



# Quaia, the Gaia-unWISE Quasar Catalog: An All-sky Spectroscopic Quasar Sample

Kate Storey-Fisher<sup>1</sup> , David W. Hogg<sup>1,2,3</sup> , Hans-Walter Rix<sup>3</sup> , Anna-Christina Eilers<sup>4</sup> , Giulio Fabbian<sup>2,5</sup> ,  
Michael R. Blanton<sup>1</sup> , and David Alonso<sup>6</sup>

<sup>1</sup> Center for Cosmology and Particle Physics, Department of Physics, New York University, 726 Broadway, New York, NY 10003, USA; [k.sf@nyu.edu](mailto:k.sf@nyu.edu)

<sup>2</sup> Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Avenue, New York, NY 10010, USA

<sup>3</sup> Max Planck Institute for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany

<sup>4</sup> MIT Kavli Institute for Astrophysics and Space Research, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

<sup>5</sup> School of Physics and Astronomy, Cardiff University, The Parade, Cardiff, Wales CF24 3AA, UK

<sup>6</sup> Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK

Received 2023 July 1; revised 2023 December 5; accepted 2023 December 5; published 2024 March 18

## Abstract

We present a new, all-sky quasar catalog, Quaia, that samples the largest comoving volume of any existing spectroscopic quasar sample. The catalog draws on the 6,649,162 quasar candidates identified by the Gaia mission that have redshift estimates from the space observatory’s low-resolution blue photometer/red photometer spectra. This initial sample is highly homogeneous and complete, but has low purity, and 18% of even the bright ( $G < 20.0$ ) confirmed quasars have discrepant redshift estimates ( $|\Delta z/(1+z)| > 0.2$ ) compared to those from the Sloan Digital Sky Survey (SDSS). In this work, we combine the Gaia candidates with unWISE infrared data (based on the Wide-field Infrared Survey Explorer survey) to construct a catalog useful for cosmological and astrophysical quasar studies. We apply cuts based on proper motions and colors, reducing the number of contaminants by approximately four times. We improve the redshifts by training a  $k$ -Nearest Neighbor model on SDSS redshifts, and achieve estimates on the  $G < 20.0$  sample with only 6% (10%) catastrophic errors with  $|\Delta z/(1+z)| > 0.2$  (0.1), a reduction of approximately three times (approximately two times) compared to the Gaia redshifts. The final catalog has 1,295,502 quasars with  $G < 20.5$ , and 755,850 candidates in an even cleaner  $G < 20.0$  sample, with accompanying rigorous selection function models. We compare Quaia to existing quasar catalogs, showing that its large effective volume makes it a highly competitive sample for cosmological large-scale structure analyses. The catalog is publicly available at [10.5281/zenodo.10403370](https://doi.org/10.5281/zenodo.10403370).

*Unified Astronomy Thesaurus concepts:* [Quasars \(1319\)](#); [Catalogs \(205\)](#); [Large-scale structure of the universe \(902\)](#); [Gaia \(2360\)](#); [Active galactic nuclei \(16\)](#)

*Supporting material:* animation

## 1. Introduction

Quasars are powerful tools for many fields of astrophysics. They are key probes of accretion physics (e.g., Sunyaev & Zeldovich 1970; Yu et al. 2020), which informs the evolution of active galactic nuclei (AGNs). The evolution of quasars and their host galaxies are intertwined, giving insight into supermassive black hole growth (e.g., Hopkins et al. 2006) as well as massive galaxy formation (e.g., Kormendy & Ho 2013). Studies of the quasar distribution can also be used to understand black hole evolution (e.g., Powell et al. 2020) and halo masses and environmental effects (e.g., DiPompeo et al. 2017). Quasars can also be utilized as background sources for cosmic phenomena such as gravitational lenses (e.g., Claeskens & Surdej 2002), and quasar spectra encode the properties of the intergalactic medium via the Ly $\alpha$  forest (e.g., Rauch 1998).

Quasars are key tracers for large-scale structure cosmology. They reside in peaks of the dark matter distribution and their clustering can be used to measure cosmological parameters, including the growth rate of structure  $f\sigma_8$  (e.g., García-García et al. 2021; Alonso et al. 2023), the Hubble distance  $D_H$  (e.g., Hou et al. 2020), primordial non-Gaussianity (e.g., Leistedt et al. 2014; Castorina et al. 2019; Krolewski et al. 2023), and

the baryon density  $\Omega_b$  (e.g., Yahata et al. 2005). Cross-correlations between quasars and other tracers provide measurements of key cosmological quantities, such as with photometric galaxy samples to measure the baryon acoustic feature (e.g., Ata et al. 2018), with cosmic microwave background (CMB) lensing to constrain quasar bias and the growth of structure (e.g., Sherwin et al. 2012), and with foreground galaxies as a probe of weak lensing (e.g., Ménard & Bartelmann 2002; Scranton et al. 2005; Zarrouk et al. 2021). They can also be used as standardizable candles to measure the expansion rate of the universe (e.g., Setti & Woltjer 1973; Risaliti & Lusso 2015; Lusso et al. 2020). Finally, given the large volume typically covered by quasar samples, the quasar distribution provides a test of the cosmological principle of isotropy and homogeneity (e.g., Secrest et al. 2021; Dam et al. 2023; D. W. Hogg et al. 2024, in preparation).

Many surveys have observed and cataloged quasars, with around 1 million spectroscopically identified and several million when including photometric samples. The Sloan Digital Sky Survey (SDSS) Data Release 16 includes a highly complete catalog of 750,414 quasars with spectroscopic redshifts (Lyke et al. 2020). Photometric surveys observe a much larger number of quasars, at the expense of low redshift accuracy; nearly 3 million quasars with reliable photometric redshifts have been cataloged (Kunsági-Máté et al. 2022), including with the Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010), which imaged the entire sky and



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.

Pan-STARRS (Chambers et al. 2016), which observed three-quarters of the sky. Shu et al. (2019) combined photometry from Gaia DR2 and unWISE (Lang 2014) to identify 2.7 million AGN candidates and estimate their photometric redshifts. Upcoming surveys will observe even more quasars: the Dark Energy Spectroscopic Instrument (DESI; Aghamousa et al. 2023) expects to obtain spectra for 3 million quasars, and the Rubin Observatory’s LSST will photometrically observe upward of 10 million quasars (Ivezić 2017). However, none of these quasar catalogs is both all-sky and contains precise redshift information. The recently released Gaia DR3 quasar candidates (Gaia Collaboration et al. 2023a) constitute a new sample that promises to fill this gap.

The Gaia quasar sample presents a new opportunity to explore these science topics. While the Gaia satellite was designed to map stars in the Milky Way (Gaia Collaboration et al. 2016), it broadly observes bright objects in the sky, which includes many extragalactic sources. Previous work identified a small number of quasars in earlier Gaia data releases, including identification based solely on their astrometric properties (Heintz et al. 2018, 2020). In DR3, the Gaia collaboration released a sample of 6,649,162 quasar candidates that were incidentally observed during the survey (Delchambre et al. 2023; Gaia Collaboration et al. 2023a, 2023b). The sources cover the entire sky and have Gaia blue photometer (BP)/red photometer (RP) spectra, low-resolution spectra covering the wavelength range of 330–1050 nm. These spectra allow for redshift estimates of the sources, with 86% having a precision of  $|\Delta z / (1+z)| < 0.01$  compared to SDSS redshifts when no processing issues affect the redshift estimation (`flags_qsoc = 0` or `flags_qsoc = 16`), which is the case for 20% of the sample; for the full sample including sources with redshift warning flags set, this percentage of high-precision redshifts decreases to 53%. While not as precise as high-resolution spectroscopic redshifts, they are significantly better than photometric redshifts. The median redshift of the sample is  $z = 1.67$ . The Gaia quasar candidate sample was constructed for completeness over purity, and has an estimated purity of 52%; the Gaia Collaboration also suggests criteria for a higher purity ( $\sim 95\%$ ) subcatalog of  $\sim 1.9$  million quasars. Overall, the sample presents an unprecedented resource for quasar science and cosmology.

There are two main issues with this raw Gaia sample. First, the sample contains a large number of non-quasar contaminants. Second, a significant fraction of the redshift estimates are catastrophic errors, due to emission line misidentification given the limitations of the low-resolution spectra. Understanding and eliminating sample contaminants matters greatly in identifying the most extreme (e.g., brightest or most luminous) quasars, which has been addressed in the AllBRICQS catalog (Onken et al. 2023) that draws on Gaia quasar candidates. In this work, we construct a clean quasar catalog across the full magnitude range with lower contamination and improved redshift estimates, with the particular goal of building a catalog appropriate for large-scale structure analyses as well as other quasar science. For both of these, we rely on crossmatches with WISE observations of the quasars (Wright et al. 2010), which adds key infrared (IR) information. To filter out contaminants, we apply color cuts based on the Gaia and WISE photometry, as well as a proper motion cut. To improve the redshifts, we identify quasars that are also observed by SDSS, for which we have highly precise spectroscopic redshifts, and train a  $k$ -

Nearest-Neighbors ( $k$ NN) model based on their photometry and Gaia redshift estimates. Further, the Gaia quasar candidate sample has strong systematic imprints from various observational effects, such as Galactic dust. To model these systematics so that their effects can be mitigated in the analyses of the catalog, we fit a model for the selection function based on observational templates using a Gaussian process. We release both the catalog and selection function as publicly accessible data products.

This paper is organized as follows. In Section 2, we describe the initial data sets used in the construction of the catalog. The construction of the catalog is detailed in Section 3. In Section 4, we present the final catalog and perform verification and comparisons to other samples, and outline the data format. We summarize the catalog and describe the access to the data in Section 5.

## 2. Initial Data Sets

### 2.1. Gaia DR3 Quasar Candidate Sample

While performing its all-sky survey of the Milky Way, the Gaia satellite (Gaia Collaboration et al. 2016) also observed millions of extragalactic objects. These sources—both quasar and galaxy candidates—were first released in Gaia DR3 (Gaia Collaboration et al. 2023a, 2023b). Gaia obtained BP/RP spectra of the sources, which are low-resolution spectra with relatively narrow wavelength ranges; the BP covers 330–680 nm and has  $30 \leq R \leq 100$  and the RP covers 640–1050 nm (Carrasco et al. 2021) with  $70 \leq R \leq 100$ . The raw spectra are not released by Gaia (besides a small subsample—the rest will be released in Gaia DR4), but redshift estimates and other derived information are contained in the catalogs.

The quasar candidates were selected based on multiple classifiers and criteria, described in detail in Gaia Collaboration et al. (2023a). The majority (5.5 million) of the quasar candidates were identified with the Discrete Source Classifier (DSC) module (detailed in Delchambre et al. 2023, a machine-learning model that takes as input the source’s BP/RP spectrum,  $G$ -band magnitude,  $G$ -band variability, parallax, and proper motion, and outputs a class label trained on SDSS spectroscopic classifications. Given these SDSS labels, the results of this module will inherit many of the same selection effects as SDSS. DSC is estimated to have a completeness of over 90% and a purity of around 24% for quasars. Another machine learning model selected over 1 million sources based on their variability, as active nuclei have time-variable accretion; the model inputs were statistics of time series data in all Gaia bands as well as photometric and astrometric quantities, as detailed in Rimoldini et al. (2023). Additionally, a set of nearly 1 million sources was selected based on their surface brightness profile; this selection used existing major quasar catalogs to compile an initial list of sources, which were then processed by the Gaia surface brightness profile module (Ducourant et al. 2023). This module included quasars in the candidates catalog which passed certain criteria, including having Gaia observations covering  $>86\%$  of the source’s surface area and a confident assessment (positive or negative) of host galaxy presence. Finally, the 1.6 million sources used to define the Gaia-CRF3 celestial reference frame were contributed, which are based on crossmatches of Gaia to external quasar catalogs. A large fraction of sources are identified as quasars by multiple of these methods; the overlapping

Fraction of objects in:

That are in:	Gaia DR3 Quasar Candidates (N=6,649,162)	Gaia DR3 Quasar Candidates with redshift estimates (N=6,375,063)	Gaia DR3 'Purer' (N=1,942,825)	Quaia superset (N=1,518,782)	Quaia clean (N=1,414,385)	SDSS DR16Q Quasars (N=638,083)
Gaia DR3 Quasar Candidates (N=6,649,162)	1.00	1.00	1.00	1.00	1.00	0.51
Gaia DR3 Quasar Candidates with redshift estimates (N=6,375,063)	0.96	1.00	0.89	1.00	1.00	0.50
Gaia DR3 'Purer' (N=1,942,825)	0.29	0.27	1.00	0.90	0.95	0.50
Quaia superset (N=1,518,782)	0.23	0.24	0.70	1.00	1.00	0.39
Quaia clean (N=1,414,385)	0.21	0.22	0.69	0.93	1.00	0.38
SDSS DR16Q Quasars (N=638,083)	0.05	0.05	0.16	0.16	0.17	1.00
unWISE (N=2,214,734,224)	0.30	0.29	0.90	1.00	1.00	0.78

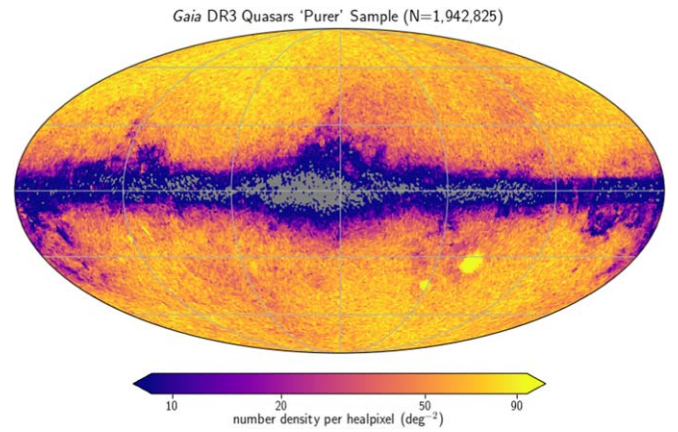
**Figure 1.** A summary of the overlaps between the various data sets and subsamples used in this work. The values describe the fraction of objects in each column's sample that are in each row's sample. Note that we only list unWISE as a row because the inverse is not relevant to this work.

contributions are shown in Figure 3 of Gaia Collaboration et al. (2023a). The full quasar candidate sample contains 6,649,162 sources,<sup>7</sup> selected for high completeness, but with a low purity estimated to be around 52% (Gaia Collaboration et al. 2023a). We show the overlaps between this Gaia quasar candidate sample and other samples and subsamples used and constructed in this work in Figure 1.

Most of the quasar candidates (6,375,063) are assigned redshifts using the Quasar Classifier (QSOC) module, which uses a chi-squared approach on the quasars' BP/RP spectra compared to composite spectra from SDSS DR12Q (Delchambre et al. 2023). We refer to these Gaia redshift estimates as  $z_{\text{Gaia}}$ . Many of these redshifts are determined by a single line due to the narrow spectral range, resulting in aliasing issues when lines are misidentified (see Figure 15 in Delchambre et al. 2023). An estimated 63.7% of the redshifts have  $|\Delta z| < 0.1$ , increasing to 97.6% for quasar candidates with no redshift warning flags (this is the case for nearly 80% of quasars with  $G < 18.5$ , but decreases to less than 20% for  $G > 19.5$ ).

Gaia Collaboration et al. (2023a) provide a query to select a purer subsample of the quasar candidates. It requires higher quasar probability thresholds from the various classifiers and excludes surface-brightness-selected galaxies that have close neighbors. This results in 1,942,825 sources with an estimated purity of 95%; 1.7 million of these have Gaia redshifts. The sky distribution of this sample, which we call the Gaia DR3 *purer* sample, is shown in Figure 2. The Gaia DR3 *purer* sample has a low density in the Galactic plane; we speculate that this is largely due to dust extinction making sources too faint to

<sup>7</sup> The Gaia DR3 quasar candidates sample (and all other Gaia data) can be downloaded at <https://gea.esac.esa.int/archive> with table name `gaiadr3.qso_candidates`.



**Figure 2.** Sky distribution of the quasar candidates in the Gaia DR3 *purer* quasar sample, in Galactic coordinates and displayed using a Mollweide projection.

observe at low Galactic latitudes. Gaia DR3 *purer* also has significant overdensities around the LMC and SMC, as the sample still contains stellar contaminants.

For our analysis, we start with the full quasar candidate sample, rather than the Gaia DR3 *purer* sample or cutting on other Gaia pipeline flags, to allow greater completeness and minimize reproducing biases; we compare our catalog with the Gaia Collaboration et al. (2023a) Gaia DR3 *purer* subsample in Section 4.3. We construct a *superset* of our catalog (which is a subset of the Gaia quasar candidates sample) that contains all the information needed for catalog construction: we require that sources are in the Gaia quasar candidates table, have Gaia  $G$ , BP, and RP measurements, unWISE W1 and W2 observations (described in Section 2.2), Gaia-estimated QSOC redshifts, and a maximum  $G$  magnitude of  $G < 20.6$ . This magnitude cut was chosen to be slightly deeper than our desired catalog magnitude limit of  $G < 20.5$ , in order to provide a buffer for redshift estimation. This results in a superset with 1,518,782 sources. We call our final catalog Quaia, so we refer to this as the Quaia superset.

## 2.2. unWISE Quasar Sample

We use the unWISE reprocessing (Lang 2014; Meisner et al. 2019) of WISE (Wright et al. 2010) to contribute IR photometry to Gaia sources. The unWISE coadds combine data from NEOWISE (Mainzer et al. 2011) with the original WISE survey, providing a time baseline 15 times longer. Compared to the original AllWISE catalog, unWISE has deeper imaging and improved modeling of crowded fields. The unWISE catalog (Schlafly et al. 2019) contains measurements in the W1 (3.4  $\mu\text{m}$ ) and W2 (4.6  $\mu\text{m}$ ) bands for over 2 billion sources. We do not use the W3 and W4 bands as these do not go as deep as we need. We perform a crossmatch of the Gaia quasar candidate sample to unWISE sources within 1".<sup>8</sup> We also crossmatch the SDSS training and validation samples (Sections 2.3, 2.4) to unWISE.

When combined with optical photometry, unWISE IR color information is very useful to identify quasars and distinguish them from contaminants. This photometry also contains useful redshift information; recent approaches to estimate redshifts

<sup>8</sup> We use NOIRLab's crossmatch service to perform this operation, available at <https://datalab.noirlab.edu/xmatch.php>.

from photometry with neural networks achieve a mean  $|\Delta z| \sim 0.22$  (Yang et al. 2017; Jin et al. 2019; Kunsági-Máté et al. 2022). In our case of redshift estimates from narrow-range BP/RP spectra, we expect IR photometry to add information that can break line identification degeneracies in order to improve estimates. We incorporate the W1 and W2 bands into both our quasar selection (Section 3.1) and redshift estimation (Section 3.2) procedures.

### 2.3. SDSS DR16 Quasar Sample

The SDSS released the largest spectroscopic quasar catalog in DR16<sup>9</sup> (Lyke et al. 2020). It combines new sources from the extended Baryon Oscillation Spectroscopic Survey (eBOSS), part of SDSS-IV, with previously observed sources from earlier SDSS campaigns. The catalog contains 750,414 quasars, with an estimated 99.8% completeness (compared to the SDSS-III/SEQUELS sample of Myers et al. 2015, which has higher signal-to-noise spectra) and 98.7%–99.7% purity. We remove sources with redshift warnings,  $z_{\text{WARNING}} = 0$ , as well as a handful of sources with unreasonably low or negative redshift estimates ( $z < 0.01$ ). This results in 638,083 sources, which is the sample shown in Figure 1. We crossmatch these with the Gaia catalog, as well as unWISE (Section 2.2), using a maximum separation of  $1''$  on the sky. We remove sources with fewer than five observations in BP ( $\text{phot\_bp\_n\_obs}$ ) or RP ( $\text{phot\_rp\_n\_obs}$ ), following Bailer-Jones (2021), as well as sources that are duplicated in the SDSS star or galaxy samples (Section 2.4). This results in 343,074 sources with both Gaia and unWISE observations that pass these criteria.

We use these to calibrate the cuts to decontaminate our sample (Section 3.1); for this purpose, we only keep sources that are also in the Quiaia superset (sources that are in the Gaia quasar candidates table, have all necessary Gaia and unWISE photometry, Gaia-estimated QSOC redshifts, and  $G < 20.6$ ). This sample contains 246,122 quasars. We also use this sample (after applying the cuts described in Section 3.1) to train our redshift estimation model (Section 3.2). While this spectroscopic sample has quite high completeness and accurate redshift information, we note that it is still imperfect, contains selection effects, and represents only a particular definition of a quasar; these issues will propagate to our catalog.

### 2.4. Contaminant Samples: Galaxies and Stars

To guide the decontamination of our catalog (Section 3.1), we compile known contaminant samples, namely galaxies and stars. For the galaxy sample, we use SDSS spectroscopic galaxies from DR18.<sup>10</sup> Following Bailer-Jones (2021), we include all galaxies with class label GALAXY in the `SpecObj` table, exclude galaxies with subclass labels AGN or AGN BROADLINE, and exclude sources with redshift warnings,  $z_{\text{WARNING}} = 0$ . We crossmatch these with Gaia DR3 and unWISE with a  $1''$  radius, and remove sources with fewer than five observations in BP or RP, as for the SDSS quasars. We also remove apparent stellar contaminants from the galaxies sample with the cut in  $G - \text{RP}$  and  $\text{BP} - G$  from Equation (1) of Bailer-Jones et al. (2019), and additionally remove sources duplicated in the SDSS quasar or star samples. This leaves

600,897 crossmatched SDSS galaxies in our sample; 1316 of these are in the Quiaia superset.

For the star sample, we also use SDSS DR18 sources, selecting objects with the class label STAR in the `SpecObj` table. As for the quasars and galaxies, we crossmatch these with Gaia DR3 with a  $1''$  radius and remove sources with fewer than five observations in BP or RP, and remove sources duplicated in the other samples. This results in a stellar sample with 482,080 crossmatched SDSS-Gaia stars, with 2276 of these in the superset.

For the decontamination procedure, we also compile a sample of sources in or near the LMC or SMC, as most of these will be stellar contaminants but have different properties than the SDSS star sample. To do this, we select all sources in the Gaia quasar candidates table that are within  $3^\circ$  of the center of the LMC or  $1.5^\circ$  from the center of the SMC. While this may include stars not actually in the LMC or SMC, we have chosen these fairly narrow radii in order to capture mostly LMC and SMC stars and a few potential quasars. Additionally requiring that these have unWISE photometry, this gives 11,770 LMC- and SMC-adjacent stars; 9927 are in the superset.

## 3. Catalog Construction

### 3.1. Decontamination with Proper Motions and unWISE Colors

The full Gaia quasar candidate sample is known to contain a significant fraction of contaminants (stars and other non-quasars, such as galaxies). The stellar contaminants might include sources such as brown dwarfs, which have similar colors as high-redshift quasars, and potentially blue horizontal branch stars, blue stragglers, and white dwarfs, which are UV bright like lower-redshift quasars. To remove stellar contaminants, we make an initial cut on proper motion  $\mu$ , as quasars should have negligible proper motions due to their large distances. The value of  $\mu$  has a dependence on  $G$ , so we make a cut in this space. To guide this cut, we use labeled sources: SDSS quasars, SDSS galaxies, SDSS stars, and Gaia LMC- and SMC-adjacent stars, as described in Sections 2.3 and 2.4. The  $G - \mu$  distributions of these sources are shown in the top panel of Figure 3. In the middle panel, we show the intersection of these labeled sources with our Quiaia superset, which consists of sources in the Gaia quasar candidates table that have Gaia redshift estimates, complete Gaia, and unWISE photometry, and are below  $G < 20.6$ . We see that the SDSS quasars tend to have much smaller proper motions than the other types of sources, with a very linear edge to the  $G$  dependence at the high proper motion side of the distribution. Based on this, we choose the cut

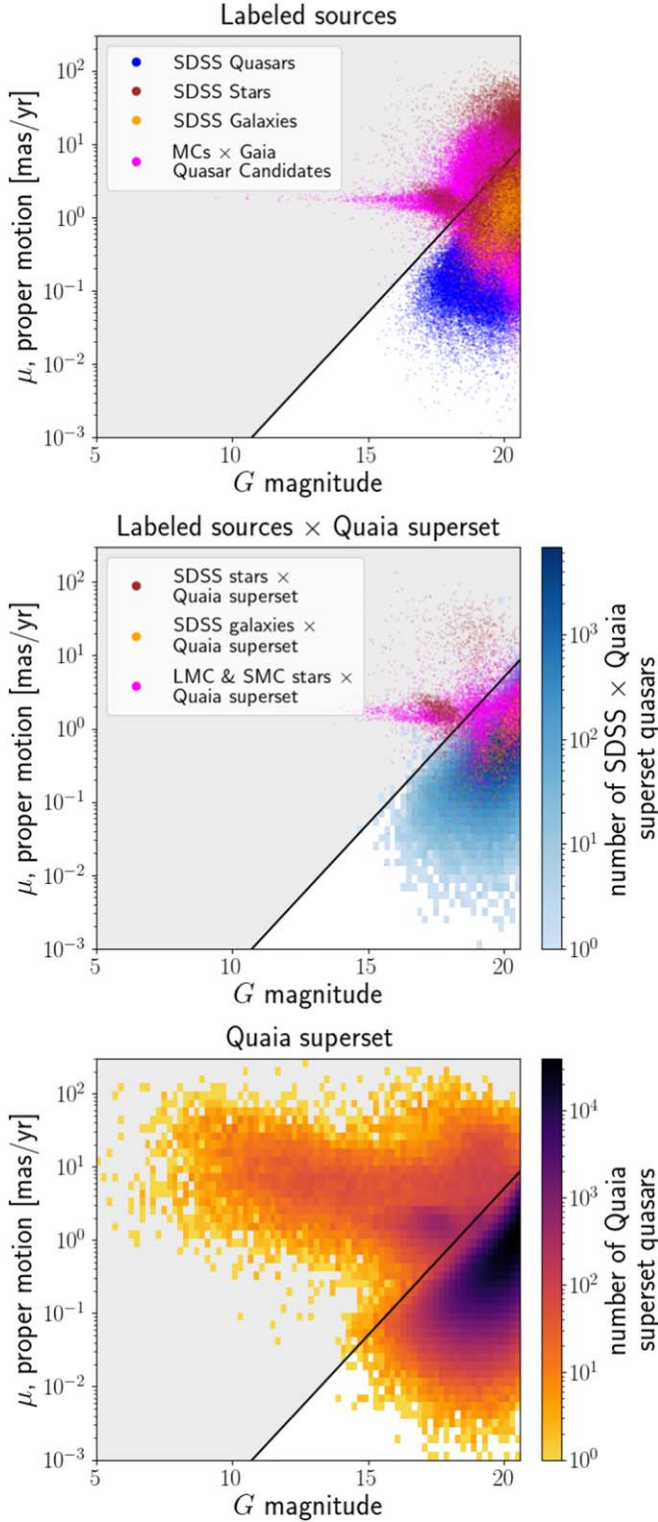
$$\mu < 10^{0.4(G-18.25)} \text{ mas yr}^{-1}. \quad (1)$$

At  $G = 18.25$ , this corresponds to  $\mu \lesssim 2.5 \text{ mas yr}^{-1}$ , and allows for less severe cuts at deeper magnitudes given the typically less precise astrometry. This is related to the proper motion *uncertainty* as a function of  $G$ , which has been quantified by Gaia (Gaia Collaboration et al. 2021). We show this cut overlaid on the Quiaia superset in the lower panel of Figure 3; based on the labeled data, we can clearly pick out the populations. The proper motion cut excludes 39,470 sources, 2.6% of the superset.

Next, we determine the color cuts based on Gaia and unWISE photometry. Generally, stars and galaxies are dim in redder, IR wavelengths compared to AGN. For instance, the

<sup>9</sup> The SDSS DR16Q quasar catalog is publicly available at [https://www.sdss.org/dr16/algorithms/qso\\_catalog](https://www.sdss.org/dr16/algorithms/qso_catalog).

<sup>10</sup> SDSS DR18 data can be accessed at <https://skyserver.sdss.org/CasJobs/jobdetails>.



**Figure 3.** Proper motion  $\mu$  vs.  $G$  magnitude for two different sets of sources. The black line shows the cut we make; the shaded gray region is excluded from the catalog. Top: the sources for which we have labels (SDSS data as well as sources near the LMC and SMC in the Gaia quasar candidates sample) that are also in the Quaia superset (Gaia DR3 quasar candidates that have all necessary photometry, Gaia redshift estimates, and  $G < 20.6$ ). Middle: sources in the top row that are also in the Quaia superset. Bottom: the superset of quasar candidates from which Quaia is constructed. The proper motion cut includes nearly all SDSS quasars in the superset while excluding a large number of stars.

eBOSS quasar target selection (Myers et al. 2015) involved linear cuts in the optical-IR, involving the SDSS  $g$ ,  $r$ , and  $i$  bands and the WISE W1 and W2 bands.

In Figure 4, we show color-color distributions for the same samples as in Figure 3. The left panel shows  $W1 - W2$  versus  $G - W1$  color, and the right column shows  $G - RP$  versus  $BP - G$  color. The top row, with the full labeled samples, shows that different types of sources tend to be localized to different areas of this parameter space (we show only a subset of each type for clarity). In particular, the colors involving unWISE (left panel) separate out the source types relatively clearly, demonstrating the importance of the unWISE crossmatch: SDSS quasars have very red  $W1 - W2$ , and intermediate  $G - W1$  color, while galaxies have bluer  $W1 - W2$  and redder  $G - W1$  compared to quasars, and stars (both SDSS stars and stars near the LMC and SMC) are bluer in both colors. In Gaia color-color space, galaxies tend to have bluer  $BP - G$  and redder  $G - RP$  colors than the other types of sources. In the middle row of Figure 4, showing the intersection of the labeled sources with the Quaia superset, we see that the superset restrictions have eliminated many of the sources, especially SDSS galaxies and stars, though a significant number remain. (We note that it is possible that some of these SDSS galaxies do host AGN though they were not classified as such by SDSS.) The Quaia superset is shown in the bottom panel; we can see clear populations of quasars, stars, and galaxies lining up with the labeled sources. Importantly, we can see the effect of the stricter SDSS color selection in the red (high  $G - W1$ ) region of parameter space into which the Gaia quasar candidates extend, but are not represented in the SDSS sample in the above panels.

We choose to apply linear cuts in these colors to decontaminate the sample. While other works (e.g., Hughes et al. 2022) train classifiers to determine which objects are true quasars using SDSS-classified quasars as labels, we opt for simpler cuts for ease of reproducibility and to mitigate the propagation of SDSS selection effects, which may include color- and magnitude-dependent effects. We choose four cuts based on the distribution of sources in color-color space. The first is in  $W1 - W2$ , which has been shown to be useful for distinguishing quasars; for instance, Nikutta et al. (2014) demonstrated that a small crossmatched SDSS quasar sample has very red  $W1 - W2 = 1.2 \pm 0.16$ , while other types of objects—namely star-forming and AGN galaxies, luminous red galaxies and stars—have bluer  $W1 - W2$ . Stars tend to have the bluest  $W1 - W2$ , with a mean of  $W1 - W2 = -0.04 \pm 0.03$ , so a cut in  $W1 - W2$  is a reliable way to filter out stellar contaminants. We add a cut in  $G - W1$  to filter out the bulk of the stars (including the LMC and SMC), and another in  $BP - G$  to cut out the galaxy contaminants. Finally, we find that these single color cuts were not sufficient to remove all of the LMC and SMC, so we add an additional diagonal cut in  $W1 - W2$  and  $G - W1$ , choosing a reasonable slope.

We optimize the values (intercepts) of these four cuts with a grid search, trying values spaced out by 0.1 mag. We note that while we show the full samples in Figure 4, in practice we make the proper motion cut before optimizing the color cuts. We choose the color cuts that maximize our objective function  $\mathcal{L}$ ,

$$\mathcal{L} = N_q - \lambda_s N_s - \lambda_g N_g - \lambda_m N_m, \quad (2)$$

where  $N_q$  is the number of true quasars that make it into the catalog,  $N_s$  SDSS stars,  $N_g$  SDSS galaxies, and  $N_m$  LMC and

SMC stars, and the  $\lambda$  parameters balance the relative ratios of each. We choose  $\lambda_s = 3$ ,  $\lambda_m = 5$ , and  $\lambda_g = 1$ .

The optimal cuts for the objects to keep in the catalog are

$$\begin{aligned} (G - W1) &> 2.15 \\ (W1 - W2) &> 0.4 \\ (BP - G) &> -0.3 \\ (G - W1) + 1.2(W1 - W2) &> 3.4. \end{aligned} \quad (3)$$

These are shown as the black lines in all panels of Figure 4, with the gray shading indicating exclusion regions. These cuts, as well as the proper motion cuts described above, exclude  $\sim 7\%$  of the superset, resulting in 1,414,385 quasars in our *decontaminated* sample. We apply an additional magnitude cut of  $G < 20.5$  to reduce edge effects in our redshift estimation; this constitutes our deep sample, with 1,295,502 sources. We refer to this as Quaia in the rest of this work. However, the catalog becomes less clean and reliable as we push to deeper magnitudes—due to less precise measurements and stronger systematics, notably the Gaia scanning pattern—so we produce a version of the catalog with  $G < 20.0$  to ensure a cleaner sample. This brighter catalog has 755,850 sources, and we report most of our results on this sample throughout the rest of this work.

### 3.2. Spectrophotometric Redshifts with unWISE and SDSS

We use unWISE and SDSS data to improve the redshift estimation of the sources. Figure 5(a) shows the redshifts estimated by the Gaia QSOC pipeline  $z_{\text{Gaia}}$  compared to the SDSS redshifts  $z_{\text{SDSS}}$  for a test sample of sources from Quaia with  $G < 20.5$ ; note that the 2D histogram is plotted in log-space to show the outliers more clearly. We find that of the Gaia redshifts  $z_{\text{Gaia}}$ , 82% (81%) agree to  $|\Delta z/(1+z)| < 0.2$  (0.1). A significant fraction of  $z_{\text{Gaia}}$  are highly precise: 75% agree with SDSS to  $|\Delta z/(1+z)| < 0.01$ . We also clearly see bands of incorrect estimation due to line aliasing issues. Additionally, in the crossmatched sample, nearly all of the very high  $z_{\text{Gaia}}$  estimates ( $z > 4.5$ ) are shown to be incorrect in comparison to SDSS. We note that the redshift estimation is much more accurate for sources that have no redshift warning flags set (`flags_qsoc=0`), as discussed in Section 2.1, but this is only true for 21% of the sources in Quaia ( $G < 20.5$ ), and even including sources with `flags_qsoc = 16` this leaves only 39% of sources.

We train a  $k$ NN model on Quaia sources to estimate improved redshifts. (We also tested other models including XGBoost and a multilayer perceptron, and found that the  $k$ NN outperformed both by a small margin.) We include all sources in our decontaminated catalog (Section 3.1), which goes out to  $G < 20.6$ , in order to have a buffer beyond our desired  $G < 20.5$  sample to reduce edge effects from the training set. The features that we train on are the Gaia redshift  $z_{\text{Gaia}}$ , colors constructed using Gaia and WISE photometry ( $G - \text{RP}$ ,  $\text{BP} - G$ ,  $\text{BP} - \text{RP}$ ,  $G - \text{W1}$ ,  $\text{W1} - \text{W2}$ ), the Gaia  $G$ -band magnitude, and the dust reddening  $E(B - V)$  at the location of the source. (We find that including the rest of the photometry does not make a difference in the results.) The reddening is determined with the Corrected Schlegel, Finkbeiner, & Davis dust map introduced by Chiang (2023), which corrects the standard Schlegel et al. (1998) dust map by subtracting off the contribution from the cosmic infrared background (CIB). (We also include the appropriate correction factor given by

Schlafly & Finkbeiner (2011).)<sup>11</sup> The labels are the SDSS redshifts,  $z_{\text{SDSS}}$ .

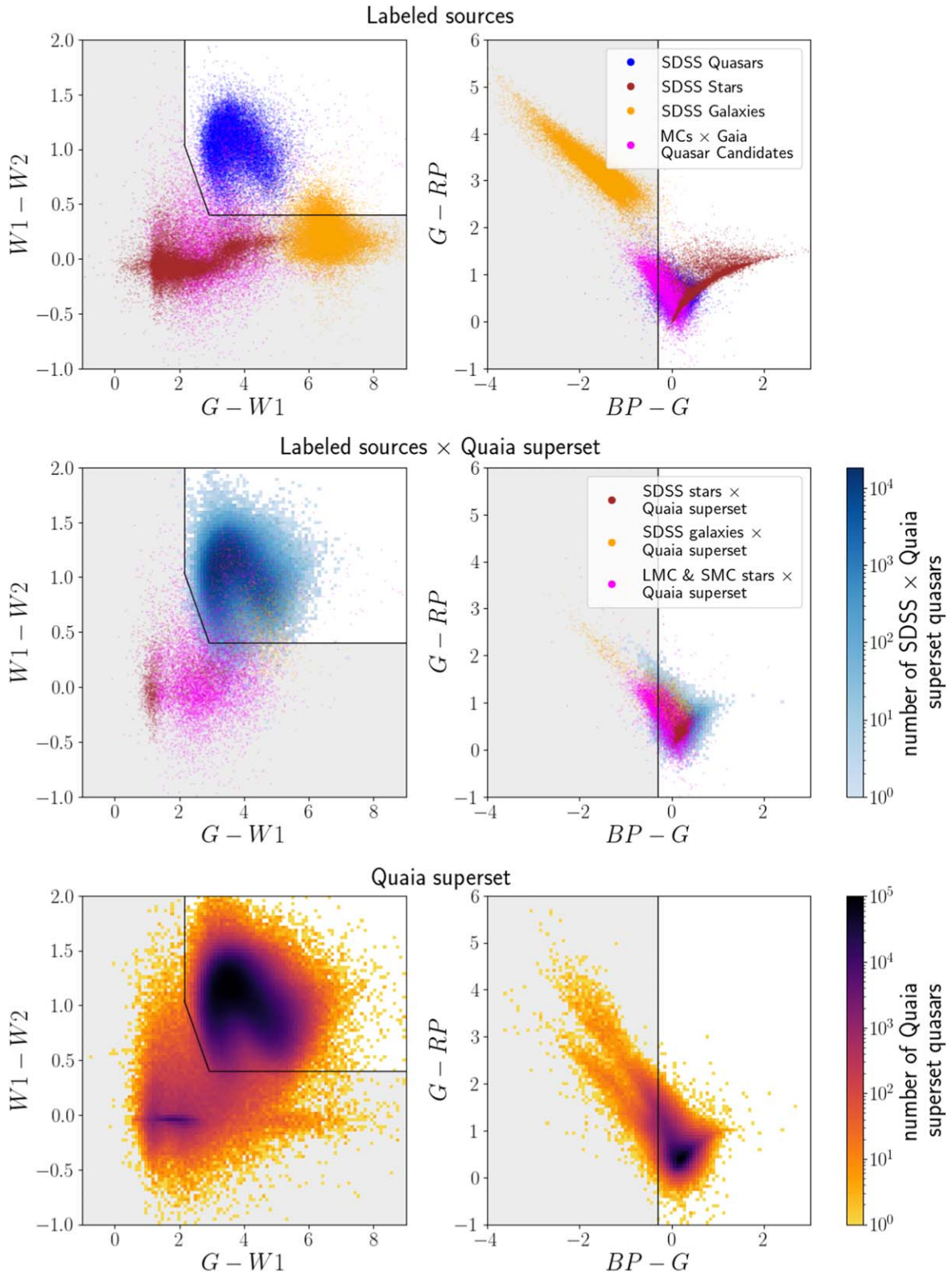
We use as our labeled data sources from the crossmatched SDSS DR16Q sample (Section 2.3) that are also in our decontaminated catalog Quaia, so that we train on sources drawn from the same distribution to which we will apply the model; this is 243,206 sources. We apply a 70%/15%/15% train/validation/test split. We build a  $k$ -d tree on the training set features using the `KDTree` implementation of `sklearn`. At the prediction stage, we access the  $K$  nearest neighbors of each input feature vector, first excluding neighbors with zero distance in feature space (i.e., neighbors that are in the training set). We assign the predicted label to be the median  $z_{\text{SDSS}}$  of the  $K$  nearest neighbors and the uncertainty to be the symmetrized inner 68% error of those neighbors. We use the validation set to tune  $K$ , and choose the value that maximizes the fraction of predicted redshifts with  $|\Delta z/(1+z)| < 0.1$ , which is  $K = 27$ ; we note that this value only varies at the  $\sim 1\%$  level for values  $15 < K < 50$ , and is similar for other choices of  $|\Delta z/(1+z)|$ . Finally, we apply the model to the full Quaia and output  $k$ NN redshift estimates,  $z_{k\text{NN}}$ , for each source.

The results are shown in Figure 6, which shows the cumulative distribution of errors  $|\Delta z/(1+z)|$  for  $z_{k\text{NN}}$  compared to that of  $z_{\text{Gaia}}$  (with  $z_{\text{SDSS}}$  as the truth) for the test set with  $G < 20.0$ . (The shapes are similar for  $G < 20.5$ , just shifted to somewhat lower accuracy.) We find that the  $z_{k\text{NN}}$  estimates have far fewer outliers than  $z_{\text{Gaia}}$ . However, the  $z_{\text{Gaia}}$  estimates tend to be more precise, as they use the full spectral information, while the  $k$ NN is essentially smoothing over the likeliest neighboring sources in feature space. We thus choose to combine the properties of both of these redshift estimates to obtain our final *spectrophotometric* (SPZ) redshifts  $z_{\text{Quaia}}$  in the following way. For sources for which  $z_{\text{Quaia}}$  and  $z_{\text{Gaia}}$  agree to  $|\Delta z/(1+z)| < 0.05$ , we assign  $z_{\text{Quaia}} = z_{\text{Gaia}}$  to preserve the precision of the Gaia estimate. For sources for which  $z_{\text{Quaia}}$  and  $z_{\text{Gaia}}$  differ by  $|\Delta z/(1+z)| > 0.1$ , we assign  $z_{\text{Quaia}} = z_{k\text{NN}}$  to preserve accuracy. In between these thresholds, we apply a smooth, linear transition to avoid hard features in our estimates. These  $z_{\text{Quaia}}$  estimates are also shown in Figure 6 compared to the true (spectroscopic, taken as truth for our purposes) SDSS redshifts, and we can see that these achieve nearly as high precision as  $z_{\text{Gaia}}$  while maintaining the high accuracy of  $z_{k\text{NN}}$ .

Our  $z_{\text{Quaia}}$  results for the test set are shown in Figure 5(b) compared to  $z_{\text{SDSS}}$ , shown here for the full catalog depth  $G < 20.5$ . We find that 91% (84%) of our SPZ redshifts agree to  $|\Delta z/(1+z)| < 0.2$  (0.1), and 62% highly agree to  $|\Delta z/(1+z)| < 0.01$ . We also give the bias (mean redshift error) and scatter ( $\sigma_{68}$ , the symmetrized inner 68% region of the redshift errors) of  $|\Delta z/(1+z)|$  in the figure; our SPZ redshifts significantly decrease the bias and scatter. The SPZ estimation corrected all of the very high- $z$  Gaia estimates, and some of the intermediate-outlying aliasing effects. We still have some catastrophic outliers due to line aliasing, but with our SPZ redshifts, we find a reduction in the number of  $|\Delta z/(1+z)| > 0.2$  (0.1) outliers by approximately three times (approximately two times) compared to the Gaia redshift estimates.

We investigate the dependence of the redshift error on the  $G$ -band magnitude in Figure 7. The fraction of redshifts with an error above various thresholds is shown as a function of

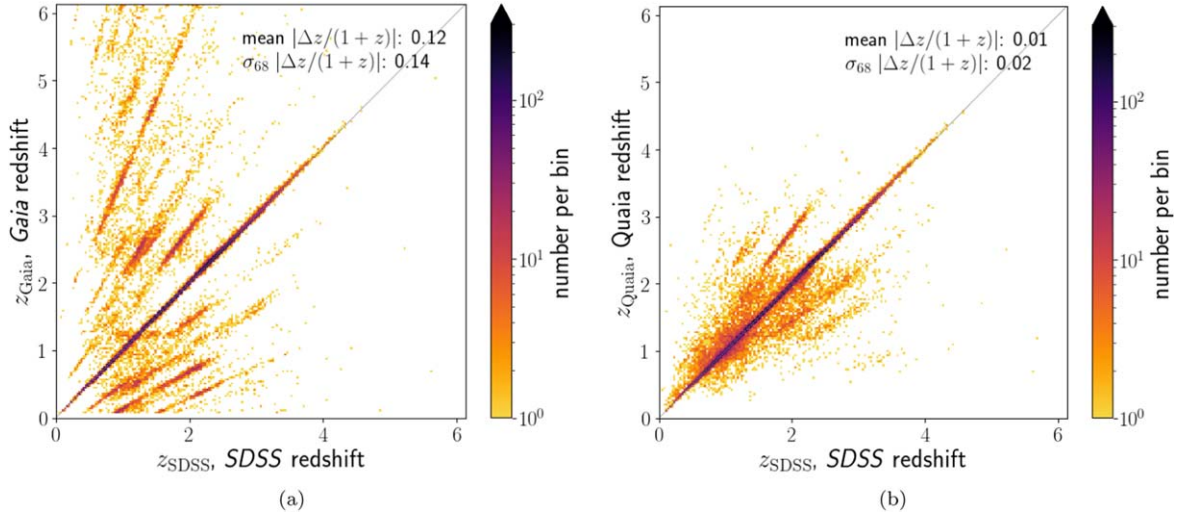
<sup>11</sup> The dust map was accessed with the Python package <https://dustmaps.readthedocs.io>.



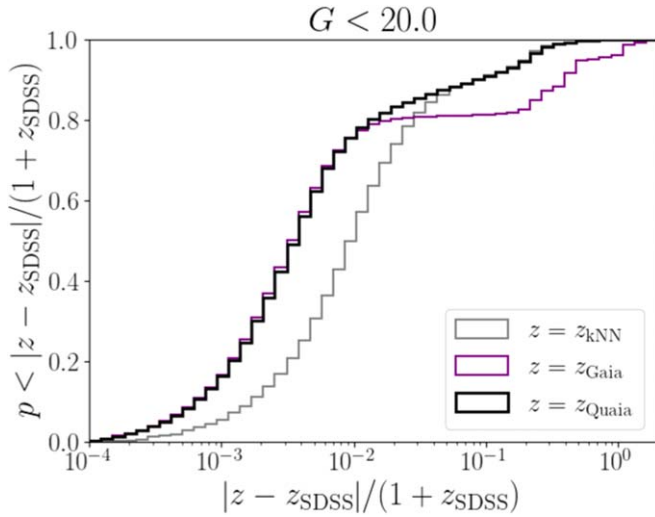
**Figure 4.** Color-color plots of three different sets of sources. The left column shows  $W1 - W2$  vs.  $G - W1$  color, and the right column shows  $G - RP$  vs.  $BP - G$  color. The black lines show the cuts we make; the shaded gray region is excluded from the catalog. The rows have the same samples as in Figure 3, except that in the top row, only 20,000 of each type of SDSS source is shown for clarity. In both color-color projections, the labeled sources are mostly localized in particular regions of parameter space, and we can see these populations somewhat clearly in the Quaia superset.

samples with the given cut on  $G$ . The errors are lowest at a bright magnitude cut of  $G < \sim 19.0$ ; in this sample, sources with SPZ redshift estimates inaccurate to  $|\Delta z / (1 + z)| > 0.2$  (0.1) comprise only 3% (4%) of the sample, and to the more

stringent requirement of  $|\Delta z / (1 + z)| > 0.01$ , 12%. This outlier fraction increases steadily as fainter sources are included. For  $G < 20.0$ , 6% (10%) are inaccurate to  $|\Delta z / (1 + z)| > 0.2$  (0.1), and 25% for  $|\Delta z / (1 + z)| > 0.01$ . Compared to the Gaia



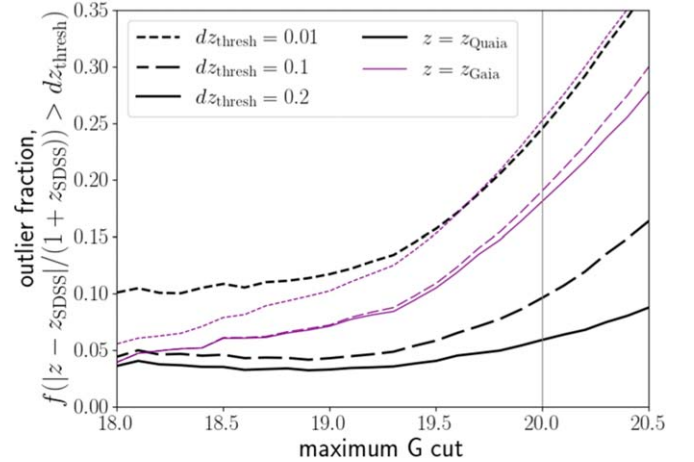
**Figure 5.** (a) Gaia redshift estimate  $z_{\text{Gaia}}$  vs. SDSS (true) redshift  $z_{\text{SDSS}}$  for a test set of sources in our quasar catalog Quiaia with  $G < 20.5$ . (b) Our estimated SPZ redshifts  $z_{\text{Quiaia}}$ , which are based on a  $k\text{NN}$  model, vs.  $z_{\text{SDSS}}$  for the same sample. The bias (mean redshift error) and scatter ( $\sigma_{68}$ , the symmetrized inner 68% region) of the redshift estimates compared to  $z_{\text{SDSS}}$  are shown in the panels. The  $z_{\text{Quiaia}}$  redshifts significantly decrease both the bias and scatter, as well as catastrophic outliers and unreasonably high-redshift estimates. The one-to-one line (perfect accuracy) is shown in gray; note that the color bar is on a log scale, and that a majority of the sources in both cases lie along this line.



**Figure 6.** The cumulative distribution of redshift errors for Quiaia test set sources with  $G < 20.0$ , considering SDSS spectroscopic redshifts  $z_{\text{SDSS}}$  as the ground truth, for estimates directly from our  $k\text{NN}$  model (gray), the original  $z_{\text{Gaia}}$  redshifts (purple), and our final  $z_{\text{Quiaia}}$  estimates (black) based on a combination of the other two. Our SPZ redshifts have far fewer outliers and similar precision compared to the Gaia estimates.

redshift estimates, the SPZ estimates  $z_{\text{Quiaia}}$  reduce the number of  $|\Delta z/(1+z)| > 0.2$  (0.1) outliers by approximately three times (approximately two times). The choice of  $G$  cut to use in a given analysis will depend on the nature of the analysis and its sensitivity to outliers.

We note that our finding that the unWISE IR information significantly improves redshift estimates, compared to only the optical information used in the Gaia QSOC estimates, is consistent with other photometric redshift work. For instance, DiPompeo et al. (2015) showed that including WISE mid-IR photometry in the redshift estimation of SDSS-imaged quasars results in a significant improvement in the estimates, even more so than including both GALEX near- and far-UV data and UKIDSS near-IR data. More recently, Yang & Shen (2023)



**Figure 7.** The fraction of outlying redshifts with  $|\Delta z/(1+z)| > (0.01, 0.1, 0.2)$ , as a function of  $G$  magnitude, for our redshift estimation test set. The SPZ redshifts are shown in black, and the Gaia redshifts in purple. The fraction of outliers increases steeply with increasing  $G$  for  $G > 19.5$  for both  $z_{\text{Quiaia}}$  and  $z_{\text{Gaia}}$ , though the fraction of catastrophic outliers for  $z_{\text{Quiaia}}$  is significantly lower (and the dependence less steep) compared to  $z_{\text{Gaia}}$ .

compiled a photometric quasar catalog from the Dark Energy Survey (DES) DR2, combining DES optical photometry with near-IR photometry as well as unWISE mid-IR photometry; they obtained photo- $z$ s with 92% having  $|\Delta z/(1+z)| < 0.1$  when all IR bands are used compared to 72% with only optical data. Additional photometric information at other wavelengths could further improve our estimates (as well as catalog decontamination), but is not currently available for enough sources in our Quiaia catalog to be worthwhile. For instance, for the UV all-sky survey GALEX (Martin et al. 2005), cross-matches to Quiaia sources are only available for 32% of the Quiaia objects for near-UV observations, and when including far-UV only 16%; this significant discrepancy is largely due to the faint end of Quiaia, where GALEX observations do not reach deep enough. The Pan-STARRS1 survey (Chambers et al. 2016) covers only three-quarters of the sky, with



crossmatches to 75% of Quiaia sources. We tested adding Pan-STARRS1 data to the redshift estimation feature set and found only a small improvement, and thus chose to prioritize keeping the full sky span of Quiaia, though we make note that incorporating Pan-STARRS1 may be useful for certain applications.

### 3.3. Selection Function Modeling

Observational and astrophysical effects impact which sources we observe and their properties; this is known as the selection function. As Gaia is a space-based mission, it avoids many of the observational issues of ground-based surveys, such as seeing and airmass. However, there are still significant selection effects: for our model, we consider dust, the source density of the parent surveys, and the scan patterns of the parent surveys.

We fit a selection function model to a particular version of the catalog, namely, a particular maximum  $G$ . For the fiducial selection function we work only in terms of sky position. We make a healpix map of the catalog with NSIDE = 64 and count the number of observed catalog sources in each healpix pixel. We choose this NSIDE, which results in 49,152 pixels each with an area of  $\sim 0.84 \text{ deg}^2$ , to balance constructing a map with reasonably high resolution with ensuring a sufficient number of sources in the pixels for stable fits, as well as fitting within memory limitations for the Gaussian process fit. In the case of no selection effects (and under the assumption of isotropy), we would expect each pixel to contain roughly the same number of sources. Our goal is to model the dependence between the number of sources per pixel and the various systematics.

The systematics maps (templates) we use are shown in Figure 8. We use the dust map of Chiang (2023), and convert it to a healpix map of NSIDE = 64. To do this, we evaluate the reddening  $E(B - V)$  at the centers of pixels of a high-resolution NSIDE = 2048 healpixelization of the sphere, and apply the 0.86 correction factor proposed by Schlafly & Finkbeiner (2011). We convert these to extinction values by multiplying by  $R_V = 3.1$ , and then take the mean of all of these values within each healpixel target NSIDE = 64 map. This produces a smoothed dust extinction map on the desired scale. The result is shown in Figure 8(a); the extinction is highest around the Galactic plane, with structure extending outward.

For the stellar distribution, we randomly select  $\sim 10.6$  million Gaia sources with  $18.5 < G < 20$ , the magnitude range of most of our quasar sample. The vast majority of these will be true stars. (While this sample will contain some other types of objects, including possibly some quasars and other extragalactic sources, these will be orders of magnitude less numerous than stars.) We count the number of stars per NSIDE = 64 healpixel; this is shown in Figure 8(b). We also include a template of the unWISE source distribution, for which we randomly selected  $\sim 10.6$  million unWISE sources (1% of the catalog) that have flux in both W1 and W2, and have primary status ( $\text{Prim} = 1$ ). We count the number of these sources per NSIDE = 64 healpixel, as shown in Figure 8(c).

In initial fits we found that the regions of the LMC and SMC are particularly poorly modeled, and that the fit is improved by including separate templates of just the LMC and SMC source density for both the Gaia and unWISE sources; this gives the model the freedom to assign different coefficients to these regions than to the overall survey source density. (The need for different coefficients could be for a number of reasons, such as

a difference in stellar density, contamination, or magnitude distribution; we leave a deeper investigation of this to future work and just use this empirical finding to improve our model.) For the LMC/SMC templates, we cut out a wide region around the LMC and SMC ( $9^\circ$  in radius around the LMC and  $5^\circ$  around the SMC), and subtract the background, which we approximate using the region at the same latitude but opposite longitude (mirrored across the  $l = 0^\circ$  line) of the given source distribution map. We do not show these maps here as they are visually similar to the stellar and unWISE source density maps in the LMC and SMC regions (though with the background subtracted).

For the Gaia completeness, we use the quantity  $M_{10}$  introduced by Cantat-Gaudin et al. (2023).<sup>12</sup>  $M_{10}$  is the median magnitude in a given sky patch of the Gaia sources with  $\leq 10$  transits across the Gaia field of view; it incorporates the effects of both the scanning law and source crowding. The actual completeness map derived by Cantat-Gaudin et al. (2023) depends on both  $M_{10}$  and  $G$ -band magnitude; this completeness is very close to 1 for nearly all of the sky for  $G = 20.0$ , with some non-negligible incompleteness for  $G = 20.5$ . However, this completeness model is based on the full Gaia source catalog, while we expect the selection function of our quasar sample to be different. We thus use the  $M_{10}$  map directly in our fit to capture the effects of the Gaia scanning law and source crowding specific to Quiaia. We downsample the map to NSIDE = 64; this is shown in Figure 8(c).

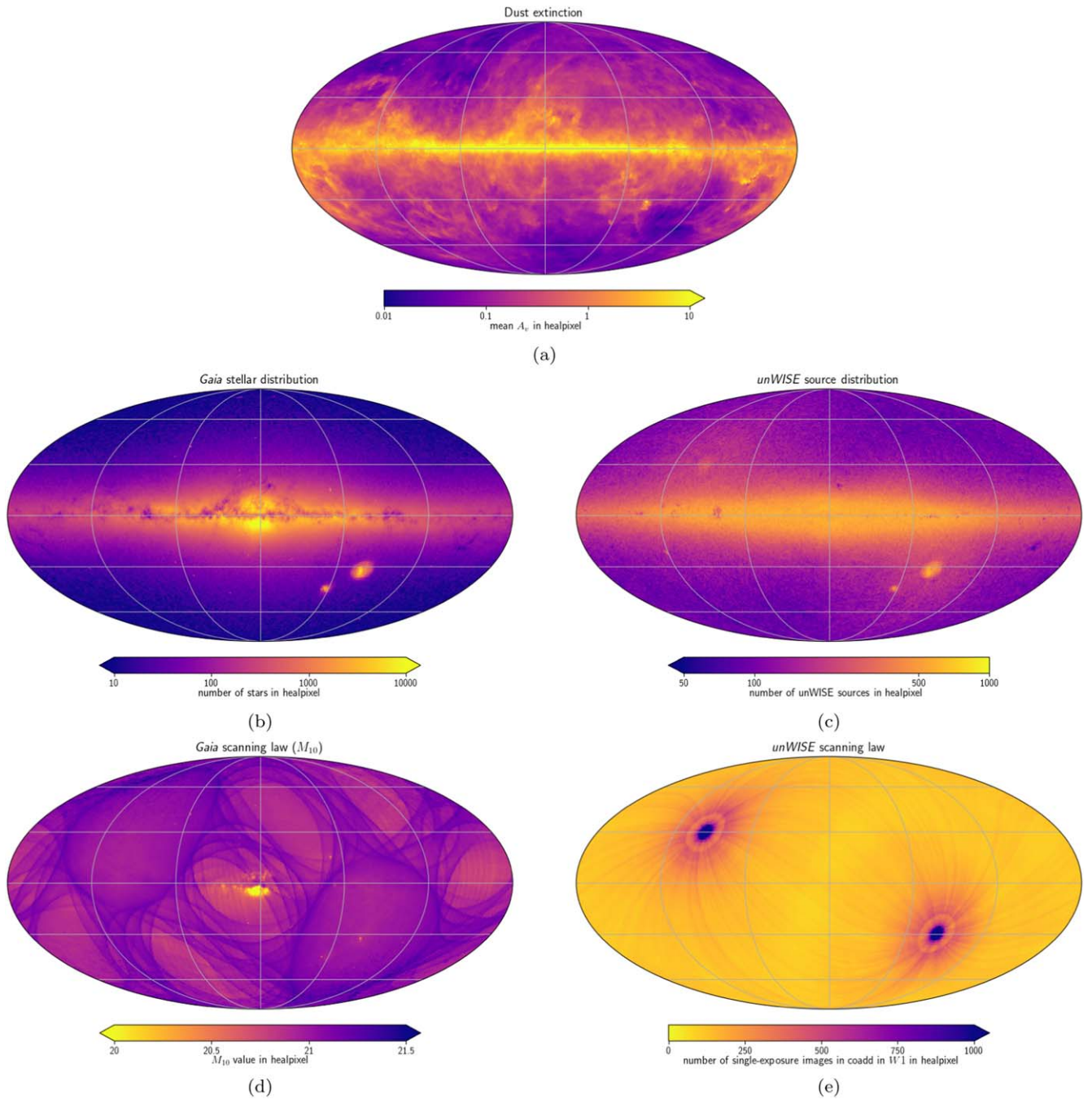
For the unWISE scanning law, using the  $\sim 10.6$  million unWISE sources described above, we take the mean number of single-exposure images in the coadd in W1 for the sources in each NSIDE = 64 healpixel. This is shown in Figure 8(e); we can see that the scan is in strips of constant ecliptic latitude, and that there is a significant increase in observations at the ecliptic poles.

To model the selection function we use a Gaussian process, a flexible machine-learning method for regression; for a detailed treatment, see Rasmussen & Williams (2005). (We first tried a linear model and found that it gave a very poor fit, because there are significant nonlinearities between the systematics and the catalog number density.) We first scale the data: for the labels (number of Quiaia sources per pixel) we work in their logarithm, and only fit for the pixels with a nonzero number of sources. For the Gaia stellar distribution, the unWISE source distribution, the unWISE scan pattern, and LMC/SMC map templates, we also take the log of the number of quasars per pixel; for the LMC/SMC map, we first replace zeros with a very small value. For all of the input feature maps, we take the mean-subtracted systematics values. We assume a Poisson error on the labels (and apply the appropriate log transformation). For the Gaussian process, we use the `george` software package (Ambikasaran et al. 2016). We use an exponential squared kernel  $k$  of the form

$$k(r^2) = \exp\left(\frac{-r^2}{2}\right), \quad (4)$$

where  $r$  is the distance between points in feature space. We train the Gaussian process on all of the data, optimizing the parameter vector using the BFGS solver (Fletcher 1987); this includes fitting for the mean of the labels. We finally evaluate

<sup>12</sup> This map can be accessed with the `gaiaunlimited` package, <https://gaiaunlimited.readthedocs.io>.



**Figure 8.** The systematics maps used in the selection function model: (a) dust extinction from Chiang (2023); (b) the stellar distribution based on  $\sim 10.6$  million randomly selected Gaia sources with  $18.5 < G < 20$ ; (c) the unWISE source distribution based on  $\sim 10.6$  million randomly selected unWISE sources; (d) the quantity  $M_{10}$ , the median magnitude of sources with  $\leq 10$  Gaia transits, which encodes the Gaia scanning law and source crowding; and (e) the unWISE scan pattern given by the mean number of single-exposure images of the sky region in the coadd. Note that the color bar on the  $M_{10}$  and unWISE scanning law maps are reversed, as high values indicate a cleaner region, the inverse of the other maps. We also include separate templates for only sources in the LMC and SMC regions for both the stellar and unWISE source densities, with the background subtracted. All templates are discussed in more detail in the text.

the predicted number of sources in each pixel. Where there were no Quia sources in the label map, we fix the prediction to zero.

To convert this to a selection function in terms of the relative completeness, we first identify *clean* pixels in the map having low dust extinction ( $A_V < 0.03$  mag), low star counts ( $N_{\text{stars}} < 15$ ), low unWISE source counts ( $< 150$ ), no stars or unWISE sources in the LMC or SMC, and high  $M_{10}$  ( $M_{10} > 21$  mag) and unWISE coadds ( $> 150$ ); this results in 479 pixels. We take the mean predicted number of quasars in these clean

pixels, and add two times the standard deviation in these pixels to encompass the scatter. We then normalize the predicted source numbers by this value, which ensures that all final values end up being less than 1 for clarity. The result is a selection function map in terms of the *relative* probability of a source at a given location being included in the catalog. We emphasize that this is relative; we have not normalized it to an absolute probability so as not to make the selection function map extremely sensitive to the maximum value. We also note that this fit must be redone for each version of the catalog

because it depends on the particular number density and distribution of sources.

There will be a dependence of the selection function on the  $G$ -band magnitude, as well as other quantities such as redshift. While we do not include these in our modeling or fiducial selection function map, we do release selection functions for a redshift split version of the catalog, using two redshift bins, which is important for certain cosmological analyses. The code to generate the selection function for any input catalog is also provided so that users can construct maps that meet their needs.

## 4. Catalog: Results and Verification

### 4.1. Properties of the Catalog

Quaia, the Gaia-unWISE Quasar Catalog, consists of 755,850 (1,295,502) quasar candidates with  $G < 20.0$  ( $20.5$ ). The sky distribution of Quaia for each of these magnitude limits is shown in Figure 9. The catalog covers the full sky, besides the Galactic plane, including the southern sky—most of which is not well covered by other surveys (discussed further in Section 4.3). The sky distribution is remarkably uniform, and the nonuniform imprints visually follow the selection effects that we incorporated into our selection function map, most notably the dust distribution (Figure 8(a)). Quaia also does not show an obvious overdensity around the LMC and SMC (as the Gaia DR3 *purser* sample does) because we have removed these with our decontamination procedure. In fact, there is now a slight underdensity of sources near the LMC; this makes sense because some quasars in that sky region are obscured by dust and confusion in the LMC, though it is possible we have also somewhat overcorrected for this and removed some true quasars.

The dearth of quasars in the Galactic plane is due largely to dust extinction and stellar crowding, as well as the fact that the SDSS training set quasars (for both the original Gaia DR3 quasar candidates sample and our decontamination procedure) are not representative of quasars in this dust-reddened region. If we exclude the regions with very high extinction  $A_V > 0.5$  mag, the quasars nearly uniformly cover the remaining sky area, which comprises  $30,277.52 \text{ deg}^2$  ( $f_{\text{sky}} = 0.73$ ). Based on this area we can also compute the effective volume  $V_{\text{eff}}$  covered by the quasars, which depends on the number density as a function of redshift and the power spectrum value  $P(k)$ , integrated over the physical volume. We assume a  $P(k)$  of  $4 \times 10^4 (h^{-1} \text{ Mpc})^3$ , based on the value for the eBOSS clustering catalog of quasars at around  $k \sim 0.01$  (Mueller et al. 2021). This gives an effective volume of  $7.67 (h^{-1} \text{ Gpc})^3$  ( $3.19 (h^{-1} \text{ Gpc})^3$ ) for the  $G < 20.5$  ( $G < 20.0$ ) sample.

We show a 3D map of the Quaia catalog in Figure 10, using our  $z_{\text{Quaia}}$  redshift estimates converted to spatial coordinates with a fiducial Planck cosmology. We also show a 3D map of the full SDSS quasar sample for comparison; Quaia spans a much larger volume than SDSS. We note that for SDSS large-scale structure analyses, the eBOSS quasar clustering catalog is used, which contains fewer sources than the full SDSS catalog as it spans only the intermediate (UV-excess) redshift range and is designed to be uniform across the sky (described in more detail in Section 4.3).

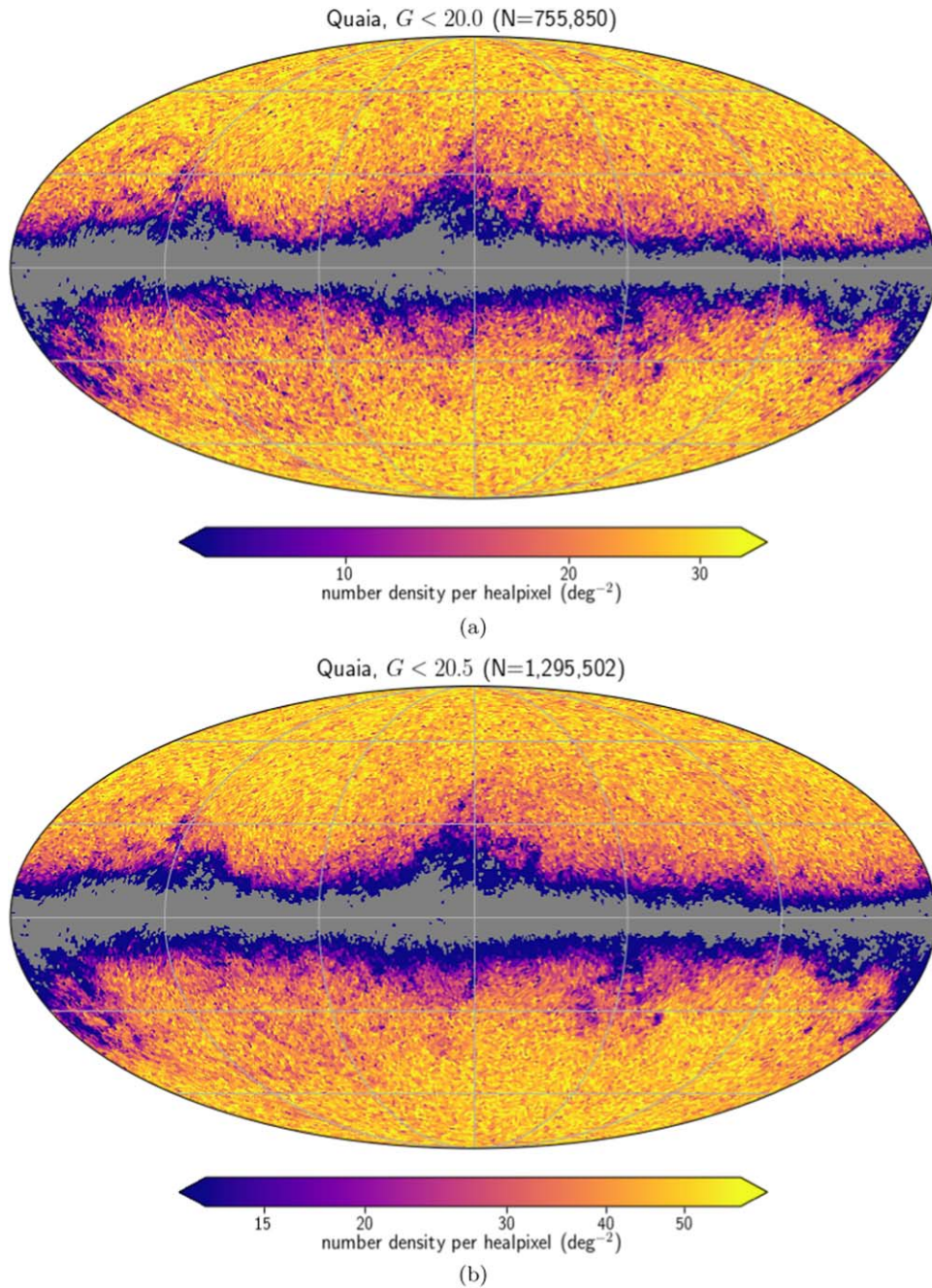
We show the redshift distribution of Quaia in Figure 11. The distribution of our Gaia-unWISE-SDSS spectrophotometric redshift estimates,  $z_{\text{Quaia}}$ , for the full  $G < 20.5$  catalog is shown in black. We compare this to other samples, cut to the same

$G$  limit where relevant: the Gaia redshifts  $z_{\text{Gaia}}$  for the same sample;  $z_{\text{Gaia}}$  for sources in the full Gaia quasar candidates sample with  $G < 20.5$  (that have redshift estimates);  $z_{\text{Gaia}}$  for sources in the Gaia DR3 *purser* sample with  $G < 20.5$  (that have redshift estimates); and  $z_{\text{SDSS}}$  for the SDSS DR16Q sources that have Gaia crossmatches, with  $G < 20.5$ . We see that the Quaia SPZ redshifts have a smoother distribution than the others, with a clear peak around  $z = 1.5$ ; the median value is 1.47. These SPZ estimates have also greatly reduced the high- $z$  tail present in the Gaia redshifts. There are still a significant amount of intermediate- $z$  objects; 10% ( $N = 132,417$ ) of the sources in the full  $G < 20.5$  Quaia catalog have  $z > 2.5$  (for the  $G < 20.0$  catalog, this is also 10% ( $N = 77,337$ ) of sources). We note that the  $z_{\text{Gaia}}$  redshift distribution for the Gaia DR3 *purser* sample is very similar to those same redshift estimates for Quaia; this is partially because a very high fraction of the objects in Quaia are also in the larger Gaia DR3 *purser* sample (see Figure 1).

We see a slight bump in the  $z_{\text{Quaia}}$  distribution around  $z \sim 2.3$ , the same location as the peak in the SDSS DR16Q quasar distribution. In the SDSS distribution, this feature is most prominent in the SDSS-III campaign quasars (see Figure 6 of Lyke et al. 2020), which targeted higher-redshift sources. To check the robustness of our redshift estimation, we reconstruct the sample and retrain the redshifts using the eBOSS quasar clustering catalog (Ross et al. 2020). This is the sample used for large-scale structure clustering analyses (e.g., Mueller et al. 2021; Rezaie et al. 2021), which has a smooth redshift distribution peaked around  $z = 1.5$ . It does still have a slight step around  $z \sim 2.3$ . We find that the  $z_{\text{Quaia}}$  redshift distribution does not change significantly when trained on this sample, and that the feature at  $z \sim 2.3$  remains. We hypothesize that this feature is thus a real feature of Gaia-selected quasars, rather than an imprint from the training set, likely related to details of the optical color selection around that redshift. We also find that compared to the full SDSS-trained sample, the sample trained on the eBOSS quasar clustering catalog produces a redshift distribution that is less smooth at low redshifts, possibly because of the lower number of low- $z$  eBOSS quasars; similarly, the high- $z$  tail is shorter. For these reasons, we choose to use the full SDSS sample (as described in Section 2.3) for the spectroscopic quasar training sample for our fiducial Quaia catalog, but confirm that the redshift distribution (and the source selection) is broadly robust to this choice.

We show the  $G$ -band magnitude distribution of Quaia in Figure 12, in comparison to the other Gaia and SDSS quasar samples described above. We see that our catalog (as well as the Gaia DR3 *purser* sample) has removed all of the sources with excessively bright (for quasars) magnitudes  $G < 12.5$  that are present in the full Gaia sample, as well as many sources with  $12.5 < G < 16$ . For the Gaia DR3 and SDSS samples, the number of quasars drops off sharply after  $G \sim 20.75$ ; to avoid the complicated selection effects at these depths, we limit our catalog to  $G < 20.5$  as shown. We also note that the SDSS DR16 quasars do not extend as bright as Quaia, and this extrapolation past the training set could bias the results in this regime, though in practice this affects very few sources.

We note that some of the Quaia sources may technically be considered lower-luminosity AGNs, or Seyfert-like galaxies, rather than quasars. We estimate the fraction of these sources using the criterion of Schneider et al. (2010): sources are



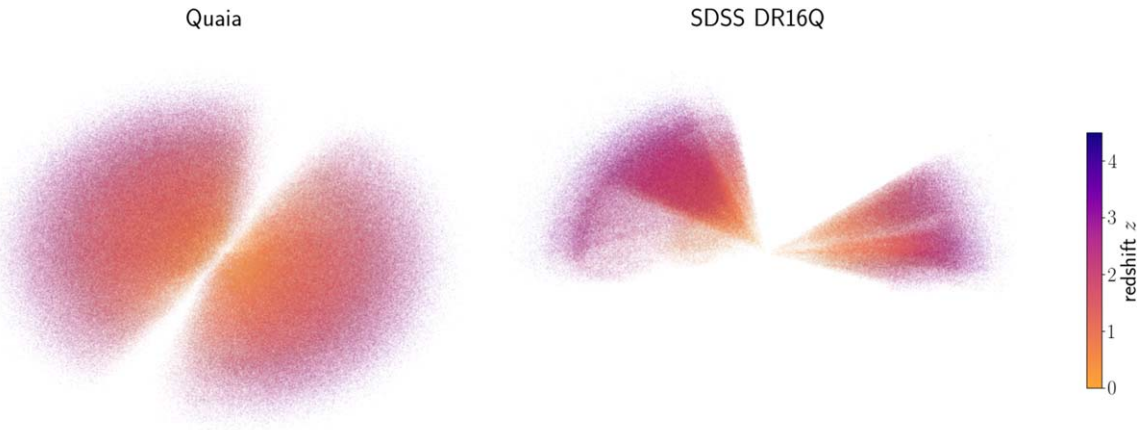
**Figure 9.** Sky distribution of the Quiaia quasar catalog, in Galactic coordinates and displayed using a Mollweide projection. Panel (a) shows sources with  $G < 20.0$ , the cleaner version with more reliable redshifts, and (b) shows the catalog down to its magnitude limit of  $G < 20.5$ .

considered true quasars if they have SDSS  $i$ -band luminosity  $M_i$  brighter than  $M_i = -22.0$ . To estimate the  $i$ -band magnitude for our Gaia sources, we compute the median  $G - i$  color for the subset of Quiaia sources with SDSS crossmatches, where  $G$  is the Gaia  $G$  band, and then subtract this value from the  $G$ -band magnitudes to obtain an effective  $i$ -band magnitude for all Quiaia sources. We convert these to absolute magnitudes  $M_i$  assuming a flat  $\Lambda$ CDM cosmology with  $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$ ,  $\Omega_m = 0.3$ , and  $\Omega_\Lambda = 0.7$ , following Schneider et al. (2010), and assuming a value of dust reddening of  $A_v/E(B - V) = 1.698$  corresponding to the SDSS  $i$  band and  $R_v = 3.1$ . We find that a small fraction, 8%, of Quiaia sources have effective  $M_i < -22.0$  and thus do not meet this standard luminosity criterion for being true quasars. This distinction may be important for certain studies, though

may not be relevant for others, and should be kept in mind for analyses of Quiaia.

#### 4.2. Selection Function Model

We show the results of our selection function modeling (Section 3.3) for the  $G < 20.0$  catalog in Figure 13. The selection function map is shown in Figure 13(a), where the values are the *relative* completeness; note that these should not be interpreted as a probability, and users may choose to normalize these values in different ways. The relationship of the selection function model to the dust and source density maps is clear visually. In Figure 13(b), we show the fractional residuals between a random catalog downsampled by this



**Figure 10.** Left: a projection of the 3D map of the full Quaia catalog ( $G < 20.5$ ). Right: the same projection for the quasars in SDSS DR16Q, the largest spectroscopic quasar catalog (note that it is a superset of SDSS quasars from multiple campaigns and as such is not intended to be uniform). The color bar shows the redshifts of the quasars ( $z_{\text{Quaia}}$  for Quaia,  $z_{\text{SDSS}}$  for SDSS), which have been converted to distances with a fiducial cosmology. Quaia spans a significantly larger volume than the SDSS sample.

(An animation of this figure is available.)

selection function and the true quasar catalog. The residuals generally look like homogeneous noise, indicating a good fit; the root mean squared fractional error is 0.49.

Around the edges of the Galactic plane the residuals show a slight bias to positive values (meaning the completeness there was predicted to be higher than it actually is); in the region around zero Galactic longitude just above the Galactic plane, the residuals are slightly biased to negative values (meaning the completeness there was predicted to be lower than it is). These discrepancies indicate that our templates are not fully capturing the selection effects in these regions. As these are largely limited to the region around the Galactic plane, the issue could be circumvented by applying a latitude cut for sensitive analyses. The underdensity around the LMC is well modeled by the selection function, with no clear residual in that region. The selection function map for the  $G < 20.5$  catalog (not shown) is similar with some moderate differences, and is also provided as a data product.

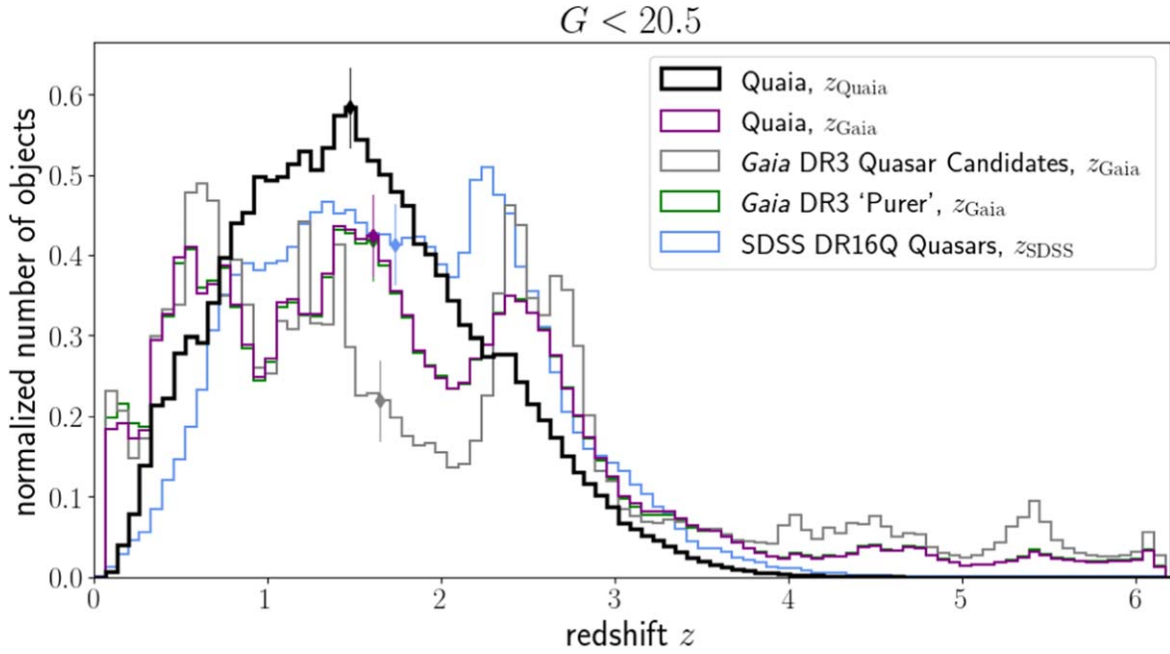
The selection function may change more significantly for different subsets of the catalog, such as redshift bins. The selection function should be refit for a given sample to be analyzed; we provide a code to fit the selection function for any other subset of the catalog. We note that depending on the subsample, certain regions may be more poorly modeled, and in particular, the regions around the LMC and SMC; users should check the residuals and may choose to mask the regions around the LMC and SMC to be more conservative.

#### 4.3. Comparison to Existing Quasar Catalogs

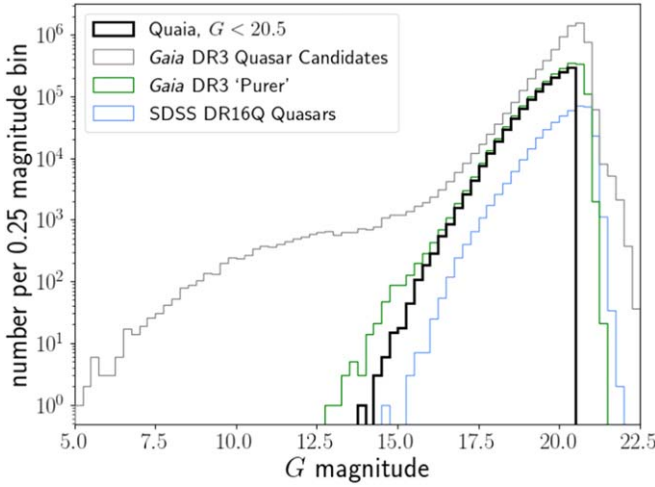
We compare Quaia to other existing quasar catalogs: Projections of these catalogs are shown in Figure 14. We show the Gaia DR3 *purser* sample (Figure 14(a)); a cross-matched catalog of WISE and Pan-STARRS (WISE-PS1), a current leading large-area photometric redshift quasar sample (Figure 14(b)); the SDSS DR16Q catalog, the current best spectroscopic sample of quasars (Figure 14(c)); the eBOSS quasar clustering catalog, the subsample of SDSS DR16Q intended for clustering analyses (Figure 14(d)); and Milliquas, a meta-catalog compiling confirmed quasars from the literature (Figure 14(e)).

The Gaia DR3 *purser* sample is described in Section 2.1; here we include only sources with QSOC redshift estimates ( $z_{\text{Gaia}}$ ). The WISE-PS1 sample was constructed by Beck et al. (2022), based on the Source Types and Redshifts with Machine learning (STRM) algorithm by Beck et al. (2020). The quasar catalog with updated photometric redshifts is presented by Kunsági-Máté et al. (2022); here we include only those quasars with redshifts labeled *reliable*, which is 59% of the sample. The SDSS DR16Q quasar catalog is the one described in Section 2.3, from Lyke et al. (2020), which compiles sources from eBOSS as well as previous SDSS campaigns (and is intended as a superset of SDSS quasars rather than a uniform sample). The eBOSS quasar clustering catalog is detailed in Ross et al. (2020); it is a subsample of SDSS DR16Q selected for large-scale structure clustering analyses, and as such is much more homogeneous than the full catalog. For the eBOSS clustering catalog, we have included both eBOSS and legacy SDSS quasars ( $\text{IMATCH} = 1$  or 2) and applied the clustering cuts of requiring sectors to have  $>0.5$  completeness ( $\text{COMP\_BOSS}$ ) and redshift success rate ( $\text{sector\_SSR}$ ); we have additionally removed sources with  $\text{ZWARNING!} = 0$ . The Milliquas catalog was compiled by Flesch (2021); a significant portion of the sources are from SDSS and AllWISE. For each of these samples, we have shown quasars brighter than a limiting magnitude of  $G \sim 20.5$ ; for the non-Gaia catalogs we convert to  $G$  from the survey's  $r$ -band magnitude using the conversion in Equation (2) of Proft & Wambsganss (2015), which is based on the SDSS  $r'$  band. While this should give a reasonable estimate for the SDSS sample (using  $r_{\text{SDSS}}$ ) and the WISE-PS1 sample (using  $r_{\text{PS1}}$  which is very similar to  $r_{\text{SDSS}}$ ), it may not be as reliable for Milliquas which catalogs *red* magnitudes from various sources, as well as for sources with  $z > 3$ , which were not included in the Proft & Wambsganss (2015) fit.

A summary of the catalogs is shown in Table 1, for the full catalogs (limited to sources with reliable redshifts) as well as the  $G_{\text{eff}} < 20.5$  subsamples. We exclude Milliquas from this comparison given its very heterogeneous nature; we do include SDSS DR16Q, though it is also not intended to be uniform, to show the comparison of Quaia to this large spectroscopic catalog of quasars. For these quantifications, we exclude areas

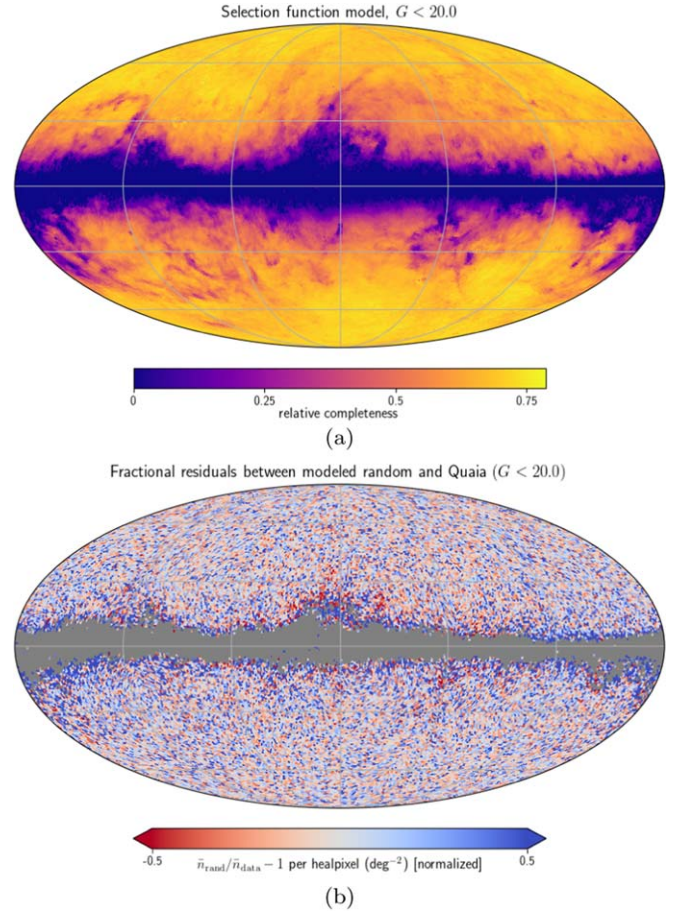


**Figure 11.** Redshift distribution of Quiaia for our spectrophotometric redshift estimates  $z_{\text{Quaia}}$  (black), normalized to the total number of objects. For comparison, we also show the normalized distributions of other samples, cut to the  $G < 20.5$  limiting magnitude of Quiaia where relevant: the Gaia redshift estimates  $z_{\text{Gaia}}$  for the same Quiaia sources (purple);  $z_{\text{Gaia}}$  for the sources in the full Gaia quasar candidate sample with  $G < 20.5$  (gray);  $z_{\text{Gaia}}$  for the Gaia DR3 *purer* subsample with  $G < 20.5$  (green); and the SDSS redshifts  $z_{\text{SDSS}}$  for the SDSS DR16Q quasar sample that have Gaia crossmatches, with  $G < 20.5$  (blue). The median redshift of each distribution is shown by the diamond and vertical line in the respective color.

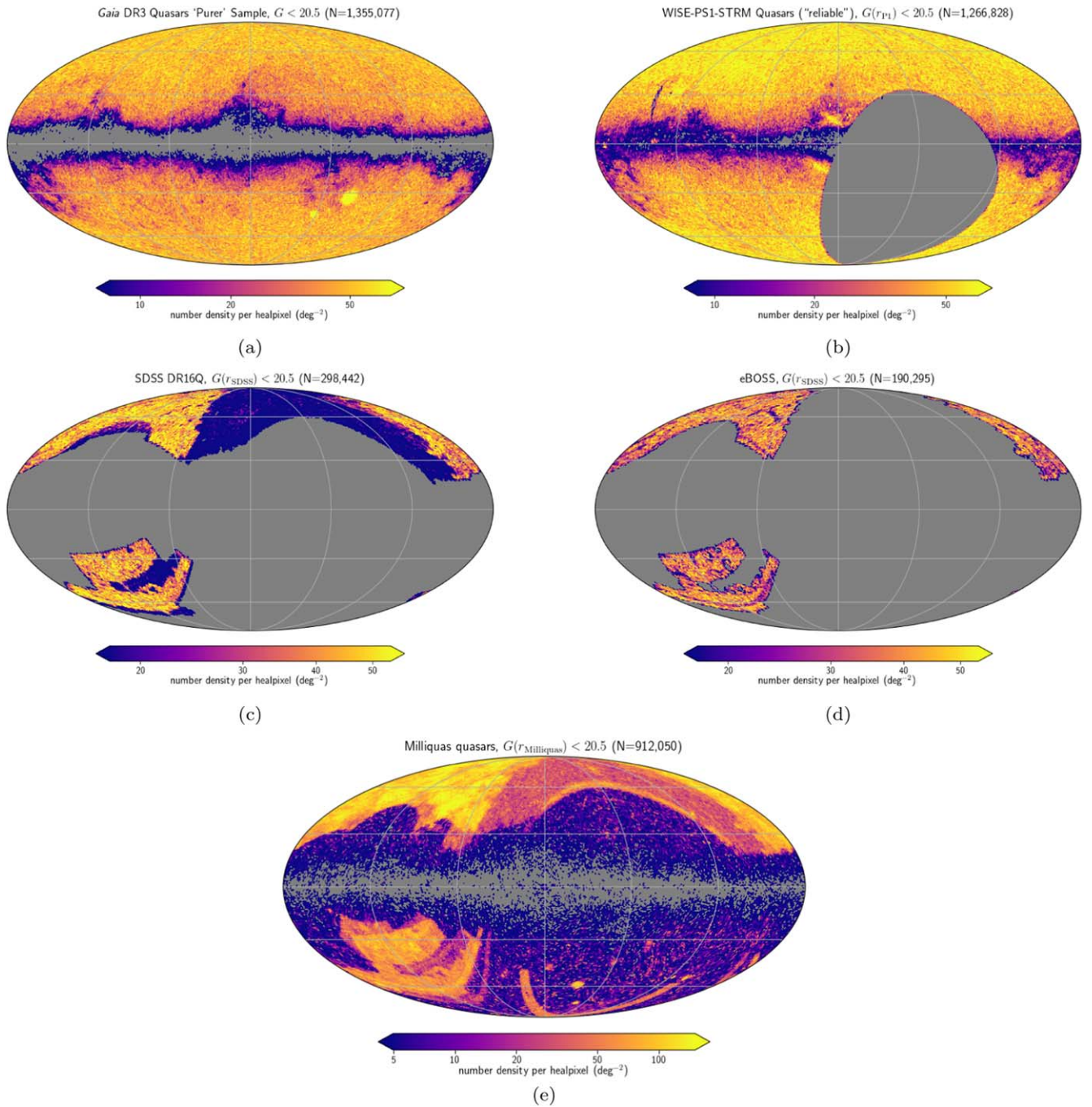


**Figure 12.** Distribution of  $G$  magnitudes of Quiaia (black), compared to the full Gaia candidates sample (gray), the Gaia DR3 *purer* sample (green), and the SDSS DR16Q quasar sample (blue).

that have  $A_V > 0.5$  mag, as well as healpixels with no quasars. For the sky fraction  $f_{\text{sky}}$ , we see that Quiaia and Gaia DR3 *purer* are limited only by the dusty regions, and cover over 30% more area than WISE-PS1 (which is limited by Pan-STARRS), nearly three times that of SDSS DR16Q, and over five times that of the eBOSS quasar clustering catalog. Compared to the Gaia DR3 *purer* sample, Quiaia has a slightly smaller number of sources, but due to its redshift distribution gives a slightly higher effective volume. The on-sky number density is similar for all of the catalogs when limiting them to similar magnitudes, with WISE-PS1 slightly higher because it has a similar number of objects to the Gaia catalogs but over a smaller area, and SDSS DR16Q and the eBOSS clustering catalog slightly lower. When including



**Figure 13.** (a) The selection function map for the  $G < 20.0$  subset of Quiaia, based on a Gaussian process model of the dust, stellar distribution, and  $M_{10}$ . (b) The fractional residuals between a random catalog downsampled by the modeled selection function and the true Quiaia  $G < 20.0$  catalog.



**Figure 14.** Other current quasar catalogs for comparison with Quaia. All are shown for sources with  $G < 20.5$  or the equivalent converted from another band, in Galactic coordinates and displayed using a Mollweide projection. The catalogs are (a) the Gaia DR3 *puer* sample, (b) the WISE-PS1-STRM catalog, (c) the SDSS DR16Q catalog, (d) the eBOSS quasar clustering catalog, and (e) the Milliquas catalog. Note that the color bars have different scales in each panel.

faint sources for the catalogs, WISE-PS1 has two and a half times the on-sky number density as Quaia, and SDSS DR16Q and the eBOSS clustering catalog have one and a half to two times.

For the volume comparison, we compute two different volumes. The first is a simple *spanning* volume,  $V_{\text{span}}$ , which is just the comoving volume in the sky area of the survey (as given by  $f_{\text{sky}}$  of the full sky area) in a redshift range  $0.8 < z < 2.2$ , a typical redshift range for clustering analyses (taken from the range of the eBOSS quasar clustering catalog). Thus, it compares in the same way as the survey areas, but gives an idea of the physical volume the catalogs span. The

second is the effective volume, described in Section 4.1; we use that same  $P(k) = 4 \times 10^4 (h^{-1} \text{Mpc})^3$  for the volume calculation for all catalogs. We see that the effective volume of WISE-PS1 is much larger (nearly three times) than that of Quaia as a result of its larger number of sources, though when considering samples with the same limiting magnitude, WISE-PS1 and Quaia have comparable effective volumes. The effective volume of Quaia is nearly twice as large as that of SDSS DR16Q, and  $6\times$  for the magnitude-limited sample; compared to the eBOSS quasar clustering catalog, the effective volume of Quaia is over twice as large, and  $7\times$  for the magnitude-limited sample.

**Table 1**  
Comparison between Quiaia and Other Existing Quasar Catalogs, Detailed in the Text

	$N$	$f_{\text{sky}}$	$\bar{n}$ , deg $^{-2}$	$V_{\text{span}}$ , ( $h^{-1}$ Gpc) $^3$	$V_{\text{eff}}$ , ( $h^{-1}$ Gpc) $^3$	$z_{\text{med}}$	$f( \delta z  < 0.01)$	$f( \delta z  < 0.1)$
Quaia	1,234,715	0.73	40.78	143.78	7.08	1.48	0.63	0.84
Gaia <i>purser</i>	1,647,311	0.73	54.42	143.76	9.24	1.63	0.53	0.62
$G < 20.5$	1,286,788	0.73	42.51	143.76	6.50	1.61	0.62	0.70
WISE-PS1	2,386,121	0.56	103.89	109.08	20.88	1.38	0.11	0.71
$G_{\text{eff}} < 20.5$	1,130,925	0.56	49.25	109.06	7.32	1.41	0.12	0.76
SDSS DR16Q	637,371	0.26	60.18	50.30	4.16	1.77	$\sim 1$	$\sim 1$
$G_{\text{eff}} < 20.5$	297,940	0.26	28.17	50.23	1.18	1.67	$\sim 1$	$\sim 1$
eBOSS clustering	409,286	0.14	72.52	26.80	3.21	1.60	$\sim 1$	$\sim 1$
$G_{\text{eff}} < 20.5$	190,263	0.14	33.96	26.61	1.01	1.49	$\sim 1$	$\sim 1$

**Note.** We show the quantities for the full catalogs (for sources with reliable redshifts) as well as the catalogs limited to  $G < 20.5$  or the rough equivalent converted from another band. For all quantities and catalogs shown, we exclude areas with high dust extinction ( $A_V > 0.5$  mag); this excludes  $\sim 5\%$  of sources for Quaia and Gaia DR3 *purser*,  $\sim 18\%$  of the full WISE-PS1 sample, and a negligible number of sources for SDSS DR16Q and the eBOSS clustering catalog. We note that the SDSS DR16Q catalog is a superset of quasars from many SDSS campaigns and is not intended to be uniform, which should be considered in particular for the sky fraction and spanning volume quantities. We show the number of sources  $N$ , the fraction of sky area covered  $f_{\text{sky}}$ , the mean number density per square degree  $\bar{n}$ , the spanning volume between  $0.8 < z < 2.2$   $V_{\text{span}}$ , the effective volume  $V_{\text{eff}}$ , the median redshift  $z_{\text{med}}$ , and the fraction of objects with  $|\delta z| \equiv |\Delta z / (1 + z)| < 0.01$  and  $< 0.1$  (where applicable).

The catalogs all have a similar median redshift, of around  $1.4 < z < 1.7$ , extending to 1.77 for SDSS DR16Q when including faint sources. However, they have significantly different redshift precision; in Table 1 we show outlier fractions estimated from comparisons to spectroscopic redshifts. We see that both of the Gaia catalogs have a similar fraction of high-precision redshifts ( $|\Delta z / (1 + z)| < 0.01$ ), but Quaia has a much higher fraction of redshifts that are not strong outliers ( $|\Delta z / (1 + z)| < 0.1$ ) compared to Gaia DR3 *purser*. WISE-PS1 falls between Quaia and Gaia DR3 *purser* in terms of strong outliers, but has an extremely low fraction of high-precision redshifts as it is a photometric survey. We note that for both Gaia DR3 *purser* and WISE-PS1, the redshift precision is significantly lower when considering the full catalog compared to samples limited to  $G_{\text{eff}} < 20.5$  like Quaia; we show both for a fair comparison. The SDSS DR16Q catalog and the eBOSS quasar clustering catalog have spectroscopic redshifts, so these are almost all very high precision; Lyke et al. (2020) estimated from a visual inspection that less than 1% of the SDSS DR16Q redshifts are outliers with  $\Delta v > 3000$  km s $^{-1}$  ( $|\Delta z| > 0.01$ ), independent of redshift; note that this is a slightly different sample than the eBOSS clustering catalog, but we can expect it to be similar. The SDSS DR16Q quasar sample has typical statistical redshift errors of  $|\Delta z| \sim 0.001$ .

To give more of an idea of the redshift precision of Quaia, we compare it to existing all-sky photometric galaxy catalogs. A common statistic to summarize photometric redshift uncertainty robust to outliers is the SMAD, scaled median absolute deviation, defined as  $1.4826 \times \text{med}(|\Delta z - \text{med}(\Delta z)|)$ , where  $\Delta z = z_{\text{phot}} - z_{\text{spec}}$  (the scaling factor adjusts the MAD such that SMAD is approximately equal to the standard deviation for normalized data). The SMAD of the full Quaia catalog ( $G < 20.5$ ) is  $\text{SMAD}(\Delta z) = 0.023$ , and the normalized SMAD of the redshift errors with the  $(1 + z)$  factor divided out is  $\text{SMAD}(\Delta z / (1 + z)) = 0.008$ . For comparison, the WISE  $\times$  SuperCOSMOS catalog of 20 million galaxies with  $z_{\text{med}} = 0.2$  (Bilicki et al. 2016) has an  $\text{SMAD}(\Delta z)$  of  $\sim 0.04$  and an  $\text{SMAD}(\Delta z / (1 + z))$  of  $\sim 0.035$ . The Two Micron All Sky Survey Photometric Redshift (2MPZ) catalog has around 1 million galaxies with a similar median redshift

(Bilicki et al. 2013), which have an SMAD ( $\Delta z$ ) of  $\sim 0.015$ . Quaia thus falls in between these common photometric galaxy samples in terms of overall redshift precision; however, we note that it is difficult to capture the redshift error of Quaia in a single statistic, given both its large number of highly precise redshifts and non-insignificant number of outliers.

We also note that the ongoing DESI survey (Aghamousa et al. 2023; DESI Collaboration et al. 2024) will observe a high density of quasars over a large sky area (Chaussidon et al. 2023), which will be competitive with and complementary to Quaia.

#### 4.4. Catalog Format

The complete Quaia catalog contains our decontaminated quasar sample with computed redshift information, relevant Gaia properties, and crossmatched catalog information. The complete catalog format with column names, units, column descriptions, and an example entry is shown in Table 2. Additional information for the sources can be obtained by joining the catalog with the relevant data source with the associated identifier (Gaia or unWISE). We include only sources with  $G < 20.5$  in the catalog; we also publish a version limited to  $G < 20.0$ , along with the selection function models fit to each (Section 4.2) and *random* catalogs generated from the selection functions. The catalog includes our SPZ redshifts  $z_{\text{Gaia}}$  along with  $1\sigma$  redshift errors, sky position, Gaia photometry, unWISE photometry, and proper motion information. The catalog is in FITS format (Wells et al. 1981), and units and descriptions are provided for each column.

#### 4.5. Limitations

While the Quaia catalog presents a highly useful quasar sample, it does have various limitations. We reiterate and discuss the main ones here.

We estimate spectrophotometric redshifts for the quasars, which are generally more accurate than the Gaia estimates, but are still low precision compared to spectroscopic redshifts. The uncertainties on these redshifts should be taken into account for any measurements, and the rate of catastrophic redshift errors



**Table 2**  
Format and Column Descriptions of Quaia, Published as a FITS Data File (Wells et al. 1981)

Column Name	Symbol	Units	Description	Example Entry Value
source_id	...	...	Gaia DR3 source identifier	6459630980096
unwise_objid	...	...	unWISE DR1 source identifier	0453p000o0014479
redshift_quaia	$z_{\text{Quaia}}$	...	Spectrophotometric redshift estimate	0.416867
redshift_quaia_err	...	...	$1\sigma$ uncertainty on spectrophotometric redshift estimate	0.060812
ra	...	deg	Barycentric R.A. of the source in ICRS at 2016.0	44.910498
dec	...	deg	Barycentric decl. $\delta$ of the source in ICRS at 2016.0	0.189649
l	...	deg	Galactic longitude	176.659434
b	...	deg	Galactic latitude	-48.835164
phot_g_mean_mag	$G$	mag	Gaia $G$ -band mean magnitude	20.173105
phot_bp_mean_mag	BP	mag	Gaia integrated BP mean magnitude	20.200150
phot_rp_mean_mag	RP	mag	Gaia integrated RP mean magnitude	18.871586
mag_w1_vg	W1	mag	unWISE W1 magnitude	14.774343
mag_w2_vg	W2	mag	unWISE W2 magnitude	13.923867
pm	$\mu$	mas yr <sup>-1</sup>	Total proper motion	0.383797
pmra	$\mu_{\alpha*}$	mas yr <sup>-1</sup>	Proper motion in R.A. $\mu_{\alpha*} \equiv \mu_{\alpha} \cos \delta$ of the source in ICRS at 2016.0	0.217806
pmdec	$\mu_{\delta}$	mas yr <sup>-1</sup>	Proper motion in decl. $\mu_{\delta}$ of the source in ICRS at 2016.0	-0.316007
pmra_error	$\sigma_{\mu_{\alpha*}}$	mas yr <sup>-1</sup>	Standard error of proper motion in R.A. direction	0.679419
pmdec_error	$\sigma_{\mu_{\delta}}$	mas yr <sup>-1</sup>	Standard error of proper motion in decl. direction	0.608799

**Note.** For an example entry, we show the first catalog row.

(not necessarily captured by the redshift uncertainty) should be considered when thinking about possible uses of the catalog.

The selection function model has multiple potential limitations. While it broadly captures the selection effects that affect the quasar sample, it has significantly lower accuracy around the galactic plane; precision measurements may require masking this region. The regions around the LMC and SMC are also more poorly modeled; users may want to mask this area. We also note that we are not fitting the healpixels with zero quasars, which may result in a slight bias toward populated regions, and fixes the zero-probability region of the selection function. Our selection function map depends only on-sky position and not other properties such as magnitude or redshift (besides fitting it to the appropriate subsample); a treatment incorporating these dependencies may be important for certain uses. The gold standard for completeness estimation is data injection and recovery tests. Unfortunately, the Gaia instrumentation has black-box elements, such as onboard image segmentation, onboard object detection, and onboard downlink prioritization, that make it impossible to perform end-to-end injection tests, so we rely on a data-driven approach, which may be less robust and more sensitive to modeling choices. Given this, it is possible that we are overfitting the selection function. Finally, the selection function depends on the assumption of isotropy, which we know to be broken to some extent by the kinematic dipole (Stewart & Sciamia 1967; Secrest et al. 2021); we will explore and measure this in an upcoming work (see Section 4.6). Users employing the selection maps or generating their own selection function for some subset of the catalog should take note of these potential issues.

Generally, Quaia has a relatively low number density (e.g., compared to the SDSS sample). This means that it may not be ideal for certain cosmological measurements, which may be shot noise dominated.

Finally, we note that this catalog is based on the Gaia quasar candidates sample, and it will inherit many of the limitations of that sample (Gaia Collaboration et al. 2023a). We are also limited to the Gaia-derived properties (e.g., the Gaia redshifts

that are a feature for our estimates). In upcoming Gaia data releases, the collaboration will release more BP/RP spectra and we will have the opportunity to work directly from the spectral data to improve the catalog.

#### 4.6. Potential Applications

Quasars are highly biased tracers of the cosmic web that trace the matter distribution at higher redshift than galaxies and in the mildly nonlinear regime. Given the Quaia catalog's sampling of quasars to deep magnitudes and across a large volume, and its reduced systematic contamination allowed by space-based observations, Quaia lends itself to large-scale structure analyses, many of which are currently ongoing.

Thanks to its large volume and well-characterized selection function, Quaia is perhaps the best current sample for testing homogeneity and isotropy in the Universe (D. W. Hogg et al. 2024, in preparation), and relatedly for measuring the dipole in the quasar distribution (A. Williams et al. 2024, in preparation), which recent measurements have consistently found to be in mild tension with the kinematic interpretation in the  $\Lambda$ CDM model. Quaia's volume also makes it a good sample for a measurement of the matter-radiation equality scale,  $k_{\text{eq}}$  (e.g., Bahr-Kalus et al. 2023).

The catalog is particularly well suited for cross-correlations with other all-sky observations of projected tracers of the large-scale structure, which are less sensitive to redshift errors compared to 3D ones. Examples of this are the CMB, the CIB, or maps of the thermal Sunyaev-Zel'dovich effect. Aghamousa et al. (2023) used the cross-correlation between CMB lensing and Quaia to constrain the growth of matter fluctuations via the parameter  $S_8$ , achieving competitive constraints as well as showing that Quaia can break the degeneracy between  $\Omega_m$  and  $\sigma_8$ . An analysis of primordial non-Gaussianity (parameterized by  $f_{\text{NL}}$ ) from this cross-correlation with CMB lensing is also underway. Analyses of the cross-correlation with CMB temperature to measure the Integrated Sachs-Wolfe effect, and with the CIB to constrain the star formation history at high redshifts (e.g., Jogo et al. 2023), are currently under investigation. Another measurement enabled by the catalog is

the cross-correlation of quasar proper motions with the large-scale structure, which directly estimates the cosmological quantity  $Hf\sigma_8$  (Duncan et al. 2024). Additionally, cross-correlations of Quaia with galaxy surveys may allow for measurements of the baryon acoustic feature (Patej & Eisenstein 2018; Zarrouk et al. 2021) and quasar environments (Padmanabhan et al. 2009; Shen et al. 2013).

Quaia is also useful for void studies, including constraining core cosmological parameters with the void size distribution; this investigation is underway (N. Arsenov et al. 2024, in preparation). The catalog is additionally relevant to astrophysical analyses of quasar properties, given its large sky coverage and multiband photometry, such as the role of galaxy interactions on AGN activity. Quaia sources may also be used to study the potential of quasars as standard candles. Further, Quaia provides perhaps the best quasar coverage of the southern sky, which may be important for a variety of applications such as identifying interesting sources there, adding new information to known sources, or calibrating surveys in that sky region. Finally, while a 3D clustering analysis of Quaia may be limited by the catalog’s relatively low number density and moderate redshift precision, a careful analysis may yield useful constraints, especially using techniques targeted at wide-field surveys (e.g., Lanusse et al. 2015).

The latter is comparable or better than other state-of-the-art galaxy and quasar samples used in large-scale structure analyses, but not enough to necessarily allow an accurate interpretation.

## 5. Summary and Data Access

We have constructed a new quasar catalog, Quaia, the Gaia-unWISE Quasar Catalog, designed for cosmological studies, derived from the Gaia DR3 quasar candidates sample and using unWISE photometry to remove contaminants and derive precise redshifts. Our key contributions and the features of the catalog are as follows:

1. We have decontaminated the Gaia DR3 quasar candidates sample with proper motion cuts and optimized color cuts based on Gaia and unWISE photometry. This reduced the number of known contaminants by approximately four times, while only excluding 1.2% of known quasars with respect to the superset of Gaia quasar candidates (that have unWISE photometry, Gaia redshifts, and a  $G$ -magnitude cut of  $G < 20.6$ ).
2. The catalog extends to a limiting magnitude of  $G < 20.5$  and contains 1,295,502 sources; we also release a brighter, cleaner sample limited to  $G < 20.0$ , which includes 755,850 sources.
3. Quaia covers the entire sky, only limited by selection effects near the Galactic plane; excluding highly dust-extincted regions ( $A_V > 0.5$  mag), this results in an area of  $30277.52 \text{ deg}^2$  ( $f_{\text{sky}} = 0.73$ ).
4. We have improved the Gaia redshift estimates using a  $k$ NN model trained on these redshifts and Gaia and unWISE colors with SDSS spectroscopic redshift labels, producing spectrophotometric redshifts. The median redshift of the  $G < 20.0$  catalog is  $z_{\text{med}} = 1.45$ , with 94% (75%) of redshifts within  $|\Delta z / (1 + z)| < 0.2$  (0.01) of SDSS redshifts. This is a reduction in the number of catastrophic outliers by approximately three times

(approximately two times) compared to the Gaia redshift estimates.

5. We produced a data-driven model of the selection function, which includes the systematic effects of dust, the source density of the parent surveys Gaia and unWISE, and the scanning laws of the parent surveys. We used this to generate random catalogs of Poisson-distributed points with similar selection effects to Quaia.

The catalog, selection function, and related data products are publicly available at [10.5281/zenodo.10403370](https://doi.org/10.5281/zenodo.10403370), along with documentation. The code used to generate this catalog is open source and available at <https://github.com/kstoreyf/gaia-quasars-lss>.

## Acknowledgments

The authors are grateful to the members of the Gaia Collaboration, in particular Coryn Bailer-Jones, Morgan Fouesneau, Anthony Brown, Ludovic Delchambre, Tristan Cantat-Gaudin, and Arvind Hughes. The authors also thank Lyuba Slavcheva-Mihova, Nestor Arsenov, Andras Kovacs, Anže Slosar, Giulia Piccirilli, Iain Duncan, Abby Williams, Dustin Lang, Mehdi Rezaie, Alex Malz, Lehmann Garrison, and Nathan Secrest for very helpful discussions. Additionally, the authors thank the members of the Astronomical Data group at the Center for Computational Astrophysics for useful feedback. The authors are grateful to the anonymous referee, whose feedback has significantly strengthened this work. This project was developed in part at the Gaia Fête, hosted by the Flatiron Institute Center for Computational Astrophysics in 2022 June. This work has made use of data from the European Space Agency 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 WISE, which is 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. We specifically use the unWISE coadds, produced by D. Lang, A.M. Meisner, and D.J. Schlegel. K.S.F. is supported by the NASA FINESST program under award No. 80NSSC20K1545. G.F. acknowledges the support of the European Research Council under the Marie Skłodowska Curie actions through the Individual Global Fellowship No. 892401 PiCOGAMBAS. D. A. acknowledges support from the Beecroft Trust, and from the John O’Connor Research Fund, at St. Peter’s College, Oxford. This research made use of computational resources at New York University (NYU); the authors thank the NYU high-performance computing team.

*Software:* Astropy (The Astropy Collaboration et al. 2013, 2018, 2022), NumPy (Harris et al. 2020), IPython (Perez & Granger 2007), SciPy (Virtanen et al. 2020), matplotlib (Hunter 2007), healpy (Gorski et al. 2005; Zonca et al. 2019), george (Ambikasaran et al. 2016).

## ORCID iDs

Kate Storey-Fisher  <https://orcid.org/0000-0001-8764-7103>  
 David W. Hogg  <https://orcid.org/0000-0003-2866-9403>  
 Hans-Walter Rix  <https://orcid.org/0000-0003-4996-9069>

Anna-Christina Eilers  <https://orcid.org/0000-0003-2895-6218>

Giulio Fabbian  <https://orcid.org/0000-0002-3255-4695>

Michael R. Blanton  <https://orcid.org/0000-0003-1641-6222>

David Alonso  <https://orcid.org/0000-0002-4598-9719>

## References

- Aghamousa, A., Aguilar, J., Ahlen, S., et al. 2023, arXiv:1611.00036
- Alonso, D., Fabbian, G., Storey-Fisher, K., et al. 2023, *JCAP*, **2023**, 043
- Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O’Neil, M. 2016, *ITPAM*, **38**, 252
- Ata, M., Baumgarten, F., Bautista, J., et al. 2018, *MNRAS*, **473**, 4773
- Bahr-Kalus, B., Parkinson, D., & Mueller, E. M. 2023, *MNRAS*, **524**, 2463
- Bailer-Jones, C. A. L. 2021, DSC ApsisOps3.2 Validation Report, ESA Gaia, [https://dms.cosmos.esa.int/COSMOS/doc\\_fetch.php?id=1445248](https://dms.cosmos.esa.int/COSMOS/doc_fetch.php?id=1445248)
- Bailer-Jones, C. A. L., Fouesneau, M., & Andrae, R. 2019, *MNRAS*, **490**, 5615
- Beck, R., Dodds, S. C., & Szapudi, I. 2022, *MNRAS*, **515**, 4711
- Beck, R., Szapudi, I., Flewelling, H., et al. 2020, *MNRAS*, **500**, 1633
- Bilicki, M., Jarrett, T. H., Peacock, J. A., Cluver, M. E., & Steward, L. 2013, *ApJS*, **210**, 9
- Bilicki, M., Peacock, J. A., Jarrett, T. H., et al. 2016, *ApJS*, **225**, 5
- Cantat-Gaudin, T., Fouesneau, M., Rix, H. W., et al. 2023, *A&A*, **669**, A55
- Carrasco, J. M., Weiler, M., Jordi, C., et al. 2021, *A&A*, **652**, A86
- Castorina, E., Hand, N., Seljak, U., et al. 2019, *JCAP*, **2019**, 010
- Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, The Pan-STARRS1 Surveys, arXiv:1612.05560
- Chaussidon, E., Yéche, C., Palanque-Delabrouille, N., et al. 2023, *ApJ*, **944**, 107
- Chiang, Y.-K. 2023, *ApJ*, **958**, 118
- Claeskens, J. F., & Surdej, J. 2002, *A&ARv*, **10**, 263
- Dam, L., Lewis, G. F., Brewer, B. J., et al. 2023, *MNRAS*, **525**, 231
- Delchambre, L., Bailer-Jones, C. A. L., Bellas-Velidis, I., et al. 2023, *A&A*, **674**, A31
- DESI Collaboration, Adame, A. G., Aguilar, J., et al. 2024, *AJ*, **167**, 62
- DiPompeo, M. A., Bovy, J., Myers, A. D., & Lang, D. 2015, *MNRAS*, **452**, 3124
- DiPompeo, M. A., Hickox, R. C., Eftekharzadeh, S., & Myers, A. D. 2017, *MNRAS*, **469**, 4630
- Ducourant, C., Krone-Martins, A., Galluccio, L., et al. 2023, *A&A*, **674**, A11
- Duncan, I., Alonso, D., Slosar, A., & Storey-Fisher, K. 2024, *OJAp*, **7**, 9
- Flesch, E. W. 2021, arXiv:2105.12985
- Fletcher, R. R. 1987, *Practical Methods of Optimization* (New York: Wiley)
- 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, Prusti, T., & de Bruijne, J. H. J. 2016, *A&A*, **595**, A1
- Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. 2023b, *A&A*, **674**, A1
- García-García, C., Ruiz-Zapatero, J., Alonso, D., et al. 2021, *JCAP*, **2021**, 030
- Gorski, K. M., Hivon, E., Banday, A. J., et al. 2005, *ApJ*, **622**, 759
- Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, *Natur*, **585**, 357
- Heintz, K. E., Fynbo, J. P. U., Geier, S. J., et al. 2020, *A&A*, **644**, A17
- Heintz, K. E., Fynbo, J. P. U., Høg, E., et al. 2018, *A&A*, **615**, L8
- Hopkins, P. F., Hernquist, L., Cox, T. J., et al. 2006, *ApJS*, **163**, 1
- Hou, J., Sánchez, A. G., Ross, A. J., et al. 2020, *MNRAS*, **500**, 1201
- Hughes, A. C. N., Bailer-Jones, C. A. L., & Jamal, S. 2022, *A&A*, **668**, A99
- Hunter, J. D. 2007, *CSE*, **9**, 90
- Ivezic, Z. 2017, in *IAU Symp. 324, New Frontiers in Black Hole Astrophysics* (Cambridge: Cambridge Univ. Press), 330
- Jego, B., Ruiz-Zapatero, J., García-García, C., Koukoufilippas, N., & Alonso, D. 2023, *MNRAS*, **520**, 1895
- Jin, X., Zhang, Y., Zhang, J., et al. 2019, *MNRAS*, **485**, 4539
- Kormendy, J., & Ho, L. C. 2013, *ARA&A*, **51**, 511
- Krolewski, A., Percival, W. J., Ferraro, S., et al. 2023, arXiv:2305.07650
- Kunsági-Máté, S., Beck, R., Szapudi, I., & Csabai, I. 2022, *MNRAS*, **516**, 2662
- Lang, D. 2014, *AJ*, **147**, 108
- Lanusse, F., Rassat, A., & Starck, J. L. 2015, *A&A*, **578**, A10
- Leistedt, B., Peiris, H. V., & Roth, N. 2014, *PhRvL*, **113**, 221301
- Lusso, E., Risaliti, G., Nardini, E., et al. 2020, *A&A*, **642**, A150
- Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, *ApJS*, **250**, 8
- Mainzer, A., Bauer, J., Grav, T., et al. 2011, *ApJ*, **731**, 53
- Martin, D. C., Fanson, J., Schiminovich, D., et al. 2005, *ApJL*, **619**, L1
- Meisner, A. M., Lang, D., Schlafly, E. F., & Schlegel, D. J. 2019, *PASP*, **131**, 124504
- Ménard, B., & Bartelmann, M. 2002, *A&A*, **386**, 784
- Mueller, E. M., Rezaie, M., Percival, W. J., et al. 2021, arXiv:2106.13725
- Myers, A. D., Palanque-Delabrouille, N., Prakash, A., et al. 2015, *ApJS*, **221**, 27
- Nikutta, R., Hunt-Walker, N., Nenkova, M., Ivezic, Z., & Elitzur, M. 2014, *MNRAS*, **442**, 3361
- Onken, C. A., Wolf, C., Hon, W. J., et al. 2023, *PASA*, **40**, e010
- Padmanabhan, N., White, M., Norberg, P., & Porciani, C. 2009, *MNRAS*, **397**, 1862
- Patej, A., & Eisenstein, D. J. 2018, *MNRAS*, **477**, 5090
- Perez, F., & Granger, B. E. 2007, *CSE*, **9**, 21
- Powell, M. C., Urry, C. M., Cappelluti, N., et al. 2020, *ApJ*, **891**, 41
- Proft, S., & Wambsganss, J. 2015, *A&A*, **574**, A46
- Rasmussen, C. E., & Williams, C. K. 2005, *Gaussian Processes for Machine Learning* (Cambridge, MA: MIT Press)
- Rauch, M. 1998, *ARA&A*, **36**, 267
- Rezaie, M., Ross, A. J., Seo, H. J., et al. 2021, *MNRAS*, **506**, 3439
- Rimoldini, L., Holl, B., Gavras, P., et al. 2023, *A&A*, **674**, A14
- Risaliti, G., & Lusso, E. 2015, *ApJ*, **815**, 33
- Ross, A. J., Bautista, J., Tojeiro, R., et al. 2020, *MNRAS*, **498**, 2354
- Schlafly, E. F., & Finkbeiner, D. P. 2011, *ApJ*, **737**, 103
- 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., Richards, G. T., Hall, P. B., et al. 2010, *AJ*, **139**, 2360
- Scranton, R., Ménard, B., Richards, G. T., et al. 2005, *ApJ*, **633**, 589
- Secrest, N., von Hausegger, S., Rameez, M., et al. 2021, *ApJL*, **908**, L51
- Setti, G., & Woltjer, L. 1973, *ApJL*, **181**, L61
- Shen, Y., McBride, C. K., White, M., et al. 2013, *ApJ*, **778**, 98
- Sherwin, B. D., Das, S., Hajian, A., et al. 2012, *PhRvD*, **86**, 083006
- Shu, Y., Koposov, S. E., Evans, N. W., et al. 2019, *MNRAS*, **489**, 4741
- Stewart, J. M., & Sciamia, D. W. 1967, *Natur*, **216**, 748
- Sunyaev, R., & Zeldovich, Y. 1970, *Ap&SS*, **7**, 3
- The Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, *ApJ*, **935**, 167
- The Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, *AJ*, **156**, 123
- The Astropy Collaboration, Robitaille, T. P., & Tollerud, E. J. 2013, *A&A*, **558**, A33
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *NatMe*, **17**, 261
- Wells, D. C., Greisen, E. W., & Harten, R. H. 1981, *A&AS*, **44**, 363
- Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, *AJ*, **140**, 1868
- Yahata, K., Suto, Y., Kayo, I., et al. 2005, *PASJ*, **57**, 529
- Yang, Q., & Shen, Y. 2023, *ApJS*, **264**, 9
- Yang, Q., Wu, X.-B., Fan, X., et al. 2017, *AJ*, **154**, 269
- Yu, Z., Martini, P., Davis, T. M., et al. 2020, *ApJS*, **246**, 16
- Zarrouk, P., Rezaie, M., Raichoor, A., et al. 2021, *MNRAS*, **503**, 2562
- Zonca, A., Singer, L., Lenz, D., et al. 2019, *JOSS*, **4**, 1298