Neural Ratio Estimators Meet Distributional Shift and Mode Misspecification: A Cautionary Tale from Strong Gravitational Lensing

Andreas Filipp [1 2 3] **Yashar Hezaveh** [1 2 3 4 5 6] **Laurence Perreault-Levasseur** [1 2 3 4 5 6]

Abstract

In recent years, there has been increasing interest in the field of astrophysics in applying Neural Ratio Estimators (NREs) to large-scale inference problems where both amortization and marginalization over a large number of nuisance parameters are needed. Here, in order to assess the true potential of this method to produce unbiased inference on real data, we investigate the robustness of NREs to distribution shifts and model misspecification in the specific scientific application of the measurement of dark matter population-level parameters using strong gravitational lensing. We investigate the behaviour of a trained NRE for test data presenting distributional shifts inside the bounds of training, as well as out of distribution, both in the linear and non-linear parameters of this problem. While our results show that NREs perform when tested perfectly in distribution, we find that they exhibit significant biases and drawbacks when confronted with slight deviations from the examples seen in the training distribution. This indicates the necessity for caution when applying NREs to real astrophysical data, where underlying distributions are not perfectly known and models do not perfectly reconstruct the true underlying distributions.

1. Introduction

One of the most striking open problems in modern astrophysics is the fact that the nature of $\sim 80\%$ of the matter content of the universe is currently completely unknown (Hinshaw et al., 2013; Planck Collaboration et al., 2020). This mysterious form of matter, known as dark matter, is believed to be a new, completely invisible particle which interacts with regular matter only through gravity. Inferring the nature of this new particle is one of the main goals of the upcoming decade in the field of astrophysics and cosmology.

One of the most promising probes to tackle this problem is strong gravitational lensing, which is the formation of multiple images of distance light sources due to the deflection of their light rays by the gravity of some intervening structures. Both the lens and the line of sight between the observer and the source are populated with dark matter halos, which are clumps of dark matter. Due to its purely gravitational effect, strong lensing is sensitive to the matter distribution on sub-galactic scales and can independently probe the mass distribution of matter, regardless of its light-emitting properties. Therefore, studying small deflections in strong lenses can allow us to map out the distribution of dark matter inside lensing galaxies, which, in turn, can inform us about the particle properties of dark matter.

Measuring the effect of these small halos on lensed images, however, is a very difficult, non-linear inverse problem, as the signal is very weak and suffers from multiple degeneracies with other nuisance parameters (for example, the morphology of the background source). Therefore, while it is possible to constrain the total mass in the lens, referred to as the Einstein radius, with a few percent uncertainty (e.g., Koopmans et al., 2006; Treu et al., 2009; Auger et al., 2010; Brownstein et al., 2012; Bolton et al., 2012; Sonnenfeld et al., 2013; Shu et al., 2016; 2017; Talbot et al., 2021), individual detections of dark matter halos have only been detected twice (Vegetti et al., 2010; Hezaveh et al., 2016b) even though of the order of 10,000s are predicted to be in every Milky Way-sized galaxy.

With a new generation of astrophysical surveys like the Legacy Survey of Space and Time (LSST) at the Vera C. Rubin Observatory and the Euclid space telescope in the upcoming decade, the number of known strong lensing systems is expected to grow from a few hundred to approximately 170,000 systems (Serjeant, 2014; Collett, 2015). This increasing number has the potential to allow putting very tight constraints on the particle properties of dark matter given

[1]Ciela Institute, Montréal, Canada [2]Mila - Quebec Artificial Intelligence Institute, Montréal, Canada [3]Department of Physics, Université de Montréal, Montréal, Canada [4]Center for Computational Astrophysics, Flatiron Institute, New York, USA [5]Perimeter Institute for Theoretical Physics, Waterloo, Canada [6]Trottier Space Institute, McGill University, Montréal, Canada. Correspondence to: Andreas Filipp <andreas.filipp@umontreal.ca>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria.

an increased statistical power.

Inferring dark matter population parameters with strong lensing is a high-dimensional problem characterized by an intractable likelihood. This makes it difficult to apply Bayesian statistics to the analysis methods. The properties of the individual subhalos correspond to a high-dimensional space that needs to be marginalized to obtain dark matter properties at a population level. The marginalization is challenging due to the diversity of observed lenses and the complexity involved in computing likelihoods for different dark matter models in such a high-dimensional space. Some data analysis methods rely on detecting and studying the amplitude of spatial fluctuations through a power spectrum decomposition (e.g., Hezaveh et al., 2016a; Cyr-Racine et al., 2016; Brennan et al., 2019), or employ summary statistic approaches (e.g., Birrer et al., 2017). These methods all attempt to reduce the dimensionality of the likelihood function, often based on intuition. The reduction and underlying assumptions can lead to biased results in the inference.

With the upcoming data increasing in volume as well as in complexity, simple parametric traditional lens analysis methods come to their limits, both because they are computationally expensive and not sufficiently expressive to capture the complexity of the data produced by these new observatories.

As a result, the use of machine learning methods, and in particular simulation-based inference methods, has significantly increased in recent years (e.g., Perreault Levasseur et al., 2017; Hezaveh et al., 2017; Cyr-Racine et al., 2019; Ostdiek et al., 2020; Legin et al., 2023; Adam et al., 2023a;b; Karchev et al., 2023; Wagner-Carena et al., 2023; 2024). In the field of dark matter inference, an alternative approach involves probing the population-wide effects of dark matter on a sample of strong gravitational lenses, rather than searching for individual detections of dark matter substructure.

Machine learning methods, which approximate the intractable likelihood or directly learn the posterior distribution from a simulated dataset, emerge as powerful tools for analyzing the high-dimensional data spaces without reducing the dimensionality and risking the loss of valuable information in the observations. Neural networks, such as neural ratio estimators (NREs) and neural posterior estimators (NPEs), can be trained to obtain the intractable likelihood or the posterior of parameter distributions directly, without reducing the dimensionality of the input data. Both approaches can successfully marginalize nuisance parameters and return the likelihood or posterior for the parameter of interest. These neural networks types have become very popular in strong gravitational lensing and are a promising avenue to circumvent the calculation of the intractable likelihood (e.g., Brehmer et al., 2020; 2019; Coogan et al., 2022; Mishra-Sharma, 2022; Zhang et al., 2022; Karchev et al., 2023; Wagner-Carena et al., 2023; 2024; Zhang et al., 2024).

In this work, we investigate and demonstrate the limitations of NREs to data originating from out-of-distribution or small deviations from the training distribution. As an example case, we do this in the context of strong lens analysis to determine population parameters of the nature of dark matter. We illustrate the effect on the posterior calculation from NRE outputs in cases of imperfect knowledge of the underlying data distribution and data originating from parameters slightly outside the training distribution.

2. Methods

2.1. Data Generation

Strong gravitational lensing is the formation of multiple images of distance light sources due to the deflection of their light rays by the gravity of intervening structures, the lens. To produce simulated images of this effect, both a background source image and a matter distribution for the lens need to be generated. The lensed imaged is then obtained using a ray-tracing simulation through the lens density, which we accomplish using the simulation package Caustics (Stone et al., 2024).

To model the light of the background source, we use an ensemble of Sérsic profiles (Sérsic, 1963). We combine between 5 to 50 Sérsic profiles with a variety in sizes and shapes to a single source, to mimic the complexity and diversity of real lensed galaxies. For the main deflector, we use a singular isothermal ellipsoid (SIE) profile (Barkana, 1998), and for the dark matter halos, we use Narrov-Frank-White (NFW) profiles (Navarro et al., 1997). All dark matter halos are sampled from a subhalo function that accounts only for subhalos in the main deflector.

Dark matter subhalos are concentrations of dark matter orbiting larger mass halos, typically encompassing galaxies or clusters of galaxies. The size of these dark matter subhalos is comparable to satellite galaxies, but do not necessarily contain stars or other visible matter. Understanding substructures in galaxies, including the particle properties of dark matter and their potential redshift evolution, is a complex and ongoing challenge. Individual subhalos and their population properties are influenced by the underlying cosmology, dark matter particle properties, the host galaxy, and other environmental factors. Modelling all these aspects of dark matter subhalos is beyond the scope of this work. Instead, we approximate the subhalo mass function with an analytical power-law profile, to focus on the methodology and to facilitate the study effects of distributional shifts on NRE predictions.

The standard model of cosmology, the ΛCDM model, pre-

dicts a scale-invariant power spectrum of primordial fluctuations. From that subhalos follow approximately a logarithmic mass distribution

$$\frac{\mathrm{d}n}{\mathrm{d}\log m_{\mathrm{halo}}} = \alpha \cdot M_{\mathrm{host}} \cdot m_{\mathrm{halo}}^{\beta} \tag{1}$$

α is a normalization constant of the profile, M_{host} is the mass of the host galaxy, and m_{halo} is the subhalo mass. To quantify the abundance of dark matter, we introduce the parameter f_{sub}, which determines the ratio of mass in the dark matter subhalos to the mass in the host galaxy:

$$f_{\mathrm{sub}} = \frac{\int dm_{\mathrm{halo}} \frac{\mathrm{d}n}{\mathrm{d}m_{\mathrm{halo}}}}{M_{\mathrm{host}}} \tag{2}$$

This simplified dark matter halo mass function allows efficient sampling of dark matter subhalos across different dark matter models for our study. The parameters f_{sub} and β define the dark matter population. We allow to vary them and aim to infer those from simulated images using an NRE. The predicted parameters for the power-law approximation from ΛCDM are $f_{\mathrm{sub}} \approx 0.05$ and $\beta \approx -0.9$ (Dalal & Kochanek, 2002; Madau et al., 2008; Springel et al., 2008; Hiroshima et al., 2018; Hsueh et al., 2020).

We aim to obtain posterior predictions on dark matter parameters, based on the anticipated data quality from the full 10-year survey of LSST. LSST is a ground-based survey designed to map the entire southern hemisphere, producing approximately 20 TB of data each night, with images taken in the optical wavelengths across the u-, g-, r-, i-, z-, and y-bands. The expected pixel size is 0.2 arcseconds, and the effective zenith seeing in LSST in r-band is poised to be approximately 0.83, which can be approximated by a Gaussian point spread function (PSF) with a mean of $\mu_{\mathrm{PSF}} = 0.83$. The PSF is convolved with the simulated image to imitate the atmospheric blurring. The instrumental zero-point in the r-band is approximated at a magnitude of 28.36, and the average night sky brightness in the r-band is expected to be $mag_{\mathrm{sky}} = 21.20$. The instrumental readout noise per pixel can be approximated by $\sigma^2_{\mathrm{inst}} = (\mathrm{readnoise}^2 + (DC * t)) * n_{\mathrm{exp}}$ [1], with an upper limit on the readout noise of readnoise $= 8.8$ electrons/second/pixel, a maximum dark current of $DC = 0.2$ electrons/second/pixel, a planned exposure time of $t = 15$ seconds and approximately 300 visits, each with 2 exposures, leading to $n_{\mathrm{exp}} = 600$ (Ivezić et al., 2019).

Figure 1 shows an example of all relevant components in our simulation of the training data. From left to right, the figure displays a convergence map κ_{SIE} generated by an SIE profile within the training data distribution, a convergence map of a sample of dark matter halos κ_{NFW} created using the dark matter parameters $f_{\mathrm{sub}} = 0.05$ and $\beta = -0.9$, which align with the predictions of ΛCDM, a realization of the source light, the lensed image derived from the provided kappa maps and source light without PSF convolution and noise, and finally, the lensed image with PSF convolution and noise. As noted earlier, the data quality of the final image, with noise and PSF convolution, is based on the expected LSST data quality in the r-band.

Table 1 illustrates the training parameter distribution for the lens and source profiles used to generate the training sample for the NRE. The parameters are drawn from either a uniform or normal distribution. The notation $\mathcal{N}(\mu, \sigma)$ indicates a normal distribution with mean μ and standard deviation σ, and $\mathcal{U}[a - b]$ denotes a uniform distribution between a and b.

Table 1: Parameter distribution for the simulation of the main deflector and the source light. $\mathcal{N}(\mu, \sigma)$ indicates a normal distribution with mean μ and standard deviation σ, and $\mathcal{U}[a - b]$ denotes a uniform distribution between a and b.

Parameter	**Distribution**
Lens galaxy	
Einstein radius θ_{E}	$\mathcal{U}[1.0 - 1.5]$
Axis ratio q_{SIE}	$\mathcal{U}[0.5 - 0.99]$
Orientation angle ϕ_{SIE}	$\mathcal{U}[0.0 - \pi]$
Lens center $(\hat{x}_{\mathrm{SIE}}, \hat{y}_{\mathrm{SIE}})$	$(0, 0)$
Source light	
Number of sources N	$\mathcal{U}[5 - 50]$
Magnitude mag_{source}	$\mathcal{N}(23.5, 0.1)$
Sérsic index $n_{\mathrm{sérsic}}$	$\mathcal{N}(2.5, 0.5) \geq 0.8$
Axis ratio $q_{\mathrm{sérsic}}$	$\mathcal{U}[0.5 - 0.99]$
Orientation angle $\phi_{\mathrm{sérsic}}$	$\mathcal{U}[0.0 - \pi]$
Sérsic radius $R_{\mathrm{sérsic}}$	$\mathcal{N}(0.5, 0.3) \geq 0.05$
Source center $\hat{x}_{\mathrm{source}}, \hat{y}_{\mathrm{source}}$	$\mathcal{N}(0.0, 0.1)$

Figure 2 displays 40 random samples of strong lens images generated from the training data parameter distribution outlined in Table 1. The images show the variety of possible strong lens configurations that the network learns to marginalize. The lens systems include closed Einstein rings, doubly imaged sources, as well as quadruply imaged sources. This diversity in strong lens populations can also be seen in real observations of lensed systems.

2.2. Posterior Estimation

To obtain the posterior distribution $p(\vartheta|\{x_i\})$ for a set of observed images $\{x_i\}$, which are generated with the param-

[1] https://smtn-002.lsst.io/v/v2.0/index.html

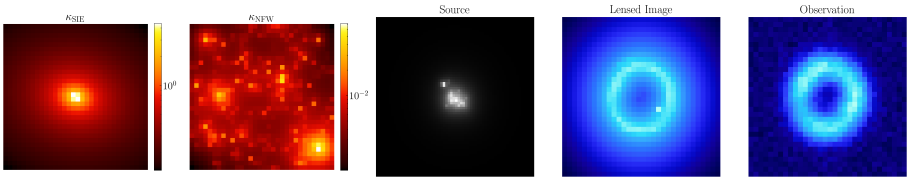

Figure 1: From the left to the right: a convergence map $\kappa_{\rm SIE}$ created by an SIE profile within the training data distribution, a convergence map of a sample of dark matter halos $\kappa_{\rm NFW}$ created with the dark matter parameters $f_sub = 0.05$ and $\beta = -0.9$ corresponding to ΛCDM, a realization of the source light, the lensed image without PSF convolution and noise, and the lensed image with PSF convolution and noise.

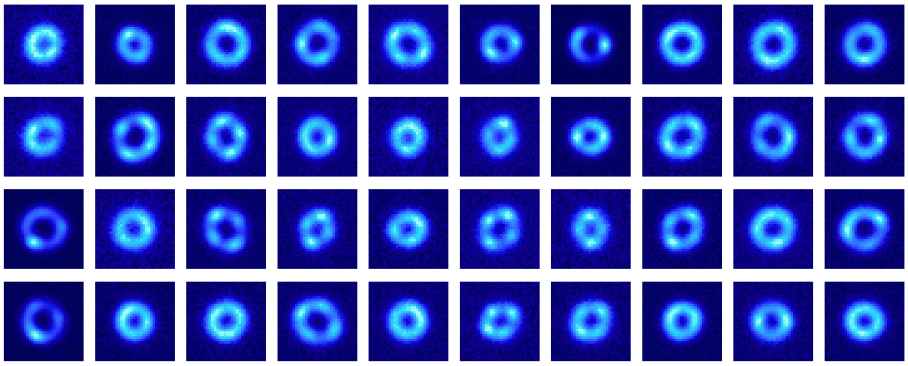

Figure 2: A sample of 40 random strong lens configurations generated from the training data parameter distribution. The lens systems include a wide variety of closed Einstein rings, doubly imaged sources, as well as quadruply imaged sources. The broad diversity is also seen in real observed lens systems.

eters of interest ϑ, we apply Bayes' theorem:

$$p(\vartheta|\{x_i\}) = \frac{p(\vartheta)\prod_i p(x_i|\vartheta)}{\int \mathrm{d}\vartheta' p(\vartheta')\prod_i p(x_i|\vartheta')} = p(\vartheta)\left[\int \mathrm{d}\vartheta' p(\vartheta')\prod_i \frac{p(x_i|\vartheta')}{p(x_i|\vartheta)}\right]^{-1} \quad (3)$$

with $p(\vartheta)$ is the prior on the parameters of interest.

The observations $\{x_i\}$ can explicitly depend on intermediate model parameters θ, which in turn depend on the parameters of interest ϑ. This is the case for the inference of dark matter population parameters in strong lensing, where the parameters defining the lens are the intermediate model parameters and those defining the dark matter population are of primary interest. The likelihood $p(x_i|\vartheta)$ is then the marginal of the likelihood $p(x_i, \theta|\vartheta)$ over all intermediate model parameters θ

$$p(x_i|\vartheta) = \int \mathrm{d}\theta p(x_i, \theta|\vartheta) \quad (4)$$

The marginalization integral is intractable and therefore the likelihood $p(x_i|\vartheta)$ is intractable as well, due to the high-dimensional nature of the likelihood $p(x_i, \theta|\vartheta)$:

$$p(x, \theta|\vartheta) = p_{\rm lens}(\theta_{\rm lens}) \times Pois(n_h|\bar{n}_h(\vartheta))\prod_i^n [p_{\rm mass}(m_{h,i}|\vartheta)p(r_i)] \times p_{\rm obs}(x|f(\text{lens}, \text{halos})) \quad (5)$$

with $p_{\rm lens}$ the likelihood of the latent lens parameters $\theta_{\rm lens}$, $\bar{n}_h(\vartheta)$ the expected number of subhalos as a function of the parameter of interest $\vartheta = (f_{\rm sub}, \beta)$ defining the subhalo mass function. n_h is the realized number in a specific simulation, and $m_{h,i}$ and r_i are the masses and position of the i-th subhalo in the simulated image. $p_{\rm mass}$ is the normalized subhalo mass function, and $p_{\rm obs}$ is the probability of observing the specific image x, taking into account Poisson fluctuation and the point spread function.

NREs can learn to marginalize all latent parameters θ during training and thus circumvent the calculation of the intractable integral, enabling the posterior calculation from equation 3.

2.3. Neural Ratio Estimator

The NRE architecture employed in this work follows the approach of a ResNet with MLPs, as used in previous works

(e.g., Brehmer et al., 2019; 2020; Zhang et al., 2022; Coogan et al., 2022; Mishra-Sharma, 2022; Karchev et al., 2023; Zhang et al., 2024). In the context of analyzing dark matter population parameters from strongly lensed images, the parameters handled by the NRE are categorized as follows:

Nuisance parameters θ: These include parameters defining the shape of the lensed images, such as the main deflector parameters, the source parameters, as well as the positions, masses, and numbers of individual dark matter subhalos.

Parameter of interest ϑ: These include β and $f_{\rm sub}$, which define the dark matter mass function from equations 1 and 2.

Observables x: These are the simulated images of strong lens systems, generated with θ and ϑ. Examples of these are shown in figure 2.

Rather than directly learning the likelihood of the intractable integral in equation 4 or the posterior from equation 3, the NRE estimates the marginal likelihood-to-evidence ratio $r(x|\vartheta)$

$$r(x|\vartheta) = \frac{p(x|\vartheta)}{p_{\rm ref}(x)} \tag{6}$$

with $p_{\rm ref}(x)$ the evidence, which is the reference likelihood of an observation x occurring under any possible dark matter parameter ϑ.

$$p_{\rm ref}(x) = \int d\vartheta' \pi(\vartheta') p(x|\vartheta') \tag{7}$$

here, $p(x|\vartheta)$ is the likelihood of an observation x occurring given a specific set of the dark matter parameters ϑ, and $\pi(\vartheta')$ is the proposal distribution of the dark matter parameters, used in training data generation.

Predicting the marginal likelihood-to-evidence ratio $r(x|\vartheta)$, rather than the intractable likelihood or the posterior directly, allows for more efficient training. Given the ratio by equation 6, the likelihood from equation 5 simplifies to only the term in the second line because the other terms cancel out, as they do not depend on the parameters of interest ϑ. This simplification can be used as the training objective of the NRE to more efficiently train the NRE on the non-marginalized likelihood-to-evidence ratio of a given specific observation $r(x, \theta|\vartheta) = \frac{p(x,\theta|\vartheta)}{p_{\rm ref}(x,\theta)}$ if the mass distribution of the dark matter halos is analytic and the probability of sampled subhalo masses for a given lens, and the probability to find them in their specific positions, are easily to infer. This approach has been successfully implemented in previous works (e.g., Brehmer et al., 2020; 2019).

The marginal likelihood-to-evidence ratio $r(x|\vartheta)$ can be used to calculate the posterior from equation 3 by replacing the fraction of likelihoods with the fraction of marginal likelihood-to-evidence ratios:

$$\frac{p(x_i|\vartheta')}{p(x_i|\vartheta')} = \frac{\frac{p(x_i|\vartheta')}{p_{\rm ref}(x)}}{\frac{p(x_i|\vartheta)}{p_{\rm ref}(x)}} = \frac{r(x_i|\vartheta')}{r(x_i|\vartheta)} \tag{8}$$

Now the integral leading to the posterior is tractable and low dimensional. The parameter space that needs to be integrated over, defined by the parameter of interest ϑ, is two-dimensional and the prior on ϑ is known and can even be varied.

The NRE, in theory, predicts the marginal likelihood-to-evidence ratio perfectly in the limit of unlimited training data and perfect learning. However, this limit is unreachable. Nonetheless, to ensure the reliability and accuracy of the NRE at inference, the trained NRE can and needs to be calibrated. Calibration addresses biases, over- or under-confidences, and discrepancies in the uncalibrated outputs produced by the NRE. The calibration procedure follows previous works (e.g., Cranmer et al., 2016; Brehmer et al., 2020; 2019; Karchev et al., 2023).

3. Results and Discussion

We evaluate the calibrated NRE on lenses drawn from the training distribution as well as several variations. The resulting posteriors are shown in Figure 3. All images were generated with the ground truth parameters, denoted by the red star, at $f_{\rm sub} = 0.05$ and $\beta = -0.9$, in line with ΛCDM predictions (Dalal & Kochanek, 2002; Madau et al., 2008; Springel et al., 2008; Hiroshima et al., 2018; Hsueh et al., 2020).

For the tests, all lens and source parameters are sampled from the training distribution from Table 1, with modifications applied to only single parameter distributions. All changes in the parameter distributions are within reasonable and minor deviations from the training distribution. For most modifications, the changes in the observed image are minimal to non-visible. The specific modified parameters for each example in Figure 3 and their new, modified distribution are listed in Table 2.

The left-hand side of Figure 3 shows variations in the lens plane and noise levels. All rows show the posterior inferred from 50, 100, and 500 lens systems. The top row on the left side illustrates the posterior inferred from lenses generated from the training distribution for 50, 100, and 500 lens systems combined. The inferred posterior for lenses generated with parameters taken exactly from the training distribution is unbiased and recovers the ground truth well, meaning that even when combining the inference over thousands of lensed systems (which is significantly more than shown here) the inference remains unbiased.

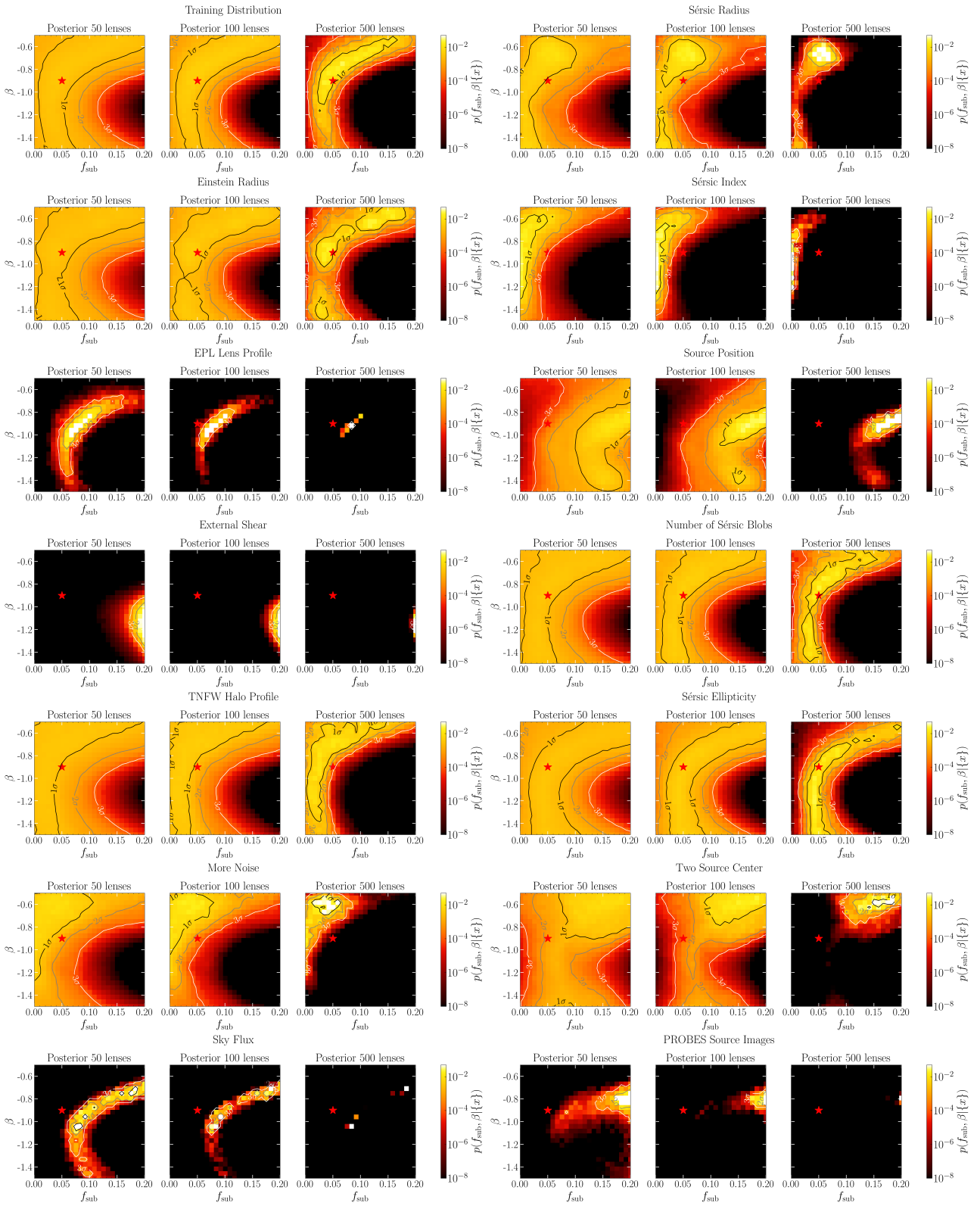

Figure 3: Evaluation of the NRE on lenses drawn from the training distribution and minor variations in the underlying parameter distributions. All evaluation datasets were generated with the ground truth, denoted by the red star, at $f_{\rm sub} = 0.05$ and $\beta = -0.9$, which corresponds to ΛCDM predictions. The modification of the parameter distributions of the individual plots can be found in Table 2

Table 2: Variation of parameter distributions used in Figure 3 in comparison to the training distribution from 1. First, the left side of the Figure is explained, then on the right side, from top to bottom.

Plot Title	**Modified Parameter**	**New Distribution**
Training Distribution	None	
Einstein Radius	Einstein radius θ_E	$\mathcal{N}(1.4, 0.2) \geq 1.0, \leq 1.5$
EPL Lens Profile	Power-law slope of lens $\gamma_{\rm EPL}$	$\mathcal{N}(2.0, 0.02)$
External Shear	Added external shear components $\gamma_{1/2}$	$\mathcal{N}(0.0, 0.1)$
TNFW Halo Profile	Truncation of NFW, truncation scale τ	5
More Noise	Noise in the image at 1% of the sky flux	
Sky Flux	Different magnitude of sky background $mag_{\rm sky}$	21.4
Sérsic Radius	Sérsic radius $R_{\rm sérsic}$	$\mathcal{N}(0.2, 0.2) \geq 0.05$
Sérsic Index	Sérsic index $n_{\rm sérsic}$	$\mathcal{N}(5.5, 1.5) \geq 0.8$
Source Position	Source centers $(\hat{x}_{\rm source}, \hat{y}_{\rm source})$	$\mathcal{N}(0.05, 0.15)$
Number of Sérsic Blobs	Number of Sérsic blobs in the source N	$\mathcal{U}[40 - 45]$
Sérsic Ellipticity	Source axis ratio $q_{\rm sérsic}$	$\mathcal{N}(0.75, 0.2) \geq 0.5, \leq 0.99$
Two Source Center	Source position $(\hat{x}_{\rm source1}, \hat{y}_{\rm source1}), (\hat{x}_{\rm source2}, \hat{y}_{\rm source2})$	$\mathcal{N}_1(0.0, 0.1), \mathcal{N}_2(-0.2, 0.1)$
Probes Dataset as Sources	Analytic Sérsic source profiles to Probes dataset	

In the second row of Figure 3, we vary the Einstein radius distribution from a uniform distribution to a normal distribution with mean at $\mu = 1.4$ and standard deviation of $\sigma = 0.2$, but maintaining the same boundaries as in the uniform training distribution. The inferred posterior for 50 and 100 lenses accurately recovers the ground truth but for the case of 500 systems, multiple posterior modes develop. This is an artifact of the distribution shift and causes the posterior to have a bias detectable through coverage tests such as TARP (Lemos et al., 2023).

The third row shows the posterior for the case of changing the main deflector model from a SIE to an elliptical power law (EPL) profile, with the mass slope drawn from a normal distribution with mean $\mu = 2.0$ and standard deviation $\sigma = 0.02$. For the case of the mass slope of $\gamma_{\rm EPL} = 2.0$ the EPL profile corresponds exactly to an SIE profile. Even though the lens profile exhibits minimal variation, with the standard deviation smaller than uncertainties on the mass slope obtained from traditional lens modelling (e.g., Koopmans et al., 2006; Treu et al., 2009; Auger et al., 2009; 2010; Brownstein et al., 2012; Bolton et al., 2012; Sonnenfeld et al., 2013; Shu et al., 2017; Chen et al., 2019; Talbot et al., 2021), the posterior for as little as 50 lenses is biased, and the ground truth lies not even within the 3σ area.

For the posterior in the fourth row, we introduce a small external shear on top of the main deflector. The external shear is defined by two components $\gamma_{1,2}$. We add the shear with a median of $\mu = 0.0$ and standard deviation of $\sigma = 0.1$. In traditional strong lens analysis, the external shear is highly degenerate with the ellipticity of the lens profile and is usually used to model the environment of the lens (e.g., Oguri et al., 2005; Koopmans et al., 2006; Auger et al., 2009; Sonnenfeld et al., 2013; Shu et al., 2016; 2017; Talbot et al., 2021). The small addition of external shear significantly biases the network already for only 50 strong lenses heavily, pushing the obtained posterior to the boundaries of the parameters of interest.

The fifth row changes the substructure mass profiles from NFW profiles to truncated NFW profiles with a truncation scale of 5, i.e. the radius at which the mass profile gets truncated is five times the scale radius. This modification converts the infinite expanded NFW profiles into finite profiles. The resulting posterior shows clear biases, with the ground truth outside the 1σ area even when combining as little as 50 strong lenses.

In the sixth row, we examine the effect of adding Gaussian noise at a one percent level of the sky flux. The minor variation in noise biases the network output strongly, excluding the ground truth in the 1σ area for already a sample of only 50 lens systems.

The seventh and last row on the left side shows the impact of reducing the sky flux in the simulations by setting it to a higher magnitude, resulting in higher signal-to-noise in observations. The posterior estimation from that is biased, with the network compressing towards a single value. This outcome highlights the sensitivity of the network to variations in sky background, which can vary widely across different nights and sky areas.

The right-hand side of Figure 3 displays variations in the source plane, with posterior results obtained from combining 50, 100, and 500 strong lens systems.

The first row varies the Sérsic radius of the individual Sérsic source components to a normal distribution with a lower

mean than the training distribution. This makes the source images appear more structured and in general smaller. One can see a noticeable bias on the obtained posteriors from the network's output.

In the second row, the distribution from which the Sérsic indices for individual source components are sampled is varied. We generate the source images from a normal distribution centred at a slightly different value than in the training distribution. A higher Sérsic index leads to a smoother profile and more flux associated with outer regions of the galaxy, resulting in a less structured appearance of the source. The posterior obtained is biased and over 3σ offset from the ground truth for 100 combined lens systems.

For the posteriors in the third row, the mean of the source position is set a little bit off-center, with a higher variation than in the training distribution. This affects the source images by producing a less focused source. This small shift leads to a highly biased posterior estimation.

In the fourth row, we modify the number of Sérsic blobs used to generate our sources from a range of 5-50 blobs to 40-45 blobs. Although this variation is within the training distribution, it results in a higher inclusion of more structured sources, yielding a generally broader posterior than for the training distribution. The ground truth still can be found within the 1σ interval, but the posterior still appears biased compared to that obtained from the training distribution.

The fifth row shows the impact when changing the prior on the source axis ratio from a uniform to a normal prior, centred in the middle of the distribution with the same boundaries as the uniform training distribution. Compared to the posterior obtained from the correct training distribution, the posterior is slightly biased but still contains the ground truth within the 1σ area. This very subtle and small change in the underlying distribution of the ellipticity of the source profile components has a visible and non-neglectable effect on the posterior prediction.

In the sixth row, the source images are generated with two different centers for the Sérsic profiles, distributed randomly in two locations: one remaining centred as in the training distribution at $(\hat{x}\mathrm{source1}, \hat{y}\mathrm{source1}) = (0.0, 0.0)$ arcsec, and the other slightly offset to a mean position of $(\hat{x}\mathrm{source2}, \hat{y}\mathrm{source2}) = (-0.2, -0.2)$ arcsec, both with a standard deviation of $\sigma = 0.1$. The structure of the source images in this setup is more peanut-shaped than round. The results of the posterior estimation are strongly biased.

The seventh and last row on the right side shows the posterior if we use pixelated source images based on real data from the PROBES dataset, a high-quality sample of 2059 galaxies with 256 x 256 pixels (Stone & Courteau, 2019; Stone et al., 2021). This approach introduces significantly more structure and realism than the training dataset, resulting in out-of-distribution data and biased posterior estimations.

Since variations of any individual parameter already produce biases in the inferred posteriors, we do not include tests on variations of multiple parameters at the same time. The comparison of these tests reveals that very small variations with different effects on the lensed image can have the similar effect of very strongly biasing the predicted posterior, whereas variations with very similar effects can introduce highly different effects on the bias and shape of the posterior prediction. It is important to stress that the initial parametrization used to train the NRE is within the range of usual state-of-the-art parametrization for lens modelling and that the deviations tested here are mostly physically very small and could very well be expected to happen in a realistic data analysis setting where the ground truth is not available. Meanwhile, the effect of those changes in the underlying parameter distributions on the NRE predictions is highly non-linear and unpredictable, making it difficult to trust the posteriors inferred with this model. This makes NREs only applicable to cases where both the underlying distribution of parameters and the physical model are perfectly known. For strong lensing and astrophysical data analysis more generally, this is very rarely the case.

4. Conclusion

In conclusion, our findings demonstrate that, when analyzing strong lensing data to measure dark matter properties, NREs perform effectively provided the testing data originates from a distribution and a physical model identical to that of the training and calibration data. In this case, they provide unbiased posteriors that can be combined in population-level inference over a large number of individual observations, significantly outperforming traditional methods. However, they fail to generalize to even slightly out-of-distribution data or even to data that are within the distribution but with distributions drawn from slightly different priors. These limitations highlight the need for caution when applying NREs to real astrophysical data, where such ideal conditions are seldom met.

This indicates that NREs are a well-suited tool for research requiring only simulated data, where the underlying distribution is perfectly known. For instance, NREs are excellent for making sensitivity predictions on for specific signals e.g. in survey data, since they seamlessly allow for amortized implicit marginalization over nuisance parameters.

Beyond the fields of strong lensing and astrophysics, this case study is also an important lesson for other fields of science where accuracy of inference is important and the true data-generating process is not perfectly known.

Acknowledgements

This work is partially supported by Schmidt Futures, a philanthropic initiative founded by Eric and Wendy Schmidt as part of the Virtual Institute for Astrophysics (VIA). The work is in part supported by computational resources provided by Calcul Quebec and the Digital Research Alliance of Canada. Y.H. and L.P. acknowledge support from the Canada Research Chairs Program, the National Sciences and Engineering Council of Canada through grants RGPIN-2020-05073 and 05102, and the Fonds de recherche du Québec through grants 2022-NC-301305 and 300397. A.F. acknowledges the support from the Bourse J. Armand Bombardier.

References

Adam, A., Perreault-Levasseur, L., Hezaveh, Y., and Welling, M. Pixelated Reconstruction of Foreground Density and Background Surface Brightness in Gravitational Lensing Systems Using Recurrent Inference Machines. , 951(1):6, July 2023a. doi: 10.3847/1538-4357/accf84.

Adam, A., Stone, C., Bottrell, C., Legin, R., Hezaveh, Y., and Perreault-Levasseur, L. Echoes in the Noise: Posterior Samples of Faint Galaxy Surface Brightness Profiles with Score-Based Likelihoods and Priors. *arXiv e-prints*, art. arXiv:2311.18002, November 2023b. doi: 10.48550/arXiv.2311.18002.

Auger, M. W., Treu, T., Bolton, A. S., Gavazzi, R., Koopmans, L. V. E., Marshall, P. J., Bundy, K., and Moustakas, L. A. The Sloan Lens ACS Survey. IX. Colors, Lensing, and Stellar Masses of Early-Type Galaxies. , 705(2): 1099–1115, November 2009. doi: 10.1088/0004-637X/705/2/1099.

Auger, M. W., Treu, T., Bolton, A. S., Gavazzi, R., Koopmans, L. V. E., Marshall, P. J., Moustakas, L. A., and Burles, S. The Sloan Lens ACS Survey. X. Stellar, Dynamical, and Total Mass Correlations of Massive Early-type Galaxies. , 724(1):511–525, November 2010. doi: 10.1088/0004-637X/724/1/511.

Barkana, R. Fast Calculation of a Family of Elliptical Mass Gravitational Lens Models. , 502(2):531–537, August 1998. doi: 10.1086/305950.

Birrer, S., Amara, A., and Refregier, A. Lensing substructure quantification in RXJ1131-1231: a 2 keV lower bound on dark matter thermal relic mass. , 2017(5):037, May 2017. doi: 10.1088/1475-7516/2017/05/037.

Bolton, A. S., Brownstein, J. R., Kochanek, C. S., Shu, Y., Schlegel, D. J., Eisenstein, D. J., Wake, D. A., Connolly, N., Maraston, C., Arneson, R. A., and Weaver, B. A. The BOSS Emission-Line Lens Survey. II. Investigating Mass-density Profile Evolution in the SLACS+BELLS Strong Gravitational Lens Sample. , 757(1):82, September 2012. doi: 10.1088/0004-637X/757/1/82.

Brehmer, J., Mishra-Sharma, S., Hermans, J., Louppe, G., and Cranmer, K. Mining for Dark Matter Substructure: Inferring Subhalo Population Properties from Strong Lenses with Machine Learning. , 886(1):49, November 2019. doi: 10.3847/1538-4357/ab4c41.

Brehmer, J., Louppe, G., Pavez, J., and Cranmer, K. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10):5242–5249, 2020. doi: 10.1073/pnas.1915980117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1915980117.

Brennan, S., Benson, A. J., Cyr-Racine, F.-Y., Keeton, C. R., Moustakas, L. A., and Pullen, A. R. Quantifying the power spectrum of small-scale structure in semi-analytic galaxies. , 488(4):5085–5092, October 2019. doi: 10.1093/mnras/stz1607.

Brownstein, J. R., Bolton, A. S., Schlegel, D. J., Eisenstein, D. J., Kochanek, C. S., Connolly, N., Maraston, C., Pandey, P., Seitz, S., Wake, D. A., Wood-Vasey, W. M., Brinkmann, J., Schneider, D. P., and Weaver, B. A. The BOSS Emission-Line Lens Survey (BELLS). I. A Large Spectroscopically Selected Sample of Lens Galaxies at Redshift ~0.5. , 744(1):41, January 2012. doi: 10.1088/0004-637X/744/1/41.

Chen, Y., Li, R., Shu, Y., and Cao, X. Assessing the effect of lens mass model in cosmological application with updated galaxy-scale strong gravitational lensing sample. , 488(3):3745–3758, September 2019. doi: 10.1093/mnras/stz1902.

Collett, T. E. The Population of Galaxy-Galaxy Strong Lenses in Forthcoming Optical Imaging Surveys. , 811(1): 20, September 2015. doi: 10.1088/0004-637X/811/1/20.

Coogan, A., Montel, N. A., Karchev, K., Grootes, M. W., Nattino, F., and Weniger, C. One never walks alone: the effect of the perturber population on subhalo measurements in strong gravitational lenses, 2022.

Cranmer, K., Pavez, J., and Louppe, G. Approximating likelihood ratios with calibrated discriminative classifiers, 2016.

Cyr-Racine, F.-Y., Sigurdson, K., Zavala, J., Bringmann, T., Vogelsberger, M., and Pfrommer, C. ETHOS—an effective theory of structure formation: From dark particle physics to the matter distribution of the Universe. , 93(12): 123527, June 2016. doi: 10.1103/PhysRevD.93.123527.

Cyr-Racine, F.-Y., Keeton, C. R., and Moustakas, L. A. Beyond subhalos: Probing the collective effect of the Universe's small-scale structure with gravitational lensing. , 100(2):023013, July 2019. doi: 10.1103/PhysRevD.100.023013.

Dalal, N. and Kochanek, C. S. Direct Detection of Cold Dark Matter Substructure. , 572(1):25–33, June 2002. doi: 10.1086/340303.

Hezaveh, Y., Dalal, N., Holder, G., Kisner, T., Kuhlen, M., and Perreault Levasseur, L. Measuring the power spectrum of dark matter substructure using strong gravitational lensing. , 2016(11):048, November 2016a. doi: 10.1088/1475-7516/2016/11/048.

Hezaveh, Y. D., Dalal, N., Marrone, D. P., Mao, Y.-Y., Morningstar, W., Wen, D., Blandford, R. D., Carlstrom, J. E., Fassnacht, C. D., Holder, G. P., Kemball, A., Marshall, P. J., Murray, N., Perreault Levasseur, L., Vieira, J. D., and Wechsler, R. H. Detection of Lensing Substructure Using ALMA Observations of the Dusty Galaxy SDP.81. , 823 (1):37, May 2016b. doi: 10.3847/0004-637X/823/1/37.

Hezaveh, Y. D., Perreault Levasseur, L., and Marshall, P. J. Fast automated analysis of strong gravitational lenses with convolutional neural networks. , 548(7669):555–557, August 2017. doi: 10.1038/nature23463.

Hinshaw, G., Larson, D., Komatsu, E., Spergel, D. N., Bennett, C. L., Dunkley, J., Nolta, M. R., Halpern, M., Hill, R. S., Odegard, N., Page, L., Smith, K. M., Weiland, J. L., Gold, B., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Tucker, G. S., Wollack, E., and Wright, E. L. Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results. , 208(2):19, October 2013. doi: 10.1088/0067-0049/208/2/19.

Hiroshima, N., Ando, S., and Ishiyama, T. Modeling evolution of dark matter substructure and annihilation boost. , 97(12):123002, June 2018. doi: 10.1103/PhysRevD.97.123002.

Hsueh, J. W., Enzi, W., Vegetti, S., Auger, M. W., Fassnacht, C. D., Despali, G., Koopmans, L. V. E., and McKean, J. P. SHARP - VII. New constraints on the dark matter free-streaming properties and substructure abundance from gravitationally lensed quasars. , 492(2):3047–3059, February 2020. doi: 10.1093/mnras/stz3177.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., Alonso, D., AlSayyad, Y., Anderson, S. F., Andrew, J., Angel, J. R. P., Angeli, G. Z., Ansari, R., Antilogus, P., Araujo, C., Armstrong, R., Arndt, K. T., Astier, P., Aubourg, É., Auza, N., Axelrod, T. S., Bard, D. J., Barr, J. D., Barrau, A., Bartlett, J. G., Bauer, A. E., Bauman, B. J., Baumont, S., Bechtol, E., Bechtol, K., Becker, A. C., Becla, J., Beldica, C., Bellavia, S., Bianco, F. B., Biswas, R., Blanc, G., Blazek, J., Blandford, R. D., Bloom, J. S., Bogart, J., Bond, T. W., Booth, M. T., Borgland, A. W., Borne, K., Bosch, J. F., Boutigny, D., Brackett, C. A., Bradshaw, A., Brandt, W. N., Brown, M. E., Bullock, J. S., Burchat, P., Burke, D. L., Cagnoli, G., Calabrese, D., Callahan, S., Callen, A. L., Carlin, J. L., Carlson, E. L., Chandrasekharan, S., Charles-Emerson, G., Chesley, S., Cheu, E. C., Chiang, H.-F., Chiang, J., Chirino, C., Chow, D., Ciardi, D. R., Claver, C. F., Cohen-Tanugi, J., Cockrum, J. J., Coles, R., Connolly, A. J., Cook, K. H., Cooray, A., Covey, K. R., Cribbs, C., Cui, W., Cutri, R., Daly, P. N., Daniel, S. F., Daruich, F., Daubard, G., Daues, G., Dawson, W., Delgado, F., Dellapenna, A., de Peyster, R., de Val-Borro, M., Digel, S. W., Doherty, P., Dubois, R., Dubois-Felsmann, G. P., Durech, J., Economou, F., Eifler, T., Eracleous, M., Emmons, B. L., Fausti Neto, A., Ferguson, H., Figueroa, E., Fisher-Levine, M., Focke, W., Foss, M. D., Frank, J., Freemon, M. D., Gangler, E., Gawiser, E., Geary, J. C., Gee, P., Geha, M., Gessner, C. J. B., Gibson, R. R., Gilmore, D. K., Glanzman, T., Glick, W., Goldina, T., Goldstein, D. A., Goodenow, I., Graham, M. L., Gressler, W. J., Gris, P., Guy, L. P., Guyonnet, A., Haller, G., Harris, R., Hascall, P. A., Haupt, J., Hernandez, F., Herrmann, S., Hileman, E., Hoblitt, J., Hodgson, J. A., Hogan, C., Howard, J. D., Huang, D., Huffer, M. E., Ingraham, P., Innes, W. R., Jacoby, S. H., Jain, B., Jammes, F., Jee, M. J., Jenness, T., Jernigan, G., Jevremović, D., Johns, K., Johnson, A. S., Johnson, M. W. G., Jones, R. L., Juramy-Gilles, C., Jurić, M., Kalirai, J. S., Kallivayalil, N. J., Kalmbach, B., Kantor, J. P., Karst, P., Kasliwal, M. M., Kelly, H., Kessler, R., Kinnison, V., Kirkby, D., Knox, L., Kotov, I. V., Krabbendam, V. L., Krughoff, K. S., Kubánek, P., Kuczewski, J., Kulkarni, S., Ku, J., Kurita, N. R., Lage, C. S., Lambert, R., Lange, T., Langton, J. B., Le Guillou, L., Levine, D., Liang, M., Lim, K.-T., Lintott, C. J., Long, K. E., Lopez, M., Lotz, P. J., Lupton, R. H., Lust, N. B., MacArthur, L. A., Mahabal, A., Mandelbaum, R., Markiewicz, T. W., Marsh, D. S., Marshall, P. J., Marshall, S., May, M., McKercher, R., McQueen, M., Meyers, J., Migliore, M., Miller, M., Mills, D. J., Miraval, C., Moeyens, J., Moolekamp, F. E., Monet, D. G., Moniez, M., Monkewitz, S., Montgomery, C., Morrison, C. B., Mueller, F., Muller, G. P., Muñoz Arancibia, F., Neill, D. R., Newbry, S. P., Nief, J.-Y., Nomerotski, A., Nordby, M., O'Connor, P., Oliver, J., Olivier, S. S., Olsen, K., O'Mullane, W., Ortiz, S., Osier, S., Owen, R. E., Pain, R., Palecek, P. E., Parejko, J. K., Parsons, J. B., Pease, N. M., Peterson, J. M., Peterson, J. R., Petravick, D. L., Libby Petrick, M. E., Petry, C. E., Pierfederici, F., Pietrowicz, S., Pike, R., Pinto, P. A., Plante, R., Plate, S., Plutchak, J. P., Price, P. A., Prouza, M., Radeka, V., Rajagopal, J., Rasmussen, A. P., Regnault,

N., Reil, K. A., Reiss, D. J., Reuter, M. A., Ridgway, S. T., Riot, V. J., Ritz, S., Robinson, S., Roby, W., Roodman, A., Rosing, W., Roucelle, C., Rumore, M. R., Russo, S., Saha, A., Sassolas, B., Schalk, T. L., Schellart, P., Schindler, R. H., Schmidt, S., Schneider, D. P., Schneider, M. D., Schoening, W., Schumacher, G., Schwamb, M. E., Sebag, J., Selvy, B., Sembroski, G. H., Seppala, L. G., Serio, A., Serrano, E., Shaw, R. A., Shipsey, I., Sick, J., Silvestri, N., Slater, C. T., Smith, J. A., Smith, R. C., Sobhani, S., Soldahl, C., Storrie-Lombardi, L., Stover, E., Strauss, M. A., Street, R. A., Stubbs, C. W., Sullivan, I. S., Sweeney, D., Swinbank, J. D., Szalay, A., Takacs, P., Tether, S. A., Thaler, J. J., Thayer, J. G., Thomas, S., Thornton, A. J., Thukral, V., Tice, J., Trilling, D. E., Turri, M., Van Berg, R., Vanden Berk, D., Vetter, K., Virieux, F., Vucina, T., Wahl, W., Walkowicz, L., Walsh, B., Walter, C. W., Wang, D. L., Wang, S.-Y., Warner, M., Wiecha, O., Willman, B., Winters, S. E., Wittman, D., Wolff, S. C., Wood-Vasey, W. M., Wu, X., Xin, B., Yoachim, P., and Zhan, H. LSST: From Science Drivers to Reference Design and Anticipated Data Products. , 873(2):111, March 2019. doi: 10.3847/1538-4357/ab042c.

Karchev, K., Trotta, R., and Weniger, C. SICRET: Supernova Ia Cosmology with truncated marginal neural Ratio EsTimation. , 520(1):1056–1072, March 2023. doi: 10.1093/mnras/stac3785.

Koopmans, L. V. E., Treu, T., Bolton, A. S., Burles, S., and Moustakas, L. A. The Sloan Lens ACS Survey. III. The Structure and Formation of Early-Type Galaxies and Their Evolution since z ~1. , 649(2):599–615, October 2006. doi: 10.1086/505696.

Legin, R., Hezaveh, Y., Perreault-Levasseur, L., and Wandelt, B. A Framework for Obtaining Accurate Posteriors of Strong Gravitational Lensing Parameters with Flexible Priors and Implicit Likelihoods Using Density Estimation. , 943(1):4, January 2023. doi: 10.3847/1538-4357/aca7c2.

Lemos, P., Coogan, A., Hezaveh, Y., and Perreault-Levasseur, L. Sampling-Based Accuracy Testing of Posterior Estimators for General Inference. *40th International Conference on Machine Learning*, 202:19256–19273, January 2023. doi: 10.48550/arXiv.2302.03026.

Madau, P., Diemand, J., and Kuhlen, M. Dark Matter Subhalos and the Dwarf Satellites of the Milky Way. , 679 (2):1260–1271, June 2008. doi: 10.1086/587545.

Mishra-Sharma, S. Inferring dark matter substructure with astrometric lensing beyond the power spectrum. *Machine Learning: Science and Technology*, 3(1):01LT03, March 2022. doi: 10.1088/2632-2153/ac494a.

Navarro, J. F., Frenk, C. S., and White, S. D. M. A Universal Density Profile from Hierarchical Clustering. , 490(2): 493–508, December 1997. doi: 10.1086/304888.

Oguri, M., Keeton, C. R., and Dalal, N. The impact of lens galaxy environments on the image separation distribution. , 364(4):1451–1458, December 2005. doi: 10.1111/j.1365-2966.2005.09697.x.

Ostdiek, B., Diaz Rivero, A., and Dvorkin, C. Extracting the Subhalo Mass Function from Strong Lens Images with Image Segmentation. *arXiv e-prints*, art. arXiv:2009.06639, September 2020. doi: 10.48550/arXiv.2009.06639.

Perreault Levasseur, L., Hezaveh, Y. D., and Wechsler, R. H. Uncertainties in Parameters Estimated with Neural Networks: Application to Strong Gravitational Lensing. , 850(1):L7, November 2017. doi: 10.3847/2041-8213/aa9704.

Planck Collaboration, Aghanim, N., Akrami, Y., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A. J., Barreiro, R. B., Bartolo, N., Basak, S., Battye, R., Benabed, K., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bond, J. R., Borrill, J., Bouchet, F. R., Boulanger, F., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J. F., Carron, J., Challinor, A., Chiang, H. C., Chluba, J., Colombo, L. P. L., Combet, C., Contreras, D., Crill, B. P., Cuttaia, F., de Bernardis, P., de Zotti, G., Delabrouille, J., Delouis, J. M., Di Valentino, E., Diego, J. M., Doré, O., Douspis, M., Ducout, A., Dupac, X., Dusini, S., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Fantaye, Y., Farhang, M., Fergusson, J., Fernandez-Cobos, R., Finelli, F., Forastieri, F., Frailis, M., Fraisse, A. A., Franceschi, E., Frolov, A., Galeotta, S., Galli, S., Ganga, K., Génova-Santos, R. T., Gerbino, M., Ghosh, T., González-Nuevo, J., Górski, K. M., Gratton, S., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Handley, W., Hansen, F. K., Herranz, D., Hildebrandt, S. R., Hivon, E., Huang, Z., Jaffe, A. H., Jones, W. C., Karakci, A., Keihänen, E., Keskitalo, R., Kiiveri, K., Kim, J., Kisner, T. S., Knox, L., Krachmalnicoff, N., Kunz, M., Kurki-Suonio, H., Lagache, G., Lamarre, J. M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Le Jeune, M., Lemos, P., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Lilley, M., Lindholm, V., López-Caniego, M., Lubin, P. M., Ma, Y. Z., Macías-Pérez, J. F., Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marcos-Caballero, A., Maris, M., Martin, P. G., Martinelli, M., Martínez-González, E., Matarrese, S., Mauri, N., McEwen, J. D., Meinhold, P. R., Melchiorri, A., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M. A., Molinari, D., Montier, L., Morgante, G., Moss, A., Natoli, P., Nørgaard-Nielsen, H. U., Pagano, L., Paoletti, D., Partridge, B., Patanchon, G., Peiris, H. V.,

Perrotta, F., Pettorino, V., Piacentini, F., Polastri, L., Polenta, G., Puget, J. L., Rachen, J. P., Reinecke, M., Remazeilles, M., Renzi, A., Rocha, G., Rosset, C., Roudier, G., Rubiño-Martín, J. A., Ruiz-Granados, B., Salvati, L., Sandri, M., Savelainen, M., Scott, D., Shellard, E. P. S., Sirignano, C., Sirri, G., Spencer, L. D., Sunyaev, R., Suur-Uski, A. S., Tauber, J. A., Tavagnacco, D., Tenti, M., Toffolatti, L., Tomasi, M., Trombetti, T., Valenziano, L., Valiviita, J., Van Tent, B., Vibert, L., Vielva, P., Villa, F., Vittorio, N., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Zacchei, A., and Zonca, A. Planck 2018 results. VI. Cosmological parameters. , 641:A6, September 2020. doi: 10.1051/0004-6361/201833910.

Serjeant, S. Up to 100,000 Reliable Strong Gravitational Lenses in Future Dark Energy Experiments. , 793(1):L10, September 2014. doi: 10.1088/2041-8205/793/1/L10.

Sérsic, J. L. Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Boletin de la Asociacion Argentina de Astronomia La Plata Argentina*, 6:41–43, February 1963.

Shu, Y., Bolton, A. S., Mao, S., Kochanek, C. S., Pérez-Fournon, I., Oguri, M., Montero-Dorta, A. D., Cornachione, M. A., Marques-Chaves, R., Zheng, Z., Brownstein, J. R., and Ménard, B. The BOSS Emission-line Lens Survey. IV. Smooth Lens Models for the BELLS GALLERY Sample. , 833(2):264, December 2016. doi: 10.3847/1538-4357/833/2/264.

Shu, Y., Brownstein, J. R., Bolton, A. S., Koopmans, L. V. E., Treu, T., Montero-Dorta, A. D., Auger, M. W., Czoske, O., Gavazzi, R., Marshall, P. J., and Moustakas, L. A. The Sloan Lens ACS Survey. XIII. Discovery of 40 New Galaxy-scale Strong Lenses. , 851(1):48, December 2017. doi: 10.3847/1538-4357/aa9794.

Sonnenfeld, A., Treu, T., Gavazzi, R., Suyu, S. H., Marshall, P. J., Auger, M. W., and Nipoti, C. The SL2S Galaxy-scale Lens Sample. IV. The Dependence of the Total Mass Density Profile of Early-type Galaxies on Redshift, Stellar Mass, and Size. , 777(2):98, November 2013. doi: 10.1088/0004-637X/777/2/98.

Springel, V., Wang, J., Vogelsberger, M., Ludlow, A., Jenkins, A., Helmi, A., Navarro, J. F., Frenk, C. S., and White, S. D. M. The Aquarius Project: the subhaloes of galactic haloes. , 391(4):1685–1711, December 2008. doi: 10.1111/j.1365-2966.2008.14066.x.

Stone, C. and Courteau, S. The Intrinsic Scatter of the Radial Acceleration Relation. , 882(1):6, September 2019. doi: 10.3847/1538-4357/ab3126.

Stone, C., Courteau, S., and Arora, N. The Intrinsic Scatter of Galaxy Scaling Relations. , 912(1):41, May 2021. doi: 10.3847/1538-4357/abebe4.

Stone, C., Adam, A., Coogan, A., Yantovski-Barth, M. J., Filipp, A., Setiawan, L., Core, C., Legin, R., Wilson, C., Missael Barco, G., Hezaveh, Y., and Perreault-Levasseur, L. Caustics: A Python Package for Accelerated Strong Gravitational Lensing Simulations. *arXiv e-prints*, art. arXiv:2406.15542, June 2024. doi: 10.48550/arXiv.2406.15542.

Talbot, M. S., Brownstein, J. R., Dawson, K. S., Kneib, J.-P., and Bautista, J. The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: a catalogue of strong galaxy-galaxy lens candidates. , 502(3):4617–4640, April 2021. doi: 10.1093/mnras/stab267.

Treu, T., Gavazzi, R., Gorecki, A., Marshall, P. J., Koopmans, L. V. E., Bolton, A. S., Moustakas, L. A., and Burles, S. The SLACS Survey. VIII. The Relation between Environment and Internal Structure of Early-Type Galaxies. , 690(1):670–682, January 2009. doi: 10.1088/0004-637X/690/1/670.

Vegetti, S., Koopmans, L. V. E., Bolton, A., Treu, T., and Gavazzi, R. Detection of a dark substructure through gravitational imaging. , 408(4):1969–1981, November 2010. doi: 10.1111/j.1365-2966.2010.16865.x.

Wagner-Carena, S., Aalbers, J., Birrer, S., Nadler, E. O., Darragh-Ford, E., Marshall, P. J., and Wechsler, R. H. From Images to Dark Matter: End-to-end Inference of Substructure from Hundreds of Strong Gravitational Lenses. , 942(2):75, January 2023. doi: 10.3847/1538-4357/aca525.

Wagner-Carena, S., Lee, J., Pennington, J., Aalbers, J., Birrer, S., and Wechsler, R. H. A Strong Gravitational Lens Is Worth a Thousand Dark Matter Halos: Inference on Small-Scale Structure Using Sequential Methods. *arXiv e-prints*, art. arXiv:2404.14487, April 2024. doi: 10.48550/arXiv.2404.14487.

Zhang, G., Mishra-Sharma, S., and Dvorkin, C. Inferring subhalo effective density slopes from strong lensing observations with neural likelihood-ratio estimation. , 517 (3):4317–4326, December 2022. doi: 10.1093/mnras/stac3014.

Zhang, G., Şengül, A. Ç., and Dvorkin, C. Subhalo effective density slope measurements from HST strong lensing data with neural likelihood-ratio estimation. , 527(2): 4183–4192, January 2024. doi: 10.1093/mnras/stad3521.