A&A, 704, A344 (2025)
https://doi.org/10.1051/0004-6361/202556142

Astronomy & Astrophysics

Counterpart identification and classification for eRASS1 and characterisation of the active galactic nuclei content

M. Salvato[1,2,⋆], J. Wolf[1,2,3], T. Dwelly[1], H. Starck[1], J. Buchner[1], R. Shirley[1], A. Merloni[1], A. Georgakakis[4], F. Balzer[1], M. Brusa[5,6], A. Rau[1], S. Freund[1], D. Lang[7], T. Liu[1], G. Lamer[8], A. Schwope[8], W. Roster[1], S. Waddell[1], M. Scialpi[9,10,11], Z. Igo[1], M. Kluge[1], F. Mannucci[11], S. Tiwari[1], D. Homan[12,8], M. Krumpe[8], A. Zenteno[13], D. Hernandez-Lang[14], J. Comparat[1], M. Fabricius[1], J. Snigula[1], D. Schlegel[15], B. A. Weaver[16], R. Zhou[17], A. Dey[16], F. Valdes[16], A. Myers[18], S. Juneau[16], H. Winkler[19], I. Marquez[20], F. di Mille[21], S. Ciroi[22], M. Schramm[7], D A. H. Buckley[23,24,25], J. Brink[8], M. Gromadzki[26], J. Robrade[27], and K. Nandra[1]

(Affiliations can be found after the references)

Received 27 June 2025 / Accepted 3 September 2025

ABSTRACT

Context. Accurately accounting for the Active Galactic Nucleus (AGN) phase in galaxy evolution requires a large, clean AGN sample. This is now possible with SRG/eROSITA, which completed its first all-sky X-ray survey (eRASS1) on June 12, 2020. The public Data Release 1 (DR1, Jan 31, 2024) includes 930,203 sources from the western Galactic hemisphere.

Aims. The data enable the selection of a large AGN sample and the discovery of rare sources. However, scientific return depends on accurate characterisation of the X-ray emitters, requiring high-quality multi-wavelength data. This paper presents the identification and classification of optical and infrared counterparts to eRASS1 sources.

Methods. Counterparts to eRASS1 X-ray point sources were identified using Gaia DR3, CatWISE2020, and Legacy Survey DR10 (LS10) with the Bayesian NWAY algorithm and trained priors. Sources were classified as Galactic or extragalactic via a machine-learning model combining optical/IR and X-ray properties, trained on a reference sample. For extragalactic LS10 sources, photometric redshifts were computed using CIRCLEZ.

Results. Within the LS10 footprint, all 656,614 eROSITA/DR1 sources have at least one possible optical counterpart; ~570 000 are extragalactic and likely AGN. Half are new detections compared to AllWISE, Gaia, and Quaia AGN catalogues. Gaia and CatWISE2020 counterparts are less reliable, due to the survey's shallowness and the limited amount of features available to assess the probability of being an X-ray emitter. In the Galactic plane, where the overdensity of stellar sources also increases the chance of associations, using conservative reliability cuts, we identified approximately 18 000 Gaia and 55 000 CatWISE2020 extragalactic sources.

Conclusions. We have released three high-quality counterpart catalogues – plus the training and validation sets – as a benchmark for the field. These datasets have many applications, but in particular, they empower researchers to build AGN samples tailored for completeness and purity, accelerating the hunt for the Universe's most energetic engines.

Key words. methods: data analysis – methods: statistical – catalogs – surveys – galaxies: active – X-rays: general

1. Introduction

Since the work of the Nuker team[1], and in particular following the paper of Magorrian et al. (1998), it has become widely accepted that galaxies hosting an actively accreting supermassive black hole (SMBH) in their centre are not exceptions in the extragalactic population, but rather the norm. This recognition introduced the challenge of incorporating the active galactic nuclei (AGN) phase in galaxy evolution models, and highlighted the need to understand how the evolution of galaxies and their central SMBH are connected, and how the two influence each other. This interconnection is reflected in a series of more or less tight scaling relations (e.g. Gebhardt et al. 2000; Ferrarese & Merritt 2000), and is often attributed to AGN feedback. Through various mechanisms, this feedback can regulate, or even suppress, star formation, thereby influencing the growth of both the galaxy and its central black hole (e.g. Harrison & Ramos Almeida 2024, see review of and references therein). Among others, Kormendy & Ho (2013a,b) provided comprehensive reviews of the relationship between SMBHs and their host galaxies. Their studies, however, challenge a simplistic interpretation, showing that the properties of the SMBHs correlate differently with various galaxy components, pointing to a complex evolutionary interplay.

Completeness (i.e. how fully a sample represents the intended target or population without significant omissions) and purity (the degree to which a sample is free from contaminating interlopers) are critical to investigate observationally the scaling relations between galaxies and their central black holes. X-ray emission is primarily produced by the accretion of material onto supermassive black holes at the centres of AGN. This direct relationship allows for a more straightforward identification of AGN compared to other methods, such as optical or infrared selection, which may include contributions from star formation or other Galactic processes (e.g. Eckart et al. 2010). However, the method is biased against the most obscured sources, which, in contrast, are more readily identified through infrared (IR) selection criteria, which generally offer a higher completeness but lower purity (e.g. Hickox & Alexander 2018), i.e. the efficacy in isolating AGN activity from other

⋆ Corresponding author: mara@mpe.mpg.de
[1] https://en.wikipedia.org/wiki/Nuker_Team

sources of emission, minimising contamination from unrelated phenomena.

Most AGN samples used to date originate either from deep observations in small-area surveys (e.g. Civano et al. 2012; Hsu et al. 2014; Ni et al. 2021) or from shallow observations covering wide areas (e.g. ROSAT/2RXS Boller et al. 2016). The relation between the growth and galaxy evolution is complex and multi-parametric, and therefore, large and complete samples are needed to reveal underlying covariances among the key physical parameters.

The situation has changed thanks to the X-ray all-sky survey of the extended ROentgen Survey with an Imaging Telescope Array (eROSITA) (Predehl et al. 2021) on board the Spectrum-RG mission (Sunyaev et al. 2021). Already within the first 6 months of the eROSITA all-sky survey (eRASS1; Merloni et al. 2024), nearly one million sources had been detected in the western Galactic hemisphere, the vast majority of which are AGN. Notably, eRASS1 uncovered many rare AGN, such as luminous or high redshift objects (e.g. Wolf et al. 2021; Medvedev et al. 2021), which are typically missed by narrow, deep, pencil-beam surveys. The resulting AGN sample is highly complementary to those selected at other wavelengths and offers new insights into AGN evolution and their connection to host galaxy properties. Nevertheless, X-ray observations alone cannot drive the AGN/galaxy co-evolution studies. Establishing reliable associations with multi-wavelength counterparts is a critical step towards enabling the detailed characterisation of these sources, for example, distinguishing between Galactic and extragalactic origins and deriving their redshifts via spectroscopy and photometry. For all-sky surveys, this is particularly challenging, due to the varying quality of multi-band data across the sky, the complex geometry of surveys, and the relatively low spatial resolution of eROSITA (point spread function half energy width of ∼30″ and corresponding 1σ median positional uncertainty of ∼4.5″); for the fainter eRASS1 sources, positional uncertainties exceeding 15″ are not uncommon (see Merloni et al. 2024). Once counterparts are identified, their classification as extragalactic is necessary, so that redshift can be acquired with spectroscopy or computed via photometric redshift (see review by Salvato et al. 2018a). Only then can well-defined AGN samples be assembled, with selection controlled for various quality parameters, enabling robust statistical and physical analyses.

This paper describes both the methodology used to assign reliable counterparts to all eROSITA sources presented in the 0.2–2.3 keV selected Main catalogue of the public Data Release 1 (DR1) Merloni et al. (2024), the redshift determination and the subsequent definition of the AGN sample derived from these associations, including comparisons with other AGN samples. Papers that have already made use of the catalogues released with this paper span a large variety of topics: from the identification of unresolved clusters in the eROSITA point-source catalogue Balzer et al. (2025), to identification of BLAZARs (Hämmerich et al., in prep.); from the identification of eROSITA/DR1 variable sources Boller et al. (2025) and tidal disruption events (Grotova et al. 2025) to the identification of AGN in the Euclid Q1 fields (Euclid Collaboration 2025); from the study of the evolution of the X-ray and UV relation (Chira et al. 2025) to the study the X-ray properties of BAL QSOs (Hiremath et al, 2025) and soft X-ray eccess (Chen et al. 2025); from the targeting of eROSITA sources with SDSS-V (Kollmeier et al. 2025) to the spectra analysis of AGN (Aydar et al. 2025; Pulatova et al. 2025) in SDSS-V/DR19 (SDSS Collaboration 2025). A complementary analysis for the hard X-ray-selected sample was provided in Waddell et al. (2024). In Section 2, we briefly summarise the characteristics of the eRASS1 catalogue and the supporting multi-wavelength catalogues. The determination of the counterparts is detailed in Section 3, while Section 4 present a basic comparison with the counterparts to ROSAT/2RXS (Boller et al. 2021). Sections 5 and 6 focus on disentangling Galactic and extragalactic sources and on the redshift computation for the extragalactic population. The definition of the final eRASS1 AGN sample, along with the comparison to other AGN samples, is presented in Section 7. Section 8 describes the six catalogues released with this work: catalogues of eRASS1 counterparts using different optical catalogues, AGN samples in high and low Galactic latitude regions, and the training datasets used throughout the analysis. The summary and an outlook conclude the work in Section 9.

Throughout the paper, we provide AB magnitudes unless stated otherwise. In order to allow direct comparison with existing works from the literature on X-ray surveys, we adopt a flat ΛCDM cosmology with $h = H_0/[100\,\mathrm{km\,s^{-1}Mpc^{-1}}] = 0.7$, $\Omega_M = 0.3$, and $\Omega_\Lambda = 0.7$.

2. The data

Below and in Table 1, we summarise the key characteristics of the main eROSITA catalogue and the supporting multi-wavelength datasets used for the identification and characterisation of counterparts.

2.1. eROSITA/eRASS1 catalogue

Merloni et al. (2024) presented the Main and Hard-selected samples of X-ray sources detected with eROSITA in the first 6 months of the survey (eRASS1) over half of the sky (western Galactic hemisphere), the so-called eROSITA-DE region. In the Main catalogue, a uniform flux limit (at 50% completeness) of $F_{0.5-2\mathrm{keV}} > 5 \times 10^{-14}$ erg s^{-1} cm^{-2} is achieved over most of the extragalactic sky, with deeper coverage towards the Ecliptic poles. To facilitate the data analysis, the sky has been split first into overlapping tiles of 3.6 by 3.6 sq. degrees, and then into contiguous tiles of 3 by 3 sq. degrees (see Figure 2 in Merloni et al. 2024). From now on, we use the latter definition.

Here, we focus on the eRASS1 Main sample, which includes 930,203 sources detected in the 0.2–2.3 keV band at various depths depending on the sky location (deeper at the South Ecliptic Pole, shallower at the ecliptic plane (see Figure 1). 903,521 X-ray sources (97.17%) are classified as point-like in the eROSITA image (`EXT_LIKE`=0), out of which 23,648 are flagged as potentially spurious. The catalogue also provides the detection likelihood `DET_LIKE_0`, measured by PSF-fitting; at the lower detection likelihood, `DET_LIKE_0`=6, the probability that the source is a spurious detection is estimated to be 14% (Seppi et al. 2022), decreasing to 4% for sources at `DET_LIKE_0` ≥ 8. We will discuss the quality of the associations as a function of `DET_LIKE_0` in various sections of the paper. More technical details about the characteristics of the X-ray catalogue are presented in Merloni et al. (2024).

2.2. Supporting optical/IR data

The reliable identification and classification of the counterparts strongly depend on the availability of ancillary data and their quality. Most wide-area optical imaging surveys emphasise observations of the extragalactic sky and thus avoid the Galactic plane, where an excess of Galactic sources and thus

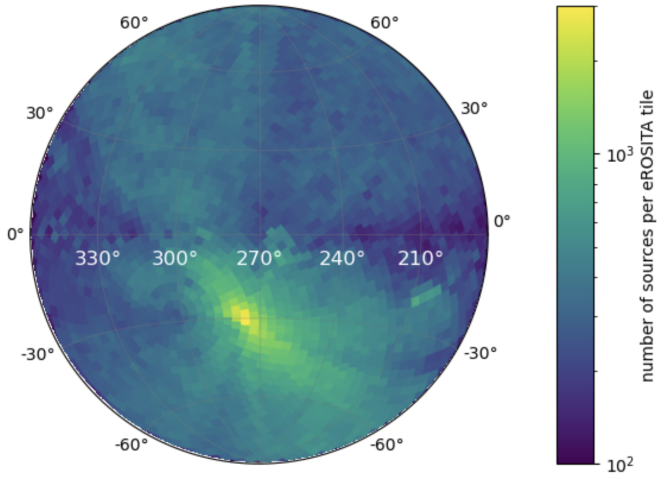

Fig. 1. Number of X-ray sources detected in the eRASS1 in contiguous eROSITA tile (3 deg × 3 deg), over the entire eROSITA_DE region in zenithal equal area (ZEA) projection. The highest density is at the South Ecliptic Pole (SEP). The map is shown in Galactic coordinates.

source confusion make the region challenging for extragalactic studies. In general, no single optical/IR survey is sufficiently wide, deep and complete for our purposes. Therefore, we have searched for counterparts to eRASS1 sources from a small set of complementary catalogues, described below.

2.2.1. Data Release 10 of the Legacy Survey imaging in support of DESI

The Data Release 10 of the Legacy Survey imaging in support of DESI (LS10) was made public in January 2023, and while the data analysis is similar to Data Release 9 (Dey et al. 2019), two significant differences characterise LS10. First, the survey covers the entire southern extragalactic sky (see Figure 1 of Saxena et al. 2024). This was possible thanks to the inclusion, among others, of DECam data taken under the DeROSITAS programme (P.I.Zenteno; Zenteno et al. 2025), designed to provide ancillary data to eROSITA for cosmological studies (e.g. Bulbul et al. 2024; Ghirardini et al. 2024), outside the area of DES (The Dark Energy Survey Collaboration 2005). Second, the survey includes *i* band data, which increases the number of detected sources[2].

LS10 includes the photometry in the WISE near and mid-infrared bands measured at the position of the optical sources (Lang 2014), using, for W1 and W2, the data acquired through year 7 of NEOWISE-Reactivation[3] and the AllWISE data for W3 and W4. For sources that are detected also by Gaia (e.g.Gaia Collaboration 2016), the basic astrometric and photometric Gaia DR3 columns are also included in the catalogue. The projected sky density of the LS10 catalogue per eROSITA tile is shown in the left panel of Figure 2.

2.2.2. Gaia DR3 (GDR3)

Gaia (e.g. Gaia Collaboration 2016) provides an all-sky catalogue with precise coordinates of optical point-like sources (i.e. stellar objects, QSOs, and compact nuclei of galaxies with a limiting magnitude of about G = 20.7, in three bands (G, BP, RP). In addition, it provides reliable information on variability (Eyer et al. 2023), proper motion, and parallax. Last but not least, the third data release (GDR3; Gaia Collaboration 2023b) also provides low-resolution spectra from the blue (BP) and red (RP) photometers (Montegriffo et al. 2023). Based on spectra, photometry and astrometry, five different machine-learning algorithms (Delchambre et al. 2023a) are used (a) to divide the sources in a probabilistic manner into broad astrophysical classes (single stars, white dwarfs, binaries, galaxies and quasars); (b) to determine the redshift of the sources depending on their classification; (c) to estimate the total Galactic[4] extinction. In this way, GDR3 provides a classification for 1.592 billion sources; approximately 11.4 million sources are classified as extragalactic, 7.8 million of which have a redshift estimation (see Delchambre et al. 2023a). The projected sky density of the GDR3 catalogue per eROSITA tile used here is shown in the middle panel of Figure 2.

2.3. CatWISE2020 (CW2020)

The CatWISE2020 survey (CW2020; Marocco et al. 2021) covers the entire sky with the WISE observations from 2010 and 2018 in W1 and W2. The catalogue is generated using CROWDSOURCE (Schlafly 2021) and its 90% completeness depth is 17.7 and 17.5 magnitude (in Vega system) for W1 and W2, respectively. As for eROSITA, also the depth of CW2020 varies substantially across the sky, being deeper at the Ecliptic poles and shallower in the Ecliptic plane. In addition, CW2020 is about 1 magnitude shallower in the Galactic plane than elsewhere due to source confusion. The catalogue includes 1.89 billion sources, and the projected sky density of the CW2020 catalogue per eROSITA tile used here is shown in the right panel of Figure 2.

3. Determination of counterparts

It is important to remember that one of the parameters that determines the reliability of the counterparts is the number density of the sources (see Figure 2) in the multi-wavelength supporting data (the higher the source density, the larger is the chance of random associations). In turn, the number density depends on the wavelength of the survey, the depth, and the sky coordinates. This, combined with the smoothly increasing depth of eROSITA from the Galactic plane to the Ecliptic poles, challenges our ability to distinguish the correct counterparts from chance associations. In this section, we discuss how we determined the most likely counterparts to the eROSITA sources.

Given the different nature of the X-ray emission, the identification of the counterparts is made separately for sources classified as extended and point-like in the eROSITA images. Extended X-ray sources are usually indicate the presence of clusters of galaxies (Kluge et al. 2024) or supernova remnants (SNR), while point sources identify stellar objects in the Milky Way or in the very nearby galaxies (Large and Small Magellanic Cloud), active black holes in galaxies (AGN, QSO) or high-redshift, extragalactic transients of various origins (e.g. Grotova et al. 2025), and unresolved clusters (Balzer et al. 2025).

There are two approaches to the determination of the counterparts. In the first case, one is interested in determining all sources of a certain type. Within the eROSITA-DE collabora-

[2] About 30% more sources in the test area in the COSMOS field are detected using *griz*, instead of *grz* (DESI, priv. comm.).
[3] https://wise2.ipac.caltech.edu/docs/release/neowise/neowise_2021_release_intro.html
[4] Indicated as Milky Way, MW in this paper.

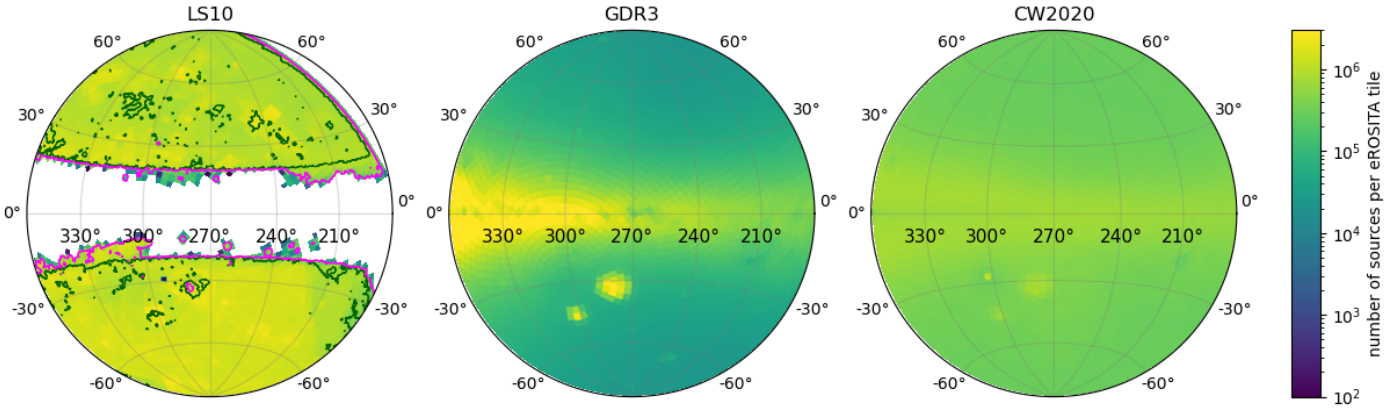

Fig. 2. Source density in each eROSITA sky tile (3 deg × 3 deg) of LS DR10 (LS10; left panel), Gaia DR3 (GDR3; middle panel), and CatWISE2020 (CW2020; right panel). On the LS10 map, the dark green (magenta) regions of InAllLS10 (InAnyLS10), indicating whether all (at least one of) the LS10 bands reach the nominal depth of the survey (see Section 3.3), are overplotted.

Table 1. Summary properties of the surveys used in this paper.

Survey	Area coverage (square degrees)	Depth (Approx.)	Number of sources
eROSITA/DR1	20,626.5	$F_{0.5-2\,\mathrm{keV}} > 5 \times 10^{-14}$ erg s^{-1} cm^{-2}	930 203
LS10	9,582	r = 23.3	2.81×10^9
GDR3	41,252.961 (all sky)	G = 20.7	1.59×10^9
CW2020	41,252.961 (all sky)	W1 = 20.4	1.89×10^9

Table 2. Final list of features adopted to identify an X-ray emitter in LS10, GDR3, and CW2020, respectively.

LS10	Gaia DR3	CW2020
g, r, i, z, W1, W2 MW extinction corrected model fluxes	Gaia *G, BP, RP* mag, corrected for MW extinction	W1 and W2 fluxes
Model flux errors	Gaia colours, corrected for MW extinction	W1 and W2 Flux errors
All possible colours MW extinction corrected	Gaia proper motion and error	W1-W2 colour
Shape_r, shape_e1, shape_e2, n_{Sersic}	Gaia parallax and error	W1 and W2 apertures photometry
Gaia photometry	Gaia astrometric_excess_noise	proper motion and error
Gaia S/N		
Gaia proper motion and error		
Gaia parallax and error		
Recall fraction 87.4%	Recall fraction 75.6%	Recall fraction 85.1%
Leakage 0.7%	Leakage 0.4%	Leakage 6%

tion, this is the approach followed in Kluge et al. (2024) (focusing on the identification of clusters), Hämmerich et al., (in prep; focusing on the identification of Blazars), Schwope et al.; Schwope et al. (2024a; 2024b, focusing on the identification of Cataclysmic variables, CVs), Avakyan et al., (in prep; focusing on X-ray binaries), and Freund et al. (2024, focusing on the search for coronal stars). This approach is extremely reliable for defining samples of sources with high purity. However, it does not account for the possibility that the emission comes from a completely different type of source (e.g. AGN).

The second approach is to account for every type of X-ray emitter within a certain distance from the X-ray source and attempt a classification only afterwards. This is the approach adopted in the past by our group for ROSAT and XMM-Slew2 (Salvato et al. 2018a), in eROSITA/eFEDS Salvato et al. (2022, 2018a), for Euclid/Q1 (Euclid Collaboration 2025) and in this paper. In all three cases, the Bayesian algorithm `NWAY` (Salvato et al. 2018b) was used, in combination with a prior that effectively predicts, from non-X-ray data alone, for each source the probability to be an X-ray emitter. For the complete description of the algorithm, readers are referred to section B4 in the Appendix of Salvato et al. (2018a), while the principles and the method for constructing the prior are described in detail in Salvato et al. (2022) (S22 hereafter). We provide here a summary of the concept, followed by its application in order to determine the eRASS1 counterparts in LS10, GDR3, and CW2020.

3.1. eRASS1 counterparts using NWAY

We use the Bayesian cross-matching algorithm `NWAY` to generate counterpart catalogues for the eRASS1 point sources in three supporting surveys independently: LS10, GDR3, and CW2020.

For each source in the primary eRASS1 catalogue NWAY ranks all the sources in the supporting catalogues within 60″ from the X-ray position, combining the probabilities a) based on the astrometric configuration (i.e. the positional error-normalised distances separating X-ray centroid and counterpart candidates and accounting for local source density) and b) on the candidates' non-astrometric properties such as photometry, colours, morphology, or parallax. The two components of the final probability are then combined in the probability p_i, following Section 3.1 of S22.

The non-astrometric information is condensed in a random forest model, assigning a probability of being X-ray detected to each counterpart candidate. Following S22, the model was constructed using as training samples (a) secure counterparts (regardless of whether the sources are compact objects, stars, AGN, or QSO) from 4XMM DR13 (Webb et al. 2020) and *Chandra* CSC2.0[5] and (b) field sources within 60″ of the X-ray position, after removing the correct counterparts. The construction of the training samples is described in Appendix A and released with this paper (see Section 8).

3.1.1. Strength of the non-astrometric information

As described in Appendix A.3, for the training of the random forest priors, we have extracted all LS10, GDR3, and CW2020 catalogue sources within 60″ of the X-ray centroids of the 4XMM and Chandra training samples, thereby including the real multi-wavelength counterparts as well as non-associated field sources. For each catalogue, a random forest (sklearn implementation, Pedregosa et al. 2011) we formally defined a sample of true X-ray counterparts (astrometrically best match counterpart to the 4XMM or Chandra source with a maximal separation of <1″) and field sources (separation >1″) as the input training sample. For each training sample, a subset of columns was extracted from the catalogues (e.g. fluxes, colours, signal-to-noise, etc.). Using these features, we train the model as a classifier to correctly identify true counterparts and field sources, also accounting for the missing features for some of the sources. The final set of features, listed in Table 2, was refined during model tuning by sorting for feature importance (feature_importances_). The models are optimised for precision and recall fraction[6], defined as true positives/(true positives + false negatives), and fractional leakage, defined as false positives/(false positives + true negatives). Our final classifiers on a test sample reach a minimum recall fraction of 75.6% and a maximum fractional leakage of 0.7%, as shown in Figure 3. These trained models are then used as a pre-processing step of the NWAY run by predicting the probability of each candidate counterpart to be an X-ray emitting source. Note that while the recall leakage fractions presented in Figure 3 are based on discrete labels (threshold set at 0.5), we feed into NWAY a continuous variable.

Not surprisingly, the survey that provides the more reliable prior is LS10 (with a recall fraction of 87.4%), due to the depth and resolution of the data and the number of features that can be used for separating X-ray and non-X-ray emitting objects, thus essentially able to parametrise the SED of the sources. This is illustrated in Fig. 3, where the confusion matrices resulting from the random forest prediction on a validation test set using the final set of features from LS10, GDR3 and CW2020 are reported. CW2020 has almost the same recall fraction as LS10, but the number of false identifications is ten times higher, probably due to the limited number of features available. It is important to note that the number of sources available in each training sample is different, with the number being the highest in LS10, making the associations more reliable under any aspect.

3.1.2. Combining astrometry and priors

One of the main characteristics of NWAY is that it is based on Bayesian statistics, allowing one to combine via multiplication the probabilities of a source to be the correct counterpart, determined in independent ways. So, given the coordinates of an eRASS1 source, for all the sources of a survey within 60″ from those coordinates, the probability of being the counterpart on the basis of astrometry is multiplied by the probability of being an X-ray emitter (assuming the corresponding random forest model). The impact that the prior has on the final probability to be the correct counterpart is expressed in the quantity bias_Xray_proba (see Appendix B.6 in Salvato et al. 2018b). A value of 1 indicates no change in the probability of being the right counterpart, i.e. the probability is based solely on the distance to the X-ray position, positional uncertainties, and source number densities. Values below 1 indicate that the prior has degraded the probability, while values above 1 indicate that the prior has reinforced the association. The left panel of Figure 4 shows this impact (i.e. weight) on the counterparts from the three surveys. The sources with bias above 1 are 78%, 55% and 70% for LS10, GDR3, and CW2020, respectively. While this indicates that for LS10 and CW2020, a large fraction of the counterparts would have been correctly selected from geometry, this is not the case for GDR3. For the latter, almost half of the GDR3 sources had a high probability of being the right counterparts based on distance from the X-ray position, but did not have the features typical of an X-ray emitter, indicating that they were just chance associations. The weighting has an ever stronger impact for the randomised eRASS1 catalogue (right panel of Figure 4), where in all three surveys, the weighting is below 1 for most sources.

The final probability provided by NWAY is encapsulated in two values, p_any, and p_i. For each X-ray source in the eRASS1 catalogue, the p_any value gives the probability that at least one of the sources in the supporting multi-wavelength catalogues is the correct counterpart. At the same time, for each source in the supporting catalogue within 60″ from the eRASS1 detection, p_i provides the relative probability of the object being the correct counterpart[7].

The source in the supporting catalogues with the higher p_i is assigned as the counterpart. With this paper, we provide catalogues of counterparts using the three surveys separately (see Section 8) but recommend that the user uses LS10 when possible.

3.1.3. Quantifying chance associations

In S22, the procedure for computing the counterparts with NWAY in combination with machine-learning for determining the prior was introduced for the first time. In the same paper, the authors quantified the NWAY ability to recover the correct counterpart of an X-ray source, by considering a sample of about 3500 *Chan-*

[5] primary_source table via the database query web API (https://cxc.cfa.harvard.edu/csc/cli/index.html)

[6] In this context, a true positive is a correctly identified X-ray emitter, a true negative is a mislabelled X-ray emitter.

[7] For the analytical demonstration of the connection between p_any and p_i see the Appendix in Salvato et al. (2018a).

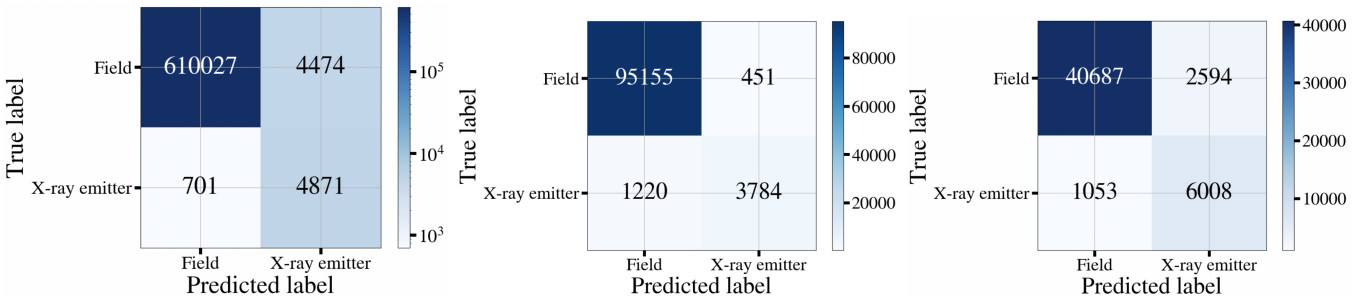

Fig. 3. Confusion matrices resulting from the Random Forest classification on the validation test sets are shown for LS10 (left panel), GDR3 (middle panel), and CW2020 (right panel). The colour scale in the panels represents the number of sources per cell, with darker shades indicating higher counts. The entries along the main diagonal (top-left to bottom-right) indicate correctly classified sources. The different total number of sources across the three matrices reflects the variations in survey depth, spatial resolution, and source density among the three surveys.

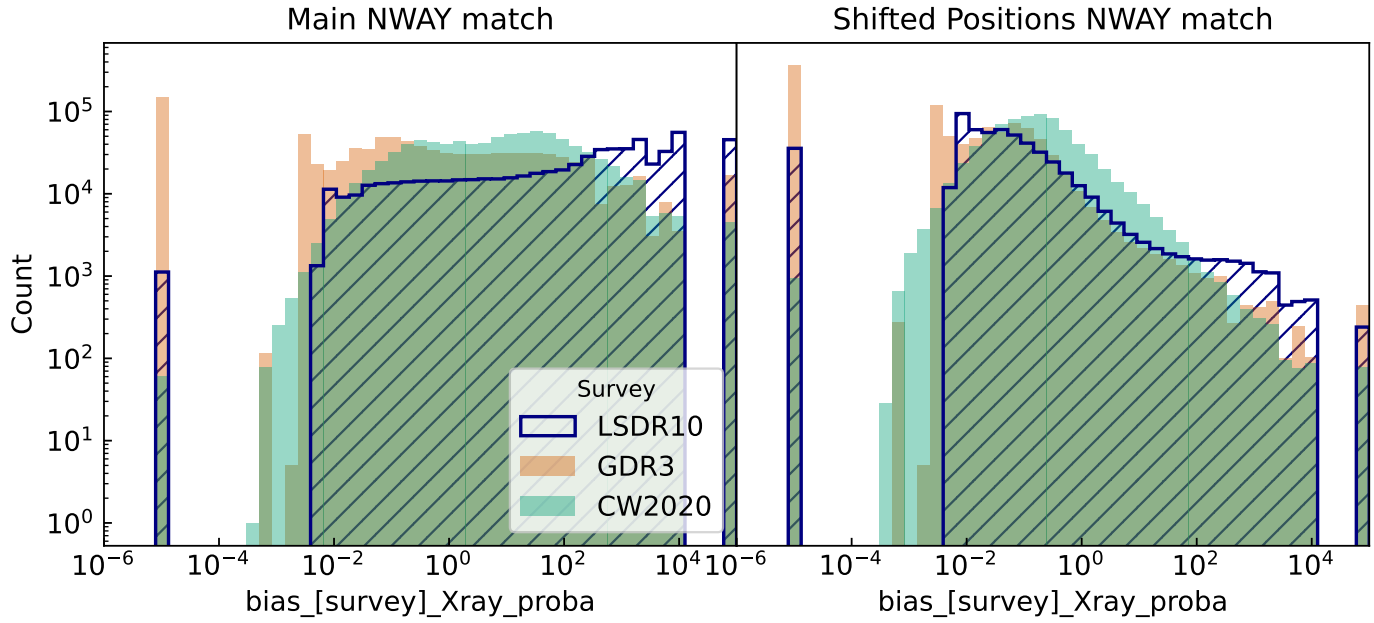

Fig. 4. Histogram distribution of the probability weighting (i.e. bias), introduced by the priors for the LS10, GDR3 and CW2020 counterparts to eRASS1. The left panel shows the distribution of the bias for the actual eRASS1 counterparts, while the right panel displays the distribution for counterparts to the eRASS1 shifted positions. A value of 1 indicates no change in the probability of being the right counterpart, i.e. the probability is based solely on the distance to the X-ray position, positional uncertainties, and source number densities. Values above(below) 1 indicate that the prior has degraded(reinforced) the probability. Almost 50% of the counterparts from GDR3 got degraded, after considering the prior. More details in the main text (Section 3.1.2).

dra sources[8], and modifying the original very small positional error by assigning an eROSITA positional error randomly sampling from the astrometric uncertainties listed in the core eFEDS source catalogue (see for details Section 4.1 of Salvato et al. 2018a). It was demonstrated there that the correct counterpart was recovered in 94% of the cases.

However, in that experiment, the distribution of p_any in the case of chance associations was not quantified. Only knowing that, the user can decide which p_any threshold and control of completeness and purity of a sample.

In order to assess at any given p_any the fraction of possible spurious associations, the coordinates of the X-ray sources have been randomised[9], and the sources with new coordinates falling within 60″ from a known (real) eRASS1 detection, have been removed. More precisely, the sources have been shifted by 3' in R.A. and DEC. In this way, most sources remain within the original eROSITA tile, so that the densities of primary and secondary catalogues are preserved and the probabilities can be compared.

Then, NWAY is re-run on this eRASS1 random catalogue with the same settings defined for the real sample for the three supporting catalogues and the same priors. This gives us the distribution of p_any values for the association of the LS10, GDR3 and CW2020 sources to a random position. Given that the new positions are random and do not fall on existing X-ray real positions, the p_anyshould remain very small. This is because the random forest model decreases p_i probability of being an X-ray emitter for the sources that could be considered counterparts from pure astrometric considerations. Figure 4 shows the weight of the prior in the selection of the counterparts for the real eRASS1 sources and for the randoms. If the probability of the association is unaffected by the prior, the bias will be 1, while the bias will move close to zero or to large positive values if the prior decreases or increases the probability, respectively.

[8] With their very small positional error, the counterparts are easily identified.
[9] We preserved the original position uncertainties.

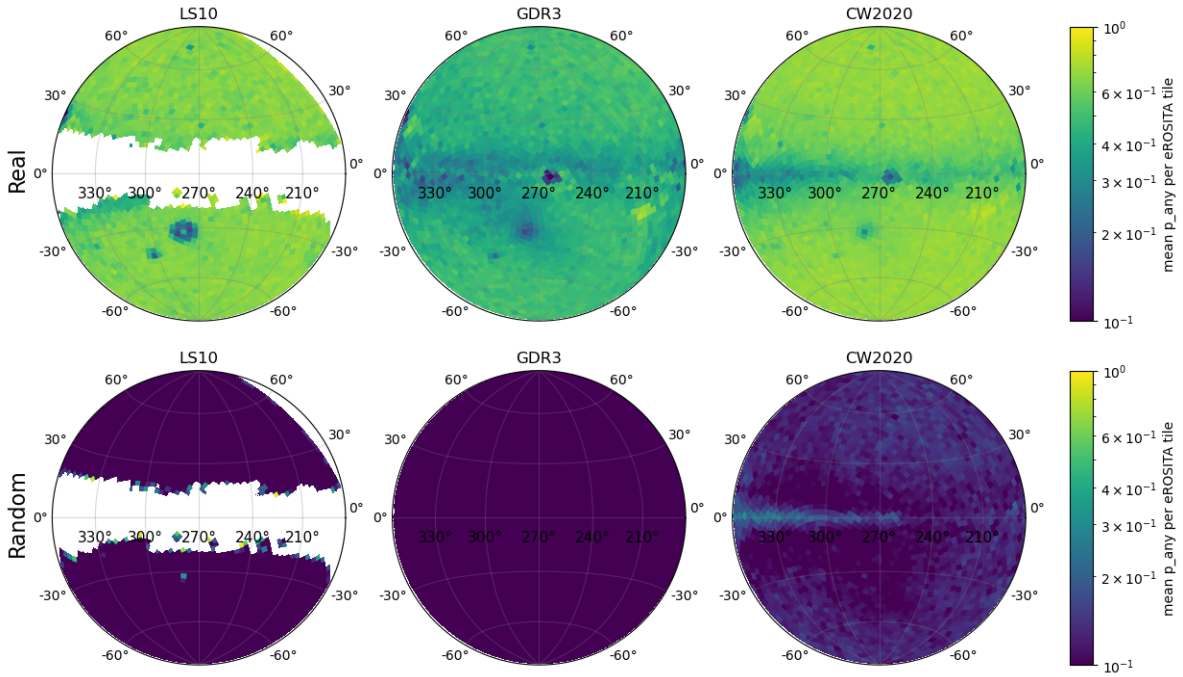

Fig. 5. Mean p_any distribution per eROSITA tile for real (top) and random (bottom row) eROSITA coordinates, using ancillary data from LS10 (left), GDR3 (middle) and CW2020 (right). The colour scale is consistent across all panels to facilitate comparison. While the p_any values for the real sources are generally higher than those for the random positions, they can approach similar levels in regions of high source density, indicating an increased risk of chance associations (see the main text for further discussions).

The large majority of the counterparts in the randoms have a bias close to zero, indicating that the probability assigned on the basis of closeness to the X-ray coordinates where heavily rescaled when accounting for the typical physical properties of an X-ray emitter.

Figure 5 provides a further visualisation of the difference between the counterparts associated with the real and random eRASS1 sources. There, the mean p_any per eROSITA tile is reported for each ancillary catalogue, both for the real eRASS1 sources (top) and for the random (bottom). Overall, the p_any values are typically much higher for the eRASS1 than for the randoms, with the regions of the Magellanic Clouds and the Galactic plane having the lower values. However, the typical p_any in the randoms is much lower, so that in general also in these two regions the risk of chance association is limited, with the exception of the Galactic plane for the CW2020 (see lower right panel of Figure 5).

3.1.4. Completeness and purity of NWAY associations

For a given ancillary catalogue, a threshold on the p_any value will determine the completeness (the fraction of real eRASS1 sources that have p_any above any given value) and purity (1 – the fraction of sources that have p_any above any given value in the corresponding random catalogue). The threshold depends on the scientific purpose that the user has in mind. For example, one could be interested in gathering as many counterparts to eRASS1 as possible, thus maximising the completeness. Others could be interested in creating a smaller but purer sample, thus reducing the number of chance associations as much as possible. Figure 6 shows in orange the completeness for LS10 (right panel), GDR3 (central panel) and CW2020 (left panel), respectively, depending on the DET_LIKE_0 values of the X-ray sources averaged over the entire eRASS1 survey. Similarly, in blue, the purity curves are constructed by subtracting from 1 the completeness in the random catalogue. This gives the probability that, at a given p_any, the association is due to chance. The purity increases with p_any (given that it is rare that, among the randoms, a high p_any is assigned), at the cost of reducing the sample size (completeness). It was already mentioned in S22 for eROSITA/eFEDS that the lower p_any and lower purity are for sources with low DET_LIKE_0, indicating that the X-ray source itself could be a spurious detection. The same trend is found for eRASS1.

The intersection between the completeness and purity curves provides the value of p_any for which the two quantities are the same. This is the threshold more commonly adopted (e.g. Brusa et al. 2007; Civano et al. 2012; Marchesi et al. 2016) for creating samples of reliable associations. While this threshold was useful for pencil-beam surveys, where the quality of the data and the number of sources in the primary and secondary catalogues remain the same across the survey area, it is misleading in the case of an all-sky survey, as Figure 1 and 2 have already shown. The consequence is that, depending on the location in the sky, the same p_any value would refer to different levels of completeness and purity. In the catalogues that we are releasing with this paper, we provide the p_any value at the intersection (threshold[6,7,8]) together with the value of the intersection itself (dubbed ComPur[6,7,8]), as a function of DET_LIKE_0 > 6,7,8. In addition, we provide the value of completeness (completeness[6,7,8]) and purity (purity[6,7,8]) at the p_any of each source, computed within the eROSITA tile to which the source belongs. An example of how this information can be used for creating a sample of AGN with high purity is described in Section 7. The maps of ComPur[6,7,8] and corresponding threshold[6,7,8] per each ancillary survey and as a function of DET_LIKE_0 are reported in Figures A.1 for clarity.

Figure 6 also shows that at the intersection, completeness and purity are higher for LS10, and this happens already at very low values of p_any, confirming the power of the LS10 model used for determining the correct counterpart. CW2020 reaches high values of purity and completeness, but at a higher value of p_any, thus substantially limiting the size of the final sample.

3.2. Separation and magnitude distributions

The distribution of the observed X-ray to optical/IR separations normalised by the X-ray positional uncertainty is shown in Figure 7 as a function of the *r*, *G*, and W1 magnitudes of the counterparts. The separation between the X-ray position and the assigned LS10, GDR3, and CW2020 counterparts is smaller than 15″ in 95%, 73% and 95% of the cases, with a mean of 1.24σ, 1.23σ and 1.19σ for LS10, GDR3, and CW2020, respectively. The faint clouds in all the plots correspond to the brightest objects, most of which are stars. As for eFEDS (S22), there is a trend for larger average X-ray-optical separations at smaller values of DET_LIKE_0, because at lower detection likelihood sources typically have larger X-ray positional uncertainty (Brunner et al. 2022; Merloni et al. 2024). The distribution is well comparable to the expectation of a Rayleigh distribution with a scale factor equal to 1, shown in green in the plots. This agreement adds confidence to the overall methodology adopted in assigning the counterparts. For GDR3 counterparts, the strongly declining Gaia completeness for objects fainter than G = 21 mag imprints a sharp cutoff onto the recovered sample.

3.3. Limitations of counterpart associations

One point to keep in mind is that the reliability of the counterparts is also bound to the reliability of the ancillary catalogues, which changes depending on the quality/depth of the images and the location in the sky. Artefacts and the presence of very nearby and bright sources will affect the source detection algorithms and, thus, the identification of the counterpart. We recommend that the user pay attention to the flags activated on the ancillary catalogues when deciding whether to trust a specific association.

3.3.1. LS10 quality masks

While GDR3 and CW2020 are all-sky surveys, LS10 covers only the extragalactic sky. At the border of the survey, the data are incomplete and shallow. The bottom-left panel of Figure 5 shows clearly how in these regions the reliability of the association is much lower (high p_any for the randoms). Taking advantage of the fact that LS10 provides depth estimates for PSF-like sources for each 0.25 × 0.25 deg2 'brick' and in each filter, we have added two flags:

- InAllLS10: when the counterpart is a brick where the *g, r, i, z* nominal depth (23.5, 23.3, 22.8 or 22.3) is reached in all bands.
- InAnyLS10: when the counterpart is in a brick where at least one of the *g, r, i, z* reached the nominal depth.

The masks are indicated in dark-green and magenta colours in the left panel of Figure 2.

Unlike typical pencil-beam surveys, an all-sky survey naturally includes very nearby, large, and resolved galaxies for which the standard way of measuring optical photometry is not reliable. For this reason, in our catalogues we have added an additional flag, inHEC, indicating whether or not the eROSITA coordinates fall

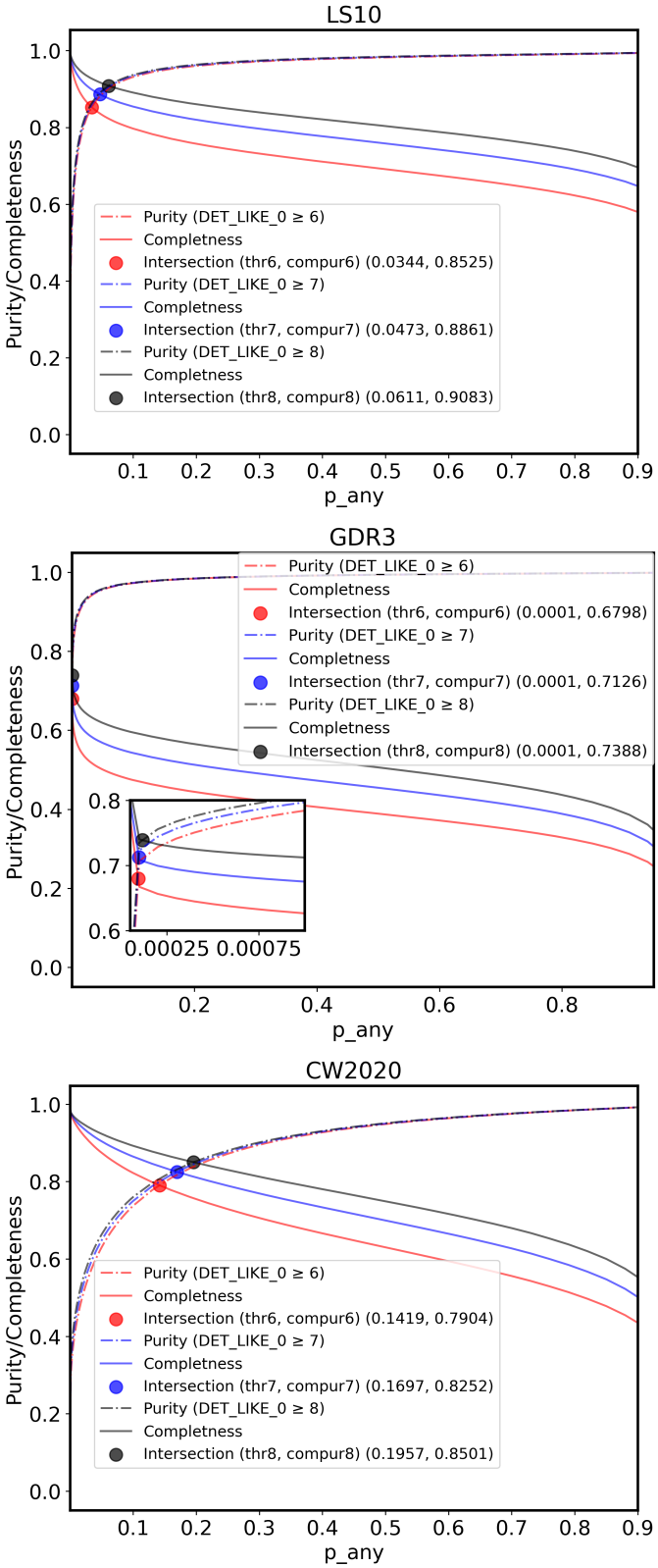

Fig. 6. Mean purity and completeness as a function of p_any for LS10, GDR3, and CW2020, averaged over their respective surveys' footprints (i.e. considering eROSITA sources outside the Galactic plane in the case of LS10). Different colours indicate different detection likelihood thresholds (DET_LIKE_0) for the eROSITA sources. The catalogues include the p_any thresholds and the completeness/purity intersection averaged compur[6,7,8] computed per eROSITA tile as a function of DET_LIKE_0.

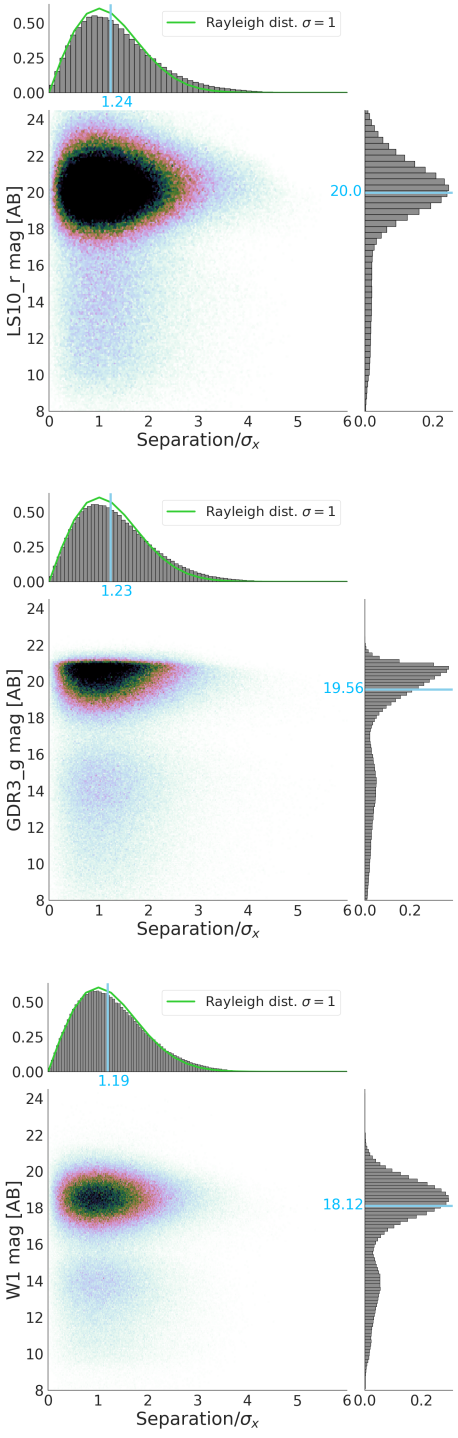

Fig. 7. Normalised separation between the eRASS1 X-ray position and the selected counterpart as a function of *r*-band (LS10), *G*-band (GDR3) and W1 (CW2020) magnitude, for sources with secure counterparts (`p_any`>=threshold6). The separation is normalised by the one-dimensional positional uncertainty of the eRASS1 source. The expected 1σ Rayleigh distribution for normalised separations (Rosen et al. 2016) is shown in green. Lines indicate the median values of the parameters.

within the area covered by the galaxy included in the Heraklion Extragalactic CATaloguE (HECATE[10]; Kovlakas et al. 2021). The catalogue lists about 205k nearby galaxies within a distance of about 200 Mpc, and it is based on the HyperLEDA Database[11]. Dedicated studies on eROSITA detected sources within galaxies in the HECATE catalogue are presented in Kyritsis et al. (2025) for the relation between X-ray and star formation.

The three columns (`inAnyLS10`, `inAllLS10` and `inHEC`) are reported in all three catalogues of counterparts to facilitate the comparison.

3.3.2. eRASS1 sources with zero or multiple counterparts

Not all the eRASS1 sources have a counterpart (`match_flag`= Null) in each of the three ancillary catalogues, and, at the same time, there are eRASS1 sources for which more than one plausible counterpart is identified within the same ancillary catalogue[12]. In our catalogues, all these cases will be listed. Here, it is interesting to look at the reasons for these behaviours.

In the entire eRASS1 catalogue, after removing the flagged sources, there are 10,906 (1.2%) sources with zero Gaia counterparts within 60″. The LS10 images at the location of these eRASS1 sources typically show the presence of very faint and/or extended sources, which will be missed from the Gaia catalogue, which is shallow and lists only bright compact or point sources.

On the other hand, only 2267 eRASS1 sources are without CW2020 counterparts (0.2%). In these cases, the WISE images typically show artefacts (e.g. stripes, saturation spikes) in the WISE/NEOWISE images; these artefacts would prevent the detection of the CW2020 sources, and they would then be missed from the catalogue.

To quantify the number of sources missing LS10 counterparts is complicated by the fact that LS10 does not cover eRASS1 completely. For that, we apply the same `InAllLS10` and `InAnyLS10` masks as described in Section 3.3. All the eRASS1 sources that fall in the footprint of LS10 do have at least one possible counterpart. This is ascribed to the depth and resolution of the data, but also to the strength of the multi-wavelength prior, which is defined using a larger number of features, when compared with GDR3 and CW2020 (see Table 2). In summary, the lack of a counterpart is a direct consequence of the quality of the supporting data.

However, there are also 5636 sources that have between two and seventeen alternative GDR3 counterparts, with the large majority of the cases located in the Galactic plane, on the Magellanic Clouds, or in star-forming regions. Similarly, there are 4644 sources with a number of possible counterparts between two (65% of the cases) and nine counterparts in CW2020. The eRASS1 sources with multiple CW2020 counterparts are once again more concentrated on the Galactic plane, on the Magellanic Clouds, or in star-forming regions, where the data are confusion-limited, also due to the coarse angular resolution of the CW2020 images. The same is happening for LS10, where 8933 sources have between two and seventeen counterparts. In all three cases, more than 50% of the sources have a very low `p_any` and `DET_LIKE_0`<7, suggesting the possibility that these X-ray sources are actually spurious detections (see discussion in Section 3.1.4).

3.4. Comparison between LS10, GDR3, and CW2020 associations

While we consider the counterparts determined with LS10 to be more reliable, the data are not available over the whole sky, and

[10] https://hecate.ia.forth.gr/

[11] http://atlas.obs-hp.fr/hyperleda/

[12] `match_flag`= 2 if $\texttt{p\_i}_j/\texttt{p\_i}_{\rm best} > 0.5$, with $\texttt{p\_i}_j$ being the `p_i` of the j source in the supporting catalogue

it is important to quantify under which assumptions the associations made with GDR3 and CW2020 can be trusted in that region. For this test, we consider only sources not flagged in the original X-ray catalogue (see details in Merloni et al. 2024) and outside the area covered by each HECATE galaxy. We cross-matched the LS10 counterparts with those from the GDR3 and the CW2020 catalogues via eRASS ID, and examined how the agreement varies as a function of X-ray detection likelihood, DET_LIKE_0, within the inAllLS10 and inAnyLS10 footprints. The results are summarised in Table 3 and described in greater detail below.

3.4.1. Within the footprint of LS10

At increasing eROSITA detection likelihood, the overall number of sources within the LS10 footprint decreases, but the fraction of sources with the same counterparts in GDR3 and CW2020 increases. This is because there is a correlation between the detection likelihood and the brightness of the X-ray source, which overall corresponds to a smaller positional uncertainty and brighter counterparts. Then, due to the depth of the ancillary data, the agreement is higher between LS10 and CW2020 than between LS10 and GDR3. In this latter case, if we consider only LS10 sources brighter than 20.7 magnitude in G band[13], the agreement between LS10 and GDR3 increases up to 94.4% at DET_LIKE_0>8. A similar trend is seen for the comparison between LS10 and CW2020, limiting the LS10 sample to sources brighter than 20.4 in W1 (thus simulating the average depth of CW2020). It is reassuring to see the very good agreement between the associations, despite the different features used to identify X-ray emitters in each survey. Finally, it is also reassuring that the fraction of sources in agreement is consistent between the area within inAllLS10 and inAnyLS10, respectively. This allows us to conclude that the prior adopted for the determination of the counterpart in the LS10 region is also robust when the quality of the data is not optimal.

3.4.2. Outside the footprint of LS10

For the area outside LS10 (defined by inAnyLS10=0), the comparison between GDR3 and CW2020 is presented in Table 4. At any detection likelihood, there are more eRASS1 sources with a counterpart only from CW2020 than from GDR3. This is consistent with the lower reliability of the Random Forrest prior based on GDR3, presented in Section 3. It means that in many cases, the GDR3 counterparts are only chance associations, and a higher threshold for the purity must be applied.

3.5. Comparison with HamStar

Freund et al. (2024) used the Bayesian algorithm HamStar (Freund et al. 2018), to identifying the GDR3 counterparts to eRASS1 associated with a coronally emitting star. They found 149,290 reliable associations ($p^{i}_{\rm stellar} > 0.53$), out of which 68 614 (58 573) are in the region InAnyLs10(inAllLS10). Of these, 91.8% (92.9%) are the same counterparts identified here and classified as Galactic in Section 5.3.1. For the remaining 8.2% (7.1%) of the sources, the counterparts provided in this work have high p_i. In addition, they are classified as extragalactic, an option that by construction is not considered by HamStar. For these sources, it is likely that the X-ray emission is produced by both counterparts. In our catalogue, we will provide a flag indicating whether or not the source coincides with the counterpart determined in Freund et al. (2024).

Table 3. Fraction of sources in GDR3 and CW2020 sharing the same counterpart as identified using LS10 for the inAllLS10 region (top) and the inAnyLS10 region (bottom).

DET_LIKE_0	inAllLS10	same LS10-GDR3	same LS10-CW2020
≥6	598,893	56.3%	82.5%
≥6 & $g < 20.7$	304 566	93.4%	–
≥6 & W1 < 20.4	511 222	–	89.6%
≥7	470,757	62.4%	86.3%
≥7 & $g < 20.7$	266 102	94.3%	–
≥7 & W1 < 20.4	417 013	–	91.6%
≥8	388,292	67.2%	88.8%
≥8 & $g < 20.7$	237 317	94.9%	–
≥8 & W1 < 20.4	352 289	–	93.0%
DET_LIKE_0	inAnyLS10	same LS10-GDR3	same LS10-CW2020
≥6	651,935	55.9%	81.9%
≥7	511,742	62.6%	85.7%
≥8	421,784	67.3%	88.2%

Notes. Only the first best counterpart(match_flag=1) is considered in this comparison, and no cut on p_any is applied.

Table 4. Comparison of counterpart associations obtained using GDR3 and CW2020 in the region outside the LS10 footprint.

DET_LIKE_0	only CW2020	only GDR3	same CTP
≥6	4401/220 132	827/220 132	58.1%
≥7	3034/170 244	647/170 244	63.6%
≥8	2038/137 566	541/137 566	68.6%

Notes. We report the number of sources that have a counterpart only from one supporting survey and the fraction of sources for which the counterpart is the same.

4. Consistency check with ROSAT

In Salvato et al. (2018b), the counterparts to the 135 118 ROSAT X-ray sources detected over the whole sky in the newly reanalysed data (ROSAT/2RXS Boller et al. 2016) were identified using NWAY and a prior based on AllWISE (Wright et al. 2010). The goal here is to quantify the fraction of sources with the same counterpart as in our eRASS1 catalogues. This is relevant for time domain studies, for which the ROSAT data point can help in identifying long-term variability. For the match, we have considered the 2RXS-AllWISE counterparts and the eRASS1-LS10, GDR3, and CW2020 counterparts computed here, allowing for a maximum distance of 2″. We found a match for 22 716, 25 416, and 25 364 sources, respectively. Using the classification criteria described in the next section, 67% of the sources in common with LS10 are extragalactic, while 54% of the GDR3 sources have a probability higher than 80% [14] to be extragalactic. A star and galaxy classification is not available for CW2020, but using a threshold in W1 and W2 photome-

[13] The magnitude limit of Gaia.

[14] Using PGAL>0.8 or PQSO>0.8).

try as discussed in Section 7.3, 66% can be considered extragalactic. Interestingly, in all three catalogues, at least 80% of the matched sources have `p_any` > 0.8, making the associations highly secure. Future works will have the counterparts of ROSAT/2RXS directly recomputed with the method presented in this paper. At this stage, in the catalogues that we release, two columns with the ROSAT ID from Boller et al. (2016) and AllWISE ID (from Wright et al. 2010) will indicate whether the source is also associated with the ROSAT/2RXS catalogue.

5. Galactic and extragalactic classification

In this section, we look at the multi-wavelength properties of the counterparts, intending to classify them as Galactic or extragalactic.

5.1. Classification in GDR3

For the catalogues of counterparts determined via Gaia, the user can adopt the Galactic or extragalactic classification from Gaia (`PQSO`, `PGAL`, `Pstar`, `PWD`, `Pbin` for the probability to be a QSO, a galaxy, a star, a white dwarf or a binary, respectively Gaia Collaboration 2023b). These probabilities are included in the catalogue released with this paper for the convenience of the user. When available, the Gaia redshift (either original or from Quaia[15], Storey-Fisher et al. 2024) is also reported.

5.2. Classification in CW2020

Unlike GDR3, CW2020 does not provide a star or galaxy classification. However, stars should have mid-IR colours close to zero in the Vega system (Stern et al. 2012). In Section 7.3, we further defined the cut and demonstrated that a cut at $W1–W2 < 0.3$ in the Vega system includes most of the Galactic sources with a minimal contamination from extragalactic sources. Nevertheless, there are extragalactic sources with $W1–W2 < 0.3$ as there are Galactic sources with $W1–W2 < 0.3$ (see more details in Section 7.3 and Figure therein. The users will be able to further clean the selection by using more stringent cuts, or by combining information from other catalogues.

5.3. Classification in LS10

For the counterparts determined using LS10, the Galactic or extragalactic nature of the source can be extrapolated using various criteria. S22 demonstrated that, at the X-ray depth of the eROSITA/eFEDS sample, the Galactic or extragalactic nature of X-ray sources can be predicted well from their optical/IR colours; they used a training sample of nearly 8000 spectroscopically classified X-ray sources to define an optimal boundary in g-r vs. z-$W1$ colour-colour space that most reliably separates extragalactic and Galactic sources. In particular, the cut creates a complete sample of extragalactic sources, although it includes many stars. Complementarily, the WISE vs X-ray fluxes used in Salvato et al. (2018b) were reliable in selecting stars at the cost of missing nearby extragalactic sources. For this reason, we introduce here STAREX, a more efficient machine-learning algorithm that we developed to distinguish Galactic sources from the extragalactic ones, using Legacy Survey DR10 data.

[15] Sources thought to be AGN based on their Wise colours got their redshift recomputed using a training sample of similar sources with spectra from SDSS. See more in Section 7.4.

5.3.1. STAREX

We have created a training sample of 131 000 eRASS1 sources, split almost 50/50 between Galactic and extragalactic objects. The former are defined as secure eRASS1 counterparts with PSTAR >90% in GDR3, while the extragalactic sources have a secure spectroscopic redshift z > 0.002 from DESI (DESI Collaboration 2025). For this (almost) ground truth X-ray sample, we have retrieved the LS10 catalogue information.

The optimal separation between Galactic and extragalactic sources is found with machine-learning. We tested Fisher's discriminant analysis, random forest and logistic regression. The features used for classification are an expanded set of those used in Salvato et al. (2022). This includes the AB magnitude from LS10 in the $g,r,z,$ and W1 bands, and the morphological type which is encoded as integers from 0 to 5 (`PSF, REX, EXP, DEV, COMP, SER`) by the complexity of the best-fit profiles (see Dey et al. 2019). We also included the log of X-ray flux (in erg/s/cm^2 in the 0.2–2.3 keV band as available from eROSITA). For sources in the training sample detected by Gaia, parallax, inverse variance, mean magnitude in BP, RP, and G as listed in LS10 are also included. These features are then used to train a supervised machine-learning classifier. However, not all LS10 sources have Gaia information due to different depths. To take advantage of this information when available, we train two classifiers, one appropriate for sources with Gaia information (11 input features) and another that can be applied to any X-ray LS10 source (6 input features). Their training is otherwise identical.

Training proceeded as follows: The ground truth sample was split into a training set (75%) and a test set (25%), the latter used for validation. The training-validation data is normalised to have a mean of zero and a standard deviation of one, and the normalisation information is stored for application to the test data and later application. The training-validation data is then split with K = 10 folds into training and validation data sets For each of these ten training samples, a classifier is then trained on the input features and ground truth label (galactic=0, extragalactic=1). For simplicity and interpretability, we chose logistic regression, where the probability of the extragalactic class given input feature vector x is

$$p(x) = \frac{1}{1 - e^{\alpha+\beta x}} = \frac{1}{1 - e^{-(x-\mu)/s}}.$$

The weight vector μ quantifies the importance of each feature, and the scale parameter s determines the sharpness of the probabilistic transition from one class to the other. Equivalently, the linear dependence can be quantified with intercept $\alpha = \mu/s$ and weights $\beta = 1/s$. Training then maximises the log-likelihood function

$$\log \mathcal{L}(\alpha,\beta) = w_{\rm gal} \sum_{i\in{\rm gal}} \log(1 - p(x_i)) + w_{\rm exgal} \sum_{i\in{\rm exgal}} \log p(x_i). \quad (1)$$

The best-fit parameters (intercept α, weights β) of the ten classifiers for each fold are then combined by averaging. This gives a final logistic classifier less dependent on the specific training sample. The linear separation planes ($0 = \alpha + \beta x$), defined with and without the Gaia features, are

$$0 = 0.14\,{\rm TYPE} - 2.25g + 0.96r - 1.60{\rm w1} + 1.58z - 4.54bp + 2.16G + 4.07rp + 1.17\log_{10}(F_{0.5-2\,{\rm keV}}) - 0.47\mu + 0.0\mu\_ivar + 13.73 \quad (2)$$

and

$$0 = 0.71\,{\rm TYPE} - 2.19g + 1.82r - 0.60{\rm w1} + 1.75z + 0.93\log_{10}(F_{0.5-2\,{\rm keV}}) - 0.52 \quad (3)$$

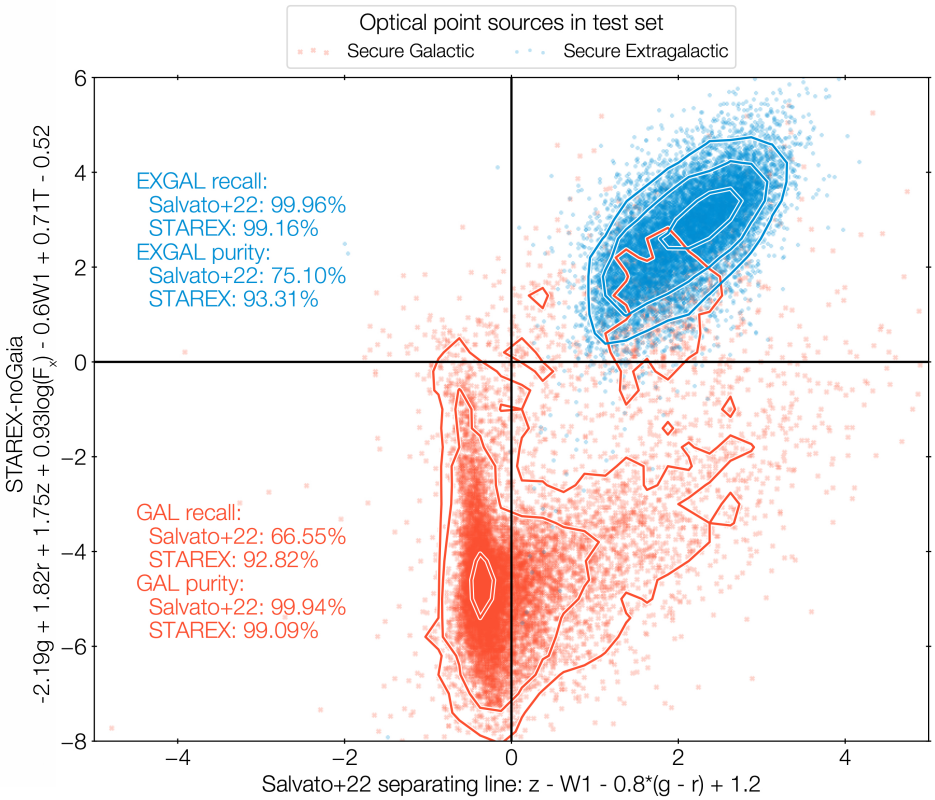

Fig. 8. Classification performance on the test sample, restricted to sources appearing point-like in LS10 images. The horizontal axis shows the distance from the S22 photometric classification criterion (with >0 indicating extragalactic and <0 Galactic classification). The vertical axis shows our new separating line, for the case where no Gaia information is available (STAREX-noGaia), using the same magnitudes but also the optical type (0 here for point-like) and the soft-band X-ray flux. The recall for secure extragalactic (blue) and Galactic (red) test sources is reported in the left panels, followed by the corresponding purity values. The contours enclose 25%, 75% and 90% of the sources.

Table 5. Source classification performance on the test set.

Subset	metric	S22	STAREX-noGaia
Point sources	Recall exgal	100.00%	99.10%
	Purity exgal	69.00%	**95.64%**
	Recall gal	67.67%	**96.75%**
	Purity gal	100.00%	99.34%
All sources	Recall exgal	99.97%	99.13%
	Purity exgal	71.84%	**95.37%**
	Recall gal	67.03%	**97.34%**
	Purity gal	99.96%	99.24%
w. Gaia info.	Metric	S22	STAREX-Gaia
Point sources	Recall exgal	99.96%	99.16%
	Purity exgal	75.10%	**93.31%**
	Recall gal	66.55%	**92.82%**
	Purity gal	99.94%	99.09%
All sources	Recall exgal	99.72%	**96.60%**
	Purity exgal	77.91%	**91.39%**
	Recall gal	63.31%	**93.31%**
	Purity gal	**99.44%**	95.24%

where TYPE corresponds to the morphological type (0 to 5).

Figure 8 illustrates the performance of our new separating line for the case without Gaia information (which is the case for most of the eRASS1 sources) on the vertical axis. For comparison, the horizontal axis of Figure 8 shows the S22 separating criterion. As already shown in the original paper, the distribution of the Galactic sources, in red, extends into the extragalactic region. These sources are the reason for a modest recall of Galactic sources. All statistics comparing S22 and STAREX are listed in Table 5, including the recall (selection completeness) and purity, for optical point-sources and all sources. The separation with STAREX is better, with high recall and sample purity both above 95%. For extragalactic sources, the recall is essentially complete with both methods (>99%), but STAREX increases the purity from 69% to 95%. The performance of STAREX further improves when Gaia information is available and used[16].

5.3.2. Final Galactic or extragalactic classification for LS10 sources

In addition to STAREX or the line separators introduced S22 and Salvato et al. (2018b), we also use the information on parallax (μ) and proper motion (π) provided by Gaia alone to identify secure Galactic sources. In particular, we consider

$$SNR\_\pi = \pi * sqrt(\pi\_IVAR). \quad (4)$$

$$SNR\_\mu = sqrt(\mu_{\rm RA}^2 \times (\mu_{\rm RA}\_IVAR) + (\mu_{\rm DEC}^2 \times \mu_{\rm DEC}\_IVAR)). \quad (5)$$

The same two quantities, computed for the 206k sources in the DR16Q catalogue (Lyke et al. 2020) detected by Gaia and with

[16] Note that the samples used on the top and bottom of Table 5 are different.

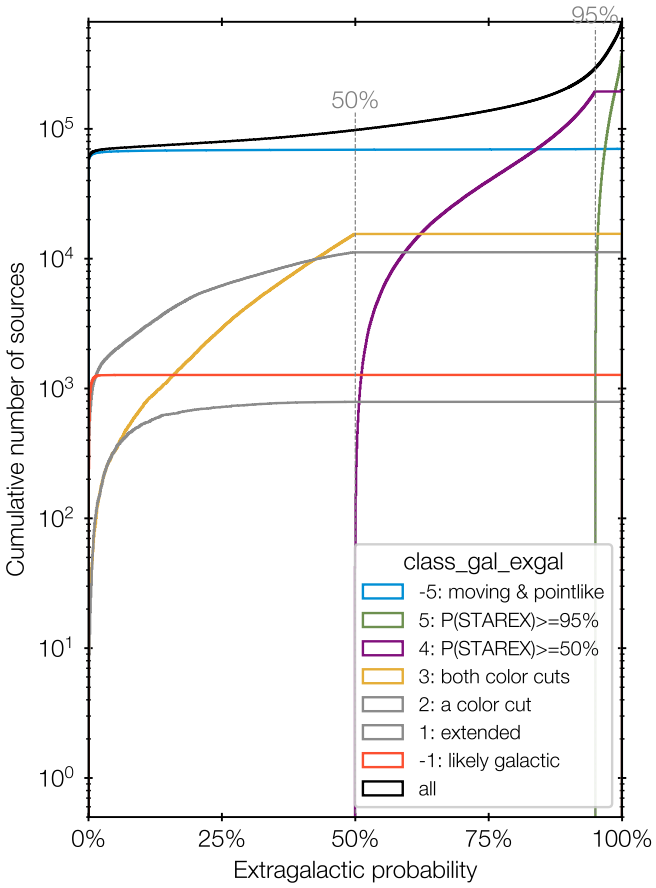

Fig. 9. Cumulative distribution of sources within the LS10 footprint sorted in order of reliability of the star and galaxy classification. Sources with a STAREX probability to be extragalactic below 50% are further split into classes (from 3 to −1) using a combination of STAREX output and diagnostic colour-colour diagrams (grzW1 and W1X) as described in 5.3.2.

visually inspected redshifts, show that 99.95% of the sources have either `SNR_parallax` > 5 or `SNR_PM` > 5. We use these thresholds combined with `TYPE`=PSF to select secure Galactic sources based on their motions.

There are cases where STAREX, the line separators from *grzW1* and/or W1X, and the proper motion and parallax from Gaia provide conflicting Galactic or extragalactic classification. For this reason, we introduced the quantity `class_gal_exgal`, which provides information on the reliability of the classification.

With `class_gal_exgal` = −5 we identify sources that are classified as Galactic because they move and are point-like (`TYPE`=PSF). Sources with a STAREX probability of being extragalactic above 95% (50%) have `class_gal_exgal`= 5(4). For STAREX probability lower than 50% we assign `class_gal_exgal` = 3(2) to the sources classified as extragalactic based on both (at least one) line separators. Finally, we assigned `class_gal_exgal` = −1 to the remaining sources. To summarise, negative numbers indicate stars, positive numbers indicate AGN, and higher absolute values indicate higher confidence.

To double-check our final classification, we first looked at the internal consistency. Only 191 sources in the area within `inAllLS10` are considered Galactic (`class_gal_exgal` = −5), but STAREX would have classified them as extragalactic. It is known that compact objects, such as Cataclysmic variables (CVs), occupy the same locus of AGN in many parameter spaces, such as those used in our classification. In fact, the large majority of these sources are either already classified or are candidate CVs following Schwope et al. (2024b), with which we have a 98.6% agreement in the identification of the counterparts. An additional test was done, checking the spectroscopic redshifts from DESI DR1 (DESI Collaboration 2025) when available. There are about 58 000 eRASS1 sources in the footprint of LS10 with a DESI redshift. Among the 86 599 eRASS1 sources classified as Galactic within the footprint of LS10, 1274 have a reliable redshift from DESI, out of which 79 have redshift z > 0.002. On the other hand, there are 569 778 eRASS1 sources classified as extragalactic, 57 162 with redshift from DESI, with only 114 of them having z < 0.002. A visual inspection of these cases reveals, for the most part, sources very close to a saturated star to which the DESI spectrum is probably related.

Figure 9 shows the breakdown of the best LS10 counterpart to the eRASS1 point sources (after removing those flagged in Merloni et al. 2024) in the various classes, as a function of the probability to be extragalactic from STAREX. To further subclassify the sample, we add further information. Firstly, we add the column `class_jetted`, which indicates whether the LS10 position is in proximity to a blazar (jetted=1) in the BzCAT (Massaro et al. 2015), a flat-spectrum radio source (jetted=2) in CRATES (Healey et al. 2007), or a blazar in Simbad (jetted=3), or 0 otherwise. We added the column `simbad_known_galactic` to report whether a Simbad positional match within 1″ identified a source with main type X-ray binary, cataclysmic variable, High-mass or Low-mass X-ray binary.

6. Redshifts

For all the eRASS1 sources classified as extragalactic in the entire LS10 footprint, we also provide redshifts or redshift estimates. First, we used the LS10 coordinates to perform a match (maximum distance of 1″) to a large compilation of spectroscopic redshifts taken from the literature (Igo et al., in prep., Kluge et al. 2024), including, among others: Quaia (Storey-Fisher et al. 2024), 2dF QSO catalogue (Croom et al. 2009), 6dFGS (Jones et al. 2009) and the latest release of DESI DR1 (DESI Collaboration 2025). The compilation has many astrophysical objects with more than one redshift reported in the literature, with redshift values in disagreement. This is often due to the inclusion of lower-quality spectra or due to differences in the way redshifts have been derived from the spectra. We decided to remove all those sources in the compilation with redshifts in disagreement larger than 0.1. This is done for two reasons. First, we are not in a position to decide which redshift is the correct one. Secondly, we use the spectroscopic sample to create a clean reference sample to measure the fraction of outliers and the accuracy for the photometric redshifts that we computed with the algorithm CIRCLEZ (Saxena et al. 2024) for the sources without spectroscopic redshift. In addition to the cleaned public compilation, we added new 137 redshifts acquired by eROSITA members over the years with various instruments at different telescopes and 217 redshifts that are provided by eROSITA collaborators from ongoing ESO campaigns (Scialpi et al., in prep). In total 196,083 LS10 counterparts to eRASS1 have a spectroscopic redshift. For the rest of the sources, we rely on photometric redshifts.

Deriving photo-z for AGN is notoriously difficult, due to the a priori unknown host/active nucleus contribution at a given

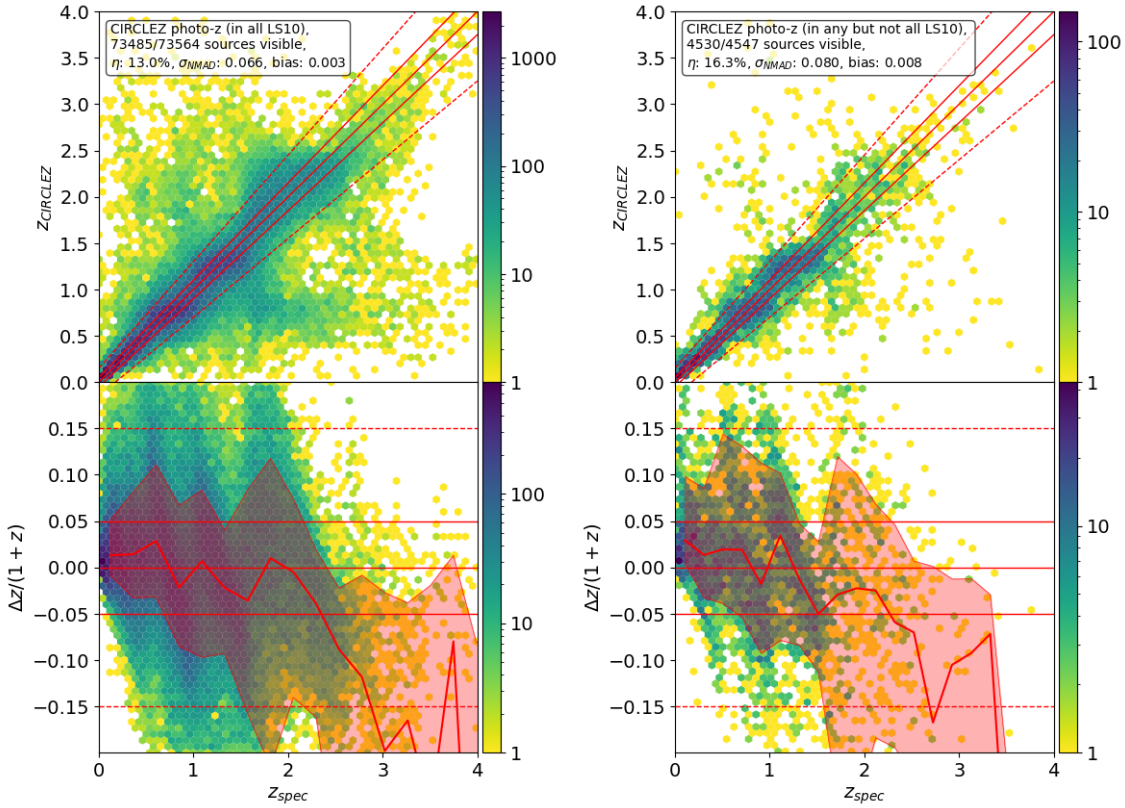

Fig. 10. Comparison between photometric and spectroscopic redshifts for sources within (left panels) and outside (right panels) the inAllLS10 area. In the top panel, the lines represent the one-to-one relation and the thresholds adopted to define outliers (see text and footnotes). The bottom panel shows the normalised residuals $(z_{\rm phot} - z_{\rm spec})/(1 + z_{\rm spec})$. The red lines in the lower panels indicate the moving medians and the 1 sigma region, revealing a negative bias above redshift 2. See main text for more details.

wavelength. Classical photometric redshift techniques use the total fluxes of a source and explicitly (SED fitting) or implicitly (machine-learning) the colour-redshift relation characteristic of a source to determine the redshift (see Salvato et al. 2018a, for a review). CIRCLEZ is a machine-learning algorithm based on a fully connected neural network (FCNN) that estimates the redshift using as features, not only the total fluxes, but also the fluxes and the residuals after subtracting the best fitting morphological type[17], in addition to the aperture photometry and colours within apertures. Here we used the same model that was defined in Saxena et al. (2024) using a sample of 14 000 X-ray detected extragalactic sources with secure spectroscopic redshift. A direct comparison between the spectroscopic and photometric redshift is shown in Figure 10. For the 74 089 sources in the `inAllLS10` with spectroscopic redshifts, we obtained an accuracy[18] of 0.067, with only 12.7% outliers fraction[19]. Worth noticing is the fact that the bulk of the outliers are close to the 0.15 separation line used in literature, but arbitrarily defined. The accuracy and the fraction of outliers are very similar at any detection likelihood, and it is superior to the one reached in eROSITA/eFEDS (S22), thanks to the new approach implemented in CIRCLEZ. In fact, for the sources outside the `inAllLS10` area (i.e. the region with shallower LS10 data), the accuracy (0.078) and bias (0.001) are very similar to those with the best photometry, and only the fraction of outliers (15.4%) increases slightly. In addition, many of the sources at $z_{\rm phot} > 3$, are outliers, due to their faintness and larger photometric errors. To interested readers, this is discussed in detail in Saxena et al. (2024) and Euclid Collaboration (2025). For these types of sources, rather than the photometric redshift, the redshift distribution function (PDZs) should be used. Unlike PDZs from SED fitting that are too optimistic (i.e. producing errors that are too narrow also for sources that are known outliers, (e.g. Brescia et al. 2019)), the PDZs obtained with CIRCLEZ are realistic (see Saxena et al. 2024). The PDZs for eRASS1 are available upon request. Only for 454 (about 0.008%) of the sources classified as extragalactic in LS10, the photometric redshift could not be constrained due to a lack of sufficient information (i.e. low number of photometric points).

7. Defining eRASS1 AGN samples

We release with this paper the counterparts to all point sources in the main catalogue of eRASS1, identified using LS10, GDR3, and CW2020, respectively (see next section). The users will then be free to select their subsamples of sources based on any (combination of) criteria (e.g. completeness, purity, depth/quality of X-ray or ancillary data).

The more basic sample of AGN could be generated by simply selecting all the sources that are classified as extragalactic. However, this would create an inhomogeneous sample of AGN in LS10, given that the depth of the ancillary data varies over the sky. Alternatively, controlled AGN samples can be generated by accounting for the quality of the ancillary data and the reliability

[17] Described in https://www.legacysurvey.org/dr10/description/

[18] $\sigma_{NMAD} = 1.48 * {\rm median}(\Delta z/(1 + z_{\rm spec}))$.

[19] Defined as fraction of sources η with $|z_{\rm phot} - z_{\rm spec}|/(1 + z_{\rm spec}) > 0.15$

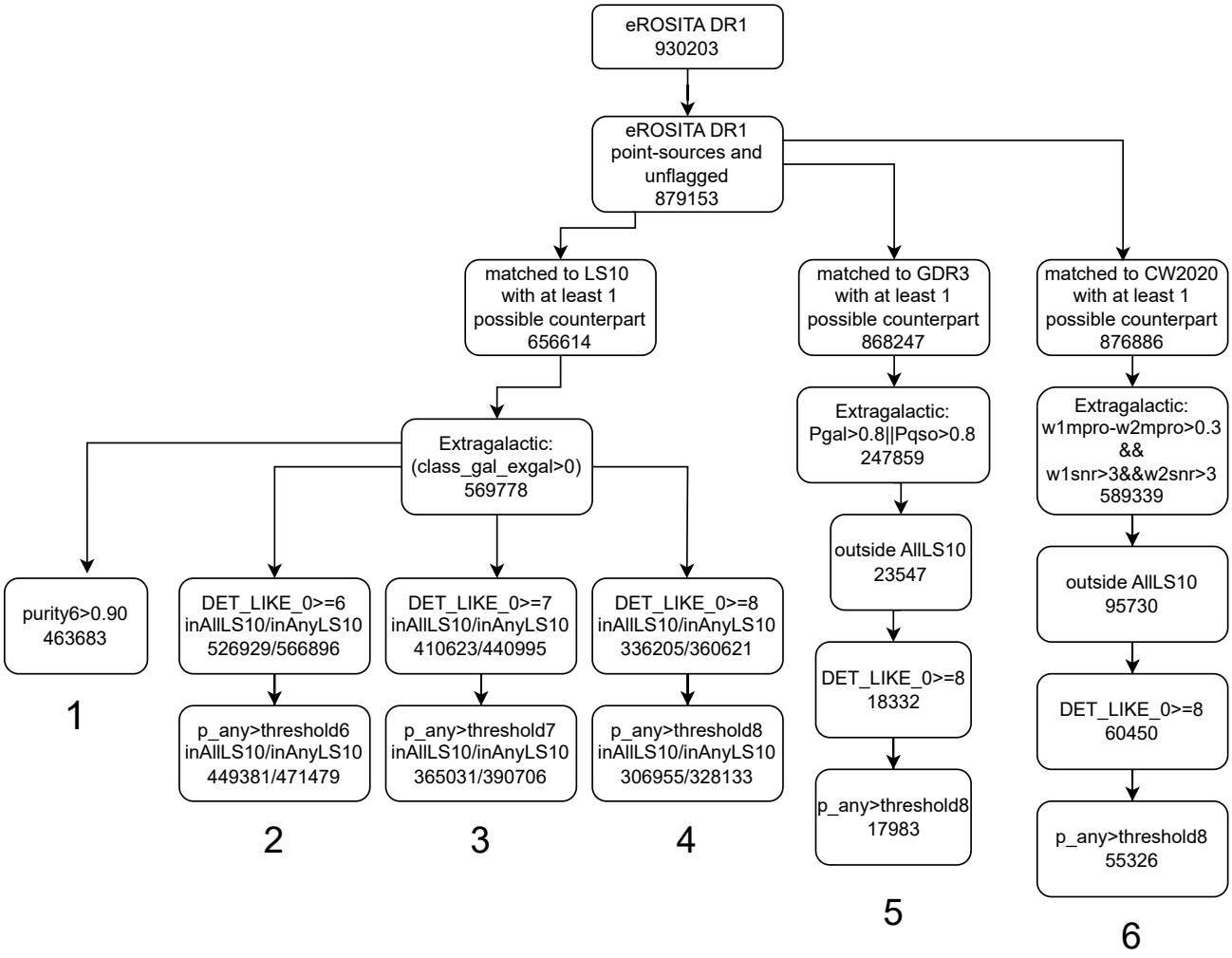

Fig. 11. Flowchart describing the construction of six samples of candidate AGN with different supporting data, on different regions and with different levels of completeness and purity. For the detailed construction of the samples, see Section 7.

of the X-ray detection. In this section, we follow the second path and create controlled samples of AGN using the counterparts from LS10, GDR3, and CW2020 to understand the complementarity of X-ray-selected AGN with other selection methods.

As an example, we define 6 high-purity samples following the steps listed below and represented in Figure 11, together with the number of sources remaining after each step. We describe here the construction of samples 1 (purity level higher than 90%, everywhere in LS10), 5 (AGN selected outside the footprint of LS10 using GDR3, with the same completeness and purity, for eRASS1 sources with `DET_LIKE_0` > 8) and 6 (same as 5 but using cw2020). Samples 2, 3, and 4 (using LS10, with same completeness and purity, for or eRASS1 sources with `DET_LIKE_0` > 6,7,8]) are described in Appendix B. We will then compare Sample 1 of AGN with those selected using AllWISE, Gaia, and a combination of WISE and Gaia.

7.1. eROSITA AGN within LS10

We start to construct our AGN catalogue from the point-like (`EXT_LIKE`=0), unflagged (`FLAG_*`=0, see Merloni et al. 2024) in the X-ray catalogue and with the counterpart from LS10 classified as extragalactic (`class_gal_exgal` > 0). A first selection (Sample 1 in Figure 11) could be applied by taking into account the spatial dependence of the purity of crossmatches (see Section 3.3) by applying a local `p_any` threshold (set for each eROSITA tile), such that the purity[6] is greater than > 90%. In such a sample, the population would differ from tile to tile, depending on the depth of both the eRASS1 and the LS10 available there. We call this sample DR1_LS10_EXGAL_DET6P90.

Figure 12 shows the X-ray luminosity and redshift distribution for the sample DR1_LS10_exgal_DET6P90, where photometric redshifts are used for the 383 349 sources without spectroscopic redshift (75 496; see Section 6). The figure suggests that a limited fraction of sources in the sample are at luminosity lower than 10^{41} erg s^{-1}, a typical threshold adopted for separating galaxies powered by AGN or star formation (e.g. Lehmer et al. 2012; Brightman & Nandra 2011; Bauer et al. 2004). Only a detailed analysis will determine the dominant contribution of these sources.

7.2. eRASS1 AGN in the Galactic plane using GDR3

The scheme used for the construction of this sample is represented in Figure 11, following the path for Sample 5. Because the prior adopted for determining the counterparts using GDR3

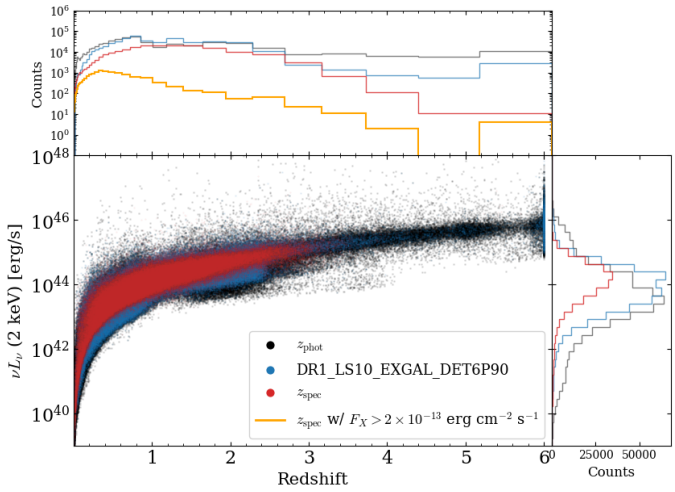

Fig. 12. X-ray (2keV) luminosity-redshift distribution for eRASS1 sources within the LS10 footprint, compared to the same distribution from the high purity and reliability AGN sample DR1_LS10_EXGAL_DET6P90 (blue). Symbols indicate the redshift: spectroscopic (red) and photometric (black). The apparent excess of sources at z = 6 is attributed to unreliable photometric redshift estimates (see Section 6). For completeness, the histogram distributions for the redshift (top) and the X-ray luminosity (right) are also shown. In the redshift distribution we also show, in orange, the spectroscopic distribution for sources at the depth of ROSAT ($F_{0.5-2\,\mathrm{keV}} > 2 \times 10^{-13}$ erg s^{-1} cm^{-2}).

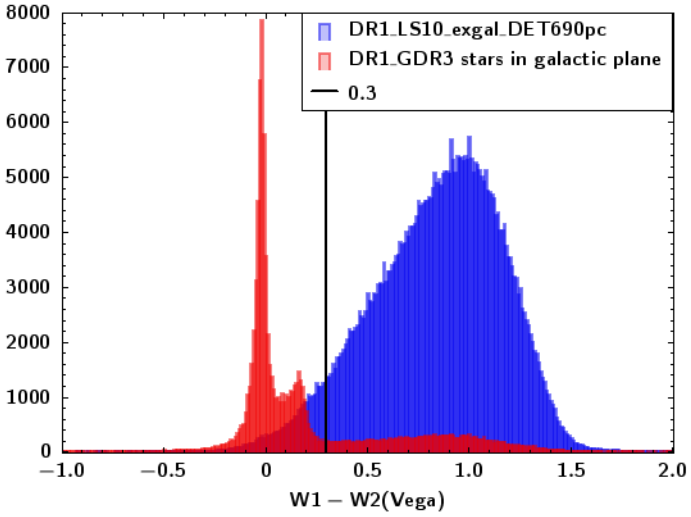

Fig. 13. CW2020 W1−W2 (Vega) colour distribution for the extragalactic sources in the clean extragalactic sample within LS10 (blue) and a clean sample of Galactic sources in the Galactic plane from the GDR3 counterparts to eRASS1. The vertical line indicates the cut to apply to define a sample of extragalactic sources among the CW2020 counterparts to eRASS1, least contaminated by Galactic sources.

is less reliable than for LS10, we decided to consider only sources with DET_LIKE_0 ≥ 8, so that the positional errors are typically smaller, limiting the area search for the correct counterpart. For the area outside LS10 (inAnyLS10=0), we consider as extragalactic the sources that have a probability of being a galaxy (Pgal) or a QSO (Pqso) larger than 80% from GDR3. Above the threshold that identifies the p_any point at which purity and completeness are maximal, there are 17,983 sources, mostly on the Galactic plane. The mean value of purity and completeness of the sample at the threshold8 (see Section 3) values is 74.4%. We will call this sample DR1_GDR3_EXGAL_GP_DET8CP74.

7.3. eRASS1 AGN in the Galactic plane using CW2020

The scheme used for the construction of this sample is represented in Figure 11, following the path for Sample 6. The AllWise W1−W2[20] colour cut at 0.7 (e.g. Assef et al. 2013, 2018) is usually adopted for selecting AGN in the Mid-infrared. However, the cut does not fully account for the fact that the instrument has a coarse angular resolution (6.1″in W1 and 6.4″in W2, with an image pixel scale of 1.375″) and that the source detection had limited deblending capability. Thus, more than one source could be contributing to the mid-infrared flux. This is even more true at the depth of CW2020 (0.8 and 1.6 magnitude deeper than AllWISE in W1 and W2, respectively), and in the Galactic plane, where the number of sources reaches the confusion limit.

To define a W1−W2 cut based on CW2020 photometry, able to separate candidate AGN among the CW2020 sources associated with eRASS1, we matched the clean sample of extragalactic sources from LS10, DR1_LS10_exgal_DET6P90 to CW2020 with a radius search of 3″, considering only those with a unique counterpart (92% of the sample). We repeated the procedure with a clean sample of stars (PSS > 0.9) in the Galactic plane from the DR1_GDR3 sample. The distribution of the sources in W1−W2 is shown in Figure 13. A cut at W1−W2 > 0.3 provides an extragalactic sample that is 95% complete (only 5% of the DR1_LS0_exgal_DET6P90 sample have W1−W2 < 0.3. However, the sample is not pure as 25% of the stars in DR1_GDR3 have W1−W2 > 0.3. The CW2020 counterparts to eRASS1 with W1−W2>0.3, with SNR in W1 and W2 larger than 3, outside the LS10 area, with DET_LIKE_0 > 8 and with p_any above the threshold are 55 326 and will be considered extragalactic. The mean value of maximum purity and completeness reached at the threshold for DET_LIKE_0 > 8 is 82.8%. We will call this sample DR1_CW2020_EXGAL_GP_DET8CP83.

Figure 14 shows the distribution of the three AGN samples in the sky: compared to the shallow GDR3, CW2020 does allow great access to low Galactic latitudes, but large gaps in coverage are evident, also due to source confusion.

7.4. Comparison between eRASS1 AGN within the LS10 footprint and those selected in Gaia, Quaia, and AllWise AGN

Now that we have defined clean subsamples of eRASS1 extragalactic sources that are bona fide AGN, selected in the area with the best data quality, we can compare their number density and basic properties with the AGN identified at other wavelengths, such as in the mid-infrared (with AllWISE, Assef et al. 2018), and in the optical (Gaia Gaia Collaboration 2023b) and in mid-infrared and optical simultaneously (with Quaia, Storey-Fisher et al. 2024).

The AllWISE sample of candidates AGN (R90) relies on a cut in W1−W2 colour determined empirically using a sample of spectroscopically confirmed, optically bright AGN originally detected in the IRAC channels Ch1 and Ch2 (very similar to W1 and W2 in WISE; Stern et al. 2005, 2012). The mid-infrared selection is less biased towards obscured AGN, e.g. those presenting narrow emission lines in the optical. However, at increasing depth of the WISE data, the fraction of star-forming galaxies

[20] Vega system.

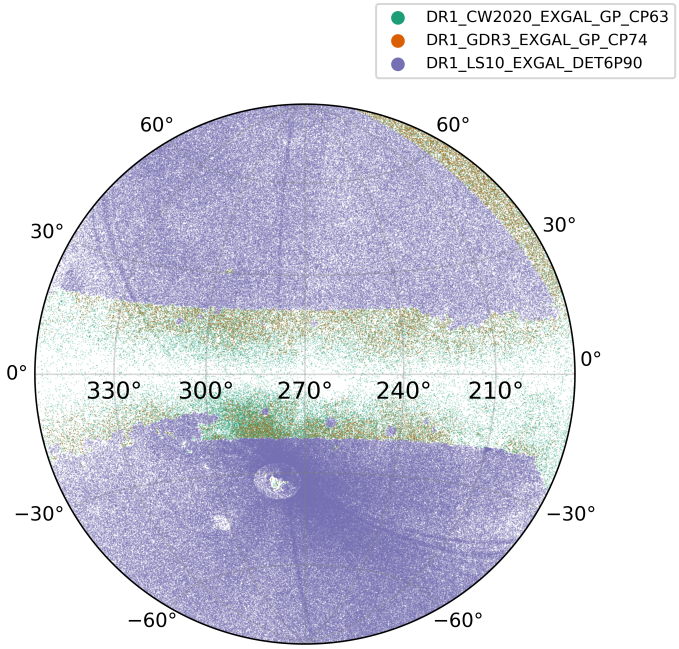

Fig. 14. Sky distribution of AGN Samples 1 (LS10; purple), 5 (GDR3; orange), and 6 (CW2020; green). Per constructions, the AGN Samples from GDR3 and CW2020 occupy the area outside the footprint of LS10.

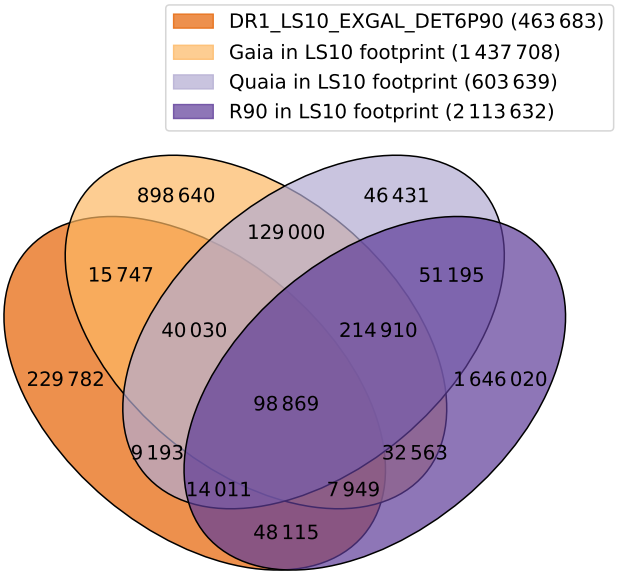

Fig. 15. Venn diagram distribution for the AGN selected via eROSITA and LS10 (Sample 1), GDR3 high-reliability QSOs, AllWise/R90 and Quaia within the footprint of LS10. In the legend, the number of sources in a given sample, within the footprint of LS10, is reported. In all cases, at least 50% of the candidate AGN are unique to that selection.

increases, making the selection less pure. Assef et al. (2018) provides four catalogues with completeness (C) and reliability (R) at 75% or 90%. For the purpose of the comparison, we are interested in purity more than in completeness, and for this reason, we selected the R90 catalogue. In the footprint of LS10, the R90 catalogue includes 2,113,681 sources.

We consider as Gaia AGN sample those listed in the catalogue of Gaia QSOs with a probability to be QSO (`Pqso`) higher than 80% (see Section 2.2.2). This is because the original sample is highly complete but with a purity of only 52% (Gaia Collaboration 2023a). The sample includes ∼2/3 of the original 6.6M sources, further limited to 1 437 708 when considering only those in the footprints of LS10.

The Quaia AGN sample is formed by the subsample of Gaia QSO (regardless of the value of `Pqso`) that is also selected in unWISE. The catalogues include 755 850 quasar candidates with G < 20.0 and 1 295 502 with G < 20.5. For these sources, the redshifts are determined on GAIA spectra using machine-learning trained on similar sources with spectra from SDSS. The Quaia sample in the footprint of LS10 lists 603 639 sources.

Figure 15 shows the Venn diagram of the AGN distribution in the four catalogues. Despite the large size of the samples analysed, the overlap remains minimal, with only 98 773 sources common to all selections. It is interesting then to understand what the properties of the sources in common are and the properties of about 50% of the eROSITA DR1 sources that are unique detections[21]. For that, it is important to keep in mind that the Gaia selection, by construction, retains only sources that are best fit by a Gaussian profile, so preferentially point sources or very compact objects. Thus, there is a strong bias towards QSO or the core of local Seyfert 1 galaxies, with a strong contrast between the active nucleus and the host. The bias is retained in the Quaia selection, which requires a detection in unWISE in addition to WISE. Neither the Gaia QSO sample nor Quaia will be able to identify galaxies hosting a low luminosity AGN or at low redshift. In addition, given the shallowness of the data, faint QSOs will not be detected. The AllWISE selection of AGN can also detect sources at lower redshift (regardless of the morphology) or faint sources at higher redshift, but only if dust is heated up sufficiently by the nucleus, assuming dust is actually present. Thus, AllWISE/R90 will not contain low luminosity AGN nor those hosted in early-type galaxies, as they lack dust to be heated up.

All this is confirmed in Figure 16 and Figure 17. Figure 16 shows the eROSITA/DR1 extragalactic sources that are either detected also by other surveys, or unique, plotted on the basis of they *rgzW1* colours, as in S22, together with the colour-colour tracks of a few basic SEDs (from Polletta et al. 2007; Salvato et al. 2009), as a function of redshift. The dominance of QSOs and spiral galaxies among the sources in common between eROSITA and the other surveys is clear. The additional QSOs that only eROSITA detects will soon be identified by surveys as LSST (Ivezić et al. 2019), using faint photometry or variability (LSST, ZTF QSO sample; Nakoneczny et al. 2025). Instead, the early-type galaxies hosting an AGN (red and green tracks in the figure) will remain the prerogative of X-ray surveys also in the future, although it will not be surprising if a fraction of the sources will also be detected with the currently ongoing radio surveys (see e.g. Igo et al. 2024; Bulbul et al. 2022, for the comparison between LOFAR and eROSITA/eFEDS). Figure 17 shows the histogram of the redshift distribution of the extragalactic eROSITA DR1 sources, split between those detected also by Gaia, Quaia, and AllWISE/R90 or unique to eROSITA/DR1. As discussed above, more than half of the eROSITA-only detected sources (65%) are indeed at redshift z < 1.

[21] A similar analysis can be done for the 2/3 of the sources that are unique to the AllWISE and Gaia, respectively, but it is not the focus of this paper.

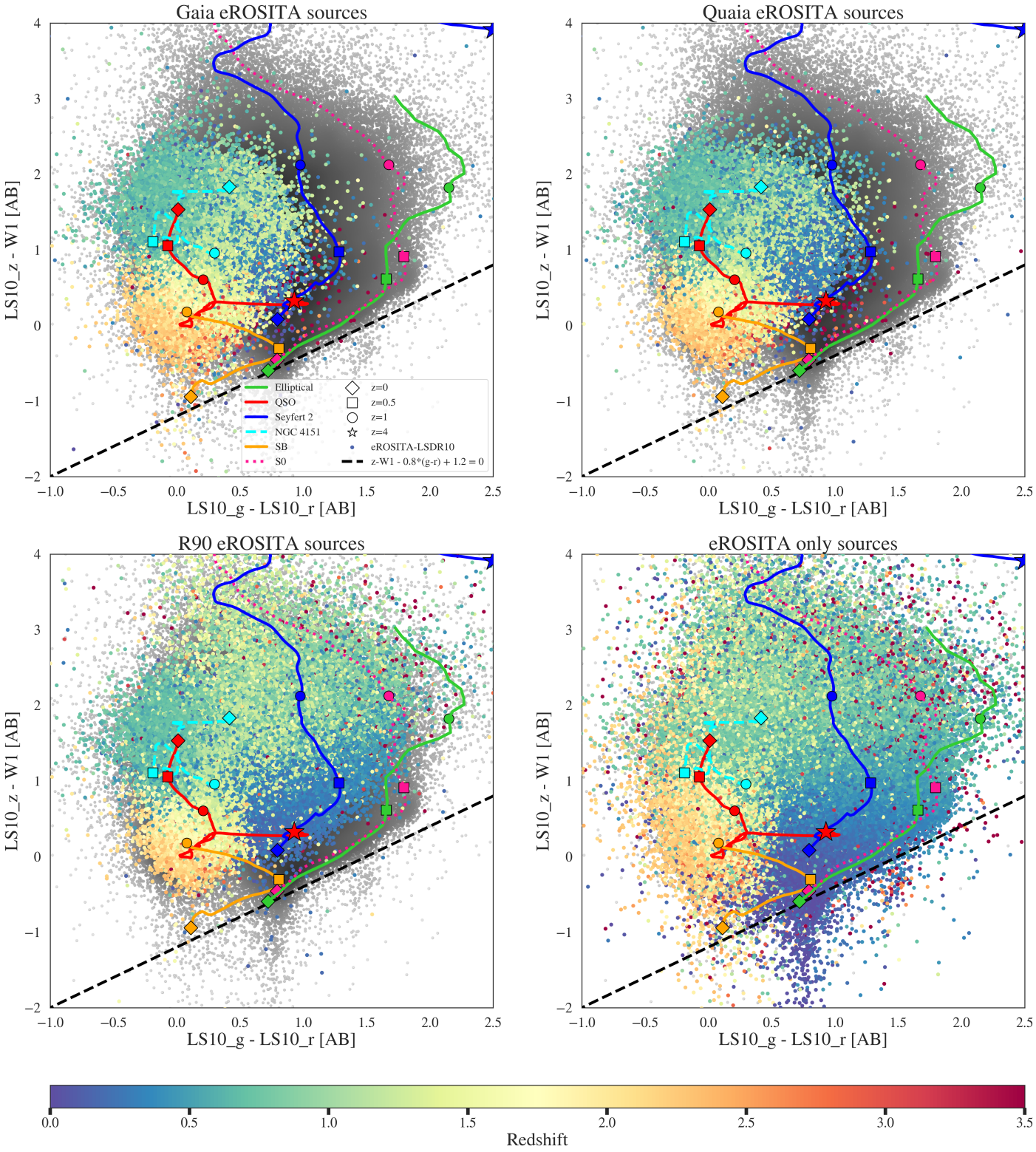

Fig. 16. *g-r* vs *z*-W1 colour-colour plots for the eROSITA sample DR1_LS10_EXGAL_DET6P90 eROSITA DR1 sources (grey), detected also in Gaia (top left), Quaia (right left), and AllWISE/R90 (bottom left), respectively. The bottom right panel shows the distribution of the sources that are unique to eROSITA DR1. The sources in all panels are colour-coded as a function of redshift. Overlapped are the tracks of a few examples of typical SEDs, which show that many of the sources that are detected only by eROSITA are early-type and Seyfert 2 at redshift below 1.

8. Catalogues format and data access

With this paper, we release the three catalogues of counterparts from LS10, GDR3, and CW2020 for the point sources in the Main catalogue of eRASS1 (Merloni et al. 2024). They will be made available via Vizier and at the eROSITA website[22] (files 4–12), together with the data model, and the description of the

[22] https://erosita.mpe.mpg.de/dr1/AllSkySurveyData_dr1/Catalogues_dr1/

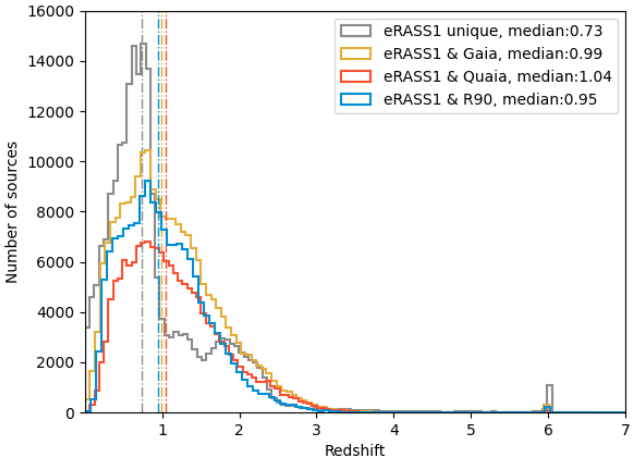

Fig. 17. Redshift distribution of the eROSITA/DR1 extragalactic sources in the footprint of LS10, split among the sources that are also detected in Gaia, Quaia, and AllWISE/R90, respectively. The majority of the sources that are unique to eRASS1 are at low redshift, where the contribution from the host galaxy dominates the SED in optical and NIR.

columns. The catalogues include basic columns from the X-ray catalogue and all the relevant columns from the supporting optical/IR catalogues, together with the basic information from NWAY like `p_any`, `p_i`, etc. In addition, we also provide `p_any` threshold[6,7,8] at which the completeness and purity curves cross, thus being at the maximum value in that specific eROSITA tile, for a specific `DET_LIKE_0`, that value being `compur`[6,7,8]. We also provide the value of completeness and purity (completeness[6,7.8] and purity[6,7,8]) corresponding to the `p_any` of the source, computed using all the sources in that specific eROSITA tile at the specific `DET_LIKE_0`. In this way, we enable the user to create a sample of sources with a specific completeness and purity over the entire sky. For example, A user could be interested in selecting all the sources with `DET_LIKE_0`>7 and purity7 > 0.9. This possibility is relevant for studies in which complete control over biases and selection effects is relevant, like in, for example, luminosity function studies. A Jupyter notebook is also provided via Zenodo at `https://zenodo.org/records/17404798`, with instructions to recreate samples 1–6 following the flow chart of Figure 11, which can be modified and used to create any other sample.

The priors for the determination of the counterparts, their classification in STAREX, and the model used for the photometric redshifts, as followed in Saxena et al. (2024), are based on a well-constructed set of training samples, which are also released with this paper. The constructions of these six catalogues are described in the Appendix A. In short, the training samples are from 4XMM (Webb et al. 2020) and CSC2 (Evans et al. 2024) that have counterparts in LS10, GDR3, and CW2020 (label 1), together with the field sources, i.e. the LS10/GDR3/CW2020 sources that are within 60″ from the 4XMM/CSC2 position and that are not associated to the X-ray emission (label 0).

9. Conclusions

In this paper, we describe the construction of the three catalogues listing the most reliable counterparts to the point sources detected by eROSITA (PSF Half Energy Width of ∼30″ and corresponding 1σ median positional uncertainty of ∼4.5″and a uniform flux limit at 50% completeness of $F_{0.5-2\,\mathrm{keV}} > 5 \times 10^{-14}$ erg s^{-1} cm^{-2}) in the first all-sky pass over the western hemisphere (eROSITA_DE), using LS10, Gaia DR3, and CW2020, respectively. Each ancillary catalogue (see Section 2.2) offers specific advantages and limitations depending on its depth, number of photometric bands, and wavelength coverage. Counterpart identification (Section 3) was performed using the `NWAY` algorithm, enhanced with a prior constructed from a Random Forest classifier trained to estimate the probability that a given source is an X-ray emitter based on its SED. The training sample consisted of securely identified X-ray emitters from XMM and Chandra observations. The most reliable counterparts are found within the LS10 footprint, given the depth and the broad multi-band coverage, which provides a rich set of features for defining the SED. Nevertheless, by applying appropriate probability thresholds, the Gaia and CW2020-based matches also yield reliable counterparts, especially in the regions not covered by LS10.

After identifying the counterparts, a source needs to be classified as Galactic or extragalactic (see Section 5). For Gaia, we adopt the classification probabilities (galaxy, star, quasar) provided by the Gaia collaboration, derived from a combination of five methods (Delchambre et al. 2023a). These probabilities allow users to apply their own selection criteria based on appropriate cuts. For CW2020, we use a simple cut in W1−W2 colour, calibrated by comparing the colour distributions of confirmed extragalactic and Galactic sources. For LS10, we have considered a combination of methods. First, we identify likely stellar sources based on Gaia detections with high proper motion or parallax and point-like morphology. Those are considered of stellar origin. Then, a machine-learning based algorithm, STAREX, was developed within the eROSITA collaboration. This algorithm combines LS10 photometry and X-ray flux information to assign a probability of the source being extragalactic, using a training set composed of securely identified Galactic and extragalactic X-ray emitters. Finally, for the sources with STAREX extragalactic probabilities below 50%, apply additional diagnostics based on colour-colour and magnitude–X-ray flux diagrams, following the methods described in Salvato et al. (2018b, 2022), as well as morphological classification from LS10. Following this approach, we classify 86,836 sources in the LS10 footprint as stars and 569 778 as extragalactic.

As the final step, we assign redshifts to the extragalactic sources (see Sect. 6). First, we compute the photometric redshifts for all extragalactic sources detected in LS10, using the CIRCLEZ code (Saxena et al. 2024). In addition, we matched the sources with a compilation of publicly available spectroscopic redshifts, allowing us to assess the quality of our photometric redshifts. In regions with high-quality photometry across all bands, we achieve an accuracy of 0.067 and an outlier fraction of 12.7%. Outside these areas, the corresponding values are 0.078 and 15%, respectively. For the eRASS1_GDR3 we provide the redshift from Gaia or Quaia, if available. We also cross-matched the sample with the redshift compilation and with our photometric redshift. For the eRASS1_CW2020 sample, we cross-matched the sources with the same compilation and with the computed photometric redshifts. In all cases, the user may request from the authors the redshift probability distribution function (PDZ) as computed from CIRCLEZ.

All three catalogues enable the construction of well-defined samples of X-ray selected sources over the entire sky, which can be compared to samples of similar sources selected using different techniques or at other wavelengths. In the paper (see Section 7), we demonstrate how to construct six AGN samples with varying levels of completeness and purity, both over the extragalactic sky and within the Galactic plane. Focusing on a

sample of 463,683 AGN within the LS10 footprint, with an estimated purity of 90%, we then compare its overlap with three other AGN samples; a Gaia-selected sample, which preferentially identifies the cores of Seyfert 1 galaxies and QSOs; a sample combining Gaia and unWISE photometry, spectroscopically confirmed via Gaia spectroscopy; a sample based on AllWISE photometry, with a purity of 90%.

In the three cases above, purity refers to the reliability of the sources classified as AGN. However, for AGN detected in the eRASS1 there is a 14% probability of being a spurious X-ray detection. We showed that the counterparts to spurious detections typically have very low p_any, comparable to those found in random catalogues.. Applying a threshold of at `purity6` > 0.9 effectively eliminates such spurious detections. The overlap among the four catalogues is marginal, confirming once more that a complete census of AGN requires the combination of multiple selection techniques. Many sources falling in the *grz*W1 locus of QSO are not detected by Gaia due to its limited depth. Meanwhile, a significant fraction of AGN detected only by eROSITA with LS10 counterparts are characterised by being at low redshift, with the emission diluted/hidden by the emission of the host galaxy, or they reside in galaxies with minimal dust reprocessing, which could reveal the presence of the AGN in the mid-infrared.

Data availability

A copy of the catalog is available at the CDS via https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/704/A344

All catalogues constructed in this paper are released to the public (see Section 8) via Vizier, Zenodo at https://zenodo.org/records/17404798 and the eROSITA web pages at https://erosita.mpe.mpg.de/dr1/AllSkySurveyData_dr1/Catalogues_dr1/ (files 4-12).

References

Assef, R. J., Stern, D., Kochanek, C. S., et al. 2013, ApJ, 772, 26
Assef, R. J., Stern, D., Noirot, G., et al. 2018, ApJS, 234, 23
Aydar, C., Merloni, A., Dwelly, T., et al. 2025, A&A, 698, A132
Balzer, F., Bulbul, E., Kluge, M., et al. 2025, A&A, 701, A283
Bauer, F. E., Alexander, D. M., Brandt, W. N., et al. 2004, AJ, 128, 2048
Boller, T., Freyberg, M. J., Trümper, J., et al. 2016, A&A, 588, A103
Boller, T., Liu, T., Weber, P., et al. 2021, A&A, 647, A6
Boller, T., Salvato, M., Buchner, J., et al. 2025, A&A, 700, A61
Brescia, M., Salvato, M., Cavuoti, S., et al. 2019, MNRAS, 489, 663
Brightman, M., & Nandra, K. 2011, MNRAS, 414, 3084
Brunner, H., Liu, T., Lamer, G., et al. 2022, A&A, 661, A1
Brusa, M., Zamorani, G., Comastri, A., et al. 2007, ApJS, 172, 353
Bulbul, E., Liu, A., Pasini, T., et al. 2022, A&A, 661, A10
Bulbul, E., Liu, A., Kluge, M., et al. 2024, A&A, 685, A106
Chen, S.-J., Buchner, J., Liu, T., et al. 2025, A&A, 701, A144
Chira, M., Georgakakis, A., & Ruiz, A. 2025, MNRAS
Civano, F., Elvis, M., Brusa, M., et al. 2012, ApJS, 201, 30
Croom, S. M., Richards, G. T., Shanks, T., et al. 2009, MNRAS, 392, 19
de Vaucouleurs, G., de Vaucouleurs, A., Corwin, Herold G., J., , et al. 1991, Third Reference Catalogue of Bright Galaxies
Delchambre, L., Bailer-Jones, C. A. L., Bellas-Velidis, I., et al. 2023a, A&A, 674, A31
DESI Collaboration (Abdul-Karim, M., et al.) 2025, ArXiv e-prints [arXiv:2503.14745]
Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168
Eckart, M. E., McGreer, I. D., Stern, D., Harrison, F. A., & Helfand, D. J. 2010, ApJ, 708, 584
Euclid Collaboration (Roster, W., et al.) 2025, A&A, in press https://doi.org/10.1051/0004-6361/202554616
Evans, I. N., Evans, J. D., Martínez-Galarza, J. R., et al. 2024, ApJS, 274, 22
Eyer, L., Audard, M., Holl, B., et al. 2023, A&A, 674, A13
Ferrarese, L., & Merritt, D. 2000, ApJ, 539, L9
Freund, S., Robrade, J., Schneider, P. C., & Schmitt, J. H. M. M. 2018, A&A, 614, A125
Freund, S., Czesla, S., Predehl, P., et al. 2024, A&A, 684, A121
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration (Brown, A. G. A., et al.) 2021, A&A, 649, A1
Gaia Collaboration (Bailer-Jones, C. A. L., et al.) 2023a, A&A, 674, A41
Gaia Collaboration (Vallenari, A., et al.) 2023b, A&A, 674, A1
Gebhardt, K., Bender, R., Bower, G., et al. 2000, ApJ, 539, L13
Ghirardini, V., Bulbul, E., Artis, E., et al. 2024, A&A, 689, A298
Grotova, I., Rau, A., Baldini, P., et al. 2025, A&A, 697, A159
Harrison, C. M., & Ramos Almeida, C. 2024, Galaxies, 12, 17
Healey, S. E., Romani, R. W., Taylor, G. B., et al. 2007, ApJS, 171, 61
Hickox, R. C., & Alexander, D. M. 2018, ARA&A, 56, 625
Hoffleit, D., & Warren, W.H.Jr 1995, VizieR Online Data Catalog: V/50
Høg, E., Fabricius, C., Makarov, V. V., et al. 2000, A&A, 355, L27
Hsu, L.-T., Salvato, M., Nandra, K., et al. 2014, ApJ, 796, 60
Igo, Z., Merloni, A., Hoang, D., et al. 2024, A&A, 686, A43
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jones, D. H., Read, M. A., Saunders, W., et al. 2009, MNRAS, 399, 683
Kluge, M., Comparat, J., Liu, A., et al. 2024, A&A, 688, A210
Kollmeier, J. A., Rix, H. W., Aerts, C., et al. 2025, ArXiv e-prints [arXiv:2507.06989]
Kormendy, J., & Ho, L. C. 2013a, ARA&A, 51, 511
Kormendy, J., & Ho, L.C. 2013b, ArXiv e-prints [arXiv:1308.6483]
Kovlakas, K., Zezas, A., Andrews, J. J., et al. 2021, MNRAS, 506, 1896
Kyritsis, E., Zezas, A., Haberl, F., et al. 2025, A&A, 694, A128
Lang, D. 2014, AJ, 147, 108
Lehmer, B. D., Xue, Y. Q., Brandt, W. N., et al. 2012, ApJ, 752, 46
Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, ApJS, 250, 8
Magorrian, J., Tremaine, S., Richstone, D., et al. 1998, AJ, 115, 2285
Marchesi, S., Civano, F., Elvis, M., et al. 2016, ApJ, 817, 34
Marocco, F., Eisenhardt, P. R. M., Fowler, J. W., et al. 2021, ApJS, 253, 8
Massaro, E., Maselli, A., Leto, C., et al. 2015, Ap&SS, 357, 75
Medvedev, P., Gilfanov, M., Sazonov, S., Schartel, N., & Sunyaev, R. 2021, MNRAS, 504, 576
Merloni, A., Lamer, G., Liu, T., et al. 2024, A&A, 682, A34
Montegriffo, P., De Angeli, F., Andrae, R., et al. 2023, A&A, 674, A3
Nakoneczny, S. J., Graham, M. J., Stern, D., et al. 2025, ApJ, 992, 153
Ni, Q., Brandt, W. N., Chen, C.-T., et al. 2021, ApJS, 256, 21
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Polletta, M., Tajer, M., Maraschi, L., et al. 2007, ApJ, 663, 81
Predehl, P., Andritschke, R., Arefiev, V., et al. 2021, A&A, 647, A1
Pulatova, N. G., Rubtsov, E., Chilingarian, I. V., et al. 2025, A&A, 702, A67
Rosen, S. R., Webb, N. A., Watson, M. G., et al. 2016, A&A, 590, A1
Salvato, M., Hasinger, G., Ilbert, O., et al. 2009, ApJ, 690, 1250
Salvato, M., Buchner, J., Budavári, T., et al. 2018a, MNRAS, 473, 4937
Salvato, M., Ilbert, O., & Hoyle, B. 2018b, Nat. Astron., 3, 212
Salvato, M., Wolf, J., Dwelly, T., et al. 2022, A&A, 661, A3
Saxena, A., Salvato, M., Roster, W., et al. 2024, A&A, 690, A365
Schlafly, E. F. 2021, Astrophysics Source Code Library [record ascl:2106.004]
Schwope, A., Kurpas, J., Baecke, P., et al. 2024a, A&A, 686, A110
Schwope, A. D., Knauff, K., Kurpas, J., et al. 2024b, A&A, 690, A243
SDSS Collaboration (Adamane Pallathadka, G., et al.) 2025, ArXiv e-prints [arXiv:2507.07093]
Seppi, R., Comparat, J., Bulbul, E., et al. 2022, A&A, 665, A78
Stern, D., Eisenhardt, P., Gorjian, V., et al. 2005, ApJ, 631, 163
Stern, D., Assef, R. J., Benford, D. J., et al. 2012, ApJ, 753, 30
Storey-Fisher, K., Hogg, D. W., Rix, H.-W., et al. 2024, ApJ, 964, 69
Sunyaev, R., Arefiev, V., Babyshkin, V., et al. 2021, A&A, 656, A132
Taylor, M. B. 2005, ASP Conf. Ser., 347, 29
Taylor, M. B. 2006, ASP Conf. Ser., 351, 666
The Dark Energy Survey Collaboration 2005, ArXiv e-prints [arXiv:astro-ph/0510346]
Waddell, S. G. H., Nandra, K., Buchner, J., et al. 2024, A&A, 690, A132
Webb, N. A., Coriat, M., Traulsen, I., et al. 2020, A&A, 641, A136
Wolf, J., Nandra, K., Salvato, M., et al. 2021, A&A, 647, A5
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Zenteno, A., Kluge, M., Kharkrang, R., et al. 2025, A&A, 698, A171

[1] Max-Planck-Institut für extraterrestrische Physik, Giessenbachstr. 1, 85748 Garching, Germany

[2] Exzellenzcluster ORIGINS, Boltzmannstr. 2, D-85748 Garching, Germany
[3] Max-Planck-Institut für Astronomie, Königstuhl 17., D-69117 Heidelberg, Germany
[4] Institute for Astronomy and Astrophysics, National Observatory of Athens, V. Paulou and I. Metaxa 11532, Greece
[5] Dipartimento di Fisica e Astronomia "Augusto Righi", Università di Bologna, Via Gobetti 93/2, 40129 Bologna, Italy
[6] INAF – Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Gobetti 93/3, 40129 Bologna, Italy
[7] Perimeter Institute for Theoretical Physics, 31 Caroline St. North, Waterloo, ON N2L 2Y5, Canada
[8] Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany
[9] University of Trento, Via Sommarive 14, I-38123 Trento, Italy
[10] Universitá di Firenze, Dipartimento di Fisica e Astronomia, Via G. Sansone 1, 50019 Sesto F.no, Firenze, Italy
[11] INAF – Osservatorio Astrofisico di Arcetri, Via Largo E. Fermi 5, 50125 Firenze, Italy
[12] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[13] Cerro Tololo Inter-American Observatory/NSF's NOIRLab, Casilla 603, La Serena, Chile
[14] Faculty of Physics, Ludwig-Maximilians-Universität München, Scheinerstr. 1, 81679 Munich, Germany
[15] Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
[16] NSF National Optical/Infrared Research Laboratory, 950 N. Cherry Ave, Tucson, AZ 85719, USA
[17] Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
[18] Department of Physics & Astronomy, University of Wyoming, 1000 E. University, Department 3905, Laramie, WY 8207, USA
[19] Department of Physics, University of Johannesburg, Johannesburg, South Africa
[20] Instituto de Astrofísica de Andalucia (IAA-CSIC), Glorieta de la astronomia S/N, 18008 Granada, Spain
[21] Las Campanas Observatory – Carnegie Institution for Science Av. Raul Bitran 1200, La Serena, Chile
[22] Universitá di Padova, Dipartimento di Fisica e Astronomia Galileo Galilei, Via Francesco Marzolo, 8, 35131 Padova, Italy
[23] South African Astronomical Observatory, PO Box 9, Observatory, Cape Town 7935, South Africa
[24] Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa
[25] Department of Physics, University of the Free State, PO Box 339, Bloemfontein 9300, South Africa
[26] Astronomical Observatory, University of Warsaw, Al. Ujazdowskie 4, 00-478 Warszawa, Poland
[27] University of Hamburg, Gojenbergsweg 112, 21029 Hamburg, Germany

Appendix A: Construction of a reference sample of X-ray sources having secure optical/IR counterparts

In this section, we describe the construction of a sample of point-like X-ray sources having secure optical/IR counterparts from the DESI Legacy Imaging Surveys catalogue (Dey et al. 2019), the Gaia EDR3 main source catalogue (Gaia Collaboration 2021), and the CatWISE2020 catalogue (Marocco et al. 2021). This reference sample is designed to span the X-ray flux range probed by the eRASS1 (Merloni et al. 2024) and eFEDS (Brunner et al. 2022) point source catalogues, and is used to train and validate our cross-matching algorithm. The selection criteria are tuned to give a high-purity sample, with completeness sacrificed when necessary.

A.1. 4XMM sample

We start with the 4XMM DR11 catalogue of detections (Webb et al. 2020), and retain only pointlike entries (SC_EXT_ML<4.0 and SC_EXTENT<3.0″), having a soft band flux (0.5–2 keV, computed as the sum of the native 0.5–1 and 1–2 keV bands) greater than 2×10^{-15} erg s^{-1} cm^{-2}, and with a signal-to-noise ratio on that flux measurement greater than 10. We then apply quality cuts on detection likelihood (SC_DET_ML>10.0) and positional precision (SC_POSERR<1.5″), and keep only isolated sources (DIST_NN>10.0″and CONFUSED=False), that were observed with low background (HIGH_BACKGROUND=False) with reasonable exposure time (EP_ONTIME>5 ks), are not flagged (SC_SUM_FLAG≤15), and which were not too far-off axis (<15′ in at least one instrument). Finally, we apply spatial filtering, removing any detections that lie at low Galactic latitudes ($|b| < 10$ deg), near the centres of the Large Magellanic Cloud (5 deg radius), the Small Magellanic Cloud (3 deg radius) or M31 (1 deg radius), within 0.05 deg of stars from the Yale Bright Star catalogue (Hoffleit & Warren 1995), within 0.05 deg of stars in Tycho-2 (Høg et al. 2000) having $VT < 9$ mag, or within the disc (radius derived from $D25$) of nearby bright ($B\_T < 12$ mag) galaxies from the RC3 catalogue (de Vaucouleurs et al. 1991). After retaining a single detection per unique X-ray source, we are left with a clean sample of 47 973 4XMM sources.

A.2. CSC2 sample

We retrieved a filtered sample of sources from the CSC2.0 primary_source table via the database query web API[23]. We selected bright ACIS detections with 0.5–2 keV aperture fluxes $> 2 \times 10^{-15}$ erg s^{-1} cm^{-2}, having a detection significance>6, that were point-like (major_axis_b<1 arcsec), well localised (err_ellipse_r0<1 arcsec), observed with reasonable exposure time (>1 ks), and those that were not flagged as piled up, confused, or associated with a read-out streak. As before, we excluded any sources which lie at low Galactic latitudes, near the LMC, SMC or M31, or those that were close to very bright stars or bright galaxies. After these filtering steps, our CSC2.0 sample contains 7426 sources.

A.3. Association with optical/IR counterpart catalogues

In order to provide samples of clean X-ray sources with secure optical/IR counterparts, we carried out the following procedure to match each X-ray reference sample with the DESI Legacy Imaging Survey DR10[24], Gaia EDR3 and CatWISE2020 catalogues. The following was repeated for each combination of X-ray and optical/IR catalogue. Firstly, we retrieved an initial catalogue of all objects in the optical/IR catalogue within a radius of 60″ of the X-ray positions - these initial catalogues serve the dual purposes of i) being the parent sample from which we find optical/IR counterparts, and ii) providing a pool of optical/IR objects which we can be certain are *not* strong X-ray emitters. Given the ∼20-year baselines spanned by the X-ray samples, we used proper motion estimates in the optical/IR catalogues to propagate the coordinates to the epoch of the nearest X-ray detection. Then, using only spatial information (positions, uncertainties), the NWAY tool (Salvato et al. 2018a) was used to find the best optical/IR counterpart (match_flag=1) for each X-ray source. Since we are interested only in a highly pure sample, we further filtered the samples to retain only secure cross-matches where p_any>0.3 and p_i>0.5. As a further refinement, an even higher quality sub-sample was identified, consisting of cross-matches having p_any>0.9 and p_i>0.8. The number of objects in our training samples is given in Table A.1.

To inform the cross-matching algorithm presented in this work, we generated a paired 'field catalogue' (i.e. non-X-ray emitting optical/IR objects) for each combination of X-ray and optical/IR catalogue. Each field catalogue is made up of all the optical/IR sources within an annulus (inner radius 15″, outer 60″) around the X-ray source positions, further excluding 15″ radius regions near *any* other X-ray sources found in the full 4XMM DR11 catalogue, or the CSC2 catalogue (considering only significantly detected compact sources with 0.5–2 keV aperture fluxes $> 5 \times 10^{-16}$ erg s^{-1} cm^{-2}). This approach ensures that the field catalogue probes a very similar distribution of survey conditions (both astrophysical and those associated with spatial variations in data quality), as the X-ray reference samples themselves.

Table A.1. Number of X-ray training sample sources (from 4XMM and CSC2) having secure counterparts in each of the three optical/IR catalogues considered here. The numbers in brackets are for the highest quality sub-sample.

O/IR Catalogue	N_{4XMM}	N_{CSC2}
LS10	39636 (32900)	5888 (5643)
Gaia EDR3	27219 (26129)	3477 (3379)
CatWISE2020	44218 (40399)	6204 (5707)

Appendix B: More examples of construction of AGN samples within the footprint of LS10

In Section 7.1 we described the construction of a highly pure (purity>90%) sample of eRASS1 AGN with counterparts in LS10. However, that selection included sources with different optical properties, depending on the depth of the LS10 data available in that position. A user could be interested in generating a sample with similar optical data. For that, samples 2,3, and 4 in Figure 11 would be more suitable. The first steps are the same as for Sample 1, but then we would consider the sources that are within the area with all(at least one) optical bands with nominal depth (inAllLS10=1/inAnyLS10=1).

[23] https://cxc.cfa.harvard.edu/csc/cli/index.html

[24] We used an early internally released version of LS DR10, which is identical to the public version in > 90% of survey bricks.

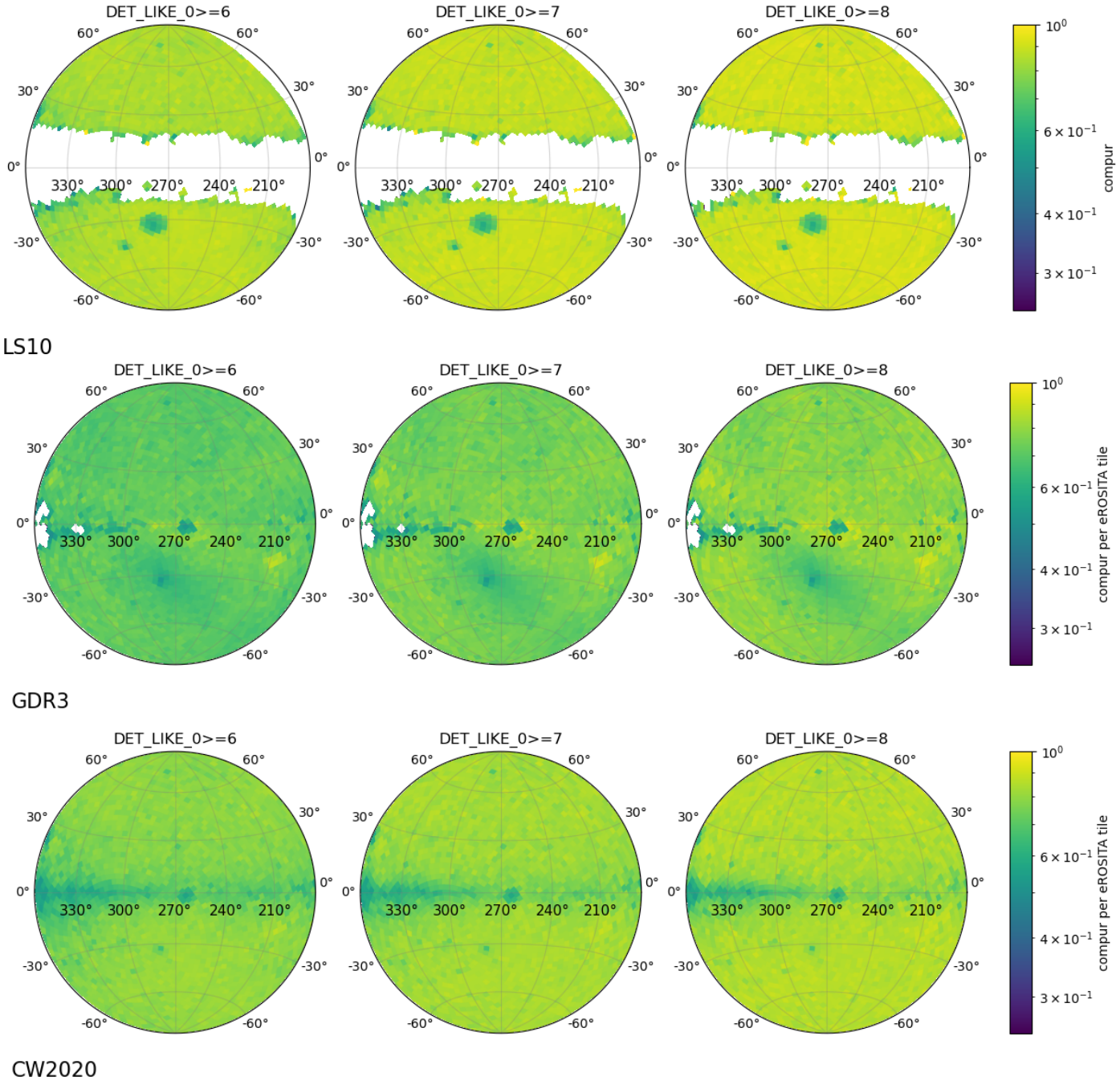

Fig. A.1. Map of the mean value per eROSITA tile of completeness and purity (`compur[6,7,8]` in the catalogues) at the intersection point (see Figure 6 for a visualisation) as a function of DET_LIKE_0, for LS10 (left panel), GDR3 (middle panel), and CW2020 (right panel). With the exception of the Magellanic Clouds and the Galactic plane, completeness and purity are high, and they increase with the detection likelihood. There are a few tiles in the Gaia maps where the completeness and purity curves do not intersect due to the limited number of sources.

To preserve high completeness, we can choose to select the sources that have `DET_LIKE_0`≥6 (Sample 2). As reported in Seppi et al. (2022) at this threshold, we expect up to 14% spurious detections, but the fraction is highly reduced once only sources with p_any above the threshold (see Section 3) are considered. However, the fraction of spurious X-ray detections could also be limited by selecting a higher detection likelihood (`DET_LIKE_0`≥7 or `DET_LIKE_0`≥8) and limiting the sources to those with p_any above the threshold that simultaneously maximise completeness and purity at that detection likelihood.

Appendix C: Acknowledgments

We are grateful to the referee who read the manuscript carefully and provided helpful suggestions. AG acknowledges funding from the Hellenic Foundation for Research and Innovation (HFRI) project "4MOVE-U" grant agreement 2688, which is part of the programme "2nd Call for HFRI Research Projects to support Faculty Members and Researchers". Ms and WR acknowledge the support (Förderkennzeichen 50002207) from Deutsches Zentrum für Luft- und Raumfahrt (DLR). ZI, JW, AM, and MS acknowledge the support by the Excellence Cluster

ORIGINS, which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2094 – 390783311. IM acknowledges financial support from the Severo Ochoa grant CEX2021-001131-S funded by MCIN/AEI/10.13039/501100011033 and by the Spanish Ministry of Science, Innovation y University (MCIU) under the grant PID2022-140871NB-C21. JR acknowledges support from DLR under 50QR2505. This work is based on data from eROSITA, the soft X-ray instrument aboard SRG, a joint Russian-German science mission supported by the Russian Space Agency (Roskosmos), in the interests of the Russian Academy of Sciences represented by its Space Research Institute (IKI), and the Deutsches Zentrum für Luft- und Raumfahrt (DLR). The SRG spacecraft was built by Lavochkin Association (NPOL) and its subcontractors, and is operated by NPOL with support from the Max Planck Institute for Extraterrestrial Physics (MPE).

The development and construction of the eROSITA X-ray instrument was led by MPE, with contributions from the Dr. Karl Remeis Observatory Bamberg & ECAP (FAU Erlangen-Nuernberg), the University of Hamburg Observatory, the Leibniz Institute for Astrophysics Potsdam (AIP), and the Institute for Astronomy and Astrophysics of the University of Tübingen, with the support of DLR and the Max Planck Society. The Argelander Institute for Astronomy of the University of Bonn and the Ludwig Maximilians Universität Munich also participated in the science preparation for eROSITA.

We have made use of TOPCAT and STILTS (Taylor 2005, 2006), scipy, astropy, scikit-learn, matplotlib.

Based on observations made with ESO Telescopes at the La Silla Paranal Observatory under programme IDs 177.A-3016, 177.A-3017, 177.A-3018 and 179.A-2004, and on data products produced by the KiDS consortium. The KiDS production team acknowledges support from: Deutsche Forschungsgemeinschaft, ERC, NOVA and NWO-M grants; Target; the University of Padova, and the University Federico II (Naples).

The data presented here were obtained in part with ALFOSC, which is provided by the Instituto de Astrofísica de Andalucía (IAA-CSIC) under a joint agreement with the University of Copenhagen and NOT. Some of the observations were collected at the Centro Astronómico Hispano Alemán (CAHA) at Calar Alto, operated jointly by the Instituto de Astrofísica de Andalucía (IAA-CSIC) and Junta de Andalucía.

The Legacy Surveys consist of three individual and complementary projects: the Dark Energy Camera Legacy Survey (DECaLS; Proposal ID 2014B-0404; PIs: David Schlegel and Arjun Dey), the Beijing-Arizona Sky Survey (BASS; NOAO Prop. ID 2015A-0801; PIs: Zhou Xu and Xiaohui Fan), and the Mayall z-band Legacy Survey (MzLS; Prop. ID 2016A-0453; PI: Arjun Dey). DECaLS, BASS and MzLS together include data obtained, respectively, at the Blanco telescope, Cerro Tololo Inter-American Observatory, NSF's NOIRLab; the Bok telescope, Steward Observatory, University of Arizona; and the Mayall telescope, Kitt Peak National Observatory, NOIRLab. The Legacy Surveys project is honoured to be permitted to conduct astronomical research on Iolkam Dúag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation.

This research used data obtained with the Dark Energy Spectroscopic Instrument (DESI). DESI construction and operations are managed by the Lawrence Berkeley National Laboratory. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High-Energy Physics, under Contract No. DE–AC02–05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI was provided by the U.S. National Science Foundation (NSF), Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF's National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Humanities, Science and Technology of Mexico (CONAHCYT); the Ministry of Science and Innovation of Spain (MICINN), and by the DESI Member Institutions: www.desi.lbl.gov/collaborating-institutions. The DESI collaboration is honoured to be permitted to conduct scientific research on I'oligam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation, the U.S. Department of Energy, or any of the listed funding agencies.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

Funding for the Sloan Digital Sky Survey V has been provided by the Alfred P. Sloan Foundation, the Heising-Simons Foundation, the National Science Foundation, and the Participating Institutions. SDSS acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. SDSS telescopes are located at Apache Point Observatory, funded by the Astrophysical Research Consortium and operated by New Mexico State University, and at Las Campanas Observatory, operated by the Carnegie Institution for Science. The SDSS website is www.sdss.org.

SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration, including Caltech, The Carnegie Institution for Science, Chilean National Time Allocation Committee (CNTAC) ratified researchers, The Flatiron Institute, the Gotham Participation Group, Harvard University, Heidelberg University, The Johns Hopkins University, L'Ecole polytechnique fédérale de Lausanne (EPFL), Leibniz-Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Extraterrestrische Physik (MPE), Nanjing University, National Astronomical Observatories of China (NAOC), New Mexico State University, The Ohio State University, Pennsylvania State University, Smithsonian Astrophysical Observatory, Space Telescope Science Institute (STScI), the Stellar Astrophysics Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Illinois at Urbana-Champaign, University of Toronto, University of Utah, University of Virginia, Yale University, and Yunnan University.

This paper uses observations made at the South African Astronomical Observatory (SAAO).