Communications of BAO, Vol. 73, Issue 2, 2026, pp. 328-333

Cross-galaxy prediction of circumstellar OH masers: from the Milky Way to the Magellanic Clouds

M. Faraji*[1], J.Th. van Loon[2], and A. Javadi[1]

[1]School of Astronomy, Institute for Research in Fundamental Sciences (IPM), P.O. Box 19395-5531, Tehran, Iran
[2]Lennard-Jones Laboratories, Keele University, ST5 5BG, UK

Abstract

Circumstellar 1612 MHz OH masers are associated with evolved stars undergoing intense mass loss. They are well studied in the Milky Way and confirmed in the Large Magellanic Cloud (LMC), but no circumstellar OH maser has yet been detected in the Small Magellanic Cloud (SMC). We present a machine-learning approach to identify new maser candidates in both Clouds. An XGBoost classifier is trained on Galactic evolved stars from the Hamburg database of circumstellar OH masers, using photometric colours from Gaia, 2MASS and WISE as the only input features. To transfer the model to the Magellanic Clouds, we apply an unsupervised, band-by-band median alignment of the photometric distributions. On a held-out Milky Way test set, the classifier reaches an area under the ROC curve (AUC) of 0.85, and on an independent sample of confirmed LMC detections and non-detections, the two classes are fully separated (AUC = 1.00). We present ranked candidate lists for the LMC and SMC and refine them with independent OGLE-III variability and chemical classifications, retaining oxygen-rich Mira variables as the most promising targets for radio follow-up.

Keywords: *masers – stars: AGB and post-AGB – circumstellar matter – Magellanic Clouds – methods: machine learning*

1. Introduction

1.1. What is a maser?

A maser (Microwave Amplification by Stimulated Emission of Radiation) is the microwave analogue of a laser. An energy source pumps molecules into an excited state until the population of the upper level exceeds that of the lower level, a condition known as population inversion. A photon of energy $h\nu$ that interacts with an excited molecule then triggers stimulated emission: the molecule emits a second photon identical to the first, with the same frequency, phase, direction, and polarisation. These photons stimulate further molecules, so the radiation is amplified exponentially along its path, producing intense and very narrow radio emission lines.

1.2. OH masers in the Milky Way and the Magellanic Clouds

Thousands of OH maser observations have been made in the Milky Way since intense OH emission was first found towards infrared-bright late-type stars (Engels & Bunzel, 2015, Wilson & Barrett, 1968). Circumstellar OH masers are associated with late-type evolved stars undergoing intense mass loss, and the 1612 MHz satellite transition is particularly characteristic of circumstellar OH/IR stars. In these stars, OH is produced in the dense, dusty outer regions of the circumstellar envelope, and the maser is pumped by infrared radiation from the dust (Elitzur et al., 1976, Goldreich & Scoville, 1976); OH masers, therefore, trace these outer envelope regions.

OH masers are also known and confirmed in the Large Magellanic Cloud (LMC) (Goldman et al., 2017, Marshall et al., 2004, Wood et al., 1992, van Loon et al., 2001), where they provide a laboratory for studying maser production at lower metallicity than in the Milky Way. In the Small Magellanic Cloud (SMC), however, no circumstellar OH maser has been confirmed, even in deep dedicated searches (Goldman

*marzieh.faraji10@gmail.com, Corresponding author

et al., 2018). This raises the question of whether the difference is caused by metallicity, dust production, mass-loss properties, evolutionary state, or selection effects. In this work, we use machine learning to identify the most promising OH maser candidates in both Clouds.

2. Data and data pre-processing

We use four catalogues, summarised in Table 1. The classifier is trained on the Milky Way sample from the Hamburg Circumstellar OH Maser Database (Engels & Bunzel, 2015), which contains both detections and non-detections. An independent LMC validation set was compiled from the literature (Goldman et al., 2017, Marshall et al., 2004, Wood et al., 1992, van Loon et al., 2001) and contains 21 stars observed at 1612 MHz: 11 detections, 7 non-detections, and 3 sources with uncertain (conflicting) status. For the candidate search, we queried SIMBAD for evolved-star populations within 6° of the LMC and 3° of the SMC.

Table 1. The catalogues used in this work. For the Milky Way and the candidate-search catalogues, the number of sources before and after pre-processing is given.

Catalogue	Purpose	Data / selection	Sources	Labelled
Milky Way	Training	Hamburg OH maser database	16,879 obs. → 3,028 stars	Yes
LMC	Validation	Literature compilation	21 (11 det., 7 non-det., 3 unc.)	Yes
LMC	Candidate search	SIMBAD, 6° around LMC	125,211 → 40,039	No
SMC	Candidate search	SIMBAD, 3° around SMC	23,682 → 11,121	No

The catalogues were prepared through the following steps:

1) *Initial source information.* For each source we collected the coordinates, distance, object and spectral type, and cross-survey identifiers.

2) *Archival photometry.* We retrieved Gaia DR3 G, $G_{\rm BP}$, $G_{\rm RP}$ (Gaia Collaboration et al., 2023), 2MASS J, H, $K_{\rm s}$ (Cutri et al., 2003), WISE $W1$–$W4$ (Wright et al., 2019) and IRAS 12, 25, 60 and 100 μm photometry.

3) *Restriction to circumstellar sources.* The OH masers in the database include both circumstellar and interstellar populations. We retained evolved-star types (AGB and RGB stars, long-period variables and Miras, OH/IR stars, post-AGB stars, planetary nebulae, etc.) and removed interstellar sources and extragalactic objects.

4) *Restriction to 1612 MHz.* We focused on the 1612 MHz satellite line, which is characteristic of circumstellar OH masers in evolved oxygen-rich stars, and excluded the 1665/1667 MHz main lines to obtain a uniform detection definition.

5) *Detection labeling.* Each source was labeled as a detection ($y = 1$) or a non-detection ($y = 0$) on the basis of the blue-peak flux density reported in the database.

6) *Band-level completeness.* We required at least 50% completeness for each photometric band; the $UBVRI$ bands were removed because of insufficient completeness.

7) *Source-level completeness.* We required at least 75% of the ten classifier bands (Gaia G, $G_{\rm BP}$, $G_{\rm RP}$; 2MASS J, H, $K_{\rm s}$; WISE $W1$–$W4$) to be populated, which ensures reliable colour features with minimal missing data.

3. Method

3.1. OH maser classification with XGBoost

We treat the identification of OH maser hosts as a supervised binary classification problem, detection ($y = 1$) versus non-detection ($y = 0$), and use the gradient-boosted decision-tree algorithm XGBoost (Chen

doi: https://doi.org/10.52526/25792776-26.73.2-328

& Guestrin, 2016). The input features are photometric colours only. Colours are, to first order, independent of distance, so a model trained on Galactic stars can be applied to the Magellanic Clouds. The model learns the multi-dimensional colour-space signature of OH masers and ranks candidate sources by their predicted maser probability, $P_{\rm maser}$.

3.2. Cross-galaxy domain alignment

A model trained on the Milky Way and applied to the Magellanic Clouds faces a domain shift: the observed colour distributions of the galaxies differ because of distance, survey effects, metallicity, dust, and stellar populations. We assume that the relation $P(\text{maser}\,|\,\text{colour})$ is approximately transferable between the galaxies (Ben-David et al., 2010), and apply an unsupervised median shift: the median magnitude of each photometric band b in the LMC or SMC is aligned to that of the Milky Way training sample,

$$m_b^{\rm MC,\,aligned} = m_b^{\rm MC} + \left[\,\mathrm{med}\left(m_b^{\rm MW}\right) - \mathrm{med}\left(m_b^{\rm MC}\right)\right]. \tag{1}$$

No maser labels are used, so the alignment is non-circular and reproducible. Because a constant shift is applied in each band, the transformation is a rigid translation in colour space: it preserves the pairwise differences between sources and the shape and spread of the distributions, and only changes the location of the population.

3.3. Feature selection

The initial features were the photometric bands G, $G_{\rm BP}$, $G_{\rm RP}$, J, H, $K_{\rm s}$ and $W1$–$W4$, from which candidate colours were constructed. We selected the final colour set by an ablation (Table 2): each candidate set was used to train the classifier on the Milky Way and to rank the full LMC catalogue, and was evaluated both in-domain (Milky Way test AUC) and on transfer (LMC AUC on the confirmed sources, and recovery of the 11 confirmed masers and 7 confirmed non-detections among the top-ranked candidates). The adopted nine-colour set (H) performs best on every metric. Physically, the colours carry three kinds of information: $W1-W3$ and $W2-W3$ trace the warm inner dust, $G-G_{\rm RP}$ traces the photospheric temperature and reddening, and the remaining colours trace the total circumstellar dust obscuration.

Table 2. Feature-set ablation. Sets D–H are built from the single colours A–C. The adopted set is H.

Set	Colours	N	MW AUC (in-domain)	LMC AUC (transfer)	Masers recovered top 500 / top 1000		Non-det. top 1000
A	$G_{\rm RP}-W4$	1	0.77	0.66	11/11	11/11	4/7
B	$J-W3$	1	0.74	0.80	2/11	2/11	0/7
C	$W2-W3$	1	0.62	0.64	1/11	3/11	0/7
D	A+B	2	0.80	0.94	10/11	10/11	4/7
E	A+C	2	0.77	0.94	9/11	9/11	1/7
F	B+C	2	0.75	0.94	2/11	4/11	0/7
G	A+B+C	3	0.82	0.97	10/11	10/11	1/7
H	**G + 6 colours[a] (adopted)**	**9**	**0.85**	**1.00**	**10/11**	**11/11**	**0/7**

[a] $G_{\rm BP}-W3$, $W1-W3$, $G-G_{\rm RP}$, $H-W3$, $H-W2$ and $K_{\rm s}-W3$.

4. Results

4.1. Model performance

On the held-out Milky Way test set the model achieves an area under the receiver operating characteristic (ROC) curve of AUC = 0.85, with an accuracy of 77%, a precision of 72% and a recall of 70% at a probability threshold of 0.5 (Figure 1, left). Applied to the LMC in the aligned frame, the Milky-Way-trained model separates the confirmed detections and non-detections completely (AUC = 1.00; Figure 1, right): the median predicted probability is 0.92 for the confirmed masers and 0.73 for the confirmed non-detections. Since the confirmed LMC sample is small, this result should be read as an indication of clean separation rather than as a precise figure of merit.

doi: https://doi.org/10.52526/25792776-26.73.2-328

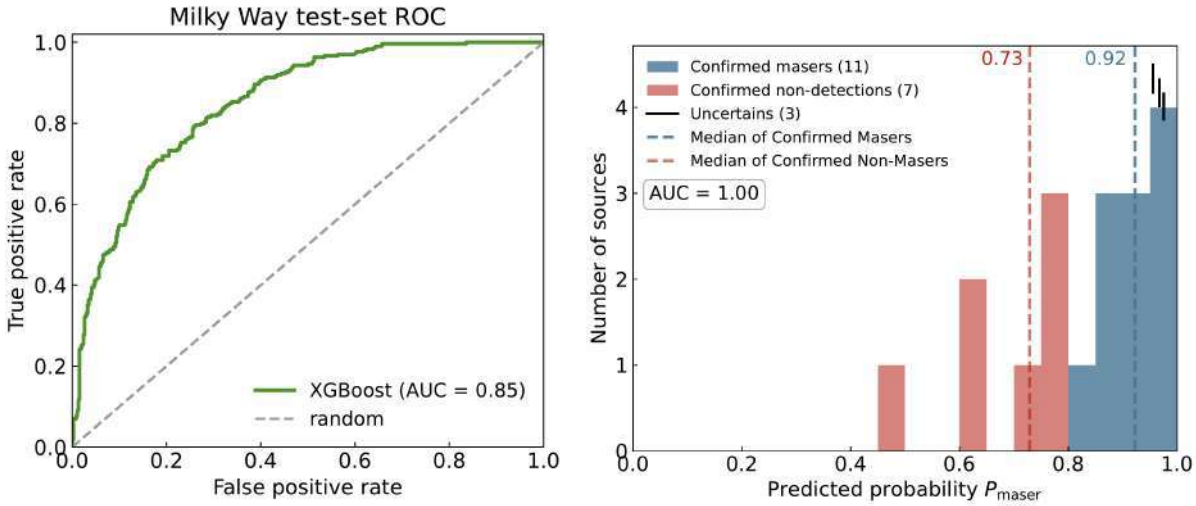

Figure 1. *Left:* ROC curve of the XGBoost classifier on the held-out Milky Way test set (AUC = 0.85). *Right:* predicted OH maser probability for the confirmed LMC sources, from the Milky-Way-trained model applied in the aligned frame. Confirmed masers (blue) and confirmed non-detections (red) are fully separated (AUC = 1.00); dashed lines mark the medians of the two classes and black ticks mark the three uncertain sources.

4.2. The model's verdict on the confirmed LMC masers

Table 3 lists the predicted probability and rank of each confirmed LMC source within the full catalogue of 40,039 LMC evolved stars. All confirmed detections lie within the top 2% of the ranking, and every confirmed detection is ranked above every confirmed non-detection. The three sources with conflicting literature reports score among the detections.

Table 3. Predicted maser probability $P_{\rm maser}$ and rank of the confirmed LMC sources within the full LMC catalogue (40,039 stars). The model received no labels for these stars.

Source	$P_{\rm maser}$	Rank	Top (%)	Source	$P_{\rm maser}$	Rank	Top (%)
Confirmed detections				*Uncertain (conflicting literature)*			
IRAS 04407−7000	98.3%	1	<0.01	IRAS 04509−6922	97.5%	12	0.03
IRAS 05003−6712	98.3%	3	0.01	IRAS 04516−6902	96.7%	32	0.08
IRAS 04498−6842	98.3%	4	0.01	IRAS 05294−7104	95.6%	63	0.16
IRAS 05558−7000	97.6%	10	0.02	*Confirmed non-detections*			
MSX LMC 811	94.9%	82	0.20	IRAS 04530−6916	79.9%	1050	2.62
IRAS 04553−6825	92.3%	141	0.35	IRAS 04553−6933	78.5%	1349	3.37
IRAS 05402−6956	92.0%	149	0.37	IRAS 04537−6922	76.8%	1861	4.65
IRAS 05280−6910	88.9%	238	0.59	IRAS 05216−6753	72.9%	3643	9.10
MSX LMC 807	88.8%	243	0.61	IRAS 05346−6949	62.1%	8908	22.25
IRAS 04545−7000	87.7%	283	0.71	MSX LMC 1207	62.0%	8930	22.30
IRAS 05298−6957	81.7%	766	1.91	MSX LMC 815	49.7%	13504	33.73

4.3. Candidates in the LMC and SMC

Applying the model to the full LMC and SMC catalogues returns a ranked list of candidates in each galaxy. The twenty highest-ranked candidates are listed in Table 4. In the LMC, the list includes four of the confirmed masers, recovered blindly, and one of the uncertain sources; the remaining stars have not previously been observed in OH. Since no circumstellar OH maser has been detected in the SMC, all SMC entries are new candidates. In both galaxies the highest-ranked candidates are Miras and long-period variables.

Table 4. The twenty highest-ranked OH maser candidates in the LMC and SMC. Types follow SIMBAD: Mi*, Mira; LP*, long-period variable.

	LMC				SMC			
Rank	Source	RA (deg)	Dec (deg)	Type	Source	RA (deg)	Dec (deg)	Type
1	IRAS 04407−7000[a]	70.1187	−69.9204	Mi*	OGLE SMC-LPV-3315	10.2317	−72.7970	Mi*
2	IRAS 05003−6712[a]	75.0791	−67.1327	Mi*	OGLE SMC-LPV-1168	8.0510	−74.8013	Mi*
3	IRAS 04498−6842[a]	72.4227	−68.6308	Mi*	OGLE SMC-LPV-1042	7.8551	−74.0555	Mi*
4	OGLE LMC-LPV-53224	81.2867	−69.6401	LP*	IRAS F01066−7332	17.0429	−73.2645	Mi*
5	SHV 0453000−682708	73.2170	−68.3720	LP*	OGLE J010603.28−722231.9	16.5138	−72.3756	Mi*
6	OGLE LMC-LPV-32336	78.7716	−68.9684	Mi*	OGLE SMC-LPV-18776	19.0224	−72.6814	Mi*
7	OGLE LMC-LPV-76059	84.5416	−69.0619	LP*	OGLE SMC-LPV-2110	9.2836	−72.4435	Mi*
8	OGLE LMC-LPV-77855	84.9247	−69.7986	LP*	OGLE SMC-LPV-17998	17.7468	−72.6272	Mi*
9	IRAS 05558−7000[a]	88.8376	−70.0008	–	OGLE SMC-LPV-12167	14.0911	−73.4276	Mi*
10	OGLE LMC-LPV-78468	85.0642	−70.1583	LP*	OGLE SMC-LPV-19166	20.2338	−71.3862	Mi*
11	IRAS 04509−6922[b]	72.6686	−69.2921	–	OGLE SMC-LPV-1513	8.5833	−74.3357	Mi*
12	2MASS J05110209−6834423	77.7587	−68.5783	LP*	OGLE J005915.79−722754.4	14.8158	−72.4651	Mi*
13	OGLE LMC-LPV-12167	75.3904	−70.6374	Mi*	2MASS J00415184−7138541	10.4660	−71.6483	LP*
14	2MASS J04463115−7211471	71.6298	−72.1964	LP*	MSX SMC 105	11.2589	−72.8733	Mi*
15	OGLE LMC-LPV-9756	74.7484	−67.0830	Mi*	OGLE SMC-LPV-2875	9.9176	−74.1498	Mi*
16	OGLE LMC-LPV-6530	73.6889	−68.0707	Mi*	OGLE SMC-LPV-18580	18.6447	−72.8921	LP*
17	OGLE LMC-LPV-74614	84.2646	−69.0759	LP*	OGLE SMC-LPV-17729	17.4561	−73.8254	Mi*
18	MSX LMC 880	83.3136	−67.6823	LP*	OGLE SMC-LPV-929	7.6697	−73.7125	Mi*
19	2MASS J06145884−6616547	93.7451	−66.2818	LP*	OGLE SMC-LPV-2981	9.9928	−72.9387	LP*
20	OGLE LMC-LPV-27195	78.0753	−70.6627	Mi*	OGLE J010719.67−724620.6	16.8321	−72.7724	Mi*

[a] Confirmed OH maser, recovered blindly. [b] Source of uncertain status (conflicting literature reports).

4.4. Independent analysis with OGLE data

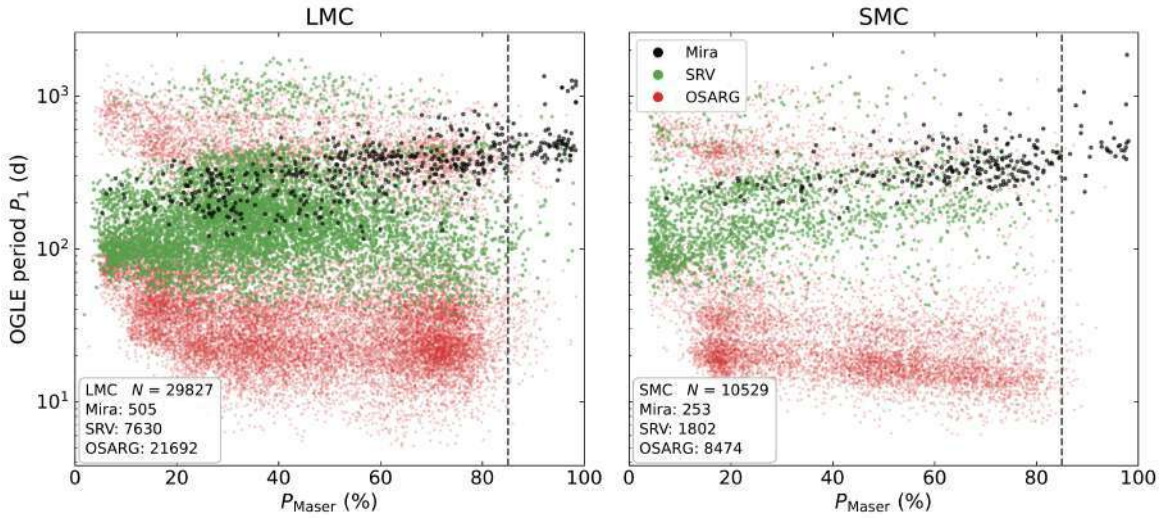

Figure 2. Primary OGLE-III pulsation period P_1 against the predicted maser probability for all LMC (left) and SMC (right) sources with an OGLE-III counterpart, colour-coded by variability subtype: Miras (black), SRVs (green) and OSARGs (red). The dashed line marks $P_{\rm maser} = 0.85$.

To test the candidates independently, we cross-matched the LMC and SMC catalogues with the OGLE-III catalogues of long-period variables (Soszyński et al., 2009, 2011), which classify stars as Miras, semi-regular variables (SRVs) or OGLE small-amplitude red giants (OSARGs) and provide pulsation periods and chemical (oxygen- or carbon-rich) classifications. Figure 2 shows the primary OGLE pulsation period against the predicted maser probability. No time-domain quantity is used by the model, so the two axes are independent. In both galaxies the Mira fraction and the typical period increase monotonically with predicted probability, and above $P_{\rm maser} = 0.85$ the matched sample consists almost entirely of Miras: the long-period, fundamental-mode pulsators expected to host OH masers.

We then refined the high-probability candidates ($P_{\rm maser} > 0.85$) using the OGLE-III classifications as independent physical priors, retaining every source that OGLE classifies both as a Mira variable and as oxygen-rich. These two criteria correspond to the two ingredients of a circumstellar 1612 MHz maser that

doi: https://doi.org/10.52526/25792776-26.73.2-328

colour alone does not guarantee: a large-amplitude pulsation that can sustain a dense wind, and an oxygen-rich chemistry in which OH can form. Of the 462 LMC sources with $P_{\rm maser} > 0.85$, 334 have an OGLE-III counterpart; 208 of these are classified as Miras and 224 as oxygen-rich. In the SMC, 67 of the 70 sources with $P_{\rm maser} > 0.85$ have an OGLE-III counterpart, of which 30 are Miras and 38 are oxygen-rich. Sources that satisfy both criteria form the final candidate lists.

5. Conclusions and outlook

We have trained an XGBoost classifier on Galactic evolved stars using photometric colours alone and transferred it to the Magellanic Clouds with an unsupervised median alignment. The model separates the confirmed LMC masers from the confirmed non-detections, ranks the known masers among its top candidates, and selects long-period Mira variables although no variability information is used. Combined with OGLE-III classifications, it provides prioritised lists of oxygen-rich Mira candidates in the LMC and SMC. These candidates are natural targets for the GASKAP-OH survey with the Australian Square Kilometre Array Pathfinder (ASKAP), which is mapping the 1612 MHz ground-state OH transition across the Magellanic Clouds (Dickey et al., 2013), and for deeper pointed observations, which may lead to the first detection of a circumstellar OH maser in the SMC.

Acknowledgements

This research has made use of the SIMBAD database and the VizieR catalogue access tool, operated at CDS, Strasbourg, France. 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). This publication makes use of data products from the Two Micron All Sky Survey, the Wide-field Infrared Survey Explorer, and the OGLE project.

References

Ben-David S., Blitzer J., Crammer K., Kulesza A., Pereira F., Vaughan J. W., 2010, Machine Learning, 79, 151

Chen T., Guestrin C., 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp 785–794, doi:10.1145/2939672.2939785

Cutri R. M., Skrutskie M. F., van Dyk S., et al., 2003, 2MASS All Sky Catalog of point sources, NASA/IPAC Infrared Science Archive

Dickey J. M., McClure-Griffiths N., Gibson S. J., et al., 2013, PASA, 30, e003

Elitzur M., Goldreich P., Scoville N., 1976, ApJ, 205, 384

Engels D., Bunzel F., 2015, A&A, 582, A68

Gaia Collaboration Vallenari A., Brown A. G. A., Prusti T., et al., 2023, A&A, 674, A1

Goldman S. R., et al., 2017, MNRAS, 465, 403

Goldman S. R., et al., 2018, MNRAS, 473, 3835

Goldreich P., Scoville N., 1976, ApJ, 205, 144

Marshall J. R., van Loon J. T., Matsuura M., Wood P. R., Zijlstra A. A., Whitelock P. A., 2004, MNRAS, 355, 1348

Soszyński I., Udalski A., Szymański M. K., et al., 2009, Acta Astron., 59, 239

Soszyński I., Udalski A., Szymański M. K., et al., 2011, Acta Astron., 61, 217

Wilson W. J., Barrett A. H., 1968, Science, 161, 778

Wood P. R., Whiteoak J. B., Hughes S. M. G., Bessell M. S., Gardner F. F., Hyland A. R., 1992, ApJ, 397, 552

Wright E. L., Eisenhardt P. R. M., Mainzer A. K., et al., 2019, AllWISE Source Catalog, IPAC, doi:10.26131/IRSA1

van Loon J. T., Zijlstra A. A., Bujarrabal V., Nyman L.-Å., 2001, A&A, 368, 950