A&A, 711, A62 (2026)
https://doi.org/10.1051/0004-6361/202658965
© The Authors 2026

Astronomy & Astrophysics

# Performance analysis of extragalactic classifications in *Gaia* Data Release 4

Sara Jamal[1,⋆], Coryn A. L. Bailer-Jones[1,⋆], Ruth Carballo[2], and Orlagh L. Creevey[3]

[1] Max-Planck Institute für Astronomy, Königstuhl 17, 69117 Heidelberg, Germany
[2] Dpto. de Matemática Aplicada y Ciencias de la Computación, Univ. de Cantabria, ETS Ingenieros de Caminos, Canales y Puertos, Avda. de los Castros s/n, 39005 Santander, Spain
[3] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice Cedex 4, France

Received 14 January 2026 / Accepted 21 May 2026

**ABSTRACT**

The Discrete Source Classifier (DSC) provides probabilistic classifications of sources in *Gaia* Data Release 4 (GDR4) based on empirically trained Bayesian classifiers. Using the *Gaia* astrometry, photometry, and low-resolution spectra (XP), the DSC classifies all sources into one of three classes: quasar, galaxy, or star. The DSC comprises three trained neural networks and three different combinations of their probabilities. When evaluated as a function of brightness and sky position on a test set that excludes the Magellanic Clouds, the DSC purity in GDR4 has improved but at the cost of a small loss in completeness. The average performance of the best classifiers at magnitudes brighter than $G$=20 is at least 88% completeness and 96% purity for the extragalactic classes, namely the quasar and galaxy classes. At fainter magnitudes, the performance is lower due to increased noise. The average performance at magnitudes of $20 \leq G < 20.5$ is a minimum of 55% completeness and 71% purity for the extragalactic classes. At $G > 20.5$ mag, the completeness is considerably reduced, primarily for the models that depend on the XP spectra. Furthermore, we trained additional models on the *Gaia* optical data together with mid-infrared photometry from the CatWISE2020 catalogue. Inclusion of infrared photometry increases the completeness of extragalactic samples at $G > 20$ mag between 9 and 29 percentage points at the cost of reducing the purity between 1 and 9 percentage points. In the GDR4, the best DSC-combined classifier prioritising completeness identifies three million quasars and two million galaxies, but with expected high contamination among fainter sources. In contrast, the combined classifiers prioritising purity identify approximately two million quasars and 1.3 million galaxies with an expected lower level of contamination. At the end of this paper, we provide recommendations for enhancing the purity of the DSC extragalactic selection by applying quality cuts to the *Gaia* photometry and astrometry.

**Key words.** methods: data analysis – methods: statistical – surveys – stars: general – galaxies: general – quasars: general

## 1. Introduction

Over the past few decades, numerous ground- and space-based surveys have compiled extensive catalogues of extragalactic sources, including quasars and galaxies, such as the Sloan Digital Sky Survey Quasar catalogue (SDSS DR16Q; York et al. 2000; Lyke et al. 2020), the Large Quasar Astrometric Catalogue (LQAC-6, Souchay et al. 2024), the Wide-field Infrared Survey Explorer (WISE) active galactic nucleus (AGN) catalogues (R90 and C75; Assef et al. 2018), the 2dF QSO Redshift Survey (2QZ; Croom et al. 2004), the AllBRICQS Survey (Onken et al. 2023), the Milliquas Catalogue (Flesch 2023), and the *Gaia* Data Release 3 extragalactic candidates (Gaia Collaboration 2023a). The extragalactic sky will be mapped in even greater detail, reaching fainter magnitudes, thanks to various other surveys, such as the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezić 2016; Ivezić et al. 2019), the Dark Energy Spectroscopic Instrument survey (DESI; DESI Collaboration 2025a,b,c), the *Euclid* ESA mission (Euclid Collaboration 2025), and the *Gaia* ESA mission (Gaia Collaboration 2023b).

A quasar – a quasi-stellar object (QSO) – is an extremely luminous object driven by an accreting supermassive black hole located at the centre of its host galaxy. The identification of quasars and galaxies is essential to constraining galaxy evolutionary models (Hopkins et al. 2006), characterising the growth of supermassive black holes (Di Matteo et al. 2005; Hopkins et al. 2005; Volonteri & Rees 2006), and tracing the evolution history of structures in the early Universe and the epoch of reionisation (Fan et al. 2006; Mortlock et al. 2011; Bañados et al. 2018), to mention a few examples. Furthermore, cosmological probes with detectable signatures in the distribution of quasars and galaxies – such as distortions in the redshift space (RSDs) and baryon acoustic oscillations (BAOs) – are used to investigate the growth history of large-scale structures and the role of dark energy in the late-time acceleration of the expansion of the Universe. Redshift space distortions emerge as an anisotropic clustering in the spatial distribution of galaxies caused by the expansion of the Universe and the peculiar velocities of galaxies. Their observation provides a direct measurement of the linear growth rate of large-scale structures (Beutler et al. 2012; Zarrouk et al. 2018). Conversely, the imprint from the BAOs on the late-time clustering of galaxies in the early Universe is used to measure the distance scale and trace the expansion history through measurements of the angular comoving distances and the Hubble parameter as functions of the redshift (Eisenstein & Hu 1998; Seo & Eisenstein 2003; Linder 2003; Blake & Bridle 2005; Eisenstein et al. 2005; Dolney et al. 2006; Beutler et al. 2011). Likewise, the Ly$\alpha$ forest in the spectra of quasars traces

⋆ Corresponding authors: jamal@mpia.de; calj@mpia.de



Open Access article, published by EDP Sciences, under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
This article is published in open access under the Subscribe to Open model. Open access funding provided by Max Planck Society.

the matter distribution and clustering and serves as a powerful probe to infer the expansion history of the Universe (Busca et al. 2013; Slosar et al. 2013; Font-Ribera et al. 2014; du Mas des Bourboux et al. 2020; Bautista et al. 2017). Therefore, the construction of high-purity quasar and galaxy catalogues lies at the core of several cosmological studies.

Primarily designed to map stars in the Milky Way, the *Gaia* ESA satellite has observed more than two billion sources, including millions of quasars and galaxies, across the entire sky down to very faint magnitudes (Gaia Collaboration 2023a). Several modules within the *Gaia* Data Processing and Analysis Consortium (DPAC) perform classification using machine learning (ML) techniques. One of these modules is the Discrete Source Classifier (DSC), part of the Coordination Unit 8 (CU8) for the astrophysical parameters inference system (Apsis) chain (Bailer-Jones et al. 2013; Andrae et al. 2018; Creevey et al. 2023; Fouesneau et al. 2023; Delchambre et al. 2023). The DSC uses supervised classification on *Gaia* photometry, astrometry, and low-resolution spectroscopy to identify a source as a quasar, a galaxy, or a star. The module computes posterior probabilities across the target classes and uses a prior function that represents our expectation of the overall distribution of classes in the sky (Delchambre et al. 2023). In *Gaia* Data Release 3 (GDR3), the DSC achieved a high completeness of 92% but a low purity of 22% for the quasar and galaxy classes. For the upcoming *Gaia* DR4 (GDR4), we aim to improve the classifications and produce quasar and galaxy catalogues with the DSC of a higher purity than was achieved in GDR3.

In the literature, sources are classified using information from spectra, photometry and astrometry to determine their type. The spectra of quasars typically exhibit a flat continuum alongside broad emission lines such as H$\alpha$ and Ly$\alpha$ and narrow emissions due to ionisation [OII] and [OIII] (Alexander et al. 2023). Spectral features of quasars can reveal evidence of strong star-formation activity (Lacy et al. 2007; Zakamska et al. 2008), outflows of ionised gas (Villar-Martín et al. 2016), and streams or tidal effects resulting from interactions and major mergers (Bessiere et al. 2012; Pierce et al. 2023). Quasar signatures can be detected across multiple wavelengths, including ultraviolet (UV), optical, and infrared (IR). However, these key signatures in the spectrum shift towards longer wavelengths with increasing redshift. This result means that very distant quasars at redshifts greater than $z\approx6.5$ can only be detected via their Ly$\alpha$ line in the infrared or radio spectrum (Mortlock et al. 2012). The spectra of galaxies are dominated by a strong continuum component superimposed on absorption lines characteristic of the stellar population as well as emission lines from ionised gas that indicate whether a galaxy is undergoing active star formation, is a host of an AGN, or is in a quiescent phase. Stellar spectra show a flat continuum with absorption lines that provide information about the star's evolutionary phase and key properties, such as effective temperature, luminosity, and element abundances. In the colour space, the different classes occupy distinct regions. For instance, quasars appear bluer than stars or galaxies, and stars tend to be concentrated along a distinct locus (Richards et al. 2001, 2002).

Historical methods for quasar-galaxy-star classification relied on colour cuts in UV, optical, or IR photometry (Newberg & Yanny 1997; Fan 1999; Richards et al. 2002; Lacy et al. 2004; Trammell et al. 2007; Hutchings & Bianchi 2010; Wu & Jia 2010; Stern et al. 2012; Wang et al. 2016). Nevertheless, these approaches are limited by the low dimensionality of the data and the arbitrary selection of thresholds applied to a small sample of sources. An improved classification requires exploring large, highly dimensional datasets. To this end, ML methods have been recognised as powerful tools for processing large datasets and detecting patterns and outliers. Various methods have been applied extensively to the quasar-galaxy-star classification, such as tree-based estimators (Weir et al. 1995; Carrasco et al. 2015; Jin et al. 2019; Schindler et al. 2017; Bai et al. 2018; Nakazono et al. 2021; Zhang et al. 2021; Golob et al. 2021; Li et al. 2021; Baqui et al. 2021; Hughes et al. 2022; Rimoldini et al. 2023; Stoppa et al. 2023; Delchambre et al. 2023; Cunha et al. 2024; Fu et al. 2024, 2025), Bayesian kernel density estimators (Richards et al. 2004; Kirkpatrick et al. 2011; Peters et al. 2015; Richards et al. 2015; Bailer-Jones et al. 2019), support vector machines (SVMs; e.g. Gao et al. 2008; Peng et al. 2012; Małek et al. 2013; Wang et al. 2016; Nakazono et al. 2021), and neural networks (NNs; Yèche et al. 2010; Tuccillo et al. 2015; Kim & Brunner 2017; Burke et al. 2019; He et al. 2021; Guo et al. 2022; Martínez-Solaeche et al. 2023; Stoppa et al. 2023; Merz et al. 2023; Rodrigues et al. 2023; Chaini et al. 2023; Cunha et al. 2024). Several studies have used a combination of multi-wavelength observations to improve quasar selection (Wu & Jia 2010; Bovy et al. 2012; DiPompeo et al. 2015; Carrasco et al. 2015; Richards et al. 2015; Cunha et al. 2024; Fu et al. 2024, 2025).

In GDR3, the DSC used a combination of a tree-based algorithm, called Extratrees, and a density-based method, called Gaussian mixture models, to classify sources as quasars, galaxies, single stars, white dwarfs, or binaries based on *Gaia* astrometry, photometry, and low-resolution spectroscopy. For GDR4, we have explored various classification algorithms and refined the construction of the labelled datasets used to train the models. We now use a combination of neural network models, known as multi-layer perceptrons, which have proven highly efficient at rejecting stellar false detections, albeit at a reduced completeness. Moreover, the DSC in GDR4 provides independent classifications combining *Gaia* optical data with mid-infrared photometry from the CatWISE2020 catalogue (Marocco et al. 2021). Throughout this paper, results obtained using the *Gaia* data are referred to as the '*Gaia* only' mode, whereas those obtained using the *Gaia* data alongside the CatWISE infrared photometry are identified as the '*Gaia*-IR' mode. This paper is organised as follows. Section 2 provides a brief description of the data used for classification. In Section 3, we describe the methodology, algorithms, and architecture. In Section 4, we evaluate classification performance in terms of completeness and purity on a labelled test set relative to the sky positions and brightnesses of the sources. We also present the full range of extragalactic classifications from the DSC in GDR4 and discuss the physical properties of quasar and galaxy candidates. Finally, we provide recommendations for quality cuts to the *Gaia* photometry and astrometry to improve the purity of the DSC extragalactic sample. We summarise our key results in Section 5.

## 2. Data

From its launch in December 2013 to its decommissioning in March 2025, the *Gaia* ESA satellite observed over two billion sources across the entire sky. Examples of the scientific results stemming from *Gaia* include detailed multi-dimensional maps of the Milky Way (Gaia Collaboration 2023b) and the identification of new black holes (El-Badry et al. 2023; Gaia Collaboration 2024).

The upcoming GDR4 is based on 66 months of data collected by the satellite between 2014 and 2020. The previous

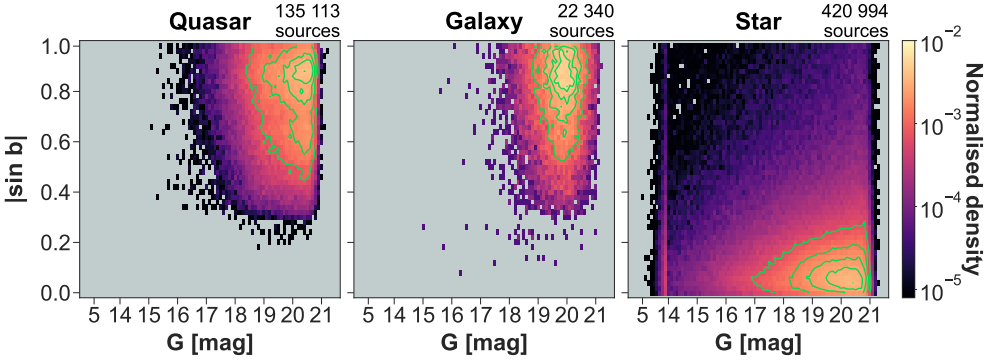



**Fig. 1.** Representation in 2D of the distribution of sources as a function of Galactic latitude and brightness of the training dataset. Each distribution has been normalised by the total number of sources in that panel. The colour scale is set such that brighter colours refer to regions of higher density compared to darker colours, which have fewer sources. The 2D representation of the density of sources is defined on a 50×91 bin grid in $|\sin b|$ and $G$. Binning into magnitude bins compresses the magnitude range at $G$<14. Contours (green lines) show the normalised density on a log scale of the highest-density regions.

GDR3 (Gaia Collaboration, 2016; Gaia Collaboration; 2023b) published data for approximately 1.9 billion sources down to $G\simeq20$ mag, while the GDR4 will publish all the processed objects, approximately 2.8 billion sources down to slightly fainter magnitudes. The new catalogue will contain astrometry, photometry, low-resolution BP/RP (XP) spectroscopy, and classifications, including the DSC results. A summary of the expected content of GDR4 is available via the public ESA link[1]. In the GDR4, the DSC provides posterior probabilities and class labels in the main table (`gaia_source`), the astrophysical parameter tables (`ap_class` and `ap_class_ir`), as well as the extragalactic tables (`qso_candidates` and `galaxy_candidates`). The `ap_class` reports the classifications using the *Gaia* data, while the `ap_class_ir` table provides classifications using a combination of *Gaia* data and mid-infrared photometry from CatWISE.

The CatWISE2020 catalogue (Marocco et al. 2021) has observed over 1.8 billion sources across the entire sky selected from the Wide-field Infrared Survey Explorer (WISE) and the NEOWISE survey data collected between 2010 January 7 and 2018 December 13. The catalogue provides an extensive list of infrared sources with improved magnitude detection limits in the $W1$ (3.4 μm) and $W2$ (4.6 μm) infrared passbands, reaching 90% completeness depth at $W1$ = 17.7 mag and $W2$ = 17.5 mag. We combine the mid-infrared photometry bands $W1$ and $W2$ of CatWISE2020 with the optical data from *Gaia* to classify a source in the *Gaia*-IR mode.

## 3. Classification methodology

### 3.1. An outline of the DSC in GDR4

The DSC is composed of six classifiers that compute posterior class probabilities: three baseline classifiers trained using a set of labelled sources, `Allosmod`, `Specmod`, and `Specmod-Supp`, complemented with three combined classifiers, `Combmod`, `Combmod-`$\alpha$, and `Combmod-Supp`. `Specmod` is trained exclusively on the low-resolution XP spectra, `Allosmod` is trained on various astrometric and photometric features, not the XP spectra, and `Specmod-Supp` is trained on both the XP spectra and various astrometric and photometric features. In *Gaia*-IR mode, the CatWISE mid-infrared photometry is supplemented to train the auxiliary classifiers `Allosmod(*)` and `Specmod-Supp(*)`[2]. The list of features is summarised in Appendix A. The combined classifiers are defined as follows. `Combmod` combines the `Allosmod` and `Specmod` posterior probabilities as a weighted product, as was done in GDR3 (Bailer-Jones 2021; Ulla et al. 2022). `Combmod-`$\alpha$ is a parametric additive combiner that weights the contributions of `Allosmod` and `Specmod` probabilities. The optimal parameter set $\alpha$ maximises the classification score[3] of the quasar and galaxy classes (Jamal & Bailer-Jones 2024). `Combmod-Supp` selects the probabilities of `Specmod-Supp` when available, otherwise `Specmod`. `Combmod-Supp` is equal to `Specmod-Supp` if all inputs are available; otherwise, it is equal to `Specmod`.

Our approach of using multiple classifiers allows for greater flexibility when processing sources with missing input data. For example, sources lacking parallaxes and proper motions, but with valid BP and RP spectra, are classified using `Specmod` and not processed using `Allosmod` or `Specmod-Supp`. We also decided to withdraw the white dwarf and binary star classes due to their poor classification performance in GDR3 (Delchambre et al. 2023). Consequently, we retained a supervised classification scheme that identifies a source as a quasar, a galaxy, or a star. In GDR3, we used a decision tree-based classifier, Extra-trees (Geurts et al. 2006), for Specmod, and Gaussian Mixture Models (Reynolds 2015) for Allosmod. In GDR4, all three DSC trained models are neural networks. To compute posterior probabilities, the DSC uses a global prior that accounts for the rarity of the quasar and the galaxy classes compared to the larger star class. The prior probabilities of any source being a quasar or a galaxy are 1/1000 and 1/5000, respectively, corresponding to the same values used in GDR3.

Our reference dataset, with labelled sources used for training and testing, is built from reliable external catalogues and data releases. The quasar and galaxy classes correspond to a sample of spectroscopically confirmed sources from the SDSS (Sloan Digital Sky Survey) DR17 catalogue (Abdurro'uf et al. 2022). The crossmatch to the SDSS catalogue within a one arcsecond radius yields a large number of sources (563 233 quasars and 1 403 392 galaxies). Our star class corresponds to a random sample of 1 097 000 sources from *Gaia*, excluding the extragalactic selection. We reduced this selection as we require

[1] https://www.cosmos.esa.int/web/gaia/dr4

[2] The symbol *(*)* indicates the models using the *Gaia* data alongside infrared photometry from CatWISE.

[3] The score is a metric that measures the classifier's performance between 0 and 1, where 1 indicates an ideal result.

valid XP spectra and all features to be available for training and testing (no missing data). We further cleaned the sample by requiring at least five transits to contribute to the mean BP and RP spectra. Finally, we removed quasar and galaxy candidates found in the following non-*Gaia* catalogues from our star class: Milliquas v7.2-v8 (Flesch 2015, 2023), LAMOST DR7 (Luo et al. 2022), SDSS DR17 photometry (Abdurro'uf et al. 2022), the R90 and C75 WISE AGN catalogues (Assef et al. 2018) as well as a private communication of the *Gaia* DPAC CU4-EO (Extended Objects) list of candidates. The final reference set contains 270 436 quasars, 44 894 galaxies and 857 977 stars.

The selected set is randomly divided into equal-sized training and testing subsets. When assessing performance, the test set excludes the Magellanic Clouds. These regions have a higher level of stellar contamination, which masks the full range of extragalactic classifications across the entire sky. We remove sources within 9° and 6° radii of the centres of the Large Magellanic Cloud (LMC) and the Small Magellanic Cloud (SMC), respectively (see Appendix B, Gaia Collaboration 2023a). Because the SDSS footprint does not cover the LMC and SMC regions, only our star class includes sources in those regions. We excluded these regions from the assessment of the performance on the test set, not from the training data. Figure 1 displays the latitude and magnitude distribution of the training set. In our reference set, the relative density of the quasars and galaxies peaks at fainter magnitudes at higher latitudes, while the density of stars peaks closer to the Galactic plane and at very bright magnitudes $G>14$. Since the SDSS footprint shows a disparity between the northern and southern Galactic hemispheres, we randomised the Galactic latitudes $b$ by drawing from a uniform distribution over $|\sin b|$ during the training only to emulate a uniform distribution of quasars and galaxies across the sky and prevent the classifier from learning the SDSS footprint for these objects. This randomisation strategy was also employed in the DSC in GDR3 (Delchambre et al. 2023). We do not use longitudes because the impact of sky position on classification is already encoded in Galactic latitudes.

### 3.2. Methods

This section outlines the methods and the training strategies employed by the DSC. We present the neural network architecture and the combined classifiers.

#### 3.2.1. Multi-layer perceptron model

Neural networks (NNs) are models in artificial intelligence that use specialised blocks to process input data and propagate information to produce an output. In its simplest form, the multilayer perceptron consists of an input layer, an output layer, and fully connected hidden layers (Rumelhart et al. 1986). The advantage of NNs lies in their flexible design and the selection of an objective function. However, designing an optimal model remains an empirical process that involves a grid search over hyperparameters, such as the number of layers, activation functions, and loss functions. For the DSC, we trained multiple MLP configurations and selected the models that perform best on the test set (i.e. the dataset never seen during training). The training set was split 80:20 for training and validation. At each epoch, a stopping criterion monitored the validation score, a metric that measures performance on the validation dataset, ranging from 0 to 1 (where 1 indicates an ideal result). The MLP architectures and the list of hyperparameters for the DSC models are provided in Appendix B. A comparison of the different algorithms and classification strategies explored for the DSC is available in a separate technical note (Jamal & Bailer-Jones 2026).

#### 3.2.2. Combined classifiers

The combined classifiers exploit the individual results of the three baseline classifiers, `Allosmod`, `Specmod`, and `Specmod-Supp`. As in GDR3, `Combmod` probabilities are computed from a product of the `Specmod` and `Allosmod` posterior probabilities, assuming that the two classifiers use the same prior and are independent (Eq. (1a)). We adapt the GDR3 definition to account for three classes. The parametric combiner `Combmod-`$\alpha$, introduced in Jamal & Bailer-Jones (2024), is computed from a weighted addition of the `Specmod` and `Allosmod` probabilities, such that the combination maximises the classification score (Eq. (1b)). The probabilities of the combined classifiers were normalised across all classes.

$$P_k^c = P_k^s P_k^a \frac{1}{\pi_k}, \tag{1a}$$

$$P_k^{c_\alpha} = (1-\alpha_k) P_k^s + \alpha_k P_k^a, \tag{1b}$$

where $k$ refers to the class, $\pi_k$ the class prior, $\alpha_k$ the parameter of the class $k$, and ($P_k^c$, $P_k^{c_\alpha}$, $P_k^s$, $P_k^a$) the posterior probabilities for `Combmod`, `Combmod-`$\alpha$, `Specmod` and `Allosmod`, respectively.

For the parametric combiner, we performed a grid search within a limited discrete space over the parameters $\alpha_k \in [0,1]$ and selected the optimal set achieving the highest geometric score (Eq. (2b)). For the quasar, galaxy and star classes, we found that the optimal result is given by the values $\boldsymbol{\alpha}=\{0.9,0.8,0.3\}$. These values imply a greater contribution of the `Allosmod` probabilities in defining the `Combmod-`$\alpha$ probabilities. In the case of missing data, `Combmod` and `Combmod-`$\alpha$ are equal to the probabilities of `Specmod`. The `Combmod-Supp` probabilities are set to the probabilities of `Specmod-Supp` when available; otherwise, they are set to the probabilities of `Specmod`.

## 4. DSC performance in GDR4

### 4.1. Evaluation on the test set

#### 4.1.1. Performance metrics

To evaluate the performance of a classifier, we compute two metrics from the confusion matrix, namely the completeness and the purity. The completeness is the proportion of correct predictions in each class relative to the total number of true members of that class. The purity is the proportion of correct predictions in each class relative to the total number of predicted members of that class. The geometric score is defined as the geometric mean of the completeness and the purity of the quasar and galaxy classes. The metrics are defined as follows:

$$\text{completeness}_\text{k} = \frac{\text{TP}_\text{k}}{\text{TP}_\text{k} + \text{FN}_\text{k}}, \qquad \text{purity}_\text{k} = \frac{\text{TP}_\text{k}}{\text{TP}_\text{k} + \text{FP}_\text{k}}, \tag{2a}$$

$$\text{score}_\text{geometric} = \Big(\text{completeness}_\text{quasar} \times \text{completeness}_\text{galaxy} \times \text{purity}_\text{quasar} \times \text{purity}_\text{galaxy}\Big)^{1/4}, \tag{2b}$$

where $k$ refers to the class, and TP, FN, and FP refer to the true positives, the false negatives, and the false positives, respectively. Both high completeness and high purity are desirable,

**Table 1.** Summary of the overall classification performance evaluated on the test dataset for models trained on *Gaia* data only.

| Magnitude bins | Classifiers | Overall completeness | | Overall purity | | Geometric score |
|---|---|---|---|---|---|---|
| | | Quasar | Galaxy | Quasar | Galaxy | |
| All magnitudes ($q$:135 323, $g$:22 447) | `Specmod` | 0.702 (+29 pp) | 0.722 (−13 pp) | 0.691 (+25 pp) | 0.618 (+26 pp) | 0.682 (+19 pp) |
| | `Allosmod` | 0.869 (+3 pp) | 0.773 (−18 pp) | 0.737 (+7 pp) | 0.703 (+27 pp) | 0.768 (+7 pp) |
| | `Specmod-Supp` | 0.862 | 0.723 | 0.749 | 0.778 | 0.776 |
| | `Combmod` | 0.972 (+5 pp) | 0.941 (−2 pp) | 0.608 (+20 pp) | 0.373 (+20 pp) | 0.675 (+14 pp) |
| | `Combmod-`$\alpha$ | 0.843 (+2 pp) | 0.762 (−17 pp) | 0.762 (+6 pp) | 0.741 (+23 pp) | 0.776 (+5 pp) |

**Notes.** The number of quasars ($q$) and galaxies ($g$) in the test set is indicated. The performance of the `Combmod-Supp` classifier is equivalent to `Specmod-Supp` in the test set, given the availability of all required inputs. Results highlighted in grey, light red, and dark red correspond to values below 0.75, 0.50, and 0.25, respectively. The numbers in pp (percentage points) show the change in performance relative to that achieved in GDR3 as evaluated on the same test set. Negative and positive percentage points are shown in blue and red colours, respectively. The DSC stellar predictions from GDR3 are merged into one star class for this purpose.

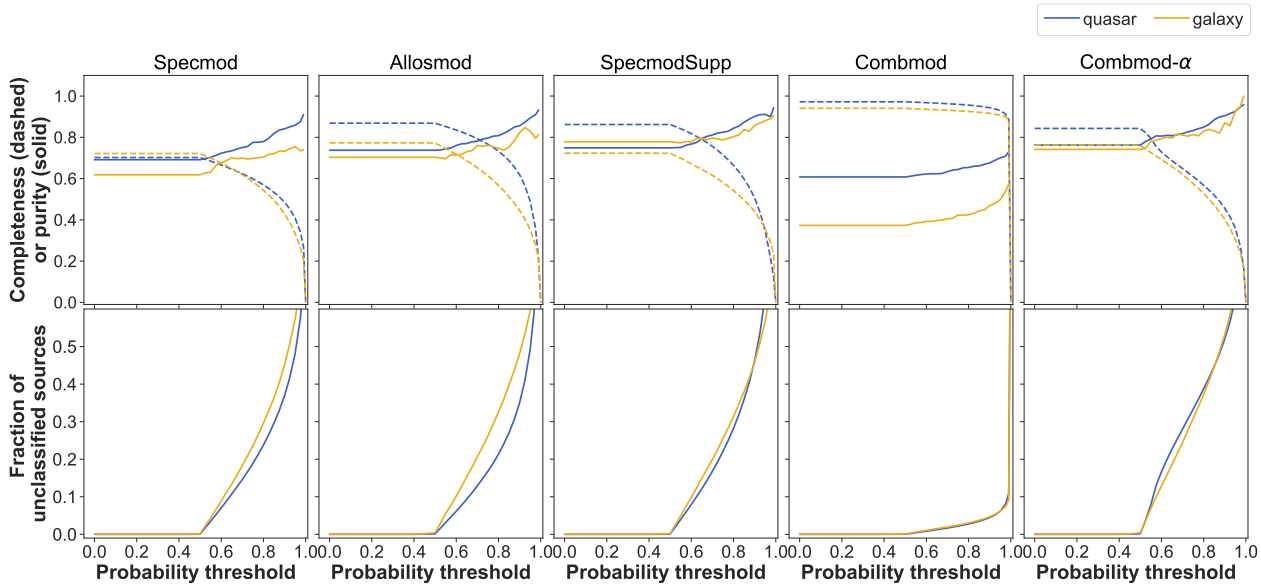



**Fig. 2.** *Top*: variation in the completeness and the purity of each class in the test set as a function of the posterior probability threshold used to assign a class. *Bottom*: variation in the proportion of the unclassified sources with probabilities below the threshold relative to the total number of sources in each class. The $y$-axis is limited to 0.5 for visualisation purposes only.

but there is an unavoidable trade-off between them. Low completeness but high purity indicates a classifier that rejects false positives (i.e. sources from other classes incorrectly classified as the target class) but overlooks true objects. High completeness but low purity, in contrast, reflects a classifier that can identify the true classes but also introduces false detections. In the DSC, a source is assigned to the class that achieves the highest classification probability above a fixed threshold. If this maximum probability is below the threshold, the source is marked as unclassified. A high threshold rejects more false positives but reduces completeness by also rejecting true positives with probabilities below the threshold.

In classification tasks, class imbalance affects performance on minority classes because the loss function is overweighted by the dominant class during training. In our case, a classifier can easily overlook the rare quasar and galaxy classes in favour of the much larger star class. We addressed class imbalance using an approach similar to that developed in Bailer-Jones et al. (2019) and adopted in Delchambre et al. (2023) and Jamal & Bailer-Jones (2024). First, we applied a prior function to the probabilities to reflect the expected fraction of each class in the data. Second, we adjusted the confusion matrix to align the test set with the overall class distributions expected in the *Gaia* survey. This adjustment affects only the purities, not the completeness. Throughout this work, we report adjusted metrics. Ignoring these adjustments (as is often the case in the literature) would give the false appearance of better performance. We do not report the performance for the star class, since its completeness and purity are always close to unity due to the dominance of stars. Classifying everything as a star would also produce purity and completeness close to unity. `Combmod-Supp` is equal to `Specmod-Supp` if all inputs are available; otherwise, it is equal to `Specmod`.

#### 4.1.2. Overall performance

We assign objects to classes with maximum probabilities above a threshold. Figure 2 shows the variation in the overall performance evaluated on the test set as a function of the posterior probability threshold. As the threshold increases, completeness

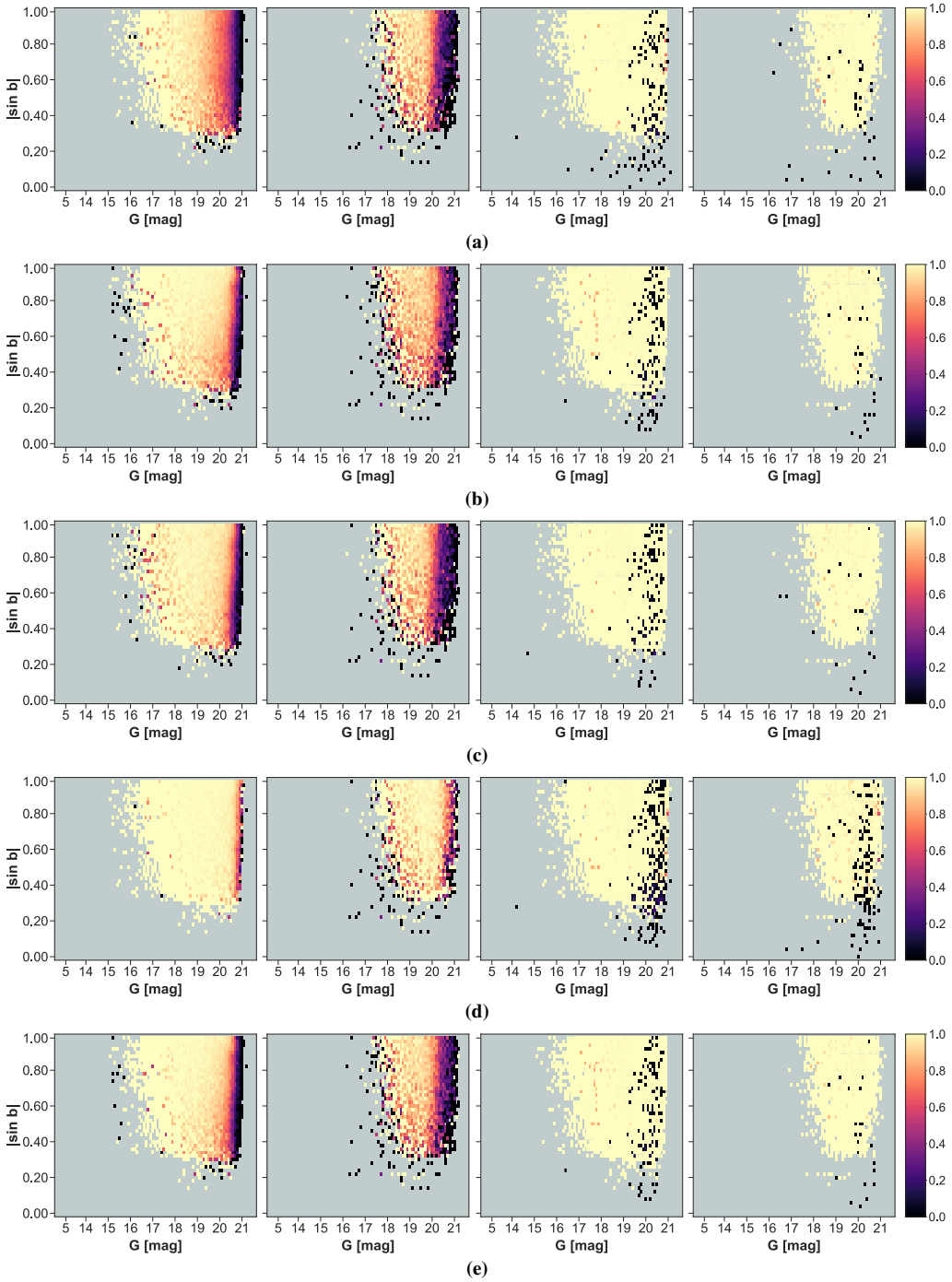
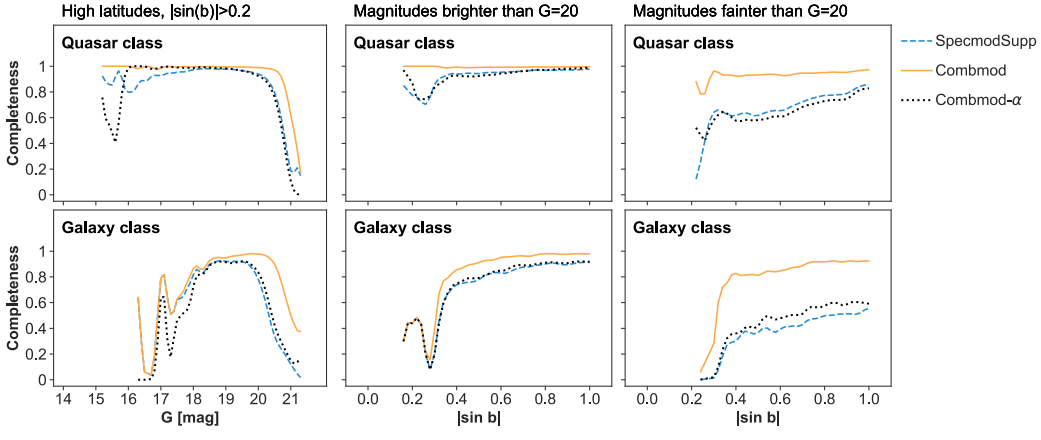



**Fig. 3.** The Discrete Source Classifier performance on the test dataset as a function of source Galactic latitude and G-magnitude. Predicted labels were obtained by selecting the class with the highest posterior probability using the global prior. Left to right, the completenesses and the purities of the quasar and galaxy classes. Rows (*a*), (*b*), (*c*), (*d*), and (*e*) refer to `Specmod`, `Allosmod`, `Specmod-Supp`, `Combmod`, and `Combmod-`$\alpha$, respectively.

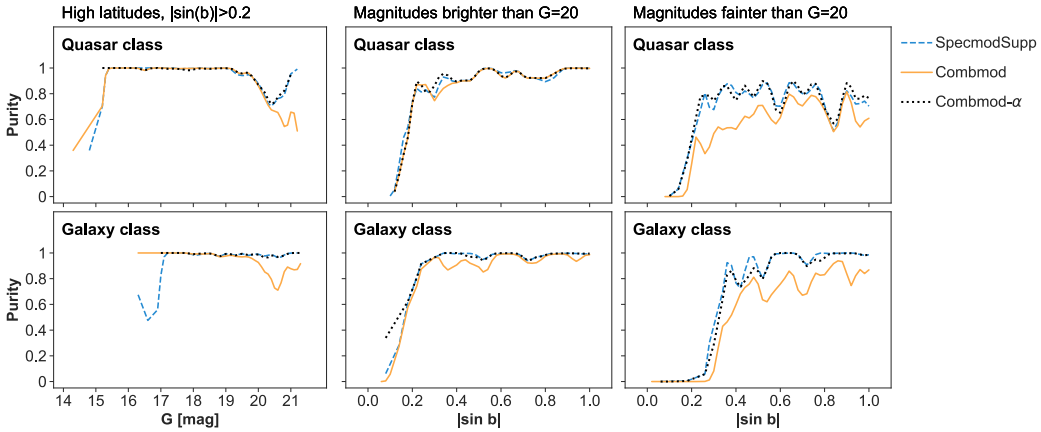
decreases while purity increases. The results of the different classifiers show `Combmod` as the best-performing classifier in terms of completeness of the extragalactic classes. The fractions of unclassified sources (i.e. sources with probabilities below the threshold) in each class are comparable in the other classifiers, which indicates a smoother probability distribution than that of Combmod, where probabilities are close to unity. Not applying a threshold to the maximum probability is equivalent to setting the threshold to zero.

In this study, we assess the performance of the DSC without applying a probability threshold, thereby maintaining the highest possible completeness of the extragalactic classes. The overall performance, as evaluated on the test set, is summarised in Table 1. `Allosmod` achieves high completenesses of 87% and 77% and purities of 73% and 70% for quasars and galaxies, respectively. `Specmod` performs less well, with completenesses of 16 percentage points (pp)[4] and 5 pp lower, and purities of 5 pp and 9 pp lower, for quasars and galaxies, respectively. `Specmod-Supp` has a similar overall performance to Allosmod,

[4] Percentage points, here noted as pp, are equal to the arithmetic difference between two percentages.




(a)




(b)

**Fig. 4.** Variation in performance as a function of magnitude for sources at high latitudes as a function of absolute Galactic latitude for two brightness ranges (middle and right) for each class (rows). The one-dimensional plots are a marginalisation of the two-dimensional representations in Figure 3. *Top*: completeness. *Bottom*: purity. The left panels show completeness and purity as functions of magnitude, averaged over latitudes with $|\sin b| > 0.2$. The middle and right panels show completeness and purity as functions of latitude, averaged over magnitudes brighter than $G$=20 and fainter than $G$=20, respectively.

the largest differences being a 5 pp decrease in completeness and a 7 pp increase in purity for the galaxy class. Combmod achieves the highest completeness among the models, at 97% for quasars and 94% for galaxies, but at the cost of very low purity (37%) for galaxies. `Combmod-`$\alpha$ and `Specmod-Supp` show the best compromise in terms of the geometric score defined by Equation (2b).

In GDR3, `Specmod` achieved low completeness for the quasars and low purity for the extragalactic classes, whereas `Allosmod` and `Combmod` achieved high completeness but low purity (Delchambre et al. 2023). Compared to these results, we now achieve an increase in quasar completeness in all classifiers but a decrease in the galaxy completeness. However, purity has significantly improved in all classifiers for both classes, which is one of the main goals of this work. The percentage points in Table 1 compare these results to those in GDR3 on the same test set. For this purpose, we mapped the three GDR3 stellar classes into one star class. The highest increase in completeness for quasars is achieved by `Specmod` as a result of changing the algorithm from Extratrees to an MLP. The highest increase in purity is 25 pp for quasars by `Specmod` and 27 pp for galaxies by `Allosmod`. Despite the reduced completeness of the galaxy class in Combmod-$\alpha$, the performance for the combined classifier shows a large increase in purity for both classes.

The analysis of the feature importance, described in Appendix C, shows that the most informative features are proper motions, astrometric excess noise, the colour `bp_g`, relative variability in the $G$ band, and the colour excess factor. Moreover, the analysis of each BP/RP pixel input to `Specmod` shows that feature importance is distributed across a range of wavelengths, reflecting the model's reliance on diverse spectral signatures to identify the source type.

#### 4.1.3. Discussion on calibration

The high completeness and low purity of `Combmod` using the same thresholds as the other classifiers may suggest that its outputs do not behave as well-calibrated probabilities and should perhaps instead be interpreted more as confidence-like scores. Interestingly, the `Combmod-`$\alpha$ output is a weighted sum of two probabilities, so it cannot be strictly interpreted as a probability, and yet this does not exhibit the same significant drop in purity seen in `Combmod`. We note that some of our models take as inputs both the BP/RP spectra and the colours formed from the integrated BP and RP fluxes, or they combine probabilities from other models that use these inputs. Those models, therefore, have inputs that are not strictly independent. This may lead to overconfident outputs that are not strictly posterior probabilities.

**Table 2.** Summary of the 2D classification performance evaluated on the test dataset at different magnitude limits for the DSC models trained on *Gaia* data only.

| Magnitude bins | Classifiers | Weighted average 2D completeness | | Weighted average 2D purity | | Geometric score |
|---|---|---|---|---|---|---|
| | | Quasar | Galaxy | Quasar | Galaxy | |
| $G$<19 ($q$:22 557, $g$:2217) | Specmod | 0.959 | 0.890 | 0.998 | 0.983 | 0.957 |
| | Allosmod | 0.981 | 0.849 | 0.996 | 0.990 | 0.952 |
| | Specmod-Supp | 0.975 | 0.894 | 0.998 | 0.986 | 0.962 |
| | Combmod | 0.996 | 0.920 | 0.998 | 0.978 | 0.972 |
| | Combmod-$\alpha$ | 0.987 | 0.878 | 0.996 | 0.988 | 0.961 |
| 19≤$G$<20 ($q$:47 868, $g$:11 415) | Specmod | 0.841 | 0.884 | 0.953 | 0.986 | 0.914 |
| | Allosmod | 0.956 | 0.880 | 0.955 | 0.986 | 0.943 |
| | Specmod-Supp | 0.961 | 0.878 | 0.943 | 0.988 | 0.942 |
| | Combmod | 0.993 | 0.974 | 0.947 | 0.961 | 0.969 |
| | Combmod-$\alpha$ | 0.957 | 0.895 | 0.949 | 0.988 | 0.947 |
| $G$<20 ($q$:70 425, $g$:13 632) | Specmod | 0.879 | 0.885 | 0.969 | 0.986 | 0.929 |
| | Allosmod | 0.964 | 0.875 | 0.968 | 0.987 | 0.947 |
| | Specmod-Supp | 0.965 | 0.881 | 0.961 | 0.987 | 0.948 |
| | Combmod | 0.994 | 0.965 | 0.963 | 0.963 | 0.971 |
| | Combmod-$\alpha$ | 0.966 | 0.893 | 0.964 | 0.988 | 0.952 |
| 20≤$G$<20.5 ($q$:38 027, $g$:6824) | Specmod | 0.633 | 0.550 | 0.811 | 0.957 | 0.721 |
| | Allosmod | 0.885 | 0.701 | 0.750 | 0.964 | 0.818 |
| | Specmod-Supp | 0.878 | 0.557 | 0.739 | 0.982 | 0.772 |
| | Combmod | 0.982 | 0.958 | 0.711 | 0.832 | 0.864 |
| | Combmod-$\alpha$ | 0.855 | 0.646 | 0.762 | 0.972 | 0.800 |
| $G$≥20 ($q$:64 898, $g$:8815) | Specmod | 0.510 | 0.469 | 0.820 | 0.955 | 0.658 |
| | Allosmod | 0.765 | 0.617 | 0.750 | 0.964 | 0.764 |
| | Specmod-Supp | 0.750 | 0.479 | 0.753 | 0.980 | 0.718 |
| | Combmod | 0.949 | 0.902 | 0.670 | 0.821 | 0.828 |
| | Combmod-$\alpha$ | 0.710 | 0.560 | 0.772 | 0.971 | 0.739 |
| $G$≥20.5 ($q$:26 871, $g$:1991) | Specmod | 0.337 | 0.190 | 0.845 | 0.936 | 0.474 |
| | Allosmod | 0.596 | 0.326 | 0.750 | 0.959 | 0.611 |
| | Specmod-Supp | 0.569 | 0.215 | 0.785 | 0.961 | 0.551 |
| | Combmod | 0.902 | 0.713 | 0.607 | 0.770 | 0.740 |
| | Combmod-$\alpha$ | 0.504 | 0.267 | 0.799 | 0.965 | 0.568 |

**Notes.** Same as Table 1.

We emphasise that the combined classifiers are primarily intended as an aggregation tool to optimise performance, rather than as a rigorous probabilistic merger. We nonetheless chose to retain the term 'probabilities' for their outputs to remain consistent with the mathematical framework of the DSC pipeline. As we do not implement any calibration on the DSC outputs, a comprehensive examination of calibration is beyond the scope of this study, although we will discuss it in a technical note that will accompany the data release. The thresholding approach for the DSC outputs selects the class with the highest probability above a defined limit. We recommend that users consider these values carefully when conducting studies that require precise probability calibration.

#### 4.1.4. Magnitude and latitude-based assessment of the performance

As noted also in Hughes et al. (2022) and Jamal & Bailer-Jones (2024), the summary performance evaluated on the entire test set does not fully reflect a classifier's strengths and weaknesses. We therefore evaluate the classification as a function of the sources' brightness and Galactic latitude. We binned the test data on a 2D grid of 4550 bins in $|\sin b| \in [0,1]$ and $G \in [5,21]$, with coarser binning in the bright magnitude regime of $G$<14. In each bin, we adjusted the confusion matrix to account for class imbalance and compute completeness and purity.

Figure 3 shows this 2D completeness and purity for the quasar and galaxy classes. We use grey for undefined values, light colours for high values, and black for very small to zero values. Undefined and zero values can be explained by Eq. (2a). Zero completeness occurs when TP is zero, but TN is not, meaning that all objects in the class are misclassified. Conversely, zero purity refers to zero TP but non-zero FP, as the classifier only predicts false detections. An undefined completeness occurs when TP and TN are both null, meaning that there are no objects from the class to predict. An undefined purity occurs when TP and FP are both null, meaning that the classifier does not introduce any contamination (FP null) nor any correct prediction (TP null). The performance is unaffected when the metrics are undefined, unlike zero values, which indicate very poor classifier performance.

**Table 3.** Summary of the 2D classification performance evaluated on the test dataset at different magnitude limits for the DSC models trained on *Gaia* data and mid-infrared photometry from CatWISE.

| Magnitude bins | Classifiers | Weighted average 2D completeness | | Weighted average 2D purity | | Geometric score |
|---|---|---|---|---|---|---|
| | | Quasar | Galaxy | Quasar | Galaxy | |
| $G$<19 ($q$:20 529, $g$:1743) | Specmod | 0.959 (–) | 0.956 (–) | 0.999 (–) | 0.987 (–) | 0.975 (–) |
| | Allosmod(*) | 0.985 (–) | 0.963 (+5 pp) | 0.998 (–) | 0.987 (–) | 0.983 (+1 pp) |
| | SpecmodSupp(*) | 0.983 (+1 pp) | 0.950 (−1 pp) | 0.998 (–) | 0.994 (+1 pp) | 0.981 (–) |
| | Combmod(*) | 0.997 (–) | 0.977 (+1 pp) | 0.998 (–) | 0.984 (–) | 0.989 (–) |
| | Combmod-$\alpha$(*) | 0.991 (–) | 0.964 (+2 pp) | 0.999 (–) | 0.989 (–) | 0.986 (+1 pp) |
| 19≤$G$<20 ($q$:42 290, $g$:8857) | Specmod | 0.843 (–) | 0.900 (–) | 0.965 (–) | 0.994 (–) | 0.924 (–) |
| | Allosmod(*) | 0.979 (+2 pp) | 0.924 (+3 pp) | 0.952 (−2 pp) | 0.991 (–) | 0.961 (+1 pp) |
| | SpecmodSupp(*) | 0.969 (+1 pp) | 0.951 (+6 pp) | 0.961 (–) | 0.991 (–) | 0.968 (+2 pp) |
| | Combmod(*) | 0.996 (–) | 0.989 (–) | 0.960 (–) | 0.974 (–) | 0.980 (–) |
| | Combmod-$\alpha$(*) | 0.980 (+2 pp) | 0.923 (+1 pp) | 0.961 (–) | 0.991 (–) | 0.963 (+1 pp) |
| $G$<20 ($q$:62 819, $g$:10 600) | Specmod | 0.881 (–) | 0.910 (–) | 0.977 (–) | 0.992 (–) | 0.939 (–) |
| | Allosmod(*) | 0.981 (+2 pp) | 0.930 (+4 pp) | 0.967 (−1 pp) | 0.991 (–) | 0.967 (+1 pp) |
| | SpecmodSupp(*) | 0.974 (+1 pp) | 0.951 (+5 pp) | 0.973 (–) | 0.992 (–) | 0.972 (+1 pp) |
| | Combmod(*) | 0.996 (–) | 0.987 (–) | 0.973 (–) | 0.975 (–) | 0.983 (–) |
| | Combmod-$\alpha$(*) | 0.984 (+2 pp) | 0.929 (+1 pp) | 0.973 (–) | 0.991 (–) | 0.969 (+1 pp) |
| 20≤$G$<20.5 ($q$:32 893, $g$:4990) | Specmod | 0.636 (–) | 0.544 (–) | 0.862 (–) | 0.972 (–) | 0.734 (–) |
| | Allosmod(*) | 0.959 (+7 pp) | 0.717 (+2 pp) | 0.837 (+1 pp) | 0.983 (–) | 0.867 (+2 pp) |
| | SpecmodSupp(*) | 0.952 (+7 pp) | 0.800 (+25 pp) | 0.851 (+2 pp) | 0.959 (−3 pp) | 0.888 (+10 pp) |
| | Combmod(*) | 0.992 (+1 pp) | 0.970 (–) | 0.804 (–) | 0.857 (−1 pp) | 0.902 (–) |
| | Combmod-$\alpha$(*) | 0.949 (+9 pp) | 0.650 (+1 pp) | 0.826 (−1 pp) | 0.993 (–) | 0.843 (+2 pp) |
| $G$≥20 ($q$:55 687, $g$:6231) | Specmod | 0.514 (–) | 0.471 (–) | 0.866 (–) | 0.968 (–) | 0.671 (–) |
| | Allosmod(*) | 0.919 (+15 pp) | 0.628 (+1 pp) | 0.797 (−2 pp) | 0.981 (–) | 0.820 (+3 pp) |
| | SpecmodSupp(*) | 0.915 (+16 pp) | 0.692 (+22 pp) | 0.809 (−2 pp) | 0.959 (−3 pp) | 0.837 (+10% |
| | Combmod(*) | 0.976 (+3 pp) | 0.916 (–) | 0.746 (−2 pp) | 0.845 (−2 pp) | 0.866 (–) |
| | Combmod-$\alpha$(*) | 0.884 (+17 pp) | 0.563 (–) | 0.797 (−4 pp) | 0.990 (–) | 0.792 (+3 pp) |
| $G$≥20.5 ($q$:22 794, $g$:1241) | Specmod | 0.338 (–) | 0.177 (–) | 0.876 (–) | 0.931 (–) | 0.470 (–) |
| | Allosmod(*) | 0.862 (+27 pp) | 0.268 (−1 pp) | 0.733 (−7 pp) | 0.959 (–) | 0.635 (+4 pp) |
| | SpecmodSupp(*) | 0.862 (+29 pp) | 0.259 (+8 pp) | 0.743 (−9 pp) | 0.962 (–) | 0.632 (+10 pp) |
| | Combmod(*) | 0.954 (+5 pp) | 0.699 (–) | 0.658 (−5 pp) | 0.786 (−2 pp) | 0.766 (−1 pp) |
| | Combmod-$\alpha$(*) | 0.789 (+29 pp) | 0.213 (−2 pp) | 0.747 (−9 pp) | 0.959 (–) | 0.589 (+4 pp) |

**Notes.** *(*)* indicates the models trained with *Gaia* data and infrared photometry from CatWISE. The numbers in pp are equal to the difference in performance between models trained with *Gaia* data and CatWISE, and models trained only with *Gaia* data, evaluated on the exact test baseline of sources with valid infrared data. The symbol *(-)* indicates no change. Colour code is similar to Table 1.

The performance decreases towards fainter magnitudes in Figure 3, as we would expect. For Specmod, the quasar class completeness begins to decrease at magnitudes fainter than $G$=19, whereas Allosmod and Specmod-Supp show a decrease only from $G$≥20. This result indicates a limitation of Specmod, which uses only the XP spectra to predict a label, and these spectra will generally be noisier than the features used by Allosmod. The galaxy completeness remains comparable over the three baseline classifiers, with a continuous decrease from $G$≥20. The purities are high overall, with contamination mostly localised at magnitudes fainter than $G$=20. Across all magnitudes, Combmod achieves the highest completenesses amongst the classifiers, but the lowest purities, particularly for the galaxy class. Combmod-$\alpha$ achieves lower completeness but a higher gain in purity, compared to Combmod.

Figure 4 illustrates how these metrics change with latitude across two distinct magnitude ranges. We compute these one-dimensional representations by averaging the two-dimensional representations along one axis. In Figure 4, the completeness at $G$<20 mag is high for all classifiers when far from the Galactic plane. At magnitudes fainter than $G$=20 mag, however, the completeness and purities decrease. Purity improves at higher latitudes where there is less dust extinction, crowding and stellar density, with the best purities reached by Allosmod, Combmod-$\alpha$ and Specmod-Supp. Overall, the completeness of the quasar class remains higher compared to that of the galaxy class. As previously stated, Combmod achieves the highest completeness amongst the classifiers, but the lowest purity for the galaxy class.

To summarise the performance across magnitude and latitude, we compute the average metrics, weighted by the denominators in Eq. (2a), i.e. the number of true sources per bin for completeness and the number of predictions per bin for purity. We emphasise that since the adjustment only affects purity, the average value does not correspond to the global metric, unlike the average completeness. Therefore, the adjustment made to the overall metrics cannot be compared with the per-bin adjustments

**Table 4.** Quasar and galaxy candidates classified by the DSC across all magnitudes and latitudes but excluding the LMC and SMC.

| | Classifier | Quasar candidates | | | | | Galaxy candidates | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Total | G<20 | 20≤*G*<20.5 | 20.5≤*G*<21 | *G*≥21 | Total | G<20 | 20≤*G*<20.5 | 20.5≤*G*<21 | *G*≥21 |
| (a) | `Specmod` | 2 108 425 | 40 (32) | 22 (18) | 18 (12) | 20 (1) | 1 885 485 | 22 (17) | 20 (18) | 21 (12) | 38 (11) |
| | `Allosmod` | 1 692 662 | 46 (43) | 33 (32) | 20 (19) | 0.4 (0.1) | 623 087 | 52 (45) | 35 (30) | 10.6 (7) | 3 (1) |
| | `Specmod-Supp` | 1 651 764 | 51 (45) | 32 (30) | 17 (16) | 0.1 (0.03) | 551 551 | 59 (54) | 33 (31) | 6 (5) | 2 (0.8) |
| | `Combmod` | 3 010 203 | 29 (26) | 23 (22) | 33 (27) | 15 (1) | 2 302 881 | 19 (16) | 29 (26) | 27 (19) | 25 (4) |
| | `Combmod-`$\alpha$ | 2 021 603 | 39 (36) | 26 (24) | 16 (11) | 19 (0.1) | 1 326 305 | 27 (22) | 19 (16) | 14 (4) | 40 (4) |
| | `Combmod-Supp` | 2 156 151 | 40 (34) | 25 (23) | 17 (13) | 18 (0.1) | 1 236 203 | 29 (24) | 16 (14) | 13 (3) | 42 (4) |
| (b) | `Specmod` | 1 084 516 | 53 (51) | 28 (27) | 16 (15) | 3 (1) | 763 628 | 30 (29) | 28 (27) | 19 (18) | 23 (18) |
| | `Allosmod(*)` | 1 726 112 | 39 (37) | 30 (29) | 30 (28) | 0.8 (0.4) | 386 739 | 56 (54) | 37 (35) | 6 (6) | 0.8 (0.6) |
| | `Specmod-Supp(*)` | 1 755 423 | 38 (37) | 30 (28) | 31 (29) | 1 (0.6) | 413 421 | 52 (49) | 39 (38) | 8 (8) | 1 (1) |
| | `Combmod` | 1 995 096 | 35 (33) | 28 (26) | 35 (32) | 3 (1) | 1 196 593 | 22 (21) | 38 (37) | 32 (30) | 8 (5) |
| | `Combmod-`$\alpha$ | 1 600 652 | 41 (40) | 31 (30) | 26 (25) | 2 (0.3) | 440 195 | 50 (48) | 33 (31) | 6 (5) | 12 (6) |
| | `Combmod-Supp` | 1 782 062 | 38 (36) | 29 (28) | 31 (28) | 2 (1) | 468 333 | 46 (44) | 35 (33) | 8 (7) | 12 (6) |
| (c) | `Allosmod(*)` | 420 567 | 10 (9) | 20 (18) | 67 (60) | 3 (1) | 79 638 | 31 (29) | 55 (51) | 12 (10) | 3 (2) |
| | `Specmod-Supp(*)` | 540 010 | 10 (8) | 23 (20) | 64 (56) | 4 (2) | 141 003 | 18 (15) | 60 (58) | 18 (17) | 4 (3) |
| | `Combmod` | 150 774 | 6 (5) | 12 (10) | 71 (58) | 11 (5) | 125 606 | 3 (2) | 14 (13) | 68 (63) | 15 (14) |
| | `Combmod-`$\alpha$ | 366 773 | 10 (9) | 24 (22) | 63 (59) | 2 (1) | 44 283 | 26 (23) | 58 (54) | 13 (12) | 4 (3) |

**Notes.** The symbol *(*)* indicates models trained using *Gaia* data and infrared photometry from CatWISE. Rows (*a*), (*b*), and (*c*) refer to candidates identified in *Gaia* only mode, *Gaia*-IR mode or exclusive to *Gaia*-IR mode, respectively. Detections exclusive to *Gaia*-IR mode refer to the candidates identified only in this mode and not in the *Gaia* only mode. The percentages of sources in each magnitude bin are shown in the columns. The numbers in parentheses refer to the percentage of sources in the magnitude bin at higher galactic latitudes $|\sin b|>0.2$ after the application of the photometric quality cut `phot_[bp;rp]_n_obs`≥10. The number of sources in each magnitude bin is equal to the percentage of sources in that bin multiplied by the total number of sources. `Combmod-Supp(*)` predictions identified exclusively in the *Gaia*-IR mode are equal to the candidates from `Specmod-Supp(*)`, as the data required for the classifier is available.

applied to the 2D metrics. The average metrics for different magnitude ranges are reported in Table 2. The classifiers perform best at magnitudes brighter than $G$=20, achieving an average 2D completeness of 87–99% and an average 2D purity of 96–98% for the extragalactic classes. At magnitudes of 20<$G$≤20.5, purity decreases rapidly, demonstrating the classifiers' inability to reject the false positives. Also, while `Allosmod` and the combined classifiers maintain a 2D completeness of at least 75% for quasars, it drops considerably for the galaxies.At magnitudes fainter than $G$=20.5, both completeness and purity are significantly reduced. The results also show that `Specmod` performance is more limited at faint magnitudes than that of the other classifiers, indicating a limited use of XP spectra for reliable classifications at such magnitudes. The results also show that, unlike `Combmod`, which prioritises completeness, `Combmod-`$\alpha$ and `Specmod-Supp` improve purity, particularly at $G$≥20. At magnitudes of 20<$G$≤20.5, the latter two classifiers increase purity by 5 and 15 pp, but reduce completeness by 13 and 50 pp, for quasars and galaxies, respectively. At magnitudes fainter than $G$=20.5, they largely reduce stellar contamination by 18 to 20 pp at the cost of missing out on true classes, especially galaxies.

### 4.2. Augmented DSC predictions with IR data

We trained separate classifiers using mid-infrared CatWISE photometry and *Gaia* data. We supplemented the inputs with the mid-infrared photometry bands $W1$ and $W2$ to train the `Allosmod`(*) and `Specmod-Supp`(*) classifiers (see Section 3). The other *Gaia* inputs were the same as before (see Appendix A). The `Specmod` classifier remained unchanged (no IR inputs). We combined the posterior probabilities of these three classifiers, as done before, to compute new combined *Gaia*-IR classifiers. As shown in Tables 2 and 3, models in *Gaia*-IR mode used a different training and test dataset composed of sources with valid CatWISE photometry than models in *Gaia* only mode.

The performance on the infrared-augmented test set at different magnitude limits is provided in Table 3. The percentage points provide a direct comparison between the *Gaia*-IR and *Gaia*-only modes on the same test set that requires valid infrared photometry. The classification results demonstrate that combining the *Gaia* optical data with mid-IR data from CatWISE improves the completeness of quasars and galaxies at the expense of a slight reduction in purity. The improvement in completeness at bright magnitudes $G$<20 reaches a maximum increase of 5 pp. This effect is more significant at fainter magnitudes, where the increase in completeness is substantial at the slight expense of a reduction in purity by 1 to 9 pp. Here, the maximum increase in completeness reaches 25 pp for galaxies at magnitudes 20≤$G$<20.5 by `Specmod-Supp`, and 29 pp for quasars at $G$≥20.5 by `Specmod-Supp` and `Combmod-`$\alpha$. The analysis of feature importance for `Allosmod`(*) in Appendix C confirms that the infrared colour $W1$-$W2$ is a key feature for the completeness of quasars, alongside the *Gaia* colours, proper motions and the astrometric excess noise. To summarise, adding the IR photometry to the *Gaia* data significantly improves DSC completeness, primarily for faint quasars at magnitudes $G$≥20.

### 4.3. Application of DSC to all 2.8 billion sources in GDR4

#### 4.3.1. Quasar and galaxy candidates

We apply the DSC to the entire *Gaia* data and report the anticipated classification statistics for GDR4. Table 4 provides the total number of quasar and galaxy candidates, excluding sources

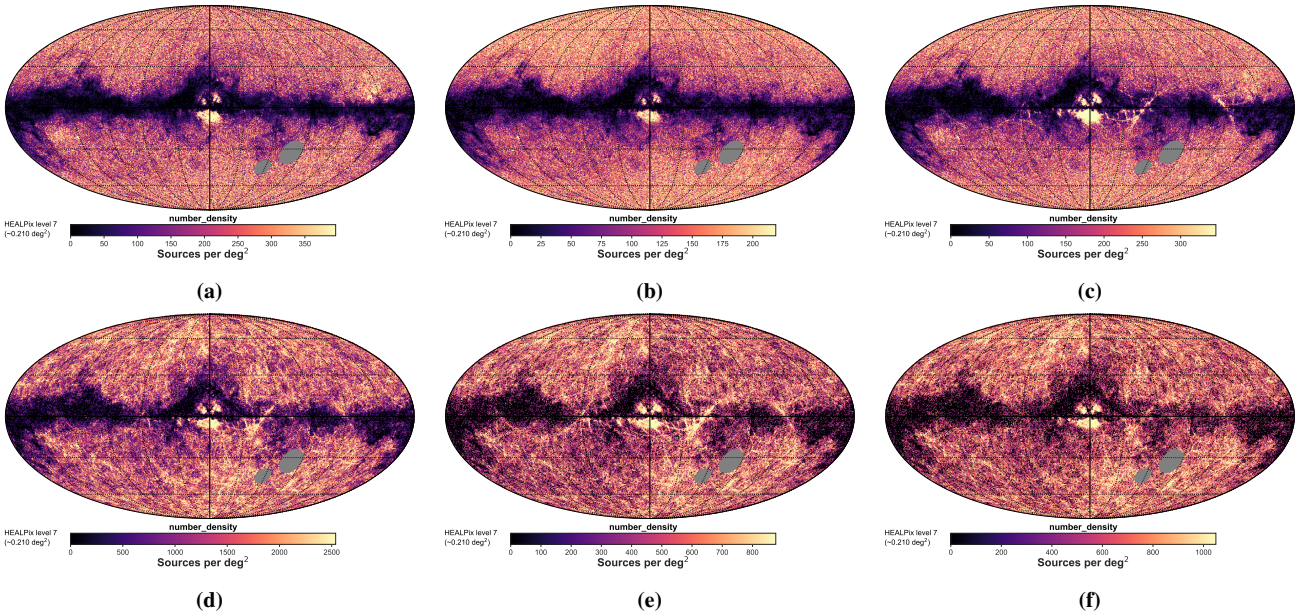



**Fig. 5.** Galactic sky distribution of sources at magnitudes $G$<20.5 classified from maximum probabilities by the DSC combined classifiers using the global prior at HEALpixel level 7 in Mollweide projection. *Top*: quasars. *Bottom*: galaxies. The colour map uses bright colours for high-density regions, while darker colours refer to regions with fewer observations. The Magellanic Clouds (LMC and SMC) are masked in grey. `Combmod` identifies 1 568 861 quasars (*a*) and 1 114 782 galaxies (*d*). `Combmod-`$\alpha$ identifies 1 313 150 quasars (*b*) and 608 781 galaxies (*e*). `Combmod-Supp` identifies 1 400 549 quasars (*c*) and 550 413 galaxies (*f*).

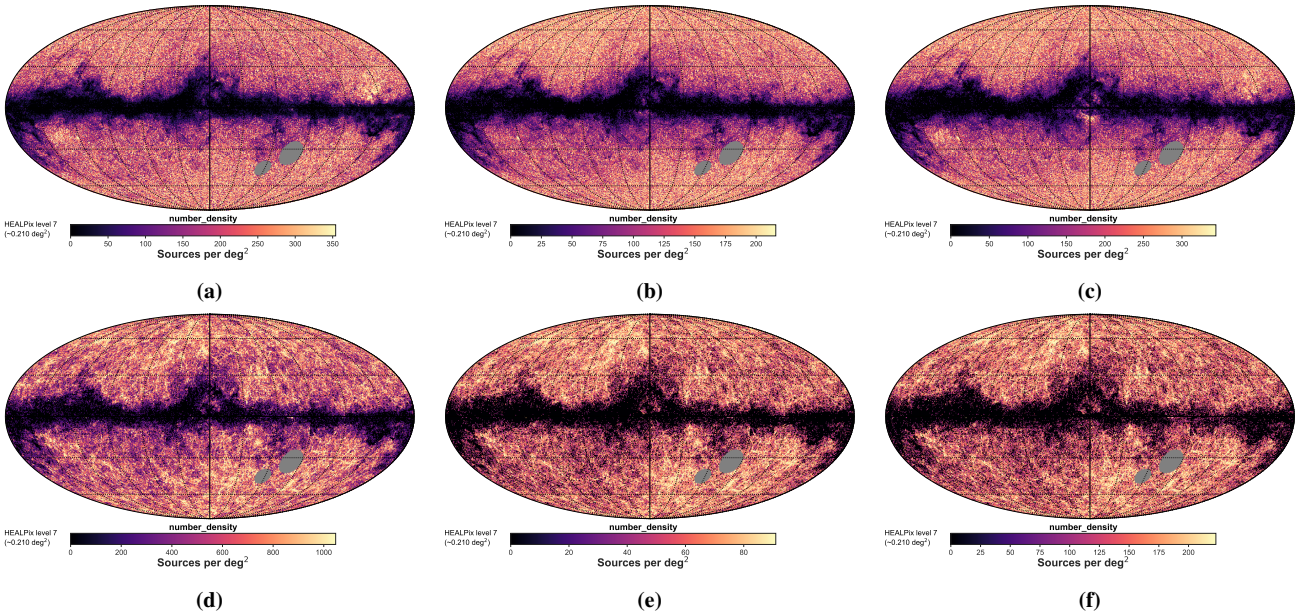
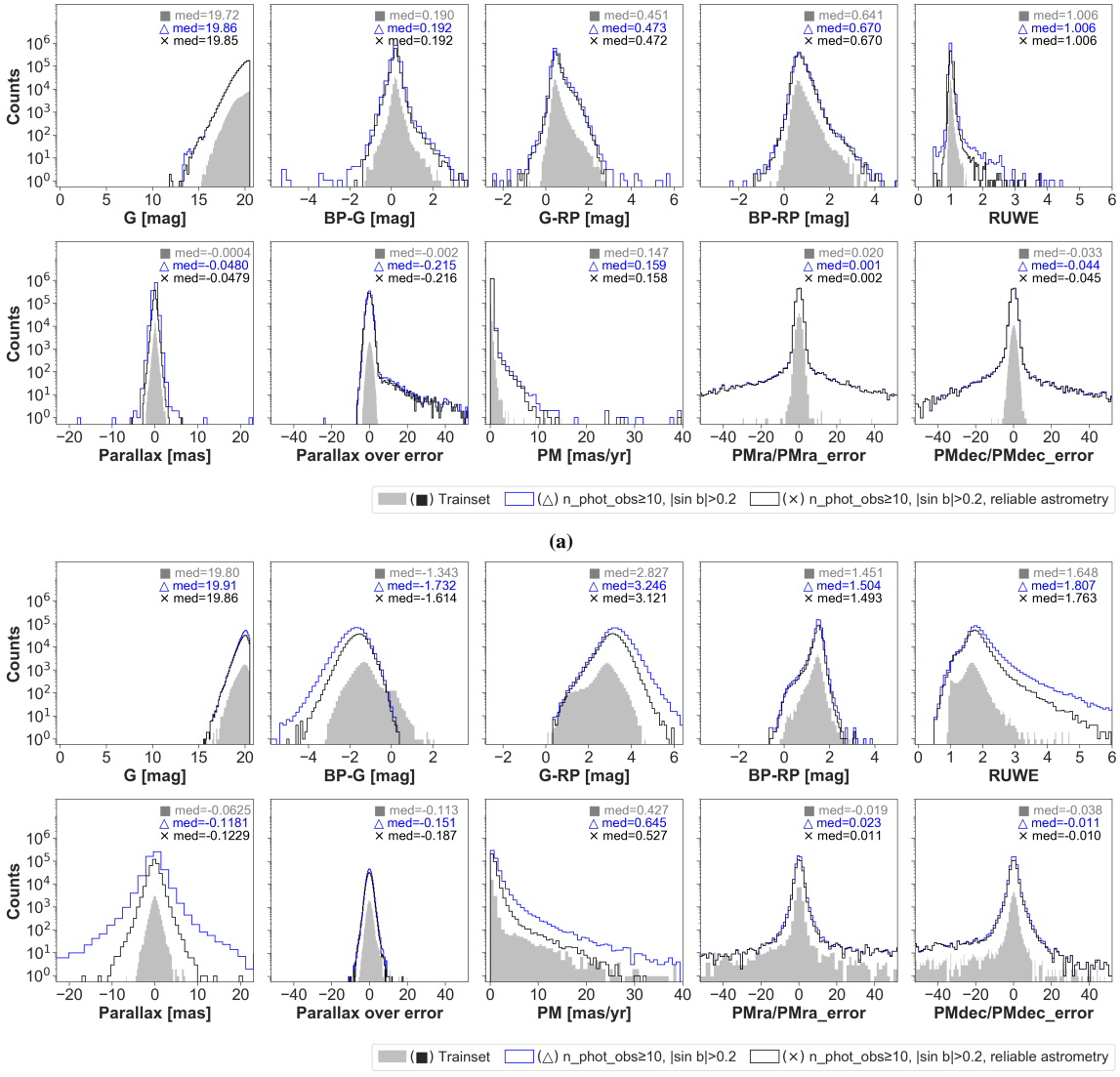



**Fig. 6.** Same as Figure 5 but with a quality cut imposed on the number of observations contributing to the integrated photometry in BP and RP `phot_[bp,rp]_n_obs`≥10. *Top*: quasars. *Bottom*: galaxies. `Combmod` identifies 1 513 811 quasars (*a*) and 1 031 405 galaxies (*d*). `Combmod-`$\alpha$ identifies 1 268 369 quasars (*b*) and 526 707 galaxies (*e*). `Combmod-Supp` identifies 1 294 542 quasars (*c*) and 494 717 galaxies (*f*).

in the Magellanic Clouds. Recall that we assign the class that achieves the highest probability, no matter how large or small that probability is. Candidates located in the LMC and SMC comprise about 2–6% of the total extragalactic predictions, except for `Specmod` and `Combmod`, which account for 15% and 9% of extragalactic predictions, respectively. We expect most classifications in dense regions to be false detections. Therefore, we excluded the LMC and SMC from our assessment.

Figure 5 shows the Galactic sky distributions of the quasar and galaxy candidates of the combined classifiers, namely `Combmod`, `Combmod-`$\alpha$ and `Combmod-Supp`. Due to stellar contamination, we expect to see Galactic features in the extragalactic maps. If the contamination rate were constant across the sky, the pattern of the Galaxy would be visible in the number of contaminants (i.e. stars), even when added to a flat distribution of true detections. The sky distributions show a pattern in latitude that follows regions of high dust extinction and high stellar density, as observed by *Gaia*. In particular, we observe a depletion of candidates along the Galactic plane far from the bulge but a higher density in the bulge due to stellar contamination. False extragalactic detections result from a limitation of the classifier or from a reduced quality of the




**(a)**




**(b)**

**Fig. 7.** Feature distributions of extragalactic candidates identified in the *Gaia* only mode by `Combmod-`$\alpha$ using the global prior at magnitudes $G$<20.5 and higher latitudes $|\sin b|$>0.20. Sources located in the LMC and SMC are excluded. (*a*) Quasar candidates. (*b*) Galaxy candidates. The quality cut was applied to *Gaia* photometry and astrometry. The photometric quality cut requires a minimum of ten photometric observations in the BP- and RP- bands (`phot_[bp,rp]_n_obs`≥10). The astrometric quality cut uses a Boolean flag, `unreliable_astrometry`, in the GDR4 table `all_source_astrometry`. The flag is set to `'false'` to indicate reliable astrometric measurements. The training dataset of the *Gaia* only mode is shown in grey and is associated with the square symbol ■. The list of candidates after the photometric quality cut is displayed in blue and is associated with the triangle symbol △. Lastly, the list of candidates after the photometric and astrometric quality cut is displayed in black and is associated with the cross symbol ×. The median value for each dataset is indicated. The term RUWE stands for renormalised unit weight error. The terms PMra and PMdec refer to the proper motions in the right ascension and declination, respectively. In (*a*), the classifier identifies as quasars, 1 215 830 sources (△) and 1 207 285 sources (×). In (*b*), the classifier identifies as galaxies, 502 209 sources (△) and 344 184 sources (×).

astrometry and photometry. Far from the Galactic plane, the overall sky distribution of quasars appears smoother, with no peculiar aggregates in distinct regions. For the galaxies, however, a clustering pattern is noticeable across the whole sky. This pattern is not new and can be observed in the GDR3 results (Delchambre et al. 2023; Jamal & Bailer-Jones 2024).

There are regions of the sky that have been observed only a few times by *Gaia*, resulting in few photometric observations or transits and noisy measurements. To filter out peculiar patterns, we impose a quality cut on the number of observations contributing to the integrated photometry in the BP and RP bands. Setting a minimum number of photometric observations to `phot_[bp,rp]_n_obs` =10 significantly reduces the densities at the bulge and the scanning law patterns, as shown in Figure 6. The Appendix Figures E.1–E.2 show the Galactic sky distributions of the quasar and galaxy candidates at different magnitude limits for `Combmod-`$\alpha$. The pattern from the scanning law, which is detectable at all magnitude limits, is reduced by the quality cut. A large number of sources are filtered at fainter magnitudes $G$≥20.5. The density at the bulge is also filtered out when a higher threshold on the number of photometric observations is imposed. At magnitudes fainter than $G$=21, the quality cut rejects most of the noise, but the predictions are still affected by the scanning law and crowding near the Galactic plane. The DSC results in the GDR4 astrophysical parameters tables (`ap_class` and `ap_class_ir`) are not subject to the quality cut, enabling users to apply custom filters when querying the *Gaia* catalogue. However, as a conservative approach, a

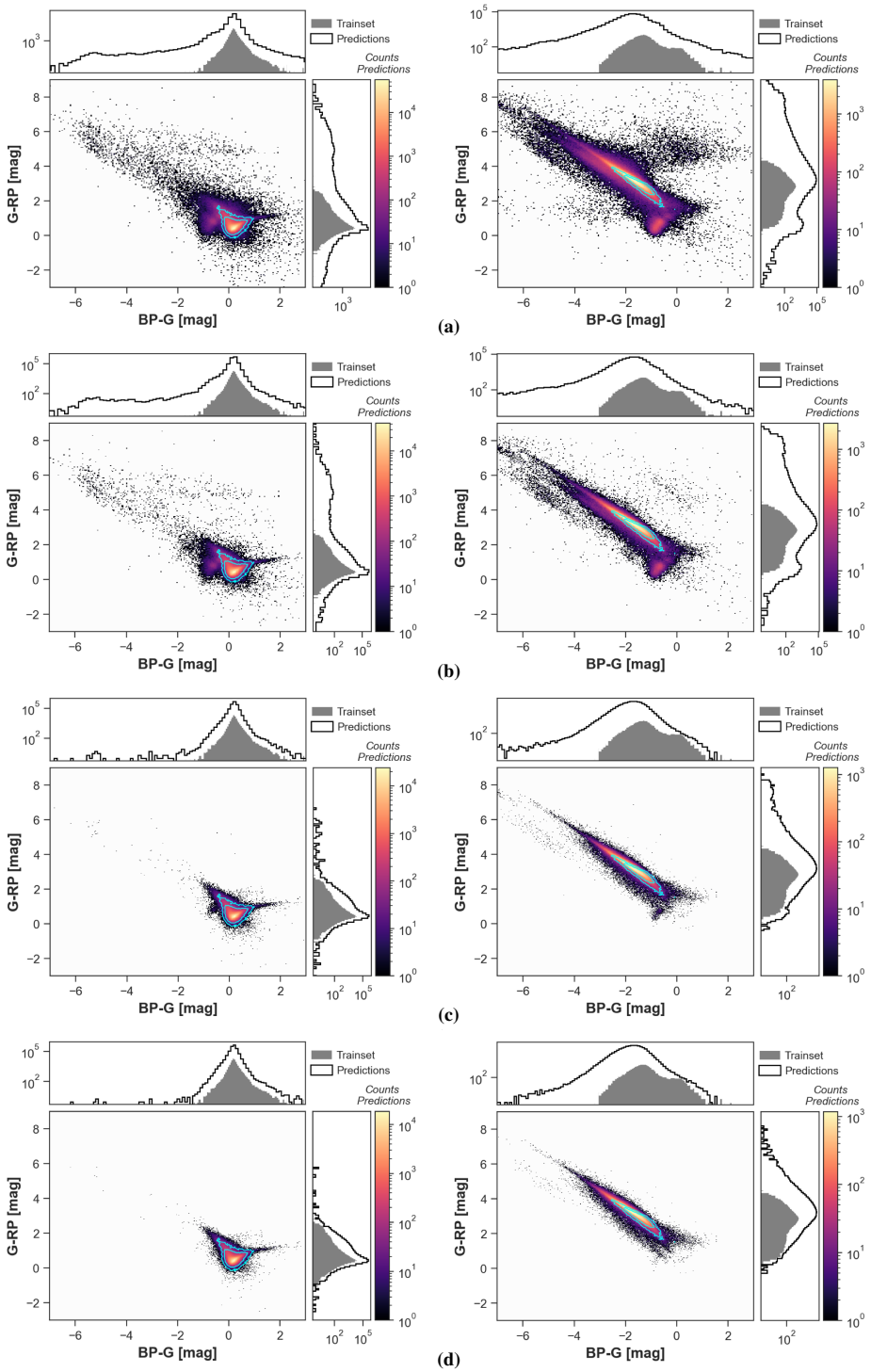



**Fig. 8.** Colour-colour distributions of sources classified by `Combmod-`$\alpha$ using the global prior at magnitudes of $G$<20.5. *Left*: quasars. *Right*: galaxies. Sources located in the LMC and SMC are excluded. Contours (cyan lines) show the normalised density on a log scale of the highest-density regions for the trainset. In *(a)*, the classifier identifies 1 313 150 quasars and 608 781 galaxies across the whole sky. In *(b)*, the classifier identifies 1 234 753 quasars and 531 634 galaxies at higher latitudes $|\sin b|$>0.20. In *(c)*, the classifier identifies 1 218 707 quasars and 504 906 galaxies at higher latitudes $|\sin b|$>0.20 and `phot_[bp,rp]_n_obs`≥6. In *(d)*, the classifier identifies 1 215 830 quasars and 502 209 galaxies at higher latitudes $|\sin b|$>0.20 and `phot_[bp,rp]_n_obs`≥10.

quality cut of `phot_[bp,rp]_n_obs`≥6 is applied to the DSC `Combmod` results reported in the GDR4 main source catalogue (`gaia_source`).

Figures 7 and 8 show the feature distributions of the quasar and galaxy candidates identified by the combined classifier `Combmod-`$\alpha$, at magnitudes $G$<20.5. The feature distribution is overlaid with our training data. We require a minimal number of photometric observations in the BP- and RP- bands and also use the Boolean flag `unreliable_astrometry` provided in GDR4, set to `'false'` to indicate reliable AGIS (Astrometric Global Iterative Solution) processing. The filtered set of candidates exhibits properties similar to those of the spectroscopically

confirmed quasars and galaxies in our SDSS DR17 training set. However, we observe deviations from the training set in the signal-to-noise ratios of the proper motions and parallaxes (proper motions over errors and parallaxes over errors). Compared to the DSC trainset, the excess of negative values in $G_{BP}$-$G$ and positive values in $G$-$G_{RP}$ indicates the presence of extended sources identified as galaxies. A deviation from our training data can be attributed either to a broader range of sources beyond our controlled training sample or to contamination. As illustrated in Appendix Figure D.1, applying a higher threshold to the probabilities produces cleaner samples. The filtering is evident in Appendix Figure D.2, showing reduced scatter in the astrometry, especially for quasars. We note, however, that even after implementing a quality cut and applying a higher threshold to the probabilities, the galaxy candidates still exhibit an excess of negative values in $G_{BP}$-$G$ and positive values in $G$-$G_{RP}$ compared to our training sample of galaxies, which indicates these extended sources as reliable candidates. The difference with the DSC trainset may be due to a selection bias in the galaxies identified in SDSS.

#### 4.3.2. Extragalactic candidates identified using *Gaia* and CatWISE

In *Gaia*-IR mode, the DSC provides classifications for *Gaia* sources crossmatched to the CatWISE2020 catalogue. Table 4 reports the total number of quasar and galaxy candidates, excluding sources in the LMC and SMC. As explained in Section 4.3.1, contamination is reduced by implementing quality cuts and excluding sources located at the Galactic plane. Appendix Figure F.1 shows the candidates identified exclusively in the *Gaia*-IR. At magnitudes brighter than $G$=21, the colour-colour distribution of the identified quasars and galaxies is similar to that of the extragalactic selection of the training set. However, at magnitudes fainter than $G$=21, the colour-colour distribution reveals redder sources that are most likely galaxies misidentified as quasars. This result may be due to misclassifications between AGN-type quasars and redder galaxies. Nevertheless, this confusion affects only a few sources to such a magnitude. The level of stellar contamination introduced by the addition of infrared photometry for classification is low and can be evident in the deviation from the training set and along the stellar locus in the colour-colour plane at magnitudes $G$<20.5 (panel (*a*) in Appendix Figure F.1). This result can also be seen in the feature distributions in Appendix Figure F.2, where a fraction of sources likely erroneously identified as quasar candidates display bright magnitudes ($G$<12) and higher proper motions.

### *4.4. Recommended selections and caveats*

In this section, we summarise quality cuts recommended to improve the selection of quasar and galaxy candidates selected by the DSC. To reject peculiar patterns due to the scanning law and crowding close to the Galactic bulge, we recommend setting the minimum number of photometric observations to `phot_[bp,rp]_n_obs`≥10 (or 6 for a less restrictive cut). Further rejection of stellar contaminants in the Galactic plane can be achieved by selecting only sources at higher Galactic latitudes $|\sin b|$>0.20. In this work, we allocated classes based simply on the maximum probability. A user could instead allocate a class only if the maximum probability is above some threshold. This approach removes contaminants at the expense of missing true detections. We suggest a probability threshold of $P$>0.50 for a less restrictive cut, while a threshold of $P$>0.90 provides a stricter cut. When available, we also recommend using any reliability flags provided in the GDR4 tables[5] on the *Gaia* astrometry and the photometry (`all_source_astrometry` and `all_source_photometry`). The astrometric reliability flag rejects sources with spurious astrometric measurements that show large negative parallaxes. A technical note on the DSC will show how the astrometric flag correction improves the classifications (Jamal & Bailer-Jones 2026).

Finally, the DSC processes all sources with valid BP and RP spectra across all magnitudes. However, classifications at magnitudes fainter than $G$=21 are highly contaminated by stellar objects due to increased noise in the data. In the bright regime of magnitudes $G$<21, stellar contamination is expected, particularly for the DSC models trained on the *Gaia* XP spectra, `Specmod` and `Specmod-Supp`, as well as the combined classifiers, `Combmod` and `Combmod-Supp`. We recommend setting the magnitude range to exclude magnitudes fainter than $G$=21.

## 5. Conclusions

We have improved purity at the expense of a moderate loss of completeness for the extragalactic classes identified by the DSC in GDR4 compared to GDR3. The best improvement in completeness has been achieved for quasars by the MLP `Specmod` classifier with an increase of 29 percentage points (pp) compared to GDR3 Extratrees `Specmod`, while the other classifiers have shown a moderate gain of only 2 to 5 pp. However, galaxy completeness has decreased, indicating greater difficulty in identifying extended sources. Nevertheless, the largest increase in purity has reached ∼25 pp for both quasars and galaxies. When re-evaluated as a function of magnitude and sky position, the best DSC models have achieved average completeness of 87–99% and purity of 96–98% for the extragalactic classes, at magnitudes brighter than $G$=20. At magnitudes fainter than $G$=20, the performance drops, with an even greater reduction at magnitudes fainter than $G$=20.5. When comparing all DSC models, the `Combmod` classifier prioritises completeness, unlike `Combmod-`$\alpha$ and `Specmod-Supp`, which improve purity, particularly at $G$≥20. We emphasise that the DSC outputs are not calibrated probabilities, and the underlying assumptions of the combined models may lead to overconfidence. But we have shown that thresholds on DSC outputs can be used effectively to identify classes.

For GDR4, we also trained independent models combining *Gaia* optical data with mid-infrared photometry from the CatWISE2020 catalogue for classification. This combination has been shown to improve the completeness of extragalactic data, particularly for quasars, albeit with a slight reduction in purity, primarily at fainter magnitudes. Compared to models trained on *Gaia* data alone, including infrared data has increased quasar completeness by up to 29 pp at magnitudes fainter than $G$=20.5.

Overall, the quasar and galaxy candidates identified in GDR4 by the DSC exhibit physical properties close to those of our controlled training set. Any deviations from the training set are either due to noisy measurements or to classifier limitations that lead to false stellar detections. The DSC candidate selection has improved when selecting sources at higher Galactic latitudes, far from the Galactic plane. Further improvements have been proposed through applying quality cuts to *Gaia* photometry and astrometry. These quality cuts have been shown to effectively reduce stellar contamination and correct for the *Gaia* scanning law pattern. In the *Gaia* IR mode, we have assessed the detected extragalactic candidates, exclusively identified by the DSC when combining *Gaia* data with infrared photometry from

[5] https://www.cosmos.esa.int/web/gaia/dr4

CatWISE, as valid candidates despite the moderate increase in stellar contamination.

*Acknowledgements.* This work has used 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. This publication also uses data products from the Wide-field Infrared Survey Explorer (WISE), a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration. The authors would like to thank the referee for their insightful comments and suggestions, which helped improve the quality of this manuscript. The authors would also like to thank Morgan Fouesneau and René Andrae for useful discussions, and Antonella Vallenari, Anthony Brown, David Teyssier, and Johannes Sahlmann for reviewing the manuscript, as well as the technical staff and collaborators at DPAC, CNES (Centre National d'Études Spatiales) and ESA. SJ contributed to the development and implementation of the models, technical and scientific validation, documentation, analysis of the results and wrote the manuscript. CBJ contributed to the project definition, method design, results analysis, and manuscript editing. OC contributed to the overall coordination of the DPAC CU8 Apsis data model, documentation, and interface with the DPCC (CNES Data Processing Centre), as well as to the technical and scientific validation and revision of the manuscript. RC contributed to the revision of the manuscript and the analysis of the results in relation to the presence of different extragalactic populations. This work was funded in part by the DLR (German space agency) via grant 50 QG 2102.

## References

Abdurro'uf, Accetta, K., Aerts, C., et al. 2022, ApJS, 259, 35
Alexander, D. M., Davis, T. M., Chaussidon, E., et al. 2023, AJ, 165, 124
Andrae, R., Fouesneau, M., Creevey, O., et al. 2018, A&A, 616, A8
Assef, R. J., Stern, D., Noirot, G., et al. 2018, ApJS, 234, 23
Bai, Y., Liu, J., Wang, S., & Yang, F. 2018, AJ, 157, 9
Bailer-Jones, C. A. L. 2021, gAIA-C8-TN-MPIA-CBJ-094
Bailer-Jones, C. A. L., Andrae, R., Arcay, B., et al. 2013, A&A, 559, A74
Bailer-Jones, C. A. L., Fouesneau, M., & Andrae, R. 2019, MNRAS, 490, 5615
Baqui, P. O., Marra, V., Casarini, L., et al. 2021, A&A, 645, A87
Bautista, J. E., Busca, N. G., Guy, J., et al. 2017, A&A, 603, A12
Bañados, E., Venemans, B. P., Mazzucchelli, C., et al. 2018, Nature, 553, 473
Bessiere, P. S., Tadhunter, C. N., Ramos Almeida, C., & Villar Martín, M. 2012, MNRAS, 426, 276
Beutler, F., Blake, C., Colless, M., et al. 2011, MNRAS, 416, 3017
Beutler, F., Blake, C., Colless, M., et al. 2012, MNRAS, 423, 3430
Blake, C., & Bridle, S. 2005, MNRAS, 363, 1329
Bovy, J., Myers, A. D., Hennawi, J. F., et al. 2012, ApJ, 749, 41
Burke, C. J., Aleo, P. D., Chen, Y.-C., et al. 2019, MNRAS, 490, 3952
Busca, N. G., Delubac, T., Rich, J., et al. 2013, A&A, 552, A96
Carrasco, D., Barrientos, L. F., Pichara, K., et al. 2015, A&A, 584, A44
Chaini, S., Bagul, A., Deshpande, A., et al. 2023, MNRAS, 518, 3123
Chollet, F. et al. 2015, Keras, https://keras.io
Creevey, O. L., Sordo, R., Pailler, F., et al. 2023, A&A, 674, A26
Croom, S. M., Smith, R. J., Boyle, B. J., et al. 2004, MNRAS, 349, 1397
Cunha, P. a. C., Humphrey, A., Brinchmann, J., et al. 2024, A&A, 687, A269
Delchambre, L., Bailer-Jones, C. a. L., Bellas-Velidis, I., et al. 2023, A&A, 674, A31
DESI Collaboration (Abdul Karim, M., et al.) 2025a, Phys. Rev. D, 112, 083515
DESI Collaboration (Abdul Karim, M., et al.) 2025b, Phys. Rev. D, 112, 083514
DESI Collaboration (Andrade, U., et al.) 2025c, Phys. Rev. D, 112, 083512
Di Matteo, T., Springel, V., & Hernquist, L. 2005, Nature, 433, 604
DiPompeo, M. A., Bovy, J., Myers, A. D., & Lang, D. 2015, MNRAS, 452, 3124
Dolney, D., Jain, B., & Takada, M. 2006, MNRAS, 366, 884
du Mas des Bourboux, H., Rich, J., Font-Ribera, A., et al. 2020, ApJ, 901, 153
Eisenstein, D. J., & Hu, W. 1998, ApJ, 496, 605
Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, ApJ, 633, 560
El-Badry, K., Rix, H.-W., Cendes, Y., et al. 2023, MNRAS, 521, 4323
Euclid Collaboration (Mellier, Y., et al.) 2025, A&A, 697, A1
Fan, X. 1999, AJ, 117, 2528
Fan, X., Strauss, M. A., Becker, R. H., et al. 2006, AJ, 132, 117
Flesch, E. W. 2015, PASA, 32, e010
Flesch, E. W. 2023, Open J. Astrophys., 6, 49
Font-Ribera, A., Kirkby, D., Busca, N., et al. 2014, JCAP, 2014, 027
Fouesneau, M., Frémat, Y., Andrae, R., et al. 2023, A&A, 674, A28
Fu, Y., Wu, X.-B., Li, Y., et al. 2024, ApJS, 271, 54
Fu, Y., Wu, X.-B., Bouwens, R. J., et al. 2025, ApJS, 279, 54
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, 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
Gaia Collaboration (Panuzzo, P., et al.) 2024, A&A, 686, L2
Gao, D., Zhang, Y.-X., & Zhao, Y.-H. 2008, MNRAS, 386, 1417
Geurts, P., Ernst, D., & Wehenkel, L. 2006, Mach. Learn., 63, 3
Golob, A., Sawicki, M., Goulding, A. D., & Coupon, J. 2021, MNRAS, 503, 4136
Guo, X., Liu, C., Qiu, B., et al. 2022, MNRAS, 517, 1837
He, Z., Qiu, B., Luo, A.-L., et al. 2021, MNRAS, 508, 2039
Hopkins, P. F., Hernquist, L., Cox, T. J., et al. 2005, ApJ, 630, 705
Hopkins, P. F., Hernquist, L., Cox, T. J., et al. 2006, ApJS, 163, 1
Hughes, A. C. N., Bailer-Jones, C. A. L., & Jamal, S. 2022, A&A, 668, A99
Hutchings, J. B., & Bianchi, L. 2010, AJ, 140, 1987
Ivezić, v. 2016, Proc. Int. Astron. Union, 12, 330
Ivezić, v., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jamal, S., & Bailer-Jones, C. A. L. 2024, A&A, 687, A258
Jamal, S., & Bailer-Jones, C. A. L. 2026, GAIA-C8-TN-MPIA-SJA-003, Tech. rep.
Jin, X., Zhang, Y., Zhang, J., et al. 2019, MNRAS, 485, 4539
Kim, E. J., & Brunner, R. J. 2017, MNRAS, 464, 4463
Kingma, D. P., & Ba, J. 2017, Adam: A Method for Stochastic Optimization
Kirkpatrick, J. A., Schlegel, D. J., Ross, N. P., et al. 2011, ApJ, 743, 125
Lacy, M., Storrie-Lombardi, L. J., Sajina, A., et al. 2004, ApJS, 154, 166
Lacy, M., Sajina, A., Petric, A. O., et al. 2007, ApJ, 669, L61
Li, C., Zhang, Y., Cui, C., et al. 2021, MNRAS, 506, 1651
Linder, E. V. 2003, Phys. Rev. D, 68, 083504
Luo, A. L., Zhao, Y. H., Zhao, G., & et al. 2022, VizieR Online Data Catalog, V/156
Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, ApJS, 250, 8
Marocco, F., Eisenhardt, P. R. M., Fowler, J. W., et al. 2021, ApJS, 253, 8
Martínez-Solaeche, G., Queiroz, C., Delgado, R. M. G., et al. 2023, A&A, 673, A103
Małek, K., Solarz, A., Pollo, A., et al. 2013, A&A, 557, A16
Merz, G., Liu, Y., Burke, C. J., et al. 2023, MNRAS, 526, 1122
Mortlock, D. J., Warren, S. J., Venemans, B. P., et al. 2011, Nature, 474, 616
Mortlock, D. J., Patel, M., Warren, S. J., et al. 2012, MNRAS, 419, 390
Nakazono, L., Mendes de Oliveira, C., Hirata, N. S. T., et al. 2021, MNRAS, 507, 5847
Newberg, H. J., & Yanny, B. 1997, ApJS, 113, 89
Onken, C. A., Wolf, C., Hon, W. J., et al. 2023, PASA, 40, e010
Peng, N., Zhang, Y., Zhao, Y., & Wu, X.-b. 2012, MNRAS, 425, 2599
Peters, C. M., Richards, G. T., Myers, A. D., et al. 2015, ApJ, 811, 95
Pierce, J. C. S., Tadhunter, C., Ramos Almeida, C., et al. 2023, MNRAS, 522, 1736
Reynolds, D. 2015, in Encyclopedia of Biometrics, eds. S. Z. Li, & A. K. Jain (Boston, MA: Springer US), 827
Richards, G. T., Fan, X., Schneider, D. P., et al. 2001, AJ, 121, 2308
Richards, G. T., Fan, X., Newberg, H. J., et al. 2002, AJ, 123, 2945
Richards, G. T., Nichol, R. C., Gray, A. G., et al. 2004, ApJS, 155, 257
Richards, G. T., Myers, A. D., Peters, C. M., et al. 2015, ApJS, 219, 39
Rimoldini, L., Holl, B., Gavras, P., et al. 2023, A&A, 674, A14
Rodrigues, N. V. N., Raul Abramo, L., Queiroz, C., et al. 2023, MNRAS, 520, 3494
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, Nature, 323, 533
Schindler, J.-T., Fan, X., McGreer, I. D., et al. 2017, ApJ, 851, 13
Seo, H.-J., & Eisenstein, D. J. 2003, ApJ, 598, 720
Slosar, A., Iršič, V., Kirkby, D., et al. 2013, J. Cosmol. Astropart. Phys., 2013, 026
Souchay, J., Secrest, N., Sexton, R., & Barache, C. 2024, A&A, 683, A112
Stern, D., Assef, R. J., Benford, D. J., et al. 2012, AJ, 753, 30
Stoppa, F., Bhattacharyya, S., Austri, R. R. d., et al. 2023, A&A, 680, A109
Trammell, G. B., Vanden Berk, D. E., Schneider, D. P., et al. 2007, AJ, 133, 1780
Tuccillo, D., González-Serrano, J. I., & Benn, C. R. 2015, MNRAS, 449, 2818
Ulla, A., Creevey, O. L., Álvarez, M. A., et al. 2022, Gaia DR3 documentation, Chapter 11: Astrophysical parameters, Tech. rep.
Villar-Martín, M., Arribas, S., Emonts, B., et al. 2016, MNRAS, 460, 130
Volonteri, M., & Rees, M. J. 2006, ApJ, 650, 669
Wang, F., Wu, X.-B., Fan, X., et al. 2016, ApJ, 819, 24
Weir, N., Fayyad, U. M., & Djorgovski, S. 1995, AJ, 109, 2401
Wu, X.-B., & Jia, Z. 2010, MNRAS, 406, 1583
York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579
Yèche, C., Petitjean, P., Rich, J., et al. 2010, A&A, 523, A14
Zakamska, N. L., Gómez, L., Strauss, M. A., & Krolik, J. H. 2008, AJ, 136, 1607
Zarrouk, P., Burtin, E., Gil-Marín, H., et al. 2018, MNRAS, 477, 1639
Zhang, Y., Zhao, Y., & Wu, X.-B. 2021, MNRAS, 503, 5263

## Appendix A: List of features from the *Gaia* astrometry and photometry used to train DSC

This section provides a summary of the features used to train `Allosmod` and `Specmod-Supp`. We provide the field names as they will appear in the GDR4. Infrared photometry features are only used in *Gaia*-IR mode in `Allosmod(*)` and `Specmod-Supp(*)`.

Astrometric features
- Sine of the Galactic latitude, $\sin b$
- parallax, `parallax`, and its error, `parallax_error`
- total proper motion, `pm`
- proper motion in the right ascension direction, `pmra`, and its error, `pmra_error`
- proper motion in the declination direction, `pmdec`, and its error, `pmdec_error`
- renormalised unit weight error, `ruwe`
- unrescaled excess noise of the source, `agis_astrometric_excess_noise`, and its significance measuring the fit to a normal distribution $\mathcal{N}(0, 1)$.

Image Parameter Determination (IPD) parameters
- fraction of successful IPD windows with one peak and high GoF (Goodness of Fit), `ipd_frac_high_gof`
- fraction of successful IPD windows with multiple peaks, `ipd_frac_multi_peak`
- fraction of transits with truncated windows or multiple gates, `ipd_gof_frac_odd_win`
- amplitude of the IPD Gof versus the position angle of scan, `ipd_gof_harmonic_amplitude`
- phase of the IPD Gof versus the position angle of scan, `ipd_gof_harmonic_phase`

Photometric features
- $G$ band magnitude, `phot_g_mean_mag`
- colour $G_{BP} - G$, `bp_g`
- colour $G - G_{RP}$, `g_rp`
- colour $G_{BP} - G_{RP}$, `bp_rp`
- relative variability in the G-band (relvarg), $\sqrt{\texttt{phot_g_n_obs}/\texttt{phot_g_mean_flux_over_error}}$
- error on the G-band mean flux, `phot_g_mean_flux_error`
- error on the BP-band mean flux, `phot_bp_mean_flux_error`
- error on the RP-band mean flux, `phot_rp_mean_flux_error`
- BP/RP flux excess factor with the colour dependency removed. The colour excess is computed from a ratio of the mean fluxes in the BP-, RP- and G-bands, (`phot_bp_mean_flux` + `phot_rp_mean_flux`)/ `phot_g_mean_flux`

Features combining mid-infrared photometry from CatWISE2020 and *Gaia* data
- colour $G - W1$
- colour $W1 - W2$

## Appendix B: Summary of the MLP architectures in the DSC

This section summarises the hyperparameters used for the MLP models in the DSC. Several models are trained with various hyperparameter configurations (number of layers, activation function, learning rate, and batch size), including different preprocessing and normalisation strategies. When normalised, XP fluxes are divided by the total sum of the fluxes in BP and RP, similar to GDR3. When preprocessed, features from astrometry and photometry are standardised. The best-performing models achieve the highest completeness and purity of the extragalactic classes.

**Table B.1.** MLP configuration of hyperparameters.

| Classifier | MLP architecture | $n$ | Selected hyperparameters | Common hyperparameters |
|---|---|---|---|---|
| `Specmod` | 1 *Gaia* input layer (size: 220)<br>3 dense layers (units: 128, 64, 32)<br>1 softmax layer (units: 3) | 38 723 | Learning rate $10^{-4}$<br>Batch size: 32<br>Activations: `SELU` | Optimizer: `Adam` (Kingma & Ba 2017)<br>Loss: categorical crossentropy<br>Maximum training epochs: 200 |
| `Allosmod` | 1 *Gaia* input layer (size: 23)<br>2 dense layers (units: 128, 64)<br>1 softmax layer (units: 3) | 11 523 | Learning rate $10^{-4}$<br>Batch size: 32<br>Activations: `RELU` | Validation data fraction : 20%<br>Early stopping on the validation loss<br>Random seed : 42 |
| `Specmod-Supp` | 1 *Gaia* input layer (size: 220 + 23)<br>3 dense layers (units: 128, 64, 32)<br>1 softmax layer (units: 3) | 41 667 | Learning rate $10^{-4}$<br>Batch size: 32<br>Activations: `SELU` | |
| `Allosmod(*)` | 1 *Gaia* input layer (size: 23)<br>2 dense layers (units: 128, 64)<br>1 IR input layer (size: 2)<br>1 concatenate layer,<br>1 softmax layer (units: 3) | 11 529 | Learning rate $10^{-4}$<br>Batch size: 512<br>Activations: `SELU` | |
| `Specmod-Supp(*)` | 1 *Gaia* input layer (size: 220 + 8)<br>3 dense layers (units: 128, 64, 32)<br>1 IR layer (size: 2)<br>1 concatenate layer<br>1 softmax layer (units: 3) | 39 753 | Learning rate $10^{-3}$<br>Batch size: 256<br>Activations: `SELU` | |

**Notes.** $n$ refers to the total number of hyperparameters in the model. The symbol *(*)* indicates the models trained with *Gaia* data and infrared photometry from CatWISE. `Specmod-Supp` is trained using the XP flux spectra (220 pixels) and the list of features (either 23 features or 8 features). Open-source libraries include the `python` `keras` library (Chollet et al. 2015) and the `java` `Eclipse` `deeplearning4j` library (https://deeplearning4j.konduit.ai/). The best-performing models identified in the *Gaia*-IR mode use preprocessed inputs, whereas those in the *Gaia* only mode use the raw inputs.

## Appendix C: Features importance analysis of the DSC `Allosmod` models

This section shows how the input features contribute to the classification performance of the trained DSC `Allosmod` model. We evaluate the importance of each input feature to each trained model by randomly permuting each input in turn. The change in performance on the test set provides a measure of the importance of the perturbed input without changing the model architecture or introducing unusual values of the input. By resampling each input across all classes, we eliminate the information that a feature may carry about a specific class.

For every model, we run ten random permutations of each input in the test set and compute the completeness and purity of the extragalactic classes in the modified test set. The results, presented in Figures C.1a-C.1b for `Allosmod` and `Allosmod(*)`, show the change in performance, defined as the difference between the modified results (averaged across all permutations) and our initial results (refer to Table 1). A negative change indicates a loss in performance, and the larger the change, the larger the amount of information that the feature appears to carry for that class. We see that the most informative features are proper motions, astrometric excess noise, the colour `bp_g`, relative variability in the $G$ band, and the colour excess factor. In the *Gaia*-IR mode, the infrared colour $W1$-$W2$ is also a key feature in the completeness of quasars. The analysis for each BP/RP pixel input to Specmod (figure not shown) indicates that feature importance is distributed across a range of wavelengths, suggesting that it uses spectral lines at specific redshifts.

We explored various MLP configurations and selected the best-performing models with the highest completeness and purity for the extragalactic classes. The performance of models varies with the neural network architecture and its training. Therefore, the results discussed here are specific to the models under presentation.

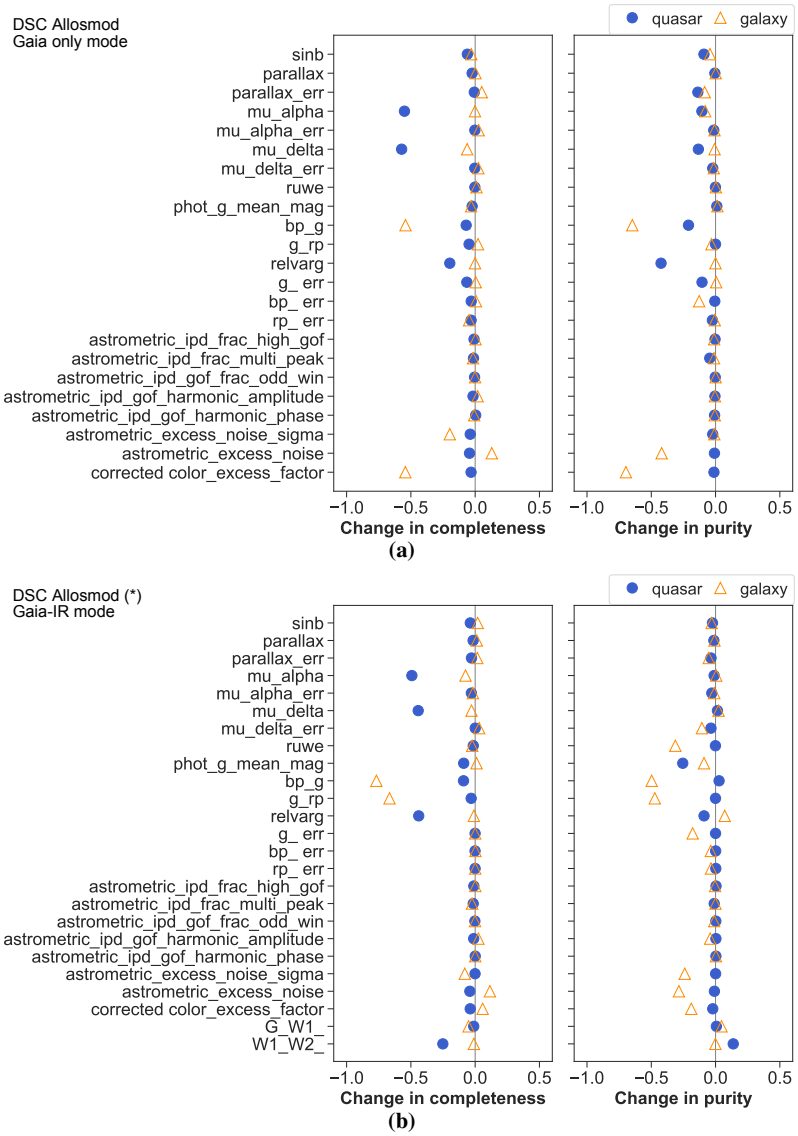



**Fig. C.1.** Feature importance for the DSC models trained using *Gaia* astrometry and photometry. The $x$-axis shows the variation of the completeness (left panel) and purity (right panel) of the extragalactic classes. A negative value indicates a loss in performance, identifying the feature as informative. The $y$-axis lists the input of the trained models. *(a)* and *(b)* show the change in performance for the `Allosmod` in the *Gaia* only mode and `Allosmod(*)` in the *Gaia*-IR mode, respectively.

## Appendix D: Quasar and galaxy candidates identified by the DSC Combmod-$\alpha$ with probabilities over a fixed threshold using *Gaia* data

This section shows how the quasar and galaxy candidates identified by the Combmod-$\alpha$ classifier are distributed in colour-colour space, as a function of the threshold applied to the maximum posterior probability used for class assignment. Figure D.1 shows the effect of a probability threshold of 0.5, 0.7 and 0.9 on the candidates. Setting a stricter probability threshold improves the purity by rejecting more stellar contaminants, albeit at the cost of a slightly reduced completeness.

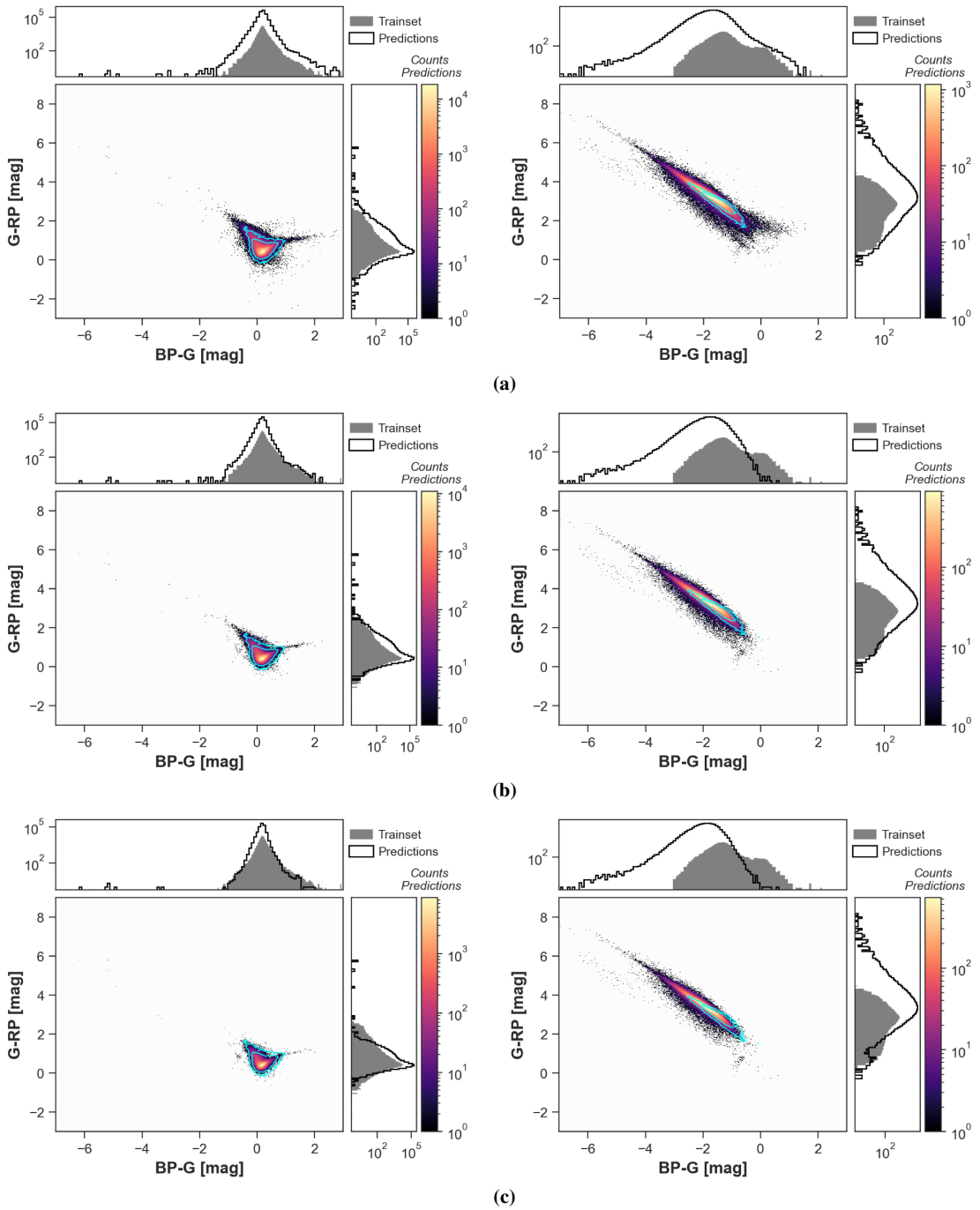
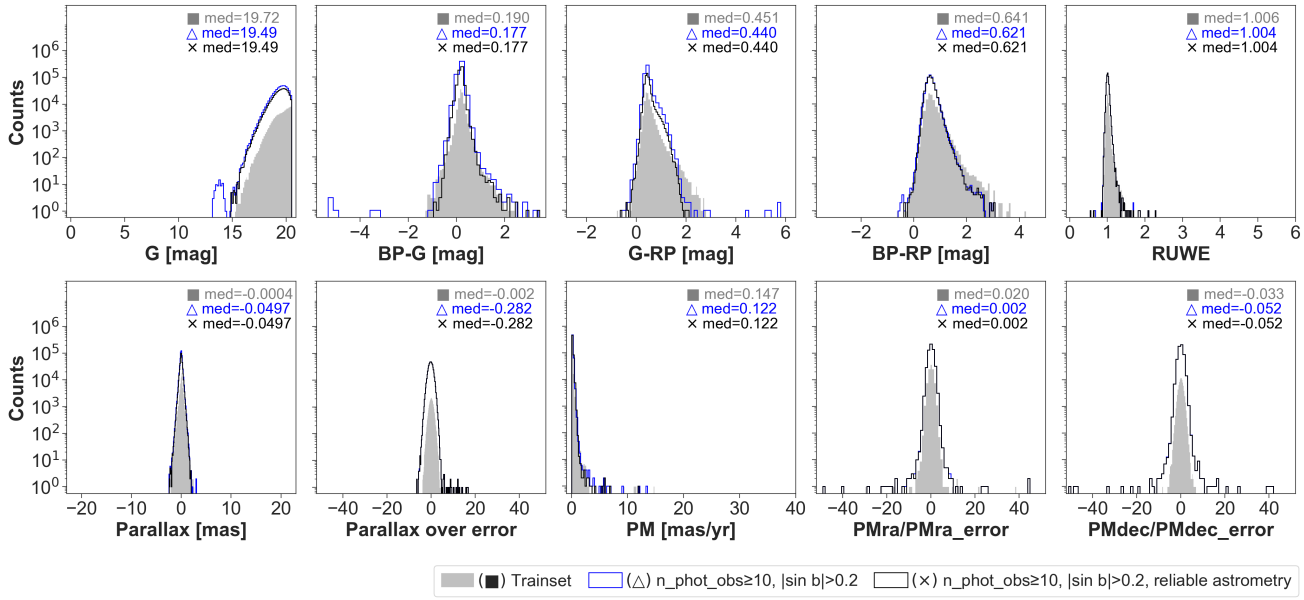
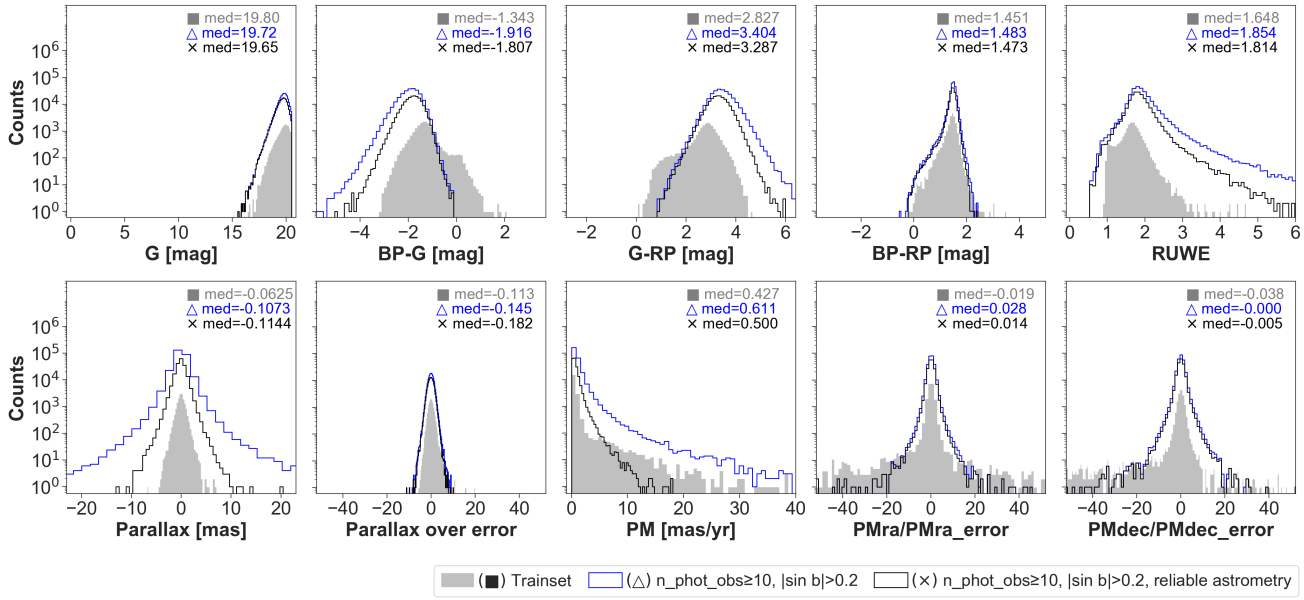



**Fig. D.1.** Same as Figure 8 showing quasar and galaxy candidates at magnitudes $G$<20.5. *Left*: quasars. *Right*: galaxies. Candidates are identified from the class with the highest probability above a fixed threshold. The quality cut was applied for sources at higher latitudes $|\sin b|$>0.20 and phot_[bp,rp]_n_obs≥10. Contours (cyan lines) show the normalised density on a log scale of the highest-density regions for the trainset. In *(a)*, the classifier identifies 1 212 536 quasars and 501 082 galaxies classified with a maximum probability of $P$>0.5. In *(b)*, the classifier identifies 881 217 quasars and 396 346 galaxies classified with a maximum probability of $P$>0.7. In *(c)*, the classifier identifies 557 949 quasars and 268 554 galaxies classified with a maximum probability of $P$>0.9.




**(a)**



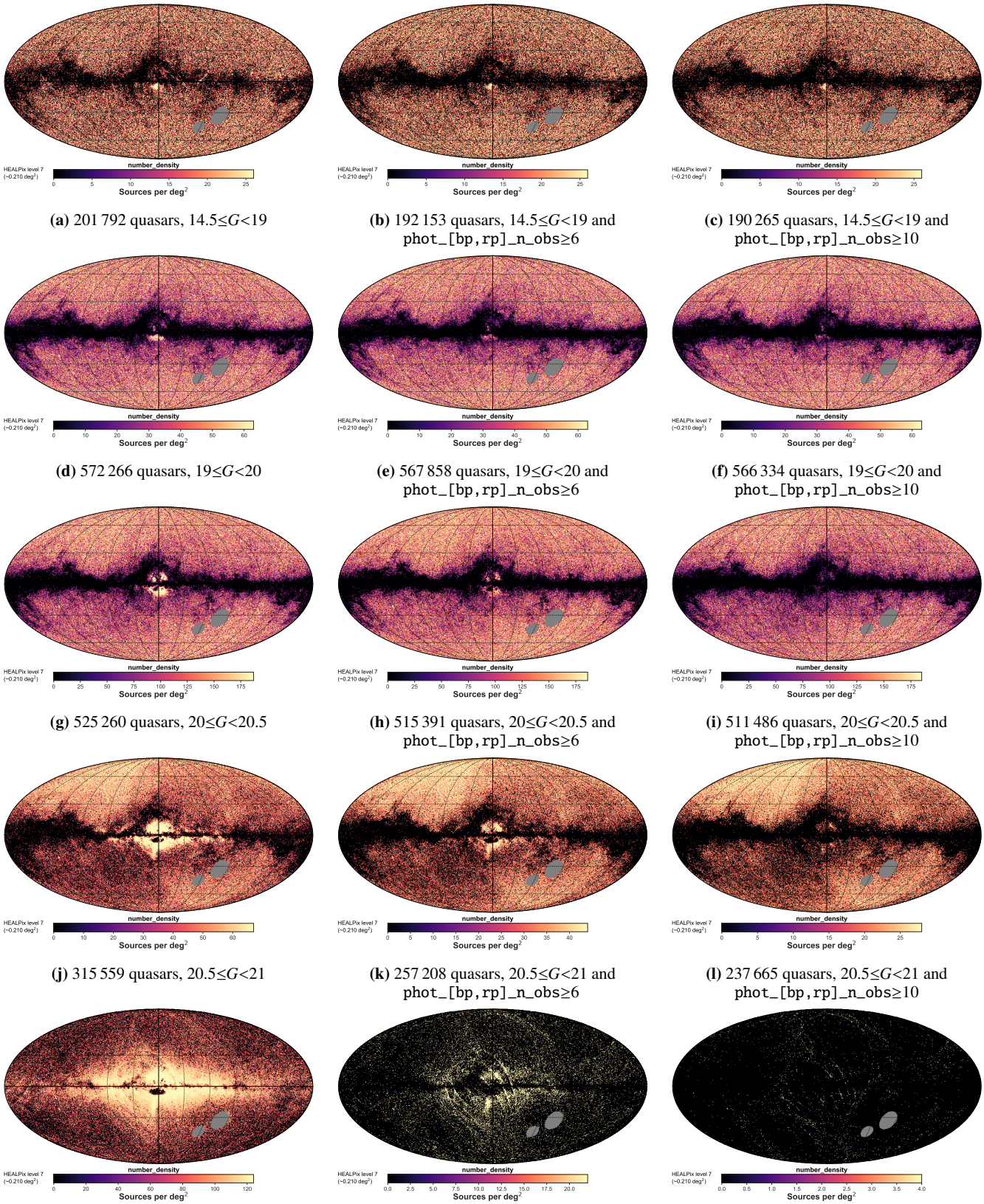

**(b)**

**Fig. D.2.** Same as Fig. 7 but for the extragalactic candidates identified in the *Gaia* only mode by Combmod-$\alpha$ using the global prior, at magnitudes $G$<20.5, with a maximum probability of $P$>0.9. (*a*) Quasar candidates. (*b*) Galaxy candidates. In (*a*), the classifier identifies as quasars, 557 949 sources (△) and 557 163 sources (×). In (*b*), the classifier identifies as galaxies, 268 554 sources (△) and 187 764 sources (×).

## Appendix E: Quasar and galaxy candidates identified by the DSC Combmod-$\alpha$ in the GDR4 per magnitude bins

This section shows how the extragalactic candidates identified by the Combmod-$\alpha$ classifier are distributed across the sky in different magnitude ranges. Quasar and galaxy candidates are shown in Figures E.1 and E.2, respectively.

**Combmod-$\alpha$ classifier predictions of quasar candidates**




**(a)** 201 792 quasars, 14.5≤$G$<19

**(b)** 192 153 quasars, 14.5≤$G$<19 and phot_[bp,rp]_n_obs≥6

**(c)** 190 265 quasars, 14.5≤$G$<19 and phot_[bp,rp]_n_obs≥10

**(d)** 572 266 quasars, 19≤$G$<20

**(e)** 567 858 quasars, 19≤$G$<20 and phot_[bp,rp]_n_obs≥6

**(f)** 566 334 quasars, 19≤$G$<20 and phot_[bp,rp]_n_obs≥10

**(g)** 525 260 quasars, 20≤$G$<20.5

**(h)** 515 391 quasars, 20≤$G$<20.5 and phot_[bp,rp]_n_obs≥6

**(i)** 511 486 quasars, 20≤$G$<20.5 and phot_[bp,rp]_n_obs≥10

**(j)** 315 559 quasars, 20.5≤$G$<21

**(k)** 257 208 quasars, 20.5≤$G$<21 and phot_[bp,rp]_n_obs≥6

**(l)** 237 665 quasars, 20.5≤$G$<21 and phot_[bp,rp]_n_obs≥10

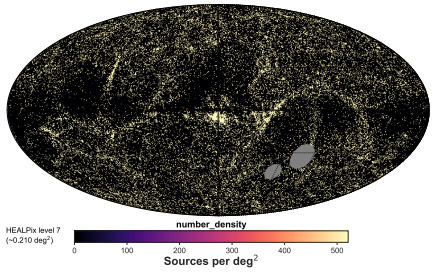
**(m)** 392 863 quasars, 21≤$G$<30

**(n)** 30 105 quasars, 21≤$G$<30 and phot_[bp,rp]_n_obs≥6

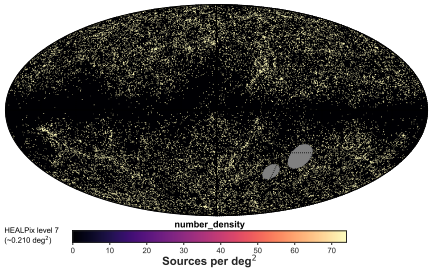
**(o)** 3379 quasars, 21≤$G$<30 and phot_[bp,rp]_n_obs≥10

**Fig. E.1.** Galactic sky distribution of sources classified as quasars from maximum probabilities by Combmod-$\alpha$ using the global prior at HEALpixel level 7 in Mollweide projection. The Magellanic Clouds (LMC and SMC) are masked in grey. Sky distributions show predicted sources at different magnitudes after applying a quality cut to the number of photometric observations in the BP- and RP- bands.

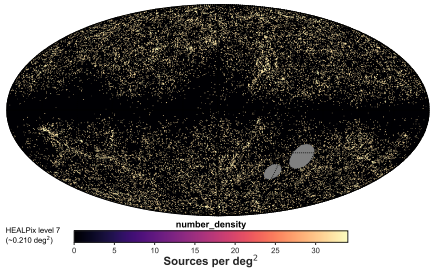
**Combmod-$\alpha$ classifier predictions of galaxy candidates**



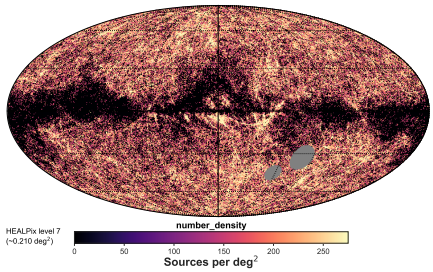

**(a)** 39 642 galaxies, 14.5≤$G$<19



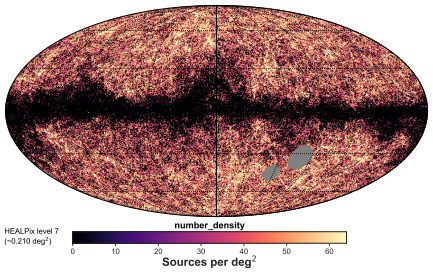

**(b)** 29 603 galaxies, 14.5≤$G$<19 and phot_[bp,rp]_n_obs≥6



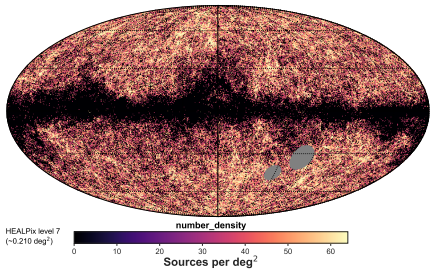

**(c)** 29 023 galaxies, 14.5≤$G$<19 and phot_[bp,rp]_n_obs≥10



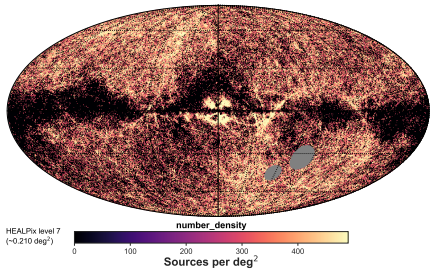

**(d)** 299 315 galaxies, 19≤$G$<20



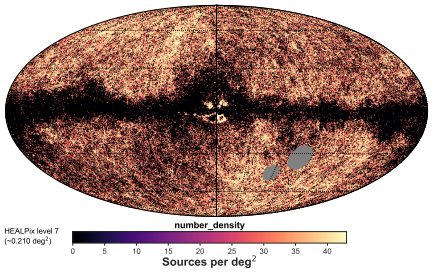

**(e)** 279 898 galaxies, 19≤$G$<20 and phot_[bp,rp]_n_obs≥6



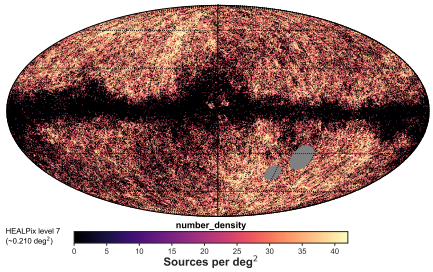

**(f)** 278 549 galaxies, 19≤$G$<20 and phot_[bp,rp]_n_obs≥10



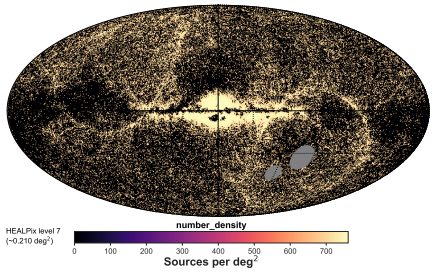

**(g)** 250 590 galaxies, 20≤$G$<20.5



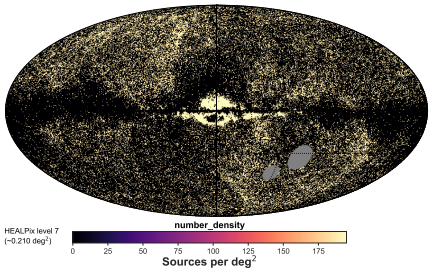

**(h)** 222 869 galaxies, 20≤$G$<20.5 and phot_[bp,rp]_n_obs≥6



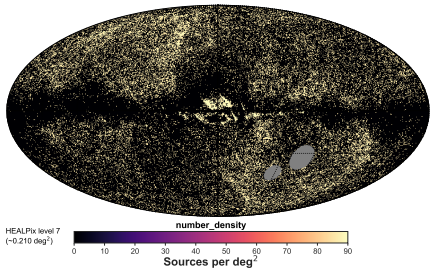

**(i)** 219 017 galaxies, 20≤$G$<20.5 and phot_[bp,rp]_n_obs≥10



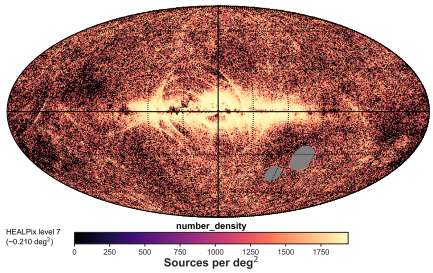

**(j)** 189 843 galaxies, 20.5≤$G$<21



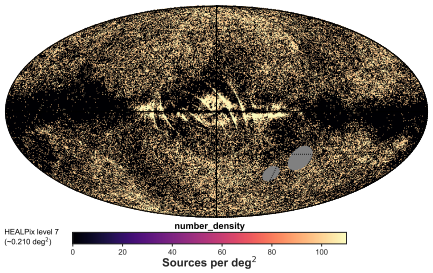

**(k)** 80 159 galaxies, 20.5≤$G$<21 and phot_[bp,rp]_n_obs≥6



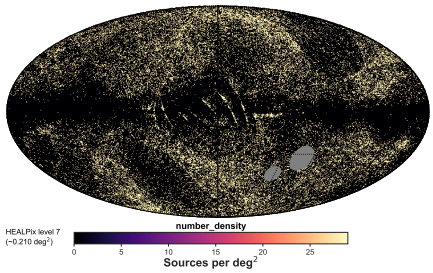

**(l)** 56 856 galaxies, 20.5≤$G$<21 and phot_[bp,rp]_n_obs≥10




**(m)** 527 316 galaxies, 21≤$G$<30




**(n)** 108 431 galaxies, 21≤$G$<30 and phot_[bp,rp]_n_obs≥6




**(o)** 51 586 galaxies, 21≤$G$<30 and phot_[bp,rp]_n_obs≥10

**Fig. E.2.** Same as Fig. E.1, but for sources classified as galaxies from maximum probabilities by Combmod-$\alpha$ using the global prior at HEALpixel level 7 in Mollweide projection.

## Appendix F: Quasar and galaxy candidates identified by the DSC Combmod-$\alpha$ using *Gaia* data and infrared photometry from CatWISE

This section illustrates how the quasar and galaxy candidates, as identified by the Combmod-$\alpha$ classifier in *Gaia*-IR mode, are distributed in the feature space across different magnitude ranges (Figures F.1 and F.2). The results show good completeness but slightly reduced purity, particularly among galaxy candidates that show contamination in the stellar locus. At magnitudes $G \geq 21$, the confusion between quasars and galaxies is more pronounced.

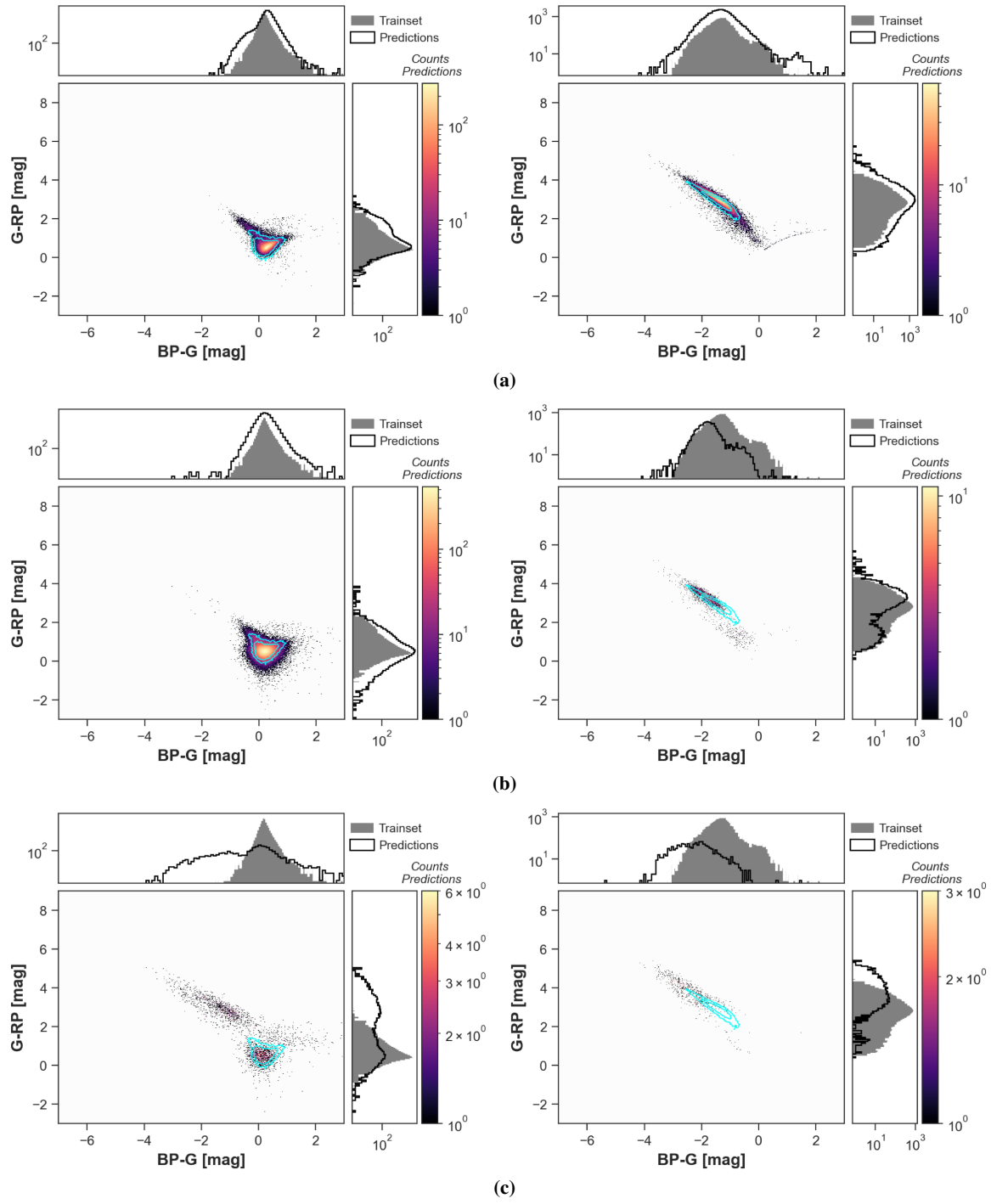
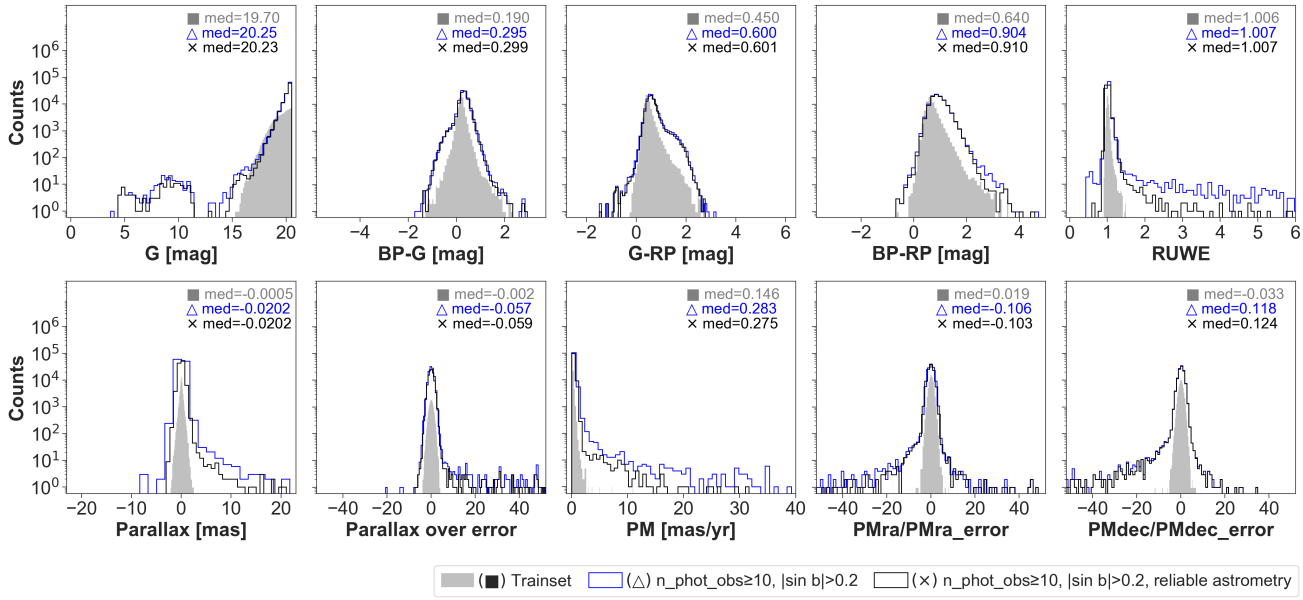
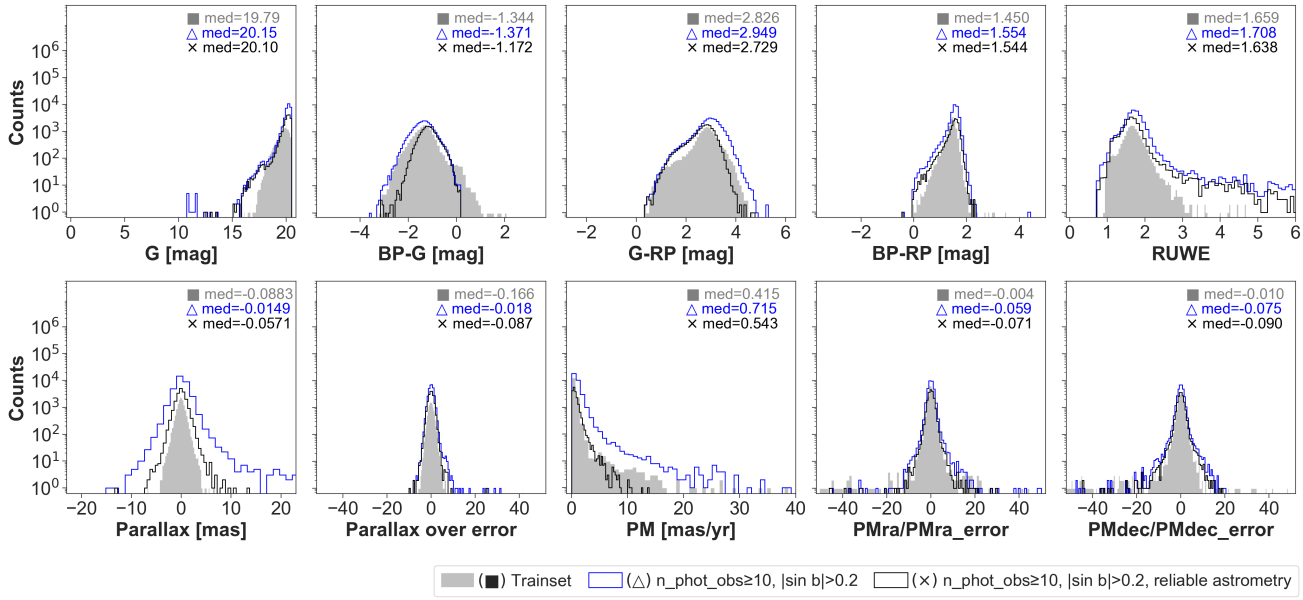



**Fig. F.1.** Colour-colour distributions of sources, exclusive to the *Gaia*-IR mode, identified by Combmod-$\alpha$ and classified from maximum probabilities using the global prior. *Left*: quasars. *Right*: galaxies. Sources located in the LMC and SMC are excluded. The quality cut was applied for sources at higher latitudes $|\sin b| > 0.20$ and phot_[bp,rp]_n_obs$\geq$10. Top-to-bottom panels show the distributions at different magnitude limits. Contours (cyan lines) show the normalised density on a log scale of the highest-density regions for the trainset. In *(a)*, the classifier identifies 113 636 quasars and 34 185 galaxies at magnitudes $G < 20.5$. In *(b)*, the classifier identifies 215 330 quasars and 5290 galaxies at magnitudes $20.5 \leq G < 21$. In *(c)*, the classifier identifies 3198 quasars and 1290 galaxies at magnitudes $21 \leq G < 30$.




**(a)**




**(b)**

**Fig. F.2.** Same as Fig. 7 but for the extragalactic candidates identified exclusively in the *Gaia*-IR mode by `Combmod`-$\alpha$ using the global prior, at magnitudes $G$<20.5. (*a*) Quasar candidates. (*b*) Galaxy candidates. In (*a*), the classifier identifies as quasars, 113 636 sources (△) and 108 497 sources (×). In (*b*), the classifier identifies as galaxies, 34 185 sources (△) and 19 434 sources (×).