The Astrophysical Journal Supplement Series, 269:10 (22pp), 2023 November https://doi.org/10.3847/1538-4365/acdd68

Classification of Chandra X-Ray Sources in Cygnus OB2

Vinay L. Kashyap[1], Mario G. Guarcello[2], Nicholas J. Wright[3], Jeremy J. Drake[1], Ettore Flaccomio[2], Tom L. Aldcroft[1], Juan F. Albacete Colombo[4], Kevin Briggs[5], Francesco Damiani[2], Janet E. Drew[6], Eduardo L. Martin[7], Giusi Micela[2], Tim Naylor[8], and Salvatore Sciortino[2]
[1] Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA; vkashyap@cfa.harvard.edu
[2] INAF—Osservatorio Astronomico di Palermo, Piazza del Parlamento 1, I-90134, Palermo, Italy
[3] Astrophysics Group, Keele University, Keele ST5 5BG, UK
[4] Universidad de Rio Negro, Sede Atlántica—CONICET, Viedma CP8500, Argentina
[5] Hamburger Sternwarte, Germany
[6] University of Hertfordshire, Center for Astrophysics Research, Hatfield AL10 9AB, UK
[7] Instituto de Astrofisica de Canarias, Tenerife, Spain
[8] School of Physics, University of Exeter, Stocker Road, Exeter EX4 4QL, UK
Received 2018 May 27; revised 2022 May 31; accepted 2022 June 1; published 2023 October 25

Abstract

We have devised a predominantly Naive Bayes−based method to classify X-ray sources detected by Chandra in the Cygnus OB2 association into members, foreground objects, and background objects. We employ a variety of X-ray, optical, and infrared characteristics to construct likelihoods using training sets defined by well-measured sources. Combinations of optical photometry from the Sloan Digital Sky Survey (*riz*) and Isaac Newton Telescope Photometric Hα Survey ($r_i i_i$Hα), infrared magnitudes from United Kingdom Infrared Telescope Deep Sky Survey and Two-Micron All Sky Survey (*JHK*), X-ray quantiles and hardness ratios, and estimates of extinction A_v are used to compute the relative probabilities that a given source belongs to one of the classes. Principal component analysis is used to isolate the best axes for separating the classes for the photometric data, and Gaussian component separation is used for X-ray hardness and extinction. Errors in the measurements are accounted for by modeling as Gaussians and integrating over likelihoods approximated as quartic polynomials. We evaluate the accuracy of the classification by inspection and reclassify a number of sources based on infrared magnitudes, the presence of disks, and spectral hardness induced by flaring. We also consider systematic errors due to extinction. Of the 7924 X-ray detections, 5501 have a total of 5597 optical/infrared matches, including 78 with multiple counterparts. We find that ≈6100 objects are likely association members, ≈1400 are background objects, and ≈500 are foreground objects, with an accuracy of 96%, 93%, and 80%, respectively, with an overall classification accuracy of approximately 95%.

Unified Astronomy Thesaurus concepts: Star forming regions (1565); Catalogs (205); Astrostatistics (1882); Astrostatistics techniques (1886); Bayesian statistics (1900); Open star clusters (1160); OB associations (1140); X-ray stars (1823); Standard stars (1564)

Supporting material: machine-readable table

1. Introduction

Nearby star-forming regions provide opportunities for studying the characteristics of young stellar objects and the star formation process itself. Growing realization that exoplanets are very common in the Galaxy has also provided impetus to explore the sites of planet formation and how this process might be affected by astrophysical environments. Star-forming regions in the solar vicinity, such as those found along the Gould Belt within 500 pc or so (e.g., Comeron et al. 1992), have proven fruitful resources for exploitation and form much of the observational basis of our current picture of star and planet formation. However, the Gould Belt represents fairly modest star formation activity, with its clusters typically containing only a few to tens of massive stars of early B or O spectral type. In order to study truly massive sites of star formation, we need to look further afield.

This is the aim of the Chandra Cygnus OB2 Legacy Survey. Cygnus OB2 is one of the largest sites of recent star formation in our Galaxy (Knödlseder 2000; Hanson 2003; Wright & Drake 2009; Wright et al. 2015), hosting tens of O stars and hundreds of OB stars and with an estimated stellar mass of $\sim 3 \times 10^4\, M_\odot$ (e.g., Massey & Thompson 1991; Hanson 2003; Drew et al. 2008; Wright et al. 2010). The Survey comprises a mosaic of Chandra/ACIS-I observations covering the central square degree of the Cygnus OB2 association. X-ray observations and the X-ray source catalog are described in Wright et al. (2023a), while the survey sensitivity and resulting completeness in terms of X-ray luminosity and stellar mass are discussed by Wright et al. (2023b) and Flaccomio et al. (2023).

The X-ray survey aims to exploit the comparative X-ray brightness of young low-mass stars in the T Tauri phase as a means of distinguishing the true association members from a plethora of foreground and background objects in the Galactic plane. A total of 7924 X-ray point sources were detected (Wright et al. 2023a), and while the majority are expected to be in Cygnus OB2 itself, a significant population of interlopers comprising mostly background active galactic nuclei (AGNs)

and active late-type stars in the foreground are expected. In order to be successful, we must solve the problem of separating out these populations and correctly identifying the sources located in the association. This problem is especially challenging for Cygnus OB2, lying as it does behind the Great Cygnus Rift and being subject to significant extinction ($E(B-V)>1.2$)) that varies considerably across the field (e.g., Massey & Thompson 1991; Hanson 2003; Guarcello et al. 2023a; Wright et al. 2015). Guarcello et al. (2023a) have correlated the X-ray catalog with new optical and existing infrared (IR) photometric catalogs, such that we have a wealth of multiwavelength data on which to draw. Here we exploit these data in order to form a statistical basis for classification of survey objects as Cygnus OB2 members or as foreground or background interlopers.

We describe the structure of our approach in Section 2. In Section 3 we describe the various information streams we use in the classification, and in Section 4 we describe the likelihoods we develop based on the data. We discuss the limitations and features of our method and present the classifications in Section 5, and we summarize in Section 6. We provide a detailed description of how measurement error bars are incorporated into the analysis in Appendix A and a discussion of the different choices of extinction and its effect on the classification in Appendix B.

2. Naive Bayes Classification

There are several methods available to cluster multi-dimensional data sets: machine-learning (ML) methods like *k*-means clustering (e.g., Stein et al. 2015), Gaussian process modeling (e.g., Gopalan et al. 2015), neural network (NN) based deep-learning (see Benavente et al. 2017, and references therein) probabilistic methods like multivariate Gaussian clustering (e.g., Stampoulis et al. 2019), and Naive Bayes (e.g., Broos et al. 2011). We have chosen the Naive Bayes approach since it allows us to incorporate relevant information and account for measurement uncertainties in a straightforward manner. It allows us to construct and take advantage of scientifically meaningful likelihoods, generated using training sets in a manner similar to that used with neural nets, but with the advantage of their operational mechanism not being hidden. It is well suited for *wide* data sets (those with large numbers of variables but with few categories, as here, in contrast to *tall* data sets, which have small numbers of observations but large numbers of categories): the manner of construction naturally circumvents the curse of dimensionality when observations in a given category are expected to be correlated. The value of the Naive Bayes approach has been previously demonstrated by its application in the MYStIX survey (Kuhn et al. 2013). Note that, as in typical ML and NN methods, our procedure also relies on setting up and using training sets to define the likelihoods for classification. However, the likelihoods are set up to be easily interpretable as being physically meaningful, and furthermore, unlike ML and NN methods, Bayesian methods are not subject to arbitariness in deciding how many iterations to run, or how many layers to include, or which activation functions to use.

We compute the classification of X-ray sources detected in the Chandra Cygnus OB2 field—whether they are in the foreground, are association members, or are in the background—by computing the likelihood that they belong to each class and choosing the class for which the probability is >0.5. In principle, this excludes weak classifications where the sum of the probabilities for two classes is greater than the class with the highest probability, but as a practical matter that situation is never encountered in our analysis. For a description of the application of Naive Bayes Classification (NBC) to an astronomical survey data set, see Broos et al. (2011). The method hinges on computing the probability that an object is of a given class given the associated data available for that object. Formally, if the class is represented by $\theta = \{\text{foreground, member, background}\}$ and D are various data sets that range from X-ray fluxes to optical magnitudes, we seek to compute the probability of each of the classes θ given the data D, $p(\theta|D)$. By Bayes's theorem,

$$p(\theta|D) \propto p(D|\theta)p(\theta), \quad (1)$$

where $p(D|\theta)$ are the likelihoods, that is, the probability of observing the data D for a given class θ, and $p(\theta)$ codify our prior belief about the relative fractions of the classes.

We describe our choices of the data D, the forms of the likelihoods, and the choices of priors in Section 4.

3. Data

3.1. X-Ray

The X-ray observations that make up the Chandra Cygnus OB2 Legacy Survey consist of a grid of 6×6 Chandra ACIS-I pointings, offset from each other by half the width of the ACIS-I field of view. In addition, two previous Chandra observations (Albacete Colombo et al. 2007; Wright & Drake 2009) are included in the data processed. The data were processed following standard Chandra data reduction procedures, including source detection, photon extraction, and background subtraction, to generate a catalog of 7924 X-ray sources (Wright et al. 2023a). An analysis of the completeness of the observations is presented by Wright et al. (2023b).

For each detected source, we calculate several measures of spectral shape. First, we compute spectral quartiles (Q_{25}, Q_{50}, Q_{75}—the energies corresponding to the 25th, 50th, and 75th cumulative percentiles in the spectrum; see Hong et al. 2004). Based on the combined counts obtained in source and background regions in the soft (S: 0.5−2 keV) and hard (H: 2–7 keV) passbands, we also compute the fractional hardness ratio ($\mathrm{HR} = \frac{\mathrm{H}-\mathrm{S}}{\mathrm{H}+\mathrm{S}}$) and X-ray colors ($C = \log \mathrm{S}/\mathrm{H}$) for each source (Park et al. 2006). These measures are used to supplement the optical and IR photometric measurements (see Section 3.2) to classify sources.

3.2. OIR

The optical−IR (OIR) catalog compiled for the Chandra Cygnus OB2 Legacy Survey counts 328,540 sources across the region of the survey, for which photometry from the following catalogs is available:

1. 65,349 sources with photometry in r, i, z across the central $41' \times 41'$ region (Guarcello et al. 2012) from specific observations with the Optical System for Imaging and low Resolution Integrated Spectroscopy (OSIRIS), mounted on the 10.4 m Gran Telescopio CANARIAS of the Spanish Observatorio del Roque de los Muchachos in La Palma (Cepa et al. 2000);

2. 24,072 sources with photometry in r_I, i_I, Hα bands from the second release of the Isaac Newton Telescope Photometric Hα Survey catalog obtained from observations with the Wide Field Camera (WFC) on the 2.5 m Isaac Newton Telescope (IPHAS; Drew et al. 2005; Barentsen et al. 2014);
3. 27,531 sources from the Sloan Digital Sky Survey (SDSS) catalog (DR8, which covers the Chandra field of view fully; Aihara et al. 2011) with photometry in u, g, r, i, z bands;
4. 273,473 sources with photometry in the *JHK* bands from the United Kingdom Infrared Telescope Deep Sky Survey's (UKIDSS) Galactic Plane Survey catalog (Hewett et al. 2006; Lucas et al. 2008), from observations taken with the Wide Field Camera (WFCAM; Casali et al. 2007) on the United Kingdom InfraRed Telescope (UKIRT), and compiled using a new photometric procedure (King et al. 2013) based on the UKIDSS images (Dye et al. 2018);
5. 43,485 sources with photometry in *JHK* from the Two-Micron All Sky Survey (2MASS) point-source catalog (Cutri et al. 2003);
6. 149,381 sources from the Spitzer/IRAC catalog with photometry in the 3.6, 4.5, 5.8, and 8.0 μm and MIPS 24 μm bands, from the Spitzer Legacy Survey of the Cygnus X region Spitzer (Beerer et al. 2010).

These catalogs have been combined into the OIR catalog in Guarcello et al. (2013), adopting a matching procedure that can be divided into three steps. First, a combined optical catalog was produced by matching the OSIRIS, IPHAS, and SDSS catalogs pairwise for all combinations. Second, an IR catalog was created similarly by matching UKIDSS, 2MASS, and Spitzer data. Each pair of catalogs was combined by using a close-neighbor method with specific matching radii defined in order to minimize the expected number of spurious coincidences and maximize the matched real pairs (see Guarcello et al. 2013, for details). In the last step, these two catalogs were merged into a unique OIR catalog. All the data used here, except those from OSIRIS, are available over the entire area surveyed with Chandra/ACIS-I.

3.3. X-Ray/OIR Matching

The adopted matching procedure between the X-ray and the multiwavelength OIR catalog resulted in 2433 X-ray sources (≈30%) with no OIR counterparts (Guarcello et al. 2023a). While we expect most background X-ray source AGNs to indeed have no OIR counterparts owing to the large extinctions in this direction, some deeply embedded members of Cygnus OB2 are also likely to have no OIR counterparts. We discuss their effect on the prior and classification in Sections 4.1 and 5.2.3, and we consider the possible presence of false negatives in more detail in Appendix C.

4. Likelihoods

A large variety of measurements are available to use to determine the data vector D, and we limit ourselves at the outset to

$$
\begin{aligned}
D = \{ & \\
& r \equiv \text{SDSS } r \text{ band magnitude}, \\
& i \equiv \text{SDSS } i \text{ band magnitude}, \\
& z \equiv \text{SDSS } z \text{ band magnitude}, \\
& \text{H}\alpha \equiv \text{IPHAS H}\alpha \text{ magnitude}, \\
& r_I \equiv \text{IPHAS } r \text{ band magnitude}, \\
& i_I \equiv \text{IPHAS } i \text{ band magnitude}, \\
& J \equiv \text{UKIDSS/2MASS } J \text{ band magnitude}, \\
& H \equiv \text{UKIDSS/2MASS } H \text{ band magnitude}, \\
& K \equiv \text{UKIDSS/2MASS } K \text{ band magnitude}, \\
& Q_{25} \equiv \text{X-ray 25th percentile quantile}, \\
& Q_{50} \equiv \text{X-ray 50th percentile quantile}, \\
& Q_{75} \equiv \text{X-ray 75th percentile quantile}, \\
& \text{HR} \equiv \text{X-ray hardness ratio } \frac{\text{H}-\text{S}}{\text{H}+\text{S}}, \\
& C \equiv \text{X-ray color log(S/H)}, \\
& A_v \equiv \text{absorption estimate} \\
\} & \qquad . \qquad (2)
\end{aligned}
$$

This enables us to use IRAC photometry for verification of the classification and avoids confusion with the analysis of disk-bearing stars (see Section 5.2.1).

A general assumption made in this type of analysis is that the likelihoods are independent of each other. That is, given any two data components, say, D_1, D_2, their individual likelihoods are independent of the other, i.e., $p(D_1|D_2, \theta) = p(D_1|\theta)$ and $P(D_2|D_1, \theta) = p(D_2|\theta)$. This is not strictly true, as trends and correlations in the data components do exist and conditional independence is not assured, and classifications made using such systems will not be optimal. In practice, however, even if full independence is not achieved, the Naive Bayes classifier is highly tolerant of attribute dependences (Domingos & Pazzani 1997). Nevertheless, in order to minimize cross talk between components, we carry out principal component analysis (PCA) on several subgroups of data attributes. Several efforts have been made to use PCA-type methods to reduce the complexity of astronomical spectra (see, e.g., Heavens et al. 2000; Hojnacki et al. 2007; Sasdelli et al. 2016a, 2016b; Davis et al. 2017; Waddell & Gallo 2020; Patil et al. 2022). Here we consider principal components (PCs) of the attributes $\{\boldsymbol{d}\} \equiv \{r, i, z\}$, $\{H, K, J\}$, $\{Q_{25}, Q_{50}, Q_{75}\}$, $\{r_I, i_I, \text{H}\alpha\}$ (see Table 1; A_v and C

Table 1
Projections onto Principal Component Axes

Attributes	Component Used in Classification	Projections[a]
(r, i, z)	2	(+0.1970, +0.0035, −0.2008)
(r, i, z)	3	(−0.0403, +0.0772, −0.0382)
$(r_I, i_I, \text{H}\alpha)$	3	(−0.0483, −0.0114, +0.0588)
$(r_I - i_I, r_I - \text{H}\alpha)$	2	(−0.3640, +0.3640)
(H, K, J)	2	(+0.0494, +0.1931, −0.2463)
(J, K)	2	(+0.2363, −0.2363)
(J, K)	1[b]	(+0.9717, +0.9717)
(Q_{25}, Q_{50}, Q_{75})	1	(+0.9433, +0.9964, +0.9294)

Notes.
[a] The corresponding eigenvalues are the summed squares of the projections.
[b] Used only if likelihood for foreground object > likelihood derived from PC2 [J, K].

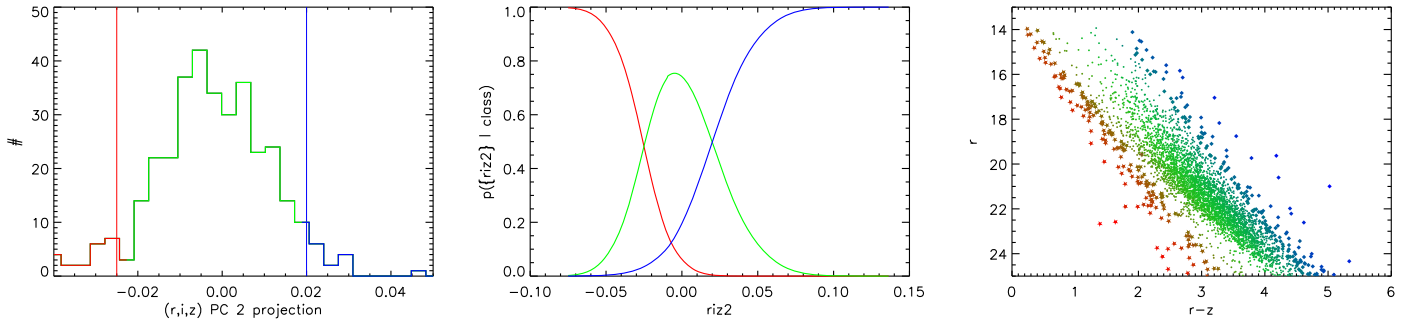

Figure 1. Demonstrating the process of generating the likelihoods of foreground, association member, and background classification using PCs for $\{r, i, z\}$. A subset with well-measured magnitudes is used as a training set, and the histogram of the projections onto the second PC is shown at left (see Table 1), with vertical lines denoting the approximate separation between the different classes, along with the likelihoods constructed from the normalized profiles (middle). A nominal classification to illustrate these likelihoods, carried out using noninformative priors, is also shown as extrapolated to the full data set (right). Each source is color coded as an RGB tuple with the relative probability of it belonging to the foreground (red stars), cluster (green circles), or background (blue diamonds) class.

(the latter based on HR) are used separately and by themselves) and select seven PCs that provide the best discriminatory power (note that in no case does a subgroup contribute more PCs to the classification stream than there are attributes). PCA has been used often in astronomical analysis, though usually as an empirical classification technique (Hojnacki et al. 2007) or a compression technique (Lee et al. 2011; Xu et al. 2014). Here we use it primarily as a scheme to find linear transformations of data subgroups that allows us to efficiently separate the foreground, member, and background sources.

We create vector spaces

$$d_{2i} = \left\{ \frac{X_i - \bar{X}}{\sqrt{\sum_i (X_i - \bar{X})^2}}, \frac{Y_i - \bar{Y}}{\sqrt{\sum_i (Y_i - \bar{Y})^2}} \right\}$$

of doublets, or

$$d_{3i} = \left\{ \frac{X_i - \bar{X}}{\sqrt{\sum_i (X_i - \bar{X})^2}}, \frac{Y_i - \bar{Y}}{\sqrt{\sum_i (Y_i - \bar{Y})^2}}, \frac{Z_i - \bar{Z}}{\sqrt{\sum_i (Z_i - \bar{Z})^2}} \right\}$$

of triplets, where X, Y, Z represent magnitudes or colors for a given object i. By analyzing the correlation matrix, we then obtain PC projections as $v_i^{(k)} = \sum_j c_{jk} d_{ji}$, where the summation is over the dimensions of the subspace, carried out separately for each object (see Table 1). The components k represent successive projections that account for the largest variances in the data, and $\{c_{jk}\}$ represent a rotational transformation that projects the data set onto a new axis. Note that the c_{jk} are not subscripted by the object index i but are dependent on the subsample chosen to compute the PCs. In the following, we drop the subscript i for the sake of brevity unless its absence is ambiguous. The distribution of the projections of the different data points defined by these components is then sifted into separate classes, with the boundaries defined as one-sided Gaussians. A typical assignment, described here for illustrative purposes (see, e.g., the middle panel of Figure 1), is one where the Foreground class is $\propto 1$ for $v^{(k)} < v_{\rm F}$ and decreases as the Gaussian $N(v_{\rm F}, \sigma^2_{\rm F\to M})$ for $v^{(k)} \geqslant v_{\rm F}$, the Members class increases as $N(v_{\rm M}, \sigma^2_{\rm M\to F})$ over the range $[v_{\rm F}, v_{\rm M}]$ and decreases as $N(v_{\rm M}, \sigma^2_{\rm M\to B})$ over the range $[v_{\rm M}, v_{\rm B}]$, and the Background class increases as $N(v_{\rm B}, \sigma^2_{\rm B\to M})$ for $v^{(k)} \leqslant v_{\rm B}$ and is $\propto 1$ for $v^{(k)} > v_{\rm B}$. The intervals and widths are chosen separately for the kth PC based on training set data as described below, and in some cases the directions of the transitions could be reversed. The sum of the components is normalized to 1 at each projected value $v^{(k)}$.

We then construct likelihoods empirically as a mixture of these three smooth components representing the foreground, association member, and background classes. To compute the likelihood for a given object, the data defining it in the vector space of interest are projected onto the component of interest, and the probability of observing those data is defined by the relative values of the likelihood curves for the different classes. This exercise is repeated for different data streams, generating a series of independent likelihoods of obtaining the data given the class. This enables us to expand the likelihood factor in Equation (1) as the product of the likelihoods obtained for each of these independent vectors. The final probability for each class is then the product of these likelihoods for a given class, multiplied by the corresponding prior, and further normalized such that the sum across the classes is 1. The likelihood that an object has the observed data values for a given membership class can then be expanded as

$$\begin{aligned} p(D \mid &\text{membership class}) \\ &\propto p(v^{(2)}(r, i, z)|\text{membership class}) \\ &\times p(v^{(3)}(r, i, z)|\text{membership class}) \\ &\times p(v^{(2)}(r_I, i_I, \mathrm{H}\alpha)|\text{membership class}) \\ &\times p(v^{(2)}(H, K, J)|\text{membership class}) \\ &\times p(v^{(1)}(J, K)|\text{membership class}) \\ &\times p(v^{(2)}(J, K)|\text{membership class}) \\ &\times p(v^{(1)}(Q_{25}, Q_{50}, Q_{75})|\text{membership class}) \\ &\times p(\text{HR}|\text{membership class}) \\ &\times p(A_v|\text{membership class}), \end{aligned} \quad (3)$$

where a component is used if and only if data are available for that source, and each component is defined as the normalized conjoined Gaussians as described above. The probability of membership in a given class, given the data, is then computed by multiplying by the prior and normalizing the sum to 1,

$$p(\text{class}|D) = \frac{p(D|\text{class}) \cdot p(\text{class})}{\sum_{c=\{\mathrm{F,M,B}\}} p(\text{class} = c|D) \cdot p(\text{class} = c).} \quad (4)$$

As a consequence of this process, each object is normalized separately, and if some objects are missing some part of the

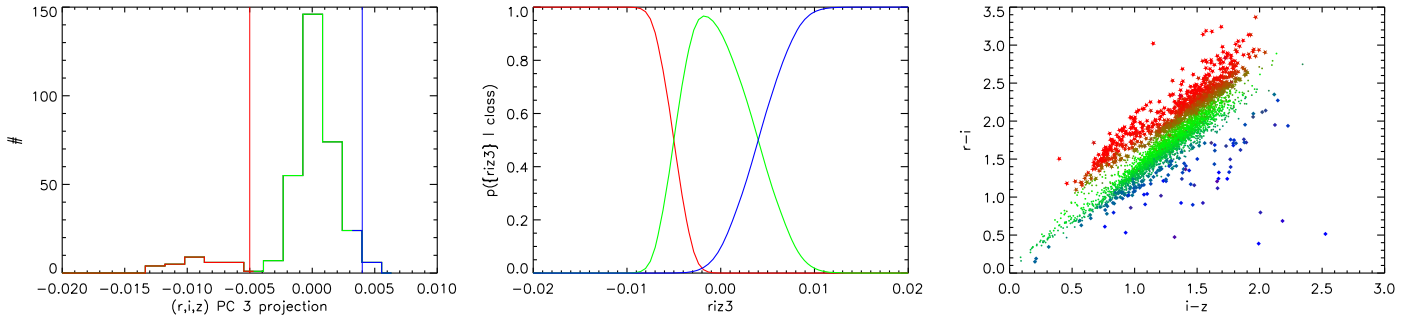

Figure 2. Same as Figure 1, but for the third PC of {*r*, *i*, *z*}.

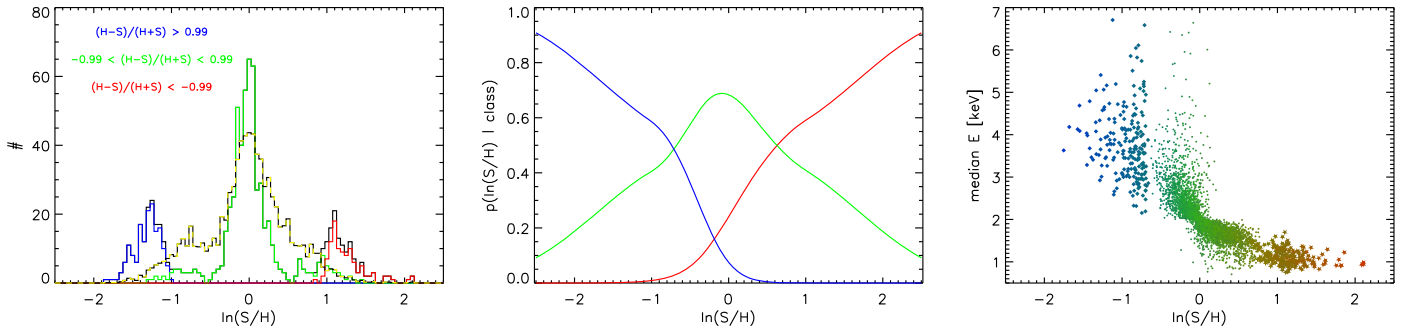

Figure 3. Likelihood generation for X-ray colors. Left: histogram of X-ray colors filtered on hardness ratios for a well-measured subsample, with the histogram for the full sample shown as the dashed black/yellow line normalized to the same number of objects and overplotted. Middle: the likelihoods generated by square-root transformation of the Gaussian components, with the tails modified to be linearly decreasing. Right: class assignments with points color coded as RGB tuples as in Figure 1.

data, those missing parts have no effect on the assigned probabilities.

The ranges and profiles of the Gaussians are modified using expert domain knowledge to segment the training sample into regions where one classification dominates, and thus they generate acceptable classifications for each vector separately. We define the locations and widths of the one-sided Gaussians by using subsets of well-measured points, i.e., observations with small error bars, treated as a training sample, from each vector. Note that this approach tends to pick out the brighter objects, which could result in biases in the likelihoods applied to fainter objects, if the faint population is qualitatively different from the bright population. However, we define class boundaries for the training samples by comparing how the selected classes project back into physically meaningful color–magnitude spaces (see, e.g., the right panels of Figures 1 and 2). Since these boundaries are generally insensitive to the magnitudes of the errors in the data set, the main effect of increased uncertainties is to decrease the contrast between the class boundaries. Thus, all changes made to the class boundaries in PC spaces are rooted to the color–magnitude spaces; the influence of using small error subsamples is minimized. In the one case where we observe the class boundaries shifting with fainter sources (for X-ray hardness ratios; see Section 4.5), we use likelihoods that are designed to be less informative. Our training sample is also highly diversified, with each subset typically having an overlap of $\approx\frac{1}{3}$ with the remaining set (except for the (*J*, *K*) set being a proper subset of the (*H*, *J*, *K*) set). The number of objects chosen for the training set is in the range of a few $\times$ 10^2, compared to 1620 unique objects in the union of the training samples. This variety in the choice of training set population prevents potential biases in any one stream from affecting the overall calculations. Note that we set the bounds independently for each subset, by evaluating the projected distributions in color–magnitude spaces over large scales. We ignore deviations that may be present at smaller scales in color–magnitude diagrams, and thus classifications derived from a single stream are necessarily crude (this point is illustrated in Section 5.1 and Figure 5). We rely on the combination of several streams of data to compute a final classification probability. This is further supplemented by manual inspection and correction to account for special cases that the broad-scale classification misses (see Section 5.2.1).

4.1. A Priori Expectations

Bayesian analysis requires that priors be defined in order to convert the likelihoods, which are probabilities defined as functions of the data, to posterior probabilities, which are probabilities defined for the parameters or classes of interest. In the absence of any prior information, a flat distribution is the best option, which in our case corresponds to $p(\text{foreground}) = p(\text{member}) = p(\text{background}) = \frac{1}{3}$. This is the choice used in order to demonstrate the effect of the adopted likelihood function for each of the component streams (Figures 1–5 and D1–D6). While this choice is appropriate to illustrate how the likelihood maps to variables of interest, for the combined analysis we can make a more informed choice. We define informative priors by estimating the number of X-ray sources that may be obtained from background quasars and from foreground Galactic sources.

We estimate the number of AGNs expected in the X-ray sample by assuming that their numbers are distributed as a broken power law with indices $\beta_1 = 1.34$ for $f_X \leqslant 8.1 \times 10^{-15}$ erg cm^{-2} s^{-1} and $\beta_2 = 2.23$ at higher f_X (see Equation (5) of Lehmer et al. 2012). Further assuming a nominal power-law

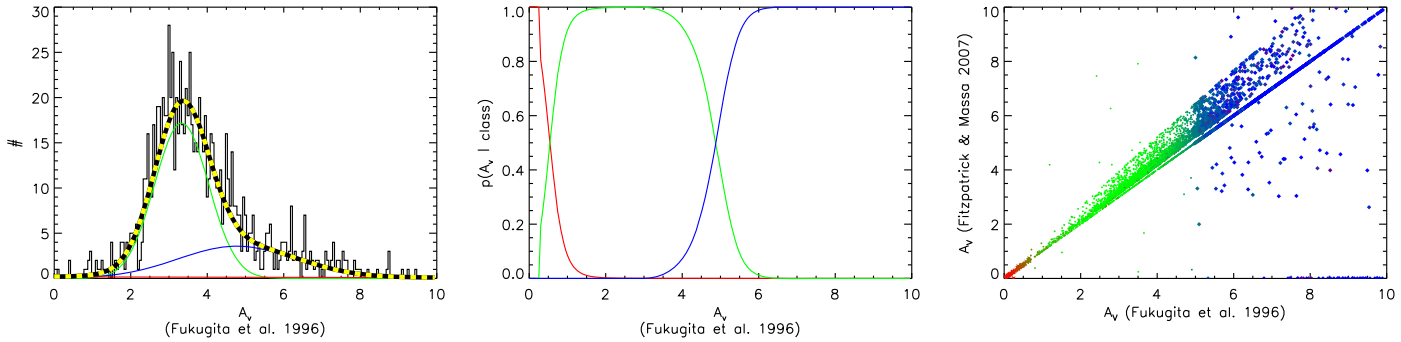

Figure 4. Separating Cygnus OB2 sources using extinction, A_v (based on F96; see Figure B1 for the FM case). The observed histogram (left) is modeled as a mixture of three Gaussians, with the components that predominate at low, medium, and high A_v assumed to represent the foreground (red), member (green), and background (blue) objects, respectively. These Gaussian components, after removal of the population size effect and renormalization, are identified with the corresponding likelihoods (middle). A scatter plot showing the difference in estimated A_v when using the extinction laws of F96 and FM is shown in the right panel, with the points color coded as RGB tuples as in Figure 1.

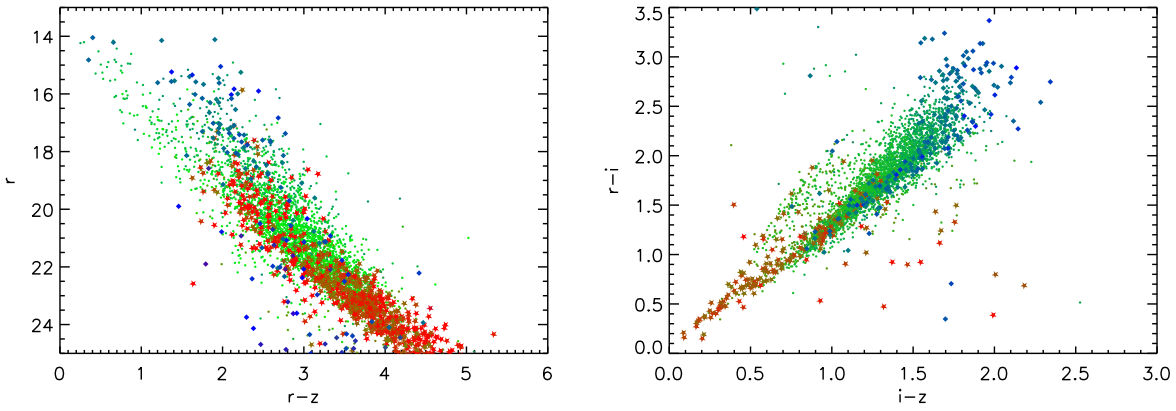

Figure 5. Demonstrating the necessity of using multiple channels of data to derive class memberships. The left panel shows r vs. $r-z$ as in the right panel of Figure 1, but using the third PC as in Figure 2. The right panel shows the reverse case, i.e., $r-i$ vs. $i-z$ as in Figure 2, but using class probabilities derived using the second PC as in Figure 1.

spectrum with index $\Gamma = 1.8$ to determine a flux-to-counts conversion factor, we find that for the specific exposure map of the Chandra Cygnus OB2 survey $\approx$ 1200–1800 sources would be present with net counts >5 for $N_{\rm H} = (1-4) \times 10^{22}$ cm^{-2}. The sensitivity of the survey varies across the field, as does the diffuse background (Albacete-Colombo et al. 2023) and the absorbing column density, with lower detection thresholds or lower $N_{\rm H}$ potentially yielding more background X-ray sources. The variation in $N_{\rm H}$ over the field can be evaluated by considering the range of A_v estimated for the various objects in our data set (see Section 4.6). The A_v ranges from $\approx$0.1 to $>$10, suggesting that $N_{\rm H}$ varies from $\approx 10^{20}$ cm^{-2} to a few $\times$ 10^{22} cm^{-2} (cf. Predehl & Schmitt 1995).

We estimate the approximate number of foreground stars using a dynamical model of the Galaxy (TRILEGAL v1.6; Girardi et al. 2012),[9] which produces a representation of the stellar population along the line of sight toward Cygnus OB2 assuming an exponential thin disk, a squared hyperbolic secant thick disk, an oblate spheroid halo, and a triaxial bulge. The number of expected X-ray detections varies by over an order of magnitude for different assumptions about the X-ray luminosity function of the field stars and the local limiting sensitivity. For assumed fixed $L_{\rm X} = 10^{27}$, 10^{28}, and 10^{29} erg s^{-1}, for a limiting sensitivity of 10^{-15} erg s^{-1} cm^{-2}, we expect $\approx$30, 800, and 10,000 X-ray sources, respectively, ranging from there being almost no foreground sources to accounting for essentially all the detected sources. It is plausible, however, for realistic luminosity functions with median $L_{\rm X} \sim 10^{28}$ erg s^{-1} to produce 500–1500 X-ray detections.

Thus, it is reasonable to assume that of the $\approx$8000 sources considered, about 2500 should be some combination of foreground and background sources. We therefore adopt prior probabilities on the classes

$$p(\text{foreground}) = p(\text{background}) = 0.15,$$

that is, if no information were available, the probability that an X-ray source in the Cygnus OB2 field of view is a member of the association is 0.7. The specific values of the prior are important only when the data are not informative, and when they are, the posterior estimates are driven by the likelihoods. We also point out that it is theoretically unjustifiable to change the priors after the analysis, e.g., by iteratively adjusting them until the posterior counts match the adopted fractions.[10] Nevertheless, we have explored the sensitivity of the class assignments to the prior by considering how many foreground objects are "lost" when $p(\text{foreground}) = 0.05$ and how many background objects are "gained" when $=p(\text{background}) = 0.2$, commensurate with the fractions that are found upon carrying

[9] http://stev.oapd.inaf.it/cgi-bin/trilegal

[10] As it happens, we find (see Table 2) the relative fractions of the foreground, member, and background objects to be $\approx$6%, 77%, and 17%, respectively.

out the classification procedure (see Section 5.2). We find that the changes are small: in the former case the number of foreground objects decreases by ≈20 (3%), and in the latter case the number of background objects increases by ≈100 (5%), similar to the magnitude of the systematic uncertainties present in the process (see Section 5.2.1).

The X-ray properties of sources matched with the OIR catalog differ substantially from those with no matches (Guarcello et al. 2023a). In particular, the sources with no matches exhibit a bimodal distribution in Q_{50}, with nearly half of the population exhibiting $Q_{50} > 3$ keV. It is thus reasonable to consider whether different values of $=p$(background) should be used for the samples of matched versus unmatched sources. However, as noted above, the precise values of the priors are not a significant factor in the classification, so for simplicity of calculation we maintain the same prior for the whole sample. Nevertheless, because sources with no OIR counterparts are classified entirely through their X-ray properties alone, we flag them in the final catalog (see Section 5.2.3) as such, and we also indicate how robust that classification is.

For each object, the likelihood that it belongs to a given class is computed independently for each data stream (the PCs in Table 1, and the mixture components of A_v and HR) while accounting for measurement errors (see Appendix A). The product of these likelihoods and the priors is then computed for each class and renormalized such that the sum adds up to 1. The probability values span a continuum between [0, 1], but for the sake of specificity we assign a specific class to each object as that which has the highest of the three probabilities. These assigned classes are then reviewed, and those that are clearly misclassified are reassigned (see Section 5.2.1).

4.2. r, i, z

The usual way to sift sources into different classes is to display them on color–color diagrams and identify regions where there is a higher propensity for members of one class to appear. When the $\{r, i, z\}$ triplet is available, for example, foreground objects stand out along the leftward edge in the r versus $(r - z)$ diagram and along the upper edge of the $(r - i)$ versus $(i - z)$ diagram. We extract this information using the second and third PCs of a subset of $\{r, i, z\}$ data points. Notice that we do not use the first component, even though it accounts for the largest fraction of the variance in the data set, as it is not informative for the purpose of separating the different classes. We apply similar judgments for other subgroups and exclude all cases where the projections of the classes overlap significantly or are not easily modeled (e.g., $\frac{f_X}{f_{opt}}$). For the training sample, we compile a list of 200 sources that have the smallest error bars in each band, which results in a total of 348 unique sources.[11] The projected components are shown as a histogram in the left panels of Figures 1 and 2, colored according to which range is preferentially dominated by members of which class (red denotes foreground, green denotes association members, and blue denotes background). Note that at this stage the sources are sifted into different classifications by construction, that is, according to our expectation of how they are likely to be distributed among the different classes. The boundary between the classes is not sharp, and the transition from one dominant class to another is assumed to occur smoothly. The distributions of the number of objects, projected along the PC, for each class are assumed to be approximated as Gaussians and smoothly varying. They are then normalized such that each component has a maximum of 1, and they are further renormalized at each point along the PC such that their sum adds up to 1. The likelihoods thus obtained are shown in the middle panels with the same color scheme. These likelihoods are then applied to the full $\{r, i, z\}$ data set (that is, not just the training sample), incorporating uncertainties (see Appendix A). The resulting classifications (using flat priors set to one-third for each class), constructed using *only this data stream and no others*, are shown in the right panels, with color hues ranging from red (denoting foreground) to green (denoting members) to blue (denoting background). We discuss the necessity of using multiple components in Section 5.1.

4.3. J, H, K

For the combination of $\{J, H, K\}$, we first consider the color–magnitude diagram J versus $(J - K)$. For this, we carry out PCA on a sample of 287 objects with the best-measured $\{J, K\}$. Ultimately, we primarily use the second component, but we use the higher of the computed foreground likelihoods for foreground objects, since a bright J strongly suggests that the object is in the foreground.

We also carry out a PCA on a subset of 332 of the best-measured triples $\{J, H, K\}$, using the second component that separates foreground and background objects in $(J - H)$ versus $(H - K)$ diagrams (Figure D3).

4.4. r_I, i_I, $H\alpha$

We consider the triple $\{r_I, i_I, \mathrm{H}\alpha\}$ as a group since these are measured with IPHAS. We compute PCs for both the triple and the paired colors $(r_I - \mathrm{H}\alpha)$ versus $(r_I - i_I)$ using a training sample of 212 well-measured objects. The third component of the former (Figure D5) and the second component of the latter (Figure D4) sift the region into similar groupings, but with slight differences that indicate a complex projection of components. We use both components. We note that (r_I, r) and (i_I, i) are strongly correlated (Pearson's $\rho \approx 1$) but display complex behavior in their errors (e.g., the ratio of the errors is correlated with $r - i$ with $\rho \approx 0.5$), and thus they are able to provide additional discriminatory power to the PCs used in Section 4.2.

4.5. X-Ray Spectral Shape

ACIS spectra typically have >800 usable pulse height channels, but their shapes are effectively characterized by counts in a small number (∼5) of passbands (similar to the number of free parameters in thermal spectra usually used while fitting the X-ray spectra of weak sources; e.g., Flaccomio et al. 2023). We use three measures of quantiles (defining the energies that include 25%, 50%, and 75% of all the observed counts within the source regions) and two measures of hardness ratios (extremes of fractional hardness $\mathrm{HR} = \frac{\mathrm{H} - \mathrm{S}}{\mathrm{H} + \mathrm{S}}$ to motivate a model that is applied to color $C = \ln \mathrm{S}/\mathrm{H})$ as proxies to characterize the dispersion that describes X-ray spectral shapes.

We use the first PC of the analysis of the best-measured sample of 357 objects of the triple $\{Q_{25}, Q_{50}, Q_{75}\}$, as it groups typically unabsorbed thermally emitting foreground objects as

[11] We employ the same method to compile training samples in Sections 4.3, 4.4, and 4.5, where PCA is used for likelihood generation.

having a soft spectrum and background objects that are likely to be power-law sources and heavily absobed as having a hard spectrum (Figure D6). For detailed discussions of the spectra, see Flaccomio et al. (2023).

Though there is overlap in the information codified by the quantiles and hardness ratios (and A_v; see Section 4.6), comparisons of typical model grids show that they encode the spectral shape information differently. Furthermore, there is a dearth of background objects in the training sample made using the quantile data (see, e.g., Figure D6), which makes it necessary to include a broader measure. We thus complement the separation obtained from the quantiles with color $C = \ln S/H$, computed using Bayesian estimation of hardness ratios (BEHR; Park et al. 2006). Because C is one-dimensional and the counts in individual bands are sensitive to normalization, we cannot use PCA to determine the optimal axes as above. Instead, we model it as a mixture of Gaussians. We build a weakly informative likelihood model by considering the behavior of the distribution of C in extreme cases and extrapolating the model to more ambiguous cases. Considering only the extremely soft ($HR \leqslant -0.99$; likely foreground) and extremely hard ($HR \geqslant +0.99$; likely background) sources, we see that the distribution of the posterior modes of the colors C of such sources splits into two distinct and well-separated components (see left panel of Figure 3, red and blue curves, respectively). In contrast, a sample of 823 of the best-measured sources with $-0.99 < HR < +0.99$ (likely members) occupy a third component in between the extremes (green curve). The presence of such distinct components suggests a simple parameterization of the likelihood function centered on each of the components. However, this measure is subject to imperfect domain adaptation: as larger samples are examined, the outer peaks move inward, eventually smoothing out the trimodal structure, suggesting that the extreme values do not form a high-fidelity training set (dashed black histogram). We therefore seek to construct a measure that accounts for the gross separation without addressing the detailed shape or the changes in the distribution of colors as samples with larger uncertainties are included, i.e., we seek to avoid overfitting to the distributions by choosing a deliberately imprecise scheme. We thus choose Gaussians centered at $C = (-1, 0, +1)$, with widths corresponding to the standard deviation of the subsets, to describe the different classes, further dilute the sharp divisions between the components by using a square-root transformation on the Gaussians to reduce the contrast, and enforce a linear decline in the tails of the central component to avoid numerical instability at large deviations. Since we expect there to be a great deal of mixing between the different classes owing to the large dynamic ranges present in X-ray luminosities, these corrections avoid the X-ray colors being the dominant contributor to the class likelihood assignments. The result of these modifications is shown in the middle panel of Figure 3 (which shows that the transitions between classes are not sharp but the extremes are indeed unambiguously assigned), and the corresponding class assignments to the full data set are shown in the right panel.

4.6. Extinction, A_v

Low extinction is a strong diagnostic of whether a source is in the foreground of the cluster or not, and conversely, high extinction suggests that an object is in the background. Because of this high sensitivity of class assignment to extinction, we compute A_v from optical and IR data and incorporate it as an additional data stream. Note that while this information is partially included in the spectral shape data extracted from the X-ray colors (see Section 4.5), extinction affects spectra at different temperatures differently, and an independently generated data stream can provide additional information to define the classification. However, because extinction is not spatially uniform across the cluster, a simple cut across A_v is not a good method to assign classes. As in the case of X-ray color C (see Section 4.5), A_v is also a one-dimensional data stream, and we model a subset of 1003 well-estimated A_v as a mixture of three Gaussians, each directly representing the three classes of interest (see Figures 4 and B1).

Our A_v values are estimated based on J, H, r, i, z, as well as 2MASS and UKIDSS magnitudes. (Note that due to the nonlinear interplay between the optical and IR magnitudes, A_v forms another complementary combination of these attributes that were used in linear combinations with PCA above.) Individual extinctions are calculated using one of two methods. The main method consists of calculating the displacement of stars in the $r-i$ versus $i-z$ diagram along the extinction vector from the 3.5 Myr isochrone. For those stars with unreliable or absent optical photometry (see Guarcello et al. 2012), individual extinctions are calculated from 2MASS and UKIDSS photometry using the Near-Infrared Color Excess Revisited (NICER) algorithm (Lombardi & Alves 2001). Errors are calculated propagating the photometric uncertainties in these two colors. We have considered both the extinction vector in the SDSS bands from Fukugita et al. (1996 hereafter F96) and the more recent one from Fitzpatrick et al. (2007, hereafter FM). Individual extinctions obtained with these two laws typically differ by $\approx 7\% \pm 10\%$, with the A_v distribution obtained with the F96 law peaking at about 5.5 mag and that with the FM law at about 4 mag. Our classification scheme primarily relies on the F96 extinction law. We discuss the difference using the different extinction laws makes on our classifications in Appendix B.

5. Discussion

5.1. Validation

A question that arises is whether it is necessary to go through all these steps to obtain a classification, instead of placing cuts in a color–color diagram, say, for $r-i$ versus $(i-z)$, as in Figure 2. The necessity of a comprehensive analysis is best demonstrated by example, as in Figure 5. Here the probability assignments made using each of the two $\{r, i, z\}$ PCs (see Section 4.2) are applied to the other, that is, the classification from the second component, trained using r versus $(r-z)$, is shown plotted with $(r-i)$ versus $(i-z)$, and the classification from the third component, trained using $(r-i)$ versus $(i-z)$, is shown plotted with r versus $(r-z)$. These plots show that there is considerable mixing evident between the different classes when viewed with different projections. It is therefore crucial that a combination of data streams be used, encompassing different pieces of information, in order to obtain reliable class assignments.

Standard methods of validation like cross-validation using k-fold training or bootstrapping on test samples are impractical for this analysis, since our training sample is small, and due to data incompleteness, the overlap between components in the training sample is also small. Furthermore, since any single

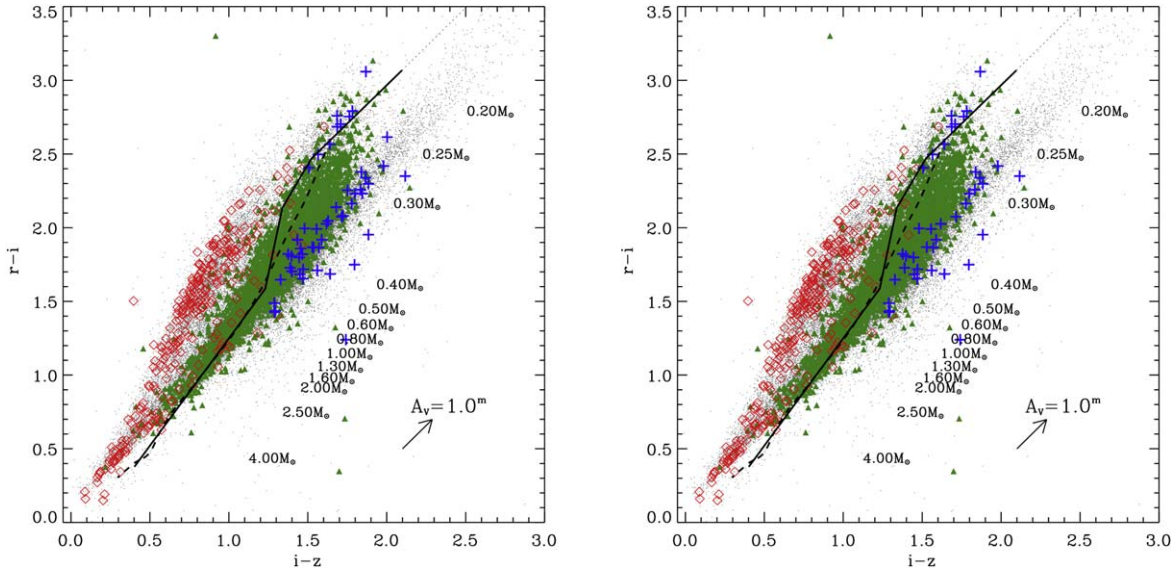

Figure 6. $r-i$ vs. $i-z$ diagrams of all the sources with good OSIRIS or SDSS photometry. The solid line is the 2.5 Myr isochrone from Siess et al. (2000) after applying the F96 transformation between the *UBVRI* and the $u_I\ g_I\ r_I\ i_I\ z_I$ photometric systems, and the dashed line is from the MIST database. The extinction vector is obtained from O'Donnell (1994). The corresponding masses are indicated with a horizontal displacement from the isochrone. The classifications are according to the automated NBC method (left) and after manual reclassification (right) and show foreground objects (red diamonds), members (green circles), and background objects (blue crosses).

Table 2
Post Hoc Reclassifications

Original	Reclassified			Subtotal Original
	Foreground	Member	Background	
Foreground	420	199	0	619
Member	71	5861	2	5934
Background	0	109	1358	1467
Subtotal reclassified	491	6169	1360	8020

Note. Number of sources of each class, with Naive Bayes approach and after reclassification, showing how many in each class change classifications.

data stream can predict classes that are inconsistent with the information present in the rest of the data, cross-validation of the process by leaving out one component is not a workable method of testing the results. Instead, we manually screened the classifications in various color–magnitude spaces, including some not used during the classification process, such as IRAC colors like $[4.5]-[5.8]$ versus $[5.8]-[8.0]$, and other color projections like g and $g-r$ versus $r-i$, to expose outliers, which we reassigned where necessary as explained in more detail below (see Section 5.2.1). Based on this manual reclassification (see Table 2), as well as comparing the effects of different extinction laws on the classification (see Appendix B), we estimate that the error rate in our classification is ≈5%.

5.2. Classification

5.2.1. Reclassification

Our classification can be affected by some stellar properties. For instance, during X-ray flares the stellar X-ray spectrum becomes harder, which mimics the expected X-ray energy quantiles typical of background sources. Futhermore, compared to naked photospheres, stars with circumstellar disks have red IR colors that can be confused with background sources at large extinction, or they may have blue optical colors typical of foreground stars because of the accretion process or the presence of scattered light (Guarcello et al. 2010). Besides, the adopted NBC classification scheme is expected to have issues for very bright stars, whose photometry is typically saturated. For these reasons, the results of the automated classification described above (Section 4) have been retested using OIR color–color and color–magnitude diagrams and the X-ray energy quantiles, which provides further leverage to separate the three classes. We also forced the classification of known stars with disks as selected by Guarcello et al. (2013, 2023b) to be association members. Overall, a total of 381 objects changed classification (see Table 2), including 49 (out of 833) of those that had originally been included in the training sets with optical and IR colors and magnitudes (and an additional 208 of 1796 of those used only with X-ray quantiles, color, and A_v analyses). This yields an overall error rate in the automated classification of ≈6% from among those X-ray sources with optical matches, and ≈5% for all objects (note that this includes cases with multiple matches). The nominal accuracy of classification of cluster membership is ≈96%. Because of the large size of this subset (6169 objects), the majority of the reclassification involves cluster members. This class gains 199 and loses 71 to the foreground class and gains 109 and loses 2 to the background class. Because of the much smaller sizes of the foreground (491 objects) and background (1360 objects) classes, the nominal accuracy of their assignment is dominated by the cluster membership accuracy, at ≈74% and ≈92%, respectively.

We compare the differences between the automated classification and manual reclassification in Figure 6 and Figures E1–E5. In each panel of the figures, the small gray points mark the colors or magnitudes of the sources in the survey area with good-quality photometry and errors smaller than 0.15 for colors and 0.1 for magnitudes. In each figure, the X-ray sources classified according to the automated NBC

scheme are shown in the left panel, and the revised version based on manual inspection is shown in the right panel. The diagrams shown in Figures 6, E1, and E2 allow us to separate the sources with low (foreground), intermediate (members), and high (background) extinction and thus select stars that are likely wrongly classified by the NBC method, such as the foreground and background stars that populate the same locus as the candidate members in these diagrams. Even after the revision, a small number of stars appear to lie on the "wrong" part of these diagrams, such as the background sources between the $E_{B-V}=0$ and $E_{B-V}=1$ main sequences in the left panel of Figure E1. These are cases where no clear indications come from the diagrams and the X-ray photon energy quantiles, for instance, because of mismatches between counterparts of different catalogs. In these cases, we have kept the classification from the NBC method. It is also important to note that the position of candidate members with disks in these diagrams can be affected by accretion and/or scattered light, for instance, increasing the $r_I - \mathrm{H}\alpha$ color or decreasing the $g-r$ color, thus pushing the sources above the $E_{B-V}=0$ main sequence in Figure E1 or below the $A_v=0$ isochrone in Figure E2.

The $H-K$ versus $J-H$ diagrams shown in Figure E3 allow us to separate sources with different extinction, and thus stars in the foreground, those in the background, and those within the association. In this diagram it is also clear that the NBC method fails to classify as members very bright stars that are clearly massive members of Cygnus OB2 according to their photometric properties, soft X-ray spectra, and in some cases existing spectral classification. Furthermore, these stars are clearly clustered in the various subclusters of Cygnus OB2.

Given that extinction does not seriously affect the Spitzer/IRAC colors, the diagrams of IRAC colors in Figure E4 are not useful for separating stars affected by different extinction. However, together with the set of diagrams used in Guarcello et al. (2013), they are useful for selecting candidate stars with disks, typically populating the regions corresponding to colors larger than 0.5^m, and background galaxies populating loci that are empirically defined by various authors. Recall that we enforced the classification as "members" of all the candidate stars with disks selected by Guarcello et al. (2013).

The spatial distributions of the X-ray sources classified before and after the revision are shown in Figure E5. Comparing the two panels, it is evident that several stars whose classification has been turned into "members" are indeed clustered in the center of the region or in other subclusters. In both panels regions with dense nebulosity can be identified by the contours of the IRAC [8.0] μm continuum emission levels. Several X-ray sources classified as "background" objects fall within high-extinction regions such as within DR 18, at approximately $\alpha=308.7$, $\delta=41.2$. These sources are likely to be embedded young stars detected in X-rays, but their classification has not been changed owing to the lack of a good-quality OIR counterpart.

5.2.2. Impact of Gaia

Unlike our method, which is based on a probabilistic weighing of proxy information, precise measurements of distances to matched OIR sources can fix the classification exactly. The Gaia DR2 catalog has parallaxes for over a billion stars (Gaia Collaboration et al. 2018) and could be a source of such distance measurements. Of the X-ray and matched OIR sources, 1128 (60 foreground, 749 members, 319 background; ≈14% of the catalog) have probable counterparts in Gaia DR2 based on overlaps of position error circles. Of these, choosing the nearest Gaia counterpart, 329 (52 foreground, 252 members, 25 background) have nonzero parallaxes, and only 140 (47 foreground, 91 members, 2 background) have well-measured distances (parallaxes measured at better than 3σ). The change in the mix of classification (from ≈6% foreground to ≈35% foreground) is consistent with nearer objects having better distance measurements. It is reasonable to expect that foreground stars will be the ones best characterized by Gaia, and any misclassifications in the probabilistic scheme will be dominated by these sources. Indeed, we find that 41 sources are apparently misclassified, of which 31 are classified as members but are at distances <1.2 kpc, and are plausibly foreground stars. In addition, one Gaia-matched source classified as background is likely a member (distance ≈1.3 kpc), and four sources classified as foreground stars are at distances >1.2 kpc and could be considered association members or background sources. This is consistent with the error rate of the classification scheme (see Tables 2, B1, and B2). Thus, the Gaia DR2 release has a negligible effect on both our method and our results. Here we report the potential changes in classification due to Gaia parallax measurements as part of the catalog (see Section 5.2.3), but otherwise we do not incorporate it in our procedure, and we defer a more detailed analysis that looks at the individual matches and an assessment of the systematic uncertainties in the Gaia catalog at large distances and large A_v to a later work.

5.2.3. Catalog

The final classifications are shown in Figure 7 using suitable color–color and color–magnitude diagrams where the foreground, background, and association sources can be distinguished. In each panel the small gray points mark those sources in the survey area with good-quality photometry and errors smaller than 0.15 in the colors and 0.1 in magnitudes. (The criteria for "good photometry data" in these catalogs are described in Guarcello et al. 2012, 2013.) The OIR+X-ray sources are marked with colors coding their final classification. We also mark a possible locus of candidate members of Cygnus OB2 using the 2.5 Myr isochrone from Siess et al. (2000) (converted into the $u_I\,g_I\,r_I\,i_I\,z_I$ photometric system), the MIST isochrones (Choi et al. 2016; Dotter 2016), and in the IPHAS color–color diagram the ZAMS defined by Drew et al. (2005), assuming the distance of 1400 pc found by Rygl et al. (2012).

In all the diagrams, with the exception of the IRAC color–color diagram, given the direction of the reddening vector, the foreground stellar population can be separated from the stars at the distance of the association or those farther away. This is particularly evident in the IPHAS color–color diagram and in the SDSS diagrams, the latter mainly for low-mass stars. The vast majority of the sources sorted as "foreground stars" populate these low-extinction loci. The background population is more evident in the *JHK* diagram, mainly in a high-extinction locus (faint and red stars). In the IRAC color–color diagram some of the background objects lie in the locus where stars with disks are typically found. These stars have been excluded from the list of stars with disks and classified as "background contaminants" by Guarcello et al. (2013). The cluster locus lies between the foreground and background

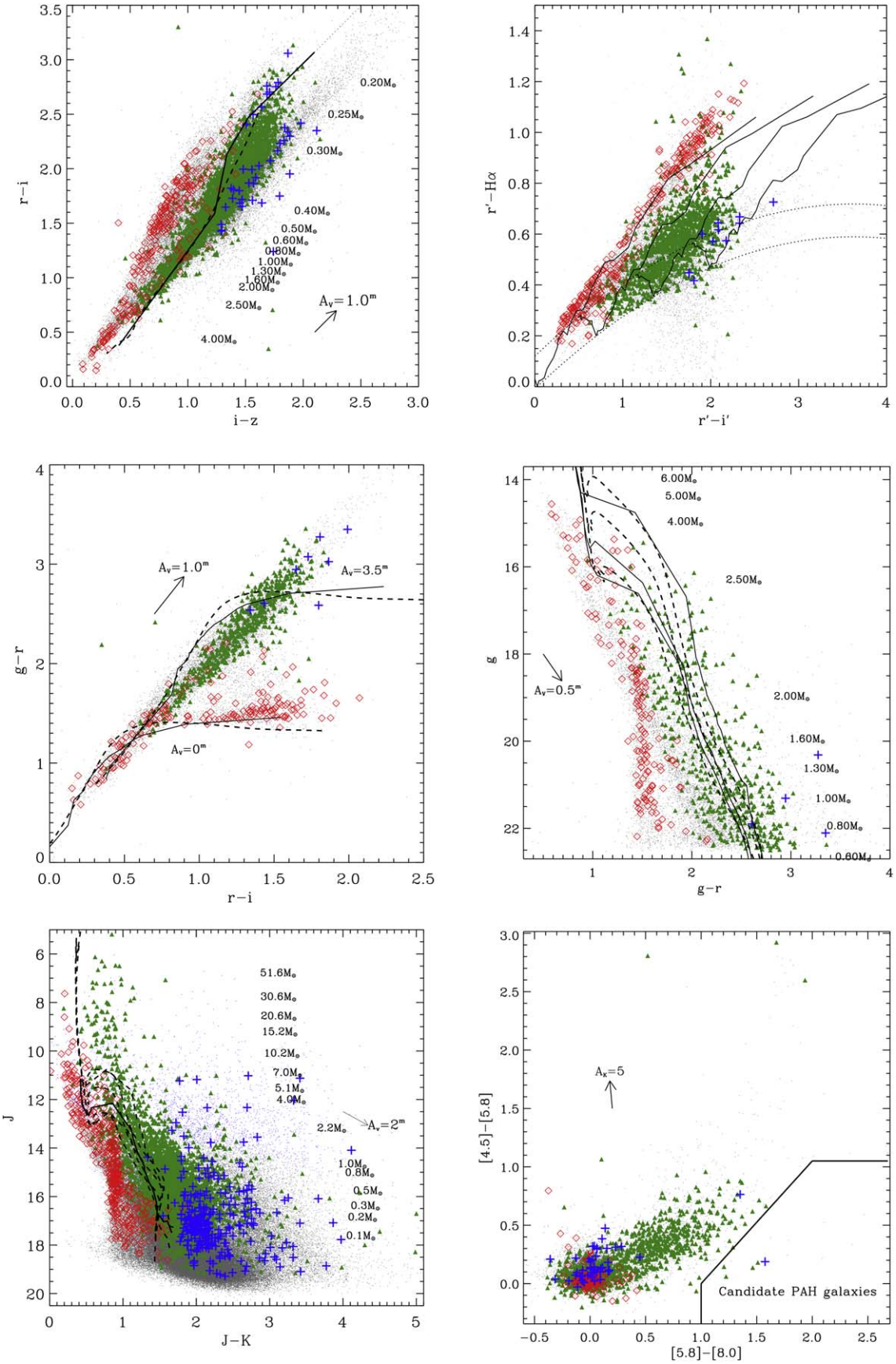

Figure 7. Set of diagrams showing colors and magnitudes of all the sources of our OIR catalog with good photometry in the involved bands (gray circles), candidate OIR+X-ray foreground sources (red diamonds), Cygnus OB2 members (green circles), and background sources (blue crosses). We also show the extinction vectors and the 2.5 Myr isochrone from Siess et al. (2000; solid lines) and the MIST database (dashed line), assuming a distance of 1400 pc and $A_V = 3.5^m$ with corresponding mass values labeled with a horizontal displacement. In the top right panel the solid lines are instead ZAMS at increasing E_{B-V} (0, 1, 2, 3).

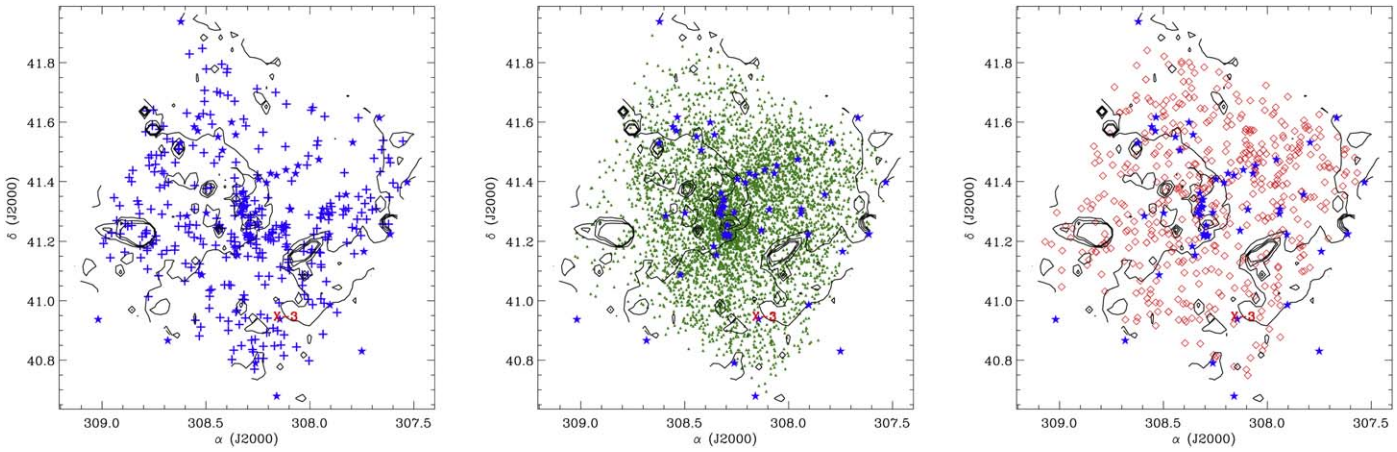

Figure 8. Spatial distribution of background (left), association member (middle), and foreground (right) sources with OIR matches. The contours mark continuum emission levels at [8.0] μm Spitzer band, denoting the presence of dust and high A_v. The position of Cyg X-3 is marked in red. The filled blue stars denote O stars.

Table 3
Catalog of Classifications[a]

CXO ID	OIR Match	$p_{\rm foreground}$	$p_{\rm member}$	$p_{\rm background}$	Classification
1	Yes	0.00	1.00	0.00	member
4	No	0.00	0.03	0.97	background
5	Yes	0.00	0.99	0.01	member[b]
6	Yes	0.00	0.00	1.00	background[b]
9	Yes	0.00	1.00	0.00	foreground[c]
11	Yes	0.13	0.87	0.00	member[d]
15	Yes	0.09	0.91	0.00	foreground[c d]
17	Yes	0.00	0.60	0.40	member[d b])
37	Multiple	1.00	0.00	0.00	foreground
	Multiple	0.00	1.00	0.00	member[b]
81	Yes	1.00	0.00	0.00	member[c e]
227	Yes	1.00	0.00	0.00	foreground[e]

Notes.
[a] Representative subset, for demonstration. Full catalog available online.
[b] Classification changed by inspection.
[c] Used as part of training set.
[d] Matched to a Gaia DR2 source within position error.
[e] Final classification inconsistent with Gaia distance.

(This table is available in its entirety in machine-readable form.)

population in most of the diagrams. Mainly in the IPHAS color–color diagram it lies within the $1 \leqslant E_{B-V} \leqslant 3$ range and in the *JHK* diagram within the $3 \leqslant A_V \leqslant 10$ range. Of those objects classified as likely members based on their IR excess by Guarcello et al. (2013), 439 (of 1843) are detected in X-rays. Among the corresponding 510 optical matches, there are 365 Class II, 16 Class I, 58 flat spectrum, 43 transition and pretransition disks, and 22 accretors according to their Hα line, and 6 have blue excesses.

The spatial distributions of the foreground, member, and background objects are shown in Figure 8. Objects classified as members show a clear concentration that corresponds to the association, while the distribution of foreground objects is more isotropic. Objects classified as background show a small enhancement over regions of small A_v (see also Albacete-Colombo et al. 2023).

A short exemplar version of the catalog of classification is in Table 3 (the full catalog is available online) and lists the CXO ID, whether it has a corresponding OIR match, the computed probabilities of classification from the NBC method, and the final classification.

6. Summary

We have classified the X-ray sources observed toward Cygnus OB2 as being foreground objects, members of the association, or background objects using a variety of associated data, including optical and IR magnitudes, X-ray quantiles and hardness ratios, and extinction estimates. We adopt a Naive Bayes method to obtain automated classifications. We use domain knowledge of expected distributions of well-measured stars observed in different passbands to construct likelihoods that are then applied to the full data set. Likelihoods are constructed by using a semisupervised training method that uses objects with well-measured magnitudes, transformed using a PCA or modeled with mixtures of Gaussians to perform efficient separations for each channel. The probability that each source belongs to a given class is then computed, and sources are sifted into the appropriate class. This is then

augmented with visual inspection of several color–color and IRAC magnitude diagrams and correlated against known properties like the presence of disks that can cause systematic misassignments in the automated classification. We consider the effects of measurement, as well as systematic uncertainties due to extinction, and estimate that the residual error in our classification is ≈5%.

We construct a catalog that includes a probabilistic assessment of the class that each source belongs to. Adopting a hard threshold that states that the highest of the triad of {p(foreground), p(member), p(background)} determines the class, we find that ≈75% of the catalog sources are members of the Cygnus OB2 association, ≈5% are foreground stars, and the remainder are background objects.

Acknowledgments

We thank the anonymous referee for a careful reading of the paper and for comments that significantly improved its clarity. This work was supported by Chandra grant GO0-11040X. V.L. K., J.J.D., and T.L.A. were supported by NASA contract NAS8-03060 to the Chandra X-ray Center and thank the directors, B. Wilkes and P. Slane, and the CXC science team for continuing support and advice. M.G.G. and N.J.W. were supported by Chandra grant GO0-11040X during the course of this work. M. G.G. also acknowledges the grant PRIN-INAF 2012 (P.I. E. Flaccomio). N.J.W. acknowledges a Royal Astronomical Society Research Fellowship. J.F.A.C. is a researcher of CONICET and acknowledges their support. This work has made use of data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC; https://www.cosmos.esa.int/web/gaia/dpac/consortium); funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement. The UKIDSS project is defined in Lawrence et al. (2007). UKIDSS uses the UKIRT Wide Field Camera (WFCAM; Casali et al. 2007). The photometric system is described in Hewett et al. (2006), and the calibration is described in Hodgkin et al. (2009). The pipeline processing and science archive are described in Hambly et al. (2008).

Facility: CXO

Software: PINTofALE (Kashyap & Drake 2000), TRILEGAL (Girardi et al. 2012), IDL (v8.4)

Appendix A
Accounting for Uncertainties in Evaluating Likelihoods

Naive Bayes analyses allow for computing the probability of a class by evaluating the likelihood at an observed value. However, observations often have large errors, and often the size of the error bars is comparable to the scale at which the likelihoods vary. Point evaluations do not account for such uncertainties and can introduce a bias in the classification. We have developed a method to incorporate measurement uncertainties into the classification probabilities.

We model likelihoods locally as fourth-degree polynomials around the estimate of a given observed measure and weight them with an assumed normal error distribution to obtain an uncertainty-weighted likelihood. The likelihood $p(D|\theta)$, where D represents the data and θ is a parameter of interest, is typically a slowly varying function, especially in the context of Naive Bayes applications. Around a specific value θ_0, it can be expanded as a Taylor series,

$$p(D|\theta) \approx p(D|\theta_0) + \sum_{k=1} (\theta-\theta_0)^k \frac{\partial^k p(D|\theta)}{\partial\theta^k}\bigg|_{\theta_0}. \quad \text{(A1)}$$

It can thus be locally approximated as a Kth-degree polynomial,

$$\mathcal{L}(x) = \sum_{k=0}^{K} a_k x^k, \quad \text{(A2)}$$

where $x = \theta - \theta_0$ is the variable of interest.

A.1. Quartic Polynomials

Consider evaluations of $\mathcal{L}(x)$ at various values surrounding $x = 0$, $x = 0$, $\pm\delta$, $\pm 2\delta$. Suppose that it has the values

$$\mathcal{L}(-2\delta) = \mathcal{L}_{nn},\quad \mathcal{L}(-\delta) = \mathcal{L}_{n},\quad \mathcal{L}(0) = \mathcal{L}_{0},\quad \mathcal{L}(+\delta) = \mathcal{L}_{p},\quad \mathcal{L}(+2\delta) = \mathcal{L}_{pp}.$$

A polynomial that goes through these five points has the form

$$\begin{aligned}\mathcal{L}(x) &= a + b\,x + c\,x^2 + d\,x^3 + e\,x^4 \\ &\equiv A+ \\ &\quad B(x+2\delta)+ \\ &\quad C(x+2\delta)(x+\delta)+ \\ &\quad D(x+2\delta)(x+\delta)(x)+ \\ &\quad E(x+2\delta)(x+\delta)(x)(x-\delta). \end{aligned} \quad \text{(A3)}$$

After some algebra, we obtain the coefficients as expressions of $\mathcal{L}(\cdot)$,

$$\begin{aligned} A &= \mathcal{L}_{nn} \\ B &= \frac{1}{\delta}(-\mathcal{L}_{nn} + \mathcal{L}_{n}) \\ C &= \frac{1}{2\delta^2}(\mathcal{L}_{nn} - 2\,\mathcal{L}_{n} + \mathcal{L}_{0}) \\ D &= \frac{1}{6\delta^3}(-\mathcal{L}_{nn} + 3\,\mathcal{L}_{n} - 3\,\mathcal{L}_{0} + \mathcal{L}_{p}) \\ E &= \frac{1}{24\delta^4}(\mathcal{L}_{nn} - 4\,\mathcal{L}_{n} + 6\,\mathcal{L}_{0} - 4\,\mathcal{L}_{p} + \mathcal{L}_{pp}) \end{aligned} \quad \text{(A4)}$$

and

$$\begin{aligned} a &= \mathcal{L}_{0} \\ b &= \frac{1}{12\delta}[\mathcal{L}_{nn} - 8\mathcal{L}_{n} + 8\mathcal{L}_{p} - \mathcal{L}_{pp}] \\ c &= \frac{1}{24\delta^2}[-\mathcal{L}_{nn} + 16\mathcal{L}_{n} - 30\mathcal{L}_{0} + 16\mathcal{L}_{p} - \mathcal{L}_{pp}] \\ d &= \frac{1}{12\delta^3}[-\mathcal{L}_{nn} + 2\mathcal{L}_{n} - 2\mathcal{L}_{p} + \mathcal{L}_{pp}] \\ e &= \frac{1}{24\delta^4}[\mathcal{L}_{nn} - 4\,\mathcal{L}_{n} + 6\,\mathcal{L}_{0} - 4\,\mathcal{L}_{p} + \mathcal{L}_{pp}]. \end{aligned} \quad \text{(A5)}$$

A.2. Weighting by Normal

We assume that the error distributions on the data points are normal, of the form

$$\mathcal{E}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}. \quad \text{(A6)}$$

The weighted likelihood estimate is then

$$p = \int_{-\infty}^{+\infty} dx\ \mathcal{L}(x)\,(x), \quad \text{(A7)}$$

since the process can be thought of as averaging over an ensemble of observations, with $\mathcal{E}(x)$ providing an importance weight for independent draws.

Noting that $\int_{-\infty}^{+\infty} dx\, x^{2n} f(x) = 2\int_{0}^{\infty} dx\; x^{2n}\, f(x)$ and $\int_{-\infty}^{+\infty} dx\, x^{2n+1} f(x) = 0$ for symmetric $f(x)$ and integer n and using known calculations of the integral (see, e.g., Gradshteyn & Ryzhik, Equations (3.461.2–3)), we obtain the uncertainty-weighted likelihood,

$$\begin{aligned}\mathcal{L}|_{\mathcal{E}} &= a \;+\; c\sigma^2 \;+\; 3e\sigma^4 \\ &= \mathcal{L}_0 + \frac{\sigma^2}{24\delta^2}(-\mathcal{L}_{nn} + 16\mathcal{L}_n - 30\mathcal{L}_0 + 16\mathcal{L}_p - \mathcal{L}_{pp}) \\ &\quad + \frac{3\sigma^4}{24\delta^4}(\mathcal{L}_{nn} - 4\mathcal{L}_n + 6\mathcal{L}_0 - 4\mathcal{L}_p + \mathcal{L}_{pp}).\end{aligned} \tag{A8}$$

Choosing the natural scale in the problem, $\delta = \sigma/z$,

$$\begin{aligned}\mathcal{L}|_{\mathcal{E}} = \frac{1}{24}[&\mathcal{L}_{nn}(3z^4 - z^2) + \mathcal{L}_n(-12z^4 + 16z^2) + \mathcal{L}_0(18z^4 - 30z^2 + 24) \\ &+ \mathcal{L}_p(-12z^4 + 16z^2) + \mathcal{L}_{pp}(3z^4 - z^2)].\end{aligned} \tag{A9}$$

For $z = 1$, the above reduces to

$$\mathcal{L}|_{\mathcal{E}} = \frac{1}{24}[2\mathcal{L}_{nn} + 4\mathcal{L}_n + 12\mathcal{L}_0 + 4\mathcal{L}_p + 2\mathcal{L}_{pp}]. \tag{A10}$$

Notice that these expressions have some desirable mathematical properties: the error-weighted likelihood is positive definite; the coefficients of each x^k are symmetric in how $\mathcal{L}(.)$ at $\pm\delta$ and $\pm 2\delta$ are included; if the likelihood function is flat, all coefficients of x^k for $k > 0$ vanish; and finally, if the likelihood function is flat, $\mathcal{L}|_{\mathcal{E}} \equiv \mathcal{L}_0$.

Appendix B
Extinction

B.1. Fukugita+ versus Fitzpatrick and Massa

As discussed in Section 4.6, we primarily use the extinction law of Fukugita et al. (1996, F96) to compute A_v. However, the extinction law based on the newer study of Fitzpatrick et al. (2007, 2009, FM) is a viable alternative. We do not use the latter as our primary reference because the peak of the distribution of A_v is shifted lower by ≈1.5 mag. Here we consider the effect of changing the extinction law on the classification (see Figure B1). In Table B1, we show how many sources change their Naive Bayes−based classification based on this change. We then carry out the same analysis as in Section 5.2.1, reclassifying sources manually, and show how many sources are reclassified in Table B2. In both cases, we find a similar fraction of changes, suggesting that our statistical classification is effectively at the limit defined by potential systematic errors in the data sets.

B.2. A_v across the Field of View

Individual extinction for stars associated with Cygnus OB2 is calculated with a similar approach to that of Guarcello et al. (2012). In that paper, individual extinction of the X-ray sources with an optical counterpart from the OSIRIS or SDSS catalogs is calculated from the displacement along the extinction vector from the $A_V = 0$ mag, 3.5 Myr isochrone from Siess et al. (2000) in the $r - i$ versus $i - z$ diagram. This method is feasible thanks to the almost monotonic shape of the isochrone in this color space, but it requires the use of suitable color transformations from the Johnson−Cousins *UBVRI* to the $u_I\, g_I\, r_I\, i_I\, z_I$ photometric system. Guarcello et al. (2012) adopted the transformations from F96.

In this paper, in order to avoid the use of any photometric transformation, we adopted the MIST isochrones that are provided in several photometric systems, one of which is $u_I\, g_I\, r_I\, i_I\, z_I$.[12] Besides, these isochrones span a wider range of stellar mass, allowing the calculation of individual extinction also for stars more massive than 7 $M_\odot$, which is the upper mass limit in the Siess et al. (2000) isochrones. Additionally, in this paper we calculate the individual extinction of those 2MASS/UKIDSS sources without a good optical counterpart using the NICER method (Lombardi & Alves 2001), based on the $H - K$ color.

The left panel of Figure B2 shows the distribution of the resulting individual extinctions for the X-ray sources associated with Cygnus OB2. The median value is 4.2 mag, with the 10% quantile equal to 2.7 mag and the 90% quantile to 7.9 mag, quite similar to previous estimates (e.g., Drew et al. 2008; Sale

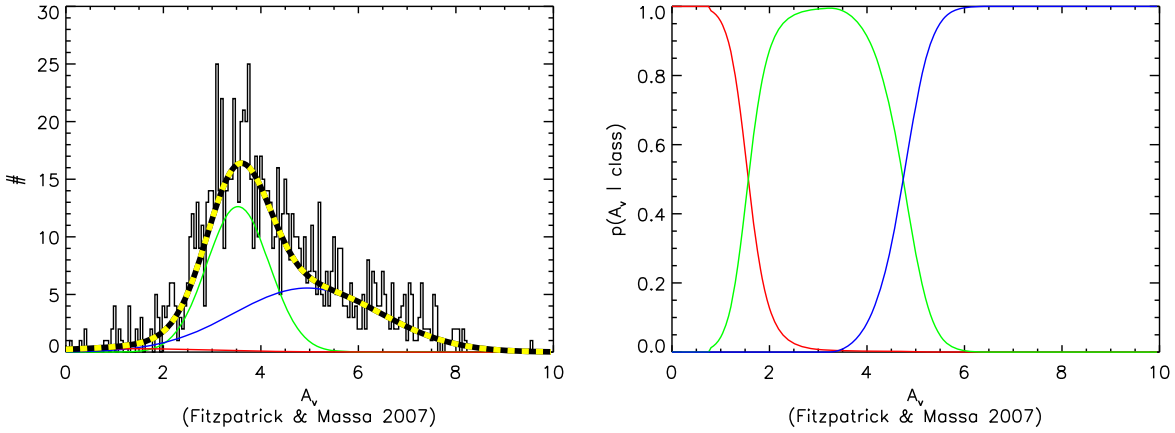

Figure B1. As in the first two panels of Figure 4, but for extinctions derived from FM.

[12] MIST isochrones can be downloaded from http://waps.cfa.harvard.edu/MIST/index.html.

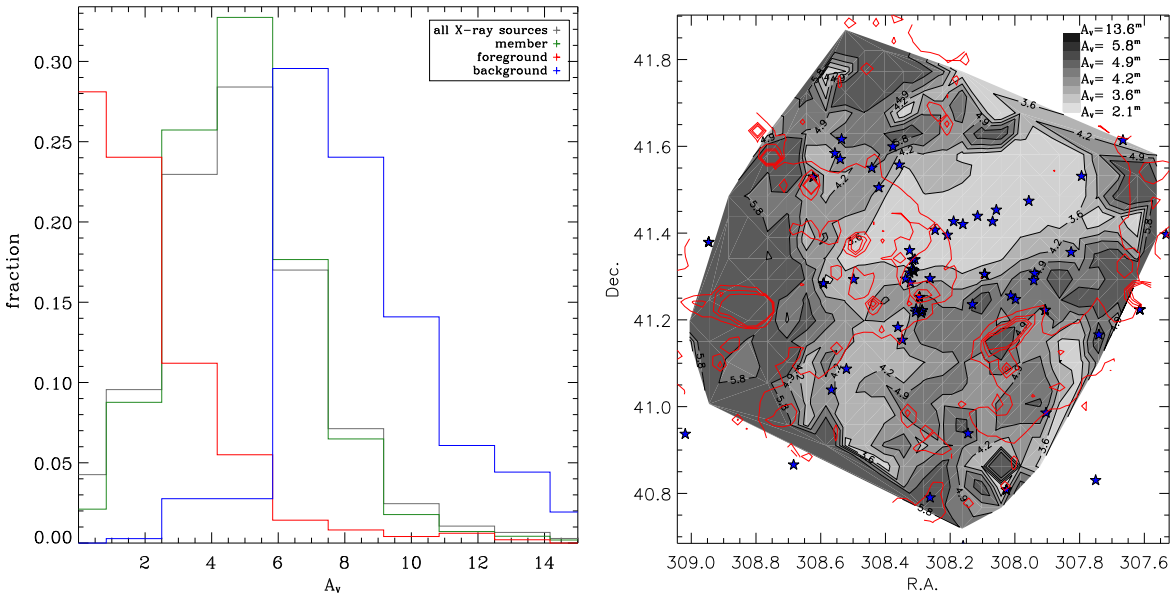

Figure B2. Left: distribution of the individual extinctions of the sources in the catalog. Right: map of stellar extinction across the area of the survey. The red contours mark the continuum emission levels at 8.0 μm. The blue stars mark the positions of known O stars.

Table B1
Changes in Naive Bayes Classification upon Changing Extinction Law from F96 to FM

F96	FM			Subtotal F96
	Foreground	Member	Background	
Foreground	610	7	2	619
Member	70	5740	124	5934
Background	3	83	1381	1467
Subtotal FM	683	5830	1507	8020

Table B2
Post Hoc Reclassifications (as in Table 2) but Based on FM Extinction Law

Original	Reclassified			Subtotal Original
	Foreground	Member	Background	
Foreground	491	192	0	683
Member	60	5768	2	5830
Background	0	110	1397	1507
Subtotal reclassified	551	6070	1399	8020

et al. 2009; Wright et al. 2010). The right panel shows the spatial map of extinction across the area of the survey, comparing it with the level of continuum emission at 8.0 μm that marks the dust emission. The well-known low-extinction region in the northwest is evident, as well as the large extinction regions corresponding to some of the dusty structures in the cloud.

Appendix C
X-Ray Sources with No OIR Matches

A detailed description of the matching procedure to determine OIR counterparts to the X-ray sources is given in Guarcello et al. (2023a). For the 1428 X-ray sources classified here as members but that have no counterparts, Guarcello et al. (2023a) list the nearest match in a supplemental table. Here we consider some properties of this population.

The presence of different populations in the sample of the X-ray sources without an OIR counterpart is evident by looking at the distribution of their photon median energy and the parameter *prob-no-source* evaluated by ACIS Extract (Broos et al. 2010). The latter parameter indicates the reliability of the source in terms of probability that it is a background fluctuation. The distribution of the photons' median energy of the X-ray sources with no OIR counterpart is clearly different (see Figure C1) than that of the whole sample, being almost flat while the latter peaks at about 1.8 keV. The high-energy tail of the X-ray sources with no OIR counterpart is due to background sources. The nature of the soft sources in this sample can be investigated by looking at the distribution of the prob-no-source parameter. For soft X-ray sources with no OIR counterpart (red distribution) the distribution of this parameter clearly peaks at high values, while the low-value bins are populated by sources with higher median photon energy, suggesting that the soft X-ray sources with no OIR counterpart are mainly candidate spurious sources, while those with higher photon energy are mainly genuine detections.

Figure C2 shows the spatial distributions of the X-ray sources without OIR counterparts classified as members and those classified as sources separately. While the latter are more uniformly distributed, the candidate members show some level of clustering corresponding to various Cygnus OB2 subclusters, as expected for real background and clusters sources. Both samples also show a clear halo of sources around the position of Cygnus X-3, which are more likely spurious X-ray sources that survived the pruning process of the X-ray catalog.

To further verify the nature of the X-ray sources with no OIR counterparts that are classified as members, we have inspected their positions in various optical and IR diagrams and the optical images of the closest source in the OIR catalog of each of these sources, selecting 46 candidate false negatives

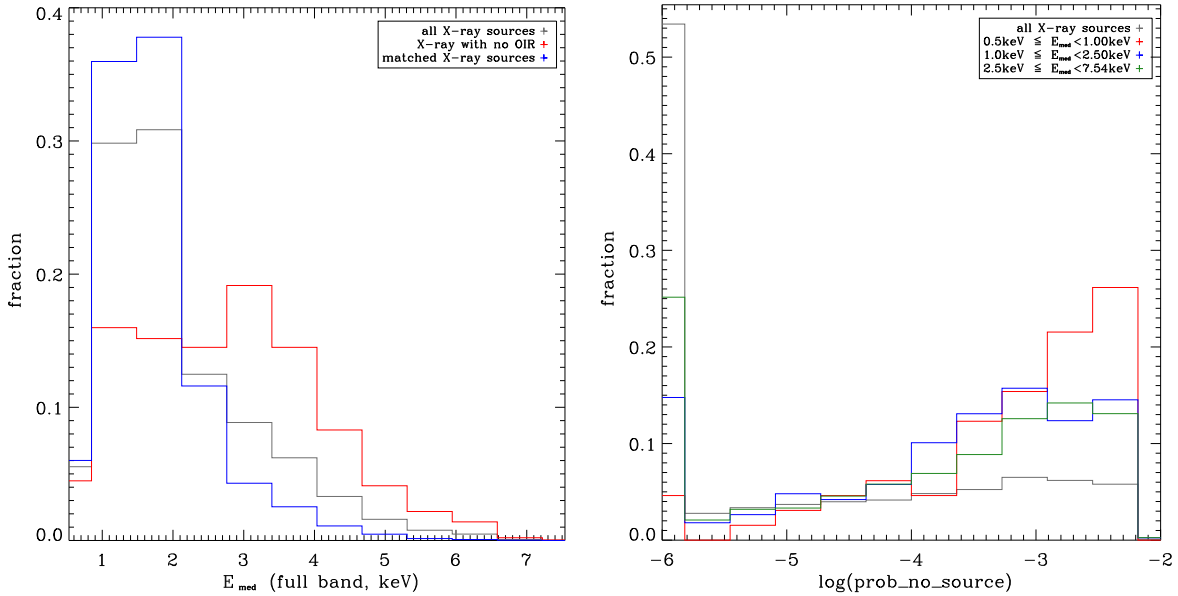

Figure C1. Left: distributions of the median energy of the detected photons for all the X-ray sources detected in our survey (gray), for those matched with OIR counterparts (blue), and for the X-ray sources without an OIR counterpart (red). Right: distribution of the ACIS Extract parameter *prob-no-source* for the X-ray sources with no OIR counterpart separated according to their median photon energies, with soft (0.5–1 keV; red), medium (1–2.5 keV; blue), hard (2.5–7.5 keV; green), and all (black).

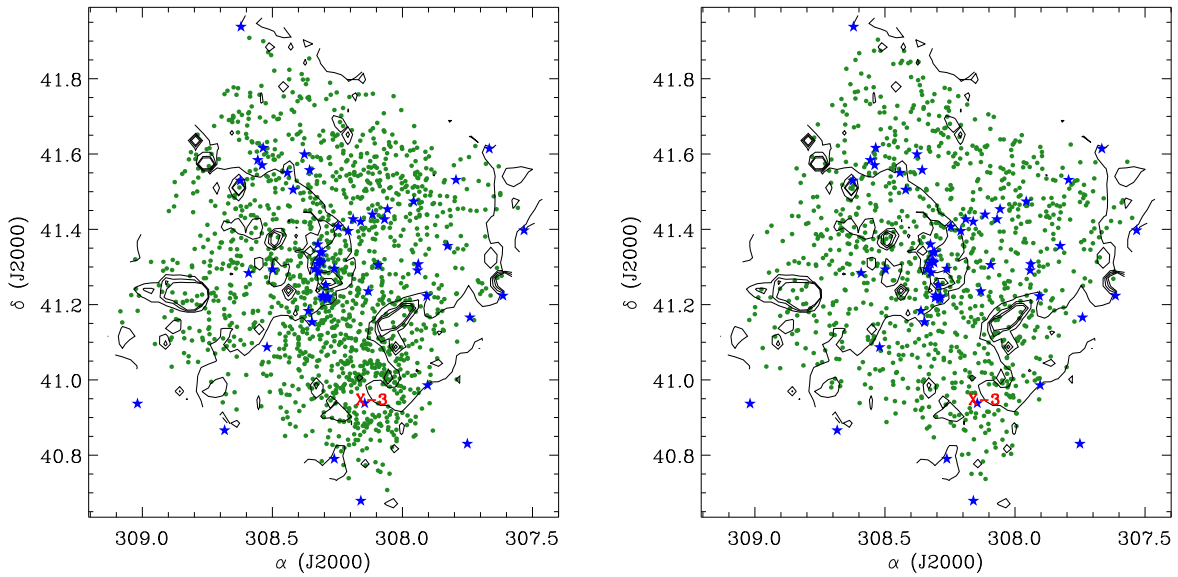

Figure C2. Spatial distributions of the X-ray sources with no OIR counterpart classified as members (left panel) and background sources (right panel). In each panel, the blue stars mark the positions of the known O stars; the contours mark the background brightness at 8.0 μm, tracing the dust emission. The position of Cygnus X-3 is also indicated.

produced in the match between the X-ray and the OIR catalogs. Table C1 lists their CXO-IDs and the separation in arcseconds from the closest OIR source. Note that some of these sources have an OIR star closer than 1″. The listed sources are divided into four categories. In four cases the closest OIR source is a known star with disk classified by Guarcello et al. (2013). In the remainder the closest OIR source falls into the loci defined by Cyg OB2 members in the various diagrams. Five stars have $J < 12$ mag (candidate bright members), 19 sources have 13 mag $< J <$ 16 mag (candidate members), and 18 sources have 16 mag $< J <$ 19 mag (candidate low-mass or highly extinguished members). In particular, the OIR source 0.″1 from the X-ray source 3532 is compatible with being a member in IR but not in optical, likely being a false coincidence between a low-extinction optical source and a high-extinction IR source; the X-ray source 4675 is close to the O8.5V star MT91-8D, which has been matched with the X-ray source 4673.

Table C1
Candidate X-Ray vs. OIR False Negatives

CXO ID	Separation (arcseconds)
Stars with Disks	
56	1.9
2327	2.4
3099	1.6
3624	2.2
Candidate OIR-bright Members	
796	2.2
3692	1.9
4602	1.7
4675	1.9
7115	1.2
Candidate Members	
121	3.2
1791	2.1
1822	1.8
2214	2.3
2397	1.2
2788	0.6
2910	5.4
3056	1.5
3141	1.8
3185	1.5
3759	1.0
3942	0.6
4590	2.8

Table C1
(Continued)

CXO ID	Separation (arcseconds)
5130	5.1
5430	1.8
5442	0.2
6905	2.0
7655	2.7
7699	3.1
Candidate Faint Members	
102	2.3
222	0.8
2558	2.5
2850	0.6
3006	5.1
3118	4.2
3346	6.5
3532	0.1
3838	2.4
4005	0.7
4835	1.6
5573	0.6
6167	2.5
6518	1.9
6593	1.4
6681	0.9
6861	0.5
7201	3.0

Appendix D
Adopted Likelihoods

Here we show the various combinations of magnitudes that were used to define the likelihoods for classification in Section 4. The combinations PCA(1){J, K}, PCA(2){J, K}, PCA(2){H, J, K}, PCA(2){$r_I - i_I$, $r_I - \mathrm{H}\alpha$}, PCA(3){r_I, i_I, Hα}, and PCA(1){Q_{25}, Q_{50}, Q_{75}} are shown in Figures D1–D6.

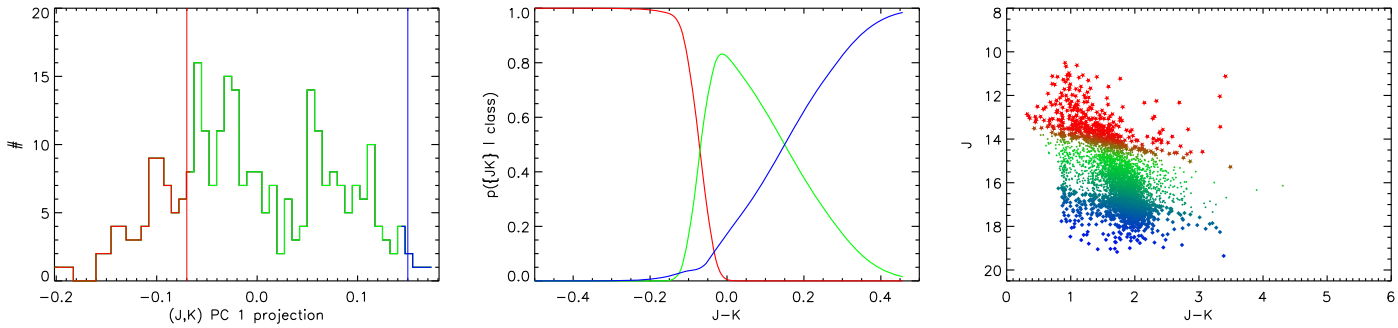

Figure D1. Same as Figure 1, but for the first PC of {J, K}.

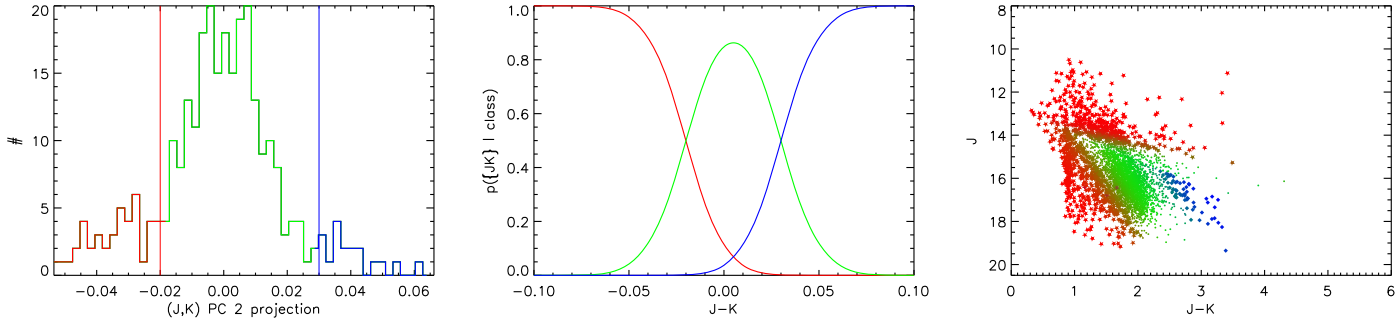

Figure D2. Same as Figure 1, but for the second PC of {J, K}. Additionally, during the classification (right panel), the higher of the foreground likelihoods between the first (Figure D1) and second components is chosen.

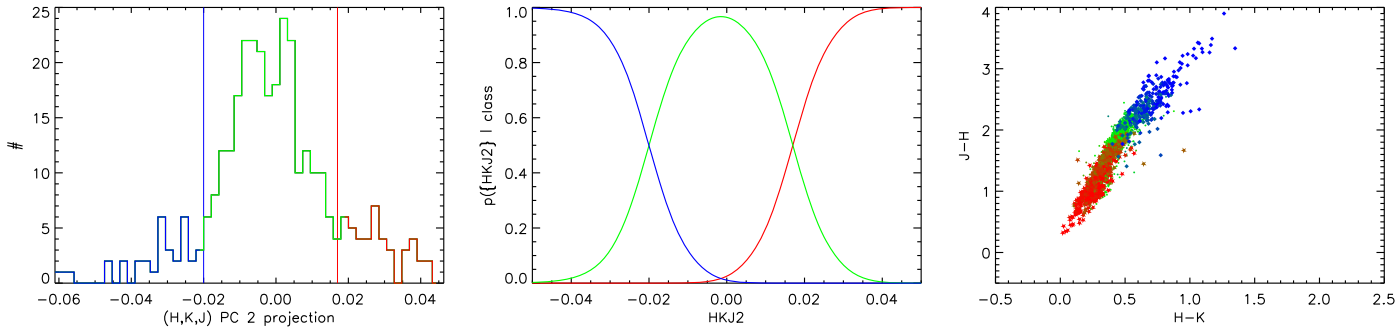

Figure D3. Same as Figure 1, but for the second PC of {H, J, K}.

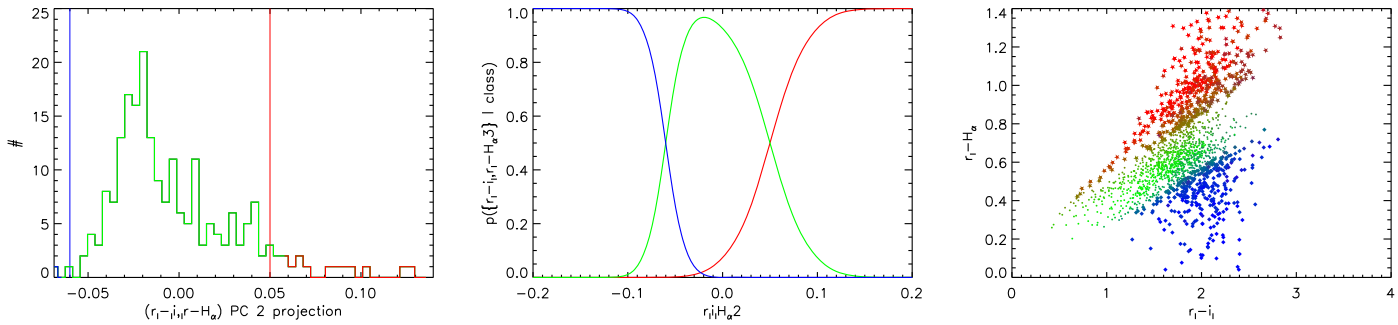

Figure D4. Same as Figure 1, but for the second PC of {$r_I - i_I$, $r_I - \mathrm{H}\alpha$}.

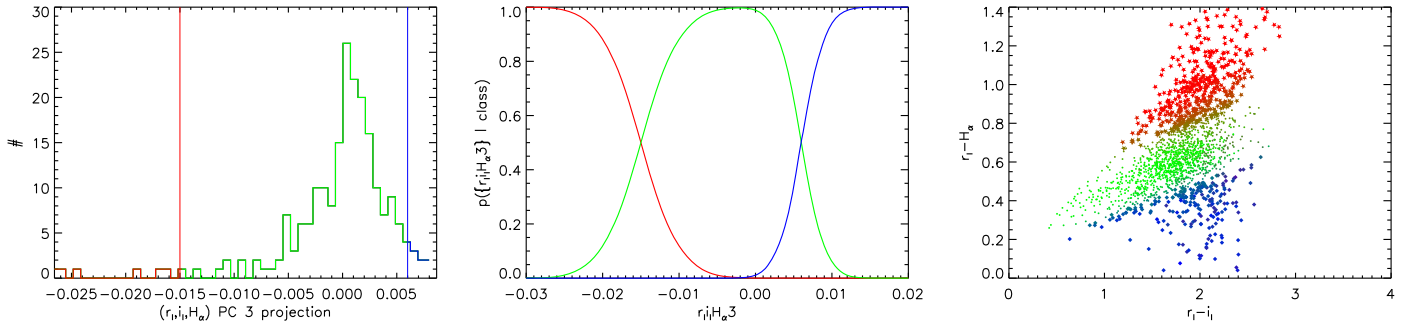

Figure D5. Same as Figure 1, but for the third PC of $\{r_I, i_I, \mathrm{H}\alpha\}$.

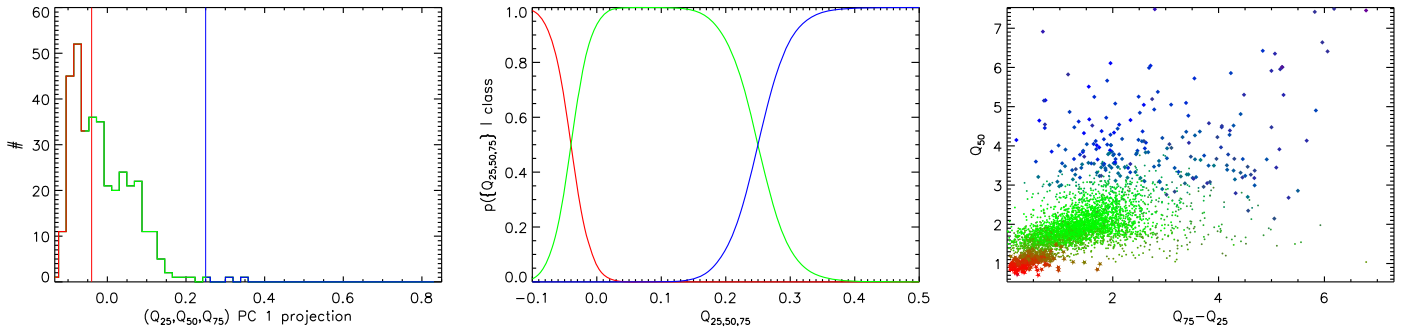

Figure D6. Same as Figure 1, but for the first PC of $\{Q_{25}, Q_{50}, Q_{75}\}$.

Appendix E
The Effect of Reclassification

Here we show various plots of the properties of the cataloged sources comparing the classes as derived from the automated Naive Bayes approach and after manual reclassification (see Section 5.2.1). In all figures, red points mark sources classified as foreground, green points mark those classified as members of the association, and blue points mark background objects. The diagrams for $r_I - \mathrm{H}\alpha$ versus $r_I - i_I$, $g_I - r_I$ versus $r_I - i_I$, J versus $J - K$, Spitzer [4.5] − [5.8] versus [5.8] − [8.0], and their spatial distributions are shown in Figures E1–E5.

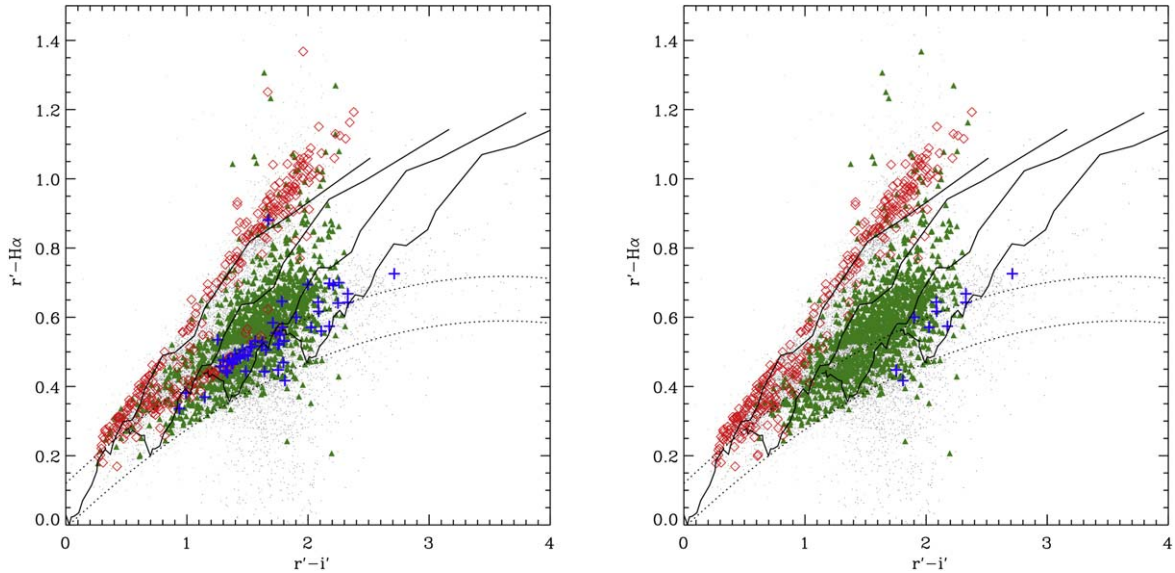

Figure E1. Same as Figure 6, but for $r_I - \mathrm{H}\alpha$ vs. $r_I - i_I$. The black lines are ZAMS with increasing extinction: from $E_{B-V} = 1$ mag to $E_{B-V} = 4$ mag from Drew et al. (2005). The curved dotted lines mark the locus typically populated by A stars.

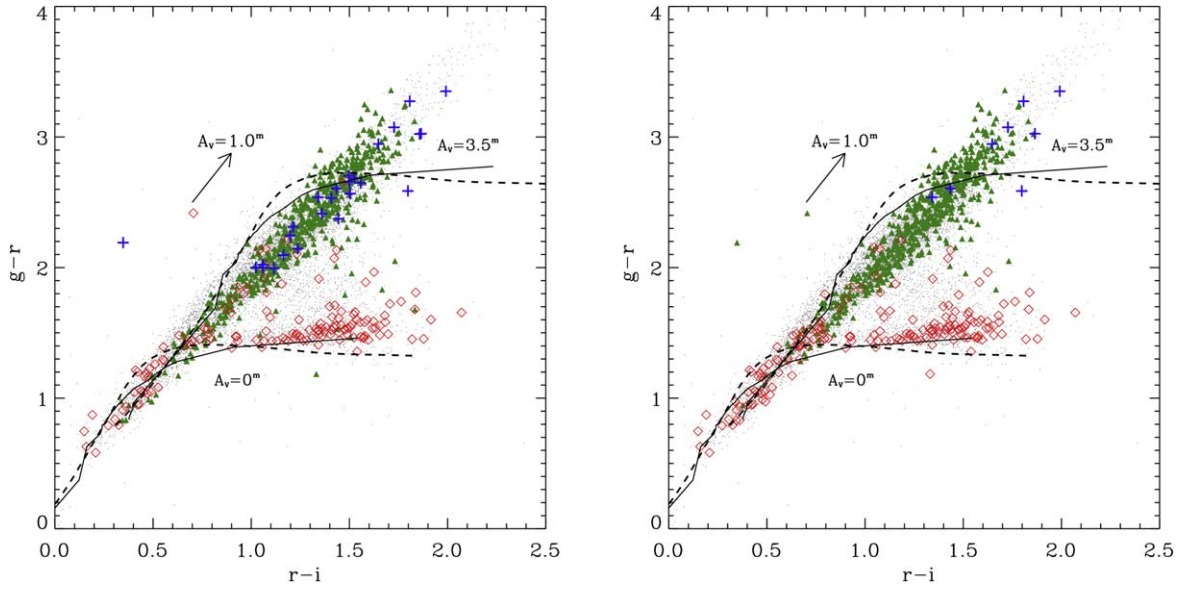

Figure E2. Same as Figure 6, but for $g_I - r_I$ vs. $r_I - i_I$ of sources with good-quality SDSS photometry. The 2.5 Myr isochrones from Siess et al. (2000) with $A_v = 0$ mag and $A_v = 3.5$ mag are shown as solid lines; the 2.5 Myr MIST isochrones are shown as dashed lines.

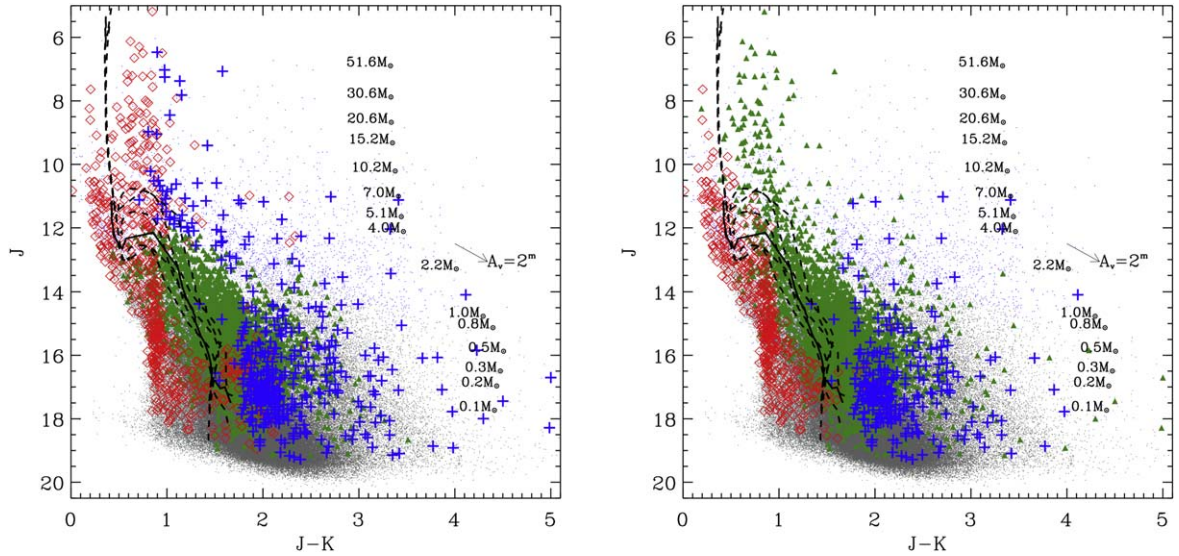

Figure E3. Same as Figure 6, but for J vs. $(J - K)$ for sources with good 2MASS or UKIDSS photometry. The solid line is the 2.5 Myr isochrone with $A_v = 3.5$ mag from Siess et al. (2000), while the dashed line marks the MIST isochrone.

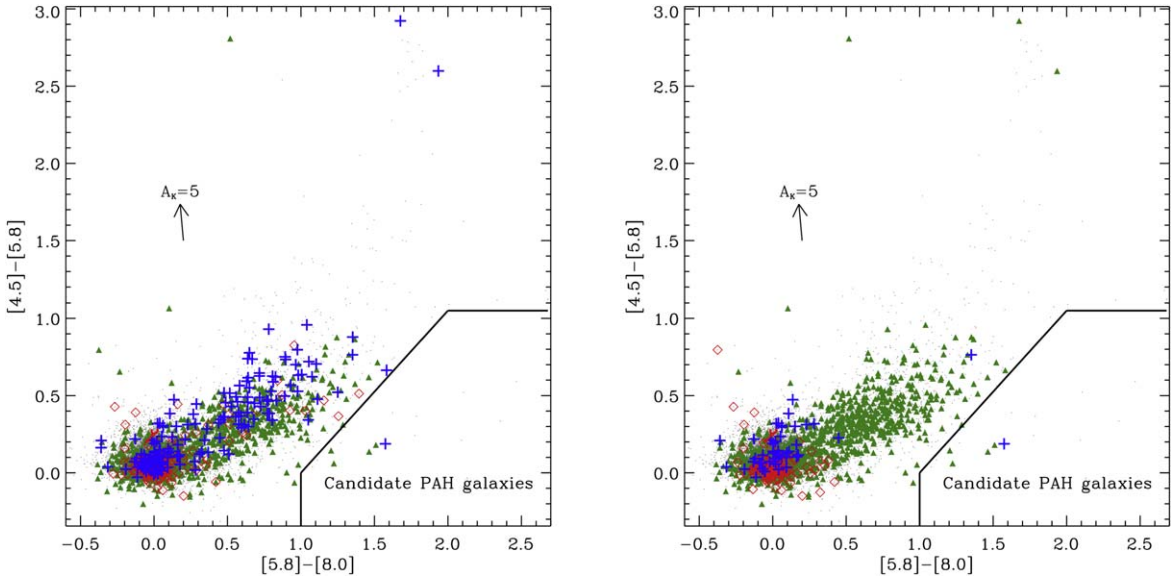

Figure E4. Same as Figure 6, but for Spitzer [4.5] − [5.8] vs. [5.8] − [8.0] colors of sources with good IRAC photometry. The solid line delimits the locus typically populated by PAH galaxies.

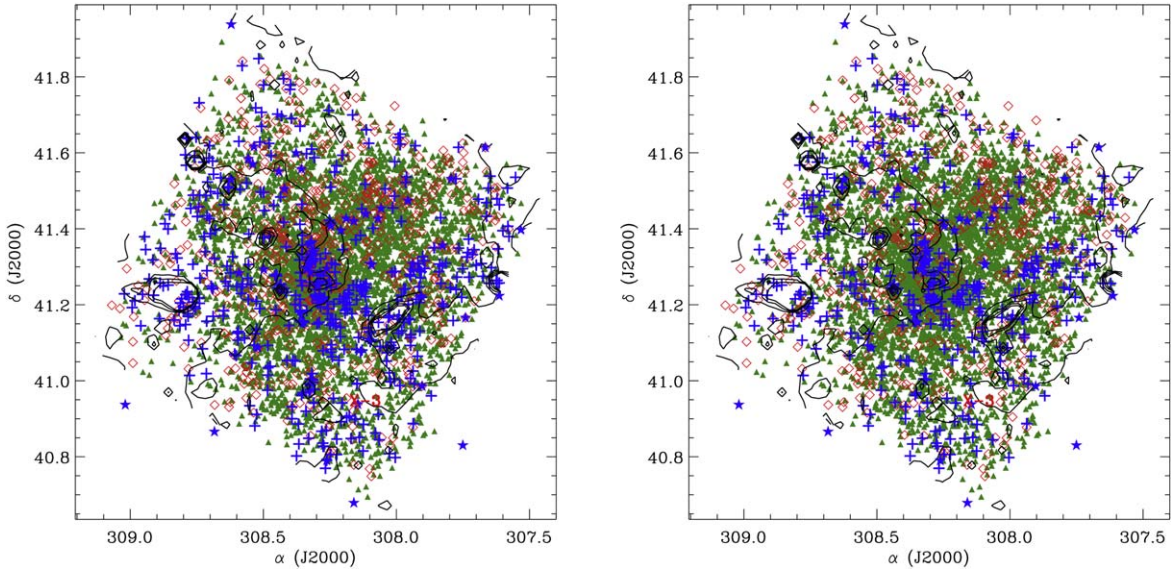

Figure E5. Spatial distribution of the X-ray sources detected in our survey and classified according to the automated NBC scheme (left) and after manual reclassification (right). The black contours delimit continuum emission levels at [8.0] μm Spitzer band. The colors and symbols are as in Figure 6.

ORCID iDs

Vinay L. Kashyap https://orcid.org/0000-0002-3869-7996
Mario G. Guarcello https://orcid.org/0000-0002-3010-2310
Jeremy J. Drake https://orcid.org/0000-0002-0210-2276
Ettore Flaccomio https://orcid.org/0000-0002-3638-5788
Juan F. Albacete Colombo https://orcid.org/0000-0001-8398-0515
Francesco Damiani https://orcid.org/0000-0002-7065-3061
Eduardo L. Martin https://orcid.org/0000-0002-1208-4833
Giusi Micela https://orcid.org/0000-0002-9900-4751
Tim Naylor https://orcid.org/0000-0002-0506-8501
Salvatore Sciortino https://orcid.org/0000-0001-8691-1443

References

Aihara, H., Allende Prieto, C., An, D., et al. 2011, ApJS, 193, 29
Albacete Colombo, J. F., Caramazza, M., Flaccomio, E., Micela, G., & Sciortino, S. 2007, A&A, 474, 495
Albacete-Colombo, J. F., Drake, J. J., Flaccomio, E., et al. 2023, ApJS, 269, 14
Barentsen, G., Farnhill, H. J., Drew, J. E., et al. 2014, MNRAS, 444, 3230
Beerer, I., Koenig, X. P., Hora, J. L., et al. 2010, ApJ, 720, 679
Benavente, P., Protopapas, P., & Pichara, K. 2017, ApJ, 845, 147
Broos, P., Getman, K., & Povich, M. 2011, ApJS, 194, 4
Broos, P., Townsley, L., & Feigelson, E. 2010, ApJ, 714, 1582
Casali, M., Adamson, A., Alves de Oliveira, C., et al. 2007, A&A, 467, 777
Cepa, J., Aguiar, M., Escalera, V. G., et al. 2000, Proc. SPIE, 4008, 623
Choi, J., Dotter, A., & Conroy, C. 2016, ApJ, 823, 102
Comeron, F., Torra, J., & Gomez, A. 1992, Ap&SS, 187, 187
Cutri, R., Skrutskie, M. F., van Dyk, S., et al. 2003, yCat, II/246
Davis, A. B., Cisewski, J., Dumusque, X., Fischer, D. A., & Ford, E. B. 2017, ApJ, 846, 59
Domingos, P., & Pazzani, M. 1997, Mach. Learn., 29, 103
Dotter, A. 2016, ApJS, 222, 8
Drew, J., Greimel, R., Irwin, M., & Sale, S. 2008, MNRAS, 386, 1761
Drew, J., Greimel, R., Irwin, M. J., et al. 2005, MNRAS, 362, 753
Dye, S., Lawrence, A., Read, M. A., et al. 2018, MNRAS, 473, 5113
Fitzpatrick, E., & Massa, D. 2007, ApJ, 663, 320

Fitzpatrick, E., & Massa, D. 2009, ApJ, 699, 1209
Flaccomio, E., Albacete Colombo, J. F., Drake, J. J., et al. 2023, ApJS, 269, 12
Fukugita, M., Ichikawa, T., & Gunn, J. 1996, AJ, 111, 1748
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1
Girardi, L., Barbieri, M., Groenewegen, M. A. T., et al. 2012, Red Giants as Probes of the Structure and Evolution of the Milky Way (Berlin: Springer), 165
Gopalan, G., Vrtilek, S., & Bornn, L. 2015, ApJ, 809, 40
Guarcello, M. G., Damiani, F., Micela, G., et al. 2010, A&A, 521, A18
Guarcello, M. G., Drake, J. J., Wright, N. J., et al. 2013, ApJ, 773, 135
Guarcello, M. G., Drake, J. J., Wright, N. J., et al. 2023a, ApJS, 269, 9
Guarcello, M. G., Drake, J. J., Wright, N. J., et al. 2023b, ApJS, 269, 13
Guarcello, M. G., Wright, N., & Drake, J. 2012, ApJS, 202, 19
Hambly, N. C., Collins, R. S., Cross, N. J. G., et al. 2008, MNRAS, 384, 637
Hanson, M. 2003, ApJ, 597, 957
Heavens, A. F., Jimenez, R., & Lahav, O. 2000, MNRAS, 317, 965
Hewett, P., Warren, S., Leggett, S., & Hodgkin, S. 2006, MNRAS, 367, 454
Hodgkin, S. T., Irwin, M. J., Hewett, P. C., et al. 2009, MNRAS, 394, 675
Hojnacki, S. M., Kastner, J. H., Micela, G., Feigelson, E. D., & LaLonde, S. M. 2007, ApJ, 659, 585
Hong, J., Schlegel, E. M., & Grindlay, J. E. 2004, ApJ, 614, 508
Kashyap, V., & Drake, J. J. 2000, BASI, 28, 475
King, R., Naylor, T., Broos, P., Getman, K., & Feigelson, E. 2013, ApJS, 209, 28
Knödlseder, J. 2000, A&A, 360, 539
Kuhn, M., Getman, K., Broos, P., Townsley, L., & Feigelson, E. 2013, ApJS, 209, 27
Lawrence, A., Warren, S. J., Almaini, O., et al. 2007, MNRAS, 379, 1599
Lee, H., Kashyap, V. L., van Dyk, D. A., et al. 2011, ApJ, 731, 126
Lehmer, B., Xue, Y. Q., Brandt, W. N., et al. 2012, ApJ, 752, 46
Lombardi, M., & Alves, J. 2001, A&A, 377, 1023
Lucas, P., Hoare, M. G., Longmore, A., et al. 2008, MNRAS, 391, 136
Massey, P., & Thompson, A. 1991, AJ, 101, 1408
O'Donnell, J. 1994, ApJ, 422, 158
Park, T., Kashyap, V., & Siemiginowska, A. 2006, ApJ, 652, 610
Patil, A. A., Bovy, J., Eadie, G., & Jaimungal, S. 2022, ApJ, 926, 51
Predehl, P., & Schmitt, J. H. M. M. 1995, A&A, 500, 459
Rygl, K., Brunthaler, A., Sanna, A., et al. 2012, A&A, 539, 79
Sale, S., Drew, J. E., Unruh, Y. C., et al. 2009, MNRAS, 392, 497
Sasdelli, M., Ishida, E. E. O., & Hillebrandt, W. 2016a, MNRAS, 460, 373
Sasdelli, M., Ishida, E. E. O., Vilalta, R., et al. 2016b, MNRAS, 461, 2044
Siess, L., Dufour, E., & Forestini, M. 2000, A&A, 358, 593
Stampoulis, V., van Dyk, D. A., Kashyap, V. L., & Zezas, A. 2019, MNRAS, 485, 1085
Stein, N., van Dyk, D., & Kashyap, V. 2015, Stat. Interface, 9, 535
Waddell, S. G. H., & Gallo, L. C. 2020, MNRAS, 498, 5207
Wright, N. J., & Drake, J. J. 2009, ApJS, 184, 84
Wright, N. J., Drake, J. J., Drew, J. E., & Vink, J. S. 2010, ApJ, 713, 871
Wright, N. J., Drake, J. J., Guarcello, M. G., et al. 2023a, ApJS, 269, 7
Wright, N. J., Drake, J. J., Guarcello, M. G., Kashyap, V. L., & Zezas, A. 2023b, ApJS, 269, 8
Wright, N. J., Drew, J. E., & Mohr-Smith, M. 2015, MNRAS, 449, 741
Xu, J., van Dyk, D. A., Kashyap, V. L., et al. 2014, ApJ, 794, 97