A Catalog of Candidate Double and Lensed Quasars from Gaia and WISE Data

Valeri V. Makarov © and Nathan J. Secrest ©

U.S. Naval Observatory, 3450 Massachusetts Ave NW, Washington, DC 20392-5420, USA; valeri.makarov@gmail.com

Received 2022 July 29; revised 2022 September 14; accepted 2022 October 4; published 2022 December 14

Abstract

Making use of strong correlations between closely separated multiple or double sources and photometric and astrometric metadata in Gaia Early Data Release 3 (EDR3), we generate a catalog of candidate double- and multiply imaged lensed quasars and active galactic nuclei (AGNs), comprising 3140 systems. It includes two partially overlapping parts: a sample of distant (redshifts mostly greater than 1) sources with perturbed data; and systems that have been resolved into separate components by Gaia at separations less than 2″. For the first part, which is roughly one-third of the published catalog, we synthesized 0.617 million redshifts using multiple machine-learning prediction and classification methods, using independent photometric and astrometric data from Gaia EDR3 and the Wide-field Infrared Survey Explorer, with accurate spectroscopic redshifts from the Sloan Digital Sky Survey (SDSS) as a training set. Using these synthetic redshifts, we estimate a 4.9% rate of interlopers with spectroscopic redshifts below 1 in this part of the catalog. Unresolved candidate double and dual AGNs and quasars are selected as sources with a marginally high BP/RP excess factor (phot_bp_rp_excess_factor), which is sensitive to source extent, limiting our search to high-redshift quasars. For the second part of the catalog, additional filters on measured parallax and near-neighbor statistics are applied to diminish the propagation of the remaining stellar contaminants. The estimated rate of the positives (double or multiple sources) is 98%, and the estimated rate of dual (physically related) quasars is greater than 54%. A few dozen serendipitously found objects of interest are discussed in more detail, including known and new lensed images, planetary nebulae, young IR stars of peculiar morphology, and quasars with catastrophic redshift errors in SDSS.

Unified Astronomy Thesaurus concepts: Active galactic nuclei (16); Double quasars (406); Astrometry (80); Gaia (2360)

Supporting material: machine-readable table

1. Introduction

Dual active galactic nuclei (AGNs) and quasars are of tremendous importance to several fields of astrophysics. In extragalactic astronomy, there seems to be a clear evolutionary path for physically interacting binary supermassive black holes to emerge via mergers of early galaxies or their building blocks (Kauffmann & Haehnelt 2000; Colpi & Dotti 2011). They are believed to eventually coalesce, generating the most powerful bursts of gravitational radiation in the universe. The details of this process are far from clear, however. The initial separation, on a scale of $10^3$–$10^4$ pc, and the initial relative velocity are too high for direct interactions between the BHs. Dynamical friction within a gaseous medium is believed to be responsible for the initial hardening, which would bring the BHs within 1 pc or less (e.g., Mayer 2013). The roles of stars and star formation are not well understood, nor are the timescales and conditions for both of the BHs to form independent accretion disks and be active at the same time. It is not known if widely separated double AGNs are much more common at earlier cosmological epochs than in the local universe. The rate of galactic mergers appears to peak at redshifts $z \sim 2$–3, where it is 5 times higher than the rate at $z \sim 6$ (e.g., Ventou et al. 2019). Mergers at moderate redshifts have also been shown to have significantly higher rates of AGNs than single galaxies (e.g., Gao et al. 2020), supporting the evolutionary picture of galaxy mergers triggering BH activity (e.g., Hopkins et al. 2008).

Previously published collections of candidate physically double (lensed or dual) quasars have explored data from the Sloan Digital Sky Survey (SDSS; Inada et al. 2006, 2012) and been limited to $\sim 10^2$ objects, with even fewer confirmed duals at close separations (see Figure 8 in Satyapal et al. 2017). A promising method that leverages the massive statistical power of SDSS spectroscopy, selection on double-peaked [O III] emission, is nonetheless very inefficient, with most objects being revealed as single AGNs in follow-up studies (e.g., Rosario et al. 2010; Müller-Sánchez et al. 2015; Foord et al. 2020). Aside from successful X-ray campaigns focusing on nearby AGNs (Koss et al. 2012), high-angular resolution radio studies with the Very Long Baseline Array and Very Large Array have generally been required to confirm the presence of dual, compact, flat-spectrum radio AGNs (e.g., Rodriguez et al. 2006; Fu et al. 2011; Kharb et al. 2017), although pointed observations are observationally inefficient and the majority of AGNs are not expected to be radio-bright. Nonetheless, research on dual quasars and AGNs is gaining momentum, because of their importance as the progenitors of powerful gravitational-wave events that serve as “standard sirens,” the gravitational equivalent of standard candles such as Type Ia supernovae (SNe; e.g., Holz & Hughes 2005; Centrella et al. 2010). As quasars generally reside in galaxies, they have peculiar motions that can be several hundreds of kilometers per second relative to the Hubble flow. The resulting astrometric proper motions should be negligible for most of them, because of the great distances separating the sources from the observer.
and indeed this is a major reason for their use in creating the International Celestial Reference Frame (Charlot et al. 2020), the physical realization of the International Celestial Reference System. For a standard flat ΛCDM cosmology with \( H_0 = 70 \, \text{km} \, \text{s}^{-1} \, \text{Mpc}^{-1} \) and \( \Omega_m = 0.3 \), 1 mas subtends 8 pc at a typical quasar redshift of \( z = 1 \), so a quasar with a peculiar velocity of 100 km s\(^{-1}\) will have an intrinsic proper motion of \( \sim 0.01 \, \mu\text{as yr}^{-1} \), three orders of magnitude below even the most precise proper motions available from Gaia.

Multiplicity is one of the important factors that can perturb precision astrometry of quasars (Makarov et al. 2012). The closer AGNs at redshifts \( z \lesssim 0.5 \) usually have resolved host galaxies associated with them, which are in general asymmetric to some degree. Quasars found in double systems at large physical separations (Inada et al. 2006; Wang et al. 2009; Comerford et al. 2012) presumably originate from mergers of galaxies with individual central AGNs. The typical scale of the projected separations is \( \sim 1 \, \text{kpc} \), with a wide distribution of a few orders of magnitude. Cosmological simulations predict that the incidence of dual AGNs at moderate redshift is of the order of one to a few percent (for a review, see De Rosa et al. 2019). At a much lower rate, quasars can be gravitationally lensed by foreground galaxies with double- or multiply imaged configurations (Delchambre et al. 2019). The lensed images of the currently known systems are mostly packed within a few arcseconds. An as yet unknown fraction of quasars at high redshifts are multiply imaged systems lensed by more distant and less massive lenses that subtend 1″ or less. These features are likely to produce astrometric position offsets. Bogus proper motions can be measured for unresolved lensed images of variable quasars in some cases, because of the time lag of the light curve, which is due to light travel time and gravitational delay effects (Koptelova et al. 2012) and resulting photocenter shifts. Quasars that exhibit apparent proper motions are rare objects, and have only recently begun to be studied (Makarov & Secrest 2022; Souchay et al. 2022).

The main objective of this paper is select candidate dual AGNs (CDAGNs) from mid-IR AGNs (MIRAGNs) in the catalog of Secrest et al. (2015), crossmatched with Gaia Early Data Release 3 (EDR3). Using an MIR-selected initial sample favors obscured quasars at higher redshifts, and therefore results in a helpful bias toward galactic mergers (Gao et al. 2020). Smaller test samples of quasars with spectroscopically determined redshifts from SDSS (Lyke et al. 2020) and available Pan-STARRS images (Kaiser et al. 2010) are used in this paper for independent verification of our detection criteria. We also compute the near-neighbor distance statistics of resolved companions within 11″ for a large sample of 0.632 million quasars that are present in Gaia EDR3 to confirm the presence of real binary AGNs. Our chosen method of selecting unresolved CDAGNs is vulnerable to perturbations in the Gaia data that are caused by extended structures around central sources at low to moderate redshifts. Our search is therefore limited to \( z > 1 \) objects. Since accurate spectroscopic redshifts are only available for \( \sim 23\% \) of the MIRAGN sample that is located within the SDSS footprint, we develop and employ a number of machine-learning (ML) techniques to estimate redshifts from the more widely available data from Gaia EDR3 and the Wide-field Infrared Survey Explorer (WISE). In both the prediction and classification regimes, the ML-generated data are trained on and compared with the SDSS redshifts.

### 2. Methodology

#### 2.1. Initial Quasar Sample

We crossmatch the catalog of 1.4 million MIRAGNs from Secrest et al. (2015) with the Gaia EDR3 catalog, using a match tolerance of 0″.5 for reliability. This tight search radius results in the loss of some genuine matches, due to the limited astrometric precision of WISE, but it helps to reduce the rate of source confusion in resolved pairs. This produced 621,946 matches, 551,482 of which have valid proper-motion and parallax measurements. For verification and ML training purposes, we also match the Gaia counterpart coordinates with the SDSS specObj-dr16.fits table,\(^1\) allowing only spectra with \( \text{ZWARNING} = 0 \) or \( \text{ZWARNING} = 4 \), the latter of which can be the case for spectra with broad lines.\(^2\) To ensure spectroscopic fiber coverage of the Gaia counterpart for this auxiliary test subsample, we allow BOSS spectra within 1″ and SDSS spectra within 1″, resulting in 126,343 objects. The high reliability of the initial selection—prior to matching with the SDSS spectra table—is illustrated by the extremely low rate of stellar contaminants, with only 72 sources having negative SDSS redshifts. The total number of objects spectroscopically classified as “STAR” is 135 (0.1%).

#### 2.2. Redshift Training Parameters

Because the SDSS footprint is about one-quarter of the full sky, accurate spectroscopic redshifts are available for just 23% of the initial sample. To increase the output of our study, we generate synthetic redshifts using statistically correlated data from WISE and Gaia as training parameters. Ideal dependencies would include monotonic functions of redshift with a negligibly small dispersion. Needless to say, there are no astrometric or photometric parameters that possess such properties, because of the multitude of complex physical phenomena that contribute to the available observables. We reviewed dozens of parameters available in the WISE and Gaia catalogs, looking for any correlations with redshift in the sample of 0.126 million sources, and selected the few most promising ones. For the prediction ML training set, we settled on just four purely photometric parameters.

Figure 1 shows the dependencies of two photometric training parameters from the WISE mission data on SDSS-determined redshifts. These MIR magnitudes have been used to generate the input catalog of MIRAGNs (Secrest et al. 2015), which serves as our starting sample. The overall distribution of MIR colors is quite red by selection. The plots are generated by sorting the sample of 0.126 million sources by spectroscopic redshift, dividing into 30 nonoverlapping bins of equal size, then computing for each bin the median redshift and the median color (shown by the solid broken line with data points) as well as the \([0.16, 0.84]\) quantiles (shown by the dashed lines) and the \([0.05, 0.95]\) quantiles (dotted lines). The former pair of quantiles is the robust statistics analog of the ±1σ interval for a Gaussian distribution of probabilities. In both cases, the scatter of the color around the median value for a given redshift is larger than \( \pm 0.1 \, \text{mag} \), which limits the performance of the ML prediction, since it deteriorates most notably for the nearest sources. While the W2 − W3 color displays a nearly monotonic dependence at \( z > 1 \), the W1−W2 color has a distinct turnover, peaking at \( z = 1.5 \). Unfortunately, the gradients

---

\(^1\) https://www.sdss.org/dr16/spectro/spectro_access/

\(^2\) https://www.sdss.org/dr16/spectro/caveats/#zstatus
are rather small, preventing a strong performance of the ML prediction on just these two parameters.

A more pronounced gradient is the $G - W_1$ color with redshift, albeit with larger scatter, shown in the left panel of Figure 2. The nearby AGNs are distinctly redder than their more distant counterparts. The median $G$ magnitude, on the other hand, is quite flat across the range of redshift. We interpret this drop in the $G - W_1$ color by approximately 2.5 mag as an observational selection effect. A large fraction of the AGNs are heavily obscured at all redshifts and, due to the relatively shallow optical magnitude limit of Gaia ($G \lesssim 20$), become too faint or completely invisible with increasing distance. The fourth parameter used for the ML redshift prediction is the Gaia EDR3 optical color $G_{BP} - G_{RP}$, given as the $bp_{-}rp$ column in the Gaia catalog (Figure 2, right). It is probably the weakest discriminator, because of the fairly flat and nonmonotonic dependence at $z > 0.5$. We include it in the hope of obtaining better performance for objects at the smallest redshifts, which exhibit the highest rate of large errors, as explained below.

The main source of perturbation of Gaia astrometric and photometric measurements is not double or multiple morphology (which is a relatively rare occurrence), but the extended character of the host galaxies at small redshifts. All the sources in Gaia EDR3 are indiscriminately processed as point-like stellar images, using a unified set of line spread functions. The drastic deterioration caused by the extended structures at $z < 0.5$ is illustrated by the analysis of the photometric excess parameter $\text{phot}_{bp}_{-}rp_{-}excess\_factor$ in Makarov & Secrest (2022). Because of the different methods used in EDR3 to estimate the broadband magnitudes $G$ and the narrower $G_{BP}$ and $G_{RP}$ magnitudes, the ratio of the corresponding fluxes may deviate upward from the most common value slightly above 1 (Riello et al. 2021). While the majority of the more distant and luminous quasars are tightly grouped around a well-defined lower envelope at $\text{phot}_{bp}_{-}rp_{-}excess\_factor \approx 1.1$, the closer AGNs show a dramatic increase both in their median and dispersion values. The more luminous AGNs that are detectable at high redshift have higher ratios of AGN/galaxy emission, so the contribution of the underlying host structures quickly tapers off with $z$. On the other hand, as demonstrated in Makarov & Secrest (2022), $\text{phot}_{bp}_{-}rp_{-}excess\_factor$ is the most reliable indicator of resolved or unresolved multiplicity of quasar images at $z > 1$. As the goal of this investigation is to remove sources at $z < 1$ as cleanly as possible, without removing double or multiple...
sources at higher redshift, we do not use phot_bp_rp_excess_factor as a parameter in our ML training set.

2.3. ML Prediction of z

Having selected the four photometric training parameters, we generate a sample of 0.126 million sources with accurate SDSS spectroscopic redshifts, which is a 20.5% subset of our working sample, counting 0.617 million objects. The subset of sources with spectroscopically determined redshifts is randomly divided into two equally sized parts. The first is the training set, on which the ML classifier is tuned to predict the observed \( z \) from the data vector \( \{W1−W2, W2−W3, G−W1, G_{BP}−G_{RP}\} \). The second is the test set for assessing the performance of the different ML methods discussed below. This assessment is based on two statistical metrics of interest. The first is the robust standard deviation proxy, 1.5 times the median absolute deviation (MAD) of the measured versus predicted redshifts \( \Delta z_{\text{obs}} − \Delta z_{\text{pre}} \). This allows for the determination of the dispersion of \( \Delta z_{\text{obs}} − \Delta z_{\text{pre}} \), independent of the presence of statistical outliers. The rate of interlopers is the second statistical metric of interest, which we define as objects with \( z_{\text{pre}} > 1 \) and \( z_{\text{obs}} < 1 \), and we separately quantify the rate of more dangerous contaminants with \( z_{\text{pre}} > 1 \) and \( z_{\text{obs}} < 0.5 \). Note that while the training is performed on a set of 0.063 million randomly selected sources with observed redshifts, the trained prediction or classification is applied to the entire working sample of 0.617 million objects, including the training and test subsets.

We experimented with ML classifiers using six different prediction methods: Linear Regression, Gradient Boosted Trees, Decision Trees, Nearest Neighbors, Random Forest, and Neural Network. Multiple ML training and prediction runs were implemented with randomly selected training sets and different methods. The performance of each run was evaluated using the two metrics. The best results were obtained with the Nearest Neighbors and the Neural Network, and the latter was accepted for processing the entire sample. The best trials yielded a rate of interlopers of 5.64% and a rate of contaminants of 0.93%. Figure 3 depicts the distribution of \( z_{\text{pre}} \) versus \( z_{\text{obs}} \) for the test set. The robust standard deviation of the differences is 0.235. With 63,000 objects in the test set, we note a statistically significant bias of the predicted values with a median \( z_{\text{obs}} − z_{\text{pre}} \) of −0.036. If such objects propagate into our final selection at a significant rate, they could dominate the final selection of candidate double quasars.

Since we are only interested in sources with redshifts greater than 1, we also employed a range of ML classification methods, which provide a binary classification (yes or no). The choice of available techniques was wider, and the logic was somewhat different. Therefore, the classification and prediction outcomes were not identical for the same data sample and training set. This redundancy helped us to achieve slightly higher reliability for the selected sample.

2.4. ML Classification of z

A wider range of ML classification methods was applied to the training set in order to separate the sources into the categories \( x_{\text{cla}} = 0 \) and \( x_{\text{cla}} = 1 \), which are defined as \( z < 1 \) and \( z > 1 \), respectively. To improve the performance of this classification, an additional astrometric control from the Gaia EDR3 catalog was used. Figure 4 shows the statistical dependence of the astrometric_gof_al astrometric goodness of fit on redshift for the training set. This Gaia parameter indicates the degree of astrometric perturbation in the postfit residuals with respect to the standard five-parameter model (position components, parallax, and proper-motion components). An ideal fit to the data, with a dispersion of residuals in agreement with the formal errors, would yield a goodness-of-fit value equal to 0. The median dependence is quite flat for \( z > 1 \), but progressively elevated at lower redshift, indicating increasingly more unexplained astrometric variance. A dramatic excess is evident for the nearest AGNs, with even the lower 0.05 quantile taking positive values for \( z < 0.3 \). The astrometric_gof_al parameter thus provides additional leverage on the nearest sources, which are the most dangerous for this study. The presence of bright extended structures accounts for this behavior, making the Gaia astrometric measurements extremely noisy (Makarov et al. 2017b).

Figure 3. ML-predicted redshifts (with the Neural Network method) vs. SDSS spectroscopic redshifts for a test sample of 0.063 million quasars.

Figure 4. The statistical dependence of the Gaia EDR3 parameter astrometric_gof_al on spectroscopic redshift. The broken solid line with dots shows the median values in 30 equal bins of sorted redshifts. The dashed lines show the 0.16 and 0.84 quantiles of colors in the same binned subsamples. The dotted lines represent the 0.05 and 0.95 quantiles.
study also benefits from the apparent lack of correlation between this metadata type and the crucial \texttt{phot_bp_rp\_excess\_factor}. The two parameters come from different parts of the Gaia pipeline. While \texttt{phot_bp_rp\_excess\_factor} is sensitive to the registered image structure (multiplicity and extended component), \texttt{astrometric\_gof\_al} mostly reflects how well the higher-level astrometric measurements comply with the adopted astrometric model.

The Gradient Boosted Trees ML method yielded slightly better classification results than its close competitors the Neural Network and Nearest Neighbors methods. The rate of interlopers (sources with \(s_{\text{cla}} = 1\) and \(z_{\text{obs}} < 1\)) came to 7.60\%, while the loss rate of objects with \(z_{\text{obs}} > 1\) in \(s_{\text{cla}} = 0\) amounted to 9.56\%. Although the performance numbers are lower than the ML prediction output (Section 2.3), we use these two conceptually independent ML methods to achieve the greater reliability of the final catalog. The ML process is intrinsically stochastic, and the borderline objects with redshifts around 1 can have different results, because of the dispersion of classifier values at a given redshift. Our strategy is to err on the safe side and reject as many of such ambiguous cases as possible, at a cost to the completeness of our selection.

### 2.5. Generating a Large Sample of High-redshift Quasars

Both of the ML prediction and classification algorithms were applied to a sample of 617,093 MIRAGN/EDR3 objects that have all five of the training parameters considered in this work (the four photometric parameters, plus astrometric \texttt{gof\_al}). For each object, two values are obtained: the predicted redshift \(z_{\text{pre}}\) and the classification flag \(s_{\text{cla}}\). Since the prediction and classification runs are based on different ML methods and use different training sets, and nonidentical classifiers, the results are not entirely fully consistent for individual objects, in that the two parameters may be contradictory. In fact, consistent results were obtained for 590,147 sources, including 303,806 sources with both indicators placing the redshift above 1 and 286,341 sources with both indicators signaling a redshift below 1. Aiming for the greatest reliability of the final selection, we only accepted the sources in the first category for the catalog. Thus, the general sample is split into four nonoverlapping subsets, and we use two of these subsets in the following processing. The first one includes objects with \(z_{\text{pre}} > 1\) and \(s_{\text{cla}} = 1\), hereafter called the Z11 sample. The other one, which is only used for verification purposes, includes objects with both \(z_{\text{pre}} < 1\) and \(s_{\text{cla}} = 0\), hereafter called the Z00 sample. Using the subset of sources with SDSS spectroscopic determinations again, a 6.5\% rate of interlopers and a 0.8\% rate of contaminants is estimated for the Z11 sample (see Section 2.3). At this point, the quality of our Z11 selection is improved by removing 4699 sources with observed redshifts below 1 (i.e., known interlopers) and by adding 8122 objects with observed redshifts above 1 that were missed by our ML methods (i.e., known losses). The estimated rate of remaining interlopers in the resulting Z11 sample of 307,239 sources is 4.9\%. A similar cleaning procedure is applied to the Z00 sample.

Figure 5 displays separate distributions of the \texttt{phot_bp_r-\_p\_excess\_factor} parameter, which has a crucial role in this study, for the Z00 (left plot) and Z11 (right plot) samples. We observe significant differences in the widths and shapes of these histograms. Even though the most frequent values are between 1 and 2 for the nearer sources, the tail of high \texttt{phot_bp_r-\_p\_excess\_factor} values is much larger than for the more distant quasars, presenting clear evidence for a secondary population peaked at \(\sim 5\), which is absent from the high-\(z\) objects. This confirms the effectiveness of our ML-based selection and filtering of nearby sources with perturbing extended structures.

### 3. Selection of CDAGNs by Photometric Excess

Makarov & Secrest (2022) provided evidence that elevated values of \texttt{phot_bp_r-\_p\_excess\_factor} at \(z > 1\) are often caused by double or multiple sources in MIRAGNs, including optical pairs with Galactic stars, physical dual quasars, and multiply imaged gravitational lenses. We use the Z11 sample of sources with mostly ML-synthesized redshifts and use \texttt{phot_bp_r-\_p\_excess\_factor} to select candidate double objects that are brighter than \(G = 20\) mag. This upper limit on the optical magnitude is meant to avoid bogus positives among faint sources, where both the photometric ML controls and \texttt{phot_bp_r-\_p\_excess\_factor} are perturbed (Fabricius et al. 2021).

The lack of bimodality in the distribution of \texttt{phot_bp_r-\_p\_excess\_factor} for high-\(z\) quasars in the right panel of Figure 5 precludes the use of a simple threshold criterion. The small excess at very high values, above \(\sim 5\), is likely to be caused by a small fraction of nearby galaxies and stars with peculiar photometric properties that managed to propagate into the Z11 sample (see the Appendix). We therefore performed a series of visual tests by sorting the entire Z11 sample by

---

**Figure 5.** Histograms of the \texttt{phot_bp_r-\_p\_excess\_factor} parameter for the Z00 (left) and Z11 (right) samples. Note the logarithmic scale of both axes.
phot_bp_rp_excess_factor and investigating the rate of Gaia-resolved sources and the appearance of objects in the Pan-STARRS mosaic images at various levels. In this manner, we settled on an empirical sample quantile of 0.995 (leaving only 0.5% of the sample with the most perturbed values), which corresponds to a threshold value of phot_bp_rp_excess_factor = 1.545. This is a conservative cut, because tightly resolved sources continue to appear at even smaller parameter values, including the candidate gravitational lenses listed by Delchambre et al. (2019).

The resulting sample, hereafter called Sample 1, counts 1023 sources. We investigated a few possible options for improving its reliability by using additional parameters. These include the phot_bp_n_blended_transits and phot_rp_n-blended_transits values, as well as their sum, which may be correlated with the presence of close companions. The distributions are very broad, without obvious structure. Visual inspections of sources that have the sum of the two parameters below 5 (459 in total) helped to detect a few objects with grossly incorrect SDSS spectroscopic redshifts (Appendix), without revealing any obvious relation to multiplicity of these sources. Objects with very red optical colors G_{BP} - G_{RP} > 2.6 that are atypical for quasars at high redshifts were also reviewed. They could be the remaining stellar interlopers mimicking quasars in the MIR region of the spectra. Indeed, four IR Galactic objects were identified out of a total of 20, as discussed in Appendix A.1. No additional filtering has been applied, because of the low rate of identifiable false positives.

4. Selection of CDAGNs Resolved in Gaia EDR3

The second group of CDAGNs (hereafter Sample 2) comprises quasars and AGNs that have been directly resolved in Gaia EDR3, irrespective of their redshift. It is straightforward to select all MIRAGN sources with near neighbors. The main problem is how to avoid stellar companions overwhelming this selection, given that the majority of the optical pairs include stars. As discussed in Makarov & Secrest (2022), the number of targets without companions within 11″ can be used to estimate the effective number density of the sources in Gaia EDR3. If N sources are randomly positioned on the celestial globe, the probability of not having at least one companion within an angular radius r (in radians) is

\[ P_{\text{empty}}(r) = \left(1 - \frac{\pi r^2}{4\pi}\right)^N. \]

This probability is accurately estimated as the ratio of crossmatched MIRAGN objects without any companions to the total number of targets. For r = 11″, we obtain N = 3.246 \times 10^8. We can now compute the expected probability of having a random companion within r = 2″, 1 - P_{\text{empty}}(r = 2″) = 0.0076. We note that such random companions, i.e., optical pairs, would mainly be Galactic stars, because stars dominate in the Gaia catalogs. However, the estimated rate of stellar contaminants within 2″ may be biased upward with respect to the modal value, by the strongly asymmetric distribution of the rate on the celestial sphere. Indeed, while 79% of quasars do not have any companions within r = 11″, 0.0011% of the sample have 16 neighbors within the same area. A small fraction of sources in crowded areas (closer to the Galactic plane or projected against globular clusters, etc.) is responsible for a large fraction of chance alignments. We devised a selection approach leading to the inevitable loss of genuine double quasars, but producing fewer optical pair contaminants.

A total of 113,241 MIRAGN objects have at least one companion within 11″, with 98,432 having only one or two neighbors within this radius. The remaining sources are mostly in areas with high local number density, and their multiple neighbors are likely stars. To eliminate such crowded areas, we filter only sources with neighbors within 5″, but without any neighbors between r = 5″ and 11″. This filtering should remove 1 - 5″/11″ \sim 81% of the chance alignments in crowded areas, leaving 12,049 candidate sources. For this search of dual AGNs, we are only interested in the tighter separations. The Gaia EDR3 resolving capabilities become impacted at separations below 2″ (Fabricius et al. 2021), which is our chosen upper limit for inclusion in the CDAGN catalog. The number of selected sources is 2363.

This selection procedure is expected to remove 1-2″/11″ \sim 96% of stellar contaminants randomly aligned with quasars within 2″. However, if the rate of physical dual quasars within 2″ is much lower than the rate of chance stellar neighbors, the selected sample may still include many false positives. To estimate the rate of stellar contamination, the normalized parallaxes of the primaries (quasars) and all the nearest companions are investigated. The distributions are shown in Figure 6. The cut for the primaries is just one of our filters applied at an earlier stage, to discard some of the stellar interlopers. Both distributions are sharply peaked at \approx 0, although the histogram of the neighbors seems to be slightly shifted to a positive normalized parallax. We find a weak sign of a secondary bump at \approx 4 for the neighbors, which may be caused by the remaining stellar companions. The total number of neighbors with \approx 4 is 144. Very approximately, the actual number of stellar contaminants may be twice this number, if the parallaxes are perturbed symmetrically around the modal value. The estimated rate of stellar companions from the observed distributions of parallax is 12%. We do not use the parallaxes of resolved neighbors to clean the final selection, however, because this astrometric parameter is often not available at close separations.

As an additional and independent verification method, we employ near-neighbor statistics. The intermediate sample of 12,049 MIRAGN–Gaia sources with at least one neighbor within 5″, but no neighbors inside the ring between 5″ and 11″, is used. As previously discussed, this selection favors sources that are located in low-density areas, which greatly improves the reliability of the statistical estimation. The sample sources have 12,713 neighbors, i.e., most of them have only one neighbor. Figure 7 displays the histogram of the separations for this sample, with a bin width of 0″/2. The absence of neighbors closer than \approx 0″/6 is caused by the hard limit on angular separation in Gaia. The straight line originating at the limiting separation represents the rate of random neighbors, assuming that the neighbors between 4″ and 5″ are all random. We see a definite excess of neighbors at separations below \approx 3″. Counting the number of neighbors above the expected rate with separations \approx 2″, the estimated fraction of statistical interlopers is 46%, leaving 54% of Sample 2 as physical dual sources. The estimate for the rate of stellar interlopers is much

\[ \text{https://gea.esac.esa.int/archive/documentation/GDR3/Gaia_archive/chap_datamodel/sec_dm_main_source_catalogue/sec_dm_gaia_source.html} \]
higher than the previously obtained 12%, possibly because the near-neighbor estimation is overly conservative. It is known that the resolving capabilities of Gaia EDR3 are strongly degraded already at separations below $\lesssim 2''$ (Fabricius et al. 2021). The actual rate of chance neighbors is a concave function, rather than a straight line. Also, some of the neighbors at $\gtrsim 4''$ may still be genuine double AGNs, because Sample 2 is not limited to high-redshift sources.

5. The Merged CDAGN Catalog

The two output samples of sources from the high-$z$ selection with elevated phot_bp_rp_excess_factor values (Sample 1, 1023 sources) and objects with Gaia-resolved neighbors within 2" (Sample 2, 2363 sources) are merged to yield the final CDAGN catalog. At this final stage of catalog production, we identified 23 sources in Sample 1 that have $z_{\text{obs}} - z_{\text{pre}} > 2$. Visual inspection of the images available through the online Pan-STARRS cutout service and the SDSS spectra available through the online SDSS DR16 service revealed that many of these cases in fact resulted from gross errors in the SDSS redshifts, as discussed in Appendix A.2.

Furthermore, 219 objects were found to be present in both samples, and the duplicates were removed.

At the final stage of filtering, we checked the CLASS metadata field, which contains the spectroscopic classification determination for those sources that are in the SDSS survey. Four sources emerged with a classifier STAR in that field. Source J081308.58+480643.2 at position (123°28584401, 48°11200684) has a 16.8 mag companion resolved by Gaia at 1°8, which is indeed a fast-moving white dwarf. Several decades ago, it was much closer to the object of interest, because its proper motion is generally away from it. The 1°042 companion of J095438.95+443356.3 is brighter than the primary target by about 1 mag and is also roughly moving away from it. Spectroscopically classified as a carbon star, the spectrum may be of composite nature, with broad absorption features and a single powerful emission line. The source J154653.67+573533.9 is an interesting example of a legitimate quasar classified as BAL by Trump et al. (2006), whose spectrum could be contaminated by a likely stellar companion separated by 1°511. Some of the objects in the intriguing class of quasars with flat and featureless spectra may be such blends. The 1°797 companion to the quasar J223223.70+135434.6 is a high-proper-motion star (42 mas yr$^{-1}$) of unknown type. After the removal of the four identified stellar contaminants, the final merged sample of 3140 sources constitutes the CDAGN catalog, which is published online.

Table 1 describes the format of the CDAGN catalog. Not all of the values are available in all of the columns. The phot_bp_rp_excess_factor parameter is not available in Gaia EDR3 for 479 sources, which come from Sample 2, as phot_bp_rp_excess_factor was not used in the selection process. The spectroscopic redshifts in column 9 are not available for 2814 sources. The ML-predicted and ML-classified redshifts in columns 10 and 11 are not specified for 521 objects originating from Sample 2. The separations in column 12 are not available for the entire Sample 1 contribution of 781 objects. The $(0.158655, 0.841345)$ quantiles of $z_{\text{obs}} - z_{\text{pre}}$ for 301 sources with both values of redshift available, which correspond to $\pm \sigma$ for a Gaussian distribution, are $\{-0.24, +0.51\}$. The strong asymmetry is caused by the skewed sample distribution having a heavy tail toward positive differences.
6. Discussion and Summary

The published CDAGN catalog includes 3140 objects. One-quarter of its entries (781) have \(-1\) in column 12 for separation to the nearest companion, i.e., these objects come from Sample 1 and are not resolved by Gaia at small angular separations. The main risk associated with this part of the catalog is that some of the Sample 1 objects are not dual or double at all, and the elevated level of \textit{phot\_bp\_rp\_excess\_factor} is caused by other circumstances. Indeed, a small number of Galactic planetary nebulae (PNs) and Herbig Ae/Be stars have sneaked into our selection (Appendix A.1), with the very high values of \textit{phot\_bp\_rp\_excess\_factor} apparently being triggered by the presence of bright emission structures surrounding the central stars. Our ML redshift determination produced incorrect results for such exotic objects because of the unusual photometric properties (large MIR excess coupled with blue optical colors). Several percent of this sample may also be nearby AGNs with extended image components, such as substrate galaxies, which propagated into the final catalog due to “regular” errors of the ML redshift prediction and classification. A smaller fraction represents nearby AGNs with grossly incorrect spectroscopic redshifts (Appendix A.2). The user is advised to pay attention to discrepant values of \textit{obs} redshift and \textit{spec} redshift, which may indicate such problems. Given these risks and the relatively small representation of Sample 1, one may ask if its inclusion is warranted. Our motivation is that this part of the catalog may capture candidate dual sources at separations beyond the current angular resolution capabilities of Gaia EDR3. The smallest separation recorded in the catalog is 0.0388; however, the resolving capabilities of Gaia EDR3 are known to be severely degraded at separations below 1. We hope that this selection captures some dynamically evolved dual AGNs that have already converged to the gravitational potential well minimum, which is a barely known area at present. Checking extant and collecting new high-quality images of CDAGNs is also quite helpful in avoiding any remaining contaminants of Sample 1 provenance, because the extended structures surrounding AGNs at \(z < 0.5\) are often clearly visible.

A higher degree of reliability for the CDAGNs from Sample 2 appears to be guaranteed by the fact that these sources have already been resolved in Gaia EDR3 with separations within 2′. However, a double source is not necessarily a dual AGN. The fraction of optical pairs that must be, on statistical grounds, mostly chance alignments with foreground stars, is difficult to estimate accurately. We use two indirect methods of estimation, which produce somewhat different results. This fraction amounts to roughly 12% from the apparent distribution of parallaxes for the near neighbors and to 46% from the general near-neighbor distance considerations. Very conservatively, we conclude that more than half of the resolved sources in Sample 2 are genuine dual—i.e., physical—AGNs. As for the overall rate of double sources (including the remaining stellar companions), we note that 2359 of 3140 sources come from Sample 2 (that is, they are already resolved at separations < 2′), while not more than \(\sim 8\%\) of the remaining 781 objects originating from Sample 1 have redshifts smaller than 1, which may trigger a false positive. Therefore, not more than 2% of the final catalog is expected to constitute single sources that are relatively nearby.

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

This research has made use of Astropy (http://www.astropy.org), a community-developed core Python package for Astronomy (Astropy Collaboration et al. 2013, 2018), and TOPCAT (Taylor 2005).

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions.

SDSS-IV acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is www.sdss.org.

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics | Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für
The process of compiling the CDAGN catalog and verifying its contents, we have encountered objects of special interest and oddities of different kinds. The frequency of such cases is elevated in CDAGNs, because the filters that are applied tend to select most perturbed objects with unusual properties, which are otherwise quite rare in the original MIRAGN sample. In this section, some of the cases are listed and briefly discussed, irrespective of whether they have been included in the final catalog or rejected.

A.1. Stellar Objects Mimicking Quasars

A small fraction of Galactic stars is present in the initial MIRAGN/EDR3 sample. These are very rare objects with unusual photometric properties, namely, extremely red W1–W2 and W2–W3 colors. Some of them are also associated with marginally high \( \text{phot}\_bp\_rp\_excess\_factor \) values, as we discovered in the process of visual inspecting the most conspicuous cases.

1. Source J200236.33+173650.6 at position (300°56141639, 17°61410948) is the known planetary nebula (PN) G056.8–06.9, which is also known as K 3–51 (Acker et al. 1992). Gaia EDR3 lists multiple sources around this object within 10′, which are likely triggered by clumps in the nebula. The nebula has a ringlike shape, with a looped filament and a nearby small structure, which may be a very red companion. The \( \text{phot}\_bp\_rp\_excess\_factor \) value is exceptionally high at 2.73, possibly due to the optically bright extended image of the nebula.

2. Source J070006.62+121441.5 at position (105°02763308, 12°2479321) is the PN G202.9+07.4 = KN 60, of remarkable bipolar structure, and a nearby companion separated by 1″332 from the central star, recently identified in the Gaia data (Chornay & Walton 2020; González-Santamaria et al. 2021). The combination of an optically bright two-lobe nebula stretching over at least 7″ and a relatively bright neighbor produced one of the highest values \( \text{phot}\_bp\_rp\_excess\_factor = 4.99 \) in the initial sample. This object has previously been misidentified as an AGN in a number of papers.

3. Source J034732.98+350248.8 at position (56°88742456, 35°04684135) is the well-known IC 351 (e.g., Kohoutek 2001), of ringlike elongated shape and a few clumps, which resulted in three bogus companions in Gaia EDR3 being resolved within 3″. Like many other PNs, it is a rather bright radio source with a peak flux of ~30 mJy listed in the NVSS (Condon & Kaplan 1998).

4. Source J184530.00-251214.6 at position (28°37501212, −25°20409959) is an apparently new PN not listed in SIMBAD or any relevant catalogs. It has a remarkable bipolar structure and an hourglass shape stretching over at least 7″. As with the previous cases, Gaia EDR3 resolves a few close neighbors, which are likely some clumps within the bright nebula.

5. Source J181933.56+035449.6 = IRAS 18170+0353 at position (274°88982401, 3.91378495) is an extreme IR object and a known SiO and OH maser (Deguchi et al. 2010; Engels & Bunzel 2015). It has a stellar morphology, without any resolved structure, in the available images. Gaia EDR3 presents a statistically significant proper motion for this object, while the measured parallax is statistically consistent with zero.

6. Source J060654.83+203916.2 at position (91°72849084, 20°65453833) is not listed in SIMBAD, but it has been incorrectly included in a number of AGN lists, as well as in the catalog of suspected young stellar objects (YSOs) by Marton et al. (2016). It has recently been classified as a new Herbig Ae/Be star by Vioque et al. (2019), using an ML approach. Composite Pan-STARRS images suggest the presence of a shell-like structure around the central point source. This object seems to be an example of an apparently isolated young star, as there are no other known YSOs in the close vicinity.

7. Source J185555.28+081701.7 at position (283°98035277, 8°28384173) is another seemingly isolated Herbig Ae/Be object not listed in SIMBAD (Vioque et al. 2019). Composite Pan-STARRS images reveal at least three faint companions, which may be bright nebulae or protostar cores.

8. Source J004449.74+632251.0 = IRAS 00418+6306 at position (11°20729618, 63°38083536) has been identified as a Herbig Ae/Be star (Vioque et al. 2019) and an OH maser (Engels & Bunzel 2015). There is a hint of a blue shell around the star in composite images.

9. Source J160028.60-562843.9 = IRAS 15564-5620 at position (240°11916884, −56°47886695) has been identified as
a Herbig Ae/Be star (Vioque et al. 2019) and an OH maser (Engels & Bunzel 2015). It is listed in Gaia EDR3 with a small parallax 0.23 mas, but a surprisingly high proper motion 6.38 mas yr\(^{-1}\). Its color is an outstanding \(G_{\text{BP}}-G_{\text{RP}} = 4.00\) mag. Its location in a relatively crowded part of the sky close to the Galactic plane, plus the presence of other very red objects with similar proper motions, suggest a young association or a star-forming region.

A.2. Quasars with Erroneous SDSS Spectroscopic Redshifts

This category of sources came under our consideration as a result of the procedures used in collecting Sample 1. This was initially based on ML-predicted and ML-classified redshifts above 1, but we also decided to back-substitute all objects from the SDSS sample with spectroscopic redshifts above 1. There may be several cases of greatly overestimated SDSS redshifts in the CDAGNs catalog. A possible explanation is the blending of relatively nearby galaxies with AGNs with a well-aligned foreground star.

1. Source J003005.38+295708.0 at position (7°525241384, 29°95221128) has a continuum spectrum peaked at approximately 550 nm and a single broad emission line at 755 nm, which was incorrectly interpreted as a Ly\(\alpha\) line in the SDSS pipeline. This radio-loud AGN undoubtedly resides in a relatively nearby spherical galaxy, as revealed by the available images. The spectroscopic SDSS-determined redshift 5.2 is incorrect, while the ML-predicted redshift 0.15 is closer to the truth. This implies that the dominating emission feature is the H\(\alpha\) line.

2. Source J142813.79+311417.1 at position (21°705751109, 31°23810507) presents a similar case to the previous object. Its continuum is not normal for a quasar rising toward the red end, and it is interspersed with narrow absorption and emission lines. The single broad emission line at 740 nm has been misinterpreted as a Ly\(\alpha\) line, leading to a redshift value of 5.1. Pan-STARRS images show an extended host galaxy, with a hint of some structure in the color distribution. The ML-predicted redshift is 0.50.

3. Source J083333.66+234943.6 at position (128°39027 1593514, 23°8287737329836) has a stellar-like spectrum, with a bump in the blue part and a single very broad and complex emission line at 930 nm. Interpreted as a Ly\(\alpha\) line, it leads to one of the highest SDSS-determined redshifts, of 6.7. The ML-predicted redshift is 0.64.

A.3. Known and New Gravitational Lenses

The filtering procedures used for the selection of Sample 2, of resolved CDAGNs (Section 4), selected objects with a few tightly packed images within a circle of 5″ radius, but no close neighbors outside this circle. This favored multiply imaged strong gravitational lenses, which are rare and important objects in their own right. It is not surprising that we find an elevated rate of lensed images following a limited visual inspection of the Sample 2 objects. The list below includes only some of the objects in this category, with short descriptions as appropriate.

1. Source J111816.92+074558.5 = QSO 1115+080 at position (169°57066749, 7°76625047) is a known lensed system of images (Cao et al. 2012; Williams et al. 2017; Ducournet et al. 2018; Delchambre et al. 2019) at \(z = 1.73\). Available Hubble Space Telescope NICMOS/NIC2 images show the complex configuration, with a row of closely packed images on the east side and two wider and fainter images on the west side of the lens, which is barely discernible in the Pan-STARRS images. Gaia has resolved four of these images.

2. Source J125149.16-122217.1 at position (12°954872187, −12°37147381) is likely to be a previously undetected gravitational lens. Composite Pan-STARRS images reveal an extended image that is probably composed of at least three images, with a complex color distribution. Gaia resolves only two sources separated by 0″9, however, and the northern companion has a large measured proper motion of 8.4 mas yr\(^{-1}\). The ML-predicted redshift is 1.11.

3. Source J205144.29+252736.9 at position (31°293464628, 25°46031899), with three images resolved within 5″ in Gaia, has been selected as a candidate strong lens by Delchambre et al. (2019). However, Pan-STARRS images show that the two closely separated (0″703) components have very different colors, with the blue one being the quasar and the red one at position angle 204° potentially being an interloping star. The astrometric measurements for the red companion are quite uncertain. The quasar with \(z_{\text{pec}} = 2.58\) is known to be optically variable (Usatov 2018), so follow-up observations on the light curves of the components could confirm or rule out this suggested lens.

4. Source J054535.89+582805.8 at position (86°39963598, 58°468300681) is likely a new gravitational lens. It has not been selected by previous Gaia-based search engines, because only two close images have been resolved at separation 0″817—however, more seem to be present in the available images. The measured proper motion of the northern component is a high 9.06 mas yr\(^{-1}\), but the Gaia proper motions are known to be strongly perturbed beyond the error budget by close unresolved multiplicity (Makarov et al. 2017a).

5. Source J053530.35+091052.5 at position (88°27226393, 9°18136749) is similar to the previous case, in that only two images have been resolved in Gaia EDR3, at separation 1″106, but the Pan-STARRS images suggest more components. It may be a new gravitational lens.

6. Source J121547.10+293409.8 at position (183°94628002, 29°56940657) was selected by Inada et al. (2012) as a candidate lensed quasar, based on SDSS data. Gaia EDR3 resolved this object into two images at one of the closest separations 0″469.

ORCID iDs

Valeri V. Makarov @ https://orcid.org/0000-0003-2336-7887
Nathan J. Secrest @ https://orcid.org/0000-0002-4902-8077

References

Acker, A., Marcout, J., Ochsenbein, F., et al. 1992, The Strasbourg-ESO Catalogue of Galactic Planetary Nebulae. Parts I, II (Garching: European Southern Observatory)
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, \AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, \A&A, 558, A33
Cao, S., Covone, G., & Zhu, Z.-H. 2012, \ApJ, 755, 31
