The Astrophysical Journal, 942:75 (22pp), 2023 January 10 https://doi.org/10.3847/1538-4357/aca525

© 2023. The Author(s). Published by the American Astronomical Society.
OPEN ACCESS

From Images to Dark Matter: End-to-end Inference of Substructure from Hundreds of Strong Gravitational Lenses

Sebastian Wagner-Carena[1,2], Jelle Aalbers[1,2], Simon Birrer[1,2], Ethan O. Nadler[3,4], Elise Darragh-Ford[1,2], Philip J. Marshall[1,2], and Risa H. Wechsler[1,2]

[1] Kavli Institute for Particle Astrophysics and Cosmology, Department of Physics, Stanford University, Stanford, CA 94305, USA; swagnerc@stanford.edu
[2] SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA
[3] Carnegie Observatories, 813 Santa Barbara Street, Pasadena, CA 91101, USA
[4] Department of Physics & Astronomy, University of Southern California, Los Angeles, CA 90007, USA

Received 2022 March 17; revised 2022 October 19; accepted 2022 November 21; published 2023 January 13

Abstract

Constraining the distribution of small-scale structure in our universe allows us to probe alternatives to the cold dark matter paradigm. Strong gravitational lensing offers a unique window into small dark matter halos ($<10^{10}\,M_\odot$) because these halos impart a gravitational lensing signal even if they do not host luminous galaxies. We create large data sets of strong lensing images with realistic low-mass halos, Hubble Space Telescope (HST) observational effects, and galaxy light from HST's COSMOS field. Using a simulation-based inference pipeline, we train a neural posterior estimator of the subhalo mass function (SHMF) and place constraints on populations of lenses generated using a separate set of galaxy sources. We find that by combining our network with a hierarchical inference framework, we can both reliably infer the SHMF across a variety of configurations and scale efficiently to populations with hundreds of lenses. By conducting precise inference on large and complex simulated data sets, our method lays a foundation for extracting dark matter constraints from the next generation of wide-field optical imaging surveys.

Unified Astronomy Thesaurus concepts: Strong gravitational lensing (1643); Cosmology (343); Dark matter (353); Convolutional neural networks (1938); Hierarchical models (1925); Dark matter distribution (356)

1. Introduction

The concordance model in cosmology, ΛCDM, includes the presence of cold, collisionless dark matter (CDM). One of the predictions of the CDM model is the presence of approximately self-similar dark matter halos ranging in mass from fractions of a solar mass to $10^{15}\,M_\odot$ (Navarro 1996; Navarro et al. 1997; Green et al. 2004; Wang et al. 2020). The formation of these structures is hierarchical, with massive halos forming from mergers and accretion of less massive halos (White & Rees 1978; Moore et al. 1999). The CDM model further predicts that the number density of halos is inversely related to their mass, producing an abundance of low-mass ($<10^{10}\,M_\odot$) halos, both as gravitationally isolated structures and as "subhalos" of larger "host" halos. Many popular alternatives to CDM impact the predicted distribution, abundances, and profiles of low-mass halos (Bode et al. 2001; Kaplinghat 2005; Bullock & Boylan-Kolchin 2017; Buckley & Peter 2018; Tulin & Yu 2018). Therefore, constraining dark matter halos at these small scales provides an important test of the CDM paradigm.

Over the past decade, a number of observational probes have placed constraints on low-mass dark matter halos. In the Local Group, measurements of Milky Way satellites have been used to rule out or limit the parameter space of alternative dark matter models (Macciò & Fontanot 2010; Kennedy et al. 2014; Nadler et al. 2019, 2021a; Dekker et al. 2022; Newton et al. 2021), and early measurements of perturbations to the Milky Way's stellar streams are providing complementary constraints (Bonaca et al. 2019; Banik et al. 2021a, 2021b). At higher redshifts, Lyα forest measurements have been able to constrain dark matter models that impact small-scale structure formation in the early universe (Viel et al. 2013; Iršič et al. 2017a, 2017b; Rogers & Peiris 2021), as have measurements of the UV galaxy luminosity function (Menci et al. 2016; Rudakovskyi et al. 2021). With the exception of stellar stream perturbations, all of these probes depend on the emission or absorption of light by baryons in the halos; connecting luminous tracers to their underlying dark matter requires accurately modeling the baryonic physics and therefore introduces large uncertainties. For Milky Way satellites, the most sensitive probe to date, modeling the galaxy–halo connection at low halo masses ($<10^{9}\,M_\odot$), is one of the dominant observational uncertainties (Nadler et al. 2021a). In order to further constrain small-scale structure, we require tracers that are comparable in sensitivity to existing probes but less dependent on accurate modeling of the baryonic physics.

Strong gravitational lensing is a promising low-mass halo probe, as it directly measures the gravitational signal of dark matter structure. In the case of galaxy–galaxy lenses, light from a distant source galaxy passes by a massive "main deflector" galaxy and is refocused to produce multiple images. Smaller halos along the line of sight and subhalos within the main deflector also deflect the light and can cause detectable perturbations in the lensing image. These low-mass halos[5] can generate a signal even if they have no luminous counterparts (see Nadler et al. 2020 for recent upper bounds on the mass of these "dark halos" in a CDM context). The number of strong lenses available for analysis is also poised to grow rapidly, with over 1000 known lenses to date (Sonnenfeld et al. 2013) and tens of thousands expected to be discovered with next-generation wide-field optical imaging surveys (Collett 2015). Given modeling tools capable of extracting

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

[5] Throughout, we refer to line-of-sight halos and subhalos collectively as "low-mass halos," and we use "subhalos" when specifically considering low-mass halos in the main lens.

the low-mass halo signal, this sizable data set provides an opportunity for state-of-the-art dark matter sensitivity.

Strong lensing studies often focus on the population of subhalos within the main lens, specifically the abundance of subhalos per unit mass, referred to as the subhalo mass function (SHMF). Constraining the SHMF at masses below $<10^{10}\,M_\odot$ allows us to measure deviations from the CDM predictions for low-mass halos. Broadly speaking, there are two frameworks through which we can model the SHMF in strong lensing. The first is to attempt to detect the signal of individual subhalos in a strong gravitational lens, often called direct detection. Traditionally, this method first models a smooth main deflector and the source light. Subhalos are then added to the model, and if the improvement in the fit is sufficient—often measured through the Bayesian information criterion—the subhalos are considered detected (Mao & Schneider 1998; Moustakas & Metcalf 2003; Koopmans 2005; Vegetti & Koopmans 2009; Hezaveh et al. 2013a). Direct detection modeling has been used to identify subhalos in three systems and place constraints on the fraction of dark matter in subhalos (Vegetti et al. 2010, 2012; Hezaveh et al. 2016a; Vegetti et al. 2018; Çağan Şengül et al. 2022). Direct detection returns a concrete picture of the exact quantity, position, and mass of the subhalos detected. However, scaling this framework is challenging; modeling only the most massive subhalos is computationally demanding and yet cannot capture the thousands of halos in the 10^7–$10^9\,M_\odot$ mass range. While these lower-mass subhalos cannot be individually detected, they can collectively produce an observable signal. Additionally, with only one subhalo detection per lens, constraining the SHMF requires substantial assumptions about the subhalo and line-of-sight halo populations.

The second approach, often called statistical detection, directly models the properties of the subhalo population. In statistical detection, the main deflector and source may still be explicitly modeled, but no attempt is made to detect individual subhalos. Instead, this framework measures the signal created by "all" of the subhalos and connects it to the population statistics (i.e., the SHMF; Dalal & Kochanek 2002; Cyr-Racine et al. 2016; Hezaveh et al. 2016b; Birrer et al. 2017a; Rivero et al. 2018a, 2018b; Brennan et al. 2019). This approach reduces the number of free parameters required to describe the full mass range from thousands to dozens and can be used to directly infer the parameters that describe alternate dark matter models. However, the same SHMF can admit both subhalo configurations that are excellent and poor descriptors of the data. Additionally, because the full mass range includes thousands of subhalos, rigorously sampling the allowed configuration space of even a single SHMF is a sizable computing task. The net effect is that directly evaluating the likelihood is intractable, making traditional Markov Chain Monte Carlo inference unworkable.[6]

Instead, statistical detection work has focused on simulation-based inference, a family of methods that circumvent the need for a tractable likelihood by leveraging access to a simulator. The most well known of these methods is approximate Bayesian computation (ABC; Rubin 1984; Beaumont et al. 2002). In ABC, simulated data are generated by sampling from a prior on the parameter space. If the summary statistics of the simulated data approximately match the observed data, the sample is kept. With a sufficiently strict matching criteria, ABC returns a faithful sampling of the likelihood. In the strong lensing context, ABC has been used to constrain warm dark matter models, the halo mass–concentration relation, and the primordial power spectrum (Birrer et al. 2017b; Gilman et al. 2020a, 2020b; Nadler et al. 2021b; Gilman et al. 2022). However, ABC has its own drawbacks as an inference methodology. First, for image data sets, ABC requires reducing the images to low-dimensional summary statistics that discard information and therefore limits the constraining power of the data. Second, ABC inference does not scale to large data sets; each observed lens requires an independent inference chain with hundreds of thousands of personalized simulations.

In light of the drawbacks to ABC, modern simulation-based inference has shifted toward using neural networks as density estimators. In this approach, a network is trained on a single training set to predict either the posterior (Lueckmann et al. 2017), the likelihood (Papamakarios et al. 2019), or a likelihood ratio (Mohamed & Lakshminarayanan 2016) given an input data point. Unlike ABC, the density estimator can take advantage of the full information content of the data. The density estimator is also scalable; after the initial training, the cost of conducting inference on a new lens is negligible. In strong lensing, Brehmer et al. (2019) demonstrated that a likelihood ratio estimator can accurately extract the signal of subhalos on a population of 100 simulated lenses, and recent work has also applied simulation-based inference techniques to the direct detection framework (Ostdiek et al. 2022a, 2022b; Coogan et al. 2020; Lin et al. 2020; Diaz Rivero & Dvorkin 2020). Outside of the subhalo context, several studies have shown the ability of neural density estimators to constrain the parameters of the main deflector (Perreault Levasseur et al. 2017; Pearson et al. 2021; Wagner-Carena et al. 2021) and infer the Hubble constant from strong lensing time-delay measurements (Park et al. 2021).

The recent literature suggests that applying neural density estimators to strong lensing images will allow us to place tight constraints on alternatives to CDM. Despite this, neural density estimators of subhalo parameters have never been applied to observed strong lensing images. The principal limitation is the simulations. The existing proof-of-concept studies make a number of simplifications; these include smooth source models, ignoring line-of-sight halos, and simplistic detector responses. Using networks trained on simplified simulations is likely to produce biased inference, and it may be that adding these complexities will wash out the constraining power of the networks. With an eye toward pushing neural density estimator techniques toward the data, we set out to answer the following questions: given simulations of strong lensing images generated using realistic assumptions for the sources, subhalos, line-of-sight halos, and Hubble Space Telescope (HST) observational effects, can simulation-based inference with neural networks constrain the subhalo population on individual lenses? Specifically, can the network constrain the normalization of the SHMF? And then, can a network trained to make predictions on individual lenses be used to hierarchically infer the SHMF of a population of lenses?

In this paper, we address these questions by using realistic, complex simulations of galaxy–galaxy strong lenses to train and test a neural posterior estimator of the SHMF normalization. We pull our sources from observed galaxies, marginalize over well-motivated theoretical uncertainties when sampling our low-mass halos, and include the effects of the

[6] Probabilistic cataloging methods conduct transdimensional Bayesian inference to overcome this limitation but remain computationally limited to the most massive $\mathcal{O}(10)$ low-mass halos (Brewer et al. 2016; Daylan et al. 2018).

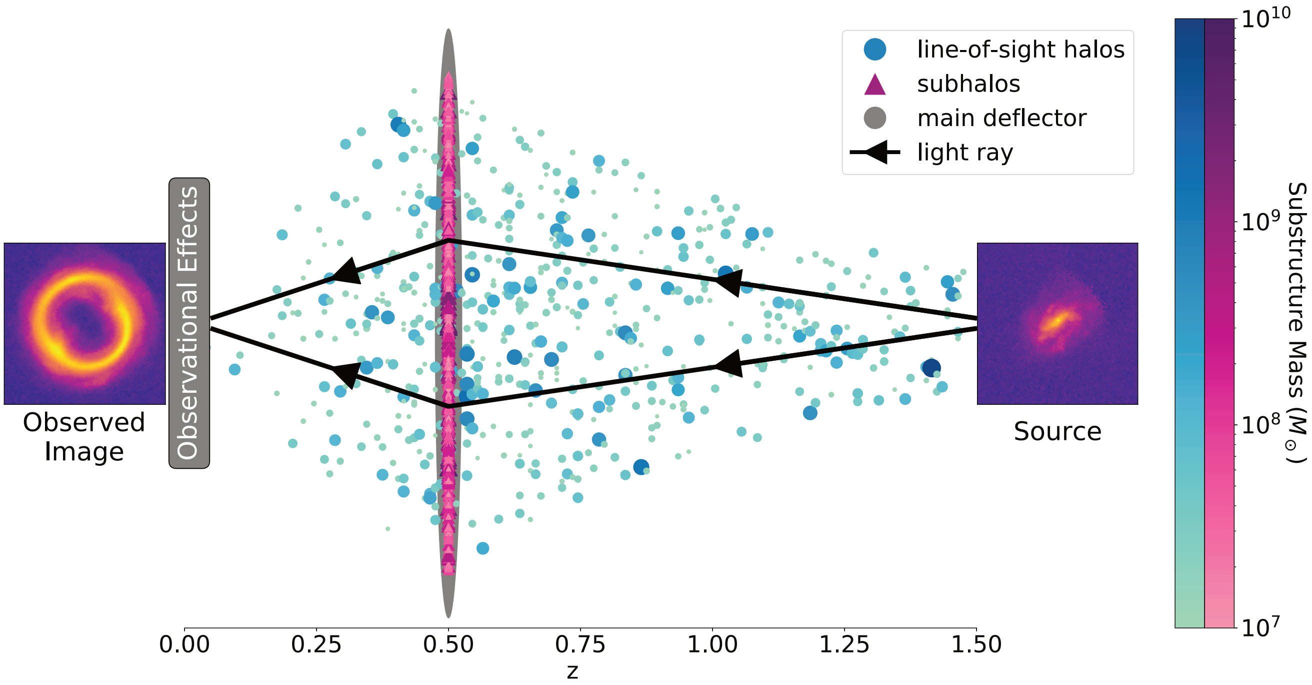

Figure 1. Schematic summary of the ingredients of our strong gravitational lensing simulation. The light rays leave the source (Section 2.4) and are perturbed by the line-of-sight halos (Section 2.3), bent by the main deflector (Section 2.1), and further perturbed by the subhalos (Section 2.2) and remaining line-of-sight halos to finally be measured by our detector. On top of the ray tracing, there are also observational effects from the detector (Section 2.5) that further distort the image. The final observed image is shown on the far left.

HST imaging pipeline. We test our neural posterior estimator's ability to accurately extract the SHMF signal on individual simulated lenses, including when the underlying sources are not seen during training. For this study, we confine our analysis to the ΛCDM model with cosmological parameters fixed to the best-fit values from Planck Collaboration (2020). We combine our network with a hierarchical inference pipeline to reconstruct the SHMF normalization for test sets that are statistically distinct from our training set. We find that our simulation-based inference approach is capable of accurately and efficiently inferring the statistic of a lensing population with hundreds of observations. Our method is sensitive to SHMF normalizations spanning over an order of magnitude and can account for systematic offsets between the distribution of training and test populations.

This paper is organized as follows. In Section 2, we discuss the simulated strong lensing images used in this work, with a particular focus on the choices surrounding the low-mass halos, the sources, and the HST observational effects. We then introduce our simulated data sets, neural posterior estimator, and hierarchical inference scheme in Section 3. We demonstrate the ability of the trained model to predict the normalization of the SHMF on individual lenses in Section 4.1. In Section 4.2, we run our model on 20 sets of lensing images, each with a self-similar SHMF, and explore the constraints we can extract with HST-quality photometry. We discuss the limitations of our analysis and potential future extensions in Section 5. Finally, in Section 6, we summarize our results and discuss the implications of our work for the future of dark matter constraints with strong lensing.

As part of this publication, we are releasing our strong lensing simulation package PALTAS.[7] The package builds on the simulation code LENSTRONOMY[8] (Birrer & Amara 2018; Birrer et al. 2021) to produce large and realistic strong lensing image data sets. The code has been designed to be modular, scalable, and easy to configure. Our hope is that it will provide an effective tool for the future development of simulation-based inference in strong lensing. The PALTAS repository includes all of the code and dependencies necessary to reproduce the results in this paper along with a set of comprehensive JUPYTER notebooks that help familiarize users with the code.

2. Simulation Methods

Extracting unbiased information with simulation-based inference requires representative simulations of the data. In the following subsections, we describe the parameterizations we use to simulate realistic strong lensing images that incorporate our existing theoretical uncertainties. We break our simulation choices down into five components: the main deflector (Section 2.1), the subhalos of the main deflector (Section 2.2), the line-of-sight halos (Section 2.3), the source (Section 2.4), and the observational and numerical parameters for the simulation (Section 2.5). A schematic representation of the model can be found in Figure 1. While the distribution of parameter values we use varies between the training, validation, and test sets, the parameterizations themselves remain consistent throughout this work. Table 1 lists all parameters of our model, along with their distribution on each image set.

For many of the cosmology calculations in our simulations, we use the COLOSSUS[9] package (Diemer 2018). For

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

[8] https://github.com/sibirrer/lenstronomy

[9] https://bitbucket.org/bdiemer/colossus

Table 1
Distribution of Simulation Parameters in the Training, Validation, and Test Sets

Component	Training/Validation Distribution	$\Sigma_{\rm sub}$ Test Distributions 1, … 20
Main Deflector—Section 2.1		
x-coordinate lens center (arcsec)	$x_{\rm lens} \sim \mathcal{N}(\mu\text{: } 0, \sigma\text{: } 0.16)$	$x_{\rm lens} \sim \mathcal{N}(\mu\text{: Unif } (-0.08,0.08),$[b] $\sigma\text{: } 0.016)$
y-coordinate lens center (arcsec)	$y_{\rm lens} \sim \mathcal{N}(\mu\text{: } 0, \sigma\text{: } 0.16)$	$y_{\rm lens} \sim \mathcal{N}(\mu\text{: Unif } (-0.08,0.08),$[b] $\sigma\text{: } 0.016)$
Einstein radius (arcsec)	$\theta_{\rm E} \sim \mathcal{N}\ (\mu\text{: } 1.1, \sigma\text{: } 0.15)$[a]	$\theta_{\rm E} \sim \mathcal{N}\ (\mu\text{: Unif}(1.025,1.175),$[b] $\sigma\text{: } 0.015)$[a]
Power-law slope	$\gamma_{\rm lens} \sim \mathcal{N}\ (\mu\text{: } 2.0, \sigma\text{: } 0.1)$[a]	$\gamma_{\rm lens} \sim \mathcal{N}\ (\mu\text{: Unif}(1.95,2.05),$[b] $\sigma\text{: } 0.01)$[a]
x-direction ellipticity eccentricity	$e_1 \sim \mathcal{N}(\mu\text{: } 0, \sigma\text{: } 0.1)$	$e_1 \sim \mathcal{N}\ (\mu\text{: Unif}(-0.05, 0.05),$[b] $\sigma\text{: } 0.01)$
xy-direction ellipticity eccentricity	$e_2 \sim \mathcal{N}(\mu\text{: } 0, \sigma\text{: } 0.1)$	$e_2 \sim \mathcal{N}\ (\mu\text{: Unif}(-0.05, 0.05),$[b] $\sigma\text{: } 0.01)$
Main halo critical mass ($M_\odot$)	$m_{\rm host} = 10^{13}$	$m_{\rm host} = 10^{13}$
Main halo redshift	$z_{\rm lens} = 0.5$	$z_{\rm lens} = 0.5$
x-direction shear	$\gamma_1 \sim \mathcal{N}(\mu\text{: } 0, \sigma\text{: } 0.05)$	$\gamma_1 \sim \mathcal{N}\ (\mu\text{: Unif}(-0.025, 0.025),$[b] $\sigma\text{: } 0.005)$
xy-direction shear	$\gamma_2 \sim \mathcal{N}(\mu\text{: } 0, \sigma\text{: } 0.05)$	$\gamma_2 \sim \mathcal{N}\ (\mu\text{: Unif}(-0.025, 0.025),$[b] $\sigma\text{: } 0.005)$
Mass–Concentration—Appendix A		
Concentration normalization	$c_0 = \text{Unif}(16, 18)$	$c_0 = \text{Unif}(16, 18)$
Redshift power-law slope	$\zeta = \text{Unif}(-0.3, -0.2)$	$\zeta = \text{Unif}(-0.3, -0.2)$
Peak height power-law slope	$\beta = \text{Unif}(0.55, 0.85)$	$\beta = \text{Unif}(0.55, 0.85)$
Mass–concentration power-law pivot mass ($M_\odot$)	$m_{\rm pivot,conc} = 10^8$	$m_{\rm pivot,conc} = 10^8$
Concentration dex scatter	$\sigma_{\rm conc} = \text{Unif}(0.1, 0.16)$	$\sigma_{\rm conc} = \text{Unif}(0.1, 0.16)$
Cosmology		
Cosmology assumption	ΛCDM from Planck 2018	ΛCDM from Planck 2018
Subhalos—Section 2.2		
SHMF power-law index	$\gamma_{\rm sub} \sim \text{Unif}(-1.92, -1.82)$	$\gamma_{\rm sub} \sim \text{Unif}(-1.92, -1.82)$
SHMF normalization (kpc^{-2})	$\Sigma_{\rm sub} \sim \mathcal{N}(\mu\text{: } 2 \times 10^{-3}, \sigma\text{: } 1.1 \times 10^{-3})$	$\Sigma_{\rm sub} \sim \mathcal{N}(\mu\text{: } 2 \times 10^{-4} \times i, \sigma\text{: } 1.5 \times 10^{-4})$[c]
Subhalo power-law pivot mass ($M_\odot$)	$m_{\rm pivot,sub} = 10^{10}$	$m_{\rm pivot,sub} = 10^{10}$
SHMF minimum mass ($M_\odot$)	$m_{\rm min,sub} = 10^7$	$m_{\rm min,sub} = 10^7$
SHMF maximum mass ($M_\odot$)	$m_{\rm max,sub} = 10^{10}$	$m_{\rm max,sub} = 10^{10}$
Subhalo truncation pivot mass ($M_\odot$)	$m_{\rm pivot,trunc} = 10^7$	$m_{\rm pivot,trunc} = 10^7$
Subhalo truncation pivot radius (kpc)	$r_{\rm pivot,trunc} = 50$	$r_{\rm pivot,trunc} = 50$
LOS Halos—Section 2.3		
LOS mass function normalization	$\delta_{\rm los} \sim \mathcal{N}(\mu\text{: } 1.0, \sigma\text{: } 0.6)$	$\delta_{\rm los} \sim \mathcal{N}(\mu\text{: } 1.0, \sigma\text{: } 0.6)$
Mass function minimum mass ($M_\odot$)	$m_{\rm min,los} = 10^7$	$m_{\rm min,los} = 10^7$
Mass function maximum mass ($M_\odot$)	$m_{\rm max,los} = 10^{10}$	$m_{\rm max,los} = 10^{10}$
Minimum LOS redshift	$z_{\rm min,los} = 0.01$	$z_{\rm min,los} = 0.01$
LOS redshift bin width	$\Delta_{z,\rm los} = 0.01$	$\Delta_{z,\rm los} = 0.01$
LOS cone opening angle (arcsec)	$\theta_{\rm los} = 8.0$	$\theta_{\rm los} = 8.0$
Minimum two-halo term radius (kpc)	$r_{\rm 2halo,min} = 0.5$	$r_{\rm 2halo,min} = 0.5$
Maximum two-halo term radius (kpc)	$r_{\rm 2halo,max} = 10.0$	$r_{\rm 2halo,max} = 10.0$
Deflection angle correction redshift bin width	$\Delta_{z,\rm correction} = 0.05$	$\Delta_{z,\rm correction} = 0.05$
Source: COSMOS Catalog—Section 2.4		
Source redshift	$z_{\rm source} = 1.5$	$z_{\rm source} = 1.5$
Maximum catalog redshift	$z_{\rm catalog,max} = 1.0$	$z_{\rm catalog,max} = 1.0$
Faintest catalog apparent magnitude	$\text{mag}_{\rm faint} = 20$	$\text{mag}_{\rm faint} = 20$
Minimum source size (pixels)	$\text{size}_{\rm min,pix} = 50$	$\text{size}_{\rm min,pix} = 50$
Minimum half-light radius (pixels)	$r_{1/2} = 10$	$r_{1/2} = 10$
Source rotation angle	$\phi_{\rm source} \sim \text{Unif}(0, 2\pi)$	$\phi_{\rm source} \sim \text{Unif}(0, 2\pi)$
x-coordinate source center (arcsec)	$x_{\rm source} \sim \mathcal{N}(\mu\text{: } 0, \sigma\text{: } 0.16)$	$x_{\rm source} \sim \mathcal{N}(\mu\text{: } 0, \sigma\text{: } 0.16)$
y-coordinate source center (arcsec)	$y_{\rm source} \sim \mathcal{N}(\mu\text{: } 0, \sigma\text{: } 0.16)$	$y_{\rm source} \sim \mathcal{N}(\mu\text{: } 0, \sigma\text{: } 0.16)$
Number of galaxy images	Training: 2163/validation: 99	99

Notes. For a detailed discussion of each parameter, see Section 2. The subsets of galaxy images used for the training and test sets are disjoint from one another. In this table, $\mathcal{N}$ is the normal distribution, and Unif is the uniform distribution. The SHMF normalization is highlighted, since it is our main parameter of interest throughout this work. For the test set distributions, any values in gray indicate that they are identical to the choices made on the training set.

[a] The distribution is capped at values larger than zero.

[b] For these parameters, each test set has a mean drawn from the uniform distribution specified.

[c] For test set i, the mean of $\Sigma_{\rm sub}$ is set to $2 \times 10^{-4} \times i$, so for test set 4, the mean is 8×10^{-4}.

the lensing calculations, we use the lens modeling package LENSTRONOMY.

2.1. Main Deflector

In our simulations, the main deflector is a power-law elliptical mass distribution (PEMD) profile with external shear. The PEMD (Kormann et al. 1994; Barkana 1998) profile is described by the convergence:

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

Here $\gamma_{\rm lens}$ is the logarithmic slope, $\theta_{\rm E}$ is the Einstein radius, and $q_{\rm lens}$ is the axis ratio of the lens. The profile described by Equation (1) assumes that the coordinate system for x and y is defined along the major and minor axes of the deflector. Therefore, there are three remaining parameters to fully describe the profile: the main deflector center position ($x_{\rm lens}$, $y_{\rm lens}$) and the main deflector rotation angle $\phi_{\rm lens}$. The additional external shear component is described by an orientation angle $\phi_{\rm ext}$ and modulus $\gamma_{\rm ext}$ (Keeton et al. 1997).

The angles $\phi_{\rm lens}$ and $\phi_{\rm ext}$ are cyclic parameters, which would complicate inference. We therefore work in the eccentricity/Cartesian coordinates for our ellipticity/shear:

$$e_1 = \frac{1 - q_{\rm lens}}{1 + q_{\rm lens}}\cos(2\phi_{\rm lens}), \quad (2)$$

$$e_2 = \frac{1 - q_{\rm lens}}{1 + q_{\rm lens}}\sin(2\phi_{\rm lens}), \quad (3)$$

$$\gamma_1 = \gamma_{\rm ext}\cos(2\phi_{\rm ext}), \quad (4)$$

$$\gamma_2 = \gamma_{\rm ext}\sin(2\phi_{\rm ext}). \quad (5)$$

Finally, the main deflector has redshift $z_{\rm lens}$ and host mass $m_{\rm host}$. The host mass uses the M_{200c} definition (White 2001).[10] There is no exact mapping from the host mass to the Einstein radius; therefore, the two values are left uncorrelated in our simulations.

2.2. Subhalos

The subhalos of the main deflector in our simulations follow the parameterization introduced by Gilman et al. (2020a) with some slight modifications. We draw our subhalos from the following mass function:

$$\frac{d^2N_{\rm sub}}{dA\ dm_{\rm sub}} = \Sigma_{\rm sub}\frac{m_{\rm sub}^{\gamma_{\rm sub}}}{m_{\rm pivot,sub}^{\gamma_{\rm sub}+1}}, \quad (6)$$

where $\Sigma_{\rm sub}$ is the normalization of the SHMF, $m_{\rm sub}$ is the subhalo mass using the M_{200c} definition (White 2001; see footnote 10), dA is the differential area element, and $m_{\rm pivot,sub}$ is the pivot mass. We render subhalos within the mass range [$m_{\rm min,sub}$, $m_{\rm max,sub}$].[11]

The SHMF as written contains no explicit dependence on host properties. Any scaling by, for example, the host mass or redshift has been absorbed into our definition of $\Sigma_{\rm sub}$. We expect our network to be sensitive to the projected number of subhalos in the main deflector, which for a fixed slope $\gamma_{\rm sub}$ is best captured by $\Sigma_{\rm sub}$. For a realistic/observed population of lenses, we therefore expect our framework to return a distribution of $\Sigma_{\rm sub}$ values that must be interpreted in the context of a model with host-dependent scaling.

The subhalos themselves are modeled as a truncated Navarro–Frenk–White (NFW) radial density profile (Baltz et al. 2009). The profile can be defined in terms of a mass, $m_{\rm sub}$; a concentration, $c_{\rm sub}$; and a truncation radius r_t. A detailed discussion of how these are drawn in our simulation can be found in Appendix A. For the positions of the subhalos, we follow Gilman et al. (2020a). Specifically, outside of the host's scale radius, $r_{s,\rm host}$, the subhalos follow the host's mass profile; within $r_{s,\rm host}$, the subhalos are uniformly distributed. To keep the simulations numerically tractable without altering the signal, we render subhalos within a projected radius of $3\theta_{\rm E}$, where $\theta_{\rm E}$ is the Einstein radius of the main deflector. The z-coordinates of the subhalos are also constrained to be within the interval [$-R_{200c}$, R_{200c}]. Here R_{200c} is the smallest radius such that the host halo's enclosed mass has a mean density of 200 times $\rho_{\rm crit}(z)$, the critical density of the universe at redshift z. Outside of this radius, potential halos are considered line-of-sight halos and accounted for in the two-point halo correlation (see Section 2.3).

2.3. Line-of-sight Halos

Historically, several studies of galaxy–galaxy strong lenses have ignored the contributions from line-of-sight halos (Vegetti et al. 2010, 2012). However, Despali et al. (2018) and Şengül et al. (2020) showed that, for certain lensing configurations, line-of-sight halos can produce a signal on par with the subhalos of the main deflector. In fact, Despali et al. (2018) and Çağan Şengül et al. (2022) demonstrated that one of the two existing subhalo detections can be better explained by a line-of-sight perturber. Therefore, while we do not explore the ability of strong lensing to constrain line-of-sight halos in this work, we do model the line-of-sight halos in our simulations self-consistently within the same CDM framework. This allows us to marginalize over their uncertainties and calculate their impact on the SHMF signal, as has been done by previous work (Gilman et al. 2020a, 2020b).[12]

As with the subhalos, the line-of-sight halos in our simulations closely follow the parameterization introduced in Gilman et al. (2020a). We draw our line-of-sight halos in discrete redshift bins ranging from a minimum redshift, $z_{\rm min,los}$, to the redshift of the source, $z_{\rm source}$. The width of the redshift bins is set by the parameter $\Delta_{z,\rm los}$. Within each redshift bin, we draw our line-of-sight halos from a modified Sheth–Tormen halo mass function (Sheth et al. 2001):

$$\frac{d^2N_{\rm los}}{dV\ dm_{\rm los}} = \delta_{\rm los}(1 + \xi_{2\ \rm halo}(r, m_{\rm host}, z_{\rm host})) \times \left[\frac{d^2N_{\rm los}}{dV\ dm_{\rm los}}\right]_{\rm ST}. \quad (7)$$

Here $\delta_{\rm los}$ is a scaling parameter that accounts for uncertainties in the overall normalization of the line-of-sight mass function, $\xi_{2\ \rm halo}(r, m_{\rm host}, z_{\rm host})$ is a contribution from the two-point halo

[10] Our definition of M_{200c} uses the critical density at the redshift of the subhalo, not the critical density at redshift zero.

[11] We assume that more massive halos would host sufficient baryons to be visible in our images and therefore modeled individually. The lower limit is set below the sensitivity of our inference (see Appendix F).

[12] It is also possible to use our simulations to marginalize over the subhalo contribution to constrain the line-of-sight halo population, although we do not do that here.

correlation function, and $\left[\frac{d^2N_{\rm los}}{dV\ dm_{\rm los}}\right]_{\rm ST}$ is the traditional Sheth–Tormen halo mass function. Each of these components is described in further detail in Appendix B. We render our line-of-sight halos in the mass range $[m_{\rm min,los}, m_{\rm max,los}]$ (see footnote 11).

The line-of-sight halos themselves are parameterized by an NFW profile (Navarro et al. 1997) with the radial density profile

$$\rho_{\rm NFW}(r) = \frac{\rho_{\rm los}}{r/r_{s,\rm los}(1 + r/r_{s,\rm los})^2}. \tag{8}$$

Here $\rho_{\rm los}$ is the amplitude of the NFW density profile in units of $M_\odot$ kpc^{-3}, r is the radial position in units of kpc, and $r_{s,\rm los}$ is the scale radius in units of kpc. As with the subhalo profile, the scale radius, $r_{s,\rm los}$, and amplitude of the density profile, $\rho_{\rm los}$, are calculated from the mass, $m_{\rm los}$, and concentration, $c_{\rm los}$, of the line-of-sight halo. The mass–concentration relation being used is fully detailed in Appendix A.

The z-coordinate of each line-of-sight halo is set by the redshift slice it is in, and the x- and y-coordinates are bound within a double cone. The cone is defined by an opening angle from the observer, $\theta_{\rm los}$, and peaks in radius at the main deflector redshift $z_{\rm lens}$. From there, the cone closes with an angle set by the requirement that the radius of the cone at the source redshift, $z_{\rm source}$, must be 0.2 of the radius at $z_{\rm lens}$. Within the cone, the x- and y-coordinates of the line-of-sight halo are sampled uniformly, as we show in Figure 1.

Finally, we must also add negative convergence sheets to the lensing potential to cancel the mean expected convergence from the line-of-sight halos. This procedure avoids rendering lines of sight that are systematically overdense relative to the matter density of the universe (Birrer et al. 2017a). At a fixed redshift, the deflection angles generated by our line-of-sight halos add linearly to one another. Therefore, we calculate the mean expected convergence by generating the deflection angles of an NFW halo with mean mass and concentration and then convolving the deflection angles with the uniform disk onto which we render our line-of-sight halos. The negative of these convolved deflection angles then defines the convergence sheet that must be added. Because this operation is expensive, we use a slightly wider binning in redshift, $\Delta_{z,\rm correction}$, for the convergence sheet calculations. We choose $\Delta_{z,\rm correction}$ so that the error in the average convergence is negligible.

2.4. Source

Previous studies using neural posterior estimators for strong lensing inference have focused on simple, parametric source models (Perreault Levasseur et al. 2017; Brehmer et al. 2019; Wagner-Carena et al. 2021). Here the sources in our simulation are drawn from 2262 real galaxy images taken by the HST COSMOS survey (Koekemoer et al. 2007). The images were taken using the HST Advanced Camera for Surveys (ACS; Ryon 2021) between 2003 October and 2005 June using the F814W filter. From this larger survey, we use the subsample of postage stamp images generated for the GREAT3 gravitational lensing challenge (Mandelbaum et al. 2012, 2014) and distributed with the package GALSIM[13] (Rowe et al. 2015). A more detailed discussion of this data set can be found in Mandelbaum et al. (2014), and we summarize the important points in Appendix C. Our pipeline takes this image catalog and imposes a few additional selection cuts to ensure that we are only using well-resolved galaxies: a minimum cutout size in pixels, size$_{\rm min,pix}$; a faintest apparent magnitude, mag$_{\rm faint}$; a maximum redshift, $z_{\rm catalog,max}$; and a minimum estimated half-light radius, $r_{1/2}$, in units of pixels. After inspecting all images that passed these cuts, we removed 110 images without a well-imaged galaxy. These mainly showed strong point sources or had much of their light masked out by preprocessing algorithms. We kept images with multiple blended or nearby galaxies. A sample of the 2262 remaining source images can be seen in Figure 2. The values we use for the cut parameters can be found in Table 1.

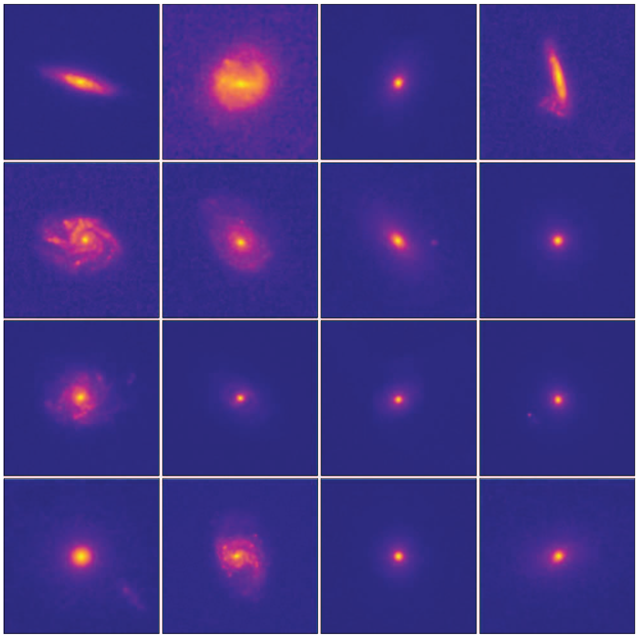

Figure 2. Sample of the HST COSMOS (Koekemoer et al. 2007) galaxy images being used as sources by the PALTAS pipeline. The cutouts shown here have been generated using the procedure discussed in Mandelbaum et al. (2014), as well as the additional cuts described in Section 2.4.

Our model for the light is then a linear interpolation of one of these 2262 images. This interpolation introduces three additional degrees of freedom to our source model: the rotation angle of the source, $\phi_{\rm source}$, and the x- and y-coordinates of the source, $x_{\rm source}$ and $y_{\rm source}$. For this source model, we keep the absolute luminosity and physical size of the galaxy fixed to what is observed in the COSMOS images. To do this, we first scale both the angular size and measured flux of the galaxy to the values that would be measured at the source redshift, $z_{\rm source}$. Then we convert from the electron count units of the F814W filter on the ACS detector to the electron count units of our target detector using the offset in the AB magnitude zero-points. For the F814W filter on ACS images, we assume an AB zero-point of 25.95, which is the average zero-point over the COSMOS survey period (Koekemoer et al. 2007; Mandelbaum et al. 2014; Ryon 2021).

2.5. Observational and Numerical Parameters

Our simulated lens images are made assuming observations by the HST Wide Field Camera 3 (WFC3) UVIS channel with the F814W filter. The low-mass halo signal we want to detect is

[13] The image database can be found at https://github.com/GalSim-developers/GalSim/wiki/RealGalaxy%20Data.

nearly at the level of the noise. Therefore, the distortions and correlated noise generated by the HST camera and pipeline are important systematics. Our simulations take care to include observational effects that are as realistic as possible.

We use the UVIS pixel size of $0\farcs040\,\mathrm{pixel}^{-1}$[14] (Gennaro 2018, Section 1.1), a CCD gain of 1.58 (Gennaro 2018, Section 5.1.1), and a read noise of $3e^{-}$ (Gennaro 2018, Section 5.1.2). For the AB magnitude zero-point, we take the 2020 corrected measurement of 25.127 (Calamida et al. 2021).[15] To calculate the expected combined sky noise and dark current, we use the HST Exposure Time Calculator,[16] which gives a total brightness of 21.83 mag arcsec^{-2} for average zodiacal light and earthshine conditions.

Our point-spread function (PSF) is pulled from the WFC3 PSF Database (Dauphin et al. 2021).[17] These PSF models have been empirically constructed from dithered observations of the star cluster Omega Centauri and are supersampled by a factor of 4 relative to the pixel size of the CCD. The modeling allows for spatial variation of the PSF across the detector, and we select the best empirical model for the center of UVIS chip 1. We degrade the PSF to a supersampling factor of 2 (i.e., twice the resolution of the CCD) in order to match the resolution at which we simulate our light rays.

The PALTAS simulations we use in this work also account for the effects of HST's drizzling pipeline. The HST science images are not just the raw detector output. Rather, the detector measurements are first run through the DRIZZLEPAC pipeline.[18] Besides standard corrections for the sky background and cosmic rays, the DRIZZLEPAC pipeline is also responsible for correcting the geometric distortions caused by the tilt of the focal plane relative to the optical axis of the detector. To do this, the DRIZZLEPAC pipeline uses the drizzle algorithm (Fruchter & Hook 2002) to combine the information from multiple dithered exposures into a final geometrically corrected image. Because the algorithm uses multiple exposures that are offset by subpixel intervals, the information content is at a higher resolution than the native resolution of the detector. Therefore, the images output by the drizzle algorithm can be at a smaller pixel scale than the detector. In fact, some amount of supersampling is important for minimizing both the aliasing and shape variation in the corrected image (Rhodes et al. 2007). From the simulation standpoint, the drizzling pipeline imposes two major challenges. First, the PSF is a property of the image in the detector plane; therefore, empirical PSF measurements are based on the resolution of the detector, not the resolution of the output of the drizzle algorithm. Second, the information in each pixel of the detector plane is spread out among several pixels in the final drizzled image. This leads to correlated noise that can mimic the small-scale deflection signal produced by low-mass halos.

In order to best capture the effect the DRIZZLEPAC pipeline will have on our sensitivity to the SHMF, PALTAS uses the DRIZZLE[19] software package to run the ray-traced images through the drizzling algorithm. Here ray tracing refers to the process of simulating the path of the emitted light rays from the source to the observer. The detailed steps can be found in Appendix D. The addition of the noise and the PSF convolution happen before the drizzling, giving the final image both a realistically distorted PSF and correlated noise.[20] For each of the four dithered images used in the DRIZZLE pipeline, we simulate a 23 minute exposure, equivalent to a total of two orbits of HST time per lens (Dressel 2021, Section 10.4.4). We choose a postage stamp of 128×128 pixels in the detector plane so that the final size of our images after drizzling is 170×170 pixels, or $5\farcs1 \times 5\farcs1$. Additionally, a mask of radius $0\farcs5$ is placed at the center of the image to represent the region that must be discarded due to a lens light subtraction process.

3. Inference Methods

Our simulation-based inference pipeline has three steps.

1. Generate a 500,000 image training set with wide distributions on the parameters of interest (Section 3.1).
2. Train a neural density estimator to estimate the posterior distribution of the parameters of individual lenses (Section 3.2).
3. Hierarchically combine the single-image posteriors of a large population of lenses to infer the distribution of SHMF normalizations (Section 3.3).

The following three sections go into more detail on each of these steps, and the pipeline is further summarized in Figure 3.

3.1. Simulated Data Sets

For the results presented in this work, we have simulated 22 data sets: one training data set, one validation data set, and 20 test data sets. The parameter distributions for the training, validation, and test data sets can be found in Table 1. The training data set is used to fit the network parameters and consists of 500,000 strong gravitational lensing images following the simulation choices outlined in Section 2. This includes a PEMD main deflector with external shear, subhalos of the main deflector, halos along the line of sight, sources pulled directly from HST COSMOS observations, and the observational properties of HST's WFC3. The validation data set is drawn from the same distributions as the training data set but includes only 1000 images. The validation data set is used to evaluate the trained network's performance on individual lenses.

The 20 test data sets have been constructed to test our pipeline's ability to constrain a range of SHMF normalizations. Each test data set contains 1000 images generated using the same models as the training data set; however, the distribution of the eight parameters that describe the main deflector and the parameter controlling the normalization of the SHMF have been modified. These are the nine parameters that our network infers (see Section 3.2). The mean SHMF normalizations linearly span the interval $\Sigma_{\mathrm{sub}} \in [2 \times 10^{-4}\,\mathrm{kpc}^{-2}, 4 \times 10^{-3}\,\mathrm{kpc}^{-2}]$. For the eight main deflector parameters, the test distributions are narrower than the training data set parameter distributions and shifted uniformly by half a sigma. These shifts will be used to confirm that our inference of the SHMF normalization is not

[14] Note that due to drizzling, the final resolution of our images is $0\farcs030\,\mathrm{pixel}^{-1}$, not $0\farcs040\,\mathrm{pixel}^{-1}$.

[15] Summary of results available on the STScI website.

[16] https://etc.stsci.edu/etc/input/wfc3uvis/imaging

[17] A selection of PSFs can be found at https://www.stsci.edu/hst/instrumentation/wfc3/data-analysis/psf, and the specific PSF map used by PALTAS can be found at https://github.com/swagnercarena/paltas/blob/main/data sets/hst_psf/emp_psf_f814w.fits.

[18] https://stsci.edu/scientific-community/software/drizzlepac.html

[19] https://github.com/spacetelescope/drizzle

[20] Ding et al. (2018) captured the effects of the drizzle algorithm using a similar pipeline.

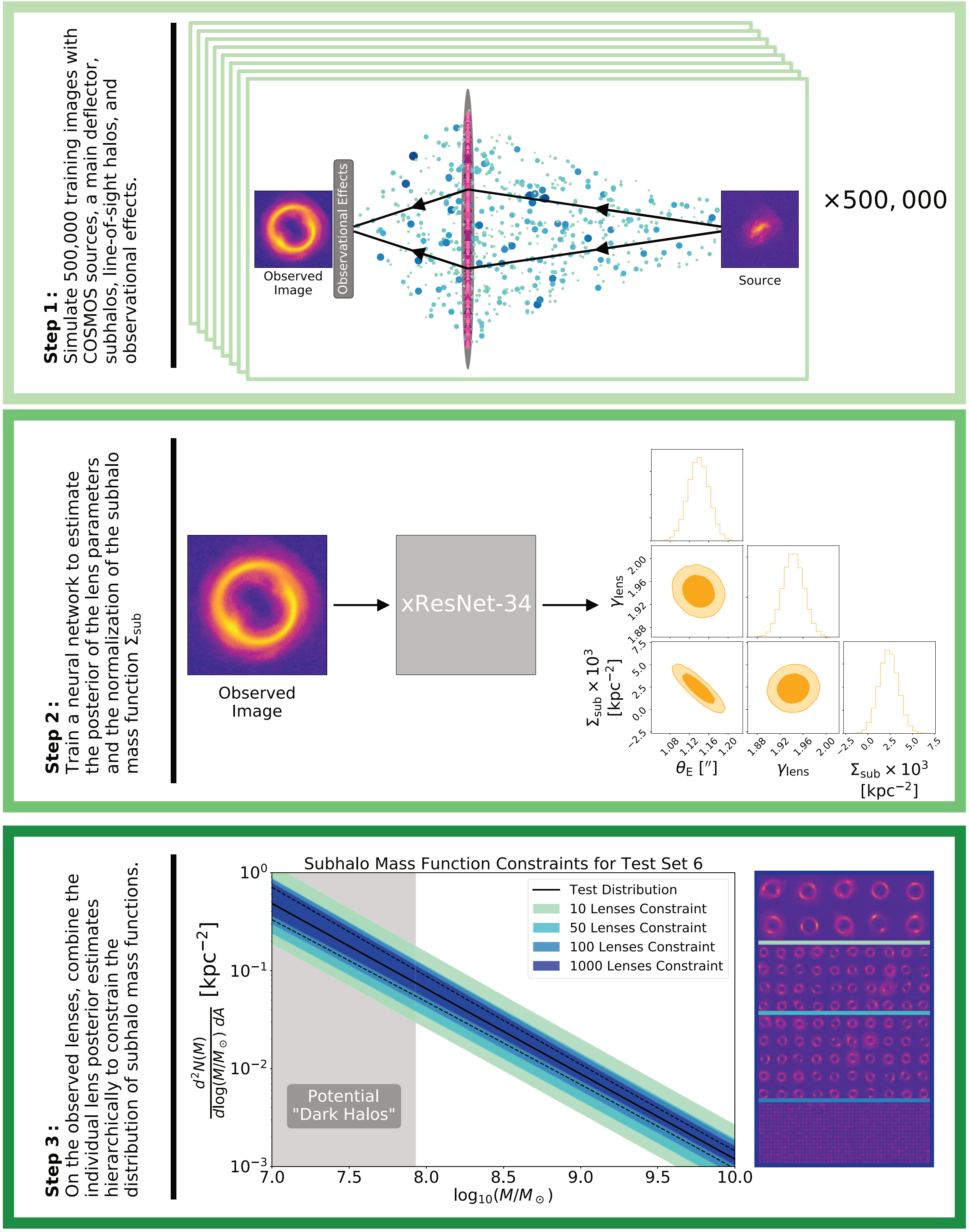
Step 1 :
Simulate 500,000 training images with COSMOS sources, a main deflector, subhalos, line-of-sight halos, and observational effects.
Observed Image
Observational Effects
Source
×500, 000
Step 2 :
Train a neural network to estimate the posterior of the lens parameters and the normalization of the subhalo mass function Σsub
Observed Image
xResNet-34
γlens
1.88
1.92
1.96
2.00
Σsub × 10^3 [kpc^−2]
−2.5
0.0
2.5
5.0
7.5
1.08
1.12
1.16
1.20
θE ["]
1.88
1.92
1.96
2.00
γlens
−2.5
0.0
2.5
5.0
7.5
Σsub × 10^3 [kpc^−2]
Step 3 :
On the observed lenses, combine the individual lens posterior estimates hierarchically to constrain the distribution of subhalo mass functions.
Subhalo Mass Function Constraints for Test Set 6
Test Distribution
10 Lenses Constraint
50 Lenses Constraint
100 Lenses Constraint
1000 Lenses Constraint
d²N(M) / dlog(M/M⊙) dA [kpc^−2]
10^0
10^−1
10^−2
10^−3
Potential "Dark Halos"
7.0
7.5
8.0
8.5
9.0
9.5
10.0
log10(M/M⊙)

Figure 3. Summary of our simulation-based inference pipeline. The top panel depicts the first step: generating 500,000 simulated lenses that include realistic sources, dark matter structure, and observational effects (see Section 3.1). Second, we use these 500,000 lenses to train a neural posterior estimator. This network constrains the lensing parameters on individual strong lensing images; the middle panel shows the posteriors returned by the trained network for a single lens image (see Section 3.2). Finally, we combine these individual lens posteriors hierarchically to constrain the SHMF for a population of lenses. The bottom panel shows the SHMF constraints derived from lens images drawn from a different lens parameter, source, and SHMF distribution than the training set (see Section 3.3).

biased by the choices of the training set. We will discuss this further in Sections 3.3 and 4.2.

The COSMOS catalog parameters have been chosen to use galaxies that are relatively nearby and better resolved. We impose these cuts because we are interested in exploring the degeneracy between galaxy morphology and the subhalo signal. However, this also means that for this analysis, we assume the population of large, nearby galaxies resembles the high-redshift population. In Section 5, we will discuss alternatives to this assumption for future analysis. The cuts we impose are the same for the training, validation, and test data sets; however, of the 2262 galaxies that survive our cuts, 2163 are reserved for training, and 99 are reserved for validation and test images. Therefore, at test time, we are conducting inference on sources the network has never seen.

For the training, validation, and test data sets, the SHMF slope is drawn from a wide uniform distribution to represent marginalizing over theoretical uncertainties estimated using cosmological CDM simulations and semianalytic models (Benson 2020). Similarly, we account for theoretical uncertainty in the line-of-sight structure mass function by drawing its normalization uniformly between zero (no line-of-sight structure) and twice the theoretical expectation. We also draw our mass–concentration relation parameters from a wide uniform distribution to cover values favored by modern large-scale structure simulations (Bullock et al. 2001; Prada et al. 2012; Ludlow et al. 2016; Diemer & Joyce 2019). Finally, we draw the mass–concentration relation scatter from a uniform distribution that covers the results of cosmological CDM simulations (Dutton & Maccio 2014; Diemer & Kravtsov 2015; Diemer & Joyce 2019).

Both the SHMF normalization and the line-of-sight mass function normalization are allowed to take on negative values in the distributions we have assigned. These negative values are not physically meaningful; when simulating, we treat a negative normalization as though it were zero. We chose to include these negative values rather than truncate our normal distributions because it allows us to conduct our hierarchical inference analytically[21] (see Section 3.3).

The remaining parameters presented in Table 1 are discussed in Appendix F.

3.2. Posterior Distribution for Individual Lenses

For each individual lens, our goal is to estimate the posterior of the parameters of interest given the image. To do this, we employ simulation-based inference using a neural density estimator. The estimator, $q_{F(d,\phi)}(\xi)$, approximates the posterior, $p(\xi|d,\ \Omega_{\text{int}})$. Here F is our neural network, q is a density function, ϕ are the parameters of that network, d is the strong lensing image, ξ are the physical parameters that we wish to constrain, and Ω_{int} is a prior distribution on the parameters ξ. The generation of the training set described in Section 3.1 can be thought of as first sampling lensing parameters from our training distribution $\xi_k \sim p(\xi|\Omega_{\text{int}})$ and then using our simulator, g, to generate an image, $d_k \sim g(\xi_k)$. To train the network weights, ϕ, we then minimize the loss function:

$$L(\phi) = -\sum_{k=1}^{N} \log q_{F(d_k,\phi)}(\xi_k). \tag{9}$$

In the limit where $N \to \infty$ and q is sufficiently flexible, Equation (9) guarantees that the network will learn $q_{F(d,\phi)}(\xi) \to p(\xi|d,\ \Omega_{\text{int}})$ (for a proof of this statement, see Papamakarios & Murray 2016, Appendix A). While the prior Ω_{int} is never explicitly enforced in our loss function, it is implicitly learned via the distribution of parameters ξ_k that the network is exposed to. We have chosen to label this distribution Ω_{int} because it is an interim choice; it is a prior that is optimized for training, but it is not necessarily the distribution we expect real lenses to follow. In Section 3.3, we will discuss our hierarchical inference methodology for extracting the true distribution of a population of lenses.

For the model architecture, $F(d,\ \phi)$, we implement xResNet-34 (He et al. 2016, 2019). For the density function q, we use a multivariate Gaussian with a full precision matrix (see Section 3.1 of Wagner-Carena et al. 2021 for implementation details). The final fully connected layer of our xResNet-34 architecture is modified to predict the 54 parameters of our multivariate Gaussian (nine means and 45 free parameters of the precision matrix), and the first layer's filters are adapted to read monochromatic images. All layer weights are randomly initialized using the Xavier uniform initialization.

The model presented here is implemented using the TENSORFLOW (Abadi et al. 2015) library in Python and trained on a NVIDIA GeForce RTX 2080 Ti GPU. The model is first trained for 100 epochs to predict the mean and the diagonal elements of the precision matrix. Then the model is trained for a further 100 epochs with the full precision matrix, but only the parameters for the final layer of the network are updated. This training scheme helps reduce the instability generated by the precision matrix terms in the loss function. We use a batch size of 256 and the Adam optimizer. The learning rate is set to 1×10^{-5}, and the default Adam parameter values of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-7}$ are used. A decay rate of 0.98 is applied to the learning rate at the end of each epoch. In total, training takes ≈96 hr.

Each individual image is normalized to have a standard deviation of 1, and the network is trained to predict output parameters that are normalized to have mean zero and standard deviation 1 on the full training set. The constants used for this normalization are saved so that the network outputs can be translated back to the physical parameters. As an additional training augmentation, each batch of images is randomly rotated before being fed into the network, and the lensing parameters are corrected accordingly. At inference time, we find that repeating this same random rotation improves the quality of inference. Therefore, all of the mean predictions, except those for x_{lens}, y_{lens}, are averaged over 100 random rotations of the input image. The covariance predictions are not averaged over rotations.

3.3. Hierarchical Inference

The final goal of our analysis is to constrain the distribution governing the normalization of the SHMF, Σ_{sub}, given a set of strong lensing images. We can write this as $p(\Omega|\{d\})$, where Ω is the population-level distribution for our lensing parameters $\{\xi\}$, and $\{d\}$ is our set of strong lensing images. As we introduced in Section 3.2, the model does not approximate the likelihood $p(\xi|d)$ but rather the posterior $p(\xi|d,\ \Omega_{\text{int}})$. We must chose Ω_{int} at training time; therefore, we cannot vary it to calculate $p(\Omega|\{d\})$. Instead, we must reweight our network's posterior estimates, $q_{F(d,\phi)}(\xi)$, by the ratio of the likelihood of drawing a parameter ξ given the proposed distribution Ω versus the training distribution Ω_{int}:

$$p(\Omega|\{d\}) = \underbrace{p(\Omega)}_{\Omega\ \text{prior}} \times \underbrace{\prod_{k}^{N_{\text{lens}}} \frac{p(d_k|\Omega_{\text{int}})}{p(\{d\})}}_{\text{normalizing factor}} \times \underbrace{\prod_{k}^{N_{\text{lens}}} \int \frac{p(\xi|\Omega)}{p(\xi|\Omega_{\text{int}})} q_{F(d_k,\phi)}(\xi)\ d\xi}_{\text{importance}-\text{sampling integral}}. \tag{10}$$

A more detailed derivation of Equation (10) can be found in Appendix C of Wagner-Carena et al. (2021). Note that only

[21] We tried inferring the logarithm of these normalization parameters instead but saw increased non-Gaussianity in the joint posteriors with other parameters, such as θ_{E}.

the first and third terms of the right-hand side depend on Ω. The first term is simply a hyperprior on the distribution Ω,[22] the second term is a constant normalizing factor, and the third term encodes all of the constraining power provided by the population of images. The ratio of $p(\xi|\Omega)$ to $p(\xi|\Omega_{\rm int})$ is often called an importance-sampling weighting. To keep the integral tractable, we must chose $\Omega_{\rm int}$ such that the term $p(\xi|\Omega)/p(\xi|\Omega_{\rm int})$ is finite for all ξ. Practically speaking, this means that the training distribution $\Omega_{\rm int}$ must be broader than any distribution Ω we would like to infer. Similarly, while it is possible to evaluate the integral numerically (see Wagner-Carena et al. 2021), having an analytic solution for the integral allows for faster sampling of $p(\Omega|\{d\})$. These two constraints inform our choice of training distribution (broad Gaussian distribution) and test distributions (narrow Gaussian distributions).

We use Equation (10) to place constraints on the distribution of $\Sigma_{\rm sub}$. To do this, we first pass the lens images in our test set through our trained network. This gives us the posterior prediction $q_{F(d_k,\phi_j)}$ for each lens k and network j. We then sample Equation (10) to get the posterior on Ω. For the experiments conducted in this work, both Ω and $\Omega_{\rm int}$ are described by nine means and nine standard deviations. These means and standard deviations describe the distribution of our eight main deflector parameters and the SHMF normalization, $\Sigma_{\rm sub}$. To sample the posterior, we use an ensemble sampler with affine invariance (Goodman & Weare 2010) implemented through the EMCEE package[23] (Foreman-Mackey et al. 2013).

4. Results

4.1. Performance on Individual Lenses

As an initial test of the performance of our neural posterior estimator, we run the network on our validation set. As discussed in Section 3.1, our validation set draws from the same underlying parameter distributions used for the training set but uses a held-out set of 99 galaxy sources. In Figure 4, we show the posterior predictions for one randomly selected validation image. The most notable feature is that the posteriors on $\Sigma_{\rm sub}$ are nearly as wide as the training distribution. The neural posterior estimator predicts a standard deviation of $1.0 \times 10^{-3}\,{\rm kpc}^{-2}$ for the SHMF normalization, $\Sigma_{\rm sub}$, compared to the standard deviation of $1.1 \times 10^{-3}\,{\rm kpc}^{-2}$ used to draw the SHMF normalization in the training data set. The network posterior also includes a strong covariance between the SHMF normalization and the Einstein radius, $\theta_{\rm E}$. This covariance is a product of our definition of the Einstein radius; as we add more subhalos to our simulation, we increase the effective mass of our main deflector and therefore increase the observed radius of the ring in the image. However, the "true" value of the Einstein radius only accounts for the mass in our PEMD main deflector. Therefore, simultaneously decreasing the Einstein radius of the main deflector and increasing the mass in subhalos produces a similar observation. The negative correlation output by our network is a consequence of that degeneracy.[24]

Figure 5 compares the predicted mean values of SHMF normalization and the Einstein radius to the true value across the entire validation data set. The Einstein radius appears to be well constrained by the data, with the predicted mean being tightly correlated to the true value ($\rho = 0.991$).[25] While not visualized in Figure 5, the same is true for all eight of the main deflector parameters our network predicts (see Appendix I). The SHMF normalization, however, is rather poorly constrained, with the mean predictions clustering around the training set mean of $2.0 \times 10^{-3}\,{\rm kpc}^{-2}$ and weakly correlated to the true value ($\rho = 0.281$). In agreement with previous work (Perreault Levasseur et al. 2017; Pearson et al. 2021; Wagner-Carena et al. 2021), the network is capable of precisely and accurately constraining the main deflector parameters. But the estimated posteriors for the SHMF normalizations are dominated by the training prior, with little information extracted from any individual lens. The signal produced by a fixed SHMF normalization is highly stochastic (see Appendix E for examples), so the comparatively weak correlation produced by our network likely reflects the poor information content of the data.

However, as we will explore in Section 4.2, a weakly correlated but statistically consistent posterior can be sufficient to extract population constraints. Figure 5 colors the predicted means according to their distance from the truth in units of the predicted standard deviation. Because the one-dimensional predictions are Gaussian, a statistically consistent posterior would have the true values Gaussian distributed around the mean. That is what we find on the validation set: 69%, 95%, and 99% of the true values fall within one, two, and three standard deviations of the mean, respectively.

4.2. Reconstructing the SHMF

To test the network's ability to constrain a population of lenses, we run our hierarchical inference pipeline on the 20 test sets described in Section 3.1. We first pass the 1000 images in each of our test sets through our neural posterior estimator. Then, we use Equation (10) to simultaneously constrain the population distribution of the SHMF normalization and the eight main deflector parameters. Figures 6 and 7 show the inferred SHMF distribution for test sets 12 ($\Sigma_{\rm sub,pop} = 2.4 \times 10^{-3}$) and 4 ($\Sigma_{\rm sub,pop} = 0.8 \times 10^{-3}$), respectively. The constraints on the SHMF include the uncertainty on the inferred population mean, $\Sigma_{{\rm sub},\mu}$, and standard deviation, $\Sigma_{{\rm sub},\sigma}$, of the SHMF normalization. The constraining power at very low subhalo mass ($<10^8\,M_\odot$) is driven by our prior assumption about the SHMF slope, $\gamma_{\rm sub}$. For test set 12, 10 lenses are sufficient to strongly disfavor the existence of halos without subhalos. As we push the analysis from 10 to 1000 lenses, the inferred SHMF converges toward the true test distribution, with the uncertainty in the mean and scatter in the 1000 lens analysis being subdominant to the intrinsic variation in the lens population. For test set 4, the 10 lens analysis fails to exclude the existence of halos without subhalos within its 95% confidence interval, but the 50 lens analysis already strongly disfavors that scenario. Unlike test set 12, we can see an upward bias in the inferred SHMF for the 1000 lens analysis, but the 10, 50, and 100 lens analyses are consistent with the truth.

In Figure 8, we show the constraints on the population mean of each test set as a function of the true population

[22] This hyperprior allows for us to enforce physical constraints on the proposed distribution, including constraining the inferred population mean of the SHMF normalization to be greater than zero.

[23] https://emcee.readthedocs.io

[24] Due to the strength of this degeneracy, we include additional tests in Appendix G that confirm that the inferred $\Sigma_{\rm sub}$ for a test population of lenses is not systematically biased by shifts in the underlying distribution of $\theta_{\rm E}$.

[25] Here ρ is the Pearson correlation coefficient.

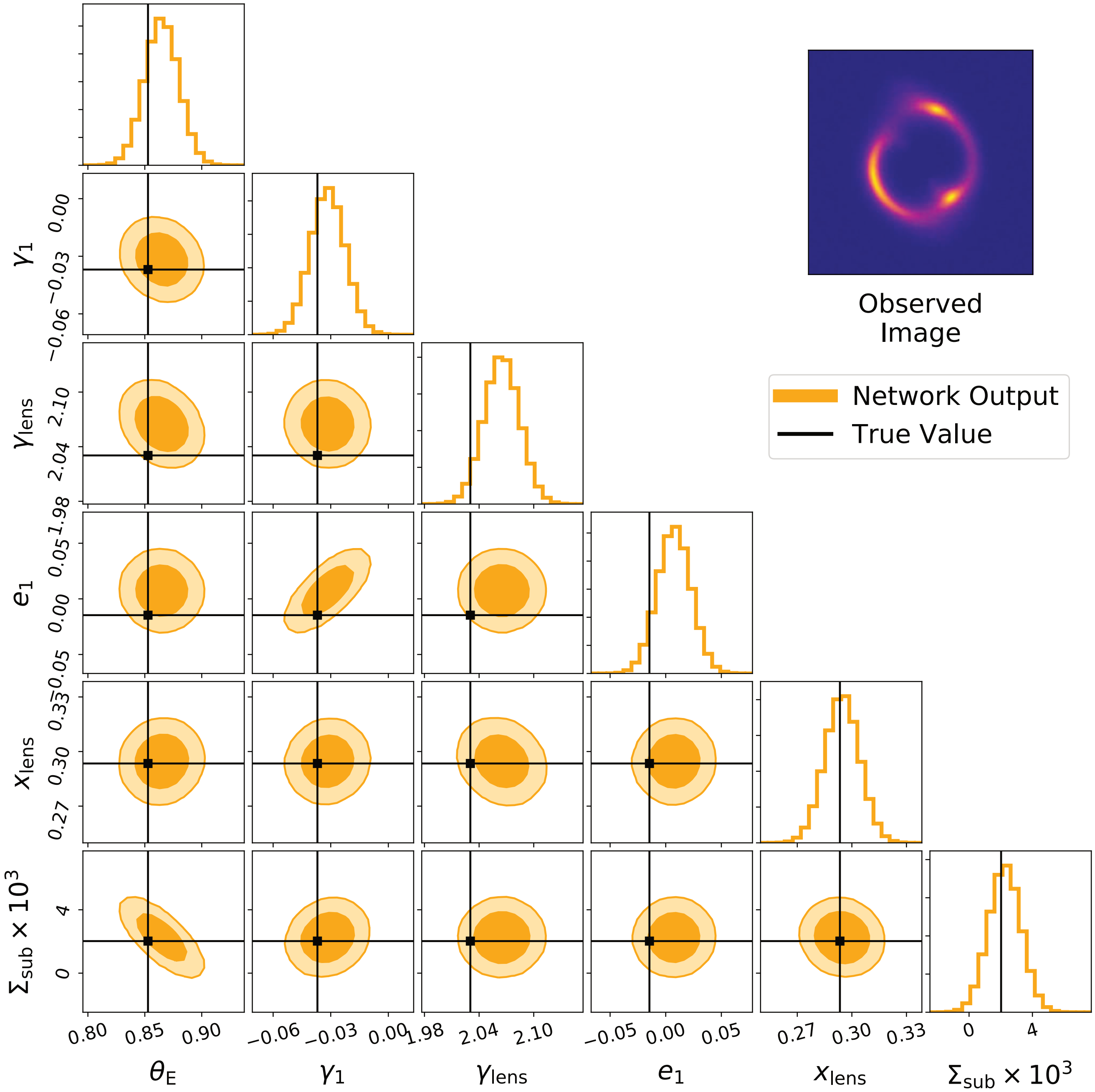

Figure 4. Example of the estimated posteriors output by our network for a simulated observation. The black points, labeled "true value," represent the input parameters used to simulate the image in the upper right. The yellow contours represent the multivariate Gaussian output by our neural posterior estimator. The darker and lighter contours correspond to the 68% and 95% confidence intervals, respectively. The posterior output for γ_2, e_2, and $y_{\rm lens}$ has been omitted to avoid visual clutter.

mean. Despite the weak correlation discussed in Section 4.1, the analysis pipeline returns a nearly linear response between the true and inferred population means. For the 10, 50, and 100 lens analyses, we find that the pipeline returns unbiased results for all but the largest values of $\Sigma_{\rm sub,pop}$. For $\Sigma_{\rm sub,pop} > 3.6 \times 10^{-3}\,{\rm kpc}^{-2}$, the inferred population mean for 10, 50, and 100 lenses is an underestimate. When we push the analysis to 1000 lenses, the inferred mean is still linearly determined by the truth, but it now exhibits a systematic bias for nearly all values of $\Sigma_{\rm sub,pop}$. Specifically, it pushes its estimates toward the training prior, overestimating values of $\Sigma_{\rm sub,pop}$ less than 2×10^{-3} and underestimating values of $\Sigma_{\rm sub,pop}$ greater than 2×10^{-3}. This pattern of bias suggests that the network has an imperfect understanding of the training prior $\Omega_{\rm int}$. In order to map from the prior-dominated constraints shown in Figure 5 to the hierarchical results in Figure 8, we divide by the interim prior as shown in Equation (10). This division should cancel the influence of the training prior in the network's posterior. However, the network learns the training prior implicitly from the training data set. If the network's learned estimate of the training prior is too wide, the division will undercorrect the network posteriors and bias the overall estimate toward the training prior mean. This is consistent with the shift we see in Figure 8 for the 1000 lens analysis. The discrepancy in the learned and true training prior is sufficiently small that it does not appear to meaningfully impact the 10, 50, or 100 lens analyses.

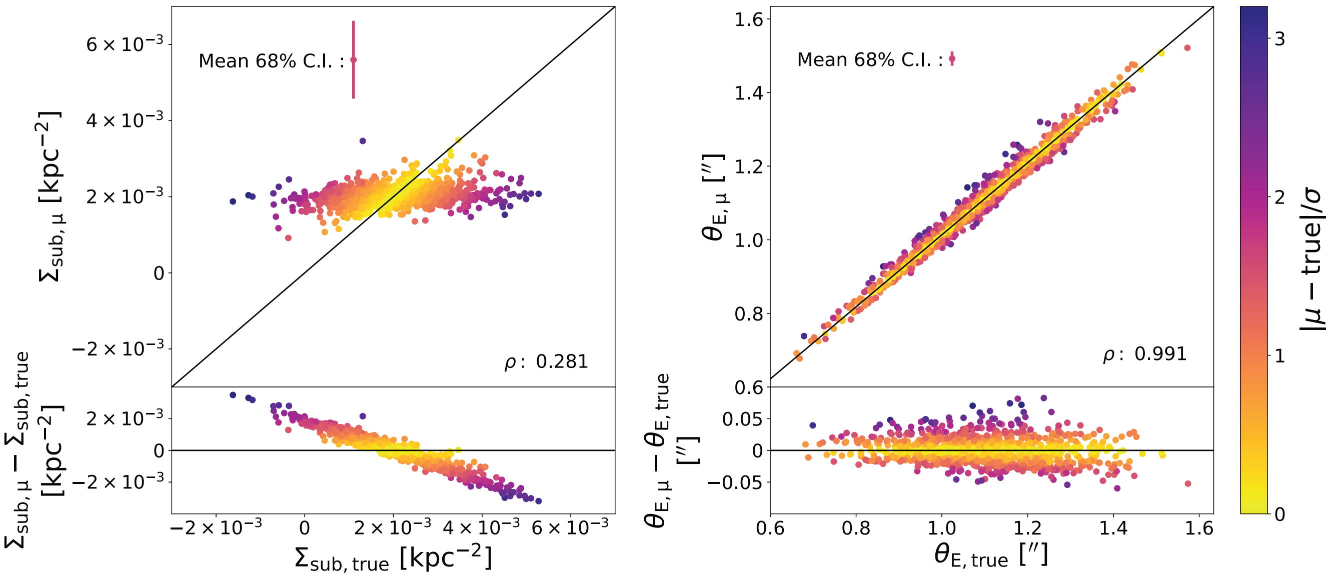

Figure 5. Comparison of the network predicted means and the true input values for all of the lenses in our validation set. The left panel shows the comparison for the SHMF normalization $\Sigma_{\rm sub}$, and the right panel shows the comparison for the Einstein radius $\theta_{\rm E}$. For both panels, one point corresponds to the predictions of one lens. The color of the point is defined by the distance between the predicted mean and the truth in units of the predicted standard deviation. The color bar for both panels is on the far right. The upper left includes a visualization of the average standard deviation being predicted by the network (equivalent to the 68% confidence interval). The ρ value printed in the plots represents the Pearson correlation coefficient between the predicted mean and the truth. The remaining parameters are shown in Figure 14.

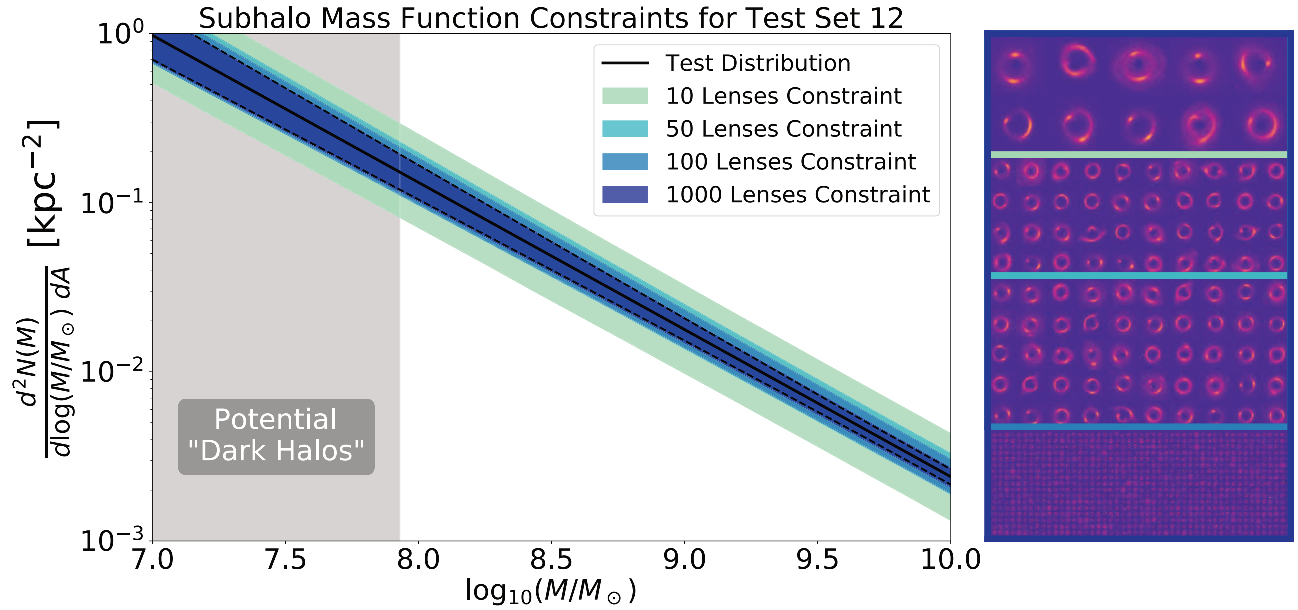

Figure 6. The SHMF constraints for test set 12 as a function of the number of lenses used in the analysis. In the left panel, the black solid and dashed lines show the range containing 95% of the SHMFs coming from the test distribution. The colored regions show the hierarchical inference estimate, which includes marginalizing over the uncertainty in the mean and standard deviation of $\Sigma_{\rm sub}$. The gray region labeled "potential dark halos" shows the current limits on halos without luminous counterparts from Milky Way satellites (Nadler et al. 2020). The right panel shows the lenses included in the 10, 50, 100, and 1000 lens analyses with the colors corresponding to the regions in the left panel.

Figure 8 also includes the range of SHMF normalizations for $\approx 10^{13}\,M_\odot$ halos found in a suite of high-resolution cosmological dark matter–only (DMO) zoom-in simulations (E. Nadler et al. 2022, in preparation). The SHMF normalization has been measured at a redshift of $z = 0.5$ and includes the subhalos within the projected virial radius with masses in the range $[9.4 \times 10^8\,M_\odot,\ 1 \times 10^{10}\,M_\odot]$. We plot the region containing 68% and 100% of the SHMF normalizations. The "host-to-host" scatter in the measured SHMF normalization is connected to the secondary properties of the main deflector halo and its assembly history (see Mao et al. 2018 for a review). Distinguishing between a mean SHMF normalization in the upper and lower halves of this range would enable meaningful measurements of the concentration of main deflector halos and the environment in which they form. Alternatively, if a strong lensing analysis returned a $\Sigma_{\rm sub,hier}$ constraint that did not overlap with this region, then there would exist a tension between strong lensing measurements and DMO predictions that could not be explained by selection effects. The 50, 100, and 1000 lens analyses produce

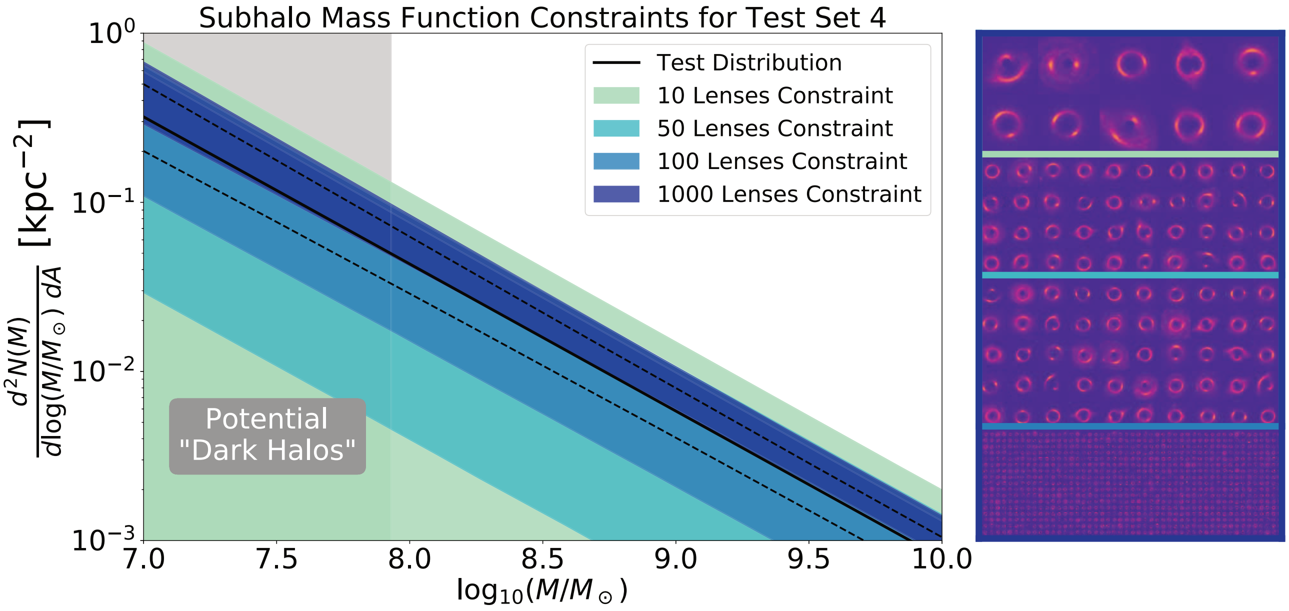

Figure 7. Same SHMF constraints as in Figure 6 but for test set 4.

constraints smaller than the DMO-predicted host-to-host variation in the SHMF normalization, and even the 10 lens analysis can detect a tension between strong lensing measurements and DMO predictions for sufficiently high $\Sigma_{\rm sub,pop}$. In terms of detecting the existence of subhalos (i.e., inferring $\Sigma_{\rm sub,hier} > 0$), 50 lenses are sufficient for the full range of SHMF normalizations present in the DMO simulation suite. The 50, 100, and 1000 lens analyses provide a 32%, 22%, and 6% precision measurement of the normalization in the region favored by the DMO simulation suite.

5. Discussion

We have presented a pipeline for inferring the SHMF normalization of a large population of lenses and probed the robustness of that pipeline on test data sets with a wide range of SHMF normalizations and main deflector parameters. In this section, we will discuss the limitations of our analysis, ways in which these limitations can be addressed, and avenues for future work.

Despite weak, prior-dominated constraints on the level of individual lenses, analyzing the neural posterior estimator output hierarchically allows us to accurately reconstruct the SHMF normalization across a wide portion of parameter space. This includes inference on data sets that are built with COSMOS sources not seen during training and marginalizing over the effects of line-of-sight halos and the mass–concentration relation. The results also scale well to large data sets, as adding more lenses to our analysis consistently improves the constraining power of our network and has a negligible impact on the computational time of the pipeline. However, for the 1000 lens analysis and the high SHMF normalization region, the normalizations inferred by the pipeline show systematic bias. Therefore, with the current choices, the analysis we present is limited to lens populations with ≈100 lenses or inferred SHMF normalizations that are well sampled by the training set. There are a number ways to potentially circumvent this limitation. For example, we could conduct sequential neural posterior estimation (SNPE), a scheme of iterative retraining where the inferred SHMF population is used as a prior for a new training set (see Greenberg et al. 2019 for a review of SNPE implementations). By construction, this additional training would ensure that the inferred SHMF normalizations fall in a region well sampled by the training distribution, reducing the tension between the test population and the assumptions made during training. We leave introducing retraining to our methodology for future work.

Additionally, while we have made significant efforts to create realistic and complex simulations, there are still assumptions in our current analysis that will need to be modified for observed lenses. For example, the use of COSMOS galaxies for our sources enables us to marginalize over realistic morphological source complexity but requires cutting on nearby, well-resolved galaxies. This selection assumes that low-redshift galaxies are representative of high-redshift galaxies. The population-level differences in the morphology of high- and low-redshift galaxies may be degenerate with the small-scale subhalo perturbations and could generate a bias on the inferred SHMF. One approach would be to develop more realistic simulated images of high-redshift source galaxies that can be realized at the required superresolution and input into our simulation pipeline.

We have also assumed a fixed cosmology throughout our analysis. Modifying the cosmological parameters would affect our halo profiles, the volume from which we sample line-of-sight halos, and the distances that set the angular size of the galaxies and halos in our simulation. Similarly, we have assumed broad priors on the mass–concentration parameters, the slope of the SHMF, and the line-of-sight halo mass function.[26] This allows us to probe the sensitivity of our analysis when it is marginalized over the existing theoretical uncertainties, but a broad, mis-centered prior can systematically shift the inferred SHMF normalization. On real data, we could choose a carefully constrained set of theoretical priors and recognize that the analysis is conditioned on those assumptions. Alternatively, the parameters governing the source population, SHMF slope, mass–concentration parameters, and line-of-sight

[26] In Appendix H, we explore how changing the distribution of line-of-sight halo normalizations affects the inferred SHMF normalization.

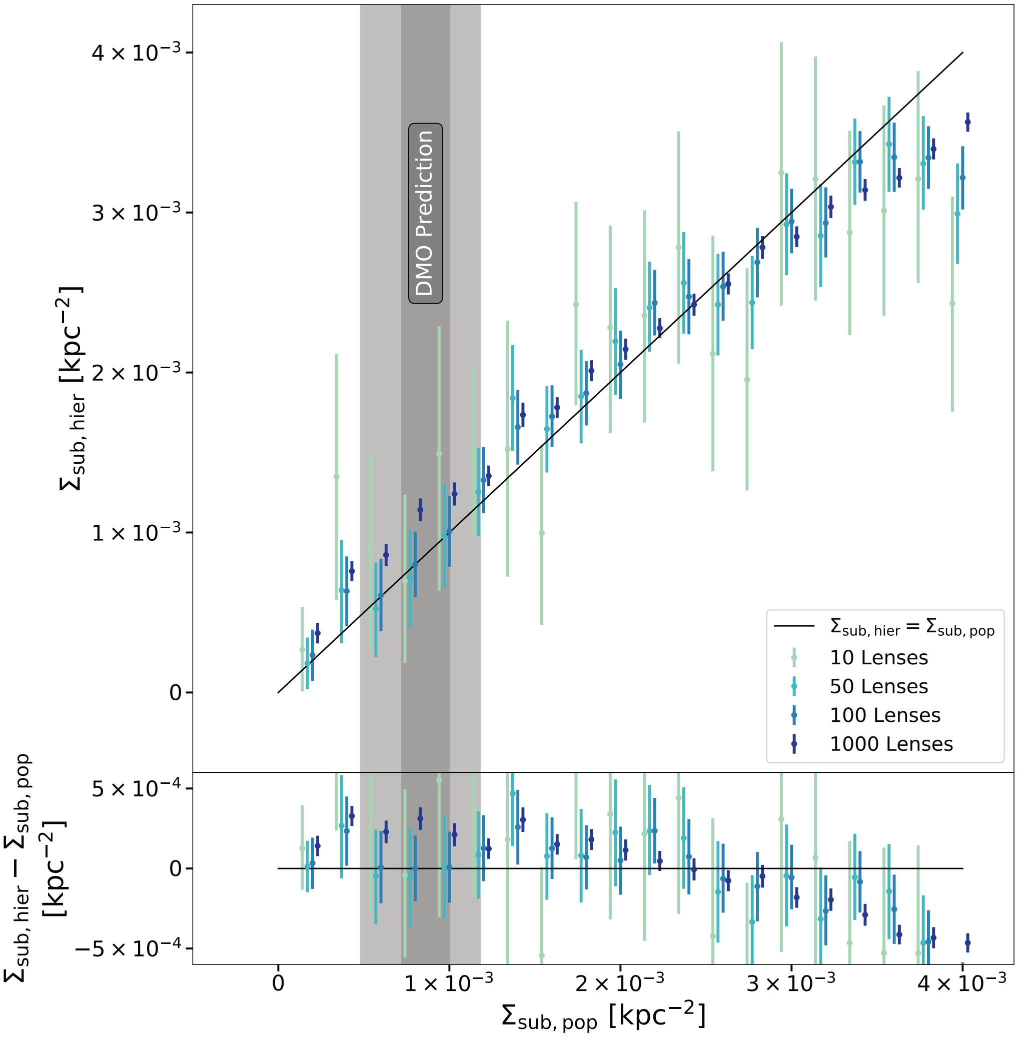

Figure 8. Inferred ($\Sigma_{\rm sub,hier}$) vs. true ($\Sigma_{\rm sub,pop}$) mean of the SHMF normalization distribution (top) and the difference between these quantities (bottom). Each set of bars represents constraints using 10, 50, 100, or 1000 lenses from one of our 20 test sets. For a given test set, all analyses have the same $\Sigma_{\rm sub,pop}$, but the coordinates of the plotted points have been slightly offset for visual clarity. The gray shaded regions contain 68% and 100% of the SHMF normalizations for $\approx 10^{13}\,M_\odot$ host halos at redshift $z = 0.5$ measured on a suite of high-resolution cosmological DMO zoom-in simulations (E. Nadler et al. 2022, in preparation). The error bars on the inferred $\Sigma_{\rm sub,hier}$ show the 68% interval derived from the hierarchical inference pipeline.

halos can be fit on a lens-by-lens level as we do for the main deflector parameters. We could then infer these parameters hierarchically from the data. The multivariate Gaussian posterior we use is not sufficiently expressive for this task, but a network leveraging normalizing flows or another more flexible posterior could circumvent this limitation. For the parameters on which we assume broad priors, either more constrained priors or a hierarchical inference approach would yield tighter constraints than the marginalization we use in this work. In that respect, the sensitivity to the SHMF normalization discussed in Section 4.2 is conservative. We leave exploring this extension of our hierarchical approach for future work.

Lastly, while our simulation and inference pipelines account for the full complexity of low-mass halos, we have focused our analysis on the normalization of the SHMF. Similarly, we have only considered the use of HST images. These choices allow us to illustrate our ability to infer dark matter parameters from large samples of lenses but leave two major avenues for future work: exploring the use of different imaging data sets and quantifying low-mass halo parameters beyond the SHMF normalization. In terms of data sets, interferometers like the Atacama Large Millimeter/submillimeter Array are capable of returning higher-resolution images of strong lenses, albeit with the addition of significant data complexity (Hezaveh et al. 2013b). No part of our methodology is inherently resolution-limited. Incorporating interferometry measurements would require either modifying the observational effects to include the artifacts introduced from Fourier space measurements or modifying the network architecture to take the Fourier space signal as input. In turn, these higher-resolution images could yield substantially more constraining power on the SHMF. In terms of the parameters we measure, future studies will also want to constrain a low-mass cutoff for both the SHMF and line-of-sight mass functions. This cutoff captures the behavior expected in several alternative dark matter models and would allow us to better quantify what mass range our model is sensitive to as a function of the number of lenses in our analysis.[27]

[27] For some preliminary discussion of our sensitivity to the mass cutoff, see Appendix E.

6. Conclusions

We present a simulation-based inference methodology for measuring the SHMF normalization of strong lensing systems. Leveraging our simulation package, PALTAS, we have trained a neural posterior estimator on a set of 500,000 synthetic lens systems with sources pulled directly from the COSMOS field, realistic low-mass halo distributions, and HST observational effects. We have tested this network against a series of test data sets and demonstrated its ability to generalize to COSMOS galaxies not seen in the training data set. Our network returns SHMF normalization constraints that are dominated by the prior on the individual lens level, but our hierarchical inference can extract unbiased constraints for populations with 10, 50, and 100 lenses across a wide range of normalizations, including the SHMF normalizations predicted by DMO simulations. Analyses with 1000 lenses are computationally accessible and improve the precision at the cost of some systematic bias. We discuss how iterative retraining of the network using simulated data sets that match the inferred population would likely alleviate this bias.

Galaxy–galaxy strong gravitational lenses are sensitive to dark matter on the small scales that best constrain alternatives to CDM. Previous work with strong lenses, including in the subhalo context, has shown that neural density estimators can extract accurate and precise parameter estimates and scale to populations with thousands of images. Extending these analysis techniques to observed lenses requires simulation tools capable of capturing the complexity of real data, as well as rigorous testing of the robustness of the networks. We believe that this work makes significant contributions to both of these challenges; our simulated data sets have been carefully constructed to be realistic and are produced with a publicly available and well-documented software package. The combined neural posterior estimator and hierarchical inference method we propose scales to large lens populations and has been tested for robustness against unseen COSMOS galaxies, shifts in the main deflector parameter distribution, and variations in the underlying SHMF. We have shown that our method performs well across nearly all of these robustness tests and discussed improvements that can be made to future analyses to address the shortcomings that do exist. We are confident that the simulation-based inference pipeline presented in this work is capable of constraining dark matter substructure using both the HST strong lensing images that exist today and the data sets that will become available with the next generation of wide-field optical imaging surveys.

We would like to thank Phil Mansfield for numerous helpful discussions during the development of this work. We would also like to thank María Luz Raggio for helping with the design of the PALTAS logo. S.W.C. was supported by NSF award DGE-1656518 and a Stanford Data Science Fellowship.

This work received support from the Kavli Institute for Particle Astrophysics and Cosmology through Kavli Fellowships to S.B. and J.A. and the U.S. Department of Energy under contract No. DE-AC02-76SF00515 to SLAC National Accelerator Laboratory.

Appendix A Subhalo Profiles

The subhalos in our simulation are modeled as a truncated NFW radial density profile (Baltz et al. 2009),

$$\rho_{\mathrm{tNFW}}(r) = \frac{\rho_{\mathrm{sub}}}{\frac{r}{r_{s,\mathrm{sub}}}\left(1+\frac{r}{r_{s,\mathrm{sub}}}\right)^2}\frac{r_t^2}{r^2+r_t^2}, \tag{A1}$$

where ρ_{sub} is the amplitude of the NFW density function in units of $M_\odot$ kpc^{-3}, r is the radial position in units of kpc, $r_{s,\mathrm{sub}}$ is the scale radius in units of kpc, and r_t is the truncation radius in units of kpc. Both $r_{s,\mathrm{sub}}$ and ρ_{sub} are calculated from the mass, m_{sub}, and concentration, c_{sub}, of the subhalo. The mass is drawn from Equation (6), and the concentration is drawn from the mass–concentration relation presented in Gilman et al. (2020b),

$$c_{\mathrm{sub}}(m, z) = c_0(1+z)^{\zeta}\left(\frac{\nu(r_{\mathrm{peak}}(m_{\mathrm{sub}}), z_{\mathrm{sub}})}{\nu(r_{\mathrm{peak}}(m_{\mathrm{pivot,conc}}), 0)}\right)^{-\beta}, \tag{A2}$$

where c_0 is the normalization of the mass–concentration relation, ζ is the redshift power-law slope, β is the peak height power-law slope, and $m_{\mathrm{pivot,conc}}$ is the mass–concentration pivot mass. The function ν is the peak height function (Doroshkevich 1970; Peebles 1980; see Mo et al. 2010 for the equations) at the subhalo redshift z_{sub} for peak radius $r_{\mathrm{peak}}(m_{\mathrm{sub}})$, defined by

$$r_{\mathrm{peak}}(m_{\mathrm{sub}}) = \left(\frac{3m_{\mathrm{sub}}}{4\pi\rho_{m,0}}\right)^{1/3}. \tag{A3}$$

Here $\rho_{m,0}$ is the matter density at redshift zero. To calculate the peak height, we use the power spectrum derived from the Eisenstein & Hu (1998) transfer function. In line with the literature (Dutton & Maccio 2014; Diemer & Kravtsov 2015; Diemer & Joyce 2019), we add an additional scatter, σ_{conc}, to Equation (A2). Given the mass and concentration, the scale radius and amplitude of the density function are given by

$$r_{s,\mathrm{sub}} = \frac{1}{c}\left(\frac{3\ m_{\mathrm{sub}}}{4\pi(200\rho_{\mathrm{crit}}(z_{\mathrm{sub}}))}\right)^{1/3}, \tag{A4}$$

$$\rho_{\mathrm{sub}} = \frac{m_{\mathrm{sub}}}{4\pi r_{s,\mathrm{sub}}^3\left(\log(1+c)-\frac{c}{1+c}\right)}, \tag{A5}$$

where $\rho_{\mathrm{crit}}(z)$ is the critical density of the universe at the redshift z_{sub} of the subhalo. The truncation radius is determined by the subhalo's position in the host using

$$r_t = 1.4\left(\frac{m_{\mathrm{sub}}}{m_{\mathrm{pivot,trunc}}}\right)^{1/3}\left(\frac{r_{\mathrm{sub}}}{r_{\mathrm{pivot,trunc}}}\right)^{2/3}, \tag{A6}$$

where $m_{\mathrm{pivot,trunc}}$ is the truncation pivot mass in units of $M_\odot$, r_{sub} is the radial distance of the subhalo from the host center in units of kpc, and $r_{\mathrm{pivot,trunc}}$ is the truncation pivot radius in units of kpc.

Appendix B
Line-of-sight Mass Function

We make two modifications to the Sheth–Tormen halo mass function. The first is a scaling parameter, $\delta_{\rm los}$, that accounts for uncertainties in overall normalization of the line-of-sight mass function. We also include a contribution from the two-point halo correlation function $\xi_{2\ {\rm halo}}(r,\ m_{\rm host},\ z_{\rm host})$. This term accounts for the overdensity of halos relative to the mean matter density of the universe near a massive halo. On large scales, we can describe the two-point halo correlation correlation function as

$$\xi_{2\ {\rm halo}}(r,\ m_{\rm host},\ z_{\rm host}) = b(m_{\rm host},\ z)\xi_{\rm lin}(r,\ z_{\rm host}). \tag{B1}$$

Here $b(m_{\rm host},\ z)$ is a mass-dependent linear halo bias parameter (Kaiser 1984; Bardeen et al. 1986; Mo & White 1996; Jing 1998; see Desjacques et al. 2018 for a review) that must be included because dark matter halos are biased tracers of the underlying distribution of matter. The dependence of the bias parameter on the host mass takes into account high-mass halos being comparatively more likely in overdense regions and low-mass halos being comparatively more likely in underdense regions. In this work, we use the bias model presented in Tinker et al. (2010). The remaining term, $\xi_{\rm lin}(r,\ z_{\rm host})$, represents the linear matter–matter correlation function at redshift $z_{\rm host}$. We use the correlation function derived from the Eisenstein & Hu (1998) transfer function. This correlation function contribution is included for halos within the range $[r_{\rm 2halo,min},\ r_{\rm 2halo,max}]$ of the host. Recent work has argued that lensing can produce clustering that exceeds the two-point halo correlation (Lazar et al. 2021), but we do not model that additional signal here.

With everything included, the line-of-sight halo mass function is given by the equation

$$\frac{d^2N_{\rm los}}{dV\ dm_{\rm los}} = \delta_{\rm los}(1 + \xi_{2\ {\rm halo}}(r,\ m_{\rm host},\ z_{\rm host}))\left[\frac{d^2N_{\rm los}}{dV\ dm_{\rm los}}\right]_{\rm ST}, \tag{B2}$$

where the Sheth–Tormen halo mass function is given by

$$\left[\frac{d^2N_{\rm los}}{dV\ dm_{\rm los}}\right]_{\rm ST} = -\frac{1}{3}\frac{\nu f(\nu)}{m^2}\frac{d\log\sigma(r_{\rm peak})}{d\log r_{\rm peak}}\rho_m. \tag{B3}$$

The definition of the radius $r_{\rm peak}$ is given in Appendix A, with the dependence on $m_{\rm los}$ and $z_{\rm los}$ left implicit in Equation (B3) for conciseness. The peak height, ν, is related to the rms variance of the linear density field $\sigma(r_{\rm peak})$ by

$$\nu(r_{\rm peak}(m_{\rm los}),\ z_{\rm los}) = \frac{\delta_c(z_{\rm los})}{\sigma(r_{\rm peak}(m_{\rm los}),\ z_{\rm los})}, \tag{B4}$$

with $\delta_c(z_{\rm los})$ being the linear overdensity threshold for halo collapse at redshift $z_{\rm los}$ (see Mo et al. 2010 for a derivation). The functional form of $\nu f(\nu)$ is given in Sheth et al. (2001),

$$\nu f(\nu) = 2A_{\rm ST}\left(1 + \frac{1}{\nu'^{2q_{\rm ST}}}\right)\frac{\nu'}{(2\pi)^{1/2}}\exp\left(-\frac{\nu'^2}{2}\right), \tag{B5}$$

$$\nu' = \sqrt{a_{\rm ST}}\,\nu, \tag{B6}$$

with $A_{\rm ST} = 0.32218$, $q_{\rm ST} = 0.3$, and $a_{\rm ST} = 0.707$. The Sheth–Tormen halo mass function is nontrivial to draw from; however, for a relatively small mass range, it can be well approximated by a power law. Therefore, in practice, we draw from a power law with normalization and slope set by minimizing the log squared distance to the Sheth–Tormen halo mass function in the range $[m_{\rm min,los},\ m_{\rm max,los}]$. The parameters $m_{\rm min,los}$ and $m_{\rm max,los}$ also set the minimum and maximum line-of-sight halo mass that will be rendered.

Appendix C
COSMOS Images

A more detailed summary of the GREAT3 data set we use can be found in Mandelbaum et al. (2014). The important points for our use are as follows.

1. The COSMOS images have been processed using MULTIDRIZZLE (see Fruchter & Hook 2002 for a summary of the drizzle algorithm and Gonzaga et al. 2012 for a discussion of a modern implementation of the full pipeline). This pipeline deals with the geometric distortion, sky subtraction, and cosmic-ray rejection. As part of this process, it combines several exposures dithered at subpixel intervals. It is therefore capable of returning a smaller pixel scale in the final coadded images, which for the COSMOS patch is set to 0.″03.
2. Within the COSMOS patch, the sources are selected using the strategy outlined in Leauthaud et al. (2007). On top of the selection cuts described there, there are additional selection cuts made to reject nongalaxy objects, galaxies with imaging defects, and galaxies that do not have a reliable photometric redshift. All objects with an F814W magnitude above 25.2 are also removed.
3. The cutout for each source is placed at the estimated source center and extends to approximately five times the half-light radius of the galaxy (the exact formula can be found in Mandelbaum et al. 2012). An additional masking and deconvolution step is applied to the images. The deconvolution takes advantage of the TinyTim PSF estimates (Krist et al. 2011), and the masked pixels are replaced with correlated noise.

Appendix D
DRIZZLE Pipeline

To capture the effects of the DRIZZLEPAC pipeline, PALTAS conducts the following procedure.

1. Using LENSTRONOMY, a supersampled ray-tracing image is generated at twice the resolution of the detector (0.″02 pixel scale). This version of the image does not include the detector noise or PSF and is simulated in the sky plane.
2. Using ASTROPY,[28] the supersampled image is mapped to four dithered detector images with half-pixel offsets (0.″04 pixel scale), corresponding to the standard four-point dithering strategy (Gonzaga et al. 2012). This is done through the World Coordinate System (WCS) of the supersampled image and the dithered images, allowing for geometric distortion coefficients to be included in the mapping. However, in this work, we do not include any geometric distortion coefficients.

[28] https://www.astropy.org/

3. These four dithered images each represent the detector output for one exposure. Therefore, we convolve these dithered images with the empirical PSF model and add the expected sky and detector noise.
4. Using the DRIZZLE package, we drizzle these four images onto the output WCS with the corresponding output 0.″03 pixel scale.

Appendix E Examples of Subhalo Signal

Before training a network to predict the SHMF, it is important to develop intuition for how varying the SHMF impacts the signal in our image. This will help us interpret our network's outputs in Section 4.1 and the results of our hierarchical analysis in Section 4.2. To do this, we simulate a set of lens images for different SHMF realizations and model the potential of their main deflector through forward modeling. We then visualize the remaining "residual" lensing signal that cannot be described by the smooth potential of the main deflector. The strength of this residual signal is a direct indicator of the subhalos present in the lensing field. To keep the forward modeling tractable for this example, we make a number of simplifications in this section. We keep the full complexity of the main deflector and subhalos, but we do not render the line-of-sight halos or use the full drizzling and empirical PSF process. Rather, we assume an angular resolution of 0.″04 and a Gaussian PSF with an FWHM of 0.″04. We also use a COSMOS source drawn from the same process described in Section 2.4, but we assume full knowledge of the source in our simplified model fitting. In the more realistic scenario we tackle in Section 4, the PSF, source uncertainty, and line-of-sight halos will all reduce the strength of the residual subhalo lensing signal. Regardless, for this toy problem, separating out those uncertainties will help focus on the effects of the subhalos.

In Figure 9, we show the residual signal generated by three different SHMFs. All three SHMFs follow the parameterization outlined in Section 2.2 but with different values for the SHMF normalization and slope. The SHMFs 1, 2, and 3 have normalizations of $\Sigma_{\rm sub} = 2.4 \times 10^{-4}$, 1.2×10^{-3}, and 1.2×10^{-3} (kpc)$^{-2}$ and slopes of $\gamma_{\rm sub} = -1.83$, -1.83, and -2.00, respectively. We show the residual plots for these three SHMFs and the case of no subhalos. As expected, without subhalos and with perfect knowledge of the source light, the entire lensing signal can be described by our best-fitting main deflector model. As we introduce a small number of subhalos in SHMF 1, we see a small residual signal emerge that cannot be described by our smooth main deflector model (χ_ν^2: −1.02); however, that signal is nearly subdominant to even the simplified noise model we are using in this toy example. With a factor of 5 larger normalization than SHMF 1, SHMF 2 produces a much stronger residual signal that can be visually identified over the noise of the instrument (χ_ν^2: −1.11). Finally, increasing the steepness of our SHMF, as we do for SHMF 3, introduces many more low-mass subhalos and leads to an even stronger residual signal (χ_ν^2: −1.20). While the simulations in Figure 9 are simplified, they suggest that for large enough normalizations and slopes, the presence of subhalos in our simulated images should be detectable by our modeling tools. Additionally, we should be able to distinguish between SHMFs

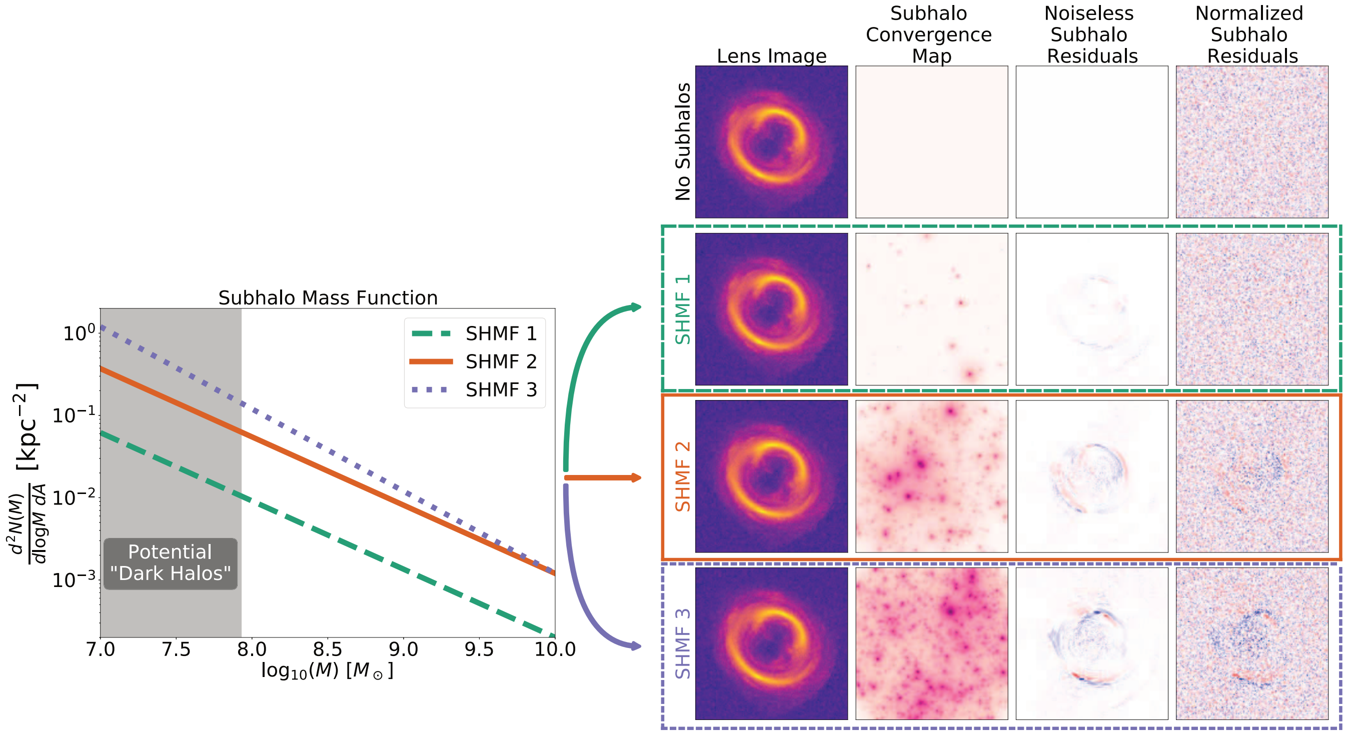

Figure 9. Comparison of the signal imprinted by subhalos for different possible SHMFs. The left panel plots the number of subhalos per unit area as a function of mass for the three SHMFs. The right panel shows how subhalos drawn from these three functions impact the lens image (first column), the convergence induced by subhalos (second column), the noiseless residual signal over a smooth main deflector (third column), and the residual signal when normalized by the noise (fourth column). Even a small number of subhalos is enough to impart some residual signal over the smooth main deflector model; however, that signal is subdominant to the noise of the observation. As we increase the slope and normalization of the SHMF, we get more subhalo convergence and therefore a larger residual signal. However, the detectable signal is localized to the brightest regions of the lens.

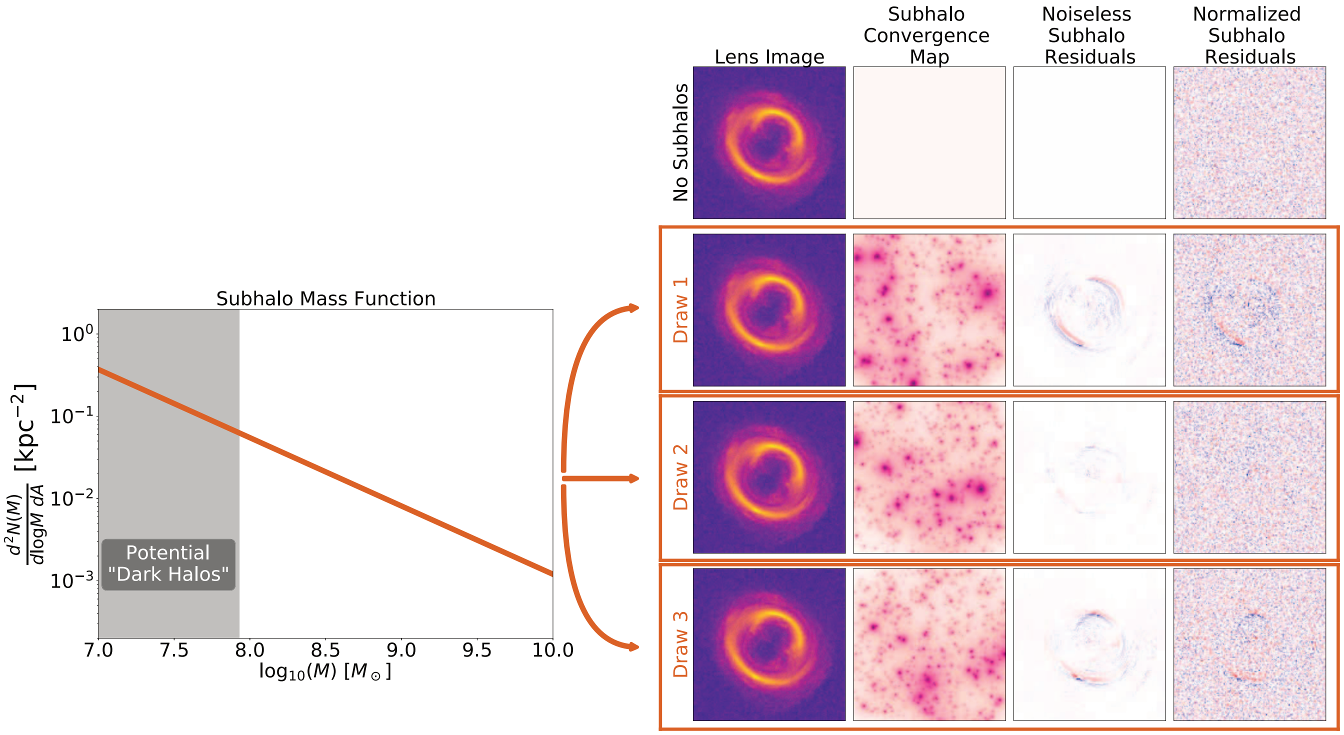

Figure 10. Same as Figure 9 but with several draws from the same underlying SHMF. Even for a fixed SHMF, the residual signal can vary drastically depending on how many subhalos overlap with the brightest parts of the image.

with larger slopes and normalizations based on the strength of this residual signal.

Figure 9 shows only one draw per SHMF. If instead we draw multiple times from the same SHMF, as we do for SHMF 2 in Figure 10, a very different picture emerges. With a fixed amplitude and slope, the convergence maps that are produced look very similar. However, the exact position of those subhalos with respect to the path of the lensed light leads to very different residual signals. Draw 1 returns a strong residual signal (χ^2_ν: −1.16), almost comparable with that of SHMF 3 in Figure 9; draw 2 returns almost no residual signal (χ^2_ν: −1.05), comparable with SHMF 1 in Figure 9; and draw 3 is the only one that returns a signal that looks equivalent to what we produced for SHMF 2 in Figure 9 (χ^2_ν: −1.12). The stochastic nature of the residual signal suggests that any modeling technique will return a large uncertainty on our SHMF parameters for an individual lens. This uncertainty is a consequence of attempting to constrain the SHMF, which is a statistical summary of the subhalos, rather than attempting to constrain the thousands of deterministic parameters that describe the subhalos. While the stochastic nature of the residual signal may limit what we can say about an individual lens, it may also allow us to improve our sensitivity to SHMF with very few subhalos. For example, on average, draws from SHMF 1 produce residual signals that are subdominant to the noise. But occasionally, a draw from SHMF 1 will return a subhalo population that produces a detectable residual signal and is therefore inconsistent with having no subhalos. Hierarchically combining these constraints could therefore allow us to make a meaningful detection of SHMF 1.

We will explore the signal we can extract for a more realistic population of lenses in Section 4.2. For now, this toy example leaves us with two takeaways.

1. The stochastic nature of the residual signal for a fixed SHMF means that we should expect large uncertainties on constraints derived from individual lenses.
2. These large uncertainties will require us to conduct hierarchical inference on a population of lenses in order to accurately constrain the SHMF. For more details on this hierarchical inference, see Section 3.3.

Appendix F
Data Set Parameters

For the simulated data sets we use in this work, a number of the parameters associated with the resolution of the simulation are kept fixed. These include the line-of-sight redshift bin widths, the mass boundaries for our subhalos and line-of-sight halos, and the opening angle of our cone. We also assume that the size and intensity of the COSMOS images are large enough to ignore the detector noise in our source model. We test the effect of these choices by generating test data sets with the same parameter distributions as our training data set but with modified resolution parameters. In particular, we generate one test data set where we decrease the minimum subhalo and line-of-sight halo mass rendered to $5 \times 10^6\,M_\odot$, one test data set where the minimum mass is increased to $5 \times 10^7\,M_\odot$, one test data set where we increase the opening angle of our cone to 10°, one test data set where we smooth the source images by a Gaussian PSF with an FWHM of 0.″05, and one test data set where we increase the line-of-sight bin width to 0.02 in redshift space. We run the same hierarchical inference pipeline as in Section 4.2 to see if any of these changes in the simulation choices affect the accuracy of our inference. We only consider 10, 50, and 100 lens analyses, since we have found that the 1000 lens analysis can have systematic bias. The results are found in Figure 11 and show that none of the changes in the

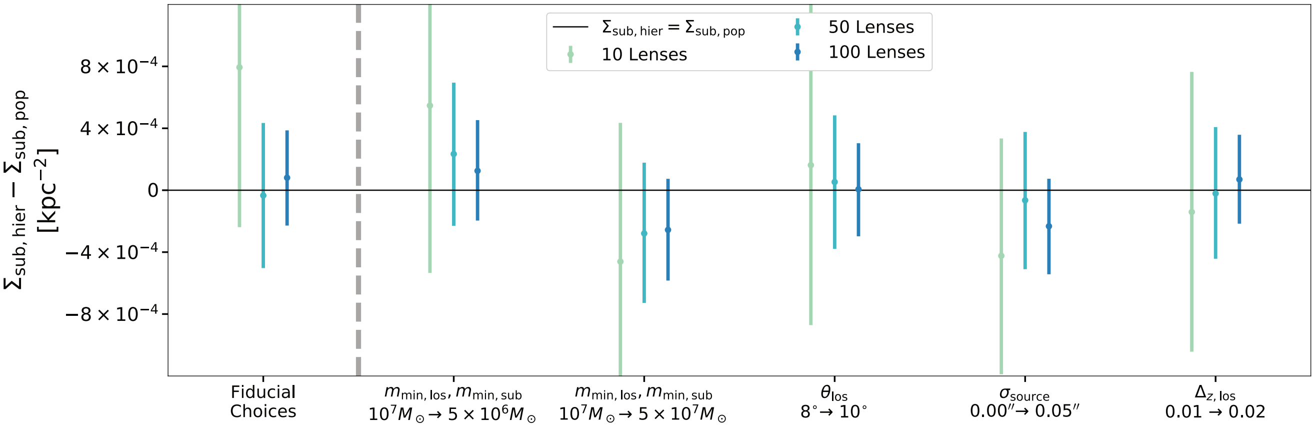

Figure 11. Difference between the inferred population mean, $\Sigma_{\rm sub,hier}$, and the true population mean, $\Sigma_{\rm sub,pop}$, for six test sets. The "Fiducial Choices" test set has the resolution parameters set to the values used throughout this work. The remaining five test sets each change one resolution assumption. The error bars on the inferred $\Sigma_{\rm sub,hier}$ show the 68% interval derived from the hierarchical inference pipeline. All six test sets have a value of $\Sigma_{\rm sub,pop} = 2 \times 10^{-3}$ kpc^{-2}.

resolution choices produce an inferred population mean more than 1σ from the true population mean. Each test data set is statistically independent, so fluctuations within the inference uncertainties are to be expected. This suggests that the results of the analysis in this paper are not significantly impacted by our resolution choices.

Additionally, for this work, we assume a fixed redshift for the source and main deflector. We also assume a fixed main halo mass $m_{\rm host}$. The main halo mass is independent of the lensing parameters of the main deflector, and we have absorbed the dependence of the SHMF on host mass into our definition of $\Sigma_{\rm sub}$ (see Section 2.2). Therefore, the main impact of this choice in our simulation is to fix the contribution from our two-halo term (see Section 2.3).

Appendix G
Further Hierarchical Inference Tests for the Einstein Radius

As we discuss in Section 3.1, the test sets have an underlying distribution for the main deflector parameters that is shifted from the training set. The purpose of this shift is to demonstrate that our final inference on the SHMF normalization is not biased when the training set assumptions are violated. This is particularly important for the Einstein radius, since Figure 4 shows that the network outputs a strong correlation between the inferred Einstein radius and the inferred SHMF normalization. In order to further test our dependence on the assumed distribution of Einstein radii, we have designed five additional test sets that have identical distributions to the training set except for the Einstein radius, which is given a fixed mean value for all lenses. We vary this fixed value, $\theta_{{\rm E},\mu}$, to span the 2σ contours of the training set prior. The inferred SHMF normalization can be seen in Figure 12. We find no evidence of systematic bias in the inferred mean SHMF normalization as we vary the Einstein radius of the lens population. If our network were only sensitive to the Einstein radius and not the underlying signal of low-mass halos, we would not be able to reconstruct an accurate and precise mean for the SHMF normalization on these test sets. Therefore, we conclude that the information our network is extracting from the images goes beyond the observed radius of the Einstein ring.

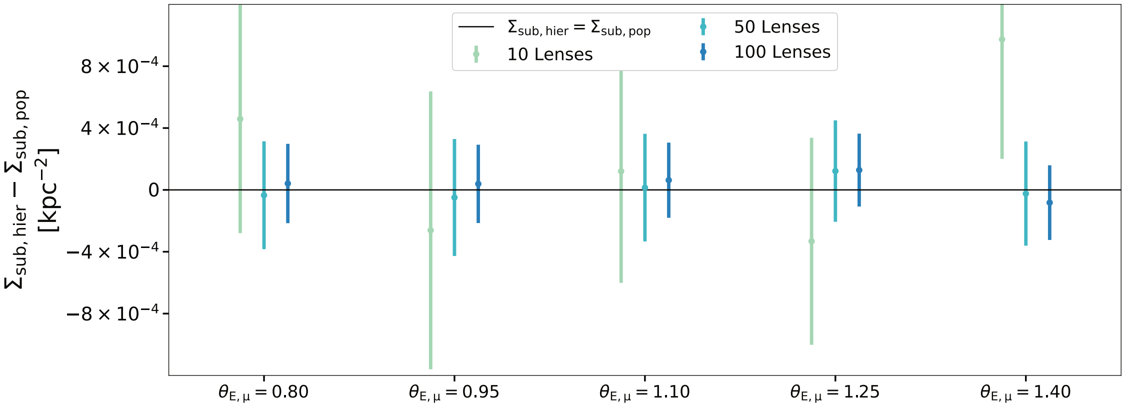

Figure 12. Difference between the inferred population mean, $\Sigma_{\rm sub,hier}$, and the true population mean, $\Sigma_{\rm sub,pop}$, for five test sets with varying fixed Einstein radius, $\theta_{{\rm E},\mu}$. Only the distribution of $\theta_{\rm E}$ has been changed, and all other parameters are distributed as in the training set. The error bars on the inferred $\Sigma_{\rm sub,hier}$ show the 68% interval derived from the hierarchical inference pipeline. All five test sets have a value of $\Sigma_{\rm sub,pop} = 2.0 \times 10^{-3}$ kpc^{-2}.

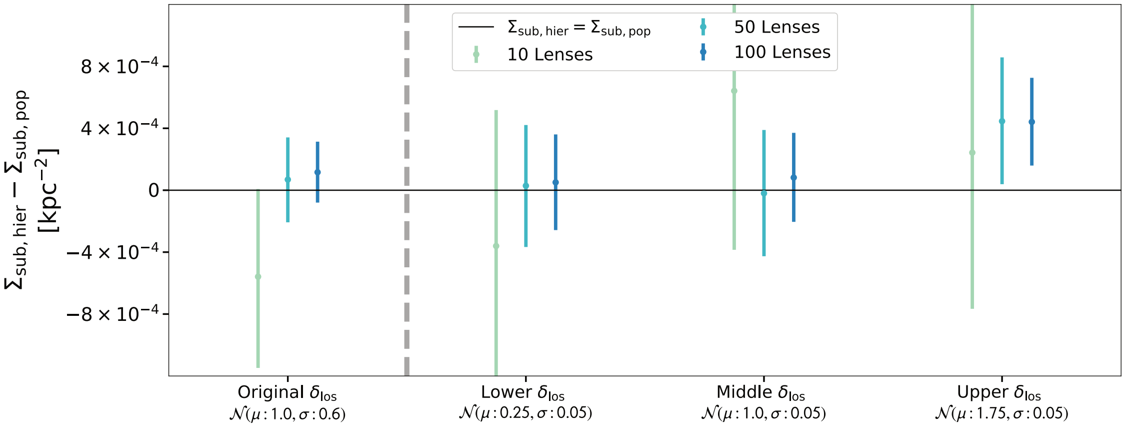

Figure 13. Difference between the inferred population mean, $\Sigma_{\rm sub,hier}$, and the true population mean, $\Sigma_{\rm sub,pop}$, for four test sets with varying distributions of the line-of-sight mass function normalization, $\delta_{\rm los}$. The "Original" test set corresponds to test set 7 from the main text. For the three test sets to the right of the gray dashed line, only the distribution of $\delta_{\rm los}$ has been changed, and all other parameters are distributed as in test set 7. The error bars on the inferred $\Sigma_{\rm sub,hier}$ show the 68% interval derived from the hierarchical inference pipeline. All four test sets have a value of $\Sigma_{\rm sub,pop} = 1.4 \times 10^{-3}$ kpc^{-2}.

Appendix H
Line-of-sight Halo Normalization

Throughout this work, we have made a fixed assumption about the distribution of line-of-sight halo normalizations, $\delta_{\rm los}$. As we discuss in more detail in Section 2.3, previous studies have suggested that the line-of-sight halos can produce deflections that are comparable to the subhalos. To better understand how constraining our assumed line-of-sight halo distribution is to our inference, we generate three test sets with much narrower line-of-sight distributions that span the "lower," "middle," and "upper" regions of the training distribution. This corresponds to a distribution of $\delta_{\rm los} \sim \mathcal{N}(\mu$: 0.25, σ: 0.05), $\mathcal{N}(\mu$: 1.0, σ: 0.05), and $\mathcal{N}(\mu$: 1.75, σ: 0.05), respectively. The goal of these test sets is to probe how a systematic error in the average line of sight translates to a bias in the inferred SHMF normalization, $\Sigma_{\rm sub,hier}$. To help make a direct comparison, we fix all of the other parameters for these test sets to the same values used for training set 7 (see Section 3.1 for more details). As with the tests in Appendix F, we only consider 10, 50, and 100 lens analyses. The results of the hierarchical inference are shown in Figure 13. For the lower and middle test sets, the change in the line-of-sight distribution does not shift the inferred mean more than 1σ from the true population mean. Only for the upper test set is the shift more than 1σ, and the final result is still within the 95% confidence interval. We conclude that, for the shifts in the $\delta_{\rm los}$ distributions we explore here, the induced bias in $\Sigma_{\rm sub,hier}$ is subdominant to the uncertainties for 10, 50, and 100 lenses.

Appendix I
Validation Predictions for the Remaining Parameters

In Figure 14, we present the predicted mean values as a function of the true values for the lens parameters not shown in Figure 5. The points shown here span the entire validation set. As with the Einstein radius, the network's predictions for the remaining main deflector parameters are strongly correlated with the truth ($\rho >= 0.98$).

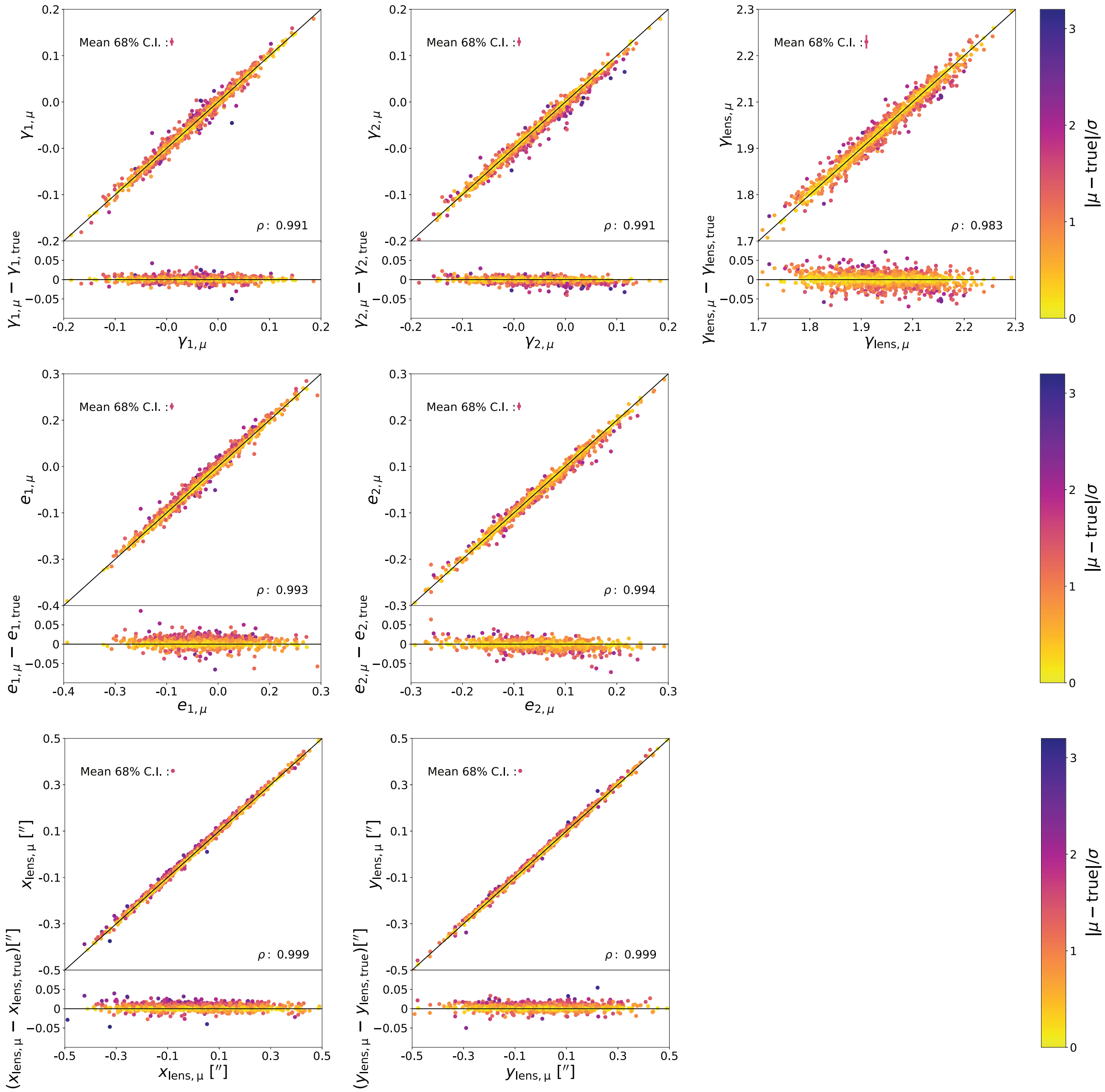

Figure 14. Comparison of the network predicted means and the true input values for all of the lenses in our validation set. The panels show all of the parameters not covered in Figure 5. Going left to right and then top to bottom, these are the x-direction shear, γ_1; xy-direction shear, γ_2; power-law slope, γ_{lens}; x-direction ellipticity eccentricity, e_1; xy-direction ellipticity eccentricity, e_2; x-coordinate lens center, x_{lens}; and y-coordinate lens center, y_{lens}. For all of the plots, one point corresponds to the predictions of one lens. The color of the point is defined by the distance between the predicted mean and the truth in units of the predicted standard deviation. The color bar for all of the panels is the same and included on the far right of each row. The upper left region of each panel includes a visualization of the average standard deviation being predicted by the network (equivalent to the 68% confidence interval). The ρ value printed in the panels represents the Pearson correlation coefficient between the predicted mean and the truth.

ORCID iDs

Sebastian Wagner-Carena https://orcid.org/0000-0001-5039-1685
Simon Birrer https://orcid.org/0000-0003-3195-5507
Ethan O. Nadler https://orcid.org/0000-0002-1182-3825
Elise Darragh-Ford https://orcid.org/0000-0002-8800-5652
Risa H. Wechsler https://orcid.org/0000-0003-2229-011X

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, arXiv:1603.04467

Baltz, E. A., Marshall, P., & Oguri, M. 2009, JCAP, 2009, 015

Banik, N., Bovy, J., Bertone, G., Erkal, D., & de Boer, T. 2021a, MNRAS, 502, 2364

Banik, N., Bovy, J., Bertone, G., Erkal, D., & de Boer, T. 2021b, JCAP, 2021, 043

Bardeen, J. M., Bond, J. R., Kaiser, N., & Szalay, A. S. 1986, ApJ, 304, 15

Barkana, R. 1998, ApJ, 502, 531

Beaumont, M. A., Zhang, W., & Balding, D. J. 2002, Genetics, 162, 2025

Benson, A. J. 2020, MNRAS, 493, 1268

Birrer, S., & Amara, A. 2018, PDU, 22, 189

Birrer, S., Amara, A., & Refregier, A. 2017b, JCAP, 2017, 037

Birrer, S., Shajib, A. J., Gilman, D., et al. 2021, JOSS, 6, 3283

Birrer, S., Welschen, C., Amara, A., & Refregier, A. 2017a, JCAP, 2017, 049

Bode, P., Ostriker, J. P., & Turok, N. 2001, ApJ, 556, 93

Bonaca, A., Hogg, D. W., Price-Whelan, A. M., & Conroy, C. 2019, ApJ, 880, 38

Brehmer, J., Mishra-Sharma, S., Hermans, J., Louppe, G., & Cranmer, K. 2019, ApJ, 886, 49

Brennan, S., Benson, A. J., Cyr-Racine, F.-Y., et al. 2019, MNRAS, 488, 5085

Brewer, B. J., Huijser, D., & Lewis, G. F. 2016, MNRAS, 455, 1819

Buckley, M. R., & Peter, A. H. 2018, PhR, 761, 1

Bullock, J. S., & Boylan-Kolchin, M. 2017, ARA&A, 55, 343

Bullock, J. S., Kolatt, T. S., Sigad, Y., et al. 2001, MNRAS, 321, 559

Calamida, A., Mack, J., Medina, J., et al. 2021, Space Telescope WFC Instrument Science Report, 4

Çağan Şengül, A., Dvorkin, C., Ostdiek, B., & Tsang, A. 2022, MNRAS, 515, 4391

Collett, T. E. 2015, ApJ, 811, 20

Coogan, A., Karchev, K., & Weniger, C. 2020, arXiv:2010.07032

Cyr-Racine, F.-Y., Moustakas, L. A., Keeton, C. R., Sigurdson, K., & Gilman, D. A. 2016, PhRvD, 94, 043505

Dalal, N., & Kochanek, C. S. 2002, ApJ, 572, 25

Dauphin, F., Anderson, J., Bajaj, V., et al. 2021, Space Telescope WFC Instrument Science Report, 12

Daylan, T., Cyr-Racine, F.-Y., Rivero, A. D., Dvorkin, C., & Finkbeiner, D. P. 2018, ApJ, 854, 141

Dekker, A., Ando, S., Correa, C. A., & Ng, K. C. 2022, PhRvD, 106, 123026

Desjacques, V., Jeong, D., & Schmidt, F. 2018, PhR, 733, 1

Despali, G., Vegetti, S., White, S. D. M., Giocoli, C., & van den Bosch, F. C. 2018, MNRAS, 475, 5424

Diaz Rivero, A., & Dvorkin, C. 2020, PhRvD, 101, 023515

Diemer, B. 2018, ApJS, 239, 35

Diemer, B., & Joyce, M. 2019, ApJ, 871, 168

Diemer, B., & Kravtsov, A. V. 2015, ApJ, 799, 108

Ding, X., Treu, T., Shajib, A. J., et al. 2018, arXiv:1801.01506

Doroshkevich, A. 1970, Ap, 6, 320

Dressel, L. 2021, Wide Field Camera 3 Instrument Handbook, v13.0 (Baltimore, MD: STScI), https://www.stsci.edu/files/live/sites/www/files/home/hst/documentation/_documents/wfc3/wfc3_ihb_cycle29.pdf

Dutton, A. A., & Maccio, A. V. 2014, MNRAS, 441, 3359

Eisenstein, D. J., & Hu, W. 1998, ApJ, 496, 605

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306

Fruchter, A. S., & Hook, R. N. 2002, PASP, 114, 144

Gennaro, M. 2018, WFC3 Data Handbook, v4.0 (Baltimore, MD: STScI), https://www.stsci.edu/files/live/sites/www/files/home/hst/documentation/_documents/wfc3/wfc3_dhb_v4.pdf

Gilman, D., Benson, A., Bovy, J., et al. 2022, MNRAS, 512, 3163

Gilman, D., Birrer, S., Nierenberg, A., et al. 2020a, MNRAS, 491, 6077

Gilman, D., Du, X., Benson, A., et al. 2020b, MNRAS, 492, L12

Gonzaga, S., Hack, W., Fruchter, A., & Mack, J. 2012, The DrizzlePac Handbook (Baltimore, MD: STScI)

Goodman, J., & Weare, J. 2010, CAMCS, 5, 65

Green, A. M., Hofmann, S., & Schwarz, D. J. 2004, MNRAS, 353, L23

Greenberg, D., Nonnenmacher, M., & Macke, J. 2019, in Proc. the 36th Int. Conf. on Machine Learning, 97, 2404, https://proceedings.mlr.press/v97/greenberg19a.html

He, K., Zhang, X., Ren, S., & Sun, J. 2016, in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Red Hook, NY: Currant Associates Inc.), 770, https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

He, T., Zhang, Z., Zhang, H., et al. 2019, in IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (Red Hook, NY: Currant Associates Inc.), 558, https://openaccess.thecvf.com/content_CVPR_2019/html/He_Bag_of_Tricks_for_Image_Classification_with_Convolutional_Neural_Networks_CVPR_2019_paper.html

Hezaveh, Y., Dalal, N., Holder, G., et al. 2013a, ApJ, 767, 9

Hezaveh, Y., Dalal, N., Holder, G., et al. 2016b, JCAP, 2016, 048

Hezaveh, Y., Marrone, D. P., Fassnacht, C., et al. 2013b, ApJ, 767, 132

Hezaveh, Y. D., Dalal, N., Marrone, D. P., et al. 2016a, ApJ, 823, 37

Iršič, V., Viel, M., Haehnelt, M. G., et al. 2017b, PhRvD, 96, 023522

Iršič, V., Viel, M., Haehnelt, M. G., Bolton, J. S., & Becker, G. D. 2017a, PhRvL, 119, 031302

Jing, Y. P. 1998, ApJL, 503, L9

Kaiser, N. 1984, ApJL, 284, L9

Kaplinghat, M. 2005, PhRvD, 72, 063510

Keeton, C. R., Kochanek, C. S., & Seljak, U. 1997, ApJ, 482, 604

Kennedy, R., Frenk, C., Cole, S., & Benson, A. 2014, MNRAS, 442, 2487

Koekemoer, A. M., Aussel, H., Calzetti, D., et al. 2007, ApJS, 172, 196

Koopmans, L. V. E. 2005, MNRAS, 363, 1136

Kormann, R., Schneider, P., & Bartelmann, M. 1994, A&A, 284, 285

Krist, J. E., Hook, R. N., & Stoehr, F. 2011, Proc. SPIE, 8127, 81270J

Lazar, A., Bullock, J. S., Boylan-Kolchin, M., et al. 2021, MNRAS, 502, 6064

Leauthaud, A., Massey, R., Kneib, J.-P., et al. 2007, ApJS, 172, 219

Lin, J.-Y.-Y., Yu, H., Morningstar, W., Peng, J., & Holder, G. 2020, arXiv:2010.12960

Ludlow, A. D., Bose, S., Angulo, R. E., et al. 2016, MNRAS, 460, 1214

Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., et al. 2017, arXiv:1711.01861

Macciò, A. V., & Fontanot, F. 2010, MNRAS, 404, L16

Mandelbaum, R., Hirata, C. M., Leauthaud, A., Massey, R. J., & Rhodes, J. 2012, MNRAS, 420, 1518

Mandelbaum, R., Rowe, B., Bosch, J., et al. 2014, ApJS, 212, 5

Mao, S., & Schneider, P. 1998, MNRAS, 295, 587

Mao, Y.-Y., Zentner, A. R., & Wechsler, R. H. 2018, MNRAS, 474, 5143

Menci, N., Sanchez, N. G., Castellano, M., & Grazian, A. 2016, ApJ, 818, 90

Mo, H., Van den Bosch, F., & White, S. 2010, Galaxy Formation and Evolution (Cambridge: Cambridge Univ. Press)

Mo, H. J., & White, S. D. M. 1996, MNRAS, 282, 347

Mohamed, S., & Lakshminarayanan, B. 2016, arXiv:1610.03483

Moore, B., Ghigna, S., Governato, F., et al. 1999, ApJL, 524, L19

Moustakas, L. A., & Metcalf, R. B. 2003, MNRAS, 339, 607

Nadler, E., Drlica-Wagner, A., Bechtol, K., et al. 2021a, PhRvL, 126, 091101

Nadler, E. O., Birrer, S., Gilman, D., et al. 2021b, ApJ, 917, 7

Nadler, E. O., Gluscevic, V., Boddy, K. K., & Wechsler, R. H. 2019, ApJL, 878, L32

Nadler, E. O., Wechsler, R. H., Bechtol, K., et al. 2020, ApJ, 893, 48

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493

Navarro, J. F. 1996, IAU Symp. 171, The Structure of Cold Dark Matter Halos (Cambridge: Cambridge Univ. Press), 255

Newton, O., Leo, M., Cautun, M., et al. 2021, JCAP, 2021, 062

Ostdiek, B., Rivero, A. D., & Dvorkin, C. 2022a, A&A, 657, L14

Ostdiek, B., Rivero, A. D., & Dvorkin, C. 2022b, ApJ, 927, 83

Papamakarios, G., & Murray, I. 2016, Advances in Neural Information Processing Systems, 29 (NIPS 2016), ed. D. Lee et al. (Red Hook, NY: Curran Associates Inc.), 1036, http://papers.nips.cc/paper/6084-fast-free-inference-of-simulation-models-with-bayesian-conditional-density-estimation

Papamakarios, G., Sterratt, D., & Murray, I. 2019, The 22nd Int. Conf. on Artificial Intelligence and Statistics, 89, ed. K. Chaudhuri & M. Sugiyama,, 837, https://proceedings.mlr.press/v89/papamakarios19a.html

Park, J. W., Wagner-Carena, S., Birrer, S., et al. 2021, ApJ, 910, 39

Pearson, J., Maresca, J., Li, N., & Dye, S. 2021, MNRAS, 505, 4362

Peebles, P. J. E. 1980, The Large-scale Structure of the Universe, Vol. 98 (Princeton, NJ: Princeton Univ. Press)

Perreault Levasseur, L., Hezaveh, Y. D., & Wechsler, R. H. 2017, ApJL, 850, L7

Planck Collaboration 2020, A&A, 641, A6

Prada, F., Klypin, A. A., Cuesta, A. J., Betancort-Rijo, J. E., & Primack, J. 2012, MNRAS, 423, 3018

Rhodes, J. D., Massey, R. J., Albert, J., et al. 2007, ApJS, 172, 203

Rivero, A. D., Cyr-Racine, F.-Y., & Dvorkin, C. 2018b, PhRvD, 97, 023001

Rivero, A. D., Dvorkin, C., Cyr-Racine, F.-Y., Zavala, J., & Vogelsberger, M. 2018a, PhRvD, 98, 103517

Rogers, K. K., & Peiris, H. V. 2021, PhRvL, 126, 071302

Rowe, B. T. P., Jarvis, M., Mandelbaum, R., et al. 2015, A&C, 10, 121

Rubin, D. B. 1984, Ann. Statist., 12, 1151

Rudakovskyi, A., Mesinger, A., Savchenko, D., & Gillet, N. 2021, MNRAS, 507, 3046

Ryon, J. E. 2021, Acs Instrument Handbook, v20.0 (Baltimore, MD: StScI)

Sheth, R. K., Mo, H. J., & Tormen, G. 2001, MNRAS, 323, 1

Sonnenfeld, A., Gavazzi, R., Suyu, S. H., Treu, T., & Marshall, P. J. 2013, ApJ, 777, 97

Tinker, J. L., Robertson, B. E., Kravtsov, A. V., et al. 2010, ApJ, 724, 878

Şengül, A. Ç., Tsang, A., Diaz Rivero, A., et al. 2020, PhRvD, 102, 063502

Tulin, S., & Yu, H.-B. 2018, PhR, 730, 1

Vegetti, S., Despali, G., Lovell, M. R., & Enzi, W. 2018, MNRAS, 481, 3661

Vegetti, S., & Koopmans, L. V. E. 2009, MNRAS, 392, 945

Vegetti, S., Koopmans, L. V. E., Bolton, A., Treu, T., & Gavazzi, R. 2010, MNRAS, 408, 1969

Vegetti, S., Lagattuta, D. J., McKean, J. P., et al. 2012, Natur, 481, 341

Viel, M., Becker, G. D., Bolton, J. S., & Haehnelt, M. G. 2013, PhRvD, 88, 043502

Wagner-Carena, S., Park, J. W., Birrer, S., et al. 2021, ApJ, 909, 187

Wang, J., Bose, S., Frenk, C. S., et al. 2020, Natur, 585, 39

White, M. 2001, A&A, 367, 27

White, S. D. M., & Rees, M. J. 1978, MNRAS, 183, 341