Catalogues of active galactic nuclei from Gaia and unWISE data

MNRAS **000**, 1–18 (2019) Preprint 6 September 2019 Compiled using MNRAS LATEX style file v3.0

Catalogues of Active Galactic Nuclei From Gaia and unWISE Data

Yiping Shu,[1⋆†] Sergey E. Koposov,[2,1] N. Wyn Evans,[1] Vasily Belokurov,[1] Richard G. McMahon,[1,3] Matthew W. Auger[1,3] and Cameron A. Lemon[1,3]
[1] *Institute of Astronomy, Madingley Rd, Cambridge, CB3 0HA, UK*
[2] *McWilliams Center for Cosmology, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA*
[3] *Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge, CB3 0HA, UK*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We present two catalogues of active galactic nucleus (AGN) candidates selected from the latest data of two all-sky surveys – Data Release 2 (DR2) of the *Gaia* mission and the unWISE catalogue of the *Wide-field Infrared Survey Explorer* (*WISE*). We train a random forest classifier to predict the probability of each source in the *Gaia*-unWISE joint sample being an AGN, $P_{\rm RF}$, based on *Gaia* astrometric and photometric measurements and unWISE photometry. The two catalogues, which we designate C75 and R85, are constructed by applying different $P_{\rm RF}$ threshold cuts to achieve an overall completeness of 75% (≈90% at *Gaia* $G \leq 20$ mag) and reliability of 85% respectively. The C75 (R85) catalogue contains 2,734,464 (2,182,193) AGN candidates across the effective 36,000 deg^2 sky, of which ≈0.91 (0.52) million are new discoveries. Photometric redshifts of the AGN candidates are derived by a random forest regressor using *Gaia* and *WISE* magnitudes and colours. The estimated overall photometric redshift accuracy is 0.11. Cross-matching the AGN candidates with a sample of known bright cluster galaxies, we identify a high-probability strongly-lensed AGN candidate system, SDSS J1326+4806, with a large image separation of 21.″06. All the AGN candidates in our catalogues will have ∼5-year long light curves from *Gaia* by the end of the mission, and thus will be a great resource for AGN variability studies. Our AGN catalogues will also be helpful in AGN target selections for future spectroscopic surveys, especially ones in the southern hemisphere. The C75 catalogue can be downloaded at https://www.ast.cam.ac.uk/~ypshu/AGN_Catalogues.html

Key words: catalogues – galaxies: active – quasars: general

1 INTRODUCTION

Active galactic nuclei (AGNs) are compact cores in active galaxies that emit strong electromagnetic radiation over a broad wavelength range. They are believed to be powered by the accretion activities of the central supermassive black holes (e.g., Lynden-Bell 1969; Rees 1984; Tanaka et al. 1995). Very luminous AGNs can also be referred to as quasars (also known as QSOs). Large samples of AGNs are of great importance in astrophysics. They can be used to define celestial reference frames (e.g., Ma et al. 1998; Fey et al. 2015; Mignard et al. 2016; Gaia Collaboration et al. 2018b). The variability from AGNs has been used to constrain the properties of supermassive black holes and the fuelling mechanisms (e.g., Blandford & McKee 1982; Vanden Berk et al. 2004; Liu et al. 2008; Li & Cao 2008; MacLeod et al. 2010; Shen et al. 2015; LaMassa et al. 2015; Yang et al. 2018). Among the most luminous sources in the sky, AGNs have been detected back to within the first billion years of the Universe and help to understand the growth of supermassive black holes (e.g., Fan et al. 2006; Wu et al. 2015; Wang et al. 2018b; Pons et al. 2019; Shen et al. 2019). In addition, AGNs have been suggested to play an important role in regulating the formation and evolution of host galaxies (e.g., Silk & Rees 1998; Kang et al. 2006; Fabian 2012; Dubois et al. 2013). Furthermore, spectroscopic observations of AGNs across a wide redshift range can probe the neutral hydrogen fraction in the intergalactic medium and mass distribution in general, which further constrain the history of reionization and cosmological parameters (e.g., Mortlock et al. 2011; Delubac et al. 2015; Bautista et al. 2017; Bañados et al. 2018; Zhao et al. 2019).

⋆ E-mail: yiping.shu@ast.cam.ac.uk,nwe@ast.cam.ac.uk
† Royal Society – K. C. Wong International Fellow

AGNs can be selected based on X-ray observations or by ultraviolet (UV), infrared (IR) or optical photometry and spectroscopy. Each has different biases that affect the resulting samples. Optical identification militates against heavily obscured AGNs, whilst X-ray selected samples are more robust against obscuration. Mid-IR and optical identification can be hampered by the host galaxy's emission, and this is known to bias against AGNs accreting at low fractions of the Eddington limit. Mid-IR and X-ray observations are usually space-based because of the Earth's atmosphere, though the latter require significantly longer exposure time.

The advent of data from the *Wide-field Infrared Survey Explorer* (*WISE*) (Wright et al. 2010) spurred the construction of AGN catalogues based solely on mid-IR data. The *WISE* mission imaged the entire sky in four mid-IR bands, centred at 3.4, 4.6, 12, and 22 μm, referred to as $W1$, $W2$, $W3$, and $W4$, respectively. As noticed in previous work (e.g., Lacy et al. 2004; Stern et al. 2005, 2012; Nikutta et al. 2014), AGNs tend to have redder $W1-W2$ colours relative to stars and inactive, low-redshift galaxies. As a result, a number of work relied on the $W1$-$W2$ colour in selecting AGNs from the AllWISE Data Release (e.g., Assef et al. 2013; Secrest et al. 2015; Assef et al. 2018). Very recently, Schlafly et al. (2019) provided an enhanced unWISE catalogue of roughly 2.03 billion objects that is based on significantly deeper imaging from use of coadds of all publicly available *WISE* data (Lang 2014; Meisner et al. 2017a,b) and has a superior treatment of crowding. This paper provides the first AGN catalogues using the unWISE data.

Nevertheless, the mid-IR-only selection techniques have some limitations. The first is the generally poor imaging resolution of mid-IR data ($\sim 6''$ in *WISE* $W1$ and $W2$ bands). As a result, source blending can become a considerable issue and lead to mis-classifications or render the blended data unusable. Secondly, some non-AGNs have similarly red $W1-W2$ colours as AGNs, which are difficult to distinguish with mid-IR data alone. For example, high-redshift ($z \gtrsim 1.2$) early-type galaxies also have red $W1-W2$ colours because of the rest-frame 1.6 μm stellar bump being shifted beyond the $W1$ band at $z \gtrsim 1.2$ (e.g., Papovich 2008; Papovich et al. 2010; Galametz et al. 2012; Yan et al. 2013). This type of contamination is not significant in the AllWISE data because the characteristic magnitude of high-redshift early-type galaxies in the $W2$ band is about 16.7 mag (e.g., Mancone et al. 2010), at which the AllWISE catalogue is very incomplete. However, it becomes more pronounced in the deeper unWISE catalogue, which reaches ≈50% complete at $W2 = 16.7$ mag (Schlafly et al. 2019). In addition, young stellar objects (YSOs), dusty asymptotic giant branch (AGB) stars, and extended planetary nebulae are also found to have similar $W1-W2$ colours as AGNs (e.g., Rebull et al. 2010; Koenig et al. 2012; Nikutta et al. 2014; Assef et al. 2018).

Optical data have also been used to efficiently select AGNs, through mostly the "UV excess" method or multi-colour cuts (e.g., Sandage & Wyndham 1965; Warren et al. 1987; Hewett et al. 1995; Richards et al. 2002, 2004; Smith et al. 2005; Schneider et al. 2010; Bovy et al. 2011; Myers et al. 2015). Furthermore, the combination of optical and IR data is found to improve the success rate of AGN selections (e.g., Wu & Jia 2010; Maddox et al. 2012; McGreer et al. 2013; Richards et al. 2015; Wang et al. 2016). High-redshift galaxies, YSOs, and AGB stars can also be better identified with the inclusion of optical data. Hitherto, the sky coverage has been limited due to the lack of an all-sky optical survey. However, the European Space Agency's *Gaia* space telescope, launched in 2013, is delivering precise astrometry and optical photometry for more than a billion sources across the entire sky for the first time (Gaia Collaboration et al. 2016). *Gaia* measures three broadband photometry (Evans et al. 2018), i.e. G band (330–1050 nm), the blue prism photometer (BP, 330–680 nm), and the red prism photometer (RP, 630–1050 nm). On the 25$^{\rm th}$ April 2018, *Gaia* delivered its second data release (Gaia DR2, Gaia Collaboration et al. 2018a) containing astrometry and photometry for 1.69 billion sources, based on the first 22 months of operation.

In this paper, we construct new all-sky AGN catalogues based on the combination of these two latest catalogues from *Gaia* and unWISE. This paper is organised as follows. Section 2 describes some properties of the *Gaia*-unWISE sample. Section 3 explains the methods and procedures used to classify AGNs and estimate their photometric redshifts. Section 4 presents our catalogues of AGN candidates. Discussions and conclusion are given in Sections 5 and 6. Throughout the paper, we adopt a cosmological model with $\Omega_m = 0.308$, $\Omega_\Lambda = 0.692$, and $H_0 = 67.8\,{\rm km\,s^{-1}\,Mpc^{-1}}$ (Planck Collaboration et al. 2016). All the magnitudes are given in the Vega system, unless otherwise noted.

2 SAMPLE PROPERTIES

2.1 Data Preparation

To build the *Gaia*-unWISE sample for AGN selection, we perform a nearest neighbour cross-match between the *Gaia* DR2 catalogue (the leading catalogue) and the unWISE catalogue using a matching radius of $2''$. In the cross-match process, we only consider sources with non-zero fluxes in both $W1$ and $W2$ bands. As will be shown later, this requirement reduces the number of AGNs in the sample by $\sim 2.6\%$ relative to requiring non-zero flux in $W1$ alone. We take into account the proper motions of sources (as provided by *Gaia* DR2) in the cross-match process because the source positions in the *Gaia* DR2 catalogue and the unWISE catalogue are given at different reference epochs. The *Gaia*-unWISE sample thus includes 641,266,363 unique *Gaia* sources (corresponding to 564,948,465 unique unWISE sources). One thing to note is that due to the design of the *Gaia* mission, mostly point-like objects can be detected by *Gaia*, so the *Gaia*-unWISE sample consists of stars, AGNs, as well as bright and compact (presumably star-forming) regions in extended galaxies.

2.2 Completeness & Depth of the *Gaia*-unWISE Sample

It is known that the *Gaia* completeness and limiting magnitude exhibit complex spatial variation patterns, primarily related to the *Gaia* scanning law (e.g. Arenou et al. 2018). However, the completeness and limiting magnitude for the *Gaia*-unWISE sample is still unclear. We thus compute the peak and 99*th* percentile in the *Gaia* G-band magnitude distribution in individual spatial bins for all the ≈ 567 million

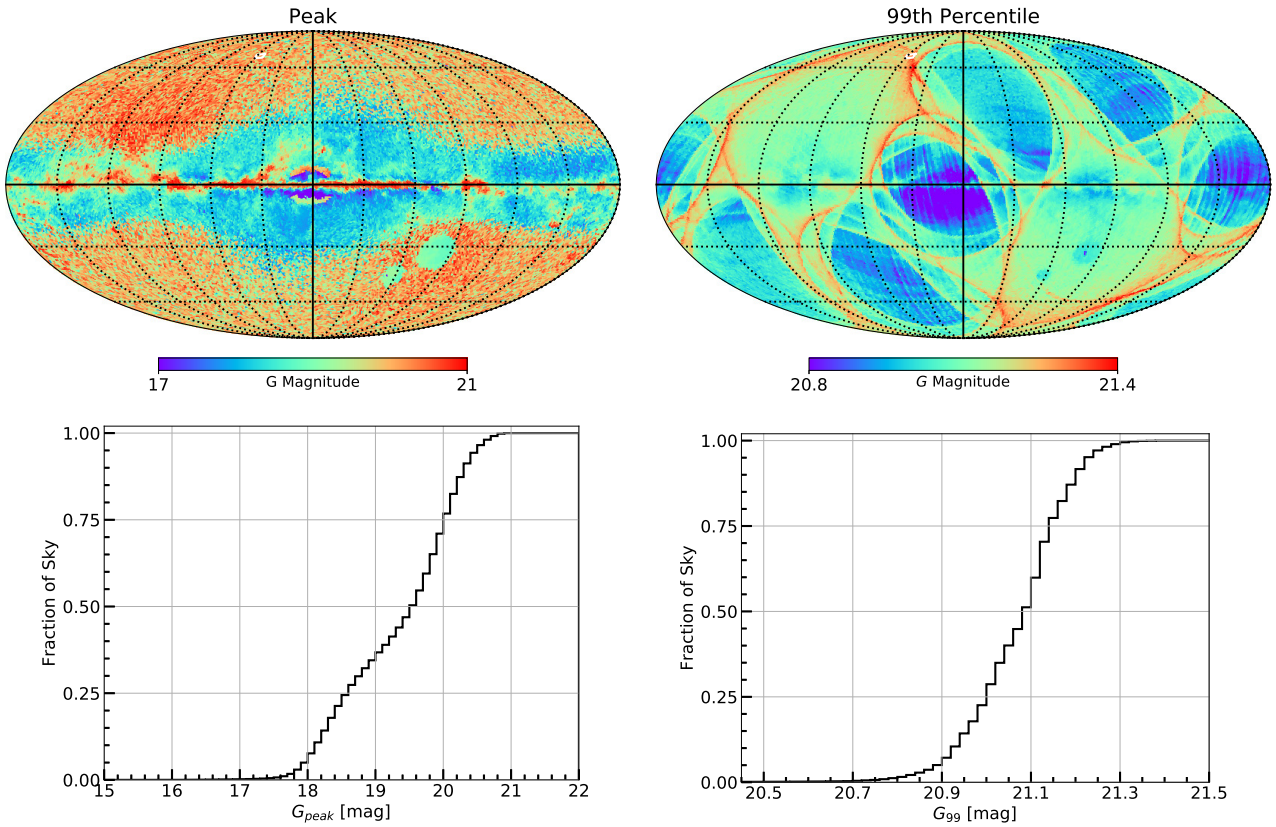

Figure 1. *Top:* Spatial distributions in Mollweide projection (cell size of ≈0.84 deg^2) of $G_{\rm peak}$ (*left*) and G_{99} (*right*) for the *Gaia*-unWISE subsample with $G \geq 16$ mag. The white polygon indicates the location of the Boötes field (at $l \approx 57°$, $b \approx 67°$). *Bottom:* One-dimensional cumulative sky coverage histograms (bin size of 0.1 mag) of $G_{\rm peak}$ (*left*) and G_{99} (*right*) for the same *Gaia*-unWISE subsample.

sources in the *Gaia*-unWISE sample with $G \geq 16$ mag. The peak G magnitude, $G_{\rm peak}$, should be a good indicator of the completeness, and the 99*th* percentile in G, G_{99}, has been used to quantify the limiting magnitude (Arenou et al. 2018). Figure 1 shows the spatial distributions and one-dimensional cumulative sky coverage histograms of $G_{\rm peak}$ and G_{99} for the *Gaia*-unWISE subsample. We point out that although shown in the *Gaia* G-band magnitude, these maps and histograms have also taken into account the incompleteness and limiting magnitudes in $W1$ and $W2$ of the unWISE catalogue. In particular, the brighter $G_{\rm peak}$ structures at low latitudes and towards the bulge and Magellanic Clouds are primarily caused by the brighter incompleteness limits in $W1$ and $W2$ (see Figures A1 and A2). We also find that the *Gaia*-unWISE sample is complete at $G \approx 19.5$ mag in more than 50% of the sky. The G_{99} map clearly shows the *Gaia* scanning law, where the limiting magnitude is deeper in regions that have more repeated observations by *Gaia*. This is primarily because faint sources that have more repeated observations by *Gaia* tend to have more precise astrometric and photometric measurements and are more likely to be included in the *Gaia* DR2 catalogue relative to sources in the less *Gaia*-scanned regions. The faintest limiting magnitude of the *Gaia*-unWISE sample is about $G = 21.4$ mag, and more than 50% of the sky has a limiting magnitude fainter than $G \approx 21.1$ mag. We note that the overall limiting magnitude and completeness for the *Gaia*-unWISE sample will improve in the near future as more repeated *Gaia* observations will be conducted across the whole sky.

The completeness of the expected AGNs in the *Gaia*-unWISE sample needs to be assessed separately. Figure A1 shows that the unWISE catalogue is complete at $W1 \approx 16.5$ mag in more than 90% of the sky. As will be shown later, the expected AGNs in the *Gaia*-unWISE sample generally have $G - W1 > 3$ mag. Considering the $G_{\rm peak}$ distribution in Figure 1, it is suggested that the expected AGNs will be complete at $G \simeq 19.5$ mag in more than 50% of the sky (mostly at high latitudes of $|b| > 20°$).

2.3 AGN Density in the *Gaia*-unWISE Sample

To estimate the expected AGN number density in the *Gaia*-unWISE sample, we use the deep and wide Boötes field of the NOAO Deep Wide-Field Survey (NDWFS, Jannuzi & Dey 1999). The Boötes field is a ~ 9.2 deg^2 region centred at approximately R.A.= 218°, Decl. = 34° (indicated by the small, white polygon in Figure 1) with deep observations in a broad range of (up to 17) filter bands from UV to mid-IR, and therefore has been used for quantifying the performance of AGN selection techniques and AGN studies in general (e.g., Assef et al. 2010, 2013; Chung et al. 2014; Assef et al. 2018; Williams et al. 2018). In particular, we make use of the catalogue from Chung et al. (2014) that contains 431,038 sources extracted from the Boötes field, re-

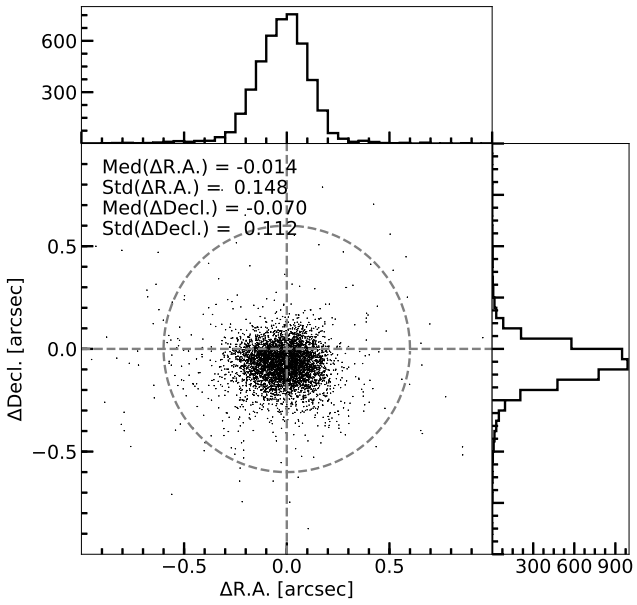

Figure 2. Positional offsets from *Gaia* DR2 to the Boötes catalogue for the 4,564 matched sources using a matching radius of 1″. Matches with separations ≤0.″6 (enclosed by the grey dashed circle) are considered as true matches.

ferred to as the Boötes source catalogue, down to $R \lesssim 23.9$ mag, which should be complete towards the faint end for our purpose as the limiting magnitude of the *Gaia*-unWISE sample is $G \sim 21$ mag. At the bright end, the typical saturation limit of the NDWFS survey is $R \simeq 17$ mag (Chung et al. 2014), which roughly corresponds to $G \approx 17$ mag. For every source in the Boötes source catalogue, Chung et al. (2014) fitted its observed spectral energy distribution (SED) with stellar, galaxy, and galaxy+AGN spectral templates, based on which one can decide whether the source is a star, a galaxy, or an AGN.

We first select a $2^\circ \times 2^\circ$ sub-region centred at R.A.= 218°, Decl. = 34° from the Boötes field, which contains 159,754 sources from the Boötes source catalogue. We perform a nearest neighbour cross-match between these sources and the *Gaia*-unWISE sample with a matching radius of 1″. Considering that the source positions are given at different reference epochs, we apply a correction to the *Gaia* DR2 positions in the cross-match process for sources with well-measured proper motions (i.e. S/N > 5), and obtain 4,564 matched sources. The unmatched ones are mostly either extended or fainter sources that are not catalogued in *Gaia* and/or unWISE. Figure 2 shows offsets from positions in *Gaia* DR2 to positions in the Boötes catalogue for the matched 4,564 sources after proper-motion corrections. We find that the median positional offsets are -0.″014 in the right ascension direction and -0.″07 in the declination direction. More than 99% (4523) of the matched sources have absolute positional offsets less than 0.″6, which are considered as true matches. Further removing sources whose SEDs are better fitted by stellar templates instead of the galaxy+AGN templates as indicated by the reduced χ^2_ν values, i.e. $\chi^2_\nu(\mathrm{star}) \leq \chi^2_\nu(\mathrm{galaxy+AGN})$, we obtain 718 extragalactic sources in this sub-region.

To determine how many of the extragalactic sources are AGNs, we consider two metrics that have been previously used for the Boötes source catalogue. The first is the F ratio derived from the reduced χ^2_ν values and degrees of freedom by Chung et al. (2014). They suggested that a threshold of $F > 10$ should yield a reasonably complete and clean AGN sample. On the other hand, Assef et al. (2018) defined a parameter $\hat{a}$, which is the AGN contribution to the total luminosity based on the SED fitting results, and used $\hat{a} > 0.5$ for selecting AGN candidates. To decide which AGN-selection criterion is appropriate for our purpose, we consider the Sloan Digital Sky Survey (SDSS) DR14 QSO catalogue (Pâris et al. 2018), based on which our AGN classification is calibrated (as will be shown later). In the $2^\circ \times 2^\circ$ sub-region, there are 89 DR14 QSOs that are in the *Gaia*-unWISE sample and have been assigned an F ratio and an $\hat{a}$ value by Chung et al. (2014). We find that requiring $F > 10$ or $\hat{a} > 0.5$ alone only recovers 76 or 80 DR14 QSOs, while requiring ($F > 10$ OR $\hat{a} > 0.5$) can recover 84 DR14 QSOs (i.e. ≈95%). We therefore assume that sources with either $F > 10$ or $\hat{a} > 0.5$ can be considered as AGNs that will be detected in this work.

315 of the 718 extragalactic sources in the Boötes sub-region satisfy the requirement of ($F > 10$ OR $\hat{a} > 0.5$) and are considered as AGNs, which implies that the AGN number density in the *Gaia*-unWISE sample is ∼100 deg^{-2} in the Boötes field. Considering that the Boötes field is among the deepest and most complete regions in the current *Gaia*-unWISE sample with $G_{\mathrm{peak}} \approx 20.1$ mag and $G_{99} \approx 21.2$ mag, the overall AGN number density in the *Gaia*-unWISE sample is expected to be less than ∼100 deg^{-2}. It also suggests that ≈99.5% of the 641 million *Gaia*-unWISE sources will be non-AGNs. An efficient and clean way of selecting AGNs from the *Gaia*-unWISE sample is thus highly necessary.

3 METHODOLOGY

3.1 Random Forest Algorithm

In this work, we use the random forest (RF) algorithm for AGN/non-AGN classification and AGN photometric redshift estimation. The RF is a widely used, supervised machine learning algorithm that has been shown to generate robust models and work efficiently with large data sets.

The RF algorithm relies on an ensemble of decision trees to make predictions for both classification and regression problems (Breiman 2001). The decision trees are built independently based on features (i.e. source properties in our case) of input data sets, which are different bootstrap samples of the original training set. The decision tree is grown in a top-down fashion. At each node of a decision tree, the data set is split into two subsets according to the feature among a randomly-selected subset of all features that gives the highest information gain. The nodes are grown recursively until the stopping criterion is met. In a classification problem, each tree will calculate the probability (1 or 0) of an input object belonging to a particular class, and the mean class probability of all the trees is returned. In a regression problem, each tree will make a prediction on the unknown quantity that we are interested (photometric redshift in our

Table 1. Features considered in the AGN classification.

Feature	Description
PLXSIG	parallax significance defined as $\lvert \frac{\texttt{PARALLAX}}{\texttt{PARALLAX\_ERROR}} \rvert$, set to -999 if null
PMSIG	proper motion significance defined as $\sqrt{(\frac{\texttt{PMRA}}{\texttt{PMRA\_ERROR}})^2 + (\frac{\texttt{PMDEC}}{\texttt{PMDEC\_ERROR}})^2}$, set to -999 if null
G	Extinction-corrected *Gaia* G-band mean magnitude (`PHOT_G_MEAN_MAG`)
G_VAR	Variation in *Gaia* G-band flux defined as $\sqrt{\texttt{PHOT\_G\_N\_OBS}} \times \frac{\texttt{PHOT\_G\_MEAN\_FLUX\_ERROR}}{\texttt{PHOT\_G\_MEAN\_FLUX}}$
BP-G	Extinction-corrected *Gaia* BP-G colour (`BP_G`), set to 999 if null
G-RP	Extinction-corrected *Gaia* G-RP colour (`G_RP`), set to 999 if null
BPRP	Extinction-corrected *Gaia* BP-RP colour (`BP_RP`), set to 999 if null
BPRP_EF	BP/RP excess factor (`PHOT_BP_RP_EXCESS_FACTOR`)
AEN	Excess noise of the source (`ASTROMETRIC_EXCESS_NOISE`)
GOF	Goodness-of-fit statistic of the astrometric solution (`ASTROMETRIC_GOF_AL`)
CNT1	Number of *Gaia* sources within a 1″-radius circular aperture
CNT2	Number of *Gaia* sources within a 2″-radius circular aperture
CNT4	Number of *Gaia* sources within a 4″-radius circular aperture
W1-W2	unWISE $W1$-$W2$ colour
G-W1	Extinction-corrected G-$W1$ colour
GW_SEP	Separation (in arcsec) between a *Gaia* source and its unWISE counterpart

case), and the average value from all the trees is used as the final estimation.

The RF algorithm has been successfully applied to a variety of tasks in astronomy (e.g., Carliles et al. 2010; Dubath et al. 2011; Richards et al. 2012; Carrasco Kind & Brunner 2013; Wyrzykowski et al. 2014; Jayasinghe et al. 2019; Chen et al. 2019), including AGN classification and photometric redshift estimation (e.g., Pichara et al. 2012; Carrasco et al. 2015; Schindler et al. 2017; Nakoneczny et al. 2019; Jin et al. 2019). We note that our work is the first RF-assisted AGN classification across the whole sky.

3.2 AGN Classification

3.2.1 Training and Test Sets

We use `RandomForestClassifier` provided in the `scikit-learn` package (Pedregosa et al. 2011) for AGN classification. We build the AGN data set for the RF classifier from the largest spectroscopically confirmed quasar sample — the SDSS DR14 QSO catalogue (DR14Q, Pâris et al. 2018). We perform a nearest neighbour cross-match between *Gaia* DR2 and DR14Q using a matching radius of 0.″5, and find that 354,586 of the 526,356 quasars in DR14Q are detected and catalogued in *Gaia* DR2. The unmatched DR14Q quasars are mostly fainter than $i \sim 20.2$ mag, beyond which *Gaia* is significantly incomplete. Requiring unWISE counterparts within 2 arcsecs with non-zero fluxes in $W1$ results in 348,252 quasars, of which 339,194 (i.e. 97.4%) further have non-zero fluxes in $W2$. We notice that some of the matched DR14Q quasars appear to have significant *Gaia* parallaxes or proper motions, inconsistent with the fact that they should be stationary. After visual inspections of the images and spectra, we find that the majority of those "moving" quasars have close companions mostly due to projection effects, which affect the estimation of their parallaxes and proper motions. Consequently, parallax, proper motion, and photometry of those objects are no longer reliable, and may confuse the RF classifier. We therefore remove the 220 DR14Q quasars that have parallax or proper motion significance larger than 5σ. The remaining 338,974 quasars comprise the AGN data set and are also referred to as the *Gaia*-unWISE-DR14 QSO sample.

To build the non-AGN data set, we randomly select 10 million objects from the *Gaia*-Pan-STARRS1 crossmatch table and cross-match them with the unWISE catalogue using a matching radius of 2 arcsecs, which results in 2,351,443 objects with unWISE counterparts with non-zero $W1$ and $W2$ fluxes. Obviously, we need to further clean this non-AGN data set by identifying and removing as many AGNs as possible. We therefore put together a known AGN compilation including almost 29 million known AGNs and AGN candidates (duplicates not removed) from the million quasar catalogue, version 5.7 (MILLIQUAS, Flesch 2015), the AllWISE two-colour selected AGN catalogue (Secrest et al. 2015), and the AllWISE R90 and C75 AGN catalogues (Assef et al. 2018). We then remove the 10,902 objects in the non-AGN data set that have counterparts in the known AGN compilation within an aggressive matching radius of 5 arcsecs and are therefore potential AGNs. This number is consistent with the expectation based on the AGN/non-AGN fraction found in Section 2.3, which suggests ∼11500 AGNs in this data set. The cleaned non-AGN data set now has 2,340,541 objects.

The AGN data set and the cleaned non-AGN data set together make up the full data set for the RF classifier, which includes 2,679,515 objects. The full data set is shuffled and randomly split so that 80% is used as the training set and the remaining 20% is used as the test set. The training set contains 271,218 AGNs and 1,872,394 non-AGNs, while the test set contains 67,756 AGNs and 468,147 non-AGNs.

3.2.2 Feature Selection

The RF classifier relies on a set of features (i.e. source properties) to determine whether a source is an AGN or not. In this work, we consider 16 features that we think are relevant in separating AGNs from stars and compact star-forming regions in galaxies. The features are summarised and explained in Table 1. Most of the features are directly available from the *Gaia* DR2 catalogue and the unWISE catalogue, and

more detailed descriptions can be found in Lindegren et al. (2018), the *Gaia* DR2 datamodel[1], and Schlafly et al. (2019). We apply extinction corrections to the *Gaia* G, BP, and RP magnitudes according to the extinction laws in Cardelli et al. (1989) and O'Donnell (1994), with the E($B-V$) value along each sight line extracted from the extinction map in Schlegel et al. (1998). *Gaia* DR2 does not report parallax or proper motion for some sources (see Lindegren et al. (2018) for details), and BP or RP under certain circumstances (see Riello et al. (2018) for details). We flag those null parallaxes and proper motions as −999, and null BP−G, G−RP, or BP−RP colours as 999. In the full data set, 61,332 AGNs and 198,208 non-AGNs do no have parallaxes and proper motions, 21,273 AGNs and 197,300 non-AGNs do not have BP-G colours, 21,252 AGNs and 196,511 non-AGNs do not have G−RP colours, and 21,285 AGNs and 197,637 non-AGNs do not have BP−RP colours.

Following Belokurov et al. (2017), we construct one feature, G_VAR, from direct measurements as

$$\mathrm{G\_VAR} = \sqrt{\texttt{PHOT\_G\_N\_OBS}} \times \frac{\texttt{PHOT\_G\_MEAN\_FLUX\_ERROR}}{\texttt{PHOT\_G\_MEAN\_FLUX}}, \quad (1)$$

in which `PHOT_G_N_OBS` is the number of observations contributing to G photometry, `PHOT_G_MEAN_FLUX` is the G-band mean flux, and `PHOT_G_MEAN_FLUX_ERROR` is the standard deviation of the G-band flux divided by $\sqrt{\texttt{PHOT\_G\_N\_OBS}}$. Clearly, 2.5×G_VAR/ln(10) is equivalent to the variation in the G-band magnitude. It is therefore helpful to include this feature, which should encode a source's variability information during the observing epochs. However, some other technical effects can also lead to a substantial variation in the G-band flux, for instance, a mix of different *Gaia* scanning directions, especially for extended sources with non-circular surface brightness distributions. For each source, we compute the numbers of *Gaia* sources (the target source is included) within circular apertures of 1″, 2″, and 4″ radii and denote them as CNT1, CNT2, and CNT4, respectively. These three features, together with the separation between a *Gaia* source and its unWISE counterpart, GW_SEP, provide a measure of the local crowding effect and the robustness of the *Gaia* astrometric solution and *Gaia* and unWISE photometric measurements, and thus help in better classifying a source.

To select the most important/relevant features for AGN classification, we first train a `RandomForestClassifier` with its default parameter choices with the training set using all the features listed in Table 1, and record its performance on the test set as measured by the `f1_score` metric. The F1 score is defined as

$$\mathrm{F1} = 2 \times \frac{\mathrm{completeness} \times \mathrm{reliability}}{\mathrm{completeness} + \mathrm{reliability}}. \quad (2)$$

For example, suppose a data set contains 100 AGNs and 10000 non-AGNs. For a classifier that mis-classifies 1 AGN and 10 non-AGNs, the F1 score is 0.947. We choose to optimise the classifier for the F1 score because it measures both the completeness and reliability. For this baseline model using 16 features, the F1 score is 0.9875. The relative importance of the 16 features is returned by the `feature_importances_` attribute of the `RandomForestClassifier` method. We remove 4 features (i.e., AEN, GOF, CNT2, and CNT1) that have a cumulative importance less than 0.01, and re-train the model. The F1 score of the new model is 0.9874, i.e. nearly as good as the baseline model. We therefore only use the remaining 12 features for the AGN classification.

3.2.3 *Classifier Tuning & Performance*

RF classifiers require specification of a number of parameters describing what kinds of trees may be built. Fortunately, we find that we can obtain clean samples of AGNs over a wide range of RF parameters. Nevertheless, we select the best possible RF parameters by optimising the RF performance over the four parameters, i.e. *max_features*, *max_depth*, *class_weight*, and *criterion*, that are most relevant to the classifier's performance in our case. We refer interested readers to the `scikit-learn` documentation[2] for a full description of the role of the parameters. We consider *max_features* = [3, 4, 5, 6], *max_depth* = [None, 25, 50], *class_weight* = [None, balanced, {0:1, 1:100}, {0:1, 1:200}, {0:1, 1:500}, {0:1, 1:1000}, {0:1, 1:10000}], and *criterion* = [entropy, gini]. The remaining parameters of `RandomForestClassifier` are set to their default values. We find that the combination of parameters that gives the highest F1 score is *max_features* = 3, *max_depth* = 50, *class_weight* = {0:1, 1:200}, and *criterion* = entropy. We therefore adopt these choices and obtain the best-trained AGN classifier after training on the training set. Nevertheless, we note that changes in the F1 score for the considered various parameter combinations are very tiny, on the level of 0.001.

The relative importance, in descending order, of the 12 features used in the best-trained AGN classifier is shown in the upper corner of Figure 3. The lower corner of Figure 3 shows the two-dimensional distributions and one-dimensional histograms of the 12 features, ordered by the importance, for AGNs and non-AGNs in the training set. The first thing to notice is the clear separation between AGNs and non-AGNs in the $W1$-$W2$ colour, which confirms again the effectiveness of $W1$-$W2$ colour in distinguishing AGNs from stars and galaxies. The PMSIG distribution is also different for AGNs and non-AGNs, with non-AGNs having an extended tail towards large PMSIG due to the presence of moving stars. The $G-W1$ colour of AGNs in the training set peaks around 4, with more than 95% having (extinction-uncorrected) $G-W1 > 2.9$. The $G-W1$ colour of non-AGNs show a bimodal distribution, with the bluer component contributed mostly by stars and the redder component mostly by galaxies. Recent work by Lemon et al. (2019) showed that one can efficiently distinguish QSOs and strongly-lensed QSOs from stars using the combination of $W1-W2$ and $G-W1$ colours. As expected, AGNs generally have larger G_VAR with a peak value of ≈ 0.12, or 0.13 mag, while non-AGNs peaks at G_VAR ≈ 0.01.

Table 2 presents the performance of the best-trained AGN classifier when applied to the test set. The true positive rate (TPR, equivalent to completeness) is the frac-

[1] https://gea.esac.esa.int/archive/documentation/GDR2/Gaia_archive/chap_datamodel/sec_dm_main_tables/ssec_dm_gaia_source.html

[2] https://scikit-learn.org/stable/documentation.html

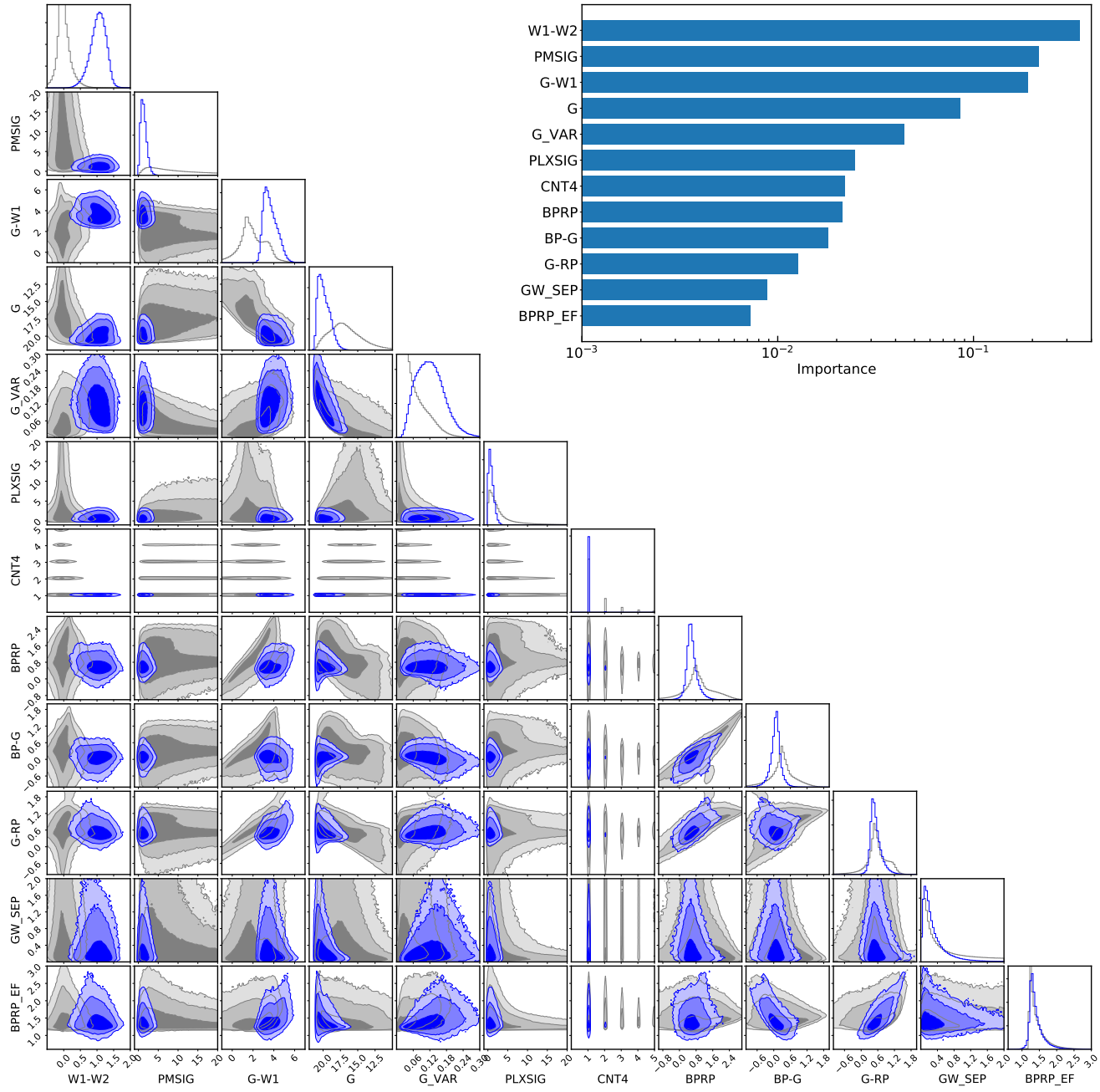

Figure 3. *Upper corner:* Relative importance of the 12 features used by the best-trained AGN classifier. *Lower corner:* Two-dimensional distributions and one-dimensional histograms of AGNs (blue) and non-AGNs (grey) in the training set in various feature spaces. The contours enclose 68%, 95%, and 99% of AGNs and non-AGNs. The features are ordered by the relative importance.

tion of AGNs that are classified as AGNs, while the false positive rate (FPR) is the fraction of non-AGNs that are mis-classified as AGNs. A good classifier should deliver a high TPR and maintain a low FPR at the same time. We show two sets of results that correspond to two different AGN probability thresholds, which, as will be shown later, yield 75% completeness ($P_{\rm RF} \geq 0.69$) and 85% reliability ($P_{\rm RF} \geq 0.94$) respectively. For the test set, the best-trained AGN classifier achieves a TPR of $\gtrsim 93\%$, and the FPR is 0.08%–0.15%.

To illustrate the advantage of combining *Gaia* (optical) and *WISE* (mid-IR) data in identifying AGNs, we apply the *WISE*-only AGN selection criteria used in Stern et al. (2012) and Assef et al. (2018) to the same test set. More specifically, sources are classified as AGNs if they satisfy $W1-W2 \geq 0.8$ (Stern et al. (2012)), or $W1-W2 > 0.71$ (the C75 criterion used by Assef et al. (2018) to achieve 75% completeness), or

$$W1-W2 > \begin{cases} 0.650 \times e^{[0.153\times(W2-13.86)^2]}, & W2 > 13.86 \\ 0.650, & W2 \leq 13.86 \end{cases}$$

(the R90 criterion used by Assef et al. (2018) to achieve 90% reliability). It is clear that using optical and mid-IR data, the TPR becomes significantly higher, i.e. more AGNs can be identified. More importantly, our FPRs are lower by

Table 2. Performance of our AGN classifier based on the test set. The definitions of the true positive rate or TPR and false positive rate or FPR are explained in the text. A good classifier should deliver a high TPR and maintain a low FPR at the same time. For comparison, we also show the results of applying the *WISE*-only AGN selection criteria used in Stern et al. (2012) and Assef et al. (2018) to the same test set. Our AGN classifier delivers significantly better performance.

	TPR	FPR
This work, $P_{\rm RF} \geq 0.69$	98.10%	0.15%
This work, $P_{\rm RF} \geq 0.94$	92.73%	0.08%
Stern et al. (2012)	84.03%	0.34%
Assef et al. (2018), C75	90.63%	0.58%
Assef et al. (2018), R90	60.67%	0.17%

0.25% on average than those of the *WISE*-only criteria (the R90 criterion in Assef et al. (2018) achieves a comparably small FPR, but at the cost of a substantially lower TPR). Although the improvement of ∼0.25% in the FPR seems tiny, it will lead to a huge improvement in the reliability because the number of non-AGNs in the *Gaia*-unWISE sample is almost 640 million. If assuming the non-AGN test set is representative of the non-AGNs in the *Gaia*-unWISE sample, an improvement of 0.25% in the FPR can prevent ≈1.6 million non-AGNs being mis-classified as AGNs.

3.3 Photometric Redshift Estimation

We use `RandomForestRegressor` provided in the `scikit-learn` package for the photometric redshift estimation. 80% of the *Gaia*-unWISE-DR14 QSO sample (271,179 AGNs) is randomly chosen as the training set, and the remaining 20% is used as the test set. The 10 features that are used in the RF regressor are G, W1, BP-G, BP-RP, G-RP, G-W1, RP-W1, W1-W2, G_VAR, and GW_SEP. The RP-W1 feature is derived from G-W1 and G-RP. Again, we find that similar photometric-redshift accuracy can be achieved for a wide range of RF parameters. Nevertheless, we optimise the choices for the two parameters, i.e. *max_features* and *max_depth*, that are usually most relevant to a regressor's performance. In particular, we consider *max_features* = [2, 3, 4, 6, 8, 10] and *max_depth* = [None, 10, 25, 50]. The remaining parameters of `RandomForestRegressor` are set to their default values.

We use the standard R^2 score to evaluate the performance of the RF regressor. Assuming the true, spectroscopic redshifts are denoted as $z^i_{\rm spec}$, the mean of $z^i_{\rm spec}$ is denoted as $\bar{z}$, and the predicted redshifts are denoted as $z^i_{\rm phot}$, the R^2 score (also known as the coefficient of determination) is defined as

$$R^2 \equiv 1 - \frac{\sum_i (z^i_{\rm spec} - z^i_{\rm phot})^2}{\sum_i (z^i_{\rm spec} - \bar{z})^2}. \quad (3)$$

Clearly, the best R^2 score is 1. The combination of parameters that gives the highest R^2 score of 0.752 is *max_features* = 4 and *max_depth* = 25. Nevertheless, changes in the R^2 score for the considered parameter combinations are very tiny. For example, a RF regressor with all its parameters set to default values delivers a R^2 score of 0.749. In the best-trained RF regressor, the most important feature is RP-W1 (relative importance of 0.22), followed by G-W1, W1-W2, W1, BP-G, GW_SEP, BP-RP, G, G_VAR, and G-RP.

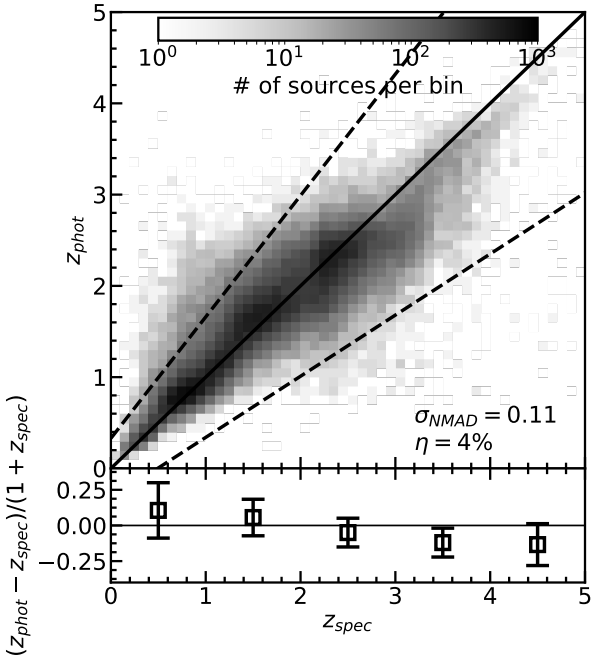

Figure 4. *Top*: Comparison between the estimated photometric redshift and the spectroscopic redshift for the test set. The overall two-dimensional histogram follows the solid one-to-one line, and the photometric redshift accuracy $\sigma_{\rm NMAD}$ is 0.11. 4% of the objects fall outside the region bounded by the two dashed lines, and are referred to as catastrophic outliers. *Bottom*: The mean and 1σ dispersion of the fractional difference $(z_{\rm phot} - z_{\rm spec})/(1+z_{\rm spec})$ in five redshift bins. A mild bias at $\lesssim 1\sigma$ level is seen, suggesting the photometric redshifts tend to be over-estimated for low-redshift AGNs and under-estimated for high-redshift AGNs.

Following the convention in the literature (e.g., Ilbert et al. 2009; Ananna et al. 2017; Fotopoulou & Paltani 2018), we estimate the photometric redshift accuracy using the normalised median absolute deviation defined as

$$\sigma_{\rm NMAD} = 1.48 \times {\rm median}\ (\frac{|z^i_{\rm phot} - z^i_{\rm spec}|}{1 + z^i_{\rm spec}}). \quad (4)$$

The top panel in Figure 4 shows the comparison between $z_{\rm phot}$ from the best-trained RF regressor and $z_{\rm spec}$ of the test set. The overall distribution is centred on the one-to-one relation (solid black line), and $\sigma_{\rm NMAD} = 0.11$. We estimate the rate of catastrophic outliers η as the fraction of sources that have

$$\frac{|z_{\rm phot} - z_{\rm spec}|}{1 + z_{\rm spec}} > 3 \times \sigma_{\rm NMAD} = 0.33. \quad (5)$$

The two dashed lines indicate the boundary where $|z_{\rm phot} - z_{\rm spec}| > 0.33 \times (1 + z_{\rm spec})$. The rate of catastrophic outliers is $\eta = 4\%$. We further divide the test set into five equally-spaced redshift bins from 0 to 5, and find that the mean and standard deviation of $(z_{\rm phot} - z_{\rm spec})/(1 + z_{\rm spec})$ is 0.10 ± 0.19 for $0 < z_{\rm spec} \leq 1$, 0.06 ± 0.13 for $1 < z_{\rm spec} \leq 2$, -0.05 ± 0.10 for $2 < z_{\rm spec} \leq 3$, -0.12 ± 0.10 for $3 < z_{\rm spec} \leq 4$, -0.13 ± 0.15 for $4 < z_{\rm spec} \leq 5$ (bottom panel in Figure 4), which implies a mild bias (at $\lesssim 1\sigma$ level) in the sense that

our best-trained regressor tends to over-estimate the redshifts for AGNs at $z \lesssim 2$ and under-estimate the redshifts for AGNs at $z \gtrsim 3$. We have tried two other commonly used, machine-learning based regression methods, i.e. XGBoost (Chen & Guestrin 2016) and Support Vector Regression. They deliver very similar photometric redshift accuracy as the RF regressor, and the bias persists. It suggests that this bias is due to the intrinsic uncertainties in the AGN photometric redshift estimation rather than the choices of the regression method or the parameter settings, especially when only broadband colours are used.

Nevertheless, the photometric redshift accuracy is comparable to performances of recent work on AGN photometric redshift estimation, most of which use more colours than our photometric redshift estimator (e.g., Maddox et al. 2012; Chung et al. 2014; Schindler et al. 2017; Jin et al. 2019). We thus use the best-trained RF regressor to estimate the photometric redshifts of our AGN candidates.

4 RESULTS

We apply the best-trained AGN classifier to the *Gaia*-unWISE sample of 641,266,363 sources and obtain 3,175,537 sources with AGN probability $P_{\rm RF} \geq 0.5$, which we refer to as AGN candidates. Upon visual inspections, we notice significant over-densities of AGN candidates towards the directions of the Large Magellanic Cloud (LMC) and Small Magellanic Cloud (SMC). Querying against the SIMBAD database finds that the majority of those AGN candidates are actually YSOs and AGB stars in the LMC and SMC that have AGN-like $W1-W2$ colours (e.g., Nikutta et al. 2014). Because of the extremely high source densities in these nearby galaxies, the *Gaia* and *WISE* photometry become less reliable. We therefore remove AGN candidates that are located within twice the radius of LMC, SMC, and M31, which is the nearest big galaxy to the Milky Way. The central positions and radii of LMC, SMC, and M31 are taken from the Catalog and Atlas of the Local Volume Galaxies (Karachentsev et al. 2013). This step removes an area of 541 deg^2. The total number of AGN candidates with $P_{\rm RF} \geq 0.5$ is reduced to 3,104,739, which is referred to as the raw AGN catalogue.

In this work, we will construct two AGN catalogues out of the raw AGN catalogue that are optimised for completeness and reliability respectively. We now explain how this can be achieved by imposing simple cuts on $P_{\rm RF}$.

4.1 C75 And R85 AGN Catalogues

We use the Boötes field, which is among the deepest fields in the *Gaia*-unWISE sample, as a reference to estimate the overall completeness and reliability of the final AGN catalogue at different $P_{\rm RF}$ thresholds. We construct a reference sample including all the 6,703 sources in the *Gaia*-unWISE sample that fall within the previously-defined $2^\circ \times 2^\circ$ subregion in the Boötes field (denoted as the reference field). For every source in the reference sample, we obtain its AGN probability $P_{\rm RF}$ from the best-trained AGN classifier. On the other hand, a nearest neighbour cross-match using a matching radius of $0\farcs6$ finds that 4,523 sources in the reference sample are also in the Boötes source catalogue, for which

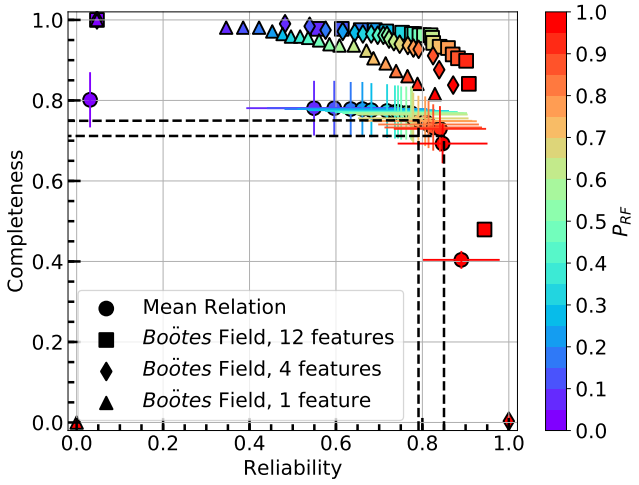

Figure 5. The mean completeness-reliability relation (filled circles) derived from 100 spatially randomly-distributed test fields. The error bars correspond to 1σ variations in completeness and reliability. Squares represent the completeness-reliability relation in the Boötes reference field obtained by the best-trained RF classifier. Diamonds and triangles represent the same relation obtained by two other RF classifiers using fewer features. The symbols are colour-coded according to the $P_{\rm RF}$ threshold. The dashed lines highlight two $P_{\rm RF}$ thresholds at which the mean completeness reaches 75% ($P_{\rm RF} \geq 0.69$) and the mean reliability reaches 85% ($P_{\rm RF} \geq 0.94$).

we can decide whether they are AGNs based on the F ratio and $\hat{a}$ parameter requirement. The unmatched ones are mostly bright objects that were not included in the Boötes source catalogue due to the saturation limit/incompleteness, which we conservatively assume to be non-AGNs. At any given $P_{\rm RF}$ threshold, we can compute the number of sources in the reference sample that have $P_{\rm RF}$ larger than or equal to the threshold (denoted as $N_{\rm candidate}$) and the number of sources among those candidates that satisfy the ($F > 10$ OR $\hat{a} > 0.5$) criterion (denoted as $N_{\rm AGN}$). In addition, we know from Section 2.3 that the total number of AGNs in this reference field is 315. The completeness is therefore given by $N_{\rm AGN}/315$, and the reliability is given by $N_{\rm AGN}/N_{\rm candidate}$. The squares in Figure 5 correspond to the completeness-reliability relation at different $P_{\rm RF}$ thresholds in the Boötes reference field. We note that the actual reliability should be slightly higher than the inferred values because of the adopted conservative treatment of the unmatched objects in the reference sample.

Clearly, the completeness-reliability relation derived from the deep Boötes reference field will be optimistic for the final AGN catalogue. Nevertheless, due to the lack of Boötes-like fields with sufficient and representative sky coverage, we choose to estimate the overall completeness and reliability of the final AGN catalogue through simulations. In particular, we select 100 test fields with the same area as the reference field that are randomly distributed across the high-latitude sky (i.e. $|b| > 20^\circ$). We adopt this requirement because the majority of the raw AGN catalogue are distributed at $|b| > 20^\circ$. For each test field, we first obtain the *Gaia* G-band magnitude distribution, $\mathrm{d}N/\mathrm{d}G$ (test), for all the *Gaia*-unWISE sources therein. A mock sample is gen-

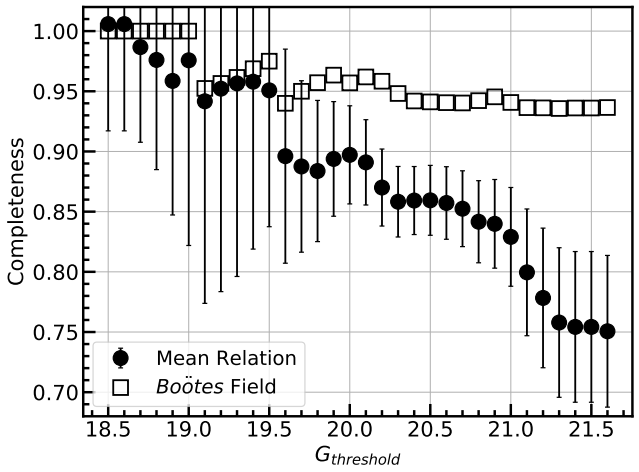

Figure 6. Mean completeness at $P_{\rm RF} \geq 0.69$ of the 100 test fields (filled circles) and the completeness of the Boötes reference field (squares) as a function of the G magnitude threshold.

erated by resampling the reference sample to match dN/dG (test). Because the source density in the test field can be different from that of the reference field, we adjust the relative weight of non-AGNs to AGNs in the reference sample to $w_{\rm non-AGN} \equiv [N_{\rm source}({\rm test}) - 315]/[N_{\rm source}({\rm ref}) - 315]$, in which $N_{\rm source}$(test) and $N_{\rm source}$(ref) are the total number of *Gaia*-unWISE sources in the test field and reference field and 315 is the total number of AGNs in the reference field. As a result, the probability of selecting a non-AGN from the reference sample is a factor of $w_{\rm non-AGN}$ larger than the probability of selecting an AGN in the resampling process. For each test field, 100 independent mock samples are generated. We compute the completeness-reliability relation for each mock sample following the above procedures for the Boötes reference field, and take the mean completeness-reliability relation as the relation for this test field. This process is done for all the 100 test fields.

The circles in Figure 5 show the mean completeness-reliability relation for the 100 test fields, and the error bars represent the 1σ standard deviations in completeness and reliability. The colour of the circles corresponds to the $P_{\rm RF}$ threshold. The completeness and reliability vary significantly across the test fields, on the levels of $\sim$7% and $\sim$13% respectively, due to the spatial variations of source density and *Gaia*-unWISE completeness and limiting magnitude. We find that the mean completeness reaches at least 75% (mean reliability $\sim$ 79%) at $P_{\rm RF} \geq 0.69$, and the mean reliability reaches at least 85% (mean completeness $\sim$ 71%) at $P_{\rm RF} \geq 0.94$. We therefore construct two AGN catalogues, denoted as C75 and R85, by selecting AGN candidates of $P_{\rm RF} \geq 0.69$ and $P_{\rm RF} \geq 0.94$ respectively. The C75 AGN catalogue contains 2,734,464 sources, and the R85 AGN catalogue contains 2,182,193 sources. It is obvious that the R85 catalogue is a subset of the C75 catalogue. The C75 AGN catalogue is publicly available as a FITS file at `https://www.ast.cam.ac.uk/~ypshu/AGN_Catalogues.html`.

The completeness of our AGN catalogues is sensitive to the G magnitude threshold. We can see from Figure 6 that although the estimated overall completeness is 75%, the C75 catalogue is $\approx$ 95% complete for AGN candidates at $G \leq 19.5$ mag and $\approx$ 90% complete for AGN candidates at $G \leq 20$ mag. For the Boötes field that is among the deepest regions in the current *Gaia*-unWISE sample, the completeness at $P_{\rm RF} \geq 0.69$ is about 93–100%, and it varies very little with the G magnitude threshold. We thus expect the overall completeness of AGN catalogues built from later *Gaia* data releases to improve substantially to that of the Boötes field as more repeated *Gaia* observations across the whole sky will be conducted.

To assess by how much the performance of the RF classifier degrades when fewer features are used, we consider two other RF classifiers that are trained on the top four most important features W1−W2, PMSIG, G−W1, and G and on the most important feature W1−W2 alone. The diamond and triangle symbols in Figure 5 show the completeness-reliability relations in the Boötes reference field using $P_{\rm RF}$ values given by these two other RF classifiers respectively. RF classifiers trained on fewer features generally deliver lower completeness and reliability values. At $P_{\rm RF} \geq 0.69$, the RF classifier using 4 features has the same completeness of 93.6% as the best-trained RF classifier using 12 features, while the RF classifier using only 1 feature has a lower completeness of 90.8%. At $P_{\rm RF} \geq 0.94$, the best-trained RF classifier achieves a reliability of 90.7%, while the other two RF classifiers deliver lower reliability of 86.0% and 82.4% respectively.

4.2 Demographics of the AGN Candidates

Figure 7 shows the spatial density distributions of the C75 and R85 AGN catalogues in the Galactic coordinate system. The colour scale is chosen such that white corresponds to an AGN density of 100 deg^{-2} as estimated from the Boötes field, and redder or bluer colour corresponds to higher or lower densities. The first thing to notice is that the AGN density distributions of the C75 and R85 catalogues strongly correlate with the Galactic extinction distribution, and the AGN densities drop quickly to zero towards the Galactic plane and the bulge region, primarily because the high extinction in those regions prevents faint AGNs being detected when optical data are involved. In addition, this could be partially related to a selection bias in our model. The mean E($B-V$) value of the AGN training set is about 0.03 mag and more than 99% AGNs in the training set have E($B-V$) $\leq$ 0.13 mag, while the mean E($B-V$) value in the region within 15° of the Galactic plane is almost 1 mag. As a result, even if there were AGNs behind the high-extinction regions that are bright enough to be detected in *Gaia*, they would tend to have brighter extinction-corrected *Gaia* G magnitudes than AGNs in the training set, and hence smaller $P_{\rm RF}$ values. The effective sky coverage is taken as the total area containing at least one AGN candidate from the R85 catalogue, which is approximately 36,000 deg^2. The average AGN number densities in the C75 and R85 catalogues are 76 deg^{-2} and 61 deg^{-2} respectively. Another clear feature in the spatial distributions of our AGN catalogues is the imprint of the *Gaia* scanning law, i.e. the patchy or filamentary structures in Figure 7. As explained in Section 2.2, the *Gaia* limiting magnitude is deeper in regions that have more repeated *Gaia* observations. As a result, the catalogue completeness and hence AGN density distribution show correla-

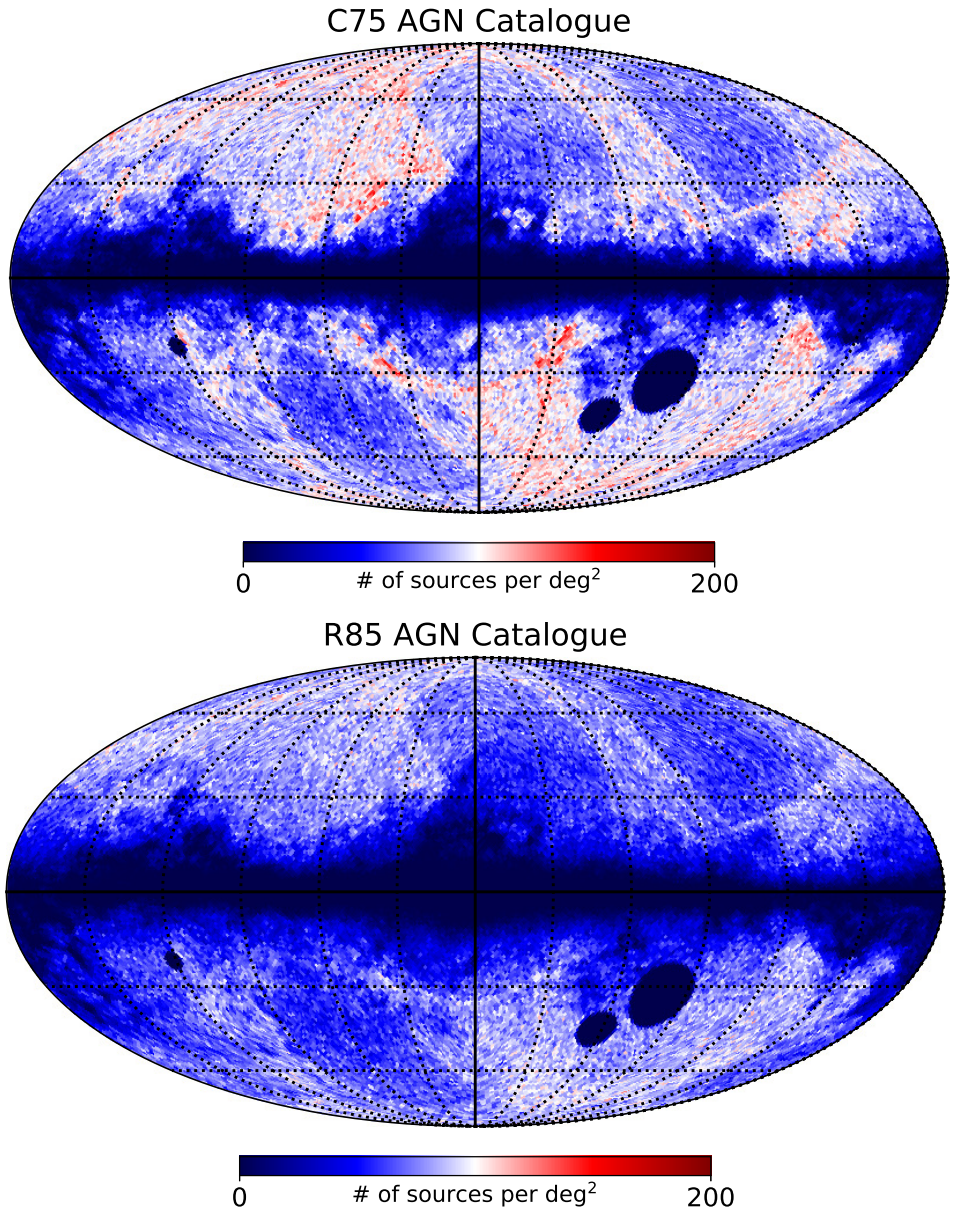

Figure 7. Spatial distributions (in Mollweide projection) of AGN candidates in the C75 (*top*) and R85 (*bottom*) AGN catalogues in the Galactic coordinate system.

tions with the *Gaia* scanning law. We expect this to improve in later *Gaia* data releases.

The top two panels in Figure 8 show the normalised histograms of parallax and overall proper motion for the C75 (solid red) and R85 (dashed blue) AGN catalogues. AGN candidates with null parallaxes or proper motions are not included in the histograms. The two catalogues have very similar parallax distributions. Ignoring AGN candidates with null parallaxes, the mean and median parallax of the C75 (R85) catalogue are −0.019 (−0.026) mas and −0.022 (−0.026) mas. The mean parallax of the more reliable R85 catalogue is consistent with the global parallax zero point of −0.029 mas found for *Gaia* DR2, considering the typical parallax uncertainty of 0.03–0.7 mas (Lindegren et al. 2018).

The bottom left panel in Figure 8 shows the normalised, extinction-corrected *Gaia* G-band magnitude distributions for the C75 (solid red) and R85 (dashed blue) AGN catalogues. At the faint end, the distributions for both samples drop sharply beyond $G \sim 20.6$ mag. We find that the C75 catalogue has a larger fraction of objects in faint magnitude bins compared to the R85 catalogue, implying that the contamination rate in the C75 catalogue becomes higher in fainter magnitude bins.

We apply the best-trained photometric redshift estimator to the C75 catalogue, and the bottom right panel in Figure 8 shows the normalised histograms of the estimated redshifts for AGN candidates in the C75 (solid red) and R85 (dashed blue) catalogues. 76,620 (28,929) AGN candidates in the C75 (R85) catalogue are predicted to be at

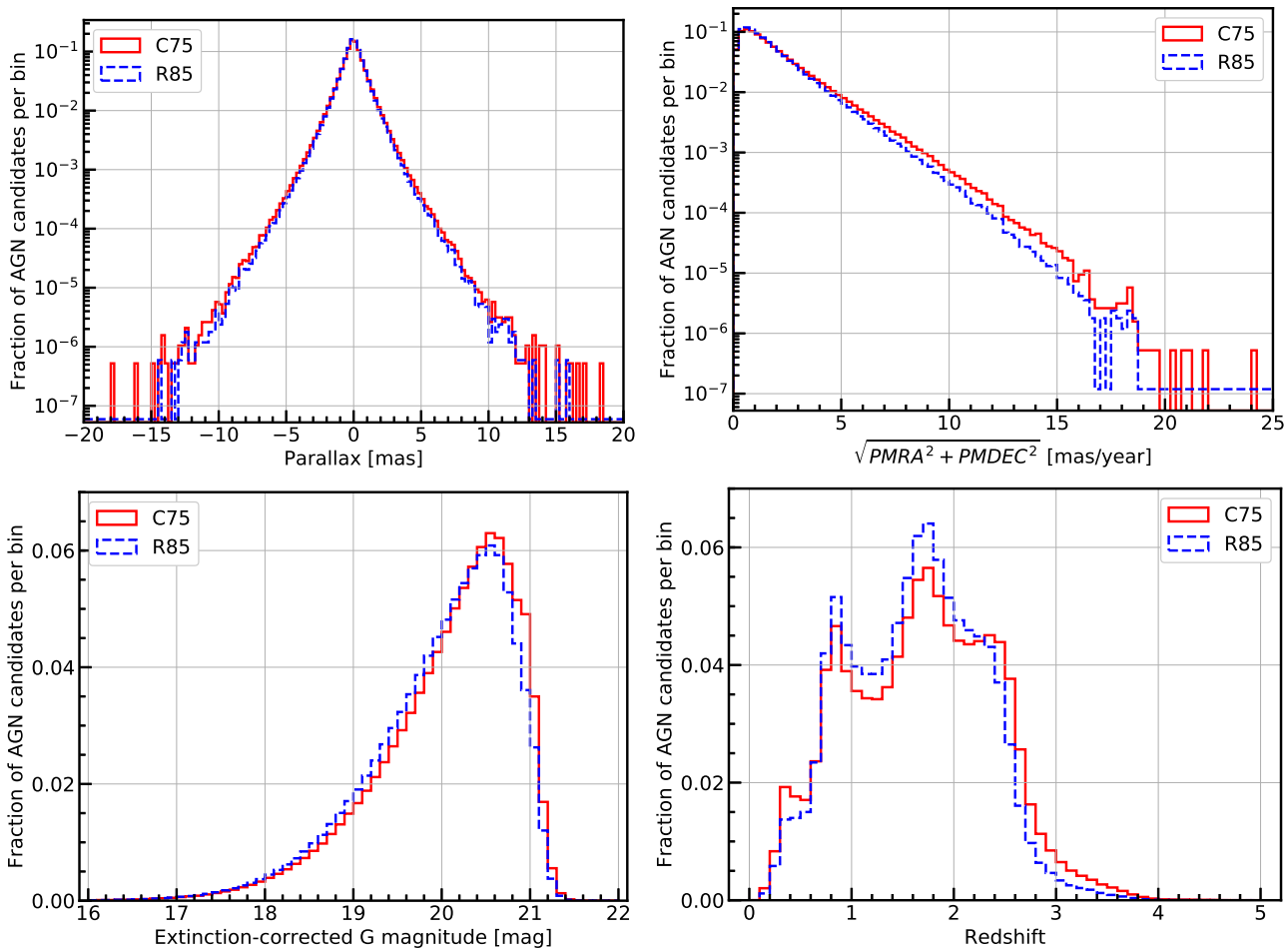

Figure 8. Normalised histograms of parallax (*top left*), proper motion (*top right*), extinction-corrected G-band magnitude (*bottom left*), and photometric redshift (*bottom right*) for the C75 (solid red) and R85 (dashed blue) AGN catalogues.

$z_{\rm phot} \geq 3$, and 1,602 (193) AGN candidates in the C75 (R85) catalogue are predicted to be at $z_{\rm phot} \geq 4$. Considering the photometric-redshift bias found using the test set, we expect the number of high-redshift ($z \gtrsim 3$) AGNs in our catalogues being higher than suggested by the estimated redshifts.

5 DISCUSSIONS

5.1 Comparisons with other AGN catalogues

We build the AGN training set from the DR14Q catalogue because it is the largest spectroscopically confirmed AGN sample to date. To examine whether our RF classifier inherits any selection bias from this choice of training set, we compare our AGN catalogues with some known, large AGN catalogues selected in different wavelength domains and by various techniques in the literature.

The MILLIQUAS catalogue (version 5.7, Flesch 2015) is a compendium of almost 2 million AGNs and high-confidence AGN candidates including the DR14Q sample, the 2-degree Field QSO sample (2QZ, Croom et al. 2004), QSO catalogues from the Large Sky Area Multi-object Fiber Spectroscopic Telescope (LAMOSTQ, Ai et al. 2016; Dong et al. 2018; Yao et al. 2019), the NBCKDE v3 catalogue (NBCKv3, Richards et al. 2015), the SDSS-XDQSO catalogue (XDQSO, Bovy et al. 2011), the All-WISE AGN catalogue (WISEA, Secrest et al. 2015), the Million Optical-Radio/X-ray Associations Catalogue (MORX, Flesch 2016), with the remaining from various other discovery papers[3]. Cross-matching the MILLIQUAS catalogue with the *Gaia*-unWISE sample using a matching radius of $0\farcs5$ results in 1,166,573 matches, which are referred to as the MILLIQUAS-Gaia-unWISE sample. We find that 94.7% and 89.4% of the MILLIQUAS-Gaia-unWISE sample are successfully recovered in our C75 and R85 catalogues. We note that these recovery rates should not be directly compared to the completeness levels of the C75 and R85 catalogues because the MILLIQUAS-Gaia-unWISE sample is not complete in the first place. Instead, the overall, high recovery rates demonstrate the effectiveness of our RF classifier.

Breaking the MILLIQUAS-Gaia-unWISE sample apart, we find that the recovery rates for the DR14Q, 2QZ, and LAMOSTQ samples are higher than the above overall rates, at ≈98% (C75) and ≈95% (R85) respectively. The bulk of

[3] A complete list of the MILLIQUAS input catalogues and references can be found at https://heasarc.gsfc.nasa.gov/W3Browse/all/milliquas.html

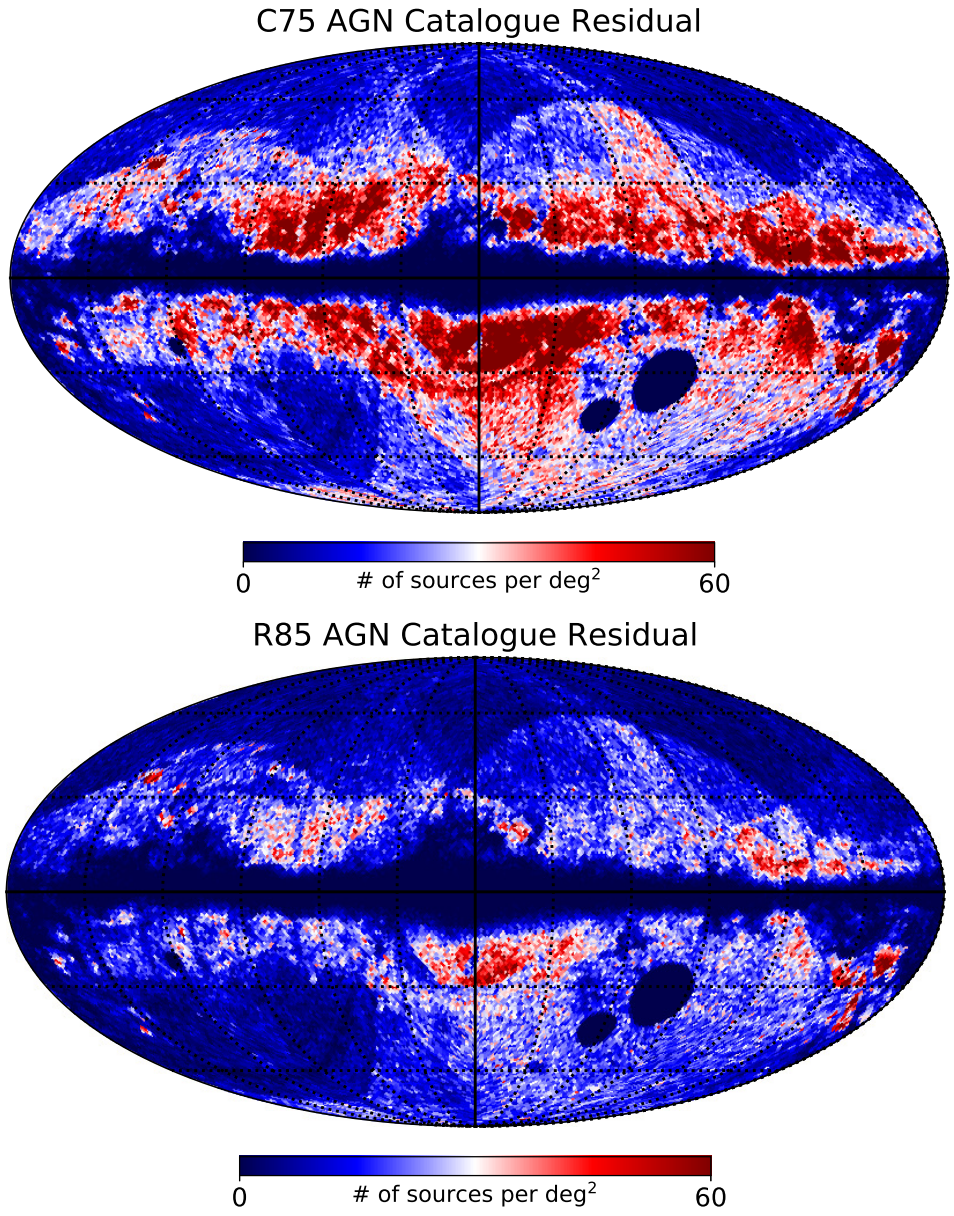

Figure 9. Spatial distribution (in Mollweide projection) of new AGN candidates in the C75 and R85 AGN catalogues in the Galactic coordinate system after removing overlaps with the known AGN compilation.

the DR14Q sample are used in the training process, so its recovery rates are expected to be higher than average. The similarly high recovery rates for the 2QZ and LAMOSTQ samples may be attributed to their target selections being similar to what are used for the DR14Q sample. The 2QZ quasars are selected based on optical ub_Jr colours (Smith et al. 2005), which is similar to how some of the SDSS DR7 quasars (a subset of the DR14Q sample) are selected. The LAMOSTQ sample is primarily selected using optical-infrared colours (Wu & Jia 2010; Wu et al. 2012; Ai et al. 2016) together with the extreme deconvolution (Bovy et al. 2011) and kernel density estimation (Richards et al. 2009) techniques. The CORE sample in the DR14Q is selected based on the extreme deconvolution technique, and a part of the BONUS sample in the SDSS DR12 QSO catalogue (a subset of the DR14Q sample) is selected based on the extreme deconvolution and kernel density estimation techniques. For the MORX sample, the recovery rates are 82% (C75) and 71% (R85), significantly lower than the overall rates. The MORX sample included in the MILLIQUAS catalogue corresponds to AGNs that are discovered in radio/X-ray (Flesch 2016). Considering that radio/X-ray observations are less affected by dust obscuration compared to optical, the lower-than-average recovery rates for the MORX sample may indicate that our RF classifier is less efficient in selecting obscured AGNs. It is also possible that the MORX sample has a higher contribution from host galaxy emission which would result in redder $G - W1$ and bluer $W1 - W2$ colours compared to AGNs in the training set (e.g., Ostrovski et al. 2017; Lemon et al. 2019).

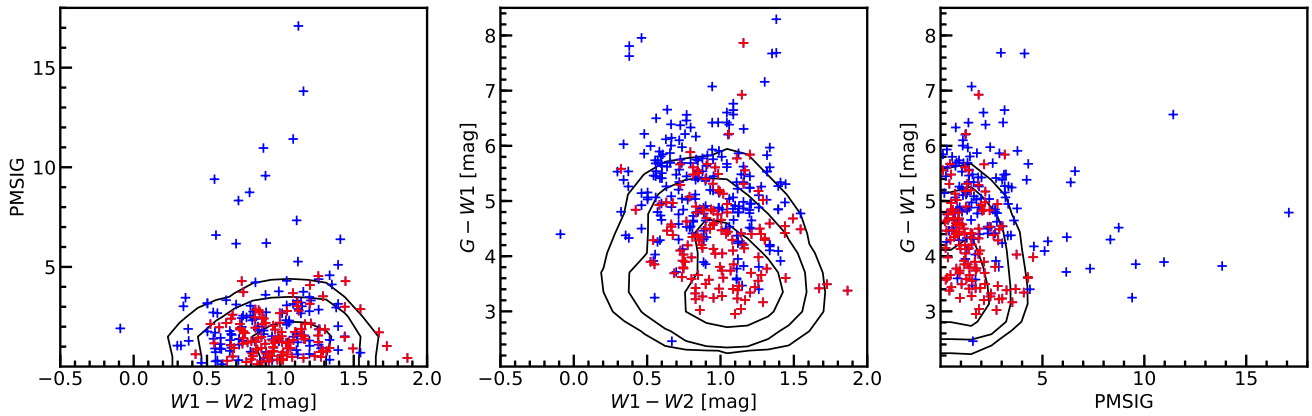

Figure 10. Distributions of the top three most important features, W1−W2, PMSIG, and G−W1, for the 333 known lensed quasar images in the *Gaia*-unWISE sample. The black contours correspond to distributions of AGNs in the training set. Compared to AGNs in the training set or lensed quasar images recovered in the C75 catalogue (red symbols), lensed quasar images that are not in the C75 catalogue (blue symbols) tend to have smaller W1−W2 and larger PMSIG and G−W1.

To determine the number of new AGN candidates in our catalogues, we cross-match the C75 and R85 catalogues with the known AGN compilation using an aggressive matching radius of 5″. We find that at least 911,622 and 515,246 AGN candidates in our C75 and R85 catalogues are previously unknown. Figure 9 shows the spatial distributions of these new AGN candidates, which we refer to as residual maps. Within the extensively observed and studied SDSS footprint, there are few new AGN candidates because our catalogues are limited by the *Gaia* detection limit, which is brighter than those of the known AGN catalogues in this field. Although there have been searches for AGNs outside the SDSS footprint (mostly using the all-sky *WISE* data), our AGN catalogues still find, on average, 30–50 new AGN candidates per deg^2 in those regions, demonstrating the high completeness of our AGN selection technique (e.g. Table 2). Comparing the residual maps of the C75 and R85 catalogues, we find that the number densities of the low-probability AGN candidates close to the Galactic plane and bulge are higher than average, which we think is due to the higher overall source densities therein.

Lastly, we examine how many known strongly-lensed quasars are recovered in our AGN catalogues. To date there are 204 known strongly-lensed quasar systems according to the Gravitationally Lensed Quasar Database[4] (Lemon et al. 2019). In total, 333 lensed quasar images in 168 known systems are in the *Gaia*-unWISE sample, of which 126 lensed quasar images in 104 systems have large enough $P_{\rm RF}$ values to be included in the C75 catalogue. The recovery rate is much lower than found above for AGNs in general. Figure 10 shows the top three most important features, W1−W2, PMSIG, and G−W1, for the 333 known lensed quasar images. We can see that the un-recovered lensed quasar images (blue symbols) generally have smaller W1−W2 and larger PMSIG and G−W1 than the recovered lensed quasar images (red symbols) or AGNs in the training set (black contours). From imaging data, we find that those un-recovered lensed quasar images are usually close to the lensing galaxies or clustered within small separations. They have lower $P_{\rm RF}$ values because 1) their $W1-W2$ and $G-W1$ colours are contaminated by the nearby lensing galaxies (e.g., Lemon et al. 2019); 2) their proper motions and parallaxes are inaccurately inferred, perhaps due to *Gaia* mis-assigning nearby images at each epoch; 3) they generally have CNT4 > 1, which makes them less similar to AGNs in our training set where more than 99% of AGNs have CNT4 = 1. We note that finding highly-clustered AGNs on small scales ($\lesssim$10″) in the presence of nearby, bright galaxies is essentially a different task from building a large and clean sample of AGNs, and a separate classifier/approach might be needed.

[4] https://www.ast.cam.ac.uk/ioa/research/lensedquasars/

5.2 A wide-separation, strongly-lensed AGN candidate

Although our AGN catalogues are not effective in finding small-separation, strongly-lensed AGN systems, they are useful in finding wide-separation strong-lens systems. It has been shown that strongly-lensed AGNs with wide image separations (> 10″) are valuable cosmological probes (e.g., Narayan & White 1988; Turner 1990; Fukugita et al. 1990; Wambsganss et al. 1995; Kochanek 1995, 1996; Lopes & Miller 2004; Oguri et al. 2004; Li et al. 2007; Oguri et al. 2012). However, only four known strongly-lensed AGNs have maximum image separations larger than > 10″ (Inada et al. 2003, 2006; Dahle et al. 2013; Shu et al. 2018). We thus carry out a search for wide-separation, strongly-lensed AGNs by identifying brightest cluster galaxies (BCGs) that have at least two AGN candidates from our C75 catalogue located within a circular aperture of 30″ radius. The BCG sample we use is compiled from Wen & Han (2011, 2015); Wen et al. (2018); Wen & Han (2018), which contains 209,419 BCGs (duplicates not removed) up to redshift of one. 57 unique BCGs with at least two neighbouring AGN candidates are found, and their optical images are visually inspected. We re-discover two previously known wide-separation, strongly-lensed quasar systems SDSS J1004+4112 (Inada et al. 2003) and SDSS J1029+2623 (Inada et al. 2006). The other two known wide-separation, strongly-lensed quasar systems, SDSS J0909+4449 (Shu et al. 2018) and SDSS J2222+2745

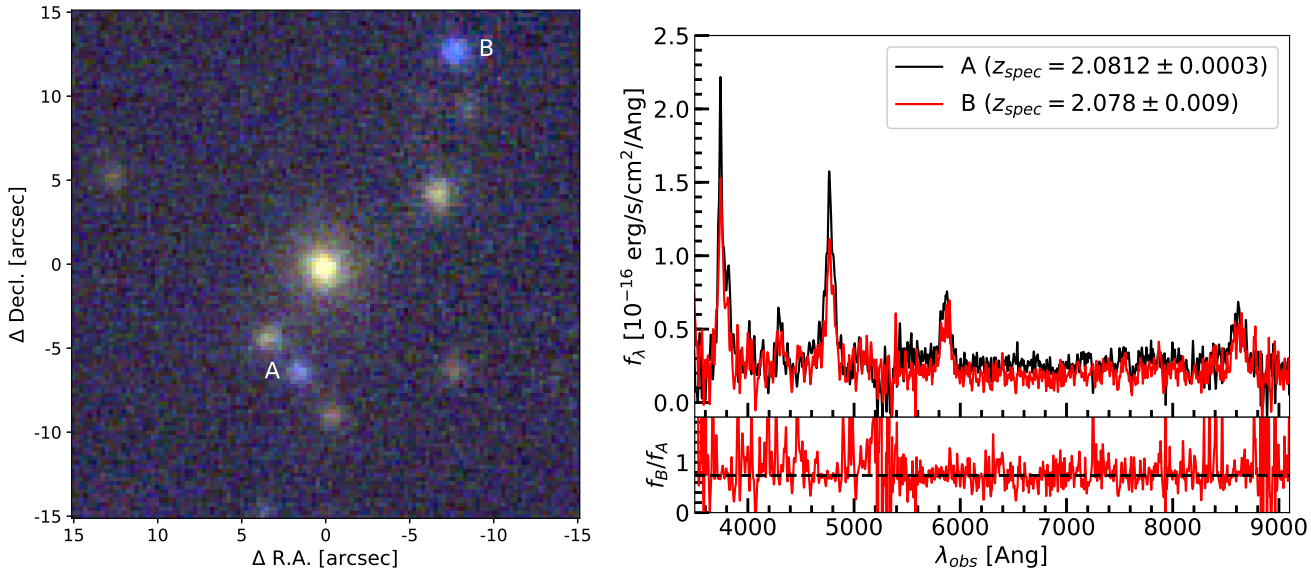

Figure 11. *Left*: PanSTARRS imaging data of the new wide-separation strongly-lensed quasar SDSS J1326+4806. The bright object in the center is a BCG at $z = 0.396$. Object A is a spectroscopically confirmed quasar at $z_A = 2.0812 \pm 0.0003$. Object B is classified as an AGN in our catalogue with $P_{\rm RF} = 0.93$. The separation between A and B is $21\farcs06$. *Top right*: Smoothed WHT spectra of A (black) and B (red). Fitting the spectrum of B confirms it to be a quasar at $z_B = 2.078 \pm 0.009$. *Bottom right*: Flux ratio of B to A. The median flux ratio is 0.74, as indicated by the black dashed line.

(Dahle et al. 2013), are not recovered because they only have zero and one lensed quasar image detected in *Gaia* DR2. In addition, we identify a high-probability strongly-lensed AGN candidate — SDSS J1326+4806. The majority of the rest of the BCGs have AGN candidates with significantly different optical colours, and therefore unlikely to be images of the same AGN, or the BCG does not lie between the AGN candidates.

The left panel in Figure 11 shows a colour cutout centered on the BCG of SDSS J1326+4806 made from *gri* imaging data from the Panoramic Survey Telescope and Rapid Response System (PanSTARRS) survey (Chambers et al. 2016). The BCG, at R.A.=201.50006°, Decl.=48.11208°, is an SDSS spectroscopically-confirmed massive early-type galaxy at $z = 0.396$. Two blue, point-like sources, labeled as A and B, are located on either side of the BCG, consistent with the image configuration of a doubly lensed system. The separation between A and B is $21\farcs06$. Our AGN classifier suggests that A and B are very likely to be AGNs with $P_{\rm RF}$ (A) = 0.99 and $P_{\rm RF}$ (B) = 0.93. In fact, source A was spectroscopically confirmed to be a $z_A = 2.0812 \pm 0.0003$ AGN by the Baryon Oscillation Spectroscopic Survey (Bolton et al. 2012).

To determine the nature and redshift of B, we obtained low-resolution spectra for A and B with the Intermediate-dispersion Spectrograph and Imaging System on the William Herschel Telescope (WHT) on the night of February 11, 2019. The R158R (1.81 Å pixel^{-1}) and R300B (0.86 Å pixel^{-1}) gratings were used on the red and blue arms, respectively, along with the standard 5300 Å dichroic and GG495 second-order cut filter in the red arm. The right panel in Figure 11 shows the smoothed, reduced spectra for A (black) and B (red), which confirms that B is indeed an AGN with a spectral profile that appears to be similar to A. Fitting the spectrum of B using a linear combination of quasar eigenspectra following Bolton et al. (2012) further suggests $z_B = 2.078 \pm 0.009$, consistent with the spectroscopic redshift of A.

Both A and B have experienced substantial variations in brightness over the past $\sim$ 16 years. The SDSS data in 2003 showed that the *g*-band AB magnitude of A and B were about 21 mag and 22 mag respectively, with A being brighter than B. The multi-epoch photometry from PanSTARRS DR2 taken between the year of 2011 and 2014 showed significant brightness variations, with the largest change reaching more than 1 magnitude. In particular, B was brighter than A when averaging over the PanSTARRS period, as indicated in the left panel of Figure 11. The PanSTARRS *g*-band mean AB magnitude of A and B were about 21.6 mag and 21 mag respectively. The median flux ratio of B to A from recent WHT spectroscopic data is 0.74, indicating that A now has become brighter than B again. Nevertheless, no clear correlation between brightness variations in A and B is detected.

We consider a simple lens model for SDSS J1326+4806 consisting of a singular isothermal sphere (SIS) mass distribution in an external shear field. The total number of free parameters is 7 (assuming the SIS mass component and the external shear field are co-centred). Considering the substantial brightness variations in A and B, we only use the relative positions of the BCG, A, and B as constraints, but not the flux ratios between A and B. As a result, the number of free parameters is more than the number of constraints, and no unique lens model can be determined. Nevertheless, the goal of this procedure is to examine whether the image configuration of SDSS J1326+4806 can be explained by a typical lens model with reasonable parameters. We optimize the model parameters with the `lensmodel` toolkit (Keeton 2001), and find that the relative positions can be perfectly recovered (as expected for this under-constrained problem). All the model parameters have reasonable values. The best-fit Einstein ra-

dius of the SIS component is $10\farcs3$, consistent with the $21\farcs06$ separation between A and B. It suggests that the total projected mass within the Einstein radius is $\approx 2.1 \times 10^{13} M_{\odot}$. On the other hand, Wen et al. (2012) estimated the r_{200} radius of this cluster to be 1.51 Mpc. Assuming that the dark matter distribution of this cluster follows a simple Navarro-Frenk-White (NFW) profile (Navarro et al. 1996, 1997), the total dark-matter mass within the sphere of radius r_{200} is approximately $M_{200} = 5.6 \times 10^{14} M_{\odot}$. The typical concentration for dark-matter halos of this mass scale at $z \sim 0.4$ is about 5 (e.g., Duffy et al. 2008; Macciò et al. 2008; Zhao et al. 2009; Klypin et al. 2011; Prada et al. 2012; Auger et al. 2013; Diemer & Kravtsov 2015). The total projected dark-matter mass within the Einstein radius (57 kpc in physical unit) is thus $2.0 \times 10^{13} M_{\odot}$, in close agreement with the required mass by strong gravitational lensing.

Based on the analyses above, SDSS J1326+4806 has a very high probability of being a strongly-lensed AGN. Follow-up higher-resolution spectroscopic and deeper imaging data could pin down the lensing nature of this system. If confirmed, SDSS J1326+4806 will be the second most widely-separated strongly-lensed AGN discovered so far. More wide-separation, strongly-lensed AGN systems are expected to be discovered by cross-matching the C75 AGN catalogue with other catalogues of galaxy groups and clusters.

5.3 Future Prospects

It is worth mentioning that as more repeated *Gaia* observations will be conducted in the coming years, we expect the overall limiting magnitude of future *Gaia* data releases to become similar to the current value of the Boötes field or even deeper in some regions. Considering that in the Boötes reference field, the current completeness at the C75 threshold is 93.6% and the reliability at the R85 threshold is 90.7%, we expect the quality of AGN catalogues built from future *Gaia* data releases to improve substantially both in completeness and reliability. In addition, the sample size and quality in astrometry and photometry of future *Gaia* data releases are also expected to improve with beneficial effects for future AGN catalogues.

On average, *Gaia* will measure astrometrically each of its targets $\sim$ 70 times over the nominal five-year operation period since 2013, and 10 photometric measurements in the G band are made during each astrometric measurement (Gaia Collaboration et al. 2016). In total, every *Gaia* source will therefore have $\sim$ 700 G-band measurements in five years. In *Gaia* DR2 (data from the first 22 months of operation), the average and highest number of G-band measurements for AGNs in the *Gaia*-unWISE-DR14 QSO sample is 211 and 1100 respectively. However, *Gaia* will not release the multi-epoch photometric data until the end of the mission, at which point all the AGN candidates in our catalogues will have *Gaia* light curves spanning a time scale of five years. These light curves will be helpful in identifying variable AGNs and even optical changing-look AGNs. These are AGNs that show optical spectral feature transitions involving appearance and disappearance of broad emission lines on time scales of years or decades. There are a few tens of known optical changing-look AGNs so far (e.g., Denney et al. 2014; LaMassa et al. 2015; Ruan et al. 2016; MacLeod et al. 2016; Gezari et al. 2017; Yang et al. 2018; Wang et al. 2018a). The physical mechanisms responsible for the transitions are still not fully understood. A large sample of variable AGNs and changing-look AGNs with a wide range of properties including redshift, luminosity, and black hole mass can help to better understand the structure of the accretion disc and broad line region and the evolution of AGNs. Our AGN catalogues, which include AGNs up to redshift $\sim$ 4, can be a useful input catalogue for future spectroscopic surveys that study AGNs and large scale structures, especially ones in the southern hemisphere, for example, 4MOST (de Jong et al. 2019; Merloni et al. 2019; Richard et al. 2019).

6 CONCLUSION

In this work, we perform an AGN/non-AGN classification of more than 641 million sources in the *Gaia*-unWISE sample across the entire sky using astrometric and photometric data from the latest data releases of *Gaia* and *WISE*. We use the supervised machine learning algorithm random forest (RF) to estimate the probability of a source being an AGN, $P_{\rm RF}$. We construct two AGN catalogues, C75 and R85, by applying two different $P_{\rm RF}$ threshold cuts that deliver an overall completeness of 75% ($\approx$ 90% at $G \leq 20$ mag) and an overall reliability of 85% respectively. The C75 catalogue contains 2,734,464 AGN candidates with $P_{\rm RF} \geq 0.69$, of which 2,182,193 AGN candidates with $P_{\rm RF} \geq 0.94$ comprise the R85 catalogue (Figure 7). We estimate the photometric redshifts of the AGN candidates using a RF regressor. We find that 76,620 and 1,602 AGN candidates in the C75 catalogue are predicted to be at redshifts higher than 3 and 4 respectively.

Comparing to *WISE*-only AGN selection techniques used in Stern et al. (2012) and Assef et al. (2018), our RF classifier using both optical and mid-IR data achieves significantly better true positive and false positive rates when applied to the *Gaia*-unWISE sample (see Table 2). Among the 1,166,573 known AGNs and high-confidence AGN candidates in the MILLIQUAS that are also catalogued in the *Gaia*-unWISE sample, 94.7% and 89.4% are successfully recovered in our C75 and R85 catalogues. Cross-matching against the known AGN compilation including almost 29 million AGNs and AGN candidates with an aggressive matching radius of $5''$, we find that at least $\approx$0.91 (0.52) million AGN candidates in our C75 (R85) catalogue are new discoveries.

The large sample of AGN candidates provided in this work is a useful resource for many applications. As an example, we have identified a strongly-lensed AGN candidate, SDSS J1326+4806, with an image separation of $21\farcs06$ by cross-matching the C75 catalogue with a sample of known brightest cluster galaxies or BCGs (Figure 11). The BCG in SDSS J1326+4806 is at $z = 0.396$, and the two AGN candidates on either side of the BCG are spectroscopically confirmed to be true AGNs at $z \sim 2.08$ with similar spectral profiles. A simple singular isothermal sphere plus external shear lens model can explain the relative positions between the BCG and the two AGNs. The total mass within the inferred Einstein radius required by strong gravitational lensing is in close agreement with the mass of dark mat-

ter within the same aperture when assuming dark matter in SDSS J1326+4806 following a simple NFW profile. Follow-up imaging and spectroscopic data will pin down the lensing nature of this system.

Moreover, all the AGN candidates in our catalogue will eventually have light curves consisting of, on average, ∼70-epoch photometry across five years from *Gaia*, which are very helpful for identifying highly-variable AGNs and changing-look AGNs. Our AGN catalogues are also useful for future spectroscopic surveys such as 4MOST.

ACKNOWLEDGEMENTS

We thank Qiusheng Gu, Paul Hewett, George Lansbury, Peter McGill, Leigh Smith, and Zhonglue Wen for helpful discussions. We thank the anonymous referee for a thoughtful report. Y.S. has been supported by the Royal Society – K.C. Wong International Fellowship (NF170995). SK is partially supported by NSF grant AST-1813881, Heising-Simon's foundation grant 2018-1030. This work made used of the Whole Sky Database (wsdb) created by Sergey Koposov and maintained at the Institute of Astronomy, Cambridge by Sergey Koposov, Vasily Belokurov and Wyn Evans with financial support from the Science & Technology Facilities Council (STFC) and the European Research Council (ERC). This software made use of the Q3C software (Koposov & Bartunov 2006). This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC,`https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This publication makes use of data products from the *Wide-field Infrared Survey Explorer*, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, and NEOWISE, which is a project of the Jet Propulsion Laboratory/ California Institute of Technology. *WISE* and NEOWISE are funded by the National Aeronautics and Space Administration.

REFERENCES

Ai Y. L., et al., 2016, AJ, 151, 24
Ananna T. T., et al., 2017, ApJ, 850, 66
Arenou F., et al., 2018, A&A, 616, A17
Assef R. J., et al., 2010, ApJ, 713, 970
Assef R. J., et al., 2013, ApJ, 772, 26
Assef R. J., Stern D., Noirot G., Jun H. D., Cutri R. M., Eisenhardt P. R. M., 2018, ApJS, 234, 23
Auger M. W., Budzynski J. M., Belokurov V., Koposov S. E., McCarthy I. G., 2013, MNRAS, 436, 503
Bañados E., et al., 2018, Nature, 553, 473
Bautista J. E., et al., 2017, A&A, 603, A12
Belokurov V., Erkal D., Deason A. J., Koposov S. E., De Angeli F., Evans D. W., Fraternali F., Mackey D., 2017, MNRAS, 466, 4711
Blandford R. D., McKee C. F., 1982, ApJ, 255, 419
Bolton A. S., et al., 2012, AJ, 144, 144
Bovy J., et al., 2011, ApJ, 729, 141
Breiman L., 2001, Machine Learning, 45, 5
Cardelli J. A., Clayton G. C., Mathis J. S., 1989, ApJ, 345, 245
Carliles S., Budavári T., Heinis S., Priebe C., Szalay A. S., 2010, ApJ, 712, 511
Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483
Carrasco D., et al., 2015, A&A, 584, A44
Chambers K. C., et al., 2016, arXiv e-prints, p. arXiv:1612.05560
Chen T., Guestrin C., 2016, arXiv e-prints, p. arXiv:1603.02754
Chen B.-Q., et al., 2019, MNRAS, 483, 4277
Chung S. M., et al., 2014, ApJ, 790, 54
Croom S. M., Smith R. J., Boyle B. J., Shanks T., Miller L., Outram P. J., Loaring N. S., 2004, MNRAS, 349, 1397
Dahle H., et al., 2013, ApJ, 773, 146
Delubac T., et al., 2015, A&A, 574, A59
Denney K. D., et al., 2014, ApJ, 796, 134
Diemer B., Kravtsov A. V., 2015, ApJ, 799, 108
Dong X. Y., et al., 2018, AJ, 155, 189
Dubath P., et al., 2011, MNRAS, 414, 2602
Dubois Y., Gavazzi R., Peirani S., Silk J., 2013, MNRAS, 433, 3297
Duffy A. R., Schaye J., Kay S. T., Dalla Vecchia C., 2008, MNRAS, 390, L64
Evans D. W., et al., 2018, A&A, 616, A4
Fabian A. C., 2012, ARA&A, 50, 455
Fan X., et al., 2006, AJ, 131, 1203
Fey A. L., et al., 2015, AJ, 150, 58
Flesch E. W., 2015, Publ. Astron. Soc. Australia, 32, e010
Flesch E. W., 2016, Publ. Astron. Soc. Australia, 33, e052
Fotopoulou S., Paltani S., 2018, A&A, 619, A14
Fukugita M., Futamase T., Kasai M., 1990, MNRAS, 246, 24P
Gaia Collaboration et al., 2016, A&A, 595, A1
Gaia Collaboration et al., 2018a, A&A, 616, A1
Gaia Collaboration et al., 2018b, A&A, 616, A14
Galametz A., et al., 2012, ApJ, 749, 169
Gezari S., et al., 2017, ApJ, 835, 144
Hewett P. C., Foltz C. B., Chaffee F. H., 1995, AJ, 109, 1498
Ilbert O., et al., 2009, ApJ, 690, 1236
Inada N., et al., 2003, Nature, 426, 810
Inada N., et al., 2006, ApJ, 653, L97
Jannuzi B. T., Dey A., 1999, in Weymann R., Storrie-Lombardi L., Sawicki M., Brunner R., eds, Astronomical Society of the Pacific Conference Series Vol. 191, Photometric Redshifts and the Detection of High Redshift Galaxies. p. 111
Jayasinghe T., et al., 2019, MNRAS, 486, 1907
Jin X., Zhang Y., Zhang J., Zhao Y., Wu X.-b., Fan D., 2019, MNRAS, 485, 4539
Kang X., Jing Y. P., Silk J., 2006, ApJ, 648, 820
Karachentsev I. D., Makarov D. I., Kaisina E. I., 2013, AJ, 145, 101
Keeton C. R., 2001, ArXiv Astrophysics e-prints,
Klypin A. A., Trujillo-Gomez S., Primack J., 2011, ApJ, 740, 102
Kochanek C. S., 1995, ApJ, 453, 545
Kochanek C. S., 1996, ApJ, 466, 638
Koenig X. P., Leisawitz D. T., Benford D. J., Rebull L. M., Padgett D. L., Assef R. J., 2012, ApJ, 744, 130
Koposov S., Bartunov O., 2006, in Gabriel C., Arviset C., Ponz D., Enrique S., eds, Astronomical Society of the Pacific Conference Series Vol. 351, Astronomical Data Analysis Software and Systems XV. p. 735
LaMassa S. M., et al., 2015, ApJ, 800, 144
Lacy M., et al., 2004, ApJS, 154, 166
Lang D., 2014, AJ, 147, 108
Lemon C. A., Auger M. W., McMahon R. G., 2019, MNRAS, 483, 4242
Li S.-L., Cao X., 2008, MNRAS, 387, L41
Li G. L., Mao S., Jing Y. P., Lin W. P., Oguri M., 2007, MNRAS, 378, 469
Lindegren L., et al., 2018, A&A, 616, A2
Liu H. T., Bai J. M., Zhao X. H., Ma L., 2008, ApJ, 677, 884

Lopes A. M., Miller L., 2004, MNRAS, 348, 519
Lynden-Bell D., 1969, Nature, 223, 690
Ma C., et al., 1998, AJ, 116, 516
MacLeod C. L., et al., 2010, ApJ, 721, 1014
MacLeod C. L., et al., 2016, MNRAS, 457, 389
Macciò A. V., Dutton A. A., van den Bosch F. C., 2008, MNRAS, 391, 1940
Maddox N., Hewett P. C., Péroux C., Nestor D. B., Wisotzki L., 2012, MNRAS, 424, 2876
Mancone C. L., Gonzalez A. H., Brodwin M., Stanford S. A., Eisenhardt P. R. M., Stern D., Jones C., 2010, ApJ, 720, 284
McGreer I. D., et al., 2013, ApJ, 768, 105
Meisner A. M., Lang D., Schlegel D. J., 2017a, AJ, 153, 38
Meisner A. M., Lang D., Schlegel D. J., 2017b, AJ, 154, 161
Merloni A., et al., 2019, The Messenger, 175, 42
Mignard F., et al., 2016, A&A, 595, A5
Mortlock D. J., et al., 2011, Nature, 474, 616
Myers A. D., et al., 2015, ApJS, 221, 27
Nakoneczny S., Bilicki M., Solarz A., Pollo A., Maddox N., Spiniello C., Brescia M., Napolitano N. R., 2019, A&A, 624, A13
Narayan R., White S. D. M., 1988, MNRAS, 231, 97p
Navarro J. F., Frenk C. S., White S. D. M., 1996, ApJ, 462, 563
Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
Nikutta R., Hunt-Walker N., Nenkova M., Ivezić Ž., Elitzur M., 2014, MNRAS, 442, 3361
O'Donnell J. E., 1994, ApJ, 422, 158
Oguri M., et al., 2004, ApJ, 605, 78
Oguri M., et al., 2012, AJ, 143, 120
Ostrovski F., et al., 2017, MNRAS, 465, 4325
Papovich C., 2008, ApJ, 676, 206
Papovich C., et al., 2010, ApJ, 716, 1503
Pâris I., et al., 2018, A&A, 613, A51
Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825
Pichara K., Protopapas P., Kim D.-W., Marquette J.-B., Tisserand P., 2012, MNRAS, 427, 1284
Planck Collaboration et al., 2016, A&A, 594, A13
Pons E., McMahon R. G., Simcoe R. A., Banerji M., Hewett P. C., Reed S. L., 2019, MNRAS, 484, 5142
Prada F., Klypin A. A., Cuesta A. J., Betancort-Rijo J. E., Primack J., 2012, MNRAS, 423, 3018
Rebull L. M., et al., 2010, ApJS, 186, 259
Rees M. J., 1984, ARA&A, 22, 471
Richard J., et al., 2019, The Messenger, 175, 50
Richards G. T., et al., 2002, AJ, 123, 2945
Richards G. T., et al., 2004, ApJS, 155, 257
Richards G. T., et al., 2009, ApJS, 180, 67
Richards J. W., Homrighausen D., Freeman P. E., Schafer C. M., Poznanski D., 2012, MNRAS, 419, 1121
Richards G. T., et al., 2015, ApJS, 219, 39
Riello M., et al., 2018, A&A, 616, A3
Ruan J. J., et al., 2016, ApJ, 826, 188
Sandage A., Wyndham J. D., 1965, ApJ, 141, 328
Schindler J.-T., Fan X., McGreer I. D., Yang Q., Wu J., Jiang L., Green R., 2017, ApJ, 851, 13
Schlafly E. F., Meisner A. M., Green G. M., 2019, ApJS, 240, 30
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525
Schneider D. P., et al., 2010, AJ, 139, 2360
Secrest N. J., Dudik R. P., Dorland B. N., Zacharias N., Makarov V., Fey A., Frouard J., Finch C., 2015, ApJS, 221, 12
Shen Y., et al., 2015, ApJS, 216, 4
Shen Y., et al., 2019, ApJ, 873, 35
Shu Y., Marques-Chaves R., Evans N. W., Pérez-Fournon I., 2018, MNRAS, 481, L136
Silk J., Rees M. J., 1998, A&A, 331, L1
Smith R. J., Croom S. M., Boyle B. J., Shanks T., Miller L., Loaring N. S., 2005, MNRAS, 359, 57
Stern D., et al., 2005, ApJ, 631, 163
Stern D., et al., 2012, ApJ, 753, 30
Tanaka Y., et al., 1995, Nature, 375, 659
Turner E. L., 1990, ApJ, 365, L43
Vanden Berk D. E., et al., 2004, ApJ, 601, 692
Wambsganss J., Cen R., Ostriker J. P., Turner E. L., 1995, Science, 268, 274
Wang F., et al., 2016, ApJ, 819, 24
Wang J., Xu D. W., Wei J. Y., 2018a, ApJ, 858, 49
Wang F., et al., 2018b, ApJ, 869, L9
Warren S. J., Hewett P. C., Irwin M. J., McMahon R. G., Bridgeland M. T., 1987, Nature, 325, 131
Wen Z. L., Han J. L., 2011, ApJ, 734, 68
Wen Z. L., Han J. L., 2015, ApJ, 807, 178
Wen Z. L., Han J. L., 2018, MNRAS, 481, 4158
Wen Z. L., Han J. L., Liu F. S., 2012, ApJS, 199, 34
Wen Z. L., Han J. L., Yang F., 2018, MNRAS, 475, 343
Williams W. L., et al., 2018, MNRAS, 475, 3429
Wright E. L., et al., 2010, AJ, 140, 1868
Wu X.-B., Jia Z., 2010, MNRAS, 406, 1583
Wu X.-B., Hao G., Jia Z., Zhang Y., Peng N., 2012, AJ, 144, 49
Wu X.-B., et al., 2015, Nature, 518, 512
Wyrzykowski Ł., et al., 2014, Acta Astron., 64, 197
Yan L., et al., 2013, AJ, 145, 55
Yang Q., et al., 2018, ApJ, 862, 109
Yao S., et al., 2019, ApJS, 240, 6
Zhao D. H., Jing Y. P., Mo H. J., Börner G., 2009, ApJ, 707, 354
Zhao G.-B., et al., 2019, MNRAS, 482, 3497
de Jong R. S., et al., 2019, The Messenger, 175, 3

APPENDIX A: UNWISE COMPLETENESS AND LIMITING MAGNITUDE MAPS

Figures A1 and A2 show the spatial distributions and one-dimensional cumulative sky coverage histograms of $W1_{\rm peak}$, $W1_{99}$, $W2_{\rm peak}$, and $W2_{99}$ for the unWISE sub-samples.

APPENDIX B: DESCRIPTION OF THE CATALOGUE

The C75 AGN catalogue is publicly available as a FITS file at `https://www.ast.cam.ac.uk/~ypshu/AGN_Catalogues.html`. Descriptions of all the columns in the FITS file are summarised in Table B1. The R85 AGN catalogue can be constructed from the C75 AGN catalogue by applying a probability threshold cut of PROB_RF ≥ 0.94.

This paper has been typeset from a TEX/LATEX file prepared by the author.

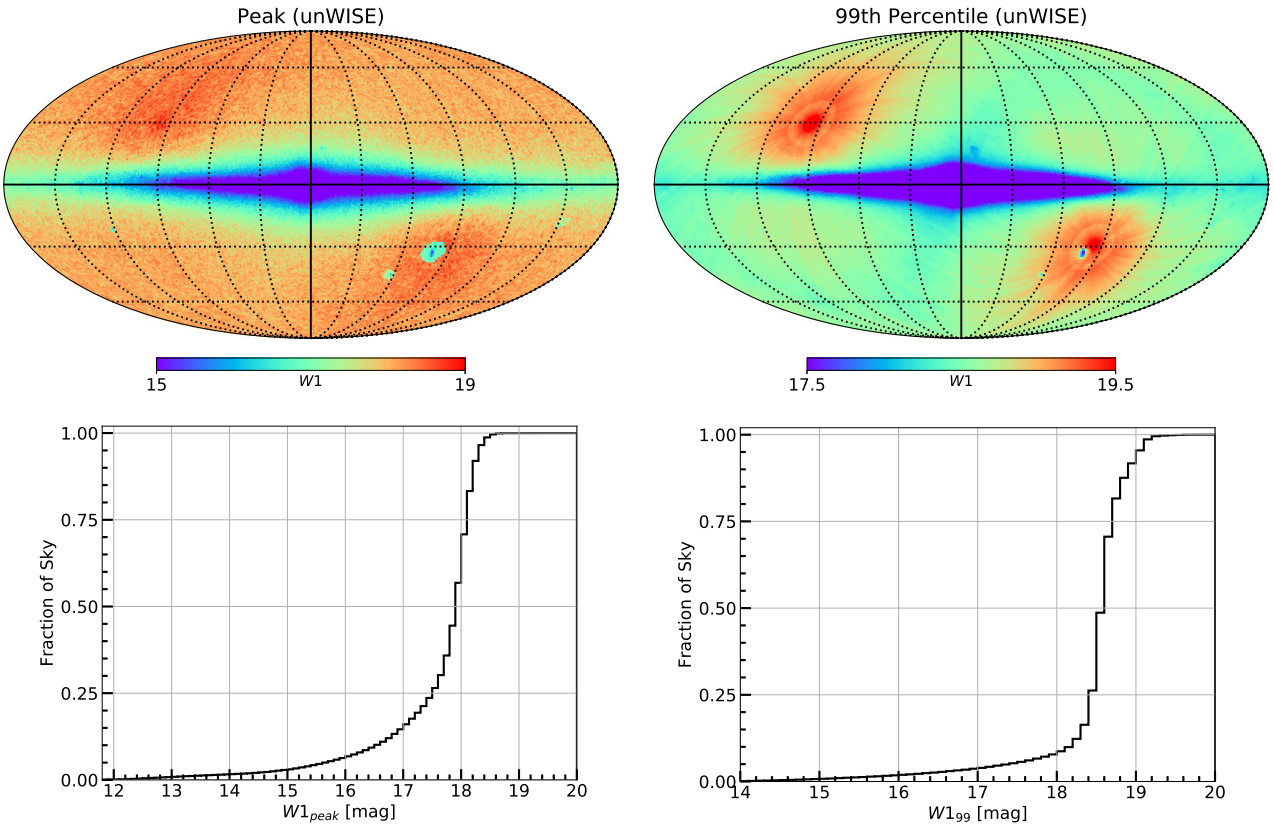

Figure A1. The same set of plots as Figure 1, but for $W1$ for 2,094,307,508 unWISE sources with $W1 \geq 8$ mag.

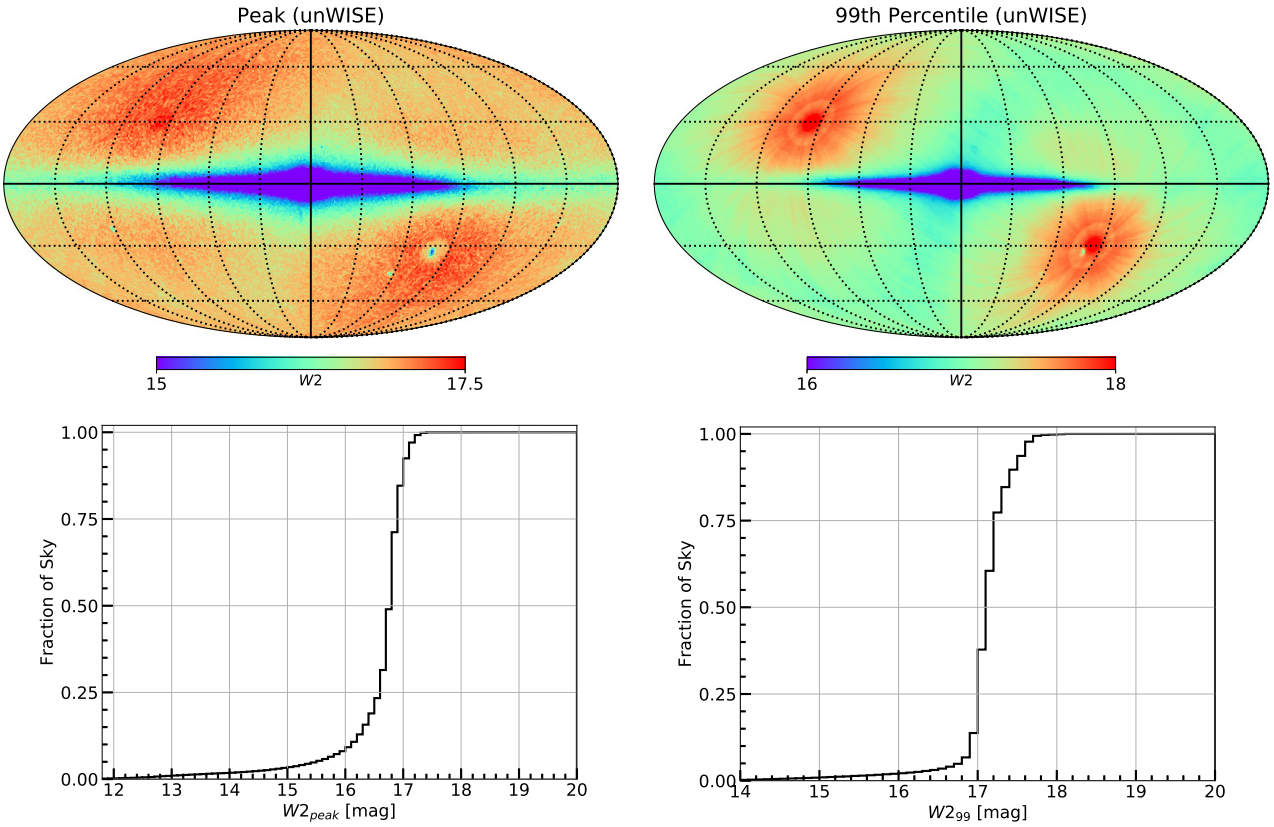

Figure A2. The same set of plots as Figure 1, but for $W2$ for 1,180,720,229 unWISE sources with $W2 \geq 8$ mag.

Table B1. Format of the AGN catalogue FITS file.

Column	Name	Description
1	RA	Right ascension in decimal degrees from *Gaia* DR2 (J2015.5)
2	DEC	Declination in decimal degrees from *Gaia* DR2 (J2015.5)
3	GAIA_SOURCEID	Unique Gaia source identifier `source_id`
4	UNWISE_OBJID	Unique unWISE source identifier `unwise_objid`
5	PLX	Parallax in milli-arcsec (mas) from *Gaia* DR2, set to -999 if null
6	PLX_ERR	Error in parallax in mas from *Gaia* DR2, set to -999 if null
7	PMRA	Proper motion in right ascension direction (mas/year) from *Gaia* DR2, set to -999 if null
8	PMRA_ERR	Error in proper motion in right ascension direction (mas/year) from *Gaia* DR2, set to -999 if null
9	PMDEC	Proper motion in declination direction (mas/year) from *Gaia* DR2, set to -999 if null
10	PMDEC_ERR	Error in proper motion in declination direction (mas/year) from *Gaia* DR2, set to -999 if null
11	PLXSIG	Parallax significance defined as $\mid \frac{\text{parallax}}{\text{parallax\_error}} \mid$, set to -999 if null
12	PMSIG	Proper motion significance defined as $\sqrt{(\frac{\text{pmra}}{\text{pmra\_error}})^2 + (\frac{\text{pmdec}}{\text{pmdec\_error}})^2}$, set to -999 if null
13	EBV	Galactic E(B-V) reddening from Schlegel et al. (1998)
14	N_OBS	Number of observations contributing to G photometry
15	G	*Gaia* DR2 G-band mean magnitude (extinction corrected)
16	BP	*Gaia* DR2 BP-band mean magnitude (extinction corrected)
17	RP	*Gaia* DR2 RP-band mean magnitude (extinction corrected)
18	W1	unWISE $W1$-band magnitude
19	W2	unWISE $W2$-band magnitude
20	BP_G	*Gaia* DR2 BP-G colour (extinction corrected), set to 999 if null
21	BP_RP	*Gaia* DR2 BP-RP colour (extinction corrected), set to 999 if null
22	G_RP	*Gaia* DR2 G-RP colour (extinction corrected), set to 999 if null
23	G_W1	*Gaia* DR2 G - unWISE $W1$ colour (extinction corrected)
24	GW_SEP	Separation (in arcsec) between a *Gaia* source and its unWISE counterpart
25	W1_W2	unWISE $W1$-$W2$ colour
26	G_VAR	Variation in *Gaia* G-band flux defined as $\sqrt{\texttt{PHOT\_G\_N\_OBS}} \times \frac{\texttt{PHOT\_G\_MEAN\_FLUX\_ERROR}}{\texttt{PHOT\_G\_MEAN\_FLUX}}$
27	BPRP_EF	BP/RP excess factor from *Gaia* DR2 (`PHOT_BP_RP_EXCESS_FACTOR`)
28	AEN	Astrometric excess noise from *Gaia* DR2 (`ASTROMETRIC_EXCESS_NOISE`)
29	GOF	Goodness-of-fit statistic of the astrometric solution from *Gaia* DR2 (`ASTROMETRIC_GOF_AL`)
30	CNT1	Number of *Gaia* DR2 sources within a $1''$-radius circular aperture
31	CNT2	Number of *Gaia* DR2 sources within a $2''$-radius circular aperture
32	CNT4	Number of *Gaia* DR2 sources within a $4''$-radius circular aperture
33	CNT8	Number of *Gaia* DR2 sources within a $8''$-radius circular aperture
34	CNT16	Number of *Gaia* DR2 sources within a $16''$-radius circular aperture
35	CNT32	Number of *Gaia* DR2 sources within a $32''$-radius circular aperture
36	PHOT_Z	Photometric redshift
37	PROB_RF	AGN probability