



The CatSouth Quasar Candidate Catalog for the Southern Sky and a Unified All-sky Catalog Based on Gaia DR3

Yuming Fu (傅煜铭)^{1,2}, Xue-Bing Wu^{3,4,5}, R. J. Bouwens¹, Karina I. Caputi^{2,6}, Yuxuan Pang^{3,4}, Rui Zhu^{3,4}, Da-Ming Yang¹, Jin Qin^{3,4}, Huimei Wang^{3,4}, Christian Wolf^{7,8}, Yifan Li¹, Ravi Joshi⁹, Yanxia Zhang⁵, Zhi-Ying Huo⁵, and Y. L. Ai^{10,11}

¹ Leiden Observatory, Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands; yfu@strw.leidenuniv.nl

² Kapteyn Astronomical Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands

³ Department of Astronomy, School of Physics, Peking University, Beijing 100871, People's Republic of China; wuxb@pku.edu.cn

⁴ Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing 100871, People's Republic of China

⁵ National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, People's Republic of China

⁶ Cosmic Dawn Center (DAWN), Copenhagen, Denmark

⁷ Research School of Astronomy and Astrophysics, Australian National University, Canberra ACT 2611, Australia

⁸ Centre for Gravitational Astrophysics, Australian National University, Canberra ACT 2600, Australia

⁹ Indian Institute of Astrophysics, Koramangala, Bangalore 560034, India

¹⁰ College of Engineering Physics, Shenzhen Technology University, Shenzhen 518118, People's Republic of China

¹¹ Shenzhen Key Laboratory of Ultraintense Laser and Advanced Material Technology, Shenzhen 518118, People's Republic of China

Received 2025 March 27; revised 2025 June 17; accepted 2025 June 18; published 2025 August 6

Abstract

The Gaia DR3 has provided a large sample of more than 6.6 million quasar candidates with high completeness but low purity. Previous work on the CatNorth quasar candidate catalog has shown that including external multiband data and applying machine learning methods can efficiently purify the original Gaia DR3 quasar candidate catalog and improve the redshift estimates. In this paper, we extend the Gaia DR3 quasar candidate selection to the Southern Hemisphere using data from SkyMapper, CatWISE, and Visible and Infrared Survey Telescope for Astronomy surveys. We train an XGBoost classifier on a unified set of high-confidence stars and spectroscopically confirmed quasars and galaxies. For sources with available Gaia BP/RP spectra, spectroscopic redshifts are derived using a pretrained convolutional neural network (RegNet). We also train an ensemble photometric redshift estimation model based on XGBoost, TabNet, and FT-Transformer, achieving a root mean square error of 0.2256 and a normalized median absolute deviation of 0.0187 on the validation set. By merging CatSouth with the previously published CatNorth catalog, we construct the unified all-sky CatGlobe catalog with nearly 1.9 million sources at $G < 21$, providing a comprehensive and high-purity quasar candidate sample for future spectroscopic and cosmological investigations.

Unified Astronomy Thesaurus concepts: Active galactic nuclei (16); Astrostatistics techniques (1886); Catalogs (205); Classification (1907); Quasars (1319); Redshift surveys (1378)

Materials only available in the online version of record: machine-readable tables

1. Introduction

Quasars are luminous active galactic nuclei (AGNs) powered by accretion onto supermassive black holes. Observable across vast cosmic distances, quasars are ideal probes for many astrophysical and cosmological studies. Comprehensive quasar samples enable investigations into the formation and evolution of supermassive black holes (e.g., X.-B. Wu et al. 2015; E. Bañados et al. 2018; K. Inayoshi et al. 2020; X. Fan et al. 2023), the coevolution of black holes and galaxies (e.g., T. Di Matteo et al. 2005; J. Kormendy & L. C. Ho 2013), and the structure and composition of the intergalactic medium (e.g., R. J. Weymann et al. 1981; M. J. Rees 1986; J. R. Trump et al. 2006). Additionally, quasar distribution traces the large-scale structure of the Universe, providing critical constraints on cosmological models (e.g., D. J. Eisenstein et al. 2011; K. S. Dawson et al. 2013; M. R. Blanton et al. 2017). Quasars also serve as reference sources for celestial frames with their

small parallaxes and proper motions (e.g., C. Ma et al. 2009; F. Mignard et al. 2016; Gaia Collaboration et al. 2018, 2022).

Until 2023, nearly 1 million quasars had been spectroscopically identified (see, e.g., The Million Quasar Catalog; E. W. Flesch 2023). Most of these quasars were identified by the Sloan Digital Sky Surveys (e.g., R. Ahumada et al. 2020; B. W. Lyke et al. 2020; A. Almeida et al. 2023). Other representative quasar surveys include the 2dF QSO Redshift Survey (2QZ; S. M. Croom et al. 2004), the 2dF-SDSS LRG and QSO survey (2SLAQ; S. M. Croom et al. 2009), the LAMOST quasar survey (Y. L. Ai et al. 2016; X. Y. Dong et al. 2018; S. Yao et al. 2019; J.-J. Jin et al. 2023), and the Early Data Release of the Dark Energy Spectroscopic Instrument (DESI EDR; DESI Collaboration et al. 2024). More recently, DESI Data Release 1 has published approximately 1.6 million spectroscopically classified quasars (DESI Collaboration et al. 2025). Nevertheless, the completeness and identification efficiency of the existing quasar samples are still restricted by factors such as the bias of candidate selection methods, instrument performance, and ground-based observation conditions.

Combining (candidate) quasars selected with different methods can effectively increase sample completeness. A



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

good example is the Gaia DR3 quasar candidate catalog¹² (hereafter GDR3 QSO candidate catalog; Gaia Collaboration et al. 2023a, 2023c) with 6.6 million quasar candidates selected by at least one of the several different classification modules, including the Discrete Source Classifier (DSC), the Quasar Classifier (QSOC), the variability classification module, the surface brightness profile module, and the Gaia DR3 Celestial Reference Frame source table. In particular, the DSC uses the Gaia BP/RP spectrum together with the mean G -band magnitude, the G -band variability, the parallax, and the proper motion to classify each Gaia source probabilistically, which is less biased than color cuts in selecting quasars. Although highly complete, the GDR3 QSO candidate catalog has low purity ($\sim 52\%$ as estimated by Gaia Collaboration et al. 2023a), which limits its further application in astrophysical and cosmological studies.

To extract purer subsamples from the GDR3 QSO candidate catalog, approaches incorporating photometric data from external mid-infrared and optical surveys have been proposed. For example, by applying cuts on Gaia and unWISE (D. Lang 2014) colors and Gaia proper motions to remove nonquasar contaminants (stars and galaxies), K. Storey-Fisher et al. (2024) constructed the Quaia quasar catalog with nearly 1.3 million sources at $G < 20.5$. Using data from Gaia, Pan-STARRS1 (PS1; K. C. Chambers et al. 2016), and CatWISE2020 (F. Marocco et al. 2021) and machine learning methods, Y. Fu et al. (2024) presented CatNorth, another improved Gaia DR3 quasar candidate catalog with more than 1.5 million sources down to the Gaia limiting magnitude in the 3π sky of PS1 footprint ($\delta > -30^\circ$). Y. Fu et al. (2024) have shown that the machine learning–based CatNorth catalog has higher completeness than Quaia while obtaining similar purity. This proves that the machine learning method can better disentangle different celestial objects in the high-dimensional space than cuts on two-dimensional planes. In addition, the inclusion of PS1 photometry in Y. Fu et al. (2024) also leads to higher photometric redshift accuracy in CatNorth than in Quaia.

In contrast to the Northern Hemisphere, where systematic surveys such as the Sloan Digital Sky Survey (SDSS) and DESI have yielded extensive quasar catalogs covering large sky areas, the southern sky has long suffered from limited multiband coverage and less homogeneous spectroscopic follow-up. Early southern efforts, including 2QZ, the Hamburg/ESO Survey for bright QSOs (L. Wisotzki et al. 2000), and the 6dF Galaxy Survey (D. H. Jones et al. 2009), provided valuable quasar identifications but did not achieve the depth or uniformity seen in the north. Recently, efforts have been made in finding the brightest quasars in the Southern Hemisphere (e.g., G. Calderone et al. 2019; K. Boutsia et al. 2020; G. Cupani et al. 2022; C. A. Onken et al. 2022, 2023), and systematic selections of quasar candidates in the Dark Energy Survey (Q. Yang & Y. Shen 2023) and the KMTNet Synoptic Survey of Southern Sky (Y. Kim et al. 2024).

In this paper, we extend the machine learning purification of GDR3 QSO candidates to the southern equatorial hemisphere, which is only partly covered by PS1 and CatNorth. To do so we utilize optical to mid-infrared data from external data sets including the fourth data release of the SkyMapper Southern

Survey (SMSS DR4; C. A. Onken et al. 2024), the second data release of the NOIRLab Source Catalog (NSC DR2; D. L. Nidever et al. 2021), the Visible and Infrared Survey Telescope for Astronomy (VISTA; J. Emerson et al. 2006) surveys, and CatWISE2020 in addition to Gaia DR3. We publish the results as the CatSouth quasar candidate catalog, and present a unified all-sky quasar candidate catalog (CatGlobe) by combining CatNorth and CatSouth.

This paper is organized as follows. In Section 2, we introduce the data used in this work. Section 3 describes the machine learning–based selection procedure of reliable quasar candidates. Section 4 describes the photometric redshift estimation and Gaia spectral redshift determination of the selected sample. Section 5 presents the contents and characteristics of the CatSouth and the all-sky CatGlobe quasar candidate catalogs. We summarize the paper in Section 6. Throughout this paper we adopt a flat Λ CDM cosmology with $\Omega_\Lambda = 0.7$, $\Omega_M = 0.3$, and $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$. The z -band magnitude does not appear alone and will not be confused with the redshift symbol z .

2. Data

The input data of this work is the Gaia DR3 quasar candidate catalog (the `qso_candidates` table) from Gaia Collaboration et al. (2023a). We combine optical and infrared photometric data from Gaia DR3, SkyMapper DR4, CatWISE2020/AllWISE, and astrometric data from Gaia DR3 to improve both purity and redshift estimation of the GDR3 QSO candidate catalog. We also retrieve samples of spectroscopically identified extragalactic objects from SDSS, and stellar samples from Gaia and additional catalogs to build well-defined training/validation sets.

When querying the photometric data, all magnitudes are presented in their original system, i.e., Gaia and SkyMapper in the AB system, and WISE, 2MASS, and VISTA in the Vega system. During the machine learning classification (Section 3) and redshift estimation (Section 4), we adopt all magnitudes in the AB system.

2.1. Astrometric and Photometric Data

2.1.1. Gaia DR3 Astrometric and Astrophysical Data

Gaia DR3 (Gaia Collaboration et al. 2023c) contains celestial positions, proper motions, parallaxes, and broadband photometry in the G , G_{BP} (330–680 nm), and G_{RP} (630–1050 nm) passbands for 1.8 billion sources at $G < 21$ that have been present in the Early Third Data Release (Gaia EDR3; Gaia Collaboration et al. 2021). Furthermore, the Gaia DR3 catalog incorporates about 1 million mean spectra from the radial velocity spectrometer, about 220 million low-resolution blue and red prism photometer BP/RP mean spectra, variability results of 10 million sources across 24 variability types, astrophysical parameters for about 470 million source, and source class probabilities for 1500 million sources, including stars, galaxies, and quasars.

2.1.2. CatWISE2020 and AllWISE Catalogs

The Wide-field Infrared Survey Explorer (WISE; E. L. Wright et al. 2010) is a NASA Medium Class Explorer mission that conducted an imaging survey of the entire sky in the 3.4, 4.6, 12, and 22 μm mid-infrared bands (W1, W2, W3, and W4). The AllWISE source catalog (E. L. Wright et al. 2019)

¹² The original Gaia DR3 quasar candidate catalog is available at the Gaia archive <https://gea.esac.esa.int/archive> with table name `gaiadr3.qso_candidates`.

was built by combining data from the WISE cryogenic and NEOWISE (A. Mainzer et al. 2011) postcryogenic survey phases, providing positions, proper motions, four-band fluxes and flux variability statistics for over 747 million objects.

The CatWISE2020 catalog (F. Marocco et al. 2021, 2020) consists of nearly 1.9 billion sources over the entire sky selected from the WISE cryogenic and NEOWISE postcryogenic survey data at W1 and W2 bands collected from 2010 January 7 to 2018 December 13. CatWISE2020 has 6 times as many exposures spanning over 16 times as large a time baseline as the AllWISE catalog. The 5σ limits for the CatWISE2020 Catalog in the Vega system are $W1 = 17.43$ mag and $W2 = 16.47$ mag (F. Marocco et al. 2021). The 5σ limiting magnitudes in the Vega system for the AllWISE catalog are 16.9, 16.0, 11.5, and 8.0 mag in W1, W2, W3 and W4, respectively.¹³

Because CatWISE2020 has deeper W1 and W2 data than AllWISE, we retrieve the point-spread function (PSF) fitting magnitudes of W1 and W2 from CatWISE2020 for Gaia DR3 objects using prematched tables from NOIRLab datalab¹⁴ with a matching radius of $1''.5$. Through an outer join between this query and the AllWISE catalog, we obtain the W3 PSF magnitude from AllWISE as auxiliary data.

We also set some constraints on the CatWISE2020 data. All sources should be (i) not too bright to avoid possible saturation ($w1mpro_pm > 7$ & $w2mpro_pm > 7$) and (ii) significantly detected in W1 and W2 bands ($w1snr_pm > 5$ & $w2snr_pm > 5$).

2.1.3. SkyMapper Southern Survey DR4

The 1.3 m SkyMapper telescope at Siding Spring Observatory, Australia, has been conducting the SMSS (S. C. Keller et al. 2007; C. Wolf et al. 2018; C. A. Onken et al. 2019) since 2014. The SkyMapper telescope has a 5.7 deg^2 field of view, and the SMSS includes six optical filters: u , v , g , r , i , and z (M. Bessell et al. 2011). The fourth data release (DR4) of SMSS (C. A. Onken et al. 2024) covers a sky area of $26,000 \text{ deg}^2$ from the South Celestial Pole to $\delta = +16^\circ$, with some fields of partial coverage reaching as far north as $\delta \sim 28^\circ$. The 10σ depth in the g band of SMSS DR4 is 20.5 mag (AB magnitude).

We obtain SMSS DR4 PSF photometry for the Gaia DR3 sources using prematched tables at NOIRLab datalab with a matching radius of $1''.5$. To select sources with good photometry, we require the i -band PSF magnitude $i > 10$ (not saturated), and $i_{\text{err}} < 0.2171$ (equivalent to signal-to-noise ratio $(S/N) > 5$).

2.1.4. VISTA Surveys

VISTA (J. Emerson et al. 2006) is a 4.1 m specialized wide-field survey telescope for the Southern Hemisphere. VISTA is equipped with a near-infrared camera VIRCAM (G. B. Dalton et al. 2026) with a 1.65 diameter field of view, five available broadband filters at Z , Y , J , H , K_s , and three narrow band filters at 0.98 , 0.99 , and $1.18 \mu\text{m}$. VISTA has been conducting six large public surveys, covering different sky areas to different depths. These surveys include the VISTA Hemisphere Survey (VHS; R. G. McMahon et al. 2013), the VISTA Kilo-Degree

Infrared Galaxy Survey (VIKING; A. Edge et al. 2013), the VISTA Magellanic Survey (VMC; M. R. Cioni et al. 2011; M. R. L. Cioni et al. 2011), the VISTA Variables in the Via Lactea survey (VVV; D. Minniti et al. 2010), the VISTA Deep Extragalactic Observations Survey (M. J. Jarvis et al. 2013a, 2013b), and UltraVISTA (H. J. McCracken et al. 2012).

The VHS is imaging the entire southern hemisphere of the sky, except the areas covered by the VIKING, VMC, and VVV surveys. Therefore, combining data from VHS and other VISTA surveys will improve the sky coverage of near-infrared photometric data in the Southern Hemisphere. We perform outer joins between our samples and the VHS, VIKING, VMC, and VVV surveys, each from a specific data release and source: VHS DR5 (ESO version 3) from the NOIRLab Data Lab;¹⁵ VVV DR4.2 from Vizier¹⁶ through the Vizier TAP service;¹⁷ VIKING DR4 via the VSA TAP service;¹⁸ and VMC DR6 (ESO version 5) through the ESO TAP service.¹⁹ A matching radius of $1''.5$ is used when joining our samples with the VHS, VIKING, and VMC catalogs, and a matching radius of $1''.0$ is used when joining our samples with the VVV catalog to avoid mismatches in the crowded Galactic plane.

We use the default point-source aperture photometry (APERMAG3) from VISTA, which has been aperture corrected and is suitable for point sources even in crowded fields (C. González-Fernández et al. 2018). To increase the query efficiency and ensure the data quality of the extremely large VIKING, VMC, and VVV catalogs, we set the following constraints when querying the databases:

```
(YAPERMAG3 BETWEEN 11 AND 22)
AND (JAPERMAG3 BETWEEN 11 AND 22)
AND (KSAPERMAG3 BETWEEN 11 AND 22)
AND MERGEDCLASS NOT IN (0, -9)
AND JAPERMAG3ERR < 0.2171
AND YPPERRBITS < 256
AND JPPERRBITS < 256
AND KSPERRBITS < 256.
```

2.1.5. NOIRLab Source Catalog DR2

The second data release (DR2) of the NSC (D. L. Nidever et al. 2018, 2021) is a catalog of over 3.9 billion sources from public imaging data in NOIRLab's Astro Data Archive.²⁰ Most of the images used in NSC DR2 are taken with CTIO-4 m Blanco + DECam (340,952 exposures). In addition, there are 41,561 exposures from KPNO 4 m Mayall + Mosaic3 (the majority from the Mayall z -band Legacy Survey, MzLS; A. Dey et al. 2016) and 29,603 exposures from the Steward Observatory Bok-2.3 m + 90Prime (from the Beijing–Arizona Sky Survey; H. Zou et al. 2017, 2018, 2019). A large fraction of the images are data obtained by the Dark Energy Survey (DES; B. P. Abbott et al. 2017) and the Legacy Survey imaging projects (A. Dey et al. 2019). NSC DR2 includes

¹³ https://wise2.ipac.caltech.edu/docs/release/allwise/expsup/sec2_3a.html

¹⁴ <https://datalab.noirlab.edu/>

¹⁵ <https://datalab.noirlab.edu/vhsdr5.php>

¹⁶ <https://cdsarc.cds.unistra.fr/viz-bin/w/VizieR?-source=II/376>

¹⁷ <https://tapvizier.u-strasbg.fr/adql/>

¹⁸ <http://tap.roe.ac.uk/vsa>

¹⁹ https://archive.eso.org/tap_cat

²⁰ <https://astroarchive.noirlab.edu/>

Table 1Relative Extinction Coefficients of Passbands Used in This Work ($R_V = 3.1$)

Band	λ_p (Å)	A_{λ_p}/A_V	R_{λ_p}
<i>g</i>	5075.19	1.1100	3.4411
<i>r</i>	6138.44	0.8535	2.6459
<i>i</i>	7767.98	0.5851	1.8139
<i>z</i>	9145.99	0.4382	1.3585
<i>G</i>	6217.59	0.8373	2.5957
<i>G_{BP}</i>	5109.71	1.1005	3.4116
<i>G_{RP}</i>	7769.02	0.5850	1.8135
<i>Y</i>	10210.71	0.3565	1.1051
<i>J</i>	12524.83	0.2336	0.7240
<i>H</i>	16432.45	0.1331	0.4127
<i>K_s</i>	21521.52	0.0762	0.2361
W1	33682.21	0.0301	0.0934
W2	46179.06	0.0157	0.0486
W3	120718.09	0.0021	0.0067

photometry in seven bands, namely *u*, *g*, *r*, *i*, *z*, *Y*, and *VR*, with different sky coverages and depths.

We utilize data of NSC DR2 *griz* bands to impute missing values of SMSS *griz* bands. We do not apply any transformation to the NSC DR2 *griz* photometry when imputing the missing values because the magnitude offsets between the two data sets vary across different sky regions (see Figure 19 of C. A. Onken et al. 2024) and are sensitive to extinction. Nevertheless, such missing data imputation is still very helpful for training and applying machine learning models, because for a list of sources, valid values from a similar NSC band are more informative and discriminative than a single mean or median value of the whole sample.

2.1.6. 2MASS

The Two Micron All Sky Survey (2MASS; M. F. Skrutskie et al. 2003, 2006) has uniformly scanned the entire sky in *J* (1.25 μm), *H* (1.65 μm), and *K_s* (2.17 μm) bands using two 1.3 m telescopes, one at Mount Hopkins, Arizona, and one at CTIO, Chile. Because the AllWISE Source Catalog that we use (Section 2.1.2) includes the closest entries from the 2MASS Point Source Catalog (PSC) within 3'' from the AllWISE positions,²¹ we adopt the 2MASS photometry from AllWISE directly. We use the 2MASS *JHK_s* magnitudes to impute missing values of VISTA photometry by converting 2MASS magnitudes to VISTA ones. To do so, the offsets between the median values of 2MASS bands and those of the VISTA bands are calculated and subtracted from 2MASS magnitudes for each of the training/validation samples (stars, galaxies, quasars), and the test sample (GDR3 QSO candidates).

2.1.7. Relative Extinction Coefficients

For each photometric band we use in this work, we compute its relative extinction coefficient R_{λ_p} , defined as

$$R_{\lambda_p} = \frac{A_{\lambda_p}}{A_V} \times R_V, \quad (1)$$

where A_{λ_p} and A_V are the extinctions in the band with pivot wavelength λ_p and *V* band, and $R_V = A_V/E(B - V)$. The pivot

wavelength λ_p (J. Koornneef et al. 1986) is calculated as

$$\lambda_p = \sqrt{\frac{\int_{\lambda} T(\lambda) \lambda d\lambda}{\int_{\lambda} T(\lambda) d\lambda/\lambda}}, \quad (2)$$

where $T(\lambda)$ is the throughput (transmission curve) of the filter as a function of the wavelength.

Using the definition of passbands from E. L. Wright et al. (2010), M. Bessell et al. (2011), C. González-Fernández et al. (2018), and M. Riello et al. (2021), and the optical to mid-IR extinction law from S. Wang & X. Chen (2019) assuming $R_V = 3.1$, we calculate the extinction coefficients as listed in Table 1. The coefficients are used for Galactic extinction correction during the photometric redshift estimation procedure.

2.2. Stellar Sample

A robust stellar sample is essential for constructing the training set of our machine learning classifier and for minimizing contaminant misclassifications. In CatSouth we adopt a strategy similar to that in Y. Fu et al. (2024): we select stars from Gaia DR3 within the SMSS DR4 footprint and complement this primary sample with additional very low-mass stars (VLMS), white dwarfs (WD), and carbon stars to form the combined master stellar sample.

2.2.1. O-to-M Type Stars from Gaia DR3

We first extract a representative sample of O-to-M type stars from Gaia DR3 following the methods in Y. Fu et al. (2024). In particular, we purify the original Gaia DR3 OBA gold sample by excluding sources with tangential velocity (v_{tan}) higher than 180 km s⁻¹ as suggested by Gaia Collaboration et al. (2023b), and increase the completeness of the FGKM sample using a less strict selection than the one adopted by Gaia Collaboration et al. (2023b). Only sources with declinations satisfying $\delta < 16^\circ$ are selected to effectively match the SMSS DR4 footprint. The detailed ADQL queries used for the Gaia selections of O-to-M type stars are provided in the Appendix. Finally, these stars are crossmatched with other catalogs in Section 2.1 with the corresponding photometric quality constraints.

2.2.2. Additional VLMS, White Dwarfs, and Carbon Stars

To account for atypical/underrepresentative stars that may contaminate the quasar selection, we supplement the O-to-M type sample with additional VLMS, WDs, and carbon stars compiled from the literature. Major sources for these additional samples include J. Li et al. (2021), A. A. West et al. (2011), and A. Alksnis et al. (2001), and we refer to Table 1 in Y. Fu et al. (2024) for a complete list of the catalogs. These extra stars are crossmatched with catalogs in Section 2.1, and are merged with the O-to-M type stars to form the master stellar sample. This combined stellar sample contains more than 1.84 million sources satisfying the photometric quality constraints in Section 2.1, providing a diverse and comprehensive training set for our classification model.

²¹ https://wise2.ipac.caltech.edu/docs/release/allwise/expsup/sec2_2.html

2.3. Quasar Sample

We combine known quasars from the Million Quasars catalog (Milliquas v8; E. W. Flesch 2023), and highly reliable quasar candidates from the CatNorth quasar candidate catalog (Y. Fu et al. 2024) to build the quasar sample for training the machine classification model.

Milliquas v8 is a compilation of quasars and quasar candidates from the literature up to 2023 June 30, which includes 907,144 type 1 QSOs and AGNs, 66,026 high-confidence ($p_{\text{QSO}} = 99\%$) photometric quasar candidates, 2814 BL Lac objects, and 45,816 type 2 objects. We select the subsample of Milliquas by requiring that the sources are (i) located at $\delta < 16^\circ$ and (ii) spectroscopically identified type 1 quasars with valid redshifts (labeled as “Q” in the “TYPE” column of Milliquas, and $z > 0$).

Among the 290,294 selected quasars, about 75% are identified by SDSS (e.g., R. Ahumada et al. 2020; B. W. Lyke et al. 2020; A. Almeida et al. 2023); other major sources come from the DESI EDR (DESI Collaboration et al. 2024), the 2dF QSO Redshift Survey (2QZ; S. M. Croom et al. 2004), the LAMOST quasar survey (Y. L. Ai et al. 2016; X. Y. Dong et al. 2018; S. Yao et al. 2019; J.-J. Jin et al. 2023), and the 2dF-SDSS LRG and QSO survey (2SLAQ; S. M. Croom et al. 2009).

CatNorth is a quasar candidate catalog built upon the GDR3 QSO candidate catalog using photometric data from Gaia, PS1, CatWISE, and low-resolution spectral information from Gaia. To complement the Milliquas sample, we select a highly reliable subsample of CatNorth in the SMSS footprint using the following criteria:

1. $p_{\text{QSO_mean}} > 0.95$,
2. $\delta < 16^\circ$,
3. $|z_{\text{ph}} - z_{\text{Gaia}}|/(1 + z_{\text{ph}}) < 0.02$,

where $p_{\text{QSO_mean}}$ is the machine learning predicted probability of a source being a quasar, z_{ph} is the photometric redshift given by CatNorth, and z_{Gaia} is the redshift in the Gaia DR3 quasar candidate catalog (`redshift_qsoc`). The selection yields 323,626 high-confidence quasar candidates.

After being crossmatched with Gaia DR3, SMSS DR4, and CatWISE2020 and filtered with the corresponding quality constraints, 105,413 sources remain in the Milliquas v8 subsample, and 276,814 sources remain in the CatNorth subsample. Combining the two subsets gives the final quasar sample containing 282,322 unique sources.

2.4. Galaxy Sample

The galaxy sample is built upon the combination of the spectroscopic galaxy catalogs of the Seventeenth Data Release of the SDSS (DR17; Abdurro’uf et al. 2022), the 2dF Galaxy Redshift Survey (2dFGRS; M. Colless et al. 2001), the 6dF Galaxy Survey (6DFGS; D. H. Jones et al. 2009), and the 2MASS Redshift Survey (2MRS; J. P. Huchra et al. 2012). In particular, the SDSS DR17 galaxy sample is selected from the SpecObj table²² using the following criteria:

1. The objects are spectroscopically classified as galaxies without broad emission lines ($\sigma_{\text{line}} > 200 \text{ km s}^{-1}$)

detected at the 5σ level: `CLASS == "GALAXY" AND SUBCLASSNOT LIKE "BROADLINE."`

2. The spectra are primary detections with good observational conditions and high S/N, and no issues are found in fitting the redshifts: `SPECPRIMARY == 1 AND PLATEQUALITY == "good" AND SN_MEDIAN_ALL > 5 AND ZWARNING == 0`.

For the other three catalogs (2dFGRS, 6DFGS, and 2MRS), we select galaxies that are at $\delta < 16^\circ$ with valid redshifts ($z > 0$) and not classified as quasars/broad-line AGNs by any of the following catalogs:

1. Milliquas v8.
2. A uniformly selected, all-sky, optical AGN catalog by I. Zaw et al. (2019).
3. An extensive list of broad-line AGNs in the 6dF Galaxy Survey (W. J. Hon et al. 2025).

We merge all galaxies satisfying the photometric quality constraints in Section 2.1 into one sample, which contains 382,303 unique sources using an inner match radius of $1''.5$. Finally, the labeled stars, quasars, and galaxies are combined as a unified training/validation set with more than 2.5 million sources.

3. Machine Learning Selection of Quasar Candidates

3.1. XGBoost Classification of Stars, Galaxies, and Quasars

We use XGBoost (T. Chen & C. Guestrin 2016), a gradient-boosting decision tree algorithm to train the machine learning classifier using the training/validation set, and reclassify the input Gaia DR3 quasar candidates as quasars, stars, and galaxies. Gradient-boosting tree algorithms including XGBoost have shown exceptional performance in astronomical studies, especially those on quasar candidate selections and photometric redshifts (e.g., X. Jin et al. 2019; Y. Fu et al. 2021, 2022, 2024; C. Li et al. 2022; A. C. N. Hughes et al. 2022; W.-B. Kao et al. 2024; G. Ye et al. 2024).

We follow the feature selection strategy of Y. Fu et al. (2024, Section 3), which combines two types of features: (1) broadband colors that capture differences in spectral energy distributions (SEDs) among quasars, stars, and galaxies, and (2) morphological indicators that help distinguish point-like sources (such as stars and quasars) from extended sources (such as galaxies). Similarly, we construct a multidimensional feature set for training the classification model in this work using multiwavelength data described in Section 2. The color and magnitude features include W1 (the W1 magnitude), and a set of color indices, $g - r$, $r - i$, $i - z$, $i - J$, $J - K_s$, $i - W1$, $z - W1$, $J - W1$, $W1 - W2$, $W2 - W3$, $G_{\text{BP}} - G_{\text{RP}}$, $G_{\text{BP}} - G$, and $G - G_{\text{RP}}$, all in AB magnitudes. We do not apply extinction corrections to the magnitudes and colors because such corrections depend on source types and distances, which are not known for the sources in the application (test) set. We also compute two morphological indicators, $\log(\chi^2_{\text{PSF}})$ from SMSS DR4, and $\log(1 + C^*)$ from Gaia DR3, which are discussed later in this section. Together, 16 features are used to train the XGBoost classifier.

The magnitude and colors we choose capture key differences in the SEDs of quasars versus stars and galaxies. In particular, the mid-infrared color $W1 - W2$ has been proven effective in selecting AGNs due to their power-law SEDs and

²² <https://data.sdss.org/sas/dr17/sdss/spectro/redux/specObj-dr17.fits>

hot dust emission (e.g., D. Stern et al. 2012; X.-B. Wu et al. 2012; R. J. Assef et al. 2018). The near-infrared color $J - K_s$ is also powerful in separating quasars at various redshifts from stars (e.g., N. Maddox et al. 2008; X.-B. Wu & Z. Jia 2010; J.-T. Schindler et al. 2017). The combination of the optical color indices ($g - r$, $r - i$, $i - z$) can help reduce the overlap between quasars and the stellar loci on two-dimensional color-color diagrams (Y. Fu et al. 2021).

To complement the color and magnitude information, we include two morphological features sensitive to source extent. The first parameter is χ_{PSF}^2 , defined as the maximum chi-squared value from the PSF photometry across the available SMSS DR4 bands (see Section 6.7.4 of C. A. Onken et al. 2024). Point sources typically yield χ_{PSF}^2 values between 1 and 3, while extended objects have substantially larger values, e.g., up to a few thousand for galaxies. The second morphological feature is the corrected BP/RP flux excess factor, C^* , from Gaia DR3 (M. Riello et al. 2021).²³ The original BP/RP flux excess factor, C , is the ratio of the sum of the integrated BP and RP fluxes to the flux in the G band: $C = (I_{\text{BP}} + I_{\text{RP}})/I_G$. Because the detection windows (apertures) of BP and RP bands are wider than that of the G band, extended sources tend to have larger flux excess factors than the point sources do (see, e.g., C. Liu et al. 2020). The corrected BP/RP flux excess factor C^* removes the color dependence present in the original C , enabling more robust characterization of source extent (M. Riello et al. 2021). We use the logarithmically transformed morphological features, $\log(\chi_{\text{PSF}}^2)$ and $\log(1 + C^*)$, to compress the dynamic ranges of the original values and mitigate the influence of extreme outliers.

Figure 1 shows the two-dimensional representations of some selected combinations of features of the training/validation set. Stars form narrow, concentrated loci on color-color diagrams that blend with quasars in the optical. The stellar loci become more distinctly separated from quasars in the infrared bands. Quasars from the training set produce unimodal probability distributions with smooth boundaries on the color-color and morphology-color diagrams, indicating no contamination from stars. However, the outermost contours of galaxies from the training set show spikes that overlap with stellar loci, indicating minor contributions from stars. We do not purify the galaxy sample further because our primary goal is to optimize quasar selection; since both galaxies and stars are treated as contaminants, the small level of stellar contamination in the galaxy sample does not impact the performance of the classifier in identifying quasars.

We use a few metrics to evaluate the model performance. For binary classification problems, with true positive denoted as TP, true negative as TN, false positive as FP, and false negative as FN, the metrics are defined as

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (5)$$

In the case of a multiclass problem, the classification task is treated as a collection of binary classification problems, one for each class. The metrics above can be calculated for each binary classification problem (each class). The metrics of the multiclass problem are the average metrics of all classes.

Hyperparameter optimization is performed with Optuna (T. Akiba et al. 2019) via five-fold cross validation, minimizing the multiclass log loss as the objective function among 500 trials. We refer to Y. Fu et al. (2024) for a detailed hyperparameter tuning procedure. After obtaining the optimal hyperparameters, the whole input data is split into a training set and a validation set according to a 4:1 ratio, and the optimal classifier is trained on the entire training set.

The normalized confusion matrix for the three-class problem, calculated on the validation set and shown in Figure 2, demonstrates the outstanding performance of the trained XGBoost classifier. For example, 99.40% of galaxies, 99.92% of quasars, and 99.98% of stars are correctly identified, while the off-diagonal elements indicate very low misclassification rates. The actual fraction of galaxies misclassified as stars is expected to be even lower than 0.41% shown in the confusion matrix, because the training/validation sample of galaxies is contaminated by stars. The good performance is also reflected in the overall F_1 score of 0.9989, which confirms that both precision and recall are extremely high for all classes.

We apply the optimal XGBoost classifier to the input test sample of Gaia DR3 quasar candidates in this work and assign probability estimates (p_{QSO} , p_{star} , p_{galaxy}) for each source. By default, a source is labeled as the class that receives the highest probability.

3.2. Additional Filtering to Remove Contaminants

To further reduce stellar contamination in our quasar candidate sample, we apply a probabilistic proper-motion cut based on the likelihood that a source has zero proper motion. Following the method described in Y. Fu et al. (2021, 2024), the probability density function of zero proper motion, f_{PM0} , is defined as

$$f_{\text{PM0}} = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x}{\sigma_x}\right)^2 - \frac{2\rho xy}{\sigma_x\sigma_y} + \left(\frac{y}{\sigma_y}\right)^2\right]\right\}, \quad (6)$$

where x is the proper motion in R.A. (pmra), y is the proper motion in decl. (pmdec), ρ is the correlation coefficient between x and y (pmra_pmdec_corr), and σ_x and σ_y are the proper-motion uncertainties. In practice, for a given level of uncertainty, sources with smaller proper motions yield higher values of f_{PM0} . To facilitate comparison across different samples, we work with the logarithm of this quantity, $\log(f_{\text{PM0}})$. As demonstrated in Y. Fu et al. (2024), a threshold of $\log(f_{\text{PM0}}) \geq -4$ effectively excludes over 99.9% of stellar contaminants while retaining more than 99.8% of the quasar sample.

To account for the higher source density and increased contamination in the regions around the Large and Small Magellanic Clouds (LMC and SMC), we define two circular regions centered on these objects. Specifically, we construct a 10° radius region centered at LMC ($\alpha = 80.8942^\circ$,

²³ <https://github.com/agabrown/gaiaedr3-flux-excess-correction>

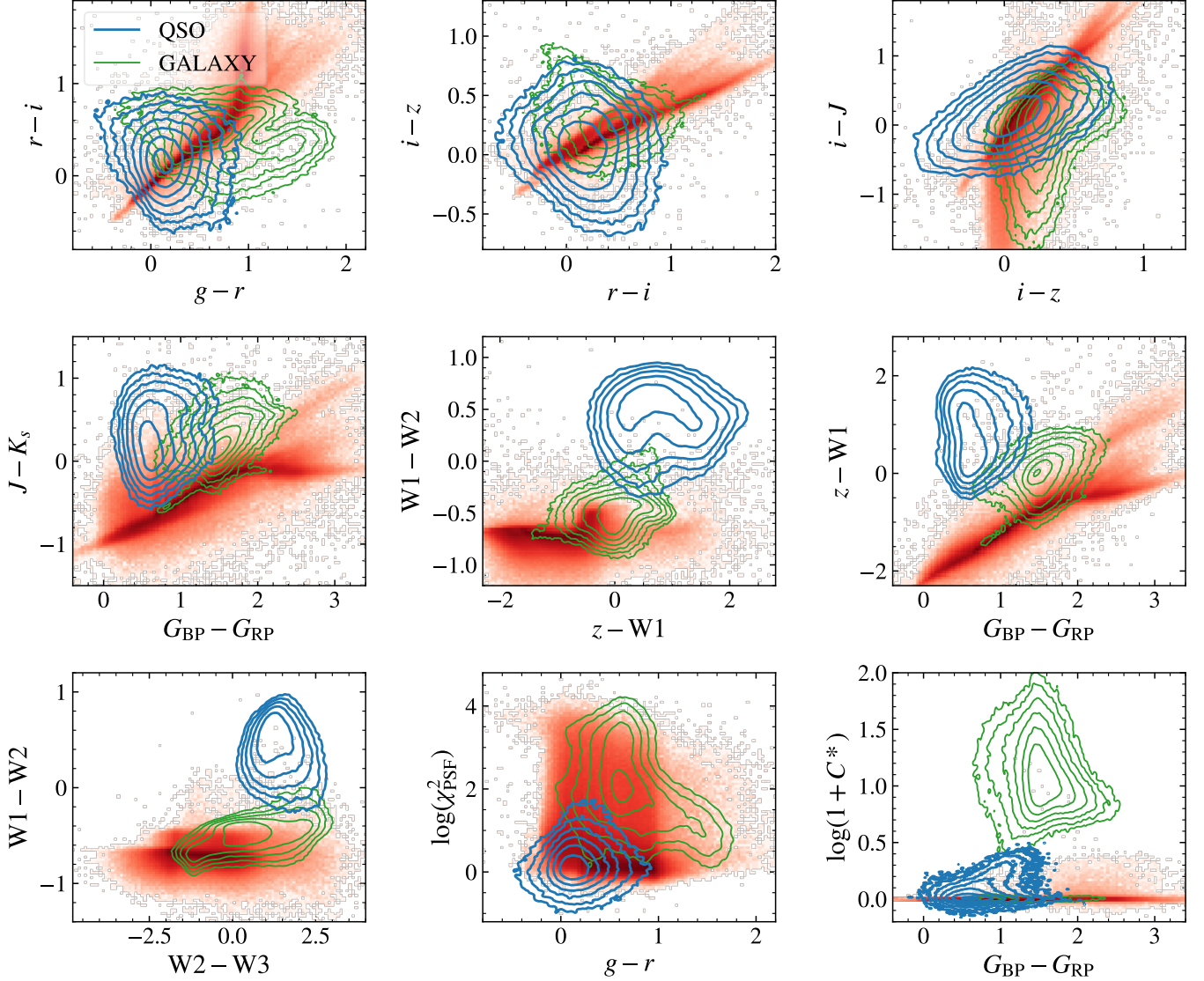


Figure 1. Two-dimensional feature representations (color-color and morphology-color diagrams) of sources in the training sample. Quasars are shown as blue contours, galaxies as green contours, and stars as red-shaded density plots. All magnitudes are in the AB system and not dereddened.

$\delta = -69.7561$) and a 5° radius region centered at SMC ($\alpha = 13.1583$, $\delta = -72.8003$) using multiorder coverage (MOC) maps generated with MOCpy (M. Baumann et al. 2024). We then determine which sources in our catalog fall within these regions by verifying whether their coordinates are contained in the corresponding MOCs.

Based on the spatial locations, we apply distinct selection criteria to select reliable quasar candidates. For sources outside the LMC/SMC regions, where contamination is lower, we require $\log f_{\text{PMO}} > -4$ and that the predicted class is “QSO” (i.e., $p_{\text{QSO}} > p_{\text{star}}$ and $p_{\text{QSO}} > p_{\text{galaxy}}$). In contrast, for sources within the LMC or SMC regions, which are known to exhibit higher stellar crowding and potential misclassification, we enforce a stricter threshold by requiring $\log f_{\text{PMO}} > -1$ and $p_{\text{QSO}} > 0.9$. This dual-threshold strategy ensures a more reliable selection of quasar candidates across different sky regions, maintaining high purity in crowded fields while preserving completeness in less contaminated areas. In total, 921,528 sources out of 1,174,509 input GDR3 QSO candidates are selected as reliable quasar candidates.

4. Redshift Estimation for Quasar Candidates

We estimate photometric redshifts for all quasar candidates selected from above and spectroscopic redshifts for quasar candidates with available Gaia BP/RP spectra. For both regression models, we adopt the root mean square error (RMSE), the normalized median absolute deviation of errors (σ_{NMAD}), and the catastrophic outlier fraction (f_c) as evaluation metrics for the redshift estimation in the training/validation sets. These metrics are defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (z_i - \hat{z}_i)^2}, \quad (7)$$

$$\sigma_{\text{NMAD}} = 1.48 \times \text{median} \left(\left| \frac{\Delta z - \text{median}(\Delta z)}{1 + z} \right| \right), \quad (8)$$

$$f_c = \frac{1}{n} \times \text{count} \left(\left| \frac{\Delta z}{1 + z} \right| > 0.15 \right), \quad (9)$$

where z is the true redshift, \hat{z} is the predicted redshift, $\Delta z = z - \hat{z}$, and n is the total number of sources. The RMSE

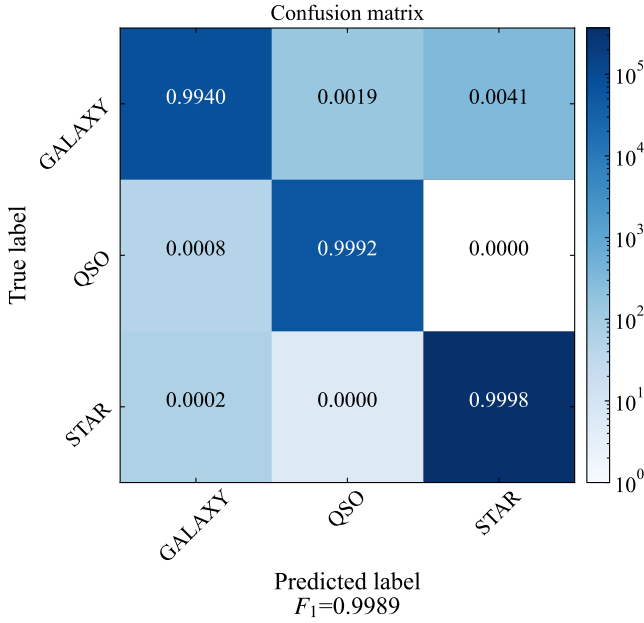


Figure 2. Normalized confusion matrix of the XGBoost classifier computed on the validation set. The matrix is color coded by the number of sources in each cell. Diagonal entries show the fraction of correctly classified objects (i.e., recall or completeness) for each class, while off-diagonal entries indicate the misclassification rates.

is widely used in regression analysis to quantify the difference between the true and predicted values. The σ_{NMAD} measures the statistical dispersion of the normalized errors $\Delta z' = \Delta z / (1 + z)$ (O. Ilbert et al. 2006; G. B. Brammer et al. 2008). The f_c represents the percentage of objects for which the redshift estimate deviates significantly from the true redshift.

4.1. BP/RP Spectroscopic Redshift Determination

In Y. Fu et al. (2024), we have trained a convolutional neural network (CNN) regression model (RegNet) for quasar redshift determination using Gaia BP/RP spectra of quasars with known redshifts, which achieved $\text{RMSE} = 0.1427$, $\sigma_{\text{NMAD}} = 0.0304$, and $f_c = 2.46\%$ on the validation set. Because Gaia BP/RP spectra do not rely on external photometric surveys, the spectroscopic redshifts of CatSouth sources with Gaia BP/RP spectra can be easily determined using the pretrained RegNet model.

The BP/RP spectra of CatSouth sources are retrieved via the `astroquery.gaia` module, then calibrated and resampled over the wavelength range [4000 Å, 10000 Å] in 20 Å intervals using the `GaiaXPy` package (D. Ruz-Mieres 2023). Each spectrum is represented by 300 data points and normalized to the [0, 1] range before inference with RegNet. The redshift estimates from RegNet are denoted as $z_{\text{xp_nn}}$.

4.2. Ensemble Photometric Redshift Estimation with XGBoost, TabNet, and FT-Transformer

Photometric redshift estimation with machine learning algorithms requires a training sample with good data quality in both feature columns and spectroscopic redshift. Starting from the 105,413 quasars from Milliquas that satisfy the photometric quality constraints in Section 2.3, we perform the

Table 2

Scores of all Photometric Redshift Regression Models (XGBoost, TabNet, FT-Transformer, and the Ensemble Model) on the Validation Set

Metric	XGBoost	TabNet	FT-Transformer	Ensemble
RMSE	0.2370	0.2329	0.2320	0.2256
σ_{NMAD}	0.0225	0.0212	0.0224	0.0187
f_c	7.68%	7.13%	6.91%	6.92%

following procedures to build a subsample with accurate and precise spectroscopic redshifts:

1. Because Milliquas only preserves three digits for the redshift (column “Z”), we replace the Milliquas redshift values (z_{MQ}) with the ones with higher precision from SDSS DR16Q (z_{sys} from Q. Wu & Y. Shen 2022), SDSS DR18 (A. Almeida et al. 2023), and DESI EDR (DESI Collaboration et al. 2024) when available. The spectroscopic redshift values from other origins are kept as they are. The newly adopted redshift is denoted as z_{cat} .
2. For DR16Q sources in the subset, we keep those with robust redshift estimates by requiring $z_{\text{sys}} > 0$, $z_{\text{sys_error}} \neq -1$, $z_{\text{sys_error}} \neq -2$, $z_{\text{sys_error}} / (1 + z_{\text{sys}}) < 0.002$, and $|z_{\text{sys}} - z_{\text{DR16Q}}| / (1 + z_{\text{sys}}) < 0.002$, where z_{DR16Q} is the final redshift from B. W. Lyke et al. (2020).
3. Milliquas has made corrections for a small fraction of quasars with wrong redshifts from SDSS or DESI. Therefore, we use $|z_{\text{MQ}} - z_{\text{cat}}| / (1 + z_{\text{cat}}) < 0.002$ to select quasars with good redshifts in both Milliquas and original catalogs.

The resulting spectroscopic Milliquas subsample contains 96,181 sources. To complement this spectroscopic sample, we select additional quasars from CatSouth that have CNN-derived redshifts in very close agreement with the Gaia redshifts by requiring $|z_{\text{xp_nn}} - z_{\text{Gaia}}| / (1 + z_{\text{Gaia}}) < 0.01$. For these additional sources, we compute the target redshift z_{cat} by averaging the CNN-derived $z_{\text{xp_nn}}$ and the original Gaia redshift (z_{Gaia}), i.e.,

$$z_{\text{cat}} = \frac{z_{\text{xp_nn}} + z_{\text{Gaia}}}{2}. \quad (10)$$

The combined training sample for the photometric redshift estimation contains 108,885 unique sources.

We choose a total of 18 features for training and deploying the machine learning models: $W1$, $g - r$, $r - i$, $i - z$, $i - J$, $J - K_s$, $i - W1$, $z - W1$, $J - W1$, $W1 - W2$, $W2 - W3$, $G_{\text{BP}} - G_{\text{RP}}$, $G_{\text{BP}} - G$, $G - G_{\text{RP}}$, $\log(\chi^2_{\text{PSF}})$, $\log(1 + C^*)$, $\log(1 + z_{\text{low}})$, and $\log(1 + z_{\text{up}})$. Here, z_{low} (`redshift_qsoc_lower`) and z_{up} (`redshift_qsoc_upper`) are the lower and upper confidence intervals of z_{Gaia} taken at 0.15866 and 0.84134 quantiles, respectively. All magnitudes are in the AB system, and corrected for Galactic dust extinction using the Planck Collaboration et al. (2016) dust map and the extinction coefficients in Table 1. Missing values are imputed with the median of the training sample.

After splitting the sample into training and validation sets (4:1 ratio), we train three independent regression models using different algorithms: XGBoost, TabNet (S. Ö. Arik & T. Pfister 2021), and FT-Transformer (Y. Gorishniy et al. 2021). Each model is trained with its optimal hyperparameters found by optuna. The final ensemble redshift z_{ph} is obtained by averaging the predictions of the three models.

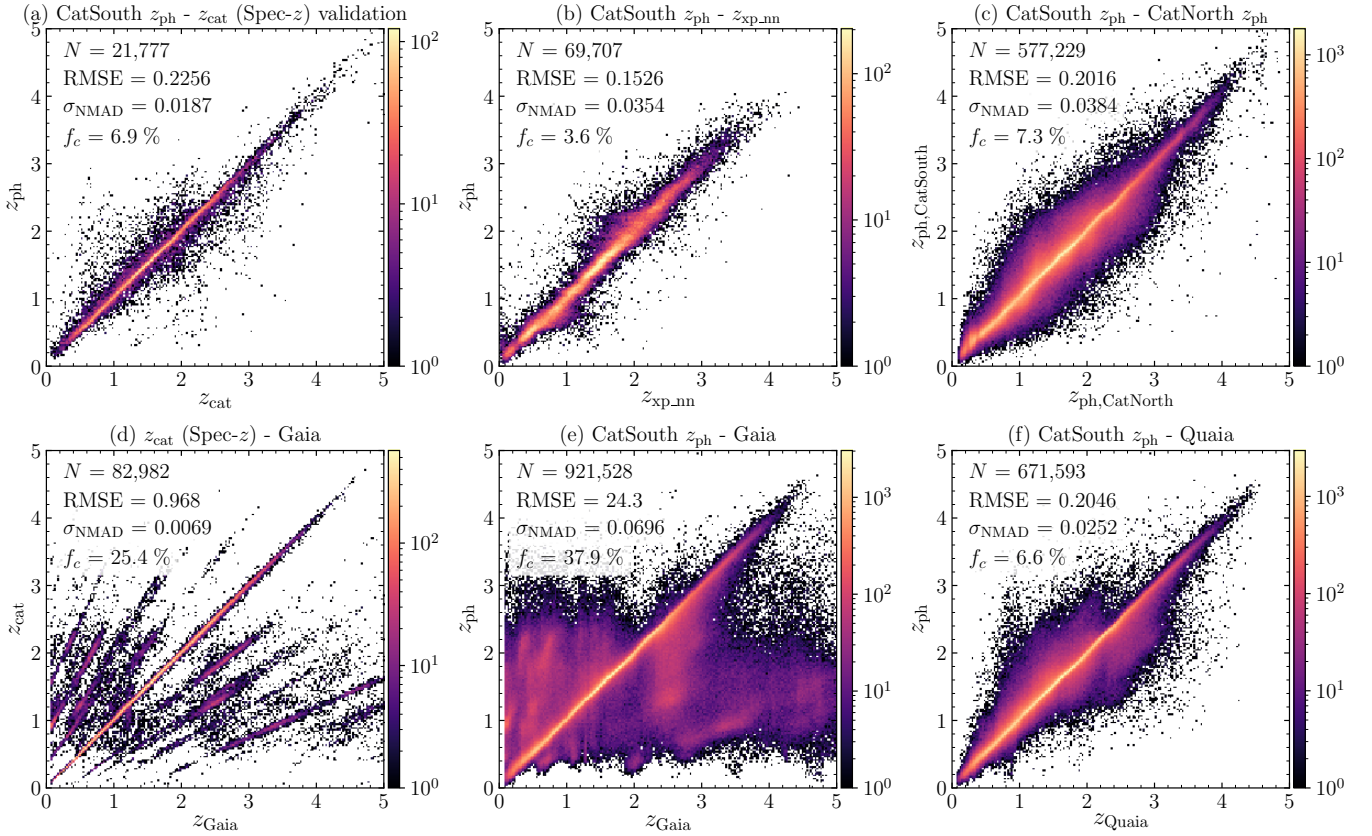


Figure 3. Top row: ensemble photometric redshift (z_{ph}) against spectral redshift (z_{cat}) of the validation set with 21,777 quasars (a), z_{ph} against RegNet redshift ($z_{\text{xp_nn}}$) (b), and z_{ph} of CatSouth vs. that in CatNorth for sources in common (c). Bottom row: comparisons between the spectral redshift (z_{cat}) in the training/validation sample and the Gaia redshift (d), CatSouth z_{ph} and Gaia (e), and CatSouth z_{ph} and Quaia (f). The plots are color coded with two-dimensional densities (number counts in the pixels), the values of which are indicated in the color bars.

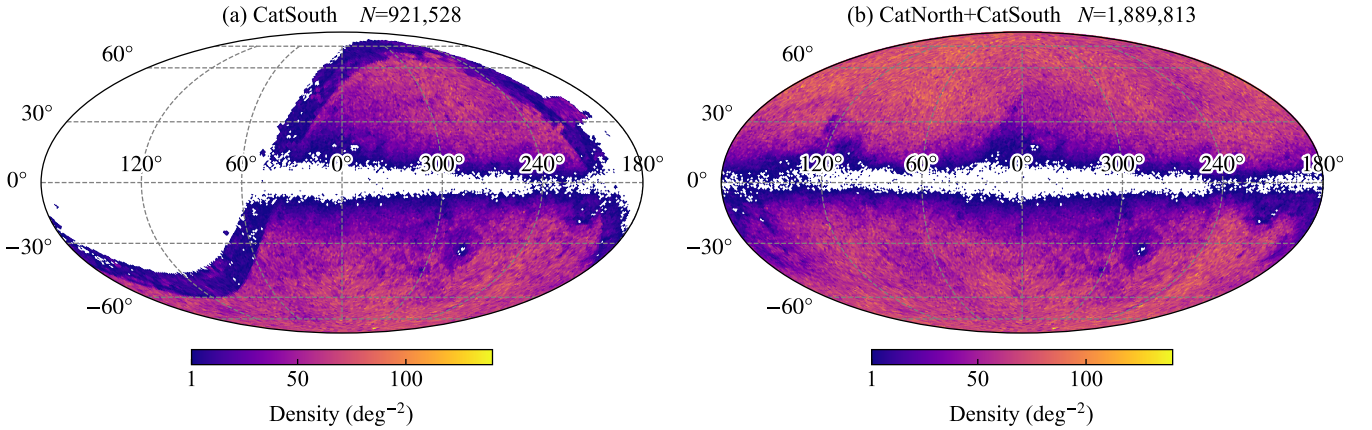


Figure 4. HEALPix (K. M. Górski et al. 2005) sky density maps of the CatSouth quasar candidate catalog (a) and the CatGlobe (CatNorth+CatSouth) quasar candidate catalog (b). The maps are plotted in Galactic coordinates, with parameter $N_{\text{side}} = 64$ and an area of 0.839 deg^2 per pixel.

The scores of the three regression models and the ensemble model on a validation set of 21,777 sources are listed in Table 2. Among the three base models, FT-Transformer achieves the lowest RMSE (0.2320) and f_c (6.91%), and TabNet achieves the lowest σ_{NMAD} (0.0212). Averaging the three base models produces an ensemble model with even lower RMSE (0.2256) and σ_{NMAD} (0.0187), and a moderately low f_c (6.92%). Because ensemble models not only reduce overfitting (variance) but also lower the estimation bias (see, e.g., O. Sagi & L. Rokach 2018), we expect the ensemble model to be more robust than the individual base models.

Figure 3 shows the performance of the redshift regression models on the validation sets, and the comparisons between CatSouth redshift estimates and those from the GDR3 QSO candidate catalog, the CatNorth catalog, and the Quaia catalog.

5. Results: The CatSouth and CatGlobe Quasar Candidate Catalogs

We compile the CatSouth quasar candidate catalog by including the Gaia DR3 photometry and astrometry, SMSS DR4 photometry, near-infrared photometry from VISTA

Table 3
Format of the CatSouth Quasar Candidate Catalog

Column	Name	Type	Unit	Description
1	source_id	long	...	Gaia DR3 unique source identifier
2	ra	double	deg	Gaia DR3 R.A. (ICRS) at Ep = 2016.0
3	dec	double	deg	Gaia DR3 decl. (ICRS) at Ep = 2016.0
4	l	double	deg	Galactic longitude
5	b	double	deg	Galactic latitude
6	parallax	double	mas	Parallax
7	parallax_error	double	mas	Standard error of parallax
8	pmra	float	mas yr ⁻¹	Proper motion in R.A. direction
9	pmra_error	float	mas yr ⁻¹	Standard error of pmra
10	pmdec	float	mas yr ⁻¹	Proper motion in decl. direction
11	pmdec_error	float	mas yr ⁻¹	Standard error of pmdec
12	pmra_pmdec_corr	float	...	Correlation between pmra and pmdec
13	phot_bp_mean_mag	float	mag	Integrated BP mean magnitude
14	phot_g_mean_mag	float	mag	G-band mean magnitude
15	phot_rp_mean_mag	float	mag	Integrated RP mean magnitude
16	bp_rp	float	mag	BP–RP color
17	phot_bp_rp_excess_factor	float	...	BP/RP excess factor
18	smss_id	long	...	SMSS DR4 unique object id
19	ra_smss	double	deg	SMSS DR4 R.A. (ICRS)
20	dec_smss	double	deg	SMSS DR4 decl. (ICRS)
21	chi2_psf	float	...	Maximum chi-squared from photometry table
22	g_psf	float	mag	Weighted mean SMSS <i>g</i> -band PSF magnitude
23	e_g_psf	float	mag	Error in <i>g</i> _psf
24	r_psf	float	mag	Weighted mean SMSS <i>r</i> -band PSF magnitude
25	e_r_psf	float	mag	Error in <i>r</i> _psf
26	i_psf	float	mag	Weighted mean SMSS <i>i</i> -band PSF magnitude
27	e_i_psf	float	mag	Error in <i>i</i> _psf
28	z_psf	float	mag	Weighted mean SMSS <i>z</i> -band PSF magnitude
29	e_z_psf	float	mag	Error in <i>z</i> _psf
30	yapermag3	float	mag	Default point source <i>Y</i> aperture corrected Vega mag (2'' diameter)
31	yapermag3err	float	mag	Error in yapermag3
32	japermag3	float	mag	Default point source <i>J</i> aperture corrected Vega mag (2'' diameter)
33	japermag3err	float	mag	Error in japermag3
34	hapermag3	float	mag	Default point source <i>H</i> aperture corrected Vega mag (2'' diameter)
35	hapermag3err	float	mag	Error in hapermag3
36	ksapermag3	float	mag	Default point source <i>K_s</i> aperture corrected Vega mag (2'' diameter)
37	ksapermag3err	float	mag	Error in ksapermag3
38	catwise_id	string	...	CatWISE2020 source id
39	ra_cat	double	deg	CatWISE2020 R.A. (ICRS)
40	dec_cat	double	deg	CatWISE2020 decl. (ICRS)
41	pmra_cat	float	arcsec yr ⁻¹	CatWISE2020 proper motion in R.A. direction
42	pmdec_cat	float	arcsec yr ⁻¹	CatWISE2020 proper motion in decl. direction
43	e_pmra_cat	float	arcsec yr ⁻¹	Uncertainty in pmra_cat
44	e_pmdec_cat	float	arcsec yr ⁻¹	Uncertainty in pmdec_cat
45	snrw1pm	float	...	Flux S/N in band-1 (W1)
46	snrw2pm	float	...	Flux S/N in band-2 (W2)
47	snrw3	float	...	Flux S/N in band-3 (W3) from AllWISE
48	w1mpropm	float	mag	WPRO magnitude in band-1 (Vega)
49	e_w1mpropm	float	mag	Uncertainty in w1mpropm
50	w2mpropm	float	mag	WPRO magnitude in band-2 (Vega)
51	e_w2mpropm	float	mag	Uncertainty in w2mpropm
52	w3mpro	float	mag	WPRO magnitude in band-3 from AllWISE (Vega)
53	e_w3mpro	float	mag	Uncertainty in w3mpro from AllWISE
54	phot_bp_rp_excess_factor_c	float	...	Corrected phot_bp_rp_excess_factor
55	log_fpm0	float	...	Logarithm probability density of zero proper motion ($\log f_{\text{pm0}}$)
56	in_lmc	boolean	...	Set to <code>True</code> if within 10° from the center of LMC; <code>False</code> otherwise
57	in_smc	boolean	...	Set to <code>True</code> if within 5° from the center of SMC; <code>False</code> otherwise
58	z_gaia	float	...	Redshift estimate from Gaia DR3 QSO candidate table
59	z_gaia_low	float	...	Lower confidence interval of <i>z_gaia</i> taken at 0.15866 quantile
60	z_gaia_up	float	...	Upper confidence interval of <i>z_gaia</i> taken at 0.84134 quantile
61	p_gal	float	...	XGBoost probability of the object being a galaxy
62	p_qso	float	...	XGBoost probability of the object being a quasar
63	p_star	float	...	XGBoost probability of the object being a star

Table 3
(Continued)

Column	Name	Type	Unit	Description
64	z_ph_xgb	float	...	Photometric redshift predicted with XGBoost
65	z_ph_tab	float	...	Photometric redshift predicted with TabNet
66	z_ph_ftt	float	...	Photometric redshift predicted with FT-Transformer
67	z_ph	float	...	Ensemble photometric redshift (mean of z_ph_xgb, z_ph_tab, and z_ph_ftt)
68	z_xp_nn	float	...	Spectral redshift predicted with RegNet using Gaia low-resolution spectroscopy

Note. This table is also available on the PaperData Repository of the National Astronomical Data Center of China at doi:[10.12149/101575](https://doi.org/10.12149/101575).

(This table is available in its entirety in machine-readable form in the [online article](#).)

Table 4
Format of the CatGlobe Quasar Candidate Catalog

Column	Name	Type	Unit	Description
1	source_id	long	...	Gaia DR3 unique source identifier
2	ra	double	deg	Gaia DR3 R.A. (ICRS) at Ep = 2016.0
3	dec	double	deg	Gaia DR3 decl. (ICRS) at Ep = 2016.0
4	l	double	deg	Galactic longitude
5	b	double	deg	Galactic latitude
6	parallax	double	mas	Parallax
7	parallax_error	double	mas	Standard error of parallax
8	pmra	float	mas yr ⁻¹	Proper motion in R.A. direction
9	pmra_error	float	mas yr ⁻¹	Standard error of pmra
10	pmdec	float	mas yr ⁻¹	Proper motion in decl. direction
11	pmdec_error	float	mas yr ⁻¹	Standard error of pmdec
12	pmra_pmdec_corr	float	...	Correlation between pmra and pmdec
13	phot_bp_mean_mag	float	mag	Integrated BP mean magnitude
14	phot_g_mean_mag	float	mag	G-band mean magnitude
15	phot_rp_mean_mag	float	mag	Integrated RP mean magnitude
16	bp_rp	float	mag	BP – RP color
17	phot_bp_rp_excess_factor	float	...	BP/RP excess factor
18	catwise_id	string	...	CatWISE2020 source id
19	ra_cat	double	deg	CatWISE2020 R.A. (ICRS)
20	dec_cat	double	deg	CatWISE2020 decl. (ICRS)
21	snrw1pm	float	...	Flux S/N in band-1 (W1)
22	snrw2pm	float	...	Flux S/N in band-2 (W2)
23	w1mpropm	float	mag	WPRO magnitude in band-1 (Vega)
24	e_w1mpropm	float	mag	Uncertainty in w1mpropm
25	w2mpropm	float	mag	WPRO magnitude in band-2 (Vega)
26	e_w2mpropm	float	mag	Uncertainty in w2mpropm
27	phot_bp_rp_excess_factor_c	float	...	Corrected phot_bp_rp_excess_factor
28	log_fpm0	float	...	Logarithm probability density of zero proper motion ($\log f_{\text{PM0}}$)
29	z_gaia	float	...	Redshift estimate from Gaia DR3 QSO candidate table
30	p_gal	float	...	XGBoost probability of the object being a galaxy
31	p_qso	float	...	XGBoost probability of the object being a quasar
32	p_star	float	...	XGBoost probability of the object being a star
33	z_ph_xgb	float	...	Photometric redshift predicted with XGBoost
34	z_ph_tab	float	...	Photometric redshift predicted with TabNet
35	z_ph_ftt	float	...	Photometric redshift predicted with FT-Transformer
36	z_ph	float	...	Ensemble photometric redshift (mean of z_ph_xgb, z_ph_tab, and z_ph_ftt)
37	z_xp_nn	float	...	Spectral redshift predicted with RegNet using Gaia low-resolution spectroscopy

Note. This table is also available on the PaperData Repository of the National Astronomical Data Center of China at doi:[10.12149/101575](https://doi.org/10.12149/101575).

(This table is available in its entirety in machine-readable form in the [online article](#).)

surveys, WISE photometry from CatWISE2020 and AllWISE, and derived quantities (source probabilities, photometric and spectroscopic redshifts) from this work. Descriptions of the format of the catalog are shown in Table 3.

The sky density distribution of CatSouth sources is shown in Figure 4. The low density around LMC and SMC is due to the additional source filtering described in Section 3.2. Apart from the Galactic plane and the Magellanic Clouds, regions with

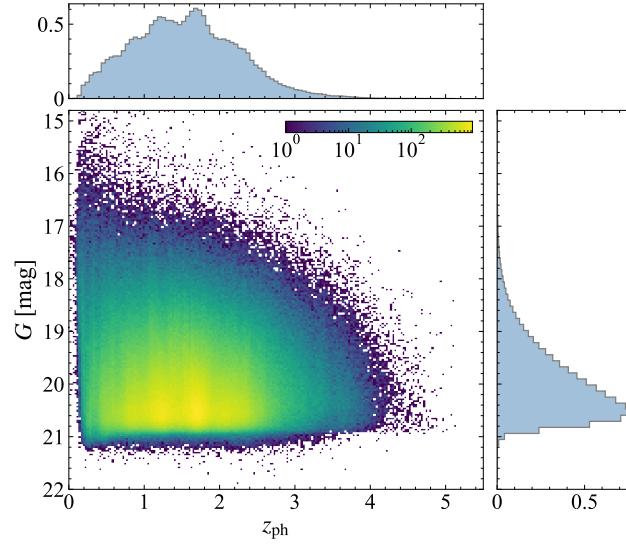


Figure 5. Apparent G magnitude and photometric redshift distribution of CatGlobe quasar candidates. The G magnitude is not extinction-corrected.

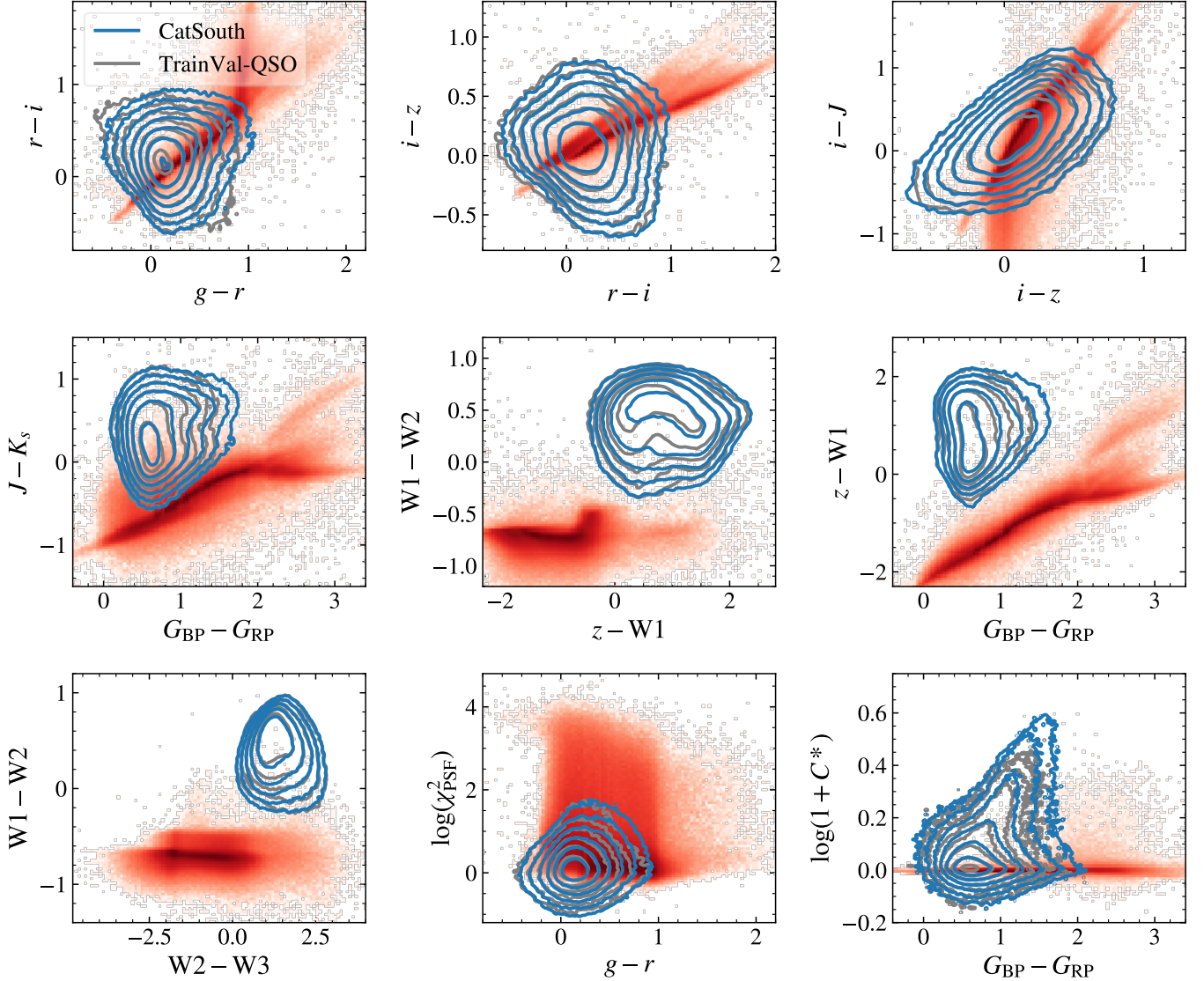


Figure 6. Two-dimensional feature representations (color-color and morphology-color diagrams) of sources in the CatSouth quasar candidate catalog (blue contours), quasars from the training/validation sample (gray contours), and stars from the training/validation sample (red-shaded density plots). To avoid clutter in the figures, galaxies are not plotted. All magnitudes are in the AB system and not dereddened.

$\delta \gtrsim 0^\circ$ show low source density because of the lack of sky coverage of SMSS (see Section 6.3 of C. A. Onken et al. 2024).

The median sky density of CatSouth is 41.7 deg^{-2} , which is approximately two-thirds of that of CatNorth (61.96 deg^{-2}). This difference in sky density is mainly because the PS1 DR1 has a 1-to-2 magnitude deeper detection limit than SMSS DR4, and CatNorth is built on the former. CatNorth and CatSouth have 577,229 sources in common, which are mainly located at $-30^\circ \lesssim \delta \lesssim 16^\circ$. Despite that CatNorth is more complete than CatSouth, CatSouth contains 14,543 objects in the PS1 footprint that are not in CatNorth. In addition, crossmatching CatSouth with the full Milliquas v8 using a radius of $1''.5$ gives 128,596 sources in common, leaving 792,940 sources new to Milliquas. Such inclusion of new candidates indicates that different surveys and selection methods complement each other in building a more complete quasar sample.

By combining CatNorth and CatSouth, we generate ‘‘CatGlobe,’’ a unified all-sky quasar candidate catalog based on Gaia DR3. We keep the CatNorth entry in the CatGlobe catalog when a source is in both CatNorth and CatSouth because CatNorth offers better depth. The final CatGlobe catalog includes 1,889,813 unique sources, whose table description is listed in Table 4, and sky density distribution is shown in Figure 4. The distributions of the apparent G magnitudes and photometric redshifts (z_{ph}) of the CatGlobe quasar candidates are displayed in Figure 5. The photometric redshift ranges from 0 to approximately 5. Similar to the CatNorth catalog, the CatSouth and CatGlobe catalogs probe the bright end of quasars.

As has been shown in Figure 14 of Y. Fu et al. (2024) for CatNorth, and Figure 6 in this work for CatSouth, our machine learning classification method recovers a quasar candidate sample with color–color and morphology–color distributions that are well-aligned with the bona fide quasar samples from the training/validation sets.

In addition, we examine the proper-motion distributions of the original GDR3 QSO candidate sample and the CatGlobe catalog in Figure 7. The original GDR3 QSO candidate sample shows asymmetric proper-motion distributions in both R.A. and decl. directions, many sources with relatively large proper motions (e.g., pmra or pmdec greater than 10 mas yr^{-1}), and overdensities in off-zero regions. Such behavior indicates modestly high stellar contamination in the GDR3 QSO candidates. In contrast, the CatGlobe quasar candidates show highly symmetric proper-motion distributions in both directions, with only a small number of sources showing large proper motions. The improvement in the proper-motion distributions suggests a significant enhancement of the purity of quasars in CatGlobe over the original GDR3 QSO candidates.

6. Summary and Conclusions

In this paper, we present the CatSouth quasar candidate catalog, an improved Gaia DR3 quasar candidate catalog in the southern sky. By combining Gaia DR3 astrometry and photometry with complementary data from SkyMapper DR4, VISTA surveys, and CatWISE2020, we implement a machine learning classification selection method that effectively purifies the original Gaia quasar candidate catalog. We construct robust training/validation sets using spectroscopically confirmed quasars and high-quality CatNorth sources

whose CNN-derived redshifts closely agree with the original Gaia estimates. With a set of carefully selected photometric and morphological features, the XGBoost classifier produces a high-purity sample of quasar candidates in the southern regions.

For quasar candidates with available Gaia BP/RP spectra, we directly derive spectroscopic redshifts using the pretrained convolutional neural network (RegNet) from CatNorth. We train an ensemble photometric redshift model for the full sample based on XGBoost, TabNet, and FT-Transformer algorithms. Our ensemble photometric redshifts demonstrate competitive performance, with validation metrics indicating a significant improvement over the original Gaia redshifts and high consistency with the CNN-based spectroscopic redshifts.

The CatSouth catalog has limiting magnitudes of approximately $G \lesssim 21$ and $i \lesssim 21$. CatSouth has a median sky density of 41.7 deg^{-2} , which is lower than that of CatNorth (61.96 deg^{-2}) due to the shallower depth of SMSS DR4 than PS1. Nevertheless, CatSouth complements CatNorth and other existing quasar catalogs, especially in the Southern Hemisphere. By merging CatSouth with CatNorth, we produce the unified all-sky CatGlobe catalog, which provides a valuable resource for spectroscopic follow-up surveys, cosmological studies, and the construction of future Gaia celestial reference frames.

Our results demonstrate the effectiveness of combining multiwavelength data and advanced machine learning techniques in the selection and redshift estimation of quasar candidates. The CatSouth catalog extends the quasar candidate sample to the Southern Hemisphere and improves the overall quality and reliability of redshift estimates, paving the way for future spectroscopic campaigns and enhanced celestial reference frame construction.

Acknowledgments

We acknowledge the support of the National Key R&D Program of China (2022YFF0503401). We are thankful for the support from the National Science Foundation of China (12133001) and the science research grant from the China Manned Space Project with No. CMS-CSST-2021-A06. The work is supported by the High-Performance Computing Platform of Peking University. We thank the referee for helpful suggestions to improve this paper.

K.I.C. acknowledges funding from the Dutch Research Council (NWO) through the award of the Vici Grant VI.C.212.036 and funding from the Netherlands Research School for Astronomy (NOVA). 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 uses data products from the SkyMapper Southern Survey data releases. The national facility capability for SkyMapper has been funded through ARC LIEF grant LE130100104 from the Australian Research Council, awarded to the University of Sydney, the Australian National University, Swinburne University of Technology, the University of Queensland, the University of Western Australia, the University of Melbourne, Curtin University of Technology, Monash University, and the Australian Astronomical

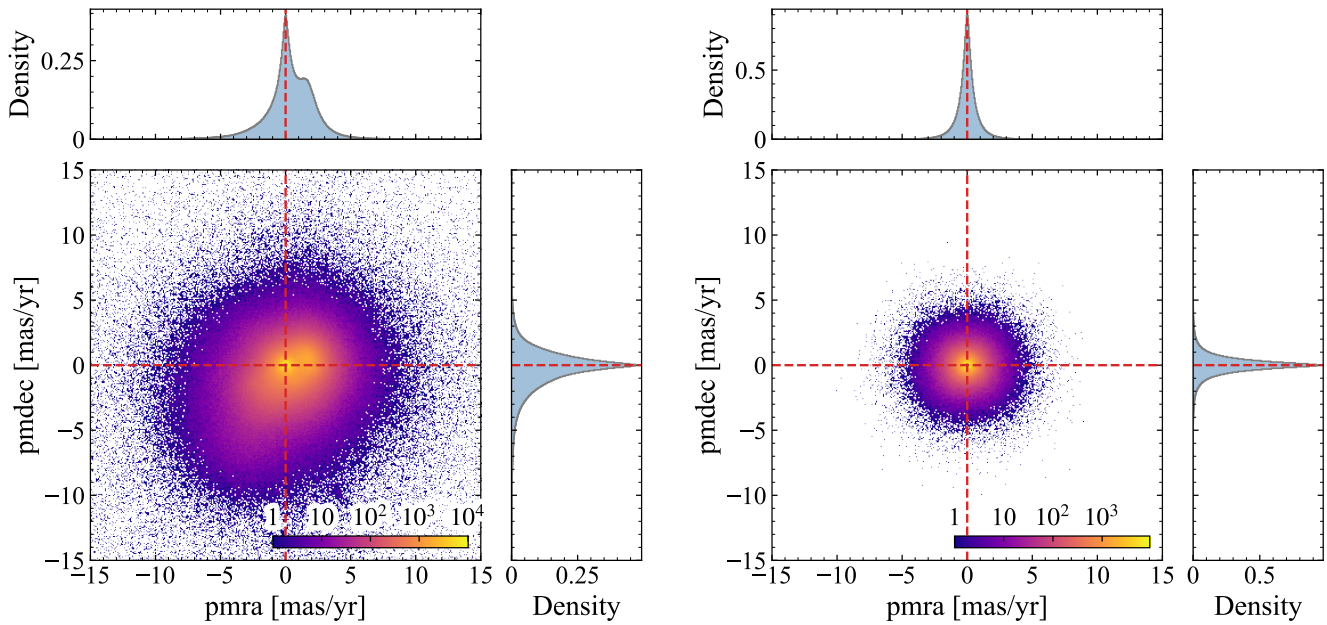


Figure 7. Left: joint and marginal density distributions of proper motions of the original GDR3 QSO candidates in R.A. and decl. directions. The two-dimensional joint density plot is color coded with the number of sources. Right: same as the left panel, but for CatGlobe quasar candidates from this work.

Observatory. SkyMapper is owned and operated by The Australian National University’s Research School of Astronomy and Astrophysics. The survey data were processed and provided by the SkyMapper Team at ANU. The SkyMapper node of the All-Sky Virtual Observatory (ASVO) is hosted at the National Computational Infrastructure (NCI). Development and support of the SkyMapper node of the ASVO has been funded in part by Astronomy Australia Limited (AAL) and the Australian Government through the Commonwealth’s Education Investment Fund (EIF) and National Collaborative Research Infrastructure Strategy (NCRIS), particularly the National eResearch Collaboration Tools and Resources (NeCTAR) and the Australian National Data Service Projects (ANDS). This research uses services or data provided by the Astro Data Lab, which is part of the Community Science and Data Center (CSDC) Program of NSF NOIRLab. NOIRLab is operated by the Association of Universities for Research in Astronomy (AURA), Inc. under a cooperative agreement with the U.S. National Science Foundation. This publication uses data products from the Wide-field Infrared Survey Explorer, 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. Funding for the Sloan Digital Sky Survey V has been provided by the Alfred P. Sloan Foundation, the Heising-Simons Foundation, the National Science Foundation, and the Participating Institutions. SDSS acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS website is www.sdss.org. SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration, including the Carnegie Institution for Science, Chilean National Time Allocation Committee (CNTAC) ratified researchers, the Gotham Participation Group, Harvard University, Heidelberg University, The Johns Hopkins University, L’Ecole polytechnique fédérale de Lausanne (EPFL), Leibniz-Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für

Extraterrestrische Physik (MPE), Nanjing University, National Astronomical Observatories of China (NAOC), New Mexico State University, The Ohio State University, Pennsylvania State University, Smithsonian Astrophysical Observatory, Space Telescope Science Institute (STScI), the Stellar Astrophysics Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Illinois at Urbana-Champaign, University of Toronto, University of Utah, University of Virginia, Yale University, and Yunnan University. This research has made use of the VizieR catalog access tool, CDS, Strasbourg Astronomical Observatory, France (doi:[10.26093/cds/vizier](https://doi.org/10.26093/cds/vizier)).

Facilities: Gaia, SkyMapper, VISTA, WISE.

Software: astropy (Astropy Collaboration et al. 2013, 2018, 2022), astroquery (A. Ginsburg et al. 2019), dustmaps (G. Green 2018), FT-Transformer (Y. Gorishniy et al. 2021), GaiaXPpy (D. Ruz-Mieres 2023), healpy (A. Zonca et al. 2019), HEALPix (K. M. Górski et al. 2005), KDEpy (T. Odland 2018), MOCpy (P. Fernique et al. 2022; M. Baumann et al. 2024), optuna (T. Akiba et al. 2019), pandas (W. McKinney 2010; The Pandas Development Team 2022), pytorch (J. Ansel et al. 2024), scikit-learn (F. Pedregosa et al. 2011), TabNet (S. Ö. Arik & T. Pfister 2021), TOPCAT (M. B. Taylor 2005), XGBoost (T. Chen & C. Guestrin 2016).

Appendix ADQL Queries for Selecting Gaia DR3 Stellar Samples

Here we provide the ADQL queries for selecting Gaia DR3 stellar samples from the Gaia Science Archive.²⁴

A.1. The Gaia DR3 OBA Sample

```
SELECT gs.source_id, gs.ra, gs.dec, l, b,
       parallax, parallax_error, parallax_over_error,
       pm, pmra, pmra_error, pmdec, pmdec_error,
```

²⁴ <http://gea.esac.esa.int/archive/>





```
pmra_pmdec_corr, phot_g_mean_mag,
phot_bp_mean_mag, phot_rp_mean_mag, phot_bp_rp_ex-
cess_factor, astrometric_excess_noise, astro-
metric_excess_noise_sig, astrometric_params_solved,
ruwe, ipd_frac_multi_peak, s.vtan_flag, gs.teff_gsp-
phot, gs.distance_gspphot, ap.teff_esphs, ap.tef-
f_esphs_uncertainty, ap.spectraltype_esphs, ap.
flags_esphs FROM gaiadr3.gaia_source AS gs
JOIN gaiadr3.gold_sample_oba_stars
AS s USING (source_id) JOIN gaiadr3.astro-
physical_parameters AS ap USING (source_id) WHERE gs.
dec < 16 AND phot_g_mean_mag > 8 AND ruwe < 1.4
AND astrometric_params_solved = 31
AND parallax_over_error > 10
AND ipd_frac_multi_peak < 6
AND phot_bp_n_blended_transits < 10 AND ap.
teff_esphs > 7000 AND gs.
classprob_dsc_combmod_star > 0.9 AND s.vtan_flag = 0
```

A.2. The Gaia DR3 FGKM Sample

```
SELECT gs.source_id, gs.ra, gs.dec, l, b,
parallax, parallax_error, parallax_over_error,
pm, pmra, pmra_error, pmdec, pmdec_error,
pmra_pmdec_corr, phot_g_mean_mag,
phot_bp_mean_mag, phot_rp_mean_mag, phot_bp_rp_ex-
cess_factor, astrometric_excess_noise, astro-
metric_excess_noise_sig, astrometric_params_solved,
ruwe, ipd_frac_multi_peak, gs.teff_gspphot, aps.
teff_gspphot_marcs, aps.teff_gspphot_phoenix
FROM gaiadr3.gaia_source AS gs JOIN gaiadr3.astro-
physical_parameters AS ap USING (source_id)
JOIN gaiadr3.astrophysical_parameters_supp
AS aps USING (source_id) WHERE gs.dec < 16
AND phot_g_mean_mag > 8 AND ruwe < 1.4
AND random_index BETWEEN 0 AND 450000000
AND astrometric_params_solved = 31
AND parallax_over_error > 15
AND ipd_frac_multi_peak < 6
AND phot_bp_n_blended_transits < 10 AND gs.
teff_gspphot > 2500 AND gs.teff_gspphot < 7500 AND gs.
distance_gspphot < 1000 / (parallax * parallax_error)
AND gs.distance_gspphot > 1000 / (parallax * paral-
lax_error) AND (gs.libname_gspphot = 'MARCS' OR gs.
libname_gspphot = 'PHOENIX') AND ap.
logposterior_gspphot > -4000 AND gs.
classprob_dsc_combmod_star > 0.9 AND gs.mh_gspphot > -
0.8 AND ABS(aps.teff_gspphot_marcs - aps.
teff_gspphot_phoenix + 65) < 150
AND radius_gspphot < 100 AND mg_gspphot < 12
AND phot_bp_n_obs > 19 AND phot_rp_n_obs > 19
AND phot_g_n_obs > 150
```

ORCID iDs

Yuming Fu (傅煜铭)  <https://orcid.org/0000-0002-0759-0504>
Xue-Bing Wu  <https://orcid.org/0000-0002-7350-6913>
R. J. Bouwens  <https://orcid.org/0000-0002-4989-2471>
Karina I. Caputi  <https://orcid.org/0000-0001-8183-1460>
Yuxuan Pang  <https://orcid.org/0009-0005-3823-9302>
Rui Zhu  <https://orcid.org/0000-0002-0792-2353>
Da-Ming Yang  <https://orcid.org/0000-0002-6769-0910>
Jin Qin  <https://orcid.org/0000-0002-7533-1264>
Huimei Wang  <https://orcid.org/0000-0001-8803-0738>
Christian Wolf  <https://orcid.org/0000-0002-4569-016X>
Yifan Li  <https://orcid.org/0009-0002-9096-2299>

Ravi Joshi  <https://orcid.org/0000-0002-5535-4186>
Yanxia Zhang  <https://orcid.org/0000-0002-6610-5265>
Zhi-Ying Huo  <https://orcid.org/0009-0003-3066-2830>

References

- Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017, *ApJL*, **848**, L13
Abdurro'uf, Accetta, K., Aerts, C., et al. 2022, *ApJS*, **259**, 35
Ahumada, R., Allende Prieto, C., Almeida, A., et al. 2020, *ApJS*, **249**, 3
Ai, Y. L., Wu, X.-B., Yang, J., et al. 2016, *AJ*, **151**, 24
Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. 2019, in Proc. of the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, ed. A. Teredesai, V. Kumar, Y. Li et al. (New York: Association for Computing Machinery), 2623
Alksnis, A., Balklavs, A., Dzervitis, U., et al. 2001, *BaltA*, **10**, 1
Almeida, A., Anderson, S. F., Argudo-Fernández, M., et al. 2023, *ApJS*, **267**, 44
Ansel, J., Yang, E., He, H., et al. 2024, in Proc. of the 29th ACM Int. Conf. on Architectural Support for Programming Languages and Operating Systems, Vol. 2, ed. N. Abu-Ghazaleh et al. (New York: Association for Computing Machinery), 929
Arik, S. Ö., & Pfister, T. 2021, in Proc. 35th AAAI Conf. on Artificial Intelligence, ed. K. Leyton-Brown (Palo Alto, CA: AAAI Press), 6679
Assef, R. J., Stern, D., Noirot, G., et al. 2018, *ApJS*, **234**, 23
Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, *ApJ*, **935**, 167
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, *AJ*, **156**, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, *A&A*, **558**, A33
Bañados, E., Venemans, B. P., Mazzucchelli, C., et al. 2018, *Natur*, **553**, 473
Baumann, M., Marchand, M., Pineau, F.-X., et al. 2024, cds-astro/mocpy: v0.17.1, Zenodo, doi:[10.5281/zenodo.14205461](https://doi.org/10.5281/zenodo.14205461)
Bessell, M., Bloxham, G., Schmidt, B., et al. 2011, *PASP*, **123**, 789
Blanton, M. R., Bershad, M. A., Abolfathi, B., et al. 2017, *AJ*, **154**, 28
Boutsia, K., Grazian, A., Calderone, G., et al. 2020, *ApJS*, **250**, 26
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, *ApJ*, **686**, 1503
Calderone, G., Boutsia, K., Cristiani, S., et al. 2019, *ApJ*, **887**, 268
Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv:1612.05560
Chen, T., & Guestrin, C. 2016, in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD'16, ed. B. Krishnapuram & M. Shah (New York: Association for Computing Machinery), 785
Cioni, M. R., Clementini, G., Girardi, L., et al. 2011, *AJ*, **144**, 25
Cioni, M. R. L., Clementini, G., Girardi, L., et al. 2011, *A&A*, **527**, A116
Colless, M., Dalton, G., Maddox, S., et al. 2001, *MNRAS*, **328**, 1039
Croom, S. M., Richards, G. T., Shanks, T., et al. 2009, *MNRAS*, **399**, 1755
Croom, S. M., Smith, R. J., Boyle, B. J., et al. 2004, *MNRAS*, **349**, 1397
Cupani, G., Calderone, G., Selvelli, P., et al. 2022, *MNRAS*, **510**, 2509
Dalton, G. B., Caldwell, M., Ward, A. K., et al. 2006, *Proc. SPIE*, **6269**, 62690X
Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, *AJ*, **145**, 10
DESI Collaboration, Abdul-Karim, M., Adame, A. G., et al. 2025, arXiv:2503.14745
DESI Collaboration, Adame, A. G., Aguilar, J., et al. 2024, *AJ*, **168**, 58
Dey, A., Rabinowitz, D., Karcher, A., et al. 2016, *Proc. SPIE*, **9908**, 99082C
Dey, A., Schlegel, D. J., Lang, D., et al. 2019, *AJ*, **157**, 168
Di Matteo, T., Springel, V., & Hernquist, L. 2005, *Natur*, **433**, 604
Dong, X. Y., Wu, X.-B., Ai, Y. L., et al. 2018, *AJ*, **155**, 189
Edge, A., Sutherland, W., Kuijken, K., et al. 2013, *Msngr*, **154**, 32
Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, *AJ*, **142**, 72
Emerson, J., McPherson, A., & Sutherland, W. 2006, *Msngr*, **126**, 41
Fan, X., Bañados, E., & Simcoe, R. A. 2023, *ARA&A*, **61**, 373
Fernique, P., Nebot, A., Durand, D., et al. 2022, MOC: Multi-order Coverage Map, v2.0 IVOA Recommendation 2022-07-27, International Virtual Observatory Alliance, doi:[10.5479/ADS/bib/2022ivoa.spec.0727F](https://doi.org/10.5479/ADS/bib/2022ivoa.spec.0727F)
Flesch, E. W. 2023, *OJAp*, **6**, 49
Fu, Y., Wu, X.-B., Li, Y., et al. 2024, *ApJS*, **271**, 54
Fu, Y., Wu, X.-B., Yang, Q., et al. 2021, *ApJS*, **254**, 6
Fu, Y., Wu, X.-B., Jiang, L., et al. 2022, *ApJS*, **261**, 32
Gaia Collaboration, Bailer-Jones, C. A. L., Teyssier, D., et al. 2023a, *A&A*, **674**, A41
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021, *A&A*, **649**, A1
Gaia Collaboration, Creevey, O. L., Sarro, L. M., et al. 2023b, *A&A*, **674**, A39

- Gaia Collaboration, Klioner, S. A., Lindegren, L., et al. 2022, *A&A*, **667**, A148
- Gaia Collaboration, Mignard, F., Klioner, S. A., et al. 2018, *A&A*, **616**, A14
- Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. 2023c, *A&A*, **674**, A1
- Ginsburg, A., Sipőcz, B. M., Brasseur, C. E., et al. 2019, *AJ*, **157**, 98
- González-Fernández, C., Hodgkin, S. T., Irwin, M. J., et al. 2018, *MNRAS*, **474**, 5459
- Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. 2021, Advances in Neural Information Processing Systems 34, ed. M. Ranzato et al. (NeurIPS), 18932, https://papers.nips.cc/paper_files/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html
- Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, *ApJ*, **622**, 759
- Green, G. 2018, *JOSS*, **3**, 695
- Hon, W. J., Webster, R. L., & Wolf, C. 2025, *MNRAS*, **536**, 3611
- Huchra, J. P., Macri, L. M., Masters, K. L., et al. 2012, *ApJS*, **199**, 26
- Hughes, A. C. N., Bailer-Jones, C. A. L., & Jamal, S. 2022, *A&A*, **668**, A99
- Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, *A&A*, **457**, 841
- Inayoshi, K., Visbal, E., & Haiman, Z. 2020, *ARA&A*, **58**, 27
- Jarvis, M. J., Bonfield, D. G., Bruce, V. A., et al. 2013a, *MNRAS*, **428**, 1281
- Jarvis, M. J., Häußler, B., & McAlpine, K. 2013b, *Msngr*, **154**, 26
- Jin, J.-J., Wu, X.-B., Fu, Y., et al. 2023, *ApJS*, **265**, 25
- Jin, X., Zhang, Y., Zhang, J., et al. 2019, *MNRAS*, **485**, 4539
- Jones, D. H., Read, M. A., Saunders, W., et al. 2009, *MNRAS*, **399**, 683
- Kao, W.-B., Zhang, Y., & Wu, X.-B. 2024, *PASJ*, **76**, 653
- Keller, S. C., Schmidt, B. P., Bessell, M. S., et al. 2007, *PASA*, **24**, 1
- Kim, Y., Kim, M., Im, M., et al. 2024, *ApJS*, **275**, 46
- Koornneef, J., Bohlin, R., Buser, R., Horne, K., & Turnshek, D. 1986, *Hia*, **7**, 833
- Kormendy, J., & Ho, L. C. 2013, *ARA&A*, **51**, 511
- Lang, D. 2014, *AJ*, **147**, 108
- Li, C., Zhang, Y., Cui, C., et al. 2022, *MNRAS*, **509**, 2289
- Li, J., Liu, C., Zhang, B., et al. 2021, *ApJS*, **253**, 45
- Liu, C., Côté, P., Peng, E. W., et al. 2020, *ApJS*, **250**, 17
- Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, *ApJS*, **250**, 8
- Ma, C., Arias, E. F., Bianco, G., et al. 2009, *ITN*, **35**, 1
- Maddox, N., Hewett, P. C., Warren, S. J., & Croom, S. M. 2008, *MNRAS*, **386**, 1605
- Mainzer, A., Bauer, J., Grav, T., et al. 2011, *ApJ*, **731**, 53
- Marocco, F., Eisenhardt, P. R. M., Fowler, J. W., et al. 2020, CatWISE2020 Catalog, IPAC, doi:10.26131/IRSA551
- Marocco, F., Eisenhardt, P. R. M., Fowler, J. W., et al. 2021, *ApJS*, **253**, 8
- McCracken, H. J., Milvang-Jensen, B., Dunlop, J., et al. 2012, *A&A*, **544**, A156
- McKinney, W. 2010, Proc. 9th Python in Science Conf. (SciPy 2010), ed. S. van der Walt & J. Millman, (SciPy), 56
- McMahon, R. G., Banerji, M., Gonzalez, E., et al. 2013, *Msngr*, **154**, 35
- Mignard, F., Klioner, S., Lindegren, L., et al. 2016, *A&A*, **595**, A5
- Minniti, D., Lucas, P. W., Emerson, J. P., et al. 2010, *NewA*, **15**, 433
- Nidever, D. L., Dey, A., Fasbender, K., et al. 2021, *AJ*, **161**, 192
- Nidever, D. L., Dey, A., Olsen, K., et al. 2018, *AJ*, **156**, 131
- Odland, T. 2018, tommyod/KDEpy: Kernel Density Estimation in Python, v0.9.10, Zenodo, doi:10.5281/zenodo.2392268
- Onken, C. A., Wolf, C., Bessell, M. S., et al. 2024, *PASA*, **41**, e061
- Onken, C. A., Wolf, C., Bian, F., et al. 2022, *MNRAS*, **511**, 572
- Onken, C. A., Wolf, C., Bessell, M. S., et al. 2019, *PASA*, **36**, e033
- Onken, C. A., Wolf, C., Hon, W. J., et al. 2023, *PASA*, **40**, e010
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *JMLR*, **12**, 2825
- Planck Collaboration, Aghanim, N., Ashdown, M., et al. 2016, *A&A*, **596**, A109
- Rees, M. J. 1986, *MNRAS*, **218**, 25P
- Riello, M., De Angeli, F., Evans, D. W., et al. 2021, *A&A*, **649**, A3
- Ruz-Mieres, D. 2023, gaia-dpci/GaiaXPY: GaiaXPY v2.0.1, Zenodo, doi:10.5281/zenodo.7566303
- Sagi, O., & Rokach, L. 2018, *WIREs Data Min. Knowl. Discov.*, **8**, e1249
- Schindler, J.-T., Fan, X., McGreer, I. D., et al. 2017, *ApJ*, **851**, 13
- Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2003, 2MASS All-Sky Point Source Catalog, IPAC, doi:10.26131/IRSA2
- Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, *AJ*, **131**, 1163
- Stern, D., Assef, R. J., Benford, D. J., et al. 2012, *ApJ*, **753**, 30
- Storey-Fisher, K., Hogg, D. W., Rix, H.-W., et al. 2024, *ApJ*, **964**, 69
- Taylor, M. B. 2005, in ASP Conf. Ser. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert (San Francisco, CA: ASP), 29
- The Pandas Development Team 2022, pandas-dev/pandas: Pandas, v1.5.0, Zenodo, doi:10.5281/zenodo.7093122
- Trump, J. R., Hall, P. B., Reichard, T. A., et al. 2006, *ApJS*, **165**, 1
- Wang, S., & Chen, X. 2019, *ApJ*, **877**, 116
- West, A. A., Morgan, D. P., Bochanski, J. J., et al. 2011, *AJ*, **141**, 97
- Weymann, R. J., Carswell, R. F., & Smith, M. G. 1981, *ARA&A*, **19**, 41
- Wisotzki, L., Christlieb, N., Bade, N., et al. 2000, *A&A*, **358**, 77
- Wolf, C., Onken, C. A., Luvaul, L. C., et al. 2018, *PASA*, **35**, e010
- Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, *AJ*, **140**, 1868
- Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2019, AllWISE Source Catalog, IPAC, doi:10.26131/IRSA1
- Wu, Q., & Shen, Y. 2022, *ApJS*, **263**, 42
- Wu, X.-B., Hao, G., Jia, Z., Zhang, Y., & Peng, N. 2012, *AJ*, **144**, 49
- Wu, X.-B., & Jia, Z. 2010, *MNRAS*, **406**, 1583
- Wu, X.-B., Wang, F., Fan, X., et al. 2015, *Natur*, **518**, 512
- Yang, Q., & Shen, Y. 2023, *ApJS*, **264**, 9
- Yao, S., Wu, X.-B., Ai, Y. L., et al. 2019, *ApJS*, **240**, 6
- Ye, G., Zhang, H., & Wu, Q. 2024, *ApJS*, **275**, 19
- Zaw, I., Chen, Y.-P., & Farrar, G. R. 2019, *ApJ*, **872**, 134
- Zonca, A., Singer, L., Lenz, D., et al. 2019, *JOSS*, **4**, 1298
- Zou, H., Zhang, T., Zhou, Z., et al. 2018, *ApJS*, **237**, 37
- Zou, H., Zhou, X., Fan, X., et al. 2017, *PASP*, **129**, 064101
- Zou, H., Zhou, X., Fan, X., et al. 2019, *ApJS*, **245**, 4