Strong Gravitational Lenses in the Era of Wide-Field Surveys

Philip Holloway

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

DECLARATION

I declare that no part of this thesis has been accepted, or is currently being submitted, for any degree or diploma or certificate or any other qualification in this University or elsewhere. Parts of this thesis have been previously published in conference proceedings or peer-reviewed journals as follows:

- Chapter 2 was published in its original form in the *Monthly Notices of the Royal Astronomical Society* as Holloway et al. (2023).
- The work underlying Chapter 3 appeared in conference proceedings Holloway et al. (2024a) and was subsequently published in the *Monthly Notices of the Royal Astronomical Society* as Holloway et al. (2024b).
- Chapter 4 has undergone internal review by the *Euclid* collaboration and has been submitted to the *Astronomy and Astrophysics* journal. It appears in its original form as a pre-print in Euclid Collaboration: Holloway et al. (2025).

The work underlying Chapter 5 has been submitted for internal review in the LSST Dark Energy Science Collaboration (DESC). Elements of the work in this thesis have been carried out in collaboration; my contributions are detailed below:

- Chapter 3 presents a strong lens search in the Dark Energy Survey (DES) using a Vision Transformer (VT) and citizen science (Section 3.5.5). As part of this search I ran the Space Warps Analysis Pipeline (SWAP), and facilitated the launch of the citizen science search but was not involved in the development or application of the VT. I had previously developed the automated retirement of systems from the Space Warps classification stream which was first used in this search and again in the *Euclid* lens search (below). The combination of citizen science and VT classifiers into an ensemble, which is the main focus of Section 3.5.5, is my own work. The method and results of this DES lens search are presented in González et al. (2025).
- Chapter 4 presents the results of a strong lens search in *Euclid* data. This collaborative project by the *Euclid* Strong Lens Working Group involved multiple teams working on image cutout generation, training set development, machine learning algorithms, the citizen science search and expert grading of lens candidates. As in the DES search, I ran the SWAP pipeline and helped with the preparation, running and analysis for the Space Warps citizen science search. The search strategy and main results of this search are presented in Euclid Collaboration: Walmsley et al. (2025). The focus of Chapter 4 is ensemble classification and consequences for lens finding in *Euclid*, which is my own work.

Philip Holloway *June 30, 2025*

Acknowledgements

Firstly, thank you so much to my supervisor Aprajita for all your help and guidance through the past four years. I have learnt a huge amount on strong lensing, citizen science and astrophysics through your supervision and have had a great time along the way. I can't wait to continue collaborating in my post-doc years!

I'd also like to thank the KIPAC strong lensing group, with special thanks to Phil Marshall for your supervision during my LTA and beyond, and Sydney Erickson and Padma Venkatraman for making me so welcome at Stanford. I had a truly amazing time and it was an experience I'll never forget.

Thanks must also go to the many people who have guided me throughout my DPhil, in Oxford and beyond; Anupreeta More, Matt Jarvis, Imogen Whittam, Catherine Hale, Ian Heywood and Matthias Tecza, thank you!

I'd also like to thank Ashling Gordon for all of your help in navigating the DPhil course, Leanne O'Donnell for your help with organising the Oxford strong lensing conference and Jonathan Patterson for all your help with Glamdring, especially on the inevitable occasions when I deleted something I didn't mean to!

To my officemates Matthew, Fergus, Rohan, James, David and Haochuan, thank you for being such wonderful friends, and thank you for all the laughs. To Matthew and Fergus for being such amazing flatmates, to Rohan for instilling in me that now is always a good time to go climbing, and James for the rollercoaster ride of Wikipedia rabbit holes which never failed to liven up a quiet afternoon. To the members of Pat and the Roche Lobes, thank you for being part of a true highlight of my time in Oxford, I'm not sure anyone will forget the Christmas Party or Breakthrough Discuss anytime soon! And to the wider Astro department for creating such an great place to work; Henry, Luke, Emma, Katie, Alex, Madalina, Casey, James G and M, Tom and everyone else, thanks to you all!

And finally, thank you to my family, Mum, Dad and Loz, for all your love, support and encouragement. I couldn't have imagined I would be sitting here now, about to submit my thesis at Oxford, when I first wondered about becoming an astrophysicist. DPhil's are a long road with many ups and downs - I simply couldn't have done it without you.

Abstract

Gravitational lensing is the deflection of a light beam due to the distortion of space-time caused by a massive object. Strong gravitational lensing occurs when this deflection is sufficient to produce multiple images of the same background source as viewed by an observer. The applications of such strong lens systems are widespread, from studying the initial mass function to measurements of the Hubble Constant. In the coming years, strong lens science will undergo a revolution with key data releases from the Legacy Survey of Space and Time (LSST) and Euclid Wide Survey (EWS) which are each expected to identify $\sim 100\,000$ lensed systems, increasing the number of known candidates by two orders of magnitude. In this thesis, I describe the opportunities that this data will bring.

Typical searches for strong lenses have to date focussed on visible, sub-millimetre and radio wavelengths. However, the advent of large format, sensitive, Near-Infrared (NIR) detectors as seen in *Euclid* and the *James Webb Space Telescope (JWST)* will enable searches in high-resolution NIR surveys. I demonstrate that *JWST* will identify lensed galaxies at higher redshift than ever seen before and that *Euclid* will allow the NIR lens population to be studied at scale.

The vast data volume from wide-field surveys such as *Euclid* and LSST presents a false positive problem, whereby high-scoring samples from lens classifiers are dominated by false positives. I show that the current performance of strong lens classifiers can be improved by combining these into an ensemble. Moreover, I demonstrate that even the state-of-the-art classifiers will still produce heavily contaminated or incomplete samples of lenses when applied to wide-field surveys. Given the vast majority of the lenses identified in such surveys will not receive spectroscopic confirmation, analysis of the complete lens candidate datasets will require the possibility of contamination to be accounted for. I demonstrate that cosmological parameter inference can still be conducted even in the presence of such contamination. To do so requires accepting a known contamination rate and knowledge of the probability that each system is a lens but, as I demonstrate, such data can be obtained from current lens classifiers.

The coming years will be a truly exciting time for strong lens science; this thesis aims to confront some of the challenges we will face.

Contents

List of Abbreviations **ix**

1 Introduction **1**
1.1 The Standard Model of Cosmology: ΛCDM 2
1.1.1 A Homogeneous and Isotropic Universe 2
1.1.2 Evolution of the ΛCDM Universe 4
1.1.3 Cosmological Distance Metrics 6
1.2 Gravitational Lensing 7
1.2.1 Early History and First Discoveries 7
1.2.2 Theory of Gravitational Lensing 8
1.2.3 Gravitational Lensing Regimes 11
1.2.4 Strong Lens Discovery through Time 13
1.2.5 Astrophysical Applications of Strong Lenses 18
1.2.6 Cosmological Applications of Strong Lenses 20
1.2.7 Strong Lens Modelling 24
1.2.8 Strong Lensing Today: Challenges and Opportunities 28
1.3 Thesis Overview 30

2 The Occurrence Rates of Strong Lenses in NIR Surveys **32**
2.1 Introduction 33
2.2 Data 35
2.2.1 Adaptions to the Galaxy Catalogue 37
2.3 Method 39
2.3.1 Frequency of Galaxy-Galaxy Conjunctions 39
2.3.2 Calculating Lensing Properties 41
2.3.3 Detectability Constraints 42
2.4 Results 46
2.4.1 Verifying the Simulations 46
2.4.2 The Observable Lens Population 48
2.4.3 Lens and Source Population Properties 53
2.4.4 Extrapolating to Wide-Field Surveys 57
2.5 Discussion 59

2.5.1 Number Density of Detectable Lens Systems 59
2.5.2 Properties of the Detected Strong Lenses 62
2.5.3 Validation of Lens Occurrence Rates with Recent Strong Lens Discoveries 64
2.5.4 Potential Further Improvements 70
2.6 Conclusion 71

3 A Bayesian Approach to Strong Lens Finding 74
3.1 Introduction 75
3.2 Data 78
3.3 Method 81
3.3.1 Summary of Calibration Methods 82
3.3.2 Application of Calibration Methods 84
3.3.3 Summary of Combination Methods 88
3.4 Results 92
3.4.1 Testing the Bayesian Combination Approaches 92
3.4.2 Applying the Bayesian Combination Methods 93
3.5 Discussion 96
3.5.1 Comparison with Previous Work 96
3.5.2 Comparison of Citizen Science versus a Network Ensemble . 97
3.5.3 Effect of Ground-Truth Selection on Classifier Performance . 98
3.5.4 Expectations and Implications for LSST 100
3.5.5 Application to the Dark Energy Survey 103
3.6 Conclusion 105

4 Lens Searches in Contemporary Wide Area Surveys 108
4.1 Introduction 109
4.2 Data 111
4.2.1 The *Euclid* Q1 Lens Search 111
4.2.2 The *Euclid* Q1 Lens Classifiers 113
4.3 Method 115
4.3.1 Calibration of Strong Lens Classifiers 115
4.3.2 Combination of Classifiers into an Ensemble 120
4.4 Results and Discussion 121
4.4.1 Ensemble Classifier Performance 121
4.4.2 Systems Identified by Citizens or Ensemble 125
4.4.3 Outlook for *Euclid* DR1 and Future Data Releases 127
4.4.4 Optimising Lens Searches in Wide Area Surveys 130
4.5 Conclusions 133

5 Lens Modelling and Cosmological Inference from an Impure Sample of Galaxy-Galaxy Strong Lenses 135
5.1 Introduction . . . 136
5.2 Data . . . 139
5.2.1 Network Training Set . . . 141
5.3 Method . . . 142
5.3.1 Neural Network Training . . . 142
5.3.2 COSMIC-BEAMS Formulation . . . 144
5.3.3 Generation of Inference Data Vectors . . . 151
5.4 Results . . . 155
5.4.1 Lens Modelling of LSST Lenses . . . 155
5.4.2 Lens Modelling of False Positives . . . 156
5.4.3 Posterior Images for Lenses and False Positives . . . 158
5.4.4 Cosmological Inference from an Impure Sample of Strong Lenses 159
5.5 Discussion . . . 161
5.5.1 Modelling of LSST Lens Candidates . . . 161
5.5.2 Inference with Impure Samples of Strong Lenses . . . 164
5.6 Conclusion . . . 166

6 Conclusions 169

Appendices

A Supplementary Details on Ensemble Methodology 176
A.1 Invariance of Dependent Bayesian Method to the Choice of Classifier 176
A.2 Relative dependence of HSC and *Euclid* Classifiers . . . 180

B Supplementary Plots from COSMIC-BEAMS Analysis 183
B.1 Lens-Source Redshift Dependence for the LSST Lens Population . . 183
B.2 Posteriors for all Lens Parameters . . . 183
B.3 Network Behaviour to False Positives across Lens Parameters 186
B.4 Cosmological Precision with Different Datasets . . . 186
B.5 Predicted Posterior Images from `paltas` . . . 186

References 192

List of Abbreviations

ΛCDM ..... Lambda Cold Dark Matter
2dFGRS ..... Two-degree-Field Galaxy Redshift Survey
2MASS ..... Two Micron All-Sky Survey
4MOST ..... 4-metre Multi-Object Spectroscopic Telescope
4SLSLS ..... 4MOST Strong Lensing Spectroscopic Legacy Survey
ABC ..... Approximate Bayesian Computation
AGN ..... Active Galactic Nuclei
ALMA ..... Atacama Large Millimeter Array
APM ..... Automatic Plate Measurement
AUROC ..... Area Under ROC
BAO ..... Baryonic Acoustic Oscillations
CANDELS Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey
CDFS ..... Chandra Deep Field-South
CDM ..... Cold Dark Matter
CFHT ..... Canada-France-Hawaii Telescope
CFHTLS ..... Canada-France-Hawaii Telescope Legacy Survey
CLASS ..... Cosmic Lens All-Sky Survey
CMB ..... Cosmic Microwave Background
CNN ..... Convolutional Neural Network
COBE ..... Cosmic Background Explorer
COSMOS ..... Cosmic Evolution Survey
COWLS ..... COSMOS-Web Lens Survey
DES ..... Dark Energy Survey
DESI ..... Dark Energy Spectroscopic Instrument
DR1 ..... Data Release 1
DREaM ..... Deep Realistic Extragalactic Model

DSPL ..... Double Source Plane Lens

ELT ..... Extremely Large Telescope

ERO ..... Early Release Observations

EWS ..... Euclid Wide Survey

FLRW ..... Friedmann–Lemaître–Robertson–Walker

FPR ..... False Positive Rate

FWHM ..... Full Width Half Maximum

GAMA ..... Galaxy And Mass Assembly

GJ ..... Galaxy Judges

GMM ..... Gaussian Mixture Model

HOLISMOKES ..... Highly Optimised Lensing Investigations of Supernovae, Microlensing Objects, and Kinematics of Ellipticals and Spirals

HSC SSP ..... Hyper-Suprime Cam Subaru Strategic Program

HSC ..... Hyper-Suprime Cam

HST ..... Hubble Space Telescope

IMF ..... Initial Mass Function

IR ..... Infrared

JADES ..... JWST Advanced Deep Extragalactic Survey

JAGUAR ..... JADES Extragalactic Ultra-deep Artificial Realization

JWST ..... James Webb Space Telescope

KiDS ..... Kilo-Degree Survey

KLIEP ..... Kullback Leibler Importance Estimation Procedure

LSST ..... Legacy Survey of Space and Time

MACHOS ..... Massive Compact Halo Object

MCMC ..... Markov Chain Monte Carlo

MIR ..... Mid-Infrared

ML ..... Machine Learning

MVN ..... Multi-Variate Normal

NFW ..... Navarro-Frenk-White

NIR ..... Near-Infrared

NISP ..... Near Infrared Spectrometer and Photometer

NPE ..... Neural Posterior Estimation
PDF ..... Probability Density Function
PEMD ..... Power-law Elliptical Mass Distribution
PGM ..... Probabilistic Graphical Model
PSF ..... Point Spread Function
ROC ..... Receiver Operating Characteristic
SBI ..... Simulation Based Inference
SDSS ..... Sloan Digital Sky Survey
SED ..... Spectral Energy Distribution
SHAM ..... Subhalo Abundance Matching
SIE ..... Singular Isothermal Ellipsoid
SIS ..... Singular Isothermal Sphere
SL2S ..... Strong Lensing Legacy Survey
SLACS ..... Sloan Lens ACS
SNR ..... Signal to Noise Ratio
SPS ..... Stellar Population Synthesis
SuGOHI ..... Survey of Gravitationally lensed Objects in HSC Imaging
SVM ..... Support Vector Machine
SWAP ..... Space Warps Analysis Pipeline
TPR ..... True Positive Rate
UKIRT ..... United Kingdom Infrared Telescope
VDF ..... Velocity Dispersion Function
VIDEO ..... VISTA Deep Extragalactic Observations
VISTA ..... Visible and Infrared Survey Telescope for Astronomy
VT ..... Vision Transformer
WMAP ..... Wilkinson Microwave Anisotropy Probe

1 Introduction

Contents

1.1 The Standard Model of Cosmology: ΛCDM 2
1.1.1 A Homogeneous and Isotropic Universe 2
1.1.2 Evolution of the ΛCDM Universe 4
1.1.3 Cosmological Distance Metrics 6
1.2 Gravitational Lensing 7
1.2.1 Early History and First Discoveries 7
1.2.2 Theory of Gravitational Lensing 8
1.2.3 Gravitational Lensing Regimes 11
1.2.4 Strong Lens Discovery through Time 13
1.2.5 Astrophysical Applications of Strong Lenses 18
1.2.6 Cosmological Applications of Strong Lenses 20
1.2.7 Strong Lens Modelling 24
1.2.8 Strong Lensing Today: Challenges and Opportunities . . 28
1.3 Thesis Overview 30

1.1 The Standard Model of Cosmology: ΛCDM

1.1.1 A Homogeneous and Isotropic Universe

The basis of our understanding of the Universe and the laws which describe it is the Cosmological Principle that, when seen on sufficiently large scales, the Universe is homogeneous (the same at each point) and isotropic (looks the same in all directions). While originally posited for mathematical convenience, measurements of the Cosmic Microwave Background (CMB) by the Cosmic Background Explorer satellite (COBE, Smoot et al., 1992; Fixsen et al., 1996), and subsequently by the Wilkinson Microwave Anisotropy Probe (WMAP, Hinshaw et al., 2013) and *Planck* (Planck Collaboration et al., 2020) experiments, have since shown the early universe was close to isotropic with only minor fluctuations. Similarly, measurements of the galaxy distributions in wide-field surveys such as the Two-degree-Field Galaxy Redshift Survey (2dFGRS, Percival et al., 2001; Lahav, 2002), Two Micron All-Sky Survey (2MASS, Alonso et al., 2015) and Sloan Digital Sky Survey (SDSS, Pandey and Sarkar, 2015) have found excellent agreement with this principle. From the observations of Lemaître (1927) and Hubble (1929), we also know that the Universe is currently expanding; galaxies appear redshifted, with greater recessional velocities at greater distances. Measuring this expansion, with its implication that galaxies were closer together at earlier times, led to the concept of the Big Bang.

The standard, 'concordance' model of cosmology is ΛCDM, in which the Universe originated from a singularity at the Big Bang, before expanding rapidly during a period of inflation. Over-densities in the primordial density field led to the formation of dark matter halos in which galaxies could evolve, eventually forming the 'cosmic-web', a complex network of filaments and voids which trace high and low density regions of galaxies. The ΛCDM universe consists of baryonic matter and radiation, the properties of which are well known, and two additional components: dark matter and dark energy.

In the concordance model, dark energy takes the form of a cosmological constant, Λ. This was originally invoked by Einstein to balance out the effects of gravity

and thus allow for a static universe, i.e., one which did not expand or contract over time. Such a static state would be unstable and the cosmological constant was dropped following the observations of Hubble that the Universe was expanding (i.e., not static). However, it is now used as a possible mathematical explanation to the observed rapid expansion of the Universe at the present day. Evidence for a cosmological constant built up during the 1990's, for example, through the study of number counts of faint galaxies (Fukugita et al., 1990) and measurements of the angular correlation function of galaxies in the Automatic Plate Measurement galaxy survey (APM, Maddox et al., 1990; Efstathiou et al., 1990). This was supported further by studies comparing the measured luminosity distances and redshifts of (standardisable) Type Ia supernovae (Riess et al., 1998; Perlmutter et al., 1998) which favoured an expanding universe with a non-zero cosmological constant.

The second unexplained component is that of dark matter. Zwicky postulated the existence of such a non-luminous 'dark' form of matter from the discrepancy between observed velocity dispersions in galaxy clusters (e.g., the Coma cluster), and that which would be expected from a system in virial equilibrium. (Zwicky, 1933; Zwicky, 1937). Further evidence of the existence of dark matter came from observations of flat galaxy rotation curves out to large radii (Rubin and Ford, 1970; Rubin et al., 1980), and observations of the signatures of gravitational lensing (e.g., Mandelbaum et al., 2006; Gavazzi et al., 2007). The leading formulation of dark matter is that of Cold Dark Matter (CDM). This is a collisionless, non-relativistic form of matter which does not interact electromagnetically but does interact through gravity.

To date, the ΛCDM model show has shown excellent agreement with many cosmological measurements, such as anisotropies in the CMB (Planck Collaboration et al., 2020), measurements of 3x2pt correlations (Abbott et al., 2022) and Baryonic Acoustic Oscillations (BAO, Alam et al., 2021). However, there a number of observed challenges to this model. Firstly, the present-day expansion rate of the universe derived from late-universe probes (e.g., Type 1a supernovae) is in 5σ tension with those from the early universe (namely the CMB), known as the 'Hubble Tension'. Moreover, measurements of the amplitude of matter-density fluctuations are in

$2-3\sigma$ tension between early- and late-universe probes (termed the 'S_8 tension'). So-called Stage-IV probes (Albrecht et al., 2006) such as *Euclid* (Euclid Collaboration: Mellier et al., 2024) and the Dark Energy Spectroscopic Instrument (DESI) are paving the way for the era of 'precision cosmology' in which the ΛCDM model can be rigorously challenged. In the following sections I will briefly outline the theory underlying the evolving ΛCDM universe, and the distance metrics relevant to this thesis used to probe its composition.

1.1.2 Evolution of the ΛCDM Universe

In General Relativity[1] (Einstein, 1916), the geometry of space-time is described by a metric. Such a homogeneous, isotropic and expanding universe is described by the Friedmann–Lemaître–Robertson–Walker (FLRW) metric[2]

$$ds^2 = -c^2dt^2 + a(t)^2[d\chi^2 + S_k(\chi)d\Omega^2] \tag{1.2}$$

which evolves according to Einstein's Field Equations. Solving these equations provides the Friedmann Equation

$$\left(\frac{\dot{a}}{a}\right)^2 - \frac{8}{3}\pi G\rho - \frac{1}{3}\Lambda c^2 = -\frac{kc^2}{a^2}\,. \tag{1.3}$$

Here, $a(t)$ describes the scale factor of the Universe as a function of time t, where at the present epoch, $a(t_0) \equiv a_0 = 1$. In this equation, ρ is a density parameter (including matter and radiation), k describes the spatial curvature and Λ is the cosmological constant. The Einstein field equations also produce the fluid equation

$$\frac{d}{dt}(c^2\rho a^3) = -P\frac{da^3}{dt}\,. \tag{1.4}$$

[1] Alternative theories of gravity have been suggested (see reviews by Joyce et al., 2016; Ferreira, 2019), but are not considered further in this thesis.

[2] Here, χ and Ω are spatially co-moving coordinates, t indicates time, $a(t)$ is the scale factor and S_k is given by

$$S_k(\chi) = \begin{cases} \frac{c}{\sqrt{\Omega_k}H_0}\sinh\left(\frac{H_0\sqrt{\Omega_k}}{c}\chi\right) & \Omega_k > 0, \\ \chi & \Omega_k = 0, \\ \frac{c}{\sqrt{|\Omega_k|}H_0}\sin\left(\frac{H_0\sqrt{|\Omega_k|}}{c}\chi\right) & \Omega_k > 0. \end{cases} \tag{1.1}$$

To incorporate this into the Friedmann Equation requires knowledge of the equation of state parameter w (where $P = w\rho c^2$) for each of the components of the universe; matter, radiation and vacuum energy (in the form of a cosmological constant in Eqn. 1.3). For matter, radiation and a cosmological constant these are given by $w_m = 0, w_r = 1/3$ and $w_\Lambda = -1$ respectively. It follows that the density parameters evolve following $\rho_m \propto a^{-3}, \rho_r \propto a^{-4}$ and $\rho_\Lambda \propto a^0$.

The Hubble Parameter, $H(t) \equiv \dot{a}/a$, which describes the expansion of the Universe over time can be written as the sum of the energy density components scaled by the critical density, $\rho_{cr} = \frac{3H_0^2}{8\pi G}$ giving

$$H^2(t) = -\frac{kc^2}{a^2} + H_0^2 \sum \Omega_i \,. \tag{1.5}$$

Applying the constraint that at the present time, $H(t_0) \equiv H_0$, defining $\Omega_{k,0} = -kc^2/H_0^2$ and applying the scaling relations above gives the evolution of the Hubble parameter over time

$$H^2 = H_0^2 \left[\Omega_{k,0}\, a^{-2} + \Omega_{m,0}\, a^{-3} + \Omega_{r,0}\, a^{-4} + \Omega_\Lambda\right] . \tag{1.6}$$

Therefore, the Hubble parameter evolves according to the relative energy densities of the known components of the universe; matter (both baryonic and dark), radiation and dark energy, as well as a contributing term from spatial curvature. Current evidence (e.g., from the CMB, Planck Collaboration et al., 2020 and BAO, Alam et al., 2021) suggests that the Universe has close to zero curvature, with $\Omega_m \sim 0.3$, $\Omega_r \sim 10^{-5}$ and $\Omega_\Lambda \sim 0.7$. Consequently, the Universe evolved from a radiation dominated era, to a matter dominated era and subsequently became dark energy dominated. There are many other theoretical models of dark energy such as quintessence or coupled dark energy. To account for such models, the dark energy term can be described by a time-varying equation of state parameter $w(a)$. In the Chevallier-Polarski-Linder (CPL) parametrisation (Chevallier and Polarski, 2001; Linder, 2003), this is a Taylor expansion of the scale parameter, $w = w_0 + w_a(1-a)$ where ΛCDM corresponds to $w_0 = -1, w_a = 0$. Measurements confirming a time-varying equation of state, as suggested by recent evidence from DESI (Adame et al.,

2025; DESI Collaboration et al., 2025a; DESI Collaboration et al., 2025b), would break the current concordance model of cosmology and are thus of great interest; one method to measure this is the subject of Chapter 5.

1.1.3 Cosmological Distance Metrics

In General Relativity, light travels along null geodesics. As the universe expands, the wavelength of the light[3] changes according to the scale factor

$$\frac{\lambda_r}{\lambda_e} = \frac{a(t_r)}{a(t_e)}\,. \tag{1.7}$$

When observed at $a(t_0) = 1$, the light is redshifted according to

$$z \equiv \frac{\lambda_r - \lambda_e}{\lambda_e} = \frac{1}{a(t_e)} - 1 \tag{1.8}$$

or equivalently $a(z) = \frac{1}{1+z}$.

There are two distances of importance within cosmology and the science to come: the luminosity distance and the angular diameter distance. The former can be derived from considering a source with luminosity L which is received at a comoving distance χ as a flux F. The luminosity distance D_L is defined such that the following relation holds

$$F = \frac{L}{4\pi D_L^2} \tag{1.9}$$

Due to cosmological redshift, the energy of the arriving photons $E = hc/\lambda$ is reduced by a factor $1 + z$, and due to time dilation the photons arrive at lower frequency, by an additional $1 + z$ factor. Therefore, the received flux is given by

$$F = \frac{L}{4\pi S_k(\chi)^2(1+z)^2} \tag{1.10}$$

and thus the luminosity distance is given by $D_L = S_k(\chi)(1 + z)$.

Secondly, the angular diameter distance is defined such that an object of length l subtends an angle θ at a distance D_A. Integrating the FLRW metric at fixed time and radial distance gives $l = a(t)S_k(\chi)\theta$, and thus

$$D_A \equiv \frac{l}{\theta} = a(t)S_k(\chi) = \frac{S_k(\chi)}{1+z} \tag{1.11}$$

[3]Here I am considering a scenario in which the light is emitted at time t_e and received at time t_r.

Both angular diameter and luminosity distances depend on the Hubble parameter, and thus on the cosmological parameters Ω_m, Ω_k, Ω_λ and Ω_r. Therefore, these parameters can be inferred from combining redshift measurements with knowledge of a system's luminosity or size, i.e., standard candles (e.g., Type 1a supernovae) or standard rulers (e.g., BAO) respectively.

1.2 Gravitational Lensing

Gravitational lensing, the deflection of light due to mass, has emerged as a effective tool for probing our understanding of both galaxy evolution and cosmology. The theory behind gravitational lensing stems from that of General Relativity, but can be described efficiently by simple optics equations. Here I describe the history, mathematical foundations and applications of gravitational lenses.

1.2.1 Early History and First Discoveries

The theory behind gravitational lensing has a long history. In the 1700's, Newton postulated that light formed corpuscles (particles) which followed the laws of classical mechanics (Newton, 1704). Shortly afterwards, Michell and de Laplace separately suggested that the mass of stars could be measured by the deceleration of such particles as they travel past one another (Michell, 1784; de Laplace, 1796). The first mention of gravitation of light came in the 1800's, from Cavendish. In 1905, Einstein published his theory of Special Relativity. From this he derived the deflection angle of light due to gravity and found the same value as in Newtonian mechanics (Einstein, 1911), previously derived by Cavendish (identified by Will, 1988) and Soldner (Soldner, 1804)[4]:

$$\hat{\alpha}_{\rm SR} = \frac{2GM}{c^2 R} \tag{1.12}$$

A few years later Einstein repeated this calculation based on his theory of General Relativity (Einstein, 1915; Einstein, 1916). This increased the deflection angle

[4]Here we consider a light particle passing a point-mass of mass M at radius R with speed c.

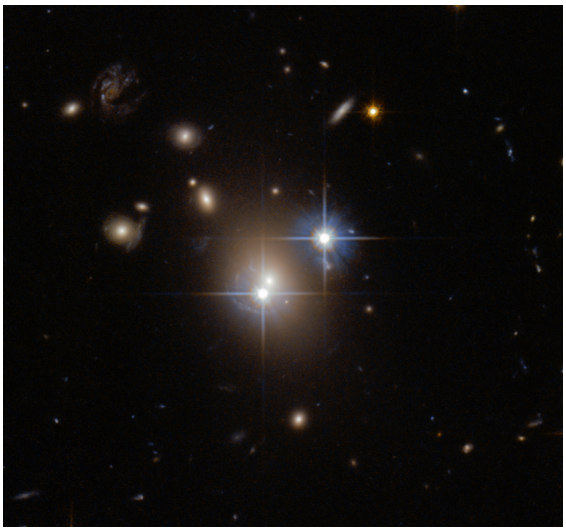


Figure 1.1: Hubble Space Telescope (HST) image of the first gravitationally lensed quasar, QSO B0957+0561, named the 'Twin Quasar' (Walsh et al., 1979). Credit: ESA/Hubble & NASA.

by a factor of 2

$$\hat{\alpha}_{\mathrm{GR}} = \frac{4GM}{c^2 R} \tag{1.13}$$

which then became a key test of the theory, verified in Eddington's expedition to Príncipe Island. In 1924, Chwolson (1924) suggested that a lensed ring may form if two objects are sufficiently aligned (now called an Einstein ring) but did not provide detailed calculations. Einstein was famously pessimistic about the prospects of gravitational lensing, but around a similar time to Link (Link, 1936), he developed the mathematical framework of gravitational lensing from an optics perspective (Einstein, 1936). It was not until 1976 that the first gravitational lens was discovered (Walsh et al., 1979), shown in Figure 1.1. Since then, the study of gravitational lenses has flourished, helping us to understand the evolution, composition and laws of the Universe.

1.2.2 Theory of Gravitational Lensing

The Lens Equation

The typical gravitational lens configuration is depicted in Figure 1.2. Given the distances between lensing galaxies is typically significantly larger than the size of

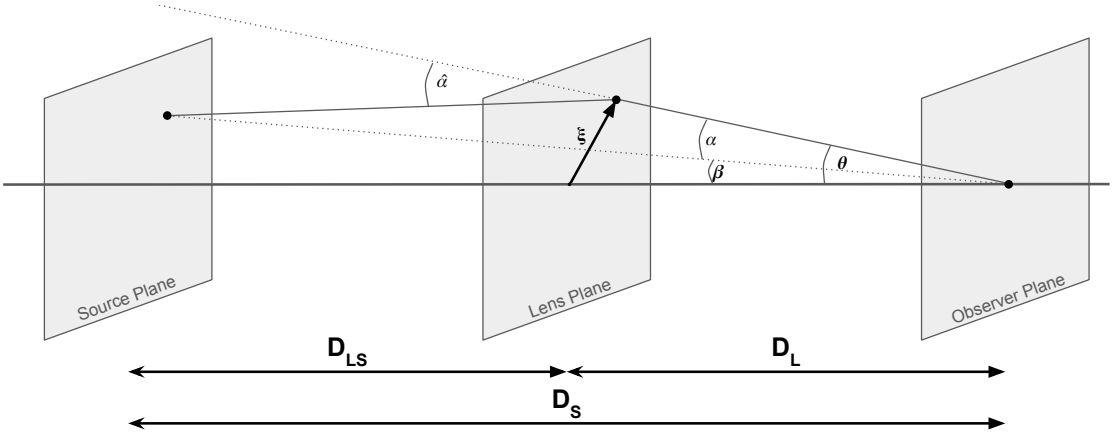


Figure 1.2: The typical gravitational lens set-up. Figure adapted from Meneghetti (2021).

a galaxy, it is common to use the *thin lens approximation*, whereby the lens is treated as a thin sheet of matter described by its surface density. From Figure 1.2, we can infer the lens equation

$$\theta D_{\rm S} = \beta D_{\rm S} + \hat{\alpha} D_{\rm LS} \tag{1.14}$$

where I note that $D_{\rm S} \neq D_{\rm L} + D_{\rm LS}$, given all three are angular diameter distances $D_A(z)$. Under this approximation, and since the deflection angle is linear in mass, the deflection angle can be treated as the sum of that from multiple point sources, which for a continuous mass distribution is given by

$$\hat{\alpha}(\boldsymbol{\xi}) = \frac{4G}{c^2} \int \Sigma(\boldsymbol{\xi'}) \cdot \frac{\boldsymbol{\xi} - \boldsymbol{\xi'}}{|\boldsymbol{\xi} - \boldsymbol{\xi'}|^2} d^2\xi' \tag{1.15}$$

where the enclosed mass is treated as function of the surface density

$$\Sigma(\boldsymbol{\xi}) = \int \rho(z, \boldsymbol{\xi}) dz. \tag{1.16}$$

For convenience, two further quantities are often defined: the critical density $\Sigma_{\rm crit} = \frac{c^2}{4\pi G} \cdot \frac{D_{\rm S}}{D_{\rm L} D_{\rm LS}}$ and the convergence, $\kappa(\boldsymbol{\xi}) = \Sigma(\boldsymbol{\xi})/\Sigma_{\rm crit}$. For an axially symmetric lens, the deflection angle is given by

$$\hat{\alpha}(\xi) = \frac{4GM(\xi)}{c^2 \xi} \tag{1.17}$$

where $M(\xi)$ denotes the lens mass enclosed in a cylinder of radius ξ. In the case that the lensed source is located directly behind the lens, $\beta = 0$, and since $\xi = D_{\rm L}\theta$ we have

$$\theta \equiv \theta_E = \sqrt{\frac{4GM(\theta_E)}{c^2}\frac{D_{\rm LS}}{D_{\rm S}D_{\rm L}}}. \tag{1.18}$$

This is called the *Einstein radius* of a lens, within which the mean convergence $\bar{\kappa}$ is equal to unity. The lens equation (Eq. 1.14) can have multiple solutions, depending on the exact mass profile of the lensing galaxy. As a consequence, gravitational lensing can produce multiple images of the same source. Furthermore, since the light forming these images travels different paths, the arrival time of each image differs. The total time-delay is dependent on this path difference combined with the difference in gravitational potential which the photons experience (Shapiro, 1964).

Lensing Potential and Magnification

The lensing potential is given by

$$\hat{\Psi}(\xi) = \frac{D_{\rm LS}}{D_{\rm L}D_{\rm S}} \cdot \frac{2}{c^2} \cdot \int \Phi(\xi, z)dz \tag{1.19}$$

where $\Phi(\xi, z)$ denotes the Newtonian gravitational potential due to the lens galaxy at position (ξ, z). The lensing potential satisfies two important properties

$$1.\ \nabla_\theta \hat{\Psi}(\theta) = \alpha(\theta)$$

$$2.\ \Delta_\theta \hat{\Psi}(\theta) = 2\kappa(\theta)$$

where $\alpha \equiv \frac{D_{\rm LS}}{D_{\rm S}}\hat{\alpha}$ and $\beta = \theta - \alpha(\theta)$.

The lensing magnification is defined as the distortion of the lensed image with respect to its unlensed counterpart, which can be derived from the Jacobian:

$$A_{ij} = \frac{\partial \beta_j}{\partial \theta_i} \tag{1.20}$$

which from the lensing equation gives:

$$A_{ij} = \delta_{ij} - \frac{\partial \alpha_j}{\partial \theta_i} = \delta_{ij} - \hat{\Psi}_{ij} \tag{1.21}$$

where I have defined $\hat{\Psi}_{ij} \equiv \frac{\partial^2 \hat{\Psi}}{\partial x_i \partial x_j}$. The magnification matrix can be split into two intuitive components

$$A_{ij} = (1-\kappa)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} - \gamma\begin{pmatrix}\cos 2\phi & \sin 2\phi\\ \sin 2\phi & -\cos 2\phi\end{pmatrix} \tag{1.22}$$

i.e., a convergence term for which the deflection is isotropic (i.e., uniform rescaling), and a shear term, $\gamma^2 = \frac{1}{4}(\hat{\Psi}_{11} - \hat{\Psi}_{22})^2 + \hat{\Psi}_{12}^2$ with a preferred direction, which changes the apparent shape of the source. The magnification is given by the inverse of the determinant of this matrix $M = 1/\det A$, and is infinite when $1 - \kappa \pm \gamma = 0$. These define the critical lines (caustics) in the lens (source) plane. While in practice infinite magnifications are not reached, magnification of $\geq 10\times$ can occur, enabling the identification of sources beyond usual detection limits. A consequence of the above is that light rays passing at different distances from the lensing mass are deflected by different amounts, changing the apparent shape of the lensed source. The shape of the unlensed source can be determined through lens modelling (Section 1.2.7).

1.2.3 Gravitational Lensing Regimes

Depending on the relative position of the lens and source galaxies, as well as the gravitational potential of the lens, gravitational lensing can be observed in one of three regimes; strong, weak and micro-lensing, described below.

Strong Lensing: Strong lensing occurs when gravitational lensing produces multiple images of the background source. Strongly lensed systems are rare because they require close angular alignment between lens and source galaxies for multiple images to be formed; in particular, the source needs to be located within one of the lens caustics. The significant magnification and presence of multiple images opens up many science pathways including the study of high-redshift galaxies and time delay cosmography. Furthermore, since the deflection is dependent on the total enclosed mass within the Einstein radius (Eqn. 1.18), the modelling of gravitational lens systems can help constrain the dark matter and stellar mass profiles within the lensing galaxy. I discuss such applications in detail in Sections 1.2.5 and 1.2.6.

Weak Lensing: The weak lensing regime is characterised by small changes in the shapes of galaxies, without the production of multiple images. This can be seen around galaxy clusters or field galaxies, or to a smaller degree by objects along the line of sight (termed *cosmic shear*). The large scale structure along the line of sight can result in correlation in the apparent shapes of galaxies which can be used to probe the density of matter in the Universe. Due to the small scale of the lensing signal, wide-field surveys such as the Kilo-Degree Survey (KiDS, Wright et al., 2025), DES (DES Collaboration et al., 2025) and Hyper-Suprime Cam (HSC, Dalal et al., 2023; Li et al., 2023) are used for cosmological studies. Weak lensing is most sensitive to the $S_8 \equiv \sigma_8 \sqrt{\Omega_m/0.3}$ parameter, which reflects the amplitude of matter density fluctuations and has been a source of debate due to differences with measurements from the early universe, namely the CMB (the S_8 tension, see Abdalla et al., 2022 for a review).

Microlensing: Finally, microlensing occurs when the deflection of light into multiple images is not resolvable, but the magnification changes over time due to the relative velocities of the lens and source objects. In these cases, the lens typically has lower mass than in the weak or strong lensing cases, and may be a star or stellar-mass black hole. Typically, the light from the lensed source follows a characteristic magnification curve. Like strong lensing (Collett, 2015; Ferrami and Wyithe, 2024), microlensing events are also rare (Wyrzykowski et al., 2023); consequently, they are typically found by survey telescopes such as *Gaia* (Gaia Collaboration et al., 2016). Microlensing has a number of applications, including searching for exoplanets (Bond et al., 2004; Tsapras, 2018) and dark matter variants such as Massive Compact Halo Object (MACHOS, Alcock et al., 2000; Calchi Novati et al., 2013). Exoplanet discoveries via microlensing will likely become numerous (Wright and Gaudi, 2013), but are hindered by the fact that such events are chance encounters, and do not repeat. Microlensing is also a source of variation in the light curves of strongly lensed quasars/supernovae, which must be accounted for when measuring the time-delay between lensed images for cosmography (e.g., Tewes et al., 2013; Liao, 2021).

1.2.4 Strong Lens Discovery through Time

Strong lenses are intrinsically rare systems. Furthermore, they can take a wide range of configurations dependent on the position of the source galaxy with respect to the caustics, and complexities in the lens mass distribution. The combination of these effects means that strongly lensed systems are challenging to find. Methods for searching for such systems vary depending on three primary factors: the data type and quality (e.g., spectroscopic, image or catalogue-level data), the target lens type (galaxy-galaxy lenses, cluster-scale lenses, or lensed transients) and the data volume. The data volume has been the primary cause of the gradual automation of lens searches over time.

Figure 1.3 shows the number of high-grade lens candidates identified by selected[5] lens searches over time. The nature of strong lens discoveries has gradually changed, from serendipitous finds and small-scale searches by individuals to automated classification of hundreds of millions of images by Machine Learning (ML) algorithms. This has been driven by technological advancements in detector sensitivity (e.g., in the NIR, Rieke, 2007) as well as survey design (Ivezić et al., 2019), which have allowed for combined wide-field sensitive surveys. In the coming years, the large-scale surveys of LSST (Ivezić et al., 2019) and *Euclid* (Euclid Collaboration: Mellier et al., 2024) will revolutionise the field where hundreds of thousands of strong lens systems are anticipated (Chapter 2 & Collett, 2015). This will be a step change for strong lensing, but will come with a number of challenges in relation to scalable lens detection and the expected sparsity of follow-up data which is addressed in this thesis.

[5]Ref: Walsh et al. (1979), Weymann et al. (1980), Huchra et al. (1985), Hewitt et al. (1988), Patnaik et al. (1992), Hewitt et al. (1992), Myers et al. (1995), Fassnacht et al. (2004), Bolton et al. (2006), Faure et al. (2008), Bolton et al. (2008), Treu et al. (2011), Gavazzi et al. (2012), Brownstein et al. (2012), Collett (2015), More et al. (2016), Petrillo et al. (2019), Jacobs et al. (2019), Sonnenfeld et al. (2020), Li et al. (2021), Cañameras et al. (2021), González et al. (2025), and Schuldt et al. (2025).

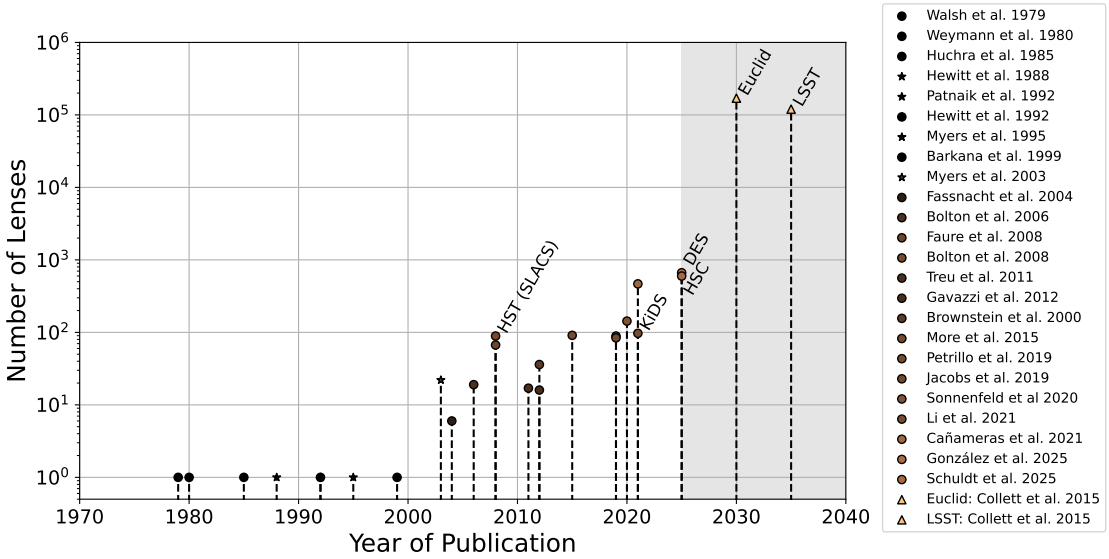


Figure 1.3: A plot of the number of high-grade lenses identified by selected lens searches, as a function of their publication date. Radio surveys are marked by a ⋆, forecasts with a ▲, and key survey searches are highlighted. The number of lens candidates identified by systematic searches has increased significantly with wide-field surveys such as DES, HSC and KiDS, and will undergo a step-change with LSST and *Euclid*.

Visual Inspection

The first strong lens system discovered was a doubly imaged quasar, identified serendipitously by Walsh et al. (1979). The next lenses that were identified by Weymann et al. (1980) and Huchra et al. (1985) were also quasars found by chance. As more lenses were identified over time, lens searches gradually became more systematic. With sufficiently small data volumes, visual inspection of a whole survey area (or all likely lenses contained within it) could feasibly be undertaken by an individual or team of researchers. Early examples include the Hubble Ultra Deep Fields (e.g., Barkana et al., 1999) and the Cosmic Evolution Survey (COSMOS) field (Faure et al., 2008; Jackson, 2008).

With larger data volumes from wide-field surveys, expert inspection has now shifted to confirmation and grading of lens candidates identified by automated and/or scalable methods. Even with strong lens experts, such classification/grading remains highly subjective. Analysis by Rojas et al. (2023) found that averaging the grades of a small team of experts would not necessarily lead to the same grade as

that from averaging over a larger team, i.e., the average grade from the smaller team had not yet stabilised. In this study, the grades were assigned a numerical score, where certain lenses were given a score of 1, probable lens: 2/3, improbable lenses: 1/3 and very unlikely lenses were assigned a score of 0. These scores were then averaged across different teams of experts of different sizes. Rojas et al. (2023) suggested ≥ 6 expert classifications were required per system to mitigate against this. For marginal systems (those receiving a score of $0.3 - 0.8$), this reduced the standard deviation between the average score from teams of 6 experts versus that from a team of ~ 20 experts to ≤ 0.1. Lens modelling can also help with this classification (see Chapter 2) but is not always readily available at this point in the classification pipeline. For ground-based imaging, spectroscopic confirmation of two distinct redshifts from the lens/source galaxy are typically required to confirm unambiguously that a system is strongly lensed.

Citizen Science

Citizen science, the public participation in scientific research, has been used across a wide range of research fields from ecology (e.g., Mac Aodha et al., 2018) to ancient history (e.g., Williams et al., 2014). In recent years, the Zooniverse[6] platform, which originated from the citizen science galaxy morphology project Galaxy Zoo (Lintott et al., 2008), has become a central hub for such projects, including searches for strong lenses. Following serendipitous lens discoveries in the Galaxy Zoo project, a dedicated strong lens project, *Space Warps*, was introduced by Marshall et al. (2016) and More et al. (2016). This combined the citizens' classifications with a Bayesian model (via the Space Warps Analysis Pipeline, SWAP, Marshall et al., 2016) to determine lens probabilities (scores) based on the performance of the citizens on training images. In this model, the citizens are assigned a skill based on their performance when classifying training subjects with a known ground truth, and each system to be classified (known as a 'subject') is assigned a corresponding score. This score changes each time the subject is classified by a user, with the

[6] https://www.zooniverse.org/

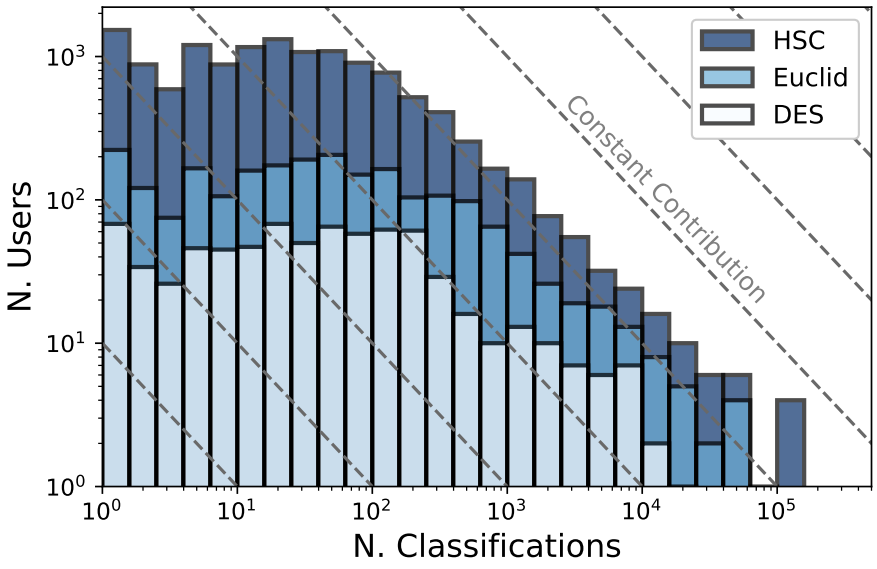


Figure 1.4: The distribution of classification contributions in three recent Space Warps lens searches: HSC: Sonnenfeld et al. (2020), DES: González et al. (2025), *Euclid*: Euclid Collaboration: Walmsley et al. (2025). The contributions roughly follow the 'Pareto Principle', for example the $\mathcal{O}(1)$ users who make $\mathcal{O}(10^5)$ classifications make a roughly equal contribution (neglecting the difference in allocated user skill) to the combined efforts of $\mathcal{O}(1000)$ users who each make $\mathcal{O}(100)$ classifications.

degree of change in the score dependent on the user's skill. Due to the lack of known lenses, More et al. (2016) used simulated lens systems painted[7] into real Canada-France-Hawaii Telescope Legacy Survey (CFHTLS) images as training systems, a strategy which has continued in subsequent searches. Due to the rarity of strong lens systems, Space Warps projects have primarily prioritised classification speed rather than detailed multi-level classifications (contrasting with Galaxy Zoo, Lintott et al., 2008), with a single 'yes/no' classification enabling individual users to classify thousands of systems. As shown in Figure 1.4, it is common for a large number of citizens to each make a small contribution, and a much smaller number to classify tens of thousands of systems. To date, Space Warps has conducted lens searches in data from CFHTLS (Marshall et al., 2016; More et al., 2016), the VISTA-CFHT Stripe 82 (Geach et al., 2015), HSC (Sonnenfeld et al., 2020), and as will be detailed in this thesis, DES (González et al., 2025) and *Euclid* (Euclid Collaboration: Walmsley et al., 2025), totalling ~ 22 million classifications. Volunteers have proven

[7]Using `SIMCT`, https://github.com/anumore/simct/tree/main/code

adept at identifying strong lens systems, outperforming state-of-the-art machine learning methods and identifying novel/rare lens configurations, such as red lensed Einstein rings (Geach et al., 2015) which may not appear in typical ML training sets. Inspection by citizens significantly increases the total human inspection budget and can provide an invaluable verification step following machine learning searches which may flag out-of-distribution artifacts as lenses. The prospects of combining the expertise of citizens with machine learning models and expanding such lens searches to wide-field surveys are discussed in detail in Chapters 3 and 4.

Machine Learning

Machine learning offers a mechanism to process rapidly large quantities of data, and is thus well suited to strong lens classification. The network architectures can vary significantly between searches, though Convolutional Neural Networks (CNN) are the most common, with Vision Transformers (VT, Andika et al., 2023; González et al., 2025), Support Vector Machines (SVM, Hartley et al., 2017) and decision trees (Khramtsov et al., 2019) having also been used. They are trained on large datasets of lenses and non-lenses using a penalty term ('loss') which is optimised by gradually updating the weights/biases within the network. As in citizen science searches, true lens systems cannot be used exclusively for training due to their rarity and so training images are often simulated. These simulations have to be realistic, with accurate survey noise properties, to prevent the network from confusing the presence of a lens with simulated structures/noise in the cutout. Over time, such machine learning networks have become more complex, including multi-band images (e.g., Cañameras et al., 2024; He et al., 2025; Shu et al., 2022), images with different resolution/pixel-scales (Melo et al., 2024), and real survey images for the lens and source galaxies (e.g., Cañameras et al., 2020; Rojas et al., 2022). While ML models have significant advantages for lens classification (in search speed and performance), their high scoring samples remain heavily contaminated by non-lenses. Furthermore, they can also lack interpretability (*why* a particular system is given a high/low score can be difficult to determine), and can be confused

by systems which do not appear in their training set. Therefore, the training sets need to contain the full variety of lens configurations, artifacts and non-lenses which may appear in the survey images. Strategies for improving overall lens classification performance are the subject of Chapters 3 and 4.

1.2.5 Astrophysical Applications of Strong Lenses

Strong lenses can be an invaluable tool for a range of astrophysics applications. Here I detail some of the primary ways in which lenses can be used.

Baryon and Dark Matter Profiling

The Einstein radius of a strong lens is dependent on the total mass contained within that radius (Eqn. 1.18). Furthermore, the exact deflection is dependent on the mass profile (Eqn. 1.15) with, for example, narrower lensed arcs forming for steeper mass profiles (Marshall et al., 2007; Galan et al., 2024). The combination of lens modelling and photometry (via Stellar Population Synthesis, SPS, models) can put constraints on both the baryonic (stellar) and dark matter mass profiles within the lens galaxy. The combined stars+dark matter mass profile in early-type galaxies has been found to be very close to isothermal (Treu et al., 2006; Koopmans et al., 2009), colloquially known as the 'bulge-halo conspiracy', even though neither profile is isothermal individually. There is differing evidence as to whether the mass density slope evolves with redshift, with some lensing studies favouring a steepening over time (Sonnenfeld et al., 2013; Chen et al., 2019; Geng et al., 2025) and cosmological simulations favouring the opposite (Xu et al., 2017; Remus et al., 2017). A larger lens sample (as expected from *Euclid* and LSST) will help resolve this tension, but will have to account for the selection effects discussed in Section 1.2.8.

Stellar Initial Mass Function

The stellar initial mass function (IMF) describes the mass distribution of stars within a galaxy. The Milky Way is well described by a Chabrier IMF (Chabrier, 2003), while galaxies with a greater number of (faint) dwarf stars are described as 'bottom-heavy' closer to a Salpeter IMF. Our current understanding of the IMF

between galaxies is still developing, with the current belief that such a function may not be universal between galaxies (Spiniello et al., 2014; Martín-Navarro et al., 2023), or even within a single galaxy (Martín-Navarro et al., 2015; Barbosa et al., 2021). Strong lensing measurements can help constrain this. The lensing deflection is dependent on all matter within the galaxy, and so is influenced by the mass of dwarf stars which may not contribute significantly to the galaxy's light, but are sufficiently numerous to contribute significantly to its mass (Conroy and van Dokkum, 2012). The lens galaxy in a strong lens system is typically at $z \sim 0.5 - 1$, but can extend to $z \sim 2$ (see Chapter 2). Comparing mass measurements from lensing with stellar mass measurements from SPS models (with a fixed choice of IMF) can help constrain the IMF of the galaxy as well as its possible evolution.

The High-redshift and Dusty Universe

Strong lensing brings two primary benefits to high-redshift astrophysics. Firstly, the magnification effect allows faint objects to be studied which otherwise would be undetectable, and secondly this magnification increases the resolution with which such objects can be studied, by increasing their angular size on the sky. NIR telescopes (or telescopes with NIR bands) such as *JWST*, *Euclid* and *Roman* will be able to utilise such benefits to probe redder galaxies, i.e., those at high redshift or with significant dust or older stellar populations. The sensitivity of these surveys combined with the magnification due to lensing will enable the investigation of more typical galaxies than would otherwise be detectable. For example, identification of lensed dusty star forming galaxies would be of interest. These galaxies are undergoing intense star formation (SFR $\sim 100 - 1000\,\mathrm{M}_\odot\,\mathrm{yr}^{-1}$), but due to their dust content much of this is obscured, and thus only visible at longer wavelengths. By probing such galaxies we can understand more about galaxy evolution at cosmic noon, including the formation of early type elliptical galaxies (Toft et al., 2014) and the evolution of Active Galactic Nuclei (AGN, Stacey et al., 2018, see Casey et al., 2014 for a review) for which strong lensing can be a valuable tool. The

prospects and opportunities for identifying such lensed systems in the above NIR telescopes are discussed in Chapter 2.

1.2.6 Cosmological Applications of Strong Lenses

Time Delay Cosmography

One feature of strong lensing, multiple imaging, is caused by light rays travelling different paths from the lensed source to the observer. This difference in path, and the difference in the paths' potentials, causes the multiple images to arrive at different times, and due to the distance of the lensed source (typically at $z \sim 1$), this time-delay is sensitive to the Universe's expansion. This has given rise to the technique of time-delay cosmography, whereby the light curves from each lensed image of a transient object (AGN or supernova) are measured and compared to infer the Hubble constant.

Measuring the time delay constrains the time delay distance (see e.g., Birrer et al., 2020)

$$D_{\Delta t} \equiv (1 + z_{\rm L}) \cdot \frac{D_{\rm L} \cdot D_{\rm S}}{D_{\rm LS}} \tag{1.23}$$

which is inversely proportional to the Hubble constant and depends on the measured time delay Δt via

$$\Delta t = \frac{D_{\Delta t}}{c}[\phi(\theta_A, \beta) - \phi(\theta_B, \beta)] \tag{1.24}$$

where ϕ (termed the Fermat Potential) is given by $\phi(\theta_i) = [\frac{1}{2}(\theta_i - \beta)^2 - \hat{\Psi}(\theta_i)]$ for lensed images $i \in \{A, B\}$. The precision on the Hubble constant achievable through strong lensing varies dependent on assumptions made on the mass profile of the lens. For example, Birrer et al. (2020) adopted a conservative approach achieving $H_0 = 67.4^{+4.1}_{-3.2}\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (i.e., 6% precision) with 7 lenses, while Wong et al. (2020) used more restrictive lens mass model assumptions to measure $H_0 = 73.3^{+1.7}_{-1.8}\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (2.4% precision) with 6 lensed quasars. By comparison, measurements of the CMB suggest $H_0 = 67.4 \pm 0.5\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Planck Collaboration et al., 2020) which is in significant tension with local distance ladder

measurements which measure $72.53 \pm 0.99\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Riess et al., 2022). While the precision of strong lensing estimates varies, they are typically closer to those of the local distance ladder measurements (Di Valentino et al., 2021; Treu et al., 2022). Time delay cosmography relies on precise mass models (illustrated by the ϕ terms above) with follow-up kinematic data and accurate measurements of the time delays from high-cadence observations. With the forthcoming LSST survey, it is anticipated that $\mathcal{O}(10^3)$ lensed quasars and $\mathcal{O}(10^2)$ lensed supernovae will be detectable (Oguri and Marshall, 2010; Arendse et al., 2024). By using these as well as data from *Euclid*, *Roman* and follow-up facilities, Treu et al. (2022) anticipate that a precision of $\sim 1\%$ (i.e., comparable to the CMB and distance ladder measurements) is achievable by the end of the decade.

Galaxy-Galaxy Lenses

As shown by Eqn. 1.18, the Einstein radius of a strong lens is dependent on the angular diameter distances to the lens and source. Such distances are dependent on cosmological parameters (as demonstrated in Figure 1.5), which can therefore be inferred from a sample of galaxy-galaxy lenses (Grillo et al., 2008). Such inference relies on accurate measurements of the mass of the lens (typically through measurements of the velocity dispersion), Einstein radius, lens/source redshifts and assumptions on the mass model of the lens. For a singular isothermal model, this relation is given by (Grillo et al., 2008)

$$r \equiv \frac{D_{\mathrm{LS}}(z_{\mathrm{L}}, z_{\mathrm{S}}, \boldsymbol{\Omega})}{D_{\mathrm{S}}(z_{\mathrm{S}}, \boldsymbol{\Omega})} = \frac{c^2 \theta_E}{4\pi\sigma_v^2}\,. \tag{1.25}$$

Given their relative abundance in comparison to double source-plane lenses (below) and lensed transients, the constraints obtainable by static galaxy-galaxy lenses are comparable to those from other lensed systems (Shajib et al., 2024). Li et al. (2024) produced a cosmological forecast based on 10 000 strongly lensed systems imaged by *Euclid* (based on the number expected to receive spectroscopic follow-up), finding that such lenses would provide the tightest single-probe constraint on w to date. This probe is the subject of Chapter 5, in which I extend this forecast to the

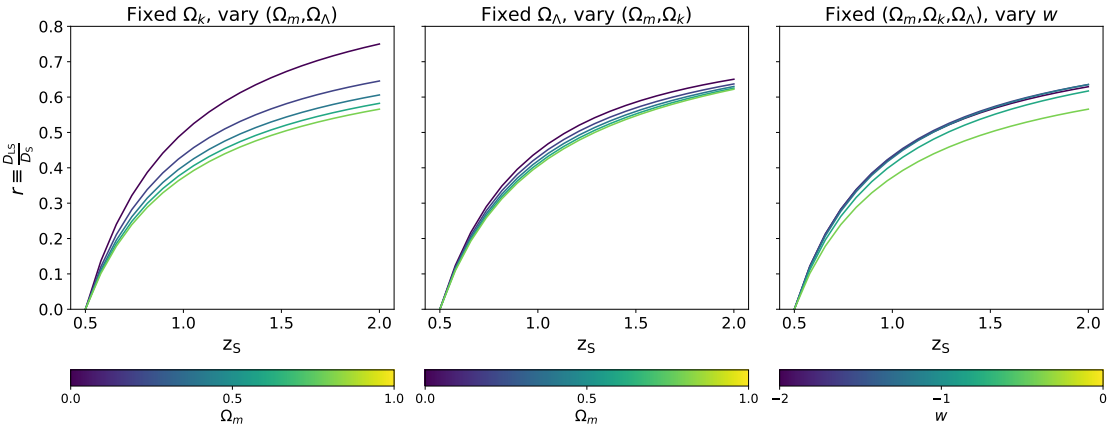


Figure 1.5: Plot of the dependence of the lens-source distance ratios with cosmology, plotted for a fiducial lens system at $z_L = 0.5$. Figure adapted from Grillo et al. (2008).

much larger photometric sample of strong lenses (i.e., those which do not receive spectroscopic follow-up) accounting for the effects of contamination by non-lenses.

Double Source Plane Lenses

Double source plane lenses (DSPL), whereby a single lens galaxy lenses two or more source galaxies, can remove some of the uncertainties faced by single galaxy-galaxy lens systems, for example by probing the enclosed mass at two distinct radii and thus measuring the mass density slope. While cluster lenses often lens multiple background galaxies, the complex mass distributions in the lensing cluster make them much more challenging subjects for cosmology. The ratio of the two Einstein radii for a galaxy-galaxy lens is independent of the Hubble Constant and only weakly dependent on the mass model. This ratio is given by

$$\beta \equiv \frac{\theta_1}{\theta_2} = \frac{D_{\mathrm{LS},1} D_{\mathrm{S},2}}{D_{\mathrm{LS},2} D_{\mathrm{S},1}} \tag{1.26}$$

for an Singular Isothermal Sphere profile (SIS, Collett et al., 2012). However, these systems are very rare with roughly 1 in 80 strong lenses in HST being a compound lens (Gavazzi et al., 2008). Only one triple-plane lens (Collett and Smith, 2020; Smith and Collett, 2021) has been identified to date, however this is expected to increase to $\sim 1000 - 2000$ DSPL's (Euclid Collaboration: Li et al., 2025) and ~ 40 triple-source plane lenses (Collett and Auger, 2014) in the *Euclid* survey.

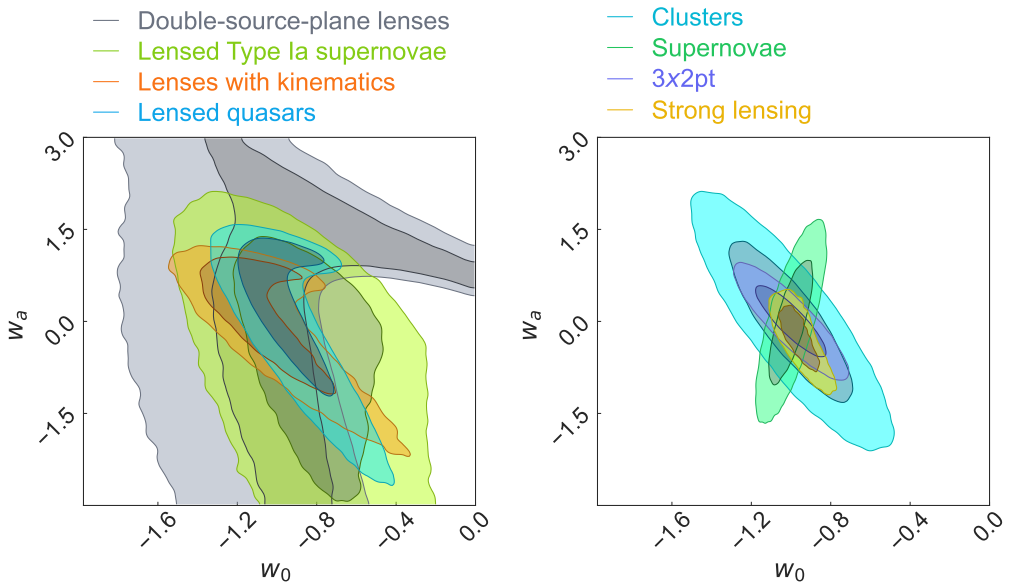


Figure 1.6: Cosmological constraints expected for 10 years of LSST data, from individual strong-lensing probes (a), and from combining multiple probes (b). Figure taken from Shajib et al. (2024).

Combining Strong Lensing Probes

Cosmological probes are often combined to achieve tighter constraints than can be reached with single probes alone. The differences in methodology and lens populations between the above probes (for example, differences in the redshift distributions between DSPL's and single-plane galaxy-galaxy lenses) give rise to different degeneracies between cosmological parameters, as shown in Figure 1.6. Assuming that the constraints between lens probes are independent, they can simply be multiplied together; however, in reality, combining probes must account for the possibility of shared covariances between probes to prevent over-estimating their combined constraining power. The combined strong lensing probes are expected to provide the tightest constraints on w_0, w_a in LSST (Figure 1.6, Shajib et al., 2024); however, this will rely on accurate and scalable modelling, the techniques for which are described in the following section.

1.2.7 Strong Lens Modelling

To achieve the science goals of the astrophysical and cosmological applications described above, a precise and accurate lens model is often required. Lens modelling requires a model for the lens and source light profiles and the lens mass distribution, as well as accounting for any external factors (such as neighbouring objects or line-of-sight effects). These models can either take a parametric form or be free-form (determined on a pixel-by-pixel basis). The approach taken for a particular system is governed by the available data. For example, single galaxy-galaxy lenses imaged from a ground based telescope may be well fitted with a single parametric form for the lens and source galaxies, whereas space-based imaging of a lensing cluster would require a much more complex model.

Common Lens Models

The most common parametric lens models are SIS (or Singular Isothermal Ellipsoid, SIE), Power-law Elliptical Mass Distribution (PEMD), and Navarro-Frenk-White (NFW)+Sérsic profiles. The convergence of an SIS profile is given by

$$\kappa(\theta_x, \theta_y) = \frac{1}{2} \cdot \frac{\theta_E}{\sqrt{\theta_x^2 + \theta_y^2}} \tag{1.27}$$

This automatically encapsulates the 'bulge-halo conspiracy', and notably has infinite density at its centre. A slightly more complex distribution is provided by the PEMD mass model (Barkana, 1998), in which the mass density slope is allowed to vary from the isothermal case of $\gamma_{\rm lens} = 2$. The convergence of the PEMD profile is given by

$$\kappa(\theta_x, \theta_y) = \frac{3 - \gamma_{\rm lens}}{2} \left(\frac{\theta_E}{\sqrt{q_{\rm lens}\theta_x^2 + \theta_y^2/q_{\rm lens})}} \right)^{\gamma_{\rm lens}-1} \tag{1.28}$$

where $q_{\rm lens}$ denotes the axis ratio. Changing the density slope of a lens changes the lensing cross-section (Sonnenfeld et al., 2023), and steeper slopes have narrower lensed arcs (Marshall et al., 2007; Galan et al., 2024). Such a profile has often been used when the available imaging is space-based (Collett and Auger, 2014; Wagner-Carena et al., 2023; Erickson et al., 2024; Sheu et al., 2024) given this is more flexible than the isothermal model above.

Parametrised models can lead to spurious results when the parametrisation is not a good fit for the system. Investigations by Etherington et al. (2024) found that measurements of external shear values measured with a PEMD model on HST data were inconsistent with those from weak lensing. More complex models can (and should) therefore be used when the data allows; a common choice is to split the mass components into two; a Sérsic and an NFW profile. The 2D Sérsic profile follows[8]

$$I(R) = I_0 \exp\left[-b_n\left(\left(\frac{R}{R_e}\right)^{1/n} - 1\right)\right] \tag{1.29}$$

while a circular NFW profile is described by two free parameters ρ_0, R_s, following

$$\rho(r) = \frac{\rho_0}{\frac{r}{R_s}(1+\frac{r}{R_s})^2} \tag{1.30}$$

and thus $\rho \propto r^{-1}$ at small radii. Here, the stars are represented by the Sérsic profile, and the dark matter modelled by the NFW.

There are two main approaches to lens modelling: forward modelling and machine learning. I discuss each of these techniques, their advantages and disadvantages below.

Conventional Forward Modelling

Conventional forward modelling explicitly evaluates a lens model against data. To do so requires a likelihood (or posterior) function, and regularisation if the lens/source models are pixelated rather than parametrised. Such regularisation can be used to prevent large curvature or ensure smoothness in the source light profile. For a model parametrised by variables $\boldsymbol{\nu}$, the posterior $P(\boldsymbol{\nu}|\boldsymbol{d})$ for data $\boldsymbol{d}$ is given by Bayes Theorem

$$P(\boldsymbol{\nu}|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|\boldsymbol{\nu})P(\boldsymbol{\nu})}{P(\boldsymbol{d})} \tag{1.31}$$

where $P(\boldsymbol{\nu})$ is termed the prior, $P(\boldsymbol{d}|\boldsymbol{\nu})$ is the likelihood, and $P(\boldsymbol{d})$ is the evidence. For Gaussian errors, the likelihood is given by

$$P(\boldsymbol{d}|\boldsymbol{\nu}) \propto \exp\left(-\frac{1}{2}\{\boldsymbol{d}-\boldsymbol{m}(\boldsymbol{\nu})\}^T \cdot \Sigma^{-1} \cdot \{\boldsymbol{d}-\boldsymbol{m}(\boldsymbol{\nu})\}\right) \tag{1.32}$$

[8]Here, n denotes the Sérsic index, R_e is the half-light radius and b_n obeys $\Gamma(2n) = 2\gamma(2n, b_n)$ with Γ and γ representing the complete and incomplete gamma function, respectively.

where $\boldsymbol{m}(\boldsymbol{\nu})$ denotes the pixel values of the image produced by model parameters $\boldsymbol{\nu}$, and Σ is the data covariance. Lens modelling codes may simply find the maximum of the likelihood, or sample over all possible lens variables (e.g., via Markov Chain Monte Carlo, MCMC) to give a full posterior, which may also require marginalising over nuisance variables to generate realistic uncertainties.

Since evaluating the likelihood may require reconstructing a lens image (or lensed image positions) for a given lens model, this method can be slow (taking of order hours/lens, Rojas et al., 2022; Schmidt et al., 2023). Furthermore, such models often require manual refinement, such as masking neighbouring unlensed objects in the image, which further increases the time-cost and reduces the scalability to large datasets. However, for a sufficiently complex functional form for the lens and source, this method can usually provide an accurate model for a given lens. Numerous examples of forward modelling software exist, such as `Lenstronomy` (Birrer and Amara, 2018), `GLEE` (Suyu and Halkola, 2010; Suyu et al., 2012), `LENSTOOL` (Jullo et al., 2007; Kneib et al., 2011), `GRAVLENS` (Keeton, 2001; Keeton, 2011), and `pyAutolens` (Nightingale et al., 2018; Nightingale et al., 2021). Given the growing number of strong lens candidates, attention is now moving to gradient-based, GPU accelerated forward modelling methods such as `Herculens` (Galan et al., 2022a; Galan et al., 2022b), `GIGA-Lens` (Gu et al., 2022) and `TinyLensGPU` (Cao et al., 2025). These can offer significant reductions in computation time while still following a Bayesian or maximum-likelihood approach, reaching ~ 1 minute per system. Even so, such methods would still require months of computation time to model the $\mathcal{O}(10^5)$ galaxy-scale lenses expected in forthcoming surveys (Collett, 2015, and see Chapter 2), which can be reduced by machine learning methods.

Modelling via Machine Learning

Machine learning lens modelling algorithms are trained by iteratively updating the weights and biases of a neural network in response to a loss function. This loss function measures the performance of the network against a set of training data with known model parameters. The trained network can then be applied to real

images to estimate lens parameters. Using ML algorithms for lens modelling was pioneered by Hezaveh et al. (2017), and has since become widespread. While these models usually require large training sets, $\mathcal{O}(10^5)$ images and take of-order hours to train (e.g., Erickson et al., 2024), once trained they can be evaluated rapidly ($\lesssim 1$ second/lens, Hezaveh et al., 2017; Erickson et al., 2024) over a large dataset, giving a significant advantage over forward-modelling techniques.

Simple convolutional neural network modelling involves producing point estimates of lens parameters. Such models have been trained on single- and multi-band images (Hezaveh et al., 2017; Pearson et al., 2019; Schuldt et al., 2021; Gawade et al., 2024), for a range of different telescopes including HST, HSC, LSST and *Euclid*. However, without accurate uncertainty estimates (and the unknown behaviour of the network to unexpectedly complex test data), such model parameters are of limited use for subsequent analysis.

Simulation Based Inference (SBI) offers a method of determining a complete posterior. The simplest form of SBI is Approximate Bayesian Computation (ABC), whereby models are drawn from a prior distribution and retained if they closely match the data, or else discarded. The parameters of those which remain form an approximate posterior (e.g., see Grazian and Fan, 2019, for a review). However, this method does not scale well to large test datasets. Neural networks can offer a solution via neural density estimation techniques. These methods can be divided into Neural Likelihood Estimators, Neural Ratio Estimators (evaluating the likelihood-evidence ratio) and Neural Posterior Estimators. For Neural Posterior Estimation (NPE), the network parameters, ϕ are tuned during training through the loss function,

$$L(\phi) = -\sum_{i}^{N_{\text{train}}} \log q_\phi(\boldsymbol{\nu}_i|\boldsymbol{d}_i) \tag{1.33}$$

which is summed over N_{train} training data and which is minimised when $q_\phi(\boldsymbol{\nu}|\boldsymbol{d}) = p(\boldsymbol{\nu}|\boldsymbol{d})$, the posterior (Papamakarios and Murray, 2018).

NPE and SBI have been applied to a range of strong lens image data, including lens modelling of HST (Erickson et al., 2024) and simulated DES (Poh et al., 2025) images, as well as determination of the subhalo mass function (Wagner-Carena

et al., 2023) - I adopt the NPE method in Chapter 5. While these methods can offer significant speed improvements compared to conventional methods, careful thought must be put into the training set. The networks must be trained on realistic simulations encompassing all likely lens configurations to ensure that at test-time, the network continues to provide accurate models.

1.2.8 Strong Lensing Today: Challenges and Opportunities

With the arrival of LSST and *Euclid*, there will be a step-change in the number of strong lens candidates (Figure 1.3, Collett, 2015, Chapter 2). This will precipitate a significant change in lens finding, modelling and subsequent analysis. It is likely that strong lens candidates will have tiered levels of additional data; a small subset of systems will receive significant follow-up enabling forward-modelling with complex parametrisations, while the remainder (likely the vast majority) will be the focus of more easily scalable modelling and analysis methods. Tying these subsets together will provide the greatest scientific benefit, but will require careful consideration of all the systematics involved, which are discussed below.

Accounting for Systematics

Mass-Sheet Degeneracy: The mass-sheet degeneracy stems from a transformation in the lens equation whereby the lensed images are unchanged when inserting a uniform mass sheet into the lens plane and correspondingly adjusting the source position. The mass-sheet transformation adjusts the lens equation (Eqn. 1.14) to

$$\lambda\beta = \theta - \lambda\alpha(\theta) - (1-\lambda)\theta \tag{1.34}$$

with convergence

$$\kappa_\lambda(\theta) = \lambda\kappa(\theta) + (1-\lambda) \tag{1.35}$$

Such a transformation adjusts the inferred time delay (and thus inferred Hubble constant), but does not affect the enclosed mass at the Einstein radius (Unruh et al., 2017). This degeneracy can be broken with additional information on the mass profile of the lens, such as stellar kinematics (e.g., Treu and Koopmans, 2002).

Without this, the mass-sheet degeneracy can be accounted for by including the mass-sheet parameter λ as a free parameter and marginalising over its uncertainties as part of the inference. Alternatively, asserting a particular functional form for the lens mass distribution automatically removes the degeneracy since the mass-transformed mass distribution will not be of the same functional form. Which of these two methods is suitable (the former being the much more conservative option) will depend on the particular science case and the degree to which prior information is chosen to be used.

Selection Function: Strong lenses are biased tracers of the general galaxy population, and strong lens candidates resulting from a lens search are typically biased tracers of the detectable lens population. Using strong lenses as a tool to analyse galaxies as a whole requires accounting for these affects.

Regarding the former, Sonnenfeld et al. (2023) identified that lens galaxies are biased towards more massive, compact galaxies, and higher values of $\alpha_{\rm sps}$ (defined as the ratio of the galaxy's true stellar mass to that which would be measured from photometry alone). Sonnenfeld (2024) and Sonnenfeld (2025) took this further, debiasing the population-level parameters inferred from the Sloan Lens ACS (SLACS) lenses (Bolton et al., 2006). In particular, Sonnenfeld (2024) found that such debiasing could significantly affect (or remove entirely) the inferred evolution of the mass density profile of galaxies, perhaps alleviating discrepancies with theoretical predictions. Consequently, accounting for selection effects can be crucial to unbiased inference of the underlying galaxy population.

Beyond the intrinsic lens selection bias, lens searches produce a biased sample of the detectable lenses in a given survey (e.g., Euclid Collaboration: Walmsley et al., 2025). Which lenses are identified (and given which grade) will depend on the properties of the lens search (for example whether or not lens subtracted images were used, or which filters were applied) as well as the particular lens classifier (e.g., an ML classifier, citizen science or expert inspection). A multi-stage search would encompass a different selection function at each stage. Quantifying this requires

realistic simulated lenses spanning the complete detectable range, to determine the likelihood of a given lens being identified in a search, as discussed in the *Euclid* lens search described in Chapter 4 (Euclid Collaboration: Walmsley et al., 2025).

1.3 Thesis Overview

The work in this thesis aims to prepare for the challenges and opportunities which the new era of strong lensing will bring. The structure of the thesis is outlined below.

In Chapter 2, I estimate the occurrence rate of strong lenses in a range of NIR surveys. This is achieved by combining an empirical galaxy catalogue with realistic strong lens simulations and detectability thresholds. I discuss the opportunities which lens searches in the NIR will bring, in particular with deep surveys from JWST which will extend the strong lens population to a much higher redshift than previously observed.

In Chapter 3, I turn to strong lens classification. The paucity of strong lens systems means that strong lens searches currently suffer from a 'false positive problem', whereby high-scoring systems from lens classifiers are often overwhelmingly false positives. In this chapter I present a possible solution, combining multiple lens classifiers of different types into an ensemble, and discuss methods to produce calibrated probabilities that a given system is strongly lensed.

In Chapter 4, I outline the results of a systematic lens search in data from the Euclid Wide Survey (EWS). I then apply the ensemble methodology from Chapter 3 and demonstrate that, while current state-of-the-art lens classifiers will identify tens of thousands of lenses, even space-based *Euclid* data is affected by the false positive problem. I finish by presenting mitigations to this problem and expectations for the full EWS.

In Chapter 5, I turn to cosmological inference. As discussed in Section 1.2.6, large samples of strong lenses can provide tight constraints on the evolution of the dark energy parameter $w(z)$. However, as will be demonstrated in Chapters 3 and 4, the large lens samples expected from LSST and *Euclid* will be accompanied by significant numbers of false positives. In this chapter I model a large sample of

simulated LSST strong lenses via NPE and present a framework to infer unbiased cosmological parameters from an impure dataset of strong lenses. Such a framework could be extended easily to infer other hyperparameters in the strong lens population, and represents a significant first step in large-scale inference from the photometric (i.e., unconfirmed) sample of strong lenses.

In Chapter 6, I present my conclusions and discuss future work to prepare for and exploit forthcoming strong lens data.

2

The Occurrence Rates of Strong Lenses in NIR Surveys

The basis of this chapter was first published in the journal article 'On the detectability of strong lensing in near-infrared surveys', Holloway et al., 2023.

Contents

2.1 Introduction 33
2.2 Data 35
2.2.1 Adaptions to the Galaxy Catalogue 37
2.3 Method 39
2.3.1 Frequency of Galaxy-Galaxy Conjunctions 39
2.3.2 Calculating Lensing Properties 41
2.3.3 Detectability Constraints 42
2.4 Results 46
2.4.1 Verifying the Simulations 46
2.4.2 The Observable Lens Population 48
2.4.3 Lens and Source Population Properties 53
2.4.4 Extrapolating to Wide-Field Surveys 57
2.5 Discussion 59
2.5.1 Number Density of Detectable Lens Systems 59
2.5.2 Properties of the Detected Strong Lenses 62
2.5.3 Validation of Lens Occurrence Rates with Recent Strong Lens Discoveries 64
2.5.4 Potential Further Improvements 70
2.6 Conclusion 71

In the coming years, telescopes such as JWST, Euclid and Roman will allow for dedicated strong lens searches in the NIR. In this chapter I present estimates for the occurrence rates and properties of strong lenses in these surveys, verified against published lens samples at visible wavelengths, and discuss recent discoveries in JWST and early Euclid data.

2.1 Introduction

Historically, lens searches have been conducted primarily through selection at visible, sub-mm and radio wavelengths, for example HSC (visible, e.g., Sonnenfeld et al., 2020), the Herschel Astrophysical Terahertz Large Area Survey (H-ATLAS, sub-mm, Negrello et al., 2010) and the Cosmic Lens All-Sky Survey (CLASS, radio, Myers et al., 2003; Browne et al., 2003). Searches in the visible are typically based on the properties of the deflector galaxy, i.e., selecting the most massive galaxies (e.g., early-type galaxies, or bulges of spiral galaxies) hence generating so-called deflector/lens-selected samples. On the other hand, lensed sub-mm and radio galaxies can be identified by selecting sources at the bright end of the flux distribution (a so-called source-selected sample, such as identified by Negrello et al., 2010) with contaminants, radio blazars (AGN) and low-redshift galaxies, easily removed. For a high enough threshold, such fluxes can only be explained by lensing rather than star formation, producing a very pure lens sample.

By contrast, the intervening NIR regime has not typically been the focus of strong lensing searches, primarily due to the historic lower sensitivity of NIR detectors limiting the depth of such surveys. However, lens searches in the NIR may reveal higher redshift lenses, extending probes on the IMF (Spiniello et al., 2011; Sonnenfeld et al., 2019) and mass density slope (Etherington et al., 2023; Sheu et al., 2024) and reveal high-redshift lensed sources. Moreover, NIR lens searches can help to reveal lensed passive galaxies, such as identified by Muzzin

et al. (2012). In addition, NIR counterparts to lensed dusty star-forming (sub-mm) galaxies may be detectable which might otherwise be missed in a visible lens search. Such galaxies are believed to evolve to form elliptical galaxies as seen in the local universe (Toft et al., 2014; Simpson et al., 2014) and play a key role in our understanding of galaxy evolution. Bright sub-mm galaxies are highly dust-obscured with high star-formation rates (e.g., Swinbank et al., 2013; Reuter et al., 2020). The magnification provided by gravitational lensing increases the achievable spatial resolution allowing for more detailed analysis, for example, through continuum imaging and identifying fainter Infrared (IR) spectral lines with the Atacama Large Millimeter Array (ALMA, e.g., Spilker et al., 2014; ALMA Partnership et al., 2015; Rybak et al., 2015; Maresca et al., 2022) even reaching comparable spatial resolution at $z \sim 3$ to that achievable at low redshift (Rybak et al., 2015). In summary, NIR lenses offer an opportunity to explore new regions of the redshift-dust-age parameter space, in particular high-redshift, dusty or older lensed galaxies.

The dramatic improvement in NIR detectors over the last two decades has paved the way for wide-area NIR sensitive surveys e.g., with the United Kingdom Infrared Telescope (UKIRT) and Visible and Infrared Survey Telescope for Astronomy (VISTA). The arrival of *JWST* and *Euclid* telescopes, together with the forthcoming *Roman* telescope, will allow for detailed study of the intervening regime between lenses identified in sub-mm wavelengths and those from searches in the visible. The narrow surveys of *JWST* will allow a small number of lensed systems and lensing clusters to be analysed in detail (e.g., studying the SMACS 0723 cluster Caminha et al., 2022; Pascale et al., 2022), while the wider-area NIR data from *Euclid* and *Roman* will provide much larger statistical samples of lenses with diffraction-limited resolution. In this chapter I explore the strong lens discovery prospects in the NIR with simulations that are well suited to medium-deep VISTA surveys and those conducted by *JWST*. I also present illustrative figures for *Euclid* and *Roman* surveys (Section 2.4.4).

Predictions for strong lens detection have previously been made by Collett (2015) who focused on surveys in the visible and used distinct lens and source

populations, the former restricted to elliptical galaxies. While this is a sensible choice given that most lenses will be massive ellipticals, here I present estimated lens frequencies based on a less restrictive prior population (see Section 2.2) and extend such estimates to the NIR. These simulations could include less common strong lenses (e.g., the bulges of late-type spirals, Treu et al., 2011) and would allow us to understand the complex lensing selection function for an untargeted search in a given survey (Sonnenfeld, 2022). In principle, any foreground galaxy can lens a background source. Therefore, I drew lenses and sources from the same generated catalogue, allowing detectability and strong-lensing criteria to determine which lens-source pairs would be identified in a given survey. This study presents a framework to assess the detectability of strong lenses for a chosen survey using self-consistent Spectral Energy Distribution (SED) and morphological information.

The chapter is structured as follows. In Section 2.2 I describe the data I used to generate the strong-lensing frequency estimates by the method described in Section 2.3. I detail the results in Section 2.4 including verifying the simulations against existing lens searches and detailing the properties of the lens and source populations in detectable lensing systems. In Section 2.5 I discuss prospects for *JWST* and lens searches in the NIR and I present the conclusions in Section 2.6.

In this chapter I assume a flat ΛCDM cosmology from the Planck Collaboration (Planck Collaboration et al., 2020) with $H_0 = 67.66$ km s^{-1} Mpc^{-1}, $\Omega_\Lambda = 0.6889$ and $\Omega_m = 0.3111$ and I use AB magnitudes throughout.

2.2 Data

To build a representative sample of lensing systems, I selected a suitable galaxy catalogue. I used a single catalogue for both the lens and source galaxies, rather than merging two distinct catalogues. This enabled any galaxy to act as a lens of any other, freeing the predictions from biases to lens and source selection, and ensuring the corresponding galaxy properties (masses, magnitudes, redshifts etc.) were all self-consistent. The catalogue was required to reach sufficiently high stellar masses (the lensing cross-section increases with lens mass), to span a wide redshift range (lenses

selected in the visible are typically found at $z \sim 0-1$, Auger et al., 2009; Shu et al., 2022, and sources at even higher redshifts) and to reach suitably faint magnitudes. The magnitude limit of the catalogue would ideally surpass the depth of the target survey since lensing can magnify an otherwise undetectable source galaxy above the magnitude limit. Therefore, I selected the JADES Extragalactic Ultra-deep Artificial Realization catalogue (JAGUAR, Williams et al., 2018) as the basis catalogue for my simulations. JAGUAR is based on empirical rather than simulated data and contained full Spectral Energy Distribution and morphological information (Sérsic indices, stellar masses, effective radii etc.) for each galaxy. The galaxy number counts are fitted to empirical stellar mass and luminosity functions for both star-forming and quiescent galaxies (Tomczak et al., 2014; Bouwens et al., 2015; Oesch et al., 2018) and so their proportions in my simulations should be representative of observations. The catalogue reaches depths of $\sim$ 30mag, spans a redshift range of $0.2 < z < 15$ and a stellar mass range of $10^{5.9} < \log(M_*/M_\odot) < 10^{11.7}$. Consequently, it is suitable for both source and lens populations. It comprises of 10 realisations of 11x11 arcmin^2 area. I verified that the observed number-counts in the VISTA Deep Extragalactic Observations (VIDEO, Jarvis et al., 2013) and UltraVISTA (McCracken et al., 2012) surveys were well matched to the JAGUAR catalogue (Figure 2.1). The total area of the JAGUAR catalogue (0.34 deg^2) is relatively small - this leads to a greater uncertainty in the number of high-mass galaxies (and thus uncertainty in the number of detectable high-mass lenses, which would be prior dependent). In Figure 2.1, the galaxy number counts predicted by JAGUAR slightly underestimate the number of the brightest galaxies seen in the VIDEO survey. This is an important consideration for the results presented in this study. I discuss application to wider area surveys in Section 2.4.4; however, the primary focus in this work was on smaller area ($\lesssim$ 10 deg^2) medium-deep ($m > 25$) NIR surveys with *JWST* and VISTA, for which this catalogue was most suited.

One disadvantage of adopting a catalogue founded on empirically derived relations is that it lacks information on the dark matter profiles which would be available from a cosmological simulation. Therefore, I augmented the physical

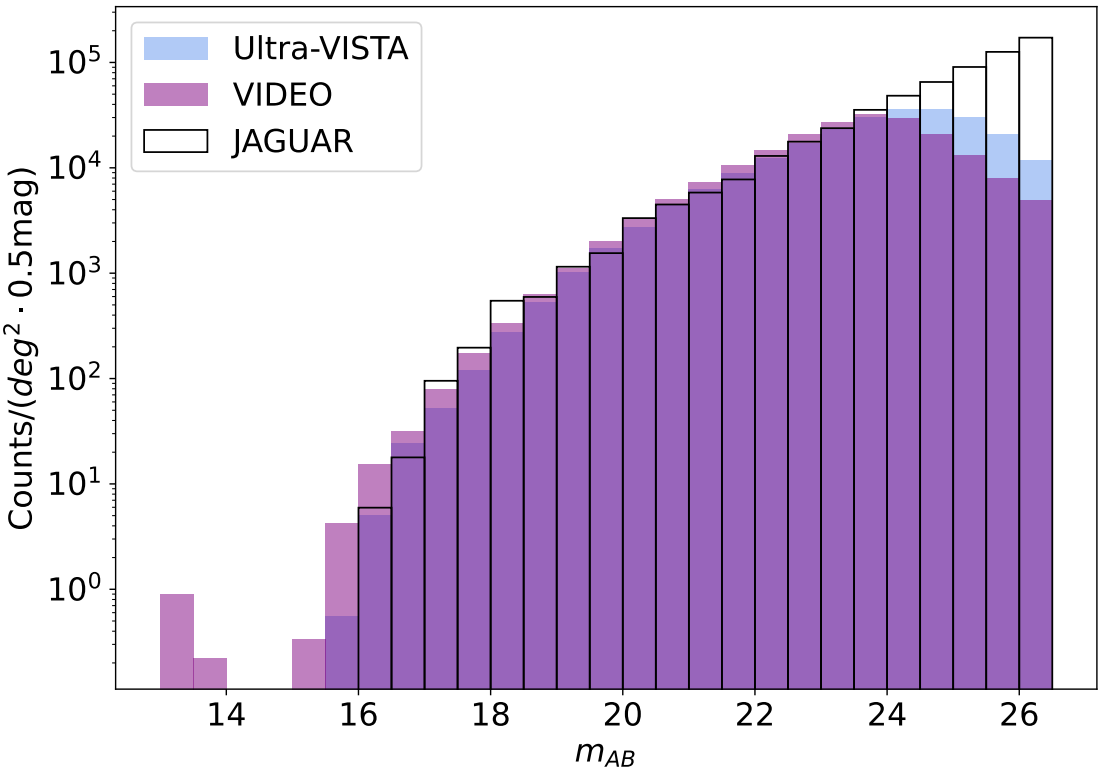


Figure 2.1: Plot of the differential H-band number counts in the NIR surveys VIDEO (XMM and CDFS fields) and UltraVISTA compared to the JAGUAR catalogue. The JAGUAR catalogue aligns well with the NIR surveys up to their respective completeness limits but misses some of the most massive galaxies.

parameters from JAGUAR with dark matter properties (profiles and masses) from the Deep Realistic Extragalactic Model (DREaM) galaxy catalogue (Drakos et al., 2022), used for simulating the Roman Ultra-Deep Field (see Section 2.2.1). This is a 1 deg^2 catalogue with a magnitude limit similar to JAGUAR reaching $z \sim 12$ with the dark matter component simulated from GADGET-2 (Springel, 2005).

2.2.1 Adaptions to the Galaxy Catalogue

Modifications to JAGUAR Morphological Parameters

Although the JAGUAR catalogue was well suited to requirements, there were two areas in which I adapted the supplied values to better reflect realistic galaxy properties, namely the effective radii ($R_{\rm eff}$) and Sérsic indices (n_S).

The values of $R_{\rm eff}$ given by the JAGUAR catalogue depend on the galaxy type and redshift. The radii of $z < 4$ star-forming galaxies (SFGs) and quiescent galaxies

(QGs) at all redshifts are based on Wel et al. (2014), using 3D-HST+CANDELS empirical data based at a rest-frame wavelength of 500nm. The values for star-forming galaxies at $z > 4$ were assigned via a M_{UV}-size relation from Shibuya et al. (2015) which predominantly used *J-H HST* bands at such redshifts. The JAGUAR values for the Sérsic indices use results derived from Wel et al. (2012), who in turn used *HST* F160W Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS) imaging. The wavelength dependence on the galaxy profile of changing the effective radius and Sérsic index simultaneously is small (Kelvin et al., 2012), so I simply aimed to ensure the n_s and $R_{\rm eff}$ values were evaluated at the same observed wavelength. This was particularly important for lower redshifts, where the lenses are likely located, since differences in these parameters would change the lensing cross-section. Therefore, I shifted the $R_{\rm eff}$ values to the observed H band, using the $R_{\rm eff}$-λ relation from Vulcani et al. (2014), so they were consistent with the JAGUAR Sérsic indices.

Within a given redshift range, the Sérsic indices in Williams et al. (2018) were randomly assigned, regardless of galaxy type. Since typical lens galaxies are quiescent, I ensured their mass and light profiles reflected their galaxy type. Therefore, I reallocated the Sérsic indices within the JAGUAR catalogue, binned by both redshift and galaxy type, where the n_S distribution for quiescent galaxies was chosen to follow the distribution observed by Gu et al. (2020) from 3D-HST+CANDELS data.

Inclusion of Dark Matter

The halo masses from the DREaM catalogue galaxies were assigned to JAGUAR galaxies via Subhalo Abundance Matching (SHAM). I ordered JAGUAR galaxies in order of stellar mass and DREaM halos in order of the maximum circular velocity over the halo's accretion history, $V_{\rm peak}$, as this has been shown to be a good predictor of stellar mass (Chaves-Montero et al., 2016; Contreras et al., 2021), then paired them accordingly. A dark-matter+baryon simulation would provide a more realistic

lens population, however the current method allowed us to draw sources and lenses from the same population over a large redshift range.

2.3 Method

My method for determining strong lens occurrence rates can be summarised as follows. I first selected the number of close galaxy pairs ($\Delta\theta < 5$") in a patch of sky (assuming random angular distribution). I then determined the number of these for which the foreground galaxy strongly lensed the background galaxy based on magnification and multiple-imaging requirements. Finally, I imposed constraints on the detectability of such systems based on the properties of the survey of interest.

I did not explicitly calculate the lensing cross-section for each galaxy, rather, I used a numerical approach, allowing any galaxy to act as a lens (determined by realistic mass information provided in the base catalogue). The lensing cross-section was indirectly accounted for by constraints on the magnification and relative lens-source positions as detailed in Section 2.3.3, that determined which strong lenses would be identified in a typical search. I describe the methodology in detail in the following sections.

2.3.1 Frequency of Galaxy-Galaxy Conjunctions

To calculate the number of possible lens systems (neglecting for the moment the Einstein radius of the lens and any detectability and significant magnification criteria, which are applied later) I assumed the galaxies were distributed randomly in the sky, i.e., I neglected the effect of galaxy clustering. For strong lensing to occur the lens and source galaxies must be closely aligned on the sky. When considering potential strong lens candidates, I chose an upper limit of $5''$ of separation to select galaxy pairs since this would encompass the vast majority of galaxy-galaxy strong lenses (see fig. 1 in Collett, 2015) while remaining computationally tractable.

The derivation for the number of galaxy-galaxy conjunctions is as follows. Consider a set of N objects, distributed randomly within a square of length L. I asserted that, for a source to be lensed, it must be located within an angular

distance R of the lens. The probability that an object A is within a radius R of a distinct object, B, placed randomly within the box, is $\frac{\pi R^2}{L^2}$. The mean number of galaxies within radius R of galaxy A is then $(N-1)\,\frac{\pi R^2}{L^2}$. Now excluding galaxy A to prevent double counting, the mean number of galaxies close to galaxy B is $(N-2)\,\frac{\pi R^2}{L^2}$. By extension, the number of galaxies in proximity to each other is therefore $\frac{\pi R^2}{L^2}\sum_{i=1}^{N-1} i$. Based on the JAGUAR catalogue scaled to 12 deg^2, I used $N = 1.1 \times 10^8$ objects, with $L = \sqrt{12}$ deg and chose $R = 5''$ as described above. One of the original motivations of this work was investigating the strong lens population in medium-to-deep NIR surveys with VISTA. Hence, I matched the simulated area (12 deg^2) to that of the VIDEO survey (Jarvis et al., 2013). I emphasise here that the R parameter is not the Einstein radius of the lens and has no impact on the lensing potential - it merely limits the number of galaxy pairs to be modelled to reduce computational time; the Einstein radii of the systems are calculated later based on the mass and redshift properties of the lens and source.

The above calculation gives 3.1×10^9 lens-source pairs for which to determine lensing properties. This simulation was easily scalable to smaller area surveys such as those of *JWST*. To ensure that a given galaxy does not produce a detectable lens system an unrealistic number of times for the total simulation size (i.e., that a given lens galaxy in the catalogue isn't double-counted, so the predictions are consistent with the empirical galaxy number counts verified previously), I subdivided the simulation into JAGUAR-sized boxes (0.34 deg^2) and only counted one lensed system from a given lens galaxy from each box; however, this only had a minor effect. To aid computation time, I limited the lens masses to $> 10^9 M_\odot$ although this made a negligible difference to the resultant lensing frequencies, since typical lenses have much higher mass than this limit.

For each galaxy-galaxy conjunction, I drew two redshifts randomly from the redshift population of the JAGUAR catalogue, and assigned the lower redshift galaxy (along with its corresponding properties e.g., stellar mass) as the lens, and the higher redshift galaxy as the source. The resultant redshift distribution is shown in Fig. 2.2; this was essentially the prior on the lens and source redshifts.

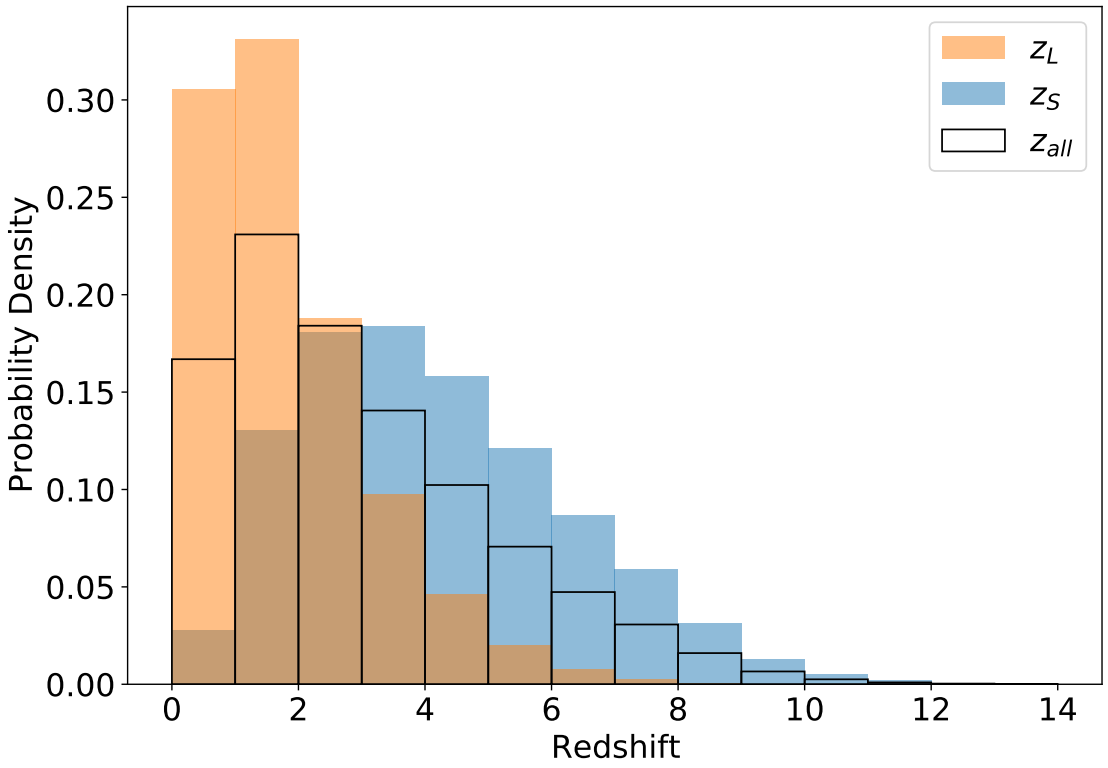


Figure 2.2: Redshift distribution for the original JAGUAR mock catalogue (unfilled), the sample of lenses (orange) and sources (blue). This distribution is before the application of any constraints for detectability or significant lensing and is effectively the prior on the lens and source redshifts.

2.3.2 Calculating Lensing Properties

To calculate lensing properties (Einstein radius, deflection angle etc.) and to generate simulated lensed images, I modelled each lens galaxy as a circular Sérsic stellar component with an elliptical NFW (Navarro et al., 1996) halo profile randomly orientated along the line of sight. Given that most of the simulated lens systems had low magnifications (see Figure 2.4), the overall effect of ellipticity was of the order of low tens of percent (Lapi et al., 2012), and thus would not affect the qualitative results in comparison to a spherical halo. Lapi et al. (2012) investigated the effect of ellipticity on the cross-section of singular isothermal sphere and ellipsoid models. They found that for ellipticities of $e \sim 0.4$ (typical for my simulations), in the case of magnifications $\mu < 8$ (the median magnification the systems shown in Figure 2.4 is $\mu = 5$), the cross section of an elliptical lens was reduced by around 30%

compared to that of a spherical lens. However, such ellipticity does add additional realism which can be required for some lens systems (Suyu et al., 2012). I modelled the lens and source light with a Sérsic profile; in the case of the lensing galaxy I used the same Sérsic parameters as the mass profile, i.e., assuming a constant stellar mass-to-light ratio. An analytic expression for the lensing properties of a 2D elliptical NFW projected profile is not known (e.g., see Gomer et al., 2023) so I used a cored steep ellipsoid as a basis function as described in Oguri (2021) to generate such a profile. Using Sérsic and NFW combined profiles allowed a wide variation in lens morphologies and ensured that the simulated galaxies were realistic. I verified (below) that the properties of the detectable lenses (e.g., Einstein radii and density slopes) agreed with well studied lenses. I used the morphological information (i.e., n_S, $R_{\rm eff}$, M_* etc.) available in the JAGUAR/DREaM catalogues, so no further assumptions were required to generate the profiles. Requiring all the simulated galaxies to have an isothermal profile, as is often done, would have been unrealistic in the cases where the foreground galaxies were not massive ellipticals and would neglect lensing from e.g., the bulges of late-type spirals. Furthermore, as demonstrated by Lapi et al. (2012) and Gavazzi et al. (2007), the difference in lensing cross-section between a de Vaucouleurs+NFW profile compared to an isothermal one is very small and can still provide a good fit to early-type lenses, so one would not expect this to have a material impact on the results.

2.3.3 Detectability Constraints

In order to ascertain which lens systems would be detectable in a given survey, I followed the procedure detailed below. Mock images were generated of size 1001×1001 pixels covering a $20'' \times 20''$ field of view. These were then resampled to the pixel scale of each of the surveys of interest and convolved with the appropriate Point Spread Function (PSF) for the survey and bandpass. This resulted in noiseless simulated lensed images for each potential lens. I describe accounting for the effect of both the photon noise from the lens galaxy and the survey noise below. I show examples of detectable lensing systems from the simulations in Fig.

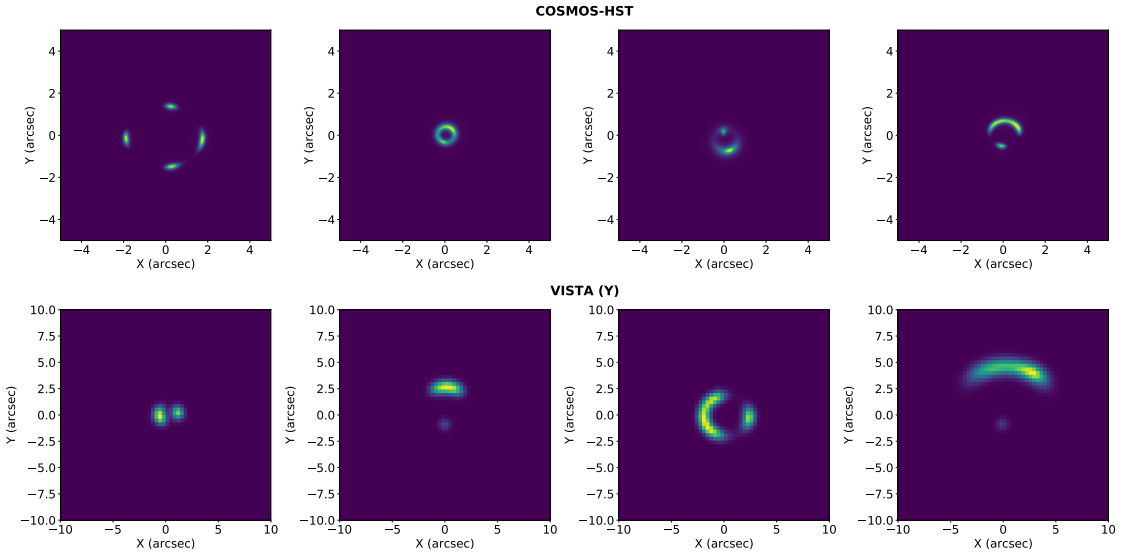


Figure 2.3: Examples of detectable systems in COSMOS-HST and VISTA (Y). In both cases the images are resampled to the survey pixel size and convolved with the relevant PSF.

2.3. To determine whether a given galaxy-galaxy pair would be detectable and classified as a strong lens candidate system, I applied the following constraints from Collett (2015) to the lensing parameters and generated image: 1) $\mu R_{\rm source} > s$, 2) $\theta_E^2 > R_{\rm source}^2 + (s/2)^2$, 3) $\mu > 3$. Here, μ is the total magnification, $R_{\rm source}$ is the unlensed source half-light radius, s is the PSF (or seeing, for a ground-based telescope) and θ_E is the Einstein radius. These constraints correspond to the detection of tangential shear, the ability to resolve the lensed image and requiring significant lensing, respectively. I further required the source position to be located within the radial caustic of an NFW+Sérsic profile, so that multiple images may be produced, 4) $\theta_{\rm sep} < \theta_{\rm caustic}$ where $\theta_{\rm sep}$ is the unlensed galaxy separation.
I added cuts for resolution, 5) $\theta_E > 1.5 r_{\rm pixelscale}$ (i.e., 3 pixels in diameter), and depth 6) $m_{\rm lens} < m_{\rm cut}$, (see Table 2.2), with values tuned to the target survey. Finally, I added a cut for the detectability of the lensed image, 7) signal-to-noise$>$ 20. The signal-to-noise calculation is described in the following section.

Signal to Noise Calculation

I calculated the number of photons incident for a given survey, exposure time and filter using the zeropoints in Table 1. A zeropoint for the *Roman* telescope was

not available so this was calculated in the same manner as in Euclid Collaboration: Schirmer et al. (2022), using the filter transmission curve provided in the *Roman* documentation[1]. The total number of photons collected by the telescope due to an object of AB magnitude m is given by:

$$n = 10^{\left(\frac{ZP-m}{2.5}\right)} \cdot t_{\rm exp} \tag{2.1}$$

where $t_{\rm exp}$ is the exposure time of the survey.

Adding the photon noise from the lens and source galaxies to the background noise of the survey in quadrature, the total noise in a pixel is:

$$N' = \sqrt{\left(\sqrt{n_{\rm lens}}\right)^2 + \left(\sqrt{n_{\rm source}}\right)^2 + \sigma_b^2} \tag{2.2}$$

where n denotes the number of photons per pixel from a given source and σ_b is the background noise from the survey.

I calculated the background noise of the survey from available magnitude limits in the literature. Where only point source, rather than blank-field depths, were available (namely for *Euclid* Near Infrared Spectrometer and Photometer (NISP), *Roman* and *JWST* data), I applied aperture correction to a radius equal to the 80% encircled energy from the PSF and from this, calculated the noise per pixel. This process produced a noise map, tending to the constant survey limit towards the image edges, but accounting for the increased noise due to the lensing galaxy in the centre. I did not include noise associated with imperfect lens subtraction. Using the generated lensed system images, I then calculated the signal-to-noise ratio for each pixel in the image. For all regions of connected (i.e., adjacent) pixels in the array with SNR ≥ 1, I calculated the total value according to:

$$\mathrm{SNR} = \frac{\sum_{\rm i} \mathrm{S_i}}{\sqrt{\sum_{\rm i} \mathrm{N_i'^2}}} \tag{2.3}$$

[1] https://roman.gsfc.nasa.gov/science/WFI_technical.html
[2] https://acszeropoints.stsci.edu/
[3] Sutherland et al. (2015)
[4] https://jwst-docs.stsci.edu/jwst-near-infrared-camera/nircam-performance/nircam-absolute-flux-calibration-and-zeropoints
[5] https://jwst-docs.stsci.edu/jwst-near-infrared-camera/nircam-instrumentation/nircam-detector-overview/nircam-detector-performance
[6] Collett (2015)
[7] Euclid Collaboration: Schirmer et al. (2022)

Tel.	Filter	ZP ($1e^{-}s^{-1}$)	Ref
HST	F814W	25.95	[2]
VISTA	Y	25.68	[3]
	J	26.29	
	H	26.83	
	Ks	26.46	
JWST	F070W	27.23	[4,5]
	F090W	27.54	
	F115W	27.59	
	F150W	27.89	
	F200W	28.08	
	F277W	27.98	
	F356W	28.14	
	F444W	28.16	
Euclid	$I_{\rm E}$	25.50	[6]
	$Y_{\rm E}$	25.04	[7]
	$J_{\rm E}$	25.26	
	$H_{\rm E}$	25.21	
Roman	J129	26.40	

Table 1: Survey zeropoints used for the lens frequency estimates. A zeropoint was not available for the *Roman* telescope so was calculated here in the same manner as in Euclid Collaboration: Schirmer et al. (2022) using the filter transmission curves provided in the *Roman* documentation.

where I summed over all the connected pixels in a given region. If any region had SNR>20, the lens system was deemed detectable. In this manner, I accounted for the difficulty in identifying a source close to very bright lenses, as well as the variation in depth with each survey. By focusing only on connected pixels, I also accounted for the fact that collections of adjacent brighter pixels would be easier to identify than if those pixels were dispersed. Many lens searches (Jacobs et al., 2017; Faure et al., 2008; More et al., 2016) are based on images which are not lens subtracted. This makes lensed source identification more difficult due to blending with the lens galaxy. I estimated the results for searches without lens subtraction by masking out pixels from the source image which are fainter than the lens at that position, then calculating the SNR in the same manner as described above. I also used the SNR calculation to determine the number of systems in which multiple

lensed images were detectable, described further in Section 2.4.3.

2.4 Results

2.4.1 Verifying the Simulations

In this section, I detail verification of my simulations. I measured the density slopes of simulated lenses detectable in *HST* to test whether the simulations tallied with the bulge-halo conspiracy. I further compared the Einstein radii of SLACS galaxies with those of similar galaxies in the simulated *HST* catalogue to establish if the choice of lensing profile had undue influence on the resultant lens properties.

Measurement of the Mass-Density Slope

In line with the existence of the bulge-halo conspiracy, I verified that the resultant surface density gradients, $\gamma_{\rm surf}$ from the composite NFW+Sérsic profiles were on average isothermal. The measured values of the surface density slope depend on the method of measurement. I first defined the surface density slope $\gamma_{\rm surf} = -d\log_{10} M(R)/d\log_{10} R$ where $M(R)$ is the mass enclosed by a cylindrical radius R. I then used four methods as follows:

- The mean value of $\gamma_{\rm surf}$ for R values between $10^{-4}\theta_E$ and θ_E with $\Delta R = 10^{-4}\theta_E$,
- The best-fit of a straight line in log-log space for R vs M(R), with R values between $10^{-4}\theta_E$ and θ_E and $\Delta R = 10^{-4}\theta_E$,
- The best-fit of a straight line in log-log space for R vs M(R), with R values between $10^{-4}\theta_E$ and θ_E with $R_{i+1} = (1 + 10^{-4})R_i$ (i.e., giving greater weight to lower values),
- The weighted mean value of $\gamma_{\rm surf}$, weighted by the enclosed mass in an cylindrical annulus at each radius, for R values between $10^{-4}\theta_E$ and θ_E with $\Delta R = 10^{-4}\theta_E$. This is the 2-dimensional equivalent to the mass-averaged slope given by Dutton and Treu (2014).

The slopes of the simulated lenses were typically marginally shallower than isothermal; this would have the effect of reducing the lensing cross-section (e.g., Mandelbaum et al., 2009), making my estimates more conservative. The median values for each method (top-bottom) were: 0.822 ± 0.179, 0.926 ± 0.206, 1.077 ± 0.349, and 0.859 ± 0.190. For comparison, an isothermal profile has surface density gradient of $\gamma_{\rm surf} = 1$ with which the simulated slopes are consistent within the scatter. Sonnenfeld et al. (2013) found that the density slope varies with stellar mass density and redshift, so one would not expect all the galaxies to be exactly isothermal. Koopmans et al. (2009) identified a scatter of $\sigma_\gamma \leq 0.2$ for SLACS lenses. Sonnenfeld et al. (2013) calculated the power law slope for the Strong Lensing Legacy Survey (SL2S) lens sample; the raw scatter in the slope was 0.2, reduced to an intrinsic scatter of $\sigma_\gamma = 0.12$, after accounting for the evolution in γ with mass and redshift. I did not account for dependence on such factors, so the measurements of the scatter in the mass-density slope are also consistent with those observed.

Confirming the Presence of Realistic Lensing Galaxies

To verify the adapted JAGUAR catalogue could reproduce observed lenses I searched for analogues to the lenses found in the SLACS Survey (Bolton et al., 2008) and compared the Einstein radii measurements to the SLACS lenses as analysed by Grillo et al. (2009). I identified comparable lensing galaxies to SLACS in the simulated galaxy catalogue (matching the mass and redshift of the SLACS lenses to galaxies in the catalogue to within a factor of $0.9 - 1.1$ in mass and ± 0.2 in lens redshift) and singled out the galaxy with the closest stellar-mass fraction to those in Grillo et al. (2009). 55 out of 57 galaxies in Grillo et al. (2009) could be matched to similar galaxies in the simulation in this way. i.e., most of these lensing galaxies are present within the simulations. The remaining two did not pass these tests due to a lack of sufficiently massive galaxies in the catalogue. For those which passed, I then calculated the corresponding Einstein radii finding a good agreement with the SLACS values from Grillo et al. (2009) with a mean Einstein radius ratio between matched lenses of 0.96 ± 0.07.

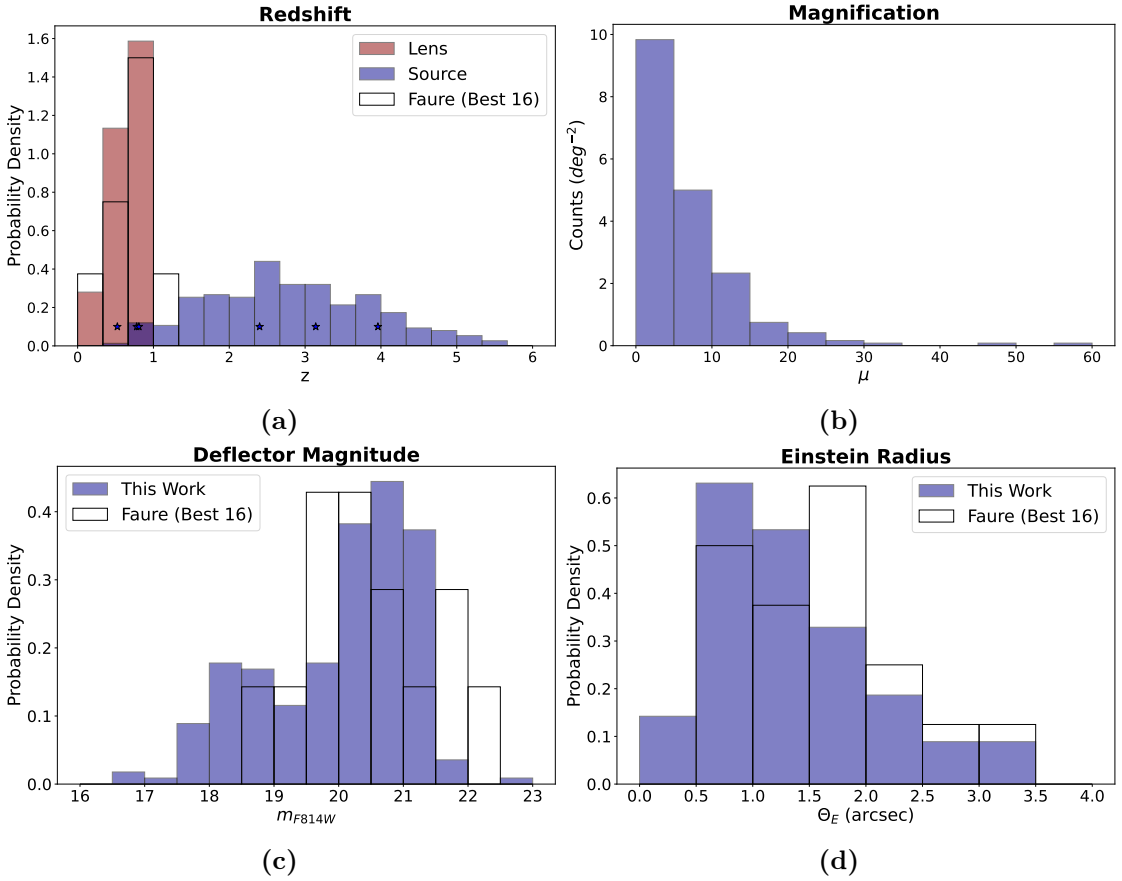


Figure 2.4: Properties of the detectable lens population at 20σ a) The simulated and observed lens/source redshift population of lenses detectable in COSMOS. The majority of the observed sources have not been spectroscopically confirmed, however I plot redshifts of the sources where known (starred). b) The magnification distribution for the simulated COSMOS lens systems. c) The simulated and observed F814W magnitude of the lenses in COSMOS d) The Einstein radii of simulated and observed lenses in the COSMOS survey.

2.4.2 The Observable Lens Population

Existing and forthcoming surveys investigated

I investigated the strong lens populations in a range of ground-based and space-based surveys spanning a broad range of depths and wavelengths. I first compared my results to 3 existing *HST* searches of the COSMOS field (Faure et al., 2008; Jackson, 2008) and archival images (Pawase et al., 2014). I then focused on surveys for which lens searches had/have not yet taken place. I compared the strong lensing population in COSMOS-HST to that which might be found using the ground-based UltraVISTA survey (McCracken et al., 2012), a 1.8 deg^2 survey that uses the *Y, J, H* and *Ks* bands on the VISTA telescope within the COSMOS field. I also

provide estimates here for the same bands in the VIDEO survey (Jarvis et al., 2013), a 12 deg^2 NIR survey of the ELAIS-S1, XMM-Newton and extended Chandra Deep Field-South (CDFS) fields. The UltraVISTA survey is split into 'deep' and 'ultradeep' stripes, the former being of comparable depth to the VIDEO survey so I just considered the ultradeep regions in this work, which cover half the area. I then investigated strong-lens prospects for *JWST*. The largest (and relatively shallow by *JWST* standards) survey COSMOS-Web (Casey et al., 2023) has point-source depths $m_{\rm F115W} \sim 27$mag, covering an area of ~ 1900 arcmin^2. The JADES medium and deep program (Rieke, 2019) by comparison are significantly deeper ($m_{\rm F115W} = 29.6 - 30.6$mag for the medium and deep surveys respectively) but cover a much smaller area (190 and 40 arcmin^2 respectively). These surveys were chosen since they span a representative range of *JWST* surveys (see Casey et al., 2023 for an overview). Finally, I comment on expectations for the forthcoming EWS (Euclid Collaboration: Scaramella et al., 2022) and Roman High Latitude Wide-Area surveys[8] which are discussed further in Section 2.4.4. I do not consider LSST here, since this is a wide-field visible survey so it was less suited to my simulations. I discuss the trade-off between survey depth and area in Section 2.5.1.

Comparison with Existing Lens Searches

The estimated values for the number of strong lenses detectable in the above surveys are shown in Table 2.2a (verification to previous lens searches), 2.2b (estimates for smaller survey areas) and 2.2c (extrapolation to wider area surveys).

[8] https://roman.gsfc.nasa.gov/high_latitude_wide_area_survey.html

Tel.	Survey	Filter	PSF ($''$)	Area (deg^2)	$m_{\rm cut}$ (lens)	m_{lim}	$N_{\rm deg}$	N_{tot}	Ref.
HST	COSMOS (F)	F814W	0.12	1.6	25.0[9]	30.42	21	34 (31)	[10]
	COSMOS (J)	F814W	0.12	1.6	25.0[11]	30.42	49	80 (75)	[12]
	COSMOS (All)	F814W	0.12	1.6	26.7	30.42	54	88 (82)	[13]
	Archive	F814W	0.12	6.0	25.4	28.70	17	100 (91)	[14]

(a) Lensing frequency estimates for detectable lenses (at SNR=20, suitable for a visual search) using the *HST* telescope for comparison to existing lens searches. m_{cut}, m_{lim}, N_{deg} and N_{tot} refer to the lens magnitude limit, the 5σ survey depth ($m_{AB}\,\mathrm{pix}^{-1}$), the number of detectable lenses per square degree, and the total number detectable in the survey respectively. The bracketed terms in the penultimate column refers the lensing frequency which might be expected in a lens search using images which have not been lens subtracted. (F) and (J) refers to the lens searches in the COSMOS field by Faure et al. (2008) and Jackson (2008) respectively, and the archive search refers to that of Pawase et al. (2014). COSMOS (All) refers to a theoretical lens search of the COSMOS field in which none of the constraints imposed by Faure et al. (2008) or Jackson (2008) were applied, i.e., an untargeted search.

VISTA	VIDEO	*Y*	0.8	12	25.2	26.97	7.0	84 (53)	[15]
		J	0.8	12	24.7	26.49	5.4	65 (40)	
		H	0.8	12	24.2	25.99	4.2	50 (33)	
		Ks	0.8	12	23.8	25.59	3.3	40 (29)	
	UltraVISTA (UD)	*Y*	0.77	0.9	25.8	27.89	15	13 (8.0)	[16]
		J	0.77	0.9	25.6	27.69	15	14 (7.7)	
		H	0.76	0.9	25.2	27.29	14	13 (7.8)	
		Ks	0.78	0.9	24.9	26.99	13	11 (7.1)	
JWST	COSMOS-Web	F115W	0.040	0.54	26.8	29.34	32	17 (16)	[17]
		F150W	0.050	0.54	27.1	29.44	47	25 (22)	
		F277W	0.092	0.54	27.5	29.59	110	59 (48)	
		F444W	0.145	0.54	27.4	29.89	110	62 (47)	
	JADES-Medium	F070W	0.029	0.053	28.8	31.53	200	10 (9.2)	[18]
		F090W	0.033	0.053	29.4	32.03	340	18 (15)	
		F115W	0.040	0.053	29.6	32.16	400	21 (16)	
		F150W	0.050	0.053	29.7	32.04	390	21 (15)	
		F200W	0.066	0.053	29.8	32.33	460	24 (16)	
		F277W	0.092	0.053	29.4	31.47	400	21 (13)	
		F356W	0.116	0.053	29.4	31.67	400	21 (12)	
		F444W	0.145	0.053	29.1	31.57	340	18 (10)	
	JADES-Deep	F090W	0.033	0.011	30.3	32.93	640	7.2 (5.3)	[19]
		F115W	0.040	0.011	30.6	33.16	780	8.6 (5.9)	
		F150W	0.050	0.011	30.7	33.04	740	8.2 (5.0)	
		F200W	0.066	0.011	30.7	33.23	790	8.7 (4.6)	
		F277W	0.092	0.011	30.3	32.37	660	7.3 (3.6)	
		F356W	0.116	0.011	30.2	32.47	620	6.9 (3.2)	
		F444W	0.145	0.011	29.9	32.37	520	5.7 (2.7)	

(b) As above, for forthcoming and existing NIR surveys.

Euclid	Wide	I_E	0.17	15,000	26.2	27.41	6.3	95,000 (86,000)	
		Y_E	0.22	15,000	24.3	25.49	1.8	28,000 (28,000)	[20]
		J_E	0.30	15,000	24.5	25.69	3.8	58,000 (51,000)	
		H_E	0.36	15,000	24.4	25.59	4.1	61,000 (55,000)	
Roman	High Latitude Wide-Area	J129	0.1	1,700	27.1	29.12	52	88,000 (72,000)	[21]

(c) As above, for wider area surveys included for comparison to previous predictions. These values represent extrapolations from our dataset as discussed in detail in Section 2.4.4.

Table 2.2

The values in Table 2.2a show simulated lensing frequencies for previously undertaken lens searches of *HST* imaging, along with those of an untargeted search of the COSMOS field (i.e., with no prior redshift or magnitude cuts applied). Unless explicitly stated, the results of my simulations for a COSMOS-HST search in this work refer to this theoretical search, i.e., without the constraints applied by previous studies. In this current section (2.4.2), I *do* apply such constraints to allow comparison with the results of previous studies.

Faure et al. (2008) detail a visual search of the 1.64 deg^2 *HST* COSMOS survey. They present 67 candidates, of which 20 display multiple images or large arcs. My prediction for the COSMOS survey (34 of which 31 are detectable

[9] To make the results comparable with Faure et al. (2008), I adopted constraints of $0.2 < z_L < 1.0$, $M_V < -20$ and $m_{\mathrm{F814W}} < 25$.
[10] Taniguchi et al. (2009)
[11] To make the results comparable with Jackson (2008), I adopted constraints of $\theta_E < 2.5$" and $m_{\mathrm{F814}} < 25$.
[12] Taniguchi et al. (2009)
[13] Taniguchi et al. (2009)
[14] Pawase et al. (2014)
[15] Varadaraj et al. (2023)
[16] Moneti et al. (2023) and McCracken et al. (2012)
[17] Casey et al. (2023)
[18] Rieke (2019)
[19] Rieke (2019)
[20] Euclid Collaboration: Scaramella et al. (2022)
[21] https://roman.gsfc.nasa.gov/science/ETC2/ExposureTimeCalc.html, using a PSF-fitted point source and zodiacal light 1.4x the minimum.

without subtraction) is within the range of highly graded lens candidates in Faure et al. (2008).

Jackson (2008) conducted a visual search of all galaxies in COSMOS using *HST* imaging with $i < 25$mag and found 2 certain, 1 probable and $\mathcal{O}(100)$ possible (but unlikely) lens systems beyond those of Faure et al. (2008). Their cutout size limited the detectable Einstein radii to $\theta_E < 2.5$" and they only identified 50% of the 20 best lenses (those with multiple images or large arcs) identified by Faure et al. (2008) through their manual search. Loosening the constraints to match those of Jackson (2008) produced 80 lenses from the simulations. Due to the completeness of the manual search compared to the more targeted inspection of Faure et al. (2008), it is perhaps unsurprising that the estimates here are higher than the number found by Jackson (2008), although they are in line with the original expectations from Jackson (2008) of ~ 100.

Pawase et al. (2014) conducted a lens search using all the available archival *HST/ACS* F814W-band imaging data. Together these covered a wider area than COSMOS (6.03deg^2) but had a range of depths; I used an average 5σ depth of 25.45. They identified 13 A-grade, 18 B-grade and 9 C-grade lens systems (totalling 40). Pawase et al. (2014) attributed the reduced number density of lenses in this field compared to COSMOS down to cosmic variance and the reduced number of inspectors compared to Faure et al. (2008) which could reduce the number identified compared to my estimates. The simulations for this data predict a higher value of 100 lenses (91 without lens subtraction) which may be attributed to the broad range of depths of the archival images, or indicate more lenses yet to be found. It should also be noted there is often significant disagreement between strong lens experts when grading/classifying lens candidates (Rojas et al., 2023) and $\gtrsim 6$ graders is required to compensate for this, while Pawase et al. (2014) was limited to 1 grader for the majority of the search area and the search by Faure et al. (2008) was limited to 4.

The masses, magnifications, magnitudes and redshifts of the simulated lens and source populations detectable in the COSMOS survey are shown in Fig. 2.4 and where possible compared to the observed population from Faure et al. (2008).

The initial COSMOS search did not use lens-subtracted images so in these plots I applied the constraints for a non-lens subtracted search as described in Section 2.3.3. This ensured the images deemed detectable were on the outskirts of the lens galaxy light distribution where they would be more easily identifiable. The Einstein radii agree reasonably with the Faure et al. (2008) population extending to the same maximum radii and a two-sample Kolmogorov–Smirnov test gives a p-value of 0.17 suggesting that the null hypothesis, that the populations are drawn from the same population, cannot be rejected. The simulated lens magnitudes are typically brighter than observed (median magnitude difference $\Delta m \sim 0.2$). This could be due to cosmic variance in a small field and small number statistics for the total number of lenses (I only used the best 16 systems identified by Faure et al. (2008) here to prevent contamination by false positives).

2.4.3 Lens and Source Population Properties

Figure 2.5 shows the distribution of redshifts, stellar masses and Einstein radii for a range of surveys studied in this work. Most of the detectable lenses are located at $z \sim 1$ (Fig. 2.5a) and have high stellar mass ($\sim 10^{11} M_{\odot}$, Fig. 2.5b), although the deeper space-based surveys (*JWST*/*HST*) can detect lower mass lenses. The majority of lenses ($\sim 70\%$) are quiescent, as flagged by the JAGUAR mock catalogue. The lenses flagged as star-forming are redder ($\Delta(g-r) \sim 1$mag) and much more massive (by a factor of $\sim 10^{4.5}$) than typical SFGs in the simulation at those redshifts.

Figure 2.6 shows the distributions of mass, size and luminosity for the detectable lensed sources for a range of *JWST* surveys as a function of redshift. The sources predominantly lie on the star-forming main sequence at redshifts $z \sim 2-4$ (Fig. 2.5a) and are intrinsically brighter than the general population (Fig. 2.6). The deeper programs target higher redshift sources with lower surface brightness at each redshift. The source galaxies become more typical of the general population with depth; this can be seen in the right-hand column of Fig. 2.6 where the modal bins of F115W magnitude with redshift are brighter by $\Delta m_{\mathrm{F115W}} \sim 5, 3$ and 1

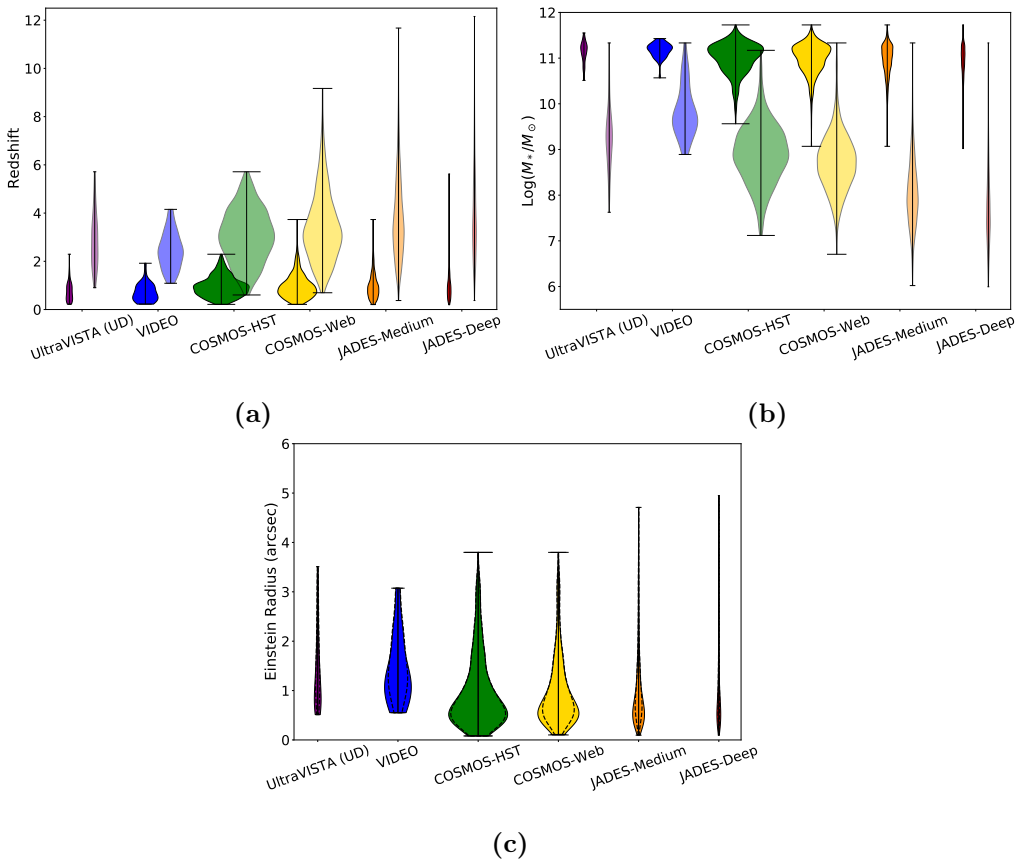


Figure 2.5: a) Redshift, b) stellar mass and c) Einstein radius distribution of detectable lenses across all filters in the given programs. The widths of the violin plots are scaled by the number of lenses expected in each program. The source galaxy values for redshift and mass are shown by the faded plots. The dashed lines in (c) refer to systems detectable without lens subtraction. Note, the surveys are of significantly different sizes so the extreme values will be very unlikely to be observed, especially in the smaller JADES-Medium/Deep surveys.

than the general JAGUAR population for COSMOS-Web, JWST Advanced Deep Extragalactic Survey (JADES)-Medium and JADES-Deep respectively.

Lens searches can be targeted (e.g., Faure et al., 2008), for example, selecting targets based on their magnitude, or untargeted (e.g., More et al., 2016). Such selections can significantly speed up a lens search but can come at the expense of reduced completeness. From the simulations one can measure the effect of such a selection. Fig. 2.7 shows the cumulative fraction of detectable lens systems with F115W magnitude across the *JWST* programs. A simple lens selection of $m_{\rm F115W} \lesssim 22$ would identify 80% of the available systems in *JWST* while reducing the number of galaxies to be searched by a factor > 4 compared to a more conservative

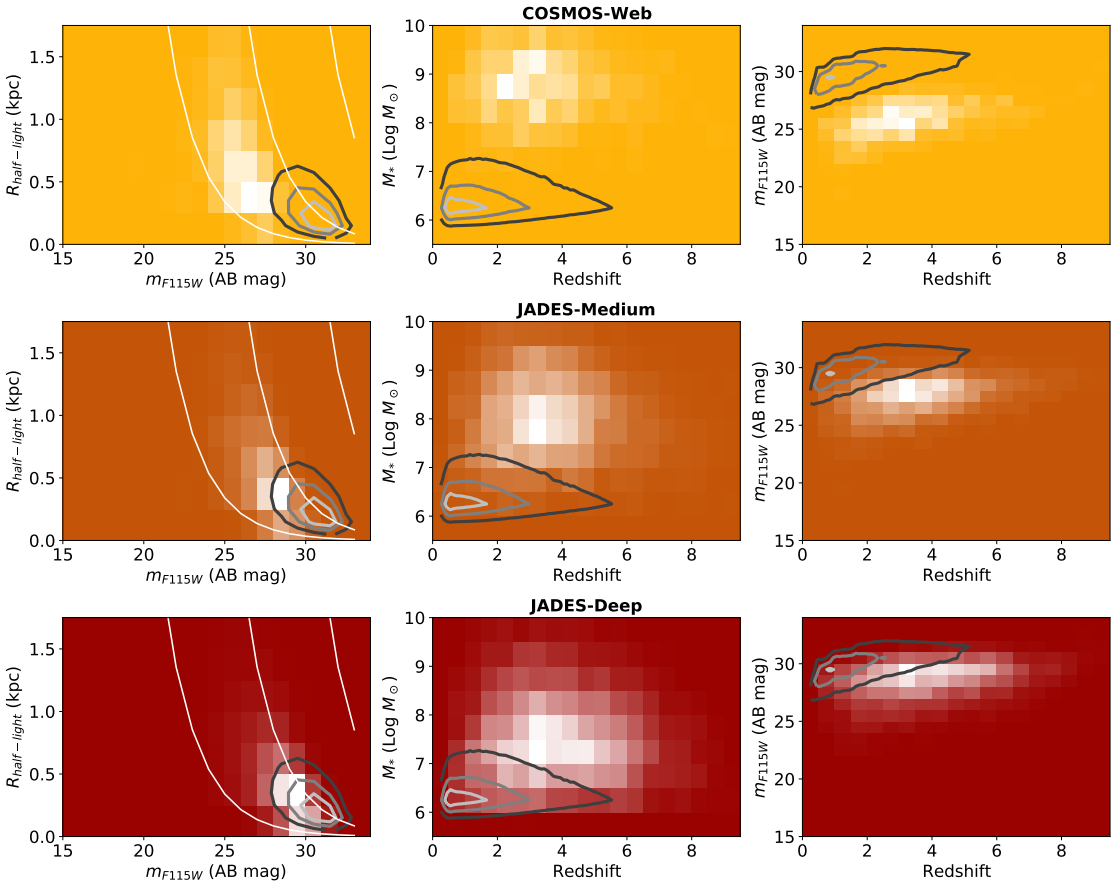


Figure 2.6: Properties of the source population of all detectable systems in the COSMOS-Web, JADES-Medium and JADES-Deep (top-bottom) systems, across all filters. The contours reflect the properties of the whole population in the JAGUAR catalogue, while the 2D histogram displays the detectable sources in each survey. The white lines in the first column show curves of constant surface brightness.

cut of $m_{\mathrm{F115W}} < 24$. Out of the simulated lens systems detectable in COSMOS, 96% had $m_{F814W} < 25$ (i.e., the cut used by Faure et al., 2008 and Jackson, 2008), indicating this cut would only have a minor affect on completeness.

Detection of Multiple Imaging

Using my simulations, I estimated the frequency of detecting multiple imaging in the mock galaxy-galaxy strong lenses. For clarity, here I refer to one of the multiple images from the same source as a sub-image, e.g., a single arc in an image containing two arcs from the same source. To split up the lensed image into sub-images I performed gradient ascent from all pixels with SNR ≥ 1 (per pixel) to find their associated maxima. I labelled all pixels which trace to the same maxima as part

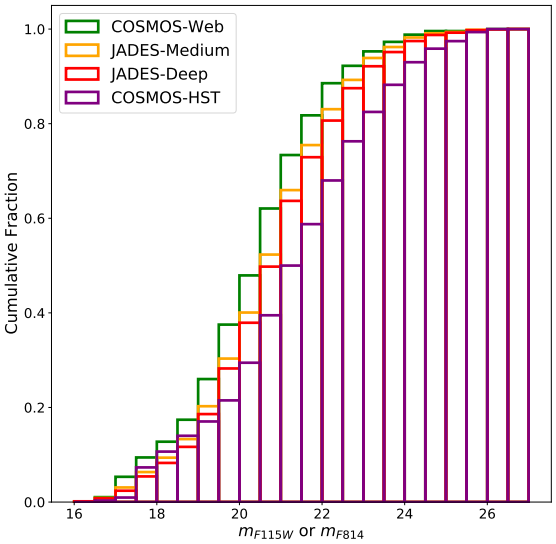


Figure 2.7: Cumulative histogram of the lens magnitudes of the detectable lenses in COSMOS-HST (m_{F814W} band) and *JWST* programs (m_{F115W} band). Roughly 80% of the lenses would be detected in a search of galaxies with $m_{\mathrm{F115W}} < 22$ for *JWST* and $m_{\mathrm{F814W}} < 23$ for *HST*.

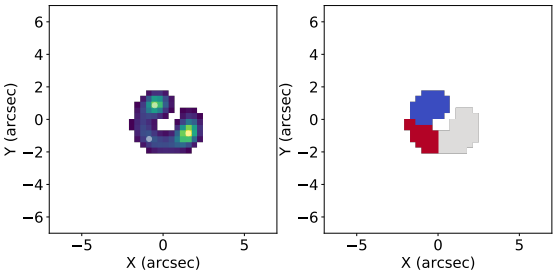


Figure 2.8: An example of a lensed image (L) along with its labelled sub-images (R). The pixels with SNR < 1 have been masked out. The white points in the left image indicate the local maxima.

of the same sub-image. For multiple imaging to be observed, the sub-images must each be detectable and resolvable from each other. To be detectable, I required at least one sub-image to have SNR > 20 and any counter-image to have SNR > 5. To be resolvable I required a sub-image to be positioned further than the Full Width Half Maximum (FWHM) of the PSF (or seeing) from another sub-image's peak. An example of such sub-images is shown in Fig. 2.8.

The proportion of simulated systems which have detectable multiple imaging is $\sim 55 - 80\%$ with deeper surveys having a larger proportion. Imposing a SNR > 20 constraint on all sub-images reduces this to $\sim 20 - 50\%$. These are systems where

the multiple images would be easily visible and so would likely be included in the 'A-grade' candidates in a visual search. The presence of multiple imaging adds confidence to lens identification and would provide tighter constraints for lens modelling. For a COSMOS-HST search akin to Faure et al. (2008), I found ~ 14 of the simulated detectable systems would have detectable multiple images (with a SNR > 5 limit for counter-images) and ~ 23 following lens subtraction; this is in agreement with the number of systems identified by Faure et al. (2008) (20, some of which were identified after lens modelling).

2.4.4 Extrapolating to Wide-Field Surveys

In this section, I discuss extrapolations to forthcoming wide-field surveys based on my simulations with the caveat that, due to the small area of JAGUAR, the most massive galaxies and therefore most likely lenses, are absent from the simulations. I accounted for this incompleteness by determining the proportion of massive lens galaxies (i.e., the fraction with $M_* > 10^{11.5} M_\odot$) in the published strong lens samples of SuGOHI (Sonnenfeld et al., 2019) and KiDS (Petrillo et al., 2017) with $z > 0.2$ to match the lower limit of the JAGUAR catalogue. These suggest that, from the mass distributions alone, the wide-area results would underestimate the true occurrence rate by a factor of 1.3 (KiDS) to 3 (SuGOHI). This should be considered as a lower limit since one notes that both KiDS and SuGOHI will have their own incompletenesses that are not counted for in the factors above, and if $z < 0.2$ lenses are included, one might expect another factor of 2 based on the fraction of $z < 0.2$, $M_* > 10^{11.5} M_\odot$ lens galaxies in SLACS (Auger et al., 2009). In particular, since the lenses used in Sonnenfeld et al. (2019) and Petrillo et al. (2017) were identified in part using neural networks or expert visual inspection with different selection effects (e.g., colour or spectroscopic requirements), they do not necessarily represent a complete sample of lenses by themselves. Overall, while these estimates are approximate, they remain useful lower estimates of the number of strong lenses that can be expected in forthcoming NIR surveys.

The EWS will cover an area of 15 000 deg^2. Using `LensPop`, Collett (2015) predicted 170 000 lenses could be detectable in the *Euclid* $I_{\rm E}$ band. The model presented here suggests 95 000 strong lenses in the $I_{\rm E}$ channel, i.e., approximately half of Collett's number. This apparent discrepancy arises from differences in the methods adopted as well as differences in the properties of the basis catalogues. For example, the JAGUAR catalogue misses massive galaxies (a comparison of velocity dispersion distributions show that ∼30% of `LensPop` galaxies with the highest velocity dispersions are absent from the JAGUAR catalogues), and JAGUAR has a higher completeness at faint magnitudes than the LSST catalogue; at the faint-end, the galaxies in JAGUAR outnumber those in the LSST catalogue used in `LensPop` by a factor of ∼3. These differences combined with differences in background noise level/image threshold in this work with respect to those adopted by Collett (2015, approx ∼2.5 lower than the limit I adopted to define detectability) are sufficient to explain the discrepancy. The estimates determined with this method are conservative due to the high threshold assumed and lack of the most massive galaxies in JAGUAR. `LensPop` provides a mid-range estimate owing to a lower threshold but also potentially underestimates faint sources (as the JADES Rieke et al., 2023 results would suggest). Therefore, a larger number of strong lenses than predicted by `LensPop` may be found in the *Euclid* $I_{\rm E}$ band, but they are likely to be at the faint end in source magnitude and may be harder to be securely identified as strong lens candidates.

Considering the *Euclid* NISP bands together, I anticipate ∼ 70 000 strong lenses will be detectable (noting again my estimates are conservative). Interestingly 40% of these are not detectable in the $I_{\rm E}$ band and I discuss this further in the Section 2.5.2. Considering the forthcoming wide-area *Roman* surveys, previous estimates by Weiner et al., 2020 (based on `LensPop`) suggested that ∼ 17 000 strong lenses would be detectable in the J129 band in 2000 deg^2. In contrast, my estimate (∼ 90 000) in 1700 deg^2 is significantly higher. This can be attributed to the difference in the zeropoints used (26.4 in this study vs 23.9 in Weiner et al., 2020). My zeropoint is consistent with the published transmission curves for *Roman* (Section 2.3.3).

Overall, for the survey depths provided in Table 2.2c, my simulations suggest that we can expect to detect $\sim 100\,000$ lenses in each of *Euclid* and *Roman* wide area surveys. I note that these estimates are significant extrapolations from the JAGUAR area, but are broadly consistent with previous estimates also given the differences in detectability criteria. Given the large numbers of lenses expected from such surveys, it will be unfeasible to obtain spectra for all of them. Measuring spectra for lens systems must be on a source-by-source basis (unlike systematic galaxy surveys such as DESI through a multi-object spectrograph), and some source galaxies will be outside the range of the [OII] doublet from the 4-metre Multi-Object Spectroscopic Telescope (4MOST) restricting source redshifts to $z_{\rm S} < 1.5$. However, the remaining systems will still be beneficial for population studies such as discussed in Chapter 5.

2.5 Discussion

In this section I focus on the detectable lens systems in forthcoming *JWST* surveys, before generalising to the other surveys investigated in this study. The properties of these systems and those of their respective samples as a whole are discussed in detail, as well as comparison of these estimates to recent lens discoveries.

2.5.1 Number Density of Detectable Lens Systems

Prospects for *JWST* Surveys

The JAGUAR-based simulated catalogue is best suited for small-area, deep surveys such as those of *JWST*. I generated lensing predictions for COSMOS-Web, JADES-Medium and JADES-Deep that span a wide range of depths and areas and thus provide indicative results for typical *JWST* surveys, present or future. Of the three surveys considered here, my simulations suggest the largest area *JWST* survey, COSMOS-Web, (Casey et al., 2023) will contain the most strong lens systems (~ 65 systems across all filters). The estimates for this survey (shown in Table 2.2b) are broadly consistent with the preliminary estimates from Casey et al. (2023) which predicted $\mathcal{O}(100)$ lenses based on a purely statistical argument without strong detectability cuts. I discuss the results of a recent lens search in COSMOS-Web

data in Section 2.5.3. Even though the JADES-Medium and Deep programs have a much smaller area than COSMOS-Web, they are still expected to include ~ 25 and ~ 10 detectable systems respectively, extending to fainter source magnitudes at higher redshift than COSMOS-Web.

Even though NIRCam can only produce pencil-beam surveys, the cumulative area observed across all its surveys will inevitably lead to serendipitous lens discoveries. The total area of *JWST* surveys for Early Release Science and in Cycle 1 was > 1 deg^2 (Windhorst et al., 2022), doubling that of COSMOS-Web alone and now likely extending to several square degrees. Thus it is likely that the number of galaxy-galaxy lenses currently identifiable in *JWST* data is of order several hundred, a non-negligible fraction of the total number of lenses currently known, with corresponding excellent image quality and photometry in the near- and mid-infrared. A future archival search akin to Pawase et al. (2014) would help capture the full range of strong lenses in the *JWST* survey fields.

Prospects for VIDEO, UltraVISTA, Euclid & Roman

Strong lensing science will change significantly with the arrival of wide-area surveys, such as *Euclid* and *Roman*, identifying hundreds of thousands of lenses. These large samples of lenses will shrink the statistical uncertainties for a wide range of astrophysical and cosmological studies, for example constraints on the stellar IMF (Sonnenfeld et al., 2019), constraining the typical dark matter profile (Sonnenfeld and Cautun, 2021), and the equation of state of dark energy (Li et al., 2024, and see Chapter 5). Fig. 2.9 shows the number density of lens systems detectable for a range of existing and future surveys. Although the *JWST* programs mark a step-change in survey depth at NIR/Mid-Infrared (MIR) wavelengths and the number density of detectable lenses is an order of magnitude or so greater than that expected for the *Euclid* $I_{\rm E}$ band, their survey areas are $\mathcal{O}(10^5)$ smaller. However, regarding number density these small, deep surveys naturally surpass the wide area surveys which I consider. For example, the number of detectable lenses in COSMOS-Web is expected to be similar to that in the 22$\times$ larger area VIDEO survey and greater than

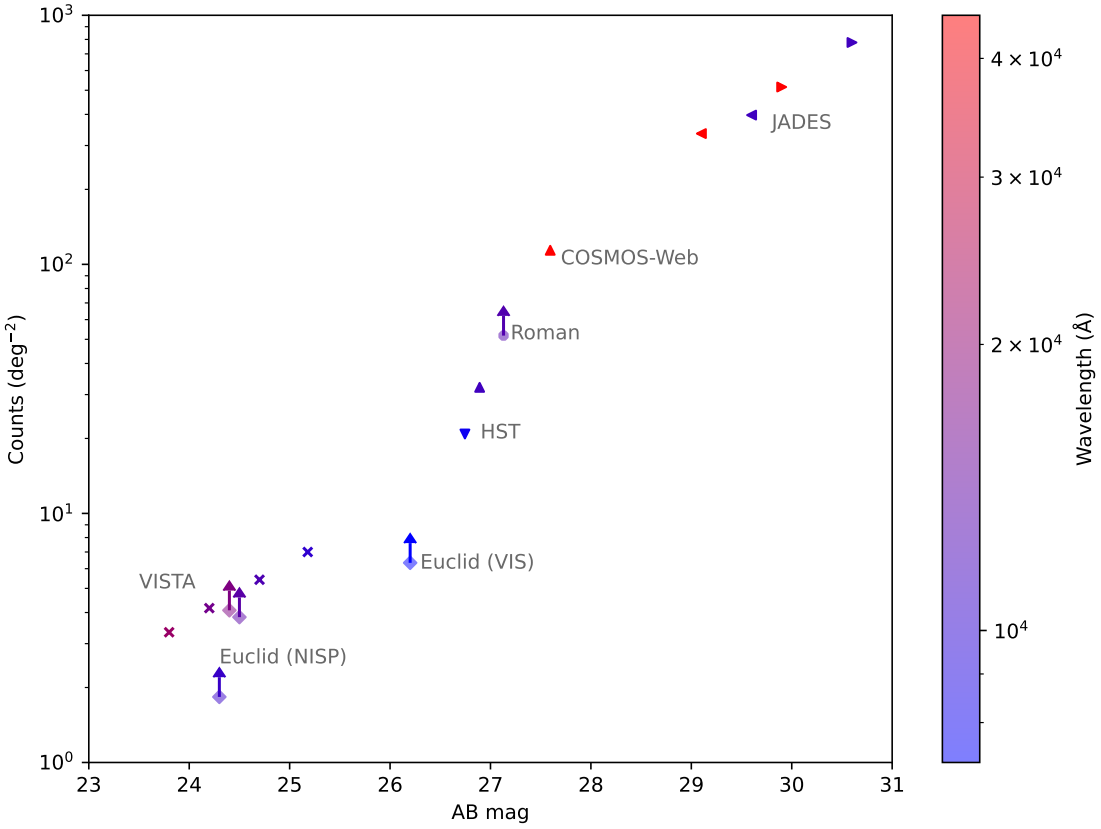


Figure 2.9: A comparison of the number density (lens systems per deg^2) of lenses detectable in a range of existing and future surveys. The data points are coloured by their central wavelength, with the following markers: HST: ▼, COSMOS-Web: ▲, JADES-Medium: ◀, JADES-Deep: ▶, *Euclid*: ♦, *Roman*: ●.

that in UltraVISTA. The high image-quality (PSF $\sim 0.05''$) of *JWST* enables more accurate identification of lens candidates and also allows for precision modelling of the most interesting individual systems. Naturally, the numbers discussed here are merely a guide of what to expect, and what fraction of these are scientifically useful will depend on the particular astrophysical or cosmological goals to be investigated.

Variation in the Number of Detectable Lenses with Waveband

The band-pass used for the lens search also influences the number of detectable lenses. This can vary due to many factors including survey depth, PSF and pixel-scale variation, and the source galaxy SED's. For COSMOS-Web, the F444W band has $\sim 3.5\times$ more detectable systems than the F115W band. In the VIDEO survey, the greater depth of the Y band reveals $\sim 2.1\times$ more detectable lenses compared

to the *Ks* band. Using multi-band imaging when conducting a lens search improves the ability of lens detection methods to separate the lens and source galaxies (e.g., Metcalf et al., 2019). Furthermore, more precise photometry in the NIR will give more accurate derived properties such as photometric redshifts and stellar masses.

2.5.2 Properties of the Detected Strong Lenses

Beyond the number of lenses detectable in different surveys, I explored how the properties of these lens systems varied by utilising the SED and photometry information present in the parent catalogue.

Variation of z_L, M_* and θ_E with survey

The lens stellar-masses and redshifts shown in Figure 2.5 are broadly unchanged by survey, all clustered around $10^{11} M_\odot$ and $z_\mathrm{L} \sim 1$. Unsurprisingly, the space-based surveys (*HST* and *JWST*) are found to probe smaller θ_E values than the ground-based VIDEO and UltraVISTA programs. This is due to the smaller PSF and pixel scale of *JWST* and *HST* compared to the typical seeing of ground-based surveys.

Comparing the lens systems detectable in the three COSMOS field surveys (COSMOS-Web, COSMOS-HST and UltraVISTA), the smaller angular resolution of *HST* and *JWST* allows these telescopes to identify a larger number of systems, extending to lower lens masses ($\sim 10^{10.5} M_\odot$) though the majority of lenses in all cases still have masses $\sim 10^{11} M_\odot$. The benefit of utilising lens-subtracted images also varies depending on the survey, as shown in Fig. 2.5c. Lens subtraction has the most use for the ground-based VIDEO search (an average gain of 50%) as the lenses tend to be brighter and thus more easily able to shroud any lensed images.

Prospects for Detecting Lensed High-Redshift Sources with JWST

One of the goals of the *JWST* blank field surveys is to detect galaxies at the very highest redshifts. While many such galaxies have been located in cluster fields ($z \sim 10$, Zitrin et al., 2014; Coe et al., 2013), here I look at high-redshift ($z > 6$) galaxy-galaxy lenses which were present in the simulations and could be detected in blank field surveys. Using these simulations one can determine the

probability of detecting galaxy-galaxy lenses at high-redshift. Drawing N source redshifts randomly 100 000 times from the simulated detectable source redshift distribution (where N is the number of detectable lenses in the survey from Table 2.2b), I found a 90% likelihood of a $z > 6$ lensed object in COSMOS-Web. In comparison to the JAGUAR population as a whole, such sources are more massive and luminous ($M_* \sim 10^{8.5}, m_{\rm F115W} \sim 26.5$) than typical high-redshift galaxies, as shown in Fig. 2.6. Although the source-redshift distributions of the deeper JADES programs also extend to high-redshift and contain more typical JAGUAR high-z sources ($M_* \sim 10^{7.5}, m_{\rm F115W} \sim 29$), their small survey areas mean few if any of these are likely to be observed.

Comparing Detectable NIR versus Visible Lens Systems

Based on my simulations, with full visible-to-NIR SEDs, one can explore whether strong lens searches in the NIR bring additional information over the traditional searches historically conducted in the visible (e.g., Sonnenfeld et al., 2020; Jacobs et al., 2019). First I consider the narrow-field *JWST* surveys compared to *HST*. Owing to the greater depth and wider wavelength coverage, nearly all those detectable with *HST* would be detected by *JWST* while only $\sim 40\%$ of the simulated systems detectable in COSMOS-Web would be also detectable in COSMOS-HST. Therefore, *JWST* will help to provide verification (and/or rejection) of candidate lenses previously discovered in the COSMOS-HST field. The sources absent from *HST* but detectable by COSMOS-Web are typically fainter by $\Delta m_{\rm F814} \sim 1.3$ (in lensed magnitude) and higher redshift ($\Delta z_S \sim 0.3$).

In contrast to *JWST* vs *HST*, the *Euclid* $I_{\rm E}$ band is substantially deeper and has better image quality than its NISP channels (i.e., $Y_{\rm E}$, $J_{\rm E}$, $H_{\rm E}$ bands, PSF $\sim 0.3''$ versus $\sim 0.17''$ in $I_{\rm E}$) which increases the number of lenses detectable in the visible compared to the NIR. Despite the lower image quality, the NISP bands add colour information, which makes strong lens identification easier than using single-band imaging. For example, verifying that all the images surrounding the possible lens galaxy have the same colours can lend confidence to these being lensed images of the

same source. Since lensed systems commonly feature early-type elliptical galaxies lensing late type spiral galaxies, identification of blue arcs around a central redder galaxy can also help with lens classification. This would be true for even those $I_{\rm E}$-identified strong lenses that lie below the SNR > 20 threshold for detectability in the NISP channels. The NISP bands will offer new systems too. Across all the lens systems detectable with *Euclid*, 43% are only detectable in the $I_{\rm E}$ channel, 35% in both $I_{\rm E}$ and NISP channels and 22% are detectable only in the NISP channels. The lensed sources not detectable in the $I_{\rm E}$ channel are typically at lower redshifts ($\Delta z_S \sim 0.8$) and have redder colours ($\Delta(I_{\rm E} - Y_{\rm E}) \sim 0.9$). On average, they are also dustier (the V-band dust attenuation differing by $\Delta\tau_V \sim 0.3$) and have older stellar populations ($\Delta Log(t_{\rm max}/yr) \sim 0.5$) where $t_{\rm max}$ is the maximum age of stars in the galaxy. This suggests that (a) even accounting for the lower image quality in the NISP channels, it is worthwhile running lens searches utilising the NISP bands alongside *Euclid* $I_{\rm E}$ and (b) such a search would include a component of dusty, moderate redshift lensed sources (e.g., Geach et al., 2015). Lensed, ultracompact quiescent galaxies (e.g., Muzzin et al., 2012) may also be detected. The frequency of $z > 2$ lensed quiescent galaxies is estimated to be $0.5 - 1$ deg^{-2} (Muzzin et al., 2012) so would be unlikely to be detected in *JWST* surveys of similar size to COSMOS-Web, but highly likely to be observed with *Roman* and *Euclid*.

2.5.3 Validation of Lens Occurrence Rates with Recent Strong Lens Discoveries

Following the completion of this work, strong lenses have been identified in two of the surveys of interest, through COSMOS-Web *(JWST)* and *Euclid* data. I discuss each of these in turn in this section.

COSMOS-Web

A systematic search for strong lenses in the COSMOS-Web survey was published in Nightingale et al. (2025) as part of the COSMOS-Web Lens Survey (COWLS). They identified 79 high-grade lens systems and 160 mid-high grade systems. Of these, 17 were labelled as 'Spectacular' and are discussed in detail in Mahler et al.

(2025). My estimates ($\sim$ 65 in COSMOS-Web) are within 20% of the number of high-grade systems, but if the mid-grade systems are truly lenses would only account for 40% of the total identified, which may be expected given the conservative nature in my assumptions. Based on the modelling and photometry of these systems (Nightingale et al., 2025; Hogg et al., 2025), in Figure 2.10 I plot a comparison of the measured Einstein radii, lens magnitudes and photometric lens redshifts (source redshifts are not yet available). There is excellent agreement between the lens redshift distribution of the observed lens candidates and the simulated COSMOS-Web lens population. This confirms the prediction that the COSMOS-Web data will contain higher redshift lenses than previously identified. Extending known lenses up to $z \sim 2$ (exemplified by the COSMOS-Web ring, below) will enhance studies of the IMF (e.g., Sonnenfeld et al., 2019) and mass distribution within galaxies (e.g., Etherington et al., 2023; Sheu et al., 2024) helping to probe whether these properties evolve with redshift. There is also good agreement between the lens magnitudes of the predictions and lens candidates. Brighter lens systems (indicating more massive lenses) were typically given higher-grades in the COWLS search, with the predictions matching the high-best candidate distributions the closest. The simulation slightly underestimates the proportion of systems with small Einstein radii. This would support the hypothesis that the detection criteria applied to the simulations were too strict here, since the small Einstein radius systems would be the most difficult to identify (and as shown by Figure 2.10, were on-average given a lower grade by the COWLS team). However, lens modelling applied by the COWLS search helped to identify a significant proportion of the lens candidates which otherwise may have been missed (Nightingale et al., 2025). 38.3% of highly scoring candidates were edge-cases, only flagged by one inspector (out of 4/5) as potential lenses prior to the inspectors being provided with lens modelling information. This demonstrates the significant benefit of modelling to identify lens candidates, and suggests the stringent detectability requirements assumed here can be relaxed when searches include modelling. It confirms that the estimates presented here are conservative, and that lens modelling provides greater confidence

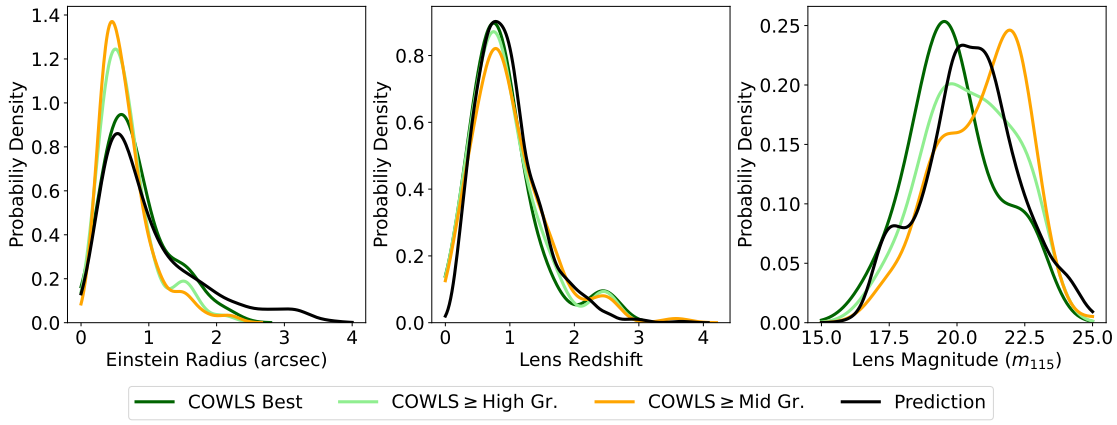


Figure 2.10: Comparison of the distribution of Einstein radii, lens redshift and lens magnitude from my simulations in comparison to COWLS (Nightingale et al., 2025) lens candidates. In each plot, the Spectacular ('COWLS Best'), high-grade ('≥ High Gr.)' and mid-high ('≥ Mid Gr.)' systems are plotted separately. There is good agreement between the lens candidates and simulation for the lens redshifts and magnitudes, while the simulation slightly underestimates the proportion of systems with small Einstein radii.

in the validity of ambiguous edge-case (e.g., low SNR) lens candidates. If such modelling were used at scale (e.g., in *Euclid* lens searches), it is likely that a greater number of high-grade lens candidates would be identified than currently predicted.

A high-redshift strong lens system (the 'COSMOS-Web Ring') was identified in COSMOS-Web imaging by Mercier et al. (2024) and van Dokkum et al. (2024). The latter measured photometric redshifts of $z_L = 1.94^{+0.13}_{-0.17}$ and $z_S = 2.98^{+0.42}_{-0.47}$, with a stellar mass $M_*(< \theta_E) = 1.1^{+0.2}_{-0.3} \times 10^{11} M_\odot$ with a Chabrier IMF. The former combined *JWST*/*HST* photometry with ground based imaging and measured photometric redshifts of $z_\mathrm{L} = 2.02 \pm 0.02$, $z_\mathrm{S} = 5.48 \pm 0.06$, with a stellar mass $M_*(< \theta_E) = 1.25 \times 10^{11} M_\odot$ (for the same IMF) and Einstein radius $\theta_E = 0.78''$. The source redshift was recently spectroscopically confirmed by Shuntov et al. (2025) using NOEMA (Northern Extended Millimeter Array) and Keck/MOSFIRE (Multi-Object Spectrometer For Infra-Red Exploration) data to be at $z = 5.10$. Therefore, this system contains both the highest redshift lens and highest redshift lensed source galaxy known to date. Figure 2.11 shows the location of this system in comparison to the redshift distribution of detectable lenses predicted in this work. While at the extreme of the redshift distribution, such a system would be expected within the $\mathcal{O}(100)$ lenses identified by COWLS. The source redshift measurements of van

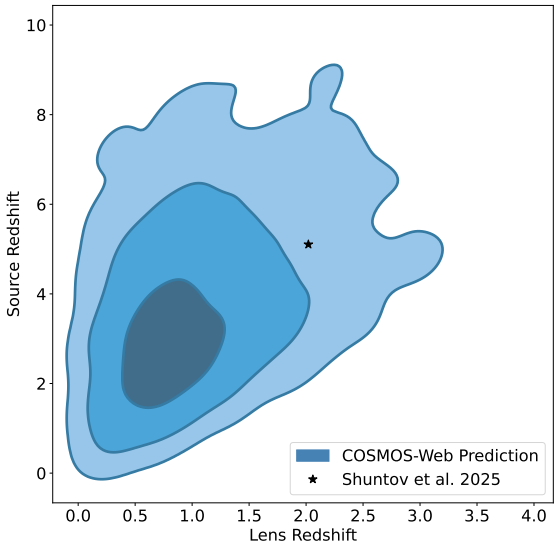


Figure 2.11: Plot of the distribution of lenses predicted to be detectable in COSMOS-Web versus the redshift of the COSMOS-Web Ring (Mercier et al., 2024; van Dokkum et al., 2024; Shuntov et al., 2025). The contours show 1, 2 and 3σ levels which in 2 dimensions enclose 39.3%, 86.5% and 98.9% of data points.

Dokkum et al. (2024) and Mercier et al. (2024) are significantly different. While van Dokkum et al. (2024) used photometry from *JWST* and *HST*, Mercier et al. (2024) used additional ground-based imaging from HSC, UltraVISTA, the Canada-France-Hawaii Telescope (CFHT) and Subaru Suprime Cam (Taniguchi et al., 2007; Taniguchi et al., 2015), which could more tightly constrain the SED. Furthermore, dependent on the photometric redshift fitting code and the photometry used (e.g., a single lensed image versus the whole ring), Mercier et al. (2024) obtain a source redshift in the range 4.78-5.48, which encompasses the spectroscopic measurement of Shuntov et al. (2025); the systematic uncertainties due to these choices are therefore larger than the presented statistical uncertainties in this case. Overall, while the uncertainties presented are too small in both studies, the much broader range of photometry used by Mercier et al. (2024) allowed their result to be more accurate.

These systems would make ideal targets for follow-up with the adaptive-optics assisted Extremely Large Telescope (ELT). This would provide high-resolution lens kinematics for measuring galaxy mass distributions (as mentioned above) and would aid with source characterisation to probe the morphologies of high-redshift galaxies, benefiting from the additional lensing magnification. The deeper, pencil-beam

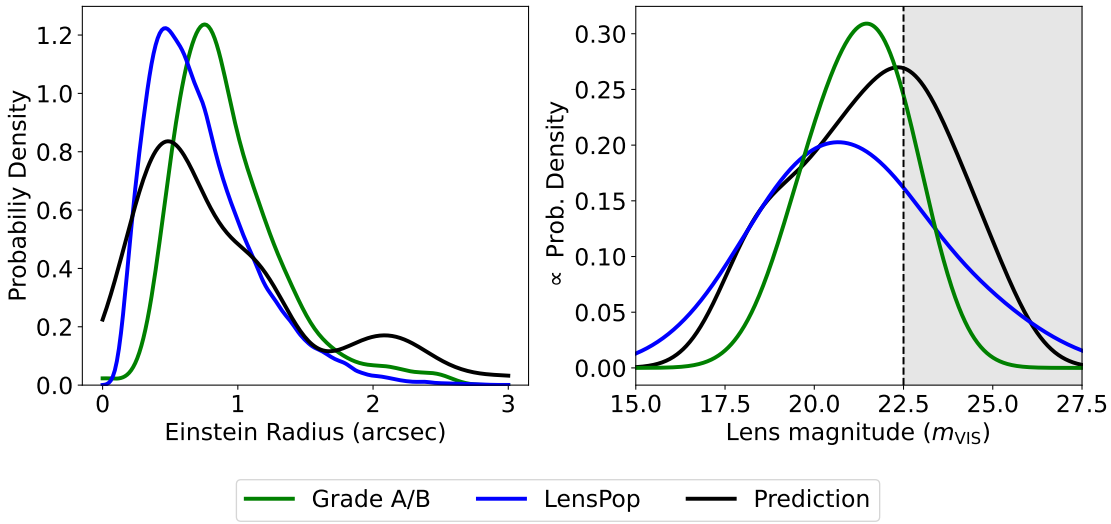


Figure 2.12: Comparison of the distribution of Einstein radii (left), and lens magnitude (right) from my simulations in comparison to the A/B grade lens candidates identified by the Strong Lens Discovery Engine (Euclid Collaboration: Walmsley et al., 2025). In the right-hand plot, the curves have been scaled to have the same area below $I_{\rm E} < 22.5$ for ease of comparison, reflecting the selection cut made in the Q1 lens search.

JWST surveys (such as JADES-Deep) which currently probe small areas, have the potential to extend these redshift limits further as their cumulative area grows.

The NIR bands of *Euclid* and *JWST* allow new regions of parameter space to be explored, in particular very red galaxies which may not be detectable in visible bands due to their dust content, stellar age or redshift. Early releases of *JWST* data of the COSMOS field revealed a sub-mm lens candidate (Pearson et al., 2024), with $z_L = 0.360$ and $z_S = 3.4 \pm 0.4$. Extrapolating their findings to the *Euclid* survey, Pearson et al. (2024) suggest $62\,000 \pm 44\,000$ sub-mm lenses are detectable in NISP but undetectable in $I_{\rm E}$ data. This is consistent with, but more optimistic than, my estimates[22] of $\sim$ 26 000 NISP-only systems. In either scenario, such lensed counterparts of sub-mm galaxies will enable us to investigate the sources that dominate the star formation density at cosmic noon at high-resolution (Zavala et al., 2021).

Euclid

Early data releases from the Euclid Wide Survey have revealed numerous lens candidates. Lens searches were conducted in two parts (1) using Early Release Observations (ERO) as discussed in Acevedo Barroso et al. (2024) and (2) using Quick Data Release 1 (Q1), discussed in Euclid Collaboration: Walmsley et al. (2025). The latter was a systematic search combining ML and Citizen Science and is discussed in detail in Chapter 4. This search identified 497 A/B grade lens candidates over an area of 63.1 $\deg^2$. As noted by Euclid Collaboration: Lines et al. (2025), we estimated a completeness of 98% for Grade A and 47% for Grade B systems in this search. Extrapolating these results to *Euclid* DR3 would suggest we will find $\sim 118\,000$ lens systems (of $\sim 185\,000$ accounting for the incompleteness). In comparison, my estimates suggest $\sim 108\,000$ across *Euclid* bands, while Collett (2015) predicts 170 000 detectable lenses in the *Euclid* $I_{\rm E}$ band (though the latter assumes lens subtraction which was not used in the Q1 search). Considering that the estimates presented in this work were based on a small-area simulated catalogue and that some of the grade B candidates in the search will be false positives, one can consider the estimates presented in this chapter and in Collett et al. (2012) to be largely consistent with the grade A/B lenses discovered in the Euclid Q1 data.

In Figure 2.12 I plot a comparison of the Einstein radius and lens magnitude from my predictions compared to those found in the Q1 lens search, and the predictions of LensPop (Collett, 2015). Based on the selection criteria of the search ($I_{\rm E} < 22.5$) and the image scaling (cutout intensity was dependent only on $I_{\rm E}$ flux[23]), I only plot simulated systems detectable in the $I_{\rm E}$ band. My predictions underestimate the number of larger Einstein radius systems, $\theta_E \gtrsim 1''$ which given the small area of the JAGUAR catalogue is unsurprising although the lens magnitude distributions show relatively good agreement for $I_{\rm E} < 22.5$. My predictions, and those of Collett (2015) suggest loosening the lens magnitude threshold beyond $I_{\rm E} < 22.5$ would

[22]While I do not include explicit uncertainties here, I direct readers to Sections 2.4.4 and 2.5.4 for a discussion on the primary uncertainties associated with these estimates.

[23]Due to the different pixel scales of the $I_{\rm E}$ and NISP bands, to maintain the resolution of the $I_{\rm E}$ image the brightness of the cutout was determined by the $I_{\rm E}$ flux while the NISP bands were used to inform the hue and saturation.

identify a non-negligible sample of additional lenses. Given that the Q1 search primarily probed $I_{\rm E}$-detectable systems, my occurrence rate estimates and those of Pearson et al. (2024) would suggest that a NIR-focussed lens-search in *Euclid* data would also be a very beneficial next step.

2.5.4 Potential Further Improvements

The main shortcoming of the calculations in this work is the small area of the original JAGUAR galaxy catalogue on which the simulations were based. This brings with it the uncertainties of small-number statistics (in particular from the brightest galaxies which are the most likely lenses in wide-area surveys such as *Euclid* and *Roman*), and cosmic variance[24] (estimated $\leq 30\%$ for the surveys considered here, and $\leq 10\%$ for COSMOS-Web, Trenti and Stiavelli, 2008). The original catalogue was chosen so that the lens and source galaxies could be determined self-consistently, and to broaden the range of possible deflectors to include late-type spirals. However, a wider area catalogue would allow more rigorous comparisons with other predictions such as those in Collett (2015) and Ferrami and Wyithe (2024). Both of these lensing estimates (Collett, 2015; Ferrami and Wyithe, 2024) explicitly utilised the lensing cross-section to determine the number of detectable lenses, based on a chosen Velocity Dispersion Function (VDF) and source galaxy catalogue. Such models offer significant flexibility (e.g., allowing measurement of the effect of differing VDF's). However, adopting a single catalogue/simulation for the lens and source population (as in this work) permits analysis which is more agnostic to deflector type (i.e., not restricted to early-type galaxies) and allows more detailed analysis of the properties of individual lens systems than was assumed in these works.

In this work, I adopted the detectability constraints of Collett (2015). Most lens candidates identified to date have undergone expert inspection to verify that

[24]The cosmic variance here includes both the Poisson noise and the variance due the large scale structure which becomes significant for small scale surveys. The latter is calculated by integrating the dark matter two point correlation function over the redshift volume of interest (Trenti and Stiavelli, 2008). The upper limit of 30% was calculated based on the number of lenses detectable in the *Euclid* $Y_{\rm E}$ band (i.e., the band with the smallest number of detectable lenses per square degree of those considered here) and was evaluated for the total area of the JAGUAR catalogue ($0.34\,{\rm deg}^2$) over the redshift range of detectable lenses in this band.

they are truly lensed, while in the future this verification may be left to machine learning algorithms or citizen scientists (see Chapter 4). Such a selection function can be modelled by incorporating simulated lenses within the lens grading process as in Euclid Collaboration: Walmsley et al. (2025) and would need to be accounted for before performing an in-depth comparison between the predictions here and a real lens search. Ideally this would also account for the help which lens modelling provides to grading edge-case lens candidates, as was the case in the COWLS search.

An ideal lens occurrence rate estimator would be based on a wide-field, $\mathcal{O}(10^4)$ deg^2, deep ($z \lesssim 8$) catalogue incorporating galaxy clustering and full SEDs (and thus galaxy properties), as well as self-consistent baryonic+dark-matter profiles. It would further need to incorporate an empirical lens selection function based on existing lens searches to determine lens detectability in each survey. Nonetheless, simpler approaches can still provide significant benefit for planning lens search strategies and for determining which systems may have been missed after a search has been completed.

2.6 Conclusion

In this chapter I generated lensing frequency estimates for strong lenses detectable in existing and forthcoming telescope surveys. To do this I used the JAGUAR and DREaM galaxy catalogues to generate a realistic galaxy population from which to draw lens and source galaxies. The resultant galaxy catalogue was small in area and suitable for studying the lens population detectable in $\lesssim 10\,\mathrm{deg}^2$ surveys such as *JWST*, VIDEO and UltraVISTA. I explored the number and properties of strong lenses in surveys utilising *JWST*, *HST*, VISTA, *Euclid* and *Roman* telescopes. My conclusions are as follows:

- *JWST* contains $\gtrsim 100$ lenses in the *JWST* Early Release Science and Cycle 1 data. This has been confirmed by the recent COWLS lens search in COSMOS-Web, identifying $\mathcal{O}(100)$ lens systems. Out of the *JWST* programs investigated, chosen to be illustrative of the range of surveys *JWST* will undertake, my

simulations suggest COSMOS-Web contains the largest number of detectable lenses (~ 65 across all bands), with a further ~ 25 across JADES-Medium and JADES-Deep.

- The resolution of *JWST* will make these lens candidates high-value targets for spectroscopic follow-up and modelling. As exemplified by the COSMOS-Web Ring, *JWST* surveys will extend the redshift limits of lens and source galaxies, likely reaching $z_S > 6$ sources. An archival search of available *JWST* data would likely reveal hundreds more lens candidates and could feasibly be conducted by a citizen science search.

- Out of the surveys I investigated, my model suggests multiple imaging could be detectable (at SNR > 5) in $55 - 80\%$ of the detectable strong lens systems. Deeper surveys have higher proportions of lensed systems with detectable multiple images. My simulations also suggest that lens subtraction is more beneficial for ground-based surveys compared to space-based surveys, e.g., increasing the yield by $\sim 50\%$ for the VIDEO survey.

- Of all the strong lens systems detectable by *Euclid*, $\sim 20\%$ would only be detected by a lens search which included the NISP channels. These could include dusty $z \sim 2$ lensed galaxies which would be missed in a search in the visible. The anticipated large population of NIR counterparts of sub-mm galaxies hinted here and suggested by Pearson et al. (2024) will be valuable for studies of galaxy evolution at cosmic noon.

- The wide-field *Euclid* and *Roman* telescopes will provide a step-change in the number of lenses known. Their resolution will allow much more complex lens modelling than can be undertaken with current ground-based wide-field surveys (HSC, DES, KiDS), while their NIR filters will extend the detectable lens population to higher redshift. It is likely that spectra will only be available for a subset of these systems. Given this, the photometric lens population (i.e., systems without spectroscopic confirmation) has the potential to play an important role in constraining population-level parameters in the future.

- Simulated populations of strong lenses, as produced in this chapter, remain a valuable tool for defining strategies (such as survey bands or search area) for lens searches and prioritising the most valuable surveys for inspection (e.g., COSMOS-Web versus JADES). Furthermore, such simulations can help demonstrate which lenses were missed from the detectable population, informing future searches.

In summary, the next decade will see a substantial increase in strong lensing data across both small and large scale surveys. However, utilizing such data will require changes in our detection and analysis methods, the subject of the subsequent chapters of this thesis.

3
A Bayesian Approach to Strong Lens Finding

The basis of this chapter was first published in the journal article 'A Bayesian approach to strong lens finding in the era of wide-area surveys', Holloway et al., 2024b. My contributions to the journal article González et al., 2025 (currently in review), are also discussed.

Contents

3.1 Introduction 75
3.2 Data 78
3.3 Method 81
3.3.1 Summary of Calibration Methods 82
3.3.2 Application of Calibration Methods 84
3.3.3 Summary of Combination Methods 88
3.4 Results 92
3.4.1 Testing the Bayesian Combination Approaches 92
3.4.2 Applying the Bayesian Combination Methods 93
3.5 Discussion 96
3.5.1 Comparison with Previous Work 96
3.5.2 Comparison of Citizen Science versus a Network Ensemble 97
3.5.3 Effect of Ground-Truth Selection on Classifier Performance 98
3.5.4 Expectations and Implications for LSST 100
3.5.5 Application to the Dark Energy Survey 103
3.6 Conclusion 105

The wide-field surveys of LSST and Euclid present a challenge to the strong lens community; strong lenses are rare and difficult to find, and the data volume to search through will be vast. In this chapter I present two methods to address these challenges, creating an ensemble of different types of classifiers, and producing calibrated lens probabilities for large-scale analysis. This is applied to HSC data as a proxy for LSST, and subsequently to data from DES as further verification of the method.

3.1 Introduction

Strong lens systems are intrinsically rare, requiring the close angular alignment between lens and source galaxies. Furthermore, the lens and source galaxies can be diverse and have different configurations. This combination of rarity and variety means it is very challenging to identify strong lens systems in wide-field surveys in which the search volumes can be considerable. As such, lens searches can be plagued with false positives i.e., non-lens systems which receive high scores from lens classifiers.

In early lens searches (e.g., using HST data, Faure et al., 2008; Jackson, 2008), the data volume was sufficiently small such that images could be inspected by hand by a small number of lensing experts (see Section 1.2.4). As data volumes have grown, more scalable methods have been required to allow inspection of a larger number of images. Citizen science, in particular through the Space Warps project (Marshall et al., 2016; More et al., 2016), has received widespread interest from the public and provided balance against automated methods which can miss unusual systems. Automated algorithms such as `Arcfinder` (Seidel and Bartelmann, 2007), `RINGFINDER` (Gavazzi et al., 2014) and `YATTALENS`, (Sonnenfeld et al., 2018) have also helped with scalability, identifying extended ring structures within an image. In recent years, these have been overtaken by machine learning methods such as

Convolutional Neural Networks (CNN's) and Vision Transformers (VT's) which have been adopted for numerous lens searches.

Although these automated methods are continually improving over time, contemporary lens searches still identify large numbers of false positives. To mitigate this, the high scoring samples typically undergo visual classification by strong lens researchers ('expert inspection') to grade likely lens candidates and remove the false positives. In recent lens searches (Sonnenfeld et al., 2020; Cañameras et al., 2021; Rojas et al., 2022) the number of high-scoring candidates which underwent such expert inspection exceeded the resulting number of A-B grade lenses by factors $\geq 10\times$, i.e., the number of false positives significantly outnumbered the number of lens candidates. Even with current medium-deep surveys of tens-hundreds of square degrees, expert inspection is time-consuming and difficult; strong lens identification is not clear-cut, particularly with seeing-limited ground-based imaging or when using single-band imaging for classification. Investigations by Rojas et al. (2023) found that ~ 6 expert graders were required before the expert classification converged with classification error[1] below 0.1 (on a $0-1$ grading system), due to the frequency of disagreement between experts on which systems were truly lensed. The forthcoming LSST and *Euclid* surveys are expected to identify $\mathcal{O}(10^5)$ lenses however expert visual inspection of the $> 100\,000$ lens candidates identified in these surveys will be an unenviable task. The work described here tackles the above challenges through the following aims:

- **To produce calibrated probabilities that a given galaxy system is a lens**. I define a calibrated score X as one for which systems with this score are indeed lenses X% of the time, as verified on a distinct test set. In this work I test multiple methods for performing this calibration, and demonstrate their respective differences. For wide area surveys, follow-up or visual inspection of all lens candidates will not be possible. Therefore, it will be important to have accurate probabilities for lens candidates in

[1]Defining this error as the standard deviation of the difference between the average grade of different teams of a given size compared to the 'true' grade from the average of ~ 20 graders.

order to perform unbiased statistical analysis at the population level, the alternative being significantly reducing the sample-size to those which have been spectroscopically or visually confirmed. Furthermore, having accurate probabilities allows direct comparison of lens candidates across different lens classifiers.

- **To create an ensemble strong lens classifier**. A given survey may be targeted by multiple lens finding methods, each with their own strengths and weaknesses. Neural network ensembles have been investigated previously (e.g., Schaefer et al., 2018; Cañameras et al., 2024 for galaxy-galaxy lensing and Andika et al., 2023 for lensed quasars), simply averaging over the individual network scores. However, such a heuristic approach would not necessarily provide an optimal ensemble and may even degrade the overall performance. Similarly, combining scores from different types of classifier (e.g., ML+citizens) simply through averaging, would have no guarantee of improving the results. Here, I take the ensemble approach further by incorporating the calibrated probabilities generated above in a Bayesian framework, and use a citizen science classifier to diversify the ensemble beyond neural networks. This is the first time that such an ensemble has been created for strong lens searches. Combining multiple classifier scores into an ensemble has a number of advantages beyond helping to achieve the primary goal of improving classification performance. Firstly, it presents a structured mechanism in which to prioritise expert visual inspection and follow-up, both of which have limited capacity. Furthermore, combining multiple classifiers helps mitigate the potential biases of any particular classifier. For example, artifacts such as stars which may receive high scores from a network due to being unrepresented in the training set would be very likely to receive a low score from citizens, which would reduce the resultant ensemble score.

In the process of reaching these goals, I aim to answer the following questions:

1. Is there a significant improvement in performance using an ensemble method compared to a single classifier?

2. Are there regions of parameter space which suit certain methods best?

3. How does the degree of improvement change when combining only neural network classifiers, compared to combining neural networks with a citizen science classifier?

The work in this chapter aims to alleviate the false positive challenge presented by wide-field surveys such as LSST and *Euclid* by building upon the particular strengths of individual classifiers though an ensemble. This chapter is structured as follows. I detail the data used in this work in Section 3.2. In Sections 3.3.1 and 3.3.2, I summarise and apply different classifier calibration methods. In Section 3.3.3, I combine these calibrated probabilities into a single ensemble classifier. My results are presented in Section 3.4 and discussed further in Section 3.5, including implications for forthcoming lens searches in LSST and other wide-field surveys in Section 3.5.4. I conclude in Section 3.6.

3.2 Data

To develop an ensemble classifier, I initially used the classification scores from six strong lens finders applied to Hyper-Suprime Cam Subaru Strategic Program data (HSC SSP, Aihara et al., 2022). This, being a ground based survey over hundreds of square degrees, was a useful proxy for LSST-like data. The wide-area survey by HSC has to date been conducted in *grizy* bands, covering $\sim 1200\,\mathrm{deg}^2$. The HSC S17A and PDR2 data releases (Aihara et al., 2018; Aihara et al., 2019) covered $1026\,\mathrm{deg}^2$ (225) and $1114\,\mathrm{deg}^2$ (305) respectively in at least one band (all bands) to a 5σ depth of $i \sim 26.2$. Only the *gri* bands were used in this work. The strong lens classifiers used are described below, along with the selections employed by their respective lens searches. Each classifier gave the systems in the search dataset a score corresponding to the likelihood that they were a lens candidate:

- **Citizen Science** (Sonnenfeld et al., 2020): A citizen science search by the Space Warps project (see Section 1.2.4), which used the HSC S17A data release (Aihara et al., 2018; Aihara et al., 2019) and provided citizen classifications for $\sim 300\,000$ objects. The galaxies selected for this search had photometric redshifts $0.2 < z < 1.2$ and inferred stellar masses $\log(M_*/M_\odot) > 11.2$.

- **Neural Network 1** (Cañameras et al., 2021): A ResNet classifier, which used HSC PDR2 data (Aihara et al., 2019) over a much larger sample of objects compared to the citizen science search (network scores were available for 5.4×10^7 objects). The objects chosen for this search were required to have i-band Kron radius $\geq 0.8''$, to narrow the sample to be searched (this cut implies the minimum image separation for any potential lensed image would be greater than the median seeing).

- **Neural Network 2** (Shu et al., 2022): A ResNet classifier from Lanusse et al. (2018), using HSC PDR2 data and, in particular, targeting high-redshift lens systems. The primary selections for this search were colour cuts: $0.6 < g-r < 3.0$ and $2.0 < g-i < 5.0$ to identify red, high-redshift galaxies. These cuts were taken from Jacobs et al. (2019) to select lens galaxies with $z \geq 0.8$. The lenses in the training set of this network lay in the range $0.1 \lesssim z \lesssim 1.0$.

- **Neural Network 3**: The second classifier presented by Shu et al. (2022), differing from the first in having a higher fraction of $z > 0.6$ lenses in the training set and altering the image pre-processing to include square-root image stretch and normalisation. The lenses in the training set of this network lay in the range $0.4 \lesssim z \lesssim 1$.

- **Neural Networks 4** (Ishida et al., 2025) **and 5** (Jaelani et al., 2023): These networks were both CNN's applied to HSC PDR2. They shared the same training set, in particular targeting smaller Einstein radii down to $0.5''$ (the distribution of Einstein radii in the training distribution of Einstein radii decreased roughly exponentially from $0.5''$ to $3.0''$ while Networks 1-3 focused

on $\theta_E \gtrsim 0.75''$). The galaxies were originally selected based on three criteria: stellar mass $> 5 \times 10^{10} M_{\odot}$, specific star formation rate $< 10^{-10}\,\mathrm{yr}^{-1}$, and redshift range $0.2 < z < 1.2$. However, in this work, these CNN's were applied to the sample of galaxies with classifier scores available from all of the first 4 classifiers listed.

The objects in each catalogue were cross-matched, with a maximum separation of 1", to produce a sample of 126 312 unique galaxies with both citizen science (CS) and neural network (NN) scores. The main cuts from the parent samples which affected the resulting cross-matched sample were therefore:

- Photometric Redshift: $0.2 < z < 1.2$ (CS)
- Stellar Mass: $\log(M_*/M_{\odot}) > 11.2$ (CS)
- Kron Radius (i-band): $R_{\mathrm{Kron,i}} \geq 0.8''$ (NN 1)
- Colour cut 1: $0.6 < g - r < 3.0$ (NN 2 + 3)
- Colour cut 2: $2.0 < g - i < 5.0$ (NN 2 + 3)

A subset (109 128 objects), with ≥ 5 citizen classifications, was used in the following analysis. The classifiers in this work had a wide range of training sets and original target selections. Therefore the resulting ensemble could benefit from the individual strengths of each classifier. In this work I did not favour lenses with certain properties (e.g., higher-redshift, Einstein radius etc.) over others, and measured the performance of each classifier simply using the cross-matched data set. The properties of the lens candidates (such as redshift, stellar mass, or colour) identified by each classifier were similar, even though the parent distributions varied between classifiers. Therefore, it was reasonable to combine these classifiers into an ensemble. In order to calibrate the output scores of the classifiers, I required a 'ground-truth', i.e., a list of classified objects of known lenses/non-lenses. I collated such a list using expert grades from four sources: SuGOHI VI (Sonnenfeld et al., 2020), Highly Optimised Lensing Investigations of Supernovae, Microlensing Objects, and

Kinematics of Ellipticals and Spirals (HOLISMOKES) VI (Cañameras et al., 2021), HOLISMOKES VIII (Shu et al., 2022) and the online SuGOHI database[2] of lens candidates. Since the expert grades from these reference datasets sometimes differed for a given system, I assigned each cross-matched object with a final grade according to the following order: HOLISMOKES VI, SuGOHI VI, SuGOHI database and HOLISMOKES VIII. These grades (G) were indications of the confidence a given object was a lens. They were defined in the range $[0, 3]$ where $G = 3$ indicated 'certain lens' (e.g., multiple images in the correct configuration), $G = 2$ indicated 'probable lens' (most features consistent with a strong lens), and $G = 1$ indicated 'possible lens' (e.g., a single arc which could well be a contaminant). Here I refer to A-B grade lenses as those with $G \geq 1.5$, and A-C grade lenses as those with $G \geq 1.0$. There were 34 objects whose grades differed by ≥ 2 which I inspected and re-assigned an appropriate grade (typically taking the average of the relevant grades except in clear-cut cases). In total, 3744 cross-matched objects had assigned grades, of which 189 were graded A or B and were treated as true lenses.

In addition to the above-mentioned classifiers from HSC, I also applied the same methodology to two lens classifiers as part of a lens search in DES, as described in González et al. (2025). In this search, 2.36×10^8 cutouts were analysed by a VT algorithm, of which the highest scoring 22 564 systems were inspected by citizens as part of a Space Warps project[3]. I describe the results of applying the ensemble methodology to this data in Section 3.5.5.

3.3 Method

My first aim was to produce calibrated scores for each classifier individually. The graded objects were split into three sets: a training ('calibration') set, a validation set and a test set, in ratios 2:1:1. I considered A+B grade lenses as true lenses and the remainder (graded or otherwise) as non-lenses. Excluding the ungraded objects would lead to unrealistic values for the receiver operating characteristic

[2] http://www-utap.phys.s.u-tokyo.ac.jp/~oguri/sugohi/

[3] https://www.zooniverse.org/projects/aprajita/space-warps-des-vision-transformer

(ROC) curve and purity-completeness summary statistics and it was likely the vast majority of true lenses were identified by at least one of the sources of expert grades. While using simulated lenses rather than lens candidates would eliminate the uncertainty in the true nature of lens candidate systems (e.g., whether grade B lenses are actually false positives), this could bias the network performances, favouring the networks trained on similar simulations (i.e., those with a smaller distribution shift between their training set and the test set). This would result in an overestimation of the sample completeness of a potential lens search. By using real lens candidate systems, I ensured that the systems denoted as true lenses were sufficiently complex/realistic by definition, and thus provide an accurate representation of the performances of each classifier on real data.

3.3.1 Summary of Calibration Methods

The purpose of calibration was to produce accurate probabilities that a given object was a lens. The true distribution of (raw) classifier scores which a large population of lenses (and non-lenses) would receive from a given classifier is generally not known but could be inferred from a sample of data (this is known as density estimation). The calibration methods used here are similar to such density estimation. However, here I produce a continuous distribution function of the proportion of objects with a given score which are lenses, inferred from a finite sample of classifier scores and their classifications. These are then tested against a separate test set to verify that the calibration is robust. The calibration methods considered in this work were selected either from common calibration methods in the literature, or from experimentation via toy models (described below). These are:

- Isotonic regression (Zadrozny and Elkan, 2002): A non-parametric, discontinuous fit of a monotonically increasing curve to the classifier output score distribution.

- Platt scaling (Platt, 2000): A parametric calibration method of the form:

$$P(L|x) = \frac{1}{1 + \exp(Ax + B)} \tag{3.1}$$

where A and B are fitted to the data and $P(L|x)$ refers to the probability that a system is a lens ($=L$) given a classifier score x. Although commonly used, this requires the calibration curve to closely match a sigmoid function, which *a priori* may not be true.

- Kullback Leibler Importance Estimation Procedure (KLIEP, Sugiyama et al., 2008): This minimises the Kullback–Leibler (KL) divergence of the ratio $f(x|L)/f(x)$ with probability density functions $f(x|L)$ and $f(x)$ of a classifier score x (given the subject is a lens in the first case). From Bayes' Theorem, weighting this by $P(L)$ gives the probability that the subject is a lens, given its classifier score. KLIEP uses a Gaussian Mixture Model (GMM), with Gaussians of fixed width placed at the positions of the lenses in classifier score space. Their respective weights are tuned to minimise the KL divergence.

- Variable Bin Fitting: This was designed to account for the small number of lenses in my sample, compared to the much larger number of non-lenses. I created a modified histogram made up of overlapping bins and rather than the bins having fixed width, I required 5 lenses per bin. Bins at lower scores were wider (due to the lower occurrence of lenses) and correspondingly narrower at higher scores. The calibration curve was then formed from the fraction of lenses in each bin compared to the total bin occupation. This provided higher resolution calibration in regions where there were lots of lenses (i.e., for high classifier scores), while averaging out the calibration where lenses were sparse (i.e., for low classifier scores).

I demonstrate these methods for calibration via a toy model shown in Figure 3.1. For this, 10^4 model systems were generated, with classifier scores (x) assigned uniformly in the range $[0, 1]$ and a class (lens or not) which was allocated randomly with probability P_{true} given by a quartic function of its score, $P_{true}(x)$. This function, which is shown in Figure 3.1a, represents perfect calibration as for each score it maps to the true probability a given system is a lens. The calibration curves for this toy data using the methods described above are shown in the remaining panels

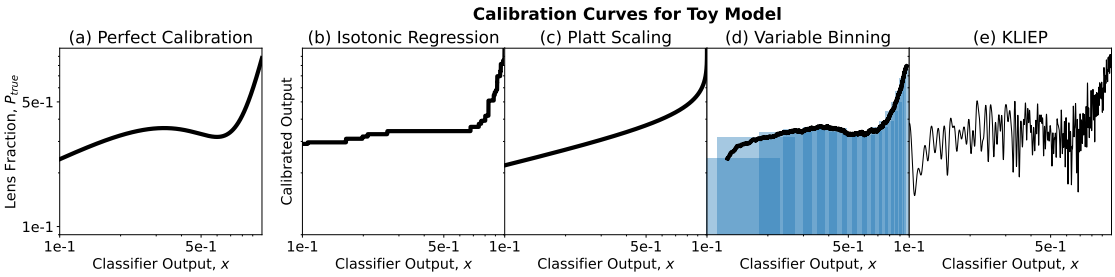


Figure 3.1: Calibration curves applied to a toy model for the range of calibration methods considered. The perfect calibration curve is shown on the far left. A subset of the bins used to generate the calibration curve in (d) are shown; the bin widths become smaller for larger classifier outputs as the fraction of lenses increases.

of this same figure. Platt scaling is the worst performing as the class distribution does not follow a sigmoid shape and thus is not considered further in this work.

3.3.2 Application of Calibration Methods

I applied the remaining calibration methods to the data. A key feature of the data was the significant class imbalance: only $\sim 0.2\%$ of all objects used in this work were lenses and these are more concentrated at high classifier scores, the latter allowing higher-resolution calibration for high scoring subjects.

For each calibration method I applied the calibration algorithm to the rank values of the classifier scores rather than the scores themselves. The ranks were defined as the ordered positions of the systems from lowest to highest raw classifier score for a given classifier; they did not depend on expert grade. I then interpolated back from rank to classifier score to produce the calibration map. To be widely applicable across classifiers, I aimed for the calibration method to be independent of the type of classifier used and the distribution of its scores. The distribution of scores can change significantly between different classifiers acting on the same objects; however, an excess of high (raw) scoring objects from a particular classifier should not be treated as all of these objects having high calibrated probabilities. Using the ranks removed this effect and simply assumes that a higher classifier score (qualitatively) implies higher confidence of a lens.

The KLIEP and variable binning methods had hyperparameters which I fixed prior to calibration. In the case of the KLIEP method, I used a Gaussian kernel of

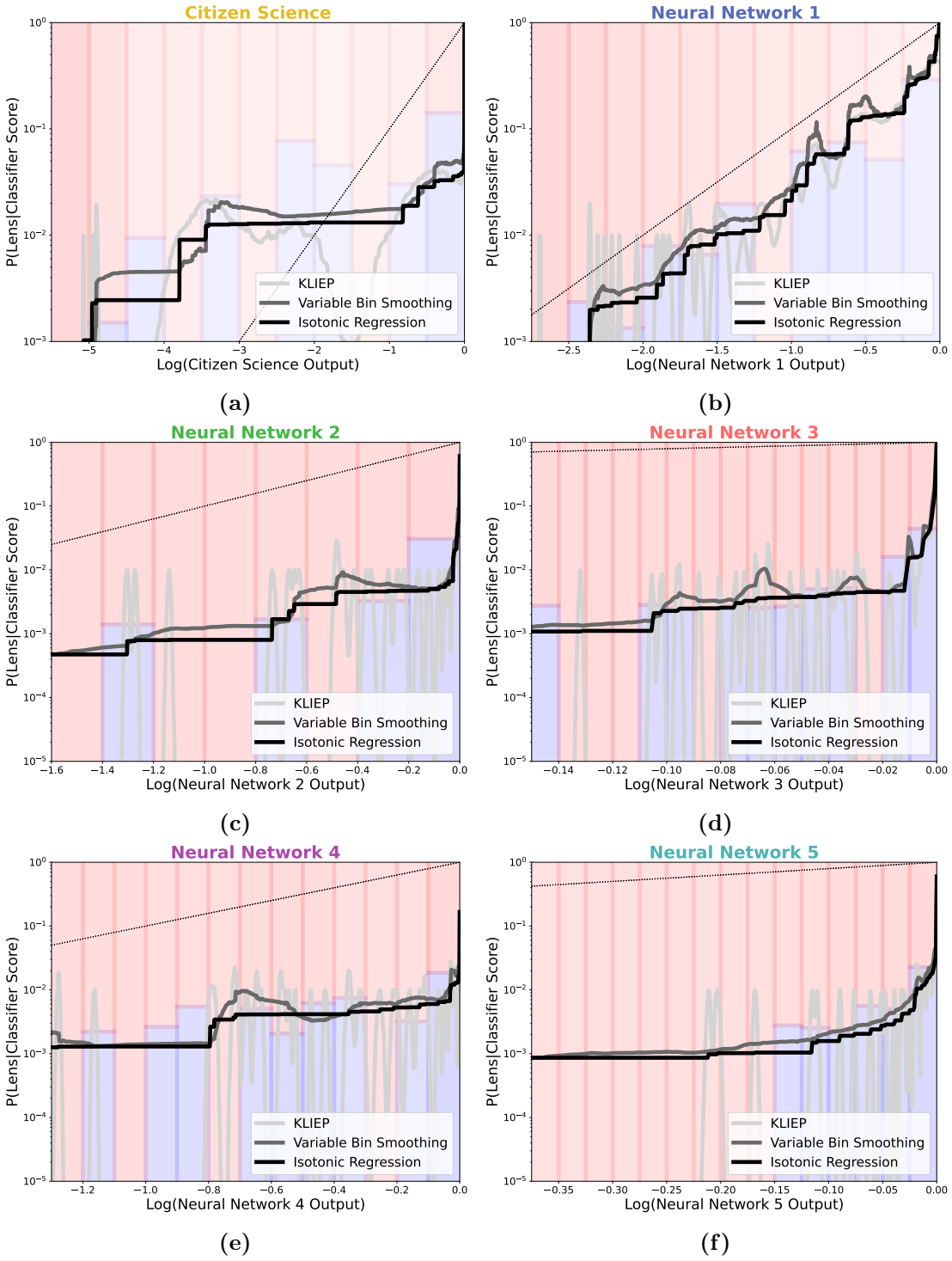


Figure 3.2: Calibration mapping for a range of calibration methods for each of the classifiers used in this work. For clarity, only the top 20% of classifier scores are shown in each panel. The red and blue histograms refer to the fraction of lenses (blue) and non-lenses (red) in each bin. The mappings should be expected to track these histograms but not always, for example, at the highest classifier scores where the fraction of true lenses increases rapidly within a single bin. The bars are shaded by the number of objects in each bin. The dashed line is the $y = x$ line, which the calibration curves would follow if the outputs of the classifiers were already calibrated.

width $\sigma = 40$; a balance between overfitting to the training data (occurring from a smaller kernel) and preventing over-smoothing at high scores (from a larger kernel). The KLIEP GMM had the same number of kernels as lenses. Figure 3.2 shows the calibration curves for each of the methods applied. As shown, the fraction of objects which are lenses increases significantly for high scores: using the rank values permitted finer tuning to these regions while not overfitting at lower classifier scores. In all cases the curves steepen for high scores: choice of score threshold will have a significant impact on purity in this region. The calibration curves do not follow the $y = x$ line (dotted), indicating that (as expected) the original classifier scores were not already calibrated. The mappings from the variable bin and isotonic regression method are relatively similar to each other across all classifiers. The KLIEP method differs more significantly as the Probability Density Function (PDF) only becomes non-zero close to the positions of the lenses in the training set. This effect would be reduced by increasing the kernel size, but would lead to underfitting at the highest classifier scores. Since I was primarily focussed on high-probability candidates, I prioritised reducing this underfitting. However, this demonstrates a limitation of this method, in that it can easily overfit to the calibration data in the regions in which the number of lenses is sparse (i.e., at low classifier rankings). The alternative methods in Figure 3.2 provide greater smoothing in such regions and so suffer less from this overfitting. An alternative, for example replacing the GMM distribution with a spline fit, still minimising the KL divergence (as in the KLIEP method), could reduce this overfitting by having fewer free parameters. This in turn would reduce the amount of data required for the calibration set.

I validated these calibration curves against a separate test set, shown in Figure 3.3. These show that spanning a wide range on $\log p$, the calibration mapping can produce accurate probabilities. The isotonic calibration produced the smallest variation upon bootstrapping, so is used for the subsequent analysis.

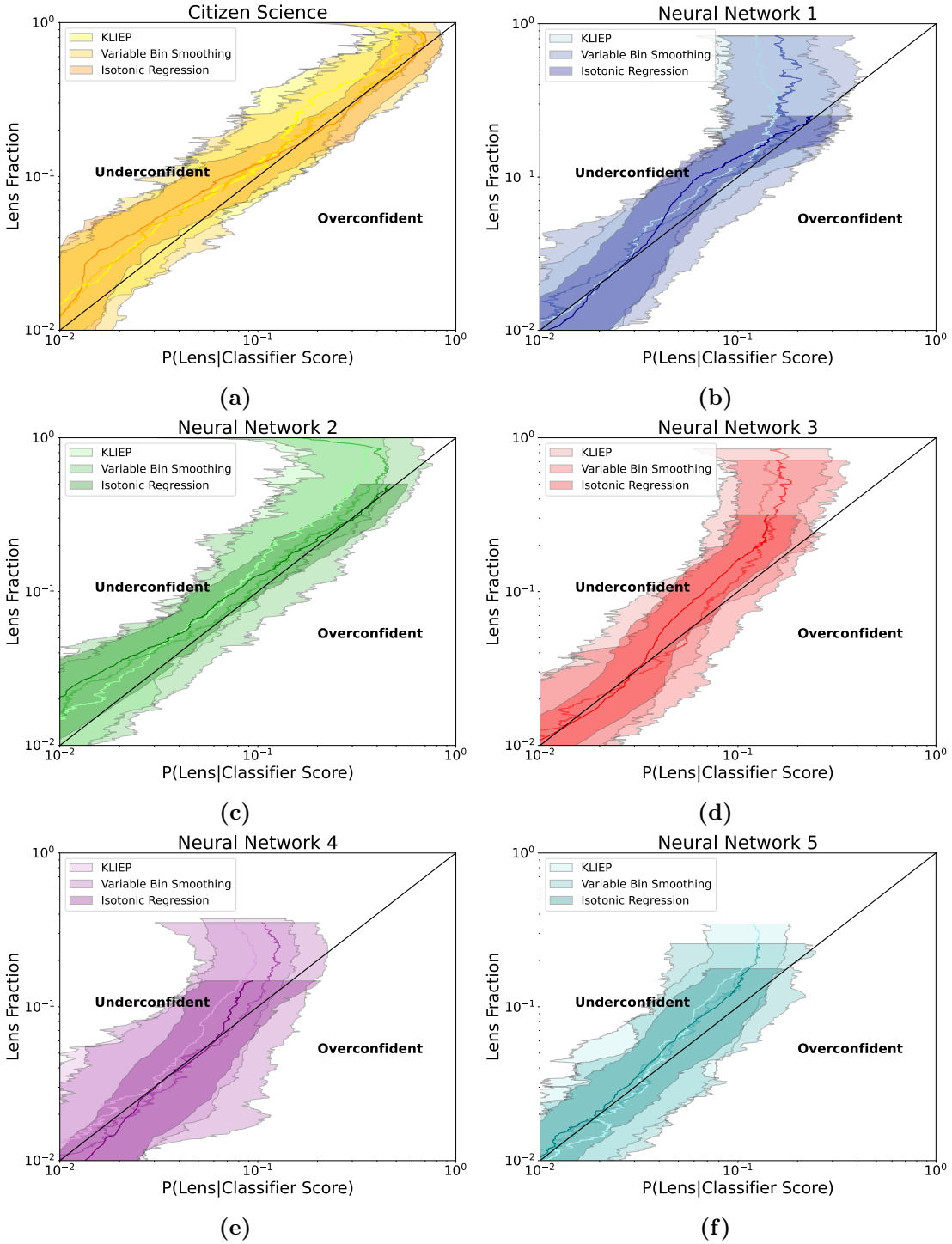


Figure 3.3: Validation of the calibration curves, applied to a separate test set of graded images. The y–axis is determined using the variable-binning method, and the error-bars are determined from bootstrapping. Each method shows good calibration (perfect calibration would be along the $y = x$ line), but are occasionally underconfident particularly for the highest scores.

3.3.3 Summary of Combination Methods

Given calibrated classifier scores, I considered methods to combine them to produce a single score. The combined score will not necessarily be calibrated, thus a further calibration stage using one of the methods above may be required. The aim of this classifier combination was to maximise the purity of the resultant sample. I tested 3 different methods: a generalised mean, dependent Bayesian combination and independent Bayesian combination. The latter two methods produce a posterior probability that a given object is a lens, given the calibrated scores of the individual classifiers. The former is a simple *ad hoc* method for combining multiple scores but does not strictly produce a posterior probability.

Generalised Mean

I first considered a simple generalised mean of the form:

$$P_{combo} = \left(\frac{1}{N} \sum_{i=1}^{N} p_i^{\alpha} \right)^{\frac{1}{\alpha}} \tag{3.2}$$

where N denotes the total number of classifiers in the ensemble, and p_i is the calibrated probability for a particular object from the ith classifier. This takes on a variety of useful functions for different values of α, in particular the arithmetic mean, harmonic mean, minimum and maximum of p_i across the N classifiers for $\alpha = 1, -1, -\infty$ and $+\infty$ respectively.

Dependent Bayesian Classifier Combination

Although I have generated calibrated probabilities for each classifier, the dependence (agreement/correlation) between each classifier on another has not yet been quantified. For example if two identical classifiers both gave probabilities of 0.9 for the same object, the combined probability should be 0.9 (the second, identical, classifier adds no new information), but if those classifiers were independent, one would expect the posterior to be > 0.9. I here outline a Bayesian approach to model this dependence as follows. I calculate the posterior $P(L|\{R_i\})$, where

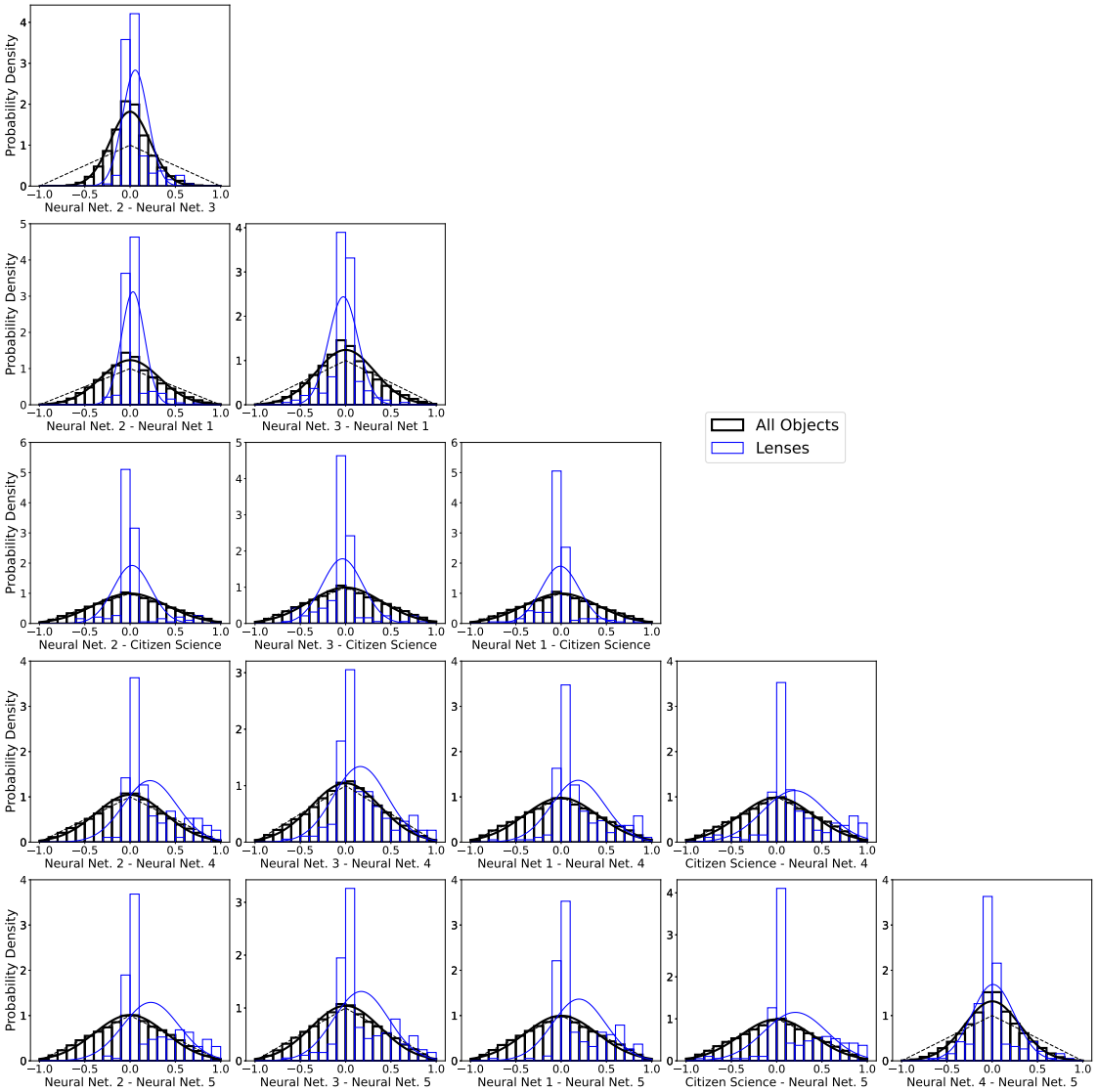


Figure 3.4: Distributions of the difference in ranking of the classifier scores for the cross-matched objects. The ranks were first normalised to take values in the range [0,1]. The black and blue histograms show the overall and lens distributions respectively. The respective curves show the projections of the best-fit multivariate gaussian to the data. The dashed curve shows a triangle distribution (which has not been fitted to the data) which results from taking the difference between two uniformly distributed variables.

$\{R_i\} \equiv \{R_1, ...R_N\}$ is the set of score rankings for a given object from each classifier and the ground truth is labelled L and $\bar{L}$ for 'lens' and 'non-lens' respectively. I denote the corresponding set of calibrated classifier scores as $\{C_i\}$. From Bayes' Theorem one can determine this posterior probability that a given system is a lens, as a function of the score rankings from all N classifiers:

$$P(L|\{R_i\}) = \frac{f(\{R_2, ...R_N\}|L, R_1)P(L|R_1)}{f(\{R_2, ...R_N\}|R_1)} \tag{3.3}$$

where f denotes a multi-dimensional probability density function, P denotes the probability of a discrete random variable and N denotes the total number of classifiers in the ensemble. $P(L|R_1)$ is known already, given a 1-1 relation between rank and calibrated score (i.e., the calibration mapping is strictly monotonic), as it is given by the calibrated score: $P(L|R_1) = C_1$. The choice of which classifier is denoted as C_1 is free and, as I will subsequently show, does not affect the result. I re-write Equation 3.3, in terms of the difference between the classifier scores, with respect to classifier C_1: $\{\Delta_i\} = \{R_i\} - R_1$:

$$P(L|\{R_i\}) = \frac{f(\{\Delta_2, ...\Delta_N\}|L, R_1)}{f(\{\Delta_2, \ldots, \Delta_N\}|R_1)} \cdot C_1 \tag{3.4}$$

I now model both numerator and denominator PDFs as Gaussian distributions, constant with respect to R_1 (i.e., the mean and covariance parameters in the Gaussian distributions in Eqn. 3.4 were fixed, and did not vary as a function of R_1). For independent classifiers, the denominator in Equation 3.4 tends towards a triangle distribution (see Figure 3.4); the distribution of $X - Y$, where X and Y are two random variables (i.e., the ranks) drawn from a uniform distribution, is a triangle distribution. If the classifiers are not completely independent the distribution will deviate from a triangle distribution; however, both are well approximated by a Gaussian. Therefore, I used score ranking, rather than the classifier score itself since this provides a better fit to this distribution but does not change the result. One now has:

$$P(L|\{R_i\}) = \frac{n(\{\Delta_2, ...\Delta_N\}|\boldsymbol{\mu}_{\mathbf{lens}}, \boldsymbol{\Sigma}_{\mathbf{lens}})}{n(\{\Delta_2, ...\Delta_N\}|\boldsymbol{\mu}_{\mathbf{full}}, \boldsymbol{\Sigma}_{\mathbf{full}})} \cdot C_1 \tag{3.5}$$

where n denotes the Gaussian function. For 6 classifiers, the multidimensional Gaussians in Eqn. 3.5 are 5 dimensional. These multivariate normal distributions are independent of the permutation of classifiers (i.e., which classifier is chosen to be C_1 does not change the best fitting Gaussian distribution, further detailed in Appendix A.1). However, this would change the value of C_1 (assuming all classifiers are not in exact agreement). With no reason, *a priori*, to favour one classifier over another, it is reasonable to average over the classifiers. Equation 3.5 becomes:

$$P(L|\{R_i\}) = \left(\frac{n(\{\Delta_2, \Delta_3, \Delta_4, \Delta_5, \Delta_6\}|\boldsymbol{\mu}_{\mathbf{lens}}, \boldsymbol{\Sigma}_{\mathbf{lens}})}{n(\{\Delta_2, \Delta_3, \Delta_4, \Delta_5, \Delta_6\}|\boldsymbol{\mu}_{\mathbf{full}}, \boldsymbol{\Sigma}_{\mathbf{full}})} \right) \cdot \langle C_i \rangle \tag{3.6}$$

Note, this stems from our (strong) assumption that the ratio of PDFs in Equation 3.4 can be modelled as a ratio of two fixed multivariate Gaussians; more complex functions would not suffer this problem. More flexible models (2 and 3-component mixture models of multivariate Gaussian distributions) were also tested, but did not provide significant improvements in the resultant combined calibrated score. The cases where the bracketed term in Eqn. 3.6 exceeds 1 refer to when there is greater agreement between the individual classifiers than would be expected from the overall distribution (i.e., when the blue curve exceeds the black curve in Figure 3.4) so the subject is likely to belong to the lens class.

Independent Bayesian Classifier Combination

I also considered the case where the results from each classifier were entirely independent of each other, since there was little correlation observed between most of the classifiers. From Bayes' Theorem, for a single classifer:

$$P(L|C_1) = \frac{f(C_1|L) \cdot P(L)}{f(C_1)} = \frac{f(C_1|L) \cdot P_0}{f(C_1|L) \cdot P_0 + f(C_1|\bar{L}) \cdot (1 - P_0)} \tag{3.7}$$

For a set of calibrated probabilities, $\{C_i\}$, for a given object:

$$P(L|\{C_i\}) = \frac{P_0 \cdot \prod_{i=1}^{N} f(C_i|L)}{P_0 \cdot \prod_{i=1}^{N} f(C_i|L) + (1 - P_0) \cdot \prod_{i=1}^{N} f(C_i|\bar{L})}. \tag{3.8}$$

Since $\{C_i\}$ are in fact calibrated probabilities, we know:

$$C_i = \frac{N_{\mathrm{L}} \cdot f(C_i|L)}{(N_{\mathrm{L}} + N_{\mathrm{NL}}) \cdot f(C_i)} \tag{3.9}$$

where $N_{\rm L}$ and $N_{\rm NL}$ refer to the number of lenses and non-lenses in the (training) sample. One can therefore simplify Eqn. 3.8:

$$P(L|\{C_i\}) = \frac{P_0 \cdot \prod_{i=1}^{N} \frac{C_i}{N_{\rm L}}}{P_0 \cdot \prod_{i=1}^{N} \frac{C_i}{N_{\rm L}} + (1 - P_0) \cdot \prod_{i=1}^{N} \frac{(1-C_i)}{N_{\rm NL}}} \tag{3.10}$$

For an accurate prior: $P_0 = N_{\rm L}/(N_{\rm L} + N_{\rm NL})$:

$$P(L|\{C_i\}) = \frac{N_{\rm L}^{1-N} \cdot \prod_{i=1}^{N} C_i}{\left(N_{\rm L}^{1-N} \cdot \prod_{i=1}^{N} C_i\right) + \left(N_{\rm NL}^{1-N} \cdot \prod_{i=1}^{N} (1 - C_i)\right)} \tag{3.11}$$

where N denotes the number of classifiers in the ensemble.

3.4 Results

3.4.1 Testing the Bayesian Combination Approaches

I generated a toy model using simulated classifier scores to test the differences between the dependent and independent combination methods described above. A set of 6 'independent' classifiers was generated as follows ('Method A'):

a) A simulated sample of classifier scores was drawn from the distribution $y = 2x$ (for $x \in [0, 1]$). These scores were assigned as belonging to true lenses.

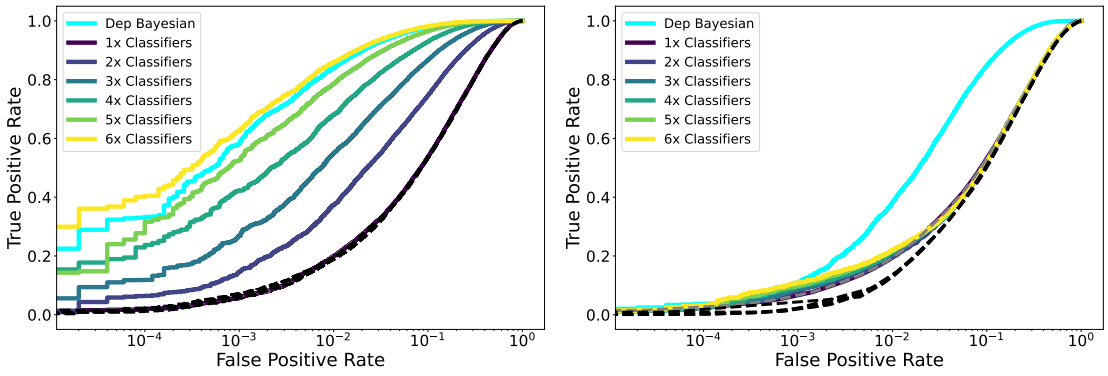
b) A second, equally sized sample of scores was drawn from a $y = 2(1 - x)$ distribution, and assigned as non-lenses. By construction, the combined sample of scores was calibrated as I used an equal number of lenses and non-lenses.

c) This was repeated for N classifiers ($N = 6$ in this case).

Sets of 6 'dependent' classifiers were generated as follows ('Method B'):

a) A set of scores from distributions $y = 2x$ and $y = 2(1 - x)$ was drawn, as in stages a and b above. This was defined as Classifier 1.

b) 5 further classifiers were generated by adding random noise to the Classifier 1 scores. Varying the level of noise altered the degree of dependence (agreement) between classifiers.



(a) Independent Classifiers **(b)** Dependent Classifiers

Figure 3.5: ROC curves for combining (a) independent and (b) dependent classifiers. The curves in the left-hand panel were generated using different data to those in the right-hand panel (Method A versus Method B respectively, above). The ROC curve for an ensemble constructed via the Dependent Bayesian method is denoted by 'Dep Bayesian', while ensembles constructed via the Independent Bayesian method with increasing number of classifiers are denoted '$X\times$ Classifiers'. The ROC curves for the individual classifiers are shown as black dashed lines. These ROC curves demonstrate that if the classifiers are independent (left-hand panel), then the Independent Bayesian method produces the best ensemble, while if classifiers give more correlated results (right-hand panel), the Dependent Bayesian method can provide better performance.

c) Each classifier was then calibrated via isotonic regression.

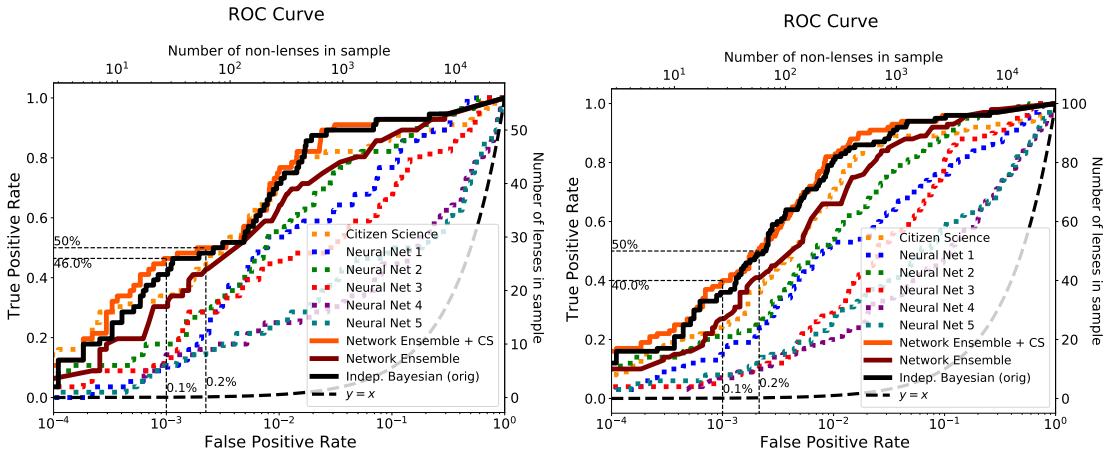
I then compared the ROC curves measured using the Dependent and Independent Bayesian methods, as shown in Figure 3.5. For independent classifiers (Figure 3.5a), the Independent Bayesian classifier combination provided the best ensemble (in Area Under ROC, AUROC, and completeness at a False Positive Rate, FPR, of 10^{-3}), while for dependent classifiers (Figure 3.5b), the best combination method (Dependent or Independent) depended on the degree of noise added in stage b) above. Therefore, when applied to real classifiers, it would be worthwhile applying both methods to verify which performs best.

3.4.2 Applying the Bayesian Combination Methods

I investigated the dependence of the six classifiers used in this analysis. Figure 3.4 shows the binned distributions of the difference in classifier ranking (with rankings normalised to 1), along with a multivariate Gaussian fit. The greatest correlation (for the whole cross-matched sample) is seen between the two HOLISMOKES VIII

neural network classifiers (Networks 2 and 3); this is perhaps expected as these networks have the same architecture (a ResNet developed by He et al., 2016b) and similar non-lens training data. The distributions (black) shown in Figure 3.4 which include the citizen science classifier (i.e., 'Neural Net. X - Citizen Science') are near-triangular. This demonstrates these rankings are nearly uncorrelated and suggests the networks and citizen classifiers found different objects easier/more difficult to classify (otherwise, the same objects would have received similar rankings from each). While the near-triangle distributions in Figure 3.4 do not guarantee the 2D distributions of classifier rank are uniform, in practice I found that they closely resemble uniform distributions, depicted in Appendix A.2. There was much greater agreement between classifiers when presented with a true lens, than there was with non-lenses. It is this property which is used by the Dependent Bayesian classifier combination method described in Section 3.3.3. While the Gaussian model fit to the whole sample (including non-lenses, black) was a good fit to the data shown in 3.4, the fit to the lens distribution (blue) was poorer. A more flexible model may have provided an improved fit in this case, in particular, one which could fit the sharp central peak along with the flatter wings. The discovery of much larger lens samples in the future will allow more complex distributions to be tested against each other which could allow for further improved classifier combination methods.

Upon inspecting the ensemble probabilities for a small number of spectroscopically confirmed lenses, I observed that in some cases, while the citizen science classifier would correctly identify the lens, the scores from the networks would not be sufficiently high to map to high probabilities. Consequently, the networks could effectively 'outvote' the citizen science classifier in the Independent Bayesian ensemble. Therefore, I tested generating a network-only ensemble, recalibrating this, then further combining this ensemble with the citizen science classifier. I show the ROC curve for this method as 'Network Ensemble + CS' in Figure 3.6. This figure shows the ROC curves for the individual classifiers along with the best performing ensemble methods. These are plotted using two different ground-truths, the effect of which is discussed in Section 3.5.3. When considering the A-B lenses (Figure



(a) Including A-B grade lenses as 'true lenses', without excluding cluster-scale candidates.

(b) Including A-C grade lenses as 'true lenses' and excluding cluster-scale candidates.

Figure 3.6: Receiver operating characteristic (ROC) curve for the individual lens classifiers (dashed) and combined methods, applied to a separate test set to that which the calibration methods were tuned. Ungraded subjects were treated as non-lenses. The dashed guidelines show the 50% completeness and 0.1% false positive rate. The ROC curve for a random classifier is shown as the dashed curve ($y = x$). Note the y-axes are linear, while the x-axes are logarithmic.

3.6a), each ensemble method shows improved classification over their individual constituent classifiers. I find 46% completeness can be achieved with a false positive rate of 10^{-3}; by comparison, the best individual classifier achieved 34% completeness on the same dataset. How this relates to the sample purity is discussed in Section 3.5.4. In Figure 3.7, I show the ROC curves for an ensemble classifier combining the classifiers using the methods described in Section 3.3.3. The ensemble performances across classifier-combination methods are broadly similar (with the exception of combination via harmonic mean, $\alpha = -1$) though in the low false-positive region (i.e., the regime of interest for strong lensing searches), the Independent Bayesian methods provide the greatest completeness. There is little difference in ensemble classification performance with/without first recalibrating a network ensemble before combining with the citizen classifier (also seen in Figure 3.6b).

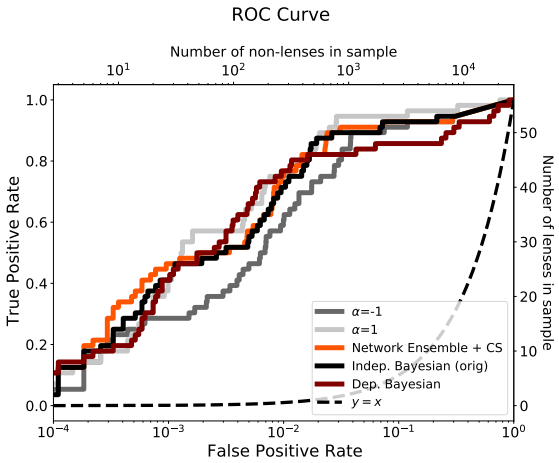


Figure 3.7: ROC Curves for a selection of different classifier-combination methods. The best performing methods at the low false-positive end were the Independent Bayesian methods (with/without generating a network-only ensemble first), which are used in the rest of the paper. The ROC curve for a random classifier is shown as the dashed curve ($y = x$). The α labels refer to the generalised mean combination from Section 3.3.3.

3.5 Discussion

3.5.1 Comparison with Previous Work

The overlap in lenses found by machine learning, citizen science and spectroscopy was investigated by Knabel et al. (2020) using Galaxy And Mass Assembly (GAMA) and KiDS data. They found very little overlap between the 3 methods: out of 107 lenses identified, only two were identified by more than one method (ML + citizen science). They attributed these to differences in the parent sample (e.g., different redshift cuts) and particular behaviours of each method (e.g., ML typically finding lenses similar to its training set). My results are derived from a different parent sample (and with different training sets) but demonstrate that citizen science and machine learning can align to a greater extent. However, there are two significant differences between our methods. Firstly, Knabel et al. (2020) employed Galaxy Zoo (Lintott et al., 2008; Holwerda et al., 2019; Kelvin et al. in prep.) as their citizen science classifier, which uses a question tree to identify the overall galaxy morphology, including the presence of lensing whereas Sonnenfeld et al. (2020) only looked for strong lenses. Secondly, in this work I only compare cross-matched objects which both techniques classified as opposed to objects simply in the same field, removing

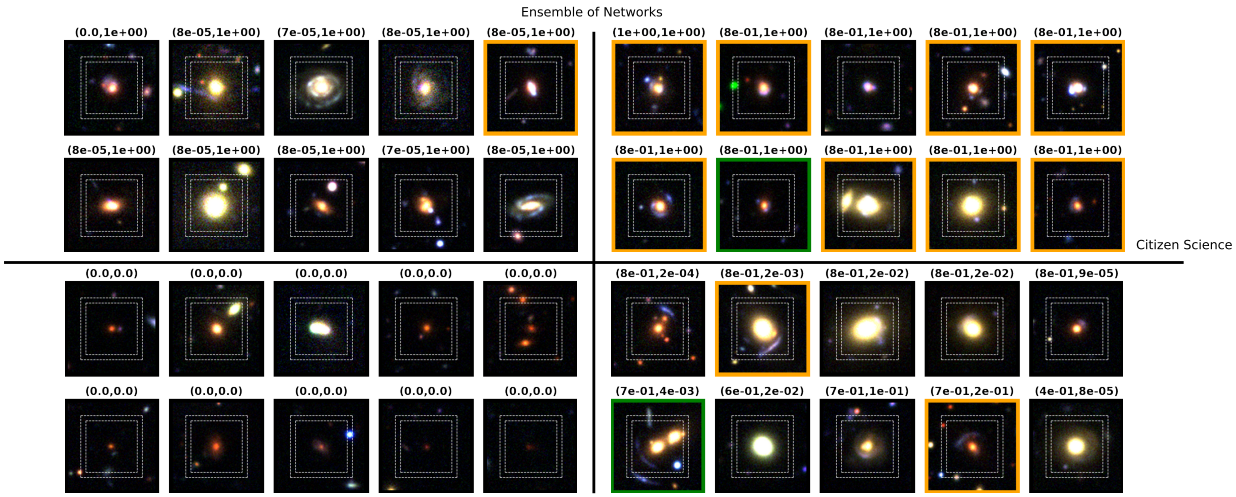


Figure 3.8: Example cutouts of high and low scoring objects from the citizen science search and the ensemble of 5 neural network classifiers. Objects with expert grades $1.5 < G < 2$ are outlined in orange, and those with $2 \leq G$ are outlined in green. The outer and inner dashed lines show the respective cutout sizes shown to the networks in Cañameras et al. (2021) and Shu et al. (2022) respectively, while the whole cutout was shown to the citizen scientists. The (P_1, P_2) values above each cutout indicate the calibrated probabilities from the Citizen Science classifier and the calibrated posterior for the Network-only ensemble respectively.

the effect of differing selection functions for each sample. The difference in my results highlights the importance of object selection when conducting lens searches; a narrow selection could significantly reduce the number of lenses identified.

3.5.2 Comparison of Citizen Science versus a Network Ensemble

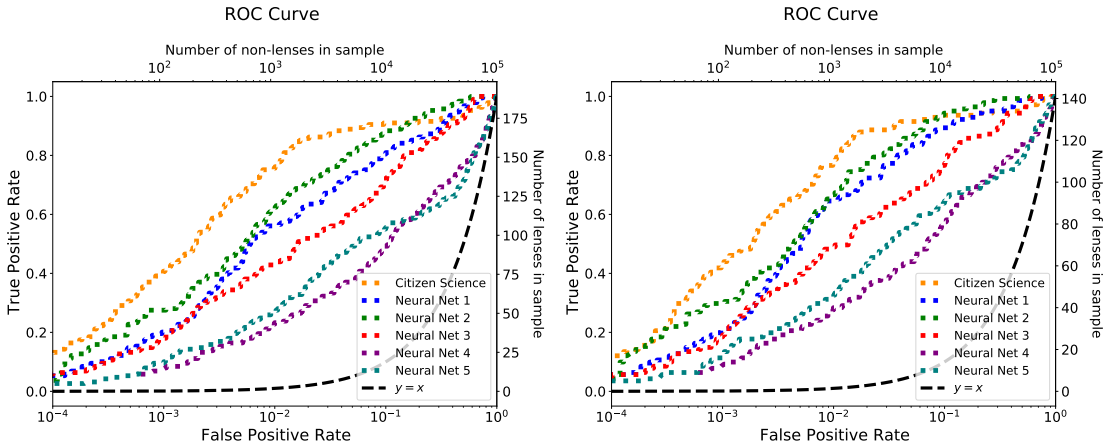
I investigated the qualitative properties of the galaxy systems identified and rejected by the citizen science classifier compared to those receiving high/low scores from an ensemble of 5 neural networks. Figure 3.8 shows a selection of such cutouts. The upper and lower quadrants show subjects which received a high and low posterior probability from the Network-only ensemble respectively. The right and left-hand quadrants show subjects which received high and low calibrated probabilities from the citizen science search respectively. The vast majority of those which received very high (low) probabilities from both the ensemble and citizen science were correctly identified as (non) lenses. Those which received high citizen scores but low ensemble scores contained a mix of true lenses and interlopers (according to the expert grades

used). There is no clear trend in these objects, but the presence of many bright bulges suggests that the network classifiers may have learnt to reject these; it should be noted that the networks did not have access to lens-subtracted images (unlike the citizen scientists) which would make some images, for example those with bright bulges but small Einstein radii, harder for the networks to classify. Furthermore, a small number of candidates were groups or had lensing features outside of the cutouts provided to the network, so it is unsurprising the networks rejected these - I demonstrate the effect of this in Section 3.5.3. The top-left quadrant of Figure 3.8 shows a selection of objects which received high ensemble probabilities but low citizen science probabilities. These contained a number of face-on spiral galaxies where some networks may have misidentified the spiral arms as lensed arcs.

3.5.3 Effect of Ground-Truth Selection on Classifier Performance

As demonstrated in Section 3.5.2, some of the objects to which the network ensemble assigned a low probability were galaxy clusters, which would not have been included in their training sets. However, these would have been identifiable in a citizen science search. I identified galaxy clusters which had been assigned previously as 'true lenses' as follows. I cross-matched the A-C grade lenses in the Masterlens database[4] (Moustakas, 2012) with my object sample and retrieved those with a 'CLUST-GAL' flag from the database. Since not all the objects in my sample were also in the database, I conducted a further visual inspection of the A-C graded candidates in my sample, flagging the cluster-scale lenses. Any objects identified by either method which had received a grade greater than the relevant cutoff ($\geq$ B grade, or $\geq$ C grade, as specified) were removed from the sample. These accounted for $\sim 25\%$ of the lens systems in my sample leaving 136 A-B and 373 A-C grade systems. I show the corresponding ROC curves for the individual classifiers in this work with/without the clusters removed in Figure 3.9. There is a small narrowing in the performance difference between the citizen science and

[4] https://test.masterlens.org/



(a) Including A-B grade lenses as 'true lenses', without excluding cluster-scale candidates.

(b) Including A-B grade lenses as 'true lenses', and excluding cluster-scale candidates.

Figure 3.9: Receiver operating characteristic (ROC) curve for the individual lens classifiers, applied to the whole cross-matched dataset. The ROC curve for a random classifier is shown as the dashed curve ($y = x$). Note the y-axes are linear, while the x-axes are logarithmic.

neural network classifiers when the cluster-scale lenses are excluded; however, a combination of the classification method and use of lens subtraction means the citizen science classifier still outperforms the networks for my object sample.

Since removing cluster-scale lenses reduced the ground-truth sample size, I also investigated the inclusion of C-grade lenses as 'true lenses' in the ensemble. In reality, while a sizeable fraction may not be lenses, this mimics the effect of increasing the sample size of lenses for the calibration, as may be available for a wider survey. This improved the calibration of the classifiers. Validation plots for this set-up are shown in Figure 3.10 showing that the calibration can accurately reach higher probabilities; having a greater number of 'true lenses' enables the calibration to be more accurate. I found that this improved the performance of the classifier ensemble compared to the individual classifiers. Figure 3.6b shows the ROC curves using A-C grade lenses as true lenses and with cluster-scale candidates removed. The Network-only ensemble provided substantial improvement over the best individual network and performs better (in AUROC, FPR at 50% completeness and True Positive Rate, TPR, at FPR $= 10^{-3}$) than the individual citizen science classifier. When citizen science was included, the ensemble improved further, increasing the

completeness from 27% (Network-only) to 40% at FPR = 10^{-3}.

3.5.4 Expectations and Implications for LSST

In Figure 3.11, I compare the number of true and false positives expected for an equivalent ground-based survey of LSST scale. In these plots, I have scaled the total number of strong lenses to 10^5, and used the true and false positive rate functions stemming from the ensemble and individual classifiers applied to the HSC data in this work. I find that, when including citizen science in the sample, a 40% complete sample can be achieved with 49% purity, compared to 32% purity for the best individual classifier. However, for higher completeness (towards the left of each plot), the sample would remain overwhelmingly false positives. The benefit of an ensemble is more substantial when the citizen science classifier is excluded (i.e., a network-only ensemble) nearly doubling the purity to 28% for a 40% complete sample. I discuss possible improvements to lens classification in Chapter 4 and discuss the implications and mitigation of this incompleteness for population-level analysis in Chapter 5.

For such large samples, expert grading of all but the highest ranked candidates becomes intractable. Thus, it becomes even more important that the scores of lens-finders are calibrated to allow statistical analysis of large samples of strong lenses including a known proportion of false positives. Figure 3.12 shows the effective sample size of both the individual calibrated classifiers and the ensemble. It demonstrates the ensemble can retrieve a larger effective sample of lenses from the data, as well as a clear plateau, the 'knee' of which would be a useful starting point for a statistical sample of uncertain lenses. Calibrated probabilities also allow the comparison of objects inspected by different classifiers. They permit the rank-ordering across different samples of objects, not necessarily all seen by the same classifier, which could be used for identifying candidates for follow-up and may be useful for selecting the forthcoming 4MOST sample (Collett et al., 2023).

A key component of the methodology described above is the presence of a ground-truth from which to base the calibration. In this work, I compiled expert

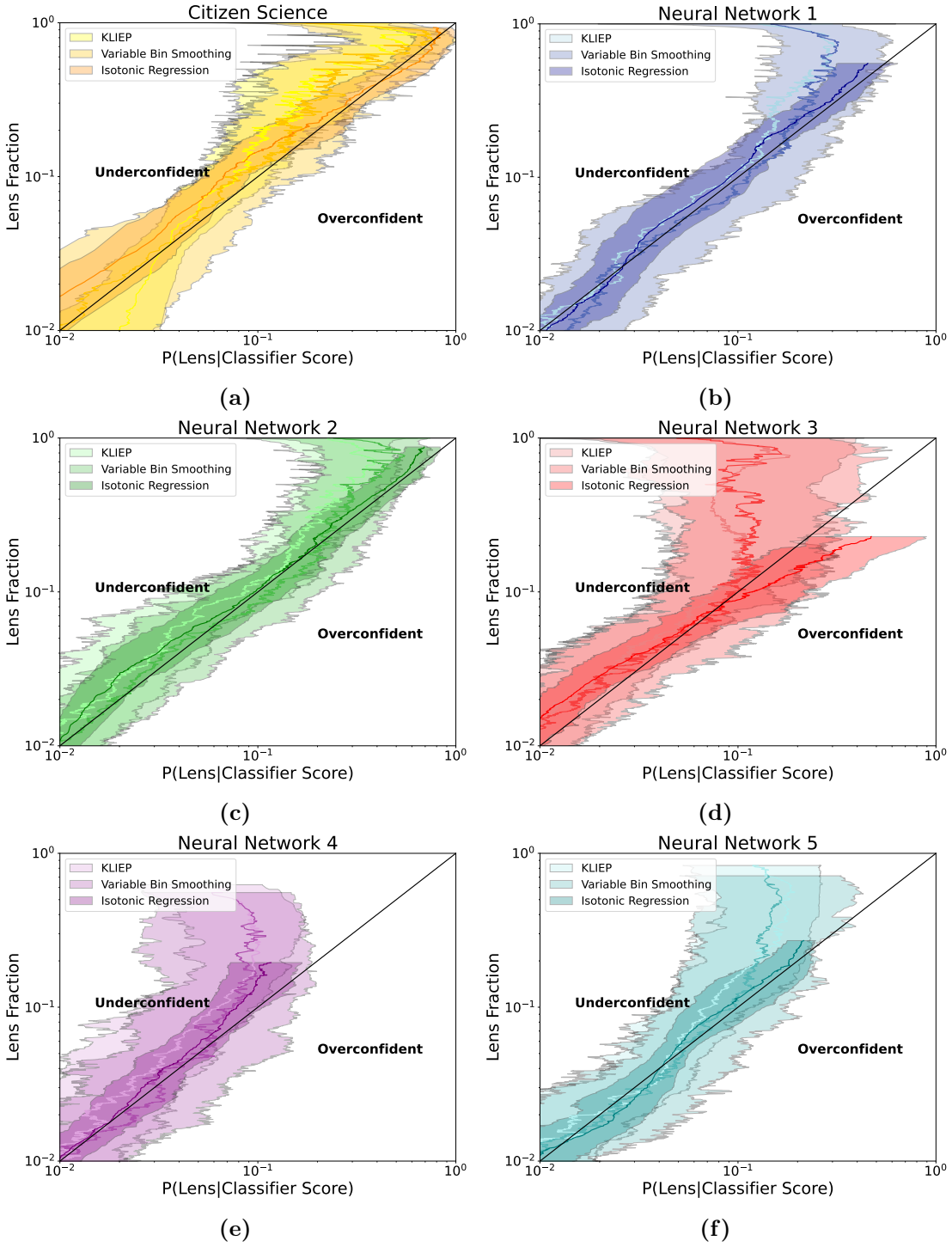
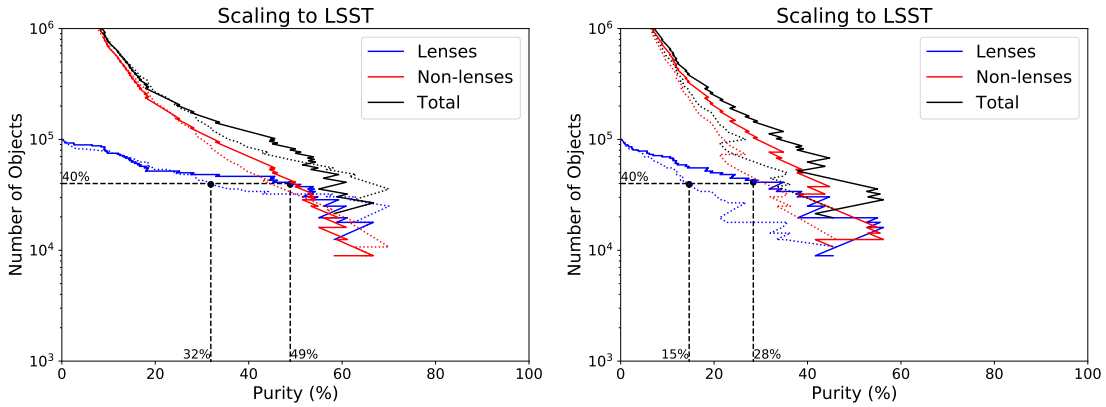


Figure 3.10: Validation of the calibration curves, applied to a separate test set of graded images. A-C grade lenses are counted as 'true lenses' for these plots, and cluster-scale candidates receiving A-C grades have been excluded.



(a) An ensemble of 6 classifiers (solid) compared to only the citizen science classifier (dashed).

(b) An ensemble of the 5 neural networks (solid) compared to a single network (Neural Network 2, dashed).

Figure 3.11: A plot of the expected number of true and false positives for LSST; the total number of lenses has been fixed to 10^5. These plots use the false-positive and true-positive rates as a function of the score threshold for the HSC classifiers used in this work (with A-B grade lenses as 'true' lenses), assuming the classifier performance in LSST matches those applied to HSC here. The curves have been cutoff when there were < 5 lenses or non-lenses remaining in the test set, to reduce the effect of small-number statistics.

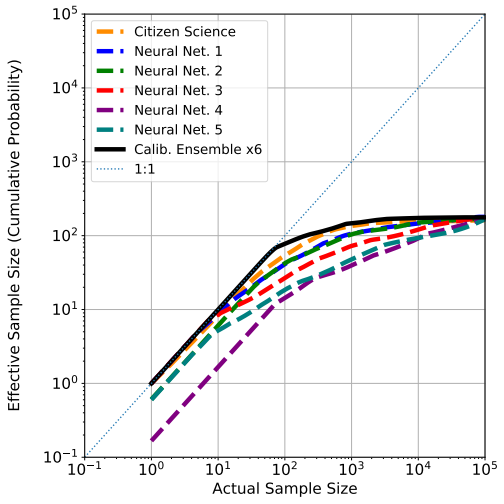


Figure 3.12: A plot of the effective sample size of the calibrated individual and ensemble classifiers presented above, as a function of total sample size. The effective sample size was determined by rank-ordering the objects in order of calibrated probability, then cumulatively summing them. The ensemble classifier used here is comprised of all 6 classifiers, and has been further recalibrated by isotonic regression after forming the ensemble. The ensemble classifier can produce the largest effective sample size out of all the classifiers.

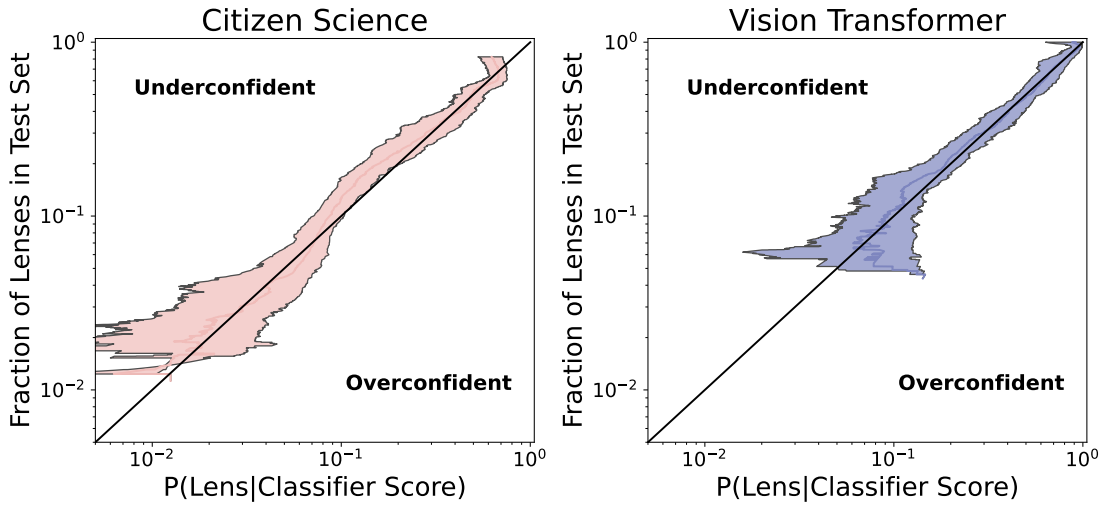


Figure 3.13: Validation of the calibration of the Vision Transformer and citizen science classifiers described in Section 3.5.5. Both classifiers can be calibrated up to high purity values, indicating high performance for both. The uncertainties have been determined via bootstrapping. Reproduced from González et al., 2025.

grades from a range of sources in order to maximise the number of objects with assigned grades. While it may appear initially that having a known ground-truth for a random sample of objects would be optimal for calibration, in practice the high-score regime is where calibration is most important and the calibration mapping changes the most rapidly. Citizen science could be used here to provide high quality lens candidates to use as a ground truth for network calibration. For LSST, this ground-truth could then be used for calibrating automated methods such as neural networks across the whole survey area.

3.5.5 Application to the Dark Energy Survey

So far, the results in this chapter have been drawn from combining the classifiers from multiple separate lens searches in HSC data. More recent lens searches have provided an opportunity to apply this methodology to lens searches which used multiple lens classifiers from the outset. The first combined machine learning+ citizen science search for strong lenses was undertaken recently using data from DES and is described in González et al. (2025). This was a two-stage search, first applying a VT to 2.36×10^8 DES cutouts, before a Space Warps citizen science search was

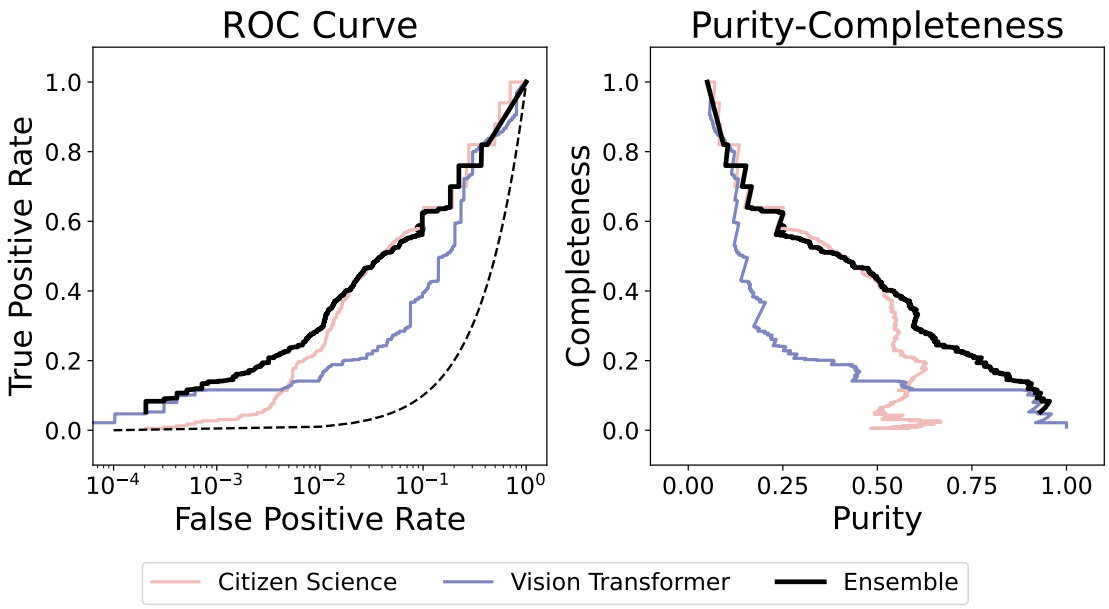


Figure 3.14: ROC and purity-completeness curves for the individual and ensemble classifiers, applied to the test dataset. The ensemble provides similar or improved classification over the individual classifiers across the purity range. Reproduced from González et al., 2025.

conducted to inspect the highest scoring 22 564 systems. Given the multi-classifier nature of this search, this was an excellent opportunity to test the calibration and ensemble methodology on an additional ground-based lens search. This is the first time that a calibrated ensemble method has been applied to a multi-classifier lens search. A more extensive lens search utilising multiple lens classifiers was conducted using data from the *Euclid* survey, and is the subject of the next chapter.

To generate an ensemble of the two DES classifiers (the VT and Space Warps) I first split the dataset of 22 564 high scoring VT systems into two equally sized calibration and test sets and then applied the isotonic regression calibration method. I classed systems receiving an expert grade $G \geq 1.25$ as a true lens, and the remaining systems as non-lenses. This calibration is validated in Figure 3.13, showing good calibration for both classifiers. I then combined these calibrated classifiers under the assumption of independence, via Eqn. 3.11. The resulting ROC and purity-completeness curves are shown in Figure 3.14. The citizen science classifier provided the highest completeness for purity values $\sim 20 - 50\%$, while the VT provided a more complete sample in the highest purity regime. However,

the ensemble classifier provides as good or improved classification over the best classifier for a given FPR (or purity). Through k-fold resampling, I also generated ensemble scores for the complete dataset inspected by citizens, thus generating a ranked list of candidates in order of lens probability. Such a list could be used in the future for prioritising time-consuming lens modelling or follow-up.

Overall the performance of the DES ensemble provides good supporting evidence that the calibration and ensemble methods described in this chapter are robust to different classifiers and lens searches. Furthermore, since the ensemble can provide improvement even if the number of classifiers is small it is a valuable tool for identifying/prioritising high-quality candidates with reduced human input.

3.6 Conclusion

This chapter had two aims: 1) to provide calibrated probabilities for a sample of galaxy cutouts that a given galaxy is a strong lens system and 2) to combine neural network and citizen science classifiers into an ensemble classifier, to maximise the purity of the resulting sample, without compromising on completeness. Initially, I used 6 classifiers (1 citizen science search and 5 neural networks) previously applied to HSC data, chosen as a proxy for the forthcoming LSST, before extending this to classifiers for DES data. Having achieved these aims, my conclusions are as follows:

1. It is possible to calibrate the scores of a given lens classifier to produce accurate probabilities that a given classified object is a lens. It is not necessary to tune a classifier to produce probabilities from the outset (i.e., during training), as this can be done through post-processing the outputs, so the methods applied here are flexible to classifier type. The original scores of typical lens classifiers are not *a priori* calibrated probabilities, and should not be treated as such in a statistical analysis.

2. There was very little correlation between HSC classifiers of the scores of non-lenses. Combining classifiers into an ensemble can take advantage of this since different classifiers find certain systems easier/more difficult to classify.

The lens sample produced by a given search is significantly affected by the original object selection; too narrow a selection can significantly reduce the number of lenses identified in a search.

3. An ensemble classifier can provide improved classification above its constituent components. For an FPR of 10^{-3}, the HSC ML + citizen science ensemble classifier increased the completeness of the resultant sample from 34% to 46%. In a lens search of LSST data, it is likely that many lens classifiers will be applied to the same dataset; combining these into an ensemble would produce a purer and more complete lens sample than could be achieved from a single classifier.

4. The calibration and ensemble methods described in this chapter are robust to different types and numbers of classifier, improving classification performance in both HSC and DES survey data. Therefore, they could be widely applied to future lens searches with minimal alteration, improving the resultant lens classification and providing a mechanism for combining lens classifiers regardless of training set, architecture or performance.

5. Given $\sim 10^5$ strong lenses in LSST, further improvements in scalable lens finding methods will be needed in order to achieve completenesses $> 50\%$ without significant contamination by false positives. Incorporating more classifiers into the ensemble would help with this, in particular if they are trained on different training sets, so they can recognise different features in the lensed images. The citizen science classifier (which notably used lens-subtracted images) had the best performance out of the HSC classifiers; which images the citizens are shown would impact on how many lenses are found and is discussed in the next chapter.

The work presented in this chapter, while applied to the field of strong lensing, is much more broadly applicable to classification tasks in general. The strong lens classifiers in this work could be replaced with any other quantitative classifier

without changing the methodology. The principles of calibrating the classifiers could, for example, be applied to Type 1a supernovae classification which can require such probabilities during cosmological inference (as will be discussed in Chapter 5). Similarly, the process of combining classifiers into an ensemble, accounting for the differing performances between classifiers could improve classification performance across a range of scientific fields, in particular where multiple different techniques are used for classification (e.g., neural networks+citizen science, or neural networks+k-nearest-neighbour classification), where the classifiers are likely to be relatively uncorrelated.

4

Lens Searches in Contemporary Wide Area Surveys

The basis of this chapter first appeared in 'Euclid Quick Data Release (Q1) – The Strong Lensing Discovery Engine E – Ensemble Classification of Strong Gravitational Lenses: Lessons for Data Release 1', Euclid Collaboration: Holloway et al., 2025 (submitted).

Contents

4.1 Introduction 109
4.2 Data 111
4.2.1 The *Euclid* Q1 Lens Search 111
4.2.2 The *Euclid* Q1 Lens Classifiers 113
4.3 Method 115
4.3.1 Calibration of Strong Lens Classifiers 115
4.3.2 Combination of Classifiers into an Ensemble 120
4.4 Results and Discussion 121
4.4.1 Ensemble Classifier Performance 121
4.4.2 Systems Identified by Citizens or Ensemble 125
4.4.3 Outlook for *Euclid* DR1 and Future Data Releases . . . 127
4.4.4 Optimising Lens Searches in Wide Area Surveys 130
4.5 Conclusions 133

Early data from the Euclid Wide Survey presents an opportunity to test the current performance of lens finders and to optimise the lens search strategy for wide-field surveys. In this chapter I discuss the results from the Euclid Quick Release 1 lens search and lessons for the first major Euclid data release (DR1), where the time investment of both experts and citizens must be managed carefully.

4.1 Introduction

The *Euclid* satellite (Euclid Collaboration: Mellier et al., 2024), which launched on 1 July 2023, aims to survey 14 000 deg^2 of the sky. Based on current estimates (Chapter 2, Collett, 2015), the EWS will identify a similar or larger number of strongly lensed systems to LSST, the main focus of the previous chapter. Compared to LSST, the EWS will provide higher-resolution imaging (0.16″ in $I_{\rm E}$, Euclid Collaboration: Cropper et al., 2024; Euclid Collaboration: McCracken et al., 2025) over a comparable area (LSST[1,2]: $r \simeq 26.9$ over $\sim 20\,000$ deg^2, *Euclid*: $I_{\rm E} \simeq 26.2$ over $\sim 14\,500$ deg^2, Euclid Collaboration: Scaramella et al., 2022). This will allow *Euclid* to probe smaller Einstein radius systems (i.e., lower mass lenses) while in the *Euclid*+LSST overlap region their combined photometric bands extending into the NIR will help improve the accuracy of photometric redshift/mass estimates. Given the greater resolution, one might expect the performance of lens classifiers on *Euclid* data to be significantly better (i.e., with higher purity and completeness) than those applied to LSST data. The *Euclid* Q1 data release (Euclid Quick Release Q1, 2025), comprising 63.1 deg^2 of imaging, provided an excellent dataset to test this hypothesis and to begin to tune the *Euclid* lens finding procedure in preparation for forthcoming larger data releases.

Following the release of the Q1 data, the *Euclid* Strong Lens Working Group conducted a systematic lens search. This search utilised multiple classifiers applied to

[1] https://pstn-054.lsst.io/
[2] https://survey-strategy.lsst.io/baseline/wfd.html

the Q1 dataset and thus provided a valuable opportunity to test an ensemble classifier at scale. In Chapter 3, the benefits of an ensemble classifier were investigated using data from HSC and DES, whereby multiple lens classifiers (both ML and citizen science) were combined together to provide a purer and more complete lens sample. In this chapter, I extend this analysis to *Euclid* data, and investigate the scalability of the current *Euclid* strong lens discovery pipeline to the full EWS. This method has the potential for broad application by the strong lens community. Given the number and varied nature of the lens classifiers used, it can be difficult to determine which candidates are the most likely lenses, since each classifier ranks each system differently. The ensemble methodology discussed in the previous chapter solves this by combining the results of the individual classifiers into a single score. Testing this on *Euclid* data lends confidence to the method before the arrival of forthcoming data releases in which ensemble classification may be used to determine which systems are shown to citizens or experts. The systems shown to citizens or experts can have a material effect on the number (and type) of lens candidates identified; therefore, such prioritisation requires careful consideration and testing. The work in this chapter aims to answer the following questions:

1. How does such an ensemble of strong lens classifiers perform when applied to high-resolution space-based data?

2. How does this classifier performance translate into the resulting purity and completeness, and based on this, how would this translate into the number of systems requiring inspection in forthcoming *Euclid* data releases?

3. Given the large number of strong lens candidates anticipated in Data Release 1 (DR1) and future releases, is it possible to use inspection by citizens in lieu of using strong lensing experts?

4. What will the best strategy be for future lens searches to make best use of the individual strengths of a diverse range of strong lens classifiers?

This chapter is structured as follows. In Section 4.2 I describe the Q1 strong lens search strategy, the individual classifiers applied to these data, and the expert-grading of lens candidates. In Section 4.3.1 I calibrate each of these classifiers, and combine them into an ensemble in Section 4.3.2. I discuss results in Section 4.4, including performance of a range of ensemble classifiers (Section 4.4.1), and the strengths and weaknesses of citizen and ensemble classifiers (Section 4.4.2). I then discuss the outlook for *Euclid* DR1 (Section 4.4.3) and possible improvements for lens finding in future searches (Section 4.4.4) and conclude in Section 4.5.

This work formed part of a series outlining the findings of a strong lens search in *Euclid* Q1 data. An overview of the search procedure and main results is given in Euclid Collaboration: Walmsley et al., 2025 (hereafter Paper A). Paper B (Euclid Collaboration: Rojas et al., 2025) detailed lens candidates identified after pre-selecting high-velocity galaxies from SDSS and DESI. The primary machine learning models used are detailed in Euclid Collaboration: Lines et al., 2025 (Paper C) while Euclid Collaboration: Li et al., 2025 (Paper D) describes the Q1 DSPL candidates.

4.2 Data

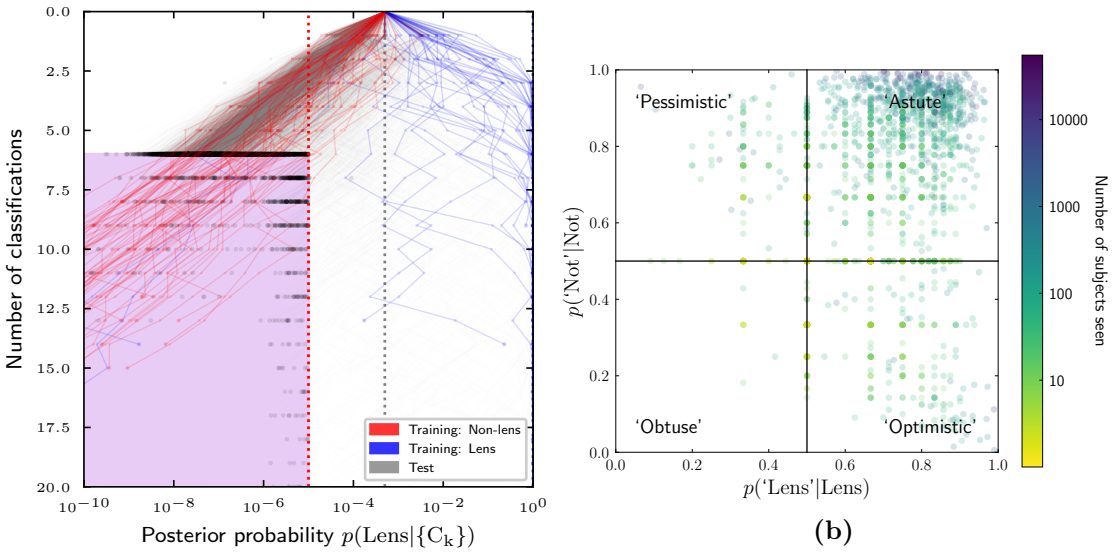
4.2.1 The *Euclid* Q1 Lens Search

To generate image cutouts of galaxies in the Q1 area, we selected $I_{\rm E}$-detected extended objects with $I_{\rm E} < 22.5$ which did not have *Gaia* counterparts from the Q1 multiwavelength MERged (MER) catalogue (Euclid Collaboration: Romelli et al., 2025). This selection produced 1.09×10^6 sources. Cutout images were then generated using the ESA Science Archive Service and the ESA Datalabs platform (ESA Datalabs, 2024). Given the volume of data the subsequent lens search took a staged approach. Firstly, multiple machine learning models were applied to the complete selection of cutouts. Then a large scale citizen science project was launched through Space Warps. Given the high performance of citizens in lens classification, having citizen scores for a large sample of systems was very useful in identifying lens candidates. Due to the large number of systems shown to the citizens (around 100 000), it was crucial that the citizens' classifications were used as efficiently as

possible. Therefore, the Space Warps Analysis Pipeline pipeline (SWAP, see Section 1.2.4) was run in real time, updating the citizen skill values and the system scores continuously. Furthermore, systems were 'retired' (removed from the platform) automatically once they received 30 classifications, or dropped below a lower score threshold indicating they were very unlikely to be a lens. This combination enabled the citizens' classifications to be used very effectively, ensuring they were not wasted on systems unlikely to be lenses. Given the rarity of lenses and the finite inspection budget of the citizens, such efficiency was essential. Plots of the evolution in system score and final user skill distribution are shown in Figure 4.1. The majority of users were assigned high skill values, indicating their adeptness at identifying lenses, and the vast majority of test subjects were rapidly retired (as expected given the rarity of strong lenses). The vast majority of simulated lenses received high scores from the citizens, demonstrating their skill in correctly identifying likely lens configurations.

Alongside the Space Warps search, we conducted an expert grading stage, termed Galaxy Judges (GJ). Here, strong lens researchers were presented with high-scoring systems from both the ML and citizen science classifiers. In particular the highest scoring 1000 systems from each of the networks, along with roughly 2700 high-scoring systems from Space Warps (with Space Warps score $p_{\rm SW} > 10^{-5}$) underwent expert inspection. These systems were given a grade A+, A, B, C or X grade, with A+ denoting a certain lens which was an interesting lens system in its own right, and A-C denoting increasingly less confident lens candidates. These grades were subsequently averaged to give a final grade. Finally, lens modelling of the high-grade lens candidates was conducted. This provided lens parameters and also acted as an additional verification step for lens candidates.

Overall, 7700 systems were inspected by GJ, identifying 250 grade A systems (confident lens) and 247 grade B systems, i.e., $\sim 7.9\,\mathrm{lenses}/\,\mathrm{deg}^2$. By comparison, a citizen science search conducted by Garvin et al. (2022) used archival HST data over roughly half the area ($27\,\mathrm{deg}^2$) of the *Euclid* Q1 data, identifying 167 A/B-grade candidates ($\sim 6.2\,\mathrm{lenses}/\,\mathrm{deg}^2$) while the HSC search of Sonnenfeld et al. (2020) identified 14+129 A+B grade candidates over $442\,\mathrm{deg}^2$ ($\sim 0.3\,\mathrm{lenses}/\,\mathrm{deg}^2$).



(a) (b)

Figure 4.1: a) Score trajectory plots of a random sample of training and test systems from the Space Warps lens search. Each system was assigned an initial prior probability of being a lens ($p_{\mathrm{SW}} = 5 \times 10^{-4}$) which was updated as the citizens classified. Simulated lenses used for training are shown in blue, non-lenses training subjects are shown in red, and test subjects are shown in grey. The systems were retired if they reached the region highlighted in purple. b) Distribution of user skills. These were calculated based on the proportion of correctly classified training subjects, split into lenses (x-axis) and non-lenses (y-axis). These skill values were updated in real time throughout the lens search. Citizens with greater skill could cause larger changes in the scores of the test and training subjects. Figures reproduced from Euclid Collaboration: Walmsley et al. (2025).

4.2.2 The *Euclid* Q1 Lens Classifiers

To generate an ensemble of *Euclid* classifiers, I used eight ML networks produced by five different teams applied to these sources, along with classifications for approximately 10% of these from the Space Warps search[3]. One network from each of these teams (Models 1, 2a, 3a, 4, and 5 listed below) was given priority when assigning which high-scoring systems to show to the citizens; I refer to these as the primary networks. In this work I also used 3 additional networks (Models 2b, 3b, and 3c described below) prepared by these teams.

Each classifier was shown 10″ cutouts. These cutouts were generated using one or both of two colour scalings: 'arcsinh' and 'MTF' (i.e., 'midtone transfer function', defined in Paper A), using different combinations of the *Euclid* band-passes. A

[3] https://www.zooniverse.org/projects/aprajita/space-warps-esa-euclid

summary of each of the classifiers is given below and presented in Table 4.1.

- Model 0: Space Warps. The Space Warps strong lens search involved more than 1000 citizens who made a total of 800 000 classifications. Unlike the machine learning classifiers, who were shown all the images in the data set, the citizens classified a subset of 115 000 cutouts, which were either high-scoring objects from the ML classifiers (80 000), or randomly drawn from the complete data set (40 000 including overlap). Each cutout was classified an average of 7.2 times by citizens; low scoring objects were removed from the platform after six classifications. The citizens were shown both arcsinh and MTF colour settings, using the $I_{\rm E}$, $Y_{\rm E}$, and $J_{\rm E}$ bands.

- Model 1: This network (adapted from Domínguez Sánchez et al., 2018; Manjón-García, 2021) was a 4-layer Convolutional Neural Network, trained using a range of simulated lens systems, including 64 grade A and B lens candidates from the Galaxy Zoo (Lintott et al., 2008) *Euclid* and Cosmic-Dawn projects. The network was applied to the $I_{\rm E}$-band-only data set, where the networks' outputs using the MTF and arcsinh colour settings were averaged to produce the final result.

- Model 2 (a,b): These OU-100 Convolutional Neural Networks (adapted from Wilde et al., 2022) were trained using (a) $I_{\rm E}$-band-only, and (b) $I_{\rm E}$ and $J_{\rm E}$ bands, respectively, with the MTF colour setting. The training set totalled 32 000 non-lenses and simulated lenses, of which 12% were lenses. In addition to simulated lens systems, the training set included around 200 grade A and B lens candidates identified in *Euclid* imaging. These real *Euclid* lens candidates were identified in searches of Early Release Observations (ERO, Acevedo Barroso et al., 2024), through inspection of galaxies with spectroscopic data (Paper B), through Galaxy Zoo Cosmic Dawn and Galaxy Zoo *Euclid* projects, as well as from serendipitous discoveries. Class weights were applied to ensure that non-lenses and lenses were weighted equally overall.

- Model 3 (a-c): These networks were adapted from Euclid Collaboration: Leuzzi et al. (2024) and had `IncNet` (a, Szegedy et al., 2015; Szegedy et al., 2016), `ResNet` (b, He et al., 2016a; Xie et al., 2017), and `VGG` (c, Simonyan and Zisserman, 2015) architectures respectively. They were trained using 40 000 $I_{\rm E}$ band-only images (non-lenses and simulated lenses) using the arcsinh colour setting.

- Model 4: `Zoobot` (see Paper C and Walmsley et al., 2023): This Bayesian Neural Network was pre-trained using 9.2×10^7 morphological classifications from the Galaxy Zoo project (Lintott et al., 2008), and subsequently fine-tuned using verified non-lenses in DESI (Euclid Collaboration: Rojas et al., 2025) and simulated lenses. This classifier was shown $I_{\rm E}$-band-only arcsinh cutouts.

- Model 5: This network (derived from Chen et al., 2020; Oquab et al., 2023; Smith et al., 2024) was pre-trained using self-supervised contrastive learning with a VT backbone using 14×14 patches on 80 000 simulated lens images, and 80 000 non-lens images, including ring galaxies, mergers, and spirals. The training images used the $I_{\rm E}$ band and arcsinh scaling.

In summary, the classifiers were diverse and adopted different training configurations. These were ideal models to combine into an ensemble.

.

4.3 Method

4.3.1 Calibration of Strong Lens Classifiers

To combine multiple classifiers into an ensemble, their scores need to be calibrated. For this chapter, as in Chapter 3, I used the grade A and B lenses from GJ as 'true lenses', and assert that all other systems (ungraded or graded as a non-lens) are not lenses. Grade C candidates were also treated as non-lenses since in reality these are typically not lensed systems. Therefore, in this chapter, the probabilities produced following calibration reflect the probability that a system is a grade

Model	Type	Bands	N_{train}	Scaling	P_{50}
0	CS	I_{E}, Y_{E}, J_{E}	12	MTF & arcsinh	68
1	CNN (4-Layer)	I_{E}	3×10^4	MTF & arcsinh	2.1
2a	CNN (OU100)	I_{E}	3×10^4	MTF	0.36
2b	CNN (OU100)	I_{E}, J_{E}	3×10^4	MTF	0.34
3a	CNN (IncNet)	I_{E}	4×10^4	arcsinh	0.14
3b	CNN (ResNet)	I_{E}	4×10^4	arcsinh	0.08
3c	CNN (VGG)	I_{E}	4×10^4	arcsinh	0.07
4	CNN (BNN)	I_{E}	9×10^7	arcsinh	7.3
5	VT	I_{E}	1.6×10^5	arcsinh	0.15

Table 4.1: Summary of the models used in this work, as well as a summary metric of their performance, the purity (%) at 50% completeness (P_{50}), as measured on the test set. The P_{50} value for Model 0 was calculated using the systems in the 'random' 40 000 data set described in Sect. 4.2 for representative comparison. I denote the approximate training set size by N_{train} – in the case of Model 0 (Space Warps), this is given by the median number of training images seen by the citizens.

A or B quality lens. Due to the large number of classifiers applied to this lens search, and because high-scoring systems from all of these were inspected by lens experts and passed to Space Warps, it was reasonable to assume the majority of the lens systems were identified. Analysis in Paper C suggests 65% of grade A and B lenses were found in the Q1 lens search, with the vast majority of those missing being grade B. Therefore, the analysis in this chapter should be seen as conservative. Since some true lenses will have been labelled here as non-lenses (because non-inspected systems are classed as non-lenses in this work), the model performance will be underestimated. This is likely to be most significant at lower model scores which were not all inspected by experts.

The best calibration method identified in Chapter 3 was isotonic regression (Zadrozny and Elkan, 2002) which assumes that the mapping from the score of the classifier to the lens probability is a monotonically increasing function. Since I did not adjust the calibration methodology from Chapter 3, I did not use a separate validation set. I split the data into two equal sized data sets, forming a 'calibration set' (on which the calibration curves were calculated), and a 'test set'. I show the calibration mappings from model rank to calibrated probability produced by this method in Figure 4.2. To apply these calibration mappings to the test set, I

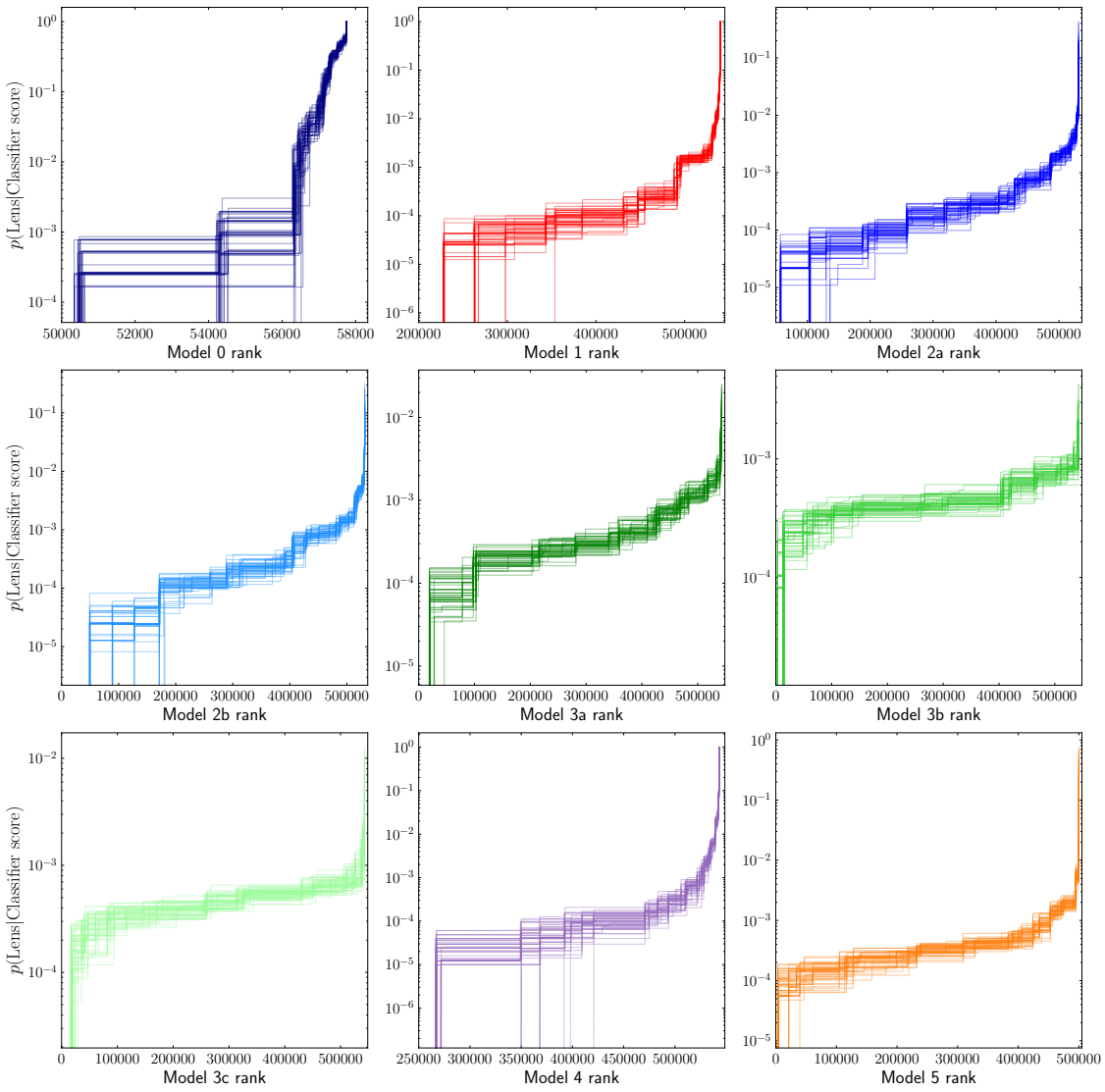


Figure 4.2: Mapping from the ranking of each object in the test set to the calibrated probability, based on the isotonic regression (Zadrozny and Elkan, 2002) procedure. I show 50 calibration curves for each model, generated by bootstrapping. The lower x-axis limit is trimmed for clarity to where the mapping is non-zero. The limits for Model 0 are more restricted, since the citizens were only shown a subset of the whole data set, and a high proportion of the lowest scoring systems received a calibrated probability of 0.

interpolated these rank values to the original model scores, such that the calibration function could map between model score and calibrated probability on new data. The best performing models could achieve calibrated probabilities $\mathcal{O}(1)$, indicating their highest scoring systems have a high lens purity. I validated the calibration on the test set, as shown in Fig. 4.3, where I measure the ratio of the number of lenses to the total number of systems with a given calibrated probability.

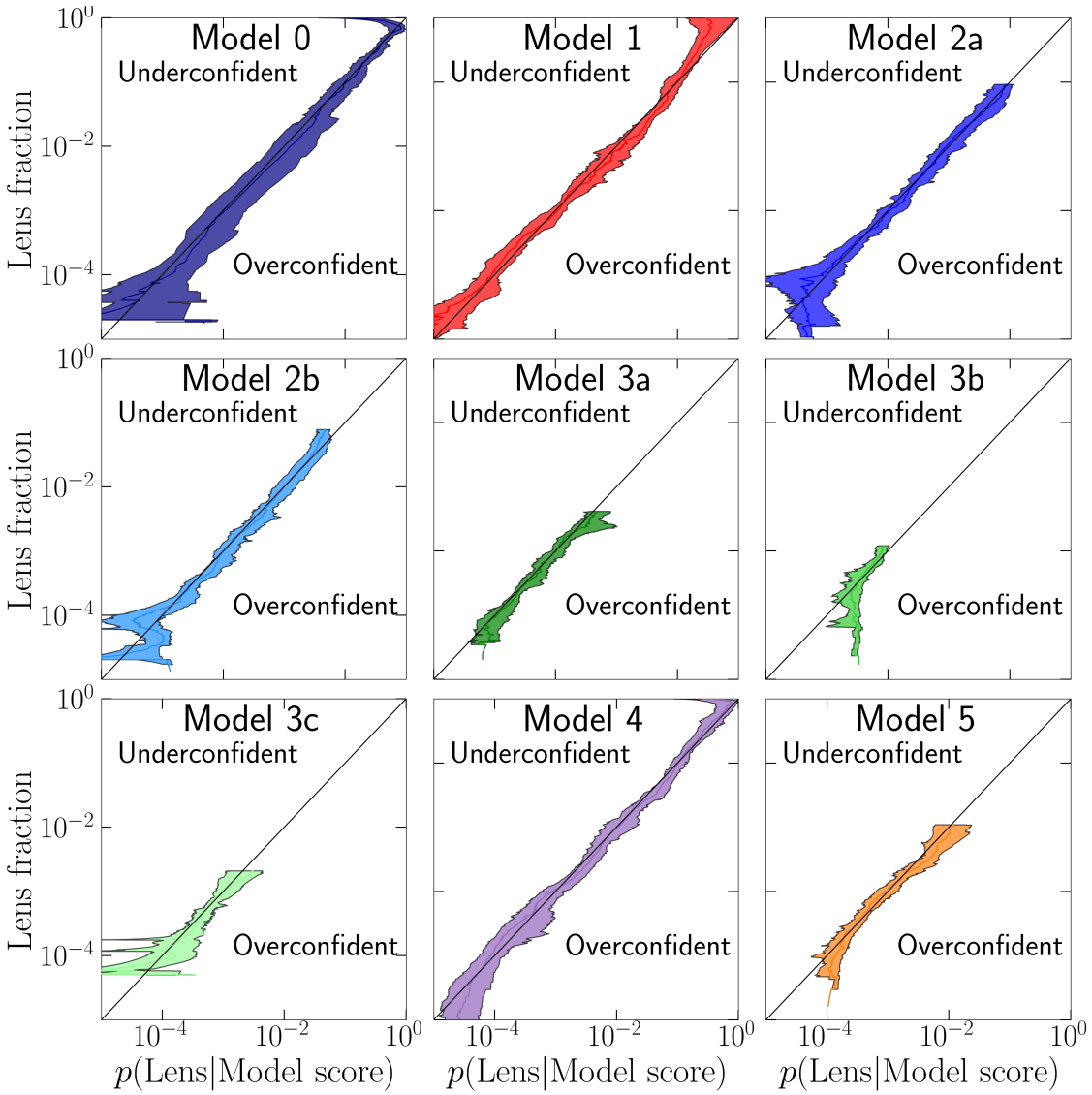


Figure 4.3: Validation of calibration curves, applied to the distinct test set of data. The best performing classifiers can be calibrated up to high probabilities, since their highest-ranked candidates have a high purity. Curves that follow the $y = x$ line are indicative of accurate calibration, where the calibrated probabilities (x-axis) match the fraction of grade A+B systems with that score in the test set (y-axis). 1σ uncertainties (shaded regions) are calculated via bootstrapping on the test set.

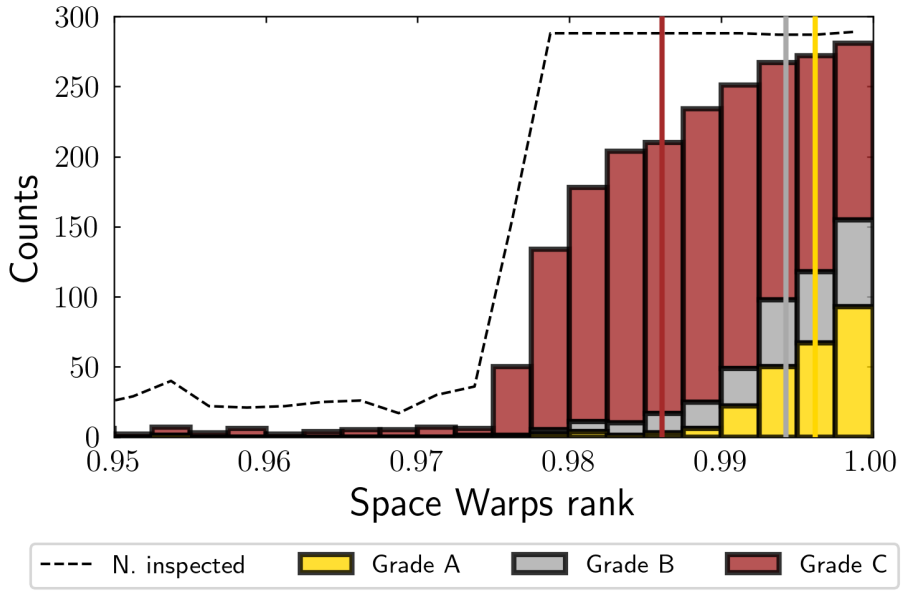


Figure 4.4: GJ grades of the top 5% ranked subjects from the Space Warps search; the bins are stacked, so the envelope represents the combined purity of grade A+B+C candidates out of the total inspected (dashed line) in each bin. The vertical lines indicate the median rank of lens candidates of each grade; higher-grade lens candidates received successively higher scores from the citizens.

As shown, the calibration is accurate across the range of lens classifiers and across many orders of magnitude. The best performing classifiers (Models 0, 1 and 4) can be calibrated up to high-lens fractions $\mathcal{O}(1)$, while models such as 3b and 3c can only be calibrated over a smaller range, since their highest scoring systems contain a comparatively low fraction of lenses. There is some underconfidence in Model 0 at high lens fractions, indicating some inflexibility in the calibration which assumes the lens fraction increases monotonically with model score. Having calibrated each classifier, I proceeded to combine them into an ensemble, described in Section 4.3.2.

I also investigated citizen scientist grades as an alternative source of ground truth. Given the number of systems that lensing experts can grade is limited, the possibility of citizens providing equivalent grading was of interest. Figure 4.4 shows the distribution of grade A, B, and C lens candidates as identified by GJ, in the highest scoring 5% of systems identified by Space Warps. The citizens were most confident at identifying higher-grade lens systems, which received successively higher scores. Given this and the strong performance of the citizen scientists (see Model 0 in Fig. 4.5), I investigated whether it was possible to use this classifier as a ground

truth for the purposes of calibrating the ML classifiers. I trialled two methods for this to determine if a more rigorous method performed significantly better.

1. Calibrating the output scores of Model 0 (Space Warps) using the original GJ ground truth as previously, but then calibrating the remaining models (1–5) using the now-calibrated Model 0 probabilities as a ground truth. To do this, for the second calibration of Models 1–5, I drew samples from the calibration set assigning binary classification values according to the calibrated probabilities from Model 0 each time. I then averaged over the calibration curves produced by each set of samples. In this manner, I accounted for the fact that the calibration of Model 0 produced probabilities, rather than binary classifications.

2. Using a simple threshold in score for Model 0, defining all systems above this threshold as lenses and vice versa.

I tested both of these methods on the test set using the original ground truth from GJ, the results of which are described in Section 4.4.

4.3.2 Combination of Classifiers into an Ensemble

Having calibrated the individual classifiers, they were then combined into an ensemble. The range of classifier types, training data, and network architectures produced outputs that typically showed little correlation for the vast majority of the objects in the data set that were not lenses. Following the calibration of each network in Section 4.3.1, I used a Bayesian approach to combine the individual networks, and treated each classifier as 'independent' as described in Chapter 3.

The volunteers participating in the Space Warps search were shown a mix of high-scoring lens systems from the ML classifiers, and random systems from the whole data set. Therefore, for the majority of objects, only eight classifiers were available to form an ensemble. However, scores from all nine classifiers ($8 \times$ ML + Space Warps) were available for the systems that were most likely be lenses.

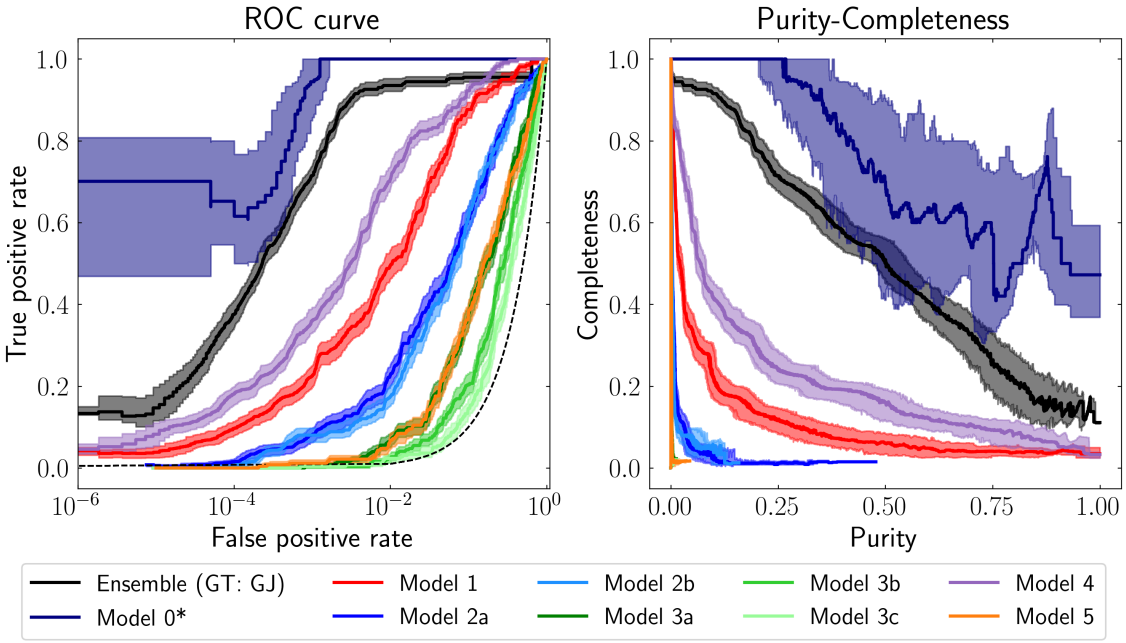


Figure 4.5: ROC (left) and purity-completeness curves (right) of the individual models and an ensemble of all nine models in this work, applied to the test data set. This ensemble was calibrated using the calibration set, assuming a ground truth of GJ grade A and B lenses. 1σ uncertainties (shaded regions) are calculated via bootstrapping on the test set. The dotted line in the ROC curve indicates the performance of a random classifier. *Model 0 (Space Warps) was applied to a mix of high-scoring ML systems and randomly selected galaxies. The ROC curve for Model 0 plotted here is generated using only the random subset (40 000), and hence has larger statistical noise. The completeness measurement for this classifier in particular is taken as if the citizens saw the whole data set, which would be unfeasible for future data releases (see Sect. 4.4.3). However, the curve for the ensemble is generated using the full test set.

4.4 Results and Discussion

4.4.1 Ensemble Classifier Performance

I found that the ensemble classifier made from all nine classifiers provided significant improvement in purity and completeness over the individual ML classifiers.

Figure 4.5 shows the ROC curve and purity-completeness curves of the individual classifiers, and that of the Space Warps+ML ensemble. To produce these, I used the expert-grades from GJ as a ground truth (in particular defining grade A and B systems to be lenses, and all other systems, graded or otherwise, to be non-lenses). In this case, the ensemble provides significant improvement over the ML classifiers, achieving 52% completeness at 50% purity. If grade C systems are excluded entirely

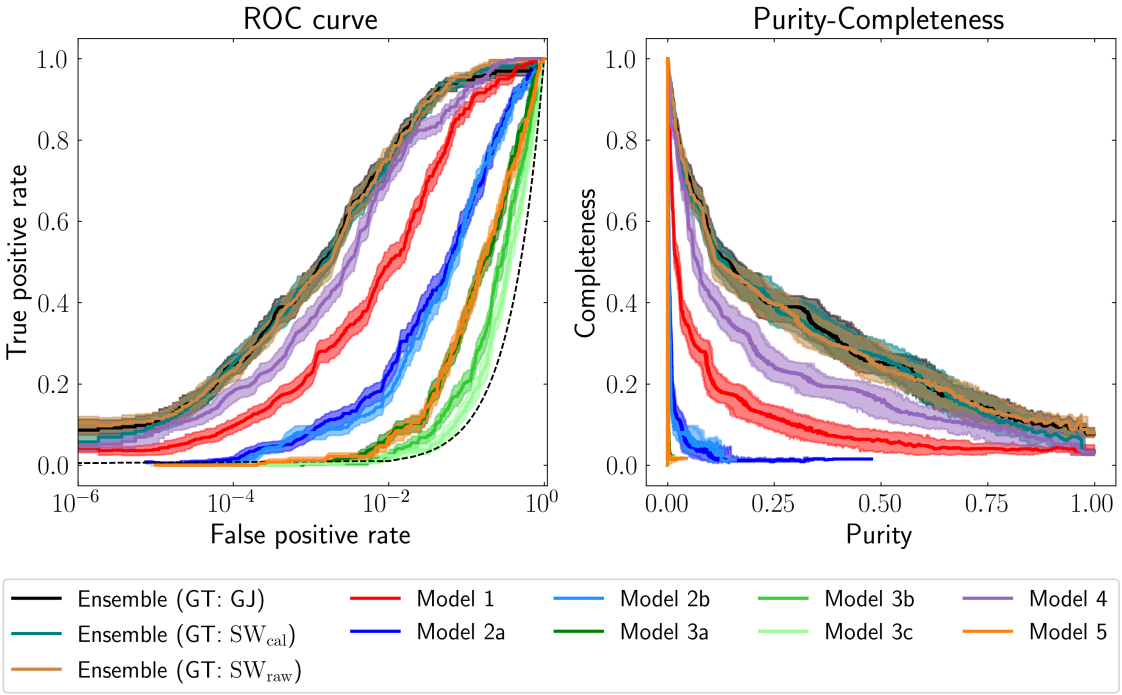


Figure 4.6: ROC (left) and purity-completeness curves (right) of the ML classifiers, and three iterations of an ensemble generated using Model 1, 2a and 4. These ensembles were calibrated using a ground truth of GJ (GT: GJ), Space Warps calibrated probabilities (GT: SW_{cal}), and Space Warps binary outputs (GT: SW_{raw}, using a raw SW score threshold of $p = 1 - 1.5 \times 10^{-8}$)

from the analysis (i.e., not treated as non-lenses), this metric further improves to 61% completeness, indicating some remaining high-scoring ensemble systems are grade C candidates. The low purity achieved by some models is indicative of the rarity of strong lens systems; false positive rates $\lesssim 10^{-3}$ are required for the resulting sample not to be dominated by non-lenses which is difficult to achieve. The Space Warps classifier (Model 0) is limited in the number of systems it can classify (to about 100 000 in this search) but Figure 4.5 shows that it performs very well, achieving $\gtrsim 70\%$ completeness at 50% purity (in the right-hand panel). However, since lenses are very rare and the total citizen-inspection budget is limited, in the future, the majority of systems will have to be pre-screened by ML models to maximise the number of lenses identified (see Section 4.4.3). The flexibility of the ensemble to account for different numbers of classifiers for each system means it can provide a ranked list of lens candidates across the *whole* data set rather than a subset, and with performance close to that which would be achieved if

the citizens had inspected the complete data set. The Model 0 curve in Figure 4.5 has much larger statistical noise since it is evaluated on the 40 000 systems in the 'random' subset shown to the citizens as opposed to the $\sim$ 500 000 systems in the test dataset used for the other classifiers. This curve shows much greater stochasticity in the high purity region in particular where the small number of A/B grade lens systems means this effect is most impactful.

Figure 4.6 shows the ROC and purity-completeness curves for only the ML classifiers, along with the results from a range of ML-only ensembles. These were generated using the same test set (and with the same GJ ground truth to define the FPR and TPR) as in Fig. 4.5. I find the best ML-only ensembles (which are plotted in Fig. 4.6) are generated using a subset of the ML classifiers (in particular, Models 1, 2a, and 4). This permutation of models was identified by adding these classifiers one by one to the ensemble until the performance peaked. The difference between an ensemble of Models 1, 2a and 4 versus Models 1-5 was marginal and within the statistical uncertainty. As described above, I generated these ensembles by calibrating the networks using three different ground truths; the original GJ grades, the Space Warps probabilities ($\mathrm{SW_{cal}}$, themselves calibrated using GJ), and the raw Space Warps score ($\mathrm{SW_{raw}}$, calibrating the networks by defining a 'lens' to be all systems with $p \geq 1 - 1.5 \times 10^{-8}$, based on the position of the 'knee' in the Space Warps ROC curve). I found the benefits of combining the networks to be smaller than that of combining the ML and Space Warps classifiers in Fig. 4.5. However, there was an improvement in classification using all these ground truths for calibration. This hints at the possibility of citizens substituting for expert grading at the larger scales of the forthcoming data releases, which I discuss further in Section 4.4.3.

In Figure 4.7 I demonstrate that calibration is a necessary step prior to combining different models into an ensemble. This figure shows the ROC curves generated from simply averaging the uncalibrated model scores compared to first calibrating each model before combining them in a Bayesian manner. I found that simply averaging the uncalibrated model scores produced a much lower performing classifier than first applying calibration and then combining them via the Bayesian framework

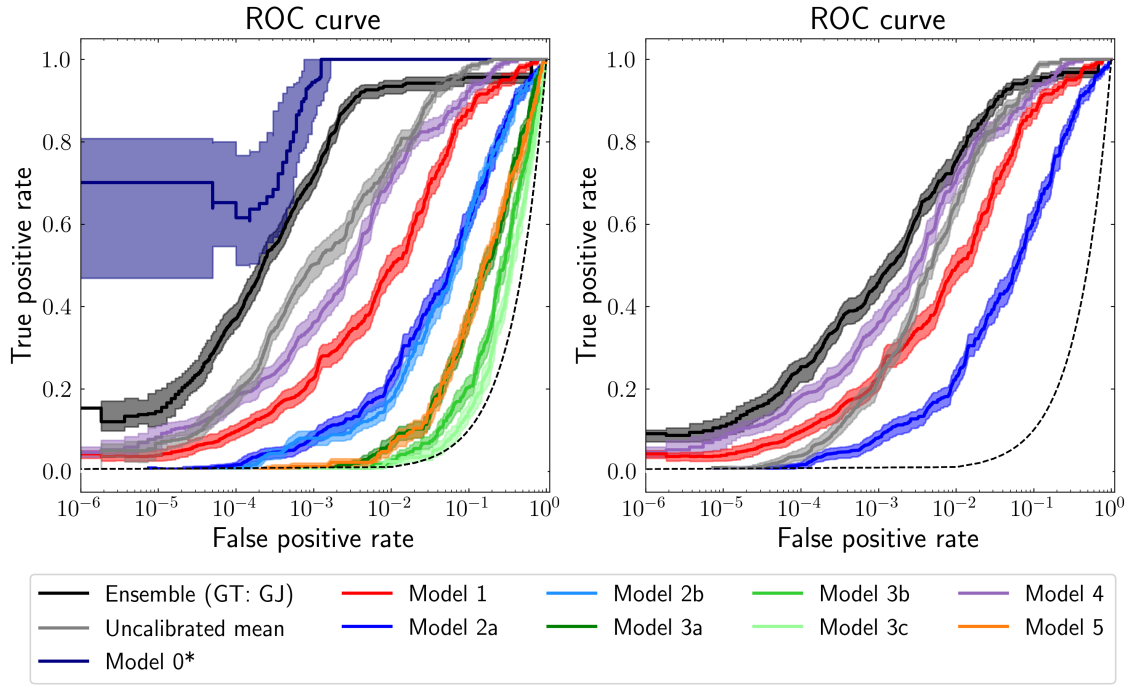


Figure 4.7: Comparison of the ensemble ROC curves generated via calculating the mean average of the uncalibrated model outputs (grey), versus Bayesian combination following calibration (black). The ensemble of Models 1, 2a and 4 is shown on the right, and the ensemble of all models is shown on the left. The ROC curves of the individual classifiers making up these ensembles are also shown as previously.

used in this work. Furthermore, the uncalibrated classifier performed significantly worse than its constituent parts in the case of the ML-only ensemble.

I also investigated the scatter between the calibrated probabilities produced by the models. I found that this scatter loosely correlated with the error ($|\text{Truth} - \text{Pred}|$) on the ensemble probability, i.e., systems were more likely to be misclassified by the ensemble when there was greater disagreement between constituent models. However, many high-grade systems classified correctly by the ensemble also had large scatter in calibrated probability between models. This derived primarily from the varying performance between models - only the best performing models could be calibrated up to high-probability values (see Fig. 4.3). This resulted in high scatter by default for likely lens candidates, since only the best models could assign probabilities $\mathcal{O}(1)$. I found that the highest correlation in model scores was between classifiers which shared a common training set (Models 3a, b, c and Models 2a and 2b). I also found correlation between Models 1 and 4 (the

best performing ML models), with Spearman's rank correlation $\rho = 0.56$. This correlation decreased ($\rho = 0.25$) when only considering grade A and B lenses, suggesting they still identified different types of lens.

4.4.2 Systems Identified by Citizens or Ensemble

ML and citizen science classifiers are naturally very different, and have their own strengths and weaknesses. These may arise from the particular data sets used to train both sets of images (for example, citizens only see a small fraction of the training images that an ML classifier would see), and the intrinsic strengths of the classification method (ML classifiers are excellent at rapid pattern detection, but may struggle with systems which are out-of-distribution such as rare artifacts). Figure 4.8 shows the ensemble posterior probability from a ML-only ensemble (from Models 1, 2a, and 4), versus that of a ML + Space Warps ensemble (all nine classifiers), along with a selection of cutouts for those only identified by the ML-only and ML + Space Warps ensembles. As expected, systems for which both the ML-only and Space Warps + ML ensembles have high posteriors are good lens candidates, including many grade A's. Systems for which the networks produced high scores but which were rejected by citizens have a range of morphologies. These include those with similar arc-like features (such as face-on spiral galaxies), artefacts, and very bright stars that may not have commonly featured in training sets. There were some systems which were ranked highly by citizens but received lower scores from the networks. Considering just the random sample (i.e., that inspected by both ML and citizen classifiers), 2 (16) A/B grade systems were ranked within the top 10% (1%) of systems by citizens but did not appear in the top 10% (1%) of any network. These included candidates that were mis-centred and those in crowded fields, and were typically B-grade.

Based on the `PyAutoLens` modelling (Nightingale et al., 2021) discussed in Paper A, I investigated correlations in classifier score and modelling properties. Here I restricted the calculations to lens candidates with successful lens models, judged to be lenses (see Paper A). This modelling provided estimates of the lensed/unlensed

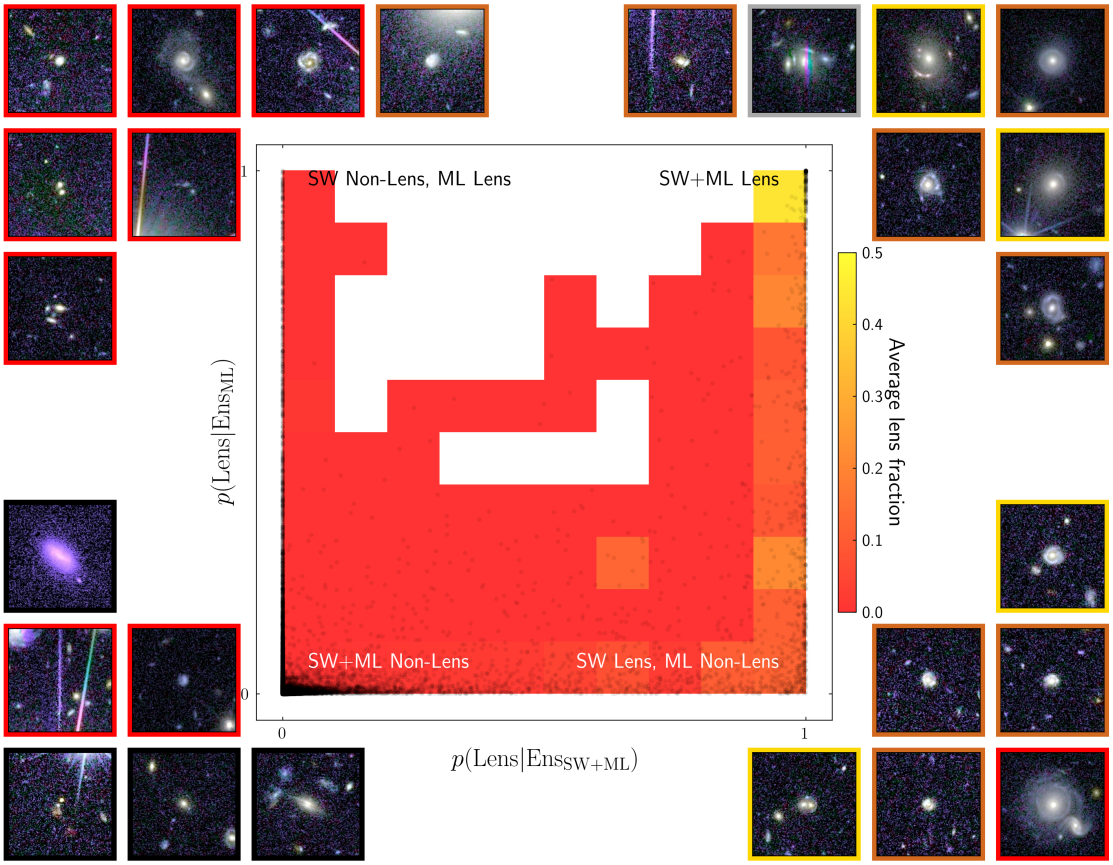


Figure 4.8: Plot of the lens posteriors from an ML-only ensemble compared to those produced by an ML+Space Warps ensemble. In the central plot, the 2D histogram depicts the average lens (grade A+B) fraction while the individual posterior values are over-plotted as scatter points (black). Regions with no data points are shown in white. Most data points are clustered towards the edges of the plot, i.e., very high or low posterior values. Systems that receive both a high ML-only posterior and a high ML+Space Warps posterior are the most likely lens candidates. A selection of systems identified by both the ML-only and ML+Space Warps ensembles are shown in the top right, while those only identified by the ML ensemble (ML+Space Warps ensemble) are depicted in the top left (bottom right). Examples of systems rejected by both ensembles are depicted in the bottom left. Cutouts are highlighted by their grade from GJ where available (A: gold, B: silver, C: bronze, Non-lens: red, Ungraded: black). The central plot demonstrates that the highest purity can be achieved when both ensembles are in agreement.

magnitudes, Einstein radius, and signal-to-noise ratio. I found the ML + Space Warps ensemble classifier score correlated most significantly with magnified source magnitude ($\rho = 0.5$), along with total signal-to-noise ($\rho = 0.47$). The Space Warps classifier was most closely correlated with the Einstein radius ($\rho = 0.3$), implying the human inspectors are more likely to identify large Einstein radius systems. The selection function of the human inspectors was measured as part of the lens search, and is discussed in Paper A.

4.4.3 Outlook for *Euclid* DR1 and Future Data Releases

In this work, I have used expert grades as a ground truth to generate an ensemble. Given the Q1 expert inspection, it will not be necessary to conduct a further expert inspection on DR1 data prior to generating a novel ensemble to be applied to DR1, since any new classifiers could be recalibrated on the existing Q1 dataset for which a sizeable lens sample has been identified, graded and modelled. This is a significant time saving, and means an ensemble can be applied to the DR1 dataset from the outset.

Given the multitude of lens classifiers expected to be applied to DR1 and beyond, such a method for combining scores will be very useful. The ensemble method enables scalable lens candidate prioritisation even if only a small sample of systems have been inspected. The ranked scores of an ensemble classifier in DR1 could be used to prioritise which systems to follow up in the forthcoming 4MOST Strong Lensing Spectroscopic Legacy Survey survey (4SLSLS, Collett et al., 2023). When this survey starts, in 2025-2026, it will become possible to calibrate the lens classifiers using spectroscopically confirmed systems. Determining the proportion of A and B-grade lens candidates which are truly lenses would then allow the expert-inspected sample to be used for accurate calibration. Benchmarking the expert grades against the true lens fraction from spectroscopy would enable the calibrated probabilities to be used in any subsequent inference using impure lens samples which is the subject of Chapter 5.

Beyond DR1, the calibration method described here could be tuned to particular strong lens science cases. For example, high-redshift lenses could be identified by recalibrating the classifiers through a smaller set of known high-redshift lenses, then applying this fine-tuned ensemble across the dataset. This would not necessarily require retraining the individual models (though this may be done anyway), as the recalibration would account for any changes in classification performance to this new target dataset.

Figure 4.9 shows the number of true lenses and false positives which would be expected as a function of purity in the full EWS, based on the performance of the Space Warps + ML and ML-only ensemble classifiers. For the former ensemble, a small majority of the A/B-grade lenses (52%) would be identified in a 50% pure sample. Given my definition of 'lens' to be grade A or B candidates in this work, it is likely that some of the other 50% would be grade C candidates. The best ML-only ensemble would achieve 25% completeness for the same purity, highlighting the value of combining citizen and ML approaches. To achieve significantly higher completeness would involve expert inspection of an increasingly large number of false positives which would rapidly become intractable.

How best to 'spend' the visual inspection budgets in DR1

Unlike ML classifiers, the number of images that both citizens and strong lensing experts can inspect is limited. Therefore, it is crucial to carefully manage what images are shown to citizens to optimise lens identification. For context, in the Q1 Space Warps lens search, around 1000 volunteers made 800 000 classifications of 100 000 cutouts over a period of $\sim$ 1 month. Wider advertisement of the project to the public could boost these numbers. For example, a Space Warps strong lens search using HSC data (Sonnenfeld et al., 2020) was featured on a US national radio channel and received 2.5×10^6 classifications from 10 000 volunteers over a 2 month period. Furthermore, more relaxed time constraints for DR1 inspection and analysis compared to Q1 will help to increase both the citizen and expert inspection budgets. However, there remains a limit to the number that can be inspected. The

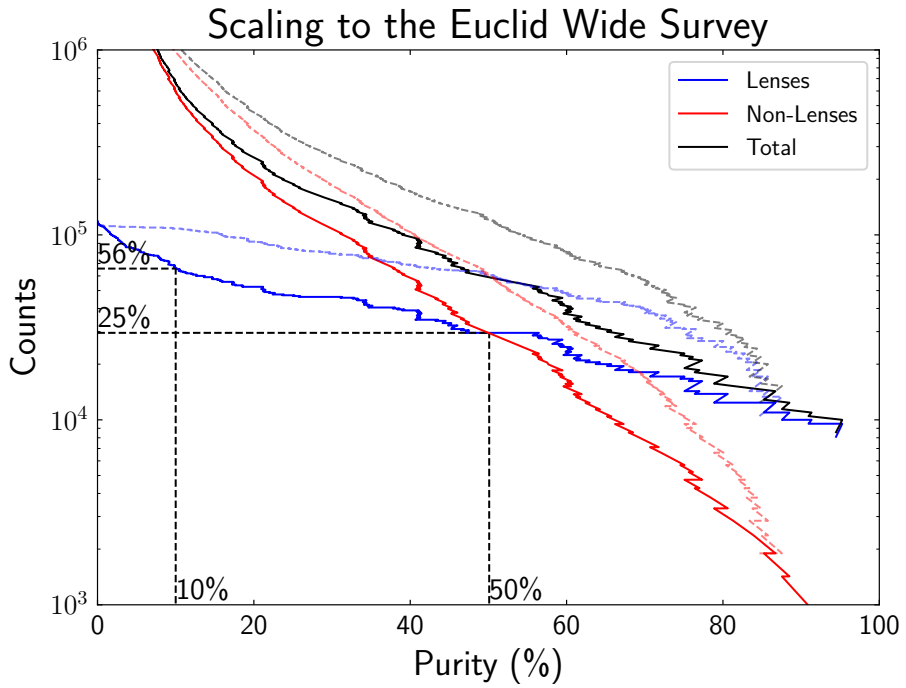


Figure 4.9: Estimate of the number of true and false positives that would be identified by a classifier with the performance of the ensemble classifiers in this work. The bold curves show the performance of the network-only ensemble, while the faint curve shows that of the full nine-classifier ensemble; the latter being an optimistic scenario, given citizen classifications likely would not be available for a data set significantly larger than the Q1 data set used in this work. The total number of lens systems has been scaled to reflect the number of grade A and B lenses found in Q1. Completeness values for two illustrative purity thresholds are shown by the dotted lines.

citizens in the Q1 lens search were shown a mix of cutouts of high-scoring systems, and random cutouts from Q1 data. While showing a small number of randomly selected galaxies can help to identify unusual lens configurations, a purely random selection of systems to the volunteers would not include the vast majority of lens systems (only about 70 lenses in a random selection of 100 000 cutouts using the Q1 pre-selection). This supports a two-stage approach, whereby ML classifiers are applied to the full DR1 data set, of which a subset (likely a few thousand) are inspected by experts to verify the performance of the networks. The ML classifiers would then be collated into an ensemble, and the resulting ranked list could be used to inform which cutouts are shown to citizens. The ensemble methodology provides a natural mechanism for prioritising which systems to show to the citizens, averaging over the incompletenesses and selection biases of the individual classifiers and thus

increasing the efficiency of lens classification. While this strategy is likely to produce the most lens candidates, an additional untargeted search by the citizens would be of significant value - while it would not identify many lenses it would help to determine the completeness/selection function of the galaxy pre-selection and search strategy.

The strong lensing experts who took part in GJ inspected and graded 7000 images over approximately 3 months. The DR1 data set will be 36 times larger than the Q1 data release, but the inspection capacity of expert graders is likely to remain similar. Given the longer timescale, it is possible that experts will inspect a larger sample so I consider two search strategy scenarios here. I first scale the results by a factor of 36× to that of the DR1 area (in particular the number of lenses/non-lenses above a given model threshold), then make cuts based on realistic inspection limits to calculate the total number of lenses which may be found. Based on the test set used in this work (and assuming for the moment that a similar ensemble was produced, perhaps via the calibrated Q1 networks), the highest-ranked 100 000 systems from an ML-only ensemble would contain 9100 grade A/B lenses, out of a total of 15 000 detectable systems in DR1. By comparison, 7300 would be identified using the same cut on Model 1 alone. If the former were shown to citizens, the highest ranked 5000 from a ML + Space Warps ensemble would contain around 3900 grade A/B systems (i.e., a fairly pure but incomplete sample). A simple cut on Space Warps score would produce 3500 systems in the top 5000. In an optimistic scenario, in which the citizens inspect 1 000 000 systems from an ML ensemble, of which the highest-scoring 15 000 are passed to experts, I find that approximately 7600 would be identified if the scores from all classifiers were first combined into a citizen+ML ensemble. Therefore, combining the scores into an ensemble at each stage would be worthwhile to produce a higher purity sample.

4.4.4 Optimising Lens Searches in Wide Area Surveys

Although measured using different lens classifiers and surveys, the purity-completeness curves in Figures 3.11 (LSST forecast), 3.14 (DES) and 4.5 (*Euclid* forecast) are striking in their similarity. Without significant improvement in lens classification,

the lens samples in LSST and *Euclid* will be hindered by either low purity ($\lesssim 50\%$), low completeness ($\lesssim 50\%$) or both. However, there is significant opportunity for improving the current performance of these classifiers. Due to the huge data volume anticipated with forthcoming surveys, such improvements must limit the need for manual intervention, while making best use of citizen scientists who are the current best-performing classifiers.

Individual lens finding algorithms will continue to improve as ML training sets become more complex; the best network in Cañameras et al. (2024) further improves upon that of Cañameras et al. (2021) used in Chapter 3 (the former was not applied to the whole HSC survey) and suggest $\mathrm{TPR}_0 \sim 60\%$ could be achievable. The combination of the much larger sample of real *Euclid* strong lens systems now available (Paper A) for training and validation and the more relaxed timescale for classifier fine-tuning for *Euclid* DR1 versus Q1 mean it is likely that classification performance will naturally improve further prior to DR1. The large number of classifications of *Euclid* cutouts now available from both citizens and expert graders provide a labelled data set with which to re-train the ML classifiers, particularly on difficult false positive systems. This is likely to occur iteratively following each future data release allowing the models to continuously improve over time. Active learning, whereby a ML model is retrained iteratively based on new labels from the most informative systems (see for example Walmsley et al., 2020; Walmsley et al., 2022) would enable such improvements to occur concurrently with future lens searches. Furthermore, rapid, large-scale modelling of lens systems (e.g., Poh et al., 2022; Gentile et al., 2023; Schuldt et al., 2023b; Erickson et al., 2024, Busillo et al. in prep., Venkatraman et al. in prep.) will provide an additional measure for the plausibility of candidate lens systems; therefore, I anticipate that higher completeness values than estimated here are achievable.

With respect to ensemble classifiers, the large differences in the rank ordering of objects between classifiers (Figure 3.4) suggest classifiers trained on a diverse range of training data find different types of non-lenses easier/more difficult to classify. This suggests that larger ensembles could offer further improvement. This

is also reflected in the toy model shown in Figure 3.5, where continual classification improvement is shown when adding independent classifiers to the ensemble (Figure 3.5a), but much more meagre improvement if these classifiers are dependent (Figure 3.5b). I investigated this with the classifiers applied to the HSC data in Chapter 3; I measured the AUROC, FPR_{50} (false positive rate at 50% completeness) and TPR_{-3} (completeness at an FPR of 10^{-3}) as a function of number of classifiers in the ensemble, averaged over the combinations of available classifiers. I found the primary benefit in these metrics was achieved when adding ≥ 3 classifiers, although the ensemble continued to improve up to the 6 used in this work. In contrast, the best performing network-only ensemble applied to *Euclid* data used in this chapter did not use all the networks. The difference in performance here was well within the uncertainty and dependent on the performance metric, but was also likely to be affected by the wide variation in performance between *Euclid* lens classifiers not seen in the HSC search.

As shown in Section 4.4.3, the visual inspection budget of experts and citizens limits the completeness of the resulting sample. This could be improved by adding a classification refinement stage to citizen inspection as done in Marshall et al. (2016), whereby citizens were shown high-scoring candidates for a second round of classification. This could also involve the citizens grading systems in line with typical lens searches (A/B/C/X), and calibrating their classifications to match those of experts. Additionally, applying the SWAP methodology (Marshall et al., 2016) to the final expert inspection, whereby the final grade was a weighted average of the experts' grades based on their performance on a chosen training set, would increase the efficiency (and thus reduce the time burden) of the grading.

A priority for *Euclid* DR1 will be to obtain a sufficient dataset ($\sim 10\,000$ systems) of high-grade strong lens candidates for follow-up through 4SLSLS. This goal is likely to be met, based on the current performance of the lens classifiers and anticipated improvements. Beyond this, the importance of purity versus completeness will vary depending on the particular science case. Typical lens studies have focussed on spectroscopically confirmed systems. However, if the contamination rate is

known, unbiased inference can be performed on impure datasets (e.g., Kunz et al., 2007; Roberts et al., 2017 and Chapter 5).

4.5 Conclusions

In this chapter, I have produced an ensemble strong lens classifier using the *Euclid* Q1 data release. In answer to the questions set out in Section 4.1, I summarise my conclusions below.

1. An ensemble classifier for lens classification in space-based imaging provides significant improvement in classification when both ML and citizen science classifiers are used in the ensemble. In particular, the ensemble can still be used across the whole data set, providing posterior probabilities that each system is a lens, even when some classification data is incomplete (for example where citizens are only shown a subset of the data).

2. The *Euclid* ensemble classifier composed of neural networks and citizen scientists produced a 52% complete sample at 50% purity and a 91% complete sample at 10% purity. The ensemble comprised of only ML classifiers produced a 25% complete sample, with 50% purity and a 56% complete sample at 10% purity. Due to limited inspection budgets, it is likely that future expert inspected samples will have much higher purity than at present (e.g., 6.8% in this work).

3. Citizen classification can produce a high-purity sample of lens candidates, and higher-grade lenses receive progressively higher scores from citizens. Citizens could stand in for expert graders in future searches although care will need to be taken to account for the possibility of misclassifications and the total citizen-inspection budget.

4. Showing citizens a random selection of cutouts would only result in a small fraction of lens systems being identified, since the vast majority of cutouts would not contain a lens system, and the number of images citizens can

inspect is limited. Using a two-stage approach via an ML-only ensemble, whereby citizens are only shown highly-scored systems from this ensemble would significantly increase the total number of lenses identified.

5. Fine-tuning machine learning classifiers by using ensemble scores from this Q1 search as labels within their training sets would likely diversify the range of lenses that these automated methods could identify. Furthermore, given anticipated lens search campaigns with future data releases, such fine-tuning could be undertaken iteratively as more lenses and non-lenses are classified.

With predicted improvements in lens classification following this lens search, I anticipate more than 10 000 A/B grade lenses will be identified in *Euclid* DR1, heralding the start of a new era for strong lens science.

5

Lens Modelling and Cosmological Inference from an Impure Sample of Galaxy-Galaxy Strong Lenses

The basis of this chapter appears in 'Lens Modelling and Cosmological Inference from an Impure Sample of Galaxy-Galaxy Strong Lenses', Holloway et al. (in review)

Contents

5.1 Introduction 136
5.2 Data 139
5.2.1 Network Training Set 141
5.3 Method 142
5.3.1 Neural Network Training 142
5.3.2 COSMIC-BEAMS Formulation 144
5.3.3 Generation of Inference Data Vectors 151
5.4 Results 155
5.4.1 Lens Modelling of LSST Lenses 155
5.4.2 Lens Modelling of False Positives 156
5.4.3 Posterior Images for Lenses and False Positives 158
5.4.4 Cosmological Inference from an Impure Sample of Strong Lenses 159
5.5 Discussion 161
5.5.1 Modelling of LSST Lens Candidates 161
5.5.2 Inference with Impure Samples of Strong Lenses 164
5.6 Conclusion 166

The prospect of using large samples of strong gravitational lenses as alternative probes of cosmological parameters will be realised within the next decade. In this chapter, I present the first consideration of the practicalities of deriving cosmological parameters with large but contaminated samples of galaxy-galaxy strong lenses.

5.1 Introduction

Strong lens systems are rich in cosmological information. Lensed quasars, supernovae, clusters, double-source-plane lenses and single-plane galaxy-galaxy lenses have all been used to derive cosmological parameters (e.g., Collett and Auger, 2014, Birrer et al., 2020, Caminha et al., 2022, Li et al., 2024, Pascale et al., 2025). The Einstein radius of a strong lens is dependent, via angular diameter distances, on cosmology. Earlier works such as Marshall et al. (2005) and Grillo et al. (2008) have used static galaxy-galaxy lenses as a cosmological probe through this dependence, however, due to their small sample sizes they were limited to restrictive priors on the assumed cosmology or weak constraints. The large lens samples from LSST and EWS will provide a much more stringent test on cosmology and demonstrate the potential of strong lenses as cosmological probes. A small but significant proportion of these samples (up to 10 000 systems), are expected to be spectroscopically confirmed by the 4SLSLS survey (Collett et al., 2023), that aims to provide lens and source redshifts, as well as velocity dispersion measurements for these systems. However, this will leave the majority of the LSST and *Euclid* lens populations without spectroscopic confirmation. Li et al. (2024) determined the precision with which cosmological parameters (Ω_m, Ω_k, Ω_Λ, w) can be determined from a spectroscopic sample of 10 000 lenses expected from the EWS and 4SLSLS survey, finding that w should be determined to a greater precision than any other single-probe measurement ($\sigma_w = 0.11$, $\sim 30\%$ tighter than constraints from BAO and Type Ia supernovae, Alam et al., 2021; Brout et al., 2022). These constraints will tighten further when combined with the other strong lensing probes (from time-delay cosmography and

DSPL's) as discussed in Section 1.2.6. In this chapter, I extend the analysis of Li et al. (2024) to include the 'photometric sample' of strong lenses, a much larger sample of systems which may help to provide even tighter cosmological constraints than obtainable from only lensed systems with spectroscopic confirmation. If the strong lensing constraints (combining this Einstein ring probe with time-delay cosmography and DSPL's) together with other cosmological probes (CMB, BAO, Type 1a supernovae etc.) were to find deviations from ΛCDM, as hinted by the recent results from the DESI (DESI Collaboration et al., 2025a; DESI Collaboration et al., 2025b), this would represent a significant shift in our understanding of cosmology.

The success of galaxy-galaxy lenses as cosmological probes will require curation of lens candidates and the derivation of accurate lens parameters from modelling. Due to the number of strong lens discoveries anticipated in the coming years, such modelling must be fast to be feasibly scaled to the number of lens candidates. Given their rapid evaluation time once trained, machine learning methods such as neural networks are a natural fit for such a challenge and have been tested at scale and with high model complexity (Hezaveh et al., 2017; Pearson et al., 2019; Pearson et al., 2021; Schuldt et al., 2021; Poh et al., 2022; Schuldt et al., 2023a; Gentile et al., 2023; Erickson et al., 2024; Poh et al., 2025, Venkatraman et al. 2025 in prep.). As discussed in Section 1.2.7, one method for this is NPE (Lueckmann et al., 2017; Papamakarios and Murray, 2018); training a network to predict the posterior distribution (often with a fixed e.g., Gaussian functional form) for the parameters of interest. Having a full posterior distribution, rather than a point estimate, allows for more rigorous analysis and is used in this work to constrain lens parameters.

This chapter is divided into two parts. The first part focusses on strong lens modelling via NPE (in particular, using `paltas`[1], Wagner-Carena et al., 2023) to determine the precision with which lens parameters can be determined from realistic simulated images of lens systems for the LSST survey. The second part uses these results to infer cosmological parameters from a realistic sample of strong lenses comprising those which are spectroscopically-confirmed (hereafter

[1] https://github.com/swagnercarena/paltas

the spectroscopic dataset) and lens candidates that have only photometric data (and thus photometrically derived properties such as redshifts, hereafter termed the photometric dataset). I focus on LSST rather than *Euclid* since the optical bands of LSST will provide more accurate photometric redshifts than those solely from *Euclid*, and the photometric sample is the one of primary interest in this work. I discuss the possible differences in my results when applied to *Euclid* data in Section 5.5.

Part 1: I aim to:

- Determine the precision and accuracy with which lens parameters (in particular the Einstein radius θ_E and mass density slope γ) can be measured for these systems. I will also determine for which systems these properties can be measured the most accurately.

- Evaluate the behaviour of the neural network when faced with realistic false positives (i.e., images which do not contain a lens system but which previously confused a lens classifier). By doing this, I will determine whether fast modelling will be useful as a first step in lens confirmation.

Part 2: The photometric sample of strong lenses in forthcoming LSST and *Euclid* surveys is expected to be much larger than the spectroscopic sample (by roughly $10:1$). In Chapter 3, it was shown that even with an ensemble of strong lens classifiers the number of false positives will outnumber true positives in the LSST survey for a sample completeness $\gtrsim 40\%$. Therefore, to utilize the full strong lens sample, one must account for both increased measurement uncertainties and the effect of contamination from false positives. Similar circumstances affect Type 1a supernova cosmology, for which the number of spectroscopically confirmed 1a supernovae is much smaller than that for which only photometric data exists (e.g., Hlozek et al., 2012). With this in mind, the Bayesian Estimation for Multiple Species (BEAMS) framework was developed (Kunz et al., 2007; Hlozek et al., 2012; Kunz et al., 2013), to account for impurities in the sample while inferring unbiased cosmological parameters. This was extended by Roberts et al. (2017) to account for redshift uncertainties. In this chapter, I adapt this framework to the galaxy-galaxy

strong lensing case, including uncertainties on velocity dispersion, redshifts and lens/non-lens classification. Given the results from Part 1, I then:

- Develop a framework to incorporate unconfirmed strong lens candidates in cosmological analysis. I term this COSMIC-BEAMS (COntaminated Strong lensing Measurements to Infer Cosmology - Bayesian Estimation Applied to Multiple Species).

- Determine the precision with which Ω_m, Ω_Λ, w_0 and w_a can be measured from a combined sample of 100 000 strong lens candidates and 10 000 spectroscopically confirmed systems. This uses the lens model parameter uncertainties determined in Part 1, as well as the behaviour of the neural network to false positives.

The chapter is structured as follows. In Section 5.2 I describe the simulated LSST lens population used throughout this work along with the generation of cutouts and network training sets. In Section 5.3 I describe the networks trained for lens modelling (Sect. 5.3.1), the formalism for deriving cosmological parameters from impure datasets (Sect. 5.3.2) and the generation of data vectors to apply this (Sect. 5.3.3). I present my results in Section 5.4, including the modelling performance on simulated LSST lenses (Sect. 5.4.1), the network's behaviour when applied to false positives (Sect. 5.4.2) and cosmological inference from an impure dataset (Sect. 5.4.4). I discuss these results and further applications of the techniques used in Section 5.5 and conclude in Section 5.6. The data flow in this work is depicted in Figure 5.1.

5.2 Data

I created a realistic catalogue of strong lenses detectable in LSST, then generated single-band LSST-like images for subsequent analysis.

The lens catalogue was generated using `LensPop` (Collett, 2015), which used distinct lens and source populations, with a velocity dispersion function based on SDSS (Choi et al., 2007) and a singular isothermal ellipsoid lens mass model.

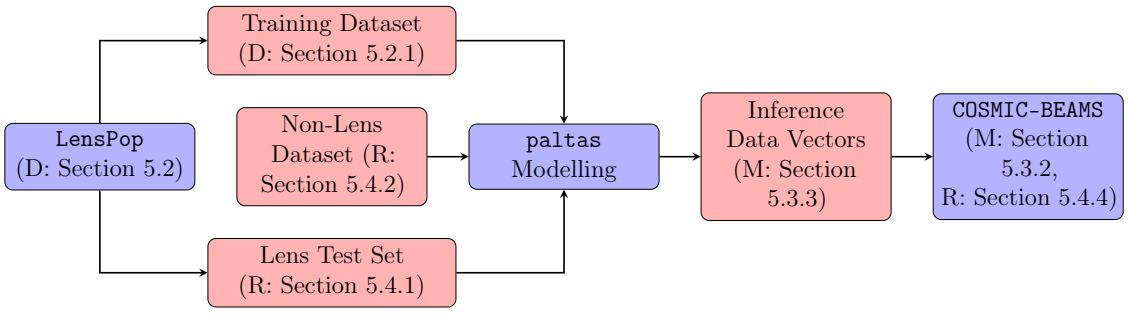


Figure 5.1: Flowchart of the data generation and analysis described in this work. I label data, method and result sections as *D*, *M* and *R* respectively. Datasets are highlighted in red and data manipulation is shown in blue.

The source properties were taken from a simulated LSST catalogue described in Connolly et al. (2010), complete to $i \sim 27.5$. `LensPop` also included stringent cuts on the detectability of each lens including SNR $= \sum \mathrm{S}/\sqrt{\sum \mathrm{N}^2} > 20$ and a magnification cut of $\mu > 3$.

When generating the above lens catalogue I configured `LensPop` to match the DP0.2 LSST simulation (LSST DESC et al., 2021; Korytov et al., 2019). DP0.2 is a large-scale end-to-end simulation over $\sim 300\,\mathrm{deg}^2$ of the simulated LSST sky, incorporating the survey cadence, multi-band imaging, image processing and catalogue generation expected from 5 years of LSST data. For this I adopted 100×30s i-band exposures, with seeing $0.83''$, gain of $0.7e$/ADU, zeropoint of 31.8 (1 count/exposure-time), with a SNR ≥ 20 threshold. Around 20 000 systems were produced which passed these cuts, with SNR ≥ 20 being the most influential detectability constraint (Collett, 2015). Note that expectations of $\sim$ 100 000 detectable systems in LSST (Collett, 2015) are for optimally stacked coadds (i.e., including only good seeing exposures) from 10 years of LSST data in g, r and i-bands (Collett, 2015); I concentrated on systems detectable in the i-band without optimal stacking (i.e., in full-coadds) for this work.

Images of the lens systems described above were then simulated by `paltas` and injected into DP0.2 coadds using the LSST Science Pipeline within `SLSim`[2]. Since the DP0.2 simulation produced 5-year coadds, the image noise, depth and corresponding modelling precision achieved in this work will be conservative compared to the

[2] https://github.com/LSST-strong-lensing/slsim

complete LSST survey. Lenses were injected into random patches of this simulated LSST sky and thus the PSF, pixel-exposure maps and pixel-noise maps varied between coadds. I used cutouts of 60 pixels (12″ on a side), with the LSST pixel scale of 0.2″. To ensure lenses were not injected on top of existing galaxies, cutouts with flux 2σ above the noise level within the central 26 pixels (accounting for the typical lens size and offsetting) were not used for injection. This was to ensure that the simulations only included single-plane lenses (rather than DSPL's or systems with deflectors at different redshifts), which are the focus of the cosmological inference in this chapter. Multi-plane lens systems would need to be treated differently, both during simulation and inference.. The injected lens and source galaxies included Poisson noise calculated using the exposure map of the corresponding DP0.2 cutout. Therefore, the noise in each coadd pixel reflected the number of single-exposures making up that particular pixel.

5.2.1 Network Training Set

For the training set, I fitted a Multi-Variate Normal (MVN) distribution to key system parameters (e.g., redshifts, Einstein radii, and lens/source magnitudes/sizes) of the `LensPop` test set, and generated a training set by drawing parameter values from a MVN distribution 20% wider than that of the test set. This ensured the training set encompassed a wide variety of lenses beyond simply those in the test set. The performance of the networks improved when a magnification cut of $\mu \geq 3$ was included in the training data (in particular with reduced bias on measuring the Einstein radius), but the training set incorporated none of the co-variances from the test set beyond this and the requirement that $z_{\rm L} < z_{\rm S}$ (as shown in Figure 5.2). This exposed the networks to a wide range of lens configurations, without fine-tuning them to the test data.

The lens and source light followed an elliptical Sérsic profile, while the lens mass distribution in the training set followed a PEMD profile (Barkana, 1998), with density slope γ. For the test set, the power-law index was fixed at $\gamma = 2$ (isothermal) consistent with `LensPop`. This ensured that the test set systems were

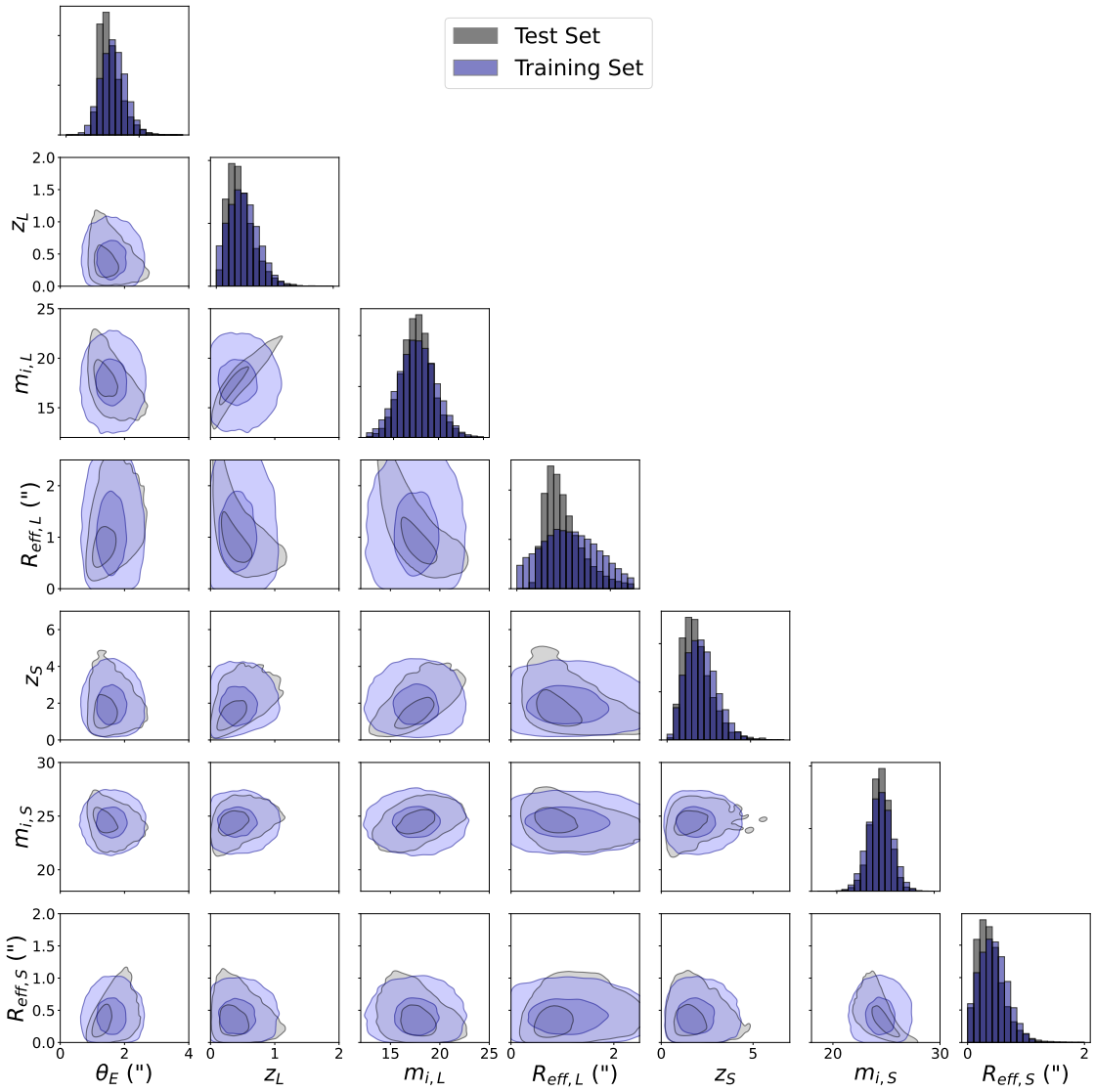


Figure 5.2: Plot of the distributions of lensing parameters of the training set (blue) and test set (grey). With the exception of a magnification cut $\mu > 3$, the detection cuts applied to the test set Collett (2015), were not applied to the training set, which subsequently shows a wider range of lensing systems.

significantly magnified/strongly lensed, while allowing the realistic precision of γ measurements to be determined.

5.3 Method

5.3.1 Neural Network Training

Four networks were generated using `paltas`, each differing by the training set or learning parameters as follows. During training, the input data for these networks

were LSST-like images (described below) containing simulated lens systems, along with their true lens parameters. Except where specified, the networks were trained to infer the mass density slope, shear, lens ellipticity and position, and Einstein radius of the lens. In the testing phase, the input to the networks were the LSST-like images in the test set. The networks produced parameters describing the posterior, which had a multivariate Gaussian functional form. Therefore, the output was the posterior mean and width for each of the parameters of interest. The multivariate Gaussian posterior produced by the networks had a diagonal covariance matrix, i.e., the posteriors for each parameter were independent. The NPE loss function (Eqn. 1.33) minimises the Kullback–Leibler divergence between the true posterior and the approximate posterior produced by the network allowing both the posterior mean and width to be accurately determined, using the training set as a prior. The networks had a ResNet architecture (in particular, xResNet-34), which were developed to overcome the degradation problem seen with deep networks [3] (He et al., 2016a; He et al., 2018), thus allowing more complex parameters to be learnt. The following four networks were trained:

1. **'Fiducial' (`FID`):** Lensed images were injected into the 5-year DP0.2 cutouts. Poisson-limited lens subtraction was assumed.

2. **'Best Seeing' (`SEE`):** As in the 'Fiducial' case; however, here the 5-year DP0.2 coadd cutouts consisted only of the top 1/3 best seeing single exposures, rather than all single exposures at a given position. This improved the resolution of the combined image (by $\sim 0.1''$) but reduced the depth.

3. **'No Neighbours' (`NNB`):** As in the 'Fiducial' case; however, the simulated lensed images were not injected into cutouts from the DP0.2 simulation and consequently did not have neighbouring (unlensed) galaxies in the image. Given the DP0.2 cutouts included LSST noise properties, for this network I replicated this by introducing equivalent sky/background noise via `paltas`.

[3]The degradation problem, illustrated in He et al. (2016a), occurs when deep neural networks show reduced training and test performance compared to shallower networks. This can be overcome by adding a 'residual' mapping of the identity function between layers of the network.

Therefore, these cutouts still had the noise properties expected from 5-year LSST coadds. This network was used to test whether the presence of neighbouring objects in the cutout was a significant source of confusion for network modelling.

4. **'No Lens Subtraction' (LLT)**: As in the 'Fiducial' case, but without Poisson-limited lens subtraction. It was this network that was subsequently applied to the false positive systems from the HSC survey (as described later). Additional parameters (size and magnitudes of the lens and source, and source position) were included in a retrained version of this network in Section 5.4.3.

Each network was trained on 500 000 training images, using the Adam Optimizer (Kingma and Ba, 2017) with a learning rate of 5×10^{-4} and a batch size of 256. A diagonal covariance matrix was assumed for the functional form of the NPE posterior since this produced more stable evolution in the network loss during training than when using a full covariance matrix. This removed any degeneracies between the posteriors of the learned parameters, but did not lead to a reduction in network precision and so was adopted here. I show example cutouts from the test sets of each network in Figure 5.3.

5.3.2 COSMIC-BEAMS Formulation

Here I present a formulation to infer unbiased cosmological parameters from an impure sample of strong lenses. This was originally inspired by the zBEAMS (Roberts et al., 2017) and BEAMS (Kunz et al., 2013; Hlozek et al., 2012; Kunz et al., 2007) methodologies, but has been adapted extensively to the strong lensing case. I split the science cases into 3 categories: Spectroscopic, Photometric and Photometric with contamination. The Probabilistic Graphical Model (PGM) for the latter scenario is shown in Figure 5.4. I denote $\mathbf{\Omega}$ to be the cosmological parameters I wish to constrain, $\mathbf{\Omega} = \{\Omega_m, \Omega_k, w_0, w_a\}$, and $\mathbf{D}$ to be the data. I wish to infer the posterior $P(\mathbf{\Omega}|\mathbf{D})$.

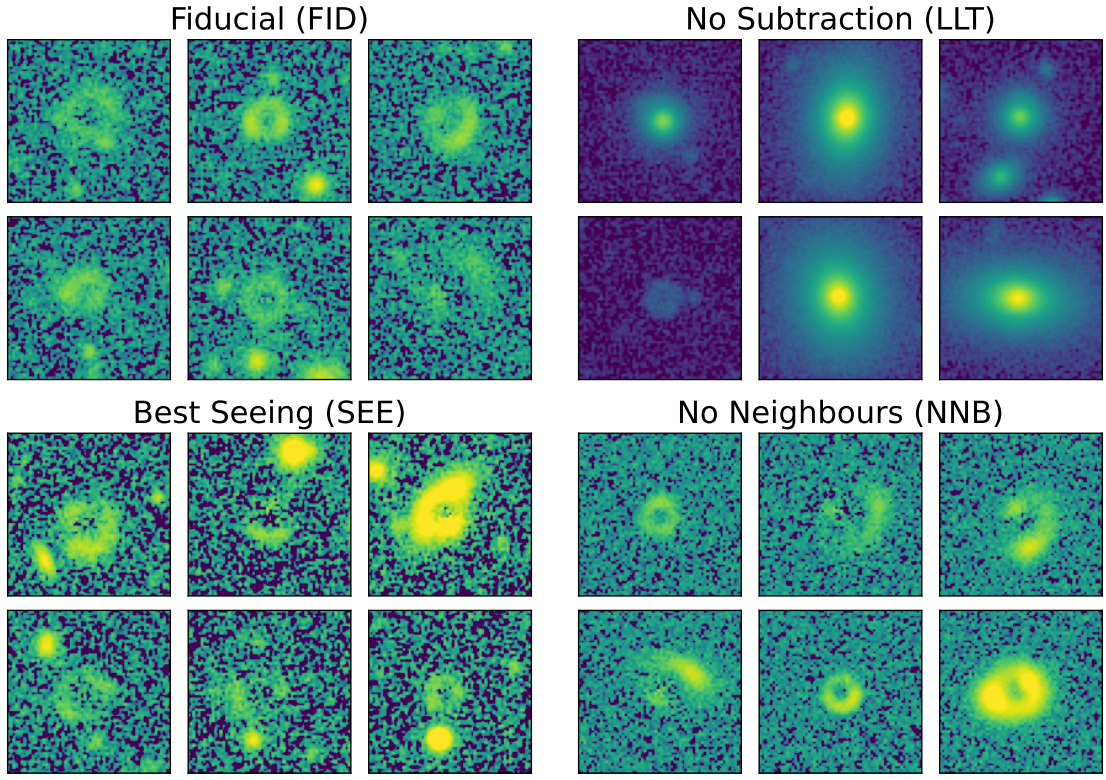


Figure 5.3: Example images used as a test set for each of the 4 networks in this work. Note the `NNB` network did not use cutouts from the LSST DP0.2 simulation, and so does not contain neighbouring (unlensed) objects in the cutout which are visible in the cutouts of the other networks.

For the cosmological inference, I adopted a SIS mass profile. Isothermal mass profiles have been found to fit closely lens galaxies (the 'bulge-halo conspiracy'), although small deviations have been found (e.g., Auger et al., 2010; Etherington et al., 2023). Given the image quality in ground-based surveys, small deviations from an isothermal profile would be difficult to measure directly and, in reality, such uncertainty would need to be marginalised over as part of the inference. In the isothermal case, the Einstein radius (θ_E), velocity dispersion (σ_v), lens and source redshifts (z_L, z_S) are related by

$$r \equiv \frac{D_\mathrm{LS}(z_\mathrm{L}, z_\mathrm{S}, \mathbf{\Omega})}{D_\mathrm{S}(z_\mathrm{S}, \mathbf{\Omega})} = \frac{c^2 \theta_E}{4\pi\sigma_v^2} \tag{5.1}$$

The data vector for the ith lens system, D_i, consists of the lens and source redshifts, ($z_{\mathrm{L,obs},i}$, $z_{\mathrm{S,obs},i}$), and the ratio of Einstein radii/velocity dispersions ($r_{\mathrm{obs},i}$) defined above: $D_i = \{z_{\mathrm{L,obs},i}, z_{\mathrm{S,obs},i}, r_{\mathrm{obs},i}\}$.

Spectroscopic Case

In this chapter, I take 'spectroscopic' to indicate perfect redshift measurements (i.e., with no uncertainty). While in practice spectroscopic measurements have a small uncertainty, the redshift errors for the spectroscopic lens sample are insignificant ($\sim 1\%$) compared to the kinematic and θ_E measurements.

From Bayes' Theorem, the posterior from one system, $P(\mathbf{\Omega}|D_i)$ is proportional to

$$
\begin{aligned}
&P(\mathbf{\Omega}|D_i) \propto P(\mathbf{\Omega}) \cdot P(D_i|\mathbf{\Omega}) \\
&\propto P(\mathbf{\Omega}) \cdot \int\int\int P(r_{\mathrm{obs},i}, z_{\mathrm{L,obs},i}, z_{\mathrm{S,obs},i}, z_{\mathrm{L},i}, z_{\mathrm{S},i}, r_i|\mathbf{\Omega}) \cdot dz_{L,i} dz_{S,i} dr_i \\
&\propto P(\mathbf{\Omega}) \cdot \int\int\int P(r_{\mathrm{obs},i}|r_i) P(z_{\mathrm{L,obs},i}|z_{\mathrm{L},i}) P(z_{\mathrm{S,obs},i}|z_{\mathrm{S},i}) P(r_i|z_{\mathrm{L},i}, z_{\mathrm{S},i}, \mathbf{\Omega}) P(z_{\mathrm{L},i}) P(z_{\mathrm{S},i}) \\
&\cdot dz_{L,i} dz_{S,i} dr_i \\
&\propto P(\mathbf{\Omega}) \cdot P(r_{\mathrm{obs},i}|r(z_{\mathrm{L,True},i}, z_{\mathrm{S,True},i}, \mathbf{\Omega}))
\end{aligned}
$$

where I have used the fact that in the spectroscopic case, $P(z_{\mathrm{obs},i}|z_i) = \delta(z_{\mathrm{obs},i} - z_i)$, $P(z_i) = \delta(z_i - z_{\mathrm{True},i})$ and that $P(r|z_{\mathrm{L}}, z_{\mathrm{S}}, \mathbf{\Omega}) = \delta(r - r(z_{\mathrm{L}}, z_{\mathrm{S}}, \mathbf{\Omega}))$. Given each lens system is independent, the combined posterior for N systems (in each of the scenarios considered here) is given by:

$$
P(\mathbf{\Omega}|\mathbf{D}) \propto P(\mathbf{\Omega}) \prod_i^N P(D_i|\mathbf{\Omega}) \tag{5.2}
$$

Photometric Case

For the photometric scenario, I marginalize over the redshift distributions of each system. The lens population is treated hierarchically in this work. The lens redshift distribution was modelled as a 4-component GMM of the form:

$$
f(x, \{\mu_i\}, \{\sigma_i\}, \{w_i\}) = \sum_{i=1}^{4} \frac{w_i}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \tag{5.3}
$$

where the weights follow $\sum_{i=1}^{4} w_i = 1$. For clarity I group these hyperparameters into a vector $\boldsymbol{\beta_L} = \{\{\mu_i\}, \{\sigma_i\}, \{w_i\}\}$. Such a distribution could fit the redshift distributions while minimizing the number of inferred hyperparameters. I modelled the lens-source redshift relation $P(z_{\mathrm{S},i}|z_{\mathrm{L},i})$ with a redshift-dependent log-normal distribution, with hyperparameters $\boldsymbol{s} = \{\sigma_c, \sigma_m, q_c, q_m\}$ which define a linear relation

with respect to lens redshift ($q_i = q_c + z_{\mathrm{L},i} \cdot q_m$, $\sigma_i = \sigma_c + z_{\mathrm{L},i} \cdot \sigma_m$). The log-normal distribution is given by

$$f(x, \mu, \sigma, q) = \frac{1}{q(x-\mu) \cdot \sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2q^2} \log^2\left(\frac{x-\mu}{\sigma}\right)\right\} \tag{5.4}$$

where I defined $x = z_{\mathrm{S}} - z_{\mathrm{L}}$ and fixed $\mu = 0$ which constrained $z_{\mathrm{L},i} < z_{\mathrm{S},i}$, as required. I found that such a log-normal distribution could fit the source redshift distribution, even though such a distribution was not used in the construction of the lens dataset. This is depicted in Appendix B.1. The posterior for a single system is proportional to

$$\begin{aligned}
&P(\boldsymbol{\Omega}|D_i) \propto P(\boldsymbol{\Omega}) \cdot P(D_i|\boldsymbol{\Omega}) \\
&\propto P(\boldsymbol{\Omega}) \int P(r_{\mathrm{obs},i}, z_{\mathrm{L,obs},i}, z_{\mathrm{S,obs},i}, z_{\mathrm{L},i}, z_{\mathrm{S},i}, r_i, \boldsymbol{\beta_L}, \mathbf{s}|\boldsymbol{\Omega}) \cdot dz_{\mathrm{L},i}\, dz_{\mathrm{S},i}\, dr_i\, d\boldsymbol{\beta_L}\, d\mathbf{s} \\
&\propto P(\boldsymbol{\Omega}) \int P(r_{\mathrm{obs},i}|r_i) P(r_i|z_{\mathrm{L},i}, z_{\mathrm{S},i}, \boldsymbol{\Omega}) P(z_{\mathrm{L,obs},i}|z_{\mathrm{L},i}) P(z_{\mathrm{S,obs},i}|z_{\mathrm{S},i}) P(z_{\mathrm{L},i}|\boldsymbol{\beta_L}) \\
&\cdot P(z_{\mathrm{S},i}|z_{\mathrm{L},i}, \mathbf{s}) P(\boldsymbol{\beta_L}) P(\mathbf{s}) \cdot dz_{\mathrm{L},i}\, dz_{\mathrm{S},i}\, dr_i\, d\boldsymbol{\beta_L}\, d\mathbf{s} \\
&\propto P(\boldsymbol{\Omega}) \int P(r_{\mathrm{obs},i}|r(z_{\mathrm{L},i}, z_{\mathrm{S},i}, \boldsymbol{\Omega})) P(z_{\mathrm{L,obs},i}|z_{\mathrm{L},i}) P(z_{\mathrm{S,obs},i}|z_{\mathrm{S},i}) P(z_{\mathrm{L},i}|\boldsymbol{\beta_L}) \\
&\cdot P(z_{\mathrm{S},i}|z_{\mathrm{L},i}, \mathbf{s}) P(\boldsymbol{\beta_L}) P(\mathbf{s}) \cdot dz_{\mathrm{L},i}\, dz_{\mathrm{S},i}\, d\boldsymbol{\beta_L}\, d\mathbf{s}
\end{aligned} \tag{5.5}$$

Photometric Case, with Contamination

Without spectroscopic confirmation of a lens system, it is possible (or even likely) that false positives will be included in a large sample of lenses candidates, such as those anticipated from LSST and *Euclid*. Without accounting for this possibility, the resulting posterior, $P(\boldsymbol{\Omega}|\mathbf{D})$, would be biased. To account for contamination, one requires probabilities that a given system is a lens, such as those discussed in Chapters 3 and 4. Here I denote the lenses as 'L', non-lenses (false positive) as '$\hat{\mathrm{L}}$', and the binary variable denoting a lens (or not) τ (i.e., $\tau = \mathrm{L}$ or $\tau = \hat{\mathrm{L}}$). I define further parent hyperparameters for the false positive population, $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$, to describe the distributions of r_{obs}, $z_{\mathrm{L,obs}}$, and $z_{\mathrm{S,obs}}$ measurements respectively. These were each parametrised by 4-component GMMs of the same form as Eqn. 5.3 giving sufficiently flexibility to fit the range of distributions. The posterior,

$P(\mathbf{\Omega}|D_i)$, is proportional to:

$$
\begin{aligned}
P(\mathbf{\Omega}|D_i) &\propto P(\mathbf{\Omega}) \cdot P(D_i|\mathbf{\Omega}) \\
&\propto P(\mathbf{\Omega}) \sum_{\tau} P(r_{\mathrm{obs},i}, z_{\mathrm{L,obs},i}, z_{\mathrm{S,obs},i}, \tau|\mathbf{\Omega}) \\
&\propto P(\mathbf{\Omega}) \cdot \\
&\Bigg[\Bigg\{ P_{\mathrm{L},i} \cdot \int P(r_{\mathrm{obs},i}|r(z_{\mathrm{L},i}, z_{\mathrm{S},i}, \mathbf{\Omega}, \mathrm{L}))\ P(z_{\mathrm{L,obs},i}|z_{\mathrm{L},i}) P(z_{\mathrm{S,obs},i}|z_{\mathrm{S},i}) P(z_{\mathrm{L},i}|\boldsymbol{\beta_L}) \\
&\cdot P(z_{\mathrm{S},i}|z_{\mathrm{L},i}, \mathbf{s}) P(\mathbf{s}) P(\boldsymbol{\beta_L}) \cdot dz_{\mathrm{L},i}\, dz_{\mathrm{S},i}\, d\boldsymbol{\beta_L}\, d\mathbf{s} \Bigg\} + \\
&\Bigg\{ (1 - P_{\mathrm{L},i}) \cdot \int P(r_{\mathrm{obs},i}|\boldsymbol{\alpha}, \hat{\mathrm{L}}) \cdot P(z_{\mathrm{L,obs},i}|\boldsymbol{\beta}, \hat{\mathrm{L}}) \cdot P(z_{\mathrm{S,obs},i}|\boldsymbol{\gamma}, \hat{\mathrm{L}}) \cdot P(\boldsymbol{\alpha}) P(\boldsymbol{\beta}) P(\boldsymbol{\gamma}) \\
&\cdot d\boldsymbol{\alpha}\, d\boldsymbol{\beta}\, d\boldsymbol{\gamma} \Bigg\} \Bigg]
\end{aligned}
\tag{5.6}
$$

where I have omitted steps identical to the previous scenarios and defined $P(\tau = \mathrm{L}) \equiv \mathrm{P_L}$. In this chapter I have assumed that the lens probabilities P_{L} are accurate, which could be achieved by using a subset of spectroscopically confirmed systems for calibration. However, if this lens probability calibration was performed using a different selection to that for which the inference was applied, this could alter the lens probabilities. In this case, the PGM presented in Figure 5.4 would need to be adjusted to account for the selection function of the follow-up spectroscopy. For example, this could be done using the probability a given system receives spectroscopic follow-up, i.e., the spectroscopic selection function, $P(\text{follow-up})$, and adjusting P_{L} to

$$
P(\text{Lens}|\text{follow-up})\mathrm{P}(\text{follow-up}) + \mathrm{P}(\text{Lens}|\text{no follow-up}) \cdot \mathrm{P}(\text{no follow-up}) \tag{5.7}
$$

where $P(\text{Lens}|\text{no follow-up})$ could be approximated from the presence/absence of a realistic lens model.

Computational Efficiency

Equation 5.6 includes marginalisation over a large number of parameters, namely the true redshifts $z_{\mathrm{L}}, z_{\mathrm{S}}$ which scale with the number of systems, and hyperparameters modelling the lens and false positive populations ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta_L}$, $\mathbf{s}$). For a realistic

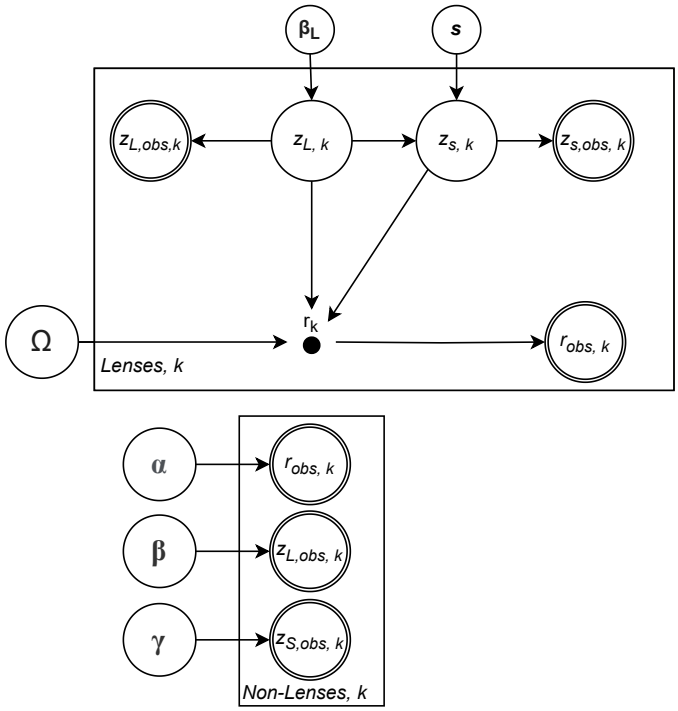


Figure 5.4: Probabilistic Graphical Models (PGM) for the photometric sample formalism, including the effects of contamination. In this diagram, measurements are shown by double circles and inferred or nuisance parameters drawn from a PDF are shown by single circles. However, in the case that this PDF is a delta function (i.e., the r parameter, which for a lens is fixed for given values of $z_{\rm L}, z_{\rm S}$ and $\mathbf{\Omega}$) these variables are depicted with a dot. The plates represent multiple independent systems with the same parametrisation.

number of lens systems, this marginalisation introduces significant memory costs. However, on the assumption that the measurements are independent, the posterior includes the likelihood product of the individual systems which in theory could be parallelised. For example, a dataset could be split into batches, for which samples of the per-batch posteriors are drawn. These could then be combined by using kernel density estimation (KDE) on these samples to obtain weighted samples of the final posterior. This process is non-trivial, as the KDE process is affected by 'curse of dimensionality' with such a large number of hyperparameters. Therefore, I tested whether inferring the hyperparameters $\boldsymbol{\alpha}$,$\boldsymbol{\beta}$,$\boldsymbol{\gamma}$,$\boldsymbol{\beta_L}$,$\mathbf{s}$ separately before treating these as known constants in the cosmology inference would be a plausible solution. I found that such treatment still provides accurate posteriors. Given that the shape of the posterior remained the same, and there was negligible change in the precision, I adopted this method for this work, splitting the photometric datasets into 4 batches.

Parameter	Distribution	Arguments
Ω_m	U(min,max)	(0,1)
Ω_k	U(min,max)	(0,1)
w_0	U(min,max)	(-3,1)
$\alpha(\mu)$	U(min,max)	(0,10)
$\alpha(\sigma)$	Log-U(min,max)	(0.001,2)
$\alpha(w)$	Dirichlet(N_{comp})	4
$\beta(\mu)$	U(min,max)	(0,2)
$\beta(\sigma)$	Log-U(min,max)	(0.001,1)
$\beta(w)$	Dirichlet(N_{comp})	4
$\gamma(\mu)$	U(min,max)	(0,10)
$\gamma(\sigma)$	Log-U(min,max)	(0.001,2)
$\gamma(w)$	Dirichlet(N_{comp})	4
$\beta_L(\mu)$	U(min,max)	(0,2)
$\beta_L(\sigma)$	Log-U(min,max)	(0.001,1)
$\beta_L(w)$	Dirichlet(N_{comp})	4
q_c	U(min,max)	(0.01,1)
q_m	U(min,max)	(-1,1)
σ_c	U(min,max)	(0.1,5)
σ_m	U(min,max)	(0,6)

Table 5.1: Priors used for cosmological inference. Uniform and Log-Uniform distributions are denoted U and Log-U respectively.

The integration over redshifts $z_{\rm L}$, $z_{\rm S}$ in Equation 5.6 cannot be achieved by traditional MCMC sampling (though methods such as reversible jump MCMC, Green, 1995, may allow this), since the posterior term involves the sum of two components $P(D_i|\mathbf{\Omega}, \text{Lens})$ and $P(D_i|\mathbf{\Omega}, \text{Non-Lens})$, of which only one includes integration over the true redshift values. Unlike typical likelihood functions involving the product of terms which could simply be sampled from, the absolute values of the two components is important here as this governs the weighting of the lens and false positive terms. Therefore, I calculated the integral over the lens and source redshifts numerically at each MCMC step. In practice, for each draw of cosmological parameters $\mathbf{\Omega}$ taken by the sampler, the redshift-dependent integrand terms were evaluated over a grid in redshift-space, allowing the likelihood for each system (contained in Eqn. 5.6) to be approximated numerically via the trapezium rule.

To draw samples from the posterior, I used the JAX based `numpyro` package (Phan et al., 2019; Bingham et al., 2019), and the No-U-Turn Sampler (NUTS,

Homan and Gelman, 2014). NUTS is an adaption of Hamiltonian Monte-Carlo (HMC) sampling, which explores a given potential by using the framework of Hamiltonian dynamics. The standard HMC algorithms require a choice of a step-size and the number of steps with which to explore the potential. The NUTS algorithm sets the number of steps taken by the sampler based on when a U-turn is made as it traces the potential. This can produce very efficient sampling, but requires use of the gradient of the potential. Therefore, automatic-differentiation methods such as JAX are required for this to be feasible. Table 5.1 shows the priors used for the cosmological parameters and other hyperparameters.

5.3.3 Generation of Inference Data Vectors

For the purposes of inferring cosmological parameters, I used a larger test set, now including both lenses and false positives, than I used for testing the networks' performances. The COSMIC-BEAMS formalism requires 5 measurements for each system: the Einstein radius, the lens and source redshifts, the lens velocity dispersion and the probability that the system is a strong lens. In each case, I generated realistic measurements/uncertainties to reflect those obtainable in LSST data. Naturally, these differed between the systems designated as being spectroscopic compared to photometric.

I used a 50:50 lens/false positive dataset in the photometric sample. While this ratio would vary depending on the particular lens classifier(s) used and the classifier score threshold chosen, this is an illustrative and realistic level to demonstrate the methodology and the effect of impurities in the sample. I note that from Chapter 3 a 50:50 purity threshold would give a low completeness ($\sim 40\%$); however, given subsequent developments in lens classification (e.g., Cañameras et al., 2024; Schuldt et al., 2025; Euclid Collaboration: Walmsley et al., 2025), a more complete sample than this is likely.

Lens Systems:

We expect to identify $\mathcal{O}(10^5)$ lenses in LSST. I emulated this larger sample by

modelling the parameter space of the test set described in Section 5.2 via KDE (in particular over redshift, velocity dispersion, effective radius and Einstein radius parameters) and drew 100k lens systems from this distribution.

The redshift measurements differed between the photometric and spectroscopic systems. In the spectroscopic case, I treated the redshift uncertainties as negligible (Eq. 5.3.2). For the photometric systems, I used uncertainties of $0.02 \cdot (1 + z)$ based on the LSST Science Requirements Document (Ivezić and The LSST Science Collaboration, 2011), and investigated in Graham et al. (2018).

I combined the observed velocity dispersion and Einstein radius to give a single cosmology-dependent term $r_{\rm True}$ as shown in Figure 5.4. In the photometric case, I assigned an uncertainty to these measurements of $\sigma_r = 0.4$ based on the typical uncertainty resulting from (1) the Einstein radii measurements of the `LLT` network (i.e., without lens subtraction) applied to the LSST test set and (2) the velocity dispersion which could be estimated using photometry alone. This was derived using the typical scatter in the Fundamental Plane, seen for typical lens galaxies from the SLACS survey (Auger et al., 2010). For the spectroscopic systems I used $\sigma_r = 0.085$ based on the uncertainty from the Einstein radii measurements of the `LLT` network and a $10\,\mathrm{km\,s^{-1}}$ uncertainty in the velocity dispersion (chosen to match the value in Li et al., 2024).

False Positive Systems:

To generate a population of false positives (i.e., non-lenses), I used example false positives from multiple lens classifiers applied to Hyper-Suprime Cam (HSC, Aihara et al., 2019; Aihara et al., 2022) survey data, in particular, classifiers used in Chapter 3 (Sonnenfeld et al., 2020; Cañameras et al., 2021; Jaelani et al., 2023; Ishida et al., 2025). Being a ground-based wide-area survey, HSC was the closest available survey to LSST, and thus the false positives and lenses identified previously in this survey are likely to be similar to those in LSST. These classifiers were trained/tested on HSC images (which, in the case of the Citizen Science ('CS') classifier, Sonnenfeld et al., 2020 had the central galaxy light subtracted). In this work I selected the top

0.1% of objects from each classifier from the cross-matched catalogue described in Chapter 3, and from these identified systems with an expert grade of 0, indicating they were not real lenses. i-band cutouts of these systems were downloaded from the HSC archive[4], and pixel-matched to the LSST pixel scale ($0.2''$). I used the performance of the `LLT` network on the false positives (in particular, measurements of their 'Einstein radii') from these classifiers to generate a dataset of realistic measurements for false positives. For automated lens model methods such as NPE, the modeller will always produce lens parameters (such as θ_E measurements) regardless of whether the cutout it is shown contains a lens system or not. Any systematic differences in such measurements between the lens and false positive samples can then be used as part of the inference to distinguish lenses and false positives. Due to the similarity between HSC and LSST, such measurements should be representative of those obtained from automated lens modelling of LSST systems.

The inference data vectors also required redshift measurements. Strong lens searches are often 'targeted', whereby colour and/or magnitude cuts are applied to a galaxy catalogue to identify early-type galaxies to which lens classifiers are subsequently applied. To avoid confusion, here I will refer to the target galaxy of such searches (i.e., the lens galaxy, in the case of a true positive) as the primary object, and any surrounding objects (i.e., the source galaxy, in the case of a true positive) as a secondary object. I assumed that the primary objects for both true and false positives will have the same redshift and velocity dispersion distributions due to the targeted nature of typical lens searches. This assumption for redshifts would not hold for the secondary objects; true positives require $z_{\rm L} < z_{\rm S}$ but false positives do not. For the false positives, I assumed that a secondary object would still be present, be that spiral arms confused for lensed arcs, or neighbouring objects which were not lensed. In practice, systems for which a secondary object could not be identified/analysed would likely be removed from any subsequent analysis so this assumption is realistic. For this work, I assumed that for the false positives the primary and secondary object redshifts were uncorrelated.

[4] https://hsc-release.mtk.nao.ac.jp/doc/

Therefore, for the secondary objects I drew redshift values randomly from the photo-z measurements in the HSC-DEEP galaxy catalogue (Aihara et al., 2022), with the 5-yr LSST i-band depth applied.

In all cases, I assumed the measurements were on average unbiased but imprecise, with representative uncertainties based on the literature. In particular, I assumed measurements of the velocity dispersion from the fundamental plane were correct to within realistic uncertainties described above. In all cases, 'measurements' were drawn from Normal (or truncated-Normal where relevant) distributions, with mean and standard deviation given by the true value and relevant uncertainty respectively.

Choice of Prior for $P_{\rm L}$

The prior lens probability, $P_{\rm L}$ represents the initial confidence that each system is strongly lensed. In practice this would be derived from strong lens finders and/or automated modelling and is discussed further in Section 5.5.2 (see also Chapters 3 and 4). I found that the hyperparameters fixed prior to conducting the cosmological inference required a certain confidence to be determined accurately. In particular, with lower $P_{\rm L}$ values, the uncertainty in the hyperparameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\beta_L}$ and $\mathbf{s}$ becomes more significant. Fixing these parameters as constants during inference could then introduce bias in the inferred cosmology and thus a relatively high degree of confidence in lens/non-lens classification (i.e., $P_{\rm L}$ values close to 0 or 1) was required such that this uncertainty was subdominant. The photometric inference test set consisted of 50 : 50 lenses:false positives, totalling 200 k systems. I assigned the photometric sample values of $P_{\rm thres}$ and $1 - P_{\rm thres}$ in equal proportions. I found $P_{\rm thres} \geq 0.9$ was required for the fixed-hyperparameters discussed in Section 5.3.2 not to bias the resulting cosmological inference and thus fixed $P_{\rm thres} = 0.9$. I did not update the $P_{\rm L}$ values during the inference (i.e., they were kept fixed), to ensure they remained accurately calibrated.

5.4 Results

5.4.1 Lens Modelling of LSST Lenses

I first present the results of the modelling networks on realistic simulated LSST lens systems. Figure 5.5 shows the network recovery of the Einstein radii in the test sets, along with the network uncertainties. The networks achieve a mean precision on θ_E of $3.7, 5.0, 2.2$ and 9.6% for networks `FID`, `SEE`, `NNB` and `LLT` respectively. The best performing network was the `NNB` network, which was the most simple, but least realistic, simulation. As may be expected, the lens systems with the highest precision are those with the highest magnification and brightest source galaxies, which produce more visible lensed arcs. The `FID` network is $2.6\times$ more precise than the `LLT` network; for the inference I use precision values based on the latter network as I use unsubtracted HSC cutouts for the false positive systems, but highlight that the inference could be further improved by using systems with the central galaxy's light subtracted. The network uncertainties are well calibrated, with $69.2 - 72.1\%$ of predicted θ_E values within 1σ of the ground truth (Fig. 5.5). Modelling results for the additional lens parameters are plotted in Appendix B.2 and the precision and mean bias achieved by each network is shown in Table 5.2.

While it may be expected that forming a coadd image of only the best seeing (top 1/3 of) single exposures would improve the model precision (i.e., the `SEE` network), I find that the precision is decreased compared to the `FID` case. This is likely because of the corresponding decrease in depth (by ~ 0.6mag) from reducing the number of single exposures contributing to the coadd image. However, I note that in practice this result may change depending the choice of 'best-seeing' single exposures used to generate the coadd.

I tested the `LLT` network on true A-grade lens systems from HSC which had measurements of θ_E available in the literature (having adjusted the cutout pixel scale of these HSC systems to that of LSST). I used θ_E measurements provided in the SuGOHI catalogue[5] and in HOLISMOKES X (Schuldt et al., 2023b) as

[5] http://www-utap.phys.s.u-tokyo.ac.jp/~oguri/sugohi/ accessed 9/4/2024.

Parameter	Network	Precision	Bias	Bias (σ)
Einstein radius, θ_E ($''$)	Fiducial	0.05	-0.01	-0.2
	Best Seeing	0.07	-0.02	-0.3
	No Neighbours	0.03	-0.008	-0.3
	No Lens Subtraction	0.1	-0.04	-0.3
Density Slope, γ	Fiducial	0.2	0.07	0.4
	Best Seeing	0.2	0.07	0.3
	No Neighbours	0.1	0.07	0.4
	No Lens Subtraction	0.2	0.06	0.3
Lens Shear,γ_1	Fiducial	0.04	-0.001	-0.01
	Best Seeing	0.05	-5e-4	2e-4
	No Neighbours	0.03	-8e-4	-0.02
	No Lens Subtraction	0.06	-0.002	-0.04
Lens Ellipticity, e_1	Fiducial	0.08	-0.001	-0.004
	Best Seeing	0.09	-4e-4	0.003
	No Neighbours	0.05	-9e-4	-0.009
	No Lens Subtraction	0.1	-0.007	-0.06
Lens Center, x (")	Fiducial	0.06	-0.001	-0.02
	Best Seeing	0.08	6e-4	0.02
	No Neighbours	0.04	3e-4	-0.003
	No Lens Subtraction	0.003	-1e-5	0.07

Table 5.2: Mean precision and bias values for each LSST realization. We find there is consistent benefit to performing lens subtraction prior to lens modelling, while the `SEE` network typically performs slightly worse than the `FID` network. Note that all except 'No Lens Subtraction' use Poisson-limited lens subtracted images.

comparison values. In general, there is good agreement as shown in Figure 5.6 which compares the Einstein radii estimated by the `LLT` network with the values from the literature for these A-grade systems.

5.4.2 Lens Modelling of False Positives

I applied the `LLT` network to cutouts of false positives as described in Section 5.3.3 to determine the behaviour of the modeller when shown objects that were not lenses.

Figure 5.7 shows the distribution of network-predicted Einstein radii for the false positives compared to the distribution of predicted values for lens candidates from HSC. The θ_E distribution differs between the false positive and true lens systems and also differs from the network training prior, i.e., when shown a false positive the network does not default to the most likely solution from its training set.

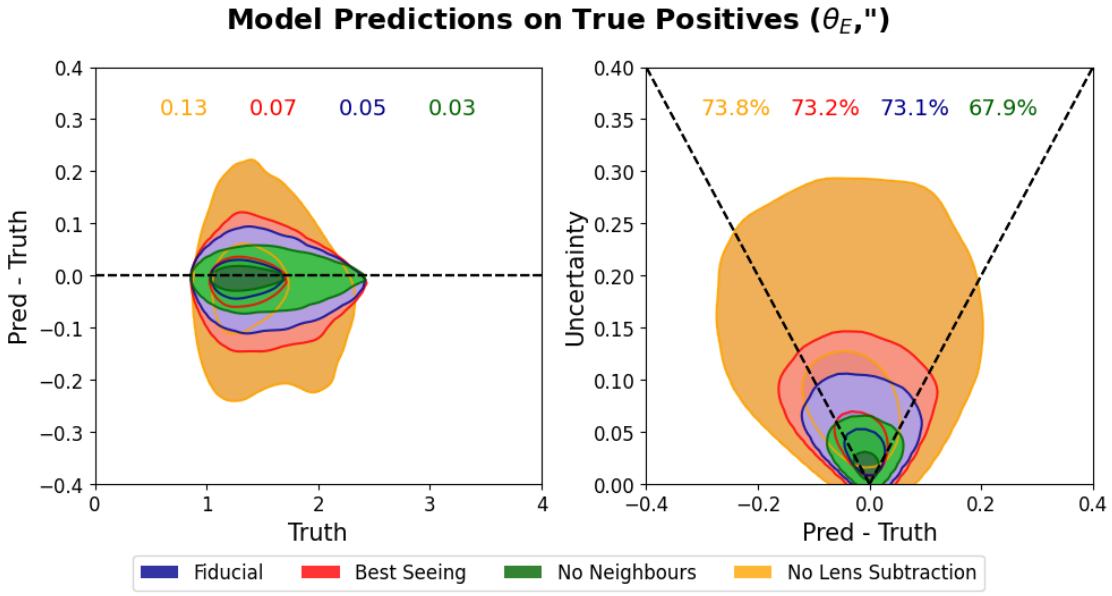


Figure 5.5: A comparison of the performance of the neural networks when trained using different training images. The best performance is seen when not injecting into DP0.2 images, and the second best is seen from the Fiducial model using full-epoch coadds (rather than down-selecting the best-seeing single exposures). In general, the uncertainties produced by the networks are well calibrated. The precision values for each network are highlighted in the left-hand plot, while the proportion of systems with model predictions less than 1-sigma from the truth are listed in the right-hand plots.

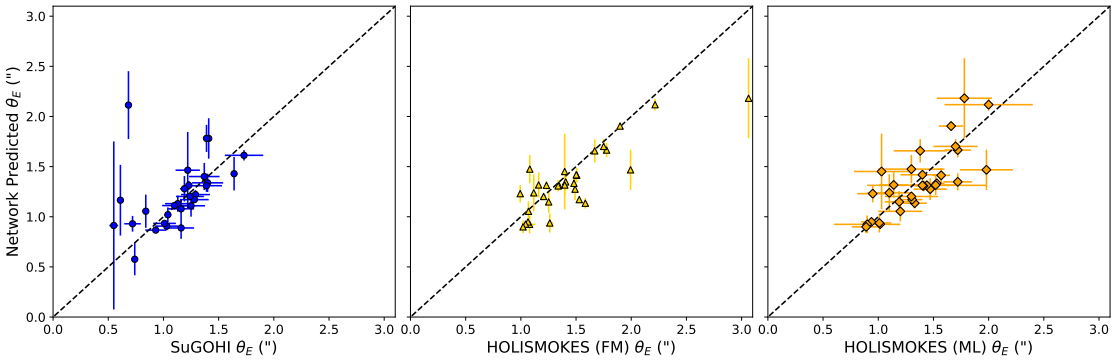


Figure 5.6: Comparison of the θ_E values from the literature (SuGOHI catalogue and Schuldt et al., 2023b) compared to the predictions of the `LLT` network. In the centre and right-hand panels I plot predictions from a machine learning ResNet model ('ML') and the forward-modelling code `GLEE & GLAD` ('FM') from Schuldt et al. (2023b). In general I find good agreement between the `paltas` predictions and literature values, even though the network in this work was not trained on HSC data.

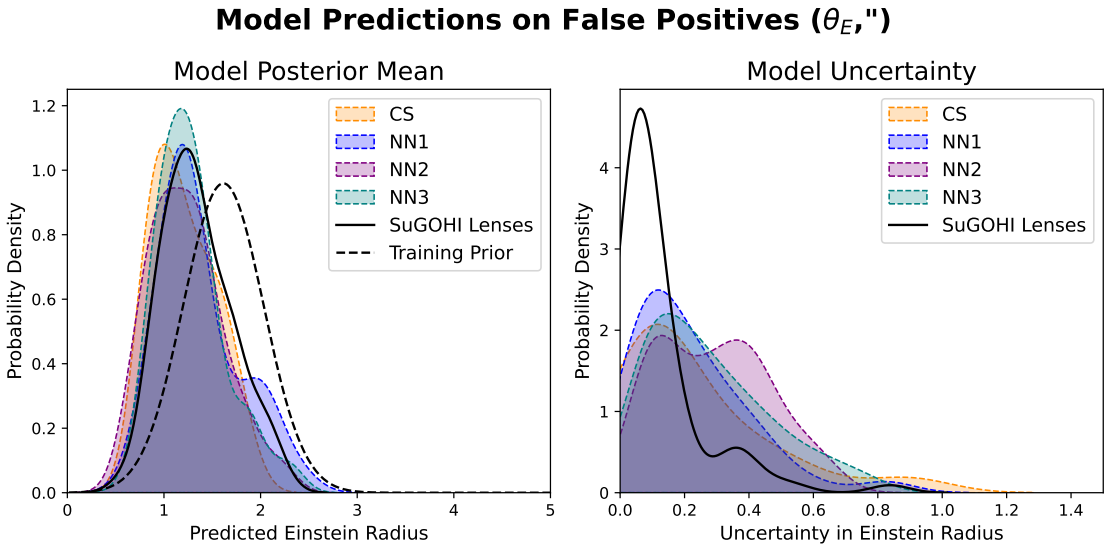


Figure 5.7: Distribution of θ_E values predicted by the `LLT` network for a range of false positives from HSC searches. These false positives were drawn from the following searches: NN1: Cañameras et al., 2021, NN2: Ishida et al., 2025, NN3: Jaelani et al., 2023, CS: Sonnenfeld et al., 2020. The left-hand plot shows the distribution of the network output posterior means, and the right-hand plot shows the network uncertainty distribution. The solid-black curve shows the same distributions, but for A-grade lens candidates in HSC. The dashed curve (left) shows the distribution of Einstein radii from which the training set was drawn.

The network is significantly less confident about its estimate of θ_E when presented with a false positive compared with a lens system. Such differences continue when considering other parameters learned by the network, for example, the network typically predicted shallower mass-density profiles ($\gamma_{\rm lens}$) for false positives (i.e., closer to a mass-sheet) than for true lenses, as shown in Appendix B.3.

5.4.3 Posterior Images for Lenses and False Positives

Given lens parameter posterior distributions, one can reconstruct 'posterior images' by drawing lens parameters from these posteriors. To achieve this, I retrained the `LLT` network with additional parameters (size and magnitudes of the lens and source, and source position). This did not cause a significant change in uncertainty or error in the primary parameters of interest (θ_E and $\gamma_{\rm lens}$). Examples of these posterior images are provided in Appendix B.5, (Fig. B.5 and B.6) for a sample of HSC lens candidates and false positives. The reconstructions for false positive

systems often follow the orientations of objects in the nearby environment, for example, neighbouring galaxies or spiral arms. Reconstructions of lens candidates from HSC replicate the true images well. In the case of a quad (Fig. B.5, row 3, left column), individual draws of the lens parameters produced varied orientations of the lensed image, and the median reconstruction does not have four clear images. This is likely because the training set contained mostly doubly imaged systems and did not contain point sources. Cases of single or double arcs are generally well replicated in the reconstruction (e.g., top left system in Fig. B.5), with scatter originating in the differences in magnification between posterior samples rather than in arc number or orientation.

5.4.4 Cosmological Inference from an Impure Sample of Strong Lenses

The primary dataset for inferring cosmological parameters in this work is composed of both lens and false positive systems. Figure 5.8 shows the cosmological posteriors obtainable from a sample of $100\,000 + 100\,000$ photometric lenses+false positives as well as 5000 spectroscopic systems, and the constraints from combining these datasets. Posteriors from different samples from the inference test set are given in Appendix B.4. I find unbiased cosmological parameters can be obtained even with such significant contamination; the precision and biases found for a range of datasets are shown in Tables 5.3 and 5.4 respectively.

Figure 5.9 shows a comparison of the precision in w which can be obtained for a range of sizes $(625 - 10\,000)$ of the spectroscopic dataset (which will depend on the fraction of systems for which source redshifts can be measured). The contaminated photometric dataset provides roughly equivalent precision in w to 2500 spectroscopic systems while the uncontaminated photometric dataset (i.e., with no false positives) provides improved precision, equivalent to 3500 spectroscopic systems. The precision increases with the larger photometric samples anticipated in later years of the survey. The combined photometric+spectroscopic precision improves upon the spectroscopic precision for all spectroscopic dataset sizes considered (by $\Delta\sigma_w = 0.05$, 0.02, 0.01 for

wCDM	Ω_m	Ω_k	Ω_Λ	w	
Phot. (TP-only)	0.1	0.11	0.03	0.12	/
Phot. (TP+FP)	0.1	0.12	0.03	0.15	/
Spec. (2k)	0.11	0.13	0.02	0.16	/
Spec. (5k)	0.07	0.08	0.01	0.1	/
Spec. (10k)	0.05	0.06	0.01	0.07	/
Spec. (2k) + Phot. (TP+FP)	0.08	0.09	0.02	0.11	/
Spec. (5k) + Phot. (TP+FP)	0.06	0.07	0.01	0.08	/
Spec. (10k) + Phot. (TP+FP)	0.05	0.05	0.01	0.07	/

w_0w_aCDM				w	w_a
Phot. (TP-only)	0.11	0.12	0.04	0.18	0.86
Phot. (TP+FP)	0.11	0.12	0.04	0.2	0.91
Spec. (2k)	0.12	0.13	0.05	0.22	1.0
Spec. (5k)	0.09	0.09	0.05	0.16	0.87
Spec. (10k)	0.08	0.06	0.05	0.12	0.76
Spec. (2k) + Phot. (TP+FP)	0.09	0.09	0.03	0.16	0.84
Spec. (5k) + Phot. (TP+FP)	0.08	0.07	0.03	0.14	0.81
Spec. (10k) + Phot. (TP+FP)	0.07	0.06	0.03	0.12	0.71

Table 5.3: Median precision on cosmological parameters for different datasets and cosmological models.

2 k, 5 k and 10 k spectroscopic systems respectively), with the greatest photometric benefit seen for $N_{\rm spec} \lesssim 5000$.

Cosmological parameter inference can be biased when the hierarchical model is not sufficiently flexible to accurately model the data. In this analysis I have assumed no evolution in the lens sample parameters (e.g., evolution in the mass density slope). This was analysed in detail by Li et al. (2024) who found that assuming no evolution could bias the inferred cosmological parameters if evolution was present in the true sample. Therefore, such a possibility would need to be accounted for when applying this method to real data, for example, by inferring the scatter and evolution in the density slope simultaneously with the cosmological inference (as in Li et al., 2024).

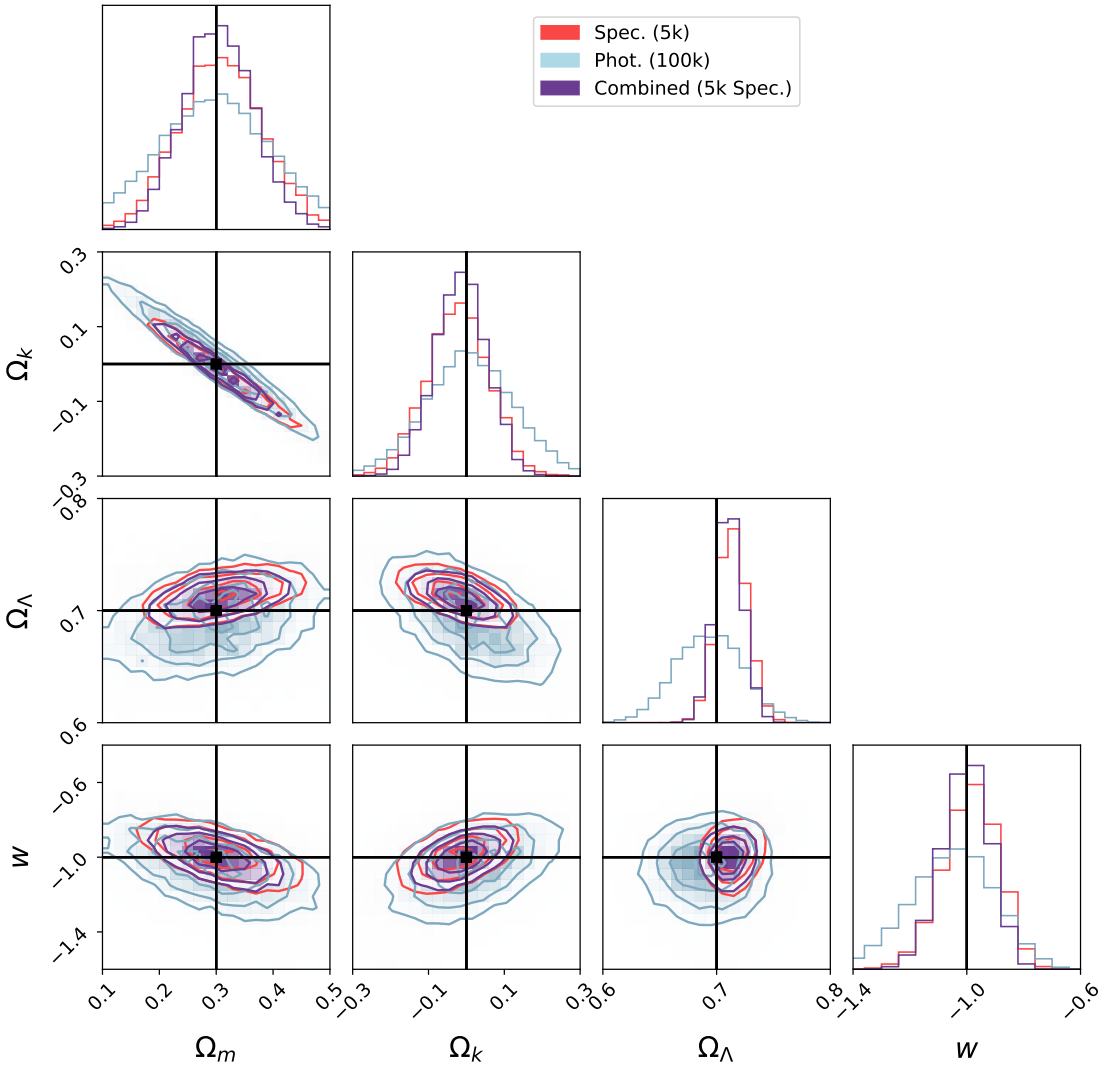


Figure 5.8: Cosmological constraints for photometric (100k TP + 100k FP), spectroscopic (5k TP) and combined photometric + spectroscopic datasets. The ground truth cosmology ($\Omega_m = 0.3, \Omega_k = 0, w = -1$) is shown by the black lines. Unbiased cosmological parameters can be inferred even with a contaminated dataset, and the cosmological precision improves when combining the photometric and spectroscopic datasets.

5.5 Discussion

5.5.1 Modelling of LSST Lens Candidates

In this chapter I trained and tested four different ML networks using realistic LSST-like simulations of i–band images. As shown by Table 5.2, the trained networks can provide accurate measurements of a range of lens parameters, including the Einstein radius, the main focus of this work. The best performing network was the

wCDM	Ω_m	Ω_k	Ω_Λ	w	
Phot. (TP-only)	0.63	0.54	0.26	0.79	/
Phot. (TP+FP)	0.57	0.58	0.41	0.64	/
Spec. (2k)	0.81	0.78	0.64	0.93	/
Spec. (5k)	0.99	0.97	0.76	0.89	/
Spec. (10k)	0.71	0.66	0.52	0.60	/
Spec. (2k) + Phot. (TP+FP)	0.66	0.65	0.67	0.73	/
Spec. (5k) + Phot. (TP+FP)	0.83	0.83	0.77	0.77	/
Spec. (10k) + Phot. (TP+FP)	0.71	0.67	0.58	0.61	/

w_0w_aCDM				w_0	w_a
Phot. (TP-only)	0.98	0.96	0.16	0.79	0.58
Phot. (TP+FP)	0.60	0.59	0.55	0.75	0.71
Spec. (2k)	0.57	0.69	0.38	0.74	0.31
Spec. (5k)	0.76	0.93	0.48	0.87	0.54
Spec. (10k)	0.66	0.63	0.56	0.46	0.59
Spec. (2k) + Phot. (TP+FP)	0.57	0.67	0.49	0.73	0.81
Spec. (5k) + Phot. (TP+FP)	0.85	1.1	0.49	0.94	0.83
Spec. (10k) + Phot. (TP+FP)	0.60	0.61	0.56	0.69	0.65

Table 5.4: Mean absolute bias, in units of σ for the same range of datasets and cosmological models.

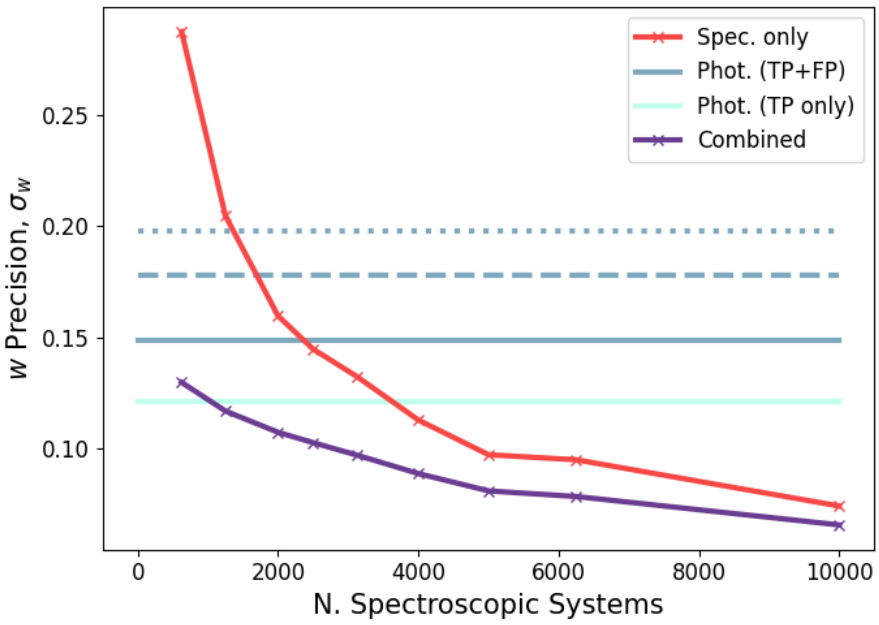


Figure 5.9: The median cosmological precision from a spectroscopic vs photometric dataset. The dotted, dashed and solid blue lines show the photometric sample precision from years 2, 5 and 10 respectively.

`NNB` network. The small decrease in precision between `NNB` and `FID` ($\sigma_{\theta_E} = 0.03''$ versus $\sigma_{\theta_E} = 0.05''$) is indicative of the confusion of the network in identifying lensed sources in cutouts featuring neighbouring objects. It is likely that such confusion would be reduced (and the overall precision improved) by training the networks on multiple wavebands (e.g., shown by a $\sim 20\%$ improvement in accuracy with multi-band modelling by Pearson et al., 2019).

I assumed Poisson-limited lens subtraction in the case of networks `FID`, `SEE` and `NNB` to evaluate the benefits of removing the lens light for lens modelling. Lens subtraction can aid lens classification by reducing the effects of blending, and would likely reveal smaller θ_E systems in LSST images than would otherwise be detectable. While tools are available to do this automatically (e.g., `YATTALENS`, Sonnenfeld et al., 2018), imperfect lens subtraction could potentially bias the inferred lens model parameters. To account for such a possibility, networks trained for real LSST data may benefit from including residual lens light in the training images to ensure network resilience. Notably the `LLT` network could accurately measure the Einstein radius of grade A lens candidates from HSC (Figure 5.6), even though it was not trained on HSC data, suggesting a degree of resilience for 'easy' lens parameters. There was little agreement when compared to literature values for parameters such as ellipticity which may be more sensitive to variations in the PSF.

Beyond the Einstein radius, one parameter of interest was $\gamma_{\rm lens}$, the mass density slope. This was fixed to $\gamma_{\rm lens} = 2$ (i.e., isothermal) in the test set (following the `LensPop` population), but allowed to vary during network training. Given the seeing-limited imaging, $\gamma_{\rm lens}$ was a challenging parameter for the networks to learn, however the median precision in the `FID` network was substantially smaller than the training prior (0.16 versus $\sigma_{\rm train} = 0.26$) indicating it was possible to learn this parameter even in ground-based imaging. Furthermore, while the median precision for the `FID` network was 0.16, this reduced to 0.09 for the brightest 10% of sources which would likely be those graded highest in a lens search.

One benefit of machine-learning lens modelling is speed and thus scalability. Following training ($\sim$ 10 hours), producing models for 100 000 lens systems took $\sim$ 17

minutes on a CPU. Such rapid modelling, along with differences in model behaviour between lenses and false positives (Figure 5.7), could allow for a 'model-informed' classification of lens candidates to distinguish further true lenses from false positives.

I investigated the failure modes of the `FID` network, in particular for which systems the network significantly over/under-predicted the Einstein radius. I found that over-predictions of the Einstein radius typically occurred when the unlensed source magnitude was very faint (~ 27) and the magnification was simultaneously low (~ 3). By contrast, under-predictions of the Einstein radius often occurred when the Einstein radius was very large, where the neighbouring objects in the cutout (from the DP0.2 simulation) could easily be confused with the lensed source. Such outliers reduced significantly (by a factor of $3\times$) in the `NNB` network. However, overall the proportion of such outliers was very low, with 0.36% of test subjects having a fractional error > 0.2, and 0.18% a fractional error < -0.2.

5.5.2 Inference with Impure Samples of Strong Lenses

In this work I have demonstrated that unbiased cosmological parameters can be inferred even with a dataset contaminated by 50% false positives and with photometric redshift uncertainties.

Table 5.3 shows the precision obtained for a range of datasets and cosmology models, with the mean absolute bias shown in Table 5.4. I note that the precision values obtained here are more optimistic than those identified by Li et al. (2024) for a spectroscopic dataset of the same size. This likely derives from the difference in mass model (isothermal versus a power-law profile including anisotropy) utilized by each method, as well as differences in the θ_E precision expected between LSST (this work) and *Euclid* (Li et al., 2024). A more flexible mass model, such as presented in Li et al. (2024) would have the effect of broadening these constraints. This will be presented as part of a complete forecast of the cosmological precision achievable from the combined strong lensing probes (in prep.). However, for context, the constraining power on w for a lens system at $z_{\rm L} = 0.4, z_{\rm S} = 1.8$ (typical redshifts for the LSST population used here), weakens by $\sim 3\%$ when using the PEMD mass model of Li

et al. (2024) versus an isothermal one. Under the current isothermal mass model, the achievable precision from the photometric+spectroscopic dataset ($\sigma_w = 0.07$ for $N_{\rm spec} = 10\,{\rm k}$) is broadly comparable to that from the recent (Flat)-wCDM results of DESI ($\sigma_w = 0.078$, DESI Collaboration et al., 2025b), eBOSS ($\sigma_w = 0.15$, Alam et al., 2021) and DES Year 5 supernovae results ($\sigma_w = 0.15$, DES Collaboration et al., 2024). The combined LSST probes are expected to provide significantly tighter constraints than the individual probes (LSST DESC et al., 2018; Shajib et al., 2024).

The photometric sample provides the greatest improvements in w-precision for $N_{\rm spec} \lesssim 5000$. Without spectroscopy, the year-2 lens sample would provide a precision of $\sigma_w = 0.2$ (again highlighting the above caveats regarding mass model assumptions), and the full 10-year photometric lens sample would be required for such constraints to be competitive with other probes (Figure 5.9). The difference in precision between the contaminated and uncontaminated photometric datasets ($\sigma_w = 0.15$ versus $\sigma_w = 0.12$) demonstrates the case for continuing to improve lens classification methods since increased purity (for the same number of lenses) materially improves the precision. The number of systems for which spectroscopic data is available will depend on a number of factors, including the proportion of lenses detectable in LSST versus those identified in *Euclid*. Lenses located in the overlapping footprint between *Euclid* and LSST will be of interest. The greater image quality of *Euclid* VIS band will improve the lens modelling precision (although this was not the greatest source of uncertainty in this work), while also helping to remove false positives from the photometric lens candidate sample. The *Euclid* data could also help to identify smaller θ_E systems which could otherwise be ambiguous candidates with LSST alone. Finally, these overlapping lens candidates will also benefit from the improved photometry (and photo-z's) from the LSST optical bands.

While cosmology has been the primary focus, this method could be adapted easily to other analyses of the strong lens population. This work involved two primary sources of uncertainty: the velocity dispersion and photometric redshifts. Investigations that did not require either or both of these would gain more significantly from the inclusion of the photometric sample. For example, inferring the evolution of

the power-law slope (which may be best suited to the overlapping data from *Euclid* and LSST) would not require an estimate of the velocity dispersion. This has been discussed at length in the literature (e.g., Ruff et al., 2011; Sonnenfeld et al., 2013; Xu et al., 2017; Remus et al., 2017), and would benefit from the anticipated large photometric dataset from these surveys. I leave such investigations to future work.

Given the size of the photometric dataset, I find any inference is sensitive to small biases in the assumed parent hyperparameters, or correlations in the dataset. The inference presented here makes some simplifying assumptions (for example, combining θ_E and σ_v into a single r_{true} with fixed uncertainty, and fixing the parent hyperparameters prior to cosmological inference) which would need to be managed carefully when applied to real data. However, this work demonstrates that contamination by false positives is a tractable problem with significant potential for population inference.

One assumption of the COSMIC-BEAMS formulation has been the availability of probabilities that each system is a strong lens. Extremes of $P_{\rm L}$ values would be easily obtainable (either from the 4SLSLS survey for confirmed lenses, or expert-inspected false positives). Interim values could be derived from expert-graded (or citizen-graded, see Chapter 4) datasets, using the fraction of these confirmed by spectroscopic follow-up to map these grades to true probabilities that each system is a lens. A broader dataset of expert-graded systems could be obtained using available strong lens databases (e.g., SLED[6] and Masterlens[7], Moustakas, 2012), which could be used to calibrate lens classifiers at scale.

5.6 Conclusion

In this work I have detailed a formulation to infer cosmological parameters from an impure sample of strong lenses. In addition, I have demonstrated the precision with which lens parameters can be measured at LSST-scale using Neural Posterior Estimation. My conclusions are as follows:

[6] https://sled.amnh.org/
[7] http://admin.masterlens.org/index.php

- **Lens Parameter Precision:** I found for the Fiducial (lens-subtracted) dataset, Einstein radii could be measured to a precision of 3.6%. The most accurate measurements were obtained from systems with high magnification and bright source galaxies. Furthermore, model precision was optimised when using full coadd images, rather than when using coadds generated using only single-exposures with the best seeing.

- **Behaviour of NPE network to False Positives:** I found the lens modelling network was more uncertain when presented with systems that were not lenses. Given the rapid speed with which NPE can estimate lens parameters, such differences in behaviour between true lenses and false positives could become a useful tool for automated lens classification, sifting likely lenses from false positives based on their measured lens parameters. For the 'Einstein radius' measurement, the network commonly mis-identified alternative light sources in the cutout to be the Einstein ring when applied to systems which were not strong lenses.

- **Cosmological precision from an impure sample of strong lenses:** Unbiased cosmology can be inferred from an impure sample of strong lenses when the false positive population is accurately accounted for (i.e., via calibrated lens probabilities). The following w-precision would be obtainable for LSST lens candidate systems under wCDM and the assumption of an isothermal mass profile:

  - 100k phot. (TP-only): $\sigma_w = 0.12$
  - 100k+100k phot. (TP+FP): $\sigma_w = 0.15$
  - 100k+100k phot. + 10k spec.: $\sigma_w = 0.07$

  The photometric sample of lenses is comparable to 2500-3500 spectroscopic systems (depending on sample purity), and provides significant improvement in w-precision over the spectroscopic sample alone up to $N_{\rm spec} \sim 5\,000$ systems.

- The primary source of uncertainty in this work derives from the velocity dispersion (which for the photometric dataset is assumed to come from the scatter in the Fundamental Plane). Given the large photometric dataset anticipated from LSST (and similarly *Euclid*), population analysis which did not require a measure of the velocity dispersion would benefit significantly from this dataset, even with contamination up to 50%.

6
Conclusions

The forthcoming decade will revolutionise the field of strong lensing. In this chapter I summarise the main conclusions from the thesis, and outline future work.

In Chapter 2 I produced estimates for the number of detectable lenses in a range of NIR surveys, including JWST, VIDEO and extrapolations to the Euclid Wide Survey. Low detector sensitivity has previously stymied lens searches in the NIR, however *Euclid*, *JWST* and *Roman* will enable large scale searches. Most notably I found that the *JWST* fields observed so far are likely to contain hundreds of lens candidates which is a significant fraction of the total number of strong lenses discovered to date, and will extend the lens population to higher lens and source redshifts. The first evidence of this population has now been identified through the COWLS lens search (Nightingale et al., 2025), and detection of the highest redshift lens known to date (van Dokkum et al., 2024; Mercier et al., 2024; Shuntov et al., 2025). My estimates and these initial results demonstrate that a lens search across the *JWST* archive would be very worthwhile. Given that the COWLS search was conducted by expert inspection, an extended search would easily be in the realms of a citizen science project. Higher redshift lens galaxies will enable us to probe possible evolution in the IMF and mass density profile while higher redshift sources will help to probe the drivers of galaxy evolution out to $z \sim 6$. My estimates for *Euclid* are in agreement with previous forecasts (Collett,

2015; Ferrami and Wyithe, 2024) in finding $\mathcal{O}(10^5)$ lenses will be identified which, when combined with those identified in LSST, will provide a step change in the number of lenses that can be used for population analysis.

In Chapters 3 and 4, I presented the methodology and results for combining multiple lens classifiers into an ensemble as applied to lens searches in HSC, DES and *Euclid* data. This work was motivated by the current performance of lens classifiers and the limited inspection budget of experts required to verify lens candidates. With the arrival of data from the *Euclid* and LSST surveys, current methods would produce a volume of false positives that would be intractable for inspection, leaving only a small proportion of systems suitable for strong lensing science and thus forfeiting a significant part of the opportunity presented by these wide-field surveys. The work in these chapters presents two solutions: firstly, combining multiple classifiers into an ensemble can provide significant improvement in classification performance over the individual classifiers. Secondly, producing calibrated probabilities that each system is strongly lensed allows population-level analysis which doesn't require spectroscopic confirmation for the whole sample. While committees of neural networks have been used previously in strong lens classification, this is the first time in which an initial calibration step has been applied, that a Bayesian method for combining the classifiers has been used, and that multiple different types of classifiers (ML and citizens) have been combined into an ensemble. As such, this represents an important development in lens classification. This work has potential beyond that of strong lensing. Given the intentionally agnostic nature of the calibration and ensemble methodology to the type of classifier, such work could be applied easily in other scientific fields in which classification performance is a limiting factor. It should be highlighted that simply averaging the results of multiple (uncalibrated) classifiers did not yield improved results, and the calibration was an important step required to account for the difference in performance observed between the classifiers. In this work I have primarily used grade A and B lenses as 'true lenses' due to the low numbers of spectroscopically confirmed systems. Therefore, such calibration represents the

probability that the lens candidates are high-grade systems, rather than truly lensed (although these should be highly correlated). Prior to using these probabilities for large-scale analysis, they would need to be rescaled to account for the (likely small) proportion of A/B grade lenses which are false positives which could be achieved using a small calibration set of A/B grade lenses with spectroscopic follow-up. In Chapter 4 I demonstrated that even when using multiple classifiers applied to space-based imaging from *Euclid*, the false positive problem persists. This is likely to reduce with the improvement of ML classifiers prior to DR1, but will not disappear entirely. With $> 100\,000$ lenses anticipated to be detectable in *Euclid*, and the requirement for ≥ 6 experts to inspect each system to account for grade scatter, it may be unfeasible for experts to inspect the complete *Euclid* lens sample, and even doing so would not be a guarantee that the high-grade systems are all strongly lensed. Therefore, it will be necessary to account for the possibility of contamination by false positives when analysing the complete lens sample, as demonstrated subsequently in Chapter 5. While citizen scientists are currently the best performing strong lens classifier, which images the citizens are shown will have a material impact on how many lenses are found in forthcoming *Euclid* lens searches. Combining ML classifiers into an ensemble before showing the highest ensemble-score candidates to citizens will help to maximise the number of lenses identified. Furthermore, given the high performance of citizens, using 'citizen grades' as a proxy for expert grades would remove the expert inspection budget as a limiting factor to having large graded samples of lens candidates. In the coming years, such samples could be used to retrain ML models and we are likely to see an improvement in lens classification performance following such retraining. Ensemble classification will remain useful, however, given the diversity of ML classifiers and the broadly fixed inspection budgets.

In Chapter 5, I investigated the cosmological potential of the photometric sample of strong lenses in LSST. This sample will be much larger than the number of lensed systems with spectroscopy. Given the tight wCDM constraints forecast for the spectroscopic sample, the larger photometric sample warranted investigation. I

began by analysing the precision and accuracy with which detectable strong lenses in LSST may be modelled. I found the main parameter of interest, θ_E, could be inferred accurately and with well-calibrated uncertainties. Furthermore, I demonstrated that increased depth (through full stacking of single exposures) and lens subtraction are worthwhile when modelling, and can significantly improve the modelling precision. I applied one of the trained networks to lens candidates and false positives from HSC and found that the network displayed different behaviours when applied to these two classes. Such a difference alludes to model-informed classification, which could be applied at scale (in the case of NPE or other ML methods) to all lens candidates in *Euclid* and LSST. Given the recent discovery of lens candidates from the *Euclid* Q1 lens search, a compelling extension to this would be to analyse the behaviour of an NPE modelling network to the lens candidates from *Euclid*, as a function of grade and ensemble probability. An evolution in this behaviour could be tuned to provide a further lens classification step, while providing lens model parameters with no additional cost. Furthermore, comparing the modelling results produced by NPE to those produced by forward modelling will be an important further test of the performance of NPE, before applying it to the larger *Euclid* datasets as they become available. Demonstrating that the uncertainties remain accurate and the posterior means remain unbiased will be crucial for larger scale NPE modelling. The remainder of this chapter discussed cosmological inference. The formalism I presented produced unbiased cosmological posteriors from an impure sample of galaxy-galaxy strong lenses. It required lens probabilities for each lens candidate system which could be produced using the methods discussed in the previous chapters. I found the photometric lens sample could provide significant improvement in cosmological precision over the spectroscopic sample alone up to $N_{\rm spec} \sim 5000$. Small differences between cosmological probes (e.g., H_0, S_8), or from the standard model are only revealed with tight constraints; combining the constraining power from all available data will be an important part of this. The combined strong lensing probes are forecast to provide the tightest w_0, w_a constraints in LSST and the photometric sample of strong lenses will help to ensure these

constraints are met. This is the first time that contamination by non-lenses has been investigated, which is a crucial step in utilising the complete photometric sample of lenses that we will identify in the coming years. In the future this analysis will need to be extended to include a more complex mass model and marginalisation over the uncertainties in the mass-density slope which is more difficult to measure with ground-based imaging. The primary uncertainty in this analysis was that in the velocity dispersion, stemming from the scatter in the fundamental plane relation. Analysis which did not require this, such as in the evolution of the power-law slope, would gain the most from the full photometric sample of lenses. I leave this as an important component of future work.

The arrival of *Euclid* DR1 and the first LSST data later in 2025 will fire the starting gun on a new era of strong lens discovery. In the coming months and years I will use the methods described in this thesis to help the collaborative efforts of the *Euclid* and LSST consortia in discovering the largest lens samples known to date. These lens samples will also include a smaller number of rare lens configurations (DSPL's, edge-on spirals, high-redshift lenses) which will merit individual study. The potential of citizen scientists grading lens candidates and masking regions in lens cutouts prior to modelling is an exciting possibility which can be applied at scale with the coming data. I plan to model the complete lens candidate samples from LSST and *Euclid* though NPE. While the modelling conducted in Chapter 5 was undertaken using single-band imaging, extending this to multiple bands would improve modelling precision, as would the combination of LSST and *Euclid* imaging. Finally, I will study the complete photometric lens candidate (i.e., without restriction to those with spectroscopy) which offers significant potential. This will require careful consideration of systematics, in particular in the lens probabilities, which could be tested on early data from LSST or the existing *Euclid* Q1 lens sample.

In this thesis, I have demonstrated the benefits that forthcoming surveys will bring to strong lensing, and that strong lensing will bring to the wider scientific field. I have detailed a number of solutions to the challenges of lens detection and population inference that we will face in the coming years and avenues for

future inquiry. By the close of the decade we will have identified hundreds of thousands of lens candidates, found higher redshift lenses than ever before and used strong lenses to constrain the Hubble constant to within 1%. Furthermore, strong lenses combined with cosmological probes such as Type 1a supernovae, galaxy clustering and 3×2pt data have the potential to validate the tentative results of DESI and thus break the current concordance model of cosmology. This is a truly exciting time for astrophysics.

Appendices

A Supplementary Details on Ensemble Methodology

A.1 Invariance of Dependent Bayesian Method to the Choice of Classifier

Below I demonstrate that the method used to verify the Dependent Bayesian classifier method outlined in Section 3.3.3 is independent of the classifier chosen as C_1. This was verified using `Mathematica`, but here I outline how this was achieved. The equation for a multivariate Normal distribution of dimensions k with mean and covariance $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ is given by

$$N(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-k/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \cdot \Sigma^{-1} \cdot (\mathbf{x} - \boldsymbol{\mu})\right) \tag{A.1}$$

Consider 6 classifiers, defined as $\{C_i\} = \{C_1, C_2, C_3, C_4, C_5, C_6\}$. The difference between score rankings for a particular subject from the ith classifier, R_i, with respect to the rank of that subject from Classifier 1, is given by

$$\mathbf{X} = \begin{pmatrix} X \\ Y \\ Z \\ Q \\ R \end{pmatrix} \equiv \begin{pmatrix} R_2 - R_1 \\ R_3 - R_1 \\ R_4 - R_1 \\ R_5 - R_1 \\ R_6 - R_1 \end{pmatrix} \tag{A.2}$$

The distribution of $\mathbf{X}$ values was parametrised by a multivariate Normal distribution in the Dependent Bayesian method described in Section 3.3.3. These rank-differences with respect to Classifier 1 are linear combinations of the rank differences with respect to Classifier 2. I denote these 'transformed' values with a subscript 'T'

$$\mathbf{X}_T = \begin{pmatrix} X_T \\ Y_T \\ Z_T \\ Q_T \\ R_T \end{pmatrix} \equiv \begin{pmatrix} R_1 - R_2 \\ R_3 - R_2 \\ R_4 - R_2 \\ R_5 - R_2 \\ R_6 - R_2 \end{pmatrix} = \begin{pmatrix} -X \\ Y - X \\ Z - X \\ Q - X \\ R - X \end{pmatrix} \tag{A.3}$$

The covariance matrix, $\mathbf{\Sigma}$, for the multivariate Normal distribution describing the distribution of the differences between classifier ranks with respect to C_1 across all systems can be expressed as

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{xx}^2 & \cdots & \sigma_{rx}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{xr}^2 & \cdots & \sigma_{rr}^2 \end{pmatrix} = \begin{pmatrix} \mu_{xx} - \mu_x^2 & \mu_{xy} - \mu_x\mu_y & \mu_{xz} - \mu_x\mu_z & \mu_{xq} - \mu_q\mu_x & \mu_{xr} - \mu_r\mu_x \\ \mu_{xy} - \mu_x\mu_y & \mu_{yy} - \mu_y^2 & \mu_{yz} - \mu_y\mu_z & \mu_{yq} - \mu_q\mu_y & \mu_{yr} - \mu_r\mu_y \\ \mu_{xz} - \mu_x\mu_z & \mu_{yz} - \mu_y\mu_z & \mu_{zz} - \mu_z^2 & \mu_{qz} - \mu_q\mu_z & \mu_{rz} - \mu_r\mu_z \\ \mu_{xq} - \mu_q\mu_x & \mu_{yq} - \mu_q\mu_y & \mu_{qz} - \mu_q\mu_z & \mu_{qq} - \mu_q^2 & \mu_{qr} - \mu_q\mu_r \\ \mu_{xr} - \mu_r\mu_x & \mu_{yr} - \mu_r\mu_y & \mu_{rz} - \mu_r\mu_z & \mu_{qr} - \mu_q\mu_r & \mu_{rr} - \mu_r^2 \end{pmatrix} \tag{A.4}$$

where I have used μ_{xy} to denote the expectation (mean) value of XY and σ_{xy} to denote the covariance between variables X and Y. The covariance matrix of the transformed multivariate Normal distribution, $\mathbf{\Sigma_T}$, can be written in terms of the untransformed variables, using the fact that e.g., $\sigma^2_{x_T r_T} = \mu(X_T R_T) - \mu(X_T)\mu(R_T)$:

$$\mathbf{\Sigma_T} = \begin{pmatrix} \sigma^2_{x_T x_T} & \cdots & \sigma^2_{r_T x_T} \\ \vdots & \ddots & \vdots \\ \sigma^2_{x_T r_T} & \cdots & \sigma^2_{r_T r_T} \end{pmatrix} \tag{A.5}$$

$$= \begin{pmatrix} \mu_{xx} - \mu_x^2 & (\mu_{xx} - \mu_x^2) - (\mu_{yx} - \mu_x\mu_y) & \dots \\ (\mu_{xx} - \mu_x^2) - (\mu_{xy} - \mu_x\mu_y) & (\mu_{xx} - \mu_x^2) + 2\mu_x\mu_y + (\mu_{yy} - \mu_y^2) - 2\mu_{yx} & \dots \\ (\mu_{xx} - \mu_x^2) - (\mu_{xz} - \mu_x\mu_z) & (\mu_{xx} - \mu_x^2) - (\mu_{xz} - \mu_x\mu_z) - (\mu_{yx} - \mu_x\mu_y) + (\mu_{yz} - \mu_y\mu_z) & \dots \\ (\mu_{xx} - \mu_x^2) - (\mu_{qx} - \mu_q\mu_x) & -(\mu_{qx} - \mu_q\mu_x) + (\mu_{yq} - \mu_q\mu_y) + (\mu_{xx} - \mu_x^2) - (\mu_{yx} - \mu_x\mu_y) & \dots \\ (\mu_{xx} - \mu_x^2) - (\mu_{rx} - \mu_r\mu_x) & -(\mu_{rx} - \mu_r\mu_x) + (\mu_{yr} - \mu_r\mu_y) + (\mu_{xx} - \mu_x^2) - (\mu_{yx} - \mu_x\mu_y) & \dots \end{pmatrix}$$

$$\begin{matrix} \dots & (\mu_{xx} - \mu_x^2) - (\mu_{xz} - \mu_x\mu_z) & \dots \\ \dots & (\mu_{xx} - \mu_x^2) - (\mu_{xz} - \mu_x\mu_z) - (\mu_{yx} - \mu_x\mu_y) + (\mu_{yz} - \mu_y\mu_z) & \dots \\ \dots & (\mu_{xx} - \mu_x^2) - 2(\mu_{xz} - \mu_x\mu_z) + (\mu_{zz} - \mu_z^2) & \dots \\ \dots & -(\mu_{qx} - \mu_q\mu_x) + (\mu_{zq} - \mu_q\mu_z) + (\mu_{xx} - \mu_x^2) - (\mu_{xz} - \mu_x\mu_z) & \dots \\ \dots & -(\mu_{rx} - \mu_r\mu_x) + (\mu_{zr} - \mu_r\mu_z) + (\mu_{xx} - \mu_x^2) - (\mu_{xz} - \mu_x\mu_z) & \dots \end{matrix}$$

$$\begin{matrix} \dots & (\mu_{xx} - \mu_x^2) - (\mu_{qx} - \mu_q\mu_x) & (\mu_{xx} - \mu_x^2) - (\mu_{rx} - \mu_r\mu_x) \\ \dots & -(\mu_{qx} - \mu_q\mu_x) + (\mu_{yq} - \mu_q\mu_y) + (\mu_{xx} - \mu_x^2) - (\mu_{yx} - \mu_x\mu_y) & -(\mu_{rx} - \mu_r\mu_x) + (\mu_{yr} - \mu_r\mu_y) + (\mu_{xx} - \mu_x^2) - (\mu_{yx} - \mu_x\mu_y) \\ \dots & -(\mu_{qx} - \mu_q\mu_x) + (\mu_{zq} - \mu_q\mu_z) + (\mu_{xx} - \mu_x^2) - (\mu_{xz} - \mu_x\mu_z) & -(\mu_{rx} - \mu_r\mu_x) + (\mu_{zr} - \mu_r\mu_z) + (\mu_{xx} - \mu_x^2) - (\mu_{xz} - \mu_x\mu_z) \\ \dots & (\mu_{qq} - \mu_q^2) - 2(\mu_{qx} - \mu_q\mu_x) + (\mu_{xx} - \mu_x^2) & (\mu_{qr} - \mu_q\mu_r) - (\mu_{qx} - \mu_q\mu_x) - (\mu_{rx} - \mu_r\mu_x) + (\mu_{xx} - \mu_x^2) \\ \dots & (\mu_{qr} - \mu_q\mu_r) - (\mu_{qx} - \mu_q\mu_x) - (\mu_{rx} - \mu_r\mu_x) + (\mu_{xx} - \mu_x^2) & (\mu_{rr} - \mu_r^2) - 2(\mu_{rx} - \mu_r\mu_x) + (\mu_{xx} - \mu_x^2) \end{matrix} \Bigg)$$

It can be shown (verified via `Mathematica`) that $(\mathbf{X}-\boldsymbol{\mu})^T \cdot \boldsymbol{\Sigma}^{-1} \cdot (\mathbf{X}-\boldsymbol{\mu}) = (\mathbf{X_T}-\boldsymbol{\mu_T})^T \cdot {\boldsymbol{\Sigma}_\mathbf{T}}^{-1} \cdot (\mathbf{X_T}-\boldsymbol{\mu_T})$, i.e.:

$$\begin{pmatrix} X-\mu_x \\ Y-\mu_y \\ Z-\mu_z \\ Q-\mu_q \\ R-\mu_r \end{pmatrix}^T \cdot \boldsymbol{\Sigma}^{-1} \cdot \begin{pmatrix} X-\mu_x \\ Y-\mu_y \\ Z-\mu_z \\ Q-\mu_q \\ R-\mu_r \end{pmatrix} = \begin{pmatrix} -(X-\mu_x) \\ (Y-\mu_y)-(X-\mu_x) \\ (Z-\mu_z)-(X-\mu_x) \\ (Q-\mu_q)-(X-\mu_x) \\ (R-\mu_r)-(X-\mu_x) \end{pmatrix}^T \cdot {\boldsymbol{\Sigma}_\mathbf{T}}^{-1} \cdot \begin{pmatrix} -(X-\mu_x) \\ (Y-\mu_y)-(X-\mu_x) \\ (Z-\mu_z)-(X-\mu_x) \\ (Q-\mu_q)-(X-\mu_x) \\ (R-\mu_r)-(X-\mu_x) \end{pmatrix} \tag{A.6}$$

The determinants of these covariance matrices, $\det(\boldsymbol{\Sigma})$ and $\det(\boldsymbol{\Sigma_T})$ are also equal (again verified by `Mathematica` using the formulation above). Therefore, since the Gaussian exponents and the determinant's are unchanged when using either the $\mathbf{X}$ or $\mathbf{X_T}$ parametrisation, the multivariate Normal distributions used in the Dependent Bayesian method are independent of the choice of classifier for C_1.

A.2 Relative dependence of HSC and *Euclid* Classifiers

The independent Bayesian method presented in Chapter 3 (Eqn. 3.11) assumes that each classifier is independent of each other. In practice this is true for most classifiers, though classifiers trained on similar datasets can show a higher degree of dependence. Figures A.1 and A.2 show the distribution of ranks of systems shown to each of the classifiers in the HSC and *Euclid* lens searches respectively. The greatest degree of dependency is seen between Neural Networks 2 and 3 (Figure A.1 and Models 2a and 2b (Figure A.2) which were developed by the same team in both cases. Interestingly, there is above-average correlation between the two best performing ML classifiers in Figure A.2 (Models 1 and 4) suggesting they may have both learnt (similar) features which allowed them to perform well.

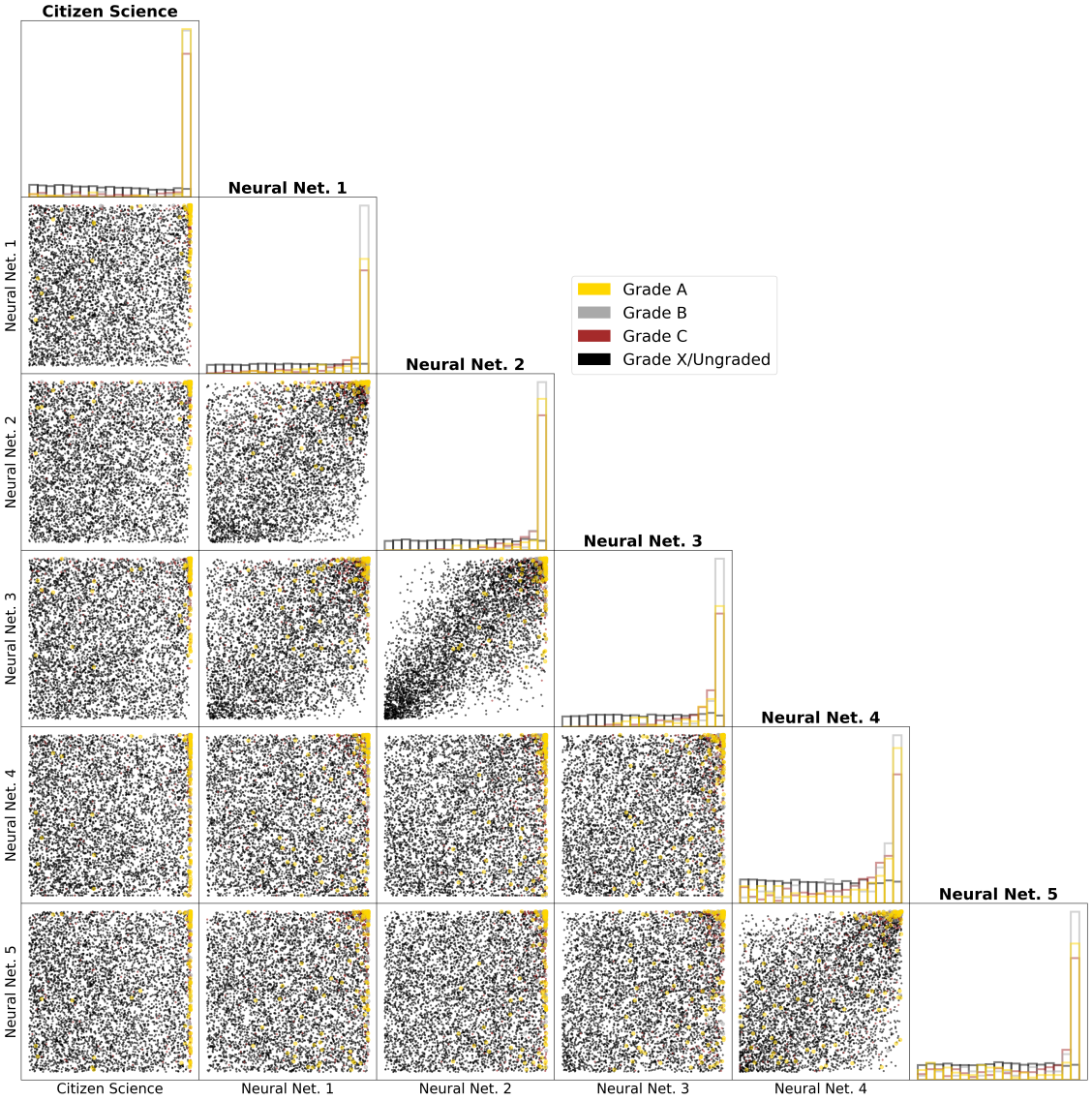


Figure A.1: Scatter plots of the distribution of the ranks of systems used in the HSC ensembles (Chapter 3). A random subset of the Grade X/Ungraded systems are shown, along with all higher grade systems. The histograms are normalised such that the total area of each grade bin are the same.

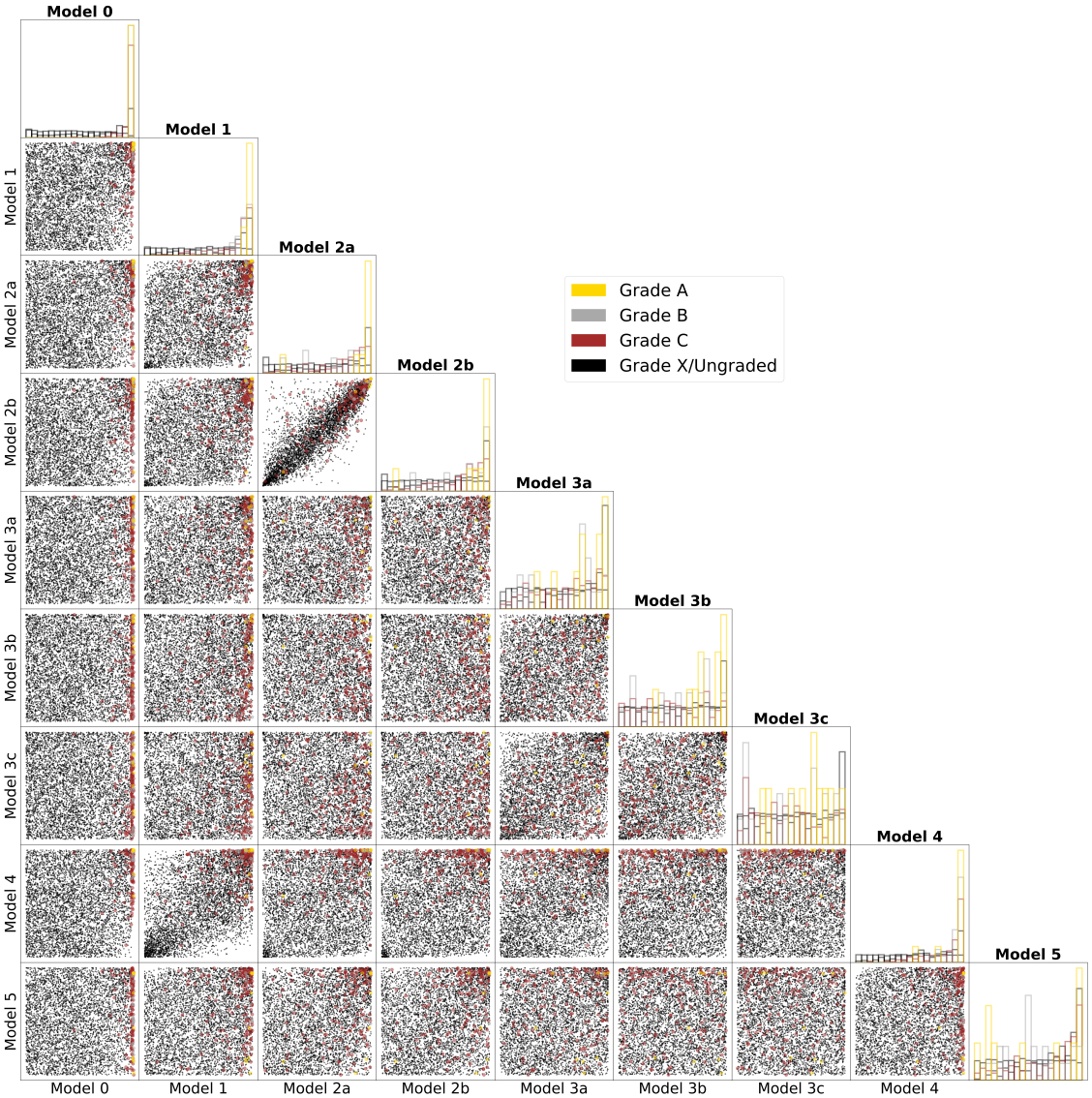


Figure A.2: As above, for the *Euclid* classifiers used in Chapter 4. In this figure, only systems included in the random 40 k subsample shown to the citizen scientists are included in all plots for fair comparison.

B
Supplementary Plots from COSMIC-BEAMS Analysis

B.1 Lens-Source Redshift Dependence for the LSST Lens Population

The COSMIC-BEAMS analysis presented in Chapter 5 assumes a log-normal distribution for the distribution of the lens-source redshift difference, with hyper-parameters which depend linearly on the lens redshift. In Figure B.1, the inferred log-normal parameters are plotted along with the redshift distributions from the detectable LSST lens population. In general, the fit is very good even though such a log-normal distribution was not included in the original population generation, and the deviations seen at high redshift (partly caused by the wider width of the last redshift bin plotted) are not sufficient to bias the resultant inferred cosmology.

B.2 Posteriors for all Lens Parameters

Figure B.2 shows the performance of each network to the remaining lens parameters. Histograms are plotted in the case of $\gamma_{\rm lens}$ and shear, for which the test set systems adopted a single value ($\gamma_{\rm lens} = 2, \gamma = 0$). The `NNB` network consistently has the highest performance except in the case of learning the lens position, in which the

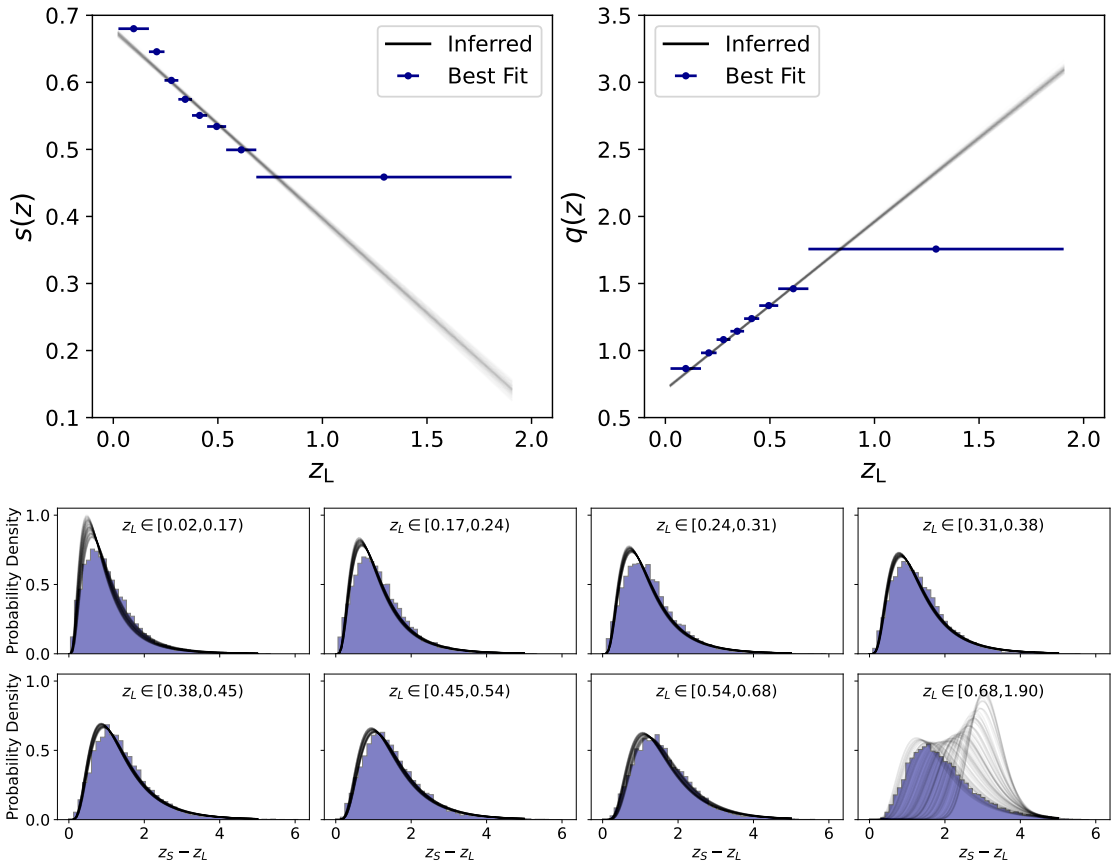


Figure B.1: Top: The inferred redshift dependence of the hyperparameters of Eqn. 5.4 compared to the best fit parameters in each redshift bin, demonstrating the parameters are inferred accurately. The bins are of equal occupancy (i.e., have the same number of systems per bin), hence the last bin is the widest. The error bars indicate the width of each bin. Bottom: The inferred log-normal distribution for each redshift bin, compared to the true distribution of $z_S - z_L$ for the LSST lens population. Overall, the log-normal form provides sufficient flexibility to fit these distributions well. The final, highest redshift, bin is the widest and thus has larger scatter in the inferred log-normal distribution.

`LLT` network unsurprisingly performs better. This demonstrates the importance of using realistic training/test sets, as oversimplified test sets (as known to be the case for `NNB`) can produce over-optimistic results. The `FID` network consistently performs better than the `SEE` and `LLT` networks, demonstrating that signal-to-noise (from greater depth) and lens subtraction are important considerations for optimizing modelling precision.

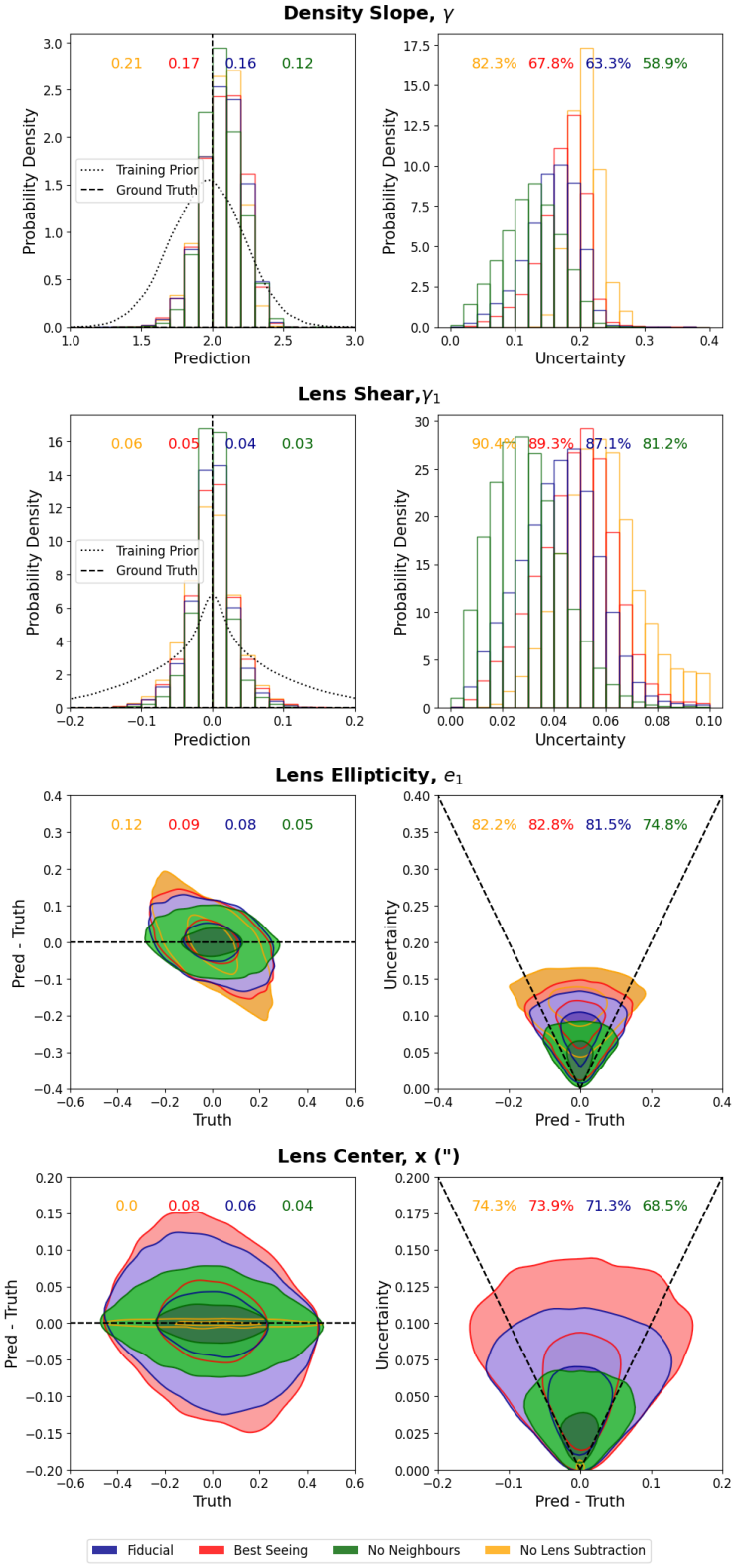


Figure B.2: Model performance for each of the networks across the other learned parameters. The precision values for each network are highlighted in the left-hand plots, while the proportion of systems with model predictions less than 1-sigma from the truth are listed in the right-hand plots.

B.3 Network Behaviour to False Positives across Lens Parameters

Figure B.3 shows the distribution of lens model parameters predicted by the `LLT` network for false positives in comparison to true lenses. Most notable is the difference in predicted $\gamma_{\rm lens}$ values; the network predicts lower $\gamma_{\rm lens}$ values for the false positives than the SuGOHI lens sample. This could be because the network cannot identify lensing features in the false positive cutouts (as expected), and thus reduces the power-law slope to be closer to that of a non-lens (a $\gamma_{\rm lens}$ value of 1 would indicate a mass-sheet, which would not strongly lens a background source). There is also greater uncertainty in the lens shear for the false positives but the remaining distributions are similar to those of the true lenses.

B.4 Cosmological Precision with Different Datasets

Figure B.4 shows the cosmology posteriors for multiple photometric+spectroscopic datasets. These datasets were generated by drawing different samples of photometric ($100\,\mathrm{k} + 100\,\mathrm{k}$ TP+FP) and spectroscopic ($5\,\mathrm{k}$) systems from the inference test set distribution. While there is relatively large scatter between the posteriors, on average they are unbiased. The mean bias values are given in Table 5.4. It is possible that a more complex mass model (or inclusion of evolution in the power-law slope) would increase the uncertainties in these posteriors, as discussed in Sections 5.4.4 and 5.5.2.

B.5 Predicted Posterior Images from `paltas`

Figures B.5, B.6 and B.7 show a sample of generated posterior images for the true A-grade HSC lenses, the false positives and the simulated LSST test set respectively. The Einstein radii from the literature (white) or catalogue ground truth (blue) are shown where relevant, in comparison to the predictions from the network (red annuli). There is large pixel-by-pixel scatter in these posterior images but the median predicted lensed image typically encompasses the lensed arc. Lens systems with clear arcs (e.g., top left Fig. B.5) show good agreement with the original HSC

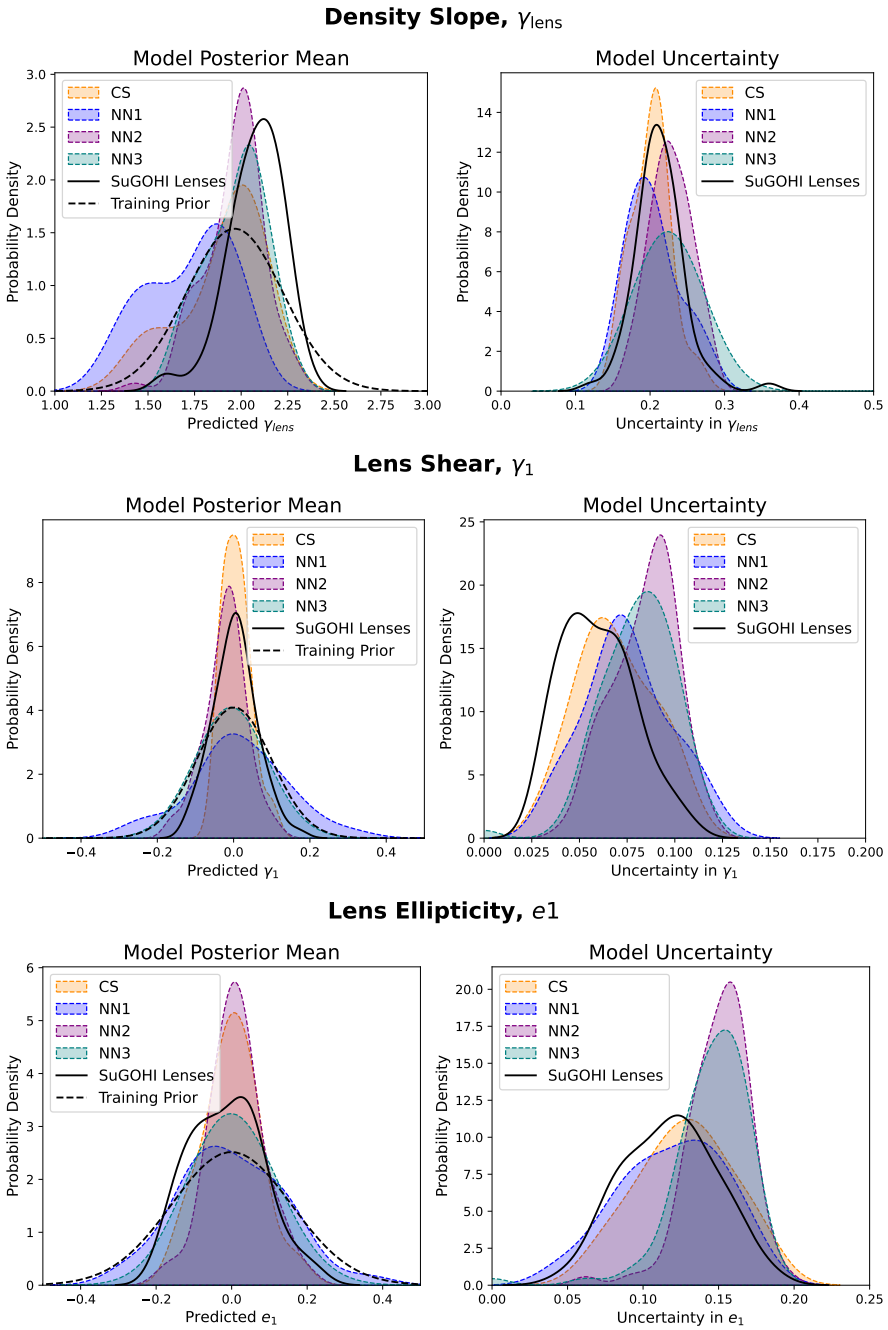


Figure B.3: Comparison of the response of the LLT network to false positives and true lens candidates for additional lens parameters. The false positives were taken from the following searches: NN1: Cañameras et al., 2021, NN2: Ishida et al., 2025, NN3: Jaelani et al., 2023, CS: Sonnenfeld et al., 2020. The distributions of the parameters of the A-grade SuGOHI lenses and training set systems are shown by the solid and dashed black lines respectively.

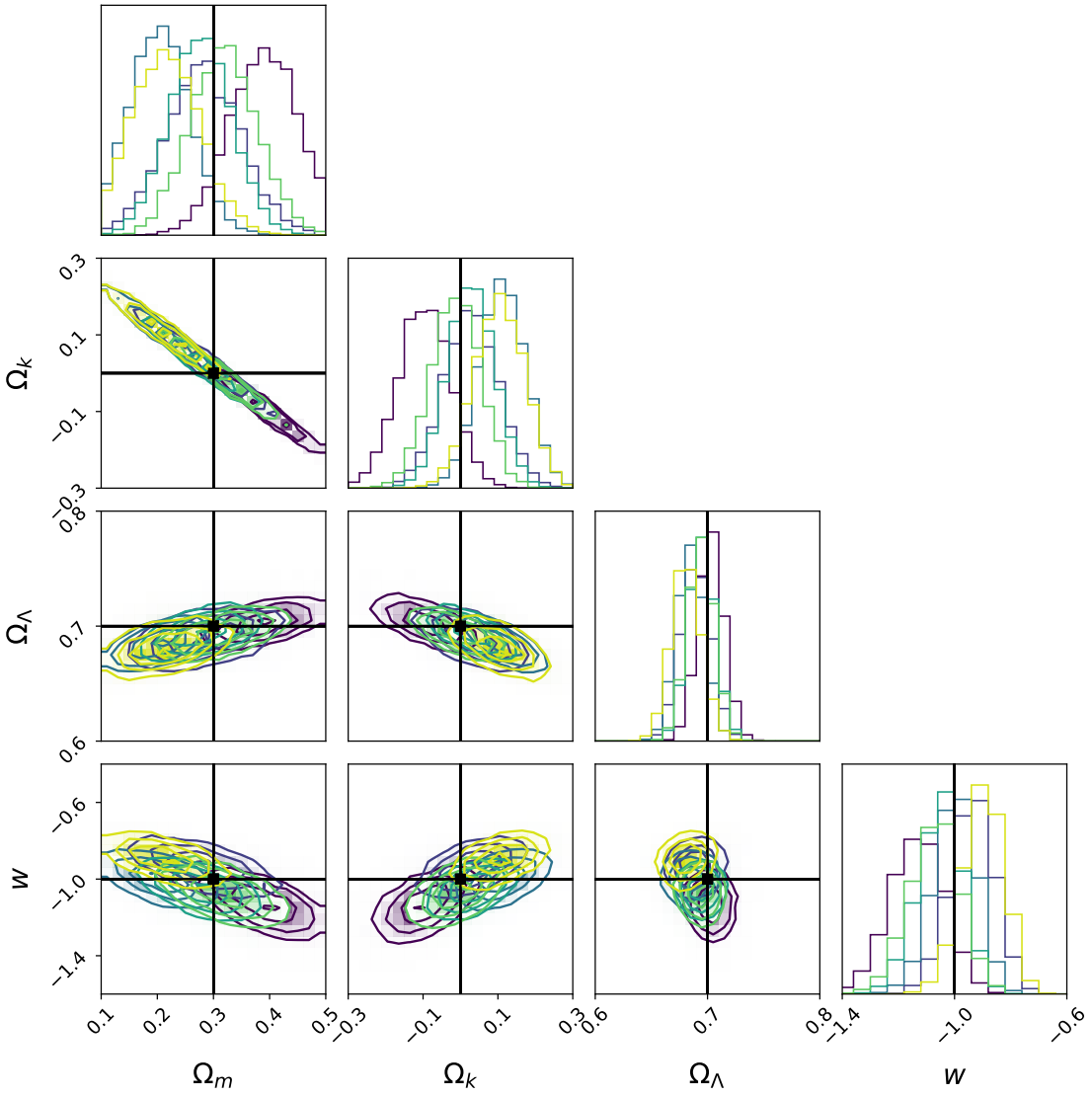


Figure B.4: Cosmological constraints from different random datasets of combined photometric (100k TP + 100k FP) + spectroscopic (5k) lenses. The ground-truth cosmology ($\Omega_m = 0.3, \Omega_k = 0, w = -1$) is shown by the black lines.

cutouts. In the case of the false positives, the per-pixel scatter reveals the network's attempts to recreate the HSC image (such as in row-2,col-1 of Figure B.6) in which neighbouring objects in the field are confused for the lensed source. This produces a systematic difference in behaviour of the network to true lens systems compared to false positives, as demonstrated in Figures 5.7 and B.3.

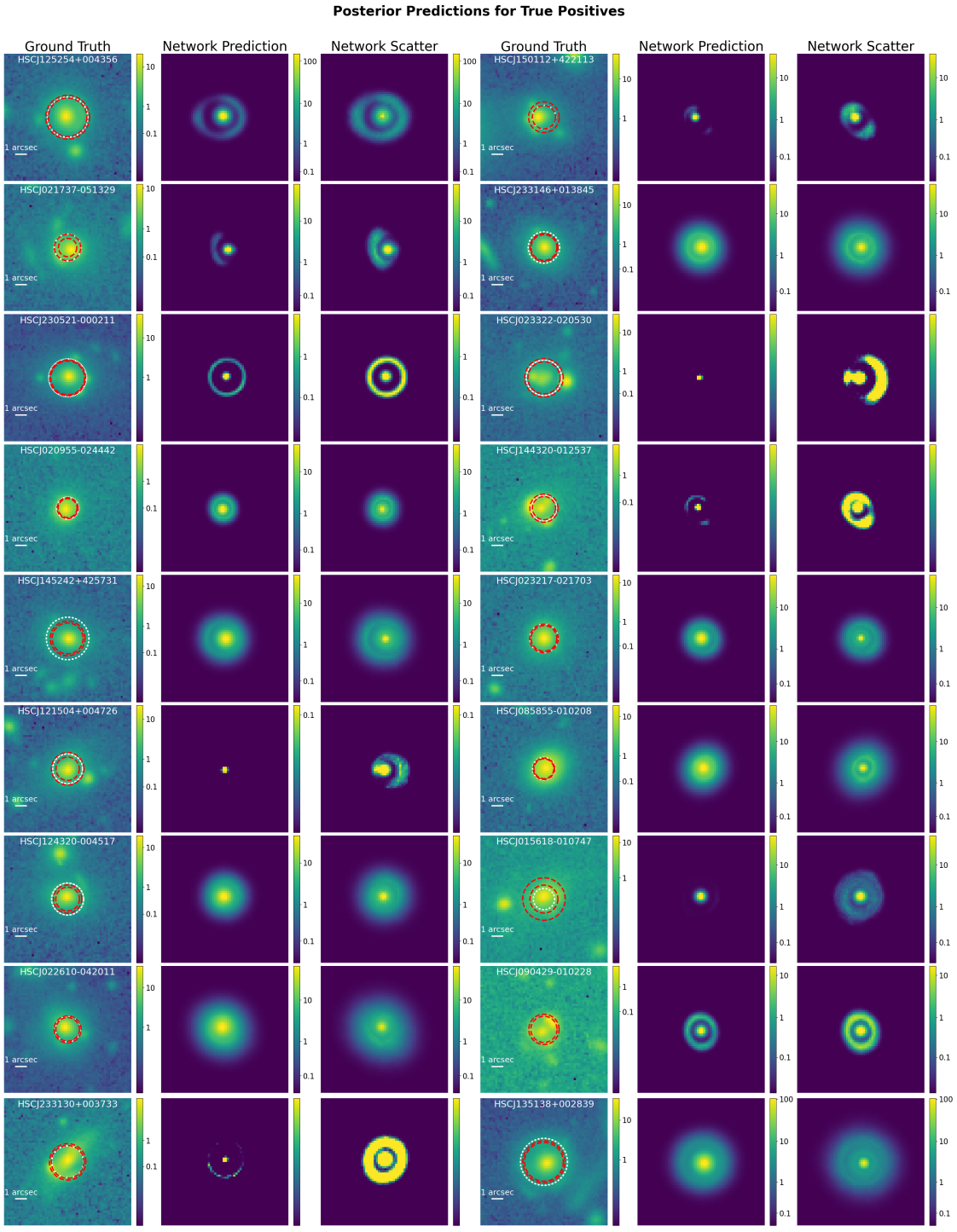


Figure B.5: `paltas` posterior images for lens candidates from HSC, based on random draws from the lens parameter posteriors. Left - Right: Original image, per-pixel median prediction, per-pixel scatter in image prediction. The red annuli show the `LLT` θ_{E} prediction with width 2σ, and the white circles show Einstein radii values drawn from the literature for comparison.

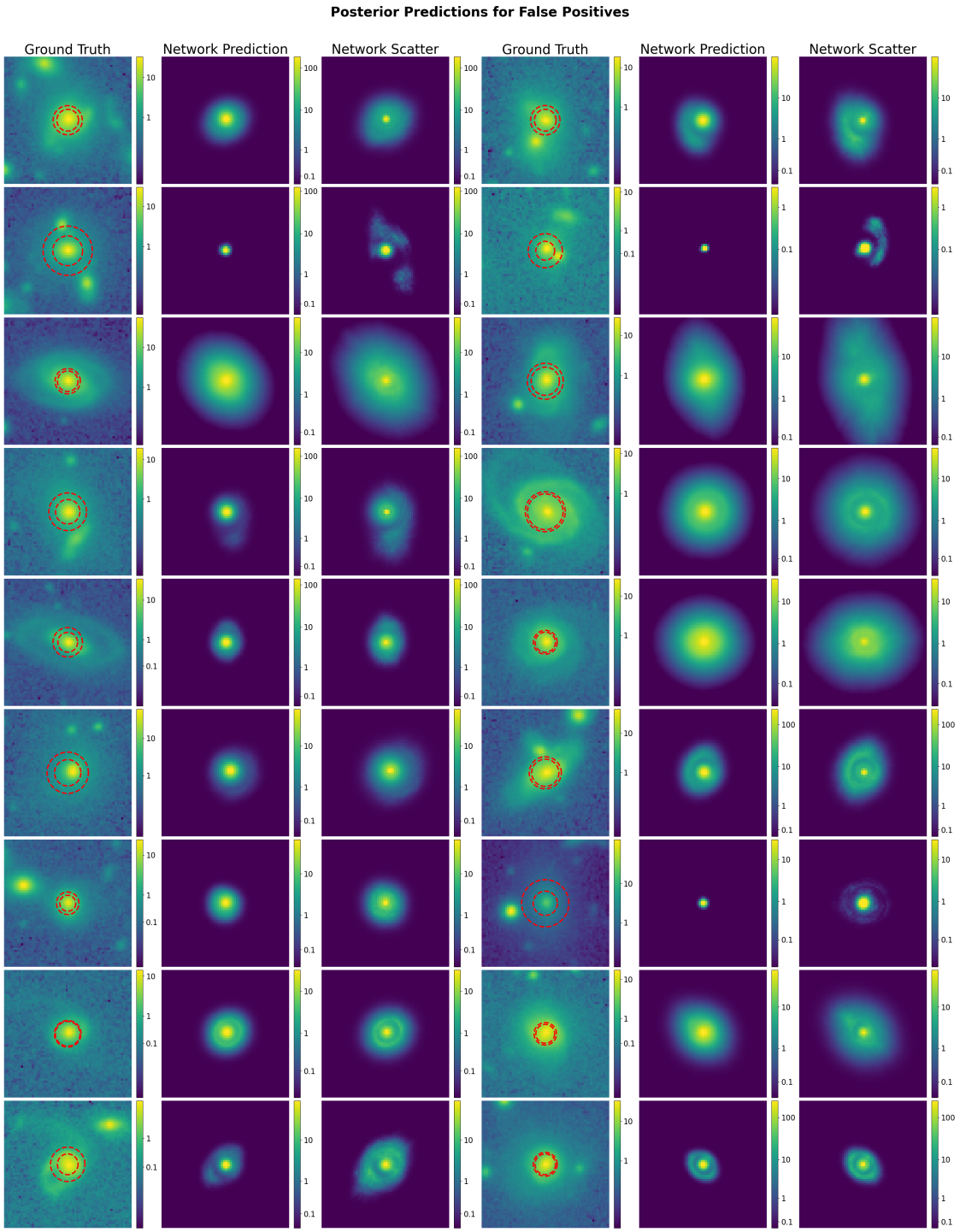


Figure B.6: `paltas` posterior images for false positives (non-lenses) from HSC, based on random draws from the lens parameter posteriors. Left - Right: Original image, per-pixel median prediction, per-pixel scatter in image prediction. The red annuli show the `LLT` Einstein radius prediction with width 2σ.

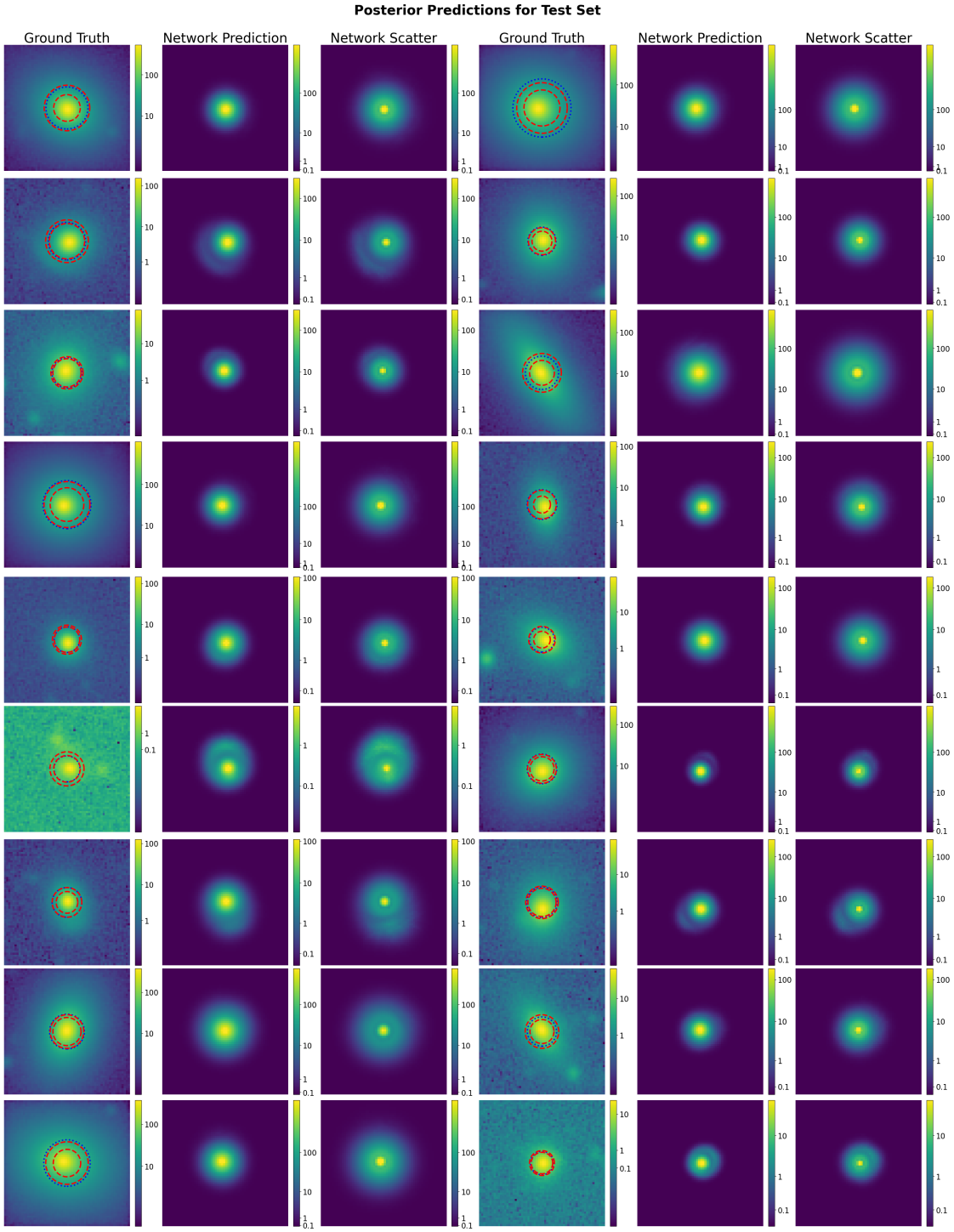


Figure B.7: `paltas` posterior images for the simulated LSST test set based on random draws from the lens parameter posteriors. Left - Right: Original image, per-pixel median prediction, per-pixel scatter in image prediction. In the cutout image, the network predicted Einstein radius is plotted (red) with a 2σ annulus along with the true value (blue).

References

Abbott, T. M. C. et al. (2022). Phys. Rev. D 105.2, 023520.

Abdalla, E. et al. (2022). *Journal of High Energy Astrophysics* 34, pp. 49–211.

Acevedo Barroso, J. A. et al. (2024). *arXiv pre-print*, arXiv:2408.06217.

Adame, A. G. et al. (2025). J. Cosmology Astropart. Phys. 2025.2, p. 021.

Aihara, H. et al. (2018). PASJ 70, S8.

Aihara, H. et al. (2019). PASJ 71.6, p. 114.

Aihara, H. et al. (2022). PASJ 74.2, pp. 247–272.

Alam, S. et al. (2021). Phys. Rev. D 103.8, 083533.

Albrecht, A. et al. (2006). *arXiv pre-print*, astro-ph/0609591.

Alcock, C. et al. (2000). ApJ 542.1, pp. 281–307.

ALMA Partnership et al. (2015). ApJ 808.1, L4.

Alonso, D. et al. (2015). MNRAS 449.1, pp. 670–684.

Andika, I. T. et al. (2023). A&A 678, A103.

Arendse, N. et al. (2024). MNRAS 531.3, pp. 3509–3523.

Auger, M. W. et al. (2010). ApJ 724.1, pp. 511–525.

Auger, M. et al. (2009). ApJ 705.2, pp. 1099–1115.

Barbosa, C. E. et al. (2021). A&A 645, L1.

Barkana, R. (1998). ApJ 502.2, pp. 531–537.

Barkana, R., Blandford, R., and Hogg, D. W. (1999). ApJ 513.2, L91-L94.

Bingham, E. et al. (2019). *J. Mach. Learn. Res.* 20, 28:1–28:6.

Birrer, S. et al. (2020). A&A 643, A165.

Birrer, S. and Amara, A. (2018). *Physics of the Dark Universe* 22, pp. 189–201.

Bolton, A. S. et al. (2006). ApJ 638.2, pp. 703–724.

Bolton, A. S. et al. (2008). ApJ 682.2, pp. 964–984.

Bond, I. A. et al. (2004). ApJ 606.2, L155-L158.

Bouwens, R. J. et al. (2015). ApJ 803.1, p. 34.

Brout, D. et al. (2022). ApJ 938.2, p. 110.

Browne, I. et al. (2003). MNRAS 341.1, pp. 13–32.

Brownstein, J. R. et al. (2012). ApJ 744.1, p. 41.

Calchi Novati, S. et al. (2013). MNRAS 435.2, pp. 1582–1597.

Caminha, G. et al. (2022). A&A 666, L9.

Cañameras, R. et al. (2020). A&A 644, A163.

Cañameras, R. et al. (2021). A&A 653, L6.

Cañameras, R. et al. (2024). A&A 692, A72.

Cao, X. et al. (2025). *arXiv pre-print*, arXiv:2503.08586.

Casey, C. M., Narayanan, D., and Cooray, A. (2014). Phys. Rep. 541.2, pp. 45–161.

Casey, C. M. et al. (2023). ApJ 954.1, p. 31.

Chabrier, G. (2003). PASP 115.809, pp. 763–795.

Chaves-Montero, J. et al. (2016). MNRAS 460.3, pp. 3100–3118.

Chen, T. et al. (2020). *A Simple Framework for Contrastive Learning of Visual Representations*.

Chen, Y. et al. (2019). MNRAS 488.3, pp. 3745–3758.

Chevallier, M. and Polarski, D. (2001). *International Journal of Modern Physics D* 10.2, pp. 213–223.

Choi, Y.-Y., Park, C., and Vogeley, M. S. (2007). ApJ 658.2, pp. 884–897.

Chwolson, O. (1924). *Astronomische Nachrichten* 221, p. 329.

Coe, D. et al. (2013). ApJ 762.1, p. 32.

Collett, T. E. et al. (2012). MNRAS 424.4, pp. 2864–2875.

Collett, T. E. et al. (2023). *The Messenger* 190, pp. 49–52.

Collett, T. E. and Auger, M. W. (2014). MNRAS 443, pp. 969–976.

Collett, T. E. (2015). ApJ 811.1 (1), p. 20.

Collett, T. E. and Smith, R. J. (2020). MNRAS 497.2, pp. 1654–1660.

Connolly, A. J. et al. (2010). In: *Modeling, Systems Engineering, and Project Management for Astronomy IV*. Ed. by G. Z. Angeli and P. Dierickx. Vol. 7738. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 77381O.

Conroy, C. and van Dokkum, P. G. (2012). ApJ 760.1, p. 71.

Contreras, S., Angulo, R. E., and Zennaro, M. (2021). MNRAS 508, pp. 175–189.

Dalal, R. et al. (2023). Phys. Rev. D 108.12, 123519.

de Laplace, P. S. (1796). *Exposition du système du monde*.

DES Collaboration et al. (2024). ApJ 973.1, L14.

DES Collaboration et al. (2025). *arXiv pre-print*, arXiv:2503.13632.

DESI Collaboration et al. (2025a). *arXiv pre-print*, arXiv:2503.14739.

DESI Collaboration et al. (2025b). *arXiv pre-print*, arXiv:2503.14738.

Di Valentino, E. et al. (2021). *Classical and Quantum Gravity* 38.15, 153001.

Domínguez Sánchez, H. et al. (2018). MNRAS 476.3, pp. 3661–3676.

Drakos, N. E. et al. (2022). ApJ 926, p. 194.

Dutton, A. A. and Treu, T. (2014). MNRAS 438.4, pp. 3594–3602.

Efstathiou, G., Sutherland, W. J., and Maddox, S. J. (1990). Nature 348.6303, pp. 705–707.

Einstein, A. (1911). *Annalen der Physik* 340.10, pp. 898–908.

Einstein, A. (1916). *Annalen der Physik* 354.7, pp. 769–822.

Einstein, A. (1915). *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, pp. 844–847.

Einstein, A. (1936). *Science* 84.2188, pp. 506–507.

Erickson, S. et al. (2024). *arXiv pre-print*, arXiv:2410.10123.

ESA Datalabs (2024). https://doi.org/10.1007/978-981-97-0041-7_1.

Etherington, A. et al. (2023). MNRAS 521.4, pp. 6005–6018.

Etherington, A. et al. (2024). MNRAS 531.3, pp. 3684–3697.

Euclid Collaboration: Cropper, M. S. et al. (2024). *arXiv pre-print*, arXiv:2405.13492, arXiv:2405.13492.

Euclid Collaboration: Holloway, P. et al. (2025). *arXiv pre-print*, arXiv:2503.15328.

Euclid Collaboration: Leuzzi, L. et al. (2024). A&A 681, A68.

Euclid Collaboration: Li, T. et al. (2025). *arXiv pre-print*, arXiv:2503.15327.

Euclid Collaboration: Lines, N. E. P. et al. (2025). *arXiv pre-print*, arXiv:2503.15326.

Euclid Collaboration: McCracken, H. J. et al. (2025). *arXiv pre-print*, arXiv:2503.15303, arXiv:2503.15303.

Euclid Collaboration: Mellier, Y. et al. (2024). *arXiv pre-print*, arXiv:2405.13491, arXiv:2405.13491.

Euclid Collaboration: Rojas, K. et al. (2025). *arXiv pre-print*, arXiv:2503.15325.

Euclid Collaboration: Romelli, E. et al. (2025). *arXiv pre-print*, arXiv:2503.15305, arXiv:2503.15305.

Euclid Collaboration: Scaramella, R. et al. (2022). A&A 662, A112.

Euclid Collaboration: Schirmer, M. et al. (2022). A&A 662, A92.

Euclid Collaboration: Walmsley, M. et al. (2025). *arXiv pre-print*, arXiv:2503.15324.

Euclid Quick Release Q1 (2025). https://doi.org/10.57780/esa-2853f3b.

Fassnacht, C. D. et al. (2004). ApJ 600.2, L155-L158.

Faure, C. et al. (2008). *ApJS* 176.1, pp. 19–38.

Ferrami, G. and Wyithe, S. (2024). *arXiv pre-print*, arXiv:2404.03143.

Ferreira, P. G. (2019). ARA&A 57, pp. 335–374.

Fixsen, D. J. et al. (1996). ApJ 473, p. 576.

Fukugita, M. et al. (1990). ApJ 361, L1.

Gaia Collaboration et al. (2016). A&A 595, A1.

Galan, A. et al. (2022a). A&A 668, A155.

Galan, A. et al. (2024). A&A 692, A87.

Galan, A. et al. (2022b). *Herculens: Differentiable gravitational lensing*. Astrophysics Source Code Library, record ascl:2209.002.

Garvin, E. O. et al. (2022). A&A 667, A141.

Gavazzi, R. et al. (2007). ApJ 667.1, p. 176.

Gavazzi, R. et al. (2008). ApJ 677.2, pp. 1046–1059.

Gavazzi, R. et al. (2012). ApJ 761.2, p. 170.

Gavazzi, R. et al. (2014). ApJ 785.2, p. 144.

Gawade, P. et al. (2024). *arXiv pre-print*, arXiv:2404.18897.

Geach, J. E. et al. (2015). MNRAS 452 (1), pp. 502–510.

Geng, S. et al. (2025). A&A 694, A196.

Gentile, F. et al. (2023). MNRAS 522.4, pp. 5442–5455.

Gomer, M. R. et al. (2023). A&A 679, A128.

González, J. et al. (2025). *arXiv pre-print*, arXiv:2501.15679.

Graham, M. L. et al. (2018). AJ 155.1, p. 1.

Grazian, C. and Fan, Y. (2019). *A review of Approximate Bayesian Computation methods via density estimation: inference for simulator-models*.

Green, P. J. (1995). *Biometrika* 82.4, pp. 711–732.

Grillo, C., Lombardi, M., and Bertin, G. (2008). A&A 477.2, pp. 397–406.

Grillo, C. et al. (2009). A&A 501 (2), pp. 461–474.

Gu, A. et al. (2022). ApJ 935.1, p. 49.

Gu, Y. et al. (2020). *PASP* 132 (1011), pp. 1–14.

Hartley, P. et al. (2017). MNRAS 471 (3), pp. 3378–3397.

He, K. et al. (2016a). *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 770.

He, K. et al. (2016b). *Identity Mappings in Deep Residual Networks*.

He, T. et al. (2018). *arXiv pre-print*, arxiv:1812.01187.

He, Z. et al. (2025). ApJ 981.2, p. 168.

Hewitt, J. N. et al. (1988). Nature 333.6173, pp. 537–540.

Hewitt, J. N. et al. (1992). AJ 104, p. 968.

Hezaveh, Y. D., Perreault Levasseur, L., and Marshall, P. J. (2017). Nature 548.7669, pp. 555–557.

Hinshaw, G. et al. (2013). ApJS 208.2, p. 19.

Hlozek, R. et al. (2012). ApJ 752.2, p. 79.

Hogg, N. B. et al. (2025). *arXiv pre-print*, arXiv:2503.08785.

Holloway, P., Marshall, P. J., and Verma, A. (2024a). In: *IAU Symposium*. Ed. by H. Stacey, A. Sonnenfeld, and C. Grillo. Vol. 381. IAU Symposium, pp. 35–37.

Holloway, P. et al. (2023). MNRAS 525.2, pp. 2341–2354.

Holloway, P. et al. (2024b). MNRAS 530.2, pp. 1297–1310.

Holwerda, B. W. et al. (2019). AJ 158.3, p. 103.

Homan, M. D. and Gelman, A. (2014). *J. Mach. Learn. Res.* 15.1, pp. 1593–1623.

Hubble, E. (1929). *Proceedings of the National Academy of Science* 15.3, pp. 168–173.

Huchra, J. et al. (1985). AJ 90, pp. 691–696.

Ishida, Y. et al. (2025). PASJ 77.1, pp. 105–117.

Ivezić, Ž. and The LSST Science Collaboration (2011). *LSST Science Requirements Document, LSST Document LPM-17*. https://ls.st/LPM-17.

Ivezić, Ž. et al. (2019). ApJ 873.2, p. 111.

Jackson, N. (2008). MNRAS 389.3 (3), pp. 1311–1318.

Jacobs, C. et al. (2017). MNRAS 471.1, pp. 167–181.

Jacobs, C. et al. (2019). MNRAS 484.4, pp. 5330–5349.

Jaelani, A. T. et al. (2023). *arXiv pre-print*, arXiv:2312.07333.

Jarvis, M. J. et al. (2013). MNRAS 428 (2), pp. 1281–1295.

Joyce, A., Lombriser, L., and Schmidt, F. (2016). *Annual Review of Nuclear and Particle Science* 66.1, pp. 95–122.

Jullo, E. et al. (2007). *New Journal of Physics* 9.12, p. 447.

Keeton, C. R. (2001). *arXiv pre-print*, astro-ph/0102340.

Keeton, C. R. (2011). *GRAVLENS: Computational Methods for Gravitational Lensing*. Astrophysics Source Code Library, record ascl:1102.003.

Kelvin, L. S. et al. (2012). MNRAS 421 (2), pp. 1007–1039.

Khramtsov, V. et al. (2019). A&A 632, A56.

Kingma, D. P. and Ba, J. (2017). *Adam: A Method for Stochastic Optimization*.

Knabel, S. et al. (2020). AJ 160 (5), p. 223.

Kneib, J.-P. et al. (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.

Koopmans, L. et al. (2009). *ApJL* 703.1, L51-L54.

Korytov, D. et al. (2019). ApJS 245.2, p. 26.

Kunz, M., Bassett, B. A., and Hlozek, R. A. (2007). Phys. Rev. D 75.10, 103508.

Kunz, M. et al. (2013). In: *Astrostatistical Challenges for the New Astronomy*, p. 1013.

Lahav, O. (2002). *Classical and Quantum Gravity* 19.13, pp. 3517–3526.

Lanusse, F. et al. (2018). MNRAS 473.3, pp. 3895–3906.

Lapi, A. et al. (2012). ApJ 755.1, p. 46.

Lemaître, G. (1927). *Annales de la Société Scientifique de Bruxelles* 47, pp. 49–59.

Li, R. et al. (2021). ApJ 923.1, p. 16.

Li, T. et al. (2024). MNRAS 527.3, pp. 5311–5323.

Li, X. et al. (2023). Phys. Rev. D 108.12, 123518.

Liao, K. (2021). ApJ 906.1, p. 26.

Linder, E. V. (2003). Phys. Rev. Lett. 90.9, 091301.

Link, F. (1936). *Bulletin Astronomique* 202, p. 917.

Lintott, C. J. et al. (2008). MNRAS 389.3, pp. 1179–1189.

LSST DESC et al. (2018). *arXiv pre-print*, arXiv:1809.01669.

LSST DESC et al. (2021). ApJS 253.1, p. 31.

Lueckmann, J.-M. et al. (2017). *Flexible statistical inference for mechanistic models of neural dynamics*.

Mac Aodha, O. et al. (2018). *PLoS computational biology* 14.3, e1005995.

Maddox, S. J. et al. (1990). MNRAS 243, pp. 692–712.

Mahler, G. et al. (2025). *arXiv pre-print*, arXiv:2503.08782.

Mandelbaum, R., Van De Ven, G., and Keeton, C. R. (2009). MNRAS 398.2, pp. 635–657.

Mandelbaum, R. et al. (2006). MNRAS 368.2, pp. 715–731.

Manjón-García, a. (2021). PhD thesis. University of Cantabria (Spain).

Maresca, J. et al. (2022). MNRAS 512.2, pp. 2426–2438.

Marshall, P., Blandford, R., and Sako, M. (2005). New A Rev. 49, pp. 387–391.

Marshall, P. J. et al. (2016). MNRAS 455.2, pp. 1171–1190.

Marshall, P. J. et al. (2007). ApJ 671.2, pp. 1196–1211.

Martín-Navarro, I. et al. (2015). MNRAS 447.2, pp. 1033–1048.

Martín-Navarro, I. et al. (2023). MNRAS 521.1, pp. 1408–1414.

McCracken, H. et al. (2012). A&A 544, A156.

Melo, A. et al. (2024). *arXiv pre-print*, arXiv:2411.18694.

Meneghetti, M. (2021). *Introduction to Gravitational Lensing; With Python Examples*. Vol. 956.

Mercier, W. et al. (2024). A&A 687, A61.

Metcalf, R. B. et al. (2019). A&A 625, A119.

Michell, J. (1784). *Philosophical Transactions of the Royal Society of London Series I* 74, pp. 35–57.

Moneti, A. et al. (2023). *VizieR Online Data Catalog*, II/373.

More, A. et al. (2016). MNRAS 455.2, pp. 1191–1210.

Moustakas, L. (2012). *The Master Lens Database and The Orphan Lenses Project*. HST Proposal ID 12833. Cycle 20.

Muzzin, A. et al. (2012). ApJ 761.2, p. 142.

Myers, S. T. et al. (1995). ApJ 447, L5.

Myers, S. et al. (2003). MNRAS 341.1, pp. 1–12.

Navarro, J. F., Frenk, C. S., and White, S. D. (1996). ApJ 462, p. 563.

Negrello, M. et al. (2010). *Science* 330.6005, p. 800.

Newton, I. (1704). *Opticks: Or, A Treatise of the Reflexions, Refractions, Inflexions and Colours of Light. Also Two Treaties of the Species and Magnitude of Curvilinear Figures*. Sam. Smith & Benj. Walford.

Nightingale, J. W., Dye, S., and Massey, R. J. (2018). MNRAS 478.4, pp. 4738–4784.

Nightingale, J. et al. (2025). *arXiv pre-print*, arXiv:2503.08777.

Nightingale, J. et al. (2021). en. *The Journal of Open Source Software* 6.58, p. 2825.

Oesch, P. A. et al. (2018). ApJ 855.2, p. 105.

Oguri, M. (2021). *PASP* 133.1025, 074504.

Oguri, M. and Marshall, P. J. (2010). MNRAS 405.4, pp. 2579–2593.

Oquab, M. et al. (2023). *arXiv pre-print*, arXiv:2304.07193.

Pandey, B. and Sarkar, S. (2015). MNRAS 454.3, pp. 2647–2656.

Papamakarios, G. and Murray, I. (2018). *Fast ϵ-free Inference of Simulation Models with Bayesian Conditional Density Estimation*.

Pascale, M. et al. (2022). *ApJL* 938.1, L6.

Pascale, M. et al. (2025). ApJ 979.1, p. 13.

Patnaik, A. R. et al. (1992). MNRAS 259, 1P-4.

Pawase, R. et al. (2014). MNRAS 439.4, pp. 3392–3404.

Pearson, J., Li, N., and Dye, S. (2019). MNRAS 488.1, pp. 991–1004.

Pearson, J. et al. (2021). MNRAS 505.3, pp. 4362–4382.

Pearson, J. et al. (2024). MNRAS 527.4, pp. 12044–12052.

Percival, W. J. et al. (2001). MNRAS 327.4, pp. 1297–1306.

Perlmutter, S. et al. (1998). Nature 391.6662, pp. 51–54.

Petrillo, C. et al. (2017). MNRAS 472.1, pp. 1129–1150.

Petrillo, C. et al. (2019). MNRAS 484.3, pp. 3879–3896.

Phan, D., Pradhan, N., and Jankowiak, M. (2019).

Planck Collaboration et al. (2020). A&A 641, A6.

Platt, J. (2000). *Adv. Large Margin Classif.* 10.

Poh, J. et al. (2022). *arXiv pre-print*, arXiv:2211.05836.

Poh, J. et al. (2025). *arXiv pre-print*, arXiv:2501.08524.

Remus, R.-S. et al. (2017). MNRAS 464.3, pp. 3742–3756.

Reuter, C. et al. (2020). ApJ 902.1, p. 78.

Rieke, G. H. (2007). ARA&A 45.1, pp. 77–115.

Rieke, M. (2019). *Proceedings of the International Astronomical Union* 15.S352, pp. 337–341.

Rieke, M. J. et al. (2023). ApJS 269.1, p. 16.

Riess, A. G. et al. (1998). AJ 116.3, pp. 1009–1038.

Riess, A. G. et al. (2022). ApJ 934.1, L7.

Roberts, E. et al. (2017). J. Cosmology Astropart. Phys. 2017.10, p. 036.

Rojas, K. et al. (2022). A&A 668, A73.

Rojas, K. et al. (2023). MNRAS 523.3, pp. 4413–4430.

Rubin, V. C., Ford Jr., W. K., and Thonnard, N. (1980). ApJ 238, pp. 471–487.

Rubin, V. C. and Ford Jr., W. K. (1970). ApJ 159, p. 379.

Ruff, A. J. et al. (2011). ApJ 727.2, p. 96.

Rybak, M. et al. (2015). *Monthly Notices of the Royal Astronomical Society: Letters* 451.1, L40-L44.

Schaefer, C. et al. (2018). A&A 611, A2.

Schmidt, T. et al. (2023). MNRAS 518.1, pp. 1260–1300.

Schuldt, S. et al. (2021). A&A 646, A126.

Schuldt, S. et al. (2023a). A&A 671, A147.

Schuldt, S. et al. (2023b). A&A 673, A33.

Schuldt, S. et al. (2025). *arXiv pre-print*, arXiv:2503.07733.

Seidel, G. and Bartelmann, M. (2007). A&A 472.1, pp. 341–352.

Shajib, A. J. et al. (2024). *arXiv pre-print*, arXiv:2406.08919.

Shapiro, I. I. (1964). Phys. Rev. Lett. 13.26, pp. 789–791.

Sheu, W. et al. (2024). *arXiv pre-print*, arXiv:2408.10316.

Shibuya, T., Ouchi, M., and Harikane, Y. (2015). *ApJS* 219 (2).

Shu, Y. et al. (2022). A&A 662, A4.

Shuntov, M. et al. (2025). *arXiv pre-print*, arXiv:2502.20136.

Simonyan, K. and Zisserman, A. (2015). *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun.

Simpson, J. M. et al. (2014). ApJ 788.2, p. 125.

Smith, M. J. et al. (2024). *AstroPT: Scaling Large Observation Models for Astronomy*.

Smith, R. J. and Collett, T. E. (2021). MNRAS 505.2, pp. 2136–2140.

Smoot, G. F. et al. (1992). ApJ 396, L1.

Soldner, J. (1804). *Astronomisches Jahrbuch für 1804*, pp. 161–172.

Sonnenfeld, A. (2022). A&A 659, A132.

Sonnenfeld, A. (2024). A&A 690, A325.

Sonnenfeld, A. (2025). *arXiv pre-print*, arXiv:2501.02054.

Sonnenfeld, A. and Cautun, M. (2021). A&A 651, A18.

Sonnenfeld, A. et al. (2013). ApJ 777.2, p. 98.

Sonnenfeld, A. et al. (2018). *PASJ* 70, S29 (SP1), pp. 29–30.

Sonnenfeld, A. et al. (2019). A&A 630, A71.

Sonnenfeld, A. et al. (2020). A&A 642, A148.

Sonnenfeld, A. et al. (2023). A&A 678, A4.

Spilker, J. S. et al. (2014). ApJ 785.2, p. 149.

Spiniello, C. et al. (2011). MNRAS 417.4, pp. 3000–3009.

Spiniello, C. et al. (2014). MNRAS 438.2, pp. 1483–1499.

Springel, V. (2005). MNRAS 364.4, pp. 1105–1134.

Stacey, H. R. et al. (2018). MNRAS 476.4, pp. 5075–5114.

Sugiyama, M. et al. (2008). *Annals of the Institute of Statistical Mathematics* 60 (4), pp. 699–746.

Sutherland, W. et al. (2015). A&A 575, A25.

Suyu, S. H. and Halkola, A. (2010). A&A 524, A94.

Suyu, S. H. et al. (2012). ApJ 750.1, p. 10.

Swinbank, A. M. et al. (2013). MNRAS 438.2, pp. 1267–1287.

Szegedy, C. et al. (2015). *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 1.

Szegedy, C. et al. (2016). *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2818.

Taniguchi, Y. et al. (2007). ApJS 172.1, pp. 9–28.

Taniguchi, Y. et al. (2009). ApJ 701.2, pp. 915–944.

Taniguchi, Y. et al. (2015). PASJ 67.6, p. 104.

Tewes, M., Courbin, F., and Meylan, G. (2013). A&A 553, A120.

Toft, S. et al. (2014). ApJ 782.2, p. 68.

Tomczak, A. R. et al. (2014). ApJ 783.2, p. 85.

Trenti, M. and Stiavelli, M. (2008). ApJ 676.2, pp. 767–780.

Treu, T. and Koopmans, L. V. E. (2002). MNRAS 337.2, L6-L10.

Treu, T., Suyu, S. H., and Marshall, P. J. (2022). A&A Rev. 30.1, p. 8.

Treu, T. et al. (2006). ApJ 640.2, pp. 662–672.

Treu, T. et al. (2011). MNRAS 417 (3), pp. 1601–1620.

Tsapras, Y. (2018). *Geosciences* 8.10, p. 365.

Unruh, S., Schneider, P., and Sluse, D. (2017). A&A 601, A77.

van Dokkum, P. et al. (2024). *Nature Astronomy* 8.1, pp. 119–125.

Varadaraj, R. G. et al. (2023). MNRAS 524.3, pp. 4586–4613.

Vulcani, B. et al. (2014). MNRAS 441 (2), pp. 1340–1362.

Wagner-Carena, S. et al. (2023). ApJ 942.2, p. 75.

Walmsley, M. et al. (2020). MNRAS 491.2, pp. 1554–1574.

Walmsley, M. et al. (2022). MNRAS 509.3, pp. 3966–3988.

Walmsley, M. et al. (2023). *The Journal of Open Source Software* 8.85, p. 5312.

Walsh, D., Carswell, R. F., and Weymann, R. J. (1979). Nature 279, pp. 381–384.

Weiner, C., Serjeant, S., and Sedgwick, C. (2020). *Research Notes of the AAS* 4.10, p. 190.

Wel, A. V. D. et al. (2012). *ApJS* 203 (2).

Wel, A. V. D. et al. (2014). ApJ 788 (1).

Weymann, R. J. et al. (1980). Nature 285.5767, pp. 641–643.

Wilde, J. et al. (2022). MNRAS 512.3, pp. 3464–3479.

Will, C. M. (1988). *American Journal of Physics* 56.5, pp. 413–415.

Williams, A. C. et al. (2014). In: *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 100–105.

Williams, C. C. et al. (2018). *ApJS* 236 (2), p. 33.

Windhorst, R. A. et al. (2022). AJ 165.1, p. 13.

Wong, K. C. et al. (2020). MNRAS 498.1, pp. 1420–1439.

Wright, A. H. et al. (2025). *arXiv pre-print*, arXiv:2503.19441.

Wright, J. T. and Gaudi, B. S. (2013). "Exoplanet Detection Methods". In: *Planets, Stars and Stellar Systems. Volume 3: Solar and Stellar Planetary Systems*. Ed. by T. D. Oswalt, L. M. French, and P. Kalas, p. 489.

Wyrzykowski, Ł. et al. (2023). A&A 674, A23.

Xie, S. et al. (2017). *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 5987.

Xu, D. et al. (2017). MNRAS 469.2, pp. 1824–1848.

Zadrozny, B. and Elkan, C. (2002). *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Zavala, J. A. et al. (2021). ApJ 909.2, p. 165.

Zitrin, A. et al. (2014). *ApJL* 793.1, L12.

Zwicky, F. (1933). *Helvetica Physica Acta* 6, pp. 110–127.

Zwicky, F. (1937). ApJ 86, p. 217.