sample of 1300 systems drawn from the OM10 (Oguri & Marshall 2010) catalog of simulated lenses with AGN sources at $z < 3.1$ in order to test a key aspect of the analysis pipeline, that of the lens modeling. We realize the lenses as power law elliptical mass distributions and simulate 5-year LSST i-band coadd images. From every image, we infer the lens mass model parameters using neural posterior estimation (NPE). Focusing on the key model parameters, $\theta_{\rm E}$ (the Einstein Radius) and $\gamma_{\rm lens}$ (the projected mass density profile slope), with consistent mass-light ellipticity correlations in test and training data, we recover $\theta_{\rm E}$ with less than 1% bias per lens, 6.5% precision per lens and $\gamma_{\rm lens}$ with less than 3% bias per lens, 8% precision per lens. We find that lens light subtraction prior to modeling is only useful when applied to data sampled from the training prior. If emulated deconvolution is applied to the data prior to modeling, precision improves across all parameters by a factor of 2. Finally, we combine the inferred lens mass models using Bayesian Hierarchical Inference to recover the global properties of the lens sample with less than 1% bias under consistent train-test mass-light correlations.

Keywords: Strong Gravitational Lensing, Neural Posterior Estimation, Bayesian Hierarchical Inference

1. INTRODUCTION

Corresponding author: Padma Venkatraman
vpadma23@slac.stanford.edu

Constraining the expansion of our Universe has been a long-standing challenge in cosmology. The two dominant probes measuring the current expansion rate of the Universe (Hubble Constant, H_0) from the early universe (such as CMB-BAO oscillations etc.) (e.g Abdalla et al. 2022) and late universe (such as Type 1a Supernovae etc.) (e.g Riess et al. 2022) grow more precise, yet dis-

crepant, leading to a 5σ difference on H_0 , known as the Hubble tension (e.g Verde et al. 2019). Additionally, using alternate methods to calibrate H_0 measurements from late-time probes like Type Ia SNe, for example, using the tip of the red giant branch (TRGB), yields consistent results with early and late time probes (e.g Freedman et al. 2019; Freedman et al. 2025). In this conjecture, it is crucial to test the systematics of different probes using independent competitive and complementary methods.

Time-delay cosmography (TDC) uses light travel times around massive deflectors to constrain angular diameter distances (Refsdal 1964). This technique has emerged as a competitive probe with independent systematics that can be used to help resolve Hubble tension (e.g Treu et al. 2022; Birrer et al. 2024).

TDC is applied to any system with a time-variable, bright, point-like source behind a massive deflector. H_0 has been measured to 2.4% precision with 7 lensed AGN (Wong et al. 2020; Millon et al. 2020). Non time-delay lenses are used to increase the precision of cosmological inferences made using time-delay lenses (e.g Birrer et al. 2020; TDCOSMO Collaboration et al. 2025), primarily by constraining the lens mass models. Surveys such as the Legacy Survey of Space and Time (LSST) (Ivezić et al. 2019), are expected to increase our cosmology grade time-delay sample size from ~ 40 lensed AGN (LAGN) (Treu et al. 2022) to $\mathcal{O}(100)$ LAGN with measured time delays and $\mathcal{O}(1000)$ LAGN in total (Oguri & Marshall 2010; Yue et al. 2022; Abe et al. 2025). In particular, LAGN brighter than 22 mag with variability detectable in LSST will be able to contribute to TDC cosmology constraints (Taak & Treu 2023).

Modeling the deflector galaxy's gravitational potential is a key component of the TDC analysis pipeline. We anticipate a "Gold" sample of LAGN in LSST that will consist of $\mathcal{O}(100)$ systems with all data required for high precision lens modeling including space-based imaging. The remaining $\mathcal{O}(1000)$ LAGN, constituting a "Silver" sample, will only have LSST imaging. We posit that combining the population constraints from this larger Silver sample with the Gold sample will improve the cosmological constraining power through an improved understanding of the mass distributions of deflector galaxies and a finer sampling of the distance-redshift relation. Within this framework, the inferred population-level distribution of lens mass model parameters from ground-based data becomes a key benchmark for our pipeline.

In this work we carry out LSST-scale modeling of ground-based imaging of simulated LAGN. Constructing mass models given an image of a strongly lensed system is a computationally expensive task. Current state of the art forward modeling techniques take $10^4 - 10^6$ CPU hours (Shajib et al. 2020), and several investigator hours (i.e. researcher analysis time), for a single lens system (Ding et al. 2021). With further automation as shown in Ertl et al. (2023); Schmidt et al. (2022) and Shajib et al. (2025), this is brought down to days, and in the best cases, hours. In the best-case scenario, GPU-accelerated models have reached modeling times down to 5 GPU hours per lens (e.g Huang et al. 2025; Galan et al. 2022).

In order to scale up to LSST, and model thousands of lenses within a reasonable analysis time, we turn to deep learning methods to infer lens mass models, as first introduced by Hezaveh et al. (2017); Perreault Levasseur et al. (2017). Previously, such machine learning methods have been validated on large samples of simulated space-based imaging data of LAGN in Park et al. (2021), and applied to real HST data of LAGN taken from STRIDES (Schmidt et al. 2022) in Erickson et al. (2025), achieving comparable accuracy to forward modeling. Similarly, they have been validated against simulated ground-based data of galaxy-galaxy lenses (G-G lenses) (e.g Pearson et al. 2019; Wagner-Carena et al. 2021; Poh et al. 2022), and tested on samples of space-based and ground-based imaging of G-G lenses (e.g Schuldt et al. 2021; Gawade et al. 2024; Gentile et al. 2023; Schuldt et al. 2023b; Euclid Collaboration et al. 2025). In this work, we verify the use of machine learning based lens modeling on a large realistic simulated sample of ground-based mock LSST imaging of LAGN.

We use a Simulation Based Inference (SBI) technique called Neural Posterior Estimation (NPE) (Papamakarios & Murray 2016; Cranmer et al. 2020) to produce the lens mass models. Using the open-source software `paltas` (Wagner-Carena et al. 2023)[1], we produce approximate multivariate Gaussian posterior probability distributions (PDFs) describing the lens model parameters learned from 5×10^5 mappings of images to lens model parameters. Once we have mass models for all lens systems, we combine them in a Hierarchical Bayesian Inference (HBI) framework to learn the properties of the sample. This process is outlined in Figure 1.

We aim to answer the following questions:

- How accurately can we measure the key lens model parameters $\theta_{\rm E}$ (the Einstein Radius) and $\gamma_{\rm lens}$ (the

[1] https://github.com/swagnercarena/paltas

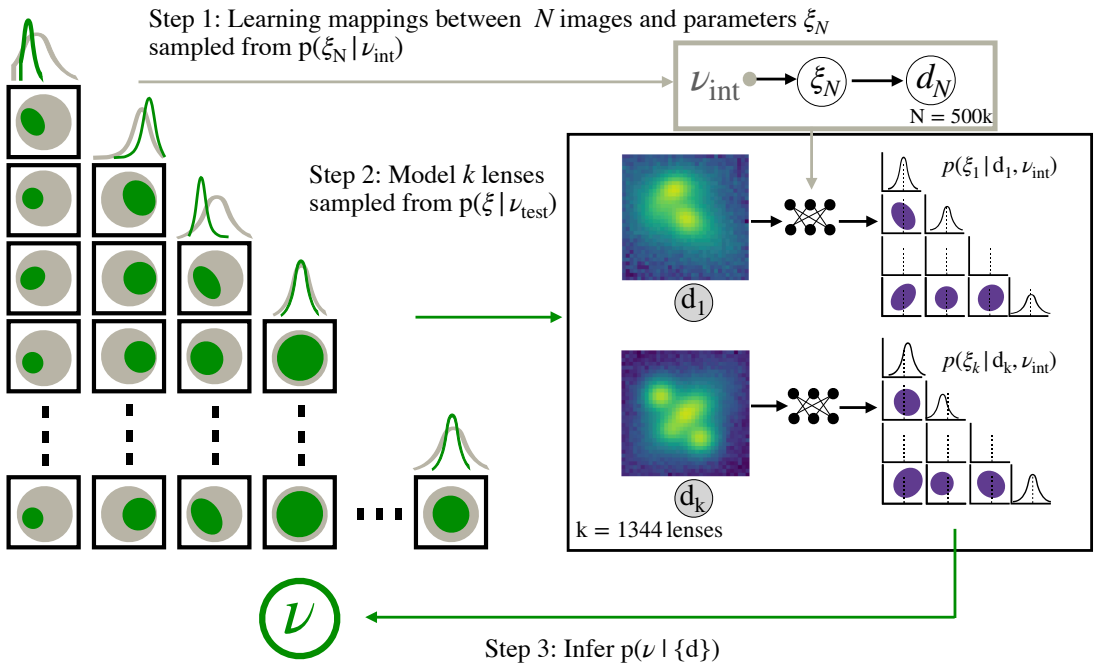

Figure 1. Probabilistic Graphical Model (PGM) depicting the levels of inference in this project. We assert broad priors on lensing parameters ν_{int} (shown in gray) and sample N = 500k lens systems with model parameters ξ_N. This is used to optimize model weights that we apply to the test image (d_k) for each of the k LAGN systems. From the data, we recover lens model parameters $p(\xi_k|d_k, \nu_{int})$ (shown in purple contours), and infer the lens population model $p(\nu|\{d\})$ (shown as overlaid green contours).

density profile slope) from the LSST 5-year coadd imaging alone?

- To what extent does this performance depend on the subtraction of the lens galaxy light, and the AGN point image light? How does deconvolution applied to the data prior to modeling improve the accuracy of measured lens models?
- How do distribution shifts in learned and latent parameters impact the inference of lens mass model parameters?
- How does the presence of mass-light ellipticity correlations in the training data impact the accuracy of inferred key lens model parameters, specifically γ_{lens} ?
- How accurately can we recover the population model for key lens mass model parameters?

This paper is organized as follows. Section 2 presents the expected LSST lensed AGN catalog constituting our sample, our lens modeling choices, and image data specifications. Section 3 describes the NPE method used to model lenses followed by the HBI inference pipeline used to infer the population. Section 4 presents the lens mass model parameter recovery results, on individual lenses and on the population-level parameters. Section 5 presents a discussion of the method and outlines the future investigation required. Section 6 presents our key conclusions.

2. DATA

To create a mock dataset of LAGN expected from LSST, we start from the forecasted lensed quasar catalog from Oguri & Marshall (2010, hereafter OM10), which contains lens systems with a deflector galaxy and background AGN. We detail the profile assumptions and modifications we make to the existing catalog in Section 2.1. We then perform a matching with the COSMODC2 mock galaxy catalog to assign an AGN host galaxy to each system, as described in Section 2.2. The end result is an updated catalog of 1344 LAGN. From this catalog, we create simulated LSST images with multiple "preparations," as detailed in Section 2.4.

2.1. *OM10 Lens Systems*

We start with the catalog of forecast strongly lensed AGN produced in OM10 [2]. The OM10 catalog assumes a singular isothermal ellipsoid (SIE) model for the lens mass distribution. It also contains AGN apparent brightness, K-correction and redshift information.

[2] https://github.com/drphilmarshall/OM10

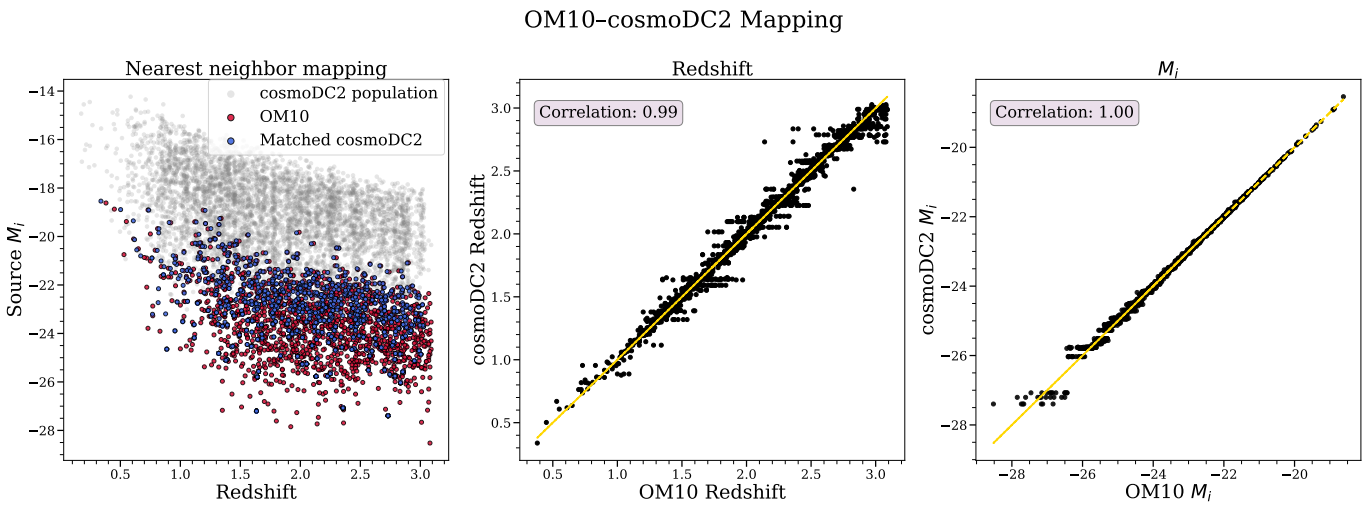

Figure 2. OM10-cosmoDC2 mapping. We "paint on" host galaxies from cosmoDC2 to systems in the OM10 catalog. Nearest neighbor mapping: To obtain a mapping between the two catalogs, we map AGN in OM10 to AGN in cosmoDC2 using their redshift and absolute magnitude and use their corresponding host galaxy. **First panel:** The red OM10 systems are mapped to the blue cosmoDC2 systems. The blue points are a subset of the gray points. **Second panel:** We demonstrate the scatter of the OM10-cosmoDC2 system mapping along AGN redshift axis. **Third panel:** We demonstrate the scatter of the OM10-cosmoDC2 system mapping along AGN i-band absolute brightness (AB mag) axis.

First, we make some cuts on the catalog to select systems observable by LSST. We require the image separation of the lensed systems to be greater 0.7 arcsec. This is a little less than the expected median PSF FWHM, or image quality, in the LSST coadd images. In Section 2.3, we revisit the image-separation cut after applying a realistic PSF to images of our modified lens models. We also require the brightness of third image of a quadruply imaged LAGN (or second image of a doubly LAGN) to have $i_3 < 23.3$ (to ensure that image is detected at 10-sigma significance, on average, in every visit image, Oguri & Marshall 2010). This results in 2818 LAGN systems selected from a large sample of 15658 LAGN systems. In the following sections, we lay out the changes we made to the lens mass, lens light, and source light profiles in order to make our mock lens systems more realistic.

2.1.1. *Lens Mass Model*

In OM10, the lens mass model is a singular isothermal ellipsoid (SIE). We re-parameterize the deflector galaxy as a power-law elliptical mass distribution (PEMD) (Barkana 1998). While SIE assumes a fixed projected radial density slope, PEMD allows the slope to vary. The lens modeling process is an inference of the convergence κ, or normalized surface mass density, of the deflector galaxy in a lens system using lensing observables. The lens model is combined with lensed source position to predict the Fermat potential, which is the quantity of interest for cosmology. Using the PEMD mass model, we can parameterize convergence as:

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

where $\gamma_{\rm lens}$ is the negative logarithmic projected mass density power-law slope and θ_E is the Einstein radius – the radius within which the deflector galaxy's integrated convergence is unity (assuming no ellipticity), x and y are defined in a coordinate system aligned with the major and minor axis of the lens, and $q_{\rm lens}$ is the minor/major axis ratio.

Instead of learning the axis ratio $q_{\rm lens}$ and orientation angle $\phi_{\rm lens}$, we parameterize the two-component ellipticity of the lens as:

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

This is because inference of periodic parameters, like $\phi_{\rm lens}$, is unstable.

We include external shear coming from the environment of the deflector and fit for these parameters. These parameters add an additional component to the deflection angle, but do not change the convergence κ.

We also change the parameterization of external shear from strength ($\gamma_{\rm ext}$) and orientation (ϕ_γ) to γ_1 and γ_2 where:

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

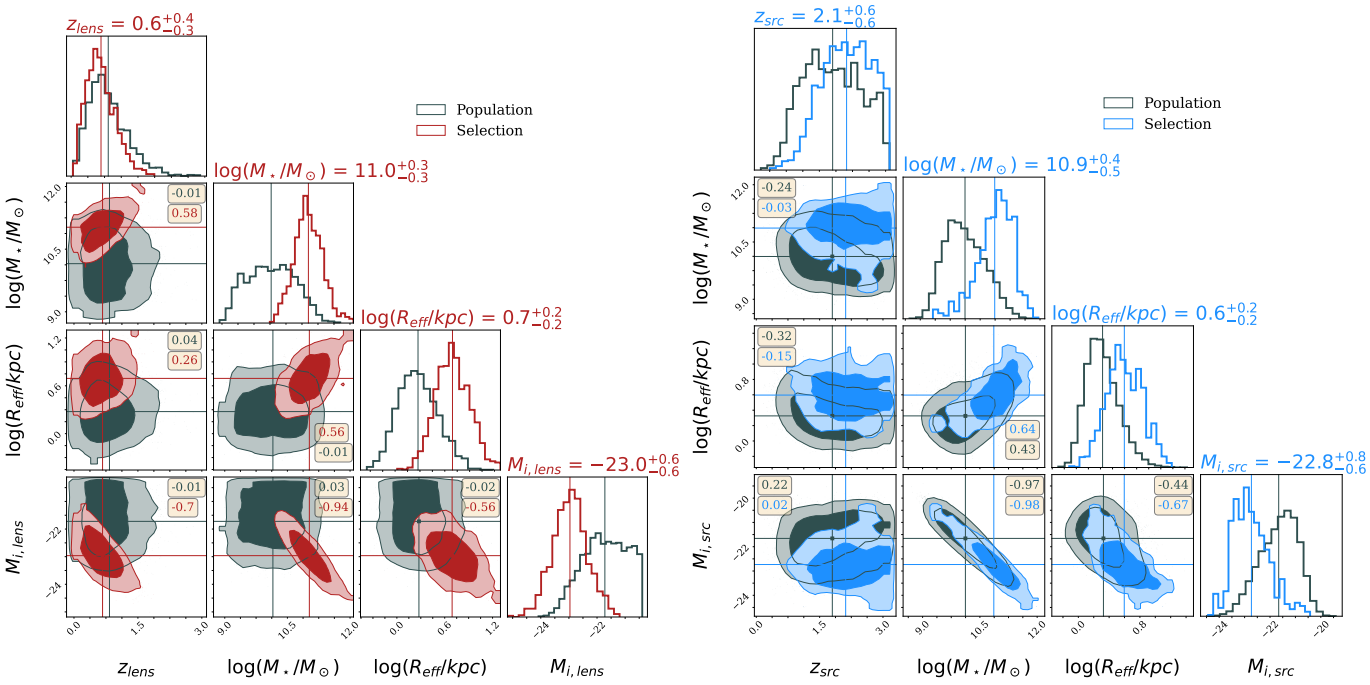

Figure 3. Deflector Galaxy and AGN Host Galaxy Selection. We display the following four galaxy properties: Redshift (z), stellar mass ($\log(M_*/M_\odot)$), effective size ($\log(R_{\rm eff}/{\rm kpc})$) and absolute brightness (M_i). **Left** – In gray we show properties of a sample of randomly selected galaxies from a cone. In red we show massive elliptical galaxies that constitute the lenses in our sample. The selection of galaxies is intrinsic: i.e more massive galaxies act as lenses. **Right** – In gray, we show the population of galaxies that contain an AGN. In blue, we show the AGN containing galaxies in our sample that are lensed by a massive foreground galaxy. The selection on AGN host galaxies in this case arises from cuts placed to ensure that these systems are detectable in LSST. Thus, this selection of AGN host galaxies in blue is observationally imposed, rather than intrinsic.

We obtain the following mass model parameters from OM10: θ_E, q, ϕ_e, γ, ϕ_γ, x, y. We randomly assign each deflector a density profile sampled from a Gaussian distribution centered on 2, with a scatter of 0.16 ($\gamma_{\rm lens} \sim {\rm N}(2, 0.16)$) because massive galaxies are approximately isothermal ($\gamma_{\rm lens} = 2$) with values ranging from 1.6 to 2.4 as shown in (Auger et al. 2010).

PEMD is often adopted in the context of TDC (Suyu et al. 2013) as it allows for flexibility in jointly constraining the density slope and H_0. When there is no arc information the density slope is degenerate with H_0. This is a manifestation of the mass-sheet degeneracy (MSD) detailed in Falco et al. (1985). This is constrained by kinematics information about the lensing galaxy or by knowing lensed AGN host galaxy properties such as brightness and size. (Birrer et al. 2020; Treu et al. 2022).

2.1.2. *Lens Light Model*

We model the deflector's surface brightness as an elliptical Sérsic profile with $\rm n_{sersic} = 4$ since a larger fraction of massive elliptical galaxies exhibit this profile (Caon et al. 1993). This is expressed as:

$$I_*(x,y) = I_* \exp\left[-b\left(\left[\frac{R}{R_*}\right]^{\frac{1}{\rm n_{sersic}}} - 1\right)\right], \quad \text{where } {\rm R} = \sqrt{{\rm qx}^2 + {\rm y}^2/{\rm q}}. \tag{4}$$

where I_* is the initial amplitude, R is the elliptical radius, b scales the equation so that R_* is the effective radius of the light profile, q is the axis ratio of the lens light and x, y are the galaxy position.

For the light profile of deflector galaxies in our catalog, we start with the spherical light profile in OM10 then introduce ellipticity. We do this by mapping OM10 deflectors to random galaxies sampled from a cone in the synthetic sky catalog, cosmoDC2, produced by LSST-DESC (Korytov et al. 2019, hereafter cosmoDC2). For a given galaxy in OM10, we find the nearest neighbor in the cosmoDC2 catalog by mapping along absolute magnitude, redshift and light axis ratio axes. This matching process results in deflector galaxies whose mass and light are exactly aligned, but which have a small scatter between the axis ratio of the lens mass and light. We show the lens sample properties in Figure 3, left panel.

2.1.3. *Source Light Model*

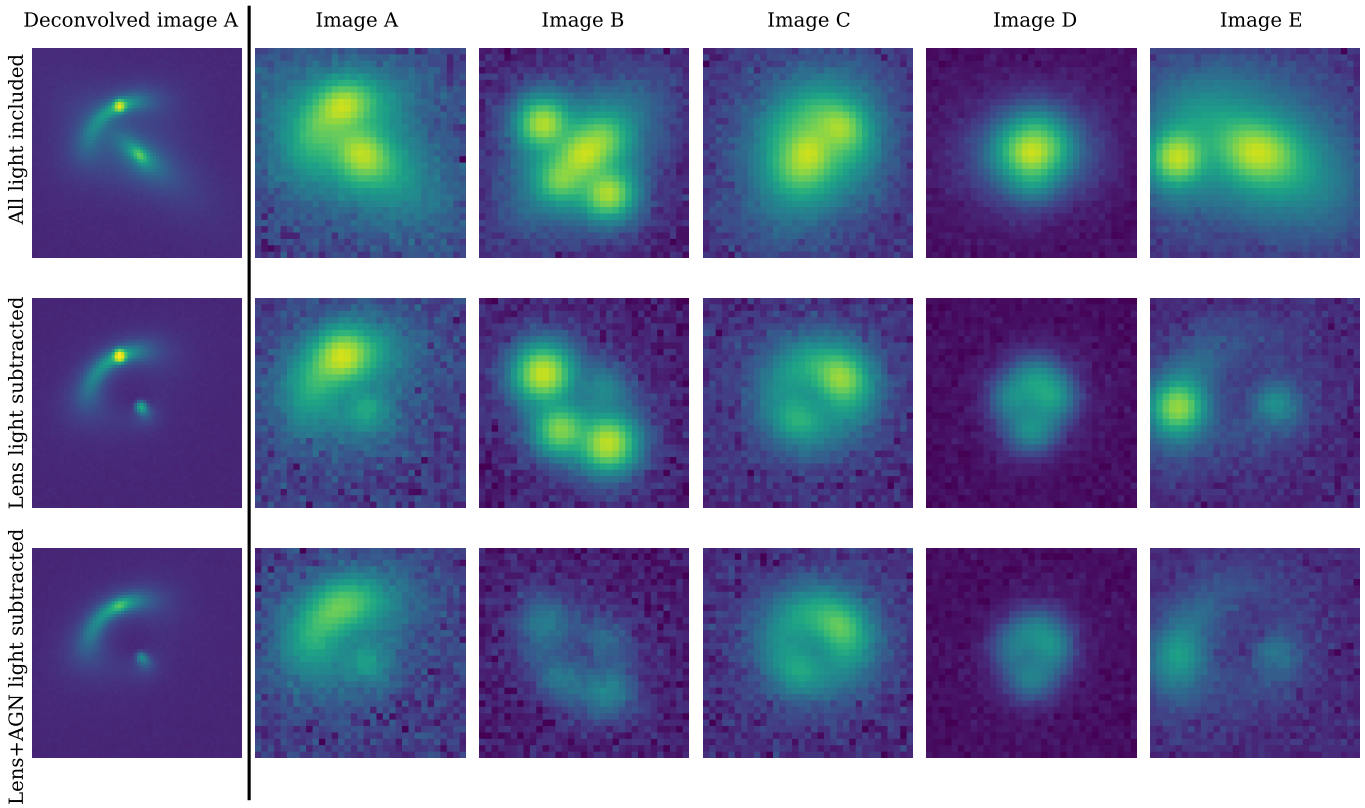

Figure 4. 5 simulated LSST lensed quasar systems in the i-band. We test our inference pipeline on 3 different preparations of the data. Subtraction of different light components have been shown to improve the performance of the modeling pipeline in lens parameter recovery. We discuss this further in Section 2.4. In the first column (to the left of the black line), we show the deconvolved version of the second column of images (to the right of the black line). Deconvolution also helps in the lens modeling, detailed also in Section 2.4. **Top row:** include lens light, AGN light and host galaxy light. **Middle row:** Lens light subtracted. Poisson noise due to lens light remains. **Bottom row:** Lens light and AGN light subtracted. Poisson noise due to all light remains. The background Gaussian noise also remains constant across 3 different preparations.

Source light consists of light from the AGN and the AGN's host galaxy. The AGN is modeled as a point source. For the AGN host galaxy, we assume that the majority of its light comes from the bulge component of the galaxy.

2.2. *Painting on AGN Host Galaxies*

In the OM10 catalog, we need to assign each lens systems with realistic AGN host galaxy properties. To do this, we map AGN properties (i.e. AGN redshift and brightness) provided in the OM10 catalog to AGN in cosmoDC2, as shown in Figure 2, and use their corresponding host galaxies which are decomposed into bulge and disk components. The quasar-host relation in cosmoDC2 is implemented as follows (reproduced for completion): the mass of the black hole ($M_{\rm BH}$) at the center of a galaxy and the Eddington rate ($\lambda_{\rm edd}$) are related to the bulge mass of the galaxy as follows (Equation 15 and 16, Section 5.4.2, Korytov et al. (2019)):

$$M_{\rm BH} = 0.0049 M_{\rm bulge}(M_{\rm bulge}/M_0)^{\alpha},$$
$$P(\lambda_{\rm edd}|z) = 0.00071\frac{(1+z)}{(1+z_0)^{\gamma_z}}\lambda_{\rm edd}^{\gamma_e}. \quad (5)$$

where $M_0 = 10^{11}M_{\odot}, \alpha = 0.15, \gamma_z = 3.47, \gamma_e = -0.65, z_0 = 0.6$, and different power-law relations are empirically obtained to account for variation in $\lambda_{\rm edd}$ as a function of redshift. The i-band magnitude (M_i) of the AGN is then assigned using the empirical relation between M_i, $\lambda_{\rm edd}$ and M_{BH} presented in Figure 15, MacLeod et al. (2010).

We only model light from the bulge as this is the dominant component, and model the bulge with a Sérsic surface brightness profile shown in Equation 4 with $n_{\rm sersic} = 4$. We display the selection on some important quantities in Figure 3. The bulges of spiral galaxies generally exhibit Sérsic surface brightness profiles with a typical Sérsic index of 4, which is the value adopted in cosmoDC2. We use the half-light radius for the galaxy's bulge, taking the sum of the effective radii along the

Table 1. Assumed LSST Camera i-band image specifications

Property	Value
Pixel scale (arcsec/pixel)	0.2
CCD gain (electron/ADU)	2.3
Instrumental noise/exposure	9
i-band instrumental zeropoint	28.17
Sky brightness (mag/arcsec^2)	20.46
Exposure time (s)	15
Number of Exposures	150 [5 year-coadd]

Note—We describe each property in Section 2.3. Values are taken from dp0.2 [a].

[a] https://smtn-002.lsst.io/

major axis and minor axis in quadrature. We show the AGN host galaxy sample properties in Figure 3, right panel.

2.3. *LSST Images*

We take the lens models described above and produce LSST-like images using them. In addition to the lens models, we provide LSST observation specifications shown in Table 1 to `lenstronomy` (Birrer & Amara 2018; Birrer et al. 2021) which computes the lensing image positions and convergence following our model assumptions. To simulate the whole catalog of lenses we use `paltas` (Wagner-Carena et al. 2023) (described further in Section 3.1) which also builds on `lenstronomy`. All images have 33×33 pixels to cover 6.6 arcsec on a side, given the LSST pixel scale of 0.2 arcsec. All images are simulated in LSST i-band, and we defer multi-band modeling to future work (Section 5.8). All noise properties and zeropoints pertain to 5-year coadd i-band imaging, shown in Table 1.

Ground-based images are heavily impacted by the atmosphere which blurs the light we receive on the telescope and this effect is captured in the PSF. We use a kernel-based point spread function (PSF) extracted from LSST simulated images in Data Preview 0.2 (DP0.2). After we realize the lensed AGN systems in our modified catalog described in Section 2.1 as LSST images, we eliminate systems that formed only a single image or unresolved images. While our initial cut simply removed systems with image separation < 0.7, changing the mass model, and PSF model improves the simulation and places a stronger cut on eligible systems. This brings our sample from 1557 lenses down to 1344 lenses.

2.4. *Image Preparations*

In order to understand how lens modeling performance depends on the preparation of image data, we simulate the catalog data under three different preparations (see Image 4).

Under the hood of `paltas`, use `lenstronomy`'s Image Simulation[3] module to make all three preparations and selectively include the different LAGN light components in each.

First, for our fiducial test set, we simulate images that include all light from all components of a lensed AGN system: deflector, lensed AGN, and lensed AGN host galaxy.

In a second test set, we emulate perfect lens light subtraction, including Poisson noise from all components but omitting the surface brightness of the deflector itself. De-blending and removing the lens light makes the point sources and arcs more visible, which can help in inferring the structure of the lensing system. We are motivated by previous works that indicate that removing lens light is beneficial for lens parameter recovery (e.g Pearson et al. 2019; Madireddy et al. 2022).

In a third test set, we assume subtraction of both the lens light and the AGN point source light. This last preparation aims to highlight information coming from the lensed arc of the AGN host galaxy which constrains model parameters like the density slope of the deflector, $\gamma_{\rm lens}$ (Birrer 2021). AGN point source light can remove arc information, since a) blending between the AGN point source and the lens reduces the contrast of the lensed arcs, and b) the high Poisson noise due to the bright AGN lowers the SNR of fainter arcs.

In all three preparations described above, the background Gaussian noise and Poisson noise due to flux from the deflector, AGN, and AGN host galaxy remain in the image, as it would in real data. These different preparations are illustrated in Figure 4.

Beyond the three preparations, we are also interested in the impact that image data quality has on the the NPE modeling of lensed AGN images. We expect to see that with better data quality, all features of the system are more visible and separable leading to better modeling results. One of the key parameters of interest is the projected mass density slope $\gamma_{\rm lens}$, and this parameter can be constrained with high-quality image data with sharp arc features. Obtaining higher-resolution images of a system we have ground based data for is possible through dedicated follow-up through a space-based telescope, or through deconvolution. In this project, we emulate the deconvolution of our LSST images, and study

[3] lenstronomy/ImSim GitHub link

an additional three preparations (deconvolved-fiducial, deconvolved-lens light subtracted and deconvolved-lens and AGN light subtracted) shown in column 1, Figure 4. We emulate the deconvolution process by simulating images with a Gaussian PSF with 2 pixel FWHM, at 0.1"/pixel (subsampled by a factor of 2 compared to the original data). This is motivated by recent successful application of techniques such as STARlet REgularized Deconvolution (STARRED Millon et al. 2024).

3. METHOD

Our goal is to measure lens mass models from 1344 simulated ground-based images of lensed AGN, and then recover sample properties of the 1344 lenses. We are primarily interested in learning the following lens mass model parameters used in Equation 1: Einstein radius (θ_E), density profile slope (γ), external shear parameters (γ_1 & γ_2), mass ellipticity parameters (e_1 & e_2), lens mass center ($x_{\rm lens,M}$ & $y_{\rm lens,M}$). When only the arc information is retained in the images with sufficient signal (in cases where the lensed AGN host galaxy is visible after lens and AGN subtraction), we also learn extended source parameters: apparent magnitude of source (m_i) and source effective radius or half light radius ($R_{\rm eff}$).

We use a SBI method called Neural Posterior Estimation (NPE) to measure individual lens mass model posterior PDFs from LSST images, outlined in Section 3.1. We use Hierarchical Bayesian Inference (HBI) to learn the underlying distribution the lenses were sampled from, detailed in Section 3.3.

3.1. *Individual Lens Mass Models: Neural Posterior Estimation*

In this section, we outline how we use Neural Posterior Estimation (NPE) to obtain the posterior PDF, represented as $p(\xi_k|d_k)$, for the PEMD lens mass model parameters ξ_k, given an image of the lensed AGN, d_k.

Deep convolutional neural networks (CNNs) can be used to map images to features. NPE guides the CNN output features to approximately describe a posterior PDF for the parameters listed above – an input image d_k, is mapped to an output $q_\phi(\xi_k|d_k)$, where ϕ represents the model weights (Papamakarios & Murray 2016; Cranmer et al. 2020). We assume $q_\phi(\xi_k|d_k)$ to be a multivariate Gaussian with a full covariance matrix over model parameters. For each image d_k, the network returns parameter means (μ) and the Cholesky factor (L) of the covariance matrix (L: lower triangular matrix such that $LL^T = \Sigma$) defining the posterior PDF.

The NPE output $q_\phi(\xi_k|d_k)$ is an approximation of the true lens model posterior PDF, $p_\phi(\xi_k|d_k)$. The multivariate Gaussian posterior distribution q_ϕ is shifted, made narrow or broader along different parameter axes in order to minimize the Kullback-Leibler (KL) divergence between the approximate posterior (q_ϕ) and the true posterior (p_ϕ). As shown in Papamakarios & Murray (2016), this corresponds to minimizing the loss below:

$$L(\phi) = -\sum_{k=1}^{N} \log q_\phi(\xi_k|d_k). \quad (6)$$

In the limit of infinite training data, given an expressive network and sufficiently flexible functional form of q_ϕ (Papamakarios & Murray 2016), $q_\phi(\xi_k|d_k) \approx p_\phi(\xi_k|d_k)$.

3.2. *Network Training*

In the following sections, we elaborate on the training data and procedure used to construct individual lens mass model posteriors using NPE.

3.2.1. *Training Data*

To learn the posterior PDF of lens mass model parameters from an image, the estimator q_ϕ requires multiple and repeated exposures to image-parameter mappings. We sample from $N = 5 \times 10^5$ lensing configurations from the broad training prior ($\xi_N \sim p(\xi_N|\nu_{\rm int})$) and simulate an image for each configuration. Here, $\nu_{\rm int}$ refers to hyperparameters governing the broad training prior, and $p(\xi_N|\nu_{\rm int})$ is the training distribution. We present this distribution for each of the model parameters and for latent parameters that we do not infer (such as lens apparent magnitude) in Table 2. These image-parameter mappings ($d_N; \xi_N$) becomes our training data.

We use `paltas` to rapidly sample the training prior, reject parameter configurations that do not form detectable lensing systems and simulate the training images. `paltas` uses `lenstronomy` simulation code to make the images, so our train and test images are produced in a consistent way.

Choice of the training prior is important because the quality of lens parameter inference relies on what the network sees during training. The training prior distribution is chosen to be wider than the test distribution for all learned and latent parameters. For certain parameters, the training distribution is truncated at limits beyond which the system is unphysical. The only physical correlation we include in our training data is between the lens mass and lens light ellipticity as shown Table 2. Further discussion on realism of training distribution and including physical correlations between parameters in the training data is available in Section 5.3.

We generate 6 sets of 5×10^5 images corresponding to the image preparations described in Section 2.4. To

Table 2. Training Distribution

Parameter	Distribution
Lens Mass Model: PEMD + External Shear	
Einstein Radius (")	$\theta_E \sim \mathcal{N}_{tr}(\sigma_{low} = -0.5, \sigma_{high} = \infty, \mu = 0.8, \sigma = 1)$
Density Profile Slope	$\gamma_{\rm lens} \sim \mathcal{N}(\mu = 2, \sigma = 0.2)$
External Shear	$\gamma_{1,2} \sim \mathcal{N}(\mu = 0, \sigma = 0.1)$
Mass Ellipticity ($e_{\rm mass}$)	$e_{1,2} \sim \mathcal{N}(\mu = 0, \sigma = 0.2)$
Lens Mass position (")($xy_{\rm mass}$)	$x_{\rm lens,M}, y_{\rm lens,M} \sim \mathcal{N}(\mu = 0, \sigma = 0.06)$
Lens Light Model: Single Sersic Ellipse	
Light Ellipticity ($e_{\rm light}$)	$e_{1,2} \sim e_{\rm mass} + 0.05 * N(0, 1)$
Lens Light position	$x_{\rm lens,L}, y_{\rm lens,L} \sim xy_{\rm mass} + 0.001 * N(0, 1)$
Sersic Index	$n_{\rm lens} \sim \mathcal{N}(\mu = 4, \sigma = 0.005)$
Effective Radius (")	$R_{\rm lens} \sim \mathcal{N}_{tr}(\sigma_{low} = -0.5, \sigma_{high} = \infty, \mu = 0.7, \sigma = 1)$
Apparent Magnitude (AB mag)	$m_i \sim \mathcal{N}(\mu = 20.5, \sigma = 2)$
Source Light Model: Single Sersic Ellipse	
Light Ellipticity (e_{src})	$e_{1,2} \sim \mathcal{N}(0, 0.1)$
Source Light position (")	$x_{\rm src}, y_{\rm src} \sim \mathcal{N}(0, 0.4)$
Sersic Index	$n_{\rm src} \sim \mathcal{N}(\mu = 4, \sigma = 0.001)$
Effective Radius(")	$R_{\rm lens} \sim \mathcal{N}_{tr}(\sigma_{low} = -0.5, \sigma_{high} = \infty, \mu = 0.7, \sigma = 1)$
Apparent Magnitude (AB mag)	$m_i \sim \mathcal{N}(\mu = 24, \sigma = 2)$

Note—This table contains the parameter distributions we sampled from to generate the 500k training images. All parameters are described in Section 2.1. $x_{\rm lens,M}$ refers to the lens mass x-position, and $x_{\rm lens,L}$ is the lens light x-position, and same notation follows for y-position. γ_1, γ_2 refers to the Cartesian shear components and e_1, e_2 refer to Cartesian ellipticity components introduced in Section 2.1.1. $\mathcal{N}$ is the Normal distribution and $\mathcal{N}_{tr}$ is the Truncated Normal distribution truncated at σ_{low} and σ_{high}.

generate and store such large datasets, we utilize CPU compute and memory allocation at the National Energy Research Scientific Computing Center (NERSC)[4].

3.2.2. *Training Procedure*

We uses a CNN based on xResNet-34 architecture (He et al. 2015) to model image data and extract lensing parameters. All training is done using GPU (NVIDIA A100) resources at NERSC. We start with a learning rate of 1e-4 and use an exponential decay schedule. We apply random rotations to the images, and use a fixed seed of 2 for all training runs. We train for 200 epochs, revisiting the data with batches of 1024 randomly sampled images. The network training loss shown in is optimized using the gradient-descent optimizer Adam (Kingma & Ba 2017). We use early stopping, defined as terminating training if the validation loss does not improve for 10 consecutive epochs, to select the model weights. We find that this consistently yields more calibrated and robust results compared to picking model weights with the lowest validation loss.

Our test set consists of expected $k = 1344$ LSST lensed AGN sample described in Section 2. Applying the selected model weights to this dataset yields a multivariate gaussian posterior describing the lens mass model for each image. The selected weights are optimized using $d_N; \xi_N$ pairs sampled from the broad training prior $\nu_{\rm int}$ – this means that the lensing parameters we learn from images are implicitly conditioned on $\nu_{\rm int}$. Thus the posterior PDF we infer from images is $\mathrm{p}(\xi_{\rm k}|\mathrm{d_k}, \nu_{\rm int})$. This corresponds to Step 2 in Figure 1. We compare the means and widths of the predicted posterior PDF to the known ground truth in order to assess the measurement accuracy. We repeat this experiment for each of the image preparations outlined in Section 2.4.

3.3. *Sample-level property inference: Hierarchical Bayesian Inference*

Once we have lens mass models for all the images in the expected LSST lensed AGN sample of $\mathcal{O}(1000)$ lenses, we can infer sample properties. Sample information can be folded into hierarchical analysis to infer

[4] NERSC: https://nersc.gov/

cosmology parameters as shown in TDCOSMO-IV (Birrer et al. 2020).

We are interested in inferring the distribution of lensing parameters i.e. $p(\xi_k|\nu)$. This is the conditional PDF or cPDF that we use to model the lens parameter distribution. We can infer the hyperparameters ν that govern the true underlying distribution given a sample of LSST images, $\{d\}$, i.e. $p(\nu|\{d\})$ using hierarchical Bayesian inference (e.g Erickson et al. 2025).

To infer properties of ν, we start with Bayes' Theorem. We assume each lens provides an independent constraint, and thus multiply the likelihood from each modeled lens k:

$$p(\nu|\{d\}) = p(\nu) \prod_k \int \frac{p(d_k|\nu)}{p(\{d\})}. \quad (7)$$

We elaborate this expression to include lens parameter information pertaining to each system i.e. ξ_k, and then marginalize over ξ_k, as shown:

$$p(\nu|\{d\}) = p(\nu) \prod_k \int d\xi_k \frac{p(d_k|\xi_k,\nu)p(\xi_k|\nu)}{p(\{d\})}. \quad (8)$$

We have access to the lens model posterior, $p(\xi_k|d_k,\nu_{int})$, for each lens k, from the NPE modeling. We note that the lens parameters ξ_k, is conditioned on both the image, d_k, and the previously assumed model during training ν_{int} (as detailed in Section 3.2.1). The inferred lens models are conditioned on the model assumed during training ν_{int}, and not the actual population model ν. Therefore, we need to rewrite the posterior using the following rearrangement:

$$p(\xi_k|d_k,\nu_{int}) = \frac{p(d_k|\xi_k,\nu_{int})p(\xi_k|\nu_{int})}{p(d_k|\nu_{int})}. \quad (9)$$

Rearranging Equation 9, and inserting it into Equation 8, we find that:

$$p(\nu|\{d\}) = p(\nu) \prod_k \int d\xi_k \; p(\xi_k|d_k,\nu_{int}) \times \frac{p(d_k|\nu_{int})}{p(\{d\})} \frac{p(\xi_k|\nu)}{p(\xi_k|\nu_{int})}. \quad (10)$$

In Equation 10, the last term in the integrand, $\frac{p(\xi_k|\nu)}{p(\xi_k|\nu_{int})}$, serves to divide out the interim prior and replace it with the cPDF, which can be thought of as the prior we would like to have assigned in the first place. Formalism for this is also presented in (e.g Wagner-Carena et al. 2021; Foreman-Mackey et al. 2014). We infer ν by adapting on scripts in `lens-npe` (Erickson 2024) which builds on the MCMC package `emcee`. We can now return to distribution of lensing parameters given that we know the true underlying distribution, i.e. $p(\xi_k|\nu)$, the cPDF. We model the cPDF as a multivariate Gaussian distribution with a diagonal covariance matrix. We start with a simple cPDF to study the available precision of the hyperparameters under such a simple assumption as well as any bias incurred by it. In future work, we will investigate more expressive cPDFs and population model. Given our population model, we test how accurately we can recover the test distribution that the images were sampled from. We elaborate on this in Section 4, and Figure 10.

In order to infer cosmological parameters such as H_0 correctly, we must infer H_0 and ν jointly as is done in Birrer et al. (2020). Biases on particular parameters like the population mean of the density profile slope will translate to a bias on Fermat potential, and thus H_0.

4. EXPERIMENTS AND RESULTS

In this section, we present lens modeling results we obtain from the NPE method and population-level model inference.

For all tests, including the fiducial (Section 4.1), we show results with and without lens light subtraction (details on image preparation in Section 2.4). After running the fiducial, we construct experiments to investigate the performance drivers. First, we test the input images to the method. We investigate the selection of systems with bright host galaxies (Section 4.2), and the resolution of the images (Section 4.3). Second, we investigate our method. We try removing mass-light correlations in training data (Section 4.4), and removing distribution shift between the test and training data (Section 4.5). We are consistently interested in recovery of θ_E and γ_{lens} , which are the easiest and most difficult parameter to measure from imaging alone Erickson et al. (2025). We present NPE recovery for all experiments for these parameters in Table 4, and population-level parameter recovery in Table 5. Further, we present NPE posterior recovery of all parameters in Appendix D in Table 7, and population-level parameter recovery in Table 6.

4.1. *Fiducial Results*

In this section, we present results on test data constructed from the OM10-cosmoDC2 catalog created in Section 2.1. We simulate LSST 5-year coadd images in the i-band with all light included (Section 2.4). These data are selected as the baseline, since they will not require any additional processing beyond survey data products. First, we assess the inference of mass model posteriors for individual lens parameters. We use the following metrics:

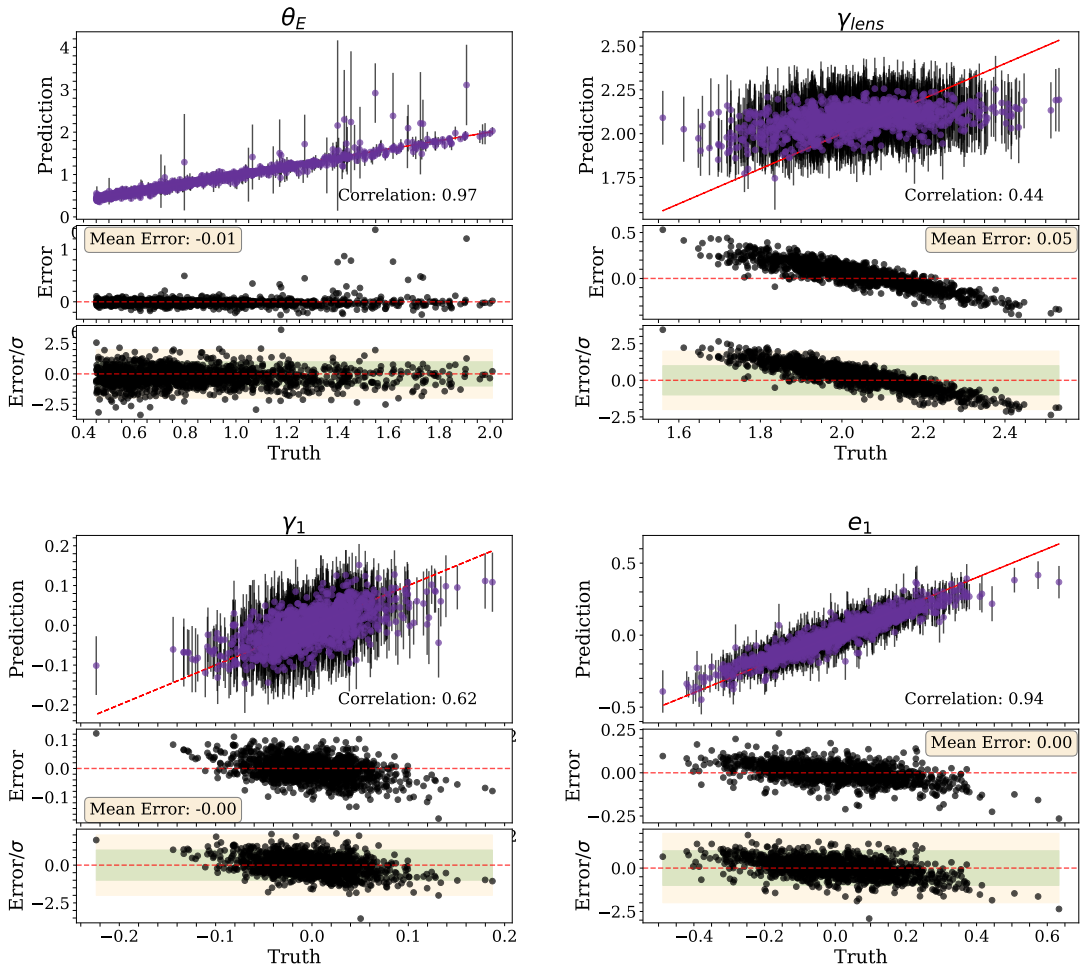

Figure 5. Comparison of 1344 predicted Gaussian posteriors (obtained as described in Section 3.1) for fiducial data (refer: Section 2.3) against ground truth. Purple points correspond to the mean of the posteriors, and black lines correspond to the posterior $1\,\sigma$ width or standard deviation. In the bottom most panel of each parameter subplot, the green shaded region represents the $1\,\sigma$ area and and yellow shaded region represents the $2\,\sigma$ region.

Table 3. Individual lens mass model posterior results on fiducial test set

Preparation	Metric	θ_E	$\gamma_{\rm lens}$	e_1	e_2	γ_1	γ_2	$x_{\rm lens}$	$y_{\rm lens}$
		(arcsec)						(arcsec)	(arcsec)
All Light [Fiducial]	ME	-0.01	0.05	0.	-0.	-0.	-0.	-0.	-0.
	MAE	0.02	0.1	0.03	0.03	0.02	0.02	0.01	0.01
	Precision	0.05	0.16	0.09	0.09	0.06	0.06	0.02	0.02
Lens Light Subtracted	ME	-0.01	0.07	0.	-0.	0.01	-0.	-0.	-0.
	MAE	0.02	0.1	0.06	0.06	0.03	0.03	0.02	0.02
	Precision	0.04	0.16	0.13	0.13	0.07	0.06	0.04	0.04

Note—Fiducial experiment results, outlined in Section 4.1. ME = Average deviation of posterior means from truth, indicator of bias. MAE = Median absolute error of posterior means from truth, indicator of scatter. Precision = Median uncertainty of posteriors.

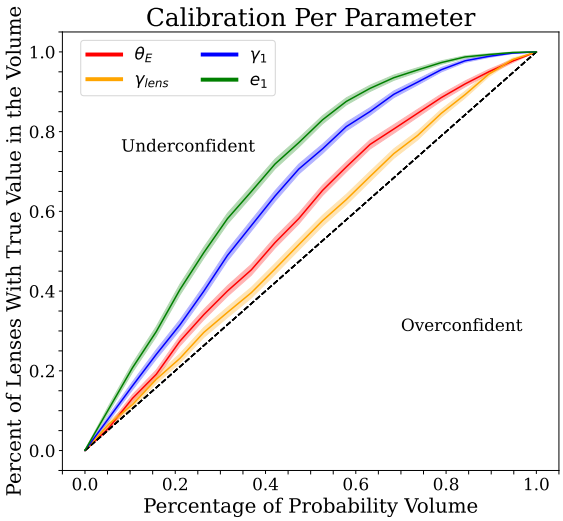

Figure 6. Calibration for 4 parameters in the fiducial case (Section 4.1.)

1. Mean error (ME): ME for a given parameter is the average difference between the mean of the posterior and the ground truth value for that parameter across all lenses. ME tracks bias by testing for systematically low or high predictions.

2. Median absolute error (MAE): MAE for a given parameter is the median absolute difference between the mean of the posterior and the ground truth value for that parameter across all lenses. MAE tracks accuracy as lower values indicate that predictions are clustered more tightly around the truth.

3. Precision: Precision of a given parameter is the median uncertainty or 1-σ posterior width across all lenses.

For parameters $\theta_{\rm E}$ and $\gamma_{\rm lens}$, we also present the relative precision (Precision/lens(%) = Median(σ/truth) $\times$ 100) and relative mean error percent (Mean error/lens(%) = Mean(error/truth) $\times$ 100) in the text to better compare across different experiments.

We are additionally interested in the uncertainty calibration of the posteriors. We test whether the predicted uncertainties of the approximate posteriors are consistent by verifying that x% of posterior volume contains the truth x% of the time. If less than x% of the truth is contained in x% of the posterior volume, this volume is too large and the predicted uncertainties are too big, the network is underconfident. Correspondingly, if more than x% of the truth is in x% of the volume, the predicted uncertainties are too small, the network is overconfident. This assesses calibration over the entire test set. This formalism was introduced in (Park et al. 2021).

We present recovery of $\theta_{\rm E}, \gamma_{\rm lens}, \gamma_1, {\rm e}_1$ in Figure 5. We present recovery of all model parameters in Table 3. First we present results when all light components are included in the image. On the individual mass models, we are able to recover key mass model parameter θ_E with 6.5% precision per lens, and a mean error of -1% per lens. This corresponds to a median precision of 0.05 and average error of -0.01 overall on the entire test set as shown in Table 3. We recover the density profile slope $\gamma_{\rm lens}$ with 8% precision per lens and 3% error per lens. This corresponds to 0.16 median precision and 0.05 average error on the whole test set as shown in Table 3. We also note that the network picks up on additional information not present in the training data: there is a strong correlation in the uncertainties measured for ellipticity and shear parameters. This could be because the network is picking up on the shear-ellipticity degeneracy (e.g Johnson et al. 2024; Fleury et al. 2021). We show calibration for individual model parameters in Figure 6. In the fiducial case, we see that the network calibration is under-confident overall which means that we overestimate the uncertainties. We note that, in the fiducial case, we predict 71 NPE posteriors that are wider than the training distribution. We discard these ill-constrained samples before inferring the population-level parameters.

Next, we apply a neural network trained on lens light subtracted images, to the lens light subtracted dataset (described in Section 2.4). We present the NPE posterior metrics in Table 3. We find that lens light subtraction has little impact on recovery of θ_E. We find that it worsens recovery of ellipticity parameters by a factor of 2, compared to the case where all light components are included. We hypothesize that this is because the network has been trained under mass-traces-light framework, seen in the gray scatter in Figure 11. Subtracting lens light removes this information. Finally, we see that with lens light subtraction, bias on γ_{lens} increases from the fiducial case by 40%. We discuss the reason for this in Section 5.2.

In addition to evaluating performance on individual lens models, we assess recovery of the population model. As described in Section 3.3, we run a hierarchical inference to infer the population distribution over lens parameters. This results in a posterior over the hyper-parameters, $p(\nu|d)$. Our inferred $p(\nu|d)$ for the fiducial test set is shown in Figure 7. We also compute the error between the inferred hyper-parameters and their ground truth values. These values are shown in Table 5. Overall, we find that the inferred population-level

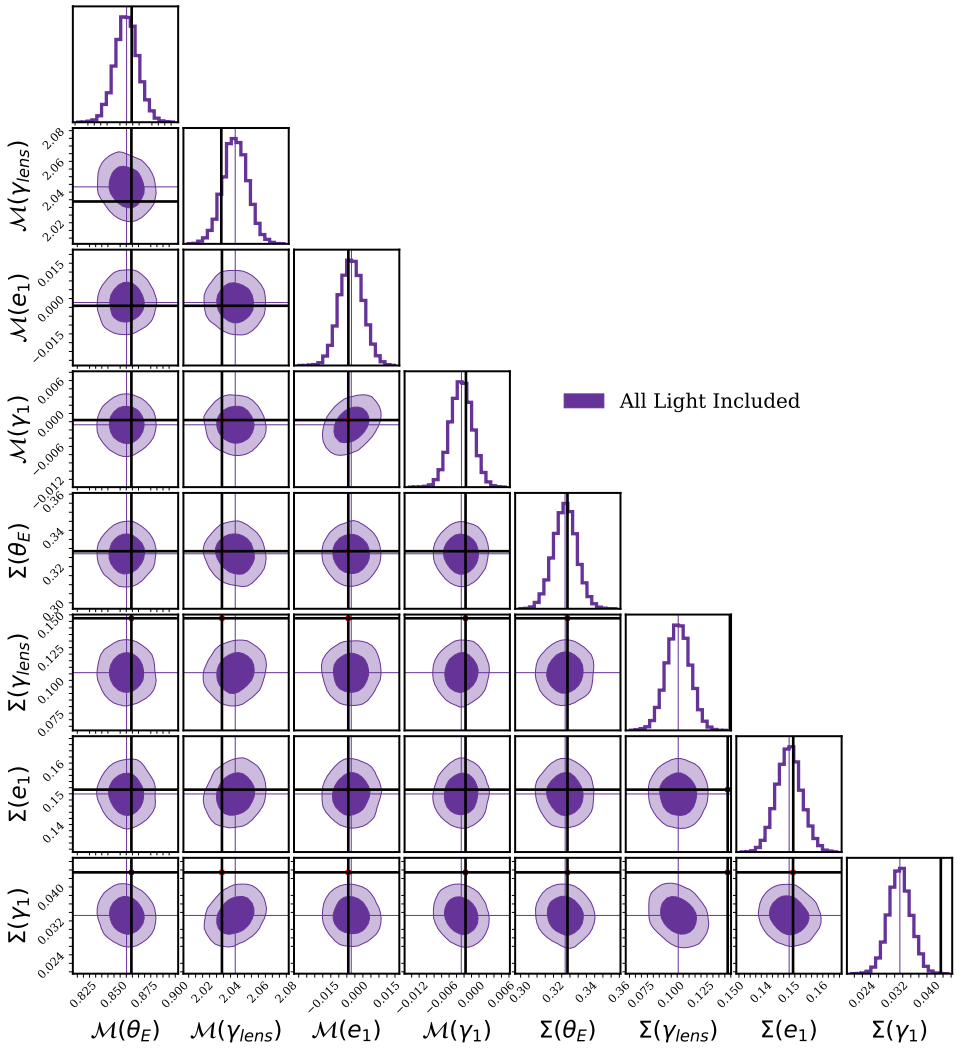

Figure 7. Population level model recovery for $\theta_E, \gamma_{\rm lens}, e_1$ and γ_1 using posteriors recovered from the fiducial test set. The black solid line is the ground truth of the test set, and the purple lines indicate the mean value of the converged chains. Note that we have removed 71 objects from our sample that had wider posteriors than the training distribution along one or more parameter axes as this breaks the HBI framework (refer Section 3.3). We show results for e_1 and γ_1 only, as e_2 and γ_2 reproduce the same result. We note that the population model learns a relationship between ellipticity and shear encoded into the test. The population mean of γ_{lens} is well constrained by the fiducial test set, and we go on to demonstrate how this can change with different model assumptions or data quality in Figure 9 and Figure 11.

model is within 1-σ of the truth for population mean parameter $\mathcal{M}(\theta_E), \mathcal{M}(\gamma_{\rm lens}), \mathcal{M}(e_{1\&2}), \mathcal{M}(\gamma_{1\&2})$ as shown in Figure 7. We outline the percentage bias in the recovery of population mean and uncertainty in Table 5. We are unable to learn intrinsic scatter in the sample of our lenses, quantified by Σ_ξ, specifically $\Sigma(\gamma_{\rm lens})$ and $\Sigma(\gamma_{1/2})$. Under-confident NPE posteriors (seen in Figure 6) yield a population model with low intrinsic scatter. We discuss this further in Section 5.6. We also tried modeling the posterior PDF for a given lens as diagonal multivariate Gaussian distribution. Given that many uncertainties are correlated, we find that this produces worse results (higher average error on parameters like $\gamma_{\rm lens}$) consistently than using a full covariance matrix, as expected.

4.2. *Bright AGN Host Galaxies*

We are interested in how the presence of a lensed arc impacts the NPE lens modeling performance, and subsequent population inference. To test this factor, we isolate the signal from the lensed arcs by removing both the lens light and the AGN light (more details about this preparation in Section 2.4). In some systems, the source galaxy is faint, and this leaves no signal above the noise floor. We only test on systems with a bright enough AGN host galaxy to produce arc light above the noise floor. We define signal as the fraction of pixels

above the noise level: $S = (N > \sigma)/N_{tot}$ where σ is the background noise level and N is the number of pixels, similar to the lensing information criterion introduced in Tan et al. (2024). Our selection criteria for bright hosts is images with $S > 0.5$, i.e more than half the pixels contain signal higher than noise – this results in 757 systems. We re-train the network to operate on arc-light only, and present results on the NPE posteriors in Table 4. While average error and MAE for θ_E and γ_{lens} remain similar to the fiducial test, we see that systems with bright host galaxies provides better precision in all light included, and lens light subtracted cases. When only the arcs remain in the image, we find that precision on θ_E and γ_{lens} is highest available out of experiments testing on LSST imaging, with the realistic train-test distribution shift. This shows that most of the information on these parameters is retained in the arcs and points to the need for future modeling techniques that isolate the arc before modeling.

We find that population recovery of $\mathcal{M}(\theta_E), \mathcal{M}(e_1)$ improve on the selection of bright host galaxies, as seen in Table 5.

4.3. *Image Deconvolution*

Here we explore how parameter recovery is affected by data quality, in particular whether the network performs better on deconvolved images. The deconvolved images are sampled on pixel grid twice as small as the original data, and have a known PSF corresponding to Gaussian with pixel full-width half maximum. We apply emulated deconvolution to train and test images under all three preparations detailed in Section 2.4. We show results on deconvolution test in Table 4. We find that this improves parameter recovery to 3% precision/lens on $\theta_{\rm E}$(compared to 6.5% precision reported on simulated LSST imaging) and 5% precision/lens on $\gamma_{\rm lens}$(compared to 8% on simulated LSST imaging). In the most optimal case, where we emulate deconvolution and isolate the arcs in the images via lens and AGN light subtraction, we obtain no bias and 4% precision/lens on γ_{lens}. Unlike θ_E, that sees marginal improvement in parameter recovery between LSST and emulated deconvolved imaging, performance on $\gamma_{\rm lens}$ improves substantially. We find that calibration for all parameters is still underconfident but improved from the fiducial case, showing that better resolution also improves calibration.

When only lens light is subtracted, we find that the average error on $\gamma_{\rm lens}$ is 0.05, which is still high. We discuss this in Section 5.1.

Regarding population parameter recovery, we find that the uncertainty on all inferred parameters, $\mathcal{M}(\xi)$ and $\Sigma(\xi)$, is lower. The emulated deconvolved data has better constraining power that simulated ground-based imaging. Despite lower uncertainties, we still see that bias on inferred $\mathcal{M}(\theta_E)$ is higher compared to the Fiducial case, as seen in Table 5. We hypothesize that this could be because with higher data resolution, the inference is more sensitive to the mis-specified functional form. We present a more in-depth discussion on this in Sections 5.4 and 5.6.

4.4. *Removing Mass-Light Correlations in the Training Data*

We hypothesize that one of the main factors affecting performance is the correlation of mass and light encoded into the training set. In our baseline set-up, the network is trained on lenses where ellipticity of the mass and light profiles are correlated (shown in Table 2). In this experiment, we regenerate training data without the ellipticity mass-light correlations, and re-train the NPE modeler. We show posteriors generated from applying this model to the fiducial dataset where mass and light ellipticity are correlated in Table 4. We see that bias on θ_E increases from 1% to 2.5%. As $\gamma_{\rm lens}$ is a difficult parameter to measure from images, we do not find any major impact on $\gamma_{\rm lens}$ from using a less informative training prior. However, we note that ellipticity bias and precision is impacted heavily, shown in Table 7. Ellipticity bias is doubled, and precision is halved. To the contrary, calibration on ellipticity parameters is improved, while other parameters have the same calibration as the fiducial case. We note that this degradation in NPE posterior performance is later folded into the hierarchical analysis of population-level parameters, and this affects $\mathcal{M}(\gamma_{\rm lens})$, $\Sigma(e_1)$ and $\Sigma(e_2)$ inference. We discuss the reason for this, and techniques to mitigate this effect in Section 5.2 and Section 5.6. Overall, we find that performance is better when the network is trained with mass-light correlations.

4.5. *Removing distribution shift between test and training data*

The network learns to produce lens model posteriors, given an image, based on what it sees from the training distribution. Any distribution shift between the test and training distribution can lead to biased inference (Filipp et al. 2025). If the test set lies in an undersampled part of parameter space, posteriors produced will be less constraining. Fiducial results are produced from a test set shifted from the training distribution across multiple dimensions (test distribution shown in green and train distribution shown in gray in Figure 10). To test if distribution shift is the driver of bias in our results, we apply the network on a test set that is sampled

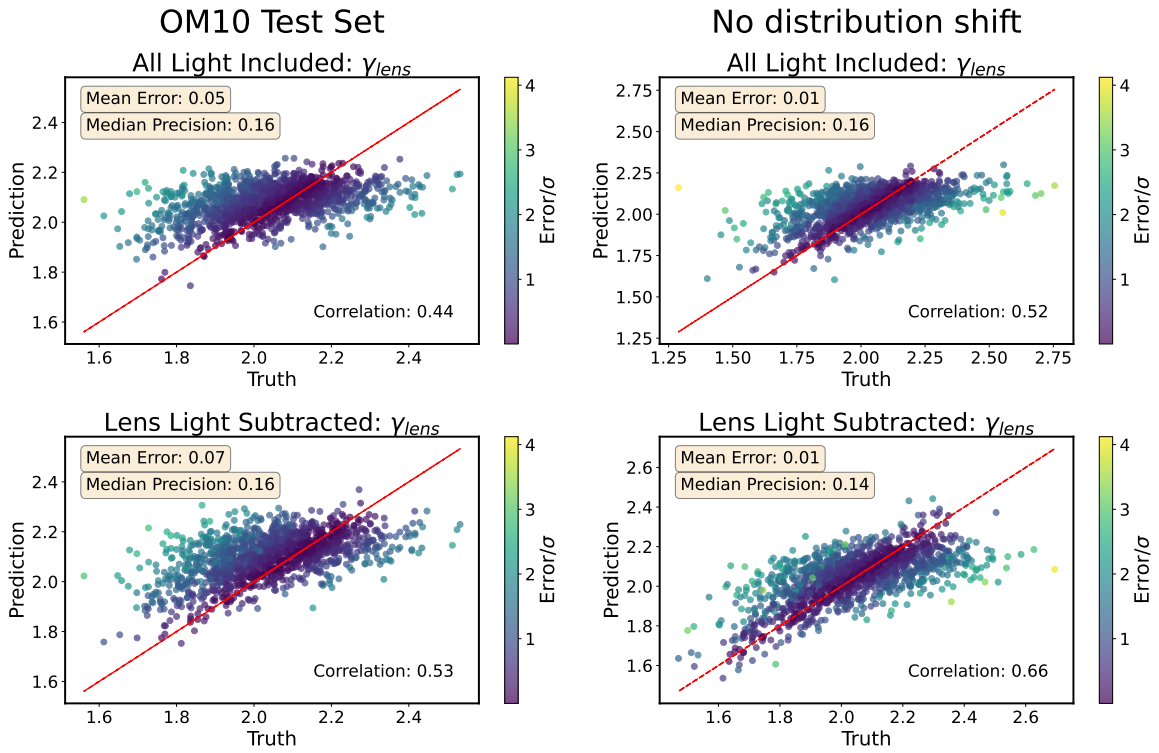

Figure 8. Top panels show recovery of $\gamma_{\rm lens}$ in presence and absence of distribution shift, when all light is included. We can see that the bias on $\gamma_{\rm lens}$ is significantly lower in the absence of distribution shift. The bottom left and right panels show the same when lens light is subtracted from the LSST quality images. We can see that in this case, not only is $\gamma_{\rm lens}$ inference less biased, it is also more accurate in the absence of distribution shift. Error is defined as predicted posterior mean - truth. The red line indictes perfect recovery.

from the training distribution. This is referred to as a 'held-out test set' going forward. Results are presented in Table 4. We show that there is still a residual bias, on parameters like θ_E and $\gamma_{\rm lens}$, and thus this is a function of misrepresented posterior PDF functional form or lack of expressivity of the network, and not just due to the distribution shift. We also find that the scatter of predictions around the true value is larger in the held-out test set. While the fiducial test has a complex selection function that makes all the lenses detectable and informative, the held-out test set lenses are completely randomly selected, and thus consist of lenses with lower lensing information. Further, the held-out test set contains some lenses that are in poorly sampled regions of the training data. This could add to the amplified MAE for all parameters seen in Table 4. Finally, we see that the calibration is perfect on all parameters in this experiment. This indicates that the under-confidence is largely a function of distribution shift, and it is insufficient to assess a network just using its calibration.

We do not propagate held-out test set posteriors into a hierarchical inference, since running the inference on a test set that matches the training set distribution would violate one of our key assumptions: that the test distribution is narrower than the training distribution (refer to Section 3.3).

5. DISCUSSION

After testing our automated modeling technique on a simulated LSST lensed AGN sample, we aim to understand our key performance drivers, and project how this technique will be adapted for future use.

5.1. *Impact of lens light subtraction on lens model recovery*

Lens light subtraction has been shown to benefit lens modeling generally (e.g Madireddy et al. 2022; Pearson et al. 2019; Pearson et al. 2021). In contrast, our fiducial experiment shows evidence lens light subtraction is not beneficial. As shown in Table 4, we see that subtracting lens light marginally improves the precision of θ_E and increases the average error of $\gamma_{\rm lens}$. One main difference between the analyses is our introduction of a distribution shift. Note that when we test without a distribution shift, we reproduce the result that lens light subtraction improves performance as seen in Figure 8. We hypothesize that, in the setting of distribution shift, lens light subtraction may no longer be helpful. Specific discussion of this effection on $\gamma_{\rm lens}$ is presented in

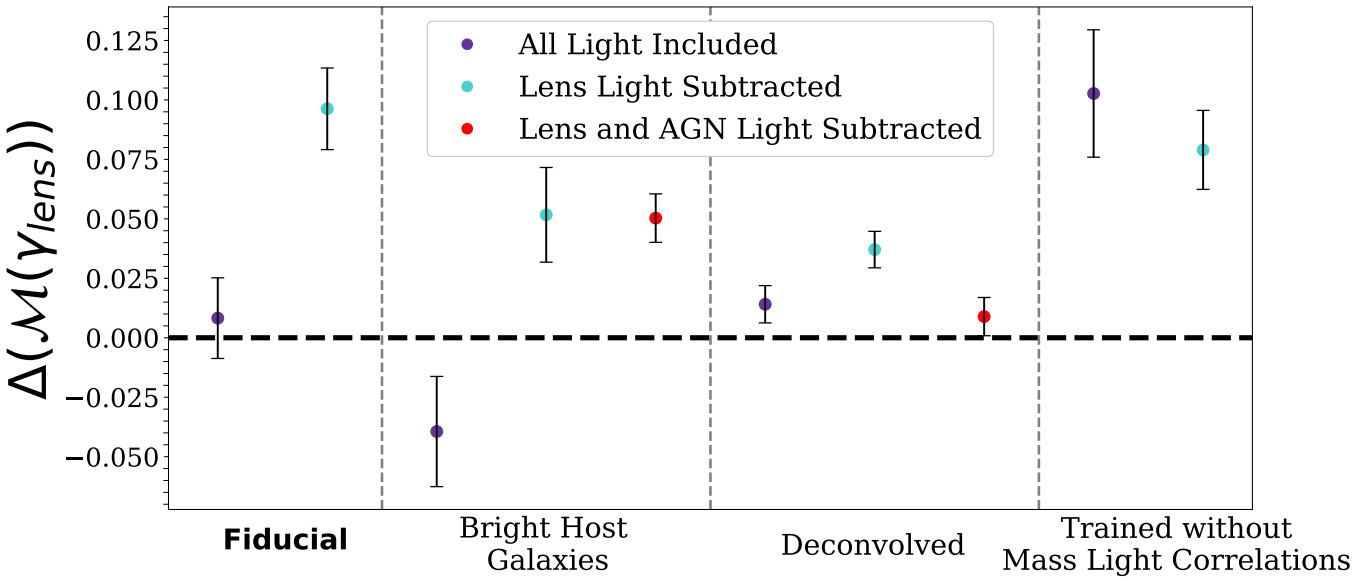

Figure 9. Recovery of $\mathcal{M}(\gamma_{\rm lens})$ under various experiments. Error bars indicate 2-σ uncertainty on the hierarchically inferred population mean. Note that the fiducial preparation captures the truth within one σ, and is lower than the average error on the individual lens mass model posteriors reported in Table 3.

Section 5.2. Also, note that the mass and light of the lens are correlated in the fiducial training. So, removing the lens light may remove a crucial piece of information about the mass profile. This loss of information may be exacerbated by the distributional shift.

5.2. $\gamma_{\rm lens}$ *Bias*

We find consistent positive bias on NPE-modelled $\gamma_{\rm lens}$, on all experiments as seen in Table 4. There is a lack of information on this parameter in imaging data (Birrer et al. 2020; Erickson et al. 2025), and it is primarily learned through the shape of the inferred mass distribution, and appearance of arcs in images. Existing degeneracies between these components of the lensing system and other parameters could introduce this bias.

This bias could be further exacerbated by the fact that we force our network to output a multivariate Gaussian PDF - a more malleable posterior functional form capable of capturing skewness and modality might help mitigate this. This is further reinforced by the fact that we sustain a mean error of 1%/lens on $\gamma_{\rm lens}$ even in the face of no distribution shift.

We demonstrate the bias on $\mathcal{M}(\gamma_{\rm lens})$ across all experiments in Figure 9, and in Table 5. We see that recovery of $\mathcal{M}(\gamma_{\rm lens})$ is also influenced by the inferred scatter on ellipticity $\Sigma(e_{1/2})$ and shear $\Sigma\gamma_{1/2}$. $\gamma_{\rm lens}$ is degenerate with the axis ratio and external shear parameters in a lensing system. The axis ratio q is the polar characterization of Gaussians e_1 and e_2, thus it is a Rayleigh distribution characterized by $\Sigma(e_{1/2})$. In the fiducial case, there is more information on these parameters, as we correlate lens mass and light ellipticity in the training, and this makes the prior more informative on these parameters. Even though we do not explicitly account for this in the conditional PDF, the fact that we learn ellipticity and shear more precisely is captured in the full covariance matrix of the individual mass model posteriors that we feed to the HBI framework. Finally, in Appendix A, we show the impact of only inferring $\mathcal{M}(\gamma_{\rm lens})$, versus all population-level parameters, showing that there is an improvement in results when we combine all the parameter posterior information.

When lens light is subtracted, and lensed arcs and point source are visible more clearly, we identify that the individual mass model posterior results consistently show higher positive bias on $\gamma_{\rm lens}$ - both in case of LSST and emulated deconvolved data. We hypothesize that this is because the network confuses light from a bright quasar for light from a large extended source being lensed and concentrated into a small region with high magnification by a deflector with a steep density profile (effect shown in Marshall et al. 2007).

In the fiducial case, there is a distribution shift between the size of quasar host galaxies in the training and test set – while the training set sees larger sources on average, most source galaxies are small in the test set. Thus, we hypothesize that the network believes that a source galaxy, on average must be bigger than the test set, and its compact appearance is due to the deflector density slope. To test this, we study the lens light subtracted results in the absence of train-test distribution shift across learned and latent parameters (Section 4.5).

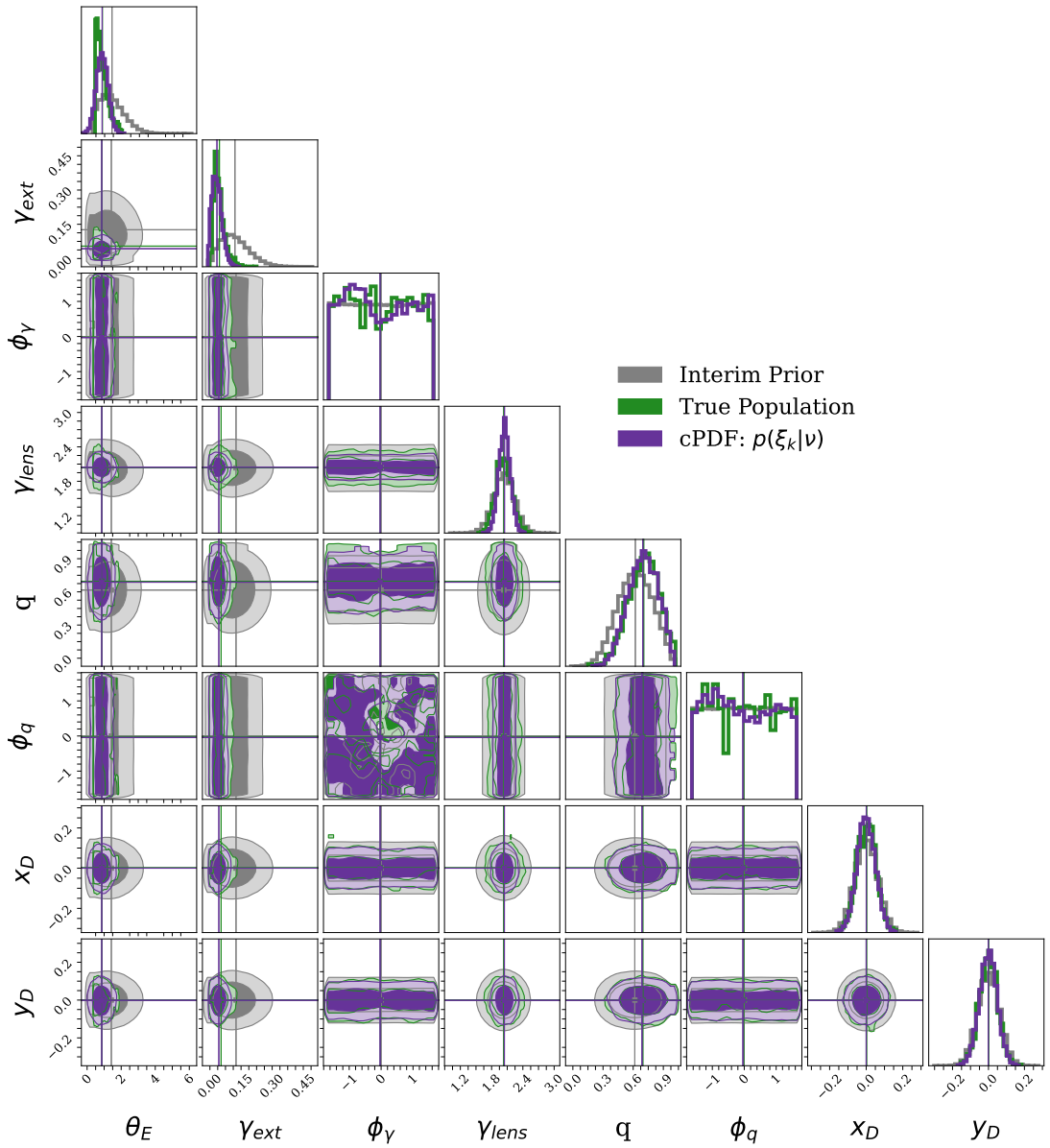

Figure 10. Here, we show the recovered conditional PDF $p(\xi_k|\nu)$ where we infer ν hierarchically. We start by showing the network 500k combinations of lensing parameters sampled from the interim prior (gray contours). We apply the network to our realistic constructed sample of OM10 lenses (green contours) realized as LSST images [fiducial], infer a multivariate Gaussian posterior PDF of the lensing parameters from each image. Then, we combine these posteriors in a hierarchical inference to learn the parameters characterizing the parent distribution they're sampled from (purple contours). We model the parent distribution as a diagonal Gaussian, and thus infer the M (mean) and Σ (scatter) at the population-level.

We show in Figure 11, that removing this distribution shift, in the case where lens light is removed, yields better accuracy on $\gamma_{\rm lens}$.

From Suyu et al. (2013); Treu & Marshall (2016), we know that mass distributions with steeper density profiles deflect light more and thus produce longer time delays, leading to measurements of shorter time-delay distances, and larger inferred values of H_0. Thus, our positive bias on the power-law density slope would translate directly to a positive bias on H_0.

5.3. *Realism of Training Data*

When we train on images with rough mass and light ellipticity correlation, and test on images that have the same relation, we see the best performance. Including loose correlations between mass and light in the training set should be done with caution. When we apply the network to test data that has any deviation from the correlation built into the training prior, we see that the calibration on ellipticity reduces. This affects the population-level parameter inference because we do not

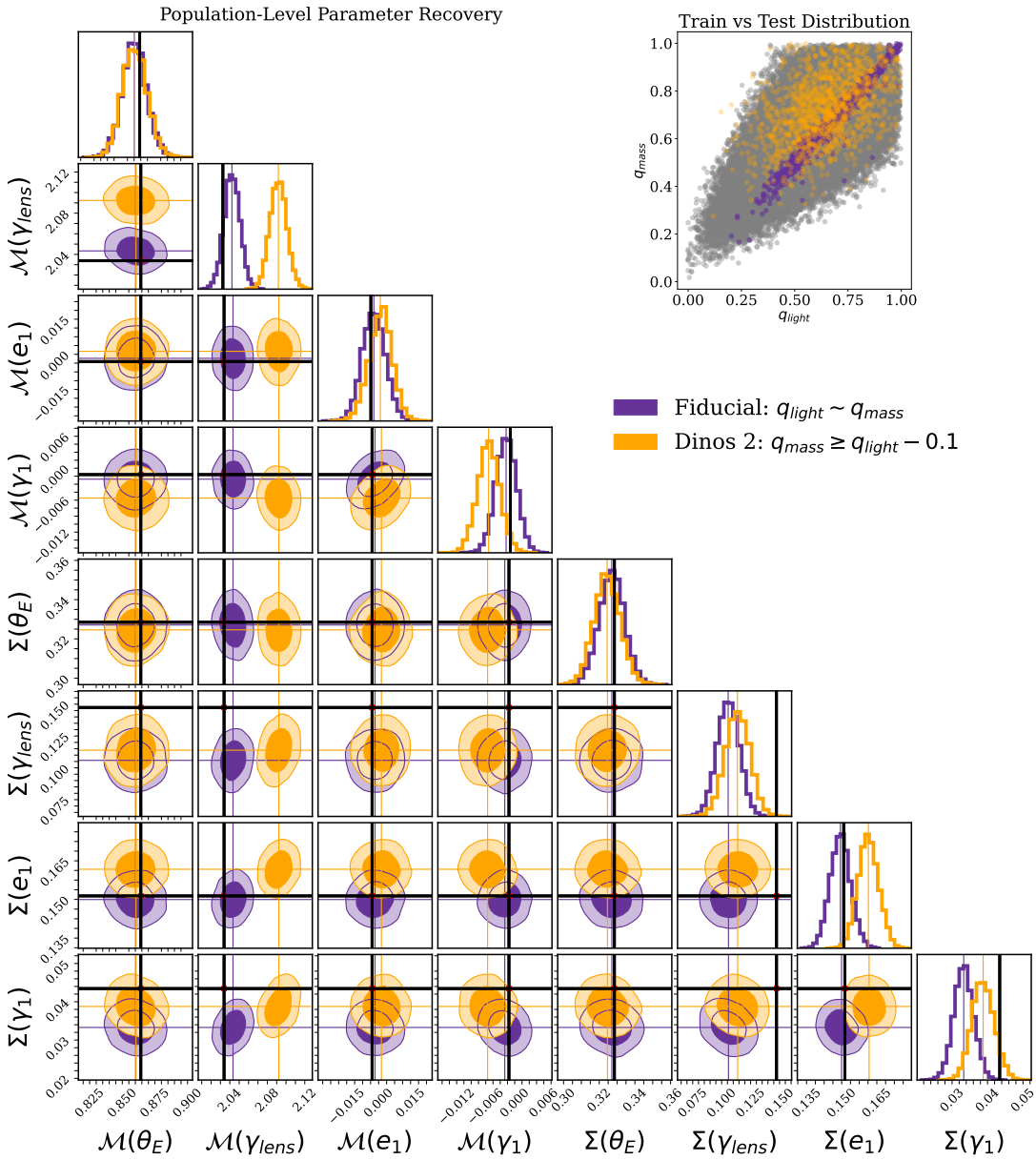

Figure 11. Recovery of population-level model parameters using 2 different test sets. In purple, we show our fiducial test set. In orange, we take the same test set and change the mass-light ellipticity relation to following $q_m \geq q_l - 0.1$ (taken from SLACS lenses (Bolton et al. 2008; Tan et al. 2024)). Inset panel: We show the two different mass-light ellipticity relations against the cloud of gray scatter points representing the mass-light ellipticity relationship in the training data. We demonstrate how sensitive the population-level parameter recovery is to such a shift, and suggest we need a more informative cPDF in order to work with such distribution shifts seen in real data.

also infer lens light parameters. In situations where we do infer the light as well, the inference pipeline will need to be provided with the information about the in-built correlation between mass and light in the modeling step.

We could also test this by including mass-light alignment taken from SLACs lenses where $\sigma_e = 0.12$ and $\sigma_\phi = 0.10°$ (Bolton et al. 2008) to the test set, and apply the fiducial network with loose mass-light correlation. This is adopted as prior in Schmidt et al. (2022) and Tan et al. (2024) who present the following relationship for the axis ratio of mass and light: $q_m \geq q_l - 0.1$, a situation that could be produced in reality by a massive elliptical galaxy residing in a more spherical dark matter halo. We construct a test-set with the exact same properties as the original catalog presented in 2, only changing the mass-light ellipticity relation to follow $q_m \geq q_l - 0.1$. We present the population-level parameter recovery in both test sets in Figure 11. Miscalibration of the network ap-

Table 4. Optimizing for data contained in images: Individual lens mass model posterior results

Experiment	Preparation	Metric	θ_E	$\gamma_{\rm lens}$
			(arcsec)	
Fiducial	ALL	ME	-0.01	0.05
		MAE	0.02	0.10
		Precision	0.05	0.16
	LLS	ME	-0.01	0.07
		MAE	0.02	0.10
		Precision	0.04	0.16
Experiments optimizing the data				
Bright Host Galaxies	ALL	ME	-0.00	0.04
		MAE	0.02	0.10
		Precision	0.04	0.16
	LLS	ME	-0.01	0.06
		MAE	0.02	0.09
		Precision	0.04	0.15
	LALS	ME	-0.01	0.06
		MAE	0.02	0.10
		Precision	0.03	0.14
Deconvolved	ALL	ME	0.01	0.03
		MAE	0.01	0.07
		Precision	0.02	0.12
	LLS	ME	-0.00	0.05
		MAE	0.01	0.05
		Precision	0.02	0.08
	LALS	ME	-0.01	0.02
		MAE	0.01	0.03
		Precision	0.01	0.05
Experiments modifying the modeling method				
No mass light correlations in training distribution	ALL	ME	-0.02	0.06
		MAE	0.03	0.10
		Precision	0.05	0.16
	LLS	ME	-0.01	0.06
		MAE	0.02	0.09
		Precision	0.04	0.16
No Distribution Shift	ALL	ME	0.01	0.01
		MAE	0.04	0.11
		Precision	0.06	0.16
	LLS	ME	-0.00	0.01
		MAE	0.03	0.09
		Precision	0.05	0.14

Note—Notes: Details about every experiment is available in Section 4. Details above every preparation is available in Section 2.4. ALL = All Light Included, LLS = Lens Light Subtracted, LALS = Lens and AGN Light Subtracted. Metrics are further defined in Section 4.

plied the test set with differing and shifted mass-light ellipticity relation results in biased inference of $\Sigma(e_{1/2})$, impacting inference of $\mathcal{M}(\gamma_{\rm lens})$.

5.4. *Lensing Information in LSST Quality Images*

While we find that precision on θ_E and $\gamma_{\rm lens}$ improves when we select for high SNR arc lenses as described in Section 4.2, it imposes a stronger selection function (narrowing our sample to 700 lenses) and distribution shift that we do not fully recover from. Currently, adopting lens and AGN light subtraction could be a good way to extract signal from these high-SNR arcs – while AGN light subtracted has not been attempted before, it could be possible with deconvolution.

From our emulated deconvolution experiment, we find that, on average, parameter recovery is better with higher resolution data. However, we become much more sensitive to functional form mis-specifications, and our population level constraints, while much tighter, are slightly biased high on $\mathcal{M}(\gamma_{\rm lens})$. Technology for deconvolving data at LSST scale is not yet available, and preparation of training data for this preparation is also challenging. Here, our results are on the most optimistic threshold - we train and test on emulated STARRED deconvolved data, and so there are no artifacts that confuse the network. If we actually apply STARRED to our simulated LSST images, we find that artifacts such as correlated noise can introduce a distribution shift in the simulation model and confuse the network.

In practice, we will not deconvolve the 5-year coadd image, but combine all the dithered exposure to reconstruct a higher resolution image of the lens, using STARRED. This would require generating individual exposures which is beyond the scope of this paper so we simply emulate the higher resolution images output by STARRED, which have a known Gaussian PSF with 2 pixel FWHM, at 0.1"/pixel (subsampled by a factor of 2 compared to the original data).

Currently, STARRED is optimized for light curve extraction of point sources. Extended sources are reconstructed on a higher resolution grid by combining dithered exposures and imposing a morphology prior through starlet regularization (Starck et al. 2015). Other methods, such as Rubin SharPy (Kaehler et al. in prep) use a Recurrent Inference Machine (RIM) to deconvolve large datasets of images with extended source at once. This could be a potential way to create a train and test dataset of deconvolved images.

5.5. *Timing*

A key benefit to machine learning lens modeling is its speed; here, we describe the timings of each piece of the

Table 5. Error in Population Model Parameter Recovery

Experiment	Preparation	$\mathcal{M}(\theta_E)$	$\Sigma(\theta_E)$	$\mathcal{M}(\gamma_{\rm lens})$	$\Sigma(\gamma_{\rm lens})$	$\mathcal{M}(e_{1\&2})$	$\Sigma(e_{1\&2})$	$\mathcal{M}(\gamma_{1\&2})$	$\Sigma(\gamma_{1\&2})$
Fiducial	ALL	**-0.004**	**-0.001**	0.01	-0.041	-0.001	-0.01	**0.001**	**-0.001**
	LLS	-0.01	-0.002	0.10	-0.004	0.01	0.001	0.02	0.03
Bright Host Galaxies	ALL	0.07	0.01	-0.05	-0.05	**0**	-0.01	-0.01	-0.01
	LLS	0.06	0.01	0.04	-0.03	0.01	**0**	0.01	0.01
	LALS	0.05	-0.3	0.04	-0.01	0.01	0.05	0.02	-0.11
Emulated Deconvolved	ALL	0.01	-0.27	0.02	**0.002**	-0.002	0.07	0.01	-0.11
	LLS	-0.002	-0.29	0.04	0.02	-0.001	0.09	-0.01	-0.103
	LALS	-0.01	-0.3	**0.009**	0.01	-0.002	0.1	-0.01	-0.1
W/o mass/light correlations	ALL	-0.03	-0.01	0.1	-0.15	0.01	0.01	0.02	0.03
	LLS	-0.01	-0.02	0.08	-0.02	0.01	0.002	0.01	0.02

NOTE—Table contains the difference between inferred population mean or scatter and fiducial test distribution mean and scatter.

lens modeling process. We generate 500 000 33x33 px size images for the training data which takes 3 hours on a CPU. We then generate 1344 images for each of OM10 lens systems for the test data, which takes 6 minutes on a CPU. Training for 200 epochs takes 1.5 hours on a GPU, and finally applying the trained network to the test set takes 5 minutes on a CPU. Overall, the modelling pipeline takes 12.5 seconds to model a single lens, for a sample of 1344 lenses. To compute the population model for each experiment, the computational cost of MCMC was 1 - 2 CPU hours. This primarily scaled with number of parameters inferred. We consistently measured 8 key parameters outlined in Section 3.

5.6. *Fiducial Test Set: Population Model Recovery*

There are two factors leading to the hypermodel recovery we see in Figure 7 and this is attributed to the lack of information provided in construction of the conditional PDF (cPDF, shown in Figure 10). Using an uninformative cPDF makes our results data-dominated, and we are information-limited in the data. The cPDF can be made more informative in two ways: (i) by building in the correlations between parameters that we expect to exist in real data, and using informative priors, and (ii) by allowing the cPDF to take on more flexible forms than a diagonal Gaussian that allow for correlations between parameters, and better characterize the distribution of parameters in the test set.

First, we use an uninformative uniform prior on the hypermodel parameters. The two parameters we are unable to learn in the hypermodel are $\Sigma(\gamma_{\rm lens})$ and $\Sigma(\gamma_{1/2})$. These are the two parameters that are hardest to measure with the data and most poorly constrained in the individual posterior level. We note that the recovery of external shear in individual mass model is still better than what has been recovered using galaxy-galaxy lenses in Schuldt et al. (2023a) (Figure 5). We hypothesize that this is due to the additional image position information in the images from the bright background point source. However, the network is underconfident about these parameters, and the scatter of prediction around the truth is high. Thus, in the hypermodel inference, where we provide information-limited posteriors, and no additional information on existing correlations between parameters or tighter constraints on parameters that are hard to measure, we find that all the scatter is attributed to individual lenses and the true scatter in the sample is underestimated. On parameters like $\mathcal{M}(\theta_E)$ and $\mathcal{M}(e_{1/2})$, there is more information in the images, and our individual posteriors are better constrained. Thus, we find that despite θ_E following a skewed distribution we are able to accurately retrieve the underlying distribution mean and scatter for these parameters.

In future work, we can obtain informative priors on $\theta_{\rm E}$ from large galaxy-galaxy lens samples consisting of lenses with similar Einstein radii (Sonnenfeld et al. 2023). Also, we expect to be able to better measure parameters like $\gamma_{\rm lens}$ using the Gold sample and can use that to make a informative prior. The Gold sample can be used in general to obtain correlations between all lens mass model parameters.

Second, our diagonal Gaussian cPDF does not represent the underlying test set as seen in Figure 10. For example, θ_E is sampled from a non-Gaussian distribution in the fiducial test set, and in several empirical datasets (Sonnenfeld et al. 2023). The lensing selection results in non-Gaussianity in lens mass model parameters, due

to observational effects (like projection effects that force more galaxies to look elliptical, and the seeing or PSF FWHM limiting the number of systems we can observe with $\theta_{\rm FWHM} > 2 \times \theta_{\rm E}$).

Learning alternate functional forms, like a log-normal distribution instead of a diagonal Gaussian could be a good first step in this case.

5.7. *Applications of Silver Sample Beyond Cosmology*

We do not find that there is any significant difference in accuracy of recovery of parameters across the range of redshifts that we probe. This means that modeling LSST imaging will provide powerful insights into the size, evolution, and mass of galaxies up to redshift 2.

5.8. *Future Work*

In this section, we highlight the main work needed in the future of SBI-based lens modeling informed by our results and discussion.

Training Data: There is limited information on certain parameters such as $\gamma_{\rm lens}$ available in ground-based images which makes the inference of these parameters prior-dominated. Although we can recover from ill-specified training priors on parameters like $\theta_{\rm E}$, we do not see the same behavior on $\gamma_{\rm lens}$. This means that we have to increase the realism of our training data in the following ways: include galaxy light-mass scaling relations, increase complexity of AGN host galaxy source structure to include features from disks, galaxy mergers etc. We also need to improve lens and environment model complexity and microlensing effects on AGN.

Normalizing flows: We note that there is a small persistent bias in parameter recovery on the held-out test set. Thus, even when the training data are fully representative of the test set, we see that the approximate inference does not sufficiently capture the true posterior PDF. We suggest that using more flexible functional forms than multivariate Gaussians and alternate methods to NPE such as normalizing flows to infer the individual mass model posteriors would be a useful next step, as demonstrated in Poh et al. (2025).

Multi-band imaging and modeling: Multi-band imaging and modeling is required on many fronts. We can learn the source structure better, we can model the deflector and subtract it out better and constrain the photometric redshifts of the systems better. From Holloway et al. 2025 (in prep), we can see that the presence of neighboring objects near the lens systems, such as nearby galaxies, degrade the precision of the predictions. This can also be mitigated by using multi-band imaging as shown in Pearson et al. (2019).

6. CONCLUSIONS

We apply Neural Posterior Estimation (NPE) on simulated LSST imaging data of lensed AGN to construct lens mass models. We go on further to use hierarchical Bayesian inference (HBI) to learn the population from which our lenses were sampled. This work represents the first application of the NPE–HBI method to a large ground-based dataset of lensed AGN systems drawn from a realistically expected parameter distribution. We have demonstrated both the strengths and limitations of the approach, with specific focus on extending this method to infering cosmology from LSST lensed AGN. We show that there is useful information in the LSST sample for cosmology and outline clear next steps for further improving individual parameter constraints and population-level parameter constraints.

Based on our investigations, we draw the following conclusions about the questions we have outlined in Section 1. All results quoted below are for the fiducial data (refer to Section 4.1). In the fiducial setup, we test on LSST lensed AGN constructed from the catalog detailed in Section 2. We apply a network trained with mass-light ellipticity correlations matching the correlations in the test set.

1. How accurately can we measure the key lens model parameters $\theta_{\rm E}$ (the Einstein Radius) and $\gamma_{\rm lens}$ (the density profile slope) from the LSST 5-year coadd imaging alone?

    - We can recover $\theta_{\rm E}$ with less than 1% bias per lens, 6.5% precision per lens and $\gamma_{\rm lens}$ with less than 3% bias per lens, 8% precision per lens.

2. To what extent does this performance depend on the subtraction of the lens galaxy light, and the AGN point image light? How does deconvolution applied to the data prior to modeling improve the accuracy of measured lens models?

    - We find that lens light subtraction yields similar performance on $\theta_{\rm E}$ and increases bias on $\gamma_{\rm lens}$. We hypothesize this drop in performance is due to the network losing information about lens mass from lack of lens light, making the network more susceptible to distribution shifts.
    - When subtracting both lens light and AGN light, leaving only the arc light from the AGN host galaxy, we see improved precision on $\theta_{\rm E}$ and $\gamma_{\rm lens}$. We hypothesize that most lensing information about $\theta_{\rm E}$ and $\gamma_{\rm lens}$ is inferred from the arcs.

- Emulated deconvolution improves the precision of inferred parameters by 30-60%. It also lowers the bias on on parameters like $\gamma_{\rm lens}$. Thus we see that deconvolution prior to modeling will be beneficial in obtaining lens models from LSST images.

3. How do distribution shifts in learned and latent parameters impact the inference of lens mass model parameters?

   - We find that removing distribution shift between the train and test data has a two-fold effect: (i) it reduces the bias on difficult to measure parameters like $\gamma_{\rm lens}$ (ii) the change in test distribution to a random sample of lenses, rather than the fiducial test set which is selected for detectability, reduces the average amount of information per lens, resulting in higher average uncertainty on parameters like $\theta_{\rm E}$. These effects demonstrate the importance of testing under a realistic distributional shift.

4. How does the presence of mass-light ellipticity correlations in the training data impact the accuracy of inferred key lens model parameters, specifically $\gamma_{\rm lens}$?

   - In the fiducial case, we include mass-light ellipticity correlations in the training data and on average, the same mass-light ellipticity correlations in the test data. We find that removing this correlation from the training data biases $\theta_{\rm E}$ and $\gamma_{\rm lens}$ recovery, demonstrating the importance of a realistic prior on mass-light ellipticity.

5. How accurately can we recover the population model for key lens mass model parameters?

   - In the fiducial case, we measure the population level parameters for $\mathcal{M}(\theta_E)$ and $\Sigma(\theta_E)$ without bias. We measure $\mathcal{M}(\gamma_{lens})$ with less than 1% positive bias and $\Sigma(\gamma_{lens})$ with 5% negative bias. We find that distributional shifts in the mass/light ellipticity relation significantly impact the recovery of $\mathcal{M}(\gamma_{lens})$.

We demonstrate the feasibility of modeling LSST-quality images of lensed AGN with automated techniques, incorporating a thorough investigation of data preparations and distributional shifts. This work builds a foundation for utilizing all LSST imaging of lensed AGN for a dark energy measurement from time-delay cosmography.

Software: lenstronomy, paltas, matplotlib, numpy, pandas, emcee

ACKNOWLEDGMENTS

This paper has undergone internal review in the LSST Dark Energy Science Collaboration. P.V would like to thank internal reviewers Sreevani Jaragula and Ayan Mitra for their feedback. We additionally thank Anowar Shajib, Risa Wechsler, Sebastian Wagner-Carena, Gautham Narayan, Decker French, Margaret Verrico and the members of the DESC Time Domain Working Group and LSST Strong Lensing Science Collaboration for useful discussions.

P.V developed datasets, built on existing tools `paltas` and `lens-npe` to scale for large datasets, ran all analysis, produced all figures and wrote the main body of the text. S.E provided invaluable feedback on all components of this work: dataset development, analysis and writing and developed tools used for HBI analysis, and post-processing of NPE analysis. P.M helped define the lensed AGN sample, design the various experiments, design statistical framework to analyze results and provided invaluable feedback on all aspects. M.M provided interpretation and context for results and provided feedback on all aspects of analysis. P.H helped benchmark results on this dataset against other datasets, advised on implementation of tools for large-scale modeling, provided feedback on all components of analysis and writeup. S.B provided feedback on analysis choices, presentation of results, framing of methods and writeup. X.H provided feedback on analysis and writeup. N.K provided feedback on lens sample used for testing, simulation of training and test images. G.M provided input on initial construction of catalog, provided feedback on broad goals and scientific context of the project. A.R provided broad feedback on all components of this work. R.K provided data products used for analysis. S.D provided context for calibration of network, useful edits to writeup. A.M and S.J were internal reviewers for this work and provided valuable feedback on all components of the writeup. K.R's work as a Rubin builder contributed to this work, especially through the realistic simulations we employ.

The work of PV was supported by KIPAC Post Baccalaureate Fellowship. SE acknowledges funding from NSF GRFP-2021313357 and the Stanford Data Science Scholars Program. MM acknowledges support by the SNSF (Swiss National Science Foundation) through mobility grant P500PT_203114 and return CH grant P5R5PT_225598. SB is supported by DoE Grant DE-

SC0026113. PH acknowledges funding from the Science and Technology Facilities Council, Grant Code ST/W507726/1. This work was performed in part under DOE Contract DE-AC02-76SF00515. 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 Discovery Alliance 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.

Code developed for this work is publicly available in the `lsst_lensed_agn_modeling` [5] repository. This research made use of `lenstronomy`, a multi-purpose gravitational lens modeling software package (Birrer et al. 2021; Birrer & Amara 2018). This work uses public software package `paltas`, a NPE tool for large-scale strong lens modeling (Wagner-Carena et al. 2023). We used hierarchical inference scripts available in `lens-npe` (Erickson 2024).

REFERENCES

Abdalla, E., Abellán, G. F., Aboubrahim, A., et al. 2022, Journal of High Energy Astrophysics, 34, 49, doi: https://doi.org/10.1016/j.jheap.2022.04.002

Abe, K. T., Oguri, M., Birrer, S., et al. 2025, The Open Journal of Astrophysics, 8, 8, doi: 10.33232/001c.128482

Auger, M. W., Treu, T., Bolton, A. S., et al. 2010, ApJ, 724, 511, doi: 10.1088/0004-637X/724/1/511

Barkana, R. 1998, The Astrophysical Journal, 502, 531–537, doi: 10.1086/305950

Birrer, S. 2021, The Astrophysical Journal, 919, 38, doi: 10.3847/1538-4357/ac1108

Birrer, S., & Amara, A. 2018, Physics of the Dark Universe, 22, 189–201, doi: 10.1016/j.dark.2018.11.002

Birrer, S., Shajib, A. J., Galan, A., et al. 2020, Astronomy & Astrophysics, 643, A165, doi: 10.1051/0004-6361/202038861

Birrer, S., Shajib, A. J., Gilman, D., et al. 2021, Journal of Open Source Software, 6, 3283, doi: 10.21105/joss.03283

Birrer, S., Millon, M., Sluse, D., et al. 2024, Space Science Reviews, 220, 48, doi: 10.1007/s11214-024-01079-w

Bolton, A. S., Treu, T., Koopmans, L. V. E., et al. 2008, ApJ, 684, 248, doi: 10.1086/589989

Caon, N., Capaccioli, M., & D'Onofrio, M. 1993, Mon. Not. Roy. Astron. Soc., 265, 1013, doi: 10.1093/mnras/265.4.1013

Cranmer, K., Brehmer, J., & Louppe, G. 2020, Proceedings of the National Academy of Sciences, 117, 30055–30062, doi: 10.1073/pnas.1912789117

Ding, X., Treu, T., Birrer, S., et al. 2021, Monthly Notices of the Royal Astronomical Society, 503, 1096

Erickson, S. 2024, Model Weights, Predictions, and Chains: Lens Modeling of STRIDES Strongly Lensed Quasars with NPE, Zenodo, doi: 10.5281/zenodo.13906030

Erickson, S., Wagner-Carena, S., Marshall, P., et al. 2025, AJ, 170, 44, doi: 10.3847/1538-3881/add99f

Ertl, S., Schuldt, S., Suyu, S. H., et al. 2023, A&A, 672, A2, doi: 10.1051/0004-6361/202244909

Euclid Collaboration, Busillo, V., Tortora, C., et al. 2025, arXiv e-prints, arXiv:2503.15329, doi: 10.48550/arXiv.2503.15329

Falco, E. E., Gorenstein, M. V., & Shapiro, I. I. 1985, ApJL, 289, L1, doi: 10.1086/184422

Filipp, A., Hezaveh, Y., & Perreault-Levasseur, L. 2025, ApJ, 989, 226, doi: 10.3847/1538-4357/adee20

Fleury, P., Larena, J., & Uzan, J.-P. 2021, JCAP, 2021, 024, doi: 10.1088/1475-7516/2021/08/024

Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. 2014, ApJ, 795, 64, doi: 10.1088/0004-637X/795/1/64

Freedman, W. L., Madore, B. F., Jang, I. S., et al. 2025, Status Report on the Chicago-Carnegie Hubble Program (CCHP): Measurement of the Hubble Constant Using the Hubble and James Webb Space Telescopes. https://arxiv.org/abs/2408.06153

Freedman, W. L., Madore, B. F., Hatt, D., et al. 2019, ApJ, 882, 34, doi: 10.3847/1538-4357/ab2f73

Galan, A., Vernardos, G., Peel, A., Courbin, F., & Starck, J. L. 2022, A&A, 668, A155, doi: 10.1051/0004-6361/202244464

[5] GitHub:https://github.com/padma18-vb/lsst_lensed_agn_modeling

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, MNRAS, 522, 5442, doi: 10.1093/mnras/stad1325

He, K., Zhang, X., Ren, S., & Sun, J. 2015, Deep Residual Learning for Image Recognition. https://arxiv.org/abs/1512.03385

Hezaveh, Y. D., Perreault Levasseur, L., & Marshall, P. J. 2017, Nature, 548, 555, doi: 10.1038/nature23463

Huang, X., Baltasar, S., Ratier-Werbin, N., et al. 2025, DESI Strong Lens Foundry I: HST Observations and Modeling with GIGA-Lens. https://arxiv.org/abs/2502.03455

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111, doi: 10.3847/1538-4357/ab042c

Johnson, D., Fleury, P., Larena, J., & Marchetti, L. 2024, Journal of Cosmology and Astroparticle Physics, 2024, 055, doi: 10.1088/1475-7516/2024/10/055

Kingma, D. P., & Ba, J. 2017, Adam: A Method for Stochastic Optimization. https://arxiv.org/abs/1412.6980

Korytov, D., Hearin, A., Kovacs, E., et al. 2019, The Astrophysical Journal Supplement Series, 245, 26, doi: 10.3847/1538-4365/ab510c

MacLeod, C. L., Ivezić, Z., Kochanek, C. S., et al. 2010, The Astrophysical Journal, 721, 1014–1033, doi: 10.1088/0004-637x/721/2/1014

Madireddy, S., Ramachandra, N., Li, N., et al. 2022, A Modular Deep Learning Pipeline for Galaxy-Scale Strong Gravitational Lens Detection and Modeling. https://arxiv.org/abs/1911.03867

Marshall, P. J., Treu, T., Melbourne, J., et al. 2007, ApJ, 671, 1196, doi: 10.1086/523091

Millon, M., Michalewicz, K., Dux, F., Courbin, F., & Marshall, P. J. 2024, AJ, 168, 55, doi: 10.3847/1538-3881/ad4da7

Millon, M., Galan, A., Courbin, F., et al. 2020, A&A, 639, A101, doi: 10.1051/0004-6361/201937351

Oguri, M., & Marshall, P. J. 2010, Monthly Notices of the Royal Astronomical Society, 405, 2579, doi: 10.1111/j.1365-2966.2010.16639.x

Papamakarios, G., & Murray, I. 2016, arXiv e-prints, arXiv:1605.06376, doi: 10.48550/arXiv.1605.06376

Park, J. W., Wagner-Carena, S., Birrer, S., et al. 2021, The Astrophysical Journal, 910, 39, doi: 10.3847/1538-4357/abdfc4

Pearson, J., Li, N., & Dye, S. 2019, MNRAS, 488, 991, doi: 10.1093/mnras/stz1750

Pearson, J., Maresca, J., Li, N., & Dye, S. 2021, Monthly Notices of the Royal Astronomical Society, 505, 4362–4382, doi: 10.1093/mnras/stab1547

Perreault Levasseur, L., Hezaveh, Y. D., & Wechsler, R. H. 2017, ApJL, 850, L7, doi: 10.3847/2041-8213/aa9704

Poh, J., Samudre, A., Ćiprijanović, A., et al. 2025, JCAP, 2025, 053, doi: 10.1088/1475-7516/2025/05/053

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, MNRAS, 128, 307, doi: 10.1093/mnras/128.4.307

Riess, A. G., Yuan, W., Macri, L. M., et al. 2022, The Astrophysical Journal Letters, 934, L7, doi: 10.3847/2041-8213/ac5c5b

Schmidt, T., Treu, T., Birrer, S., et al. 2022, Monthly Notices of the Royal Astronomical Society, 518, 1260–1300, doi: 10.1093/mnras/stac2235

Schuldt, S., Cañameras, R., Shu, Y., et al. 2023a, Astronomy & Astrophysics, 671, A147, doi: 10.1051/0004-6361/202244325

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

Shajib, A. J., Nihal, N. S., Tan, C. Y., et al. 2025, dolphin: A fully automated forward modeling pipeline powered by artificial intelligence for galaxy-scale strong lenses. https://arxiv.org/abs/2503.22657

Shajib, A. J., Birrer, S., Treu, T., et al. 2020, MNRAS, 494, 6072, doi: 10.1093/mnras/staa828

Sonnenfeld, A., Li, S.-S., Despali, G., et al. 2023, Astronomy & Astrophysics, 678, A4, doi: 10.1051/0004-6361/202346026

Starck, J.-L., Murtagh, F., & Bertero, M. 2015, Starlet Transform in Astronomical Data Processing, ed. O. Scherzer (New York, NY: Springer New York), 2053–2098, doi: 10.1007/978-1-4939-0790-8_34

Suyu, S. H., Auger, M. W., Hilbert, S., et al. 2013, ApJ, 766, 70, doi: 10.1088/0004-637X/766/2/70

Taak, Y. C., & Treu, T. 2023, Monthly Notices of the Royal Astronomical Society, 524, 5446–5453, doi: 10.1093/mnras/stad2201

Tan, C. Y., Shajib, A. J., Birrer, S., et al. 2024, Monthly Notices of the Royal Astronomical Society, 530, 1474–1505, doi: 10.1093/mnras/stae884

TDCOSMO Collaboration, Birrer, S., Buckley-Geer, E. J., et al. 2025, arXiv e-prints, arXiv:2506.03023, doi: 10.48550/arXiv.2506.03023

Treu, T., & Marshall, P. J. 2016, The Astronomy and Astrophysics Review, 24, doi: 10.1007/s00159-016-0096-8

Treu, T., Suyu, S. H., & Marshall, P. J. 2022, The Astronomy and Astrophysics Review, 30, doi: 10.1007/s00159-022-00145-y

Verde, L., Treu, T., & Riess, A. G. 2019, Nature Astronomy, 3, 891, doi: 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

Wagner-Carena, S., Park, J. W., Birrer, S., et al. 2021, The Astrophysical Journal, 909, 187, doi: 10.3847/1538-4357/abdf59

Wong, K. C., Suyu, S. H., Chen, G. C. F., et al. 2020, MNRAS, 498, 1420, doi: 10.1093/mnras/stz3094

Yue, M., Fan, X., Yang, J., & Wang, F. 2022, AJ, 163, 139, doi: 10.3847/1538-3881/ac4cb0

APPENDIX

A. RELATIONSHIP BETWEEN ACCURACY OF NPE POSTERIORS AND POPULATION MEAN

In this section we elaborate on the relationship between the individual lens parameter posterior PDFs obtained using NPE and the inferred population parameters. We combine posterior PDFs from each individual lens using HBI described in Section 3.3 to infer the population that the lenses are sampled from. In Figure 5, we show that the NPE model cannot predict $\gamma_{\rm lens}$ accurately from single lenses in the fiducial test set. The flat scatter of points around 2.03 of predicted values vs true values indicates that the network is just returning the mean of the $\gamma_{\rm lens}$. However, since we infer each lens posterior PDF with a full covariance matrix, there are higher order terms that contain useful information. In Figure 12, we show that constraining the population model using only $\gamma_{\rm lens}$ sustains the bias we see in the interim posteriors. The average error on $\gamma_{\rm lens}$ in interim posteriors is 0.045, corresponding to the 0.042 error on $\mathcal{M}(\gamma_{\rm lens})$. However, jointly inferring all lens model parameters uses the higher-order information and de-biases the constraint $\mathcal{M}(\gamma_{\rm lens})$, bringing down the error to 0.01. We note that including information from all parameter dimensions also causes the uncertainty on inferred hyperparaters to inflate as the sampler now needs to accommodate a solution that fits more dimensions.

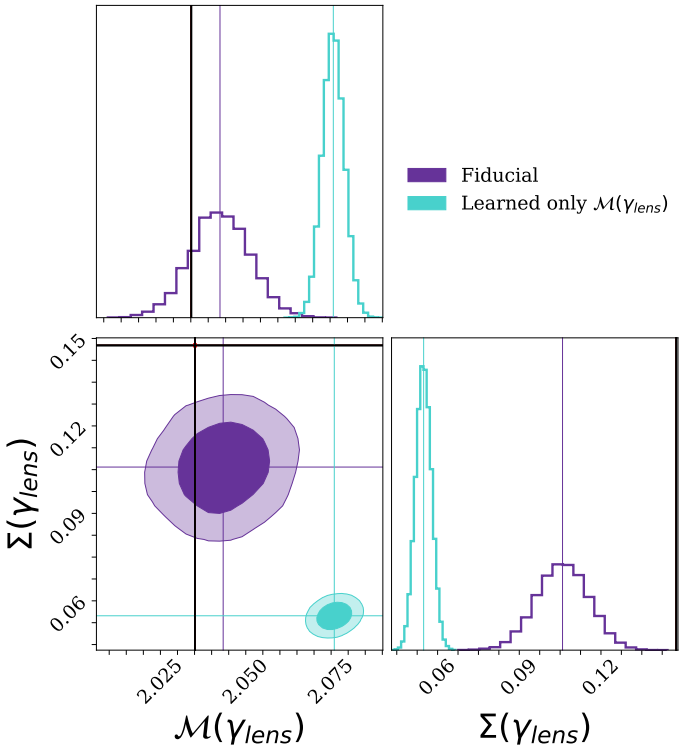

Figure 12. Black lines indicate the ground truth for the test set mean and scatter in $\mathcal{M}(\gamma_{\rm lens})$. Purple contour is reproduced from Figure 7, where we jointly learn all population model parameters for key 8 parameters. Turqoise contour is produced by only using $\gamma_{\rm lens}$ information from NPE posteriors to constrain population model for $\gamma_{\rm lens}$. The mean error on $\gamma_{\rm lens}$ NPE posteriors is 0.04, and this is perfectly reproduced in the inferred population mean.

B. FINAL POSTERIORS: RE-WEIGHTING OF NPE POSTERIORS

The lens parameter posterior PDFs we infer from images $\rm d_k$, are conditioned on the prior that we implicitly introduce by training the NPE modeler on $\xi_{\rm k}; \rm d_k$ pairs, where ξ_k is drawn from a broad distribution $\rm p(\xi|\nu_{int})$ (detailed in Section 3.2.1). We combine information gained from the full image dataset, $\{d\}$, to infer the conditional PDF $p(\xi_k|\nu)$ (detailed in Section 3.3). We can now update the lens model inferred from images, $\rm p(\xi_k|d_k, \nu_{int})$, with information from all the

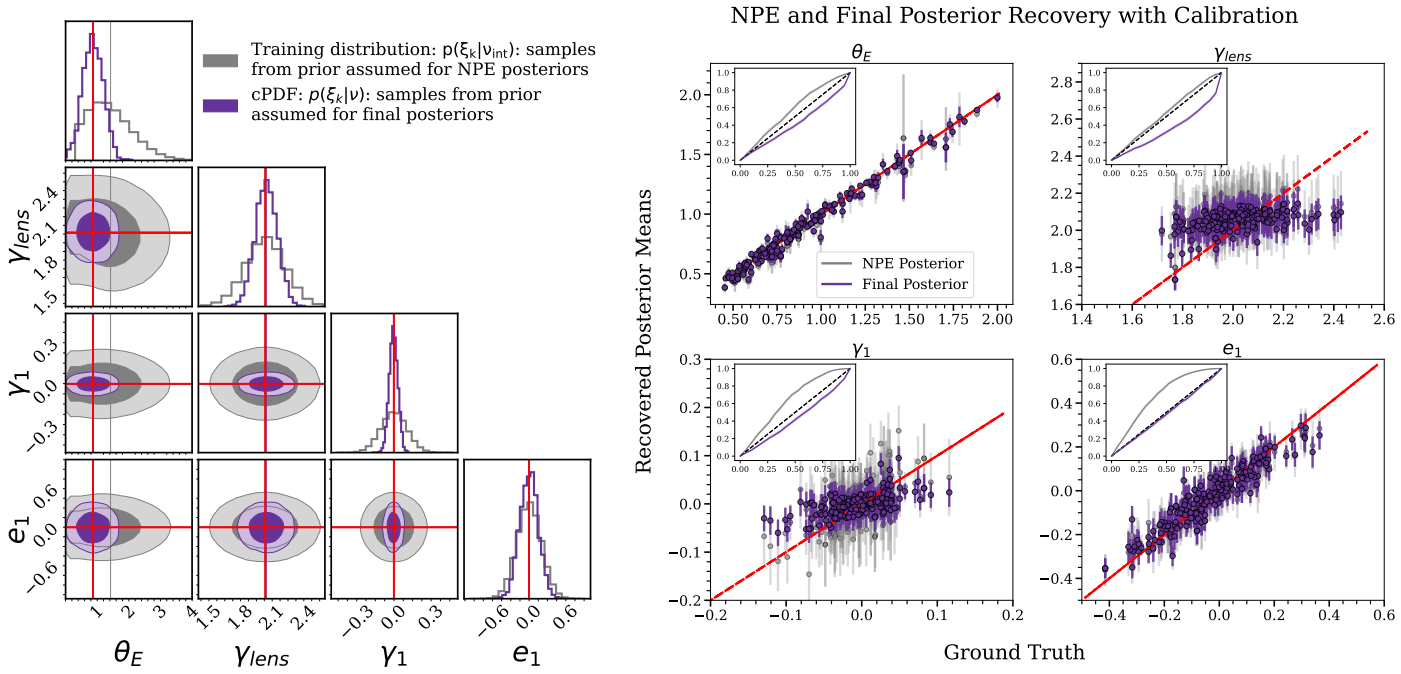

Figure 13. Here we show the impact of reweighting the NPE posteriors (Section 3.1) using the inferred cPDF (Section 3.3). Left: In gray, we show samples from the prior that the NPE posteriors are conditioned on, and in purple, we show samples from the prior that we learn and use for the final posteriors. Right: We show both NPE posteriors and final posteriors for a random sample of 200 objects. Final posteriors have smaller error-bars than the NPE posteriors, and their means are shifted towards the cPDF mean.

images $\{d\}$. Starting once again with Bayes' theorem,

$$\mathrm{p}(\xi_\mathrm{k}|\{\mathrm{d}\}) = \frac{\mathrm{p}(\{\mathrm{d}\}|\xi_\mathrm{k})}{\mathrm{p}(\{\mathrm{d}\})}. \tag{B1}$$

We can expand Equation B1, by including the dependency on ν, and integrating over choice of ν:

$$\mathrm{p}(\xi_\mathrm{k}|\{\mathrm{d}\}) = \int \mathrm{d}\nu \frac{\mathrm{p}(\{\mathrm{d}\}|\xi_\mathrm{k},\nu)\mathrm{p}(\xi_\mathrm{k}|\nu)\mathrm{p}(\nu)}{\mathrm{p}(\{\mathrm{d}\})}. \tag{B2}$$

Once again, since we have access to $\mathrm{p}(\xi_\mathrm{k}|\mathrm{d}_\mathrm{k},\nu)$ from the network, we can use the same expansion as Equation 9 in Section 3.3, to obtain:

$$\begin{aligned}\mathrm{p}(\xi_\mathrm{k}|\{\mathrm{d}\}) &= \int \mathrm{d}\nu\ \mathrm{p}(\xi_\mathrm{k}|\mathrm{d}_\mathrm{k},\nu_\mathrm{int})\mathrm{p}(\mathrm{d}_\mathrm{k}|\nu_\mathrm{int})\frac{\mathrm{p}(\xi_\mathrm{k}|\nu)}{\mathrm{p}(\xi_\mathrm{k}|\nu_\mathrm{int})}\mathrm{p}(\nu|\{\mathrm{d}\}) \\ &\propto \mathrm{p}(\xi_\mathrm{k}|\mathrm{d}_\mathrm{k},\nu_\mathrm{int})\frac{1}{\mathrm{N}}\sum_{\nu\sim\mathrm{p}(\nu|\{\mathrm{d}\})}\frac{\mathrm{p}(\xi_\mathrm{k}|\nu)}{\mathrm{p}(\xi_\mathrm{k}|\nu_\mathrm{int})}\end{aligned} \tag{B3}$$

In Equation B3, we obtain $\mathrm{p}(\nu|\{\mathrm{d}\})$ from HBI. We can drop $\mathrm{p}(\mathrm{d}_\mathrm{k}|\nu_\mathrm{int})$ which serves as a normalization constant, and N refers to the number of samples that we pull to evaluate $\frac{\mathrm{p}(\xi_\mathrm{k}|\nu)}{\mathrm{p}(\xi_\mathrm{k}|\nu_\mathrm{int})}$. An in-depth derivation of this reweighting scheme is presented in Wagner-Carena et al. (2021), Appendix C.

Our NPE posteriors are under-confident - we attribute this to distribution shift in a parameter between the test and train data, amount of information on that parameter in a ground based image, amount of information built into the training prior on a particular parameter. Overall, in cases where there is not much information in the image, network is not able to meaningfully hone in on the parameter value and the posterior is wide. This is further exacerbated by the fact that our multivariate gaussian function form is not malleable enough to capture skew/modality in the posterior PDF - so there will be instances where some posteriors are wider than they need to be. Under-confident NPE posteriors

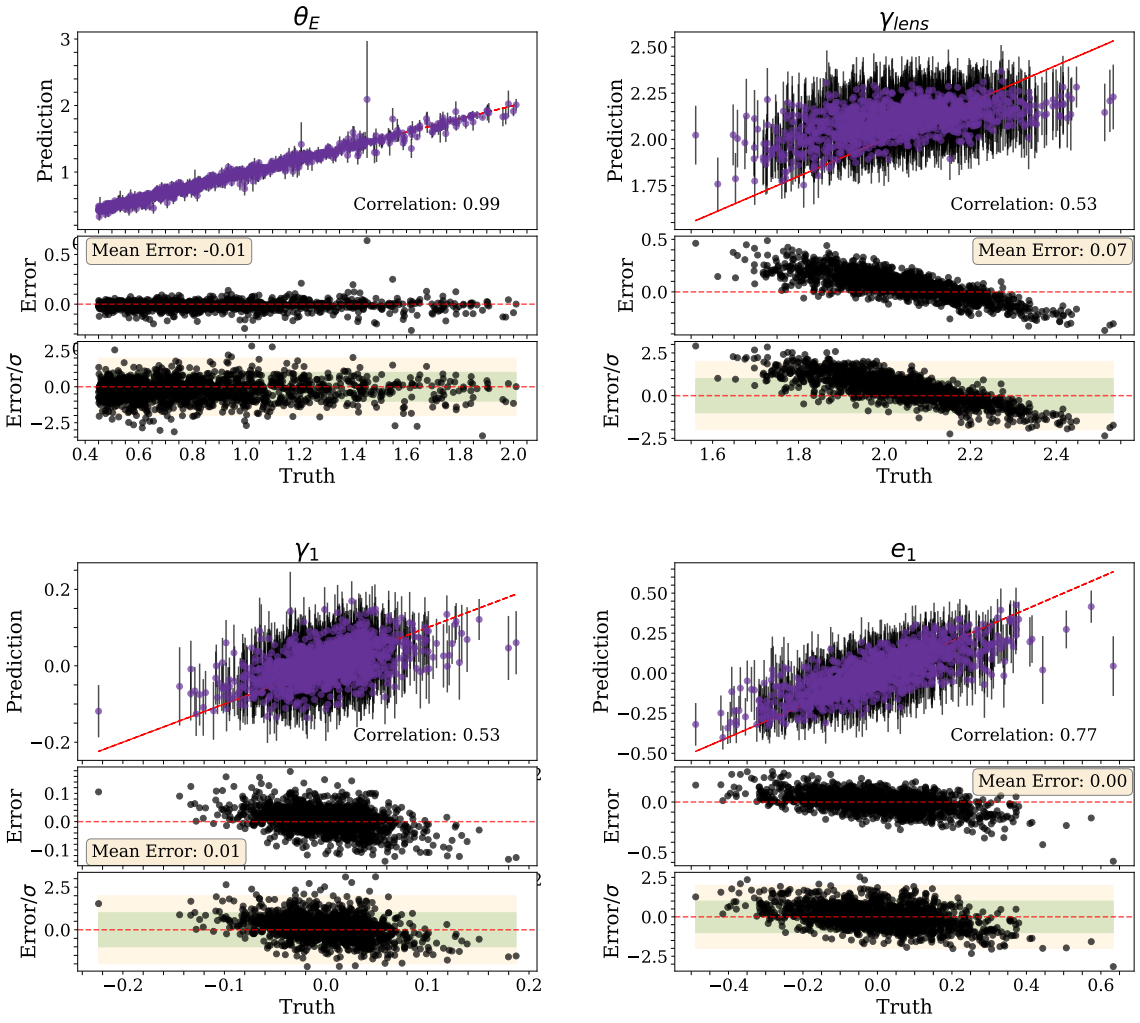

Figure 14. Here we reproduce Figure 5, but for the lens light subtracted preparation of data. Comparison of 1344 predicted Gaussian posteriors (obtained as described in Section 3.1) for lens light subtracted data (refer: Section 2.4) against ground truth. Purple points correspond to the mean of the posteriors, and black lines correspond to the posterior $1\,\sigma$ width or standard deviation. In the bottom most panel of each parameter subplot, the green shaded region represents the $1\,\sigma$ area and and yellow shaded region represents the $2\,\sigma$ region.

lead to under-estimated intrinsic scatter hyperparameters and hence an overly-narrow conditional PDF, which then over-corrects the individual posteriors such that the final posteriors are over-confident. We see this in the calibration curves in Figure 13 – the gray curves indicating underconfidence, become overconfident post re-weighting.

C. LENS LIGHT SUBTRACTED RESULTS

We present NPE posteriors for the lens light subtracted case. Further discussion on results are presented in Sections 4.1, 5.1.

D. NPE POSTERIORS FOR ALL PARAMETERS

We attach Table 7 containing NPE posterior metrics for all parameters across all preparation. Additionally, we attach a table containing the population model inference for all parameters across all experiments in Table 6.

Table 6. Population-Level Model Parameter Recovery

Experiment	Preparation	Metric	$\mathcal{M}(\theta_E)$	$\Sigma(\theta_E)$	$\mathcal{M}(\gamma_{\rm lens})$	$\Sigma(\gamma_{\rm lens})$	$\mathcal{M}(e_{1\&2})$	$\Sigma(e_{1\&2})$	$\mathcal{M}(\gamma_{1\&2})$	$\Sigma(\gamma_{1\&2})$
1		Ground Truth	0.87	0.34	2.04	0.15	-0.00	0.04	-0.00	0.15
	ALL	Inferred	$0.86^{+0.01}_{-0.01}$	$0.33^{+0.01}_{-0.01}$	$2.04^{+0.01}_{-0.01}$	$0.11^{+0.01}_{-0.01}$	$-0.00^{+0.002}_{-0.002}$	$0.03^{+0.003}_{-0.003}$	$-0.00^{+0.01}_{-0.01}$	$0.15^{+0.004}_{-0.004}$
		Error	**-0.004**	**-0.001**	0.01	-0.041	-0.001	-0.01	**0.001**	**-0.001**
	LLS	Inferred	$0.85^{+0.01}_{-0.01}$	$0.33^{+0.01}_{-0.01}$	$2.13^{+0.01}_{-0.01}$	$0.14^{+0.01}_{-0.01}$	$0.01^{+0.003}_{-0.003}$	$0.04^{+0.003}_{-0.003}$	$0.01^{+0.01}_{-0.01}$	$0.18^{+0.01}_{-0.01}$
		Error	-0.01	-0.002	0.10	-0.004	0.01	0.001	0.02	0.03
2		Ground Truth	0.87	0.34	2.04	0.16	-0.00	0.04	0.00	0.16
	ALL	Inferred	$0.94^{+0.01}_{-0.01}$	$0.34^{+0.01}_{-0.01}$	$1.99^{+0.01}_{-0.01}$	$0.11^{+0.01}_{-0.01}$	$-0.00^{+0.003}_{-0.003}$	$0.03^{+0.003}_{-0.003}$	$-0.01^{+0.01}_{-0.01}$	$0.14^{+0.01}_{-0.01}$
		Error	0.07	0.01	-0.05	-0.05	**0**	-0.01	-0.01	-0.01
	LLS	Inferred	$0.93^{+0.01}_{-0.02}$	$0.35^{+0.01}_{-0.01}$	$2.08^{+0.01}_{-0.01}$	$0.13^{+0.01}_{-0.01}$	$0.01^{+0.003}_{-0.003}$	$0.04^{+0.003}_{-0.003}$	$0.01^{+0.01}_{-0.01}$	$0.17^{+0.01}_{-0.01}$
		Error	0.06	0.01	0.04	-0.03	0.01	**0**	0.01	0.01
	LALS	Inferred	$0.93^{+0.01}_{-0.01}$	$0.04^{+0.003}_{-0.003}$	$2.08^{+0.01}_{-0.01}$	$0.15^{+0.01}_{-0.01}$	$0.01^{+0.003}_{-0.003}$	$0.09^{+0.004}_{-0.004}$	$0.02^{+0.01}_{-0.01}$	$0.04^{+0.002}_{-0.002}$
		Error	0.05	-0.3	0.04	-0.01	0.01	0.05	0.02	-0.11
3		Ground Truth	0.87	0.34	2.04	0.15	-0.00	0.04	-0.00	0.15
	ALL	Inferred	$0.87^{+0.01}_{-0.01}$	$0.04^{+0.002}_{-0.001}$	$2.05^{+0.004}_{-0.004}$	$0.15^{+0.004}_{-0.003}$	$-0.00^{+0.002}_{-0.002}$	$0.11^{+0.004}_{-0.003}$	$0.00^{+0.01}_{-0.004}$	$0.05^{+0.001}_{-0.001}$
		Error	0.01	-0.27	0.02	**0.002**	-0.002	0.07	0.01	-0.11
	LLS	Inferred	$0.87^{+0.01}_{-0.01}$	$0.04^{+0.001}_{-0.001}$	$2.07^{+0.004}_{-0.004}$	$0.16^{+0.004}_{-0.004}$	$-0.00^{+0.002}_{-0.002}$	$0.13^{+0.003}_{-0.003}$	$-0.01^{+0.01}_{-0.01}$	$0.05^{+0.001}_{-0.001}$
		Error	-0.002	-0.29	0.04	0.02	-0.001	0.09	-0.01	-0.103
	LALS	Inferred	$0.86^{+0.01}_{-0.01}$	$0.04^{+0.001}_{-0.001}$	$2.04^{+0.004}_{-0.004}$	$0.15^{+0.003}_{-0.003}$	$-0.00^{+0.001}_{-0.001}$	$0.14^{+0.003}_{-0.003}$	$-0.01^{+0.004}_{-0.004}$	$0.05^{+0.001}_{-0.001}$
		Error	-0.01	-0.3	**0.01**	0.01	-0.002	0.1	-0.01	-0.1
4		Ground Truth	0.86	0.33	2.03	0.15	-0.00	0.04	-0.00	0.15
	ALL	Inferred	$0.84^{+0.01}_{-0.01}$	$0.32^{+0.02}_{-0.009}$	$2.14^{+0.01}_{-0.01}$	$0.00^{+0.08}_{-0.0}$	$0.01^{+0.01}_{-0.01}$	$0.06^{+0.01}_{-0.01}$	$0.02^{+0.01}_{-0.014}$	$0.18^{+0.01}_{-0.01}$
		Error	-0.03	-0.01	0.1	-0.15	0.01	0.01	0.02	0.03
	LLS	Inferred	$0.85^{+0.01}_{-0.01}$	$0.32^{+0.01}_{-0.01}$	$2.11^{+0.01}_{-0.01}$	$0.13^{+0.01}_{-0.01}$	$0.00^{+0.003}_{-0.003}$	$0.05^{+0.003}_{-0.003}$	$0.00^{+0.01}_{-0.01}$	$0.17^{+0.01}_{-0.01}$
		Error	-0.01	-0.02	0.08	-0.02	0.01	0.002	0.01	0.02

Note—1 - Fiducial (4.1); 2 - Bright Host Galaxies (4.2); 3 - Emulated Deconvolved (4.3); 4 - Removing mass/light correlations in training prior (4.4). Note that we removed posteriors that were unconstrained (i.e. width of the posteriors was wider than the training prior) from the sample before computing the population parameters.

Note— ALL = All Light Included, LLS = Lens Light Subtracted, LALS = Lens and AGN Light Subtracted.

Table 7. Summary of metrics for all experiment results detailed in Section 4.

Experiment	Preparation	Metric	θ_E	$\gamma_{\rm lens}$	e_1	e_2	γ_1	γ_2
			(arcsec)					
Fiducial	All Light Included	ME	-0.01	0.05	0.	-0.	-0.	-0.
		MAE	0.02	0.1	0.03	0.03	0.02	0.02
		Precision	0.05	0.16	0.09	0.09	0.06	0.06
	Lens Light Subtracted	ME	-0.01	0.07	0.	-0.	0.01	-0.
		MAE	0.02	0.1	0.06	0.06	0.03	0.03
		Precision	0.04	0.16	0.13	0.13	0.07	0.06
Experiments optimizing the data								
Bright Host Galaxies	All Light Included	ME	-0.	0.04	0.	-0.	-0.	0.
		MAE	0.02	0.1	0.03	0.03	0.02	0.02
		Precision	0.04	0.16	0.09	0.09	0.06	0.06
	Lens Light Subtracted	ME	-0.01	0.06	0.01	-0.	0.01	-0.
		MAE	0.02	0.09	0.05	0.05	0.03	0.03
		Precision	0.04	0.15	0.11	0.11	0.06	0.06
	Lens and AGN Light Subtracted	ME	-0.01	0.06	0.02	0.03	0.01	0.01
		MAE	0.02	0.1	0.05	0.05	0.03	0.02
		Precision	0.03	0.14	0.1	0.1	0.05	0.05
Deconvolved	All Light Included	ME	0.01	0.03	0.	0.	-0.	0.
		MAE	0.01	0.07	0.03	0.03	0.02	0.02
		Precision	0.02	0.12	0.06	0.06	0.04	0.04
	Lens Light Subtracted	ME	-0.	0.05	-0.	0.01	-0.	0.
		MAE	0.01	0.05	0.03	0.03	0.02	0.02
		Precision	0.02	0.08	0.07	0.07	0.04	0.04
	Lens and AGN Light Subtracted	ME	-0.01	0.02	-0.01	-0.	-0.	-0.
		MAE	0.01	0.03	0.02	0.02	0.01	0.01
		Precision	0.01	0.05	0.04	0.04	0.02	0.02
Experiments modifying the modeling method								
W/o mass/light correlations	All Light Included	ME	-0.02	0.06	0.01	-0.	-0.	0.
		MAE	0.03	0.1	0.06	0.06	0.03	0.03
		Precision	0.05	0.16	0.14	0.14	0.07	0.07
	Lens Light Subtracted	ME	-0.01	0.06	0.	0.01	0.	0.01
		MAE	0.02	0.09	0.06	0.06	0.03	0.03
		Precision	0.04	0.16	0.13	0.12	0.06	0.06
No Distribution Shift	All Light Included	ME	0.01	0.01	0.	-0.01	-0.	-0.
		MAE	0.04	0.11	0.05	0.05	0.03	0.03
		Precision	0.06	0.16	0.09	0.09	0.06	0.06
	Lens Light Subtracted	ME	-0.	0.01	0.	-0.	0.	-0.
		MAE	0.03	0.09	0.06	0.06	0.03	0.03
		Precision	0.05	0.14	0.11	0.1	0.06	0.06

NOTE—We present tabulated results for all parameters other than lens position. We find that the network is almost always able to infer the lens position, and metrics do not provide useful insights on those parameters. Across all preparations, the bias on lens x and y position are less that 0.001", and the median precision is less than 0.01", thus, it errors and uncertainties are on a sub-pixel level for LSST. Details about every experiment is available in Section 4. Details above every preparation is available in Section 2.4. Metrics: ME = Mean error; MAE = Median absolute error; Precision - Median standard deviation of each posterior. Metrics are further defined in Section 4.