THE UNIVERSITY OF CHICAGO

STRONG GRAVITATIONAL LENSING SCIENCE IN THE ERA OF LARGE ASTRONOMICAL SURVEYS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ASTRONOMY AND ASTROPHYSICS

BY

WEI JIE JASON POH

CHICAGO, ILLINOIS

JUNE 2024

To my parents

Poh Soon Hoong and Toh Lily

"It is good to have an end to journey toward; but it is the journey that matters, in the end."

— Ursula K. Le Guin, The Left Hand of Darkness

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation was a difficult labor, and its completion would not have been possible if not for the support and help of a great number of people.

First, I would like to thank my advisor, Josh Frieman. I am far from a perfect student – thank you for your infinite patience and steady guidance throughout my tumultuous journey. My thesis committee, Mike Gladders, Richard Kron, Brian Nord, and Angela Olinto, were incredibly supportive throughout my PhD – thank you for believing in me.

I have been fortunate to work with many amazing people throughout the years, and I would like to thank these three in particular for their mentorship – Brian Nord (again!), Aleksandra Ciprijanovic, and Elizabeth Buckley-Geer. A lot of the material in this dissertation could not have happened without their input and advice.

I would also like to sincerely thank all the administrative staff at the University of Chicago, past and present, for being the lifeblood of the Astronomy department and KICP – Laticia Rebeles, Jennifer Smith, Brent Barker, Souvik De, Julia Brazas, Val Smith, Juliet Crowell, Galen Tsongas, Helen Pates, Aimee Giles, Ted Russell, Maureen Lowery, Catherine Burroughs, Carolyn Topper and Valeri and Elena Galtsev. Special thanks to Letty for being an amazing student affairs administrator and advocate for us!

The Deep Skies Collaboration has been an integral part of my time here, and I would like to acknowledge my collaborators in various projects – Ashwin Samudre, Sreevani Jarugula, and Rebecca Nevin for all the great conversations we had throughout our collaboration. There are a lot of other Deep Skies members, past and present, that I would like to thank for making it a great community: Amanda Pagul, Joao Caldeira, Sam McDermott, Ashia Lewis-Livaudais, Maggie Voetberg, Yuanyuan Zhang, Jason St. John, John Wu and Marina Dunn (I'm sure there are others I forgot – apologies!).

The community of friends I have made while at the University of Chicago is truly exceptional, and I have so many people to thank for being both great colleagues and friends:

Gourav Khullar and James Lasker (the OG ERC 450 office-mates!), Nora Shipp, Phil Mansfield, Andrew Neil, Amy Tang, Mandy Chen, Emily Simon, Maya Mallaby-Kay, Celeste Keith, Elyssa Brooks, Grace Chesmore, Fei Xu, Rebecca Diesing, Jordan Scherer, Emily Gilbert, Andrea Bryant, John Hood, Tracy Hood, David Zegeye, Abby Lee, Lucas Secco, Christine Simpson, Junran Yang, Alex Alarcon, Judit Prat, Jose Maria Ezquiaga, Anowar Shajib, Giulia Giannini, Dan Baxter, Katie Harrington, Ariane Dekker, Daniele Rogantini, Luca Orusa, Ani Chiti, Yiming Zhong, Tanvi Karwal, Hayley Macpherson, Anatasia Sokolenko, Christoph Welling, Georgios Valogiannis, Akhil Premkumar, Maria Steinrueck, Elena Pinetti, TK Chan and Michael Zhang.

I would also have not made it this far without the support of close friends and family. To my closest friends from home – Hua Sheng, Ji Wei, Lien, Zhuang Mao, Valerie – and my family – Mom, Dad, Lisa, Lilian, Nolan, Kayley, and Liam – I am sorry for missing so much of your lives by moving half a world away for more than a decade. I miss you all. Thank you always for your love and patience.

ABSTRACT

Strong gravitational lensing occurs when light from a distant source is deflected by a massive object along its line of sight to an observer, causing the source image to appear multiply imaged or distorted to the observer. The morphology of strong gravitational lensing systems depends not only on the projected mass profile of the intervening lens object but also on a combination of geometric distances between the observer, lens, and source, making them useful probes of dark matter and dark energy. Many strong lenses have been found through previous searches in wide-field surveys, and the advent of future large-scale cosmic surveys will usher in a new epoch of strong lensing discovery. In this thesis, I will present several related works involving strong gravitational lenses discovered or imaged in astronomical surveys. First, I present the spectroscopic and photometric characterization of the environment around two time-delay gravitational lenses discovered in Dark Energy Survey data. Then, I present the mass and light properties of Four Early-type galaxy-scale strong lenses discovered in the Dark Energy Survey. Finally, I present the application of two different novel deep learning techniques to strong lensing analysis in simulated wide-field survey imaging data.

CHAPTER 1

INTRODUCTION

Gravitational lensing posits that massive objects warp the fabric of spacetime, bending the path of light rays traversing through their vicinity. This gravitational deflection acts as a cosmic magnifying glass, distorting and amplifying the light emitted by distant sources. Over the decades, gravitational lensing has transcended from a theoretical curiosity to a pivotal tool in astronomy. At its core, gravitational lensing embodies the intricate dance between geometry and matter. The gravitational field of a massive object, whether it be a star, a galaxy, or even a large cluster of galaxies, serves as the lensing agent, bending and refracting light from background sources. Through mathematical formulations grounded in General Relativity and optical theory, astronomers have crafted a rich framework to describe gravitational lensing phenomena across a gargantuan range of mass and distance scales.

Observationally, gravitational lensing manifests in a myriad of observational phenomena, each offering a glimpse into fundamental astrophysical and cosmological insights. From arcs and Einstein rings to the mirages of multiply-imaged source galaxies or quasars, the observational signatures of gravitational lensing serve as powerful astrophysical and cosmological probes of the distribution of matter and light across cosmic time scales, enabling the study of star and structure formation, as well as shedding light on the elusive nature of dark matter and dark energy.

Modern astronomical surveys, armed with advanced instrumentation and computational techniques, have also recorded a large and comprehensive census of lensing phenomena, enabling large-scale statistical studies on the population of these lensing systems. By dissecting the statistical properties of lensing phenomena, cosmologists can glean invaluable insights into the distribution of matter on cosmic scales, constraining cosmological parameters with unprecedented precision.

1.1 A Very Brief History of Gravitational Lensing

The inception of gravitational lensing dates back to Albert Einstein's formulation of general relativity in 1915. In general relativity, massive objects warp the fabric of spacetime, influencing the trajectories of nearby particles and light rays. Building upon this framework, Karl Schwarzschild and Johannes Droste independently derived the first exact solution to Einstein's field equations in 1916, describing the gravitational field around a spherically symmetric mass. While these early developments laid the groundwork for understanding gravity's effects on light, it remained largely untested until 1919, when a team led by Arthur Eddington observed a solar eclipse. Eddington's team aimed to test Einstein's prediction of the deflection of starlight by the Sun's gravitational field, and the successful confirmation of this effect provided empirical support for general relativity and hinted at the potential application of gravitational lensing.

Meanwhile, theoretical discussions about gravitational lensing began to emerge, with Swiss astrophysicist Fritz Zwicky speculating in 1937 about the potential for massive galaxies to act as gravitational lenses, leading to the formation of multiple images of distant sources (Zwicky, 1937). Sjur Refsdal proposed the first scientific application of gravitational lenses, suggesting the use of time-delayed images from lensed supernovae to infer the Hubble Constant, which characterizes the local rate of cosmological expansion of the universe (Refsdal, 1964).

The discovery of the twin quasar Q0957+561 by Walsh, Carswell, and Weymann in 1979 (Walsh et al., 1979) marked a watershed moment in the field of gravitational lensing astronomy. This remarkable system, featuring two nearly identical quasar images separated by a few arcseconds, provided compelling observational evidence of gravitational lensing on extragalactic scales. The subsequent decades witnessed a surge of observational campaigns aimed at identifying and characterizing gravitational lensing phenomena, facilitated by advances in imaging techniques and the deployment of ground-based and space-based observatories.

Notable discoveries during this period included the identification of Einstein rings, multiple-image quasars, and giant arcs, each offering unique insights into the distribution of matter in the universe.

The turn of the millennium heralded a new era of gravitational lensing research, driven by technological innovations and cosmological implications. Ground-based observatories such as the Keck Observatory and space-based telescopes like the Hubble Space Telescope enabled high-resolution imaging of gravitational lensing systems, unveiling intricate details of distant galaxies and quasars distorted by intervening mass concentrations. Moreover, large-scale astronomical surveys such as the Sloan Digital Sky Survey (SDSS) and the Dark Energy Survey (DES) systematically mapped the distribution of gravitational lenses, facilitating statistical analyses and cosmological inference. Machine learning techniques have also played a pivotal role in automating the detection, classification, and modeling of gravitational lensing systems, revolutionizing the field and paving the way for large-scale cosmological studies.

As we look toward the future, gravitational lensing remains a fertile ground for exploration and discovery. With the advent of next-generation observatories and ambitious survey projects, including the Vera C. Rubin Observatory and the James Webb Space Telescope, we anticipate further advancements in our understanding of gravitational lensing phenomena and their cosmological implications. Through interdisciplinary collaborations between theoretical physicists, astronomers, and computer scientists, we strive to unlock the full potential of gravitational lensing as a tool for probing the fundamental properties of the universe.

1.2 The Basics of Gravitational Lensing

General Relativity predicts that when light from a distant source passes near an intervening massive object (the lens), the mass causes light to deflect around the object. An observer would detect this in the form of angular deflections in the observed image of the source object. The deflection is given by the lens equation,

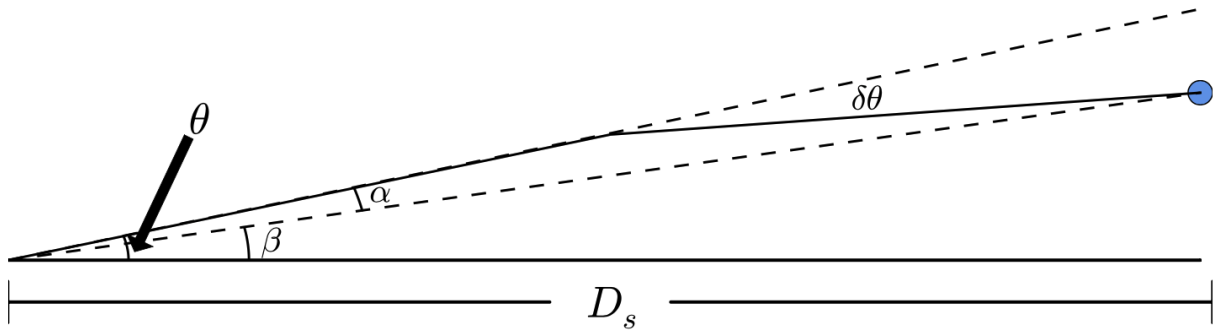

Figure 1.1 The geometry of gravitational lensing systems described by Eqn. 1.1. Light from a distant source at distance D_s would appear at angle β if not deflected. However, the deflection from gravitational lensing causes the source to appear as if it were at angle θ. The angular difference between β and θ is α. (Figure from (Dodelson, 2017)).

$$\beta = \theta - \alpha, \tag{1.1}$$

where β describes the angular coordinates of a source object if it were not lensed and θ is the angular coordinates of the lensed object as measured by an observer. The difference between the two angles – caused by gravitational lensing – is the deflection angle, α. Confusingly, there is another angle denoted by $\delta\theta$ in Figure 1.1 that is also commonly referred to as the deflection angle in the scientific literature. To disambiguate the two, we follow the prescription by Schneider et al. (2013) and call α the *scaled deflection angle*. Defining D_l as the distance from the observer to the lens, D_{ls} as the distance between the lens and source, and D_s the distance from the observer to the source, and using the small angle approximation, the true position of the source is related to its observed position and the deflection angle by

$$\beta D_s = \theta D_s - \delta\theta D_{ls}, \tag{1.2}$$

By substituting Eqn. 1.1 into Eqn. 1.2, we find that the scaled deflection angle is related to the deflection angle by

$$\alpha \equiv \frac{D_{ls}}{D_s}\,\delta\theta \tag{1.3}$$

For practical reasons, β, θ, α and $\delta\theta$ are usually described in terms of two-dimensional coordinates vectors in some observed image plane that is perpendicular to the line of sight, and so we denote these angles as coordinate vectors $\hat{\beta}$, $\hat{\theta}$, $\hat{\alpha}$ and $\hat{\delta\theta}$ going forward.

The geometry described in Figure 1.1 appears rather idealized, assuming that all the deflection occurs at one fixed point along the line of sight. This is known as the *thin lens approximation*. This approximation is valid in the regime where the physical size of the lensing agent is much smaller than the distances between observer, lens, and source. Remarkably, this condition is valid in virtually all strongly lensed gravitational lensing phenomena, including lensing by massive galaxy clusters. In this approximation, the lensing effect of any arbitrary three-dimensional mass distribution can be fully described without loss of generality by a two-dimensional projection of that mass distribution orthogonal to the line of sight. Mathematically, this *surface mass density*, Σ, is

$$\Sigma(\hat{\xi}) \equiv \int p(\hat{\xi}, z)\, dz \tag{1.4}$$

where p is the three-dimensional mass distribution, $\hat{\xi} \equiv (\xi_1, \xi_2)$ are the two coordinate vectors on the lens plane perpendicular to the line of sight, and z is the coordinate vector along the line of sight. The deflection angle $\hat{\delta\theta}$ at position $\hat{\xi}$ is given by the integral of all mass elements in the lens plane up to $\hat{\xi}$

$$\hat{\delta\theta}(\hat{\xi}) = \frac{4G}{c^2} \int \Sigma(\hat{\xi'}) \frac{(\hat{\xi} - \hat{\xi'})}{|\hat{\xi} - \hat{\xi'}|^2}\, d^2\hat{\xi'} \tag{1.5}$$

where G is the gravitational constant and c is the speed of light[1]. Using the small angle

1. The factors outside the integral are derived from calculating the deflection of a light ray in a gravitational potential in General Relativity - this is a factor of 2 larger than the Newtonian prediction

approximation $\hat{\xi} \approx \hat{\theta} D_l$, we can do a change of variable to Eqn. 1.5 to obtain

$$\hat{\delta\theta}(\hat{\theta}) = \frac{4GD_l}{c^2} \int \Sigma(\hat{\theta'}) \frac{(\hat{\theta} - \hat{\theta'})}{|\hat{\theta} - \hat{\theta'}|^2} \, d^2\hat{\theta'} \tag{1.6}$$

Finally, we relate it to the scaled deflection angle by substituting Eqn. 1.3 to get

$$\hat{\alpha}(\hat{\theta}) = \frac{4GD_l D_{ls}}{c^2 D_s} \int \Sigma(\hat{\theta'}) \frac{(\hat{\theta} - \hat{\theta'})}{|\hat{\theta} - \hat{\theta'}|^2} \, d^2\hat{\theta'} \tag{1.7}$$

The reciprocal of the constant outside the integral of Eqn. 1.7 defines a *critical surface mass density*[2] $\Sigma_{cr} \equiv \frac{c^2 D_s}{4\pi G D_l D_{ls}}$ that characterizes the dividing line between strong and weak lensing. For strong lensing phenomena, which is the main subject of this thesis, we are interested in regimes where $\Sigma(\hat{\theta'}) \geq \Sigma_{cr}$.

The fact that $\hat{\alpha}(\hat{\theta})$ is a function of $\hat{\theta}$ is what gives rise to the interesting properties of lensing systems because it makes Eqn. 1.1 generally non-linear. There can be multiple solutions for $\hat{\theta}$ in the lens equation, implying that a single source object at angular coordinates $\hat{\beta}$ may map to multiple image positions $\hat{\theta}$, resulting in multiply-imaged objects. Another important and useful feature of gravitational lensing is that in the absence of emission and absorption of photons along the lensing path, the source surface brightness is conserved due to Liouville's theorem. However, the solid angle of the source $\delta\hat{\beta}^2$ can be mapped onto a larger (smaller) observed solid angle, $\delta\hat{\theta}^2$, resulting in magnification (demagnification). This has useful astrophysical applications, which we discuss in the next section.

In summary, solving the lens equation is central to lensing science. However, analytic solutions exist for only the simplest parametric lens mass models. Fortunately, there is now a wealth of literature describing the mathematical and numerical formalism of lensing (e.g., Narayan & Bartelmann, 1996), cosmological and astrophysical applications (e.g., Treu,

2. Observant readers will notice an additional factor of π in the denominator of Σ_{cr} - this is a bookkeeping convention to account for the fact that most integrals of Eqn. 1.7 are over a circular or ellipsoidal angular surface.

2010), and solutions for some of the most common lens mass profiles used in lensing studies, such as Ellipsoidal models (e.g., Kassiola & Kovner, 1993; Kormann et al., 1994; Barkana, 1998). We use such models in our work and describe them in more detail at various points in later chapters (e.g., §3.3, §4.3, §5.2).

1.3 Applications of Strong Gravitational Lenses

The standard model of cosmology (*"concordance cosmology"*) posits that the universe consists of three major components:

1. **Baryonic Matter**: All ordinary forms of matter that interact by the standard model of particle physics.

2. **Dark Matter**: A hypothesized form of matter that interacts primarily through gravitational effects.

3. **Dark Energy**: A hypothesized form of vacuum energy that drives the accelerating expansion of the universe.

The properties of gravitational lenses make them attractive observational probes of all three major components of concordance cosmology. We provide some examples of astrophysical and cosmological applications of strong lenses to each component below.

1.3.1 Baryonic Matter

As previously mentioned, gravitational lenses can amplify the source light by increasing the observed surface angle of the source object. This has two primary benefits: (1) it increases the amount of light collected from that object, increasing the signal-to-noise ratio of the observed source flux and (2) it increases the apparent spatial size of the object, allowing

for greater observed spatial resolution. Each of these features can be exploited for different science applications.

The amplification of the source light from a galaxy or quasar from being lensed allows for the observation and analysis of high redshift sources that would otherwise be too faint to be imaged (e.g., Bayliss et al. (2014); Hainline et al. (2009); Quider et al. (2009); Smit et al. (2017); Dessauges-Zavadsky et al. (2009); Christensen et al. (2012); James et al. (2013); Yuan & Kewley (2009); Finkelstein et al. (2009); Iwata et al. (2008); Jones et al. (2013); Khullar et al. (2021)). This phenomenon can also be used on smaller stellar-mass scale lenses (where it is called *microlensing*) to detect and measure the masses of exoplanets and massive compact halo objects (see Mao (2012) and references within). Gravitational lensing also enables the study of spatially resolved distant lensed sources at resolutions down to tens of parsecs, enabling detailed studies of the morphology and astrophysics of localized regions in the source object, such as star-forming regions in distant source galaxies (e.g., Tremblay et al. (2014); Johnson et al. (2017)).

1.3.2 Dark Matter

In Eqn. 1.7, we explicated the direct relation between the extent of the observed deflection caused by gravitational lensing and the mass distribution of the lensing agent. Therefore, by modeling and reconstructing the observed lensing phenomena, we can infer the mass distribution of lensing systems. The total mass of most lensing objects is dominated by dark matter, making strong gravitational lenses well-established probes of the dark matter distributions in lensing systems at both galactic and extragalactic mass and distance scales. Another advantage of gravitational lenses is that they are a non-dynamic probe of matter, which means they do not rely on the motions of stars or galaxies in the lensing system to infer their mass.

Early-type galaxies constitute the majority of galaxy-scale lenses, and studies on samples

of Early-type galaxy lenses have yielded fruitful studies on their density profiles, dark matter mass fraction, density slope, and stellar initial mass function (e.g. Auger et al. (2010b); Barnabe et al. (2011); Newman et al. (2015, 2012a,b); Cardone et al. (2009)). To date, the mass distributions of more than a hundred early-type galaxy lenses have been studied systematically (e.g. Tan et al. (2023)), and the ever-growing catalog of modeled Early-type lenses across a range of redshifts will enable future studies of their evolutionary history. The lens modeling work in Chapter 3 is related to these efforts.

1.3.3 Dark Energy

As mentioned in §1.1, the first proposed application of gravitational lenses was to use time delays from multiply-imaged lensed supernovae (*"time-delay lenses"*) to infer the Hubble Constant (Refsdal, 1964). The theoretical basis of using gravitational lenses to infer cosmological parameters arises from the observation that the deflection angle in Eqn. 1.7 has a dependence on the distances between the observer, lens and source (recall the definitions of D_l, D_{ls}, D_s from §1.2). If the lens and source are at cosmological distances, then these distances have a cosmological dependence, and encode information about the cosmic expansion history of the universe. If we can measure D_l, D_{ls}, D_s, or some combination of those distances for a (or many) lensing system(s), then we can use the distances (or combination of distances) to infer the parameters of a cosmological model.

Time-Delay Lenses

Time-delay lenses remain the most mature and well-established gravitational lensing probe of cosmology (e.g. Suyu et al. (2010a, 2013, 2016); Bonvin et al. (2017); Birrer et al. (2016); Shajib et al. (2020)), and there is a large collaboration of scientists actively researching time-delay lenses under the aegis of TDCOSMO (Millon et al., 2020a).

The crux of time-delay cosmography is that the observed time delay between multiply-

imaged sources, Δt, depends on the *time-delay distance*

$$\Delta t \propto D_{\Delta t} \equiv \frac{D_l D_s}{D_{ls}} \tag{1.8}$$

By obtaining precise measurements of the time delays, coupled with well-fitting lens mass models and a good handle on the environment around and along the line-of-sight of the lens galaxy, we obtain a precise measurement of $D_{\Delta t}$, from which we can infer cosmological parameters using Bayesian inference. Chapter 2 describes published work that was conducted to characterize the environment of two lensed quasars (Buckley-Geer et al., 2020), and the results were used to obtain a 3.9% measurement of the Hubble constant (Shajib et al., 2020).

Emerging Strong Lensing Probes of Cosmology

One drawback of the time-delay technique is that obtaining precise time delays for each lens system requires long-term follow-up studies throughout many years (Millon et al., 2020b). This limits the number of such systems that can be studied. However, in recent years, two novel strong lensing probes of cosmology have emerged that do not have this requirement.

One such probe is double-source-plane lenses, which are systems that have two source objects at different redshifts lensed by the same lensing agent such that it produces two distinct characteristic scaled deflection angles, $\hat{\alpha_1}$ and $\hat{\alpha_2}$ (recall Eqn. 1.7), for each lensed source (Gavazzi et al., 2008; Collett & Auger, 2014; Collett et al., 2012). In these systems, taking the ratio of the two scaled deflection angles cancels out most of the terms in Eqn. 1.7, reducing to a combination of distances

$$\beta \equiv \frac{\hat{\alpha_1}}{\hat{\alpha_2}} = \frac{D_{ls_1} D_{s_2}}{D_{ls_2} D_{s_1}} \tag{1.9}$$

where D_{ls_1}, D_{ls_2} are the distances between the lens and each source, and D_{s_1}, D_{s_2} are the distances from the observer to each source.

These systems are highly informative, but are exceedingly rare, requiring a geometry where three objects are surreptitiously aligned along the same line of sight to the observer (instead of two). To date, only one system has been used to constrain cosmology, and fewer than 100 systems are forecasted to be discoverable by the Rubin observatory over 10 years (Collaboration et al., 2021).

The other novel probe is the so-called *"Distance Ratio"* method (Grillo et al., 2007; Cardone et al., 2015; Cao et al., 2015a). This method requires only a sample of galaxy-galaxy lenses, which are the most common lens configurations. By measuring the Einstein radius θ_E and a dynamical estimate of each lens (typically the central velocity dispersion of the stellar component σ_0), we can derive the distance ratio

$$\mathcal{R} \equiv \frac{D_{ls}}{D_s} \tag{1.10}$$

from which cosmological constraints can be extracted. While galaxy-scale lenses are abundant, the drawback of this method is that individually, each lens does not provide as much statistically constraining power as a single time-delay or double-source-plane lens. Existing studies have assembled catalogs of more than 100 lenses from previous lensing surveys, and follow-up programs such as the 4MOST Strong Lensing Spectroscopic Legacy Survey (Collett et al., 2023) aim to measure spectroscopic redshifts and velocity dispersions for 5000 more lenses from future cosmic surveys, enabling this method to be a competitive probe of cosmology in the future. However, this and other statistical studies of strong lenses present a scaling challenge, which we discuss in the next section.

1.4 Prospects in the Era of Large Astronomical Surveys

Achieving the science objectives described in the previous section requires precise and accurate modeling of statistical samples of lensing systems. Conventional lens modeling involves

reconstructing the lens mass distribution and source light distribution from imaging (and perhaps ancillary) data. Approximately a thousand strong lensing systems have been discovered in large cosmic imaging surveys to date (see catalogs referenced in Chen et al. (2019a); Moustakas (2012) and the Gravitationally Lensed Quasar Database[3]), and we continue to discover hundreds of new lens candidates in ongoing and recently completed surveys such as the Dark Energy Survey (DES) (Jacobs et al., 2019a, 2017) and the DECam Local Volume Exploration Survey (DELVE) (Zaborowski et al., 2023).

The Vera Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezić et al., 2019) is expected to find about 100,000 such systems, and space-based surveys such as Euclid[4] and the Nancy Grace Roman Space Telescope[5] will additionally find tens of thousands more (Collett et al., 2015; Weiner et al., 2020). While the current approach to strong lens modeling described above has been successful for samples of order 100 or less (Tan et al., 2023), it is labor- and therefore time-intensive; scaling it to the projected LSST strong lens sample would require an estimated ~ 1000 person-years of effort for the analysis, which is infeasible. This necessitates the development of more efficient techniques to model strong lensing systems.

This is still a nascent field, with many approaches that have not yet been explored. The first application of deep learning techniques to lens modeling and analysis was performed less than a decade ago (Hezaveh et al., 2017a). Since then, various deep learning models have shown promise as candidates for automation of the lens analysis process. There has also been significant progress in the development of uncertainty quantification methods in deep learning, which is crucial to strong lensing science.

Chapters 4 and 5 describe the application of two different novel deep learning techniques to strong lensing analysis in simulated wide-field survey imaging data. While most deep learning papers on strong lenses have focused on high-fidelity HST (or HST-like) imaging,

3. `https://research.ast.cam.ac.uk/lensedquasars/`

4. `https://www.cosmos.esa.int/web/euclid`

5. `https://roman.gsfc.nasa.gov`

we have chosen to focus on applying deep learning techniques to ground-based data, which is much less explored but still of great relevance due to recently completed surveys like DES and ongoing surveys like LSST and the Dark Energy Spectroscopic Instrument (DESI). Ground-based data are also typically of much lower quality than space-based data and pose additional challenges to deep learning techniques compared to space-based imaging.

1.5 Thesis Outline

This thesis compiles several related works on strong lensing science using astronomical survey data that I have led or made major contributions to. In Chapter 2, I describe the spectroscopic and photometric characterization of the environment around two time-delay gravitational lenses. In Chapter 3, I present the mass and light properties of Four Early-type galaxy-scale strong lenses discovered in the Dark Energy Survey Year 1 data. In Chapter 4, I present the first application of a state-of-the-art convolutional neural network on simulated ground-based astronomical data. In that paper, we also derive uncertainty estimates for the lens parameters using *concrete dropout* (Gal & Ghahramani, 2016), a technique arising from recent developments in Bayesian deep learning. Finally, in Chapter 5, I describe the first application of simulation-based inference techniques on simulated ground-based astronomical data. Using state-of-the-art normalizing flows, we directly infer the full posterior distribution of lens parameters from imaging data.

CHAPTER 2

SYSTEMATICS IN THE ENVIRONMENT OF TWO TIME-DELAY QUASARS

This chapter is a truncated version of Buckley-Geer et al. (2020), which is a collaborative effort between the four principal co-authors - Elizabeth J. Buckley-Geer, Huan Lin, Cristian E. Rusu, and myself. I led the analysis performed in §2.5-2.6. The reduction of spectroscopic data (§2.2-2.3) and the velocity dispersion of the main lenses (§2.4) are used in §2.5-2.6 and so I have also included those sections.

2.1 Introduction

When a source with a time-varying luminosity such as a quasar or a supernova undergoes strong gravitational lensing, the light reaching the observer from the multiple images takes different paths and hence different travel times. It was noted by Refsdal (1964) that these time delays between the images can be used to measure cosmological distances and the Hubble constant H_0. In recent years, the H_0 Lenses in COSMOGRAIL's Wellspring (H0LiCOW) collaboration have been leading an active time-delay cosmography program to measure H_0 using lensed quasars, see Suyu et al. (2017); Wong et al. (2019) and references therein.

For a recent review of the field of time-delay cosmography see Treu & Marshall (2016). The significant improvements in the uncertainty in the measurements of H_0 in the last two decades have come from the understanding of the key ingredients required to achieve an accurate measurement. In particular, three of these ingredients are: 1) determining the velocity dispersion of the lensing galaxy, 2) identifying galaxies and groups along the line of sight that are close enough to the lens and massive enough to be included in the mass model, and 3) estimating the external convergence $\kappa_{\rm ext}$ due to less massive structures that are not included explicitly in the mass model. In this work, we present results on these three

ingredients for two time-delay lensed quasar systems, DES J0408−5354 (source redshift $z_{\rm s} = 2.375$, main deflector redshift $z_{\rm d} = 0.5967$, Lin et al., 2017) and WGD 2038−4008 ($z_{\rm s} = 0.777$, $z_{\rm d} = 0.2283$, Agnello et al., 2018), as part of the STRong Lensing Insights into the Dark Energy Survey (STRIDES) campaign (Treu et al., 2018b), an external collaboration of the Dark Energy Survey.

It has been known for some time that including stellar kinematics of the lensing galaxy allows one to break the degeneracies inherent in the mass profile of the lens (Treu & Koopmans, 2002b). To obtain the stellar velocity dispersion of the lens we took spectroscopic observations of the main lensing galaxy in DES J0408−5354 and WGD 2038−4008.

If the perturbers along the line of sight are not explicitly accounted for in the lens modeling, these perturbations can result in systematic errors of order a few percent in the inferred value of H_0. To reduce such systematics, we identify galaxies and galaxy groups in the fields of DES J0408−5354 and WGD 2038−4008 that may significantly impact the lensing potential of the system. These galaxies and galaxy groups will be included in the lens models for both DES J0408−5354 (Shajib et al., 2019a, Yidirim et al. in prep) and WGD 2038−4008 (Wong et al. in prep).

To identify the most significant perturbing galaxies and galaxy groups, we use the "*flexion shift*" diagnostic proposed by McCully et al. (2014, 2017a), which has also been used in the line-of-sight analysis of the H0LiCOW lenses HE 0435−1223 (Sluse et al., 2017) and WFI 2033−4723 (Sluse et al., 2019). This diagnostic estimates the difference in lensed image positions caused by the leading order non-tidal (i.e. third-order) perturbation produced by a nearby galaxy or galaxy group. McCully et al. (2017a) showed that by explicitly modeling perturbers with flexion shifts larger than the conservative limit of $\Delta_3 x > 10^{-4}$ ″, we can constrain the bias on H_0 due to this uncertainty to the percent level.

We perform a spectroscopic survey to obtain redshifts of galaxies in the fields of DES J0408−5354 and WGD 2038−4008. This redshift data is used to identify galaxy groups

located in the environment or along the line of sight to these strong lensing systems as well as in the calculation of the flexion shifts of all galaxies and galaxy groups in the environment of both lens systems.

The structure of this paper is as follows. In Section 2.2 we describe our photometric and spectroscopic data, and in Section 2.3 we present our techniques to measure redshifts and stellar masses. In Section 2.4 we derive the stellar velocity dispersions for the main lensing galaxies in the two systems, and in Section 2.5 we describe our technique to identify galaxy groups. In Section 2.6 we identify the structures which can potentially affect the modeling of the lensing systems. Finally, we conclude in Section 2.7.

The current work represents one of a series of papers from the STRIDES collaboration, which together aims to obtain an accurate and precise estimate of H_0 with a blinded approach, from a comprehensive modeling of DES J0408−5354 and WGD 2038−4008. In particular, lens modeling is performed by two independent modeling teams (Shajib et al., 2019a, as well as Yildirim et al. in prep, Wong et al. in prep), both of which make use of the stellar velocity dispersion, environment, and line-of-sight constraints derived in the present work. Throughout this paper, we assume a flat ΛCDM cosmology with $H_0 = 70$ $\mathrm{km\,s^{-1}\,Mpc^{-1}}$, $\Omega_m = 0.3$ for convenience when estimating physical individual galaxy and galaxy group properties (§2.5-§2.6). We present all magnitudes in the AB system. We define all standard deviations as the semi-difference between the 84th and 16th percentiles.

2.2 Data

The Dark Energy Survey (DES) is a deep sky survey that was carried out using the Dark Energy Camera (DECam, Flaugher et al., 2015) located on the Blanco 4m telescope at the Cerro Tololo Inter-American Observatory in the Chilean Andes. The survey ran from 2013-2019 (Diehl, 2020) and covered $\sim$ 5100 sq. degs of the southern sky in five optical filters ($grizY$). The DES data are processed by the DES Data Management team (DESDM,

Morganson et al., 2018) to produce annual data releases that consist of coadded images and object catalogs. We have used two data sets for the work described here, the first year of DES data which is referred to as DES Year 1 (Y1), and the first three years known as DES Year 3 (Y3). The median single epoch PSF FWHM in the i-band is $0.88''$and the coadd magnitude limit in the i-band is 23.44. More details of the survey data can be found in Abbott et al. (2018a).

2.2.1 Spectroscopic Observations

Spectroscopic observations were carried out using three instruments: (1) the Gemini Multi-Object Spectrograph (GMOS-S) on the Gemini South telescope; (2) the Low Dispersion Survey Spectrograph (LDSS-3) on the Magellan Clay telescope; and (3) the Multi Unit Spectroscopic Explorer (MUSE) on the European Southern Observatory (ESO) Very Large Telescope (VLT) Unit Telescope 4 (UT4). Table 2.1 summarizes details of the spectroscopic data taken using these three telescope+instrument setups.

The GMOS-S observations were taken as part of two programs: (1) a Gemini Large and Long Program (LLP; PI: E. Buckley-Geer; program IDs GS-2015B-LP-5 and GS-2017A-LP-5) of spectroscopic follow-up for DES strong lensing systems and for DES photometric redshift (photo-z) calibrations; and (2) a dedicated program (PI: H. Lin; program ID GS-2018B-Q-220) to observe line-of-sight galaxy redshifts and lensing galaxy velocity dispersions for our two lensed quasar systems. These programs observed a total of four GMOS-S multi-object spectroscopy (MOS) masks for DES J0408−5354 and two masks for WGD 2038−4008, and the data were taken in queue mode on Gemini South during the semesters 2015B, 2017A, and 2018B. Nearly all the masks were observed using both the GMOS-S B600 and R400 gratings, to provide spectral coverage over both blue and red wavelength ranges, respectively, spanning approximately 3800Å−7500Å and 5000Å−10500Å. Multiple science exposures were taken to reject cosmic rays, and the grating central wavelength was dithered slightly for

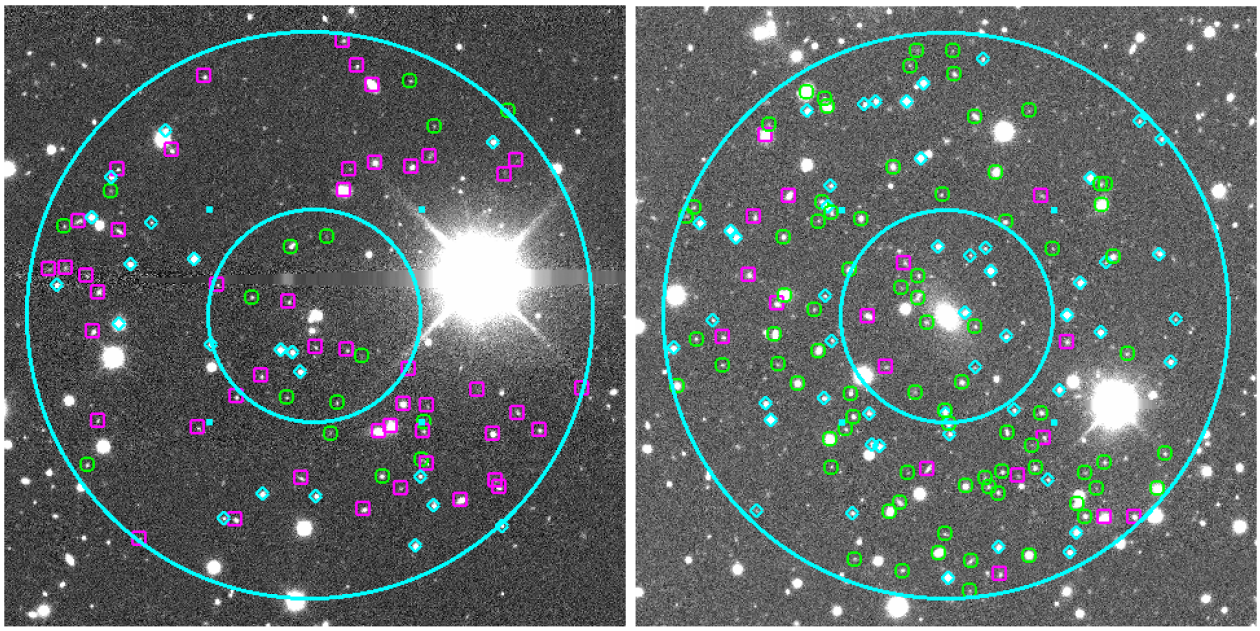

Figure 2.1 The 1000 × 1000 pixel field of view for DES J0408−5354 (left) and WGD 2038−4008 (right). North is up and East is left. The $i < 22.5$ galaxies inside the 120″ radius aperture are indicated by magenta squares for the objects with a spectroscopic redshift and green circles for the objects with no spectroscopic redshift. Stars are indicated by the cyan diamonds. The two concentric cyan circles indicate the apertures of 120″ and 45″ radius respectively.

different exposures to fill in wavelength gaps due to spatial gaps between the three GMOS-S CCDs. Flat field and wavelength calibration exposures were interspersed with the science exposures.

The LDSS-3 observations were taken as part of a semester 2018A Magellan program (PI: J. Frieman) to obtain line-of-sight galaxy spectroscopy for DES J0408−5354 and two other lensed quad quasar systems. Four LDSS-3 MOS masks were observed for DES J0408−5354 over the two nights of 2018 January 21,22 UT. Each mask was observed using the LDSS-3 VPH-All grism, with a wavelength coverage of about 3800Å−10500Å. Multiple science exposures were taken to reject cosmic rays. Flat field and wavelength calibration exposures were taken immediately before and after the sequence of science exposures.

The VLT MUSE observations were taken as part of an ESO program (0102.A-0600(E), PI: A. Agnello) to do integral field spectroscopy of DES J0408−5354 and its surrounding field. The MUSE observations were done in wide-field mode with adaptive optics corrections and were carried out over the two nights of 2019 Jan 11 and 13. The final MUSE data cube covered an area of $92'' \times 95''$, and the wavelength coverage spanned 4700Å−5803Å and 5966Å−9350Å. Additional spectroscopic analysis of the MUSE data and further details of the observations and data processing are given in Shajib et al. (2019a).

2.2.2 Spectroscopic target selection

Galaxy targets for the Gemini and Magellan masks were selected using DES photometry. The exact selection criteria changed somewhat with time, due to improvements in DES photometric measurements, star-galaxy separation, and object catalogs. Specifically, sets of selection criteria were used for the masks in Table 2.1, listed below in order from earliest to latest.

Table 2.1 Spectroscopic observations for the DES J0408−5354 and WGD 2038−4008 systems.

Telescope/Instrument Mask/Cube	Grating/ Grism	UT Date	Exposure time (sec)
DES J0408−5354			
Gemini South/GMOS-S			
(1) DESJ0408-5354_42	B600	2015 Dec 09	4×900
(2) DESJ0408-5354_42	R400	2015 Dec 09	4×900
(3) DESJ0408-5354_45	B600	2017 Apr 26−27	5×900
(4) DESJ0408-5354_45	R400	2017 Mar 30−31	4×900
(5) DESJ0408-5354_A	B600	2018 Dec 04	4×900
(6) DESJ0408-5354_A2	R400	2019 Feb 04	6×1000
(7) DESJ0408-5354_B	B600	2018 Dec 09	4×900
(8) DESJ0408-5354_B	R400	2018 Dec 09	4×900
Magellan Clay/LDSS-3			
(9) des0408a	VPH-All	2018 Jan 21	7×780
(10) des0408b	VPH-All	2018 Jan 21	6×780
(11) des0408c	VPH-All	2018 Jan 22	7×780
(12) des0408d	VPH-All	2018 Jan 22	6×780
VLT UT4/MUSE			
(13) MUSE		2019 Jan 11,13	14400
WGD 2038−4008			
Gemini South/GMOS-S			
(14) DESJ2038-4008_A	B600	2018 Nov 06	4×900
(15) DESJ2038-4008_A	R400	2018 Nov 07	4×900
(16) DESJ2038-4008_B	B600	2018 Nov 07	4×900

(A) Gemini South masks (1)-(4): Galaxies were selected from the DES Year 1 (Y1) "Y1A1 COADD" catalog (Drlica-Wagner et al., 2018), using the i-band `SExtractor AUTO` magnitude (Bertin & Arnouts, 1996) cuts $20 \leq$ `MAG_AUTO_I` < 22.5. No Milky Way extinction corrections were applied to the magnitudes before selection. Star-galaxy separation used the `SExtractor SPREAD_MODEL` classifier (Desai et al., 2012), also in the i-band: `SPREAD_MODEL_I` $>$ 0.002.

(B) Magellan masks (9)-(12): Galaxy targets were now selected from the deeper DES Year 3 (Y3) data set, specifically from the "Y3 GOLD" (Sevilla et al. in prep) version 1.0 catalog, using the i-band magnitude cuts $18 \leq$ `MAG_AUTO_I` < 23, in particular extending the faint magnitude limit fainter to aim for a deeper sample. Again no Milky Way extinction corrections were applied to the magnitudes. Star-galaxy separation used the same i-band cut: `SPREAD_MODEL_I` > 0.002.

(C) Gemini South masks (5)-(8) and (14)-(16): These masks were the latest to be designed and therefore used improved selection methods compared to the masks in (A) and (B). Galaxies were again selected from the "Y3 GOLD" data set, but using the most current version 2.2 catalog. We also changed the selection magnitudes from `AUTO` to `MOF` (Drlica-Wagner et al., 2018, Sevilla et al. in prep) magnitudes, where the latter provide improved photometry, based on simultaneous multi-epoch, multi-band, and multiobject fits. The adopted i-band cuts were $15 \leq$ `MOF_CM_MAG_CORRECTED_I` < 23, where these MOF magnitudes also included Milky Way extinction corrections and several other sub-percent photometric zero point corrections (Sevilla et al. in prep). Finally, we also updated the star-galaxy separation classifier to `EXTENDED_CLASS_MASH_MOF` (Sevilla et al. in prep), specifically using the cuts `EXTENDED_CLASS_MASH_MOF` $= 3$ (indicating "high confidence galaxies") or 2 (indicating "mostly galaxies"). For $i \leq 22.5$, this classification should yield a galaxy efficiency $> 98.5\%$

and stellar contamination $< 1\%$ (Sevilla et al. in prep).

Each Magellan LDSS-3 and Gemini South GMOS-S mask included galaxy targets distributed over about a $5.5' \times 5.5'$ sky area centered on each lensed quasar system. Generally, one to three slits on each mask were manually designed to target objects in or close to the lensed quasar system, e.g., to measure the redshift or velocity dispersion of the main lensing galaxy in each system, or to observe close nearby neighbor galaxies of the systems. The remaining targets were selected automatically by `GMMPS`[1] or `maskgen`[2], the respective GMOS-S or LDSS-3 mask design software, both of which designed masks to maximize the number of targets observed. The potential set of galaxies that could be targeted was subject to the selection criteria described above. In addition, at the time each mask was designed, we removed from the initial target list any galaxies that already had high-confidence redshifts from previous observations, or that were already targeted on companion masks designed for the same semester's (Gemini) observing queue or (Magellan) observing run. Moreover, for the Magellan LDSS-3 targets (only), we assigned targeting priorities as inputs to `maskgen`, depending on the i-band magnitude and the radius from the DES J0408−5354 quad system. Specifically, for galaxies with radius $\leq 3'$, we assigned priorities based on `MAG_AUTO_I` (as in criteria (B) above), with highest priority given to bright galaxies `MAG_AUTO_I` < 19.5, next priority to galaxies otherwise brighter than `MAG_AUTO_I` $= 22$, and lowest priority to galaxies otherwise brighter than `MAG_AUTO_I` $= 23$. Finally, for galaxies with radius $> 3'$, we assigned lower priorities than for galaxies $\leq 3'$, and these priorities were tied linearly to `MAG_AUTO_I` (with brighter galaxies at higher priority).

1. `https://gmmps-documentation.readthedocs.io/en/latest/`

2. `https://code.obs.carnegiescience.edu/maskgen`

2.3 Redshifts and stellar masses

In this section we first describe our spectroscopic redshift measurement procedure, summarize the results, and show the line-of-sight galaxy redshift distributions for each lensing system. We then describe and plot the spectroscopic redshift completeness for the overall galaxy samples in the two systems. Finally, we detail the procedures for measuring stellar masses using photometric model fitting for our galaxies.

2.3.1 Spectroscopic redshifts

The Gemini and Magellan data were processed to 2D and 1D spectra using the IRAF Gemini and COSMOS (Dressler et al., 2011; Oemler et al., 2017) reduction packages, respectively. Initial redshifts were determined automatically (without visual inspection) using the IRAF external package `rvsao` (Kurtz & Mink, 1998) to cross-correlate the 1D spectra against a set of SDSS galaxy templates. All processed 2D and 1D spectra, along with the corresponding automated redshift results, were visually inspected to assign final quality flags to the redshifts. If necessary, an automated redshift may be overridden and manually re-measured from the 2D or 1D spectrum. Only high-confidence redshifts were included in the subsequent analyses.

For DES J0408−5354, we obtained 101 high-confidence galaxy redshifts from Gemini South GMOS-S and 70 from Magellan LDSS-3. From the VLT MUSE data, we obtained another 28 redshifts that were not already among the Gemini and Magellan redshifts, thus resulting in a total of 199 high-confidence redshifts for DES J0408−5354. For WGD 2038−4008 we obtained a total of 54 high-confidence galaxy redshifts, all obtained from the Gemini South GMOS-S data. The above counts include the two redshifts of the main lensing galaxies in the systems. Figure 2.2 shows histograms of the redshift distributions for each lensing system.

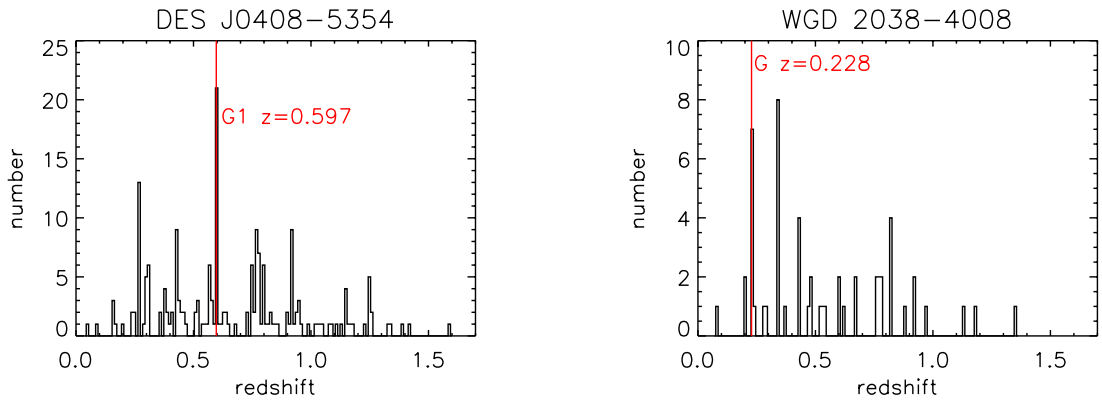

Figure 2.2 Histograms of line-of-sight galaxy spectroscopic redshifts (§2.3.1) for the (left) DES J0408−5354 and (right) WGD 2038−4008 systems. In each panel, the red vertical line indicates the redshift of the main lensing galaxy in each system.

2.3.2 Redshift completeness

We define spectroscopic redshift completeness to be the fraction of DES Y3 Gold (Sevilla et al. in prep) galaxies that have redshifts (as described in §2.3.1). To define our photometric galaxy sample, we use the latest and best available version of the Y3 Gold catalog, version 2.2, and we also use the same magnitude and star-galaxy separation cuts as listed in target selection criteria (C) of §2.2.2 above.

For DES J0408−5354, the resulting spectroscopic redshift completeness is 0.68 for 18 <= `MOF_CM_MAG_CORRECTED_I` < 23 galaxies and $5''$ <= radius < $3'$; see the top panels of Figure 2.3. For WGD 2038−4008, the redshift completeness is 0.16 for the same i-band magnitude and radius ranges, as shown in the bottom panels of Figure 2.3. Within the plotted magnitude and radius ranges, the redshift completeness stays fairly constant for both systems. The inner radius cut of $5''$ is intended to exclude the quasar images from consideration. The outer radius cut is set at $3'$ as the redshift completeness drops very rapidly beyond this radius for either system.

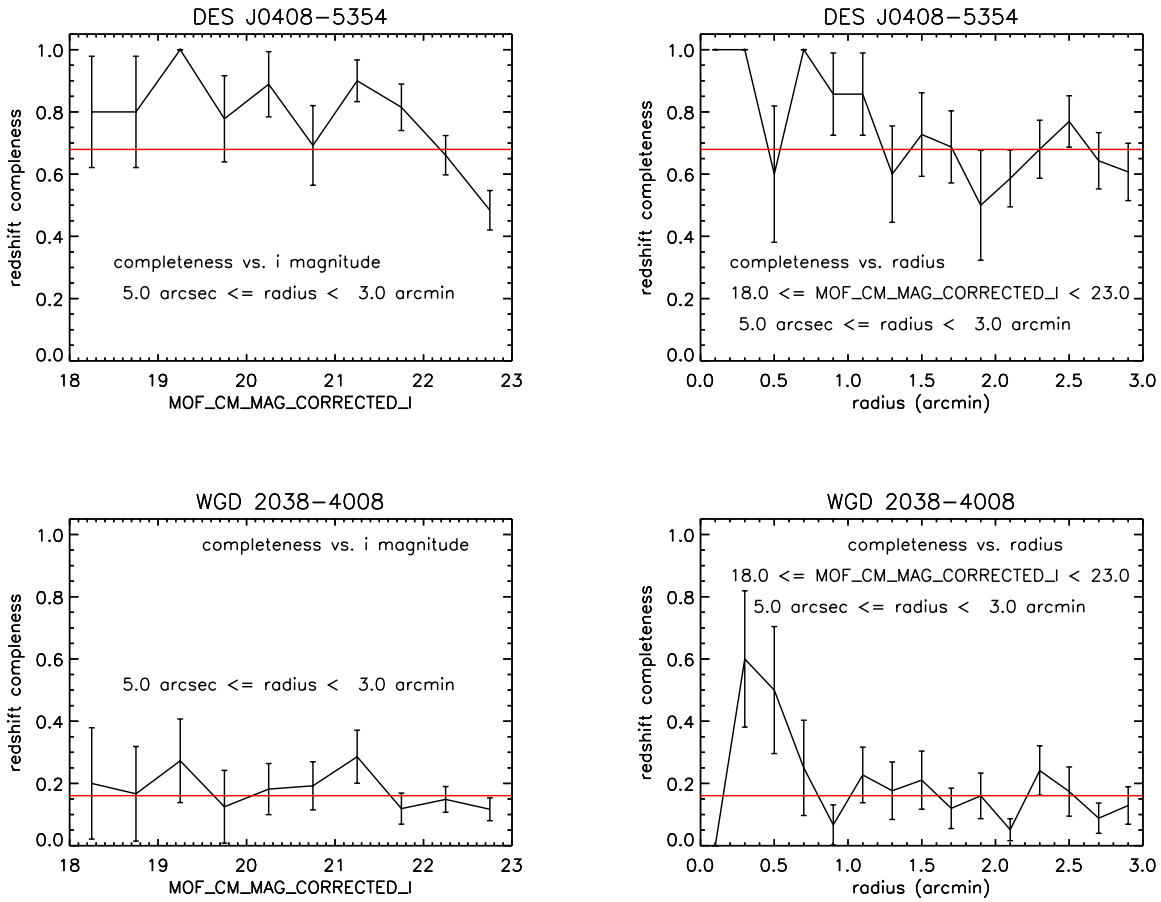

Figure 2.3 Spectroscopic redshift completeness (defined in §2.3.2) for DES J0408−5354 (top panels) and WGD 2038−4008 (bottom panels). The left panels show redshift completeness vs. i-band magnitude, within the radius range $5''$to $3'$ from the lens galaxy in each system. The right panels show redshift completeness vs. radius, within the i-band magnitude range 18 to 23. In all panels, the red horizontal line indicates the overall redshift completeness within the indicated magnitude and radius ranges for each system.

2.3.3 Stellar masses

Stellar masses necessary for the computation of the flexion shift, the criterion used to separate between the structures that need to be accounted for in the lensing model and those which can be incorporated inside $\kappa_{\rm ext}$ (see §2.6.1) were computed using the galaxy template fitting code Le PHARE (Arnouts et al., 1999; Ilbert et al., 2006). In the Le PHARE fits, we either used our spectroscopic redshifts when available or DES Y3 Gold DNF photo-z's (`DNF_ZMEAN_SOF`), combined with MOF photometry (Drlica-Wagner et al., 2018) from the DES Y3 Gold (Sevilla et al. in prep) version 2.2 catalog. Specifically, we used the `MOF_CM_MAG_CORRECTED` magnitudes and their associated errors in the *griz* filters, as these magnitudes included Milky Way extinction and other small photometric corrections (see 2.2.2).

The galaxy template set used in the Le PHARE fits was taken from the BC03 (Bruzual & Charlot, 2003) spectral energy distribution (SED) library. Specifically, we used a set of 27 BC03 simple stellar population models, computed using the Padova 1994 stellar evolution library (described in Bruzual & Charlot, 2003) and the Chabrier (2003) initial mass function. The 27 models consisted of 9 exponentially declining star formation rate (SFR) histories (with decay times $\tau = 0.1, 0.3, 1, 2, 3, 5, 10, 15, 30$ Gyr), each computed at 3 different metallicity values (0.4, 1, and 2.5 times solar). Each of the 27 models was also computed at different ages (ranging from 0.2 to 13.5 Gyr) and redshifts (up to a maximum redshift of 1.1, in steps of 0.03 in redshift). No dust extinction was included in the models, but recall the DES magnitudes already included correction for foreground Milky Way extinction. The stellar masses and uncertainties computed from the Le PHARE fit are tabulated in the appendix of the published version of this chapter (Buckley-Geer et al., 2020) under Table A1.

For a few of the close neighbor galaxies of the two lensed quasar systems, the original object deblending and resulting photometry in the DES Y3 catalog was incorrect upon visual inspection of the images. In particular, for DES J0408−5354, the neighbor galaxy with ID

number 488066768 (or name G5, in Table 2.4) had magnitudes that were too bright, and for WGD 2038−4008, the three neighbor galaxies with ID numbers 13, 14, and 15 (Table A1 in Buckley-Geer et al. (2020)) were originally merged into a single object. To correct these photometry problems, we used the galaxy image fitting code GALFIT (Peng et al., 2010) to redo the galaxy model fitting and photometry for these objects, based on the DES Y3 coadd images in the *griz* filters. These GALFIT results were used to compute stellar masses via Le PHARE.

2.4 Velocity Dispersions of Lensing Galaxies

The main lensing galaxies in both the DES J0408−5354 and WGD 2038−4008 systems were specifically targeted for stellar velocity dispersion measurements on a number of the spectroscopic masks listed above in Table 2.1. For these observations we describe below the details of the targeting on the spectroscopic slits, the procedures used to extract the lensing galaxy spectra, and the method employed to measure velocity dispersions.

2.4.1 DES J0408−5354 G1

The main lensing galaxy G1 in DES J0408−5354 was measured from four independently observed spectra: two from Magellan LDSS-3, one from Gemini South GMOS-S, and one from VLT MUSE.

G1 was targeted on two of the four Magellan LDSS-3 masks listed in Table2.1, "des0408a" and "des0408b"; hereafter these two masks will be denoted "Magellan a" and "Magellan b". The slit setup for G1 was the same on both masks, while the remaining targets were different on the two masks. Specifically, the slit was oriented so that it included both G1 and quasar image B (the naming convention for the lensing galaxy G1 and the lensed quasar images A, B, and D are shown in Lin et al., 2017). In addition, some contaminating flux from quasar image A was also visible in the 2D spectrum from the slit. To extract the 1D spectrum of

G1, we first fit the spatial profiles along the slit (at each wavelength) of G1, B, and A by two Moffat profiles and a Gaussian profile, respectively. Moffat profiles were adopted for G1 and B because Gaussian profiles gave worse fits as determined by visual inspection. We subtracted off the best-fit spatial profiles for B and A and summed the remaining flux over an extraction window along the slit of $1''$ (approximately the FWHM of G1's Moffat profile) or $2''$ (about the extent of G1's profile) to extract G1's 1D spectrum.

G1 was targeted on one of the Gemini South GMOS-S masks in Table 2.1, "DESJ0408-5354_A2" (R400 grating), hereafter denoted "Gemini A2," on a slit which also included quasar image D. The procedure to extract G1's 1D spectrum was entirely analogous to that described above for the Magellan data. To fit the spatial profiles of G1 and D, Moffat profiles were again found to be better than Gaussians.

In the VLT MUSE data cube, the quasar and source light components were first modeled and removed, and the region near C and G2 was masked out. The remaining light from G1 was then extracted to a 1D spectrum by summing the light over a $1'' \times 1''$ box or a $2.2'' \times 2.2''$ box.

The velocity dispersion of G1 was then measured from the above data using the ULySS (Koleva et al., 2009) galaxy spectral modeling package. In the rest wavelength range 4800Å to 5500Å, including the Hβ, Mg, and CaFe features, the G1 spectra were fit to Vazdekis et al. (2010) stellar population models, which used the MILES stellar library (Sánchez-Blázquez et al., 2006) and the Salpeter (1955) initial mass function. The wavelength-dependent line spread function (LSF) in the Magellan and Gemini-South data were determined from the widths of the arc lamp lines in the respective wavelength calibration spectra, while the LSF of the MUSE data were taken from the fits given in §3.1 of Guérou et al. (2017). Because the Gemini-South and in particular Magellan LSFs were noticeably non-Gaussian, we modified the ULySS package so that it could make use of an empirical LSF, instead of an analytic Gaussian (or low-order Gauss-Hermite) LSF. The resulting velocity dispersion measurements

and associated statistical errors are given in Table 2.2, showing good agreement among the results from the four independent data sets. Plots of the 1D spectra and best-fit models for the $1''$ extraction window cases are shown in Figure 2.4.

2.4.2 WGD 2038−4008 G

The main lensing galaxy G in WGD 2038−4008 was measured from one spectrum observed using Gemini South GMOS-S, specifically on mask (14) listed in Table 2.1: "DESJ2038-4008_A" (B600 grating), hereafter denoted "Gemini A" for simplicity. One slit on this mask targeted galaxy G together with the quasar images C and D (the naming convention for the lensing galaxy and the lensed quasar images are shown in Figure 1 of Agnello et al., 2018). To extract the 1D spectrum of G, we used the same method described above in §2.4.1. Specifically, we fit the spatial profiles of G, C, and D with three Moffat profiles, subtracted off the best-fit profiles of C and D, and then summed the remaining flux over extraction windows of $1''$ or $2''$ along the length of the slit. The velocity dispersion of G was then measured with ULySS, using the same rest wavelength range and the same stellar population models as above in §2.4.1. The resulting velocity dispersion measurements and statistical errors are given in Table 2.2. The 1D spectrum and best-fit model for the $1''$ extraction window case are shown in Figure 2.4.

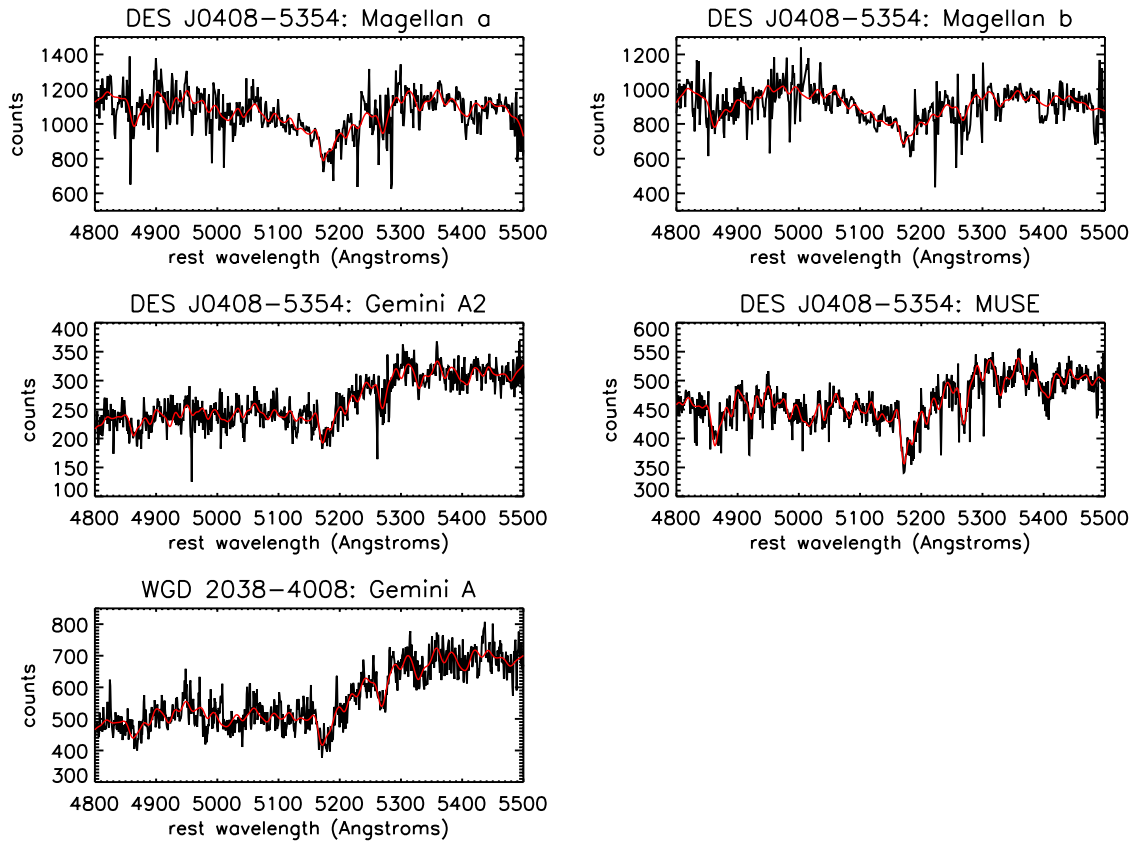

Figure 2.4 The 1D spectra and fits involved in the velocity dispersion measurements of the main lensing galaxies G1 in the DES J0408−5354 system (top and middle panels) and G in the WGD 2038−4008 system (bottom panel), as described in §2.4 and listed in Table 2.2. The black curves in the plots show the observed data, in units of counts vs. rest-frame wavelengths in Angstroms, while the red curves show the best-fit models (details in §2.4). All the data and fits shown are for the case of a $1''$ extraction window (see §2.4 and Table 2.2).

Table 2.2 Velocity dispersion results for the main lensing galaxies in the DES J0408−5354 and WGD 2038−4008 systems. Details of the measurements are given in §2.4.

Data set	Velocity dispersion (km/s)	Slit width (arcsec)	Extraction window (arcsec)	Seeing FWHM (arcsec)	Moffat index β
DES J0408−5354					
Magellan a	230 ± 37	1.0	1.0	0.68	2.97
Magellan b	236 ± 42	1.0	1.0	0.76	3.20
Gemini A2	220 ± 21	0.75	1.0	0.52	3.06
MUSE	227 ± 9	1.0	1.0	0.61	1.55
Magellan a	209 ± 37	1.0	2.0	0.68	2.97
Magellan b	230 ± 47	1.0	2.0	0.76	3.20
Gemini A2	261 ± 21	0.75	2.0	0.52	3.06
MUSE	227 ± 9	2.2	2.2	0.61	1.55
WGD 2038−4008					
Gemini A	296 ± 19	0.75	1.0	0.90	1.74
Gemini A	303 ± 24	0.75	2.0	0.90	1.74

2.5 Galaxy Group identification

2.5.1 Galaxy Group Identification Algorithm

For galaxy-group identification, we employed the same algorithm used in the spectroscopic analysis of the fields of H0LiCOW lenses HE 0435−1223 (Sluse et al., 2017) and WFI 2033−4723 (Sluse et al., 2019), which is based on the group-finding algorithms of Wilman et al. (2005) and Ammons et al. (2014). Wilson et al. (2016) uses a similar method, the results of which were used in the analysis of the H0LiCOW lens PG 1115+080 (Chen et al., 2019b). We summarize the method here and refer interested readers to Sluse et al. (2017) for a more complete description and explanation of parameter choices in this algorithm.

The first step towards identifying galaxy groups involves searching for candidate groups in the spectroscopic redshift distribution of the surveyed galaxy catalog. We begin by constructing a redshift histogram with bins of width 2000 $\mathrm{km\,s^{-1}}$. We identify redshift bins with 5 or more members as candidate groups. To ensure that candidate group members are

not split across two bins due to an arbitrary choice of bin boundaries, we construct a second redshift histogram with the bins shifted by half a width of a bin (1000 $\mathrm{km\,s^{-1}}$) and count all non-duplicate redshift peaks from both histograms as candidate groups. We include all other galaxies that are within 1500 $\mathrm{km\,s^{-1}}$of a candidate group member in that candidate group.

Once we have identified the candidate groups, we use a biweight location estimator (Beers et al., 1990) to calculate the mean (group) redshift of each candidate group. The group centroid is also calculated from the positions of the candidate group members. Since Sluse et al. (2019) found that using a luminosity-weighted scheme to calculate the centroid does not improve the match between the group centroid and brightest galaxy in this method, we do not use luminosity-weighted centroids here.

Once candidate groups have been identified, they are subjected to an algorithm that iteratively removes outliers in both redshift and angular space until the algorithm converges to a stable solution or a group membership of zero. The latter indicates that the candidate group is not gravitationally linked and is spurious. The algorithm is as follows:

1. We set the initial observer-frame velocity dispersion, σ_{obs} to 500 $\mathrm{km\,s^{-1}}$. This value will be revised in subsequent iterations.

2. Candidate group members that are further than twice the velocity dispersion away from the group redshift are excluded from the group. This corresponds to the following limit

$$\delta z_{\mathrm{max}} = n \times \sigma_{\mathrm{obs}}/c \tag{2.1}$$

where $n = 2$. This redshift limit is converted into an angular separation limit

$$\delta\theta_{\mathrm{max}} = \frac{c \times \delta z_{\mathrm{max}}}{b(1+z)\,H(z)\,D_{\theta}(z)} \tag{2.2}$$

where $H(z)$ is the Hubble parameter at redshift z and $D_\theta(z)$ is the angular diameter distance from the observer to redshift z. Following Sluse et al. (2017), we set the aspect ratio $b = 3.5$. Candidate group members that have an angular separation that is larger than $\delta\theta_{\rm max}$ from the group centroid are excluded from the group.

3. Once cuts have been made in both redshift and angular separation, we recalculate the group centroid, group redshift, and observed-frame velocity dispersion $\sigma_{\rm obs}$ from the remaining candidate group members. We obtain the latter two quantities following this framework:

    - If there are more than 10 galaxies remaining, we use the biweight location and scale estimators to calculate the group redshift and velocity dispersion, respectively (Mosteller & Tukey, 1977).
    - If there are between 4 and 10 galaxies remaining (inclusive), we use the biweight location to calculate the group redshift and the gapper estimator to calculate the velocity dispersion (Wainer & Thissen, 1976; Beers et al., 1990).
    - If there are fewer than 4 galaxies, we use the mean redshift as the group redshift and the standard deviation as the velocity dispersion.

Steps (ii) and (iii) are repeated until we reach a stable solution. Galaxies that are members of these identified groups are then used to infer group properties, such as the group redshift, centroid, velocity dispersion, and flexion shift (following the method described in §2.6.1. The rest-frame velocity dispersions are calculated from the observer-frame velocity dispersions using

$$\sigma_{\rm rest} = \frac{\sigma_{\rm obs}}{1 + \bar{z}_{\rm group}}. \tag{2.3}$$

We then estimate uncertainties in the group properties by bootstrapping (i.e. random sampling with replacement) the group members of each group 1000 times. We recalculate the group properties of the resampled groups and use the bootstrapped distribution in those quantities to estimate their uncertainties.

Because the associated measurement uncertainties of the galaxies in the spectroscopic redshift catalog ($\Delta v_{\rm err} \sim 100$ km s^{-1}, see §2.3) are of order the measured velocity dispersion of many of the identified groups, care must be given to account for these uncertainties. To this end, we forward-model the kinematic datasets to infer the velocity dispersion given measurement uncertainties, following techniques used in dwarf-galaxy studies (e.g. Koposov et al. (2011); Walker & Peñarrubia (2011); Amorisco & Evans (2012)), where it is especially relevant for systems with small numbers of discrete kinematic tracers (Martin et al., 2018; Laporte et al., 2019). We construct a generative likelihood model for the data and evaluate the posterior probability distribution for the intrinsic velocity dispersion, $\sigma_{\rm int}$. The likelihood function is

$$\mathcal{L} = \prod_{\rm i} \frac{1}{\sqrt{2\pi}\sigma_{\rm obs}} \exp\left(-0.5\left(\frac{v_i - \langle v \rangle}{\sigma_{\rm obs}}\right)^2\right) \tag{2.4}$$

where $\langle v \rangle$ is the mean velocity, the product is over all member galaxies i of the galaxy group, and $\sigma^2_{\rm obs} = \sigma^2_{\rm int} + \Delta v^2_{\rm err}$. We assume $\Delta v_{\rm err} = 100$ km s^{-1}, and a non-informative Jeffreys prior for the intrinsic velocity dispersion $\sigma_{\rm int}$ (i.e. $p(\sigma_{\rm int}) \propto 1/\sigma_{\rm int}$) over the range 1 to 1000 km s^{-1}. We also assume a uniform prior for the mean group velocity $\langle v \rangle$ over the range -500 to 500 km s^{-1}and treat it as a nuisance parameter. We then sample the posterior PDF using the `emcee` affine invariant Markov Chain Monte Carlo sampler (Foreman-Mackey et al., 2013; Goodman & Weare, 2010). We then report the median and 68th percentile confidence intervals of the posterior PDF for $\sigma_{\rm int}$ in Table 2.3. For groups where the posterior PDF for $\sigma_{\rm int}$ peaks near zero and the lower bounds are not well-constrained, we report only the 68th

percentile upper limits.

2.5.2 Identified Groups in the Environment of DES J0408−5354

We applied the galaxy-group identification algorithm to the combined catalog of 199 galaxies with high-confidence redshifts in the field of DES J0408−5354 described in §2.3. We identified 10 galaxy groups comprising 76 galaxies from this spectroscopic sample, which we label Group 1-10 in order of increasing group redshift. Their properties are summarized in Table 2.3, and Figure 2.5 shows, for each identified galaxy group, the positions of both accepted and rejected trial member galaxies of that group in right ascension and declination, as well as the distances and velocities relative to the converged group centroid.

For each group in Figure 2.5, the first plot (above) shows the positions of the candidate member galaxies associated with that group relative to the lens galaxy, with rejected group members represented as red squares and accepted group members represented as green circles. The lens galaxy DES J0408−5354 (star) and group centroid (cross) are also displayed. The R_{200} radius of the group is represented by a solid line, while the dashed circle represents the angular separation cut of the group-finding algorithm in its final iteration. The second plot (below) shows the observer-frame velocity of individual member galaxies relative to the group centroid as a function of that galaxy's angular distance from the centroid. Galaxies that passed the iterative algorithm described in §2.5.1 are shown in green, while trial galaxy members that were cut through the algorithm are shown in red. Horizontal error bars represent the measurement error for each galaxy (see §2.3). The final observer-frame velocity dispersion and angular separation cuts from the group-finding algorithm are presented as dashed and dotted lines respectively. We also show 1-D histograms and rug plots of the velocity and distance distributions of the member galaxies. The 1-D histograms are produced using a Kernel Density Estimate (KDE) with a bandwidth determined using Scott's Rule. In the 1D velocity histogram, the dashed blue line shows a Gaussian with width equal to

Table 2.3 Group properties in the field of view of DES J0408−5354 and WGD 2038−4008. The columns show the group ID, group redshift, number of spectroscopically identified galaxies in that group, the group rest-frame velocity dispersion (rounded to the nearest $10\,\mathrm{km\,s^{-1}}$), the group centroid (in RA and Dec), projected distance of the centroid to the lens ($\Delta\theta$), and median flexion shift $\log(\Delta_3 x(\mathrm{arcsec}))$ (see §2.6.1 for methodology). All reported values are median quantities of 1000 bootstrapped samples, with uncertainties given by the 16th and 84th confidence intervals of the distribution of the bootstrapped quantity. Velocity dispersion estimates are rounded to the nearest $10\,\mathrm{km\,s^{-1}}$. See §2.6.1 for further discussion.

ID	$\bar{z}_{\mathrm{group}}$	N	σ_{rest} ($\mathrm{km\,s^{-1}}$)	σ_{int} ($\mathrm{km\,s^{-1}}$)	R_{200} (Mpc)	$\mathrm{RA_{ctr}}$, $\mathrm{DEC_{ctr}}$ (deg)	error($\mathrm{RA_{ctr}}$, $\mathrm{DEC_{ctr}}$) (arcmin)	$\Delta\theta$ (arcsec)	$\log_{10}(\Delta_3 x)$ ($\log_{10}$(arcsec))
DES J0408−5354									
1	0.272	11	110 ± 20	< 70	$0.30^{+0.05}_{-0.06}$	62.096099, -53.903136	0.67, 0.46	32^{+22}_{-16}	< -5.88
2	0.307	6	290^{+60}_{-90}	300^{+160}_{-80}	$0.81^{+0.15}_{-0.24}$	62.065082, -53.919339	0.71, 0.31	94^{+10}_{-10}	$-4.79^{+0.74}_{-0.63}$
3	0.428	7	110 ± 30	< 70	$0.31^{+0.07}_{-0.09}$	62.130389, -53.936509	0.29, 0.29	156^{+10}_{-8}	< -8.04
4	0.570	5	140^{+30}_{-90}	< 160	$0.41^{+0.09}_{-0.25}$	62.042510, -53.918109	1.00, 0.52	122^{+28}_{-21}	< -6.53
5 ⋆	0.599	17	240^{+30}_{-40}	220^{+60}_{-50}	$0.68^{+0.08}_{-0.11}$	62.079749, -53.901149	0.54, 0.18	26^{+16}_{-12}	$-3.86^{+0.97}_{-0.72}$
6	0.751	3	80 ± 40	< 100	$0.21^{+0.11}_{-0.11}$	62.147320, -53.872270	0.05, 0.09	156^{+3}_{-2}	< -8.06
7	0.768	5	120^{+30}_{-50}	< 120	$0.34^{+0.10}_{-0.15}$	62.084144, -53.898689	0.28, 0.24	17^{+12}_{-7}	< -4.83
8	0.799	6	340^{+60}_{-80}	370^{+170}_{-100}	$0.96^{+0.18}_{-0.22}$	62.143634, -53.894864	0.64, 0.71	120^{+22}_{-22}	$-5.46^{+0.72}_{-0.59}$
9	0.918	9	300^{+40}_{-50}	310^{+120}_{-70}	$0.86^{+0.10}_{-0.15}$	62.051793, -53.892228	0.49, 0.45	90^{+16}_{-17}	$-5.70^{+0.58}_{-0.52}$
10	1.252	7	240^{+30}_{-50}	240^{+110}_{-70}	$0.67^{+0.09}_{-0.15}$	62.124690, -53.913336	0.69, 0.12	87^{+22}_{-21}	$-6.92^{+0.80}_{-0.71}$
WGD 2038−4008									
1 ⋆	0.229	7	200^{+70}_{-60}	200^{+90}_{-70}	$0.53^{+0.18}_{-0.16}$	309.528309, -40.126006	0.41, 0.28	62^{+23}_{-24}	$-5.30^{+0.90}_{-0.83}$
2	0.342	8	380^{+70}_{-90}	400^{+150}_{-100}	$1.05^{+0.19}_{-0.26}$	309.543495, -40.142963	0.54, 0.28	91^{+26}_{-27}	$-5.39^{+0.73}_{-0.59}$

Note: ⋆ Group contains the lens galaxy.

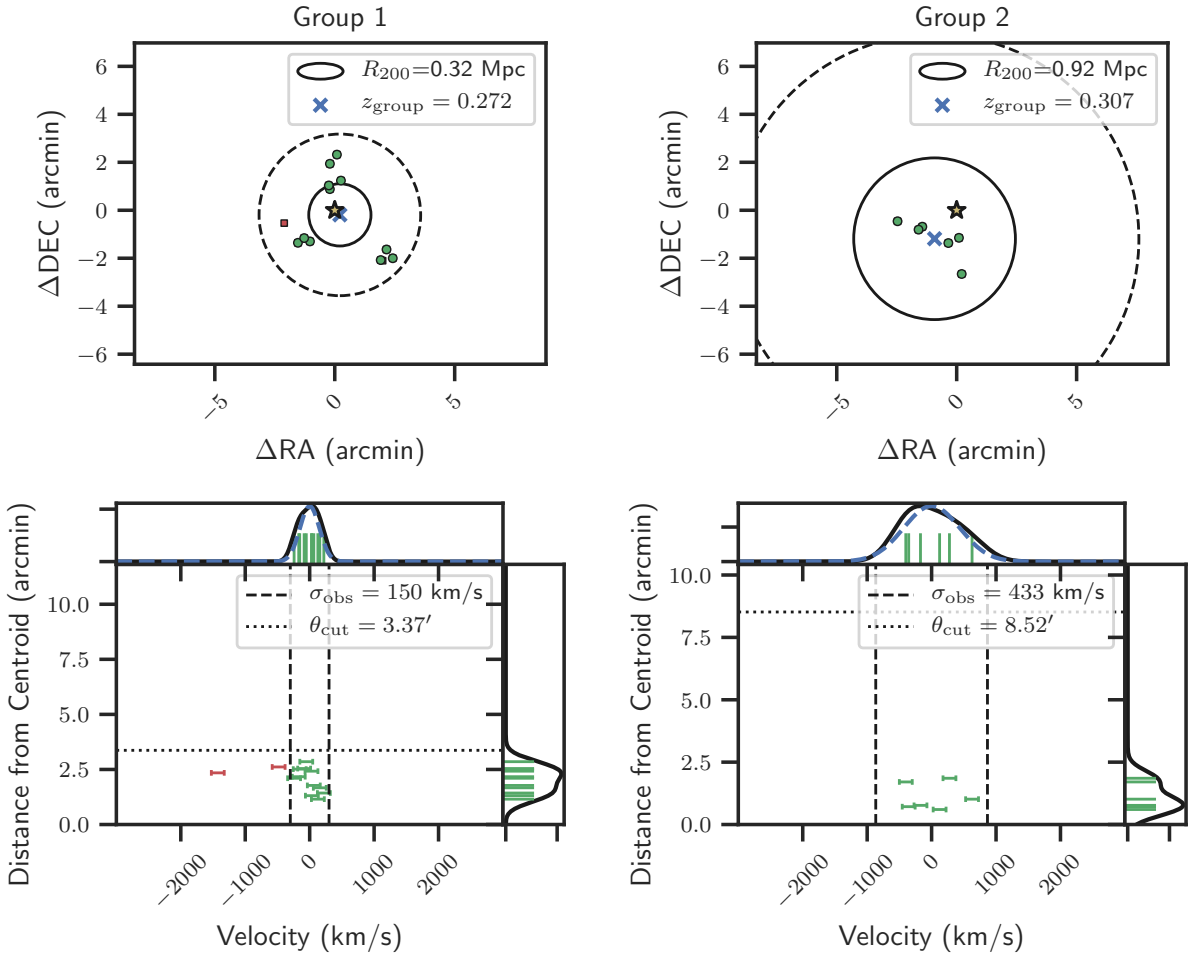

Figure 2.5 Galaxy groups identified in the spectroscopic sample of galaxies in the field of view of DES J0408−5354. See §2.5.2 for details.

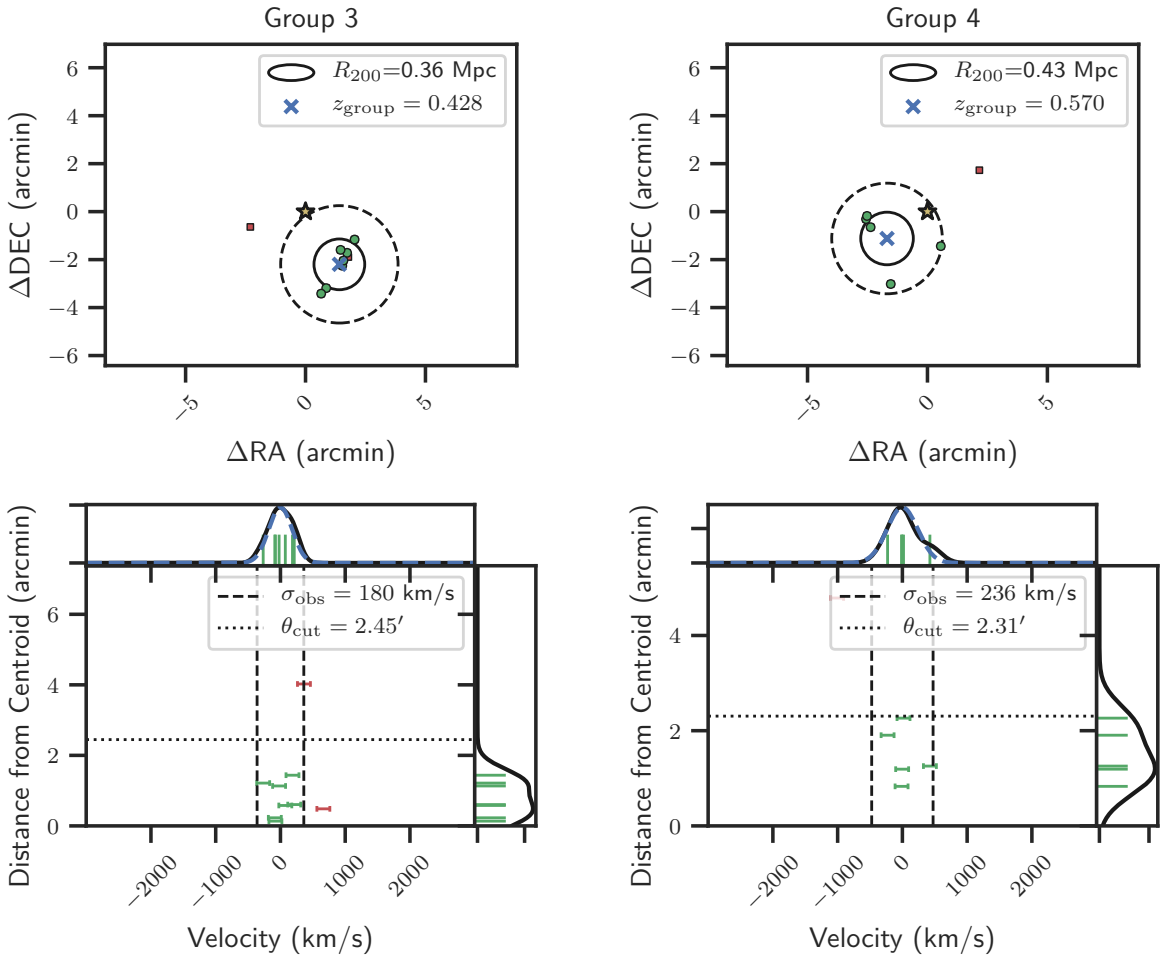

Figure 2.5 (continued)

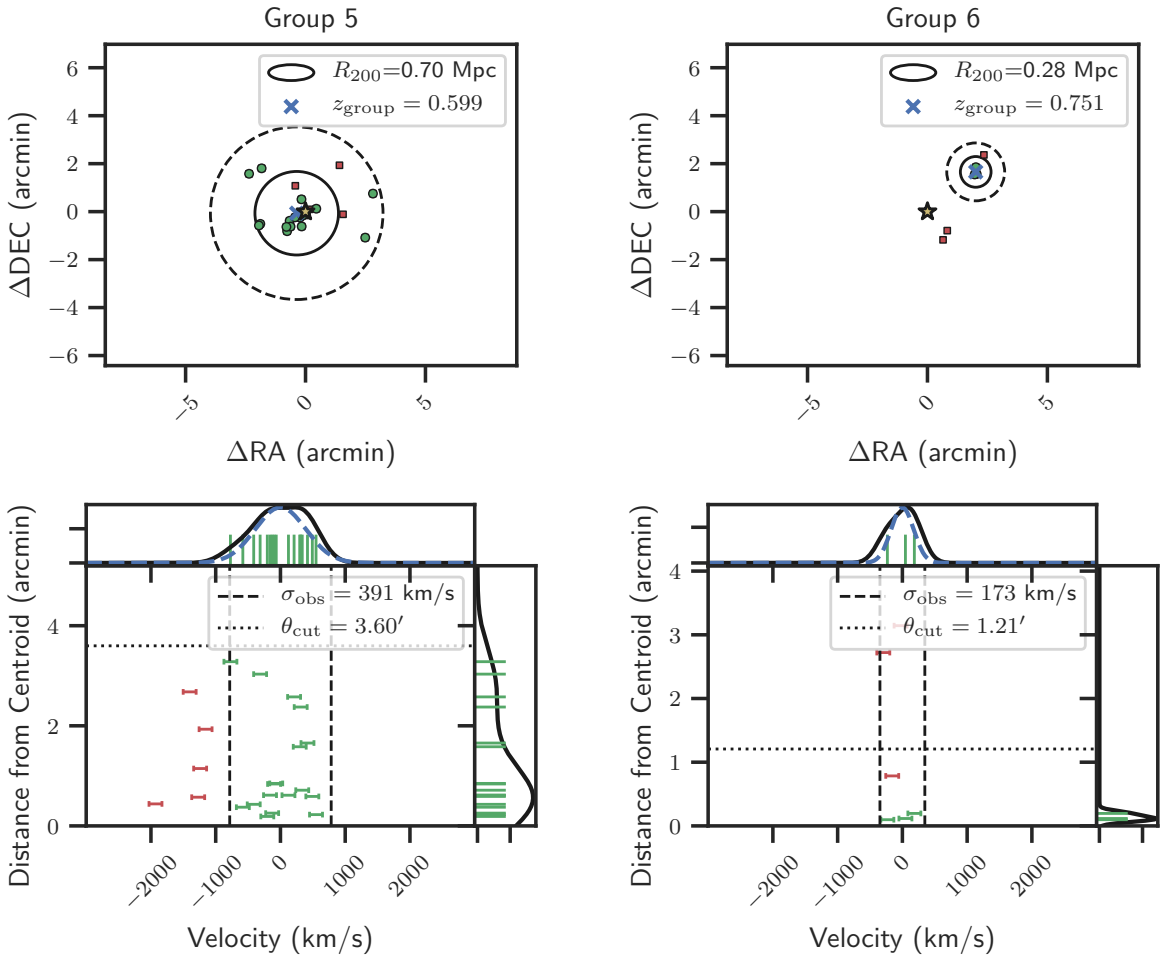

Figure 2.5 (continued)

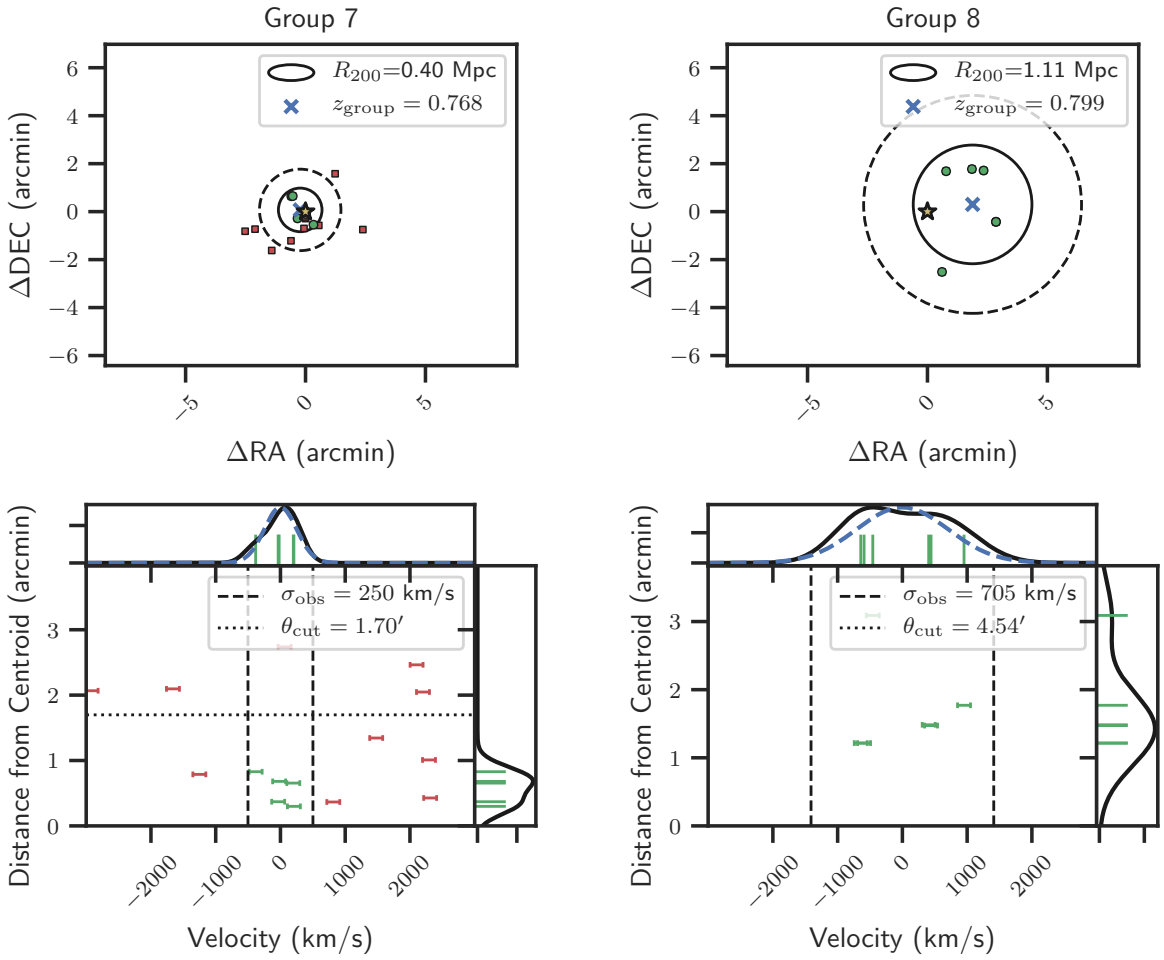

Figure 2.5 (continued)

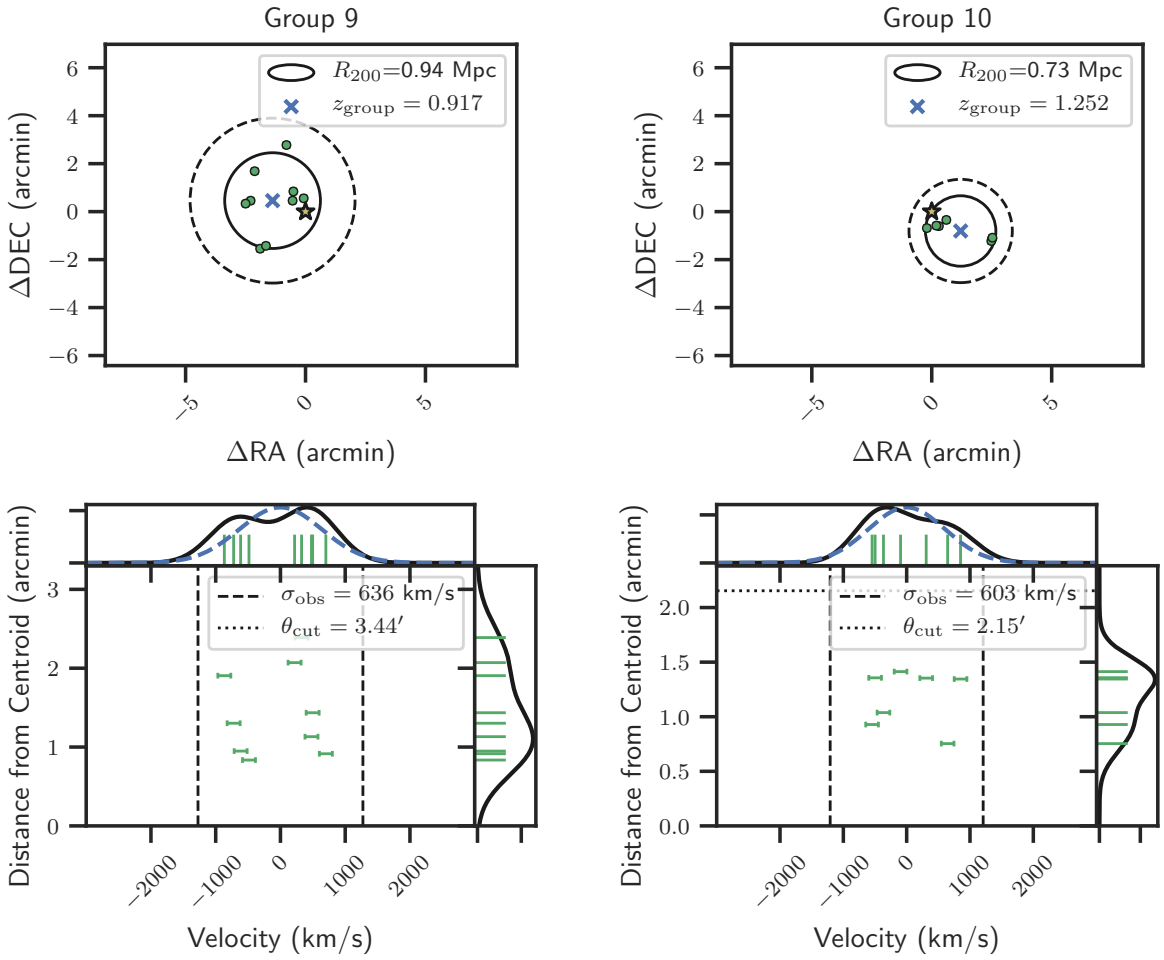

Figure 2.5 (continued)

the observer-frame velocity dispersion of the group.

The largest galaxy group identified in this spectroscopic sample is Group 5, which contains 17 member galaxies, including the lens galaxy of DES J0408−5354. The centroid of this group is close to DES J0408−5354 (26^{+16}_{-12} arcsec), which is also the most luminous member galaxy in the group.

Aside from group 5, the identified groups are generally small, with no identified group containing more than 11 member galaxies. For Groups 1, 3, 4, 6, and 7, the posterior PDFs for the intrinsic velocity dispersions peak at or near zero, and the lower limits are not well-constrained. For these distributions, we report the upper 68th percentile confidence intervals and treat it as the upper limit of the intrinsic velocity dispersion for that group.

For groups 8 and 9, the distribution of member galaxies in velocity space appears to be bimodal, with two separate subgroups separated by $\sim 1000\mathrm{km\,s^{-1}}$. However, there are not enough member galaxies in that redshift range to successfully separate these two subgroups into separate groups as none of the individual subgroups have more than 5 potential members.

The choice of parameters used in the group-finding algorithm described in §2.5.1 can impact the final membership of each galaxy group. As mentioned in §2.5.1, the choice of fiducial values for the initial observer-frame velocity dispersion $\sigma_{\rm obs} = 500$ km s^{-1}, velocity threshold $n = 2$ (Eq. 2.1), aspect ratio $b = 3.5$ (Eq. 2.2) follow that of previous group-finding analyses (Sluse et al., 2017; Sluse et al., 2019; Wilman et al., 2005). We relaxed the parameter n to $n = 3$ to investigate the effect of a more conservative (i.e. more inclusive) choice in the group finding algorithm and found that relaxing the parameters results in groups that contain outlier members, or have bi- or multimodal configurations, all of which are likely to be spurious.

As an additional check, we inspected the Chandra X-ray images of the field (PI: Pooley; Program 20419; ACIS-S; 25ks). No diffuse emission is detected (Pooley, D & Gallo, E. 2019, private communication), with an upper limit (90% CL) of $\sim 10^{44}$ erg/s within a 1Mpc radius

(with considerable error bars depending on the assumed temperature). The non-detection makes it unlikely that the lens galaxy is a member of a galaxy cluster.

2.5.3 Identified Groups in the Environment of WGD 2038−4008

We applied the group-finder to the 54 galaxies with high confidence redshifts in the field of WGD 2038−4008. From this sample, we identified 2 galaxy groups. The results and properties of these galaxy groups are summarized in Table 2.3, and Figure 2.6 show, for each identified galaxy group, the positions of both accepted and rejected member galaxies of that group, as well as the distances and velocities of individual galaxies relative to the group centroid. Group 1 in WGD 2038−4008's field contains the eponymous lens galaxy of that field.

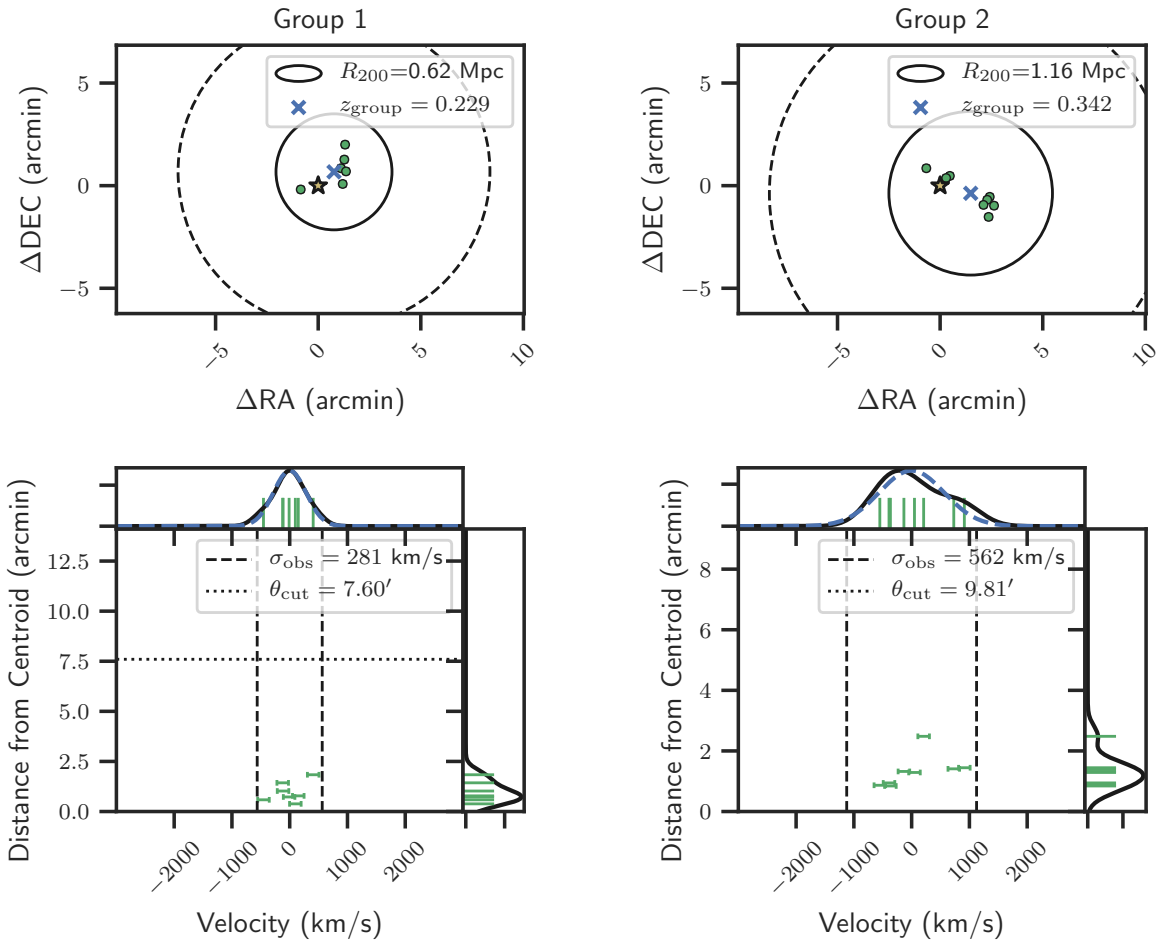

Figure 2.6 Galaxy groups identified using the spectroscopic sample of galaxies in the environment of WGD 2038−4008. See Figure 2.5 for an explanation of the figures.

2.6 Contribution of Environment Galaxies and Galaxy-Groups to the lens structure

2.6.1 Flexion Shift Formalism

A major objective of this analysis is to identify galaxies or galaxy groups along the line of sight or in the environment of lensing systems that significantly perturb the lensing potential of that system and require explicit modeling in the cosmological analysis. Weak perturbations can be treated as a tidal perturbation, producing external convergence and external shear at the position of the lens galaxy. However, perturbers that produce non-negligible third order (i.e. flexion) and higher perturbations to the lens potential cannot be well-approximated by the tidal approximation and will need to be explicitly treated in the lens model. Since the largest terms in the lens potential that are neglected in the tidal approximation are the third-order flexion terms, McCully et al. (2017a) proposes a simple diagnostic, the "*flexion shift*", which is a measurement of the difference in lensed image positions caused by the inclusion of the flexion terms from a perturber. As long as the flexion shift of a perturbing galaxy or galaxy group is much smaller than the astrometric precision of the observed lensed images, the perturber does not need to be explicitly treated in the lens model in a cosmographical analysis.

Following McCully et al. (2017a), the flexion shift is given by

$$\Delta_3 x = f(\beta) \times \frac{(\theta_{\rm E}\,\theta_{\rm E,p})^2}{\theta^3}, \tag{2.5}$$

where $\theta_{\rm E}$ and $\theta_{\rm E,p}$ are the Einstein radii of the main lens and perturber respectively, and θ is the angular separation on the sky between them. The function $f(\beta)$ is

$$f(\beta) \in \left\{ \begin{array}{ll} (1-\beta)^2 & \text{if perturber is behind the lens} \\ 1 & \text{if perturber is in the foreground} \end{array} \right\}, \tag{2.6}$$

where

$$\beta = \frac{D_{\rm dp} D_{\rm os}}{D_{\rm op} D_{\rm ds}}, \tag{2.7}$$

is a combination of angular diameter distances involving the observer (o), deflector (d), perturber (p), and source (s), where the subscripts $D_{ij} = D(z_i, z_j)$ indicate the angular diameter distance between redshifts z_1 and z_2.

This diagnostic provides a simple quantity to estimate the difference in lensed image positions caused by the leading order non-tidal (i.e. third-order) perturbation produced from a perturber. McCully et al. (2017a) showed that by explicitly modeling perturbers with flexion shifts larger than the conservative limit of $\Delta_3 x > 10^{-4\prime\prime}$, we can constrain the bias on H_0 due to this uncertainty to the percent level. We explain how the Einstein radii, $\theta_{\rm E,p}$, as well as the flexion shift uncertainties for each perturber, are determined for galaxies in §2.6.1, and for galaxy groups in §2.6.1.

We calculated the flexion shift for all galaxies in the spectroscopic survey, as well as the flexion shift of all galaxy groups identified from the survey (§2.6.1). For individual galaxies, we exclude 4 objects that are in the DES Y1 Gold catalog but not in the DES Y3 Gold catalog. We also exclude 12 galaxies with spectroscopic redshifts from MUSE because they do not have DES Y3 photometry. In addition, one galaxy in the spectroscopic sample (`488065214`) was found to have bad MOF magnitudes (`MOF_CM_MAG_CORRECTED` magnitudes of -9999 for all bands), and two other galaxies (`488069251, 488066060`) were found to have bad MOF fits, with unrealistically large sizes (`MOF_CM_T` values on the order of ∼5000 square arcseconds) and `MOF_CM_MAG_CORRECTED_I` magnitudes that are brighter than their `MAG_AUTO_CORRECTED_I` magnitudes by more than 4 magnitudes (18.638176 and 17.868253

compared to 23.349312 and 22.776083 respectively). For these three galaxies with spurious MOF photometry, we used the `MAG_AUTO_CORRECTED` photometry to calculate the stellar masses of these galaxies instead.

For completeness, we also calculated the flexion shifts of all likely galaxies with photometric redshift estimates in the DES "Y3 Gold" photometric catalog within $10'$ of the lens galaxies, excluding galaxies that are in the spectroscopic sample. To do this, we made the following selections to the DES Y3 Gold catalog: First, we selected all objects within $10'$ of the lens galaxies in the Y3 Gold catalog that satisfied `FLAGS_GOLD = 0` and `EXTENDED_CLASS_MASH_MOF` $\geq$ `2`, which selected likely galaxies. From this catalog, we excluded all galaxies that have `COADD_OBJECT_ID`s that matched galaxies already in the spectroscopic sample. We then selected objects with DNF photometric redshifts that satisfied `DNF_ZMEAN_SOF > 0` and `DNF_ZSIGMA_SOF < 10`, which removed several objects with spurious redshifts. Finally, we made the following cuts specific to the field of each lens galaxy: For DES J0408−5354, we removed objects with `COADD_OBJECT_ID = 488068193, 488069583, 488067795`, as they are features of the lens system and not galaxies in the environment of the lens. For WGD 2038−4008, we removed objects with `COADD_OBJECT_ID = 169192447, 169193208, 169192589, 169193438` as they are misclassified stars, and reincluded `169190696`, which is a galaxy misclassified as a star[3]. After applying these selection criteria, we obtained a photometric catalog of 5082 objects within $10'$ of DES J0408−5354 and 4438 objects within $10'$ of WGD 2038−4008. We then performed the same analysis on these objects as on the spectroscopic sample, using `DNF_ZMEAN_SOF` instead of spectroscopic redshift when necessary.

For the lens galaxies, we use the following quantities in our analysis: For DES J0408−5354, we use coordinates $\{\mathrm{RA}, \mathrm{DEC}\} = \{62.090417, -53.899889\}$, lens redshift $z_\mathrm{d} = 0.59671$, source redshift $z_\mathrm{s} = 2.375$, and Einstein radius $\theta_\mathrm{E} = 1.80$ (Lin et al., 2017; Shajib et al.,

3. When comparing with the available *HST* data, we found that five objects in the WGD 2038−4008 field were erroneously classified by this pipeline, mainly in the form of stars wrongly classified as galaxies.

2019b). For WGD 2038−4008, we use $\{\mathrm{RA}, \mathrm{DEC}\} = \{309.511379, -40.137024\}$, $z_\mathrm{d} = 0.22829$, $z_\mathrm{s} = 0.777$, $\theta_\mathrm{E} = 1.38$ (Agnello et al., 2018; Shajib et al., 2019b).

Individual Galaxies

We follow the general methodology described in Sluse et al. (2019) to estimate the Einstein radii of galaxies.

First, we inferred the stellar masses of galaxies from DES photometry using the galaxy template fitting code Le PHARE (see §2.3.3). We then use an empirical scaling relation to estimate the line-of-sight central velocity dispersion of the galaxy, σ. In this work, we use and compare results derived from two different scaling relations, one from Zahid et al. (2016) and another from Auger et al. (2010a). The Zahid et al. (2016) relation was derived from a sample of $\sim 3.7 \times 10^5$ SDSS elliptical galaxies at z ¡ 0.7 with stellar masses in the range $\log_{10}(M_\star/M_\odot) \in [9.5, 11.5]$. The relation is fit with a broken power law given by Eqn 5 of Zahid et al. (2016), which we rewrite here in logarithmic form:

$$\begin{aligned} \log_{10}(\sigma) &= \log_{10}(\sigma_b) + \alpha_1 \left(\log_{10}(M_\star) - \log_{10}(M_b)\right) \text{for } M_\star \leq M_b \\ \log_{10}(\sigma) &= \log_{10}(\sigma_b) + \alpha_2 \left(\log_{10}(M_\star) - \log_{10}(M_b)\right) \text{for } M_\star > M_b \end{aligned} \tag{2.8}$$

where $\log_{10}(\sigma_b) = 2.073$, $\log_{10}(M_b/M_\odot) = 10.26$, $\alpha_1 = 0.403$, and $\alpha_2 = 0.293$. Since Zahid et al. (2016) found no significant change in the scaling relation at different redshift bins, we assume that the stellar-mass-to-velocity-dispersion scaling relation does not evolve with redshift.

Alternatively, we also used the scaling relation from Auger et al. (2010a), which was obtained from fitting a sample of 73 elliptical galaxy lenses from the SLACS survey. The

best-fit relation is

$$\log_{10}(\sigma) = 0.18 \log_{10}\left(M_\star/(10^{11} M_\odot)\right) + 2.34 \tag{2.9}$$

where we have opted to use the best-fit parameters for the model that includes the intrinsic scatter, which is 0.04 ± 0.01 in the fit. The elliptical galaxies used for this fit are generally more massive compared to the sample used in the Zahid et al. (2016) analysis and have stellar masses in the range $\log_{10}(M_\star/M_\odot) \in [10.5, 12]$. We also assume that the stellar-mass-to-velocity-dispersion scaling relation does not evolve with redshift.

Once we obtain line-of-sight velocity dispersion estimates for each galaxy, we convert the velocity dispersion to the Einstein radius, $\theta_{\rm E,p}$, assuming a Singular Isothermal Sphere (SIS) model

$$\theta_{\rm E,p} = 4\pi \left(\frac{\sigma}{c}\right)^2 \frac{D_{\rm ps}}{D_{\rm os}}, \tag{2.10}$$

from which the flexion shift can be calculated with Eq.(2.5).

The uncertainties for the flexion shifts are calculated by adding two different sources of uncertainty in quadrature. The first source comes from the uncertainty in the stellar mass estimates from Le PHARE. The second source of uncertainty comes from the intrinsic scatter in the scaling relation between stellar mass and velocity dispersion. For the Zahid et al. (2016) relation, we quantify this uncertainty by taking half the difference in the central 68th percentile limits of the velocity dispersion distribution at a given stellar mass and use that as the uncertainty from the intrinsic scatter of the scaling relation (see Figure 9(A) of Zahid et al., 2016). For the Auger et al. (2010a) relation, we use their fit for the intrinsic scatter, which is $\Delta \log_{10}(\sigma) = 0.05$ (taking the more conservative limit). The two sources of uncertainties are added in quadrature and then propagated forward into an uncertainty in

the flexion shift.

Flexion shift estimates for galaxies in the spectroscopic sample with stellar mass estimates that are significantly outside the mass ranges used to derive the Zahid et al. (2016) and Auger et al. (2010a) scaling relations should be treated with caution, as the scaling relations (and errors) are extrapolated. Therefore, flexion shift estimates for galaxies in our sample with stellar masses $\log_{10}(M_\star/M_\odot) < 9.5$ should be treated with caution, as both the scaling relations from Zahid et al. (2016) and Auger et al. (2010a) may not be valid at the lower end of the stellar mass range. However, the validity of this extrapolation does not affect the main results of this study, since the most significant perturbers (i.e. galaxies that contribute the largest flexion shift contributions) tend to be more massive. The 10 galaxies with the largest flexion shift contributions at the lens positions of DES J0408−5354 and WGD 2038−4008 (Table 2.4) are within the stellar mass ranges used to derive at least one of the two scaling relations.

Galaxy Groups

For galaxy groups, we obtain a probability density function for the Einstein radius by adopting the same SIS approximation described in Eq. (2.10), and sampling 1000 values from the posterior PDF of the intrinsic velocity dispersion of the groups identified in §2.5.1 as well as from the bootstrapped PDF of the redshift of each group.

To obtain the flexion shifts of the galaxy groups and corresponding uncertainties, we use Eq. (2.5), sampling from the PDF of the Einstein radius as well as the bootstrapped group centroid position.

Table 2.4 Properties of the 10 galaxies with the largest flexion shifts, sorted in order of decreasing 68th percentile upper limits, in the field of DES J0408−5354 and WGD 2038−4008. The columns display, in order, the DES Y3 Object ID (and ID used in future papers), coordinates (RA, DEC in degrees; ICRS), redshift z, distances to the lensing galaxy, and flexion shifts, calculated using the scaling relations by Zahid et al. (2016) and Auger et al. (2010a) respectively. Galaxies marked with $*$ are not in the spectroscopic survey and only have photometric redshifts; we report the `DNF_ZMEAN_SOF` redshift value and `DNF_ZSIGMA_SOF` uncertainties. Spectroscopic redshift uncertainties are about 100 $\mathrm{km\,s^{-1}}$or 0.00033 in redshift. Stellar masses and corresponding uncertainties were calculated using the `Le PHARE` galaxy template fitting code and DES Y3 photometry (see §2.3.3). Galaxies marked with † have spurious `MOF` magnitudes. For these galaxies, we use `MAG_AUTO_CORRECTED` photometry to calculate stellar masses instead. Flexion shifts and uncertainties are calculated following the method described in §2.6.1-2.6.1. For a complete list of galaxies, see Table A1 of the published version of this chapter (Buckley-Geer et al., 2020).

ID	RA (deg)	DEC (deg)	z	i-band Magnitude	$\log_{10}(M_*)$ ($log_{10}(M_\odot)$)	$\Delta\theta$ (arcsec)	$\log_{10}(\Delta_3 x_{\mathrm{Zahid}})$ ($\log_{10}$(arcsec))	$\log_{10}(\Delta_3 x_{\mathrm{Auger}})$ ($\log_{10}$(arcsec))
DES J0408−5354								
488068102 (G3)	62.090965	-53.901634	0.76866	20.096	$11.42^{+0.10}_{-0.12}$	6.4	$-2.21^{+0.25}_{-0.29}$	$-2.20^{+0.21}_{-0.22}$
488065185 (G6)	62.095575	-53.898291	0.59441	21.867	$10.29^{+0.10}_{-0.17}$	12.4	$-3.95^{+0.42}_{-0.58}$	$-3.42^{+0.21}_{-0.24}$
488066144 (G4)	62.090243	-53.903609	0.77069	22.015	$10.66^{+0.13}_{-0.17}$	13.4	$-4.07^{+0.38}_{-0.51}$	$-3.72^{+0.22}_{-0.24}$
488066768 (G5)	62.092923	-53.900091	1.03197	23.682	$9.68^{+0.24}_{-0.25}$	5.4	$-4.93^{+0.63}_{-0.93}$	$-3.87^{+0.26}_{-0.27}$
488066462	62.083912	-53.903969	0.60048	22.304	$9.88^{+0.12}_{-0.19}$	20.2	$-5.24^{+0.52}_{-0.79}$	$-4.37^{+0.22}_{-0.25}$
MUSE6 / 488065214 †	62.077898	-53.903878	0.59797	22.303	$10.37^{+0.11}_{-0.16}$	30.2	$-5.02^{+0.41}_{-0.56}$	$-4.53^{+0.21}_{-0.23}$
488070148	62.084648	-53.885204	0.27247	18.309	$10.99^{+0.04}_{-0.04}$	54.3	$-4.81^{+0.31}_{-0.37}$	$-4.61^{+0.20}_{-0.20}$
488066886	62.090541	-53.904725	0.77568	22.925	$9.91^{+0.16}_{-0.30}$	17.4	$-5.46^{+0.53}_{-0.90}$	$-4.61^{+0.23}_{-0.30}$
488070807	62.072770	-53.910296	0.59811	19.975	$11.27^{+0.05}_{-0.06}$	53.0	$-4.69^{+0.25}_{-0.30}$	$-4.61^{+0.20}_{-0.21}$
MUSE8 / 488067782	62.103108	-53.897861	0.59832	22.100	$10.11^{+0.13}_{-0.20}$	27.9	$-5.29^{+0.48}_{-0.71}$	$-4.62^{+0.22}_{-0.25}$
WGD 2038−4008								
169192350 $*$	309.514393	-40.137815	0.349 ± 0.033	21.999	$9.53^{+0.15}_{-0.29}$	8.8	$-5.71^{+0.59}_{-1.03}$	$-4.52^{+0.22}_{-0.30}$
169193255 $*$	309.529391	-40.151522	0.240 ± 0.138	18.806	$10.76^{+0.06}_{-0.10}$	72.0	$-5.94^{+0.35}_{-0.45}$	$-5.63^{+0.20}_{-0.21}$
169190952 $*$	309.503723	-40.120144	0.277 ± 0.021	19.052	$10.78^{+0.05}_{-0.09}$	64.3	$-6.02^{+0.34}_{-0.44}$	$-5.72^{+0.20}_{-0.21}$
169192931 $*$	309.508988	-40.144869	0.421 ± 0.044	20.446	$10.62^{+0.05}_{-0.07}$	29.0	$-6.17^{+0.37}_{-0.47}$	$-5.80^{+0.20}_{-0.21}$
169191098 $*$	309.487406	-40.124006	0.238 ± 0.016	18.360	$10.69^{+0.04}_{-0.03}$	80.9	$-6.16^{+0.36}_{-0.45}$	$-5.82^{+0.20}_{-0.20}$
169191973 $*$	309.536238	-40.134562	0.331 ± 0.011	18.444	$11.15^{+0.03}_{-0.03}$	69.0	$-6.05^{+0.27}_{-0.33}$	$-5.92^{+0.20}_{-0.20}$
169193929 $*$	309.520149	-40.160086	0.250 ± 0.005	19.137	$10.63^{+0.05}_{-0.09}$	86.5	$-6.41^{+0.36}_{-0.47}$	$-6.04^{+0.20}_{-0.21}$
169190452	309.539131	-40.115839	0.22917	18.213	$10.76^{+0.04}_{-0.04}$	107.9	$-6.39^{+0.35}_{-0.43}$	$-6.08^{+0.20}_{-0.20}$
169192596	309.492815	-40.140131	0.23003	21.389	$9.28^{+0.11}_{-0.18}$	52.3	$-7.61^{+0.60}_{-0.96}$	$-6.21^{+0.21}_{-0.24}$
169191228	309.535586	-40.122881	0.22900	20.025	$10.13^{+0.07}_{-0.11}$	83.9	$-6.86^{+0.45}_{-0.62}$	$-6.21^{+0.21}_{-0.22}$

Note: $*$ Galaxy is not in the spectroscopy survey; stellar masses and flexion shift estimates are from photometric redshifts.
† Galaxy has spurious `MOF` photometry. For these galaxies, we use `MAG_AUTO_CORRECTED` photometry to calculate stellar masses instead.

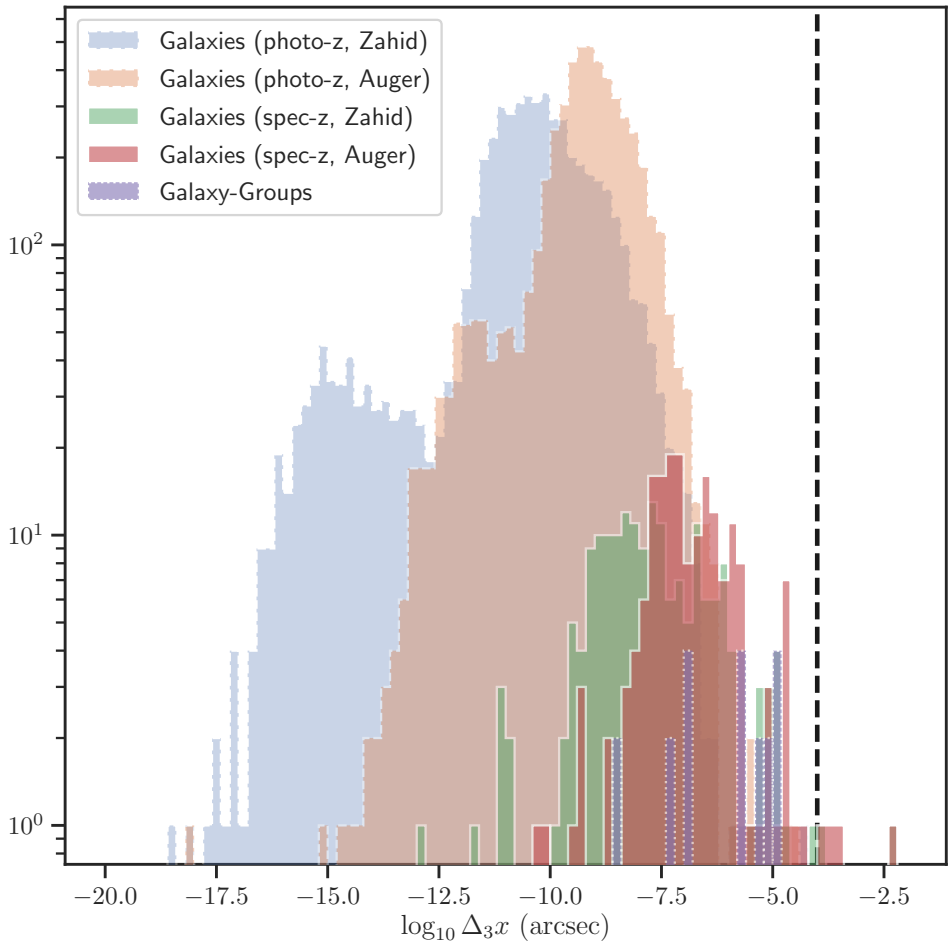

Figure 2.7 Flexion shift histograms for galaxies and galaxy groups in the environment of DES J0408−5354.

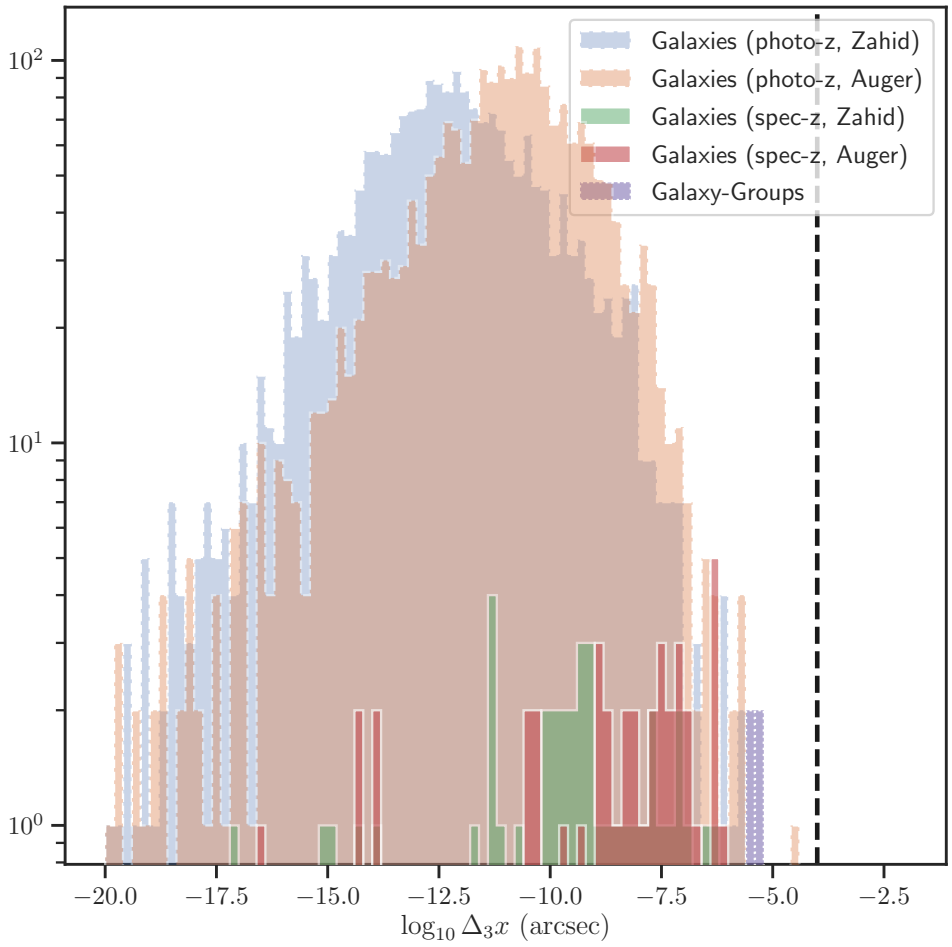

Figure 2.8 Flexion shift histogram for galaxies and galaxy groups in the environment of WGD 2038−4008.

2.6.2 Flexion shifts for galaxies and galaxy groups in the field of DES J0408−5354 and WGD 2038−4008

We present a table of the properties of the 10 galaxies with the largest flexion shifts at the lens position in the fields of DES J0408−5354 and WGD 2038−4008. Comparing the results from the two scaling relations, The scaling relation from Auger et al. (2010a) is shallower than the fit by Zahid et al. (2016) but produces larger estimates of the line-of-sight velocity dispersion for galaxies with stellar masses $\log_{10}(M_\star/M_\odot) \lesssim 11.45$. Since the majority of galaxies in the sample have smaller stellar mass estimates than that, the Auger et al. (2010a) scaling relation produces larger flexion shifts than that from the Zahid et al. (2016) for this sample and can be treated as the more conservative estimate of the two. In figures 2.7 and 2.8, we show the distribution of flexion shifts of the spectroscopic and photometric catalogs for the environment of DES J0408−5354 and WGD 2038−4008, using both the Zahid et al. (2016) and Auger et al. (2010a) scaling relations. The $\log_{10}(\Delta_3 x) > -4$ criteria is indicated by the dashed vertical line.

When calculating flexion shifts for the photometric catalog, we found some objects with spurious `MOF_CM_MAG_CORRECTED` photometry. For these objects, we instead used stellar masses computed from `MAG_AUTO` photometry, including the same Milky Way extinction and other photometric corrections as for the MOF magnitudes.

Using this criterion, for DES J0408−5354, there are four galaxies (G3, G4, G5, G6) with flexion shift contributions above $\log_{10}(\Delta_3 x) > -4$ when using the stellar masses from the Auger et al. (2010a) scaling relation. These four galaxies are explicitly modeled in the lens model analysis (Shajib et al., 2019a). In addition, the group including the lensing galaxy, group 5, has a flexion shift of $-3.86^{+0.97}_{-0.72}$ $''$. The large uncertainties in the flexion shift of group 5 are due to a combination of the proximity of the group centroid to DES J0408−5354 and the uncertainty in the centroid location, with the upper limits of the flexion shift being produced when the group centroid is near the lens galaxy.

For WGD 2038−4008, we did not identify any galaxies or galaxy groups with a flexion shift $\log_{10}(\Delta_3 x) > -4$.

2.6.3 Photometrically Identified Galaxy-Groups

RedMaPPer Clusters in the Field of the Lenses

Due to the low spectroscopic completeness of the survey of WGD 2038−4008, we complemented our spectroscopic group-identification efforts with a search of all photometrically-identified clusters with richness $\lambda > 5$ in the field of view of WGD 2038−4008 using the redMaPPer algorithm Rykoff et al. (2014). We used the sixth release of the redMaPPer cluster catalog on DES Y3A2 data (v6.4.22+2) and found two redMaPPer clusters within $3'$ of WGD 2038−4008 (with unique IDs `MEM_MATCH_ID = 62659, 138669`).

One of the clusters (`62659`) has a photometric cluster redshift of $z_{\rm photo} = 0.221 \pm 0.008$ and a richness of $\lambda = 5.1 \pm 1.7$. The photometric cluster redshift of cluster `62659` is consistent with the spectroscopically-identified Group 1 ($z_{\rm group} = 0.229$), and its algorithmically-identified central galaxy, `COADD_OBJECT_ID = 169190452`, is also a member of Group 1, suggesting that redMaPPer cluster `62659` and Group 1 are the same. Only two of the seven spectroscopically-identified group members in Group 1 (`COADD_OBJECT_ID = 169190452, 169189459`) are also members of the redMaPPer cluster `62659`. However, this could be simply due to bad photometric redshift estimates, as 4 of the 5 group members have `DNF_ZMEAN_SOF` photometric redshift estimates that range from $0.38 - 0.44$, and the lens galaxy (`169191076`) has a spurious redshift estimate of 0.00977.

The second redMaPPER cluster (`138669`) has a cluster redshift of $z_{\rm photo} = 0.405 \pm 0.017$ and a richness of $\lambda = 10.8 \pm 2.0$. None of the galaxies in this cluster share group membership with the spectroscopically-identified Group 2, suggesting that this group is distinct from Group 2.

For completeness, we also searched for redMaPPer clusters in the field of DES J0408−5354,

though the spectroscopic completeness of the field of DES J0408−5354 is much higher than that of WGD 2038−4008. However, we did not find any clusters within $3'$of DES J0408−5354. One reason for this paucity is that a nearby region of the lens has been flagged and precluded from redMaPPer analysis due to a bright star in the foreground.

Flexion Shifts for redMaPPer Clusters near the lenses

We performed the same flexion shift calculations on the redMaPPer clusters as the spectroscopic groups, following the procedure outlined in §2.6.1, using the same lens parameter quantities, and using the same conservative SIS approximation described by Eq. 2.10 for the perturber. We use the scaling relation given by Eq. (17) of Andreon & Hurn (2010) to convert the cluster richness into velocity dispersion estimates for the SIS model. From this, we obtain flexion shifts for the redMaPPer clusters. For cluster `62659`, we obtain a flexion shift of $\log_{10}(\Delta_3 x) = -5.1^{+0.3}_{-0.4}$. For cluster `138669`, we obtain a flexion shift of $\log_{10}(\Delta_3 x) = -6.0 \pm 0.2$. The estimated uncertainties come from propagating both the uncertainties in cluster richness, as well as the uncertainties in the scaling relation. The properties of both groups are summarized in Table 2.5. Neither redMaPPer group in the field of WGD 2038−4008 exceeds the threshold of $\log_{10}(\Delta_3 x) > -4$.

Table 2.5 Properties of redMaPPer clusters in the field of view of WGD 2038−4008. The columns show the cluster ID, cluster redshift, richness, velocity dispersion (rounded to the nearest $10\,\mathrm{km\,s^{-1}}$), the group centroid (in RA and Dec), projected distance of the centroid to the lens ($\Delta\theta$), and flexion shift $\log(\Delta_3 x(\mathrm{arcsec}))$. See §2.6.3 for further discussion

MEM_MATCH_ID	$\bar{z}_{\mathrm{group}}$	λ	σ_{group} ($\mathrm{km\,s^{-1}}$)	$\mathrm{RA_{ctr}}$, $\mathrm{DEC_{ctr}}$ (deg)	$\Delta\theta$ (arcsec)	$\log_{10}(\Delta_3 x)$ ($\log_{10}$(arcsec))
62659	0.221 ± 0.008	5.1 ± 1.7	340^{+60}_{-70}	309.53913, -40.11584	107.9	$-5.1^{+0.3}_{-0.4}$
138669	0.405 ± 0.017	10.8 ± 2.0	430^{+50}_{-60}	309.48707, -40.16080	108.6	-6.0 ± 0.2

2.7 Conclusions

Galaxies and groups along the line of sight that are massive enough and close to the lens can constitute a significant source of systematic uncertainty to the measurement of the Hubble constant H_0. In this chapter, we present an analysis of such potential systematics in the environment of lensed quasars DES J0408−5354 and WGD 2038−4008. This requires spectroscopic redshifts for the galaxies in the fields of the two lenses. To obtain these we have carried out spectroscopic observations using Gemini South/GMOS-S, Magellan/LDSS-3, and VLT/MUSE. As detailed in §2.3.1-2.3.2, we obtained a total of 199 high-confidence redshifts from the three instruments for DES J0408−5354, corresponding to redshift completeness of 68% for galaxies with $18 \leq i < 23$ and $5'' \leq$ radius $< 3'$. For WGD 2038−4008 we obtained 54 high-confidence redshifts from the Gemini South/GMOS-S data, with a 16% redshift completeness for the same i-band magnitude and radius ranges.

As described in §2.4, in our redshift survey observations we also set aside slits to measure the stellar velocity dispersions of the main lensing galaxies in our two systems. The velocity dispersion of the main deflector G1 in DES J0408−5354 was measured using four independent spectra from the above three instruments, with a consistent result of about 230 $\mathrm{km\,s^{-1}}$(see Table 2.2). The velocity dispersion for the main lensing galaxy G in WGD 2038−4008 was obtained from one spectrum taken using Gemini South/GMOS-S, with a resulting value of about 300 $\mathrm{km\,s^{-1}}$. The detailed velocity dispersion measurements and uncertainties are given in Table 2.2.

The galaxy group identification uses the spectroscopic redshifts described above and the same algorithm that was used for the analysis of the H0LiCOW lenses HE 0435−1223 (Sluse et al., 2017) and WFI 2033−4723 (Sluse et al., 2019). We find 10 galaxy groups in DES J0408−5354 for which we then compute the flexion shift $\Delta_3 x$ proposed by McCully et al. (2017a). McCully et al. (2017a) showed that explicitly modeling perturbers with flexion shifts larger than $\Delta_3 x > 10^{-4}{}''$ allows one to constrain the bias on H_0 due to this uncertainty to

the percent level. Out of our 10 groups in DES J0408−5354 we find one group (labelled 5 in Table 2.3) that has a flexion shift of $\log_{10}\Delta_3 x = -3.86^{+0.97}_{-0.72}$. This group has 17 members, one of which is the lensing galaxy G1 and the centroid of this group is close to the lens. However Shajib et al. (2019a) shows that the change in H_0 of including this group would result in a decrease of approximately 0.4 percent so it is not explicitly included in the mass model. For WGD 2038−4008, we find two galaxy groups from our spectroscopic redshift sample, but neither group has a flexion shift above our cut.

To calculate the flexion shift for individual galaxies we start with the general methodology described in Sluse et al. (2019) and then use two different scaling relations (Zahid et al., 2016; Auger et al., 2010a) to estimate the line-of-sight central velocity dispersion of the galaxy from its stellar mass. The stellar masses are calculated using a galaxy model fitting to DES photometry, as detailed in §2.3.3. The Auger et al. (2010a) relation between velocity dispersion and stellar mass gives more conservative, larger flexion shift values, resulting in four galaxies in DES J0408−5354 with flexion shifts larger than our $\log_{10}\Delta_3 x = -4.0$ cut; these galaxies (G3, G4, G5, and G6 in Table 2.4) are therefore selected for explicit modeling by Shajib et al. (2019a). For WGD 2038−4008, we do not find any individual galaxies with flexion shifts greater than our cut.

CHAPTER 3

MASS AND LIGHT PROPERTIES OF FOUR EARLY-TYPE GALAXY STRONG LENSES IN THE DARK ENERGY SURVEY YEAR 1 DATA

3.1 Introduction

Strong gravitational lensing is the phenomenon whereby light from a distant astrophysical source is lensed by a massive foreground object, producing a distinct morphology. The amplification of the source light (typically from a galaxy or quasar) being lensed allows for the observation and analysis of high redshift sources that would otherwise be too faint to be imaged (e.g. Bayliss et al. (2014); Hainline et al. (2009); Quider et al. (2009); Smit et al. (2017); Dessauges-Zavadsky et al. (2009); Christensen et al. (2012); James et al. (2013); Finkelstein et al. (2009)). Strong lensing systems have also been used as cosmographical distance probes of the expansion history of the universe, which can be used in turn to constrain cosmological models (e.g. Refsdal (1964); Suyu et al. (2010a, 2013, 2016); Bonvin et al. (2017); Birrer et al. (2016); Gavazzi et al. (2008); Collett & Auger (2014); Collett et al. (2012); Cardone et al. (2015); Cao et al. (2015a); Grillo et al. (2007)).

Strong gravitational lensing is also an established non-dynamical probe of the projected dark and baryonic matter distribution in lensing systems at cosmological distances. Early-type galaxies constitute the majority of galaxy-scale lenses, and studies on populations of Early-type galaxy lenses have yielded fruitful studies on their density profiles, dark matter mass fraction, density slope, and stellar initial mass function (e.g. Auger et al. (2010b); Barnabe et al. (2011); Newman et al. (2015, 2012a,b); Cardone et al. (2009))

Currently, several hundred early-type galaxy lenses have been discovered. Most of them have been found through dedicated surveys, such as the Strong Lensing Arcs Survey (SLACS), BOSS Emission-Line Survey (BELLS), Lensing Structure and Dynamics Survey (LSD), and

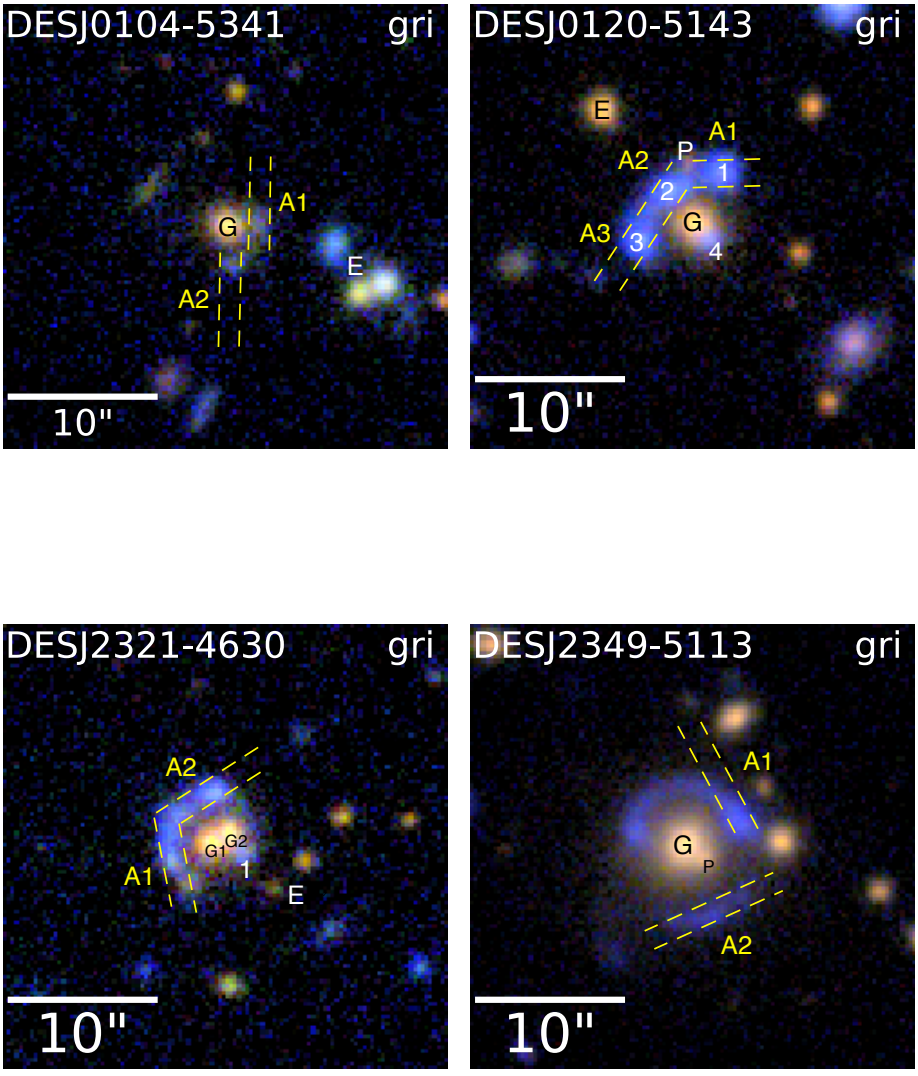

Figure 3.1 *gri*-band composite color co-added DES images of the five lensed systems described in this work. Images are oriented north up, east left and are $30'' \times 30''$ in area. Object labels beginning with 'G' denote the main lensing galaxy/galaxies, while object labels beginning with 'A' follow that of Nord et al. (2020), and refer to lensing features targeted for spectroscopy, with the dashed line marking the approximate spectroscopy slit positions. Numbered objects refer to distinct lensed images that can be visually identified. Nearby galaxies in the lens environment are labeled with the letter 'E'. For DESJ0120-5143, there is also a galaxy labeled 'P' situated between lensed images 1 and 2 that may be in the foreground or environment of the lens galaxy.

Table 3.1 Positions and redshifts for lensing systems in this study

System	Object	RA (J2000)	Dec. (J2000)	$z_{d,photo}$ or $z_{s,spec}$
DESJ0104-5341	G	16.098363	-53.699436	0.679 ± 0.022
	A1	16.097533	-53.699283	1.2318 ± 0.00012
	A2	16.098237	-53.700129	1.2318 ± 0.000057
DESJ0120-5143	G	20.175979	-51.731407	0.603 ± 0.063
	A1	20.177571	-51.731638	1.29566 ± 0.0005
	A2	20.176720	-51.730802	1.29553 ± 0.000052
	A3	20.175358	-51.730461	1.29566 ± 0.000047
DESJ2321-4630	G1	350.368403	-46.513765	0.70 ± 0.02
	G2	350.367937	-46.513625	0.64 ± 0.03
	A1	350.369406	-46.514079	1.7479 ± 0.0031
	A2	350.368353	-46.512958	1.7469 ± 0.0014
DESJ2349-5113	G	357.375254	-51.22752	0.613 ± 0.043
	A1	357.374425	-51.228842	1.3932 ± 0.0001
	A2	357.373730	-51.227069	1.3938 ± 0.0001

Notes. Photometric redshifts are provided for lensing galaxies (object labels starting with 'G'), while spectroscopic redshifts are listed for the lensed features (object labels starting with 'A'). Positions and Photometric redshifts are from the DESDM database. Additional information may be found in Nord et al. (2020).

Table 3.2 Photometric information for lensing galaxies from lens modeling

System	Object	g	r	i	z	Y
DESJ0104-5341	G	22.0 ± 0.3	20.1 ± 0.3	19.1 ± 0.3	18.2 ± 0.3	18.1 ± 0.3
DESJ0120-5143	G	20.8 ±0.05	19.4 ± 0.05	18.6 ± 0.05	17.7 ± 0.05	18.1 ± 0.05
	P	25.9 ±0.2	24.5 ±0.2	23.7 ±0.2	22.9 ±0.2	23.2 ±0.2
DESJ2321-4630	G1	21.385 ± 0.005	20.150 ± 0.005	19.009 ± 0.005	18.966 ± 0.005	18.301 ± 0.005
	G2	21.704 ± 0.009	20.437 ± 0.009	19.427 ± 0.009	19.469 ± 0.009	18.894 ± 0.009
DESJ2349-5113	G	19.209 ± 0.013	17.639 ± 0.013	16.785 ± 0.013	16.818 ± 0.013	16.569 ± 0.013

the Strong Lensing Legacy Survey (SL2S). Wide-field surveys provide prime opportunities to discover more strong lensing systems for future follow-up observations. The Dark Energy Survey (DES) is one such survey, covering 5000 square degrees in the southern hemisphere in the g, r, i, z and Y photometric bands over five years (Flaugher et al., 2015). With the large survey footprint and high image quality from the Dark Energy Camera (DECam) (Flaugher et al., 2015), DES is predicted to discover ~ 2000 strong lensing system from the final dataset, which would effectively triple the current census of confirmed strong lensing systems (Collett et al., 2015). A variety of photometry-based searches have been used to identify lens candidates in the DES data thus far (e.g. Diehl et al. (2017)). Semi-automated arc-finding (e.g. More et al. (2012b)) and ring-finding algorithms (e.g. Gavazzi et al. (2014a)) are also being employed in the DES lens search. These searches have yielded hundreds of strong lensing candidates from DES imaging data. Of these candidates, six group- and cluster-scale lensed systems have been spectroscopically confirmed in the DES Science Verification data (Nord et al., 2016), and an additional nine systems have since been spectroscopically confirmed in the Year 1 data using GMOS on Gemini South as part of the Gemini Large and Long program GS-2014B-LP-5[1](Nord et al., 2020). Additionally, eight lensed quasars have been discovered in DES survey data to date (Agnello et al., 2015; Ostrovski et al., 2017; Lin et al., 2017; Anguita et al., 2018) under the STRong-lensing Insights into Dark Energy Survey (STRIDES) program[2] (Treu et al., 2018a).

In this paper, we present the results of a lens modeling analysis of four galaxy-scale lenses spectroscopically confirmed in the DES Year 1 data. Forward modeling was conducted on DES *gri*-band image cutouts of the confirmed lensed systems from survey data to obtain estimates of their physical Einstein radius, enclosed mass, luminosity, and mass-to-light ratio. The paper is organized as follows: In §2, we briefly describe the four DES Year 1

1. `http://www.gemini.edu/?q=node/12238#Buckley`

2. `http://strides.astro.ucla.edu/`

spectroscopically confirmed lenses and their properties. In §3, we describe our lens modeling procedure. In §4 we describe our results and in §5 we compare our measurement of the dark matter mass fraction to the ones from SLACS (Cardone et al., 2009). Finally, we summarize and present our conclusions in §6. Throughout this paper, a standard flat ΛCDM cosmology of $\Omega_{\rm m} = 0.3$ and $h = 0.7$ is assumed whenever required.

3.2 Lens Systems Information

In this paper, we carried out lens modeling on four galaxy-scale lenses discovered in the Y1A1 data: DESJ0104-5341, DESJ0120-5143, DESJ2321-4630, and DESJ2349-5113. The search process that produced the four lenses described in the paper is described in detail in Diehl et al. (2017). Three of the systems, DESJ0104-5341, DESJ0120-5143, and DESJ2349-5113, were first identified in the visual scan of 7328 galaxy clusters identified by the redMaPPer galaxy clustering algorithm (Rykoff et al., 2014) in the DES Year 1 data release (Y1A1). The other system, DESJ2321-4630 was identified in a sample of DES year 1 luminous red galaxies selected using the redMaGiC algorithm (Rozo et al., 2015) in association with three or more nearby blue objects, where blue is defined as $-1 \leq g - r < 1$ and $-1 \leq r - i < 1$ and $r < 21$. Two of the systems, DESJ0120-5143 and DESJ2321-4630 were also identified by the catalog search techniques described in Diehl et al. (2017).

The full details of the discovery and spectroscopic confirmation of the strong lensing systems discovered in the DES Year 1 data are presented in a separate paper (Nord et al., 2020). Here, we summarize the relevant information on these four strong lensing systems. Fig. 3.1 shows the *gri* color co-added images of the strong lensing systems, while Tables 3.1 summarize the positions, and redshift information relevant to this work.

3.2.1 DESJ0104-5341

The central red lens galaxy (G) has a photometric redshift of $z_{\rm d} = 0.68 \pm 0.02$. There is a prominent blue arc situated close to the lens galaxy and extending around most of the southern and western quadrants of the lens galaxy (Fig. 3.1(a)). A faint counterimage may be present northeast of the lens galaxy. [OII]3727 lines were observed from two points on the lensed arc (A1 and A2) at 8318.68Å and 8318.54Å respectively, yielding a spectroscopic redshift of $z_{\rm s,spec} = 1.231800 \pm 0.000057$ and $z_{\rm s,spec} = 1.2318 \pm 0.0001$.

3.2.2 DESJ0120-5143

The central red lens galaxy (G) has a photometric redshift of $z_{\rm d} = 0.60 \pm 0.06$. The source image is quadruply lensed, with three blue images (1, 2, and 3) to the north and east of the lens galaxy and a counterimage (4) partially superimposed over the south and west of the lens galaxy (Fig. 3.1(b)). [OII]3727 emission lines were identified in follow-up spectroscopy of the lensed images at A1, A2 and A3 at 8556.6Å, 8556.13Å and 8556.62Å respectively, corresponding to spectroscopic redshifts of $z_{\rm s,spec} = 1.29566 \pm 0.000047$, $z_{\rm s,spec} = 1.29553 \pm 0.000052$ and $z_{\rm s,spec} = 1.29566 \pm 0.0005$. There is also an additional galaxy (P) situated in between images 1 and 2 that may be in the foreground. Unfortunately, due to a lack of spectroscopic redshift for P, we are unable to confirm if it is indeed a foreground or background galaxy. Therefore, in our lens modeling, we account for both possibilities by modeling the lens with and without the addition of a second perturbing mass situated at the location of P.

3.2.3 DESJ2321-4630

This lensed system has two red lensing galaxies (G1, G2), a lensing arc extending between the regions marked A1 and A2, as well as a counter-image (1) south and west of the lenses (Fig. 3.1(c)). The lensing galaxies G1 and G2 have photometric redshifts of $z_{\rm d} = 0.64 \pm 0.03$

and $z_{\rm d} = 0.70 \pm 0.02$ respectively, and the lensed features at A1 and A2 were spectroscopically confirmed using the [CII], [NeIV], [MgII], [NeV] and [OII] lines to have a redshift of $z_{\rm s,spec} = 1.7479 \pm 0.0031$ and $z_{\rm s,spec} = 1.7469 \pm 0.0014$ respectively.

3.2.4 DESJ2349-5113

The lensing system's central lens galaxy (G) has a photometric redshift of $z_{\rm d} = 0.613 \pm 0.043$, with the extended lensed source regions A1 and A2 spectroscopically confirmed to be at redshifts of $z_{\rm s,spec} = 1.3932 \pm 0.0001$ and $z_{\rm s,spec} = 1.3938 \pm 0.0001$ respectively. There appears to be a smaller galaxy (P) in the environment of the lens galaxy G.

3.3 Lens Modeling Method

We carried out lens modeling using the `pylens` and `imageSim` forward modeling packages (Auger et al., 2010a, 2013; Meng et al., 2015). The lens modeling software `lensed` (Tessore et al., 2016) was also used to quickly obtain initial estimates for the lens and light profile model parameters for each of the lensed systems.

The lens modeling requires as input (1) image cutouts of the lensed arcs, (2) pixel noise maps of the image cutouts, and (3) a point-spread function (PSF) for each photometric band. We use DES image cutouts from the first three years of DES survey data to provide this data. The pixel noise maps encode the total uncertainty per pixel from sky photons as well as Poisson noise from the sources themselves. The DESDM image processing pipeline (summarized in Balbinot et al., 2015 and described in detail in Sevilla et al., 2011; Desai et al., 2012; Mohr et al., 2012) carries out image coaddition for each band and creates the noise maps. The default DESDM noise maps only include the noise from the sky photons and not the photons contributing to the objects in the image. We therefore modified and re-ran the image processing pipeline to generate a noise map that also includes shot noise from the objects. The DESDM pipeline also fits a model of the PSF using PSFEX (Bertin,

2011; Bertin & Arnouts, 1996). The PSF model is used to generate a model star at the location of the lens.

For each system, the light of the source and lensing galaxies are modeled as elliptical Sersic profiles with a Sersic index allowed to vary between 0.5 and 8. Mass models are then specified for each lensing galaxy, and the deflected source light in the image plane is calculated for the given lens model. The model is then convolved with the PSF to obtain the reconstructed image. The difference between the flux in the model image and the observed image is divided by the noise in each pixel to produce a residual image. The square of the residual in each pixel is summed to construct a χ^2 likelihood for that band. In this work, we combine the squares of the residuals in the g, r, and i bands to construct our likelihood. The χ^2 is then minimized using the affine-invariant ensemble sampler Markov chain Monte Carlo (MCMC) `emcee` (Foreman-Mackey et al., 2012) to sample the parameter space of the lens model.

For the lensed systems in this work, we used a family of elliptical power-law models for each galaxy, with reduced surface mass density

$$\kappa(x, y) = \frac{2 - \eta}{2} \left(\frac{\theta_{\rm Ein}}{\sqrt{q(x^2 + s^2) + y^2/q}} \right)^{\eta} \tag{3.1}$$

where (x, y) are image plane coordinates oriented such that x is aligned with the semi-major axis, q is the ratio of semiminor to semimajor axes, η is the power law index, s is the core radius and $\theta_{\rm Ein}$ is the Einstein radius. In this parametrization, a power law index of $\eta = 1$ corresponds to an isothermal mass profile, while larger values of η correspond to a mass profile with a steeper cusp. The Einstein radius $\theta_{\rm Ein}$ is defined as the effective radius (in units of angular distance) within which the average convergence $\langle \kappa \rangle = 1$, and characterizes the scale length of deflections on the source light by the lensing galaxy.

From the best-fit model parameters, we obtained estimates and associated uncertainties

of the physical Einstein radius, mass, luminosity, and mass-to-light ratio through Monte Carlo analysis.

3.3.1 Physical Einstein Radius

For each lensed system, we used the small angle approximation to estimate the physical Einstein radius, $R_{\rm Ein} = \theta_{\rm Ein} D_{\rm d}$, where $\theta_{\rm Ein}$ is the Einstein radius (in units of angular distance) from the best-fit mass model (Eq. (3.1)) and $D_{\rm d}$ is the angular diameter distance to the lens galaxy. To obtain $D_{\rm d}$, we assumed the concordance ΛCDM cosmology in §3.1 and evaluated the angular diameter distance at the photometric redshift of the lens galaxies (Table 3.1). We estimated the derived physical Einstein radius and its uncertainty by sampling the Einstein radii $\theta_{\rm Ein}$ from the posterior distribution of the best-fit lens model parameters and a distribution of $D_{\rm d}$, drawn from an assumed Gaussian distribution in the lens redshift $z_{\rm d}$ with standard deviation given by the errors in Table 3.1.

3.3.2 Mass of Lensing Galaxies

The masses of the lensing galaxies were estimated by integrating the surface mass density profile of the best-fit lens model over the physical area enclosed by the Einstein radius

$$\begin{aligned} M_{\rm enc} &= \iint \Sigma(\theta) dA \\ &= \iint \Sigma_c \kappa(x,y) dA \end{aligned} \tag{3.2}$$

where $\Sigma_{\rm c}$ is the critical surface density

$$\Sigma_{\rm c} = \frac{c^2}{4\pi G} \left(\frac{D_{\rm s}}{D_{\rm d} D_{\rm ds}} \right), \tag{3.3}$$

c is the speed of light, G is the gravitational constant and $D_{\rm d}$, $D_{\rm s}$ and $D_{\rm ds}$ are the angular diameter distances between the observer and the lens galaxy, between the observer and the source object, and between the lens galaxy and the source object respectively.

The error in the enclosed mass $M_{\rm enc}$ was estimated by sampling values of $M_{\rm enc}$ and calculating the 68th percentile errors from the sample. Each value of $M_{\rm enc}$ is calculated as follows: for the relevant quantities in the reduced mass density profile $\kappa(x, y)$, values were drawn from the posterior distribution of the Monte Carlo chain; values of angular diameter distance quantities $D_{\rm d}$, $D_{\rm s}$ and $D_{\rm ds}$ were sampled from a gaussian distribution of lens and source redshifts, $z_{,\rm d}$ and $z_{\rm d}$, with mean and standard deviations in Table 3.1.

3.3.3 Luminosity of Lensing Galaxies

To obtain the luminosity of the lens galaxy, we used the photometric redshifts of the lensing galaxy to obtain an estimate of the luminosity distance D_L to that galaxy. For a given luminosity distance, we then calculated the appropriate k-correction (Hogg et al., 2002) to that galaxy in each DES photometric band using the software package `kcorrect` (Blanton & Roweis, 2006) given the DES throughput curves[3], the photometric magnitudes of the lensing galaxy in the g, r, i, z and Y bands (Table 3.1) and their photometric redshift estimates. A table of K-correction values calculated and used in this work is presented in §3.A. We also correct for Milky Way extinction using the values in the DESDM catalog (§3.A) which are calculated using the dust map of Schlegel et al. (1997). We then obtain the absolute magnitude M of the galaxy using the relation:

$$M = m - 5\log_{10}\left(\frac{D_L}{10\ \rm pc}\right) - K - E \tag{3.4}$$

where m and K are the apparent magnitude and k-correction respectively for a given photometric band, and E is the extinction due to dust. The absolute magnitude of the lensing

3. Courtesy of Douglas Tucker and William Wester

galaxy in that given band is converted into a luminosity in units of solar luminosities. We used the SDSS g-band AB absolute magnitude of $M_{\odot} = 5.12 \pm 0.02$ (Blanton & Roweis, 2006) and the SDSS-to-DES transformation equations[4] (see §3.B) to obtain the absolute magnitude of the Sun in the DES photometric bands. From there, the luminosity of the lens in each photometric band can be calculated as:

$$L_{\rm lens}/L_{\odot} = 10^{-0.4(M_{\rm lens}-M_{\odot})} \tag{3.5}$$

Finally, using Eq. (3.2) and Eq. (3.5), we computed the mass-to-light ratio for each lensing galaxy. All luminosities and mass-to-light ratios are reported in the i-band. Errors in the i-band luminosities were estimated through Monte Carlo sampling in a fashion similar to the physical Einstein radius and enclosed mass described above. The relevant errors that are accounted for and propagated in this Monte Carlo estimate come from uncertainties in (1) DES i-band magnitudes of the relevant lens galaxies (given in Table 3.1), (2) Photometric redshifts of the lens galaxy, (3) SDSS g-band AB absolute magnitude of the sun, (4) SDSS $g-r$, $r-i$, and $i-z$ color conversion for the sun (5) SDSS-to-DES i-band transformation equation. All distributions of the above quantities in the Monte Carlo sampling were assumed to be Gaussian.

Finally, the i-band mass-to-light ratios were calculated and reported, with estimated uncertainties obtained from sampling the distribution of the enclosed masses and i-band luminosities of the lens galaxies from the prior Monte Carlo analyses.

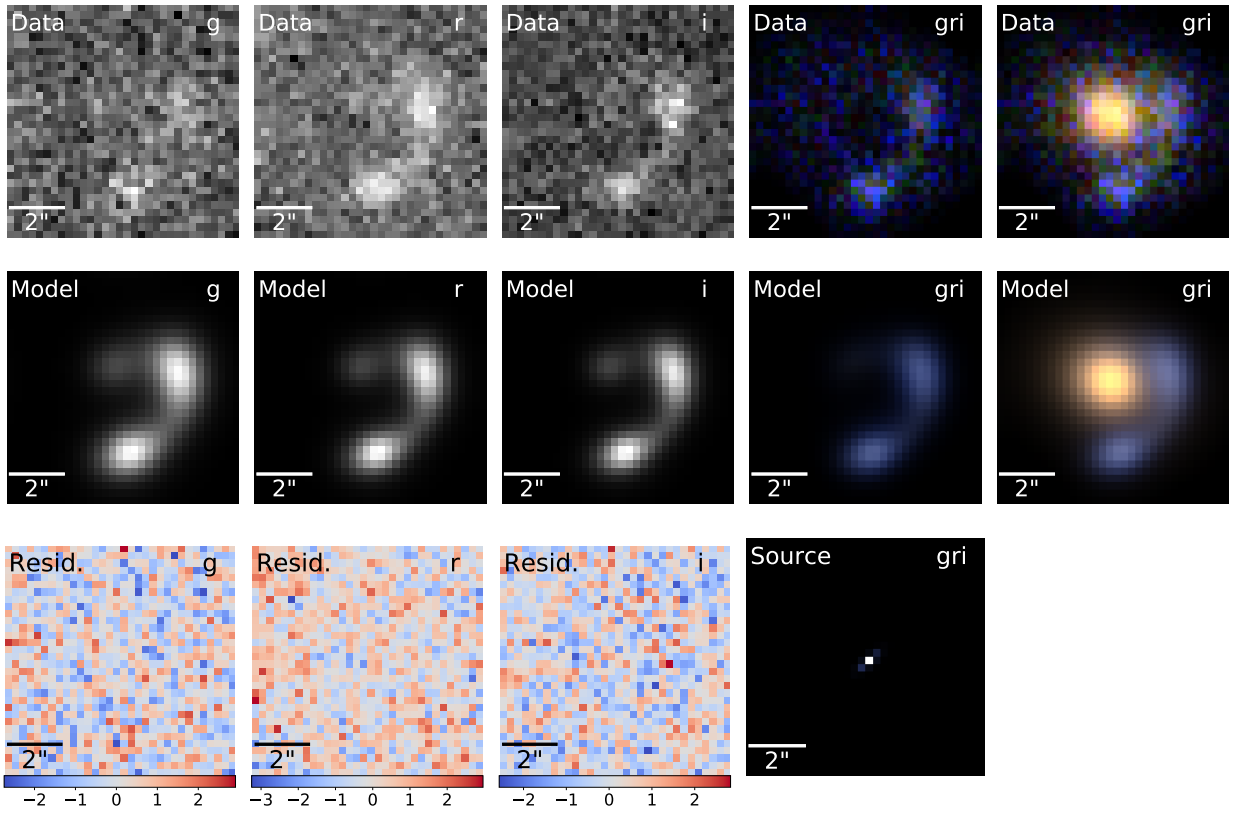

Figure 3.2 Multi-band lens-subtracted images, best-fit model, and residuals of DESJ0104-5341 in the g,r, and i bands, along with gri color composite images of the system with and without lens-subtraction. The first three columns show the lens-subtracted DES single-epoch observed image (top), best-fit model (middle), and residuals (bottom) in the DES g,r, and i bands respectively. The two rightmost columns show gri composite color images and best-fit model, with the rightmost column including the lens galaxy light, which is modeled with a single component Sersic light profile. The bottom rightmost figure shows the position of the source light profile, also modeled with a Sersic light profile.

Table 3.3 Inferred best-fit lens model parameters for lenses in this work. for a non-singular isothermal ellipsoid with external shear (NSIE+XS), described by Einstein radius $\theta_{\rm Ein}$, core radius $\theta_{\rm c}$, axis ratio (ratio of semiminor to semimajor axes) q, position angle $\theta_{\rm q}$, shear magnitude γ and associated position angle θ_γ. The position of the center of mass of this lens model is fixed to the center of light of the lensing galaxy. The position angles are measured north of west. Reported best-fit parameters are median values, with errors corresponding to 16th and 84th percentiles.

System	Model	Lens	$\theta_{\rm Ein}(''$)	$\theta_{\rm c}(''$)	q	$\theta_{\rm q}$ (°)	η	γ	θ_γ (°)
DESJ0104-5341	NSIE+XS	G	$2.8^{+0.5}_{-0.4}$	1.0 ± 0.4	0.64 ± 0.11	158 ± 5	-	$0.027^{+0.053}_{-0.019}$	164^{+49}_{-32}
DESJ0120-5143	1EPL+XS	G	2.853 ± 0.002	-	$0.718^{+0.005}_{-0.003}$	122.45 ± 0.08	$0.649^{+0.001}_{-0.005}$	$0.1143^{+0.0014}_{-0.003}$	$136.89^{+0.12}_{-0.5}$
	2EPL+XS	G	2.701 ± 0.005	-	$0.674^{+0.008}_{-0.010}$	130.6 ± 0.7	0.909 ± 0.007	0.179 ± 0.004	$133.3^{+0.4}_{-0.2}$
		P	0.27 ± 0.04	-	0.679 ± 0.022	160.1 ± 2.0	1.112 ± 0.015	-	-
DESJ2321-4630	2SIE+XS	G1	$2.029^{+0.006}_{-0.008}$	-	0.91 ± 0.02	$140.8^{+0.3}_{-0.5}$	-	0.058 ± 0.003	168.4 ± 0.4
		G2	$0.974^{+0.006}_{-0.007}$	-	0.59 ± 0.02	139.5 ± 0.4			
DESJ2349-5113	1EPL+XS	G	4.435 ± 0.002	0.151 ± 0.003	0.620 ± 0.007	171.6 ± 0.3	$0.802^{+0.004}_{-0.002}$	0.061 ± 0.003	$77.06^{+1.0}_{-0.8}$

3.4 Lens Modeling Results

3.4.1 DESJ0104-5341

A single component Sersic light profile was found to be a good fit for the lens galaxy, with only slight residuals in the i band near the center of the lens galaxy. The very close positions of the lensed source image to the lensing galaxy make it difficult to discern the position of a counter-image in the lens-subtracted image, as the light from the counter-image may be blended with the light from the lens galaxy due to the PSF.

The best-fit mass model was obtained using a non-singular (cored) isothermal ellipsoid mass profile (NSIE) (Kormann et al., 1994) with external shear. The lens model is described by the following parameters: two components for the ellipticity and its orientation, $(q, \theta_{\rm q})$, one component for the core radius $\theta_{\rm c}$ and one component for the Einstein radius, $\theta_{\rm Ein}$. An additional two components, (γ, θ_γ), describe the external shear and its orientation angle. Due to uncertainty in the position of a counter-image, we impose the additional constraint that the position of the center of the NSIE mass profile is fixed to the position of the center

4. Courtesy of Douglas Tucker and Sahar Allam

of the Sersic light profile of the lens galaxy G in our MCMC.

The best-fit parameters are listed in Table 3.3, and Fig. 3.2 shows the data, best-fit model image, and residuals. In this model, a faint counter-image is positioned northwest of the lens galaxy and extremely close to the lens, resulting in the light from the counter-image blending with the lens light. In the best-fit model, the Einstein radius of the lens model is $\theta_{\rm Ein} = 2.8^{+0.5}_{-0.4}{}''$, with core radius $\theta_{\rm c} = 1.0 \pm 0.4''$. The axis ratio and position angle of the NSIE mass profile $(q, \theta_{\rm q}) = (0.64 \pm 0.11\ ,\ 158 \pm 5^\circ)$ is consistent with the axis ratio and position angle of the Sersic light profile of the lens galaxy, $(q_{\rm s}, \theta_{\rm q,s}) = (0.743 \pm 0.025,\ 154 \pm 3^\circ)$.

A slight external shear of $(\gamma, \theta_\gamma) = (0.027^{+0.053}_{-0.019},\ 164^{+49}_{-32}{}^\circ)$ was preferred by the best-fit model. The orientation angle of 164^{+49}_{-32}counterclockwise from west (north of west) is consistent with the direction of a group of nearby galaxies in the lens environment to the south and west (labeled 'E' in Fig. 3.1(a)).

The photometric redshift of the lens galaxy, $z_{\rm d} = 0.679 \pm 0.022$ corresponds to an angular diameter distance of $D_A = 1456^{+19}_{-20}$ Mpc. The estimated physical Einstein radius is therefore $R_{\rm Ein} = 20 \pm 3$ kpc, and the enclosed mass given by Eq. (3.2) and the Einstein radius of the best-fit model is $M_{\rm enc} = (2.9^{+0.9}_{-0.7}) \times 10^{12}\ M_\odot$.

The luminosity of the lens in the i band is $L_{\rm lens} = (9^{+4}_{-3}) \times 10^{11}\ L_\odot$. This corresponds to a mass-to-light ratio of $M/L = 4.0^{+1.9}_{-1.3}\ M_\odot/L_\odot$in the i band.

3.4.2 DESJ0120-5143

Sersic light profiles were fitted to both the central red galaxy G and the satellite galaxy P. The quadruple lensed images after lens subtraction were found to be fairly well-described by an elliptical power-law (EPL) model (e.g. Tessore & Metcalf (2015); Barkana (1998)) with external shear. The elliptical power-law model is described by 6 parameters: two for position (x, y), two for ellipticity $(q, \theta_{\rm q})$, the Einstein radius $\theta_{\rm Ein}$ and power-law index η. In this parametrization, the power-law density profile slope is defined so that an isothermal

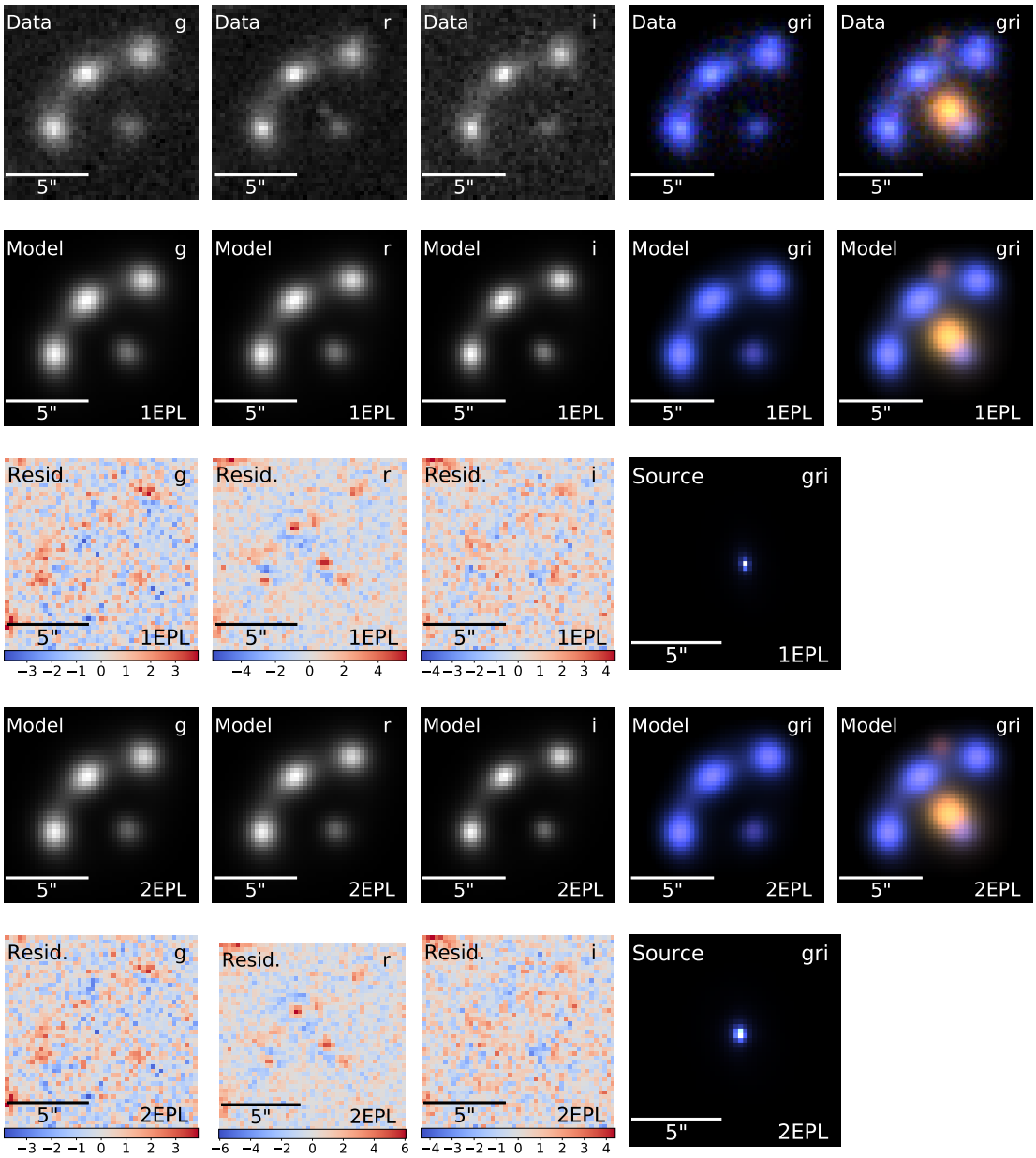

Figure 3.3 Multi-band lens-subtracted images, best-fit model, and residuals of DESJ0120-5143 in the g,r, and i bands, along with gri color composite images of the system with and without lens-subtraction. The top row shows the lens-subtracted DES single-epoch observed images in the g,r, and i bands (columns 1-3), as well as gri color composite images with (column 4) and without (column 5) lens subtraction. The second and third rows from the top show the best-fit model and residuals respectively for a 1-component elliptical power law lens model (1EPL), while the fourth and fifth rows from the top show the best-fit model and residuals for a 2-component elliptical power law lens model (2EPL). The bottom rightmost figures in the third and fifth rows show the position of the source light profile, also modeled with a Sersic light profile.

profile has a power-law index of $\eta = 1$, with higher values corresponding to steeper cusps [5]. An additional two components, (γ, θ_γ), describe the external shear and its orientation angle.

Given the uncertainty of whether P is in the foreground or background, we chose to model the system with and without accounting for the presence of an additional mass profile centered on the coordinate of P. We found that while a one-component EPL mass model profile centered on the central galaxy G, with external shear (1EPL+XS), captures the main lensing features, a two-component EPL with external shear model (2EPL+XS), with a second mass centered at the perturbing galaxy, P, results in better residuals.

The relevant best-fit parameters for both models are summarized in Table 3.3, and the observed and model images, along with residuals, are shown in Fig. 3.3. Unlike DESJ0104-5341, the position of the center of the mass profile is not fixed to the center of the light profile. However, the best-fit mass model is found to trace the light profile of the central lens galaxy, with the best-fit position of the center of the mass profile matching up within $0.1''$ to the position of the center of the light profile. The mass model axis ratio and position angle $(q, \theta_q) = (\ 0.718^{+0.005}_{-0.003}\ ,\ 122.45 \pm 0.08\ ^\circ)$ is less elliptical than the best-fit Sersic light profile of the central lens galaxy $(q_s, \theta_{\rm q,s}) = (0.783 \pm 0.014,\ 130.28^{+0.07}_{-0.06}{}^\circ)$, but oriented in approximately the same direction.

A significant external shear of $\gamma = 0.1143^{+0.0014}_{-0.003}$ with orientation angle $\theta_\gamma = 136.89^{+0.12}_{-0.5}$ $^\circ$ was required in this best-fit model. This distortion is consistent with the presence of multiple galaxies in the lens environment aligned along this axis (e.g. 'E' in Fig. 3.1(b)). The photometric redshift of the central lens galaxy, $z_{\rm d} = 0.603 \pm 0.063$ corresponds to an angular diameter distance of $D_{\rm d} = 1380^{+60}_{-70}$ Mpc. The estimated physical Einstein radius is therefore $R_{\rm Ein} = 19.1^{+0.8}_{-1.0}$ kpc, and the enclosed mass given by Eq. (3.2) and the Einstein radius of the best-fit model is $M_{\rm enc} = (3.1 \pm 0.5) \times 10^{12}\ M_\odot$. The luminosity of the lens in the i band is $L_{\rm lens} = (5.4^{+1.3}_{-1.1}) \times 10^{11}\ L_\odot$. This corresponds to a mass-to-light ratio of

5. See e.g. Eqn (1) Collett & Cunnington (2016) for the explicit form of the reduced surface mass density for the elliptical power-law profile.

$M/L = 5.9 \pm 0.4\ M_{\odot}/L_{\odot}$in the i band.

The addition of a second EPL mass component (2EPL+XS), centered at the position of the satellite galaxy, P, was found to result in reduced residuals and reduced χ^2 values. The parameters of this fit are listed in Table 3.3 and the model and residuals are presented in Fig. 3.3. In this lens model, the axis ratio of the central lens galaxy best-fit mass model is decreased from the 1EPL model from $q = 0.718^{+0.005}_{-0.003}$ to $0.674^{+0.008}_{-0.010}$, which matches the axis ratio of the best-fit Sersic light profile of the lens galaxy within the 68th percentile errors. However, a higher external shear amplitude of $\gamma = 0.179 \pm 0.004$ is required in this model. The best-fit Einstein radius is also decreased in this model, resulting in lower estimates of the physical Einstein radius, mass enclosed, and mass-to-light ratio of the central lens galaxy.

3.4.3 DESJ2321-4630

The light profiles of the two lens galaxies and source light were fitted with Sersic light profiles. Singular isothermal ellipsoid (SIE) mass models Kormann et al. (1994) were then used to model each lens galaxy, G1 and G2. Separate fits were done with and without external shear, and the best-fit models for each lens galaxy are shown in Table 3.3. The lens models are described by the following parameters: two components for the ellipticity and its orientation, $(q, \theta_{\rm q})$, and one component for the Einstein radius, $\theta_{\rm Ein}$. For the model with external shear, an additional two components, $(\gamma, \theta_{\gamma})$, describe the external shear and its orientation angle respectively.

We found that the addition of external shear did not significantly improve the fit. However, if external shear is allowed, there is a resulting trade-off between the shear amplitude and the axis ratios of the best-fit lens model. In the model where no external shear was added, the best-fit model axes ratios for G1 and G2 are $q_{\rm G1} = 0.91 \pm 0.02$ and $q_{\rm G2} = 0.59 \pm 0.02$ respectively. However, with the addition of an external shear component of amplitude and orientation angle $(\gamma, \theta_{\gamma}) = (\ 0.058 \pm 0.003\ ,\ 168.4 \pm 0.4\ ,$ the axis ratios of the best-fit SIE

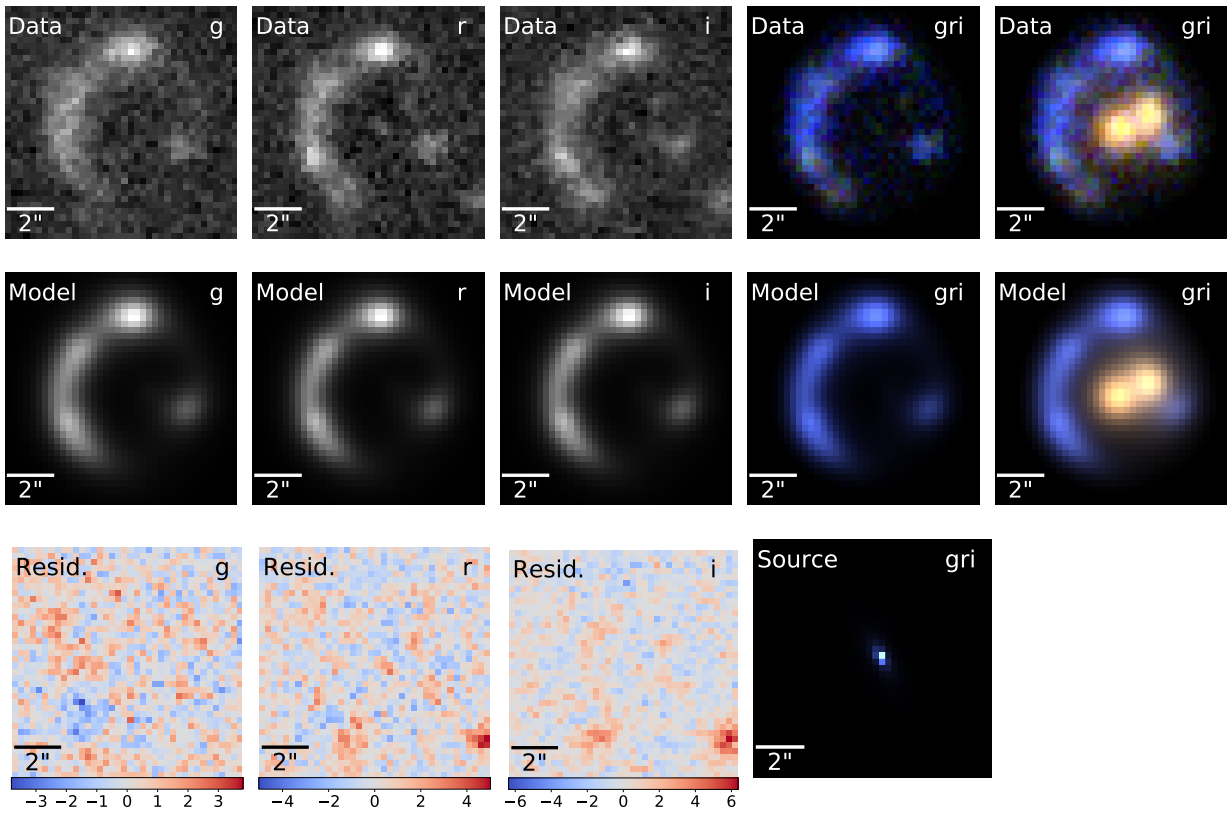

Figure 3.4 Multi-band lens-subtracted images, best-fit model, and residuals of DESJ2321-4630 in the g,r, and i bands, along with gri color composite images of the system with and without lens-subtraction. The top row shows the lens-subtracted DES single-epoch observed images in the g,r, and i bands (columns 1-3), as well as gri color composite images with (column 4) and without (column 5) lens subtraction. The second and third rows from the top show the best-fit model and residuals respectively for a two-component singular isothermal ellipsoid (2SIE), while the fourth and fifth rows from the top show the best-fit model and residuals for a 2-component elliptical power law lens model with external shear (2EPL+XS). The bottom rightmost figure in the third and fifth rows shows the position of the source light profile, also modeled with a Sersic light profile, while the bottom rightmost image shows the same source light magnified.

models are increased to $q_{\rm G1} = 0.91 \pm 0.02$ and $q_{\rm G2} = 0.59 \pm 0.02$ respectively. The orientation angle of the external shear is θ_γ oriented towards the direction of a group of nearby galaxies (e.g. 'E' in Fig. 3.1(c)).

From this model, we calculate the physical radius, enclosed mass, and mass-to-light ratio for both G1 and G2. The photometric redshifts of the lens galaxy G1 and G2 are $z_{\rm d} = 0.70 \pm 0.02$ and 0.64 ± 0.03 respectively, which corresponds to angular diameter distances of $D_d = 1474^{+16}_{-17}$ Mpc and 1420 ± 0.3 Mpc.

In the model where no external shear was added (2SIE), the estimated physical Einstein radii for G1 and G2 are $R_{\rm Ein}$ =14.5 ± 0.2 kpc and $6.70^{+0.14}_{-0.15}$ kpc respectively, and the enclosed mass given by Eq. (3.2) and the Einstein radius of the best-fit model is $M_{\rm enc} = (1.56 \pm 0.06) \times 10^{12}\ M_\odot$ for G1 and $(3.2 \pm 0.2) \times 10^{11}\ M_\odot$ for G2. The combined mass of the two lenses is $(1.86 \pm 0.1) \times 10^{12}\ M_\odot$.

When external shear is added (2SIE+XS), the best-fit physical Einstein radii for G1 and G2 are slightly modified to $R_{\rm Ein}$ =14.5 ± 0.2 kpc and $6.70^{+0.14}_{-0.15}$ kpc, within the 68th percentile errors of the 2SIE model. The corresponding enclosed mass is $M_{\rm enc} = (1.56 \pm 0.06) \times 10^{12}$ $M_\odot$ for G1 and $(3.2 \pm 0.2) \times 10^{11}\ M_\odot$ for G2, giving a combined mass of $(1.83 \pm 0.1) \times 10^{12}$ $M_\odot$.

The luminosities, $L_{\rm lens}$ of the lens in the i band for G1 for G2, are $(5.3 \pm 0.4) \times 10^{11}$ $L_\odot$ and $(2.71^{+0.20}_{-0.25}) \times 10^{11} L_\odot$ respectively. For the 2SIE model, this corresponds to a i-band mass-to-light ratio of $M/L = 2.95^{+0.15}_{-0.13}\ M_\odot/L_\odot$ for G1 and $1.18 \pm 0.06\ M_\odot/L_\odot$ for G2. When shear is added, the mass-to-light ratio for G1 is $M/L = 2.95^{+0.15}_{-0.13}\ M_\odot/L_\odot$, and $1.18 \pm 0.06\ M_\odot/L_\odot$ for G2.

3.4.4 DESJ2349-5113

This system has has main galaxy G and a smaller galaxy P in its environment. We used single-component Sersic light profiles to model both galaxies, which resulted in good fits, with

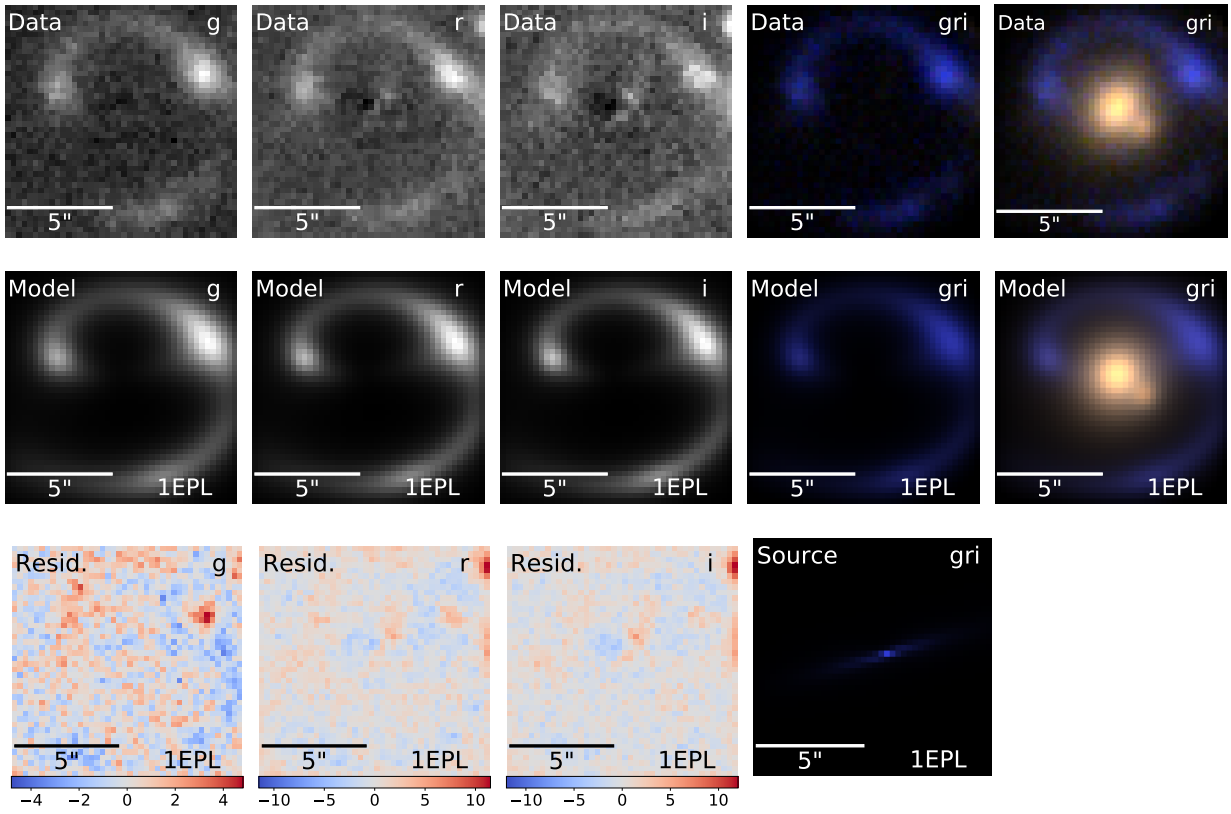

Figure 3.5 Multi-band lens-subtracted images, best-fit model, and residuals of DESJ2349-5113 in the g,r, and i bands, along with gri color composite images of the system with and without lens-subtraction. The first three columns show the lens-subtracted DES single-epoch observed image (top), best-fit model (middle), and residuals (bottom) in the DES g,r, and i bands respectively. The two rightmost columns show gri composite color images and best-fit model, with the rightmost column including the lens galaxy light, which is modeled with a single component Sersic light profile. The bottom rightmost image shows the source light, also modeled with a Sersic light profile.

only slight residuals in the i band near the center of the lens galaxy. The lensed source takes on an extended ring-like structure, with visible features that extend around the majority of the lens.

We found that a single-cored elliptical power law (cEPL) model was able to reproduce the observed image the best. We also tried using two separate elliptical models to model both the mass of the central galaxy G and the smaller galaxy P but found that it did not result in a better fit. The elliptical power-law model is described by 6 parameters: two for position (x, y), two for ellipticity $(q, \theta_{\rm q})$, the Einstein radius $\theta_{\rm Ein}$, core radius $\theta_{\rm c}$, and power-law index η. In this parametrization, the power-law density profile slope is defined so that an isothermal profile has a power-law index of $\eta = 1$, with higher values corresponding to steeper cusps [6]. An additional two components, (γ, θ_γ), describe the external shear and its orientation angle.

The best-fit parameters are listed in Table 3.3, and Fig. 3.5 shows the data, best-fit model image, and residuals. In the best-fit model, the Einstein radius of the lens model is $\theta_{\rm Ein} = 4.435 \pm 0.002$ $''$, with core radius $\theta_{\rm c}$ =0.151 ± 0.003 $''$. The axis ratio and position angle of the NSIE mass profile $(q, \theta_{\rm q}) = (0.620 \pm 0.007$, 171.6 ± 0.3 $^\circ)$ is inconsistent with the axis ratio and position angle of the Sersic light profile of the lens galaxy G, $(q_{\rm s}, \theta_{\rm q,s}) = (0.916 \pm 0.007$, 42.11 ± 0.02 $^\circ)$. A slight external shear of $(\gamma, \theta_\gamma) = (0.061 \pm 0.003$, $77.06^{+1.0}_{-0.8}$ $^\circ)$ was preferred by the best-fit model.

The photometric redshift of the lens galaxy, $z_{\rm d} = 0.345 \pm 0.032$ corresponds to an angular diameter distance of $D_A = 1010 \pm 60$ Mpc. The estimated physical Einstein radius is therefore $R_{\rm Ein}$ =$21.7^{+1.2}_{-1.4}$ kpc, and the enclosed mass given by Eq. (3.2) and the Einstein radius of the best-fit model is $M_{\rm enc} = (3.53 \pm 0.4) \times 10^{12}$ $M_\odot$. The luminosity of the lens in the i band is $L_{\rm lens} = (5.6 \pm 1.1) \times 10^{11}$ $L_\odot$. This corresponds to a mass-to-light ratio of $M/L = 6.3^{+.7}_{-0.5}$ $M_\odot/L_\odot$in the i band.

6. See e.g. Eqn (1) Collett & Cunnington (2016) for the explicit form of the reduced surface mass density for the elliptical power-law profile.

3.5 Discussion and Summary

In this paper, we presented mass models for four galaxy-scale strong lensing systems, DESJ0104-5241, DESJ0120-5143, DESJ2321-4630, and DESJ2349-5113, discovered in the DES Year 1 data (Diehl et al., 2017; Nord et al., 2020). These systems were identified through visual identification and catalog searches and confirmed through spectroscopy from GMOS on the Gemini South telescope.

DES expects to find approximately 2000 such galaxy-galaxy strong lensing systems in the final five-year dataset. A variety of photometry-based searches have been used to identify lens candidates in the DES data thus far (e.g. Diehl et al. (2017)). Semi-automated arc-finding (e.g. More et al. (2012b)) and ring-finding algorithms (e.g. Gavazzi et al. (2014a)) are also being employed in the DES lens search. However, the sheer amount of survey imaging data makes both lens searches through conventional methods of visual inspection and lens modeling through conventional techniques extremely time-intensive. This necessitates the development of both more automated techniques as well as large-scale crowdsourcing efforts to both discover and model strong lensing systems.

In terms of automated approaches to lens classification and modeling, significant efforts to develop automated lens classification pipelines using machine learning approaches with accuracy comparable to human identification have been made (e.g. Avestruz et al. (2019); Lanusse et al. (2017)). These machine learning approaches generally operate by extracting characteristic lensing features from lensed images and then using pattern recognition algorithms to classify images based on the presence (or absence) of these lensing features as to whether they contain lenses or not. These algorithms have been applied to DES data to look for high-redshift lenses with some success (Jacobs et al., 2019b). The development of automated lens modeling pipelines are also underway (e.g. Nightingale et al. (2018)).

Infrastructure for crowdsourcing approaches to the classification of lenses through visual inspection has also emerged as a viable scalable approach to lens classification. The success

of the *Space Warps* program (Marshall et al., 2015; More et al., 2015), a citizen science approach to the visual identification of strong lensing systems in Canada-France-Hawaii Telescope Science Legacy Survey (CFHTLS) survey data, in identifying new lens candidates is a testament to the viability of crowd-sourcing approaches. The development of *Space Warps* for DES survey data is also underway. On the lens modeling front, there is also a significant effort to develop a crowd-sourcing infrastructure for lens modeling (e.g. Li et al (2017); Küng et al. (2015)). The ongoing development and employment of new tools to discover and analyze lens candidates will likely further increase the scientific yield from strong lensing systems in the DES dataset in the future.

3.A K-correction and Extinction Magnitudes

Table 3.A.1 K-correction values for DES Year 1 Lensing Galaxies

System ID	Object ID	g	r	i	z	Y
DESJ0104-5341	G	2.55±0.05	1.88±0.04	1.25±0.01	1.06±0.02	0.99±0.02
DESJ0120-5143	G	1.79±0.05	0.76±0.05	0.22±0.07	0.12±0.05	0.11±0.04
DESJ2321-4630	G1	1.75±0.03	1.31±0.02	0.82±0.02	0.79±0.02	0.78±0.01
	G2	1.55±0.04	0.95±0.04	0.55±0.02	0.55±0.03	0.56±0.03
DESJ2349-5113	G	1.30±0.09	0.44±0.02	0.23±0.01	0.17±0.02	0.19±0.01

Notes. K-correction values for each lensing galaxy in the DES g, r, i, z, and Y photometric bands. Values and associated errors are generated by sampling magnitudes from the inferred lens model parameter posterior chains and redshifts, calculated with `kcorrect` (Blanton & Roweis, 2006). The Object ID refers to lensing galaxies (e.g., G1) discussed in the text and in Fig. 3.1. Refer to §3.3 for more details.

Table 3.A.2 Extinction values for DES Year 1 Lensing Galaxies

System ID	Object ID	g	r	i	z	Y
DESJ0104-5341	G	0.064062	0.042856	0.031548	0.024064	0.021086
DESJ0120-5143	G	0.037412	0.025028	0.018424	0.014053	0.012314
DESJ2321-5113	G1	0.026204	0.01753	0.012905	0.009843	0.008625
	G2	0.026141	0.017487	0.012873	0.009819	0.008604
DESJ2349-4630	G	0.051485	0.034442	0.025355	0.019339	0.016946

Notes. Extinction for each lensing galaxy in the DES g, r, i, z, and Y photometric bands, drawn from the DESDM database using the SFD98 dust model (Schlegel et al., 1997).The Object ID refers to lensing galaxies (e.g., G1) discussed in the text and in Fig. 3.1. Refer to §3.3 for more details.

3.B SDSS-to-DES transformation equations

The transformation equations to convert SDSS DR13 magnitudes to DES Y3A1 magnitudes are based on SDSS DR13 and DES Y3A1 single-epoch data. The zero-point (the constant term) in each relation was derived by comparing the observed SDSS DR13 vs. DES Y3A1 relation with its Pickles synthetic counterpart (Pickles, 1998), and then manually refining the zero-point to match the calibration of the Y3A1 FGCM standard stars (v2.5) (Burke et al., 2017). The associated R.M.S. errors are also shown on the right.

$$g_{\rm DES} = g_{\rm sdss} + 0.001 - 0.075 \times (g-r)_{\rm sdss} \qquad (\pm 0.021)$$

$$r_{\rm DES} = r_{\rm sdss} - 0.009 - 0.069 \times (g-r)_{\rm sdss} \qquad (\pm 0.021)$$

$$i_{\rm DES} = i_{\rm sdss} + 0.014 - 0.214 \times (i-z)_{\rm sdss} - 0.096 \times (i-z)^2_{\rm sdss13} \ (\pm 0.021) \qquad (\pm 0.023)$$

$$z_{\rm DES} = z_{\rm sdss} + 0.022 - 0.068 \times (i-z)_{\rm sdss} \qquad (\pm 0.025)$$

$$Y_{\rm DES} = z_{\rm sdss} + 0.045 - 0.306 \times (i-z)_{\rm sdss} \qquad (\pm 0.030)$$

The relevant SDSS color conversion factors and errors for the Sun, using transformation equations from Bilir et al. (2005) and Rodgers et al. (2006), are:

$$(g-r)_{\rm sdss} = +0.44 \qquad (\pm 0.02)$$

$$(r-i)_{\rm sdss} = +0.11 \qquad (\pm 0.02)$$

$$(i-z)_{\rm sdss} = +0.03 \qquad (\pm 0.02)$$

CHAPTER 4

DEEP LEARNING IN WIDE-FIELD SURVEYS: FAST ANALYSIS OF STRONG LENSES IN GROUND-BASED COSMIC EXPERIMENTS

4.1 Introduction

Strong gravitational lensing occurs when massive objects (e.g., galaxies and their dark-matter halos) deform spacetime, deflecting the light rays that originate at sources along the line of sight to the observer (e.g., Schneider et al., 2013; Petters et al., 2012; Mollerach & Roulet, 2002). The key signature of a strong lens is a magnified and multiply imaged or distorted image of the background source, which can only occur if the source is sufficiently closely aligned to the line of sight of the warped gravitational potential generated by the lensing mass. Strong lensing also depends on the angular diameter distances between the observer, lens, and source, which encloses information about the underlying cosmology.

Strong lensing systems are unique probes of many astrophysical and cosmological phenomena. The source light may be magnified up to a hundred times, as the light deflection due to strong lensing conserves the surface brightness while increasing the total angular size of the object. They act as "gravitational telescopes," enabling the study of distant source objects that would otherwise be too faint to observe, such as high redshift galaxies (e.g., Ebeling et al., 2018; Richard et al., 2011; Jones et al., 2010), dwarf galaxies (Marshall et al., 2007), star-forming galaxies (Stark et al., 2008), quasar accretion disks (Poindexter et al., 2008), and faint Lyman-alpha blobs (Caminha et al., 2015). Lensing systems can also be used as non-dynamical probes of the mass distribution of galaxies (e.g. Treu & Koopmans, 2002c; Treu & Koopmans, 2002a; Koopmans et al., 2006a), and galaxy clusters (e.g., Kovner, 1989; Abdelsalam et al., 1998; Natarajan et al., 2007; Zackrisson & Riehm, 2010; Carrasco et al., 2010; Coe et al., 2010), providing a key observational window on dark matter (see,

e.g., Meneghetti et al., 2004).

Strong lensing has also been used — alone or in combination with other probes — to derive cosmological constraints on the cosmic expansion history, dark energy, and dark matter (see, e.g., Jullo et al., 2010; Caminha et al., 2016; Bartelmann et al., 1998; Cooray, 1999; Golse et al., 2002; Treu & Koopmans, 2002a; Yamamoto et al., 2001; Meneghetti et al., 2004, 2005; Jullo et al., 2010; Magaña et al., 2015; Cao et al., 2015b; Caminha et al., 2016; Schwab et al., 2010; Enander & Mörtsell, 2013; Pizzuti et al., 2016). Precise and accurate time-delay distance measurements of multiply-imaged lensed QSO systems have been used to measure the expansion rate of the universe (Oguri, 2007; Suyu et al., 2010b). More recently, this technique has also been applied to multiply-imaged lensed supernovae (Kelly et al., 2015; Goobar et al., 2016). Strong lensing can also be used to constrain dark matter models (Vegetti et al., 2012; Hezaveh et al., 2014; Gilman et al., 2018; Rivero et al., 2018; Bayer et al., 2018), as well as detect dark-matter substructures along the line of sight (Despali et al., 2018; McCully et al., 2017b).

The broad range of applications has inspired many searches for strong lensing systems. Many of these searches have been carried out on high-quality space-based data from the Hubble Space Telescope (HST): Hubble Deep Field (HDF; Hogg et al., 1996), the HST Medium Deep Survey (Ratnatunga et al., 1999), the Great Observatories Origins Deep Survey (GOODS; Fassnacht et al., 2004), the HST Archive Galaxy-scale Gravitational Lens Survey (HAGGLeS; Marshall, 2009), the Extended Groth Strip (EGS; Marshall et al., 2009), and the HST Cosmic Evolution survey (COSMOS; Faure et al., 2008; Jackson, 2008).

However, there is also a plethora of ground-based imaging data that merits exploration. The majority of confirmed strong lensing systems that have been discovered to date were first discovered in ground-based surveys, such as the Red-Sequence Cluster Survey (RCS; Gladders et al., 2003; Bayliss, 2012), the Sloan Digital Sky Survey (SDSS; Estrada et al., 2007; Belokurov et al., 2009; Kubo et al., 2010; Wen et al., 2011; Bayliss, 2012), the Deep Lens

Survey (DLS; Kubo & Dell'Antonio, 2008), The Canada-France-Hawaii Telescope (CFHT) Legacy Survey (CFHTLS; Cabanac et al., 2007; More et al., 2012a; Maturi et al., 2014; Gavazzi et al., 2014b; More et al., 2016; Paraficz et al., 2016), the Dark Energy Survey (DES; e.g., Nord et al., 2015), the Kilo Degree Survey (KIDS; e.g., Petrillo et al., 2017). Strong lenses have also been discovered in follow-up observations of galaxy clusters (e.g., Luppino et al., 1999; Zaritsky & Gonzalez, 2003; Hennawi et al., 2008; Kausch et al., 2010; Furlanetto et al., 2013) and galaxies (e.g., Willis et al., 2006). Next-generation surveys, like LSST (Ivezić et al., 2008), Euclid (Laureijs et al., 2011), and WFIRST (Green et al., 2012), are projected to discover up to two orders of magnitude more lenses than what is currently known (Collett, 2015a).

Many of the current catalogs of strong lensing systems were found through visual searches. However, the increasingly large data sets projected from current and future wide-field surveys necessitate the development and deployment of automated search methods to find and classify lens candidates.

Neural networks are one class of automated techniques. Deep learning algorithms (Goodfellow et al., 2016; LeCun et al., 2015), and in particular, convolutional neural networks (CNNs; LeCun et al., 1998), are established as state of the art for many sectors in computer vision research (see, e.g. Russakovsky et al., 2015; Zhang et al., 2018; Lin et al., 2018; John et al., 2015; Peralta et al., 2018). In some cases, they have been shown to perform better than humans (e.g., He et al., 2015; Metcalf et al., 2018). Deep learning allows the development of models that can process complex and minimally processed (even raw) data from a wide variety of sources to extract relevant features, which can be effectively linked to other properties of interest. For example, in computer vision tasks, it has been successfully applied to facial recognition (Lu et al., 2017), speech detection and characterization (Vecchiotti et al., 2018; Abdel-Hamid et al., 2014), music classification (Choi et al., 2017), medical prognostics (Li et al., 2018b) and diagnostics (Hannun et al., 2019). Several recent

works have demonstrated that both traditional neural networks (Bom et al., 2017; Estrada et al., 2007) and deep neural networks (Petrillo et al., 2019b; Jacobs et al., 2019c; Petrillo et al., 2019a; Metcalf et al., 2018; Lanusse et al., 2018; Glazebrook et al., 2017; Cheng et al., 2019; Madireddy et al., 2019) can be used to identify morphological features in raw images that distinguish lenses from non-lenses, with minimal intervention from humans.

Deep learning models are not only useful for solving classification problems with discrete variables but are also used for regression problems with continuous variables (Lathuilière et al., 2018a,b; Belagiannis et al., 2015). In addition to finding catalogs of lenses, measuring the properties of lenses, like the Einstein Radius or the velocity dispersion of the lensing galaxy, typically requires follow-up observations as well as modeling. Conventionally, modeling is performed with computationally expensive maximum likelihood algorithms (e.g., Bradač et al., 2009; Diego et al., 2005; Coe et al., 2008; Oguri, 2010; Jullo et al., 2007; Metcalf & Petkova, 2014; Petkova et al., 2014), which can take up to weeks on CPUs and require manual input.

However, Hezaveh et al. (2017a) showed that deep learning techniques could also be used to produce fast automated analyses of strong lenses. The lens parameters in that work were inferred from high-quality HST simulations and images. Additional work from Pearson et al. (2019) made use of traditional CNNs to model strong lenses in LSST-like and Euclid-like imaging data, and addresses the feasibility of recovering lens features from images. On the uncertainty quantification front, Levasseur et al. (2017) produced uncertainty estimates on strong lensing parameters using dropout techniques, which evaluate the deep neural network from a Bayesian perspective (Gal & Ghahramani, 2016). The same approach was used by Morningstar et al. (2018) to estimate the uncertainties on the inferred lens parameters of strong lenses in interferometric data.

In this work, we address the problem of strong lensing analysis in ground-based wide-field optical astronomical surveys, which have lower image quality than space-based data.

We also study the feasibility of modeling the uncertainties of lens parameters in a Bayesian deep-learning manner. To develop and validate our deep neural network model, we produced a simulated catalog of strong lensing systems with DES-quality imaging. These simulations are used to train and evaluate the deep neural network model. The model is then used to infer the velocity dispersion $\sigma_{\rm v}$, lens redshift $z_{\rm l}$, source redshift $z_{\rm s}$, and Einstein Radius $\theta_{\rm E}$ of the lens. Our approach is generic – although the model is optimized for galaxy-scale lens systems (i.e., two objects, often visually blended; distorted source images, ring-like images, or multiple images from the same source), it could also be extended and optimized to analyze other species of strong lenses, such as time-delay and double-source-plane lens systems.

This paper is organized as follows: First, in §4.2, we introduce the deep learning models and uncertainty estimation formalism used in this work. Then, in §4.3, we describe the simulated data used in this work. Following that, in §4.4, we describe how we trained the deep neural network model. In §4.5, we apply our model to a test set and evaluate its performance. Finally, we conclude and present an outlook for future work in §4.6.

4.2 Regression with Uncertainty Measurements in Deep Learning Models

In this work, we use an architecture based on the *inception* module (Szegedy et al., 2015, 2016) to perform the regression task. We use the *concrete dropout* technique (Gal et al., 2017) to approximate the posterior distribution of the parameter values and use the approximate posteriors to estimate uncertainties using confidence intervals. Both the model architecture and the error estimation techniques are described below.

4.2.1 Model Architecture: Inception

Large neural network architectures are challenging to train for several reasons. First, overfitting becomes a major limitation when the size of the training set does not scale with the number of parameters. Second, a uniform increase in the number of filters in convolutional layers requires a quadratic increase in computation. Third, when new layers are added to large linear architectures, many weights become close to zero after training, and thus most of the computation time is not fruitful (Szegedy et al., 2015).

The inception module (shown in Fig. 4.2.2) addresses these challenges by having a non-linear architecture that also introduces sparsity to the network. The sparsity adds a relevant advantage by making the neural network more adaptable and stable. Additionally, a wider layer increases cardinality (Tishby & Zaslavsky, 2015; Xie et al., 2017), i.e., the number of independent paths which can provide a new way of adjusting the model. With just thousands of parameters, the inception module has been shown to outperform traditional linear VGG models that have tens of millions of parameters. Images with features at different size scales present a challenge for networks that lack the flexibility to look for features at different sizes. Inception does not require a prescription of the optimal convolutional kernel size, because the convolutions are performed in parallel, each with a different kernel size. We present the full inception network used in this work in Fig. 4.2.1.

Starting with the original architecture (Szegedy et al., 2015), we replace the regular multi-class softmax activation function (Krizhevsky et al., 2012) after the last dense layer with an unbounded linear activation that can output a single continuous value — enabling the regression task. It has three streams: the *input*, the *core*, and the *output*. The *input* stream reduces dimensionality. It is composed of two consecutive blocks of 2D convolutional layers, a batch normalization layer, a ReLU activation, and 2D max pooling layers. Both convolutional layers have a kernel size of $(5, 5)$, while the pooling layers have a size of $(2, 2)$.

The core stream is composed of four consecutive Inception blocks. Each one of these

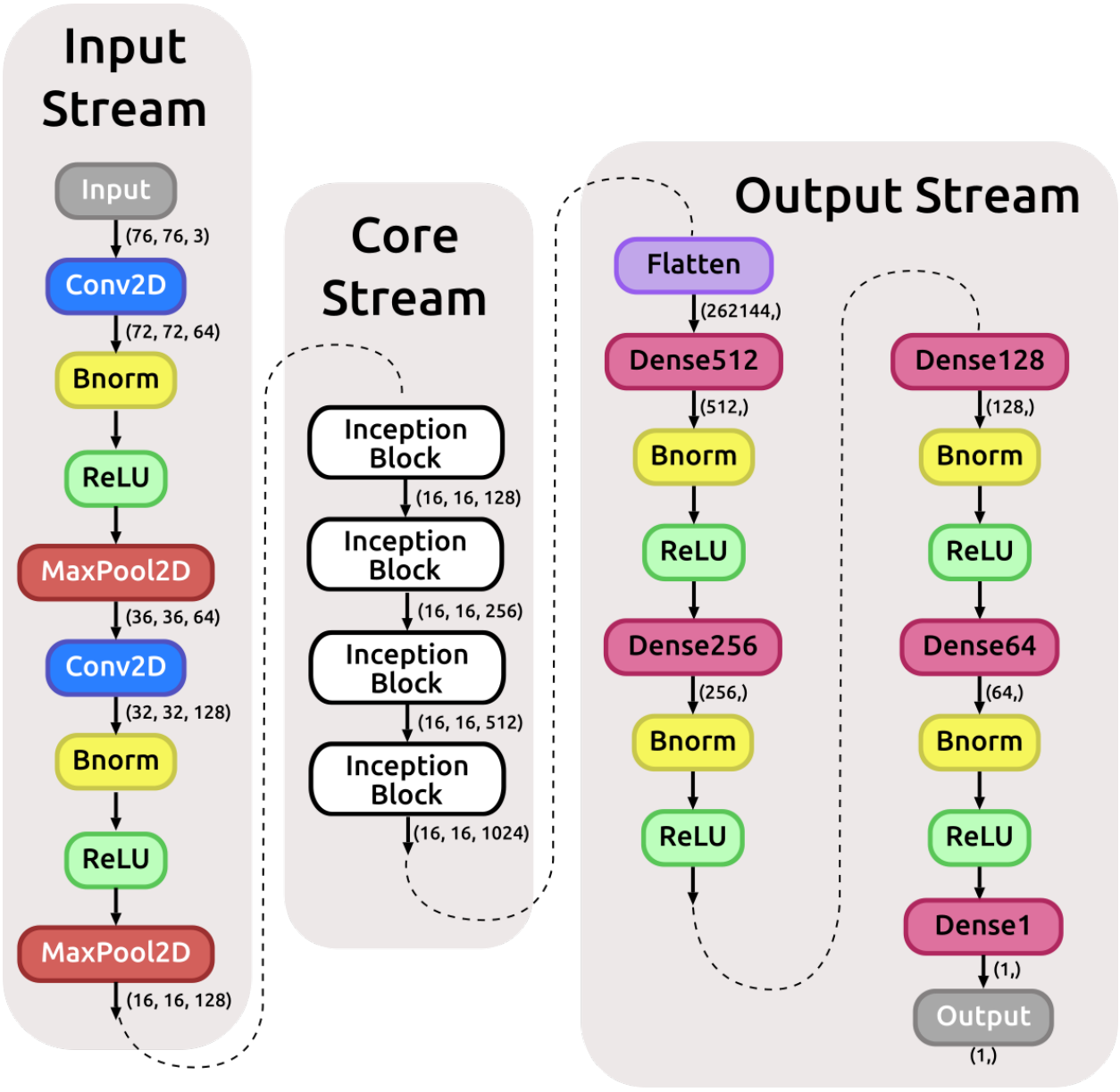

Figure 4.2.1 Diagram of the inception network used in this work. This architecture is composed of a total of 86 layers (convolutional, batch normalization, ReLU activations, max pooling, and dense). The size of the activation maps is shown at relevant locations of the net (after each layer that changes its size). The architecture is divided into three main streams: the input stream, which is used to reduce the dimensionality of input data and optimize training; the core stream, whose main purpose is to extract meaningful spatial features from the input data; and the output stream, which correlates these extracted features with the lensing property to be estimated, using a gradually decreasing feature hyperspace (512, 256, 128, 64 and finally 1).

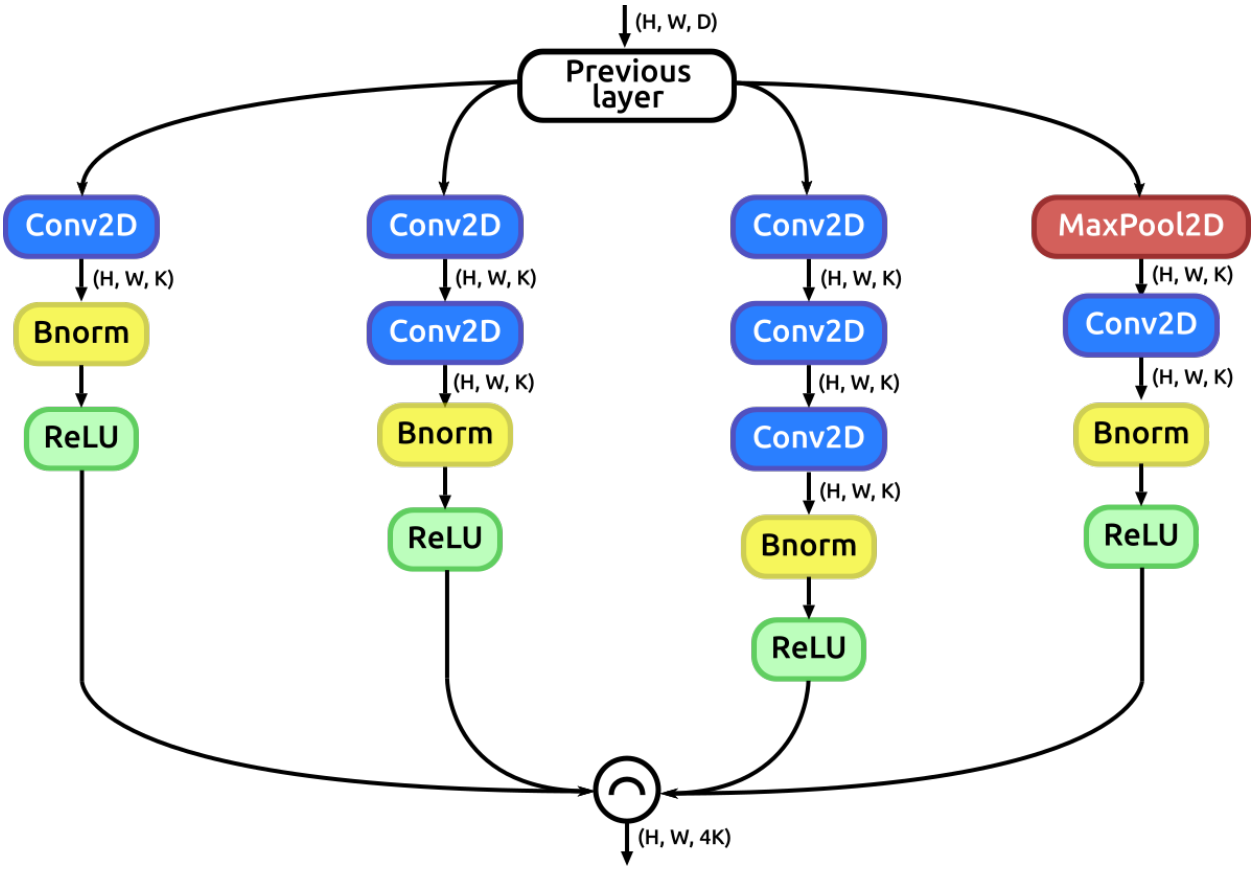

Figure 4.2.2 Diagram of the inception block used in the inception network, which is shown in Figure 4.2.1. This block is based on the inception module proposed by Szegedy et al. (2016) for the Inception-v2 architecture. This block allows the network to process data at different kernel sizes, which is useful for learning features at multiple size scales. The $(1, 1)$ convolutions (the first convolutional layer appearing on each branch) are used here to reduce image depth to make the training process less computationally expensive. The left branch only processes the input data using the $(1, 1)$ convolutional scheme. The second branch from the left has a $(1, 1)$ convolution followed by a $(3, 3)$ convolution. The third branch adds an extra $(3, 3)$ convolution to the previous schema, thus making it possible to capture image features at increasingly larger sizes. Finally, the last branch uses a 2D max pooling layer to achieve translation invariance. The activation maps of all branches are processed by batch normalization and a ReLU activation at their respective ends. Afterwards, the resulting maps are concatenated depth-wise. The dimension of the input data is (H, W, D), and K is the number of output features.

blocks, whose structure can be seen in Fig. 4.2.2, has four branches, each with a different kernel size. The $(1, 1)$ convolutions (the first convolutional layer appearing on each branch) are used for image depth reduction, which lowers computational cost. Any immediately following convolutional layer has a size of $(3, 3)$[1].

As shown in Fig. 4.2.1, apart from its depth, the size of the internal activation maps inside the core stream is not modified. Other types of architectures need to reduce activation map dimensions to extract features at different scales. However, inception modules keep the dimensions of activation maps constant, as the features at different scales are extracted using the parallel convolutional schema mentioned earlier. After the inception modules, one needs to map the relevant features into the predicted variable. Therefore, we implemented a bottleneck-structured sequence as proposed by He et al. (2016) — [Convolution (Conv) / BatchNorm (BN) / Rectified Linear Unit (ReLU)].

We tested the current architecture in two different schemes. One scheme is trained to predict a single parameter only, and the other predicts four parameters simultaneously. Before the output block, the weights are shared for each of the output parameters.

4.2.2 Error Analysis

Uncertainty estimates are critical for assessing confidence in scientific measurements. Nevertheless, cogent and interpretable methods for uncertainty estimation in deep learning remain elusive. Several types of uncertainty are useful in assessing scientific confidence. They may be broadly classified into two categories — *aleatoric* and *epistemic*. Aleatoric uncertainties encompass uncertainties that are due to the presence of stochastic errors in the input data itself. In an astrophysical context, this includes effects such as detector noise or the point spread function in astronomical seeing. These uncertainties do not decrease, for a given fixed

1. Stacking several $(3, 3)$ convolutional layers is equivalent to having a single convolutional layer with a larger kernel. For example, stacking two $(3, 3)$ filters is equivalent to using a single $(5, 5)$ kernel. However, this stacking is more computationally efficient than using single kernels (Szegedy et al., 2016).

model, when the number of data points collected in the data-acquisition process increases. Epistemic uncertainties, on the other hand, encapsulate uncertainties in the model due to a lack of knowledge about the best model and are usually caused by not having enough input data to train the model on. In the case of deep learning regression – which is a data-driven model – if one feeds the network with more data during the training process, this could lower the network's ignorance of which model is more likely to have generated the collected training data (Kendall & Gal, 2017b). Hence, epistemic uncertainties decrease when the training data size increases.

We name the total error – which includes the epistemic and aleatoric uncertainties – predictive error. Standard error propagation is currently untenable because even measurement errors in raw imaging data require some form of modeling (e.g., the transfer function of the imaging instrument). Additionally, there is no way to propagate that uncertainty, if it existed, through a deep learning model to the inferred parameters because we would need to know the uncertainties on the model parameters, and this is not well-known. Ideally, we would be able to perform uncertainty estimates of all parameters in a fully Bayesian framework. Bayesian neural networks may provide Gaussian process approximations of the variance in the output parameters (Lee et al., 2017); however, not all parameters may have Gaussian errors.

One Bayesian neural network method that has been used recently is *Concrete Dropout*, which was first described by Gal & Ghahramani (2016); Gal et al. (2017) and first used in strong lens modeling by Levasseur et al. (2017). In this method, to estimate uncertainties on the parameters of interest, we compute the probability distribution function (PDF) of the predictions using the concrete dropout technique. Concrete dropout approximates a posterior distribution $p\left(Y|X\right)$ of the predicted physical parameter Y given an input image x in a Bayesian framework. We interpret our model from the variational perspective (Jordan et al., 1999; Graves, 2011a). In this framework, dropout – which, as its name implies,

involves dropping out neural network weights randomly from a trained neural network – samples from the distribution $p(\omega|\mathbf{X}, \mathbf{Y})$ of the weights ω, which are learned via a set of inputs $\mathbf{X} = \{x_1, ..., x_N\}$ and the corresponding output parameters $\mathbf{Y} = \{y_1, ..., y_N\}$ (Gal & Ghahramani, 2016). We optimize the dropout rates during the training phase.

Epistemic (model) uncertainty

Considering that neural networks can theoretically provide universal approximations, using dropout in this way is analogous to sampling over the space of functions (Gal & Ghahramani, 2016; Gal, 2016). The error associated with this sampling is associated with the ignorance of the model; this is known as *epistemic uncertainty*. In this method, a trained deep learning model could be interpreted in a Bayesian framework by optimizing its dropout rates and sampling the posterior $p(y|x, \mathbf{X}, \mathbf{Y})$ with the forward passes. Then, once the network is trained, the sampling of predicted values is simply forward passes upon which we apply dropouts. From this, we can sample a posterior for the predictions. This technique is known as Monte Carlo dropout.

However, defining the dropout rate is not a trivial task. For example, a fixed dropout probability will penalize larger weights compared to smaller ones (Gal et al., 2017), since when larger weights are dropped, they are likely to have a bigger impact on the results. To minimize the epistemic error in this scenario, one would optimize weights such that larger weights are dropped. However, this could result in the limiting case where all weights are set to 0, resulting in predictions that would always be 0 regardless of the dropout rates. This would correspond to a situation where there is no epistemic uncertainty. However, the model would also not be predictive at all. Therefore, the aim of defining a dropout rate is not to get optimized precision but to find a point where epistemic errors can be reasonably defined. Some authors proposed a grid search for this task. However, this procedure may be prohibitive in big and complex architectures.

In the variational scheme (Jordan et al., 1999; Graves, 2011a), one may define a procedure to optimize the dropout rate. The problem can be stated as follows: for a network, we can compute the PDF of a predicted value y with input x as:

$$p(y|x, \mathbf{X}, \mathbf{Y}) = \int p(y|x, \omega)\, p(\omega|\mathbf{X}, \mathbf{Y})\; d\omega\,. \tag{4.1}$$

The posterior $p(\omega|\mathbf{X}, \mathbf{Y})$ has explicit dependence on the training data set, and its form is generally infeasible to derive. We may infer the parameters from $q_\theta(\omega)$ by maximizing the log-evidence lower bound (Fox & Roberts, 2012):

$$L = \int q_\theta(\omega)\; \log p(\mathbf{Y}|\mathbf{X}, \omega)\; d\omega - \mathrm{KL}(q_\theta(\omega)||p(\omega))\,, \tag{4.2}$$

which is our objective function (loss function) that is optimized with respect to the neural network weights and the dropout probability parameterized in $q_\theta(\omega)$. We can also define the log probability in our regression problem as

$$\log p(\mathbf{Y}|\mathbf{X}, \omega) = \sum_n \sum_k \frac{-1}{2\sigma_k^2} ||y_{n,k} - \hat{y}_{n,k}(\mathbf{x}_n, \omega)||^2 - \frac{1}{2} \log \sigma_k^2, \tag{4.3}$$

where σ_k, the observation noise parameter for the k output parameter, $y_{n,k}$ and $\hat{y}_{n,k}$ are the true values and the predicted values respectively for the n training samples. The first term of equation 4.2 corresponds to a traditional loss term, the log-likelihood of the outputs for the training set. The integral can be performed by a Monte Carlo integration procedure. The second term is the Kullback – Leibler (KL) divergence, which parameterizes the distance between the distributions $p(\omega)$ and $q_\theta(\omega)$. The training process minimizes this distance. The KL divergence term can be approximated by a L_2 regularization (Gal & Ghahramani, 2016). For a set of parameters $\theta = \{\mathbf{M}, p_l\}_{l=1}^L$ in which $\mathbf{M}_l$ are the mean weight matrices

and p are the l^{th} layer, we may choose $q_\theta(\omega)$ as:

$$\begin{aligned} q_\theta(\omega) &= \prod_l q_{\mathbf{M}_l}(\mathbf{W}_l), \\ q_{\mathbf{M}_l}(\mathbf{W}_l) &= \mathbf{M}_l \cdot \text{diag}[\text{Bernoulli}(1-p_l)^{K_l}], \end{aligned} \tag{4.4}$$

where the set of random weight matrices are $\omega = \{\mathbf{W}_{l=1}^L\}$, with L layers and the dimension of each weight matrix given by K_l. The Bernoulli variables, $z = \text{Bernoulli}(1-p_l)$, drop some neural network weights with their given probability.

In this way, a deep learning model can be interpreted in a Bayesian framework by optimizing its dropout rates and sampling the posterior $p(y|x, \mathbf{X}, \mathbf{Y})$ with the forward passes. This technique is known as Monte Carlo dropout. The epistemic uncertainty of the posterior represents the ignorance of the model given a certain training set. However, in some cases, there may be some issues in performing this optimization. It can be shown that finding the minimum of the KL divergence term in Eq. 4.2 is equivalent to maximizing the entropy of a Bernoulli random variable with probability $1 - p$ (Gal et al., 2017). This penalizes larger models trained on small amounts of data because it pushes the dropout rate close to $p = 0.5$ in comparison to smaller models. Therefore, smaller models would have lower optimized dropout rates in the low-data regime. Nevertheless, with epistemic uncertainty, the dropout rate is lowered for both large and small models if we provide the neural network with more data to train on.

There remain caveats when evaluating the derivative of the objective function with respect to a dropout rate in discrete Bernoulli distributions. Therefore, we follow the prescription from Gal et al. (2017) and replace the Bernoulli variables with a Concrete distribution (Maddison et al., 2016), which is a continuous distribution with the ability to approximate discrete random variables. We sampled from the Concrete distribution that approximates

the one-dimensional Bernoulli, equivalent to a binary random variable[2]:

$$\tilde{z} = \text{sigmoid}\left(\frac{1}{t} \cdot [\log p - \log(1-p) + \log u - \log(1-u)]\right), \tag{4.5}$$

where t is a temperature parameter and u is the uniform distribution $u \sim \mathcal{U}(0,1)$.

After training, we derive 1000 realizations for each system. We define the 68%, 95%, and 99% ($1\sigma, 2\sigma$, and 3σ) confidence intervals. We compared the confidence interval of the scatter of the medians — the scatter of different objects with the same truth value — with the confidence levels of individual lenses obtained through Concrete dropout. The confidence intervals obtained from Concrete dropout were a little wider than the confidence intervals of the scatter of the medians but correlated closely with it.

Aleatoric (statistical) uncertainties

To get the total predictive error of the neural network model, we must also account for aleatoric uncertainty. To do so, we look at the noise proprieties of the data set. For a *homeostatic* data set — in which all the data has similar noise proprieties — aleatoric uncertainty is usually accounted for by adding a term to the model or loss function that can be manually fine-tuned. However, the data we used in this work has varying levels of signal-to-noise. Therefore, we estimated the aleatoric uncertainties in a *heteroscedastic* framework: we train the neural networks to predict σ_k, the observation noise parameter for the k output parameter in Eq. 4.3. This optimizes σ_k in the loss function Eq. 4.2.

2. Note that by the time Gal et al. (2017) article was published, the method was not implemented in convolutional layers. We used the updated version in the aforementioned repository, which works for convolutional layers.

Systematic uncertainties

Systematic uncertainties can overlap with both epistemic and aleatoric uncertainties. However, the latter forms of uncertainties may not span all instances of systematic uncertainty, which would be revealed in noise-free input data. Besides aleatoric uncertainty, there might remain other sources of systematic uncertainty in the data that cannot be reduced by the collection of more data points. For instance, in wide-field survey imaging, the pixel size and PSF are typically larger than in space-based observations. There may also be degeneracies that can result in more complex and correlated uncertainties than a random scatter on the predicted values. Strong lenses may have multiple source images that can be distorted in several ways, and can also be blended with the lens galaxy. Additionally, the lensed image has a parameter space with at least ten independent variables. This might be a relevant source of systematic errors when trying to extract information from images. For example, it can be significantly easier to infer the Einstein radius, $\theta_{\rm E}$, of a strong lens system in cases where the light from the lensed source is not blended with the light from the lens galaxy than in cases where it is, even if the noise level in the images are the same.

To evaluate a possible bias or scatter in our results, we visually compared the median predictions in our training sample to the respective truth values. We observed that, even when considering the epistemic uncertainties, there was a small bias in some of our model predictions that scaled linearly. To address this problem, we adopted the following procedure: we performed a linear fit between model predictions and the truth values and then subtracted the bias in the predictions. We then used the same linear fit to remove bias in the test data set. Therefore, our model consists of a deep learning prediction and a linear correction. After the fitting procedure, we found that the percentile error in the scatter in the medians and the percentile error due to the sampling performed by dropouts were consistently symmetric around the median and in most of the range around the $y = x$ line, except for certain high and low boundaries that corresponded to regions where the model has fewer samples. We

discuss this further in §4.6.

4.3 Simulated Data

To optimally train, validate, and test a neural network for strong lensing analysis, we require a large image catalog of strong lenses. Given the paucity of known strong lenses in the current census (∼1000 lenses to date), we used simulated lenses from `LensPop`[3] (Collett, 2015a) for training, validation and testing purposes. Here, we present a brief overview of the procedure used in the `LensPop` algorithm. For a complete description of `LensPop`, we refer interested readers to Collett (2015b).

`LensPop` first generates a synthetic population of galaxy-scale strong lensing systems in the sky. For the lens population, `LensPop` assumes SIE profiles for all lenses, with masses drawn from the velocity dispersion function of SDSS galaxies (Choi et al., 2007). Observations show that elliptical galaxies, which dominate the lensing probability of the universe (see Oguri & Marshall (2010) and references within), are well-approximated by Singular Isothermal Ellipsoid (SIE) mass profiles (Auger et al., 2010a; Koopmans et al., 2006b). The redshift of the lenses is drawn independently of the mass from the differential comoving volume function. The light profile of the lens is assumed to follow a de Vaucouleurs profile that is aligned and concentric with the mass. The lens colors are assumed to follow the rest-frame SED of a galaxy whose star formation occurred 10 Gyrs ago. For the source population in `LensPop`, the source light has elliptical exponential light profiles, with magnitude, color, and redshift distributions drawn from the sky catalogs of Connolly et al. (2010).

The observing conditions of the imaging survey are then simulated and applied to the synthetic lenses to produce a mock catalog of lens imaging data that mimics that survey. In this work, we simulated the observing and instrumental capabilities of the DES survey to produce lenses with DES-like image quality. The mock images are created by first pixelating

3. `https://github.com/tcollett/LensPop`

the model lens image to the pixel scale of the detector of the survey instrument. The pixelated images are then convolved with circular atmospheric PSFs. Poisson noise from the lens, source, uniform sky background, and CCD read noise are then added to the mock images. The zero points, exposure times, number of exposures, pixel scale, read noise, filter bands, and survey area are taken from DES survey specifications. The seeing and sky brightness are stochastic variables drawn from DES data and are described in Table 1 of Collett (2015b).

Every simulated lens in our data set is deemed DES-observable. We follow the criteria set in Collett (2015b) to determine which lensing systems are detectable by DES:

Multiply imaged: All detectable lenses must be multiply-imaged. Therefore, we have

$$\theta_{\rm E}^2 > x_s^2 + y_s^2, \tag{4.6}$$

, where $\theta_{\rm E}$ is the Einstein radius, and x_s and y_s, are the unlensed source position relative to the lensing galaxy.

Resolved images: In at least one of the g, r, i bands, the image and counter-image must be resolved. Hence, we also require that:

$$\theta_{\rm E}^2 > r_s^2 + (s/2)^2, \tag{4.7}$$

where r_s^2 is the half-light radius of the source, and s is the seeing.

Resolved tangential shear: Additionally, the tangential shear of the magnified source images in the image plane must also be resolved, and the magnification has to be large enough that the source images are noticeably sheared. Following Collett (2015b), we adopt:

$$\mu_{\rm total} r_s > s \quad \text{and} \quad \mu_{\rm total} > 3, \tag{4.8}$$

where $\mu_{\rm total}$ is the total magnification of the source.

Source Signal-to-noise: Finally, the signal-to-noise ratio of the lensed source (i.e., the lens-subtracted light) in that band, S/N, must be high enough that it is feasible to identify lensing features and to determine if the above criteria are met. Therefore, following Collett (2015b), we set

$$S/N > 20. \tag{4.9}$$

In summary, detectable lenses must obey Eqs. 4.6-4.9. Using `LensPop`, we generated $18,600$ simulated DES-observable galaxy-galaxy lensing systems. The distributions of the Einstein radii ($\theta_{\rm E}$), velocity dispersion ($\sigma_{\rm v}$), and lens and source redshifts ($z_{\rm l}$, $z_{\rm s}$) in the DES simulated data set agree with that of Collett (2015b). Fig. 4.3.1 shows a representative random sample of 20 DES-observable systems from the total dataset.

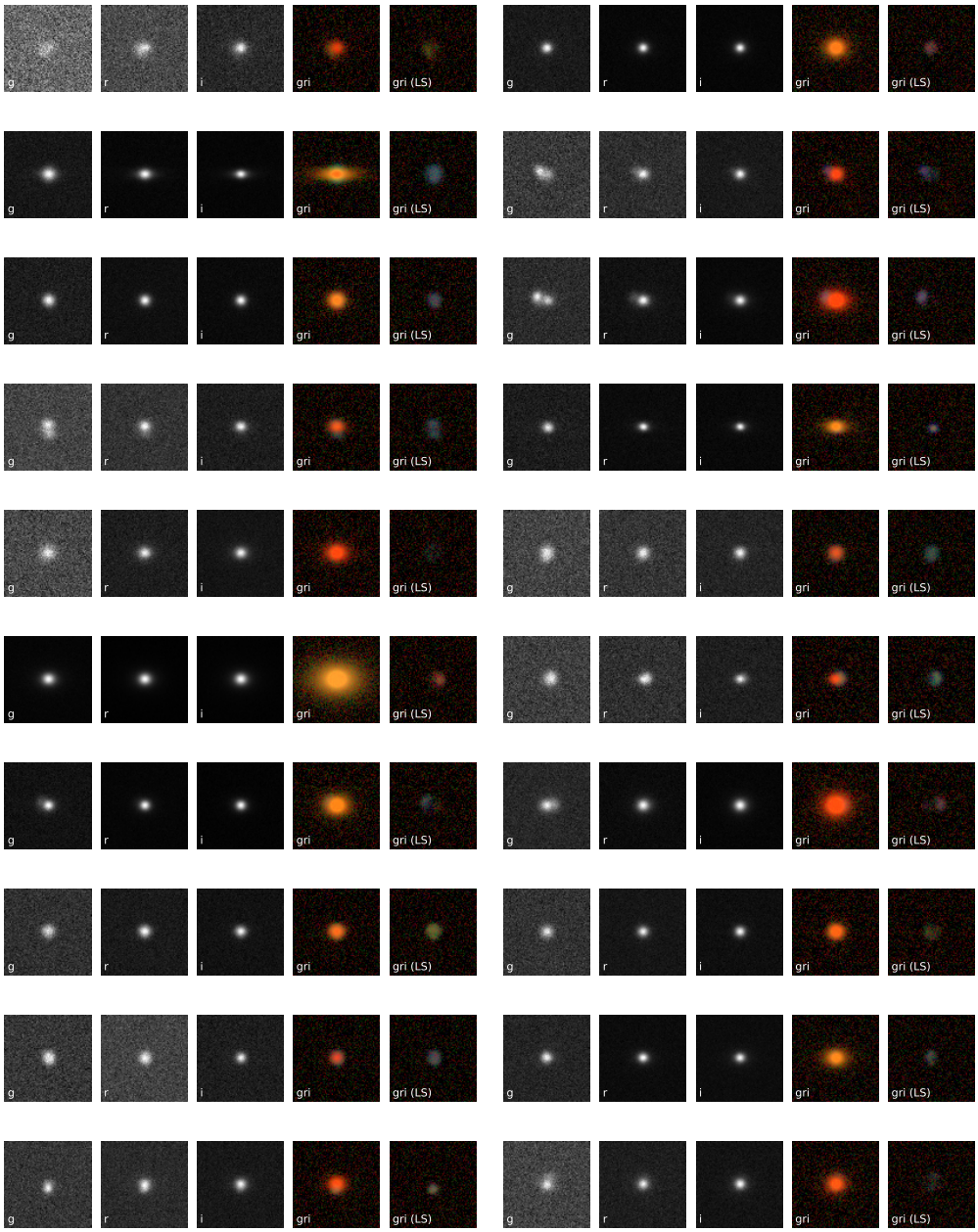

Figure 4.3.1 A random selection of 20 DES-observable lens systems from the `LensPop` simulation. For each system, the individual g, r, i, band images, as well as false-color gri composite image with lens galaxy ('gri') and without lens galaxy ('gri (LS)'), are shown from left to right. False-color images were made following Lupton et al. (2004). Images are sorted left-to-right, top-to-bottom by increasing Einstein radii. Refer to §4.3 for further information on the simulated images dataset.

4.4 Training the Inception Deep Learning Model

The strong lensing sample was divided into two groups: 80% for training, and 20% for testing/validation. The training subset is the only one used to update the weights of the network in the backpropagation algorithm (Ruder, 2016). We trained the inception architecture for each of the parameters $\theta_{\rm E}$, $\sigma_{\rm v}$, $z_{\rm l}$, and $z_{\rm s}$ individually, producing a model optimized for each parameter. The fine-tuned hyperparameters used for training the architecture were found by manually changing their values within a certain range until the (local) maximum accuracy on regression was achieved. The chosen batch size for training was $2,000$, while the maximum number of epochs was set to 400. To avoid overfitting, and to improve model convergence, both a learning rate reducer and an early stopper were used. The training was performed with an Adam optimizer. The model was trained on a 24-core Intel Xeon CPU X5670 (2.93 GHz) and a GeForce GTX 1080 GPU.

The training time of each model was $\sim$ 4 hours. In Fig. 4.4.1, we present the training diagnostic plot loss vs. epoch for the z_s, where we set a fixed number of epochs to 200 for the training and validation dataset. The plot shows the training loss on both the training and validation data set as a function of the training epoch. The trend of the training loss suggests no overfitting. The training loss curves for the $\sigma_{\rm v}$, $\theta_{\rm E}$, and z_L parameters were similar to that of z_s.

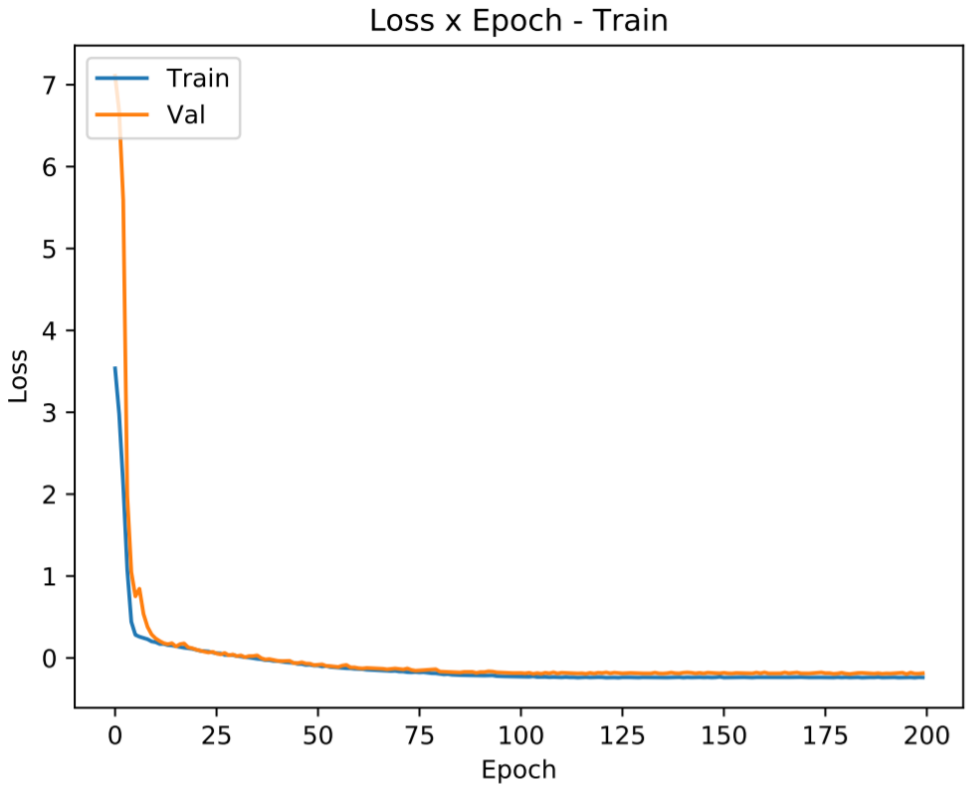

Figure 4.4.1 Training (blue) and validation (orange) loss curves for the inception network trained on the source redshift parameter, z_s.

4.5 Results

We performed regression with the trained models on our validation sample of 3720 systems. The model takes less than ~ 1 sec to sample 10^3 values of the parameter posterior distribution for each strong lens system. We then evaluate the model predictions for $\theta_{\rm E}$, $\sigma_{\rm v}$, $z_{\rm l}$, and $z_{\rm s}$. In Figs. 4.5.1, 4.5.2, 4.5.3, and 4.5.4, we present one-to-one comparisons of prediction to truth, residuals (the difference in the prediction and truth, divided by the truth), and sample distributions for each lensing system parameter.

We find that all parameters except for $z_{\rm s}$ are well predicted for a large range of values. In particular, we observe that most of the predicted values remained within 15% of the truth values, except for $z_{\rm s}$. However, there are noticeable biases and larger uncertainties at the high and low ends of the parameter ranges. All the sample distributions are non-uniform, with fewer samples at the high and low ends of the distributions. Moreover, the distributions are mostly asymmetric: for $\theta_{\rm E}$, $z_{\rm l}$, and $z_{\rm s}$, the distributions peak below the median value, while for $\sigma_{\rm v}$, they peak close to the median value. The regions of smaller sample sizes correspond to the regions of increased bias and uncertainty. The predicted distributions also correlate strongly with the input distributions, as expected.

In the top part of Fig. 4.5.5, we present the median of the fractional deviation and the 68% confidence level percentile. We do not notice any strong bias in any parameter. In the bottom part of the same figure, we evaluate the size of the high-deviation sample, which we define as the subset of the test lenses where the predicted parameter values deviate from the true values by more than 15%. For all parameters except $z_{\rm s}$, the number of lenses in the test set that were considered high-deviation was fewer than 20% of the total number of test lenses. We also compared the performance of models where the model was trained on each lensing system parameter individually with the model where all parameters were trained simultaneously. We found that whether to train the parameters individually or together produced models that performed quite similarly. However, the model where all parameters

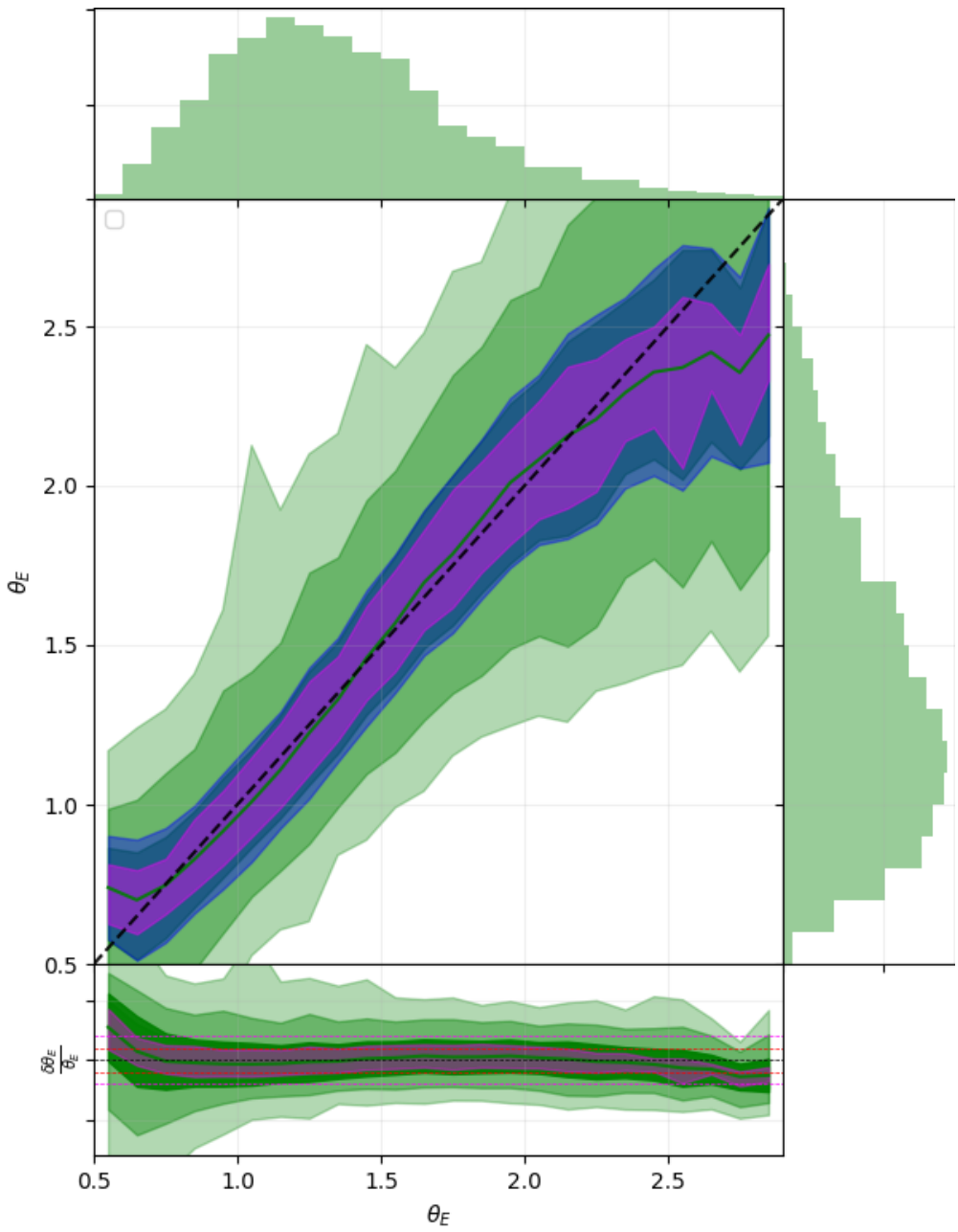

Figure 4.5.1 True vs predicted values and the residuals for the Einstein Radius, $\theta_{\rm E}$. The green regions correspond to the $1\sigma - 2\sigma - 3\sigma$ uncertainty regions, and the blue and magenta regions correspond to the standard deviation and 1σ scatter of the median predicted values respectively. The dashed lines in the residual plot (below) represent the 0.0 (black), 0.1 (red), and 0.2 (magenta) fractional difference respectively. The distributions above and on the right side of the figure show the distribution of $\theta_{\rm E}$ in the training data.

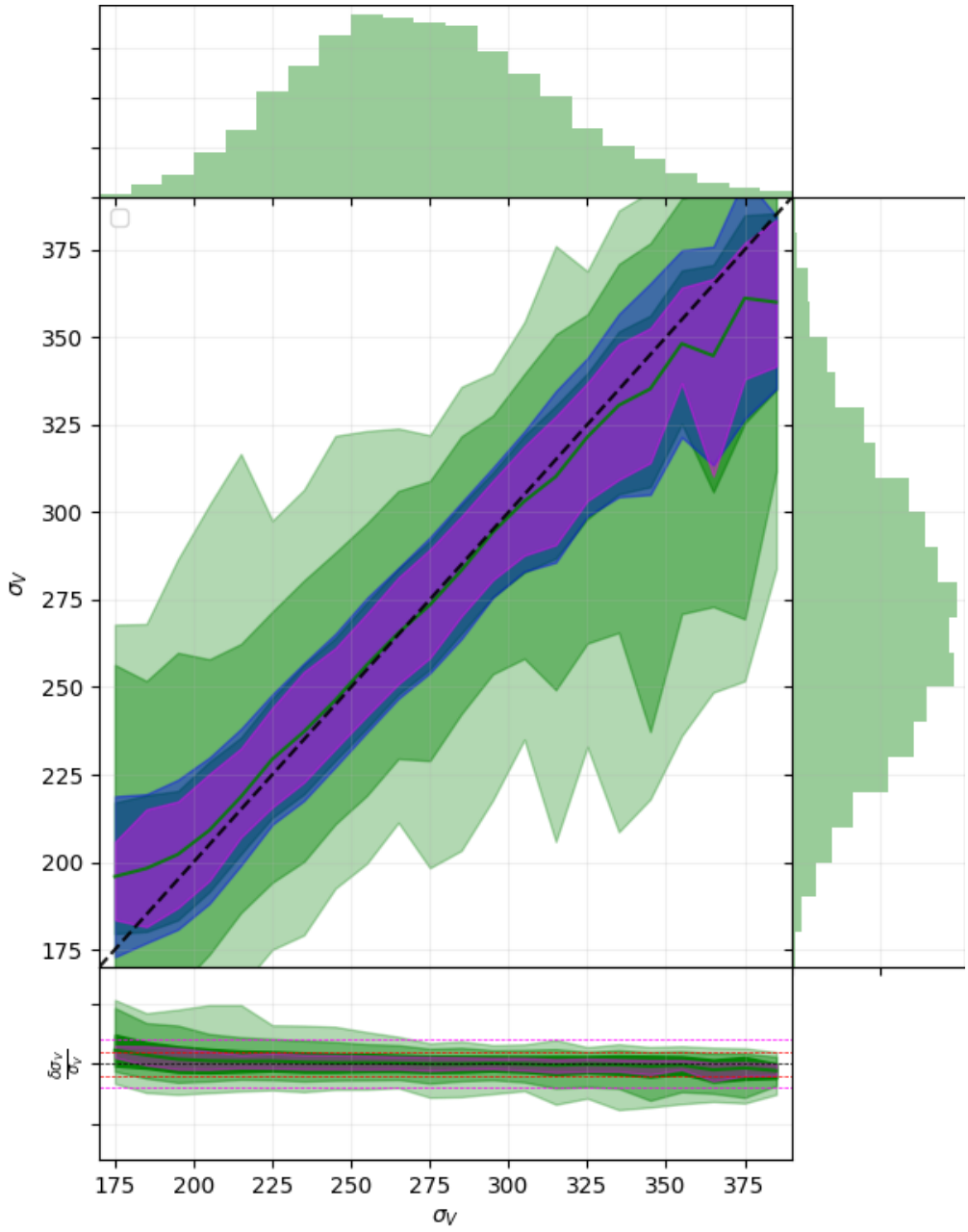

Figure 4.5.2 True vs predicted values and the residuals for the Einstein Radius, $\sigma_{\rm v}$. The green regions correspond to the $1\sigma - 2\sigma - 3\sigma$ uncertainty regions, and the blue and magenta regions correspond to the standard deviation and 1σ scatter of the median predicted values respectively. The dashed lines in the residual plot (below) represent the 0.0 (black), 0.1 (red), and 0.2 (magenta) fractional difference respectively. The distributions above and on the right side of the figure show the distribution of $\sigma_{\rm v}$ in the training data.

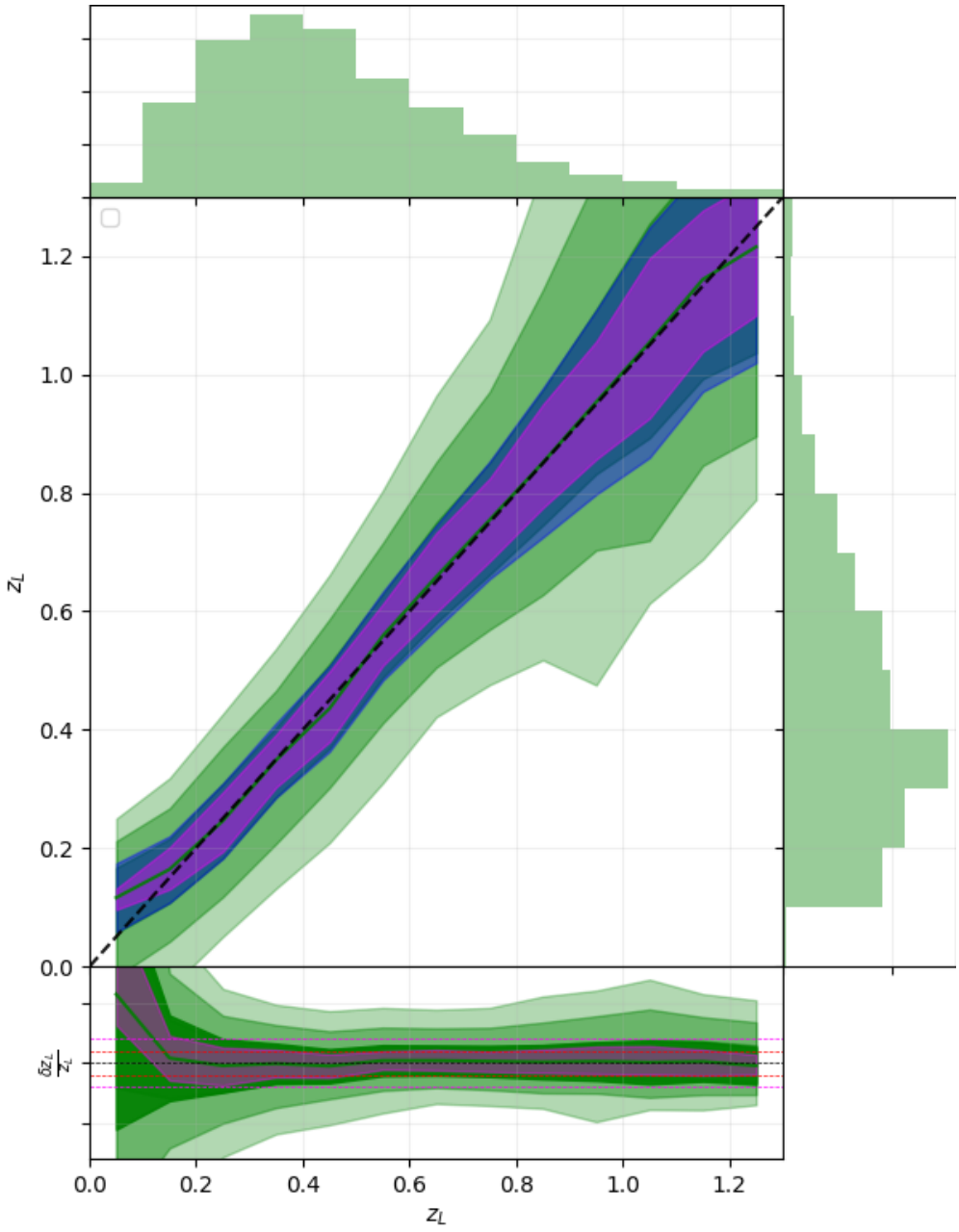

Figure 4.5.3 True vs predicted values and the residuals for the Einstein Radius, $z_{\rm l}$. The green regions correspond to the $1\sigma - 2\sigma - 3\sigma$ uncertainty regions, and the blue and magenta regions correspond to the standard deviation and 1σ scatter of the median predicted values respectively. The dashed lines in the residual plot (below) represent the 0.0 (black), 0.1 (red), and 0.2 (magenta) fractional difference respectively. The distributions above and on the right side of the figure show the distribution of $z_{\rm l}$ in the training data.

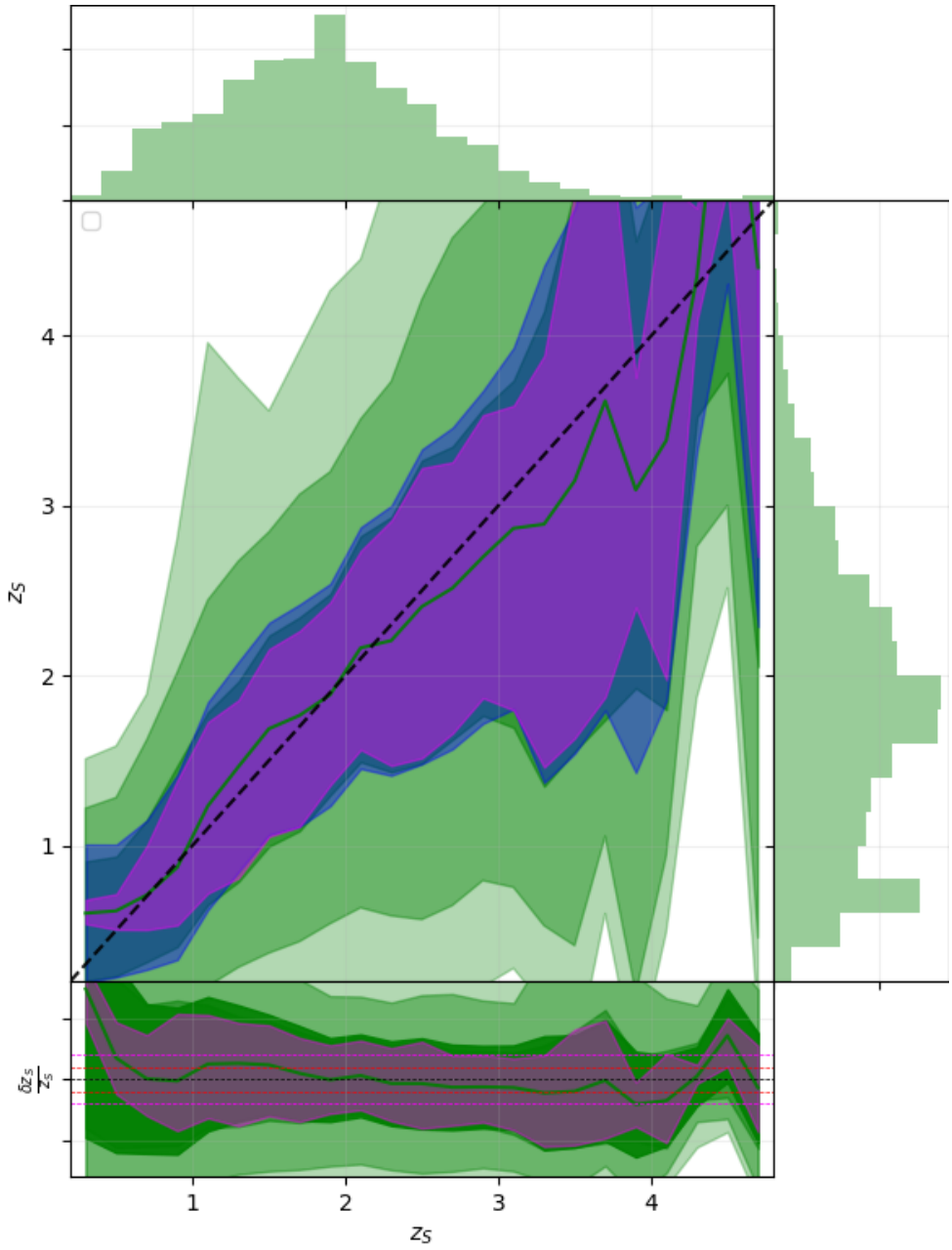

Figure 4.5.4 True vs predicted values and the residuals for the Einstein Radius, $z_{\rm s}$. The green regions correspond to the $1\sigma - 2\sigma - 3\sigma$ uncertainty regions, and the blue and magenta regions correspond to the standard deviation and 1σ scatter of the median predicted values respectively. The dashed lines in the residual plot (below) represent the 0.0 (black), 0.1 (red), and 0.2 (magenta) fractional difference respectively. The distributions above and on the right side of the figure show the distribution of $z_{\rm s}$ in the training data.

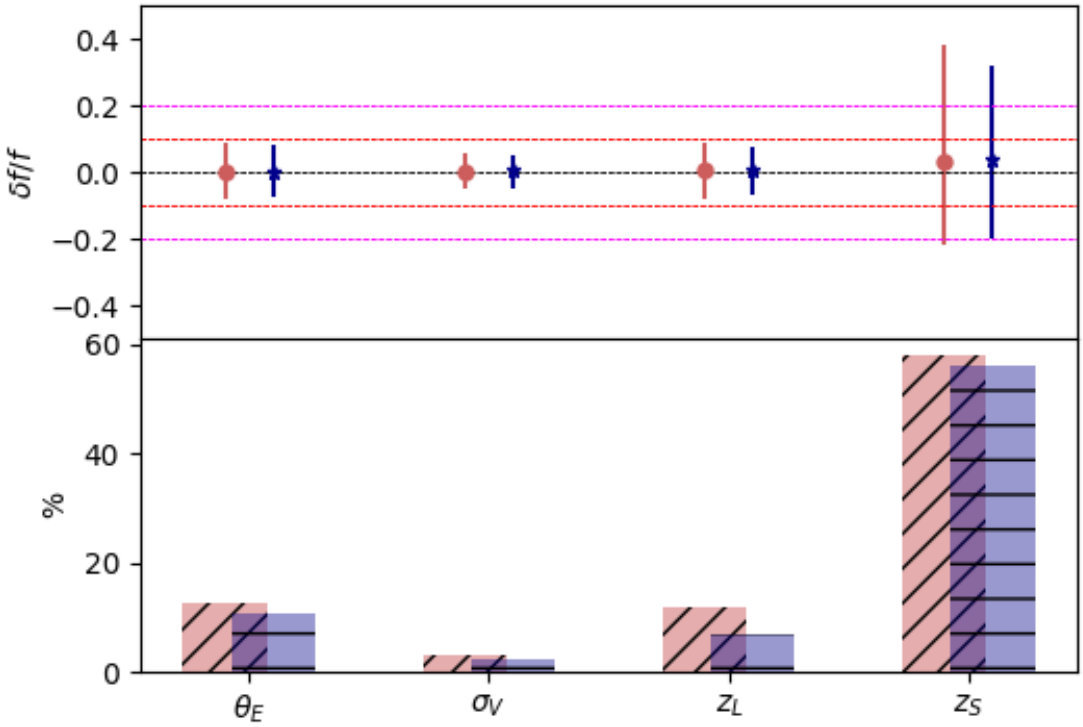

Figure 4.5.5 **Top:** The average fractional difference between the predicted and true values for each parameter in the test set. The dashed lines represent the 0.0 (black), 0.1 (red), and 0.2 (magenta) fractional difference respectively. The red circle indicates models where only a single parameter was regressed, while the blue star indicates the model where all 4 parameters were regressed together. **Bottom:** The percentage of lenses in the test set where its predicted parameter value deviates from its true value by more than 15%. The blue and red bars correspond to the models where the parameters were individually modeled and modeled together respectively.

were simultaneously modeled performed slightly better across all parameters (see Fig. 4.5.5).

We also investigated how the fractional deviation changes as a function of other physical values, like the magnitude and size of the lensing object. In most cases, we found that regions of higher bias or error corresponded to regions in parameter spaces where we had less training data, suggesting that the uncertainty or bias could be overcome with more data. However, there were still some interesting trends, such as the ones plotted in Figs. 4.5.6 and 4.5.7. From the aforementioned figures, we observed that both smaller and bigger values of $\theta_{\rm E}$ were correlated with larger fractional deviations from true values in $z_{\rm s}$, to the order of $+40\%/-60\%$. Though the error bars are wide, and it does not exclude 0% deviation (perfectly accurate predictions), it is worth mentioning that there are regions in the high $\theta_{\rm E}$ parameter space that have a lower number of examples, and the errors might be more poorly defined in that region. $\theta_{\rm E}$ is also connected to $z_{\rm s}$ though cosmological distances. In the $z_{\rm l}$ case, we observe that lower/higher velocity dispersion may be linked to $+40\%/-20\%$ deviations.

In addition to Figs. 4.5.6 and 4.5.7, we also observed that predicted errors in $\theta_{\rm E}$ are wider by a factor of ~ 2, i.e., $\sim \pm 20\%$ when the $\sigma_{\rm v}$ parameter is below 230 km/s. As $\sigma_{\rm v}$ scales with mass and is therefore connected to $\theta_{\rm E}$, this trend suggests that as the effects of strong gravitational lensing get weaker, the uncertainty in the inferred Einstein radius increases. This is an intuitive result, as the lensing features of smaller lensing masses are harder to observationally detect and measure.

At low- and high-value regions of each lens parameter, we also observed a systematic bias and higher predicted uncertainties compared to lenses in the middle of each parameter region. This is likely caused by the relatively low number of training examples in those regions of parameter space. Another source of uncertainty in systems with smaller Einstein radii arises from the blending of lens and source light, making it harder to differentiate canonical lensing features in those systems. This manifests in our results, where the fractional uncertainties

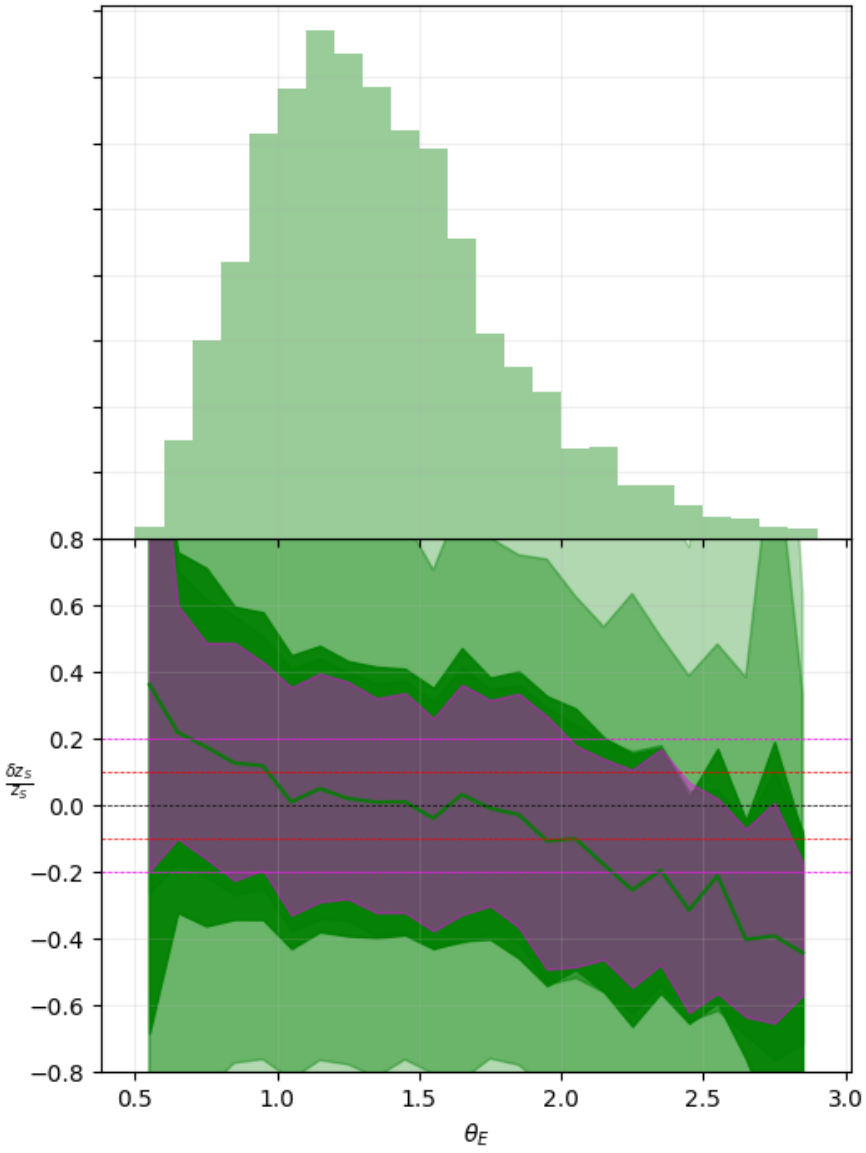

Figure 4.5.6 The fractional deviation from true values in source redshift $z_{\rm s}$ as a function of Einstein Radius θ_E. The green regions correspond to the $1\sigma - 2\sigma - 3\sigma$ uncertainty regions, and the magenta region corresponds to the 1σ scatter of the median predicted values. The dashed lines represent the 0.0 (black), 0.1 (red), and 0.2 (magenta) fractional difference respectively. The distribution above shows the distribution of θ_E in the training data.

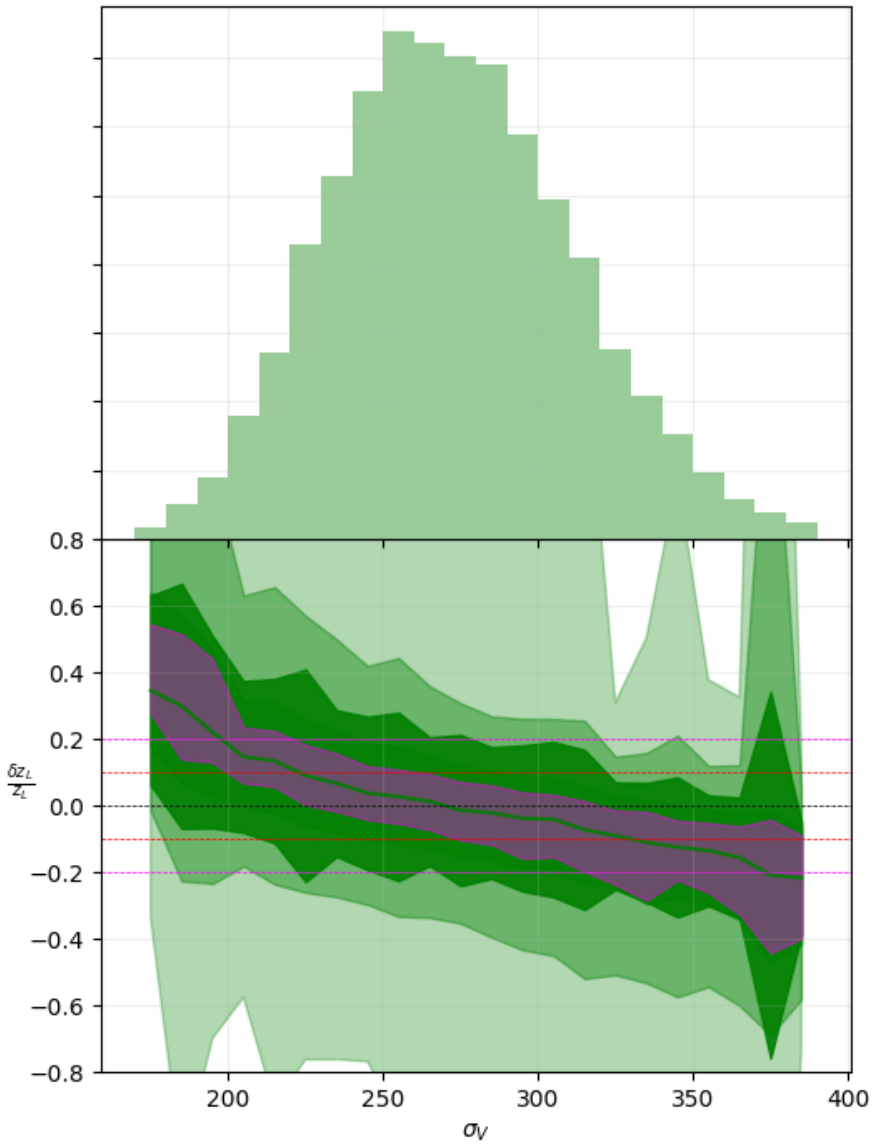

Figure 4.5.7 The fractional deviation from true values in lens redshift $z_{\rm l}$ as function of velocity dispersion $\sigma_{\rm v}$. The green regions correspond to the $1\sigma - 2\sigma - 3\sigma$ uncertainty regions, and the magenta region corresponds to the 1σ scatter of the median predicted values. The dashed lines represent the 0.0 (black), 0.1 (red), and 0.2 (magenta) fractional difference respectively. The distribution above shows the distribution of $\sigma_{\rm v}$ in the training data.

for $\theta_{\rm E}$ are $\sim 20\%$ for small Einstein radii, but decrease to $\sim 10\%$ level for larger Einstein radii. More examples in these regions of parameter space could lower the model's uncertainty. However, as the uncertainties decrease, more sources of systematic errors that were previously subdominant may now manifest. In future work, we seek to a) address the interplay and trade-offs for various kinds of uncertainties; and b) explore how changes in image quality affect these results.

4.6 Conclusion

In summary, we show that Bayesian deep neural network models can be used to predict the lensing parameters of strong lensing systems in simulated wide-survey imaging data. These parameter predictions include estimates of both epistemic and aleatoric uncertainties, and we verified that the scatter on thousands of individual systems predictions was lower than this uncertainty. For example, the velocity dispersion was constrained to lower than 10% uncertainty using only three bands. The current results demonstrate that deep learning is a promising approach for automating lens analyses without involving large amounts of computation time or highly specialized labor.

A deep learning pipeline such as the one we present in this work can in principle be used to select well-characterized systems for follow-up studies. It can also be used as a precursor to a manual MCMC pipeline, where the results of the deep learning model are used as the priors for the MCMC analysis. The results can also enable statistical studies that leverage strong lenses with $\sim 15\%$ uncertainties in its inferred lens parameters. An example of this is in galaxy-galaxy strong lensing cosmology (Cao et al., 2015b; Chen et al., 2018), where measurements of $\theta_{\rm E}$, combined with an independent measurement of $\sigma_{\rm v}$, can be used to constrain cosmological models and galaxy mass-density profiles (Li et al., 2018a).

One interesting result of this work is that the fractional deviation of the inferred velocity dispersions $\sigma_{\rm v}$ of the lenses from their true values was around 10%. This precision is compet-

itive with spectroscopic surveys such as BELLS (Bolton et al., 2012), and SLACS (Bolton et al., 2006). We note that the input data of our method include not only the photometric information from three bands but also the mass and light morphology of both lens and source respectively. The inclusion of this ancillary data may have helped the neural network model constrain the inferred velocity dispersions $\sigma_{\rm v}$ of the lensing galaxies. This is supported by the work of Davis et al. (2003), who showed that if $\theta_{\rm E}$ or angular separation θ, $z_{\rm l}$ and $z_{\rm s}$ are well-constrained, one could estimate the velocity dispersion $\sigma_{\rm v}$ with this level of accuracy, given a density profile and a cosmology. If the velocity dispersion of lens galaxies can be constrained from strong lensing in this way, it could be applied to tests of modified gravity such as those conducted by Cao et al. (2017); Schwab et al. (2009). This can be further investigated with techniques, such as the Local Interpretable Model-Agnostic Explanations (LIME; Ribeiro et al., 2016). This will be the subject of future investigation.

It should be noted that, due to the somewhat idealized nature of `LensPop`'s prescription for simulating a DES-like image dataset, the uncertainties quantified here may be lower than the uncertainty in our inferences from real imaging data. While `LensPop`'s prescription is well motivated by both theoretical and empirical considerations, it makes many simplifying assumptions about the populations of lenses and sources, as well as the simulated observing conditions of the systems (some of which are discussed in §7 of Collett (2015b)). Real strong lensing systems are likely to have characteristics that deviate from these assumptions to various extents. For example, the mass profile of a real strong lensing system may not be well described by a singular isothermal ellipsoidal profile, and the source may have a more irregular morphology that cannot be well-approximated by an elliptical exponential light profile. `LensPop` also assumes that each lensing system is found in isolation. In reality, elliptical galaxies, which constitute the majority of galaxy-scale lenses, tend to cluster, which leads to external perturbations to the lensing potential of the lens system by nearby galaxies, systematically affecting the lensing observables. These nearby galaxies can also crowd the

field of view of the lens system. `LensPop` also ignores objects that may be situated along the line of sight of the lens by coincidence. These inhomogeneous scenarios can result in other forms of heteroscedastic uncertainty when the regression is applied to real data.

In such regimes, one needs to properly estimate both random and systematic uncertainties from the data, instead of simply performing post hoc calibration to account for it. A possible way to reduce the impact of heteroscedasticity due to factors not considered in idealized simulations might be to do *transfer learning*. Transfer learning involves retraining some layers of an already-trained model (such as the one in this work) on real data. However, since there are orders of magnitude fewer strong lensing systems discovered in real data to date compared to the size of the training set used in this work, properly conducting transfer learning is a major challenge and one might also need to make use of data augmentation methods. We are currently evaluating the relevance of idealized simulations by working on simulations with increased degrees of realism. This work therefore represents a novel step towards building a more robust framework to analyze strong lensing systems found in current and future ground-based survey data.

CHAPTER 5

STRONG LENSING PARAMETER ESTIMATION ON GROUND-BASED IMAGING DATA USING SIMULATION-BASED INFERENCE

The text of this chapter is in preparation for submission to MNRAS.

5.1 Introduction

When light from a distant source is deflected by a massive object along its line-of-sight to an observer, causing the source image to appear multiply imaged or distorted to the observer, we have what is known as *strong gravitational lensing*. Strong gravitational lensing systems are well-established observational probes of astrophysics and cosmology; on the astrophysical side, the amplification of the source galaxy or quasar light due to gravitational lensing allows for the observation and analysis of high-redshift sources that would otherwise be too faint to detect (e.g., Bayliss et al., 2014; Hainline et al., 2009; Quider et al., 2009; Smit et al., 2017; Dessauges-Zavadsky et al., 2009; Christensen et al., 2012; James et al., 2013; Yuan & Kewley, 2009; Finkelstein et al., 2009; Iwata et al., 2008; Jones et al., 2013). Additionally, the morphology of lensed images provides information about the distribution of dark and baryonic matter in the lens systems (e.g., Auger et al., 2010b; Barnabe et al., 2011; Newman et al., 2015, 2012a,b). On the cosmological front, strong lensing systems with multiple images of a time-varying source (e.g., quasars and supernovae) can be used to constrain the Hubble constant as well as models of dark energy (e.g., Suyu et al., 2010a, 2013; Shajib et al., 2020; Jee et al., 2016, 2019).

Achieving these science objectives requires precise and accurate modeling of statistical samples of lensing systems. In the simplest case, a lens system comprises light from a single source object (e.g., a distant galaxy or quasar) that is lensed by a single intervening

lens object (e.g., a galaxy or a galaxy cluster) along the line of sight. Conventional lens modeling involves reconstructing the lens galaxy mass distribution and source galaxy light distribution from imaging (and perhaps ancillary) data. A well-established class of modeling methods involves Bayesian parametric forward modeling, where both the source light and the lens mass distributions are approximated by parametric models (e.g. Birrer & Amara (2018); Birrer et al. (2021); Kneib et al. (2011); Suyu et al. (2010a, 2013) and the software mentioned in Lefor et al. (2013). An explicit likelihood function is constructed from the residual (RMS) difference between the observed image and the reconstructed model image. This likelihood function is then maximized, often through random sampling methods such as Markov Chain Monte Carlo sampling. The posterior model parameter distribution is then used for lensing science applications.

Approximately a thousand galaxy-galaxy strong lensing systems have been discovered in large cosmic imaging surveys to date (e.g. Zaborowski et al. (2023); Jacobs et al. (2019a, 2017) and the catalogs referenced in Chen et al. (2019a); Moustakas (2012) and the Gravitationally Lensed Quasar Database[1]). The Vera Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezić et al., 2019) is expected to find about 100,000 such systems, and space-based surveys such as Euclid[2] and the Nancy Grace Roman Space Telescope[3] will additionally find tens of thousands more (Collett et al., 2015; Weiner et al., 2020). While the current approach to strong lens modeling described above has been successful for samples of order 100 or less (Tan et al., 2023) , it is labor- and therefore time-intensive; scaling it to the projected LSST strong lens sample would require an estimated $\sim$ 1000 person-years of effort for the analysis, which is infeasible. This necessitates the development of more efficient techniques to model strong lensing systems. While there have been some efforts to develop automated lens modeling techniques (e.g. Nightingale et al., 2018), this is still a nascent field

1. https://research.ast.cam.ac.uk/lensedquasars/

2. https://www.cosmos.esa.int/web/euclid

3. https://roman.gsfc.nasa.gov

with many approaches that have not yet been explored.

In this paper, we use Simulation-Based Inference (SBI; for an overview of these methods, see Cranmer et al., 2019) to model strong lensing systems in synthetic imaging data designed to emulate that in ground-based surveys such as DES. SBI methods leverage simulators at the core of a statistical inference procedure that can accurately infer the likelihood or posterior distributions of a scientific analysis for the experiment or observation at hand. These methods have previously been used in a variety of astronomical tasks, such as the analysis of supernova data, tomographic cosmic shear, inference of the HI ionization rate from high-redshift Lyman-α forests in Alsing et al. (2019b), dark matter substructure inference in galaxy-galaxy strong lenses (Brehmer et al., 2019a; Coogan et al., 2022; Anau Montel et al., 2023), inference of variability properties of astrophysical objects from dead time-affected light curves in X-ray observations (Huppenkothen & Bachetti, 2021), and the inference of the Hubble constant from binary neutron star mergers (Gerardi et al., 2021).

For strong lensing, there have been various efforts to automatically infer strong lensing parameters using a variety of techniques (e.g., for Neural Posterior Estimation (NPE): Legin et al. (2021); Wagner-Carena et al. (2023); for Bayesian Neural Networks (BNN): Levasseur et al. (2017); Pearson et al. (2021); Wagner-Carena et al. (2021); Park et al. (2021), for Neural Ratio Estimation (NRE): Coogan et al. (2022); Anau Montel et al. (2023); with Vision Transformers: Huang et al. (2022)). These studies have thus far focused on space-based imaging data, with a very small pixel scale and no image blurring due to atmospheric turbulence (seeing). Since the majority of new lens discoveries will come from ground-based survey data, we present a novel and complementary application of both NPE and BNN methods on simulated ground-based imaging data.

This paper is organized as follows. In Section 5.2 we give an overview of strong lens system modeling. Details about how NPE is implemented, as well as BNN that are used for comparison, are given in Section 5.3. We describe the simulator we use and discuss how we

create our datasets in Section 5.4. Metrics for evaluating model performance are discussed in Section 5.5. Our results are presented in Section 5.6, with discussion in Section 5.7 and conclusion in Section 5.8.

5.2 Strong Gravitational Lensing

General Relativity predicts that when light from a distant source passes near an intervening massive object (the lens), the mass causes light to deflect around the object. The deflection field is given by the lens equation,

$$\hat{\beta} = \hat{\theta} - \hat{\alpha}, \tag{5.1}$$

where $\hat{\beta}$ are 2D angular positions in the source plane, $\hat{\theta}$ are positions in the image plane, and $\hat{\alpha}$ is the deflection angle caused by gravitational lensing. The mapping from source plane coordinates, $\hat{\beta}$, to image plane coordinates, $\hat{\theta}$, is generally non-linear, and a single source position $\hat{\beta}$ may map to multiple image positions $\hat{\theta}$ in the image plane . The deflection angle $\hat{\alpha}$ depends on the mass distribution of the intervening lensing object(s) and the impact parameter of a given ray in the lens plane relative to the lens mass distribution. Solving this lens equation is central to lens modeling problems, but analytic solutions exist for only the simplest parametric lens mass models. There is a wealth of literature describing the mathematical formalism of lensing (e.g., Narayan & Bartelmann, 1996), cosmological and astrophysical applications (e.g., Treu, 2010), and solutions for some of the most common lens mass profiles (e.g., Kassiola & Kovner, 1993; Kormann et al., 1994; Barkana, 1998).

The population of strong lens systems predicted to be found in future surveys is dominated by galaxy-galaxy lensing systems, in which a single source galaxy is lensed by a single lensing galaxy, leading to a distorted (strongly sheared) image of the source. The lens population is dominated by massive, Early-type (elliptical) galaxies (Oguri & Marshall (2010)

and references within), the mass distributions of which are well-approximated by Singular Isothermal Ellipsoid (SIE) profiles. For the SIE profile, the reduced surface mass density (i.e., convergence) is given by

$$\kappa(x', y') = \frac{1}{2}\left(\frac{\theta_{\rm E}}{\sqrt{q_{\rm lens}x'^2 + y'^2/q_{\rm lens}}}\right), \tag{5.2}$$

where (x', y') are 2D image plane coordinates oriented such that x' is aligned with the semi-major axis of the mass distribution, $q_{\rm lens}$ is the ratio of semiminor to semimajor axis of the mass profile (the lens ellipticity), and $\theta_{\rm E}$ is the Einstein radius, defined as the effective radius within which the average convergence $\langle\kappa\rangle = 1$, and which characterizes the scale length of deflections of the source light by the lensing galaxy.

The solution for the deflection angle $\hat{\alpha}$ in Eq. 5.1 in the same coordinate system as Eq. 5.2 is given analytically for SIE lenses by (Kochanek & Schechter, 2004)

$$\alpha_{x'} = \theta_E\left(\frac{\sqrt{q_{\rm lens}}}{\sqrt{1-q_{\rm lens}^2}}\right)\arctan\left(\frac{x'\sqrt{1-q_{\rm lens}^2}}{\sqrt{q_{\rm lens}^2x'^2 + y'^2}}\right), \tag{5.3}$$

$$\alpha_{y'} = \theta_E\left(\frac{\sqrt{q_{\rm lens}}}{\sqrt{1-q_{\rm lens}^2}}\right)\operatorname{arctanh}\left(\frac{y'\sqrt{1-q_{\rm lens}^2}}{\sqrt{q_{\rm lens}^2x'^2 + y'^2}}\right). \tag{5.4}$$

where $\alpha_{x'}$ and $\alpha_{y'}$ are the x'- and y'- components of the deflection angle. To fully describe the SIE profile, we must also specify the position of the center of the lens mass (x, y) in an observer's 2D Cartesian coordinate system ($\hat{x}$,$\hat{y}$) in the image plane and the position angle of the lens $\phi_{\rm lens}$ in that system; $\phi_{\rm lens}$ is thus the rotation angle between the coordinate axes (x',y') and ($\hat{x}$,$\hat{y}$). Because $\phi_{\rm lens}$ is a cyclic parameter, which can complicate numerical analysis, it is often convenient to perform a parameter transformation from the lens ellipticity

and position angle to the eccentricity moduli:

$$l_{e1} = \frac{1 - q_{\rm lens}}{1 + q_{\rm lens}} \cos(2\phi_{\rm lens}), \tag{5.5}$$

$$l_{e2} = \frac{1 - q_{\rm lens}}{1 + q_{\rm lens}} \sin(2\phi_{\rm lens}). \tag{5.6}$$

The SIE model profile is therefore specified by five parameters: the Einstein radius ($\theta_{\rm E}$), the eccentricity moduli (l_{e1}, l_{e2}), and the 2D position of the center of mass (x, y).

In addition to the mass profile parameters, we include an external shear component to emulate the effect of external mass perturbers near the line of sight. The external shear is parameterized by a shear modulus γ and position angle ϕ_γ (Keeton & Kochanek, 1997) in the (x, y) coordinate system. However, as with the lens ellipticity and position angle, we avoid working in the cyclic coordinate ϕ_γ by changing variables to two external shear components:

$$\gamma_1 = \gamma \cos(2\phi_\gamma), \tag{5.7}$$

$$\gamma_2 = \gamma \sin(2\phi_\gamma). \tag{5.8}$$

Next, we model the source galaxy light profile as a 2D Sersic profile (Sersic, 1968):

$$I(r) = I_e \exp\left(-b_n \left(\left(\frac{r}{R}\right)^{\frac{1}{n}} - 1\right)\right), \tag{5.9}$$

where

$$r = \sqrt{q_{\rm src} x''^2 + y''^2/q_{\rm src}} \tag{5.10}$$

is the projected radial distance from the center of the light profile, $q_{\rm src}$ is the ratio of semiminor to semimajor axis (i.e., the ellipticity) of the source light profile, x'' and y'' are 2D Cartesian coordinates oriented such that x'' is aligned with the semimajor axis of the light

profile, and the source-light position angle in the (x, y) coordinate system is given by $\phi_{\rm src}$. The coefficient b_n in Eq. 5.9 is a function of the Sersic index n and is well-approximated by

$$b_n \approx 2n - \frac{1}{3} + \frac{4}{405n} + \frac{46}{25515n^2}, \tag{5.11}$$

in the range 0.5 ¡ n ¡ 10 (Ciotti & Bertin, 1999). The parameter R is the half-light radius (i.e., the radius which contains half the emitted flux of the galaxy), and I_e is the intensity at the half-light radius R, i.e., $I(R) = I_e$.

In this work, we use the source galaxy's apparent magnitude m_s in some passband in place of the source intensity I_e in that band, as the former is the quantity directly measured in astronomical observations. Converting between the two quantities requires the magnitude zero point, defined as the magnitude that produces 1 count per second on the detector in that band, which varies from telescope to telescope. We define the zero point using Dark Energy Survey observing conditions in the g-band filter, which we describe in Section 5.4.

As with the lens ellipticity (Eq. 5.5-5.6), we replace the corresponding source parameters ($q_{\rm src}$, $\phi_{\rm src}$) with the eccentricity moduli:

$$s_{e1} = \frac{1 - q_{\rm src}}{1 + q_{\rm src}} \cos(2\phi_{\rm src}), \tag{5.12}$$

$$s_{e2} = \frac{1 - q_{\rm lens}}{1 + q_{\rm src}} \sin(2\phi_{\rm src}), \tag{5.13}$$

which describes the ellipticity of the source light profile.

The lensing morphology depends only on the relative difference between the 2D positions of the centers of the lens mass and source light profiles. Therefore, we choose to fix the center of the source light profile to the origin of the (x, y) image coordinate system without any loss of generality. This reduces the total number of parameters needed to describe the source light profile to five: the source apparent magnitude m_s, the effective radius R, the Sersic index n, and the eccentricity moduli (s_{e1}, s_{e2}).

In this study, we make a few additional simplifying assumptions in our lens setup. First, we assume that the lens galaxy's light has been perfectly subtracted from the image to be modeled. The lens light is often subtracted before lens modeling, and there has been promising work done to automate the process of lens light subtraction as a pre-processing step in an automated pipeline (e.g. Hezaveh et al. (2017b)). The alternative would be to include the lens light in the model, with additional parameters. Second, we set the lens redshift to $z_l = 0.5$ and the source redshift to $z_s = 2.0$, because those are the peaks of the projected redshift distributions for galaxy-galaxy strong lenses in forthcoming surveys (Collett et al., 2015). Finally, we assume a flat ΛCDM cosmology with $H_0 = 70\,\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$ and $\Omega_M = 0.3$.

In summary, each strong lens is characterized by 12 parameters: five to describe the lens mass (θ_E, l_{e1}, l_{e2}, x_c, y_c), two to describe the external shear (γ_1, γ_2), and five to describe the source light (m_s, R, n, s_{e1}, s_{e2}). Table 5.4.2 contains further details on the distributions of these parameters.

5.3 Methodology: Posterior Estimation

5.3.1 Overview: Generative Modeling and Posterior Estimation

In traditional Bayesian inference, the posterior $p(\theta|X)$ is given by Bayes' theorem as:

$$p(\theta|X) = \frac{p(X|\theta) * p(\theta)}{p(X)}, \tag{5.14}$$

where X denotes the data vector (e.g., a set of flux values in a pixellated image), θ is the set of model parameters to be inferred, $p(X|\theta)$ denotes the likelihood, $p(\theta)$ is the prior on the parameters before acquiring the data X, $p(X)$ is the evidence (or marginal likelihood), and we are implicitly working in the context of a particular model characterized by the parameters θ. Explicit inference – where $p(X|\theta)$ is approximated by an analytic likelihood function and

computed in the inference process – is the traditional approach to estimating parameter posteriors in Bayesian inference. Examples of explicit likelihood inference methods include Markov Chain Monte Carlo (MCMC) methods, which are now ubiquitous in astronomical and cosmological inference. However, in cases where the likelihood function is unknown or computationally intractable (e.g., in cases where the model contains hundreds or even thousands of parameters), explicit likelihood methods may not be feasible.

To circumvent the challenges inherent in such likelihood calculations, Simulation-Based Inference (SBI) – a.k.a. Implicit Likelihood Inference (ILI) or Likelihood-Free Inference (LFI) – methods have gained popularity in recent years (Cranmer et al. (2019) and references within). Traditional SBI methods include Approximate Bayesian Computation (ABC; Rubin, 1984; Tavaré et al., 1997; Beaumont et al., 2002) and its frequentist analog, Approximate Frequentist Computation (AFC; Brehmer et al., 2018a,b).

ABC starts by simulating realizations of model observables X (synthetic data) that are drawn from points in the model parameter space; such simulations can be performed even in cases where the likelihood function (or the sampling distribution from which it derives) is intractable. The points are then accepted or rejected based on the model observables' similarity to the observed data up to some tolerance ϵ in some specified metric. Through this rejection sampling, the distributions of accepted points in parameter space will approximate the desired posterior distribution while circumventing explicit likelihood calculations.

Traditional SBI methods such as ABC suffer from two drawbacks. First, the inference is *not* amortized – for every new observation, the potentially computationally expensive inference process has to be repeated[4]. Second, there is an unavoidable trade-off between inference quality and sampling efficiency. Accurate inference of the posterior occurs in the limit $\epsilon \to 0$, in which case the rate of acceptance can be so low that the computational expense is prohibitively high even for low-dimensional (and certainly for high-dimensional)

4. This is also true of traditional explicit likelihood methods like MCMC.

parameter spaces.

Recent advancements in deep learning methods and SBI techniques have largely addressed these drawbacks. They have led to the adoption of novel neural network surrogates for modeling the conditional probability densities of the likelihood or posterior distribution in a given inference problem from simulations (see Cranmer et al. (2019) and references within). These *neural density estimators* are more flexible than conventional parametric density estimators and scale better to higher-dimensional parameter spaces. They are also *amortized* – once the upfront computational cost of training the density estimator is paid, the subsequent inference is fast and cheap. Advances in SBI methodology have also led to the development of more efficient learning methods that bypass the inherent dichotomy between inference quality and sampling efficiency in traditional SBI techniques such as ABC and thus enable SBI to better overcome the curse of dimensionality in higher dimensional parameter spaces.

In this work, we use and compare two different amortized approaches that leverage neural networks for posterior estimation – Neural Posterior Estimation (NPE) and Bayesian Neural Networks (BNNs).

The first method – which we refer to as NPE in this paper – is an SBI approach that leverages recent advancements in deep learning techniques to train a neural density estimator to directly learn the posterior distribution of strong lens model parameters. NPE requires three inputs – a mechanistic model (simulator), priors on parameters of the model being inferred, and data (or corresponding summary statistics) which are typically outputs of the simulator. The priors are utilized to sample the parameters and simulate synthetic data that is passed to a neural network (density estimator). The network is trained to learn the relation between simulated data and the underlying parameters and is later deployed on empirical data to obtain the posterior parameter distribution corresponding to that data. NPE avoids explicit likelihood calculations and uses simulations to train the network and to

get the relevant parameters.

The second approach, Bayesian Neural Networks, involves using stochastic neural networks trained using a Bayesian approach (Jospin et al., 2022). Traditional artificial neural networks – when trained on data (lens images in this case) and their corresponding labels (lens model parameter values) – produce the equivalent of maximum likelihood (MLE) or maximum a posteriori (MAP) point estimates for the labels when given an observed data point. Although these so-called deterministic neural network models are easy to deploy, they do not include uncertainty quantification, which is crucial in science applications. BNNs address this flaw in traditional neural networks by replacing the deterministic network weights with probability distributions. By introducing stochasticity into the network, the BNN model produces a probability density estimate instead of a point estimate as its output, thus enabling posterior density estimation and uncertainty quantification.

We note here that while the BNNs in this work are used to model the posterior and as such can technically count as a neural posterior estimation method, to prevent confusion we use neural posterior estimation to refer only to the SBI NPE method, while BNNs will always be referred to explicitly as such.

In the next section, we give detailed technical descriptions of the architecture and implementation of both methods.

5.3.2 Neural Posterior Estimation

Overview

NPE is an SBI approach that uses neural density estimators to directly learn the posterior density for a particular inference problem. So far, there have been several different approaches for this method – including the original method, NPE-A (Papamakarios & Murray, 2018) and two variants, NPE-B (Lueckmann et al., 2018) and NPE-C (Greenberg et al.,

2019)[5]. These methods differ in how they handle active learning, also known as *sequential* estimation. However, this sequential aspect is optional, and we opt not to use it in this work for reasons we discuss in §5.3.2. In the case where no sequential estimation is performed, all NPE methods reduce to the same approach described in NPE-A, which we expound on here.

In NPE, a neural density estimator model – $q_\phi(\theta|X)$, parameterized by a vector ϕ representing the neural network model parameters (e.g., network weights) – is trained on N pairs of training data $\{\theta_i, X_i\}_{i=1}^N$ from the simulator to approximate the posterior density $p(\theta|X)$, i.e. the posterior probability distribution of model parameters θ given some observed data X. The training data is generated by sampling

$$\theta_i \sim p(\theta), \qquad\qquad X_i \sim p(X|\theta_i) \tag{5.15}$$

where $p(\theta)$ is the prior and $p(X|\theta_i)$ is a statistical model representing the simulator's data-generating process given the model parameter θ_i. The model's goodness-of-fit is evaluated by maximizing the average log-likelihood

$$\mathcal{L}(\phi) = \frac{1}{N}\sum_i^N \log q_\phi(\theta_i|X_i) \tag{5.16}$$

with respect to ϕ. By the strong law of large numbers, in the limit $N \to \infty$, Eq. 5.16 converges to the expectation value of the log-likelihood, $\log q_\phi(\theta|X)$, over the true underlying joint probability distribution of the training data $\{\theta_i, X_i\} \sim p(X|\theta)p(\theta)$

5. Here we follow the NPE nomenclature of Mancini et al. (2022), which is consistent with the taxonomy of other SBI literature (e.g., Durkan et al. (2020)).

$$\begin{aligned}\mathcal{L}(\phi) &= \lim_{N\to\infty} \frac{1}{N} \sum_{i}^{N} \log q_\phi(\theta_i|X_i) \\ &= \int\int p(X|\theta)p(\theta) \log q_\phi(\theta_i|X_i) dX d\theta \\ &\equiv \mathbb{E}_{p(X|\theta)p(\theta)}[\log q_\phi(\theta|X)]. \end{aligned} \tag{5.17}$$

This is equivalent to minimizing the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) between the true and model probability distributions

$$\begin{aligned}\mathrm{KL}(p(\theta|X)||q_\phi(\theta|X)) &\equiv \int\int p(\theta|X) \log \frac{p(\theta|X)}{q_\phi(\theta|X)} dX d\theta \\ &= -\mathbb{E}_{p(X|\theta)p(\theta)}[\log q_\phi(\theta|X)] + \text{Const.} \end{aligned} \tag{5.18}$$

with respect to ϕ. In other words, maximizing Eq. 5.16 for $q_\phi(\theta|X)$ decreases its statistical distance from the true posterior underlying $p(\theta|X)$, enabling it to become a better approximation of $p(\theta|X)$.

Therefore, if $q_\phi(\theta|X)$ is a sufficiently flexible mathematical model, then as $N \to \infty$, $q_\phi(\theta|X)$ will converge to the true posterior $p(\theta|X)$ when the average log-likelihood in Eq. 5.16 is maximized. By using the negative average log-likelihood as a loss function, we can train $q_\phi(\theta|X)$ on the training data $\{\theta_i, X_i\}_{i=1}^N$ to approximate the true posterior $p(\theta|X)$.

Masked Autoregressive Flow Density Estimators

NPE works by approximating the posterior density with neural density estimators. In this work, we used Masked Autoregressive Flows (MAF; Papamakarios et al., 2017) as our choice of neural density estimator for their state-of-the-art flexibility in modeling complex density distributions. In this section, we provide an overview of how they work.

MAFs are a fusion of two state-of-the-art families of neural density estimators, namely,

autoregressive models (e.g., Uria et al. (2016)) and normalizing flows (e.g., Rezende & Mohamed (2015)). Autoregressive density estimators make use of the property that any n-dimensional probability density $p(\mathbf{v})$ where $\mathbf{v} \equiv [v_i, ..., v_n]$ can be factorized by the chain rule into a product of 1-dimensional conditional probability distributions

$$p(\mathbf{v}) = \prod_{i=1}^{n} p(v_i|v_1, ..., v_{i-1}) = \prod_{i=1}^{n} p(v_i|\mathbf{v}_{1:i-1}) \tag{5.19}$$

where $\mathbf{v}_{1:i-1}$ is the subset of the data vector containing data points with indices less than i. These conditional probabilities, where $p(v_i)$ is only conditional on data points with indices less than i, are said to be *autoregressive*. The family of autoregressive density estimation methods uses neural networks to model $p(v_i)$ for all n, which can then be combined using Eq.5.19 to get the target density $p(\mathbf{v})$.

By contrast, normalizing flows work by transforming samples from a simple base density (e.g., typically a standard normal distribution), $\mathbf{u} \sim p(\mathbf{u})$ into samples from the target density $\mathbf{v} \sim p(\mathbf{v})$ through an invertible transformation T (*"flow"*)

$$p(\mathbf{v}) = p(\mathbf{u}) \left| \frac{\partial T}{\partial \mathbf{u}} \right|^{-1} \tag{5.20}$$

where $\mathbf{u} = T^{-1}(\mathbf{v})$. $p(\mathbf{v})$ is well-defined under the transformation T if and only if T is invertible and both T and T^{-1} are differentiable (i.e. T has a tractable Jacobian). Enforcing this property ensures that the resulting transformation is both *normalized* (i.e. $\int p(\mathbf{v})d\mathbf{v} = \int p(\mathbf{u})d\mathbf{u} = 1$) and *composable* (i.e. the composition of two transformations $T = T_2 \cdot T_1$ also has a tractable Jacobian). The base density $p(\mathbf{u})$ is chosen such that it is easily evaluated for any $\mathbf{u}$, with the standard normal distribution being a very common choice in the literature, and one that we will continue to use in this work. The family of normalizing flow estimation methods uses neural networks to model the invertible transformations T.

A particular family of autoregressive models - autoregressive models that are parameter-

ized as Gaussian conditional probabilities - can also be re-interpreted as normalizing flows. In these models, the i-th autoregressive conditional can be written as:

$$p(v_i|\mathbf{v}_{1:i-1}) = \mathcal{N}(v_i|\mu_i, (\exp \sigma_i)^2), \tag{5.21}$$

where μ_i and σ_i are the mean and log standard deviation parameters of the Gaussian that v_i is sampled from, and are themselves scalar functions of previous data points $\mathbf{v}_{1:i-1}$ only. In mathematical terms,

$$\mu_i = f_{\mu_i}(\mathbf{v}_{1:i-1}) \tag{5.22}$$

$$\sigma_i = f_{\sigma_i}(\mathbf{v}_{1:i-1}) \tag{5.23}$$

The probability density model in Eq. 5.21 can be used to generate data points recursively. To generate data, the model requires a source of randomness, typically a random number generator of some kind of probability density, with the standard normal distribution being a common choice. Random numbers are first generated by that source of randomness, and are then transformed into data points in data space; for example, we generate a data point v_i by first sampling a random number u_i from a standard normal distribution $u_i \sim \mathcal{N}(0, 1)$, and then transform u_i by shifting and scaling u_i by μ_i and σ_i respectively to get the data point v_i:

$$v_i = u_i \exp \sigma_i + \mu_i. \tag{5.24}$$

Eq. 5.24 can also be interpreted as a transformation T of a vector of random numbers $\mathbf{u}$ to a vector of data points $\mathbf{v} = T(\mathbf{u})$. By construction, T is invertible and has a tractable Jacobian, which implies that the autoregressive model described in Eq. 5.21 is also definitionally a normalizing flow.

By itself, density models of the type described in Eq. 5.24 – that are solely parameterized

by Gaussian distributions – have limited flexibility. However, the insight that Eq. 5.24 also represents a normalizing flow leads to the realization that it is composable; that is, instead of having a single autoregressive model transform the initial simple base density $\mathbf{u}$ into the target density $\mathbf{v}$, we can compose stacks of autoregressive model "blocks", such that samples of the output density of the last model is used as the input of the next model. For example (and following the same notation as Papamakarios et al. (2017) for consistency), consider a model comprised of k autoregressive model blocks, $M_1, ..., M_k$, of the kind described in Eqs. 5.21-5.24. In this model, the block M_1 takes in as input samples from the density $\mathbf{u_1}$, which is the output samples modeled by M_2. The input samples of M_2, $\mathbf{u_2}$, is, in turn, the output samples of M_3, and so on, until we recurse to $\mathbf{u_k}$, which samples from the base density, which for this example we will take to be the standard normal distribution, $\mathbf{u_k} \sim \mathcal{N}(0, 1)$.

This composition greatly increases the flexibility of the density model, as samples from the initial input density – the standard normal distribution $\mathbf{u_k} \sim \mathcal{N}(0, 1)$ – can be progressively transformed into samples from increasingly more complex distributions $\mathbf{u_{<k}}$, which are in turn used as input densities for intermediate model blocks to produce even more complex output densities for the next block in the stack, and so on, until we obtain the final target data $\mathbf{v}$. This stack of models can be trained jointly in an end-to-end fashion to produce a normalizing flow that is far more flexible than its model blocks. The stacking of autoregressive model blocks in this manner is the essential structure of a MAF, and we refer to Papamakarios et al. (2017) for a full treatment of the remaining technical details.

In terms of implementation, MAFs use Masked Autoencoder for Distribution Estimation (MADE; Germain et al., 2015) to encode the autoregressive model blocks M_j described above. For disambiguation, we will use j to index the j-th model block M_j and their associated input data vector $\mathbf{u_j}$, and i to index the i-th data point in that data vector. For a given block M_j, MADEs are neural networks that model the functions $\{f_{\mu_{j,i}}, f_{\sigma_{j,i}}\}$ in Eq. 5.22-5.23 for all i. More specifically, it takes in as input the data output of the previous

block, $\mathbf{u_{j-1}} = [u_{j-1,1}, ..., u_{j-1,i}, ..., u_{j-1,n}]$ and outputs the mean $\mu_{j,i}$ and log standard deviation $\sigma_{j,i}$, which is then used to generate $u_{j,i}$ for all i using Eq. 5.24. The advantage of MADE over other autoregressive model architectures is that it avoids the need to compute the recursion in Eqs. 5.21-5.24 sequentially. By judiciously applying binary masks to some of the weight matrices in the MADE such that the autoregressive property in Eqs. 5.21-5.24 are enforced, the function $\{f_{\mu_{j,i}}, f_{\sigma_{j,i}}\}$ for all i can be calculated efficiently in a single forward pass. For a full treatment of all the remaining technical details, we refer to Germain et al. (2015).

Data Compression with Summary Statistics

For SBI problems where the number of dimensions of the data space X is relatively small ($\lesssim 100$), the data vectors can be directly used as inputs to the NPE itself. However, for situations where the dimensionality of the data is large ($\gtrsim 1000$), such as in this work or similar inference problems involving imaging data, it can become prohibitively expensive computationally to learn the sampling distribution of the data as a function of the parameters due to the curse of dimensionality. Traditional inference methods circumvent this problem by constructing summary statistics to compress the dimensionality of the data down to a smaller number of summary statistics. The inference process, which may involve comparing the summary statistics of the observed data to the simulated data, or performing density estimation of the model parameters over the summary statistics, becomes far more computationally tractable. This approach has also been extended to SBI methods and has become a common practice when performing inference on high-dimensional data (e.g. Cranmer et al. (2019); Alsing et al. (2019a); Charnock et al. (2018); Rodrigues et al. (2021)).

The choice of summary statistics can have a big impact on the inference process, as a well-chosen set of summary statistics will preserve the largest amount of information from the original data, while an ill-chosen set of summary statistics can result in significant in-

formation loss that impacts the quality of the inference. What is considered a well-chosen set of summary statistics is also highly dependent on the dataset itself. As such, summary statistics are usually prescribed by domain experts, guided by their intuition and knowledge of the subject. In recent years, however, there have been efforts to systematize the construction and evaluation of summary statistics, especially in relation to its use in SBI (see e.g. Blum et al. (2013); Prangle (2015) and references within).

In machine learning, *features* or *representations* are often synonymous with summary statistics, and great advancements have been made in developing methods to learn or extract summary statistics/features in an automated fashion in recent years (see Bengio et al. (2014) and references within). Convolutional neural networks (CNN) are a common type of neural network architecture used for this purpose, and it remains state-of-the-art in numerous machine learning applications, including, saliently, computer vision problems (Li et al., 2022).

In this work, we leverage these machine learning advancements by using a CNN to first compress the output of our simulators (lens imaging data) into a lower dimensional vector of summary statistics, which are then passed on as inputs to the NPE. The parameters of both the CNN and the NPE are then learned together when trained on the training data.

NPE implementation details

In this section, we describe what model architectures we use in this work, and what hyperparameters were chosen for those models.

We used the NPE algorithm (see §5.3.2) implemented in `sbi`[6] (Tejero-Cantero et al., 2020) to train an MAF neural density estimator (see §5.3.2) to infer the lens parameter posteriors from simulated single-band lens images. The simulated lens images are first passed through a custom CNN which compresses the higher dimensional image data into informative

6. https://sbi-dev.github.io/sbi/

summaries statistics (see §5.3.2). This CNN, also known as an embedding network, comprises six convolutional layers (Conv2d in Table 5.3.1), with each convolutional layer followed by a batch normalization layer (BatchNorm2d in Table 5.3.1). We perform a max pooling operation (Maxpool2d) after every other convolutional layer, specifically after the 2nd, 4th, and 6th convolutional layers. The output from the convolutional layers is flattened before being passed to the fully connected (Linear) layer. In Table 5.3.1, the 'Output shape' column shows the dimensionality of that layer after passing the input through it. In the 'Parameters' column, 'k' denotes the convolutional kernel size (size of the convolution window in pixels), and 's' denotes the stride for the convolutional operation (how many pixels the kernel slides in each step).

The embedding network acts as a summary statistics finder - it compresses high-dimensional image data into a lower-dimensional vector of summary statistics that is then used in the training of the neural density estimator. In all of our experiments, we output $4N$ summaries, where N is the number of model parameters that we are aiming to infer. Through trial-and-error experimentation, we found that the performance of the model did not demonstrate improvement when the number of summaries per inferred parameter was increased beyond 4 summaries per parameter for the data used in this work. In later sections, we apply the method to models with 1, 5, and 12 parameters, for which we used 4, 20, and 48 summaries respectively.

The data summaries from the embedding network are then used as the inputs to the MAF density estimator. The embedding network is trained along with the density estimator and functions as an automatic summary statistic extractor.

The MAF implementation in `sbi` allows for a variety of hyperparameters to be adjusted, with the most important two being the number of hidden units (nodes) in each hidden layer of each MADE block in the MAF, and the number of flow transformations in the MAF, as these hyperparameters change the structure of the MAF architecture. We manually tuned

both of these hyperparameters, by performing a course grid search, increasing the number of hidden units up to 1000, and the number of transformations up to 50. Each model was trained on the same data set, and evaluated on the same test set. We evaluated each model on its posterior coverage calibration until we noticed no further visual improvement in their empirical coverage (see §5.5.1 for details on that diagnostic). In the end, we settled on the following hyperparameter choices for each model: for the 1-parameter lens model, we used 100 hidden units and 10 transformations. For the 5-parameter model, we used 120 hidden units and 20 transformations. For the 12-parameter model, we used 400 hidden units and 20 transformations.

For the remaining hyperparameters, we opted to leave them at their default values as modifying them did not appear to result in any significant gains. The `Adam` optimizer (Kingma & Ba, 2017) was used, with the learning rate and batch size set to 0.0005 and 50 respectively. The validation fraction was set to 0.1 of the training set. The training was automatically terminated if there was no improvement in training loss on the validation set after 20 epochs[7].

Why not use Sequential NPE (SNPE)?

As mentioned in §5.3.2, NPE methods have an optional active learning component, also known as *sequential* estimation. This is an optional step and one that we choose not to adopt in this work. In this section, we briefly describe what sequential estimation is, the tradeoffs involved when using it, and why we ultimately chose not to use it in this work.

The *sequential* aspect of the inference algorithm involves training a density estimator iteratively over several rounds to learn a *proposal prior* distribution $\tilde{p}(\theta)$. This proposal prior distribution is then used to simulate the N pairs of training data $\{\theta_i, X_i\}_{i=1}^{N}$ used to train the target posterior density estimator.

7. An epoch is defined as a single complete pass of the training dataset through the training process, culminating in an update of the neural network's internal parameters.

Table 5.3.1 The architecture of the embedding network used in the NPE to compress the image data into summary statistics. The first column lists the layer type, the second lists the dimensionality of the output from that layer, and the third column lists the parameters of that layer, where 'k' denotes the kernel size and 's' denotes the stride. The final layer outputs data summaries. See §5.3.2 for more details.

Layer	Output shape	Parameters
Conv2d	[-1, 8, 32, 32]	k=3, s=1
BatchNorm2d	[-1, 8, 32, 32]	k=3, s=1
Conv2d	[-1, 16, 32, 32]	k=3, s=1
BatchNorm2d	[-1, 16, 32, 32]	k=3, s=1
MaxPool2d	[-1, 16, 16, 16]	k=2, s=2
Conv2d	[-1, 32, 16, 16]	k=3, s=1
BatchNorm2d	[-1, 32, 16, 16]	k=3, s=1
Conv2d	[-1, 32, 16, 16]	k=3, s=1
BatchNorm2d	[-1, 32, 16, 16]	k=3, s=1
MaxPool2d	[-1, 32, 8, 8]	k=2, s=2
Conv2d	[-1, 64, 8, 8]	k=3, s=1
BatchNorm2d	[-1, 64, 8, 8]	k=3, s=1
Conv2d	[-1, 128, 8, 8]	k=3, s=1
BatchNorm2d	[-1, 128, 8, 8]	k=3, s=1
MaxPool2d	[-1, 128, 4, 4]	k=2, s=2
Flatten	[-1, 2048]	-
Linear	[-1, 4*N]	-

The motivation behind this sequential step is as follows: if we are interested in the posterior distribution corresponding to a given observation X_0, $p(\theta|X = X_0)$, the training data that is most informative for training $q_\phi(\theta|X = X_0)$ to accurately approximate $p(\theta|X_0)$ should be drawn from a distribution that closely matches $p(\theta|X_0)$. The prior $p(\theta)$ is typically much wider than $p(\theta|X_0)$, so the vast majority of the simulated training data generated from the prior $p(\theta)$ is uninformative for training $q_\phi(\theta|X_0)$ to approximate $p(\theta|X_0)$. Therefore, by first learning a proposal prior $\tilde{p}(\theta)$ that closely matches $p(\theta|X_0)$, we require fewer simulations sampled from $\tilde{p}(\theta)$ to accurately train $q_\phi(\theta|X_0)$ to approximate $p(\theta|X_0)$.

While this sequential step improves the sampling efficiency in principle, it has two major drawbacks:

1. The method is no longer amortized. For every new data point X_i for which we want to infer the posterior $p(\theta|X_i)$, a new proposal prior has to be first learned for each datum. This additional training cost for each new datum may offset any improvements in sampling efficiency in cases where the objective is to rapidly infer the posteriors for a great number of data points (such as in this work).

2. Sequential methods have additional failure modes and complications because there is no guarantee that sequential rounds of training will result in a proposal prior that is appropriate for training a target posterior. Examples of such failure modes include posterior leakage, which can cause inference to be prohibitively expensive over multiple rounds (Deistler et al., 2022), as well as overconfident uncertainty predictions (Hermans et al., 2022).

SNPE appears to work best in cases where the number of data points that are evaluated *separately* is fairly small. However, it comes with several trade-offs, including the loss of amortization as well as potential failure modes. As such, we do not think SNPE is suitable for our use case, as our objective in this work is to develop a pipeline to model a large number

of lens systems as efficiently as possible. This is consistent with the recommendations by Lueckmann et al. (2021) in their benchmarking work. In light of these trade-offs, we opt not to use SNPE for this work.

5.3.3 Bayesian Neural Networks

To compare the results of our SBI NPE method with another deep learning-based method, we train a BNN (Neal, 1996) on the same simulated datasets. With BNNs, the deterministic weights of a standard CNN network are replaced by probability distributions, which can be used subsequently to provide a measure of how (un)certain a model is in its predictions. BNNs have already been successfully used in astrophysics and cosmology– e.g., for the classification of blazars (Butter et al., 2022), calibration of gamma-ray bursts as cosmological distance indicators using supernovae type Ia (Escamilla-Rivera et al., 2022), and galaxy morphology measurements (Walmsley et al., 2022; Tanoglidis et al., 2022), among others.

The objective of training a BNN is to find the posterior distribution $p(\phi|\mathcal{D})$ of the weights ϕ, given the training datasets $\mathcal{D} = (X, \theta)$, where X are the inputs and θ are the corresponding labels. Since inference of the BNN posterior is a very difficult task that includes calculation of an integral over all model weights, approximate methods are utilized. Here we use a common approach called variational inference to model the posterior using a simple variational distribution $q(\phi|\zeta)$, which we take to be a Gaussian (i.e., ζ consists of the mean and standard deviation of the Gaussian distribution), and fit the distribution's parameters ζ to be as close as possible to the true posterior $p(\phi|\mathcal{D})$ (Graves, 2011b; Shridhar et al., 2019). This is done by minimizing the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) between the true and variational posterior probability distributions:

$$\mathrm{KL}(q(\phi|\zeta)||p(\phi|\mathcal{D})) \equiv \int q(\phi|\zeta) \log \frac{q(\phi|\zeta)}{p(\phi|\mathcal{D})} d\phi. \tag{5.25}$$

Table 5.3.2 The Bayesian Neural Network architecture. The first column lists the layer type, the second lists the dimensionality of the output from that layer, and the third column lists the parameters of that layer, where 'k' denotes the kernel size and 's' denotes the stride. See §5.3.3 for more details.

Layer	Output shape	Parameters
Conv2dFlipout	[-1, 16, 32, 32]	k=3, s=1
MaxPool2d	[-1, 16, 16, 16]	k=2, s=2
Conv2dFlipout	[-1, 32, 16, 16]	k=3, s=1
Conv2dFlipout	[-1, 32, 16, 16]	k=3, s=1
MaxPool2d	[-1, 32, 8, 8]	k=2, s=2
Conv2dFlipout	[-1, 48, 8, 8]	k=3, s=1
Conv2dFlipout	[-1, 48, 8, 8]	k=3, s=1
MaxPool2d	[-1, 48, 4, 4]	k=2, s=2
Conv2dFlipout	[-1, 64, 4, 4]	k=3, s=1
Conv2dFlipout	[-1, 64, 4, 4]	k=3, s=1
Conv2dFlipout	[-1, 64, 4, 4]	k=3, s=1
MaxPool2d	[-1, 64, 2, 2]	k=2, s=2
Flatten	[-1, 256]	-
DenseFlipout	[-1, 2048]	-
DenseFlipout	[-1, 512]	-
DenseFlipout	[-1, 64]	-
Dense	[-1, 90]	-
MNTriL	[-1, N]	-

This gives us the final objective of BNN training, i.e., minimization of the Evidence Lower Bound (ELBO) loss that comprises the negative log-likelihood loss and the KL divergence:

$$\mathcal{L}(\mathcal{D}, \zeta) = \mathbb{E}_{q(\phi|\zeta)} \left[\log q(\phi|\zeta) - \log p(\phi) p(\mathcal{D}|\phi) \right]. \tag{5.26}$$

The outputs of the BNN are multivariate normal distributions, i.e., BNN outputs both mean and variance for each of the output parameters. BNNs are capable of capturing both *aleatoric* (statistical - related to the intrinsic randomness of the data-generating or measurement process) and *epistemic* (systematic - related to the model; reducible with more data or better model) uncertainties (Depeweg et al., 2017; Kendall & Gal, 2017a). However, the utility of BNNs is generally limited by the distributions used to model the neural network weights, which in practice are typically Gaussian or Gaussian mixture models. If the true posterior distribution follows a more complicated distribution, BNNs may not be flexible enough to approximate the correct distribution.

The BNN architecture details are provided in Table 5.3.2. Conv2dFlipout denotes the 2D convolution layer with Flipout gradient estimator (Wen et al., 2018) and DenseFlipout denotes the Dense layer with Flipout gradient estimator, which is used for decorrelating the gradients within a mini-batch by implicitly sampling pseudo-independent weight perturbations for each example. MNTriL denotes the output multivariate normal distribution. As with Table 5.3.1, the 'Output shape' column shows the dimensionality of that layer after passing the input through it. In the 'Parameters' column, 'k' denotes the convolutional kernel size (size of the convolution window in pixels), and 's' denotes the stride for the convolutional operation (how many pixels the kernel slides in each step).

We trained the network for 500 epochs as we found that increasing the number of training epochs after did not significantly improve the training loss. and use the Adadelta optimizer (Zeiler, 2012) with an initial learning rate of 0.08. The Adadelta optimizer has an adaptive learning rate and the results are insensitive to the choice of initial learning rate.

5.4 Data: Simulations of Strong Lenses

To develop our SBI NPE and BNN models, we consider 3 levels of model complexity: 1) proof of concept 1-parameter model, where we only vary the Einstein radius, θ_{E}; 2) 5-parameter lens-only model, where the external shear and source galaxy light parameters are fixed and only the lens mass parameters are varied; 3) the full 12-parameter model, where lens mass, source light, and environmental shear parameters are all varied. As we will describe in Sections 5.6 and 5.7, we have found that for more model parameters, a larger training set is needed to reach good model performance (as quantified below). We use the same training and test data for both NPE and BNN models. For the BNN models, we generate an additional validation dataset for model training. For the NPE model, 10% of the training data is used as a validation dataset, as internally set by the `sbi`package. Table 5.4.1 shows the sizes of all datasets used in our experiments.

For the 1- and 5-parameter models, we model the light from the source galaxy as a fixed circular Sersic profile with Sersic index $n = 4$, scale radius $R = 0.35''$, and magnitude $m_s = 21$, which are typical values observed in cosmic surveys (see, e.g., Bolton et al. (2008) and Collett et al. (2015)). We do not include external shear in these two experiments. Furthermore, for the simple 1-parameter model, we vary only the Einstein radii, keeping the lens mass ellipticity fixed at zero and with a fixed lens-source offset of $(x_c, y_c) = (0.2'', 0.0'')$.

Table 5.4.1 Dataset sizes used in different experiments. The full details of each dataset are described in Section 5.4.

Model	Training	BNN Validation	NPE Validation	Test
1-parameter	$200,000$	$50,000$	$20,000$	1000
5-parameter	$400,000$	$100,000$	$40,000$	1000
12-parameter	$800,000$	$200,000$	$80,000$	1000

We use `deeplenstronomy`[8] (Morgan et al., 2021), which is built on `lenstronomy`[9], to simulate strong lensing systems for training and testing (Birrer & Amara, 2018; Birrer et al., 2021). These packages together provide state-of-the-art simulations of lensing systems, including high-fidelity ray-tracing of light from source objects around lenses, emulation of cosmic survey observations, and the generation of sample distributions.

To model modern cosmic survey experiments, we emulate empirical DES Data Release 1 (DR1) survey conditions following Abbott et al. (2018b). The Dark Energy Camera (DECam) has a pixel scale of 0.263 arcsec/pixel, CCD gain of 6.083 e^-/count, and read noise of 7.0 e^-. We draw on empirical distributions of sky brightness, atmospheric seeing, and the effective number of exposures from Figures 4, 5, and 6 of Abbott et al. (2018b) respectively. We simulate g-band images with a magnitude zero point of 26.58.

In generating the simulation samples for training and testing, we draw independently from uniform prior distributions of each model parameter between minimum and maximum values based loosely on the ranges we expect to encounter in real data (see Table 5.4.2). We restrict the test-set prior range to cover only a portion of the training-set range for each parameter, to ensure that the performance of the algorithms near the limits of the test-set range is not degraded due to a lack of examples in the training set.

We also generate several out-of-distribution (OOD) test sets to evaluate the performance of the models on test sets for which the parameter distributions are significantly different from the prior distributions used to train the models. This OOD test is important, as the distribution of lens parameters for lenses in future surveys is not well-known and may diverge from theoretical predictions. As such, we want our models' performance to be robust to test sets with different priors. We provide those priors in Table 5.4.2 as well.

Figures 5.4.1 and 5.4.2 show 20 randomly selected lenses from the 5- and 12-parameter

8. https://github.com/deepskies/deeplenstronomy

9. https://github.com/sibirrer/lenstronomy

test sets, respectively. All the images used for training and testing are single-band images that are 32x32 pixels in size.

The data we used for this work are published on Zenodo[10]. We include the `deeplenstronomy` notebooks used to generate our data, as well as the notebooks and source code for training our NPE and BNN models, on our GitHub[11] page.

10. To be published when the paper is accepted

11. https://github.com/deepskies/DeepSLEEP

Table 5.4.2 Parameter distributions used to generate training and test sets for the 1, 5, and 12-parameter models. The parameters relevant to each model are indicated under the 'Applicable Models' column. Where there are common parameters across the 1, 5, or 12-parameter models (e.g., θ_E), we use the same distributions for those parameters. The full list of model parameters is as follows. For the lens mass, we use SIE mass profile, Einstein radius ($\theta_{\rm E}$), ellipticity components (le_1,le_2), and lens-source offset (x_c, y_c). For the lens environment, we use the 2 components of the external shear (γ_1, γ_2). For the source light we use a Sersic profile with apparent magnitude (m_s), half-light radius (R), Sersic index (n), and ellipticity components (se_1,se_2). Uniform distributions are denoted by $\mathcal{U}(\min, \max)$, Normal/Gaussian distributions are denoted by $\mathcal{N}(\text{mean}, \text{standard deviation})$, and lognormal distributions are denoted by $\mathcal{N}_{\log}(\text{mean}, \text{standard deviation})$.

Parameter	Applicable Models	Training Set Priors	Test Set Priors	OOD Priors Test Set 1	OOD Priors Test Set 2	OOD Priors Test Set 3
			Lens Mass Parameters			
$\theta_{\rm E}$(")	1,5,12	$\mathcal{U}(0.3, 4.0)$	$\mathcal{U}(0.5, 3.0)$	$\mathcal{N}(2.0, 0.2)$	$\mathcal{N}(1.0, 0.2)$	$\mathcal{N}(1.0, 0.2)$
le_2	5,12	$\mathcal{U}(-0.8, 0.8)$	$\mathcal{U}(-0.2, 0.2)$	$\mathcal{N}(-0.2, 0.2)$	$\mathcal{N}(-0.1, 0.2)$	$\mathcal{N}(-0.2, 0.2)$
le_2	5,12	$\mathcal{U}(-0.8, 0.8)$	$\mathcal{U}[-0.2, 0.2]$	$\mathcal{N}(-0.2, 0.2)$	$\mathcal{N}(-0.1, 0.2)$	$\mathcal{N}(0.2, 0.2)$
$x_{\rm c}$ (")	5,12	$\mathcal{U}(-2, 2)$	$\mathcal{U}(-1, 1)$	$\mathcal{N}(0.2, 0.2)$	$\mathcal{N}(-0.1, 0.2)$	$\mathcal{N}(-0.2, 0.2)$
$y_{\rm c}$ (")	5,12	$\mathcal{U}(-2, 2)$	$\mathcal{U}(-1, 1)$	$\mathcal{N}(-0.2, 0.2)$	$\mathcal{N}(-0.1, 0.2)$	$\mathcal{N}(0.2, 0.2)$
			Lens Environment Parameters			
γ_1	12	$\mathcal{U}(-0.8, 0.8)$	$\mathcal{U}(-0.05, 0.05)$	$\mathcal{N}_{\log}(-3, 1)$	$\mathcal{N}(0.00, 0.05)$	$\mathcal{N}(0.00, 0.01)$
γ_2	12	$\mathcal{U}(-0.8, 0.8)$	$\mathcal{U}(-0.05, 0.05)$	$\mathcal{N}_{\log}(-3, 1)$	$\mathcal{N}(0.00, 0.05)$	$\mathcal{N}(0.00, 0.01)$
			Source Light Parameters			
m_s	12	$\mathcal{U}(18, 25)$	$\mathcal{U}(19, 24)$	$\mathcal{N}(22, 1)$	$\mathcal{N}(21, 1)$	$\mathcal{N}(21.0, 0.5)$
R (")	12	$\mathcal{U}(0.1, 3.0)$	$\mathcal{U}(0.5, 1.0)$	$\mathcal{N}(0.7, 0.1)$	$\mathcal{N}(1.0, 0.2)$	$\mathcal{N}(0.8, 0.1)$
n	12	$\mathcal{U}(0.5, 8.0)$	$\mathcal{U}(2, 4)$	$\mathcal{N}(4, 1)$	$\mathcal{N}(3.0, 0.5)$	$\mathcal{N}(5.0, 0.1)$
se_1	12	$\mathcal{U}(-0.8, 0.8)$	$\mathcal{U}(-0.2, 0.2)$	$\mathcal{N}(-0.2, 0.2)$	$\mathcal{N}(-0.1, 0.2)$	$\mathcal{N}(0.1, 0.1)$
se_2	12	$\mathcal{U}(-0.8, 0.8)$	$\mathcal{U}(-0.2, 0.2)$	$\mathcal{N}(-0.2, 0.2)$	$\mathcal{N}(0.1, 0.2)$	$\mathcal{N}(-0.1, 0.1)$

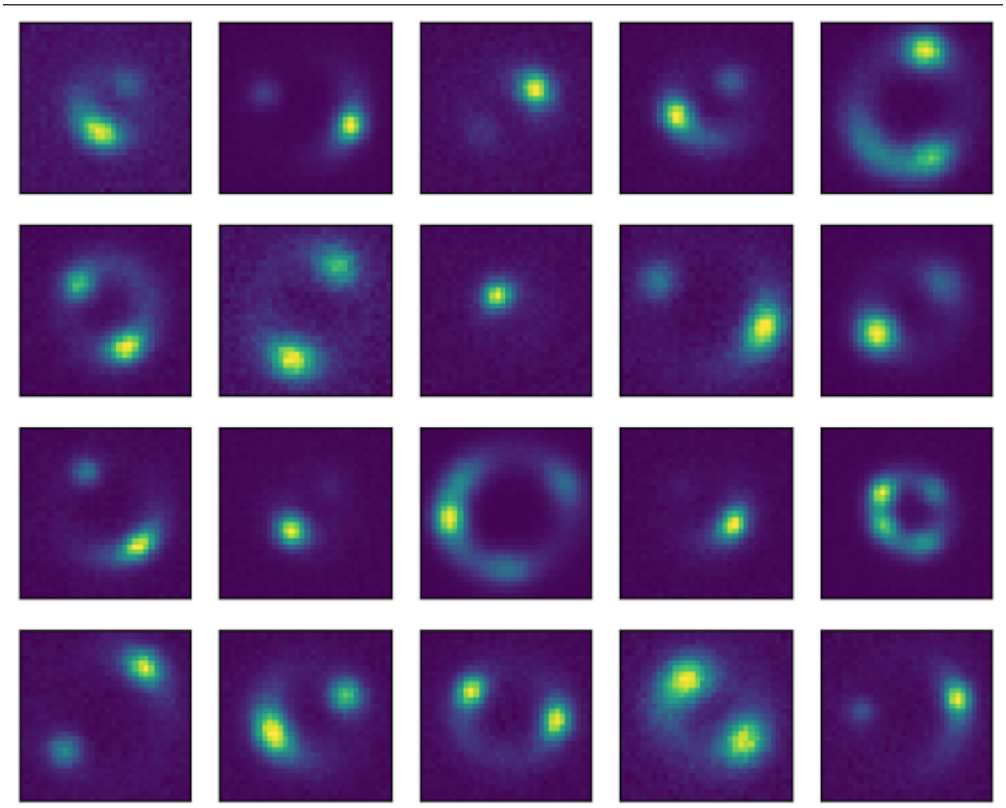

Figure 5.4.1 20 simulated g-band lensed images in the test set with the lens light perfectly subtracted in the 5-parameter model.

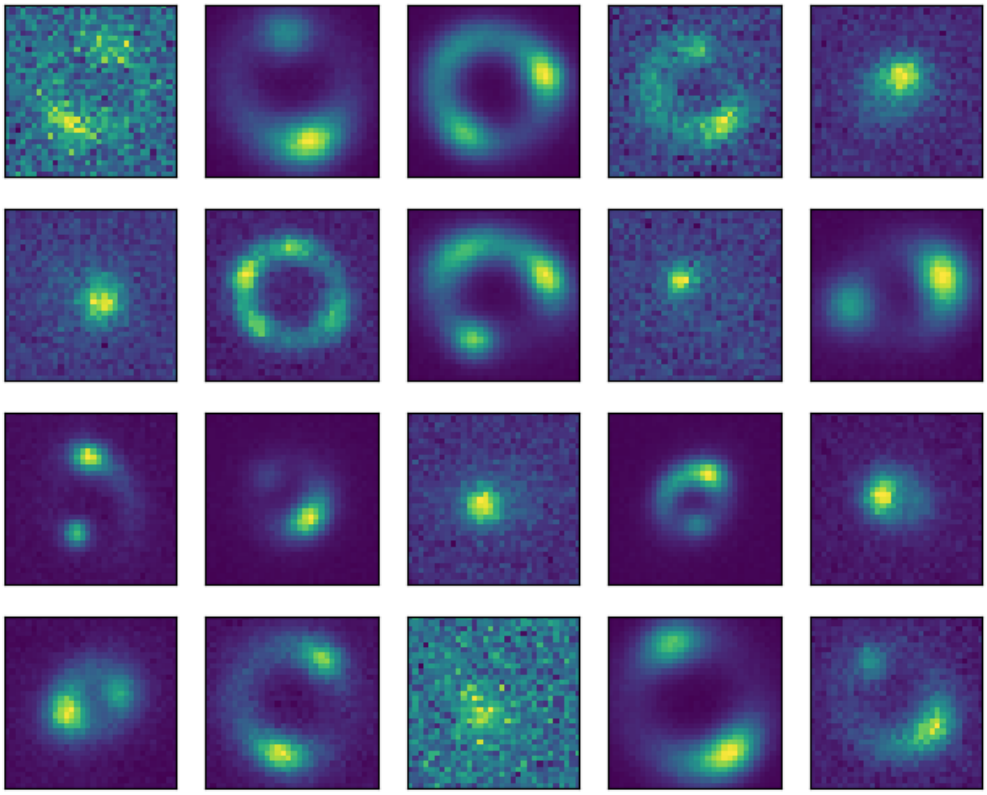

Figure 5.4.2 20 simulated g-band lensed images in the test set with the lens light perfectly subtracted in the 12-parameter model.

5.5 Model Evaluation Diagnostics

For each model that we evaluated in this work, we employed a suite of six performance diagnostics. In this section, we describe the purpose and methodology of each diagnostic, as well as how they are implemented.

5.5.1 Posterior Coverage

For a model to be trustworthy, it must produce uncertainty estimates that are credible and well-calibrated. The posterior coverage measures how well-calibrated the predicted uncertainties of the posterior models are. For a posterior distribution to be well-calibrated, it must contain the true value p% of the time in p% of the posterior probability volume. For example, a posterior distribution with 68% (1σ equivalent) confidence intervals should contain the true value within those confidence intervals 68% of the time. Moreover, these uncertainties will have to be averaged across a large ensemble of lenses in a test set, as the estimated posteriors of a single lens image alone may be not indicative of the average performance of the model.

We present the posterior coverage of our models in the form of plots which are calculated using the following steps:

1. Create a test set of 1000 lens images as described in section 5.4.

2. For each lens in the test set, use the trained posterior estimator to obtain an estimate of the posterior distribution of the model parameters for that particular lens by sampling the posterior estimator model.

3. From each posterior distribution, calculate parameter confidence intervals for that lens.

4. For each confidence interval, calculate the fraction of the lenses in the test set whose true parameter values fall within that confidence interval range.

5. For each parameter, plot the fraction of the lenses whose true parameter values fall within the confidence interval, as a function of the confidence interval.

If the model's posterior uncertainties are perfectly calibrated for the entire test ensemble, we expect to see the posterior coverage follow the 1-1 line (refer to Fig. ADD FIGURE). Otherwise, we interpret the plots as follows: if the posterior coverage falls above the 1-1 line, it is **underconfident**, as it overestimates the uncertainties (i.e. a larger fraction of the test set's true values fall within the given confidence interval than predicted by the model). Conversely, if it falls below the 1-1 line, it is **overconfident**, as it underestimates the true uncertainties.

In addition to calculating posterior coverages for individual parameters, we also calculate the distance metric from Wagner-Carena et al. (2021), defined as:

$$d(\boldsymbol{\theta}_i) = (\boldsymbol{\theta_i} - \boldsymbol{\mu_{\theta_i}}).\Sigma_{\text{data}}.(\boldsymbol{\theta_i} - \boldsymbol{\mu_{\theta_i}})^T \tag{5.27}$$

where $\boldsymbol{\mu_{\theta_i}}$ is the vector of the means of the i-th lens' posterior parameters and Σ_{data} is the empirical covariance matrix for the training set parameters. $\boldsymbol{\theta_i}$ is a vector of parameter values of a posterior sample for the i-th lens. The distance metric $d(\boldsymbol{\theta}_i)$ therefore combines a vector of parameter values (e.g. from a posterior sample) into a single objective function that takes into account the covariance between different parameters.

5.5.2 Scatter Plot of Residuals

To characterize the statistical performance of the best-fit values over an ensemble of test images, we use the scatter in the residuals as a summary statistic. We calculate it as follows: for each lens in the test set, we use our model/s to infer the lens parameter posteriors. We then subtract the best-fit (maximum likelihood) parameter values from the true value of the

lens parameters for each lens. We then plot the difference for all lenses in the test set as a multidimensional scatter (sometimes called a "corner" or "triangle") plot. In this work, we opt to show 68% and 95% contours instead of individual data points to reduce visual clutter (e.g. Figure 5.6.6).

Superficially, these plots look similar to the parameter covariance plots (also known as "corner" or "triangle" plots e.g. Bocquet & Carter (2016); Foreman-Mackey (2016); Lewis (2019)) of an MCMC sampling analysis, in that it displays scatter plots for every pair of parameters and a histogram/probability density of each parameter in some parameter space, overlaid on a grid. The difference is that in an MCMC corner plot, each scatter point represents a posterior sample from the MCMC chain. However, in the plots described here, each scatter point represents the difference between the model's best-fit estimate and the ground truth values of the parameters of one image in the test set.

The scatter characterizes the average best-fit point estimate of the model across the entire test ensemble and can be thought of as a multidimensional variant of the true-vs-predicted scatter plot visualization diagnostic that is common in machine learning literature. If the method is completely accurate (i.e. it exactly predicts the true value for every single lens), we expect a single point at (0,0). If the method produces systematically biased best-fit point estimates, we expect the scatter to shift away from the origin and if the method is not precise, we expect a large scatter in the distribution.

We opt to present this diagnostic plot over the standard true-vs-predicted scatter plot for our multidimensional data for two reasons: First, if the models are systematically biased in a way that is correlated between any pair of parameters, we will be able to observe highly skewed ("degenerate") contour plots for those pair/s of parameters. It would not be possible to visualize correlated biases in true-vs-predicted scatter plots. Second, when plotting true-vs-predicted scatter plots for a large number of test images with their associated uncertainties (in our case, 1000 data points), the visualization becomes cluttered and it is

difficult to discern structure in the data.

5.5.3 Simulation-based Calibration (SBC)

Simulation-based Calibration (SBC) (Talts et al., 2018) tests are a general method of diagnosing and validating Bayesian inference methods based on the theoretical framework of Cook et al. (2006). The general SBC procedure is as follows:

1. From the prior, we randomly generate N independent parameter samples θ_n.
2. We simulate observations of each sample, X_n (i.e. lens image)
3. From our trained SBI posterior model, we obtain M independent randomly posterior samples $p_m(\theta_n|X_n)$ from each of the N posterior models.
4. For each posterior model, count how many samples of $p_m(\theta_n|X_n)$ fall below the corresponding "true" parameter value θ_n. This number is the **rank**.

The main principle behind SBC is that if the trained posterior model is well-calibrated, the SBC ranks should be uniformly distributed between 0 and N. The reason for this is that samples from each of the randomly chosen posteriors should be distributed according to the prior. This corresponds to the claim that "the $p\%$ confidence interval of this posterior model should contain the simulated value in $p\%$ of simulations for all values of x.

Histograms of the SBC rank distributions therefore serve as a diagnostic for checking if the posterior model is correctly computing the posterior. As Talts et al. (2018) explains, deviations from uniformity in SBC rank histograms indicate how the computed posteriors are systemically biased. Specifically, if the histogram is:

- **left-skewed**: The posterior model is biased high compared to the true posterior and systematically overestimates the posterior mean.

- **right-skewed**: the posterior model is biased low compared to the true posterior and systematically underestimates the posterior mean.

- **$\cup$-shaped**: On average, the posterior model is narrower than the true posterior and systematically underestimates the uncertainty.

- **$\cap$-shaped**: On average, the model posterior is wider than the posterior model and systematically overestimates the uncertainty.

Aside from the visual indications of bias listed above, we also plot the 99% confidence interval of a uniform distribution given the number of samples provided. For a uniform distribution, we expect 1 out of 100 histogram bars to lie outside the confidence interval.

5.5.4 Out-of-Distribution (OOD) test set

It is often assumed in deep learning applications that the training set data used to train deep learning algorithms have underlying model parameter distributions that are similar to that of the real-world data that the algorithms will be applied to. However, this assumption is usually not true, especially in scientific applications where the true physical distribution of model parameters is often not well understood. In these cases, the practical value of a deep learning model comes from whether its performance is generalizable – that is, whether it can still produce accurate and precise results when applied to data that are differently distributed from the training set data. Deep learning algorithms that are not robust to out-of-distribution data might be simply learning idiosyncratic or spurious features in the training set.

Therefore, it is important to test how robust the algorithm performance is to test set data with different parameter distributions from the training set. In this work, we test the performance using three different test sets with narrow Gaussian distributions with means

centered on different values shown in Table 5.4.2. We evaluate the results with posterior coverage plots (§5.5.1) and scatter plots of the best-fit values (§5.5.2).

5.5.5 Multiple-seed initialization

There is inherent stochasticity in the way that deep learning algorithms are trained because the pre-trained network weights in the architecture are randomly initialized with different random number generator seeds. Multiple instantiations of the same deep learning architecture using different seeds, when independently trained on the same training data and applied to the same test data, may produce different results. We have observed that sporadically, some model instantiations perform significantly better or worse than others. This phenomenon has also been observed in other deep learning applications (Picard, 2021), and although Picard (2021) found that even though the effects of the choice of random seed initialization does not generally produce a large variance in the performance between models, it is not difficult to fortuitously pick outlier seeds that perform significantly different from the average. The trial-and-error aspect of model architecture and hyperparameter selection may represent an inadvertent search for well-performing seeds.

While the use of outlier seeds does not mean the results of those models are invalid, the performance of those models may not be representative of the typical performance of that particular deep learning architecture, making it challenging to determine the true performance of a model. It is therefore important to be able to diagnose the stability and variance of model performance across multiple seeds. Therefore, we train 3 different seed instantiations for every combination of deep learning architecture (and present results and diagnostic plots for all three seeds) and training set data in this work to identify outlier models.

5.6 Results

In this section, we present the results of NPE and BNN on simulated data generated by three models of varying complexity: a simple, one-parameter proof-of-concept model (Einstein radius), a five-parameter fixed-source model, and a full twelve-parameter model that includes both source and lens parameters. We employed an Nvidia RTX A5000 GPU for training purposes. For the 12-parameter model, both NPE and BNN models took about 6 hours to train, and the other models took less. In comparison, a conventional MCMC fit would take around the order of an hour per lens system. To assess the performance of each model, we utilized the diagnostics outlined in §5.5.

5.6.1 1-parameter Model

We use the one-parameter model to illustrate the diagnostics used to evaluate the effectiveness of our models. In this model, as described in Sections 5.3 and 5.4, we vary the Einstein radius of the lens while leaving all other parameters fixed. We trained the models on 200,000 images in which the Einstein radius ranged from 0.3 to 4.0". We used 100 hidden features and 10 transformations for the MAF NPE estimator. Figure 5.6.1 shows the posterior distribution of θ_E inferred by both the NPE and BNN model for a single lensed image in the test set. We see that in this simple example, both NPE and BNN models are extremely precise, predicting the true θ_E value with $\sim 10^{-3}$ arcsec and $\sim 10^{-2}$ arcsec 68% uncertainties respectively. However, we can see that the NPE method is both more accurate and $\sim$ 10 times more precise than the BNN model.

However, the performance of our models on singular (or even several) lensed images may not be indicative of their performance in general and hence can be misleading. A more representative metric would be their performance across an ensemble of test lenses. We quantify ensemble performance in two ways. First, we create posterior coverage plots according to §5.5.1. The posterior coverage plots show if the predicted *uncertainty confidence*

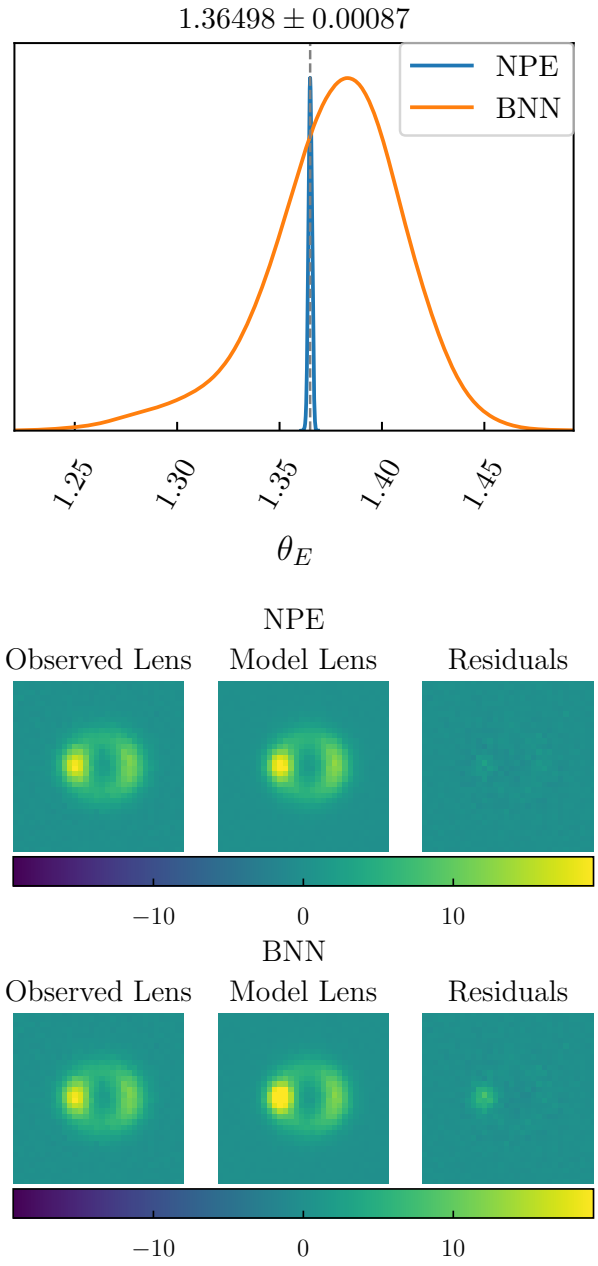

Figure 5.6.1 An example of inference on Einstein radius θ_E for a single image in a one-parameter model for NPE and BNN models. **Top:** Inferred posterior distributions for our NPE (blue) and BNN (orange) models. The value and uncertainty above the plot are for the NPE model. The true value is indicated by the dashed line. **Bottom:** "Observed" lens image from simulator, model image predicted by the best-fit Einstein radius, and residuals for NPE and BNN models. We find that inference from our NPE model is generally more performant than the BNN model, being both more accurate and precise. We provide several more examples in Appendix A.

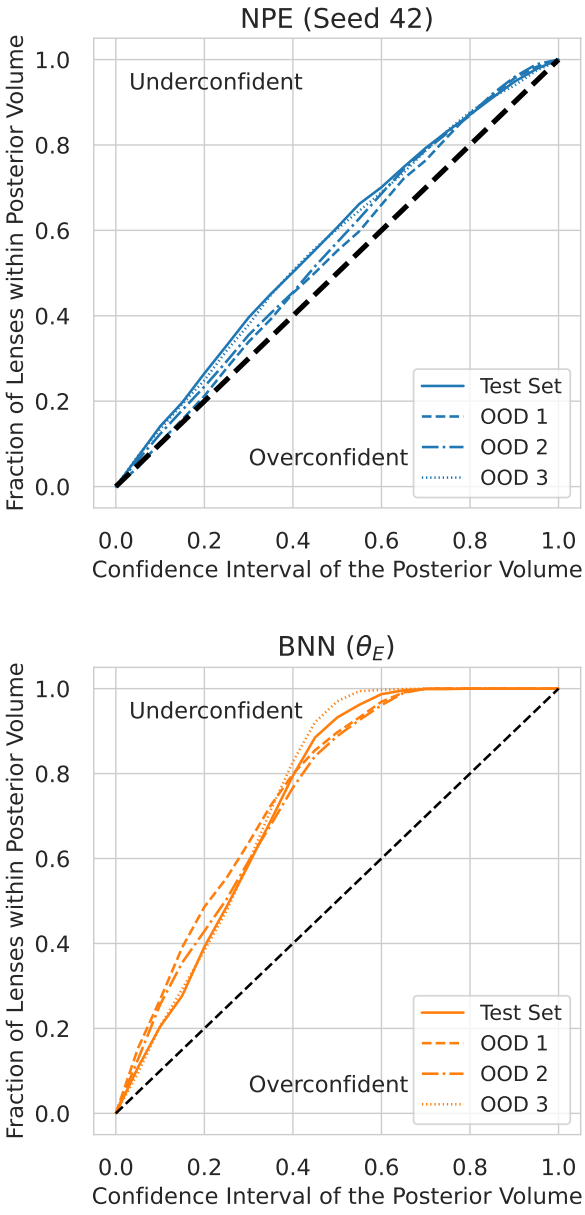

Figure 5.6.2 Posterior coverage plots of each of the lens parameters inferred using the NPE (above) and BNN (below) methods for the same test set. The 1:1 line indicates perfect uncertainty calibration and is indicated by a dashed black line. The NPE's posterior coverage is accurate, while the BNN's is extremely underconfident.

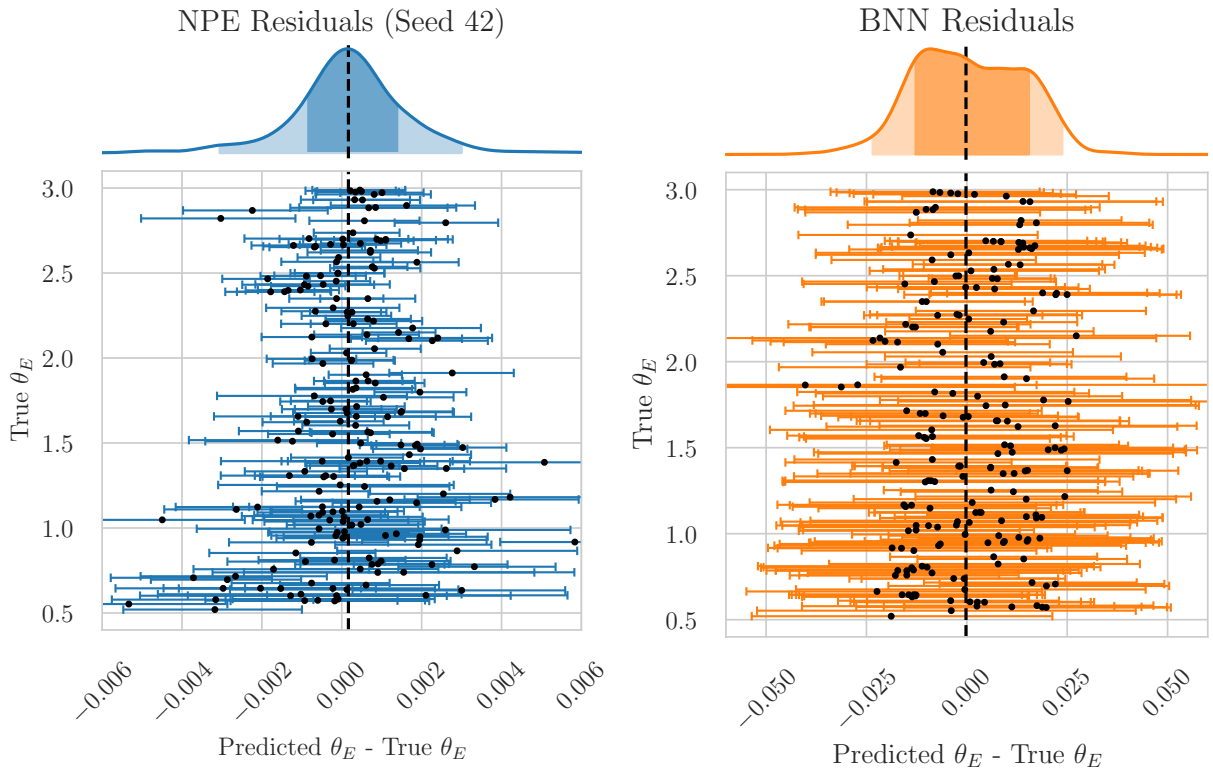

Figure 5.6.3 1-parameter scatter plots of the Einstein radius θ_E for the same 200 randomly selected images from the test set for the NPE and BNN models. The x-axis plots the difference between the best-fit (maximum log-probability) Einstein radii predicted by the neural posterior model and its true value as a function of the true θ_E (y-axis). The error bars correspond to the posterior models' 68% uncertainties for each image. Above the scatter plot, we also plot the marginal distribution of best-fit values for the entire test set of 1000 images using a kernel density estimator with kernel width selected with Scott's rule (Scot, 1992). The 68% and 95% scatter in best-fit values are indicated by the dark and lighter shaded regions respectively.

intervals from the inferred posterior distributions of the ensemble are correctly calibrated to reflect the true uncertainties in the ensemble population. Figure 5.6.2 shows the posterior coverage plots for the NPE and BNN models. The solid line represents the posterior coverage plot for the canonical test set. We see that the NPE method is slightly under-confident while the BNN model is extremely underconfident. This can be seen in the scatter plots as well. We also show the posterior coverage plots for the three out-of-distribution (OOD) test sets and find that they perform similarly to the canonical test set and are slightly underconfident.

Second, we create scatter plots of residuals as described in §5.5.2. To summarize, we calculate the difference between the best-fit (maximum likelihood) parameter values from their true values for all of the 1000 test images (we refer to this as the *residuals*) and plot these distributions of residuals pairwise by parameter. This serves as a point-estimate diagnostic for the test set that evaluates how well the models can predict the best-fit values, with a completely accurate model producing a single point at the origin (i.e. best-fit values exactly match the true values across the entire test set).

We perform kernel density estimation on the parameter distribution with a bandwidth calculated using Scott's rule (Scot, 1992), and plot the resulting 68% and 95% confidence intervals in the spread of the residuals, which we plot in the top part of Figure 5.6.3. For illustration, we also present 200 randomly selected test images and plot the posterior model's best-fit values and 68% confidence intervals. The 68% confidence intervals are $\sim 10^{-3}$ arcsec for the NPE model and $\sim 10^{-2}$ arcsec for the BNN model. This is in agreement with the single image example in Figure 5.6.1 and further validates that specific example as representative of the typical performance of both models. We also observe a slight trend where the uncertainties decrease with increasing Einstein radii in the NPE model. This may be because, at lower Einstein radii ($<$ 1arcsec), the blending effect of the point spread function of the simulated images becomes comparable to the size of the Einstein radii itself and becomes a significant source of systematic uncertainty.

We also tested the extent to which the choice of the seed initialization affected the model performance of our NPE model. We initialized our networks using 3 different seeds (42, 465, 839) and trained our models on the same training set for each seed. Upon testing, we find that for the one-parameter model, variations between model predictions of θ_E were on the order of $\sim 10^{-3}$ arcsec ($\sim$0.1%), on the same order of magnitude of the predicted model uncertainties. Figure 5.A.1 shows some of the variation in residuals for the different seed models on the same test images. In principle, this is undesirable, as it implies that the choice of random seed does significantly impact the model posterior predictions. However, we consider this an edge case, as a $\sim 10^{-3}$ arcsec modeling uncertainty is only achievable for extremely constrained and unrealistic toy models such as this 1-parameter model. In practice, modeling uncertainties for θ_E even in the most optimal real-world cases - i.e. high-resolution space-based images of simple lens systems (Ruff et al., 2011) - are on the order of 5%.

What this 1-parameter model does demonstrate is that there is perhaps a stochastic noise floor caused by the nature of how neural network model weights are randomly initialized, and that the uncertainty of this noise floor is on the order of $\sim$0.1%. In this case, what our results show is that given a sufficiently constrained model, it is possible to attain the precision and accuracy level up to the level of this noise floor. Looking forward, it may be possible to push down the noise floor further by doing large-scale ensembling of model results to average out the noise. We discuss this in §5.7.

Moving on to the posterior coverage plots (Figure 5.A.2), we find that the NPE model is generally well-calibrated, whereas the BNN model is underconfident. The posterior coverage for the out-of-distribution test sets does not significantly differ from the canonical test set, showing that the model is generalizable to an OOD test set. Repeating this experiment on 3 different seeds, we find that the posterior coverage plot fluctuates between the different seeds, ranging from underconfident in seed 42 to overconfident in seed 465. This reflects the

sensitivity of the posterior coverage diagnostic - a small change in the model's uncertainty predictions - on the order of 0.001 arcsec - will cause a visual change in the posterior coverage plot.

The SBC plots for the three NPE models with different random seed initializations show that the biases of the 1-parameter models vary between seed numbers. Visually inspecting the 3 different plots, we find that all three models in Figure 5.A.3 exhibit slight ∩-shaped distributions, indicating that the posterior models slightly overestimate the true posterior uncertainty. In addition, the seed 42 model skews slightly to the left, indicating a slight systematic bias towards lower values compared to the true posterior (underestimation) while the seed 465 model skews slightly to the right, indicating that there is a slight systematic bias towards higher values compared to the true posterior (overestimation). We note that these biases are virtually imperceptible when looking at the residual plots (Figure 5.A.1) or the posterior coverage plot (Figure 5.A.2). This demonstrates the complementarity of using multiple diagnostics when evaluating neural density estimator models.

5.6.2 5-parameter Fixed-source Model

The 5-parameter model increases the complexity of the modeling problem over the 1-parameter toy model by also additionally including 2 eccentricity moduli parameters and 2 coordinate position parameters to the inference. As mentioned in §5.4, we trained the models on 400,000 images with parameter distributions shown in Table 5.4.2. In training, the neural density estimator was instantiated with 120 hidden features and 20 transforms for the MAF. We evaluated the performance on a test set of 1000 lenses generated using the priors from the test set, also presented in Table 5.4.2.

The performance of our NPE and BNN models on this 5-parameter lens model is broadly similar to that of the 1-parameter model. Figure 5.6.4 shows a typical parameter inference on a single image. Similarly to the 1-parameter model, we see that the NPE model typically

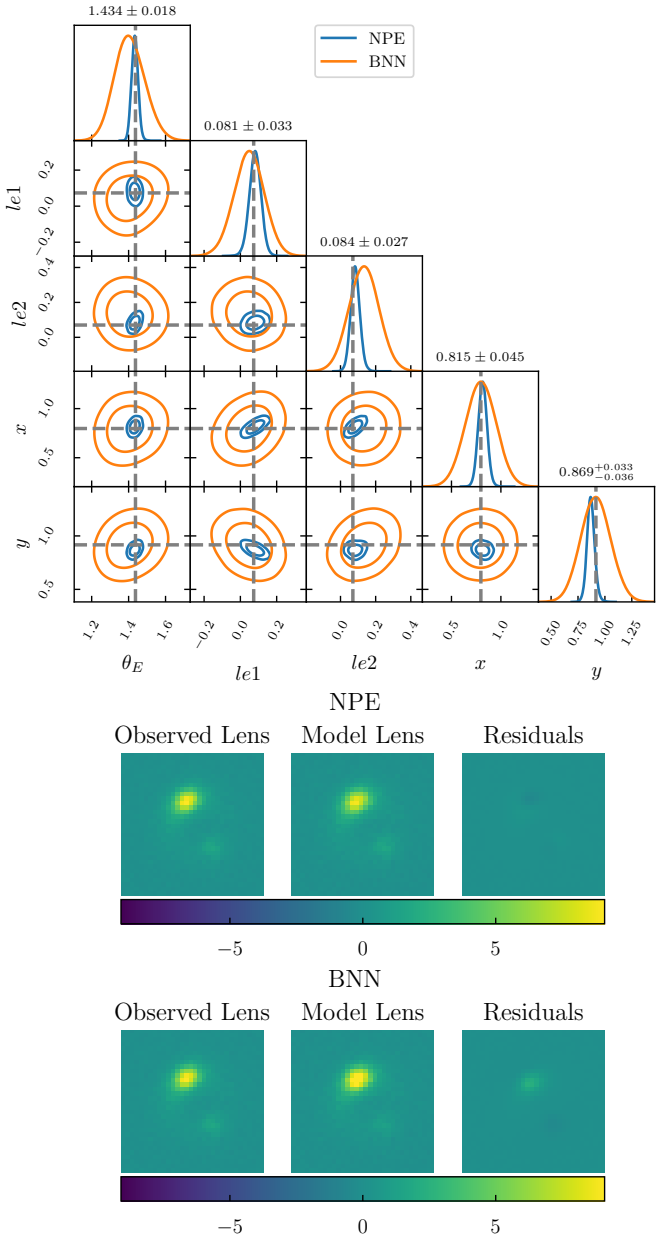

Figure 5.6.4 An example of how the NPE and BNN models perform on a single image for a 5-parameter model. **Top:** Inferred posterior distributions for our NPE (blue) and BNN (orange) models. The value and uncertainty above the plot are for the NPE model. The true value is indicated by the dashed line. **Bottom:** "Observed" lens image from simulator, model image predicted by the best-fit Einstein radius, and residuals for NPE and BNN models. We find that inference from our NPE model is generally more performant than the BNN model, being both more accurate and precise. We provide several more examples in Appendix A.

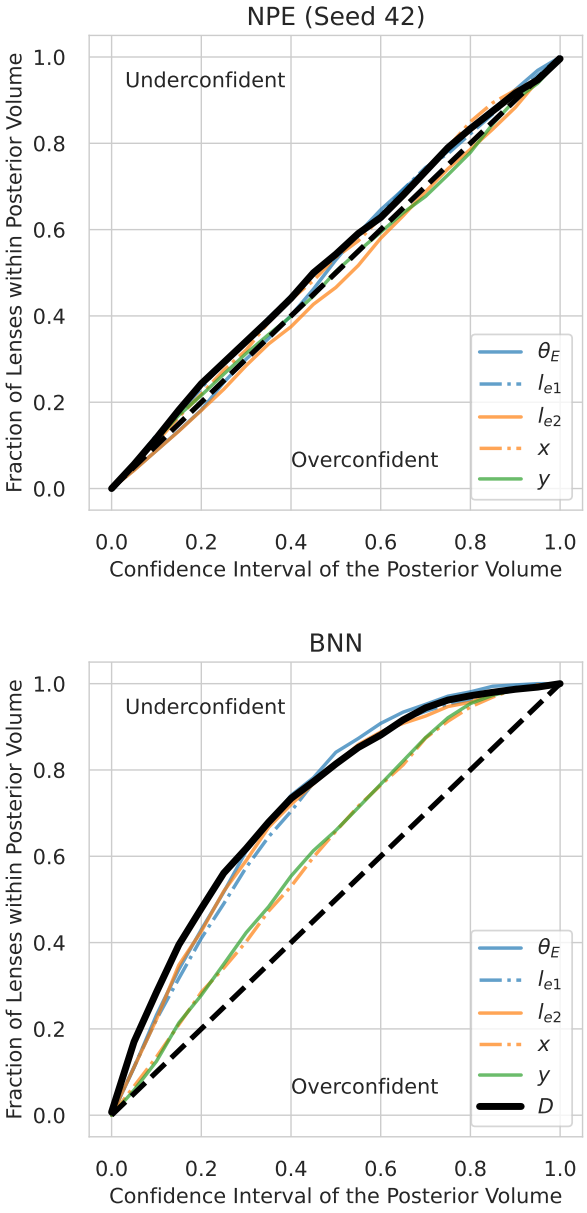

Figure 5.6.5 Posterior coverage plots of each of the 5 lens parameters inferred using the NPE (left) and BNN (right) methods for the same test set. The 1:1 line indicates perfect uncertainty calibration and is indicated by a dashed black line. The solid black line represents the distance metric described in Eq. 18 of (Wagner-Carena et al., 2021), which combines all lens parameters into a single distance metric while accounting for the empirical covariance between parameters. The NPE outperforms the BNN in every individual parameter and the combined metric shown here.

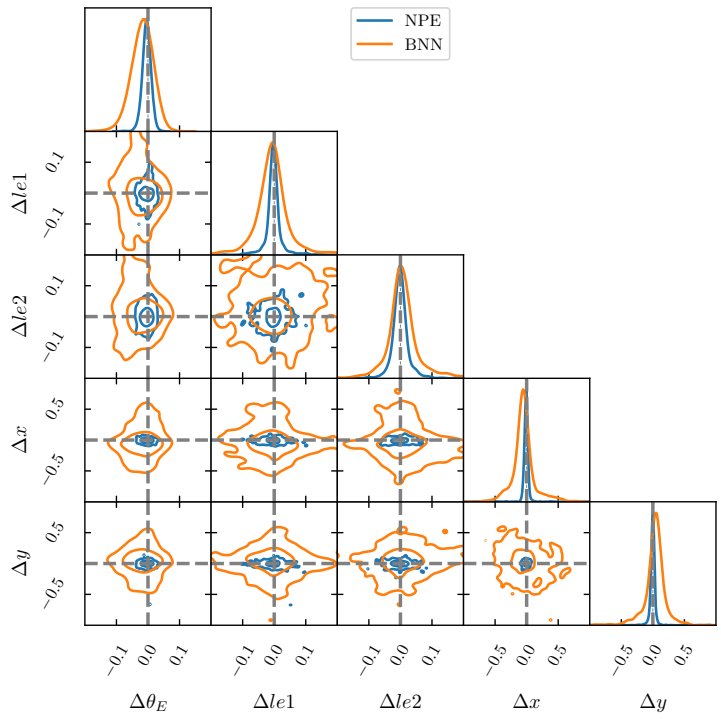

Figure 5.6.6 5-parameter scatter plots of the Einstein radius θ_E for the same test set of 1000 images for the NPE and BNN models. The x-axis plots the difference between the best-fit (maximum log-probability) Einstein radii predicted by the neural posterior model and its true value as a function of the true θ_E (y-axis). The error bars correspond to the posterior models' 68% uncertainties for each image.

produces significantly smaller uncertainties compared to the BNN model, although, unlike the 1-parameter model, the BNN uncertainties are on average a factor of 2-5 times larger than the NPE model, rather than 10 times as seen in the 1-parameter model. Compared to the 1-parameter model, the average uncertainty in the Einstein radius is about $\sim$0.02 arcsec for the NPE and $\sim$0.08 arcsec for the BNN model, compared to $\sim 10^{-3}$ arcsec and $\sim 10^{-2}$ arcsec 68% in the 1-parameter model.

The posterior coverage diagnostic (Figure 5.6.5) exhibits similar characteristics as the 1-parameter model, where the NPE model uncertainties are quite accurately calibrated whereas the BNN uncertainties are underconfident for every parameter in the model (albeit slightly less so for the position coordinate parameters).

The scatter plot diagnostic shows that the 68% percentile scatter in the inferred best-fit residuals from the NPE and BNN models are both unbiased, as the median values are consistent with zero within the 68% constraints (see Table 5.6.1 and Figure 5.6.6). However, the 68% scatter is 2-3 times wider in the BNN model than in the NPE model, showing that the BNN model is less precise at inferring the ground truth parameter values compared to the NPE.

Next, we evaluate the effect of random seed initialization on the NPE model. As with the 1-parameter model, we randomly initialize our models using 3 different seeds (42, 465, 839) and train each model on the same training data. We then evaluate the 3 models on the same test data. The resulting posterior coverage plot diagnostics, shown in Figure 5.A.4, show that

Table 5.6.1 Average 68% scatter in the difference between the best-fit lens parameters from 1000 lenses for SBI and BNN lens parameters

Parameter	SBI	BNN
$\theta_{\rm E}$(")	$0.001^{+0.016}_{-0.015}$	$-0.003^{+0.030}_{-0.037}$
le_1	0.003 ± 0.038	-0.006 ± 0.081
le_2	$-0.002^{+0.026}_{-0.021}$	$-0.003^{+0.078}_{-0.087}$
x (")	$0.016^{+0.027}_{-0.040}$	$0.07^{+0.27}_{-0.40}$
y (")	0.02 ± 0.14	$-0.04^{+0.41}_{-0.28}$

there are slight differences between the uncertainty calibration between seeds, but that the NPE model remains relatively well-calibrated across all 3 seeds. Using the combined distance metric, Eq. 5.27 as a proxy for the overall uncertainty calibration across all 5 parameters, we can see that seed 42 is just slightly underconfident, seed 465 is almost perfectly calibrated, and seed 839 is just slightly overconfident. Such slight calibration differences are unlikely to be the most significant source of uncertainty in real-life applications, but it also highlights the importance of evaluating the effects of random seed on the performance and potential stability of deep learning models. We discuss some strategies for dealing with this in §5.7.

The SBC histogram diagnostics, shown in Figure 5.A.10 and Figure 5.A.9, also offer some insight into the uncertainty calibration of individual parameters. From the empirical cumulative distribution function of the SBC ranks for each parameter and model seed in Figure 5.A.10, the empirical cumulative distributions fall within the expected deviation (shown in grey), except for the Einstein radius, θ_{E}, for seeds 42 and 465, as well as the ellipticity component, le_1, for seed 42, all of which are biased low compared to the true posterior. This is similarly visually reflected in the rank histograms of the parameters in Fig. 5.A.9 skewing right.

Finally, we evaluate the performance of both NPE (with all 3 seeds) and BNN models on the 3 OOD test sets described in Table 5.4.2. We evaluate the posterior coverages for all these models. For the NPE models, the posterior coverage plots are shown in Figures 5.A.5, 5.A.6, and 5.A.7, while the BNN model posterior coverage is shown in Figure 5.A.8. These diagnostics show that the OOD test sets do not change the main result, which is that the uncertainty calibration of the NPE models continues to be more accurate than the BNNs, which remain underconfident across all 3 OOD test sets. There are slight variations in the posterior coverage plot of each model across each OOD test set, but no indication that any model exhibits performance degradation or catastrophic failure, demonstrating that these models are robust to variations in the parameter distribution of the test set, as long that

they fall within the boundaries of the prior range of the training set.

5.6.3 12-Parameter Full Model

The 12-parameter model represents the most complex modeling challenge that we tackle in this work. In addition to the 5 lens parameters discussed in §5.6.2, we model an additional 2 components of external shear as well as 5 components of the source light, parameterized by a Sersic profile (see §5.4 for full details). We trained the models on 800,000 images with parameter distributions shown in Table 5.4.2. In training, the neural density estimator was instantiated with 400 hidden features and 20 transforms for the MAF. We evaluated the performance on a test set of 1000 lenses generated using the priors from the test set, also presented in Table 5.4.2.

Figures 5.6.7 and 5.6.8 show the inferred posterior distributions and best-fit image reconstruction and residuals for a single image respectively. Continuing the performance trends exhibited in the 1- and 5-parameter models, the NPE model was more accurate and precise than the BNN model, inferring significantly smaller uncertainties on the same test image.

The posterior coverage of the models is shown in Figure 5.6.9. We see that the NPE continues to outperform the BNN model concerning uncertainty calibration, with the empirical coverage for all parameters following the 1-1 dashed line indicating perfect calibration much more closely than the BNN model, which overestimates the uncertainties (i.e. is underconfident) for every parameter except the Sersic half-light, R, and Sersic index, n, parameters respectively. Worryingly, those two parameters exhibit an 'S' shaped coverage curve, which is usually an indication that the inferred posteriors are not accurately capturing the ground truth.

The scatter plot of the residuals of the best-fit, shown in Figure 5.6.10, provides further insight into the empirical performance of both the NPE and BNN models on predicting the best-fit values of different lens model parameters. Unlike in the 1- and 5-parameter

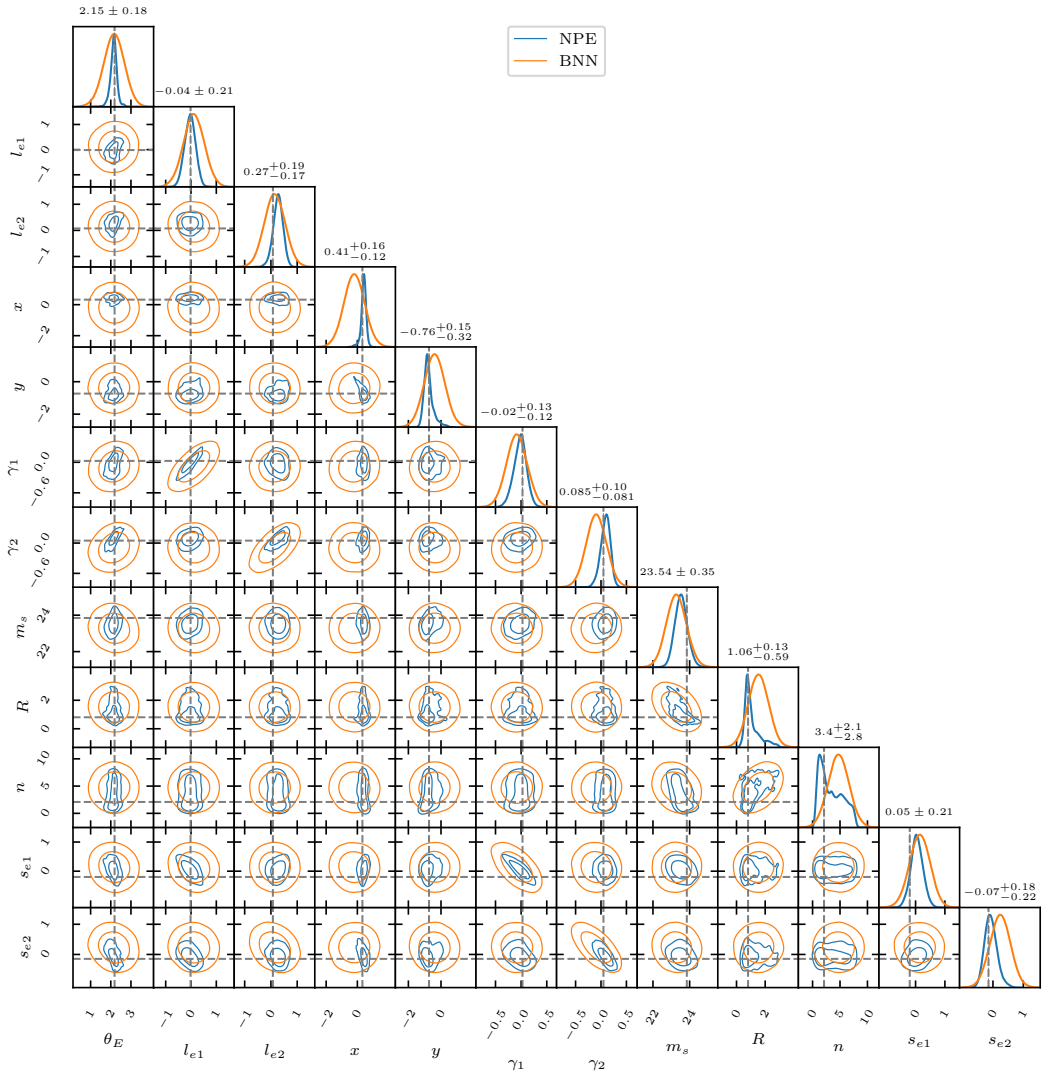

Figure 5.6.7 An example of the inferred posterior distributions for a 12-parameter lens model using NPE (blue) and BNN (orange) posterior models. The value and uncertainty above the plot are for the NPE model. The true value is indicated by the dashed line. We find that inference from our NPE model is generally more performant than the BNN model, being both more accurate and precise. We provide several more examples in Appendix A.

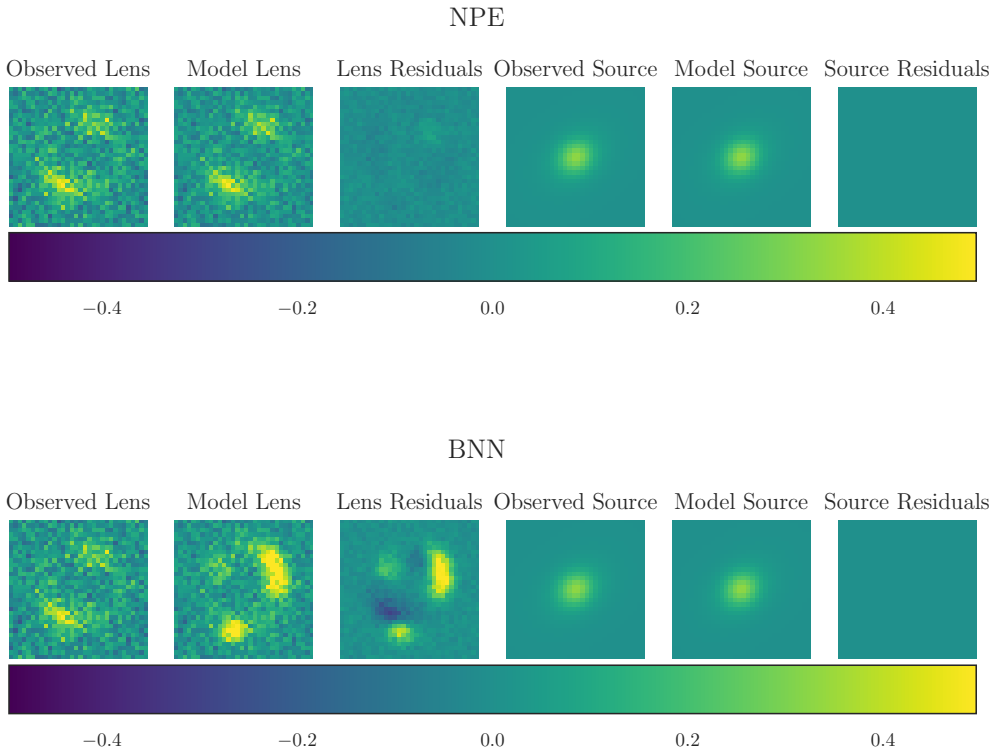

Figure 5.6.8 "Observed" lens image from the simulator, model image predicted by the best-fit lens model parameters, and residuals for NPE and BNN models for a 12-parameter model. We find that inference from our NPE model is generally more performant than the BNN model, being both more accurate and precise. We provide several more examples in Appendix A.

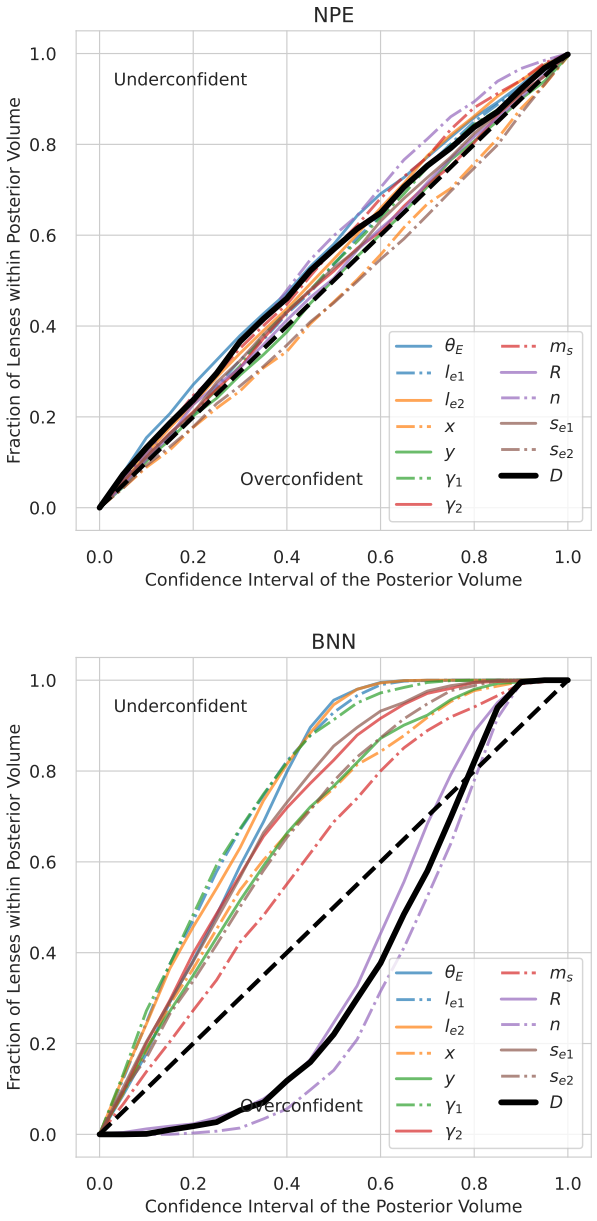

Figure 5.6.9 Posterior coverage plots of each of the 12 lens parameters inferred using the NPE (top) and BNN (bottom) methods for the same test set. The 1:1 line indicates perfect uncertainty calibration and is indicated by a dashed black line. The solid black line represents the distance metric described in Eq. 18 of (Wagner-Carena et al., 2021), which combines all lens parameters into a single distance metric while accounting for the empirical covariance between parameters. The NPE outperforms the BNN in every individual parameter and in the combined metric shown here.

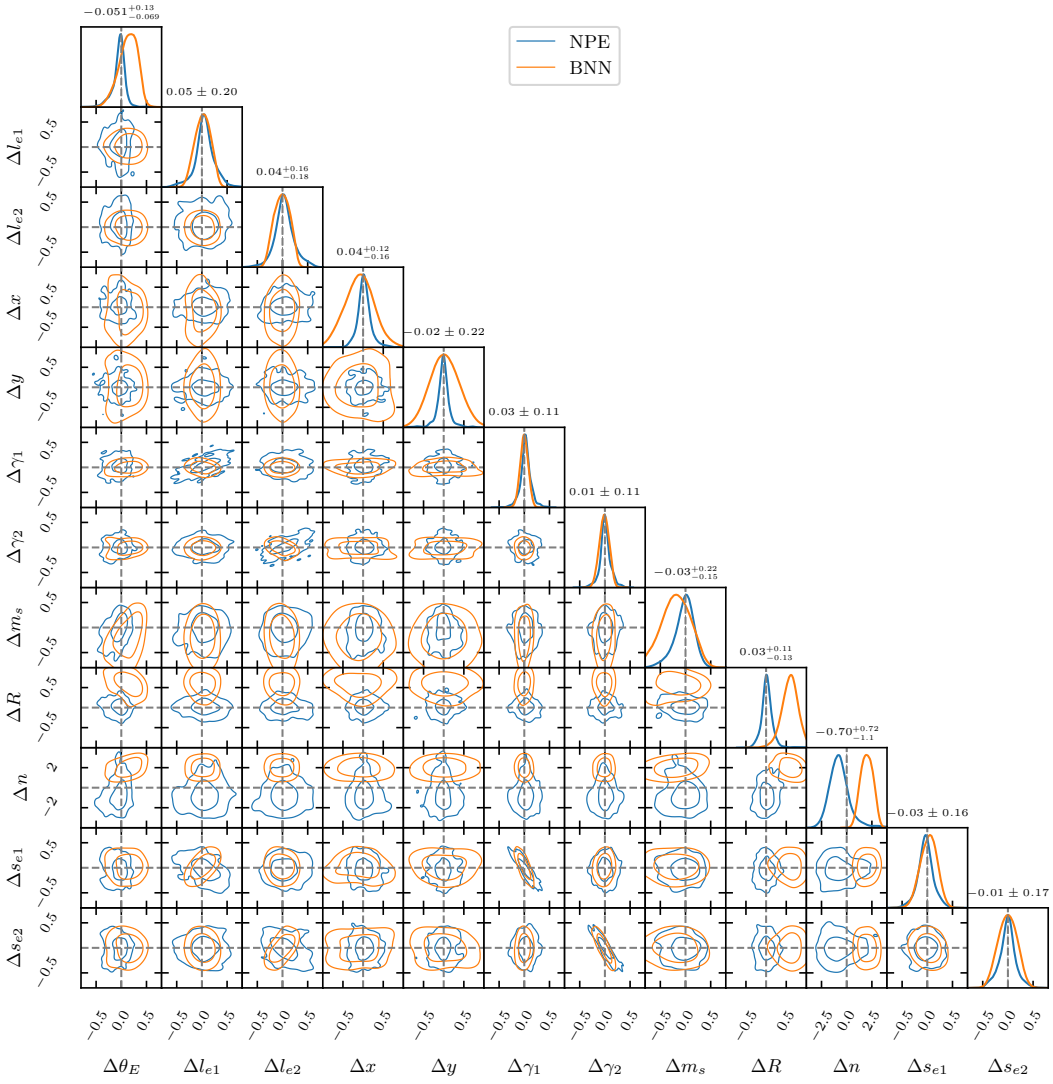

Figure 5.6.10 Scatter plot matrix of the differences between the best-fit posterior values and true value for 1000 test images for NPE (blue) and BNN (orange) methods with lens light subtracted. The contours show approximate 68th and 95th percentile uncertainties in the scatter. The dotted lines indicate the (0,0) point. Both methods produce a similar contour area suggesting roughly similar model precision, but BNN exhibits systematic bias for the source Sersic R and n parameters.

models, where the NPE model outperforms the BNN model in both accuracy (how centered each parameter residual distribution is to the zero point) and precision (how narrow the width of the residual distribution is) across all parameters, this is not the general case in the 12-parameter model. Instead, we observe the following trends: for source and plane ellipticity (l_{e_1},l_{e_2},s_{e_1},s_{e_2}) and external shear (γ_1,γ_2) parameters, both NPE and BNN models are roughly similarly precise, producing residual scatter plots of similar size, although there is a slight indication that the NPE model residuals have a longer tail than the BNN model for those parameters, with slightly wider 95% contours than the BNN.

The BNN did not perform as well compared to the NPE for the remaining parameters. For the coordinate parameters (x,y), both NPE and BNN models were unbiased (accurate), with their residuals distributed about the zero point; however, the BNN residuals exhibit significantly larger scatter (lower precision) compared to the NPE residuals, with the BNN's 68% uncertainties being similar in size to the NPE model's 95% uncertainties. For the Einstein radius (θ_{E}), source magnitude (m_s), Sersic half-light radius (R), and Sersic index (n) parameters, The BNN's inference was systematically biased; on average, it slightly overestimates $\theta_{\mathrm{E}}$and slightly underestimates m_s, as evident by the slightly shifted peaks of the residuals away from the zero point. The BNN performed worst on R and n, overestimating the true values of R and n for more than 95% of the test set. On average, it also overestimates R by more than 0.5 arcsec and n by 2, both of which are significantly large enough for the inference to be considered unsuccessful. This validates the observed 'S' shaped posterior coverage plot for these parameters in Figure 5.6.9. In comparison, the NPE was both more accurate and precise in its inference of the best-fit parameter values for all these parameters. Notably, the NPE was systematically unbiased for all the parameters except n, which it underestimates on average.

Moving on, we evaluate the effect of random seed initialization on the NPE model. As with the 1- and 5- 5-parameter models, we randomly initialize our models using 3 different

seeds (42, 465, 839) and train each model on the same training data. We then evaluate the 3 models on the same canonical test data. As with the 5-parameter model, the posterior coverage plots, shown in Figure 5.A.11, show that the NPE model exhibits slight variations in uncertainty calibration between seeds. Using the combined distance metric, Eq. 5.27 as a proxy for the overall uncertainty calibration across all 5 parameters, we can see that seed 42 is almost perfectly calibrated, seed 465 is slightly underconfident, and seed 839 is slightly overconfident.

Compared to the 5-parameter model, however, the uncertainty calibration of each parameter exhibits larger variation between seeds than in the 5-parameter model, as well as larger deviations from the 1-1 line indication of perfect calibration. For example, the least well-calibrated parameter across all 3 seeds in the 5-parameter NPE model is the Einstein radius θ_{E} for seed 839 (solid blue line on rightmost plot in Figure 5.A.4), which, at worst, exhibits about $\sim$5% confidence interval (CI) difference from perfect calibration (i.e. the empirical fraction of lenses in the test set with true values contained within the 60% CI of the inferred posterior of each lens is 0.65) (DOUBLE CHECK TRUE VALUE). In contrast, the most miscalibrated parameter across all 3 seeds in the 12-parameter NPE model is the Einstein radius θ_{E} for seed 839 (solid blue line on rightmost plot in Figure 5.A.11), which exhibits a about $\sim$10-15% confidence interval (CI) difference from perfect calibration, with 60% CI of the inferred posteriors containing 75% of the ground truth Einstein radii of the lenses in the test set, while it is less than 10% in the other 2 seeds.

This result shows that the effects of seed initialization can become increasingly significant when both the simulation complexity and dimensionality of the inference problem are increased, as in this case. The uncertainty calibration of a trained model can vary between seeds even when the model architecture (including hyperparameters), training, and test sets remain the same. This increased stochastic variation between different seeds highlights the importance of the empirical/posterior coverage as a visual diagnostic for evaluating the per-

formance of any trained density estimator whenever any stochastic training procedure is performed.

The SBC rank diagnostics for the NPE model for the 3 different seeds, shown in Figure 5.A.16 (cumulative) and Figure 5.A.17 (histograms), provide useful insight on the uncertainty calibration of the model parameters. For example, if we see a consistent systematic miscalibration across different seed initializations, this indicates a source of systematic uncertainty that cannot be corrected by, retraining the model with a different seed initialization, or by ensembling or averaging the inferred results of multiple models trained on different seeds. On the other hand, if the SBC diagnostic reveals a deviation from good calibration that doesn't persist through multiple seed initializations, this indicates that the cause of the miscalibration may be stochastic and can be reduced or corrected through either training the model again on a different seed or by ensembling models.

As with the 1- and 5-parameter SBC diagnostic plots, the grey region represents the 99% confidence intervals of a uniform rank distribution, which indicates good calibration. Across the 3 seeds, most of the parameters generally produce rank histograms or empirical cumulative distribution functions (CDF) that are consistent with uniform distributions. However, there are a few exceptions. The source magnitude, m_S, appears to be biased high compared to the ground truth in all 3 seeds. This can be visually diagnosed in the CDF plots (Figure 5.A.16), where the empirical CDF deviates from uniformity (the grey region). In the rank histograms (Figure 5.A.17), we also see that the distributions are left-skewed indicating the same deviation from uniformity. Another parameter that appears to be systematically biased across all 3 seeds is the Sersic index, n, which is biased low compared to the ground truth. This is most apparent in seeds 465 and 839, where the rank histograms are skewed to the right. This systematic bias is consistent with the scatter plot diagnostic (Figure 5.6.10), which shows that the NPE model slightly underestimates the best-fit value of n compared to the ground trust values across the test set. Aside from those two parameters, we did not

observe any obvious deviation from uniformity for any other model parameter that persisted systematically across all 3 seeds.

Finally, we evaluate the performance of both NPE (with all 3 seeds) and BNN models on the 3 OOD test sets described in Table 5.4.2. We evaluate the posterior coverage for all these models. For the NPE models, the posterior coverage plots are shown in Figures 5.A.12, 5.A.13, and 5.A.14, while the BNN model posterior coverage is shown in Figure 5.A.15. Like for the 5-parameter model, the NPE models, across all 3 seeds, remained more well-calibrated than the BNN model for all 3 OOD test sets. The BNN model also exhibited large variations in posterior coverage for each OOD test set, especially compared to the NPE models, indicating that the BNN model is not as robust and generalizable to OOD variations in the parameter distribution of the test set compared to the NPE model.

5.7 Discussion

In this paper, we strove to be as thorough as possible in defining an appropriate level of ambition on three levels: 1) the fidelity of our simulations and dimensionality of the model parameters to be inferred, 2) the complexity of our neural posterior estimation methods, and 3) the thoroughness of our model evaluation diagnostics. However, it is impossible to explore every single nuance in a single work. In this section, we discuss some of these aspects and limitations. We hope some of these insights can help guide the scope of future studies incorporating neural posterior estimation techniques.

5.7.1 Mitigating Effects of Random Seeds with Ensembling

In all our models, we found that using different random seeds to initialize the neural posterior model weights before training resulted in small differences in predictions between identical model architectures that were trained on the same training data. In the 5- and 12-parameter models, these differences were subdominant to the uncertainties predicted by the posterior

models. However, in the 1-parameter model, which was precise to a sub-percent level, the stochastic noise becomes a significant source of variation in model performance between different random seeds.

In the future, to characterize the full impact of random seeds on the model performance, we suggest using ensembling methods to mitigate the impact of random seed variations - that is - by averaging the predictions from multiple models initialized from different random seeds while keeping all other training data and model hyperparameters fixed. If possible, performing large-scale ensemble runs can also help assess the stability and variability of model performance.

5.7.2 Adopting Sequential SBI for Follow-up Studies

As mentioned in section 5.3.2, we did not use sequential estimation methods because the goal of this work is to be able to efficiently evaluate a large catalog of lenses in an amortized fashion. However, we can envision a use case for using sequential posterior estimation methods as a follow-up on a smaller sub-sample of lenses from the catalog. For example, if, after an initial analysis using the methods described in this paper and ancillary observations, we identify a smaller sub-sample of information-rich lenses (e.g. quadruply-imaged lenses or time-delay lenses) that we would like to refine the posterior estimates of, we can perform SNPE on those lenses by using the posteriors obtained from NPE in this work as a proposal prior for subsequent rounds of inference, thereby ensuring that no computation is wasted.

5.7.3 Increasing Modeling Complexity

We limited our inference to models with up to 12 parameters in this exploratory work. This level of modeling complexity is appropriate for the vast majority of lenses we expect to discover from ground-based imaging data (i.e. single-lens, single-source galaxy-galaxy lenses with $\sim$1" resolution e.g. Agnello et al. (2017); Knabel et al. (2023)). However, more

complicated lens system morphologies, such as cluster-scale lenses, multiple source-place lenses, or subhalo studies have modeling requirements that could increase by several factors, to orders of ∼30-50 parameters (e.g. Mahler et al. (2023) for cluster lenses). Even with more informative space-based follow-up imaging (e.g. from JWST), the dimensionality of the problem poses a challenge for SBI methods.

In an earlier work applying SNPE to mechanistic models of neural dynamics, Goncalves et al. (2020) observed that *'current implementations of SNPE can scale well to high-dimensional observations (e.g. ∼1000s of dimensions), but scaling SNPE to even higher-dimensional parameter spaces (above 30) is challenging (note that previous approaches were generally limited to less than 10 dimensions). Given the difficulty of estimating full posteriors scales exponentially with dimensionality, this is an inherent challenge for all approaches that aim at full inference*'. In this case, the dimensionality of the observation means the dimensionality of the dataset (e.g. in this work, the dimensionality of our observations is 1024, corresponding to the number of pixels in each image).

While advances in both SBI methodology and density estimation techniques have enabled us to overcome the limitations of the so-called curse of dimensionality in classical ABC techniques, further advances must be made for SBI techniques to credibly scale to inferring parameters in higher dimensional spaces. In addition, all validation techniques used to ensure that the model's predictions are credible will also need to scale up to those number of parameters (e.g. OOD tests, uncertainty calibration).

Fortunately, there have been advancements in recent years that may enable SBI neural posterior estimation techniques to continue to scale to higher dimensionality problems in the future:

1. The development of simulators that are not simply black boxes that produce pairs of parameters and data (θ, X), but also additional information, such as the score $\nabla_\theta p(X|\theta)$ i.e., the gradient of the likelihood with respect the parameters of interest.

By augmenting the training of the neural density estimator with useful additional information such as the score, the training can be more performant. This has been applied to strong lensing analyses in the form of a proof-of-concept analysis of dark matter subhalo population properties in strong lenses (Brehmer et al., 2019b).

2. Using SBI in a Bayesian hierarchical modeling framework for greater constraining power, especially in the inference of population-level model parameters (e.g. cosmological parameters like Ω_m). Hierarchical models complement an SBI analysis by allowing us to combine SBI outputs with other prior information about model parameters at different levels of the hierarchy in a principled way, thereby increasing the constraining power of the inference. This has been demonstrated in strong lensing applications (Wagner-Carena et al., 2021; Wagner-Carena et al., 2023).

5.7.4 Post-training Calibration

In this work, we did not apply post-training calibration to the models to evaluate their native "out-of-the-box" performance. Post-training calibration in machine learning models is an actively studied field, with much of the literature focused on calibrating neural network predictions in a classification or regression task (e.g. Guo et al. (2017). Its applicability to SBI models in an astrophysical context is as of yet largely unexplored, although beta calibration (Kull et al., 2017) has been used to calibrate BNN models used to infer cosmological parameters from Cosmic Microwave Background maps (Hortúa et al., 2020).

Conformal prediction (Angelopoulos & Bates (2022); Vovk et al. (1999)) is a promising calibration technique, as it provides theoretical guarantees that the coverage prediction is valid provided that the test set is independent and identically distributed to the validation set used for the calibration. However, the calibrated model, while valid, does not discriminate - it predicts the same uncertainties for all test set data and is unable to predict larger uncertainties for data within the test set that have higher risk (e.g. noisier lens images).

While there has been recent work to make conformal prediction both valid and discriminative (Lin et al., 2021), it is unclear how this can be combined with existing SBI methods of uncertainty prediction. We leave the exploration of calibration techniques on SBI methods to future work.

5.8 Conclusion

Our main results are as follows:

1. Both NPEs and BNNs can predict informative (constrained) posterior distributions and point estimates of lens parameters from ground-based galaxy-galaxy lenses. We validate our model's performance with a comprehensive suite of diagnostics, which, to our knowledge, has not been consistently implemented in SBI studies to date.

2. Due to their amortized nature, the NPE and BNN models can infer the posterior densities of thousands of lens systems in minutes. This meets the scale requirement of being able to infer lens model parameters from tens to hundreds of thousands of lens systems we project will be discovered by future surveys. For the 12-parameter model, both NPE and BNN models took about 6 hours to train. In comparison, a conventional MCMC fit would take around the order of an hour per lens system.

3. This work is also the first-of-its-kind demonstration of the use of Masked Autoregressive Flows (MAF) for lens parameter posterior density estimation. The NPE models are also generalizable to out-of-distribution test data without much loss in constraining power.

4. We found that our NPE model uncertainties are generally more well-calibrated than our BNN model uncertainties. In this work, our BNN models tend to be rather underconfident, overpredicting the true uncertainties of the test images. We hypothesize

that this may be because a multivariate Gaussian model may not be sufficiently flexible enough to model lens parameter posterior densities. One potential method to improve the uncertainty calibration of our BNN model is to increase the number of Gaussian components in our BNN model weights. However, this comes with the tradeoff of a steep computational cost increase, as each additional Gaussian component requires 2 additional free parameters to be trained for every weight in the network.

5. The effect of random seed initialization in NPEs is subdominant to the model's predicted uncertainties in the 5- and 12-parameter models, implying that the choice of model will not significantly impact the results. In the 1-parameter model, the choice of random seed results in a 0.1% uncertainty, which is on the same order of magnitude as the uncertainties from the posterior itself. However, this is unlikely to impact the modeling of real lenses in practice, as achieving a 0.1% uncertainty prediction is only possible in the 1-parameter toy model, which is too simple to accurately model real lenses. In practice, real lenses have uncertainties typically on the order of 5% (e.g. Ruff et al. (2011)).

6. As a general guideline (with the caveat that this does not provide theoretical guarantees!), we recommend that all applications of SBI to posterior estimation, at minimum, include diagnostics to check that: (1) both maximum-likelihood point estimates and uncertainty predictions are, when averaged across an large ensemble of data, unbiased and well-calibrated with respect to ground-truth parameter values; (2) the performance of posterior models do not significantly degrade when applied to out-of-distribution data; and (3) model performance do not exhibit large variations with the choice of random seeds. These diagnostics are essential for showing that the SBI model produces reliable and reproducible results.

We see this work as a starting point for future work in applying SBI methods to strong

lensing analysis. One potential research avenue involves using SBI in a hierarchical framework to infer population-level statistics, such as cosmological parameters. It has been previously demonstrated that cosmological constraints can be extracted from large catalogs of galaxy-galaxy lenses using inferred lens model parameters such as those described in this work (Li et al., 2023; Leaf & Melia, 2018; Cao et al., 2015c; Bom et al., 2019). However, the potential utility of applying SBI methods to such an analysis is as yet largely unexplored.

The SBI methods used in this work can also be used as part of a modular strong lens analysis pipeline in astronomical surveys. For example, it can be used as a first step to obtain lens model posterior density estimates for a large catalog of lenses. If a suitable subsample of that catalog exhibits scientifically interesting properties worth following up on, the SBI estimates for those lenses can be combined with sequential SBI methods (as discussed in §5.7), or used as a prior distribution for conventional MCMC Bayesian fitting.

While SBI methods have made great improvements in performance in recent years due to new developments in machine learning, as well as exponential gains in computational power, it is by no means a mature field with no further room for achieving greater performance gains. As discussed in §5.7, SBI methods that leverage not only the training data but additional information that characters the latent process of the simulators (e.g. the score) can greatly increase the quality of inference. However, this requires the development of simulators that are not simply black boxes but are capable of also providing additional information on top of simulated data. The development and use of such simulators should be encouraged in future work. The potential performance improvements and computational trade-offs from ensembling large numbers of SBI models should also be investigated in more detail.

There are also a lot of potential gains to be made in uncertainty quantification and calibration. Theoretically, there remains a lot of investigation to be done on designing well-motivated and theoretically sound diagnostic metrics that can be used to evaluate model performance. Additionally, SBI methods may be able to leverage the wealth of existing

literature on post-training calibration techniques in machine learning to calibrate SBI models to produce uncertainty predictions that have formal theoretical guarantees.

With the myriad of unexplored research directions in both SBI methods and their applications, we are optimistic about the utility and potential for the growth of SBI methods in astrophysics.

5.A Comparison of NPE models with different seed initializations

5.A.1 1 parameter model

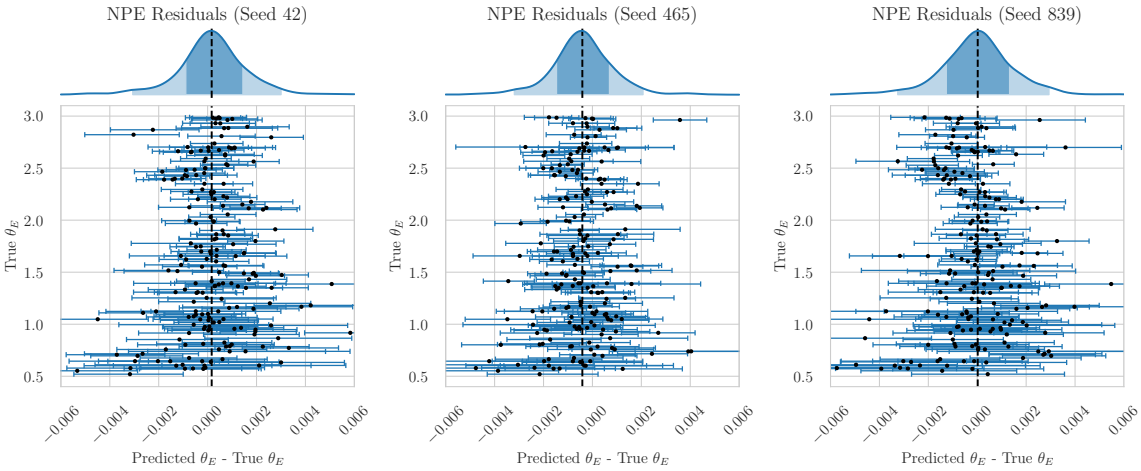

Figure 5.A.1 1-parameter scatter plots of the Einstein radius θ_E for the same 200 randomly selected images from the test set for 3 different random seed choices. The x-axis plots the difference between the best-fit (maximum log-probability) Einstein radii predicted by the neural posterior model and its true value as a function of the true θ_E (y-axis). The error bars correspond to the posterior models' 68% uncertainties for each image. Above the scatter plot, we also plot the marginal distribution of best-fit values for the entire test set of 1000 images using a kernel density estimator with kernel width selected with Scott's rule (Scot, 1992). The 68% and 95% scatter in best-fit values are indicated by the dark and lighter shaded regions respectively.

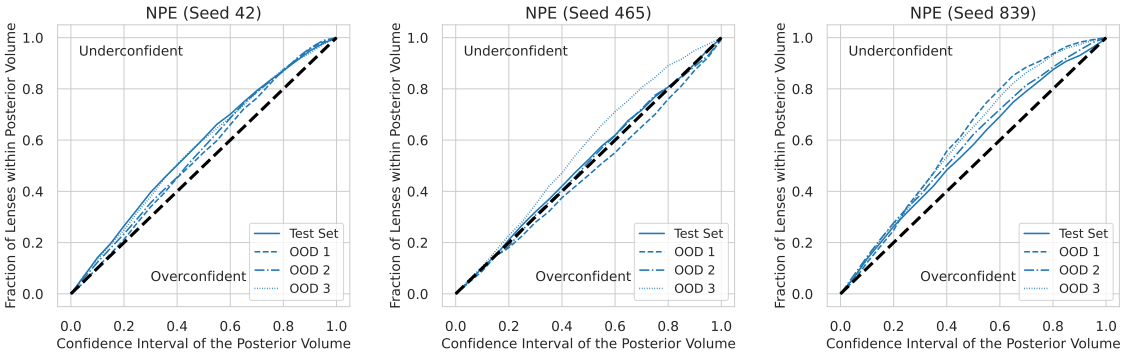

Figure 5.A.2 1-parameter model coverage plots for 3 different seed initializations. For each model trained a different seed, the posterior coverage for the canonical and three OOD test sets for that model are plotted together on the same plot.

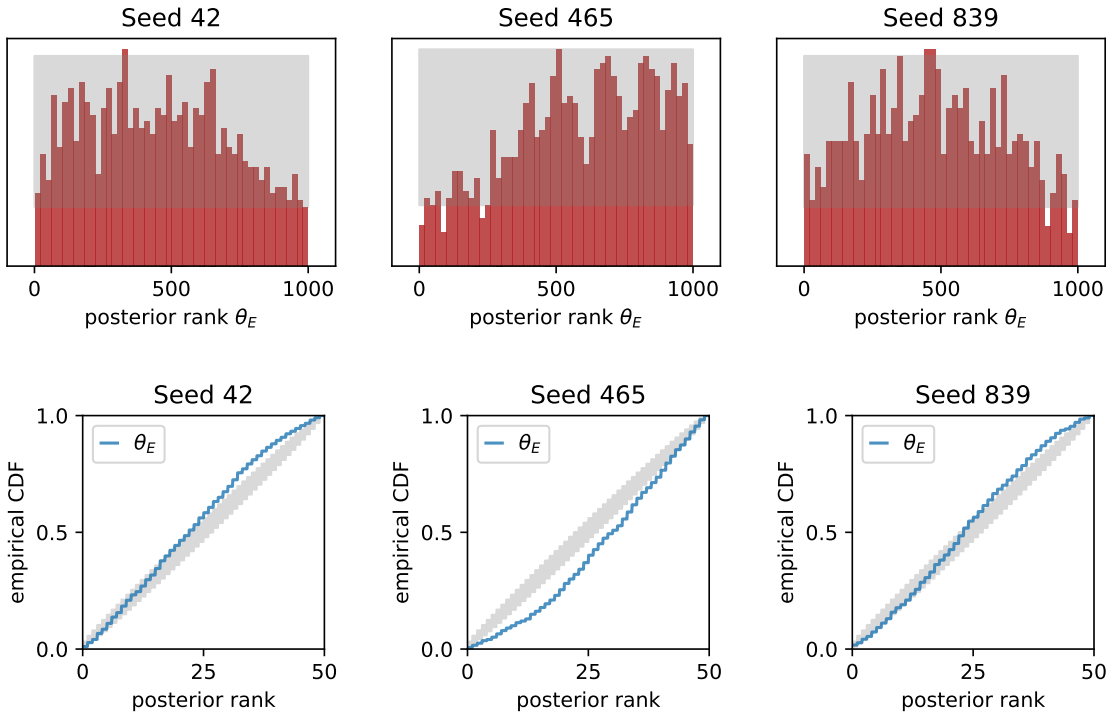

Figure 5.A.3 1-parameter model Simulation-based calibration (Talts et al., 2018) rank histogram plots (above) and cumulative density function plots (below) for the seeds 42, 465, 839. We discuss the plots in § 5.6.1. The grey regions indicate the 99% confidence interval of a uniform distribution given the number of samples provided. In a histogram, for a uniform distribution, we expect 1 out of 100 histogram bars to lie outside the confidence interval.

.

5.A.2 5 parameter model

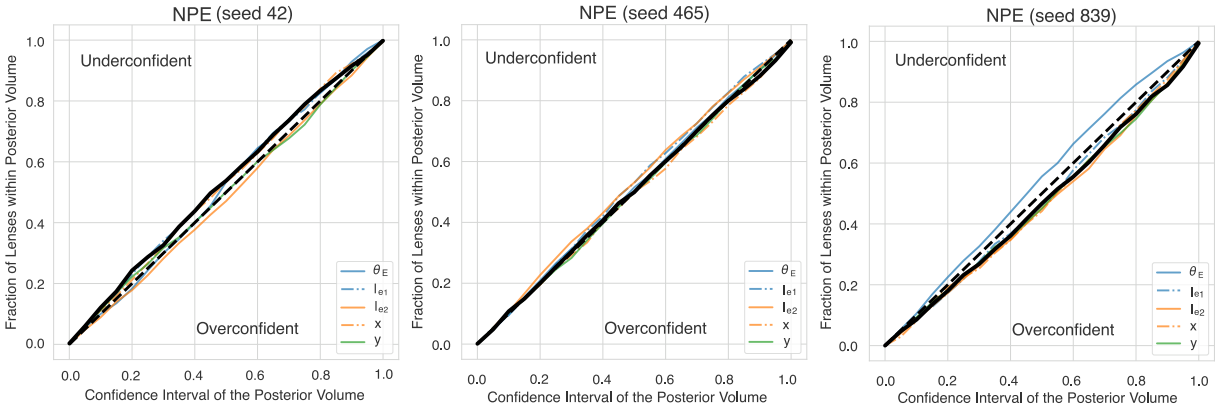

Figure 5.A.4 5-parameter model posterior coverage for 3 random seeds, canonical test set.

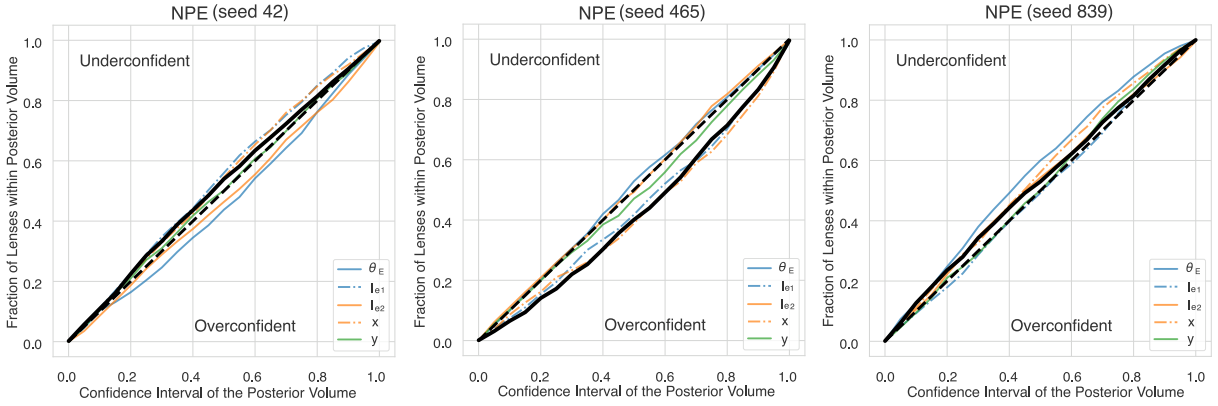

Figure 5.A.5 5-Parameter Model Posterior coverage for OOD Gaussian distribution test set 1 for seed 42, 465, 839.

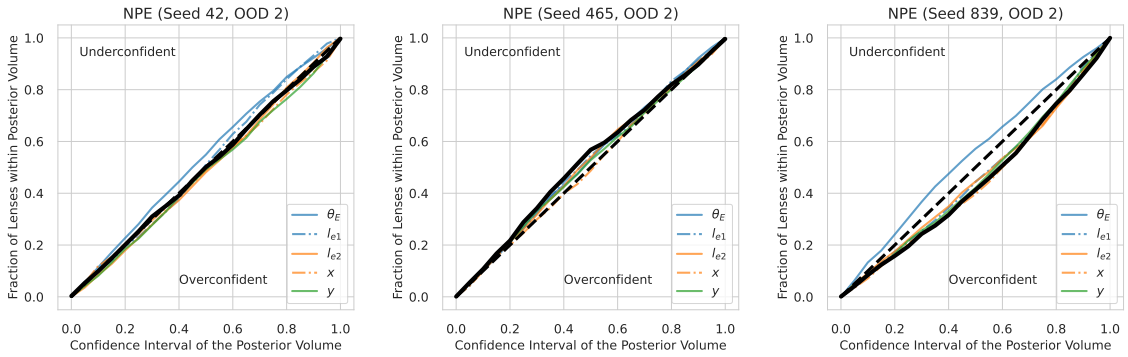

Figure 5.A.6 5-Parameter Model Posterior coverage for OOD Gaussian distribution test set 2 for seed 42, 465, 839

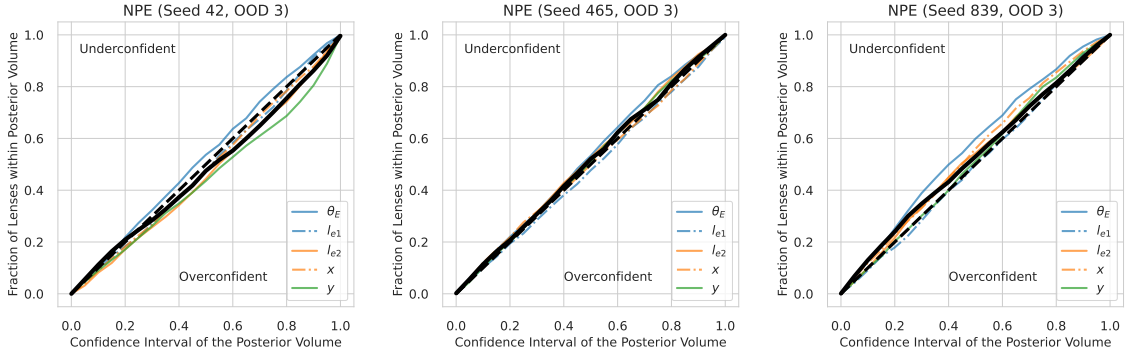

Figure 5.A.7 5 Parameter Model Posterior coverage for OOD Gaussian distribution test set 3 for seed 42, 465, 839

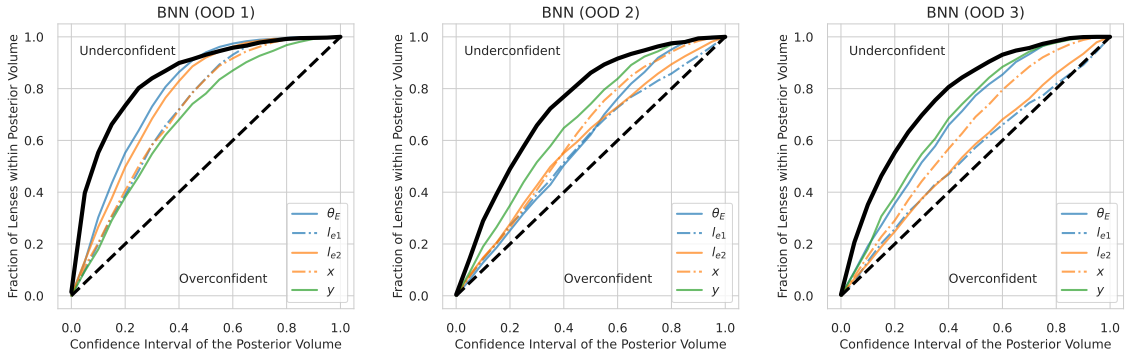

Figure 5.A.8 5-parameter model posterior coverage for the 3 OOD Gaussian distribution test sets for the BNN.

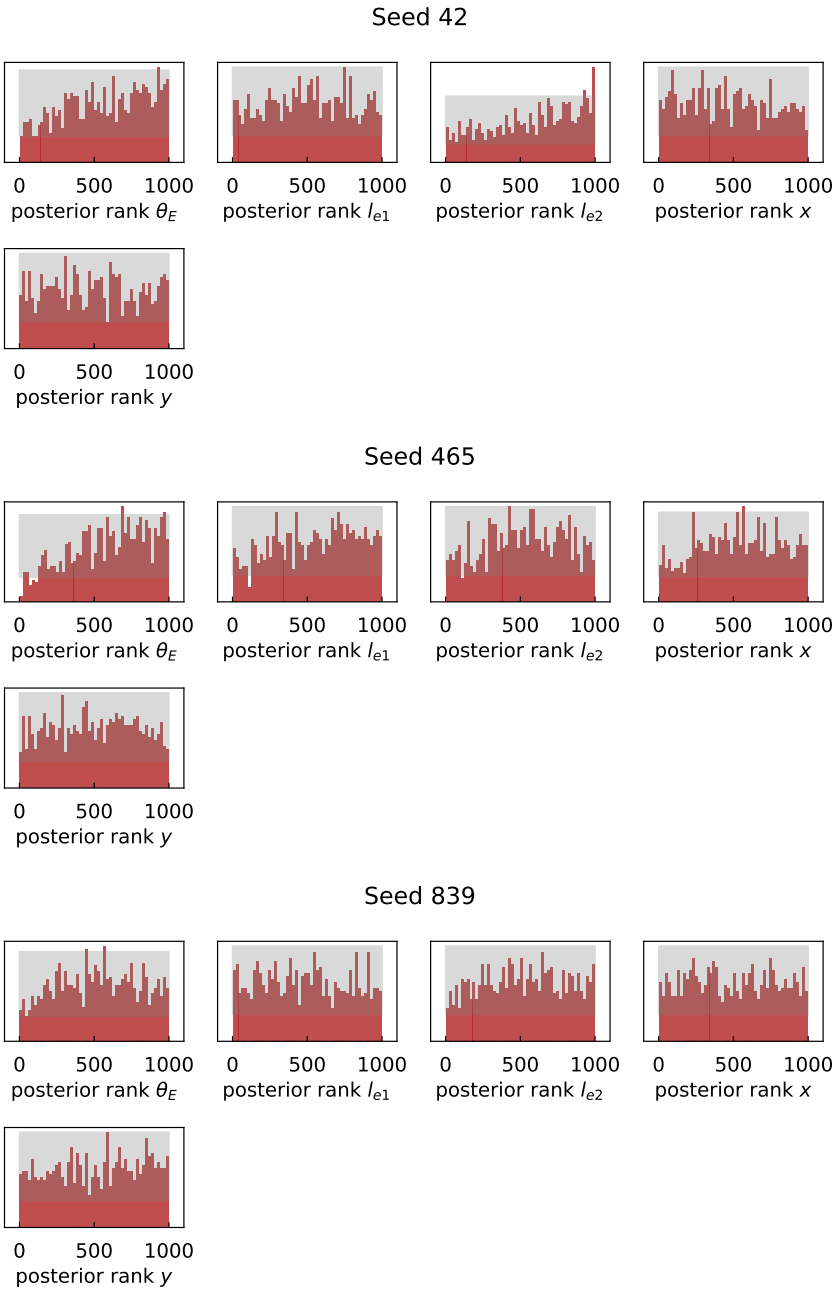

Figure 5.A.9 5-parameter model SBC histogram plots for seed 42, 465, 839. The grey regions indicate the 99% confidence interval of a uniform distribution given the number of samples provided. In a histogram, for a uniform distribution, we expect 1 out of 100 histogram bars to lie outside the confidence interval.

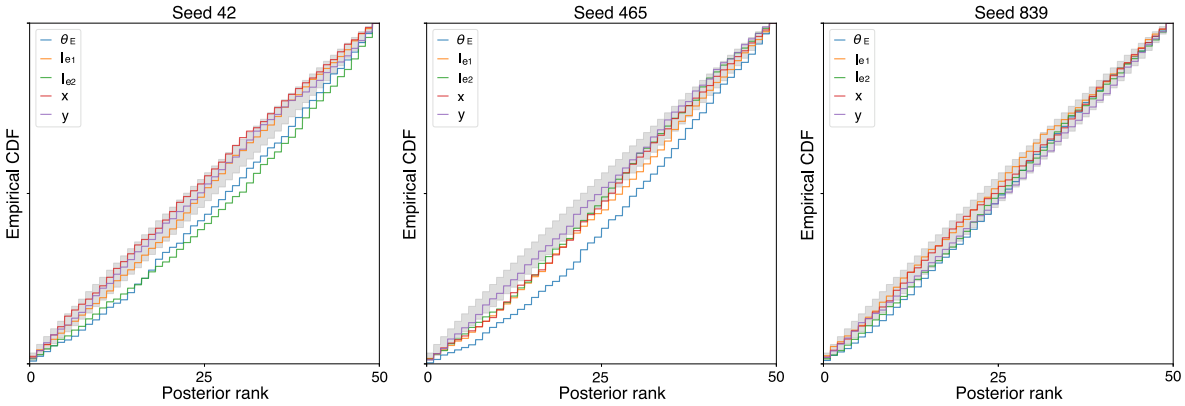

Figure 5.A.10 5-parameter model SBC CDF plots for seed 42, 465, 839. The grey regions indicate the 99% confidence interval of a uniform distribution given the number of samples provided.

5.A.3 12 parameter model

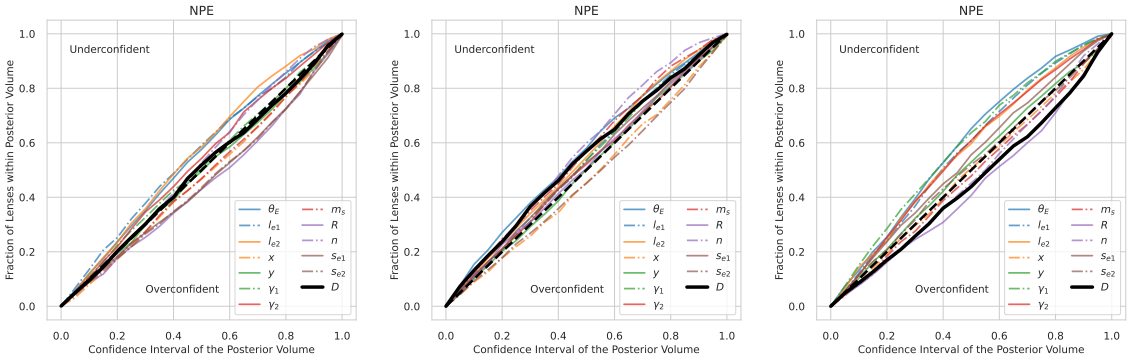

Figure 5.A.11 12 Parameter Model for 3 seeds, canonical test set

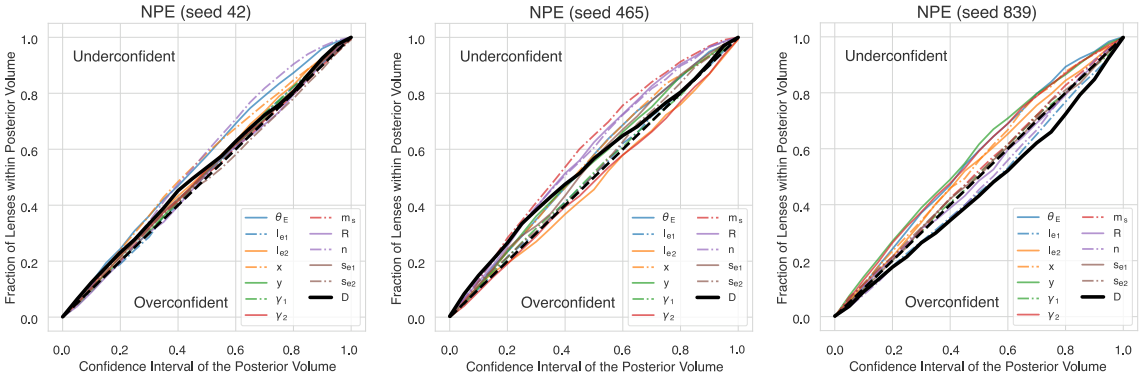

Figure 5.A.12 12-parameter model posterior coverage for OOD Gaussian distribution test set 1 for seed 42, 465, 839.

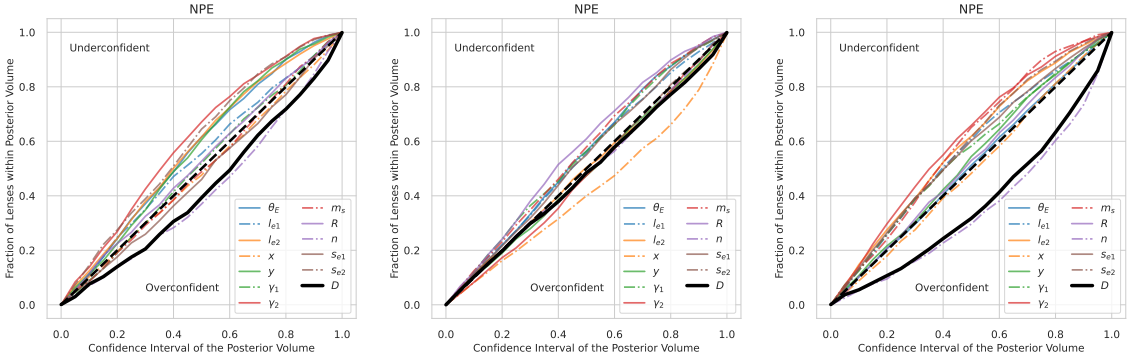

Figure 5.A.13 12-Parameter Model Posterior coverage for OOD Gaussian distribution test set 2 for seed 42, 465, 839

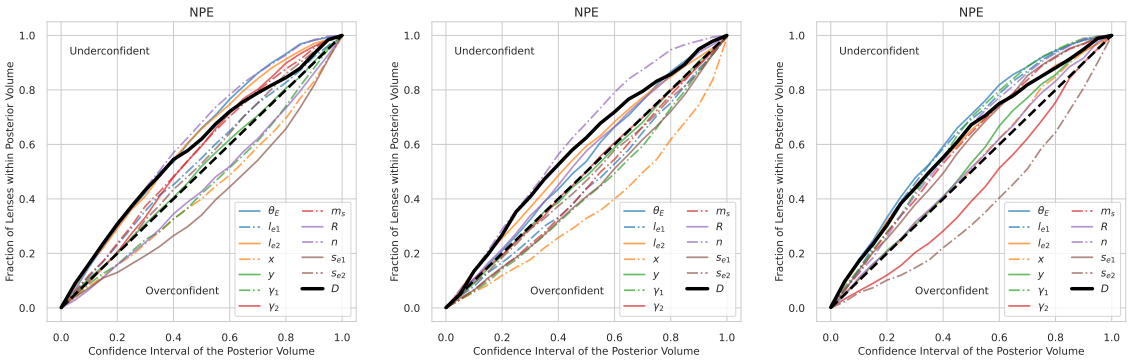

Figure 5.A.14 12 Parameter Model Posterior coverage for OOD Gaussian distribution test set 3 for seed 42, 465, 839

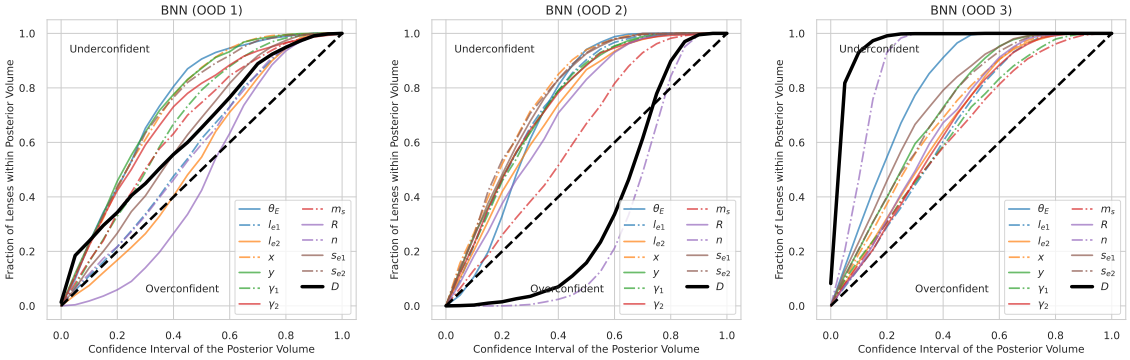

Figure 5.A.15 12 Parameter Model Posterior coverage for the BNN model for the three OOD Gaussian distribution test sets.

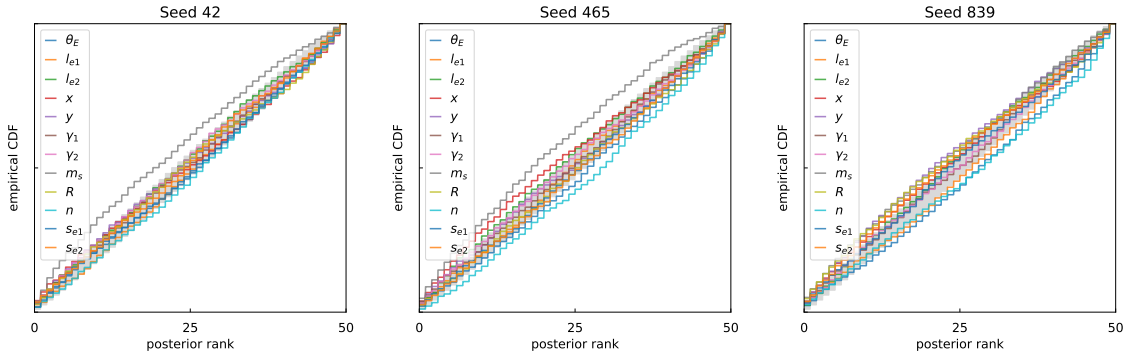

Figure 5.A.16 12-parameter model SBC CDF plots for seed 42, 465, 839. 5-parameter model SBC CDF plots for seed 42, 465, 839. The grey regions indicate the 99% confidence interval of a uniform distribution given the number of samples provided.

Figure 5.A.17 12-parameter model SBC histogram plots for seed 42, 465, 839. The grey regions indicate the 99% confidence interval of a uniform distribution given the number of samples provided. In a histogram, for a uniform distribution, we expect 1 out of 100 histogram bars to lie outside the confidence interval.

5.B Single Image Examples

5.B.1 1 parameter model

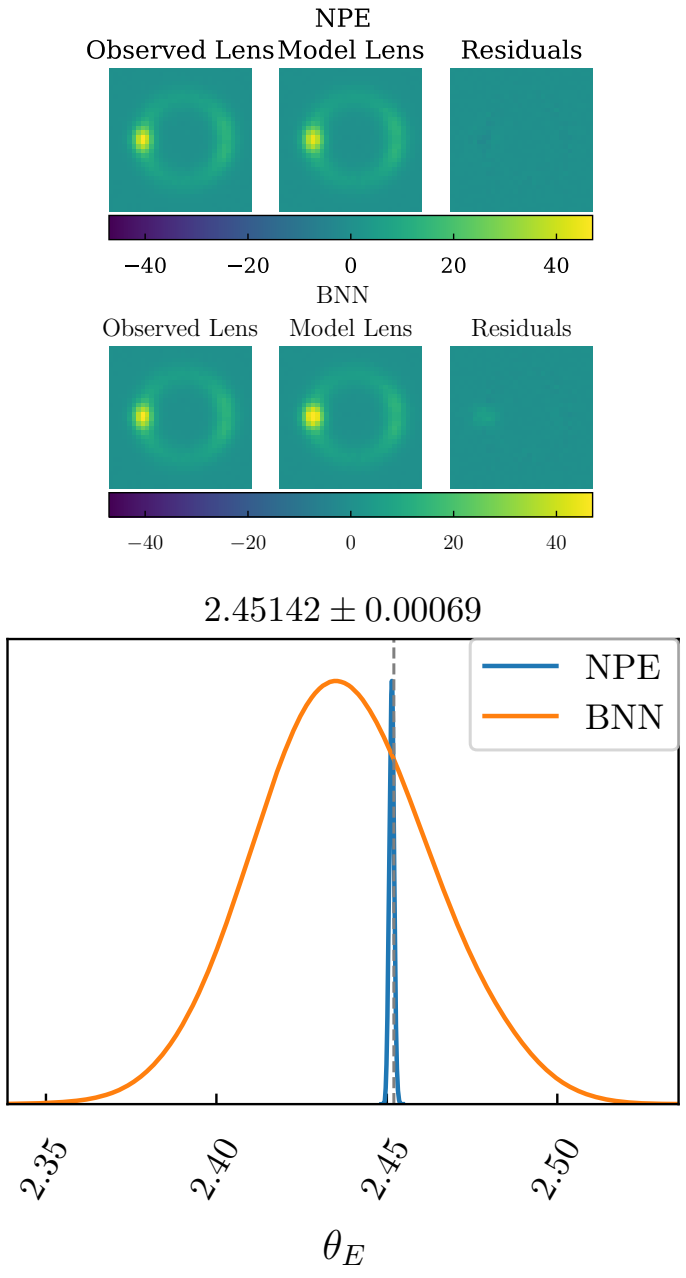

Figure 5.B.1 1-parameter model: Single Image Inference Example 1

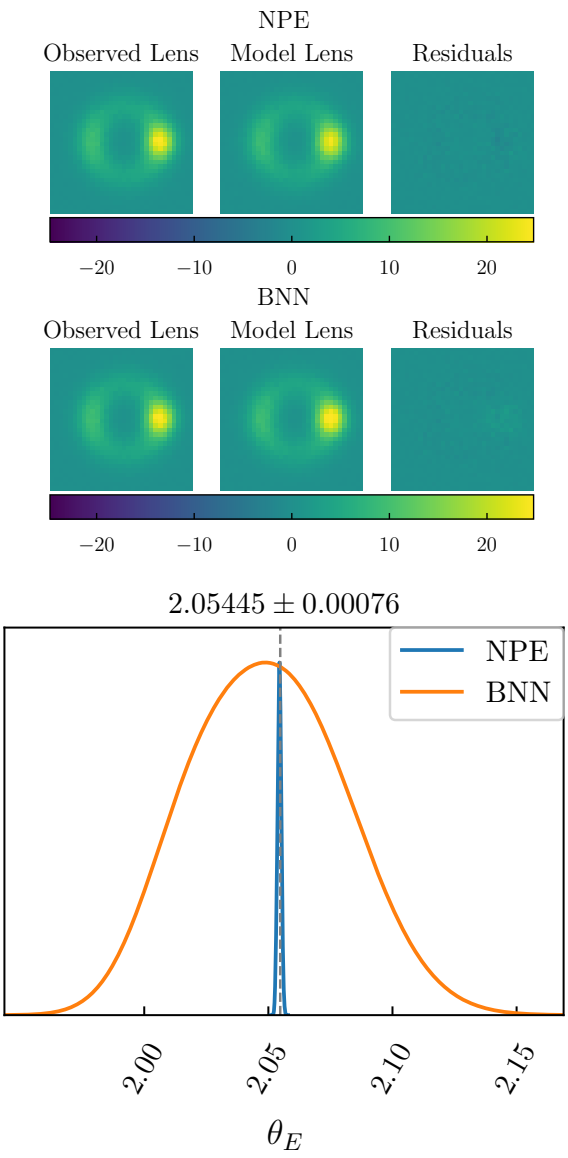

Figure 5.B.2 1-parameter model: Single Image Inference Example 2

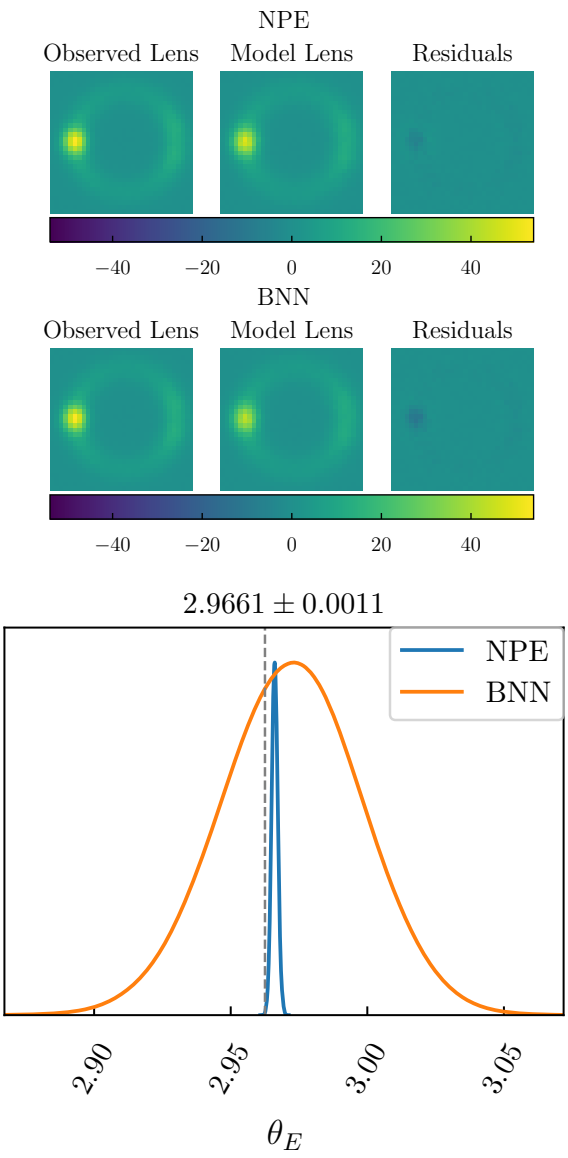

Figure 5.B.3 1-parameter model: Single Image Inference Example 3

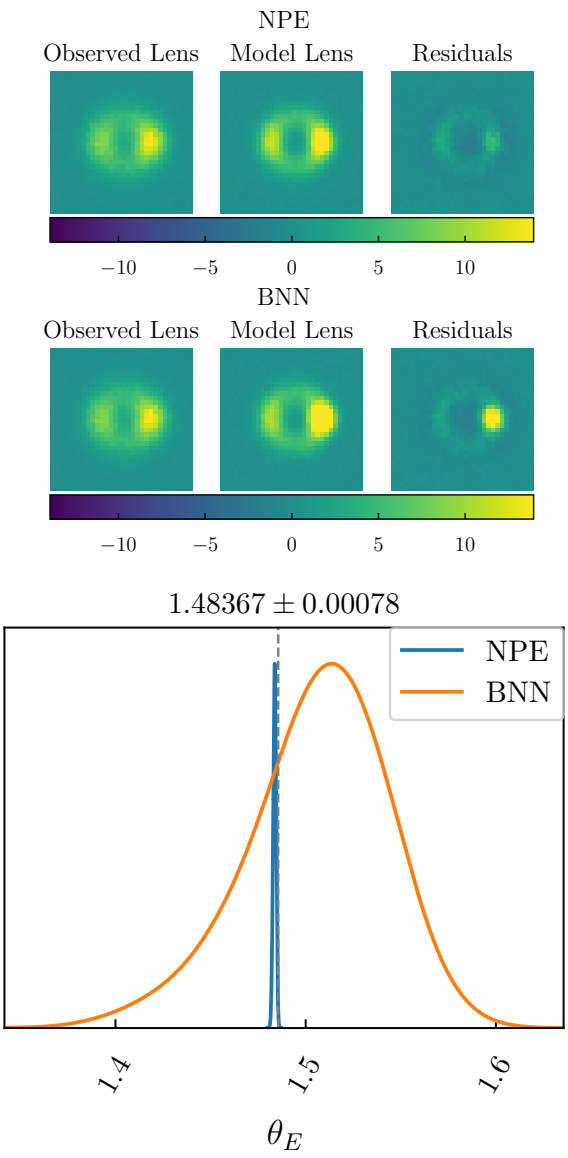

Figure 5.B.4 1-parameter model: Single Image Inference Example 4

5.B.2 5 parameter model

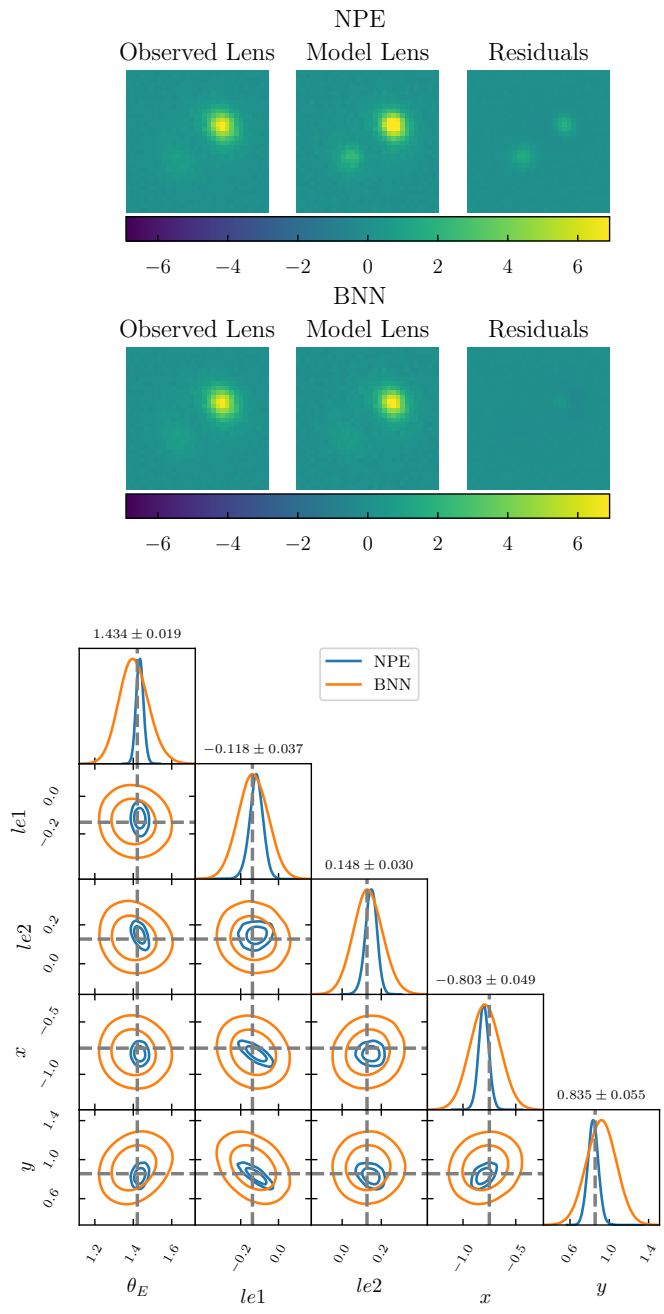

Figure 5.B.5 5-parameter model: Single Image Inference Example 1

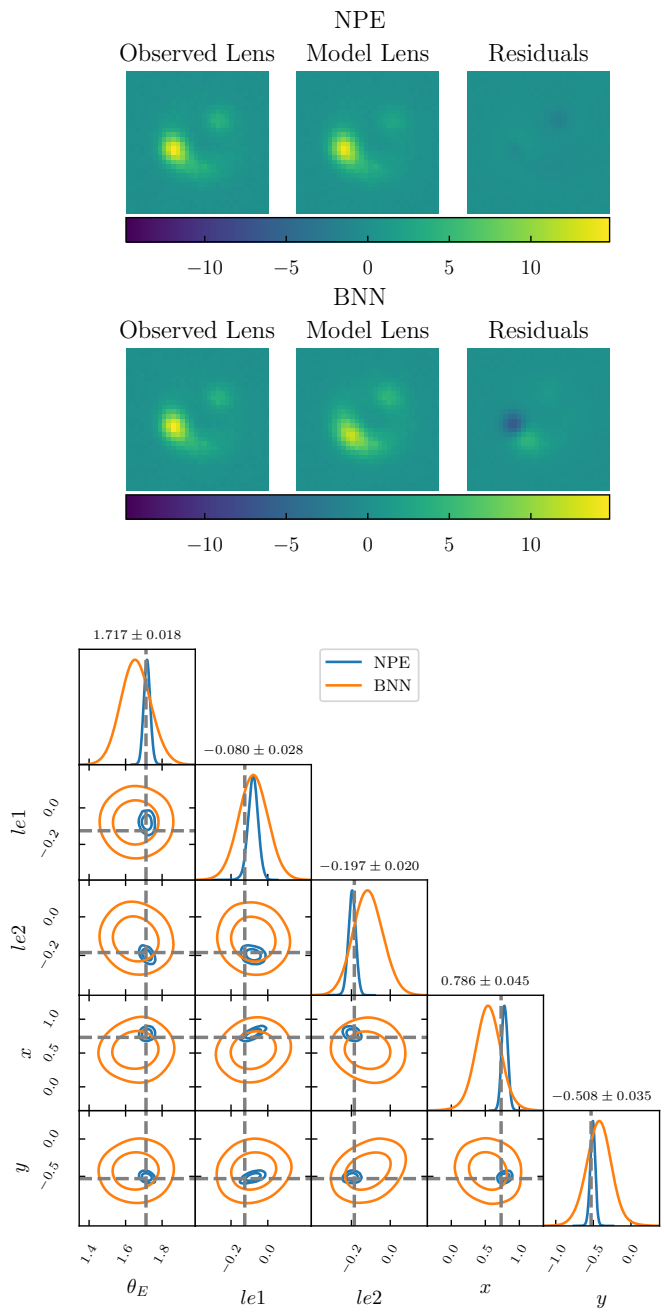

Figure 5.B.6 5-parameter model: Single Image Inference Example 2

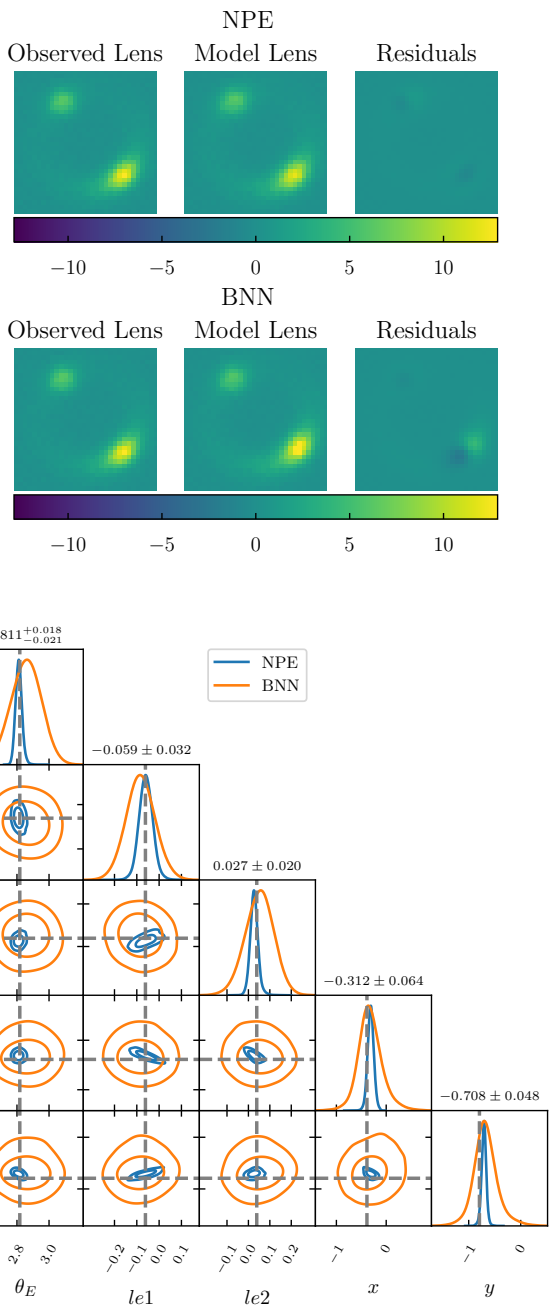

Figure 5.B.7 5-parameter model: Single Image Inference Example 3

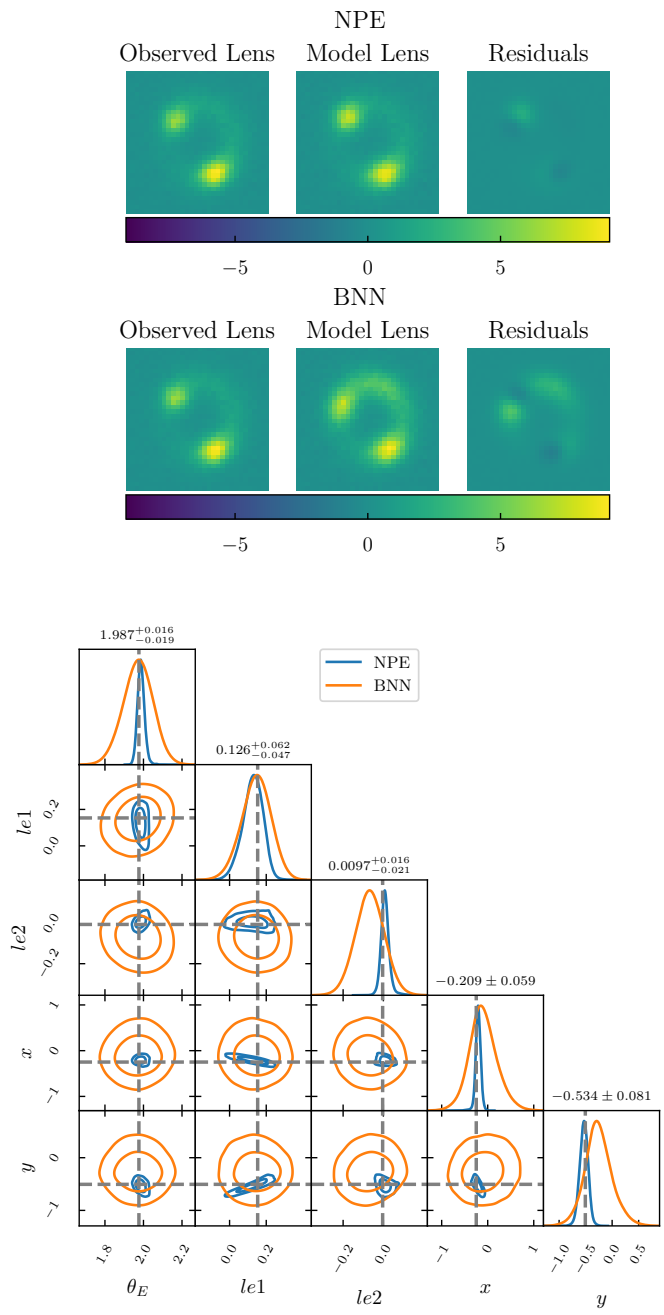

Figure 5.B.8 5-parameter model: Single Image Inference Example 4

5.B.3 12 parameter model

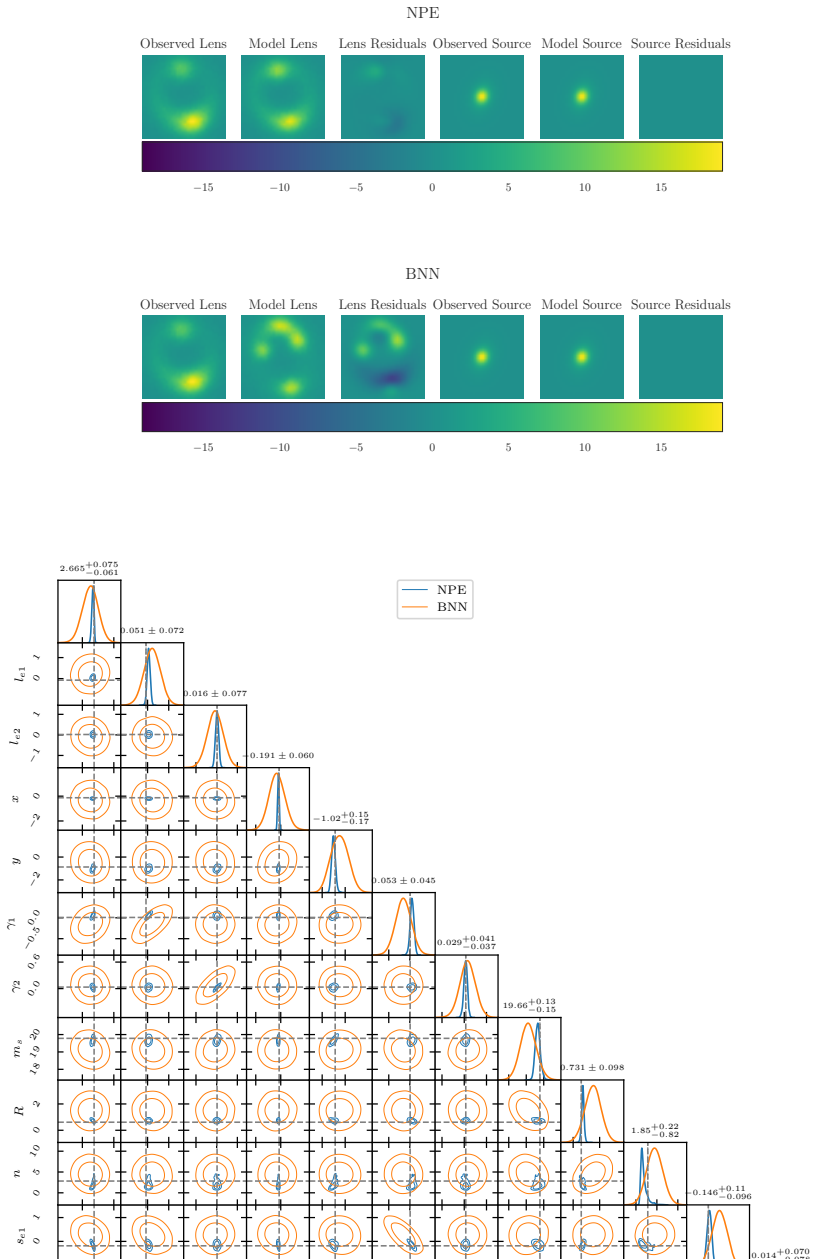

Figure 5.B.9 12-parameter model: Single Image Inference Example 1

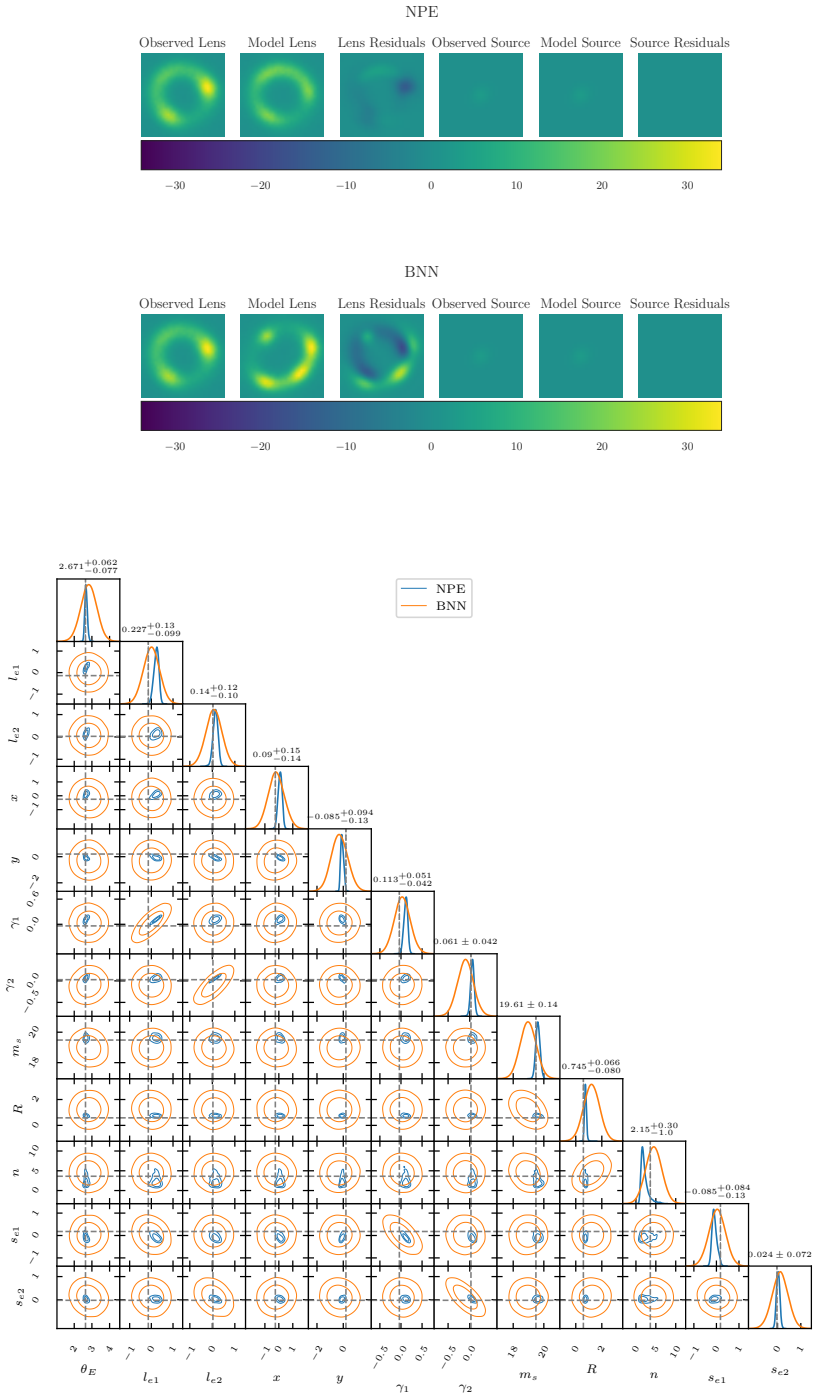

Figure 5.B.10 12-parameter model: Single Image Inference Example 2

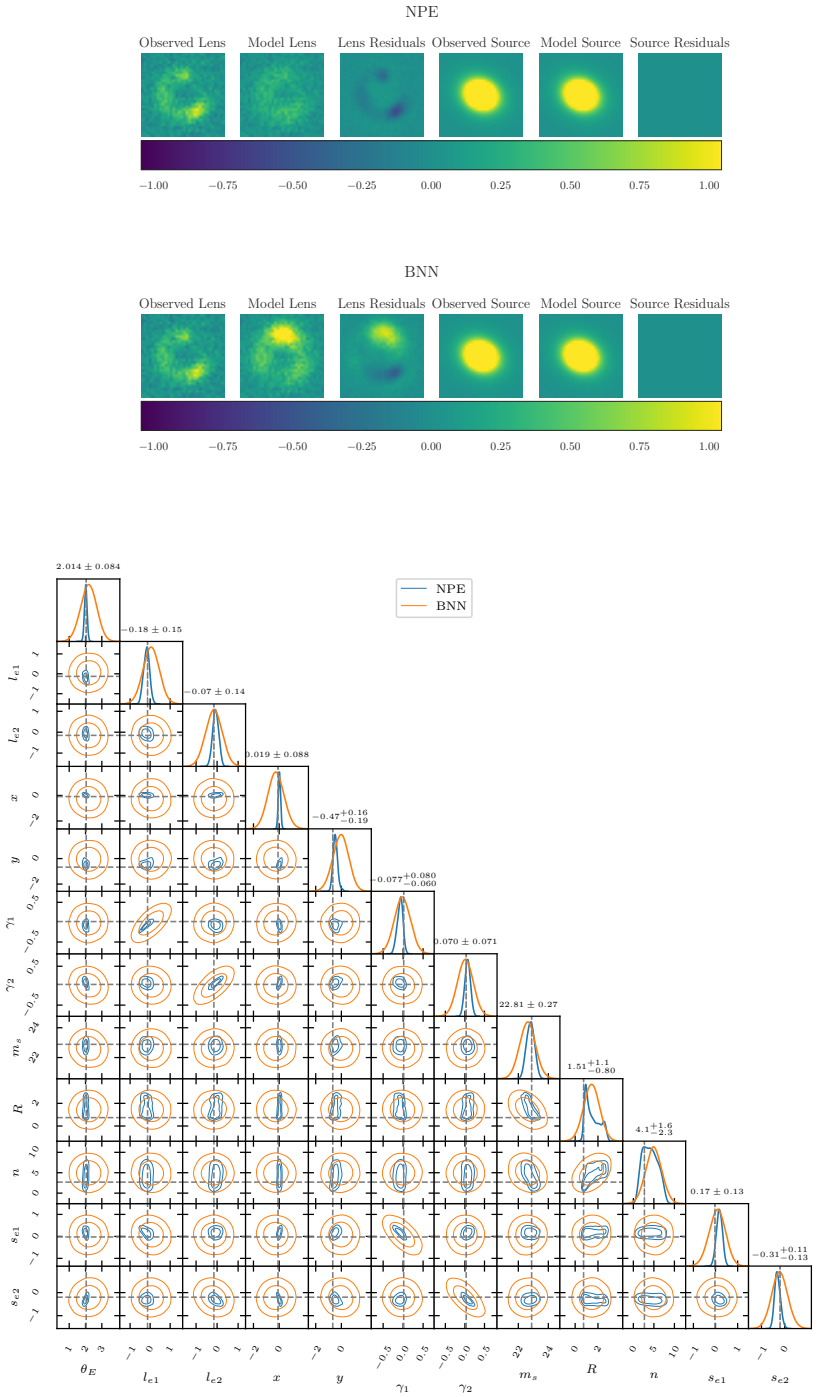

Figure 5.B.11 12-parameter model: Single Image Inference Example 3

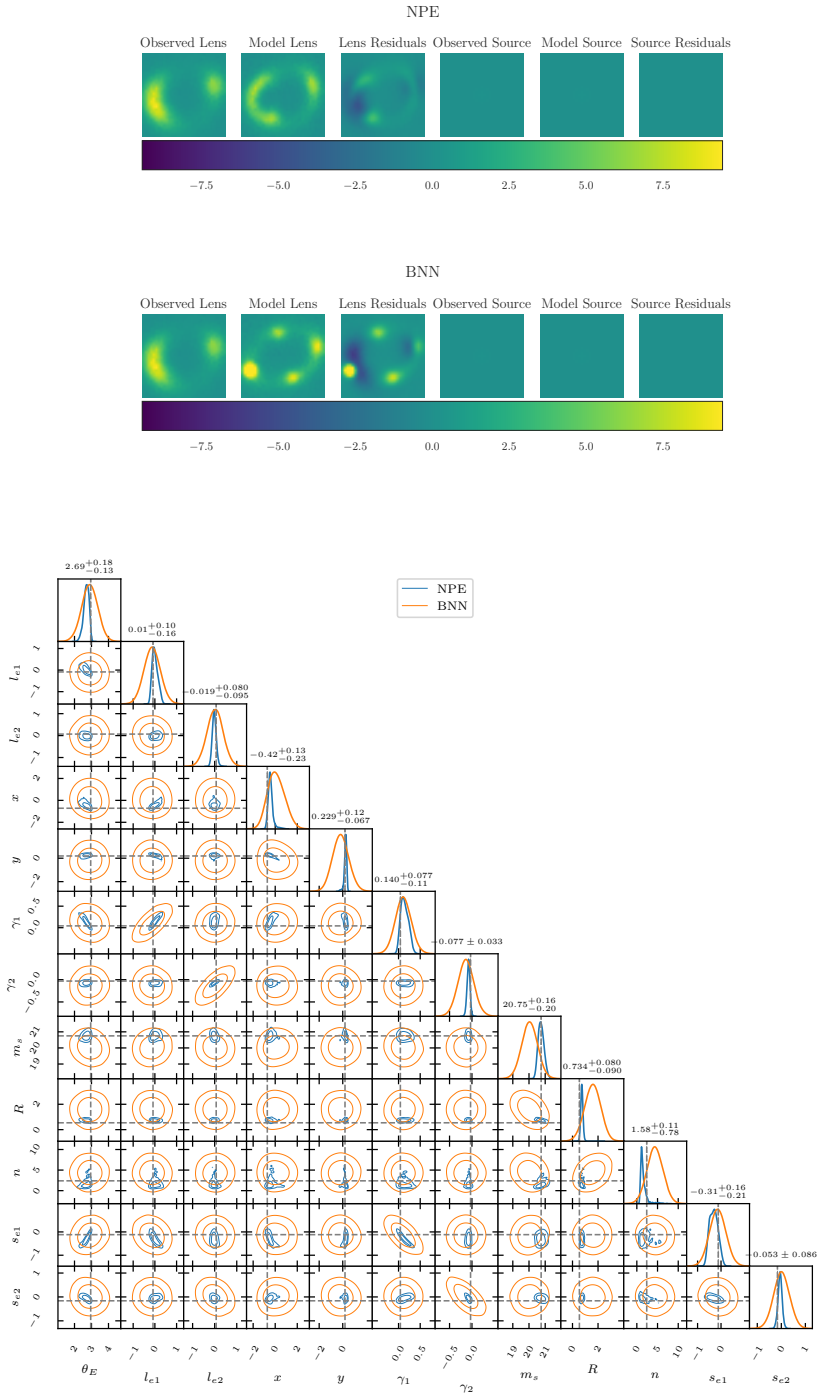

Figure 5.B.12 12-parameter model: Single Image Inference Example 4

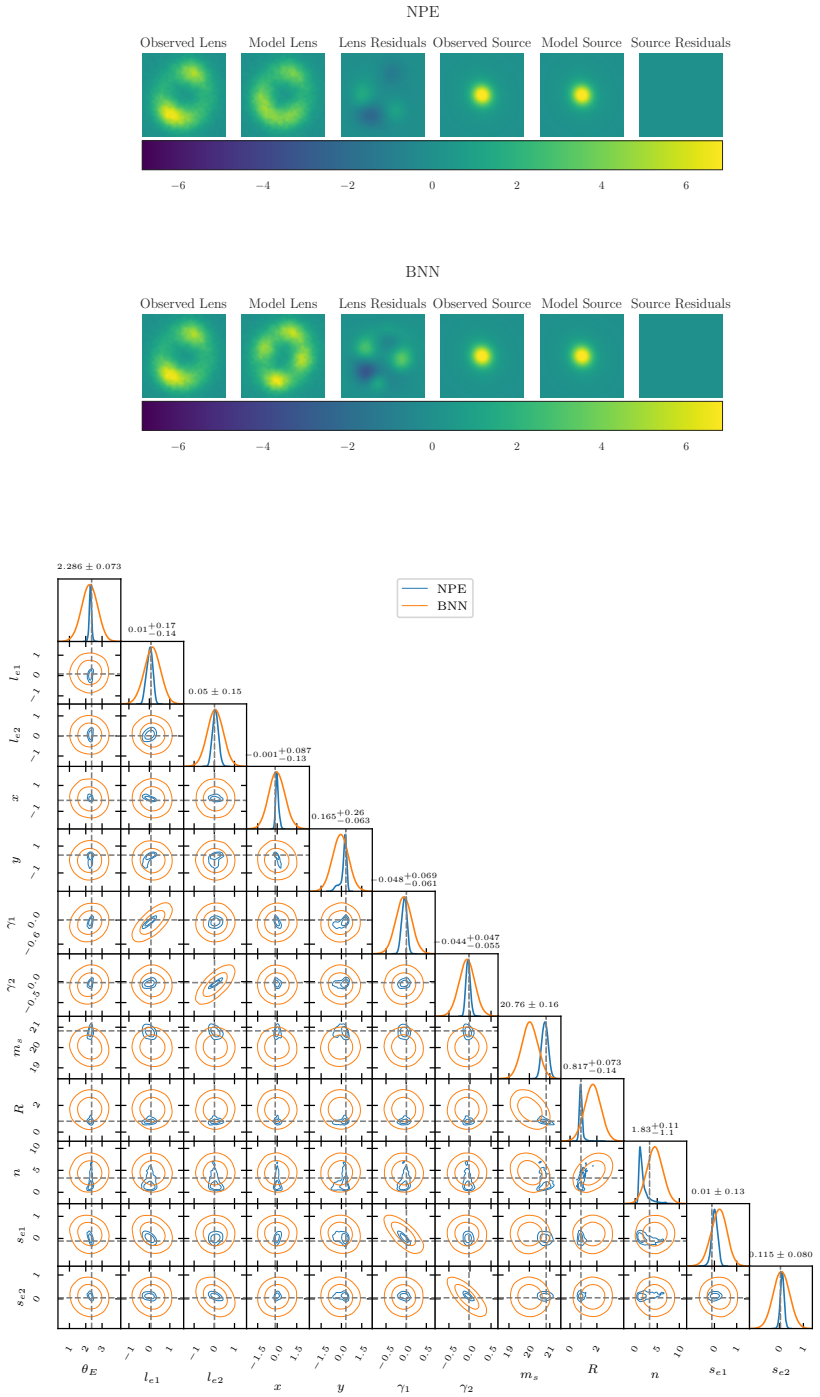

Figure 5.B.13 12-parameter model: Single Image Inference Example 5

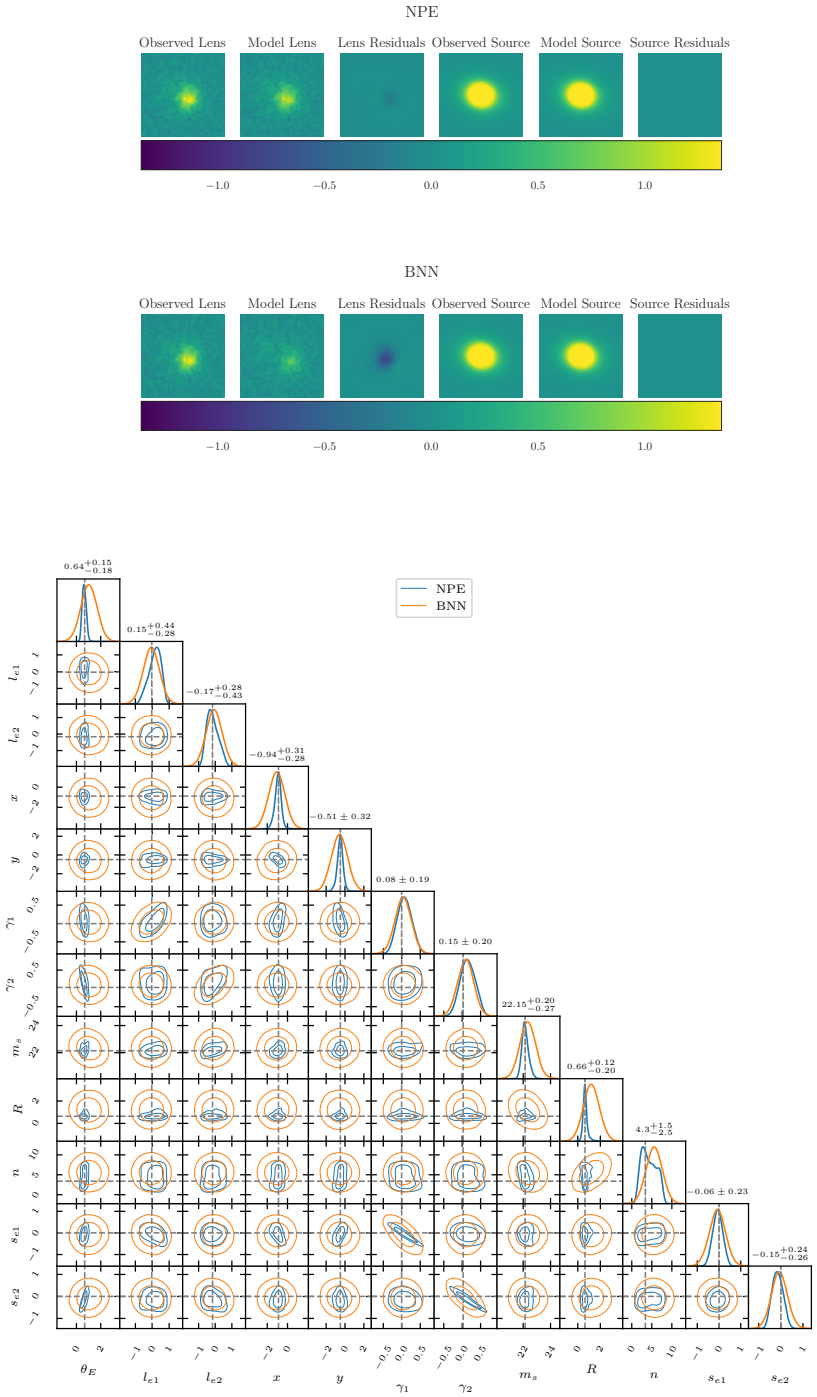

Figure 5.B.14 12-parameter model: Single Image Inference Example 6

REFERENCES

Abbott T. M. C., et al., 2018a, Astrophys. J. Suppl., 239, 18

Abbott T. M. C., et al., 2018b, Astrophys. J. Suppl., 239, 18

Abdel-Hamid O., Mohamed A.-r., Jiang H., Deng L., Penn G., Yu D., 2014, IEEE/ACM Transactions on audio, speech, and language processing, 22, 1533

Abdelsalam H. M., Saha P., Williams L. L. R., 1998, Mon. Not. Roy. Astron. Soc., 294, 734

Agnello A., et al., 2015, Monthly Notices of the Royal Astronomical Society, 454, 1260–1265

Agnello A., et al., 2017, Monthly Notices of the Royal Astronomical Society, 472, 4038

Agnello A., et al., 2018, Mon. Not. Roy. Astron. Soc., 479, 4345

Alsing J., Charnock T., Feeney S., Wandelt B., 2019a, Monthly Notices of the Royal Astronomical Society

Alsing J., Charnock T., Feeney S., Wandelt B., 2019b, Mon. Not. Roy. Astron. Soc., 488, 4440

Ammons S. M., Wong K. C., Zabludoff A. I., Keeton C. R., 2014, Astrophys. J., 781, 2

Amorisco N. C., Evans N. W., 2012, Mon. Not. Roy. Astron. Soc., 424, 1899

Anau Montel N., Coogan A., Correa C., Karchev K., Weniger C., 2023, Mon. Not. Roy. Astron. Soc., 518, 2746

Andreon S., Hurn M. A., 2010, Mon. Not. Roy. Astron. Soc., 404, 1922

Angelopoulos A. N., Bates S., 2022, A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification (`arXiv:2107.07511`)

Anguita T., et al., 2018, Mon. Not. Roy. Astron. Soc., 480, 5017

Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, Mon. Not. Roy. Astron. Soc., 310, 540

Auger M. W., Treu T., Brewer B. J., Marshall P. J., 2010a, Monthly Notices of the Royal Astronomical Society: Letters, Volume 411, Issue 1, pp. L6-L10., 411, L6

Auger M. W., Treu T., Bolton A. S., Gavazzi R., Koopmans L. V. E., Marshall P. J., Moustakas L. A., Burles S., 2010b, The Astrophysical Journal, Volume 724, Issue 1, pp. 511-525 (2010)., 724, 511

Auger M. W., Budzynski J. M., Belokurov V., Koposov S. E., McCarthy I. G., 2013, Monthly Notices of the Royal Astronomical Society, 436, 503

Avestruz C., Li N., Zhu H., Lightman M., Collett T. E., Luo W., 2019, The Astrophysical Journal, 877, 58

Balbinot E., et al., 2015, Monthly Notices of the Royal Astronomical Society, 449, 1129

Barkana R., 1998, The Astrophysical Journal, Volume 502, Issue 2, pp. 531-537., 502, 531

Barnabe M., Czoske O., Koopmans L. V. E., Treu T., Bolton A. S., 2011, Monthly Notices of the Royal Astronomical Society, Volume 415, Issue 3, pp. 2215-2232., 415, 2215

Bartelmann M., Huss A., Colberg J. M., Jenkins A., Pearce F. R., 1998, Astr. & Astroph., 330, 1

Bayer D., Chatterjee S., Koopmans L., Vegetti S., McKean J., Treu T., Fassnacht C., 2018, arXiv preprint arXiv:1803.05952

Bayliss M. B., 2012, Astrophys. J., 744, 156

Bayliss M. B., Rigby J. R., Sharon K., Wuyts E., Florian M., Gladders M. D., Johnson T., Oguri M., 2014, The Astrophysical Journal, Volume 790, Issue 2, article id. 144, 20 pp. (2014)., 790

Beaumont M. A., Zhang W., Balding D. J., 2002, Genetics, 162, 2025

Beers T. C., Flynn K., Gebhardt K., 1990, The Astronomical Journal, 100, 32

Belagiannis V., Rupprecht C., Carneiro G., Navab N., 2015, in Proceedings of the IEEE International Conference on Computer Vision. pp 2830–2838

Belokurov V., Evans N. W., Hewett P. C., Moiseev A., McMahon R. G., Sanchez S. F., King L. J., 2009, Mon. Not. Roy. Astron. Soc., 392, 104

Bengio Y., Courville A., Vincent P., 2014, Representation Learning: A Review and New Perspectives (`arXiv:1206.5538`)

Bertin E., 2011, in Astronomical Data Analysis Software and Systems XX. ASP Conference Proceedings, Vol. 442, proceedings of a Conference held at Seaport World Trade Center, Boston, Massachusetts, USA on 7-11 November 2010. Astronomical Society of the Pacific (ASP), p. 435

Bertin E., Arnouts S., 1996, Astronomy and Astrophysics Supplement Series, 117, 393

Bilir S., Karaali S., Tuncel S., 2005, Astronomische Nachrichten, Vol.326, Issue 5, p.321-331, 326, 321

Birrer S., Amara A., 2018, Physics of the Dark Universe, 22, 189

Birrer S., Amara A., Refregier A., 2016, Journal of Cosmology and Astroparticle Physics, 2016, 020

Birrer S., et al., 2021, The Journal of Open Source Software, 6, 3283

Blanton M. R., Roweis S., 2006, The Astronomical Journal, Volume 133, Issue 2, pp. 734-754., 133, 734

Blum M. G. B., Nunes M. A., Prangle D., Sisson S. A., 2013, Statistical Science, 28

Bocquet S., Carter F. W., 2016, The Journal of Open Source Software, 1

Bolton A. S., Burles S., Koopmans L. V., Treu T., Moustakas L. A., 2006, The Astrophysical Journal, 638, 703

Bolton A. S., Burles S., Koopmans L. V. E., Treu T., Gavazzi R., Moustakas L. A., Wayth R., Schlegel D. J., 2008, Astrophys. J., 682, 964

Bolton A. S., et al., 2012, The Astrophysical Journal, 757, 82

Bom C. R., Makler M., Albuquerque M. P., Brandt C. H., 2017, Astr. & Astroph., 597, A135

Bom C., Poh J., Nord B., Blanco-Valentin M., Dias L., 2019, Deep Learning in Wide-field Surveys: Fast Analysis of Strong Lenses in Ground-based Cosmic Experiments (arXiv:1911.06341)

Bonvin V., et al., 2017, Mon. Not. Roy. Astron. Soc., 465, 4914

Bradač M., et al., 2009, Astrophys. J., 706, 1201

Brehmer J., Cranmer K., Louppe G., Pavez J., 2018a, A Guide to Constraining Effective Field Theories with Machine Learning (arXiv:1805.00020)

Brehmer J., Cranmer K., Louppe G., Pavez J., 2018b, Physical Review Letters, 121

Brehmer J., Mishra-Sharma S., Hermans J., Louppe G., Cranmer K., 2019a, Astrophys. J., 886, 49

Brehmer J., Mishra-Sharma S., Hermans J., Louppe G., Cranmer K., 2019b, The Astrophysical Journal, 886, 49

Bruzual G., Charlot S., 2003, Mon. Not. Roy. Astron. Soc., 344, 1000

Buckley-Geer E. J., et al., 2020, Monthly Notices of the Royal Astronomical Society, 498, 3241

Burke D. L., Rykoff E. S., DES 2017, In preparation

Butter A., Finke T., Keil F., Krämer M., Manconi S., 2022, JCAP , 2022, 023

Cabanac R. A., et al., 2007, Astr. & Astroph., 461, 813

Caminha G. B., et al., 2015, arXiv:1512.05655,

Caminha G. B., et al., 2016, Astr. & Astroph., 587, A80

Cao S., Biesiada M., Gavazzi R., Piórkowska A., Zhu Z.-H., 2015a, The Astrophysical Journal, Volume 806, Issue 2, article id. 185, 12 pp. (2015)., 806

Cao S., Biesiada M., Gavazzi R., Piórkowska A., Zhu Z.-H., 2015b, The Astrophysical Journal, 806, 185

Cao S., Biesiada M., Gavazzi R., Pió rkowska A., Zhu Z.-H., 2015c, The Astrophysical Journal, 806, 185

Cao S., Li X., Biesiada M., Xu T., Cai Y., Zhu Z.-H., 2017, The Astrophysical Journal, 835, 92

Cardone V. F., Tortora C., Molinaro R., Salzano V., 2009, Astronomy & Astrophysics, 504, 769

Cardone V. F., Piedipalumbo E., Scudellaro P., 2015, Monthly Notices of the Royal Astronomical Society, Volume 455, Issue 1, p.831-837, 455, 831

Carrasco E. R., et al., 2010, Astrophys. J. Lett., 715, L160

Chabrier G., 2003, PASP, 115, 763

Charnock T., Lavaux G., Wandelt B. D., 2018, Physical Review D, 97

Chen Y., Li R., Shu Y., 2018, arXiv preprint arXiv:1809.09845

Chen Y., Li R., Shu Y., Cao X., 2019a, Monthly Notices of the Royal Astronomical Society, 488, 3745–3758

Chen G. C. F., et al., 2019b, Mon. Not. Roy. Astron. Soc., 490, 1743

Cheng T.-Y., Li N., Conselice C. J., Aragón-Salamanca A., Dye S., Metcalf R. B., 2019, arXiv e-prints, p. arXiv:1911.04320

Choi Y.-Y., Park C., Vogeley M. S., 2007, Astrophys. J., 658, 884

Choi K., Fazekas G., Sandler M., Cho K., 2017, in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 2392–2396

Christensen L., et al., 2012, Monthly Notices of the Royal Astronomical Society, Volume 427, Issue 3, pp. 1973-1982., 427, 1973

Ciotti L., Bertin G., 1999, Astr. & Astroph., 352, 447

Coe D., Fuselier E., Benítez N., Broadhurst T., Frye B., Ford H., 2008, Astrophys. J., 681, 814

Coe D., Benítez N., Broadhurst T., Moustakas L. A., 2010, Astrophys. J., 723, 1678

Collaboration T. L. D. E. S., et al., 2021, The LSST Dark Energy Science Collaboration (DESC) Science Requirements Document (`arXiv:1809.01669`)

Collett T. E., 2015a, Astrophys. J., 811, 20

Collett T. E., 2015b, Astrophys. J., 811, 20

Collett T. E., Auger M. W., 2014, Monthly Notices of the Royal Astronomical Society, Volume 443, Issue 2, p.969-976, 443, 969

Collett T. E., Cunnington S. D., 2016, Monthly Notices of the Royal Astronomical Society, Volume 462, Issue 3, p.3255-3264, 462, 3255

Collett T. E., Auger M. W., Belokurov V., Marshall P. J., Hall A. C., 2012, Monthly Notices of the Royal Astronomical Society, Volume 424, Issue 4, pp. 2864-2875., 424, 2864

Collett T. E., Buckley-Greer E., Lingren K. A., Nord B., Collaboration D., 2015, The Astrophysical Journal, Volume 811, Issue 1, article id. 20, 10 pp. (2015)., 811

Collett T. E., et al., 2023, The Messenger, 190, 49

Connolly A. J., et al., 2010, in Angeli G. Z., Dierickx P., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 7738, Modeling, Systems Engineering, and Project Management for Astronomy IV. p. 77381O, doi:10.1117/12.857819

Coogan A., Anau Montel N., Karchev K., Grootes M. W., Nattino F., Weniger C., 2022, arXiv e-prints, p. arXiv:2209.09918

Cook S., Gelman A., Rubin D., 2006, Journal of Computational and Graphical Statistics - J COMPUT GRAPH STAT, 15

Cooray A. R., 1999, Astr. & Astroph., 341, 653

Cranmer K., Brehmer J., Louppe G., 2019, arXiv e-prints, p. arXiv:1911.01429

Davis A. N., Huterer D., Krauss L. M., 2003, Monthly Notices of the Royal Astronomical Society, 344, 1029

Deistler M., Goncalves P. J., Macke J. H., 2022, Truncated proposals for scalable and hassle-free simulation-based inference (`arXiv:2210.04815`)

Depeweg S., Hernández-Lobato J. M., Doshi-Velez F., Udluft S., 2017, arXiv e-prints, p. arXiv:1710.07283

Desai S., et al., 2012, The Astrophysical Journal, Volume 757, Issue 1, article id. 83, 22 pp. (2012)., 757

Despali G., Vegetti S., White S. D. M., Giocoli C., van den Bosch F. C., 2018, Mon. Not. Roy. Astron. Soc., 475, 5424

Dessauges-Zavadsky M., D'Odorico S., Schaerer D., Modigliani A., Tapken C., Vernet J., 2009, Astronomy and Astrophysics, Volume 510, id.A26, 510

Diego J. M., Protopapas P., Sandvik H. B., Tegmark M., 2005, Mon. Not. Roy. Astron. Soc., 360, 477

Diehl H. T., 2020, doi:10.2172/1596042

Diehl H. T., Buckley-Greer E., Lingren K. A., Nord B., Collaboration D., 2017, In preparation

Dodelson S., 2017, Gravitational Lensing. Cambridge University Press

Dressler A., et al., 2011, PASP, 123, 288

Drlica-Wagner A., et al., 2018, Astrophys. J. Suppl., 235, 33

Durkan C., Murray I., Papamakarios G., 2020, On Contrastive Learning for Likelihood-free Inference (`arXiv:2002.03712`)

Ebeling H., Stockmann M., Richard J., Zabl J., Brammer G., Toft S., Man A., 2018, Astrophys. J., 852, L7

Enander J., Mörtsell E., 2013, Journal of High Energy Physics, 2013, 1

Escamilla-Rivera C., Carvajal M., Zamora C., Hendry M., 2022, JCAP , 2022, 016

Estrada J., et al., 2007, Astrophys. J., 660, 1176

Fassnacht C. D., Moustakas L. A., Casertano S., Ferguson H. C., Lucas R. A., Park Y., 2004, Astrophys. J. Lett., 600, L155

Faure C., et al., 2008, Astrophys. J. Suppl., 176, 19, erratum 2008, 178, 382

Finkelstein S. L., Papovich C., Rudnick G., Egami E., Floc'h E. L., Rieke M. J., Rigby J., Willmer C. N. A., 2009, The Astrophysical Journal, Volume 700, Issue 1, pp. 376-386 (2009)., 700, 376

Flaugher B., et al., 2015, The Astronomical Journal, Volume 150, Issue 5, article id. 150, 43 pp. (2015)., 150

Foreman-Mackey D., 2016, The Journal of Open Source Software, 1, 24

Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2012, Publications of the Astronomical Society of Pacific, Volume 125, Issue 925, pp. 306 (2013)., 125, 306

Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306

Fox C. W., Roberts S. J., 2012, Artificial intelligence review, 38, 85

Furlanetto C., et al., 2013, Mon. Not. Roy. Astron. Soc., 432, 73

Gal Y., 2016, PhD thesis, PhD thesis, University of Cambridge

Gal Y., Ghahramani Z., 2016, in international conference on machine learning. pp 1050–1059

Gal Y., Hron J., Kendall A., 2017, in Advances in Neural Information Processing Systems. pp 3581–3590

Gavazzi R., Treu T., Koopmans L. V. E., Bolton A. S., Moustakas L. A., Burles S., Marshall P. J., 2008, The Astrophysical Journal, Volume 677, Issue 2, article id. 1046-1059, pp. (2008)., 677

Gavazzi R., Marshall P. J., Treu T., Sonnenfeld A., 2014a, The Astrophysical Journal, Volume 785, Issue 2, article id. 144, 14 pp. (2014)., 785

Gavazzi R., Marshall P. J., Treu T., Sonnenfeld A., 2014b, Astrophys. J., 785, 144

Gerardi F., Feeney S. M., Alsing J., 2021, Phys. Rev. D, 104, 083531

Germain M., Gregor K., Murray I., Larochelle H., 2015, in Bach F., Blei D., eds, Proceedings of Machine Learning Research Vol. 37, Proceedings of the 32nd International Conference on Machine Learning. PMLR, Lille, France, pp 881–889, https://proceedings.mlr.press/v37/germain15.html

Gilman D., Birrer S., Treu T., Keeton C. R., Nierenberg A., 2018, Monthly Notices of the Royal Astronomical Society, 481, 819

Gladders M. D., Hoekstra H., Yee H. K. C., Hall P. B., Barrientos L. F., 2003, Astrophys. J., 593, 48

Glazebrook K., Jacobs C., Collett T., More A., McCarthy C., 2017, Monthly Notices of the Royal Astronomical Society, 471, 167

Golse G., Kneib J.-P., Soucail G., 2002, Astr. & Astroph., 387, 788

Goncalves P., et al., 2020, eLife, 9

Goobar A., et al., 2016, arXiv:1611.00014,

Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press

Goodman J., Weare J., 2010, Communications in Applied Mathematics and Computational Science, Vol.˜5, No.˜1, p.˜65-80, 2010, 5, 65

Graves A., 2011a, in Shawe-Taylor J., Zemel R. S., Bartlett P. L., Pereira F., Weinberger K. Q., eds, , Advances in Neural Information Processing Systems 24. Curran Associates, Inc., pp 2348–2356, http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf

Graves A., 2011b, in Shawe-Taylor J., Zemel R., Bartlett P., Pereira F., Weinberger K., eds, Vol. 24, Advances in Neural Information Processing Systems. Curran Associates, Inc., https://proceedings.neurips.cc/paper/2011/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf

Green J., et al., 2012, arXiv preprint arXiv:1208.4012

Greenberg D., Nonnenmacher M., Macke J., 2019, in International Conference on Machine Learning. pp 2404–2414

Grillo C., Lombardi M., Bertin G., 2007, Astronomy and Astrophysics, Volume 477, Issue 2, January II 2008, pp.397-406, 477, 397

Guérou A., et al., 2017, Astr. & Astroph., 608, A5

Guo C., Pleiss G., Sun Y., Weinberger K. Q., 2017, On Calibration of Modern Neural Networks (arXiv:1706.04599)

Hainline K. N., Shapley A. E., Kornei K. A., Pettini M., Buckley-Geer E., Allam S. S., Tucker D. L., 2009, Astrophys. J., 701, 52

Hannun A. Y., Rajpurkar P., Haghpanahi M., Tison G. H., Bourn C., Turakhia M. P., Ng A. Y., 2019, Nature medicine, 25, 65

He K., Zhang X., Ren S., Sun J., 2015, in Proceedings of the IEEE international conference on computer vision. pp 1026–1034

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

Hennawi J. F., et al., 2008, Astron. J., 135, 664

Hermans J., Delaunoy A., Rozet F., Wehenkel A., Begy V., Louppe G., 2022, A Trust Crisis In Simulation-Based Inference? Your Posterior Approximations Can Be Unfaithful (arXiv:2110.06581)

Hezaveh Y., Dalal N., Holder G., Kisner T., Kuhlen M., Levasseur L. P., 2014, arXiv preprint arXiv:1403.2720

Hezaveh Y. D., Levasseur L. P., Marshall P. J., 2017a, Nature, 548, 555

Hezaveh Y. D., Levasseur L. P., Marshall P. J., 2017b, Nature, 548, 555

Hogg D. W., Blandford R., Kundic T., Fassnacht C. D., Malhotra S., 1996, Astrophys. J. Lett., 467, L73

Hogg D. W., Baldry I. K., Blanton M. R., Eisenstein D. J., 2002, The K correction (arXiv:astro-ph/0210394)

Hortúa H. J., Volpi R., Marinelli D., Malagò L., 2020, Physical Review D, 102

Huang K.-W., Chen G. C.-F., Chang P.-W., Lin S.-C., Hsu C.-J., Thengane V., Lin J. Y.-Y., 2022, arXiv preprint arXiv:2210.04143

Huppenkothen D., Bachetti M., 2021, arXiv e-prints, p. arXiv:2104.03278

Ilbert O., et al., 2006, Astr. & Astroph., 457, 841

Ivezić Ž., et al., 2008, Astrophys. J., 684, 287

Ivezić Ž., et al., 2019, ApJ, 873, 111

Iwata I., et al., 2008, The Astrophysical Journal, Volume 692, Issue 2, pp. 1287-1293 (2009)., 692, 1287

Jackson N., 2008, Mon. Not. Roy. Astron. Soc., 389, 1311

Jacobs C., Glazebrook K., Collett T., More A., McCarthy C., 2017, Monthly Notices of the Royal Astronomical Society, 471, 167–181

Jacobs C., et al., 2019a, The Astrophysical Journal Supplement Series, 243, 17

Jacobs C., et al., 2019b, Mon. Not. Roy. Astron. Soc., 484, 5330

Jacobs C., et al., 2019c, Mon. Not. Roy. Astron. Soc., 484, 5330

James B. L., et al., 2013, Monthly Notices of the Royal Astronomical Society, Volume 440, Issue 2, p.1794-1809, 440, 1794

Jee I., Komatsu E., Suyu S., Huterer D., 2016, Journal of Cosmology and Astroparticle Physics, 2016, 031–031

Jee I., Suyu S. H., Komatsu E., Fassnacht C. D., Hilbert S., Koopmans L. V. E., 2019, Science, 365, 1134–1138

John V., Mita S., Liu Z., Qi B., 2015, in 2015 14th IAPR International Conference on Machine Vision Applications (MVA). pp 246–249

Johnson T. L., et al., 2017, Astrophys. J. Lett., 843, L21

Jones T. A., Swinbank A. M., Ellis R. S., Richard J., Stark D. P., 2010, Mon. Not. Roy. Astron. Soc., 404, 1247

Jones T., Ellis R. S., Schenker M. A., Stark D. P., 2013, The Astrophysical Journal, Volume 779, Issue 1, article id. 52, 8 pp. (2013)., 779

Jordan M. I., Ghahramani Z., Jaakkola T. S., Saul L. K., 1999, Machine learning, 37, 183

Jospin L. V., Laga H., Boussaid F., Buntine W., Bennamoun M., 2022, IEEE Computational Intelligence Magazine, 17, 29

Jullo E., Kneib J.-P., Limousin M., Elíasdóttir Á., Marshall P. J., Verdugo T., 2007, New Journal of Physics, 9, 447

Jullo E., Natarajan P., Kneib J.-P., D'Aloisio A., Limousin M., Richard J., Schimd C., 2010, Science, 329, 924

Kassiola A., Kovner I., 1993, The Astrophysical Journal, 417, 450

Kausch W., Schindler S., Erben T., Wambsganss J., Schwope A., 2010, Astr. & Astroph., 513, A8

Keeton C. R., Kochanek C. S., 1997, Astrophys. J., 487, 42

Kelly P. L., et al., 2015, Science, 347, 1123

Kendall A., Gal Y., 2017a, in arXiv:1703.04977.

Kendall A., Gal Y., 2017b, in Advances in neural information processing systems. pp 5574–5584

Khullar G., et al., 2021, The Astrophysical Journal, 906, 107

Kingma D. P., Ba J., 2017, Adam: A Method for Stochastic Optimization (arXiv:1412.6980)

Knabel S., et al., 2023, Monthly Notices of the Royal Astronomical Society, 520, 804

Kneib J.-P., Bonnet H., Golse G., Sand D., Jullo E., Marshall P., 2011, LENSTOOL: A Gravitational Lensing Software for Modeling Mass Distribution of Galaxies and Clusters (strong and weak regime), Astrophysics Source Code Library, record ascl:1102.004

Kochanek C. S., Schechter P. L., 2004, in Freedman W. L., ed., Measuring and Modeling the Universe. p. 117

Koleva M., Prugniel P., Bouchard A., Wu Y., 2009, Astr. & Astroph., 501, 1269

Koopmans L. V. E., Treu T., Bolton A. S., Burles S., Moustakas L. A., 2006a, Astrophys. J., 649, 599

Koopmans L. V. E., Treu T., Bolton A. S., Burles S., Moustakas L. A., 2006b, The Astrophysical Journal, Volume 649, Issue 2, pp. 599-615., 649, 599

Koposov S. E., et al., 2011, Astrophys. J., 736, 146

Kormann R., Schneider P., Bartelmann M., 1994, Astronomy and Astrophysics (ISSN 0004-6361), vol. 284, no. 1, p. 285-299, 284, 285

Kovner I., 1989, Astrophys. J., 337, 621

Krizhevsky A., Sutskever I., Hinton G. E., 2012, in Advances in neural information processing systems. pp 1097–1105

Kubo J. M., Dell'Antonio I. P., 2008, Mon. Not. Roy. Astron. Soc., 385, 918

Kubo J. M., et al., 2010, Astrophys. J. Lett., 724, L137

Kull M., Filho T. S., Flach P., 2017, in Singh A., Zhu J., eds, Proceedings of Machine Learning Research Vol. 54, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. PMLR, pp 623–631, https://proceedings.mlr.press/v54/kull17a.html

Kullback S., Leibler R. A., 1951, The annals of mathematical statistics, 22, 79

Küng R., et al., 2015, Monthly Notices of the Royal Astronomical Society, Volume 447, Issue 3, p.2170-2180, 447, 2170

Kurtz M. J., Mink D. J., 1998, Publications of the Astronomical Society of the Pacific, 110, 934

Lanusse F., Ma Q., Li N., Collett T. E., Li C.-L., Ravanbakhsh S., Mandelbaum R., Póczos B., 2017, Monthly Notices of the Royal Astronomical Society, 473, 3895–3906

Lanusse F., Ma Q., Li N., Collett T. E., Li C.-L., Ravanbakhsh S., Mandelbaum R., Póczos B., 2018, Mon. Not. Roy. Astron. Soc., 473, 3895

Laporte C. F. P., Agnello A., Navarro J. F., 2019, Mon. Not. Roy. Astron. Soc., 484, 245

Lathuilière S., Mesejo P., Alameda-Pineda X., Horaud R., 2018a, arXiv preprint arXiv:1803.08450

Lathuilière S., Mesejo P., Alameda-Pineda X., Horaud R., 2018b, in Proceedings of the European Conference on Computer Vision (ECCV). pp 202–217

Laureijs R., et al., 2011, arXiv preprint arXiv:1110.3193

LeCun Y., Bottou L., Bengio Y., Haffner P., et al., 1998, Proceedings of the IEEE, 86, 2278

LeCun Y., Bengio Y., Hinton G., 2015, nature, 521, 436

Leaf K., Melia F., 2018, Monthly Notices of the Royal Astronomical Society, 478, 5104

Lee J., Bahri Y., Novak R., Schoenholz S. S., Pennington J., Sohl-Dickstein J., 2017, arXiv preprint arXiv:1711.00165

Lefor A. T., Futamase T., Akhlaghi M., 2013, New Astronomy Reviews, 57, 1–13

Legin R., Hezaveh Y., Perreault Levasseur L., Wandelt B., 2021, arXiv e-prints, p. arXiv:2112.05278

Levasseur L. P., Hezaveh Y. D., Wechsler R. H., 2017, The Astrophysical Journal Letters, 850, L7

Lewis A., 2019, GetDist: a Python package for analysing Monte Carlo samples (arXiv:1910.13970)

Li et al N., 2017, In preparation

Li R., Shu Y., Wang J., 2018a, arXiv preprint arXiv:1805.06624

Li X., Ding Q., Sun J.-Q., 2018b, Reliability Engineering and System Safety, 172, 1

Li Z., Liu F., Yang W., Peng S., Zhou J., 2022, IEEE Transactions on Neural Networks and Learning Systems, 33, 6999

Li T., Collett T. E., Krawczyk C. M., Enzi W., 2023, Cosmology from Large Populations of Galaxy-Galaxy Strong Gravitational Lenses (arXiv:2307.09271)

Lin H., et al., 2017, Astrophys. J. Lett., 838, L15

Lin P., Li X., Chen Y., He Y., 2018, Food and Bioprocess Technology, 11, 765

Lin Z., Trivedi S., Sun J., 2021, Locally Valid and Discriminative Prediction Intervals for Deep Learning Models (arXiv:2106.00225)

Lu J., Wang G., Zhou J., 2017, IEEE Transactions on Image Processing, 26, 4042

Lueckmann J.-M., Bassetto G., Karaletsos T., Macke J. H., 2018, arXiv e-prints, p. arXiv:1805.09294

Lueckmann J.-M., Boelts J., Greenberg D. S., Gonçalves P. J., Macke J. H., 2021, Benchmarking Simulation-Based Inference (arXiv:2101.04653)

Luppino G. A., Gioia I. M., Hammer F., Le Fèvre O., Annis J. A., 1999, Astr. & Astroph., Supp., 136, 117

Lupton R., Blanton M. R., Fekete G., Hogg D. W., O'Mullane W., Szalay A., Wherry N., 2004, PASP, 116, 133

Maddison C. J., Mnih A., Teh Y. W., 2016, arXiv preprint arXiv:1611.00712

Madireddy S., Li N., Ramachandra N., Balaprakash P., Habib S., 2019, Modular Deep Learning Analysis of Galaxy-Scale Strong Lensing Images (arXiv:1911.03867)

Magaña J., Motta V., Cárdenas V. H., Verdugo T., Jullo E., 2015, Astrophys. J., 813, 69

Mahler G., et al., 2023, Precision modeling of JWST's first cluster lens SMACSJ0723.3-7327 (arXiv:2207.07101)

Mancini A. S., Docherty M. M., Price M. A., McEwen J. D., 2022, Bayesian model comparison for simulation-based inference (arXiv:2207.04037)

Mao S., 2012, Research in Astronomy and Astrophysics, 12, 947–972

Marshall P. J., 2009, in Bulletin of the American Astronomical Society. p. 377

Marshall P. J., et al., 2007, Astrophys. J., 671, 1196

Marshall P. J., Hogg D. W., Moustakas L. A., Fassnacht C. D., Bradač M., Schrabback T., Blandford R. D., 2009, Astrophys. J., 694, 924

Marshall P. J., et al., 2015, Monthly Notices of the Royal Astronomical Society, Volume 455, Issue 2, p.1171-1190, 455, 1171

Martin N. F., Collins M. L. M., Longeard N., Tollerud E., 2018, Astrophys. J. Lett., 859, L5

Maturi M., Mizera S., Seidel G., 2014, Astr. & Astroph., 567, A111

McCully C., Keeton C. R., Wong K. C., Zabludoff A. I., 2014, Mon. Not. Roy. Astron. Soc., 443, 3631

McCully C., Keeton C. R., Wong K. C., Zabludoff A. I., 2017a, Astrophys. J., 836, 141

McCully C., Keeton C. R., Wong K. C., Zabludoff A. I., 2017b, The Astrophysical Journal, 836, 141

Meneghetti M., Dolag K., Tormen G., Bartelmann M., Moscardini L., Perrotta F., Baccigalupi C., 2004, Modern Physics Letters A, 19, 1083

Meneghetti M., Jain B., Bartelmann M., Dolag K., 2005, Mon. Not. Roy. Astron. Soc., 362, 1301

Meng X.-L., Treu T., Agnello A., Auger M. W., Liao K., Marshall P. J., 2015, Journal of Cosmology and Astroparticle Physics, Issue 09, article id. 059, (2015)., 9

Metcalf R. B., Petkova M., 2014, Mon. Not. Roy. Astron. Soc., 445, 1942

Metcalf R. B., et al., 2018, arXiv e-prints, p. arXiv:1802.03609

Millon M., et al., 2020a, Astronomy & Astrophysics, 639, A101

Millon M., et al., 2020b, Astronomy & Astrophysics, 640, A105

Mohr J. J., et al., 2012, Software and Cyberinfrastructure for Astronomy II. Proceedings of the SPIE, Volume 8451, article id. 84510D, 12 pp. (2012)., 8451

Mollerach S., Roulet E., 2002, Gravitational Lensing and Microlensing. World Scientific, https://books.google.com.br/books?id=PAErrkpBYG0C

More A., Cabanac R., More S., Alard C., Limousin M., Kneib J.-P., Gavazzi R., Motta V., 2012a, Astrophys. J., 749, 38

More A., Cabanac R., More S., Alard C., Limousin M., Kneib J.-P., Gavazzi R., Motta V., 2012b, The Astrophysical Journal, 749, 38

More A., et al., 2015, Monthly Notices of the Royal Astronomical Society, Volume 455, Issue 2, p.1191-1210, 455, 1191

More A., et al., 2016, Mon. Not. Roy. Astron. Soc., 455, 1191

Morgan R., Nord B., Birrer S., Lin J., Poh J., 2021, The Journal of Open Source Software, 6, 2854

Morganson E., et al., 2018, Publications of the Astronomical Society of the Pacific, 130, 074501

Morningstar W. R., Hezaveh Y. D., Levasseur L. P., Blandford R. D., Marshall P. J., Putzky P., Wechsler R. H., 2018, arXiv preprint arXiv:1808.00011

Mosteller F., Tukey J. W., 1977, Data analysis and regression. A second course in statistics. Addison-Wesley

Moustakas L., 2012, The Master Lens Database and The Orphan Lenses Project, HST Proposal ID 12833. Cycle 20

Narayan R., Bartelmann M., 1996, eprint arXiv:astro-ph/9606001

Natarajan P., De Lucia G., Springel V., 2007, Mon. Not. Roy. Astron. Soc., 376, 180

Neal R. M., 1996, Bayesian Learning for Neural Networks. Vol. 118, Springer Science & Business Media

Newman A. B., Treu T., Ellis R. S., Sand D. J., Nipoti C., Richard J., Jullo E., 2012b, The Astrophysical Journal, Volume 765, Issue 1, article id. 24, 35 pp. (2013)., 765

Newman A. B., Treu T., Ellis R. S., Sand D. J., 2012a, The Astrophysical Journal, Volume 765, Issue 1, article id. 25, 12 pp. (2013)., 765

Newman A. B., Ellis R. S., Treu T., 2015, The Astrophysical Journal, Volume 814, Issue 1, article id. 26, 28 pp. (2015)., 814

Nightingale J. W., Dye S., Massey R. J., 2018, Monthly Notices of the Royal Astronomical Society, 478, 4738–4784

Nord B., et al., 2015, arXiv:1512.03062,

Nord B., et al., 2016, The Astrophysical Journal, 827, 51

Nord B., et al., 2020, Monthly Notices of the Royal Astronomical Society, 494, 1308–1322

Oemler A., Clardy K., Kelson D., Walth G., Villanueva E., 2017, COSMOS: Carnegie Observatories System for MultiObject Spectroscopy (ascl:1705.001)

Oguri M., 2007, Astrophys. J., 660, 1

Oguri M., 2010, glafic: Software Package for Analyzing Gravitational Lensing, Astrophysics Source Code Library (ascl:1010.012)

Oguri M., Marshall P. J., 2010, Monthly Notices of the Royal Astronomical Society, Volume 405, Issue 4, pp. 2579-2593., 405, 2579

Ostrovski F., et al., 2017, Mon. Not. Roy. Astron. Soc., 465, 4325

Papamakarios G., Murray I., 2018, Fast ϵ-free Inference of Simulation Models with Bayesian Conditional Density Estimation (`arXiv:1605.06376`)

Papamakarios G., Pavlakou T., Murray I., 2017, arXiv preprint arXiv:1705.07057

Paraficz D., et al., 2016, arXiv:1605.04309,

Park J. W., Wagner-Carena S., Birrer S., Marshall P. J., Lin J. Y.-Y., Roodman A., Collaboration L. D. E. S., et al., 2021, The Astrophysical Journal, 910, 39

Pearson J., Li N., Dye S., 2019, Monthly Notices of the Royal Astronomical Society, 488, 991

Pearson J., Maresca J., Li N., Dye S., 2021, Monthly Notices of the Royal Astronomical Society, 505, 4362

Peng C. Y., Ho L. C., Impey C. D., Rix H.-W., 2010, Astron. J., 139, 2097

Peralta D., Triguero I., García S., Saeys Y., Benitez J. M., Herrera F., 2018, International Journal of Intelligent Systems, 33, 213

Petkova M., Metcalf R. B., Giocoli C., 2014, Monthly Notices of the Royal Astronomical Society, 445, 1954

Petrillo C. E., et al., 2017, arXiv: 1702.07675,

Petrillo C. E., et al., 2019a, Mon. Not. Roy. Astron. Soc., 482, 807

Petrillo C. E., et al., 2019b, Mon. Not. Roy. Astron. Soc., 484, 3879

Petters A., Levine H., Wambsganss J., 2012, Singularity Theory and Gravitational Lensing. Progress in Mathematical Physics, Birkhäuser Boston, https://books.google.com.br/books?id=i1vdBwAAQBAJ

Picard D., 2021, arXiv e-prints, p. arXiv:2109.08203

Pickles A. J., 1998, Publications of the Astronomical Society of the Pacific, 110, 863

Pizzuti L., et al., 2016, arXiv:1602.03385,

Poindexter S., Morgan N., Kochanek C. S., 2008, Astrophys. J., 673, 34

Prangle D., 2015, Summary Statistics in Approximate Bayesian Computation (arXiv:1512.05633)

Quider A. M., Pettini M., Shapley A. E., Steidel C. C., 2009, Monthly Notices of the Royal Astronomical Society, Volume 398, Issue 3, pp. 1263-1278., 398, 1263

Ratnatunga K. U., Griffiths R. E., Ostrander E. J., 1999, Astron. J., 117, 2010

Refsdal S., 1964, Monthly Notices of the Royal Astronomical Society, 128, 307

Rezende D., Mohamed S., 2015, in International conference on machine learning. pp 1530–1538

Ribeiro M. T., Singh S., Guestrin C., 2016, arXiv preprint arXiv:1606.05386

Richard J., Jones T., Ellis R., Stark D. P., Livermore R., Swinbank M., 2011, Mon. Not. Roy. Astron. Soc., 413, 643

Rivero A. D., Cyr-Racine F.-Y., Dvorkin C., 2018, Physical Review D, 97, 023001

Rodgers C. T., Canterna R., Smith J. A., Pierce M. J., Tucker D. L., 2006, The Astronomical Journal, Volume 132, Issue 3, pp. 989-993., 132, 989

Rodrigues P. L. C., Moreau T., Louppe G., Gramfort A., 2021, HNPE: Leveraging Global Parameters for Neural Posterior Estimation (arXiv:2102.06477)

Rozo E., et al., 2015, Monthly Notices of the Royal Astronomical Society, Volume 461, Issue 2, p.1431-1450, 461, 1431

Rubin D. B., 1984, The Annals of Statistics, 12, 1151

Ruder S., 2016, arXiv preprint arXiv:1609.04747

Ruff A. J., Gavazzi R., Marshall P. J., Treu T., Auger M. W., Brault F., 2011, The Astrophysical Journal, 727, 96

Russakovsky O., et al., 2015, International Journal of Computer Vision (IJCV), 115, 211

Rykoff E. S., et al., 2014, The Astrophysical Journal, 785, 104

Salpeter E. E., 1955, Astrophys. J., 121, 161

Sánchez-Blázquez P., et al., 2006, Mon. Not. Roy. Astron. Soc., 371, 703

Schlegel D. J., Finkbeiner D. P., Davis M., 1997, The Astrophysical Journal, Volume 500, Issue 2, pp. 525-553., 500, 525

Schneider P., Ehlers J., Falco E., 2013, Gravitational Lenses. Astronomy and Astrophysics Library, Springer Berlin Heidelberg, https://books.google.com.br/books?id=XJ3zCAAAQBAJ

Schwab J., Bolton A. S., Rappaport S. A., 2009, The Astrophysical Journal, 708, 750

Schwab J., Bolton A. S., Rappaport S. A., 2010, Astrophys. J., 708, 750

Scot D. W., 1992, Multivariate density estimation

Sersic J. L., 1968, Atlas de Galaxias Australes. Cordoba, Argentina: Observatorio Astronomico

Sevilla I., et al., 2011, eprint arXiv:1109.6741

Shajib et al. 2019a, Mon. Not. Roy. Astron. Soc., submitted

Shajib A. J., et al., 2019b, Monthly Notices of the Royal Astronomical Society, 483, 5649

Shajib A. J., et al., 2020, Mon. Not. Roy. Astron. Soc., 494, 6072

Shridhar K., Laumann F., Liwicki M., 2019, arXiv preprint arXiv:1901.02731

Sluse D., et al., 2017, Monthly Notices of the Royal Astronomical Society, 470, 4838

Sluse D., et al., 2019, Mon. Not. Roy. Astron. Soc., 490, 613

Smit R., Swinbank A. M., Massey R., Richard J., Smail I., Kneib J. P., 2017, Monthly Notices of the Royal Astronomical Society, Volume 467, Issue 3, p.3306-3323, 467, 3306

Stark D. P., Swinbank A. M., Ellis R. S., Dye S., Smail I. R., Richard J., 2008, Nature, 455, 775

Suyu S. H., Marshall P. J., Auger M. W., Hilbert S., Blandford R. D., Koopmans L. V. E., Fassnacht C. D., Treu T., 2010a, The Astrophysical Journal, Volume 711, Issue 1, pp. 201-221 (2010)., 711, 201

Suyu S. H., Marshall P. J., Auger M. W., Hilbert S., Blandford R. D., Koopmans L. V. E., Fassnacht C. D., Treu T., 2010b, Astrophys. J., 711, 201

Suyu S. H., et al., 2013, The Astrophysical Journal, Volume 766, Issue 2, article id. 70, 19 pp. (2013)., 766

Suyu S. H., et al., 2016, eprint arXiv:1607.00017, 15, 1

Suyu S. H., et al., 2017, Mon. Not. Roy. Astron. Soc., 468, 2590

Szegedy C., et al., 2015, in Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1–9

Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z., 2016, in Proceedings of the IEEE conference on computer vision and pattern recognition. pp 2818–2826

Talts S., Betancourt M., Simpson D., Vehtari A., Gelman A., 2018, Validating Bayesian Inference Algorithms with Simulation-Based Calibration, doi:10.48550/ARXIV.1804.06788, `https://arxiv.org/abs/1804.06788`

Tan C. Y., et al., 2023, Project Dinos I: A joint lensing-dynamics constraint on the deviation from the power law in the mass profile of massive ellipticals (`arXiv:2311.09307`)

Tanoglidis D., Ćiprijanović A., Drlica-Wagner A., 2022, arXiv e-prints, p. arXiv:2207.03471

Tavaré S., Balding D. J., Griffiths R. C., Donnelly P., 1997, Genetics, 145, 505

Tejero-Cantero A., Boelts J., Deistler M., Lueckmann J.-M., Durkan C., Gonçalves P. J., Greenberg D. S., Macke J. H., 2020, Journal of Open Source Software, 5, 2505

Tessore N., Metcalf R. B., 2015, Astronomy & Astrophysics, Volume 580, id.A79, 6 pp., 580

Tessore N., Bellagamba F., Metcalf R. B., 2016, Monthly Notices of the Royal Astronomical Society, 463, 3115

Tishby N., Zaslavsky N., 2015, in 2015 IEEE Information Theory Workshop (ITW). pp 1–5

Tremblay G. R., et al., 2014, Astrophys. J. Lett., 790, L26

Treu T., 2010, ARA&A, 48, 87

Treu T., Koopmans L. V. E., 2002a, Mon. Not. Roy. Astron. Soc., 337, L6

Treu T., Koopmans L. V. E., 2002b, Mon. Not. Roy. Astron. Soc., 337, L6

Treu T., Koopmans L. V. E., 2002c, The Astrophysical Journal, 575, 87

Treu T., Marshall P. J., 2016, A&ARv, 24, 11

Treu T., et al., 2018a, Mon. Not. Roy. Astron. Soc., 481, 1041

Treu T., et al., 2018b, Mon. Not. Roy. Astron. Soc., 481, 1041

Uria B., Côté M.-A., Gregor K., Murray I., Larochelle H., 2016, The Journal of Machine Learning Research, 17, 7184

Vazdekis A., Sánchez-Blázquez P., Falcón-Barroso J., Cenarro A. J., Beasley M. A., Cardiel N., Gorgas J., Peletier R. F., 2010, Mon. Not. Roy. Astron. Soc., 404, 1639

Vecchiotti P., Vesperini F., Principi E., Squartini S., Piazza F., 2018, in , Multidisciplinary Approaches to Neural Computing. Springer, pp 161–170

Vegetti S., Lagattuta D. J., McKean J. P., Auger M. W., Fassnacht C. D., Koopmans L. V. E., 2012, Nature, 481, 341

Vovk V., Gammerman A., Saunders C., 1999, in Sixteenth International Conference on Machine Learning (ICML-1999) (01/01/99). pp 444–453, https://eprints.soton.ac.uk/258960/

Wagner-Carena S., Park J. W., Birrer S., Marshall P. J., Roodman A., Wechsler R. H., Collaboration L. D. E. S., et al., 2021, The Astrophysical Journal, 909, 187

Wagner-Carena S., Aalbers J., Birrer S., Nadler E. O., Darragh-Ford E., Marshall P. J., Wechsler R. H., 2023, Astrophys. J., 942, 75

Wainer H., Thissen D., 1976, Psychometrika, 41, 9

Walker M. G., Peñarrubia J., 2011, Astrophys. J., 742, 20

Walmsley M., et al., 2022, Mon. Not. Roy. Astron. Soc., 509, 3966

Walsh D., Carswell R. F., Weymann R. J., 1979, Nature, 279, 381

Weiner C., Serjeant S., Sedgwick C., 2020, Research Notes of the AAS, 4, 190

Wen Z.-L., Han J.-L., Jiang Y.-Y., 2011, Research in Astronomy and Astrophysics, 11, 1185

Wen Y., Vicol P., Ba J., Tran D., Grosse R., 2018, arXiv preprint arXiv:1803.04386

Willis J. P., Hewett P. C., Warren S. J., Dye S., Maddox N., 2006, Mon. Not. Roy. Astron. Soc., 369, 1521

Wilman D. J., Balogh M. L., Bower R. G., Mulchaey J. S., Oemler A., Carlberg R. G., Morris S. L., Whitaker R. J., 2005, Mon. Not. Roy. Astron. Soc., 358, 71

Wilson M. L., Zabludoff A. I., Ammons S. M., Momcheva I. G., Williams K. A., Keeton C. R., 2016, Astrophys. J., 833, 194

Wong K. C., et al., 2019, arXiv e-prints, p. arXiv:1907.04869

Xie S., Girshick R., Dollár P., Tu Z., He K., 2017, in Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp 5987–5995

Yamamoto K., Kadoya Y., Murata T., Futamase T., 2001, Progress of Theoretical Physics, 106, 917

Yuan T. T., Kewley L. J., 2009, The Astrophysical Journal Letters, Volume 699, Issue 2, pp. L161-L164 (2009)., 699, L161

Zaborowski E. A., et al., 2023, The Astrophysical Journal, 954, 68

Zackrisson E., Riehm T., 2010, Advances in Astronomy, 2010

Zahid H. J., Geller M. J., Fabricant D. G., Hwang H. S., 2016, Astrophys. J., 832, 203

Zaritsky D., Gonzalez A. H., 2003, Astrophys. J., 584, 691

Zeiler M. D., 2012, arXiv e-prints, p. arXiv:1212.5701

Zhang Y., Fu K., Sun H., Sun X., Zheng X., Wang H., 2018, Remote Sensing Letters, 9, 11

Zwicky F., 1937, Phys. Rev., 51, 290