The Astrophysical Journal Supplement Series, 254:33 (31pp), 2021 June https://doi.org/10.3847/1538-4365/abe465

SPICY: The Spitzer/IRAC Candidate YSO Catalog for the Inner Galactic Midplane

Michael A. Kuhn[1], Rafael S. de Souza[2], Alberto Krone-Martins[3,4], Alfred Castro-Ginard[5], Emille E. O. Ishida[6], Matthew S. Povich[1,7], and Lynne A. Hillenbrand[1]
for the COIN Collaboration

[1] Department of Astronomy, California Institute of Technology, Pasadena, CA 91125, USA; mkuhn@astro.caltech.edu
[2] Key Laboratory for Research in Galaxies and Cosmology, Shanghai Astronomical Observatory, Chinese Academy of Sciences, 80 Nandan Rd., Shanghai 200030, People's Republic of China; drsouza@shao.ac.cn
[3] Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA 92697, USA
[4] CENTRA/SIM, Faculdade de Ciências, Universidade de Lisboa, Ed. C8, Campo Grande, 1749-016, Lisboa, Portugal
[5] Institut de Ciències del Cosmos, Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, E-08028 Barcelona, Spain
[6] Université Clermont Auvergne, CNRS/IN2P3, LPC, F-63000 Clermont-Ferrand, France
[7] Department of Physics and Astronomy, California State Polytechnic University Pomona, 3801 West Temple Avenue, Pomona, CA 91768, USA
Received 2020 November 12; revised 2020 December 23; accepted 2020 December 24; published 2021 June 2

Abstract

We present ∼120,000 Spitzer/IRAC candidate young stellar objects (YSOs) based on surveys of the Galactic midplane between $\ell \sim 255°$ and 110°, including the GLIMPSE I, II, and 3D, Vela-Carina, Cygnus X, and SMOG surveys (613 square degrees), augmented by near-infrared catalogs. We employed a classification scheme that uses the flexibility of a tailored statistical learning method and curated YSO data sets to take full advantage of Spitzer's spatial resolution and sensitivity in the mid-infrared ∼3–9 μm range. Multiwavelength color/magnitude distributions provide intuition about how the classifier separates YSOs from other red IRAC sources and validate that the sample is consistent with expectations for disk/envelope-bearing pre–main-sequence stars. We also identify areas of IRAC color space associated with objects with strong silicate absorption or polycyclic aromatic hydrocarbon emission. Spatial distributions and variability properties help corroborate the youthful nature of our sample. Most of the candidates are in regions with mid-IR nebulosity, associated with star-forming clouds, but others appear distributed in the field. Using Gaia DR2 distance estimates, we find groups of YSO candidates associated with the Local Arm, the Sagittarius–Carina Arm, and the Scutum–Centaurus Arm. Candidate YSOs visible to the Zwicky Transient Facility tend to exhibit higher variability amplitudes than randomly selected field stars of the same magnitude, with many high-amplitude variables having light-curve morphologies characteristic of YSOs. Given that no current or planned instruments will significantly exceed IRAC's spatial resolution while possessing its wide-area mapping capabilities, Spitzer-based catalogs such as ours will remain the main resources for mid-infrared YSOs in the Galactic midplane for the near future.

Unified Astronomy Thesaurus concepts: Young stellar objects (1834); Milky Way disk (1050); Star formation (1569); Star forming regions (1565); Stellar associations (1582)

Supporting material: data behind figure, figure sets, FITS files

1. Introduction

The majority of young stellar objects (YSOs) in our galaxy are formed in massive star-forming complexes located near the Galaxy's midplane. The prevalence of star-forming regions in this part of the Galaxy is attested to by the spatially complex mid-infrared (mid-IR) nebulosity observed to permeate the entirety of the inner midplane and much of the outer midplane. For example, observations by the Spitzer Space Telescope (Werner et al. 2004) and the Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010) have identified more than a thousand interstellar medium bubbles in these regions, most of which are associated with star formation activity (Churchwell et al. 2006, 2007; Simpson et al. 2012; Anderson et al. 2014; Bufano et al. 2018; Jayasinghe et al. 2019). Nevertheless, apart from a few dozen well-studied star-forming regions, the YSOs in these regions remain either mostly or wholly unknown. This is a consequence of observational difficulties at low Galactic latitudes, including high dust column densities along many lines of sight, which limit optical studies, high stellar densities, which may produce source confusion and increase the number of contaminants in catalogs, and lines of sight that pass through multiple star-forming regions at different distances (Feigelson 2018).

There are many scientific applications for reliable lists of YSOs generated uniformly for large segments of the sky rather than on a region-by-region basis. For example, it remains an open question whether nearly all stars are formed in dense groups or whether there is a significant population of stars formed in low-density environments (e.g., Carpenter 2000; Bressert et al. 2010; Gieles et al. 2012; Pfalzner et al. 2012; Kuhn et al. 2015). Hence, catalogs that sample YSOs from both types of environment help to address this question. In addition, YSO catalogs, when combined with Gaia astrometric data, can be used to map out the kinematics of the youngest component of the Milky Way's thin disk. Furthermore, with an increasing number of surveys searching large areas of the sky for transients, these catalogs would help in identifying outbursting YSOs and other YSO-related variability (Hodgkin et al. 2013; Bonito et al. 2018; Graham et al. 2019).

Our goal here is to make optimum use of Spitzer survey data from the inner Galactic midplane (between $\ell \sim 255°$ and 110° and $|b| < 1°$–3°) to identify YSOs out to several kiloparsecs in distance, using IR excess selection criteria that are independent

of spatial clustering. Here, we focus on the four-channel Infrared Array Camera (IRAC; Fazio et al. 2004) because this instrument provided the highest spatial resolution of any mid-IR imager with wide-area mapping capabilities over wavelengths from 3 to 9 μm. IRAC far exceeded the point-source sensitivity of WISE in the Galactic plane, because the latter was severely limited by both detector saturation and source confusion. The extensive IRAC observations of the Galactic plane were obtained as part of the Galactic Legacy Infrared Mid-Plane Survey Extraordinaire (GLIMPSE; Benjamin et al. 2003; Churchwell et al. 2009) along with several related Spitzer/IRAC programs that followed similar observing and data processing strategies.

Spitzer has proven effective at identifying candidate YSOs (e.g., Allen et al. 2004; Hartmann et al. 2005; Harvey et al. 2007; Simon et al. 2007; Gutermuth et al. 2009; Povich et al. 2011, 2013, and many others). However, these studies use differing criteria to select YSO candidates, ranging from simple cuts in color space to empirical probabilistic classification to fitting the spectral energy distributions (SEDs) with models of circumstellar dust actively infalling or accreting onto a central stellar object (e.g., Robitaille et al. 2006, 2007; Robitaille 2017). Our study employs a hybrid approach that combines the strengths of SED fitting and principled statistical learning techniques.

An earlier GLIMPSE study ($\ell \sim 295°$–$65°$; Robitaille et al. 2008) identified ~20,000 "intrinsically red sources" ([4.5]–[8.0] $\geqslant 1$), using strict photometric brightness and quality measures to guarantee that the infrared excesses they identify are real. However, they find that this selection criterion is sensitive not only to YSOs but also to intrinsically red contaminants, largely comprised of (post-)asymptotic giant branch stars (AGBs). They find that the 24 μm Spitzer/MIPS band is helpful for distinguishing between these cases. However, this band is not available for the vast majority of point sources detected in GLIMPSE (Gutermuth & Heyer 2015). In our study, we relax these criteria to identify significantly more YSOs candidates, but use patterns in the IRAC photometry to better distinguish between YSOs and contaminants. Our survey area also overlaps YSO catalogs for Cygnus X (Beerer et al. 2010; Winston et al. 2020) and the Spitzer Mapping of the Outer Galaxy (SMOG; Winston et al. 2019) survey; we use the latter as a benchmark to compare to our results.

This paper is organized as follows. Section 2 describes the data sets. Section 3 explains our statistical methodology. Section 4 introduces our YSO catalog. Color–color and color–magnitude diagrams of candidate YSOs and probable contaminants are examined in Section 5. The next sections describe properties of YSO candidates related to environment (Section 6), spatial clustering and kinematics (Section 7), and variability (Section 8). Comparisons with other catalogs are made in Section 9. We provide our conclusions in Section 10.

2. Data

2.1. IRAC catalogs

The YSO selection is largely based on IRAC photometry from GLIMPSE (Benjamin et al. 2003; Churchwell et al. 2009) and related surveys that used similar observing strategies and data reduction methodologies.[8] These include GLIMPSE I (31,184,509 sources), GLIMPSE II (19,067,533 sources), and GLIMPSE 3D (20,403,915 sources), the Vela-Carina (2,001,032 sources) (Majewski et al. 2007; Zasowski et al. 2009), Cygnus X (3,913,559 sources) (Beerer et al. 2010), and SMOG (2,512,099 sources) (Winston et al. 2019) surveys. Spitzer's observations of the Galactic Center (Stolovy et al. 2006) were included in the GLIMPSE II Catalog. We use only the Spitzer photometry obtained during the cryogenic mission, which includes four mid-IR bands centered at 3.6, 4.5, 5.8, and 8.0 μm. We omit the GLIMPSE 360 data from the warm Spitzer mission that includes only the 3.6 and 4.5 μm bands.

The GLIMPSE team designed their reduction pipeline to provide reliable point-spread function (PSF) fitting photometry in crowded fields with spatially varying nebular emission (Benjamin et al. 2003; Kobulnicky et al. 2013)—conditions that are common in IRAC images of star-forming regions. They provide two source lists for each survey, the "Catalog," which is more reliable, and the "Archive," which is more complete.[9] Following, Povich et al. (2013), we use the "Catalog" photometry. We make no additional cuts on quality flags, but certain flags are discussed in Appendix A. The Spitzer/IRAC images have PSFs with full widths at half maximum of 1.″66 at 3.6 μm, 1.″72 at 4.5 μm, 1.″88 at 5.8 μm, and 1.″98 at 8.0 μm. This is significantly better than the ~6″ resolution provided by the WISE survey over a similar wavelength range (Wright et al. 2010), giving IRAC a distinct advantage in crowded fields in the Galactic midplane. The catalogs from the GLIMPSE team also include many stars that are missing from the Spitzer Enhanced Imaging Products (SEIP) due to GLIMPSE's better treatment of mid-IR nebulosity. PSF photometry is also more accurate than SEIP aperture photometry in regions with variable backgrounds (Fang et al. 2020).

The GLIMPSE I, II, 3D, Galactic Center, and Vela-Carina survey observations consisted of (2–5) × 1.2 s integrations at each positions, while the Cygnus X and SMOG surveys used a 0.4 + 10.4 s high dynamic range mode. Given that Cygnus X and SMOG are deeper than the rest of the data, we impose uniformity by omitting sources in these fields that are either brighter or fainter than the limits for the main GLIMPSE survey.[10] The magnitudes of our selected YSO candidates range from [3.6] = 8.3–14.9 mag, [4.5] = 7.3–13.7 mag, [5.6] = 6.4–12.9 mag, and [8.0] = 5.5–12.2 mag (1%–99% quantiles), and the median photometric uncertainties are 0.062, 0.073, 0.075, and 0.060 mag in these bands, respectively.

Even in the full GLIMPSE catalogs, the presence of red sources and several catalog artifacts can be seen in color–color diagrams (Figure 1). The highest concentration of sources have colors close to 0 (expected for normal stars without IR excess), but numerous sources form a distribution extending to the upper right in these plots, which is made up of both YSOs and other red mid-IR sources (e.g., evolved stars, galaxies, etc.; see

[8] The IRAC point-source catalogs were obtained from the NASA/IPAC archive at https://irsa.ipac.caltech.edu/data/SPITZER/GLIMPSE/overview.html.

[9] http://www.astro.wisc.edu/sirtf/docs.html

[10] The bright limits for GLIMPSE are 7, 6.5, 4.0, and 4.0 mag in the 3.6, 4.5, 5.8, and 8.0 μm bands, respectively. The 3σ detection limits are 15.5, 15.0, 13.0, and 13.0 mag in these bands, but completeness declines precipitously before reaching these limit (http://www.astro.wisc.edu/sirtf/GQA-master.pdf). Completeness in the GLIMPSE Catalog is a strong function of both crowding and background sky level, with structure in the background playing a larger role than photon noise (Kobulnicky et al. 2013). In the main GLIMPSE survey, the magnitude distribution of Catalog sources peaks at [4.5] ≈ 13.6 mag, before declining. We adopt a faint limit of [4.5] = 14.5 mag (where density has decreased by a factor of ~5) based on this magnitude distribution and because no source fainter than this is selected as a YSO candidate.

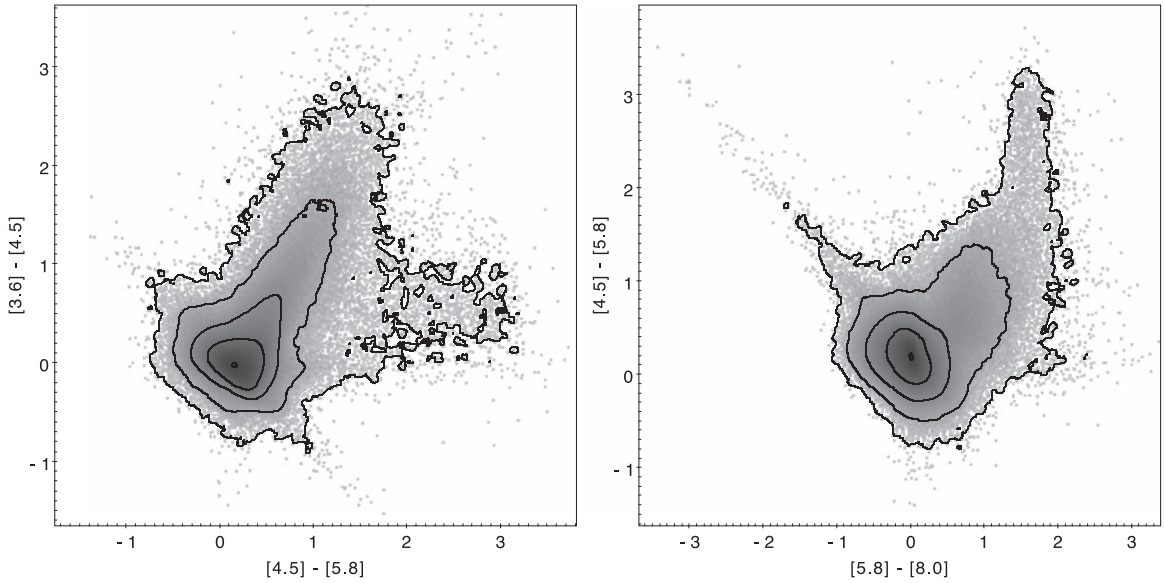

Figure 1. Colors of sources from the GLIMPSE Catalog. (Due to the high number of sources in the full tables, we display a random subsample for plotting convenience.) Contours are drawn at increases in density by a factor of 12. From these plots, we see that the highest source density is at colors ∼0, but in both the [3.6]–[4.5] vs. [4.5]–[8.0] (left) and [4.5]–[5.8] vs. [5.8]–[8.0] (right) diagrams, there is an excess of redder sources to the upper right. In the right panel, two additional features stand out. A streak from the origin to the upper left is an artifact resulting from source-extraction errors in the 5.8 μm band. To the upper right, there is a curved feature in the sources distribution, with [4.5]–[5.8] $\gtrsim$ 1.6 mag and [5.8]–[8.0] $\approx 1.6 \pm 0.25$ mag. We argue that these colors are affected by PAH emission (Section 5.7).

Appendix B). In the [4.5]–[5.8] versus [5.8]–[8.0] diagram, a streak can be seen extending from the origin to the upper left. This streak appears to be related to erroneous photometry in the 5.8 μm band for a low fraction of the GLIMPSE sources, and it extends from the origin because this is where the source density is highest. Another prominent feature in this diagram is a finger-like structure extending upward at [5.8]–[8.0] ≈ 1.6, which we attribute to polycyclic aromatic hydrocarbon (PAH) emission (Section 5.7).

2.2. Cross-matches to Near-IR Catalogs

Near-infrared JHK_s photometry from the Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006) is already included in the GLIMPSE (and extensions) data products. 2MASS has a spatial resolution of $\sim 2''$, which is comparable to the Spitzer/IRAC PSF. For our sample, 2MASS is nearly complete down to $J \sim 15.4$ mag, $H \sim 14.2$ mag, and $K_s \sim 13.0$ mag, with median photometric uncertainties of 0.038, 0.040, and 0.035, respectively. While these limiting magnitudes correspond well with the limits of the GLIMPSE surveys, in practice YSOs are often found in regions of high interstellar reddening, where 2MASS may not be deep enough to detect NIR counterparts of red GLIMPSE sources.

Deeper NIR catalogs with higher spatial resolution are available from the United Kingdom Infra-Red Telescope (UKIRT) Infrared Deep Sky Survey (UKIDSS; Lawrence et al. 2007) and the Visible and Infrared Survey Telescope for Astronomy (VISTA) Variables in the Vía Láctea survey (VVV; Minniti et al. 2010) for the northern and southern portions of the Galactic plane, respectively, with overlap around the Galactic center. These catalogs are deeper than 2MASS, but are saturated for brighter sources. We use the UKIDSS catalog from the Galactic Plane Survey (Lucas et al. 2008) and the averaged VVV photometry for multiple epochs from the VVV Infrared Astrometric Catalog (VIRAC DR1; Smith et al. 2018). For both deeper NIR surveys, the photometry extends to $J \sim 19$ mag, $H \sim 18$ mag, and $K_s \sim 16$ mag, with formal photometric uncertainties <0.01 mag.

IRAC and UKIDSS/VVV were cross-matched using a $1''$ match radius in TOPCAT (Taylor 2005). Photometric measurements from UKIDSS/VVV were omitted if they did not have the flag $mergedClass = -1$ (stellar) or if they had magnitudes $J < 11$, $H < 12$, $K < 10.5$ (UKIDSS) or $J < 12.5$, $H < 13$, $K_s < 11.5$ (VVV), for which saturation effects start to affect photometry. The higher spatial resolutions of the NIR catalogs mean that it is possible for multiple NIR sources to be associated with individual IRAC sources; however, examination of IRAC+UKIDSS matching by Morales & Robitaille (2017) revealed that the NIR flux is usually dominated by a single counterpart.

In our analysis, we perform the YSO candidate selection independently on cross-matches of IRAC+2MASS, IRAC+UKIDSS, and IRAC+VVV, and the results of these separate selections are merged.

2.3. Ancillary Data

The Gaia mission (Gaia Collaboration et al. 2016), in its second data release (Gaia DR2; Gaia Collaboration et al. 2018), has provided optical broadband photometry (Evans et al. 2018) for the whole sky along with exquisite astrometric measurements (Lindegren et al. 2018) for more than 1.3 billion stars. These data on their own can be used for selecting possible pre–main-sequence stars (e.g., Zari et al. 2018), but for our study we use them as ancillary data in order to better understand the parallax (ϖ) and proper motion ($\mu_{\alpha*}$, μ_δ), distributions of the IR-excess selected YSO candidates. From the YSOs candidate list (Section 4), 33% have Gaia counterparts with the full five-parameter astrometric solution (Lindegren et al. 2018). A match rate below 50% is expected because many YSOs are enshrouded by dust and thus not optically visible.

Longer-wavelength photometry is available from both Spitzer's MIPS Galactic Plane Survey (MIPSGAL; Carey et al. 2009; Gutermuth & Heyer 2015) at 24 μm and the WISE All-Sky Data Release (Wright et al. 2010) at 22 μm. In the Galactic plane,

AllWISE is affected by high numbers of spurious sources, particularly in the longer-wavelength bands, so we follow the catalog cleaning recommendations from Koenig & Leisawitz (2014) and apply the signal-to-noise and χ^2 quality cuts on the profile-fit photometry given in their Equations (1)–(4). For MIPSGAL, we use the "Catalog" instead of the "Archive." The WISE photometry, and to a lesser extent the MIPS photometry, is strongly affected by crowding and nebulosity in the regions around the Galactic plane that we are investigating, leading to fewer reliable source detections in these areas. The application of the quality cuts from Koenig & Leisawitz (2014) leave visible holes in the spatial distribution of WISE sources surrounding the clusters of YSO candidates that we identify with the IRAC photometry. Only 30% of our IRAC candidates have counterparts in the 24 μm MIPS band, and 8% have reliable 22 μm WISE photometry. Because of the unavailability of longer-wavelength data for the majority of our sample, we do not use these bands for identifying candidates, but only for post-selection examination of the sample. For both catalogs, we used a cross-match radius of 1.″2.

2.4. Published YSO Catalogs

YSOs identified in earlier studies of star-forming regions within GLIMPSE (and extensions) can be used to train a classifier to find similar types of objects. We use YSOs identified as part of the Massive Young Star-Forming Complex Study in Infrared and X-ray (MYStIX; Feigelson et al. 2013), in addition to a similar, earlier study of the Carina Nebula (Townsley et al. 2011). From the combined lists, we have included probable YSOs from the Carina Nebula, NGC 6611, M17, NGC 6530, M20, NGC 6357, NGC 6334, RCW 38, RCW 36, and DR 21, ranging from ~0.7 to 2.7 kpc in distance. Povich et al. (2011, 2013) performed the IR excess detection for these projects using Spitzer data, based on a strategy that included SED fitting of both reddened stellar atmospheres and the Robitaille YSO models, color cuts to remove certain types of contaminants, spatial filtering to remove objects that are not clustered, and visual examination of SEDs.

We select all objects from the MYStIX IR-excess catalogs classified as both a YSO ($Cl = 0$) and a probable member ($Mm = 1$). More recently, Gaia DR2 has become available, so for the subset of sources with Gaia parallaxes (~30% of the sample), we refine the sample further by removing any source with a parallax that is discrepant from the median parallax of the group by >2 times the reported parallax error.

In addition to the aforementioned studies of multiple regions and large areas, GLIMPSE has been used in hundreds of papers about individual (or several) star-forming regions. A few representative examples of these include Zavagno et al. (2006), Watson et al. (2008), Povich et al. (2009), Dewangan & Ojha (2013), Samal et al. (2014), Mallick et al. (2015), and Povich et al. (2016).

3. Methodology

YSOs make up a minuscule fraction of the nearly fifty million sources detected in the Spitzer/IRAC surveys included in this project. This means that selection of YSO candidates requires rejection of numerous contaminants (mostly field stars) along similar lines of sight. The first steps in our procedure, in which we reject sources that can be explained without IR excess, are nearly identical to those from Povich et al. (2013). These steps greatly reduce the sample size and are based on well-established stellar atmosphere models. However, in the next steps—classification of the remaining sources—rather than fitting models of YSO SEDs as Povich et al. (2013) do, we use their resulting MYStIX YSO sample to train our random forest classifier.

A data-driven approach offers some advantages. For instance, we use IRAC photometry of actual stars as a training set instead of artificial photometry generated from theoretical YSO models. This means that the method will tend to avoid classifying an object with unusual colors as a YSO even if these colors can be reproduced by a physically unrealistic configuration of a star, disk, and envelope that exists in a grid of theoretical YSO models. Furthermore, it takes significantly less computational time to apply a trained classifier to millions of stars than it does to fit each of them with several categories of parametric YSO model. Nevertheless, the YSO SED fitting method does play an important role in generating training sets for the classifier (Section 3.2).

3.1. Removing Sources without Significant IR Excess

In the Galactic midplane, many background stars are affected by high levels of foreground extinction, so any source that is either insufficiently red or whose red colors can plausibly be explained by reddening alone is dropped from further scrutiny.

Cuts on IRAC colors and color uncertainties can remove many objects that have no chance of being selected as reliable IR excess objects. We apply the rules recommended by Povich et al. (2011, 2013), decreasing the number of sources in our sample by a factor of ~10. All retained sources must be detected in at least four out of the seven IR bands, two of which must be 3.6 and 4.5 μm. Sources are kept if there is the suggestion of IR excess in the [3.6]–[4.5] color using the criterion

$$[3.6] - [4.5] - 0.408 > \mathrm{error}([3.6] - [4.5]), \quad (1)$$

where "error" denotes uncertainty in color, calculated by adding the photometric uncertainties for the two bands in quadrature. The value 0.408 is the expected reddening of this color with $A_V \approx 30$ mag of extinction. Sources are also kept if they have have photometric measurements in the 5.8 and 8.0 μm bands and either meet both the criteria

$$|[4.5] - -[5.8]| > \mathrm{error}([4.5] - -[5.8]) \quad (2)$$

$$[5.8] - -[8.0] > \mathrm{error}([5.8] - -[8.0]), \quad (3)$$

or

$$|[4.5] - -[5.8]| \leqslant \mathrm{error}([4.5] - -[5.8]) \quad (4)$$

$$|[5.8] - -[8.0]| \leqslant \mathrm{error}([5.8] - -[8.0]). \quad (5)$$

These rules, optimized from experience with GLIMPSE data, ensure that determination of IR excess is based on more than just the 8.0 μm band, which can occasionally give a spuriously bright measurement.

We fit the *JHK*+IRAC SEDs of the remaining sources with reddened Castelli & Kurucz (2003) stellar atmosphere models, using the Indebetouw et al. (2005) extinction law. The fitting procedure takes into account the statistical photometric uncertainties on the data, which happen to be of similar size for both 2MASS and IRAC photometry. The UKIDSS and VVV data sets provide *JHK* photometry for many objects that were not detected in 2MASS, allowing many more sources to

be included. However, the statistical measurement uncertainties for most UKIDSS and VVV sources are far more precise than for 2MASS or IRAC. Given that we are mostly interested in detecting deviations from a reddened stellar atmosphere model in the IRAC bands and we want similar selection performance for each data set, we rescale all UKIDSS and VVV error bars that are smaller than the median 2MASS error bars to be equal to the median 2MASS error bars. The sources that are poorly fit by the reddened stellar photosphere, with χ^2 per data point >4, comprise the target set for our random forest classifier.

Overall, these pruning steps leave 319,251 2MASS+IRAC, 188,701 UKIDSS+IRAC, and 257,334 VVV+IRAC sources with possible IR excess as inputs to our classification step below.

3.2. Training Sets

The training data includes both MYStIX IR-excess sources (Section 2.4) that we label "YSO" and sources unlikely to be YSOs that we label "contaminant" (see discussion of contaminants below). Although lists of members are more complete in some of the nearest star-forming regions (e.g., Ophiuchus or Taurus; see Evans et al. (2009b) and Luhman (2018)), we choose to use MYStIX because these massive star-forming complexes may better represent the regions we expect to probe in the Galactic midplane, i.e., at greater distances, with higher extinction, and in more extreme environments. Furthermore, many of the MYStIX regions lie within the survey region of GLIMPSE and its extensions, meaning that homogeneous data products are available for both the training and target sets.

Contaminants can include both sources that occur in star-forming regions (e.g., nonstellar sources such as nebular knots and shocked emission) and sources that are smoothly distributed on the sky (e.g., AGB stars and galaxies; see Robitaille et al. (2008) and Gutermuth et al. (2009)). Refer to Appendix B for discussion of these objects. Povich et al. (2011, 2013) used a variety of techniques to remove these objects from their catalogs, including SED fitting of Robitaille et al. (2007) models, color cuts, and visual inspection. We apply the term "field object" to any object within the field of view of these studies that was not classified as a probable young star by either IR or X-ray criteria.

To enlarge our sample of contaminants, we identify several fields near the Galactic plane that have no signs of star formation (Appendix B.3) and label objects in these fields as non-YSOs. These fields were selected to include lines of sight at multiple Galactic longitudes, in the midplane and up to several degrees above or below it, and with different amounts of Galactic extinction.

Training sets are generated separately for each combination of NIR+IRAC data due to the differences in NIR filters. For 2MASS+IRAC, the training set contains 2865 YSOs, 3436 field objects in the MYStIX fields, and 7718 other field objects for the 2MASS+IRAC data set. For UKIDSS+IRAC these numbers are 919, 2000, and 1128, and for VVV+IRAC they are 1266, 1459, and 2595, respectively.

The distributions of training set objects in color space are discussed in Appendix C. The full IRAC catalog of $\sim 5 \times 10^7$ sources includes a few outliers in region of color space that are not well-sampled by either the labeled YSOs or labeled non-YSOs in the training set. (The limits used to identify these outliers are given in Appendix C.) Given that we have little basis to assign such objects to either category, we opt to be cautious and do not include these objects in our final YSO list.

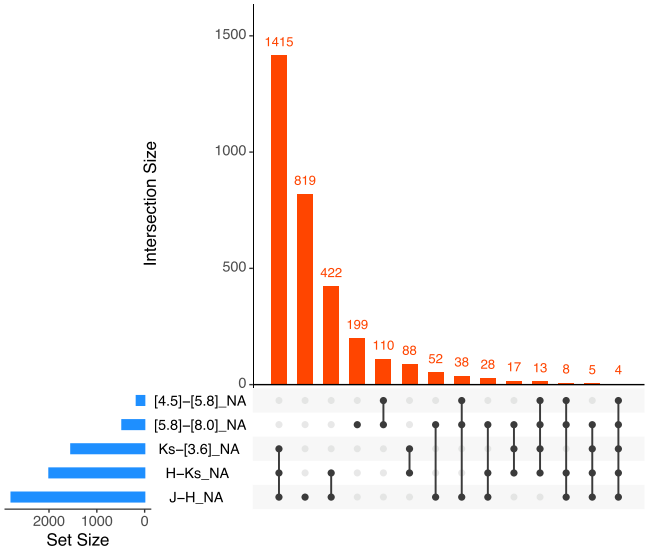

Figure 2. Missing data pattern for the labeled 2MASS+IRAC training set. Blue bars are the number of missing colors. Connected black dots indicate combinations of missing colors. Red histograms indicate the number of instances these combinations are missing. The UKIDSS and VVV diagrams are provided in the figure set.

(The complete figure set (3 images) is available.)

3.3. Missing Data

When data are combined across multiple catalogs, it is almost certain that missing data will occur, as is the case here. Figure 2 depicts the missing pattern for the 2MASS+IRAC, UKIDSS+IRAC, and VVV+IRAC training sets, from which only 57%, 46%, and 29% of objects, respectively, have complete information. About 20% of the 2MASS+IRAC objects are missing three colors at once, and JHK_s are often missing together. The VVV+IRAC data set has the most missing data, with 57% of the objects missing at least three bands. While UKIDSS+IRAC is the most complete, more than 35% of their rows have at least two missing colors. Thus, a naive removal of rows presenting missing values would throw away a non-negligible amount of valuable information.

As a final pre-processing step before training our YSO classifier, we employed a multiple-copula imputation. This decomposes joint probability distributions into their marginal distributions and a function, the copula, that couples them (Nelsen 2010). Copulas have been used previously in astronomy—for example, to construct likelihood functions for weak lensing analysis (Sato et al. 2011; Lin et al. 2016) and to infer bivariate luminosity and mass functions (Andreani et al. 2018). Previous tests suggest that this method outperforms other popular approaches, such as multiple imputation via chained equations (van Buuren & Groothuis-Oudshoorn 2011) and Amelia (Honaker et al. 2011), in terms of bias and coverage, especially in cases where the variables are not normally distributed (Hoff 2007). The underlying idea of copula imputation is to derive conditional density functions of the missing variables given the observed ones through the corresponding conditional copulas, and then impute missing values by drawing observations from them. Finally, the choice of performing imputation before training the random forest models has been previously assessed by other studies (Jaeger et al. 2020), which have shown it to reduce the variance in model error estimate, without any detectable change in

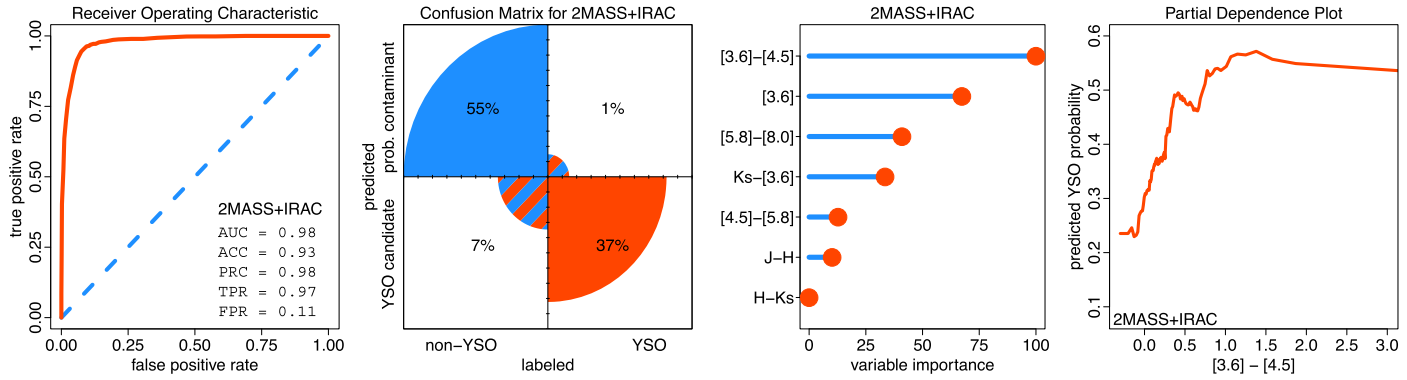

Figure 3. Diagnostic plots for the 2MASS+IRAC classifier (plots for additional variables and for the UKIDSS and VVV classifiers are included in the figure set). Far left: ROC curve. Values are provided for area under the curve (AUC), accuracy (ACC), precision (PRC), true-positive rate (TPR), and false-positive rate (FPR). Center left: Confusion matrix. Wedge areas are proportional to the numbers of objects in each category. Center right: Estimated importance of colors and magnitudes in the random forest model. Far right: Partial dependence plot for the [3.6]–[4.5] color.

(The complete figure set (30 images) is available.)

precision. The imputation method was implemented using the SBGCOP package (Hoff 2018) within the R language (Core Team 2019). Copulas were fit simultaneously to both training and target data sets.

Overall, the imputed data preserves the coverage of the original data set (Appendix C). Nevertheless, we do not advocate for the use of these colors in other contexts. They are treated here as nuisance parameters to enable classification of the entire data set.

3.4. Tree-based Classification

Decision trees are learning algorithms that resemble the natural flow of human decision making. At each node of the tree, the algorithm randomly selects one feature, and based on the distribution of training data, determines the decision boundaries that best separate different classes. Training objects are then propagated along their branch of the tree to the next node, where a new feature is selected. The process is repeated until the tree reaches a predetermined depth or until all objects in a leaf belong to the same class; for a more detailed description, see Rokach & Maimon (2014). This basic concept has given rise to successful algorithms in many different fields. However, a single decision tree trained on an entire data set is prone to overfitting, presenting low-accuracy results whenever faced with data not used in training. This problem can be overcome by randomizing different stages of the tree construction and combining many independent estimators in a more robust classifier. This type of approach belongs to the wider class of ensemble models.

Ensemble methods (e.g., Sagi & Rokach 2018) are regression algorithms, constructed from the combination of many weak classifiers that, when considered together, provide a more robust estimate than any of their constituents. Random forests (Ho 1995; Breiman 2001) are one such algorithm, composed of many decision trees, each constructed independently. The final classification is determined via majority vote, considering all trees in the forest. In this context, the probability of being a YSO is approximated by the percentage of trees in the ensemble voting for a YSO candidate—we call this probability estimate the "YSO score." Random forests have been successfully used to classify YSOs in smaller-scale studies, including with missing data imputation (e.g., Ducourant et al. 2017; Melton 2020).

We construct the YSO random forest classifiers using the following covariates: $J-H$, $H-K_s$, K_s–[3.6], [3.6]–[4.5], [4.5]–[5.8], [5.8]–[8.0], and the 3.6 μm band magnitude. To mitigate the influence of class imbalance upon tree optimization, we employed an upsampling method during the training process. Overall, this does not have a perceptible effect on the classification performance, but it does push the decision boundary probability between classes closer to 0.5. Each model was independently trained for 2MASS + IRAC, UKIDSS + IRAC, and VVV + IRAC data sets. The random forest was employed using the CARET R package, with 1500 trees, which was sufficient to guarantee stable solutions. As a sanity check, we tested few other regression models (including generalized additive models, support vector machine, gradient boosting machines, and conditional random forest), but no significant difference in the final YSO candidate set was found. This suggests that random forests (or other typical nonlinear classifiers) capture the data complexity well enough without the need for highly complex models.

3.5. Classifier Performance

We assessed the performance of the classifiers by using validation tests in which we partitioned the labeled objects into a training set (80%) and a test set (20%).

Figure 3 (far left) shows true-positive rate (TPR) as a function of false-positive rate (FPR), a curve often called the receiver operating characteristic (ROC). The area under the curve (AUC) can range from 0 to 1, where random guessing gives an area of 0.5, and higher values indicate better performance. Our classifiers all have AUC > 0.9. A threshold on the YSO score of $p = 0.5$ gives high TPRs and low FPRs for each classifier, meaning that this threshold is effective for distinguishing between the YSO and non-YSO classes. Other performance measures are listed on the plot, including accuracy (ACC), i.e., the number of true positives plus true negatives divided by the total population, and precision (PRC), i.e., the number of true positives divided by the number of true positives plus false positives. The confusion matrix (Figure 3, center left) shows few misclassifications.

Variable importance (Figure 3, center right) is evaluated via out-of-bag samples, which consist of random samplings of the data that are left out of each tree. These are calculated by measuring variations in the prediction error when the out-of-bag data are permuted solely among a specific color, leaving the others unchanged. The process is then repeated across all trees. The final result is a measure of the incremental error for a given color when compared with the unperturbed colors for all the 1500 trees over the entire forest.

Figure 3 (far right) displays a partial dependence plot (PDP; Greenwell 2017) for [3.6]–[4.5] color (plots for other variables are included in the figure set). PDPs are useful for visualizing the relationship between individual features and the response while accounting for the average effect of the other predictors in the model. The shape and steepness of the curves are indicators of the predictor's relative influence. Note the sharp behavior of [3.6]–[4.5], one of the best indicators of YSO candidates.

4. Catalog

All objects with YSO scores >50% from any of the three random forests are classified as candidate YSOs, while other sources are regarded as probable contaminants. Among the Spitzer sources with IR excess, there are 117,446 candidate YSOs and 180,997 probable contaminants. The candidates are listed in Table 1 with the designation Spitzer/IRAC Candidate YSO (SPICY).

Figure 4 shows how the candidates are distributed within the footprints of the Spitzer surveys. Many of the candidate YSOs are concentrated toward the Galactic midplane, while others form prominent clumps.

More detail is visible in the zoomed-in maps from the atlas (Figure 5), which consist of 19 panels ($\sim 6^\circ \times 12^\circ$) covering the entire survey area. In this atlas, we have also marked other features possibly connected to star formation, including H II bubbles from WISE (Anderson et al. 2014) and massive YSOs from MSX (Lumsden et al. 2013). These maps show that the YSO candidate distribution can be resolved into stellar clusters and associations, along with a non-negligible number of widely distributed objects. It also appears that dense groups of YSOs tend to coincide with the locations of bubbles as seen in the plane of the sky.

With the new YSO candidates, some previously unrecognized stellar groups become apparent. In Figure 6, we show an image containing one such group located in the Vela-Carina portion of the survey and designated G271.6-0.5 (left side of the image). To the southwest of this group, a previously identified but little-studied star-forming region, G271.1-0.8, can also be seen.

The SPICY catalog is the largest homogeneous sample of YSO candidates available to date for the inner regions of the Milky Way. It seems unlikely that this mid-IR list of YSOs will be superseded in the near future, given that no existing or planned mid-IR instrument exceeds Spitzer's spatial resolution in tandem with its wide-area mapping capabilities. The catalog is intended for use both in addressing questions about star formation on Galactic scales and for assisting in searches for interesting individual YSOs. However, some contaminants inevitably remain, and formal assessment of contamination requires follow-up observations (e.g., spectrographic surveys). Nevertheless, the properties of these stars, including their colors, the environments in which they are found, their spatial and kinematic distributions, and their photometric variability (discussed in Sections 5–8), are useful for corroborating the results of the random forest classifier and may give a qualitative sense of the level of remaining contamination.

Table 1
Candidate YSOs

Column	Column ID	Description
1	SPICY	Candidate YSO designation
2	RAdeg	ICRS R.A. coordinate in decimal degrees
3	DEdeg	ICRS decl. coordinate in decimal degrees
4	GLON	Galactic longitude
5	GLAT	Galactic latitude
6	p1	YSO random forest score[a] from IRAC+2MASS photometry
7	p2	YSO random forest score[a] from IRAC+UKIDSS photometry
8	p3	YSO random forest score[a] from IRAC+VVV photometry
9	class	YSO class[b]
10	silicate	Flag for a possible strong silicate feature
11	pah	Flag for a possible strong PAH feature
12	alpha	Spectral index used for YSO class[c]
13	alpha8	Spectral index derived from the [4.5]–[8.0] color
14	alpha24	Spectral index derived from the [4.5]–[24] color
15	alphaW4	Spectral index derived from the [4.5]−W4 color
16	env	Classification of the YSO environment[d] from the $3' \times 3'$ IRAC cutout
17	group	HDBSCAN group to which the star is assigned[e]
18	var	ZTF light-curve variability flag[f]
19	n_ZTFrmag	Number of good ZTF *r*-band observations used
20	ZTFrmag	ZTF mean magnitude in the *r*-band
21	sigma	ZTF light-curve $\sigma_{\rm var}$ standard deviation in the *r*band
22	skewness	ZTF light-curve skew in the *r*band
GLIMPSE (and Extensions) Catalog Columns		
23	Spitzer	Spitzer source designation
24	3.6mag	Spitzer/IRAC channel 1 magnitude
25	e_3.6mag	Error on Spitzer/IRAC channel 1 magnitude
26	4.5mag	Spitzer/IRAC channel 2 magnitude
27	e_4.5mag	Error on Spitzer/IRAC channel 2 magnitude
28	5.8mag	Spitzer/IRAC channel 3 magnitude
29	e_5.8mag	Error on Spitzer/IRAC channel 3 magnitude
30	8.0mag	Spitzer/IRAC channel 4 magnitude
31	e_8.0mag	Error on Spitzer/IRAC channel 4 magnitude
32	csf	Close source flag
33	n_3.6mag	Number of detections in the 3.6 μm band
34	n_4.5mag	Number of detections in the 4.5 μm band
35	n_5.8mag	Number of detections in the 5.8 μm band
36	n_8.0mag	Number of detections in the 8.0 μm band
Cross-matched Catalogs		
37	2MASS	2MASS source designation
38	UKIDSS	UKIDSS source designation
39	VIRAC	VIRAC DR1 source designation
40	GaiaDR2	Gaia DR2 source designation
41	MIPS	Spitzer/MIPS source designation
42	AllWISE	AllWISE source designation
43	ZTFDR3	ZTF DR3 source designation

Notes. In addition to the quantities derived in this paper, this table also, for the convenience of the user, provides select columns from the GLIMPSE (and extensions) catalogs.
[a] Random forest scores range from 0 to 1, where higher scores indicate a greater chance that an object is a YSO. We include sources with scores >0.5 from one of the classifiers.
[b] YSO classes are "Class I," "FS" (flat SED), "Class II," and "Class III."
[c] This α combines the results from Columns 13–15 as described in Section 5.5.
[d] The environment classes are "EnvI" (no or minimal nebulosity), "EnvII" (mixed category), "EnvIII" (cloud-like environment).
[e] The list of groups is provided in Table 2.
[f] The variability classes are 1 (weak or statistically insignificant variability), 2 (moderate variability), 3 (high variability).
(This table is available in its entirety in FITS format.)

5. Color and Magnitude Distributions

By examining the IR color and magnitude distributions for classified objects, we gain insight into how the classifier makes its decisions and how it compares to other selection criteria used in previous studies.

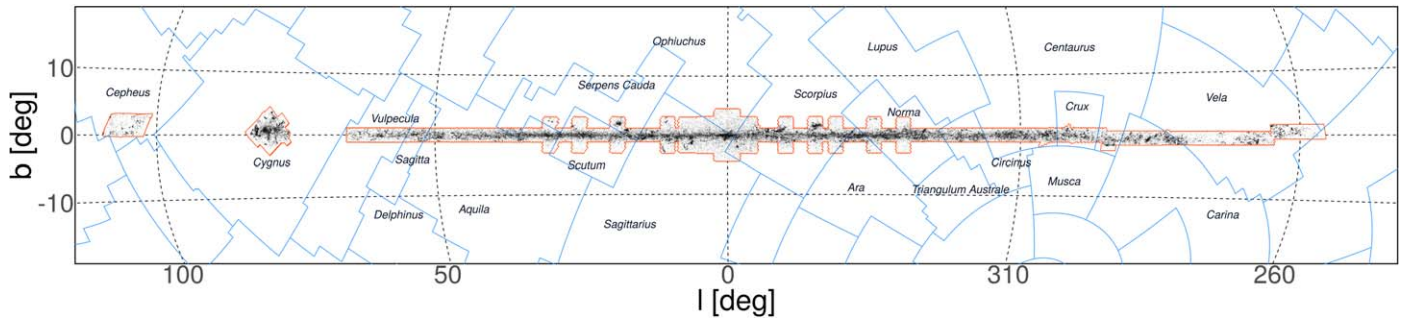

Figure 4. Spatial distribution of the candidate YSOs (black points) within the Spitzer/IRAC survey regions (outlined in red). Data are plotted in Galactic coordinates, and constellation boundaries are shown in blue. Candidate YSOs tend to be concentrated toward the midplane and/or in spatial clusters.

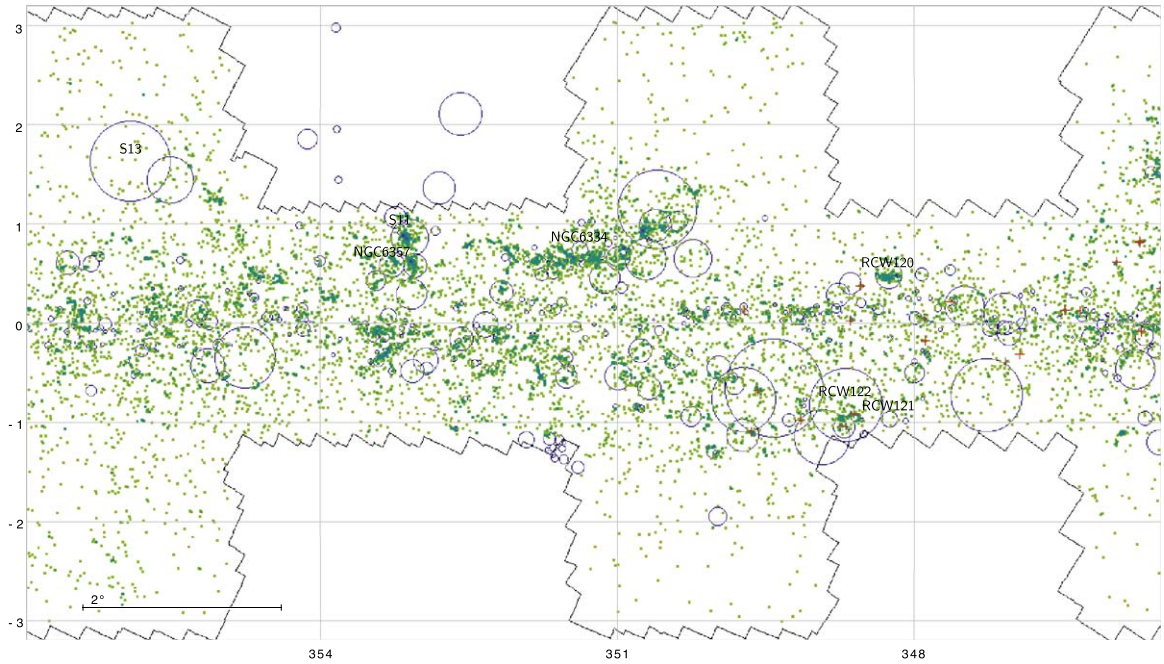

Figure 5. This figure set (19 components) provides an atlas of the Galactic midplane with locations of the YSO candidates (green points) and the boundaries of the IRAC surveys (black lines). Overlapping points produce darker shades of green, using a square-root scale and the "viridis" color palette. For context, we also show outlines of H II bubbles from WISE (blue circles; Anderson et al. 2014), massive YSOs from MSX (red crosses; Lumsden et al. 2013), and labels of select star-forming regions.

(The complete figure set (19 images) is available.)

Figure 7 shows the distribution of [4.5]–[8.0], one of the main features used in the earlier study by Robitaille et al. (2008). The sources we input into the classifier have a bimodal distribution in this color, but each of the output classes has a unimodal distribution, with the probable contaminants making up the bluer peak and the YSO candidates making up the redder peak. The densities are approximately equal at [4.5]–[8.0] ≈ 1, the threshold used by Robitaille et al. (2008), although we also find a substantial number of objects of both classes (but particularly the contaminants) crossing the threshold.

The four panels of Figure 8 show the magnitude distributions of the classified sources in each IRAC band. The input distributions are bimodal, with lower peaks near the brightness limits and higher peaks at fainter magnitudes. The peaks at bright magnitudes can be attributed to an artifact of the χ^2 fitting step, because bright sources tend to have smaller magnitude uncertainties—and thus a smaller deviation is capable of leading to a formally "bad fit." The classifier has identified the majority of sources associated with the bright peaks as probable contaminants. The distributions of the candidate YSOs are all unimodal, with peaks at fairly faint magnitudes, and heavy tails extending to brighter magnitudes. The contaminants also exhibit a second peak at magnitudes slightly fainter than the peak for YSOs.

The YSO magnitude distributions appear reasonable, given that we would expect most of them to be low-to-intermediate mass objects at distances of one to several kpc, with a low number of brighter objects that could either be massive YSOs or nearby objects. The tendency to classify the faintest objects as probable contaminants may inherit a bias from the MYStIX training set, which only includes YSOs out to $\sim$3 kpc. However, the differences in colors of the faintest objects (examined below) imply that they may be intrinsically different.

Figures 9–12 show various *JHK*+IRAC color–magnitude and color–color diagrams. Candidate YSOs (red points) overlap

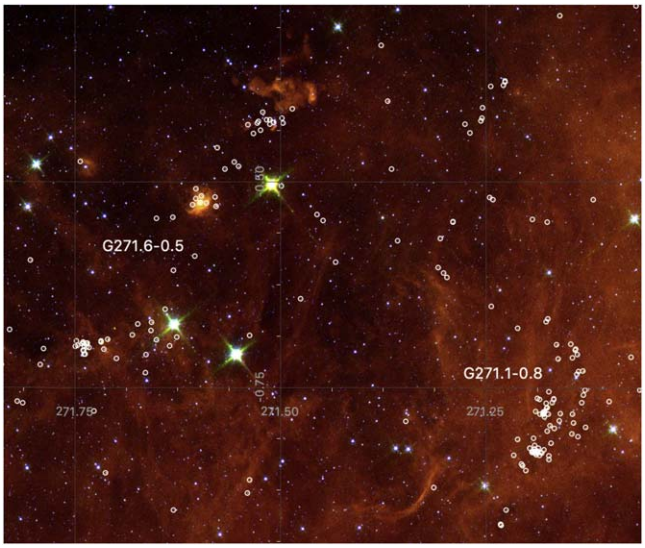

Figure 6. Spitzer/IRAC image (Vela-Carina survey) with our YSO candidates marked by the white circles. Image is composed of the 3.6 μm (blue), 5.8 μm (green), and 8.0 μm (red) images. The image captures two groups of stars: the previously unstudied group, G271.6-0.5, and a neighboring group, G271.1-0.8.

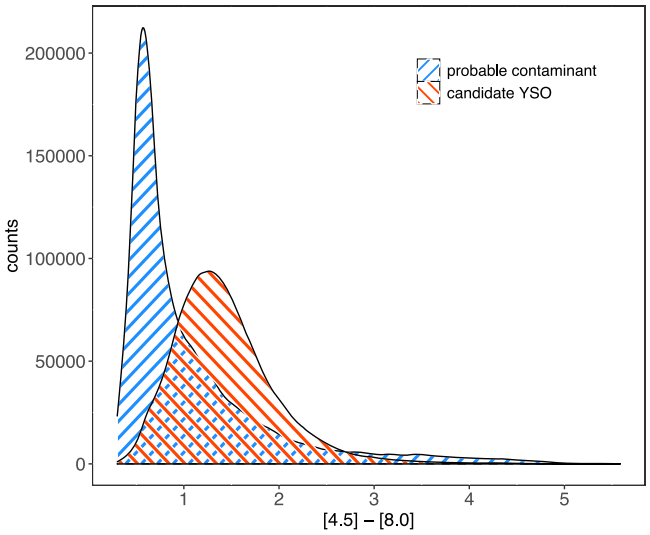

Figure 7. Distributions of [4.5]–[8.0] color for the candidate YSOs (red stripes) and the probable contaminants (blue stripes). Overall, the candidate YSOs tend to be redder than the probable contaminants. Densities of sources in both samples are approximately equal at [4.5]–[8.0] $\approx$ 1 mag, the limit imposed in the study by Robitaille et al. (2008), but in our sample 18% of the YSO candidates are bluer than this limit and 25% of the probable contaminants are redder. Probable contaminants also outnumber candidate YSOs at colors [4.5]–[8.0] $\gtrsim$ 3.5 mag.

probable contaminants (blue points) in each of these projections. Nevertheless, the locations in these diagrams with greatest source density are different for the two classes. We show reddening vectors indicating the effect of $A_K \approx 1$ mag ($\sim$9 mag in the V band) of extinction, adopting the reddening law from Rieke & Lebofsky (1985) for *JHK* and Indebetouw et al. (2005) for the IRAC bands. We also plot curves for the near-IR stellar colors for stellar models without additional IR excess. For graphical display, we have merged the 2MASS, UKIDSS, and VVV photometry, converting UKIDSS and VVV to the 2MASS system using the first-order transformations from Hodgkin et al. (2009) and Soto et al. (2013), and picking the most reliable photometry for each source.

Some of these color spaces have been used in previous studies for selecting YSOs based on cuts on color. For example, the selection boundaries between YSOs and contaminants used by Gutermuth et al. (2009) are depicted as gray lines in several of the diagrams, including [4.5] versus [4.5]–[8.0] (Figure 9, left panel), [3.6]–[4.5] versus [4.5]–[5.8], and [4.5]–[5.8] versus [5.8]–[8.0] (Figure 11, upper panels).

In the following subsections, we examine the IR criteria used for classification, evidence from Gaia that stars are pre–main-sequence, properties of the stars at 24 μm, YSO evolutionary classes, and the effects of various IR absorption and emission features.

5.1. Color–Magnitude Diagrams

On the J versus $J-H$ diagram (Figure 9, left), both candidate YSOs and probable contaminants occupy a triangular region of color–magnitude space, where the upper edge of the triangle is approximately parallel to the reddening vector. The YSO candidates are densest around $J \sim 15.5$ mag and $J-H \sim 1.3$ mag, whereas the probable contaminant distribution is multimodal, with one peak just blueward of the peak of the YSO candidates, and another strip of stars along the upper right edge of the triangle. The stars in this strip, which are more luminous than the typical YSO candidate with the same $J-H$ color, lie in the region of the diagram that would be occupied by reddened post–main-sequence stars.

On the [4.5] versus [4.5]–[8.0] diagram (Figure 9, right), the YSO candidates form a smooth distribution ranging from the bright limit at [4.5] = 6.5 mag to $\sim$14 mag, where sensitivity declines, with the peak of the distribution at [4.5] $\sim$ 12.2 mag and [4.5]–[8.0] $\sim$ 1.2. A $\sim$ 1 Myr old (pre–)main-sequence star with a mass in the range 0.4–10 $M_\odot$ at a distance of $\sim$1–2 kpc would have an unreddened photospheric magnitude $9 \lesssim [4.5] \lesssim 14$ mag (Bressan et al. 2012)—approximately where were find the bulk of the YSO candidates. The probable contaminant distribution peaks at both bright and faint magnitudes. The bright contaminants form a band that tends to be bluer than the YSOs in [4.5]–[8.0], while the faint contaminants tend to have redder [4.5]–[8.0] colors.

The gray lines on the [4.5] versus [4.5]–[8.0] diagram were defined by Gutermuth et al. (2009) to separate dusty AGNs from YSOs in their studies of nearby star-forming regions. Although many of the faintest 4.5 μm sources in our sample have been classified as probable contaminants, the region defined by Gutermuth et al. (2009) for selecting AGNs does not appear to separate our classes well. This apparent discrepancy may arise because Gutermuth et al. (2009) examined deeper Spitzer surveys of relatively nearby star-forming clouds at higher Galactic latitudes, where more AGN are expected to be detected, whereas GLIMPSE is less sensitive to this type of contaminant. Furthermore, GLIMPSE includes more distant star-forming regions in which legitimate YSOs will present fainter observed [4.5] magnitude distributions.

5.2. Color–Color Diagrams

Figure 10 shows the distributions of sources in $J-H$, $H-K_s$, and H–[4.5]. On the JHK_s diagram, we include a representative isochrone for $\sim$1 Myr unreddened stellar models. Most of the objects are shifted to the upper right from this curve, in the approximate direction of the reddening vector. However, the distribution of the YSO candidates spreads to redder $H-K_s$ colors, which would be expected for stars with

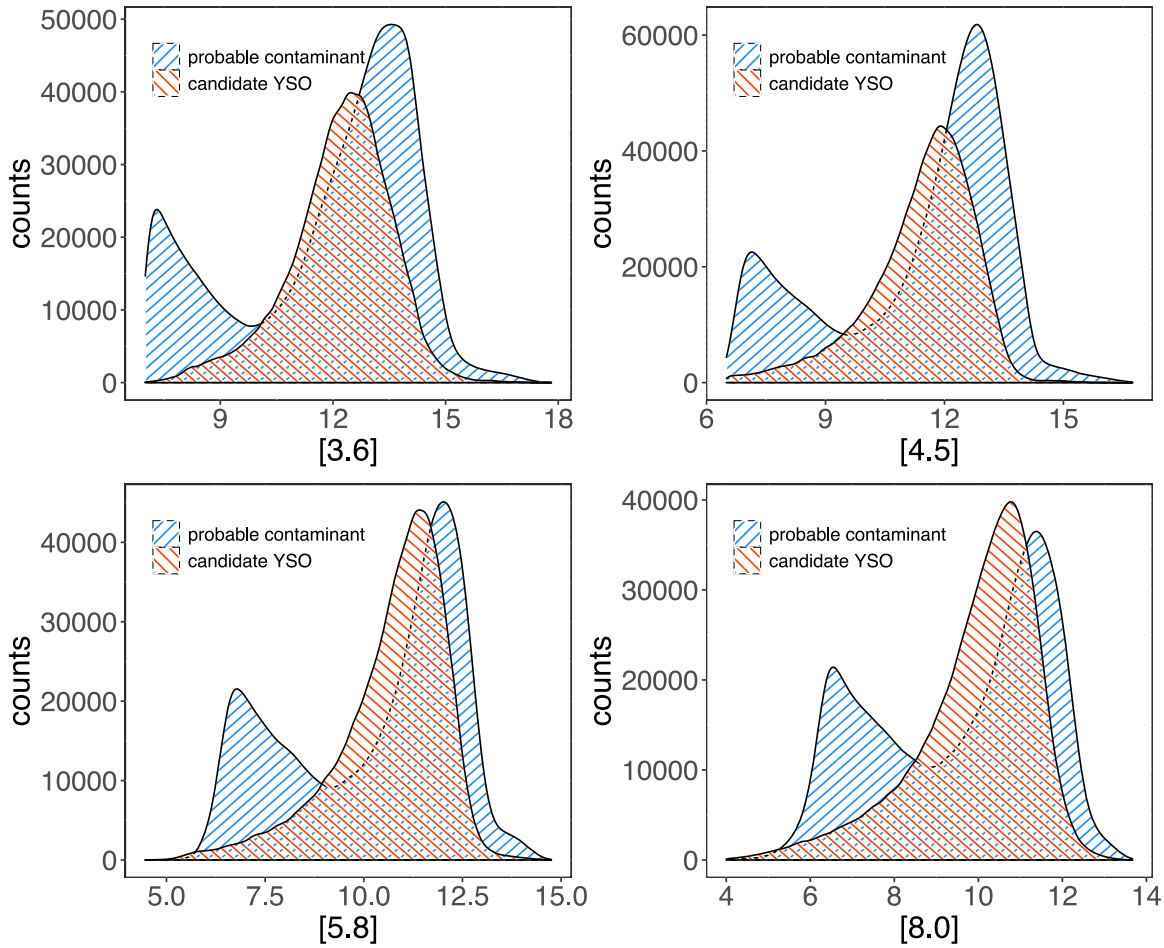

Figure 8. Distributions of IRAC magnitudes for the candidate YSOs (red stripes) and the probable contaminants (blue stripes). Distributions for probable contaminants are all multimodal. YSO candidates each have a single mode toward the fainter end of the distribution, and a heavy tail consisting of brighter sources.

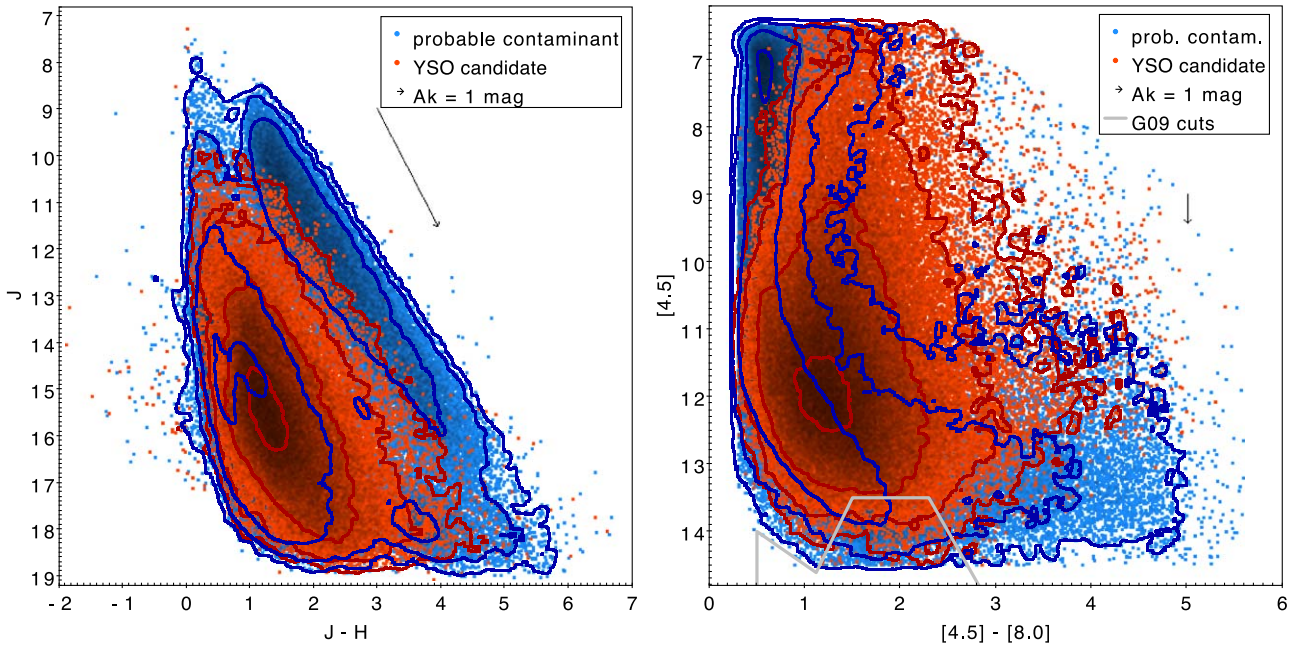

Figure 9. Infrared color–magnitude diagrams, J vs. $J - H$ (left) and [4.5] vs. [4.5]–[8.0] (right), with candidate YSOs (red) and probable contaminants (blue). In low-density parts of the scatter plot, individual points are drawn, but in areas with overlapping points, darker colors indicate higher density. We also include contours at evenly spaced logarithmic increases in density. Arrows indicate the approximate shift produced by extinction of $A_K = 1$ mag, assuming the Indebetouw et al. (2005) reddening law. Gray polygon demarcates the region used by Gutermuth et al. (2009) to select contaminants.

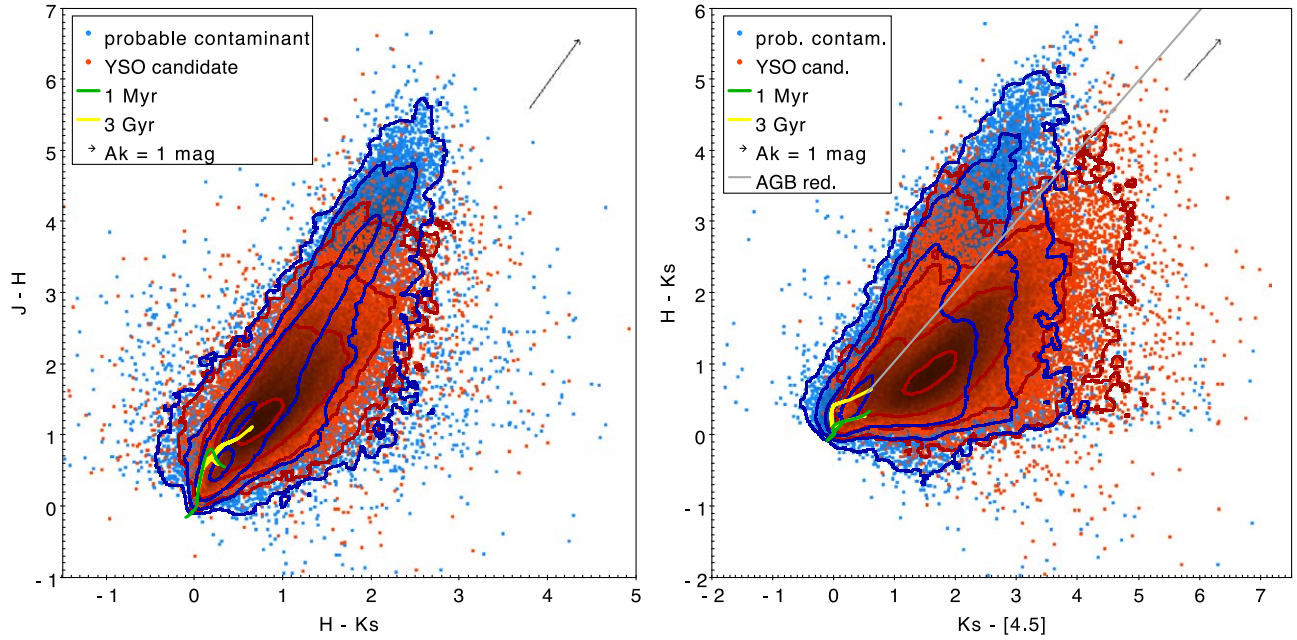

Figure 10. Color–color diagrams for $J-H$ vs. $H-K_s$ (left) and $H-K_s$ vs. K_s–[4.5] (right) with candidate YSOs (red) and probable contaminants (blue). Curves indicate 1 Myr (green) and 3 Gyr (yellow) isochrones (Bressan et al. 2012), with models for the AGB phase (Marigo et al. 2013) included in the 3 Gyr isochrone. In the right panel, a gray line parallel to the Indebetouw et al. (2005) reddening vector extends from the tip of the AGB. In our sample, objects classified as contaminants predominate above this line while objects classified as candidate YSOs are more abundant below.

K_s-band excess. Objects with very red J–$H > 5$ colors are largely classified as contaminants.

On the $H-K_s$ versus K_s–[4.5] diagram, we show both a 1 Myr isochrone for (pre–)main-sequence stars and a 1 Gyr isochrone that also includes post–main-sequence stars. The red-giant branch extends upward to stars with redder $H-K_s$ colors than the (pre–)main sequence, allowing these groups to be better separated. On this plot, the reddening vector points to the upper right. If we consider a line parallel to the reddening vector, starting from the tip of the asymptotic giant branch (as shown by the gray line in the figure), we would expect that many of the stars lying above this line could be evolved stellar contaminants. This is consistent with what the classifier finds; most objects above this line are classified as probable contaminants, while the candidate YSOs are more abundant below this line. The slope of the Indebetouw et al. (2005) reddening vector is not precisely parallel to the upper edge of the source distribution—this may arise due to systematic uncertainties in the reddening law, or it could be a property of IR colors of highly obscured evolved stars.

Figure 11 shows four projections of sources in IRAC color–color space. On the [3.6]–[4.5] versus [4.5]–[5.8] diagram, the YSO candidates are smoothly distributed, with a peak in density around [3.6]–[4.5] ~ 0.5 and [4.5]–[5.8] ~ 0.4, and a tail that extends up and to the right. In contrast, the contaminant distribution peaks slightly bluer in [3.6]–[4.5], and the distribution appears bifurcated, with some sources being redder in [3.6]–[4.5] while others are redder in [4.5]–[5.8]. This bifurcation may be related to the types of contaminants. For example, the contaminants to the upper left roughly correspond to a region of color space identified by Gutermuth et al. (2009) (and indicated by the gray boundary) as containing sources produced by knots of shocked emission from H_2, while the contaminants to the lower right were associated with knots of PAH emission. Both areas outlined by Gutermuth et al. are dominated by objects that we classify as probable contaminants, but the edges of the YSO distribution also overlap these boundaries.

On the [4.5]–[5.8] versus [5.8]–[8.0] diagram, the peak density of YSO candidates is redder in both colors than the peak density of probable contaminants. The objects with the most extreme [5.8]–[8.0] colors are nearly all classified as contaminants. These lie in a region of the diagram identified by Gutermuth et al. (2009) (gray boundary lines) as being dominated by unresolved star-forming galaxies. This diagram also includes a finger comprised of both YSO candidates and contaminants, extending to high [4.5]–[5.8] values ranging from $\sim$2 to $\sim$3.5, but with [5.8]–[8.0] colors in a restricted range ($1.5 \lesssim$ [5.8]–[8.0] $\lesssim 2.6$. Previously published YSO catalogs (e.g., Gutermuth et al. 2009; Rebull et al. 2011) have included a few YSOs in this region of color space; however, the high number of sources identified when examining the entire inner Galactic midplane makes this feature appear much more pronounced. These stars have colors similar to the PAH nebulosity found in star-forming regions (Povich et al. 2013); however, visual inspection of a sample of these sources suggests that the majority are bona fide point sources in all four IRAC bands.

In the bottom two panels of Figure 11, the peaks of the YSO candidate distributions are redder than the peaks of the probable contaminant distributions for each IRAC color. Nevertheless, while the objects with reddest [3.6]–[4.5] tend to be YSO candidates, the objects with most extreme red [4.5]–[8.0] or [5.8]–[8.0] colors are almost all classified as contaminants.

Figure 12 shows the $J-H$ colors, which are the most sensitive to extinction, versus the IRAC colors, which are the most sensitive to IR excess. In $J-H$, the peaks of the density distributions are slightly redder for the contaminants than for the YSO candidates, but in IRAC colors, the peaks are significantly redder for the YSO candidates than the contaminants. In both cases, the objects with most extreme red

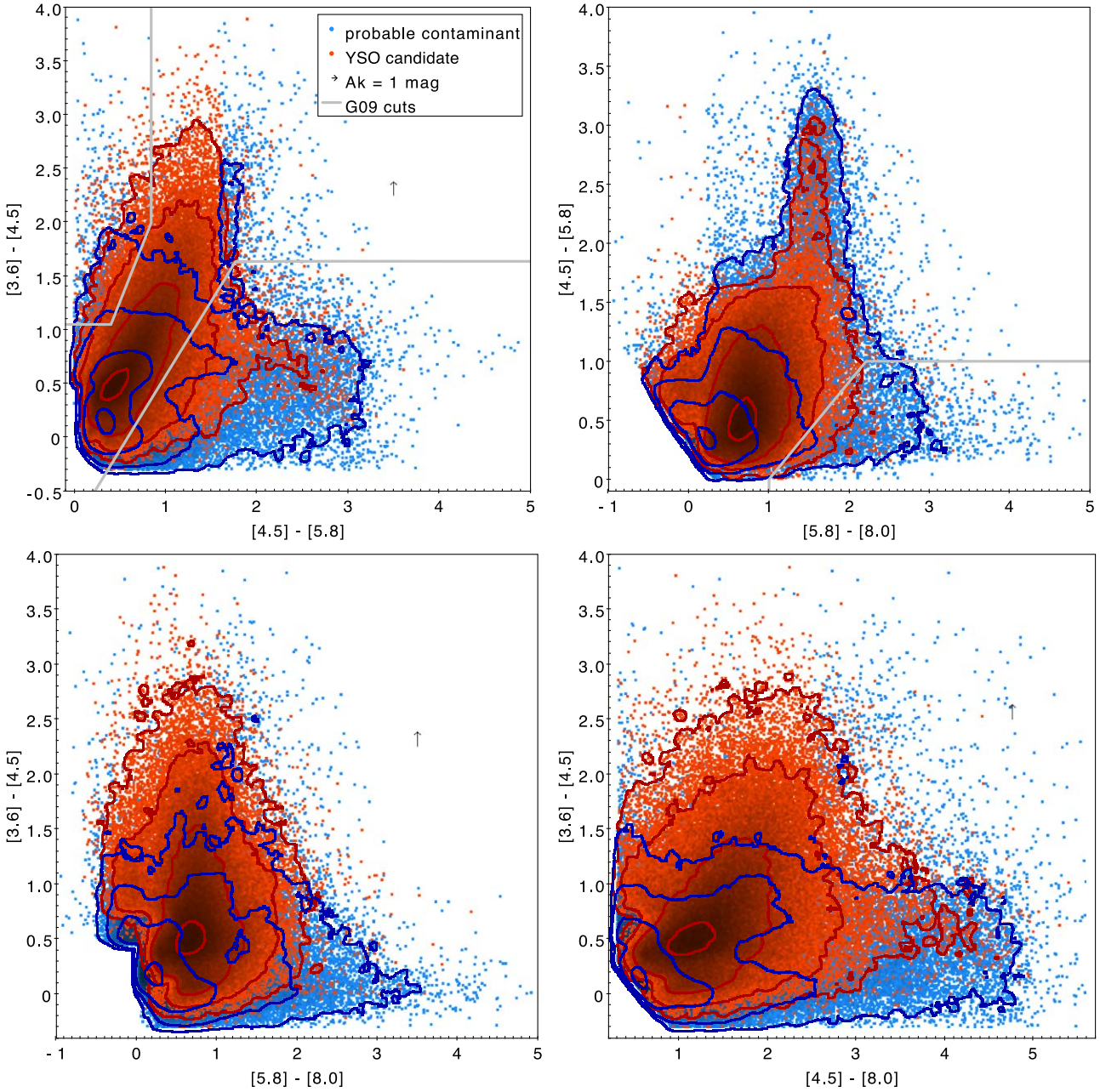

Figure 11. Color–color diagrams in the IRAC bands. YSO candidates (red points) and probable contaminants (blue points) partially overlap in each of these projections, but differences are visible in their distributions. The short length (or absence) of the $A_K = 1$ mag reddening vectors (black arrows) implies that extinction would need to be extreme to significantly change these distributions. PAH feature is distinctly visible on the [4.5]–[5.8] vs. [5.8]–[8.0] diagram. Gutermuth et al. (2009) criteria are indicated by gray lines, for comparison.

colors tend to be classified as contaminants. However, some of the reddest IRAC sources do not appear on these plots because they lack *J*-band magnitudes.

On all these color–color plots, the blue ends of the distributions are artificially truncated by the selection rules imposed to ensure that the IR excesses are real. Thus, our catalogs will not be sensitive to certain classes of YSOs, including some YSOs with anemic disks or pre–main-sequence stars without disks.

5.3. Optical Color–Magnitude Diagram

Less than half the YSO candidates are optically visible: for example, Gaia DR2 detects ∼36,000 of them, which comprise ∼30% of the entire sample. The candidates detected by Gaia tend not to be as red in the mid-IR as other candidates (e.g., [3.6]–[4.5] $\lesssim 1$ and [4.5]–[8.0] $\lesssim 1.2$).

Figure 13 shows a Gaia color–magnitude diagram for the visible YSO candidates. Absolute *G*-band magnitudes, computed using Gaia parallaxes ϖ, are plotted against Gaia $G - RP$ colors. Only sources with signal-to-noise $\varpi/\sigma_\varpi > 3$ are included, meaning that the sample of 7686 sources is small in comparison with the total number of YSO candidates. Nevertheless, this sample is useful for evaluating whether the optically bright candidates have properties consistent with pre–main-sequence stars.

On the Gaia color–magnitude diagram, we show isochrones for young stars at several ages, ranging from 1 to 50 Myr, from the Bressan et al. (2012) models. We also indicate the effects of

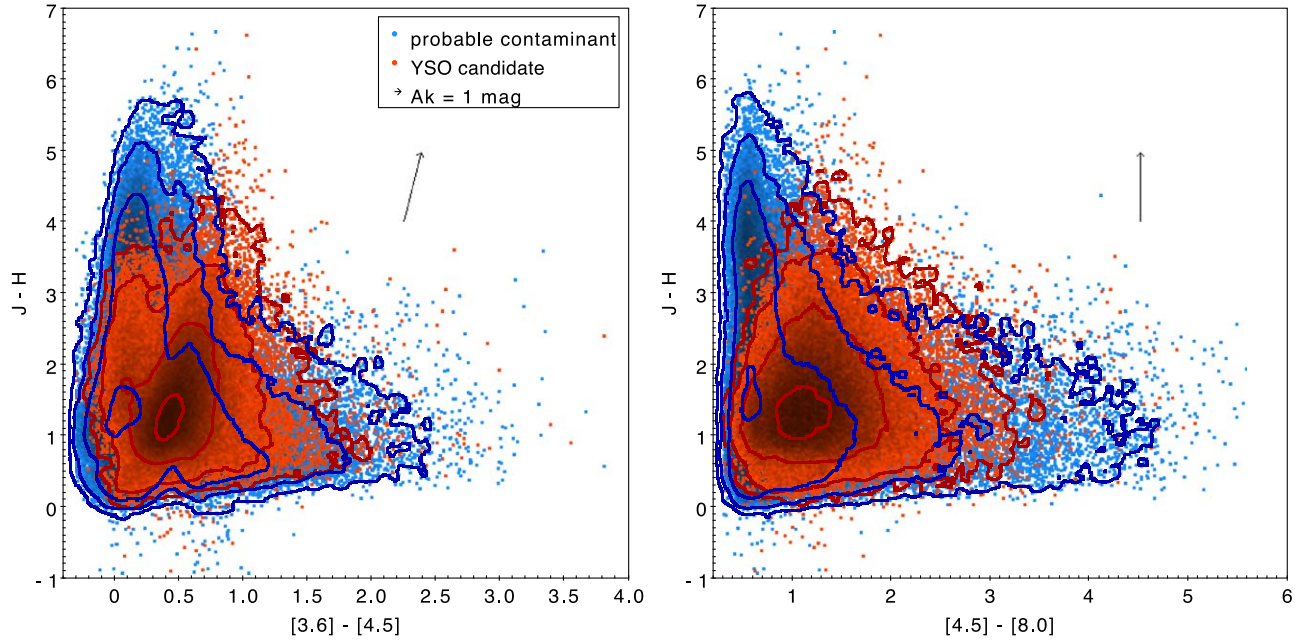

Figure 12. Color magnitude diagrams showing $J - H$ (the color most sensitive to reddening) vs. [3.6]–[4.5] (left) and [4.5]–[8.0] (right), which are both useful for selecting YSOs. Symbols and lines are the same as in Figure 9.

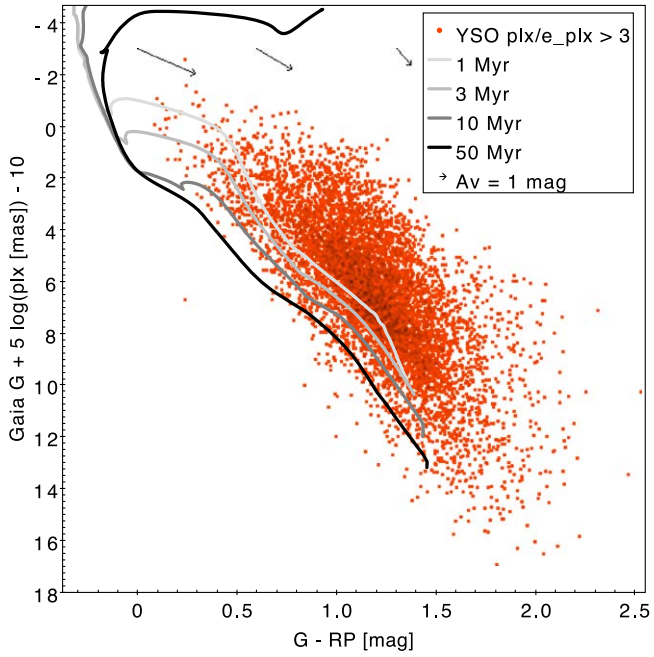

Figure 13. Absolute Gaia G-band magnitude vs. $G - RP$ color for candidate YSOs with $\varpi/\sigma_{\varpi} > 3$. Curves are unreddened isochrones with ages of 1, 3, 10, and 50 Myr from Bressan et al. (2012). Arrows indicate approximate Gaia reddening vectors using the Cardelli et al. (1989) and O'Donnell (1994) extinction curves with $R_V = 3.1$. The broad Gaia bands mean that these vectors vary with color, so we show three vectors estimated using stellar spectra with intrinsic colors of G-$RP = 0$, 0.6, and 1.3. Nearly all of these candidate YSOs are in the region of this color–magnitude diagram consistent with the pre–main sequence.

reddening, which would shift points down and to the right on this diagram. The wide Gaia bands mean that the effect of reddening depends on the spectrum of the object, so we show three approximate reddening vectors for three colors; more discussion of how this affects selection of pre–main-sequence stars can be found in Herczeg et al. (2019) and Kuhn et al. (2020). Nearly all the candidates lie above the 50 Myr isochrone, and the majority also lie above the 1 Myr isochrone, which is consistent with most of these candidates being very young pre–main-sequence stars.

5.4. 24 Micron Photometry

When photometry is available in the MIPS 24 μm band (or the W4 band at 22 μm), it can be useful for corroborating classifications based on IRAC. For example, the SED at $\sim$24 μm tends to be more steeply declining for AGB stars, where IR excess is produced in hot dusty winds, in contrast with YSOs' relatively cooler disks and envelopes.

Figure 14 shows the candidate YSOs and contaminants in $J - K_s$ versus [4.5]–[24] colors. These colors may be useful for distinguishing between AGB stars and YSOs, because the typical AGB star has a precipitous rise in the *JHK* bands followed by a drop in the mid-IR. The figure shows that the YSO candidates tend to be in the middle of the [4.5]–[24] distribution. The objects with [4.5]–[24] $\lesssim 2.4$ are almost all classified as probable contaminants; however, there is a red tail to the probable contaminant distribution, with a high percentage of objects redder than [4.5]–[24] $\gtrsim 7$ also being considered contaminants. An examination of the spatial distribution (not shown) of the probable contaminants in this red tail reveals that many of them are nonclustered, higher-latitude objects in the Galactic bulge. Of the contaminants with low [4.5]–[24], the distribution of $J - K_s$ ranges from $\sim$0 to $\sim$9, extending redward of the YSOs. Such red $J - K_s$ colors, combined with relatively blue [4.5]–[24] colors, would be consistent with our hypothesis that many of these probable contaminants are AGB stars.

5.5. SED Class

Spectral index in the infrared, defined as

$$\alpha = \frac{d \log(\lambda f_{\lambda})}{d \log \lambda}, \tag{6}$$

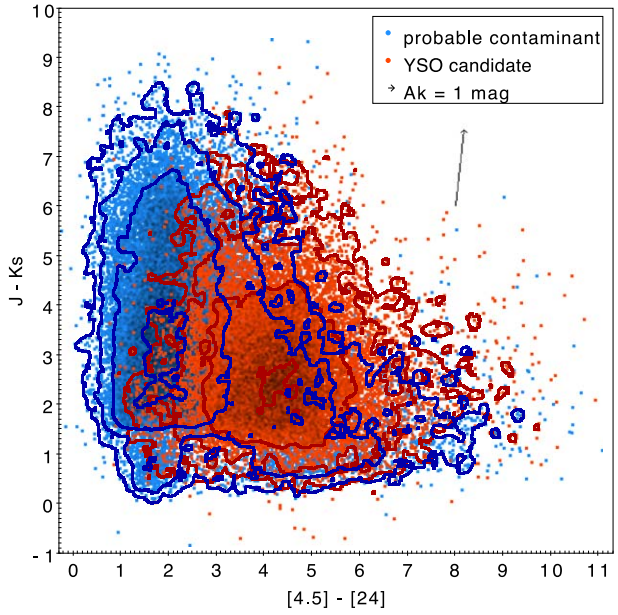

Figure 14. The $J - K_s$ vs. [4.5]–[24] color–color diagram for candidate YSOs and probable contaminants. This diagram may be useful for verifying separation between AGB stars and YSOs. AGB stars typically have steep red SED shapes in the near-IR, but turn over to a Rayleigh–Jeans tail around 24 μm. We find the sources that exhibit the reddest $J - H$ colors but not as red [4.5]–[24] colors are mostly classified as probable contaminants, consistent with being AGB stars.

is frequently used to assess the evolutionary stages of YSOs (e.g., Lada 1987; Andre & Montmerle 1994; Evans et al. 2009a; Rebull et al. 2014). However, the value of α depends on what spectral range is used, with the largest available range typically being favored by most studies. Furthermore, the calculation of spectral index may also be affected by reddening. To estimate α values that are minimally affected by reddening, we use the wavelength range from 4.5 to 24 μm, since interstellar extinction in these bands is smaller than at shorter wavelengths and the reddening curve is flatter (Indebetouw et al. 2005; McClure 2009; Xue et al. 2016). For these bands:

$$\alpha_{[4.5]-[24]} \approx 0.55([4.5] - [24]) - 2.94 \quad (7)$$

$$\alpha_{[4.5]-W4} \approx 0.58([4.5] - W4) - 2.92 \quad (8)$$

$$\alpha_{[4.5]-[8.0]} \approx 1.64([4.5] - [8]) - 2.82. \quad (9)$$

Where available, we prefer the α estimate based [4.5]–[24], followed by [4.5] − W4, and finally [4.5]–[8.0]. For YSOs suspected of having strong silicate absorption or PAH emission (Sections 5.6–5.7), we do not use the [4.5]–[8.0] color to estimate YSO class, because either feature could affect the 8.0 μm band.

Figure 15 shows the distribution of spectral indices calculated for candidate YSOs. Based on these estimates, there are 15,943 Class I ($\alpha > 0.3$), 23,810 flat spectrum ($0.3 \leqslant \alpha < -0.3$), 59,949 Class II ($-0.3 \leqslant \alpha < -1.6$), and 5352 Class III ($\alpha \leqslant 1.6$) YSOs, using the α boundaries from Greene et al. (1994). In addition, there are 12,392 candidate YSOs with uncertain classes due to missing photometry. This classification scheme roughly reflects the YSO evolutionary sequence from deeply embedded sources with massive envelopes (Class I and flat spectrum) to stars with disks (Class II) and systems where the disk has mostly dispersed (Class III). However, viewing geometry may also affect the

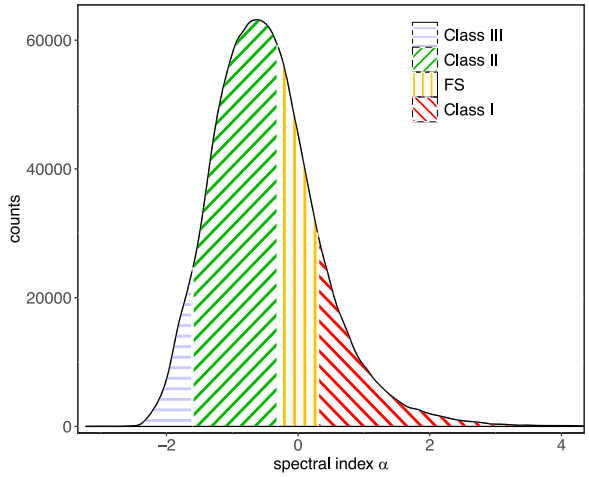

Figure 15. Distribution of spectral index α for YSO candidates, subdivided into YSO class using the customary demarcations at $\alpha = -1.6$, -0.3, and 0.3. The shape of the distribution will be the product of the prevalence of the YSO classes, with Class II/III YSOs being more common than Class I/flat SED YSOs, due to the longer lifetimes of the later evolutionary stages (e.g., Evans et al. 2009b), as well as our sensitivity to each class, which may be lower for YSOs with smaller IR excesses (e.g., Class III) and for deeply embedded YSOs (e.g., Class I).

assigned YSO class: for example, a YSO that would otherwise be considered Class II may have a Class I SED if viewed at high inclination (Williams & Cieza 2011). Finally, we clarify that, even though some classification schemes regard Class III sources as having no IR excess (see Evans et al. 2009a), in our scheme Class III implies weak but detectable excess.

5.6. Possible Silicate Absorption

Broad silicate dust absorption or emission features, centered at ∼9.7 and ∼18 μm, are frequently detected in the mid-IR spectra of YSOs (e.g., Furlan et al. 2006, 2008, 2011; Oliveira et al. 2010). The 9.7 μm feature overlaps the IRAC 8 μm band, so these features can affect YSO colors observed by IRAC.

In the color–color diagram shown in Figure 16 (left panel), a group of ∼2000 YSO candidates stand out due to their unusually blue [5.8]–[8.0] < 0 colors—these objects are flagged in Table 1. Given the lack of a red color in [5.8]–[8.0], the classification of these stars as YSO candidates was based mainly on their [3.6]–[4.5] $\gtrsim$ 0.5 and [4.5]–[5.8] $\gtrsim$ 0.5 colors, both of which tend to be redder than most of the other YSO candidates.

Figure 17 shows three example SEDs that we have fit with YSO models from Robitaille (2017)—for each source, the 10 best-fitting convolved models are indicated by the gray lines. Robitaille (2017) include multiple configurations of disks and/or envelopes, so we used the simplest model forms capable of explaining the data: a star and disk model (sp-s-i[11]) for SPICY 75228; a star, disk, and envelope models with variable inner radius (spu-hmi) for SPICY 85135; and a star and disk model with variable inner radius (sp-h-i) for SPICY 99415. Although these fits are not all formally good, given the reported photometric uncertainties, they illustrate the range of SED morphologies that could produced the colors that we observe. Each case requires a strong silicate absorption feature at 9.7 μm to reproduce the lower 8.0 μm band emission. The best models

[11] The designations correspond to models from Robitaille (2017).

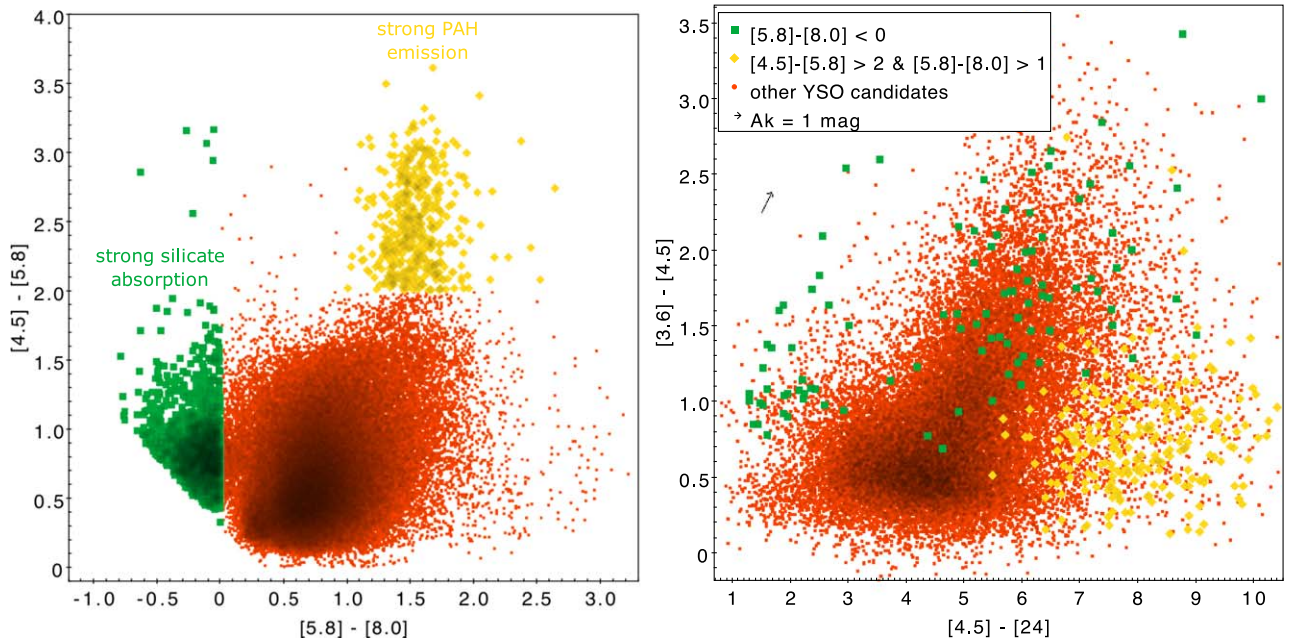

Figure 16. Left: IRAC color–color diagram with YSO candidates with possible strong silicate absorption and PAH emission labeled. Right: YSO candidates on the [3.6]–[4.5] vs. [4.5]–[24] diagram, with several subcategories selected. YSO candidates with low values of [5.8]–[8.0] colors are shown as green squares, and YSO candidates with suspected PAH emission are yellow–orange diamonds. Both groups have redder than average [4.5]–[24] colors, consistent with these classes of sources being deeply embedded.

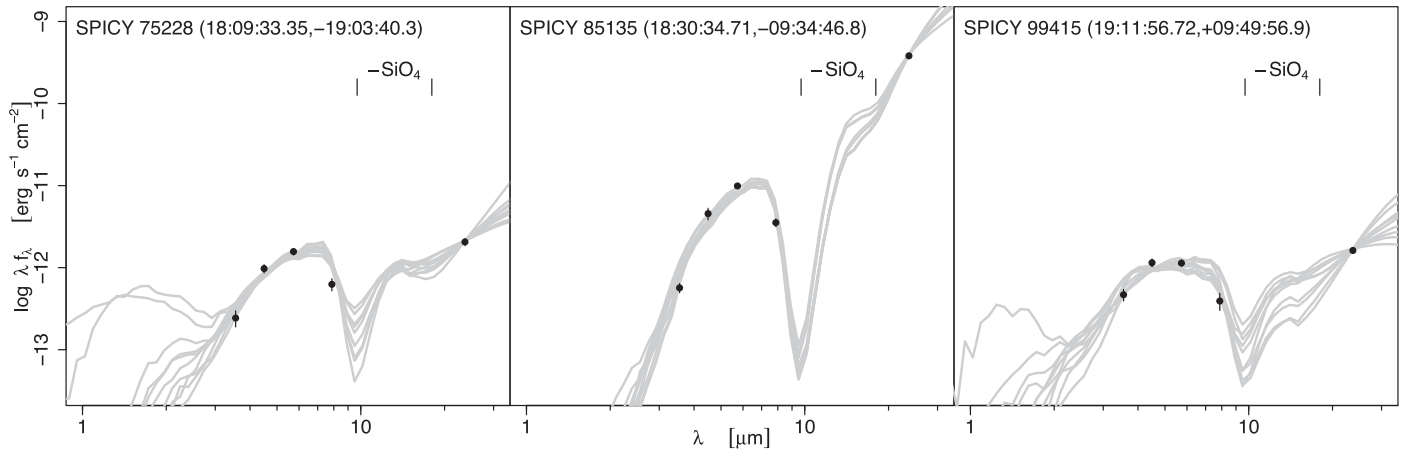

Figure 17. SEDs for three YSO candidates with [5.8]–[8.0] < 0. Black points represent the photometry in the IRAC bands, with 1σ error bars. Gray lines are the 10 best-fitting convolved YSO models from Robitaille (2017). Although not all of these models provide formally good fits, they represent the SED shapes necessary to produce the observed IRAC colors. These models demonstrate that strong silicate absorption features (centers indicated by tick marks) are necessary to reproduce observed photometry.

also tend to also have nearly edge-on inclination to provide the high absorbing column density.

Silicate absorption in YSO SEDs can come from the object itself or from foreground interstellar dust (van Breemen et al. 2011). We would expect a YSO with strong intrinsic silicate absorption to have a substantial disk or envelope that can produce the extinction, and Forbrich et al. (2010) find that YSOs with positive spectral indices are more likely to have strong silicate absorption. Figure 16 (right panel) shows our strong silicate absorption candidates on a plot of [3.6]–[4.5] versus [4.5]–[24]. The color [4.5]–[24] is a good indicator for SED spectral index that is not affected by silicate absorption. Most of the objects with possible silicate absorption have [4.5]–[24] > 5, higher than average for the YSOs, but ∼20 sources have [4.5]–[24] colors bluer than this. In the interstellar medium, the relation between the optical depth of the 9.7 μm feature and optical extinction is approximately $\tau_{9.7} \sim A_V/20$ (Roche & Aitken 1984; Chiar et al. 2007; Shao et al. 2018). Although most stars in our sample would not have sufficiently high foreground extinction for the feature to become optically thick, this can be achieved along lines of sight that pass through dense molecular clouds or near the Galactic Center.

5.7. Possible PAH Emission

Another salient feature in the [4.5]–[5.8] versus [5.8]–[8.0] diagram is the finger-like structure at [5.8]–[8.0] ≈ 1.6. Most of the sources with these colors in the full IRAC catalogs were classified as probable contaminants, but a minority (∼490 objects) were classified as YSO candidates. These colors match those expected for sources dominated by PAH emission bands. For an astronomical PAH emission spectrum, the ratio of flux

in the 5.8 μm band to the 8.0 μm band ranges from 0.31 to 0.41 (Draine & Li 2007, and references therein), corresponding to [5.8]–[8.0] = 1.6–1.9. There is little PAH emission in the 4.5 μm band, leading to a red [4.5]–[5.8] color. Candidates are flagged in Table 1 for strong PAH emission if they meet the criteria [4.5]–[5.8] > 2 and [5.8]–[8.0] > 1, which are based on the observed morphology of this feature in color space.

Although IR nebulosity in star-forming regions is dominated by PAH emission, inspection of the flagged YSO candidates suggests that most are valid point sources in all four IRAC bands, not spurious detection of nebular knots. For example, ~90% of these sources have $M = 2$ detections in both the 5.8 and 8.0 μm bands, indicating reliable detections. We examined the images of a subset of these objects by eye and found that, even in cases with surrounding nebulosity, the sources themselves appeared to match the PSF. PAH emission may be intrinsic to massive YSOs with sufficiently high ultraviolet luminosities (e.g., Whitney et al. 2013). Spitzer IRS spectroscopy of massive YSOs has shown PAH emission to be nearly ubiquitous and correlated with YSO luminosity (Oliveira et al. 2013).

In the MYStIX IR-excess catalog that we used for training, Povich et al. (2013) aggressively filtered sources with PAH emission in order to avoid contamination by nebular PAH knots. To be included, they required sources to exhibit red $K_s - [4.5]$ colors (avoiding bands with PAH emission), as would be expected for a massive YSO. This requirement will be reflected in our classifications via the random forest classifier. In the SPICY catalog, YSO candidates with possible PAH emission have median $K_s - [4.5] = 3.3$, as compared with a median for the entire sample of 2.0.

Figure 16 (right panel) shows the 24 μm emission for these objects. The YSO candidates with possible PAH emission have [4.5]–[24] colors ranging from 5 to 10, much higher than the average for YSOs. This result is consistent with these objects being massive YSOs.

6. Environment

Many dynamical processes sculpt the interstellar medium in star-forming regions and affect the spatial relationship between the clouds and young stars (Shu et al. 1987; McKee & Ostriker 2007). Infrared nebulosity, however, can be considered a strong proxy for star formation, as newly formed massive stars illuminate the primordial clouds in the star-forming complex. The Spitzer images reveal features ranging from IR dark clouds to bright PAH-dominated nebulosity, which can trace the photodissociation regions at the edges of clouds and bubbles (e.g., Churchwell et al. 2009; Pari & Hora 2020).

To facilitate the study of the local environments around YSOs, we have created an album of $3' \times 3'$ image cutouts[12] in all four Spitzer bands, with additional false-color image files ready for visual inspection or generic image processing frameworks (e.g., Yang et al. 2012). The false-color images (see examples in Figure 18) were created with a heuristic based on Lupton et al. (2004), mapping the IRAC 3.6 μm to the blue channel, 5.8 μm to green, and 8.0 μm to red. Here, we applied a hyperbolic arcsin transform to each IRAC band, and we selected the range of the color intervals from the mode of the distribution of the lower 2×10^{-2}% of the pixels for the minimum value, and the mode of the distribution of the upper 6×10^{-5}% pixels for the maximum. These values were chosen to optimize the visual experience while minimizing information loss and excluding extreme outlier pixels. The modes were estimated using the Venter (1967) estimator, as implemented by the MODEEST package (Poncet 2019).

The SPICY album comprises a total of 117,224 PNG stamps. A total of 222 YSOs candidates from the SPICY catalog miss their stamps, due to numerical problems in the original FITS files and/or a lack of response from the IPAC archive in one or more bands at the time of the album creation. All PNG and FITS files are archived long-term at Zenodo,[13] hosted at CERN facilities.

6.1. A Simple Characterization

Below, we demonstrate an example application for these cutouts, using a simple unsupervised image clustering strategy to characterize environments in which the YSOs candidates are found.

We avoid clustering in the pixel space because it is not invariant to image translations and rotations, which are properties that any proper content-based image clustering solution should have. Two candidate transforms that can introduce these properties via the power spectrum are wavelets (as used for a similar application in Krone-Martins et al. (2019)) and Fourier transforms (e.g., Kauppinen et al. 1995; van der Schaaf & van Hateren 1996); here, we adopt the latter. This is partially motivated because the Fourier power spectra is linked to the turbulent properties of the star formation medium (e.g., Elmegreen & Scalo 2004), revealing signatures of different physical phenomena.

We first compute the 2D Fourier power spectra of each cutout in each IRAC band. We then compute 1D radially medianized power spectra from each of the original 2D power spectra and concatenate these 1D power spectra to form a vector for each YSO candidate. Next, we organize the vectors of all environments into a single matrix and perform principal component analysis (PCA; Pearson 1901; Hotelling 1933), from which we select the most relevant dimensions (see also Ishida & de Souza (2013) and de Souza et al. (2014) for PCA variants), which acts as feature compression (see, e.g., Sasdelli et al. 2016). Finally, we model the distribution using a multivariate Gaussian mixture model (GMM; Pearson 1894; Scrucca et al. 2016; de Souza et al. 2017; Melchior & Goulding 2018) in the space defined by the first two principal components of the power spectra and the modes of the pixel values in each cutout, which we transform using an inverse sinh function. The distribution in this space is complex and requires many (25) Gaussian components, with model selection using the Bayesian information criterion (Schwarz 1978).

Visual inspection shows that the GMM components tend to correspond to three types of environments: those that are nebulosity-free (or have minimal nebulosity), mixed environments, and cloud-like environments. These are labeled Environments 1, 2, and 3 in Table 1. We also found outliers on the boundary of the distribution. Examination of the cutouts showed that the outliers

[12] To produce the album, we constructed an infrastructure to query the IPAC archive at http://irsa.ipac.caltech.edu, which tracks the FITS transfers and also tracks and verifies the local generation of the PNG stamps. This infrastructure makes use of a PostgreSQL database (PostgreSQL Global Development Group 2020) and is parallelized. However, we kept the number of parallel data transfers from IPAC low in order to avoid overloading their servers, enabling the extraction of all IRAC images and the construction of all the stamps in about three days.

[13] The SPICY album (251 GB uncompressed) is hosted by Zenodo (doi: 10.5281/zenodo.4462819) as a compressed tarball (187 GB).

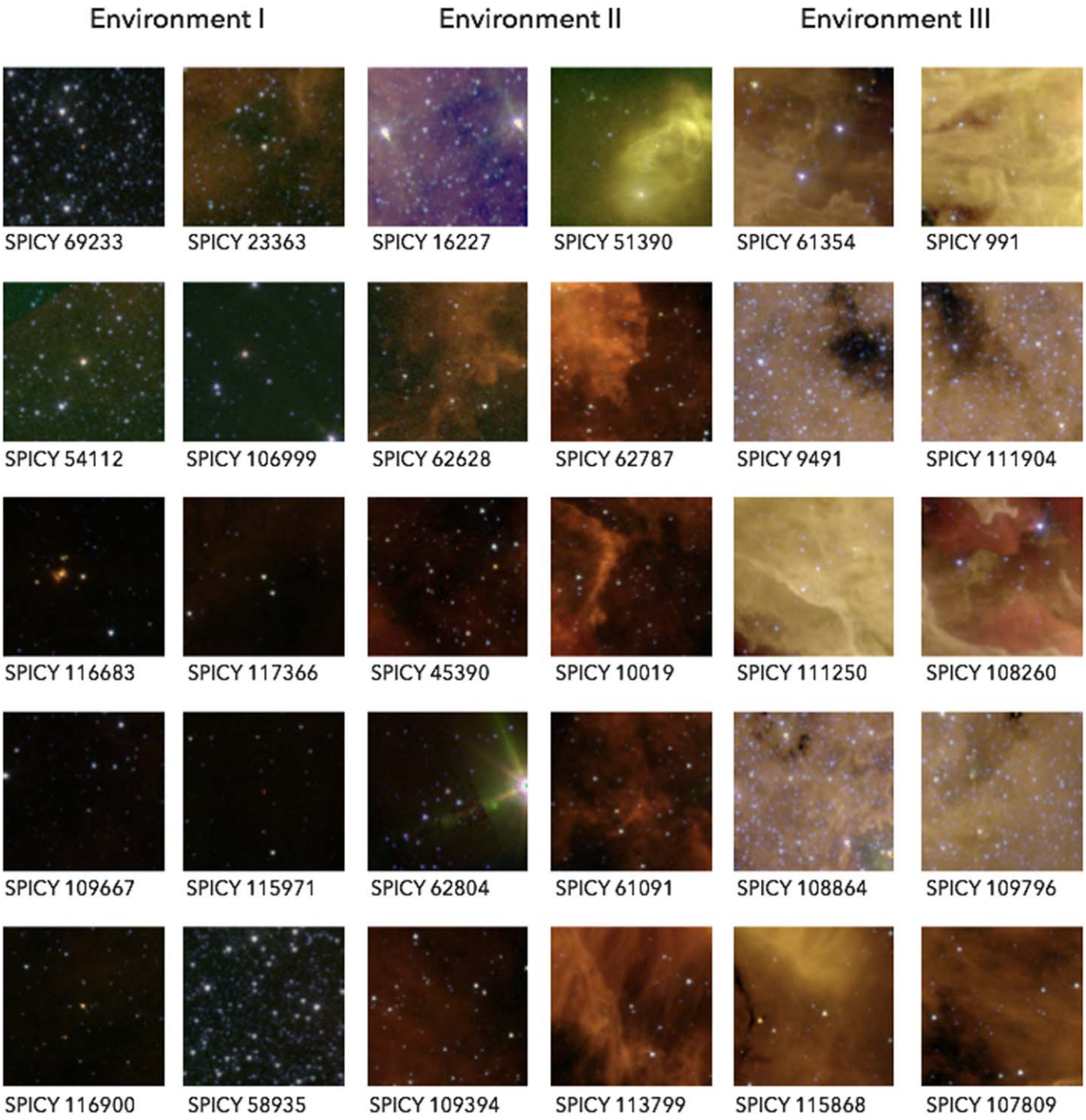

Figure 18. Color $3' \times 3'$ cutouts (IRAC 3.6 μm in blue; 5.8 μm in green; 8.0 μm in red) centered around a sample of YSO candidates from the SPICY Album. First two columns: examples of stamps from the Environment 1 class, corresponding to images with no or minimal nebulosity. Middle two columns: examples of stamps from the Environment 2 class, which is a mixed class mostly containing regions at the transition between Environments 1 and 3 or misclassifications from those two extreme classes. Last two columns: examples of regions classified as Environment 3 that clearly correspond to cloud-like environments.

correspond to severe image reconstruction errors and/or missing data in one or more bands and are located at the edges of the surveys.

A total of 66,539 stamps, or ∼57% of the valid stamps, were classified as cloud-like, while 32,790 stamps, or ∼28%, were classified as nebulosity-free. The mixed class and the outliers correspond to 15,462 and 2433 stamps, or ∼13% and ∼2%, respectively. These numbers indicate that most YSO candidates are indeed in cloud-dominated environments, as would be expected for YSOs in star-forming regions. However, the number of candidates in environments presenting diffuse nebulosity or even no detectable nebulosity is not negligible. There is a slight but statistically significant[14] correlation between spectral slope α and environment class, with stars in

[14] The Kolmogorov–Smirnov test yields a p-value <0.01 for the null hypothesis that the α values are drawn from the same distribution for cloud-dominated versus nebulosity-free regions.

nebulosity-free environments having preferentially more negative spectral indices (indicating later evolutionary stages).

The cloud-like environments are most prevalent in the inner regions of the Galaxy, between approximately $\ell = 300°$ and $50°$ and $|b| \leqslant 1°$. Cloud-like environments are also associated with large star-forming complexes outside this coordinate range, including some of the star-forming regions in Cygnus X.

The mixed environment is most prevalent further from the Galactic center (e.g., $\ell \leqslant 310°$ or $\ell \geqslant 30°$). Some stellar associations include both cloud-like and mixed classes (e.g., the Carina Nebula).

The image cutouts with no or minimal nebulosity are found throughout the entirety of the survey, except within $\sim 1°$ of the Galactic Center. Stars in this environment are the most evenly distributed—but even among these stars, several clusters can be seen. For example, the Sco OB1 Association is made up of both cloud-like and nebulosity-free classes.

This simple application can certainly be significantly improved by adopting tailored methodologies and signal representations, for instance, curvelet transforms (e.g. Candès & Donoho 2000; Starck et al. 2003), to characterize the signal power contained in filamentary structures, and customized clustering methods. Moreover, a proper physical characterization of the YSO environment requires consideration of effects of, for example, the distances to the objects, accounting for the distinct physical scales probed, differences of PSFs in different IRAC bands and their impacts on the power spectra, projection effects of the nebulous matter in the plane of the sky, etc. However, as we show here, even a simple analysis has already revealed that, curiously, a significant fraction of the YSO candidates in the SPICY catalog do not seem to be lying in environments dominated by clouds.

7. Spatial Clustering

Spatial aggregation is a well-recognized property of YSOs (e.g., van den Bergh 1964; Carpenter 2000; Allen et al. 2007) that can be observed in the distribution of our YSO candidates (Figure 5). However, the best clustering algorithm to use for stars, or even what is the most meaningful definition of star cluster, is not obvious (see Ascenso 2018; Gouliermis 2018; Kuhn & Feigelson 2019). For example, different groups may have vastly different numbers and densities of stars, and the spatial distributions are complex and often fractal-like. Thus, different cluster analysis methods that yield different segmentations may be appropriate for different scientific applications (e.g., Everitt et al. 2001).

We choose the algorithm "Hierarchical Density-Based Spatial Clustering of Applications with Noise" (HDBSCAN; Campello et al. 2013), which has been successfully applied to Gaia DR2 data to detect hundreds of new open clusters (e.g., Kounkel & Covey 2019; Castro-Ginard et al. 2020). The HDBSCAN algorithm (Campello et al. 2013) allows for groups of stars with different numbers, densities, and morphologies, it permits stars to not belong to any group, and it can be applied across the entire survey area in a uniform way, providing a reasonable-looking clustering solution. We apply this algorithm to each of the contiguous survey regions using a Python implementation.[15] The main parameter we choose is the minimum number of stars in a group, which we set to $n = 30$, and we run the algorithm using the "excess of mass" method for cutting the tree. The algorithm is run on the Galactic ℓ and b coordinates for each contiguous segment of the IRAC survey area. The resulting groups are not necessarily gravitationally bound systems, but rather collections of YSO candidates that appear to be spatially aggregated. The two groups, labeled in Figure 6, were selected using this algorithm. We provide a list of these groups, along with their properties, in Table 2.

This method found 406 stellar groups, collectively including 58,084 (=49%) of the YSO candidates. This suggests that roughly half the YSO candidates are spatially clustered, while the other half are more widely distributed. The choice of n does affect the solution, particularly whether groups are subdivided into smaller groups are unified into larger groups. However, the percentage of stars in groups stays relatively constant (i.e. 47%–56%) when n is varied from 15 to 60. The median angular diameters of the groups increase from $\sim 0\overset{\circ}{.}2$ to $\sim 0\overset{\circ}{.}7$ over this range of n, with a value of $\sim 0\overset{\circ}{.}4$ at $n = 30$. We pick $n = 30$ because the resulting solution appears to avoid chaining together unrelated stars over large areas of the sky, but the groups are large enough to include enough Gaia sources for their astrometric properties to be estimated. An example of a complicated structure identified by us as a single group is the Carina Nebula complex, an association made up of multiple star clusters. We also note that HDBSCAN collects the overdensity of candidate YSOs toward the Galactic center into a single group labeled "G0.2-0.1." These stars do not all come from the same star-forming regions, but this region of the Galaxy is challenging for our algorithms.

7.1. Galactic XY Distribution

To estimate the heliocentric distances ($d_\odot$) to each of the YSO groups, we employ a hierarchical Bayesian model (Hilbe et al. 2017) to account for the measurement errors in parallaxes as well as the presence of outliers. The use of robust statistics is particularly suitable given the non-negligible presence of unknown contaminants in each group. Normality assumptions are sensitive to noise and outliers, which may result in a biased estimate of the mean distance. Replacing a Gaussian likelihood with a t-distribution is a relatively easy fix. The t-distribution has an extra ν parameter called "degrees of freedom," which controls how closely the distribution resembles the normal distribution. Larger values $\nu > 30$ essentially recover the normal distribution, while smaller values result in a distribution with heavier tails. This extra flexibility enables it to adapt to the extra noise in the data without introducing a bias in the underlying relationship.

The model formulation for the robust estimate is given below, where we define a t-likelihood for the observed ϖ and suitably vague priors on all the model parameters: uniform for $d_\odot$ over 25 kpc, and a gamma (Γ) prior (to ensure positivity) for ν.

$$\begin{aligned} \varpi_i &\sim \mathcal{T}(1/d_\odot, \sigma^2_{\varpi_i}, \nu), \\ \nu &\sim \Gamma(2, 0.1), \\ d_\odot &\sim \text{Uniform}(0, 25), \\ i &= 1 \ldots n_{\text{Gaia}}. \end{aligned} \tag{10}$$

The index i runs over the members of each group n_{Gaia} with Gaia astrometric information. Although distance is constrained to be positive, our likelihood model permits the parallax measurements for individual stars, ϖ_i, to be either positive or

[15] https://hdbscan.readthedocs.io/en/latest/api.html

Table 2
YSO Groups from HDBSCAN

Column	Column ID	Description
1	group	Group designation
2	l0	Central Galactic longitude ℓ_0 [deg]
3	b0	Central Galactic latitude b_0 [deg]
4	plx	Mean parallax [mas]
5	e_plx	Error on mean parallax [mas]
6	pml	Mean proper motion in ℓ [mas yr^{-1}]
7	e_pml	Error on mean proper motion in ℓ [mas yr^{-1}]
8	pmb	Mean proper motion in b [mas yr^{-1}]
9	e_pmb	Error on mean proper motion in b [mas yr^{-1}]
10	n	Total number of constituents
11	nG	Number of constituents with five-parameter Gaia astrometric solutions
12	flag	Flag for potential model problems

Note. Properties of YSO groups identified from the HDBSCAN algorithm. Median astrometric properties, including group parallax and proper motion, are inferred from the hierarchical Bayesian modeling of the Gaia DR2 astrometry. The group parallaxes and proper motions in this table are in the Gaia DR2 system, with no correction for zero-point offsets. We report formal (MAD) uncertainties from our model added in quadrature to the ± 0.04 mas and ± 0.07 mas yr^{-1} spatially correlated systematic errors on DR2 zero points (Lindegren et al. 2018). Groups are flagged if potential problems could affect interpretation of the Bayesian model as described in Section 7.1.
(This table is available in its entirety in FITS format.)

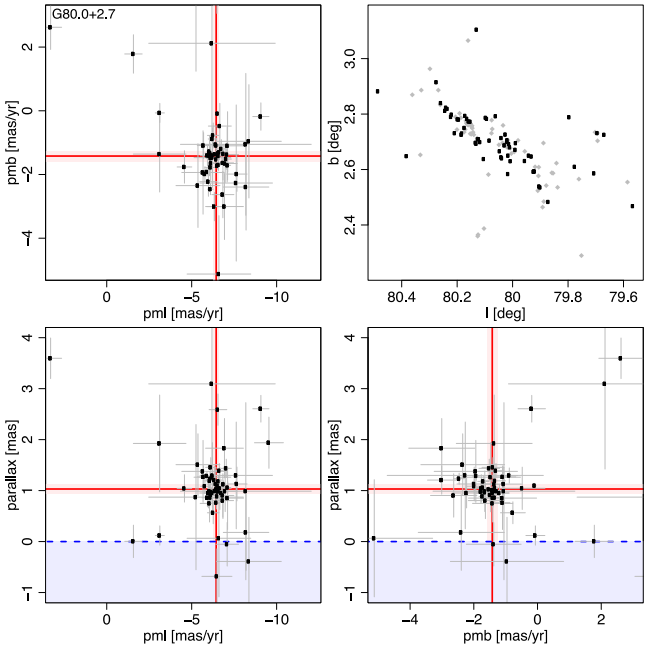

Figure 19. Scatter plots of astrometric properties for YSO groups; G80.0 + 2.7 (a group in the Cygnus X field) is shown as an example. These plots include proper motion vs. proper motion, parallax vs. proper motion, and (ℓ, b) positions. The estimated means from the hierarchical Bayesian model are shown by the red lines, and 2σ formal uncertainties are illustrated by the pink shaded areas. Values excluded by our prior are shaded blue. Stars with Gaia DR2 five-parameter astrometry are black circles with gray 1σ error bars, and stars with only position information are gray diamonds.

(The complete figure set (406 images) is available.)

negative. We evaluate the model using a Gibbs sampler, for which we use the JAGS[16] package (Plummer 2017) within the R language. We initiate three Markov Chains by starting the Gibbs sampler at different initial values. Initial burn-in phases were set to 5000 steps, followed by 5000 integration steps for each YSO group, which are sufficient to guarantee the convergence of each chain.

Table 2 provides group parallaxes and proper motions estimated from the posterior medians. Uncertainties are estimated from the mean absolute deviation (MAD) of the posterior (scaled to approximate 1σ uncertainties) and added in quadrature to the ± 0.04 mas and ± 0.07 mas yr^{-1} spatially correlated systematic errors on DR2 zero points (Lindegren et al. 2018). Out of 406 groups, 402 have some Gaia astrometry, giving at least a rough estimate of parallax and proper motion. Of these, most groups include at least $n_{\rm Gaia} = 10$ members having Gaia five-parameter astrometric solutions, enabling estimates that are more precise than those based on individual stars.

For each group, we show scatter plots of stellar proper motions, parallaxes, and positions (Figure 19), with the groups' mean parallaxes and proper motions indicated. In most cases, the stars form a single clump in $\mu_{\ell^*} - \mu_b - \varpi$ space, suggesting that most of the group members are spatially and kinematically associated. In other cases (e.g., G77.8+1.0), multiple clumps are apparent, which may imply that distinct stellar groups with chance alignment have been merged by the HDBSCAN algorithm. In the example G77.8+1.0, the estimated properties correspond to the more distant but more numerous of the two groups. We visually inspected all groups in Figure 19 and have flagged those in Table 2 for which problems—such as the suggestion of multimodality, groups that appear dominated by field stars, or a single data point with too much leverage—could affect the interpretation of the Bayesian models. We also flag groups with $n_{\rm Gaia} < 3$.

The locations of these stellar groups in heliocentric Galactic XY coordinates are plotted in Figure 20. For this plot, we converted parallaxes to distance using an average -0.0523 mas Gaia DR2 zero-point correction estimated by Leung & Bovy (2019). Thus, $X = d_\odot \cos(b)\cos(\ell)$, and $Y = d_\odot \cos(b)\sin(\ell)$. The estimated centers of spiral arms from Reid et al. (2019) are also indicated. Fainter red points are groups with $\varpi/\sigma_\varpi < 2$ or groups that have been flagged. Few YSO groups are detected within 1 kpc, but the footprints of the GLIMPSE (and GLIMPSE extensions) surveys exclude many of the nearest star-forming regions, which are located more than several degrees above or below the Galactic midplane. The bulk of the YSO candidates for which we have accurate measurements have heliocentric distances that range from 1 to 3 kpc. There may be some bias in the distances to which we are sensitive, because this range resembles the range in the distances of objects in our training set. Nevertheless, there are groups that appear to lie beyond ~3 kpc, but Gaia-based distances become more uncertain at this range.

The YSO groups are not distributed smoothly within the Galaxy, but instead reveal Galactic structure. The survey areas intersect several spiral arms, and provide crucial information about Galactic structure at the boundary between Quadrants I and IV, where structures traced by $v_{\rm lsr}$ measurements of gas become degenerate. We discuss the relation of the stellar groups to the spiral arms below.

Local (Orion) Arm: In Quadrants I and II, the Local Arm intersects both the SMOG field and the Cygnus X field. In

[16] http://cran.r-project.org/package=rjags

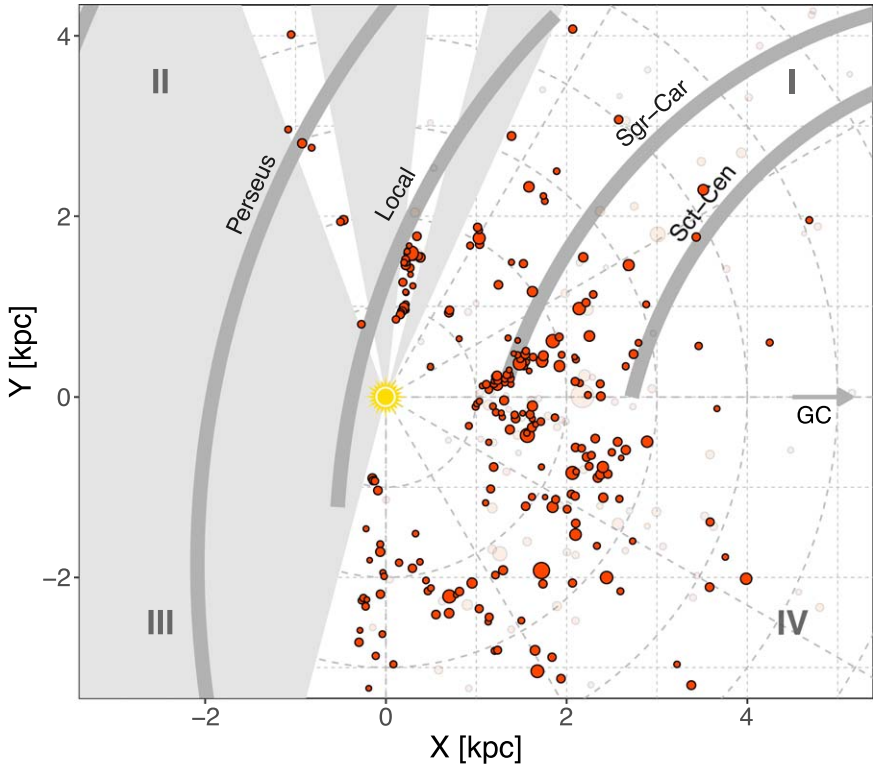

Figure 20. Spatial distribution of YSO groups in heliocentric Galactic *XY* coordinates. Groups with more reliable distances are depicted by the darker red circles, while those with more uncertain distances ($\varpi/\sigma_\varpi < 2$ or flagged) are lighter red. Circle sizes are proportional to the total numbers of members. The approximate centers of spiral arms from Reid et al. (2019) are indicated by the gray curves. The Sun's location is indicated by the yellow symbol. Wedges of the *XY* plane not covered by the catalog are shaded gray. Galactic quadrants and the direction of the Galactic Center are labeled. For conversion of Gaia DR2 parallaxes to distance in this figure, we use the −0.0523 mas zero-point offset estimated by Leung & Bovy (2019).

SMOG, there is one group at the approximate distance of this arm. We find the YSO groups in Cygnus X to be spread linearly, spanning a factor of ∼2 in distance (∼1–2 kpc). This situation is consistent with looking down the length of the arm. Cygnus X is thought to lie at an end of a long molecular filament (Alves et al. 2020; Zucker et al. 2020) that contains multiple prominent star-forming regions (e.g., Orion, Taurus, the North America Nebula). The additional length of Cygnus X may increase the total length of this structure by 50%. A similar result is reported by Xu et al. (2016). In Quadrant IV, a chain of groups is located at $X \sim 0$ and extends from ∼1 to ∼3 kpc in distance toward the constellation Vela. The orientation of this chain suggests that these groups could connect with the Local Arm.

Sagittarius–Carina Arm: Numerous star-forming regions can be found within 20° of the direction of the Galactic center, including some famous regions like the Trifid Nebula, the Lagoon Nebula, and NGC 6334. Some of these groups (including the aforementioned famous regions) form a coherent, chevron-shaped structure that has a vertex pointing toward us at a distance of ∼1.0 kpc and edges that extend away with lengths of ∼1 kpc. This collection of star-forming regions is at the approximate distance of the Sagittarius–Carina Arm from Reid et al. (2019), implying that the arm is an active site of star formation activity. However, the angles of the edges of the chevron are inconsistent with the angle of the arm predicted by Reid et al. (2019). The linear structure making up the edges of the chevron cannot be a result of the "Fingers of God" effect, because it is not oriented along our line of sight. We find relatively little sign of YSO groups associated with the Sagittarius–Carina Arm at Galactic longitudes beyond $\ell > 30°$ in Quadrant I or between $300 < \ell < 330°$ in Quadrant IV.

Scutum–Centaurus Arm: This arm is less clearly delineated by stellar groups than the others, possibly owing to large distance uncertainties at the distance of this arm. However, there is an increase in the density of groups near this arm in Quadrants I and IV.

Perseus Arm: This arm is intersected by the SMOG field, and three groups have distance estimates consistent with the center of this arm from Reid et al. (2019). The large distance of this arm may decrease our sensitivity to YSOs associated with it.

Inter-arm: There are multiple stellar groups that appear to be located between the spiral arms from Reid et al. (2019). For example, within ∼10° of the Galactic Center, many groups appear located between the Sagittarius–Carina and Scutum–Centaurus arms. In Quadrant I, several groups are located between the Local Arm and the Sagittarius–Carina Arm.

7.2. Galactic Rotation

The procedure for calculation of mean proper motions for the groups is similar to the calculation of heliocentric distances, using a weakly Gaussian prior for $\mu_{\alpha^*,0}$ and $\mu_{\delta,0}$ instead.

Figure 21 displays the proper motions in Galactic longitude and latitude as function of ℓ. The expected distribution for stars in circular Galactic orbits would be governed by Galactic

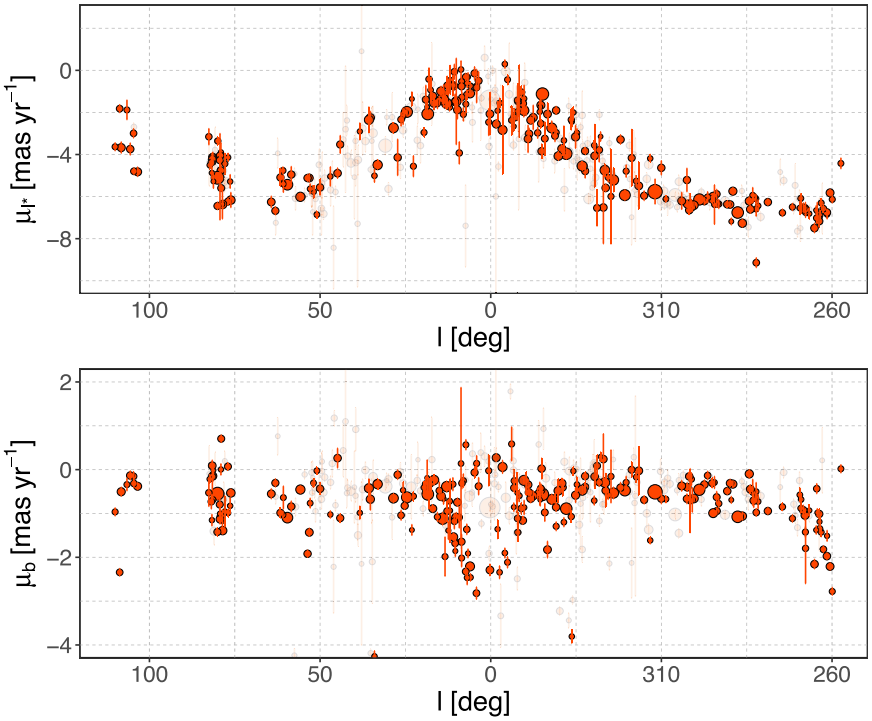

Figure 21. Position–velocity diagrams for the YSOs candidates: μ_{ℓ^*} vs. ℓ (top) and μ_b vs. ℓ (bottom). The distributions are the cumulative effects of Galactic rotation, solar motion, and peculiar velocities of YSO associations. Size and shading of circles are the same as in Figure 20. Error bars combine the statistical uncertainties on group motions with the ± 0.07 mas yr^{-1} Gaia DR2 systematic uncertainty.

rotation, parameterized by the Oort (1927) constants, and effects from solar motion, which are distance-dependent (Bovy 2017). On the μ_{ℓ^*} versus ℓ diagram, to first order in the Galactic plane, this would be a sinusoid with period 180° for Galactic rotation added to a sinusoid with period 360° for solar motion. This overall structure appears to dominate the μ_{ℓ^*} versus ℓ plot for our YSO groups, but there are hints of deviations that we will discuss in more detail in a subsequent paper.

On the μ_b versus ℓ diagram, several structures can be seen. For example, between $\ell \sim 5°$–$25°$, there is a diagonal chain of groups in position–proper-motion space. These groups correspond to one of the edges of the chevron-like structure detected in the *XY* diagram. The dispersion in μ_b is slightly higher around $\ell \approx 75°$ (Cygnus X) and around $\ell \approx 260°$ (Vela)—both of which correspond to extended structures along the line of sight.

7.3. Spatial Distribution of YSOs by Class

YSO candidates of all SED classes are spatially clustered, but the classes corresponding to earlier evolutionary stages tend to be more strongly clustered, as can be seen in the section of the Galactic midplane in Figure 22 (left panel). In this region near the young cluster NGC 6823, Class I sources mostly lie within the densest groups, but the other classes are comparatively more distributed. The *K* function (Ripley 1976) can be used to quantitatively compare the relative strength of clustering for these populations. We used the SPATSTAT package (Baddeley 2017) to estimate *K* as a function of angular separation *r* for Class I, flat spectrum, Class II, and Class III objects with 95% confidence intervals estimate using the bootstrap method of Loh (2008). On this log–log plot (Figure 22, right panel), at angular separations of several arcminutes, the slope for Class I YSOs is significantly flatter than for flat spectrum YSOs, which is also significantly flatter than those for Class II and III YSOs, implying that the earlier stages are more clustered. This finding agrees with numerous other examinations of the spatial distribution of sources by YSO class (e.g., Sung et al. 2009; Samal et al. 2010; Buckner et al. 2020). We note that even some candidate Class I YSOs appear isolated: for example, ∼100 of these objects (<1% of the Class I YSOs) are separated from their nearest neighbors in our catalog by more than 10′.

Figure 23 shows the smoothed distributions of sources of various classes in both Galactic longitude and latitude. In longitude, the normalized distributions of YSOs of all SED classes are similar, whereas in latitude, the distributions of earlier classes (e.g., Class I and flat spectrum) are more strongly concentrated near the midplane than those of the later classes (e.g., Class II and III). This may be a result of the dispersal of YSOs, if stars are born in regions nearest the midplane and then drift away. For example, a YSO traveling at a tangential velocity of ∼2 km s^{-1} at a distance of ∼2 kpc could travel 0°.25 from its point of origin in ∼5 Myr. This is enough to flatten the distribution of b shown in the figure but not the distribution of ℓ.

The sources with strong silicate absorption are more concentrated toward the Galactic center than other YSOs. This could be an effect of the higher interstellar dust column densities in this direction. The YSOs with strong PAH emission also appear to be preferentially concentrated toward the inner Galaxy, but the peak of the distribution appears to be in star-forming regions around $\ell \sim 330°$.

8. Optical Variability

Optical variability, with amplitudes ranging from several tenths of a magnitude to outbursts of multiple magnitudes, is

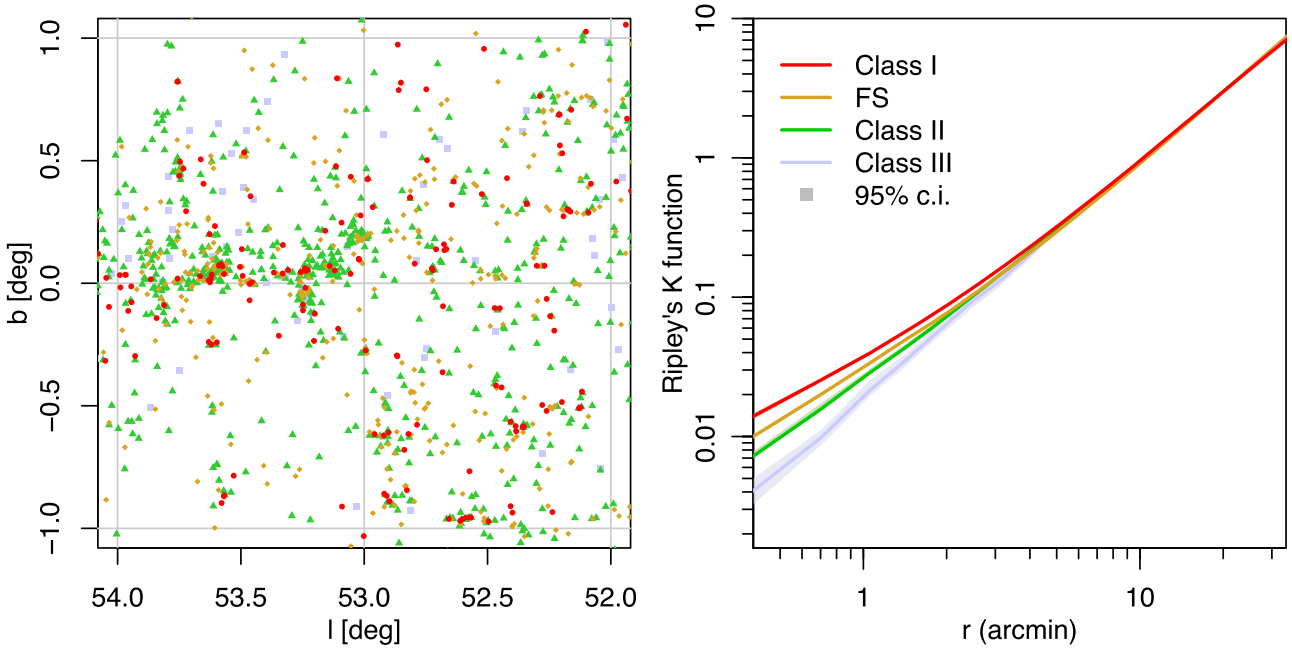

Figure 22. Diagnostics of spatial clustering for stars of different classes. The same color-coding is used to represent each class in both panels. Left: Spatial distribution of YSO candidates in a $\sim 2^\circ \times 2^\circ$ sample area in Sagitta. Right: Ripley's reduced second moment $K(r)$ function, with 95% confidence intervals calculated for all stars. The flatter slope for the Class I and flat-SED YSOs at small angular separations implies these sources are more strongly clustered than the Class II and Class III objects. All classes exhibit some spatial clustering, but the earlier evolutionary classes tend to be more clumped. Nevertheless, examples of isolated YSOs of all classes can also be found.

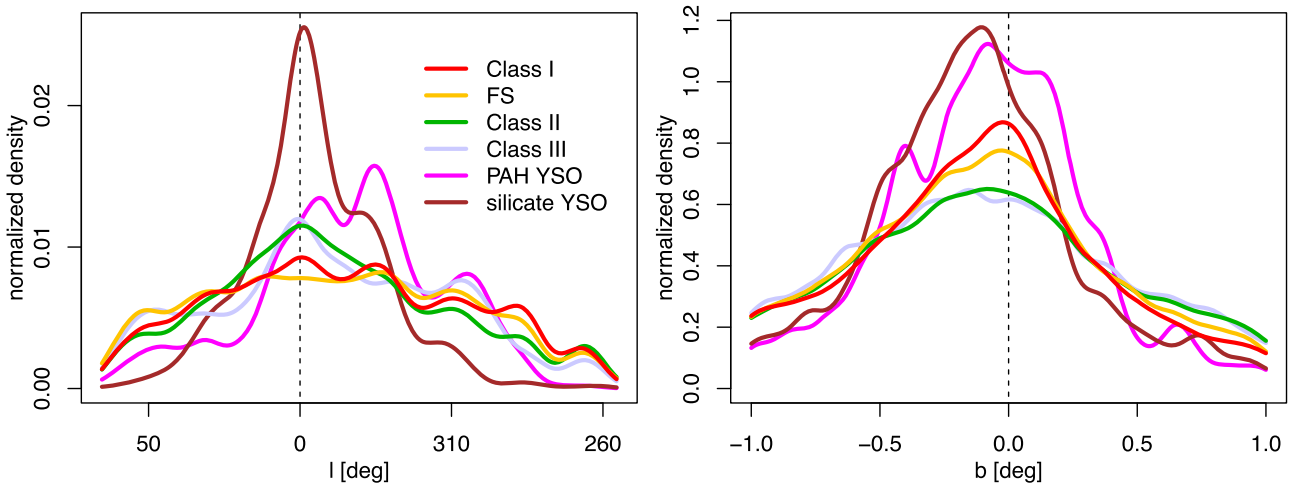

Figure 23. Distributions of ℓ and b for stars of different YSO classes and stars with probable PAH emission or silicate absorption. Densities are calculated with a 5° kernel in ℓ and a $0\overset{\circ}{.}05$ kernel in b. Only GLIMPSE I, II, 3D, and Vela-Carina data are included in this plot. In Galactic longitude, the distributions of Class I/FS/II/III stars are similar, but in Galactic latitude, the younger YSO classes are more strongly peaked toward the midplane. Sources with silicate absorption are strongly peaked near the Galactic center, while the peak in PAH source density is offset by $\sim 30^\circ$ from the center.

associated with YSOs (e.g., Joy 1945; Herbig 1954; Cody & Hillenbrand 2018) and has even been used as a criterion for identifying previously unrecognized YSOs (e.g., Contreras Peña et al. 2017). Thus, strong optical variability from YSO candidates in our Spitzer selected sample can be regarded as corroborating evidence for the youth of these objects. To investigate which sources show optical variability, we use photometric measurements from the Zwicky Transient Facility (ZTF; Bellm et al. 2019), which is sensitive to a variety of variability phenomena from YSOs, with its cadence of approximately one observation per night (Graham et al. 2019), including dips due to occultation from circumstellar dust, variations in accretion rate, magnetic flares, and rotational modulation due to large star spots.

We cross-match our YSO candidates to the ZTF DR3 (Masci et al. 2019) catalog using a match radius of $1''$, and use ZTF sources with at least 10 measurements in the r band between 2018 April and 2019 June, excluding observations from the high-cadence deep-drilling program; the median number of observations is $\sim$130. This yields 7585 YSO candidates with usable ZTF light curves. This represents a relatively small fraction of our entire catalog because many of the Spitzer sources are not detected in the optical and ZTF is only available for the Northern Hemisphere. Nevertheless, in terms of the

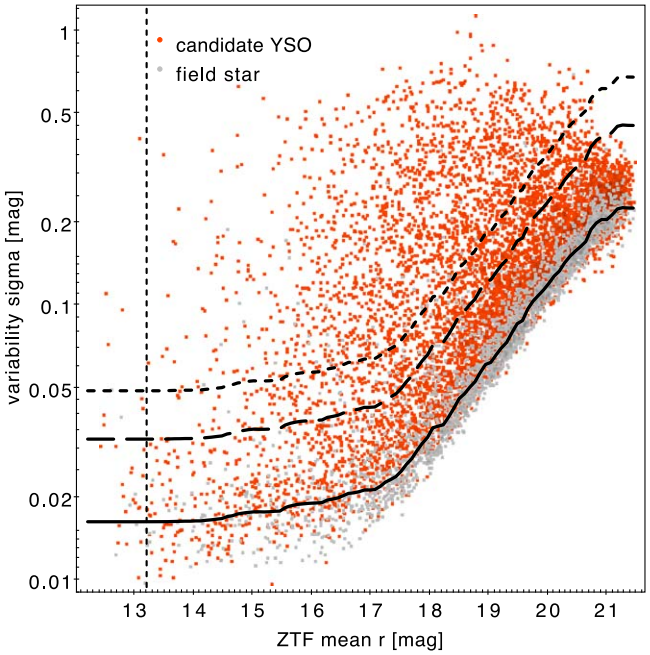

Figure 24. Standard deviation of the ZTF light curve's variability in the *r*band ($\sigma_{\rm var}$) vs. mean *r* for 7,585 YSO candidates (red) and ∼2,400 randomly selected field stars (gray) from the same areas of the sky. The solid black line is the median $\sigma_{\rm var}$ as a function of magnitude for the field stars, and the lines above are two times (dashed) and three times (dotted) this level. We define stars above the dotted line as having "strong variability," between the dashed and dotted lines "moderate variability," and below the dashed line "low or insignificant variability."

absolute numbers of sources, this sample is moderately large and useful for statistical analysis. The sample is skewed toward Class II and III YSOs, but over a thousand sources categorized as Class I or flat SED YSOs are included as well.

To characterize variables, we calculate the *r*-band light curve's standard deviation $\sigma_{\rm var}$, the mean magnitude $\bar{r}$, and skewness of the distribution. Figure 24 shows $\sigma_{\rm var}$ versus $\bar{r}$ for our YSO candidates (red points) as well as ∼2400 randomly selected field stars (gray points) from the same region of the sky. For the field stars, we have estimated the running median $\sigma_{\rm var}$ as a function of $\bar{r}$, shown as the solid black line. Eighty-eight percent of the YSO candidates have $\sigma_{\rm var}$ values greater than this line, indicating that the YSO candidates have higher variability on average than the field stars. Following Fang et al. (2020), we use the median $\sigma_{\rm var}$ for field stars to delineate variability thresholds for the YSO candidates. Objects where $\sigma_{\rm var}$ is more than 3 times the median value for field stars are considered to show high variability (above the dotted line), objects between 2–3 times the median value are considered to show moderate variability (between the dashed and dotted lines), and objects <2 times the median value have low or insignificant variability (below the dashed line). Using these definitions, we find 1695 with high variability, 914 with moderate variability, and 4976 with low or insignificant variability.

Many interesting variables have asymmetric variability, as indicated by skewness. For example, "dipper" YSOs, where fading is caused by circumstellar material passing in front of the star, tend to have positive skewness (e.g., Cody & Hillenbrand 2018). Of the highly variable stars in our sample, 76% have positive skewness, with a mean skewness of 0.3.

We find a slight tendency for strong or moderate YSO candidates to be more spatially clustered than weak or nonvariable YSO candidates. For example, 56% of stars with strong variability are members of HDBSCAN groups, while 51% of stars with moderate variability are members, and 47% of stars with weak or no variability are members.[17] If nonvariable candidates have a higher probability of being non-YSO contaminants, this could influence the observed trend, since contaminants are not expected to be clustered. However, even among the YSO candidates with high variability, 44% are not members of HDBSCAN groups, providing further evidence suggesting that many of the relatively isolated candidates may still be legitimate YSOs.

Figure 25 shows a sample of light curves from sources showing high variability. Many of these stars exhibit dipping features with sharp bottoms, a morphological feature typically associated with extinction from dust (possibly in the circumstellar disk) briefly passing in front of the stars. The timescales of the dips seen in these ZTF light curves can range from several days to multiple months. Some light curves show long-timescale trends, like the gradual brightening seen in SPICY 110421. Other stars' light curves exhibit outbursts, with SPICY 116663 shown as an example with a particularly large >3 mag amplitude. These features are similar to the categories of YSO variability identified by Cody & Hillenbrand (2018), albeit many of the structures identified in their K2 study occur on a shorter timescale than we are sensitive to with the cadence of ZTF. Some of our candidates YSOs also exhibit periodic behavior, which is thought to be associated with the rotation periods of the stars due either to star spots or material orbiting at the corotation radius (Herbst et al. 1994; Stauffer et al. 2017). SPICY 108092 is an example of one such star, which includes periodic rotation along with dips, bursts, and long-timescale changes.

The ZTF YSO light curves exhibit considerable diversity in their morphologies. Given that the objects in the SPICY catalog were selected in a uniform way independent of their variability, this data set may be useful as a training set for future efforts to develop a classifier of YSOs based on optical variability.

9. Comparison to Other YSO Catalogs

Both we and Robitaille et al. (2008) have searched GLIMPSE catalogs for YSOs, but our catalog extends the search to much fainter magnitudes. We have less stringent source quality criteria, we do not impose an ab initio [4.5]–[8.0] ⩾ 1 color cut, and most significantly, we include sources fainter than the flux limits imposed in their catalog. Their 10 mJy limit in the 8.0 μm band ([8.0] < 9.52 mag) would discount 73% of our YSOs. In the overlapping regions, the GLIMPSE I and II survey areas, we identify a number of YSO candidates >4 times greater than the number of red sources from Robitaille et al. (2008) catalog.

Unlike our catalog, most of the sources from Robitaille et al. (2008) are bright enough to have been detected by observations at 24 μm. For these objects, they use a simple heuristic set of color criteria to separate YSO and AGB candidates: sources with [4.5] ⩽ 7.8 or [8.0]–[24] < 2.5 are considered likely AGB stars, while sources with [4.5] > 7.8 and [8.0]–[24] ⩾ 2.5 are likely YSOs. They acknowledge that a division like this is likely to produce erroneous classifications in either direction. Out of 16,670 "red sources" from Robitaille et al. (2008) that they labeled either YSO (9387) or AGB (7283), 13,290 (80%) were reidentified as candidate YSOs by our analysis, including

[17] Fisher's exact test of contingency tables (Fisher 1922) indicates that, even though the effect size is small, these differences are statistically significant at the $p < 0.01$ level.

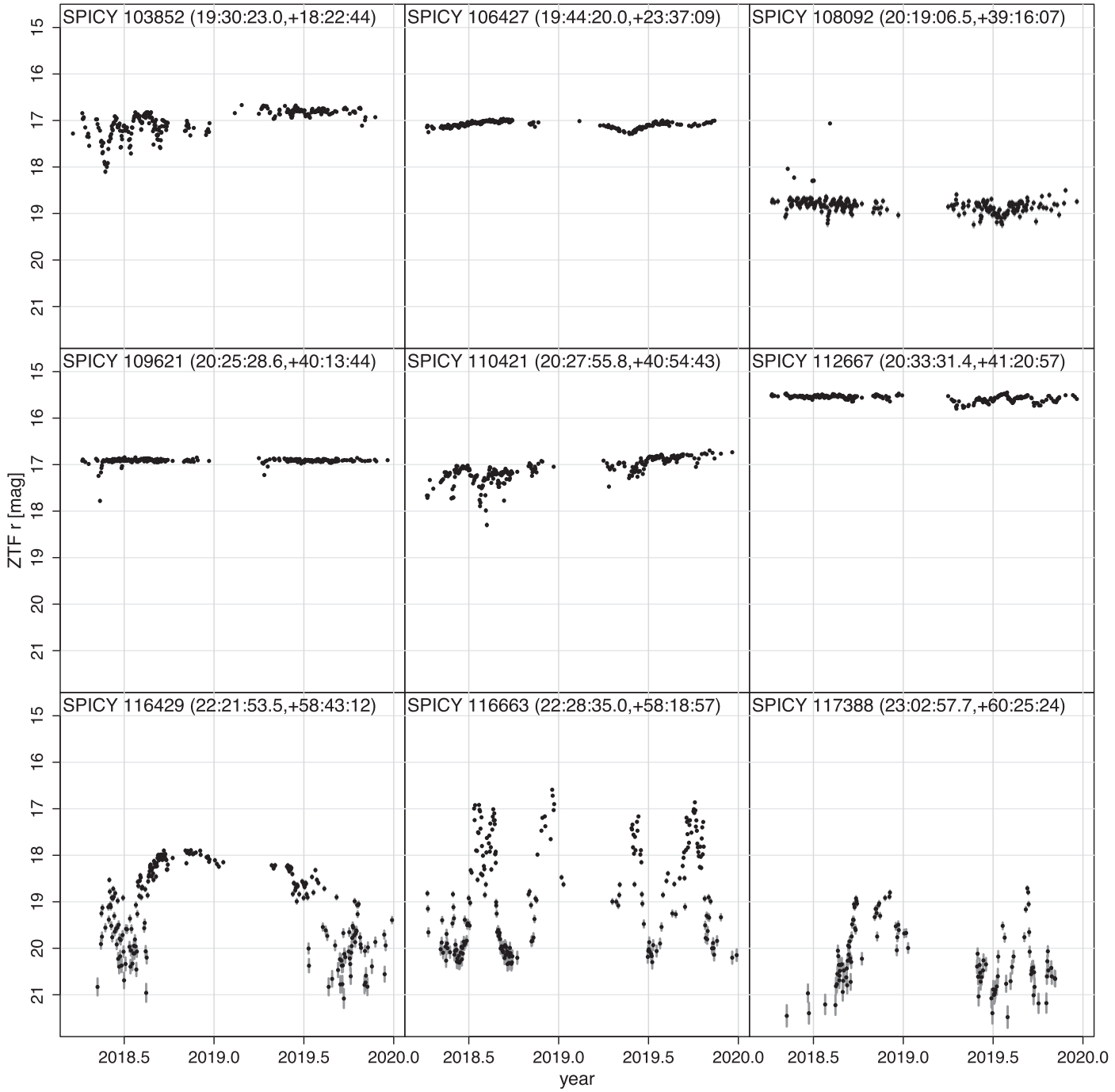

Figure 25. Sample ZTF light curves for several YSO candidates showing strong variability. These light curves exhibit diverse behaviors, including dipping, slow variation in brightness, outbursts, and periodicity. (ZTF light curves for 7585 SPICY sources are provided as data behind the figure.)

(The data used to create this figure are available.)

8637 (92%) of those labeled YSO and 4653 (64%) of those labeled AGB. Assuming the Robitaille et al. (2008) classifications are accurate, and extrapolating to fainter objects, this could suggest that up to ∼35% of our YSO candidates are misclassified. However, of the 4653 sources classified as YSOs by us but as AGB by Robitaille et al. (2008), 33% are members of the spatial clusters from Section 7, suggesting that some of the objects they label AGB stars could be YSOs.

As discussed in the introduction, a variety of strategies have been used to identify YSOs from IRAC colors. The SMOG field, which was published by Winston et al. (2019) using a modified version of the Gutermuth et al. (2009) color selection rules, provides an excellent testbed for such a comparison. Comparison between our SMOG candidates (1524 objects) and theirs (4648 objects) shows that our selection methodology is more restrictive; 97% of our YSO candidates were also classified as YSOs by Winston et al. (2019), while only 32% of their YSO candidates were classified as YSOs by us. Figure 26 (right) shows that objects from their list that are not included by us tend to be either objects with bluer colors or objects fainter

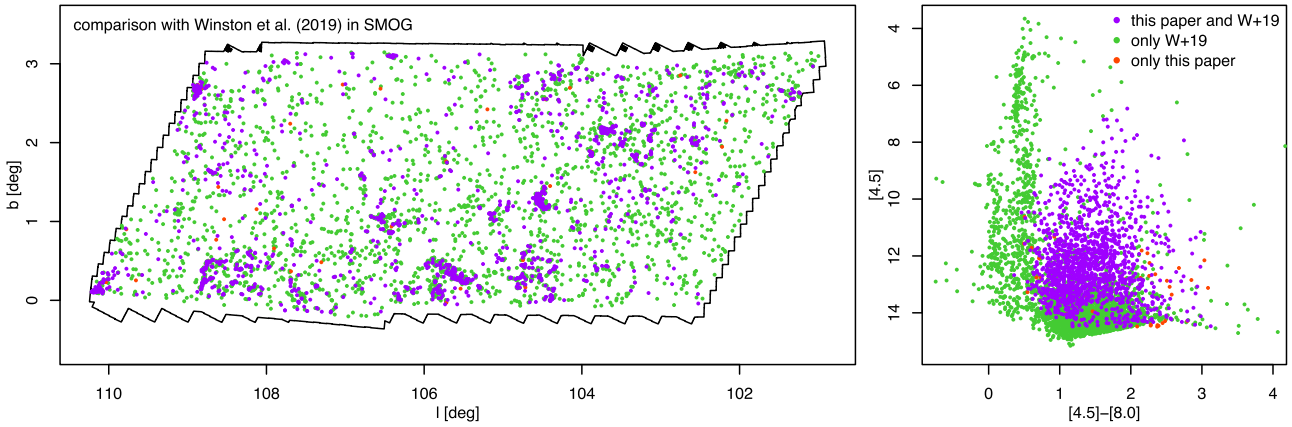

Figure 26. Comparison between our catalog and Winston et al. (2019) in the SMOG field, where Spitzer observations are deeper than the main GLIMPSE survey. YSOs in both catalogs are color-coded purple, objects only in Winston et al. (2019) are green, and a small number of objects only in our catalog are red. Left panel shows the spatial distribution. Right panel shows a color–magnitude diagram. These diagrams indicate that our criteria are more selective than Winston et al. (2019), and the sources we omit tend to either have bluer IR colors or be fainter. Consistency between the catalogs is greatest among the spatially clustered sources, but our catalog does not include many of the nonclustered objects found by Winston et al. (2019). However, formal assessment of the accuracy of either catalog would require additional information from follow-up spectroscopic observations.

than most of our sample. The difference in magnitude distributions—ours peaking at $[4.5] = 13$ mag and theirs peaking at $[4.5] = 14.5$ mag—may be a limitation related to our training set which was dominated by objects from the shallower GLIMPSE survey areas (Section 2). In spatial distribution (Figure 26, left), our candidate YSOs tend to be more spatially clustered than those from Winston et al. (2019).

The "Star Formation in the Outer Galaxy" (SFOG; Winston et al. 2020) YSO catalog was recently produced for the GLIMPSE 360 fields, observed during Spitzer's warm mission, meaning that Spitzer's 5.8 and 8.0 μm bands were unavailable. In Galactic coverage, this catalog is largely complementary to ours, but overlaps in the regions of SMOG and part of Cygnus X. The catalogs also overlap in Vela in Galactic longitude, but cover different ranges of Galactic latitude.

WISE covers similar wavelengths as Spitzer, but provides photometry for the whole sky. All-sky searches for YSOs in WISE data include Marton et al. (2016) and Marton et al. (2019), with the latter using cross-matches with Gaia. Below, we compare our catalog to the list of ∼130,000 candidate Class I–II objects from Marton et al. (2016). This paper also lists >600,000 candidate Class III sources, but we do not include these in our comparison. The reason for this is that Class III sources are a minority in our catalog, but they make up the majority of the candidates from Marton et al. (2016). Within the footprint of our catalog, Marton et al. (2016) identify ∼75,000 Class I–II WISE sources, whereas we identify ∼110,000 Class I–II IRAC sources. It is unsurprising that Spitzer can identify more YSOs in the Galactic midplane, due to IRAC's higher spatial resolution and WISE's greater susceptibility to detector saturation from bright nebulosity. Using a 2″ match radius, there are only ∼5000 sources in common between our catalog and theirs; visual inspection of the spatial distributions of the unmatched candidates reveal that we include more clustered YSO candidates (often more difficult to observe with WISE), while they include more spatially distributed candidates.

An effort to identify intermediate-mass young stars (e.g., Herbig Ae/Be stars) via machine learning was made by Vioque et al. (2020), who identify 8470 candidates using public optical and infrared catalogs. Their list, focused on the higher end of the initial mass function, includes many fewer stars than our

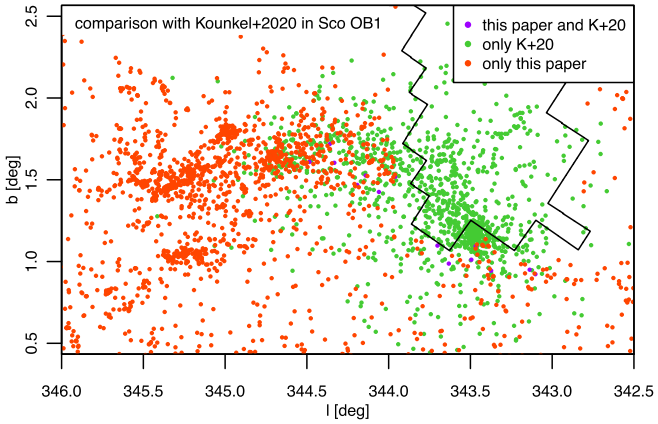

Figure 27. Comparison between our catalog and Kounkel et al. (2020) (age < 10 Myr) in a field centered around the Sco OB1 association. There are few individual stars in common (purple points) between our infrared-excess selected YSOs (red points) and their astrometrically selected sample (green points). Nevertheless, these catalogs appear complementary because they trace different components of the same stellar association. The boundary of the GLIMPSE field (black lines) excludes a section of the sky in the upper right of this figure from our survey.

catalog; however, within the spatial overlap area, most of their candidates were reselected by us.

Another relevant catalog is provided by Kounkel et al. (2020) who identify groups of comoving stars, including clusters, associations, moving groups, and stellar streams, in Gaia DR2 using a search radius of 3 kpc. Although these systems are not necessarily young, their catalog does include ∼35,000 members of groups with ages <10 Myr, many of which are located near the Galactic midplane. Only ∼300 objects are in common between our catalog and theirs, but this low fraction appears to be related to different selection biases, in particular their stringent Gaia quality cuts, which are only met by 4% of our YSO candidates. The objects in common are mostly assigned to groups with ages from Kounkel et al. (2020) between 4–10 Myr; a handful of objects with older ages may result from either errors in their age estimates or contaminants in our catalog. Visual examination suggests that the catalogs reveal complementary aspects of stellar associations, with Kounkel et al. (2020) mostly selecting diskless members and us

the disk/envelope-bearing members. In Figure 27, we show Sco OB1 as an example where the combination of both lists provides a more complete picture of the association.

10. Conclusions

We present a catalog of 117,446 candidate YSOs (∼90,000 of which are new identifications) in the Galactic midplane from the GLIMPSE survey (Benjamin et al. 2003; Churchwell et al. 2009) and extensions of this survey observed during Spitzer's cryogenic mission. We classify objects obtained from the GLIMPSE I, II, 3D, Vela-Carina (Majewski et al. 2007; Zasowski et al. 2009), Cygnus X (Beerer et al. 2010), and SMOG (Winston et al. 2019) IRAC catalogs, using ancillary data from the near-IR 2MASS (Skrutskie et al. 2006), UKIDSS (Lawrence et al. 2007), and VVV (Minniti et al. 2010) surveys, in the most comprehensive search for YSOs in the inner Galactic midplane to date. This catalog is largely restricted to the inner Galaxy, and thus is complementary to YSO searches in the outer Galaxy (e.g., Winston et al. 2020).

Classification of candidates was entirely based on near-IR and IRAC photometry, and our random forest diagnostics confirm that the IRAC bands were the most important for classification. The IRAC catalogs contain many sources not detected by MIPS or WISE; this is because IRAC, particularly when processed by the GLIMPSE pipeline, is more sensitive in regions of the Galaxy with high crowding and nebulosity. By focusing on IRAC, we are able to identify tens of thousands of new YSO candidates that we would not have been able to discern if we required additional bands. Depending on the science application, future studies that use our YSO list may wish to augment our catalog with YSOs selected using other wavelengths.

The spatial distribution of the candidates, as projected on the sky, is highly structured, with cluster-like and filament-like patterns, but also includes a substantial nonclustered population. We have not used spatial information as an input to the classifier, because the extent of spatial clustering of YSOs is still an open question and we wish to minimize the influence of selection effects on the observed spatial distributions. From the HDBSCAN algorithm, we identify ∼400 groups of YSOs and estimate their distances and proper motions from the mean astrometry of members detected by Gaia DR2.

The YSOs we identify in the Galactic midplane are mostly at distances $\gtrsim$1 kpc. Some YSO groups appear associated with the Orion, Sagittarius–Carina, and the Scutum–Centaurus arms of the Galaxy, but do not appear to closely trace the estimated arm centers found by other methods (e.g., Reid et al. 2019). Near the boundary between Galactic Quadrants I and IV, a large collection of YSO groups are located at the approximate distance of the Sagittarius–Carina Arm. However, these groups are not aligned parallel to the arm, instead forming a chevron-like shape.

From the portion of our catalog visible to the ZTF survey, our YSO candidates tend to be more variable than field stars in the same region of the Galaxy. Nearly half the stars measured have statistically significant ZTF variability. Visual examination of the sources with the highest variability amplitudes suggests that most of them have light-curve morphologies that resemble those expected for YSOs, with large dipping or bursting features. This data set provides a useful testbed for future work on statistical classification of YSO light curves.

Although the properties of our sample, including optical/infrared photometry, spatial clustering, and variability, are consistent with most of the candidates being YSOs, the level of contamination is difficult to constrain without follow-up observations. Although most objects are optically faint, the large total number of YSO candidates means that there are plenty of objects bright enough to follow up with optical spectroscopy. For example, ∼66,000 YSO candidates have $G < 19$ mag, the faint limit for future large spectroscopic surveys such as WEAVE (Dalton et al. 2014) or 4MOST (de Jong et al. 2019), and more than 85% of them are newly proposed in this paper. Furthermore, in the IR, ∼2000 YSO candidates have $H < 11.5$ mag, bright enough for an instrument like APOGEE (Blanton et al. 2017; Cottle et al. 2018), and approximately half of them are newly proposed in this paper. Our candidates have been selected with a nearly uniform methodology, so they should provide a useful statistical sample for further studies.

We thank Robert Benjamin for useful discussions about GLIMPSE and the spiral structure of the Galaxy, and Philip Lucas and Leigh Smith for assistance with the UKIDSS and VVV catalogs. This work is a result of the 6th COIN Residence Program (CRP#6; https://cosmostatistics-initiative.org/residence-programs/crp6) held in Chamonix, France in 2019 August. COIN is financially supported by CNRS as part of its MOMENTUM program over the 2018–2020 period. This work is based on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. This work has also made use of data from the European Space Agency mission Gaia, processed by the Gaia Data Processing and Analysis Consortium. Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement. This work is also based in part on observations obtained with the Samuel Oschin 48 inch Telescope at the Palomar Observatory as part of the Zwicky Transient Facility project. ZTF is supported by the National Science Foundation under grant No. AST-1440341 and a collaboration including Caltech, IPAC, the Weizmann Institute for Science, the Oskar Klein Center at Stockholm University, the University of Maryland, the University of Washington, Deutsches Elektronen-Synchrotron and Humboldt University, Los Alamos National Laboratories, the TANGO Consortium of Taiwan, the University of Wisconsin at Milwaukee, and Lawrence Berkeley National Laboratories. This research has made use of the NASA/IPAC Infrared Science Archive, which is funded by the National Aeronautics and Space Administration and operated by the California Institute of Technology. A.K.M. acknowledges support from the Portuguese Fundação para a Ciência e a Tecnologia (FCT) through grants SFRH/BPD/74697/2010, PTDC/FIS-AST/31546/2017 and from the Portuguese Strategic Programme UID/FIS/00099/2013 for CENTRA. M.A.K. acknowledges support from the Chandra grant GO9-20002X for analysis of data related to the Trifid Nebula.

Facilities: 2MASS, Gaia, Spitzer (IRAC, MIPS), UKIRT, VISTA/VIRCAM, WISE, ZTF, IRSA.

Software: caret (Kuhn 2015), mclust (Scrucca et al. 2016), hdbscan (McInnes et al. 2017), modeest (Poncet 2019), mclust (Scrucca et al. 2016), PostgreSQL (PostgreSQL Global Development Group 2020), Python, R (Core Team 2019), rjags (Plummer 2017, 2019), SAOImage DS9 (Joye & Mandel 2003), sbgcop (Hoff 2018), TOPCAT & STILTS (Taylor 2005).

Appendix A
GLIMPSE Flags

We visually examined a sample of IRAC images from crowded, nebulous regions of the Galaxy in order to investigate

what GLIMPSE flags, including the "close source flag" and the number of detections in each band, imply about the reliability of source detection. The close neighbors for YSO candidates can usually be seen in the 3.6 micron image, but rarely in the 8.0 micron images, possibly because they are fainter at this wavelength, since most neighbors lack IR excess. In such cases, it appears that the GLIMPSE pipeline has correctly identified both sources. Nearly all sources with two or more detections in a band look like a bona fide point sources in that band's images; however, some of the sources detected only once look like point sources while others do not. Thus, having fewer than two detections is an indicator that a source could be less reliable. Nevertheless, we do not filter out sources based on either the "close source flag" or the number of detections in each band, because any such cuts would remove numerous objects that appear to be good YSOs. These GLIMPSE flags are included in Table 1 for any users of our YSO catalog who would like to make alternate choices for their own scientific applications.

Table 3
Regions near the Galactic Midplane without Significant YSO Populations

$\ell_{\rm min}$ (deg)	$\ell_{\rm max}$ (deg)	$b_{\rm min}$ (deg)	$b_{\rm max}$ (deg)
275.63	277.85	−1.52	0.51
328.95	331.37	−2.97	−1.17
349.02	349.40	−2.22	−0.09
358.32	2.08	0.81	3.87
1.25	1.44	−0.33	−0.08
0.12	2.28	−4.43	−3.11
0.12	6.43	−3.12	−1.88
23.89	26.47	−3.01	−1.22
28.84	31.46	−2.97	−1.30
32.06	32.73	−1.14	−0.23
43.22	44.37	0.28	1.13
47.79	48.46	0.20	0.86

Note. This table provides the lower left and upper right boundaries of the rectangular regions from which we obtained our labeled list of "field" objects. Rectangles are drawn with lines of constant Galactic ℓ and b.

Appendix B
Red Non-YSOs in the IRAC Catalogs

B.1. Evolved Stars

Certain evolved stars, including dusty red giants, AGB stars, post-AGB stars, and red super giants (RSGs), have red IR SEDs due to dusty stellar winds (e.g., Marengo et al. 1997, 1999; Groenewegen 2012; Chun et al. 2015; Suh 2020), making such objects a significant category of contaminant in infrared YSO catalogs (e.g., Robitaille et al. 2008; Povich et al. 2013). Reiter et al. (2015) present a sample of AGB stars (including O-rich, S-rich, and C-rich stars) and RSG stars with *JHK* and IRAC photometry. In the near-IR, the $J-K$ colors of this sample ($J-K \gtrsim 0.9$) are consistent with the group of probable contaminants on the J versus $J-K_s$ diagram that are brighter and redder than the typical YSOs (Figure 9, left panel). In IRAC color space, the distribution of the AGB+RGS sources partially overlaps the distribution of YSO candidates. However, most of them have IRAC colors (e.g., $[3.6]-[4.5] \lesssim 0.5$, $[4.5]-[8.0] \lesssim 1$) bluer than the typical colors of YSOs but similar in color to the bright sources classified as probable contaminants (e.g., Figure 9, right panel).

B.2. Extragalactic Sources

Extragalactic sources, including active and star-forming galaxies, can have mid-IR colors that mimic the IR excesses of YSOs (e.g., Stern et al. 2005; Jarrett et al. 2011). These sources can contribute a significant number of possible contaminants in some YSO searches (e.g., Harvey et al. 2007; Gutermuth et al. 2008). However, in the GLIMPSE survey, extragalactic contamination is expected to be lower, owing to shallower IRAC observations and high extinction near the Galactic midplane (e.g., Kang et al. 2009). Jarrett et al. (2011) provide a sample of IRAC sources in fields dominated by extragalactic sources. The galaxies in these fields tend to have [3.6]–[4.5] ≈ 0–1.25 and [4.5]–[8.0] ≈ 1–4. This distribution more closely resembles the distribution of probable contaminants in our sample than it does our YSO candidates (Figure 11, lower right). Their extragalactic sources mostly have $[3.6] \gtrsim 14$ mag, which is fainter than most of our candidate YSOs.

B.3. Labeled Non-YSOs in the Training Set

For the random forest classifier, the sample of non-YSOs in our training set is equal in importance to the sample of YSOs. As with the labeled YSOs, the labeled field objects are obtained from the set of NIR+IRAC sources that could not be fit by a reddened stellar photosphere (Section 3.1). Thus, they also represent objects with red IR colors.

We generate our list of "non-YSOs" using both sources within the boundaries of the MYStIX star-forming regions that were classified as non-YSOs and field stars in regions of the Galaxy where there is no star formation activity. For the first category, we include all IRAC sources that lie within the MYStIX fields that show no indication of youth—requiring both rejection as a YSO based on its SED and nondetection of X-ray emission.

We also include stars from rectangular regions of the sky in areas where there is no evidence for star formation or the presence of YSOs. These regions, listed in Table 3, have been chosen to sample stars along different Galactic latitudes and longitudes and along lines of sight with different levels of extinction. Given the high numbers of stars within these fields, a random subsample of these stars are included in the training data for the classifier. Altogether, there are 14,019 labeled field objects for the 2MASS+IRAC sample, 5320 for UKIDSS +IRAC, and 4047 for VVV+IRAC.

Appendix C
Training Sets and Imputed Colors

Figure 28 shows color–color/magnitude diagrams for the training set, containing both labeled YSOs (reddish points) and non-YSOs (bluish points) and observed (dark points) and imputed (light points) colors. Here, we show only the 2MASS +IRAC training data, but the general morphologies of the distributions are similar for the target data set, as well as for the training/target UKIDSS+IRAC and VVV+IRAC data sets. For each NIR+IRAC combination, an identical copula was used for every data point regardless of label (YSO or non-YSO) or whether the data point belongs to the training or target set. Thus, any differences that emerge in the distributions of imputed data must emerge from the data itself.

In general, the distribution of the imputed data lies within the distributions traced out by the observed data, with the $J-H$ versus $H-K_s$ diagram being the main exception. In the JHK_s diagram, many sources are missing the $J-H$ color (presumably due to high extinction), and they follow a locus with a slope slightly flatter than that of objects for which both $J-H$ and $H-K_s$ have been measured. This behavior also appears when using UKIDSS or VVV *JHK* photometry, and it may suggest that there is an intrinsic difference in distributions for sources with and without measured $J-H$. The sources without measured $J-H$ also tend to have higher-than-average $H-K_s$ values, possibly influencing the distribution of the imputed sources. Nevertheless, the analysis of the random forest classifier suggests that $J-H$ color is one of the less important features for producing a classification.

None of the other diagrams reveal such large deviations between observed and imputed data, but differences are visible in the distributions of YSOs and non-YSOs. On some diagrams, there are regions with no imputed data (e.g., the lower left of the [3.6]–[4.5] versus [4.5]–[5.8] diagram) because any missing data would have meant that the source would not be under consideration (see Section 3.1).

In the target set, we observe a low number of outliers in regions of color space that are not well-populated by sources from the training set, either by sources labeled "YSO" or "non-YSO," meaning that our classifier would not be able to generate reliable classifications for these objects. These are either rare objects that arise due to the large size ($\sim 5 \times 10^7$ sources) of the full IRAC photometric catalogs or are sources unlikely to have infrared excess but that we failed to remove with our procedures in Section 3.1. Given that we have no basis for classifying these objects with a RF, we opt to be cautious and exclude them from our lists of candidate YSOs. We use the following criteria to define these outliers: [3.6]–[4.5] < -0.3, [4.5]–[5.8] < 0, [4.5]–[8.0] < 0, [3.6]–[5.8] < 0, [3.6]–[5.8] $>$ 6, [4.5]–[5.8] < 0, or [3.6]–[8.0] < 0.

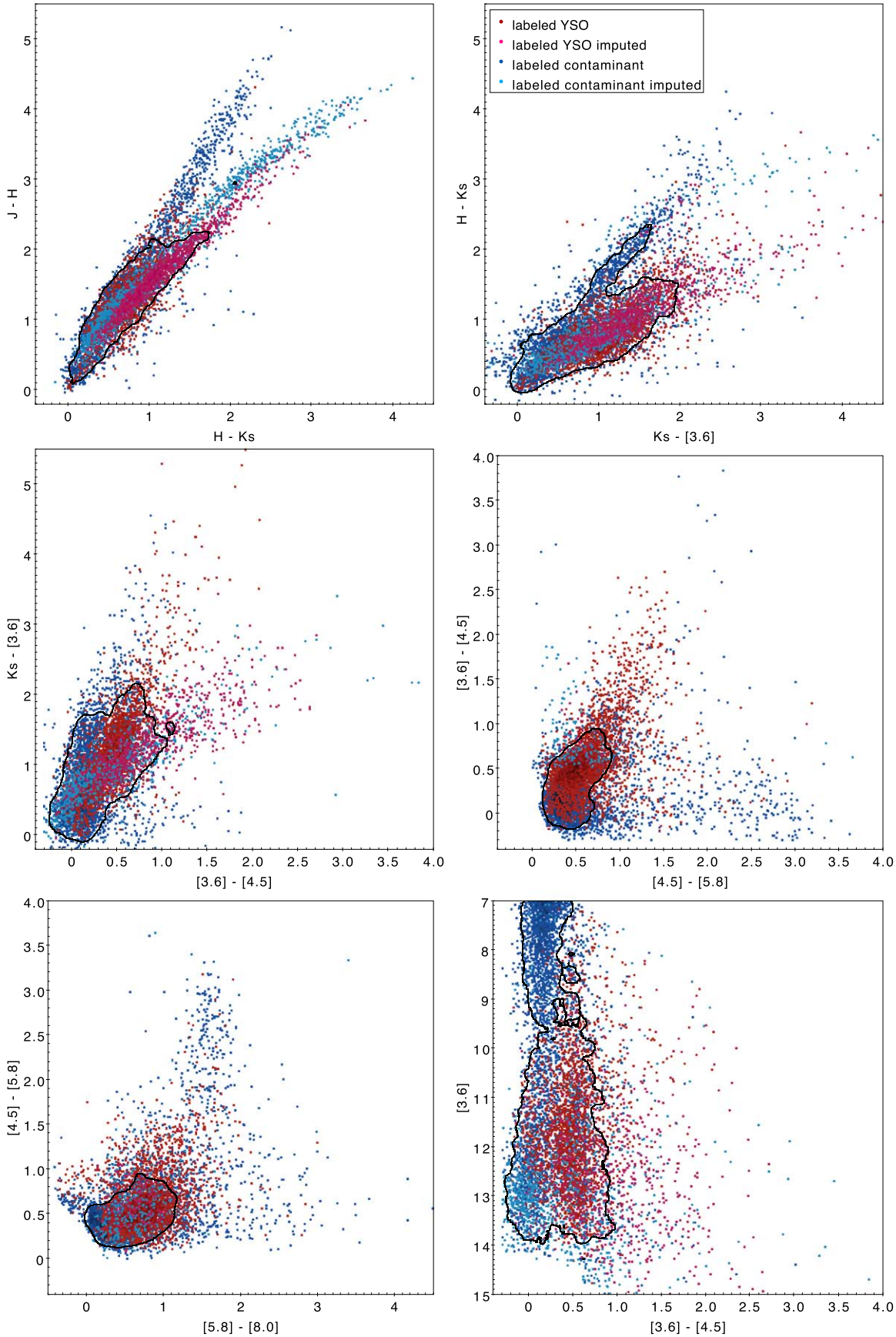

Figure 28. Scatter plots of labeled training data after copula imputation. Objects labeled "YSO" are indicated by reddish points and objects labeled "field" are indicated by bluish points. On each panel, the darker points have measurements of both colors, while points where one or both colors are imputed are marked with a lighter color as indicated by the legend.

ORCID iDs

Michael A. Kuhn https://orcid.org/0000-0002-0631-7514
Rafael S. de Souza https://orcid.org/0000-0001-7207-4584
Alberto Krone-Martins https://orcid.org/0000-0002-2308-6623
Alfred Castro-Ginard https://orcid.org/0000-0002-9419-3725
Emille E. O. Ishida https://orcid.org/0000-0002-0406-076X
Matthew S. Povich https://orcid.org/0000-0001-9062-3583

References

Allen, L., Megeath, S. T., Gutermuth, R., et al. 2007, in Protostars and Planets V, ed. B. Reipurth, D. Jewitt, & K. Keil (Tucson, AZ: Univ. Arizona Press), 361
Allen, L. E., Calvet, N., D'Alessio, P., et al. 2004, ApJS, 154, 363
Alves, J., Zucker, C., Goodman, A. A., et al. 2020, Natur, 578, 237
Anderson, L. D., Bania, T. M., Balser, D. S., et al. 2014, ApJS, 212, 1
Andre, P., & Montmerle, T. 1994, ApJ, 420, 837
Andreani, P., Boselli, A., Ciesla, L., et al. 2018, A&A, 617, A33
Ascenso, J. 2018, in The Birth of Star Clusters, ed. S. Stahler (Cham: Springer International Publishing), 1
Baddeley, A. 2017, Spatial Stat., 22, 261
Beerer, I. M., Koenig, X. P., Hora, J. L., et al. 2010, ApJ, 720, 679
Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002
Benjamin, R. A., Churchwell, E., Babler, B. L., et al. 2003, PASP, 115, 953
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
Bonito, R., Hartigan, P., Venuti, L., et al. 2018, arXiv:1812.03135
Bovy, J. 2017, MNRAS, 468, L63
Breiman, L. 2001, MachL, 45, 5
Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127
Bressert, E., Bastian, N., Gutermuth, R., et al. 2010, MNRAS, 409, L54
Buckner, A. S. M., Khorrami, Z., González, M., et al. 2020, A&A, 636, A80
Bufano, F., Leto, P., Carey, D., et al. 2018, MNRAS, 473, 3671
Campello, R. J., Moulavi, D., & Sander, J. 2013, in Pacific-Asia Conf. on Knowledge Discovery and Data Mining, ed. J. Pei et al. (Berlin: Springer), 160
Candès, E. J., & Donoho, D. L. 2000, Curvelets—A Surprisingly Effective Nonadaptive Representation For Objects with Edges (Nashville, TN: Vanderbilt Univ. Press), 1
Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, ApJ, 345, 245
Carey, S. J., Noriega-Crespo, A., Mizuno, D. R., et al. 2009, PASP, 121, 76
Carpenter, J. M. 2000, AJ, 120, 3139
Castelli, F., & Kurucz, R. L. 2003, in IAU Symp. 210, Modelling of Stellar Atmospheres, ed. N. Piskunov, W. W. Weiss, & D. F. Gray (San Francisco, CA: ASP), A20
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2020, A&A, 635, A45
Chiar, J. E., Ennico, K., Pendleton, Y. J., et al. 2007, ApJL, 666, L73
Chun, S.-H., Jung, M., Kang, M., Kim, J.-W., & Sohn, Y.-J. 2015, A&A, 578, A51
Churchwell, E., Babler, B. L., Meade, M. R., et al. 2009, PASP, 121, 213
Churchwell, E., Povich, M. S., Allen, D., et al. 2006, ApJ, 649, 759
Churchwell, E., Watson, D. F., Povich, M. S., et al. 2007, ApJ, 670, 428
Cody, A. M., & Hillenbrand, L. A. 2018, AJ, 156, 71
Contreras Peña, C., Lucas, P. W., Minniti, D., et al. 2017, MNRAS, 465, 3011
Core Team, R. 2019, R: A Language and Environment for Statistical Computing (Vienna: R Foundation for Statistical Computing), https://www.R-project.org/
Cottle, J. N., Covey, K. R., Suárez, G., et al. 2018, ApJS, 236, 27
Dalton, G., Trager, S., Abrams, D. C., et al. 2014, Proc. SPIE, 9147, 91470L
de Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, Msngr, 175, 3
de Souza, R. S., Dantas, M. L. L., Costa-Duarte, M. V., et al. 2017, MNRAS, 472, 2808
de Souza, R. S., Maio, U., Biffi, V., & Ciardi, B. 2014, MNRAS, 440, 240
Dewangan, L. K., & Ojha, D. K. 2013, MNRAS, 429, 1386
Draine, B. T., & Li, A. 2007, ApJ, 657, 810
Ducourant, C., Teixeira, R., Krone-Martins, A., et al. 2017, A&A, 597, A90
Elmegreen, B. G., & Scalo, J. 2004, ARA&A, 42, 211
Evans, D. W., Riello, M., De Angeli, F., et al. 2018, A&A, 616, A4
Evans, N., Calvet, N., Cieza, L., et al. 2009a, arXiv:0901.1691
Evans, N. J. I., Dunham, M. M., Jørgensen, J. K., et al. 2009b, ApJS, 181, 321
Everitt, B. S., Landau, S., Leese, M., & Stahl, D. 2001, Cluster Analysis (New York: Wiley)
Fang, M., Hillenbrand, L. A., Kim, J. S., et al. 2020, ApJ, 904, 146
Fazio, G. G., Hora, J. L., Allen, L. E., et al. 2004, ApJS, 154, 10
Feigelson, E. D. 2018, in The Birth of Star Clusters, ed. S. Stahler (Cham: Springer International Publishing), 119
Feigelson, E. D., Townsley, L. K., Broos, P. S., et al. 2013, ApJS, 209, 26
Fisher, R. A. 1922, J. R. Stat. Soc., 85, 87
Forbrich, J., Tappe, A., Robitaille, T., et al. 2010, ApJ, 716, 1453
Furlan, E., Hartmann, L., Calvet, N., et al. 2006, ApJS, 165, 568
Furlan, E., Luhman, K. L., Espaillat, C., et al. 2011, ApJS, 195, 3
Furlan, E., McClure, M., Calvet, N., et al. 2008, ApJS, 176, 184
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1
Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1
Gieles, M., Moeckel, N., & Clarke, C. J. 2012, MNRAS, 426, L11
Gouliermis, D. A. 2018, PASP, 130, 072001
Graham, M. J., Kulkarni, S. R., Bellm, E. C., et al. 2019, PASP, 131, 078001
Greene, T. P., Wilking, B. A., Andre, P., Young, E. T., & Lada, C. J. 1994, ApJ, 434, 614
Greenwell, B. M. 2017, The R Journal, 9, 421
Groenewegen, M. A. T. 2012, A&A, 540, A32
Gutermuth, R. A., & Heyer, M. 2015, AJ, 149, 64
Gutermuth, R. A., Megeath, S. T., Myers, P. C., et al. 2009, ApJS, 184, 18
Gutermuth, R. A., Myers, P. C., Megeath, S. T., et al. 2008, ApJ, 674, 336
Hartmann, L., Megeath, S. T., Allen, L., et al. 2005, ApJ, 629, 881
Harvey, P., Merín, B., Huard, T. L., et al. 2007, ApJ, 663, 1149
Herbig, G. H. 1954, ApJ, 119, 483
Herbst, W., Herbst, D. K., Grossman, E. J., & Weinstein, D. 1994, AJ, 108, 1906
Herczeg, G. J., Kuhn, M. A., Zhou, X., et al. 2019, ApJ, 878, 111
Hilbe, J. M., de Souza, R. S., & Ishida, E. E. O. 2017, Bayesian Models for Astrophysical Data Using R, JAGS, Python, and Stan (Cambridge: Cambridge Univ. Press)
Ho, T. K. 1995, in Proc. Third Int. Conf. on Document Analysis and Recognition, ICDAR 95 (Los Alamitos, CA: IEEE Computer Society), 278
Hodgkin, S. T., Irwin, M. J., Hewett, P. C., & Warren, S. J. 2009, MNRAS, 394, 675
Hodgkin, S. T., Wyrzykowski, L., Blagorodnova, N., & Koposov, S. 2013, RSPTA, 371, 20120239
Hoff, P. 2018, sbgcop: Semiparametric Bayesian Gaussian Copula Estimation and Imputation, https://CRAN.R-project.org/package=sbgcop
Hoff, P. D. 2007, AnApS, 1, 265
Honaker, J., King, G., & Blackwell, M. 2011, J. Stat. Softw., 45, 7
Hotelling, H. 1933, J. Educ. Psychol., 24, 417
Indebetouw, R., Mathis, J. S., Babler, B. L., et al. 2005, ApJ, 619, 931
Ishida, E. E. O., & de Souza, R. S. 2013, MNRAS, 430, 509
Jaeger, B. C., Tierney, N. J., & Simon, N. R. 2020, arXiv:2010.00718
Jarrett, T. H., Cohen, M., Masci, F., et al. 2011, ApJ, 735, 112
Jayasinghe, T., Dixon, D., Povich, M. S., et al. 2019, MNRAS, 488, 1141
Joy, A. H. 1945, ApJ, 102, 168
Joye, W. A., & Mandel, E. 2003, in ASP Conf. Ser. 295, Astronomical Data Analysis Software and Systems XII, ed. H. E. Payne et al. (San Francisco, CA: ASP), 489
Kang, M., Bieging, J. H., Povich, M. S., & Lee, Y. 2009, ApJ, 706, 83
Kauppinen, H., Seppanen, T., & Pietikainen, M. 1995, ITPAM, 17, 201
Kobulnicky, H. A., Babler, B. L., Alexander, M. J., et al. 2013, ApJS, 207, 9
Koenig, X. P., & Leisawitz, D. T. 2014, ApJ, 791, 131
Kounkel, M., & Covey, K. 2019, AJ, 158, 122
Kounkel, M., Covey, K., & Stassun, K. G. 2020, AJ, 160, 279
Krone-Martins, A., Graham, M. J., Stern, D., et al. 2019, arXiv:1912.08977
Kuhn, M. 2015, caret: Classification and Regression Training, Astrophysics Source Code Library, ascl:1505.003
Kuhn, M. A., & Feigelson, E. D. 2019, in Handbook of Mixture Analysis, ed. S. Fruhwirth-Schnatter, G. Celeux, & C. Robert (New York: Chapman and Hall/CRC), 463
Kuhn, M. A., Getman, K. V., & Feigelson, E. D. 2015, ApJ, 802, 60
Kuhn, M. A., Hillenbrand, L. A., Carpenter, J. M., & Menendez, A. R. A. 2020, ApJ, 899, 128
Lada, C. J. 1987, in IAU Symp. 115, Star Forming Regions, ed. M. Peimbert & J. Jugaku (Dordrecht: Reidel), 1
Lawrence, A., Warren, S. J., Almaini, O., et al. 2007, MNRAS, 379, 1599
Leung, H. W., & Bovy, J. 2019, MNRAS, 489, 2079
Lin, C.-A., Kilbinger, M., & Pires, S. 2016, A&A, 593, A88
Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2
Loh, J. M. 2008, ApJ, 681, 726
Lucas, P. W., Hoare, M. G., Longmore, A., et al. 2008, MNRAS, 391, 136
Luhman, K. L. 2018, AJ, 156, 271
Lumsden, S. L., Hoare, M. G., Urquhart, J. S., et al. 2013, ApJS, 208, 11

Lupton, R., Blanton, M. R., Fekete, G., et al. 2004, PASP, 116, 133
Majewski, S., Babler, B., Churchwell, E., et al. 2007, Galactic Structure and Star Formation in Vela-Carina, Spitzer Proposal, ID 40791
Mallick, K. K., Ojha, D. K., Tamura, M., et al. 2015, MNRAS, 447, 2307
Marengo, M., Busso, M., Silvestro, G., Persi, P., & Lagage, P. O. 1999, A&A, 348, 501
Marengo, M., Canil, G., Silvestro, G., et al. 1997, A&A, 322, 924
Marigo, P., Bressan, A., Nanni, A., Girardi, L., & Pumo, M. L. 2013, MNRAS, 434, 488
Marton, G., Ábrahám, P., Szegedi-Elek, E., et al. 2019, MNRAS, 487, 2522
Marton, G., Tóth, L. V., Paladini, R., et al. 2016, MNRAS, 458, 3479
Masci, F. J., Laher, R. R., Rusholme, B., et al. 2019, PASP, 131, 018003
McClure, M. 2009, ApJL, 693, L81
McInnes, L., Healy, J., & Astels, S. 2017, JOSS, 2, 205
McKee, C. F., & Ostriker, E. C. 2007, ARA&A, 45, 565
Melchior, P., & Goulding, A. D. 2018, A&C, 25, 183
Melton, E. 2020, AJ, 159, 200
Minniti, D., Lucas, P. W., Emerson, J. P., et al. 2010, NewA, 15, 433
Morales, E. F. E., & Robitaille, T. P. 2017, A&A, 598, A136
Nelsen, R. B. 2010, An Introduction to Copulas (New York: Springer)
O'Donnell, J. E. 1994, ApJ, 422, 158
Oliveira, I., Pontoppidan, K. M., Merín, B., et al. 2010, ApJ, 714, 778
Oliveira, J. M., van Loon, J. T., Sloan, G. C., et al. 2013, MNRAS, 428, 3001
Oort, J. H. 1927, BAN, 3, 275
Pari, J., & Hora, J. L. 2020, PASP, 132, 054301
Pearson, K. 1894, RSPTA, 185, 71
Pearson, K. 1901, London, Edinburgh, Dublin Phil. Mag. J. Sci., 2, 559
Pfalzner, S., Kaczmarek, T., & Olczak, C. 2012, A&A, 545, A122
Plummer, M. 2017, JAGS: Just Another Gibbs Sampler, https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/
Plummer, M. 2019, rjags: Bayesian Graphical Models using MCMC, https://CRAN.R-project.org/package=rjags
Poncet, P. 2019, modeest: Mode Estimation, https://CRAN.R-project.org/package=modeest
PostgreSQL Global Development Group 2020, PostgreSQL, https://www.postgresql.org
Povich, M. S., Churchwell, E., Bieging, J. H., et al. 2009, ApJ, 696, 1278
Povich, M. S., Kuhn, M. A., Getman, K. V., et al. 2013, ApJS, 209, 31
Povich, M. S., Smith, N., Majewski, S. R., et al. 2011, ApJS, 194, 14
Povich, M. S., Townsley, L. K., Robitaille, T. P., et al. 2016, ApJ, 825, 125
Rebull, L. M., Cody, A. M., Covey, K. R., et al. 2014, AJ, 148, 92
Rebull, L. M., Guieu, S., Stauffer, J. R., et al. 2011, ApJS, 193, 25
Reid, M. J., Menten, K. M., Brunthaler, A., et al. 2019, ApJ, 885, 131
Reiter, M., Marengo, M., Hora, J. L., & Fazio, G. G. 2015, MNRAS, 447, 3909
Rieke, G. H., & Lebofsky, M. J. 1985, ApJ, 288, 618
Ripley, B. D. 1976, J. Appl. Probab., 13, 255
Robitaille, T. P. 2017, A&A, 600, A11
Robitaille, T. P., Meade, M. R., Babler, B. L., et al. 2008, AJ, 136, 2413
Robitaille, T. P., Whitney, B. A., Indebetouw, R., & Wood, K. 2007, ApJS, 169, 328
Robitaille, T. P., Whitney, B. A., Indebetouw, R., Wood, K., & Denzmore, P. 2006, ApJS, 167, 256
Roche, P. F., & Aitken, D. K. 1984, MNRAS, 208, 481
Rokach, L., & Maimon, O. 2014, Data Mining With Decision Trees: Theory and Applications (2nd ed.; Singapore: World Scientific)
Sagi, O., & Rokach, L. 2018, WIREs Data Min. Knowl. Discov., 8, e1249
Samal, M. R., Pandey, A. K., Ojha, D. K., et al. 2010, ApJ, 714, 1015
Samal, M. R., Zavagno, A., Deharveng, L., et al. 2014, A&A, 566, A122
Sasdelli, M., Ishida, E. E. O., Vilalta, R., et al. 2016, MNRAS, 461, 2044
Sato, M., Ichiki, K., & Takeuchi, T. T. 2011, PhRvD, 83, 023501
Schwarz, G. 1978, AnSta, 6, 461
Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. 2016, The R Journal, 8, 289
Shao, Z., Jiang, B. W., Li, A., et al. 2018, MNRAS, 478, 3467
Shu, F. H., Adams, F. C., & Lizano, S. 1987, ARA&A, 25, 23
Simon, J. D., Bolatto, A. D., Whitney, B. A., et al. 2007, ApJ, 669, 327
Simpson, R. J., Povich, M. S., Kendrew, S., et al. 2012, MNRAS, 424, 2442
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Smith, L. C., Lucas, P. W., Kurtev, R., et al. 2018, MNRAS, 474, 1826
Soto, M., Barbá, R., Gunthardt, G., et al. 2013, A&A, 552, A101
Starck, J. L., Donoho, D. L., & Candès, E. J. 2003, A&A, 398, 785
Stauffer, J., Collier Cameron, A., Jardine, M., et al. 2017, AJ, 153, 152
Stern, D., Eisenhardt, P., Gorjian, V., et al. 2005, ApJ, 631, 163
Stolovy, S., Ramirez, S., Arendt, R. G., et al. 2006, JPhCS, 54, 176
Suh, K.-W. 2020, ApJ, 891, 43
Sung, H., Stauffer, J. R., & Bessell, M. S. 2009, AJ, 138, 1116
Taylor, M. B. 2005, in ASP Conf. Ser. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell et al. (San Francisco, CA: ASP), 29
Townsley, L. K., Broos, P. S., Corcoran, M. F., et al. 2011, ApJS, 194, 1
van Breemen, J. M., Min, M., Chiar, J. E., et al. 2011, A&A, 526, A152
van Buuren, S., & Groothuis-Oudshoorn, K. 2011, J. Stat. Softw., 45, 3
van den Bergh, S. 1964, ApJS, 9, 65
van der Schaaf, A., & van Hateren, J. 1996, Vis. Res., 36, 2759
Venter, J. H. 1967, Ann. Math. Stat., 38, 1446
Vioque, M., Oudmaijer, R. D., Schreiner, M., et al. 2020, arXiv:2005.01727
Watson, C., Povich, M. S., Churchwell, E. B., et al. 2008, ApJ, 681, 1341
Werner, M. W., Roellig, T. L., Low, F. J., et al. 2004, ApJS, 154, 1
Whitney, B. A., Robitaille, T. P., Bjorkman, J. E., et al. 2013, ApJS, 207, 30
Williams, J. P., & Cieza, L. A. 2011, ARA&A, 49, 67
Winston, E., Hora, J., Gutermuth, R., & Tolls, V. 2019, ApJ, 880, 9
Winston, E., Hora, J. L., & Tolls, V. 2020, AJ, 160, 68
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Xu, Y., Reid, M., Dame, T., et al. 2016, SciA, 2, e1600878
Xue, M., Jiang, B. W., Gao, J., et al. 2016, ApJS, 224, 23
Yang, Y., Nie, F., Xu, D., et al. 2012, ITPAM, 34, 723
Zari, E., Hashemi, H., Brown, A. G. A., Jardine, K., & de Zeeuw, P. T. 2018, A&A, 620, A172
Zasowski, G., Majewski, S. R., Indebetouw, R., et al. 2009, ApJ, 707, 510
Zavagno, A., Deharveng, L., Comerón, F., et al. 2006, A&A, 446, 171
Zucker, C., Speagle, J. S., Schlafly, E. F., et al. 2020, A&A, 633, A51