A&A 667, A123 (2022)
https://doi.org/10.1051/0004-6361/202243401

Astronomy & Astrophysics

TDCOSMO

IX. Systematic comparison between lens modelling software programs: Time-delay prediction for WGD 2038−4008

A. J. Shajib[1,2,3,⋆], K. C. Wong[4,5], S. Birrer[6,7], S. H. Suyu[8,9,10], T. Treu[3,⋆⋆], E. J. Buckley-Geer[11,1], H. Lin[11], C. E. Rusu[4], J. Poh[1,2], A. Palmese[12,11,⋆], A. Agnello[13], M. W. Auger-Williams[14,15], A. Galan[16], S. Schuldt[8,9], D. Sluse[17], F. Courbin[16], J. Frieman[1,2,11], and M. Millon[16]

[1] Department of Astronomy & Astrophysics, University of Chicago, Chicago, IL 60637, USA
e-mail: ajshajib@uchicago.edu
[2] Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA
[3] Department of Physics and Astronomy, University of California, Los Angeles, CA 90095, USA
[4] National Astronomical Observatory of Japan (NAOJ), National Institutes of Natural Sciences, 2-21 Osawa, Mitaka, Tokyo 181-8588, Japan
[5] Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan
[6] Kavli Institute for Particle Astrophysics and Cosmology and Department of Physics, Stanford University, Stanford, CA 94305, USA
[7] SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA
[8] Max Planck Institute for Astrophysics, Karl-Schwarzschild-Str. 1, 85748 Garching, Germany
[9] Technische Universität München, Physik-Department, James-Franck-Str. 1, 85748 Garching, Germany
[10] Institute of Astronomy and Astrophysics, Academia Sinica, 11F of ASMAB, No. 1, Section 4, Roosevelt Road, Taipei 10617, Taiwan
[11] Fermi National Accelerator Laboratory, PO Box 500, Batavia, IL 60510, USA
[12] Department of Astronomy, University of California, Berkeley, 501 Campbell Hall, Berkeley, CA 94720, USA
[13] DARK, Niels Bohr Institute, Jagtvej 128, 2200 Copenhagen, Denmark
[14] Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK
[15] Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[16] Institute of Physics, Laboratoire d'Astrophysique, École Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland
[17] STAR Institute, Quartier Agora, Allée du Six Août, 19c, 4000 Liege, Belgium

Received 23 February 2022 / Accepted 2 September 2022

ABSTRACT

The importance of alternative methods for measuring the Hubble constant, such as time-delay cosmography, is highlighted by the recent Hubble tension. It is paramount to thoroughly investigate and rule out systematic biases in all measurement methods before we can accept new physics as the source of this tension. In this study, we perform a check for systematic biases in the lens modelling procedure of time-delay cosmography by comparing independent and blind time-delay predictions of the system WGD 2038−4008 from two teams using two different software programs: GLEE and LENSTRONOMY. The predicted time delays from the two teams incorporate the stellar kinematics of the deflector and the external convergence from line-of-sight structures. The un-blinded time-delay predictions from the two teams agree within 1.2σ, implying that once the time delay is measured the inferred Hubble constant will also be mutually consistent. However, there is a $\sim 4\sigma$ discrepancy between the power-law model slope and external shear, which is a significant discrepancy at the level of lens models before the stellar kinematics and the external convergence are incorporated. We identify the difference in the reconstructed point spread function (PSF) to be the source of this discrepancy. When the same reconstructed PSF was used by both teams, we achieved excellent agreement, within $\sim 0.6\sigma$, indicating that potential systematics stemming from source reconstruction algorithms and investigator choices are well under control. We recommend that future studies supersample the PSF as needed and marginalize over multiple algorithms or realizations for the PSF reconstruction to mitigate the systematics associated with the PSF. A future study will measure the time delays of the system WGD 2038−4008 and infer the Hubble constant based on our mass models.

Key words. gravitational lensing: strong – methods: data analysis – galaxies: elliptical and lenticular, cD – distance scale

1. Introduction

The Hubble constant, H_0, is a central cosmological parameter as it sets the expansion rate of the Universe. Consequently, precise knowledge of its value is crucial for our understanding of the Cosmos, and it also has important implications in extragalactic astrophysics. However, different methods have measured the Hubble constant with discrepant values, producing the so-called Hubble tension (e.g. Freedman 2021). Mapping of the temperature fluctuations of the cosmic microwave background allows one to measure the Hubble parameter $H(z \approx 1100)$ at the last scattering surface, and then the Hubble constant, H_0, at the

⋆ NHFP Einstein fellow.
⋆⋆ Packard fellow.

current epoch is extrapolated using Λ cold dark matter (ΛCDM) cosmology. This early-Universe probe resulted in constraints of $H_0 = 67.4 \pm 0.5\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Planck Collaboration VI 2020) and $H_0 = 67.6 \pm 1.1\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Aiola et al. 2020). In the local Universe, H_0 is typically measured by building a cosmic distance ladder up to type Ia supernovae (SNe) in the Hubble flow by calibrating their absolute magnitudes with intermediate distance probes. The Supernova H_0 for the Equation of State of dark energy (SH0ES) team used Cepheids and parallax distances to calibrate the cosmic distance ladder, measuring $H_0 = 73.04 \pm 1.04\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Riess et al. 2022), which is in 5σ tension with the *Planck* measurement. The Carnegie–Chicago Hubble Project used the tip of the red giant branch to calibrate the distance ladder, measuring $H_0 = 69.6 \pm 1.9\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Freedman et al. 2019, 2020), which, interestingly, is statistically consistent with both the SH0ES and *Planck* measurements. Several other local probes strengthened the Hubble tension, for example the Megamaser Cosmology Project measured $H_0 = 73.9 \pm 3.0\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Pesce et al. 2020), the Tully–Fisher method calibrated with Cepheids measured $H_0 = 75.1 \pm 0.2 \pm 3.0\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Kourkchi et al. 2020), and the surface brightness fluctuation method measured $H_0 = 73.7 \pm 0.7 \pm 2.4\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Blakeslee et al. 2021). If systematics in these measurements can be ruled out as the source of this Hubble tension, new physics beyond the standard ΛCDM cosmology will be required to resolve the tension (e.g. Poulin et al. 2019; Knox & Millea 2020; Efstathiou 2021). Therefore, thoroughly investigating the potential systematics that are as yet unknown in each of the probes is paramount.

Strong-lensing time delays provide an independent probe of the Hubble constant (Refsdal 1964). The delays between the arrival times of photons corresponding to different images of the background source depend on the cosmological distances involved in the strong-lensing system, and thus these delays allow us to measure a combination of these distances, called the ‘time-delay distance’ (Suyu et al. 2010). The time-delay distance is inversely proportional to H_0 and weakly dependent on other cosmological parameters. Although early implementations of this method in the 1990s and the early 2000s suffered from limitations in data quality and analysis techniques, both of these aspects have improved by a large margin over the past decade (for a review with a historical perspective, see Treu & Marshall 2016). Inferring H_0 from the time delays requires: (i) measuring the time delays, (ii) measuring the redshifts of the deflector and the background quasar, (iii) modelling the mass distribution in the central deflector to compute the Fermat potential differences between the image positions, and (iv) estimating the extra lensing contribution from the line-of-sight (LOS) mass distribution between the background source and the observer. Thanks to breakthroughs in all of these factors, the H_0 Lenses In the COSMOGRAIL’s Wellspring (H0LiCOW) and the Strong-lensing High Angular Resolution Programme (SHARP) collaborations measured $H_0 = 73.3^{+1.7}_{-1.8}\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ from a sample of six strongly lensed quasar systems (Suyu et al. 2017; Bonvin et al. 2017; Birrer et al. 2019; Chen et al. 2019; Rusu et al. 2020; Wong et al. 2020). The STRong-lensing Insights into the Dark Energy Survey (STRIDES) collaboration analysed a seventh lens system to measure $H_0 = 74.2^{+2.7}_{-3.0}\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Shajib et al. 2020). It is noteworthy that six out of these seven analyses were performed blindly, with only the first one being a non-blind analysis. The H0LiCOW, STRIDES, Cosmological Monitoring of Gravitational Lenses (COSMOGRAIL), and SHARP collaborations have united under the umbrella of the Time-Delay COSMOgraphy (TDCOSMO) collaboration.

The TDCOSMO collaboration has already performed a number of tests to search for previously unknown systematics. Millon et al. (2020, TDCOSMO-I) checked for systematics arising from the current treatments of the stellar kinematics, LOS mass distribution, and the choice of lens model families, finding no evidence for unaccounted errors. Gilman et al. (2020, TDCOSMO-III) find that dark sub-halos – which are ignored in lens modelling through the assumption of smooth mass profiles – also do not systematically bias the H_0 inference, adding negligible random uncertainty. Birrer et al. (2020, TDCOSMO-IV) relaxed the assumption of the power-law mass distribution in the deflector galaxies to allow maximal degeneracy in the mass distribution under the mass-sheet transformation (MST; Falco et al. 1985). By constraining the mass distribution from the stellar kinematics only, these authors inferred $H_0 = 74.5^{+5.6}_{-6.1}\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, that is, relaxing the power-law assumption leads to an increase in H_0 uncertainty from 2.2% to 7.9% for the sample of the seven analysed systems. To regain the lost precision, TDCOSMO-IV combined an external sample of galaxy–galaxy strong lenses from the Sloan Lens ACS[1] (SLACS) survey to add more information on the galaxy mass distribution, under the assumption that the SLACS lenses and the TDCOSMO lenses belong to the same galaxy population. Adding a sample of 33 SLACS lenses improved the precision to 5.4%. Although the point estimate of H_0 shifted to $H_0 = 67.4^{+4.1}_{-3.2}\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ with the addition of the SLACS lenses, this value is still consistent with all the previous TDCOSMO measurements within 1σ. Birrer & Treu (2021, TDCOSMO-V) forecasted that a future sample of 40 time-delay lenses with spatially resolved stellar kinematics and an external lens sample of 200 non-time-delay lenses will be able to infer H_0 with 1.2–1.3% precision, which is necessary to independently settle the Hubble tension at the ~5σ confidence level. Van de Vyvere et al. (2022a, TDCOSMO-VII) find that the systematic bias in the measured H_0 arising from the boxy-ness or discy-ness of the deflector galaxy is <1% and thus insignificant. Blind data challenges are also important tests for the presence of systematics. The Time-Delay Challenge validated the robustness of the methods currently used to measure time delays from quasar light curves (Dobler et al. 2015; Liao et al. 2015). The Time-Delay Lens Modelling Challenge similarly validated the modelling techniques currently used to recover the ground truth when the shapes of the underlying galaxy mass profiles are known (Ding et al. 2021).

In this paper we present the results of an experiment to search for potential systematics in the lens modelling – within specific assumed mass profile families – that may arise from different modelling software programs used by different investigators. In this experiment, two teams using different software programs independently modelled the strongly lensed quasar system WGD 2038−4008 to the level required for cosmographic application (i.e. to the noise level; Agnello et al. 2018). The two modelling software programs being compared are GLEE[2] and LENSTRONOMY[3]. The core members of the GLEE team are K. C. Wong and S. H. Suyu; the core members of the LENSTRONOMY team are A. J. Shajib, S. Birrer, and T. Treu. Both of the software programs have previously been used for lens modelling in cosmographic analyses by the TDCOSMO

[1] Advanced Camera for Surveys.

[2] GLEE is developed by A. Halkola and S. H. Suyu (Suyu & Halkola 2010; Suyu et al. 2012).

[3] The lead developer of LENSTRONOMY is S. Birrer. LENSTRONOMY also received numerous contributions from the community. The full list of contributors is provided at: https://github.com/lenstronomy/lenstronomy/blob/main/AUTHORS.rst.

collaboration – five systems with GLEE and two systems with LENSTRONOMY. Although Birrer et al. (2016) performed a cosmographic analysis outside the TDCOSMO umbrella using LENSTRONOMY for the system RX J1131−1231, which was previously analysed by the H0LiCOW collaboration using GLEE, a systematic blind comparison between the two software programs on the same lens system has not been done previously. Both software programs perform parametric modelling of the deflector mass distribution, but they differ in the method used for source reconstruction. Whereas GLEE uses a pixel-based source reconstruction with regularization conditions (Suyu et al. 2006), LENSTRONOMY uses a basis set of parameterized profiles for source reconstruction (Birrer et al. 2015, 2021; Birrer & Amara 2018).

In addition to the software architectures, differences in the lens models may arise from modelling choices made by an investigator in such modelling processes. Our experiment also encompasses this human aspect of the modelling process by having the two teams work independently and blindly. However, to facilitate a fair comparison between the model predictions, we established a baseline model setup with minimal specifications that was agreed upon by the two teams before performing their own analyses. After each team separately completed their internal systematic checks and went through an internal review by the TDCOSMO collaboration, the lens models were frozen and the model predictions were un-blinded to make comparisons between the two teams. As the time delay for this system has not yet been measured with sufficient precision for an H_0 measurement, we leave the H_0 inference from our models to be done in the future. However, we predict the time delays for this system as a function of H_0 after marginalizing over the inferences from the two modelling software programs. As a result, our 'preemptive' lens models enforce an additional layer of blindness for the future H_0 measurement from this system.

The baseline models for comparison have two different lens model setups: (i) a power-law mass model and (ii) a two-component mass model that individually accounts for the dark and baryonic components. It is well known that conventional parametric models such as the power-law model impose assumptions that break the mass-sheet degeneracy (MSD; e.g. Birrer et al. 2020; Kochanek 2020). However, a lens model is still useful for extracting the relevant lensing information (i.e. the Fermat potential difference) from the data, which can then be processed to allow the additional freedom along the MSD following TDCOSMO-IV. Although techniques to extract lensing information without relying on parametric models have recently been proposed (e.g. Birrer 2021), they have not yet been applied to real systems for rigorous lens modelling similar to the TDCOSMO analyses. Furthermore, no evidence has so far demonstrated that the simply parametrized models are not an adequate description, and the necessity or physical reality of a mass component that acts as a physical mass sheet has not been demonstrated. For all these reasons, until new evidence is gathered to inform new choices, simply parametrized lens models are going to be the baseline in TDCOSMO analyses. Therefore, it is important to compare the modelling methods based on these software programs to check for systematic differences as performed in this paper.

In this paper we only predict the time delays for WGD 2038−4008 based on our lens models, as the actual time delays for this system are yet to be measured and thus the H_0 cannot be inferred. Measuring H_0 based on the lens models presented in this paper is left for a future paper.

This paper is organized as follows. In Sect. 2 we provide a brief review of the strong lensing formalism to establish the notations and describe the Bayesian inference framework of our model predictions. The observables in our analysis are described in Sect. 3. We present the baseline models that are common to both teams in Sect. 4. The modelling procedures and results are presented by the GLEE and LENSTRONOMY teams in Sects. 5 and 6, respectively. We compare and discuss the results from the two teams in Sect. 7 and conclude the paper in Sect. 8. Sections 1–6 were written prior to the un-blinding. After un-blinding on October 22, 2021, Sects. 7 and 8 were written and no major edits were done to Sects. 1–6, except for minor fixes for typos and grammatical errors.

2. Framework of the lens modelling

In this section we describe the theoretical framework for our analysis. We give a brief overview of the strong lensing formalism in Sect. 2.1, discuss the MSD in Sect. 2.2, explain our modelling of the stellar kinematics in Sect. 2.3, and present the Bayesian inference framework for our analysis in Sect. 2.4.

2.1. Strong lensing formalism

The goal of this section is to provide the necessary definitions in strong lensing and establish the notation. This formalism was developed in multiple previous studies (see e.g. Schneider et al. 1992; Blandford & Narayan 1992) and has been implemented in numerous previous TDCOSMO analyses (e.g. Suyu et al. 2010; Birrer et al. 2019; Shajib et al. 2020).

The delay $\Delta t_{\rm XY}$ between arrival times of photons corresponding to images labelled as X and Y is given by

$$\Delta t_{\rm XY} = \frac{1+z_{\rm d}}{c}\frac{D_{\rm d}D_{\rm s}}{D_{\rm ds}}\left[\frac{(\boldsymbol{\theta}_{\rm X}-\boldsymbol{\beta})^2}{2} - \frac{(\boldsymbol{\theta}_{\rm Y}-\boldsymbol{\beta})^2}{2} - \psi(\boldsymbol{\theta}_{\rm X}) + \psi(\boldsymbol{\theta}_{\rm Y})\right]. \tag{1}$$

Here, $D_{\rm d}$ is the angular diameter distance to the deflector, $D_{\rm s}$ is that to the source, and $D_{\rm ds}$ is that between the deflector and the source, $z_{\rm d}$ is the deflector redshift, c is the speed of light, $\boldsymbol{\theta}$ is the image position, $\boldsymbol{\beta}$ is the un-lensed source position, and ψ is the deflection potential that is related to the deflection angle as $\nabla\psi \equiv \boldsymbol{\alpha}$ and the convergence as $\nabla^2\psi = 2\kappa$. The convergence is the surface mass density scaled by the critical density as $\kappa \equiv \Sigma/\Sigma_{\rm crit}$ with

$$\Sigma_{\rm crit} = \frac{c^2 D_{\rm s}}{4\pi G D_{\rm ds} D_{\rm d}}. \tag{2}$$

The Fermat potential ϕ is defined by combining the geometric delay term with the deflection potential as

$$\phi(\boldsymbol{\theta}) \equiv \frac{(\boldsymbol{\theta}-\boldsymbol{\beta})^2}{2} - \psi(\boldsymbol{\theta}). \tag{3}$$

The so-called time-delay distance is defined as

$$D_{\Delta t} \equiv (1+z_{\rm d})\frac{D_{\rm d}D_{\rm s}}{D_{\rm ds}}. \tag{4}$$

Each distance term contains a factor of H_0^{-1}, which cancel out such that $D_{\Delta t} \propto H_0^{-1}$. Equation (1) can be written in short form as

$$\Delta t_{\rm XY} = \frac{D_{\Delta t}}{c}\left[\phi(\boldsymbol{\theta}_{\rm X}) - \phi(\boldsymbol{\theta}_{\rm Y})\right] \equiv \frac{D_{\Delta t}}{c}\Delta\phi_{\rm XY}. \tag{5}$$

2.2. Mass-sheet degeneracy

The imaging observables of the lensing phenomenon – the image positions and the flux ratios – remain invariant under the transformation

$$\begin{aligned}&\kappa(\boldsymbol{\theta}) \to \kappa_\lambda(\boldsymbol{\theta}) = \lambda\kappa(\boldsymbol{\theta}) + 1 - \lambda,\\ &\boldsymbol{\beta} \to \boldsymbol{\beta}' = \lambda\boldsymbol{\beta},\end{aligned} \tag{6}$$

which is referred to as the MST (Falco et al. 1985). The invariance of the observables under this transformation gives rise to the MSD. We note that the magnifications are not invariant under the MST (although magnification ratios are), and thus strongly lensed standard candles can break the MSD (Bertin & Lombardi 2006).

We can separate all of the mass contributing to lensing of the background source into two components as

$$\kappa_{\rm true} = \kappa_{\rm cen} + \kappa_{\rm ext}, \tag{7}$$

where $\kappa_{\rm cen}$ is the convergence from the central deflector and $\kappa_{\rm ext}$ is the convergence from all the LOS mass distribution – except the central deflector – projected onto the plane of the central deflector (i.e. the image plane). In some cases, the central deflector may have nearby companions or satellites, or nearby LOS perturbing galaxies that are explicitly accounted for in the lens model, for example RX J1131−1231, HE 0435−1223, and ES J0408−5354 (Suyu et al. 2013; Wong et al. 2017; Shajib et al. 2020). We consider these additional mass components to be included in $\kappa_{\rm cen}$. As the mass distribution of the central deflector goes to zero at very large radius, we have

$$\lim_{\theta\to\infty} \kappa_{\rm true}(\theta) = \kappa_{\rm ext}. \tag{8}$$

Therefore, $\kappa_{\rm ext}$ can be interpreted as lensing mass in the 3D space far from or 'external' to the central deflector. Let $\kappa'_{\rm model}$ be the model convergence that can reproduce the imaging observables. However, due to the MSD, $\kappa'_{\rm model}$ is not a unique solution and we cannot ascertain that $\kappa_{\rm true} = \kappa'_{\rm model}$. If we impose the condition $\lim_{\theta\to\infty}\kappa'_{\rm model} = 0$, then $\kappa'_{\rm model}$ is a mass-sheet transform of $\kappa_{\rm true}$ with the rescaling factor $\lambda = 1/(1-\kappa_{\rm ext})$ as

$$\kappa_{\rm true} \to \kappa'_{\rm model} = \frac{1}{1-\kappa_{\rm ext}}(\kappa_{\rm cen}+\kappa_{\rm ext}) - \frac{\kappa_{\rm ext}}{1-\kappa_{\rm ext}} = \frac{\kappa_{\rm cen}}{1-\kappa_{\rm ext}}. \tag{9}$$

If the external convergence $\kappa_{\rm ext}$ can be independently estimated by studying the lens environment, then the true lensing convergence $\kappa_{\rm true}$ can be recovered from $\kappa'_{\rm model}$ through the corresponding inverse MST. However, the lens model $\kappa_{\rm model}$ that we actually constrain can be an internal MST of $\kappa'_{\rm model}$ as

$$\kappa'_{\rm model} = \lambda_{\rm int}\kappa_{\rm model} + 1 - \lambda_{\rm int}. \tag{10}$$

Interestingly, both $\kappa_{\rm model}$ and $\kappa'_{\rm model}$ can go to zero at $\theta\to\infty$ by construction. In such a case, $\lambda_{\rm int}$ is not a constant and it satisfies $\lim_{\theta\to\infty} = 1$ (Schneider & Sluse 2014). We can combine Eqs. (8)–(10) to write the relation between the true mass distribution $\kappa_{\rm true}$ and the modelled mass distribution $\kappa_{\rm model}$ as

$$\kappa_{\rm true} = (1-\kappa_{\rm ext})\left[\lambda_{\rm int}\kappa_{\rm model} + 1 - \lambda_{\rm int}\right] + \kappa_{\rm ext}. \tag{11}$$

To constrain $\lambda_{\rm int}$, we require observables that rescale with the MST, for example the stellar kinematics. Although such observables rescale with $\lambda_{\rm int}(1-\kappa_{\rm ext})$, the external convergence $\kappa_{\rm ext}$ is independently estimated from the LOS properties leaving only $\lambda_{\rm int}$ to be constrained from those observables. The LOS velocity dispersion rescales with the MST as

$$\sigma_{\rm los} \to \sigma'_{\rm los} = \sqrt{\lambda}\sigma_{\rm los}. \tag{12}$$

This rescaling is only valid for a pure MST, such as the external MST, and is approximately valid for an internal MST with single aperture kinematics. However, this is not valid for internal MST with spatially resolved kinematics (Chen et al. 2021; Yıldırım et al. 2021). The time delay rescales with the MST as

$$\Delta t \to \Delta t' = \lambda\Delta t. \tag{13}$$

As a result, we need to correct the time delays $\Delta t_{\rm model}$ predicted by the model $\kappa_{\rm model}$ as

$$\Delta t_{\rm true} = (1-\kappa_{\rm ext})\lambda_{\rm int}\Delta t_{\rm model}. \tag{14}$$

In the next section, we describe our framework for the kinematics analysis.

2.3. Kinematics analysis

The stellar velocity dispersion probes the 3D mass distribution of the deflector galaxy that is deprojected from $\kappa_{\rm cen}$. We adopt the spherical Jeans equation that connects the velocity dispersion with the gravitational potential $\Phi(r)$ as

$$\frac{{\rm d}\left(l(r)\,\sigma_{\rm r}(r)^2\right)}{{\rm d}r} + \frac{2\beta_{\rm ani}(r)\,l(r)\,\sigma_{\rm r}(r)^2}{r} = -l(r)\,\frac{{\rm d}\Phi(r)}{{\rm d}r}. \tag{15}$$

Here, $l(r)$ is the 3D luminosity density, $\sigma_{\rm r}(r)$ is the radial velocity dispersion, and $\beta_{\rm ani}(r)$ is the anisotropy parameter that relates $\sigma_{\rm r}$ to the tangential velocity dispersion $\sigma_{\rm t}$ as

$$\beta_{\rm ani}(r) \equiv 1 - \frac{\sigma_{\rm t}^2(r)}{\sigma_{\rm r}^2(r)}. \tag{16}$$

The observable quantity is the luminosity-weighted LOS velocity dispersion, which we can obtain by solving the Jeans equation as

$$\sigma_{\rm los}^2(R) = \frac{2G}{I(R)}\int_R^\infty \mathcal{K}_\beta\left(\frac{r}{R}\right)\frac{l(r)\,M(r)}{r}\,{\rm d}r, \tag{17}$$

where G is the gravitational constant, $I(R)$ is the surface brightness, and $M(r)$ is the 3D enclosed mass within radius r (Eqs. (A.15) and (A.16) of Mamon & Łokas 2005). The function $\mathcal{K}_\beta(r/R)$ depends on the parameterization of $\beta_{\rm ani}(r)$. We adopt the Osipkov–Merritt parameterization given by

$$\beta_{\rm ani}(r) = \frac{r^2}{r^2 + r_{\rm ani}^2}, \tag{18}$$

where $r_{\rm ani}$ is a scaling radius (Osipkov 1979; Merritt 1985a,b). For this parameterization, the form of $\mathcal{K}_\beta(r/R)$ is given by

$$\begin{aligned}\mathcal{K}_\beta\left(u \equiv \frac{r}{R}\right) = &\frac{u_{\rm ani}^2 + 1/2}{(u_{\rm ani}+1)^{3/2}}\left(\frac{u^2 + u_{\rm ani}^2}{u}\right)\tan^{-1}\left(\sqrt{\frac{u^2-1}{u_{\rm ani}^2+1}}\right)\\ &- \frac{1/2}{u_{\rm ani}^2+1}\sqrt{1-\frac{1}{u^2}},\end{aligned} \tag{19}$$

with $u_{\rm ani} \equiv r_{\rm ani}/R$ (Mamon & Łokas 2005). The observed aperture-averaged velocity dispersion is

$$\sigma_{\rm ap}^2 = \frac{\int_{\rm ap}\left[I(R)\sigma_{\rm los}^2(R)\right] * \mathcal{S}\,{\rm d}x{\rm d}y}{\int_{\rm ap} I(R) * \mathcal{S}\,{\rm d}x{\rm d}y}, \tag{20}$$

where $\int_{\rm ap}$ denotes integration over the aperture and $*\mathcal{S}$ denotes convolution with the seeing. Thus, the lens-model-predicted LOS velocity dispersion can be written in the form

$$\sigma^2_{\rm ap,\,model} = \frac{D_{\rm s}}{D_{\rm ds}} c^2 J(\xi_{\rm lens}, \xi_{\rm light}, \beta_{\rm ani}), \tag{21}$$

where $\xi_{\rm lens}$ is the set of mass model parameters and $\xi_{\rm light}$ is the set of light distribution parameters. The internal and external MST parameters modify the lens-model-predicted velocity dispersion as

$$\sigma^2_{\rm ap,\,true} = (1 - \kappa_{\rm ext})\lambda_{\rm int}\, \sigma^2_{\rm ap,\,model}. \tag{22}$$

The dependence of $\sigma_{\rm ap}$ on the cosmology is fully captured in the $D_{\rm s}/D_{\rm ds}$ term. The function J is independent of cosmology as all of its arguments are expressed in angular units, but it should be noted that J is directly connected to the model convergence $\kappa_{\rm model}$ through the parameters $\xi_{\rm lens}$ (Birrer et al. 2016).

2.4. Bayesian inference

We denote the set of all the observables as $O \equiv \{O_{\rm img}, O_{\rm kin}\}$, where $O_{\rm img}$ is the imaging data of the lens system and $O_{\rm kin}$ is the measured stellar velocity dispersion. Although data from spectroscopic and photometric surveys of the lens environment are necessary to estimate the external convergence, we fold in the estimated external convergence as the prior $p(\kappa_{\rm ext})$ in our inference. To predict the time delay for a given cosmology, we want to infer the Fermat potential difference $\Delta\phi$ between the corresponding image pairs. The Fermat potential difference $\Delta\phi(\xi, \kappa_{\rm ext}, \lambda_{\rm int})$ is a function of the set of model parameter $\xi \equiv \{\xi_{\rm lens}, \xi_{\rm light}, r_{\rm ani}\}$ in a model family M, external convergence $\kappa_{\rm ext}$, and internal MST parameter $\lambda_{\rm int}$. Thus, to obtain $p(\Delta\phi \mid O)$, we first aim to infer $p(\xi, \kappa_{\rm ext}, \lambda_{\rm int} \mid O)$. Applying Bayes' theorem, we can write

$$\begin{aligned} p(\xi, \kappa_{\rm ext}, \lambda_{\rm int} \mid O) &\propto p(O \mid \xi, \kappa_{\rm ext}, \lambda_{\rm int})\ p(\xi, \kappa_{\rm ext}, \lambda_{\rm int}) \\ &= p(O \mid \xi, \kappa_{\rm ext}, \lambda_{\rm int})\ p(\xi, \kappa_{\rm ext})\ p(\lambda_{\rm int}) \\ &= \int p(O \mid \xi, M, S, D_{\rm s/ds}, \kappa_{\rm ext}, \lambda_{\rm int})\ p(\xi, \kappa_{\rm ext} \mid M, S) \\ &\times p(\lambda_{\rm int})\ {\rm d}S\ {\rm d}D_{\rm s/ds}\ {\rm d}M. \end{aligned} \tag{23}$$

Here, S is the set of lens model hyper-parameters that is only relevant for $O_{\rm img}$, and $D_{\rm s/ds}$ is a short notation for the distance ratio $D_{\rm s/ds} \equiv D_{\rm s}/D_{\rm ds}$. We explicitly separate the hyper-parameters S – that need to be fixed during optimizing a lens model, for example the set of pixels for computing the image likelihood, resolution of the source reconstruction – from the choice of lens model family M. The prior $p(\kappa_{\rm ext} \mid M)$ depends on the model family M, since the model-constrained shear is used to estimate $\kappa_{\rm ext}$ corresponding to M. Since $O_{\rm img}$ and $O_{\rm kin}$ are independent data, the likelihood term $p(O \mid \xi, M, S, D_{\rm s/ds}, \kappa_{\rm ext})$ can be decomposed as

$$\begin{aligned} p(O \mid \xi, M, S, D_{\rm s/ds}, \kappa_{\rm ext,\lambda_{int}}) &= p(O_{\rm img} \mid \xi, M, S) \\ \times p(O_{\rm kin} \mid \xi, M, D_{\rm s/ds}, \kappa_{\rm ext}, \lambda_{\rm int}). \end{aligned} \tag{24}$$

Then, we can first perform the following sub-integral within the right-hand side of Eq. (23):

$$\begin{aligned} &\int p(O_{\rm img} \mid \xi, M, S)\ p(\xi \mid M, S)\ p(S)\ {\rm d}S \\ &= \int p(\xi \mid O_{\rm img}, M, S)\ p(O_{\rm img} \mid M, S)\ p(S)\ {\rm d}S. \end{aligned} \tag{25}$$

Here, $p(O_{\rm img} \mid M, S)$ is the model evidence. We perform this integral in the form of the right-hand side of Eq. (25) for numerical convenience, as it allows us to first obtain the posterior $p(\xi \mid O_{\rm img}, M, S)$ using Monte Carlo sampling, and then combine the posteriors weighted by the model evidence to perform the integration in Eq. (25). We use the Bayesian information criterion (BIC) as a proxy for the model evidence in our analysis (Schwarz 1978). The BIC is defined as

$$\mathrm{BIC} = k \ln N_{\rm data} - 2 \ln \hat{\mathcal{L}}, \tag{26}$$

where k is the number of free parameters, $N_{\rm data}$ is the number of data points, and $\hat{\mathcal{L}}$ is the maximum of the likelihood function $\mathcal{L}$. Both the BIC and directly computed model evidence were used in previous analyses for Bayesian model averaging (BMA; e.g. Madigan & Raftery 1994; Hoeting et al. 1999) in the context of lens modelling for cosmographic analysis (BIC: Birrer et al. 2019; Chen et al. 2019; Rusu et al. 2020; model evidence: Shajib et al. 2020).

Specific implementations of the Bayesian inference framework presented in this section through sampling by each team are described in Sects. 5 and 6.

3. Imaging data and ancillary measurements

The system WGD 2038−4008 was discovered from a combined search in the Wide-field Infrared Survey Explorer and *Gaia* data over the Dark Energy Survey (DES) footprint (Agnello et al. 2018). The deflector redshift is $z_{\rm d} = 0.230 \pm 0.002$ and the source redshift is $z_{\rm s} = 0.777 \pm 0.001$ (Agnello et al. 2018). In this section we describe the imaging data and spectroscopic measurements used in our analysis.

3.1. HST imaging

We obtained *Hubble* Space Telescope (HST) imaging of the system (GO-15320, PI: Treu; Shajib et al. 2019) using the Wide-Field Camera 3 (WFC3). The imaging was taken in three filters: $F160W$ in the infrared (IR) channel, and $F814W$ and $F475X$ in the ultraviolet-visual (UVIS) channel. Four exposures were taken in each filter to cover the large dynamic range in surface brightness of the brighter quasar images and the fainter extended host galaxy. For the IR band, we adopted a four-point dither pattern and STEP100 readout sequence for the MULTIACCUM mode. The total exposure times are 2196.9 s, 1428.0 s, and 1158.0 s, respectively, in the three filters. We show a false-colour red-green-blue (RGB) image of the system created from the HST imaging in Fig. 1.

The point spread function (PSF) corresponding to each filter is estimated from stacking 4–6 stars that are within each corresponding HST image. These PSFs are only used as an initial estimate by both teams and they are refined to more accurately match the PSF at the quasar image positions by iterative reconstruction during the lens model optimization (see Sects. 5 and 6 for more details on the iterative reconstruction).

3.2. Stellar velocity dispersion

Buckley-Geer et al. (2020) measure the stellar velocity dispersion of the deflector from spectroscopic observation using the Gemini Multi-Object Spectrograph (GMOS-S) on the Gemini South Telescope. The measured velocity dispersion is $\sigma_{\rm los} = 296 \pm 19$ km s^{-1} from a $0\farcs75 \times 1''$ rectangular aperture, which is in agreement with a more recent measurement from the X-shooter

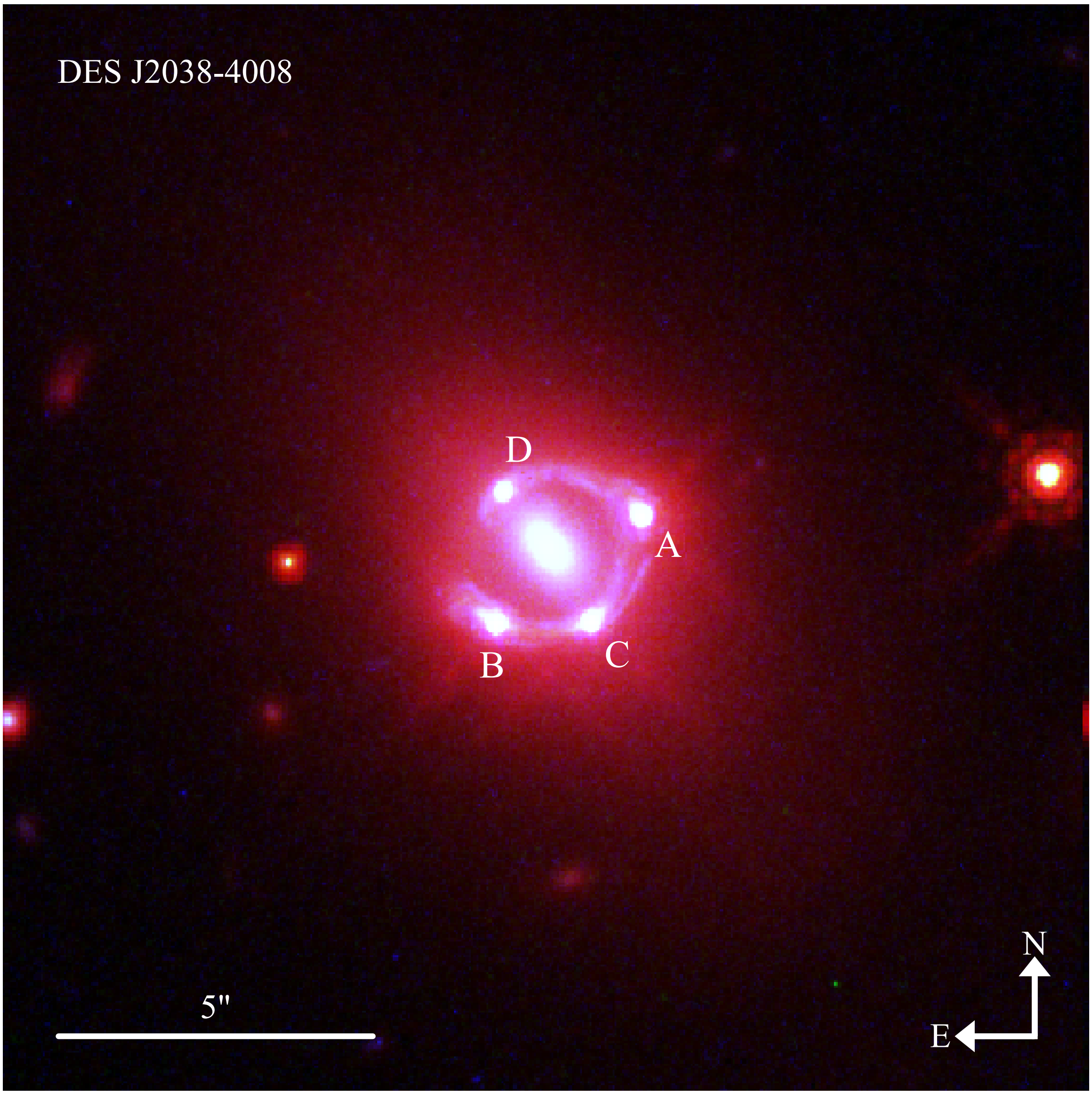

Fig. 1. False-colour image of the lens systems WGD 2038−4008. This RGB image is created from the $F160W$ (red), $F814W$ (green), and $F475X$ (blue) filters of the HST WFC3. We adjusted the relative amplitudes between the three filters to achieve a higher contrast for better visualization. The four lensed quasar images are marked as A, B, C, and D.

instrument on the Very Large Telescope (VLT; Melo et al. 2021). We used the measurement from Buckley-Geer et al. (2020) instead of the more precise measurement from Melo et al. (2021) because the latter was published after the un-blinding, when the lens models were frozen and utilized the previous measurement. The seeing full width at half maximum (FWHM) is 0$''$.9, and the exponent parameter of the Moffat PSF is $\beta = 1.74$.

3.3. LOS environment

The LOS environment of the system WGD 2038−4008 was studied by Buckley-Geer et al. (2020). These authors estimated the external convergence based on the weighted galaxy number counts approach (Greene et al. 2013; Rusu et al. 2017, 2020; Birrer et al. 2019). The weighted number counts were obtained in two separate apertures with radii 45$''$ and 120$''$ centred on the lens system from the DES multi-band imaging. The magnitude limit of counted galaxies is $I = 22.5$ mag. The counts are weighted based on simple physical quantities, such as the inverse of the distance to the lens. The spectroscopic redshifts were obtained from Gemini South GMOS-S and the photometric redshifts are based on DES multi-band photometry. Analogous number counts are also obtained within a large number of different apertures with the same sizes along random LOSs in the DES footprint. By comparing the weighted number counts for the LOS around WGD 2038−4008 with those for random LOSs, the over- or under-density is estimated in terms of a weighted number count ratio. The external convergence is then estimated by comparing the weighted number count ratio with that from statistically similar LOSs from the Millennium simulation with computed external convergence (Springel et al. 2005; Hilbert et al. 2009). If no external shear is considered, then the system WGD 2038−4008 was found to be along a LOS with approximately no overdensity within ~1% uncertainty. We provide the $\kappa_{\rm ext}$ re-weighted based on the best-fit external shear magnitudes from our lens models in Sects. 5 and 6.

Buckley-Geer et al. (2020) also find that no nearby LOS perturbers are significant enough that they need to be included explicitly in the lens mass modelling.

4. Setup of baseline models

In this section we describe the baseline models that were initially agreed upon by the two teams before performing separate and independent lens modelling. In our baseline models, we use two families of mass models for the central deflector: (i) a power-law profile, and (ii) a composite profile with an elliptical NFW potential for the dark component and a superposition of three Chameleon profiles (hereafter, triple Chameleon profile) in convergence for the luminous component. We also add external shear to both types of mass model. For the light profile of the central deflector, we adopt a triple Sérsic profile in all three bands in the models with the power-law mass profile. In the models with the composite mass profile, however, we adopt a triple Chameleon light profile in the $F160W$ band linked with the triple Chameleon mass profile and a triple Sérsic profile in the UVIS bands.

We adopted three Chameleon profiles to sufficiently account for the complexity in the light profile of the deflector. Moreover, we adopted the triple Chameleon light profile only for the $F160W$ profile, since this is the only band that is connected to the luminous component of the convergence profile.

Although both teams adopted these baseline models, individual teams were allowed to make their own choices – which may not necessarily be identical – pertaining to other model specifications, for example parameter priors and fixing parameter values.

In the next subsections we provide the definitions of the mass and light profiles in the baseline models.

4.1. Mass profiles

The two baseline lens model families we adopt are the power-law mass profile and the composite mass profile.

4.1.1. Power-law mass profile

We adopted the power-law elliptical mass distribution (PEMD; Barkana 1998) defined as

$$\kappa_{\rm PL}(\theta_1, \theta_2) \equiv \frac{3-\gamma}{2} \left[\frac{\theta_{\rm E}}{\sqrt{q_{\rm m}\theta_1^2 + \theta_2^2/q_{\rm m}}} \right]^{\gamma-1}, \tag{27}$$

where γ is the logarithmic slope, $\theta_{\rm E}$ is the Einstein radius, and $q_{\rm m}$ is the axis ratio. The coordinates (θ_1, θ_2) are in the coordinate frame that is aligned with the major and minor axes. The position angle of this frame is $\varphi_{\rm m}$ with respect to the RA–Dec frame.

4.1.2. Composite mass profile

The composite mass profile consists of two individual mass profiles for the baryonic and the dark components of the mass distribution.

For the dark matter distribution, we adopt a Navarro–Frenk–White (NFW) profile with ellipticity defined in the potential. The 3D NFW profile in the spherical case is given by

$$\rho_{\rm NFW}(r) \equiv \frac{\rho_{\rm s}}{(r/r_{\rm s})\,(1 + r/r_{\rm s})^2}, \tag{28}$$

where $\rho_{\rm s}$ is the density normalization, and $r_{\rm s}$ is the scale radius (Navarro et al. 1997). We refer to Golse & Kneib (2002) for the expressions of the lens potential and deflection angles associated with the elliptical NFW profile.

For the baryonic mass distribution, we adopt the Chameleon convergence profile. The Chameleon profile matches with the Sérsic profile within a few per cent at 0.5–3$\theta_{\rm eff}$, where $\theta_{\rm eff}$ is the half-light or effective radius of the Sérsic profile (Dutton et al. 2011). The Chameleon profile is defined as the difference between two non-singular isothermal ellipsoids:

$$\kappa_{\rm Chm}(\theta_1, \theta_2) \equiv \frac{a_0}{1+q_{\rm m}} \left[\frac{1}{\sqrt{\theta_1^2 + \theta_2^2/q_{\rm m}^2 + 4w_{\rm c}^2/(1+q_{\rm m}^2)}} - \frac{1}{\sqrt{\theta_1^2 + \theta_2^2/q_{\rm m}^2 + 4w_{\rm t}^2/(1+q_{\rm m}^2)}} \right], \tag{29}$$

where a_0 is the normalization and $w_{\rm c}$ and $w_{\rm t}$ are the core sizes for the individual non-singular isothermal components in the Chameleon profile (Dutton et al. 2011; Suyu et al. 2014). This profile is numerically convenient for computing lensing quantities using closed-form expressions unlike the Sérsic profile.

4.2. Light profiles of the deflector

4.2.1. Sérsic profile

The Sérsic profile is defined as

$$I_{\rm Sersic}(\theta_1, \theta_2) \equiv I_{\rm eff} \exp\left[-b_n \left\{ \left(\frac{\sqrt{\theta_1^2 + \theta_2^2/q_{\rm L}^2}}{\theta_{\rm eff}/\sqrt{q_{\rm L}}} \right)^{1/n_{\rm s}} - 1 \right\} \right], \tag{30}$$

where $I_{\rm eff}$ is the amplitude, $\theta_{\rm eff}$ is the effective radius along the intermediate axis, and $n_{\rm s}$ is the Sérsic index (Sérsic 1968). The factor b_n is a normalizing factor so that $\theta_{\rm eff}$ is the half-light radius.

4.2.2. Chameleon light profile

In the composite baseline model, we use the same Chameleon profile from Eq. (29) for the light profile of the deflector, but replacing the convergence amplitude κ_0 with the flux amplitude I_0.

5. GLEE modelling

In this section we describe the GLEE modelling procedure. GLEE is a software package developed by S. H. Suyu and A. Halkola (Suyu & Halkola 2010; Suyu 2012). The lensing mass distribution is described by a parameterized profile. The lensed quasar images are modelled as point sources on the image plane convolved with the PSF. The extended host galaxy of the lensed quasar is modelled on a 50×50 pixel grid with curvature regularization (Suyu et al. 2006), spanning the range of source coordinates corresponding to the pixels within a region containing the lensed arcs (hereafter, referred to as the ‘arcmask’). The quasar image amplitudes are independent of the extended host galaxy light distribution to allow for deviations due to microlensing, time delays, and substructure. The lens galaxy light distribution is represented as the sum of three Sérsic (or three Chameleon) profiles with a common centroid.

The lens model is constrained by the positions of the lensed quasar images and the surface brightness of the pixels of the lensed Einstein ring of the quasar host galaxy in the three HST bands that are fit simultaneously. The quasar positions are fixed to the positions of the point sources on the image plane (after they have stabilized) and are given a fixed Gaussian uncertainty of width $0\farcs004$ to account for offsets due to substructure in the lens or LOS. This uncertainty is small enough to satisfy astrometric requirements for cosmography (Birrer & Treu 2019). The quasar flux ratios are not used as constraints, as they can be affected by microlensing, which has been detected in this system (Melo et al. 2021). We use the initial PSF estimate in each band that was created from ~4–6 bright stars within the HST image (Sect. 3.1). We first model the lens separately in each band to iteratively update the respective PSFs using the lensed active galactic nucleus (AGN) images (Chen et al. 2016; Wong et al. 2017; Rusu et al. 2020). We then keep the 'corrected' PSFs fixed and use them in our final models that simultaneously use the surface brightness distribution in all three bands as constraints. We use the positions of the quasar images to align the cutouts in the three HST bands. We do not enforce any similarity of pixel values at the same spatial position across different bands (i.e. the model flux at any position in one band is independent of the model flux in other bands). In our Markov chain Monte Carlo (MCMC) sampling, we vary the light parameters of the lens galaxy and quasar images, the mass parameters of the lens galaxy, and the external shear. The source position is also sampled in the modelling. The quasar image positions are linked across all bands, but the other light parameters are allowed to vary independently.

We create cutouts of the HST images with dimensions of $5\farcs6 \times 5\farcs6$, which corresponds to a 140×140 pixel cutout for the UVIS/$F475X$ and UVIS/$F814W$ bands and a 70×70 pixel cutout for the IR/$F160W$ band. This conservative cutout size is chosen to include the entire region containing the lensed host galaxy arc light. We define the arcmask around the deflector galaxy in each of the three bands, which encloses the region where we reconstruct the lensed arc from the extended quasar host galaxy. The arcmask is used to calculate the likelihood involving the reconstructed lensed arc light, but the whole cutout is used for calculating the likelihood associated with the lens light. The construction of the weight images and bad pixel masking for each cutout are analogous to the procedure in Wong et al. (2017) and Rusu et al. (2020). In order to avoid biasing the modelling due to large residuals from a PSF mismatch near the AGN image positions, we rescale the weights in those regions by a power-law model such that a pixel originally given an estimated 1σ noise value of $\sigma_{\mathrm{img},i}$ is rescaled to a noise value of $A \times \sigma^{b}_{\mathrm{img},i}$. The constants A and b are chosen for each band such that the normalized residuals (the residual flux of each pixel normalized by its 1σ uncertainty) in the AGN image regions are approximately consistent with the normalized residuals in the rest of the arc region. We do not rescale the weights outside of the AGN image regions. The arcmask region and the regions around the AGN with rescaled weights are shown in the first column of Fig. 2.

5.1. Power-law model

Our fiducial power-law mass model uses the triple Sérsic parameterization for the lens galaxy light and has the additional free parameters: (i) position (θ_1, θ_2) of the mass centroid (allowed to vary independently from the centroid of the light distribution), (ii) Einstein radius θ_{E}, (iii) minor-to-major axial ratio, q_{m}, and associated position angle φ_{m} (measured east of north), (iv) 3D slope of the power-law mass distribution γ, and (v) external shear γ_{ext} and associated position angle φ_{ext} (measured east of north).

We assume uniform priors on the model parameters over a wide physical range. Figure 2 shows the data and the lens model results in all three bands for our fiducial power-law model, as well as the reconstructed sources. Our model simultaneously reproduces the surface brightness structure of the lensed AGN and host galaxy in all bands. The normalized residual in the third column shows the area within the arcmask, as well as the region interior to the arcmask. In the IR/$F160W$ band, there is an excess residual at the inner boundary of the arcmask (as well as outside of the arcmask, not shown in this figure) arising from the technical details of the PSF not being corrected outside of the arcmask. We run a test where the pixels showing excess residual outside of the arcmask are downweighted and find no significant change in the model parameters.

5.2. Composite model

Our composite model consists of a baryonic component linked to the light profile of the lens galaxy, plus a dark matter component. The composite model assumes the triple Chameleon light profile for the lens galaxy in the IR/$F160W$ band scaled by an overall mass-to-light (M/L) ratio. The Chameleon light profiles link to parameters describing the light distribution to those of the mass distribution in a straightforward way, as they are fundamentally just a combination of isothermal profiles. We keep the triple Sérsic model for the lens galaxy light in the UVIS bands to maintain consistent parameterization with the power-law models. The dark matter component is modelled as an elliptical NFW (Navarro et al. 1996) halo with the centroid linked to the light centroid in the $F160W$ band.

The fiducial composite model has the following free parameters in addition to the lens light parameters: (i) mass-to-light ratio (M/L) for the baryonic component, (ii) NFW halo scale radius, r_{s}, (iii) NFW halo normalization, $\kappa_{0,\mathrm{h}}$ (defined as $\kappa_{0,\mathrm{h}} \equiv 4\kappa_{\mathrm{s}} \equiv 4\rho_{\mathrm{s}} r_{\mathrm{s}} \Sigma^{-1}_{\mathrm{crit}}$; Golse & Kneib 2002), (iv) NFW halo minor-to-major axial ratio, q_{NFW}, and associated position angle, φ_{NFW}, and (v) external shear, γ_{ext}, and associated position angle, φ_{ext}.

A Gaussian prior for the M/L of the baryonic component is employed, using the stellar mass constraint from Agnello et al. (2018) of $\log_{10}(M_\star/M_\odot) = 11.40^{+0.01}_{-0.08}$ for a Salpeter initial mass function (IMF). Although this value is lower than our estimate derived from the photometry of our models of the lens light profile (see Sect. 6), this prior has little influence on the result, as the model prefers an almost maximal M/L with little dark matter contribution (see Sect. 5.6). We set a Gaussian prior of $r_{\mathrm{s}} = 22\farcs6 \pm 3\farcs1$ based on the results of Gavazzi et al. (2007) for a sample of lenses in the SLACS survey (Bolton et al. 2006). These lenses span a redshift and velocity dispersion range that includes WGD 2038−4008, with a mean virial mass of $\langle M_{\mathrm{vir}} \rangle = 1.4^{+0.6}_{-0.5} \times 10^{13}\, h^{-1}\, M_\odot$. All other parameters are given uniform priors. The relative amplitudes of the three Chameleon profiles representing the stellar light distribution of the lens galaxy can vary within an MCMC chain. However, their relative amplitudes in the mass model initialization are necessarily fixed (due to the way that the GLEE user interface is set up), even though they share the same global M/L parameter. To account for this, we iteratively run a series of MCMC chains for the fiducial composite model and update the relative amplitudes of the three mass components to match that of the light components after each chain. After several iterations, the predicted Fermat potential stabilizes, and we stop iterating. We subsequently ran a test fiducial model using an updated version of GLEE in which the amplitudes

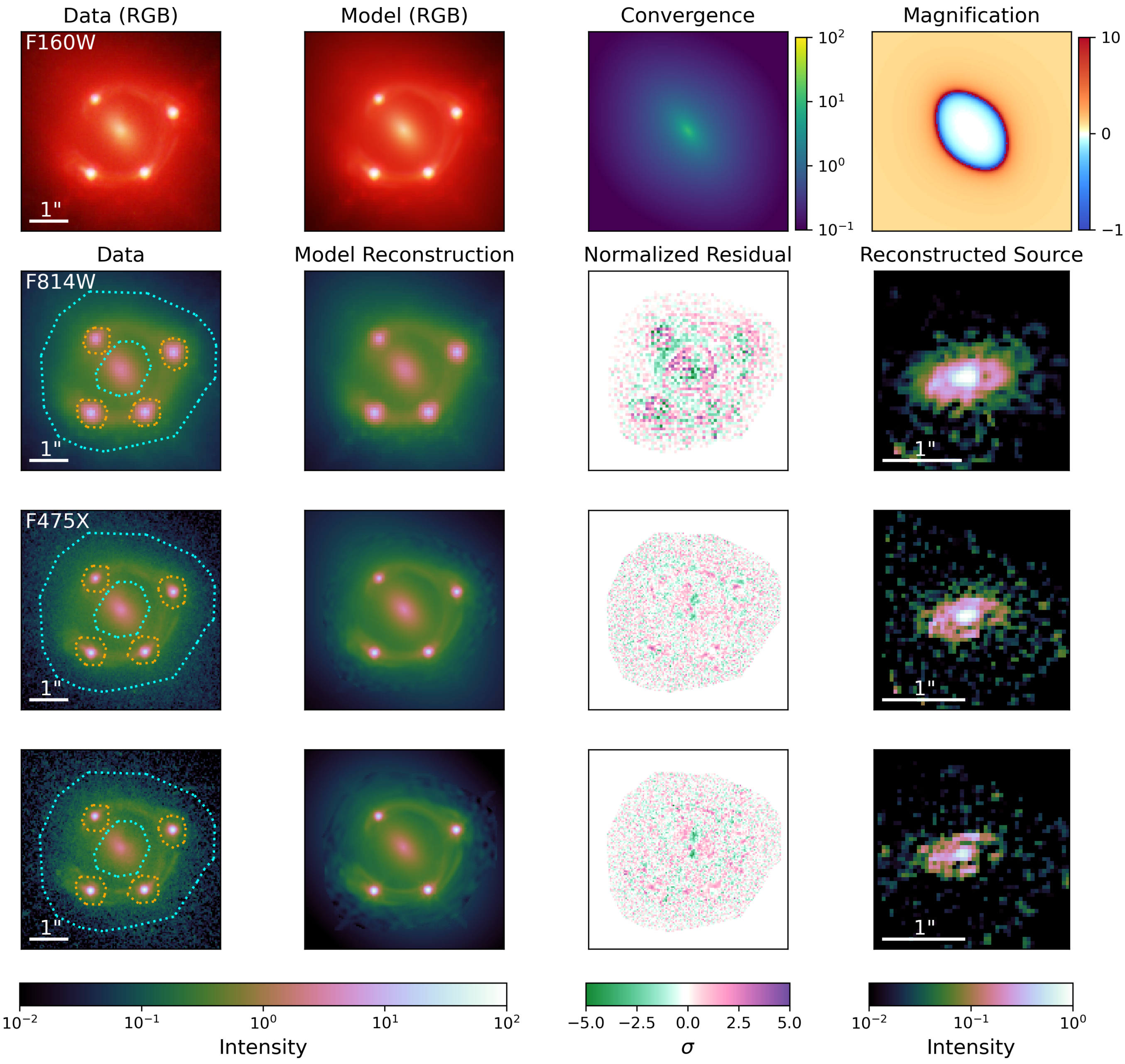

Fig. 2. Fiducial power-law model results for IR/$F160W$ (*top row*), UVIS/$F814W$ (*middle row*), and UVIS/$F475X$ (*bottom row*) from GLEE. The maximum-likelihood model in the MCMC chain is shown. Shown are the observed image (*first column*), the reconstructed image predicted by the model (*second column*), the normalized residual within and interior to the arcmask region (defined as the difference between the data and model, normalized by the estimated uncertainty of each pixel; *third column*), and the reconstructed source (*right column*). *First column*: the dotted cyan lines indicate the arcmask (donut-shaped) region used for fitting the extended source, the dotted orange lines indicate the AGN mask region where the power-law weighting is applied, and the region outside the dotted cyan arcmask is used to further constrain the foreground lens light and (partly) the AGN light (but not the AGN host galaxy light since its corresponding lensed arcs are below the noise level in this outer region). The colour bars show the scale in the respective panels. The results shown here are for the fiducial power-law model, but the results for the other systematics tests (Sect. 5.3) are qualitatively similar.

of the mass components are directly linked to the light components and found that the results were unchanged. Figure 3 shows the data and the lens model results in all three bands for our fiducial composite model, as well as the source reconstructions.

5.3. Systematics tests

In this section we describe a variety of tests of the effects of various systematics in our modelling arising from different assumptions in the way we constructed the model that might affect the posterior. In addition to the basic fiducial models described above, we perform inferences for both the power-law and composite models given the following sets of assumptions: (i) a model where the regions near the AGN images are given zero weight rather than being scaled by a power-law weighting; (ii) a model where the region near the AGN images scaled by the power-law weighting is increased by one pixel around the outer edge; (iii) a model where the reconstructed source plane

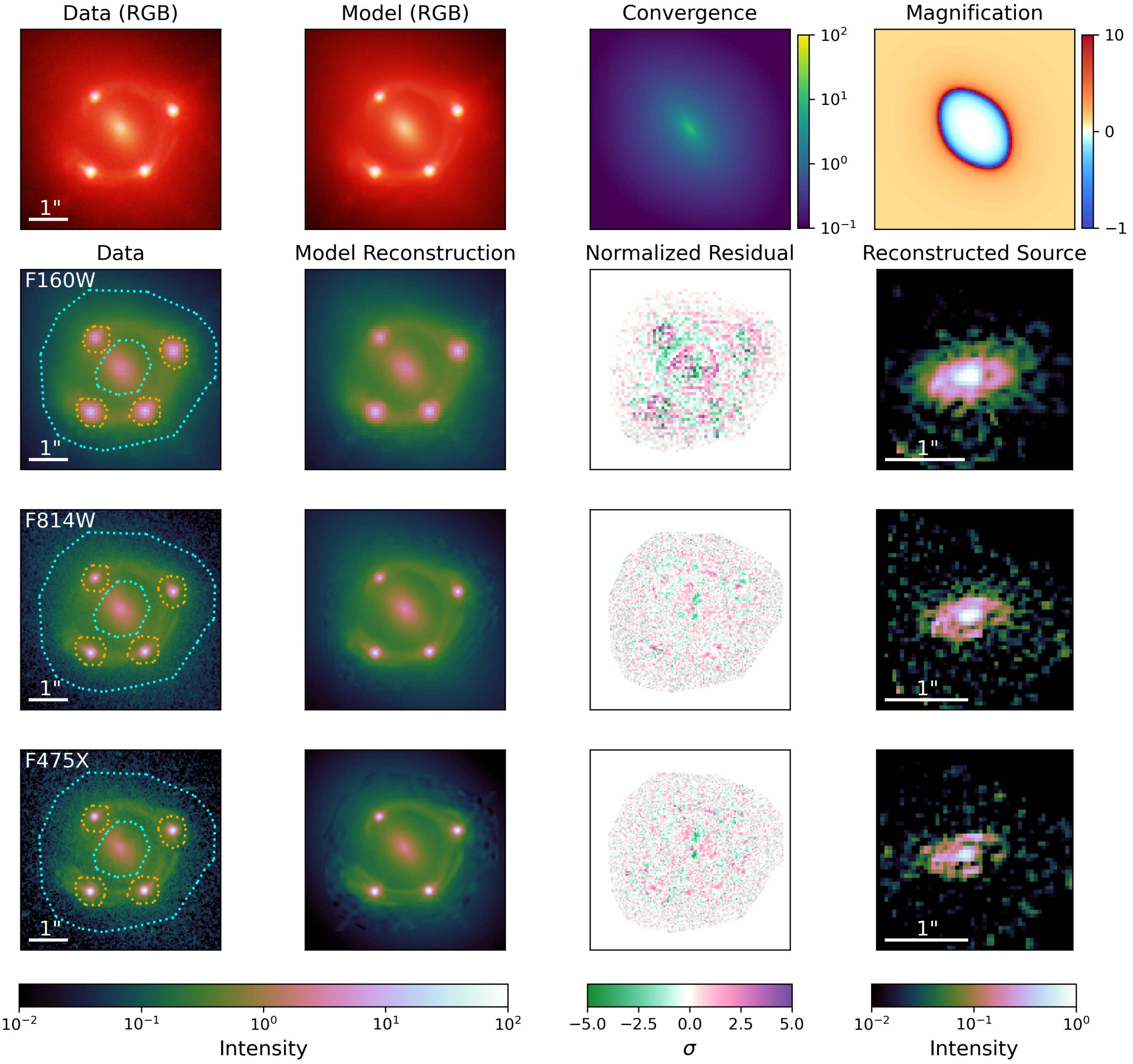

Fig. 3. Same as Fig. 2, but for the fiducial composite model from GLEE.

resolution in all bands is reduced to 40 × 40 pixels; (iv) a model where the reconstructed source plane resolution in all bands is increased to 60 × 60 pixels; and (v) a model with the arcmask region increased by one pixel on both the inner and outer edges. We combined the MCMC chains from all of these tests, weighted by the BIC (similar to Rusu et al. 2020, see Sect. 5.5).

5.4. Kinematics and external convergence

We used the kinematics and external convergence constraints from Buckley-Geer et al. (2020). We combined both LOS velocity dispersion measurements to constrain the lens models. Buckley-Geer et al. (2020) constrain the external convergence for different external shear amplitudes in steps of 0.01. For each model, we use the distribution corresponding to the external shear that is closest to the median amplitude for that model. We use importance sampling (e.g. Lewis & Bridle 2002) to simultaneously combine the velocity dispersion and external convergence distributions in a manner similar to Wong et al. (2017) and Rusu et al. (2020). For each set of lens parameters ν from our lens model chain, we draw a $\kappa_{\rm ext}$ sample from the distributions in Buckley-Geer et al. (2020) and a sample of $r_{\rm ani}$ from the uniform distribution $[0.5, 5]\theta_{\rm eff}$ ($\theta_{\rm eff}$ is calculated from the lens light distribution in the IR/$F160W$ band from the power-law model). From these together with the $D_{\rm ds}/D_{\rm d}$ ratio (that is fixed given the fixed $\Omega_{\rm m}$ value of 0.3 in flat ΛCDM), we can compute the kinematics likelihood for the joint sample $\{\nu, \Omega_{\rm m}, \kappa_{\rm ext}, r_{\rm ani}\}$ via Eq. (22) and use this to weight the joint sample. We can then combine the Fermat potential computed from our lens model parameters ν with values of $\kappa_{\rm ext}$ and $D_{\Delta t}$

to predict the time delays as a function of H_0 (via Eqs. (5) and (14)).

5.5. BIC weighting

We weight our models using the BIC, defined in Eq. (26). We take $N_{\rm data}$ (the number of data points) to be the number of pixels in the image region across all three bands that are outside the fiducial AGN mask (so that we are comparing equal areas), plus eight (for the four AGN image positions), plus one (for the velocity dispersion). k (the number of free parameters) is taken to be the number of parameters in the model that are given uniform priors, plus two (for the source position), plus one (for the anisotropy radius to predict the velocity dispersion). $\hat{L}$ (the maximum likelihood of the model from the MCMC sampling) is the product of the AGN position likelihood, the pixellated image plane likelihood, and the kinematic likelihood. The image plane likelihood is the Bayesian evidence of the pixelated source intensity reconstruction using the imaging data within the arcmask (which marginalizes over the source surface brightness pixel parameters and is thus the likelihood of the lens parameters excluding the source pixel parameters; see Eqs. (12) and (13) in Suyu & Halkola 2010) multiplied by the likelihood of the lens model parameters within the image plane region that excludes the arcmask. We evaluate the BIC using the fiducial weight image and arcmask, as the majority of the models were optimized with these.

We estimate the variance in the BIC, $\sigma^2_{\rm BIC}$, by sampling the fiducial model with source resolutions of [47, 48, 49, 50, 51, 52, 53, 54, 56, 58, 60] pixels on a side (the 50 × 50 pixel case is just the original fiducial model), keeping the arcmask the same. Changing the source resolution in this way shifts the predicted time delays stochastically, but there is no overall trend with resolution, and the degree of the shifts are smaller than the scatter among the different models in the systematics tests we run. We calculate the BIC for each of these models with different source resolutions and take the variance of this set of models as $\sigma^2_{\rm BIC}$. We find $\sigma^2_{\rm BIC} \sim 36$ for the power-law models and $\sigma^2_{\rm BIC} \sim 34$ for the composite models.

To avoid biases due to our choice of lens model parameterization, we split the samples into the power-law and composite models and calculate the relative BIC and weighting for each set separately, similar to Birrer et al. (2019) and Rusu et al. (2020). Specifically, we weight a model with a given BIC of value x by a function $f_{\rm BIC}(x)$, defined as the convolution

$$f_{\rm BIC}(x) = h(x, \sigma_{\rm BIC}) * \exp\left(-\frac{x - {\rm BIC}_{\rm min}}{2}\right), \tag{31}$$

where ${\rm BIC}_{\rm min}$ is the smallest BIC value within a set of models (power-law or composite), and h is a Gaussian centred on x with a variance of $\sigma^2_{\rm BIC}$. The exponential term is a proxy to the evidence ratio. We follow the calculation of Yıldırım et al. (2020) in evaluating the convolution integral in Eq. (31). Once we weighted time delay distributions for the power-law and composite models, we combined these two with equal weight in the final inference.

5.6. Modelling results with $\lambda_{\rm int=1}$

The marginalized parameter distributions of the power-law model are shown in Fig. 4. We show the combined distributions of all power-law models where each model is given equal weight, as well as the BIC-weighted distribution. Figure 5 shows the similar parameter distribution for the composite models. The point estimates for the mass model parameters from the GLEE models are presented and compared with those from the LENSTRONOMY models later in Sect. 7.2. The reconstructed sources of each model are qualitatively very similar, which is an important consistency check of the two models.

The power-law model has a steep mass profile slope of $\gamma = 2.30 \pm 0.01$, but the parameters are consistent with the previous model of Shajib et al. (2019). The various systematics tests do not show substantial variation. The 'island'-like feature in Fig. 4 comes from the model with a lower source plane resolution, but this model is downweighted by the BIC, so it does not affect our result. The centroid of the mass and light profiles are consistent to within $\sim$0$''$.003, and the model is able to fit the quasar positions to an rms of $\sim$0$''$.005.

The composite model fits the quasar positions to an rms of $\sim$0$''$.01, slightly worse than the power-law model. We note that the dark matter component contributes a very small fraction of the mass (of order $\sim$1%) relative to the stellar component, which has a large mass-to-light ratio. While this may appear unusual, the stellar mass enclosed within the Einstein radius determined from stellar population synthesis (SPS) models fit to the imaging data assuming a Salpeter IMF is consistent with the total enclosed mass as constrained by the lensing. In Fig. 6, we show the circularly averaged convergence of both the power-law and composite models. The effective Einstein radii (at which $\langle\kappa(<r)\rangle = 1$) of the two models agree to within less than one UVIS pixel (0$''$.04), which corresponds to $\sim$2–3%. At the Einstein radius, the composite model slope closely matches the slope of the power-law model. The magnitude of the external shear ($\gamma_{\rm ext}$) required for the power-law and composite models differs, resulting in a difference in the external convergence ($\kappa_{\rm ext}$) as determined by Buckley-Geer et al. (2020).

The relative BIC weightings of each model are provided in Table 1. The blinded distributions of Fermat potential differences are plotted individually for each model in Fig. 7. The unblinded illustrations of the BIC-weighted distributions are provided later in Sect. 7.1. Notably, the power-law and composite model have predicted time delays that are offset by $\sim$13%, indicating a difference in the two models. Contributing to this difference is the larger $\kappa_{\rm ext}$ for the composite model. As a result, the combined constraint has a larger uncertainty, reflecting this difference. Without factoring in the different $\kappa_{\rm ext}$ distributions, the power-law and composite models would be offset by $\sim$8%.

6. Lenstronomy modelling

In this section we describe the LENSTRONOMY model setups and modelling results. The software package LENSTRONOMY (Birrer & Amara 2018; Birrer et al. 2021) is a publicly available lens modelling software[4]. In contrast with GLEE, the software LENSTRONOMY uses basis sets to reconstruct the flux distribution of the background source galaxy (Birrer et al. 2015). In this section we describe the specific model settings for LENSTRONOMY on top of the baseline models from Sect. 4, then present our modelling results, and lastly combine the lens models with the measured stellar kinematics and the estimated external convergence.

[4] https://github.com/lenstronomy/lenstronomy

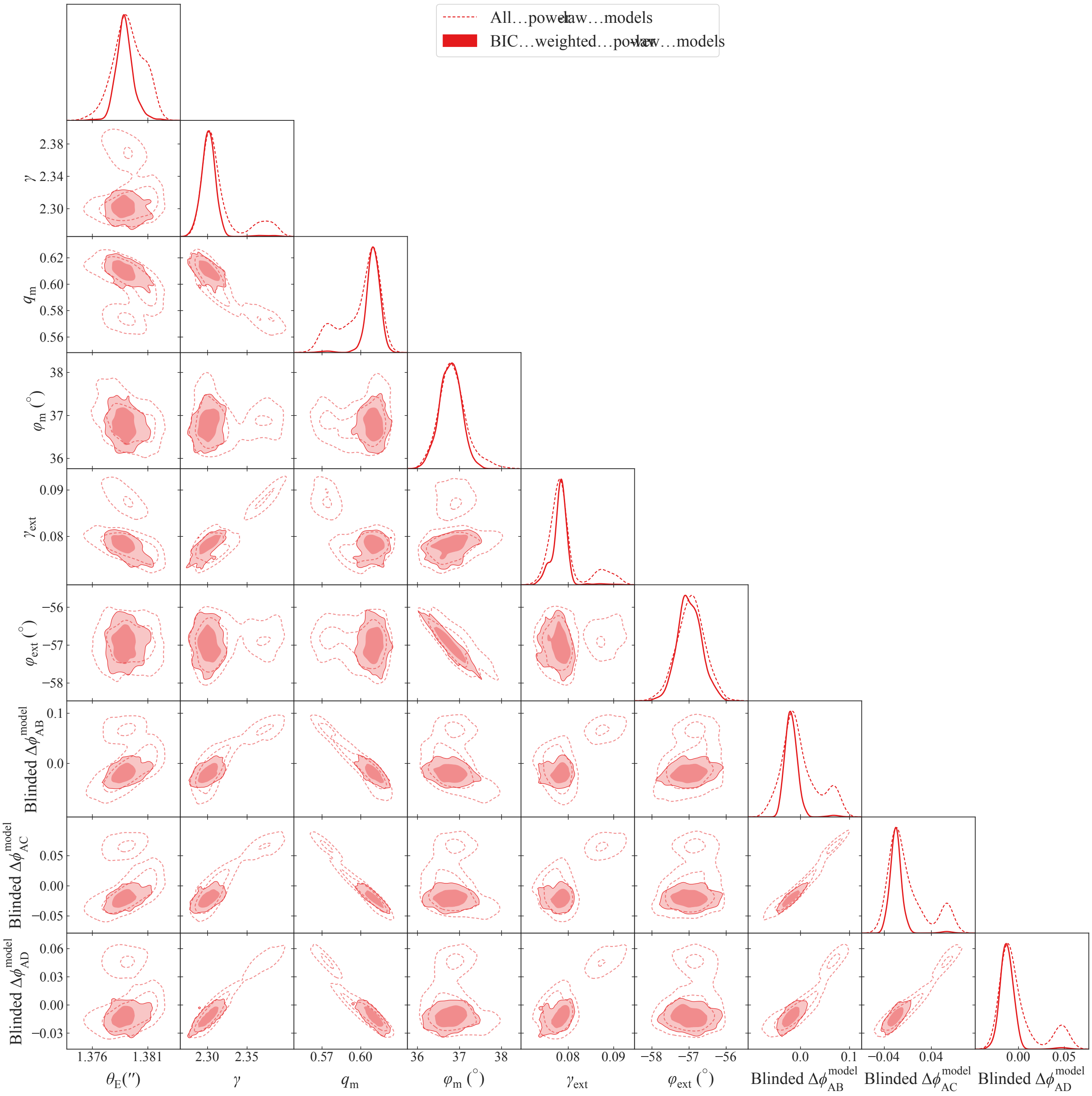

Fig. 4. Marginalized parameter distributions from our power-law lens model results from GLEE. We show the combined results from our systematics tests (dashed red contours) with each model weighted equally, as well as the BIC-weighted model results (shaded red contours). The contours represent the 68.3% and 95.4% quantiles.

6.1. LENSTRONOMY *specific model settings*

We explain particular model settings related to the mass and light profiles of the deflector galaxy in Sect. 6.1.1, the source light profiles in Sect. 6.1.2, and the image region for likelihood computation in Sect. 6.1.3. We summarize the set of all the lens models combining these different settings in Sect. 6.1.4.

6.1.1. Mass and light profiles of the deflector galaxy

We simultaneously model the HST images from all three bands. We join the centroids of the triple Sérsic profiles across the three bands in the power-law model setup, and also the centroids of the triple Chameleon profiles in the composite model setup. We join the ellipticity parameters of the light profiles only between the two UVIS bands. We let the amplitudes I_eff, effective radii θ_eff, and the Sérsic indices n_s in the three bands independently vary to allow for a colour gradient.

In the composite model setup, we adopt a Gaussian prior with mean 22$''$.6 and standard deviation 3$''$.1 for the NFW scale radius r_s based on the measurements of Gavazzi et al. (2007) for a sample of SLACS survey lens systems (Bolton et al. 2006). Since the velocity dispersion and the redshift of the central deflector of WGD 2038−4008 fall within the ranges spanned by the SLACS lenses, such a prior is appropriate (Treu et al. 2006). Similar priors were also adopted in previous H0LiCOW

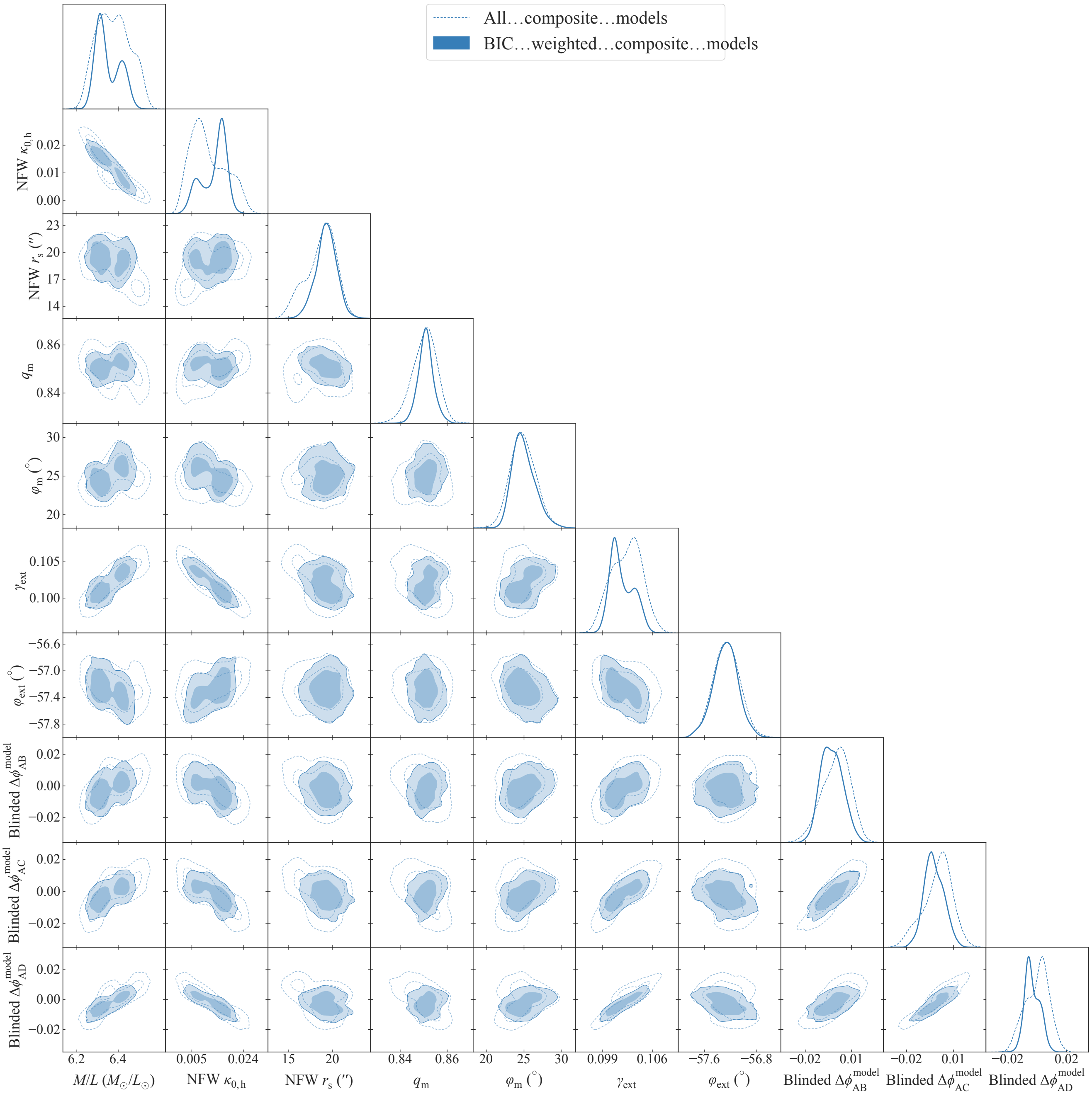

Fig. 5. Marginalized parameter distributions from our composite lens model results from GLEE. We show the BIC-weighted model (shaded blue contours) and the combined results from our systematics tests (dashed blue contours). The contours represent the 68.3% and 95.4% quantiles.

and STRIDES analyses (e.g. Wong et al. 2017; Rusu et al. 2017; Shajib et al. 2020). Although the measurement by Gavazzi et al. (2007) are reported in the physical kpc unit, we use the same fiducial cosmology as Gavazzi et al. (2007) to recover the scale in the observable angular unit. We also impose a prior on the concentration parameter using the theoretical M_{200}−c relation from Diemer & Joyce (2019) with an intrinsic scatter of 0.11 dex.

6.1.2. Source light profiles

We adopt a basis set of shapelets and one elliptical Sérsic profile to describe the flux distribution of the quasar host galaxy. The Sérsic profile describes the smooth component of the flux distribution of the host galaxy, and the shapelets account for the non-smooth features (Refregier 2003; Birrer et al. 2015). The number of shapelets $n_\mathrm{shapelets}$ depends on the maximum polynomial order n_max as $n_\mathrm{shapelets} = (n_\mathrm{max} + 1)(n_\mathrm{max} + 2)/2$, and the spatial extent of the shapelets is characterized with a scale size ς. We model the quasar images as point sources on the image plane. We treat the positions of the quasar images as free parameters throughout the model optimization and MCMC procedures. The point source positions are constrained directly through the likelihood of the pixel-level flux values in the imaging data. The four image positions give six independent relative positional parameters. We chose the option within LENSTRONOMY to solve the lens equation to constrain six parameters out of the set of the mass model parameters from these six independent relative positional

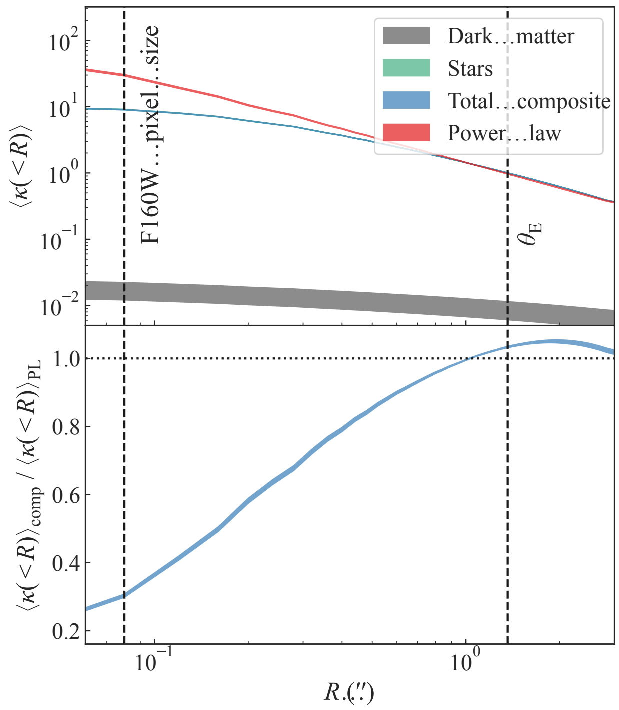

Fig. 6. Radial mass profiles of the central deflector constrained by the GLEE models. *Top*: circularly averaged convergence $\langle\kappa(<R)\rangle$ as a function of radius for the GLEE power-law model (red) and composite model (blue). The shaded regions represent the 1σ credible regions. The stellar (green) and dark matter (black) components of the composite model are plotted separately. The vertical dashed black lines mark the pixel size in the $F160W$ band and the best fit Einstein radius. *Bottom*: ratio of average convergence of the composite model to that of the power-law model as a function of radius.

Table 1. BIC weighting for different lens models from GLEE.

Model setting	ΔBIC	BIC weight
Power-law ellipsoid model		
Fiducial	0	0.661
AGN mask weight = 0	223	0.000
AGN mask + 1 pix	26	0.324
40×40 source	84	0.015
60×60 source	179	0.000
Arcmask + 1 pix	295	0.000
Composite model		
Fiducial	34	0.218
AGN mask weight = 0	252	0.000
AGN mask + 1 pix	45	0.132
40×40 source	424	0.000
60×60 source	0	0.650
Arcmask + 1 pix	137	0.000

Notes. The ΔBIC values are calculated relative to the model with the lowest BIC value.

parameters[5]. These six mass model parameters then have 'one-to-one' correspondence with the sampled quasar image positions. Therefore, they are not treated as non-linear parameters anymore in the optimization and sampling procedures. For the power-law model, the six parameters chosen are the PEMD's centroid RA and Dec, axis ratio $q_{\rm m}$, position angle $\varphi_{\rm m}$, Einstein radius $\theta_{\rm E}$, and the external shear angle $\varphi_{\rm ext}$. For the composite model, the six parameters chosen are the NFW profile's centroid RA and Dec, axis ratio $q_{\rm NFW}$, position angle $\varphi_{\rm NFW}$, density normalization $\rho_{\rm s}$, and the external shear angle $\varphi_{\rm ext}$.

We join the ellipticity parameters of the source Sérsic profiles across the three bands. The centroids of all the light profiles are also joint across the three bands. This centroid is set at the quasar position in the source plane that is constrained through solving the lens equations for the four image positions. The effective radii $\theta_{\rm eff}$, the Sérsic indices $n_{\rm s}$, the shapelet scale sizes ς for different bands are independent of each other.

We treat $n_{\rm max}$ as a hyper-parameter and fix it for a particular model optimization. A minimum number of shapelet components is necessary to describe the complex features in the lensed arcs; however, too many shapelet components will fit the noise in the imaging data. Thus, striking a balance between these two scenarios is necessary when choosing the number of shapelet components. We adopt three choices for $\{n_{\rm max}^{\rm IR}, n_{\rm max}^{\rm UVIS}\}$: $\{7, 11\}$, $\{8, 12\}$, $\{9, 13\}$.

6.1.3. HST image region for likelihood computation

We chose a circular aperture in each band encompassing the lensed arcs centred on the lens galaxy to compute the imaging likelihood. The radii of these apertures are hyper-parameters in the model. We take two sets of choices for $\{r_{\mathcal{L}}^{\rm IR},\ r_{\mathcal{L}}^{\rm UVIS}\}$: $\{2''\!.2,\ 3''\!.6\}$, $\{2''\!.3,\ 3''\!.7\}$ with $r_{\mathcal{L}}$. Some nearby objects (stars or smaller galaxies) are masked out if they fall within the likelihood computation region (see Figs. 8 or 9 for the shape and comparative size of the likelihood computation regions).

6.1.4. Model choice combinations

Summarizing the above sections, we have the hyper-parameter choices (i) for the lens galaxy mass profile: power-law and composite; (ii) for the source light $\{n_{\rm max}^{\rm IR}, n_{\rm max}^{\rm UVIS}\}$: $\{7, 11\}$, $\{8, 12\}$, and $\{9, 13\}$; and (iii) for the likelihood computation region radii $\{r_{\mathcal{L}}^{\rm IR},\ r_{\mathcal{L}}^{\rm UVIS}\}$: $\{2''\!.2,\ 3''\!.6\}$ and $\{2''\!.3,\ 3''\!.7\}$. Taking a combination of these choices, we have 12 different model setups. We perform the optimization with the same models setups twice. These twin runs are different due to stochasticity in the PSF reconstruction and MCMC sampling procedures, and help us assess random errors. As a result, we have 24 different optimized models, on which we perform BMA. The light profiles from the deflector, the lensed light profiles from the quasar host galaxy, and the point sources at the quasar image positions form a linear basis set for reconstructing the observed HST imaging. As a result, the amplitudes of these profiles are linear parameters, as they can be obtained through a linear inversion for a sampled set of non-linear parameters that describe all the mass and light profiles. There are 206−281 linear parameters and 51−54 non-linear parameters in our models.

6.2. Modelling workflow

For each model setting, we reconstruct the PSF in each HST band. The reconstruction is initiated from a PSF estimate with a corresponding error map created from ~4−6 bright stars within the HST image. The PSF reconstruction is carried out in multiple iterations with model optimization having fixed PSFs interlaced in between the PSF reconstruction iterations (see Birrer et al. 2019 for details, and also Chen et al. 2016 for a similar algorithm).

[5] Using the 'PROFILE_SHEAR' solver of LENSTRONOMY.

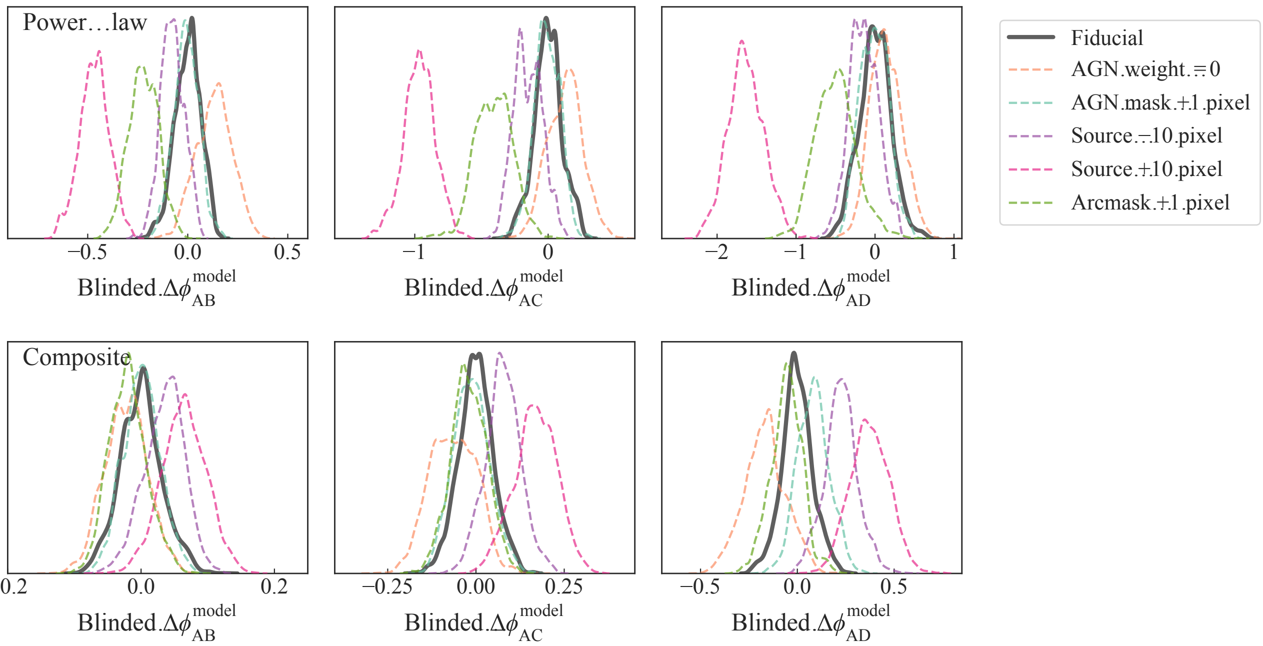

Fig. 7. Model-predicted distributions of Fermat potential differences (blinded) for each of the GLEE models tested, with power-law models (*top*) and composite models (*bottom*).

There is an offset between the recorded coordinates between IR and UVIS images. After each iteration of PSF reconstruction, we re-align the coordinate system of the IR image with that of the UVIS images using the quasar positions (Shajib et al. 2019). Thus, we have a block of three operations constructing one unit of PSF reconstruction iteration: (i) IR-band image re-alignment, (ii) lens model optimization, and (iii) PSF reconstruction.

We optimized the lens model using the particle swarm optimization (PSO) method (Kennedy & Eberhart 1995), which is implemented in LENSTRONOMY. After the PSF reconstruction procedure, we performed MCMC sampling of the model posterior using EMCEE (Foreman-Mackey et al. 2013), which is an affine-invariant ensemble sampler (Goodman & Weare 2010). We chose the number of walkers to be eight times the number of sampled parameters. We run the chain for 10 000 steps. We check for convergence of the chain by manually inspecting that the median and standard deviations of the parameters within the walkers at each step has reach equilibrium for at least 1000 steps. We take the walker distribution from the last 1000 steps of the chain to be the model posterior.

We illustrate the best-fit model from the model setup with the lowest BIC value among the power-law and composite model families in Figs. 8 and 9, respectively.

6.3. Bayesian model averaging

We have 24 models that make up our set of models $\{S, M\}$, with each lens model family from $M \equiv \{\text{powerlaw}, \text{composite}\}$ has 12 different hyper-parameter settings in S. We approximate the integral on the right-hand side of Eq. (25) as a discrete summation over S as

$$\int p(\xi \mid O_{\rm img}, M, S)\; p(O_{\rm img} \mid M, S)\; p(S)\, {\rm d}S$$
$$\approx \sum_n \Delta S_n\; p(S_n)\; p(\xi \mid O_{\rm img}, M, S_n)\; p(O_{\rm img} \mid M, S_n). \qquad (32)$$

Here, ΔS_n can be interpreted as the model space volume that represents the model S_n. Although the models $\{S_n\}$ differ from each other by discrete steps, an appropriately chosen expression for ΔS_n can account for sparse sampling from the model space, as we cannot adopt a sufficiently large number of models that are densely populated in the model space due to computational limitation. We use the BIC score of a model as a proxy for the model evidence $p(O_{\rm img} \mid M, S_n)$. Thus, $\exp(-\Delta{\rm BIC}/2)$ acts as the evidence ratio and provides the relative weight between two models. The ΔS_n term is effectively an additional weighting on top of this BIC weighting (Birrer et al. 2019; Shajib et al. 2020). We take $p(S_n) = 1$ and therefore need to effectively implement the weighted sum of $p(\xi \mid O_{\rm img}, M, S_n)$ in the right-hand side of Eq. (32) through sampling.

We tabulate the BIC values of the models in Table 2. The BIC values are computed from the maximum sampled likelihood in each MCMC chain. We estimate the sparsity of models $\{S_n\}$ by taking the variance $\sigma^{\rm model^2}_{\Delta{\rm BIC}}$ of ΔBIC between 'neighbouring' models that differ with each other by one step in only one setting (Shajib et al. 2020). We furthermore accounted for the numeric uncertainty in estimation of ΔBIC by taking the variance $\sigma^{\rm numeric^2}_{\Delta{\rm BIC}}$ of ΔBIC between identical models that we have optimized twice. These twin runs produce slightly different posteriors – and thus BIC values – due to stochasticity in the PSF reconstruction, PSO, and MCMC sampling steps, similar to what was done in Birrer et al. (2019). Thus, our total variance in ΔBIC is

$$\sigma^2_{\Delta{\rm BIC}} \equiv \sigma^{\rm model^2}_{\Delta{\rm BIC}} + \sigma^{\rm numeric^2}_{\Delta{\rm BIC}}. \qquad (33)$$

We compute that $\sigma^{\rm model}_{\Delta{\rm BIC}} = 304$ and $\sigma^{\rm numeric}_{\Delta{\rm BIC}} = 69$. To implement the ΔS_n weighting through sampling, we first follow Birrer et al. (2019) to obtain the absolute weight $W_{n,{\rm abs}}$ of the nth model by convolving the ΔBIC with the evidence ratio function $f(x)$ as

$$W_{n,{\rm abs}} = \frac{1}{\sqrt{2\pi}\sigma_{\Delta{\rm BIC}}} \int_{-\infty}^{\infty} f(x) \exp\left[-\frac{({\rm BIC}_n - x)^2}{2\sigma^2_{\Delta{\rm BIC}}}\right] {\rm d}x, \qquad (34)$$

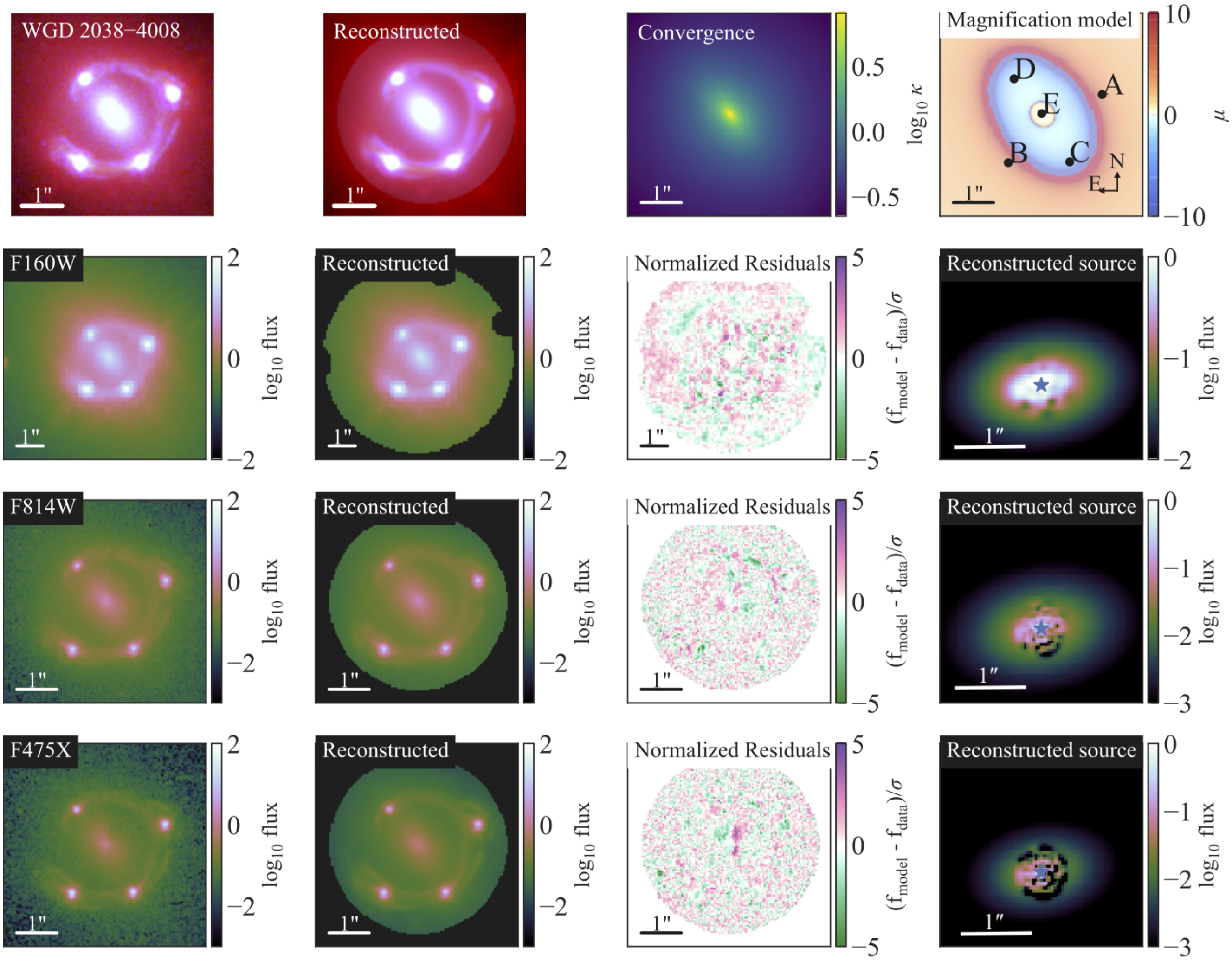

Fig. 8. Most likely LENSTRONOMY lens model and reconstructed image of WGD 2038−4008 using the power-law model. The *top row* shows, from left to right, the observed RGB image, the reconstructed RGB image, the convergence profile, and the magnification model. The next three rows show, from left to right, the observed image, the reconstructed image, the residual, and the reconstructed source for each of the HST filters. The three filters are *F*160*W* (*second row*), *F*814*W* (*third row*), and *F*475*X* (*fourth row*). All the scale bars in each panel correspond to 1″. The star symbol in the reconstructed source panels marks the position of the quasar host galaxy's centroid.

where the evidence ratio function $f(x)$ is defined using the BIC difference as

$$f(x) \equiv \begin{cases} 1 & x < \mathrm{BIC_{min}}, \\ \exp\left(\mathrm{BIC_{min}} - x\right) & x \geq \mathrm{BIC_{min}}. \end{cases} \tag{35}$$

Then, we obtain the relative weight $W_{n,\mathrm{rel}}$ by normalizing the absolute weights by the maximum absolute weight as

$$W_{n,\mathrm{rel}} = \frac{W_{n,\mathrm{abs}}}{\max\left(\{W_{n,\mathrm{abs}}\}\right)}. \tag{36}$$

Finally, we combine the individual model posteriors following Eq. (32) as

$$\sum_n \Delta S_n\, p(S_n)\, p(O_{\mathrm{img}} \mid M, S_n)\, p(\xi \mid O_{\mathrm{img}}, M, S_n)$$
$$\propto \sum_n W_{n,\mathrm{rel}}\, p(\xi \mid O_{\mathrm{img}}, M, S). \tag{37}$$

In Sect. 6.4, we compare the two mass model families after performing the above model averaging procedure within each model family.

6.4. Lensing-only constraints on the Fermat potential difference

We constrain the image positions with uncertainty 0.″002 for the power-law model, and with uncertainty 0.″004 for the composite model. Given the longest predicted time-delay for this system, these precisions are well below the astrometric requirement of ∼0.″02 uncertainty so that the astrometric uncertainty is subdominant to achieve ≤5% uncertainty in H_0 from this single system (Birrer & Treu 2019).

In Fig. 10 we compare between the model settings – namely, the choices for $n_{\max}$ and the size of the image likelihood computation region. All the posteriors within a particular lens model family (i.e. power law or composite) are statistically consistent with each other within 1σ.

We compare the combined model posteriors between the power-law and composite models in Fig. 11. The Fermat potential differences deviate by 16–21% between the power-law and composite model setups. However, the MST-invariant quantity $\xi_{\mathrm{rad}} \propto \theta_{\mathrm{E}} \alpha''_{\mathrm{E}}/(1-\kappa_{\mathrm{E}})$ is consistent between the two model setups (see Eq. (42) of Birrer 2021 for the full definition of ξ_{rad}, and also Kochanek 2020). Thus, the difference in the Fermat potential

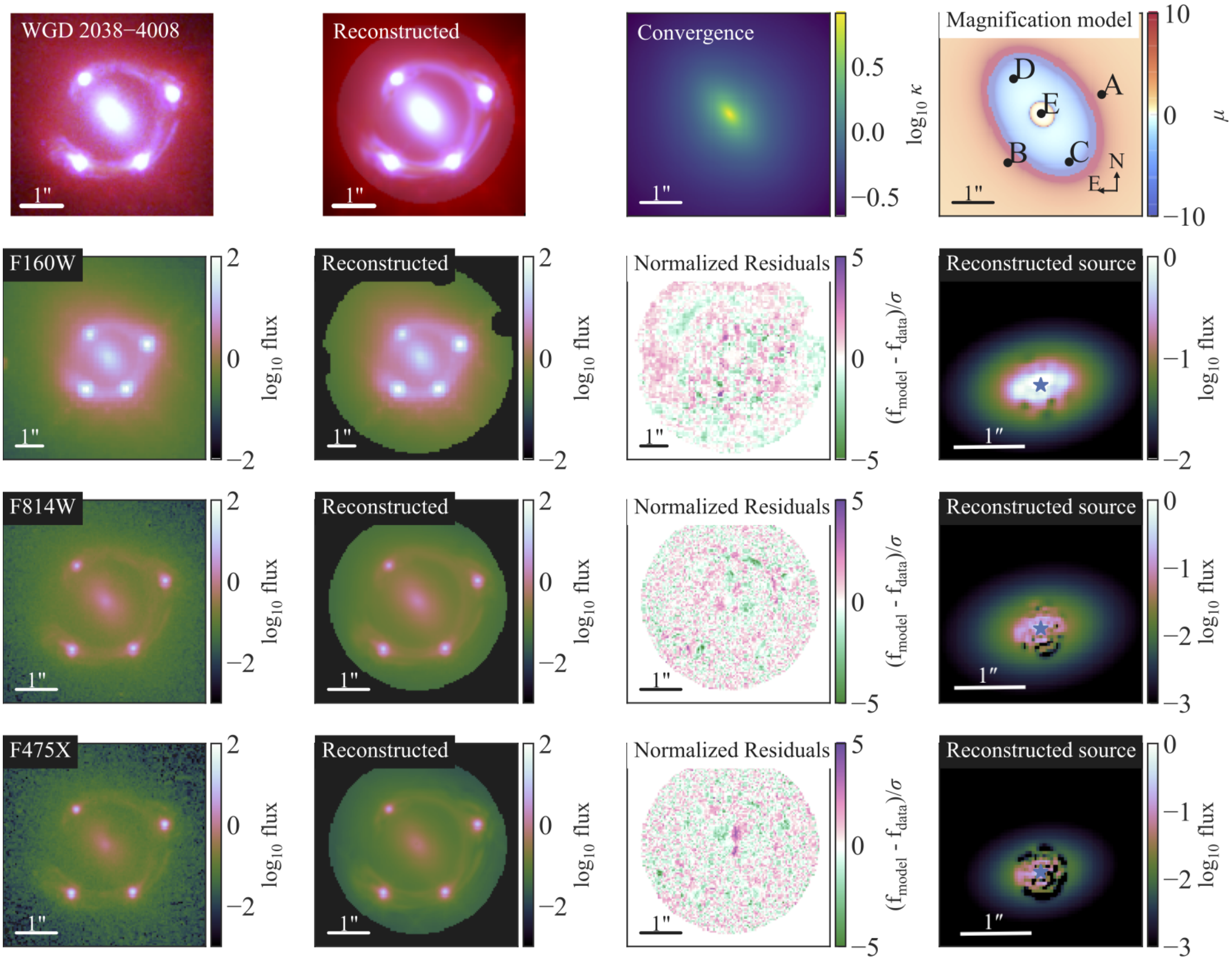

Fig. 9. Most likely LENSTRONOMY lens model and reconstructed image of WGD 2038−4008 using the composite model. The *top row* shows, from left to right, the observed RGB image, reconstructed RGB image, the convergence profile, and the magnification model. The next three rows show, from left to right, the observed image, the reconstructed image, the residual, and the reconstructed source for each of the HST filters. The three filters are $F160W$ (*second row*), $F814W$ (*third row*), and $F475X$ (*fourth row*). All the scale bars in each panel correspond to $1''$. The star symbol in the reconstructed source panels marks the position of the quasar host galaxy's centroid. In the magnification model, a central image is predicted due to a central core in the triple Chameleon light profile. However, this central image is highly de-magnified, with magnification 0.019 ± 0.02, and thus its presence cannot be ascertained in our imaging data.

from the two model setups can be interpreted as a manifestation of the internal MSD. We combine the stellar kinematics and estimated external convergence with the lens models to mitigate the internal MSD in Sect. 6.5.

However, we first check for potential unphysical properties in our best fit composite models as the source of the large difference in the Fermat potential in Sects. 6.4.1 and 6.4.2. These checks were performed prior to un-blinding the models.

6.4.1. Halo properties in the composite model

Figure 12 illustrates the M_{200}–c relation posterior for our system in comparison with the adopted prior; the median of the concentration posterior is consistent with the concentration prior within 1σ. Our combined posterior from the composite model setup provides the total halo mass $\log_{10}(M_{200}/M_\odot) = 13.04^{+0.14}_{-0.13}$ and the total stellar mass is $\log_{10}(M_\star/M_\odot) = 11.87^{+0.01}_{-0.03}$. The total stellar mass is obtained by doubling the enclosed mass within the half-light radius of $3\farcs2$ corresponding to the $F160W$ band. The projected dark matter fraction within the Einstein radius is $0.22^{+0.06}_{-0.02}$. The total baryon-to-dark-matter fraction is $0.07^{+0.03}_{-0.02}$, which is consistent with the upper limit set by the cosmic baryonic fraction 0.19 (Planck Collaboration VI 2020). In Fig. 13 we plot the azimuthally averaged convergence profiles for the power-law and composite models to illustrate the difference in the convergence slope at the Einstein radius $\theta_{\rm E}$. The inner region ($\lesssim 0\farcs2$) of the triple Chameleon profile is flat unlike the singular centre in the power-law model. The flat or cored convergence profile at the centre gives rise to an inner critical curve in the image plane (Fig. 9). This core in the centre of the stellar mass distribution follows from the stellar flux distribution, as the radial flux profile from isophotal fitting also shows a stellar core. We fit a core Sérsic profile – that is defined by Eq. (2) in Dullo (2019) – to the azimuthally averaged light profile from our isophotal fitting to obtain the stellar core radius. We obtain 0.780 ± 0.004 kpc assuming a fiducial flat ΛCDM cosmology with $H_0 = 70\,{\rm km\,s^{-1}\,Mpc^{-1}}$ and $\Omega_{\rm m} = 0.3$. This stellar core radius is consistent with the core radii measured in local elliptical galaxies (0.64–2.73 kpc; Bonfini & Graham 2016; Dullo 2019). In Fig. 14, we illustrate the velocity dispersion profiles

Table 2. BIC values for different LENSTRONOMY model setups.

Source light $n_{\max}$	Likelihood computation region size	Run number	ΔBIC	BIC weight
Power-law ellipsoid model				
{9, 13}	{2.2, 3.6}	2	0	1.00
{9, 13}	{2.2, 3.6}	1	43	0.95
{8, 12}	{2.2, 3.6}	2	209	0.76
{8, 12}	{2.2, 3.6}	1	235	0.73
{7, 11}	{2.2, 3.6}	2	606	0.37
{7, 11}	{2.2, 3.6}	1	726	0.29
{9, 13}	{2.3, 3.7}	2	2129	0.00
{8, 12}	{2.3, 3.7}	2	2350	0.00
{9, 13}	{2.3, 3.7}	1	2366	0.00
{7, 11}	{2.3, 3.7}	2	2778	0.00
{8, 12}	{2.3, 3.7}	1	2786	0.00
{7, 11}	{2.3, 3.7}	1	2793	0.00
Composite model				
{9, 13}	{2.2, 3.6}	1	0	1.00
{9, 13}	{2.2, 3.6}	2	10	0.99
{8, 12}	{2.2, 3.6}	1	318	0.64
{8, 12}	{2.2, 3.6}	2	449	0.51
{7, 11}	{2.2, 3.6}	1	604	0.37
{7, 11}	{2.2, 3.6}	2	675	0.32
{9, 13}	{2.3, 3.7}	2	2009	0.00
{9, 13}	{2.3, 3.7}	1	2191	0.00
{8, 12}	{2.3, 3.7}	2	2373	0.00
{8, 12}	{2.3, 3.7}	1	2378	0.00
{7, 11}	{2.3, 3.7}	2	2807	0.00
{7, 11}	{2.3, 3.7}	1	2912	0.00

Notes. The difference ΔBIC is calculated only within the particular mass profile family – power law or composite. The model setups are ordered from lower to higher BIC values within each mass profile family. The ΔBIC values are calculated relative to the model setup with the lowest BIC value. The relative weights for each model are obtained from ΔBIC adjusted for sparse sampling from the model space as described in Sect. 6.3.

predicted by the lens model posteriors assuming isotropic orbit and a flat ΛCDM cosmology with $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. The composite-model-predicted velocity dispersion profile decreases towards the centre by ~20% due to the flattened mass profile. Such a large decrease in the velocity dispersion has not been observed in local massive ellipticals (e.g. Cappellari 2016; Ene et al. 2019). As a result, the composite lens model posterior suggests an inconsistency with kinematic observations of the local ellipticals. Interestingly, an inner critical curve in our composite model predicts a central image. The magnifications of the 5 images are A: $-1.6^{+0.1}_{-0.2}$, B: $4.0^{+0.2}_{-0.3}$, C: $-3.6^{+0.2}_{-0.4}$, D: 4.8 ± 0.3, and central: 0.019 ± 0.002. The predicted appearance of the central image is invariant under the MST. However, the demagnified and potentially dust-extincted central image is indistinguishable in the present imaging data. Thus, we are unable to distinguish the two profile families on the basis of the presence or absence of the central image.

6.4.2. Test of potential systematics from modelling choices and priors

We checked if our composite models are robust against potential biases from our particular model choices, for example the likelihood computation region and the prior on the NFW halo scale radius. We first optimized a lens model with the power-law mass profile and the triple Chameleon profile for the light instead of the triple Sérsic profile. We then took these best fit parameters for the triple Chameleon profile and fixed them in the test composite setup. We let the overall scaling of the baryonic mass and light distributions be free, thus effectively allowing for a free mass-to-light ratio (M/L). We adopted a halo mass prior for the NFW profile dependent on the stellar mass given by

$$p(\log_{10} M_{200} \mid \log_{10} M_\star^{\mathrm{Chab}})$$
$$\equiv \mathcal{N}\left(\mu_{\mathrm{h}} + \beta_{\mathrm{h}}\left[\log_{10} M_\star^{\mathrm{Chab}} - 11.3\right],\ \sigma_{\mathrm{h}}\right), \tag{38}$$

where $M_\star^{\mathrm{Chab}}$ is the total stellar mass based on the SPS method assuming a Chabrier IMF (Sonnenfeld et al. 2018). Here, the parameters are $\mu_{\mathrm{h}} = 13.11 \pm 0.04$, $\beta_{\mathrm{h}} = 1.43 \pm 0.15$, and $\sigma_{\mathrm{h}} = 0.23 \pm 0.04$. We measure the stellar mass $M_\star^{\mathrm{Chab}}$ from the total fluxes in the three HST bands. We fit the surface brightness profile of the deflector separately in three bands from large cutouts that fully contains the light distribution of the deflector (see Fig. 15). First, we subtract lensed arcs and the quasar images from these cutouts using our best fit power-law model. Then, we fit elliptical isophotes using the PHOTUTILS software (Bradley et al. 2020). The method `photutils.isophote.Ellipse.fit_image()` allows a convenient way to ignore the overlapping objects (i.e. stars and galaxies) by sigma-clipping pixels along an isophote. Thus, we do not need to mask out these overlapping objects. We reconstructed the surface brightness profile of the deflector based on the fitted isophotes, which effectively interpolates through the pixels that are contaminated by overlapping objects. We obtain the total flux in each HST band from the reconstructed surface brightness profile using the fitted isophotes. We used PYGALEXEV[6] to obtain $M_\star^{\mathrm{Chab}}$, which is a PYTHON wrapper for GALAXEV (Bruzual & Charlot 2003). By adopting the Basel Stellar Library (BaSeL; Lejeune et al. 1998), exponentially decaying stellar formation history, and free metallicity, we obtain $\log_{10} M_\star^{\mathrm{Chab}} = 11.57^{+0.16}_{-0.13}$. Thus, from Eq. (38), our Gaussian prior on the halo mass $\log_{10} M_{200}$ has mean 13.5 and standard deviation 0.3. We additionally adopt a prior for the total stellar mass $\log_{10} M_\star$. We take an ad hoc prior that is uniform between 11.51 and 11.88 and drops off like a Gaussian function outside these limits. The range between 11.51 and 11.88 accounts for the unknown IMF and spans the range between light (e.g. Chabrier) and heavy (e.g. Salpeter) IMFs that differ by ~0.25 dex in the stellar mass. The exact form of this prior is

$$p(\log_{10} M_\star^{/\odot}) = \begin{cases} A \exp\left[-\frac{(\log_{10} M_\star^{/\odot} - 11.51)^2}{2\times 0.13^2}\right], & \log_{10} M_\star^{/\odot} < 11.51, \\ A, \quad 11.51 \leq \log_{10} M_\star^{/\odot} \leq 11.88, \\ A \exp\left[-\frac{(\log_{10} M_\star^{/\odot} - 11.88)^2}{2\times 0.16^2}\right], & \log_{10} M_\star^{/\odot} > 11.88, \end{cases} \tag{39}$$

where $M_\star^{/\odot} \equiv M_\star / M_\odot$, A is the amplitude that normalizes the probability distribution to have $\int p(\log_{10} M_\star^{/\odot})\,\mathrm{d}(\log_{10} M_\star^{/\odot}) = 1$. The actual value of A is not required for sampling in the MCMC method. We also allow an additional uncertainty of ±0.06 dex in $M_\star$ to allow 15% uncertainty on the assumed H_0 in the SPS-based stellar mass estimation.

We furthermore mask out the central region in the deflector galaxy in the test composite model setup so that the optimization does not incentivize the presence of a central image to make up

[6] https://github.com/astrosonnen/pygalaxev

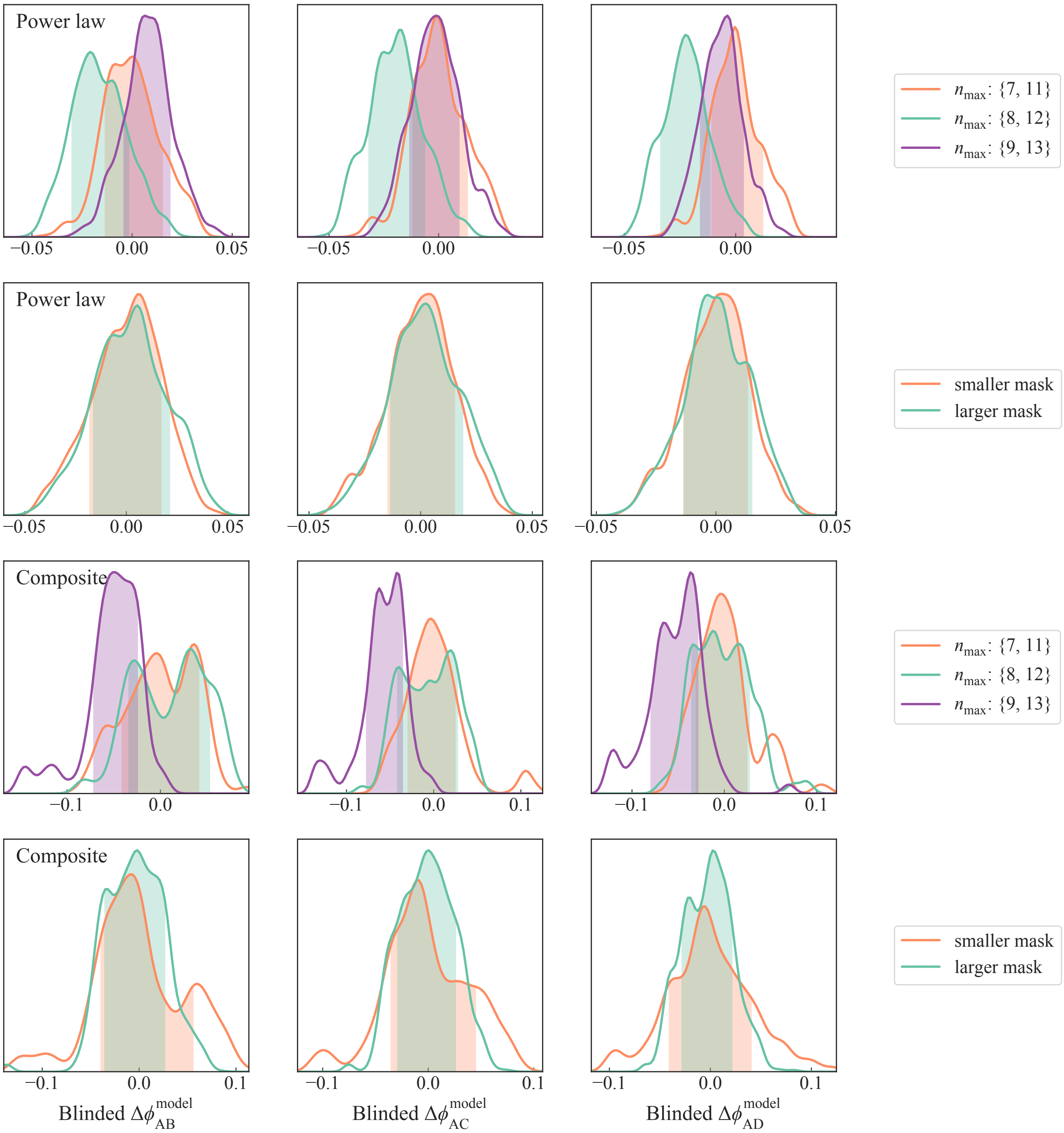

Fig. 10. Comparison between the Fermat potential difference posteriors from different LENSTRONOMY model settings. The posterior for a particular setting is obtained by averaging over models that differ in the other model settings, but within a particular model family using the procedure described in Sect. 6.3. *Top two rows*: correspond to the power-law mass model families, and *bottom two rows*: correspond to the composite mass model families. The shaded regions illustrate the 68% credible regions. The posteriors are blinded by $\Delta\phi^{\rm blinded} \equiv \Delta\phi/\overline{\Delta\phi_{\rm ref}} - 1$, where the 'ref' subscript refers to one of the compared models. The potential differences are consistent within 1σ between our adopted choices of model settings.

for residuals in the deflector light galaxy model that cannot be fully accounted by the triple Chameleon light profile.

We perform the 12 different model setups also for this test composite model and combine the posteriors based on their BIC values. We compare our primary composite model with the test composite model in Fig. 16. Although the Fermat potential differences are consistent between these two model setups within 1σ, the ones from the test setup are smaller by 4−7% than the primary setup. For the test setup the stellar mass is $\log_{10}(M_\star/M_\odot) = 11.84^{+0.02}_{-0.01}$, the halo mass is $\log_{10}(M_{200}/M_\odot) = 13.3^{+0.1}_{-0.3}$, the total baryonic fraction is $0.04^{+0.03}_{-0.01}$, and the dark matter fraction within the Einstein radius is $0.28^{+0.02}_{-0.03}$. The total halo mass increases in the test setting over the one obtained from our primary setting due to the adopted halo mass prior. As a result, the increase in the halo normalization leads to a decrease in the Fermat potential, or equivalently the shallowing of the convergence profile. This impact of the prior on the halo profile normalization is the same as observed by Shajib et al. (2021) for the

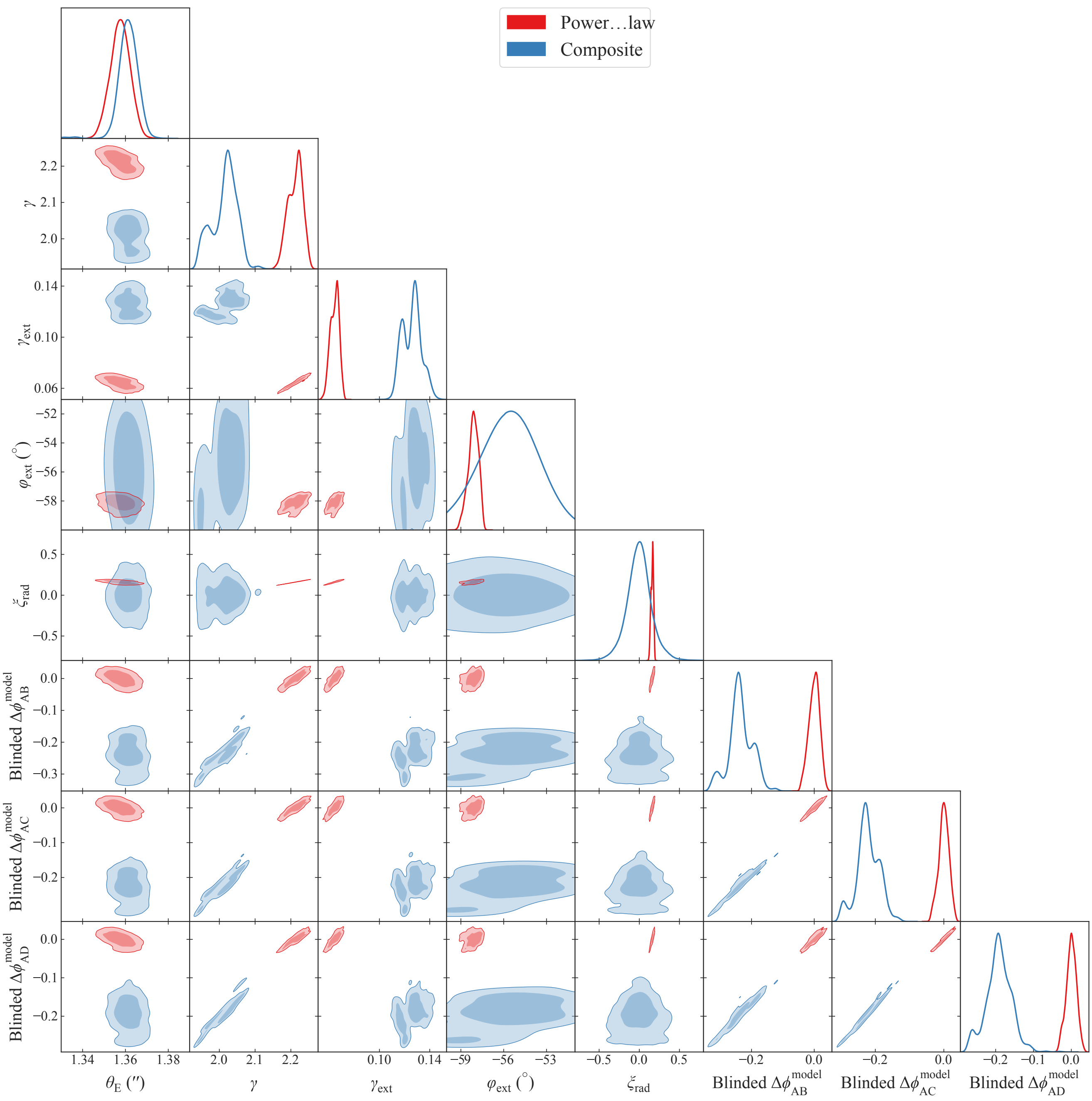

Fig. 11. Comparison of the LENSTRONOMY lens model parameters and Fermat potential difference between the power-law (red) and composite (blue) mass models. The posteriors are obtained by averaging over all the model settings following the procedure described in Sect. 6.3. The parameter γ for the composite model is computed from circularly averaging the quantity $2 - [{\rm d}\log\alpha(r)/{\rm d}\log r]_{r=\theta_{\rm E}}$. Some parameters are blinded as $p^{\rm blinded} \equiv p/\overline{p_{\rm pl}} - 1$ for $p \in \{\gamma, \Delta\phi\}$, where the subscript 'pl' refers to the power-law model posteriors. The composite model-predicted γ is approximately 8% smaller than that from the power law. Consequently, the Fermat potential differences are smaller by approximately 16% for the composite model.

SLACS lenses. As adopted priors can systematically shift Fermat potential differences, more physically motivated priors are not sufficient to explain all the differences between the composite and the power-law models.

The mass difference of $3.8 \times 10^{10}\,M_\odot$ within $0\overset{\prime\prime}{.}2$ (assuming flat ΛCDM cosmology with $H_0 = 70\,{\rm km\,s^{-1}\,Mpc^{-1}}$) between the power-law and composite models could be explained by an ultra-massive black hole (e.g. Mehrgan et al. 2019; Dullo 2019). Another potential solution that would push the Fermat potential differences from the composite models towards the ones from the power-law models is to have stellar mass-to-light ratio gradient ($\eta \sim 0.27$ with $M_\star/L \propto R^{-\eta}$) to steepen up the total mass density profile. Although Shajib et al. (2021) find no evidence for a significant mass-to-light ratio on average for SLACS lenses at the similar redshift ($\langle z\rangle \sim 0.2$) as WGD 2038−4008 ($z_{\rm d} = 0.23$), there can still be individual cases with steep mass-to-light ratio gradients. Moreover, the value $\eta \sim 0.27$ will be consistent with the constraints $\langle\eta\rangle = 0.24 \pm 0.04$ reported by Sonnenfeld et al. (2018). Adopting a mass-to-light ratio gradient for the luminous component or including a central black hole in the composite

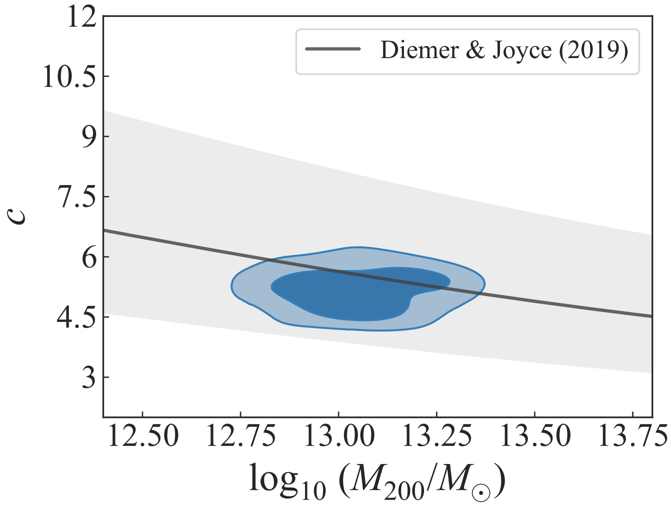

Fig. 12. Distribution of M_{200} and c_{200} parameters for the NFW halos in LENSTRONOMY composite model (blue shaded region). This distribution is averaged over all the model settings within the composite model family following the procedure described in Sect. 6.3. The 2 contours correspond to the 68% and 95% credible regions, respectively. The black solid line traces the theoretical prediction of the $M_{200}-c_{200}$ relation at $z_d = 0.230$ from Diemer & Joyce (2019) with the grey shaded region corresponding to the 68% confidence interval. We adopt this $M_{200}-c$ relation as a prior in our analysis in addition to a M_{200} prior based on Sonnenfeld et al. (2018).

lens model is beyond the scope of this study. At this point, however, there is no a priori reason to modify the composite model to push the Fermat potential differences towards the ones from the power-law models. We use the kinematics data to bridge the discrepancy between the two models or to be the decider between them next in Sect. 6.5. As the composite models are related to the true underlying mass distribution through an approximate MST, we rely on the kinematics data to appropriately adjust the Fermat potential differences along the MSD towards the true values.

6.5. Combining with stellar kinematics and external convergence

In this subsection we perform dynamical modelling of the deflector based on our lens models using Eq. (17). We need to estimate the luminosity density $l(r)$ to use in this equation by deprojecting the surface brightness profile of the deflector. The surface brightness of WGD 2038−4008 extends far beyond the size of our likelihood computation region (Fig. 15). Thus, deprojecting the light profile fit from our lens models may potentially produce early truncation in the 3D luminosity distribution along the LOS. Therefore, we use the reconstructed surface brightness profile from the fitted isophotes in the $F814W$ band from Sect. 6.1.1 (see Fig. 15). We chose the light profile from the $F814W$ band, because the velocity dispersion was measured in the optical. From this reconstructed surface brightness profile, we numerically find the circular aperture that contains half of the total light as $\theta_{\rm eff} = 2\farcs4$. The aperture size in this numeric computation can only grow by a size of a pixel, which is $0\farcs08$. We adopt a 3% Gaussian uncertainty for $\theta_{\rm eff}$, which combines one pixel size as a systematic uncertainty with a typical 2% uncertainty for $\theta_{\rm eff}$ from fitting surface brightness profiles with continuous parameters from Shajib et al. (2021). We check that adopting 20% uncertainty for $\theta_{\rm eff}$ or using the $\theta_{\rm eff}$ measured from the $F160W$ band does not significantly impact ($\lesssim$0.1%) the resultant Fermat potential difference (Fig. 17). We approximate the reconstructed surface brightness profile with an elliptical 2D multi-Gaussian expansion (MGE) series (Cappellari 2002). The MGE approximation allows for a straightforward deprojection into a 3D light profile, which we use as $l(r)$ in Eq. (17). Since we only solve the Jeans equation in the spherical case, we adopt the spherical equivalent of the elliptical Gaussian components by taking the Gaussian scales along the intermediate axes. We also apply a self-consistent 2% uncertainty to the MGE scale parameters by letting them vary with $\theta_{\rm eff}$. We adopt a uniform prior on $\log a_{\rm ani}$, where $a_{\rm ani}$ is the anisotropy scaling factor defined by $r_{\rm ani} \equiv a_{\rm ani}\theta_{\rm eff}$. Birrer et al. (2016, 2020) find that the uniform prior for $\log a_{\rm ani}$ is a less informative choice than the uniform prior on $a_{\rm ani}$.

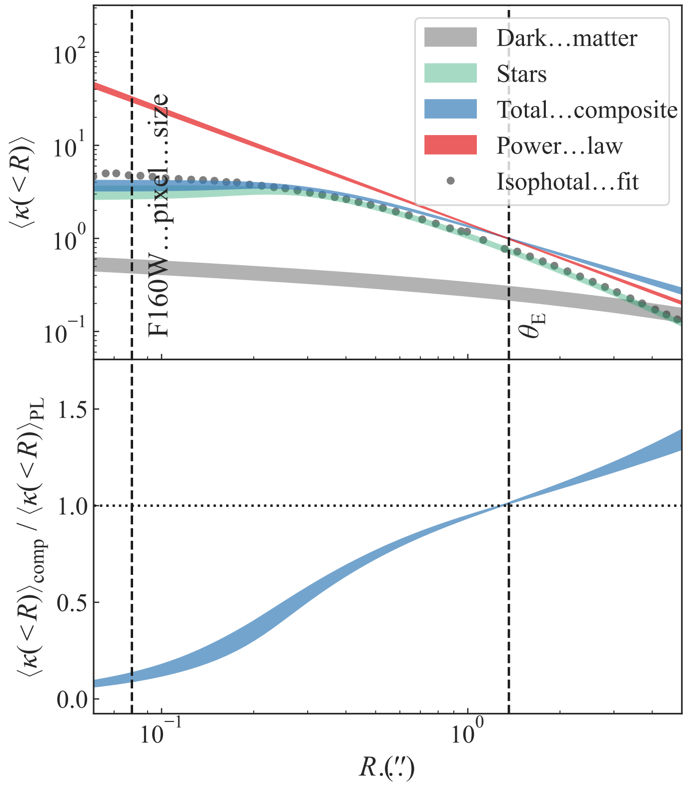

Fig. 13. Radial mass profiles of the central deflector constrained by the LENSTRONOMY models. *Top*: circularly averaged convergence $\langle\kappa(<R)\rangle$ as a function of radius from LENSTRONOMY power-law (red) and composite (blue) models. The stellar (green) and dark matter (grey) distributions in the composite model are also individually illustrated. The shaded regions encompass the 16th and 84th percentile of the sampled profiles for the corresponding case. The grey points illustrate the surface brightness profile fitted with isophotes as described in Sect. 6.1.1. The amplitude of the isophotal fit is normalized to match with the triple Chameleon profile (green shaded region) at θ_E for the purpose of this illustration. The vertical black dashed lines mark the pixel size in the $F160W$ band and the best fit Einstein radius. *Bottom*: ratio of the circularly averaged convergence profiles between the composite and the power-law models. At the Einstein radius the convergence slope deviates by 16−21% between the two model setups.

We performed a test for systematics in our velocity dispersion modelling. We adopted two test cases: (i) where the $\theta_{\rm eff}$ uncertainty is taken as 20%, and (ii) where the light profile from the $F160W$ band is used in the kinematic computation.

For the external convergence $\kappa_{\rm ext}$, we impose a selection criterion on the $P(\kappa_{\rm ext})$ estimated in Buckley-Geer et al. (2020) by

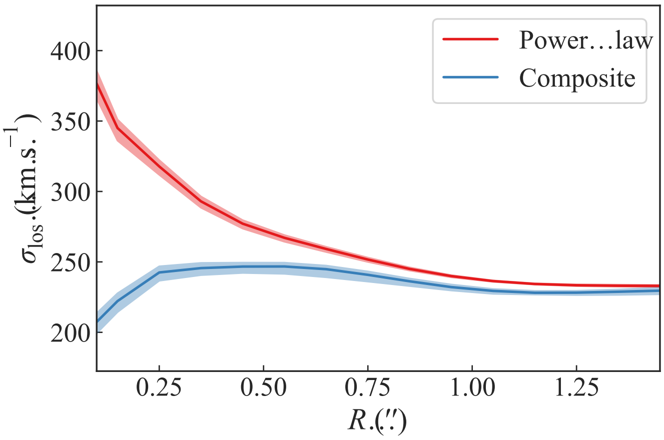

Fig. 14. Line-of-sight velocity dispersion profile (circularized) corresponding to the lens model posteriors in the power-law (red) and composite (blue) models from LENSTRONOMY. Isotropic stellar orbits are assumed in computing these velocity dispersion profiles. The solid lines correspond to the median and the shaded regions encompass the 16th and 84th percentiles. The composite profile predicts a decrease in the velocity dispersion towards the centre due to the flattened core – which is not observed in local massive ellipticals (e.g. Cappellari 2016; Ene et al. 2019) – thus pointing to the atypicality of the posterior mass profile in the composite model.

requiring that the selected LOSs also correspond to the combined (through BIC weighting) external shear value from our lens models within a lens model family (i.e. power law or composite). In Fig. 18 we illustrate the two $\kappa_{\rm ext}$ distributions consistent with the external shear values for the power-law and composite mass profiles[7].

We combine the stellar kinematics and external convergence information with the lens model posterior in two different ways: (i) with fixed $\lambda_{\rm int} = 1$ (Sect. 6.5.1), and (ii) with free $\lambda_{\rm int}$ constrained by the stellar kinematics (Sect. 6.5.2).

6.5.1. The case with $\lambda_{\rm int} = 1$

Assuming $\lambda_{\rm int} = 1$, the model-predicted velocity dispersion can be written as

$$\sigma^2_{\rm ap} = (1 - \kappa_{\rm ext}) \frac{D_{\rm s}}{D_{\rm ds}} c^2 J(\xi_{\rm lens}, \xi_{\rm light}, \beta_{\rm ani}). \tag{40}$$

This assumption when combining the stellar kinematics information with the lens model posteriors is the same as done in earlier TDCOSMO analyses prior to Birrer et al. (2020, TDCOSMO-IV), for example in Suyu et al. (2013), Wong et al. (2017), and Rusu et al. (2020). We assume a flat ΛCDM cosmology with $\Omega_{\rm m} = 0.3$ to compute the fiducial distance ratio $D_{\rm s}/D_{\rm ds}$. To combine the stellar kinematic information, we consider the kinematics likelihood function

$$\log \mathcal{L}_{\rm kin} = -\frac{(\sigma_{\rm model} - \sigma_{\rm measured})^2}{2\sigma^2_{\sigma_{\rm measured}}} - \frac{1}{2} \log \left(2\pi\sigma^2_{\sigma_{\rm measured}}\right). \tag{41}$$

We first combine the lens model posteriors from the power-law and composite models with equal weights, and then importance sample from this combined posterior with weight $\mathcal{L}_{\rm kin}$ (Lewis & Bridle 2002). We note that each of the power-law and composite models are already averaged over the various adopted model settings within each mass model family following our BMA procedure from Sect. 6.3. We illustrate the Fermat potential differences from each of the power-law and composite models in Fig. 19.

After combining the kinematics information with $\lambda_{\rm int}$, the combined posterior for the Fermat potential differences end up mostly similar to the power-law posterior, as the kinematic likelihood heavily down-weights the posterior from the composite model. Although the composite model was designed with physical motivations to mimic a real galaxy structure, the kinematics data heavily disfavours the composite lens model posterior for $\lambda_{\rm int} = 1$. Furthermore, applying more physical priors to resolve this inconsistency rather makes the kinematics data disfavour the composite model more, which suggest that our composite model is not adequate in describing the true galaxy mass distribution. Further generalization in the composite model (e.g. mass-to-light ratio gradient and generalized NFW halo) may thus be necessary for a composite model to be simultaneously consistent with the lensing data, the kinematics data, and the cosmological expectations for galaxies, e.g. the $M{-}c$ relation, baryonic fraction. The uncertainties on the combined Fermat potential differences, and thus on the predicted time delays, are approximately 4%, which is comparable with those from the previous TDCOSMO analyses under the same assumption of $\lambda_{\rm int} = 1$ (Wong et al. 2020).

6.5.2. The case with free $\lambda_{\rm int}$

Now, we treat $\lambda_{\rm int}$ as a free parameter and constrain it using the stellar kinematics by re-expressing Eq. (22) as

$$\lambda_{\rm int} = \frac{\sigma^2_{\rm ap}}{(1 - \kappa_{\rm ext})(D_{\rm s}/D_{\rm ds}) c^2 J(\xi_{\rm lens}, \xi_{\rm light}, \beta_{\rm ani})}. \tag{42}$$

Such constraining of $\lambda_{\rm int}$ from stellar kinematics is the same approach as Birrer et al. (2020, TDCOSMO-IV), albeit these authors achieved a tighter constraint on $\lambda_{\rm int}$ from a joint sample of seven time-delay lens systems and 33 non-time-delay lenses through a hierarchical Bayesian analysis.

We obtain the $D_{\rm s}/D_{\rm ds}$ distribution to use in Eq. (42) from the relative distance constrained by the Pantheon SN sample (Scolnic et al. 2018). We approximate the luminosity distance up to the Pantheon supernovae using a fourth-order Taylor expansion. The coefficients in the Taylor expansion allow increasing complexity by including the deceleration parameter q_0, the jerk parameter j_0, the snap parameter s_0, and the curvature density parameter Ω_k. We compute the model evidence using nested sampling for different size of the parameter set (Skilling 2004). We select the model that goes up to the jerk parameter j_0 based on its highest model evidence. We use the relation $D_{\rm A} = D_{\rm L}/(1+z)^2$ to convert the luminosity distance to angular diameter distance and transform the 2D posterior distribution of (q_0, j_0) to obtain the $D_{\rm s}/D_{\rm ds}$ distribution.

[7] While Buckley-Geer et al. (2020) apply a joint constraint of number counts inside the 45″ and 120″ apertures, this would lead to too few LOSs selected from the Millennium simulation once the large shear value constraint of the composite model obtained by the LENSTRONOMY team is imposed. To contain enough LOSs for a robust distribution, the LENSTRONOMY team therefore removed the 45″ aperture constraint. For consistency, this was done for both the power-law and composite mass models. However, the number counts from the 45″ aperture were still retained by the GLEE team, whose composite model shear value is smaller. This difference between the external convergence distributions used by the two teams was revealed to each other only after the un-blinding. While this creates an inconsistency between the two teams, Rusu et al. (2020) show that for large shear values, the $\kappa_{\rm ext}$ distribution is dominated by the shear constraint, and therefore the imposition of the 45″ aperture or the lack thereof is expected to have a negligible impact.

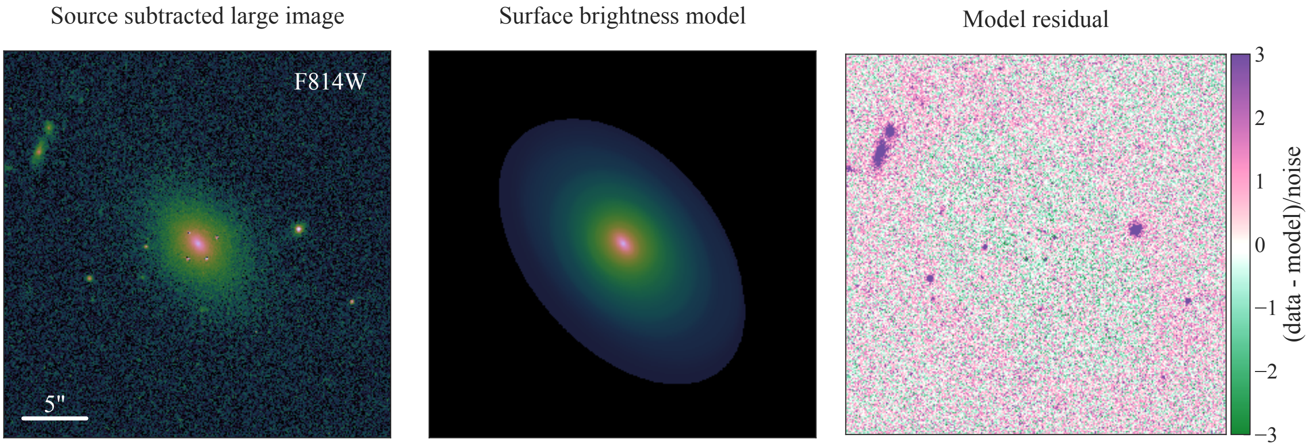

Fig. 15. Fitting the lens galaxy light profile from a much larger cutout than that used for lens modelling. *Left*: large cutout around the deflector galaxy in the $F814W$ band, with the lensed arcs and quasar images subtracted using the best-fit of a power-law lens model. The galaxy extends much further beyond the Einstein radius with $\theta_{\rm eff}/\theta_{\rm E} \approx 1.7$. *Middle*: model for the surface brightness profile of the deflector constructed using elliptical isophotes. *Right*: model residual showing that the model has captured the overall light distribution despite numerous overlapping objects. The model from *middle panel* was further approximated with an elliptical multi-Gaussian expansion (MGE; Cappellari 2002) to allow deprojection along the LOS for kinematic modelling.

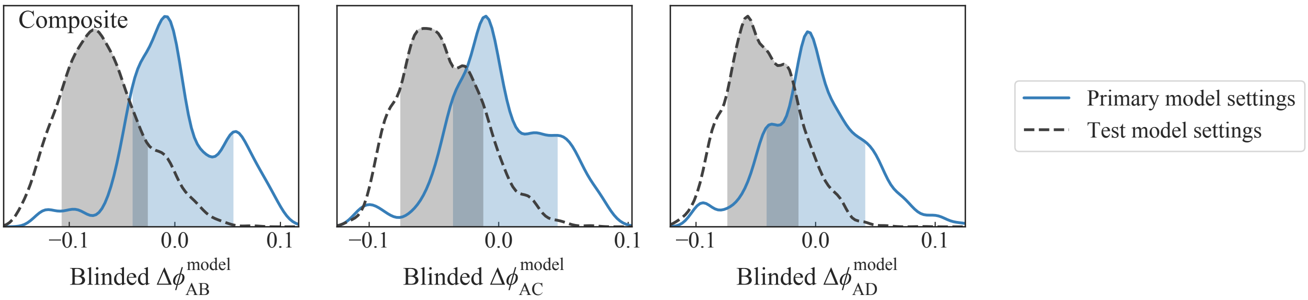

Fig. 16. Comparison of the Fermat potential posteriors between two different model settings for the LENSTRONOMY composite model. The blue solid distributions correspond to our primary model settings. In the test model setup (black dashed distributions), we fix the triple Chameleon profile for the stellar mass and light distributions to best fit values from a separately optimized model with the power-law mass profile. We also mask the central region of the deflector galaxy in the test model setup. Additionally, the prior on the halo mass profile is different. Whereas we adopt a prior on the NFW scale radius $r_{\rm s}$ in the primary setup, we adopt a combination of priors on M_{200} and $M_\star$ in the test setup. Despite multiple differences in the model settings, the Fermat potential differences are consistent within 1σ with each other.

In Fig. 20 we compare the MST-corrected Fermat potential differences between the power-law and the composite models. We find $\lambda_{\rm int}^{\rm pl} = 1.02 \pm 0.15$ for the power-law model and $\lambda_{\rm int}^{\rm comp} = 1.79 \pm 0.53$ for the composite model. This large median value of $\lambda_{\rm int}^{\rm comp}$ falls in the excluded region in Birrer et al. (2020, TDCOSMO-IV) that is based on physical arguments on the mass density distribution. However, the mass profile adopted by Birrer et al. (2020) is a cored power-law profile, which can be interpreted as the presence of a cored component in the NFW profile with its radius being much larger than the Einstein radius. Shajib et al. (2021) demonstrate that deviations from a power-law profile can be explained by shifting the normalization of the NFW profile without any core component. Whereas the MST considered by Birrer et al. (2020, TDCOSMO-IV) allows redistribution of matter only within the dark component, the composite model considered here allows redistribution of matter between dark and luminous components. Thus, large deviations of $\lambda_{\rm int}$ from 1 is still physically plausible, so the large $\lambda_{\rm int}$ produced by our composite model is not in tension with the exclusion range set by Birrer et al. (2020, TDCOSMO-IV).

The predicted Fermat potential difference between the power-law and composite models are consistent within 1σ after adjusting for the internal MST and the external convergence. Thus, we demonstrate that the two model families we adopted are linked through an approximate MST. Therefore, to predict the time delays or to measure H_0 through constraining $\lambda_{\rm int}$ from stellar kinematics, the choice of mass model family is largely irrelevant as the same result can be obtained with any of the conventional model families. In this case, the uncertainty on the Fermat potential difference is dominated by the velocity dispersion uncertainty, which is at 6.4%. As $\lambda_{\rm int} \propto \sigma_{\rm ap}^2$, the uncertainty on the $\lambda_{\rm int}$ is thus expected to be twice the uncertainty of the velocity dispersion. Our obtained uncertainties of $\lambda_{\rm int}^{\rm pl}$ and $\lambda_{\rm int}^{\rm comp}$ are consistent with this expectation. The uncertainties on the predicted time delays from the final combined posterior with free $\lambda_{\rm int}$ is 19–22%.

6.6. Discussion on LENSTRONOMY models

The predicted Fermat potential differences from LENSTRONOMY are discrepant between the power-law and composite models, with the predictions from the composite model being 16–21% lower than those from the power-law model. Both model families fit the imaging data almost equally well, with the

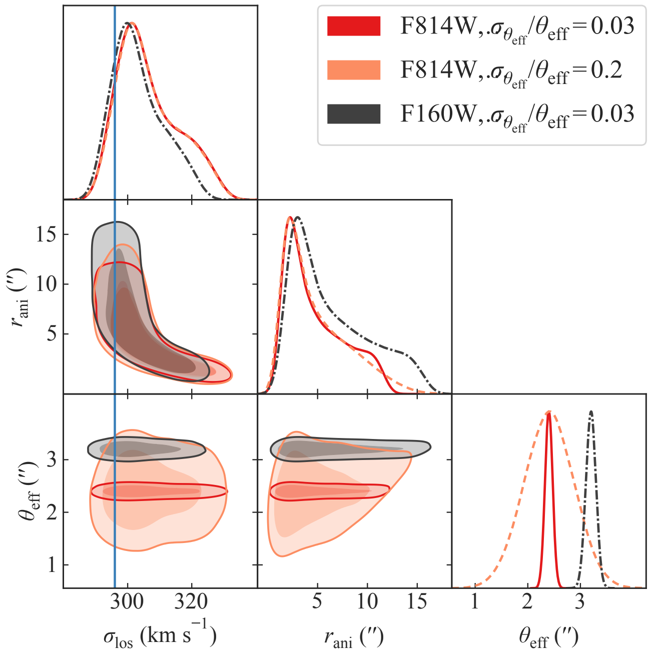

Fig. 17. Checking for systematics in the kinematic computation. Our primary settings for kinematic computation adopts the $F814W$ light profile with 3% uncertainty in $\theta_{\rm eff}$ (red). The orange contours are for the case with 20% uncertainty in $\theta_{\rm eff}$, and the black contours are for the case with $F160W$ light adopted in the kinematic computation. The vertical blue line marks the measured LOS velocity dispersion, which has an uncertainty of 19 km s^{-1}. The difference in the computed velocity dispersion between these cases is negligible, and it impacts the Fermat potential differences by $\lesssim$0.1%.

composite model providing a slightly higher likelihood value (Table 2). The discrepancy between the two model families are caused by the NFW profile normalization in the composite model, which makes the logarithmic slope of the density profile shallower at the Einstein radius (Fig. 13). We check for any potential unphysical properties in the mass profile posterior for the composite model (see Sect. 6.4.1). The size of the central core observed in stellar mass profile is consistent with previous observations (e.g. Bonfini & Graham 2016; Dullo 2019). The halo properties (e.g. the M−c relation, the total baryonic fraction, and the dark matter fraction within the Einstein radius) are consistent with previous observations of galaxy properties and cosmology. However, the predicted velocity dispersion profile (Fig. 14) from the composite mass model shows a decrease towards the centre, which has not been observed in local massive elliptical galaxies (e.g. Cappellari 2016; Ene et al. 2019), thus pointing to a potential inconsistency in the composite profile. We tested with additional physically motivated priors for the halo mass profile; however, that only amplified the discrepancy further. Alternatively, the discrepancy between the power-law and composite mass profiles can be reconciled by including an ultra-massive black hole ($M_{\rm BH}$ ∼ 3.8 × 10^{10} $M_{\odot}$), or by incorporating a stellar mass-to-light ratio gradient with exponent η ∼ 0.27, or a combination of both. These values are plausible based on previous observations (e.g. Sonnenfeld et al. 2018; Mehrgan et al. 2019; Dullo 2019).

Furthermore, the predicted central velocity dispersion from the composite model is inconsistent with the observed one. As a result, when no internal MSD is assumed (i.e. $\lambda_{\rm int}$ = 1), the kinematics likelihood largely excludes the posterior from the composite model in the final combined posterior. As a result, the final combined posteriors from LENSTRONOMY is almost entirely contributed by the power-law model.

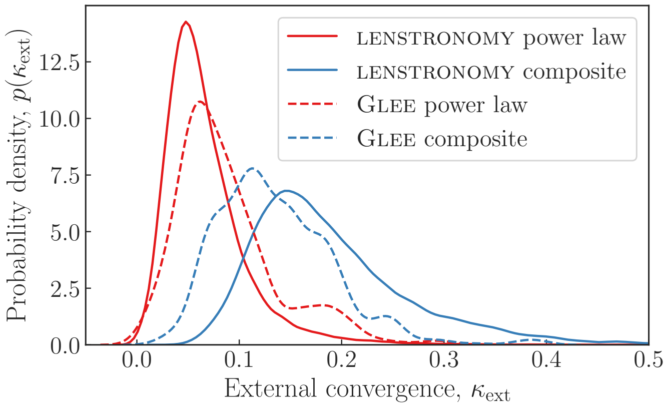

Fig. 18. External convergence distribution from Buckley-Geer et al. (2020) with additional weighting applied based on the predicted external shears for the power-law (red lines) and composite (blue lines) models from LENSTRONOMY (solid lines) and GLEE (dashed lines). Each illustrated GLEE distribution is a BIC-weighted combination of multiple $\kappa_{\rm ext}$ distributions corresponding to the external shear constraint from individual lens model setups. In contrast, each illustrated LENSTRONOMY distribution is a single $\kappa_{\rm ext}$ distribution corresponding to the combined (through BIC weighting) external shear value from all the model setups within a model family (i.e. power law or composite). The $\kappa_{\rm ext}$ distributions used by one team were not revealed to the other team before the un-blinding to maintain independence.

7. Comparison of the two software programs

The lens model posteriors from both teams were un-blinded on October 22, 2021, and no further modification to the lens models was performed afterwards. We only performed tests to investigate the differences or the lack thereof between the two modelling teams; the final time-delay predictions are kept frozen at the values during un-blinding. Table 3 compares the model parameters, derived quantities, predicted time delays between GLEE and LENSTRONOMY. We compare the un-blinded time-delay predictions from the combination of lensing, kinematics, and LOS analyses in Sect. 7.1. Then, we compare the lens model parameters and Fermat potential differences from lens modelling only in Sect. 7.2. We compare the pixelized PSF reconstructions between the software programs in Sect. 7.3 and the computational requirements in Sect. 7.4. Finally, we discuss our findings in Sect. 7.5.

7.1. Predicted time delays

We illustrate the final predicted time delay from both teams as a function of H_0 in Fig. 21, assuming a flat ΛCDM cosmology with $\Omega_{\rm m}$ = 0.3. The predictions for all image pairs are consistent with each other within ∼1σ. We further compare the time delay predictions for combined, power-law-only, and composite-only cases in Fig. 22. The combined time-delay posteriors differ the largest for the AB image pair by 11% (1.2σ). We note that the GLEE team applied kinematics weighting to the lens model posteriors only within the mass families and then combined the mass families with equal weighting, whereas the LENSTRONOMY team weighted the mass families by kinematics. As a result, the combined posteriors from the GLEE team

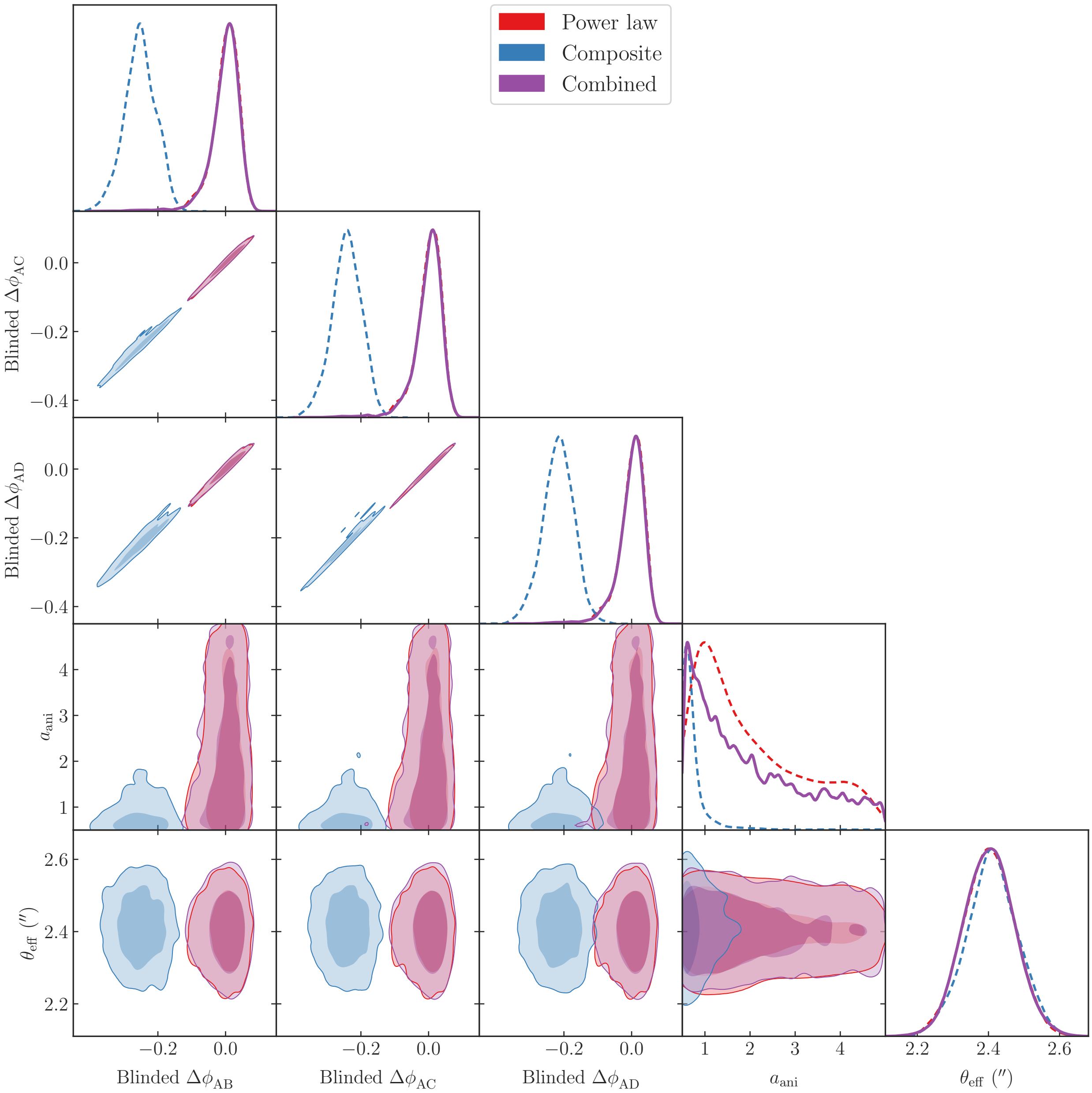

Fig. 19. Fermat potential differences $\Delta\phi = \Delta\phi^{\rm model}(1 - \kappa_{\rm ext})$ from LENSTRONOMY for the power-law (red) and composite (blue) model with the kinematics information folded into these individual model posteriors. The folding in of the kinematics information is performed through importance sampling from the posterior weighted by the kinematics likelihood (Eq. (41)). Next, the two model posteriors are joined together with equal weights and then the kinematics information is folded in. The combined posterior (purple) mostly resemble the power-law posterior, as the composite posterior is heavily down-weighted by the kinematics likelihood.

receives equal contribution from both model families giving rise to bi-modalities, where the combined posterior from the LENSTRONOMY team is almost entirely contributed by the power-law model posterior. The power-law-model-predicted time delays agree better between the two teams, with the largest difference appearing for AB image pair by 3.4% (0.6σ). The composite model predictions are more discrepant, with the largest difference appearing for AC image pair by 15% (2.1σ).

7.2. Model parameters and Fermat potential differences

The astrometric uncertainty on the constrained AGN image positions are consistent between the two teams, as both teams constrain the image positions with uncertainty $0\farcs004$ at maximum. This astrometric precision satisfies the requirement for cosmographic measurements (Birrer & Treu 2019).

In Fig. 23 we compare the lens model parameter posteriors and the predicted Fermat potential posteriors from the power-law lens models from both teams. The power-law exponent constrained by the LENSTRONOMY model is $\gamma = 2.21 \pm 0.02$, whereas that constrained by the GLEE model is $\gamma = 2.30 \pm 0.01$, which is a discrepancy at 4σ. The external shear magnitude constrained by the LENSTRONOMY model is $\gamma_{\rm ext} = 0.065 \pm 0.004$, and that by the GLEE model is $\gamma_{\rm ext} = 0.078 \pm 0.001$, which is at a 3.4σ discrepancy. We identify a degeneracy between γ and $\gamma_{\rm ext}$ internal to both models, and the discrepancy between the posteriors from the two models lie along this degeneracy (Fig. 23). The external shear magnitudes are typical for quadruply lensed quasar systems

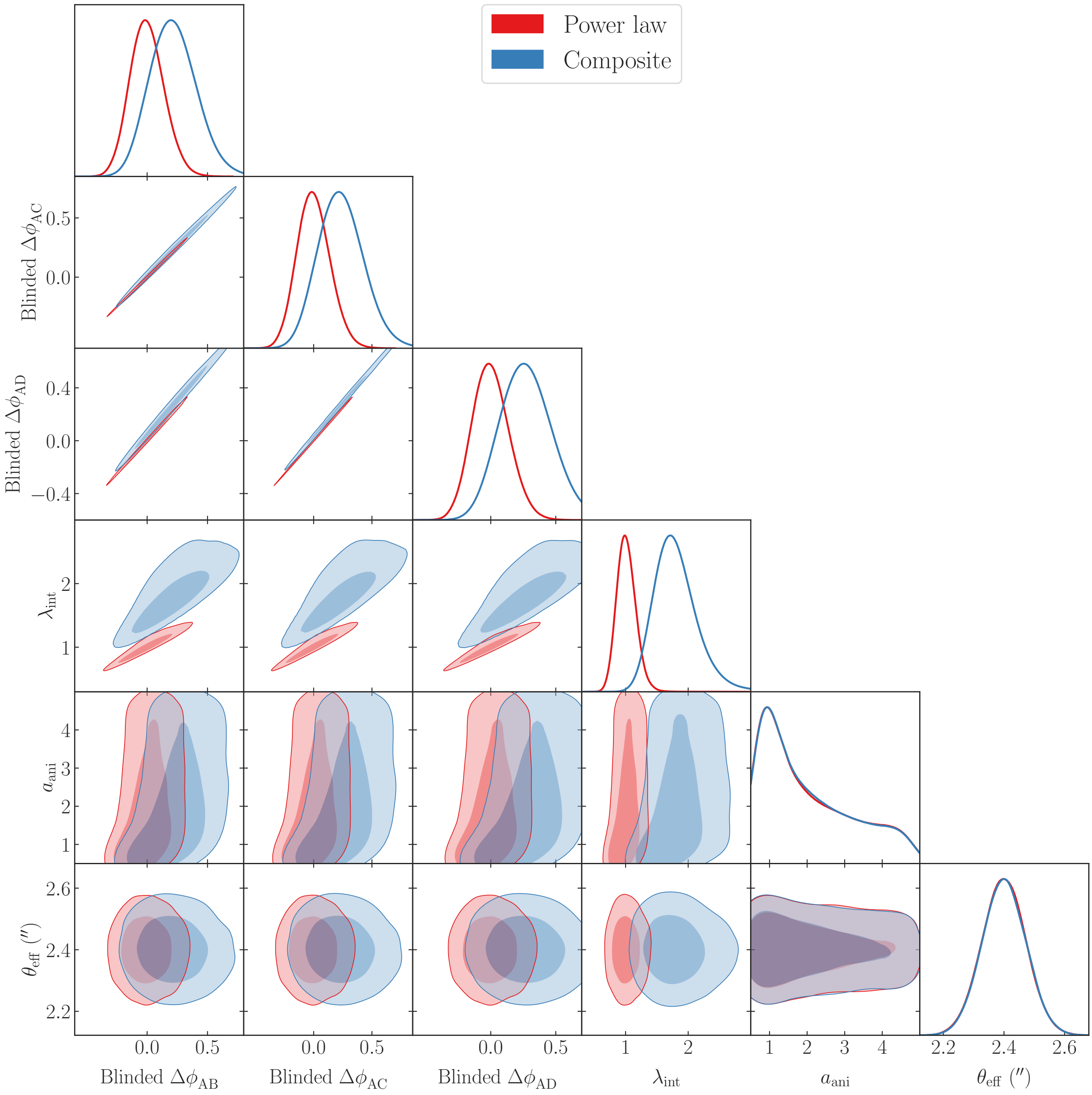

Fig. 20. MST-adjusted Fermat potential differences $\Delta\phi = \Delta\phi^{\mathrm{model}}\lambda_{\mathrm{int}}(1-\kappa_{\mathrm{ext}})$ from LENSTRONOMY for the power-law (red) and composite (blue) model families. The internal MST parameter λ_{int} is estimated by combining the lens models with the measured velocity dispersion and the external convergence estimate. After adjusting for the MST, the power-law and composite model predictions for the Fermat potential differences become consistent with each other within 1σ.

(e.g. see Schmidt et al. 2022). We investigated for deviation from the simple ellipticity description in our mass models as a potential source of the external shear. We find that boxy-ness or discy-ness in the luminous component is negligible within the Einstein radius (i.e. $\sqrt{a_4^2 + b_4^2} \lesssim 0.005$), which implies that allowing boxy-ness or discy-ness in the description of the mass profile is not required (for definitions of a_4 and b_4, and their impact on H_0 measurement, see Van de Vyvere et al. 2022a). As a result, we attribute the LOS galaxies around the central deflector to be the main source of the external shear, with additional potential contribution from the mild isophotal twist beyond the Einstein radius in the central deflector (Van de Vyvere et al. 2022b).

Next in Fig. 24, we compare the Fermat potential differences only from the power-law models of both teams without adjusting for the external convergence (top row) and with adjustment for the external convergence (bottom row). The Fermat potential differences from the lens model are discrepant, for example by 5.5σ (8.9%) for the AD image pair that has the longest predicted time delay. However, after combining the corresponding external convergence – based on selection cuts using the best fit external shear from each model – the Fermat potential differences all become consistent within 1σ, for example by 0.26σ (1%) for the AD image pair. Interestingly, the positive correlation between γ and γ_{ext} allows the estimated κ_{ext} selected on γ_{ext} to bring the time-delay posteriors closer, as higher γ_{ext} selects higher κ_{ext}.

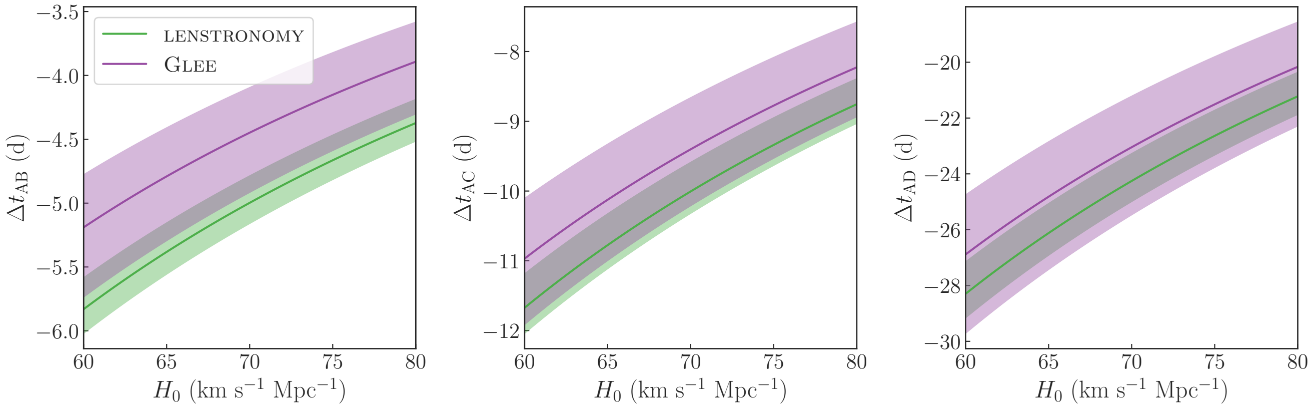

Fig. 21. Comparison between the two modelling teams for the predicted time delays (un-blinded) as a function of H_0 for the three image pairs, in flat ΛCDM cosmology with $\Omega_{\rm m} = 0.3$. Each posterior is the final combined posterior from the two mass-model setups: power-law and composite, including external convergence and stellar kinematics, and assuming $\lambda_{\rm int} = 1$.

Table 3. Comparison of GLEE and LENSTRONOMY model parameters and derived quantities.

Parameter	GLEE constraints	LENSTRONOMY constraints
	Power-law ellipsoid model	
$\theta_{\rm E}$ ($''$) $^{(a)}$	$1.379^{+0.001}_{-0.001}$	$1.380^{+0.001}_{-0.001}$
$q_{\rm m}$	$0.610^{+0.005}_{-0.005}$	$0.643^{+0.005}_{-0.005}$
$\varphi_{\rm m}$ (°)	$36.8^{+0.3}_{-0.3}$	$37.1^{+0.2}_{-0.2}$
γ	$2.30^{+0.01}_{-0.01}$	$2.22^{+0.02}_{-0.03}$
$\gamma_{\rm ext}$	$0.078^{+0.001}_{-0.002}$	$0.065^{+0.003}_{-0.004}$
$\varphi_{\rm ext}$ (°)	$-57.0^{+0.3}_{-0.3}$	$-58.1^{+0.3}_{-0.4}$
	Composite model	
Stellar M/L ($M_\odot/L_\odot$) $^{(b)}$	$6.3^{+0.1}_{-0.1}$	$2.30^{+0.06}_{-0.20}$
NFW $\kappa_{0,\rm h}$	$0.015^{+0.003}_{-0.008}$	$0.16^{+0.04}_{-0.01}$
NFW $r_{\rm s}$ ($''$)	$19.3^{+1.2}_{-1.2}$	$22.8^{+2.6}_{-3.5}$
NFW $q_{\rm m}$	$0.85^{+0.01}_{-0.01}$	$0.76^{+0.07}_{-0.04}$
NFW $\varphi_{\rm m}$ (°)	$24.8^{+1.7}_{-1.3}$	$-54.2^{+2.3}_{-3.4}$
$\gamma_{\rm ext}$	$0.101^{+0.002}_{-0.001}$	$0.128^{+0.005}_{-0.008}$
$\varphi_{\rm ext}$ (°)	$-57.3^{+0.2}_{-0.2}$	$-55.3^{+0.9}_{-1.4}$
Predicted time delays from power-law and composite models combined $^{(c)}$		
$\Delta t_{\rm AB}$ (d)	$-4.4^{+0.4}_{-0.5}$	$-5.0^{+0.2}_{-0.2}$
$\Delta t_{\rm AC}$ (d)	$-9.4^{+0.7}_{-0.8}$	$-10.0^{+0.4}_{-0.3}$
$\Delta t_{\rm AD}$ (d)	$-23.0^{+1.8}_{-2.4}$	$-24.2^{+1.0}_{-0.7}$

Notes. Reported values are medians, with errors corresponding to the 16th and 84th percentiles. Angles are measured east of north. $^{(a)}$Spherical-equivalent Einstein radius. $^{(b)}$$M/L$ for rest-frame V band. The given uncertainties are statistical uncertainties only. The stellar mass is calculated assuming $H_0 = 70\,{\rm km\,s^{-1}\,Mpc^{-1}}$, $\Omega_{\rm m} = 0.3$, and $\Omega_\Lambda = 0.7$, but changes in the cosmology affect the M/L by a negligible amount. $^{(c)}$Assuming a flat ΛCDM cosmology with $H_0 = 70\,{\rm km\,s^{-1}\,Mpc^{-1}}$, $\Omega_{\rm m} = 0.3$, and $\Omega_\Lambda = 0.7$.

However, the strength of the positive correlation between γ and $\gamma_{\rm ext}$ depends on the particular morphology of the quad lenses (Shajib et al. 2019). Thus, we cannot conclude if this effect – that the $\gamma_{\rm ext}$-selected $\kappa_{\rm ext}$ brings the time delay posteriors more into agreement – applies to all lensing systems and is not just a particular occurrence for the lens system WGD 2038−4008. A detailed analysis of a larger sample of lenses is required to reach a conclusion on this matter, and it is left for future work.

7.3. Reconstructed PSFs

The reconstructed pixelized PSFs by the GLEE modelling procedure have smaller FWHM by ∼2–7% than the ones from LENSTRONOMY: 1.7% in $F160W$, 3.9% in $F814W$, and 6.7% in $F475X$ (Fig. 25). Furthermore, the $F160W$ PSF in GLEE is supersampled with a supersampling factor of 3, whereas the $F160W$ PSF in LENSTRONOMY has the same pixel resolution ($0\farcs08$) as the drizzled image. The PSFs for the UVIS filters have the same pixel resolution as the drizzled image ($0\farcs04$) for both GLEE and LENSTRONOMY.

We test how the differences in the reconstructed PSFs contribute to the differences in the logarithmic slope parameter for the power-law model between the two software programs. The test results are illustrated in Fig. 26. In these tests, we change the adopted PSFs and optimize a fiducial lens model from each software program. For LENSTRONOMY, the fiducial model is the power-law model with the highest BIC score (Table 2). For GLEE, the fiducial model is the 'power-law fiducial model' (Table 1). We first interchanged all the PSFs between the software programs. Due to numerical requirements, the GLEE team artificially supersample the LENSTRONOMY-reconstructed PSF through interpolation for this and subsequent tests. With the PSFs interchanged, the power-law slope parameter γ constrained by one software program shifts towards the fiducial constraint from other software program. These shifts bring the γ distributions within ∼1.2–1.6σ consistency between the two software programs given the same PSFs, although still leaving some unexplained deviations.

We further interchange the weighted uncertainty maps in addition to the reconstructed PSFs between the software programs. Originally, the GLEE team creates a weighted uncertainty map by boosting the noise levels around the quasar positions (see Sect. 5), whereas the LENSTRONOMY team adds the PSF uncertainty map of the initial PSF estimate in quadrature with the data uncertainty map at the positions of the quasars. In this test, the resultant γ distributions become slightly more consistent within ∼1.0–1.5σ compared to the previous test. Therefore, the particular choice or method to estimate the weighted uncertainty maps does not significantly contribute to the deviation in the power-law γ parameter distribution from the two software programs.

In the next test, we constrained the lens models from UVIS data only with interchanged PSFs. In this test, the GLEE constraint on γ remained stable; however, the LENSTRONOMY constraint shifted towards the GLEE constraint to be consistent within

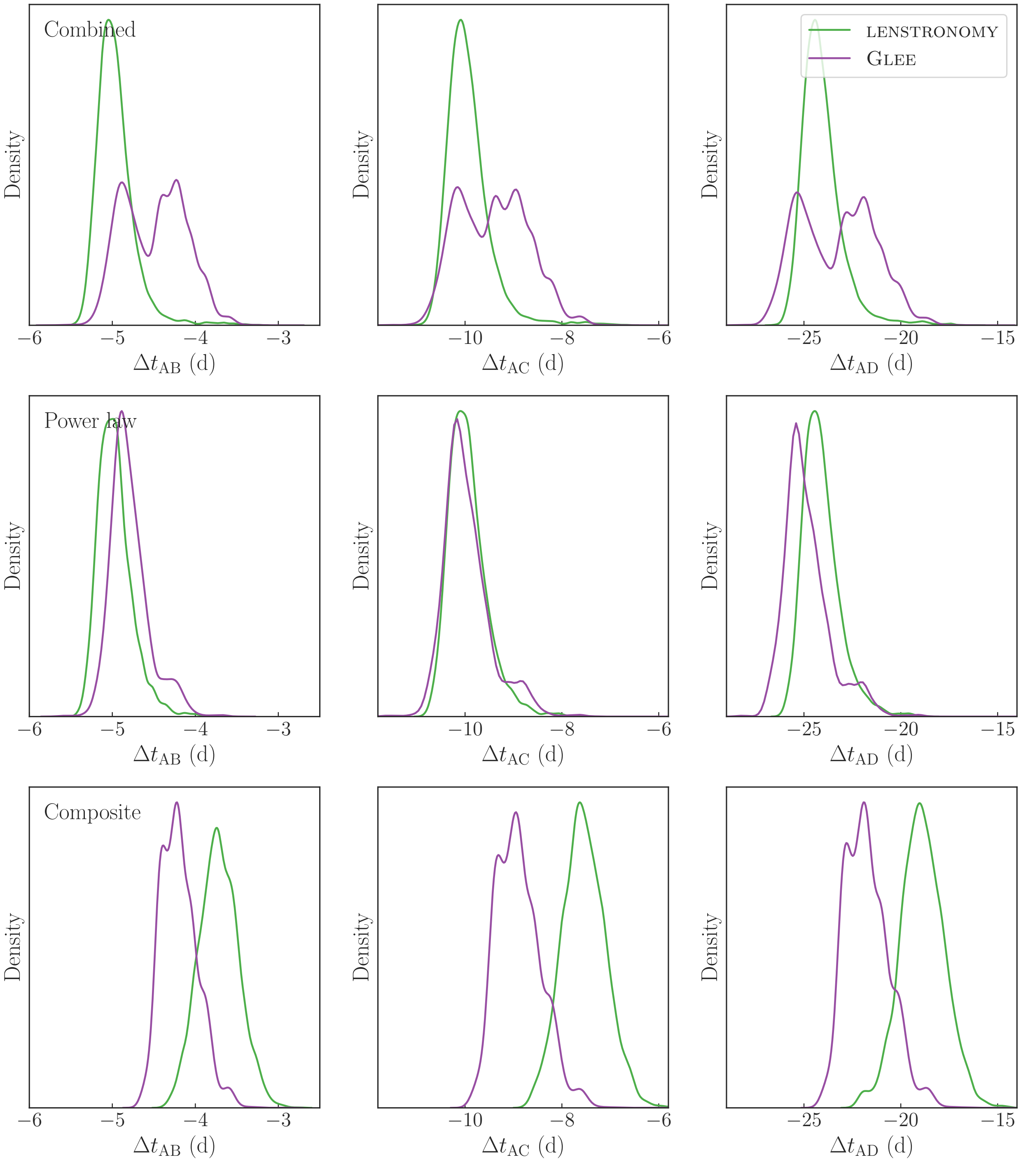

Fig. 22. Comparison of predicted time delays between the LENSTRONOMY (green) and GLEE (purple) teams. The three rows show the comparison for combined (*top*), power-law-only (*middle*), and composite-only (*bottom*) cases. The combined posteriors are consistent within $<1.2\sigma$, the power-law model's posteriors are consistent within $<0.6\sigma$, and the composite model's posteriors are consistent within $<2.1\sigma$. In percentage, the maximum deviation for the power-law model is between AB image pair by 3.4%, and for the composite model is between AC image pair by 15%. The GLEE team combined the power-law and composite model posteriors with equal weighting after applying the kinematics weighting within each model family, whereas the LENSTRONOMY team combined the two model families with kinematics weighting, leading to the final combined posterior being dominated by the power-law model.

$\sim 0.5\sigma$. As a result, we conclude that the discrepancy in the γ distribution between LENSTRONOMY and GLEE is dominated by the difference in the IR PSF.

For the last test, we add a new feature in LENSTRONOMY to reconstruct the IR PSF with a supersampled resolution and reconstruct the PSF with a supersampling factor of 3 following the GLEE team. If this supersampled IR PSF from LENSTRONOMY is used by both software programs, then the resultant γ distribution becomes consistent within $\sim 0.6\sigma$. However, the γ distribution of the GLEE model with its own supersampled PSF still differs by 2.6σ from that of the LENSTRONOMY model with its own supersampled PSF. It is not possible to

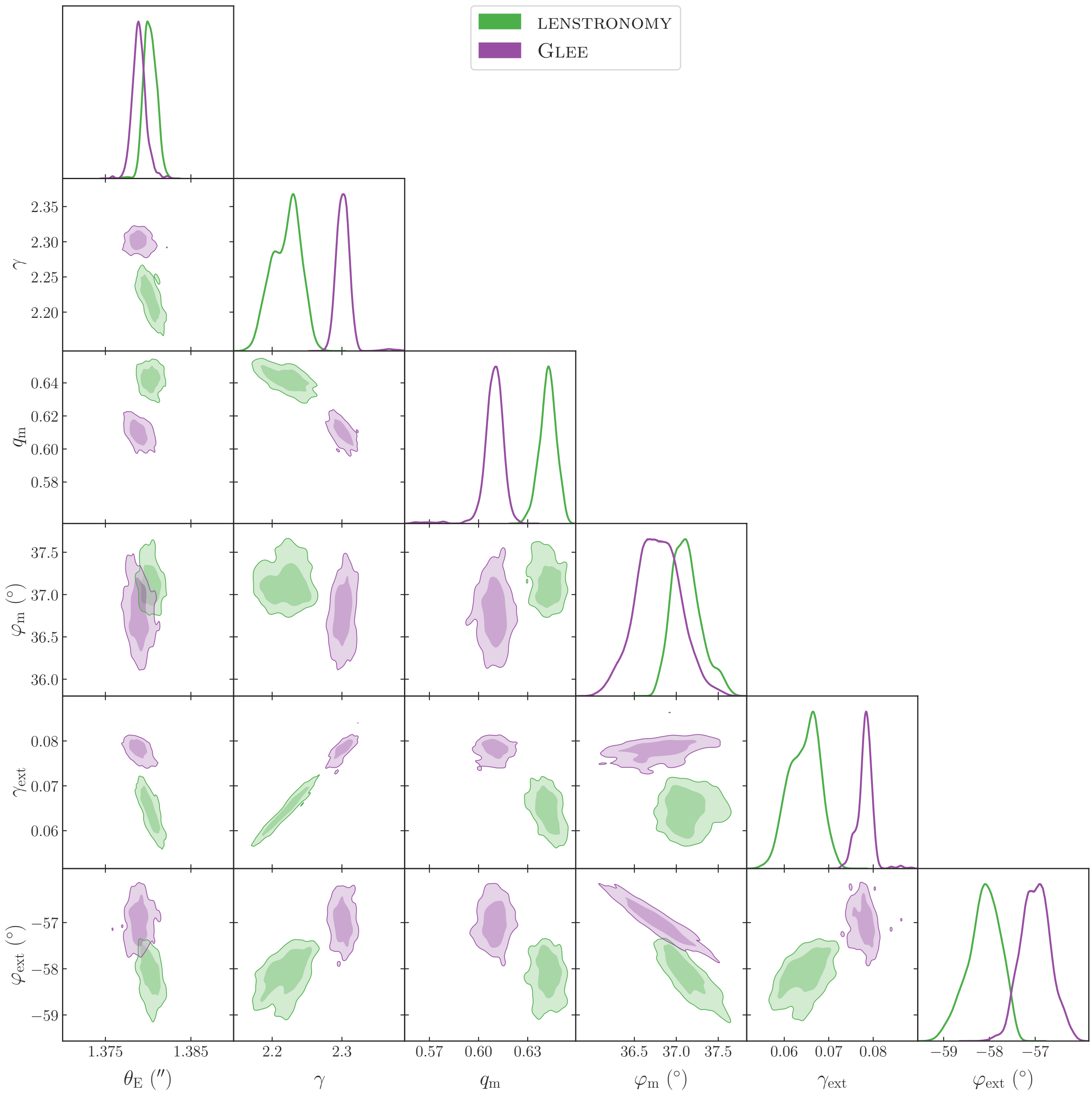

Fig. 23. Comparison of lens model parameter differences for the power-law model between the LENSTRONOMY (green) and the GLEE (purple) teams.

evaluate which reconstructed PSF is more accurate a priori. Therefore, it is recommended to marginalize over multiple PSF reconstructions to account for the stochasticity within one particular reconstruction algorithm and as well as different reconstruction algorithms. In addition, supersampled PSFs are recommended especially for the IR band with large pixels; the subsampling factor can be set to the minimal value to produce stable results while keeping computational time low.

7.4. Requirements for computational resources

The entire LENSTRONOMY modelling procedure required ~4×10^5 CPU hours including initial modelling trials, running full MCMC chains of the adopted models, post-processing of posteriors, and post-un-blinding tests. The estimated usage for GLEE models are $O(10^5)$ CPU hours only to run the final models and MCMC chains to produce the final un-blinded result. However, the total usage of CPU hours can be $O(10^6)$ CPU hours including modelling trials, robustness tests, and reruns of chains to recover lost progress due to numerical issues.

7.5. Discussion

The final un-blinded time-delay predictions agree within $<1.2\sigma$ between the two modelling teams. As a result, the inferred Hubble constants from the two teams based on the observed time delays will be consistent within $<1.2\sigma$. However, the predictions from the composite models only are less in agreement between the

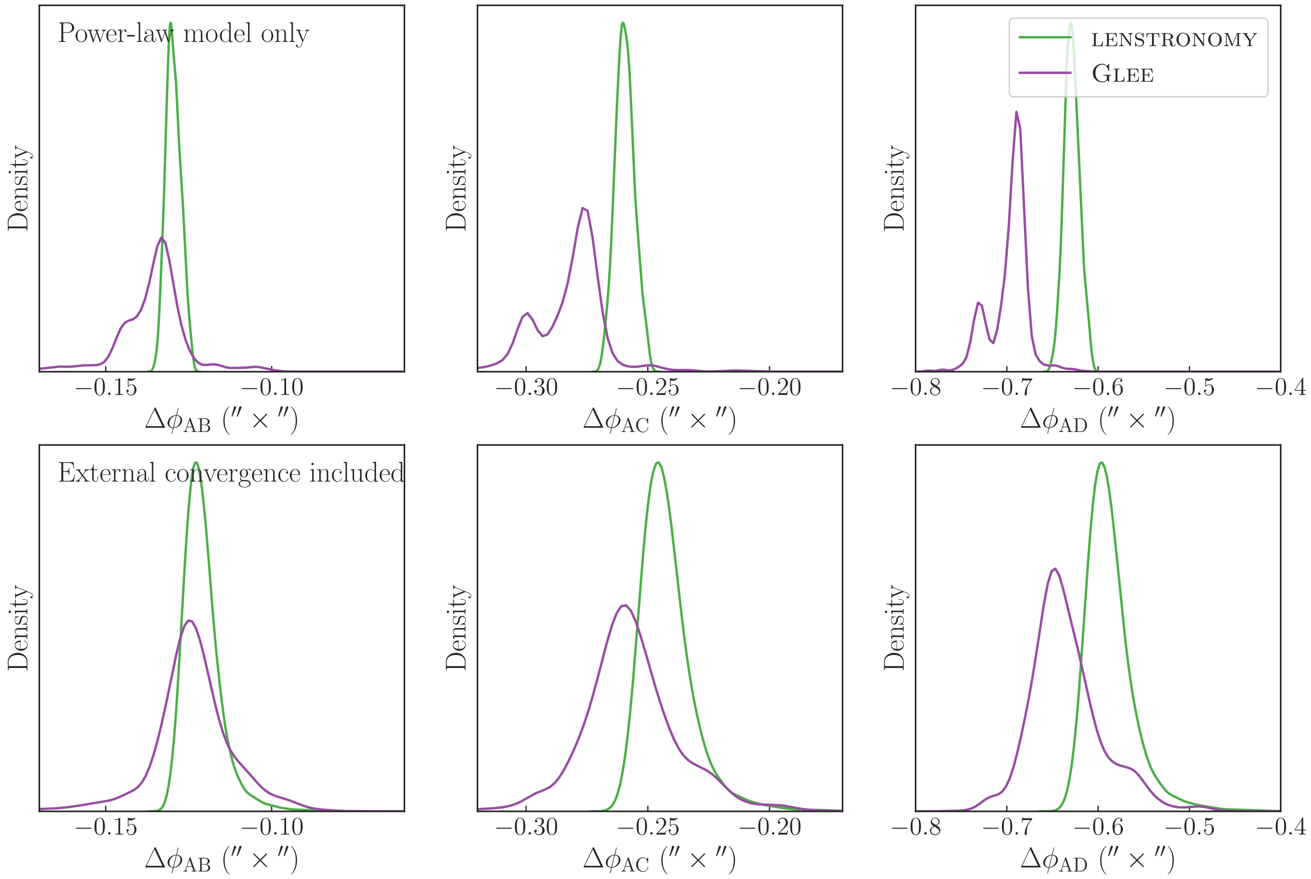

Fig. 24. Comparison of Fermat potential differences for the power-law model between the LENSTRONOMY (green) and the GLEE (purple) teams, without including the external convergence (*top row*), and with including the external convergence (*bottom row*).

two teams. Interestingly, the composite-model predictions deviate from the power-law ones towards the same direction for both teams. We were already aware prior to the un-blinding that the composite model is atypical with respect to previous observations, for example the velocity dispersion profile significantly decreases towards the centre for the one from the LENSTRONOMY (Fig. 14), the one from the GLEE team has a very low inner dark matter fraction. Although such discrepancies between composite and power-law model predictions have been observed in the previously analysed systems (e.g. Suyu et al. 2014), this system WGD 2038−4008 demonstrates the largest discrepancy to date out of the systems analysed by H0LiCOW/TDCOSMO. However, unlike the previously analysed systems, the Einstein radius of this system encompasses only the very central region of the very extended lens galaxy, and thus the imaging observables probe a different region of the elliptical galaxy: at $\sim\theta_{\rm eff}/3$ instead of $\sim\theta_{\rm eff}$. This discrepancy in the composite model illustrates that this model is not an adequate description for the mass distribution at the central region of all elliptical galaxies. Rather an appropriate combination of a mass-to-light ratio gradient and a supermassive black hole can be necessary to sufficiently describe the mass distribution at the scales considered here. In such cases when the different mass model families fit the imaging observables equally well, but lead to different predictions in the Fermat potential and kinematics, the observed kinematics is crucial to act as the differentiator between the mass model families through appropriate weighting of the kinematics likelihood. In the future, spatially resolved velocity dispersion from integral field spectra, for example from the Multi Unit Spectroscopic Explorer (MUSE) on the VLT, will be able to constrain such an improved composite model with additional degrees of freedom allowed.

For the composite model setup, the modelling teams had more freedom in choosing particular priors and model settings, allowing for discrepancies between the results from the two teams. However, the power-law models are specified with less room for independent choices to be made by the modelling teams. Therefore, we compare the results from the power-law models between the two teams to identify systematic differences at the level of the software packages.

In particular, we focus on the power-law logarithmic slope parameter γ, as this parameter is the most sensitive lens model parameter to the predicted time delays and thus the inferred Hubble constant. The γ distributions between the modelling teams are discrepant at 4σ. We identify the difference in the reconstructed PSFs, especially in the IR band, to be the dominant source of this discrepancy. Given the same supersampled PSF in the IR band and non-supersampled PSFs in the UVIS bands, both modelling softwares produce γ constraints with differences below 0.5%. The γ distributions from the two modelling software programs are not expected to be identical due to differences in the numeric implementation, for example the likelihood computation region and the source reconstruction method. Thus, we can conclude that the systematic differences in the model fitting part of the software programs are below 0.5% in γ corresponding to ~1% on H_0. However, the PSF reconstruction method can lead to systematic differences as large as ~4%. This difference is reconciled in the time-delay predictions from the two teams after combining the lens model posteriors with the kinematics

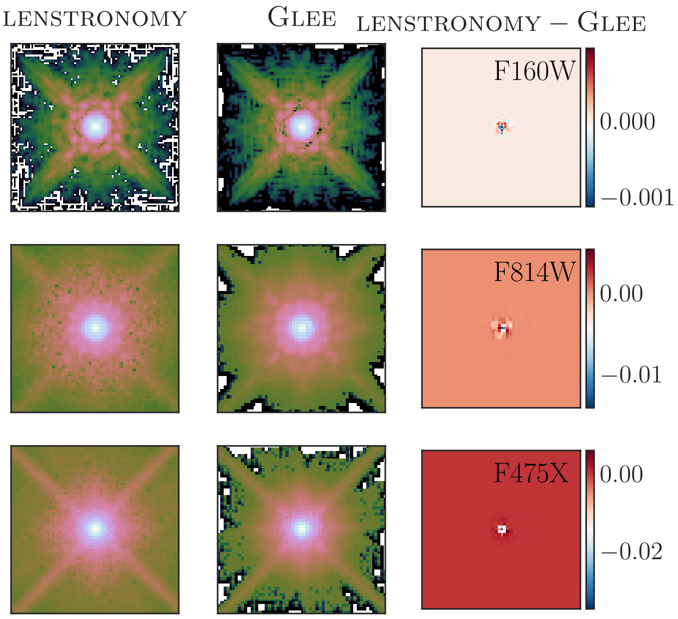

Fig. 25. Comparison of the reconstructed pixelized PSFs by the LENSTRONOMY (*first column*) and the GLEE (*second column*) modelling procedures. *Third column*: illustrates the difference between the LENSTRONOMY and GLEE PSFs. The three rows correspond to the *F*160*W*, *F*814*W*, and *F*475*X* filters from top to bottom. The *F*160*W* PSF is supersampled with a supersampling factor of 3. We note that the illustrated supersampled LENSTRONOMY PSF was not used in the pre-un-blinding models and was reconstructed after the un-blinding to perform further tests. The original reconstructed PSF in LENSTRONOMY in *F*160*W* was not supersampled. The GLEE PSF FWHMs are smaller by ~2−7% than the LENSTRONOMY ones.

data and the estimated external convergence for the particular system WGD 2038−4008. However, it is inconclusive if such differences can be similarly reconciled for all lens systems in general, and we leave this investigation with a larger lens sample for a future study. If such differences cannot be reconciled for other lens systems and these differences do not average out when a sample of lenses is considered, then these differences would be non-negligible in the long run when a large sample of lens systems are combined to infer the Hubble constant. As it is currently not possible to evaluate the appropriateness of one reconstructed PSF over the other, we recommend to marginalize over different realizations of the PSF reconstruction and also over different reconstruction algorithms to avoid any potential bias. Furthermore, a supersampled PSF in the IR band is also recommended, as the drizzled pixel scale of 0.″08 in the IR does not optimally sample the PSF.

8. Summary and conclusion

In this study two teams independently modelled the lens system WGD 2038−4008 from three-band HST imaging using two different software programs: LENSTRONOMY and GLEE. Two families of models were specified as baseline models before the modelling was performed – one family with a power-law ellipsoidal mass distribution and the other family with a two-component mass distribution that accounts for the dark and baryonic components separately. The baseline settings were pre-specified to allow a fair comparison between the modelling results. Individual teams were allowed to improve upon the baseline settings as they deemed appropriate, for example the choice of priors and

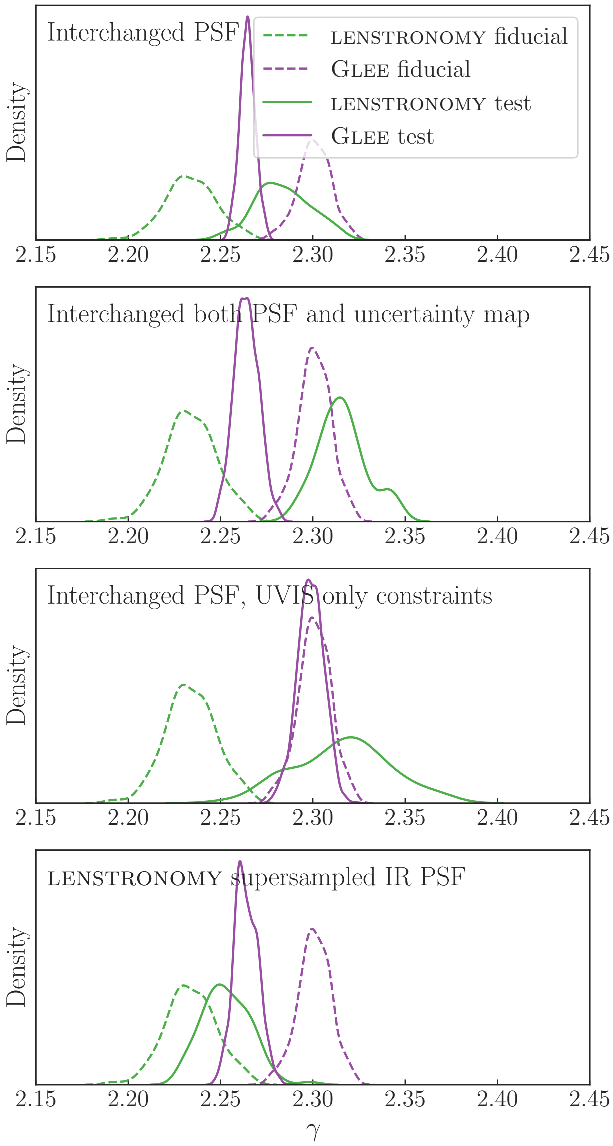

Fig. 26. Deviations in the logarithmic slope γ of the power-law model with different PSF settings. In all panels, the dashed distributions show the fiducial constraints – LENSTRONOMY in green and GLEE in purple. The LENSTRONOMY fiducial constraint is from the highest BIC value power-law model (Table 2), and the GLEE fiducial constraint is from the ‘power-law fiducial’ setup (Table 1). *First panel*: when only the reconstructed PSFs are interchanged to optimize the models, the constraint from one software program moves towards the other’s fiducial constraint to be consistent within 1.2−1.6σ. We note that GLEE model artificially creates supersampled version of the LENSTRONOMY PSF through interpolation to perform the tests here. *Second panel*: when both the PSFs and weighted uncertainty maps are interchanged, the resultant γ becomes slightly more consistent between the software programs within 1.0−1.5σ. Therefore, we conclude that the differences in the PSF itself significantly contributes to the discrepancy in the power-law γ constraint and not the particular method of weighting the uncertainty map to account for PSF uncertainty. *Third panel*: when models from both of the software programs are optimized only with UVIS data and interchanged PSFs, the GLEE constraint does not shift significantly; however, the LENSTRONOMY constraint shifts significantly towards the GLEE constraints. This result indicates that the difference between the GLEE and LENSTRONOMY fiducial models are largely created by the difference in the IR PSF. *Fourth panel*: when both software programs use a supersampled PSF with a supersampling factor of 3 reconstructed by LENSTRONOMY, the resultant γ constraint agrees very well, within 0.6σ.

the numerical settings specific to the software program being used. The two modelling procedures were carried out blindly with regards to the other team. The models were un-blinded on October 22, 2021, after an internal review by scientists from the TDCOSMO collaboration not directly involved with either modelling team, and no further modifications to the lens models were performed. The predicted Fermat potential differences from both teams were combined with the observed kinematics data and estimated external convergence to predict the time delays between the three image pairs. A future study will infer the Hubble constant by comparing the predicted time delays with the observed ones from ongoing monitoring campaigns. We investigated the observed systematic differences between the model outputs from the two teams and identify that differences in the reconstructed PSF are the dominant source of systematic differences. The main results of this study are as follows:

- The final predicted time delays from LENSTRONOMY are: $\Delta t_{\rm AB} = -5.0^{+0.2}_{-0.2}$ d, $\Delta t_{\rm AC} = -10.0^{+0.4}_{-0.3}$ d, and $\Delta t_{\rm AD} = -24.2^{+1.0}_{-0.7}$ d; and the ones from GLEE are: $\Delta t_{\rm AB} = -4.4^{+0.4}_{-0.5}$ d, $\Delta t_{\rm AC} = -9.4^{+0.7}_{-0.8}$ d, and $\Delta t_{\rm AD} = -23.0^{+1.8}_{-2.4}$ d. These values assume a flat ΛCDM cosmology with $H_0 = 70\,{\rm km\,s^{-1}\,Mpc^{-1}}$ and $\Omega_{\rm m} = 0.3$. The negative value of $\Delta t_{\rm AB}$, for example, signifies that image B lags image A. This system is currently being monitored under the COSMOGRAIL programme to measure the time delays (Eigenbrod et al. 2005). Once the time delays are measured, the mutual agreement between the predicted values will result in a mutually consistent inference of the Hubble constant.
- The logarithmic slope, γ, and the external shear, $\gamma_{\rm ext}$, of the power-law model deviate by 4σ (3.9%) between the LENSTRONOMY and GLEE models. This discrepancy is predominantly created by the difference in the reconstructed pixelized PSFs. When the same PSF is used by both modelling programs, then the resultant γ and $\gamma_{\rm ext}$ distributions agree within $\sim 0.6\sigma$ (∼0.5%), which is compatible with the ∼1% precision goal in the Hubble constant measurement from a large sample of ∼40 quad lenses (e.g. Shajib et al. 2018; Birrer & Treu 2021). The particular method of weighting the uncertainty map to account for the PSF uncertainty is non-dominant in this discrepancy.
- Our composite model posteriors are not generally in good agreement with the one from the power-law model, and the discrepancy is more prominent for LENSTRONOMY models due to the adoption of more stringent physical priors on the halo mass. This inconsistency points to the inadequacy of our composite model in describing the mass distribution for this particular lens system, WGD 2038−4008; WGD 2038−4008 is atypical compared to previously analysed TDCOSMO lenses in that the $\theta_{\rm E}/\theta_{\rm eff}$ is relatively small and thus the imaging information only probes the inner region of the deflector galaxy where the combination of NFW and mass-follows-light profiles is not an adequate model. We stress that for all the seven systems previously analysed by the TDCOSMO collaboration, the power-law and composite models were in excellent agreement. For such discrepancies between the power-law and composite models, additional data, such as the stellar kinematics, should be used to select the better model. The two models predict significantly different velocity dispersion profiles and will therefore be easily separable by spatially resolved kinematics.

In the context of the recent debate around the Hubble constant, it is paramount to thoroughly investigate for potential systematic biases in the measurement methods (e.g. Freedman 2021; Riess et al. 2022). This study performs one such crucial systematic check for time-delay cosmography to investigate the robustness of lens modelling software programs. By keeping the modelling systematics under control, future large samples of lensed quasars and SNe will robustly measure the Hubble constant to ≲1% precision (e.g. Jee et al. 2016; Birrer & Treu 2021; Birrer et al. 2022).

Acknowledgements. We thank Christopher D. Fassnacht, Veronica Motta, and Abigail Lee for useful comments and discussions that improved this manuscript. Support for this work was provided by NASA through the NASA Hubble Fellowship grant HST-HF2-51492 awarded to A.J.S. by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555. A.J.S., S.B., and T.T. were supported by the National Aeronautics and Space Administration (NASA) through the Space Telescope Science Institute (STScI) grant HST-GO-15320. A.J.S. was also supported by the Dissertation Year Fellowship from the University of California, Los Angeles (UCLA) graduate division. This research was supported by the US Department of Energy (DOE) Office of Science Distinguished Scientist Fellow Program. This work was supported by World Premier International Research Center Initiative (WPI Initiative), MEXT, Japan. This work was supported by JSPS KAKENHI Grant Number JP20K14511. S.H.S. thanks the Max Planck Society for support through the Max Planck Research Group. This research is supported in part by the Excellence Cluster ORIGINS which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2094 – 390783311. T.T. acknowledges support by the Packard Foundation through a Packard Research fellowship, by the National Science Foundation through NSF grants AST-1836016, AST-1906976, and by the Gordon and Betty Moore Foundation through grant 8548. AA's research is funded by Villum Experiment Grant Cosmic Beacons (project number 36225). This work was supported by the Swiss National Science Foundation (SNSF) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 787886). This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group. Additionally, this work was completed in part with resources provided by the University of Chicago's Research Computing Center. A.J.S. and S.H.S. acknowledge the hospitality of the Aspen Center of Physics (ACP) and the Munich Institute for Astro- and Particle Physics (MIAPP) of the Excellence Cluster "Universe", where part of this research was completed. A.C.P. is supported by National Science Foundation grant PHY-1607611. This research made use of GLEE (Suyu & Halkola 2010; Suyu 2012), LENSTRONOMY (Birrer et al. 2015; Birrer & Amara 2018; Birrer et al. 2021), FASTELL (Barkana 1999), NUMPY (Oliphant 2015), SCIPY (Jones et al. 2001), ASTROPY (Astropy Collaboration 2013, 2018), PHOTUTILS – an ASTROPY package for detection and photometry of astronomical sources (Bradley et al. 2020), JUPYTER (Kluyver et al. 2016), MATPLOTLIB (Hunter 2007), SEABORN (Waskom et al. 2014), SEXTRACTOR (Bertin & Arnouts 1996), EMCEE (Foreman-Mackey et al. 2013), COLOSSUS (Diemer 2018), GETDIST (https://github.com/cmbant/getdist), PYGALAXEV (Bruzual & Charlot 2003), and DYNESTY (Skilling 2004; Speagle 2020).

References

Agnello, A., Lin, H., Kuropatkin, N., et al. 2018, MNRAS, 479, 4345
Aiola, S., Calabrese, E., Maurin, L., et al. 2020, JCAP, 2020, 047
Astropy Collaboration (Robitaille, T. P., et al.) 2013, A&A, 558, A33
Astropy Collaboration (Price-Whelan, A. M., et al.) 2018, AJ, 156, 123
Barkana, R. 1998, ApJ, 502, 531
Barkana, R. 1999, Astrophysics Source Code Library [record ascl:9910.003]
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bertin, G., & Lombardi, M. 2006, ApJ, 648, L17
Birrer, S. 2021, ApJ, 919, 38
Birrer, S., & Amara, A. 2018, Phys. Dark Univ., 22, 189
Birrer, S., & Treu, T. 2019, MNRAS, 489, 2097
Birrer, S., & Treu, T. 2021, A&A, 649, A61
Birrer, S., Amara, A., & Refregier, A. 2015, ApJ, 813, 102
Birrer, S., Amara, A., & Refregier, A. 2016, JCAP, 8, 020
Birrer, S., Treu, T., Rusu, C. E., et al. 2019, MNRAS, 484, 4726
Birrer, S., Shajib, A. J., Galan, A., et al. 2020, A&A, 643, A165
Birrer, S., Shajib, A. J., Gilman, D., et al. 2021, J. Open Sour. Softw., 6, 3283
Birrer, S., Dhawan, S., & Shajib, A. J. 2022, ApJ, 924, 2
Blakeslee, J. P., Jensen, J. B., Ma, C.-P., Milne, P. A., & Greene, J. E. 2021, ApJ, 911, 65
Blandford, R. D., & Narayan, R. 1992, ARA&A, 30, 311

Bolton, A. S., Burles, S., Koopmans, L. V. E., Treu, T., & Moustakas, L. A. 2006, ApJ, 638, 703
Bonfini, P., & Graham, A. W. 2016, ApJ, 829, 81
Bonvin, V., Courbin, F., Suyu, S. H., et al. 2017, MNRAS, 465, 4914
Bradley, L., Sipőcz, B., Robitaille, T., et al. 2020, https://doi.org/10.5281/zenodo.4044744
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Buckley-Geer, E. J., Lin, H., Rusu, C. E., et al. 2020, MNRAS, 498, 3241
Cappellari, M. 2002, MNRAS, 333, 400
Cappellari, M. 2016, ARA&A, 54, 597
Chen, G. C.-F., Suyu, S. H., Wong, K. C., et al. 2016, MNRAS, 462, 3457
Chen, G. C. F., Fassnacht, C. D., Suyu, S. H., et al. 2019, MNRAS, 490, 1743
Chen, G. C.-F., Fassnacht, C. D., Suyu, S. H., et al. 2021, A&A, 652, A7
Diemer, B. 2018, ApJS, 239, 35
Diemer, B., & Joyce, M. 2019, ApJ, 871, 168
Ding, X., Treu, T., Birrer, S., et al. 2021, MNRAS, 503, 1096
Dobler, G., Fassnacht, C. D., Treu, T., et al. 2015, ApJ, 799, 168
Dullo, B. T. 2019, ApJ, 886, 80
Dutton, A. A., Brewer, B. J., Marshall, P. J., et al. 2011, MNRAS, 417, 1621
Efstathiou, G. 2021, MNRAS, 505, 3866
Eigenbrod, A., Courbin, F., Vuissoz, C., et al. 2005, A&A, 436, 25
Ene, I., Ma, C.-P., McConnell, N. J., et al. 2019, ApJ, 878, 57
Falco, E. E., Gorenstein, M. V., & Shapiro, I. I. 1985, ApJ, 289, L1
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Freedman, W. L. 2021, ApJ, 919, 16
Freedman, W. L., Madore, B. F., Hatt, D., et al. 2019, ApJ, 882, 34
Freedman, W. L., Madore, B. F., Hoyt, T., et al. 2020, ApJ, 891, 57
Gavazzi, R., Treu, T., Rhodes, J. D., et al. 2007, ApJ, 667, 176
Gilman, D., Birrer, S., & Treu, T. 2020, A&A, 642, A194
Golse, G., & Kneib, J.-P. 2002, A&A, 390, 821
Goodman, J., & Weare, J. 2010, Commun. Appl. Math. Comput. Sci., 5, 65
Greene, Z. S., Suyu, S. H., Treu, T., et al. 2013, ApJ, 768, 39
Hilbert, S., Hartlap, J., White, S. D. M., & Schneider, P. 2009, A&A, 499, 31
Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. 1999, Stat. Sci., 14, 382
Hunter, J. D. 2007, Comput. Sci. Eng., 9, 90
Jee, I., Komatsu, E., Suyu, S. H., & Huterer, D. 2016, JCAP, 4, 031
Jones, E., Oliphant, T., Peterson, P., et al. 2001, SciPy: Open Source Scientific Tools for Python, http://www.scipy.org
Kennedy, J., & Eberhart, R. 1995, Proceedings of ICNN'95 – International Conference on Neural Networks (IEEE)
Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in Positioning and Power in Academic Publishing: Players, Agents and Agendas, eds. F. Loizides, & B. Schmidt (Netherlands: IOS Press), 87
Knox, L., & Millea, M. 2020, Phys. Rev. D, 101, 043533
Kochanek, C. S. 2020, MNRAS, 493, 1725
Kourkchi, E., Tully, R. B., Eftekharzadeh, S., et al. 2020, ApJ, 902, 145
Lejeune, T., Cuisinier, F., & Buser, R. 1998, A&AS, 130, 65
Lewis, A., & Bridle, S. 2002, Phys. Rev. D, 66, 103511
Liao, K., Treu, T., Marshall, P., et al. 2015, ApJ, 800, 11
Madigan, D., & Raftery, A. E. 1994, J. Am. Stat. Assoc., 89, 1535
Mamon, G. A., & Łokas, E. L. 2005, MNRAS, 363, 705
Mehrgan, K., Thomas, J., Saglia, R., et al. 2019, ApJ, 887, 195
Melo, A., Motta, V., Godoy, N., et al. 2021, A&A, 656, A108
Merritt, D. 1985a, MNRAS, 214, 25P
Merritt, D. 1985b, AJ, 90, 1027
Millon, M., Galan, A., Courbin, F., et al. 2020, A&A, 639, A101
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493
Oliphant, T. E. 2015, Guide to NumPy, 2nd edn. (USA: CreateSpace Independent Publishing Platform)
Osipkov, L. P. 1979, Pisma v Astronomicheskii Zhurnal, 5, 77
Pesce, D. W., Braatz, J. A., Reid, M. J., et al. 2020, ApJ, 891, L1
Planck Collaboration VI. 2020, A&A, 641, A6
Poulin, V., Smith, T. L., Karwal, T., & Kamionkowski, M. 2019, Phys. Rev. Lett., 122, 221301
Refregier, A. 2003, MNRAS, 338, 35
Refsdal, S. 1964, MNRAS, 128, 307
Riess, A. G., Yuan, W., Macri, L. M., et al. 2022, ApJ, 934, L7
Rusu, C. E., Fassnacht, C. D., Sluse, D., et al. 2017, MNRAS, 467, 4220
Rusu, C. E., Wong, K. C., Bonvin, V., et al. 2020, MNRAS, 498, 1440
Schmidt, T., Treu, T., Birrer, S., et al. 2022, MNRAS, submitted [arXiv:2206.04696]
Schneider, P., & Sluse, D. 2014, A&A, 564, A103
Schneider, P., Ehlers, J., & Falco, E. E. 1992, Gravitational Lenses (New York: Springer-Verlag)
Schwarz, G. 1978, Ann. Stat., 6, 461
Scolnic, D. M., Jones, D. O., Rest, A., et al. 2018, ApJ, 859, 101
Shajib, A. J., Treu, T., & Agnello, A. 2018, MNRAS, 473, 210
Shajib, A. J., Birrer, S., Treu, T., et al. 2019, MNRAS, 483, 5649
Shajib, A. J., Birrer, S., Treu, T., et al. 2020, MNRAS, 494, 6072
Shajib, A. J., Treu, T., Birrer, S., & Sonnenfeld, A. 2021, MNRAS, 503, 2380
Skilling, J. 2004, in American Institute of Physics Conference Series, eds. R. Fischer, R. Preuss, & U. V. Toussaint, 735, 395
Sonnenfeld, A., Leauthaud, A., Auger, M. W., et al. 2018, MNRAS, 481, 164
Speagle, J. S. 2020, MNRAS, 493, 3132
Springel, V., White, S. D. M., Jenkins, A., et al. 2005, Nature, 435, 629
Sérsic, J. L. 1968, Atlas de Galaxias Australes (Cordoba: Observatorio Astronomico)
Suyu, S. H. 2012, MNRAS, 426, 868
Suyu, S. H., & Halkola, A. 2010, A&A, 524, A94
Suyu, S. H., Marshall, P. J., Hobson, M. P., & Blandford, R. D. 2006, MNRAS, 371, 983
Suyu, S. H., Marshall, P. J., Auger, M. W., et al. 2010, ApJ, 711, 201
Suyu, S. H., Hensel, S. W., McKean, J. P., et al. 2012, ApJ, 750, 10
Suyu, S. H., Auger, M. W., Hilbert, S., et al. 2013, ApJ, 766, 70
Suyu, S. H., Treu, T., Hilbert, S., et al. 2014, ApJ, 788, L35
Suyu, S. H., Bonvin, V., Courbin, F., et al. 2017, MNRAS, 468, 2590
Treu, T., & Marshall, P. J. 2016, A&ARv, 24, 11
Treu, T., Koopmans, L. V., Bolton, A. S., Burles, S., & Moustakas, L. A. 2006, ApJ, 640, 662
Van de Vyvere, L., Gomer, M. R., Sluse, D., et al. 2022a, A&A, 659, A127
Van de Vyvere, L., Sluse, D., Gomer, M. R., & Mukherjee, S. 2022b, A&A, 663, A179
Waskom, M., Botvinnik, O., Hobson, P., et al. 2014, https://doi.org/10.5281/zenodo.12710
Wong, K. C., Suyu, S. H., Auger, M. W., et al. 2017, MNRAS, 465, 4895
Wong, K. C., Suyu, S. H., Chen, G. C. F., et al. 2020, MNRAS, 498, 1420
Yıldırım, A., Suyu, S. H., & Halkola, A. 2020, MNRAS, 493, 4783
Yıldırım, A., Suyu, S. H., Chen, G. C. F., & Komatsu, E. 2021, A&A, submitted [arXiv:2109.14615]