Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **517,** 140–160 (2022) https://doi.org/10.1093/mnras/stac2692
Advance Access publication 2022 September 23

Massive young stellar objects in the Local Group spiral galaxy M 33 identified using machine learning

David A. Kinson ,★ Joana M. Oliveira and Jacco Th. van Loon
Lennard-Jones Laboratories, School of Chemical and Physical Sciences, Keele University, ST5 5BG, UK

Accepted 2022 September 15. Received 2022 August 23; in original form 2022 July 22

ABSTRACT

We present a supervised machine learning classification of stellar populations in the Local Group spiral galaxy M 33. The Probabilistic Random Forest (PRF) methodology, previously applied to populations in NGC 6822, utilizes both near and far-IR classification features. It classifies sources into nine target classes: young stellar objects (YSOs), oxygen, and carbon-rich asymptotic giant branch stars, red giant branch, and red super-giant stars, active galactic nuclei, blue stars (e.g. O-, B-, and A-type main sequence stars), Wolf–Rayet stars, and Galactic foreground stars. Across 100 classification runs the PRF classified 162 746 sources with an average estimated accuracy of ∼86 per cent, based on confusion matrices. We identified 4985 YSOs across the disc of M 33, applying a density-based clustering analysis to identify 68 star forming regions (SFRs) primarily in the galaxy's spiral arms. SFR counterparts to known H II regions were recovered with ∼91 per cent of SFRs spatially coincident with giant molecular clouds identified in the literature. Using photometric measurements, as well as SFRs in NGC 6822 with an established evolutionary sequence as a benchmark, we employed a novel approach combining ratios of [Hα]/[24 μm] and [250 μm]/[500 μm] to estimate the relative evolutionary status of all M 33 SFRs. Masses were estimated for each YSO ranging from 6–27$M_\odot$. Using these masses, we estimate star formation rates based on direct YSO counts of $0.63 M_\odot\ yr^{-1}$ in M 33's SFRs, $0.79 \pm 0.16 M_\odot\ yr^{-1}$ in its centre and $1.42 \pm 0.16 M_\odot\ yr^{-1}$ globally.

Key words: methods: statistical – stars: formation – stars: protostars – Galaxies: individual (M 33) – Local Group – galaxies: stellar content.

1 INTRODUCTION

Studies of the galaxy M 33 and its stellar populations began with Hubble (1926), yet nearly 100 years hence a comprehensive study of resolved star formation across the galaxy is still unavailable. M 33 is the third largest galaxy in the Local Group ($M_{gas} \sim 3 \times 10^9 M_\odot$, Corbelli 2003; $M_* \sim 5.5 \times 10^9 M_\odot$, Corbelli et al. 2014; Kam et al. 2017), after the Milky Way and M 31. M 33 lies at a distance of ∼850 kpc ($\mu_{M33} = 24.67$ mag, de Grijs & Bono 2014) and extends to an apparent size of approximately 60 × 35 arcmin (Paturel et al. 2003). Its relatively face-on inclination ($i = 54°$, de Vaucouleurs et al. 1991) makes M 33 a more favourable target to study the entirety of a spiral galaxy's disc over the larger and similarly distant M 31, which is seen nearly edge on (e.g. Ma 2001).

The metallicity of M 33 is around half-solar (e.g. Braine et al. 2018), similar to that of the LMC (see fig. 1 of Williams et al. 2021). The metallicity of M 33 varies across the disc with a negative gradient with increasing galactocentric radius well documented (e.g. Searle 1971; Cioni 2009; Magrini et al. 2010; Alexeeva & Zhao 2022); however its steepness is debated with recent results favouring a shallower slope (Alexeeva & Zhao 2022). A negative gradient supports an inside-out model of disc formation (Cioni 2009; Williams et al. 2009), supported in M 33's by the observed star formation history radial profiles (Williams et al. 2009; Javadi et al. 2017). The radial stellar age profile has been reported to reverse at radii larger than 9 kpc beyond the break in optical brightness of the disc (Williams et al. 2009; Barker et al. 2011; Mostoghiu et al. 2018). A similar break in the gas velocity profiles is observed (e.g. Corbelli et al. 2014; Kam et al. 2015), however a link between these has not been definitively made.

Whilst the outer gas distribution of M 33 is warped (Rogstad, Wright & Lockhart 1976; Corbelli et al. 2014), likely by a previous minor interaction with M 31 (Semczuk et al. 2018), the disc within 9 kpc appears relatively undisturbed (Quirk et al. 2022). M 33 is a flocculent spiral with two primary spiral arms plus four additional fragmentary arms either side of the centre branching from, and filling in between the primary arms (Humphreys & Sandage 1980). M 33 is generally not categorized as a barred galaxy, however recent observations suggest the presence of a weak bar within the bright central region (Williams et al. 2021; Lazzarini et al. 2022). Whilst there is no strong central bulge in M 33 (e.g. van den Bergh 1991), a nuclear star cluster is present with star formation thought to have occurred there inside the last 40 Myrs (Long, Charles & Dubus 2002; Javadi, van Loon & Mirtorabi 2011). The spiral arms of M 33 can be traced in the distributions of H I (Gratier et al. 2010) and CO (Druard et al. 2014; Braine et al. 2018) emission, giant molecular clouds (GMCs, Corbelli et al. 2017), and bright young clusters (Humphreys & Sandage 1980; Williams et al. 2021). The multiple arms in flocculent galaxies have been suggested to support the model of dynamic spiral formation (Dobbs & Baba 2014) over the quasi-static model (Lin & Shu 1964). GMCs studied in M 33

★ E-mail: d.a.kinson@keele.ac.uk

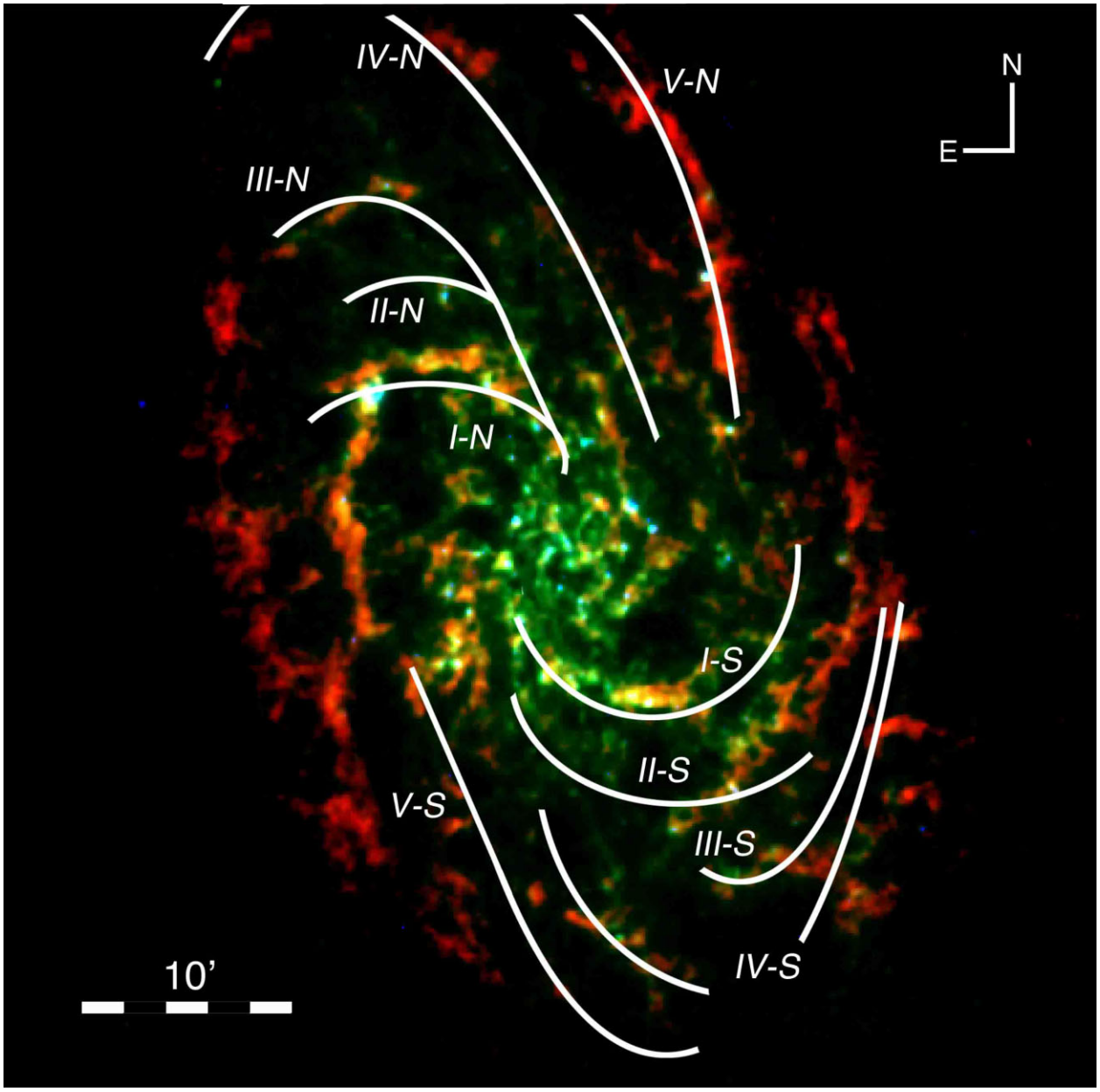

Figure 1. An RGB image of M 33, showing VLA H I (red, Gratier et al. 2010), 250 μm *Herschel*-SPIRE (green, Kramer et al. 2010), 24 μm *Spitzer*-MIPS (blue, Engelbracht, MIPS Science Team & SINGS Team 2004). The figure covers the same footprint as the near-IR WFCAM catalogue of Javadi et al. (2015). The spiral arm identifications, adapted from Humphreys & Sandage (1980), are shown in white.

however show an evolutionary progression which is associated with quasi-static arm models (Corbelli et al. 2017), as gas accumulates at the potential minimum triggering cloud collapse (Lin & Shu 1964).

The arm structure is also well traced by the distribution of H II regions (Humphreys & Sandage 1980; Alexeeva & Zhao 2022). M 33 contains many prominent H II regions, which have have been studied widely across M 33 alongside GMCs (Gratier et al. 2010; Miura et al. 2012; Corbelli et al. 2017; Alexeeva & Zhao 2022). Resolved IR observations of ongoing star formation, i.e. of massive YSOs in M 33, however have not been extended beyond NGC 604 (e.g. Fariña, Bosch & Barbá 2012). NGC 604 is the second most luminous H II region in the Local Group behind only 30 Dor in the LMC (Relaño & Kennicutt 2009; Martínez-Galarza et al. 2012). Star formation in NGC 604 has been well studied at many wavelengths (e.g. Churchwell & Goss 1999; Tabatabaei et al. 2007), including both near-IR studies of individual massive young stellar objects (YSOs) (Fariña et al. 2012) and integrated mid-IR properties (Relaño & Kennicutt 2009; Martínez-Galarza et al. 2012). Triggered star formation events have been theorized in NGC 604 (Tabatabaei et al. 2007; Tachihara et al. 2018), possibly driven by feedback from a population of around 200 O-type stars (Hunter et al. 1996).

Machine learning offers a method by which sources in large multidimensional data sets can be accurately classified. In the Local Group dwarf-irregular galaxy NGC 6822, sites of ongoing star formation were identified from wide-scale near-IR survey data

using probabilistic random forest (PRF) analysis (Kinson, Oliveira & van Loon 2021). A combination of near-IR and far-IR classification features were used in NGC 6822 to separate point sources into multiple object classes. The classifier achieved high levels of estimated accuracy (∼90 per cent) across all classes with that of massive YSOs exceeding this (Kinson et al. 2021). The existence of similar near-IR data covering the M 33 disc (Javadi et al. 2015) offers the opportunity to extend the detailed analysis of ongoing star formation, for the first time, across the entire disc of a spiral galaxy.

The paper is organized as follows. Section 2 introduces the archival data used in this work, Section 3 contains details of our PRF classification method. The results are presented in Section 4, in which the spatial distributions of the different source classes are described and star formation regions (SFRs) are identified. In Section 5, we discuss the properties of YSOs and SFRs identified in our analysis in the context of the galaxy's structure. Finally in Section 6, we summarize our findings.

2 DATA

The description of the data used in our analysis is divided in two parts: catalogues and images used for the PRF object classification (Section 3), and images used for the subsequent analysis of star forming regions across the disc of M 33 (Section 5.1).

2.1 Data for object classification

2.1.1 Near-IR images and point-source catalogue

The near-IR catalogue for M 33 was constructed by Javadi et al. (2015), using data obtained on the United Kingdom Infrared Telescope (UKIRT) using the Wide Field Camera (WFCAM, Casali et al. 2007). Four separate pointing observations were obtained to cover a ∼0.89 deg^2 sky area (∼13 × 13 kpc) at a resolution of 0.4 arcsec per pixel. Multi-epoch observations were made as part of a monitoring programme over dates from 2005 September to 2007 October. More details on the data reduction can be found in Javadi et al. (2015). They retrieved the photometric catalogues for each individual tile and epoch from the public WFCAM Science Archive (WSA)[1], and performed absolute and relative photometric calibration. In our analysis, we make use of their catalogue of mean magnitudes of point sources towards M 33 for source classification (see Section 4.2). The catalogue contains ∼245 000 sources. We set the additional requirement that a source must be detected in all three JHK_s-bands, reducing the number of near-IR sources to ∼163 000 sources. The JHK_s 5σ limiting magnitudes of the catalogue are 21.5, 20.6, and 20.5 mag, respectively. Source density of the catalogue is shown in Fig. 2, and basic photometric properties are shown in a Colour-Magnitude Diagram (CMD) in Fig. 3.

Due to the construction of the catalogue with data taken over multiple epochs and detector pointings, different regions of the science field-of-view reach varying depths. As shown in Fig. 2, the catalogue is uniform to depths of $K_s = 19.2$ mag, beyond which the varying depth between detectors becomes apparent. Whilst these artefacts in the catalogue construction will not affect the accuracy of classification for individual sources; it is important to note when analysing the spatial distribution of sources (see Section 4.3).

[1] http://wsa.roe.ac.uk/

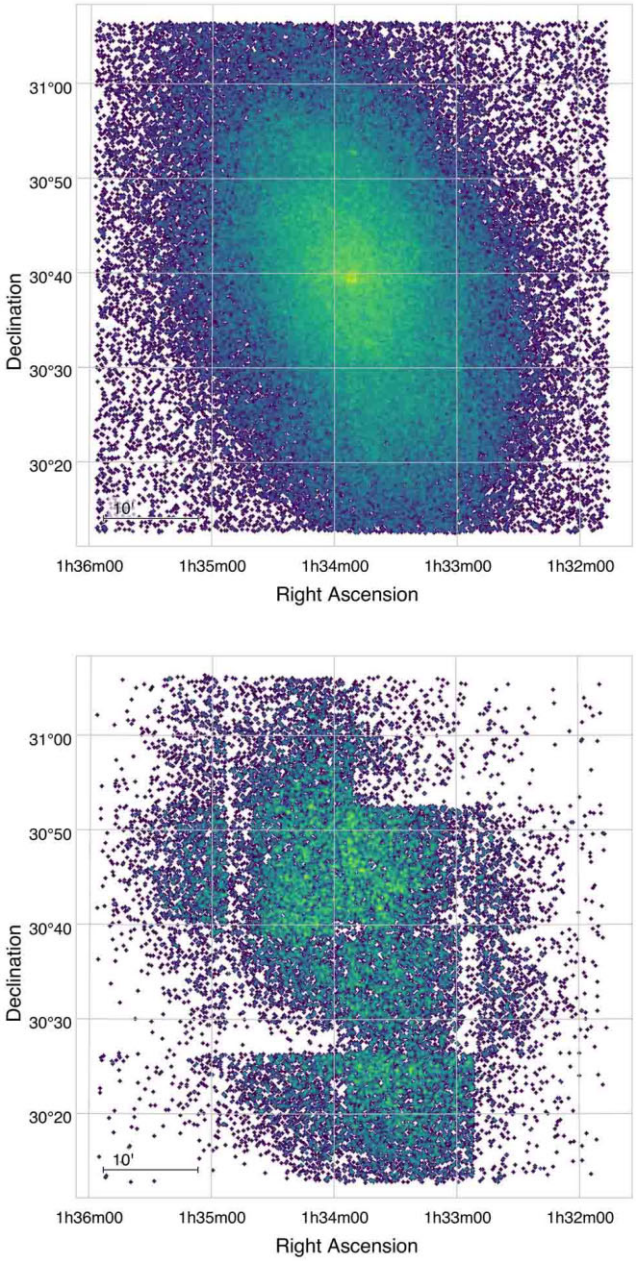

Figure 2. Hess diagrams of source density, brighter (top) and fainter (bottom) than $K_s = 19.2$ mag. The effect of variable depth in the catalogue across the field-of-view is clear at fainter magnitudes.

2.1.2 Far-IR images and measurements

Light emitted by hot young stars at UV wavelengths is reprocessed by surrounding dust and re-emitted at far-IR wavelength (e.g. Bianchi et al. 2012). To provide additional environmental information for each source in the near-IR catalogue, we use the neighbourhood far-IR brightness as an indicator of proximity to star-formation activity (Kinson et al. 2021). To this end, we used 70 and 160 μm images obtained with the Photodetector Array Camera & Spectrometer (PACS, Poglitsch et al. 2010) onboard the ESA *Herschel* Space Observatory (*Herschel*, Pilbratt et al. 2010), obtained as part of the *HERschel* M 33 Extended Survey (HERM33ES, Kramer et al.

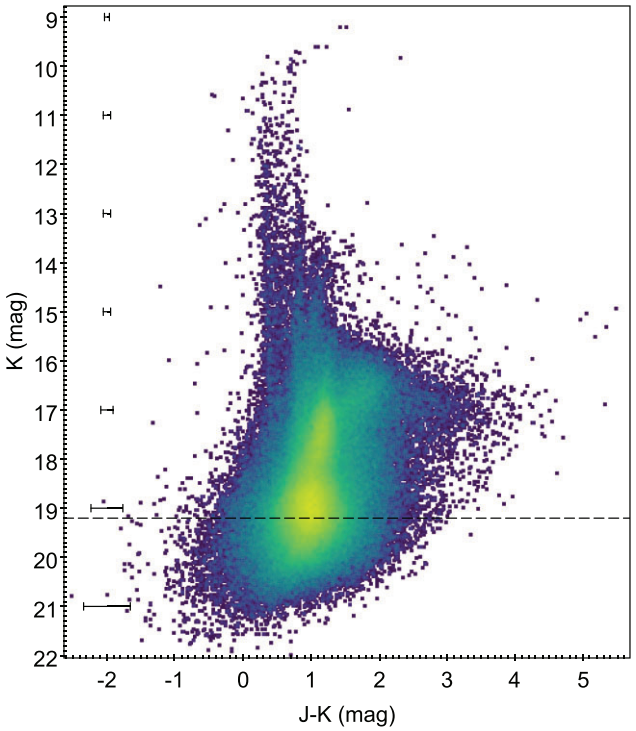

Figure 3. The near-IR catalogue presented in a CMD Hess diagram. Average error bars are shown. The dashed line at $K_s = 19.2$ mag indicates the magnitude at which the catalogue depth becomes very patchy (Fig. 2).

2010). The images were retrieved from the ESA *Herschel* Science Archive.[2]

Point sources located both in NGC 6822 and the Magellanic Clouds (MC) were used to train the PRF classifier (see Section 3.1 for full details). For NGC 6822, the 70 and 160 μm images (Galametz et al. 2010) were also retrieved from the *Herschel* Science Archive, as were the Magellanic Clouds 160 μm images (Meixner et al. 2013). The Magellanic 70 μm images (Meixner et al. 2006; Gordon et al. 2011) were obtained using the Multiband Imaging Photometer for Spitzer (MIPS, Rieke et al. 2004) onboard the *Spitzer Space Telescope* (*Spitzer*, Werner et al. 2004), retrieved from the *Spitzer* Heritage Archive.[3] Small non-astrophysical bias levels in some of the Magellanic images were corrected for as described in Kinson et al. (2021).

At the position of each K_s-band source, an aperture of 30 parsec radius (7.2 arcsec for M 33) was used to measure an average brightness. Photometry was performed using the PHOTUTILS package for PYTHON (Bradley et al. 2020). The size of this aperture is the same as used in NGC 6822 (Kinson et al. 2021), and was chosen based on the scale of emission in the far-IR images and typical molecular cloud scales (e.g. Tan et al. 2014).

2.2 Ancillary data

Archival Hα, 24 μm *Spitzer*-MIPS, and 250/500 μm *Herschel*-Spectral and Photometric Imaging Receiver (SPIRE, Griffin et al. 2010) images are used in our analysis to provide evolutionary information on the star forming regions, as discussed in Section 5.1.

[2]http://archives.esac.esa.int/hsa/whsa/
[3]https://sha.ipac.caltech.edu/applications/Spitzer/SHA/

The Hα images of both M 33 and NGC 6822, retrieved from the NASA/IPAC Extragalactic Database (NED),[4] were taken as part of a survey of Local Group galaxies (Massey et al. 2006); as described in Massey et al. (2007a) the images were reduced and calibrated in a similar way and are therefore directly comparable with one another (see their tables 1 and 2). The *Spitzer*-MIPS 24 μm mosaic images of both galaxies were retrieved from the *Spitzer* Heritage Archive (NGC 6822: Kennicutt et al. 2003; M 33: Engelbracht et al. 2004). The *Herschel* Science Archive provided the 250/500 μm SPIRE images, originally described in Kramer et al. (2010) for M 33 and Galametz et al. (2010) for NGC 6822.

3 PROBABILISTIC RANDOM FOREST (PRF)

A random forest classifier (RFC) is a robust and established tool for classification problems (Breiman 2001). We use an adaptation of the RFC developed by Reis, Baron & Shahaf (2019) called a probabilistic random forest (PRF). The PRF classifier improves on the RFC by taking into account feature uncertainties as well as allowing for the classification of sources with missing data. This both increases the accuracy of the classifier and the number of sources that can be classified (Reis et al. 2019). A more in-depth discussion of the difference in the methodologies for RFC and PRF classifiers is presented in Kinson et al. (2021); we follow their methodology that is summarized below.

To classify the sources a set of six features were used: the near-IR K_s-band magnitude, three near-IR colours ($J-H$, $H-K_s$, and $J-K_s$) and two far-IR brightnesses at 70 and 160 μm. To classify sources the PRF requires a set of sources of known type on which the algorithm is trained. These training set sources are then randomly split into training and testing samples, allowing for an estimate of the classifier's accuracy (see Section 4.1). Splitting is done on a 75 per cent training, 25 per cent test basis with the splitting applied globally to the training set rather than per each individual class. This random splitting can lead to some stochastic effects in the training data selection; these are mitigated by repeating the splitting over many runs with different random seeds. Where one class in the training set is disproportionately large, such that it dominates the randomly selected training sample, the accuracy of the classifier is negatively affected. We took steps to counteract this effect as described in Section 3.2. The following section details the sources selected for PRF training.

3.1 Sources in the training set

The training set for the PRF consists of sources from nine target classes. These are Galactic foreground stars (FG), blue stars, and yellow supergiant stars (BS), red supergiant stars (RSG), oxygen, and carbon-rich asymptotic giant branch stars (OAGB and CAGB), red giant branch stars (RGB), Wolf–Rayet stars (WR), massive young stellar objects (YSOs), and finally unresolved background galaxies (AGN). Other classes of objects are present in the M 33 stellar population, but they are either dissimilar enough from YSOs that their misclassification will not contaminate the YSO sample or are rare (e.g. planetary nebulae), and so will not significantly impact the purity of the classified YSO sample. In the case of planetary nebulae, a PRF classification in NGC 6822 misclassifies the few examples as AGN (see section. 3.4.7 of Kinson et al. 2021); including classes of rare objects would

[4]https://ned.ipac.caltech.edu

however adversely affect the accuracy of the PRF classifier (see Section 3.2).

The detailed selection criteria for each class is given in the following subsections. To maintain the purity of the training set stringent selection criteria are set, requiring sources identified in the literature to have been classified on the basis of methods other than broad-band photometry, e.g. spectroscopy, narrow-band indices, or *Gaia* proper motions. In most instances, however, the catalogues from which training set sources are drawn that do not completely cover the area of the near-IR catalogue. The M 33 sources in the training sample were crossmatched to the near-IR catalogue, using a radius of 0.5 arcsec.

Most classes include exclusively sources in M 33 with the exception of the AGN and RGB that also include sources behind the MCs and in NGC 6822. Training set YSOs come exclusively from the MCs and NGC 6822. The near-IR data for NGC 6822 and M 33 are however comparable. We therefore believe that while the near-IR catalogue to be classified may be affected by source blending, such effects are on the whole also present in the training set data, providing the PRF with effective examples on which to learn.

3.1.1 *Foreground Galactic sources*

The training set of Galactic foreground contaminants includes sources from Massey, Neugent & Smart (2016) with optical spectra consistent with Galactic dwarfs. They separate foreground dwarfs from B-, A-, F-, and G-type supergiants by the shape and strength of their Balmer series lines, and the differing strengths of metallic lines (Si, Ca, K, Ti, Mg, and Sr). Additionally, we include near-IR sources with a *Gaia* EDR3[5] (Gaia Collaboration et al. 2020) counterpart if their proper motion is greater than 0.5 mas yr^{-1} in both RA and Dec components. Near-IR colour cuts at $0.3 < J - K_s < 0.9$ mag, defined using TRILEGAL foreground simulations (Girardi et al. 2005) towards M 33, are then applied to remove spurious chance matches between the *Gaia* and near-IR catalogues. Whilst Galactic sources may be found outside these cuts to ensure purity of the FG training set we select only sources in the conspicuous vertical foreground sequences (see Fig. 4).

Foreground sources identified spectroscopically or with *Gaia* proper motions extend only to $K_s \sim 16.5$ mag; in order to accurately train the PRF however, foreground sources at magnitudes down to the limit of our near-IR catalogue ($K_s \sim 20.5$ mag) are needed. For this purpose, we used the foreground population simulated with TRILEGAL already mentioned. The simulated foreground source magnitudes were perturbed in *J*-, *H*-, and K_s by an amount consistent with the average error bar in the near-IR catalogue at similar magnitudes. Foreground stars have no preferential location in the field of view, therefore, to generate far-IR measurements for these sources, apertures were placed randomly in the far-IR images and measurements taken as described in Section 2.1.2.

3.1.2 *Active galaxies*

Active galactic nucleus (AGN) have been shown to be significant contaminants in near-IR YSO samples due to their colour similarities (e.g. Sewiło et al. 2013; Jones et al. 2017). The strength of the far-IR emission as a measure of the proximity to star formation activity can help differentiate YSOs from contaminants such as AGN, as shown by Kinson et al. (2021). We start from the AGN training sample from Kinson et al. (2021), which is comprised of 89 background galaxies behind the SMC. This sample is classified using a variety of data across multiple wavelengths including X-Ray, UV, near-IR, and radio (Pennock et al. 2021). This AGN sample was augmented with 36 sources behind M 33 taken from the latest update of the MILLIQUAS compilation (the Million Quasars Catalog, version 7.2, Flesch 2021).

[5]https://www.cosmos.esa.int/web/gaia/earlydr3

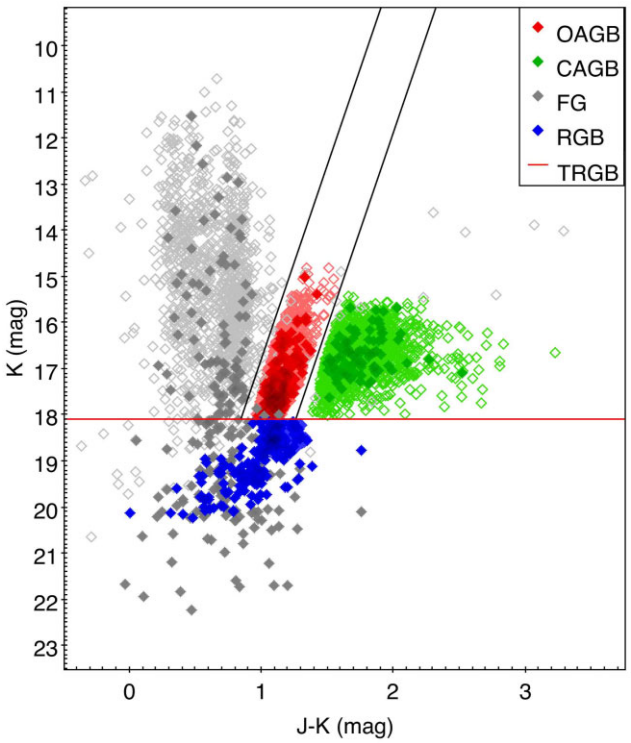

Figure 4. A CMD showing the four large classes, which were down-sampled with the full set of data shown by open symbols and the down-sampled data by filled symbols. The parameter space for each class is well represented by the down-sampled data. The TRGB magnitude ($K_s = 18.11$ mag) and AGB colour-cuts adapted from Ren et al. (2021) are shown by the red and black lines, respectively.

3.1.3 *Asymptotic giant branch stars*

Asymptotic giant branch stars (AGBs) can display near-IR colours and magnitudes similar to bright massive YSOs. OAGBs and CAGBs have distinct magnitude and colour properties due to the composition of their circumstellar dust envelopes (see Fig. 4) and thus are classified independently.

The AGB sample is based on the catalogue of *V* and *I* broadband and TiO and CN narrowband photometry towards M 33 (Rowe et al. 2005). Using $V - I$ and CN − TiO colour cuts defined by Rowe et al. (2005), we identified both OAGBs and CAGBs from their catalogue. Both classes of AGB have $V - I > 1.8$ mag with OAGBs having colours of CN − TiO < -0.2 mag and CAGBs CN − TiO > 0.3 mag. From this sample, we remove any sources with *Gaia* proper motions consistent with a Galactic foreground dwarf (see Section 3.1.1). Sources with any spectroscopic classification of another type from Massey et al. (2016) are also removed.

Ren et al. (2021) define near-IR colour and magnitude boundaries for both OAGB and CAGB sources in M 33 (see their Fig. 10), which we adopted to refine the samples from Rowe et al. (2005). These cuts

are shown in Fig. 4. We also select only sources brighter than the M 33 tip of the RGB (TRGB) magnitude, $K_s = 18.11$ mag (Ren et al. 2021). Finally for the OAGBs we apply an upper magnitude limit at $K_s = 14.8$ mag, which includes all variable AGB sources identified in Javadi et al. (2015) and thermally pulsing AGB models from Ren et al. (2021).

3.1.4 Red giants and supergiants

RGB stars are a significant population that exhibit similar colours and magnitudes to faint YSOs in M 33. We began with M-type stars identified in Rowe et al. (2005) as described in subsection Section 3.1.3. Sources were rejected from the RGB sample if their K_s-band magnitude was brighter than the TRGB magnitude ($K_s = 18.11$ mag, Ren et al. 2021). Since RGBs and Galactic foreground sources overlap in colour-space at fainter magnitudes (e.g. Kinson et al. 2021), we reject any source with a *Gaia* proper motion consistent with a Galactic star (see Section 3.1.1). A colour cut was made at $J - K_s > 0.8$ mag to remove spurious near-IR matches; this value was selected based on the TRILEGAL (Girardi et al. 2005) Galactic foreground simulation mentioned in Section 3.1.1.

The Rowe et al. (2005) sample includes only RGBs brighter than $K \sim 18.9$ mag. Therefore, the sample was augmented with additional spectroscopically confirmed fainter RGB sources from NGC 6822 ($\mu_{\rm NGC\,6822} = 23.34$ mag, Jones et al. 2019; $\mu_{\rm M33} = 24.67$ mag, de Grijs & Bono 2014), using the RGB training set compiled in Kinson et al. (2021). The process by which RGBs from both galaxies are combined to form the training class is discussed further in Section 3.2.

RSG stars are a young population (∼10–30 Myrs, Britavskiy et al. 2019) which may contaminate the brighter end of a YSO sample. They can be dusty and due to their relative youth are located near sites of star formation (e.g. Hirschauer et al. 2020; Kinson et al. 2021). RSGs were identified from optical and IR photometry using machine learning techniques in Maravelias et al. (2022), however as these sources lack further confirmation, such as spectroscopy, they are not included in the training set in order to maintain its purity. We adopt spectroscopically confirmed RSGs from the catalogue of Massey et al. (2016), confirmed based on their radial velocities and the presence of a strong Ca II triplet in their spectra. Using the RSG training set employed in NGC 6822 (Kinson et al. 2021) as a guide, colour cuts at $0.4 < J - K_s < 2.5$ mag were made to remove a small number of spurious near-IR matches.

3.1.5 Blue stars

We include a class for bright and bluer stellar sources in M 33. These include bright main-sequence stars as well as other classes not numerous enough to warrant a separate class; these are labelled collectively as 'blue stars' (BS) in our classification scheme. These stars represent a younger population in M 33 compared to e.g. AGB or RGB classes. A machine learning based photometric identification of these populations is presented in Maravelias et al. (2022) however, as with RSGs (see Section 3.1.4) we cannot utilize their catalogues to populate our BS class due to the lack of higher level classification.

The BS class is populated with spectroscopically confirmed O-, B-, and A-type main-sequence stars from the catalogues of Massey et al. (2016). Main-sequence stars were sorted into their spectral types based on the relative strengths of Balmer lines (Hδ, Hγ, and Hβ) and the presence and ratio of He lines. Additionally, we include sources they classified as Luminous Blue Variables (LBV), yellow supergiant stars (YSGs), and H II regions. Massey et al. (2016) separate LBVs from unresolved H II regions based on the presence of strong Fe II lines (Massey et al. 2007b). YSGs were identified using radial velocities and the presence of the O I triplet at $\lambda \sim 777.4$ nm to separate YSGs from foreground yellow dwarfs (Drout, Massey & Meynet 2012). Further colour cuts are set at $-0.5 < J - K_s < 0.3$ mag.

3.1.6 Wolf–Rayet stars

Wolf–Rayet (WR) stars are a relatively rare population with only ∼200 confirmed across the disc of M 33 (Neugent & Massey 2011); they can present near-IR colours similar to those of YSOs and are often located close to regions of ongoing star formation (Massey et al. 2007b; Fariña et al. 2012). Therefore, WR stars can contaminate YSO samples and are included in our classification scheme. The WR training set is comprised of spectroscopically confirmed sources from the catalogues of Massey et al. (2016) and Neugent & Massey (2011).

3.1.7 Young stellar objects

Our training sample of YSOs contains sources from both the Magellanic Clouds and NGC 6822. YSOs with scaled near-IR magnitudes brighter than the detection thresholds in Section 2.1.1 were selected from catalogues of spectroscopically confirmed YSOs, Oliveira et al. (2013) for the SMC and Jones et al. (2017) for the LMC. Near-IR data for these sources were transformed from the native IRSF photometric system (Kato et al. 2007) to the WFCAM photometric system as detailed in Kinson et al. (2021). This resulted in 69 LMC and 26 SMC sources for the YSO training class. We further include 55 YSOs in NGC 6822. These YSOs were first identified in Jones et al. (2019) and Hirschauer et al. (2020) using mid-IR photometry and spectral energy distribution (SED) fitting with evolutionary models (Robitaille et al. 2006; Robitaille 2017), and confirmed using machine learning techniques (Kinson et al. 2021).

3.2 Down-sampling of large training classes

When one or more particularly numerous classes dominate the training set, the classifier training is faced with many more examples of those classes to the detriment of sparser classes. Hence, the balance of class sizes in the training set affects classifier performance (e.g. Khoshgoftaar, Golawala & Hulse 2007; More & Rana 2017). Due to real astrophysical population differences as well as the varied selection methods, the number of sources available for each class vary from 85 for WR to ∼7000 for FG. To ensure the PRF has the highest possible accuracy across all classes it was necessary to down-sample the four most numerous training set classes, FG, RGB, OAGB, and CAGB. The positive effect of the down-sampling on classifier accuracy is shown in Section 4.1.

The RGB training sources come from two sets of data, one in M 33 and another from NGC 6822 (see Section 3.1.4). Given the very different properties of these two galaxies (namely in terms of total stellar mass: $M_{*\rm NGC6822} \sim 1.5 \times 10^8 {\rm M}_\odot$, Madden et al. 2014; $M_{*\rm M33} \sim 5.5 \times 10^9 {\rm M}_\odot$, Corbelli et al. 2014; Kam et al. 2017), and vastly different source density in CMD/CCD parameter space of confirmed RGB sources in each galaxy, these two RGB populations cannot just be added without introducing non-astrophysical biases that would affect the classifier performance. It was therefore necessary to down-sample the RGB sample from M 33 to be more comparable with that of RGBs from NGC 6822. This was done

Table 1. Number of sources for each class for the five training sets (see Section 3.1) after down-sampling of large training classes (see Section 3.2).

PRF class	Number TS sources
YSO	150
OAGB	172
CAGB	91
AGN	125
FG	283
RGB	200
RSG	180
BS	347
WR	85
Total Sources	1631

by comparing the fraction of NGC 6822 RGBs above and below the M 33 sample cut-off when scaled to the same distance (see Section 3.1.4). Reducing the M 33 RGB subsample by a factor of 1 in 24 provides homogeneity in the combined NGC 6822 and M 33 RGB subsamples across the M 33 RGB cut-off. For simplicity, the same down-sampling factor was applied to the other large classes (FG, OAGB, and CAGB).

It is somewhat inevitable that down-sampling of the large classes will introduce some stochastic selection effects. Such effects, as well as those resulting from the train/test splitting of the sample, are counteracted by repeating both down-sampling and train/test splitting multiple times. For classes which cover a large range of magnitudes such as the FG class, we checked that the down-sampling still adequately samples the parameter space (see Fig. 4).

In total, we performed the down-sampling of the four larger classes randomly five times to create different training sets for the PRF. This number was selected based on achieving a stable number of YSOs recovered in common with each down-sampled training set (see Appendix A online only material). The number of sources in each training set class are given in Table 1. Each training set was used to train a PRF classifier which was run 20 times with different random seeds for the train/test split, totalling 100 runs.

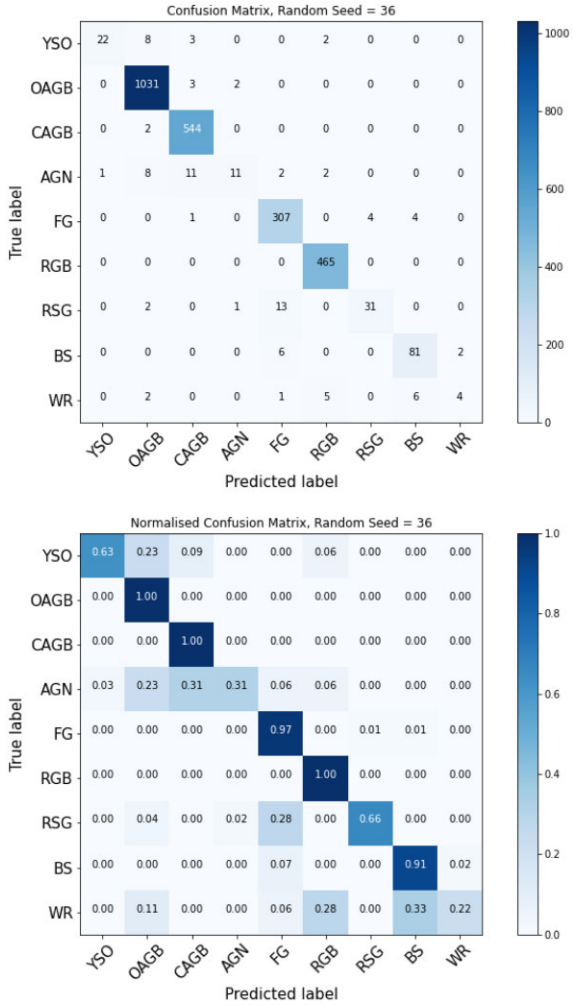

Figure 5. Non-normalized (top) and normalized (bottom) confusion matrices for an example PRF run with no class down-sampling (see text). The large classes achieve high accuracy, however for the smaller classes high levels of confusion are evident.

4 RESULTS

Using the training set defined in Section 3.1, the PRF classifier is applied 100 times to the 162 746 sources remaining in the catalogue.

4.1 Confusion matrices

Confusion matrices provide a helpful visualization of the classifier's accuracy. Each matrix shows the PRF classification of the 25 per cent of sources in the test set classified using the remaining 75 per cent of training set sources. In Fig. 5, the confusion matrices, both non-normalized and normalized, show the accuracy of a PRF classifier using the training set without any down-sampling applied. High-classification accuracy is achieved for the large classes to the detriment of all other classes: sources from the smaller classes are often misclassified into the four large classes. In particular for YSOs, without down-sampling the PRF achieves accuracies ranging from 55 to 75 per cent across the 100 runs with a median value of 66.5 per cent.

In Fig. 6, we show the PRF matrices, using the same random seed as those shown in Fig. 5 with down-sampling applied as described in Section 3.2. In general, an improvement in the overall PRF classification accuracy, exemplified by the strong diagonal feature in the normalized matrix is evident. In particular, the YSO classification accuracy significantly improves, ranging from 62 to 97 per cent with a median value of 82 per cent across all runs. Across the 100 PRF runs the median class-averaged accuracy is 87 per cent. The estimated accuracy per PRF is skewed by the WR class which performs significantly worse than all others by a large margin (see Fig. 6); we discuss source misclassification and contamination in the following section.

4.1.1 Potential misclassifications and class contamination

As already mentioned, YSOs in the training set are recovered with high accuracy (median accuracy of 82 per cent). More specifically 67 PRF runs achieve an YSO accuracy of over 80 per cent, and only 4 runs have accuracy below 70 per cent. Misclassified training set YSOs are most often placed into the OAGB, RGB, and WR classes. Some OAGB, RGB, and dusty WR stars have similar near-IR colours and magnitude to YSOs, which is the likely cause for the confusion in the PRF's classification. Additionally WR stars are likely to be associated with sites of bright far-IR emission (e.g. Fariña et al. 2012) similar to YSOs.

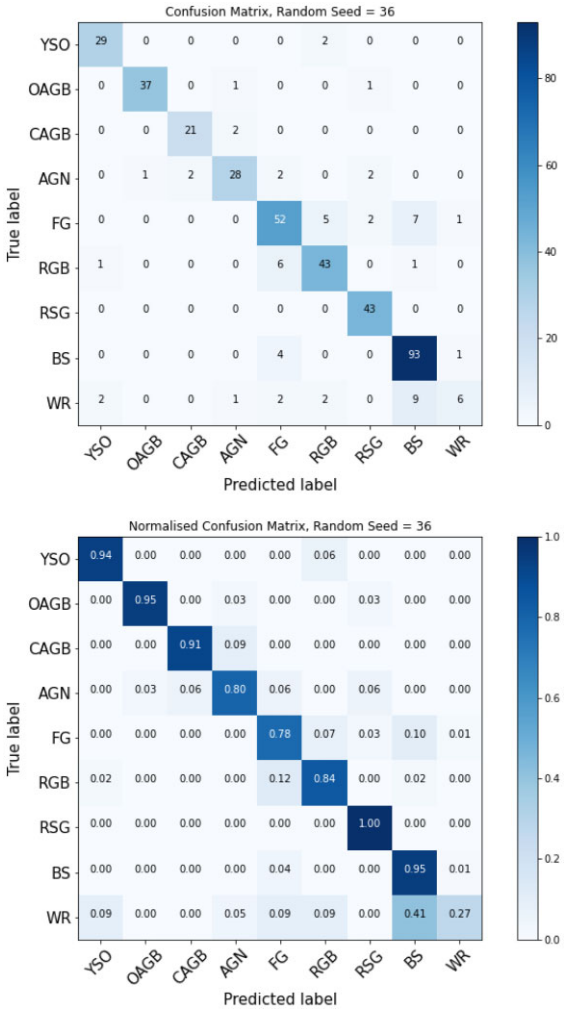

Figure 6. Non-normalized and normalized confusion matrices (respectively top and bottom) for the PRF run using the same random seed as those shown in Fig. 5, but here with class down-sampling (see text). The misclassifications for the smaller classes are very effectively reduced.

The YSO class suffers from very low levels of contamination from other classes; the highest fraction of incorrectly classified YSOs in the test sample are WRs due to the similarities noted above. Dusty WRs are however relatively rare, therefore the absolute contamination of YSOs remains very low. The opposite happens for the RGB class: their fractional contamination to the YSO class is low, however they are very numerous, meaning RGBs can still be important contaminants of the YSO sample. We use training set sources that after down-sampling are returned to the main catalogue to further investigate YSO contamination in the final classifier output (Section 4.2).

As noted previously, WR is the worse performing class. This class has the fewest training sources available (85 sources), and is misclassified into AGN, BS, FG, RGB, and YSO classes. Of these, the BS class is the dominant misclassification in some runs even outscoring the correct classification (see Fig. 6). The lower performance of the PRF in WR classification is a consequence of previously discussed similarities to other classes and the small training set size for this class.

For the AGNs, we see some confusion with the OAGB and CAGB classes, likely due to the fact that AGN can have near-IR colours similar to those of the AGB populations (Hony et al. 2011; Pennock et al. 2022). A similar effect was also seen in AGN classifications behind NGC 6822 (Kinson et al. 2021).

Table 2. Number of sources in M 33 classified into each PRF class and total number sources including those from the training set after down-sampling of the largest classes (see Section 3.2).

PRF Class	Classified sources	Training & classified sources
YSO	4985	
OAGB	18214	18387
CAGB	2086	2177
AGN	3757	3793
FG	5294	5577
RGB	27422	27498
RSG	1424	1604
BS	3111	3458
WR	82	167

4.2 Final classifier outputs

Each of the individual 100 PRF runs provides a classification for all sources not included in the training/testing sets. These 100 classifications provide a score between 0 and 100 for each source and for each class, $n_{\rm class}$ with class = YSO, FG, etc. The PRF classifies 41 per cent of sources into the same class over all runs (i.e. max $(n_{\rm class}) = 100$). These sources are included in our subsequent analysis, and are henceforth referred to as classified. The breakdown of the 66 378 classified sources into the different PRF classes is given in Table 2.

In Fig. 7, we present a CMD, colour–colour diagram (CCD), and far-IR brightness plot for both training/testing set data and classified sources. The plots show that, for every class, training and classified sources occupy a similar position in parameter space. Whilst both training and classified YSOs cover a similar range of $J - K_s$ colours from 0.5 to 5 mag, and K_s-band magnitudes from 16 to 21 mag at magnitudes fainter than $K_s = 19.5$ mag classified YSOs are seldom redder than $J - K_s = 2.5$ mag. This is primarily due to the fact that some training set YSOs can have J- and H-band magnitudes fainter than the near-IR catalogue's detection thresholds (see Section 2.1.1). This arises from practical considerations in the design of the near-IR observations with shorter wavelength images not deep enough to characterize the redder sources being these YSOs or AGBs. Therefore, the faintest YSOs we identify are not particularly red and, as expected, no classified YSOs are found outside the colour and magnitude ranges described by the training set YSOs.

Fig. 7 also shows that whereas in the training set there is a region of the CMD occupied by both OAGBs and CAGBs brighter than $K_s = 16$ mag in the classified sources this region is dominated by AGN classifications. These sources are likely missclassified due to the confusion between these classes commented upon in Section 4.1.1.

We discussed potential YSO contamination in Section 4.1.1. The confidence matrices however only provide the likelihood of contamination for a single PRF run; for a source to effectively become a contaminant of the YSO class, it needs to be consistently classified in that class 100 times. We use the sources from the training set that are returned to the catalogue for classification to quantify such effects for the most numerous astrophysical classes. In total, 655 sources are excluded from the RGB, FG, OAGB, and CAGB training sets after down-sampling (see Section 3.2). These sources with known

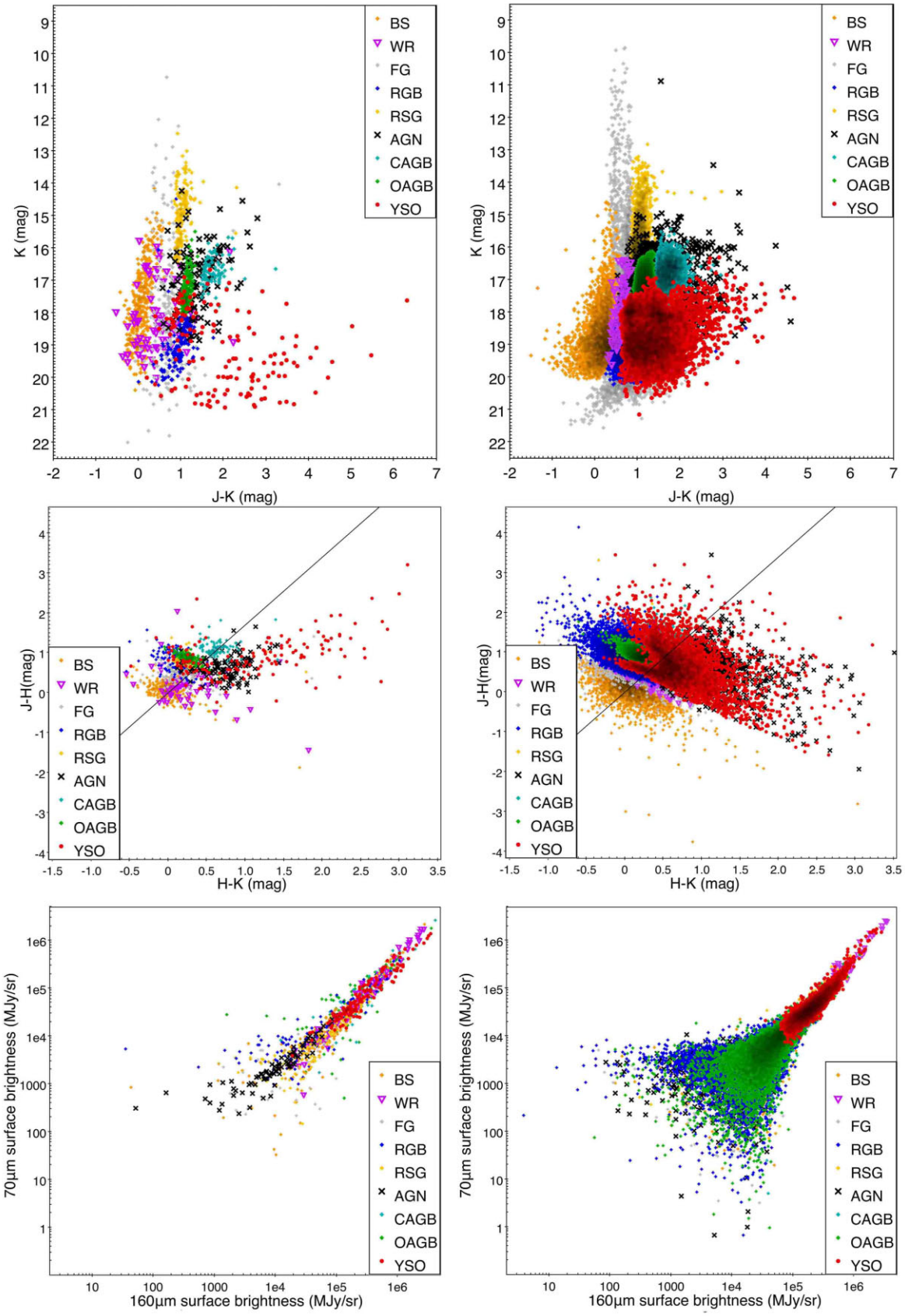

Figure 7. CMD, CCD, and far-IR brightness plots of the training set sources (left) and for the classified sources (right). Colour-coding is given in the legend. The reddening line shown in the CCD plots is derived using the coefficients from Rieke & Lebofsky (1985).

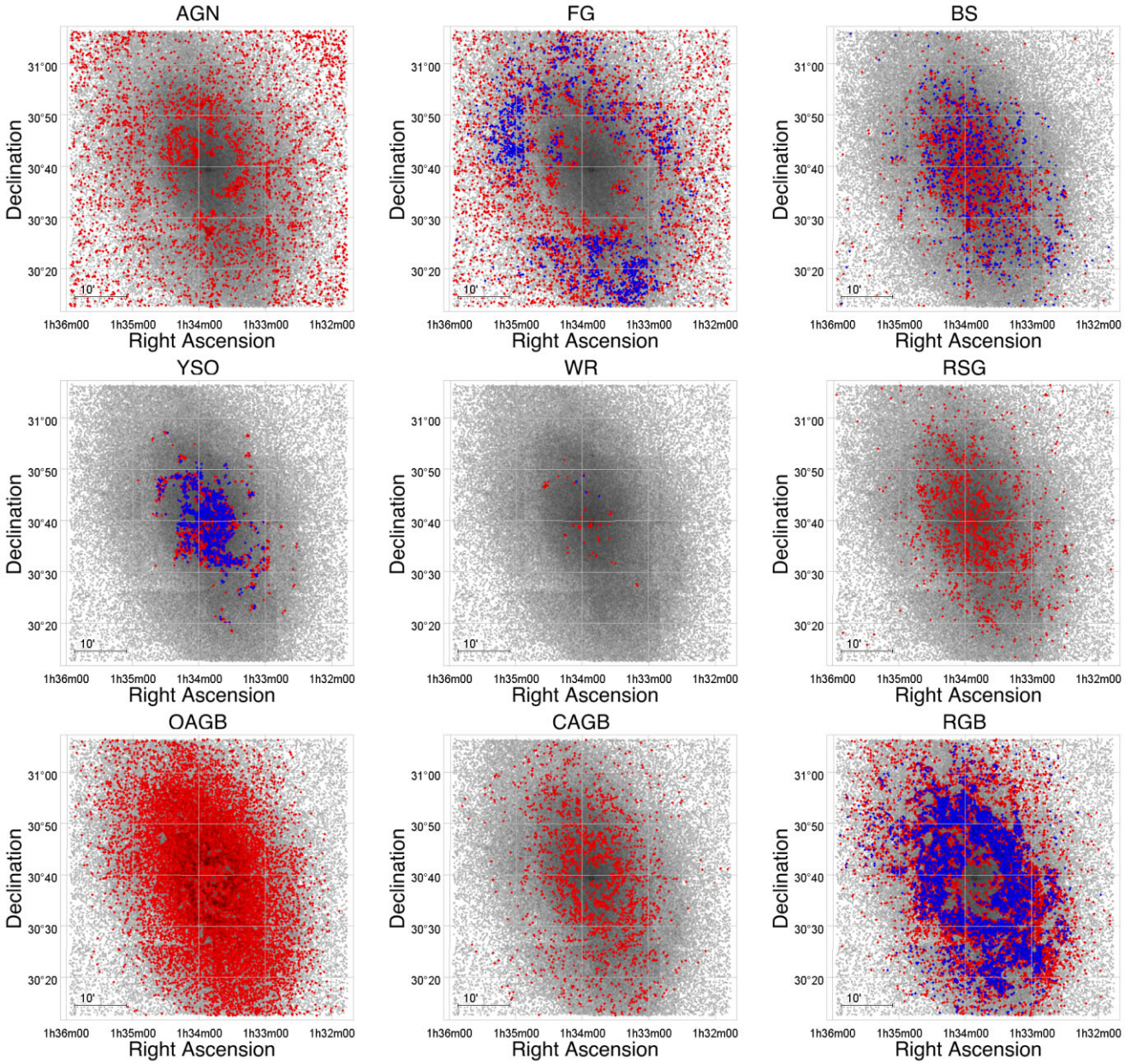

Figure 8. Spatial distributions for each PRF class. Sources with $K_s < 19.2$ mag and $K_s \geq 19.2$ mag are shown, respectively, in red and blue. The full catalogue is shown in the background.

classification are used to provide an additional estimate of class contamination alongside the statistics provided by the confusion matrices (Section 4.1). Of these 655 sources, 358 (∼55 per cent) are classified by the PRF with 330 assigned to the correct literature class (i.e. 92 per cent of classified sources are correctly classified). The 28 incorrectly classified sources (seventeen OAGBs, three CAGBs and twelve FGs) are misclassified as sixteen RSGs, eight AGNs and four BSs. None of these sources are classified as a YSO. Noteworthy is the fact that despite the considerations discussed in Section 4.1.1, none of the RGBs are misclassified, as YSO or any other class.

4.3 Spatial distributions

As noted in Section 2.1.1, sensitivity issues become apparent for $K_s > 19.2$ mag. The effects of source crowing increase significantly towards the centre of the galaxy (central $\sim 7 \times 7$ arcmin^2 region), since evolved star density profiles decrease as a function of radial distance (e.g. Rowe et al. 2005; Williams et al. 2021). Due to crowding the PRF's, classifications are less certain in the central region with a larger fraction of sources being assigned $n_{\rm class} < 100$, effectively remaining unclassified by the PRF. In the central region, 30 per cent of sources are classified compared to 42 per cent in the outer regions. While crowding affects the identification for all classes, classes dominated by fainter sources are more severely affected.

In Fig. 8, we show the spatial distributions of classified sources for each class. We briefly highlight some salient features of non-YSO distributions, however a thorough discussion is beyond the scope of this paper. We discuss the YSO distribution in Section 5.1.

Table 3. Catalogue of YSOs in M 33 classified using the PRF analysis. For YSOs assigned to a SFR by the DBSCAN analysis, the SFR ID is given. YSO mass estimates are discussed in Section 5.2. A sample of the table is provided here, the full catalogue is available as supplementary material.

RA (J2000) h:m:s	Dec (J2000) deg:m:s	J mag	$J_{\rm err}$ mag	H mag	$H_{\rm err}$ mag	K mag	$K_{\rm err}$ mag	SFR ID	mass $M_\odot$
01:33:49.16	+30:40:17.7	18.48	0.066	17.89	0.047	16.99	0.051		13.9
01:34:10.32	+30:36:40.7	19.17	0.061	18.28	0.078	17.28	0.057	26	12.9
01:34:06.18	+30:37:47.3	19.03	0.054	17.77	0.050	16.35	0.039	39	19.8
01:33:48.66	+30:44:48.3	19.48	0.143	18.55	0.107	18.04	0.087	48	20.1
01:33:37.54	+30:36:02.1	21.13	0.260	20.25	0.306	19.85	0.318	56	9.4

AGN and FG sources are fairly, evenly distributed across the field as expected. AGNs are not identified in the crowded central region of M 33, since the increased point-source density and brighter completeness limit there make it very difficult to identify background sources. Furthermore, as noted in Sections 4.1.1 and 4.2, there is some confusion between the AGN and AGB classes. These effects are most apparent in the centre of M 33, where the AGN distribution appears less uniform than in the outer regions. The FG class shows some correlation with the overall catalogue source density outside the centre of M 33, especially at fainter magnitudes. This behaviour is reversed in the central region, where FG sources are seldom classified consequence of the crowding and associated completeness issue.

The AGB and RGB classes show distributions throughout the disc of M 33 in agreement with the source density distributions previously reported (Javadi et al. 2015; Williams et al. 2021). We recover the faint two arm morphology seen in the RGB and combined OAGB and CAGB distributions in the inner $\sim$20 $\times$ 10 arcmin^2 region (Williams et al. 2021). The CAGB class does not exhibit a source density increase towards to the centre of M 33, as is seen in the OAGBs, in agreement with the density profiles observed by Rowe et al. (2005). The strong ring-like CAGB structures (at $\sim$3.5 kpc from the centre of M 33, Block et al. 2004, 2007) are not seen in our analysis. As already mentioned, in the central region, crowding affects the PRF classification with fewer classified faint sources present, as seen in particular for the RGB distribution.

The BS and RSG classes represent stellar populations younger than AGB and RGB classes. Their distributions are highly structured, more closely associated with the spiral arms. For the BS class, this morphology is in general agreement with the distribution of the young main-sequence population in the central region of M 33 (Williams et al. 2021, MS distribution in their figure 22). The RSG distribution closely resembles that found by Massey et al. (2021, see their fig. 11) and Ren et al. (2021, see their fig. 11). The WR source distribution, even though very sparse, loosely follows the distribution of YSOs (Section 5.1).

4.4 YSO distribution and clustering

The PRF identifies 4985 YSOs across the disc of M 33; their properties are listed in Table 3, and their distribution is shown in Fig. 9. As already discussed, the PRF classifies $\sim$30 to 42 per cent of sources in the catalogue; therefore this YSO sample is robust but unlikely to be complete. The YSO sources are found mostly in the central region of the galaxy and on the two major spiral arms of M 33 (I-N and I-S). Arms I-N and I-S contain $\sim$300 YSOs, each with a similar total YSO mass (see Section 5.2 for details on YSO mass estimates). The area adjacent to the base of I-S in which many YSOs are found is the base of arm IV-S (Humphreys & Sandage 1980). A small number of YSOs lie further along the other spiral arms.

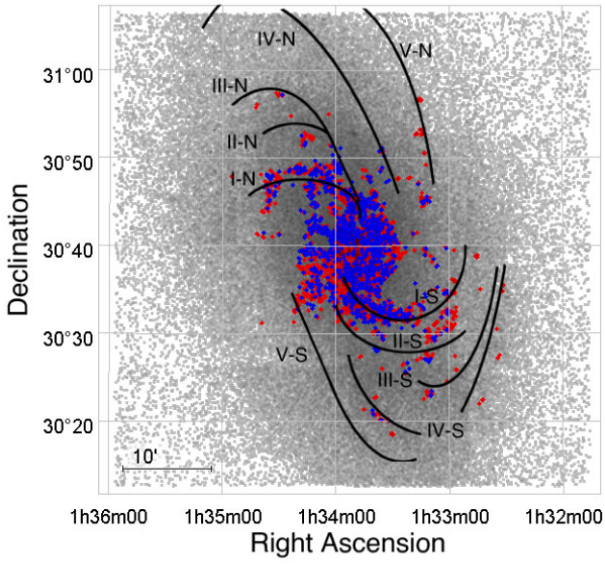

Figure 9. YSO distribution in M 33 with the spiral structure adapted from Humphreys & Sandage (1980) overlaid (colour-coding as in Fig. 8).

We identify SFRs in M 33 by examining the spatial clustering of classified YSOs. These YSO clusters were identified using a density-based spatial clustering of applications with noise (DBSCAN, Ester et al. 1996). DBSCAN is a clustering algorithm which finds density-based associations in spatial data. This process was performed using deprojected coordinates (see Appendix A, online only material, for details).

DBSCAN requires two parameters that can be tuned to the data: a minimum number of YSOs in a cluster and a distance parameter ϵ, the furthest distance at which a neighbour is selected. The minimum YSO number is set to eight, selected to avoid splitting the most apparent clusters and consistent with the value used in a similar analysis in NGC 6822 (Jones et al. 2019). We optimized the choice of ϵ using a k-nearest neighbours (k-NN) method. It analyses the distances between individual YSOs and finds the 'elbow-point' in the distance distribution which is the optimal value for ϵ (Rahmah & Sitanggang 2016).

The initial run of DBSCAN ($\epsilon = 0.1551$) identified 23 clusters but was unable to identify clusters in the central region of M 33, where the source density is much higher. To recover additional clusters, the process was repeated with progressively smaller ϵ values using those YSO sources that remained unassigned (see Table 4). This process was repeated five times, after which the ϵ distance returned by the k-NN analysis effectively plateaued. Overall, DBSCAN identifies 62 YSO clusters.

Table 4. ϵ distances used in the DBSCAN clustering analysis and the cumulative number of clusters recovered after each step (see text).

ϵ (kpc)	Identified clusters
0.1551	23
0.1064	41
0.0885	50
0.0852	58
0.0824	62

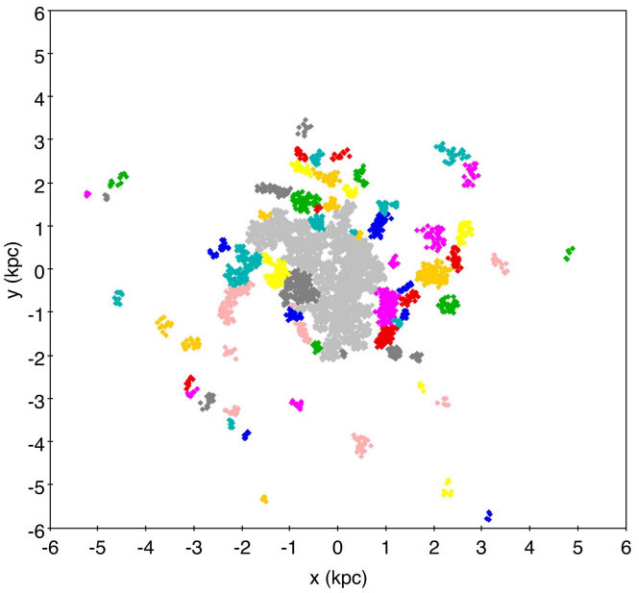

Figure 10. Clusters of YSOs identified by DBSCAN, displayed in deprojected coordinates. The central region (see text) without identified clusters is shown in light grey colour. This projection is rotated by 90 degree clockwise with respect to the sky coordinates shown in Fig. 9.

A visual inspection of the YSO source distribution revealed a small number of additional YSO clusters that did not meet the DBSCAN criteria. One example is the H II region IC 133, which has many indicators of massive star formation such as H_2O and OH maser emission (respectively Churchwell et al. 1977; Staveley-Smith et al. 1987), but was not identified by DBSCAN due to its nine YSOs being spread across a larger area (131 pc or 32 arcsec). Six more clusters were identified by eye. A total of 68 YSO clusters (henceforth referred to as SFRs) were identified across the disc of M 33, ranging in size from 31 to 550 pc (7.5 to 132 arcsec) and containing between 3 and 211 YSOs. The radii of the SFRs are broadly consistent, albeit at the higher end with the GMC sizes in M 33 analysed by Corbelli et al. (2017); we discuss the relationship between SFRs and GMCs in Section 5.1.3. The SFR spatial distribution in deprojected coordinates is shown in Fig. 10. The centre of each SFR is defined as the average of the members' positions and its radius is the largest distance from this average position. This definition of SFR size is consistent with that used by Jones et al. (2017) in NGC 6822, allowing for a direct comparison of SFR properties in both galaxies (see Section 5.1). SFR properties are listed in Table 5.

As discussed previously, in the central dense region of M 33 DBSCAN was unable to recover YSO clusters. In total 1986, YSOs were assigned to a SFR listed in Table 5, 562 were unclustered and 2437 were left in the central dense 'remnant' ($\sim$11.6 $\times$ 10.4 arcmin^2 or 2.8 $\times$ 2.5 kpc^2 in size, light grey in Fig. 10). In general, the PRF works less well in this central region with only 30 per cent of sources classified as opposed to 41 per cent overall. As already discussed, we identify fewer than expected RGB sources in this region (see Fig. 8 and Section 4.3). Given their expected distribution in the M 33 disc and strong overlap in colour-magnitude space with YSOs (see Fig. 7), RGBs are an important contaminant class (see Section 4.1.1), even if no known RGBs are misclassified as YSOs by the PRF (Section 4.2). Nevertheless, assuming in extremis that all 811 YSOs overlapping the RGB region of the CMD space are contaminants, we estimate that at most 30 per cent of YSOs could be wrongly classified in the central region. We take this into account in the analysis in Section 5.2.

In Fig. 11, the number of YSOs per SFR and the size of each SFR are shown against the deprojected radial distance to the centre of M 33: the largest and more numerous clusters are found closer to the centre of the disc.

5 DISCUSSION

5.1 The star forming regions in M33

In this section, we discuss the observed properties of SFRs in M 33 and discuss their evolutionary status, using SFRs in NGC 6822 (analysed using similar methods) as a benchmark.

5.1.1 SFR observed properties

Integrated optical to far-IR brightnesses can be used to characterize and probe the activity in SFRs. Hα emission in SFRs arises from unobscured massive YSOs and young massive stars, whilst emission at 24 μm traces warm dust associated with recent star formation activity (e.g. Kennicutt & Evans 2012). In order for Hα emission arising from massive young stars to be observed, sufficient time for the ionizing radiation and winds of those stars to clear the surrounding, obscuring dust must have passed. Hence the ratio of Hα to 24 μm provides a measure of the levels of exposed to embedded star formation respectively (e.g. Schruba et al. 2017), and from this the relative ages of SFRs can be estimated (Jones et al. 2019). Recently, both Hα and 24 μm emission have been used as indicators of youth in age estimations of stellar clusters (with ages $>$2 Myr) across the disc of M 33 (Moeller & Calzetti 2022).

The ratio of far-IR emission observed with *Herschel* has been shown to spatially correlate with other shorter wavelength tracers of star formation across many nearby galaxies (Boselli et al. 2010) including in M 33 (Tabatabaei et al. 2007; Kramer et al. 2010). Specifically, the ratio of 250–500 μm emission in H II regions across NGC 6822 correlates well with other tracers of ongoing star formation (Galametz et al. 2010), pinpointing SFRs analysed in detail in more recent studies (Jones et al. 2019; Kinson et al. 2021). Longer wavelength emission is especially valuable at tracing the earliest stages of star formation, in which light emitted at shorter wavelengths is either obscured by dust (e.g. Hα) or the dust has not been sufficiently heated to become bright at mid-IR wavelengths.

Thus optical to far-IR emission can be expected to peak at different stages of the evolution of a SFR. A higher flux at longer wavelengths compared to Hα suggests rising star formation activity (e.g. Jones et al. 2019); the opposite behaviour is expected for regions in which star formation is ending and exposed massive stars begin to move on to the main sequence (e.g. Lada & Lada 2003;

Table 5. Catalogue of SFRs in M 33 identified using DBSCAN. The evolution score is discussed in Section 5.1.2. A sample of the table is provided here, the full version is available as supplementary material.

SFR ID	RA (J2000) h:m:s	Dec (J2000) deg:m:s	Maximum radius pc	Median radius pc	YSO number	Evolution score	SFR identifiers
1	01:34:17.91	+30:37:21.5	195	88	12	−0.239	–
2	01:33:10.89	+30:29:56.6	198	92	19	0.746	–
3	01:34:35.62	+30:45:59.3	357	193	28	0.239	NGC604-S
4	01:34:13.24	+30:45:59.3	376	156	49	0.388	–
5	01:33:13.07	+30:45:12.2	218	99	26	0.209	–

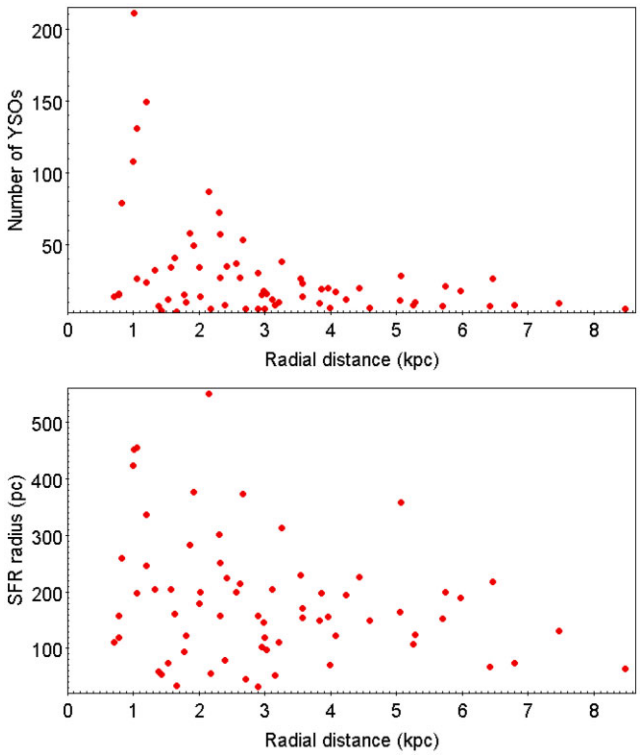

Figure 11. Number of YSOs (top) and radius (bottom) for each SFR identified by DBSCAN as a function of radial distance. A decreasing profile with increasing distance from the centre is seen in both panels.

Portegies Zwart, McMillan & Gieles 2010). Hence by comparing the ratios of Hα to 24 μm and 250–500 μm ([Hα]/[24 μm] and [250 μm]/[500 μm], respectively) for several SFRs it is possible to establish their evolutionary sequence.

For each SFR identified by the DBSCAN analysis background subtracted aperture photometry was performed in Hα, 24 μm *Spitzer*-MIPS, 250 and 500 μm *Herschel*-SPIRE images (see Section 2.2 for image details) to measure an average brightness within each aperture. The position and size of the apertures were set to the SFR centre and radius (see Table 5). In order to calibrate the properties and evolutionary status of SFRs in M 33, we used regions in NGC 6822 that have been well-characterized in the literature (Schruba et al. 2017; Jones et al. 2019; Kinson et al. 2021) as a benchmark for which we performed similar measurements. Positions and radii for NGC 6822 SFRs were taken from table 9 of Jones et al. (2019). These seven regions are the complete census of significant sites of star formation in NGC 6822. We do not include in this analysis the smaller SFRs newly identified in Kinson et al. (2021), since an established evolutionary sequence is not available for these regions.

Fig. 12 shows SFR measurements in M 33 and NGC 6822: Hα brightness against 24 μm brightness (upper panel), the far-IR 250 and 500 μm brightnesses (middle panel) and [Hα]/[24 μm] against [250 μm]/[500 μm] ratios (lower panel). The Hα and 24 μm brightnesses appear loosely correlated while the 250 and 500 μm brightnesses show a much tighter correlation (for M 33 SFRs, $r_{\rm pearson} \sim 0.24$ and 0.95, respectively). The 24 μm brightnesses for the SFRs in the two galaxies appear broadly consistent; the Hα brightnesses for SFRs in NGC 6822 are higher than those in M 33 with none falling below ∼20 counts per pixel. As noted in Section 2.2, the two Hα images are taken with similar instruments and are calibrated in a consistent way (see tables 1 and 2 of Massey et al. 2007b), hence counts can be confidently compared between images.

The higher Hα brightnesses in NGC 6822 may be a consequence of its lower metallicity (∼0.2 $Z_\odot$, e.g. Skillman, Terlevich & Melnick 1989; Richer & McCall 2007). At low metallicity, the interstellar medium (ISM) is more porous allowing for increased leakage of ionizing radiation (Madden et al. 2006; Dimaratos et al. 2015). This effect has been used to explain the observed ISM properties in many dwarf galaxies (Cormier et al. 2015, 2019). The resulting increased mean free path for far-UV photons could therefore make Hα-emitting regions in NGC 6822 larger and brighter, compared to those in M 33.

In Fig. 12 (middle panel), we show loci of theoretical modified blackbody emission for dust temperatures 20, 25, and 30K (colour-coded) and values of β the dust emissivity index ($\beta = 2$ and 1.5, solid and dashed lines respectively). β represents the frequency dependence of the dust emissivity, which modifies the blackbody emission of dusty sources (Hildebrand 1983). In NGC 6822, values of β adopted previously lie within this range (e.g. Israel, Bontekoe & Kester 1996). Tabatabaei et al. (2014) find that β varies from $\beta = 2$ in the central regions of M 33 to $\beta = 1.3$ in the outer disc; for SFRs however, a value of $\beta = 2$ seems to be more appropriate (Braine et al. 2010; Tabatabaei et al. 2014). The position of the SFRs in NGC 6822 is broadly consistent with those in M 33 with a slight offset to higher 250 μm values. This offset corresponds to an increase in temperature of ∼2K or a variation in β of ∼0.4. This offset could be due to the difference in dust properties with ISM in NGC 6822 having a smaller grain size than that in M 33 (Wang et al. 2022). Smaller grain sizes have been shown to correlate with higher grain equilibrium temperatures (Zelko & Finkbeiner 2020). Dust temperatures have been found to be higher in the lower-metallicity SMC compared to LMC (van Loon et al. 2010). Higher dust temperatures in dwarf galaxies can also lead to stronger far-IR emission per dust mass unit than in larger galaxies (Henkel, Hunt & Izotov 2022).

The symbol sizes in Fig. 12 are proportional to the number of YSOs in the SFR; YSO numbers come from the DBSCAN analysis in Section 4.4 for M 33, and from table 4 of Kinson et al. (2021) for NGC 6822 (these values are used instead of those reported by Jones et al. (2019), since PRF identification is also used). For the most populous regions in M 33, measurements other than Hα tend

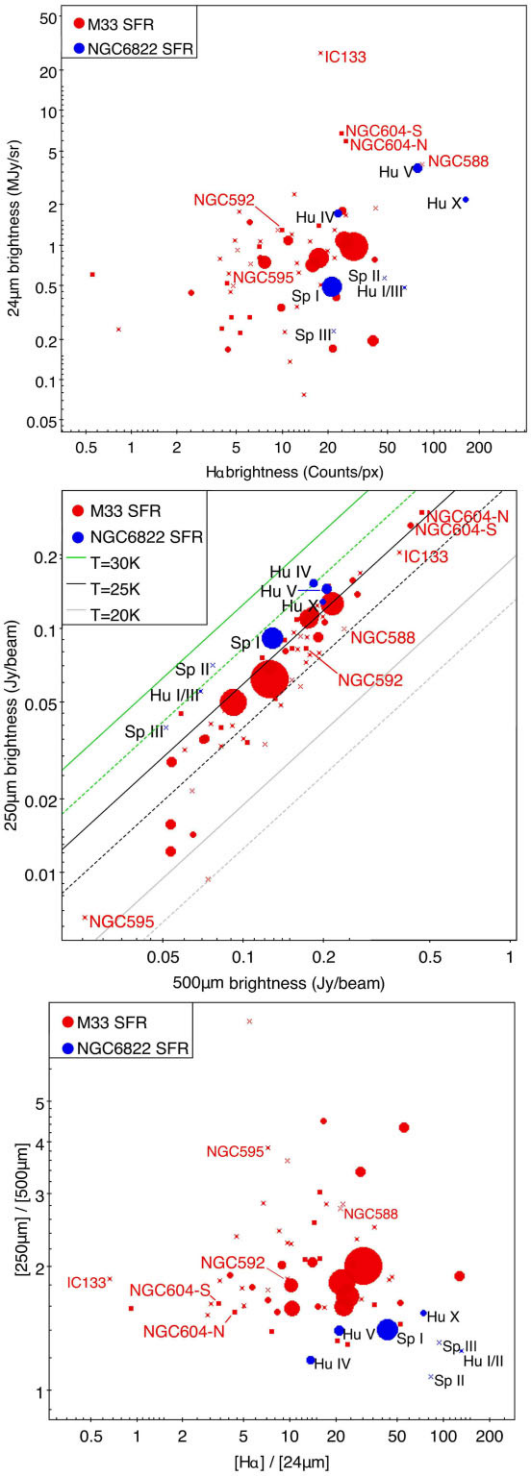

Figure 12. Photometric measurements for each SFR in M 33 and NGC 6822 (red and blue symbols, respectively): Hα and 24 μm (upper panel), 250 and 500 μm (middle), [Hα]/[24 μm] and [250 μm]/[500 μm] (lower). The symbol size is proportional to the number of YSOs in each region (crosses mark particularly small regions); SFR radii for M 33 and NGC 6822 are, respectively, from our analysis and from Kinson et al. (2021). In the middle panel, loci for modified blackbodies of different temperatures (colour-coded) and $\beta = 2$ and 1.5 (solid and dashed lines, respectively) are shown. Significant SFRs are labelled (see text).

towards the ranges' averages (Fig. 12 upper and middle panels). This could be expected if the largest SFRs identified by DBSCAN are in fact comprised of multiple smaller regions of differing properties that even out when integrated. In the Milky Way, the Orion–Eridanus superbubble contains several stellar subgroups, sites of ongoing star formation (e.g. Bally et al. 2009; Lim et al. 2021) alongside structures with older populations (e.g. Bally 2008). As the individual subgroups evolve they expand into and interact with one another (Ochsendorf et al. 2015), creating large-scale substructures that have been mapped in free-streaming Hα emission (Ochsendorf et al. 2015; Ha et al. 2022). The Orion-Eridanus superbubble when scaled to the distance of M 33 would be approximately 254 pc (61 arcsec) in size, which would place it well within the range of M 33 SFRs (see Fig. 11 and Appendix D). This may explain why the largest SFRs in M 33 have the brightest Hα emission but unremarkable overall mid- and far-IR brightness, appearing relatively evolved (see next section).

5.1.2 *SFR evolutionary status*

As previously mentioned, we utilize SFRs in NGC 6822 for which there is an established evolutionary sequence as a guide for the SFRs we identify in M 33. Given the previously discussed differences between SFRs in M 33 and NGC 6822 and the very different sample sizes, we compared the SFRs in the two galaxies using the regions' rank order in each ratio.

In Fig. 13 (upper panel), we show the rank sequence for the SFRs in NGC 6822. Using a combination of [Hα]/[24 μm] ratio and CO morphologies, Schruba et al. (2017) suggest that *Hubble* I/III and *Hubble* X are likely more evolved than *Hubble* IV and *Hubble* V. Jones et al. (2019) use similar tracers to propose that the most evolved SFR is likely *Hubble* I/III, *Spitzer* I, and *Hubble* V are the least evolved and regions *Hubble* IV and X, *Spitzer* II, and III are intermediate. This is broadly consistent with the position of the regions in Fig. 13: the least evolved regions are found towards the lower left and most evolved towards the upper right; the blue arrow indicates the sequence of evolution. While this generally agrees with the relative evolution stages from Schruba et al. (2017) and Jones et al. (2019), the exception is *Hubble* X which would appear less evolved in our analysis. Whilst the intermediate regions in NGC 6822 appear quite distant from the locus of parity between the ranked ratios (shown by the black diagonal lines in Fig. 13), this is due to the low number of SFR present. Indeed, this effect is not seen in the rank order of the SFRs in M 33 (lower panel of Fig. 13). Some of the most prominent H II regions and SFR in M 33 are discussed further in Section 5.1.4.

In order to compare the evolution stage of SFRs in M 33 and NGC 6822, we convert the distance from the locus of rank parity in Fig. 13 into a measure of evolution, normalized to the number of sources in each sample. We call this the evolution score. A negative evolution score represents a less evolved, more embedded region in which the [250 μm]/[500 μm] ratio dominates over the [Hα]/[24 μm] ratio. A positive value of the normalized evolution score reflects a region in which the ISM is being cleared by bright young massive stars and neutral gas is ionized forming H II regions, allowing shorter-wavelength photons to freely propagate.

To characterize star formation activity across the disc of M 33, we investigate the relation between galactic location and evolution score. In Fig. 14, the location of each SFR in M 33 is shown superposed on spiral arm structure; region size and evolution score are indicated by symbol size and colour, respectively. The largest regions, that are also generally the most evolved, lie at the base of the two primary spiral

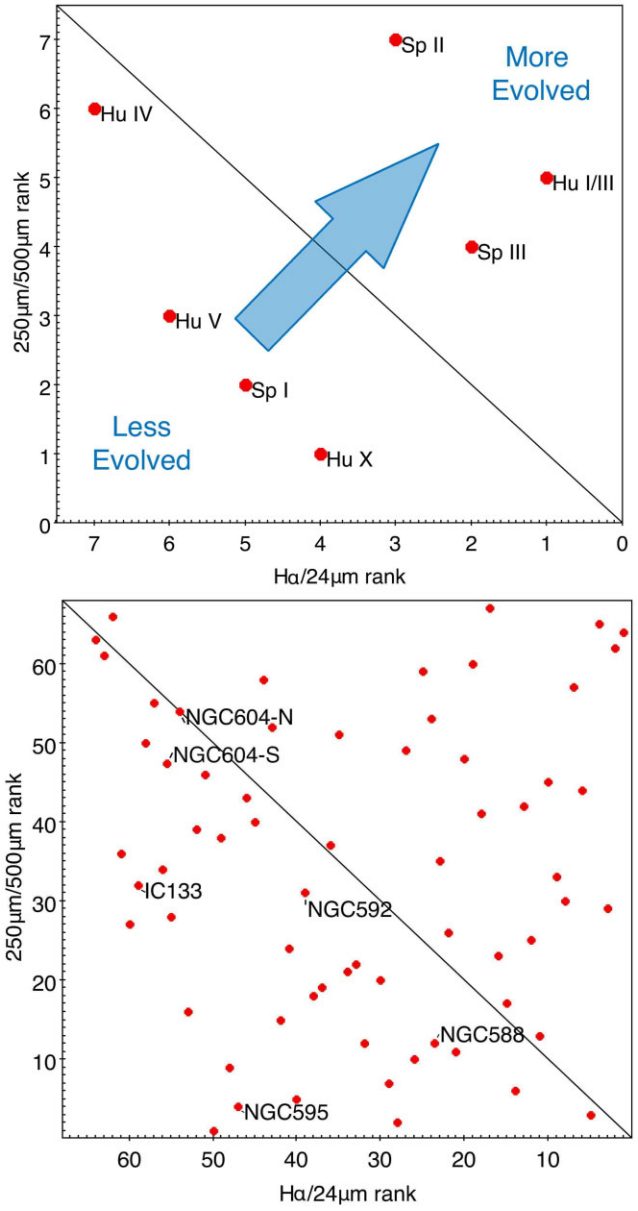

Figure 13. SFRs in NGC 6822 (upper) and M 33 (lower) shown by their relative ranks in the [Hα]/[24 μm] and [250 μm]/[500 μm] ratios. The diagonal line indicates the locus of equal rank in both ratios. In the top panel, the direction of SFR evolution is indicated by the arrow; significant SFRs are labelled (see text for more detail).

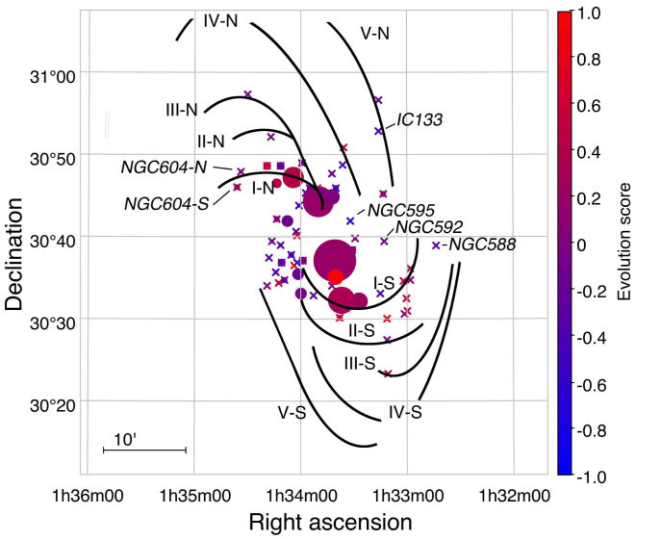

Figure 14. Galactic location of SFRs in M 33 shown with a schematic labelled spiral structure. Symbol size is proportional to the number of YSOs, colour shows the evolution score (the smallest regions are marked with a cross). The least evolved regions (purple hues) ring, the centre of the galaxy with more evolved regions (red hues) located further out in the disc (see also Fig. 15). SFRs discussed in Section 5.1.4 are labelled.

arms I-N and I-S; the least evolved SFRs mainly lie immediately surrounding the central region of the galaxy. In Fig. 15, we explore in more detail the effect of radial distance on the evolution scores of the SFRs. At radii larger than ∼4.5 kpc, most SFRs have positive evolution scores (i.e. are more evolved); obvious outliers are IC 133 in arm V-N and NGC 588, which are discussed further in Section 5.1.4.

We compare the relation between the number of YSOs in a SFR to its evolution score in both M 33 and NGC 6822 in Fig. 16. The SFRs in NGC 6822 show a decreasing number of YSOs with increasing evolution score ($r_{\rm pearson} \sim -0.71$). For M 33, the opposite trend is seen, albeit less strong ($r_{\rm pearson} \sim 0.21$), that could suggest that larger regions appear more evolved. In order to assess the similarity of the two SFR samples, we used a 2-Dimensional KS test (Peacock 1983; Fasano & Franceschini 1987). We find a low probability ($p \sim 0.29$) that the two samples are drawn from distinct parent samples with the caveat that the low number of SFRs analysed in NGC 6822 is not an effect of sampling, since these are all the significant SFRs in this galaxy. Whilst the [Hα]/[24 μm] ratio (lower panel of Fig. 12) suggests that larger regions should correlate to higher evolution scores, this is in fact not seen in Fig. 16 with the exception of the very largest regions ($n_{\rm YSOs} \geq 50$); as discussed in Section 5.1.1, these regions likely result from the combination of multiple smaller SFRs.

5.1.3 SFRs in the context of GMCs

We checked the positions of the 68 SFRs identified in our analysis against existing giant molecular cloud (GMC) catalogues. Corbelli et al. (2017) identified 566 GMCs using CO (2–1) observations and classify these according to their emission characteristics: the types A, B, and C correspond, respectively, to inactive GMCs clouds with embedded or low-mass star formation, and clouds with massive or exposed star formation, the latter associated with Hα and 24 μm emission. We find 17 type A, 16 type B, and 54 type C GMCs that have a positional overlap with 62 out of 68 SFRs (∼91 per cent), using the SFR median radii provided in Table 5 and the GMC deconvolved effective radii (see table 5 of Corbelli et al. 2017). Since significant 24 μm emission (strongly correlated with star formation, e.g. Williams, Gear & Smith 2018) is required for a type C classification, most SFRs are indeed matched to this GMC type; furthermore as discussed in Section 5.2, our analysis allows only for the identification of the most massive YSOs. Type A matches occur mostly for the largest SFRs that, in fact, include multiple GMCs of different types. Corbelli et al. (2017) find that type-B GMCs are rarely found close to the spiral arms of M 33, whereas types A and C are more closely aligned to H I filaments in the arms. We do not find an overall correlation between GMC type and SFR evolution score.

Star formation in the two primary spiral arms of M 33 has been previously studied to differing degrees. Arm I-N contains

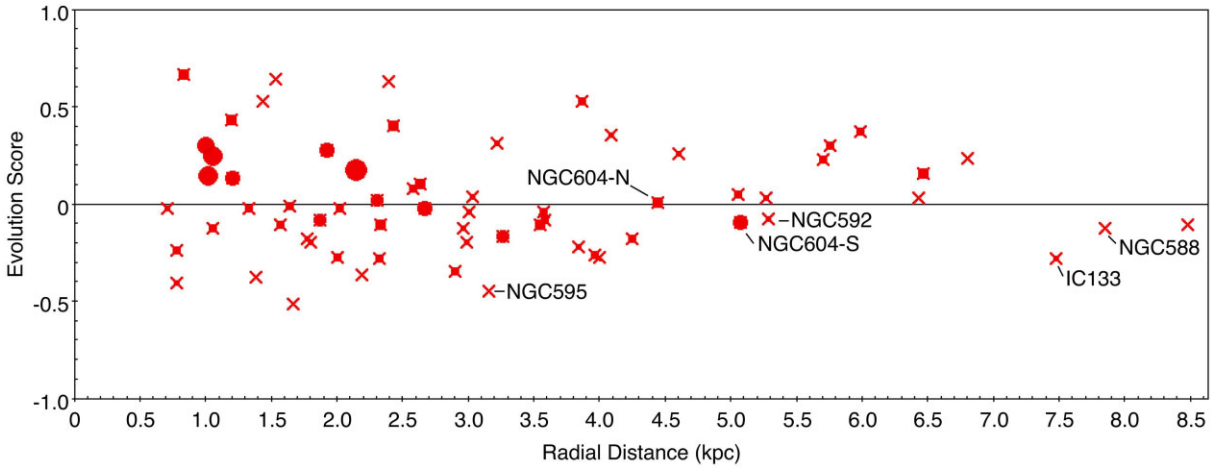

Figure 15. Normalized evolution score against radial distance for SFRs in M 33. Symbol size is proportional to the radius of each cluster. IC 133 and NGC 588 are notable outliers in that they have a low evolution score and lie far out in the disc of M 33.

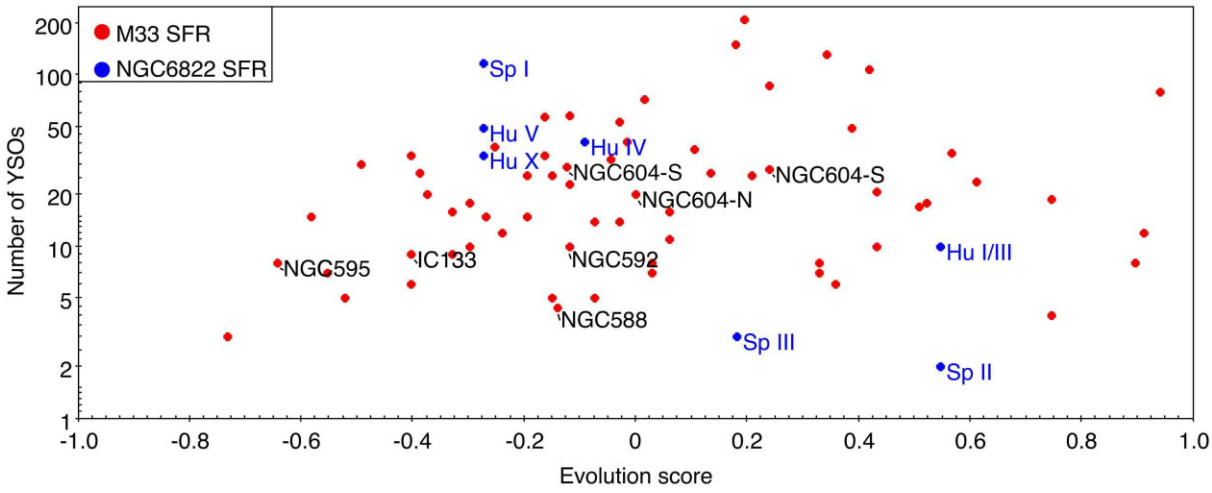

Figure 16. Number of YSOs against normalized evolution scores for SFRs in M 33 and NGC 6822. The number of YSOs for SFRs in M 33 and NGC 6822 are, respectively, from our analysis and from Kinson et al. (2021). There seems to be a slight tendency for larger SFRs to appear more evolved in M 33.

several well studied GMCs along its extension as well as the prominent H II region NGC 604. We find counterparts to GMCs also identified in the CO (3–2) observations of M 33 by Miura et al. (2012): SFRs 11 and 36 (GMC 16 and 8 respectively in their nomenclature) as well as two additional CO peaks in between these (see fig. 1 of Kondo et al. 2021), which correspond to SFRs 25 and 35. NGC 604 is recovered in our analysis as two SFRs discussed further in Section 5.1.4. These regions in I-N were studied in detail with recent ALMA observations (Kondo et al. 2021; Muraoka et al. 2020; Tokuda et al. 2020). All the SFRs in arm I-N are associated with Type-C GMCs, SFR 36 also contains a Type-B GMC. SFR 11/GMC 16 contains filamentary structure (Tokuda et al. 2020), which is not present in the comparatively inactive SFR 36/GMC 8 (Kondo et al. 2021). The lack of filamentary structure, and the presence of a Type-B GMC in SFR 36/GMC 8 would suggest it is less evolved compared to SFR 11/GMC 16, as supported by the evolution scores, −0.11 and −0.03, respectively.

Arm I-S is less disturbed than arm I-N and it seems to exhibit a clear progression from Type-A to Type-C GMCs through the arm (Corbelli et al. 2017). Due to the few SFR matches to Type-A GMCs, we cannot confirm this observation. The progression across arm I-S, as well as spatial offsets between filamentary structures and H I gas (e.g. in SFR 11/GMC 16, Tokuda et al. 2020), are consistent with the 'quasi-stationary spiral structure' model of Lin & Shu (1964). Whilst Kondo et al. (2021) find H I gas velocities in SFR 36/GMC 8, which are consistent with 'dynamic spiral' theory (Dobbs & Baba 2014), they cannot rule out an external source for the gas such as tidal interactions with M 31 (Tachihara et al. 2018).

5.1.4 Comments on individual M 33 SFRs

NGC 604 is one of the largest and brightest H II regions in the Local Group (e.g. Bosch, Terlevich & Terlevich 2002). Located around 4.8 kpc from the centre of M 33 in arm I-N star formation has been studied there at many wavelengths (e.g. Heidmann 1983; Fariña et al. 2012; Miura et al. 2012; Tachihara et al. 2018; Leitherer 2020; Muraoka et al. 2020). NGC 604 has undergone multiple star formation events (Eldridge & Relaño 2011) with earlier star formation episodes suggested to trigger the subsequent episodes (Tosaki et al. 2007; Tachihara et al. 2018).

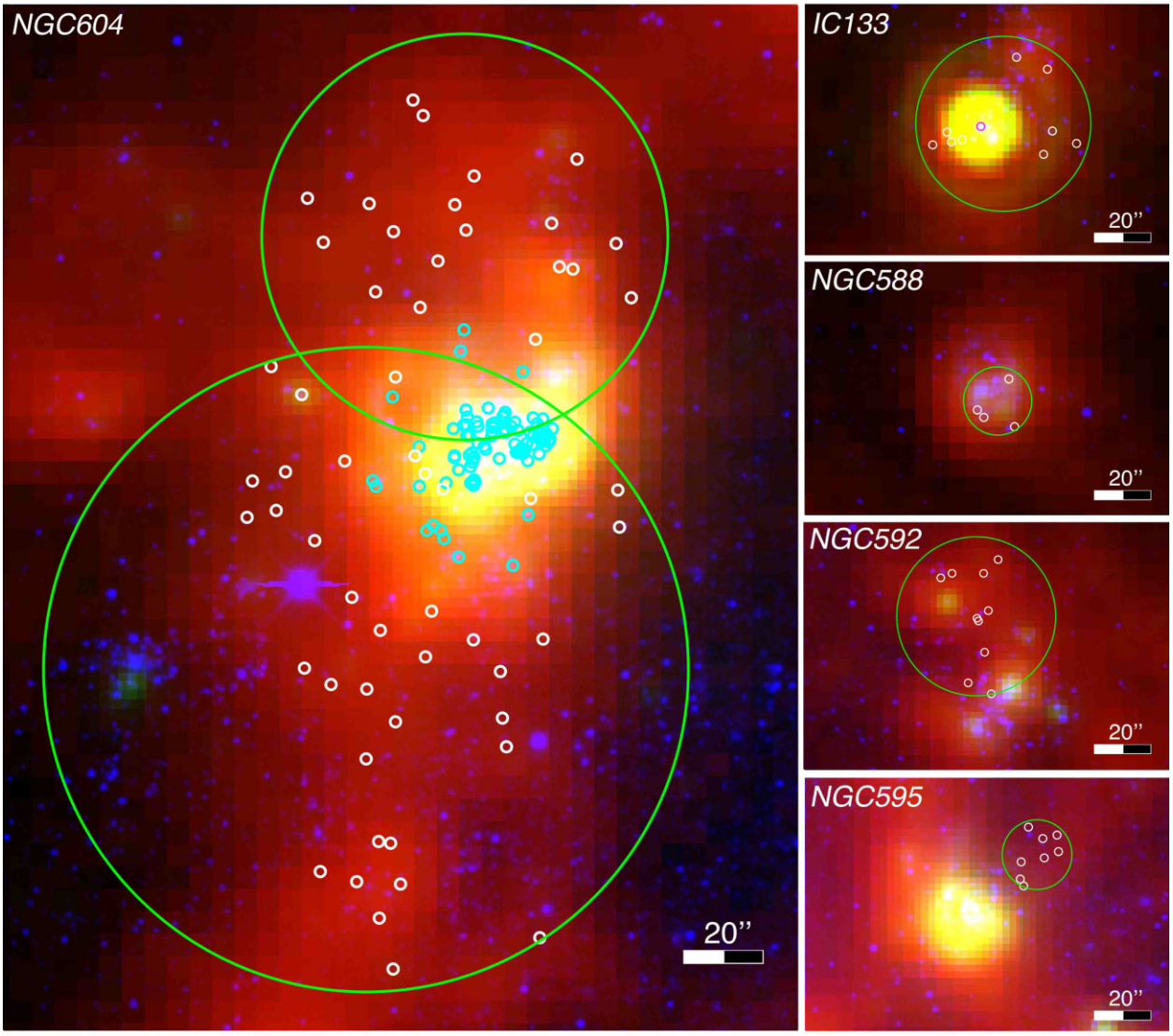

Figure 17. RGB image (250 μm *Herschel*-SPIRE, 24 μm *Spitzer*-MIPS, Hα, respectively − see Section 2.2 for image details) of NGC 604, IC 133, NGC 588, NGC 592, and NGC 595. YSOs identified in this work are shown by white circles, the extent of each SFR is shown by the green circles in NGC 604 cyan circles show YSOs identified in Fariña et al. (2012), in IC 133 the magenta circle shows the location of the maser counterpart (see text).

Using GEMINI-NIRI photometry with excellent seeing conditions (∼0.35 arcsec), Fariña et al. (2012) identified 68 massive YSOs in the central region of NGC 604 (see left-hand panel of Fig. 17). Whilst all the YSOs identified by Fariña et al. (2012) are brighter than the catalogue sensitivity limits (see Section 2.1.1), none of these sources have a counterpart within 1 arcsec in the near-IR catalogue of Javadi et al. (2015). In fact, within 30 arcsec of the centre of NGC 604 (01:34:32.1, + 30:47:01; Montiel et al. 2015), the near-IR catalogue contains only 27 sources, of which five are classified by the PRF analysis (as WRs, consistent with the young nature of the region). Likewise, the *Spitzer*-IRS pointings described in Martínez-Galarza et al. (2012) are all located in this region of sparse near-IR point sources. This is a limitation of the catalogue used in our analysis in this region of extremely bright ambient emission; the YSOs we identify in our analysis are found instead at its periphery.

The DBSCAN analysis divides NGC 604 into two SFRs, North and South of the centre of brightest emission (see Fig. 17). The two SFRs (3 and 17 in Table 5), which we refer to as NGC 604-N and -S contain 20 and 28 YSOs, respectively. Whilst the separation of NGC 604 into two SFRs may be in part driven by the paucity of near-IR data described above, this separation is supported astrophysically by the decomposition of NGC 604 into multiple components in CO (1–0) and (2–1) emission (Druard et al. 2014; Muraoka et al. 2020), and the South-East and North-West CO lobes of Wilson & Scoville (1992) which are coincident with our SFRs. We record different evolution scores respectively 0.01 and −0.09 for NGC 604-N and -S, indicative of star formation propagating from North to South in agreement with the Tosaki et al. (2007) and Muraoka et al. (2020) scenarios of triggered star formation in NGC 604. We note however that our analysis probes larger scales and in fact NGC 604-N lies outside the region discussed in those literature analyses. It is therefore more relevant to consider the larger scale H I gas interactions discussed in Tachihara et al. (2018). They identified two components of H I gas separated by $\sim$20 km s^{-1}; NGC 604-N is co-spatial with a peak in the redshifted component whilst NGC 604-S is co-spatial with the blue-shifted component (see their fig. 11). The collision of these

two large H I gas components is suggested to have triggered the star forming activity and growth of NGC 604 (Tachihara et al. 2018); such a scenario has also been proposed for other regions in arm I-N, namely SFR 11/GMC 16 and SFR 36/GMC 8 (Kondo et al. 2021). The origin of the infalling gas is not clear, however the presence of a H I stream between M 33 and M 31 (Bekki 2008; Lockman, Free & Shields 2012) due to a previous interaction between these two galaxies (Semczuk et al. 2018) offers one possible explanation (Tachihara et al. 2018).

NGC 595 (SFR 47), in which we identify eight YSOs, is the second most luminous H II region in M 33 after NGC 604 (Relaño & Kennicutt 2009) and is comparatively understudied. It lies to the North-West of the centre of M 33 towards the base of arm IV-N. Its evolution score of −0.4 suggests that NGC 595 is yet to reach peak star formation and may be amongst the youngest sites of star formation in the galaxy. The YSOs are located North-West of the bright 24 and 250 μm emission (see Fig. 17).

As noted in Section 5.1.2 and Fig. 15, the H II region IC 133 (SFR 62) has a low-evolution score (−0.28) for its large radial distance (∼7.5 kpc). IC 133 is located in arm V-N and contains nine YSOs, and a source of H_2O maser (Huchtmeier, Eckart & Zensus 1988; Greenhill et al. 1993), and OH maser (Staveley-Smith et al. 1987) emission. We identify a bright (K_s = 14.4 mag) and red ($J - K_s$ = 1.5 mag) source as the likely near-IR counterpart of the maser emission (at a distance of ∼0.″28) at coordinates 01:33:16.54, +30:52:49.7, which the PRF classifies into several classes across the 100 runs: $n_{\rm RSG}$ = 74, $n_{\rm CAGB}$ = 18, $n_{\rm YSO}$ = 5, $n_{\rm AGN}$ = 3. This suggests that a RSG classification is the most likely, since such sources are also known to harbour water maser emission (e.g. van Loon et al. 1998). The presence of an RSG source in an SFR is more likely if the H II region is more mature (>10 Myr), such that stars can evolve sufficiently to become RSGs, which is not reflected by the evolution score for IC 133. This may indicate that star formation in IC 133 is restarting after a period of hiatus.

Directly West of the centre of M 33 and not obviously linked with any spiral arm is the prominent H II region NGC 592. This is SFR 18 that contains ten YSOs. This H II region is thought to be young with age estimates from far-UV SED fitting of 4 and 5.6 Myr (respectively Pellerin 2006; Úbeda & Drissen 2009). Relaño & Kennicutt (2009) find compact knots of Hα coincident with the brightest 24 μm sources. We assign NGC 592 an evolution score of −0.08.

NGC 588, another large H II region in which star formation has been studied (e.g. Relaño & Kennicutt 2009; Monreal-Ibero et al. 2011) lies almost directly West of NGC 592 between the tips of arms I-S and III/IV-S as indicated in Fig. 14. Only four YSOs are classified within its extent (SFR 68). Alongside IC 133, NGC 588 is notable for its low evolution score (−0.12) at high radial distance (∼7.8 kpc) from the centre of M 33 (see Fig. 15).

5.2 YSO masses and star formation rate

The properties of the YSO sources analysed here are likely dominated by the most massive source in an unresolved proto-cluster (see also discussions in Oliveira et al. 2013; Ward et al. 2016, 2017). This effect on YSO model fitting analysis is discussed in Chen et al. (2010a), and accordingly Jones et al. (2019) present their mass estimates for YSOs in NGC 6822 as overestimated for the dominant source but underestimated for the total unresolved cluster. Furthermore, it is also widely accepted that most massive stars are found in binaries or multiple systems (e.g. Sana et al. 2008, 2012; Kobulnicky et al. 2014), implying that the dominant source is in turn an unresolved binary. These important caveats affect similar analysis in the literature (e.g. Sewiło et al. 2013; Jones et al. 2019 respectively in the SMC and NGC 6822) and are impossible to account for properly, and thus the mass estimates we discuss below should be taken with some caution.

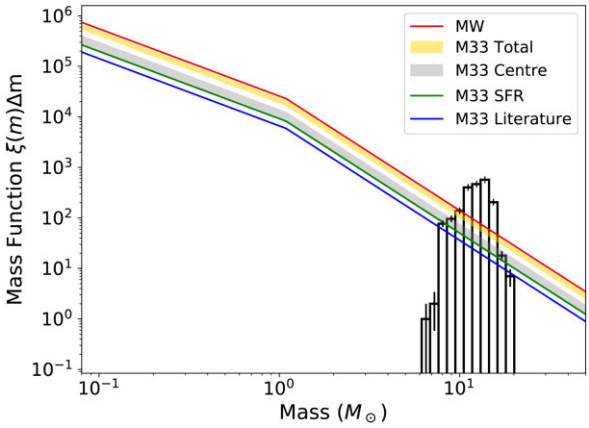

Figure 18. The mass distribution of the 1986 YSOs assigned to M 33 SFRs with scaled Kroupa (2002) IMFs overlain, see text for details. Poisson errors are indicated for each histogram bin.

Since the YSOs identified in our analysis only have photometry in the three near-IR bands, it is not feasible to obtain their masses using individual SED fitting as seen in, e.g. Whitney et al. (2008), Sewiło et al. (2013), Jones et al. (2019). We therefore use predicted near-IR K_s-band magnitudes (scaled to the distance of M 33) and $J - K$ colours estimated from the model grid of Robitaille et al. (2006) and the YSOs' positions in the CMD to assign them a model mass. For each of the 4985 YSOs identified by the PRF, we thus obtained a mass estimate as described below. Due to the depth of the near-IR catalogue (see Section 2.1.1), our analysis is likely sensitive to only the most massive YSOs. Given these sources evolve rapidly on to the main sequence once they leave their embedded stages, we use only models in the grid corresponding to Stage 0/I YSOs. We note that this model grid does not represent a realistic mass distribution in an Initial Mass Function (IMF) sense.

Each YSO is compared to models within a 0.5 mag distance in CMD space. For YSOs with at least three models in this range the median mass for the models is adopted; for YSOs with fewer models within 0.5 mag distance, the closest three models are used to compute the median model mass. This latter group of YSOs accounts for ∼11 per cent of all YSOs and ∼10 per cent of YSOs assigned to clusters; we consider these mass estimates more uncertain. YSO mass estimates range from 6–27$M_\odot$ with a median value of 13$M_\odot$.

The mass distribution of the YSOs assigned to SFRs is shown in Fig. 18 with a total mass of 2.5 × $10^4 M_\odot$. Using the commonly adopted functional form for the IMF by Kroupa (2002), scaled to match the observed mass distribution, and integrated over the range 0.08–100$M_\odot$, we estimate the total mass of YSOs in SFRs as 1.5 × $10^5 M_\odot$. Adopting a Stage 0/I lifetime of 0.2 Myr (e.g. Jones et al. 2019, and references therein), we estimate a star formation rate of 0.63$M_\odot$ yr^{-1} in M 33's SFRs (green line in Fig. 18). Due to the effects of crowding, the lower PRF classification certainty and potential contamination (see Section 4.4), we estimate the star formation rate separately for the unclustered YSOs in the central region. This rate is 0.79 ± 0.16$M_\odot$ yr^{-1} (grey shaded region in Fig. 18). Considering all YSOs, the total star formation rate is 1.42 ± 0.16$M_\odot$ yr^{-1} (gold shaded region) that overlaps with Milky Way estimates.

There are numerous determinations of global star formation rates in the Milky Way (MW), as compiled in table 1 of Chomiuk & Povich

(2011) for a range of methods (ionization rates, supernovae rates, near-IR to far-IR dust-heating ratios, nucleosynthesis rates and YSO counts), re-scaled to a Kroupa (2002) IMF; typical values are in the range $\sim 1.9 \pm 0.4 M_{\odot}\ yr^{-1}$ (see also Xiang et al. 2018). More recent work that uses Bayesian statistics to compare the rates compiled by Chomiuk & Povich (2011) favours a rate of $1.65 \pm 0.19 M_{\odot}\ yr^{-1}$ as the best fit to the data (Licquia & Newman 2015). Using direct YSO counts, Davies et al. (2011) find a rate of $1.75 \pm 0.25 M_{\odot}\ yr^{-1}$ (the average shown as the red line in Fig. 18). The rate of star formation in star forming galaxies is strongly correlated to the mass of available gas (Kennicutt & Evans 2012). It is therefore expected that M 33 ($M_{gas} \sim 3 \times 10^9\ M_{\odot}$, Corbelli 2003) has a lower star formation rate than the MW ($M_{gas} \sim 5 \times 10^{10} M_{\odot}$, Licquia & Newman 2015), as seen in Fig. 18.

Star formation rates estimated from direct YSO counts tend to be higher than those calculated with other methods that are sensitive to different star formation timescales, as documented in the MCs (e.g. Chen et al. 2010b; Carlson et al. 2012) and NGC 6822 (Jones et al. 2019), but are generally consistent (Sewiło et al. 2013). Our estimates are higher than the values calculated using the 24 μm ($0.2 M_{\odot}\ yr^{-1}$), Hα ($0.35 M_{\odot}\ yr^{-1}$), and far-UV ($0.55 M_{\odot}\ yr^{-1}$) emission maps by Verley et al. (2009) that adopted an average value of $0.45 M_{\odot}\ yr^{-1}$. More recently far-UV *Hubble Space Telescope* observations of M 33 were used by Lazzarini et al. (2022) to find a star formation rate of $0.74 M_{\odot}\ yr^{-1}$ over the last 100 Myr. The long-period variable (LPV) population gives an estimated star formation rate of $0.42 M_{\odot}\ yr^{-1}$ over the last 100 Myr (Javadi et al. 2017). Elson et al. (2019) explored star formation in M 33 at multiple scales from 49 pc to 782 pc at mid and far-IR wavelengths and estimated star formation rates of $0.44 \pm 0.1 M_{\odot}\ yr^{-1}$ (at 100 μm) and $0.34^{+0.42}_{-0.27} M_{\odot}\ yr^{-1}$ (at 12 μm). Using CO and HCN relations, Blitz & Rosolowsky (2006) inferred an integrated star formation rate in M 33 of $0.7 M_{\odot}\ yr^{-1}$. Our estimates for the star formation rate of M 33 are broadly consistent with these estimates towards the upper end.

6 CONCLUSIONS

In this work, we identified and described the YSO population across the whole disc of the flocculent spiral galaxy M 33 for the first time. We adapted the PRF classification technique which was successfully applied in NGC 6822 (Kinson et al. 2021) to better reflect the stellar populations in M 33. The PRF classifier was trained using a combination of near-IR and far-IR feature information to identify nine target classes.

In total, we applied the PRF to 162 746 sources of which 66 378 are consistently assigned to the same class across a total of 100 PRF runs. The PRF classifies with a median estimated accuracy of 86 per cent (the accuracy is based on the PRF's confusion matrices for the test runs). A total of 4985 YSOs were identified. A DBSCAN clustering analysis of the YSO population was used to identify 68 SFRs, mostly previously unknown, across the disc of M 33, containing 1986 YSOs. Most of these SFRs are located in the spiral arms. 2437 YSOs are found in the central $\sim 11.6 \times 10.4\ arcmin^2$ region that is too crowded for the clustering algorithm to be effective. The remainder 562 YSOs are seemingly isolated based on our analysis.

In total 62 out of our 68 SFRs (~91 per cent) are co-spatial with GMCs identified by Corbelli et al. (2017), mainly Type-C clouds (~87 per cent) with tracers of massive or exposed star formation. We identify SFR counterparts to the prominent H II regions IC 133, NGC 588, NGC 592, NGC 595, and NGC 604. A novel approach combining [Hα]/[24 μm] and [250 μm]/[500 μm] ratios were used to constrain the comparative evolutionary status of the M 33 SFRs, using regions in NGC 6822 as a benchmark sample. These ratios were converted into a common metric for ease of comparison. This evolution scores were used to compare SFRs in the context of radial distance in the galaxy number of YSOs and the relation to M 33's spiral structure.

We resolve the wider NGC 604 environment into two SFRs with different evolutionary status; these are co-spatial with two different H I gas components identified by Tachihara et al. (2018). The collision of these components may explain the triggering of initial star formation and progression from North to South (Tosaki et al. 2007), for which we see some evidence in our evolution score analysis. In this scenario, the in-falling H I gas is responsible for feeding the growth of NGC 604 into one of the most luminous H II regions in the Local Group. This gas component may originate from a stream connecting M 33 and M 31 arising from an earlier interaction with M 31.

We used model grids for Stage 0/I YSOs (Robitaille et al. 2006) to estimate the mass of each of the 4985 YSOs. Given that a SED fitting analysis is not feasible with just three near-IR bands, masses are derived from the models that are closest to each YSO in the colour-magnitude diagram. Estimated YSO masses range from 6–$27 M_{\odot}$ with a median value of $13 M_{\odot}$. The total mass of YSOs assigned to SFRs is $2.5 \times 10^4 M_{\odot}$. Using a Stage 0/I lifetime of 0.2 Myr, we estimate a star formation rate of $0.63 M_{\odot}\ yr^{-1}$ for M 33 spiral arms' SFRs. In the central region of M 33, we find a higher value of $0.79 \pm 0.16 M_{\odot}\ yr^{-1}$ with the caveat of less certain source classifications for this crowded region. These estimates give a total M 33 star formation rate of $1.42 \pm 0.16 M_{\odot}\ yr^{-1}$ determined from direct YSO counts. As expected from gas mass scaling relations, the star formation rate for M 33 is lower than that of the more massive MW ($1.75 \pm 0.25 M_{\odot}\ yr^{-1}$, Davies et al. 2011, also computed from YSO counts).

We have for the first time identified massive YSOs on galactic scales in a Local Group spiral galaxy, extending such analysis beyond the nearby star-forming dwarf galaxies (LMC, SMC, and NGC 6822). Machine learning approaches, as we have demonstrated, offer an invaluable tool for disentangling and classifying large data sets. The next generation of observatories such as the extremely large telescope, *James Webb* and *Roman Space Telescopes* will deliver a treasure-trove of such data extending the range of galaxies in which such studies can be conducted.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous referee for their helpful comments and suggestions that helped improve the paper. DAK acknowledges financial support from STFC via their PhD studentship programmes as well as Keele University via their Phase-2 COVID-19 funding extension programme. The authors thank A. Javadi for help with gaining access to their catalogues. *Herschel* is a European Space Agency (ESA) space observatory with science instruments provided by European-led Principal Investigator consortia and with important participation from NASA. This work is based on observations made with the *Spitzer Space Telescope*, which is operated by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA. This work has made use of data from the 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.

DATA AVAILABILITY

The data underlying this article which are not included in the supplementary online materials will be shared on reasonable request to the corresponding author.

REFERENCES

Alexeeva S., Zhao G., 2022, ApJ, 925, 76
Bally J., 2008, in Reipurth B., ed., ASP Monograph Publications, Vol. 4., Handbook of Star Forming Regions:Volume I, The Northern Sky. Astron. Soc. Pac., San Francisco, p. 459
Bally J., Walawender J., Reipurth B., Megeath S. T., 2009, AJ, 137, 3843
Barker M. K., Ferguson A. M. N., Cole A. A., Ibata R., Irwin M., Lewis G. F., Smecker-Hane T. A., Tanvir N. R., 2011, MNRAS, 410, 504
Bekki K., 2008, MNRAS, 390, L24
Bianchi L., Efremova B., Hodge P., Massey P., Olsen K. A. G., 2012, AJ, 143, 74
Blitz L., Rosolowsky E., 2006, ApJ, 650, 933
Block D. L., Freeman K. C., Jarrett T. H., Puerari I., Worthey G., Combes F., Groess R., 2004, A&A, 425, L37
Block D. L. et al., 2007, A&A, 471, 467
Bosch G., Terlevich E., Terlevich R., 2002, MNRAS, 329, 481
Boselli A. et al., 2010, A&A, 518, L61
Bradley L. et al., 2020, astropy/photutils: 1.0.0
Braine J. et al., 2010, A&A, 518, L69
Braine J., Rosolowsky E., Gratier P., Corbelli E., Schuster K. F., 2018, A&A, 612, A51
Breiman L., 2001, Machine learning, 45, 5
Britavskiy N. et al., 2019, A&A, 624, A128
Carlson L. R., Sewiło M., Meixner M., Romita K. A., Lawton B., 2012, A&A, 542, A66
Casali M. et al., 2007, A&A, 467, 777
Chen C. H. R. et al., 2010b, ApJ, 721, 1206
Chen C.-H. R. et al., 2010a, ApJ, 721, 1206
Chomiuk L., Povich M. S., 2011, AJ, 142, 197
Churchwell E., Goss W. M., 1999, ApJ, 514, 188
Churchwell E., Witzel A., Huchtmeier W., Pauliny-Toth I., Roland J., Sieber W., 1977, A&A, 54, 969
Cioni M. R. L., 2009, A&A, 506, 1137
Corbelli E., 2003, MNRAS, 342, 199
Corbelli E., Thilker D., Zibetti S., Giovanardi C., Salucci P., 2014, A&A, 572, A23
Corbelli E. et al., 2017, A&A, 601, A146
Cormier D. et al., 2015, A&A, 578, A53
Cormier D. et al., 2019, A&A, 626, A23
Davies B., Hoare M. G., Lumsden S. L., Hosokawa T., Oudmaijer R. D., Urquhart J. S., Mottram J. C., Stead J., 2011, MNRAS, 416, 972
de Grijs R., Bono G., 2014, AJ, 148, 17
de Vaucouleurs G., de Vaucouleurs A., Corwin Herold G. J., Buta R. J., Paturel G., Fouque P., 1991, Third Reference Catalogue of Bright Galaxies. Springer, New York
Dimaratos A., Cormier D., Bigiel F., Madden S. C., 2015, A&A, 580, A135
Dobbs C., Baba J., 2014, PASA, 31, e035
Drout M. R., Massey P., Meynet G., 2012, ApJ, 750, 97
Druard C. et al., 2014, A&A, 567, A118
Eldridge J. J., Relaño M., 2011, MNRAS, 411, 235
Elson E. C., Kam S. Z., Chemin L., Carignan C., Jarrett T. H., 2019, MNRAS, 483, 931
Engelbracht C. W., MIPS Science Team, SINGS Team, 2004, in BAAS, 36, 701
Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. KDD'96, AAAI Press, p. 226
Fariña C., Bosch G. L., Barbá R. H., 2012, AJ, 143, 43
Fasano G., Franceschini A., 1987, MNRAS, 225, 155
Flesch E. W., 2021, MILLIQUAS - Million Quasars Catalog, Version 7.2, preprint (arXiv:2105.12985)
Gaia Collaboration, Brown A. G. A., Vallenari A., Prusti T., de Bruijne J. H. J., Babusiaux C., Biermann M., 2020, A&A, 649, A1
Galametz M. et al., 2010, A&A, 518, L55
Girardi L., Groenewegen M. A. T., Hatziminaoglou E., da Costa L., 2005, A&A, 436, 895
Gordon K. D. et al., 2011, AJ, 142, 102
Gratier P. et al., 2010, A&A, 522, A3
Greenhill L. J., Moran J. M., Reid M. J., Menten K. M., Hirabayashi H., 1993, ApJ, 406, 482
Griffin M. J. et al., 2010, A&A, 518, L3
Ha T., Li Y., Kounkel M., Xu S., Li H., Zheng Y., 2022, ApJ, 934, 7
Heidmann J., 1983, Highlights of Astronomy, 6, 611
Henkel C., Hunt L. K., Izotov Y. I., 2022, Galaxies, 10, 11
Hildebrand R. H., 1983, QJRAS, 24, 267
Hirschauer A. S., Gray L., Meixner M., Jones O. C., Srinivasan S., Boyer M. L., Sargent B. A., 2020, ApJ, 892, 91
Hony S. et al., 2011, A&A, 531, A137
Hubble E. P., 1926, ApJ, 63, 236
Huchtmeier W. K., Eckart A., Zensus A. J., 1988, A&A, 200, 26
Humphreys R. M., Sandage A., 1980, ApJS, 44, 319
Hunter D. A., Baum W. A., O'Neil E. J. J., Lynds R., 1996, ApJ, 456, 174
Israel F. P., Bontekoe T. R., Kester D. J. M., 1996, A&A, 308, 723
Javadi A., van Loon J. T., Mirtorabi M. T., 2011, MNRAS, 414, 3394
Javadi A., Saberi M., van Loon J. T., Khosroshahi H., Golabatooni N., Mirtorabi M. T., 2015, MNRAS, 447, 3973
Javadi A., van Loon J. T., Khosroshahi H. G., Tabatabaei F., Hamedani Golshan R., Rashidi M., 2017, MNRAS, 464, 2103
Jones O. C. et al., 2017, MNRAS, 470, 3250
Jones O. C., Sharp M. J., Reiter M., Hirschauer A. S., Meixner M., Srinivasan S., 2019, MNRAS, 490, 832
Kam Z. S., Carignan C., Chemin L., Amram P., Epinat B., 2015, MNRAS, 449, 4048
Kam S. Z., Carignan C., Chemin L., Foster T., Elson E., Jarrett T. H., 2017, AJ, 154, 41
Kato D. et al., 2007, PASJ, 59, 615
Kennicutt R. C., Evans N. J., 2012, ARA&A, 50, 531
Kennicutt Robert C. J. et al., 2003, PASP, 115, 928
Khoshgoftaar T. M., Golawala M., Hulse J. V., 2007, in 19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007). IEEE, Patras, Greece, p. 310
Kinson D. A., Oliveira J. M., van Loon J. T., 2021, MNRAS, 507, 5106
Kobulnicky H. A. et al., 2014, ApJS, 213, 34
Kondo H. et al., 2021, ApJ, 912, 66
Kramer C. et al., 2010, A&A, 518, L67
Kroupa P., 2002, Science, 295, 82
Lada C. J., Lada E. A., 2003, ARA&A, 41, 57
Lazzarini M. et al., 2022, ApJ, 934, 76
Leitherer C., 2020, Galaxies, 8, 13
Licquia T. C., Newman J. A., 2015, ApJ, 806, 96
Lim W. et al., 2021, PASJ, 73, S239
Lin C. C., Shu F. H., 1964, ApJ, 140, 646
Lockman F. J., Free N. L., Shields J. C., 2012, AJ, 144, 52
Long K. S., Charles P. A., Dubus G., 2002, ApJ, 569, 204
Ma J., 2001, Chin. Phys. Lett., 18, 1420
Madden S. C., Galliano F., Jones A. P., Sauvage M., 2006, A&A, 446, 877
Madden S. C. et al., 2014, PASP, 126, 1079
Magrini L., Stanghellini L., Corbelli E., Galli D., Villaver E., 2010, A&A, 512, A63
Maravelias G., Bonanos A. Z., Tramper F., de Wit S., Yang M., Bonfini P., 2022, preprint (arXiv:2203.08125)
Martínez-Galarza J. R., Hunter D., Groves B., Brandl B., 2012, ApJ, 761, 3
Massey P., Olsen K. A., Hodge P. W., Jacoby G. H., McNeill R. T., Smith R. C., Strong S. B., 2006, in American Astronomical Society Meeting Abstracts. p. 27.01
Massey P., Olsen K. A. G., Hodge P. W., Jacoby G. H., McNeill R. T., Smith R. C., Strong S. B., 2007a, AJ, 133, 2393

Massey P., McNeill R. T., Olsen K. A. G., Hodge P. W., Blaha C., Jacoby G. H., Smith R. C., Strong S. B., 2007b, AJ, 134, 2474
Massey P., Neugent K. F., Smart B. M., 2016, AJ, 152, 62
Massey P., Neugent K. F., Levesque E. M., Drout M. R., Courteau S., 2021, AJ, 161, 79
Meixner M. et al., 2006, AJ, 132, 2268
Meixner M. et al., 2013, AJ, 146, 62
Miura R. E. et al., 2012, ApJ, 761, 37
Moeller C., Calzetti D., 2022, AJ, 163, 16
Monreal-Ibero A., Relaño M., Kehrig C., Pérez-Montero E., Vílchez J. M., Kelz A., Roth M. M., Streicher O., 2011, MNRAS, 413, 2242
Montiel E. J., Srinivasan S., Clayton G. C., Engelbracht C. W., Johnson C. B., 2015, AJ, 149, 57
More A. S., Rana D. P., 2017, in 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM) Review of random forest classification techniques to resolve data imbalance. IEEE, p. 72
Mostoghiu R., Di Cintio A., Knebe A., Libeskind N. I., Minchev I., Brook C., 2018, MNRAS, 480, 4455
Muraoka K. et al., 2020, ApJ, 903, 94
Neugent K. F., Massey P., 2011, ApJ, 733, 123
Ochsendorf B. B., Brown A. G. A., Bally J., Tielens A. G. G. M., 2015, ApJ, 808, 111
Oliveira J. M. et al., 2013, MNRAS, 428, 3001
Paturel G., Petit C., Prugniel P., Theureau G., Rousseau J., Brouty M., Dubois P., Cambrésy L., 2003, A&A, 412, 45
Peacock J. A., 1983, MNRAS, 202, 615
Pellerin A., 2006, AJ, 131, 849
Pennock C. M., van Loon J. T., Bell C. P. M., Filipović M. D., Joseph T. D., Vardoulaki E., 2021, in Proc. IAU Symp. 356, Nuclear Activity in Galaxies Across Cosmic Time. Cambridge Univ. Press, Cambridge, p. 335
Pennock C. M. et al., 2022, MNRAS, 515, 6046
Pilbratt G. L. et al., 2010, A&A, 518, L1
Poglitsch A. et al., 2010, A&A, 518, L2
Portegies Zwart S. F., McMillan S. L. W., Gieles M., 2010, ARA&A, 48, 431
Quirk A. C. N. et al., 2022, AJ, 163, 166
Rahmah N., Sitanggang I. S., 2016, IOP Conf. Ser.: Earth Environ. Sci. Vol 31, Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra. IoP Publishing, Bristol, p. 012012
Reis I., Baron D., Shahaf S., 2019, AJ, 157, 16
Relaño M., Kennicutt Robert C. J., 2009, ApJ, 699, 1125
Ren Y., Jiang B., Yang M., Wang T., Jian M., Ren T., 2021, ApJ, 907, 18
Richer M. G., McCall M. L., 2007, ApJ, 658, 328
Rieke G. H., Lebofsky M. J., 1985, ApJ, 288, 618
Rieke G. H. et al., 2004, ApJS, 154, 25
Robitaille T. P., 2017, A&A, 600, A11
Robitaille T. P., Whitney B. A., Indebetouw R., Wood K., Denzmore P., 2006, ApJS, 167, 256
Rogstad D. H., Wright M. C. H., Lockhart I. A., 1976, ApJ, 204, 703
Rowe J. F., Richer H. B., Brewer J. P., Crabtree D. R., 2005, AJ, 129, 729
Sana H., Gosset E., Nazé Y., Rauw G., Linder N., 2008, MNRAS, 386, 447
Sana H. et al., 2012, Science, 337, 444
Schruba A. et al., 2017, ApJ, 835, 278
Searle L., 1971, ApJ, 168, 327
Semczuk M., Łokas E. L., Salomon J.-B., Athanassoula E., D'Onghia E., 2018, ApJ, 864, 34
Sewiło M. et al., 2013, ApJ, 778, 15
Skillman E. D., Terlevich R., Melnick J., 1989, MNRAS, 240, 563
Staveley-Smith L., Cohen R. J., Chapman J. M., Pointon L., Unger S. W., 1987, MNRAS, 226, 689
Tabatabaei F. S. et al., 2007, A&A, 466, 509
Tabatabaei F. S. et al., 2014, A&A, 561, A95
Tachihara K., Gratier P., Sano H., Tsuge K., Miura R. E., Muraoka K., Fukui Y., 2018, PASJ, 70, S52
Tan J. C., Beltrán M. T., Caselli P., Fontani F., Fuente A., Krumholz M. R., McKee C. F., Stolte A., 2014, in Beuther H., Klessen R. S., Dullemond C. P., Henning T., eds, Protostars and Planets VI. University of Arizona Press, Tucson, p. 149,
Tokuda K. et al., 2020, ApJ, 896, 36
Tosaki T., Miura R., Sawada T., Kuno N., Nakanishi K., Kohno K., Okumura S. K., Kawabe R., 2007, ApJ, 664, L27
Úbeda L., Drissen L., 2009, MNRAS, 394, 1847
van Loon J. T., Hekkert P. T. L., Bujarrabal V., Zijlstra A. A., Nyman L.-A., 1998, A&A, 337, 141
van Loon J. T., Oliveira J. M., Gordon K. D., Sloan G. C., Engelbracht C. W., 2010, AJ, 139, 1553
van den Bergh S., 1991, PASP, 103, 609
Verley S., Corbelli E., Giovanardi C., Hunt L. K., 2009, A&A, 493, 453
Wang Y., Gao J., Ren Y., Chen B., 2022, ApJS, 260, 41
Ward J. L., Oliveira J. M., van Loon J. T., Sewiło M., 2016, MNRAS, 455, 2345
Ward J. L., Oliveira J. M., van Loon J. T., Sewiło M., 2017, MNRAS, 464, 1512
Werner M. W. et al., 2004, ApJS, 154, 1
Whitney B. A. et al., 2008, AJ, 136, 18
Williams B. F., Dalcanton J. J., Dolphin A. E., Holtzman J., Sarajedini A., 2009, ApJ, 695, L15
Williams T. G., Gear W. K., Smith M. W. L., 2018, MNRAS, 479, 297
Williams B. F. et al., 2021, ApJS, 253, 53
Wilson C. D., Scoville N., 1992, ApJ, 385, 512
Xiang M. et al., 2018, ApJS, 237, 33
Zelko I. A., Finkbeiner D. P., 2020, ApJ, 904, 38

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Figure S1. The number of $n_{\rm YSO} = 20$ sources classified in common for increasing down-sampled training set selections.
Figure S2. The distribution of the number of PRF classifications for each source across all classes.
Figure S3. Histograms showing the distribution of number of YSOs and radii for the 68 YSO clusters identified with DBSCAN.
Figure S4. Spatial distributions of classified sources (red circles) in the central region of M33, overlaid on an RGB image: VLA Hi (red, Gratier et al. 2010), 250 μm Herschel-SPIRE (green, Kramer et al. 2010), 24 μm Spitzer-MIPS (blue, Engelbracht et al. 2004).

APPENDIX A: ON-LINE MATERIALS

(i) Down-sampling of the largest training set classes
(ii) PRF classification statistics
(iii) De-projected coordinates
(iv) Additional SFR statistics
(v) Central region spatial distributions

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