MNRAS 000, 1–12 (2020) Preprint 29 January 2021 Compiled using MNRAS LATEX style file v3.0

A classifier for spurious astrometric solutions in Gaia EDR3

Jan Rybizki[1]★, Gregory M. Green[1], Hans-Walter Rix[1], Markus Demleitner[2], Eleonora Zari[1], Andrzej Udalski[3], Richard L. Smart[4], and Andy Gould[1,5]

[1]*Max Planck Institute for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany*
[2]*Astronomisches Rechen-Institut, Zentrum für Astronomie der Universität Heidelberg, Mönchhofstrasse 12-14, D-69120 Heidelberg, Germany*
[3]*Astronomical Observatory, University of Warsaw, Al. Ujazdowskie 4, 00-478 Warszawa, Poland*
[4]*INAF - Osservatorio Astrofisico di Torino, via Osservatorio 20, 10025 Pino Torinese (TO), Italy*
[5]*Department of Astronomy, Ohio State University, 4055 McPherson Laboratory, 140 West 18th Avenue, Columbus, Ohio 43210, USA*

arXiv:2101.11641v1 [astro-ph.IM] 27 Jan 2021

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The Gaia mission is delivering exquisite astrometric data for 1.47 billion sources, which are revolutionizing many fields in astronomy. For a small fraction of these sources the astrometric solutions are poor, and the reported values and uncertainties may not apply. For many analyses it is important to recognize and excise these spurious results, commonly done by means of quality flags in the Gaia catalog. Here we devise and apply a path to separating 'good' from 'bad' astrometric solutions that is an order-of-magnitude cleaner than any single flag: we achieve a purity of 99.7% and a completeness of 97.6% as validated on our test data. We devise an extensive sample of manifestly bad astrometric solutions: sources whose inferred parallax is *negative* at $\geq 4.5\sigma$; and a corresponding sample of presumably good solutions: the sources in HEALPix patches of the sky that do not contain extremely negative parallaxes. We then train a neural net that uses 14 pertinent Gaia catalog entries to discriminate these two samples, captured in a single 'astrometric fidelity' parameter. An extensive and diverse set of verification tests show that our approach to assessing astrometric fidelity works very cleanly also in the regime where no negative parallaxes are involved; its main limitations are in the very low S/N regime. Our astrometric fidelities for all EDR3 can be queried via the Virtual Observatory. In the spirit of open science, we make our code and training/validation data public, so that our results can be easily reproduced.

Key words: Galaxy: stellar content, Galaxy: kinematics and dynamics, software: public release, space vehicles: instruments, virtual observatory tools

1 INTRODUCTION

Parallax measurements contain information about the distance of astrophysical objects, and are critical to anchoring the cosmic distance ladder. At the same time, kinematic measurements – proper motions and radial velocities – provide phase-space information that is key to understanding Milky Way dynamics and external galaxies. The 1.47 billion astrometric measurements reported in *Gaia* Early Data Release 3 (Gaia Collaboration et al. 2020a, "EDR3") constitute the largest astrometric dataset ever produced.

While this astrometric catalog is of extremely high quality (Lindegren et al. 2020b), a significant fraction of astrometric solutions are spurious (Fabricius et al. 2020). Spurious astrometric solutions should are a distinct issue from negative parallaxes, which are an expected outcome of the normally distributed parallax measurement (Bailer-Jones 2015; Luri et al. 2018). Spurious solution have a biased parallax value with an incompatible parallax uncertainty reported, due to specific failure modes (Fabricius et al. 2020). This can be a particular concern when looking at sparsely populated portions of the color-magnitude diagram, or at extreme objects, such as the nearest or fastest-moving stars (i.e., those with the largest parallaxes or proper motions, respectively). For example, naively selecting all objects with measured parallaxes greater than 10 mas (corresponding to a distance of less than 100 pc) yields a catalog with an estimated 50 % of spurious parallax measurements. *Gaia* EDR3 provides a number of astrometric quality parameters that can be used to exclude such spurious solutions. The "Gaia Catalogue of Nearby Stars" (Gaia Collaboration et al. 2020b, "GCNS") uses a combination of these parameters to filter out spurious sources, obtaining a highly complete and pure subset of *Gaia* EDR3 sources lying within 100 pc. In this paper, we use a similar approach to extend this work to the entire *Gaia* EDR3 catalog.

Gaia EDR3 provides 1.47 billion astrometric measurements containing of a two-dimensional position on the sky, a two-dimensional proper motion and a parallax (in addition, a 7.2M subset also has radial velocity measurements). There are many pos-

★ E-mail: rybizki@mpia.de

© 2020 The Authors

sible sources of excess noise in these astrometric measurements. Some error modes, such as unmodeled acceleration caused by an unresolved binary companion typically introduce small residuals into the astrometric solution, which will usually be accounted for in the parallax uncertainty estimate (Lindegren et al. 2020b). However, other error modes, such as incorrect epoch cross-matches with background or spurious sources and also close source pairs, which might be partially resolved (Fabricius et al. 2020), can introduce very large residuals, scattered around the true parallax(Gaia Collaboration et al. 2020b), which are unaccounted for in the reported parallax uncertainty. Spurious astrometric solutions mainly happen in very dense parts of the sky (Fabricius et al. 2020). It is this latter class of "catastrophic" errors in the astrometric solutions (leading to errors in excess of the stated uncertainties) that we will attempt to detect.

One can try to mitigate these spurious astrometric solutions with cuts on `ruwe`, visibility_periods_used and G magnitude. These cuts are known to exclude many valid sources and also bias the sky coverage. In the GCNS the approach was to use many and only astrometric quality indicators and train a random forest on a good and a bad training sample. Since only sources with observed parallax of greater 8mas were considered, this was operating in an extremely high parallax SNR regime. For the "bad" examples the sources with parallax < -8 mas were used, exploiting the fact that spurious astrometric solutions can be expected to scatter randomly around the true parallax. For the "good" examples sources in low density regions of the sky were used that were crossmatched to 2MASS and showed consistent absolute magnitudes in Gaia and 2MASS bands with main stellar populations.

When trying to classify the spurious parallax solutions for the whole GaiaEDR3 catalog we also need to make informed decisions for low parallax SNR as they constitute 85 % of the sources (for SNR < 4.5). Since parallaxes of low SNR less stringently constrain the distance, it is harder to establish a genuine difference between valid and spurious astrometric solutions. We aim to mitigate this by training specialised models for the high- and low-SNR regime.

In the following, we will attach a single scalar measure of "astrometric fidelity" bound between 0 and 1, to all sources in eDR3.

2 OPEN SCIENCE APPROACH

This preprint is a work in progress. We have created a classifier that we believe identifies cleanly sources with spurious parallaxes, and which performs significantly better than the simple cuts advocated in the existing literature. However, we are open to suggestions for improvement – in the classifier itself, the creation of the training datasets, and ideas for validating performance. We want to allow the community to use this classifier, and welcome feedback. We are open to offering co-authorship in the final journal-submitted version of this work for any significant contribution.

All of the work in this paper, including the training of the classifier and the various validation tests, can be redone using a `Python` notebook and data that we have made available.[1].

As we update the classifier, we will continue to keep the initial version of the astrometric fidelities (v1) in the corresponding Virtual Observatory (VO) table, but will add additional columns with updated probabilities. This will allow astronomers to redo their analysis with upcoming classifiers, and to compare results across different versions of the classifier.

[1] The notebook can be found at `https://colab.research.google.com/drive/1d4KCXiCyFzLF1RzTzRGRAnVS0Uc8x3RU?usp=sharing`, while the necessary data is stored at `https://keeper.mpdl.mpg.de/d/21d3582c0df94e19921d/`

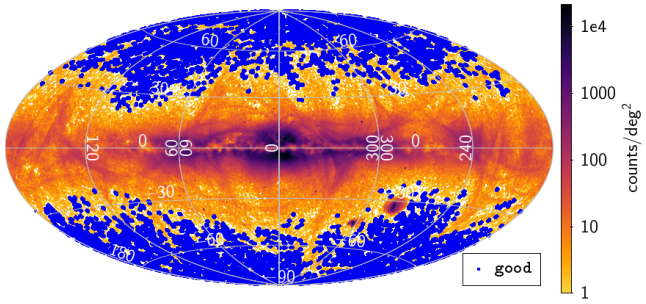

Figure 1. Density distribution of the sources identified as `bad` for our training sample by their $< -4.5\sigma$ (negative) parallaxes, shown using an Aitoff projection in Galactic coordinates (orange to black). In contrast, the regions of the sky from which we drew the `good` training sample, where such strongly negative parallaxes are absent, are shown in blue.

3 TRAINING SAMPLE GENERATION

In all of the following we neglect the zero-point parallax offset (Lindegren et al. 2020a). We name the training set for spurious (valid) astrometric solutions "`bad`" ("`good`").

3.1 Spurious sources

We construct our `bad` training sample by selecting sources with `parallax_over_error` < −4.5. We use the following query:

```
SELECT *
FROM gaiaedr3.gaia_source
WHERE parallax_over_error < -4.5
```

This returns 4.18 million sources. If all of the 1.47 billion sources in Gaia EDR3 with measured parallaxes had a true parallax of zero, and all of the measurement errors were Gaussian, then we would expect approximately 5000 stars – nearly three orders of magnitude fewer – to satisfy the above cut. Since in reality, sources have positive parallaxes, the discrepancy is even larger. Thus, even with the most pessimistic assumptions, the contamination rate of our `bad` training sample by sources with good astrometric solutions is ∼ 0.1%.

Fig. 1 shows the distribution of our `bad` sources over the sky. Dense areas such as the bulge, disc and the Magellanic clouds are apparent, but scanning law patterns are also visible, with regions of the sky that are scanned most often (notably the two rings along ecliptic latitude ≈ ±45°) having higher densities of spurious astrometric solutions. We conjecture that this is due to the many scans along a similar scanning angle, which increases the probability of spurious detections ocurring at the same place and therefore reducing the probability of being filtered out in the downstream process (Torra et al. 2020) for example due to `visibility_periods_used` < 9.

3.2 Good sources

In order to construct the good training set, we select all sources in regions of the sky that do not contain any sources with significantly negative parallaxes (in the above -4.5σ sense). This means that our good and bad training examples come from disjoint regions of the sky, as can be seen in Fig. 1. The good sample does not come from the Galactic plane (i.e., $|b| > 19^\circ$ for all good sources). In detail, we separately query each HEALPix level-6 pixel that contains no sources with `parallax_over_error` < -3.5. In all, 4197 out of 49152 pixels meet this condition. Our query for a single pixel is as follows:

```
SELECT *
FROM (SELECT dr3.*, tmass.tmass_oid,
     FLOOR(dr3.source_id/140737488355328) as hpx6
FROM gaiaedr3.gaia_source as dr3
JOIN gaiadr2.tmass_best_neighbour AS tmass
USING (source_id)
WHERE source_id BETWEEN 0 AND 562949953421311) AS
     subquery
-- Query only first HEALpix of level 6
JOIN gaiadr1.tmass_original_valid AS tm
   USING (tmass_oid)
-- only sources with a crossmatch to 2MASS are queried
```

We obtain a total of 5.24 million sources from the 4197 pixels we query in this manner. The requirement that the source is also visible in 2MASS (Skrutskie et al. 2006) ensures that we do not include spurious sources. It lowers the fraction of faint blue objects (e.g. white dwarfs) though, but this cut does not seem to propagate into our prediction, as we do not classify using photometric indicators.

We split our good training set into two subsets: a high-SNR subset with `parallax_over_error` (SNR) > 4.5 and a low-SNR sample with $-3.0 <$ SNR < 4.5. In order to further purify our good training set, we require $G - RP < 1.8$ mag (1.5 mag) for the high-SNR (low-SNR) subset. The cut on $G - RP$ requires RP photometry, which only excludes 10k sources. The 40k sources removed by this cut are unphysically red, as can be seen in the upper panel of Fig. 2. This extremely red color sometimes coincides with nearby sources and/or high `phot_bp_rp_excess_factor`. After these photometric cuts, our good training set contains 5.18 million sources.

As is apparent from the lower panel of Fig. 2, the high-SNR subsample of the good training set (in blue) resembles a low-extinction color-absolute magnitude diagram (CAMD), with many of the known subpopulations clearly distinguishable and almost no unphysical features. The small overdensity of sources between the main sequence (MS) and the white dwarf (WD) sequence are mostly due to erroneously high BP values from faint sources (Riello et al. 2020). It is also important to note that the cut in $G-RP$ does not exclude the reddest asymptotic giant branch (AGB) stars in $BP - RP$. As expected, the absolute magnitudes of the low-SNR subsample scatter to much brighter absolute magnitudes, due to their poorly constrained distance moduli. The low-SNR sample consists mainly of sources with apparent G between 15 and 20 mag.

In constructing the good sample, we do not cut on any astrometric flags that we use later as potential features in the classifier. Of course, our good and bad training sets probe quite different regimes, with the good training set coming mostly from low-extinction regions and sparse fields. Nevertheless, we hope (and later verify) that in the space of astrometric parameters and quality flags, our training sets cover the relevant feature space and will allow our classifier to have discriminative power over the entire sky, as was the case for the GCNS. However, we acknowledge that it would be desirable to add valid astrometric solutions from the Galactic plane to the good training sample.

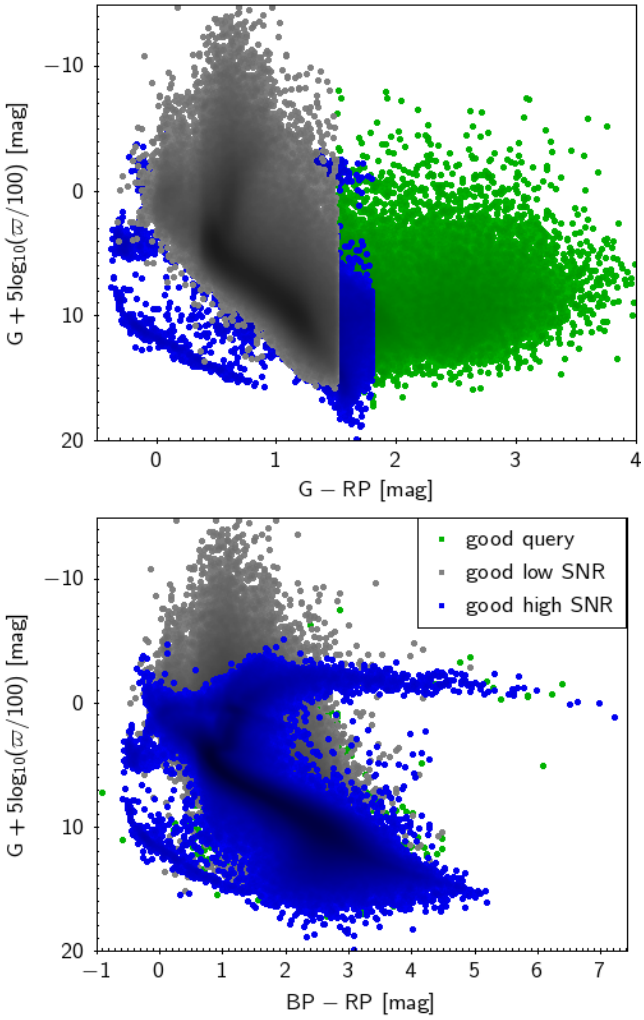

Figure 2. Color-absolute magnitude (CAMD) distribution for different versions of the good training sample, showing a naive estimate of the absolute G-band magnitude as a function of two different colors, $G - RP$ in the top panel, and $(BP - RP)$ in the bottom panel. The green points result from the initial query, before sources excessively red in $G - RP$ were removed. The gray and blue points show the low SNR and high SNR subsample, defined by $3 <$ `parallax_over_error` (SNR) < 4.5 and `parallax_over_error` (SNR) > 4.5, respectively. The plotting order is changed such that either the high-SNR subset in blue (bottom) or the low-SNR subset in grey (top) is fully visible.

3.3 Training features

We use the exact same features as in the GCNS (Gaia Collaboration et al. 2020b) marked in grey in their Table A.1. For completeness, we list them here in descending order of importance according to the Gini metric reported by the GCNS: `parllax_error`, `parallax_over_error`, `astrometric_sigma5d_max`, `pmra_error`, `pmdec_error`, `astrometric_excess_noise`, `ipd_gof_harmonic_amplitude`, `ruwe`,

visibility_periods_used, pmdec, pmra, ipd_frac_odd_win, ipd_frac_multi_peak and astrometric_gof_al. For parallax_over_error, which we abbreviate as SNR, both the GCNS and we use the absolute value as a feature.

3.4 Training for two different |SNR| regimes

We train two different classifiers, intended for use in the regimes |SNR| < 4.5 and |SNR| > 4.5. We will refer to these classifiers as the "low-SNR" and "high-SNR" classifiers, respectively. The most important difference between these two classifiers is that the high-SNR classifier uses |SNR| as a feature, while the low-SNR classifier does not. Recall that our bad training set does not include sources with $|SNR| < 4.5$. If we were to allow the low-SNR classifier to take |SNR| into account, it would learn that there are no bad sources with $|SNR| < 4.5$, which is simply an artifact of our method of identifying training data.

To train the low-SNR classifier, we use only training data with $|SNR| > 4.5$ (i.e., omitting the low-SNR good training examples). Excluding low-SNR training data while training the low-SNR classifier may seem counterintuitive, but our goal is to prevent the imbalance in coverage of SNR-space in the good and bad training sets from impacting our classifications in the low-SNR regime. In this regime, we end up with 3,964,264 good and 4,180,244 bad training examples.

To train the high-SNR classifier, we use the entire good and bad sets. In this regime, we end up with 5,184,555 good and 4,180,244 bad training examples.

3.5 Neural network training

In contrast to the work on GCNS, where spurious sources were identified using a random forest, we employ a feed-forward neural network (NN) here. Our NN model consists of 4 hidden layers, each with 64 neurons and a Rectified Linear Unit (ReLU) activation. The final layer has a single neuron with a sigmoid activation, and represents the probability that a source belongs to the good class. We use the binary cross-entropy loss function (e.g. Goodfellow et al. 2016), which is closely related to the Kullback-Leibler divergence and which measures how much additional information would needed to correct the classifier's prediction. Given input features $\vec{x}$, the classifier outputs a probability $P(\vec{x})$ that the source belongs to the good class. Denote true class (the label) by $y \in \{0, 1\}$ (where $y = 1$ signifies "good"). The binary cross-entropy is then given by

$$\mathcal{H} = -(1-y)\ln\left[1 - P(\vec{x})\right] - y \ln P(\vec{x}) \ . \quad (1)$$

As the binary cross-entropy is a measure of missing information, it can be expressed in units of bits or nats.

We implement our model in Tensorflow 2 (Abadi et al. 2016) and Keras (Chollet et al. 2015). We train for 100 epochs with an Adam optimizer (Kingma & Ba 2014), using a learning rate of 10^{-3} in the first 50 epochs, and a learning rate of 10^{-4} in the final 50 epochs. During training, we apply a dropout rate of 0.1 after each hidden layer in order to prevent over-fitting. The features are shuffled and normalised (to zero mean and unit variance) prior to the training, and we set 20% of the data aside for validation.

We train our high- and low-SNR classifiers separately. We assess our final performance by applying the low-SNR classifier to our low-SNR test dataset, and our high-SNR classifier to our high-SNR test dataset. On this combined test dataset, we achieve a loss of 0.0405 nats of entropy, with a purity of 99.7% and a completeness of 97.6%. On the high-SNR dataset, we achieve a binary cross-entropy of 0.0228 nats, with a purity of 99.6% and a completeness of 98.8%.

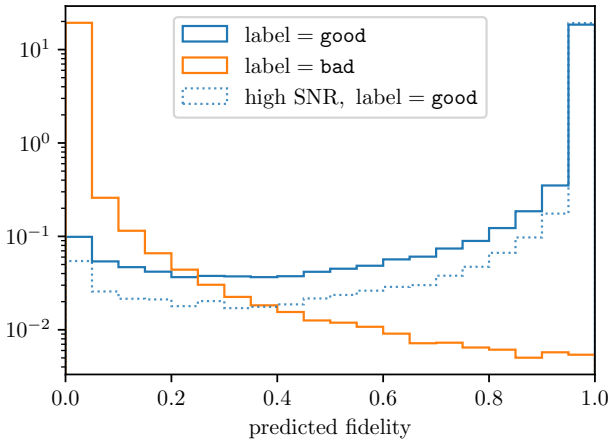

Figure 3. Histogram of the predicted classifier probabilities (of belonging to the good class) for sources in the test dataset, split by training label. As the good class contains both low- and high-SNR training data, we additionally show the classifier probabilities for the high-SNR good sources. The x-axis, is the probability output by the classifier that a given source is good, which we term the "astrometric fidelity".

In Fig. 3 we see the histograms of the astrometric fidelity prediction for the 20 % validation sources that are in the relevant SNR regime. In the top (low) panels we show the results for the high-SNR (low-SNR) model. On the left (right) side the astrometric fidelity predictions for the bad (good) sources are shown. As in the GCNS, the minimum for the truly good sources (in a training set sense) seems to be around a probability of 0.4, while for the truly bad sources the minimum is closer to 1. Because for us both classes are important we use an astrometric fidelity of 0.5 as the dividing value in the subsequent analysis. This means everything with a predicted astrometric fidelity > 0.5 will be attributed to good sources and everything with an astrometric fidelity lower than 0.5 will be attributed to bad sources. The user of the catalog can make cuts different from that in order to trade off completeness vs. purity of their sample.

3.6 Comparison with simpler classifiers

We compare the astrometric fidelity predicted by our neural network to analogous quantities obtained using simpler classifiers. First, we evaluate how cleanly simple cuts on ruwe and astrometric_excess_noise separate good and bad sources in the high-SNR test dataset. Fig. 4 shows the binary cross-entropy, purity and completeness of the cut as a function of the threshold value for each feature. For ruwe, we achieve a minimum binary cross-entropy of 1.61 nats using a cut of ruwe < 1.12, corresponding to a purity of 89.5% and a completeness of 89.0%. For astrometric_excess_noise, we achieve a minimum binary cross-entropy of 1.01 nats for a cut of astrometric_excess_noise < 0.708, corresponding to a purity of 93.2% and a completeness of 93.3%.

Next, we train a logistic model, which takes into account a

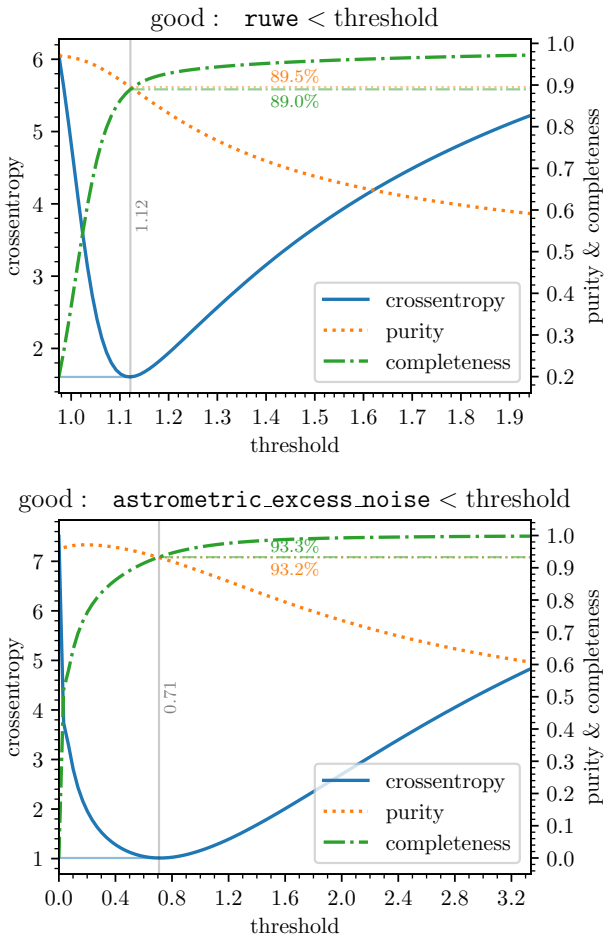

Figure 4. Performance of simple cuts on `ruwe` (top panel) and `astrometric_excess_noise` (bottom panel) in differentiating `good` and `bad` astrometric solutions. In each panel, we show how binary cross-entropy, purity and completeness depend on the threshold chosen for the cut. Our neural network predicting astrometric fidelity achieves an order of magnitude less contamination than the optimal choices for these cuts (at minimal cross-entropy).

linear combination of features. This model assigns a probability

$$P\,(\texttt{good} \mid \vec{x}) = \left[1 + e^{-(\vec{w}\cdot\vec{x}+b)}\right]^{-1} \tag{2}$$

of belonging to the `good` class to each source, where $\vec{x}$ is a vector containing the features, $\vec{w}$ is a vector containing a weight for each feature, and b, the bias, is a scalar. We use the Adam optimizer to find the weights and bias that minimize the binary cross-entropy of the predictions. On the high-SNR test dataset, we obtain a binary cross-entropy of 0.0960 nats, a purity of 96.4% and a completeness of 97.0%. This is better than what we achieve with simple cuts, but still represents more than three times the binary cross-entropy we obtain with the full neural network model.

The full neural network is not significantly more difficult to implement than these simpler classifiers, and it achieves a far more complete and pure separation of the validation data. For these reasons, we strongly favor use of the full neural network classification over simpler alternatives.

4 VALIDATION

We divide the validation for the two models in the regimes SNR $\geq$ 4.5 for the high-SNR model and SNR $<$ 4.5 for the low-SNR model.

Our classifier performs very strongly on the test dataset, which is statistically identical to the training data. In this section, we perform several tests using a variety of outside datasets, in order to determine whether our classifier generalizes beyond our training data. Here, the validation revolves around external information on the distance (e.g. membership in a cluster or the LMC), or the plausibility of the on-sky distribution, or simply external measurements (e.g. from OGLE).

4.1 Gaia Catalogue of Nearby Stars

We first apply our classifier to all sources in the "Gaia Catalogue of Nearby Stars" (GCNS). All 1.2 million sources with parallax $>$ 8 mas from eDR3 have classifications from the GCNS. Taking the GCNS classifications as a ground truth, our high-SNR model achieves a purity of 99.95% and a completeness of 99.28% on the sources with |SNR| $\geq$ 4.5 (comprising 91 % of the sample), while our low-SNR model achieves a purity of 90.8% and a completeness of 62.9% on the sources with |SNR| $<$ 4.5. Taking the GCNS classifications to be correct, our low-SNR model has lower performance than our high-SNR model. However, bad parallax determinations are fundamentally more difficult both to identify and to define in this regime, as the reported measurements are compatible with a very wide range of true parallaxes. However, note that we did not train our classifier on the GCNS sample, so it is not unsurprising that our low-SNR classifier performs worse on the GCNS dataset than on our own test dataset.

4.2 Clusters

Open and globular clusters, and the "prior" information on the distance of their likely member stars, offer a great opportunity to validate our parallax classifier. We begin with a catalog of 162,484 sources assigned to 121 clusters, coming from Bailer-Jones et al. (2018). This catalog was compiled using a method similar to that used in Gaia Collaboration et al. (2018). For each individual cluster, we calculate the variance-weighted mean parallax of its member sources, as well as the corresponding uncertainty in the mean parallax:

$$\langle\hat{\varpi}\rangle = \sum_i \frac{\hat{\varpi}_i}{\sigma_i^2}\,, \qquad \sigma_{\langle\hat{\varpi}\rangle} = \left(\sum_i \frac{1}{\sigma_i^2}\right)^{-1/2}\,. \tag{3}$$

We then select the 41 clusters for which $\sigma_{\langle\hat{\varpi}\rangle}/\langle\hat{\varpi}\rangle < 0.2$ (i.e., for which parallax is determined to better than 20%). For each source in each of these clusters, we then calculate a parallax residual, using the estimated cluster parallax as a reference:

$$\Delta\hat{\varpi} \equiv \hat{\varpi} - \langle\hat{\varpi}\rangle\,, \qquad \sigma_{\Delta\hat{\varpi}} = \left(\sigma_{\hat{\varpi}}^2 + \sigma_{\langle\hat{\varpi}\rangle}^2\right)^{1/2}\,. \tag{4}$$

The distribution of these parallax residuals (divided by the corresponding uncertainties) is shown in Fig. 5. Our classifier labels approximately 1% of sources in these clusters as `bad`. The standardized residuals of sources classified as `good` roughly follow the expected unit normal distribution, while the distribution of standardized residuals of the sources classified as `bad` is shifted negative and has much longer tails.

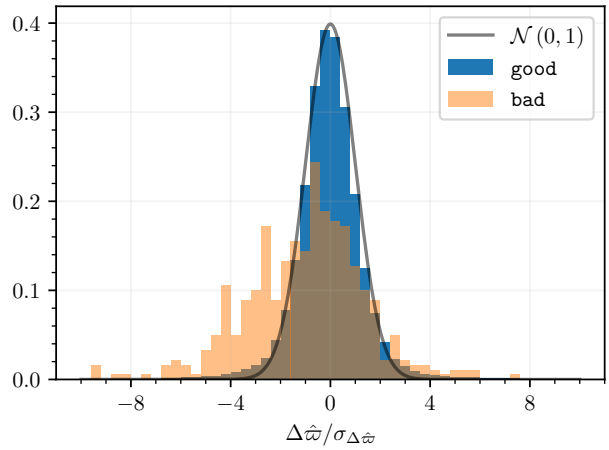

Figure 5. Validation of our astrometric fidelity prediction using open and globular clusters. The figure shows histograms of the standardized parallax residuals for good and bad sources. The true parallaxes are estimated using the variance-weighted mean of the parallaxes in each cluster. We restrict this comparison to clusters with distances determined to 20% or better. The good sources closely follow the expected unit normal distribution, in marked contrast to the standardized residuals of the bad sources.

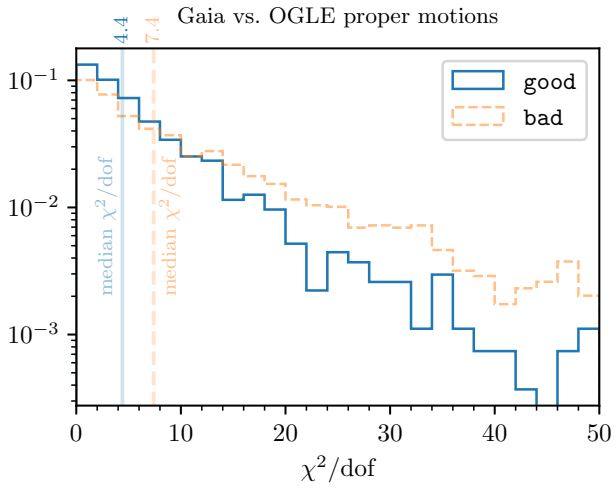

Figure 6. Validation of our astrometric fidelity classification through proper motion comparison between OGLE and Gaia EDR3. Shown is the distribution of χ^2/dof, based on a comparison of Gaia EDR3 and OGLE-IV proper motions in the Galactic Bulge, for sources labeled good and bad by our classifier. For ideal data, the median χ^2/dof would be ~ 0.69. We find that that proper-motion uncertainties are underestimated for Gaia EDR3 and/or OGLE-IV, leading to larger χ^2/dof values. However, for sources labeled good by our classifier, Gaia EDR3 and OGLE-IV proper motions match significantly better, as indicated by the lower median χ^2/dof values (4.4 for the good subsample, vs. 7.4 for the bad subsample).

4.3 OGLE proper motions

The Fourth Phase of the Optical Gravitational Lensing Experiment (OGLE-IV, Udalski et al. 2015) began observing the bulge of the Milky Way in 2010. Here, we validate our classifier using sources with proper-motion measurements from both OGLE-IV (OGLE Uranus astrometry project, Udalski et al. 2021, in preparation) and Gaia EDR3. Our assumption is that objects with spurious parallax determinations in Gaia EDR3 are more likely to have spurious proper-motion determinations. This should be reflected in the proper-motion residuals between Gaia EDR3 and OGLE-IV, with sources classified as bad in Gaia EDR3 having systematically higher χ^2 values in this comparison. We begin with a catalog of OGLE-IV sources with proper-motion measuresments, lying in a 0.15 deg × 0.15 deg box centered on $(\alpha_{\rm J2000}, \delta_{\rm J2000}) = (271.761\,\text{deg}, -26.698\,\text{deg})$. Using a matching radius of 0.2″, we obtain 14125 matching Gaia EDR3 sources with measured proper motions. Our classifier labels 2288 of these sources good.

We calculate the proper-motion residuals, $\Delta\vec{\mu} \equiv \vec{\mu}_{\rm Gaia} - \vec{\mu}_{\rm OGLE}$, as well as the covariance matrix of the residuals, $C_{\Delta\vec{\mu}} = C_{\mu,\rm Gaia} + C_{\mu,\rm OGLE}$. We then calculate $\chi^2 = \Delta\vec{\mu}^T C^{-1}_{\Delta\vec{\mu}} \Delta\vec{\mu}$ for each source. If the uncertainties are well estimated and the residuals follow a Gaussian distribution, then the χ^2 values that we obtain should follow a χ^2 distribution with two degrees of freedom. However, we find that the resulting χ^2 values are significantly larger, on average, than expected, both for sources labeled good and bad, indicating that Gaia EDR3 and/or OGLE-IV proper-motion uncertainties are underestimated in the Galactic Bulge. One could attempt to address this problem by inflating the uncertainties by a constant factor or by introducing a systematic error floor. However, these different methods of "correcting" the proper-motion uncertainties impact the distributions of χ^2 values obtained for the good and bad sources differently, as the good sources tend to have smaller estimated proper-motion uncertainties than the bad sources. In order to avoid these difficulties, we restrict our comparison to sources in a relatively small range of estimated proper-motion uncertainties, for which we assume the true proper-motion uncertainties to be similar. In particular, we select sources with $0.1\,\text{mas yr}^{-1} < \left|C_{\Delta\vec{\mu}}\right|^{1/4} < 0.2\,\text{mas yr}^{-1}$, obtaining 1192 sources labeled good and 1978 sources labeled bad. The resulting distributions of χ^2 values are displayed in Fig. 6. We find that sources labeled good by our classifier tend to have significantly lower χ^2 values than those labeled bad, with the median χ^2 per degree of freedom (dof) for the good subsample being 4.4, and the median χ^2/dof of the bad subsample being 7.4.

4.4 Large Magellanic Cloud

In the direction of the Large Magellanic Cloud (LMC), the vast majority of sources should be at a distance of ~50 kpc, corresponding to a parallax of 0.02 mas. This affords us another opportunity to validate our classifications, as almost all stars labeled good in this region of the sky should have reported parallaxes consistent with 0.02 mas. We expect the bad sources to have larger than reported residuals, and to scatter equally to positive and negative parallaxes, leading to a widened distribution of reported parallaxes centered on 0.02 mas.

We query a 0.25 deg cone in Gaia eDR3, centered on Galactic coordinates $(\ell, b) = (280.47\,\text{deg}, -32.88\,\text{deg})$, obtaining 252,115 sources, which we then run through our classifier. In this densely crowded region of the sky, only 11.2% of all sources are classified as good, while only 0.7% of high-SNR sources are classified as good. In order to model the small number of Milky Way foreground stars in this field, we compare to a control field of the same apparent size with the same Galactic latitude, and longitude reflected around $\ell = 0\,\text{deg}$. This control field has 1589 sources. In the control field, 71.6% of the sources are classified as good.

Fig 7 shows the parallax distribution of good and bad sources

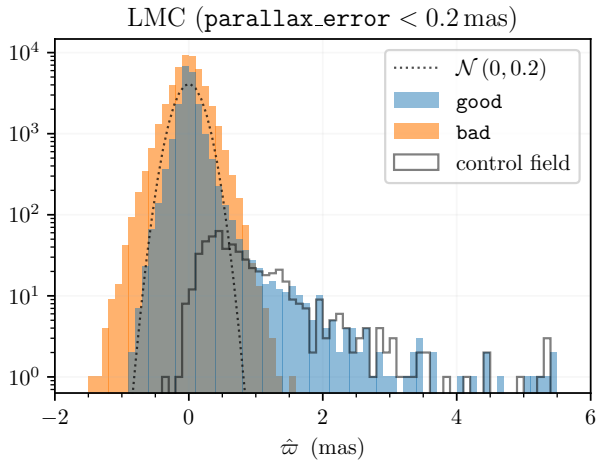

Figure 7. Validation of our astrometric fidelity classification through the parallax distribution of good and bad sources with small parallax uncertainties ($\sigma_{\hat{\varpi}} < 0.2$ mas) towards the LMC. The parallax distribution of good sources is consistent with a large population of distant ($\hat{\varpi} \approx 0$) sources (approximated by a normal distribution with zero mean and a standard deviation of 0.2 mas), along with an expected population of foreground stars (matching the distribution of parallaxes in a control field). In contrast, the bad parallaxes are consistent with a distant population of stars with parallax uncertainties that are underestimated by ~50%.

with small reported errors (parallax_error < 0.2) in our LMC field. The parallax distribution of good sources is consistent with a distant population of stars with well-measured errors, plus a small foreground population of Milky Way stars at larger parallax (matching the control field). The bad sources are consistent with a distant population of stars with significantly underestimated parallax errors. Our classifier is thus clearly identifying sources with excess parallax residuals, and even in this dense field, is still cleanly identifying foreground stars.

4.5 The Structure of the Galactic Disk in OBA stars

The catalog of O-, B-, and A-type (OBA) stars devised by Zari et al. (2021, subm.) offers another opportunity to test our classifier with an ensemble of sources at *low Galactic latitudes*. Zari et al. (*in preparation*) select stars brighter than $G = 16$ mag, with *Gaia* EDR3 and 2MASS colors consistent with (reddened) OBA-type stars. Zari et al. do not apply any condition on the parallax error, as the sample was designed to be inclusive for spectroscopic follow-up. We run our classifier on the resulting catalog consists of ~1 million stars. Fig. 8 shows the distribution of good (left, ~75% of the initial sample) and bad (right) sources in the Galactic plane ($|b| < 25°$). The distribution of sources with good astrometric solutions shows known regions of young stars and traces the spiral arm structure of the Milky Way disk, as discussed in Zari et al. (cf. their Fig. 11). The distribution of sources with bad astrometric solutions shows a ring-like feature between 2 and 3 kpc, which is physically implausible and hence presumably spurious. This is expected, as the parallax distribution of all sources in the OBA catalog peaks at around 0.3 mas (~3 kpc).

4.6 The Good/Bad Sky Distribution in Parallax Bins

As a final approach to validation, we visually inspect the projected sky distribution and CAMDs of Gaia EDR3 sources classified as good and bad in narrow bins of (catalog-reported) parallax. We refer to the parallax bins by their corresponding nominal distances. The 100 pc sample consists of the 1.2 million sources with $\hat{\varpi} > 8$ mas (the GCNS sample). The 300 pc sample consists of the 1.3 million sources with 3.3 mas $< \hat{\varpi} <$ 3.4 mas, the 1 kpc sample (1 mas $< \hat{\varpi} <$ 1.1 mas) contains 3.2 million sources, the 3 kpc sample (0.333 mas $< \hat{\varpi} <$ 0.334 mas) contains 1.3 million sources, the 10 kpc sample (0.1 mas $< \hat{\varpi} <$ 0.101 mas) contains 1.2 million sources, and the 30 kpc sample (0.0325 mas $< \hat{\varpi} <$ 0.034 mas) contains 1.4 million sources.

4.6.1 *High-SNR model*

Here we only look at sources in each parallax bin that have |SNR| ≥ 4.5, which results in strong selection effects. Fig. 9 shows the projected sky positions of good sources in four parallax slices (from top to bottom, at nominal distances of 0.3, 1, 3 and 10 kpc). In the closer distance slices, distinct overdensities that correspond to open clusters are visible. At 3 kpc, the Milky Way's overall structure becomes clearly visible, with star forming regions standing out. Not many sources at very large distances have high SNR, and highly extincted regions of the Galactic plane have no sources (and are therefore colored gray).

When looking at the sources classified as bad in Fig. 10, the bulge and disk dominate in all parallax bins. Even nominally nearby sources with high SNR are concentrated in the region of the sky corresponding to the bulge and disk. Interestingly, even a cut for reasonably high SNR does not remove spurious astrometric solutions, as can be seen by the large number of bad sources.

Fig. 11 shows the CAMD of high-SNR good solutions in each parallax bin. The stellar locus well populated in each parallax bin. The unphysically large number of seemingly pre-main sequence stars (redder and brighter than the main sequence) is due to photometric excess in the RP photometry of sources in dense regions of the sky (Riello et al. 2020). Another feature that is apparent in these CAMDs is that the red clump becomes increasingly elongated with distance, due to the greater range of dust columns probed at larger distances.

For the high SNR sources that are classified bad we see in Fig. 12 a floor of sources near to the Gaia magnitude limit for the respective parallax slices. For the 3000pc bin there might be an indication of AGB sources wrongly classified as bad. Though observational conditions for these extreme objects might resemble astrometrically a bad solution source.

4.6.2 *Low-SNR model*

Now we inspect the result of the low-SNR model on the low SNR validation data (|SNR| < 4.5). Again, the sample will be biased but now contains more sources with small parallaxes.

In the closest distance bin of the sky projection of the good sources in Fig. 13, we see a similarity to the high-SNR sample, though the bulge region is missing and scanning law patterns are visible. While the structures overlap more across different distance bins (due to the lower parallax SNR), similar structures are visible at 3 and 10 kpc as in the high-SNR sample (Fig. 9). Reassuringly, the bulge is most prominent at 10 kpc, while at 30 kpc, the Magellanic clouds are more prominent.

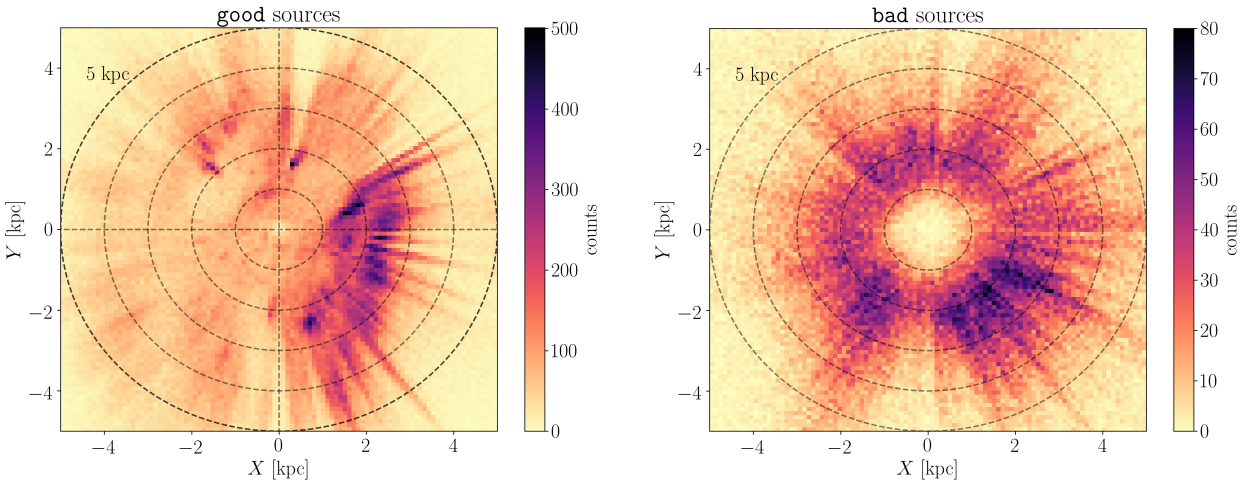

Figure 8. Validation of our astrometric fidelity classification through the astrophysical plausibility of the X-Y distribution of young stars in the Galactic plane. Shown is the distribution of `good` (left) and `bad` (right) OBA star sources in the Galactic plane, with the Sun located at $(X, Y) = (0\,\mathrm{kpc}, 0\,\mathrm{kpc})$, and the Galactic center at $(8.2\,\mathrm{kpc}, 0\,\mathrm{kpc})$. We have divided the plane into pixels 100 pc on a side. The color bar shows the number of sources per bin. The dashed circles have radii ranging from 1 to 5 kpc, in steps of 1 kpc. The `good` sources show concentrations at many known locations of young stars, and show spiral arm like morphology. The `bad` sources show a ringlike structure, exactly centered on the sun and at the (seeming) distance of the most common parallax; clearly, a far too Ptolemean distribution to be "real".

Fig. 14 shows the sky distribution of sources classified as `bad` in the low-SNR sample. Bad sources strongly outnumber the `good` sources in the 1 kpc slice. In every distance slice, the sky distribution of the `bad` sources essentially traces the highest density parts of the sky. At 100 pc the scanning law is still visible, but the sparsity is mainly due to most sources having a high SNR in this distance bin.

For the low SNR sample we only focus on the 10 kpc bin when looking at the CAMD. For the good sources in the upper panel of Fig. 15 we can see massive main sequence stars and turn-off stars, as well as the red clump and maybe sdB stars at the very blue end. For the bad sources in the lower panel, a weak signal of all those populations seems to be present and again some AGB stars might have been wrongly predicted as bad. Overall most of the physical structure can be seen in the good sample, making us confident that even at low SNR our astrometric fidelity classifier is useful.

5 ACCESS TO THE CATALOG

Our catalog is hosted at the German Astrophysical Virtual Observatory (GAVO),[2] in the table `gedr3spur.main`.[3] The simplest way to access the astrometric fidelities for a sample of stars is to cross-match directly via a Table Access Protocol (TAP) upload join in TOPCAT.[4] If the local table has Gaia EDR3 `source_id`s one can simply query:

```
SELECT src.*
FROM gedr3spur.main as src
JOIN TAP_UPLOAD.t1 AS target
-- TAP_UPLOAD.tX needs to be the table number in TOPCAT
  USING (source_id)
```

Because there is 100 MB upload limit, one can increase the number of sources queried at a time by hiding all columns except `source_id`. GAVO hosts a light version of Gaia EDR3, containing only the most commonly used columns. One can directly query this light versio nof Gaia eDR3 and simultaneously crossmatch to our astrometric fidelities. For example,

```
SELECT COUNT(*) AS ct,
ROUND(parallax/parallax_error,2) AS bin
FROM gaia.edr3lite -- only contains most important rows
JOIN gedr3spur.main using (source_id)
WHERE fidelity_v1 >= 0.5 GROUP BY bin
```

returns a histogram of the parallax distribution for the 730 million `good` sources. Requiring `fidelity_v1` < 0.5 returns the parallax distribution for the 738 million `bad` sources. Fig. 16 shows the distributions returned by these two queries. As expected, the `bad` solutions dominate the negative parallax regime. Interestingly, this is also the case for positive parallaxes in the range $0.8 < \hat{\varpi}/\mathrm{mas} < 13$. The parallax distribution of `bad` sources peaks at $\hat{\varpi} = 0.19$ mas and is almost symmetrically distributed around this point, with an excess of 35 million sources in the positive wing. This might be partly due to misclassification of sources with good astrometric solutions, but could also due in part to spurious solutions scattering around the true parallax value (Gaia Collaboration et al. 2020b).

The parallax distribution of the `good` sources peaks at $\hat{\varpi} = 0.26$ mas. For comparison, we have also plotted the distribution of observed parallaxes from 1.33 billion GeDR3mock[5] sources (Rybizki et al. 2020), which peaks at $\hat{\varpi} = 0.16$ mas. The `bad` astrometric solutions are usually real sources that have anomalously large parallax errors. One can therefore imagine how the excess GeDR3mock sources (compared to the `good` EDR3 sources) could be randomly

[2] https://dc.zah.uni-heidelberg.de/
[3] https://dc.zah.uni-heidelberg.de/browse/gedr3spur/q
[4] The TOPCAT program is described at http://www.star.bris.ac.uk/~mbt/topcat/.
[5] We have adopted the EDR3 corrected `parallax_error` from the ADQL query in Gaia Collaboration et al. (2020b) and imposed the magnitude limits from table `maglim_6` of GeDR3mock: https://dc.zah.uni-heidelberg.de/browse/gedr3mock/q.

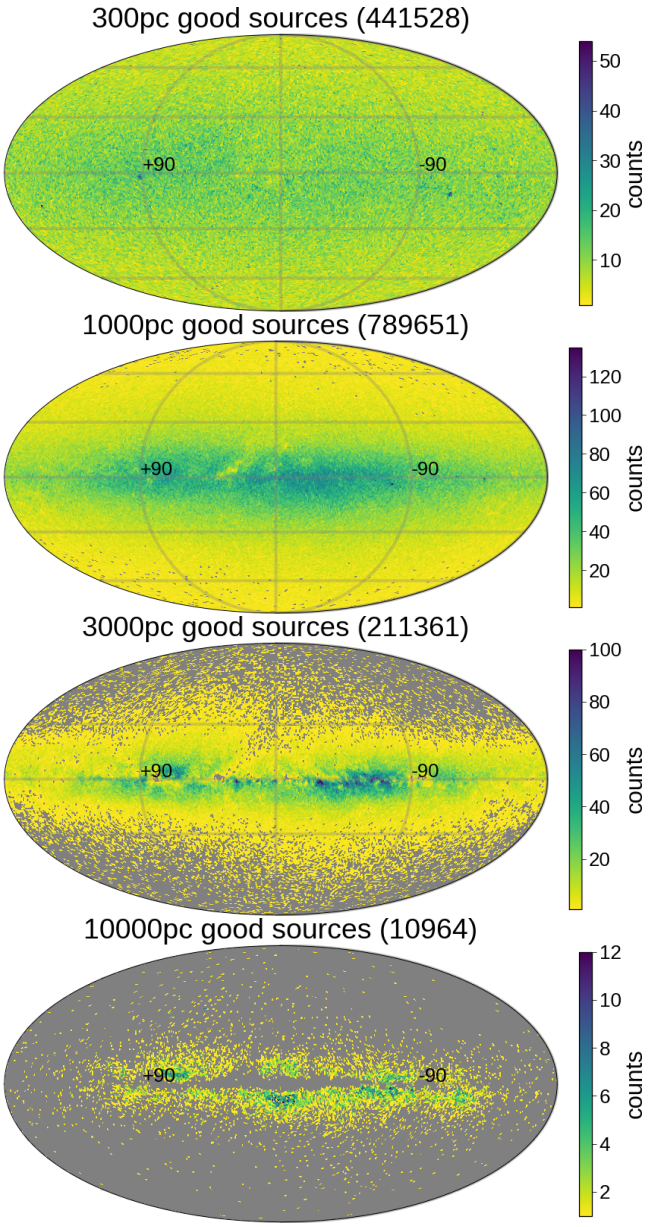

Figure 9. Distribution of good solutions for |SNR| ≥ 4.5 over the sky as a function of parallax (expressed as distance). We use a Mollweide projection of Galactic coordinates. The nearby sources show a nearly-isotropic distribution with only a slight concentration towards the Galactic plane, as expected. More distant shells show the Galactic plane; the most distant bin is sparsely populated, as most sources do not satisfy |SNR| ≥ 4.5.

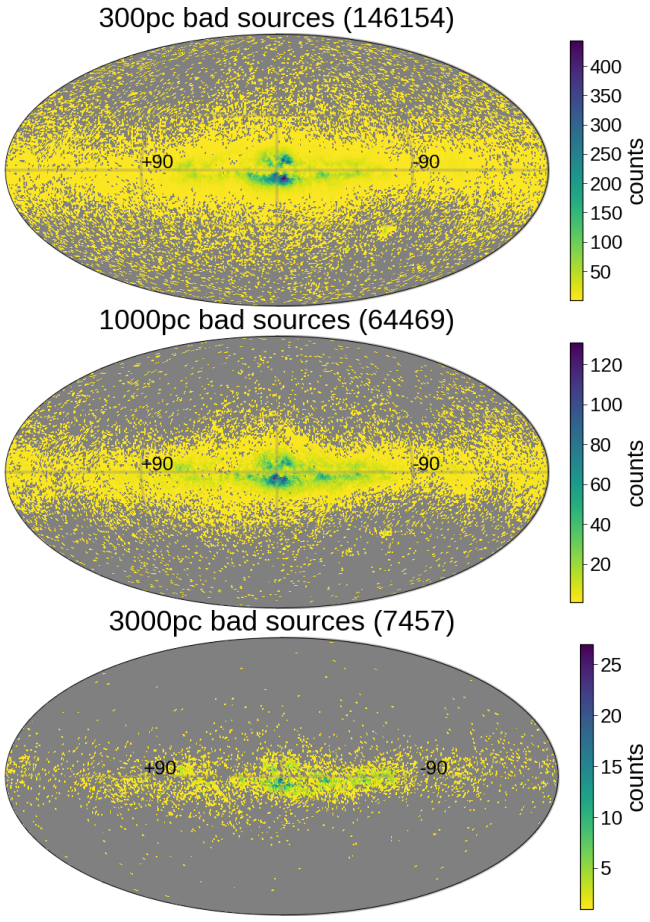

Figure 10. Distribution of bad solutions for |SNR| ≥ 4.5 over the sky, analogous to Fig. 9. We use a Mollweide projection of Galactic coordinates. The strong concentration of sources towards the Galactic center and the Galactic plane, even for sources that purportedly lie within 300 pc; this shows that their quoted parallaxes must have low astrometric fidelity.

scattered to produce the distribution of the bad sources, painting a consistent picture.

Fig. 17 shows the SNR distribution for the good and bad sources in EDR3. We see a jump in the number of sources at the |SNR| = 4.5 transition between our high- and low-SNR classifier. At SNR = 4.5, the number of bad sources increases by almost 50 %, i.e. the high-SNR classifier seems to have a higher purity. Fabricius et al. (2020) estimates the total contamination of the sources with SNR > 5 to be of the order of ∼ 3 million (equal to the number of sources with SNR < -5). Our classifier finds 12.2 million bad sources in this regime, which constitute 6 % of the 192 million sources with SNR > 5.

6 ROOM FOR IMPROVEMENT

This is a list of ideas that could be applied to improve results and might enter a future version of the classifier.

- correcting for parallax zero point
- use wide binaries for validation, use wide binary sample which have statistically sound parallax uncertainty for training
- make mock test (using gedr3mock) with photometric cleaning for sample generation
- Sources in the good sample HEALpix without a 2MASS crossmatch could be still used if a crossmatch to Pan-STARRS1 (Chambers et al. 2016) is found.
- clean classified samples from sources that are obviously wrong and feed them into training
- use different photometric cuts
- use more restrictive cuts when acquiring the good training sample, e.g. only HEALpix with no sources of SNR < −2.5.
- take into account the error coming from true parallax over error, vs measured parallax over error

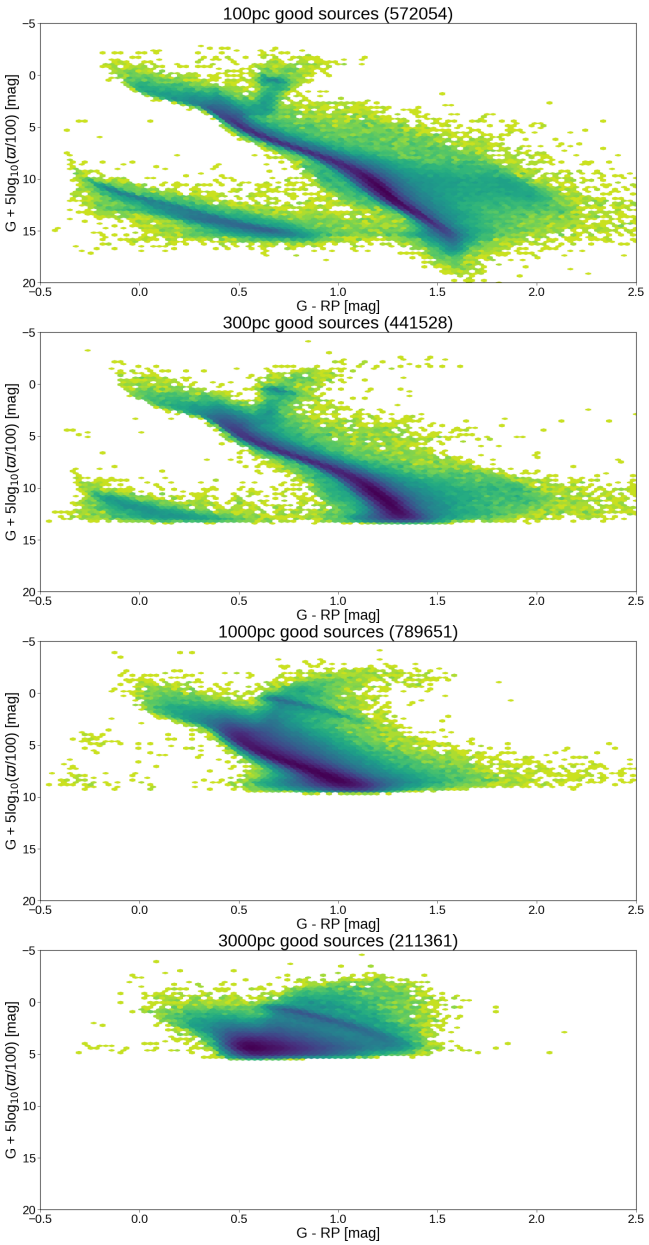

Figure 11. CAMD for good solutions for $|\mathrm{SNR}| \geq 4.5$ in different bins of parallax, expressed as distances. The top panel, covering the widest range in absolute magnitudes, shows a CAMD distribution that is astrophysically plausible, in stark contrast to the analogous panel in Fig. 12.

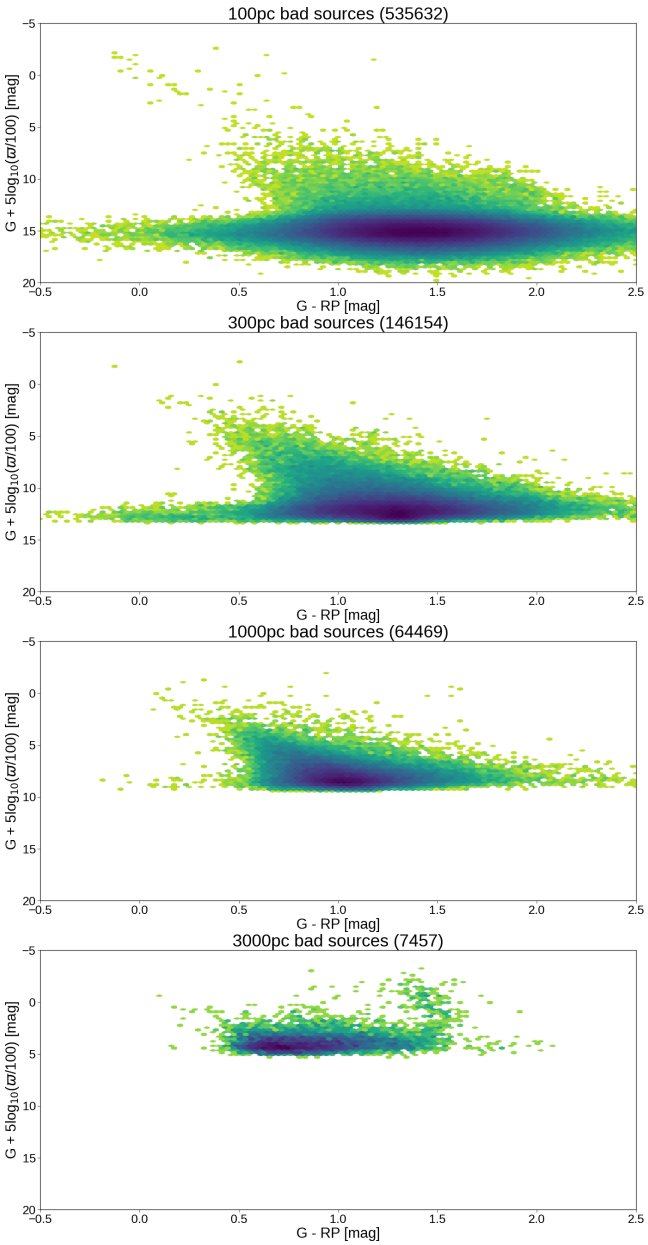

Figure 12. CAMD for bad solutions for $|\mathrm{SNR}| \geq 4.5$, in different bins of parallax, expressed as distances. In contrast to the analogous Fig. 11, the CAMDs in this figure are astrophysically unrealistic. Features such as the main sequence, the red clump and the white dwarf sequence are completely lacking, which confirms that the high-SNR bad sources overwhelmingly have spurious parallaxes.

- cut the samples at different snr levels
- Optional features which might have high predictive power:

  – `matched_transits_removed`

  – `astrometric_params_solved` (or maybe train a model for 5p and 6p solutions separately)

  – `phot_proc_mode` (available for little less sources than those with astrometric solution but more than RP, BP)

  – it might also help to add in `astrometric_excess_noise` / `astrometric_excess_noise_sig` as an estimate of the uncertainty of the excess source noise

  – distance to nearest neighbour in the catalogue

  – distance to nearest bright neighbour (e.g. $< 16\,G$) due to their spurious source creation (Fabricius et al. 2016).

  – `phot_bp_rp_excess_factor` (only available for sources with both colors)

7 CONCLUSION

We have extended the classification of valid and spurious astrometric solutions from the Gaia Catalogue of nearby stars (Gaia Collaboration et al. 2020b) to all 1.47 billion sources in Gaia EDR3

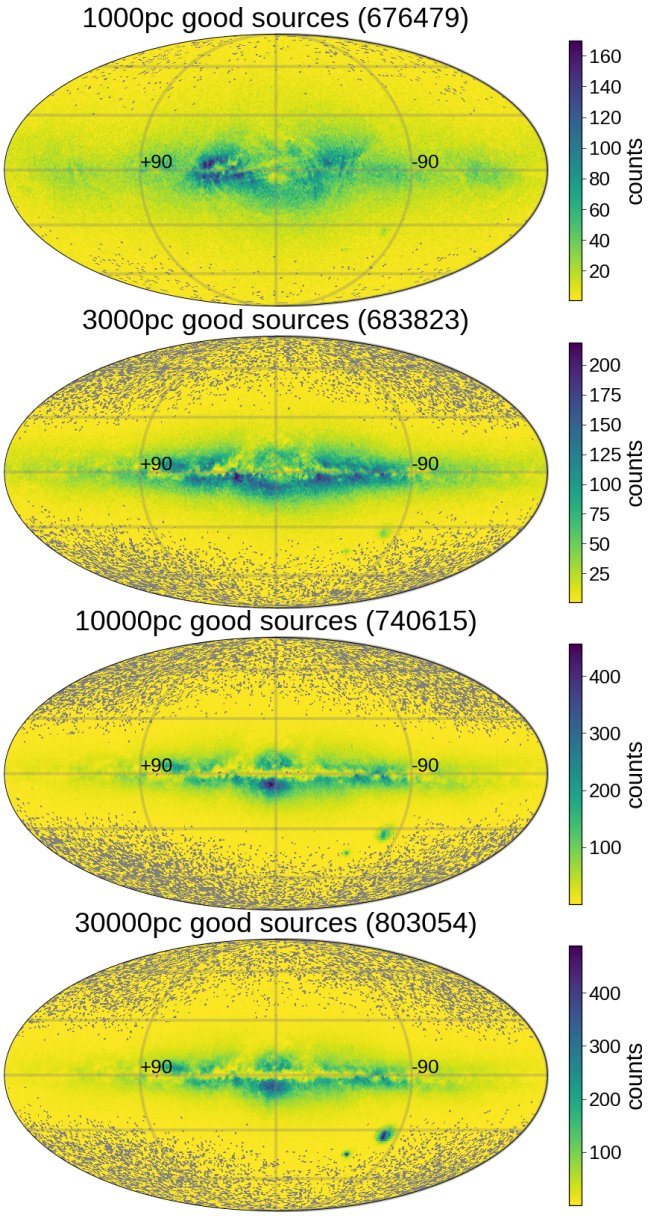

Figure 13. Sky distribution of `good` solutions for |SNR| < 4.5 over the sky, using a Mollweide projection of Galactic coordinates.

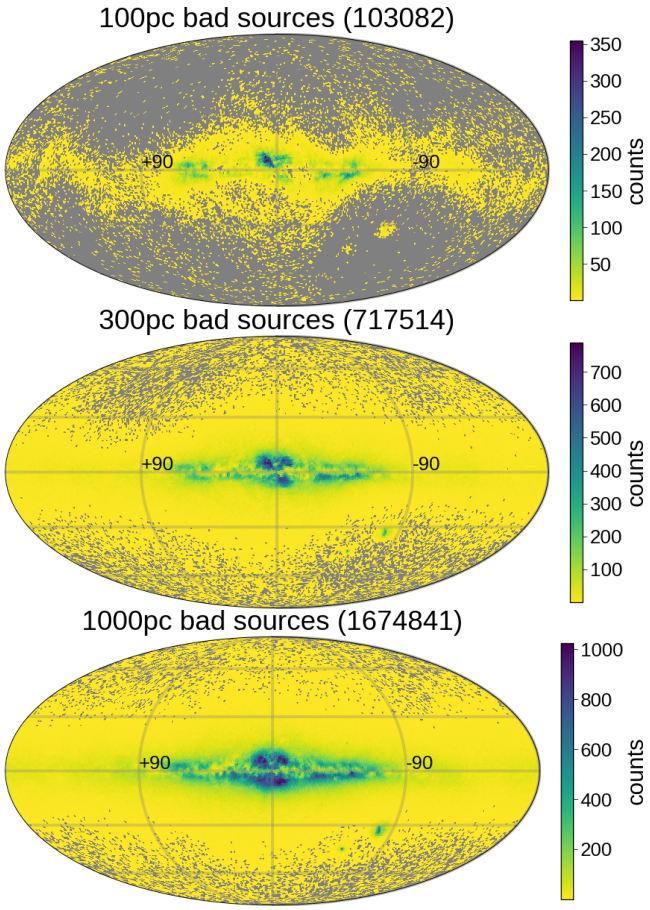

Figure 14. Sky distribution of `bad` solutions for |SNR| < 4.5 over the sky, using a Mollweide projection of Galactic coordinates.

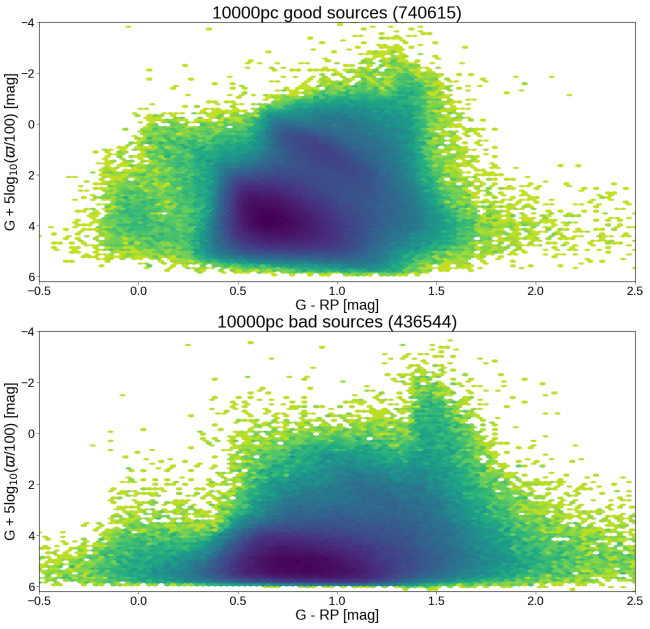

Figure 15. CAMD of `good` (top panel) and `bad` (bottom panel) solutions for |SNR| < 4.5.

with astrometry. Our training sample of spurious sources are obtained by taking all sources with `parallax_over_error` < -4.5. Our training sample of good astrometric solutions is obtained by taking all sources with a 2MASS crossmatch in parts of the sky where no sources with `parallax_over_error` < -3.5 exist. We train two neural network models, one for high parallax SNR and one for low (divided at |SNR|=4.5), which take astrometric quality parameters from EDR3 as inputs.

Our validation shows that we outperform simple cuts but also logistic models that take into account a linear combination of our features. Our `good` sources' parallaxes distribute normally with respect to clusters but also the LMC. Sources classified as `good` also have significantly lower χ^2 when comparing to OGLE proper motions. Sources classified as `bad` usually occur in high-density

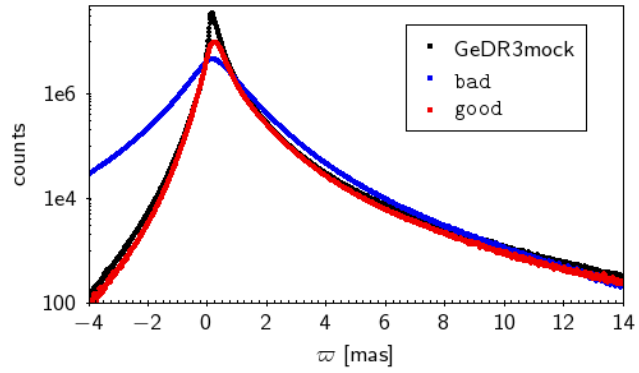

Figure 16. Parallax distribution for good and bad sources in Gaia EDR3 and for the mock observed parallaxes of GeDR3mock.

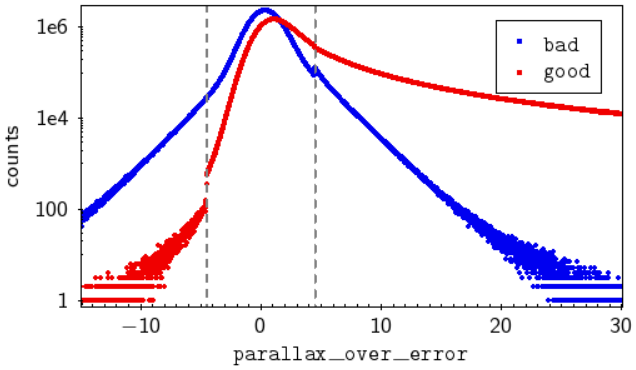

Figure 17. SNR (`parallax_over_error`) distribution for good and bad sources in Gaia EDR3. The |SNR| = 4.5, where the classifiers change, is shown in dashed grey lines.

regions, e.g. in the bulge and disc region and in the Magellanic clouds.

ACKNOWLEDGEMENTS

The authors would like to thank Douglas P. Finkbeiner and Joshua S. Speagle for helpful discussions and suggestions.

This work has made use of data from the European Space Agency (ESA) mission Gaia, processed by the Gaia Data Processing and Analysis Consortium (DPAC). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

This research or product makes use of public auxiliary data provided by ESA/Gaia/DPAC as obtained from the publicly accessible ESA Gaia SFTP.

This work was funded by the DLR (German space agency) via grant 50 QG 1403.

GG acknowledges funding from the Alexander von Humboldt Foundation, through the Sofja Kovalevskaja Award.

The OGLE project has received funding from the National Science Centre, Poland, grant MAESTRO 2014/14/A/ST9/00121 to AU.

JR will not travel anywhere by aeroplane for the purpose of promoting this paper.

Software: `TOPCAT` (Taylor 2005), `HEALpix` (Górski et al. 2005).

Data availability

The data underlying this article are available in the article and in its online supplementary material.

REFERENCES

Abadi M., et al., 2016, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems (arXiv:1603.04467)
Bailer-Jones C. A. L., 2015, PASP, 127, 994
Bailer-Jones C. A. L., Rybizki J., Fouesneau M., Mantelet G., Andrae R., 2018, AJ, 156, 58
Chambers K. C., et al., 2016, arXiv e-prints, p. arXiv:1612.05560
Chollet F., et al., 2015, Keras, https://keras.io
Fabricius C., et al., 2016, A&A, 595, A3
Fabricius C., et al., 2020, arXiv e-prints, p. arXiv:2012.06242
Gaia Collaboration et al., 2018, A&A, 616, A10
Gaia Collaboration Brown A. G. A., Vallenari A., Prusti T., de Bruijne J. H. J., Babusiaux C., Biermann M., 2020a, arXiv e-prints, p. arXiv:2012.01533
Gaia Collaboration et al., 2020b, arXiv e-prints, p. arXiv:2012.02061
Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press
Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Kingma D. P., Ba J., 2014, arXiv e-prints, p. arXiv:1412.6980
Lindegren L., et al., 2020a, arXiv e-prints, p. arXiv:2012.01742
Lindegren L., et al., 2020b, arXiv e-prints, p. arXiv:2012.03380
Luri X., et al., 2018, A&A, 616, A9
Riello M., et al., 2020, arXiv e-prints, p. arXiv:2012.01916
Rybizki J., et al., 2020, PASP, 132, 074501
Skrutskie M. F., et al., 2006, AJ, 131, 1163
Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, Astronomical Society of the Pacific Conference Series Vol. 347, Astronomical Data Analysis Software and Systems XIV. p. 29
Torra F., et al., 2020, arXiv e-prints, p. arXiv:2012.06420
Udalski A., Szymański M. K., Szymański G., 2015, Acta Astron., 65, 1

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