Draft version August 26, 2019
Typeset using LaTeX **twocolumn** style in AASTeX62

Mining for Dark Matter Substructure: Inferring subhalo population properties from strong lenses with machine learning

Johann Brehmer,[1,2] Siddharth Mishra-Sharma,[1] Joeri Hermans,[3] Gilles Louppe,[3] and Kyle Cranmer[1,2]

[1] *Center for Cosmology and Particle Physics, Department of Physics, New York University, 726 Broadway, New York, NY 10003, USA*
[2] *Center for Data Science, New York University, 60 Fifth Ave, New York, NY 10011, USA*
[3] *Montefiore Institute, University of Liège, Belgium*

ABSTRACT

The subtle and unique imprint of dark matter substructure on extended arcs in galaxy-galaxy strong lenses contains a wealth of information about the properties and distribution of dark matter on small scales and, consequently, about the underlying particle physics. Teasing out this effect poses a significant challenge however due to the high dimensionality of the underlying latent space associated with a large number of dark matter subhalos. We apply recently-developed simulation-based techniques to the problem of substructure inference in galaxy-galaxy strong lenses. By leveraging additional information extracted from the simulator, these methods can be used to train neural networks to estimate likelihood ratios associated with population-level parameters characterizing the distribution of substructure. We show through proof-of-principle application to simulated data how these methods can provide an efficient and principled way to infer substructure properties by simultaneously analyzing the large lens samples deliverable by upcoming surveys.

Keywords: strong gravitational lensing (1643) — gravitational lensing (670) — nonparametric inference (1903) — astrostatistics techniques (1886) — cosmology (343) — dark matter (353)

Contents

[JB: Better way to mark a joint first authorship?]

Corresponding author: Johann Brehmer
johann.brehmer@nyu.edu

Corresponding author: Siddharth Mishra-Sharma
sm8383@nyu.edu

1. INTRODUCTION

Dark matter (DM) accounts for nearly a quarter of the energy budget of the Universe, and pinning down its fundamental nature and interactions is one of the most pressing problems in cosmology and particle physics today. Despite an organized effort to do so through terrestrial (Aprile et al. 2018; Cui et al. 2017; Akerib et al. 2017), astrophysical (Albert et al. 2017; Chang et al. 2018; Lisanti et al. 2018), and collider searches (Aaboud et al. 2019; Sirunyan et al. 2017), no conclusive evidence

of interactions between the Standard Model (SM) and dark matter exists to-date.

An alternative and complementary approach involves studying dark matter directly through its irreducible gravitational interactions. The concordance Cold Dark Matter (CDM) framework of non-relativistic, collisionless dark matter particles provides an excellent description of the observed distribution of matter on large scales. However, many well-motivated models predict deviations from CDM on smaller scales. Fundamental dark matter microphysical properties, such as its particle mass and self-interaction cross-section, can imprint themselves onto its macroscopic distribution in ways that can be probed by current and future experiments (Drlica-Wagner et al. 2019). As motivating examples, theories where dark matter has a significant free-streaming length would lead to a dearth of subhalos at lower masses ($\lesssim 10^9\,\mathrm{M}_\odot$) (Bond & Szalay 1983; Bode et al. 2001; Dalcanton & Hogan 2001; Boyanovsky et al. 2008; Boyanovsky & Wu 2011), and self-interactions (Kaplinghat et al. 2016, 2014; Zavala et al. 2013; Peter et al. 2013; Vogelsberger et al. 2012, 2019; Kahlhoefer et al. 2019) or dissipative dynamics (Agrawal & Randall 2017; Agrawal et al. 2017; Buckley & DiFranzo 2018; Fan et al. 2013; Vogelsberger et al. 2016) in the dark sector would modify the structure of the inner core of subhalos as compared to CDM predictions.

There exist several avenues for probing the structure of dark matter on small scales. While the detection of ultrafaint dwarf galaxies through the study of stellar overdensities and kinematics (Bechtol et al. 2015; Koposov et al. 2008, 2015) can be used to make statements about the underlying dark matter properties, theoretical uncertainties in the connection between stellar and halo masses (Wechsler & Tinker 2018) and the effect of baryons on the satellite galaxy population (Brooks 2018; Fitts et al. 2018; Garrison-Kimmel et al. 2017; Errani et al. 2017) pose a challenge. Furthermore, suppressed star-formation in smaller halos means that there exists a threshold ($\lesssim 10^8\,\mathrm{M}_\odot$) below which subhalos are expected to be mostly dark and devoid of stars (Efstathiou 1992; Fitts et al. 2017; Read et al. 2017). This makes studying the imprints of gravitational interaction the *only* viable avenue for probing substructure at smaller scales. In this spirit, the study of perturbations to the stellar phase-space distribution in cold stellar streams (Bonaca & Hogg 2018; Johnston et al. 1999; Carlberg 2012; Carlberg & Grillmair 2013; Bonaca et al. 2019), and in stellar fields in the disk and halo (Buschmann et al. 2018) have been proposed as methods to look for low-mass subhalos through their gravitational interactions in the Milky Way.

Complementary to the study of locally-induced gravitational effects, gravitational lensing has emerged as an important tool for studying the distribution of matter over a large range of scales. Locally, the use of time-domain astrometry has been proposed as a promising method to measure the distribution of local substructure through correlated, lens-induced motions on background celestial objects (Van Tilburg et al. 2018). In the extragalactic regime, galaxy-scale strong lensing systems are a laboratory for studying substructure. The presence of flux-ratio anomalies in multiply-imaged quasar lenses has been used to infer the typical abundance of substructure within galaxy-scale lenses (Dalal & Kochanek 2002; Hsueh et al. 2019; Dalal & Kochanek 2002). Lensed images of extended sources have been used to find evidence for a handful of subhalos with masses $\gtrsim 10^8\,\mathrm{M}_\odot$ (Hezaveh et al. 2016b; Vegetti et al. 2010, 2012).

A complementary approach relies on probing the collective effect of sub-threshold (*i.e*, not individually resolvable) subhalos on extended arcs in strongly lensed systems. A particular challenge here is the high dimensionality of the latent parameter space associated with the large number of subhalos and their (potentially covariant) individual as well as population properties, a consequence of which is the intractability of the likelihood of high-level substructure parameters conditional on the data. Methods based on summary statistics (Birrer et al. 2017) and studying the amplitude of spatial fluctuations on different scales through power spectra (Hezaveh et al. 2016a; Díaz Rivero et al. 2018; Diaz Rivero et al. 2018; Cyr-Racine et al. 2019; Brennan et al. 2019; Chatterjee & Koopmans 2018; Cyr-Racine et al. 2016) have been proposed as ways to reduce the dimensionality of the problem and enable substructure inference in a tractable way. Trans-dimensional techniques may also be able to efficiently map out the parameter space associated with multiple sub-threshold subhalos in these systems (Brewer et al. 2016; Daylan et al. 2018). This class of methods is well-suited to studying dark matter substructure since they can be sensitive to the *population* properties of low-mass subhalos in strongly lensed galaxies which are directly correlated with the underlying dark matter particle physics. Furthermore, near-future observatories like LSST (LSST Science Collaboration et al. 2009; Drlica-Wagner et al. 2019; Verma et al. 2019) and *Euclid* (Refregier et al. 2010) are expected to find tens of thousands of galaxy-galaxy strong lenses (Collett 2015), making substructure inference in these systems (and high-resolution followups on a subset) one of the key ways to investigate dark

matter substructure and stress-test the Cold Dark Matter paradigm in the near future. This calls for methods that can efficiently analyze large samples of lensed images to infer the underlying substructure properties with minimal loss of information stemming from dimensional reduction.

In this paper we apply a powerful recently-developed class of techniques for simulation-based inference (Brehmer et al. 2018a,b,c) to the problem of extracting high-level substructure properties from an ensemble of galaxy-galaxy strong lensing images. In contrast to traditional simulation-based (or "likelihood-free") approaches, these methods do not rely on summary statistics and instead leverage additional information extracted from the simulator in order to train neural networks that can be used to efficiently estimate the likelihood ratio. This provides an elegant bridge between machine learning and the ubiquitous likelihood ratio, which is provably the most powerful statistic for hypothesis testing (Neyman & Pearson 1933).

This paper is organized as follows. In Sec. 2 we briefly review the formalism of gravitational strong lensing and describe our simulation setup, including the assumptions we make about the population of lensed sources and host galaxies, the substructure population and observational parameters. In Sec. 3 we describe the simulation-based analysis technique used and its particular application to the problem of mining substructure properties from an ensemble of extended lensed arcs. We show a proof-of-principle application of this method to simulated data in Sec. 4 and comment on how these methods can be extended to more "realistic" scenarios in Sec. 5. We conclude in Sec. 6.

2. STRONG LENSING FORMALISM AND SIMULATION SET-UP

In strong lensing systems, the background light emission source can in general be a point-like quasar or supernova, or a faint, extended "blue" galaxy. The former results in multiple localized images on the lens plane rather than extended arc-like images, providing the ability to probe substructure over a limited region on the lens plane. For this reason, we focus our method towards galaxy-galaxy lenses — systems producing images with extended arcs — since we aim to disentangle the collective effect of a population of subhalo perturbers over multiple images. Young, blue galaxies are ubiquitous in the redshift regime $z \gtrsim 1$ and dominate the faint end of the galaxy luminosity function, resulting in a much larger deliverable sample of galaxy-galaxy strong lenses compared to quadruply- and doubly-imaged quasars/supernovae.

We now briefly review the basic strong lensing formalism before describing in turn the models for the background source, lensing galaxy and population parameters of the lens systems used in this study.

2.1. *Strong lensing formalism*

We briefly review here the mathematical formalism behind strong lensing. For more details see, *e.g.*, Keeton (2001); Schneider et al. (1992). For a mass distribution with dimensionless projected surface mass density $\kappa(\mathbf{r}) = \Sigma(\mathbf{r})/\Sigma_{\rm cr}$, where $\Sigma_{\rm cr} \equiv \frac{1}{4\pi G_N}\frac{D_{\rm s}}{D_{\rm ls}D_{\rm l}}$ is the critical lensing surface density, the two-dimensional lensing potential is given by

$$\psi(\mathbf{r}) = \frac{1}{\pi}\int \mathrm{d}\mathbf{r}'\ \ln|\mathbf{r}-\mathbf{r}'|\kappa(\mathbf{y}). \tag{1}$$

The lensed position of the source can be determined through the lens equation,

$$\mathbf{u} = \mathbf{r} - \nabla\psi(\mathbf{r}) \tag{2}$$

where $\mathbf{u}$ is the position of the source and $\nabla\psi$ is typically referred to as the deflection, which we will denote as $\boldsymbol{\phi}$ for brevity. For an extended source brightness profile $f_{\rm src}$, the final lensed image can be obtained as the the source profile evaluated on the image plane,

$$f'_{\rm src}(\mathbf{r}) = f_{\rm src}\left(\mathbf{r} - \nabla\psi(\mathbf{r})\right). \tag{3}$$

For a spherically symmetric, the radial deflection field is given by

$$\phi_r(r) = \frac{2}{r}\int_0^r \mathrm{d}r'\, r'\,\kappa(r') = \frac{1}{\pi\Sigma_{\rm cr}}\frac{M_{\rm cyl}(r)}{r} \tag{4}$$

where $M_{\rm cyl}(r)$ is the mass enclosed within a cylinder or radius r. Extension to the slightly more general case of elliptical symmetry is straightforward (see, *e.g.*, Keeton (2001)).

2.2. *Lensing host galaxy*

Cosmological N-body simulations suggest that the dark matter distribution in structures at galactic scales can be well-described by a universal, spherically symmetric NFW profile. However, strong lensing probes a region of the host galaxy much smaller than the typical virial radii of galaxy-scale dark matter halo, and the mass budget here is dominated by the baryonic bulge component of the galaxy. Taking this into account, the total mass budget of the lensing host galaxy, being early-type, can be well describe by a singular isothermal ellipsoid (SIE) profile, known as the bulge-halo conspiracy since neither the dark matter nor the baryonic components are individually isothermal. The host profile is

thus described as

$$\rho(\theta_x, \theta_y) = \frac{\sigma_v^2}{2\pi G\left(\theta_x^2/q + q\theta_y^2\right)} \tag{5}$$

where σ_v is the central 1-D velocity dispersion of the lens galaxy and q is the ellipsoid axis ratio, with $q = 1$ corresponding to a spherical profile. We explicitly denote our angular coordinates as $\{\theta_x, \theta_y\}$. The Einstein radius for this profile, giving the characteristic lensing scale, is given by

$$\theta_E = 4\pi \left(\frac{\sigma_v}{c}\right)^2 \frac{D_{ls}\left(z_l, z_s\right)}{D_s\left(z_s\right)} \tag{6}$$

where D_{ls} and D_s are respectively the angular diameter distances from the source to the lens planes and from the source plane to the observer respectively.

The deflection field for the SIE profile is given by (Keeton 2001)

$$\phi_x = \frac{\theta_E q}{\sqrt{1-q^2}} \tan^{-1}\left[\frac{\sqrt{1-q^2}\theta_x}{\psi}\right] \tag{7}$$

$$\phi_y = \frac{\theta_E q}{\sqrt{1-q^2}} \tanh^{-1}\left[\frac{\sqrt{1-q^2}\theta_y}{\psi + q^2}\right] \tag{8}$$

with $\psi \equiv \sqrt{\theta_x^2 q^2 + \theta_y^2}$.

Although the total galaxy mass (baryons + dark matter) describe the macro lensing field, for the purposes of describing substructure we require being able to map the measure properties of an SIE lens onto the properties of the host dark matter halo. To do this, we relate the central stellar velocity dispersion σ_v to the mass M_{200} of the host dark matter halo. Zahid et al. (2018) derived a tight correlation between σ_v and M_{200}, modeled as

$$\log\left(\frac{M_{200}}{10^{12}\,\mathrm{M}_\odot}\right) = \alpha + \beta\left(\frac{\sigma_v}{100\,\mathrm{km\,s^{-1}}}\right) \tag{9}$$

with $\alpha = 0.09$ and $\beta = 3.48$. We model the host dark matter halo with a Navarro-Frenk-White (NFW) profile (Navarro et al. 1996, 1997)

$$\rho(r) = \frac{\rho_s}{(r/r_s)\left(1 + r/r_s\right)^2} \tag{10}$$

where ρ_s and r_s are the scale density and scale radius, respectively. The halo virial mass M_{200} describes the total mass contained with the virial radius r_{200}, defined as the radius within which the mean density is 200 times the critical density of the universe and related to the scale radius through the concentration parameter $c_{200} \equiv r_{200}/r_s$. Thus, an NFW halo is completely described by the parameters $\{M_{200}, c_{200}\}$. We use the concentration-mass relation from Sánchez-Conde & Prada (2014) assuming a log-normal distribution for c_{200} around the median inferred value given by the relation with scatter 0.15 dex.

The spherically-symmetric deflection for an NFW perturber is given by (Keeton 2001)

$$\phi_r = 4\kappa_s r_s \frac{\ln(x/2) + \mathcal{F}(x)}{x} \tag{11}$$

where $x \equiv r/r_s, \kappa_s \equiv \rho_s\, r_s/\Sigma_{\mathrm{cr}}$ with Σ_{cr} the critical surface density, and

$$\mathcal{F}(x) = \begin{cases} \frac{1}{\sqrt{x^2-1}} \tan^{-1}\sqrt{x^2-1} & (x > 1) \\ \frac{1}{\sqrt{1-x^2}} \tanh^{-1}\sqrt{1-x^2} & (x < 1) \\ 1 & (x = 1). \end{cases} \tag{12}$$

We described the population parameters we use to model the host velocity dispersion (and thus its Einstein radius and dark matter halo mass) in Secs. 2.5 and 2.6 below.

2.3. *Lensing substructure*

The ultimate goal of our method is to characterize the substructure population in strong lenses. Here we describe our procedure to model the substructure contribution to the lensing signal. Understanding the expected abundance of substructure in galaxies over a large range of epochs is complex undertaking and an active area of research. Properties of individual subhalos (such as their density profiles) as well as those that describe their population (such as the mass and spatial distribution) are strongly affected by their host environment, and accurately modeling all aspects of subhalo evolution and environment is beyond the scope of this paper. Instead, we use simple physically justifiable assumptions to model the substructure contributions in order to highlight the broad methodological points associated with the application of our method.

ΛCDM, often called the standard model of cosmology, predicts a scale-invariant power spectrum of primordial fluctuations and the existence of substructure over a broad range of masses with equal contribution per logarithmic mass interval. We parameterize the distribution of subhalo masses m_{200} in a given host halo of mass M_{200} — the subhalo mass function — as a power law distribution with a linear dependence on the host halo mass,

$$\frac{dn}{dm_{200}} = \alpha \cdot M_{200} \cdot m_{200}^{\beta} \tag{13}$$

where α encodes the overall substructure abundance, with larger α corresponding to more substructure, and the slope β encodes the relative contribution of subhalos at different masses, with more negative β corresponding to a steeper slope with more low-mass subhalos.

Theory and simulations within the framework of ΛCDM predict a slope $\beta \sim -0.9$ (Springel et al. 2008; Madau et al. 2008), giving a nearly scale-invariant spectrum of subhalos, which we assume in our fiducial setup. We parameterize the overall subhalo abundance α through the mass fraction contained in subhalos, $f_{\rm sub}$, defined as the fraction of the total dark matter halo mass contained in bound substructure in a given mass range. We have

$$f_{\rm sub} = \frac{\int_{m_{200,\rm min}}^{m_{200,\rm max}} dm_{200}\, m_{200}\, \frac{dn}{dm_{200}}}{M_{200}} \qquad (14)$$

For a given $\{f_{\rm sub}, \beta\}$ and host halo mass M_{200}, this can be used to determine α in Eq. 13. The linear scaling of the subhalo mass function with the host halo mass M_{200} in Eq. 13 is additionally described in Han et al. (2016); Despali & Vegetti (2017). In our fiducial setups, we take the minimum mass $m_{200,\rm min} = 10^7\,{\rm M}_\odot$ and $m_{200,\rm max} = 0.01\,M_{200}$ (Despali & Vegetti 2017; Hiroshima et al. 2018), and corresponding fiducial substructure fraction in this range of 5%, roughly consistent with Hiroshima et al. (2018); Hsueh et al. (2019); Dalal & Kochanek (2002) within our considered mass range.

With all parameters of the subhalo mass function specified, the total number $n_{\rm tot}$ of subhalos expected within the virial radius R_{200} of the host halo can be inferred as $\int_{m_{200,\rm min}}^{m_{200,\rm max}} dm_{200}\, \frac{dn}{dm_{200}}$. Strong lensing probes a region much smaller this scale — the typical Einstein radii for the host deflector are much smaller than the virial radius of the host dark matter halos. In order to obtain the expected number of subhalos within the lensing observations region of interest, we scale the total number of subhalos obtained from the above procedure by the ratio of projected mass within our region of interest $\theta_{\rm ROI}$ and the host halo mass M_{200} as follows. We assume the subhalos to be distributed in number density following the host NFW dark matter profile. In this case, the NFW enclosed mass function is $M_{\rm enc}(x) = M_{200}\,[\ln(x/2) + \mathcal{F}(x)]$ (Keeton 2001), where x is the angular radius in units of the virial radius, $x \equiv \theta/\theta_s$ and $\mathcal{F}(x)$ is given by Eq. 12 above. The number of subhalos within our ROI is thus obtained as $n_{\rm ROI} = n_{\rm tot}\,[\ln(x_{\rm ROI}/2) + \mathcal{F}(x_{\rm ROI})]$. We conservatively take the lensing ROI to enclose a region of angular size twice the Einstein radius of the host halo, $\theta_{\rm ROI} = 2\cdot\theta_E$.

Since strong lensing probes the line-of-sight distribution of subhalos within the host, their projected spatial distribution is approximately uniform within the lensing ROI (Despali & Vegetti 2017). We thus distribute subhalos uniformly within our ROI. The density profile of subhalos is assumed to be NFW and given by Eq. 10, with associated lensing properties as described and the concentration inferred from the relation modeled in Sánchez-Conde & Prada (2014).

We finally emphasize that we do not intent to capture all of the intricacies of the subhalo distribution, such as the effects of baryonic physics, tidal distruption of subhalos in proximity to the center of the host and redshift evolution of host as well as substructure properties. Although our description can be extended to take these into account, their precise characterization and effect is still subject to large uncertainties, and our simple model above captures the essential physics for demonstration purposes.

2.4. *Background source*

We model the emission from background source galaxies using a Sérsic profile, with the surface brightness given by

$$f_{\rm src}(\theta_r) = f_e \exp\left\{-b_n\left[\left(\frac{\theta_r}{\theta_{r,e}}\right)^{1/n} - 1\right]\right\}, \qquad (15)$$

where $\theta_{r,e}$ is the effective circular half-light radius, n is the Sérsic index, and b_n is a factor depending on n that ensures that $\theta_{r,e}$ contains half the total intensity from the source galaxy, given by (Ciotti & Bertin 1999)

$$b_n \approx 2n - \frac{1}{3} + \frac{4}{405n} + \frac{46}{25515n^2} + \frac{131}{1148175n^3} - \frac{2194697}{30690717750n^4}.$$

We assume $n = 1$ for the source galaxies, corresponding to a flattened exponential profile and consistent with expectation for blue-type galaxies at the relevant redshifts. f_e encodes the flux at half-light radius, which can be mapped onto the total flux (or magnitude) associated with a given galaxy.

The total unlensed magnitude M (in a given band) of a galaxy can be mapped on to f_e as follows. For a detector with zero-point magnitude M_0, which specifies the magnitude of a source giving 1 count ${\rm s}^{-1}$ in expectation, by definition the total counts are given by $S_{\rm tot} = 10^{0.4(M-M_0)}$. Requiring the half-light radius to contain half the expected counts, for $n = 1$ we have the relation $f_e \approx 0.526\, t_{\rm exp} S_{\rm tot}/(2\pi\theta_{r,e}^2)$ in counts arcsec^{-2}, where $t_{\rm exp}$ is the exposure length.

Treatment of the other Sérsic parameters, in particular the total emission and half-light radius, in the context of population studies is described in Secs. 2.5 and 2.6 below.

2.5. *Observational considerations*

A noted above, our method is best-suited to analyzing a statistical sample of strong lenses to search for substructure, such as those that are expected to be obtained

in the near future with optical telescopes like *Euclid* and LSST. Given the challenges associated with the precise characterization of such a sample at the present time, we describe here the observational characteristics we assume in order to build up training and testing samples to validate our inference techniques.

We largely follow the description in Collett (2015), and use the associated *LensPop* package, to characterize our mock observations. In particular, we use the detector configuration for *Euclid*, assuming a zero-point magnitude $m_{\rm AB} = 25.5$ in the single optical VIS passband, pixel size 0.1 arcsec, a Gaussian point spread function (PSF) with FWHM 0.18 arcsec, individual exposures with exposure time 1610 s, an isotropic sky background with magnitude 22.8 arcsec^{-2} in the detector passband.

These properties, in particular the exposure, sky background and PSF shape are expected to vary somewhat across the lens sample. Additionally, a given region may be imaged by multiple exposures over a range of color bands. Although these variations can easily be incorporated into our analysis, modeling this is beyond the scope of this study. We briefly comment on how this information can be taken into account later.

2.6. *Population statistics of the lens and source samples*

The fact that the strong lens population is expected to be dominated by higher-redshift ($z \gtrsim 1$) blue source galaxies lensed by intermediate-redshift ($z \sim 0.5$–1) elliptical galaxies presents significant challenges for quantifying the lens population obtainable with future observations. Specifically, planned ground-based surveys like LSST and space telescopes like *Euclid* present complementary challenges for delivering images of strong lensing systems suitable for substructure studies. LSST is expected to image in six bands, allowing efficient source selection and distinguishing source and lens emission, but at the cost of lower resolution by virtue of being a ground-based instrument. *Euclid* imaging is expected be much higher in resolution but with a single optical passband (VIS). Near-IR imaging from WFIRST may deliver a high-resolution, multi-wavelength dataset that is more suitable for substructure studies, although the lens and source populations may differ from those probed by optical telescopes.

In light of these uncertainties, we confine ourselves to a setting where the main methodological points can be made without detailed modeling of the detector capabilities and the deliverable lensing dataset, which is outside of the scope of the current paper. For concreteness, we simulate a sample of lenses with a simplified subset of host galaxy properties consistent with those deliverable by *Euclid* as modeled by Collett (2015). In particular, we assume spherical lenses, with ellipticity parameter $q = 1$ in Eq. 5. We draw the central 1-D velocity dispersions σ_v of host galaxies from a normal distribution with mean 225 km s^{-1} and standard deviation 50 km s^{-1}. Equation 9 the results of Zahid et al. (2018) are used to map the drawn σ_v to a dark matter halo mass M_{200}, and the host Einstein radius is analytically inferred with Eq. 6.

We draw the lens redshifts z_l from a log-normal distribution with mean 0.56 and scatter 0.25 dex, discarding lenses with $z_l > 1$ as these tend to have a small angular size over which substructure perturbations are relevant. The source offsets θ_x and θ_y are drawn from a normal distribution with zero mean and standard deviation 0.2. These are consistent with the lens sample generated from the *LensPop* code packaged with Collett (2015).

3. STATISTICAL FORMALISM AND SIMULATION-BASED INFERENCE

Our goal is to infer the subhalo mass function parameters from a catalog of images of observed lenses. In this section we will describe the challenges of this inference problem and our approach of simulation-based inference. For simplicity, we will use a more abstract notation, distinguishing between three sets of quantities in the lensing system:

Parameters of interest ϑ: The vector $\vartheta = (f_{\rm sub}, \beta)^T$ parameterizes the subhalo mass function given, our goal is to infer their values.

Latent variables z: A vector of all other unobservable random variables in the simulator. These include the mass M_{200}, position (θ_x, θ_y), and redshift $z_{\rm host}$ of the host galaxy, the number of subhalos in the region of interest $n_{\rm ROI}$, the position and mass m_{200} of each subhalo, and the random variables related to the point spread function and Poisson fluctuations.

Observables x: The observed lens images.

Unfortunately, the conventions in astrophysics and statistics clash a little and can cause confusion: note the distinction between the parameters ϑ from the angular positions θ_x, θ_y and the Einstein radius θ_E; between the latent variables z and the redshifts $z_{\rm source}$, $z_{\rm host}$; and between the observed image x and the argument of the NFW profile $M_{\rm enc}(x)$ used in the last section.

As described in the previous section, we have implemented a simulator for the lensing process in the "forward" direction: for given parameters ϑ, the simulator

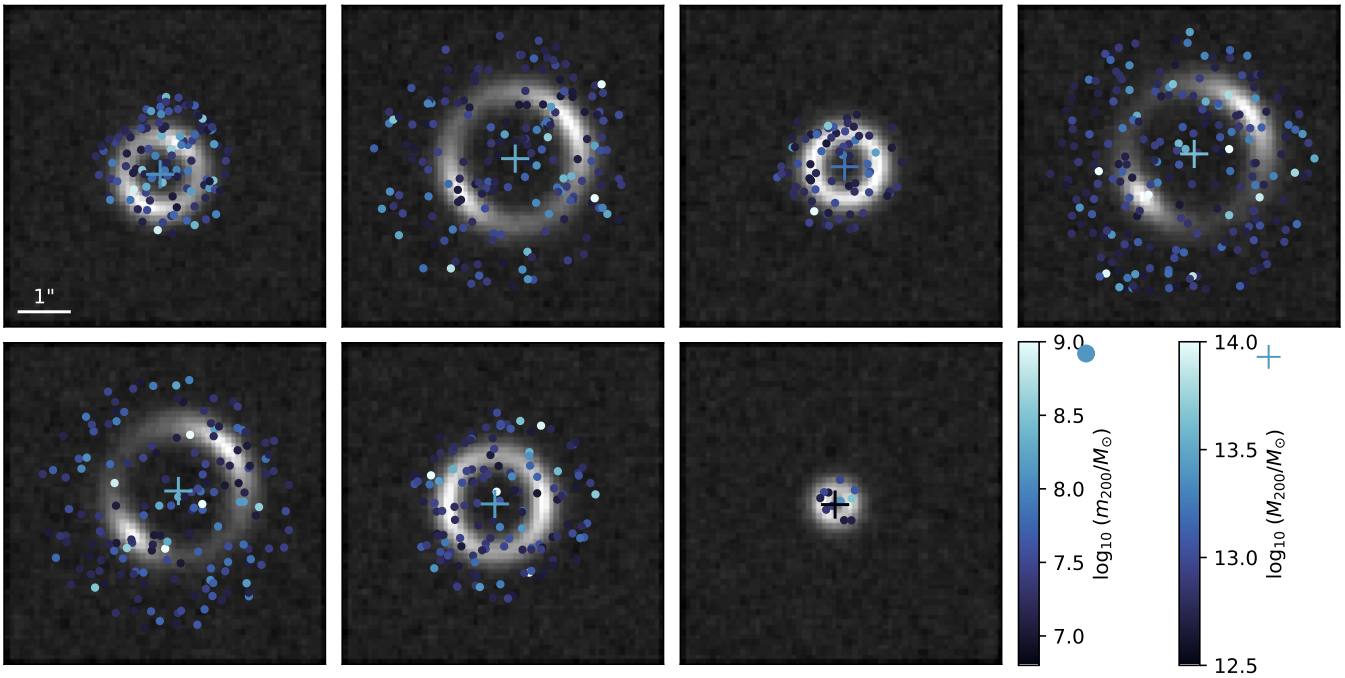

Figure 1. Simulated lenses. The cross markers show the offset of the host galaxy, its color its mass. The simulated subhalos are shown as dots, the color again indicates their masses. The greyscale images show the corresponding observed images. We show seven images randomly generated for $f_{\rm sub} = 0.05$ and $\beta = -0.9$.

samples latent variables z and finally observed images $x \sim p(x|\vartheta)$. Here $p(x|\vartheta)$ is the probability density or likelihood function of observing a lens image x given parameters ϑ. It can be schematically written as

$$p(x|\vartheta) = \int \mathrm{d}z\; p(x, z|\vartheta)\,, \tag{16}$$

where we integrate over the latent variables z and $p(x, z|\vartheta)$ is the joint likelihood of observables and latent variables:

$$\begin{aligned} p(x, z|\vartheta) = {} & p_{\rm host}(M_{200}, \theta_x, \theta_y, z_{\rm host})\;\; \mathrm{Pois}(n|\bar{n}_{\rm ROI}(\vartheta)) \\ & \prod_i^n \left[p_m\left(m_{200,i}|\vartheta\right)\;\; \mathrm{Uniform}\left(r_i\right)\right] \\ & \times p_{\rm obs}\left(x|f(M_{200}, \theta_x, \theta_y, z_{\rm host}; \{(m_{200,i}, r_i)\})\right)\,. \end{aligned} \tag{17}$$

In this schematic form $p_{\rm host}$ summarizes the distribution of the host halo parameters $z_{\rm host}$ such as its mass, offset and redshift; n is the actual number of subhalos in the region of interest, $\bar{n}_{\rm ROI}(\vartheta)$ is the mean number of subhalos in the region of interest as a function of the parameters $\vartheta = (f_{\rm sub}, \beta)^T$; $m_{200,i}$ and r_i are the subhalo masses and positions; $p_m(m|\vartheta) = 1/n\,\mathrm{d}n/\mathrm{d}m_{200}$ is the normalized subhalo mass function given in Eq. (13); and in the last line $p_{\rm obs}$ is the probability of observing an image x based on the true lensed image $f(z_{\rm host}, \{(m_{200,i}, r_i)\})$ due to the point spread function and Poisson fluctuations.

Most frequentist and Bayesian inference methods rely on evaluating the likelihood function $p(x|\vartheta)$. Unfortunately, even in our somewhat simplified simulator, each run of the simulation easily involves more than a thousand latent variables, and the integral over this enormeous space clearly cannot be computed explicitly. The likelihood function $p(x|\vartheta)$ is thus intractable, providing a major challenge for both frequentist and Bayesian inference. Similarly, inference with Markov Chain Monte Carlo (MCMC) methods based directly on the joint likelihood function $p(x, z|\vartheta)$ requires unfeasibly many simulations before converging because the latent space is so large. Systems defined through a forward simulator that does not admit a tractable likelihood are known as "implicit models", inference techniques for this case as "likelihood-free inference".

One way to tackle this issue is to reduce the high-dimensional data x to lower-dimensional summary statistics $v(x)$, for instance based on power spectra (Hezaveh et al. 2016a; Díaz Rivero et al. 2018; Diaz Rivero et al. 2018; Cyr-Racine et al. 2019; Brennan et al. 2019; Chatterjee & Koopmans 2018; Cyr-Racine et al. 2016). The likelihood $p(v|\vartheta)$ in the space of summary statistics can either be explicitly estimated through density estimation techniques such as histograms, kernel density estimation, or Gaussian processes, or replaced by a rejection probability in an Approximate Bayesian Computation (ABC) technique. While the compression

to summary statistics makes the analysis tractable, it typically loses information and hence reduces the statistical power of the analysis.

Instead, we follow an approach in which we approximate the likelihood function with a neural network, which has to be trained only once and can be evaluated efficiently for any parameter point and observed image. We will show how this turns the intractable integral in Eq. (16) into a tractable minimization problem and amortizes this marginalization. This method scales well to the expected large number of lenses expected in upcoming surveys. **[JB: Add references]** In particular, we use a new simulation-based inference technique introduced in Brehmer et al. (2018a,b,c) that consists of four steps:

1. During each run of the simulator, additional information that characterizes the subhalo population and lensing process is stored together with the simulated observed image.
2. This information is used to train a neural network to approximate the likelihood ratio function.
3. The neural network output is calibrated, ensuring that errors during training do not lead to wrong inference results.
4. The calibrated network output is then used in either frequentist or Bayesian inference techniques.

In the remainder of this section, we will explain these four steps in detail.

3.1. *Extracting additional information from the simulator*

In a first step, we generate training data by simulating a large number of observed lenses. For each lens, we first draw two parameter points from a proposal distribution, $\vartheta, \vartheta' \sim \pi(\vartheta)$. This proposal distribution should cover the region of interest in the parameter space, but does not have to be identical to the prior in a Bayesian inference setting, which allows us to be agnostic about the inference setup at this stage.

Next, the simulator is run for the parameter point ϑ, generating an observed image $x \sim p(x|\vartheta)$. In addition, we calculate and save two quantities: the joint likelihood ratio

$$r(x,z|\vartheta) = \frac{p(x,z|\vartheta)}{p_{\mathrm{ref}}(x,z)} \equiv \frac{p(x,z|\vartheta)}{\int \mathrm{d}\tilde{\vartheta}\, \pi(\tilde{\vartheta})\, p(x,z|\tilde{\vartheta})} \tag{18}$$

and the joint score

$$t(x,z|\vartheta) = \nabla_\vartheta \log p(x,z|\vartheta)\,. \tag{19}$$

The joint likelihood ratio quantifies how much more or less likely a particular simulation chain including the latent variables z is for the parameter point ϑ compared to a reference distribution. For this we choose the marginal distribution of latent variables and observables corresponding to the proposal distribution $\pi(\vartheta)$, which has support for every potential outcome of the simulation (Hermans et al. 2019). The joint score is the gradient of the joint log likelihood in model parameter space and quantifies if a particular simulation chain becomes more or less likely with infinitesimal changes of the parameters of interest. Both quantities depend on the latent variables of the simulation chain.

We compute the joint likelihood ratio and joint score with Eq. (17). Conveniently, many steps in the simulator do not explicitly depend on the parameters of interest ϑ and cancel in the joint likelihood ratio and joint score, and the remaining terms can be evaluated with little overhead to the simulation code. We also calculate the joint likelihood ratio $r(x,z|\vartheta')$ and the joint score $t(x,z|\vartheta')$ for the second parameter point ϑ' and store the parameter points ϑ and ϑ', the simulated image x, as well as the joint likelihood ratios and joint scores.

Our training samples consist of 10^6 images, with parameter points chosen from a uniform range in $0.001 < f_{\mathrm{sub}} < 0.2$ and $-1.5 < \beta < -0.5$.

3.2. *Machine learning*

How are the joint likelihood ratio and joint score, which are conditional on the latent variables z, useful to learn about the likelihood function $p(x|\vartheta)$, which only depends on the observed lens images and the parameters of interest? Consider the functional

$$\begin{aligned} L[g(x,\vartheta)] = &\int \mathrm{d}\vartheta \int \mathrm{d}\vartheta' \int \mathrm{d}x \int \mathrm{d}z\; \pi(\vartheta)\; \pi(\vartheta')\; p(x,z|\vartheta) \\ &\times \Bigg[-s \log g - (1-s)\log(1-g) - s' \log g' - (1-s')\log(1-g') \\ &+ \alpha \left\{ \left| t - \nabla_\vartheta \log \tfrac{1-g}{g}\Big|_\vartheta \right|^2 + \left| t' - \nabla_\vartheta \log \tfrac{1-g}{g}\Big|_{\vartheta'} \right|^2 \right\} \Bigg], \end{aligned} \tag{20}$$

where we have abbreviated $s \equiv s(x,z|\vartheta) \equiv 1/(1 + r(x,z|\vartheta))$, $s' \equiv s(x,z|\vartheta') \equiv 1/(1 + r(x,z|\vartheta'))$, $g \equiv g(x,\vartheta)$, $g' \equiv g(x,\vartheta')$, $t = t(x,z|\vartheta)$, and $t' \equiv t(x,z|\vartheta')$ for readability. Note that the test function $g(x,\vartheta)$ is a function of x and ϑ only. The first two lines are an improved version of the cross-entropy loss, in which the joint likelihood ratio is used to decrease the variance compared to the canonical cross-entropy Stoye et al. (2018). The last line adds gradient information, weighted by a hyperparameter α.

As shown in Stoye et al. (2018), this "ALICES" loss functional is minimized by the function

$$g^*(x,\vartheta) \equiv \arg\min_g L[g(x,\vartheta)] = \frac{1}{1+r(x|\vartheta)}\,, \tag{21}$$

one-to-one with the likelihood ratio function

$$r(x|\vartheta) \equiv \frac{p(x|\vartheta)}{p_{\rm ref}(x)} = \frac{1-g^*(x,\vartheta)}{g^*(x,\vartheta)}\,. \tag{22}$$

We demonstrate the minimization of this functional explicitly in Appendix A. This means that if we can construct the functional in Eq. (20) with the joint likelihood ratio and joint score extracted from the simulator and numerically minimize it, the resulting function lets us reconstruct the (otherwise intractable) likelihood ratio function $r(x|\vartheta)$! Essentially, this step lets us integrate out the dependence on latent variables z from the joint likelihood ratio and score, but in a general, functional form that does not depend on a set of observed images.

This is why extraction of the joint likelihood ratio and joint score has been described with the analogy of "mining gold" from the simulator (Brehmer et al. 2018c): while calculating these quantities may require some effort and changes to the simulator code, through the minimization of a suitable functional they allow us to calculate the otherwise intractable likelihood ratio function.

In practice, we implement this minimization with machine learning. A neural network plays the role of the test function $g(x,\vartheta)$, the integrals in Eq. (20) are approximated with a sum over training data sampled according to $\pi(\vartheta)\pi(\vartheta')p(x,z|\vartheta)$, and we minimize the loss numerically through a stochastic gradient descent algorithm. The neural network trained in this way provides an estimator $\hat{r}(x|\vartheta)$ of the likelihood ratio function that is exact in the limit of infinite training samples, sufficient network capacity, and efficient minimization. Note the "parameterized" structure of the network, in which a single neural network is trained to estimate the likelihood ratio over all of the parameter space, with the tested parameter point ϑ being an input to the network (Cranmer et al. 2015; Baldi et al. 2016). This approach is more efficient than a point-by-point analysis of a grid of parameter points: it allows the network to "borrow" information from neighboring parameter points, benefitting from the typically smooth structure of the parameter space.

Given the image nature of the lensing data, we choose a convolutional network architecture based on the ResNet-18 (He et al. 2016) implementation in PyTorch (Paszke et al. 2017). The parameters ϑ enter as additional inputs in the fully connected layers of the network. Compared to the original ResNet-18 architecture, we add another fully connected layer at the end to ensure that the relation between parameters of interest and image data can be modeled. All inputs are normalized to zero mean and unit variance. We train the networks by minimizing the loss in Eq. (20) with $\alpha = 2\cdot 10^{-3}$ over 100 epochs with a batch size of 128 using the Adam optimizer (Kingma & Ba 2014), exponentially decaying the learning rate from $3\cdot 10^{-4}$ to $3\cdot 10^{-6}$ with early stopping. This architecture and hyperparameter configuration performed best during a rough hyperparameter scan, though for this proof-of-concept study we have not performed an exhaustive optimization.

3.3. *Calibration*

In reality, the neural network might not learn the likelihood ratio function $r(x|\vartheta)$ exactly, for instance due to limited training data or inefficient training. To make sure that our inference results are correct even in this case, we calibrate the network output with histograms (Cranmer et al. 2015; Brehmer et al. 2018b). For every parameter point ϑ that we want to test, we simulate a set of images $\{x\} \sim p(x|\vartheta)$ from this parameter point and calculate the network prediction $\hat{r} \equiv \hat{r}(x|\vartheta)$ for each image. We also simulate a set of images $\{x\} \sim p_{\rm ref}(x)$ from the reference model, that is, drawing a new parameter points from $\pi(\vartheta)$ for each image, and again calculate the network prediction $\hat{r}$ for each of these lenses as well. We then calculate the calibrated likelihood ratio from histograms of the network predictions as

$$\hat{r}_{\rm cal}(x|\vartheta) = \frac{\hat{p}(\hat{r}|\vartheta)}{\hat{p}_{\rm ref}(\hat{r})} \tag{23}$$

where the $\hat{p}(\dots)$ denote probability densities estimated with univariate histograms.

This additional calibration stage comes with a certain computational cost that increases linearly with the number of evaluated parameter points. However, it guarantees that as long as the simulator accurately models the process, the inference results may be perfect or conservative, but not be too optimistic, even if the neural network output is substantially different from the true likelihood ratio.

3.4. *Inference*

After a neural network has been trained (and optionally calibrated) to estimate the likelihood ratio function, it provides the basic ingredient to both frequentist and Bayesian inference. For frequentist hypothesis tests, the likelihood ratio provides the most powerful test statistic (Neyman & Pearson 1933). In addition, its asymptotic properties allow us in many cases to directly translate a value of the likelihood ratio into a p-value and thus

into exclusion limits at a given confidence level (Wilks 1938; Wald 1943; Cowan et al. 2011).

For Bayesian inference, note that we can write Bayes' theorem as

$$\begin{aligned} p(\vartheta|\{x_i\}) &= \frac{p(\vartheta)\ \prod_i p(x_i|\vartheta)}{\int \mathrm{d}\tilde{\vartheta}\, p(\tilde{\vartheta})\ \prod_i p(x_i|\tilde{\vartheta})} \\ &= p(\vartheta)\left[\int \mathrm{d}\tilde{\vartheta}\, p(\tilde{\vartheta})\ \prod_i \frac{p(x_i|\tilde{\vartheta})}{p(x_i|\vartheta)}\right]^{-1} \\ &\approx p(\vartheta)\left[\int \mathrm{d}\tilde{\vartheta}\, p(\tilde{\vartheta})\ \prod_i \frac{\hat{r}(x_i|\tilde{\vartheta})}{\hat{r}(x_i|\vartheta)}\right]^{-1}, \end{aligned} \tag{24}$$

where $\{x_i\}$ is the set of observed lens images and $p(\vartheta)$ is the prior on the parameters of interest, which may be different from the proposal distribution $\pi(\vartheta)$ used during the generation of training data. The posterior can thus be directly calculated given an estimator $\hat{r}$, provided that the space of the parameters of interest is low-dimensional enough to calculate the integral, or with MCMC or variational inference techniques otherwise.

While our approach to inference is strongly based on the ideas in Brehmer et al. (2018a,b,c); Stoye et al. (2018), there are some novel features in our analysis that we would like to highlight briefly. Unlike in those earlier papers, we use a marginal model based on the proposal distribution $\pi(\vartheta)$ as reference model in the denominator of the likelihood ratio, which reduces numerical issues due to the joint likelihood ratio evaluating as zero or infinity substantially. It also allows us to include the "flipped" terms involving s' and g' in the loss function in Eq. (20); we found that this new, improved version of the ALICES loss improves the sample efficiency of our algorithms. Finally, this is the first application of the "gold mining" idea to image data, the first combination with a convolutional network architecture, and the first use for Bayesian inference.

4. RESULTS

5. EXTENSIONS

- Multiple passbands
- Per-image observational properties (e.g., exposure and PSF)
- More complicated host and source profiles
- Tidal stripping of subhalos (spatial dependence of density profile/concentration and overall depression in number density)
- Redshift-dependence of SHMF / different degrees of knowledge about redshift
- Inclusion of external shear
- Inclusion of line-of-sight substructure

6. CONCLUSIONS

We thank Simon Birrer, Christopher Fassnacht, Daniel Gilman, and Neal Weiner for useful conversations. JB and KC are partially supported by NSF awards ACI-1450310, OAC-1836650, and OAC-1841471, and the Moore-Sloan data science environment at NYU. SM is partially supported by the NSF CAREER grant PHY-1554858 and NSF grant PHY-1620727. KC is also supported through the NSF grant PHY-1505463. This work was also supported through the NYU IT High Performance Computing resources, services, and staff expertise.

Software: *Astropy* (Astropy Collaboration et al. 2013, 2018), *IPython* (Perez & Granger 2007), *LensPop* (Collett 2015), *MadMiner* (Brehmer et al. 2019), *matplotlib* (Hunter 2007), *NumPy* (van der Walt et al. 2011), *PyTorch* (Paszke et al. 2017), *SciPy* (Jones et al. 2001–).

APPENDIX

A. MINIMUM OF THE LOSS FUNCTIONAL

A central step in our inference technique is numerically minimizing the functional $L[g(x,\vartheta)]$ given in Eq. (20) to obtain an estimator for the likelihood ratio function. Here we will use calculus of variation to explicitly show that the solution given in Eq. (21) in fact minimizes this loss.

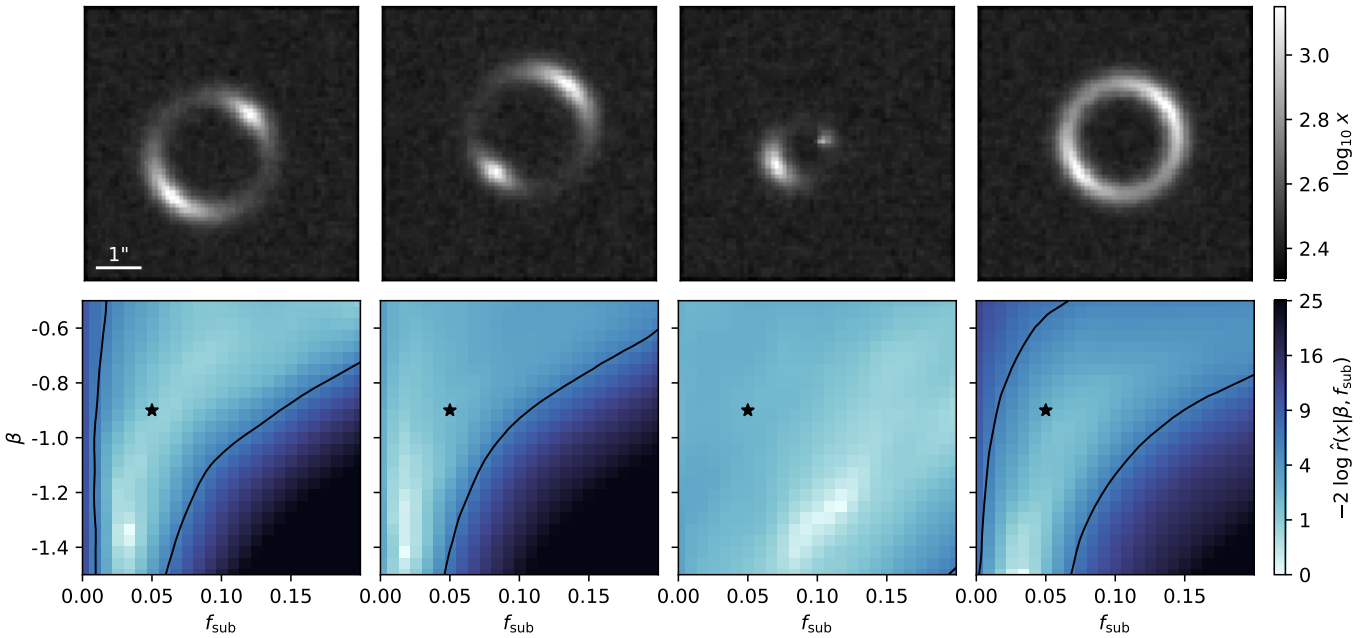

Figure 2. Four simulated lens images (upper panels) and the corresponding estimated likelihood maps (lower panels). The star marks the true point used to generate the images, the black line shows 95% CL contours in parameter space based on each image. **[JB: So far without calibration.]**

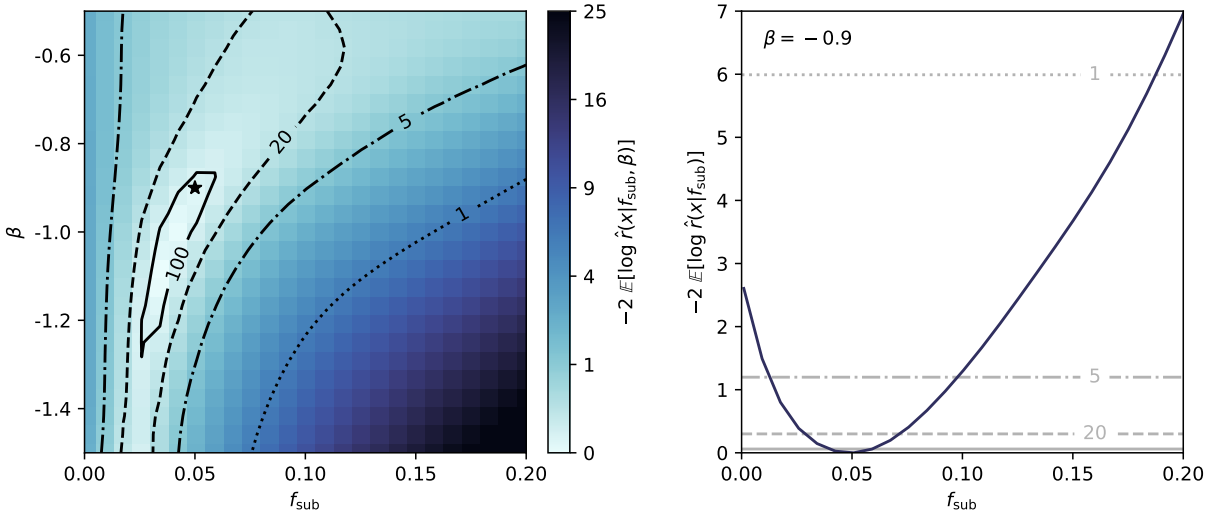

Figure 3. Expected likelihood ratio map assuming $\beta = -0.9$ and $f_{\rm sub} = 0.05$ in the two-dimensional parameter space (left) and a one-dimensional slice at $\beta = -0.9$ (right). The lines show expected 95% CL exclusion limits for 1 (dotted), 5 (dash-dotted), 20 (dashed), and 100 (solid) observed lenses. **[JB: So far without calibration.]**

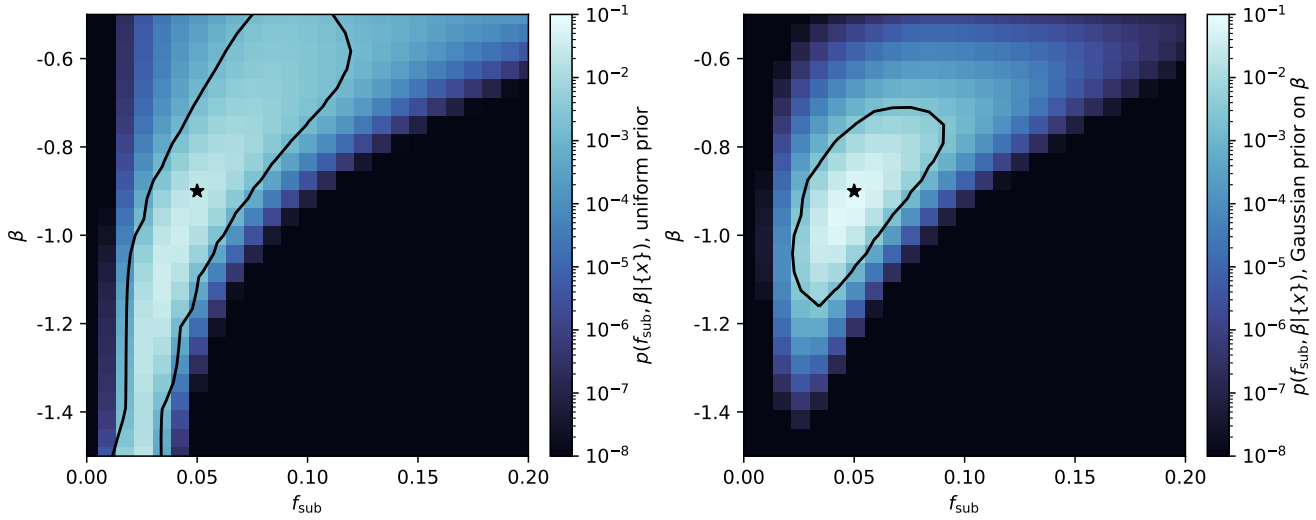

Figure 4. Expected posterior based on a uniform prior (left) or a Gaussian prior on β with mean -0.9 and standard deviation 0.1 (right) for 10 observed lenses. We assume $\beta = -0.9$ and $f_{\rm sub} = 0.05$. **[JB: So far without calibration.]**

First consider the case of $\alpha = 0$, i. e. the functional

$$\begin{aligned} L[g(x,\vartheta)] &= \int \mathrm{d}\vartheta \int \mathrm{d}\vartheta' \int \mathrm{d}x \int \mathrm{d}z\, \pi(\vartheta)\pi(\vartheta') p(x,z|\vartheta) \Big(-s \log g - (1-s)\log(1-g) - s' \log g' - (1-s')\log(1-g') \Big) \\ &= \int \mathrm{d}\vartheta \int \mathrm{d}x \underbrace{\left[\int \mathrm{d}z\, \pi(\vartheta) \Big(p(x,z|\vartheta) + p_{\rm ref}(x,z) \Big) \Big(-s \log g - (1-s)\log(1-g) \Big) \right]}_{\equiv F(x,\vartheta)} , \end{aligned} \tag{A1}$$

where we use the shorthand notation $s \equiv s(x,z|\vartheta) \equiv 1/(1+r(x,z|\vartheta))$, $s' \equiv s(x,z|\vartheta') \equiv 1/(1+r(x,z|\vartheta'))$, $g \equiv g(x,\vartheta)$, $g' \equiv g(x,\vartheta')$. The function $g^*(x|\vartheta)$ that minimizes this functional has to satisfy

$$0 \overset{!}{=} \left.\frac{\delta F}{\delta g}\right|_{g^*} = \int \mathrm{d}z\, \pi(\vartheta) \Big(p(x,z|\vartheta) + p_{\rm ref}(x,z) \Big) \Big(-\frac{s}{g^*} + \frac{1-s}{1-g^*} \Big) \tag{A2}$$

As long as $\pi(\vartheta) > 0$, this is equivalent to

$$(1-g^*) \int \mathrm{d}z \Big(p(x,z|\vartheta) + p_{\rm ref}(x,z) \Big) s = g^* \int \mathrm{d}z \Big(p(x,z|\vartheta) + p_{\rm ref}(x,z) \Big) (1-s) \tag{A3}$$

and finally

$$\begin{aligned} g^*(x|\vartheta) &= \frac{\int \mathrm{d}z \Big(p(x,z|\vartheta) + p_{\rm ref}(x,z) \Big) s(x,z|\vartheta)}{\int \mathrm{d}z \Big(p(x,z|\vartheta) + p_{\rm ref}(x,z) \Big)} \\ &= \frac{\int \mathrm{d}z \Big(p(x,z|\vartheta) + p_{\rm ref}(x,z) \Big) \frac{1}{1+p(x,z|\vartheta)/p_{\rm ref}(x,z)}}{\int \mathrm{d}z \Big(p(x,z|\vartheta) + p_{\rm ref}(x,z) \Big)} \\ &= \frac{p_{\rm ref}(x)}{p(x|\vartheta) + p_{\rm ref}(x)} = \frac{1}{1+r(x|\vartheta)} , \end{aligned} \tag{A4}$$

in agreement with Eq. (21). Note that this result is independent of the choise of $\pi(\vartheta)$, as long as this proposal distribution has support at all relevant parameter points.

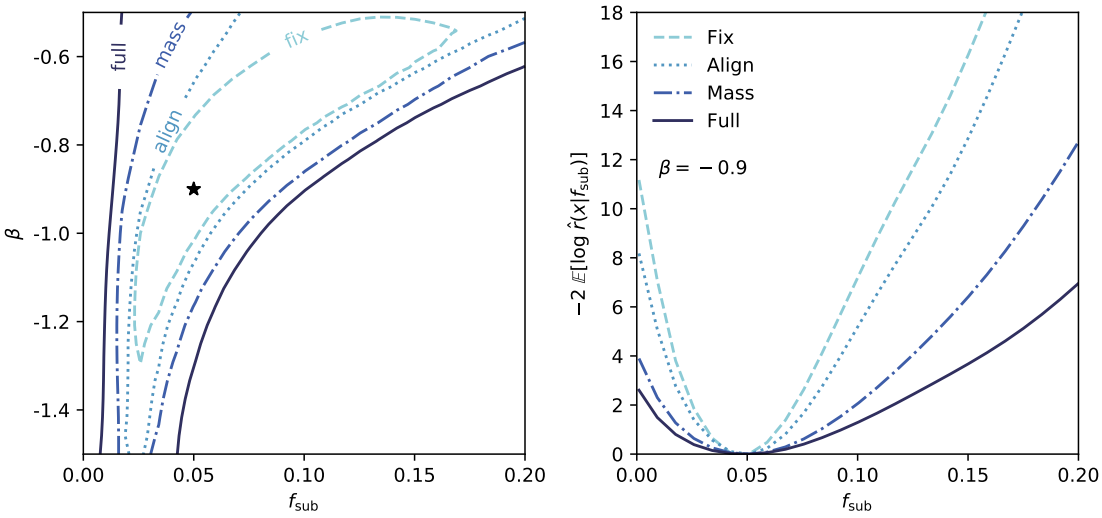

Figure 5. Left: Expected 95% CL exclusion limits for 10 observed lenses for four different levels of complexity of the simulator. Right: expected likelihood ratio along a one-dimensional slice through the parameter space at $\beta = -0.9$ for the same four simulator scenarios. In both panels we compare the "full" simulator discussed in Sec. 2, a scenario in which the host mass is varied but the offset relative to the source is fixed at zero ("mass"), a case in which the source offset is varied but the host halo mass is fixed ("align"), and a toy scenario in which both the offset and the mass of the host halo are fixed ("fix"). The data was generated for $\beta = -0.9$ and $f_{\rm sub} = 0.05$. **[JB: So far without calibration.]**

Similarly it can be shown that the gradient term in the loss functional weighted by α is minimized when the gradient of the log likelihood ratio estimated by the neural network is equal to the true score,

$$\nabla_\vartheta \log \hat{r}(x|\vartheta) \equiv \nabla_\vartheta \log \frac{1 - g^*(x,\vartheta)}{g^*(x,\vartheta)} = \nabla_\vartheta \log r(x|\vartheta)\,. \tag{A5}$$

We refer the reader to (Brehmer et al. 2018b) for the derivation. While not strictly necessary for the inference technique, including this term in the loss function substantially improves the sample efficiency of the algorithm, similar to how gradient information makes any fit converge faster. Varying the hyperparameter α, we found a good performance for $\alpha \approx 2 \cdot 10^{-3}$, with the network performance being insensitive to small variations.

B. SIMPLIFIED SCENARIOS

REFERENCES

Aaboud, M., Aad, G., Abbott, B., et al. 2019, Journal of High Energy Physics, 2019, 142, arXiv: 1903.01400, doi: 10.1007/JHEP05(2019)142

Agrawal, P., Cyr-Racine, F.-Y., Randall, L., & Scholtz, J. 2017, JCAP, 2017, 021, arXiv: 1702.05482, doi: 10.1088/1475-7516/2017/08/021

Agrawal, P., & Randall, L. 2017, JCAP, 2017, 019, arXiv: 1706.04195, doi: 10.1088/1475-7516/2017/12/019

Akerib, D. S., Alsum, S., Araújo, H. M., et al. 2017, PhRvL, 118, 021303, arXiv: 1608.07648, doi: 10.1103/PhysRevLett.118.021303

Albert, A., Anderson, B., Bechtol, K., et al. 2017, ApJ, 834, 110, arXiv: 1611.03184, doi: 10.3847/1538-4357/834/2/110

Aprile, E., Aalbers, J., Agostini, F., et al. 2018, PhRvL, 121, 111302, arXiv: 1805.12562, doi: 10.1103/PhysRevLett.121.111302

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33, arXiv: 1307.6212, doi: 10.1051/0004-6361/201322068

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123, arXiv: 1801.02634, doi: 10.3847/1538-3881/aabc4f

Baldi, P., Cranmer, K., Faucett, T., Sadowski, P., & Whiteson, D. 2016, European Physical Journal C, 76, 235, arXiv: 1601.07913, doi: 10.1140/epjc/s10052-016-4099-4

Bechtol, K., Drlica-Wagner, A., Balbinot, E., et al. 2015, ApJ, 807, 50, arXiv: 1503.02584, doi: 10.1088/0004-637X/807/1/50

Birrer, S., Amara, A., & Refregier, A. 2017, JCAP, 2017, 037, arXiv: 1702.00009, doi: 10.1088/1475-7516/2017/05/037

Bode, P., Ostriker, J. P., & Turok, N. 2001, ApJ, 556, 93, arXiv: astro-ph/0010389, doi: 10.1086/321541

Bonaca, A., & Hogg, D. W. 2018, ApJ, 867, 101, arXiv: 1804.06854, doi: 10.3847/1538-4357/aae4da

Bonaca, A., Hogg, D. W., Price-Whelan, A. M., & Conroy, C. 2019, ApJ, 880, 38, arXiv: 1811.03631, doi: 10.3847/1538-4357/ab2873

Bond, J. R., & Szalay, A. S. 1983, ApJ, 274, 443, doi: 10.1086/161460

Boyanovsky, D., de Vega, H. J., & Sanchez, N. G. 2008, PhRvD, 78, 063546, arXiv: 0807.0622, doi: 10.1103/PhysRevD.78.063546

Boyanovsky, D., & Wu, J. 2011, PhRvD, 83, 043524, arXiv: 1008.0992, doi: 10.1103/PhysRevD.83.043524

Brehmer, J., Cranmer, K., Louppe, G., & Pavez, J. 2018a, PhRvL, 121, 111801, arXiv: 1805.00013, doi: 10.1103/PhysRevLett.121.111801

—. 2018b, PhRvD, 98, 052004, arXiv: 1805.00020, doi: 10.1103/PhysRevD.98.052004

Brehmer, J., Kling, F., Espejo, I., & Cranmer, K. 2019, arXiv e-prints, arXiv:1907.10621, arXiv: 1907.10621

Brehmer, J., Louppe, G., Pavez, J., & Cranmer, K. 2018c, arXiv e-prints, arXiv:1805.12244, arXiv: 1805.12244

Brennan, S., Benson, A. J., Cyr-Racine, F.-Y., et al. 2019, MNRAS, 2044, arXiv: 1808.03501, doi: 10.1093/mnras/stz1607

Brewer, B. J., Huijser, D., & Lewis, G. F. 2016, MNRAS, 455, 1819, arXiv: 1508.00662, doi: 10.1093/mnras/stv2370

Brooks, A. M. 2018, arXiv e-prints, arXiv:1812.00044, arXiv: 1812.00044

Buckley, M. R., & DiFranzo, A. 2018, PhRvL, 120, 051102, arXiv: 1707.03829, doi: 10.1103/PhysRevLett.120.051102

Buschmann, M., Kopp, J., Safdi, B. R., & Wu, C.-L. 2018, PhRvL, 120, 211101, arXiv: 1711.03554, doi: 10.1103/PhysRevLett.120.211101

Carlberg, R. G. 2012, ApJ, 748, 20, arXiv: 1109.6022, doi: 10.1088/0004-637X/748/1/20

Carlberg, R. G., & Grillmair, C. J. 2013, ApJ, 768, 171, arXiv: 1303.4342, doi: 10.1088/0004-637X/768/2/171

Chang, L. J., Lisanti, M., & Mishra-Sharma, S. 2018, PhRvD, 98, 123004, arXiv: 1804.04132, doi: 10.1103/PhysRevD.98.123004

Chatterjee, S., & Koopmans, L. V. E. 2018, MNRAS, 474, 1762, arXiv: 1710.03075, doi: 10.1093/mnras/stx2674

Ciotti, L., & Bertin, G. 1999, A&A, 352, 447

Collett, T. E. 2015, ApJ, 811, 20, arXiv: 1507.02657, doi: 10.1088/0004-637X/811/1/20

Cowan, G., Cranmer, K., Gross, E., & Vitells, O. 2011, European Physical Journal C, 71, 1554, arXiv: 1007.1727, doi: 10.1140/epjc/s10052-011-1554-0

Cranmer, K., Pavez, J., & Louppe, G. 2015, arXiv e-prints, arXiv:1506.02169, arXiv: 1506.02169

Cui, X., Abdukerim, A., Chen, W., et al. 2017, PhRvL, 119, 181302, doi: 10.1103/PhysRevLett.119.181302

Cyr-Racine, F.-Y., Keeton, C. R., & Moustakas, L. A. 2019, PhRvD, 100, 023013, arXiv: 1806.07897, doi: 10.1103/PhysRevD.100.023013

Cyr-Racine, F.-Y., Moustakas, L. A., Keeton, C. R., Sigurdson, K., & Gilman, D. A. 2016, PhRvD, 94, 043505, arXiv: 1506.01724, doi: 10.1103/PhysRevD.94.043505

Dalal, N., & Kochanek, C. S. 2002, ApJ, 572, 25, arXiv: astro-ph/0111456, doi: 10.1086/340303

Dalcanton, J. J., & Hogan, C. J. 2001, ApJ, 561, 35, arXiv: astro-ph/0004381, doi: 10.1086/323207

Daylan, T., Cyr-Racine, F.-Y., Diaz Rivero, A., Dvorkin, C., & Finkbeiner, D. P. 2018, ApJ, 854, 141, arXiv: 1706.06111, doi: 10.3847/1538-4357/aaaa1e

Despali, G., & Vegetti, S. 2017, MNRAS, 469, 1997, arXiv: 1608.06938, doi: 10.1093/mnras/stx966

Diaz Rivero, A., Cyr-Racine, F.-Y., & Dvorkin, C. 2018, PhRvD, 97, 023001, arXiv: 1707.04590, doi: 10.1103/PhysRevD.97.023001

Díaz Rivero, A., Dvorkin, C., Cyr-Racine, F.-Y., Zavala, J., & Vogelsberger, M. 2018, PhRvD, 98, 103517, arXiv: 1809.00004, doi: 10.1103/PhysRevD.98.103517

Drlica-Wagner, A., Mao, Y.-Y., Adhikari, S., et al. 2019, arXiv e-prints, arXiv:1902.01055, arXiv: 1902.01055

Efstathiou, G. 1992, MNRAS, 256, 43P, doi: 10.1093/mnras/256.1.43P

Errani, R., Peñarrubia, J., Laporte, C. F. P., & Gómez, F. A. 2017, MNRAS, 465, L59, arXiv: 1608.01849, doi: 10.1093/mnrasl/slw211

Fan, J., Katz, A., Randall, L., & Reece, M. 2013, Physics of the Dark Universe, 2, 139, arXiv: 1303.1521, doi: 10.1016/j.dark.2013.07.001

Fitts, A., Boylan-Kolchin, M., Elbert, O. D., et al. 2017, MNRAS, 471, 3547, arXiv: 1611.02281, doi: 10.1093/mnras/stx1757

Fitts, A., Boylan-Kolchin, M., Bozek, B., et al. 2018, arXiv e-prints, arXiv:1811.11791, arXiv: 1811.11791

Garrison-Kimmel, S., Wetzel, A., Bullock, J. S., et al. 2017, MNRAS, 471, 1709, arXiv: 1701.03792, doi: 10.1093/mnras/stx1710

Han, J., Cole, S., Frenk, C. S., & Jing, Y. 2016, MNRAS, 457, 1208, arXiv: 1509.02175, doi: 10.1093/mnras/stv2900

He, K., Zhang, X., Ren, S., & Sun, J. 2016, in Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778

Hermans, J., Begy, V., & Louppe, G. 2019, arXiv e-prints, arXiv:1903.04057, arXiv: 1903.04057

Hezaveh, Y., Dalal, N., Holder, G., et al. 2016a, JCAP, 2016, 048, arXiv: 1403.2720, doi: 10.1088/1475-7516/2016/11/048

Hezaveh, Y. D., Dalal, N., Marrone, D. P., et al. 2016b, ApJ, 823, 37, arXiv: 1601.01388, doi: 10.3847/0004-637X/823/1/37

Hiroshima, N., Ando, S., & Ishiyama, T. 2018, PhRvD, 97, 123002, arXiv: 1803.07691, doi: 10.1103/PhysRevD.97.123002

Hsueh, J.-W., Enzi, W., Vegetti, S., et al. 2019, arXiv e-prints, arXiv:1905.04182, arXiv: 1905.04182

Hunter, J. D. 2007, Computing In Science & Engineering, 9, 90

Johnston, K. V., Zhao, H., Spergel, D. N., & Hernquist, L. 1999, ApJL, 512, L109, arXiv: astro-ph/9807243, doi: 10.1086/311876

Jones, E., Oliphant, T., Peterson, P., et al. 2001–, SciPy: Open source scientific tools for Python. http://www.scipy.org/

Kahlhoefer, F., Kaplinghat, M., Slatyer, T. R., & Wu, C.-L. 2019, arXiv e-prints, arXiv:1904.10539, arXiv: 1904.10539

Kaplinghat, M., Keeley, R. E., Linden, T., & Yu, H.-B. 2014, PhRvL, 113, 021302, arXiv: 1311.6524, doi: 10.1103/PhysRevLett.113.021302

Kaplinghat, M., Tulin, S., & Yu, H.-B. 2016, PhRvL, 116, 041302, arXiv: 1508.03339, doi: 10.1103/PhysRevLett.116.041302

Keeton, C. R. 2001, arXiv e-prints, astro, arXiv: astro-ph/0102341

Kingma, D. P., & Ba, J. 2014, arXiv e-prints, arXiv:1412.6980, arXiv: 1412.6980

Koposov, S., Belokurov, V., Evans, N. W., et al. 2008, ApJ, 686, 279, arXiv: 0706.2687, doi: 10.1086/589911

Koposov, S. E., Belokurov, V., Torrealba, G., & Evans, N. W. 2015, ApJ, 805, 130, arXiv: 1503.02079, doi: 10.1088/0004-637X/805/2/130

Lisanti, M., Mishra-Sharma, S., Rodd, N. L., & Safdi, B. R. 2018, PhRvL, 120, 101101, arXiv: 1708.09385, doi: 10.1103/PhysRevLett.120.101101

LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv e-prints, arXiv:0912.0201, arXiv: 0912.0201

Madau, P., Diemand, J., & Kuhlen, M. 2008, The Astrophysical Journal, 679, 1260, arXiv: 0802.2265, doi: 10.1086/587545

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563, doi: 10.1086/177173

—. 1997, ApJ, 490, 493, doi: 10.1086/304888

Neyman, J., & Pearson, E. S. 1933, Philosophical Transactions of the Royal Society of London Series A, 231, 289, doi: 10.1098/rsta.1933.0009

Paszke, A., Gross, S., Chintala, S., et al. 2017, in NIPS-W

Perez, F., & Granger, B. E. 2007, Computing in Science and Engineering, 9, 21, doi: 10.1109/MCSE.2007.53

Peter, A. H. G., Rocha, M., Bullock, J. S., & Kaplinghat, M. 2013, MNRAS, 430, 105, arXiv: 1208.3026, doi: 10.1093/mnras/sts535

Read, J. I., Iorio, G., Agertz, O., & Fraternali, F. 2017, MNRAS, 467, 2019, arXiv: 1607.03127, doi: 10.1093/mnras/stx147

Refregier, A., Amara, A., Kitching, T. D., et al. 2010, arXiv e-prints, arXiv:1001.0061, arXiv: 1001.0061

Sánchez-Conde, M. A., & Prada, F. 2014, MNRAS, 442, 2271, arXiv: 1312.1729, doi: 10.1093/mnras/stu1014

Schneider, P., Ehlers, J., & Falco, E. E. 1992, Gravitational Lenses, doi: 10.1007/978-3-662-03758-4

Sirunyan, A. M., Tumasyan, A., Adam, W., et al. 2017, Physics Letters B, 769, 520, arXiv: 1611.03568, doi: 10.1016/j.physletb.2017.02.012

Springel, V., Wang, J., Vogelsberger, M., et al. 2008, Monthly Notices of the Royal Astronomical Society, 391, 1685, arXiv: 0809.0898, doi: 10.1111/j.1365-2966.2008.14066.x

Stoye, M., Brehmer, J., Louppe, G., Pavez, J., & Cranmer, K. 2018, arXiv e-prints, arXiv:1808.00973, arXiv: 1808.00973

van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science and Engineering, 13, 22, arXiv: 1102.1523, doi: 10.1109/MCSE.2011.37

Van Tilburg, K., Taki, A.-M., & Weiner, N. 2018, JCAP, 2018, 041, arXiv: 1804.01991, doi: 10.1088/1475-7516/2018/07/041

Vegetti, S., Koopmans, L. V. E., Bolton, A., Treu, T., & Gavazzi, R. 2010, MNRAS, 408, 1969, arXiv: 0910.0760, doi: 10.1111/j.1365-2966.2010.16865.x

Vegetti, S., Lagattuta, D. J., McKean, J. P., et al. 2012, Nature, 481, 341, arXiv: 1201.3643, doi: 10.1038/nature10669

Verma, A., Collett, T., Smith, G. P., Strong Lensing Science Collaboration, & the DESC Strong Lensing Science Working Group. 2019, arXiv e-prints, arXiv:1902.05141, arXiv: 1902.05141

Vogelsberger, M., Zavala, J., Cyr-Racine, F.-Y., et al. 2016, MNRAS, 460, 1399, arXiv: 1512.05349, doi: 10.1093/mnras/stw1076

Vogelsberger, M., Zavala, J., & Loeb, A. 2012, MNRAS, 423, 3740, arXiv: 1201.5892, doi: 10.1111/j.1365-2966.2012.21182.x

Vogelsberger, M., Zavala, J., Schutz, K., & Slatyer, T. R. 2019, MNRAS, 484, 5437, arXiv: 1805.03203, doi: 10.1093/mnras/stz340

Wald, A. 1943, Transactions of the American Mathematical Society, 54, 426

Wechsler, R. H., & Tinker, J. L. 2018, ARA&A, 56, 435, arXiv: 1804.03097, doi: 10.1146/annurev-astro-081817-051756

Wilks, S. S. 1938, Annals Math. Statist., 9, 60, doi: 10.1214/aoms/1177732360

Zahid, H. J., Sohn, J., & Geller, M. J. 2018, ApJ, 859, 96, arXiv: 1804.04492, doi: 10.3847/1538-4357/aabe31

Zavala, J., Vogelsberger, M., & Walker, M. G. 2013, MNRAS, 431, L20, arXiv: 1211.6426, doi: 10.1093/mnrasl/sls053