Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **527,** 4183–4192 (2024) https://doi.org/10.1093/mnras/stad3521
Advance Access publication 2023 November 15

Subhalo effective density slope measurements from *HST* strong lensing data with neural likelihood-ratio estimation

Gemma Zhang ,★ Atınç Çağan Şengül and Cora Dvorkin
Department of Physics, Harvard University, Cambridge, MA 02138, USA

Accepted 2023 November 13. Received 2023 November 3; in original form 2023 August 24

ABSTRACT

Examining the properties of subhaloes with strong gravitational lensing images can shed light on the nature of dark matter. From upcoming large-scale surveys, we expect to discover orders of magnitude more strong lens systems that can be used for subhalo studies. To optimally extract information from a large number of strong lensing images, machine learning provides promising avenues for efficient analysis that is unachievable with traditional analysis methods, but application of machine learning techniques to real observations is still limited. We build upon previous work, which uses a neural likelihood-ratio estimator, to constrain the effective density slopes of subhaloes and demonstrate the feasibility of this method on real strong lensing observations. To do this, we implement significant improvements to the forward simulation pipeline and undertake careful model evaluation using simulated images. Ultimately, we use our trained model to predict the effective subhalo density slope from combining a set of strong lensing images taken by the *Hubble Space Telescope*. We found the subhalo slope measurement of this set of observations to be steeper than the slope predictions of cold dark matter subhaloes. Our result adds to several previous works that also measured high subhalo slopes in observations. Although a possible explanation for this is that subhaloes with steeper slopes are easier to detect due to selection effects and thus contribute to statistical bias, our result nevertheless points to the need for careful analysis of more strong lensing observations from future surveys.

Key words: gravitational lensing: strong – dark matter.

1 INTRODUCTION

The standard Lambda cold dark matter (CDM) cosmological model has been in remarkable agreement with large-scale observations, but there is scarce evidence for the nature of dark matter (DM) on small (sub-galactic) scales. Because the nature of DM remains elusive, examining various DM models using small-scale cosmological observables becomes crucial. One of the promising observables used to study DM is subhaloes, which are small DM clumps gravitationally bound to a larger halo. Probing the properties of these subhaloes can potentially shine light on the nature of DM, as subhaloes exhibit different properties under alternate DM models beyond CDM. For instance, warm dark matter (WDM) models predict a smaller number of low-mass subhaloes and more cored subhalo density profiles compared to CDM (Bode, Ostriker & Turok 2001), while self-interacting dark matter (SIDM) models generally predict more cored subhalo profiles than that of the CDM model (Spergel & Steinhardt 2000).

Because low-mass subhaloes are observed to lack luminous matter (Fitts et al. 2017; Read et al. 2017; Kim, Peter & Hargis 2018), they are typically probed through their gravitational effects. Strong gravitational lensing, a predicted phenomenon from General Relativity, is a powerful way to constrain subhalo properties. In strong gravitational lensing, light emitted by a distant source gets deflected by the gravitational field of a massive structure (lens), and properties of the lens and its substructure can be inferred by analysing the images of the source light. In this paper, we will focus on studying subhaloes in the lens galaxy of strong lensing systems in which both the lens and background source are galaxies.

To date, there have been a few claimed detections of substructure in galaxy–galaxy strong lensing observations (Vegetti et al. 2010, 2012; Hezaveh et al. 2016b). Existing analyses that use observed strong lensing images to constrain DM models primarily rely on modelling individual (often the most massive) substructure in a lens system (Vegetti et al. 2014; Ritondale et al. 2019; Şengül et al. 2022). While useful, direct substructure modelling is computationally costly, and it is often limited to capturing the effect of relatively massive subhaloes. Even though the CDM paradigm predicts a large number of subhaloes with smaller masses, they are difficult to probe through traditional analysis methods because the inclusion of more subhaloes makes sampling of the joint parameter space prohibitive. As a result, it is important to explore alternative analysis methods that can more optimally incorporate information from the large population of smaller subhaloes.

To leverage the collective effect of subhalo populations on strong lensing images, there has been significant work done to obtain statistical constraints from subhaloes (Dalal & Kochanek 2002; Cyr-Racine et al. 2016; Hezaveh et al. 2016a; Birrer, Amara & Refregier 2017; Daylan et al. 2018; Díaz Rivero, Cyr-Racine & Dvorkin 2018a; Díaz Rivero et al. 2018b; Gilman et al. 2018; Brennan et al. 2019; Cyr-Racine, Keeton & Moustakas 2019; He et al. 2022). In

★ E-mail: yzhang7@g.harvard.edu

particular, machine learning has emerged as a promising candidate to analyse subhaloes in strong lensing images for its ability to efficiently and implicitly marginalize over a large parameter space. With the upcoming large-scale imaging surveys, the number of observed strong lensing systems is expected to increase significantly (Laureijs et al. 2011; Collett 2015; McKean et al. 2015; Bechtol et al. 2019; Jacobs et al. 2019; Huang et al. 2021; Storfer et al. 2022). Machine learning has a much-needed advantage that can make inference on these large data sets feasible.

Several deep learning techniques have been demonstrated to be effective at constraining the subhalo mass function using simulated strong lensing images (Brewer, Huijser & Lewis 2016; Brehmer et al. 2019; Anau Montel et al. 2023; Ostdiek, Diaz Rivero & Dvorkin 2022a, b; Wagner-Carena et al. 2023), but so far, there has been no successful attempt at applying them to real observations. The main challenge of applying deep learning methods to observations comes from the need for the training set to closely resemble the test set, as deep learning models are known to struggle in the presence of a distribution shift between training and test sets (Recht et al. 2018, 2019). Most of the previous works on machine learning applications to strong lensing made simplifying assumptions in the forward modelling pipeline of the training set in order to demonstrate the potential suitability of a method. However, for the machine learning model to be deployed on observations, its training set needs to incorporate all possible complexities that exist in the observed data.

In this work, for the first time, we analyse subhalo properties in real strong lensing observations with a machine learning technique. We build upon the method developed in Zhang, Mishra-Sharma & Dvorkin (2022) by adding multiple layers of complexities in the forward pipeline for the training set. Through training, our model learns to infer the effective subhalo density slope (directly related to the subhalo concentration), a promising observable proposed by Şengül & Dvorkin (2022) for distinguishing DM models. Several other works have also shown that the concentration of subhaloes is an effective probe of DM (Minor et al. 2021a, b; Amorisco et al. 2022). Using our trained model on real observed strong lensing images, we found a subhalo density slope steeper than those of subhaloes predicted by the CDM model. This measurement is consistent with previous works, which also found unexpectedly large subhalo concentrations (Minor et al. 2021b; Şengül & Dvorkin 2022).

This paper is organized as follows. In Section 2, we discuss details of the forward model used to generate mock strong lensing images. In Section 3, we summarize the deep learning technique that we use for inference, discuss our inference procedure, and outline our neural network architecture. In Section 4, we evaluate our trained model and compare the model predictions on the observed data with those under the CDM model. We conclude with a summary of our results in Section 5, and discuss the implications of our work.

2 DATA

We generate simulated lensing images to train our neural network and compare our model predictions with ground truths on mock images post-training to ensure training quality. At inference time, we apply the trained model to a set of observed lensing images from the *Hubble Space Telescope* (*HST*). We discuss details of both the mock data and the real *HST* observations below.

2.1 Mock data generation

To generate our mock strong lensing images, we use the software package LENSTRONOMY (Birrer, Amara & Refregier 2015; Birrer & Amara 2018). In order to match the *HST* post-drizzling image configuration, we generate (100×100) pixel^2 images, with a resolution of 0.04 arcsec per pixel. We build upon the simulation pipeline used in Zhang, Mishra-Sharma & Dvorkin (2022) and include significantly more complexities in the modelling process so as to make the images as similar to real observations as possible. Modelling a strong lens system requires several ingredients in the forward model: a source galaxy, a main (host) lens galaxy, a population of subhaloes and line-of-sight (LoS) haloes. In addition, we specify the instrumental configuration and image pre-processing of the mock images in the forward simulation. The distributions of parameters governing the lens models of our simulated images are summarized in Table 1.

2.1.1 Source and main lens

In a galaxy–galaxy lens system, light rays of a background source galaxy get gravitationally deflected by a foreground lens galaxy en route to the detector. Strong gravitational lensing specifically refers to the case where the projected surface mass density of the lens is greater than the critical surface density $\Sigma_{\rm crit}$. In this scenario, the bending of source light is significant enough to result in characteristic arcs of light in observed images.

To simulate the source light, we use galaxy images taken by the *HST* Cosmic Evolution Survey (COSMOS; Koekemoer et al. 2007; Scoville et al. 2007) processed by PALTAS (Wagner-Carena et al. 2023). The PALTAS package takes a subsample of the *HST* COSMOS survey galaxy images (Mandelbaum et al. 2012, 2014) and filters out suitable source candidates. To simulate the source for each mock image, we randomly draw a galaxy image from the COSMOS catalogue and randomly vary the rotation angle and the source coordinates $(x_{\rm source}, y_{\rm source})$. From the 2262 available source galaxies, we use 2163 (96 per cent) for the training set, 70 (3 per cent) for the validation set, and the remainder (1 per cent) for testing and evaluation.

We model the main lens mass distribution using an elliptical power law (EPL) profile (Barkana 1998). The convergence of an EPL profile at position (x, y) on the lens plane is given as follows:

$$\kappa(x, y) = \frac{3-\gamma}{2}\left(\frac{\theta_{\rm E}}{\sqrt{q x_\phi^2 + y_\phi^2/q}}\right)^{\gamma-1}, \tag{1}$$

where $\theta_{\rm E}$ is the Einstein radius, q is the minor/major axis ratio, x_ϕ, y_ϕ are positions on the axes aligned with the major and minor axes, and γ is the power-law slope of the mass distribution. To model each main lens, we draw its $\gamma_{\rm ML}$ (γ of the main lens) from $\mathcal{N}(2, 0.1)$ and truncate the tails of the normal distribution so that the range of possible values is bounded by 1.1 and 2.9. Slope values outside of the (1, 3) interval lead to non-physical or divergent mass profiles and are thus not included in our modelling. Adding variations in $\gamma_{\rm ML}$ simulates the natural stochasticity in lens density profiles that deviate from an isothermal profile ($\gamma = 2$). Note that ϕ indicates the angle between the major/minor axes and the fixed (x, y) axes of an image. The inputs into LENSTRONOMY are the ellipticity moduli, which are directly related to q and ϕ:

$$e_1 = \tfrac{1-q}{1+q}\cos(2\phi), \tag{2}$$

$$e_2 = \tfrac{1-q}{1+q}\sin(2\phi). \tag{3}$$

In addition, we add multipole moments $m = 3, 4$ to the EPL lens mass distribution in order to more realistically model the mass distribution of more complex lenses that may deviate from an elliptical profile.

Table 1. Parameters of the main components of a galaxy–galaxy strong gravitational lensing system and their respective training distributions in our forward simulation pipeline.

Parameter	Distribution
Source	
Source redshift	$z_{\rm source} \sim \mathcal{U}(0.5, 0.7)$
x-coordinate	$x_{\rm source} \sim \mathcal{U}(-0.1\,{\rm arcsec}, 0.1\,{\rm arcsec})$
y-coordinate	$y_{\rm source} \sim \mathcal{U}(-0.1\,{\rm arcsec}, 0.1\,{\rm arcsec})$
Main lens	
Lens redshift	$z_{\rm lens} \sim \mathcal{U}(0.15, 0.25)$
x-coordinate	$x_{\rm lens} \sim \mathcal{U}(-0.2\,{\rm arcsec}), 0.2\,{\rm arcsec})$
y-coordinate	$y_{\rm lens} \sim \mathcal{U}(-0.2\,{\rm arcsec}, 0.2\,{\rm arcsec})$
Einstein radius	$\theta_{\rm E} \sim \mathcal{U}(0.9\,{\rm arcsec}, 1.3\,{\rm arcsec})$
Ellipticities	$e_1 \sim \mathcal{U}(-0.2, 0.2) \quad e_2 \sim \mathcal{U}(-0.2, 0.2)$
Multipole moments ($m = 3, 4$)	$a_m \sim \mathcal{U}(0, 0.05) \quad \phi_m \sim \mathcal{U}(-\pi, \pi)$
EPL slope of density profile	$\gamma_{\rm ML} \sim \mathcal{N}(2, 0.1)$
External shear	$\gamma_{\rm shear,1} \sim \mathcal{U}(-0.1, 0.1) \quad \gamma_{\rm shear,2} \sim \mathcal{U}(-0.1, 0.1)$
Lens light	
Apparent magnitude	$m \sim \mathcal{U}(17, 19)$
Half-light radius	$R_{\rm sersic} \sim \mathcal{N}(0.8, 0.15)$
Sérsic index	$n_{\rm sersic} \sim \mathcal{N}(2, 0.5)$
Ellipticities	$e_1 \sim \mathcal{U}(-0.1, 0.1) \quad e_2 \sim \mathcal{U}(-0.1, 0.1)$
LoS haloes	
EPL ellipticities	$e_1 \sim \mathcal{U}(-0.2, 0.2) \quad e_2 \sim \mathcal{U}(-0.2, 0.2)$
EPL slope of density profile per lens system	$\gamma \sim \mathcal{U}(1.1, 2.9)$
EPL slope of density profile per subhalo	$\gamma_i \sim \mathcal{N}(\gamma, 0.1\gamma)$
LoS halo mass	$M_{200} \in [10^7, 10^{10}]\,{\rm M}_\odot$
Halo mass function normalization	$\delta_{\rm los} \sim \mathcal{U}(0, 2)$
Subhaloes	
EPL ellipticities	$e_1 \sim \mathcal{U}(-0.2, 0.2) \quad e_2 \sim \mathcal{U}(-0.2, 0.2)$
EPL slope of density profile per lens system	$\gamma \sim \mathcal{U}(1.1, 2.9)$
EPL slope of density profile per subhalo	$\gamma_i \sim \mathcal{N}(\gamma, 0.1\gamma)$
Subhalo mass function power-law slope	-1.9
Subhalo mass	$M_{200} \in [10^7, 10^{10}]\,{\rm M}_\odot$

We also include an external shear parametrized by $\gamma_{\rm shear,1}$ and $\gamma_{\rm shear,2}$ (Keeton, Kochanek & Seljak 1997). The shear parameters $\gamma_{\rm shear,1}$ and $\gamma_{\rm shear,2}$ are the diagonal and off-diagonal terms of the shear matrix, respectively.

In Zhang, Mishra-Sharma & Dvorkin (2022), it is assumed that the light produced by the lens galaxy has already been subtracted from the original observed image through a coarse modelling process. However, for real observed images, removing the lens light may involve imperfect modelling and high computational cost. To bypass this assumption, we include lens light in our mock image modelling. We assume that the centre of the lens light coincides with the centre of its mass density profile and that the lens light takes on an elliptical Sérsic profile (Sérsic 1963), with the brightness parametrized as:

$$I(r) = I_0 \exp\left(-b_{n_{\rm sersic}}\left(\frac{r}{r_{\rm sersic}}\right)^{\frac{1}{n_{\rm sersic}}}\right), \tag{4}$$

where $n_{\rm sersic}$ is the Sérsic index, $b_{n_{\rm sersic}} \approx 1.999 n_{\rm sersic} - 0.327$. Here, $r_{\rm sersic}$ is the half-light radius, and I_0 is determined by the apparent magnitude (Birrer & Amara 2018). We draw the apparent magnitude of the lens light from a uniform distribution between 17 and 19. We choose this range to be consistent with the apparent magnitude measurements of lens galaxies in the observed images used in our analysis (Auger et al. 2009), which are discussed in detail in Section 2.2. In each mock image, we vary all parameters governing the lens model, including its centre position, Einstein radius, shear parameters, apparent magnitude, and its ellipticity. The variation ranges of these parameters are summarized in Table 1.

In simulating our training set images, we take into account the spectroscopic redshifts of the real *HST* observations used during inference. To simulate each image in our training set, the source galaxy redshift is drawn from a uniform distribution of $z_{\rm source} \sim \mathcal{U}(0.5, 0.7)$, while the lens galaxy redshift is drawn from a uniform distribution of $z_{\rm lens} \sim \mathcal{U}(0.15, 0.25)$. These redshift ranges roughly match with those of the real observations that we use for inference. We deliberately chose to work with systems with relatively low-source redshifts because they align better with the redshifts of the COSMOS galaxies that are used in our modelling pipeline, minimizing the difference between our simulated images and the real observations.

2.1.2 *Subhaloes and line-of-sight haloes*

Aside from the main lens, the observed strong lensing images of the source light are affected by additional structures: subhaloes, which are small haloes residing inside the main host halo, and LoS haloes, which are located along the LoS between the source galaxy and the observer. If these (sub)haloes are found within the bright lensed arcs in the observed image, they can leave detectable perturbations on the observed images. Analysing these perturbations provides us with information about the properties of these substructures.

In our pipeline for simulating the training and validation set images, we add subhaloes and LoS haloes that follow the EPL profile given in equation (1). The γ parameter in the EPL profile controls the steepness of the halo density profiles: a larger γ implies a denser

halo density profile. We model our training and validation set images with EPL (sub)haloes because it allows us to label each image with its underlying power-law slope, which is the ultimate parameter of interest during our inference. We include a uniform prior on γ in our training set so that we do not unnecessarily bias our model. To model the subhaloes and LoS haloes in each image, we first draw γ from a uniform distribution: $\gamma \sim \mathcal{U}(1.1, 2.9)$; we then draw normally distributed slopes $\gamma_i \sim \mathcal{N}(\gamma, 0.1\gamma)$ for the ith subhalo. This normal distribution is truncated so that γ_i is constrained between 1.01 and 2.99. The number of subhaloes added to each image is drawn from a uniform distribution $N_{\rm sub} \sim \mathcal{U}\{0, 3000\}$. Note that the upper bound of 3000 subhaloes is an overestimate of a realistic number of subhaloes for our host halo mass, but we include a higher number of subhaloes so that our neural network can successfully learn the signatures in the lensed images corresponding to the changes in the density slope γ. During model evaluation, we will use a smaller $N_{\rm sub}$ range to simulate a more realistic substructure fraction, as will be discussed in Section 4.

Because only subhaloes near the bright arcs of an image have observable effects, in our simulated images, we place subhaloes solely in pixels whose brightness is more than a fifth of the maximum brightness in the smooth model image, which is the image modelled with only the lens and source galaxies and no substructure (and in this case no lens light). The Einstein radius of each subhalo is determined by its mass M_{200}, which is drawn from a subhalo mass function $\mathrm{d}N_{\rm sub}/\mathrm{d}M_{200} \propto M_{200}^{-1.9}$. The mass M_{200} is defined as the total mass enclosed by r_{200}, which is the radius within which the average mass density is 200 times the critical density of the Universe. In our simulated images, we only add subhaloes with masses between 10^7 and $10^{10}\,\mathrm{M}_\odot$, because subhaloes heavier than this range are scarce and can often be individually modelled.

To add the LoS haloes in our modelling, we use the pipeline provided by PALTAS, with several added modifications. The properties of each LoS halo are determined by the following parameters: its mass M_{200}, density slope γ_i, ellipticities e_1, e_2, redshift $z_{\rm los}$, and position coordinates $x_{\rm los}$, $y_{\rm los}$. PALTAS determines the mass M_{200} of each LoS halo using a modified Sheth–Tormen halo mass function, which includes two additional free parameters to the mass function proposed originally in Sheth, Mo & Tormen (2001): (1) an overall scaling factor that accounts for uncertainties in the normalization of the mass function; (2) a parameter that accounts for the contribution from the two-point halo correlation function, due to the fact that DM haloes are biased tracers of the overall matter distribution. The two-point halo correlation function correction is only added for haloes along the LoS that are sufficiently close to the main halo of the lens. For a more comprehensive discussion of the modified Sheth–Tormen mass function used in PALTAS, we refer readers to Wagner-Carena et al. (2023). The density slopes and ellipticities of LoS haloes are drawn from the same uniform distributions as subhaloes, as discussed above. To determine the position of LoS haloes, we divide the space between the observer and the source galaxy into thin slices of redshift with uniform thickness. The position coordinates, $(x_{\rm los}, y_{\rm los})$, are bounded by a double cone whose bases lie in the lens plane and whose apexes lie at the observer and the source. The position of each LoS halo is sampled uniformly in the volume of the double cone.

The addition of subhaloes causes an enlarged effective Einstein radius, so to restore the effective Einstein radius to its smooth model counterpart, we add a negative mass sheet in the lens plane for the subhaloes. In addition, to avoid making the region along our LoS overdense compared to the rest of the Universe, we add a negative mass sheet in each redshift slice for the LoS haloes. The negative mass sheet is a constant sheet of convergence such that the sum over all its pixels cancels out the total convergence added by the subhaloes or LoS haloes.

2.1.3 *Instrumentation details and data pre-processing*

To make the mock images as similar to real observations as possible, we incorporate *HST* instrumentation details in the production of our training set. We model our images using the *HST* Advanced Camera for Surveys Wide Field Channel (ACS/WFC) *F*814*W* filter configuration and apply an empirical point spread function, obtained from examining the exposure of point-like stars (Anderson & King 2000). We add noise using an exposure time of 2200 s, in approximate agreement with the noise level of the observed *HST* images discussed in Section 2.2.

Moreover, in real observations, there are often bright structures close to the strong lens system of interest that can potentially distract our analysis. During training, we apply a circular mask to cover the region outside of the lensed arcs, so that our model learns to not get confused by potential confounders. To mask out the edges of the images, we set the area outside of a circular mask to zero after an image has been whitened. We vary the radius of the circular mask based on the Einstein radius of the image.

2.2 *HST* observations

We demonstrated, in Zhang, Mishra-Sharma & Dvorkin (2022), that a neural likelihood-ratio estimator is capable of extracting subhalo population density slope information from simulated strong lensing images. In this work, we apply the same method to real strong lensing data taken by the *HST*. Specifically, we use strong lens systems identified by the Sloan Lens ACS (SLACS) survey (Bolton et al. 2008) and followed up by *HST* observations.

In the SLACS strong lens systems, the redshifts of the foreground galaxies range from 0.05 to 0.5, while the redshifts of the background galaxies range from 0.2 to 1.2. For our analysis, we choose observed images that share the same set of properties, and then simulate a matching training set. The shared properties include 0.04 arcsec pixel resolution, *F*814*W* camera band, and exposure time of approximately 2200 s. We also select lens systems with source redshifts, lens redshifts, and Einstein radii that fall in a reasonably narrow range to limit the span of the overall parameter space. From the *HST* observations, we made (100×100) pixel^2 cutouts in which the lens systems of interest are located roughly at the centre. Out of these cutouts, we then selected a subset of them that contain visible lensed arcs. The selected *HST* observations share the same instrumentation details with our training set so as to avoid having an unnecessary distribution shift between training and testing. Our selection process led to a subset of 13 images that we ultimately used for inference, as shown in Fig. 1.

3 MODEL AND INFERENCE

To infer the subhalo density slopes, we use a simulation-based inference (SBI) machine learning technique. SBI methods have gained increasing popularity in parameter inference problems in cosmology because of their ability to approximate intractable likelihoods due to complicated physical processes. In our application, we train a neural likelihood-ratio estimator as a parametrized classifier to learn the likelihood function (Cranmer, Pavez & Louppe 2015; Baldi et al. 2016; Hermans, Begy & Louppe 2019).

In this section, we will give a high-level summary of our model and inference method. For a more detailed description of the theory

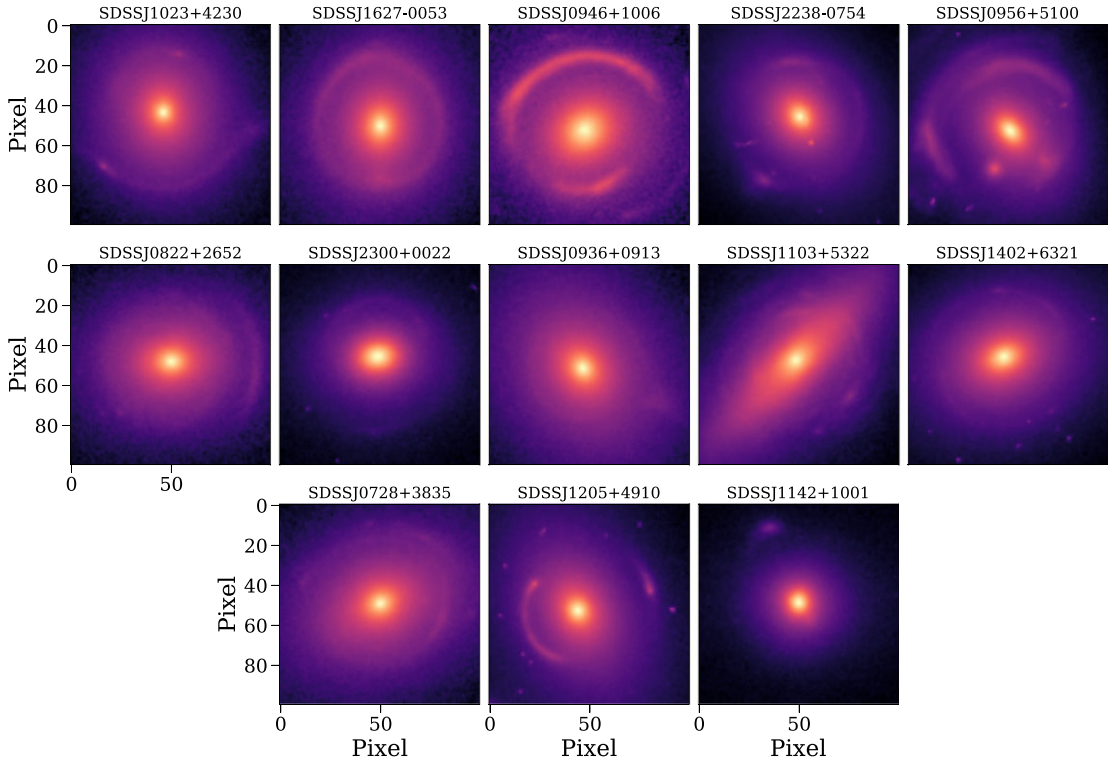

Figure 1. For our inference, we selected 13 image cutouts from observations that share similar instrumental configurations taken by the *HST F*814*W* filter. Each image cutout has 100 pixels of size 0.04 arcsec per side. The pixel values of the images are shown in logscale so that features of the lensed arcs are visible by eye. The title of each image corresponds to the name of the strong lens system.

underpinning the neural likelihood-ratio method, we refer readers to Cranmer, Pavez & Louppe (2015), while for more details on its application to analysing subhalo density slopes in strong lensing images, we refer readers to Zhang, Mishra-Sharma & Dvorkin (2022).

3.1 Inference method

Suppose θ denotes the parameters of our interest and x denotes the observed data. The core idea of likelihood-ratio estimation is training a classifier to distinguish between samples from two different probability distributions: the joint data–parameter distribution $p(x, \theta)$, which is the distribution of our interest, and the product of the marginal distributions of the data and the parameter $p(x)p(\theta)$. In our case, x corresponds to observed strong lensing images, while θ is the subhalo density slope γ underlying each image. We train a neural network as a classifier to learn the decision function $s(x, \theta) = p(x, \theta)/(p(x, \theta) + p(x)p(\theta))$, which is in one-to-one correspondence with the likelihood ratio $r(x|\theta) = p(x, \theta)/(p(x)p(\theta))$ as follows:

$$r(x \mid \theta) = \frac{s(x, \theta)}{1 - s(x, \theta)}. \tag{5}$$

This allows us to convert a likelihood inference task to a classification task (Cranmer, Pavez & Louppe 2015; Baldi et al. 2016; Mohamed & Lakshminarayanan 2016). At test time, to compute the likelihood-ratio profile as a function of γ for a given lensed image, we obtain the classifier logits for a linearly spaced array of input γ values. The likelihood-ratio estimation method is amortized: after spending an initial overhead for model training, minimal computational cost is needed during inference.

If we have an ensemble of strong lensing observations $\{x\}$ that are independently and identically distributed when conditioned on γ, then we can obtain their combined likelihood ratio by computing the product of the individual likelihood ratios,

$$\hat{r}(\{x\} \mid \gamma) = \prod_i \hat{r}(x_i \mid \gamma). \tag{6}$$

This offers a way for us to efficiently combine results of multiple observations with little additional computational cost.

3.2 Uncertainty quantification

If our likelihood-ratio estimator is a perfect classifier, then the test statistic $2\,(\ln \hat{r}_{\rm MLE} - \ln \hat{r})$ should be χ^2-distributed (Wilks 1938), where $\ln \hat{r}$ is the loglikelihood evaluated at the true γ and $\ln \hat{r}_{\rm MLE}$ is the loglikelihood at the maximum-likelihood estimate (MLE) $\gamma_{\rm MLE}$. However, we found that with our imperfect classifier, the test-statistic distribution deviates slightly from a true χ^2. Therefore, instead of quoting the 68 per cent uncertainty interval using a χ^2 distribution, we empirically determine the threshold for the 68 per cent confidence interval (CI) of the test statistic. In practical terms, we do this by computing the test statistic of many samples and then determining a threshold value under which approximately 68 per cent of the test statistic of these samples are included. Then, for a likelihood-ratio profile, the γ values whose likelihood ratios evaluate to this threshold determine the upper and lower uncertainties on the MLE. Because we found that combining different numbers of likelihood ratios leads to slightly different test-statistic distributions, this empirical threshold is determined separately for combining different numbers of images. This uncertainty quantification procedure ensures that approximately

68 per cent of the ground truths fall within the uncertainties quoted, and is used to determine the error bars presented in Section 4.

3.3 Model and training details

For our application, we use as our classifier a modified version of a common computer vision model, the ResNet-50 convolutional neural network implemented in PYTORCH (He et al. 2016; Paszke et al. 2019). We add a sigmoid projection after all of the dense layers in the ResNet in order to output the classification score $\hat{s}(x, \theta)$. At training time, we append the true γ for each input image to the latent vector after the convolutional layers in order to ensure that the model learns the true label. At test time, we instead append test γ values in order to obtain likelihood-ratio estimates over a range of γ. Our training objective is the canonical binary cross-entropy loss for classification. We provide more details of our customized ResNet-50 architecture in Appendix A.

To help with model convergence, we pre-process our training set images. We normalize image pixel values to having zero mean and unit standard deviation across the training set; in addition, we normalize the γ values to zero mean. To ensure consistency at test time, we use the training set mean and standard deviation to whiten our test data.

We use the AdamW optimizer (Kingma & Ba 2014; Loshchilov & Hutter 2017) with an initial learning rate of 10^{-3}. We follow a learning rate schedule that decays by an order of magnitude when the validation loss stagnates for three epochs, followed by a two-epoch cool-down period. We use a batch size of 1000 based on the maximum GPU memory available. There are 5000 000 mock images in our training set and 1000 in our validation set, all of which are generated using the forward model described in Section 2. Training terminates when the validation loss plateaus under a threshold of 10^{-3}. We carried out our neural network training on NVIDIA V100 GPUs for ∼20 epochs, with each epoch taking ∼5 h. We found that scaling up the size of the training set and the model complexity significantly improved the model performance during inference, and we expect there to be more improvement if the computing resource availability allows for more upscaling.

4 RESULTS

After our model has been trained, we first need to evaluate its convergence. To do this, we compare model predictions of the subhalo density slope with their ground truths using individual images in our validation set. In Fig. 2, we show the MLE (along with their 68 per cent uncertainties) compared to the true γ for 93 images with $1.2 < \gamma < 2.8$.

The validation images have parameters drawn from the same distributions as the training set, except for $N_{\rm sub}$, which is drawn from $\mathcal{U}\{0, 1800\}$ to simulate a more realistic substructure fraction. Note that because each image in our training and validation set is labelled with a true underlying γ value for EPL subhaloes, we ideally would like the neural network to predict the ground truths and be agnostic to the number of subhaloes. Therefore, having a more realistic number of subhaloes in our validation set serves as a way for us to ensure that changing the number of subhaloes does not incur a bias in our neural network predictions. The source galaxy images used in validation were held out in training. These images contain EPL subhaloes whose true underlying subhalo density slopes are known in the forward simulation pipeline, making this direct comparison possible. As shown in Fig. 2, our model predictions follow the ground truths in trend. This demonstrates that despite the addition of several layers of complexities into training images, our neural network remains sensitive to the signature imprinted on strong lensing images by changes in the subhalo density slope. However, the relatively large CIs indicate that the constraining power of individual images is limited, which makes it imperative to combine multiple images for inference. Note that the uncertainties are larger at the lower end of the γ range because smaller γ values indicate less concentrated subhaloes, which leave less detectable signatures in the lensed images.

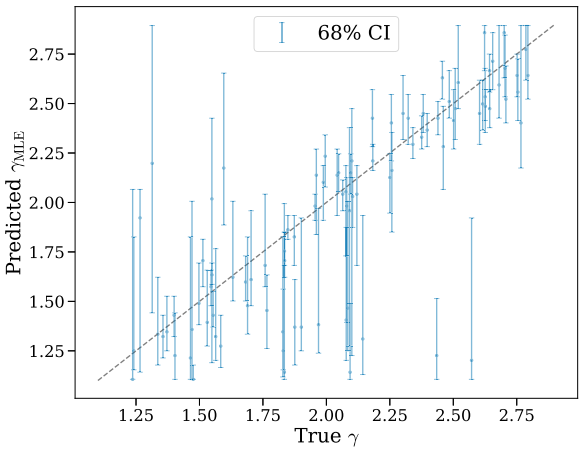

Figure 2. Scatter plot of the MLEs $\gamma_{\rm MLE}$ and their associated 68 per cent CIs (as discussed in Section 3.2) predicted using the trained likelihood-ratio estimator compared to the true underlying γ of 93 test images. The images contain EPL subhaloes with $M_{200} \in [10^7, 10^{10}]$ $\rm M_\odot$ and $N_{\rm sub} \sim \mathcal{U}\{0, 1800\}$. The model was trained on images containing EPL subhaloes with $M_{200} \in [10^7, 10^{10}]$ $\rm M_\odot$ and $N_{\rm sub} \sim \mathcal{U}\{0, 3000\}$.

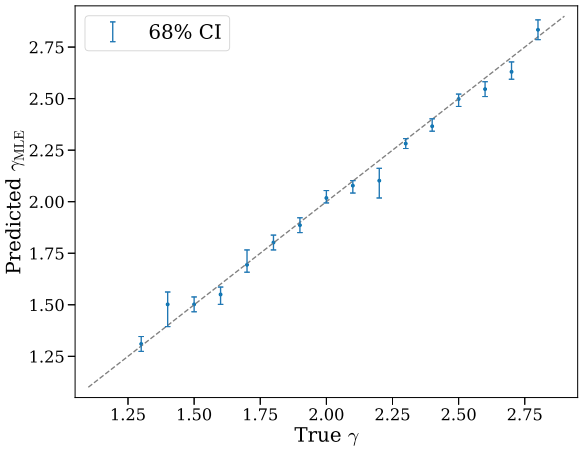

Figure 3. Maximum-likelihood estimated $\gamma_{\rm MLE}$ and associated 68 per cent CIs predicted from combined likelihoods of sets of 13 images containing EPL subhaloes compared to the true γ underlying each set of images (with $M_{200} \in [10^7, 10^{10}]$ $\rm M_\odot$ and $N_{\rm sub} \sim \mathcal{U}\{0, 1800\}$). The model was trained on images containing EPL subhaloes with $M_{200} \in [10^7, 10^{10}]$ $\rm M_\odot$ and $N_{\rm sub} \sim \mathcal{U}\{0, 3000\}$.

In addition, we check the model predictions of combining likelihood ratios of multiple images. In Fig. 3, we show $\gamma_{\rm MLE}$ compared to the ground truth γ for combined likelihoods of sets of 13 images, with each set sharing the same underlying slope γ. Note that we still simulate the natural spread in γ_i, so the slope for each subhalo varies

slightly. These images share the same parameter distributions as the images used in Fig. 2 except that the source galaxies come from the held-out set for validation. In Fig. 3, we see that the MLE predictions of our model closely follow the ground truths with relatively small error bars. Comparing the uncertainties between Figs 2 and 3, we see that combining images significantly improves constraining power.

4.1 Simulated images with NFW subhaloes

To obtain the expected subhalo density slopes under the CDM model, we simulate images containing subhaloes and LoS haloes following the Navarro–Frenk–White (NFW) profile (Navarro, Frenk & White 1997). Its radial density profile given by:

$$\rho(r) = \frac{\rho_0}{\frac{r}{r_s}\left(1+\frac{r}{r_s}\right)^2}, \tag{7}$$

where r is the distance from the centre of a subhalo, and the normalization ρ_0 and the scale radius r_s are free parameters. The NFW profile is the most common fit for the universal density profile of haloes from CDM N-body simulations. In addition to a normalization and a scale radius, the NFW profile can also be parametrized by the (sub)halo mass M_{200} and concentration c_{200}. The concentration c_{200} relates to the scale radius and virial radius r_{200} (as defined in Section 2.1.2) following $r_{200} = c_{200} r_s$. In our simulated images with NFW subhaloes, we relate M_{200} and c_{200} with a mass–concentration relation extrapolated from Dutton & Macciò (2014), which is an empirical relation determined using haloes in CDM simulations. We add a dex scatter of $\mathcal{N}(0, 0.1)$ to the mass–concentration relation for each subhalo in order to mimic the natural spread in the relationship. Note that a dex scatter of 0.1 corresponds to a ∼26 per cent variation in concentration. If we denote the CDM mass–concentration as $f_{\rm CDM}(M_{200})$, then we can modify the CDM mass–concentration relation by multiplying it by a constant factor (which will be referred to as concentration multiplicative factor) in order to simulate different density slopes of subhalo populations. In other words, this means that for a given concentration multiplicative factor a, we set the concentration of a subhalo with mass M_{200} to be $a \times f_{\rm CDM}(M_{200})$. To test the robustness of our neural network with as realistic images as possible, we add subhaloes everywhere in the image in these test sets.

In addition, due to tidal stripping from the host halo, subhaloes typically lose mass in their outer region (Hayashi et al. 2003; Diemand & Moore 2011). This means that, instead of an NFW profile, they can be more realistically modelled by a truncated NFW (tNFW) profile. The tNFW profile is parametrized by the NFW parameters as well as a truncation radius r_t:

$$\rho(r) = \frac{r_t^2 + r^2}{r_t^2} \frac{\rho_0}{r/r_s\,(1+r/r_s)^2}. \tag{8}$$

Because the truncation steepens the subhalo density profile past the truncation radius, we expect tNFW subhaloes to have steeper power-law density slopes than their NFW counterparts. To model the tNFW subhaloes in our pipeline, we first determine their NFW parameters following the procedure described for NFW subhaloes and subsequently set their truncation radii. The choice of truncation radii affects subhalo density profiles in the intermediate and outer region and thereby their measured slopes (Şengül & Dvorkin 2022). For our test images, we set the truncation radii following Wagner-Carena et al. (2023):

$$r_t = 1.4 \left(\frac{M_{200}}{m_{\rm trunc,pivot}}\right)^{\frac{1}{3}} \left(\frac{r_{\rm sub}}{r_{\rm trunc,pivot}}\right)^{\frac{2}{3}}, \tag{9}$$

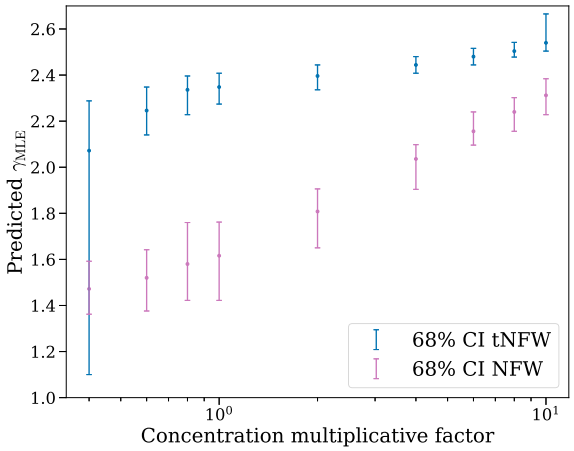

Figure 4. The median maximum-likelihood predictions (scatter points) and 68 per cent CIs (error bars) obtained by combining 13 images of $M_{200} \in [10^7, 10^{10}]\,{\rm M}_\odot$ NFW subhaloes or tNFW subhaloes as a function of the concentration multiplicative factor.

with $m_{\rm trunc,pivot} = 10^7 {\rm M}_\odot$ and $r_{\rm trunc,pivot} = 50$ kpc. Here, M_{200} is the subhalo mass and $r_{\rm sub}$ is the distance of the subhalo from the main halo centre. Note that in the test sets where subhaloes are modelled with tNFW, LoS haloes are still modelled with the NFW profile as they experience less tidal stripping than subhaloes.

One question that might arise is why our trained likelihood-ratio estimator can be applied to images with (t)NFW subhaloes and LoS haloes even though the training set only contains their EPL counterparts. The justification for this has been demonstrated in Şengül & Dvorkin (2022) and Zhang, Mishra-Sharma & Dvorkin (2022): given limited resolution and appropriate noise level, the observable changes in the surface brightness due to the presence of (t)NFW subhaloes can be well approximated by that of a power-law profile subhalo.

Because the density slopes of (t)NFW subhaloes vary with mass (*i.e.* larger masses have more extended density profiles and thereby smaller density slopes), the intrinsic stochasticity in the masses of a (t)NFW subhalo population introduces intrinsic aleatoric uncertainty into the slope measurement. To account for this uncertainty in each of our measurements, we generate 100 separate sets of images with shared properties and then obtain the MLE of each combined likelihood; using the set of 100 MLEs, we empirically determine the 68 per cent CIs.

In Fig. 4, we sample an array of varying concentration multiplicative factors and show our model MLE from the combined likelihood of 13 images that contain (t)NFW subhaloes and LoS haloes at each multiplicative factor. As expected, subhaloes with higher concentrations lead to larger γ predictions. In particular, the data points for a concentration multiplicative factor of 1 correspond to the expected subhalo density slope measurements under the CDM model, and they will be compared with the predicted slope of the *HST* observations in Section 4.2. From the figure, we find that tNFW subhaloes in general produce higher density slope measurements than NFW subhaloes, consistent with findings presented in Şengül & Dvorkin (2022).

4.2 Result with *HST* images

Having done model validation and obtained the expected density slope of CDM subhaloes, we will now use our model to infer

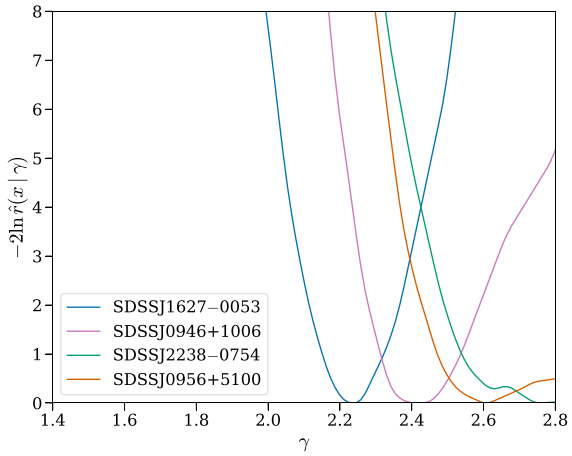

Figure 5. Representative examples of the individual likelihood-ratio test-statistic profiles for *HST* images. Each profile is labelled with the name of its corresponding lens system.

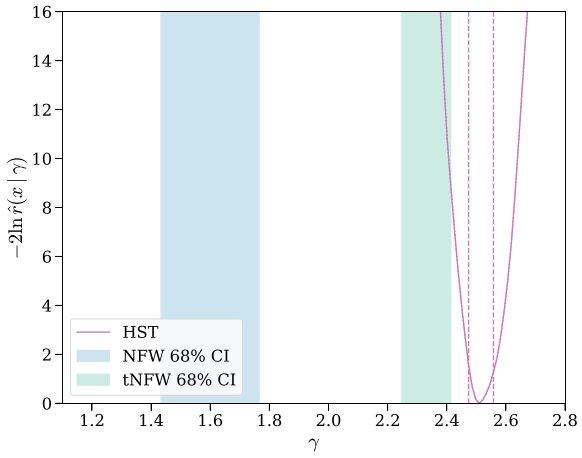

Figure 6. The 68 per cent CI of combined likelihood-ratio test-statistic profiles of 13 images containing $M_{200} \in [10^7, 10^{10}]$ $M_\odot$ NFW subhaloes and tNFW subhaloes, both with a concentration multiplicative factor of 1, as well as the combined likelihood-ratio test-statistic profile of the 13 *HST* images shown in Fig. 1. The uncertainties corresponding to the 68 per cent CI are shown in dashed lines for the likelihood-ratio test-statistic profile.

the subhalo slope of observed *HST* strong lensing images, which are described in Section 2.2. These images are masked at the edges and whitened with the training mean and standard deviation before being fed into our neural network. In Fig. 5, we show the individual likelihood-ratio test-statistic profiles for several of the *HST* observations. In comparison with the predictions of NFW and tNFW subhaloes under the CDM model, as discussed in Section 4.1, we see that the predicted slopes of these *HST* observations are larger than the predicted slopes of the CDM model.

We subsequently combine these individual likelihood ratios using equation (6) in order to obtain tighter constraints. In Fig. 6, we show the combined likelihood-ratio test-statistic profile for all of the 13 images. From this profile, we get a measurement of the subhalo density slope of $\gamma_{\rm MLE} = 2.51^{-0.04}_{+0.05}$, with the quoted uncertainties indicating the 68 per cent credible interval shown in dotted lines.

In the same figure, we also show the 68 per cent CIs for the combined likelihood-ratio test-statistic profiles of 13 images containing NFW subhaloes or tNFW subhaloes. These correspond to data points shown in Fig. 4 for a concentration multiplicative factor of 1. Comparing the slope prediction of the *HST* images with that of the simulated images with NFW subhaloes, we see that the measured density slope of the *HST* data is significantly steeper than the expected slope under the assumption that CDM subhaloes follow an NFW profile. The predicted slope of the images containing tNFW subhaloes is also less than the *HST* measurement, but their difference is less statistically significant than that with the NFW prediction. While surprising, this is in agreement with previous works that also measured a higher than expected concentration (Minor et al. 2021b; Şengül & Dvorkin 2022). In particular, our 13 *HST* images include the SDSSJ0946+1006 system analysed by Minor et al. (2021b), which measured a much higher concentration than the CDM prediction. The individual likelihood-ratio test-statistic profile for the same system in our analysis is shown in Fig. 5, and it is in broad agreement with the result in their work. It is also worth noting that our method provides a stronger constraint due to the neural network's ability to efficiently combine multiple observations. It would also be useful to compare our results with those obtained by Şengül & Dvorkin (2022) of the JVAS B1938+666 lens system, but to our knowledge, there is no suitable *HST* observation of this lens system that matches our training set configuration. Thus, we leave this for future work when more observations become available.

One possible explanation of the difference between our result and the CDM predictions lies in the assumptions made in our subhalo modelling. Several assumptions about subhalo density profiles went into modelling the lens system in the image; in particular, the density profile parametrization and the choice of mass–concentration relation affect the predicted slope measurements of subhaloes under the CDM model. Modelling these properties for subhaloes is an ongoing area of research (Green, van den Bosch & Jiang 2021), and an improved understanding of subhalo profiles may change the predicted CDM density slopes. Another possible reconciliation is accounting for the selection effects. Subhaloes with steeper density slopes are more concentrated and, therefore, are easier to detect in observations. Within our current resolution constraint and noise level, the less concentrated smaller subhaloes are not detectable, hence biasing our statistics. This effect of the selection function on slope measurements is important, and we leave a careful study of it for future work, when more observations become available from ongoing and upcoming surveys.

5 CONCLUSIONS AND OUTLOOK

Observations at sub-galactic scale are essential for examining alternate DM models and contrasting them against the standard CDM model. Among the small-scale observables, subhaloes provide a promising avenue for DM studies. In addition to constraining the subhalo mass function, studying the subhalo density slope (concentration) can help to potentially differentiate various classes of DM models. Subhalo properties can be probed by analysing strong gravitational lensing images. Traditional strong lensing image analyses model individual subhaloes through a forward modelling pipeline, but this process can only provide limited statistics; to model more subhaloes in a system or to combine statistics from many images, direct lens modelling becomes computationally infeasible.

The rapid progress in machine learning enables the development of techniques that have the power to leverage the collective effect of subhalo populations in strong lensing images, as well as to efficiently

analyse a large ensemble of observations. Despite showcases of success on simulated images, many of these machine learning methods require further validation and improvements before they can be successfully applied to real strong lensing observations.

In this work, we built upon the likelihood-ratio estimation method developed in Zhang, Mishra-Sharma & Dvorkin (2022) and trained a neural network capable of making inference from observed strong lensing images. To make the leap from mock to real images, we added numerous layers of realism in the forward pipeline of the training set. This includes complexifying the lens model to account for the lens light, multipole moments as well as external shear, incorporating realistic noise levels, and adding LoS haloes. We demonstrated that the likelihood-ratio estimator retains its sensitivity to changes in the subhalo density slope in simulated strong lensing, even after adding these layers of realism. Furthermore, we obtained the expected subhalo density slope measurements in simulations under the CDM model. This measurement comes from using our trained neural network to predict the slope of simulated lensing images containing (t)NFW subhaloes that follow a mass–concentration relation derived from CDM simulations. Finally, we measured the subhalo slopes of a set of 13 *HST* observations and statistically combined their constraints. By comparing the subhalo slope in the *HST* observations with the measurement from simulated CDM images, we found an unexpectedly high slope measurement in the *HST* observations, in tension with CDM predictions.

Several recent works in cluster lensing have also suggested that substructures in galaxy clusters are more compact than expected of the CDM model (Meneghetti et al. 2020, 2022, 2023). Combined with several similar results in the literature, our measurement has important implication for DM studies as it may motivate more careful examination of alternate DM models. The most common alternatives to CDM, the WDM model, and many SIDM models, predict a lower than CDM subhalo density slope and would exacerbate the tension that we observe (Lovell et al. 2012, 2014; Vogelsberger, Zavala & Loeb 2012; Rocha et al. 2013; Kahlhoefer et al. 2019). However, certain SIDM models [*e.g.* with large self-interacting cross-sections (Nishikawa, Boddy & Kaplinghat 2020)] also predict that SIDM subhaloes can undergo core collapses that result in unusually concentrated inner profiles in a time-scale relevant for observations today (Lynden-Bell & Wood 1968; Kochanek & White 2000; Colín et al. 2002; Elbert et al. 2015; Nadler, Yang & Yu 2023). This gravitothermal core collapse due to DM self-interactions has been suggested as a possible explanation of these high-density central regions in cluster galaxies (Yang & Yu 2021). Resolving galactic subhaloes in simulations is harder due to their lower masses. A hybrid approach in Zeng et al. (2022), which includes a combination of semi-analytical methods and *N*-body simulations has shown that some SIDM models can produce subhaloes with collapsed cores at subgalactic mass scales ($< 10^{10}\,\mathrm{M}_{\odot}$). This phenomenon provides a possible explanation for the high subhalo density slope that we measured. Based on our work, it is still not possible to pinpoint the mechanism that causes this outlier measurement from the CDM model, but there are several directions of future work that can take us closer to answering this question. For instance, one can study the subhalo slope predictions under different microphysical DM models and compare them with the predictions from observed lensing images. In addition, one can examine the effect of assumptions about CDM subhalo properties on the likelihood-ratio estimator's slope predictions. As more lensing systems are expected to be discovered with upcoming surveys (and followed up by observations), the likelihood-ratio estimator will be a valuable tool for obtaining more measurements to help elucidate the nature of DM.

ACKNOWLEDGEMENTS

We thank Simon Birrer and Siddharth Mishra-Sharma for helpful discussions. This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions; http://iaifi.org/). The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University. Some of the data presented in this paper were obtained from the Mikulski Archive for Space Telescopes (MAST). STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555.

DATA AVAILABILITY

The 13 *HST* images analysed in this work can be downloaded from https://mast.stsci.edu/portal/Mashup/Clients/Mast/Portal.html. The code used to produce the results shown in this paper is available at https://github.com/gemyxzhang/neural-subhalo-slope-data.

REFERENCES

Amorisco N. C. et al., 2022, MNRAS, 510, 2464
Anau Montel N., Coogan A., Correa C., Karchev K., Weniger C., 2023, MNRAS, 518, 2746
Anderson J., King I. R., 2000, PASP, 112, 1360
Auger M. W., Treu T., Bolton A. S., Gavazzi R., Koopmans L. V. E., Marshall P. J., Bundy K., Moustakas L. A., 2009, ApJ, 705, 1099
Baldi P., Cranmer K., Faucett T., Sadowski P., Whiteson D., 2016, Eur. Phys. J. C, 76, 235
Barkana R., 1998, ApJ, 502, 531
Bechtol K. et al., 2019, BAAS, 51, 207
Birrer S., Amara A., 2018, Phys. Dark Universe, 22, 189
Birrer S., Amara A., Refregier A., 2015, ApJ, 813, 102
Birrer S., Amara A., Refregier A., 2017, J. Cosmol. Astropart. Phys., 2017, 037
Bode P., Ostriker J. P., Turok N., 2001, ApJ, 556, 93
Bolton A. S., Burles S., Koopmans L. V. E., Treu T., Gavazzi R., Moustakas L. A., Wayth R., Schlegel D. J., 2008, ApJ, 682, 964
Brehmer J., Mishra-Sharma S., Hermans J., Louppe G., Cranmer K., 2019, ApJ, 886, 49
Brennan S., Benson A. J., Cyr-Racine F.-Y., Keeton C. R., Moustakas L. A., Pullen A. R., 2019, MNRAS, 488, 5085
Brewer B. J., Huijser D., Lewis G. F., 2016, MNRAS, 455, 1819
Colín P., Avila-Reese V., Valenzuela O., Firmani C., 2002, ApJ, 581, 777
Collett T. E., 2015, ApJ, 811, 20
Cranmer K., Pavez J., Louppe G., 2015, preprint (arXiv:1506.02169)
Cyr-Racine F.-Y., Moustakas L. A., Keeton C. R., Sigurdson K., Gilman D. A., 2016, Phys. Rev. D, 94, 043505
Cyr-Racine F.-Y., Keeton C. R., Moustakas L. A., 2019, Phys. Rev. D, 100, 023013
Dalal N., Kochanek C. S., 2002, ApJ, 572, 25
Daylan T., Cyr-Racine F.-Y., Diaz Rivero A., Dvorkin C., Finkbeiner D. P., 2018, ApJ, 854, 141
Díaz Rivero A., Cyr-Racine F.-Y., Dvorkin C., 2018a, Phys. Rev. D, 97, 023001
Díaz Rivero A., Dvorkin C., Cyr-Racine F.-Y., Zavala J., Vogelsberger M., 2018b, Phys. Rev. D, 98, 103517
Diemand J., Moore B., 2011, Adv. Sci. Lett., 4, 297
Dutton A. A., Macciò A. V., 2014, MNRAS, 441, 3359
Elbert O. D., Bullock J. S., Garrison-Kimmel S., Rocha M., Oñorbe J., Peter A. H. G., 2015, MNRAS, 453, 29
Fitts A. et al., 2017, MNRAS, 471, 3547
Gilman D., Birrer S., Treu T., Keeton C. R., Nierenberg A., 2018, MNRAS, 481, 819

Green S. B., van den Bosch F. C., Jiang F., 2021, MNRAS, 503, 4075
Hayashi E., Navarro J. F., Taylor J. E., Stadel J., Quinn T., 2003, ApJ, 584, 541
He K., Zhang X., Ren S., Sun J., 2016, IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, NV, USA, p. 770
He Q. et al., 2022, MNRAS, 511, 3046
Hermans J., Begy V., Louppe G., 2019, preprint (arXiv:1903.04057)
Hezaveh Y., Dalal N., Holder G., Kisner T., Kuhlen M., Perreault Levasseur L., 2016a, J. Cosmol. Astropart. Phys., 2016, 048
Hezaveh Y. D. et al., 2016b, ApJ, 823, 37
Huang X. et al., 2021, ApJ, 909, 27
Jacobs C. et al., 2019, ApJS, 243, 17
Kahlhoefer F., Kaplinghat M., Slatyer T. R., Wu C.-L., 2019, J. Cosmol. Astropart. Phys., 2019, 010
Keeton C. R., Kochanek C. S., Seljak U., 1997, ApJ, 482, 604
Kim S. Y., Peter A. H. G., Hargis J. R., 2018, Phys. Rev. Lett., 121, 211302
Kingma D. P., Ba J., 2014, preprint (arXiv:1412.6980)
Kochanek C. S., White M., 2000, ApJ, 543, 514
Koekemoer A. M. et al., 2007, ApJS, 172, 196
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Loshchilov I., Hutter F., 2017, preprint (arXiv:1711.05101)
Lovell M. R. et al., 2012, MNRAS, 420, 2318
Lovell M. R., Frenk C. S., Eke V. R., Jenkins A., Gao L., Theuns T., 2014, MNRAS, 439, 300
Lynden-Bell D., Wood R., 1968, MNRAS, 138, 495
McKean J. et al., 2015, Proc. Sci., Strong Gravitational Lensing with the SKA. Sissa, Trieste, PoS(AASKA14)084
Mandelbaum R., Hirata C. M., Leauthaud A., Massey R. J., Rhodes J., 2012, MNRAS, 420, 1518
Mandelbaum R. et al., 2014, ApJS, 212, 5
Meneghetti M. et al., 2020, Science, 369, 1347
Meneghetti M. et al., 2022, A&A, 668, A188
Meneghetti M. et al., 2023, A&A, 678, L2
Minor Q., Kaplinghat M., Chan T. H., Simon E., 2021a, MNRAS, 507, 1202
Minor Q., Gad-Nasr S., Kaplinghat M., Vegetti S., 2021b, MNRAS, 507, 1662
Mohamed S., Lakshminarayanan B., 2016, preprint (arXiv:1610.03483)
Nadler E. O., Yang D., Yu H.-B., 2023, preprint (arXiv:2306.01830)
Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
Nishikawa H., Boddy K. K., Kaplinghat M., 2020, Phys. Rev. D, 101, 063009
Ostdiek B., Diaz Rivero A., Dvorkin C., 2022a, A&A, 657, L14
Ostdiek B., Diaz Rivero A., Dvorkin C., 2022b, ApJ, 927, 83
Paszke A. et al., 2019, in Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., eds, Advances in Neural Information Processing Systems 32. p. 8024
Read J. I., Iorio G., Agertz O., Fraternali F., 2017, MNRAS, 467, 2019
Recht B., Roelofs R., Schmidt L., Shankar V., 2018, preprint (arXiv:1806.00451)
Recht B., Roelofs R., Schmidt L., Shankar V., 2019, preprint (arXiv:1902.10811)
Ritondale E., Vegetti S., Despali G., Auger M. W., Koopmans L. V. E., McKean J. P., 2019, MNRAS, 485, 2179
Rocha M., Peter A. H. G., Bullock J. S., Kaplinghat M., Garrison-Kimmel S., Oñorbe J., Moustakas L. A., 2013, MNRAS, 430, 81
Scoville N. et al., 2007, ApJS, 172, 1
Sérsic J. L., 1963, Bol. Asociacion Argentina de Astron. La Plata Argentina, 6, 41
Sheth R. K., Mo H. J., Tormen G., 2001, MNRAS, 323, 1
Spergel D. N., Steinhardt P. J., 2000, Phys. Rev. Lett., 84, 3760
Storfer C. et al., 2022, preprint (arXiv:2206.02764)
Şengül A. Ç., Dvorkin C., 2022, MNRAS, 516, 336
Şengül A. Ç., Dvorkin C., Ostdiek B., Tsang A., 2022, MNRAS, 515, 4391
Vegetti S., Koopmans L. V. E., Bolton A., Treu T., Gavazzi R., 2010, MNRAS, 408, 1969
Vegetti S., Lagattuta D. J., McKean J. P., Auger M. W., Fassnacht C. D., Koopmans L. V. E., 2012, Nature, 481, 341
Vegetti S., Koopmans L. V. E., Auger M. W., Treu T., Bolton A. S., 2014, MNRAS, 442, 2017
Vogelsberger M., Zavala J., Loeb A., 2012, MNRAS, 423, 3740
Wagner-Carena S., Aalbers J., Birrer S., Nadler E. O., Darragh-Ford E., Marshall P. J., Wechsler R. H., 2023, ApJ, 942, 75
Wilks S. S., 1938, Ann. Math. Stat., 9, 60
Yang D., Yu H.-B., 2021, Phys. Rev. D, 104, 103031
Zeng Z. C., Peter A. H. G., Du X., Benson A., Kim S., Jiang F., Cyr-Racine F.-Y., Vogelsberger M., 2022, MNRAS, 513, 4845
Zhang G., Mishra-Sharma S., Dvorkin C., 2022, MNRAS, 517, 4317

APPENDIX: MODEL ARCHITECTURE

We describe in this appendix the customized ResNet-50 architecture used in this work. The original ResNet-50 model used in computer vision consists of a series of convolution blocks followed by pooling and dense layers. We made two modifications to this model for our inference task. First, we append the truth label γ of each image during training to the flattened latent space vector after the convolution blocks, as indicated by the top arrow in Fig. A1. This ensures that the neural network incorporates information about γ into its prediction. In addition, we add a logistic activation function after the last layer of ResNet-50 to ensure that the final output is a valid classification score $\hat{s}(\gamma, x)$ (*i.e.* between 0 and 1). As discussed in Section 3.3, when we train the neural network as a classifier, the value given by the ResNet before the logistic activation gives us the loglikelihood estimate $\ln \hat{r}$, as indicated in Fig. A1.

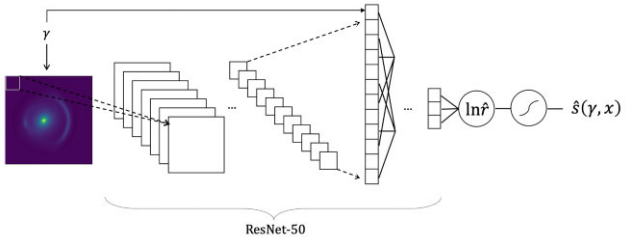

Figure A1. Graphical illustration of the neural network architecture used in this work.

This paper has been typeset from a TEX/LATEX file prepared by the author.