



UvA-DARE (Digital Academic Repository)

The effect of the perturber population on subhalo measurements in strong gravitational lenses

Coogan, A.; Anau Montel, N.; Karchev, K.; Grootes, M.W.; Nattino, F.; Weniger, C.

DOI

[10.1093/MNRAS/STAD2925](https://doi.org/10.1093/MNRAS/STAD2925)

Publication date

2024

Document Version

Final published version

Published in

Monthly Notices of the Royal Astronomical Society

License

CC BY

[Link to publication](#)

Citation for published version (APA):

Coogan, A., Anau Montel, N., Karchev, K., Grootes, M. W., Nattino, F., & Weniger, C. (2024). The effect of the perturber population on subhalo measurements in strong gravitational lenses. *Monthly Notices of the Royal Astronomical Society*, 527(1), 66-78. <https://doi.org/10.1093/MNRAS/STAD2925>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)

The effect of the perturber population on subhalo measurements in strong gravitational lenses

Adam Coogan¹,^{1,2,3} Noemi Anau Montel,³★ Konstantin Karchev¹,^{3,4} Meiert W. Grootes,⁵ Francesco Nattino⁵ and Christoph Weniger³★

¹Département de Physique, Université de Montréal, 1375 Avenue Thérèse-Lavoie-Roux, Montréal, QC H2V 0B3, Canada

²Mila – Quebec AI Institute, 6666 St-Urbain, 200, Montreal, QC, H2S 3H1, Canada,

³GRAPPA (Gravitation Astroparticle Physics Amsterdam), University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands

⁴SISSA (Scuola Internazionale Superiore di Studi Avanzati), via Bonomea 265, I-34136 Trieste, Italy

⁵Netherlands eScience Center, Science Park 402 (Matrix III), 1098 XH Amsterdam, the Netherlands

Accepted 2023 September 21. Received 2023 July 30; in original form 2022 October 18

ABSTRACT

Analyses of extended arcs in strong gravitational lensing images to date have constrained the properties of dark matter by measuring the parameters of one or two individual subhaloes. However, since such analyses are reliant on likelihood-based methods like Markov-chain Monte Carlo or nested sampling, they require various compromises to the realism of lensing models for the sake of computational tractability, such as ignoring the numerous other subhaloes and line-of-sight haloes in the system, assuming a particular form for the source model and requiring the noise to have a known likelihood function. Here, we show that a simulation-based inference method called truncated marginal neural ratio estimation (TMNRE) makes it possible to relax these requirements by training neural networks to directly compute marginal posteriors for subhalo parameters from lensing images. By performing a set of inference tasks on mock data, we verify the accuracy of TMNRE and show it can compute posteriors for subhalo parameters marginalized over populations of hundreds of substructures, as well as lens and source uncertainties. We also find that the multilayer perceptron (MLP) mixer network works far better for such tasks than the convolutional architectures explored in other lensing analyses. Furthermore, we show that since TMNRE learns a posterior function it enables direct statistical checks that would be extremely expensive with likelihood-based methods. Our results show that TMNRE is well-suited for analysing complex lensing data, and that the full subhalo and line-of-sight halo population must be included when measuring the properties of individual dark matter substructures with this technique.

Key words: gravitational lensing: strong – methods: statistical – dark matter.

1 INTRODUCTION

Determining the microphysical properties of the dark matter (DM) comprising about 85 per cent of the Universe’s mass is one of the key problems in physics. The distribution of DM on scales larger than dwarf galaxies is well-characterized and consistent with DM behaving as an approximately cold, collisionless, classical fluid (see e.g. Profumo 2017 for an overview). On the other hand, the distribution of DM on smaller scales is currently only roughly mapped out. At present, there is continued debate over whether the known abundance of dwarf galaxies and the density profiles of low-mass galaxies are in tension with the predictions of Lambda cold dark matter [Λ CDM; respectively dubbed the missing satellites problem (Klypin et al. 1999; Moore et al. 1999) and the cusp-core problem (de Blok & McGaugh 1997), reviewed in Bullock & Boylan-Kolchin 2017]. DM models which are warm instead of cold (Colin, Avila-Reese & Valenzuela 2000; Hogan & Dalcanton 2000), collisional instead of collisionless (Spergel & Steinhardt 2000), or quantum on

macroscopic scales rather than classical (Hu, Barkana & Gruzinov 2000) predict a diverse array of possible configurations of low-mass haloes and could potentially resolve these tensions (Buckley & Peter 2018).

Unfortunately, light DM haloes are difficult to probe as they are not expected to accumulate enough baryonic matter to form stars (Efstathiou 1992; Fitts et al. 2017). If DM has significant self-interactions, such haloes might be detectable by searching for the self-annihilation or decay products of DM (Adhikari et al. 2022). However, even if such interactions are not present, light haloehals can potentially be probed through their irreducible gravitational effects. In this work we study one such probe: galaxy–galaxy strong gravitational lensing.

In galaxy–galaxy strong lenses the light from a *source* galaxy is dramatically distorted into a ring shape by the mass of a *lens* galaxy lying close to the line of sight (LOS) to the source. This leads to multiple magnified and distorted images of the source, as explained by general relativity (GR). A *perturber* (i.e. a subhalo or LOS halo lying somewhere between the observer and source) positioned near one of these images contributes additional, much

* E-mail: n.anaumontel@uva.nl (NAM); c.weniger@uva.nl (CW)

more localized distortions. By carefully analysing the relationship between the multiple images of the source, the distortions from a perturber can be disentangled from possible variations in the source light and its properties can be measured. From measuring the distributions of perturbers' masses and other parameters, it is possible to infer population-level properties like the (sub)halo mass function parameters, which are dictated by the fundamental properties of DM.

Such analyses have been developed for two types of lensing systems: quadruply lensed quasars ('quads') and lenses with extended arcs. In the former, the source is a nearly point-like quasar that is lensed into four compact images. These images' positions and flux ratios comprise the summary statistics for these systems. The presence of a perturber near one of these images would cause anomalies in the ratios of their fluxes relative to what would be predicted assuming a smooth lens mass distribution. Evidence for flux ratio anomalies due to perturber was first found in Mao & Schneider (1998) and Dalal & Kochanek (2002) developed the first statistical analysis to measure perturbers' properties from flux ratios.

Here we focus on *gravitational imaging*, which refers to the analysis of lenses with extended arcs (Koopmans 2006; Vegetti & Koopmans 2009a; Vegetti & Koopmans 2009b). The observation in this case consists of a whole image. On one hand, such images cover a larger area of the sky than the four point-like images in quads, potentially providing more sensitivity to detect perturbations due to perturbers. On the other hand, extracting this information requires modelling the source galaxy's light, which generally has a complex morphology. Gravitational imaging has so far yielded several detections of $\sim 10^9 M_\odot$ perturbers using deep, high-resolution observations in the optical from the *Hubble Space Telescope* and Keck as well as in radio data from Atacama Large Millimeter/submillimeter Array (Vegetti, Czoske & Koopmans 2010a; Vegetti et al. 2010b, 2012; Hezaveh et al. 2016b; Diego et al. 2022). Near-future telescopes such as the Rubin Observatory, *Euclid*, JWST, and the Extremely Large Telescope will greatly increase the quality of data suitable for gravitational imaging analyses as well as its quantity, from $\mathcal{O}100$ to $\mathcal{O}10^5$ images (Collett 2015).

Established gravitational imaging analyses such as the method in Vegetti & Koopmans (2009a) and Hezaveh et al. (2016b) use *likelihood-based* inference to infer the properties of perturbers. Measurements and non-detections of individual perturbers can be converted to constraints on the (sub)halo mass function and thus DM's properties. The central mathematical object in such approaches is the likelihood, a probabilistic model $p(\mathbf{x}|\theta)$ for the data \mathbf{x} given some parameters $\theta = (\eta_{\text{lens}}, \eta_{\text{src}}, \vartheta_{\text{sub}}, \theta_{\text{other}})$ for the lens, source, perturber and possibly other (hyper)parameters.¹ Statistical inference of perturber parameters ϑ_{sub} such as mass and position given an observation \mathbf{x}_0 amounts to computing marginal posteriors $p(\vartheta_{\text{sub}}|\mathbf{x}_0)$ by means of Markov-chain Monte Carlo (MCMC) or nested sampling (Skilling 2004). Likelihood-based inference tools do not directly produce marginal posteriors but instead compute the *joint posterior* $p(\theta|\mathbf{x}_0)$, which must then be marginalized over.

The computational expense of sampling from the joint posterior imposes restrictions on the realism of lensing models that can be analysed. One such restriction common to most analyses is to assume no more than two perturbers are present in each image. Allowing for n perturbers would cause the joint posterior to become highly multimodal, with approximately $n!$ modes due to exact invariance of

the observation under relabelling of perturbers. Trans-dimensional MCMC methods provide an inroad into this problem by inferring the probabilities of different possible populations of perturbers (Brewer, Huijser & Lewis 2016; Daylan et al. 2018), albeit at substantial computational cost. Another approach is to circumvent measuring individual perturbers by instead engineering summary statistics such as the power spectrum of the residuals between the image and best-fitting reconstruction excluding substructure and relating them to the (sub)halo mass function parameters (Hezaveh et al. 2016a; Bayer et al. 2023; Diaz Rivero, Cyr-Racine & Dvorkin 2018; Çağan Şengül et al. 2020). It is unknown how much information such approaches discard, and more generally unknown how large an impact ignoring all but one perturber has on measurements. An alternative strategy involves linearizing the gravitational potential via a Taylor expansion of the lens equation. By employing a Taylor expansion, it becomes feasible to capture all small-scales DM substructures without the need to parameterize them directly in the likelihood function (Koopmans 2005; Vegetti & Koopmans 2009a; Galan et al. 2022). This technique is, therefore, able to account for the full DM subhalo population. However, it should be noted that this approach cannot capture the curl-component induced by multiplane lensing effects.

Likelihood-based analyses also typically assume a particular form of the noise and source model so that the source uncertainties can be excluded from the sampling and marginalized over analytically (Vegetti & Koopmans 2009a; Vegetti et al. 2010a, b, 2012; Hezaveh et al. 2016b). This makes it difficult to explore more complex source models described by e.g. generative machine learning methods or noise artifacts like cosmic ray streaks that cannot be described by an analytic likelihood.

An additional difficulty with likelihood-based analyses is that each run of MCMC or nested sampling produces posterior samples for just a single observation. Directly exploring the systematics, biases and other statistical properties of a particular lensing model is thus extremely time-consuming, necessitating rerunning posterior sampling many times for different input observations. This also makes analyses such as mapping perturber measurement sensitivity costly. It is noteworthy that recently Nightingale et al. (2023) pushed to the limits of how far one can feasibly go using likelihood-based analyses, fitting 54 images with five different mass models.

In this work, we demonstrate that a simulation-based inference (SBI; Cranmer, Brehmer & Louppe 2020) method called truncated marginal neural ratio estimation (Miller et al. 2020, 2021), from here on TMNRE, can circumvent these inference challenges to measure the properties of individual perturbers. SBI refers to a class of statistical inference methods that use the output of a stochastic simulator that need not have a known likelihood. In particular, neural ratio estimation (NRE), first presented in Hermans, Begy & Louppe (2020), trains a neural network to map from observations directly to *marginal posteriors* for a specified subset of model parameters (e.g. the position and mass of a perturber). This bypasses the requirement of likelihood-based inference to sample the joint posterior. In contrast to methods like approximate Bayesian computation, this also removes the need to engineer summary statistics (He et al. 2022a) as they are in effect learned directly from the training data. Since NRE learns a marginal posterior *function*, it is straightforward to check the statistical properties of the inference results for different observations. TMNRE further extends NRE by focusing training data generation in the regions of parameter space most relevant for analysing a particular observation over a sequence of inference rounds. This substantially reduces the number of simulations required to train the inference network as well as the required network complexity.

¹For example, the hyperparameters could include the pixel size for pixelated sources or strength of source regularization.

Several other works have applied machine learning and SBI to substructure lensing. We recently demonstrated that TMNRE can measure the cutoff in the warm DM subhalo mass function (SHMF) directly from images by combining multiple observations generated with a simple simulator (Anau Montel et al. 2023). In Zhang, Mishra-Sharma & Dvorkin (2022) a likelihood-ratio estimation technique similar to TMNRE was employed to measure density profile parameters of subhaloes from images. Wagner-Carena et al. (2023) recently applied neural posterior estimation to measure the SHMF normalization in mock lensing images using real galaxy images as sources. Brehmer et al. (2019) utilized a ‘likelihood-based’ SBI method requiring the simulator’s score² to measure the slope and normalization of a SHMF in simple mock images. In Ostdiek, Diaz Rivero & Dvorkin (2022a, b) image segmentation was used to classify whether each pixel in an image contained a subhalo in a given mass bin. Classifiers were also used in Alexander et al. (2020) to distinguish between different DM models based on their lensing signatures.

This work on measuring individual perturbers complements these efforts in several ways. First, it offers a path towards cross-checking current substructure measurements under different modelling assumptions. Secondly, inference based on perturbers provides a level of interpretability beyond measuring SHMF parameters directly from images, and moreover the opportunity to test different DM models through measuring the properties of individual subhaloes. Thirdly, measuring the heaviest subhaloes in an observation enables modelling them explicitly in lensing simulations, which could reduce the training data requirements and improve inference accuracy for direct SHMF measurements.

This paper is organized as follows: In Section 2 we explain our lensing model, which uses an analytic source and main lens in conjunction with well-motivated perturber models. In Section 3, we review TMNRE. Our analysis begins in Section 4.1, where we show that TMNRE is capable of recovering posteriors for a subhalo’s mass and position in the limit where they are analytically calculable. In the other analyses in Section 4, we gradually complexify our inference tasks, first accounting for the fact that the source and lens parameters are unknown and later by incorporating a population of light perturbers to marginalize over. This work will help form the basis for TMNRE-based measurement of light DM haloes in existing and future lensing data.

2 MODELLING STRONG LENS OBSERVATIONS

Here we summarize the source, main lens, perturbers, and instrument models we use to simulate mock images of gravitational lenses. We implement our lensing model in PYTORCH (Paszke et al. 2019) so that we can leverage GPUs to rapidly generate large numbers of observations.

Before delving into modelling details we briefly summarize the key points of the physics of gravitational lensing, referring the reader to e.g. Meneghetti (2016) for a more detailed overview. We assume that mass densities are low enough to treat the gravitational field of the matter in the image plane in the Newtonian approximation of GR. In this case the metric is fully characterized by the lens’ gravitational potential ψ . We also adopt the thin lens approximation, which assumes all the lens mass lies in a single *image plane* and all the source light is emitted from a *source plane*. We use ξ and x as

two-dimensional angular coordinates in the image and source planes respectively and use z to indicate distances along the orthogonal dimension. Since the image plane covers a small angular patch of the sky and the lensing deflections are small in the Newtonian limit, the coordinate system can be treated as Cartesian.

In this setting, the lens’ matter distribution can be described by its surface density

$$\Sigma(\xi) = \int dz \rho(\xi, z), \quad (1)$$

where ρ is the lens’ three-dimensional mass density and z is its redshift. The source-plane coordinate to which a light ray through the image plane traces back is given by the *lens equation*

$$x = \xi - \alpha(\xi). \quad (2)$$

Here α is the *deflection field* of the lens, which can be computed through the integral

$$\alpha(\xi) = \frac{4G}{c^2} \frac{D_{LS}}{D_L D_S} \int d[2](D_L \xi') \frac{\xi - \xi'}{|\xi - \xi'|^2} \Sigma(\xi'). \quad (3)$$

This expression involves the (angular diameter) distances D_{LS} (from the lens to the source), D_L (from the observer to the lens), and D_S (from the observer to the source).³ Since lensing merely alters the trajectories of photons rather than creating or destroying them, the surface brightness $B(\xi)$ in the image plane is equal to the surface brightness at the point to which it traces back in the source plane:

$$B(\xi) = B(x(\xi)). \quad (4)$$

Our lens model thus requires specifying the form of the deflection fields of lens components and the surface brightness of the source.

2.1 Source

The brightness profile of our mock sources is parametrized by the widely used Sérsic profile:

$$f(x) = I_e \exp \left\{ -k_n \left[\left(\frac{R(x)}{r_e} \right)^{1/n} - 1 \right] \right\}, \quad (5)$$

where r_e is the half-light radius and k_n is a normalization constant related to the index n . For $n > 0.36$ (Ciotti & Bertin 1999)

$$k_n \approx 2n - \frac{1}{3} + \frac{4}{405n} + \frac{46}{25515n^2} + \frac{131}{1148175n^3} - \frac{2194697}{30690717750n^4}. \quad (6)$$

For typical galaxies $1/2 < n < 10$. The radial parameter $R(x)$ is the length of the elliptical coordinate vector

$$\begin{pmatrix} R_x \\ R_y \end{pmatrix} = \begin{pmatrix} q^{1/2} & 0 \\ 0 & q^{-1/2} \end{pmatrix} \begin{pmatrix} \cos \varphi & \sin \varphi \\ -\sin \varphi & \cos \varphi \end{pmatrix} \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}, \quad (7)$$

which depends on the source’s position angle φ , axis ratio q , and position (x_0, y_0) . We fix the source’s redshift to $z_{\text{src}} = 2$.

Our source model therefore has seven parameters, $\eta_{\text{src}} \equiv (x_{\text{src}}, y_{\text{src}}, \varphi_{\text{src}}, q_{\text{src}}, n, r_e, I_e)$.

2.2 Main lens

We adopt the singular power law ellipsoid (SPLE) model for the main lens galaxy, which is capable of modelling the gravitational

²The score is the derivative of the log-likelihood for a given observation with respect to the model’s parameters.

³We compute these with ASTROPY (Astropy Collaboration 2013, 2018) using the flat cosmology from Planck (Planck Collaboration 2020).

potentials of strong lenses to near the per cent level (Suyu et al. 2009). The SPLE deflection field can be expressed in closed-form as a complex field $\alpha = \alpha_x + i\alpha_y$ (Tessore & Metcalf 2015; O’Riordan, Warren & Mortlock 2020):

$$\alpha^{\text{SPLE}}(\xi) = \theta_E \frac{2q_{\text{lens}}^{1/2}}{1 + q_{\text{lens}}} \left(\frac{\theta_E}{R} \right)^{\gamma-2} e^{i\phi} \cdot {}_2F_1 \left(1, \frac{\gamma-1}{2}, \frac{5-\gamma}{2}, -\frac{1-q_{\text{lens}}}{1+q_{\text{lens}}} e^{2i\phi} \right). \quad (8)$$

Here (R, ϕ) are elliptical coordinates, related to the Cartesian coordinates ξ through a transformation parametrized by the lens’ orientation φ_{lens} , axis ratio q_{lens} , and position $(x_{\text{lens}}, y_{\text{lens}})$:

$$\begin{pmatrix} R_x \\ R_y \end{pmatrix} = \begin{pmatrix} q_{\text{lens}}^{1/2} & 0 \\ 0 & q_{\text{lens}}^{-1/2} \end{pmatrix} \begin{pmatrix} \cos \varphi_{\text{lens}} & \sin \varphi_{\text{lens}} \\ -\sin \varphi_{\text{lens}} & \cos \varphi_{\text{lens}} \end{pmatrix} \begin{pmatrix} \xi_x - x_{\text{lens}} \\ \xi_y - y_{\text{lens}} \end{pmatrix}, \quad (9)$$

$$\tan \phi = \frac{R_y}{R_x}. \quad (10)$$

Since the hypergeometric function ${}_2F_1$ is not implemented in PYTORCH, we instead pretabulate its value as a function of ϕ , q_{lens} , and γ and interpolate, as described in Chianese et al. (2020).

The slope γ has a complicated degeneracy with the size of the source (Schneider & Sluse 2013, 2014). Roughly, larger γ values cause the spatial scale of the source to increase (Nightingale & Dye 2015, section 3.3). For simplicity, we fix $\gamma = 2.1$. We also assume the lens galaxy’s light has been perfectly subtracted, and fix its redshift to $z_{\text{lens}} = 0.5$.

To account for the weak lensing due to large-scale structure located along the LOS to the source, we also include an external shear component, which is constant across the image plane:

$$\alpha^{\text{shear}}(\xi) = \begin{pmatrix} \gamma_1 & \gamma_2 \\ \gamma_2 & -\gamma_1 \end{pmatrix} \xi. \quad (11)$$

Our main lens model thus has seven parameters: the SPLE parameters $(x_{\text{lens}}, y_{\text{lens}}, \varphi_{\text{lens}}, q_{\text{lens}}, \theta_E)$ and the external shear parameters (γ_1, γ_2) , which we denote collectively with η_{lens} .

2.3 Perturbers

2.3.1 Density profiles

We model the deflection field of subhaloes using a truncated Navarro–Frenk–White (NFW) profile (Baltz, Marshall & Oguri 2009) to account for tidal stripping by the main lens:

$$\rho_{\text{NFW}}(r) = \frac{\rho_s}{r/r_s(1+r/r_s)^2} \frac{1}{1+r_t^2/r^2}. \quad (12)$$

Here r is the distance from the centre of the subhalo, ρ_s is the density normalization, r_s is the scale radius, and r_t is the truncation radius. The deflection field for this density profile is given in Baltz et al. (2009, appendix A), and differs from that of an NFW profile for $r \gtrsim r_t$. While the value of $\tau \equiv r_t/r_s$ depends on the full history of the subhalo, it typically falls between 4 and 10 (Gilman et al. 2020); we fix $\tau = 6$ for simplicity. For simplicity, we model LOS haloes using exactly the same profile even though they typically have not undergone tidal stripping.

To generate perturber populations for our third analysis task, we must choose values for their density normalizations and scale radii. Since simulation studies typically measure the halo mass m_{sub}^4 and

⁴This is defined as the mass of the halo enclosed in a sphere where the untruncated halo’s average density is 200 times the critical density.

the concentration c , it is more convenient to sample populations from distributions over these parameters. These variables can then be mapped to the parametrization above via

$$r_s = \frac{1}{c} \left[\frac{3m_{\text{sub}}}{4\pi 200\rho_{\text{cr}}(z_{\text{lens}})} \right]^{1/3}, \quad (13)$$

$$\rho_s = \rho_{\text{cr}}(z_{\text{lens}}) \frac{1}{3} \frac{c^3}{\log(1+c) - c/(1+c)}. \quad (14)$$

For simplicity we fix $c = 15$, which is roughly the average value for perturbers in the mass range 1×10^7 to $1 \times 10^{10} M_{\odot}$ (Richings et al. 2021, fig. 7). We anticipate that accounting for scatter in the mass-concentration relation might actually improve our ability to measure subhaloes’ parameters as higher concentrations lead to substantially stronger lensing signals (Amorisco et al. 2022).

The parameters of an individual subhalo which are not fixed are thus $\mathcal{P}_{\text{sub}} \equiv (x_{\text{sub}}, y_{\text{sub}}, m_{\text{sub}})$, where the second and third components are the projected position of the subhalo. In the case of LOS haloes, the parameter set also includes the redshift z_{los} . In the next two subsections, we describe how we sample these parameters.

2.3.2 Generating subhaloes

We sample subhalo masses using a mass function of the form from Giocoli et al. (2010):

$$\frac{dn}{d \log m_{200}} = m_{200}(1+z_{\text{lens}})^{1/2} A_M m_{200}^{-\alpha} \exp \left[-\beta \left(\frac{m_{200}}{M_{200}} \right)^3 \right], \quad (15)$$

where M_{200} is the mass of the main lens. The free parameters in this function were fit to hydrodynamical cosmological simulations that included baryons in Despali & Vegetti (2017). In particular, we use the fits to EAGLE, which give $\alpha = 0.85$ (given in the text) and $(A_M, \beta) = (2.4 \times 10^{-4} M_{\odot}^{-\alpha}, 300)$ (extracted from their figures). Integrating the mass function over a given mass interval gives the expected number of subhaloes in that interval distributed throughout the whole main lens.

Despali & Vegetti (2017) found the distribution of radial coordinates in hydrodynamical simulations is well-fit by an Einasto profile, but can be approximated as uniform over the lens plane. For a given lensing system, we thus precompute \bar{n}_{sub} , the number of subhaloes expected to fall within the lens plane. Thereafter, we generate the subhalo population by sampling the number of subhaloes from Poisson(\bar{n}_{sub}), drawing their masses from the SHMF and sampling their projected positions uniformly over the lens plane. Since the vast majority of subhaloes fall outside the lens plane, we expect their lensing effect to be mostly degenerate with external shear, and thus do not simulate them. With the lens redshift we have chosen, over a $5 \text{ arcsec} \times 5 \text{ arcsec}$ image and integrating over the mass range $1 \times 10^7 M_{\odot}$ to $1 \times 10^8 M_{\odot}$, we find $\bar{n}_{\text{sub}} = 3.1$.

2.3.3 Generating line-of-sight haloes

As described in Şengül et al. (2020) and Anau Montel et al. (2023), we first compute the average number of LOS haloes in the double-pyramid geometry connecting the observer, lens-plane, and source. For each simulation we sample the number of LOS haloes from Poisson(\bar{n}_{los}). We then sample their redshifts and projected positions uniformly over the double-pyramid region and draw their masses from the mass function in Tinker et al. (2008), with Δ set to 200. For the lens and source redshifts, we have chosen $\bar{n}_{\text{los}} = 265.6$.

To avoid expensive iterative ray-tracing through the lens planes of each LOS halo, we project them as effective subhaloes into the lens plane, using the relations derived in Şengül et al. (2020) to rescale their scale radii and masses. As with subhaloes, we ignore any LOS haloes lying outside the double pyramid volume.

It should be noted that effective convergence methods, like the one we adopt (Şengül et al. 2020), do not fully capture the subtleties and degeneracies of multiplane lens analysis, by disregarding how, when a DM small-scale halo is not in the lens plane, the lens mass model can absorb its lensing signal (Fleury, Larena & Uzan 2021; Amorisco et al. 2022; He et al. 2022b). The omission of this effect may lead to an overestimate of the LOS haloes contribution and needs to be addressed before this technique can be safely used for the analysis of real data.

2.4 Instrumental effects

We generate mock data with comparable quality to *Hubble Space Telescope* observations. All images are $5 \text{ arcsec} \times 5 \text{ arcsec}$ with 0.05 arcsec resolution ($100 \text{ pixels} \times 100 \text{ pixels}$). In our simulations, we do not include a point-spread function (PSF) for simplicity, but this component cannot be disregarded in real data analysis. To account for the fact that each pixel in the image corresponds to a finite collecting region in the sky, we generate our images at a resolution eight times higher than the target resolution and downsample. In experiments we found that neglecting this effect can have a significant impact on inference results. Lastly, we add Gaussian pixel noise to our observations such that the brightest pixels are approximately 30 times the noise level.

3 INFERENCE WITH TRUNCATED MARGINAL NEURAL RATIO ESTIMATION

In the inference tasks we confront in the rest of this work, our goal is to infer two-dimensional marginals for the position and one-dimensional marginals for the mass of a subhalo. Each posterior is to be marginalized over the other perturber parameters and potentially another set of parameters η for the main lens, source, and perturber population. In this section, we review how TMNRE solves such inference problems.

To begin with, NRE (Hermans et al. 2020) is a technique for inferring the posterior $p(\vartheta|\mathbf{x})$ for a model with the joint distribution $p(\mathbf{x}, \vartheta)$, where \mathbf{x} is an observation (e.g. a lensing image) and ϑ is a parameter of interest (e.g. the mass of a subhalo). The idea is to train a classifier to distinguish between data and parameters drawn from two classes labelled by the binary variable C :

$$p(\mathbf{x}, \vartheta | C = 0) = p(\mathbf{x})p(\vartheta) \quad (16)$$

$$p(\mathbf{x}, \vartheta | C = 1) = p(\mathbf{x}, \vartheta). \quad (17)$$

These two distributions correspond respectively to simulating data from the simulator and drawing an unrelated set of parameters from the prior versus sampling parameters and data from the simulator. Sampling $C = 0$ and $C = 1$ with equal probability, the decision function for the (Bayes-)optimal classifier can be computed using Bayes' theorem:

$$p(C = 1|\mathbf{x}, \vartheta) = \frac{p(\mathbf{x}, \vartheta)}{p(\mathbf{x}, \vartheta) + p(\mathbf{x})p(\vartheta)} \equiv \sigma[\log r(\mathbf{x}, \vartheta)], \quad (18)$$

where we introduced the sigmoid function $\sigma(y) \equiv 1/(1 + e^{-y})$ and the likelihood-to-evidence ratio:

$$r(\mathbf{x}, \vartheta) \equiv \frac{p(\mathbf{x}|\vartheta)}{p(\mathbf{x})} = \frac{p(\mathbf{x}, \vartheta)}{p(\mathbf{x})p(\vartheta)} = \frac{p(\vartheta|\mathbf{x})}{p(\vartheta)}. \quad (19)$$

Therefore, by training a neural network $\hat{r}_\phi(\mathbf{x}, \vartheta)$ to estimate $r(\mathbf{x}, \vartheta)$ via this supervised classification task,⁵ we obtain an estimate of the posterior through $\hat{p}_\phi(\vartheta|\mathbf{x}) = \hat{r}_\phi(\mathbf{x}, \vartheta)p(\vartheta)$. This *ratio estimator* can be trained by minimizing the binary cross-entropy loss

$$\ell[\hat{r}_\phi] = - \int d\mathbf{x} d\vartheta \left\{ p(\mathbf{x}, \vartheta) \log \sigma[\log \hat{r}_\phi(\mathbf{x}, \vartheta)] + p(\mathbf{x})p(\vartheta) \log [1 - \sigma[\log \hat{r}_\phi(\mathbf{x}, \vartheta)]] \right\} \quad (20)$$

with respect to the ratio estimator's parameters ϕ using stochastic gradient descent techniques. Critically, training only requires the ability to generate *samples* from the simulator. This makes it straightforward to apply marginal ratio estimation in scenarios where the explicit form of the likelihood cannot be written in closed form. In practice, posterior samples can be generated by resampling prior samples (with replacement) weighted by the ratio, enabling posterior sampling even when the prior cannot be expressed in closed form.

The extension to estimating marginal posteriors is straightforward: parameters to be marginalized over must be sampled, but *not* presented to the ratio estimator. In more detail, consider a model with the joint distribution $p(\mathbf{x}, \eta, \vartheta) = p(\mathbf{x}|\eta, \vartheta)p(\eta, \vartheta)$, where η is a set of parameters to be marginalized over (e.g. the source and main lens parameters). If η is not passed to the ratio estimator, the loss function becomes

$$\ell[\hat{r}_\phi] = - \int d\mathbf{x} d\vartheta d\eta \left\{ p(\mathbf{x}, \eta, \vartheta) \log \sigma[\log \hat{r}_\phi(\mathbf{x}, \vartheta)] + p(\mathbf{x})p(\eta, \vartheta) \log [1 - \sigma[\log \hat{r}_\phi(\mathbf{x}, \vartheta)]] \right\} \quad (21)$$

$$= - \int d\mathbf{x} d\vartheta \left\{ p(\mathbf{x}, \vartheta) \log \sigma[\log \hat{r}_\phi(\mathbf{x}, \vartheta)] + p(\mathbf{x})p(\vartheta) \log [1 - \sigma[\log \hat{r}_\phi(\mathbf{x}, \vartheta)]] \right\} \quad (22)$$

where we integrated over η to obtain the second equality, proving our statement.

The ratio estimators discussed so far are fully *amortized*: that is, they attempt to learn $r(\mathbf{x}, \vartheta)$ over the whole range of the prior $p(\eta, \vartheta)$. In principle, it is useful to be able to analyse any possible observation with the same network. In practice, when the posterior $p(\eta, \vartheta|\mathbf{x}_0)$ for a particular observation \mathbf{x}_0 is much narrower than the prior, training an accurate ratio estimator requires a massive amount of training data. We instead focus on the problem of *targeted inference* of the posterior for \mathbf{x}_0 , which substantially reduces training data requirements and reduces the complexity of the function the ratio estimator must learn to model. Such an approach is also well-suited to individually targeting the small sample of lenses relevant to DM substructure measurement that exist at present ($\mathcal{O}(100)$).

Training targeted ratio estimators is achieved by replacing the prior with a *truncated prior* $p_\Gamma(\eta, \vartheta)$, where the parameters are restricted to a region Γ where they are likely to have generated \mathbf{x}_0 . Since parameters from the complement of Γ are unlikely to have generated \mathbf{x}_0 , training a ratio estimator with data generated from the truncated prior as opposed to the full prior has little impact on the posterior learned by our ratio estimators.

Since the highest probability density region of the true posterior Γ is unknown, we compute an estimate $\hat{\Gamma}$ over a sequence of inference rounds. At the beginning of each round, we sample from $p_\Gamma(\eta, \vartheta)$ (or the true prior in the first round) and train a ratio estimator. We re-estimate $\hat{\Gamma}$ by keeping only the parts of the previous truncated prior for which $\hat{r}_\phi(\mathbf{x}, \vartheta)$ exceeds a certain threshold, as described

⁵For better numerical stability, we actually train the network to learn $\log r(\mathbf{x}, \vartheta)$.

in Miller et al. (2021). This determines the truncated prior for the next round. Our final ratio estimator is obtained when $\hat{\Gamma}$ stops changing substantially between rounds. This whole procedure is called TMNRE.

TMNRE is related to other *sequential* SBI methods, such as sequential neural posterior estimation (Papamakarios & Murray 2016; Lueckmann et al. 2017; Greenberg, Nonnenmacher & Macke 2019) and sequential neural likelihood estimation (Papamakarios, Sterratt & Murray 2018). These two methods use the *posteriors* learned in each round to generate simulations for the next round rather than the truncated prior. This approach is inefficient for learning multiple marginal posteriors simultaneously, since sampling from the marginal for a particular parameter may hinder learning the marginals for other parameters.

The fact that TMNRE learns a function that can be rapidly evaluated makes it possible to perform statistical consistency checks. In this work, we perform expected coverage checks (Cole et al. 2022; Hermans et al. 2021) to test the calibration of our ratio estimators for observations generated using parameters from the truncated prior. This test measures whether credible regions of different widths have achieved their nominal coverage (i.e. whether the true parameters fall within the 68 per cent credible interval of the estimated posterior for 68 per cent of observations). Agreement between the nominal and empirically measured expected coverage is a necessary (but not sufficient) condition for the ratio estimator to be a correct estimate of the posterior. While typically expected coverage tests are a statement about the ratio estimator's properties averaged over the truncated prior, at increased computational cost the coverage can be checked in a frequentist manner as a function of the true parameters.

4 RESULTS

We now apply TMNRE to three different substructure lensing problems of increasing complexity. For all tasks we use the same general ratio estimator architecture. It consists of an initial compression network that maps the $100 \text{ pixel} \times 100 \text{ pixel}$ images into a feature vector. This feature vector is concatenated to $\boldsymbol{\vartheta}_{\text{sub}}$ (and separately to $\boldsymbol{\eta}_{\text{src}}$ and $\boldsymbol{\eta}_{\text{lens}}$ for tasks where they are also inferred). The vector is then passed to a multilayer perceptron (MLP) which outputs an estimate of the two-dimensional and one-dimensional marginal likelihood-to-evidence ratios for $(x_{\text{sub}}, y_{\text{sub}})$ and $\log_{10} m_{\text{sub}}/M_{\odot}$, respectively (with separate MLPs used to estimate the one-dimensional ratios for $\boldsymbol{\eta}_{\text{src}}$ and $\boldsymbol{\eta}_{\text{lens}}$).

For each ratio estimator we begin the first training round with 10 000 training examples. We then truncate each parameter's prior. If none of the truncated priors shrank by at least 20 per cent, we increase the number of training examples by a factor of 1.5 for the next inference round. A fresh network is then trained using simulations drawn from the truncated prior. Convergence of the ratio estimator is declared after five such consecutive increases in the training set size. For tasks in which we must infer the macromodel parameters, we first train the macromodel ratio estimator using this procedure and use the resulting truncated priors to generate training data for the subhalo ratio estimators using the same training procedure. We use the implementation of TMNRE in SWYFT⁶ (Miller et al. 2022), which is built on PYTORCH and PYTORCH-LIGHTNING.⁷

The training data for our ratio estimators differs in important ways from typical data sets studied by machine learning researchers, making the choice of a good compression network an interesting challenge. Consider, for example, the machine learning problem of classifying the content of natural images. Natural images are distinguished by a hierarchy of visual features at different scales (for example, small-scale features such as textures and edges which comprise large-scale features like the head of an animal or part of an object). A good image classifier should be translation-invariant, producing the same output regardless of the position of an image's contents. Since the deep convolutional neural network (CNN) architecture has an inductive bias towards learning a hierarchy of features and are translation invariant, CNNs are widely used in computer vision.

The training data for our ratio estimators does not share these features. Different perturber configurations produce images with slightly different relationships between the multiple images of the source galaxy. The variations between images lie near the Einstein ring, and do not show the same rich hierarchical structure of natural images. This means that inductive biases of CNNs are not necessarily beneficial in the context of substructure lensing.

In our experiments, we used CNNs in the ratio estimators for the macromodel parameters, finding their performance to be adequate. However, we found they produced much too wide two-dimensional marginals for the position of a subhalo. Instead, we found the MLP Mixer (Tolstikhin et al. 2021) to work well.⁸ Roughly, the MLP Mixer splits the image into patches, stacks the patches, and passes each pixel in the stack through an MLP, acting as a dilated convolution. Another MLP is then applied along the channel dimension of the mixed patches, and the process is iterated. The MLP Mixer thus directly examines the relationships between pixels in disparate parts of the image, which is exactly how the properties of subhaloes are imprinted. We expect that other architectures that split the image into patches such as Vision Transformer (ViT; Dosovitskiy et al. 2020) could work well for the compression network, though ViT is known to require large amounts of training data.

The architectures of our macromodel and subhalo compression networks are given in Appendix A. While we did not perform a full hyperparameter exploration, we found the batchnorm layers to be crucial for stable training of the CNN used for the macromodel ratio estimator. Since our images are roughly one-quarter the area of the images studied in the paper introducing MLP Mixer, we use a substantially smaller model than they suggest. Using dropout in the MLP Mixer and classifier MLPs improved performance. Varying the number of hidden layers and their size in the classifiers had little impact.

We used the Adam optimizer with an initial learning rate of 6×10^{-3} for the macromodel ratio estimator and 4×10^{-4} for the subhalo ratio estimator (found through a learning rate test) and a batch size of 64. The learning rate was reduced by a factor of 0.1 whenever the validation loss plateaued for three epochs. Training was run for no longer than 30 epochs.

4.1 Subhalo position inference with fixed mass, source, and lens

We first consider the case where the only free parameters in the lens are the position of a single $10^9 M_{\odot}$ subhalo, $\boldsymbol{\vartheta}_{\text{sub}} = (x_{\text{sub}}, y_{\text{sub}})$. The prior is taken to be uniform over the image plane (i.e. $\mathcal{U}(-2.5, 2.5)$)

⁶<https://github.com/undark-lab/swyft/>

⁷<https://www.pytorchlightning.ai/>

⁸The MLP Mixer implementation we use can be found at <https://github.com/lucidrains/mlp-mixer-pytorch>.

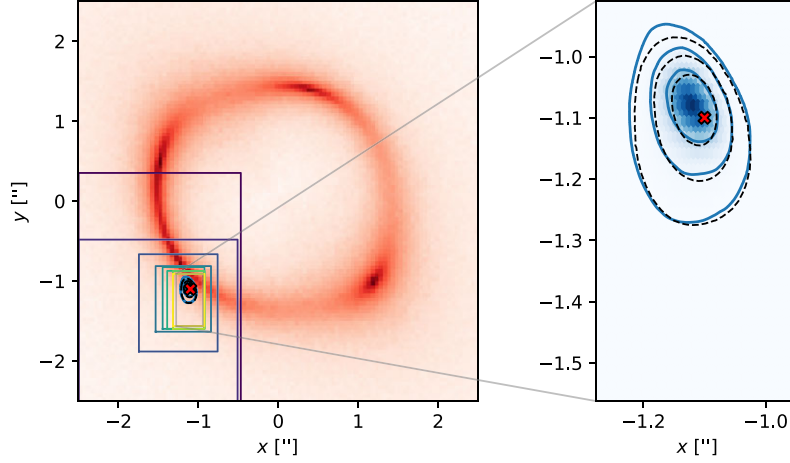


Figure 1. Validation of TMNRE through inference of the position of a subhalo, with macromodel parameters fixed to their true values and the subhalo’s mass fixed to $10^9 M_{\odot}$. The observation is shown in the left panel. The blue and dashed black contours correspond to the posterior inferred with TMNRE and computed analytically respectively, indicating the 68, 95, and 99.7 per cent credible regions. The red \times shows the subhalo’s true position. The blue through yellow boxes in the left panel show the ranges of the truncated prior based on the one-dimensional marginals for the subhalo’s coordinates. The zoom-in on the right encompasses the range of the final truncated prior. The distorted blue hex-bin histogram shows the magnitude of the inferred posterior.

for both coordinates). The posterior for ϑ_{sub} can then be computed analytically. Adopting a uniform prior over ϑ_{sub} covering the image plane and using the fact the posterior is much narrower, we have

$$\log p(\vartheta_{\text{sub}}|\mathbf{x}) \sim -\frac{1}{2} \sum_{i,j} \left(\frac{x_{ij} - f_{ij}(\vartheta_{\text{sub}})}{\sigma_n} \right)^2, \quad (23)$$

where the sum runs over pixels and we dropped terms independent of ϑ_{sub} .

Fig. 1 shows the truncation regions for each round and compares the analytically computed posterior with the posterior inferred using TMNRE. While the truncation regions and posterior estimates in early rounds are extremely broad compared to the analytically computed posterior, TMNRE successfully identifies the region of the image containing the subhalo. After 10 inference rounds the truncation region stabilizes and the inferred posteriors agree well with the true ones for each coordinate. To complement this visual check, we also check the coverage for samples from the final round of TMNRE in Fig. 2. We find the empirical and nominal coverage to be in good agreement, with our ratio estimator very slightly underestimating the width of the posterior beyond the 95 per cent confidence level.

Having validated TMNRE in this simple scenario, we now turn to more complex inference tasks where the posteriors of interest cannot be derived analytically.

4.2 Subhalo mass and position inference

Next we aim to infer the position and mass of a single subhalo, $\vartheta_{\text{sub}} = (m_{\text{sub}}, x_{\text{sub}}, y_{\text{sub}})$, in a system where the source and main lens parameters are also unknown. The priors for the 17 parameters of the model are given in Table 1. Due to the relatively low dimensionality, inference on this model is within the reach of likelihood-based tools such as MCMC or nested sampling. In addition, it can be implemented in a differentiable manner, making the application of methods such as Hamiltonian Monte Carlo (HMC) possible (Chianese et al. 2020; Gu et al. 2022). Running such expensive scans is beyond the scope of this paper.

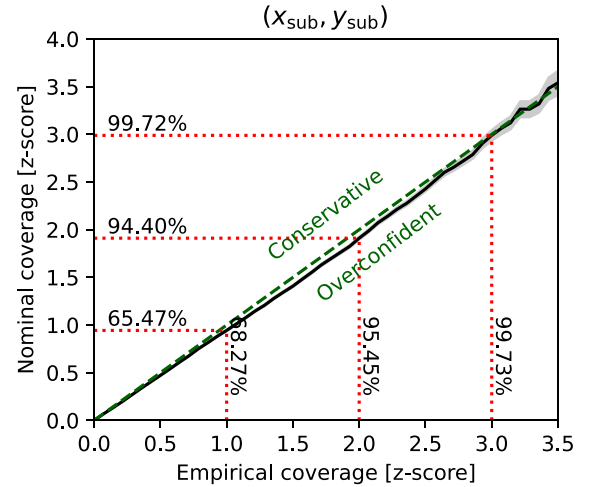


Figure 2. Coverage plot for inference task where only the subhalo’s position is free (see Fig. 1), showing our ratio estimator produces posteriors of the correct size on average. In detail, the black curve shows the empirical versus nominal coverage, estimated by computing posteriors for 10 000 observations drawn from the final truncated prior. The statistical uncertainty of this estimate is plotted in grey; its derivation is explained in detail in Cole et al. (2022). For a perfectly calibrated ratio estimator, the black curve would lie along the diagonal green dashed line. The red dashed lines indicate the empirical and nominal coverage of the 1σ – 3σ credible regions.

The final posteriors for the subhalo parameters are shown in Fig. 3. The true values of all parameters fall within the ~ 68 per cent credible intervals of the inferred posteriors. We find the effect of the uncertain macromodel is not too strong (at least for this noise realization), with the size of the subhalo position posterior being comparable to what we found in the previous inference task. Fig. 4 demonstrates that our ratio estimator has good coverage with respect to the constrained prior. In Figs 5 and 6 we display the marginal posteriors and coverage plots for all 14 source and main lens parameters, which demonstrate they are well-calibrated.

Table 1. True subhalo and macromodel parameter values and priors used in the first TMNRE inference round in our three inference tasks. The last column references the first section in which the indicated parameter is inferred rather than being fixed to its true value. The slope of the main lens is fixed to 2.1, as explained in Section 2.2. The main lens and source redshifts are set to $z_{\text{lens}} = 0.5$ and $z_{\text{src}} = 2$, respectively. For the analysis in Section 4.3 involving a population of light perturbers, we sample the number of LOS and subhaloes from Poisson distributions with means $\bar{n}_{\text{los}} = 265.6$ and $\bar{n}_{\text{sub}} = 3.1$, respectively, and restrict their masses to the range 1×10^7 to $1 \times 10^8 M_{\odot}$. The halo mass functions and redshift distributions are described in detail in Section 2.3. For all perturbers, we fix the concentration to $c = 15$ and truncation scale $\tau = r_l/r_s = 6$.

	Parameter	True value	Initial prior	First inferred in
Subhalo	$x_{\text{sub}} ["]$	-1.1	$\mathcal{U}(-2.5, 2.5)$	Section 4.1
	$y_{\text{sub}} ["]$	-1.1	$\mathcal{U}(-2.5, 2.5)$	Section 4.1
	$\log_{10} m_{\text{sub}}/M_{\odot}$	9.5	$\mathcal{U}(8, 10.5)$	Section 4.2
SPLE	$x_{\text{lens}} ["]$	-0.05	$\mathcal{U}(-0.2, 0.2)$	Section 4.2
	$y_{\text{lens}} ["]$	0.1	$\mathcal{U}(-0.2, 0.2)$	
	$\phi_{\text{lens}} [^{\circ}]$	1	$\mathcal{U}(0.5, 1.5)$	
	q_{lens}	0.75	$\mathcal{U}(0.5, 1)$	
	γ	2.1	—	
	$r_{\text{ein}} ["]$	1.5	$\mathcal{U}(1, 2)$	
Shear	γ_1	0.005	$\mathcal{U}(-0.5, 0.5)$	Section 4.2
	γ_2	-0.010	$\mathcal{U}(-0.5, 0.5)$	
Source	$x_{\text{src}} ["]$	0	$\mathcal{U}(-0.2, 0.2)$	Section 4.2
	$y_{\text{src}} ["]$	0	$\mathcal{U}(-0.2, 0.2)$	
	$\phi_{\text{src}} [^{\circ}]$	0.75	$\mathcal{U}(0.5, 1.25)$	
	q_{src}	0.5	$\mathcal{U}(0.2, 0.8)$	
	n	2.3	$\mathcal{U}(1.5, 3)$	
	$r_e ["]$	2.0	$\mathcal{U}(0.5, 3)$	
	I_e	0.6	$\mathcal{U}(0.1, 2)$	

4.3 Subhalo mass and position inference with a population of perturbers

For our final inference task we extend the previous one by aiming to infer the position and mass of a relatively heavy target subhalo while marginalizing over a population of lighter perturbers of unknown size. The priors for the perturber population are summarized in Table 1 and Section 2.3. Our lensing images contain on average about 260 LOS haloes and three subhaloes in the lens plane. This means on average about 800 parameters are required to describe such images. Likelihood-based sampling of this high-dimensional, transdimensional posterior requires techniques such as reversible-jump MCMC (Brewer et al. 2016; Daylan et al. 2018). To marginalize over the perturber population with TMNRE, their parameters are sampled over during data generation but not passed to the ratio estimator.

Since the population of perturbers can contain a member with mass greater than our target subhalo, we need to make this inference task well-defined by ‘labelling’ the subhaloes of interest. We accomplish this by making the perturber population lighter than the target subhalo, with mass restricted to the range 1×10^7 to $1 \times 10^8 M_{\odot}$. We further assume the target subhalo has been localized to a 1.4 arcsec \times 1.4 arcsec patch of the image around its true position.

The final-round inference results for ϑ_{sub} plotted in Fig. 7 show that inclusion of the perturber population has a substantial effect on the posteriors. The posterior for the subhalo’s mass peaks around the true value, but has a long tail extending towards the lower boundary of the prior. This indicates we are only able to obtain an upper bound on the subhalo mass rather than a measurement, and cannot exclude

the possibility its mass is the lowest value consistent with the prior. Having validated our analysis in simpler cases and checked our ratio estimator has good coverage, we conclude our marginal posteriors are in fact close to the true ones.

Our results are roughly in line with the image segmentation analysis of Ostdiek et al. (2022a, b), which found subhaloes of mass above roughly $10^{8.5} M_{\odot}$ were resolvable in similar mock observations. In addition, while the 68 per cent credible region for the subhalo’s position contains its true position, the 95 and 99.7 per cent credible regions cover nearly the whole prior region.

The posteriors for the source and lens parameters are shown in Fig. 8. While some of the parameters’ posteriors have comparable widths to those found in the previous inference task (namely $\phi_{\text{src/lens}}$, $q_{\text{src/lens}}$, the source index, I_e , γ_1 , and γ_2), others are measured much less precisely due to the stochastic perturber population ($x_{\text{src/lens}}$, $y_{\text{src/lens}}$, r_e , and r_{Ein}). We omit coverage plots for this analysis as they are of comparably good quality to those in the previous subsection.

5 DISCUSSION AND CONCLUSIONS

Measuring the properties of individual DM haloes on subgalactic scales is an important probe of the fundamental nature of DM. However, extracting their parameters from observations is difficult for a myriad of reasons, including the fact that lenses contain multiple perturbers (sub-/LOS haloes). In this work, we demonstrated that TMNRE enables analyses of individual perturbers’ properties in scenarios where the application of likelihood-based methods is difficult or infeasible. The key strength of TMNRE is its ability to directly learn marginal posterior functions for a set of scientifically interesting parameters from simulated data. By truncating the range of parameters used to generate the simulations, TMNRE enables precision inference of individual observations using a targeted set of training data. This enables the previously intractable marginalization over large perturber populations. Furthermore, the method is applicable to simulators with unknown likelihood functions and large or even variable numbers of input parameters. The resulting inference networks can be poked and prodded to confirm they are statistically well-behaved.

With three lensing simulators of varying complexity, we demonstrated the following characteristics of the method and perturber inference:

5.1 TMNRE can recover existing results

We verified the accuracy of TMNRE by confirming it reproduces analytically calculable posteriors in a toy lensing scenario with known macromodel parameters and subhalo mass.

5.2 TMNRE enables statistical checks

Since the inference networks learned by TMNRE are locally amortized over a range of potential observations, we were able to test their statistical consistency. Our checks confirm that TMNRE on average produces posteriors with the correct width for the macromodel and subhalo parameters. Such tests would be extremely expensive with likelihood-based inference since they would require rerunning the sampling machinery on numerous mock observations.

5.3 The perturber population matters

We demonstrated that the sensitivity with which a subhalo’s parameters are measurable can be significantly degraded when marginalizing

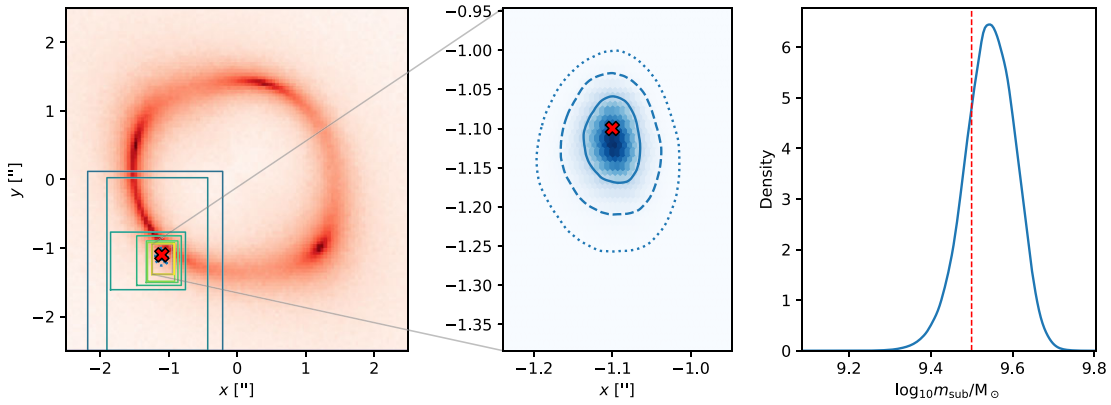


Figure 3. Marginal posteriors inferred with TMNRE for a subhalo’s two-dimensional position (left and centre) and mass (right) in a lens with unknown macromodel parameters. See the caption of Fig. 1 for further details, though note we have instead used solid, dashed, and dotted lines, respectively, to mark the 68, 95, and 99.7 per cent credible regions of the position posterior. The range of the x -axis in the right panel shows the final-round truncated prior for the subhalo’s \log_{10} -mass.

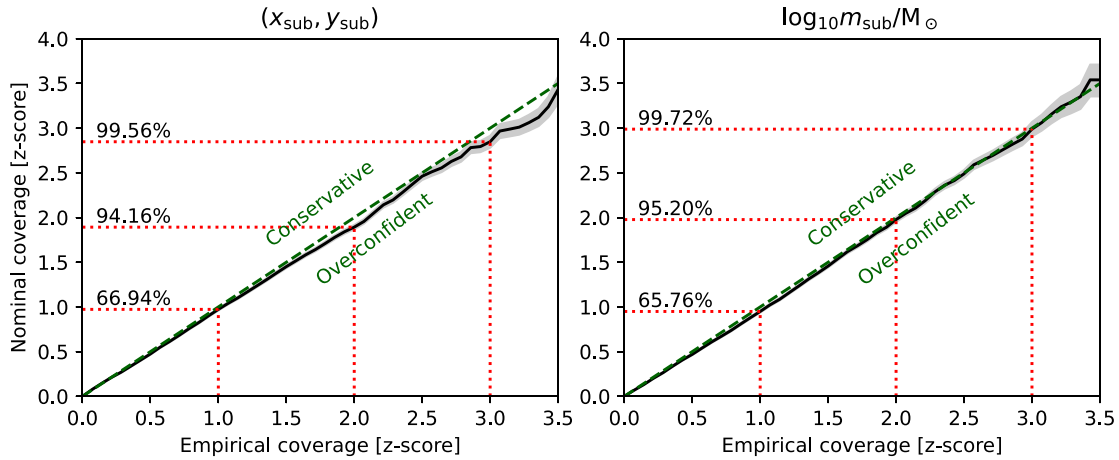


Figure 4. Coverage plots for subhalo position and mass ratio estimators learned from the observation in Fig. 3. These again indicate the estimators’ credible regions are on average the correct size for observations drawn from the final-round truncated prior. See Fig. 2 for an explanation of the format.

over a population of perturbers. While the 1σ regions of our position and mass posteriors were centred on the subhalo’s true parameters, they had heavy tails extending to the boundaries of our tight, manually fixed priors. Given our validation, statistical checks and the fact TMNRE is so far the only method capable of performing the high-dimensional marginalization required for this analysis, *our results therefore suggest that the population of light perturbers should not be neglected*. However, it is important to highlight that we cannot strictly conclude from our results that the presence of a substructure population makes the inference of the properties of an individual subhalo unfeasible. What we find is that it makes the task more challenging for our particular SBI approach and network architectures. Whether better network architectures for the ratio estimator capable of modelling the posterior more accurately than MLP Mixer, or maybe the proper handling of the problem with a full transdimensional likelihood-based MCMC method dealing with the perturber population can resolve the issue remains open, and an important question to study in future work.

While this work used simple mock lenses, TMNRE makes it possible to add realism and parameters to a simulator without significantly altering the inference procedure, or necessarily increasing the simulation budget (Cole et al. 2022). It should, therefore, be straightforward to incorporate various complexities we ignored in this work: a mass-concentration relation for the perturbers, the lens galaxy’s light, the (possibly uncertain) PSF, multiband observations, drizzling, and even complex noise with an unknown likelihood function. Our analysis can also in principle incorporate more complex source models based on (for example) shapelets (Birrer, Amara & Refregier 2015, 2017), Gaussian processes (Karchev, Coogan & Weniger 2022) or neural networks (Chianese et al. 2020). We expect source models capable of refining fine details to improve our measurement precision since the lensing distortions from substructure scale with the gradient of the source.

Another interesting direction for further work is the use of TMNRE for model comparison. While here our ratio estimators were trained to compute the likelihood-to-evidence ratio, as pointed out in Hermans et al. (2020) it is possible to learn other ratios of densities. In

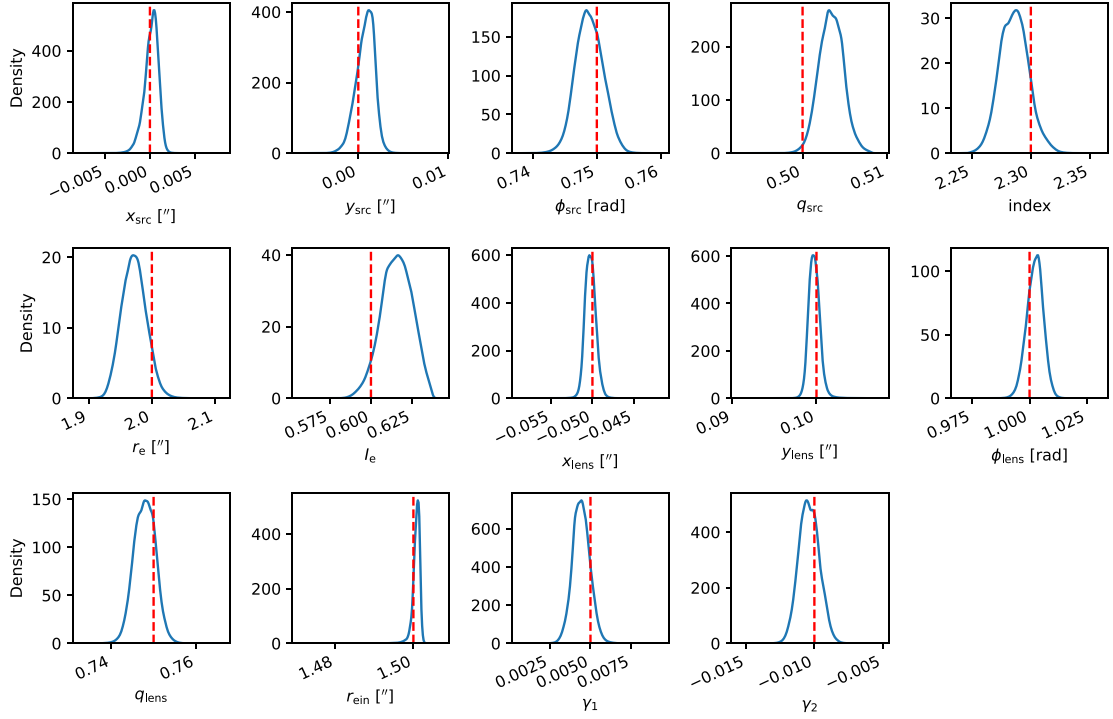


Figure 5. The one-dimensional marginal posteriors for all macromodel parameters of the lensing system shown in Fig. 3. The posteriors were computed using a CNN-based ratio estimator. The first seven panels correspond to the source parameters, the next five are for the main lens and the last two are for the external shear. All posteriors encompass the true parameter values (vertical red dashed lines) within the $\sim 2\sigma$ interval.

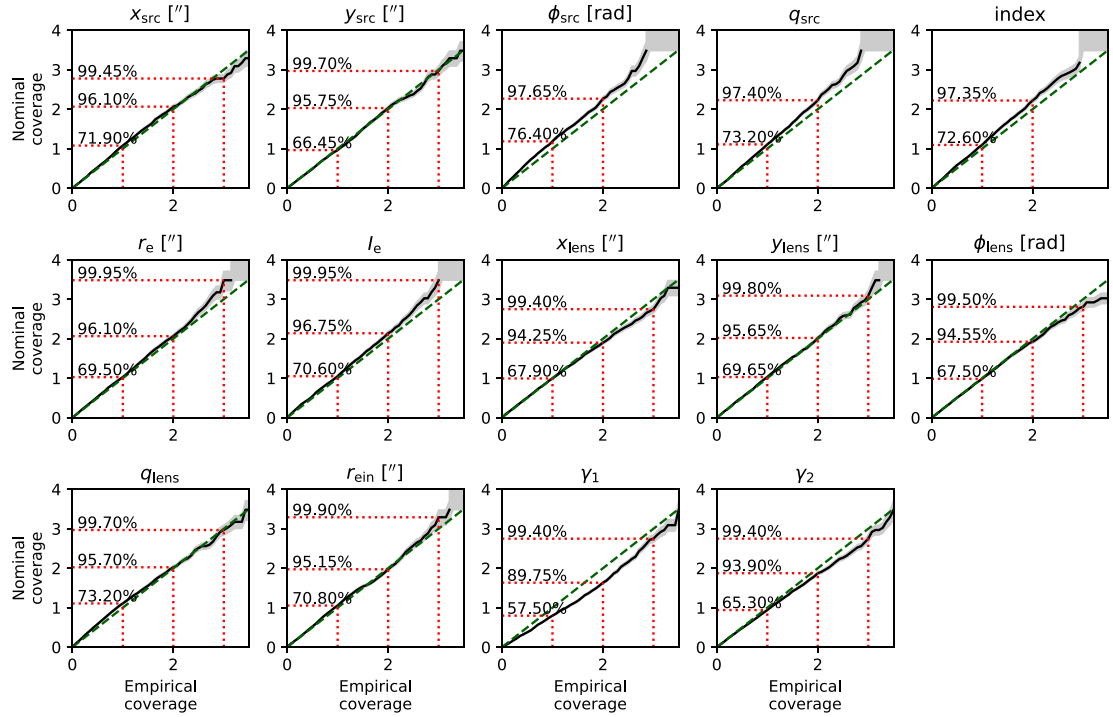


Figure 6. Coverage plots for the one-dimensional marginal macromodel parameter posteriors of the lensing system from Fig. 3, using the same format as in Fig. 2. The posteriors generally have coverage, with a few being slightly conservative (ϕ_{src} , q_{src} , and the source index) and the shear posteriors being slightly overconfident.

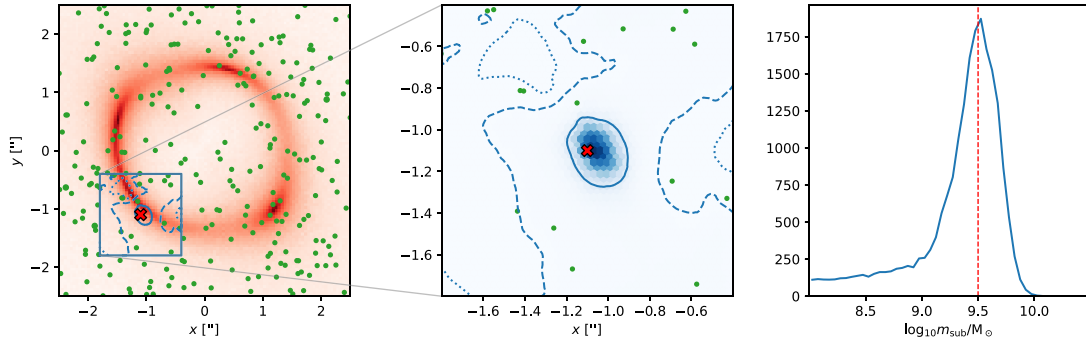


Figure 7. Subhalo position and mass posteriors obtained with TMNRE, now marginalizing over a population of $1 \times 10^7 M_{\odot}$ to $1 \times 10^8 M_{\odot}$ LOS/subhaloes (green dots) in addition to the unknown macromodel. See the caption of Fig. 3 for details. The initial prior for the subhalo’s position is indicated by the blue box in the left panel. The range of the x -axis in the right panel shows the prior on the \log_{10} of its mass. For both the subhalo’s mass and position, the width of the inferred posteriors prevents TMNRE from truncating the priors.

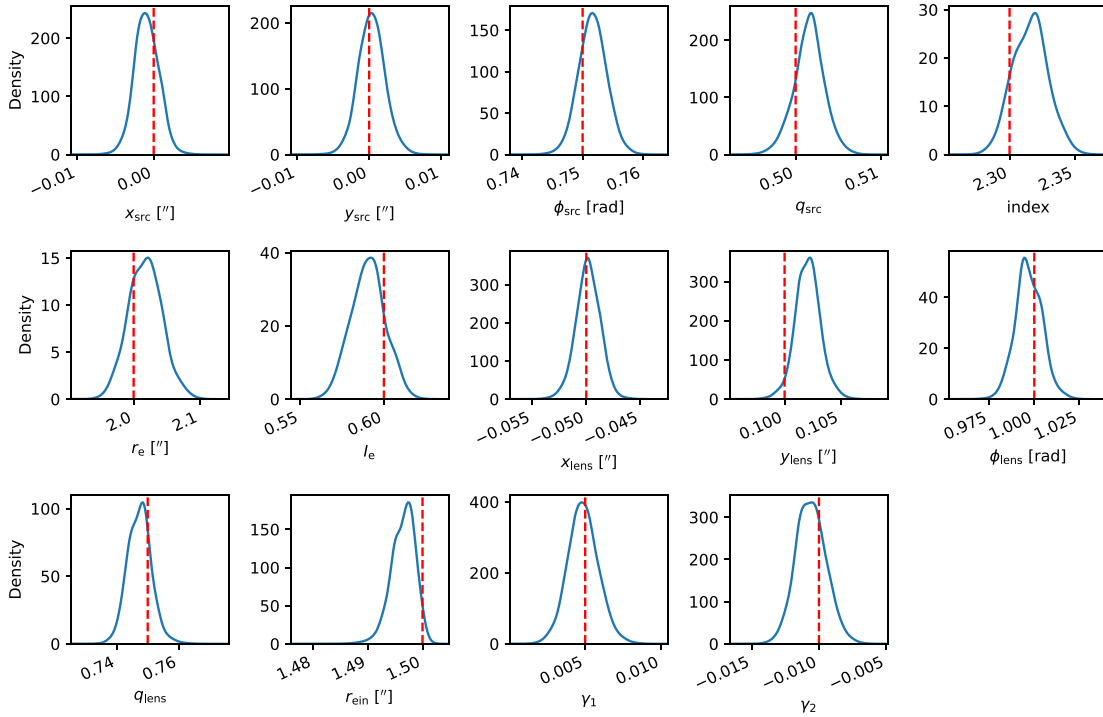


Figure 8. Macromodel one-dimensional marginal posteriors as in Fig. 6, but for the inference task where a population of 1×10^7 to $1 \times 10^8 M_{\odot}$ are present in the observation. This has the effect of broadening most of the posteriors.

particular ratio estimators can be used to learn the Bayes factor for assessing the strength of the evidence for different models. This could be used to determine whether an image contains a perturber or not, and to map the minimum-detectable perturber mass as a function of its position.

Overall, we believe using TMNRE to measure perturbers as described in this work in combination with measuring the (sub)halo mass function directly (Anau Montel et al. 2023) provides a promising path towards uncovering the identity of DM.

ACKNOWLEDGEMENTS

We thank Benjamin Kurt Miller and Elias Dubbeldam for helpful discussions.

AC received funding from the Netherlands eScience Center (grant number ETEC.2019.018) and the Schmidt Futures Foundation. This work is part of a project that has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement no. 864035).

This work was carried out on the Lisa Compute Cluster at SURFsara, which runs on 100 per cent wind energy.

Software used: astropy (Astropy Collaboration 2013, 2018), JUPYTER (Kluyver et al. 2016), MATPLOTLIB (Hunter 2007), NUMPY (Harris et al. 2020), PYTORCH (Paszke et al. 2019), PYTORCH-LIGHTNING,⁹ SEABORN (Waskom 2021), and SWYFT (Miller et al. 2022).

⁹<https://www.pytorchlightning.ai/>

DATA AVAILABILITY

The data underlying this article will be shared on request to the corresponding author.

REFERENCES

- Adhikari S. et al., 2022, preprint([arXiv:2207.10638](https://arxiv.org/abs/2207.10638))
- Alexander S., Gleyzer S., McDonough E., Toomey M. W., Usai E., 2020, *ApJ*, 893, 15
- Amorisco N. C. et al., 2022, *MNRAS*, 510, 2464
- Anau Montel N., Coogan A., Correa C., Karchev K., Weniger C., 2023, *MNRAS*, 518, 2746
- Astropy Collaboration, 2013, *A&A*, 558, A33
- Astropy Collaboration, 2018, *AJ*, 156, 123
- Baltz E. A., Marshall P., Oguri M., 2009, *J. Cosmol. Astropart. Phys.*, 2009, 015
- Bayer D., Chatterjee S., Koopmans L. V. E., Vegetti S., McKean J. P., Treu T., Fassnacht C. D., 2023, *MNRAS*, 523, 1310
- Birrer S., Amara A., Refregier A., 2015, *ApJ*, 813, 102
- Birrer S., Amara A., Refregier A., 2017, *J. Cosmol. Astropart. Phys.*, 05, 037
- Brehmer J., Mishra-Sharma S., Hermans J., Louppe G., Cranmer K., 2019, *ApJ*, 886, 49
- Brewer B. J., Huijser D., Lewis G. F., 2016, *MNRAS*, 455, 1819
- Buckley M. R., Peter A. H. G., 2018, *Phys. Rept.*, 761, 1
- Bullock J. S., Boylan-Kolchin M., 2017, *ARA&A*, 55, 343
- Chianese M., Coogan A., Hofma P., Otten S., Weniger C., 2020, *MNRAS*, 496, 381
- Ciotti L., Bertin G., 1999, *A&A*, 352, 447
- Çağan Şengül A., Tsang A., Diaz Rivero A., Dvorkin C., Zhu H.-M., Seljak U., 2020, *Phys. Rev. D*, 102, 063502
- Cole A., Miller B. K., Witte S. J., Cai M. X., Grootes M. W., Nattino F., Weniger C., 2022, *J. Cosmol. Astropart. Phys.*, 2022, 004
- Colin P., Avila-Reese V., Valenzuela O., 2000, *ApJ*, 542, 622
- Collett T. E., 2015, *ApJ*, 811, 20
- Cranmer K., Brehmer J., Louppe G., 2020, *Proc. Natl. Acad. Sci.*, 117, 30055
- de Blok W. J. G., McGaugh S. S., 1997, *MNRAS*, 290, 533
- Dalal N., Kochanek C. S., 2002, *ApJ*, 572, 25
- Daylan T., Cyr-Racine F.-Y., Diaz Rivero A., Dvorkin C., Finkbeiner D. P., 2018, *ApJ*, 854, 141
- Despali G., Vegetti S., 2017, *MNRAS*, 469, 1997
- Diaz Rivero A., Cyr-Racine F.-Y., Dvorkin C., 2018, *Phys. Rev. D*, 97, 023001
- Diego J. M., Pascale M., Kavanagh B. J., Kelly P., Dai L., Frye B., Broadhurst T., 2022, *A&A*, 665, A134
- Dosovitskiy A. et al., 2020, preprint([arXiv:2010.11929](https://arxiv.org/abs/2010.11929))
- Efstathiou G., 1992, *MNRAS*, 256, 43P
- Fitts A. et al., 2017, *MNRAS*, 471, 3547
- Fleury P., Larena J., Uzan J.-P., 2021, *J. Cosmol. Astropart. Phys.*, 2021, 024
- Galan A., Varnardos G., Peel A., Courbin F., Starck J.-L., 2022, *A&A*, 668, A155
- Gilman D., Birrer S., Nierenberg A., Treu T., Du X., Benson A., 2020, *MNRAS*, 491, 6077
- Giocoli C., Tormen G., Sheth R. K., van den Bosch F. C., 2010, *MNRAS*, 404, 502
- Greenberg D. S., Nonnenmacher M., Macke J. H., 2019, in Chaudhuri K., Salakhutdinov R., eds, Proceedings of the 36th International Conference on Machine Learning, Vol. 97, PMLR, p. 2404
- Gu A. et al., 2022, *ApJ*, 935, 49
- Harris C. R. et al., 2020, *Nature*, 585, 357
- He Q. et al., 2022a, *MNRAS*, 511, 3046
- He Q. et al., 2022b, *MNRAS*, 512, 5862
- Hermans J., Begy V., Louppe G., 2020, preprint([arXiv:1903.04057](https://arxiv.org/abs/1903.04057))
- Hermans J., Delaunoy A., Rozet F., Wehenkel A., Louppe G., 2021, *TMLR*, preprint([arXiv:2110.06581](https://arxiv.org/abs/2110.06581))
- Hezaveh Y., Dalal N., Holder G., Kisner T., Kuhlen M., Perreault Levasseur L., 2016a, *J. Cosmol. Astropart. Phys.*, 11, 048
- Hezaveh Y. D. et al., 2016b, *ApJ*, 823, 37
- Hogan C. J., Dalcanton J. J., 2000, *Phys. Rev. D*, 62, 063511
- Hu W., Barkana R., Gruzinov A., 2000, *Phys. Rev. Lett.*, 85, 1158
- Hunter J. D., 2007, *Comput. Sci. Eng.*, 9, 90
- Karchev K., Coogan A., Weniger C., 2022, *MNRAS*, 512, 661
- Kluyver T. et al., 2016, in Loizides F., Schmidt B., eds, Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, Amsterdam, p. 87
- Klypin A. A., Kravtsov A. V., Valenzuela O., Prada F., 1999, *ApJ*, 522, 82
- Koopmans L. V. E., 2005, *MNRAS*, 363, 1136
- Koopmans L. V. E., 2006, *EAS Publ. Ser.*, 20, 161
- Lueckmann J.-M., Goncalves P. J., Bassetto G., Ócal K., Nonnenmacher M., Macke J. H., 2017, *NeurIPS*, preprint([arXiv:1711.01861](https://arxiv.org/abs/1711.01861))
- Mao S.-d., Schneider P., 1998, *MNRAS*, 295, 587
- Meneghetti M., 2016, in Introduction to Gravitational Lensing, Springer Nature, Switzerland
- Miller B. K., Cole A., Louppe G., Weniger C., 2020, *NeurIPS*, preprint ([arXiv:2011.13951](https://arxiv.org/abs/2011.13951))
- Miller B. K., Cole A., Forré P., Louppe G., Weniger C., 2021, in Ranzato M., Beygelzimer A., Dauphin Y., Liang P., Vaughan J. W., eds, Advances in Neural Information Processing Systems, Vol. 34. Curran Associates, Inc., p. 129, available at: <https://proceedings.neurips.cc/paper/2021/file/01632f7b7a127233fa1188bd6c2e42e1-Paper.pdf>
- Miller B. K., Cole A., Weniger C., Nattino F., Ku O., Grootes M. W., 2022, *J. Open Source Softw.*, 7, 4205
- Moore B., Ghigna S., Governato F., Lake G., Quinn T. R., Stadel J., Tozzi P., 1999, *ApJ*, 524, L19
- Nightingale J. W., Dye S., 2015, *MNRAS*, 452, 2940
- Nightingale J. W. et al., 2023, *Scanning For Dark Matter Subhalos in Hubble Space Telescope Imaging of 54 Strong Lenses*, preprint([arXiv:2209.10566](https://arxiv.org/abs/2209.10566))
- O'Riordan C. M., Warren S. J., Mortlock D. J., 2020, *MNRAS*, 496, 3424
- Ostdiek B., Diaz Rivero A., Dvorkin C., 2022a, *A&A*, 657, L14
- Ostdiek B., Diaz Rivero A., Dvorkin C., 2022b, *ApJ*, 927, 83
- Papamakarios G., Murray I., 2016, *NeurIPS*, preprint([arXiv:1605.06376](https://arxiv.org/abs/1605.06376))
- Papamakarios G., Sterratt D. C., Murray I., 2018, Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, 89, PMLR, Naha, Okinawa
- Paszke A. et al., 2019, in Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., eds, Advances in Neural Information Processing Systems, PyTorch: An Imperative Style, High-Performance Deep Learning Library, Vol. 32, Curran Associates, Inc., p. 8024, available at: <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- Planck Collaboration, 2020, *A&A*, 641, A6
- Profumo S., 2017, An Introduction to Particle Dark Matter. World Scientific Press, Singapore
- Richings J., Frenk C., Jenkins A., Robertson A., Schaller M., 2021, *MNRAS*, 501, 4657
- Schneider P., Sluse D., 2013, *A&A*, 559, A37
- Schneider P., Sluse D., 2014, *A&A*, 564, A103
- Skilling J., 2004, in Fischer R., Preuss R., Toussaint U. V., eds, AIP Conf. Ser. Vol. 735, Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Am. Inst. Phys., New York, p. 395
- Spergel D. N., Steinhardt P. J., 2000, *Phys. Rev. Lett.*, 84, 3760
- Suyu S. H., Marshall P. J., Blandford R. D., Fassnacht C. D., Koopmans L. V. E., McKean J. P., Treu T., 2009, *ApJ*, 691, 277
- Tessore N., Metcalf R. B., 2015, *A&A*, 580, A79
- Tinker J. L., Kravtsov A. V., Klypin A., Abazajian K., Warren M. S., Yepes G., Gottlob S., Holz D. E., 2008, *ApJ*, 688, 709
- Tolstikhin I. O. et al., 2021, preprint([arXiv:2105.01601](https://arxiv.org/abs/2105.01601))
- Vegetti S., Koopmans L. V. E., 2009a, *MNRAS*, 392, 945
- Vegetti S., Koopmans L. V. E., 2009b, *MNRAS*, 400, 1583
- Vegetti S., Czoske O., Koopmans L. V. E., 2010a, *MNRAS*, 407, 225
- Vegetti S., Koopmans L. V. E., Bolton A., Treu T., Gavazzi R., 2010b, *MNRAS*, 408, 1969
- Vegetti S., Lagattuta D. J., McKean J. P., Auger M. W., Fassnacht C. D., Koopmans L. V. E., 2012, *Nature*, 481, 341

- Wagner-Carena S., Aalbers J., Birrer S., Nadler E. O., Darragh-Ford E., Marshall P. J., Wechsler R. H., 2023, *ApJ*, 942, 75
Waskom M. L., 2021, *J. Open Source Softw.*, 6, 3021
Zhang G., Mishra-Sharma S., Dvorkin C., 2022, *MNRAS*, 517, 4317

APPENDIX A: COMPRESSION NETWORK ARCHITECTURES

The compressor architectures are given in Table A1 and Table A2. Note that we standardize the images before providing them to the networks.

Table A1. The convolutional compression network used in the macromodel parameter ratio estimator. The notation is taken from PYTORCH: the arguments to Conv2d are the number of input channels, output channels, kernel size, stride, and padding, respectively. The horizontal lines demarcate where the number of channels changes. The output of the network is flattened into a vector with 128 features.

```
Conv2d(1, 4, 8, 2, 1,
bias = False)
BatchNorm2d(4)
LeakyReLU(0.2)

Conv2d(4, 8, 8, 2, 1,
bias = False)
BatchNorm2d(8)
LeakyReLU(0.2)

Conv2d(8, 16, 8, 2, 1,
bias = False)
BatchNorm2d(16)
LeakyReLU(0.2)

Conv2d(16, 32, 8, 2, 1,
bias = False)
BatchNorm2d(32)
LeakyReLU(0.2)
```

Table A2. The details of the MLP Mixer compression network in the sub-halo ratio estimator. We use the implementation from <https://github.com/lucidrains/mlp-mixer-pytorch>, with arguments given in the table.

image_size	100
channels	1
patch_size	10
dim	256
depth	4
num_classes	32
dropout	0.1

This paper has been typeset from a \TeX/L\AA\TeX file prepared by the author.