Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS 503, 1096–1123 (2021)
Advance Access publication 2021 February 22

doi:10.1093/mnras/stab484

Time delay lens modelling challenge

X. Ding,[1,2⋆] T. Treu,[1] S. Birrer,[3] G. C.-F. Chen,[4] J. Coles,[5] P. Denzel,[6,7] M. Frigo,[8] A. Galan,[9] P. J. Marshall,[10] M. Millon,[9] A. More,[2] A. J. Shajib,[1,11] D. Sluse,[12] H. Tak,[13,14,15,16,17] D. Xu,[18] M. W. Auger,[19] V. Bonvin,[9] H. Chand,[20,21] F. Courbin,[9] G. Despali,[22] C. D. Fassnacht,[4] D. Gilman,[1] S. Hilbert,[23,24] S. R. Kumar,[20] J. Y.-Y. Lin,[25] J. W. Park,[10] P. Saha,[7,6] S. Vegetti,[8] L. Van de Vyvere[12] and L. L. R. Williams[26]

Affiliations are listed at the end of the paper

Accepted 2021 February 17. Received 2021 February 17; in original form 2020 June 15

ABSTRACT

In recent years, breakthroughs in methods and data have enabled gravitational time delays to emerge as a very powerful tool to measure the Hubble constant H_0. However, published state-of-the-art analyses require of order 1 yr of expert investigator time and up to a million hours of computing time per system. Furthermore, as precision improves, it is crucial to identify and mitigate systematic uncertainties. With this time delay lens modelling challenge, we aim to assess the level of precision and accuracy of the modelling techniques that are currently fast enough to handle of order 50 lenses, via the blind analysis of simulated data sets. The results in Rungs 1 and 2 show that methods that use only the point source positions tend to have lower precision (10–20 per cent) while remaining accurate. In Rung 2, the methods that exploit the full information of the imaging and kinematic data sets can recover H_0 within the target accuracy ($|A| < 2$ per cent) and precision (<6 per cent per system), even in the presence of a poorly known point spread function and complex source morphology. A post-unblinding analysis of Rung 3 showed the numerical precision of the ray-traced cosmological simulations to be insufficient to test lens modelling methodology at the percent level, making the results difficult to interpret. A new challenge with improved simulations is needed to make further progress in the investigation of systematic uncertainties. For completeness, we present the Rung 3 results in an appendix and use them to discuss various approaches to mitigating against similar subtle data generation effects in future blind challenges.

Key words: gravitational lensing: strong – methods: data analysis – cosmology: observations.

1 INTRODUCTION

The flat Λ cold dark matter (ΛCDM) cosmological model has been remarkably successful in explaining the geometry and dynamics of our Universe. It has been able to predict/match the results of a wide range of experiments covering a wide range of physical scales (Eisenstein et al. 2005; Betoule et al. 2014; Planck Collaboration XVI 2014; Planck Collaboration XIII 2016; Riess et al. 2016; Alam et al. 2017), and the expansion of our Universe (Riess et al. 1998; Perlmutter et al. 1999).

One of the key parameters of the model is the Hubble constant (H_0) that determines the age and physical scale of the Universe. Measuring H_0 at high precision and accuracy has been one of the main goals of observational cosmology for almost a century (Freedman et al. 2001). In recent years, as the precision of the measurements has improved to a few per cent level, a strong tension has emerged between early and late universe probes. As far as early-universe probes are concerned, analysis of *Planck* data yields $H_0 = 67.4 \pm 0.5\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Planck Collaboration VI 2020), assuming a ΛCDM model. In the local universe, the Equation of State of dark energy (SH0ES) team using the traditional 'distance ladder' method based on Cepheid calibration of Type Ia supernovae by finds $H_0 = 74.03 \pm 1.42\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Riess et al. 2019), and $H_0 = 72.4 \pm 2.0\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ based on the tip of the red giant brand (Yuan et al. 2019). The Carnegie–Chicago Hubble Program calibrated the tip of the red giant branch and applied to Type Ia supernovae, finding a mid-way Hubble tension as $H_0 = 69.8 \pm 0.8$ (±1.1 per cent stat) ± 1.7 (±2.4 per cent sys) (Freedman et al. 2019). The tension between late and early universe probes ranges between 4σ and 6σ (see summary by Verde, Treu & Riess 2019). If this ∼8 per cent difference is real and not due to unknown systematics in multiple measurements, it demonstrates that ΛCDM is not a good description of the universe, and additional ingredients such as new particles or early dark energy might be needed (e.g. Arendse et al. 2020; Knox & Millea 2020). Given the potential implications of this tension, it is crucial to have several independent methods to measure H_0 each with sufficient precision to resolve the tension (e.g. 1.6 per cent to resolve the 8 per cent H_0 tension at 5σ).

Time-delay cosmography by strong gravitational lensing provides a one-step measurement of H_0 together with other cosmological parameters (Refsdal 1966; Treu & Marshall 2016). The strongly lensed source produces multiple images, corresponding to multiple paths followed by the photons through the universe. According to Fermat's principle, the lensed images arrive at the observer at different times, corresponding to the extrema of the arrival time surface. The time delays between the images depend on the absolute value of cosmological distances, chiefly through the so-called 'time-delay distance', $D_{\Delta t}$, and can thus be used to infer H_0 like any

⋆ E-mail: dingxuheng@mail.bnu.edu.cn

other distance indicator (Schechter et al. 1997; Treu & Koopmans 2002a; Suyu et al. 2010). Importantly, time delay cosmography is independent of all other probes of H_0.

At the time of writing, the H_0 Lenses in COSMOGRAIL's Wellspring (H0LiCOW) and SHARP collaborations have finished the analysis of six strong lensed quasars and obtain a joint inference for Hubble constant as $H_0 = 73.3^{+1.7}_{-1.8}\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Wong et al. 2020). In addition, as part of the STRIDES collaboration, Shajib et al. (2020b) analysed one particularly information-rich strong lens system DES J0408-5354 alone and constrained the H_0 at 3.9 per cent level, in excellent agreement with the Wong et al. (2020) result. [In the rest of the paper, we refer to H0LiCOW/SHARP/COSMOGRAIL/STRIDES collectively as TDCOSMO (Millon et al. 2020)]. Measurements of H_0 using time delay lenses also have been investigated by other collaborations (Paraficz & Hjorth 2010; Ghosh, Williams & Liesenborgs 2020).

Based on the current results, it is predicted that a 1 per cent precision in the H_0 can be achieved via the time delay cosmography alone using a sample of 40 lensed systems (Shajib, Treu & Agnello 2018). However, two issues need to be addressed before a 1 per cent measurement of H_0 can be achieved with time delay cosmography. First, the analysis and computational costs need to be reduced in order to make the larger samples tractable. Secondly, all sources of potential systematic uncertainties must be investigated in order to identify and mitigate any outstanding one.

The first issue is well illustrated by the current state of the art. At present, the analysis of each system requires approximately 1 yr of effort full time by an expert investigator. Furthermore, the analysis by Shajib et al. (2020b) required approximately 1 million hours of CPU time. Analysing 40 lenses would thus be prohibitive with current techniques, especially in terms of investigator time. Efforts are underway to automate these modelling efforts so that they can be scaled to a large number of lenses reducing the investigator time per lens (Shajib et al. 2019), but much work remains to be done to get to high precision, low cost modelling (Schmidt et al., in preparation).

Regarding the second issue, a number of efforts are under way to identify systematic uncertainties (e.g. Millon et al. 2020). All parts of the analysis need to be checked with high-quality data and independent analysis, as well as with simulated data sets.

One effective strategy to test for unknown systematic errors is to use blind analysis. In the implementation followed by the TDCOSMO collaboration, the inferred values of $D_{\Delta t}$ and H_0 are kept blind until all coauthors agree to freeze the analysis during a collaboration telecon. The inference is then unblinded and published without modification. One of the goals of the blind analysis is to avoid conscious and unconscious confirmation bias.

Another powerful strategy is to study systematic errors using realistic simulations, possibly analysed blindly. Blind analysis of simulated data sets was the strategy of the 'time delay challenge' (TDC). In the TDC (Dobler et al. 2015; Liao et al. 2015), a so-called 'Evil' team first simulated a large number of realistic 'mock' time delay light curves, including anticipated physical and experimental effects. Then, the 'Evil' team published the mock data and invited the community to extract time delay signals *blindly* using their own method. Liao et al. (2015) showed that time delays can be measured from realistic data sets with precision within 3 per cent and accuracy within 1 per cent.

The success of TDC encouraged the community to take on the next step by verifying the precision and accuracy of lens models with a time delay lens modelling challenge (TDLMC), initiated on 2018 January 8 by posting a paper on arXiv (TDLMC1, Ding et al. 2018). The challenge 'Evil' team simulated 48 systems of mock strongly lensed quasars data and provided access to the data through a weblink to the participating teams ('Good' teams) to model, blindly:[1]

(i) https://tdlmc.github.io

The 'Evil' team produced realistic simulated time-delay lens data including (i) *Hubble Space Telescope (HST)*-like lensed AGN images, (ii) lens time delays, (iii) LOS velocity dispersions, and (iv) external convergence. After the 'Good' team submitted their inferred H_0, the performance of the adopted method could be estimated by comparing them with the true values in the simulation.

The number of simulated lensed quasars was chosen to have sufficient statistics to assess the performance at the percent level (7 per cent expected per system, gives approximately a ∼1 per cent precision on the mean). We stress that this is already a huge sample for current modelling methods, and thus the challenge is exclusively testing 'fast methods'. The computational cost of lens modelling is a major hurdle that will need to be overcome in the future; thus TDLMC uses large simulated samples aiming at testing the speed and performance of these 'fast methods'.

We also note that TDLMC is limited to the study of the lens model accuracy. Other sources of uncertainty are not considered. Therefore ancillary data, including time delay, LOS velocity dispersion, and information of external convergence are provided unbiased and with true uncertainties.

This paper provides the details of the challenge design that were hidden in the challenge opening paper (Ding et al. 2018, hereafter: TDLMC1) and presents an overview of the submission results. We encourage the individual 'Good' teams to submit more detailed papers on their methods and results. The paper is structured as follows. In Section 2, we describe the details of the challenge, including hitherto hidden adopted when simulating the sample. Sections 3 includes the response from the participating teams to this challenge and a brief summary of the method(s) adopted. The analysis of the submissions for Rungs 1 and 2 is presented in Section 4. For Rung 3, we discovered post-unblinding that the numerical precision of the ray-traced simulations was insufficient to test lens model methodology at the percent level, making the results from this rung difficult to interpret. Therefore, we dedicate a full Section 5 to the subtleties of Rung 3 that will need to be addressed in a future challenge that wishes to adopt numerical simulations of galaxies as a starting point. The results of Rung 3 are given in Appendix B for completeness, even though the results should be interpreted with caution. We draw some of the implications of the results and discuss our findings in Section 6. Section 7 presents a brief summary of the paper. To facilitate the tracking of the whole process of the TDLMC, we paste the main body of the TDLMC1 in Appendix A which was used to open this challenges.

2 DETAILS OF THE TDLMC CHALLENGE DESIGN

There are three challenge ladders in TDLMC, called Rung 1, Rung 2, and Rung 3. In addition, an entry-level Rung 0 is also designed for training propose. To ensure that the 'Good' teams do not infer any information for the previous rung, we reset the H_0 at each rung. We adopt two independent codes, namely LENSTRONOMY[2] (Birrer &

[1]For an early implementation of a blind challenge, see paper by Williams & Saha (2000).

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

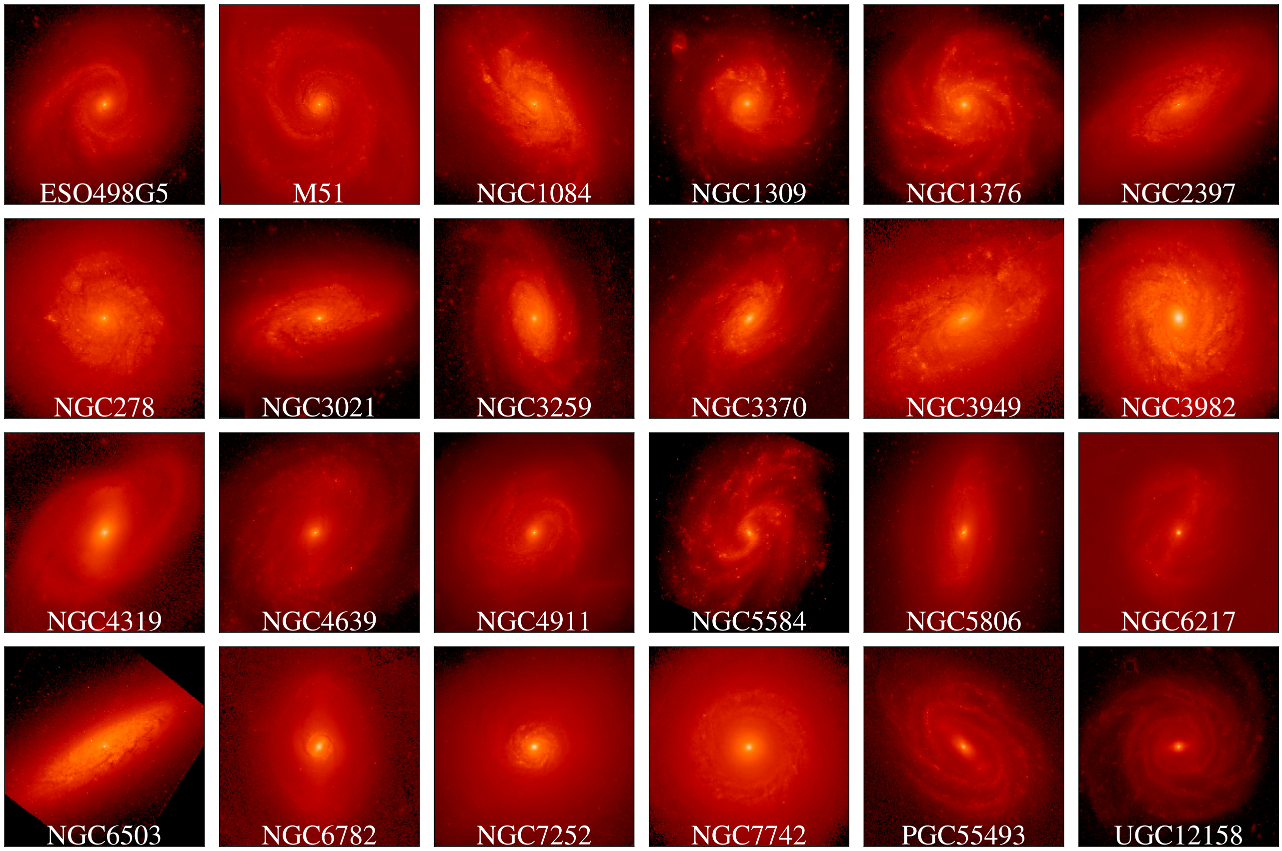

Figure 1. *HST* images of the realistic galaxies that were used in the TDLMC simulations as lensed AGN host galaxies.

Amara 2018) and PYLENS[3] (Auger et al. 2011), to simulate *HST*-like lensed AGN images (equally split). This strategy helps us to mitigate the 'home advantage', if any, in the sense that when 'Good' team happens to adopt the same code as the one used to generate the simulated images. The use of two independent codes also allows us to estimate numerical uncertainties related to the implementation of the algorithms, if present.

2.1 Challenge structure

The TDLMC begins with Rung 0, consisting of two lens systems – one *double* and one *quad*. This training rung aims to ensure that 'Good' team members understand the format of the data, and avoids any trivial coding errors or mistakes that potentially affect the results of the entire challenge.

Considering that the lens modelling process is usually time-consuming, we generated in total of 48 lensing systems, spread over three blind rungs (i.e. Rung 1, 2, and 3. There are 16 systems in each rung). The sample size is small enough to ensure it is tractable by the 'Good' teams and large enough to explore different conditions with sufficient statistics and uncover potential biases at the percent level. We increase the level of complexity from Rung 1 to Rung 3.

We reveal the details of the simulations for each blind rung in the rest of this section, including the ones that were only known to the 'Evil' team before unblinding.

[3]https://github.com/tcollett/pylens

2.2 Details of each Rung

For training purpose, Rung 0 was designed to be as simple as possible. Therefore, simple parametrized forms were adopted to describe the surface brightness of the deflector and the source galaxy (i.e. Sérsic), and the mass profile (elliptical power law) of the deflector. The true point spread function (PSF) is given, and external convergence was not considered. The Rung 0 data are released with all the input parameters, so that the 'Good' teams can validate their analysis.

In Rung 1, the increase in complexity with respect to Rung 0 is that the surface brightness of the AGN host galaxy is realistically complex, rather than described by a simply parametrized model like Sérsic. For the purpose of making realistic source galaxies, we started from high-resolution images of nearby galaxies obtained by *HST*. The digital images are downloaded from the Hubble Legacy Archive.[4] In order to get a clean galaxy image, we first removed isolated stars and background foreground objects in the field. All the processed galaxy images are shown in Fig. 1. Then, we obtained the global properties of these galaxies, using GALFIT to fit them as the Sérsic profiles so as to obtain their effective radius ($R_{\rm eff}$) in arcsec and total flux. This information is then used to rescale the galaxy size and magnitude in the source plane, as described in Section 2.3.3. A random external convergence value is also added in Rung 1 (see Section 2.4).

Rung 2 increases the complexity of Rung 1 by providing the 'Good' teams with only a guess of the PSF, instead of the actual PSF used to generate the simulations. This added complexity is meant to

[4]https://hla.stsci.edu/hlaview.html

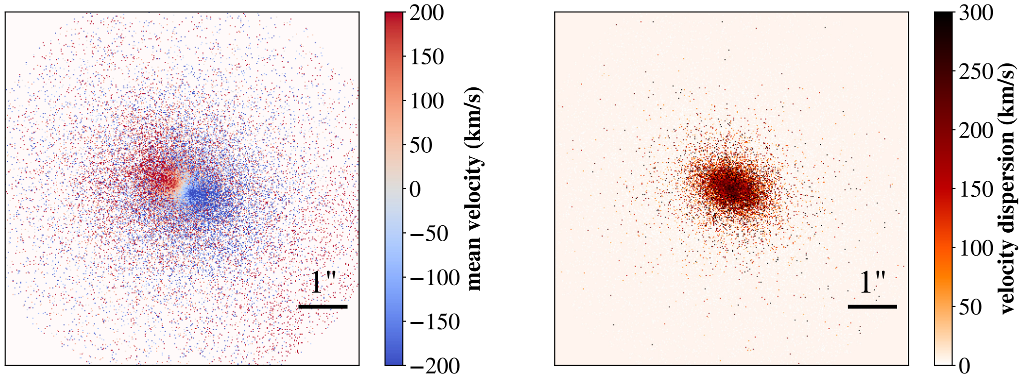

Figure 2. Example of 2D kinematic map for Rung 3. (left): LOS mean velocity ($v_{\rm ave}$) map. (right): velocity dispersion ($\sigma_{\rm ave}$) map. The image resolution is 16 times higher than *HST*/WFC3, i.e. 0.″13/16 = 0.″008125.

represent a typical situation where the observer uses a nearby star or model as an initial guess to the actual PSF and then improves on it using the quasar images themselves. In order to implement this step in a realistic manner, the 'Evil' team took one actual star observed by *HST* WFC3/F160W and constructed a high-resolution image by interpolation. This PSF image is used to carry out the simulation process described in Section 2.5. However, the PSF information based on a different star was given to the 'Good' teams.

Rung 3 was the most ambitious as we aimed to increase the complexity of the deflector mass distribution, in addition to retain the complexities of Rung 2. Assessing the effects of the complexity of the deflector mass distribution is crucial to evaluate the performance of modelling methods. For example, the mass sheet degeneracy (MSD; Falco, Gorenstein & Shapiro 1985) can be broken by adopting a power-law model to a non-power-law lens mass distribution (Schneider & Sluse 2013, 2014). The assumption of any specific mass profile can potentially result in the systematic bias to the measured Hubble constant, the magnitude of which depends on the difference between the model and the true unknown profile. This effect has been illustrated with cosmological hydrodynamic simulations (Xu et al. 2016; Tagore et al. 2018), suggesting a potential bias could be introduced due to the MSD. In an attempt to model this, the deflector galaxies in Rung 3 are based on cosmological numerical simulations. However, this is also the most conceptually difficult step because we do not have access to the 'true' mass distribution in real galaxies. For Rung 3, the 'Evil' team examined two options to produce a realistic deflector mass. The first option, following Gilman et al. (2017), is to use the surface brightness distribution of real galaxies, convert it into stellar mass, and add some dark matter components with some prescription. There are challenges to this approach; for example, it is not clear how to obtain self-consistent stellar kinematics. Thus, we discarded this option and decided to (following e.g. Xu et al. 2016) take the results of hydrodynamical simulations as our 'realistic' mass distribution [specifically Illustris (Vogelsberger et al. 2013, 2014) and the 'zoom' simulations in Frigo et al. (2019) were adopted]. This method has clear advantages but also limitations. For example, the results are only as good in terms of interpreting the real universe as the simulations are accurate, and it is well known that to simulate massive elliptical galaxies accurately is a challenge (e.g. Naab & Ostriker 2017). Furthermore, the resolution of the state-of-the-art simulations is finite, and the effects of finite resolution are important at the scale of strong lensing (Mukherjee et al. 2019; Enzi et al. 2020). We did not anticipate additional numerical issues that were discovered post-unblinding and will be discussed later in the paper.

After setting up the deflector redshift in Section 2.3.1, the Rung 3 deflector providers produced deflector maps at the corresponding redshift. These maps are at very high resolution, which is superior to *HST* by a factor of 16 (i.e. 0.″13/16 = 0.″008125 per pixel). The following information was provided by simulators to generate Rung 3 lensed images including

(i) *mass distribution*: The lensing maps include potential map (f), the deflection angles maps (including f'_x and f'_y, i.e. first-order derivation of f), and the hessian map (f''_{xx}, f''_{yy} and f''_{xy}, i.e. second-order derivation of f).

(ii) *surface brightness*: The 'observed' R band surface brightness is used to illustrate the light of the deflector in the simulation in Section 2.5. This map is also used to calculate the light-weighted LOS stellar velocity dispersion in Section 2.6.2.

(iii) *kinematics*: The kinematic maps include the LOS averaged velocity map ($V_{\rm ave}$), which accounts for the deflector rotation (Fig. 2, left-hand panel), and the averaged velocity-dispersion map ($\sigma_{\rm ave}$, see Fig. 2, right-hand panel).

2.3 Specific ingredients of the simulations

In TDLMC, the 'Evil' team intends to provide the mock data as realistic as possible. An overview of models/configuration that used to simulate the mock data have been introduced in the challenge designing paper TDLMC1, i.e. Appendix A. However, for obvious reasons, some details had to be kept blind and are presented here for the first time.

2.3.1 Redshift of deflector and source

The redshifts of the mock lenses are assumed to be distributed as for typical lenses. In Rungs 1 and 2, the 'Evil' team randomly generated their values from a normal distribution with $z_d \in$ N(0.5 ± 0.2), $z_s \in$ N(2.0 ± 0.4). In Rung 3, the lensing maps[5] are directly provided by the hydrodynamical simulation, fixing the redshift of the source at z_s = 1.5 and adopting the same deflector redshift same as provided by the simulation ($z_d \sim 0.5$).

[5]Lensing maps include the potential map, the deflection map, and the convergence map, i.e. f, f'_x, f'_y, f''_{xx}, f''_{yy}, and f''_{xy}.

2.3.2 Detailed set-ups of the lensing mass

The lensing maps are assumed to be composed of a main deflector, plus external shear and convergence. We describe the mass distribution of the deflector in this section. The deflector mass models are meant to describe typical elliptical galaxies.

In Rungs 1 and 2, the main deflector is assumed to follow a typical elliptical power-law mass distribution (see also the section 2.3.1 in TDLMC1, i.e. Appendix A2.2), with parameter distributions as listed in Table 1. In the simulations, we first draw the single isothermal sphere (SIS) velocity dispersion from the distribution in Table 1. Then, the corresponding Einstein radius can be calculated as $R_E = 4\pi v_d^2 \frac{D_{ds}}{D_s}$, where $D_{\rm ds}$, $D_{\rm s}$ are the angular diameter distance between the source and the deflector and from the source to us.

In Rung 3, the deflector mass information is provided by the two simulating teams (XD, MF, and SV) as described in Section 2.2. They also provide the velocity map of the deflector, which is used to calculate the aperture velocity dispersion in Section 2.6.2, and its surface brightness (see next section).

2.3.3 Surface brightness calculation

The surface brightness in an image is comprised of light both from the deflector and the lensed source. The main deflector surface brightness in Rungs 1 and 2 is described with the widely used Sérsic profile (as described in section 2.2.1 in TDLMC1, i.e. Appendix A2.1) with parameters distributed as shown in Table 1. In Rung 3, the (relative) R-band luminosity of stellar particles as deflector light were computed based on their age and metallicity, using the Bruzual & Charlot (2003) model. We only assume the distribution of the deflector's magnitude given in Table 1 to normalize its total flux for the purpose of achieving a realistic signal-to-noise ratio.

To define the realistic surface brightness distribution of the AGN host galaxy, we adopt a true high-resolution image taken from *HST* archive for all the blind rungs,. We first rescale the image by projecting it on the source plane, so that it has an apparent Sérsic effective radius drawn from Table 1. The magnitude of the source host galaxy is then rescaled from the observed according to the redshift of the source.

In order to obtain images similar to those used for cosmographic measurements, we assume that the active nuclei have a comparable flux to that of their host galaxy, see Table 1.

We vary the position of the source AGN so as to generate the lensing image in a range of configurations (including *cusp*, *fold*, *cross*, and *double*).

Table 1. Parameter distributions.

Simulation ingredient	Model and parameter values
(A): redshift	
Deflector redshift	$z_d \sim$ N(0.5 ± 0.2)
Source redshift	$z_s \sim$ N(2.0 ± 0.4)
(B): deflector (image plane)	
Lensing galaxy mass	Elliptical power law
SIS velocity dispersion	$v_d \sim$ N(250 ± 25) km s^{-1}
Einstein radius[a]	$R_{\rm Ein} = 4\pi v_d^2 \frac{D_{ds}}{D_s}$
Mass slope	$s \sim$ N(2.0 ± 0.1)
Ellipticity	$q \sim$ U(0.7 − 1.0)
Elliptical axis angle[b]	$\phi_m \sim$ U(0 − π)
Lensing galaxy SB	Sérsic profile
Total magnitude[c]	$mag \sim$ U(17.0 − 19.0) magnitude
Effective radius	$R_{\rm eff} = R_{\rm Ein}*$U(0.5 − 1.0)
Sérsic index	$n \sim$ U(2.0 − 4.0)
Ellipticity	$q \sim$ U(0.7 − 1.0)
Elliptical axis angle[d]	$\phi = \phi_m$U(0.9 − 1.1)
(C): AGN (source plane)	
Host galaxy SB	Realistic galaxy
Total magnitude	$mag \sim$ U(22.5 − 20.0) magnitude
Effective radius[e]	$R_{\rm eff} \sim$ U(0.″37, 0.″45), 1.0 < z_s < 1.5
	$R_{\rm eff} \sim$ U(0.″34, 0.″42), 1.5 < z_s < 2.0
	$R_{\rm eff} \sim$ U(0.″31, 0.″35), 2.0 < z_s < 2.5
	$R_{\rm eff} \sim$ U(0.″23, 0.″33), 2.5 < z_s < 3.0
Active nuclear light	Scaled point source
Source plane total flux	$f_{\rm AGN} = f_{\rm host}*$U(0.8 − 1.25)
External shear	
Amplitudes	$\gamma \sim$ U(0 − 0.05)
Shear axis angle	$\phi \sim$ U(0 − π)
External convergency	
External kappa[f]	$\kappa_{\rm ext} \sim$ N(0 ± 0.025)

Notes. Table lists the assumptions that were used to distribute the parameters for the TDLMC simulation. In Rung 3, non-parametrized deflectors (i.e. lensing galaxy mass and surface brightness) are adopted. Thus, the B part in the table is not adoptable for this rung. The distribution of 'N' means normal distribution and the 'U' means uniform distribution. Among all the parameters shown in the table, only the redshifts (with zero observation error) and unbiased estimated of external convergence $\kappa_{\rm ext} = 0 \pm 0.025$ were provided to the 'Good' teams.

[a] Using our definition, the Einstein Radius would be in the range [1.″00, 1.″20].

[b] The position angles start from the x-axis anticlockwise.

[c] The flux in cps and the magnitude value are related by the equation: mag = −2.5*log 10(flux) + zp, where zp is the filter zeropoint in AB system. For filter WFC3/F160W, zp = 25.9463.

[d] The effective radius and elliptical axis angle of the lensing light are assumed to be correlated with lensing mass at a certain level.

[e] The effective radius of the realistic galaxy is obtained by fitting Sérsic profiles using Galfit.

[f] $\kappa_{\rm ext}$ is randomly generated to calculate the time delay data. The parent distribution was provided to the 'Good' teams, but not the actual value, to mimic real analyses, see the descriptions at equation (4) for more details.

2.4 External shear and convergence

All the mass along the line of sight (LOS) contributes to the deflection of photons. In current state-of-the art analyses, this problem is made tractable by modelling the main deflector and the most massive nearby perturbers explicitly, while describing the remaining effects to first order as external shear and convergence ($\kappa_{\rm ext}$).

For simplicity, in this challenge we do not include massive perturbers, so there are just two components, the main deflector (described above) and the LOS external shear and convergence.

In Rungs 1 and 2, we add an external shear to the lensing potential with typical strength and random orientation, as shown in Table 1. External shear is not added in Rung 3 in order to keep the lensing potential self-consistent with the mass. More important is the effect of the external convergence ($\kappa_{\rm ext}$), since it affects the relative Fermat potential and time delay. As mentioned in TDLMC1, we consider the effect of $\kappa_{\rm ext}$ by drawing from a Gaussian distribution N(0 ± 0.025) for all the three Rungs.

2.5 Generating *HST*-like data

Having defined the ingredients of the simulations, we adopt two independent codes to build the pipeline that produces the mock *HST*

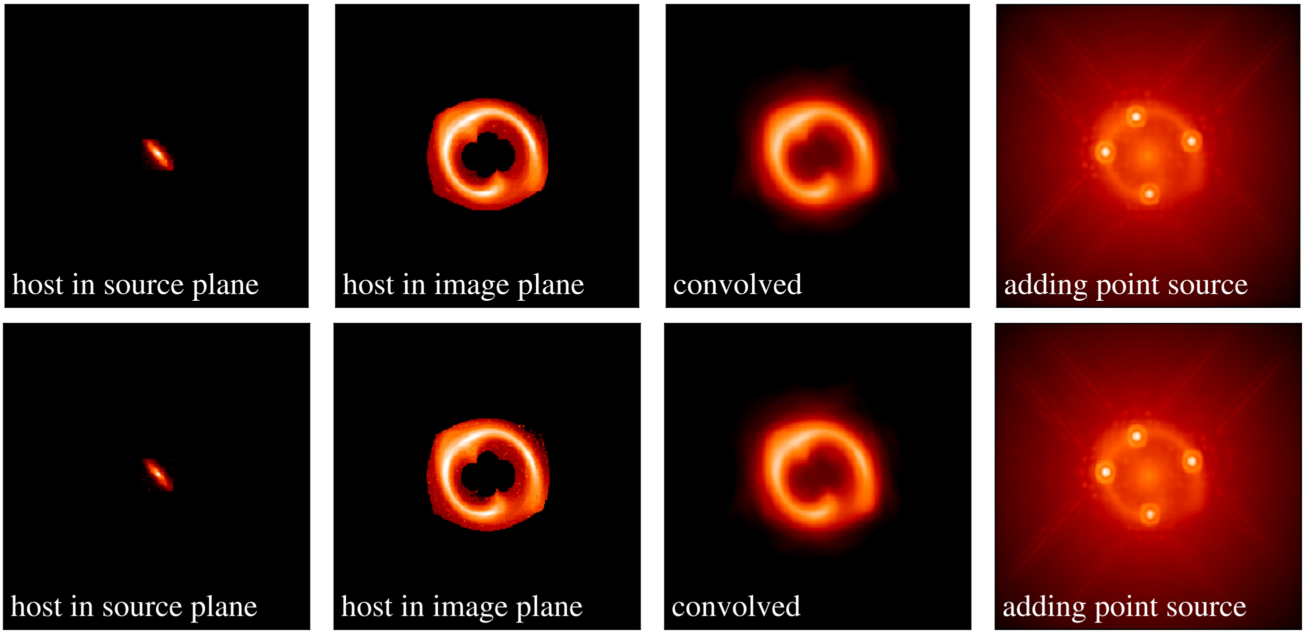

Figure 3. Illustration of the generation of a mock lensed AGN image, using LENSTRONOMY (top) and PYLENS (bottom). The image is sampled based on *HST*-WFC3/F160W at four times higher resolution (i.e. 0.″13/4). Note that the difference of numerical implementation between the two codes yield little systematic residuals, which is well below the noise level (see Fig. A2).

imaging data. We aim to simulate the image quality of typical state-of-the-art data sets, i.e. WFC3/F160W with individual exposures of 1200 s, and typical background. We use astrodrizzle to co-add eight single dithered exposures to obtain the final image with pixel sampling improved from 0.″13 to 0.″08.

The simulations are similar to those described by Ding et al. (2017a), which we refer to for more details. A brief description is given here for convenience. The simulation starts from high-resolution images with pixel scale four times smaller than the *HST* resolution (i.e. 0.″13/4). We start from actual *HST* images, as illustrated in fig. 2 of TDLMC1, i.e. Fig. A2. To numerically define the surface brightness of these actual images, PYLENS uses interpolation, and with LENSTRONOMY we chose to use shapelet decomposition (Refregier 2003; Birrer, Amara & Refregier 2015). We then rescale the image to the desired size. Then, the distortion by lensing is based on the deflection angles. We convolve the image plane surface brightness with the PSF and add scaled PSF in the position as the point sources to mimic instrumental resolution. In Rung 1, the PSF is generated with *TinyTim* (Krist, Hook & Stoehr 2011), while in Rung 2 and Rung 3 PSFs are extracted from the real *HST* images, and we use interpolation to obtain the PSF image at higher resolution. The pipeline is illustrated in Fig. 3. Note that at this step, the images are still sampled at the 0.″13/4 resolution.

In the next step, we rebin the pixels by 4 × 4 to degrade the image at *HST* resolution, i.e. 0.″13. We select eight different patterns to rebin the image, so as to mimic the dither process. In the next step, we add the noise to the data, see figs 1 and 2 in Ding et al. (2017a) for details. Finally, we use the drizzling process to co-add eight dithered images to obtain the final drizzled image at 0.″08 sampling. We present the 48 simulated images of the three rungs in Fig. 4.

In the TDLMC, the eight dithered *HST* images and the final drizzled images are all provided to the 'Good' teams including the science images, noise level maps, and a sampled PSF image.

2.6 Simulated ancillary data

In addition to the *HST* imaging data, the 'Evil' team provides time delay and aperture stellar velocity dispersion, computed as described in this section.

2.6.1 Time delay

The *true* time delay between the lensed AGN images are calculated using the following equations once the values of the simulated parameters are given by

$$\Delta t_{ij} = \frac{D_{\Delta t}}{c}\left[\phi(\theta_i) - \phi(\theta_j)\right], \tag{1}$$

where θ_j and θ_j are the coordinates of the images i and j in the image plane. $\phi(\theta_i)$ is the Fermat potential at image i and $D_{\Delta t}$ is so-called time-delay distance, defined as

$$\phi(\theta_i) = \frac{(\theta_i - \beta)^2}{2} - \psi(\theta_i), \tag{2}$$

$$D_{\Delta t} \equiv (1 + z_d)\frac{D_{\rm d} D_{\rm s}}{D_{\rm ds}}, \tag{3}$$

where $D_{\rm d}$, $D_{\rm s}$, and $D_{\rm ds}$ are, respectively, the angular distances from the observer to the deflector, from the observer to the source, and from the deflector to the source.

We consider the effects of the $\kappa_{\rm ext}$ to the observed time delay by

$$\Delta t_{\rm obs} = (1 - \kappa_{\rm ext})\Delta t_{\rm true}. \tag{4}$$

Note that the true value of $\kappa_{\rm ext}$ is assumed to be zero. It is the measured value of $\kappa_{\rm ext}$ that is scattered as $N(0 \pm 0.025)$. The effect of adding such $\kappa_{\rm ext}$ is equivalent to adding a perturbation on the observed time delays as equation (4). In principle, the external convergence effect should also shift the Einstein radius, which we did not consider in TDLMC for simplicity. That is, the $\kappa_{\rm ext}$ is only taken as a pure scatter effect on the time delay, hence H_0.

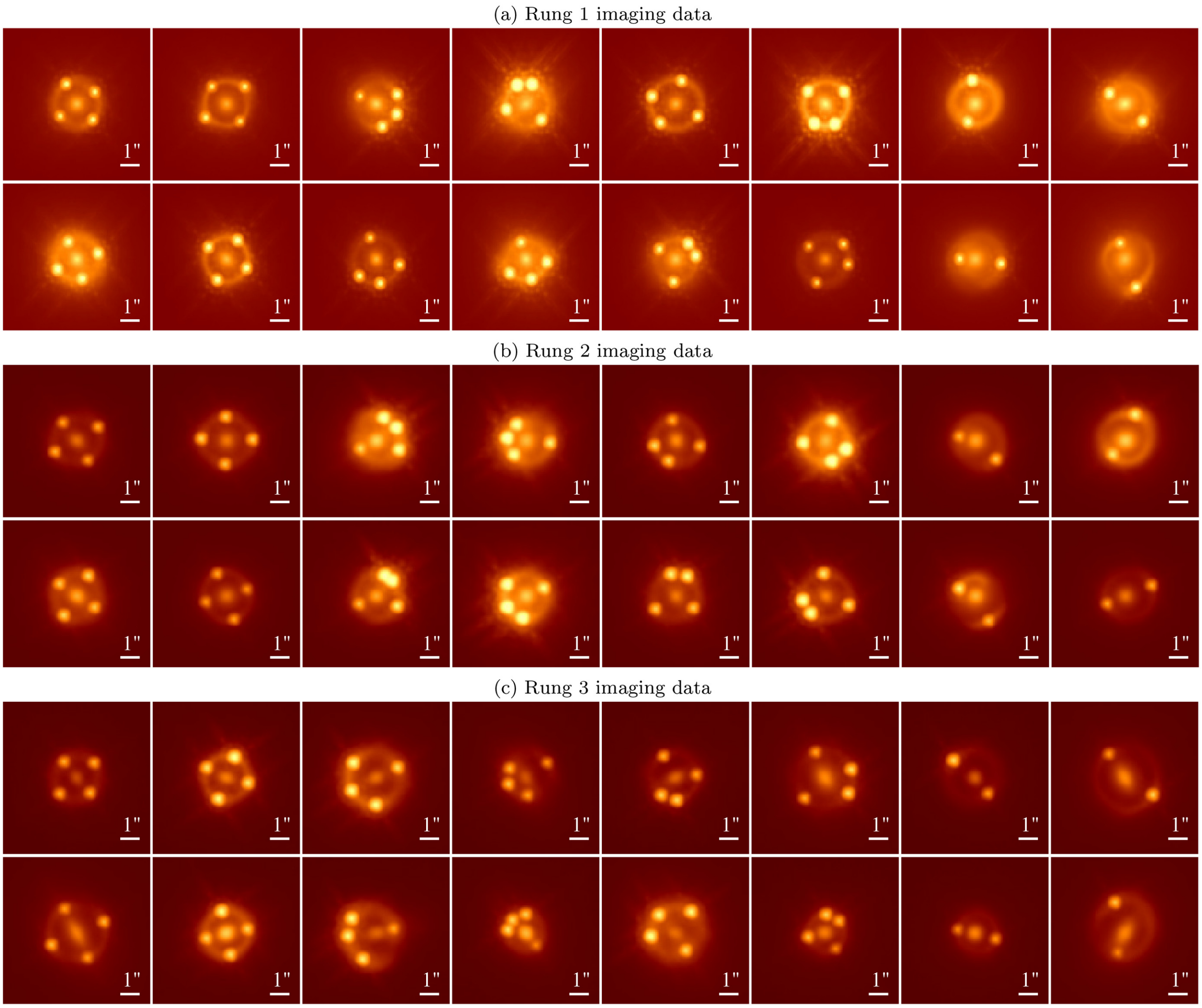

Figure 4. Mock data provided for Rung 1, Rung 2, and Rung 3. The configurations from left to right are *cross*, *cusp*, *fold*, and *double*. The images belonging to the same rung are shown with the same stretch to facilitate visual comparisons.

Assuming zero bias on the time delay, we add random error as the largest between 1 per cent and 0.25 d were adopted. We are deliberately keeping the uncertainties on the time delay as small as in the very best cases, in order not to obfuscate lens modelling errors.

2.6.2 Aperture stellar velocity dispersion

The aperture stellar velocity dispersion is helpful to break the MSD (Falco et al. 1985; Treu & Koopmans 2002b). The integrated LOS velocity dispersion is computed as the second moment of the velocity distribution weighted by surface brightness in a square aperture by 1″.0 × 1″.0, similar to the standard aperture used for real systems. Seeing conditions are also chosen to mimic the best current ground-based systems, idealized as a Gaussian kernel with a full width at half-maximum (FWHM) as 0″.6.

In Rungs 1 and 2, the deflector mass distribution is simply parametrized. Following current practice (e.g. Shajib et al. 2018), we assume that the mass distribution is related to the velocity dispersion profile through the spherical Jeans equation:

$$\frac{1}{l(r)}\frac{\mathrm{d}\left(l\sigma_{\rm r}^2\right)}{\mathrm{d}r} + 2\beta_{\rm ani}(r)\frac{\sigma_{\rm r}^2}{r} = -\frac{GM(\leq r)}{r^2}, \quad (5)$$

where $l(r)$ is the luminosity density of the deflector galaxy, $\sigma_{\rm r}$ is the radial velocity dispersion and $\beta_{\rm ani}(r)$ is the anisotropy profile and described as

$$\beta_{\rm ani}(r) = 1 - \frac{\sigma_{\rm t}^2}{\sigma_{\rm r}^2}, \quad (6)$$

where $\sigma_{\rm t}$ is the tangential velocity dispersion. The observed LOS velocity dispersion is surface brightness weighted, and thus can be calculated by solving the equation as (Mamon & Łokas 2005)

$$I(R)\sigma_{\rm los}^2(R) = 2G\int_R^\infty \mathrm{k}\left(\frac{r}{R}, \frac{r_{\rm ani}}{R}\right) l(r)M(r)\frac{\mathrm{d}r}{r}, \quad (7)$$

where $I(R)$ is the deflector surface brightness.

We adopt the Osipkov–Merritt parametrization of anisotropy $\beta_{\rm ani}(r) = 1/(1 + r_{\rm ani}^2/r^2)$ (Osipkov 1979; Merritt 1985 a, b), with

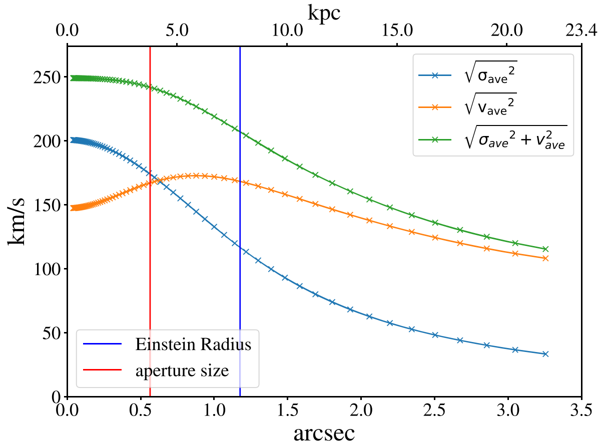

Figure 5. Surface-brightness-weighted line-of-sight stellar velocity dispersion as a function of aperture radius, based on a lens system in Rung 3. The stellar velocity dispersion is composed of the LOS mean velocity (i.e. $v_{\rm ave}$) and LOS velocity dispersion (i.e. $\sigma_{\rm ave}$), added in quadrature. The Einstein radius and the effective aperture size are also shown as the blue and red lines, respectively.

the function $\mathrm{k}(u, u_{\rm ani})$ given by

$$\mathrm{k}(u, u_{\rm ani}) = \frac{u_{\rm ani}^2 + 1/2}{(u_{\rm ani}^2+1)^{3/2}} \left(\frac{u^2 + u_{\rm ani}^2}{u}\right) \tan^{-1}\sqrt{\frac{u^2-1}{u_{\rm ani}^2+1}} - \frac{1/2}{u_{\rm ani}^2+1}\sqrt{1-1/u^2}. \quad (8)$$

The anisotropy radius $r_{\rm ani}$ is usually considered to be a free parameter with size comparable to the effective radius. In the simulation, we assume $r_{\rm ani} = R_{\rm eff}$ to calculate the velocity dispersion.

In Rung 3, the velocity dispersion is provided by the hydrodynamical simulations via high-resolution maps (16 times higher than *HST*), see Section 2.2. The aperture stellar velocity dispersion is thus a combination of the two kinematic maps by $V_{\rm aper} = \sqrt{V_{\rm ave}^2 + \sigma_{\rm ave}^2}$, where the $V_{\rm ave}$ and $\sigma_{\rm ave}$ is the LOS mean velocity and the velocity dispersion as shown in Fig. 2. The 'Evil' team calculate the 2D surface-brightness-weighted LOS dispersion and convolve it using a FWMH 0.″6 Gaussian kernel. Finally, the averaged velocity dispersion in the aperture was computed. Note that in principle the surface brightness weighting should be considered before convolving and aperture selection. However, the velocity map and the surface brightness map are both convolved using the same Gaussian kernel, making the sequence of this processing irrelevant. For illustration, the velocity dispersion as a function of aperture size is shown in Fig. 5.

A random Gaussian noise with 5 per cent standard deviation is added to the model velocity dispersion to represent high-quality measurement errors.

2.7 Metrics and expected performance

The 'Good' teams submitted their modelled H_0 of each lens system in the three rungs, and the 'Evil' team defined four standard metrics to estimate the performance of the submissions, including *efficiency* (*f*), *goodness* (χ^2), *precision* (*P*), and *accuracy* (*A*). They are defined as follows:

$$f = \frac{N}{N_{\rm total}}, \quad (9)$$

$$\chi^2 = \frac{1}{N}\sum_i \left(\frac{\tilde{H}_{0\,i} - H_0}{\delta_i}\right)^2, \quad (10)$$

$$P = \frac{1}{N}\sum_i \frac{\delta_i}{H_0}, \quad (11)$$

$$A = \frac{1}{N}\sum_i \frac{\tilde{H}_{0\,i} - H_0}{H_0}, \quad (12)$$

where N is the number of successfully modelled systems in each submission and $N_{\rm total} = 16$. δ_i is the uncertainty (1σ level) of H_0 by each systems in the submission. We identified the following targets for the metrics, based on current state of the analyses:

$$0.4 < \chi^2 < 2, \quad (13)$$

$$P < 6 \text{ per cent}, \quad (14)$$

$$|A| < 2 \text{ per cent}. \quad (15)$$

The χ^2 metric target is aimed to ensure that the estimated errors are a reasonable measure of the deviation from the truth. The *P* metric target is chosen to represent the precision of the best current measurements. The *A* metric target is set to investigate whether the fast methods can contain biases below the current reported precision by state-of-the-art analysis of samples of a few lenses. We don't set a metric target for *f*, as deciding which systems can be analysed with sufficient confidence depends on the methodology employed and thus we expect it to vary widely across submissions.

3 RESPONSE TO THE CHALLENGE

The TDLMC challenge mock data were released on 2018 January 8. The deadlines of the blind submission for the three rungs were 2018 September 8 for Rung 1, 2019 April 8 for Rung 2, and 2019 September 8 for Rung 3. Each rung was unblinded a few days after the submission deadline, to give teams a chance to learn in real time during the challenge. The 'Evil' team was especially mindful to help the 'Good' teams detect bugs and glitches that could invalidate the subsequent blind rungs, and prevent the teams from learning about their ability to tackle increased complexity.

Prior to the Rung 3 deadline, the 'Evil' team received in total 15, 17, and 24 submissions for Rung 1, Rung 2, and Rung 3, respectively, from five different participating teams ('Good' teams). We describe the method adopted by each team in the rest of this section.

3.1 Student-T team

H. Tak

This team proposes the following posterior density of H_0 designed to combine information from multiple lens systems in a simple but statistically principled way:

$$\pi(H_0 \mid D) \propto L(H_0)h(H_0). \quad (16)$$

The notation D denotes a set of the time delay estimates, Fermat potential difference estimates, and their standard errors for all unique pairs of lenses in 16 systems. Also, $L(H_0)$ represents the likelihood function and $h(H_0)$ indicates a Uniform (20, 120) prior density. This proper uniform prior guarantees posterior propriety of the resulting posterior (Tak, Ghosh & Ellis 2018). The team derives the likelihood function from a Gaussian assumption on the Fermat potential difference estimate (Marshall et al., in preparation):

$$\phi^{\rm est}_{ijk} \mid \Delta_{ijk}, H_0 \sim N\left(\phi_{ijk} = \frac{c\Delta_{ijk}}{D_\Delta(H_0)},\ \sigma^2\left(\phi^{\rm est}_{ijk}\right)\right), \quad (17)$$

where $\phi^{\rm est}_{ijk}$ denotes the Fermat potential difference estimate of the ith and jth lensed images in the kth lens system and $\sigma(\phi^{\rm est}_{ijk})$ indicates its standard error (1σ uncertainty). The notation Δ_{ijk} is the time delay between the ith and jth lensed images in the kth system ($\Delta_{ijk} = -\Delta_{jik}$). The time delay distance $D_\Delta(H_0)$ is treated as a function of only H_0 because all other information is completely given in the TDLMC.

On top of this Gaussian assumption on $\phi^{\rm est}_{ijk}$, the team adopts another Gaussian distribution for the time delay Δ_{ijk} with its mean equal to $\Delta^{\rm est}_{ijk}$ and standard error $\sigma(\Delta^{\rm est}_{ijk})$, i.e.

$$\Delta_{ijk} \sim N\left(\Delta^{\rm est}_{ijk},\ \sigma^2\left(\Delta^{\rm est}_{ijk}\right)\right). \quad (18)$$

The team also assumes that Δ_{ijk} and H_0 are independent a priori in a sense that Δ_{ijk} is typically inferred from light curves of multiple-lensed images without any information about H_0 (Tak et al. 2017).

The Gaussian assumptions in equations (17) and (18) make it simple to integrate out Δ_{ijk} analytically from their joint distribution, leading to the Gaussian distribution of $\phi^{\rm est}_{ijk}$ given only H_0:

$$\phi^{\rm est}_{ijk} \mid H_0 \sim N\left(\frac{c\Delta^{\rm est}_{ijk}}{D_\Delta(H_0)},\ \frac{c^2\sigma^2\left(\Delta^{\rm est}_{ijk}\right)}{D^2_\Delta(H_0)} + \sigma^2\left(\phi^{\rm est}_{ijk}\right)\right). \quad (19)$$

The team also assume the conditional independence among Fermat potential difference estimates within and across lensed systems given the Hubble constant H_0. Then, the likelihood function of H_0 is the product of Gaussian densities whose distributions are specified in equation (19), for every unique pair of gravitationally lensed images i and j across 16 lensed systems.

Since the posterior density function of H_0 in equation (16) is a function of only H_0, it is easy to draw an i.i.d. sample from this posterior via a grid sampling (chapter 5, Gelman et al. 2013).

On top of the posterior $\pi(H_0|D)$, the team models $\kappa_{\rm ext}$ using the relationship, $H^{\rm ext}_0 = (1-\kappa_{\rm ext})H_0$, where $H^{\rm ext}_0$ is the Hubble constant with $\kappa_{\rm ext}$ considered and H_0 is the one without $\kappa_{\rm ext}$ considered (Rusu et al. 2017). The team puts a $N(0, 0.025^2)$ prior on $\kappa_{\rm ext}$ for simplicity, which is assumed to be independent of the data. Finally, the posterior distribution of $H^{\rm ext}_0$ is derived as

$$\pi\left(H^{\rm ext}_0 \mid D\right) = \int \pi\left(H^{\rm ext}_0 \mid D, \kappa_{\rm ext}\right) g\left(\kappa_{\rm ext}\right) {\rm d}\kappa_{\rm ext}, \quad (20)$$

where g denotes the $N(0,\ 0.025^2)$ density of $\kappa_{\rm ext}$. The posterior distribution of $H^{\rm ext}_0$ in equation (20) is sampled via a Monte Carlo integration: (i) draw a random sample of $\kappa_{\rm ext}$ from $N(0,\ 0.025^2)$, (ii) sample H_0 from equation (16), and (iii) lastly set $H^{\rm ext}_0 = (1-\kappa_{\rm ext})H_0$. A Jacobian term is not needed for a deterministic transformation within a Bayesian sampling framework (Tak et al. 2020). The proposed framework does not account for the lens velocity dispersion for each lens system.

The key to the proposed approach is to obtain D to be used as a condition of the posterior distribution in equation (20) because given D, it is simple to draw a random sample of H_0. The team notes again that D is composed of time delay estimates, $\Delta^{\rm est}_{ijk}$'s, their standard errors, $\sigma(\Delta^{\rm est}_{ijk})$'s, Fermat potential difference estimates, $\phi^{\rm est}_{ijk}$'s, and their standard errors, $\sigma(\phi^{\rm est}_{ijk})$'s. The first two components are fully known in the TDLMC, and thus the remaining ingredients for sampling $H^{\rm ext}_0$ from equation (20) are $\phi^{\rm est}_{ijk}$'s and $\sigma(\phi^{\rm est}_{ijk})$'s.

For this purpose, the team uses LENSTRONOMY (version 0.4.3, Birrer & Amara 2018). In Rung 1, the team uses the elliptical Sérsic profile for the source light model and adopts one, two, or three elliptical Sérsic profiles for the lens light model. In Rungs 2 and 3, the team utilizes a superposition of a smooth power-law elliptical mass density profile (SPEMD) with external shear for the lens mass model. An elliptical Sérsic profile *with* shapelets (Birrer et al. 2015) is adopted for the source light model, and an elliptical Sérsic profile is used for the lens light model. The team fixed $n_{\rm max} = 10$ as the order of the shapelets basis for the baseline model. Also, the team uses the PSF iteration to correct the PSF model (Shajib et al. 2019). In addition, the team manually boosts the noise level by adopting one of seven different PSF error inflation rates (1 per cent, 5 per cent, 10 per cent, 15 per cent, 20 per cent, 25 per cent, and 30 per cent) to deal with additional errors in the given PSF. This means that for each unique pairs of lenses, the team fits the model by LENSTRONOMY seven times each with one of the seven PSF error inflation rates.

For each of the seven fits, LENSTRONOMY produces a posterior sample of ϕ_{ijk} that is possibly non-Gaussian. Thus, to obtain $\phi^{\rm est}_{ijk}$ and $\sigma(\phi^{\rm est}_{ijk})$, the team summarizes the posterior distribution in two ways; posterior mean and standard deviation (Summary 1); posterior median and quantile-based standard error (Summary 2). This is because the posterior mean and standard deviation can be misleading if the posterior distribution of ϕ_{ijk} is not Gaussian.

Consequently, for each pair of lensed images the team obtains the seven pairs of $(\phi^{{\rm est}(l)}_{ijk}, \sigma(\phi^{{\rm est}(l)}_{ijk}))$ for $l = 1, \ldots, 7$, according to each type of summary. Since D requires having only one representative pair of $(\phi^{\rm est}_{ijk}, \sigma(\phi^{\rm est}_{ijk}))$ for each pair of lensed images, the team takes an average of these seven pairs in three ways. The first one is a Fisher-type weighted average of $\phi^{{\rm est}(l)}_{ijk}$'s weighted by $1/\sigma^2(\phi^{{\rm est}(l)}_{ijk})$'s (Average 1). This averaging method puts more weights on the pairs with smaller standard errors. The second averaging method simply takes an arithmetic mean over seven estimates and over seven variances (Average 2). This way puts equal weights on all seven pairs regardless of their different standard errors. Finally, the third one uses the same arithmetic mean as Average 2 but sets $\sigma(\phi^{\rm est}_{ijk})$ to a sample variance of the seven estimates, $\phi^{{\rm est}(l)}_{ijk}$'s (Average 3). This one does not use the information about standard errors at all. The team briefly describes the details of each submission in Table 2.

Due to the space limitations, the detailed information of the lens modelling settings will be presented in a separate paper (Tak et al., in preparation).

3.2 EPFL team

M. Millon, A. Galan, F. Courbin, V. Bonvin

3.2.1 *modelling technique*

The EPFL team followed a streamlined version of current modelling practices applied to time delay cosmography. The main difference with respect to the analysis described by Birrer et al. (2019) and Shajib et al. (2020b) is that the challenge is known to be free of significant perturbers besides the main deflector and the LOS. Taking advantage of this information and to reduce computation costs, a smaller number of model choices was considered in the challenge as compared to real systems. In addition, in order to reduce human investigator time, the modelling was standardized as opposed to tailored to the specific of each individual lens. For this purpose, a partly automated modelling pipeline was developed by the team. A more detailed description of the pipeline may be the subject

Table 2. The details of the submissions of Student-T team. Summaries 1, 2, Averages 1, 2, 3 are defined in Section 3.1.

Rung	Algorithm	Details
1	1	Summary 1 and Average 1
	2	Summary 1 and Average 2
	3	Summary 1 and Average 3
	4	The same as Algorithm 1 except that three pairs are intentionally removed for consistency
	5	The same as Algorithm 2 except the three pairs
	6	The same as Algorithm 3 except the three pairs
2	1	Summary 1 and Average 1
	2	Summary 1 and Average 2
	3	Summary 2 and Average 1
	4	Summary 2 and Average 2
	5	An independent replication of Algorithm 1
3	1	Summary 1 and Average 1
	2	Summary 1 and Average 2
	3	Summary 2 and Average 1
	4	Summary 2 and Average 2
	5	The same as Algorithm 1 with three times more repetitions (i.e. 21 pairs instead of 7 pairs)
	6	The same as Algorithm 2 with 21 pairs
	7	The same as Algorithm 3 with 21 pairs
	8	The same as Algorithm 4 with 21 pairs
	9	The same as Algorithm 5 but without considering $\kappa_{\rm ext}$ i.e. sampling from (16) instead of (20)
	10	The same as Algorithm 6 but sampling from (16)
	11	The same as Algorithm 7 but sampling from (16)
	12	The same as Algorithm 8 but sampling from (16)

of a future paper. The standardization is a necessary step towards modelling large numbers of systems, but it may result in failures if the one-size-fits all approach is not (yet) sufficiently accurate.

For the modelling part, the team used the publicly available software LENSTRONOMY (Birrer & Amara 2018). This software is well validated and has been previously used for the modelling and cosmography analysis of real time delay strong lens systems (Birrer, Amara & Refregier 2016; Birrer et al. 2019; Shajib et al. 2020b). The entire challenge data set was used as constraints for our models, including the provided drizzled image, noise maps, and PSF; the measured time delays at lensed AGN positions $\Delta t_{\rm measu}$; the measured LOS velocity dispersion of stars in the lens galaxy $\sigma_{\rm los,\, measu}$; and the estimate of the external convergence $\kappa_{\rm ext}$.

The models are described by linear (surface brightness amplitudes) and non-linear parameters, depending on the type of profiles (see Birrer et al. 2015, for details). The team chose to add the time delay distance $D_{\Delta t}$ as a free non-linear parameter.

For a single system, the generic workflow starting from lens modelling up to H_0 inference can be divided in the three following steps.

(1) First, linear and non-linear parameters are optimized by alternating particle swarm optimizer (PSO) runs and increments of the complexity of lens models. Parameters are sampled from uniform priors, ensuring that all lenses can be modelled from the same initial set of priors. The time delay distance $D_{\Delta t}$, considered as a free non-linear parameter of the model, is constrained by the measured time delays $\Delta t_{ij,\,\rm measu}$ by enforcing the modelled time delays to be compatible with the measured ones. Modelled time delays $\Delta t_{ij,\,\rm model}$ are computed as follows:

$$\Delta t_{ij,\rm model} = (1+z_{\rm d})\,\frac{D_{\Delta t}}{c}\,\Delta\Phi_{ij,\rm model}, \quad (21)$$

where $z_{\rm d}$ is the lens redshift, $\Phi_{\rm model}$ is the model Fermat potential, c is the speed of light, and 'ij' defines the difference of the indicated quantity evaluated at the positions of two lensed AGN i and j. This procedure gives best-fitting estimates of the linear and non-linear parameters, which are then used as a starting point of an MCMC sampling. Both PSO and MCMC routines are implemented in LENSTRONOMY, based on the COSMOHAMMER package (Akeret et al. 2013) and EMCEE (Foreman-Mackey et al. 2012).

(2) For each MCMC sample, the team derived in a post-processing step the LOS velocity dispersion $\sigma_{\rm los,\,model}$ from model parameters. The team used the Osipkov–Merritt model to solve the spherical Jeans equation, again following current practices, e.g. Suyu et al. (2010), Shajib et al. (2018), with routines implemented in LENSTRONOMY. The team computed angular diameter distances from both kinematics and time delays. The sampled time delay distance gives directly the distance ratio $D_{\rm d}D_{\rm s}/D_{\rm ds}$. The modelled LOS velocity dispersion, along with the model parameters $\boldsymbol{\xi}_{\rm model}$, is used to compute the distance ratio $D_{\rm s}/D_{\rm ds}$ from the following relation (Birrer et al. 2016):

$$\sigma^2_{\rm los,\,model} = \frac{D_{\rm s}}{D_{\rm ds}}\,c^2\,J(\boldsymbol{\xi}_{\rm model},\,r_{\rm ani}), \quad (22)$$

where J captures all dependencies on model parameters and kinematics anisotropy, moving any dependencies on cosmological parameters in the distance ratio. The external convergence was also sampled as $\kappa_{\rm ext} \backsim \mathcal{N}(0,\,0.025)$, to simulate a correction to the time delay distance by any mass external to the main deflector, through: $D_{\Delta t,\,\rm eff} = D_{\Delta t}/(1-\kappa_{\rm ext})$. From the two distance ratios described above, is straightforward to extract the angular distance to the deflector, namely $D_{\rm d}$.

(3) Following Birrer et al. (2019), the inference of the Hubble constant is performed in the 2D plane defined by angular distances $D_{\Delta t,\,\rm eff}$ and $D_{\rm d}$. This plane encodes the joint constraints from imaging data, time delays, external convergence, and lens kinematics. In order to approximate the full covariance between the two $D_{\Delta t,\,\rm eff}$ and $D_{\rm d}$ posteriors, both distributions are used to evaluate the likelihood when inferring H_0. Since $\Omega_{\rm m}$ is fixed in this challenge, the only cosmological parameter being sampled is the Hubble constant.

(4) The team computed the final inferred H_0 value and associated uncertainty estimates for an entire rung in two steps. First, an outlier rejection scheme was performed, according to the following criteria, that were found to be good markers of poor models:

(a) Each individual H_0 median value must be inside the prior bounds defined by the TDLMC, i.e. inside $[50,\,90]\,{\rm km\,s^{-1}Mpc^{-1}}$;

(b) The sampled time delay distance $D_{\Delta t}$ (free parameter constrained by the lens model and time delays) and the modelled time delay distance $D_{\Delta t,\,\rm model}$ [obtained through equation (21) inversion] must be consistent with each other at the $\lesssim 1\sigma$ level;

(c) The modelled lens velocity dispersion $\sigma^2_{\rm LOS,\,model}$ must be consistent at $\lesssim 2\sigma$ level with the measured value;

(d) Each individual H_0 posterior must be consistent with each other at the $\lesssim 2\sigma$ level.

When all the above criteria were fulfilled, the team kept the model for the joint inference over the rung, for a given model family. This leads to a set of $D_{\Delta t}$ and $D_{\rm d}$ pairs of posteriors. The team then performed two joint inferences using

(a) only time-delay information. H_0 is sampled according to the ensemble of $D_{\Delta t}$ posteriors only.

(b) both time-delay and kinematics information. This follows the approach described in Birrer et al. (2019), H_0 is sampled in the 2D plane over the set of $D_{\Delta t}$ and $D_{\rm d}$ posteriors. This last option is the standard procedure used for joint inference of real lenses (e.g, Wong et al. 2020)

Note that even in the first case of inference H_0 from $D_{\Delta t}$ only, knowledge about kinematics still plays a (smaller) role, because of model selection steps are performed *before* the inference.

The joint H_0 posteriors described above are computed under the assumption that the systems do not share systematic errors. If this assumption breaks, then one should marginalize from individual distributions, instead of the joint inference. For this reason, the team also submitted H_0 posteriors that are marginalized over the selected models. Additional details specific to each rung are given in the following subsections.

3.2.2 Rung 1

In Rung 1, lens mass and light profiles are simply-parametrized. Hence, the team used SPEMD (Barkana 1998) with external shear profiles to describe the projected mass distribution, and a single Sérsic profile for the lens surface brightness. For the source, the team used a Sérsic profile superimposed to a set of shapelets (Refregier 2003; Birrer et al. 2015). The team chose $n_{\rm max} = 8$ as the maximum order of the shapelets basis for their the baseline model. When significant residuals were observed at Einstein ring location, $n_{\rm max}$ were slightly increased, typically up to $n_{\rm max} = 14$. The source galaxy centroid (Sérsic+shapelets) was fixed to the position of the quasar, itself modelled as a single point source constrained by enforcing lensed images to trace back to the same position in source plane.

The 'Evil' team kept secret any details related to kinematics modelling assumptions, including the anisotropy model they used for computing velocity dispersion. As stated above, the EPFL team used Osipkov–Merritt modelling for computing velocity dispersions (Osipkov 1979; Merritt 1985a). This model assumes a parametrized anisotropy parameter $\beta_{\rm ani} = r^2/(r^2 + r_{\rm ani}^2)$, where $r_{\rm ani}$ is the anisotropy radius, which defines the radius at which stellar orbits go from being radial (near the centre) to isotropic (equally radial and tangential). Standard practices are to sample the anisotropy space through a uniform prior on the anisotropy radius, see e.g. Suyu et al. (2012) and Shajib et al. (2018). In Rung 1, the team used a uniform prior $r_{\rm ani} \backsim \mathcal{U}(0.5,\ 5)\, r_{\rm eff}$, where $r_{\rm eff}$ is the half-light radius of the lens.

The unblinding of Rung 1 revealed that the team's submitted inference was strongly affected by one (or several) systematic error(s), as quantified by an accuracy of $A = 7.512$ per cent. The main origin of this bias was found to be a consequence of the high precision of measured time delays, which surpasses those of real time delay lenses so far, combined with small angular separation between lensed images. Indeed image separations are on average ∼1 arcsec, and time delays are of the order of dozens of days with precision 0.25 d. Typical lensed systems modelled by the TDCOSMO collaboration have on average image separations of ∼2.5 arcsec with time delays precision up to a couple of days. A particularly high precision is therefore required when modelling the position of each lensed images in the setting of the challenge, which is not the case for all real systems analysed so far. A lack of precision can propagate to a significant bias on the Hubble constant. The bias they observed in their initial Rung 1 submission allowed them to highlight such a requirement, which have been the topic of a dedicated paper by Birrer & Treu (2019). The authors introduced simple formulae that, given an expected precision on the Hubble constant, can be at first order used to estimate the astrometric requirements that must be fulfilled, from image separations and time delays precision. They refer the reader to that paper for consequences of such requirements and quantitative examples. As discussed in Section 4.2, the problem was solved by the EFPL team by introducing in LENSTRONOMY a nuisance parameter to describe the unknown difference between true and measured image positions and marginalizing over it.

For Rung 1, the team submitted a single sample of models, and related joint Hubble value, following the description above.

3.2.3 Rung 2

In Rung 2, only a guess of the PSF was provided, in order to test PSF reconstruction algorithms. The team used the iterative PSF reconstruction originally implemented in LENSTRONOMY. For a set of baseline models, the team incorporated this routine during parameter optimization, effectively alternating between PSO and PSF reconstructions. Having noticed that the PSF was degraded the same way for each of the 16 lenses of Rung 2, the team computed a median stacked PSF kernel from their best reconstructed kernels. This reconstructed PSF was then used for all of their subsequent Rung 2 modelling attempts.

Based on Rung 1 knowledge, the team took into consideration the astrometric requirements described in previous subsection, in order to mitigate a potential bias on the inferred Hubble constant. The team allowed extra degrees of freedom to model any unknown uncertainty on the position of AGN images (a.k.a. point sources), by introducing in the parameter space, two new 'offset' parameters, δ_x and δ_y, for each of the 2 or 4 images independently. These offsets actually represent the error between the (modelled) position of point sources on the image, and the (predicted) positions at which the Fermat potential is evaluated for time delays computation. These additional parameters are sampled as non-linear parameters, and constrained by time delays and imaging data. The team regularly checked that those offsets were correctly constrained, with amplitudes expected to be below the image pixel scale.

After careful analysis of post-unblinding or Rung 1, the team realized that most consistent results were obtained when $r_{\rm ani} \approx r_{\rm eff}$. Consequently, in Rung 2, the team fixed the anisotropy radius $r_{\rm ani}$ to be equal to the lens half-light radius for all the remaining submissions.

The remaining volume of the parameter space (mass and light profiles of the lens galaxy, light profiles of source galaxy, and quasar model) was identical to those of the previous rung.

The team submitted four model samples and corresponding joint value for Rung 2:

(i) `DdDdt`: the inferred H_0 was obtained through joint inference in the 2D plane $\left\{D_{\Delta t,\rm eff},\ D_{\rm d}\right\}$;

(ii) `margDdDdt`: same as `DdDdt`, except that the inferred final value was obtained by marginalization over individual H_0 posteriors, as opposed to a joint inference ;

(iii) `Ddtonly`: same as `DdDdt`, except that H_0 values were inferred only from the time delay distance $D_{\Delta t,\rm eff}$;

(iv) `margDdtonly`: same as `Ddtonly`, except that the inferred final value was obtained by marginalization over individual H_0 posteriors.

3.2.4 Rung 3

For Rung 3, the team used the exact same PSF reconstruction method as for Rung 2. For lens models, they followed the practices of the TDCOSMO collaboration, in the sense that they chose two families of models: power law and composite. The former consists of elliptical power-law mass distribution with external shear, whereas the latter distinguishes the baryonic mass and dark matter, in addition to the external shear. For the baryonic matter, they used a double Chameleon profile (see Suyu et al. 2014, for definition) to fit the lens surface brightness, and convert it to surface mass density through a constant mass-to-light ratio, introduced as a free parameter. They modelled the dark matter component as a single elliptical NFW profile.

In order to improve their efficiency in modelling Rung 3 with two model of families, which require significant amount of work, they also used double Chamelon profiles to describe the lens light in their power-law models. This allowed them to extract best-fitting lens light parameters from their power-law models, and properly initialize the corresponding composite models, for a given lens. Note that it is different than the usual TDCOSMO procedure, where the surface brightness of the lens galaxy is fitted with double Sérsic for power-law mass models. They checked that no systematic errors were introduced when using double Sérsic instead of double Chameleon profiles, which is expected as the latter is designed to be a good approximation of the former.

The rest of the procedure was similar to their submissions for Rung 2 and 3, in terms of selection criterions and joint inference. The selection was performed independently for the two model families described above, meaning that their composite and power-law submissions did not necessarily consist in the same modelled lenses, nor the same number of lenses. For each model family, they submitted two submission pairs, with H_0 inferred from (1) joint $\left\{D_{\Delta t,\,\mathrm{eff}},\ D_{\mathrm{d}}\right\}$ inference and (2) $D_{\Delta t,\,\mathrm{eff}}$ only. Additionally, they submitted a third pair of submissions with a subset of lenses whose models were coincidentally accepted with both model families, which enabled them to combine their inferences from power-law and composite models. More precisely, for a given lens, they marginalized over the two model families, prior to the final joint inference H_0 among the different lenses. To summarize, one ended up with six submissions for this rung.

3.3 Freeform team

P. Denzel, J. Coles, P. Saha, L. L.R. Williams

The lenses were reconstructed with the codes GLASS by Coles, Read & Saha (2014) and its precursor PIXELENS by Saha & Williams (2004), which are based on the free-form modelling technique. In contrast to other methods, free-form lens reconstructions are not restricted to a parametrized family of models, but rather build a lens as superpositions of a large number of mass components, e.g. mass tiles or pixels, with minimal assumptions about the form of the full lens. The price to pay for the flexibility is that the free parameters outnumber the constraints and thus regularization needs to be imposed to avoid overfitting the data.

While GLASS and PIXELENS are completely separate codes, implemented in different languages, and using different Monte Carlo sampling engines, they both share the same approach to free-form lenses. Represented as a discrete grid of pixels, the lens potential takes the following form:

$$\psi(\theta) = \sum \kappa_n \nabla^{-2} Q_n(\theta), \tag{23}$$

where κ_n is the density of the nth mass tile and $Q_n(\theta)$ is the shape integral over the nth pixel. Each tile is a square and its contribution $\kappa_n Q_n(\theta)$ to the potential at θ can be worked out analytically (AbdelSalam, Saha & Williams 1998). In both GLASS and PIXELENS, the tiles cover a circular area centred on the lensing galaxy. The radius of this area r_{p}, in pixels, determines the resolution of a model. For instance, $r_{\mathrm{p}} = 8$ places one tile at the centre and eight tiles extending left and right (17 pixels side to side) with a total of 225 pixels covering the entire circular area. The tile size in arcseconds can be set explicitly or estimated such that there are several rings of pixels outside the outermost image. Mass distributions that are assumed to be radially symmetric (doubles and some quads) are constrained to have diametrically opposite pixels of equal value, which reduces the number of pixels by half. GLASS also allows for the central pixel to be further subdivided into 3 × 3 or 5 × 5 subpixels, to capture a steeply rising cusp. In this paper, we denote the use of the subdivision with the parameter $sp = 3$ or $sp = 5$, respectively. A central pixel with no subdivision is equivalent to $sp = 1$. Both codes ensure a small region of 'pixel rings' outside the outermost image.

Quasar image positions, time delays, and redshifts are the only data input for the models. Image parities are also given but are determined solely from experience and by generating test models to verify image parity assignment. As is well-known images are located at extrema of $\nabla\psi$ and the sign of $\nabla\nabla\psi$ determines the parity.

This input is used to create a system of equations that are linear in the source position β and mass tiles κ_n. The intrinsic and well-known problem of lensing arises from the fact that there are infinitely many solutions to these linear equations. Free-form techniques usually sample from that solution space according to a few reasonable priors. Most notably they require non-negative mass tiles, limited to twice the average of all neighbouring tiles, and the local density gradient to point typically 45° from the centre; additionally, the azimuthally averaged mass profiles must not increase, which still allows for twisting isodensity contours and significantly varying ellipticities with radius. These priors ensure some minimum level of physical correctness where the density of the reasonably smooth lensing mass is increasing towards the centre. From the information provided by the 'Evil' team for each rung, further physical parameters and priors could be included:

(i) Redshifts set the distance scales (assuming a standard cosmology of $\Omega_m = 0.27$ and $\Omega_\Lambda = 0.73$).

(ii) The models allowed for external shear.

(iii) Time delays were constrained, for GLASS with uncertainties of ±0.25 d, for PIXELENS without.

(iv) The range of H_0 was limited to 50−90 km s^{-1} Mpc^{-1}.

The velocity dispersion information was not used to constrain the models, but can be derived from the models following Leier (2009).

A free-form lens model consists of an ensemble of models; ∼1000 typically provide a good cover of the solution space. A single model may contain more than one lensing system, in which case they are coupled by the requirement that H_0 must be the same for all systems.

An ensemble usually includes many different convergence maps some of which are unphysical at times. Generally, this is not a problem, as the ensemble average[6] washes out these outliers. Nevertheless, the ensemble can be filtered according to different criteria in order to optimize the ensemble average. In Rung 2 for instance, we applied such a post-processing filter based on a simplified version of the source mapping algorithm described in Denzel et al. (2020). Instead of only using quasar image positions, the entire photometric information was used to select the most probable models in the following manner. A χ^2 value was computed for each

[6]Due to the linear nature of the lens equation, a superposition of solutions also is a solution.

lens model of the ensembles by fitting a synthetic image using the drizzled image data (including science images, noise level maps, and a sampled PSF image, while masking out the lensing galaxies in the centre). For each ensemble, 300 models with the best values were retained to estimate H_0. This ensured that only the models which best fit the entire image data were used to infer H_0. Despite slight improvements on H_0 the filter was abandoned again for Rung 3, because, at the time, the methods were computationally too intensive.

Each ensemble H_0 distribution was Gaussian fitted as was demanded by the submission format of the challenge. However, it is important to note that the distributions are far from Gaussian as discussed in Denzel et al. (2021).

For each rung, model ensembles were generated for all 16 single lenses and for groups of multiple lenses (four sets of four lenses) using GLASS and PIXELENS. These submissions have the suffixes `Single` and `Multi`, respectively.

In Rung 1, all GLASS models use $p_r = 8$ but single lenses have $sp = 5$, and multilenses use $sp = 1$. In Rung 2, GLASS single lens models have a higher resolution using $r_{\rm p} = 10$ and $sp = 5$, while multilenses use $r_{\rm p} = 8$ and $sp = 1$. For Rung 3, the resolution of GLASS models was increased as high as was computationally feasible to $r_{\rm p} = 12$ for the submission `glassSingleHiRes`. The submission `glassSingleLowRes` used the standard $r_{\rm p} = 8$. Both submissions further resolved the central pixel with $sp = 3$.

Additionally, in Rung 1 `glassCherrypick` is a multilens analysis using a subset of four lenses for which the individually modelled arrival-time surfaces and mass maps subjectively appeared to be unproblematic (e.g. no additional images and a clean arrival time surface). In Rung 2, `glassSynthFiltered` used the aforementioned source mapping algorithm to select models from the `glassMulti` ensemble that best reproduced the lensed images.

3.4 Rathnakumar team

S. R. Kumar, H. Chand

The main motivation of the team was to understand to what accuracy and precision H_0 can be constrained through simple analytical modelling, constrained by point image positions and flux ratios. To this end, the team modelled the TDLMC Rung 0, Rung 1, and Rung 2 systems using GLAFIC software (Oguri 2010). In general, the mass distribution of the lensing galaxy was modelled as singular isothermal ellipsoid along with a shear component (SIE + γ). In Rung 1, some double lens systems were found to overfit ($\chi^2 << 1$). Thus, the team replaced SIE by SIS along with a shear component (SIS + γ). All the Rung 2 systems were modelled as SIE + γ, except for one system for which this model was found to result in catastrophic failure. The exceptional case was modelled as singular isothermal ellipsoid without any shear component (SIE only).

The astrometry of the lensed quasar images and the centre of the lensing galaxy were measured from the provided *HST* drizzled image for each system using 'IMEXAM' task in IRAF. The astrometric coordinates were assigned an uncertainty of $0\farcs02$. The fluxes of the lensed quasar images were also measured through aperture photometry using the same IRAF task from *HST* drizzled image. From these fluxes, the absolute flux ratio was computed for each lensed quasar image with respect to the brightest image. These flux ratios were each assigned a sufficiently large uncertainty of 0.2 (e.g. for quads, three flux ratio values were considered), in order to accommodate for factors such as intrinsic quasar variability, microlensing-induced variability, etc. Parity constraints were inferred for the lensed quasar images based on the arrival time order and the configuration, in case of quadruple lenses. The team used the velocity dispersion and relative time delay values provided along with their uncertainties as constraints during the modelling. The fitting process was done using standard procedure by implemented in GLAFIC. The background cosmology was fixed to $\Omega_m = 0.27$, $\Omega_\Lambda = 0.73$, and $w = -1$. Source and lens redshifts were fixed for each system according to the provided values. The measured H_0 for each system was taken to be that corresponded to the best-fitting model. The 1σ uncertainty of H_0 was inferred by fixing it at different values around the measured value and marginalizing all the model parameters to minimize χ^2 and noting the range where $\Delta\chi^2 < 1$, with respect to the value for the best-fitting model. The error bars in positive and negative directions were averaged. To include the LOS effects for Rungs 1 and 2 systems, 2.5 per cent was added in quadrature to the H_0 uncertainty. The team submitted only the results for those systems where H_0 was constrained to better than $20\,{\rm km\,s^{-1}\,Mpc^{-1}}$. The remaining systems were flagged as failure. The team also submitted results filtered according to cut-off values of 15 and $10\,{\rm km\,s^{-1}\,Mpc^{-1}}$ to see what effect these selections have on the TDLMC performance metrics. In order to combine all the H_0 estimates from the individual systems into one global value for a rung, the team did a simple weighted average.

3.5 H0rton team

J. W. Park, Y.-Y. Lin

The H0rton team automated the lens modelling using a Bayesian neural network (BNN), a method pioneered by Hezaveh, Levasseur & Marshall (2017). The BNN-inferred lens model posterior was then propagated into H_0 inference. Readers are referred to the accompanying method paper (Park et al. 2020) for more details. The implementation of the H0rton pipeline is available in the form of the open-source PYTHON package H0RTON.[7]

Given the drizzled image of each lens system, the BNN predicted the posterior PDF over a power-law elliptical mass model (PEMD) parameters, the source position, and the half-light radius of the Sérsic lens light (for computing the velocity dispersion likelihood). The posterior PDF was parametrized as a mixture of two Gaussians with full covariance matrices, informed by the results of Wagner-Carena et al. (2020) that the parameter recovery improved with this form of the posterior in comparison to the single uncorrelated Gaussian originally adopted by Hezaveh et al. (2017).

The training set for the BNN consisted of 200 000 images. The assumed lens mass and lens light profiles were identical to those used to generate the TDLMC data of Rungs 1 and 2, i.e. PEMD and elliptical Sérsic, respectively. The AGN host light, however, was assumed to follow an elliptical Sérsic profile in order to keep the parametrization simple. The predictive model parameters in the training set were assumed to be independently distributed, aside from selecting the magnification to be greater than 2 in order to ensure significant lensing signal. The approximate range of each parameter was inferred from the Rung 1 data set and confirmed by visual inspection on the Rung 3 images. For the PSF convolution, the simulation rotated among the 16 drizzled PSF maps provided in Rung 1. The PSF information was fed to the BNN only via the convolved image and the network was expected to process the deconvolution internally. Non-drizzled images or PSF maps were not used. The training set was generated using the team's open-source

[7]https://github.com/jiwoncpark/h0rton

Table 3. Summary table of input data.

Team	Point sources	Extended source	Kinematics
Student-T	Yes	Yes	No
EPFL	Yes	Yes	Yes
Freeform	Yes	No	No
Rathnakumar	Yes	No	Yes
H0rton	Yes	Yes	Yes

Note. Table summarizes the input data as used by the 'Good' team. In addition, all teams use time delays and redshifts, and simulated *HST* images to constrain the deflector.

PYTHON package BAOBAB,[8] which wraps around the LENSTRONOMY package (Birrer & Amara 2018).

The combined cosmographic likelihood was the product of the likelihoods of the time delays and the LOS velocity dispersion with the nuisance parameters, i.e. the external convergence, kinematic anisotropy, and the BNN-inferred model parameters, marginalized out. The velocity dispersion was modelled assuming a spherical power-law mass profile and a Hernquist lens light to solve the spherical Jeans equation, as done by Suyu et al. (2010). The kinematic computations were performed with LENSTRONOMY. Samples from the cosmographic likelihood were obtained via MCMC sampling with EMCEE (Foreman-Mackey et al. 2012). Note that, in contrast to the traditional forward modelling approach, the pixelwise image likelihood was never directly modelled. Instead, the BNN-inferred posterior entered the MCMC integration as a prior over the lens model parameters at the H_0 inference stage.

It was discovered during the analysis procedure that, when the BNN-predicted source position and lens model were directly used to solve the lens equation, the predicted number of images often did not agree with the data. These cases were traced to sources very close to the caustic, for which the precision requirements on the source position tended to be very high (see e.g. Birrer & Treu 2019). The BNN-inferred posterior was placing significant weight on models that did not produce the correct number of images. To alleviate this discrepancy, the image positions were manually estimated from the images and fed in as additional data into the MCMC sampling pipeline. A Gaussian likelihood of the image positions, when appended to the MCMC sampling objective, iteratively brought the BNN-inferred lens model closer to one that yielded the observed image positions.

The H0rton team joined the challenge late and only made a blind submission to Rung 3. The open-box data sets of Rungs 1 and 2 that were available at the time, however, informed the team's approach.

4 ANALYSIS OF RUNGS 1 AND 2 SUBMISSIONS

To summarize the input data used by each 'Good' team, we present the information in Table 3. A summary of the computation and investigator time invested in the challenge is given in Table 4. A brief analysis of the results of the submissions is presented in this section.

4.1 Basic statistics

In this section, we give an overview of the performance of the blind submissions to Rungs 1 and 2. As described in Section 2.7, four metrics are used to perform a synthetic evaluation of the submissions, even though we encourage teams to carry out more detailed studies. The metrics of each submission for Rungs 1 and 2 are shown in Table 5. Note that the 'Good' teams were allowed to adopt multiple methods based on different algorithms and submit multiple results

[8] https://github.com/jiwoncpark/baobab

Table 4. Summary of computation and investigator time.

Team	CPU time (h)	Investigators time (h)
Student-T	15 400	48
EPFL	500 000	1700
Freeform	5000	–
Rathnakumar	–	–
H0rton	–	–

Note. Estimated CPU and investigator time spent for TDLMC by the teams who provided them.

Table 5. Metrics of blind submission for Rungs 1 and 2.

Team	Algorithm	f	$\log(\chi^2)$	$P(\%)$	$A(\%)$
	Metrics of Rung 1				
Student-T	Algorithm1	0.688	0.771	4.834	1.049
Student-T	Algorithm2	0.688	0.615	5.374	1.752
Student-T	Algorithm3	0.688	0.493	8.237	2.492
Student-T	Algorithm4	0.688	0.541	6.533	0.293
Student-T	Algorithm5	0.688	0.324	7.019	1.005
Student-T	Algorithm6	0.688	0.094	10.036	1.825
EPFL	Submission	0.688	0.411	6.169	7.512
Freeform	GlassCherrypick	0.250	1.193	5.785	− 22.847
Freeform	GlassMulti	1.000	0.406	9.002	− 4.570
Freeform	GlassSingle	1.000	0.264	13.812	− 8.516
Freeform	PixelensMulti	1.000	0.349	9.299	− 7.220
Freeform	PixelensSingle	1.000	0.790	13.123	− 5.632
Rathnakumar	Cut-off10	0.125	0.024	8.429	4.112
Rathnakumar	Cut-off15	0.250	− 0.164	12.137	6.337
Rathnakumar	Cut-off20	0.375	− 0.339	15.419	3.932
Rung 1 combined metrics					
Direct average		0.654	0.522	9.140	− 1.745
Bagging			− 0.199	9.646	− 1.644
Rejection σ-median			0.219	9.639	− 1.081
Rejection σ-mean			0.205	9.649	− 0.920
Rejection widths-median			0.522	9.147	− 1.779
	Metrics of Rung 2				
Student-T	Algorithm1	0.812	− 0.161	18.215	− 4.811
Student-T	Algorithm2	0.875	− 0.672	27.764	5.161
Student-T	Algorithm3	0.812	0.845	8.531	− 6.096
Student-T	Algorithm4	0.750	0.414	12.267	− 3.663
Student-T	Algorithm5	0.750	− 0.247	18.225	− 8.014
EPFL	DdDdt	0.688	− 0.127	3.260	− 1.740
EPFL	Ddtonly	0.688	0.180	2.635	− 1.957
EPFL	MargDdDdt	0.688	− 0.127	3.260	− 1.740
EPFL	MargDdtonly	0.688	0.180	2.635	− 1.957
Freeform	GlassMulti	1.000	2.762	10.603	− 3.496
Freeform	GlassSingle	1.000	1.834	13.010	− 3.580
Freeform	GlassSynthFiltered	1.000	1.847	13.017	− 0.683
Freeform	PixelensMulti	1.000	0.053	16.335	17.095
Freeform	PixelensSingle	1.000	− 0.293	21.480	3.187
Rathnakumar	Cut-off10	0.125	− 0.249	12.304	− 2.090
Rathnakumar	Cut-off15	0.312	− 0.293	17.166	4.797
Rathnakumar	Cut-off20	0.375	− 0.311	18.382	1.461
Rung 2 combined metrics					
Direct average		0.785	1.765	13.154	− 0.309
Bagging			− 0.343	10.768	0.372
Rejection σ-median			− 0.040	14.041	1.481
Rejection σ-mean			0.660	17.170	0.870
Rejection widths-median			1.769	13.187	− 0.279
	Rung 2 post-blind submissions, see Section 4.3				
Student-T	Algorithm1	0.938	− 0.421	15.492	− 5.969
Student-T	Algorithm2	1.000	− 0.873	26.844	6.396
Student-T	Algorithm3	1.000	0.317	6.591	0.056
Student-T	Algorithm4	1.000	− 0.162	11.805	4.330

Note. Table summarizes the metrics of the blind submission for Rungs 1 and 2, together with the post-blind submissions by Student-T team (see Section 4.3).

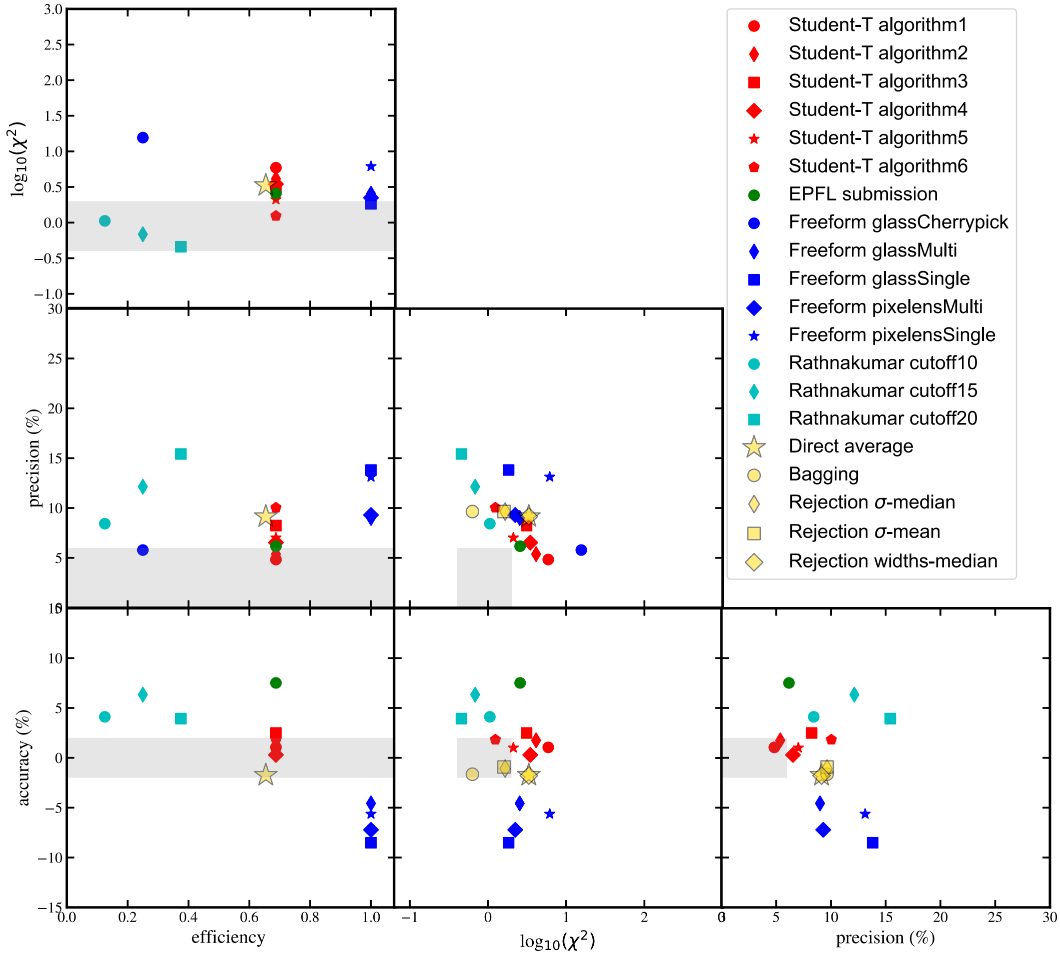

Figure 6. Results for TDLMC Rung 1, showing the four metrics for all the submissions using different algorithms, together with the combined metrics shown as yellow points. The f, χ^2, P, and A are defined in Section 2.7. The grey regions in each plot bracket the expected performance of the metrics. Note that we did not set a target performance for the *efficiency* (f) metric; the grey regions in the three left-hand panels is drawn only for the other metrics. The last four combined metrics have either reconstructed its sample or rejected the outliers, thus the efficiency metrics are also not considered.

for each rung. The metrics plots by each submission are shown in Figs 6 and 7.

'Good' teams including Student-T, EPFL, and Rathnakumar also estimated and submitted the *overall* H_0, which is their best estimation using the combination of the lens systems analysed in each rung. The Freeform team also submitted the *overall* H_0 values after unblinding, although it is based on a straightforward average of blind inferences. Following equations (11) and (12), we calculated the metrics of precision and accuracy using the values of these overall H_0 and show them in Fig. 8. Note that overall H_0 is a joint inference from the combination of the multiple lens systems; thus, the precision metric value should be, in principle, decreased by the square root of the volume of the analysed lensed systems (i.e. $\sqrt{N}$), compared to Figs 6 and 7. The combination of multiple systems could also in principle allow teams to flag and reject outliers, thus reducing the impact of overly complicated systems, i.e. those for which the modelling tool or data quality is insufficient.

Furthermore, we investigated whether there is 'wisdom in the crowd' by considering metrics combined across H_0 submissions for Rungs 1 and 2. We considered the following strategies:

(i) *Direct average*: of all the submission of H_0 without weighting;

(ii) *Bagging*: For each lens in one rung, we compute the mean H_0 across all the submissions and estimate the uncertainty via bootstrap resampling. The result is taken as the H_0 inference for each lens system. Then, we combine H_0 inference across all the lens systems in the rung to compute the metrics;

(iii) *Rejection σ-median*: We combine the entire H_0 submissions in one rung to do the bootstrap resampling. We remove the outliers before inferring the averaged metrics using the following criteria. In each bootstrap seed, we calculate the median H_0 ($H_{0,\,\mathrm{median}}$) and reject the outliers by $|H_{0,\mathrm{median}} - \tilde{H}_{0\ i}|/\delta_i > 3$;

(iv) *Rejection σ-mean*: Similar to *rejection σ-median*, we remove the outliers in each bootstrap seeding using the H_0 weighted mean value ($H_{0,\,\mathrm{mean}}$) by $|H_{0,\mathrm{mean}} - \tilde{H}_{0\ i}|/\delta_i > 3$;

(v) *Rejection widths-median*: Similar to previous rejection methods, we use the widths of the H_0 distribution in each bootstrap drawing (i.e. W_{H_0}, which is the half width of 16–84 per cent confidence interval in H_0 distribution) and remove the outliers in the bootstrapped sample by $|H_{0,\mathrm{mean}} - \tilde{H}_{0\ i}|/W_{H_0} > 3$.

The combined metrics are shown in Table 5 and Figs 6 and 7. These values can be considered as the combined performance of the entire 'Good' teams in each rung. As expected, we find that the points of these averaged metrics are in the centre of the cloud of

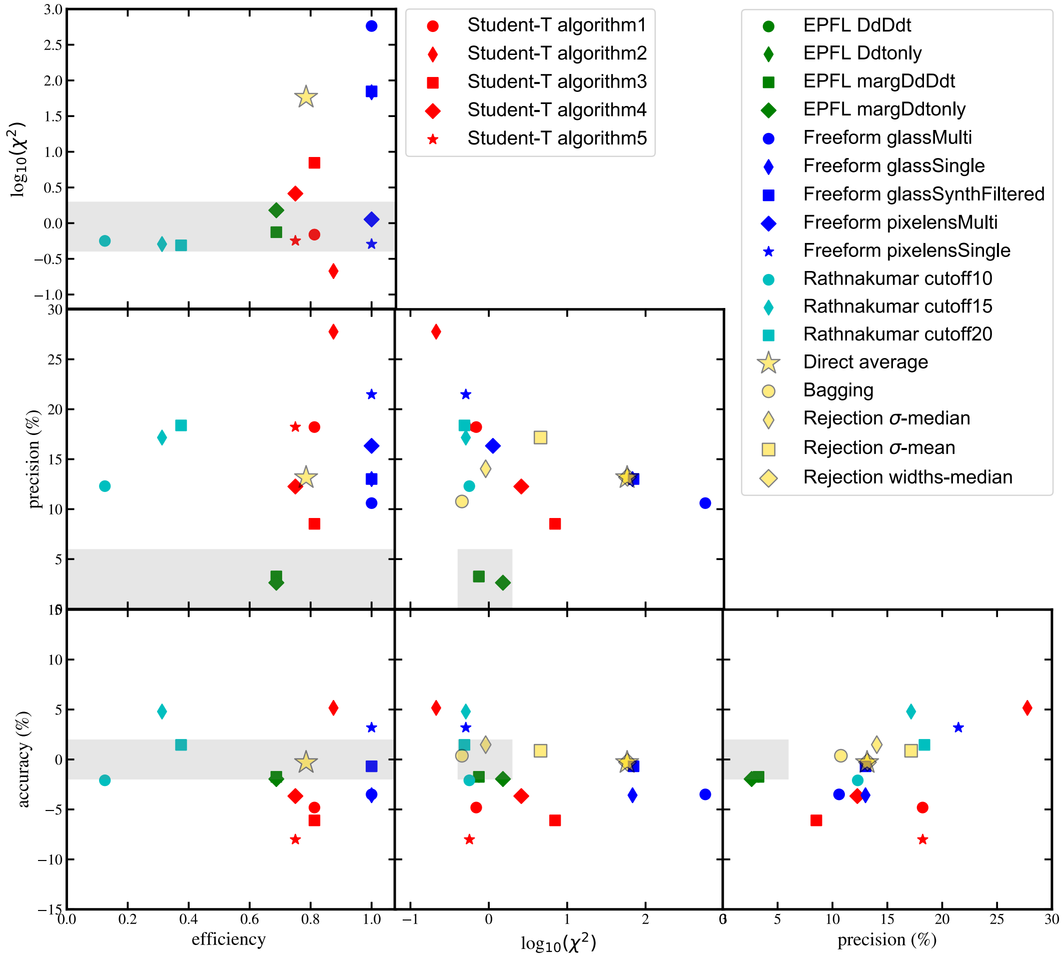

Figure 7. Same as Fig. 6, but for Rung 2's results. To demonstrate the improvement of the Student-T team's result after using the correct file (see Section 4.3), figure also shows post-blind submissions labeled by the hollow markers. We note that the *combined metrics*, i.e. the yellow points, does not include the results by post-blind algorithms.

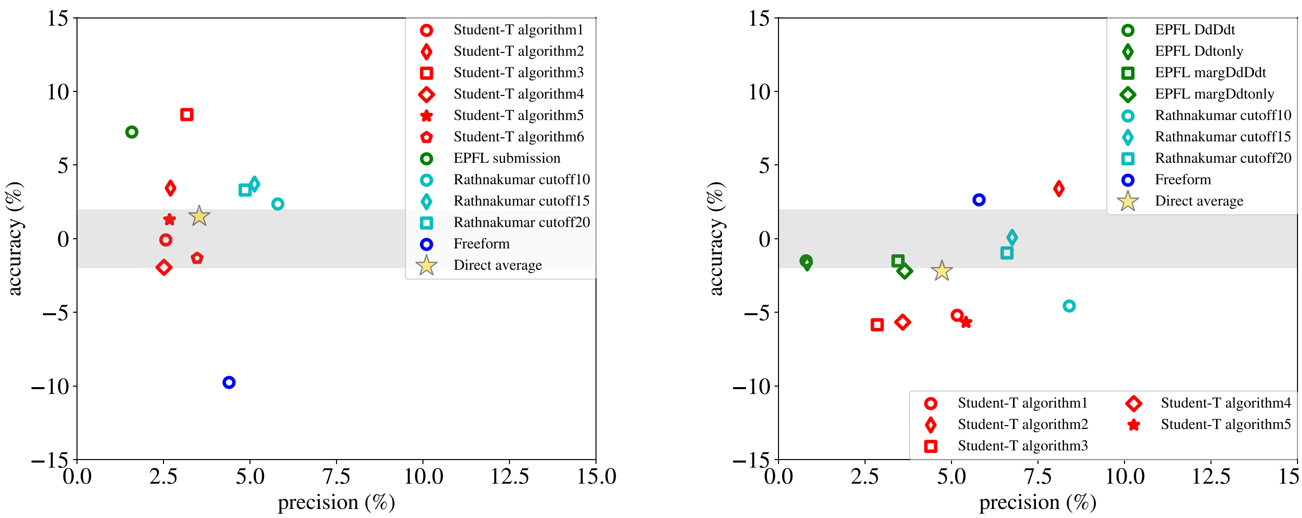

Figure 8. Results for TDLMC Rung 1 (left) and Rung 2 (right), based on the *overall* H_0 submissions by each algorithm. The Freeform team submitted the *overall* H_0 values after unblinding, although it is based on a straightforward average of blind submissions.

Table 6. Summary of the *precision* and *accuracy* by combining algorithms based on different level of information used to constrain the models in Rung 2.

Combined fitting algorithm	Precision (%)	Accuracy (%)
Everything	2.9	− 1.8 ± 0.4
Extended source:		
Blind submissions only	11.4	− 2.7 ± 1.0
Blind + post-blind	12.8	− 1.2 ± 0.8
Only post-blind for Student-T	10.1	0.02 ± 0.69
Point sources	15.2	2.5 ± 1.4

Note. 'Everything' calculates the metrics combining the algorithms that adopted point sources, extended source, and kinematics. 'Extended Source' combines the results of the algorithms that utilized the lensed arc information in the lens modelling. 'Point Sources' combines the ones that use only point sources but not lensed arcs. For cases with post-blind submissions explained in the text, we report all the permutations of blind and post-blind combinations.

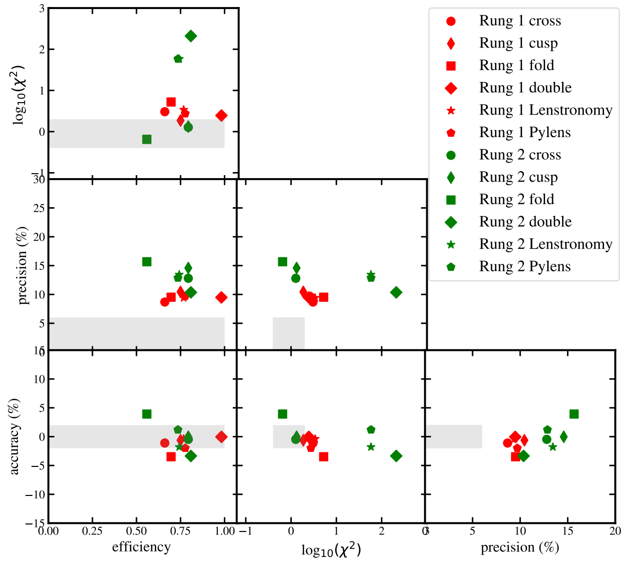

Figure 9. Figure illustrates the metrics of Rungs 1 and 2 according to different categories of lens systems using the entire submissions. Note that in Rung 2, the goodness (i.e. χ^2) is overwhelmingly dominated by the four *double* systems in the Freeform glassMulti's submission. Because these four double systems are simulated by LENSTRONOMY and PYLENS evenly, the corresponding $\log_{10}(\chi^2)$ in Rung 2 is significantly larger than the other ones. The larger goodness by Freeform team is an artefact due to the prior that H_0 is between 50 and 90 km s^{-1} Mpc^{-1}. The values which lie close to 50 have low error estimates (cut-off at 50), which results in very high χ^2 value.

the submission by the 'Good' teams. It is also encouraging that the ensemble averages show no evidence of bias, even though they are a little off the precision target. We note that these combined metrics are inferred after the unblinding in our TDLMC, but they are based on blind submissions. In future blind challenges, this kind of combined metrics could be built in from the start. We note that the *averaged metrics* are only introduced to help to 'guide the eye' to evaluate if there is 'wisdom in the crowd'. This is not a common practice in current research on this topic. Furthermore, the combined metrics are not representative and overweighting certain methods since different teams had different number of submissions.

A few trends emerge from these plots, discarding Student-T submission to Rung 2, and EPFL submission to Rung 1 for reasons discussed in the next subsections. First, most methods seem to have a realistic assessment of their uncertainties, landing on or close to the χ^2 target. Secondly, the methods constrained only by point source position and fluxes tend to produce significantly larger uncertainties than the target precision. Only the method using the full extent of the surface brightness of the host galaxy and the ancillary data hits the precision target. This trend can be confirmed by Table 6, in which the combined metrics of *precision* and *accuracy* are calculated in Rung 2 based on the algorithms using different levels of information. This finding is encouraging even though not surprising: using more data yields more precise results. Also encouraging is that even in the more challenging Rung 2 all the methods – including Student-T post blind – hit the accuracy target. Unexpectedly, the accuracy in Rung 1 seemed to have been less than in Rung 2. The improved accuracy in Rung 2 is likely due to the fact that the 'Good' teams learned from Rung 1's results to improve their algorithms and identify bugs in the codes.

To understand if the performance of the lens modelling is different between different lens configurations (i.e. *cross*, *cusp*, *fold*, and *double*) and simulating codes (i.e. LENSTRONOMY and PYLENS), we categorize the entire submissions and compare their metrics directly by plotting them together in Fig. 9. Interestingly, there is no significant evidence of difference between the different configurations (e.g. doubles and quads), which is an echo of the recent study by Birrer et al. (2019) that the precision of the cosmographic measurement with the doubly imaged AGNs could be comparable to those of quadruply imaged ones. Of course, this result should not be overinterpreted as the additional information content of the quads may just be not apparent in the configuration and regimes studied here, but relevant in other situations where, for example, the mass distribution is more complicated or the data quality is not as good, or the uncertainties are smaller. One potential explanation for the similarity is that the quads considered here are fairly more symmetric than the quads of the TDCOSMO collaborations, likely as a result of the selection function that favours systems with large ellipticity and shear since they have the highest cross-section for quads *versus* doubles. Symmetric quads have typically shorter time delays and less radial leverage when compared to more asymmetric ones, and thus provide weaker constraints on the Hubble constant. For all these reasons, the similarity between quads and doubles found in this challenge does not imply that they are equally efficient in reality. Also, the metrics are indistinguishable if we consider the LENSTRONOMY and PYLENS samples separately. This is true even if we restrict the comparison to the submissions by Student-T and EPFL teams, who used LENSTRONOMY. The lack of significant 'home advantage' is consistent with the fact that the difference of the simulated images between LENSTRONOMY and PYLENS is below the noise level (see Fig. A2).

Due to the limitations of Rung 3, as discussed in Section 5, we present the Rung 3 results in Appendix B.

4.2 Lessons form Rungs 1 and 2

The first important lesson is that the independent teams have come up with several independent techniques, including novel ones. As described above, the underlying assumptions of the techniques vary greatly, and so does the amount of information used by each technique and the flexibility of the models. As often the case in astrophysics, finding the right balance between too little and too much flexibility in the models is difficult yet vital to obtain accuracy and precision. Too little flexibility may lead to bias or underestimated error bars. Too much flexibility may lead to unphysical solutions or unnecessary

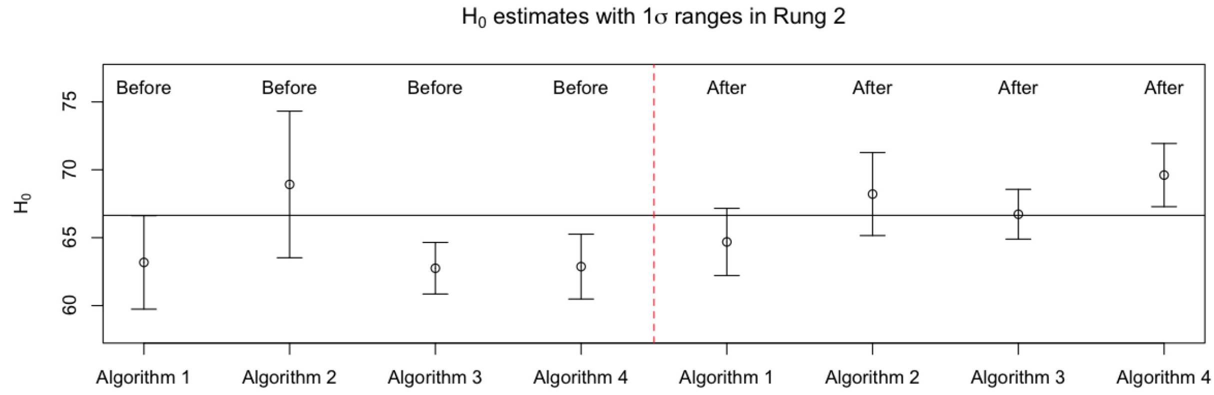

Figure 10. Post-blind improvement of the Student-T team's results using the correct PSF file for Rung 2 submissions, without changing any code or algorithm. This correction removes any bias in the inference and improves slightly the precision.

inflation of the error bars. The level of flexibility directly ties to another major obstacle to precision, lensing degeneracies. One way in which degeneracies can be quantified is by pulling multiple solutions from different families of models, and analysing the variance within that ensemble (see e.g. Gomer & Williams 2020; Saha 2000).

The second important lesson is that most methods seem to produce reasonable estimates of their uncertainties. In Rung 1, virtually all methods produced acceptable χ^2 metric distributions, while in Rung 2 the submissions that returned an answer for every system (i.e. high efficiency) sometimes paid the price in the sense that they underestimated their uncertainties.

The third important lesson is that more information translates to higher precision. Therefore, if one wishes to extract high precision from time delay measurements, it is crucial to use all the information available, not just the positions of the point sources (or their flux). However, an important caveat is that information content by itself does not necessarily guarantee accuracy if the modelling technique is not sufficiently flexible, as discussed above. Rungs 1 and 2 provide a useful test, but much remains to be done to explore the right degree of flexibility.

After unblinding Rung 1, the EPFL team discovered that small systematic uncertainties in the position of the multiply imaged quasars at the level of a fraction of a pixel could introduce a noticeable bias in the inference given the precision of the time delays. Thus, in Rung 2, the EPFL team introduced nuisance parameters to describe this uncertainty and marginalized over it. The effect is evident by comparing their blind results in Rungs 1 and 2. This is an example of the importance of modelling technique flexibility to ensure accuracy, and the fourth key lesson from Rungs 1 and 2 is that astrometric precision needs to be commensurate with the time delay precision. As discussed by Birrer & Treu (2019), the requirements can be at the level of milli-arcseconds if the time delay is known to percent precision. For *HST*-like images, the requirements correspond to a small fraction of a pixel, a challenging requirement for point sources superimposed on an extended and unknown source. It is thus important to consider explicitly this source of uncertainty and marginalize it, transforming a potential source of bias into a decrease in precision.

4.3 Notes about Student-T's submissions for Rungs 2 and 3

After unblinding, it was discovered that in Rung 2 (and Rung 3) the Student-T team used the non-drizzled PSF, drizzled lens image, and drizzled noise map, owing to clerical errors. The team's unblinded (post) analyses show that this mismatch was the main source of biases in the blinded analysis. In Fig. 10, we find that the Rung 2's result after using the correct file is much improved. The corresponding metrics of the post analysis are also given in Table 5. We stress that these post-submissions only corrects the input file; the modelling algorithms remain unchanged. These post-submissions are not used while calculating the combined (i.e. averaged) metrics.

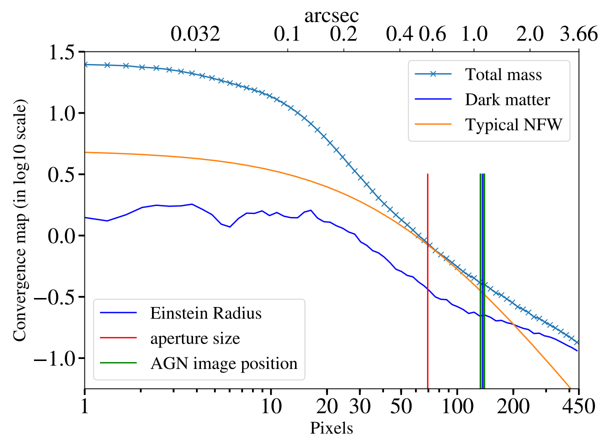

Figure 11. Mass profile of a typical deflector in Rung 3, illustrating the unphysical core in the central regions and the departure of the dark matter halo from a standard (Navarro et al. 1997) form.

5 LIMITATIONS OF RUNG 3, INCLUDING POST-UNBLINDING DISCOVERIES

Rung 3 was inconclusive because of the limitations of the procedure used to construct the lenses for this rung. We discuss here some of the limitations of the hydrodynamical simulations used to construct Rung 3. The 'Evil' team was aware of some of them while constructing the challenge, while others only became apparent post-unblinding. We introduce them in the following subsection.

5.1 Limitations known before unblinding

The main known limitations of the simulations pre-unblinding are twofold.

First, the resolution of the simulations we used is insufficient to describe the inner regions of early-type galaxies. This is illustrated in Fig. 11, where we show a typical mass profile, decomposed in dark and total mass. The total mass profile has a core of approximately 0.″1,

(a) Mock lensed arcs for Rung 2

(b) Mock lensed arcs for Rung 3

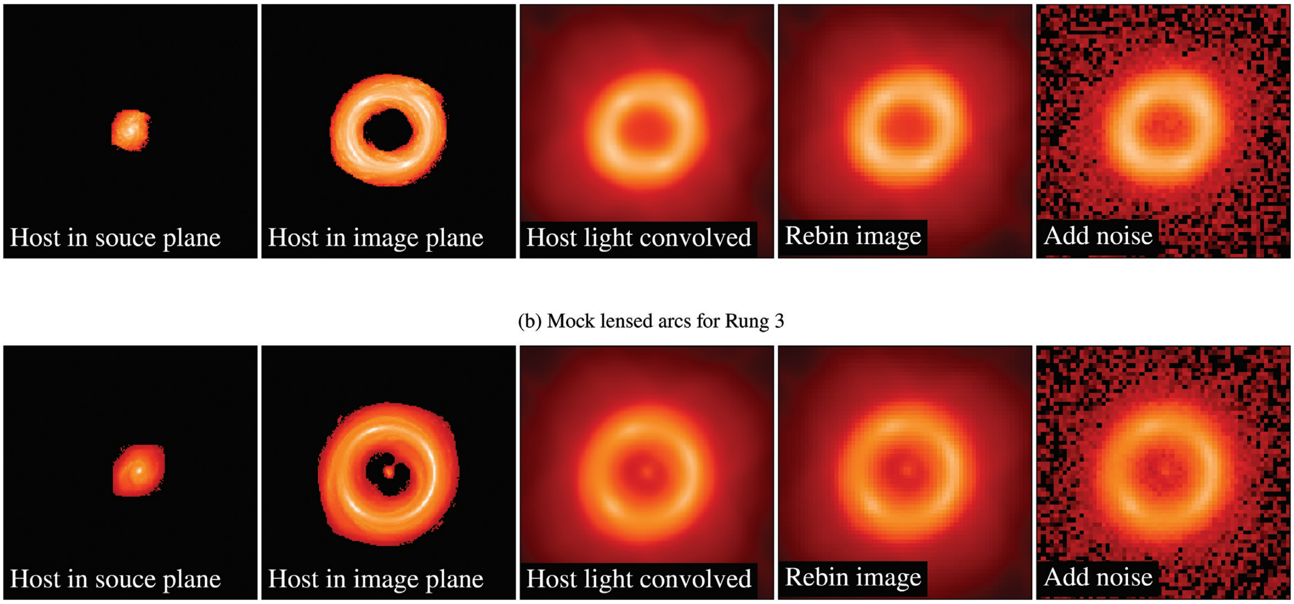

Figure 12. Mock images of the lensed arcs in the simulations of Rungs 2 and 3. Due to the unphysical core of Rung 3 deflector's mass profile, the lensed arcs appear the feature of 'central' lensed image (bottom-second plane). However, after entire simulation process this feature is not distinct anymore and would be overwhelmed by deflector light.

about half a kpc at the redshift of our sources. We also note that the adopted numerical simulations have softening lengths of 200–700 pc, which have partially contributed to the core sizes in these simulated galaxies. Despite that some cored massive elliptical galaxies have been found (Thomas et al. 2016) and could be produced in highly accurate dynamical simulations (Rantala et al. 2018), they are unlikely to be present in real lens galaxies with mass like Rung 3 ones. A recent detailed analysis of the mass density profiles of massive lens galaxies (Shajib et al. 2020a) in terms of stars and dark matter haloes shows that the dark matter halo is well described by a 'cuspy' unperturbed Navarro, Frenk & White (1997) halo and that the population of the lens galaxies' total mass density profile is close to a power-law profile (within ∼5 per cent near the Einstein radius). Although simulations have made a lot of progress in reproducing massive elliptical galaxies, Shajib et al. (2020a) show that they still fall short in simultaneously reproducing the mass density profile and the dark matter fraction of real galaxies at the level of detail needed for this test.

The main evidence against cores is from the search for central images of gravitational lenses themselves. The central slope of the mass density profile controls the magnification of the central image. The fact that the central image is almost always absent in galaxy scale lenses (not in clusters-scale lenses), is a strong argument against cores. For example, radio observations (e.g. Rusin & Ma 2001; Keeton 2003; Winn, Rusin & Kochanek 2004; Boyce et al. 2006; Zhang et al. 2007; Quinn et al. 2016) usually present a non-detection of the 'central' lensed image, which gives an upper limit of the core (<5∼100 pc). Likewise, in the TDCOSMO project, which models the high-resolution lensed AGN images based on *HST* observations, the fifth image has not been detected, although as we show below at optical/infrared wavelengths contamination by the deflector light limits the sensitivity.

A simple 'gedanken experiment' shows that the cores present in the Rung 3 simulations are unphysical, and therefore justifies our caution interpreting them. As shown in Fig. 12, Rung 3 predicts a central image, while Rung 2 does not. Unfortunately, in the optical and near-infrared such central image is difficult to disentangle from the light of the deflector.

In contrast, if we could perform the observations of the Rung 3 systems in the radio, assuming the multiple-imaged point source is radio loud, the test would be conclusive. The mean value of the magnification of the central source $\mu_{\rm c}$ for Rung 3's simulations is ∼0.032, which is significantly larger than the upper limit level reported by Keeton (2003, i.e. $\mu_{\rm c} < 0.001$). Furthermore, we calculated the ratio between the $\mu_{\rm c}$ and the magnification of the standard lensed point sources ($\mu_{\rm bright}$) and found that the mean value of $\mu_{\rm bright}/\mu_{\rm c}$ is ∼192, which is inconsistent with the values reported in the literature (Boyce et al. 2006; Zhang et al. 2007; Quinn et al. 2016, i.e. >2500, >1000, >10000, respectively). These results indicate that the core feature in Rung 3's simulations is not realistic.

The second argument to use Rung 3 with caution is that since simulations do not match perfectly the mass profile of real massive elliptical galaxies, as shown by Fig. 11, generalizing the results of such a test is always going to be complicated. For example, if the modellers were to assume the mass density profile to be cuspy in the inner regions and thus do not match the cores in the simulations, would this be a problem in analysing real galaxies, which should be cuspy? The recent study by Enzi et al. (2020) shows that without kinematic information, departures from a single power law (in this case, in the form of a core) can lead to a bias on the inference of H_0 of up to 25 per cent. A similar concern about the realism of simulations is illustrated by Xu et al. (2017), who analysed Illustris simulations and showed that the simulations do not match exactly the detailed properties of real galaxies in terms of central dark matter fraction and slope of the mass density profile (see also Shajib et al. 2020a; Wang et al. 2020).

These limitations were known to the 'Evil' team while designing the challenge. The 'Evil' team considered these limitations a 'necessary evil', to be kept in mind in the interpretation of the results. Indeed, when simulating the mock images, the 'Evil' team was aware that the Rung 3's lensed arcs demonstrated the fifth image feature, compared to Rung 2's simulation, see Fig. 12. However, this feature is not detectable in the simulated images due the contamination from the deflector light (the fifth image flux ratio is < 0.1 per cent, compared to the deflector light).

In the end, the benefit of knowing the three-dimensional 'truth' for a complex system was considered to outweigh the downside of the system not being fully realistic.

Future challenges may want to pursue some form of empirically driven models (perhaps based on observations of local massive elliptical galaxies) until the fidelity of simulations improves significantly.

5.2 Limitations discovered post-unblinding

Additional limitations were discovered post-unblinding because of collaborative efforts by the 'Evil' and 'Good' teams. However, these limitations do not necessarily invalidate the mock data or introduce a major bias to 'Good' team's inference of H_0.

5.2.1 Substructure and dynamics

In Rung 3, 12/16 simulations dynamically bound substructures (i.e. satellite haloes) were identified and removed before producing the lensing quantities. This procedure renders the kinematics inconsistent with the lensing quantities because the motion of the stars and gas was pre-computed based on the full mass distribution including substructure. Substructure accounts for approximately 1 per cent of the total mass at the relevant scales, so this is not a large effect, but can potentially introduce a bias at the percent level when combining lensing and kinematic tracers.

5.2.2 Halo truncation

For computational reasons, only the particles within the virial radius (R_{200}) or twice the virial radius were considered when projecting the mass distribution to calculate lensing quantities. This introduces two main outcomes. First, not taking into account mass beyond R_{200} may introduce a negative mass-sheet transform, biasing H_0 below the percent level. Secondly, the spherical truncation at R_{200} does not follow the isodensity contours of the mass profile, introducing an artificial shear (Van de Vyvere et al. 2020). At this radius, the artificial shear created by the truncation is small and may bias H_0 by less than 1 percent. Both truncation effects (i.e. artificial shear introduction and negative mass-sheet bias) have low amplitude for truncation at the virial radius. They then may introduce a small bias on the H_0 inference but should not be the major cause of bias in Rung 3 results.

6 DISCUSSION AND IMPLICATIONS FOR FUTURE WORK

First of all, a positive outcome of the challenge is that several teams were able to analyse a sample of 48 lenses, the sample size needed to reach subpercent precision (Shajib et al. 2018). Analysing this large sample within the time constraints of the challenge required good teams to apply fast methods as opposed to the more time and resource consuming approaches of state-of-the-art analysis of real data. These fast methods are necessary to make progress, and it is essential to test them as we did in the challenge. We note that even with the fast methods participation to the challenge was labor intensive, and the 'Evil' team extended the original deadlines set in TDLMC1 by a few months in order to allow more 'Good' teams to participate.

Rungs 1 and 2 demonstrate that current fast lens modelling technology is able to obtain precise and accurate estimates of H_0 starting from a best guess of the PSF, when using the information content of *HST*-like images. The expected complexity of the lensed host galaxy of the quasar is not an obstacle to the inference, provided that sufficiently flexible models are used to describe the source. The common practice of reconstructing the PSF starting from an empirical or theoretical best guess and the use of flexible source description is validated by the two rungs and should become the standard in future work.

Astrometry of the point sources from *HST*-like images can be a source of bias at the few percent level for extremely precise time delays. Mitigation strategies include adding nuisance parameters to describe the astrometric noise arising from poor sampling, or using higher resolution images, e.g. from adaptive optics or radio interferometers.

The conclusions about modelling the gravitational potential of the deflector are not so clear cut. Encouragingly, the teams performed well when the deflector was described by a simply parametrized analytic forms as in Rungs 1 and 2, with no evidence of inaccuracy. As discussed above, and as expected, the fast methods using more information performed better in terms of precision than the ones which used only AGN positions and flux ratios. Rung 3 was helpful in unveiling subtle effects that need to be considered if one wishes to use simulations to test gravitational lens modelling techniques for cosmological inference to high precision. Unfortunately, the same limitations – and the known limitations in resolution and realism at the beginning of the challenge – make it difficult to draw conclusions based on it. More work is needed on this front, and it will require either much higher resolution simulations than the ones adopted here or more advanced computational techniques to calculate the lensing quantities. Alternatively, a future challenge could find a way to generate high precision and realistic models, perhaps inspired by empirical data on local massive elliptical galaxies.

7 SUMMARY AND CONCLUSIONS

In this paper, we described the main results of the TDLMC. We first revealed some of the details of the construction of the simulated data sets that were kept blind during the challenge. Secondly, we gave a brief description of the methods followed by the 'Good' teams to do the inference. Thirdly, we described a number of limitations of Rung 3, including some numerical effects discovered post-unblinding that preclude inferences at the percent level required for this challenge. These limitations make Rung 3 difficult to interpret but are reported here with the aim to inform future challenges. Finally, we presented an overview of the performance of the methods against four metrics (precision, accuracy, efficiency, and goodness of fit).

The main conclusions, based on Rungs 1 and 2, can be summarized as follows:

(i) Each team came with fundamentally different methods to study a large sample of systems. In particular, methods constrained only by point-like images and using either analytic or free-form models, a novel Bayesian technique assuming a locally Gaussian Fermat potential, and modelling similar to current cosmographic analyses. A BNN approach has also been applied on unblinded data. Several teams developed fast methods that allowed them to analyse 48 lenses within the duration of this challenge (∼1–2 yr). This is a much larger number of systems per investigator time than the current state-of-the-art models, that so far require of order ∼1 year per system (not considering the process of collecting ancillary data and analysing the lens environment).

(ii) The fast methods applied to this challenge estimate their uncertainty appropriately, yielding error bars that are statistically comparable with the departure from the truth.

(iii) The fast methods that exploit the full information content of the data achieve higher precision than the ones that only utilize lensed quasars positions and fluxes to constrain the models.

(iv) The fast methods based on full image reconstruction can meet the target precision (6 per cent per system) and accuracy (2 per cent) when analysing mock images based on complex sources and starting with a guess of the PSF.

(v) Astrometric requirements on the position of the point sources can be stringent and difficult to meet for high precision time delay measurements, given the *HST* PSF and pixel size. Biases arising from the poor sampling of the PSF can be avoided by modelling the astrometric noise explicitly.

As far as Rung 3 is concerned, one generic problem was known before the challenge, i.e. if simulations do not reproduce real galaxies at the percent level precision in gravitational potential, it is difficult to generalize the outcome of the challenge. A good example of this issue is the finite resolution of cosmological hydrodynamical resolution, which introduces features like cores that are unlikely to be present in real systems. A spherical redistribution of cusp to core would not itself affect lensing observables, but it would change kinematic and other properties. If modellers assume that galaxies are cuspy, and do not detect the core in the simulations, what does it mean for real galaxies? The following additional and more subtle effects were identified post-unblinding.

(i) The kinematics of the particles in the simulations must be consistent to sub-percent level with the gravitational potential generated by the lensing data products given to the 'Good' teams. Removing substructures or other parts of the simulation when generating the lensing data may cause internal tension in the data so that the lensing and dynamical probes cannot be combined without bias.

(ii) The standard practice of truncating simulated haloes at the virial radius may lead to inconsistencies between the actual Fermat potential and the one computed from truncated maps. Lensing quantities such as the Fermat potential are non-local, and the kernel mapping convergence into potential is logarithmic. Therefore, in order to avoid biases in Fermat potential at the few percent level, one has to include all particles well beyond the virial radius and carefully consider the shape of the truncation.

In recent years, a number of works have investigated the systematic uncertainties in time-delay cosmography (e.g. Schneider & Sluse 2013; Birrer et al. 2016; Sonnenfeld 2018; Kochanek 2020; Millon et al. 2020). However, it is difficult to make a quantitative comparison between our results and those in the literature because the uncertainties depend strongly on the assumptions and methods used.

To conclude, this work shows that blind challenges on simulated data are a powerful tool to study and characterize a method, alongside blind and independent analysis of real data sets (Millon et al. 2020). The results obtained from this first TDLMC are encouraging, in the sense that accurate and precise H_0 can be derived blindly even in the presence of complex sources and unknown PSF. However, our results also demonstrate that much work remains to be done before we can have conclusive end-to-end tests based on simulations. First, state-of-the-art modelling methods exploiting the full information content of the data need to speed up so that even larger simulated data sets can be analysed within a practical time frame to explore a variety of more complicated configurations. For example, the EPFL team that used all the information employed 500 000 CPU hours and 1700 h of investigator time, almost a full year equivalent. This is significantly less time than currently employed per lens by H0LiCOW or STRIDES. However, the challenge was single plane and by design simpler in terms of satellites and perturbers along the LOS than real lenses. So, in order to analyse samples of order 100–1000 lenses with increased complexity, further speed-ups are necessary.

Secondly, improvements in numerical simulations of massive elliptical galaxies and the calculation of their lensing properties are needed before they can be used to perform lens modelling challenges to percent level precision.

ACKNOWLEDGEMENTS

We thank Kenneth C. Wong and Sherry H. Suyu for useful suggestions and supports.

TT acknowledges support by the Packard Foundation in the form of a Packard Research Fellowship. TT and CDF acknowledge support by NSF through grant 'Collaborative Research: Toward a 1 per cent Measurement of The Hubble Constant with Gravitational Time Delays' AST-1906976. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 787886). CDF and GC.-FC acknowledge support for this work from the National Science Foundation under grant agreement Nos AST-1312329 and AST-1907396. This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (COSMICLENS: grant agreement No 787886) and the Swiss National Science Foundation (SNSF). AJS was supported by the National Aeronautics and Space Administration (NASA) through the Space Telescope Science Institute (STScI) grant agreement No HST-GO-15320. This research was supported by the U.S. Department of Energy (DOE) Office of Science Distinguished Scientist Fellow Program.

HT acknowledges Simon Birrer for his sincere and tireless support on implementing LENSTRONOMY, which has enabled the team's participation in the challenge. HT also appreciates Xuheng Ding for his thoughtful comments on the team's PYTHON code submitted to the Evil team, which has significantly improved the methodology (including the discovery of the clerical error). In addition, HT acknowledges computational supports from the Institute for Computational and Data Sciences at Pennsylvania State University.

SRK and HC acknowledge financial support from SERB, DST, Govt. of India through grant agreement No PDF/2016/003848 during the course of this project. SH acknowledges support by the DFG cluster of excellence 'Origin and Structure of the Universe'. SV thanks the Max Planck Society for support through a Max Planck Lise Meitner Group, and acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (LEDA: grant agreement No 758853).

At the moment of this writing the revised version of the manuscript following the referee's comments, the mock data of TDLMC have already been used as a validating set for subsequent analyses (Birrer et al. 2020).

DATA AVAILABILITY

The data underlying this article are available in the TDLMC website, at https://tdlmc.github.io/.

REFERENCES

Abdelsalam H. M., Saha P., Williams L. L. R., 1998, MNRAS, 294, 734
Akeret J., Seehars S., Amara A., Refregier A., Csillaghy A., 2013, Astron. Comput., 2, 27
Alam S. et al., 2017, MNRAS, 470, 2617
Arendse N. et al., 2020, A&A, 639, A57
Auger M. W., Treu T., Brewer B. J., Marshall P. J., 2011, MNRAS, 411, L6
Barkana R., 1998, ApJ, 502, 531
Bartelmann M., 2010, Class. Quantum Gravity, 27, 233001
Betoule M. et al., 2014, A&A, 568, A22
Birrer S., Amara A., 2018, Phys. Dark Universe, 22, 189
Birrer S., Treu T., 2019, MNRAS, 489, 2097
Birrer S., Amara A., Refregier A., 2015, ApJ, 813, 102
Birrer S., Amara A., Refregier A., 2016, J. Cosmol. Astropart. Phys., 2016, 020
Birrer S. et al., 2019, MNRAS, 484, 4726
Birrer S. et al., 2020, A&A, 643, A165
Boyce E. R., Winn J. N., Hewitt J. N., Myers S. T., 2006, ApJ, 648, 73
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Choi E., Ostriker J. P., Naab T., Johansson P. H., 2012, ApJ, 754, 125
Ciotti L., Bertin G., 1999, A&A, 352, 447
Colbert J. W. et al., 2013, ApJ, 779, 34
Coles J. P., Read J. I., Saha P., 2014, MNRAS, 445, 2181
Collett T. E. et al., 2013, MNRAS, 432, 679
de Vaucouleurs G., 1948, Ann. Astrophys., 11, 247
Denzel P., Mukherjee S., Coles J. P., Saha P., 2020, MNRAS, 492, 3885
Denzel P., Coles J. P., Saha P., Williams L. L. R., 2021, MNRAS, 501, 784
Ding X. et al., 2017a, MNRAS, 465, 4634
Ding X. et al., 2017b, MNRAS, 472, 90
Ding X. et al., 2018, preprint (arXiv:1801.01506)
Dobler G., Fassnacht C. D., Treu T., Marshall P., Liao K., Hojjati A., Linder E., Rumbaugh N., 2015, ApJ, 799, 168
Eisenstein D. J. et al., 2005, ApJ, 633, 560
Enzi W., Vegetti S., Despali G., Hsueh J.-W., Metcalf R. B., 2020, MNRAS, 496, 1718
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. et al., 2001, ApJ, 553, 47
Freedman W. L. et al., 2019, ApJ, 882, 34
Frigo M., Naab T., Hirschmann M., Choi E., Somerville R. S., Krajnovic D., Davé R., Cappellari M., 2019, MNRAS, 489, 2702
Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., Rubin D. B., 2013, Bayesian Data Analysis. CRC Press, Boca Raton, FL
Ghosh A., Williams L. L. R., Liesenborgs J., 2020, MNRAS, 494, 3998
Gilman D., Agnello A., Treu T., Keeton C. R., Nierenberg A. M., 2017, MNRAS, 467, 3970
Gomer M., Williams L. L. R., 2020, J. Cosmol. Astropart. Phys., 2020, 045
Greene Z. S. et al., 2013, ApJ, 768, 39
Grillo C., Lombardi M., Bertin G., 2008, A&A, 477, 397
Hezaveh Y. D., Levasseur L. P., Marshall P. J., 2017, Nature, 548, 555
Hilbert S., White S. D. M., Hartlap J., Schneider P., 2007, MNRAS, 382, 121
Hilbert S., Hartlap J., White S. D. M., Schneider P., 2009, A&A, 499, 31
Hu C.-Y., Naab T., Walch S., Moster B. P., Oser L., 2014, MNRAS, 443, 1173
Jee I., Komatsu E., Suyu S. H., 2015, J. Cosmol. Astropart. Phys., 2015, 033
Jee I., Komatsu E., Suyu S. H., Huterer D., 2016, J. Cosmol. Astropart. Phys., 2016, 031
Keeton C. R., 2003, ApJ, 582, 17
Knox L., Millea M., 2020, Phys. Rev. D, 101, 043533
Kochanek C. S., 2020, MNRAS, 493, 1725
Koopmans L. V. E. et al., 2009, ApJ, 703, L51
Krist J. E., Hook R. N., Stoehr F., 2011, 20 Years of Hubble Space Telescope Optical Modeling Using Tiny Tim. p. 81270J
Leier D., 2009, MNRAS, 400, 875
Liao K. et al., 2015, ApJ, 800, 11
Mamon G. A., Łokas E. L., 2005, MNRAS, 363, 705
Merritt D., 1985a, AJ, 90, 1027
Merritt D., 1985b, MNRAS, 214, 25P
Metcalf R. B., Petkova M., 2014, MNRAS, 445, 1942
Millon M. et al., 2020, A&A, 639, A101
Mukherjee S., Koopmans L. V. E., Metcalf R. B., Tortora C., Schaller M., Schaye J., Vernardos G., Bellagamba F., 2019, preprint (arXiv:1901.01095)
Naab T., Ostriker J. P., 2017, ARA&A, 55, 59
Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
Oguri M., 2010, PASJ, 62, 1017
Oser L., Ostriker J. P., Naab T., Johansson P. H., Burkert A., 2010, ApJ, 725, 2312
Osipkov L. P., 1979, Pisma Astron. Zh., 5, 77
Paraficz D., Hjorth J., 2010, ApJ, 712, 1378
Park J. W., Wagner-Carena S., Birrer S., Marshall P. J., Yao-Yu Lin J., Roodman A., 2020, preprint (arXiv:2012.00042)
Perlmutter S. et al., 1999, ApJ, 517, 565
Petkova M., Metcalf R. B., Giocoli C., 2014, MNRAS, 445, 1954
Planck Collaboration XVI, 2014, A&A, 571, A16
Planck Collaboration XIII, 2016, A&A, 594, A13
Planck Collaboration VI, 2020, A&A, 641, A6
Quinn J. et al., 2016, MNRAS, 459, 2394
Rantala A., Johansson P. H., Naab T., Thomas J., Frigo M., 2018, ApJ, 864, 113
Refregier A., 2003, MNRAS, 338, 35
Refsdal S., 1966, MNRAS, 132, 101
Riess A. G. et al., 1998, AJ, 116, 1009
Riess A. G. et al., 2016, ApJ, 826, 56
Riess A. G., Casertano S., Yuan W., Macri L. M., Scolnic D., 2019, ApJ, 876, 85
Rusin D., Ma C.-P., 2001, ApJ, 549, L33
Rusu C. E. et al., 2017, MNRAS, 467, 4220
Saha P., 2000, AJ, 120, 1654
Saha P., Williams L. L. R., 2004, AJ, 127, 2604
Schneider P., 2006, in Meylan G., Jetzer P., North P., Schneider P., Kochanek C. S., Wambsganss J., eds, Saas-Fee Advanced Course 33: Gravitational Lensing: Strong, Weak and Micro. Springer, p. 1
Schneider P., Sluse D., 2013, A&A, 559, A37
Schneider P., Sluse D., 2014, A&A, 564, A103
Schechter P. L. et al., 1997, ApJ, 475, L85
Sersic J. L., 1968, Atlas de galaxias australes. Observatorio Astronomico, Cordoba, Argentina
Shajib A. J., Treu T., Agnello A., 2018, MNRAS, 473, 210
Shajib A. J. et al., 2019, MNRAS, 483, 5649
Shajib A. J., Treu T., Birrer S., Sonnenfeld A., 2020a, preprint (arXiv:2008.11724)
Shajib A. J. et al., 2020b, MNRAS, 494, 6072
Sonnenfeld A., 2018, MNRAS, 474, 4648
Springel V., 2005, MNRAS, 364, 1105
Suyu S. H., Marshall P. J., Auger M. W., Hilbert S., Blandford R. D., Koopmans L. V. E., Fassnacht C. D., Treu T., 2010, ApJ, 711, 201
Suyu S. H. et al., 2012, preprint (arXiv:1202.4459)
Suyu S. H. et al., 2014, ApJ, 788, L35
Tagore A. S., Barnes D. J., Jackson N., Kay S. T., Schaller M., Schaye J., Theuns T., 2018, MNRAS, 474, 3403
Tak H., Mandel K., van Dyk D. A., Kashyap V. L., Meng X.-L., Siemiginowska A., 2017, Ann. Appl. Stat., 11, 1309
Tak H., Ghosh S. K., Ellis J. A., 2018, MNRAS, 481, 277
Tak H., You K., Ghosh S. K., Su B., Kelly J., 2020, J. Comput. Graph. Stat., 29, 659

Thomas J., Ma C.-P., McConnell N. J., Greene J. E., Blakeslee J. P., Janish R., 2016, Nature, 532, 340
Treu T., 2010, ARA&A, 48, 87
Treu T., Koopmans L. V. E., 2002a, MNRAS, 337, L6
Treu T., Koopmans L. V. E., 2002b, ApJ, 575, 87
Treu T., Koopmans L. V. E., 2004, ApJ, 611, 739
Treu T., Marshall P. J., 2016, A&AR, 24, 11
Van de Vyvere L., Sluse D., Mukherjee S., Xu D., Birrer S., 2020, A&A, 644, A108
Verde L., Treu T., Riess A. G., 2019, Nat. Astron., 3, 891
Vogelsberger M. et al., 2014, Nature, 509, 177
Vogelsberger M., Genel S., Sijacki D., Torrey P., Springel V., Hernquist L., 2013, MNRAS, 436, 3031
Wagner-Carena S., Park J. W., Birrer S., Marshall P. J., Roodman A., Wechsler R. H., 2020, preprint (arXiv:2010.13787)
Wang Y. et al., 2020, MNRAS, 491, 5188
Williams L. L. R., Saha P., 2000, AJ, 119, 439
Winn J. N., Rusin D., Kochanek C. S., 2004, Nature, 427, 613
Wong K. C. et al., 2017, MNRAS, 465, 4895
Wong K. C. et al., 2020, MNRAS, 498, 1420
Xu D. D. et al., 2009, MNRAS, 398, 1235
Xu D., Sluse D., Schneider P., Springel V., Vogelsberger M., Nelson D., Hernquist L., 2016, MNRAS, 456, 739
Xu D., Springel V., Sluse D., Schneider P., Sonnenfeld A., Nelson D., Vogelsberger M., Hernquist L., 2017, MNRAS, 469, 1824
Yuan W., Riess A. G., Macri L. M., Casertano S., Scolnic D. M., 2019, ApJ, 886, 61
Zhang M., Jackson N., Porcas R. W., Browne I. W. A., 2007, MNRAS, 377, 1623

APPENDIX A: ORIGINAL DRAFT OF TDLMC1

In order to allow the reader to reconstruct exactly the information that was available to participants before unblinding, we reproduce in this appendix the original draft of the manuscript of TDLMC1 (Ding et al. 2018), as it was posted on arXiv on 2018 January 4 to open the challenge. To minimize duplications, we omit the text of section 3 of the TDLMC1 manuscript (A4 in this appendix), which is repeated in Section 2.7 of this paper.

Title: Time delay lens modeling challenge: i. experimental design

X. Ding, T. Treu, A. J. Shajib, D. Xu, G. C.-F. Chen, A. More, G. Despali, M. Frigo, C. D. Fassnacht, D. Gilman, S. Hilbert, P. J. Marshall, D. Sluse, S. Vegetti

A1 Abstract

Strong gravitational lenses with measured time delay are a powerful tool to measure cosmological parameters, especially the Hubble constant (H_0). Recent studies show that by combining just three multiple-imaged AGN systems, one can determine H_0 to 2.4 per cent precision. Furthermore, the number of time-delay lens systems is growing rapidly, enabling, in principle, the determination of H_0 to 1 per cent precision in the near future. However, as the precision increases it is important to ensure that systematic errors and biases remain subdominant. For this purpose, challenges with simulated data sets are a key component in this process. Following the experience of the past challenge on time delay, where it was shown that time delays can indeed be measured precisely and accurately at the sub-percent level, we now present the 'TDLMC' (TDLMC). The goal of this challenge is to assess the present capabilities of lens modelling codes and assumptions and test the level of accuracy of inferred cosmological parameters given realistic mock data sets. We invite scientists to model a set of simulated *HST* observations of 50 mock lens systems. The systems are organized in rungs, with the complexity and realism increasing going up the ladder. The goal of the challenge is to infer H_0 for each rung, given the *HST* images, the time delay, and a stellar velocity dispersion of the deflector, for a fixed background cosmology. The TDLMC challenge starts with the mock data release on 2018 January 8. The deadline for blind submission is different for each rung. The deadline for Rungs 0 and 1 is 2018 September 8, the deadline for Rung 2 is 2019 April 8, and the one for Rung 3 is 2019 September 8. This first paper gives an overview of the challenge including the data design, and a set of metrics to quantify the modelling performance and challenge details. After the deadline, the results of the challenge will be presented in a companion paper with all challenge participants as co-authors.

A2 Ingredients of the simulations

We briefly introduce the key ingredients including deflector/source surface brightness and deflector mass for simulating the lens image in Sections A2.1 and A2.2, respectively.

A2.1 Surface brightness

To study how results change with increasing complexity, we adopted a variety of approaches to simulate the brightness profiles of the lens and source galaxies.

(1) As a matter of convenience, in the entry level of the challenge, we choose a common simply parametrized description of the surface brightness for the lens and source galaxy. This choice is meant primarily for testing the codes, both for 'Evil' and 'Good' Teams. In the literature, the Sérsic profile (Sersic 1968) is one of the most commonly used models to describe the surface brightness of galaxies. It ranges from exponential discs to de Vaucouleurs (1948) profiles.

The Sérsic profile is parametrized by

$$I(R) = A \exp\left[-k\left(\left(\frac{R}{R_{\rm eff}}\right)^{1/n} - 1\right)\right], \tag{A1}$$

$$R(x, y, q) = \sqrt{qx^2 + y^2/q}, \tag{A2}$$

where A is the amplitude and Sérsic index n controls the shape of the radial surface brightness profile; a larger n corresponds to a steeper inner profile and a highly extended outer wing. k is a constant that depends on n so as to ensure that the isophote at $R = R_{\rm eff}$ encloses half of the total light (Ciotti & Bertin 1999) and q denotes the axial ratio.

(2) For the bulk of the challenge, we use more realistic and complex surface brightness distribution for the host galaxy of the lensed AGN. For example, we use real images of galaxies appropriately smoothed and cleaned by foreground/background contaminants as shown in Fig. A1.

A2.2 Deflector mass

Likewise, we achieve different levels for complexity of the challenge by increasing the realism of the deflector mass distribution stepping up the ladder.

(1) A common simply parametrized description of the deflector mass density profile is given by elliptical power-law models whose

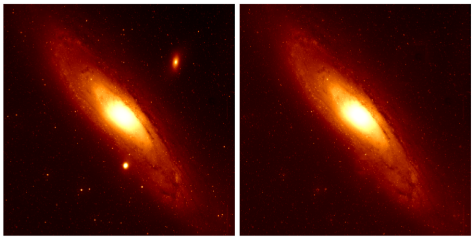

Figure A1. Illustration of real *HST* galaxy image as lensed source host. The bright point source and foreground/background galaxies in the original image (left) are replaced by interpolation of nearby pixels to obtain a clean galaxy image (right).

surface mass density is given by

$$\Sigma(x, y) = \Sigma_{cr} \frac{3-\gamma'}{2} \left(\frac{\sqrt{q_m x^2 + y^2/q_m}}{R_{\rm E}} \right)^{1-\gamma'}, \quad (A3)$$

q_m described the projected axial ratio. The so-called Einstein radius $R_{\rm E}$ is chosen such that, when $q_m = 1$ (i.e. spherical limit), it encloses a mean surface density equal to $\Sigma_{\rm cr}$. The exponent γ' is the slope of the power-law profile, for massive elliptical galaxies $\gamma' \approx 2$ (Treu & Koopmans 2002b, 2004; Koopmans et al. 2009). We refer the reader to the reviews by Schneider (2006), Bartelmann (2010), and Treu (2010) for more details.

(2) In order to achieve a more realistic deflector mass distribution, we also consider massive early-type lens galaxies produced by cosmological numerical simulations. We only consider a single deflector, and do not include the effects of the LOS other than via the external convergence introduced before. We choose systems with virial mass approximately 10^{13} $M_\odot$ yielding Einstein radii of order 1 arcsec for typical source and deflector redshift ($z_{\rm d} \approx 0.5$) and ($z_{\rm s} \approx 1.5$).

A3 Structure of the challenge

In this section, we first describe the data sets that are made available to the 'Good' Teams in Section A3.1. Then, a description of the layers (rungs) of the challenge is given in Section A3.2.

A3.1 Data sets

The mock data available to the 'Good' teams consist of deep *HST* images, time delays, stellar velocity dispersion, and external convergence, as described below. The released mock data sets have been tested by analysing a subset of them with two independent lens modelling software and verifying that the input cosmology (and lens parameters when applicable) could be recovered within the uncertainties.

(1) In order to mimic a typical observational set-up in state of the art observations, we choose to simulate high-resolution images obtained with the *HST*, using the Wide Field Camera 3 (WFC3) IR channel in the F160W band. Even though this set-up has lower resolution than optical images taken with WFC3-UVIS or ACS, we adopt it in order to minimize the effects of dust extinction and optimize the contrast between the (blue) AGN and (red) host galaxy. We adopt a range of AGN to host flux ratios so as to produce a distribution similar to that observed in real systems (Ding et al. 2017a,b). For simplicity, we do not include any dust extinction, and we assume AGN to be at the centre of the host galaxy. Also, multiband

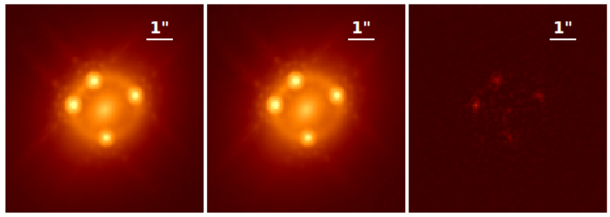

Figure A2. The left-hand and middle panels illustrate the simulated *HST*-like images based on two independent codes with same lens parameters. The right-hand panel shows the difference on the same scale. The pixel scale is 0.″08 after drizzling.

data sets or adaptive optics assisted ground-based images are left for future challenges.

In practice, the following steps are taken in order to simulate realistic lens configurations.

(a) For every set of lens and source parameters, compute high-resolution images of the lensed host, and deflector light.
(b) Convolve with the PSF appropriate for WFC3/F160W.
(c) Compute the image plane positions and fluxes of the lensed AGN images and add them as appropriately scaled PSF in the image plane.
(d) Rebin the oversampled images to the actual data resolution. Using different rebinning patterns, one can simulate eight dithering images in order to drizzle them[9] in step (vi).
(e) Add noise based on realistic observation condition, including background, readnoise, and Poisson noise from the source. The exposure time of each one of the eight images is taken to be 1200 s, thus the total exposure time is 9600 s.
(f) Drizzle the individual images to recover some of the resolution lost due to pixelization. Following common practice, we drizzle eight images into one final image; the corresponding pixel size is 0.″13 and 0.″08, before and after drizzling. This step introduces correlated noise. In order to allow 'Good' Teams to model the original data, the eight non-drizzled images of one lens system are provided in addition to the final drizzled image.

A detailed description of these steps is given by Ding et al. (2017a, section 3). In order to control for numerical issues and for implicit bias in favour of any 'Good' Team, we use two independent codes to generate the simulations (half the sample with each code). An example of mock images generated by two independent codes with the same parameters is shown in Fig. A2. The noise maps for the images are provided that contain the standard deviation of the noise.

(2) Once the values of lens parameters are set, the difference of the Fermat potential between the AGN images can be calculated with equation (2). To calculate the corresponding time delay with equation (1), we need to assume a set of cosmological parameters. For simplicity, we draw values randomly from a uniform distribution between 50 and 90 km s^{-1} Mpc^{-1} for the Hubble Constant, assuming a flat ΛCDM cosmological model with $\Omega_{\rm m} = 1 - \Omega_\Lambda = 0.27$. We then add measurement uncertainty to the time delay. As discussed in the introduction, the time delay is supposed to be known with sufficient precision and accuracy so that we can test the precision and accuracy of the models. We thus assume zero bias and the smallest random errors that can be obtained with current monitoring strategies.

[9]MULTIDRIZZLE is adopted for the drizzling, see http://www.stsci.edu/hst/wfpc2/analysis/drizzle.html for more information.

Thus, we adopt as random error the largest between 1 per cent and 0.25 d.

(3) In principle, all mass along the LOS contributes to the deflection of light rays. In practice, however, it is often the case that the lensing configuration can be approximated by a single main deflector with the addition of external shear and convergence ($\kappa_{\rm ext}$). The latter is particularly important because it does not change the image positions and relative fluxes, but it affects the relative Fermat potential, and hence the time delay, according to the equation (4). If not accounted for, the $\kappa_{\rm ext}$ can bias the inference of time delay distance and thus H_0. A common practice to constrain the $\kappa_{\rm ext}$ is to compare the distribution of mass along the LOS with numerical simulations (Hilbert et al. 2009; Suyu et al. 2010; Colbert et al. 2013; Collett et al. 2013; Greene et al. 2013; Rusu et al. 2017).

Since the focus of this challenge is single plane lens modelling, we include the effects of the LOS contribution in the following simplified manner. We randomly generate a $\kappa_{\rm ext}$ using a random Gaussian distribution with 0 and 0.025 as mean value and standard deviation, respectively. This, from the point of view of modelling, one can adopt a prior on $\kappa_{\rm ext}$ of 0 ± 0.025. The uncertainty is chosen to represent well-characterized lines of sight, and corresponds to a random uncertainty of 2.5 per cent on H_0, to first approximation.

(4) Stellar kinematic information is essential for breaking the mass-sheet degeneracy and constraining the lensing potential. Furthermore, the measurement of stellar kinematics provides extra cosmological information (Grillo, Lombardi & Bertin 2008; Jee, Komatsu & Suyu 2015; Jee et al. 2016; Shajib et al. 2018). Thus, we provide the model deflector velocity dispersion in addition to the *HST* images, time delays, and $\kappa_{\rm ext}$. An integrated LOS velocity dispersion is computed by weighting the velocity field by the surface brightness in a square aperture with 1 arcsec on a side. Typical seeing condition is rendered by convolving the surface-brightness-weighted LOS velocity dispersion image with a Gaussian kernel with a FWHM of 0.″6.

Following current practice (Wong et al. 2017; Shajib et al. 2018), a random Gaussian noise with 5 per cent standard deviation is added to the model velocity dispersion to account for typical measurement errors.

A3.2 Rungs

Lens modelling is usually time-consuming both in terms of human and computer time. Thus, the size of simulated samples is limited by practical considerations. Based on the experience of the evil team and consultations with members of the lensing community a sample size of 50 was considered a good compromise between practicality and the need to explore different conditions with sufficient statistics to uncover potential biases. Thus, we construct the challenge in the following manner. Similar to the TDC we provide an entry level zeroth rung for format checking and testing purposes. The zeroth rung consists of two simple lenses. If the teams can successfully recover H_0 from the zeroth rung, they are encouraged to participate and submit their results for rungs consisting of the real challenge. Each rung consists of 16 lenses. For each rung, 16 systems are simulated including *cusp, fold, cross*, and *double* configurations; four examples for each configuration are generated using two independent codes.

Considering the quality of the data simulated here, constraints on H_0 with a precision of ∼6 per cent should be possible and thus 48 systems would deliver H_0 to sub-percent precision that would be sufficient to uncover biases at this level. We set a global value of H_0 per rung. To ensure that the 'Good' Teams do not infer any information for the previous rung, we reset H_0 at each rung. The complexity and realism of the systems increase with rung level, thus allowing us to separate different aspects of the lens models and understand what needs improvement. Partial submissions for a subset of the rungs will be accepted.

Detailed information for each rung's design including the lens components and data provided to the 'Good' team is given in the following subsections.

A3.2.1 Rung 0 This rung is a training exercise that consists of two lens systems, one two-image (*double*) and one four-image (*quad*) configuration. The goal of this rung is to ensure that 'Good' Team members understand the format of the data and that no bugs or mistakes could potentially affect the results of the challenge for a specific method.

In view of this goal, parametric models for surface brightness model and mass profile are selected. We adopt a single Sérsic profile to describe the surface brightness of both the lens and source galaxy, and elliptical power-law models for the lens surface mass density. Also, we randomly added an external shear to the lens potential drawn from a typical range. The AGN images are added as a point sources and the PSF is provided. The lens parameters and cosmological parameters for the simulations are released with the data for the modelling team to check.

A3.2.2 Rung 1 This rung is meant to be the easiest one of the actual challenging ladder. Thus, the mocks in Rung 1 are generated in a similar way as in Rung 0, except that we use the images of real galaxies to get realistic surface brightness distribution for the lensed AGN host and the time delays are affected by external convergence (i.e. equation 4). For Rungs 0 and 1, we also provide an oversampled PSF; the pixel size is 0.″13/4 = 0.″0325. This is to mimic the oversampling that is generally achieved by combining several stars in the science image.

A3.2.3 Rung 2 This rung is meant to test PSF reconstruction features of lensing codes, in addition to the aspects tested in Rung 1. For this purpose, we only provide a guess of the PSF but not the one actually used to generate the data.

A3.2.4 Rung 3 This rung is the highest level in this challenge, and thus the simulations are intended to be the most realistic. In addition to all the complexity we have adopted for Rungs 1 and 2, the observables are generated using massive early-type galaxies selected from numerical cosmological simulations.

A4 Instructions for participation and evaluation metrics

The text of this section is omitted to avoid duplication. The information can be found in section 2.7 or in section 3 of the TDLMC1 on arXiv (Ding et al. 2018).

A5 TDLMC1 Summary

We presented the TDLMC. The structure of the challenge is as follows. The 'Evil' Team produced a set of mock lenses, meant to mimic state-of-the-art data quality. Anyone in the community is invited to participate as a 'Good' Team, by modelling the data and submitting a blind estimate of the Hubble Constant. The 'Evil' Team will compute four metrics aimed at quantifying the accuracy and precision of the estimates. The overall goal of the challenge is to assess whether current lens modelling techniques are sufficient to ultimately reach a 1 per cent measurement of H_0. The challenge is organized in rungs in order to help identify aspects of the lens modelling effort that may represent bottlenecks and may require additional improvements.

APPENDIX B: DETAILS OF RUNG 3

B1 Illustris simulations

The first group of simulated galaxies is selected from the Illustris simulation (Vogelsberger et al. 2013, 2014) with six galaxies at $z = 0.4$ and six galaxies at $z = 0.6$. All have total dark matter halo masses between 1 and 2 ($10^{13}\,M_\odot$), and velocity dispersion ranging from 250 to 320 km s^{-1}. In this challenge, we do not intend to test biases in the most severe cases where the true profiles significantly deviate away from the power-law models. For this reason, our selection was based on the fact that the selected galaxies shall distribute fairly closely around the best-fitting general mass–velocity dispersion relation. As a result, the majority of the selected galaxies are not classified as the extreme cases of deviations from power-law mass distributions; the most severe case would result in an underestimate of Hubble constant by 15 per cent (see Xu et al. 2016).

The convergence and potential maps (as well as potential's first and second derivatives) were calculated using netted-mesh based methods through FFT with an isolated boundary condition. All matter distribution of the selected galaxy halo is truncated at R200 with a spherical aperture (Xu et al. 2009). The results have been cross-checked with the public software GLAMER, which is a ray-tracing code for the simulation of gravitational lenses Metcalf & Petkova (2014) and Petkova, Metcalf & Giocoli (2014). In addition, we also calculated the same maps using a mesh-based FFT algorithm, adopting smoothed-particle hydrodynamics (SPH) kernel to smooth the simulated particles to the mesh. The two sets of results showed expected consistency within the numerical uncertainties.

The velocity maps were calculated on desired meshes; here no smoothing was used. The pixel values of mean velocity and velocity dispersions were weighted by rest-frame SDSS-*r*-band luminosities of stellar particles projected to the pixel.

B2 Zoom simulations

The second set of simulations is a sample of 'zoom' cosmological simulations, which have been previously used in Frigo et al. (2019). A 'zoom' simulation is a higher resolution re-run of a small part of the cosmological box of a large-scale simulation (like Illustris), called the 'parent' simulation. In the set we employed, the parent simulation is a 100 Mpc wide cosmological box simulated with dark matter only (Oser et al. 2010), and each zoom simulation covers the volume of a dark matter halo (at $z = 0$). The simulations were run with a modified version of GADGET2 (Springel 2005), called SPHGAL (Hu et al. 2014), which avoids some of the shortcomings of SPH codes. Unlike the parent, the zoom simulations also include gas, stars, and black hole particles. They include models for star formation (based on gas density and temperature), metal enrichment, gas cooling, stellar winds, supernova feedback (Type Ia and Type II), and AGN feedback [using the Choi et al. (2012) model]. The spatial resolution (softening length) of the simulation is 200 pc, while the mass resolution (initial mass of gas particles) is $7 \times 10^5\,M_\odot$. This is a higher resolution than Illustris, but not high enough to avoid the issues presented in Section 5. The simulations run from $z = 43$ to $z = 0$. The sample of simulated galaxies varies in mass, size, dynamical, and stellar-population properties. For the TDLMC project, we used snapshots at different redshifts ($0.3 < z < 0.5$) of the four most massive AGN galaxies, which have arcsec-size Einstein radii. More details on the simulation code and on this sample can be found in Frigo et al. (2019).

Table B1. Metrics of blind submission for Rung 3.

Team	Algorithm	f	$\log(\chi^2)$	$P(\%)$	$A(\%)$
	Metrics of Rung 3				
Student-T	Algorithm1	0.750	0.117	15.616	− 3.803
Student-T	Algorithm2	0.812	− 0.583	26.226	6.221
Student-T	Algorithm3	0.938	0.459	8.472	1.677
Student-T	Algorithm4	1.000	0.213	12.869	2.512
Student-T	Algorithm5	0.875	0.402	11.998	− 11.998
Student-T	Algorithm6	0.938	− 0.932	26.515	3.986
Student-T	Algorithm7	1.000	0.718	4.885	− 5.415
Student-T	Algorithm8	1.000	0.027	12.587	− 2.786
Student-T	Algorithm9	0.875	0.532	8.247	− 7.373
Student-T	Algorithm10	0.938	− 0.848	15.369	4.401
Student-T	Algorithm11	1.000	1.132	3.923	− 5.065
Student-T	Algorithm12	1.000	0.115	9.728	− 1.195
EPFL	Combined	0.438	0.893	4.276	− 9.963
EPFL	CombinedDdtOnly	0.438	0.879	4.584	− 9.944
EPFL	Composite	0.500	1.515	2.612	− 11.302
EPFL	CompositeDdtOnly	0.500	1.500	2.559	− 11.403
EPFL	Powerlaw	0.812	0.938	2.941	− 7.016
EPFL	PowerlawDdtonly	0.812	0.955	3.001	− 6.973
Freeform	GlassMulti	1.000	2.464	5.106	− 16.041
Freeform	GlassSingleHiRes	1.000	1.954	5.809	− 17.267
Freeform	GlassSingleLowRes	1.000	1.401	9.632	− 11.441
Freeform	PixelensMulti	1.000	− 0.695	18.866	7.626
Freeform	PixelensSingle	1.000	− 0.226	21.637	0.542
H0rton	Bayesian neural network	0.312	0.637	9.056	3.356

Note. Table summarizes the metrics of the blind submission for Rung 3.

The maps of convergence, lensing potential, and its derivatives were calculated with the post-processing ray-tracing code HILBERT (Hilbert et al. 2007, 2009). The whole high-resolution region of each simulation, roughly reaching out to twice the virial radius of the galaxy, was fed into the code and used to calculate the lensing maps. The 3D orientation of the galaxy was chosen randomly before the analysis. The kinematic maps were calculated on the same grid as the lensing maps, weighting the LOS velocity of each particle with its *R*-band luminosity.

B3 Rung 3 results

For completeness, we report here the full results of Rung 3. We caution the reader that the interpretation of these results is difficult because of the limitations and numerical issues described in Section 5.

The metrics of each submission for Rung 3 are listed in Table B1 and plotted in Figs B1 and B2.

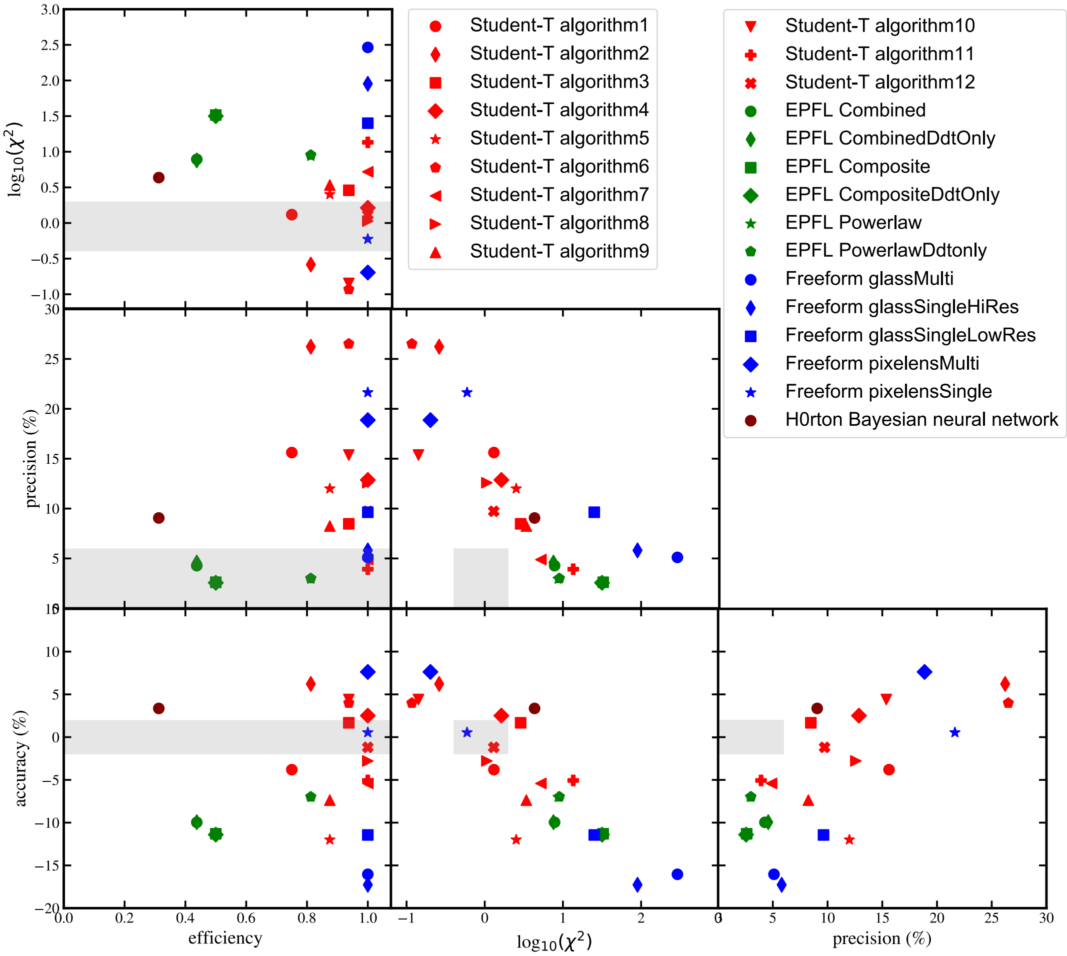

Figure B1. Metrics of Rung 3 blind submissions. Note that Rung 3 was affected by issues described in Section 5 and thus great caution should be taken in interpreting these results.

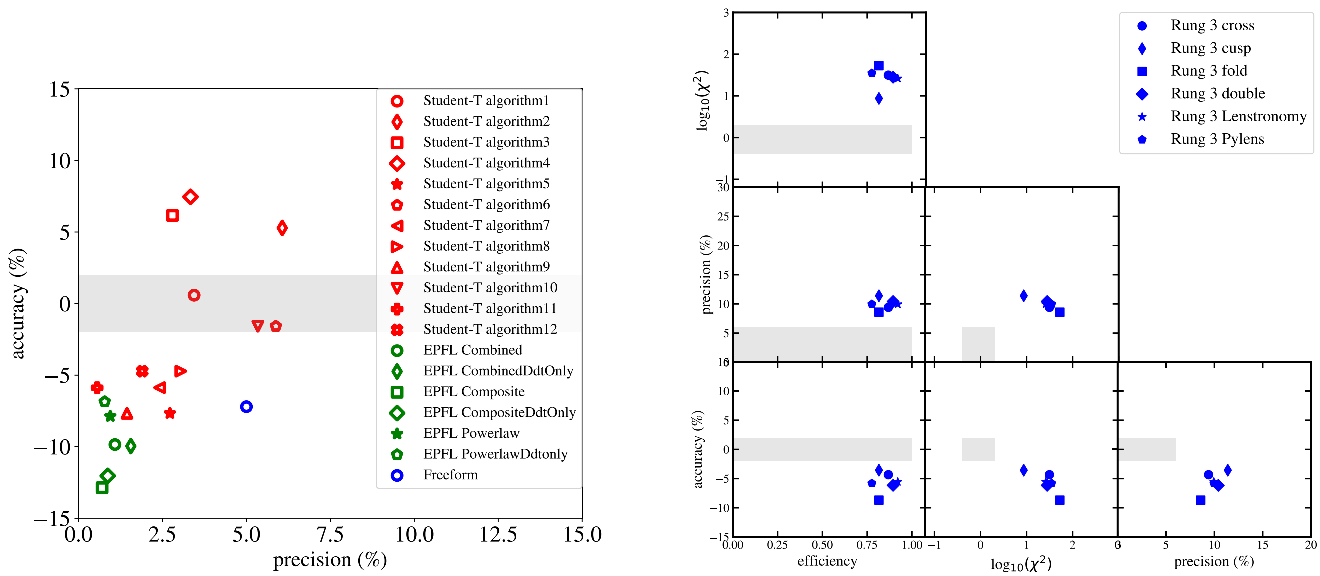

Figure B2. Panel (left) and (right) is the same as Fig. 8 and 9, separately, but for Rung 3.

[1]*Department of Physics and Astronomy, University of California, Los Angeles, CA 90095-1547, USA*
[2]*Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan*
[3]*Kavli Institute for Particle Astrophysics and Cosmology and Department of Physics, Stanford University, Stanford, CA 94305, USA*
[4]*Department of Physics, University of California, Davis, CA 95616, USA*
[5]*Physik-Department T38, Technische Universität München, James-Franck-Str 1, D-85748 Garching, Germany*
[6]*Institute for Computational Science, University of Zurich, CH-8057 Zurich, Switzerland*
[7]*Physics Institute, University of Zurich, CH-8057 Zurich, Switzerland*
[8]*Max Planck Institute for Astrophysics, Karl-Schwarzschild-Strasse 1, D-85740 Garching, Germany*
[9]*Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, CH-1290 Versoix, Switzerland*
[10]*Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 452 Lomita Mall, Stanford, CA 94035, USA*
[11]*Department of Astronomy & Astrophysics, University of Chicago, Chicago, IL 606374, USA*
[12]*STAR Institute, Quartier Agora - Allée du six Août, 19c, B-4000 Liège, Belgium*
[13]*International CHASC Astrostatistics Collaboration, Harvard University, Cambridge, MA 02138, USA*
[14]*Center for Astrostatistics, The Pennsylvania State University, University Park, PA 16802, USA*
[15]*Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA*
[16]*Department of Astronomy and Astrophysics, The Pennsylvania State University, University Park, PA 16802, USA*
[17]*Institute for Computational and Data Sciences, The Pennsylvania State University, University Park, PA 16802, USA*
[18]*Department of Astronomy, Tsinghua University, Beijing 100084, China*
[19]*Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[20]*Department of Astronomy, Aryabhatta Research Institute of observational sciencES (ARIES), Nainital 263002, India*
[21]*Department of Physics and Astronomical Sciences, Central University of Himachal Pradesh, Dharamshala 176206, India*
[22]*Zentrum für Astronomie der Universität Heidelberg, Institut für Theoretische Astrophysik, Albert-Ueberle-Str 2, D-69120 Heidelberg, Germany*
[23]*Exzellenzcluster Universe, Boltzmannstr 2, D-85748 Garching, Germany*
[24]*Ludwig-Maximilians-Universität, Universitäts-Sternwarte, Scheinerstr 1, D-81679 München, Germany*
[25]*Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, IL 61801, USA*
[26]*School of Physics and Astronomy, University of Minnesota, 116 Church Street SE, Minneapolis, MN 55455, USA*

This paper has been typeset from a TEX/LATEX file prepared by the author.