Astronomy & Astrophysics manuscript no. output
January 28, 2026

Euclid Quick Data Release (Q1)

From simulations to sky: Advancing machine-learning lens detection with real *Euclid* data

Euclid Collaboration: N. E. P. Lines[⋆1], T. E. Collett[1], P. Holloway[1], K. Rojas[2], S. Schuldt[3,4], R. B. Metcalf[5,6], T. Li[1], A. Verma[7], G. Despali[5,6,8], F. Courbin[9,10,11], R. Gavazzi[12,13], C. Tortora[14], B. Clément[15,16], N. Aghanim[17], B. Altieri[18], L. Amendola[19], S. Andreon[20], N. Auricchio[6], C. Baccigalupi[21,22,23,24], M. Baldi[25,6,8], A. Balestra[26], S. Bardelli[6], P. Battaglia[6], A. Biviano[22,21], E. Branchini[27,28,20], M. Brescia[29,14], S. Camera[30,31,32], G. Cañas-Herrera[33,34], V. Capobianco[32], C. Carbone[4], J. Carretero[35,36], M. Castellano[37], G. Castignani[6], S. Cavuoti[14,38], A. Cimatti[39], C. Colodro-Conde[40], G. Congedo[33], C. J. Conselice[41], L. Conversi[42,18], Y. Copin[43], H. M. Courtois[44], M. Cropper[45], H. Degaudenzi[46], G. De Lucia[22], H. Dole[17], F. Dubath[46], X. Dupac[18], S. Dusini[47], A. Ealet[43], S. Escoffier[48], M. Farina[49], R. Farinelli[6], F. Faustini[37,50], S. Ferriol[43], F. Finelli[6,51], M. Frailis[22], E. Franceschi[6], M. Fumana[4], S. Galeotta[22], K. George[52], B. Gillis[33], C. Giocoli[6,8], P. Gómez-Alvarez[53,18], J. Gracia-Carpio[54], A. Grazian[26], F. Grupp[54,55], S. V. H. Haugan[56], W. Holmes[57], I. M. Hook[58], F. Hormuth[59], A. Hornstrup[60,61], K. Jahnke[62], M. Jhabvala[63], B. Joachimi[64], E. Keihänen[65], S. Kermiche[48], A. Kiessling[57], B. Kubik[43], M. Kümmel[55], M. Kunz[66], H. Kurki-Suonio[67,68], A. M. C. Le Brun[69], S. Ligori[32], P. B. Lilje[56], V. Lindholm[67,68], I. Lloro[70], G. Mainetti[71], D. Maino[3,4,72], E. Maiorano[6], O. Mansutti[22], S. Marcin[2], O. Marggraf[73], M. Martinelli[37,74], N. Martinet[12], F. Marulli[5,6,8], R. J. Massey[75], E. Medinaceli[6], S. Mei[76,77], M. Melchior[78], Y. Mellier[79,13], M. Meneghetti[6,8], E. Merlin[37], G. Meylan[15], A. Mora[80], M. Moresco[5,6], L. Moscardini[5,6,8], R. Nakajima[73], C. Neissner[81,36], S.-M. Niemi[82], J. W. Nightingale[83], C. Padilla[81], S. Paltani[46], F. Pasian[22], K. Pedersen[84], W. J. Percival[85,86,87], V. Pettorino[82], S. Pires[88], G. Polenta[50], M. Poncet[89], L. A. Popa[90], L. Pozzetti[6], F. Raison[54], A. Renzi[91,47], J. Rhodes[57], G. Riccio[14], E. Romelli[22], M. Roncarelli[6], C. Rosset[76], R. Saglia[55,54], Z. Sakr[19,92,93], A. G. Sánchez[54], D. Sapone[94], B. Sartoris[55,22], J. A. Schewtschenko[33], P. Schneider[73], T. Schrabback[95], A. Secroun[48], G. Seidel[62], S. Serrano[96,97,98], C. Sirignano[91,47], G. Sirri[8], L. Stanco[47], J. Steinwagner[54], P. Tallada-Crespí[35,36], A. N. Taylor[33], I. Tereno[99,100], N. Tessore[45], S. Toft[101,102], R. Toledo-Moreo[103], F. Torradeflot[36,35], I. Tutusaus[98,96,92], J. Valiviita[67,68], T. Vassallo[22], A. Veropalumbo[20,28,27], Y. Wang[104], J. Weller[55,54], A. Zacchei[22,21], G. Zamorani[6], F. M. Zerbi[20], E. Zucca[6], M. Ballardini[105,106,6], M. Bolzonella[6], E. Bozzo[46], C. Burigana[107,51], R. Cabanac[92], M. Calabrese[108,4], A. Cappi[109,6], T. Castro[22,23,21,110], J. A. Escartin Vigo[54], L. Gabarra[7], J. García-Bellido[111], V. Gautard[112], S. Hemmati[104], M. Huertas-Company[40,113,114], J. Macias-Perez[115], R. Maoli[116,37], J. Martín-Fleitas[117], M. Maturi[19,118], N. Mauri[39,8], P. Monaco[119,22,23,21], M. Pöntinen[67], C. Porciani[73], I. Risso[20,28], V. Scottez[79,120], M. Sereno[6,8], M. Tenti[8], M. Tucci[46], M. Viel[21,22,24,23,110], M. Wiesmann[56], Y. Akrami[111,121], I. T. Andika[52,122], G. Angora[14,105], S. Anselmi[47,91,123], M. Archidiacono[3,72], F. Atrio-Barandela[124], E. Aubourg[76,125], L. Bazzanini[105,6], D. Bertacca[91,26,47], M. Bethermin[126], F. Beutler[33], A. Blanchard[92], L. Blot[127,69], M. Bonici[85,4], S. Borgani[119,21,22,23,110], M. L. Brown[41], S. Bruton[128], A. Calabro[37], B. Camacho Quevedo[21,24,22], F. Caro[37], C. S. Carvalho[100], F. Cogato[5,6], S. Conseil[43], A. R. Cooray[129], O. Cucciati[6], S. Davini[28], F. De Paolis[130,131,132], G. Desprez[133], A. Díaz-Sánchez[134], S. Di Domizio[27,28], J. M. Diego[135], P.-A. Duc[126], V. Duret[48], M. Y. Elkhashab[22,23,119,21], A. Enia[6], Y. Fang[55], P. G. Ferreira[7], A. Finoguenov[67], A. Fontana[37], A. Franco[131,130,132], K. Ganga[76], T. Gasparetto[37], E. Gaztanaga[98,96,1], F. Giacomini[8], F. Gianotti[6], G. Gozaliasl[136,67], A. Gruppuso[6,8], M. Guidi[25,6], C. M. Gutierrez[137], A. Hall[33], H. Hildebrandt[138], J. Hjorth[84], J. J. E. Kajava[139,140], Y. Kang[46], V. Kansal[141,142], D. Karagiannis[105,143], K. Kiiveri[65], J. Kim[7], C. C. Kirkpatrick[65], S. Kruk[18], M. Lattanzi[106], L. Legrand[144,145], F. Lepori[146], G. Leroy[147,75], G. F. Lesci[5,6], J. Lesgourgues[148], T. I. Liaudat[125], M. Magliocchetti[49], A. Manjón-García[134], F. Mannucci[149], C. J. A. P. Martins[150,151], L. Maurin[17], M. Miluzio[18,152], A. Montoro[98,96], C. Moretti[22,21,23], G. Morgante[6], S. Nadathur[1], K. Naidoo[1,62], P. Natoli[105,106], S. Nesseris[111], D. Paoletti[6,51], F. Passalacqua[91,47], K. Paterson[62], L. Patrizii[8], A. Pisani[48], D. Potter[146], G. W. Pratt[88], S. Quai[5,6], M. Radovich[26], W. Roster[54], S. Sacquegna[153], M. Sahlén[154], D. B. Sanders[155], E. Sarpa[24,110,23], A. Schneider[146], D. Sciotti[37,74], E. Sellentin[156,34], L. C. Smith[157], J. G. Sorce[158,17], K. Tanidis[7], C. Tao[48], F. Tarsitano[159,46], G. Testera[28], R. Teyssier[160], S. Tosi[27,28,20], A. Troja[91,47], A. Venhola[161], D. Vergani[6], G. Vernardos[162,163], G. Verza[164,165], S. Vinciguerra[12], M. Walmsley[166,41], N. A. Walton[157], and A. H. Wright[138]

(Affiliations can be found after the references)

January 28, 2026

ABSTRACT

In the era of large-scale surveys such as *Euclid*, machine learning has become an essential tool for identifying rare yet scientifically valuable objects, such as strong gravitational lenses. However, supervised machine-learning approaches require large quantities of labelled examples to train on, and the limited number of known strong lenses has led to a reliance on simulations for training. A well-known challenge is that machine-learning models trained on one data domain often underperform when applied to a different domain: in the context of lens finding, this means that strong performance on simulated lenses does not necessarily translate into equally good performance on real observations. In the *Euclid* Quick Data Release 1 (Q1), covering 63 deg^2, 500 strong lens candidates were discovered through a synergy of machine learning, citizen science, and expert visual inspection. These discoveries now allow us to quantify this performance gap and investigate the impact of training on real data. We find that a network trained only on simulations recovers up to 92% of simulated lenses with 100% purity, but only achieves 50% completeness with 24% purity on real *Euclid* data. By augmenting training data with real *Euclid* lenses and non-lenses, completeness improves by 25–30% in terms of the expected yield of discoverable lenses in the *Euclid* Data Release 1 and the full Euclid Wide Survey. Roughly 20% of this improvement comes from the inclusion of real lenses in the training data, while 5–10% comes from exposure to a more diverse set of non-lenses and false positives from Q1. We show that the most effective lens-finding strategy for real-world performance combines the diversity of simulations with the fidelity of real lenses. This hybrid approach establishes a clear methodology for maximising lens discoveries in future data releases from *Euclid* and will likely also be applicable to other surveys such as the Vera Rubin Observatory's Legacy Survey of Space and Time.

Key words. Gravitational lensing: strong – Methods: data analysis – Surveys

1. Introduction

Strong gravitational lensing is a unique and powerful probe of astrophysics and cosmology, providing direct insights into phenomena that are otherwise difficult to observe. By mapping the deflection of light, strong lenses allow a direct measurement of the total mass of the lens, tracing the combination of its luminous and dark matter. On the galaxy-scale, this provides a method of probing total dark matter contributions (Gavazzi et al. 2007; Auger et al. 2010) and the presence of dark matter subhaloes (e.g. Vegetti et al. 2010, 2012; O'Riordan et al. 2023; Ertl et al. 2024), as well as other mass components such as supermassive black holes (Melo-Carneiro et al. 2025; Nightingale et al. 2023). Cluster-scale strong lenses can additionally provide insights into dark matter on small scales (e.g. Natarajan et al. 2017; Meneghetti et al. 2020, 2022, 2023; Dutra et al. 2025) and offer exceptional magnification and resolving power for the study of distant background sources (e.g. Vanzella et al. 2017, 2020, 2023; Adamo et al. 2024; Meštrić et al. 2022; Welch et al. 2022; Fujimoto et al. 2025; Bradley et al. 2025). Time-delay measurements from strongly lensed transients on both galaxy and cluster-scales provide an independent route to constrain the Hubble constant (e.g. Kelly et al. 2023; Grillo et al. 2024; Tdcosmo Collaboration et al. 2025; Pascale et al. 2025; Suyu et al. 2025), and strong lensing cosmography can provide measurements of cosmological parameters such as the equation of state of dark energy (e.g. Jullo et al. 2010; Caminha et al. 2016, 2022; Moresco et al. 2022; Li et al. 2024). Strong lenses are scientifically valuable, but unfortunately are intrinsically rare, with only a few thousand candidate systems known.

Euclid (Euclid Collaboration: Mellier et al. 2025) is set to revolutionise strong lensing through its unique synergy of wide-field coverage (14 000 deg^2) and high angular resolution (0.″16 point-spread-function (PSF) full width half maximum in the optical filter) across the Euclid Wide Survey (EWS). Forecasts predict that around 170 000 galaxy-scale strong lenses should be detectable in the full survey (Collett 2015; Acevedo Barroso et al. 2025b). This unprecedented sample, exceeding the total number of previously known strong lenses by around two orders of magnitude, will be transformative for the field. This vast dataset will enable a wealth of new scientific insights into galaxy evolution, dark matter, and cosmology.

With these unprecedented quantities of data comes a new era of lens searching. Currently, the most reliable method for finding strong lenses is visual inspection by experts (Pearce-Casey et al. 2024; Acevedo Barroso et al. 2025b). However, visually inspecting all 1.5 billion galaxies imaged by *Euclid* to uncover these strong lenses is intractable, and hence help from automated techniques is necessary. Because of the complexity of the lens-finding challenge, deep machine-learning (ML) networks have proved to be one of the most promising methods for addressing it. In recent years, ML models have been successfully used to find strong lenses in a wealth of astronomical data. These span ground-based large-area surveys such as the Canada–France–Hawaii Telescope Legacy Survey (CFHTLS; e.g. Jacobs et al. 2017), Dark Energy Survey (DES; e.g. Jacobs et al. 2019b,a; Rojas et al. 2022; González et al. 2025), Hyper Suprime-Cam (HSC; e.g. Sonnenfeld et al. 2018, 2020; Cañameras et al. 2021; Shu et al. 2022; Wong et al. 2022; Jaelani et al. 2024; Schuldt et al. 2025a,b), Kilo-Degree Survey (KiDS; e.g. Petrillo et al. 2017, 2019; Li et al. 2020, 2021; Nagam et al. 2023, 2024; Grespan et al. 2024), Ultraviolet Near-Infrared Optical Northern Survey (UNIONS; e.g. Savary et al. 2022; Acevedo Barroso et al. 2025a), Dark Energy Spectroscopic Instrument (DESI) Legacy Survey (e.g. Huang et al. 2020, 2021; Storfer et al. 2024), and Panoramic Survey Telescope and Rapid Response System (PanSTARRS; e.g. Cañameras et al. 2020), as well as space-based imaging from the *Hubble* Space Telescope (HST; e.g. Pourrahmani et al. 2018; Teimoorinia et al. 2020).

The first major data release from the *Euclid* survey, Quick Release 1 (Q1; Euclid Collaboration: Aussel et al. 2025), covers 63 deg^2 and provides a first glimpse into what *Euclid* can achieve. In the initial Q1 search, around 500 galaxy-scale strong lens candidates were found through a combination of ML, citizen science, and expert visual inspection, covered by the paper series 'The Strong Lensing Discovery Engine' (SLDE), namely Euclid Collaboration: Walmsley et al. (2025), Euclid Collaboration: Rojas et al. (2025), Euclid Collaboration: Lines et al. (2025), Euclid Collaboration: Li et al. (2025), and Euclid Collaboration: Holloway et al. (2025). In addition, around 80 group- and cluster-scale lenses were identified from visual inspection

⋆ e-mail: natalie.lines@port.ac.uk

of cluster fields (Euclid Collaboration: Bergamini et al. 2025). Compared to future releases, a relatively large proportion of Q1 data could be visually inspected to enable this quantity of lens discoveries. Visual inspection budgets can be stretched through crowd-sourcing with citizen science to help process growing data volumes; however, the limited scalability of visual inspection necessitates complementary advances in ML approaches. Machine-learning classifiers can be combined with citizen science as an effective strategy to optimise the use of available human effort (Holloway et al. 2024). Such approaches nevertheless benefit from stronger individual ML classifiers, motivating efforts to improve standalone ML performance (Euclid Collaboration: Holloway et al. 2025).

Machine-learning algorithms are natural interpolators: trained on a fixed dataset, they typically perform well on unseen data from the same underlying distribution. However, known strong lenses are few and far between, and this scarcity poses a problem for training deep-learning models. Training a supervised ML model from scratch typically requires at least 10^4–10^5, but ideally millions, of labelled images (Sun et al. 2017). The number of identified lens candidates within any one survey is typically of the order 10^2–10^3 at most, and hence training ML models for detecting lenses generally relies on simulations. Although these simulations are made to be as realistic as possible by adding lensed arcs on top of real images of lensing galaxies, resulting in simulations that appear visually realistic to humans (e.g. see Fig. 1), it is hard to ensure that they are perfectly accurate and representative of the true lensing population. Additionally, ML models pick up subtle biases or artefacts in the simulations rather than the true astrophysical features that characterise strong lenses. As a result, an ML model that detects simulated strong lenses with high accuracy does not necessarily perform well at finding lenses in real data. This problem has been well-reported in ML strong lens finders: Jacobs et al. (2017) trained convolutional neural networks (CNNs) to find lenses in CFHTLS data using two different lens simulation approaches. They found that while the CNNs excelled at detecting their respective types of simulated lenses, they underperformed on the other type of simulated lenses and the performance did not translate well to real data. In Metcalf et al. (2019), ten different lens-finding approaches were tested on simulated ground-based data as well as real images from KiDS data. All ten performed significantly better on the simulated data than on the KiDS data. Pearce-Casey et al. (2024) found that training the same ML model on different simulated data produced very different performance, and that these networks consistently underperformed on real data compared to the simulations they were trained on. Cañameras et al. (2024) explored how varying training data (both positive and negative) impacts the ability of ML models to recover real lenses from HSC data. They found that the choice of simulated lenses for training strongly impacts performance, and that boosting the fraction of usual contaminants in the negative training class improves performance.

Although most ML models have relied on simulations, training on real data alone is not without precedent: neural networks have been successfully developed to find strong lenses in DESI Legacy Imaging Surveys, using a training set of real lens (of which there were as few as 700) and non-lens images alone (Huang et al. 2020, 2021; Storfer et al. 2024). Additionally, the limited number of real lenses available for ML training can be mitigated by using the human-in-the-loop pipeline to dynamically build the training set (e.g. Xu et al. in prep.). However, when training on real lenses alone, it is hard to guarantee that these training sets are representative of the real-world distribution of lenses. Training on these alone may introduce selection biases. Therefore, despite the domain gap, training on simulations remains beneficial.

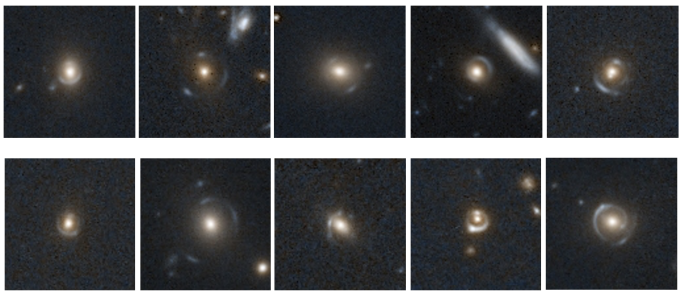

Fig. 1. *Top*: Simulated *Euclid* lenses from Euclid Collaboration: Rojas et al. (2025). *Bottom*: Real lenses found in Q1.

The discrepancy in performance when shifting between domains is a well-known and pervasive challenge in ML, extending beyond lens finding and astronomy (Zhou et al. 2023; Zhang & Gao 2024), meaning that domain adaptation has been a major topic in computer vision. Humans intuitively learn to recognise high-level semantic features in images, enabling us to generalise across variations in appearance. In contrast, ML models often rely more heavily on low-level properties such as noise patterns and synthetic artefacts. Simulations can contain subtle statistical differences that are more recognisable in high-dimensional feature spaces, where small mismatches in low-level image statistics can accumulate and become recognisable to ML networks (Zhou et al. 2023). Deep networks, with their large numbers of parameters, have a tendency to exploit these domain-specific features rather than learn generalisable, domain-invariant properties. As a result, a model trained on one domain (e.g. simulated lenses) often suffers a significant drop in performance when applied to a different domain (e.g. real *Euclid* data, Farahani et al. 2021). Domain adaptation techniques such as domain-adversarial neural networks (DANNs; Ganin et al. 2016) have been explored to address this domain shift between simulations and real observations for other astronomical tasks, such as classifying ionised nebulae (Belfiore et al. 2025) and the study of galaxy mergers (Ćiprijanović et al. 2020). Using pre-trained models has also been shown to enhance adaptation to out-of-domain data (Hendrycks et al. 2019) and improve robustness by combining knowledge from multiple domains (Niu et al. 2020). However, producing domain-invariant ML models remains very challenging, and eliminating the performance drop associated with this domain shift is an open problem.

The Q1 sample of around 500 *Euclid* lens candidates now provides the first statistically meaningful dataset for testing and improving ML algorithms on real *Euclid* data. In particular, it enables us to: (i) quantify the discrepancy between model performance on simulations and real data; (ii) assess how incorporating real lenses, as well as previously misclassified contaminants, into the training set improves model robustness; and (iii) explore how domain adaptation may scale as the number of confirmed lenses grows in future *Euclid* releases. In this paper, we explore these aspects by varying training and testing data with a fine-tuned version of the Zoobot ML model, the most efficient lens-finding network in the initial Q1 lens search (Euclid Collaboration: Lines et al. 2025). Closing the domain gap is expected to produce more accurate, reliable lens-finding models, which will be better equipped to discover lenses in future *Euclid* data releases.

The paper is structured as follows. Section 2 covers the data available prior to Q1, as well as the labelled lenses and non-lenses from Q1 now used for training and testing. In Sect. 3, we describe the ML model and our approach for using these new data to investigate ML performance. Our results are presented in Sect. 4, where we first report the results from using the Q1 data to quantify performance relative to pre-Q1 data, then report the impact of using Q1 data in training. Conclusions are presented in Sect. 5.

2. Data used for training and testing

The data used in this work consist of that available prior to Q1, which were used to develop the ML models used in Q1, and the labelled Q1 lenses and non-lenses, which resulted from the SLDE search. The former dataset ('pre-Q1 lenses and non-lenses'; see Sect. 2.1) benefits from a larger number of positive training examples, whereas the latter ('Q1 lenses and non-lenses'; see Sect. 2.2) offers the advantage of being drawn directly from real observations. The datasets are summarised in Table 1. Most of these data are introduced in Euclid Collaboration: Lines et al. (2025), which includes details on how the distribution of parameters such as the Einstein radius varies between the simulations, Q1 lenses, and forecasts.

Table 1. Overview of the datasets used in this work, along with their sample sizes.

	Dataset	Size
Pre-Q1 positives	Simulations S1	11 057
	Simulations S2	3737
Pre-Q1 negatives	Classified non-lenses	5000
Q1 positives	grade A + grade B lenses	497
Q1 negatives	Randomly selected Q1 images	40 000
	ML false positives	78 214

2.1. Pre-Q1 data

The positive and negative training data available prior to Q1 and used originally to train the ML models are detailed in Euclid Collaboration: Lines et al. (2025), but their key features are outlined here. The pre-Q1 positives consisted of simulated lenses alone, since the number of known *Euclid* lenses at the time were too few. Two sets of simulations were used. The first simulation set (S1) is from Euclid Collaboration: Rojas et al. (2025), where the simulations were made by painting lensed arcs onto high-velocity-dispersion luminous red galaxies (LRGs), selected using DESI data. A singular isothermal ellipsoid (SIE) mass model was adopted, with parameters derived from the Sérsic fit, and Einstein radius calculated using the velocity dispersion of the lens and redshifts of the lens and source. Background sources were drawn from the HST Advanced Camera for Surveys F814W high-resolution catalogue (Leauthaud et al. 2007; Scoville et al. 2007; Koekemoer et al. 2007), combined with HSC colour information (Cañameras et al. 2020). Source images were lensed based on the mass model, downsampled to match *Euclid* VIS $I_{\rm E}$ image resolution, and finally convolved with the telescope PSF after adding the lensed features to the lens images.

The second set of simulations (S2; Metcalf et al. in prep.) was created by selecting *Euclid* images of all observed galaxies with $I_{\rm E} < 22$ and applying additional cuts to remove stars and reduce the number of face-on spirals. Each image was then matched to an object in the Flagship simulation (Euclid Collaboration: Castander et al. 2025) with a nearest-neighbour algorithm in the space of all magnitudes in all four bands, ellipticity, and redshift, when available for the observed galaxy. The parameters of the Flagship galaxy and dark matter halo were then used to construct a mass model for the lens. A synthetic source, created by combining between one and four Sérsic profiles, was then placed near or in the tangential caustic. The image of the lensed object was then convolved with the local PSF, and Poisson noise was added.

Notably, both simulation sets added simulated arcs to real *Euclid* images of galaxies to include as many features of *Euclid* imaging as possible – a well-established strategy of producing the most realistic mock lenses. This means that any peculiarity of the simulations that could be learnt by the ML models must reside within the lensed arcs, or are due to the fact that lens galaxies used in simulations are not from the exact same underlying distribution as that of real *Euclid* lenses.

Additionally, prior to Q1, a catalogue of human-classified non-lenses was compiled from a visual inspection of high-velocity-dispersion galaxies (Euclid Collaboration: Rojas et al. 2025). This catalogue includes approximately 2300 spiral galaxies, 60 ring galaxies, 250 mergers, and 2700 LRGs. These objects were originally used as the negative class in training. Prior to Q1 there was no catalogue of classified non-lenses with the same selection cut as that of the Q1 lens search. Consequently, the distribution of the pre-Q1 non-lenses differs from that of the non-lenses encountered in Q1. In particular, more common false positives, such as spirals and ring galaxies, are over-represented in the pre-Q1 set relative to LRGs, compared to their distribution in the Q1 data.

2.2. Q1 data

The Q1 data correspond to seven days worth of imaging from *Euclid* and make up just 0.45% of the full EWS. We work with visible imaging data from *Euclid*'s VIS instrument (Euclid Collaboration: Cropper et al. 2025). This paper builds upon the work from the original Q1 SLDE lens search, from which 497 strong lens candidates were discovered using these data. Briefly describing the original search, the method involved reducing the original catalogue of 30 million objects to 1 086 554 objects by selecting extended sources having $I_{\rm E} < 22.5$, along with additional selection cuts to remove likely stars and artefact. These 1 086 554 objects were scored by five ML models trained using the data outlined in Sect. 2.1, and these scores informed the selection of objects that were visually inspected by citizen scientists. In total, around 115 329 objects were visually inspected, including the top 20 000 ranked Q1 objects according to Zoobot, objects highly ranked by the other ML models, and 40 000 randomly selected Q1 objects – to represent the underlying Q1 population. Around 7000 objects, considered likely to be lenses according to the citizen science project, were then graded by strong lensing experts. This resulted in a catalogue of 497 strong lens candidates (corresponding to objects classified as grade A or grade B lenses). From a visual inspection of the 40 000 randomly selected Q1 objects and 78 214 objects selected by the combined ML models, a catalogue of around 100 000 classified non-lenses was produced, including a large number of common lens contaminants by construction. This catalogue is outlined in Euclid Collaboration: Walmsley et al. (2025).

Of the Q1 strong lens candidates, many objects show very clear lensing features with no other astrophysical explanation. However, without spectroscopic information, it is hard to determine for certain if an object is a true strong lens system or not.

Unfortunately, such data are expensive to obtain and, in their absence, we rely on visual inspection by strong lensing experts as the most robust method to determine what is or is not a strong lens candidate. Rojas et al. (2023) demonstrated that averaging the grades assigned by six or more experts is a reliable method of identifying strong lenses, and hence we can be fairly confident that the majority of the candidates in the Q1 sample are likely to be true strong lenses. In the future, spectroscopic data for 10 000 strong lens candidates will be provided by the 4MOST Strong Lensing Spectroscopic Legacy Survey (4SLSLS; Collett et al. 2023), which will allow more robust ML models to be trained.

We note that some *Euclid* lenses exist beyond this SLDE Q1 sample, which we exclude for simplicity. Euclid Collaboration: Rojas et al. (2025) discovered 38 grade A and 40 grade B lenses through the visual inspection of *Euclid* imaging of high-velocity-dispersion galaxies. There is some overlap between this catalogue and that of Euclid Collaboration: Walmsley et al. (2025), but approximately 22 (28) grade A (B) lenses were not included in the sample of 497 Q1 lenses considered here. However, many of these were excluded from the Q1 search, either because they did not pass the initial selection cut or because they are outside of the Q1 area and therefore lack data processing consistent with the Q1 data. Hence, we do not include them. Following our initial Q1 lens search, there have been other searches through the Q1 data for strong lenses (e.g. Ecker et al. in prep.; Xu et al. in prep.), although these catalogues were not finalised at the time of this work and additionally originate from a different parent sample and so are not considered.

3. Method

3.1. ML approach

We explored the impact of changing the training and testing data for ML performance at lens finding. We tested this using a fine-tuned version of the Zoobot foundation model as the ML architecture (Walmsley et al. 2023). The Zoobot foundation model is pre-trained on around 100 million Galaxy Zoo morphologies across data from a range of surveys (Lintott et al. 2008) and serves as a base model that can then be fine-tuned for more specific tasks, such as detecting strong lenses. In the original Q1 search, this was the best ML approach for finding lenses (Euclid Collaboration: Lines et al. 2025), although we note that since the Q1 release, other architectures, such as pre-trained vision transformers, have been shown to perform similarly or better than the Zoobot model (Vincken et al. in prep.). While we expect the general trends to be applicable to most ML models, we note that the quantitative improvement is likely to be architecture-dependent.

The fine-tuning procedure closely follows that of Euclid Collaboration: Lines et al. (2025). We used the ConvNeXT-Nano version of the architecture with 15.6 million parameters and fine-tuned the last three blocks. We used the same image preprocessing (including VIS-only images) and fixed the hyperparameters to the best values identified during Q1 training to ensure a fair comparison, varying only the training or testing datasets. We note that further performance improvement can likely be achieved by optimising the hyperparameters for each specific version of the model, but this paper aims to isolate and study general performance trends, rather than optimising for a single best-performing model.

3.2. Test sets

When evaluating performance on the pre-Q1 data, we reserved 20% of the simulations (the combined S1 and S2 sets) and pre-Q1 non-lenses as a held-out test set, while the remaining 80% was used for training (including epoch-level validation). For performance evaluation on the Q1 data, we constructed a dedicated test set designed to provide a realistic estimate of real-world performance. The positive class consists of 20% of the Q1 lenses (110 total objects), while the negative class consists of 75% of a randomly selected Q1 non-lens sample (30 000 total objects). The ratio of positives to negatives in the test set is not reflective of the lensing rate in the Universe, and this was taken into account in all reported metrics. We chose to only include randomly selected Q1 non-lenses – and not the ML-selected false positives – in the test set to ensure the negatives are representative of the true distribution of non-lenses in *Euclid* datasets. This established test set is also used to evaluate other *Euclid* lens-finding algorithms (e.g. Vincken et al. in prep.). The additional Q1 non-lenses that were flagged for visual inspection because they received high scores from the ML models, but were ultimately identified as false positives, are incorporated into the training set, where they do not overlap with the randomly selected negatives.

When evaluating on Q1 data, we used the grades assigned by expert visual inspection as the ground truth, using the 497 grade A and grade B Q1 lenses as the positive sample. The Q1 lens search also resulted in a set of 585 grade C lens candidates that exhibit lens-like features, but could not confidently be classified as lenses, which we included in neither the positive nor the negative set for simplicity. Although humans are not perfect at recognising lenses, especially those of lower signal-to-noise ratio (S/N) or smaller Einstein radii (Rojas et al. 2023; Euclid Collaboration: Walmsley et al. 2025), humans remain substantially better at recognising lenses than ML algorithms: in the original Q1 search, citizen scientists outperformed the ML classifiers using expert scores as a ground truth (Euclid Collaboration: Holloway et al. 2025), and expert visual inspection was more reliable than raw ML scores for predicting which systems could be successfully modelled (Euclid Collaboration: Walmsley et al. 2025). This does mean that the definition of ground-truth is different for the simulations versus Q1 lenses: the Q1 lenses are objects that experts recognise as lenses, but have not been confirmed through spectroscopy and cannot necessarily be modelled using simple parametric lens models; whereas the simulations are lens systems generated from known parametric mass models and therefore, by construction, can be modelled under those assumptions. Therefore, performance on simulations versus Q1 lens candidates is not necessarily an apples-to-apples comparison, and this should kept in mind. This distinction implies that the observed performance gap is driven not solely by properties of the training images, but also by the inherent label noise in these observed data.

3.3. Performance metrics

Machine-learning performance can be well understood using the receiver operating characteristic (ROC) curve. The ROC curve is the true positive rate (TPR; the fraction of all lenses classified as lenses, also known as completeness or recall) against the false positive rate (FPR; the fraction of all non-lenses classified as lenses). In terms of the number of positives ($N_{\rm P}$), which is split into true positives ($N_{\rm TP}$) and false positives ($N_{\rm FP}$), and negatives ($N_{\rm N}$), which is split into true negatives ($N_{\rm TN}$) and false negatives

($N_{\rm FN}$), these are defined as

$$\mathrm{TPR} = \frac{N_{\rm TP}}{N_{\rm P}} = \frac{N_{\rm TP}}{N_{\rm TP} + N_{\rm FN}}\,, \tag{1}$$

$$\mathrm{FPR} = \frac{N_{\rm FP}}{N_{\rm N}} = \frac{N_{\rm FP}}{N_{\rm FP} + N_{\rm TN}}\,. \tag{2}$$

An ideal classifier can reach TPR = 1 at FPR = 0, and hence has an area under the ROC curve (AUC) of one. The TPR and the FPR are invariant to class imbalance (the relative proportion of positive and negative samples), since they are the normalised fraction of lenses and non-lenses that fall within a certain threshold range. These can be converted into purity (equivalent to precision) using the negative-to-positive ratio ($N_{\rm N}/N_{\rm P}$) as

$$\mathrm{Purity} = \frac{N_{\rm TP}}{N_{\rm TP} + N_{\rm FP}} = \frac{\mathrm{TPR}}{\mathrm{TPR} + (N_{\rm N}/N_{\rm P})\ \mathrm{FPR}}\,. \tag{3}$$

In Q1, 497 strong lenses were found from an initial sample size of 1 086 554, meaning there were approximately 2200 non-lenses per lens, $N_{\rm N}/N_{\rm P} = 2200$. Understanding the trade-off between purity and completeness allows us to better understand the number of lenses expected to be discovered through visually inspecting the top-scored images according to an ML model applied to real data.

While AUC quantifies performance integrated across the full range of thresholds, the F1 score can be used as a metric that quantifies the balance of purity and completeness at a given threshold, which translates much better into real-world lens-finding returns. It is the harmonic mean of purity and completeness, or equivalently

$$\text{F1 score} = \frac{2\,N_{\rm TP}}{2\,N_{\rm TP} + N_{\rm FP} + N_{\rm FN}}\,. \tag{4}$$

We report the maximum F1 value achieved over all possible decision thresholds. Because it reflects the trade-off between purity and completeness, this optimum typically lies in the threshold regime where both are reasonably high. This is the most relevant regime for strong lens searches, since this is the range in which candidates above this threshold are forwarded for visual inspection.

For each version of the model with different training or testing data, we repeated the full training and testing process ten times with different random initialisations to quantify the variability of the model and the uncertainty in the performance metrics. The statistical uncertainty from bootstrapping within a reserved test set is subdominant compared to the variability from different initialisations (see Euclid Collaboration: Holloway et al. 2025 for more details).

4. Results

4.1. Testing data: Simulations versus real lenses

We first explored the discrepancy in performance between an ML model evaluated on simulations (in-domain data) versus real lenses (out-of-domain data). To do this we trained an ML model on a subset of the pre-Q1 data (using simulated lenses only), evaluated it on a separate pre-Q1 test subset from the same parent dataset, and then evaluated its generalisation performance on the independent Q1 test set consisting of real lenses (see Sect. 3.2).

Figure 2 shows the ROC curve for the same model evaluated on the two different test sets. When tested on the pre-Q1 data, the model can almost perfectly distinguish the positives from the negatives, with an AUC of 0.9991 ± 0.0003. In contrast, when tested on the Q1 data, the model performs significantly worse, with an AUC of 0.94 ± 0.02. This discrepancy is evident in other metrics, shown in Table 2, including the maximum F1 score and purity at set completeness levels. The performance on the in-domain test data is significantly better than the out-of-domain performance by every metric: the same model can recover up to 92% of the simulated lenses with 100% purity (zero false-positives), while in the Q1 data the same model recovers 50% of the lenses with only 24% purity.

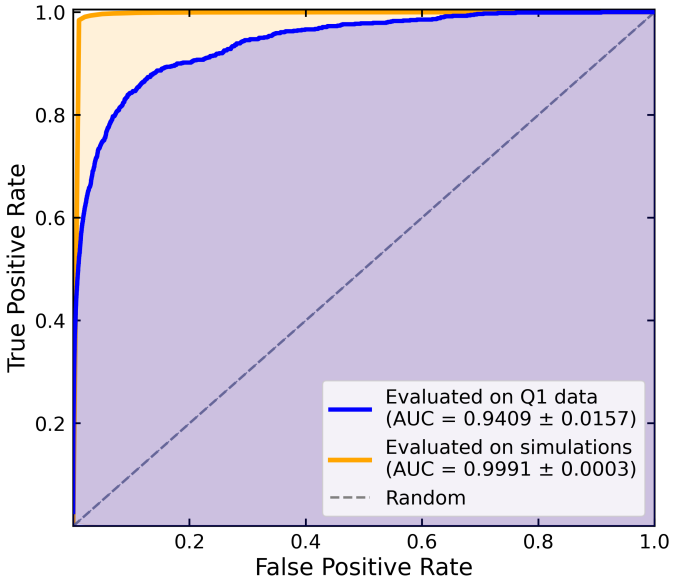

Fig. 2. ROC curve for the fine-tuned Zoobot ML model (trained only on pre-Q1 simulated lenses), evaluated on both pre-Q1 and Q1 data.

Table 2. Performance metrics of the same model evaluated on simulations in comparison to real Q1 data.

	Evaluated on simulations	Evaluated on Q1 data
AUC	0.9991 ± 0.0003	0.9409 ± 0.0157
F1 score	0.9933 ± 0.0014	0.3740 ± 0.0645
Purity at 50% completeness	1.0000 ± 0.0000	0.2361 ± 0.1013
Purity at 90% completeness	1.0000 ± 0.0000	0.0313 ± 0.0047
Purity at 100% completeness	0.9099 ± 0.0511	0.0056 ± 0.0002

This discrepancy between performance on pre-Q1 and Q1 data indicates that the data distributions are fundamentally different, both for the positive and negative classes, and that the model has implicitly overfitted to features that are not robust across domains. This highlights a fundamental challenge: simulations and curated training sets, no matter how carefully designed, cannot fully capture the diversity, complexity, and observational nuances of real survey data. As long as the true data distribution remains only partially known, standard metrics such as AUC on in-domain tests give an overly optimistic view of model reliability. The nature of lens finding creates a circular problem: accurate lens-finding models require a large representative training sample, but obtaining a large representative sample requires accurate lens-finding models. Given how realistic current simulations appear, without understanding exactly how they differ from real lenses, it is difficult to make improvements that ensure better transfer of model performance from simulated to real data. Therefore, the only reliable way to close this domain gap

is to include real lenses, along with representative non-lenses, in the training sample.

Evaluating performance on the Q1 test set may not perfectly reflect lens-finding abilities, since the Q1 lens sample is not 100% complete. Most of the Q1 lenses were found because they were scored highly by at least one of five ML models trained to find lenses in Q1. Therefore, any lenses that are particularly elusive to ML algorithms are likely to have been missed. However, the Q1 lenses from Euclid Collaboration: Walmsley et al. (2025) originate from a search using five different independently developed ML approaches, along with visual inspection of randomly selected objects. Thus, this sample already includes lenses missed by the original Zoobot model and should be relatively balanced: the ML network is not being tested solely on lenses it previously found. Even with a slight bias in the sample, we would expect performance on the Q1 test sample to be much more representative of actual performance than that on the simulations.

4.2. Augmenting pre-Q1 training data with Q1 lenses and non-lenses

We next explored the impact of adding the available Q1 data to training. We evaluated the performance on the reserved test set of 20% of the Q1 lenses and 75% of the randomly selected images classified as non-lenses. We then used the remaining 80% of the Q1 lenses and randomly selected Q1 images (as well as the objects scored as likely to be lenses but classified as non-lenses) for training. We augmented the sample of real lenses by adding four rotations of each lens in order to artificially inflate the number of lenses available for training and to encourage rotational-invariance. However, this still only resulted in 1548 lenses available for training. Given that this sample is still relatively small, we used these Q1 positives and negatives to supplement, rather than replace, the pre-Q1 training data.

Table 3 shows the same performance metrics as in Table 2, but evaluated on the reserved test set of real *Euclid* lenses and non-lenses. In Fig. 3 we use this performance on the reserved test set to extrapolate how many lenses we may find in DR1/EWS as a function of how many images need to be visually inspected. To do this, we first computed the TPR (fraction of lenses recovered) as a function of the FPR from the reserved test set of lenses and non-lenses. We then scaled these rates to the full survey areas (DR1 and EWS) by extrapolating the total numbers of images and lenses expected, based on the ratio measured in Q1. Finally, by converting the FPR into the corresponding number of images inspected, we obtained the expected number of lenses recovered as a function of images to inspect in each survey. In the most relevant range for large-scale discovery efforts (10^5–10^6 inspected images), incorporating Q1 data into the training reduces the number of images that must be inspected by an order of magnitude, while still discovering the same number of lenses. This translates to a significant increase in the fraction of lenses we could expect to discover in DR1 and the EWS using this version of Zoobot alone – 25% in the case of DR1 (from visually inspecting 500 000 images) and 30% in the case of the EWS (from visually inspecting 1 000 000 images). Visual inspection of roughly these numbers of images is planned with the help of citizen science through the Space Warps project.

To understand the trend in improvement, in Fig. 4 we show the impact of adding the non-test set Q1 data incrementally to the training data, starting from a model trained only on pre-Q1 positives and negatives. At each increment, another random fraction of the non-test set Q1 lenses are added to the training data.

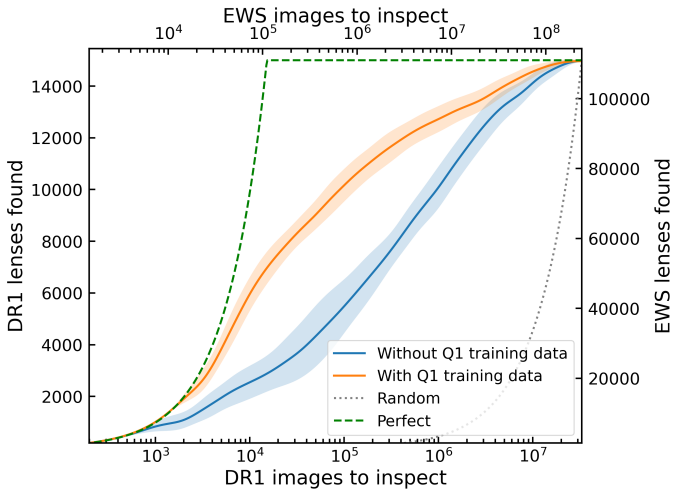

Fig. 3. Projected number of lenses discoverable in DR1 and EWS as a function of the number of images to inspect, for the network trained with and without Q1 data.

The Q1 non-lenses are added by score, so the first 20% added to training data are the top 20% of the Q1 non-lenses that the model initially thought were most likely to be lenses. This reflects how performance scales with visual inspection depth: the further down the candidate list one inspects, the more labelled non-lenses become available to improve the model. We quantified performance using the F1 score and AUC, which show a trend of improvement as more Q1 data are incorporated into the training set. To relate these metrics to practical outcomes, we used the ROC curves to estimate the fraction of lenses that could realistically be recovered in the EWS by visually inspecting the top 1 000 000 candidates, a ballpark number of images that can feasibly be reviewed. There is a general upwards trend, though roughly half the improvement is achieved by adding just the first 20% of the available Q1 data. This suggests that there may be diminishing returns in terms of the performance improvement from adding Q1 data to pre-existing training data. However, given the relatively small size of the Q1 dataset, it is hard to confidently extrapolate how much improvement we can expect using images from future *Euclid* data releases: DR1 will be around 30 times the size of Q1. If performance scales with training data size as a power law, as found to be the case in related tasks (Walmsley et al. 2024), we would expect substantial improvement once we have a DR1-size dataset to train on. Given the improvement already obtained by re-training on just a few hundred real lenses, it is worthwhile to re-train ML models as soon as new data become available, even if the sample is relatively small. As more data are acquired, it is expected that further improvements might be achievable by using the Q1 data to replace, rather than augment, the pre-existing training data. This is addressed further in Sect. 4.4.

4.3. True positives versus false positives

To understand the discrepancy between performance on the pre-Q1 versus Q1 data, it is informative to investigate how much of the improvement comes from a more representative training class of the positives versus negatives. Figure 5 shows the same metrics as in Fig. 4, but plotted as a function of the fraction of the Q1 lenses (rather than all the Q1 data) added to the training data. This is plotted for two scenarios, one in which all the Q1 non-lenses are also used in the training data, and one in which none

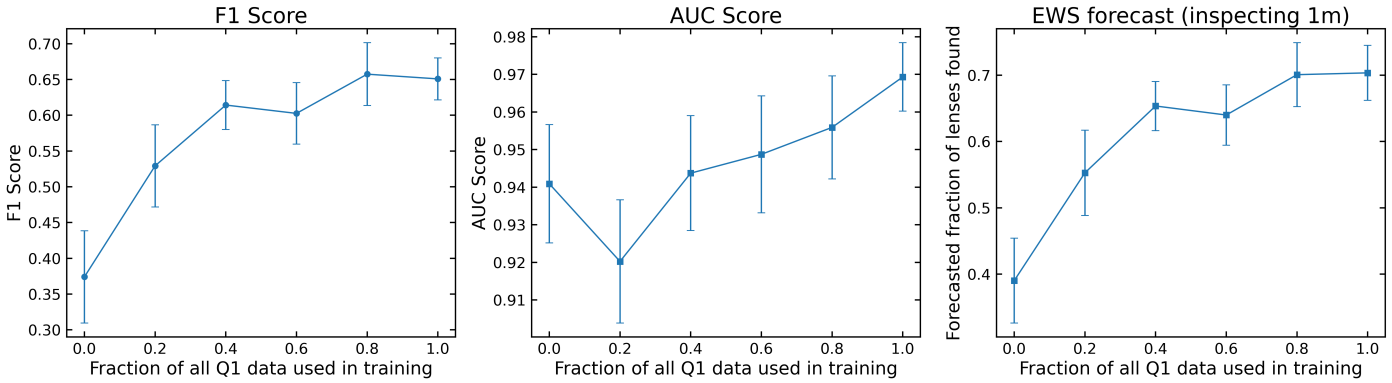

Fig. 4. Impact of augmenting the training data with available Q1 lenses and non-lenses, in terms of performance across a range of metrics. These include F1 score and AUC, as well as the projected fraction of lenses discoverable in a dataset the size of the EWS, assuming one million images can be visually inspected.

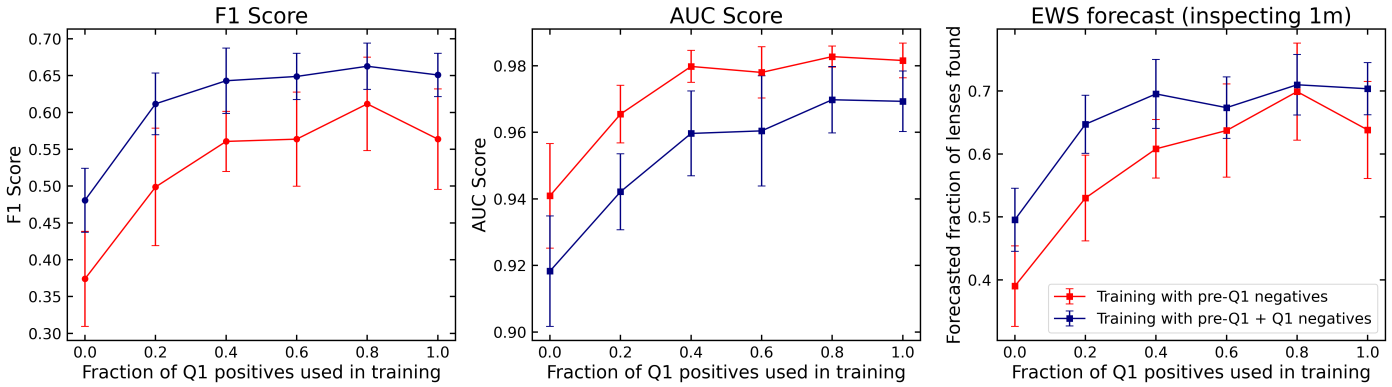

Fig. 5. Same as in Fig. 4, but isolating the impact of adding Q1 lenses and non-lenses separately. The x-axis corresponds to the addition of Q1 lenses to the pre-existing training data. The two lines show the change in performance with and without the addition of the Q1 non-lenses.

Table 3. Performance metrics evaluated on real data for two models with different training data.

	Trained on pre-Q1 data	Train on pre-Q1 and Q1 data
AUC	0.941 ± 0.016	0.969 ± 0.009
F1 score	0.374 ± 0.065	0.651 ± 0.029
Purity at 50% completeness	0.236 ± 0.101	0.838 ± 0.068
Purity at 90% completeness	0.031 ± 0.005	0.048 ± 0.036
Purity at 100% completeness	0.0056 ± 0.0002	0.0062 ± 0.0013

Notes. These metrics are the same as in Table 2, but evaluated on real data rather than simulations. The model trained on pre-Q1 data alone corresponds to the blue curve in Fig. 3, and the combination of pre-Q1 and Q1 data corresponds to the orange curve.

of the Q1 non-lenses are added to the training data. This allows us to understand if the problem lies with the simulations not being representative of real lenses, or with the Q1 data containing peculiar non-lens objects that the model has not been trained to recognise.

It can be seen that the majority of the improvement in these metrics is driven by adding the Q1 lenses to the training data: adding the Q1 non-lenses to the training data results in consistently better performance in terms of the number of expected lens discoveries by roughly 5–10%, but including the Q1 lenses in training increases the fraction of discoverable lenses by roughly 20%. This suggests that the problem primarily lies with the model misclassifying true lenses as non-lenses. Figure 6 displays examples of lenses that were originally misclassified as negatives by the pre-Q1 version of the model but successfully identified by the model after re-training with the Q1 data. These systems are predominantly grade B candidates, exhibiting fainter arcs and less complete Einstein rings, indicating that such features were underrepresented in the simulations. While adding a more diverse set of negatives to the training data can improve purity, obtaining a suitably complete sample of lenses can only be achieved through a robust understanding of the diversity of lens systems. The pre-Q1 negative set already consisted entirely of pure *Euclid* non-lenses, so the domain gap is less of an applicable problem in this case. This explains why Q1 negatives offer less new information than the positives: the Q1 negative set helps refine the training data and facilitates its understanding of atypical contaminants it might have previously struggled with, whereas the Q1 positives directly expand its knowledge of what genuine lenses look like.

It is interesting to note that, according to the AUC, using Q1 non-lenses in training performs worse than not doing so, despite other metrics showing the opposite. Additionally, the AUC score is not always well correlated with the number of lenses expected from applying the ML model to real data. This highlights a flaw with global metrics such as the AUC for highly imbalanced classification problems, where only the top-ranked candidates are relevant. As long as a model performs decently well, a higher AUC does not necessarily translate to increased lens returns (see Appendix A for a more detailed discussion). In contrast, trends in F1 score are much better correlated with the number of lenses expected to be discoverable in practice.

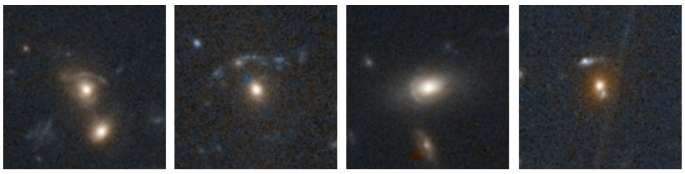

Fig. 6. Q1 lenses misclassified as non-lenses by the pre-Q1 model but correctly identified after training with Q1 data.

4.4. Training on Q1 data alone

Given the discrepancy between performance on simulations versus real data, and that lens-finding neural networks have been successfully trained on lens samples only a few hundred larger than Q1, it is interesting to investigate whether a network can be trained on Q1 data alone. To test this, we independently varied the composition of both the positive and negative training data: each can consist of pre-Q1 data only (simulations and pre-Q1 non-lenses), Q1 data only, or both combined. Figure 7 shows how the different combinations of training data impact F1 score, as evaluated on the reserved test set of Q1 lenses and random non-lenses.

Training on the Q1 data alone produces significantly better performance (F1 score = 0.512 ± 0.069) than training on the pre-Q1 data alone (F1 score = 0.374 ± 0.064), but does not perform as well as the model trained on the combination of the data (F1 score = 0.651 ± 0.029). Since the F1 score is evaluated only on Q1 lenses and non-lenses, the model trained solely on this subset should perform best, provided that no other limiting factors constrain training, since it is tested on in-domain data. The superior performance of the model trained on both Q1 and pre-Q1 data suggests that limited training-set size constrains performance when training on Q1 data alone. Once the number of Q1 lenses and non-lenses matches that of the pre-Q1 training size, the model's performance should surpass that of the current combined-data model, since it would then benefit from both sufficient training size and in-domain data. Alternative training strategies, such as the real-data-focused human-in-the-loop pipeline presented in Xu et al. (in prep.), have been proposed as promising approaches for identifying lenses without a reliance on simulations. This further reinforces our conclusion that the inclusion of real observational data is critical for maximising the scientific yield of lens-finding campaigns.

In terms of the positive sample, we find that training on Q1 lenses alone results in a better F1 score compared to training on simulations alone, despite their relatively small sample size. The fact that the model was fine-tuned from a pre-trained state, rather than from scratch, likely helps mitigate the usual limitations of smaller training datasets. The simulations contain complementary information though, since we can see that adding the simulations to the Q1 lenses results in further improvement. Training on the combination of Q1 lenses and simulations has an added benefit over training on real lenses alone: we can be more confident that the simulations cover an appropriate range of lens parameters, such as Einstein radii and S/N. Since many of the Q1 lenses were found by ML models, training on these lenses alone may mean that any selection biases that occur in the ML models will propagate and amplify in a self-reinforcing way. This is of particular concern here, since Q1 is just the start of the *Euclid* discoveries: these Q1 lenses will help find lenses in DR1, which in turn will help find lenses in the next data releases. The extent of this potential bias is difficult to quantify due to the limited number of Q1 lenses, but could be investigated further in the future. However, this effect is likely to be mitigated by using multiple ML models, since different approaches generally have different strengths and weaknesses in terms of the types of lenses they find (Euclid Collaboration: Holloway et al. 2025; Gonzalez et al. 2025; Nagam et al. 2025). Additionally, visually inspecting randomly selected *Euclid* objects will allow us to calibrate the ML models and enable the discovery of lenses without this bias. There is also the argument that using Q1 lenses in training means that the model will be trained to recognise more atypical lens systems that might not have been well accounted for in the simulations. For example, the simulated lenses consisted primarily of lensing by LRGs, while the Q1 sample contained 30–40 late-type disk lenses.

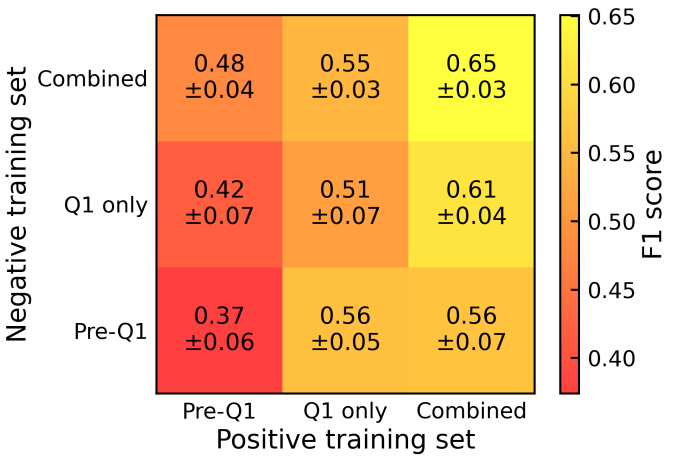

Fig. 7. F1 scores achieved by training with different training sets, as evaluated on the reserved Q1 test set.

When considering the negative data, adding the Q1 non-lenses to the training set also improves performance: training on the Q1 only negatives generally results in better performance than training on the pre-Q1 negatives alone, and training on the combination of the two typically results in the best performance. We note that this trend does not hold as well when training on the Q1 positives alone, which could likely be related to the imbalance between the positives and negatives when training in this scenario. However, when training on a sufficiently large positive set (e.g. the combined real and simulated lenses), increasing the negative sample size from $\sim 10^3$ (pre-Q1 negatives) to $\sim 10^5$ (Q1 negatives) still improves performance, indicating that the model is fairly robust to class imbalance. As highlighted in previous studies (e.g. Cañameras et al. 2024), it is more effective to construct a negative training set that inflates the fraction of common contaminants than to use one that simply mirrors the true distribution of non-lenses in the data. The negative sample obtained from the Q1 lens search contains a large number of typical lens contaminants by construction, so it makes sense that using these for training aids in the model's ability to discern lenses from non-lenses.

An alternative approach to mitigating the domain gap by incorporating real data is the use of unsupervised domain adaptation techniques, such as DANNs. These architectures include a domain classifier (in addition to the regular label predictor) attached to the feature-extraction layers, which attempts to distinguish between domains (e.g. simulated and real images). Through a gradient reversal layer, the loss from this domain classifier is used to encourage the feature extraction layers to only learn representations that are common across the domains, thereby promoting domain-invariance. This approach has

been successfully applied to other astronomical problems (e.g. Belfiore et al. 2025) and could be explored in future work to improve the robustness of strong lens detection models.

4.5. Limitations from test-set size

Throughout this work, performance gains from training on Q1 lenses have been evaluated using a reserved test set consisting of 75% of the randomly selected Q1 non-lenses (30 000 objects) and 20% of the Q1 lenses (110 objects), leaving sufficient lenses available for training. The set of 110 lenses is a sample size small enough to be influenced by statistical fluctuations. Increasing the proportion of lenses in the test set would reduce the number available for training and thus limit potential performance improvements. Therefore, to assess variability due to the limited test size, we partitioned the Q1 lenses into five disjoint 20% subsets (A–E) and alternated the subset used for testing, with the remaining 80% used for training.

Figure 8 shows the variation in performance when varying the Q1 lenses used for training versus testing, for the version trained on the combination of pre-Q1 and Q1 lenses and non-lenses (the best-performing version). Test set A corresponds to the test set that has been used throughout this paper, and its performance is represented by the blue line in Fig. 3. The variation in performance evaluated on different Q1 lenses is larger than the variation from random initialisations of the model, demonstrating that the limited size of the lens test set is a dominant source of uncertainty in the reported performance. Notably, the performance on test set A is lower than average, suggesting that the performance may have been under-reported, since this subset of the Q1 lenses is harder for the ML model to recognise compared to the broader Q1 lens population.

While the Q1 lens sample size is large enough to understand how lens-finding performance can scale in the future, its limited size means that performance estimates remain sensitive to statistical fluctuations and to the specific lenses included in the test set. The much larger sample of 10 000–15 000 lenses expected from *Euclid* DR1 will overcome these constraints, reducing uncertainties and allowing us to more confidently extrapolate performance.

4.6. Exploring the embedding space

While machine-learning models are famously difficult to interpret, exploring their embedding space can help us understand their inner workings. This can be achieved using the Uniform Manifold Approximation and Projection (UMAP; McInnes et al. 2018) algorithm, which projects high-dimensional parameter space into lower dimensions for visualisation, while preserving both local and global structure. By applying UMAP to the embeddings extracted from the layer immediately preceding the classification head, we can visualise how the model organises similar inputs and separates different classes.

Figure 9 shows the UMAP projection of the pre-Q1 fine-tuned Zoobot model, mapping simulated images, Q1 lenses, and Q1 non-lenses into a shared embedding space. The plot reveals that Q1 lenses occupy an intermediate position between the simulated lenses and non-lenses. At one extreme (bottom left) lie images with high S/N and complete Einstein rings. A large fraction of simulations fall in this region, whereas fewer real lenses exhibit the same properties. Most Q1 lenses contain less complete rings and often fainter arcs, while the simulations exhibit these characteristics less frequently. The idealised nature of the simulations means that the real lenses that lie closest to the simulated lenses are the grade A lenses, while those nearer the non-lenses tend to be grade B with fainter arcs and less complete rings.

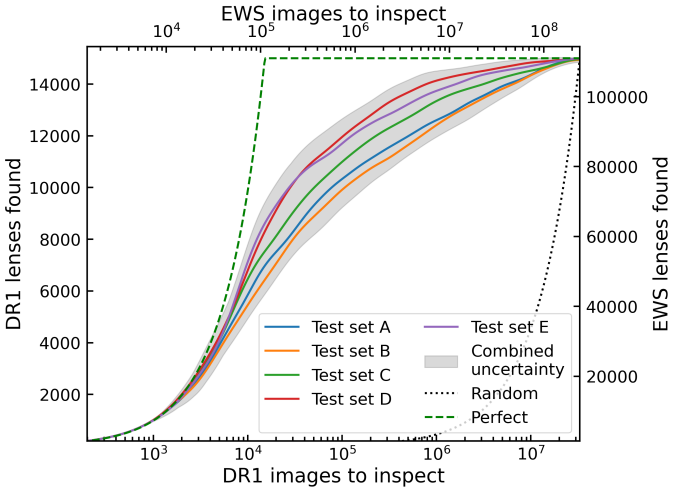

Fig. 8. Projected number of lenses discoverable in DR1 and the EWS as a function of the number of images to inspect (same as Fig. 3), as evaluated on different 20% subsets of Q1 lenses.

The Q1 non-lenses are more tightly clustered than the simulated and real lenses, with objects such as extended spirals occupying distinct regions. The greater scatter observed in the positive class relative to the negative class helps understand why incorporating real positives provides a larger performance improvement than adding real negatives: the model already has a more robust understanding of the non-lenses. Interestingly, a compact subset of the Q1 lenses forms an outlier cluster, corresponding to edge-on disk lenses. Although these on average lie closer to the negatives than the positives, the fact that the model can efficiently distinguish them suggests that a multi-class output could be beneficial and warrants future exploration. Such outliers also point to regions of the lens parameter space where simulations are lacking, motivating the inclusion of these lens types in future training sets.

The UMAP suggests that part of the difference in performance between simulated and real lenses is due to the overly idealised nature of the simulations. Training on exaggerated simulations can have benefits: such examples present clear lens features, encouraging the model to learn unambiguous characteristics and reduces the risk of confusion. Training on less clear lenses might incentivise the model to pick out non-lensing features. For example, Euclid Collaboration: Leuzzi et al. (2024) demonstrated that including a larger fraction of faint lenses in the training sample can increase misclassification. Additionally, by focusing on unambiguous lenses, the model naturally prioritises candidates that are both more likely to be genuine lenses and are often more scientifically valuable. However, this approach has limitations: the model becomes less robust to fainter or atypical lenses, meaning that achieving a representative and complete sample of the lens population is challenging. Incorporating real observations alongside simulations can therefore help capture the full diversity of lens morphologies and ensure robust performance across the range of detectable lenses.

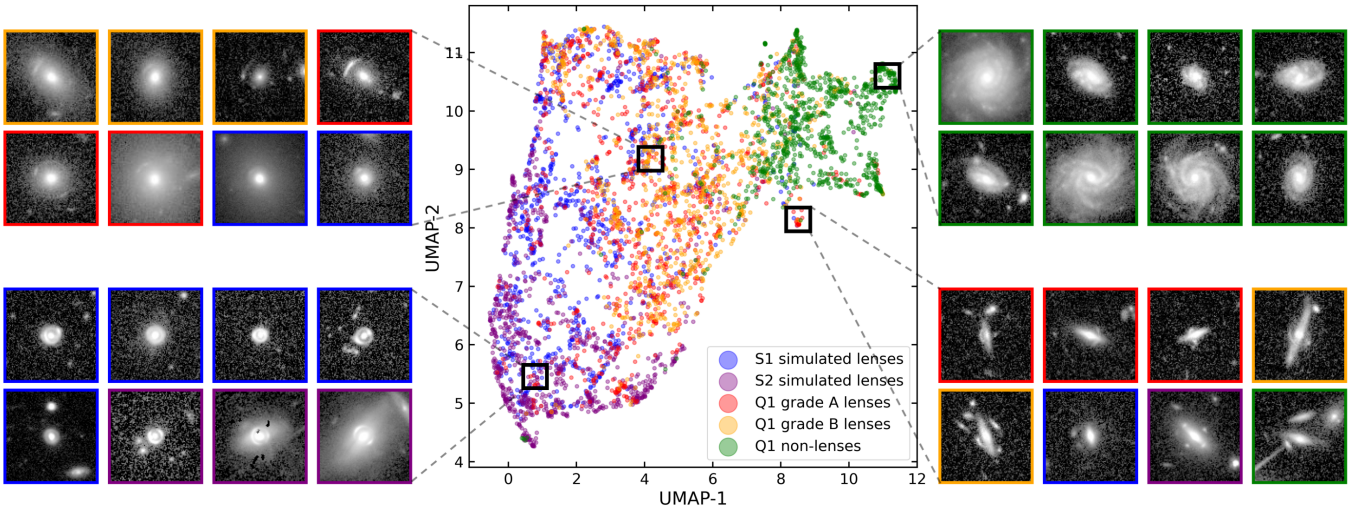

Fig. 9. UMAP projection of simulated lenses (S1 and S2), Q1 lenses (grade A and grade B), and Q1 non-lenses in the embedding space of the fine-tuned Zoobot model. Four hundred randomly selected points are plotted from each class for clarity. Example images shown to the left and right correspond to the highlighted regions of the projection, with coloured borders indicating the associated object class.

5. Conclusions

The *Euclid* mission is set to transform our understanding of cosmology and astrophysics by delivering an unprecedented dataset of approximately 100 000 galaxy-scale strong gravitational lenses, but their discovery within *Euclid* 's vast imaging presents a significant challenge. Human visual inspection is infeasible at this scale, making ML models essential for detecting the majority of these lens systems. However, training robust lens-finding models is challenging. The number of known strong lenses available for supervised learning is still limited, restricting the feasibility of purely real-data-driven approaches. Consequently, current methods rely heavily on large, diverse sets of simulated lenses, which provide statistical power but inevitably fail to capture the full complexity of real astronomical observations. It is well established that models trained exclusively on one data domain do not always transfer well to another domain, leading to a simulation-to-reality performance gap.

In this work, we addressed this gap directly by using the 497 strong lenses discovered from the initial search of the *Euclid* Q1 data release (Euclid Collaboration: Walmsley et al. 2025). These lenses provide the first statistically meaningful sample of real *Euclid* lenses, enabling for the first time a direct and quantitative comparison of how models trained on simulations perform on actual *Euclid* data. Furthermore, we investigate strategies to close this gap by combining simulated and real data to improve both completeness and purity in lens discovery. Our key findings are as follows.

1. We confirm a significant discrepancy between model performance on simulated versus real data. A model trained exclusively on simulations that achieves over 90% completeness with nearly 100% purity on a simulated test set only recovers 50% of real Q1 lenses at a purity of 24%. This underscores the limitations of relying solely on simulation-based metrics as indicators of real-world performance.
2. Augmenting the simulation-based training set with real Q1 lenses and non-lenses provides a substantial performance boost. This hybrid training approach can decrease the number of images requiring inspection in future *Euclid* data releases by a factor of ten, increasing the projected number of discoverable lenses by 25–30% for upcoming searches in *Euclid*'s Data Release 1 and the full EWS.
3. The inclusion of real lenses is the primary driver of this improvement, teaching the model the complex, high-fidelity features of observed systems that simulations fail to capture perfectly. Including real non-lenses offers a secondary benefit by helping the model reject common false positives, thereby increasing the purity of the final candidate list. The limited number of real lenses available for test sets is a major source of uncertainty in the reported performance of the ML models, but we expect this to be overcome with the growing number of lenses expected from upcoming *Euclid* data releases. There will likely be further increases in completeness and purity as the next visual inspection campaigns produce more lenses for training.
4. The optimal training strategy is a hybrid approach that combines the statistical power and diversity of large simulation sets with the fidelity of a smaller but growing sample of real lenses. While training on Q1 data alone outperforms training on simulations alone, the limited size of the current real-lens sample means that the combined dataset yields the best results. This may change in the future once more *Euclid* data have been acquired.

While the hybrid training approach results in immediate improvements, we must be wary of the risk of a self-reinforcing feedback loop. We are currently at the start of *Euclid* lens finding, with the 497 Q1 candidates representing the seed population for future training sets. As we scale from these hundreds of candidates to the number of lenses expected in the full survey, we must ensure that we understand how our training affects the selection function: any biases in the population of lenses discovered can directly impact the scientific inferences we make from such a sample, particularly if these biases are not quantified and accounted for.

This risk arises from several intersecting factors. Supervised ML models are optimised to recognise objects similar to those they are trained on, which in turn are bounded by our knowledge of what we are looking for. While simulations provide a necessary starting point, they rarely capture the full diversity of real

data. For instance, our pre-Q1 simulations did not explicitly prioritise edge-on disk lenses, yet these emerged as a notable population in the real Q1 data. If we strictly optimise our models to find only what we are expecting to find, we risk missing the unexpected discoveries that are an exciting element of the mission. Beyond the training data, the algorithms themselves possess intrinsic selectivities. For example, the design of CNNs means that they are more adept at learning local spatial correlations and hence may better learn to detect continuous features, such as lensed arcs, compared to Einstein crosses. Another source of selection bias is introduced by training on images recognised as lenses through human visual inspection. If experts consistently overlook certain types of lenses, the ML models will learn to treat these potentially legitimate lenses as negatives.

Ideally, we would detect every lens in the survey with 100% completeness and purity, resulting in a catalogue perfectly representative of the strong lensing distribution in the Universe. Since this is infeasible, understanding the completeness of our detection rates as a function of the lens parameter space is important. Several strong lensing science cases can be enabled by obtaining a subset of the lensing population with 100% completeness, at the expense of a smaller sample size (Sonnenfeld 2022; Zhou et al. 2024). To ensure a less biased catalogue and robustly characterise the selection function, there are a few potential strategies. First, we must continue the visual inspection of random images to identify lens candidates without the ML bias. While this method is still subject to human bias, it is free from the specific morphological biases of the ML models. Future spectroscopic surveys, such as 4MOST, will be able to confirm many of these lens candidates, thereby providing a more reliable and consistent definition of what constitutes a strong gravitational lens. Second, we should employ multiple independent ML approaches, as different approaches will inherently prioritise different regions of the parameter space. For example, the pipeline developed by Xu et al. (in prep.) minimises reliance on simulations in favour of iterative training on human-identified real lenses. In their application to Q1, this approach uncovered 91 previously missed grade B candidates and four new grade A candidates. This prioritising of grade B candidates illustrates the specific utility of their training strategy: while it carries a stronger imprint of human selection bias, it is effective at recovering more ambiguous candidates that constitute a large range of the lens candidate population and are more likely missed by models trained on idealised simulations. Finally, we must quantify the domain gap in the selection function. While we can easily measure selection functions en masse using simulations, we must determine if these measurements are transferable to real data. A first step is to explore whether the recoverability trends observed in simulations (e.g. improved detection at high S/N) translate equivalently to real lenses. Combining these diverse discovery channels and using simulated lenses to map their selection functions will ensure the final *Euclid* lens catalogue is both vast and scientifically valuable.

With these selection effects quantified, we can confidently scale up the search by iteratively updating our models as new data releases arrive. This approach optimises the use of visual inspection time, balancing the speed of ML with the reliability of human visual inspection, and offers the best path to build the largest sample of strong gravitational lenses to date.

6. Data availability

This paper makes use of the Euclid Quick Release 1 data (Euclid Quick Release Q1 2025), covered by Euclid Collaboration: Aussel et al. (2025). The data used for training are available on Zenodo at https://doi.org/10.5281/zenodo.15003116.

Acknowledgements. NEPL is supported through a graduate studentship from the UKRI STFC and the University of Portsmouth. Numerical computations were carried out on the SCIAMA High Performance Compute (HPC) cluster, which is supported by the ICG and the University of Portsmouth. 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. 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 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 Netherlandse Onderzoekschool Voor Astronomie, the Norwegian Space Agency, the Research Council of Finland, the Romanian Space Agency, the State Secretariat for Education, Research, and Innovation (SERI) at the Swiss Space Office (SSO), and the United Kingdom Space Agency. A complete and detailed list is available on the *Euclid* web site (www.euclid-ec.org/consortium/community/). 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.

References

Acevedo Barroso, J. A., Clément, B., Courbin, F., et al. 2025a, A&A, submitted, arXiv:2503.10610
Acevedo Barroso, J. A., O'Riordan, C. M., Clément, B., et al. 2025b, A&A, 697, A14
Adamo, A., Bradley, L. D., Vanzella, E., et al. 2024, Nature, 632, 513
Auger, M. W., Treu, T., Bolton, A. S., et al. 2010, ApJ, 724, 511
Belfiore, F., Ginolfi, M., Blanc, G., et al. 2025, A&A, 694, A212
Bradley, L. D., Adamo, A., Vanzella, E., et al. 2025, ApJ, 991, 32
Cañameras, R., Schuldt, S., Shu, Y., et al. 2024, A&A, 692, A72
Cañameras, R., Schuldt, S., Shu, Y., et al. 2021, A&A, 653, L6
Cañameras, R., Schuldt, S., Suyu, S. H., et al. 2020, A&A, 644, A163
Caminha, G. B., Grillo, C., Rosati, P., et al. 2016, A&A, 587, A80
Caminha, G. B., Suyu, S. H., Grillo, C., & Rosati, P. 2022, A&A, 657, A83
Ćiprijanović, A., Kafkes, D., Jenkins, S., et al. 2020, in 34th Conference on Neural Information Processing Systems
Collett, T. E. 2015, ApJ, 811, 20
Collett, T. E., Sonnenfeld, A., Frohmaier, C., et al. 2023, The Messenger, 190, 49
Dutra, I., Natarajan, P., & Gilman, D. 2025, ApJ, 978, 38
Ertl, S., Schuldt, S., Suyu, S. H., et al. 2024, A&A, 685, A15
Euclid Collaboration: Aussel, H., Tereno, I., Schirmer, M., et al. 2025, A&A, submitted (Euclid Q1 SI), arXiv:2503.15302
Euclid Collaboration: Bergamini, P., Meneghetti, M., Acebron, A., et al. 2025, A&A, in press (Euclid Q1 SI), https://doi.org/10.1051/0004-6361/202554577, arXiv:2503.15330
Euclid Collaboration: Castander, F., Fosalba, P., Stadel, J., et al. 2025, A&A, 697, A5
Euclid Collaboration: Cropper, M., Al-Bahlawan, A., Amiaux, J., et al. 2025, A&A, 697, A2
Euclid Collaboration: Holloway, P., Verma, A., Walmsley, M., et al. 2025, A&A, accepted (Euclid Q1 SI), arXiv:2503.15328
Euclid Collaboration: Leuzzi, L., Meneghetti, M., Angora, G., et al. 2024, A&A, 681, A68
Euclid Collaboration: Li, T., Collett, T. E., Walmsley, M., et al. 2025, A&A, in press (Euclid Q1 SI), https://doi.org/10.1051/0004-6361/202554543, arXiv:2503.15327
Euclid Collaboration: Lines, N. E. P., Collett, T. E., Walmsley, M., et al. 2025, A&A, in press (Euclid Q1 SI), https://doi.org/10.1051/0004-6361/202554542, arXiv:2503.15326
Euclid Collaboration: Mellier, Y., Abdurro'uf, Acevedo Barroso, J., et al. 2025, A&A, 697, A1
Euclid Collaboration: Rojas, K., Collett, T. E., Acevedo Barroso, J. A., et al. 2025, A&A, in press (Euclid Q1 SI), https://doi.org/10.1051/0004-6361/202554605, arXiv:2503.15325

Euclid Collaboration: Walmsley, M., Holloway, P., Lines, N. E. P., et al. 2025, A&A, accepted (Euclid Q1 SI), arXiv:2503.15324
Euclid Quick Release Q1. 2025, https://doi.org/10.57780/esa-2853f3b
Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H. R. 2021, in Advances in Data Science and Information Engineering, ed. R. Stahlbock, G. M. Weiss, M. Abou-Nasr, C.-Y. Yang, H. R. Arabnia, & L. Deligiannidis (Cham: Springer International Publishing), 877–894
Fujimoto, S., Ouchi, M., Kohno, K., et al. 2025, Nature Astronomy, 9, 1553
Ganin, Y., Ustinova, E., Ajakan, H., et al. 2016, J. Mach. Learn. Res., 17, 2096–2030
Gavazzi, R., Treu, T., Rhodes, J. D., et al. 2007, ApJ, 667, 176
Gonzalez, J., Collett, T., Rojas, K., et al. 2025, ApJ, submitted, arXiv:2510.23782
González, J., Holloway, P., Collett, T., et al. 2025, ApJ, submitted, arXiv:2501.15679
Grespan, M., Thuruthipilly, H., Pollo, A., et al. 2024, A&A, 688, A34
Grillo, C., Pagano, L., Rosati, P., & Suyu, S. H. 2024, A&A, 684, L23
Hendrycks, D., Lee, K., & Mazeika, M. 2019, Proceedings of the International Conference on Machine Learning
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
Jacobs, C., Collett, T., Glazebrook, K., et al. 2019a, ApJS, 243, 17
Jacobs, C., Collett, T., Glazebrook, K., et al. 2019b, MNRAS, 484, 5330
Jacobs, C., Glazebrook, K., Collett, T., More, A., & McCarthy, C. 2017, MNRAS, 471, 167
Jaelani, A. T., More, A., Wong, K. C., et al. 2024, MNRAS, 535, 1625
Jullo, E., Natarajan, P., Kneib, J.-P., et al. 2010, Science, 329, 924
Kelly, P. L., Rodney, S., Treu, T., et al. 2023, Science, 380, eabh1322
Koekemoer, A. M., Aussel, H., Calzetti, D., et al. 2007, ApJS, 172, 196
Leauthaud, A., Massey, R., Kneib, J.-P., et al. 2007, ApJS, 172, 219
Li, R., Napolitano, N. R., Spiniello, C., et al. 2021, ApJ, 923, 16
Li, R., Napolitano, N. R., Tortora, C., et al. 2020, ApJ, 899, 30
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
McInnes, L., Healy, J., Saul, N., & Grossberger, L. 2018, The Journal of Open Source Software, 3, 861
Melo-Carneiro, C. R., Collett, T. E., Oldham, L. J., et al. 2025, MNRAS, 541, 2853
Meneghetti, M., Cui, W., Rasia, E., et al. 2023, A&A, 678, L2
Meneghetti, M., Davoli, G., Bergamini, P., et al. 2020, Science, 369, 1347
Meneghetti, M., Ragagnin, A., Borgani, S., et al. 2022, A&A, 668, A188
Metcalf, R. B., Meneghetti, M., Avestruz, C., et al. 2019, A&A, 625, A119
Meštrić, U., Vanzella, E., Zanella, A., et al. 2022, MNRAS, 516, 3532
Moresco, M., Amati, L., Amendola, L., et al. 2022, Living Reviews in Relativity, 25, 6
Nagam, B. C., Acevedo Barroso, J. A., Wilde, J., et al. 2025, A&A, in press, https://doi.org/10.1051/0004-6361/202554132, arXiv:2502.09802
Nagam, B. C., Koopmans, L. V. E., Valentijn, E. A., et al. 2023, MNRAS, 523, 4188
Nagam, B. C., Koopmans, L. V. E., Valentijn, E. A., et al. 2024, MNRAS, 533, 1426
Natarajan, P., Chadayammuri, U., Jauzac, M., et al. 2017, MNRAS, 468, 1962
Nightingale, J. W., Smith, R. J., He, Q., et al. 2023, MNRAS, 521, 3298
Niu, S., Liu, Y., Wang, J., & Song, H. 2020, IEEE Transactions on Artificial Intelligence, 1, 151
O'Riordan, C. M., Despali, G., Vegetti, S., Lovell, M. R., & Moliné, Á. 2023, MNRAS, 521, 2342
Pascale, M., Frye, B. L., Pierel, J. D. R., et al. 2025, ApJ, 979, 13
Pearce-Casey, R., Nagam, B. C., Wilde, J., et al. 2024, A&A, accepted, arXiv:2411.16808
Petrillo, C. E., Tortora, C., Chatterjee, S., et al. 2017, MNRAS, 472, 1129
Petrillo, C. E., Tortora, C., Vernardos, G., et al. 2019, MNRAS, 484, 3879
Pourrahmani, M., Nayyeri, H., & Cooray, A. 2018, ApJ, 856, 68
Rojas, K., Collett, T. E., Ballard, D., et al. 2023, MNRAS, 523, 4413
Rojas, K., Savary, E., Clément, B., et al. 2022, A&A, 668, A73
Savary, E., Rojas, K., Maus, M., et al. 2022, A&A, 666, A1
Schuldt, S., Cañameras, R., Andika, I. T., et al. 2025a, A&A, 693, A291
Schuldt, S., Cañameras, R., Shu, Y., et al. 2025b, A&A, 699, A350
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Shu, Y., Cañameras, R., Schuldt, S., et al. 2022, A&A, 662, A4
Sonnenfeld, A. 2022, A&A, 659, A132
Sonnenfeld, A., Chan, J. H. H., Shu, Y., et al. 2018, PASJ, 70, S29
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
Sun, C., Shrivastava, A., Singh, S., & Gupta, A. 2017, in 2017 IEEE International Conference on Computer Vision (ICCV), 843–852
Suyu, S. H., Acebron, A., Grillo, C., et al. 2025, A&A, submitted, arXiv:2509.12319
Tdcosmo Collaboration, Birrer, S., Buckley-Geer, E. J., et al. 2025, A&A, 704, A63
Teimoorinia, H., Toyonaga, R. D., Fabbro, S., & Bottrell, C. 2020, PASP, 132, 044501
Vanzella, E., Calura, F., Meneghetti, M., et al. 2017, MNRAS, 467, 4304
Vanzella, E., Loiacono, F., Bergamini, P., et al. 2023, A&A, 678, A173
Vanzella, E., Meneghetti, M., Pastorello, A., et al. 2020, MNRAS, 499, L67
Vegetti, S., Koopmans, L. V. E., Bolton, A., Treu, T., & Gavazzi, R. 2010, MNRAS, 408, 1969
Vegetti, S., Lagattuta, D. J., McKean, J. P., et al. 2012, Nature, 481, 341
Walmsley, M., Allen, C., Aussel, B., et al. 2023, Journal of Open Source Software, 8, 5312
Walmsley, M., Bowles, M., Scaife, A. M. M., et al. 2024, arXiv e-prints, arXiv:2404.02973
Welch, B., Coe, D., Diego, J. M., et al. 2022, Nature, 603, 815
Wong, K. C., Chan, J. H. H., Chao, D. C. Y., et al. 2022, PASJ, 74, 1209
Zhang, L. & Gao, X. 2024, IEEE Transactions on Neural Networks and Learning Systems, 35, 23
Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. 2023, IEEE Transactions on Pattern Analysis and Machine Intelligence, 45, 4396
Zhou, Q., Sonnenfeld, A., & Hoekstra, H. 2024, A&A, 690, A390

[1] Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 3FX, UK
[2] University of Applied Sciences and Arts of Northwestern Switzerland, School of Computer Science, 5210 Windisch, Switzerland
[3] Dipartimento di Fisica "Aldo Pontremoli", Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy
[4] INAF-IASF Milano, Via Alfonso Corti 12, 20133 Milano, Italy
[5] Dipartimento di Fisica e Astronomia "Augusto Righi" - Alma Mater Studiorum Università di Bologna, via Piero Gobetti 93/2, 40129 Bologna, Italy
[6] INAF-Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, 40129 Bologna, Italy
[7] Department of Physics, Oxford University, Keble Road, Oxford OX1 3RH, UK
[8] INFN-Sezione di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
[9] Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, 08028 Barcelona, Spain
[10] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig de Lluís Companys 23, 08010 Barcelona, Spain
[11] Institut de Ciencies de l'Espai (IEEC-CSIC), Campus UAB, Carrer de Can Magrans, s/n Cerdanyola del Vallés, 08193 Barcelona, Spain
[12] Aix-Marseille Université, CNRS, CNES, LAM, Marseille, France
[13] Institut d'Astrophysique de Paris, UMR 7095, CNRS, and Sorbonne Université, 98 bis boulevard Arago, 75014 Paris, France
[14] INAF-Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131 Napoli, Italy
[15] Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland
[16] SCITAS, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
[17] Université Paris-Saclay, CNRS, Institut d'astrophysique spatiale, 91405, Orsay, France
[18] ESAC/ESA, Camino Bajo del Castillo, s/n., Urb. Villafranca del Castillo, 28692 Villanueva de la Cañada, Madrid, Spain
[19] Institut für Theoretische Physik, University of Heidelberg, Philosophenweg 16, 69120 Heidelberg, Germany
[20] INAF-Osservatorio Astronomico di Brera, Via Brera 28, 20122 Milano, Italy
[21] IFPU, Institute for Fundamental Physics of the Universe, via Beirut 2, 34151 Trieste, Italy
[22] INAF-Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, 34143 Trieste, Italy
[23] INFN, Sezione di Trieste, Via Valerio 2, 34127 Trieste TS, Italy
[24] SISSA, International School for Advanced Studies, Via Bonomea 265, 34136 Trieste TS, Italy
[25] Dipartimento di Fisica e Astronomia, Università di Bologna, Via Gobetti 93/2, 40129 Bologna, Italy

[26] INAF-Osservatorio Astronomico di Padova, Via dell'Osservatorio 5, 35122 Padova, Italy
[27] Dipartimento di Fisica, Università di Genova, Via Dodecaneso 33, 16146, Genova, Italy
[28] INFN-Sezione di Genova, Via Dodecaneso 33, 16146, Genova, Italy
[29] Department of Physics "E. Pancini", University Federico II, Via Cinthia 6, 80126, Napoli, Italy
[30] Dipartimento di Fisica, Università degli Studi di Torino, Via P. Giuria 1, 10125 Torino, Italy
[31] INFN-Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy
[32] INAF-Osservatorio Astrofisico di Torino, Via Osservatorio 20, 10025 Pino Torinese (TO), Italy
[33] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[34] Leiden Observatory, Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands
[35] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avenida Complutense 40, 28040 Madrid, Spain
[36] Port d'Informació Científica, Campus UAB, C. Albareda s/n, 08193 Bellaterra (Barcelona), Spain
[37] INAF-Osservatorio Astronomico di Roma, Via Frascati 33, 00078 Monteporzio Catone, Italy
[38] INFN section of Naples, Via Cinthia 6, 80126, Napoli, Italy
[39] Dipartimento di Fisica e Astronomia "Augusto Righi" - Alma Mater Studiorum Università di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
[40] Instituto de Astrofísica de Canarias, E-38205 La Laguna, Tenerife, Spain
[41] Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK
[42] European Space Agency/ESRIN, Largo Galileo Galilei 1, 00044 Frascati, Roma, Italy
[43] Université Claude Bernard Lyon 1, CNRS/IN2P3, IP2I Lyon, UMR 5822, Villeurbanne, F-69100, France
[44] UCB Lyon 1, CNRS/IN2P3, IUF, IP2I Lyon, 4 rue Enrico Fermi, 69622 Villeurbanne, France
[45] Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK
[46] Department of Astronomy, University of Geneva, ch. d'Ecogia 16, 1290 Versoix, Switzerland
[47] INFN-Padova, Via Marzolo 8, 35131 Padova, Italy
[48] Aix-Marseille Université, CNRS/IN2P3, CPPM, Marseille, France
[49] INAF-Istituto di Astrofisica e Planetologia Spaziali, via del Fosso del Cavaliere, 100, 00100 Roma, Italy
[50] Space Science Data Center, Italian Space Agency, via del Politecnico snc, 00133 Roma, Italy
[51] INFN-Bologna, Via Irnerio 46, 40126 Bologna, Italy
[52] University Observatory, LMU Faculty of Physics, Scheinerstr. 1, 81679 Munich, Germany
[53] FRACTAL S.L.N.E., calle Tulipán 2, Portal 13 1A, 28231, Las Rozas de Madrid, Spain
[54] Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, 85748 Garching, Germany
[55] Universitäts-Sternwarte München, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstr. 1, 81679 München, Germany
[56] Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029 Blindern, 0315 Oslo, Norway
[57] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA, 91109, USA
[58] Department of Physics, Lancaster University, Lancaster, LA1 4YB, UK
[59] Felix Hormuth Engineering, Goethestr. 17, 69181 Leimen, Germany
[60] Technical University of Denmark, Elektrovej 327, 2800 Kgs. Lyngby, Denmark
[61] Cosmic Dawn Center (DAWN), Denmark
[62] Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany
[63] NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
[64] Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK
[65] Department of Physics and Helsinki Institute of Physics, Gustaf Hällströmin katu 2, University of Helsinki, 00014 Helsinki, Finland
[66] Université de Genève, Département de Physique Théorique and Centre for Astroparticle Physics, 24 quai Ernest-Ansermet, CH-1211 Genève 4, Switzerland
[67] Department of Physics, P.O. Box 64, University of Helsinki, 00014 Helsinki, Finland
[68] Helsinki Institute of Physics, Gustaf Hällströmin katu 2, University of Helsinki, 00014 Helsinki, Finland
[69] Laboratoire d'etude de l'Univers et des phenomenes eXtremes, Observatoire de Paris, Université PSL, Sorbonne Université, CNRS, 92190 Meudon, France
[70] SKAO, Jodrell Bank, Lower Withington, Macclesfield SK11 9FT, UK
[71] Centre de Calcul de l'IN2P3/CNRS, 21 avenue Pierre de Coubertin 69627 Villeurbanne Cedex, France
[72] INFN-Sezione di Milano, Via Celoria 16, 20133 Milano, Italy
[73] Universität Bonn, Argelander-Institut für Astronomie, Auf dem Hügel 71, 53121 Bonn, Germany
[74] INFN-Sezione di Roma, Piazzale Aldo Moro, 2 - c/o Dipartimento di Fisica, Edificio G. Marconi, 00185 Roma, Italy
[75] Department of Physics, Institute for Computational Cosmology, Durham University, South Road, Durham, DH1 3LE, UK
[76] Université Paris Cité, CNRS, Astroparticule et Cosmologie, 75013 Paris, France
[77] CNRS-UCB International Research Laboratory, Centre Pierre Binétruy, IRL2007, CPB-IN2P3, Berkeley, USA
[78] University of Applied Sciences and Arts of Northwestern Switzerland, School of Engineering, 5210 Windisch, Switzerland
[79] Institut d'Astrophysique de Paris, 98bis Boulevard Arago, 75014, Paris, France
[80] Telespazio UK S.L. for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[81] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona), Spain
[82] European Space Agency/ESTEC, Keplerlaan 1, 2201 AZ Noordwijk, The Netherlands
[83] School of Mathematics, Statistics and Physics, Newcastle University, Herschel Building, Newcastle-upon-Tyne, NE1 7RU, UK
[84] DARK, Niels Bohr Institute, University of Copenhagen, Jagtvej 155, 2200 Copenhagen, Denmark
[85] Waterloo Centre for Astrophysics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[86] Department of Physics and Astronomy, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[87] Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada
[88] Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, 91191, Gif-sur-Yvette, France
[89] Centre National d'Etudes Spatiales – Centre spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[90] Institute of Space Science, Str. Atomistilor, nr. 409 Măgurele, Ilfov, 077125, Romania
[91] Dipartimento di Fisica e Astronomia "G. Galilei", Università di Padova, Via Marzolo 8, 35131 Padova, Italy
[92] Institut de Recherche en Astrophysique et Planétologie (IRAP), Université de Toulouse, CNRS, UPS, CNES, 14 Av. Edouard Belin, 31400 Toulouse, France
[93] Université St Joseph; Faculty of Sciences, Beirut, Lebanon
[94] Departamento de Física, FCFM, Universidad de Chile, Blanco Encalada 2008, Santiago, Chile
[95] Universität Innsbruck, Institut für Astro- und Teilchenphysik, Technikerstr. 25/8, 6020 Innsbruck, Austria

[96] Institut d'Estudis Espacials de Catalunya (IEEC), Edifici RDIT, Campus UPC, 08860 Castelldefels, Barcelona, Spain
[97] Satlantis, University Science Park, Sede Bld 48940, Leioa-Bilbao, Spain
[98] Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, 08193 Barcelona, Spain
[99] Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, PT1749-016 Lisboa, Portugal
[100] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Tapada da Ajuda, 1349-018 Lisboa, Portugal
[101] Cosmic Dawn Center (DAWN)
[102] Niels Bohr Institute, University of Copenhagen, Jagtvej 128, 2200 Copenhagen, Denmark
[103] Universidad Politécnica de Cartagena, Departamento de Electrónica y Tecnología de Computadoras, Plaza del Hospital 1, 30202 Cartagena, Spain
[104] Caltech/IPAC, 1200 E. California Blvd., Pasadena, CA 91125, USA
[105] Dipartimento di Fisica e Scienze della Terra, Università degli Studi di Ferrara, Via Giuseppe Saragat 1, 44122 Ferrara, Italy
[106] Istituto Nazionale di Fisica Nucleare, Sezione di Ferrara, Via Giuseppe Saragat 1, 44122 Ferrara, Italy
[107] INAF, Istituto di Radioastronomia, Via Piero Gobetti 101, 40129 Bologna, Italy
[108] Astronomical Observatory of the Autonomous Region of the Aosta Valley (OAVdA), Loc. Lignan 39, I-11020, Nus (Aosta Valley), Italy
[109] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice cedex 4, France
[110] ICSC - Centro Nazionale di Ricerca in High Performance Computing, Big Data e Quantum Computing, Via Magnanelli 2, Bologna, Italy
[111] Instituto de Física Teórica UAM-CSIC, Campus de Cantoblanco, 28049 Madrid, Spain
[112] CEA Saclay, DFR/IRFU, Service d'Astrophysique, Bat. 709, 91191 Gif-sur-Yvette, France
[113] Université PSL, Observatoire de Paris, Sorbonne Université, CNRS, LERMA, 75014, Paris, France
[114] Université Paris-Cité, 5 Rue Thomas Mann, 75013, Paris, France
[115] Univ. Grenoble Alpes, CNRS, Grenoble INP, LPSC-IN2P3, 53, Avenue des Martyrs, 38000, Grenoble, France
[116] Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 2, 00185 Roma, Italy
[117] Aurora Technology for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[118] Zentrum für Astronomie, Universität Heidelberg, Philosophenweg 12, 69120 Heidelberg, Germany
[119] Dipartimento di Fisica - Sezione di Astronomia, Università di Trieste, Via Tiepolo 11, 34131 Trieste, Italy
[120] ICL, Junia, Université Catholique de Lille, LITL, 59000 Lille, France
[121] CERCA/ISO, Department of Physics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA
[122] Technical University of Munich, TUM School of Natural Sciences, Physics Department, James-Franck-Str. 1, 85748 Garching, Germany
[123] Laboratoire Univers et Théorie, Observatoire de Paris, Université PSL, Université Paris Cité, CNRS, 92190 Meudon, France
[124] Departamento de Física Fundamental. Universidad de Salamanca. Plaza de la Merced s/n. 37008 Salamanca, Spain
[125] IRFU, CEA, Université Paris-Saclay 91191 Gif-sur-Yvette Cedex, France
[126] Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, 67000 Strasbourg, France
[127] Center for Data-Driven Discovery, Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan
[128] California Institute of Technology, 1200 E California Blvd, Pasadena, CA 91125, USA
[129] Department of Physics & Astronomy, University of California Irvine, Irvine CA 92697, USA
[130] Department of Mathematics and Physics E. De Giorgi, University of Salento, Via per Arnesano, CP-I93, 73100, Lecce, Italy
[131] INFN, Sezione di Lecce, Via per Arnesano, CP-193, 73100, Lecce, Italy
[132] INAF-Sezione di Lecce, c/o Dipartimento Matematica e Fisica, Via per Arnesano, 73100, Lecce, Italy
[133] Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands
[134] Departamento Física Aplicada, Universidad Politécnica de Cartagena, Campus Muralla del Mar, 30202 Cartagena, Murcia, Spain
[135] Instituto de Física de Cantabria, Edificio Juan Jordá, Avenida de los Castros, 39005 Santander, Spain
[136] Department of Computer Science, Aalto University, PO Box 15400, Espoo, FI-00 076, Finland
[137] Instituto de Astrofísica de Canarias, E-38205 La Laguna; Universidad de La Laguna, Dpto. Astrofísica, E-38206 La Laguna, Tenerife, Spain
[138] Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute (AIRUB), German Centre for Cosmological Lensing (GCCL), 44780 Bochum, Germany
[139] Department of Physics and Astronomy, Vesilinnantie 5, University of Turku, 20014 Turku, Finland
[140] Serco for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[141] ARC Centre of Excellence for Dark Matter Particle Physics, Melbourne, Australia
[142] Centre for Astrophysics & Supercomputing, Swinburne University of Technology, Hawthorn, Victoria 3122, Australia
[143] Department of Physics and Astronomy, University of the Western Cape, Bellville, Cape Town, 7535, South Africa
[144] DAMTP, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, UK
[145] Kavli Institute for Cosmology Cambridge, Madingley Road, Cambridge, CB3 0HA, UK
[146] Department of Astrophysics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[147] Department of Physics, Centre for Extragalactic Astronomy, Durham University, South Road, Durham, DH1 3LE, UK
[148] Institute for Theoretical Particle Physics and Cosmology (TTK), RWTH Aachen University, 52056 Aachen, Germany
[149] INAF-Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, 50125, Firenze, Italy
[150] Centro de Astrofísica da Universidade do Porto, Rua das Estrelas, 4150-762 Porto, Portugal
[151] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, PT4150-762 Porto, Portugal
[152] HE Space for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[153] INAF - Osservatorio Astronomico d'Abruzzo, Via Maggini, 64100, Teramo, Italy
[154] Theoretical astrophysics, Department of Physics and Astronomy, Uppsala University, Box 516, 751 37 Uppsala, Sweden
[155] Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, HI 96822, USA
[156] Mathematical Institute, University of Leiden, Einsteinweg 55, 2333 CA Leiden, The Netherlands
[157] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[158] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, 59000 Lille, France
[159] Institute for Particle Physics and Astrophysics, Dept. of Physics, ETH Zurich, Wolfgang-Pauli-Strasse 27, 8093 Zurich, Switzerland
[160] Department of Astrophysical Sciences, Peyton Hall, Princeton University, Princeton, NJ 08544, USA

[161] Space physics and astronomy research unit, University of Oulu, Pentti Kaiteran katu 1, FI-90014 Oulu, Finland
[162] Department of Physics and Astronomy, Lehman College of the CUNY, Bronx, NY 10468, USA
[163] American Museum of Natural History, Department of Astrophysics, New York, NY 10024, USA
[164] International Centre for Theoretical Physics (ICTP), Strada Costiera 11, 34151 Trieste, Italy
[165] Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, 10010, New York, NY, USA
[166] David A. Dunlap Department of Astronomy & Astrophysics, University of Toronto, 50 St George Street, Toronto, Ontario M5S 3H4, Canada

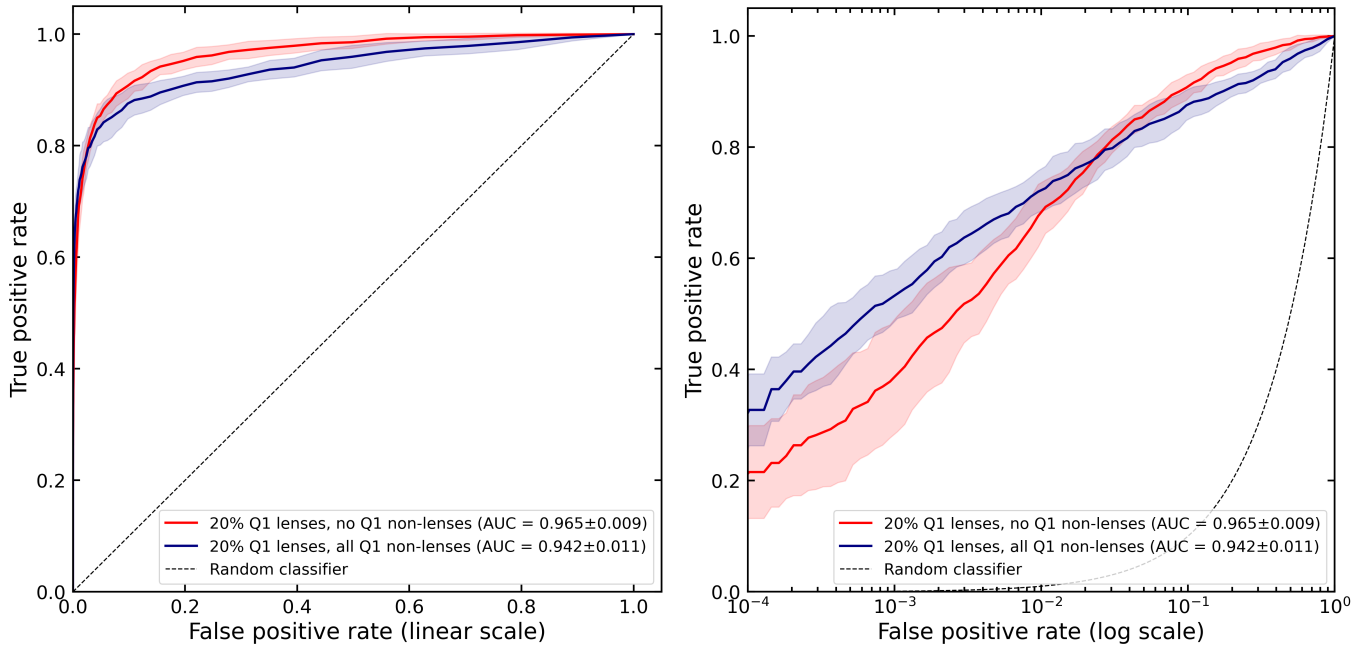

Fig. A.1. ROC curves linear scale (*left*) and log scale (*right*) showing the performance of two versions of the Zoobot network, each with different augmentations to the training data: one using 20% of Q1 lenses and no Q1 non-lenses; the other using 20% of Q1 lenses and all Q1 non-lenses.

Appendix A: Translating ROC curves into expected lens discoveries

Figure A.1 shows two ROC curves for models with two different training datasets. The first augments the pre-Q1 training data with 20% of the Q1 lenses and none of the Q1 non-lenses, and the second with 20% of the Q1 lenses and all of Q1 non-lenses. The former has a better AUC (0.965 ± 0.009) than the latter (0.942 ± 0.011). However, the version that includes the Q1 non-lenses in the training data outperforms the other version in the range TPR < 0.8. Despite representing a very small fraction of the ROC curve, it is the range of interest for lens finding. Even visually inspecting within the range FPR < 0.1 means that 10% of all the negatives have to be visually inspected: this translates to around 3 million images in a DR1-size sample. The discrepancy in the AUC comes from the difference in ability to recover lenses in the range $0.1 <$ FPR < 0.6, a range that lies beyond the scope of what could be visually inspected. For this reason, a higher AUC does not directly correlate with better lens-finding performance.