

Euclid Quick Data Release (Q1)

XXX. The Strong Lensing Discovery Engine E – Ensemble classification of strong gravitational lenses: lessons for Data Release 1[★]

Euclid Collaboration: P. Holloway^{1,2,★★}, A. Verma², M. Walmsley^{3,4}, P. J. Marshall^{5,6}, A. More⁷, T. E. Collett¹, N. E. P. Lines¹, L. Leuzzi^{8,9}, A. Manjón-García¹⁰, S. H. Vincken¹¹, J. Wilde¹², R. Pearce-Casey¹³, I. T. Andika^{14,15}, J. A. Acevedo Barroso¹⁶, T. Li¹, A. Melo^{15,14}, R. B. Metcalfe^{8,9}, K. Rojas^{11,1}, B. Clément^{16,17}, H. Degaudenzi¹⁸, F. Courbin^{12,19}, G. Despali^{8,9,20}, R. Gavazzi^{21,22}, S. Schuldt^{23,24}, B. C. Nagam^{25,26}, D. Sluse²⁷, C. Tortora²⁸, H. Domínguez Sánchez²⁹, K. Finner³⁰, A. Galan^{14,15}, C. Giocoli^{9,20}, L. Guzzo^{23,31,32}, N. B. Hogg³³, K. Jahnke³⁴, S. Kruk³⁵, G. Mahler^{27,36,37}, M. Millon³⁸, P. Nugent³⁹, J. Pearson¹³, L. R. Ecker^{40,41}, A. Sainz de Murieta¹, C. Scarlata²⁵, S. Serjeant¹³, A. Sonnenfeld⁴², C. Spiniello², T. T. Thai⁴³, L. Ulivi^{44,45,46}, L. Weisenbach¹, M. Zumalacarregui⁴⁷, N. Aghanim⁴⁸, B. Altieri³⁵, A. Amara⁴⁹, S. Andreon³¹, N. Auricchio⁹, H. Aussel⁵⁰, C. Baccigalupi^{51,52,53,54}, M. Baldi^{55,9,20}, A. Balestra⁵⁶, S. Bardelli⁹, P. Battaglia⁹, R. Bender^{41,40}, A. Biviano^{52,51}, A. Bonchi⁵⁷, E. Branchini^{58,59,31}, M. Brescia^{60,28}, J. Brinchmann^{61,62}, S. Camera^{63,64,65}, G. Cañas-Herrera^{66,67,68}, V. Capobianco⁶⁵, C. Carbone²⁴, V. F. Cardone^{69,70}, J. Carretero^{71,72}, M. Castellano⁶⁹, G. Castignani⁹, S. Cavuoti^{28,73}, K. C. Chambers⁷⁴, A. Cimatti⁷⁵, C. Colodro-Conde⁷⁶, G. Congedo⁷⁷, C. J. Conselice⁴, L. Conversi^{78,35}, Y. Copin⁷⁹, H. M. Courtois⁸⁰, M. Cropper⁸¹, A. Da Silva^{82,83}, G. De Lucia⁵², A. M. Di Giorgio⁸⁴, C. Dolding⁸¹, H. Dole⁴⁸, F. Dubath¹⁸, C. A. J. Duncan⁴, X. Dupac³⁵, S. Dusini⁸⁵, A. Ealet⁷⁹, S. Escoffier⁸⁶, M. Farina⁸⁴, R. Farinelli⁹, F. Faustini^{69,57}, S. Ferriol⁷⁹, F. Finelli^{9,87}, S. Fotopoulou⁸⁸, M. Frailis⁵², E. Franceschi⁹, M. Fumana²⁴, S. Galeotta⁵², K. George⁴⁰, B. Gillis⁷⁷, P. Gómez-Alvarez^{89,35}, J. Gracia-Carpio⁴¹, B. R. Granett³¹, A. Grazian⁵⁶, F. Grupp^{41,40}, S. V. H. Haugan⁹⁰, J. Hoar³⁵, W. Holmes⁹¹, F. Hormuth⁹², A. Hornstrup^{93,94}, P. Hudelot²², M. Jhabvala⁹⁵, B. Joachimi⁹⁶, E. Keihänen⁹⁷, S. Kermiche⁸⁶, A. Kiessling⁹¹, B. Kubik⁷⁹, M. Kümmel⁴⁰, M. Kunz⁹⁸, H. Kurki-Suonio^{99,100}, Q. Le Boulc'h¹⁰¹, A. M. C. Le Brun¹⁰², D. Le Mignant²¹, S. Ligi⁶⁵, P. B. Lilje⁹⁰, V. Lindholm^{99,100}, I. Lloro¹⁰³, G. Mainetti¹⁰¹, D. Maino^{23,24,32}, E. Maiorano⁹, O. Mansutti⁵², S. Marcin¹⁰⁴, O. Marggraf¹⁰⁵, M. Martinelli^{69,70}, N. Martinet²¹, F. Marulli^{8,9,20}, R. Massey³⁷, S. Maurogordato¹⁰⁶, E. Medinaceli⁹, S. Mei^{107,108}, M. Melchior¹¹, Y. Mellier^{109,22,†}, M. Meneghetti^{9,20}, E. Merlin⁶⁹, G. Meylan¹⁶, A. Mora¹¹⁰, M. Moresco^{8,9}, L. Moscardini^{8,9,20}, R. Nakajima¹⁰⁵, C. Neissner^{111,72}, R. C. Nichol⁴⁹, S.-M. Niemi⁶⁶, J. W. Nightingale¹¹², C. Padilla¹¹¹, S. Paltani¹⁸, F. Pasian⁵², K. Pedersen¹¹³, W. J. Percival^{114,115,116}, V. Pettorino⁶⁶, S. Pires⁵⁰, G. Polenta⁵⁷, M. Poncet¹¹⁷, L. A. Popa¹¹⁸, L. Pozzetti⁹, F. Raison⁴¹, R. Rebolo^{76,119,120}, A. Renzi^{121,85}, J. Rhodes⁹¹, G. Riccio²⁸, E. Romelli⁵², M. Roncarelli⁹, R. Saglia^{40,41}, Z. Sakr^{122,123,124}, D. Sapone¹²⁵, B. Sartoris^{40,52}, J. A. Schewtschenko⁷⁷, P. Schneider¹⁰⁵, A. Secroun⁸⁶, G. Seidel³⁴, S. Serrano^{126,127,128}, P. Simon¹⁰⁵, C. Sirignano^{121,85}, G. Sirri²⁰, L. Stanco⁸⁵, J. Steinwagner⁴¹, P. Tallada-Crespí^{71,72}, A. N. Taylor⁷⁷, I. Tereno^{82,129}, S. Toft^{130,131}, R. Toledo-Moreo¹³², F. Torradeflot^{72,71}, I. Tutusaus¹²³, E. A. Valentijn²⁶, L. Valenziano^{9,87}, J. Valiviita^{99,100}, T. Vassallo^{40,52}, G. Verdoes Kleijn²⁶, A. Veropalumbo^{31,59,58}, Y. Wang¹³³, J. Weller^{40,41}, A. Zacchei^{52,51}, G. Zamorani⁹, F. M. Zerbi³¹, E. Zucca⁹, V. Allevato²⁸, M. Ballardini^{134,135,9}, M. Bolzonella⁹, E. Bozzo¹⁸, C. Burigana^{136,87}, R. Cabanac¹²³, A. Cappi^{9,106}, D. Di Ferdinando²⁰, J. A. Escartin Vigo⁴¹, L. Gabarra², M. Huertas-Company^{76,137,138,139}, J. Martín-Fleitas¹¹⁰, S. Matthew⁷⁷, N. Mauri^{75,20}, A. A. Nucita^{140,141,142}, A. Pezzotta^{143,41}, M. Pöntinen⁹⁹, C. Porciani¹⁰⁵, I. Risso¹⁴⁴, V. Scottez^{109,145}, M. Sereno^{9,20}, M. Tenti²⁰, M. Viel^{51,52,54,53,146}, M. Wiesmann⁹⁰, Y. Akrami^{147,148}, S. Anselmi^{85,121,149}, M. Archidiacono^{23,32}, F. Atrio-Barandela¹⁵⁰, C. Benoist¹⁰⁶,

[★] This publication has been made possible by the participation of more than a thousand volunteers in the Space Warps project. Their contributions are individually acknowledged at <https://www.zooniverse.org/projects/aprajita/space-warps-esa-euclid/about/team>

^{★★} Corresponding author: philip.holloway@port.ac.uk

[†] Deceased.

K. Benson⁸¹, P. Bergamini^{23,9}, D. Bertacca^{121,56,85}, M. Bethermin¹⁵¹, A. Blanchard¹²³, L. Blot^{152,149}, S. Borgani^{153,51,52,53,146}, M. L. Brown⁴, S. Bruton¹⁵⁴, A. Calabro⁶⁹, F. Caro⁶⁹, C. S. Carvalho¹²⁹, T. Castro^{52,53,51,146}, F. Cogato^{8,9}, A. R. Cooray¹⁵⁵, O. Cucciati⁹, S. Davini⁵⁹, F. De Paolis^{140,141,142}, G. Desprez²⁶, A. Díaz-Sánchez¹⁰, J. J. Diaz¹³⁷, S. Di Domizio^{58,59}, J. M. Diego²⁹, P.-A. Duc¹⁵¹, A. Enia^{55,9}, Y. Fang⁴⁰, A. G. Ferrari²⁰, P. G. Ferreira², A. Finoguenov⁹⁹, A. Fontana⁶⁹, A. Franco^{141,140,142}, K. Ganga¹⁰⁷, J. García-Bellido¹⁴⁷, T. Gasparetto⁵², V. Gautard¹⁵⁶, E. Gaztanaga^{128,126,1}, F. Giacomini²⁰, F. Gianotti⁹, G. Gozaliasi^{157,99}, C. M. Gutierrez¹³⁷, A. Hall⁷⁷, W. G. Hartley¹⁸, C. Hernández-Monteagudo^{120,76}, H. Hildebrandt¹⁵⁸, J. Hjorth¹¹³, J. J. E. Kajava^{159,160}, Y. Kang¹⁸, V. Kansal^{161,162}, D. Karagiannis^{134,163}, K. Kiiveri⁹⁷, C. C. Kirkpatrick⁹⁷, J. Le Graet⁸⁶, L. Legrand^{164,165}, M. Lembo^{134,135}, F. Lepori¹⁶⁶, G. Leroy^{36,37}, G. F. Lesci^{8,9}, J. Lesgourgues¹⁶⁷, T. I. Liaudat¹⁶⁸, A. Loureiro^{169,170}, J. Macias-Perez¹⁷¹, G. Maggio⁵², M. Magliocchetti⁸⁴, E. A. Magnier⁷⁴, F. Mannucci⁴⁶, R. Maoli^{172,69}, C. J. A. P. Martins^{173,61}, L. Maurin⁴⁸, M. Miluzio^{35,174}, P. Monaco^{153,52,53,51}, A. Montoro^{128,126}, C. Moretti^{54,146,52,51,53}, G. Morgante⁹, K. Naidoo¹, A. Navarro-Alsina¹⁰⁵, S. Nesseris¹⁴⁷, F. Passalacqua^{121,85}, K. Paterson³⁴, L. Patrizii²⁰, A. Pisani^{86,175}, D. Potter¹⁶⁶, S. Quai^{8,9}, M. Radovich⁵⁶, S. Sacquegna^{140,141,142}, M. Sahlén¹⁷⁶, D. B. Sanders⁷⁴, E. Sarpa^{54,146,53}, A. Schneider¹⁶⁶, D. Sciotti^{69,70}, E. Sellentin^{177,68}, L. C. Smith¹⁷⁸, K. Tanidis², G. Testera⁵⁹, R. Teyssier¹⁷⁵, S. Tosi^{58,59,31}, A. Troja^{121,85}, M. Tucci¹⁸, C. Valieri²⁰, A. Venhola¹⁷⁹, D. Vergani⁹, G. Vernardos^{180,181}, G. Verza¹⁸², P. Vielzeuf⁸⁶, N. A. Walton¹⁷⁸, and D. Scott¹⁸³

(Affiliations can be found after the references)

Received 18 March 2025 / Accepted 5 January 2026

ABSTRACT

The Euclid Wide Survey (EWS) is expected to identify in the order of 100 000 galaxy-galaxy strong lenses across 14 000 deg². The Euclid Quick Data Release (Q1) of 63.1 deg² *Euclid* images provides an excellent opportunity to test our lens-finding ability, and to verify the anticipated lens frequency in the EWS. Following the Q1 data release, eight machine learning networks from five teams were applied to approximately one million images. This was followed by a citizen science inspection of a subset of around 100 000 images, of which 65% received high network scores, with the remainder randomly selected. The top scoring outputs were inspected by experts to establish confident (grade A), likely (grade B), possible (grade C), and unlikely lenses. In this paper we combine the citizen science and machine learning classifiers into an ensemble, demonstrating that a combined approach can produce a purer and more complete sample than the original individual classifiers. Using the expert-graded subset as ground truth, we find that this ensemble can provide a purity of $52 \pm 2\%$ (grade A/B lenses) with 50% completeness (for context, due to the rarity of lenses a random classifier would have a purity of 0.05% and the best machine learning network in this work achieved 7.3% purity for the same completeness). We discuss future lessons for the first major *Euclid* data release (DR1), where the big-data challenges will become more significant and will require analysing more than ~ 300 million galaxies, and thus the time investment of both experts and citizens must be carefully managed.

Key words. gravitational lensing: strong – methods: data analysis – methods: statistical

1. Introduction

Strong gravitational lensing, whereby light is deflected by gravity to form multiple images of the background source, provides a useful probe for a wide range of science cases, including cosmology (e.g. Shajib et al. 2023; Li et al. 2024), probing dark matter (e.g. Powell et al. 2023; Wagner-Carena et al. 2024), galaxy evolution (Sonnenfeld et al. 2019; Etherington et al. 2023), and probing the early Universe (e.g. van Dokkum et al. 2024). With the commencement of the Euclid Wide Survey (EWS, Euclid Collaboration: Scaramella et al. 2022), and the forthcoming start of the Legacy Survey of Space and Time (LSST, Ivezić et al. 2019), the field of strong lensing will undergo significant advances, with the number of strong lens systems known expected to increase to around 100 000 (Collett 2015; Holloway et al. 2023; Pearson et al. 2024). Furthermore, the large-scale spectroscopic confirmation of perhaps 10 000 of these systems using the 4MOST Strong Lensing Spectroscopic Legacy Survey (4SLSLS, Collett et al. 2023) will allow for population-level analysis on a scale not seen to date.

The *Euclid* satellite (Euclid Collaboration: Mellier et al. 2025), which launched on 1 July 2023, aims to survey 14 000 deg² of the sky. In comparison to

LSST, a ground-based seeing-limited (0''.8) survey, the EWS will provide higher-resolution imaging (0''.16 in I_E , Euclid Collaboration: Cropper et al. 2025; Euclid Collaboration: McCracken et al. 2026) over a comparable area, but to a shallower depth (LSST: $r \simeq 26.9^1$, *Euclid*: $I_E \simeq 26.2$, Euclid Collaboration: Scaramella et al. 2022). The *Euclid* Quick Data Release (Q1, Euclid Quick Release Q1 2025), comprising 63.1 deg² of imaging, provides an excellent testbed for current strong-lens-detection algorithms.

To date, strong lenses have been found by a range of different methods and classifiers (see Lemon et al. 2024 for a review), including arc- and ring-finding algorithms (e.g. Seidel & Bartelmann 2007; Sonnenfeld et al. 2018), machine learning (hereafter ML), such as neural network classifiers (e.g. Huang et al. 2020, 2021; Pearce-Casey et al. 2025; More et al. 2024; Melo et al. 2025; Cañameras et al. 2024; Storfer et al. 2024; Nagam et al. 2025), through visual inspection by strong lens experts (Faure et al. 2008; Jackson 2008; Pawase et al. 2014), and crowd-sourced through citizen science (e.g. Marshall et al. 2016; More et al. 2016; Sonnenfeld et al. 2020; Garvin et al. 2022; González et al. 2025). The citizen

¹ <https://www.lsst.org/scientists/keynumbers>

science search conducted by Garvin et al. (2022) used *Hubble* Space Telescope (HST) archival data on roughly half the area (27 deg^2) of the Q1 data, identifying 167 A or B-grade candidates. Due to the wide range of possible lens configurations, the rarity of strong lens systems, and the high rate of strong-lens mimics (e.g. high-redshift or faint spiral galaxies, ring galaxies, chance alignments, etc.) that vastly outnumber true lenses, state-of-the-art lens classifiers still suffer from a ‘false positive problem’ (Holloway et al. 2024), where non-lenses dominate the high-scoring sample from lens classifiers. To date, these false positives have had to be identified and removed manually.

The EWS presents a challenge to the strong lensing community. The methods used to find the lens systems must be scalable to the billions of galaxy systems this survey will detect. Furthermore, since strong lensing is a rare phenomenon, these methods must either have sufficiently low false positive rates ($\text{FPR} \lesssim 10^{-3}$) that their lens samples are not dominated by non-lenses, or any subsequent analysis must be done in a manner that accounts for possible contamination (Holloway et al., in prep.). In Holloway et al. (2024), the benefits of an ensemble classifier were investigated using data from the Hyper-Suprime Cam (HSC) Subaru Strategic Program (Aihara et al. 2019), whereby multiple lens classifiers (both ML and citizen science) were combined to provide a purer and more complete lens sample. In this work, we extend this analysis to *Euclid* data, and investigate the scalability of the *Euclid* Strong Lens Discovery Engine to the full EWS.

Ensemble classifiers have been used in previous lens searches (e.g. Schaefer et al. 2018; Andika et al. 2023; Cañameras et al. 2024; Nagam et al. 2025), typically averaging the scores of individual ML classifiers. Bayesian ensemble methods have also been used previously in other fields. For example, Bayesian neural networks (BNN’s) have been used for classification (e.g. Thummel et al. 2024), in which the neural network weights are assigned a distribution rather than a point value, allowing estimation of the classification uncertainty. Furthermore, multiple point-estimate predictions from neural networks can be used as samples of a full posterior (see, e.g. Lakshminarayanan et al. 2017; Wilson & Izmailov 2020; Hoffmann & Elster 2021) as a Bayesian approximation with little additional cost. In this work, we adopt a post-processing approach, using distinct calibration and Bayesian combination steps. This has the benefit of being able to account for the variation in performance between classifiers, data incompleteness, and different classifier types, for example, neural networks or citizen science. This paper aims to answer the following questions.

1. How well can 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 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 paper is part of a series outlining the findings of a strong lens search in the Q1 data. Euclid Collaboration: Walmsley et al. (2026, hereafter Paper A), details the source selection, search procedure, and final strong lens candidate catalogue. Euclid Collaboration: Rojas et al. (2026, Paper B) details lens

candidates identified in Q1 data after pre-selecting high-velocity galaxies from the *Sloan* Digital Sky Survey (SDSS) and from Dark Energy Spectroscopic Instrument (DESI) spectra, in addition to the generation of one of the training sets used for the ML and citizen science searches. Euclid Collaboration: Lines et al. (2026, hereafter Paper C) details the primary machine learning models used for finding lenses in this search. Euclid Collaboration: Li et al. (2026, Paper D) focuses on the double-source plane lens (DSPL) candidates, including modelling of these systems and the outlook for the full DSPL *Euclid* sample.

This paper is structured as follows. In Sect. 2 we describe the data used in the Q1 strong lens search, the individual classifiers applied to these data, and the expert-grading of lens candidates. In Sect. 3.1 we calibrate each of these classifiers and combine these into an ensemble in Sect. 3.2. We discuss results in Sect. 4, including performance of a range of ensemble classifiers (Sect. 4.1), and the strengths and weaknesses of citizen and ensemble classifiers (Sect. 4.2). We then discuss the outlook for *Euclid* DR1 (Sect. 4.3) and conclude in Sect. 5.

2. Data

For the Q1 strong lens search, we selected I_E -detected extended objects with $I_E < 22.5$ that did not have *Gaia* counterparts from the Q1 MER catalogue (Euclid Collaboration: Romelli et al. 2026, see Paper A for the complete selection). 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). In this work, we use eight ML networks produced by five different teams, applied to these sources, along with classifications for roughly 10% of these from the Space Warps citizen science search² (as described in Paper A).

For clarity, we briefly describe the individual classifiers used in this ensemble below, but direct readers to Paper A for a comprehensive overview of the source selection, search procedure, and final catalogue of identified Q1 lenses, and Paper C for detailed descriptions of the performance and architectures of the primary five networks (Models 1, 2a, 3a, 4, and 5 listed below). In this work we also use three 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, as described below and summarised in Table 1.

- Model 0: Space Warps (see Paper A). The Space Warps strong lens search involved around 1000 citizens who made a total of 800 000 classifications. Unlike the machine learning classifiers, who were shown all the images in the dataset, 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 dataset (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_E , Y_E , and J_E bands.
- Model 1: This network (adapted from Domínguez Sánchez et al. 2018; Manjón-García 2021)

² <https://www.zooniverse.org/projects/aprajita/space-warps-esa-euclid>

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_E -band-only dataset, 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_E -band-only and (b) I_E and J_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 Observation (ERO) data (Acevedo Barroso et al. 2025), through inspection of galaxies with spectroscopic data (Paper B), Galaxy Zoo Cosmic Dawn and Galaxy Zoo *Euclid* projects, as well as 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, 2016), ResNet (b, He et al. 2016; Xie et al. 2017), and VGG (c, Simonyan & Zisserman 2015) architectures respectively. They were trained using 40 000 I_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 Dark Energy Spectroscopic Instrument data (DESI, Euclid Collaboration: Rojas et al. 2026) and simulated lenses. This classifier was shown I_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 Vision Transformer (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_E band and arcsinh scaling. Two augmented views of each image were generated using transformations, including resizing, horizontal flips, colour jitter, and Gaussian blur. A contrastive loss function, scaled by a temperature parameter, was applied to maximise similarity between positive pairs (different views of the same image) and minimise similarity between negative pairs (different images). During fine-tuning, curriculum learning was applied with an initial warm-up phase of five epochs with uniform sampling, followed by dynamic sample weighting and hard-example mining. Hard examples were identified using misclassification or low-confidence predictions, with their weights adjusted progressively to prioritise challenging samples in later epochs.

Following the application of the five primary networks (Models 1, 2a, 3a, 4, and 5) and while the Space Warps search was ongoing, the highest scoring 1000 systems from each of the networks, along with roughly 2700 high-scoring systems from Space Warps (with Space Warps score $p_{\text{SW}} > 10^{-5}$), were inspected by strong lensing experts in a phase called ‘Galaxy Judges’ (hereafter GJ, see Paper A). In total, 7700 systems were inspected, identifying 250 grade A systems (confident lens) and 247 grade B systems.

3. Method

3.1. Calibration of strong lens classifiers

To combine multiple classifiers into an ensemble, their outputs need to be calibrated. Here we define calibration as the function that maps the output score of a classifier to the probability that a system with that score is a lens. This calibration accounts for any over- or under-confidence in any particular classifier, perhaps due to differences in training sets, activation functions, or data processing in the images they were shown. The calibration of these classifiers requires a ground truth, a set of objects for which their true type (i.e. lens or non-lens) is known. For this work, since strong lens systems are rare and the majority do not have spectroscopic confirmation, unless stated otherwise, we use the grade A and B lenses from GJ described in Paper A as ‘true lenses’, and assert that all other systems (ungraded or graded as a non-lens) are not lenses. Unless stated explicitly, grade C candidates were also treated as non-lenses since, in reality, these are typically not lensed systems. Therefore, in this work, the probabilities produced following calibration reflect the probability that a system is a grade 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 is reasonable to assume the majority of the lens systems were identified. The analysis in Paper C suggests that 65% of grade A and B lenses were found in the Q1 lens search, the vast majority of those missing being grade B. The analysis in this paper should therefore be seen as conservative. Since some true lenses will have been labelled here as non-lenses (since non-inspected systems are classed as non-lenses in this work), the model performance will be underestimated. This would likely be most significant at lower model scores, which were not all inspected by experts.

It should be noted that since the ground truth used in this work relies upon expert graded systems, it may contain a systematic bias towards lens systems recognisable by eye. In turn, this may favour the performance of the citizen science search (since it is likely the volunteer inspectors on Space Warps carry similar selection biases to the expert inspectors) over the machine learning classifiers trained on simulated lenses. The training sets of Models 1, 2a/b, and 5 also included lens candidates identified in part through expert visual inspection, so could also include a bias towards human-identifiable lenses. In the future, such bias may be mitigated by using large samples of spectroscopically confirmed lenses (e.g. through 4SLSLS, Collett et al. 2023).

The best calibration method identified in Holloway et al. (2024) was isotonic regression (Zadrozny et al. 2002). This calibration method assumes that the mapping from the output of the classifier to the lens probability is a monotonically increasing function, i.e. the higher the score that a classifier gives to a particular system, the more likely the system is strongly lensed. We split the data into two equal sized datasets, forming a ‘calibration set’ (on which the calibration curves were calculated), and a ‘test set’. Since we did not adjust the calibration methodology from Holloway et al. (2024) we did not use a separate validation set. We show the calibration mappings produced by this isotonic regression method in Fig. A.1. We validated the calibration on the test set, as shown in Fig. 1, where we measure the ratio of the number of lenses to the total number of systems with a given calibrated probability. If the calibration is accurate, this ratio would be equal to the calibrated probability. We find 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 $O(1)$, while

Table 1. Summary of models used in this work, as well as a summary metric of their performance, purity (%) at 50% completeness, P_{50} , as measured on the test set.

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

Notes. The P_{50} value for Model 0 was calculated using systems in the ‘random’ 40 000 dataset described in Sect. 2 for representative comparison. We 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.

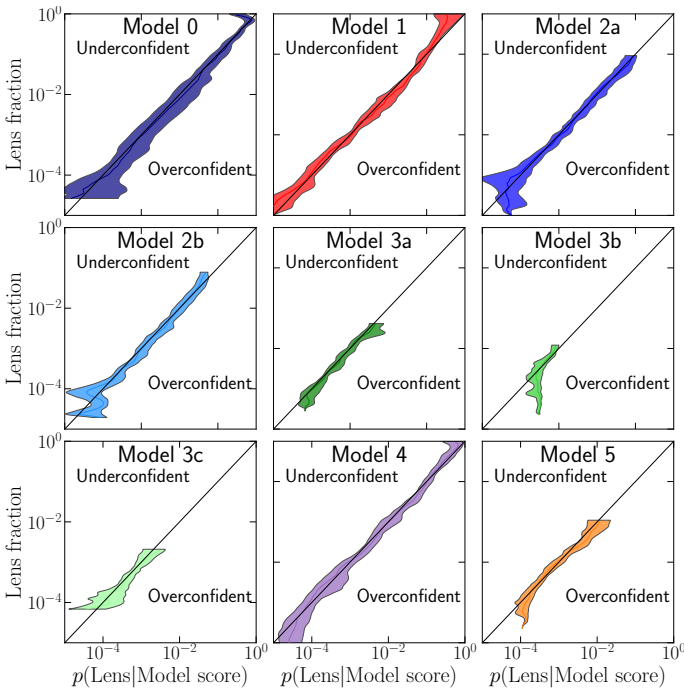


Fig. 1. Validation of calibration curves, applied to the distinct test set of data. The best performing classifiers can be calibrated up to high probabilities, since their highest-ranked candidates have a high purity. Curves that follow the $y = x$ line are indicative of accurate calibration, where the 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.

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. We find some underconfidence in Model 0 at high lens fractions, indicating some inflexibility in the calibration, which assumes that the lens fraction increases monotonically with model score. We calibrated each classifier, then combined them into an ensemble, described in Sect. 3.2.

We also investigate citizen scientist grades as an alternative source of ground truth. Given that the number of systems that lensing experts can grade is limited, the possibility of citizens providing equivalent grading is of interest. Figure 2 shows the distribution of lens candidates of grades A, B, and C as identified

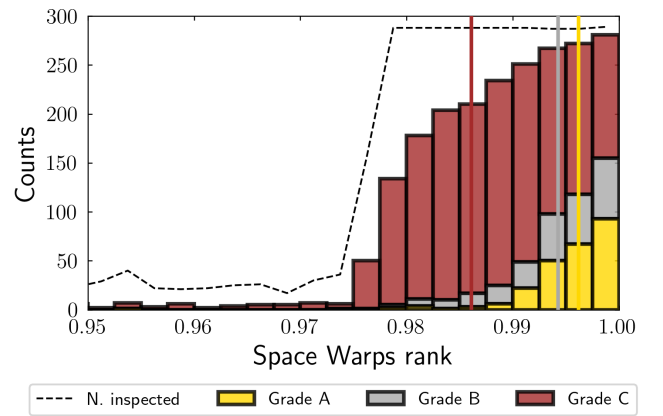


Fig. 2. GJ grades of the top 5% ranked subjects from the Space Warps search; the bins are stacked, so the envelope represents the combined purity of grade A+B+C candidates out of the total inspected (dashed line) in each bin. The vertical lines indicate the median rank of lens candidates of each grade; higher-grade lens candidates receive successively higher scores from the citizens.

by GJ, in the highest scoring 5% of systems identified by Space Warps. We find that the citizens are most confident in identifying higher-grade lens systems, which receive successively higher scores. Given this and the strong performance of the citizen scientists (see Model 0 in Fig. 3), we investigated if it was possible to use this classifier as a ground truth for the purposes of calibrating the ML classifiers. While the calibration of Model 0 was overconfident at low lens-fractions, in the primary region of interest (i.e. high probability), the calibration was accurate. We trialled two methods for this to determine whether 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 we drew samples from the calibration set assigning binary classification values according to the calibrated probabilities from Model 0 each time. We then averaged over the calibration curves produced by each set of samples. In this manner, we accounted for the fact that the calibration of Model 0 produced probabilities, rather than binary classifications.

2. Using a simple threshold in the score for Model 0, defining all systems above this threshold as ‘lens’ and vice versa. We tested both of these methods on the test set, using the original ground truth from GJ and describe the results in Sect. 4.

3.2. Combination of classifiers into an ensemble

The individual classifiers were calibrated and were then combined into an ensemble. The range of classifier types, training data, and network architectures produced outputs that typically show little correlation for the vast majority of objects in the dataset that are not lenses. A non-lens system that ranks relatively highly in one classifier could receive a very low score from a different classifier. This is beneficial, since combining each of these classifiers can build upon their individual strengths and weaknesses (eight classifiers that were always in complete agreement would provide no more information than one classifier). Following the calibration of each network in Sect. 3.1, we used a Bayesian approach to combine the individual networks and treated each classifier as effectively independent. From Holloway et al. (2024), the posterior probability that a given system is strongly lensed (denoted below by L), having received calibrated probabilities $\{C_i\}$ from classifiers $\{1, \dots, N\}$ is given by

$$P(L|\{C_i\}) = \frac{N_L^{1-N} \prod_N C_i}{N_L^{1-N} \prod_N C_i + N_{NL}^{1-N} \prod_N (1 - C_i)}, \quad (1)$$

where we have assumed a prior $p_0 = \frac{N_L}{N_L + N_{NL}}$, where N_L and N_{NL} are the number of lenses and non-lenses in the calibration set (249 and 542 965, respectively, in this work).

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 dataset. Therefore, for the majority of objects, only eight classifiers were available to form an ensemble. However, scores from all nine classifiers ($8 \times \text{ML} + \text{Space Warps}$) were available for the systems that were most likely be lenses.

4. Results and discussion

4.1. Ensemble-classifier performance

We find that the ensemble classifier made from all nine classifiers provides significant improvement in purity and completeness over the individual ML classifiers.

Figure 3 shows the receiver operating characteristic (ROC) and purity-completeness curves of the individual classifiers and those of the ensemble. To produce these, we 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 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 $\leq 10^{-3}$ are required for the resulting sample not to be dominated by non-lenses, which is difficult to achieve. 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 dataset. In Fig. 3, we plot the Model 0 results for

only the random subset to allow comparison between classifiers. Due to the smaller size of the random dataset (and correspondingly fewer lenses), there is much larger statistical noise in this model’s curve. The Space Warps classifier is limited in the number of systems it can classify (to about 100 000 in this search). Figure 3 shows that Model 0 performs very well on the random dataset, achieving $\geq 70\%$ completeness at 50% purity (in the RH panel). However, since lenses are very rare and the total citizen-inspection budget is limited, in the future systems will have to be pre-screened by ML models to maximise the number of lenses identified (see Sect. 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 dataset rather than a subset, and with performance close to that if the citizens had inspected the complete dataset.

Figure 4 shows the ROC and purity-completeness curves for only the ML classifiers, along with the results from a range of ML-only ensembles. These are generated using the same test set as in Fig. 3, as outlined in Sect. 3.1. Here we focus on the ensemble performance and refer readers to Paper C for greater discussion of the performance of the individual ML models. We find that the best ML-only ensembles are generated using a subset of the ML classifiers (in particular, Models 1, 2a, and 4). We discuss this finding further in Appendix B, and the effect of retraining Model 4 (the best performing ML classifier in Appendix C). Given the very rapid timescale in which the networks had to be trained and optimised for Q1 data (about 1 week), the relative performance of the networks used in this work would likely change, and the individual networks would further improve significantly, in future data releases. As described in Sect. 3.1, we generated three ML-only ensembles by calibrating the networks using different ground truths; the original GJ grades, the Space Warps probabilities (SW_{Cal} , themselves calibrated using GJ), and the raw Space Warps output (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). The performance of these ML-only ensembles, shown in Fig. 4, is evaluated using the original GJ ground truth in all cases, for fair comparison. We see an improvement in classification using all of 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 we discuss further in Sect. 4.3. However, overall we find the benefits of combining the networks to be smaller than those of combining the ML and Space Warps classifiers in Fig. 3.

In Appendix D we demonstrate that calibration is a necessary step prior to combining different models into an ensemble. We find that simply averaging the uncalibrated model outputs produces a much lower performing classifier than first applying calibration and then combining them via the Bayesian framework used in this work.

We also investigated the scatter between the calibrated probabilities produced by the models. We 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. 1). This resulted in high scatter by default for likely lens candidates, since only the best models could assign

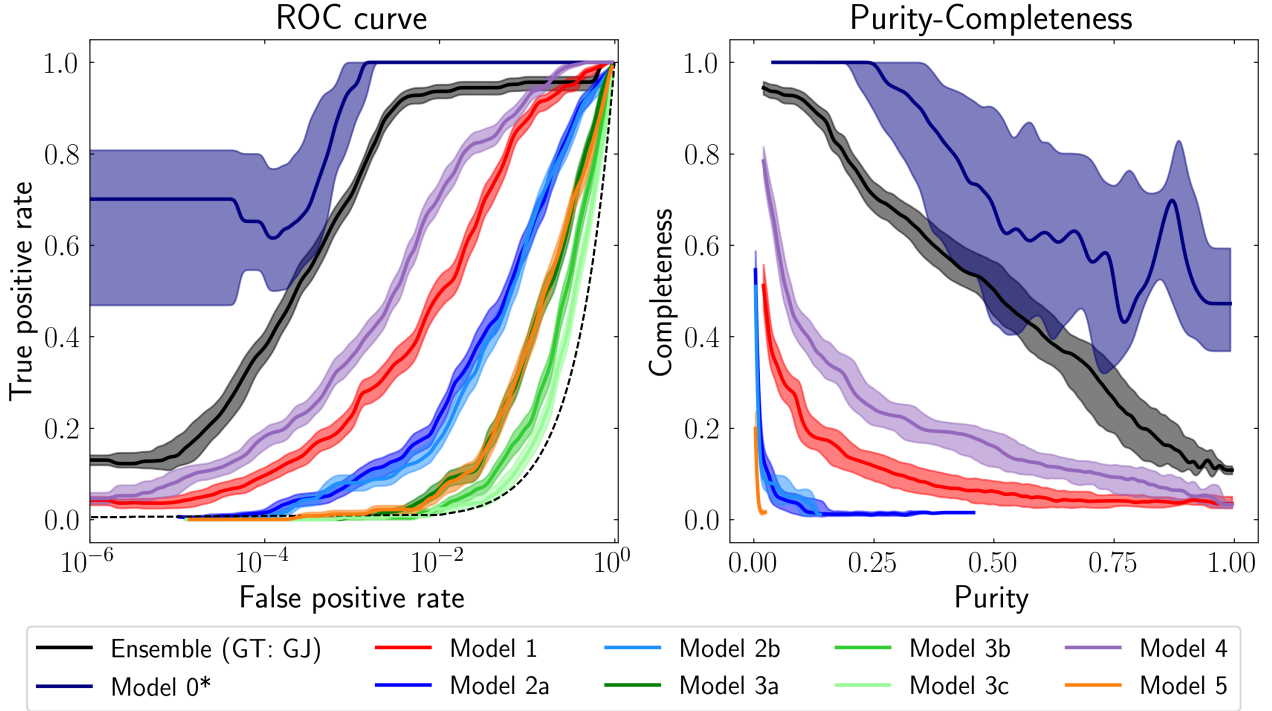


Fig. 3. ROC (*left*) and purity-completeness curves (*right*) of the individual models and an ensemble of all nine models in this work, applied to the test dataset. This ensemble was calibrated using the calibration set, assuming a ground truth of GJ grade A and B lenses. 1σ uncertainties (shaded regions) 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) of the full 1 million sample, and hence has larger statistical noise. The completeness measurement for this classifier in particular is taken as if the citizens saw the whole dataset, which would be unfeasible for future data releases (see Sect. 4.3). The curve for the ensemble is generated using the full test set, however.

probabilities $O(1)$. We found that the highest correlation in model output was between classifiers that shared a common training set (Models 3a, b, c, and Models 2a and 2b). We also found a 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 grade A and B lenses were considered, suggesting they still identified different types of lens.

4.2. Systems identified by citizens or ensemble

The ML and citizen science classifiers are naturally very different, and have their own strengths and weaknesses. These may arise from the particular datasets 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 that are out-of-distribution, such as rare artefacts). Figure 5 shows the ensemble posterior probability from an ML-only ensemble (from Models 1, 2a, and 4), versus that of an ML + Space Warps ensemble (all nine classifiers), along with a selection of cutouts of those 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 fea-

tured in training sets. There were some systems that were ranked highly by citizens, but that 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 include candidates that were mis-centred and those in crowded fields, and are typically B-grade.

Based on the PyAutoLens modelling (Nightingale et al. 2021) discussed in Paper A, we investigated correlations in classifier score and modelling properties. Here we restricted our calculations to lens candidates with successful lens models, judged to be lenses (see Paper A). This modelling provided estimates of the lensed and unlensed magnitudes, Einstein radius, and signal-to-noise ratio. We find the ML + Space Warps ensemble classifier score correlates most significantly with magnified source magnitude ($\rho = 0.5$), along with total signal-to-noise ratio ($\rho = 0.47$). The Space Warps classifier is most correlated with the Einstein radius ($\rho = 0.3$), implying that 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.

4.3. Outlook for DR1 and future data releases

In this work, we have used expert grades as a ground truth. With 4SLSLS (Collett et al. 2023) it will become possible to use spectroscopically confirmed (and spectroscopically refuted) systems instead, to aid with the calibration of lens classifiers. Prior to this, the ranked outputs of an ensemble classifier in DR1 could

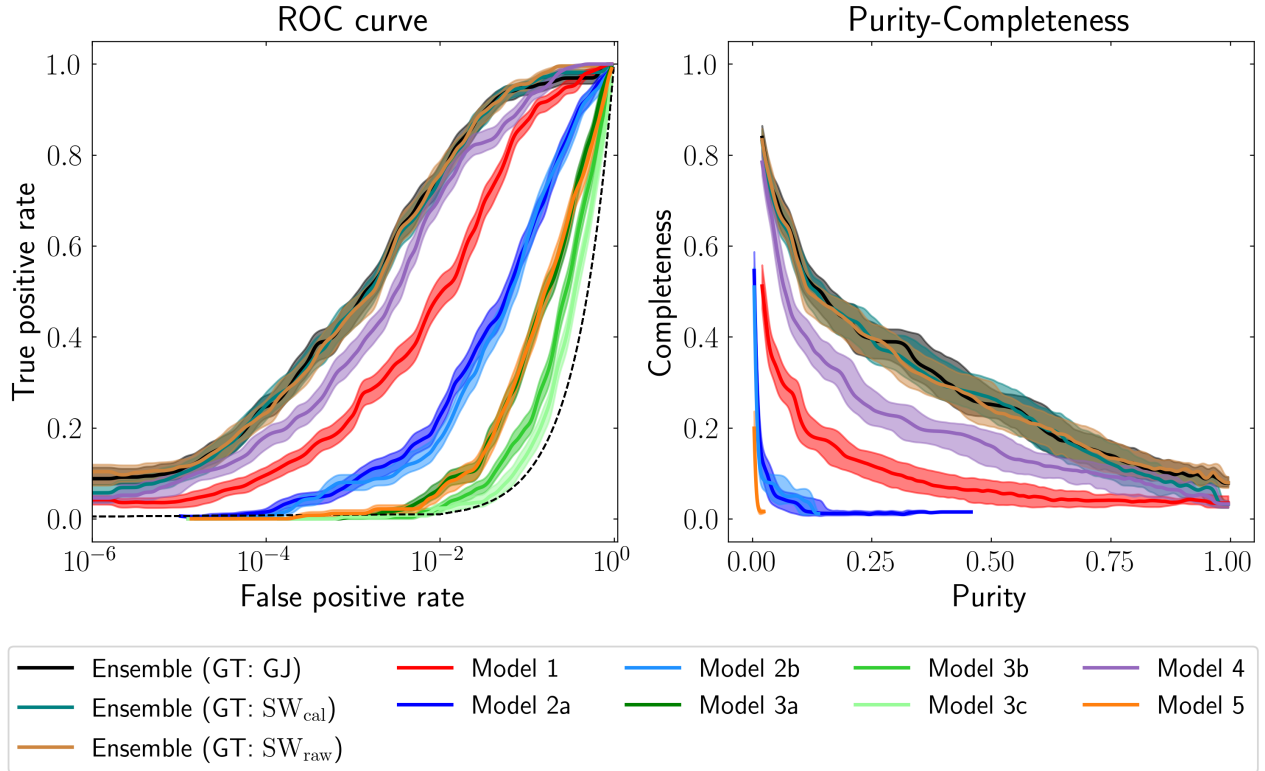


Fig. 4. ROC (*left*) and purity-completeness curves (*right*) of the ML classifiers, and three iterations of an ensemble generated using Model 1, 2a, and 4. To generate these three ensembles, the constituent ML models were calibrated using three different ground truths, namely GJ (GT: GJ), Space Warps calibrated probabilities (GT: SW_{cal}), and Space Warps binary outputs (GT: SW_{raw}, using a raw SW score threshold of $p = 1-1.5 \times 10^{-8}$), as described in Sect. 3.1. The models were then combined into ensembles via Eq. (1). In this figure, the performances of these ensembles has been evaluated using the original GJ ground truth for fair comparison. We find the effect of calibrating the models using different ground truths to be small.

also be used to prioritise 4SLSLS follow-up. Figure 6 shows the number of true lenses that would be identified as a function of the number that would need to be inspected in the full EWS, based on the performance of the Space Warps + ML and ML-only ensemble classifiers. For the former ensemble, we find that a small majority of the A and B-grade lenses (52%) would be identified in a 50% pure sample. Given our 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. However, given the rapid improvement of lens classifiers over time, the much larger catalogue of *Euclid* strong lens systems now available (Paper A) for training and validation, and the fact that the networks applied to Q1 had to be optimised over a very short timescale, it is likely that classification performance will improve further prior to DR1. The large number of classifications of *Euclid* cutouts now available from both citizens and expert graders will provide a rich dataset from which to re-train the ML classifiers, particularly on difficult false-positive systems. This will likely occur iteratively following each future *Euclid* data release, allowing the models to continuously improve over time. Similarly, 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, 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. 2023; Erickson et al. 2025; Busillo et al. in prep.; Venkatraman et al. in prep.) will provide an additional measure for the plausibility of candidate lens systems, and therefore we anticipate that a higher completeness than estimated here is achievable.

4.3.1. 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 identify the most lenses. For context, in the Q1 Space Warps lens search, around 1000 volunteers made 800 000 classifications of 100 000 cutouts over a period of 2 months. 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, also over a 2 month period. Furthermore, more relaxed time constraints for DR1 inspection and analysis compared to Q1 will help increase both the citizen and expert inspection budgets. However, there remains a limit to the number that can be inspected. The 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

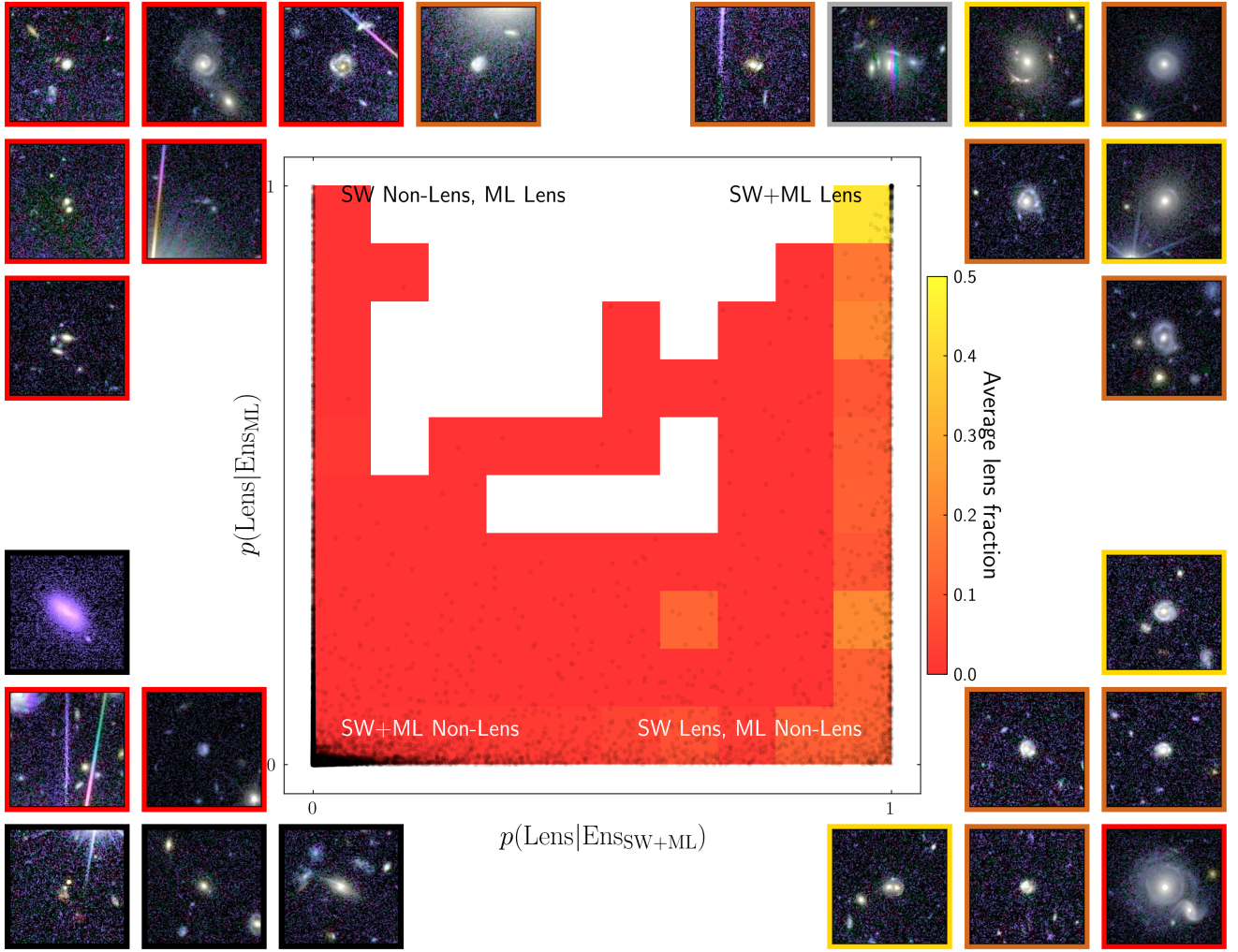


Fig. 5. Plot of the lens posterior from an ML-only ensemble compared to from an ML+Space Warps ensemble. Systems that receive both a high ML-only posterior and a high ML+Space Warps posterior are the most likely lens candidates. A selection of systems identified by both the ML-only and ML+Space Warps 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). In the central plot, we depict the average (grade A+B) lens fraction, showing that the highest purity can be achieved when both ensembles are in agreement.

randomly selected galaxies can help identify unusual lens configurations, showing 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 therefore supports a two-stage approach, whereby ML classifiers are applied to the full DR1 dataset, 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 strong lensing experts who took part in GJ inspected and graded 7000 images over roughly 3 months. The DR1 dataset will be 36 times larger than the Q1 data release, but the inspection capacity of expert graders will likely remain similar. Given the longer timescale, it is possible that experts will inspect a larger sample, so we consider two scenarios here. We first scale our results by a factor of $36\times$ to that of the DR1 area (in particular on the number of lenses and 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, we find that approximately 7600 would be identified if the scores from all classifiers were first combined into a citizen+ML ensemble. Combining the scores into an ensemble at each stage would therefore be worthwhile to produce a higher purity sample.

Without significantly increasing the inspection capacity of experts or citizens, the resulting completeness would be

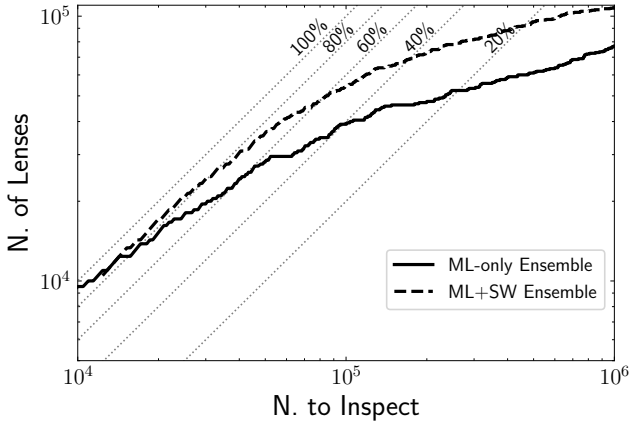


Fig. 6. Estimate of the number of true strong lenses that would be identified in the full EWS as a function of the number required to inspect, for both the ML-only and ML+Space Warps ensembles. The latter shows an optimistic scenario, given citizen classifications likely would not be available for a dataset significantly larger than the Q1 dataset used in this work. The total number of lens systems expected in the EWS has been scaled from the number of grade A and B lenses found in Q1. The diagonal lines (dotted) show lines of constant purity (as labelled). This figure demonstrates that achieving high completeness comes at the cost of lower purity, and requires large samples for inspection.

relatively low (<50%). A priority for DR1 will be obtaining a sufficient dataset (around 10 000 systems) of high-grade strong lens candidates for follow-up through 4SLSLS, which based on the current performance of the lens classifiers and anticipated improvements will likely be met. Beyond this, the importance of purity versus completeness will vary depending on the particular science case. While typical lens studies have focused on spectroscopically confirmed systems, if the contamination rate is known, unbiased inference can be performed on impure datasets (e.g. Kunz et al. 2007; Roberts et al. 2017; Holloway et al. in prep.). The completeness could be improved through the automated modelling of candidates, or by adding a refinement stage to citizen inspection as done in Marshall et al. (2016). The latter of these 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 methodology of the Space Warps Analysis Pipeline (Marshall et al. 2016) to the final 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 expert grading.

5. Summary and conclusions

In this work, we have produced an ensemble strong lens classifier using the Q1 data release. In answer to the questions set out in Sect. 1, we summarise our conclusions below.

1. An ensemble classifier 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 dataset, providing posterior probabilities that each system is a lens, even when some classification data are incomplete (for example, where citizens are only shown a subset of the data).
2. An ensemble composed of neural networks and citizen scientists produces a 52% complete sample at 50% purity, or 91% complete at 10% purity. An ensemble comprised of

only ML classifiers produced a 25% complete sample, with 50% purity, or 56% complete 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). In the case of the ML-only ensemble, performance was primarily driven by Model 4, the best performing ML classifier (see Fig. B.1).

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 ML 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 done iteratively as more lenses and non-lenses are classified.

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

Acknowledgements. We gratefully acknowledge the Zooniverse team for their invaluable work on the host platform for Space Warps. This project would not be possible without the immense contributions of citizen scientists. Those contributing classifications for this project are listed on [The Team](#). We greatly appreciate their commitment, enthusiasm, and time. This work has made use of the *Euclid* Quick Release Q1 data from the *Euclid* mission of the European Space Agency (ESA), 2025, <https://doi.org/10.57780/esa-2853f3b>. This research makes use of ESA Datalabs (datalabs.esa.int), an initiative by ESA's Data Science and Archives Division in the Science and Operations Department, Directorate of Science. The Euclid Consortium acknowledges the European Space Agency and a number of agencies and institutes that have supported the development of *Euclid*, in particular the Agenzia Spaziale Italiana, the Austrian Forschungsförderungsgesellschaft funded through BMIMI, the Belgian Science Policy, the Canadian Euclid Consortium, the Deutsches Zentrum für Luft- und Raumfahrt, the DTU Space and the Niels Bohr Institute in Denmark, the French Centre National d'Etudes Spatiales, the Fundação para a Ciência e a Tecnologia, the Hungarian Academy of Sciences, the Ministerio de Ciencia, Innovación y Universidades, the National Aeronautics and Space Administration, the National Astronomical Observatory of Japan, the Nederlandse Onderzoekschool Voor Astronomie, the Norwegian Space Agency, the Research Council of Finland, the Romanian Space Agency, the Swiss Space Office (SSO) at the State Secretariat for Education, Research, and Innovation (SERI), and the United Kingdom Space Agency. A complete and detailed list is available on the *Euclid* web site (www.euclid-ec.org/consortium/community/). PH acknowledges funding from the Science and Technology Facilities Council, Grant Code ST/W507726/1. This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (LensEra: grant agreement No 945536). TEC is funded by the Royal Society through a University Research Fellowship. C.T. acknowledges the INAF grant 2022 LEMON. LL thanks the support from INAF theory Grant 2022: Illuminating Dark Matter using Weak Lensing by Cluster Satellites. SS has received funding from the European Union's Horizon 2022 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101105167 – FASTIDIoUS. The Dunlap Institute is funded through an endowment established by the David Dunlap family and the University of Toronto. GD acknowledges the funding by the European Union – NextGenerationEU, in the framework of the HPC project – “National Centre for HPC, Big

Data and Quantum Computing” (PNRR – M4C2 – I1.4 – CN00000013 – CUP J33C22001170001).

References

- Acevedo Barroso, J. A., O’Riordan, C. M., Clément, B., et al. 2025, *A&A*, **697**, A14
- Aihara, H., AlSayyad, Y., Ando, M., et al. 2019, *PASJ*, **71**, 114
- Andika, I. T., Suyu, S. H., Cañameras, R., et al. 2023, *A&A*, **678**, A103
- Cañameras, R., Schuldt, S., Shu, Y., et al. 2024, *A&A*, **692**, A72
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020, *A Simple Framework for Contrastive Learning of Visual Representations*
- Collett, T. E. 2015, *ApJ*, **811**, 20
- Collett, T. E., Sonnenfeld, A., Frohmaier, C., et al. 2023, *The Messenger*, **190**, 49
- Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. L. 2018, *MNRAS*, **476**, 3661
- Ericksen, S., Wagner-Carena, S., Marshall, P., et al. 2025, *AJ*, **170**, 44
- ESA Datalabs 2024, https://doi.org/10.1007/978-981-97-0041-7_1
- Etherington, A., Nightingale, J. W., Massey, R., et al. 2023, *MNRAS*, **521**, 6005
- Euclid Collaboration (Scaramella, R., et al.) 2022, *A&A*, **662**, A112
- Euclid Collaboration (Leuzzi, L., et al.) 2024, *A&A*, **681**, A68
- Euclid Collaboration (Cropper, M., et al.) 2025, *A&A*, **697**, A2
- Euclid Collaboration (Mellier, Y., et al.) 2025, *A&A*, **697**, A1
- Euclid Collaboration (Li, T., et al.) 2026, *A&A*, **711**, A29 (Euclid Q1 SI)
- Euclid Collaboration (Lines, N. E. P., et al.) 2026, *A&A*, **711**, A28 (Euclid Q1 SI)
- Euclid Collaboration (McCracken, H. J., et al.) 2026, *A&A*, **711**, A2 (Euclid Q1 SI)
- Euclid Collaboration (Rojas, K., et al.) 2026, *A&A*, **711**, A27 (Euclid Q1 SI)
- Euclid Collaboration (Romelli, E., et al.) 2026, *A&A*, **711**, A4 (Euclid Q1 SI)
- Euclid Collaboration (Walmsley, M., et al.) 2026, *A&A*, **711**, A26 (Euclid Q1 SI)
- Euclid Quick Release Q1 2025, <https://doi.org/10.57780/esa-2853f3b>
- Faure, C., Kneib, J.-P., Covone, G., et al. 2008, *ApJS*, **176**, 19
- Garvin, E. O., Kruk, S., Cornen, C., et al. 2022, *A&A*, **667**, A141
- Gentile, F., Tortora, C., Covone, G., et al. 2023, *MNRAS*, **522**, 5442
- González, J., Holloway, P., Collett, T., et al. 2025, *ApJ*, submitted [arXiv:2501.15679]
- He, K., Zhang, X., Ren, S., & Sun, J. 2016, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770
- Hoffmann, L., & Elster, C. 2021, arXiv e-prints [arXiv:2105.13283]
- Holloway, P., Verma, A., Marshall, P. J., More, A., & Tecza, M. 2023, *MNRAS*, **525**, 2341
- Holloway, P., Marshall, P. J., Verma, A., et al. 2024, *MNRAS*, **530**, 1297
- Huang, X., Storfer, C., Gu, A., et al. 2021, *ApJ*, **909**, 27
- Huang, X., Storfer, C., Ravi, V., et al. 2020, *ApJ*, **894**, 78
- Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, *ApJ*, **873**, 111
- Jackson, N. 2008, *MNRAS*, **389**, 1311
- Kunz, M., Bassett, B. A., & Hlozek, R. A. 2007, *Phys. Rev. D*, **75**, 103508
- Lakshminarayanan, B., Pritzel, A., & Blundell, C. 2017, *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17* (Red Hook, NY, USA: Curran Associates Inc.), 6405
- Lemon, C., Courbin, F., More, A., et al. 2024, *Space Sci. Rev.*, **220**, 23
- Li, T., Collett, T. E., Krawczyk, C. M., & Enzi, W. 2024, *MNRAS*, **527**, 5311
- Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, *MNRAS*, **389**, 1179
- Manjón-García, A. 2021, *Automated Techniques for the Detection and Analysis of the Gravitational Lensing Effect in New Generation Galaxy Surveys*, <https://hdl.handle.net/10902/24245>
- Marshall, P. J., Verma, A., More, A., et al. 2016, *MNRAS*, **455**, 1171
- Melo, A., Cañameras, R., Schuldt, S., et al. 2025, *A&A*, **698**, A264
- More, A., Verma, A., Marshall, P. J., et al. 2016, *MNRAS*, **455**, 1191
- More, A., Cañameras, R., Jaelani, A. T., et al. 2024, *MNRAS*, **533**, 525
- Nagam, B. C., Acevedo Barroso, J. A., Wilde, J., et al. 2025, *A&A*, **702**, A130
- Nightingale, J. W., Hayes, R. G., Kelly, A., et al. 2021, *J. Open Source Softw.*, **6**, 2825
- Oquab, M., Darcet, T., Moutakanni, T., et al. 2023, *Transactions on Machine Learning Research* [arXiv:2304.07193]
- Pawase, R. S., Courbin, F., Faure, C., Kokotanekova, R., & Meylan, G. 2014, *MNRAS*, **439**, 3392
- Pearce-Casey, R., Nagam, B. C., Wilde, J., et al. 2025, *A&A*, **696**, A214
- Pearson, J., Serjeant, S., Wang, W.-H., et al. 2024, *MNRAS*, **527**, 12044
- Poh, J., Samudre, A., Čipriani, A., et al. 2022, arXiv e-prints [arXiv:2211.05836]
- Powell, D. M., Vegetti, S., McKean, J. P., et al. 2023, *MNRAS*, **524**, L84
- Roberts, E., Lochner, M., Fonseca, J., et al. 2017, *JCAP*, **10**, 036
- Schaefer, C., Geiger, M., Kuntzer, T., & Kneib, J. P. 2018, *A&A*, **611**, A2
- Schuldt, S., Cañameras, R., Shu, Y., et al. 2023, *A&A*, **671**, A147
- Seidel, G., & Bartelmann, M. 2007, *A&A*, **472**, 341
- Shajib, A. J., Mozumdar, P., Chen, G. C. F., et al. 2023, *A&A*, **673**, A9
- Simonyan, K., & Zisserman, A. 2015, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015*
- Smith, M. J., Roberts, R. J., Angeloudi, E., & Huertas-Company, M. 2024, arXiv e-prints [arXiv:2405.14930]
- Sonnenfeld, A., Chan, J. H. H., Shu, Y., et al. 2018, *PASJ*, **70**, S29
- Sonnenfeld, A., Wang, W., & Bahcall, N. 2019, *A&A*, **622**, A30
- Sonnenfeld, A., Verma, A., More, A., et al. 2020, *A&A*, **642**, A148
- Storfer, C., Huang, X., Gu, A., et al. 2024, *ApJS*, **274**, 16
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. 2016, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818
- Szegedy, C., Wei, L., Yangqing, J., et al. 2015, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1
- Thummel, L., Bose, B., Pourtsidou, A., & Lombriser, L. 2024, *MNRAS*, **535**, 3141
- van Dokkum, P., Brammer, G., Wang, B., Leja, J., & Conroy, C. 2024, *Nat. Astron.*, **8**, 119
- Wagner-Carena, S., Lee, J., Pennington, J., et al. 2024, *ApJ*, **975**, 297
- Walmsley, M., Smith, L., Lintott, C., et al. 2020, *MNRAS*, **491**, 1554
- Walmsley, M., Lintott, C., Geron, T., et al. 2022, *MNRAS*, **509**, 3966
- Walmsley, M., Allen, C., Aussel, B., et al. 2023, *J. Open Source Softw.*, **8**, 5312
- Wilde, J., Serjeant, S., Bromley, J. M., et al. 2022, *MNRAS*, **512**, 3464
- Wilson, A. G., & Izmailov, P. 2020, in *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20* (Red Hook, NY, USA: Curran Associates Inc.)
- Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. 2017, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5987
- Zadrozny, B., & Elkan, C. 2002, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694

- ¹ Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 3FX, UK
- ² Department of Physics, Oxford University, Keble Road, Oxford OX1 3RH, UK
- ³ David A. Dunlap Department of Astronomy & Astrophysics, University of Toronto, 50 St George Street, Toronto Ontario M5S 3H4, Canada
- ⁴ Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK
- ⁵ Kavli Institute for Particle Astrophysics & Cosmology (KIPAC), Stanford University, Stanford, CA 94305, USA
- ⁶ SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA
- ⁷ The Inter-University Centre for Astronomy and Astrophysics, Post Bag 4, Ganeshkhind, Pune 411007, India
- ⁸ Dipartimento di Fisica e Astronomia “Augusto Righi” – Alma Mater Studiorum Università di Bologna, Via Piero Gobetti 93/2, 40129 Bologna, Italy
- ⁹ INAF-Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, 40129 Bologna, Italy
- ¹⁰ Departamento Física Aplicada, Universidad Politécnica de Cartagena, Campus Muralla del Mar, 30202 Cartagena, Murcia, Spain
- ¹¹ University of Applied Sciences and Arts of Northwestern Switzerland, School of Engineering, 5210 Windisch, Switzerland
- ¹² Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, 08028 Barcelona, Spain
- ¹³ School of Physical Sciences, The Open University, Milton Keynes MK7 6AA, UK
- ¹⁴ Technical University of Munich, TUM School of Natural Sciences, Physics Department, James-Frank-Str. 1, 85748 Garching, Germany
- ¹⁵ Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85748 Garching, Germany
- ¹⁶ Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland

- 17 SCITAS, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
- 18 Department of Astronomy, University of Geneva, ch. d’Ecogia 16, 1290 Versoix, Switzerland
- 19 Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig de Lluís Companys 23, 08010 Barcelona, Spain
- 20 INFN-Sezione di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
- 21 Aix-Marseille Université, CNRS, CNES, LAM, Marseille, France
- 22 Institut d’Astrophysique de Paris, UMR 7095, CNRS, and Sorbonne Université, 98 bis boulevard Arago, 75014 Paris, France
- 23 Dipartimento di Fisica “Aldo Pontremoli”, Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy
- 24 INAF-IASF Milano, Via Alfonso Corti 12, 20133 Milano, Italy
- 25 Minnesota Institute for Astrophysics, University of Minnesota, 116 Church St SE, Minneapolis, MN 55455, USA
- 26 Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700, AV Groningen, The Netherlands
- 27 STAR Institute, University of Liège, Quartier Agora, Allée du six Août 19c, 4000 Liège, Belgium
- 28 INAF-Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131 Napoli, Italy
- 29 Instituto de Física de Cantabria, Edificio Juan Jordá, Avenida de los Castros, 39005 Santander, Spain
- 30 Caltech/IPAC, 1200 E. California Blvd., Pasadena, CA 91125, USA
- 31 INAF-Osservatorio Astronomico di Brera, Via Brera 28, 20122 Milano, Italy
- 32 INFN-Sezione di Milano, Via Celoria 16, 20133 Milano, Italy
- 33 Laboratoire univers et particules de Montpellier, Université de Montpellier, CNRS, 34090 Montpellier, France
- 34 Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany
- 35 ESAC/ESA, Camino Bajo del Castillo s/n, Urb. Villafranca del Castillo, 28692 Villanueva de la Cañada, Madrid, Spain
- 36 Department of Physics, Centre for Extragalactic Astronomy, Durham University, South Road, Durham DH1 3LE, UK
- 37 Department of Physics, Institute for Computational Cosmology, Durham University, South Road, Durham DH1 3LE, UK
- 38 Institute for Particle Physics and Astrophysics, Dept. of Physics, ETH Zurich, Wolfgang-Pauli-Strasse 27, 8093 Zurich, Switzerland
- 39 Lawrence Berkeley National Laboratory, One Cyclotron Road., Berkeley CA 94720, USA
- 40 Universitäts-Sternwarte München, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstr. 1, 81679 München, Germany
- 41 Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, 85748 Garching, Germany
- 42 Department of Astronomy, School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China
- 43 National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan
- 44 University of Trento, Via Sommarive 14, I-38123 Trento, Italy
- 45 Dipartimento di Fisica e Astronomia, Università di Firenze, Via G. Sansone 1, 50019 Sesto Fiorentino, Firenze, Italy
- 46 INAF-Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, 50125 Firenze, Italy
- 47 Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Am Mühlenberg 1, D-14476 Potsdam-Golm, Germany
- 48 Université Paris-Saclay, CNRS, Institut d’astrophysique spatiale, 91405 Orsay, France
- 49 School of Mathematics and Physics, University of Surrey, Guildford, Surrey GU2 7XH, UK
- 50 Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, 91191 Gif-sur-Yvette, France
- 51 IFPU, Institute for Fundamental Physics of the Universe, Via Beirut 2, 34151 Trieste, Italy
- 52 INAF-Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, 34143 Trieste, Italy
- 53 INFN, Sezione di Trieste, Via Valerio 2, 34127 Trieste, TS, Italy
- 54 SISSA, International School for Advanced Studies, Via Bonomea 265, 34136 Trieste, TS, Italy
- 55 Dipartimento di Fisica e Astronomia, Università di Bologna, Via Gobetti 93/2, 40129 Bologna, Italy
- 56 INAF-Osservatorio Astronomico di Padova, Via dell’Osservatorio 5, 35122 Padova, Italy
- 57 Space Science Data Center, Italian Space Agency, Via del Politecnico snc, 00133 Roma, Italy
- 58 Dipartimento di Fisica, Università di Genova, Via Dodecaneso 33, 16146 Genova, Italy
- 59 INFN-Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy
- 60 Department of Physics “E. Pancini”, University Federico II, Via Cinthia 6, 80126 Napoli, Italy
- 61 Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, PT4150-762 Porto, Portugal
- 62 Faculdade de Ciências da Universidade do Porto, Rua do Campo de Alegre, 4150-007 Porto, Portugal
- 63 Dipartimento di Fisica, Università degli Studi di Torino, Via P. Giuria 1, 10125 Torino, Italy
- 64 INFN-Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy
- 65 INAF-Osservatorio Astrofisico di Torino, Via Osservatorio 20, 10025 Pino Torinese, (TO), Italy
- 66 European Space Agency/ESTEC, Keplerlaan 1, 2201 AZ Noordwijk, The Netherlands
- 67 Institute Lorentz, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands
- 68 Leiden Observatory, Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands
- 69 INAF-Osservatorio Astronomico di Roma, Via Frascati 33, 00078 Monteporzio Catone, Italy
- 70 INFN-Sezione di Roma, Piazzale Aldo Moro, 2 – c/o Dipartimento di Fisica, Edificio G. Marconi, 00185 Roma, Italy
- 71 Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avenida Complutense 40, 28040, Madrid, Spain
- 72 Port d’Informació Científica, Campus UAB, C. Albareda s/n, 08193 Bellaterra, Barcelona, Spain
- 73 INFN section of Naples, Via Cinthia 6, 80126 Napoli, Italy
- 74 Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, HI 96822, USA
- 75 Dipartimento di Fisica e Astronomia “Augusto Righi” – Alma Mater Studiorum Università di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
- 76 Instituto de Astrofísica de Canarias, E-38205 La Laguna, Tenerife, Spain
- 77 Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
- 78 European Space Agency/ESRIN, Largo Galileo Galilei 1, 00044 Frascati, Roma, Italy
- 79 Université Claude Bernard Lyon 1, CNRS/IN2P3, IP2I Lyon, UMR 5822, Villeurbanne F-69100, France
- 80 UCB Lyon 1, CNRS/IN2P3, IUF, IP2I Lyon, 4 rue Enrico Fermi, 69622 Villeurbanne, France
- 81 Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK
- 82 Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, PT1749-016 Lisboa, Portugal
- 83 Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Campo Grande 1749-016, Lisboa, Portugal
- 84 INAF-Istituto di Astrofisica e Planetologia Spaziali, Via del Fosso del Cavaliere 100, 00100, Roma, Italy
- 85 INFN-Padova, Via Marzolo 8, 35131 Padova, Italy
- 86 Aix-Marseille Université, CNRS/IN2P3, CPPM, Marseille, France

- ⁸⁷ INFN-Bologna, Via Irnerio 46, 40126 Bologna, Italy
- ⁸⁸ School of Physics, HH Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, UK
- ⁸⁹ FRACTAL S.L.N.E., calle Tulipán 2, Portal 13 1A, 28231 Las Rozas de Madrid, Spain
- ⁹⁰ Institute of Theoretical Astrophysics, University of Oslo, PO Box 1029, Blindern 0315, Oslo, Norway
- ⁹¹ Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA
- ⁹² Felix Hormuth Engineering, Goethestr. 17, 69181 Leimen, Germany
- ⁹³ Technical University of Denmark, Elektrovej 327, 2800 Kgs. Lyngby, Denmark
- ⁹⁴ Cosmic Dawn Center (DAWN), Kongens Lyngby, Denmark
- ⁹⁵ NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
- ⁹⁶ Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK
- ⁹⁷ Department of Physics and Helsinki Institute of Physics, Gustaf Hållströmin katu 2, University of Helsinki, 00014 Helsinki, Finland
- ⁹⁸ Université de Genève, Département de Physique Théorique and Centre for Astroparticle Physics, 24 quai Ernest-Ansermet, CH-1211 Genève 4, Switzerland
- ⁹⁹ Department of Physics, PO Box 64, University of Helsinki 00014 Helsinki, Finland
- ¹⁰⁰ Helsinki Institute of Physics, University of Helsinki, Gustaf Hållströmin katu 2, 00014 Helsinki, Finland
- ¹⁰¹ Centre de Calcul de l'IN2P3/CNRS, 21 Avenue Pierre de Coubertin, 69627 Villeurbanne Cedex, France
- ¹⁰² Laboratoire d'étude de l'Univers et des phénomènes eXtremes, Observatoire de Paris, Université PSL, Sorbonne Université, CNRS, 92190 Meudon, France
- ¹⁰³ SKAO, Jodrell Bank, Lower Withington, Macclesfield SK11 9FT, UK
- ¹⁰⁴ University of Applied Sciences and Arts of Northwestern Switzerland, School of Computer Science, 5210 Windisch, Switzerland
- ¹⁰⁵ Universität Bonn, Argelander-Institut für Astronomie, Auf dem Hügel 71, 53121 Bonn, Germany
- ¹⁰⁶ Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice Cedex 4, France
- ¹⁰⁷ Université Paris Cité, CNRS, Astroparticule et Cosmologie, 75013 Paris, France
- ¹⁰⁸ CNRS-UCB International Research Laboratory, Centre Pierre Binétruy, IRL2007, CPB-IN2P3, Berkeley, USA
- ¹⁰⁹ Institut d'Astrophysique de Paris, 98bis Boulevard Arago, 75014 Paris, France
- ¹¹⁰ Aurora Technology for European Space Agency (ESA), Camino bajo del Castillo s/n, Urbanización Villafranca del Castillo, Villanueva de la Cañada 28692, Madrid, Spain
- ¹¹¹ Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra, Barcelona, Spain
- ¹¹² School of Mathematics, Statistics and Physics, Newcastle University, Herschel Building, Newcastle-upon-Tyne NE1 7RU, UK
- ¹¹³ DARK, Niels Bohr Institute, University of Copenhagen, Jagtvej 155, 2200 Copenhagen, Denmark
- ¹¹⁴ Waterloo Centre for Astrophysics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
- ¹¹⁵ Department of Physics and Astronomy, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
- ¹¹⁶ Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada
- ¹¹⁷ Centre National d'Etudes Spatiales – Centre spatial de Toulouse, 18 Avenue Edouard Belin, 31401 Toulouse Cedex 9, France
- ¹¹⁸ Institute of Space Science, Str. Atomistilor, nr. 409 Măgurele, Ilfov 077125, Romania
- ¹¹⁹ Consejo Superior de Investigaciones Científicas, Calle Serrano 117, 28006, Madrid, Spain
- ¹²⁰ Universidad de La Laguna, Dpto. Astrofísica, E-38206 La Laguna, Tenerife, Spain
- ¹²¹ Dipartimento di Fisica e Astronomia “G. Galilei”, Università di Padova, Via Marzolo 8, 35131 Padova, Italy
- ¹²² Institut für Theoretische Physik, University of Heidelberg, Philosophenweg 16, 69120 Heidelberg, Germany
- ¹²³ Institut de Recherche en Astrophysique et Planétologie (IRAP), Université de Toulouse, CNRS, UPS, CNES, 14 Av. Edouard Belin, 31400 Toulouse, France
- ¹²⁴ Université St Joseph; Faculty of Sciences, Beirut, Lebanon
- ¹²⁵ Departamento de Física, FCFM, Universidad de Chile, Blanco Encalada 2008, Santiago, Chile
- ¹²⁶ Institut d'Estudis Espacials de Catalunya (IEEC), Edifici RDIT, Campus UPC, 08860, Castelldefels Barcelona, Spain
- ¹²⁷ Satlantis, University Science Park, Sede Bld, 48940 Leioa-Bilbao, Spain
- ¹²⁸ Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans s/n, 08193 Barcelona, Spain
- ¹²⁹ Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Tapada da Ajuda, 1349-018 Lisboa, Portugal
- ¹³⁰ Cosmic Dawn Center (DAWN), Copenhagen, Denmark
- ¹³¹ Niels Bohr Institute, University of Copenhagen, Jagtvej 128, 2200 Copenhagen, Denmark
- ¹³² Universidad Politécnica de Cartagena, Departamento de Electrónica y Tecnología de Computadoras, Plaza del Hospital 1, 30202 Cartagena, Spain
- ¹³³ Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA
- ¹³⁴ Dipartimento di Fisica e Scienze della Terra, Università degli Studi di Ferrara, Via Giuseppe Saragat 1, 44122 Ferrara, Italy
- ¹³⁵ Istituto Nazionale di Fisica Nucleare, Sezione di Ferrara, Via Giuseppe Saragat 1, 44122 Ferrara, Italy
- ¹³⁶ INAF, Istituto di Radioastronomia, Via Piero Gobetti 101, 40129 Bologna, Italy
- ¹³⁷ Instituto de Astrofísica de Canarias, E-38205 La Laguna; Universidad de La Laguna, Dpto. Astrofísica, E-38206 La Laguna, Tenerife, Spain
- ¹³⁸ Université PSL, Observatoire de Paris, Sorbonne Université, CNRS, LERMA, 75014 Paris, France
- ¹³⁹ Université Paris-Cité, 5 Rue Thomas Mann, 75013 Paris, France
- ¹⁴⁰ Department of Mathematics and Physics E. De Giorgi, University of Salento, Via per Arnesano, CP-I93, 73100 Lecce, Italy
- ¹⁴¹ INFN, Sezione di Lecce, Via per Arnesano, CP-193, 73100 Lecce, Italy
- ¹⁴² INAF-Sezione di Lecce, c/o Dipartimento Matematica e Fisica, Via per Arnesano, 73100 Lecce, Italy
- ¹⁴³ INAF – Osservatorio Astronomico di Brera, Via Emilio Bianchi 46, 23807 Merate, Italy
- ¹⁴⁴ INAF-Osservatorio Astronomico di Brera, Via Brera 28, 20122, Milano, Italy, and INFN-Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy
- ¹⁴⁵ ICL, Junia, Université Catholique de Lille, LITL, 59000 Lille, France
- ¹⁴⁶ ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data e Quantum Computing, Via Magnanelli 2, Bologna, Italy
- ¹⁴⁷ Instituto de Física Teórica UAM-CSIC, Campus de Cantoblanco, 28049, Madrid, Spain
- ¹⁴⁸ CERCA/ISO, Department of Physics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA
- ¹⁴⁹ Laboratoire Univers et Théorie, Observatoire de Paris, Université PSL, Université Paris Cité, CNRS, 92190 Meudon, France
- ¹⁵⁰ Departamento de Física Fundamental. Universidad de Salamanca., Plaza de la Merced s/n, 37008 Salamanca, Spain
- ¹⁵¹ Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, 67000 Strasbourg, France
- ¹⁵² Center for Data-Driven Discovery, Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan

- ¹⁵³ Dipartimento di Fisica – Sezione di Astronomia, Università di Trieste, Via Tiepolo 11, 34131 Trieste, Italy
- ¹⁵⁴ California Institute of Technology, 1200 E California Blvd, Pasadena, CA 91125, USA
- ¹⁵⁵ Department of Physics & Astronomy, University of California Irvine, Irvine, CA 92697, USA
- ¹⁵⁶ CEA Saclay, DFR/IRFU, Service d’Astrophysique, Bat. 709, 91191 Gif-sur-Yvette, France
- ¹⁵⁷ Department of Computer Science, Aalto University, PO Box 15400, Espoo FI-00 076, Finland
- ¹⁵⁸ Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute (AIRUB), German Centre for Cosmological Lensing (GCCL), 44780 Bochum, Germany
- ¹⁵⁹ Department of Physics and Astronomy, University of Turku, Vesilinnantie 5, 20014 Turku, Finland
- ¹⁶⁰ Serco for European Space Agency (ESA), Camino bajo del Castillo s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada 28692, Madrid, Spain
- ¹⁶¹ ARC Centre of Excellence for Dark Matter Particle Physics, Melbourne, Australia
- ¹⁶² Centre for Astrophysics & Supercomputing, Swinburne University of Technology, Hawthorn, Victoria 3122, Australia
- ¹⁶³ Department of Physics and Astronomy, University of the Western Cape, Bellville, Cape Town 7535, South Africa
- ¹⁶⁴ DAMTP, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, UK
- ¹⁶⁵ Kavli Institute for Cosmology Cambridge, Madingley Road, Cambridge, CB3 0HA, UK
- ¹⁶⁶ Department of Astrophysics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
- ¹⁶⁷ Institute for Theoretical Particle Physics and Cosmology (TTK), RWTH Aachen University, 52056 Aachen, Germany
- ¹⁶⁸ IRFU, CEA, Université Paris-Saclay, 91191 Gif-sur-Yvette Cedex, France
- ¹⁶⁹ Oskar Klein Centre for Cosmoparticle Physics, Department of Physics, Stockholm University, Stockholm SE-106 91, Sweden
- ¹⁷⁰ Astrophysics Group, Blackett Laboratory, Imperial College London, London SW7 2AZ, UK
- ¹⁷¹ Univ. Grenoble Alpes, CNRS, Grenoble INP, LPSC-IN2P3, 53, Avenue des Martyrs, 38000 Grenoble, France
- ¹⁷² Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 2, 00185 Roma, Italy
- ¹⁷³ Centro de Astrofísica da Universidade do Porto, Rua das Estrelas, 4150-762 Porto, Portugal
- ¹⁷⁴ HE Space for European Space Agency (ESA), Camino bajo del Castillo s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada 28692, Madrid, Spain
- ¹⁷⁵ Department of Astrophysical Sciences, Peyton Hall, Princeton University, Princeton, NJ 08544, USA
- ¹⁷⁶ Theoretical astrophysics, Department of Physics and Astronomy, Uppsala University, Box 516, 751 37 Uppsala, Sweden
- ¹⁷⁷ Mathematical Institute, University of Leiden, Einsteinweg 55, 2333, CA Leiden, The Netherlands
- ¹⁷⁸ Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
- ¹⁷⁹ Space Physics and Astronomy Research Unit, University of Oulu, Pentti Kaiteran katu 1, FI-90014 Oulu, Finland
- ¹⁸⁰ Department of Physics and Astronomy, Lehman College of the CUNY, Bronx, NY 10468, USA
- ¹⁸¹ American Museum of Natural History, Department of Astrophysics, New York, NY 10024, USA
- ¹⁸² Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, 10010 New York, NY, USA
- ¹⁸³ Department of Physics and Astronomy, University of British Columbia, Vancouver, BC V6T 1Z1, Canada

Appendix A: Calibration curves for each model

The mapping from model rank to calibrated probability for each model is shown Fig. A.1. To apply these calibration mappings to the test set, we interpolate these rank values to the original model scores, such that the calibration function can map between model score and calibrated probability on new data. The best performing models can achieve calibrated probabilities $O(1)$, indicating that their highest scoring systems have a high lens purity.

Appendix B: Choice of classifiers to combine into an ensemble

The main ML ensemble presented in this work was composed of Models 1, 2a, and 4. This permutation of models was identified by adding these classifiers one by one to the ensemble (in decreasing order of classifier performance) until the ensemble performance peaked. The ROC curves in Fig. B.1 demonstrate the performance of different ensembles using a range of different classifiers. In particular, the pink curves show ensembles composed of an increasing number of ML classifiers, added in decreasing order of their performance. The best individual ML classifier (Model 4) is shown for ease of comparison. The best ML-only ensemble is found from using 3 classifiers, though in practice this is very similar to an ensemble of 4 ML classifiers. In Fig. B.1 we investigate different methods for combining the ML and Model 0 (citizen science) classifiers. The black curve shows an ensemble of Model 0 and the ML classifiers, following Eq. (1). The blue curve shows the performance where the ML ensemble is replaced by the Model 0 posterior where available, i.e. excluding information from the ML models if citizen classifications are available. We find the former classifiers consistently show improved performance over the latter classifiers in the low FPR region (the region of interest).

Figure B.2 shows the performance of a range of ensembles now including Model 0 (citizen science). Having combined Model 0 and Model 4, there are only small further changes in the performance of the ensemble, likely reflecting the wide range in performance of the constituent ML classifiers. The ensemble that includes all models provides slightly better performance in the low false positive rate region (the region of interest) and is thus adopted in the remainder of this work.

Appendix C: The effect of retraining Model 4 with different seeds

Given the strong performance of Model 4, we investigated the benefit of retraining Model 4. Figure C.1 depicts the performance of 10 versions of this model, trained on different seeds. The original Model 4 is also shown for comparison. We find there is a relatively large spread in the performance of this model dependent on seed. Accordingly, there is some benefit to be gained from combining these different initialisations. The right-hand panel of Fig. C.1 depicts a range of ensembles utilising these additional initialisations. Applying the Bayesian combination method to combine the Model 4 initialisations would not be well posed given that the initialisations have highly correlated results, and thus the resultant posteriors from the Bayesian ensemble method used (which assumes independence) would be too optimistic. Therefore, a simple mean average is used in this case and is referred to subsequently as Model 4'. A weighted average score from these different initialisations (based on a separate validation set) could also have been used, but is not explored here. Model 4' has a comparable performance to the

original ML-only ensemble (using the original Model 4). An updated ML-only ensemble, swapping the original Model 4 for Model 4', further improves the performance over the original ML-only ensemble and Model 4'. However, this improvement is small, likely due to the greater performance difference between Model 4' and the next best performing ML classifier. Finally, the all-classifier ensemble (orange) now replacing Model 4 with Model 4' is shown to have the best performance. This improvement alludes to a more general point, that in cases where a network shows varying performance across different seed initialisations, it would be beneficial to present and use an ensemble of these initialisations (i.e. Model 4' rather than Model 4 here) to maximise performance and to ensure that the results presented are not affected by the choice of seed.

Appendix D: Effect of calibration on classifier combination

Figure D.1 shows the ROC curves generated by simply averaging the uncalibrated model outputs compared to first calibrating each model before combining them in a Bayesian manner. The calibration provides significant improvement over the uncalibrated ensembles. Furthermore, the uncalibrated classifier performs much worse than its constituent parts in the case of the ML-only ensemble.

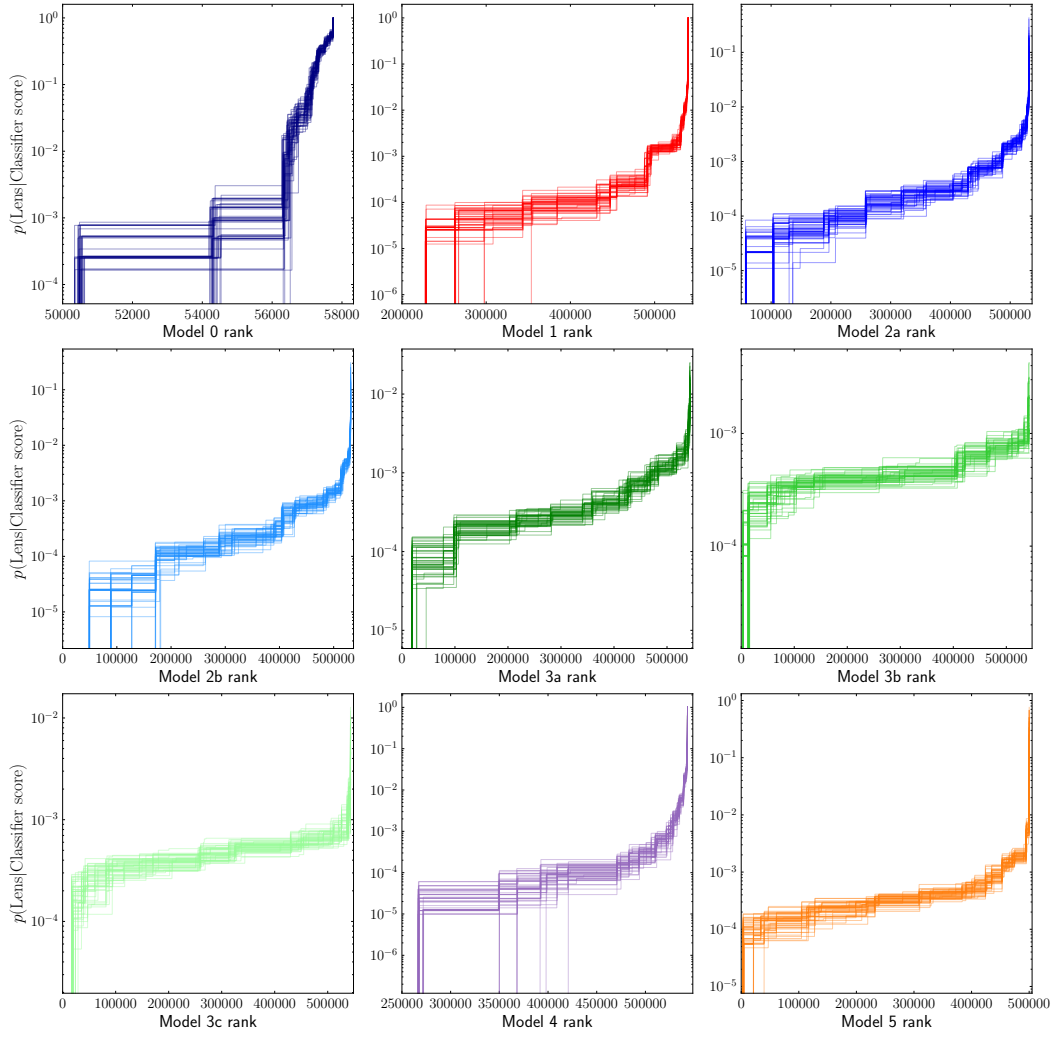


Fig. A.1. Mapping from the ranking of each object in the test set to calibrated probability, based on the isotonic regression (Zadrozny et al. 2002) procedure. We show 50 calibration curves for each model, generated by bootstrapping. The lower x -axis limit is trimmed for clarity to where the mapping is non-zero. The limits for Model 0 are more restricted, since the citizens were only shown a subset of the whole dataset, and a high proportion of the lowest scoring systems received a calibrated probability of 0.

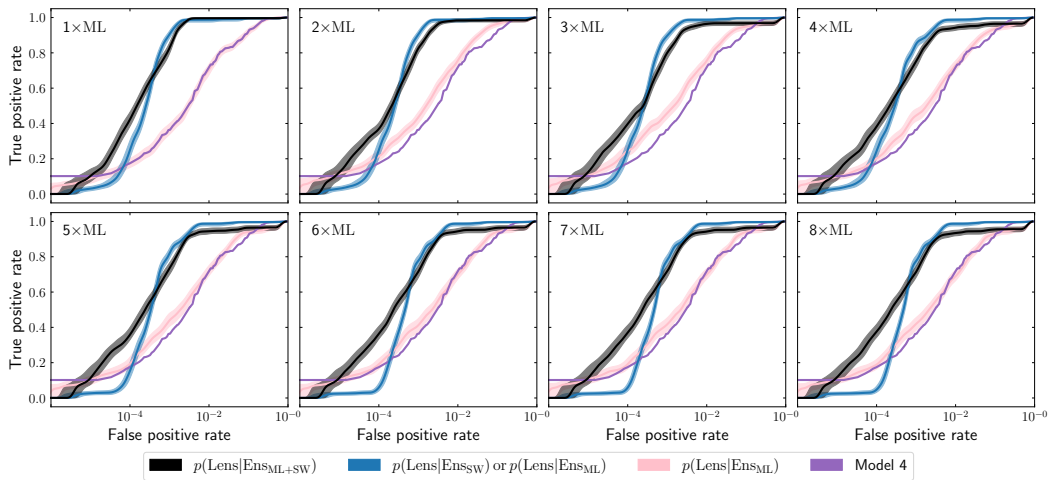


Fig. B.1. ROC curves for a range of ensembles, generated by sequentially adding ML classifiers in descending order of their performance (specifically 4, 1, 2a, 2b, 5, 3a, 3b, 3c), from only one classifier (trivial ensemble, *top left*) to an ensemble of all ML classifiers (*bottom right*). The pink curve shows the ML-only ensemble in each panel. The black curve depicts the performance when this ensemble is further combined with Model 0 (Space Warps). The blue curve shows the performance of a classifier where the ML+Model 0 ensemble posterior is replaced by the Model 0 posterior (alone) where available, i.e. excluding information from the ML models if citizen classifications are available. Finally, the best performing ML classifier (Model 4) is shown for comparison.

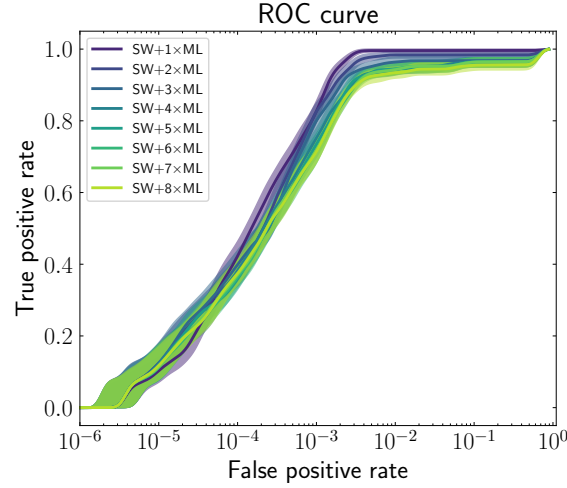


Fig. B.2. ROC curves for a range of ensembles, generated from combining Model 0 with an increasing number of ML models, added in decreasing order of their performance (specifically 4, 1, 2a, 2b, 5, 3a, 3b, 3c). Having combined Model 0 and Model 4, the performances of these ensembles are similar, likely reflecting the relatively wide difference in performance between Model 0 and the next best ML classifier.

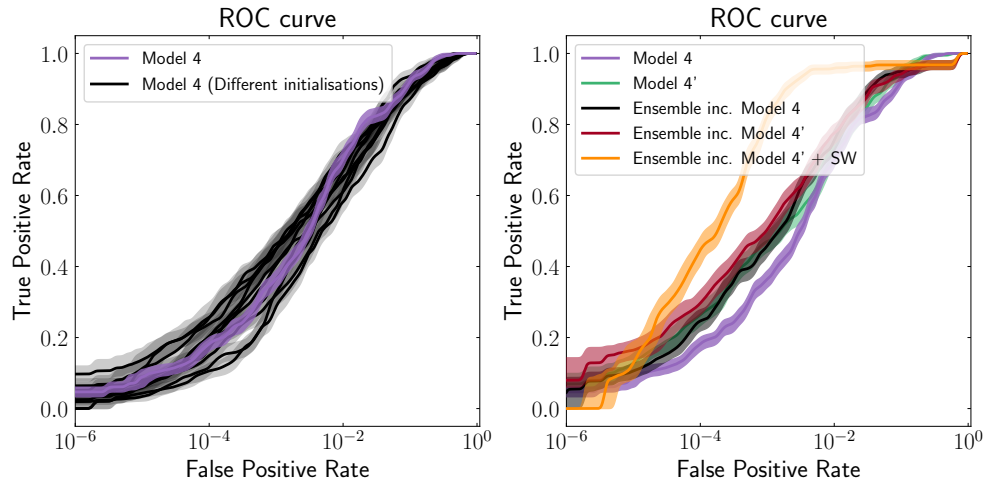


Fig. C.1. ROC curves demonstrating the effect of retraining Model 4 using different seed initialisations. *Left:* ROC curves (black) of the individual initialisations of Model 4, applied to the test set. The original Model 4 is also shown (purple) for comparison in both panels. *Right:* ROC curves for a range of different ensembles incorporating Model 4' (the mean score of the different initialisations). The ROC curve for Model 4' is shown in green, which has comparable performance to the original ML-only ensemble (black, composed of Model 1, 2a, and 4). An updated ML-only ensemble (composed of Model 1, 2a, and 4') is shown in red, demonstrating further improvement in performance. The ensemble performance when including Model 0 (citizen science) is shown in orange, which has the best performance.

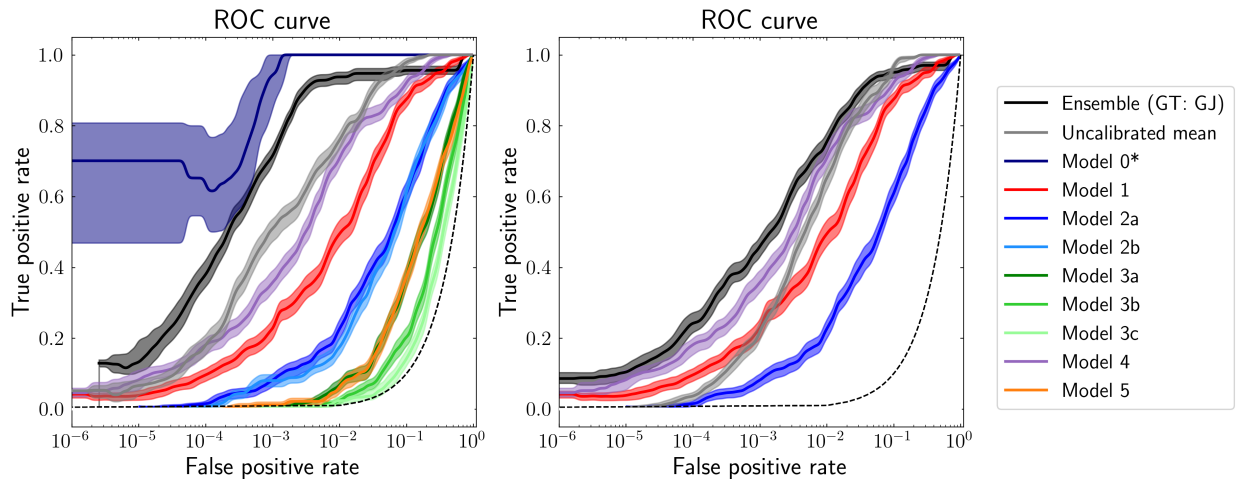


Fig. D.1. Comparison of the ensemble ROC curves generated via calculating the mean average of the uncalibrated model outputs (grey), versus Bayesian combination following calibration (black). The ensemble of Models 1, 2a, and 4 is shown on the *right*, and the ensemble of all models is shown on the *left*.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