

Probing the nature of dark matter by forward modeling flux ratios in strong gravitational lenses

Daniel Gilman^{1*}, Simon Birrer¹, Tommaso Treu¹, Charles R. Keeton², Anna Nierenberg³

¹*Department of Physics and Astronomy, University of California, Los Angeles, CA 90095, USA*

²*Department of Physics and Astronomy, Rutgers, the State University of New Jersey, 136 Frelinghuysen Road, Piscataway, NJ 08854*

³*Department of Physics and Astronomy, University of California Irvine, Irvine CA, 92697-4575*

Accepted . Received

ABSTRACT

The free streaming length of dark matter particles determines the abundance of structure on sub-galactic scales. We present a statistical technique, amendable to any parameterization of subhalo density profile and mass function, to probe dark matter on these scales with quadrupole image lenses. We consider a warm dark matter particle with a mass function characterized by a normalization and free streaming scale m_{hm} . We forecast bounds on dark matter warmth for 120-180 lenses, attainable with future surveys, at typical lens (source) redshifts of 0.5 (1.5) in early-type galaxies with velocity dispersions of 220-270 km/sec. We demonstrate that limits on m_{hm} deteriorate rapidly with increasing uncertainty in image fluxes, underscoring the importance of precise measurements and accurate lens models. For our forecasts, we assume the deflectors in the lens sample do not exhibit complex morphologies, so we neglect systematic errors in their modeling. Omitting the additional signal from line of sight halos, our constraints underestimate the true power of the method. Assuming cold dark matter, for a low normalization, corresponding the destruction of all subhalos within the host scale radius, we forecast 2σ bounds on m_{hm} (thermal relic mass) of $10^{7.5}$ (5.0), 10^8 (3.6), and $10^{8.5}$ (2.7) $M_{\odot}$ (keV) for flux errors of 2%, 4%, and 8%. With a higher normalization, these constraints improve to $10^{7.2}$ (6.6), $10^{7.5}$ (5.3), and $10^{7.8}$ (4.3) $M_{\odot}$ (keV) with 120 systems. We are also able to measure the normalization of the mass function, which has implications for baryonic feedback models and tidal stripping.

Key words: [gravitational lensing: strong - cosmology: dark matter - galaxies: structure - methods: statistical]

1 INTRODUCTION

Dark matter models make testable predictions regarding the abundance and mass profiles of substructure in galactic dark matter halos. In the standard Λ CDM picture, structure grows bottom-up at practically all length scales (Schneider et al. 2013), resulting in a scale-free mass function for subhalos (Springel et al. 2008; Gao et al. 2012; Fiacconi et al. 2016), and density profiles fit by the Navarro Frenk and White (NFW) profile (Navarro et al. 1997). In contrast, in warm dark matter (WDM) models, free streaming washes out small density perturbations, resulting in a paucity of structure below a certain scale, which depends on the free streaming scale of the dark matter particle(s) (Schneider et al. 2012; Pullen et al. 2014; Viel et al. 2013;

Lovell et al. 2014a; Bose et al. 2016; Menci et al. 2016). Thermal relics and sterile neutrinos represent two WDM candidates (Kusenko 2009; Abazajian 2017), and under certain assumptions produce similar mass functions. Self-interacting dark matter alters the density profiles of individual subhalos, producing cores rather than cusps in the center of halos (Schneider et al. 2017; Vogelsberger et al. 2016; Kamada et al. 2017).

These alternative models to Λ CDM have gained traction, motivated by apparent failures of the Λ CDM picture on small scales (see Bullock & Boylan-Kolchin 2017). For example, by invoking WDM one explains the non-detection of low-mass satellites in the Milky Way, the “missing satellites problem” (Klypin et al. 1999), by reducing the expected number of subhalos. In another example, rotation curves of satellite galaxies imply shallower-than-isothermal inner den-

* gilmanda@ucla.edu

sity profiles (de Blok 2010), characteristic of the mass profiles associated with self interacting dark matter.

These apparent anomalies are based on observations of luminous structure and rely on assumptions about the connection between baryons and dark matter halos. Under different combinations of models for star formation and kinematic data in satellites, some small-scale challenges to the Λ CDM picture can be resolved (e.g. Kim et al. 2017). Unfortunately, it is difficult to measure kinematics of low mass galaxies owing to the small number of stars that can be used for this purpose. This leads to large uncertainties in the inferred halo masses and densities. Different models for tidal disruption and baryonic feedback (Despali & Vegetti 2017; Garrison-Kimmel et al. 2017), differing luminosity functions for dark subhalos (Nierenberg et al. 2016), and large scatter in stellar-mass-halo-mass relations below $10^8 M_\odot$ (Munshi et al. 2017) complicate constraints on the nature of dark matter. An independent and direct probe, which does not rely on assumptions regarding the physics of star formation in low mass galaxies, is needed to disentangle the physics of baryons and the nature of dark matter.

Gravitational lensing offers a direct probe of dark substructure below halo masses of $10^8 M_\odot$. Lensing relates a set of three observables - time delays, positions, and magnifications - to the gravitational potential along a path traversed by light emitted by a background source. As the observables depend only on the gravitational potential of the deflector and the potential along the line of sight, lensing offers a tool to study dark matter substructure directly, without relying on baryons as tracers.

Various techniques employ strong lensing as a probe of dark matter structure. When the light from a spatially extended background source is warped by a foreground deflector into a highly magnified arc, substructure can distort the arc. By simultaneously modeling the mass distribution in the lens plane and reconstructing the luminosity distribution of the background source, one can infer the mass of a perturber and constrain the subhalo mass function (Koopmans 2005; Vegetti et al. 2010; Vegetti et al. 2012, 2014; Li et al. 2016; Hezaveh et al. 2016b; Birrer et al. 2017; Vegetti et al. 2018).

Flux ratios in lensed quasars offer an alternative probe of dark substructure (Mao & Schneider 1998; Metcalf & Madau 2001; Dalal & Kochanek 2002; Chiba 2002). The sensitivity of this observable derives from the compact size of a quasar and the fact that lensing magnifications can be perturbed by a subhalo whose deflection angle is comparable to or larger than the source size (Dobler & Keeton 2006). From a theoretical standpoint, several forecasting studies (e.g. Xu et al. 2009, 2012; Graus et al. 2017) use N-body simulations to anticipate the lensing signal from CDM substructures. From an observational and lens modeling point of view, given observed flux ratios, one can add a subhalo to the lens model and vary its properties to infer the mass (e.g., Fadely & Keeton 2012; MacLeod et al. 2013; Nierenberg et al. 2014) or rule out the presence of substructure near the images (e.g., Nierenberg et al. 2017). This lens modeling technique, and the direct detection of subhalos via gravitational imaging, comprise a class of observations that yield constraints on individual substructures.

Recently, authors have called attention to a potential bias present in flux ratios, wherein the morphology of deflec-

tor in the lens plane produces flux ratio anomalies reminiscent of those induced by substructure. Gilman et al. (2017) showed that realistic deflectors with a luminous mass component drawn from HST images of nearby galaxies occasionally produce flux ratio anomalies with respect to a simplified smooth lens model. Hsueh et al. (2017a) performed a similar analysis with galaxies produced in the Illustris simulation (Vogelsberger et al. 2014), and reached similar conclusions. In two observed lenses, Hsueh et al. (2016b, 2017b) argued that a disk component in the deflector can explain observed flux ratio anomalies. These effects may contribute to frequency of observed flux ratio anomalies in strong lenses, which occur more frequently than one would expect in a CDM scenario, as pointed out by Xu et al. (2015). Among non-dark matter sources of flux ratio anomaly, microlensing can also induce drastic fluctuations in image magnifications, although this effect can be mitigated by using flux ratios measured from the more spatially extended narrow-line region of a background quasar (Moustakas & Metcalf 2003; Nierenberg et al. 2014, 2016).

In the context of dark matter, it is important to note that models predict large ensembles of dark subhalos, which could act together to affect a lensed image. In contrast to single subhalo models, other methods attempt to probe the collective impact of numerous perturbers whose individual effects are not statistically significant, but which together produce a measurable signal (Dalal & Kochanek 2002; Fadely & Keeton 2012; Hezaveh et al. 2016a; Cyr-Racine et al. 2016; Daylan et al. 2017; Chatterjee & Koopmans 2017; Birrer et al. 2017). Since these methods do not require high significance detections of individual perturbers, they extract information from a larger area around Einstein rings.

In an example, Birrer et al. (2017) quantify substructure in the lens RX J1131–1231 by modeling surface brightness anomalies detected in HST imaging data. Through a forward modeling approach that relies on generating an extensive suite of realistic simulations, they are able to constrain models of dark matter statistically. They could rule out WDM mass functions with thermal relics below the mass of 2 keV at 2σ .

In this work, we present a statistical method that utilizes the flux ratios from an ensemble of multiply imaged quasars to distinguish between dark matter models. Our technique takes as input data from a sample of strong lens systems and a prescription for rendering substructure realizations for a dark matter theory, and returns posterior probability distributions for the parameters describing the substructure population. We use the technique of Approximate Bayesian Computing (ABC; Rubin 1984), also applied in Birrer et al. (2017), which enables an application of Bayesian statistics to the problem of substructure lensing. In our framework, we are able to efficiently explore the parameter space spanned by dark matter models with different predictions regarding the nature of substructure without explicitly computing a likelihood function, which in substructure lensing is a computationally and analytically daunting task. ABC permits one to circumvent calculation of formidable likelihoods through the use of summary statistics, which quantify agreement between an observation and data computed in a forward model. Our method also naturally accommodates joint inference from multiple strong lens systems. The method can be applied to any parameteriza-

tion of dark substructure, provided one specifies the mass function, spatial distribution, and density profile of individual subhalos. Since the method relies on flux ratio statistics rendered in a forward model, accurate lens models and control over systematic errors in flux ratios are crucial for attaining robust constraints. Finally, by omitting line of sight substructure that is expected to contribute significantly to flux ratio anomalies (Chen et al. 2003; Inoue 2016; Despali et al. 2017), we do not capture the full information content of each lens, so our results can be interpreted as conservative theoretical bounds.

We present the formalism of our method and illustrate its general capabilities via a case study in which we distinguish between two simplified dark matter models. We consider a subhalo mass function with variable normalization and damping below a free-streaming scale, and provide forecasts for the constraints afforded under different flux precisions with up to 180 systems, which is a sample size attainable with future surveys such as Euclid, LSST and WFIRST (Oguri & Marshall 2010). With our forward modeling framework, we forecast the possible constraints on a WDM subhalo mass function using flux ratios, and quantify how these constraints scale with the number of lenses, the uncertainty in fluxes, and the overall normalization of the mass function.

This paper is structured as follows. Section 2 poses the problem of substructure lensing in a Bayesian framework and reviews the basics of Approximate Bayesian Computing. In section 3, we describe our parameterization for the subhalo mass function, our method for creating mock data sets, and the procedure to compute posterior distributions from the forward model. In Section 4, we examine how the signal from different substructure models appears surfaces as flux ratio anomalies, and provide forecasts for constraints on the half-mode mass for under various levels of precision in image flux measurements. We use a cosmology with $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$ and $h = 0.7$. We use the software package `lensmodel` to solve the lens equation and fit smooth models to lensing observables (Keeton 2011).

2 BAYESIAN INFERENCE ON THE SUBHALO MASS FUNCTION

In this section we describe how we infer the parameters describing subhalo populations within a Bayesian framework, and propose an Approximate Bayesian Computing (ABC) algorithm to contend with the highly stochastic nature of substructure lensing. Section 2.1 derives the expression for the posterior distribution of dark matter model parameters given a set of lensing observables. In Section 2.2, we briefly review the technique of Approximate Bayesian Computing.

2.1 Connecting dark matter model parameters to lensing observables

The observables from a strong lens system are image time delays $\mathbf{t}$, positions $\mathbf{x}$, and fluxes $\mathbf{f}$.¹ The data vector for

the n th lens can be written $\mathbf{d}_n = \{\mathbf{x}_n, \mathbf{t}_n, \mathbf{f}_n\}$, and we represent the dataset for a sample of N lenses as set $\mathbf{D} = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_N\}$.

Given a dark matter model with global properties described by a set of hyper-parameters $\mathbf{q}_{\text{sub}}$, the desired posterior distribution for the parameters $\mathbf{q}_{\text{sub}}$ is given by

$$p(\mathbf{q}_{\text{sub}} | \mathbf{D}) \propto \mathcal{L}(\mathbf{D} | \mathbf{q}_{\text{sub}}) \pi(\mathbf{q}_{\text{sub}}) \quad (1)$$

where $\pi(\mathbf{q}_{\text{sub}})$ is the prior probability for the parameters. In practice, $\mathbf{q}_{\text{sub}}$ describes the shape and normalization of the subhalo mass function, the spatial distribution of subhalos, the density profile of subhalos, etc.

Since the data from each lens is independent, the joint likelihood in Equation 1 can be written as a product of the likelihoods for each lens

$$\mathcal{L}(\mathbf{D} | \mathbf{q}_{\text{sub}}) = \prod_{n=1}^N \mathcal{L}(\mathbf{d}_n | \mathbf{q}_{\text{sub}}). \quad (2)$$

We specify the model for each lens as a combination of two mass components. The first is a macromodel, which accounts for most of the mass of the deflector and its environment. For the n th lens, we denote this component $\mathbf{q}_{\text{mac}(n)}$. The second component is the substructure population described by $\mathbf{q}_{\text{sub}}$. With these definitions, the components of $\mathbf{q}_{\text{mac}(n)}$ are nuisance parameters which are marginalized out of the posterior

$$\mathcal{L}(\mathbf{d}_n | \mathbf{q}_{\text{sub}}) \propto \int p(\mathbf{d}_n | \mathbf{q}_{\text{mac}(n)}, \mathbf{q}_{\text{sub}}) \pi(\mathbf{q}_{\text{mac}(n)}) d\mathbf{q}_{\text{mac}(n)}, \quad (3)$$

where we have assumed that the macromodel $\mathbf{q}_{\text{mac}(n)}$ is independent of the dark matter parameters $\mathbf{q}_{\text{sub}}$, and introduce the prior $\pi(\mathbf{q}_{\text{mac}(n)})$. The assumption that $\mathbf{q}_{\text{mac}}$ and $\mathbf{q}_{\text{sub}}$ are independent is not formally correct, as parameters such as the Einstein radius may be informative of the total halo mass and the normalization of subhalo mass function. Working with real datasets, this information would need to be incorporated in the analysis. For the purpose of forecasting the possible constraints on $\mathbf{q}_{\text{sub}}$, we examine the case of two fixed normalizations that span the expected range of substructure abundance for the halo masses implied by the distribution of Einstein radii in our mock data (see Section 3). As we will demonstrate in Section 4, the information content of each lens scales with the overall normalization, such that the bounds on a sample of lenses with diverse halo masses will be bound by the two limiting cases of the overall normalization we analyze.

Dark matter models do not directly map $\mathbf{q}_{\text{sub}}$ to a set of lensing observables. Rather, $\mathbf{q}_{\text{sub}}$ specifies statistical distributions for the masses, positions, density profiles, etc. of the subhalos. Defining a vector $\mathbf{m}_{\text{sub}}$ that specifies a specific substructure realization, the likelihood becomes

$$\mathcal{L}(\mathbf{d}_n | \mathbf{q}_{\text{sub}}) \propto \int \mathcal{L}(\mathbf{d}_n | \mathbf{m}_{\text{sub}}, \mathbf{q}_{\text{mac}(n)}) p(\mathbf{m}_{\text{sub}} | \mathbf{q}_{\text{sub}}) \times \pi(\mathbf{q}_{\text{mac}(n)}) d\mathbf{m}_{\text{sub}} d\mathbf{q}_{\text{mac}(n)} \quad (4)$$

To make progress in evaluating Equation 4, we note that the astrometric and time delay perturbations from substructure are generally small and can be partially absorbed by small adjustments in $\mathbf{q}_{\text{mac}(n)}$.² The flux ratios,

¹ To impose constraints, we actually use flux ratios in order to divide out the unknown source flux.

² Positions and time delays have some ability to probe substructure.

Table 1. Definitions and descriptions for parameters relevant to Equation 7

parameter	definition	description
$\mathbf{d}_n$	data from the n th lens	positions, time delays, flux ratios
$\mathbf{q}_{\text{sub}}$	vector of hyper-parameters describing the global subhalo mass function	(A_0, m_{hm}) , spatial distribution
A_0	normalization of subhalo mass function	1% substructure mass fraction at $R_{\text{Ein}} \approx A_0 = 2 \times 10^8 M_{\odot}^{-1}$ mass range $10^6 \leq M_{200} \leq 10^{10} [M_{\odot}]$
m_{hm}	half-mode mass	number of subhalos below m_{hm} is strongly suppressed
$\mathbf{m}_{\text{sub}}$	defines parameters for an individual substructure realization	subhalo positions, masses, density profiles
$\mathbf{q}_{\text{mac}(n)}^*$	maximum-likelihood macromodel for n th lens	fits n th positions, time delays in presence of substructure
$\mathbf{f}_n^*$	flux ratios computed in the forward model for n th lens	computed with $\mathbf{q}_{\text{mac}(n)}^*$ as opposed to true mass distribution

on the other hand, are determined by the second derivative of the gravitational potential near an image, thus the effects of substructure are difficult to reproduce by adjustments in $\mathbf{q}_{\text{mac}(n)}$. With this in mind, we write image positions and time delays separately from the flux ratios, writing $\mathbf{d}_n = \{\mathbf{d}_{\text{tx}(n)}, \mathbf{f}_n\}$, where $\mathbf{d}_{\text{tx}(n)}$ denotes the positions and time delays $\{\mathbf{t}, \mathbf{x}\}$. To relate $\mathbf{q}_{\text{sub}}$ to $\mathbf{d}_{\text{tx}(n)}$ and $\mathbf{f}_n$, our strategy is to forward model simulated data sets of image positions, time delays, and fluxes $\{\mathbf{d}'_{\text{tx}(n)}(\mathbf{q}_{\text{mac}(n)}, \mathbf{m}_{\text{sub}}), \mathbf{f}'_n(\mathbf{q}_{\text{mac}(n)}, \mathbf{m}_{\text{sub}})\}$, which depend on $\mathbf{q}_{\text{sub}}$ through the realizations of substructure $\mathbf{m}_{\text{sub}}$. The likelihood of observing $\mathbf{d}_n$ is therefore

$$\mathcal{L}(\mathbf{d}_n | \mathbf{m}_{\text{sub}}, \mathbf{q}_{\text{mac}(n)}) = \mathcal{L}(\mathbf{d}_{\text{tx}(n)} | \mathbf{d}'_{\text{tx}(n)}) \mathcal{L}(\mathbf{f}_n | \mathbf{f}'_n). \quad (5)$$

Next, we note that most choices of $\mathbf{q}_{\text{mac}(n)}$, with a wide prior distribution, yield the incorrect positions and time delays, and therefore do not contribute substantially to the integral in Equation 4. We therefore approximate the marginalization over the macromodel parameters by fixing the macromodel in a certain configuration $\mathbf{q}_{\text{mac}(n)}^*$, which fits the image positions and time delays. This step avoids sampling the potentially vast parameter space of $\mathbf{q}_{\text{mac}}$. Explicitly, $\mathbf{q}_{\text{mac}(n)}^*$ is defined by the relation

$$\mathbf{d}_{\text{tx}(n)} = \mathbf{d}'_{\text{tx}(n)}(\mathbf{q}_{\text{mac}(n)}^*, \mathbf{m}_{\text{sub}}). \quad (6)$$

The procedure of re-optimizing the macromodel was also employed in Dalal & Kochanek (2002).

By evaluating the flux ratios only with respect to $\mathbf{q}_{\text{mac}(n)}^*$, we effectively take a derivative, while formally an integral is required to marginalize over $\mathbf{q}_{\text{mac}}$. The procedure of optimizing, rather than marginalizing, the macromodel will yield a good approximation to the true likelihood as long as the image fluxes do not vary significantly over the range of macromodel parameters space for which the image positions and time delays are fit. We verify that the

variation in image fluxes for different macromodel configurations that fit the observed image positions is smaller than the typical 5–8% variations derived empirically in (Gilman et al. 2017) by fitting smooth lens models to realistic deflectors.³ In macromodel parameterizations more complicated than power-law ellipsoids that possess additional degrees of freedom, the fluxes may vary substantially even for configurations of $\mathbf{q}_{\text{mac}}$ with fixed image positions, and the macromodel may be better able to absorb flux perturbations from substructure. Stellar disks fall into this category (Hsueh et al. 2016b; Gilman et al. 2017), as do models with extreme angular structures (Congdon & Keeton 2005), but the former are unlikely to be present in a sample of massive elliptical deflectors, and the latter are unphysical. If external sources of error were reduced such that the dominant source of flux uncertainty stemmed from marginalizing the macromodel, one would need to explicitly do the marginalization. The procedure outlined here should be amended to sample prior distributions of $\mathbf{q}_{\text{sub}}$ and $\mathbf{q}_{\text{mac}}$ constructed on a lens-by-lens basis when working with real data, but for the purpose of computational expediency and making approximate forecast statements we leave this level of detail for future work.

After replacing the integral over $\mathbf{q}_{\text{mac}}$ with $\mathbf{q}_{\text{mac}(n)}^*$, Equation 4 becomes

$$\mathcal{L}(\mathbf{d}_n | \mathbf{q}_{\text{sub}}) \propto \int \mathcal{L}(\mathbf{f}_n | \mathbf{f}_n^*) p(\mathbf{m}_{\text{sub}}, \mathbf{q}_{\text{sub}}) d\mathbf{m}_{\text{sub}} \quad (7)$$

where we introduce the notation $\mathbf{f}_n^* = \mathbf{f}'_n(\mathbf{q}_{\text{mac}(n)}^*, \mathbf{m}_{\text{sub}})$. The parameters relevant to Equation 7 are summarized in Table 1.

³ We find that constraining the magnitude of the external shear, external shear angle, axis ratio and position angle to 0.01, 5°, 0.05, 5°, as is possible with detailed modeling (e.g. Wong et al. 2017), is sufficient to ensure the flux variations associated with marginalizing the macromodel is below 4% per image. We also verify that uncertainty associated with the power-law slope of the main deflector does not incur a serious bias in our forecasts (see Appendix E).

ture (Chen et al. 2007; Keeton & Moustakas 2009), but flux ratios experience stronger perturbations that are our focus here.

At this step, a Markov Chain Monte Carlo integration scheme would be inefficient, as the flux ratios corresponding to the overwhelming majority of realizations $\mathbf{m}_{\text{sub}}$ would not match those observed in the data. Rather than computing Equation 7 directly, we employ a computational method that allows us to efficiently explore the parameter space spanned by $\mathbf{q}_{\text{sub}}$.

2.2 Approximate Bayesian Computing

Approximate Bayesian Computing (ABC) is a computational algorithm rooted in Bayesian statistics that circumvents the direct calculation of intractable likelihoods, and enables inferences from simulated data sets computed in a forward model. For details in addition to those presented in this section, see e.g. Turner & Zandt (2012); Csillry et al. (2010); Lintusaari et al. (2017). In recent years, ABC has seen applications across a wide range of problems in cosmology and astrophysics (Weyant et al. 2013; Robin et al. 2014; Hahn et al. 2017; Birrer et al. 2017; Herbel et al. 2017).

In an implementation of ABC, one draws samples from a prior probability distribution, creates a forward model of simulated data from the samples, compresses the data sets into summary statistics (optional, but often necessary to keep computation costs low), and accepts or rejects the samples based on the similarity of the simulated to the observed data. An implementation of the algorithm therefore proceeds as follows:

1. Sample from a set of model parameters θ .
2. From the samples θ , forward-model a set of simulated data $\mathbf{d}'$.
3. The data vector $\mathbf{d}'$ is often multi-dimensional, but in many cases the relevant information that will discriminate between different parameters θ is contained in only a subset of the data. To reduce the dimensions of the problem, introduce the summary statistics $S(\mathbf{d}')$ and $S(\mathbf{d})$, which compress the relevant information contained in $\mathbf{d}'$ and $\mathbf{d}$.
4. Introduce a distance metric $R(S(\mathbf{d}'), S(\mathbf{d}'))$ - for instance, the Euclidean distance between the summary statistics in N-dimensions - and accept the proposition θ under the requirement

$$R(S(\mathbf{d}'), S(\mathbf{d})) \leq \epsilon \quad (8)$$

for some tolerance ϵ .

5. Repeat steps [1-4] until the distribution of accepted samples is stable to changes in the total number of samples computed, and the total number of samples retained in the posterior.

Formally, when implementing ABC one obtains samples from the posterior density

$$p(\theta | \mathbf{R}(\mathbf{d}', \mathbf{d}) \leq \epsilon), \quad (9)$$

with the property

$$p(\theta | \mathbf{d}) = \lim_{\epsilon \rightarrow 0} [p(\theta | \mathbf{R}(\mathbf{d}', \mathbf{d}) \leq \epsilon)]. \quad (10)$$

Thus, assuming the summary statistic contains the information necessary to distinguish between different models, the distribution of accepted samples from θ converges to the

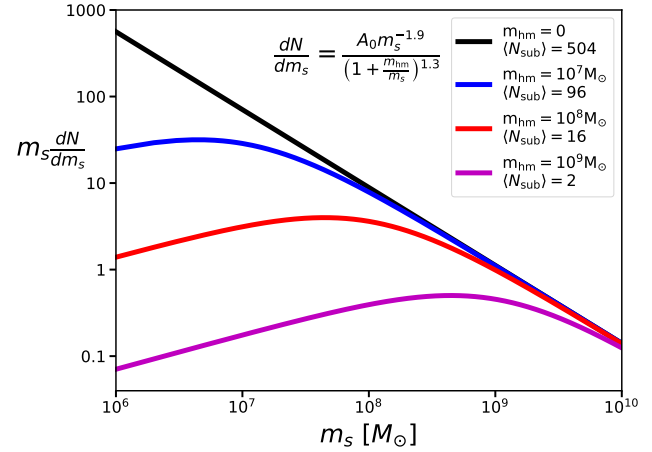


Figure 1. The subhalo mass function of Equation 13 which we use to generate substructure realizations. In the figure we vary the half-mode mass m_{hm} with fixed normalization $A_0 = 1.2 \times 10^8 \text{M}_\odot^{-1}$. We generate subhalo populations in such a way that the amplitude of a CDM-like and a WDM-like mass functions are identical for masses $m \gg m_{\text{hm}}$, rendering subhalos in projection to a radius of 18.6 kpc.

true posterior as ϵ tends to zero. Put another way, the relative number of accepted samples between multiple competing models reflects the relative probabilities of these models as $\epsilon \rightarrow 0$. In practice, one must compromise between an ϵ large enough to ensure timely convergence of the ABC procedure, and a value stringent enough to ensure the distribution of accepted samples is representative of the true posterior.

Crucial steps in the implementation of ABC include the choice of summary statistic $S(\mathbf{d})$, and the acceptance criterion ϵ . A summary statistic which erodes the discriminating information contained in the data will not converge to the true posterior. In a similar vein, an acceptance threshold too lax will result in a posterior distribution too broad, with the extreme limit of accepting all samples from θ and returning the prior. For this reason, assuming the algorithm has converged, the joint posterior distribution for the model parameters approximated in ABC will always be conservative, in the sense that it contains more volume than the true posterior.

In the context of substructure lensing, we compute a summary statistic for each realization based on the observed flux ratios, or the fluxes of three images normalized by the flux of the fourth. The summary statistic we use based on the observed flux ratios $\mathbf{f}_{\mathbf{n}}$ and the forward model flux ratios $\mathbf{f}_{\mathbf{n}}^*$ is given by

$$S(\mathbf{f}_{\mathbf{n}}, \mathbf{f}_{\mathbf{n}}^*) = \sqrt{\sum_{i=1}^3 (f_{n(i)} - f_{n(i)}^*)^2}, \quad (11)$$

where the summation runs over the three flux ratios of the n th lens, making use of the full information contained in these data. For an example of results using a different summary statistic, see Appendix C.

3 SIMULATION SETUP

In this section, we describe our lensing simulations, including the ingredients necessary to render substructure realizations and how we implement ABC to constrain the subhalo mass function. Section 3.1 describes our prescription for modeling substructure populations for both cold and warm dark matter scenarios. Section 3.2 describes the mock data sets we use in our simulations, including lens and source redshift configurations. In section 3.3, we explain how we use the information contained in the forward model within the ABC framework to make inferences on $\mathbf{q}_{\text{sub}}$.

3.1 Subhalo density profile, mass definition, and mass function

When quoting subhalo masses, we refer to the mass inside a sphere of radius r_{200} , M_{200} . We model subhalos as tidally truncated NFW profiles (Baltz et al. 2009)

$$\rho(r) = \frac{\rho_0}{x(1+x)^2} \frac{\tau^2}{x^2 + \tau^2} \quad (12)$$

where $x = \frac{r}{r_s}$, $\tau = \frac{r_t}{r_s}$, r_t is the truncation radius and r_s is a scale radius.⁴ The finite, truncated mass can differ from M_{200} to a varying degree depending on the truncation radius and the concentration, but the effect on image flux ratios is primarily determined by the central density, with the truncation playing a sub-dominant role provided $r_t > r_s$.

We render subhalos with M_{200} (denoted m_s) between $10^6 \leq \frac{m_s}{M_\odot} \leq 10^{10}$, drawing from a subhalo mass function written as a broken power law,

$$\begin{aligned} \frac{dN_{\text{wdm}}}{dm_s} &= \frac{dN_{\text{cdm}}}{dm_s} \left(1 + \frac{m_{\text{hm}}}{m_s}\right)^{-1.3} \\ &= A_0 \left(\frac{m_s}{M_\odot}\right)^{-1.9} \left(1 + \frac{m_{\text{hm}}}{m_s}\right)^{-1.3}, \end{aligned} \quad (13)$$

a functional form which resembles the subhalo mass function of a WDM particle (Schneider et al. 2012; Lovell et al. 2014b). We restrict our analysis to this specific parameterization, and do not attempt to constrain the exponent -1.3 appearing in Equation 13 or the -1.9 slope of the CDM function. More complex scenarios, such as multi-component dark matter in which only a fraction of the dark matter is warm, will require a more careful treatment of the parameterization (see Vegetti et al. 2018). For reference, $A_0 \approx 2 \times 10^8 M_\odot^{-1}$ corresponds to a convergence in substructure of 0.005, or a mass fraction of 1% at the Einstein radius (for more details see Appendix A).

The parameter m_{hm} , the half-mode mass, denotes the mass scale at which the WDM power spectrum is damped with respect the CDM case by one-half. Assuming a thermal relic particle of mass m comprises the dark matter, one can establish scaling relation $m_{\text{hm}} \propto m^{-3.33}$ (Schneider et al. 2012). We normalize this relation to the $2 \times 10^8 M_\odot h^{-1} \sim$

⁴ While experimenting with Pseudo-Jaffe profiles, we find that our results depend sensitively on how convergence is partitioned between the truncation radius and the central density. Since the central density dominates the flux ratio signal, different choices for the truncation and central density normalization yield significantly different results.

3.3keV result of Viel et al. (2013), and translate between the two parameters as

$$m_{\text{hm}}(m) = 10^{10} \left(\frac{m}{1\text{keV}}\right)^{-3.33} M_\odot h^{-1}. \quad (14)$$

In our simulations, we convert to physical masses in the lens plane using $h = 0.7$. With this metric, the 2 keV result from Birrer et al. (2017) corresponds to $m_{\text{hm}} = 10^{8.8} M_\odot$. As shown in Figure 1, we normalize the mass function such that m_{hm} does not affect the amplitude at scales $m_s \gg m_{\text{hm}}$, yielding the same numbers of very massive subhalos, on average.

We combine this mass definition with a form for the mass-concentration relation for warm dark matter halos presented by Bose et al. (2016) (see also Macciò et al. 2008; Ludlow et al. 2016)

$$c(m_s) = 6 \left(\frac{m_s}{10^{12} M_\odot}\right)^{-0.098} \left(1 + 60 \frac{m_{\text{hm}}}{m_s}\right)^{-0.17}. \quad (15)$$

which results in lower central densities at a given mass for warm dark matter models. The relation between concentration and m_{hm} reflects the later collapse epoch of small WDM subhalos, which prevents them from building up their concentrations over time.⁵

Given an M_{200} drawn from the mass function in Equation 13, and a concentration from Equation 15, we compute the normalization ρ_0 and the scale radius r_s . To obtain the truncation radius, and the lensing properties associated with the mass profile in Equation 12, we generate subhalos in a 3-D sphere of radius 250 kpc (see Appendix A for details regarding the spatial distribution). Given r_{3d} , we compute the truncation radius (Cyr-Racine et al. 2016)

$$r_t = \left(\frac{m_s r_{3d}^2}{2\Sigma_{\text{crit}} R_{\text{Ein}}}\right)^{\frac{1}{3}}, \quad (16)$$

where Σ_{crit} is the critical density and where $R_{\text{Ein}} \approx 1''$ is a typical Einstein radius.

3.2 Mock Data Sets

We consider three subhalo mass functions: a WDM mass function with m_{hm} of $10^8 M_\odot$ and normalization of $10^8 M_\odot^{-1}$, and two CDM mass functions with normalizations of $8 \times 10^7 M_\odot^{-1}$ and $40 \times 10^7 M_\odot^{-1}$. The two normalizations in the CDM simulations correspond to projected mass fractions at the Einstein radius of 0.4% and 2%, respectively and bracket a plausible range than spans the theoretical uncertainties associated with the connection to halo mass (e.g. Jiang & van den Bosch 2017) and the tidal destruction of subhalos by the host galaxy (Despali & Vegetti 2017; Garrison-Kimmel et al. 2017). The low normalization case corresponds to the scenario in which all subhalos inside the halo NFW scale radius are destroyed, while the latter corresponds to no destruction (see the discussion in Appendix A for details on obtaining these numbers).

When rendering subhalos to create mock data sets, we

⁵ We do not model the scatter in the mass-concentration relation. In a careful measurement, this feature should be included.

solve the lens equation and ray trace with every subhalo between 10^6 and $10^{10} M_\odot$ included in the computation. We distribute substructures over an SIE+shear macromodel with randomly oriented shear and ellipticity. Ellipticity (shear) is sampled from a Gaussian with mean 0.2 (0.05) and standard deviation 0.05 (0.01), and Einstein radii sampled from a Gaussian with mean $1''$ and variance $0.2''$. We randomly sample source positions to produce equal numbers of cusp, fold and cross configurations. In Appendix B, we compare the sensitivities of the different image configurations by using each on separately to infer $\mathbf{q}_{\text{sub}}$. We take the deflector to lie at a typical redshift $z_d = 0.5$, and put the source at $z_s = 1.5$. The source in both the data and forward model is parameterized by a circular Gaussian with width of 10 pc, or 1.2 m.a.s. To check to what degree this source size plays a role in our analysis, using source sizes as large as 30 pc we generate distributions of thousands of flux ratios from identical mass functions, and verify the distributions are nearly identical.

Even in the presence of identical subhalo populations, the observed and simulated flux ratios will differ due to the underlying macromodel and measurement errors. Examining mock lens systems with luminous mass components from galaxies in the Virgo and Coma clusters, in Gilman et al. (2017) we found flux ratio anomalies in mock deflectors composed of a NFW halo with a galaxy at its center, with respect to an SIE+shear model. Ray tracing through galaxies formed in the Illustris simulations, the authors of Hsueh et al. (2017a) also conclude that the incorrect macromodel can contribute to a measured flux ratio anomaly.

To ensure that modest deviations away from an isothermal-ellipsoid macromodel parameterization does not bias the precision of our inference on $\mathbf{q}_{\text{sub}}$, we have simulated mock lenses with with power law slopes drawn from a distribution offset from isothermal and modeled them with SIE profiles in the forward model. Encouragingly, the precision of our forecast is not degraded when subjected to this source of error. See Appendix E for more details regarding this test. In practice, systematic error associated with the macromodel can be dealt with by sampling additional parameters in the forward model. For instance, macromodel deviations around an isothermal profile can be handled by simultaneously sampling the power law slope and $\mathbf{q}_{\text{sub}}$.⁶

Even for samples of morphologically simple deflectors (i.e. no disks or other prominent morphological features), we expect deviations in flux ratios at the percent level from measurement errors, and residual uncertainties in the image fluxes caused by the parameterization of the macromodel. To incorporate these uncertainties in our simulations, we add perturbations to the fluxes in our mock data sets. We model flux anomalies as Gaussian, and perturb each flux F as

$$\begin{aligned} F &\rightarrow F + \delta F; \\ \delta F &= \mathcal{N}(0, \delta \times F), \end{aligned} \quad (17)$$

examining specific cases of $\delta = 0.02$, $\delta = 0.04$ and $\delta = 0.08$, which correspond to flux errors of 2, 4 and 8%. Conceptu-

ally, one can interpret this operation as erasing information at the δ level, which enables one to track the sensitivity of lensing constraints on small sources of flux perturbations. These perturbations lump together all deviations in image fluxes away from those of an SIE+shear model fit to the data that are not caused by dark substructure, including measurement errors and the baryonic structure of the deflector, and bracket the range of errors one expects for morphologically simple deflectors. These perturbations are empirically motivated by the flux residuals we encounter in (Gilman et al. 2017) fitting smooth lens models to lenses build from high resolution images of galaxies in the nearby Virgo cluster. We reiterate that extreme morphological features like stellar disks may produce larger systematic perturbations than those we are mimicking with the δ perturbations, but these prominent features are unlikely to be present in massive ellipticals (Auger et al. 2010; Sonnenfeld et al. 2013).⁷ Furthermore, deflectors likely to contain disks can be readily identified based on velocity dispersion and stellar mass, in addition to high resolution imaging. Since we add these flux errors independent of perturbations to the image positions and time delays, we assume that the macromodel and measurement-error induced flux ratio anomalies are independent from astrometric and time delay anomalies, a conservative choice as correlations provide additional information that can be used to identify these features in the data.

For reference, current techniques using measurements of narrow-line fluxes achieve precision of roughly 4 – 6% (Nierenberg et al. 2014, 2017). The 8% errors can therefore be interpreted as pessimistic case, with errors induced by the use of a simplified macromodel compounding measurement errors, while 2% simulates an optimistic future precision. The 4% curve serves to illustrate how the bounds evolve between these two extremes. Finally, we add measurement errors of 3 m.a.s. to the mock image positions, typical of astrometric uncertainties with current instruments.

3.3 Constraining the subhalo mass function

In our simulations, the subhalo mass function is defined by the free parameters A_0 and m_{hm} ; our goal is to relate these parameters to the observed data from N simulated lenses, **D**. To do so, for each A_0 and m_{hm} , we render ≈ 2000 substructure realizations $\mathbf{m}_{\text{sub}}$ per proposed set of parameters $\mathbf{q}_{\text{sub}}$, sampling the prior uniformly in A_0 and uniformly in $\log_{10}(m_{\text{hm}})$, yielding in $\approx 10^6$ realizations per lens. In Appendix D, we perform tests to verify convergence with this number of samples. For each realization, we use `lensmodel` to re-optimize the macromodel parameters $\mathbf{q}_{\text{mac}(n)}$ that satisfy Equation 6, and ray-trace through a grid sampled at 0.4 m.a.s. per pixel to obtain image magnifications and flux ratios with an extended background source modeled as a circular Gaussian with a full-width at half-maximum of 10 pc. When re-optimizing the macromodel, we assume uncertainties of 3 m.a.s. and 2 days on positions and time delays. We use flux constraints weak enough that they do not impact

⁶ In principle, this approach could be extended in the forward modeling framework to more complex morphological features on a lens by lens basis.

⁷ In fact, studies of massive elliptical lensing galaxies find their mass profiles are well modeled by nearly isothermal power law ellipsoids (Sonnenfeld et al. 2015; Shankar et al. 2017; Gilman et al. 2017).

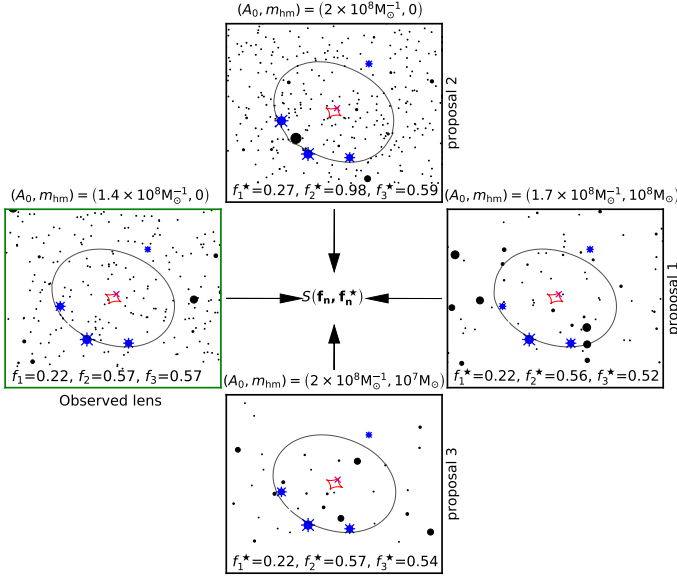


Figure 2. This figure presents an illustration of the forward modeling analysis. We use one realization of substructure to create mock data, as shown in the panel outlined in green (left). Lensed images (blue) have sizes corresponding to their flux, while subhalos (black) have sizes corresponding to M_{200} . We compute flux ratios with respect to the middle image in the triplet; f_1 denotes the image in the upper right, while f_2 and f_3 denote images in the left and right of the triplet, respectively. We then discriminate between different parameters describing the subhalo mass function by drawing many substructure realizations from proposal mass functions, three examples of which are shown here (panels outlined in black). For each realization we re-optimize the macromodel and compute the summary statistic $S(\mathbf{f}_n, \mathbf{f}_n^*)$ from Equation 11, which we use to accept or reject the realization and the parameters describing the mass function it came from. The procedure visualized here is repeatedly applied in the full analysis shown in Figure 3.

the re-optimization of the macromodel, allowing us to impose different flux perturbations in our mock data sets after running the simulation.

To deal with the flux errors we add to our mock data sets, we add perturbations of the same form as Equation 17 to the forward model fluxes. With the perturbed fluxes in hand, we evaluate the statistic in Equation 11. After computing the summary statistics, each proposal of $\mathbf{q}_{\text{sub}}$ in each of the N lenses has a set metric distances associated with it. At this juncture, we apply a rejection criterion to select the most probable models, including the 1,800 samples with smallest corresponding summary statistics. In principle, because the systems are statistically independent, we could apply this criteria to each one individually and multiply the resulting distributions. Practically, however, multiplying a large number of probability densities together computed on a discretely sampled space is numerically unstable. To handle these numerical issues, we first reduce the dimensionality adding the summary statistics for pairs of lenses. In the limit of an infinite number of realizations and infinitely stringent acceptance criteria, this is equivalent to multiplying the likelihoods. To each of the resulting $\frac{N}{2}$ probability densities, we apply a Gaussian Kernel Density Estimator (KDE), and multiply them. We verify that, with these choices, our

Forward Model

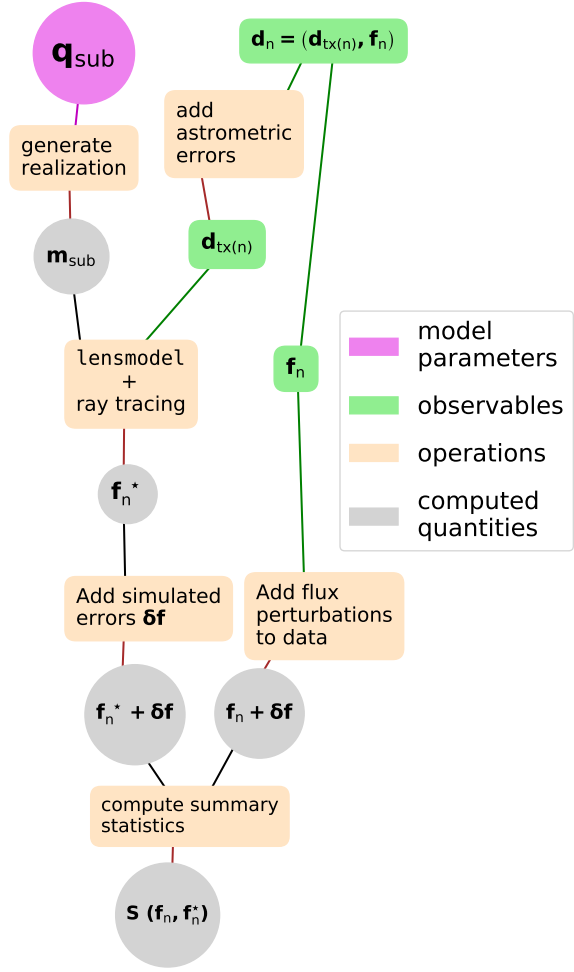


Figure 3. A schematic overview of the forward model used to compare flux ratios in the simulated data sets to flux ratios derived from substructure realizations drawn from $\mathbf{q}_{\text{sub}}$.

algorithm satisfies a rudimentary test of convergence (see Appendix D).

In summary: the forward model $\mathbf{f}_n^*$ contains the information needed to discriminate between different dark matter models through the realizations $\mathbf{m}_{\text{sub}}$, drawn from the hyper-parameters $\mathbf{q}_{\text{sub}}$. For any parameterization of $\mathbf{q}_{\text{sub}}$, Equation 7 relates the flux ratios $\mathbf{f}_n'(\mathbf{q}_{\text{mac}(n)}, \mathbf{m}_{\text{sub}})$ to the observed data by evaluating the flux ratios at fixed image positions, enforced by first re-optimizing the macromodel by fitting a smooth lens model to the data $\mathbf{d}_{\text{tx}(n)}$, and then computing the flux ratios $\mathbf{f}_n^*$ for this lens model in the presence of the substructure realization $\mathbf{m}_{\text{sub}}$.

We account for flux ratio anomalies caused by measurement errors, and by the imperfect SIE+shear macromodel fit to realistic deflectors, by adding random Gaussian perturbations to the forward model fluxes. We then compute a summary statistic which reflects the degree to which the observed flux ratios match those computed in the forward model. The final posterior probability distribution is composed of the set samples from the prior $\pi(\mathbf{q}_{\text{sub}})$ whose corre-

sponding realizations $\mathbf{m}_{\text{sub}}$ yield summary statistics closest to those computed from the flux ratios in the observed data. To keep computational costs low, in the forward model we only render substructures below masses of $10^{7.5}M_{\odot}$ if they lie within 0.5 arcseconds of an image. All higher mass subhalos are included regardless of position.

4 RESULTS

Based on the method we outlined in the previous two sections, we are able to quantify the effect of substructure on image flux ratios, and forecast the constraining power of this method. We analyze three scenarios: a CDM mass function with a mass fraction in substructure of 0.4% substructure at the Einstein radius ($A_0 = 8 \times 10^7 M_{\odot}^{-1}$), a CDM mass function with a mass fraction in substructure of 2% at the Einstein radius ($A_0 = 4 \times 10^8 M_{\odot}^{-1}$), and a WDM mass function with a mass fraction in substructure of 0.5% substructure at the Einstein radius ($A_0 = 10^8 M_{\odot}^{-1}$) and a half-mode mass of $10^8 M_{\odot}$ corresponding to a 3.6 keV thermal relic.⁸ We begin in Section 4.1 by discussing how variations in the normalization and half-mode mass impact distributions of flux ratios. In 4.2, we show the results of simulations in which the mock data sets are free from errors; these simulations serve to determine the constraints achievable with the best possible data. In 4.3, we add measurement errors and uncertainties to the fluxes in the mock data sets, account for them in the forward model, and quantify their effect on ones inference of $\mathbf{q}_{\text{sub}}$.

4.1 Flux ratio signal from structure in the lens plane

In Figure 4, we plot the cumulative distribution of flux ratio anomalies for 10,000 substructure realizations. Several trends emerge which help to understand the signal coming from substructure in image flux ratios.

If we consider modest anomalies whose strength is $< 60\%$ (summed in quadrature), we see that the normalization and half-mode mass both affect the frequency of anomalies. This suggests a degeneracy between the two mass function parameters, which indeed surfaces in a joint inference. The tails of the distributions, shown in the lower panel of Figure 4, tell a different story: the curves behave similarly for anomalies whose strength is $> 140\%$ (summed in quadrature), except for the most extreme WDM case (shown in magenta). Together, these results suggest that the most massive substructures, which survive the free-streaming cutoff and are present in the black, blue, grey, and red curves, are responsible for the largest flux ratio anomalies. The frequency of flux ratio anomalies from the model with $m_{\text{hm}} = 10^9 M_{\odot}$ suggests that substructures with masses between $10^7 M_{\odot}$ and $10^9 M_{\odot}$ dominate the lensing cross section for the largest flux ratio anomalies, while the cross section for small flux ratio anomalies is dominated by subhalos of mass $\lesssim 10^7 M_{\odot}$.

⁸ Due to the demanding computational resources required per lens in the forward modeling procedure, we limit our present analysis to these cases.

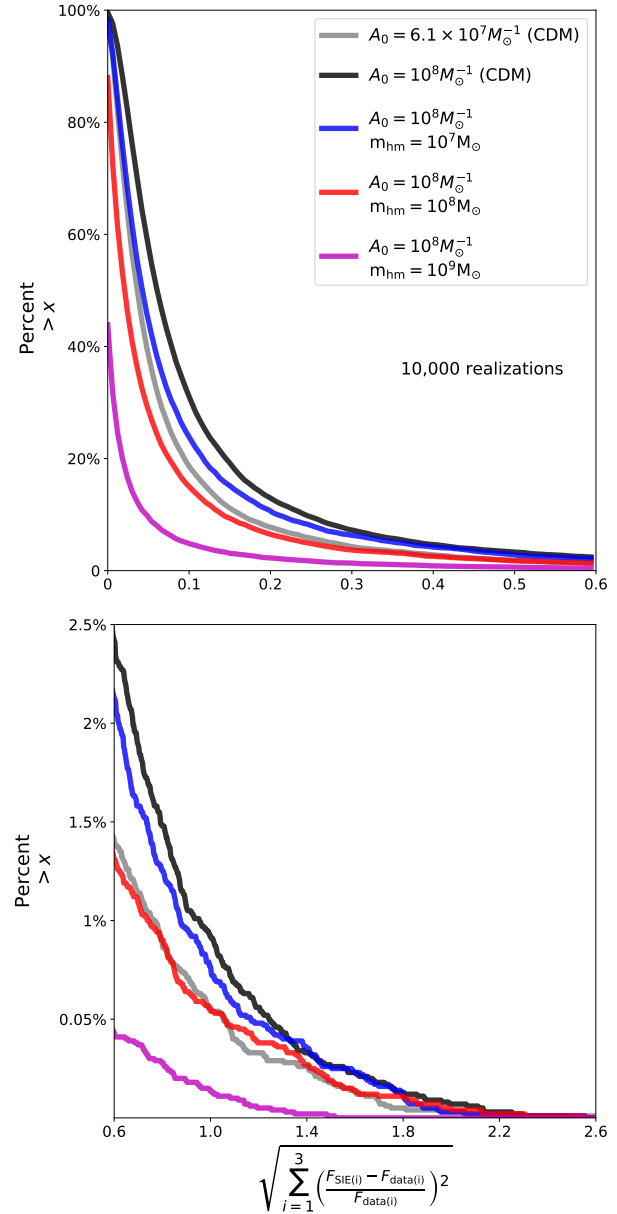


Figure 4. Cumulative distributions of flux ratio anomalies summed in quadrature for different subhalo mass functions with varying normalization (grey, black), and identical normalizations but varying m_{hm} (black, blue, red, magenta). The lower panel shows a zoom-in of the long, low probability tail of the distributions. Models with higher normalization (black vs. grey) produce more frequent flux ratio anomalies. Models with high m_{hm} produce less frequent anomalies than the black curve with $m_{\text{hm}} = 0$ for flux ratio anomalies < 1.4 .

4.2 Inference on subhalo mass function with idealized data sets

Before adding simulated errors to the measured flux ratios, as will be present in a real sample of lenses, we first perform the inference on data sets where the flux ratios in the data and the forward model are un-perturbed. Effectively, in these simulations, the only unknowns are the properties of the underlying subhalo mass function, as the macromodel in both the data and the forward model is the same. They

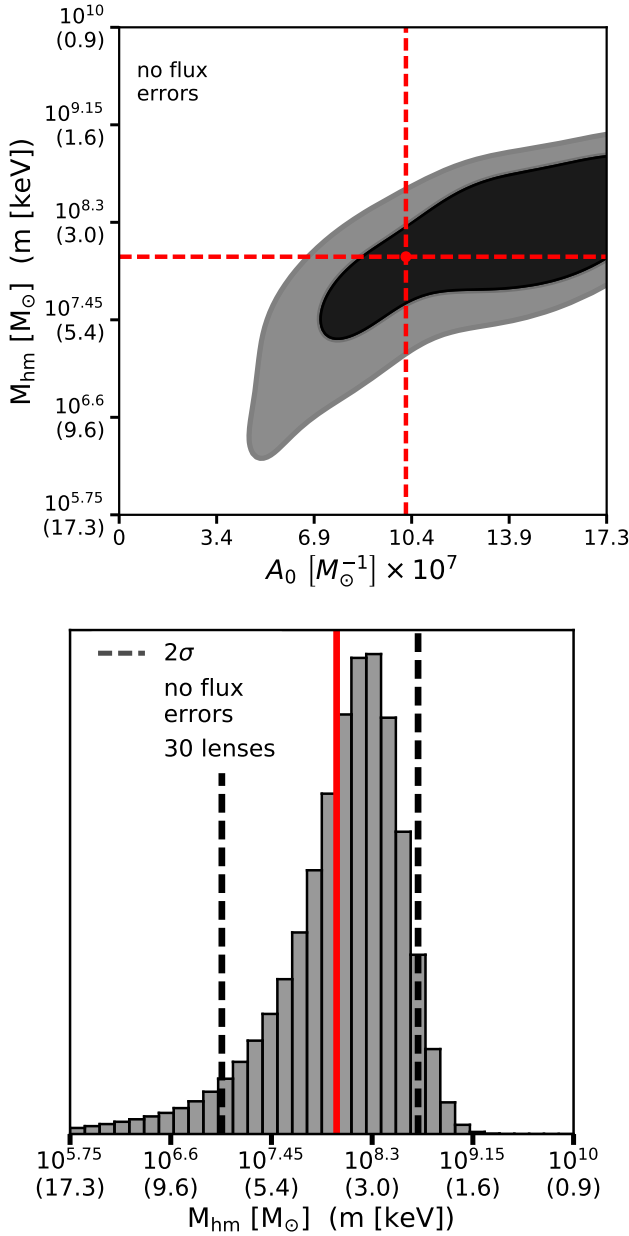


Figure 5. (Top) Joint posterior distribution for data with input $(A_0, m_{\text{hm}}) = (1.04 \times 10^8 M_\odot^{-1}, 10^8 M_\odot)$, marked by red lines. The closing of the 1σ contour from the bottom, ruling out CDM mass functions at 2σ , demonstrates the sensitivity of the flux ratios, and this method, for probing substructure on scales $M_{200} \approx 10^{6.5} M_\odot$. (Bottom) Marginalized constraints on m_{hm} . The 2σ bounds correspond to $10^7 M_\odot < m_{\text{hm}} < 10^{8.7} M_\odot$.

demonstrate the best one could hope to do by modeling subhalos only in the lens plane.

First, Figure 5 shows the joint posterior distribution for a simulated data set of 30 lenses with $\mathbf{q}_{\text{sub}} = (A_0, m_{\text{hm}}) = (10.4 \times 10^7 M_\odot^{-1}, 10^8 M_\odot)$. This idealized calculation provides a useful limit to the sensitivity of the flux ratio anomaly method. The closing of the 2σ contour around $10^{6.5} M_\odot$ demonstrates that the signal in flux ratios is sensitive to subhalos of this mass. On their own, these objects

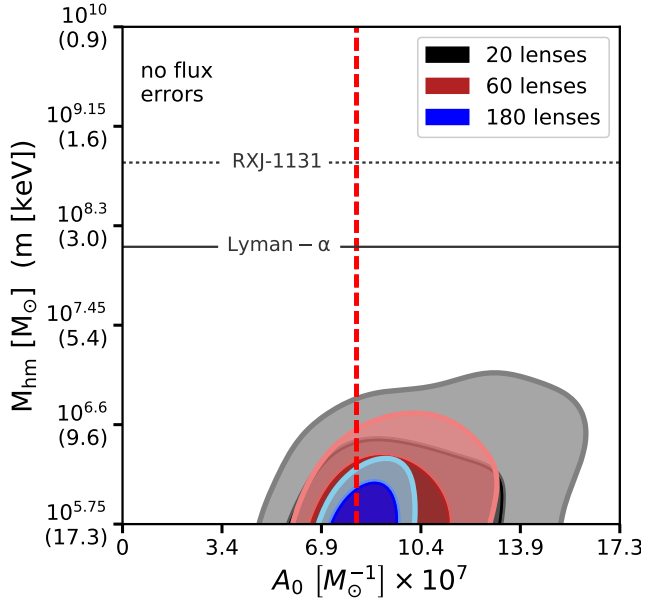


Figure 6. Posterior distribution for a CDM-like mass function with input mass function parameters $(A_0, m_{\text{hm}}) = (8.2 \times 10^7 M_\odot^{-1}, 0)$. In this idealized simulation, we do not add measurement errors or any other perturbations to fluxes in the simulated data sets. Effectively, the only unknown variables are A_0 and m_{hm} , which describe the subhalo mass function. In such an idealized case, flux ratios probe scales below $10^{6.5} M_\odot$, ruling out WDM models with $m_{\text{hm}} > 10^{6.1} M_\odot$ and $m_{\text{hm}} > 10^{6.4} M_\odot$ at 1 and 2σ , respectively. For reference, grey solid and dashed lines show the 2σ bounds on a WDM particle mass from Viel et al. (2013) and Birrer et al. (2017), respectively.

produce a very weak lensing signal, but they create collective effects that should make it possible to distinguish between a CDM scenario in which they are abundant and a WDM scenario in which their numbers are depleted. In this case shown here, the 2σ bounds on the half-mode mass are $10^7 M_\odot < m_{\text{hm}} < 10^{8.7} M_\odot$, which correspond to bounds on the WDM particle mass of 7.3 and 2.3 keV, respectively.

In a second simulation, we use a data set composed of 180 systems with an input mass function with $(A_0, m_{\text{hm}}) = (8.2 \times 10^7 M_\odot^{-1}, 0)$. Figure 6 shows the joint posterior distribution on this data set as the number of lenses is increased. We interpret the 2σ bound of $10^{6.4} M_\odot$ as the best one could do with 180 systems.⁹ As we show in the next section, these bounds weaken significantly under flux ratio errors of 2%, 4% and 8%, which mimic the signal in flux ratios produced by the smallest subhalos, or by subhalos far from an image.

4.3 Inference on subhalo mass function with simulated uncertainties

In this Subsection we demonstrate the effect of flux errors on our inference.

We start by using the the same data drawn from the

⁹ In each of the simulations with mock data sets containing CDM mass functions, we only quote the upper bound because the 2σ lower bound is set by the limits of the prior assigned to m_{hm} .

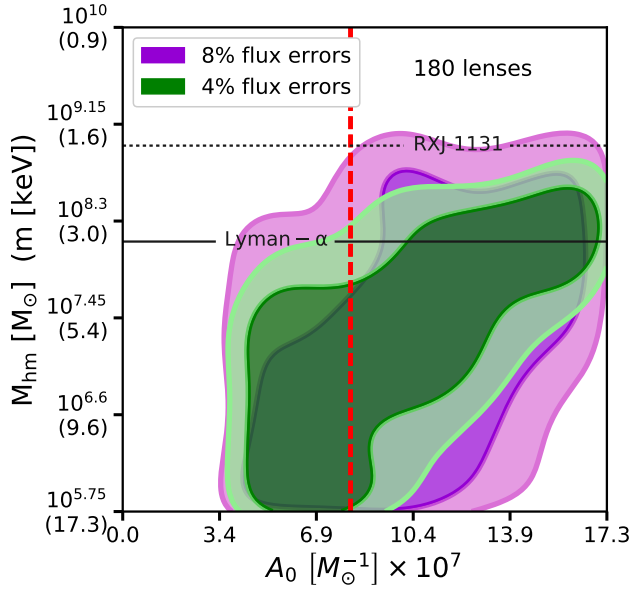


Figure 7. Similar to Figure 6, but in this case the fluxes in the mock data sets receive 4% and 8% uncorrelated Gaussian errors. On their own, these errors look like substructure, biasing the inference to models with more subhalos. However, we are able to account for this uncertainty by introducing perturbations to the fluxes in the forward model that match those applied to the mock data. Adding noise to the fluxes washes out the signal from the smallest subhalos, and the 2σ constraining power on m_{hm} is diminished by over an order of magnitude. This figure shows the result of a single draw of Gaussian errors in the fluxes. In Figure 9, we compute 2σ bounds on m_{hm} averaged over many draws of these errors.

low subhalo mass normalization scenario as is plotted in Figure 6. We add flux ratio errors of 4 and 8 percent to the data, and add errors of the same form to the forward model. Random flux errors applied to the data and forward model weaken the constraint on A_0 by a factor of ~ 2 and on m_{hm} over an order of magnitude, as shown in Figure 7. This is in part due to the loss of signal from subhalos below masses $10^8 M_\odot$. The 2σ posterior probability contours slope upwards, mirroring the degeneracy seen in Figure 5. We also explore the effect of flux errors in the case of a higher normalization (Figure 8). A higher normalization results in more frequent high-significance flux perturbations from substructure, which translates to stronger constraints on m_{hm} for a fixed number of lenses, and for fixed flux uncertainties.

In the case of low normalization, the inference with 4% errors rules out models with no substructure and very warm mass functions, but with this specific realization of flux errors applied to the data does not quite surpass the bounds set by Viel et al. (2013) with the Lyman- α forest. If more subhalos survive in the lens plane than the most extreme models for tidal stripping suggest, the constraints on the free streaming length improve to ≈ 5.5 keV, as shown in Figure 8. Of course neither of these scenarios incorporate the boost to the overall normalization which comes from line of sight structure, so the low normalization case may be seen as a lower limit on the signal we expect. In a future paper we will consider the effects of this additional signal.

We note that our method can robustly distinguish be-

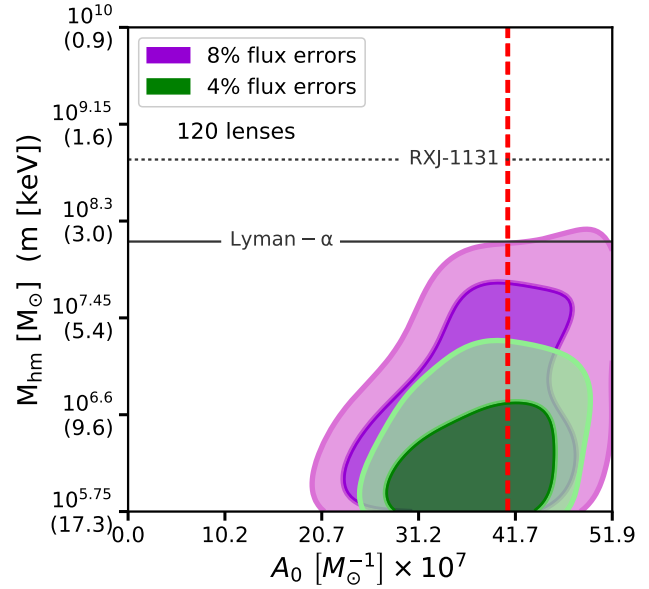


Figure 8. Similar to Figure 7, but an inference performed on a dataset with a subhalo mass function normalization corresponding to the expected mass fraction in substructure if there is no tidal disruption of subhalos within the host scale radius (see Appendix A). The higher abundance of substructures in this scenario results in a higher probability of observing flux perturbation larger than the effective detection threshold, which is determined by the precision of the image fluxes. Neither the high nor low normalization scenarios includes the expected boost from line of sight structure which we will consider in a future work.

tween the two normalization scenarios even in the case of 8% flux errors. This is important as it indicates that we are able to measure the difference between WDM and a high normalization and CDM and a low normalization. This can potentially provide useful input to simulations of tidal disruption in massive halos, although the signal from the lens plane will be somewhat diminished by the boost from line-of-sight structure.

Figures 7 and 8 each show one realization of flux perturbations applied to the data. In Figure 9, we average over many realizations of flux perturbations to properly account for the constraints possible from N lenses. Stepping through samples of N lenses in increments of 20, we compute 200 bootstraps for each sample, and repeat this procedure for errors of 2%, 4%, and 8%. The resulting curves in reflect a compromise between more precise flux measurements and increasing the lens sample size. Both options improve the constraints afforded by image flux ratios on the subhalo mass function, but increasing the flux precision by better handling systematic errors in the lens modeling and by more accurate measurements results yields better marginal gains. The curves do not appear to level off, suggesting that these bounds may improve by including more lenses in the inference, but without simulating more systems it is difficult to make definitive statements for the regime $N > 180$. The normalization is also observed to play a key role, as it effectively boosts the signal to noise ratio in the data and enables measurements more robust to deviations in the fluxes at the percent level.

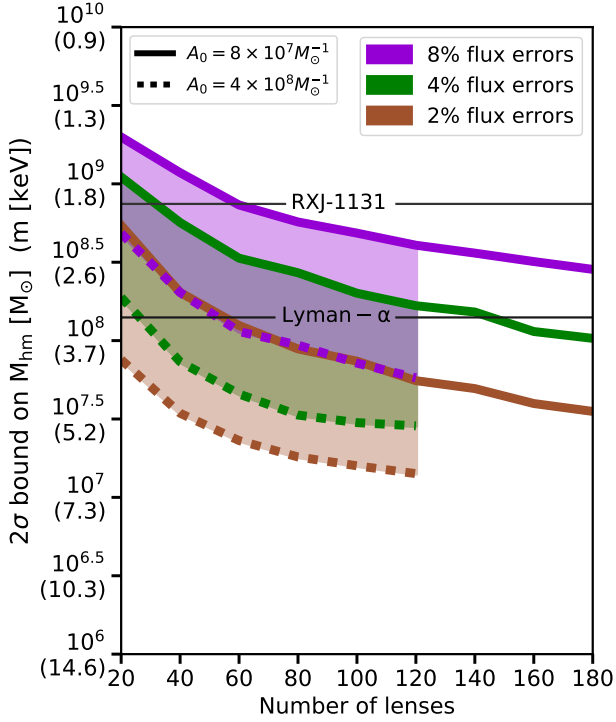


Figure 9. Curves show the dependence of the 2σ bounds on m_{hm} , and the mass m of a corresponding thermal relic dark matter particle, as a function of the number of lenses. For each plotted point, we randomly sample different combinations of N lenses, each with a random draw of flux errors of 2, 4, and 8 percent. We iterate this procedure 200 times, and compute the mean of the two sigma bounds over the 200 iterations. With 180 lenses, the 2σ bound on m_{hm} is $10^{6.4}M_{\odot}$, $10^{7.5}M_{\odot}$, $10^8M_{\odot}$, $10^{8.4}M_{\odot}$ for fluxes with errors controlled at 0%, 2%, 4%, and 8%, respectively. For the higher normalization case, (dotted curves) the signal to noise ratio in the data is higher, allowing tighter constraints for a fixed sample size. For reference, horizontal lines show bounds on the mass of a thermal from the Lyman- α forest (Viel et al. 2013) and an analysis of the strong lens RXJ-1131 (Birrer et al. 2017).

Conceptually, one can interpret the dependence of the 2σ bounds on different flux errors δF as tracers of the probability distribution $p(\delta F|\mathbf{m}_{\text{sub}})$. Supposing that $m_{\text{s}*}$ defines the subhalo mass scale that dominates this probability density, it follows that adding random noise at the δF level will erode the sensitivity to mass scales $\approx m_{\text{s}*}$, so the fast deterioration of constraining power on m_{hm} tracks the loss of sensitivity to mass scales $\approx m_{\text{s}*}$. In this context, a higher normalization increases the probability at each scale $m_{\text{s}*}$ of observing a flux ratio anomaly δF , counteracting the loss of signal induced by adding flux errors at the δF level.

5 DISCUSSION AND CONCLUSIONS

We have introduced a new method to infer the nature of dark matter from observations of flux ratios in quadruply lensed quasars. The method uses an Approximate Bayesian Computing algorithm to statistically infer the input parameters describing the subhalo mass function without directly computing a likelihood function. We have illustrated the method

by performing simulations of strong lenses systems with substructure populations of NFW subhalos rendered in the lens plane, in the case of cold and warm dark matter, and for two different normalizations of the subhalo mass function.

While a real sample of lenses will be diverse in both redshift distribution and host halo mass, the lens and source redshift will primarily impact the contribution from the line of sight, while the connection between the normalization and host halo mass, and the effect on lensing observables, is subject to considerable theoretical uncertainty (Despali & Veggetti 2017; Garrison-Kimmel et al. 2017). We handle this uncertainty by considering two limiting cases of the normalization, corresponding to scenarios with and without complete subhalo disruption within the scale radius of the parent halo. As the normalization of each lens effectively weights the information content available, the limiting cases we analyze bound the constraints from a sample of lenses with diverse halo masses.

Our main results can be summarized as follows:

- In an idealized scenario, where the macromodel is known to high precision and other sources of flux ratio errors are mitigated, the only source of flux ratio perturbation comes from dark substructure and flux ratios probe the mass function at scales below $10^7M_{\odot}$. With flux uncertainty at the level 2%, 4% and 8%, the bounds on the half-mode mass $[M_{\odot}]$ (thermal relic mass [keV]) are $m_{\text{hm}} = 10^{7.5}$ (5.0), $m_{\text{hm}} = 10^8$ (3.6) and $m_{\text{hm}} = 10^{8.4}$ (2.7) with 180 systems. For the higher normalization case, the improvement in the signal to noise ratio in the data yields constraints of $m_{\text{hm}} = 10^{7.2}$ (6.6), $m_{\text{hm}} = 10^{7.5}$ (5.3) and $m_{\text{hm}} = 10^{7.8}$ (4.3) with just 120 lenses. In a WDM scenario, we find with no uncertainty in flux ratios that we can measure the position of the free-streaming cutoff in the subhalo mass function with just 30 lenses, constraining it to between $10^7M_{\odot}$ and $10^{8.7}M_{\odot}$ at 2σ . With less control over systematics and degraded measurement precision, more than 30 lenses would be required to achieve these constraints, but our simulations suggest that this method can, in principle, measure the warmth of dark matter should CDM be the incorrect model. Provided one controls for systematic errors in flux ratios associated with incorrect macromodels, these constraints will likely improve after adding the contribution from line of sight structure, which contributes substantial additional signal.

- The 2σ bound on m_{hm} improves rapidly with increasing flux uncertainties, and falls slowly after $N \approx 100$ lenses. This reflects the sensitivity of flux ratios to low mass subhalos, which impart deviations at the level of a few percent. In terms of overall strategy for the study of strong lens systems, this establishes the necessity of measuring fluxes precisely and controlling for systematic errors arising from the parameterization of the macromodel. The simple SIE+shear parameterization implemented in this work may not be sufficient for systems in which additional mass components are present in the main deflector, such as stellar disks. In practice, identifying morphological complexity in the main deflector can be achieved by deep imaging of the lensing galaxy to identify luminous mass components, and preferentially analyzing slow-rotators with high central velocity dispersions.

- The frequency and magnitude of flux ratio anomalies

differentiates between different dark matter models. In this work, we have explored the effect of a varying normalization and half-mode mass. While the half-mode mass scale carries information regarding the nature of dark matter, baryonic effects can impact the normalization, which plays a crucial role as it effectively determines the signal to noise ratio in the data. This translates into better constraints on the shape of the mass function, making it easier to distinguish WDM from CDM. There is some degeneracy between the normalization of the mass function and the half-mode mass, but flux ratios are sensitive enough to break this degeneracy and probe mass scales below $10^8 M_\odot$, where CDM and WDM subhalo abundance differs significantly.

Recent analysis has shown the contribution from line-of-sight subhalos is substantial (Despali et al. 2017). Since inclusion of the line-of-sight structure will likely improve our projected constraints, we interpret our results as understated limits on the power of substructure lensing. We leave the extension of our method to include line-of-sight structure to future work.

Comparing the posterior probability distributions in Figures 7 and 8, although there is some degeneracy between the normalization and the half-mode mass these parameters can be constrained simultaneously with ≈ 100 lenses.¹⁰ The normalization has important implications for dark matter and baryonic physics through its connection to total halo mass and tidal stripping of subhalos, respectively, and thus potentially will provide an important constraint for theoretical models.

Our method hinges on accurate measurement of image flux ratios and controlling for systematic errors in their modeling. To quantify the impact of small uncertainties in these observables, we simulate observations with different errors applied to the image fluxes to erase information at the 2%, 4%, and 8% level, and find the projected constraints are extremely sensitive to loss of information at this level. Case in point: the difference between perfect models and perfect measurements, and an observational scenario with 2% uncertainties in image fluxes is an order of magnitude in m_{hm} . To achieve the required level of measurement precision, we will need flux ratios computed from the narrow-line emission of the background quasar (Nierenberg et al. 2014, 2017), which yield measurements accurate to 4 – 6% in flux, and are resilient to microlensing. The presence of systematic errors in the modeling can also be mitigated by restricting analysis to deflectors with high velocity dispersions and no complicated morphological features like stellar disks, assuring that dark substructure dominates as the source of flux ratio anomaly computed with respect to simple SIE macromodels. Alternatively, deep imaging of the deflector and its stellar mass distribution may enable the construction of lens models that map the luminous structure of the deflector in detail, as was done in Hsueh et al. (2016a, 2017b). For these more complicated systems, additional observable information in the

form of deep imaging is required to constrain the mass distribution of the deflector.

Finally, we note that our method accommodates any arbitrary dark matter model, provided it specifies the form of the subhalo mass function and the density profiles of individual substructures. Possible extensions of our method can explore subhalo populations from mixed or self-interacting dark matter (Rocha et al. 2013; Todoroki & Medvedev 2017), and models in which a fraction of the dark matter is composed of primordial black holes (e.g. Cotner & Kusenko 2017).

ACKNOWLEDGMENTS

We thank Adriano Agnello, Andrew Benson, Michael Boylan-Kolchin, Francis-Yan Cyr-Racine, Chris Fassnacht, Stacy Kim, Alex Kusenko, Phil Marshall, Leonidas Moustakas, Annika Peter, Simona Vegetti, and Dandan Xu for helpful suggestions and interesting discussions throughout the course of this project. We also thank the anonymous referee for feedback that improved the quality of the paper.

DG, TT, and SB acknowledge support by the US National Science Foundation through grant AST-1714953. CRK acknowledges support by the NSF through grant AST-1716585. This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by the UCLA Institute for Digital Research and Education’s Research Technology Group.

REFERENCES

- Abazajian K. N., 2017, preprint, ([arXiv:1705.01837](#))
- Auger M. W., Treu T., Bolton A. S., Gavazzi R., Koopmans L. V. E., Marshall P. J., Moustakas L. A., Burles S., 2010, *The Astrophysical Journal*, 724, 511
- Baltz E. A., Marshall P., Oguri M., 2009, *JCAP*, 1, 015
- Birrer S., Amara A., Refregier A., 2017, *JCAP*, 5, 037
- Bose S., Hellwing W. A., Frenk C. S., Jenkins A., Lovell M. R., Helly J. C., Li B., 2016, *MNRAS*, 455, 318
- Bullock J. S., Boylan-Kolchin M., 2017, *Annual Review of Astronomy and Astrophysics*, 55, 343
- Chatterjee S., Koopmans L. V. E., 2017, preprint, ([arXiv:1710.03075](#))
- Chen J., Kravtsov A. V., Keeton C. R., 2003, *ApJ*, 592, 24
- Chen J., Rozo E., Dalal N., Taylor J. E., 2007, *ApJ*, 659, 52
- Chiba M., 2002, *ApJ*, 565, 17
- Congdon A. B., Keeton C. R., 2005, *MNRAS*, 364, 1459
- Cotner E., Kusenko A., 2017, *PhysRevD*, 96, 103002
- Csillry K., Blum M. G., Gaggiotti O. E., Franois O., 2010, *Trends in Ecology & Evolution*, 25, 410
- Cyr-Racine F.-Y., Moustakas L. A., Keeton C. R., Sigurdson K., Gilman D. A., 2016, *PhysRevD*, 94, 043505
- Dalal N., Kochanek C. S., 2002, *ApJ*, 572, 25
- Daylan T., Cyr-Racine F.-Y., Diaz Rivero A., Dvorkin C., Finkbeiner D. P., 2017, preprint, ([arXiv:1706.06111](#))
- Despali G., Vegetti S., 2017, *MNRAS*, 469, 1997
- Despali G., Vegetti S., White S. D. M., Giocoli C., van den Bosch F. C., 2017, preprint, ([arXiv:1710.05029](#))
- Dobler G., Keeton C. R., 2006, *MNRAS*, 365, 1243

¹⁰ In fact, the differences between the low normalization and high normalization scenarios become apparent with fewer than 100 lenses, but the extent to which one can differentiate the two depends on the degree to which one controls for systematic errors in flux ratios, and the difference between the two normalizations.

- Fadely R., Keeton C. R., 2012, MNRAS, 419, 936
- Fiacconi D., Madau P., Potter D., Stadel J., 2016, ApJ, 824, 144
- Gao L., Navarro J. F., Frenk C. S., Jenkins A., Springel V., White S. D. M., 2012, MNRAS, 425, 2169
- Garrison-Kimmel S., et al., 2017, MNRAS, 471, 1709
- Gilman D., Agnello A., Treu T., Keeton C. R., Nierenberg A. M., 2017, MNRAS, 467, 3970
- Graus A. S., Bullock J. S., Boylan-Kolchin M., Nierenberg A. M., 2017, preprint, ([arXiv:1710.11148](#))
- Hahn C., Vakili M., Walsh K., Hearin A. P., Hogg D. W., Campbell D., 2017, MNRAS, 469, 2791
- Han J., Cole S., Frenk C. S., Jing Y., 2016, MNRAS, 457, 1208
- Herbel J., Kacprzak T., Amara A., Refregier A., Bruderer C., Nicola A., 2017, JCAP, 8, 035
- Hezaveh Y., Dalal N., Holder G., Kisner T., Kuhlen M., Perreault Levasseur L., 2016a, JCAP, 11, 048
- Hezaveh Y. D., et al., 2016b, ApJ, 823, 37
- Hsueh J.-W., Fassnacht C. D., Vegetti S., McKean J. P., Spingola C., Auger M. W., Koopmans L. V. E., Lagattuta D. J., 2016a, preprint, ([arXiv:1601.01671](#))
- Hsueh J.-W., Fassnacht C. D., Vegetti S., McKean J. P., Spingola C., Auger M. W., Koopmans L. V. E., Lagattuta D. J., 2016b, MNRAS, 463, L51
- Hsueh J.-W., Despali G., Vegetti S., Xu D., Fassnacht C. D., Metcalf R. B., 2017a, preprint, ([arXiv:1707.07680](#))
- Hsueh J.-W., et al., 2017b, MNRAS, 469, 3713
- Inoue K. T., 2016, MNRAS, 461, 164
- Jiang F., van den Bosch F. C., 2017, MNRAS, 472, 657
- Kamada A., Kaplinghat M., Pace A. B., Yu H.-B., 2017, Phys. Rev. Lett., 119, 111102
- Keeton C. R., 2011, GRAVLENS: Computational Methods for Gravitational Lensing, Astrophysics Source Code Library ([ascl:1102.003](#))
- Keeton C. R., Moustakas L. A., 2009, ApJ, 699, 1720
- Keeton C. R., Gaudi B. S., Petters A. O., 2003, ApJ, 598, 138
- Keeton C. R., Gaudi B. S., Petters A. O., 2005, The Astrophysical Journal, 635, 35
- Kim S. Y., Peter A. H. G., Hargis J. R., 2017, preprint, ([arXiv:1711.06267](#))
- Klypin A., Kravtsov A. V., Valenzuela O., Prada F., 1999, ApJ, 522, 82
- Koopmans L. V. E., 2005, MNRAS, 363, 1136
- Kusenko A., 2009, Phys. Rept., 481, 1
- Li R., Frenk C. S., Cole S., Gao L., Bose S., Hellwing W. A., 2016, MNRAS, 460, 363
- Lintusaari J., Gutmann M. U., Dutta R., Kaski S., Corander J., 2017, Systematic Biology, 66, e66
- Lovell M. R., Frenk C. S., Eke V. R., Jenkins A., Gao L., Theuns T., 2014a, MNRAS, 439, 300
- Lovell M. R., Frenk C. S., Eke V. R., Jenkins A., Gao L., Theuns T., 2014b, MNRAS, 439, 300
- Ludlow A. D., Bose S., Angulo R. E., Wang L., Hellwing W. A., Navarro J. F., Cole S., Frenk C. S., 2016, MNRAS, 460, 1214
- MacLeod C. L., Jones R., Agol E., Kochanek C. S., 2013, ApJ, 773, 35
- Macciò A. V., Dutton A. A., van den Bosch F. C., 2008, MNRAS, 391, 1940
- Mao S., Schneider P., 1998, MNRAS, 295, 587
- Menci N., Sanchez N. G., Castellano M., Grazian A., 2016, ApJ, 818, 90
- Metcalf R. B., Madau P., 2001, ApJ, 563, 9
- Moustakas L. A., Metcalf R. B., 2003, MNRAS, 339, 607
- Munshi F., Brooks A. M., Applebaum E., Weisz D. R., Governato F., Quinn T. R., 2017, preprint, ([arXiv:1705.06286](#))
- Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
- Nierenberg A. M., Treu T., Wright S. A., Fassnacht C. D., Auger M. W., 2014, MNRAS, 442, 2434
- Nierenberg A. M., Treu T., Menci N., Lu Y., Torrey P., Vogelsberger M., 2016, MNRAS, 462, 4473
- Nierenberg A. M., et al., 2017, preprint, ([arXiv:1701.05188](#))
- Oguri M., Marshall P. J., 2010, MNRAS, 405, 2579
- Pullen A. R., Benson A. J., Moustakas L. A., 2014, ApJ, 792, 24
- Robin A. C., Reylé C., Fliri J., Czekaj M., Robert C. P., Martins A. M. M., 2014, AA, 569, A13
- Rocha M., Peter A. H. G., Bullock J. S., Kaplinghat M., Garrison-Kimmel S., Oñorbe J., Moustakas L. A., 2013, MNRAS, 430, 81
- Rubin D. B., 1984, Ann. Statist., 12, 1151
- Schneider A., Smith R. E., Macciò A. V., Moore B., 2012, MNRAS, 424, 684
- Schneider A., Smith R. E., Reed D., 2013, MNRAS, 433, 1573
- Schneider A., Trujillo-Gomez S., Papastergis E., Reed D. S., Lake G., 2017, MNRAS, 470, 1542
- Shankar F., et al., 2017, ApJ, 840, 34
- Sonnenfeld A., Gavazzi R., Suyu S. H., Treu T., Marshall P. J., 2013, ApJ, 777, 97
- Sonnenfeld A., Treu T., Marshall P. J., Suyu S. H., Gavazzi R., Auger M. W., Nipoti C., 2015, ApJ, 800, 94
- Springel V., et al., 2008, MNRAS, 391, 1685
- Todoroki K., Medvedev M. V., 2017, preprint, ([arXiv:1711.11078](#))
- Turner B. M., Zandt T. V., 2012, Journal of Mathematical Psychology, 56, 69
- Vegetti S., Koopmans L. V. E., Bolton A., Treu T., Gavazzi R., 2010, Monthly Notices RAS, 408, 1969
- Vegetti S., Lagattuta D. J., McKean J. P., Auger M. W., Fassnacht C. D., Koopmans L. V. E., 2012, Nat, 481, 341
- Vegetti S., Koopmans L. V. E., Auger M. W., Treu T., Bolton A. S., 2014, MNRAS, 442, 2017
- Vegetti S., Despali G., Lovell M. R., Enzi W., 2018, preprint, ([arXiv:1801.01505](#))
- Viel M., Lesgourgues J., Haehnelt M. G., Matarrese S., Riotto A., 2005, PhysRevD, 71, 063534
- Viel M., Becker G. D., Bolton J. S., Haehnelt M. G., 2013, PhysRevD, 88, 043502
- Vogelsberger M., et al., 2014, MNRAS, 444, 1518
- Vogelsberger M., Zavala J., Cyr-Racine F.-Y., Pfrommer C., Bringmann T., Sigurdson K., 2016, MNRAS, 460, 1399
- Weyant A., Schafer C., Wood-Vasey W. M., 2013, ApJ, 764, 116
- Wong K. C., et al., 2017, MNRAS, 465, 4895
- Xu D. D., et al., 2009, MNRAS, 398, 1235
- Xu D. D., Mao S., Cooper A. P., Gao L., Frenk C. S., Angulo R. E., Helly J., 2012, MNRAS, 421, 2553

Xu D., Sluse D., Gao L., Wang J., Frenk C., Mao S., Schneider P., Springel V., 2015, MNRAS, 447, 3189
 de Blok W. J. G., 2010, Advances in Astronomy, 2010, 789293

APPENDIX A: MORE ON GENERATING SUBSTRUCTURE REALIZATIONS

A1 Spatial distribution and truncation

As discussed in Section 3, we tidally truncate the subhalos according to their 3-d position in the halo r_{3d} according to

$$r_t = \left(\frac{m_s r_{3d}^2}{2\Sigma_{\text{crit}} R_{\text{Ein}}} \right)^{\frac{1}{3}} \quad (\text{A1})$$

(see Cyr-Racine et al. 2016), where m_s refers to M_{200} at redshift $z_d = 0.5$. Although strong lensing quantities typically live in projection in the lens plane (omitting the line of sight), the third spatial dimension enters through the truncation, via the 3-d position $r_{3d} = \sqrt{z^2 + r_{2d}^2}$.

To include this effect in our simulations, we begin by noting that in projection, in the inner portions of a galactic halo where strong lensing takes place, subhalos appear distributed uniformly in two dimensions (Xu et al. 2015). We therefore assign each subhalo a projected position r_{2d} with a spatially uniform probability density out to $R_{\text{max}} = 18.6\text{kpc}$, or 3 arcseconds at the lens redshift.

To obtain the 3-dimensional z coordinate for a subhalo, we start with a two dimensional distribution that is uniform (to a very good approximation) with in 18.6 kpc

$$p(r_{2d}|r_c) \propto \left(1 + \frac{r_{2d}^2}{r_c^2} \right)^{-1} \quad (\text{A2})$$

with $r_c = 75\text{kpc}$. We then de-project this 2-d density into the third dimension to obtain a density for the z coordinate, out to a maximum 3-d radius $R = 250\text{kpc}$. The corresponding distribution for the z coordinate, given a 2-d position, becomes

$$p(z|r_{2d}, r_c) \propto \left(1 + \frac{z^2 + r_{2d}^2}{r_c^2} \right)^{-1.5}. \quad (\text{A3})$$

We note that this has the same asymptotic form $\propto r^{-3}$ as an NFW profile for large z . The z coordinate affects the lensing only indirectly through the truncation radius. For an NFW profile, this does not significantly impact the image magnification, as this observable is principally determined by the central density which is unchanged. We verify that our truncation scheme consistently yields $r_t > r_s$. Typical values for $\frac{r_t}{r_s}$ range between 5-30, depending on the concentration of the subhalo.

A2 Mass Function

Focusing first on the form of the mass function $\frac{dN}{dm}$, numerical simulations of cold dark matter halos (Springel et al. 2008; Gao et al. 2012; Fiacconi et al. 2016) suggest a scale free mass function $\frac{dN}{dm_s} \propto m_s^{-1.9}$ for $10^6 \leq \frac{m_s}{M_\odot} \leq 10^{10}$. We acknowledge that tidal disruption may alter this prediction significantly in the inner portions of galactic halos, but we

do not address this concern here, as our principle aim is to demonstrate the method rather than focus on the most realist mass function.

When normalizing the subhalo mass function, we wish to compare CDM realizations with WDM realizations with the same amplitude at mass scales far above the half-mode mass m_{hm} . To accomplish this, we start with a scale free CDM mass function

$$\frac{dN_{\text{cdm}}}{dm_s} = A_0 \left(\frac{m_s}{M_\odot} \right)^{-\alpha} \quad (\text{A4})$$

taking $\alpha = 1.9$. In the regions of dark matter halos, the spatial distribution of substructure is approximately uniform in projection (Xu et al. 2015). Uniformly distributing subhalos in a plane with radius R_{max} , we relate the substructure convergence at the Einstein radius, κ_{sub} , to the mass in substructure between m_L and $m_s \leq m_H$

$$\Sigma_{\text{crit}} \kappa_{\text{sub}} \pi R_{\text{max}}^2 = \int_{m_L}^{m_H} m_s \frac{dN_{\text{cdm}}}{dm_s} dm_s. \quad (\text{A5})$$

This yields the normalization A_0 in terms of κ_{sub} , and the mean number of subhalos

$$\begin{aligned} A_0 &= \frac{(2-\alpha) \pi R_{\text{max}}^2 \Sigma_{\text{crit}} \kappa_{\text{sub}}}{M_\odot^\alpha (m_H^{2-\alpha} - m_L^{2-\alpha})} \\ \langle N_s \rangle &= \frac{A_0}{1-\alpha} (m_H^{1-\alpha} - m_L^{1-\alpha}) M_\odot^\alpha. \end{aligned} \quad (\text{A6})$$

We then draw N_s subhalos from a Poisson distribution with average value $\langle N_s \rangle$.

A3 Normalization

The normalization of the subhalo mass function depends both on the accretion history and evolution of subhalos in the lens halo, and the effects of baryonic physics in the central regions of the halo. The former effect, the accretion history of dark matter halos as a function of halo mass, has been well studied, and here we adopt the result of (Han et al. 2016) and assume a total surviving halo mass fraction of $f_{\text{sub,halo}} \sim 6\%$. There is scatter in the predictions depending on halo accretion history and redshift which can raise this value by as much as a factor of 2 (e.g. Fiacconi et al. 2016; Jiang & van den Bosch 2017), however we conservatively adopt the lower normalization for this estimate, and lower the overall subhalo mass normalization for both the extreme and no tidal disruption cases by 30%, bringing the assumed total halo mass fraction in substructure to $\sim 4\%$.

The effect of baryonic physics is more uncertain. Recent state of the art hydrodynamic simulations indicate that in a Milky Way mass host halo, the central disk may destroy all subhalos within the central 20 kpc of the host, in addition to reducing the total number of subhalos by $\sim 30\%$ compared to a dark matter only run (Garrison-Kimmel et al. 2017). In these simulations, the destruction appears independent of subhalo mass.

In order to bracket the range of possibilities and to demonstrate how tidal disruption in the central region would affect our inference, we consider two scenarios. In the first, we assume that dark matter subhalos follow the NFW density profile of the host (Han et al. 2016), which is seen in dark matter only simulations. In the second case we mimic

the effects of central tidal disruption by assuming that the destruction radius scales with the host scale radius. Dedicated simulations of lens mass halos will be necessary to calibrate this effect in more detail.

For a lower bound on the normalization, we assume that all tidal disruption destroys all subhalos within the scale radius of the host (for a typical lens halo of mass $\sim 10^{13} M_\odot$ this radius is ~ 150 kpc). For an upper bound, we assume that the subhalo number density follows the density profile of the host. In each case, we compute the projected mass density in substructure along the longitudinal virial radius of the host inside a projected cylinder of 18 kpc (~ 3 arcseconds at $z = 0.5$), and obtain $(\kappa_{\text{sub}}) = 0.002$ (0.01) in the case of extreme (minimal) tidal stripping. These values correspond to values of $A_0 = 8$ (40) $\times 10^7 M_\odot^{-1}$.

A4 Extension to WDM

The half mode mass m_{hm} corresponds to a characteristic length scale at which the linear matter power spectrum with pure WDM is damped with respect to that of CDM by one-half. For details, see e.g. Viel et al. (2005); Schneider et al. (2012). We wish to compare a range of WDM subhalo populations with varying m_{hm} according to the mass function

$$\frac{dN_{\text{wdm}}}{dm_s} = \frac{dN_{\text{cdm}}}{dm_s} \left(1 + \frac{m_{\text{hm}}}{m_s}\right)^{-1.3} \quad (\text{A7})$$

while preserving the amplitude of the mass function for masses high above m_{hm} to isolate the effect of m_{hm} from the normalization A_0 . To do this, we first generate subhalos according to Equation A5. This results in scale free mass function, which we deplete by removing subhalos probabilistically with probability

$$P \propto \left(1 + \frac{m_{\text{hm}}}{m_s}\right)^{-1.3}. \quad (\text{A8})$$

This yields substructure populations obeying the curves plotted in Figure 1.

APPENDIX B: SENSITIVITY OF CUSP, FOLD, AND CROSS CONFIGURATIONS

The response of an image magnification to small scale structure is heightened if it lies close to a critical curve. Similarly, images close to one another may be affected by the same substructures, introducing a correlation between flux anomalies in different images.

Cusp (fold) configurations are characterized by three (two) images straddling the critical curve, and by the three (two) images close in proximity to each other. One therefore expects a ranking in sensitivity to substructure of cusp, fold, cross, in descending order. In Figure B1, we show that this is indeed the case. Interestingly, the degeneracy between a warm mass function with a high normalization, and a cold mass function with low normalization is reduced in cusp and fold configurations compared to crosses.

To identify image configurations, we adopt the following classification scheme based on the Einstein radius R_{Ein} and the image separations. If the smallest image separation is greater than $0.7R_{\text{Ein}}$, the lens is immediately classified as

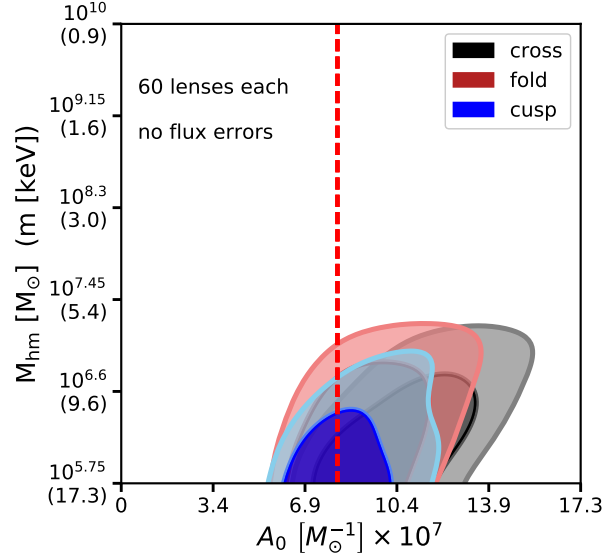


Figure B1. Response of cusp, fold and cross image configurations to small scale structure. For a fixed number of lenses, cusps yield the strongest constraints, followed by folds and crosses.

a cross. If the second largest separation is $< 1.2R_{\text{Ein}}$, we classify it as a cusp, and otherwise it is a fold.

APPENDIX C: USE OF OTHER SUMMARY STATISTICS

The summary statistic in Equation 11 is closely the likelihood, or a χ^2 value. It penalizes models which do not reproduce the anomalies observed in the data in the correct images. There is, however, no ‘correct’ choice of summary statistic, only ones that perform better than others.

We experimented with using other statistics, including R_{cusp} and R_{fold} (see Keeton et al. 2003; Keeton et al. 2005). An advantageous feature of these parameters is that they can be computed directly from the observed image fluxes, and do not rely on a lens model to identify anomalies. After experimenting with other summary statistics, however, we find the summary statistic in Equation 11 yields the strongest constraints, because it does not discard information by adding and subtracting fluxes from different images.

As an example, consider the summary statistic

$$S(\mathbf{f}_n, \mathbf{f}_n^*) = \left| \sum_{i=1}^3 (f_{n(i)} - f_{n(i)}^*) \right|. \quad (\text{C1})$$

Using this equation instead of Equation 11, we perform an inference on the same simulation as in Figure 6. The results are shown in Figure C1. Equation 11 performs better.

APPENDIX D: CONVERGENCE OF ABC SIMULATIONS AND POSTERIOIRS

To assess convergence of the ABC algorithm, we compare two inferences made on the same data set in which one has half the number of realizations as the other. To produce the

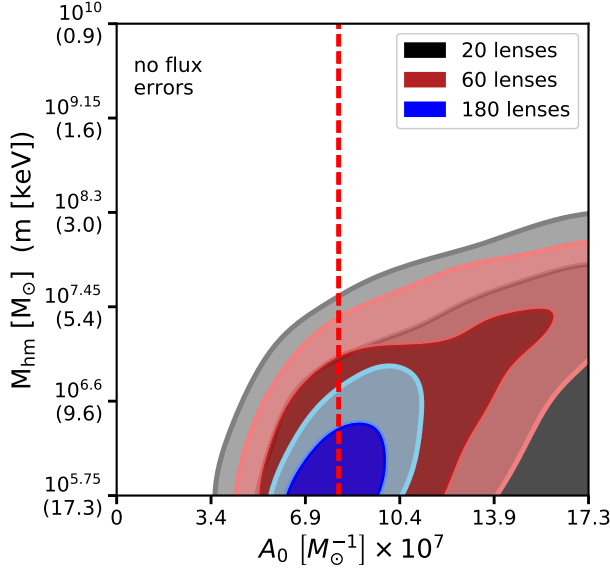


Figure C1. Inference made using the statistic in defined in Equation C1, rather than Equation 11.

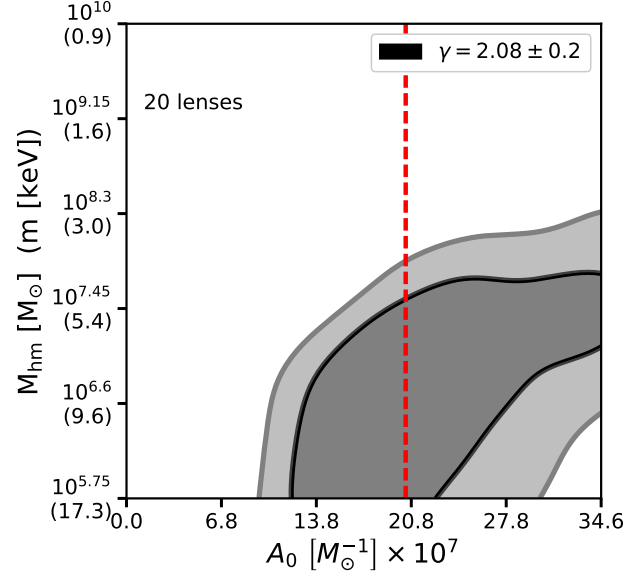


Figure E1. We test the effect of an incorrectly parameterizing the macromodel by generating mock data with non-isothermal ellipsoids, and modeling the data with an SIE in the forward model. The power law slope γ is sampled from a Gaussian with mean 2.08 and variance 0.2.

APPENDIX E: FITTING SIE MACROMODELS TO DATASETS BUILT WITH NON-ISOTHERMAL POWER LAWS

We test the effect of incorrectly parameterizing the macromodel by created datasets with non-isothermal power law ellipsoids, and fitting them with SIEs in the forward model. When generating mock lenses, we sample power law slopes from a Gaussian distribution offset from isothermal at $\gamma = 2.08 \pm 0.2$, consistent with the findings of Shankar et al. (2017), who find the power law slopes for massive ellipticals are consistently steeper than isothermal. The results of this exercise are shown in Figure E1. The effect of varying the power law slope in the data is seen to not significantly degrade the inference on A_0 and m_{hm} , and therefore does not affect the precision of our forecast statements.

By completely neglecting the presence of this systematic, we exaggerate its potential bias in the inference. In practice, this systematic should be properly dealt with by sampling different power laws slopes in the forward model. A prior γ may be constructed on a lens by lens basis based on measurements of the central velocity dispersion, if available.

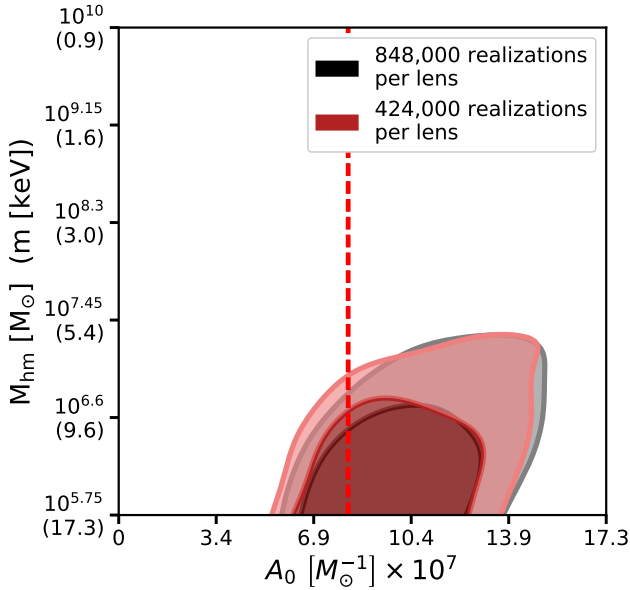


Figure D1. A convergence test in which we discard half of the realizations before applying the acceptance criterion. Each distribution is composed of the samples from $\mathbf{q}_{\text{sub}}$ associated with the lowest 1,800 summary statistics.

black distribution, we retain the draws from $\mathbf{q}_{\text{sub}}$ associated with the lowest 1,800 summary statistics. We then discard half of the realizations, and reject all but the lowest 1,800 summary statistics from the depleted simulation. Under-sampling by a factor of two, we recover the same bounds on A_0 and m_{hm} to a high degree of precision, indicating the inference made with ≈ 2000 realizations per proposal $\mathbf{q}_{\text{sub}}$ has converged.