Machine learning identification of massive young stellar objects in Local Group galaxies

David Andrew Kinson

Doctor of Philosophy

Faculty of Natural Sciences, Keele University

June 2023

Abstract

This thesis presents the development and implementation of a machine learning classification of massive Young Stellar Objects (YSOs) in two Local Group galaxies, NGC 6822 and M 33. Using archival near- and far-IR data, point sources in both galaxies are classified into multiple stellar classes using a Probabilistic Random Forest classifier (PRF) trained on objects of known types. The spatial distributions of all classes are discussed. YSOs are classified with a high level of confidence (up to 97 per cent) in both galaxies. In NGC 6822, 125 YSOs are confirmed and 199 are newly identified. All major star forming regions (SFRs) in NGC 6822 are recovered and, additionally smaller SFRs are newly identified. In M 33 4985 YSOs were identified across the disk of M 33 and, applying a density-based clustering analysis 68 SFRs were identified primarily in the galaxy's spiral arms. SFRs associated with known H II regions were recovered, with $\sim$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, I employed a novel approach combining, into one metric, ratios of [Hα]/[24μm] and [250μm]/[500μm] to estimate the relative evolutionary status of all M 33 SFRs. By comparing the YSOs identified in M 33 with model grids for mass determination, a star formation rate is estimated for the first time from direct YSO counts; ($1.42 \pm 0.16\,\mathrm{M_\odot\,yr^{-1}}$) that is lower than that of the more massive Milky Way as expected. This project for the first time identifies 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). The techniques developed offer an invaluable tool for classifying large data sets.

Acknowledgements

I'd like to begin by thanking my supervisors Joana Oliveira and Jacco van Loon for their guidance, support and encouragement to explore new ideas during this project. I am deeply grateful for their patience in answering my many questions and offering valuable advice which has improved not only the work in this project but also my skills as a researcher.

I also want to thank those who shared images and pre-publication catalogues with me, which enabled much of the research in this project to be conducted.

I would like to thank the whole Astrophysics department for their warm welcome to Keele and for enabling me to transition from working on campus to at home. My thanks go especially to my fellow students, who's support and provision of much needed breaks have been invaluable. On a similar note, I must thank my many friends in the Athletic Union, especially those in the Lacrosse and Snowsports clubs. They have been key in keeping me going, and without whom my time at Keele would have been far less enjoyable. I also must thank my friends outside of the university for days away as well as hours of conversation online about everything and nothing.

Finally, I want to thank my family, especially my parents, for their constant support not only during my time at Keele but in all the preceding years. None of this would be possible without you. I'd like to dedicate this thesis to those members of my family who are no longer here to see its completion.

Contents

List of Figures

List of Tables

1 Introduction

The formation of massive stars plays a significant role in shaping their environment, with feedback from ionising UV radiation, strong winds and outflows acting to sculpt the interstellar medium (ISM). Massive stars are the factories in which heavy elements, up to iron, are produced. As these stars evolve, they enrich the ISM with these heavy elements. Whilst star formation in the low mass regime is well understood (see for example Shu et al., 1987, for a classical review), the mechanisms of high mass star formation are less well understood. Star formation has been extensively studied in our Galaxy and its satellites the Magellanic Clouds (MCs). Whilst our location within the Milky Way (MW) means that much of our understanding of massive star formation and its impact has been shaped in a spiral galaxy, this position within the MW means a complete census and an overall view of the spiral arms is not possible. To fully understand massive star formation and its impact on the evolution of the host galaxy we must probe different environments by studying galaxies at greater distances.

It has long been established that the dusty environment of star formation leads to an infrared emission excess which can be detected observationally (Lynden-Bell & Pringle, 1974). The surrounding dust envelopes act to block radiation at visible wavelengths, making observations in that regime difficult, as Young Stellar Objects (YSOs) are often heavily embedded, especially at early evolutionary stages. Hence studying in the infrared and longer wavelengths, able to pierce the nebulous obscuration and where their emission spectrum peaks, affords the best window into the processes these objects are undergoing. The availability of archival infrared surveys of Local Group galaxies therefore offers an excellent avenue by which to study high mass star formation across the entirety of a galaxy.

1.1 Star formation

In this section the general process of forming a star will be outlined before proceeding to consider factors specific to the high mass regime.

1.1.1 Collapsing clouds

The Kennicut-Schmidt law (Schmidt, 1959; Kennicutt, 1989) empirically shows that on large scales the star formation rate in a region is proportional to the gas density. The densest regions in the ISM, molecular clouds or giant molecular clouds (GMCs), have long been known to be the predominant sites of star formation both in the MW (Zuckerman & Palmer, 1974; Mooney & Solomon, 1988), and nearby galaxies (e.g. Engargiola et al., 2003; Blitz et al., 2007). GMCs range in size from $10-100\,\mathrm{pc}$ with typical masses of orders $10^4-10^6\,\mathrm{M}_\odot$ (e.g. Chevance et al., 2022).

Star formation in a GMC is governed by the balance of gravitational collapse against several mechanisms supporting the cloud. The virial theorem describes this balance of energies as shown below, following the method of Ward-Thompson & Whitworth (2015).

An isolated cloud can be described as a spherical collection of particles each with mass m_i, radial position $\mathbf{r}_i$, and velocity $\mathbf{v}_i$ moving in their collective gravitational field. The moment of inertia $\mathcal{I}$ of the cloud is given by

$$\mathcal{I} = \sum_i (m_i\,\mathbf{r}_\mathrm{i}\cdot\mathbf{r}_\mathrm{i}). \tag{1.1}$$

Setting the criterion that the cloud is in equilibrium, the derivative of $\mathcal{I}$ with respect to time must be zero. Given that $\dot{m}=0$ and $\dot{\mathbf{r}_i}=\mathbf{v}_i$, $\dot{\mathcal{I}}$ can be expressed as

$$\dot{\mathcal{I}} = 2\sum_i (m_i\,\mathbf{r}_\mathrm{i}\cdot\mathbf{v}_\mathrm{i}) = 0. \tag{1.2}$$

Equilibrium further demands that $\ddot{\mathcal{I}}=0$. Using Newton's second law $\mathbf{F}_i=m_i\dot{\mathbf{v}}_\mathrm{i}$, to describe the force on the ith particle, $\ddot{\mathcal{I}}$ is

$$\begin{aligned}\ddot{\mathcal{I}} &= 2\sum_i (m_i \dot{\mathbf{v}}_\mathrm{i} \cdot \dot{\mathbf{r}}_\mathrm{i}) + \sum_i (m_i \mathbf{v}_\mathrm{i} \cdot \mathbf{v}_\mathrm{i}) = 0 \\ &= 2\sum_i \mathbf{F_i} \cdot \mathbf{r}_\mathrm{i} + 4\mathcal{K},\end{aligned} \tag{1.3}$$

where $\mathcal{K} = \frac{1}{2}\sum_i (m_i\, v_i^2)$ is the total kinetic energy of the cloud. Clouds are supported by energies from multiple sources, $\mathcal{K}$ encompasses the contributions from turbulence, rotation and thermal motions. For the isolated cloud model $\mathbf{F_i}$ represents the net force acting on particle i from all other cloud particles j, i.e.

$$\mathbf{F_i} = \sum_{j\neq i} (\mathbf{F_{ij}}). \tag{1.4}$$

Combining Eqs. 1.3 and 1.4 gives,

$$\begin{aligned}\ddot{\mathcal{I}} &= 2\sum_i \sum_{j\neq i} (\mathbf{F_{ij}} \cdot \mathbf{r}_\mathrm{i}) + 4\mathcal{K} \\ &= \sum_i \sum_{j\neq i} (\mathbf{F_{ij}} \cdot \mathbf{r}_\mathrm{i} + \mathbf{F_{ji}} \cdot \mathbf{r}_\mathrm{j}) + 4\mathcal{K}.\end{aligned} \tag{1.5}$$

Newton's third law tells us that $\mathbf{F_{ij}} = -\mathbf{F_{ji}}$ and hence Eq. 1.5 becomes

$$\ddot{\mathcal{I}} = \sum_i \sum_{j\neq i} (\mathbf{F_{ij}} \cdot [\mathbf{r}_\mathrm{i} - \mathbf{r}_\mathrm{j}]) + 4\mathcal{K}. \tag{1.6}$$

Neglecting short range forces that act when $\mathbf{r}_\mathrm{i} - \mathbf{r}_\mathrm{j} \ll 1$, $\mathbf{F_{ij}}$ is

$$\mathbf{F_{ij}} = -\frac{G m_i m_j}{|\mathbf{r_i} - \mathbf{r_j}|^3} (\mathbf{r_i} - \mathbf{r_j}). \tag{1.7}$$

Substituting Eq. 1.7 into Eq. 1.6 gives,

$$\begin{aligned}\ddot{\mathcal{I}} &= \sum_i \sum_{j\neq i} \left(\frac{G m_i m_j}{|\mathbf{r_i} - \mathbf{r_j}|^3} (\mathbf{r_i} - \mathbf{r_j}) + 4\mathcal{K}\right. \\ &= 2\,\phi_G + 4\mathcal{K},\end{aligned} \tag{1.8}$$

where ϕ_G is the self-gravitational potential energy of the cloud. The factor of 2 arises from the summation of every particle pairing twice in Eq. 1.6.

Taking the condition that in equilibrium $\ddot{\mathcal{I}} = 0$ and Eq. 1.8 the virial condition for equilibrium is

$$0 = \phi_G + 2\mathcal{K}. \tag{1.9}$$

In an isolated, stable case where bulk motions due to rotation turbulence are not present, there are no additional external pressures and neglecting support contributions from magnetic fields, $\mathcal{K}$ simply becomes the thermal energy, U. The cloud of N many particles can be treated as an ideal monatomic gas with pressure P, volume V and, temperature T such that,

$$U = \frac{3}{2}PV = \frac{3}{2}Nk_BT. \tag{1.10}$$

The gravitation potential ϕ_G for the cloud, assuming uniform density, is given as

$$\phi_G = -\frac{3}{5}\frac{GM^2}{R}, \tag{1.11}$$

where M and R are the total mass and radius of the cloud respectively. Setting $N = \frac{M}{\mu m_H}$, with μ the mean molecular weight, m_H the atomic mass of hydrogen and defining the average cloud density as ρ the condition to trigger cloud collapse can be expressed as,

$$M > M_J = \left(\frac{5k_bT}{G\mu m_H}\right)^{\frac{3}{2}}\left(\frac{3}{4\pi\rho}\right)^{\frac{1}{2}}. \tag{1.12}$$

M_J is the Jeans Mass, the critical mass at which a self-gravitating cloud will undergo collapse (Jeans, 1902). An approximation for the Jeans mass can be found by substituting $\rho = n\mu m_H$, the isothermal sound speed $c_s = \left(\frac{k_bT}{\mu}\right)^{\frac{1}{2}}$ and evaluating known constants such as m_H, G and k_B. Doing so gives

$$M_J \approx \frac{T^{\frac{3}{2}}}{n^{\frac{1}{2}}}M_\odot. \tag{1.13}$$

Typical GMC values for T and n ($\sim 30\,\mathrm{K}$ and $10^3\,\mathrm{cm}^{-3}$, Schulz 2012) in units of $10\,\mathrm{K}$ and $10^4\,\mathrm{cm}^{-3}$ respectively lead to a typical Jeans mass of $\sim 10^2\,\mathrm{M}_\odot$. Com-

paring this to the typical GMC masses of $10^4 - 10^6\,\mathrm{M}_\odot$ (e.g. Chevance et al., 2022), a difference of several orders of magnitude is seen suggesting that all GMCs should be collapsing. The time taken for an unsupported cloud to collapse under self-gravity can be found using the free-fall timescale (t_{ff}). It is estimated by considering a mass element collapsing from the surface to the centre

$$\frac{d^2r}{dt^2} = -\frac{GM_r}{r^2}, \tag{1.14}$$

as,

$$t_{\mathrm{ff}} = \sqrt{\frac{3\pi}{32G\rho}}. \tag{1.15}$$

Using again typical GMC values gives $t_{\mathrm{ff}} \sim 10^5\,\mathrm{yrs}$ and comparing this to typical GMC lifetimes ($\sim 10^6\,\mathrm{yrs}$, Chevance et al., 2020) it becomes clear that additional support mechanisms are necessary to explain the process of cloud collapse. The additional support can be explained by considering magnetic fields as well as turbulence.

Turbulence contributes to the energy supporting a cloud (e.g. Dobbs et al., 2014; Burkhart, 2018). While generically included in the $\mathcal{K}$ term in Eq. 1.18, it is not considered in the stationary, isolated model described here. Whilst at large scales the ISM turbulence and gravity are roughly balanced (e.g. Kim & Ostriker, 2017), observational evidence of turbulence in GMCs from line broadening suggests that the thermal kinetic motion alone is not sufficient to balance against the gravitational attraction (Shu et al., 1987). Furthermore, as the turbulent dissipation timescale is comparable to the crossing time (Stone et al., 1998; Mac Low, 1999), for clouds massive enough to overcome the magnetic component of support, collapse still occurs on timescales comparable to t_{ff} (Chevance et al., 2022, and references therein). Given the expected rapid decay of turbulence (Goldreich & Kwan, 1974), external sources of energy are required to maintain it. In the rotating disk of a spiral galaxy turbulence can be generated, and energy is thus imparted to a GMC, from instabilities and fragmentation in the spiral arm (e.g. Inoue & Yoshida, 2018; Ramírez-Galeano et al., 2022).

Assessment of the magnetic field strength in GMCs suggests that they are not

the dominant support process (Crutcher, 2012). Nevertheless, they contribute to the energy balance of the cloud and can significantly affect the motion of charged particles which are bound to the magnetic field lines. The pressure exerted by a magnetic field is given by

$$\mathcal{P}_{\mathcal{M}} = \frac{B^2}{8\pi}, \tag{1.16}$$

where B is the magnetic field strength. Integrating this expression over a volume gives the total magnetic energy in that volume,

$$\mathcal{M} = \frac{1}{8\pi} \int_V B^2 dV. \tag{1.17}$$

This energy contribution can be incorporated into the virial equilibrium in Eq. 1.9 to give

$$0 = \phi_G + 2\mathcal{K} + \mathcal{M}. \tag{1.18}$$

Given that Eq. 1.17 is independent of cloud radius, the energy contribution from $\mathcal{M}$ is expected to remain constant throughout the collapse process. Clouds with masses less than that which can be magnetically supported (magnetically subcritical) are not typically observed (Crutcher, 2012), however during the collapse of a magnetically supercritical cloud as ϕ_G decreases, $\phi_G \lesssim \mathcal{M}$ can occur. This raises the problem of how collapse can continue, that can be solved if magnetic field is decreasing due to ambipolar diffusion. Charged particles in a cloud are bound to magnetic field lines, and in doing so collide with non-charged particles which move freely across the magnetic field. These collisions between neutral and charged particles in the cloud allow magnetic field energy to be kinetically dissipated, weakening the $\mathcal{M}$ term in Eq. 1.18. The rate at which ambipolar diffusion takes place is linked to the ratio of ionised to neutral particles, with a small $n_{ionised}/n_{neutral}$ beneficial. Such conditions are observed in GMCs (e.g. Schulz, 2012), hence $\mathcal{M}$ can decrease and gravitational collapse continue as ϕ_G also decreases. The problem of GMC support and collapse remains under active discussion, and a comprehensive review is found in Chevance et al. (2022).

1.1.2 From molecular clouds to YSOs

In the previous section gravitational collapse was treated as an isothermal process, the cloud densities being sufficiently low for energy released from gravitational collapse to be radiated away. In the isothermal case any region of the cloud with $M > M_J$ can independently collapse. This leads to fragmentation wherein the collapsing regions of a cloud form over-densities known as cloud cores. In this section the process of turning a core into a star is detailed.

As the average density of the cloud ($\bar{\rho}$) rises to $\sim 10^{-13}\,\mathrm{g\,cm^{-3}}$ the internal thermal pressure becomes significant (e.g. Schulz, 2012). At this point the opacity of the cloud (κ_R) is sufficient to trap its own radiation and cooling mechanisms become inefficient. This leads to rising temperature which in turn increases the critical mass M_J (see Eq. 1.12) halting further fragmentation. Cloud opacity is reduced in clouds with low metallicities (Elsender & Bate, 2021). The balance of thermal pressure against gravitational collapse leads to the creation of the first hydrostatic core.

With an initial thermal core formed and fragmentation ended, the core continues to accrete surrounding material from its envelope. This material free-falls onto the core at a velocity v_{ff}

$$v_{\mathrm{ff}} = \sqrt{\frac{2GM}{r}}, \tag{1.19}$$

which follows from Eq. 1.15. At the boundary between the core and the in-falling material a shock front of radius r is formed, where excess kinetic energy is dissipated via emission of photons. The associated luminosity of these shock front photons is called the accretion luminosity (L_{acc}) and is given by

$$L_{\mathrm{acc}} \sim \frac{1}{2}\dot{M}v_{\mathrm{ff}}^2 \sim \frac{GM\dot{M}}{r}, \tag{1.20}$$

where $\dot{M}$ is the mass accretion rate. Once an initial thermal core has been established the timescale for collapse changes (Larson, 1973). This new timescale known as the Kelvin-Helmholtz or thermal timescale (t_{KH}) is dependent on the luminosity of

the core (L_R) and ϕ_G

$$t_{KH} = \frac{|\phi_G|}{L_R} \approx 7 \times 10^{-5} \kappa_R \frac{M_R^2}{R^3 T^4}. \tag{1.21}$$

Given typical solar neighbourhood values for GMC densities and metallicity it can be shown that the time taken for thermal adjustment in the core exceeds the free-fall timescale by two orders of magnitude (Schulz, 2012) with $\mathrm{t_{ff}} \sim 10^5$yrs and $\mathrm{t}_{KH} \sim 10^7$yrs (Ward-Thompson & Whitworth, 2015). The free-fall timescale t_{ff} is dependent only on the core density not the mass, hence it is comparable across mass regimes. t_{KH} however is dependent on T^4 and thus it becomes much shorter as T increases implying that for high mass star formation t_{KH} can be very short. This presents particular challenges that will be described in Sect. 1.1.4.

The initial core mass is a fraction of its eventual mass ($M \sim 10^{-3} - 10^{-2}\,\mathrm{M}_\odot$, $R \sim 1{-}10\,\mathrm{AU}$, Bhandare et al. 2018). The core is however luminous enough ($T \sim 1000\,\mathrm{K}$) for molecular hydrogen to become excited in both vibrational and rotational modes. This excitation leads to H_2 molecular dissociation at $\mathrm{T} \sim 2000\,\mathrm{K}$, dissipating some energy which otherwise would contribute to further heating. This dissipation of energy allows for a second collapse to occur until hydrostatic equilibrium is re-established at a smaller core radius. This process can repeat cyclically as the mass and temperature of the core increases, ionising heavier elements with each cycle. As this progresses the core will become optically thick and begin to accrete from a disk of surrounding material (see phases b & c in Fig. 1.1).

During the collapse of a cloud fragment of radius R down to a second hydrostatic core of radius r conservation of angular momentum implies that where $R > r$ then the angular velocity Ω must increase i.e. $\Omega_r > \Omega_R$. Rotation generates a centrifugal force which is proportional to the square of angular velocity

$$F_C = m\Omega_r^2 r. \tag{1.22}$$

From the conservation of angular momentum it can be seen that $\Omega \propto r^{-2}$; using this relation and Eq. 1.22 then $F_C \propto r^{-3}$. Comparing this to the force due to gravity,

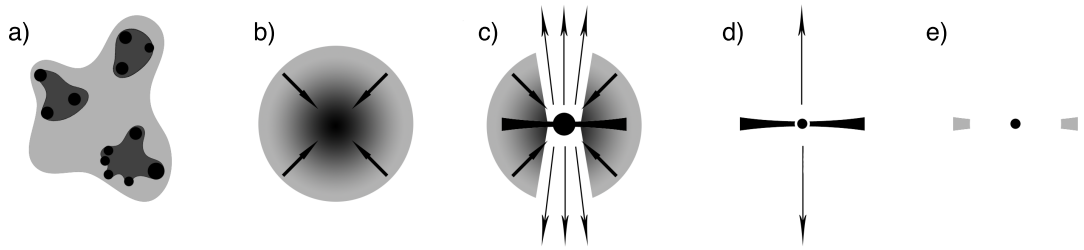

Figure 1.1: Phases of YSO formation in low mass regimes. a) A GMC in which dense cores form, b) gravitational collapse into a hydrostatic core, c) an embedded YSO accreting from a disk with bipolar outflows, d) the YSO had cleared most surrounding envelope and continues to accrete from its disk, observed as a strong lined T-Tauri star, e) the YSO's disk dissipates, detected as a weak lined T-Tauri star.

$F_G \propto r^{-2}$, clearly even small initial values of Ω can lead to significant outwards support, sufficient to prevent collapse in the rotational plane. The force however offers no support along the rotational axis and this leads to the formation of disk structures (see c & d in Fig. 1.1). These disks provide a mechanism by which angular momentum can be removed from the central core, via either friction and outwards transference of momentum (Schulz, 2012) or magnetic braking (e.g. Ireland et al., 2021). Removal of angular momentum is necessary to allow further accretion onto the core and to match observed rotation velocities of young main sequence stars (e.g. Gallet & Bouvier, 2013). Rotationally braked material is then further accreted from the disk onto the protostar. This process can be highly variable (e.g. Vorobyov, 2009; Meyer et al., 2017), with $\dot{M}$ $\sim 10^{-4}$–$10^{-7}\,\mathrm{M_\odot yr^{-1}}$ (e.g. Schulz, 2012). Clumpiness in the disk may be an explanation for variability in accretion rates and is thought to be a key factor in the formation of planetary systems (e.g. Janson et al., 2012).

Circumstellar disks drive bipolar outflows along the rotational axis of the YSO (e.g. Krumholz, 2015a). These outflows provide a mechanism by which angular momentum from the disk can be removed and envelopes dissipated (Frank et al., 2014). Outflows can also drive the formation of ionised H II regions where they interact with their environments, this is especially notable in massive stars which are key in sculpting

their wider environments (see Sect. 1.1.4).

When the circumstellar disks are mostly dissipated (see e in Fig. 1.1) accretion ceases, marking the beginning of the transition on to the main sequence. As the YSO continues to contract hydrogen burning for sources with $M \gtrsim 0.075\,M_{\odot}$ begins (e.g. Dantona & Mazzitelli, 1985).

1.1.3 Observing YSOs

The process of low-mass star formation described above can be subdivided into four stages as per Shu et al. (1987) and Lada & Shu (1990),

- the formation of a dense core in a cloud as gravitational attraction overcomes support mechanisms;
- forming a YSO at the centre of the collapsing core;
- a rotational disk and bipolar outflows are formed;
- cessation of infall as the YSO moves onto the main sequence.

Linking these stages to observed properties allows sources to be categorised reflecting their evolutionary stage. As previously noted, optical observations of YSOs can be inhibited by the thick envelopes from which they form. Short wavelength emission radiated from the central source is reprocessed by the surrounding dust and is re-emitted at infra-red (IR) wavelengths. Using IR fluxes between $2 - 20\,\mu$m Lada (1987) defines the spectral index (α) as,

$$\alpha = \frac{d\log(\lambda F_{\lambda})}{d\log\lambda}, \tag{1.23}$$

which allows YSOs to be separated into three classes,

- Class I: $\alpha \geqslant 0.3$;
- Class II: $-1.6 \leqslant \alpha \leqslant -0.3$;

- Class III: $\alpha < -1.6$.

Andre et al. (1993) add a fourth class, Class 0, for the least evolved and therefore most heavily embedded YSOs which are not detectable at 20 μm. The progression from Class 0 to III therefore covers the full range from initial core (Class 0), through an in-falling envelope onto a core with an accretion disk (Class I), the dissipation of the outer envelope leaving an accreting core and disk (Class II), to a YSO with a weak disk contracting onto the main sequence (Class III). In the low mass regime ($M \lesssim 2\,M_\odot$) Classes II and III are observed as classical and weak-lined T Tauri stars respectively. At intermediate masses ($2 - 8\,M_\odot$) Class II and III sources are identified as Herbig Ae/Be stars (Herbig, 1960). Class II and III YSOs are not observed at masses higher than $\sim 8\,M_\odot$, as these sources evolve rapidly onto the main sequence effectively bypassing these stages (see Sect. 1.1.4 for more details).

1.1.4 High mass star formation

While the process of star formation in the low mass regime is fairly well understood, at high masses ($M \gtrsim 8\,M_\odot$) our understanding is less well developed.

To form a high mass star requires the collapse of an equivalently massive cloud or region of a cloud. As shown in Sect. 1.1.1 any region of a cloud with $M > M_J$ will collapse if unsupported. Collapsing a massive reservoir of material would therefore form multiple low mass cores (Krumholz, 2015b). This behaviour is also seen in simulations which generally fail to produce massive stars (e.g. Dobbs et al., 2005), and has been called "fragmentation-induced starvation" (Peters et al., 2010; Girichidis et al., 2012; Prole et al., 2022). These simulations however are mostly dependent on gravity and hydrodynamics, and do not consider contributions from magnetic fields and radiation pressure. The effects of magnetic fields in supporting a cloud against collapse and removing angular momentum from circumstellar material, allowing faster accretion rates, have previously been discussed (see Sects. 1.1.1 and 1.1.2) and play an equally important role in the high mass regime (Tan et al., 2014). Luminosity scales as $L \propto M^n$,

and for stars with masses between $10-100\,\mathrm{M}_{\odot}$ n ranges between $3.1-1.6$, therefore radiation pressure becomes a significant factor for massive YSOs. In turn T relates to L following the Stephan-Boltzmann law for blackbodies,

$$L = 4\pi\sigma_B R^2 T^4. \tag{1.24}$$

Hence for massive stars the radiation heating of surrounding material may considerably raise the local Jeans mass (see Eq. 1.12). This effect has been reproduced in simulations (e.g. Krumholz et al., 2011; Bate, 2012; Matsukoba et al., 2022), however whether it alone can account for observed massive star rates is not certain. These models of monolithic collapse are grouped into "turbulent core" models (McKee & Tan, 2003). The problems caused by fragmentation are circumvented in models of competitive accretion (Bonnell et al., 2001; Bonnell & Bate, 2006).

Competitive accretion models propose a funnelling of material from large reservoirs along filamentary structures. Indeed cores have been observed along filamentary structures (e.g. Könyves et al., 2015). In this model those cores at the centre of the structure are able to accrete inflowing material at higher rates (illustrated in steps $1-3$ of Fig. 1.2). The high rate of mass inflow to the centre floods those cores with material before fragmentation effects can dominate. Cores outside the centre are limited in mass by fragmentary processes whilst those in the centre become high mass stars. This, by necessity, creates a scenario in which massive stars lie in the centre of a region, surrounded by low mass stars. This mass segregated arrangement is observed in some regions (Evans & Oh, 2022; Zhang et al., 2022), but is not ubiquitous with some massive stars forming in less dense regions (e.g. Wright et al., 2016). Furthermore evidence from observations (Dib et al., 2018) and simulations (Spera et al., 2016; Guszejnov et al., 2022) suggest that mass segregation can rapidly take place from less segregated natal arrangements.

As mentioned in Sect. 1.1.3, high mass sources evolve rapidly. By adopting typical values of luminosity ($\sim 10{,}000\,\mathrm{L}_{\odot}$), radius ($\sim 4\,\mathrm{R}_{\odot}$) and mass ($\sim 10\,\mathrm{M}_{\odot}$) for a B0

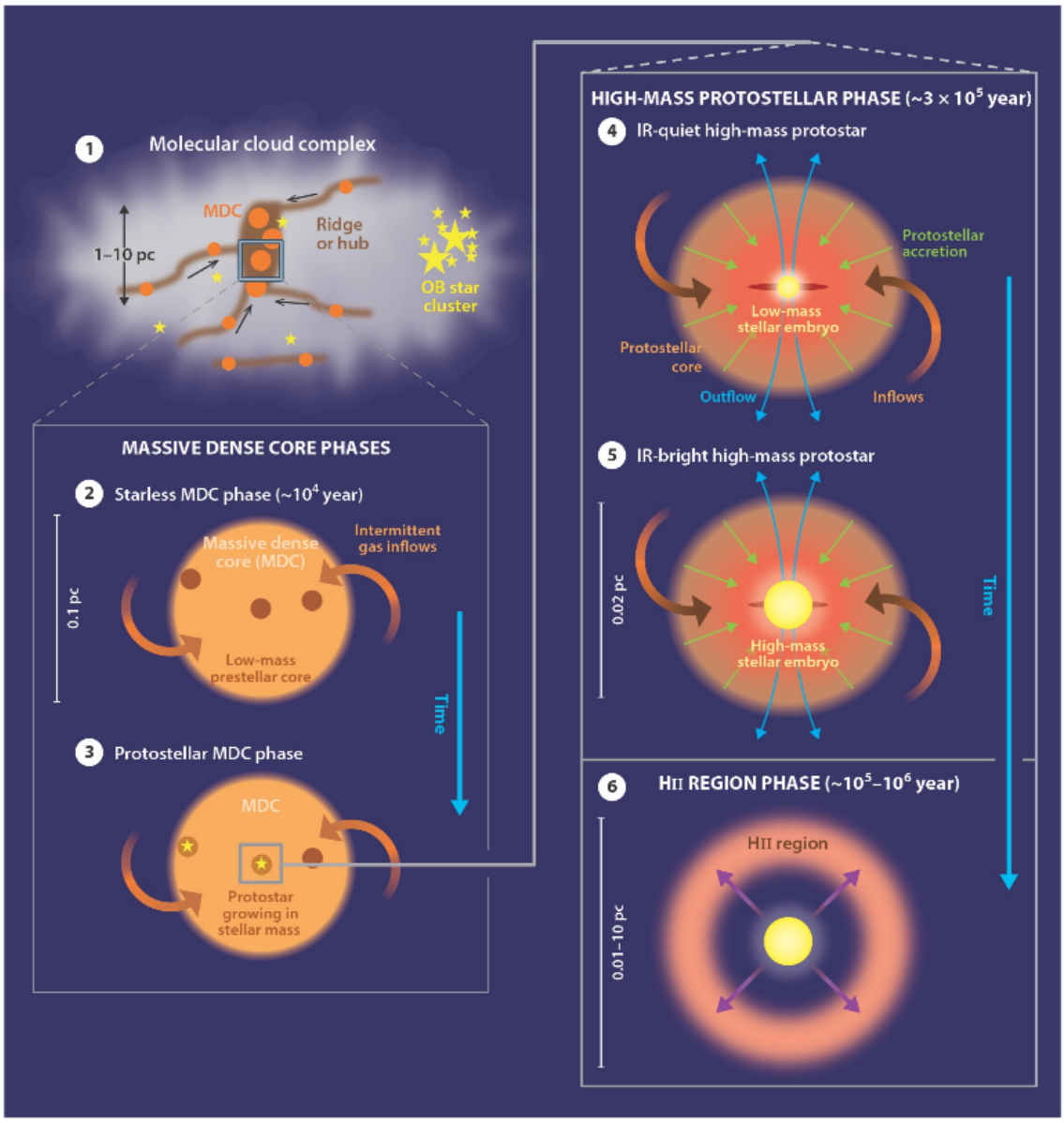

Figure 1.2: A diagram of the stages of massive star formation showing the inflow of material onto the central core in a cluster as proposed in competitive accretion models. Adopted from figure 5 of Louvet (2018).

star (Pecaut & Mamajek, 2013, updated in 2022 [1]) and using Eq. 1.21, a value of $t_{KH} \sim 10^5$yrs is found. This is comparable to the time necessary to accrete $\sim 10\,\mathrm{M}_\odot$ of material assuming a generous $\dot{M}_\odot \sim 10^{-4}\,\mathrm{M}_\odot\mathrm{yr}^{-1}$. Comparing these timescales for a YSO which will continue to contract as it evolves massive YSOs must begin hydrogen burning, and hence arrive on the main sequence, whilst still accreting. This accounts for why Class II and III low mass YSOs do not have analogues above $\mathrm{M} \sim 8\,\mathrm{M}_\odot$.

Given the above, and that massive stars (OB-types) are typically embedded for a significant fraction of their lifetimes ($\sim$15 per cent, Churchwell, 2002; Lumsden et al., 2013) it is necessary to accrete material onto already very luminous sources. Comparing the outwards pressure due to radiation with the force due to gravity gives a limit mass at which accretion is prevented,

$$\frac{GM}{r^2} > \frac{\kappa_R L}{4\pi r^2 c}. \tag{1.25}$$

Rearranging, with units of $\kappa_R = 10\,\mathrm{cm}^2\mathrm{g}^{-1}$ and solar units the point at which the forces balance is,

$$\left(\frac{L}{M}\right) = 1300 \left(\frac{L_\odot}{M_\odot}\right) \kappa_R^{-1}, \tag{1.26}$$

which occurs for stars with $\mathrm{M} \sim 20\,\mathrm{M}_\odot$. This would represent the maximum possible stellar mass. However observed stars with initial masses an order of magnitude higher than this are known in both the MW and MCs (e.g. Kashi & Soker, 2010; Crowther et al., 2016). The limiting mass given by Eq. 1.26 assumes spherically symmetric accretion, which is unrealistic given the effects of angular momentum conservation described by Eq. 1.22. Opacity in the disk shields material at greater radii from the energetic photons from the central source (Fig. 1.3). With the envelope collapsed towards the equatorial plane energetic jets can escape at both poles. The escape of radiation pressure parallel to the poles shown in the right-hand panel of Fig. 1.3 allows for increased accretion flows from the envelope onto the heated dusty disk. The dusty

[1] `http://www.pas.rochester.edu/~emamajek/EEM_dwarf_UBVIJHK_colors_Teff.txt`

disks produced via the conservation of angular momentum are therefore a necessity for, rather than a consequence of, high mass star formation.

Using this geometry, simulations have shown mass accretion can continue up to $40 - 140\,\mathrm{M}_\odot$ (e.g. Kuiper et al., 2010). The theoretical upper mass limit for star formation is highly dependent on cloud opacity and therefore metallicity (Jeřábková et al., 2018; Vink, 2018), with lower values of metallicity enabling higher mass star formation (Eq. 1.25). Stars with masses $\gtrsim 150\,\mathrm{M}_\odot$ are extremely rare (Weidner & Kroupa, 2004) hence only a few examples have been reported in distant MW star forming regions (SFR) and nearby galaxies (Martins et al., 2008; Crowther et al., 2010).

Evidence for disks around massive YSOs in the MW has been found as double-peaked velocity dispersions in narrow-band Brγ and spectroscopic observations (Blum et al., 2004; Kaper et al., 2011; Navarete et al., 2015; Van Gelder et al., 2020) and from direct imaging using high resolution millimeter observations (Ginsburg et al., 2018; Ilee et al., 2018; Maud et al., 2019). Simulations of circumstellar disks show varying levels of fragmentation within the disk (see section 9 of Oliva & Kuiper, 2020, for a review of disk fragmentation). Fragmentation in the disk has been linked to observed variations in accretion rates (Caratti o Garatti et al., 2017; Oliva & Kuiper, 2020).

Bipolar outflows or jets have been observed for over a century (Burnham, 1890). Where they interact with the surrounding interstellar medium ionised fronts are formed, these optically bright structures are known as Herbig-Haro objects (Reipurth & Heathcote, 1997). As previously noted outflows provide a mechanism by which angular momentum is removed from the YSO and help dissipate the circumstellar envelope. For massive YSOs the outflows can be highly energetic driven by accretion and magnetic fields (e.g. Commerçon et al., 2022), in turn causing material they interact with to exibit maser emission (e.g. Felli & Palagi, 1995; Cyganowski et al., 2009). Some types of maser emission are solely driven by the radiative excitation of material by high energy photons, i.e. radiatively pumped rather than via kinetic excitation (Gray, 2012). The detection of radiatively pumped OH and CH_3OH (methanol) masers associated with outflows (De Buizer, 2003) can therefore provide confirmation of an energetic driver

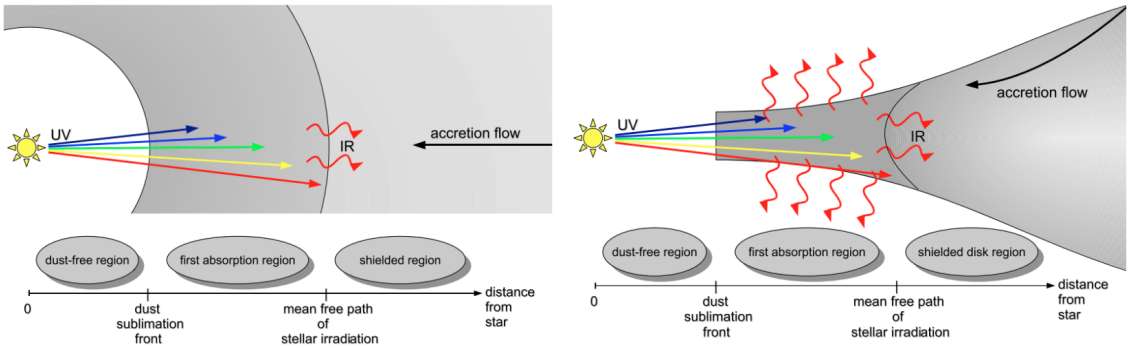

Figure 1.3: A diagram of the radiative forces on accretion in a spherical (left) and disk (right) system geometry. Adopted from figure 4 of Louvet (2018).

and for YSOs thereby confirm a high mass nature (De Villiers et al., 2015).

Massive YSOs which have begun hydrogen burning emit short wavelength photons capable of ionising their environments. This occurs around each YSO in a Strömgren sphere (Kuiper et al., 1937; Strömgren, 1939), where atomic hydrogen is stripped of electrons to form an ultra-compact ($r \leqslant 0.1\,\mathrm{pc}$, Beuther et al., 2007; Hoare et al., 2007) region of H II plasma (UCH II, see stage 6 in Fig. 1.2). The initial UCH II forms rapidly (e.g. Mottram et al., 2011) and slowly expands outwards. In regions where multiple massive YSOs are found these expanding H II regions can become very large and merge leading to photodissociation regions where they interact with surrounding material (Hollenbach & Tielens, 1997, 1999). This interaction can impart external forces onto otherwise quiescent molecular gas triggering further star formation (e.g. Hester & Desch, 2005; Tosaki et al., 2007, see also Fig. 1.4).

1.2 Massive YSO surveys

Early IR surveys identified massive YSOs in the MW from O-type star populations associated with UCH II regions which exhibited an IR excess. Using this methodology Wood & Churchwell (1989) identified 1646 embedded massive stars in the MW and 71

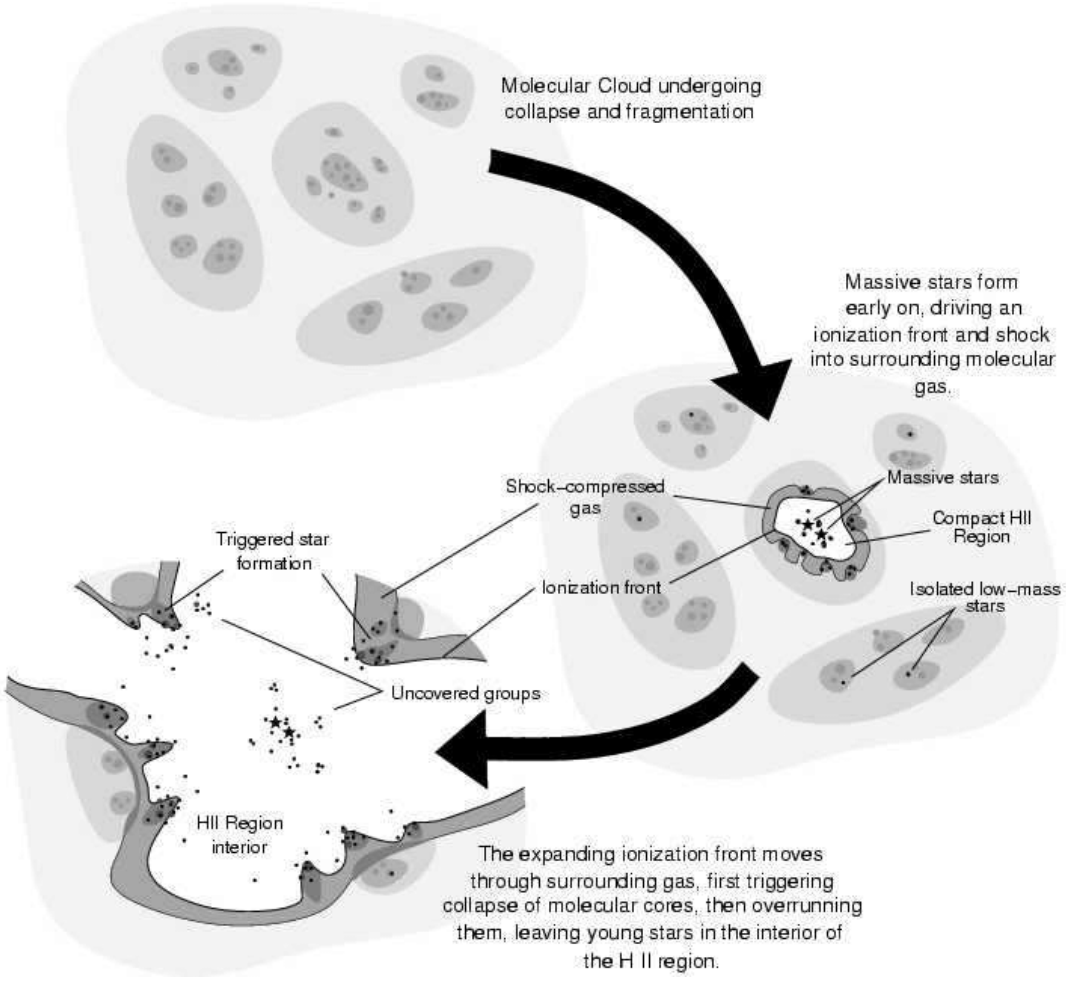

Figure 1.4: A diagram of the evolution of a H II region, showing the progression from initial formation of a population of massive stars through to the expansion of the H II region triggering another generation of star formation. Taken from figure 8 of Hester & Desch (2005).

in the MCs using *InfraRed Astronomical Satellite* (IRAS) all sky data. Their selection criteria required that sources had already begun to form a UCH II region, thus excluding YSOs at an earlier stage of formation. By using the same IR excess criteria as Wood & Churchwell (1989) but without the UCH II region association, Sridharan et al. (2002) expanded the earlier sample to include 69 less evolved sources within the MW. Whilst Prusti et al. (1992) utilise IRAS photometry to identify 5086 candidate YSOs within the MW, many of these are low mass sources.

Using the mid-IR *Midcourse Space Experiment* (MSX) point source catalogue (Egan et al., 2003) alongside near-IR Two Micron All Sky Survey (2MASS) data, Lumsden et al. (2013) created the Red MSX Survey (RMS survey), the largest survey of massive YSOs at that time. The RMS survey is estimated to be complete for YSOs more luminous than a B0 star at the distance of the Galactic centre (~ 8.5 kpc, Lumsden et al., 2013; Miville-Deschênes et al., 2017). In total ~ 2000 candidate YSOs were identified, of which later observations including near-IR spectroscopy have confirmed ~ 600 as massive YSOs and ~ 500 as UCH II regions (Cooper et al., 2013).

Extragalactic surveys of massive YSOs have become possible in recent years with both the Large and Small MCs (LMC and SMC respectively) the ideal initial targets due to their proximity ($d_{LMC} \sim 50$ kpc, Pietrzyński et al. 2013 and $d_{SMC} \sim 62$ kpc, De Grijs & Bono 2015). Using *Spitzer* (Werner et al., 2004) photometry, Whitney et al. (2008) first identified ~ 1000 candidate YSOs across the LMC, 458 of which are assigned high probability. Using new selection criteria on the same data, Gruendl & Chu (2009) recover an additional 1172 candidate YSOs, with 855 definitely categorised. The sources identified in Gruendl & Chu (2009) were rejected in the analysis of Whitney et al. (2008) due to their complex environments, causing them to fail their strict point source criteria. They further suggest that $20-30$ per cent of YSO candidates from Whitney et al. (2008) are in fact unresolved background galaxies. Carlson et al. (2012) recover 127 YSOs from earlier studies and identify a further 918 YSOs in the LMC not identified in previous works located in nine LMC H II regions, many of which are of lower mass. Similar photometric YSO surveys of the SMC have been conducted. Building on the techniques applied by Whitney et al. (2008) in the LMC, Sewiło et al.

(2013) identified 1007 candidate SMC YSOs.

A fraction of these photometric YSO candidates from MC studies were selected for IR spectroscopic follow-up to confirm their nature. This process is time intensive and so spectroscopic confirmation is rarely performed for all candidates from an earlier work. Follow up spectroscopy was performed with *Spitzer*-InfraRed Spectrograph (IRS, Houck et al., 2004) at mid-IR wavelengths. Mid-IR features of YSO spectra used to confirm the nature of candidate YSOs vary as the YSO evolves. At early, embedded stages (Class 0 or phases a and b in Fig. 1.1), YSO spectra display absorption features due to ice molecules such as H_2O, CO and, CO_2 (e.g. Oliveira et al., 2011). As the still embedded YSO begins to increase its temperature (Class I or phase c in Fig. 1.1), ices are destroyed and silicate absorption becomes the characteristic spectral feature (e.g. Jones et al., 2017). Once feedback from the YSO begins to disperse the circumstellar envelope the YSO is able ionise its surroundings (Class II or phase c in Fig. 1.1). As this occurs silicate absorption features become less clear as emission from polycyclic aromatic hydrocarbons (PAHs) becomes apparent (phase d in Fig. 1.1; see also, Jones et al., 2017). For YSOs of intermediate mass which are analogues of Galactic Herbig Ae/Be objects silicate emission can be detected (Oliveira et al., 2013; Jones et al., 2017).

Oliveira et al. (2009) and Seale et al. (2009) present *Spitzer*-IRS observations of LMC YSOs; a compilation of the classifications of all LMC point sources is presented in Jones et al. (2017). Using a combination of *Spitzer*-IRS spectroscopy and photometry, alongside optical spectroscopy and radio data, Oliveira et al. (2013) confirm the YSO nature for $\sim$30 high mass SMC sources; a compilation of SMC *Spitzer* classifications is presented by Ruffle et al. (2015). Near-IR spectroscopy, also used in the MW by Cooper et al. (2013), was used to further characterise massive YSOs in the MCs (Ward et al., 2016, 2017; Reiter et al., 2019; Jones et al., 2022).

Beyond the MC, studies to identify YSOs in NGC 6822 and M 33 has been recently performed, discussed in Sects. 1.4.1 and 1.4.2 respectively. Spectroscopically confirmed MC YSOs are used to train the machine learning algorithm as detailed in Sects. 3.3, 4.2 and 5.1.

1.3 Star formation on galactic scales

The energetic processes at the birth and death of high mass stars are significant drivers of change in their environments (Russeil, 2003; Louvet, 2018; Zari et al., 2021; Chevance et al., 2022). By identifying sites of star formation in relation to the structure of the host galaxy theories linking galaxy structure and interactions with star formation can be constrained. As briefly mentioned, star formation is known to correlate with the available gas on kiloparsec scales, following the Kennicutt-Schmidt relation (Schmidt, 1959; Kennicutt, 1989). The locations of star formation activity are therefore often inferred at large scales via tracers such as the gas required for star formation or other features associated with young populations such as H II regions. In Fig. 1.5 the distributions of H_2 gas surface density and H II regions in the MW are shown in relation to the MW spiral arms. Whilst as previously noted due to our location within the MW's disk a full survey of it is difficult, where data is available a strong correlation between arm structure and these tracers of star formation is seen.

1.3.1 Spiral galaxies

Since redshift $z \sim 2$ most stars have formed in the disk of spiral galaxies (Casey et al., 2014; Dobbs & Baba, 2014). The most apparent feature of a spiral galaxy are its eponymous spiral arms. Spiral galaxies can have a few distinct arms (grand design spirals) or multiple, less distinct arms (flocculant spirals), and can be with or without a central bar. Whilst it is often observed in grand design spiral galaxies that the arms begin at the ends of the central bar, and grand design spirals are more frequently observed in barred galaxies (Elmegreen et al., 2011), the connection between bar and spiral arms is not fully understood (e.g. Mo et al., 2010; Dobbs & Baba, 2014).

Analysis of stellar populations shows that spiral structure is only weakly represented in older populations and is strongest in gas and young stars (e.g. Elmegreen, 2011, see also Fig. 1.5). The process of star formation however is complex and the relation to this disk structure is not fully understood (Baba et al., 2016) and is being

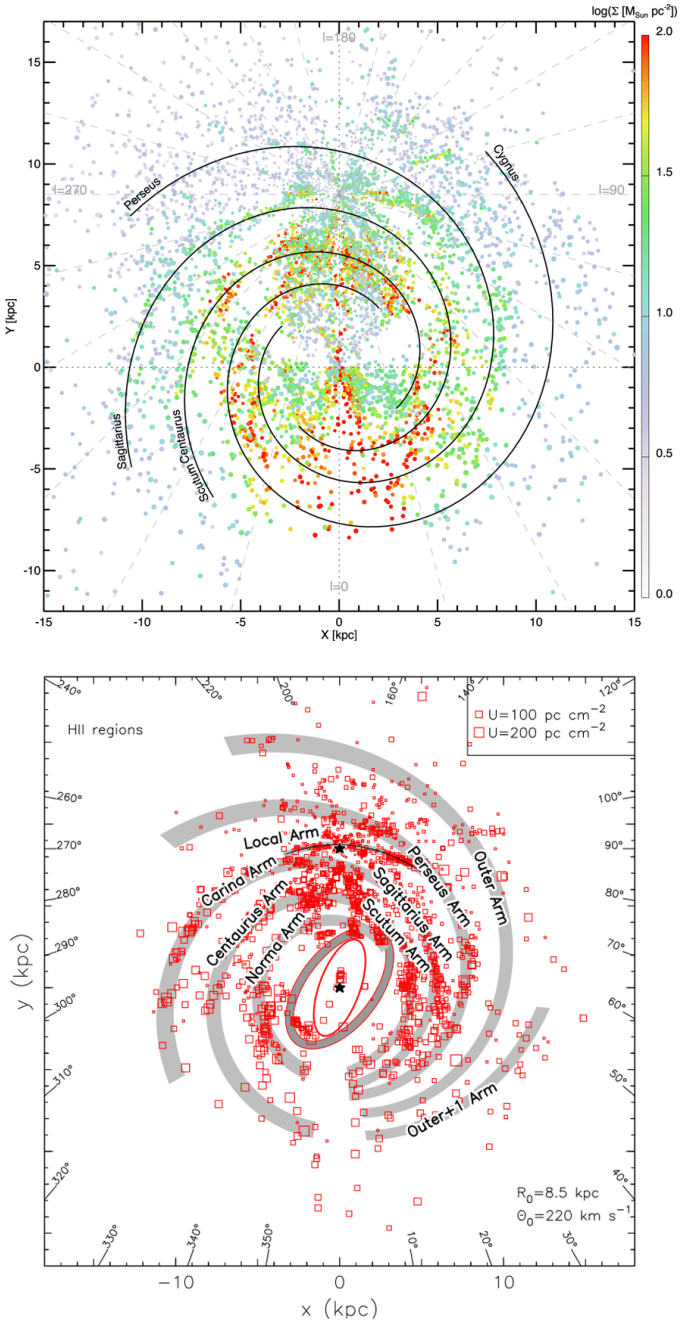

Figure 1.5: Distributions of MW H_2 surface density (top, figure 10 from Miville-Deschênes et al. 2017), and H II regions (bottom, figure 5 of Hou & Han 2014). Both diagrams show the major MW spiral arm structure in overlays, the location of which correlate strongly with gas and young population distributions.

actively explored (e.g. Dobbs et al., 2022). The two dominant mechanisms proposed for generating spiral structure in galaxies are quasi-static spirals or density wave theories (hereafter QSS, Lin & Shu, 1964), and transient or dynamic spiral formation (DSF, Goldreich & Lynden-Bell, 1965; Dobbs & Baba, 2014). Whilst these models are not strictly mutually exclusive, each describes behaviour on different timescales and appear to be favoured by different types of spiral galaxy (Dobbs & Baba, 2014).

QSS models predict spiral structure to be long lived ($\gtrsim 1$ Gyr), longer than the typical rotation period of the disk (e.g. Bertin & Lin, 1996). Therefore, material in the disk will overtake multiple spiral arms and interact with the shock fronts along the arms (Roberts, 1969; Shu et al., 1972; Lee & Shu, 2012; Lee, 2014) within two arm transits (Woodward, 1975; Wada & Koda, 2004; Wada, 2008). The shock front arises from a piling up of upstream material in the potential minimum of the arm (Toomre, 1977) akin motorway traffic jams (Dobbs & Baba, 2014) in which the braking of the first vehicle causes a progressing tailback in those following. In QSS it is expected that the gas velocity changes suddenly along the shock front, as has long been observed in nearby spiral galaxies M 51 (e.g. Tully, 1974) and M 81 (e.g. Visser, 1980). Another prediction of QSS is the transition, across the width of an arm, from the precursors of star formation ahead of the arm, via active star formation to older populations in the wake of the arm. QSS simulations predict monotonic progression of the shock upstream in relation to the arm with increased galactocentric radius (Gittins & Clarke, 2004; Baba et al., 2015); this progressive offset may also be observed in the stellar and pre-stellar populations. Whilst QSS is most easily considered in grand design spirals, by invoking shorter timescales, density standing waves have been reproduced which more closely match multi-armed, flocculent spiral galaxies (Bertin et al., 1989).

Contrary to the long lived spirals in QSS, DSF predicts spiral structures that change on timescales of a few hundred Myrs in all spiral types (Sellwood, 2011; Baba, 2015; Kumamoto & Noguchi, 2016). Rather than a model in which material is rotating at different rates to spiral structure and is therefore swept through an arm, DSF proposes that material falls onto the arms from both what in QSS would be up and downstream directions (Dobbs & Bonnell, 2008; Wada et al., 2011). There is there-

fore no offset of star formation to the potential minimum of an arm in DSF (Baba et al., 2015), and no systematic progression of star formation across the arm. DSF is typically used to explain flocculant spirals which emerge more readily from DSF simulations than grand design spirals (Dobbs & Baba, 2014). Using gas velocities as a tracer, Baba et al. (2016) predict observational patterns in simulations which can be used to differentiate between QSS and DSF models.

Other theories such as self-propagating stochasic star-formation (SPSSF, Mueller & Arnett, 1976; Gerola & Seiden, 1978) and star formation linked to tidal interactions (Holmberg, 1941) have also been proposed as mechanisms of spiral formation, however these are not currently thought to be the primary drivers of spiral structure. Both SPSSF and tidal interactions may however play a more significant role in dwarf galaxies.

1.3.2 Irregular dwarf galaxies

SPSSF may explain the progression of star formation in irregular dwarf galaxies or localised regions of sequentially triggered star formation such as very large and multigenerational H II regions (Tosaki et al., 2007; Dobbs et al., 2011; Gusev & Shimanovskaya, 2019). In dwarf galaxies a few large H II regions may represent a significant fraction of the total sites of star formation activity in the galaxy (e.g. Jones et al., 2019). Understanding star formation in these regions can therefore give a reasonably complete overview of star formation in low mass galaxies. For an isolated dwarf galaxy, SPSSF therefore offers a mechanism by which sites of star formation can be traced back over multiple generations of star formation. However, in practice, this idealised model is often complicated by external factors such as interactions with other galaxies.

One of the best studied occurrences of dwarf galaxy interactions is that of the MC, with current consensus pointing towards a direct collision, widely accepted to have occurred ~ 0.2 Gyr ago (Oey et al., 2018; Zivick et al., 2019; Schmidt et al., 2020). Interactions between galaxies impart tidal forces upon the ISM within each galaxy and can transfer new material into an otherwise quiescent galaxy. The two most active sites of high-mass star formation in the MC, the large H II regions 30 Doradus and N 44

Table 1.1: Properties of the galaxies discussed in this work. Values are taken from: [(1)] De Grijs & Bono (2014), [(2)] Braine et al. (2018), [(3)] Corbelli et al. (2014), [(4)] Lee et al. (1993), [(5)] Richer & McCall (2007), [(6)] Madden et al. (2014), [(7)] De Grijs & Bono (2015), [(8)] Skillman et al. (1989), [(9)] Besla (2015b), [(10)] Pietrzyński et al. (2013), [(11)] Williams et al. (2021), [(12)] Van der Marel (2006).

Galaxy name	Distance (kpc)	Distance modulus (μ, mag)	Metallicity ($Z_\odot$)	Stellar Mass ($M_\odot$)
M 33	$850^{(1)}$	$24.67^{(1)}$	$0.5^{(2)}$	5.5×10^{9} [(3)]
NGC 6822	$490^{(4)}$	$23.34^{(4)}$	$0.2^{(5)}$	1.5×10^{8} [(6)]
SMC	$62^{(7)}$	$18.97^{(7)}$	$0.2^{(8)}$	3.1×10^{8} [(9)]
LMC	$50^{(10)}$	$18.49^{(10)}$	$0.4^{(11)}$	2.7×10^{9} [(12)]

(both within the LMC) have been suggested to be tidally triggered (Fukui et al., 2017; Tsuge et al., 2019) by SMC gas slamming onto the LMC's disk. 30 Dor itself is one of the largest H II regions in the Local Group (e.g. Lebouteiller et al., 2008). Tidal forces and flows transferring material between the MCs have also been proposed as the drivers of star formation in the SMC regions NGC 602 (Fukui et al., 2020), N 83 and N 84 (Ohno et al., 2020).

1.4 The Galaxies

The main goal of this project is to characterise star formation across a whole spiral galaxy, by identifying individual YSOs and star forming regions. M 33 was chosen for this analysis due to its relatively face-on inclination ($i = 54°$, De Vaucouleurs et al. 1991) making it a more favourable target over the similarly distant M 31 which is seen nearly edge on (e.g. Ma, 2001). NGC 6822, a dwarf-irregular galaxy, was used to develop the analysis techniques, owing to its convenient intermediate distance between the MC and M 33 and the fact that individual YSOs have been characterised there previously. In this section I describe properties of NGC 6822 and M 33. Selected properties of both NGC 6822, M 33 and the MC are provided in Table. 1.1 for reference.

1.4.1 NGC 6822

As noted in the previous section, resolved star formation has been extensively studied on large scales in both the MW and MCs. Stepping out to a distance of $\sim 490\,\mathrm{kpc}$ (Lee et al., 1993; Mateo, 1998) NGC 6822 is the closest star-forming dwarf irregular galaxy to the MW beyond the MCs. With no known companions (see for example De Blok & Walter, 2000), and no interactions with large Local Group spiral galaxies (e.g. McConnachie et al., 2021), NGC 6822 presents itself as a non-tidally disrupted analogue to the SMC. By understanding how star formation progresses in NGC 6822 the impact of tidal interactions on triggering star formation can be better constrained. NGC 6822 has a metallicity approximately equal to that of the SMC ($\sim$0.2 $\mathrm{Z}_{\odot}$, e.g. Skillman et al., 1989; Richer & McCall, 2007). Understanding massive star formation in a metal poor environment has implications for studies of the early universe, NGC 6822 and the SMC are analogues for typical star forming galaxies at $z \sim 2$ (Hirschauer et al., 2020).

NGC 6822 is relatively gas-rich, with very conspicuous large scale East-West 'wings' of H I gas (Volders & Högbom, 1961). The total H I mass is estimated to be $1.38 \times 10^8\,\mathrm{M}_{\odot}$ (Mateo, 1998), and the molecular and dust masses are respectively $M_{\mathrm{mol}} < 1 \times 10^7\,\mathrm{M}_{\odot}$ (Gratier et al., 2010a) and $M_{\mathrm{dust}} = 2.9^{+2.8}_{-0.8} \times 10^5\,\mathrm{M}_{\odot}$ (Rémy-Ruyer et al., 2015). Using these mass estimates, Schruba et al. (2017) find a gas-to-dust ratio of $480\,^{+170}_{-240}$. The total stellar mass is $1.5 \times 10^8\,\mathrm{M}_{\odot}$ (Madden et al., 2014), giving an observed baryonic mass of $\sim 2.9 \times 10^8 \mathrm{M}_{\odot}$. Weldrake et al. (2003) find a total dark matter mass to 5 kpc (the extent of the H I disk) of $\sim 3.2 \times 10^9\,\mathrm{M}_{\odot}$, implying that NGC 6822 is heavily dark-matter-dominated.

The H I gas distribution in NGC 6822 has a very intricate structure. It is dominated by a large under-density or cavity seen to the South-East of the main galaxy body (e.g. Gottesman & Weliachew, 1977; De Blok & Walter, 2000, see also Figs. 1.6 and 4.1). The inner rim of this cavity is edged by optical emission that could be linked to its origin in large-scale stellar feedback (Cannon et al., 2012), although no agreement exists on the mechanism responsible (De Blok & Walter, 2000). Opposing this feature on the North-West wing of the main H I distribution there is a large over-density of

gas. It has been suggested that this over-density is due to the presence of a putative interacting companion (e.g. De Blok & Walter, 2000, 2003). However this hypothesis is not supported by stellar population studies across NGC 6822 (Cannon et al., 2012). A likely explanation for the complex extended H I structure in NGC 6822 is a warped disk inclined with respect to the line of sight (e.g. Cannon et al., 2012).

Clearly apparent in NGC 6822 is the central bar which runs nearly perpendicular to the H I gas distribution in a North-South direction for ~ 1.4 kpc (~ 10 arcmin; see Fig. 1.6). This central bar is host to the young stellar component of the galaxy, with older populations more elliptically distributed (e.g. Letarte et al., 2002; Hirschauer et al., 2020). The central bar is boxed at either end by bright SFRs first identified by Hubble (1925) with ages up to 10 Myr (Efremova et al., 2011; Bianchi et al., 2012). Attempts to find sites of star formation beyond the bar, namely in the H I over-dense region, have so far been unsuccessful (Schruba et al., 2017) despite promising indicators in the distribution of H I gas (De Blok & Walter, 2000, 2003).

CO emission is often used as a proxy for molecular hydrogen (which does not emit at radio wavelengths) due to their general spatial coincidence. No CO maps of the entirety of the central bar of NGC 6822 have yet been produced, with published studies focusing on the brightest SFRs (Gratier et al., 2010a; Schruba et al., 2017). Schruba et al. (2017) produced ALMA high-resolution maps of several small (110×110 arcsec^2) fields in CO (2−1), four of which are centred on the most prominent SFRs: Hubble I/III, IV, V and X. They find CO cores with typical sizes of ~ 2.3 pc and propose that such small scales could be the cause of the low levels of CO emission seen in many dwarf galaxies, due to poor beam filling at lower resolutions.

Previous studies of resolved YSO populations in NGC 6822 on a galaxy wide basis have used established colour-cuts (Jones et al., 2019) or basic statistical (Hirschauer et al., 2020) classification criteria. In Jones et al. (2019) candidate YSOs were found using a series of mid-infrared (mid-IR) colour-magnitude diagram (CMD) cuts developed by Whitney et al. (2008) and Sewiło et al. (2013) for the MCs. The spectral energy distributions (SEDs) of those candidates were fitted initially using stellar atmosphere models (Castelli & Kurucz, 2003) in order to remove contaminant objects. The

Figure 1.6: A European Southern Observatory (ESO) composite image of NGC 6822 comprised of Atacama Large Milimeter/submilimetre Array (ALMA), Very Large Array (VLA) and 2.2−metre ESO telescope WFI optical images. WFI optical B, V, R, Hα are blue, green, yellow and red respectively. H I gas detected by VLA is shown by diffused blue halo extending beyond the central bar of the galaxy. ALMA CO observations are shown in orange and together with Hα emission reveal the locations of major star forming regions. This image covers an area of approximately $23 \times 26\,\mathrm{arc\,min}^2$.

sources remaining were then compared to YSO model grids (Robitaille et al., 2006; Robitaille, 2017). Sources were assigned to one of three confidence levels based on the goodness-of-the-fit to the best-fit model and colour-cut criteria.

In addition to the four well-known SFRs already mentioned (Hubble I/III, IV, V and X), Jones et al. (2019) studied in detail three other significant SFRs, which they label Spitzer I, II and III (their table 9 provides the positions and sizes). To the South of Spitzer I lie regions identified in Hubble (1925): Hubble VI and VII, a young open star cluster (Chandar et al., 2000) and a globular cluster (Huxor et al., 2013) respectively, while Spitzer II borders Hubble IX, a cluster of undetermined age (Huxor et al., 2013). Jones et al. (2019) remove from their YSO candidate lists any sources within the half-light radius of the globular cluster Hubble VII. Spitzer I is particularly prominent with an infrared excess noted by Cannon et al. (2006) and CO (2−1) emission identified by Gratier et al. (2010a). This region seems to be more active in terms of star formation than the other optically brighter Hubble regions (Jones et al., 2019).

Using the same near-infrared (near-IR) and mid-IR catalogues, Hirschauer et al. (2020) applied colour cuts developed using kernel density estimate techniques to separate different stellar populations. YSO candidates were identified based on consistent CMD positions as well as being located within one of the SFRs discussed in Jones et al. (2019). The major SFRs were all recovered in the resulting YSO distribution, however fewer YSOs were identified compared to Jones et al. (2019) due to different limiting magnitude cuts applied to the classifications. My own work, described in Kinson et al. (2021) and Chap. 4, identified YSOs in all the SFR detailed above.

1.4.2 M 33

Studies of 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 ($M_{\rm gas} \sim 3 \times 10^9\,{\rm M}_\odot$, Corbelli 2003; $M_* \sim 5.5 \times 10^9\,{\rm M}_\odot$, Corbelli et al. 2014; Kam et al. 2017), after the Milky Way and M 31. M 33 lies at a distance of $\sim 850\,{\rm kpc}$ ($\mu_{\rm M33} = 24.67\,{\rm mag}$, De Grijs

& Bono 2014) and extends to an apparent size of approximately 60 arcmin × 35 arcmin (Paturel et al., 2003).

The metallicity of M 33 is around half-solar (e.g. Braine et al., 2018), similar to that of the LMC (see figure 1 of Williams et al., 2021, and Table. 1.1). The metallicity of M 33 varies across the disk, 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 is consistent with an inside-out model of disk formation (Cioni, 2009; Williams et al., 2009), supported by the observed M 33 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 disk (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 et al., 1976; Corbelli et al., 2014), likely by a previous minor interaction with M 31 (Semczuk et al., 2018), the disk 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 categorised 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 within the last 40 Myrs (Long et al., 2002; Javadi et al., 2011).

The spiral arms of M 33 can be traced in the distributions of H I (Gratier et al., 2010b) 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). GMCs studied in M 33 show an evolutionary progression

Figure 1.7: An optical RGB (g, r, Hα respectively) ESO VLT Survey Telescope (VST) image of M 33. Sites of Hα emission, including regions of star formation are revealed in red. This image covers an area of approximately 57×68 arc mins^2.

which is associated with quasi-static arm models (Corbelli et al., 2017), more suggestive of QSS than DSF typically associated with flocculant spirals (see Sect. 1.3.1).

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 alongside GMCs (Gratier et al., 2010b; 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 et al., 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 YSOs (Fariña et al., 2012) and integrated mid-IR properties (Relaño & Kennicutt, 2009; Martínez-Galarza et al., 2012). Triggered star formation events have been theorised 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). Using GEMINI Near Infrared Imaging and Spectrometer (NIRI) JHK_s images with excellent seeing (FWHM ~ 0.35 arcsec) Fariña et al. (2012) identified 68 massive YSO and 11 Wolf-Rayet (WR) candidates within NGC 604. My own work, described in Kinson et al. (2022) and Chap. 5, identified YSOs across M 33 including NGC 604.

1.5 Project Objectives

My project aims to meet the following goals:

- **To develop a machine learning technique to identify YSOs from wide-scale IR survey data**

  Previously discussed surveys of entire galaxies to identify YSOs have employed a piece-wise approach based on individual magnitudes and colours, or are biased to known regions of star formation. Machine learning techniques provide

a more holistic method to classify YSOs, which does not rely on, and takes into account, interdependancies and degeneracies between observable features. To achieve a machine learning classification of YSOs a combination of near-IR and far-IR features is used in order to classify point sources via a Probabilistic Random Forest (PRF, see Sect. 2.2.2). Near-IR colours and magnitudes provide information on individual sources, whilst far-IR surface brightnesses sampled across the surrounding area inform the classifier about the environment in which the source lies. This will enable separation of Galactic foreground sources, background galaxies and different stellar populations in the target galaxy. Machine learning techniques are scalable and therefore offer an invaluable tool for disentangling and classifying large data sets. They can be applied across whole galaxies, avoiding bias towards only known sites of star formation.

- **To apply and validate machine learning techniques to NGC 6822, a galaxy with well-studied star formation activity**

  NGC 6822 offers a perfect laboratory in which to validate the machine learning techniques I developed. At around half the distance to M 33, it bridges the $\sim$17 factor in distance between the MC, the furthest distance at which large numbers of confirmed YSOs are available, and M 33. Massive YSOs have recently been identified in several major star forming regions across the centre of NGC 6822 (see Sect. 1.4.1). By applying a non-spatially biased machine learning technique I was able to classify YSOs across NGC 6822. Through the recovery of known star forming regions and individual literature YSOs, the reliability of this technique was assessed.

- **To identify YSOs across the whole disk of M 33**

  Applying the machine learning techniques across the disk of M 33, where a galaxy-wide study of resolved ongoing star-formation has not yet been performed, allows me to classify YSOs and identify regions of star formation by

examining the spatial clustering of classified sources. Using YSO model grids, mass estimates for each YSO are found and a star formation rate is estimated from direct YSO counts, the first such estimate for M 33. The spatial distributions of all stellar classes in M 33 are compared to the literature and models of spiral structure formation (described in Sect. 1.3.1). I also compare properties of SFRs identified in NGC 6822 and M 33, to infer their relative evolutionary status and context in terms of their large-scale galaxy structure.

2 Machine Learning Techniques

As discussed in Sects. 1.4.1 and 1.4.2, identification of YSOs has been performed in NGC 6822 and M 33 on the basis of piece-wise colour cuts and theoretical YSO models (e.g. Fariña et al., 2012; Jones et al., 2019; Hirschauer et al., 2020). These methods do not take into account interdependancies and degeneracies between observable properties (features) and, in the case of fitting models to spectral energy distributions, may be computationally expensive when scaled to large data-sets. Machine learning techniques provide a solution to both of these limitations.

Machine learning techniques can be separated into supervised – requiring human labelling of sources or other intervention, and unsupervised – purely data driven with little to no user direction. Both sets of techniques can be used for a variety of purposes including what machine learning terminology calls 'classification' problems. These are scenarios in which a data-set is organised into two or more discrete classes using feature information from each source. Both supervised and unsupervised methods are used in this project to classify the sources from their observed properties, this section outlines the basic principles of the methods employed.

2.1 Unsupervised Machine Learning

The absence of human intervention in unsupervised machine learning allows previously unknown relations in data to be found, and can also be useful in classification of data where labels may be unreliable. Any relations found by such unsupervised methods arise entirely from properties inherent to the data and therefore avoid potentially introduced biases which can be present in supervised machine learning.

There are many unsupervised machine learning approaches to data classification. One category of unsupervised methods are 'self organising maps', wherein unlabelled data is arranged by similarity allowing classifications to be made. One such method which has been applied previously to identify different stellar populations in astronomy

(e.g. Pennock et al., 2022b; Rim et al., 2022; Santiago et al., 2022) is t-distributed Stochastic Neighbour Embedding (t-SNE) (Van der Maaten & Hinton, 2008).

2.1.1 t-Stochastic Neighbour Embedding

A t-SNE implementation produces a 2D map of higher dimensional data in which sources of similar properties are often clustered. This is achieved by calculating the joint probability that any two data entries lie beneath a Student's t-distribution of one another in high dimensional space. The technique then attempts to minimise the divergence between the two-dimensional positions of the pair of entries and their higher dimensional counterparts. This results in a 2D map in which each source is located nearby to those with similar higher dimensional characteristics. t-SNE maps have been shown to be effective in separating unlabelled photometric sources in catalogue data (e.g. Steinhardt et al., 2020).

A consequence of the higher dimensionality pairing of probabilities is that t-SNE analysis requires all features to be present for a source to be mapped. Other limitations of the t-SNE method are the prohibitive run times for large catalogues and corresponding memory limitations of the system to hold the large volume of high dimensional information during the calculation process. Consequently t-SNE run times scale non-linearly with the increase in data set size and modest increases in data set size can lead to prohibitively long computation times. Whilst other unsupervised methods such as Principal Component Analysis (PCA, Hotelling, 1936) scale better for large data-sets, t-SNE offers significant advantages over PCA. PCA is inferior in maintaining the small scale structure of the data (i.e. clusters) and is more easily affected by outlier data as it preserves the global structure (variance) unlike t-SNE. Furthermore, unlike PCA, t-SNE does not assume linear relations between features, which is important in this case given the features used (see Sect. 4.1). Given these limitations, for any data-set a 'sweet spot' exists where good separation of the data is achieved without excessive run time.

A t-SNE implementation relies on several fine tuning parameters which can affect

the outcome of the map and run time including the number of iterations in the calculations and the number of neighbouring sources compared to in each calculation of the divergence difference, known as the perplexity value. This project uses the SKLEARN t-SNE implementation for PYTHON (Pedregosa et al., 2011) to create t-SNE maps as this allows for simple control of the iteration and perplexity parameters. Application of t-SNE maps in NGC 6822 are discussed in Sect. 4.7.

2.2 Supervised Machine Learning

Supervised machine learning is a useful tool in situations where a relationship in one set of data with a priori labels can be applied to a second set with the same measurements but unknown classification. Often in astronomy supervised machine learning involves training on a set of sources with previously confirmed properties and applying this to another set of data in which a specific object class is of interest, e.g. evolved stars in Hernandez et al. (2021b) or YSOs in Cornu & Montillaud (2020). One of the best established supervised classifiers is the Random Forest Classifier (RF, Breiman, 2001).

2.2.1 Random Forest Classifiers

An RF classifier in its simplest form is a set of randomised decision trees each of which return a classification for each source in the data. At each node in the tree a threshold value for a set of features is implemented which splits the decision path for the data input. This is repeated over a large number of randomly generated trees. The majority decision amongst all the trees is then given as the RF classification for each object.

The RF is trained on a subset of "known" sources against which its classification accuracy can be estimated. This is done by splitting the training data set into a sample for training and a sample against which the trained classifier is tested. Most commonly this is done randomly/pseudo-randomly with a random seed; the latter is the method used in this work (Sect. 4.3). The accuracy of the classifier on the test set is taken as

an estimate of the classification accuracy on the unlabelled data. 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. I took steps to counteract this effect as described in Section 5.2.

One method to visualise these estimates on a class-wise basis is to employ confusion matrices, a standard tool in supervised machine learning. Each matrix shows the statistics of the actual and predicted labels for the test sample sources. A confusion matrix for a perfect classifier will show a diagonal of 100 per cent accurate classifications. In practice however some classes may perform better than others; the matrices identify those classifications that are the most significant cause of confusion. Example matrices are presented in Sects. 4.3.1 and 5.3.

The accuracy of the classifier is inherently linked to the quality of the training data in both the extent of the feature parameter space covered by the objects in each class in the training set and the similarity of the training set to the data to be classified. Another consideration is the reliability of the classifications in the training set. The more sources with incorrect target classes in the training set the worse the RF will perform in classification. This can be minimised via conservative construction of the training set samples (Sects. 4.2 and 5.1) and using a random forest classifier which can account for uncertainties in the training data target class labelling. A probabilistic random forest classifier offers this capability, as well as being able to account for uncertainties in feature data and missing data.

2.2.2 Probabilistic Random Forest

A probabilistic random forest is a variation on the traditional RF approach which takes uncertainties into account in the features. For a source (i) in feature (j) with a mea-

surement (x) and associated uncertainty ($\Delta\, x$), a normal ($\mathcal{N}$) probability distribution (X) is generated such that:

$$X_{i,j} \sim \mathcal{N}(x_{i,j}, \Delta\, {x^2}_{i,j}) \tag{2.1}$$

The combined probability distribution in the relevant features is taken into account at each node of the randomised decision trees in the forest. The probability that source i propagates down either path (Ps) is split at each node based on $X_{i,j}$, rather than using a threshold condition against which each source is judged; an illustration of this is shown in Fig. 2.1.

Unlike a traditional RF model, which requires all feature data to be present, if feature information is missing in the data for a given node a PRF can propagate to the next nodes on an even split basis, i.e. $\mathrm{Ps} = 0.5$ for both paths. In this way a source with incomplete data can be classified with the caveat that such a classification is less reliable.

Figure 2.2 illustrates how an RF propagates only one path per source (yellow branches), whereas a PRF propagates all branches and finds a probability (p) at the end of each path (labelled 'Ideal PRF'). The probabilities calculated are such that for the theoretical PRF $\Sigma\, \mathrm{p} = 1$. The path with the highest probability is then given as the most likely classification. To aid computation time the PRF implementation of Reis et al. (2019), utilised in this project, does not propagate probabilities further down branches which fall below a threshold p value (branches with red crosses in the 'Approx. PRF' diagram of Fig. 2.2). In the numerical tests of Reis et al. (2019) no reduction in classification accuracy was seen with a threshold value up to 5 per cent, the default value in their implementation and the one used in this project.

Reis et al. (2019) showed that their PRF implementation offers an increase in classification accuracy over a RF in a toy data-set. Furthermore they show a PRF is able to outperform a RF in classification accuracy where class labels may be impure. The benefits of a PRF over a RF in the context of this project are explored in the following section using data from NGC 6822.

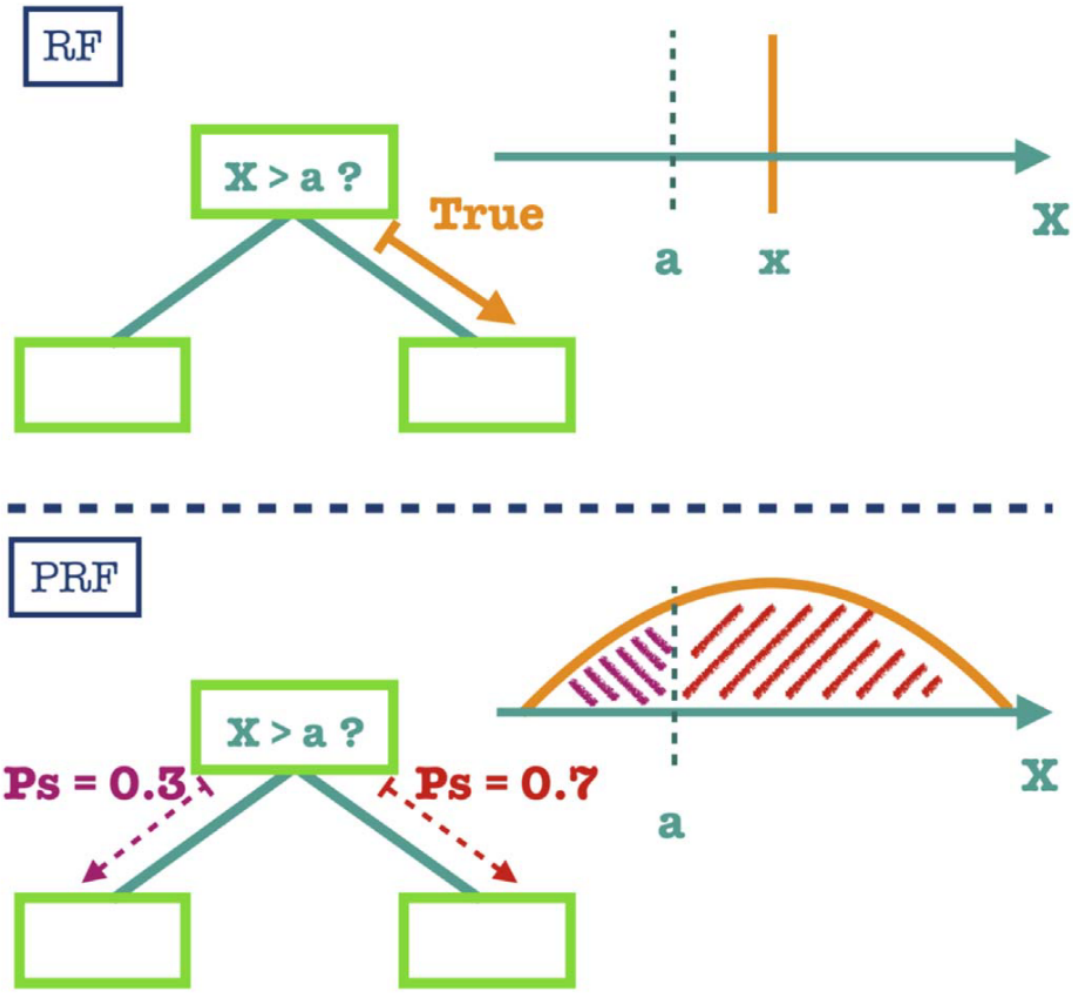

Figure 2.1: A diagram showing the different approach to path splitting at each decision node in an RF and PRF classifier. Whereas in an RF only the path which meets the threshold criteria 'X > a' is propagated, in a PRF the probability distribution for that source and feature is used to assign a likelihood of propagation down each branch from the node. This figure is reproduced from fig. 1 of Reis et al. (2019).

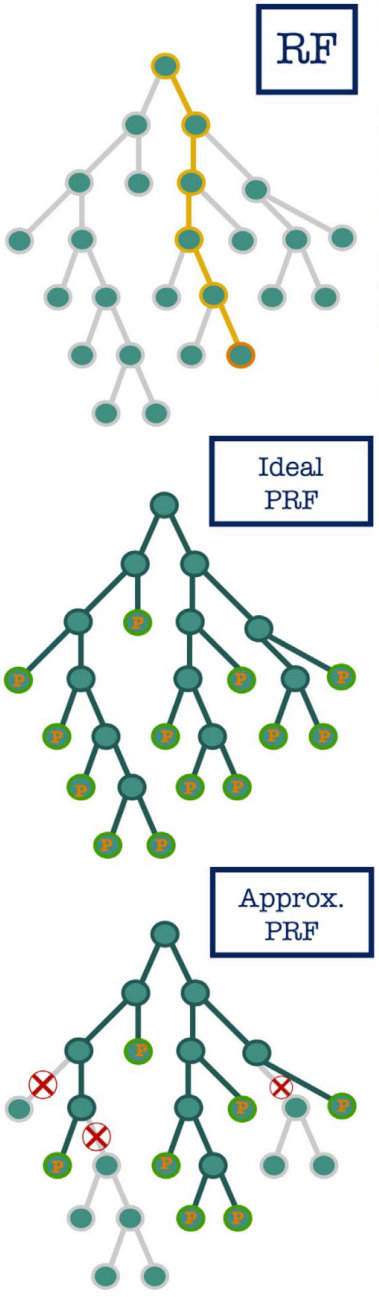

Figure 2.2: An example decision tree shown for an RF (top), an ideal – theoretical PRF (centre) and approximated – implemented PRF (bottom). In the latter case branches with probabilities below a threshold value are discounted from further propagation to aid computation time, see text for details. This figure is adapted from fig. 2 of Reis et al. (2019).

2.2.3 RF vs PRF comparisons

Both RF and PRF classifiers were run on a selection of the most important classes using the data from NGC 6822. This reduction in complexity was made to aid rapid comparison of the classifiers without replicating the full analysis performed in Chap. 4. The RF and PRF test runs were conducted multiple times with pseudo-randomly selected train/test splits to ensure any stochastic effects were accounted for.

In testing, no general improvement in average classifier accuracy over all classes was found by implementing a PRF rather than a RF, with both achieving ~ 89 per cent estimated accuracy. There was however a minor improvement in classes with fewer available training sources such as YSOs, the main goal of the classification. The use of a PRF over a RF improved the average classification accuracy of YSOs from 89 per cent to 92 per cent.

As previously discussed one benefit of employing a PRF over a RF is the ability to handle missing feature information. In NGC 6822 there are several sources with one missing feature due to the properties of the near-IR catalogue, hence a PRF increases the number of sources available to be classified from ~ 8000 to $\sim 12{,}000$ (for further details of the data see Chap. 3). In M 33 the near-IR catalogue does not have similar issues. Given, that the near-IR data in M 33 is taken from average catalogues (see Chap. 3), the inclusion of uncertainties with measurements is of increased importance. This is a key benefit of a PRF over a RF in the context of this project.

For the reasons of increased accuracy in the main class of interest, the ability to handle uncertainties, and the larger volume of data available for classification in NGC 6822, a PRF is the classifier chosen. The PRF implementation developed by Reis et al. (2019) is used as the main classifier for this project.

3 Data

This chapter describes the images and catalogues used in the analysis of the young stellar populations in NGC 6822 and M 33, as well as ancillary data used for the PRF classification of point sources and scientific interpretation.

3.1 NGC 6822 data

3.1.1 Near-IR images and point-source catalogues

3.1.1.1 Near-IR catalogues

The aperture photometry catalogue from Sibbons et al. (2012) is used in the NGC 6822 analysis. It was constructed using images obtained on the United Kingdom Infrared Telescope (UKIRT) using the Wide Field Camera (WFCAM, Casali et al., 2007). The focal plane array of WFCAM is comprised of four Rockwell Hawaii-II detectors (Casali et al., 2007). To fill in the gaps between the detectors four exposures are required, resulting in a tile image covering $\sim 0.89\,\mathrm{deg}^2$ (see Fig. 3.2) at a resolution of 0.4 arcsec per pixel. Several tiled images were used to construct the catalogue of Sibbons et al. (2012).

The catalogue contains $\sim$375,000 sources over an area of $3\,\mathrm{deg}^2$ centred on NGC 6822. The catalogue is estimated to be complete to depths of $J = 19.5$ mag and $K_s = 18.7$ mag (Sibbons et al., 2012). Full details on the data acquisition, reduction and catalogue generation can be found in Sibbons et al. (2012).

Star formation activity in NGC 6822 occurs in an area approximately 14×16 kpc (e.g. Letarte et al., 2002; Jones et al., 2019; Hirschauer et al., 2020). This area contains several significant star forming regions (see Fig. 4.1). The analysis in this project is not restricted only to these regions, but rather uses their recovery as a means of technique validation (see Chap. 4). A total area of approximately $0.07\,\mathrm{deg}^2$, covered by a single WFCAM detector, is considered in NGC 6822 (the yellow dashed region in Fig. 4.1).

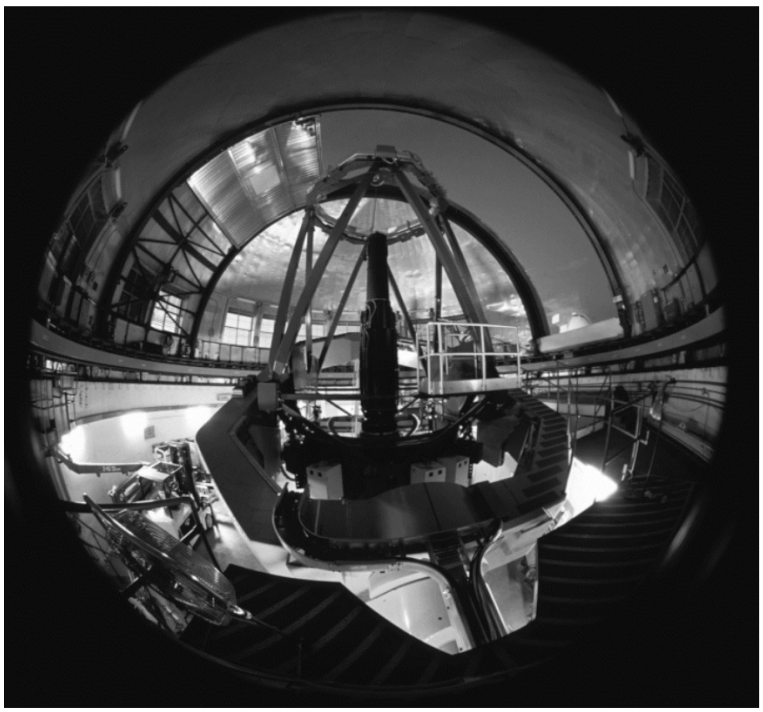

Figure 3.1: A photograph of WFCAM mounted on UKIRT, reproduced from fig. 1 of Casali et al. (2007). The telescope is pointed towards the zenith and the camera is the dark cylindrical element rising from the centre of the primary mirror.

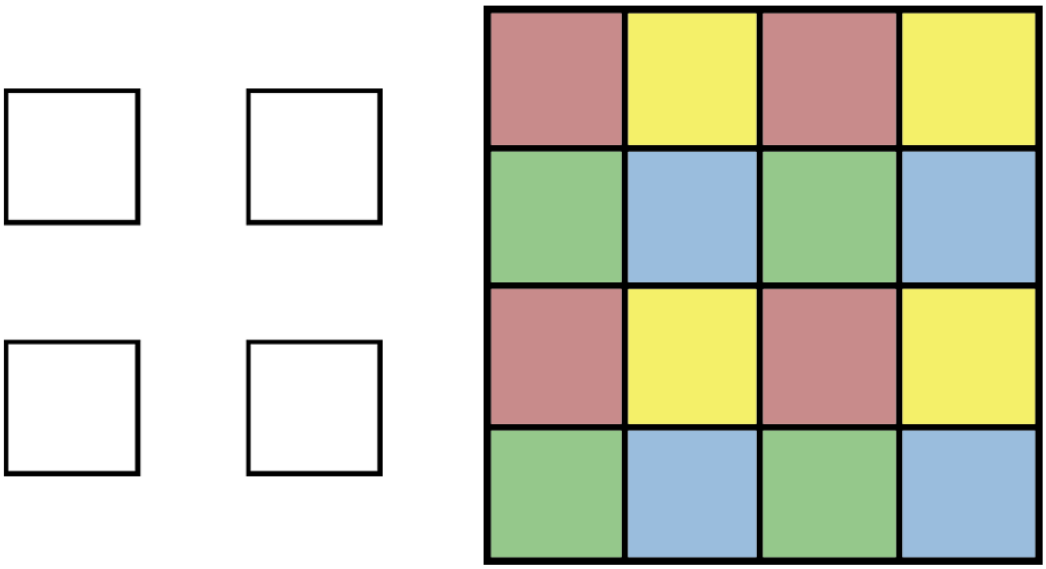

Figure 3.2: A representation of the layout of the four Rockwell Hawaii-II detectors on the WFCAM focal plane (left) and how four exposures are combined to achieve one tiled image (right). The colours on the right illustrate those portions of the tile imaged simultaneously.

The Sibbons et al. (2012) near-IR catalogues contains $\sim$15,000 point sources in this area.

3.1.1.2 Additional near-IR aperture photometry

Upon close inspection of both the catalogue and images it was apparent that the catalogue did not include aperture photometry of point-sources towards the central regions of the bright SFRs. In particular several *Spitzer* sources identified as YSOs in Jones et al. (2019) were seen in the images but did not have entries in the Sibbons et al. (2012) near-IR catalogue. It was thus necessary to extract additional aperture photometry in these regions. The JHK_s images were retrieved from the WFCAM Science Archive (WSA), fully processed using the standard WFCAM pipeline by CASU [2].

Aperture photometry was performed using the PhotUtils package for Python (Bradley et al., 2020). Using the WSA standard radius of 3.57 pixels, apertures were placed at the position of known *Spitzer* sources from Jones et al. (2019). In addition, to calibrate the new photometry apertures were placed on $\sim$9000 sources with near-IR photometry in the Sibbons et al. (2012) catalogue. A 1-σ dispersion of 0.055 mag or less was found in each band between the new photometry and that in the published catalogue for these calibration sources.

This process recovered near-IR magnitudes for an additional 54 sources located in bright SFRs which were added to the photometric catalogue. Magnitudes and uncertainties of the final catalogue are shown in Fig. 3.3. The photometric uncertainties for the added sources are at the higher end of the range of values seen in the extant near-IR data, reflecting the high background levels generally present in bright SFR.

In Fig. 3.4 the final near-IR catalogue, containing 11,341 sources, used in the analysis of NGC 6822 is presented as a Hess diagram in CMD space.

[2] https://research.ast.cam.ac.uk/vdfs/documentation.html#wsystem

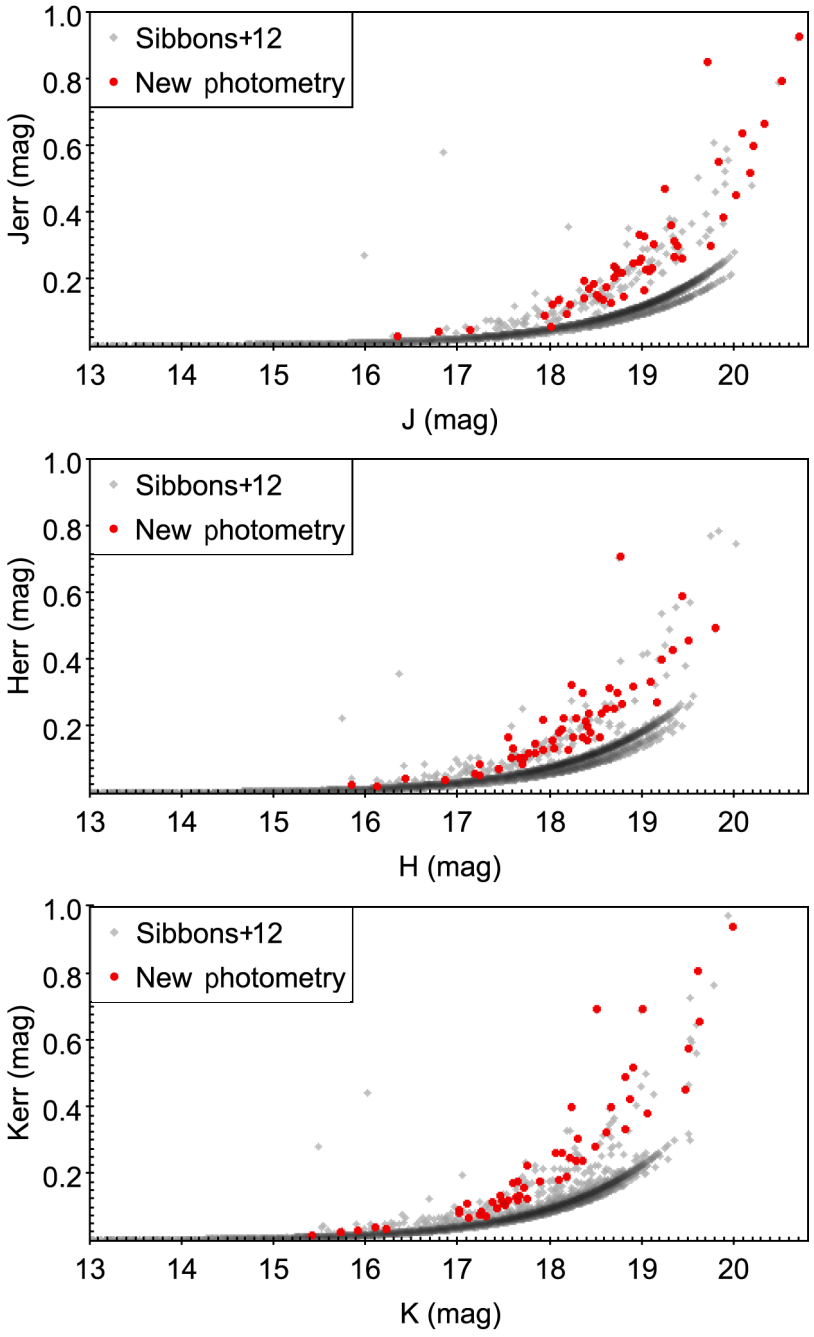

Figure 3.3: Magnitudes and uncertainties of the new aperture photometry (red circles) compared to those in the catalogue of Sibbons et al. (2012, grey circles). The reader is referred to the source paper for any data issues in that catalogue.

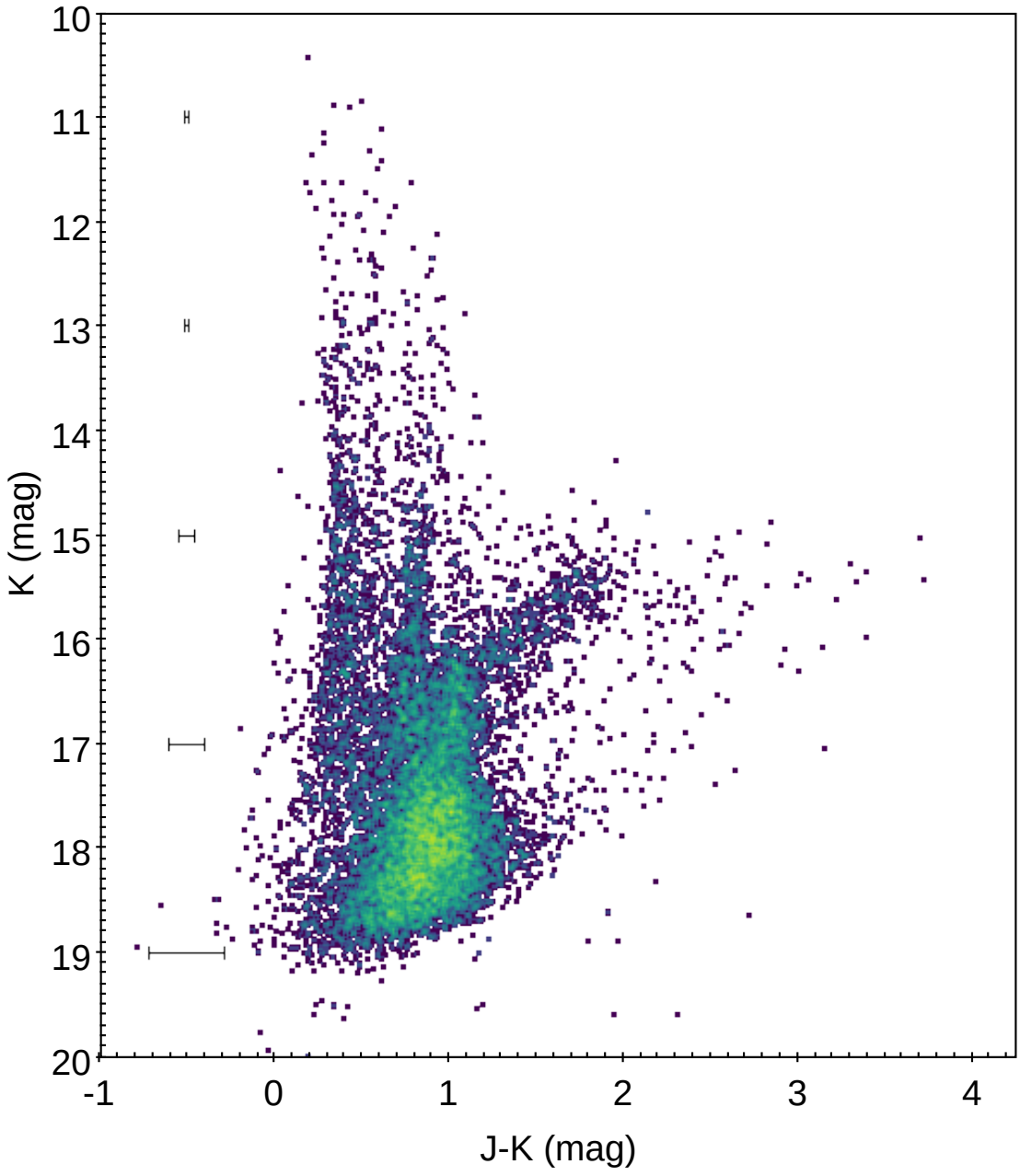

Figure 3.4: The near-IR catalogue for NGC 6822 shown as a Hess diagram in CMD space. Average error bars are shown.

3.1.2 Far-IR images and measurements

Light emitted by 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 near-IR source the neighbourhood far-IR brightness is used as an indicator of proximity to star-formation activity.

Galametz et al. (2010) obtained far-IR images of NGC 6822 using the Photodetector Array Camera & Spectrometer (PACS, Poglitsch et al., 2010) on the *ESA Herschel Space Observatory* (Pilbratt et al., 2010). The fully reduced PACS images at 70 and 160 μm were retrieved from the ESA *Herschel* Science Archive[3].

At the position of each K_s-band source, apertures were placed on the far-IR images using a large radius. The average count value is measured an aperture in both images using the same aperture radius. A radius equivalent to 30 pc (12.7 arcsec) around each source was chosen based on the typical scales from theoretical predictions of IR dark cloud sizes (Tan et al., 2014), and comparison with the CO emission tracing dust in NGC 6822 (Schruba et al., 2017).

3.2 M 33 data

3.2.1 Near-IR images and point-source catalogue

The near-IR catalogue for M 33 was constructed by Javadi et al. (2015) also using WFCAM data. A single tile of four separate pointing observations was obtained to cover a $\sim 0.89\,\mathrm{deg}^2$ sky area ($\sim 13\,\mathrm{kpc} \times 13\,\mathrm{kpc}$ at a distance of $\sim 850\,\mathrm{kpc}$, De Grijs & Bono 2014). Multi-epoch observations were made as part of a monitoring programme, from September 2005 to October 2007. 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) [4] and performed absolute

[3] `http://archives.esac.esa.int/hsa/whsa/`
[4] `http://wsa.roe.ac.uk/`

and relative photometric calibration. In my analysis, I used their catalogue of mean magnitudes of point sources towards M 33 for source classification (see Sect. 5.4). The catalogue contains $\sim$245,000 sources. I set the additional requirement that a source must be detected in all three JHK_s-bands, reducing the number of near-IR sources to $\sim$163,000 sources. The JHK_s, 5σ limiting magnitudes are 21.5, 20.6 and 20.5 mag respectively. Source density is shown in Fig. 3.5 and basic photometric properties are shown in a CMD in Fig. 3.6.

Since data was taken over multiple epochs and detector pointings, different regions of the science field-of-view reach varying depths. As shown in Fig. 3.5, 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 do not affect the accuracy of classification for individual sources, they become important when analysing the spatial distribution of sources of different types in M 33 (see Sect. 5.5).

3.2.2 Far-IR images and measurements

Far-IR brightness is used in M 33 as an indicator of proximity to star-formation activity. To this end, *Herschel*-PACS images at 70 and 160 μm obtained as part of the *HERschel M 33 Extended Survey* (HERM33ES, Kramer et al., 2010) were used, as retrieved from the ESA *Herschel* Science Archive[5].

At the position of each K_s-band source in M 33, an aperture of 30 parsec radius (7.2 arcsec) was used to measure an average brightness. Photometry was once again performed using the PhotUtils package for Python (Bradley et al., 2020).

[5] http://archives.esac.esa.int/hsa/whsa/

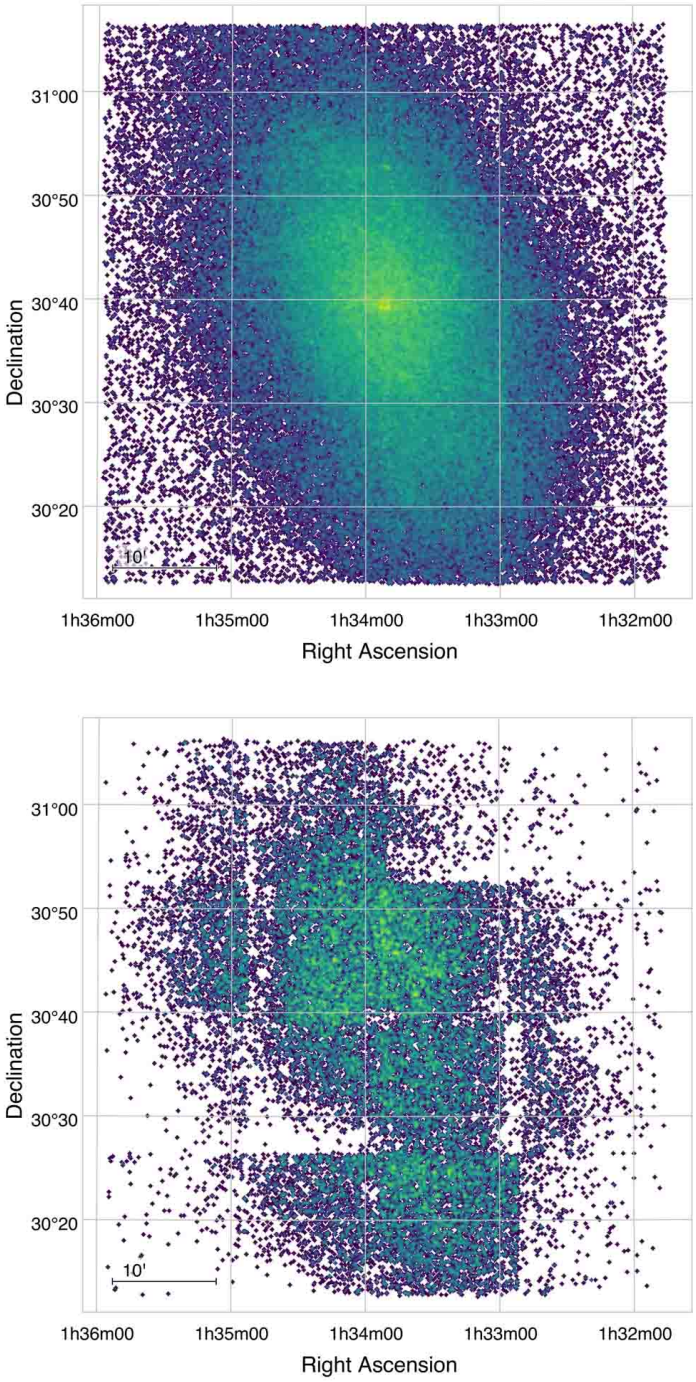

Figure 3.5: Hess diagrams of source density in M 33, brighter (top) and fainter (bottom) than $K_s = 19.2$ mag. The effects of variable depth in the catalogue across the field-of-view is clear at fainter magnitudes.

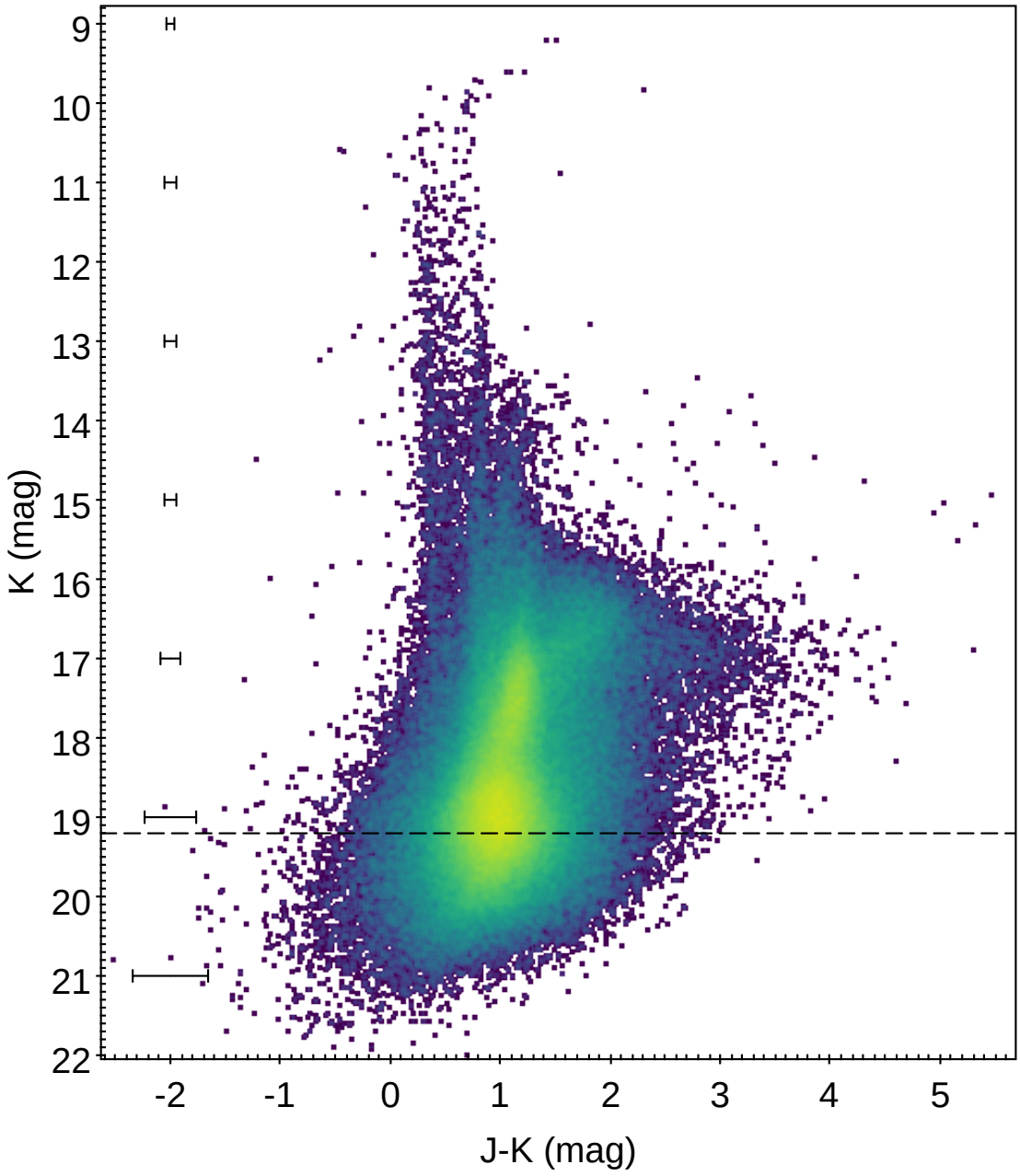

Figure 3.6: The M 33 near-IR catalogue presented in a CMD Hess diagram. Average error bars are shown. The dashed line at $K_s = 19.2$ mag indicates the magnitude at which the catalogue depth becomes very patchy (Fig. 3.5).

3.3 Magellanic Cloud data

My approach to designing the PRF training sets relies on samples of sources whose nature is constrained by methods other than photometric cuts. As a consequence there are not enough confirmed YSOs in NGC 6822 and M 33 to sample the YSO parameter space. It was therefore necessary to include additional confirmed YSO sources in the training data. I considered the Galactic YSO catalogue from the RMS Survey (Lumsden et al., 2013) of which a subsample have been confirmed as YSOs using near-IR spectroscopy (Cooper et al., 2013). Considering this latter sample, only a small fraction (< 20 per cent) of sources occupy the same magnitude range as the near-IR catalogue of the target galaxies (see Sects. 4.2.6 and 5.1.7). Furthermore the majority of those brighter RMS YSOs are extremely red (80 per cent of sources have $J - K_s > 4$). Systematically higher extinction is measured towards Galactic YSOs compared to e.g. Magellanic YSOs (Ward et al., 2017), likely due to differences in metallicity and dust properties (e.g. Jones et al., 2022). Consequently, I used instead the reasonably numerous sample of spectroscopically confirmed YSOs in the MCs, described below.

3.3.1 The Magellanic YSO sample

Whilst YSOs have previously been identified in NGC 6822 (Jones et al., 2019; Hirschauer et al., 2020) and in NGC 604 within M 33 (Fariña et al., 2012), these analyses rely on *Spitzer* colour cuts and/or SED fitting. To ensure classifier accuracy maintaining the purity of training set data is paramount. Hence the requirement for additional confirmation of their YSO status, beyond those based on broadband photometry, are set (see Sect. 4.2 and 5.1 for more details).

Instead the YSO training set was constructed from spectroscopically confirmed YSOs in the SMC (Oliveira et al., 2013) and LMC (Jones et al., 2017): massive YSOs from embedded Stage I sources to more evolved ultracompact H II regions. The spectroscopic classification of these YSOs relies mostly on *Spitzer*-IRS (Houck et al., 2004)

spectra, and uses a variety of spectral features in the $5-20\,\mu$m range, see Sect. 1.2.

3.3.2 Near-IR catalogues

Near-IR (JHK_s) photometric data for sources in or behind the MC was obtained as part of the MCs survey conducted using the SIRIUS camera on the InfraRed Survey Facility (IRSF) at the South African Astronomical Observatory; full details of the data acquisition and reduction, as well as catalogue construction are reported in Kato et al. (2007). IRSF magnitudes were converted to the WFCAM photometric system.

The transformations applied to convert from the IRSF photometric system to the WFCAM system are:

$$\begin{aligned} K_{\rm WFCAM} &= K_{\rm IRSF} - 0.014 \\ (J-H)_{\rm WFCAM} &= 0.923 \times (J-H)_{\rm IRSF} + 0.036 \\ (H-K)_{\rm WFCAM} &= (H-K)_{\rm IRSF} + 0.055 \times (J-K)_{\rm IRSF} - 0.04 \end{aligned}$$

These were obtained by using the conversions from IRSF to 2MASS and WFCAM to 2MASS available respectively in Kato et al. (2007) and Hodgkin et al. (2009).

3.3.3 Far-IR images and measurements

The MC sources required far-IR measurements similar to those performed in the target galaxies described in Sect .3.1.2 and 3.2.2. To obtain these measurements, 70 μm images from the Multiband Imaging Photometer for Spitzer (MIPS, Rieke et al., 2004) onboard the *Spitzer Space Telescope* (*Spitzer*, Werner et al. 2004) and 160 μm *Herschel*-PACS images of the SMC and LMC (Meixner et al., 2006; Gordon et al., 2011; Meixner et al., 2013) were used.

Distance modulii of $\mu = 23.34$ mag for NGC 6822 (Jones et al., 2019) and of 18.49 and 18.90 mag are adopted for the LMC and the SMC respectively (see Table. 1.1, Pietrzyński et al., 2013; Hilditch et al., 2005). The same 30 pc radius used in NGC 6822

and M 33 results in apertures sizes equivalent to ~ 103 and 124 arcsec respectively at the distances of the SMC and the LMC.

The 160 μm images of the MCs, obtained with *Hercshel*-PACS have residual (non-astrophysical) bias levels that needed to be corrected for before the large aperture brightness measurements can be performed. Such offsets result from the complexity and challenges of processing these datasets obtained very early in the *Herschel* mission (see Meixner et al. 2013 for full details, and more recently Clark et al. 2021). These zero-level corrections were taken from pixel value histograms for each image which are shown in Figs. 3.7 and 3.8. The 160 μm images required offsets of +4.50 and +8.25 MJy sr^{-1} in each pixel for the SMC and LMC respectively.

For consistency, the 70 μm *Spitzer* MIPS images of the SMC and LMC were checked and accordingly very small correction of -0.14 and -0.05 MJy sr^{-1} were applied respectively. The NGC 6822 and M 33 images did not require any such corrections.

3.4 Ancillary data

In order to compare the evolutionary status of star forming regions in NGC 6822 and M 33, archival Hα, 24 μm *Spitzer*-MIPS, and 250/500μm *Herschel*-Spectral and Photometric Imaging Receiver (SPIRE, Griffin et al. 2010) images are used. These wavelengths trace levels of emission from unobscured massive stars (Hα) in the oldest regions to cold dust in the least evolved (500μm) (Sect. 5.7 describes the full methodology).

The Hα images of both M 33 and NGC 6822, retrieved from the NASA/IPAC Extragalactic Database (NED) [6], 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

[6] `https://ned.ipac.caltech.edu`

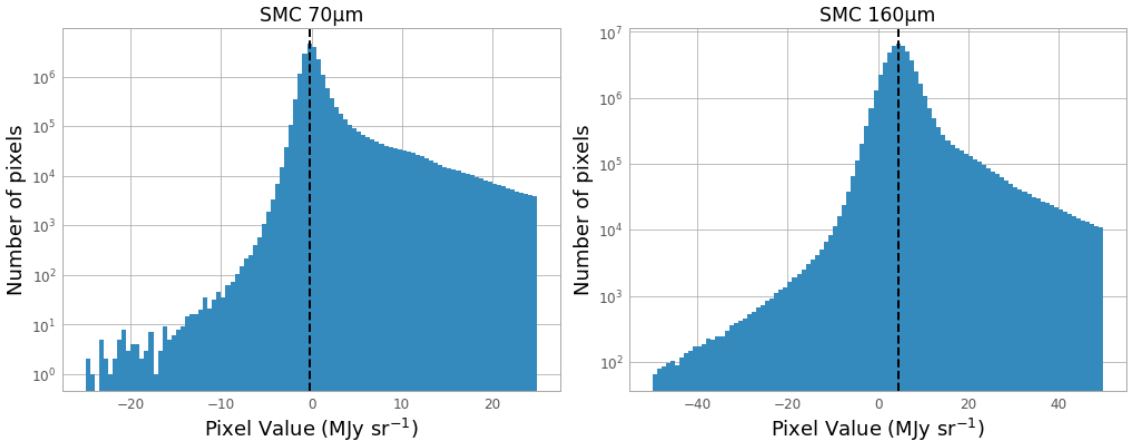

Figure 3.7: Histograms of the pixel values for the far-IR image of the SMC. Vertical dashed lines show the correction value applied. Corrections of -0.14 and $+4.50\,\mathrm{MJy\,sr^{-1}}$ were applied in 70 and 160 μm images respectively.

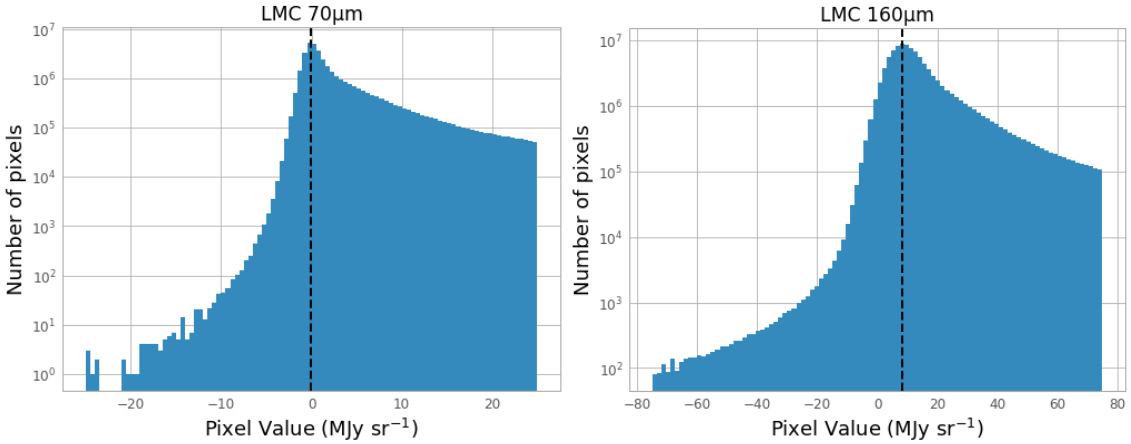

Figure 3.8: Histograms of the pixel values for the far-IR image of the LMC. Vertical dashed lines show the correction value applied. Corrections of -0.05 and $+8.25\,\mathrm{MJy\,sr^{-1}}$ were applied in 70 and 160 μm images respectively.

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.

4 NGC 6822

The work presented in this Chapter has been published in Kinson et al. (2021), with tables available on the VizieR database. Some minor adjustments were necessary to incorporate the paper into this document. These changes do not affect the methods or results presented.

NGC 6822 is a dwarf-irregular galaxy which lies at a distance approximately half-way between the MC and M 33 (see Table. 1.1). YSOs have been identified in several key star forming regions in NGC 6822 (Jones et al., 2019; Hirschauer et al., 2020). Applying the machine learning techniques described in Chap. 3 enables classification of point sources across NGC 6822 in a spatially unbiased way. YSOs identified in this process can be compared with those in the literature, providing a method by which the effectiveness of the selected machine learning techniques can be assessed. These techniques use a selection of measurements to classify each source, which are referred to as features.

4.1 Classification features

Both machine learning methods used in this analysis (supervised and unsupervised, see Chap. 3) were trained on six features: 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 (see Sect. 3.1).

Sources were classified using the PRF on a minimum of two out of four near-IR features. This allows for one missing band from the near-IR JHK_s data, for example a missing H-band value would affect two near-IR features $J-H$ and $H-K_s$. Those sources which lack the sufficient features were removed from the catalogue. Sources which presented clear issues in many features such as un-physical colours (e.g. $J-K_s \ll 0$) or excessive error bars (e.g. $(J-K_s)_{err} > 1$ mag) were also removed. In total

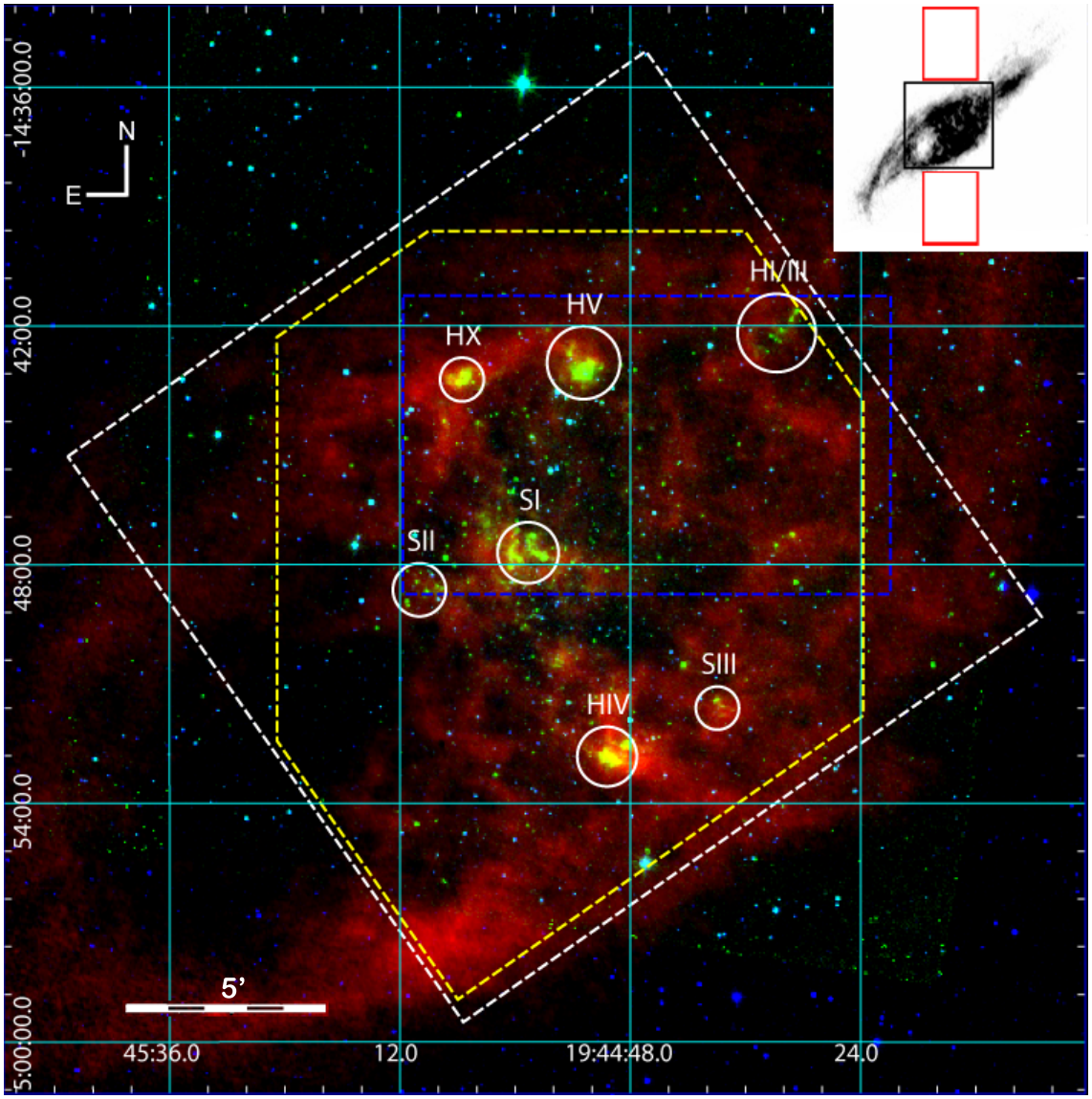

Figure 4.1: An RGB image of NGC 6822 showing H I gas emission (red, Schruba et al., 2017), 8 μm *Spitzer* IRAC (green, Kennicutt et al., 2003) and 2MASS K-band(blue, Skrutskie et al., 2006) images. The area covered by this study is shown by the dashed yellow line. The coverage of the far-IR *Herschel* PACS images is given by the white dashed line. CO (2−1) coverage from Gratier et al. (2010a) is shown by the blue dashed rectangle. Major SFRs are identified. The cavity in NGC 6822's H I distribution can be seen in the lower left of the image. Note the H I coverage extends far beyond the area of the main image, see inset upper right. The off-galaxy fields used for Galactic foreground comparison in Sects. 4.2.3 and 4.4.2 are indicated by the red outlines in the inset H I image.

Table 4.1: Positions, measurements and their uncertainties (where available) as well as source classification for the training set sources. A single row for each training set class is shown here, the full version is available in the online supplementary material of Kinson et al. (2021). Near-IR magnitudes are presented in the WFCAM photometric system.

RA (J2000)	Dec (J2000)	J	J_{err}	H	H_{err}	K_s	$K_{s\,err}$	[70]	[160]	Target
h:m:s	deg:m:s	mag	mag	mag	mag	mag	mag	MJy sr^{-1}	MJy sr^{-1}	Class
05:04:51.69	−66:38:07.4	18.83	0.043	18.02	0.040	17.07	0.034	21578.2	138989.8	YSO
19:44:34.63	−14:55:52.0	17.52	0.036	16.61	0.024	16.41	0.026	1818.6	22382.3	OAGB
19:44:32.41	−14:56:30.8	17.86	0.047	16.85	0.029	16.27	0.023	12720.6	600.6	CAGB
00:37:04.67	−73:22:29.6	18.05	0.050	17.29	0.070	16.47	0.060	405.8	837.7	AGN
19:44:26.62	−14:56:38.2	16.90	0.022	16.55	0.022	16.43	0.028	0	0	FG
19:44:47.45	−14:54:28.9	19.09	0.131	18.21	0.094	18.10	0.107	5316.9	35.2	RGB
19:44:55.70	−14:51:55.9	13.26	0.003	12.58	0.002	12.36	0.002	9917.9	44576.6	RSG
19:45:03.02	−14:54:27.1	18.00	0.053	17.95	0.075	18.21	0.117	401.7	7192.1	MMS

~ 2.5 per cent of the sources in the original near-IR catalogue were removed. This left a catalogue of 11,341 sources remaining.

PRF classification relies upon both measurement and uncertainty information in each feature (see Sect. 2.2.2). This information was included for all intrinsic source features (i.e. all near-IR colours and K_s-band magnitude), however uncertainty values were not included for the far-IR features which provide only environmental information for each source. The far-IR apertures sample the same astrophysical scale (see Sects. 3.1.2, 3.2.2 and 3.3.3) but the number of pixels within the aperture can vary by a factor of 10^3 between the MCs and target galaxies. Due to this variation the subsequent error function considered by the PRF (see Sect. 2.2.2) may be distorted and significantly different for the different galaxies. Furthermore, given that all features are given the same weight, this choice also keeps the appropriate balance between four *intrinsic* features and two *environmental* features.

4.2 Sources in the training set

The training set for the PRF was constructed from various extant catalogues containing sources in eight target classes: YSOs, Oxygen-rich Asymptotic Giant Branch stars (OAGBs), Carbon Asymptotic Giant Branch stars (CAGBs), Red Giant Branch stars (RGBs), Red Super-Giant stars (RSGs), Active-Galactic Nuclei (AGNs), Massive Main Sequence stars (MMSs) and Galactic Foreground stars (FGs). The observed properties of sources in each of these classes are shown in the CMD, colour-colour diagram (CCD) and far-IR brightness plots in Fig. 4.2.

As previously noted in Chap. 3, the performance of the classifier is linked to the numerical size of the data set, how much parameter space each class samples and the labelling accuracy of the training data. To ensure the highest reliability of test sample target labels I included only sources identified in the literature using methods in addition to broad-band photometry, e.g. spectroscopy, narrow-band indices or *Gaia* proper motions. The training set sources in NGC 6822 are matched to the near-IR catalogue of Sibbons et al. (2012) using a 1 arcsec search radius. A summary of the information for each training set class is provided in Table 4.2. For some classes the number of sources is relatively small, however the parameter space occupied by that class is often small and thus the sampling remains good (see Fig. 4.2). It is important to note that given that there is not an 'unknown' class, all sources in the catalogue must be assigned to one of the training set classes. This will inevitably lead to classification contamination, which is discussed in Sect. 4.4.2. The individual classes in the training set are described in detail below.

4.2.1 Asymptotic giant branch stars

It is important to include a well defined AGB training set as these stars can have similar near-IR colours and magnitudes to massive YSOs (see Fig. 4.2). The AGB training samples consist solely of previously classified sources in NGC 6822. Most of the AGB sources originate from Sibbons et al. (2012) identified initially with near-IR

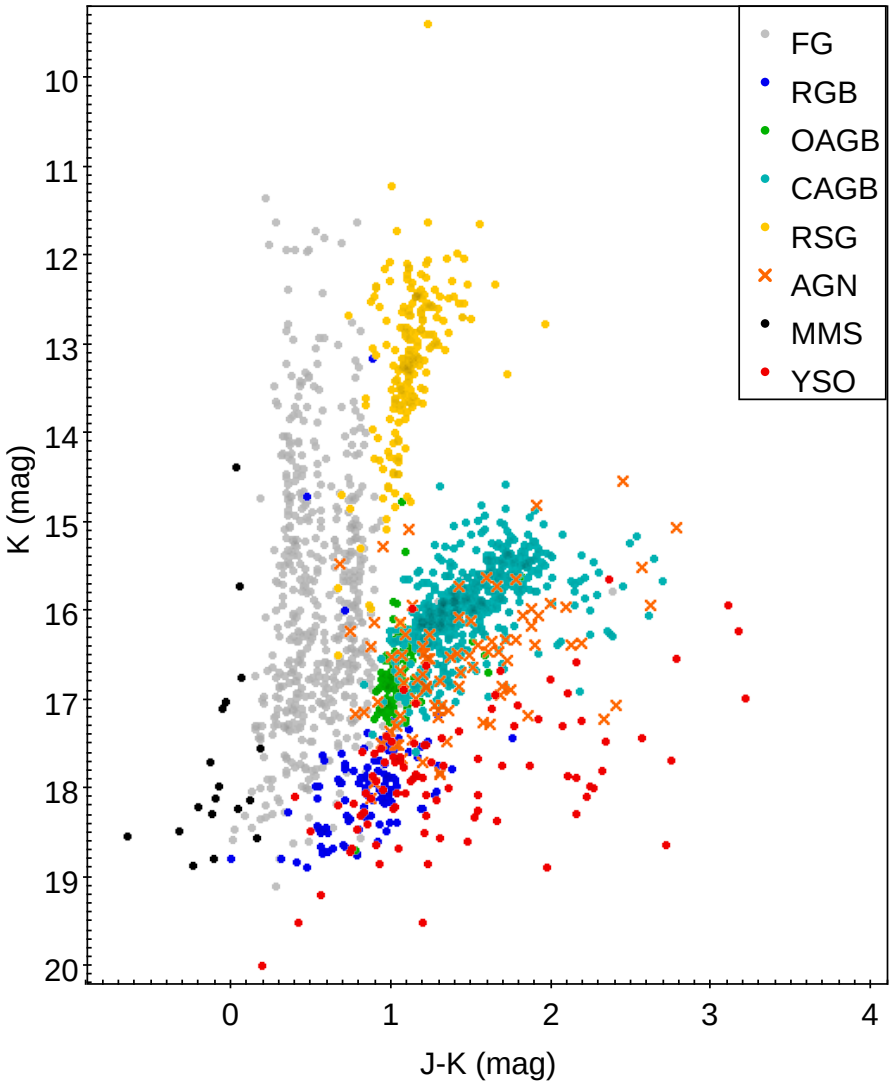

Figure 4.2: CMD plot for the sources in the initial training set.

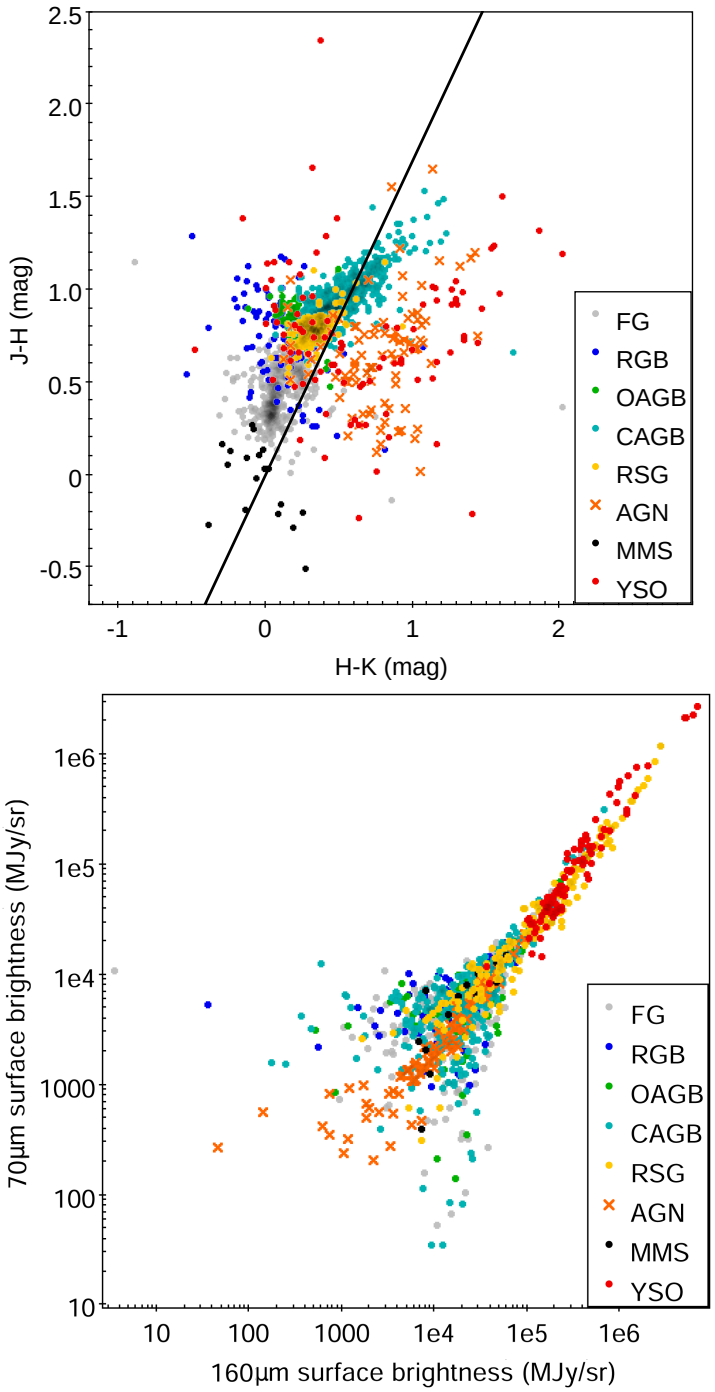

Figure 4.3: CCD (top) and far-IR brightness (bottom) plot for the sources in the initial training set. The reddening line shown in the CCD is calculated from the values given in Rieke & Lebofsky (1985).

Table 4.2: Information on the eight target classes included in the training set. The classification method and reference are given, as well as the number of sources in each class. The AGN sample are identified using a variety of methods. More details of all these classes are provided in Sect. 4.2.

Class	**Number of Sources**	**Identification Method**
YSO	43	*Spitzer*-IRS spectra
OAGB	99	VIS/NIR spectra, narrowband indices
CAGB	461	VIS/NIR spectra, narrowband indices
AGN	89	Various
FG	500	*Gaia* proper motions
RGB	124	*Spitzer*-IRS spectra
RSG	192	Optical & *Spitzer*-IRS spectra
MMS	18	*Gaia* proper motions

photometry and further confirmed with spectroscopy (Sibbons et al., 2015). Additional AGB sources come from the four-band catalogue (R, I, CN and TiO) from Letarte et al. (2002) and the spectroscopic catalogue from Kacharov et al. (2012), which utilises low-resolution VIMOS spectroscopy to confirm AGB nature. I distinguish between O- and C-rich AGBs which present different colours due to their distinct atmospheric molecular composition. These classifications are used to create two AGB target classes in the training set. The training data includes 560 AGBs, split between 461 CAGBs and 99 OAGBs; this difference in class size does not significantly affect the PRF's training since they occupy distinct and reasonably compact regions of parameter space, as shown in Fig. 4.2.

4.2.2 Red giant and supergiant stars

Red giant and supergiant stars are two different populations which contaminate YSO samples at opposite ends in terms of magnitude. Red supergiants are a bright, dusty (similar to AGBs) and young population ($\sim 10 - 30$ Myrs, Britavskiy et al., 2019) which may be located close to sites of recent star formation. Red giant branch (RGB) stars are an older, more dynamically evolved population that tends to be more smoothly

distributed over the body of a galaxy (see for example Cioni et al. 2000 in the SMC and Hirschauer et al. 2020 in NGC 6822) and therefore are less likely to be tightly correlated with sites of far-IR emission. Whilst RGB stars are rarely dusty (Van Loon, 2008) they will likely contribute significantly to the YSO contaminants towards the sensitivity limit which is ~ 2 mag below the tip of the RGB (TRGB, $K_s = 17.36$ mag, Hirschauer et al., 2020).

The training sample RGBs come from three spectroscopic catalogues. A sample of RGBs in Local Group dwarf galaxies are used in Kirby et al. (2013) to constrain the galaxies' metallicities, by determining the Fe/H ratio from spectroscopy; I include the NGC6822 RGBs into the training set. The catalogues of Tolstoy et al. (2001) and Swan et al. (2016) both use spectra containing the Ca II triplet centred at 850 nm to quantify the metallicity ratios in RGB stars. This training class contains 124 sources.

The RSG class for the training set is drawn from catalogues in NGC 6822 and the LMC. In NGC 6822 the spectroscopically confirmed samples of Massey (1998) and Massey et al. (2007a) include 22 sources. Given this small number, it is augmented by including LMC RSG sources, from the catalogues of Jones et al. (2017) which are based on *Spitzer*-IRS spectroscopy, as well as some additional sources from Neugent et al. (2020) identified with spectroscopy focused on Balmer and TiO lines from 340 nm to 1 μm. The training class contains a total of 170 LMC RSGs, giving a total of 192 sources.

4.2.3 Foreground Galactic sources

To define a training class of foreground Galactic contaminants I began by crossmatching the *Gaia* EDR3 catalogue[1] (Gaia Collaboration et al., 2020) with the near-IR data (1 arcsec matching radius). This recovered 5007 near-IR sources with *Gaia* counterparts. Subsequently proper motion (PM) measurements were employed to identify high-reliability Galactic contaminants (see also Sect. 4.2.4 for similar analysis for MMS

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

stars) with sources in the MW displaying higher proper motions than those within NGC 6822.

The distribution of PMs were analysed in both right ascension (RA) and declination (Dec) components separately in order to disentangle sources in NGC 6822 from Galactic foreground objects. Using individual PM components rather than a combined velocity allows for a better identification of Galactic sources, with a high proper motion in either component incompatible with a source in the target galaxies. To achieve this I placed conservative limits on the PM component values as shown in Fig. 4.4; these limits are intended to obtain *clean samples* of FG (and MMS) sources rather than *complete samples*. I also compared the PM distributions of sources in the direction of NGC 6822 with those of two neighbouring off-galaxy areas with the same size, to the North and South (see Fig. 4.1). The two off-field regions extend from 19:44:21 to 19:45:26 in right ascension; in declination the Northern field runs from −14:20:00 to −14:39:30 and the Southern from −14:59:00 to −15:17:50. Using the PM histograms (Fig. 4.4), I set limits for inclusion in the FG training set that the measurement with associated uncertainty must be outside the range of −3 and 3 mas/yr in RA and −5 and 3 mas/yr in Dec. A sample of 500 foreground sources is identified and included in the training set; any remaining foreground sources without reliable *Gaia* PMs will be classified by the machine learning processes. I discuss the foreground training and recovered sets further in Sect. 4.4.2.

4.2.4 Massive main-sequence stars

Massive main-sequence stars in NGC 6822 come from the catalogue of Bianchi et al. (2001). I took the bluest sources ($B - V < 0.4$ mag) as suggested by these authors. I further applied a near-IR cut ($J - K_s < 0.2$ mag), based on the intrinsic colours of O- and B-type stars (Zombeck, 2006) and the average reddening estimates of $E(B - V) = 0.35$ mag (Bianchi et al., 2001). This sample was then matched to the *Gaia* EDR3 catalogue to obtain PMs in a similar process as for the FG class above. Using the PM histograms (Fig. 4.4) I set limits for inclusion into the MMS training set of the

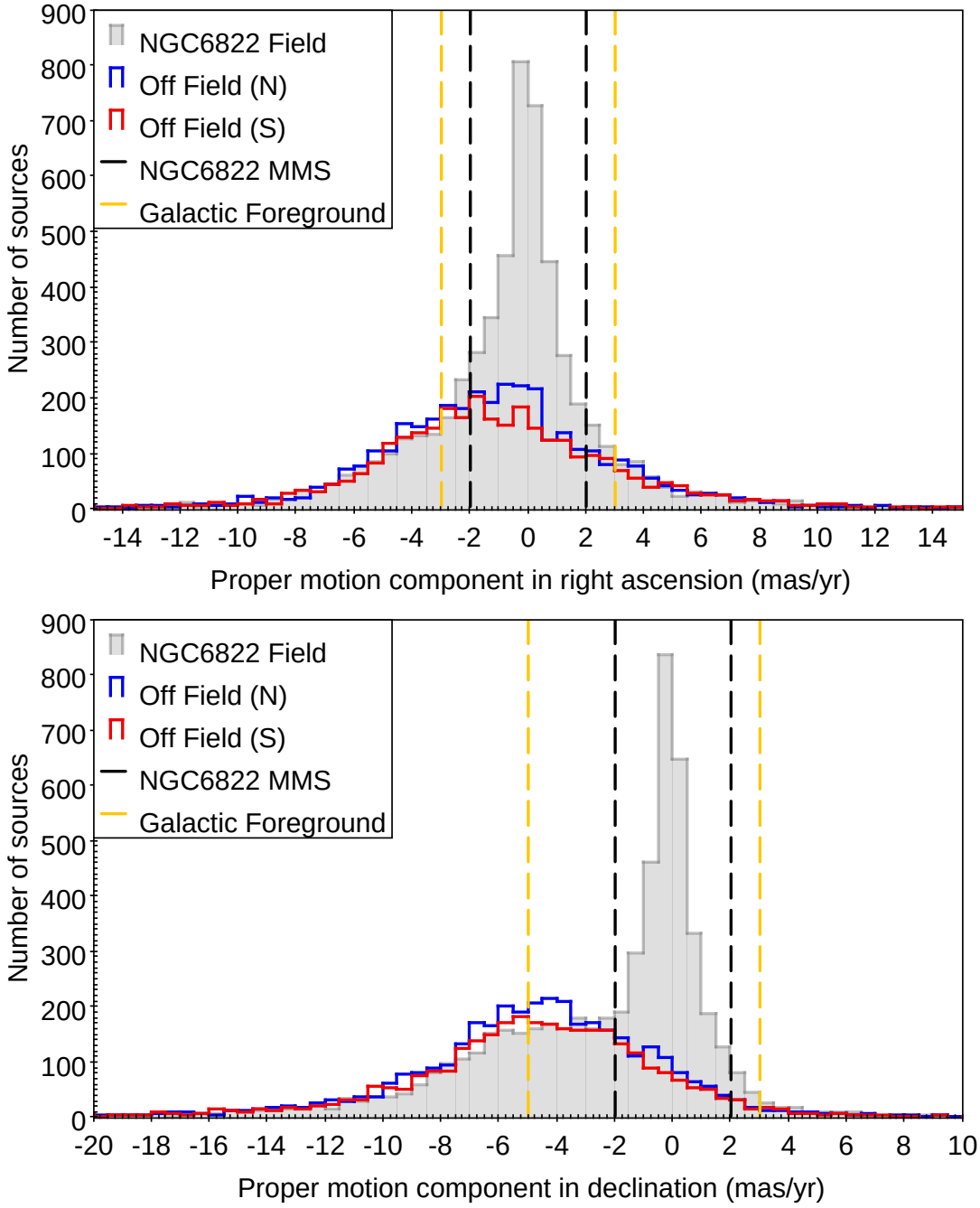

Figure 4.4: Histograms of proper motion components in RA (top) and Dec (bottom) with the limits for training set inclusion for MMS and FG classes shown. Off-galaxy comparison fields to the North (N) and South (S) are shown by the blue and red histograms respectively.

measurement with uncertainty laying between −2 and 2 mas/yr in both RA and Dec. These limits ensure that only sources with PMs consistent with NGC 6822 membership are included in the training set.

The near-IR catalogue samples the brightest main-sequence sources, therefore only nineteen sources for this class were identified. This is the smallest class in the training set; Sect. 4.4.1 shows the effect this has on the training process.

4.2.5 Active galactic nuclei

Active Galactic Nuclei (AGN) are also known contaminants of YSO samples (e.g. Whitney et al., 2008; Sewiło et al., 2013; Jones et al., 2017) and their near-IR colours show considerable overlap (see Figs. 4.2 and 4.3). The large aperture far-IR measurements are thus crucial to differentiate between YSOs which are strongly correlated with nearby far-IR emission and AGN which as background objects have no such preferential correlation on large scales. The recent update of the MILLIQUAS compilation (the Million Quasars Catalog, version 7.2, Flesch, 2021) does not include any spectroscopically confirmed AGNs in my field of analysis. Therefore I choose for the training set AGNs located behind the SMC; this sample is analysed in detail in Pennock et al. (2021). It was compiled from a variety of surveys employing different methods including: Magellanic Quasars Survey (Kozłowski et al., 2011); MACHO Spectroscopy (Geha et al., 2003), *Chandra* observations and OGLE optical to near-IR photometry (Dobrzycki et al., 2003), *XMM-Newton* and *WISE* mid-IR photometry (Maitra et al., 2018) as well as VLT/FORS2 spectra (Ivanov et al., 2016). Near-IR photometry for these sources originates from the IRSF catalogue of Kato et al. (2007) and was converted to the WFCAM system using the transformations given in Sect. 3.3.2. There are a total of 89 sources with sufficient feature data which are taken into the training set.

4.2.6 Young stellar objects

As previously discussed YSO candidates have been identified within SFRs in the central bar of NGC 6822 (Jones et al., 2019; Hirschauer et al., 2020). These analyses were based on *Spitzer* colour cuts and/or SED fitting. To maintain the purity of the training set therefore required additional confirmation of an object's nature and as such these samples are not automatically included into the training set. Furthermore, to validate my approach I aim to independently classify YSOs in these samples. The initial YSO training set was constructed from spectroscopically confirmed YSOs in the SMC (Oliveira et al., 2013) and LMC (Jones et al., 2017): these include sources from embedded Stage I sources through to ultracompact H II regions unresolved in *Spitzer* observations. Spectroscopic classification of these YSOs relies mostly on *Spitzer*-IRS spectra, using spectral features in the $5 - 20\,\mu$m range. For further detail of the spectral features used for each classification see Chap. 3.

After conversion to the WFCAM photometric system (see Sect. 3.3.2), the magnitudes of the Magellanic YSOs were scaled to the distance of NGC 6822. Furthermore these sources were selected such that they are brighter than the detection threshold for the NGC 6822 data of $K_s \sim 19.5$ mag. In total 43 MC YSOs are included in the training set, 39 from the LMC and four from the SMC. These MC YSOs are by design amongst the most massive, but are well-matched to the sample that can be identified with the present near-IR survey.

4.2.7 Exclusion of planetary nebulae from classification

YSO samples can also be contaminated by planetary nebulae (PN). I considered the PN candidates from the analysis of Leisy et al. (2005) which surveyed a large area in NGC 6822, a seventeen-strong sample that the authors state is complete down to 3.5 mag below the brightest PN. However, only one PN candidate has a counterpart in the near-IR catalogue. Therefore the Sibbons et al. (2012) near-IR catalogue seems in fact too shallow to detect all but the very brightest PN in NGC 6822.

Nevertheless, I further considered the samples of Leisy et al. (1997) and Jones et al. (2017) in the LMC. This resulted in 29 PN which would be detectable in the near-IR catalogue when shifted to the distance of NGC 6822. Upon closer inspection it was found that these LMC PN are of rare types (e.g. proto-PN), or their PN nature is questionable or ambiguous. Introducing these sources into the training set would lead to a significant bias in the classifier towards potentially rarer or uncertain types. Furthermore, taking into account the stellar mass of the LMC and NGC 6822, respectively $2.7 \times 10^9\,\mathrm{M}_\odot$ (Besla, 2015a) and $1.5 \times 10^8\,\mathrm{M}_\odot$ (Madden et al., 2014) few such objects would be expected in NGC 6822. This reinforces the conclusion above that very few if any PN are present in the near-IR catalogue, and therefore a PN class is not needed in the training set and in fact including it would adversely affect classification accuracy for other classes. For completeness I note that the single PN with a near-IR counterpart in NGC 6822 is classified as an AGN by the PRF classifier.

4.3 Initial PRF outcomes

With the training set defined for each of the eight target classes I ran the PRF on the remaining catalogue data. The classifier was run twenty times with different random seeds for the train/test splitting to eliminate any stochastic effects in training data selection. This splitting is done on a global rather than class-wise basis, leading to some unevenness in testing data class sizes (see Sect. 4.3.1). It was done on a 75 per cent training, 25 per cent test split, which provides a robust sample to train on in all target classes even those with a low total number of sources in the training set data.

This method is somewhat similar to a k-fold cross-validation approach to training classifiers (for a theoretical introduction see Mosteller & Tukey, 1968). However here all features are included in each PRF run rather than excluding one per fold as a way of estimating feature importance and classifier performance. This was done since the classification uses a relatively small number of features, all equally important as established from testing with t-SNE maps with individual features removed (Sect. 4.7).

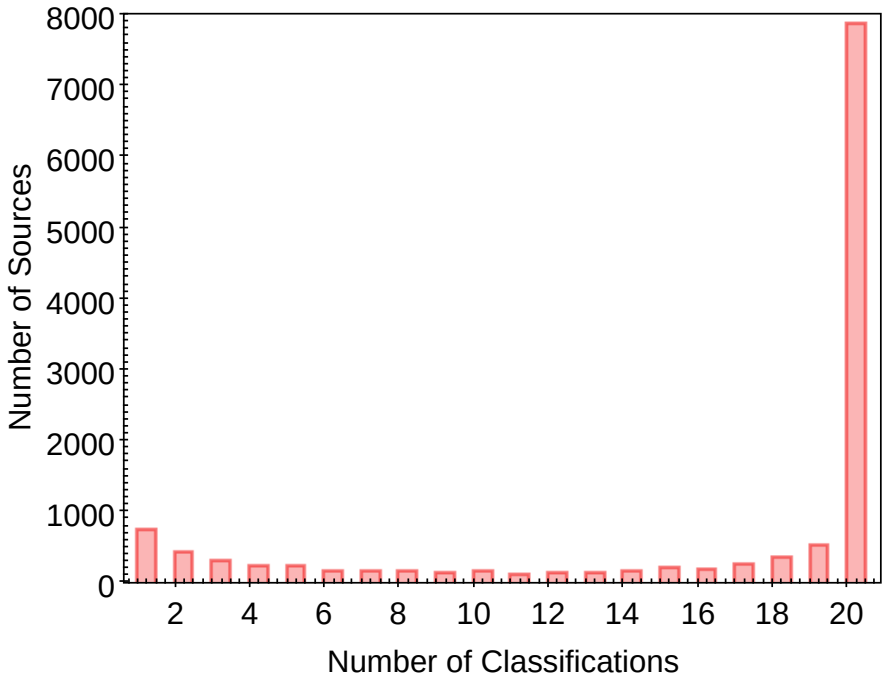

Figure 4.5: A histogram of the PRF classifications across the eight classes and twenty runs. Most sources (~ 79 per cent) are consistently classified in the same class ($n_{\rm class} = 20$).

Applying the PRF in multiple runs risks encountering issues associated with overfitting of the data especially in small target classes. However given the accuracies returned (Sect. 4.3.1) for each class overfitting is not thought to be an issue in this application.

For each PRF run a list of source classifications as well as a set of confusion matrices (Sect. 4.3.1) are generated. Using the ACCURACY_SCORE function in SKLEARN each run returns an estimated accuracy of correct classification across all classes. For the twenty runs of the PRF this varies from 84 to 91 per cent.

For every source a value $n_{\rm class}$ for each class is obtained: the number of runs a source is classified into that class. This $n_{\rm class}$ value allows me to assess the confidence for the object to belong to each particular class. For the training data, most test sources are consistently classified into the same (correct) class. Due to the random nature of the train/test sampling, each source in the training set is effectively classified

a different number of times, and therefore it is not meaningful to assign them global $n_{\rm class}$ values. For the classification of the rest of the catalogue ~ 79 per cent of sources are identified consistently into the same target class over all twenty runs (Fig. 4.5). This is indicative of a robust classification system which is independent of biases induced by random sampling effects and the sources included in the training set. It also shows that the classifier is able to effectively distinguish between target classes.

4.3.1 Confusion matrices

For each run of the classifier two matrices are generated: one with the raw number of sources for each class and one which is normalised by the variable number of sources in each test class (Fig. 4.6). The un-normalised matrices allows me to track any potential imbalances between the number of sources in each training set classes, while the normalised matrices provide an easy to interpret measure of classification accuracy for each target class.

In the normalised confusion matrices a high rate of correct identification is seen for most classes. Issues arise only between sources of similar observed properties such as OAGBs and CAGBs: for instance some OAGBs (presumably the dustiest) are sometimes classed as CAGBs due to the similarities in their colours. Additionally for both AGB classes there is some confusion with RSGs and AGNs due to the fact that fainter RSGs can have similar features to massive AGB stars (Fig. 4.2). Some AGNs have SEDs that peak at mid-IR wavelengths and thus can exhibit IR colours similar to AGB stars (Hony et al., 2011; Van Loon & Sansom, 2015), and spatially are also uncorrelated with large-scale far-IR emission; therefore some classifier confusion between these classes is not unexpected.

The highest degree of misclassification of any class occurs for the MMSs that are classed as FG, a likely consequence of the similarities in near-IR colours. Such misidentifications are not seen in reverse (i.e. FG to MMS) suggesting that this effect is exacerbated by the small number of MMS sources in the test portion of the training set (Fig. 4.2).

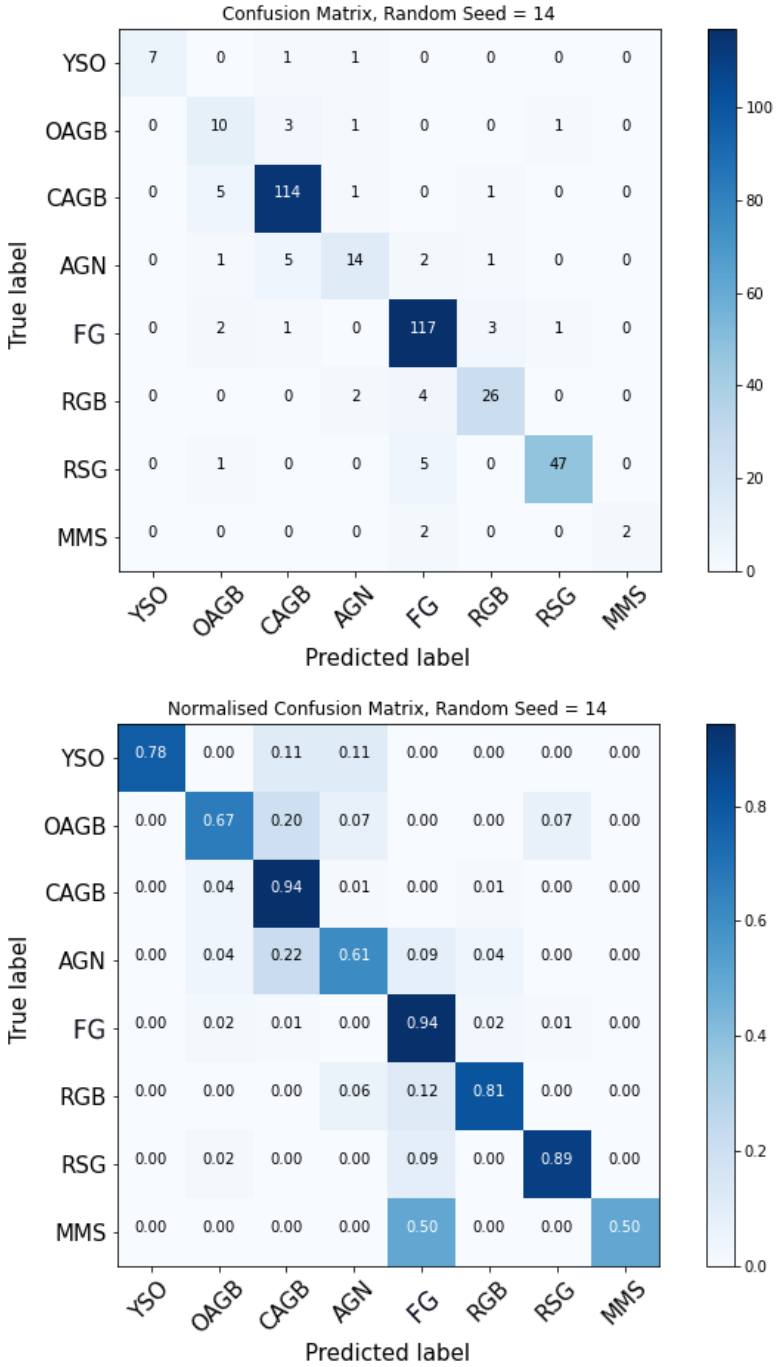

Figure 4.6: A non-normalised (top) and normalised (bottom) confusion matrix for a single run of the PRF classifier using the initial training set. Both matrices were generated from the run with random seed = 14.

FG sources are well recovered, with a small level of confusion into the RGB class. A greater number of RGBs are incorrectly identified as FG sources. This occurs at fainter magnitudes beyond the depth at which *Gaia* counterparts could be found (see Sect. 4.4.2 for further discussion).

The YSO class does not suffer from any contamination from other classes (see first column of the matrices in Fig. 4.6). YSO misclassifications occur into the CAGB and AGN classes (top row of the matrices in Fig. 4.6). Fig. 4.2 shows that these classes overlap significantly in near-IR features with YSOs. The inclusion of the far-IR features in the classification scheme clearly added discriminating power, reducing any confusion to low levels, $\sim$11 per cent for both classes. The matrix values shown in Fig. 4.6 are representative of all seeds, with significant variations occurring only for runs in which the sampling of a particular class is poor.

4.3.2 Extending the YSO training set

The promising results from the initial PRF runs, with very successful classification for MC YSOs (Fig. 4.6), motivated the application of the PRF classifier to the near-IR counterparts of YSO candidates in the catalogues of Jones et al. (2019) and Hirschauer et al. (2020), with the intention of confirming their nature and expanding the YSO training set.

An initial description of the YSO identification in Jones et al. (2019) and Hirschauer et al. (2020) is presented in Sect. 1.4.1. These YSOs are split into three confidence tiers, which properties summarised below. Using a series of CMDs and CCDs individual YSOs are assigned scores based against their position relative to colour and magnitude cuts. All selected YSO sources have high CMD scores, suggesting that their colours are consistent with a YSO nature. The 105 high-confidence YSOs further have low reduced-χ^2 fits to YSO models (Robitaille et al., 2006; Robitaille, 2017). The 88 medium-confidence YSOs have SEDs relatively poorly fit by YSO models. Finally there are 555 lower-confidence YSOs classified in Jones et al. (2019). These sources were excluded from their SED fitting analysis due to insufficient mid-IR data points, a

disjointed SED or indication of a stellar photosphere; some of these sources may still be bona-fide YSO candidates but their nature could not be appropriately constrained. Of these three types, respectively 23, 18 and 195 have counterparts in my near-IR catalogue.

Hirschauer et al. (2020) focuses on identifying a variety of dusty stellar populations in NGC 6822 and therefore does not provide YSO confidence levels in the same way as Jones et al. (2019). Hirschauer et al. (2020) identify 310 YSO candidates, 59 of which are distinct from those classified in Jones et al. (2019). Of these 59 unique sources 41 have a near-IR counterpart.

Whilst the PRF is capable of classifying a source with missing features, as described in Sect. 4.1, the quality of these classifications will be reduced owing to the increased number of nodes in each tree at which an even split rather than a probabilistic decision is made. From the YSO candidates of all confidence levels in the Jones et al. (2019) and Hirschauer et al. (2020) catalogues there were 277 sources out of 807 for which enough features for the PRF to make a meaningful classification were available.

The PRF classifies many of these 277 sources with a high level of certainty: as shown in Fig. 4.7 the $n_{\rm class} \geqslant 19$ bins contain 40 per cent of the sources. Of the 277 sources, 82 were classified as YSOs for some of the PRF runs, 55 of which have $n_{\rm YSO} \geqslant 19$. Of these 55 sources, 47 are from the tables of Jones et al. (2019) with 10, 4 and 33 coming from their high, medium and lower reliability classifications respectively. This represents 43 per cent of the highest confidence YSOs from Jones et al. (2019) used in my analysis. The remaining eight are sources unique to the YSO classifications of Hirschauer et al. (2020). These 55 sources are added to the training set, boosting the number of YSOs from 43 to 98. The PRF classifier was retrained on this extended training set for its application across the full NGC 6822 catalogue. The remaining 222 literature YSO candidates which do not meet the threshold of $n_{\rm YSO} \geqslant 19$ are included in the catalogue for the PRF classification; their final classifications are discussed in Sect. 4.8.1.

The YSOs originating from the MCs are on average redder than those from

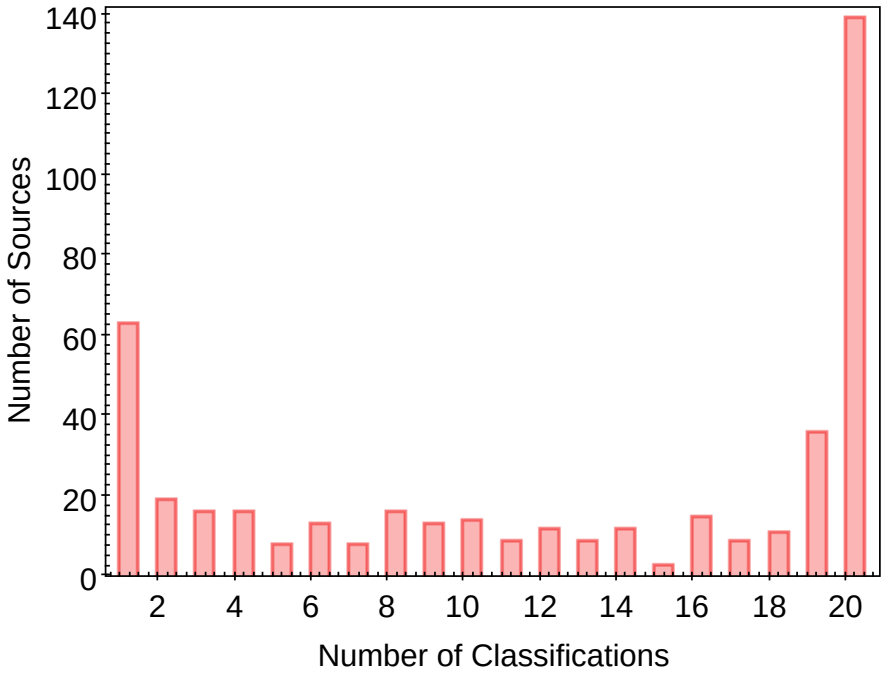

Figure 4.7: A histogram of the n_{class} values across all eight target classes for the literature YSOs from Jones et al. (2019) and Hirschauer et al. (2020) considered for extension of the training set.

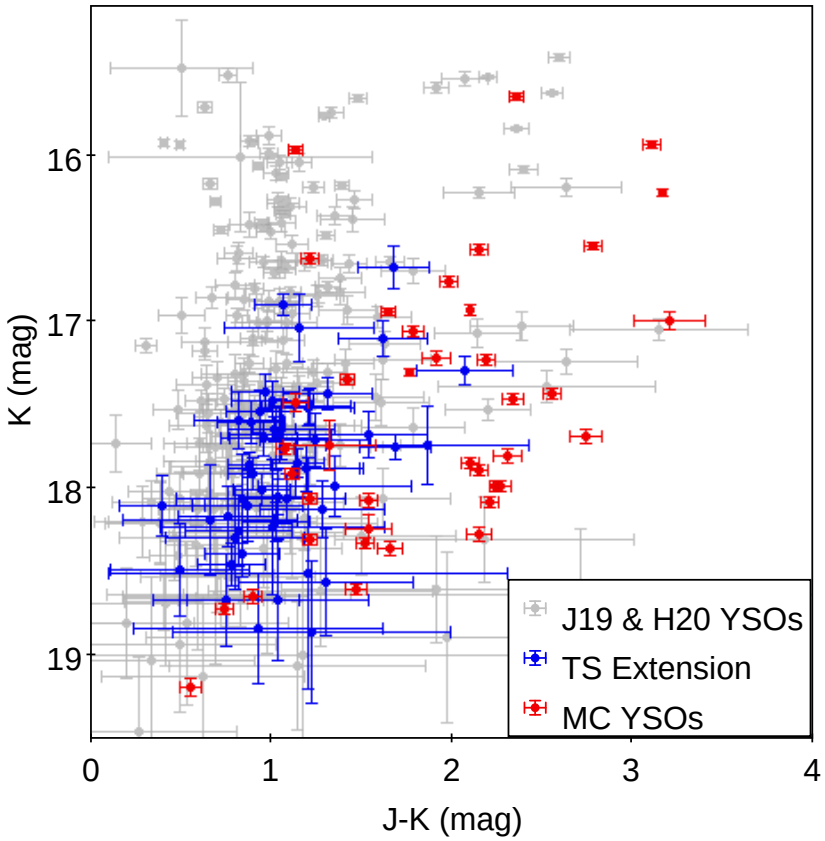

Figure 4.8: CMD of the YSOs considered for the training set extension. The YSOs from Jones et al. (2019, J19) and Hirschauer et al. (2020, H20) are shown in grey. The sources identified for inclusion in the extension of the YSO training set are shown in blue. Training set YSOs from the MCs (red circles) have been scaled to the distance of NGC 6822.

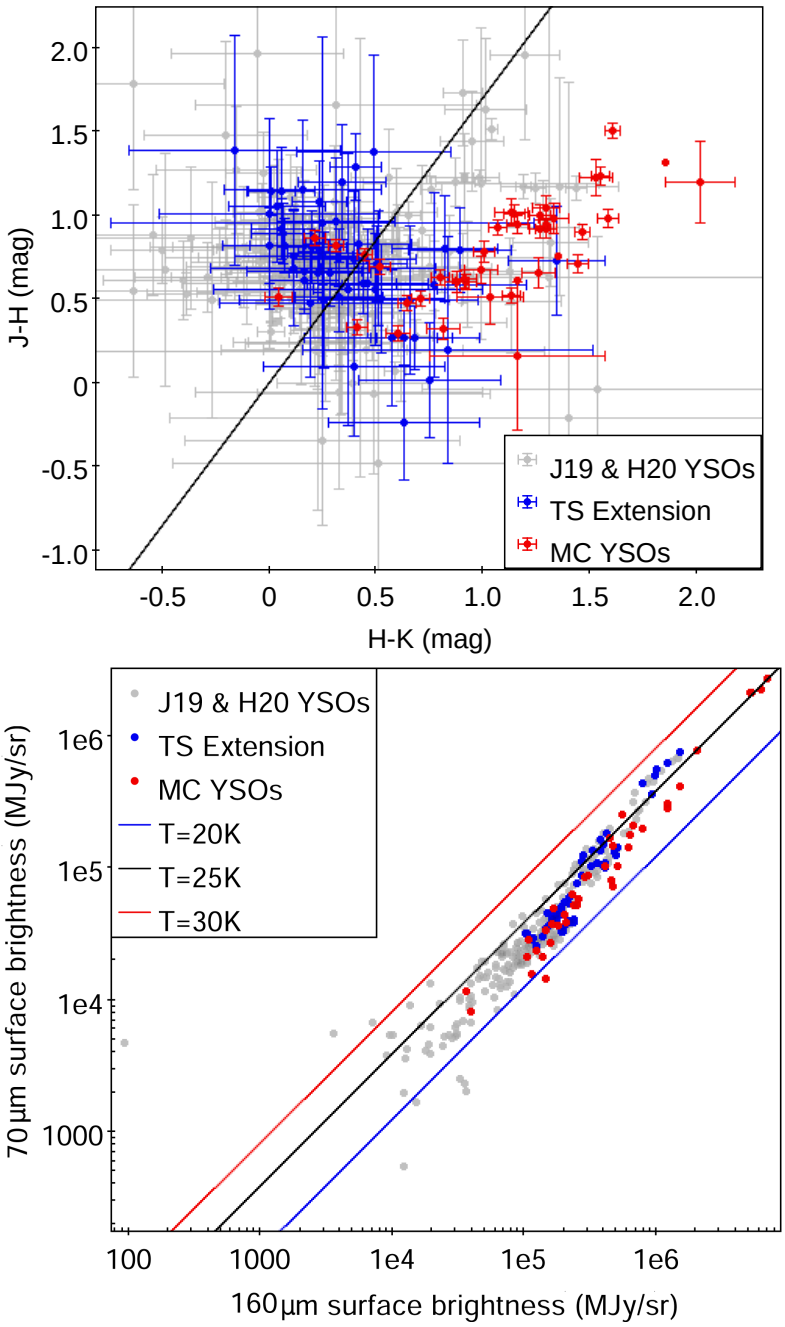

Figure 4.9: CCD (top) and far-IR brightness plot (bottom) of the YSOs considered for the training set extension. Colour-coding as is Fig. 4.8. The reddening line in the CCD is the same as that in Fig. 4.3. In the far-IR brightness plot (bottom) theoretical loci for dusty blackbodies at various temperatures are shown.

NGC 6822 (Figs. 4.8 and 4.9). This is unsurprising given that the MC sample was selected for $5-30\,\mu$m *Spitzer*-IRS spectroscopy, from which their classification is derived. The bottom panel of Fig. 4.9 shows the far-IR brightnesses for the YSOs in both NGC 6822 and the MCs. The plot includes the loci of dusty sources at various temperatures (adopting a dust emissivity coefficient $\beta = 1.5$), generated using 1D blackbody models in Astropy (Price-Whelan et al., 2018). All YSOs generally follow the locus for a dust temperature of $T \sim 25\,$K. There may be a slight hint that the MC YSO far-IR brightnesses could be consistent with a marginally lower dust temperature. This would be expected given the differences in metallicity between the LMC (from which most MC YSOs originate) and NGC 6822 (see Table. 1.1), as metallicity and dust temperature have been shown to be anti-correlated (Van Loon et al., 2010). However, such effect if present seems modest. As already mentioned (see Sect. 3.3) most Galactic RMS YSOs are fainter than the detection threshold of the near-IR catalogue and, due to the observed much higher extinction, are even redder than the MC sample. Such red sources are not present in the NIR catalogue.

This enhanced YSO training set covers a wider region in parameter space for all the used features; furthermore it provides a training set that is now dominated by sources in NGC 6822, mitigating any potential issues relating to differing YSO properties in these galaxies. Given the goal of characterising the YSO population in NGC 6822 and M 33 (see Chap. 5) with a range of sub-solar metallicities (see Table. 1.1) this enhanced training set is more robust to modest variations in dust properties associated with metallicity.

4.4 Enhanced PRF classifier

The PRF classifier with the extended YSO training set was trained and applied twenty times for the classification of the full catalogue. This was done with the same twenty seeds to determine the split in train/test data as used for the original PRF runs, allowing me to assess the improvement in classification directly. The range of accuracy

scores for these new runs is between 87 and 92 per cent. This is a minor improvement overall, however by comparing the normalised confusion matrices it is clear that for some classes (including YSOs) the improvement is more pronounced.

4.4.1 New confusion matrices

In the same manner as the initial runs, confusion matrices were generated for each PRF run. The example normalised confusion matrix in Fig. 4.10 shows that the PRF identifies well all classes except AGN and MMS in the training data. A clear boost in the rate of correctly classified YSOs can be seen by comparing the values in Figs. 4.6 and 4.10: an increase from 78 per cent to 95 per cent. Some misclassification of AGN sources remains. FG sources are confused for sources in NGC 6822 only in a very small number of cases. The asymmetric confusion between FG and RGB classes is still present (further discussion in Sect. 4.4.2).

All confusion matrices both normalised and non-normalised generated in the PRF runs using the extended training set in Figs. A.3 and A.6 for completeness. Across all classes and runs the PRF has a predicted average accuracy of 90 per cent, with class to class variations, exceeding 96 per cent for YSOs.

4.4.2 Galactic foreground estimation

As previously discussed both for the initial and improved runs of the PRF, a confusion between FG and RGB sources is seen in the training/test data. This confusion is investigated in the final classification and is compared to simulated foreground models.

Simulations of Galactic populations using TRILEGAL (Girardi et al., 2005) were used to estimate the predicted number of foreground sources for the same area on the sky as covered by the near-IR catalogue, using the detection limit at $K_s = 19.5$ mag. The modelled foreground suggests that there should be ~ 2978 Galactic sources above this threshold.

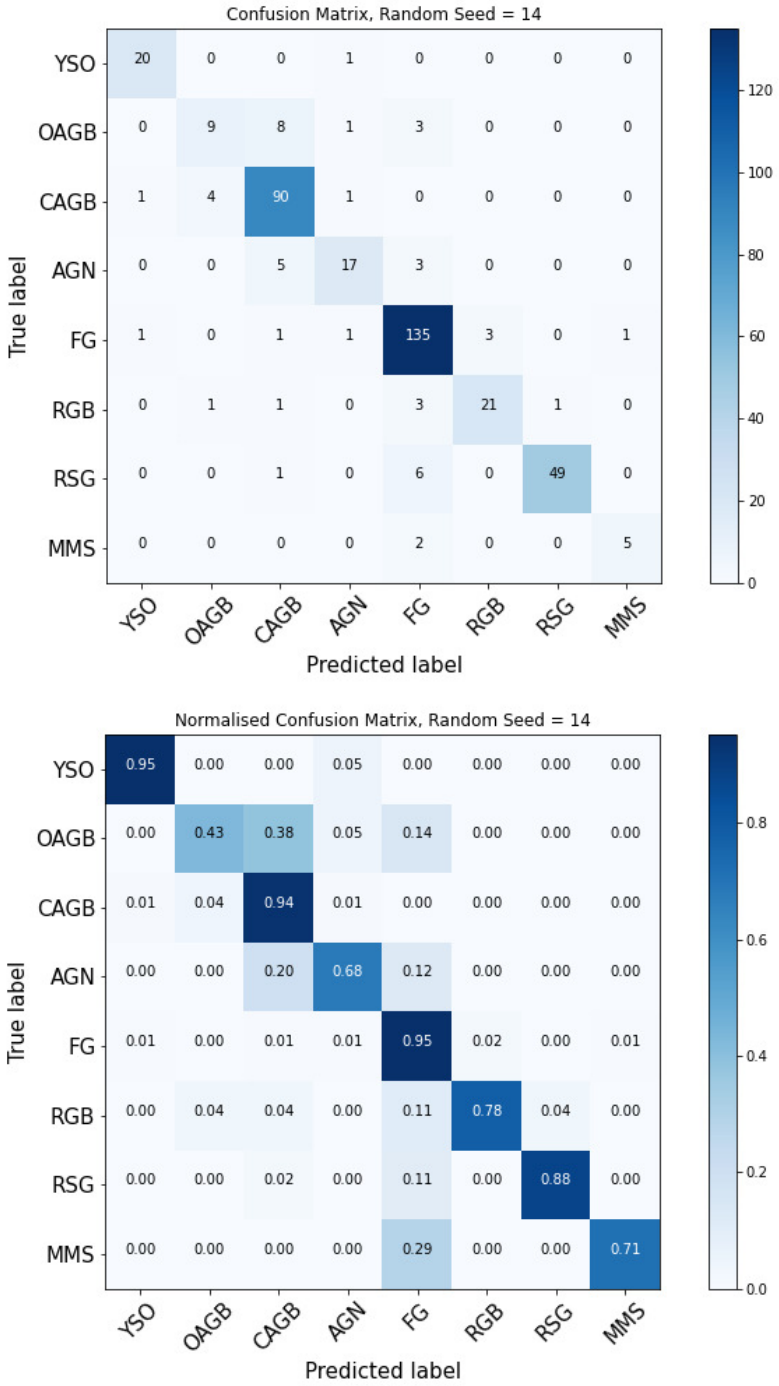

Figure 4.10: A non-normalised (top) and normalised (bottom) confusion matrix for a run of the PRF classifier using the extended training set. The random seed used is the same as that for the matrices in Fig. 4.6.

The PRF runs classify a total of 3082 sources as Galactic foreground in one or more runs, with 2511 classified with $n_{\rm FG} = 20$. Taking only the most certain Galactic sources, those from the training set and those classified with $n_{\rm FG} = 20$, 3011 foreground contaminants are obtained compared to the 2978 predicted by TRILEGAL.

Due to the limiting depth of the *Gaia* data (that corresponds to $K_s \sim 18.6$ mag), the FG training set is restricted to brighter magnitudes. This limitation is reflected in the FG classifications by the PRF which drop off rapidly below $K_s \sim 17.5$ mag (Fig. 4.11). As seen from the confusion matrices, misclassified FG sources are often classed as RGB. I compared the number of output FG and RGB sources to the TRILEGAL model and a Northern off-field region of equal area to the target field (more details in Sect. 4.2.3) in Fig. 4.11. I focus on a range of colours centred on the vertical CMD sequence in the foreground data in which the confusion with RGBs is expected to be more prevalent, $0.6 \leqslant J{-}K_s \leqslant 0.9$ mag (Fig. 4.12). There are 1537 $n_{\rm RGB} = 20$ sources and 1237 $n_{\rm FG} = 20$ sources within this colour range.

The number of $n_{\rm FG} = 20$ sources closely matches what is seen in the off-field data (with a slight excess compared to the TRILEGAL model predictions) down to $K_s \sim 17.5$ mag. At fainter magnitudes, the number of model sources continues to grow, overtaking the detected sources as the completeness limit is reached. Below 17.5 mag, as the number of FG sources drops off sharply, the RGB class indeed begins to dominate below the TRGB at $K_s = 17.36$ mag (Hirschauer et al., 2020).

The comparison above confirms that in this colour range a significant number of faint FG sources are misclassified as RGB sources. Using Fig. 4.11, I estimate that the contamination in this colour and magnitude range of the RGB class by Galactic sources is ~ 54 per cent. Taking 54 per cent of the $n_{\rm RGB} = 20$ sources within the colour range $0.6 \leqslant J{-}K_s \leqslant 0.9$ mag in addition to the most certain FG sources gives a total estimate of 3840 Galactic foreground sources.

In the off-target field in the colour range where FG sources are classified by the PRF ($0.2 \leqslant J{-}K_s \leqslant 0.9$ mag, see Figs. 4.12 and 4.13), there are 3877 sources. This agrees remarkably closely with the estimated Galactic foreground from the classification once the RGB class contamination is accounted for. The estimated foreground, while

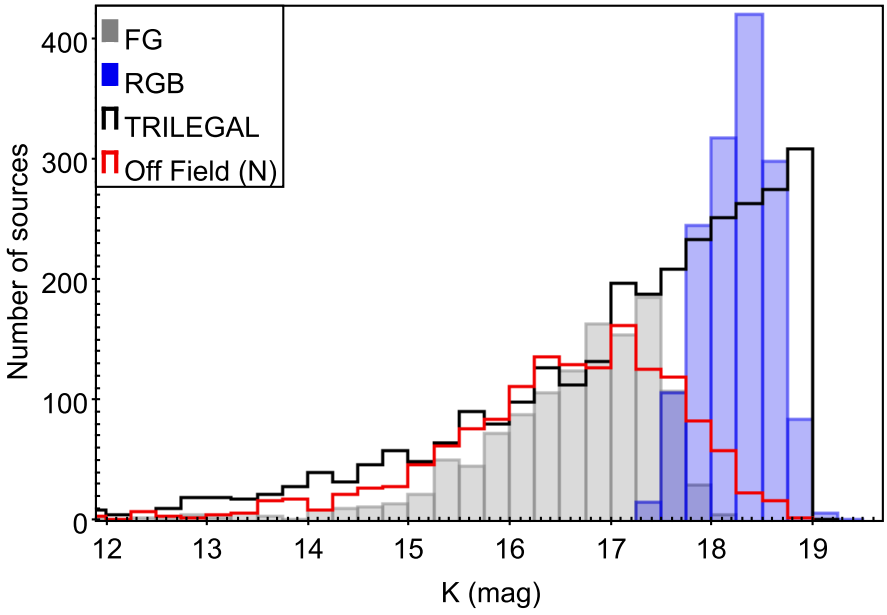

Figure 4.11: A histogram of K_s-band magnitudes for $n_{\rm class} = 20$ FG and RGB sources in the colour interval $0.6 \leqslant J{-}K_s \leqslant 0.9$ mag. Foreground estimates from the Northern off-target field and TRILEGAL are indicated.

higher, is consistent with the TRILEGAL simulation which uses an approximate parameterised model for the Galaxy (Girardi et al., 2005).

4.5 Comparing classifier outputs to the training set

In this section I compare the properties of the PRF classified sources and the training data. Figs. 4.12 and 4.13 shows the CMD, CCD and far-IR brightness plots for the sources which are always classified in the same class in all runs ($n_{\rm class} = 20$); these figures can be directly compared to Figs. 4.2 and 4.3 for the training set.

Even though all classes occupy similar positions in the individual diagrams for the training and output data sets, there are however some differences. The RGB class, whilst fairly sparsely populated in the training set plots (Figs. 4.2 and 4.3), is the most

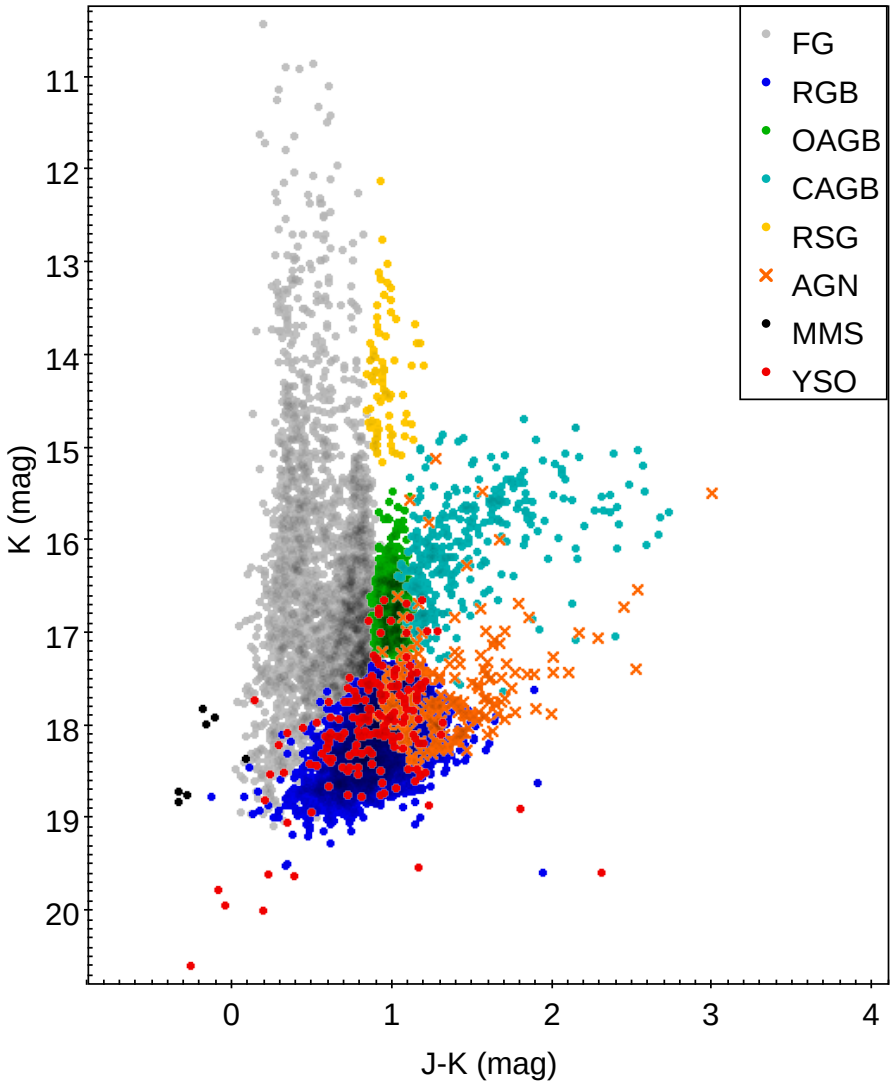

Figure 4.12: CMD plot of the $n_{\mathrm{class}} = 20$ sources from the improved PRF classification.

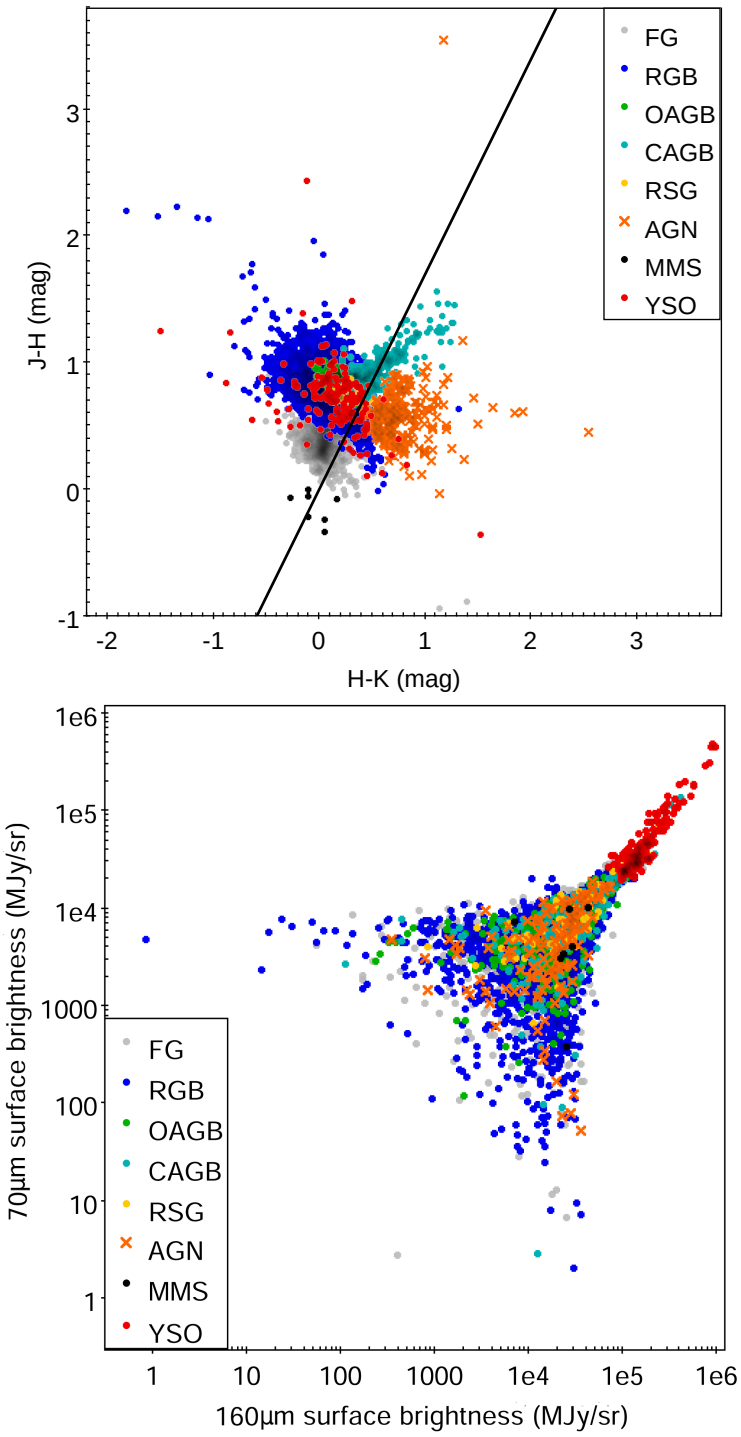

Figure 4.13: CCD (top) and far-IR brightness (bottom) plots of the $n_{\rm class} = 20$ sources from the improved PRF classification.

numerous class in the PRF's output (Figs. 4.12 and 4.13). As discussed in the previous section, all faint FG sources are misclassified into this class.

Furthermore, as discussed in Sect. 4.2 each source must be classified into one of the eight target classes. In the training set, RGB sources occupy a region of near-IR parameter space shared by the bulk of sources in the NGC 6822 catalogue with no strong relation to far-IR emission. Hence a source with no extreme features is likely to be classified by the PRF classifier as an RGB star.

The bottom panel in Fig. 4.13 shows which classes have the strongest association with the far-IR emission. YSOs dominate for very high far-IR values. RSG are also relatively young and as such are still expected to be spatially associated with far-IR emission; indeed the RSGs' brightnesses extend to higher than average values. Confusion between these two classes is however unlikely, since RSGs are bright and well separated in the CCD from the YSOs (see also the matrices in Fig. 4.10).

4.6 Spatial distributions

In Figs. 4.14 and 4.15 I show the spatial distribution of sources for each target class: sources with $n_{\mathrm{class}} = 20$ and those for which the given class is the largest but $n_{\mathrm{class}} < 20$ are colour coded. I highlight a few salient qualitative points below, however a detailed study of the spatial distribution of classes other than YSOs and young RSGs is beyond the scope of this work.

The AGB sources show a decrease in number with increasing galactocentric radius and, while globally correlated with the known galactic structure, appear less constrained to the central bar compared to classes of younger stellar populations (RSG and YSO). This is consistent with the roughly spheroidal distribution of AGB stars described by Letarte et al. (2002) and seen in Hirschauer et al. (2020).

RGB sources are fairly evenly distributed across the field, the source density gradient between central and outer regions apparently consistent with a population older and more dynamically evolved than the AGB classes (as seen also in e.g. SMC,

Cioni et al., 2000). Very few RGB sources are identified inside the major SFRs, likely due to increased crowding that makes such regions less complete to faint RGB stars. It is important to note that contamination between YSO and RGB classes is not seen in the confusion matrices (Fig. 4.10).

The distribution of MMS sources is difficult to comment on due to the low number of sources in this class. As expected FG and AGN sources are evenly distributed across the field. The FG distribution shows a weak hint of the galaxy bar; as shown also from the confusion matrices (see Sect. 4.4.1 and Fig. 4.10) the FG is expected to be contaminated by RSGs and RGBs at the brighter end. The spatial distribution of RSGs is discussed in Sect. 4.6.1 and the distribution of the YSOs in Sect. 4.8.2.

4.6.1 RSG distribution

The RSGs classified by the PRF represent a young ($\sim 10 - 30$ Myrs, Britavskiy et al., 2019) population in NGC 6822. The locations of the most certain RSGs ($n_{\rm RSG} = 20$) to some extent trace past star formation in the galaxy.

The RSGs occupy the bar of the galaxy filling in between the major SFRs indicated by the YSOs (Fig. 4.16). This is in agreement with existing models of the star formation history in NGC 6822 which suggest a bar-centric burst of star formation in the last 200 Myr (De Blok & Walter, 2000). The current SFRs could be evolutionarily linked to this slightly older population, I discuss the relative ages of the SFRs in Sect. 4.9.

An additional spur in the RSG distribution is seen to the South-East of the bar; this feature is present but not discussed in fig. 12 of Hirschauer et al. (2020). This region borders the large cavity in the H I emission (e.g. De Blok & Walter, 2000) and has been linked to both far-UV (Bianchi et al., 2012) and Hα (Massey et al., 2007a) emission suggesting young populations are present. The far-IR emission is modest, suggesting lack of short-wavelength emission reprocessed by dust, and furthermore no YSOs are classified in this region. The detection of RSGs could help tighten estimates of the age of this H I feature, if as discussed in De Blok & Walter (2000) its origin is

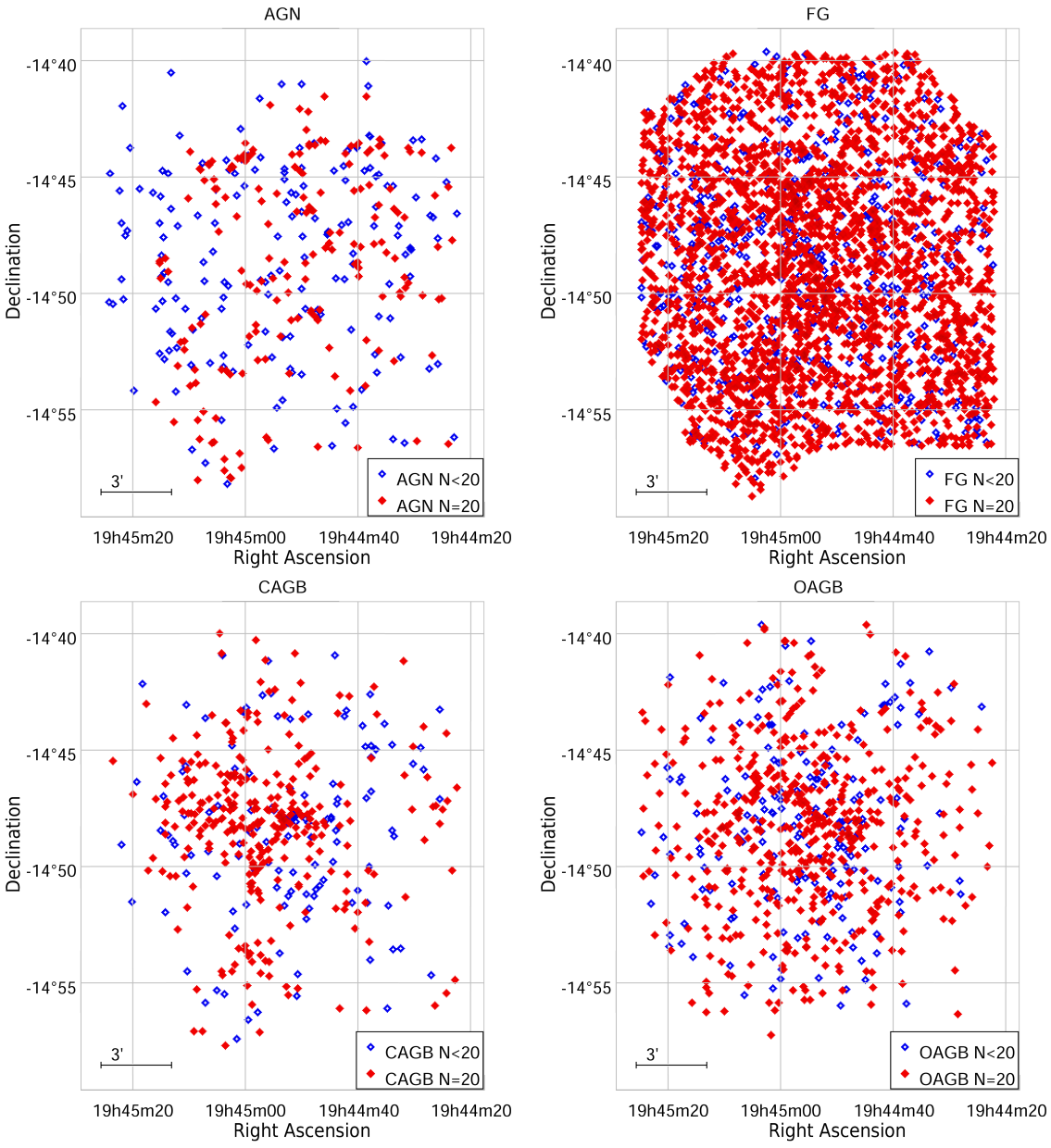

Figure 4.14: Spatial distributions on the sky of AGN, FG, CAGB and, OAGB target classes from the enhanced classification. Sources with $n_{\rm class} = 20$ and $10 < n_{\rm class} < 20$ are shown in filled red and open blue diamonds respectively.

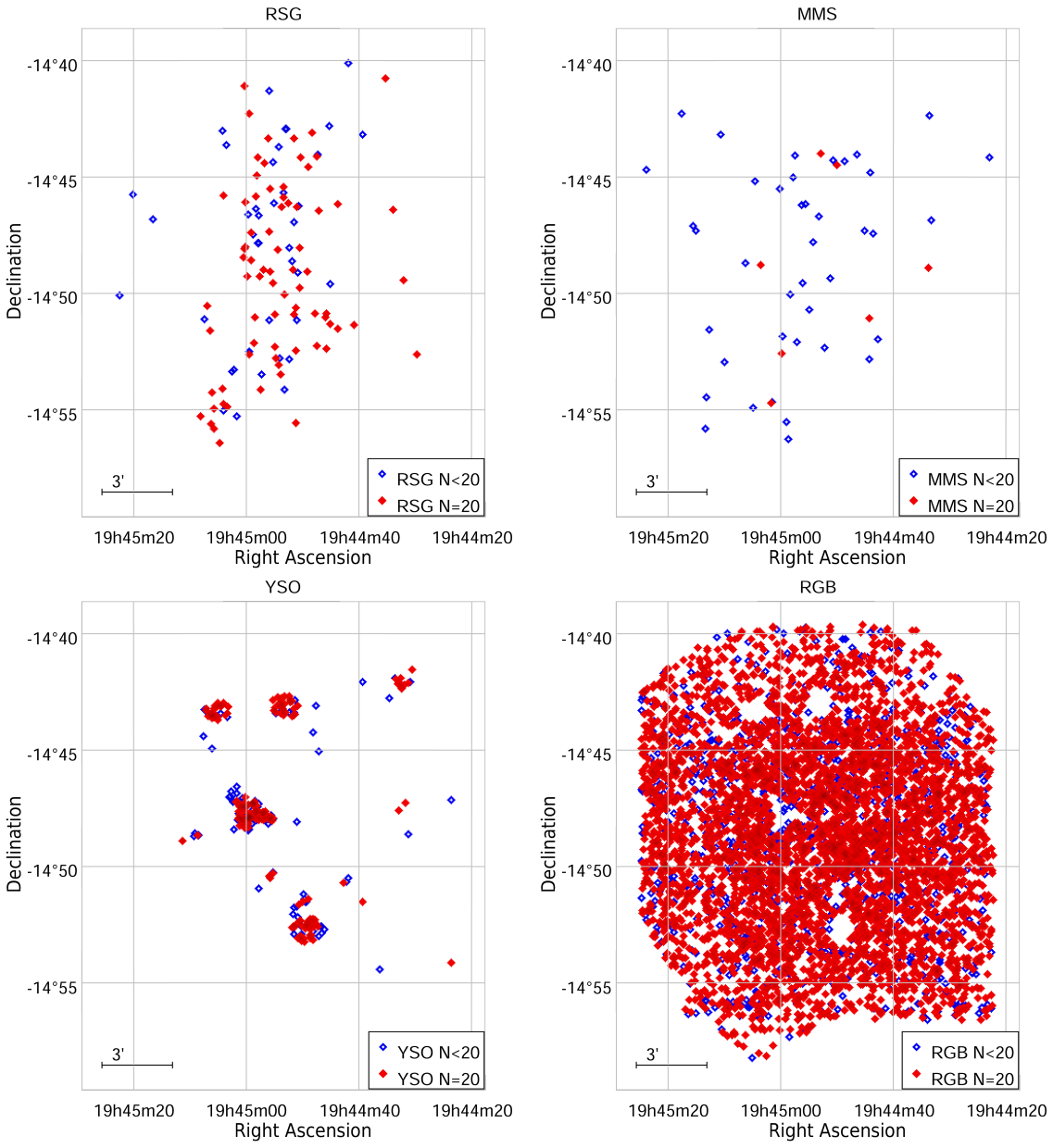

Figure 4.15: Spatial distributions on the sky of RSG, MMS, YSO and, RGB target classes from the enhanced classification. Sources with $n_{\rm class} = 20$ and $10 < n_{\rm class} < 20$ are shown in filled red and open blue diamonds respectively.

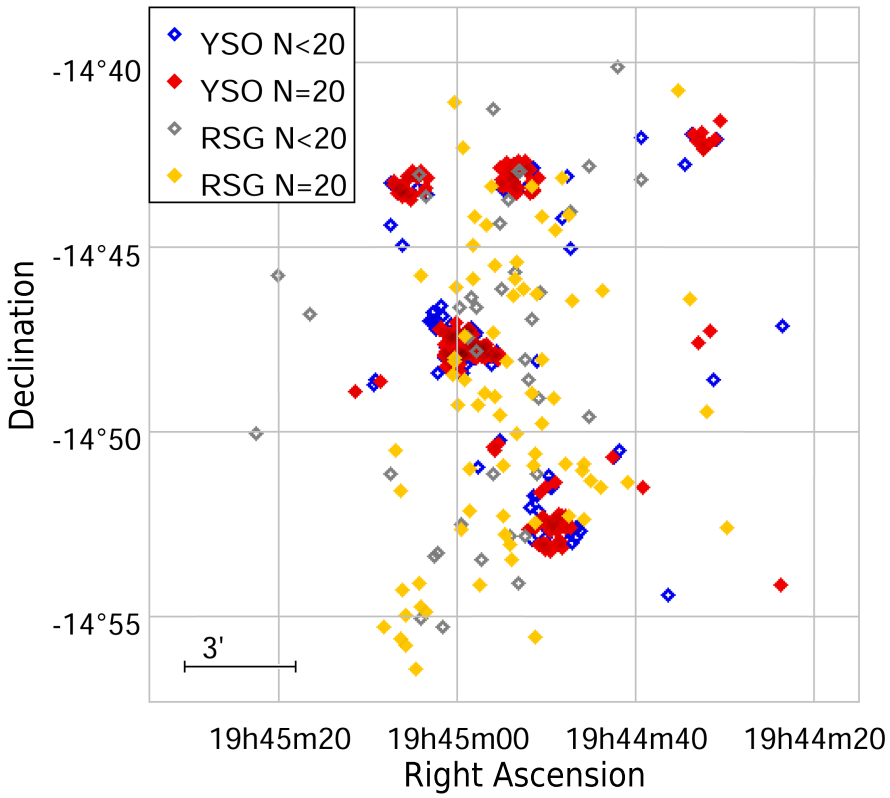

Figure 4.16: The spatial distribution on the sky of YSO and RSG sources. YSOs and RSGs with $n_{class} = 20$ are shown in red and gold respectively; sources with $10 < n_{class} < 20$ are shown in blue and grey, respectively.

linked to a burst of previous star formation.

4.7 t-SNE maps

Using all features in the PRF catalogue, a t-SNE map of the data was created to compare the class separations in a purely data driven, unsupervised machine learning application. Including both far-IR brightnesses, which trace different temperature dust, improves the separation of target classes in parameter space. A worsening in the class separations was observed when any of the near-IR colours or the K_s-band magnitude were removed. In this analysis the perplexity and number of iterations parameters are set to 200 and 500 respectively; these values were selected to maximise visual separation of classes in the output plots whilst maintaining a run-time which is not excessive.

Figure 4.17 shows the t-SNE maps from the training set and catalogue classified by the PRF. The training set sources and PRF classification outputs with $n_{class} = 20$ are shown respectively on the left- and right-hand panels, colour-coded according to the target class. In both maps it can be seen that some classes are tightly grouped (e.g. CAGB) and others are spread over large areas (e.g. RGB). Whilst the area covered by the RGB class is similar in both maps the sparseness of the training set over its parameter space is clear when compared to the similarly sized training class for OAGB sources which are distributed over a smaller map area. The AGB classes occupy a spur to the bottom of the diagram in both the training and output classifications. The CAGB class in particular is very well separated from other classes, suggesting that the use of a t-SNE map with these features could be a useful tool for future studies of evolved stars.

The RSG sources in the training set are tightly packed in a spur off the lobe occupied by the FG sources (see below) and are well separated from AGB sources in the training set. However, there are some newly classified OAGB sources in the same area in the output map. This reflects the inherent difficulties in separating faint RSGs and bright OAGBs in regions of the parameter space where these classes naturally

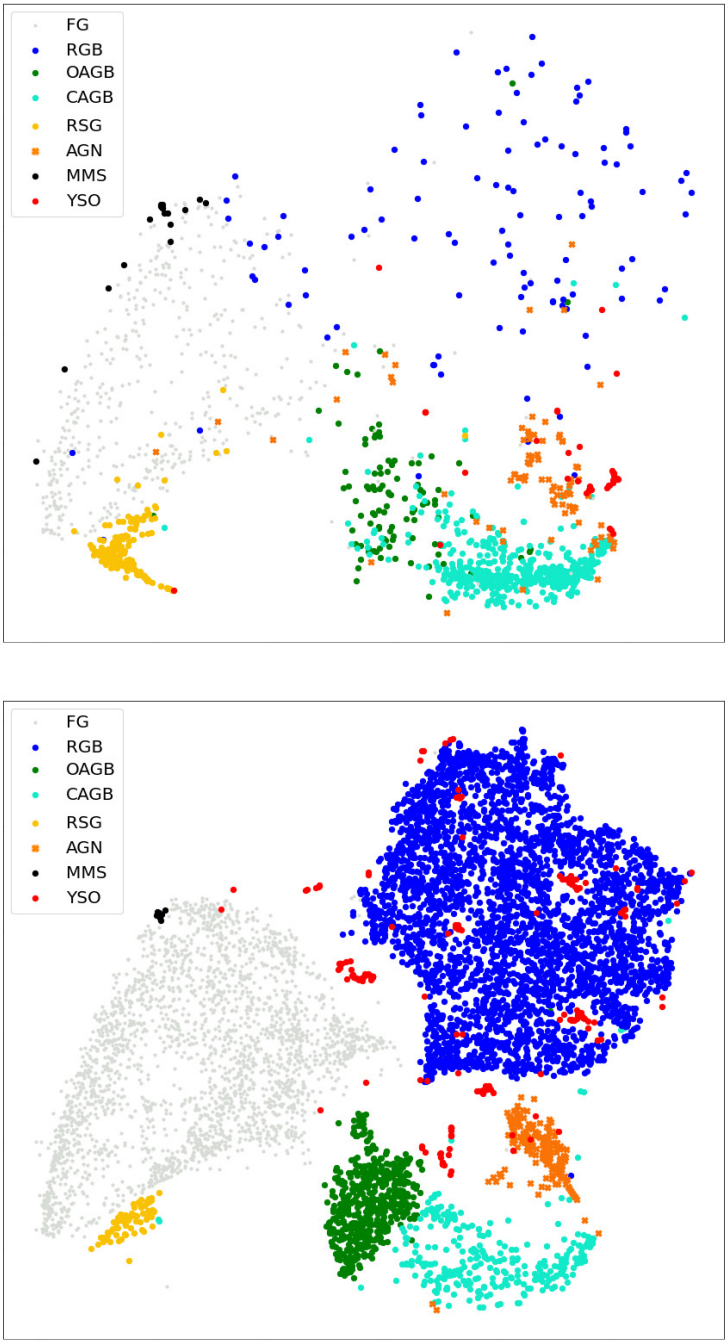

Figure 4.17: t-SNE maps for the training set data (top) and PRF classification outputs with $n_{class} = 20$ (bottom), colour-coded as Figs. 4.2, 4.3, 4.12 and 4.13. Note that the axes for a t-SNE plot are unitless.

overlap (Figs. 4.12 and 4.13); in fact even the spectra of such sources are often similar (see sect. 5.2 of Jones et al., 2017, and references therein).

YSOs are located in several tightly defined clumps. In the training set the vast majority of YSOs occupy an island at the bottom-right of the map, the remaining YSOs are scattered outside of the lobe dominated by FG sources, with some sources seen at the tip of the AGB spur. In the $n_{class} = 20$ map, the main island of YSOs from the training set map is partially recovered, and smaller groupings of YSOs are also co-located with scattered YSOs in the training set map. The YSOs seen at the tip of the AGB spur in the training set map are not present in the $n_{class} = 20$ map. CAGBs are some of the reddest sources in the catalogue, and the YSO training set contains objects that are redder than the vast majority of YSOs recovered by the PRF (see Figs. 4.2, 4.3, 4.12 and, 4.13). The lack of YSO sources in this area in the $n_{class} = 20$ map could be a reflection of this difference.

FG sources in the training set occupy the lobe to the lower left. In this same area of the output map (Fig. 4.17, right) the vast majority of sources are classified as $n_{\rm FG} = 20$ sources. It can also be seen that $n_{\rm FG} = 20$ exist outside of the lobe dominated by FG in the training set map. The area dominated by $n_{\rm RGB} = 20$ sources blends into the region in which $n_{\rm FG} = 20$ sources appear due to the contamination between these classes (Sect. 4.4.2). AGN occupy several locations within the extra-galactic regions of the t-SNE map for the training set, perhaps unsurprising given their wide range of physical properties, however most AGN are concentrated on an island above the tip of the AGB spur. In the t-SNE map for classified sources however the $n_{\rm AGN} = 20$ sources are tightly concentrated on that same location with no scattered points; this is a likely consequence of the PRF classification methodology since here only $n_{AGN} = 20$ sources are plotted and these are more likely to be those which match the bulk of the training set in parameter space.

Based on Fig. 4.17, it is clear that an unsupervised machine learning method like the t-SNE using near-IR and far-IR features can be used to identify some classes of objects like RSG, OAGB and CAGB. Even though YSOs do cluster in such diagrams, such clusters are distributed across the map. Therefore, using the t-SNE maps to

Table 4.3: Catalogue of YSOs and YSO candidates in NGC 6822 classified using the PRF analysis. For sources previously identified as YSOs, the reference is provided in the last column, either Jones et al. (2019, J19) or Hirschauer et al. (2020, H20). Sources included in the training set extension are marked with *. A sample of the table is provided here, the full catalogue is available in the online material of Kinson et al. (2021).

RA (J2000) h:m:s	Dec (J2000) deg:m:s	J mag	J_{err} mag	H mag	H_{err} mag	K_s mag	K_{serr} mag	YSO status	Previous Identification
19:45:04.36	−14:43:04.9	18.41	0.070	17.84	0.052	17.55	0.059	YSO	
19:44:49.22	−14:52:26.7	20.02	0.452	18.78	0.271	19.62	0.660	YSO	H20
19:44:54.21	−14:43:18.2	18.70	0.240	18.14	0.226	17.64	0.180	YSO*	J19/H20
19:45:00.39	−14:47:40.1	17.65	0.075	16.74	0.051	16.54	0.050	YSO candidate	J19

identify YSOs would be more difficult without any additional information and thus inherently less reliable. Due to the nature of t-SNE maps the positions of classes shown here apply within the constraints of this data set and the input parameters used (Sect. 2.1.1).

4.8 The YSO population of NGC 6822

The PRF identifies 368 sources with $n_{\rm YSO} > 0$; of these 269 have $n_{\rm YSO}$ as their largest $n_{\rm class}$ value. These sources are broken down into two categories based on the number of PRF runs in which they are identified as a YSO. There are 182 sources identified as YSOs in all the PRF runs ($n_{\rm YSO} = 20$); these are classified as highest probability YSOs. A further 87 sources have $20 > n_{\rm YSO} > 10$, and are classed as (probable) YSO candidates.

The catalogue of 324 YSOs and YSO candidates (including the 55 NGC 6822 sources from the training set extension) is provided in Table 4.3; 173 sources were identified for the first time using my PRF methodology, 111 YSOs and 62 YSO candidates, as described below. Next, I compare this YSO catalogue to those in previous works and comment on the physical properties of the sources.

4.8.1 The classifications of known YSOs

The YSO candidates from the catalogues of Jones et al. (2019) and Hirschauer et al. (2020) with near-IR counterparts not included in the extended training set were included in the catalogue of sources to be classified. For these 222 sources, their output classifications are examined in greater detail; 126 sources have $n_{class} = 20$ and the remaining 96 sources have $n_{class} < 20$.

For the 126 $n_{class} = 20$ sources, over a third are classified by the PRF as YSOs (Fig. 4.18, upper panel). This is followed by RGB, CAGB and FG classifications that together account for another third of the classifications. The lowest reliability YSOs from Jones et al. (2019) dominate most classes in Fig.4.18 (shown in grey), unsurprising given that such sources vastly outnumber higher confidence YSOs. The highest reliability sources from Jones et al. (2019) with $n_{class} = 20$ are classified mainly into the YSO class, with a smaller number in the RGB, FG and AGN classes. Of the 126 $n_{class} = 20$ sources, 7, 7, and 91 come from the high, medium and low reliability tables of Jones et al. (2019) respectively, the remainder are sources from Hirschauer et al. (2020). Overall ~ 38 per cent of the sources identified as YSOs in either Jones et al. (2019) and Hirschauer et al. (2020) with near-IR counterparts are classified by the PRF into another class with high confidence, i.e. $n_{class} = 20$.

The remaining 96 sources, with $10 < n_{class} < 20$, are classified into more than one class across the 20 PRF runs but have a majority consensus classification. The lower panel of Fig. 4.18 shows the target class with the greatest number of classifications for each source. As above, sources from the low confidence table in Jones et al. (2019) are the most numerous. Most sources are classified as YSOs, followed by CAGBs and FGs. Sources from the highest confidence table of Jones et al. (2019) are categorised into YSO, CAGB, and AGN classes. The YSO candidates from Hirschauer et al. (2020) are classified as RGBs and FGs less often than those from Jones et al. (2019), however the opposite is true for the CAGB class.

From a total of 807 YSO candidates from Jones et al. (2019) and Hirschauer et al. (2020) there is sufficient feature information to classify 277 sources. The YSO nature

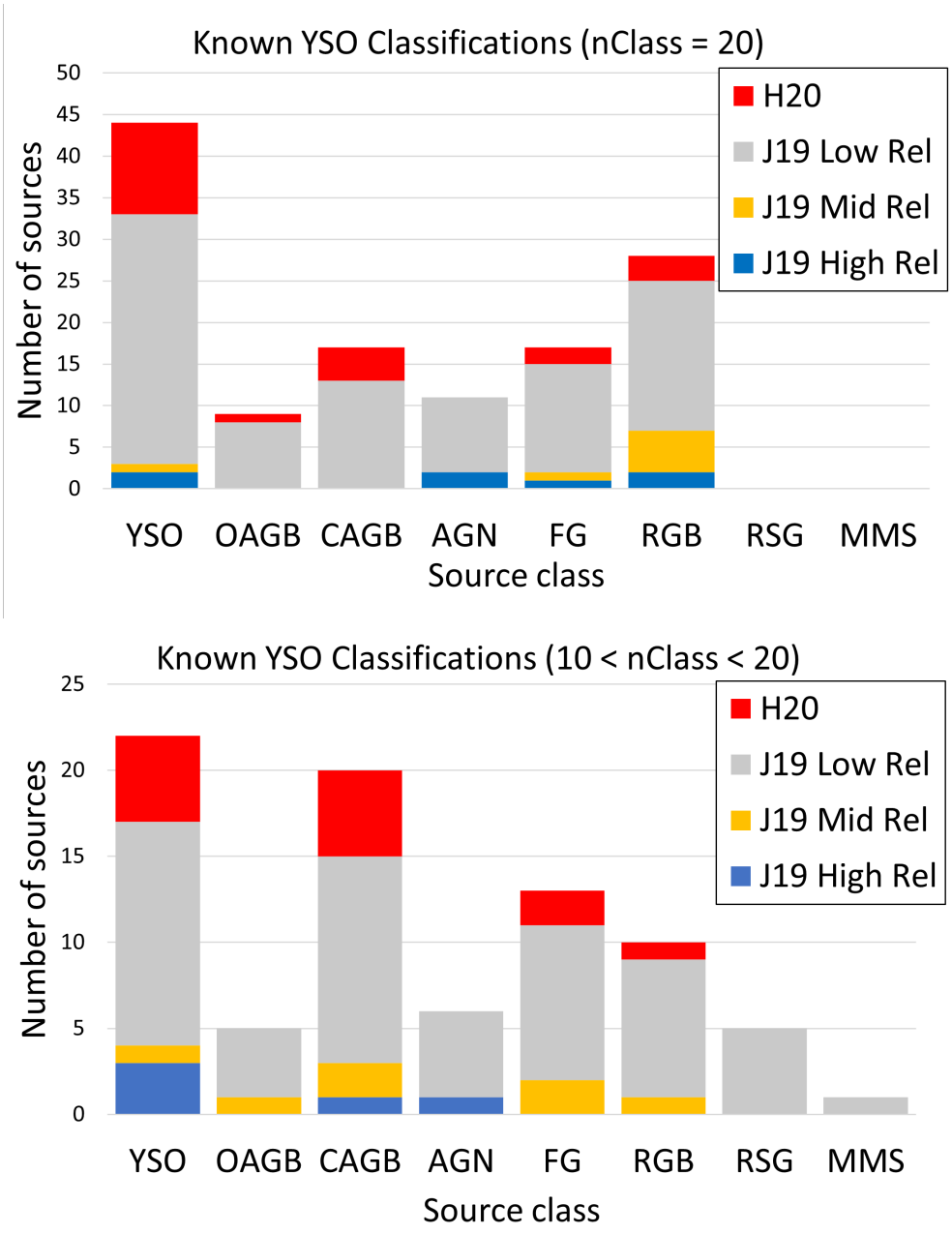

Figure 4.18: PRF classifications for previously known YSO candidates from Jones et al. (2019) and Hirschauer et al. (2020) with $n_{class} = 20$ (top) and $10 < n_{class} < 20$ (bottom, showing the majority consensus classification). The reliability levels from Jones et al. (2019, J19) and YSO candidates unique to Hirschauer et al. (2020, H20) are colour-coded.

for 125 of these is confirmed (77 as YSOs and 48 as candidates) with 55 included in the training set extension and the remainder classified by the subsequent PRF runs: in detail 15/23 high-, 6/18 mid- and 76/195 low-reliability sources from Jones et al. (2019), and 28/41 sources from Hirschauer et al. (2020). Overall the confirmation rate of previously known YSO candidates is ~ 44 per cent, ~ 65 per cent for the higher-reliability sample. Furthermore 82 out of the 277 literature YSO candidates are classified with high degree of confidence ($n_{class} = 20$) in the following PRF classes: seventeen FGs, twenty eight RGBs, eleven AGNs, seventeen CAGBs and nine OAGBs. There are a further two literature sources with no majority classification (no class has $n_{class} > 10$).

Of the 215 unique YSOs identified with the PRF (including training set extension sources), 76 and 31 sources were previously classified as YSOs respectively by Jones et al. (2019) and Hirschauer et al. (2020); of the 87 YSO candidates 42 and 15 were also identified in those papers. This accounts for ~ 40 per cent of the YSOs and candidates classified. Therefore I classify for the first time 199 sources, 136 of which are YSOs and 63 are candidates.

4.8.2 YSO properties

Jones et al. (2019) provide YSO masses and evolutionary stages derived using the SED models of Robitaille et al. (2006) and Robitaille (2017) for their high-confidence sample. All YSOs in common with my sample are best fitted by Stage I models (i.e. still relatively embedded). By comparing the K_s-band magnitude range of the PRF-identified YSOs with that of the YSO candidates from Jones et al. (2019) with mass determinations, I estimate a mass range for the newly identified YSOs between $15 - 50\,\mathrm{M}_{\odot}$. These massive YSOs are more likely the dominant source in an unresolved cluster (Oliveira et al., 2013; Ward et al., 2016; Stephens et al., 2017; Ward et al., 2017). Indeed Jones et al. (2019) note the effect of multiplicity on a comparable YSO model fitting analysis from Chen et al. (2010a) and hence present their mass estimates as overestimated for the dominant source but underestimated for the total unresolved cluster.

The CO (2−1) map of Gratier et al. (2010a) covers the Northern section of the

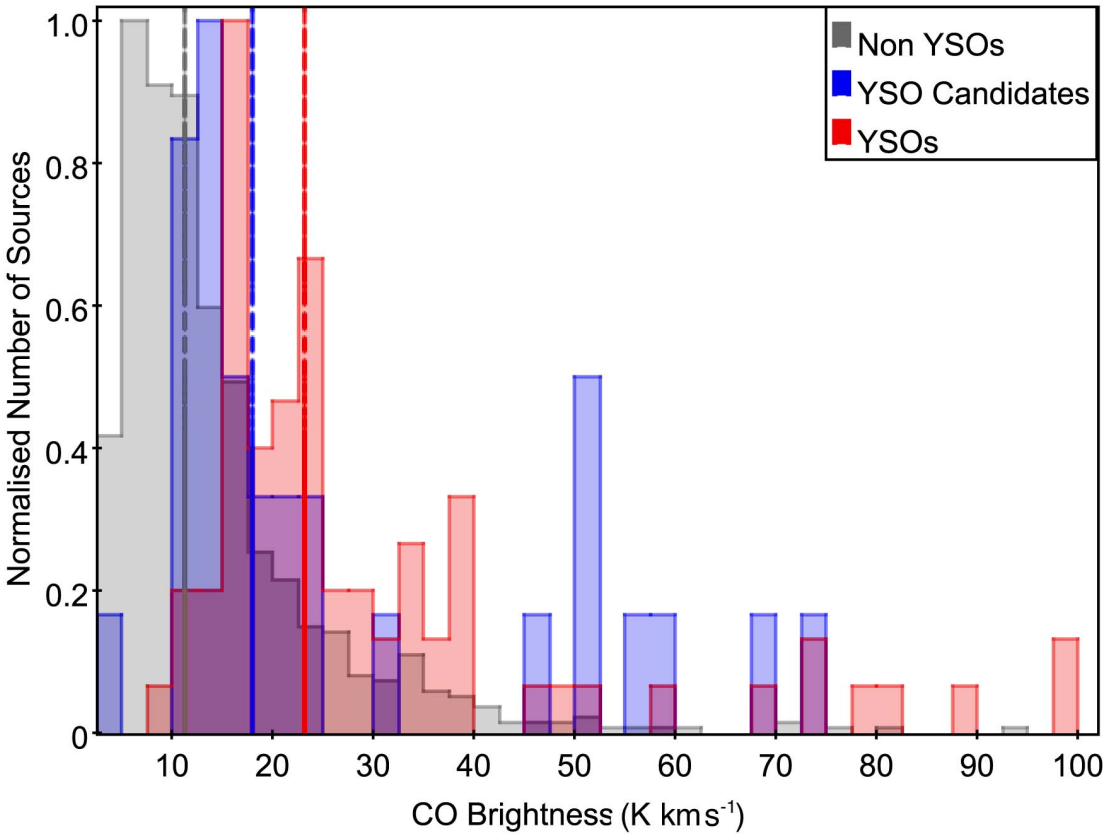

Figure 4.19: A normalised histogram of CO brightness for YSOs, candidates and non-YSO sources. The median value for each group (24.66, 17.58 and 11.25 respectively) is shown by the vertical dashed line of the same colour.

galaxy's bar (Fig. 4.1), with significant gaps in the coverage between the major SFRs. To explore the potential use of CO emission as a feature in the PRF identification of YSOs I performed large aperture photometry in the same way as described in Sect. 3.1.2 for the far-IR data.

CO brightnesses were measured for 1061 sources, 71 of which are YSOs and 30 YSO candidates (Fig. 4.19). YSOs exhibit on average higher CO brightness, with slightly lower average values seen for candidate YSOs. CO brightnesses for non-YSO sources are on average even lower: the median CO brightness values for YSOs and candidates are higher than that for non-YSOs by factors ~ 2 and ~ 1.5 respectively. I conclude that unresolved CO brightness can be a powerful discriminant between YSOs and other stellar populations over large areas. In this project however I do not use CO brightness as a classification feature for M 33 (Chap. 5) due to the spatial similarities to far-IR emission already included as features.

4.9 The star formation environment in NGC 6822

I classified YSOs and YSO candidates in all the major (known) SFRs identified in Fig. 4.1 as well as outside these regions in smaller numbers (Figs. 4.16 and 4.20). The number of YSOs and candidates classified in each of the SFRs is provided in Table 4.4. None of the YSOs or YSO candidates are coincident with the globular or young clusters in the region (see Sect. 1.4.1).

The ratio of Hα emission from less embedded young stars to mid-IR emission from warm dust surrounding embedded sources is greater in older, more evolved regions as radiative feedback from massive young stars clears the interstellar medium. Using a comparison of CO and Hα morphologies as well as the Hα/24 μm emission ratio, Schruba et al. (2017) suggest a more embedded stage of star formation for Hubble IV and V indicative of a younger age, while Hubble I/III and X, that present fewer signatures of embedded star formation (more dispersed morphologies and a higher Hα/24 μm) are more evolved.

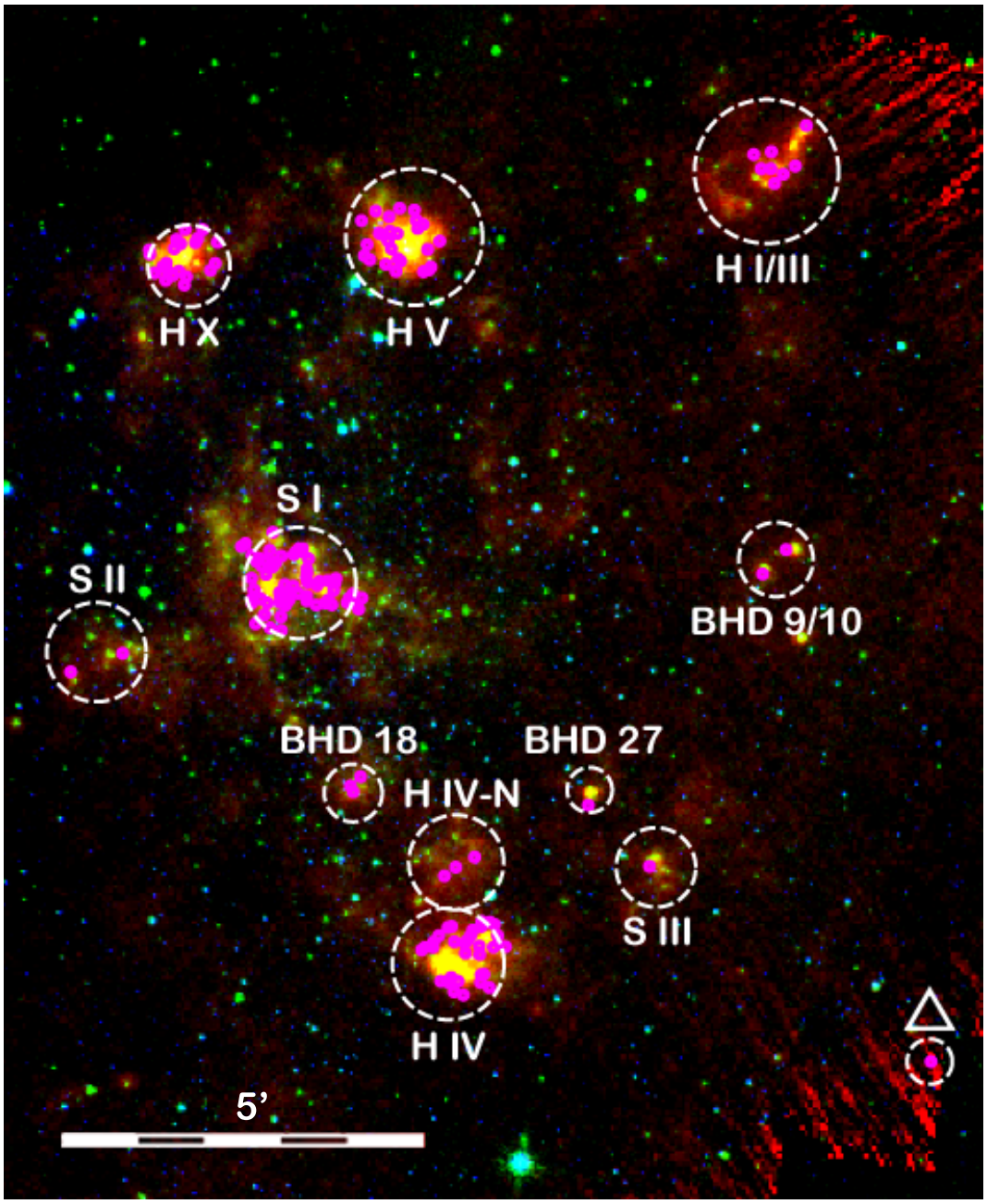

Figure 4.20: RGB image of NGC 6822 (respectively *Herschel* PACS 160 μm, *Spitzer* IRAC 8 μm and WFCAM J-band) with $n_{\rm YSO} = 20$ sources identified (magenta squares). The seven SFRs are shown with the radii given by Jones et al. (2019). The regions BHD 9/10, 18, 27 and Hubble IV−N are newly identified in this work as star formation sites. The region marked with an upright triangle shows the position of the single YSO discussed in the final paragraph of Sect.4.9.

Using a similar methodology as Schruba et al. (2017), Jones et al. (2019) suggest an evolutionary stage for Spitzer II and Spitzer III (newly identified in their work) similar to that of Hubble IV and V, even though the former two look relatively inconspicuous in their fig. 10. They propose that Spitzer I is the youngest region due to faint Hα and UV emission and strong mid-IR emission. Since many embedded YSOs are identified in this region they conclude that the rate of star formation is near its peak. Spitzer I is also the region with the highest number of YSOs in my catalogue. In Fig. 4.20 the green glow of 8 μm emission tracing warm dust is far greater compared to, for example, Hubble I/III which is thought to be the more evolved SFR in NGC 6822.

Since I find a low number of YSOs and candidates in Spitzer II (see Table. 4.4), I discuss this region in a little more detail. In Spitzer II there are 30 literature YSO candidates of which all but four originate from the low reliability sample of Jones et al. (2019); the remainder are one high and one medium reliability sources from Jones et al. (2019) and two YSO candidates from Hirschauer et al. (2020). I find near-IR counterparts for eighteen literature YSO candidates in this region, of which fifteen are classified with $n_{class} = 20$ by the PRF analysis, but only two are classified as YSOs (counterparts to one high- and one low-reliability YSO candidates, Jones et al., 2019). In addition I identity a further three YSO candidates ($10 < n_{YSO} < 20$), two of which are in the low-reliability sample. The remaining $n_{class} = 20$ sources in Spitzer II are classified as five RGBs, four OAGBs, two CAGBs and two FGs. Therefore, given that most literature YSO candidates were considered to be of low-reliability, it is not surprising that I only find two YSOs and three candidates in Spitzer II.

Using *Herschel* far-IR and *Spitzer* mid-IR data Galametz et al. (2010) found that region-integrated SEDs between $10 - 100\,\mu$m show signatures of evolution for some of the H II regions in NGC 6822. They propose that the $250\,\mu$m/$500\,\mu$m emission ratio in particular correlates with the $24\,\mu$m emission and thus traces star formation activity. According to that ratio (see their figure 3b), Hubble V would be the most active region followed by Hubble IV, Spitzer I and Hubble X. These are the regions with the largest number of YSOs both in my analysis and that of Jones et al. (2019). Given that Hubble X and V have strong Hα emission, they would have evolved past the peak of

star formation activity. Spitzer I and Hubble IV on the other hand would be at their peak; for Spitzer I this is supported by the largest number of YSOs, however this is less clear for Hubble IV.

Looking at the positions within the SFRs of the PRF-classified YSOs, in Hubble IV and V the literature YSOs are more centrally concentrated. This is likely due to limitations in recovering point sources within the centre of these bright SFR in the near-IR images (see Sect. 3.1). In the other regions there is no significant difference in the locations of the YSOs within SFRs. In Hubble I/III the PRF classifies YSOs primarily in Hubble I and at the interface of the regions, in agreement with Jones et al. (2019) who suggest that Hubble I is more actively forming stars.

One of the two unnamed clusterings of YSOs noted (but not discussed in detail) in Jones et al. (2019) and recovered in my classification is coincident with a pair of H II regions; the first was designated Kα (Kinman et al., 1979) and the second HK1 (Hodge et al., 1988). These regions are listed in the catalogue of Brunthaler et al. (2006) as BHD 9 and 10; they are located around 19:44:32, −14:47:26, almost directly South of Hubble I/III and to the West of the bar, away from the bulk of star formation activity (Fig. 4.20). The other unnamed region is found to the North-East of Spitzer III (at 19:44:42, −14:50:39) within 10 arcsec of the H II region BDH 27 (Brunthaler et al., 2006), in which only a single YSO and two candidates are identified. In a region in the bar approximately equidistant from Spitzer I and Hubble IV (at 19:44:55, −14:50:24) I find three YSOs and one candidate. This area is bright at 8 μm and in H I emission. The candidate YSO is coincident with a very bright Hα point source or small bubble (BHD 18, Brunthaler et al., 2006). These three regions are clearly seen in the 250 μm/500 μm ratio map of Galametz et al. (2010), suggesting star formation activity is present. I label these newly identified regions of star formation using the Brunthaler et al. (2006) denominations (BHD 9/10, 18, and 27) in Fig. 4.20 and list the number of YSOs in each region in Table 4.4.

I classify three YSOs and six candidates directly North of Hubble IV but outside the SFR radius defined by Jones et al. (2019). The YSOs trace a line at the centre of this region which I name as Hubble IV-N (located at 19:44:50.00 −14:51:31.0). Hubble IV-N

is bright in both 8 and 160 μm emission (see Fig. 4.20) but comparatively faint at 24 and 70 μm. This region has no visible large-scale Hα emission and it is also relatively bright in the 250 μm/500 μm ratio map of Galametz et al. (2010). The location of these newly identified sources along with that of those found in BHD 18 are very suggestive of additional star-formation activity in the bar of NGC 6822 between the major regions Hubble IV and Spitzer I. Given that this analysis deals only with the most massive YSOs ($15-50\,\mathrm{M}_{\odot}$, see Sect. 4.8.2) there is potential for more, lower-mass, YSOs to be found in this bar region.

There is a YSO to the South-West of my field with $n_{YSO}=20$. At this location, there is no UV emission in the images of Hunter et al. (2010), but the Hα image from Massey et al. (2007a) shows a point source. Mid-infrared emission is also unremarkable with a point-source source visible at 3.6 μm but not at 8 μm. Far-IR emission is not prominent but this location is close to the edge of these images. This source is identified with $\triangle$ in Fig. 4.20 (located at 19:44:23.64, −14:54:07.9). This source could represent an isolated YSO, perhaps at the lower mass limit of the current detection range.

In addition, there are a further thirteen isolated YSO candidates located outside the SFRs in Fig. 4.20. These candidates are less certain ($n_{\mathrm{YSO}}<15$), and thus I do not discuss them further.

4.10 NGC 6822 Summary

With a combination of near-IR and far-IR features I have used machine learning algorithms based on a probabilistic random forest classifier (PRF) and t-distributed stochastic neighbour embedding (t-SNE) to classify stellar populations in the main bar of NGC 6822, covering all prominent SFRs.

The PRF was trained using three near-IR colours ($J-H$, $H-K_s$ and $J-K_s$), K_s-band magnitude and two far-IR brightnesses (at 70 and 160 μm) and classifies sources into eight target classes (YSO, OAGB, CAGB, RGB, RSG, MMS, FG and AGN) with an estimated accuracy of 91 per cent across all classes rising to 96 per cent for YSOs

Table 4.4: The number of YSOs ($n_{YSO} = 20$), candidate YSOs ($10 < n_{YSO} < 20$), and training set extension YSO sources (see Sect. 4.3.2) classified in each of the previously known SFRs in NGC 6822, as well as in newly identified YSO groupings (see discussion in the text).

SFR	**YSO number**	**YSO candidate number**	**Training Set Extension YSOs**
Hubble I/III	9	4	1
Hubble IV	35	14	6
Hubble V	36	11	13
Hubble X	29	5	5
Spitzer I	90	49	26
Spitzer II	2	3	0
Spitzer III	2	0	1
BHD 9/10	4	1	2
BHD 18	3	0	1
BHD 27	1	2	0
Hubble IV-N	3	6	0

(based on the PRF confusion matrices of the test sample).

I used the same near- and far-IR features to construct (unsupervised) t-SNE maps to identify stellar populations. Such maps are very effective in picking AGB stars (with a clear differentiation between OAGBs and CAGBs), AGN and RSG stars. Without additional information, the t-SNE maps seem however less powerful in identifying other classes of sources, including YSOs.

The spatial distributions of most stellar populations are essentially as expected. RSG stars, that trace the recent star formation history, occupy the bar of NGC 6822, linking the more conspicuous SFRs. An extension of the bar to the South-East, into a region which has indicators of youth (e.g. De Blok & Walter, 2000) is seen in the RSG distribution, however no YSOs or candidates are classified there.

I classify a total of 324 YSOs and candidates. I confirm the nature of 125 out of 277 literature YSO candidates with enough feature information. Additionally 136 YSOs and 63 YSO candidates are classified for the first time in my analysis. There was no requirement imposed that YSOs and candidates need to be located in main SFRs

(as was done in previous works), and have detected YSOs in the bar of NGC 6822 between the major SFRs. YSOs classified by the PRF have mass estimates between $\sim 15 - 50\,M_{\odot}$, representing the most massive YSO population in NGC 6822. Another 82 out of 277 literature YSO candidates are definitively classified as non-YSOs by the PRF analysis.

I have identified YSOs in all known major star formation complexes in NGC 6822 (Hubble I/III, Hubble IV, Hubble V, Spitzer I, Spitzer II and Hubble X), but also in smaller star formation sites: the HII regions BHD 9/10 and 27 (Jones et al., 2019), as well as new regions of star formation BHD 18 and a region to the North but physically distinct from Hubble IV, that I name Hubble IV−N. The detection of massive YSOs in new regions, especially in BHD 18 and Hubble IV−N, is very suggestive of additional star formation occurring in the bar of NGC 6822 between the major previously known SFRs. The prospect of detecting further YSOs in the bar region in the mass regime below the sensitivity of this analysis remains to be explored.

5 M 33

The work presented in this Chapter has been published in Kinson et al. (2022), with tables available on the VizieR database. Some minor adjustments were necessary to incorporate the paper into this document. These changes do not affect the methods or results presented.

M 33 is the third largest spiral galaxy in the Local Group, its favourable inclination ($i = 54°$, De Vaucouleurs et al. 1991) makes M 33 a more favourable target over the similarly distant but near edge on M 31 (e.g. Ma, 2001). M 33 contains several prominent H II regions and GMCs which have have been studied widely (Gratier et al., 2010b; Miura et al., 2012; Corbelli et al., 2017; Alexeeva & Zhao, 2022). Using near-IR photometry massive YSOs have been identified in the largest of these regions, NGC 604 (Fariña et al., 2012). I apply a PRF classifier (Chap. 2) using a methodology tested in NGC 6822 (Chap. 4), to classify point sources across M 33. This allows me to identify massive YSOs across the whole disk of a spiral galaxy outside the MW for the first time.

The PRF methodology was applied to M 33 point sources, using a set of six features: 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. The following section details the sources selected for PRF training.

5.1 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

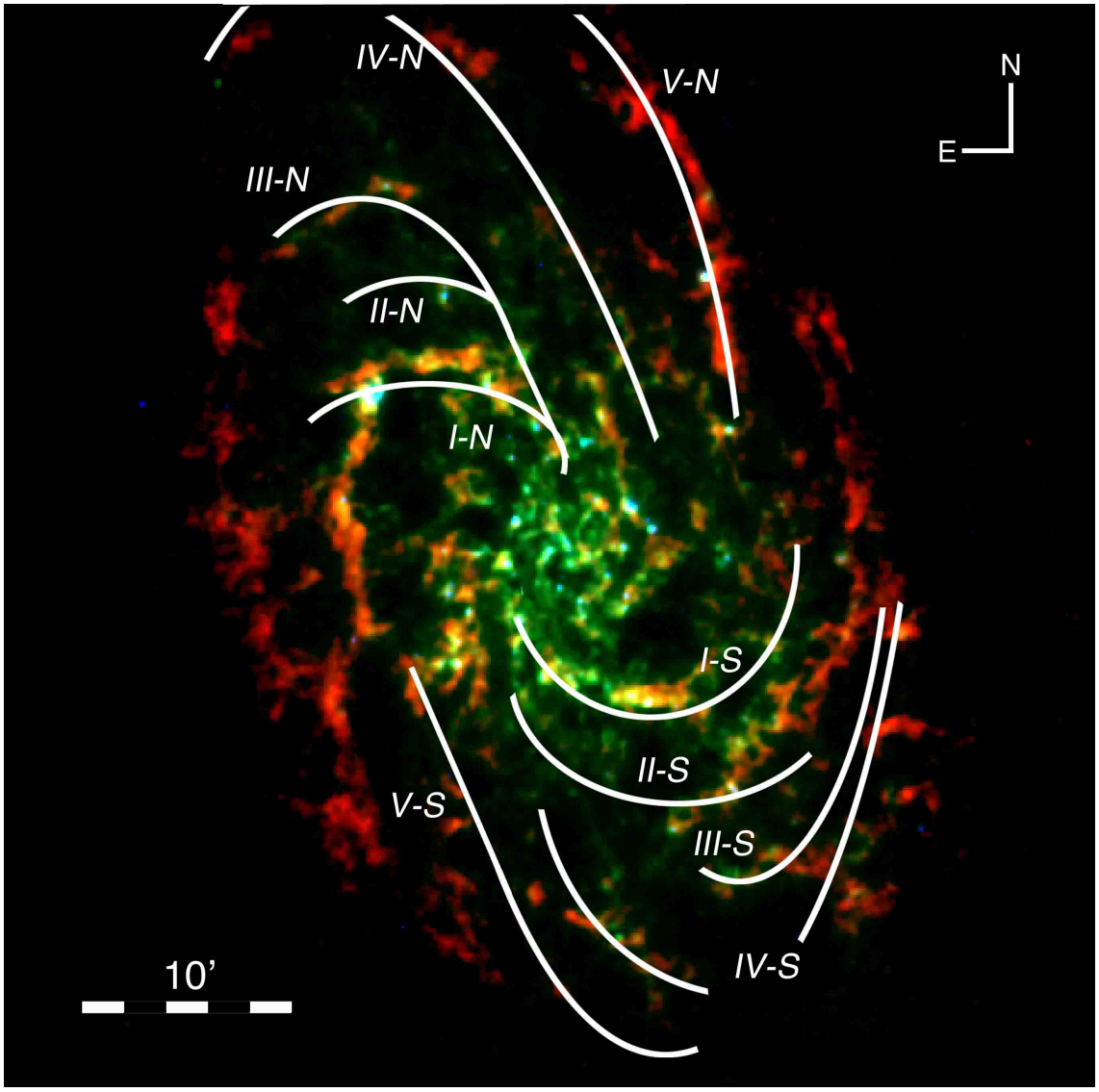

Figure 5.1: An RGB image of M 33, showing VLA H I (red, Gratier et al., 2010b), 250 μm *Herschel*-SPIRE (green, Kramer et al., 2010), 24 μm *Spitzer*-MIPS (blue, Engelbracht et al., 2004). The figure covers the same footprint as the near-IR WFCAM catalogue of Javadi et al. (2015). The spiral arm identifications, adapted from Humphreys & Sandage (1980), are shown in white.

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. Spectroscopically confirmed PN in M 33 from the catalogues of Magrini et al. (2009) and Bresolin et al. (2010) were compiled (see also Delgado-Inglada et al. 2020). Crossmatching the PN to the near-IR catalogue gave only four sources with a counterpart within a radius of 0.5 arcsec; hence insufficient sources were available to construct a training set for a PN class. WR are included in the classification for M 33 which are not considered in NGC 6822 (Sect. 4.2) due to their rarity; including classes of rare objects would however adversely affect the accuracy of the PRF classifier (see Sect. 5.2). The BS class is analogous to the MMS class in Chap. 4, including additional stellar types not numerous enough to warrant a separate class (Sect. 5.1.5).

The detailed selection criteria for each class is given in what follows. 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 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. Therefore, it is the case that whilst 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.

5.1.1 Foreground Galactic sources

The training set of Galactic foreground contaminants includes sources from Massey et al. (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, using a similar PM histogram analysis as described in Sect. 4.2.3 for NGC 6822, I include near-IR sources with a *Gaia* EDR3[1] (Gaia Collaboration et al., 2020) counterpart if their proper motion and associated error is greater than $0.5\,\mathrm{mas\,yr^{-1}}$ in both RA and Dec components. Near-IR colour cuts at $0.3 < J - K_s < 0.9\,\mathrm{mag}$, 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 I select only sources in the conspicuous vertical foreground sequences (see Fig. 5.3).

Foreground sources identified spectroscopically or with *Gaia* proper motions extend only to $K_s \sim 16.5\,\mathrm{mag}$; in order to accurately train the PRF, foreground sources at magnitudes down to the limit of the near-IR catalogue ($K_s \sim 20.5\,\mathrm{mag}$) are needed. For this purpose, I 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 sampled from a Gaussian distribution based on 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 3.2.2.

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

5.1.2 Active galaxies

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). I 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).

5.1.3 Asymptotic giant branch stars

AGB stars 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. 5.3) 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 $\mathrm{CN}-\mathrm{TiO}$ colour cuts defined by Rowe et al. (2005) I identified both OAGBs and CAGBs from their catalogue. Both classes of AGB have $V-I>1.8\,\mathrm{mag}$, with OAGBs having colours of $\mathrm{CN}-\mathrm{TiO}<-0.2\,\mathrm{mag}$ and CAGBs $\mathrm{CN}-\mathrm{TiO}>0.3\,\mathrm{mag}$. From this sample any sources with *Gaia* proper motions consistent with a Galactic foreground dwarf (see Sect.5.1.1) were removed. Sources with a different classification in the spectroscopic samples of 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 figure 10) which are adopted to refine the samples from Rowe et al. (2005). These cuts are shown in Fig. 5.3. I also select only sources brighter than the M 33 TRGB magnitude, $K_s = 18.11\,\mathrm{mag}$ (Ren et al., 2021). Finally

for the OAGBs I 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), to separate from RSGs.

5.1.4 Red giants and supergiants

RGB stars are a significant population that exhibit similar colours and magnitudes to faint YSOs. I began with M-type stars identified in Rowe et al. (2005) as described in subsection Sect. 5.1.3. 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), any source with a *Gaia* proper motion consistent with a Galactic star (see Sect.5.1.1) were rejected. 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 Sect.5.1.1.

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 (scaled to M 33 distance, see table. 1.1), 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 5.2.

RSG stars are a young population ($\sim 10 - 30$ 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. I 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.

5.1.5 Blue stars

I 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 my 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 Sect. 5.1.4) their catalogues cannot be used to populate the 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 included are sources they classified as Luminous Blue Variables (LBV), yellow super giant 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 et al., 2012). Further colour cuts are set at $-0.5 < J - K_s < 0.3$ mag.

5.1.6 Wolf-Rayet stars

Wolf-Rayet (WR) stars are a relatively rare population with only ~ 200 confirmed across the disk 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 my classification scheme. The WR training set is comprised

of spectroscopically confirmed sources from the catalogues of Massey et al. (2016) and Neugent & Massey (2011).

5.1.7 Young stellar objects

The 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 Sect. 3.2.1 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 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. I 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 SED fitting with evolutionary models (Robitaille et al., 2006; Robitaille, 2017), and confirmed using machine learning techniques (Kinson et al., 2021, see Sect. 4.3.2).

5.2 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 et al., 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 $\sim$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 Sect. 5.3.

The RGB training sources come from two sets of data, one in M 33 and an-

Table 5.1: Number of sources for each class for the five training sets (see Sect. 5.1) after down-sampling of large training classes (see Sect. 5.2).

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

other from NGC 6822 (see Sect. 5.1.4). 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, 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 by comparing the fraction of NGC 6822 RGBs above and below the M 33 sample cut-off when scaled to the same distance (see Sect. 5.1.4). Reducing the M 33 RGB sub-sample by a factor of 1 in 24 provides homogeneity in the combined NGC 6822 and M 33 RGB sub-samples 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).

In total I 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. The selection of sources in each training class was checked to ensure that the parameter space for each class was fully represented in each down-sampling (e.g. Fig. 5.3). Nevertheless it is somewhat inevitable that down-sampling of the large classes

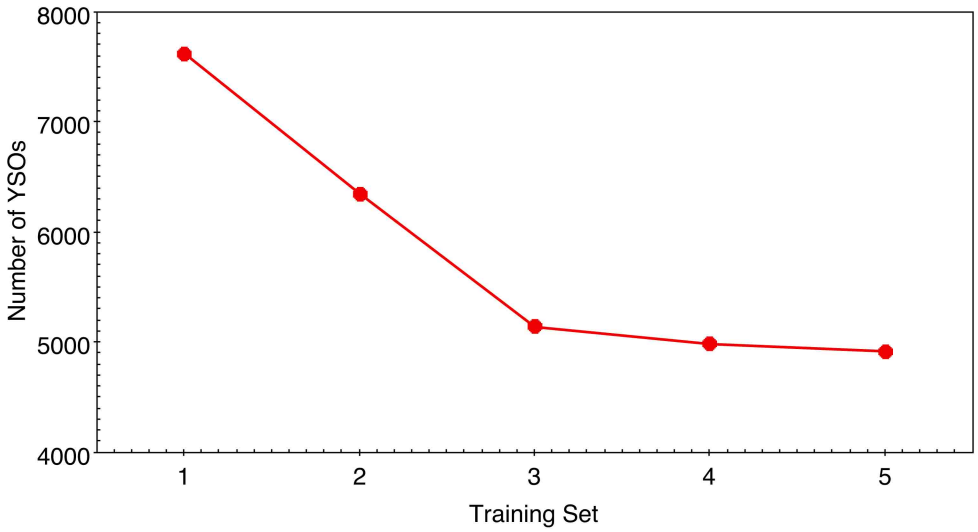

Figure 5.2: The number of $n_{\rm YSO} = 20$ sources classified in common for increasing down-sampled training set selections.

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. If the stochastic effects from the selection of training data dominate, then the number of sources in common classified for a particular class will decrease with each additional down-sampling selection.

Each down-sampled training set was used for 20 PRF runs, hence a maximum score of $n_{\rm class} = 20$ is available for a source per down-sampled training set. Figure. 5.2 shows the number of sources which achieve $n_{\rm YSO} = 20$ for distinct down-sampling selections; beyond four selections the number of YSOs returned in common between the runs does not decrease significantly. This indicated that five distinct down-sampling selections were sufficient to robustly account for any stochastic effects.

The number of sources in each training set class are given in Table 5.1. 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. With the training set in

M 33 defined above, the PRF classifier was applied 100 times to the remaining 162,746 sources with all three $J\,H\,K_s$ bands in the catalogue.

5.3 Confusion matrices

Confusion matrices provide a helpful visualisation 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. 5.4 the confusion matrices, both non-normalised and normalised, 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.

Figure. 5.5 shows the PRF matrices, using the same random seed as those shown in Fig. 5.4, with down-sampling applied as described in Sect. 5.2. In general an improvement in the overall PRF classification accuracy, exemplified by the strong diagonal feature in the normalised 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. 5.5); source misclassification and contamination are discussed in the following section.

5.3.1 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

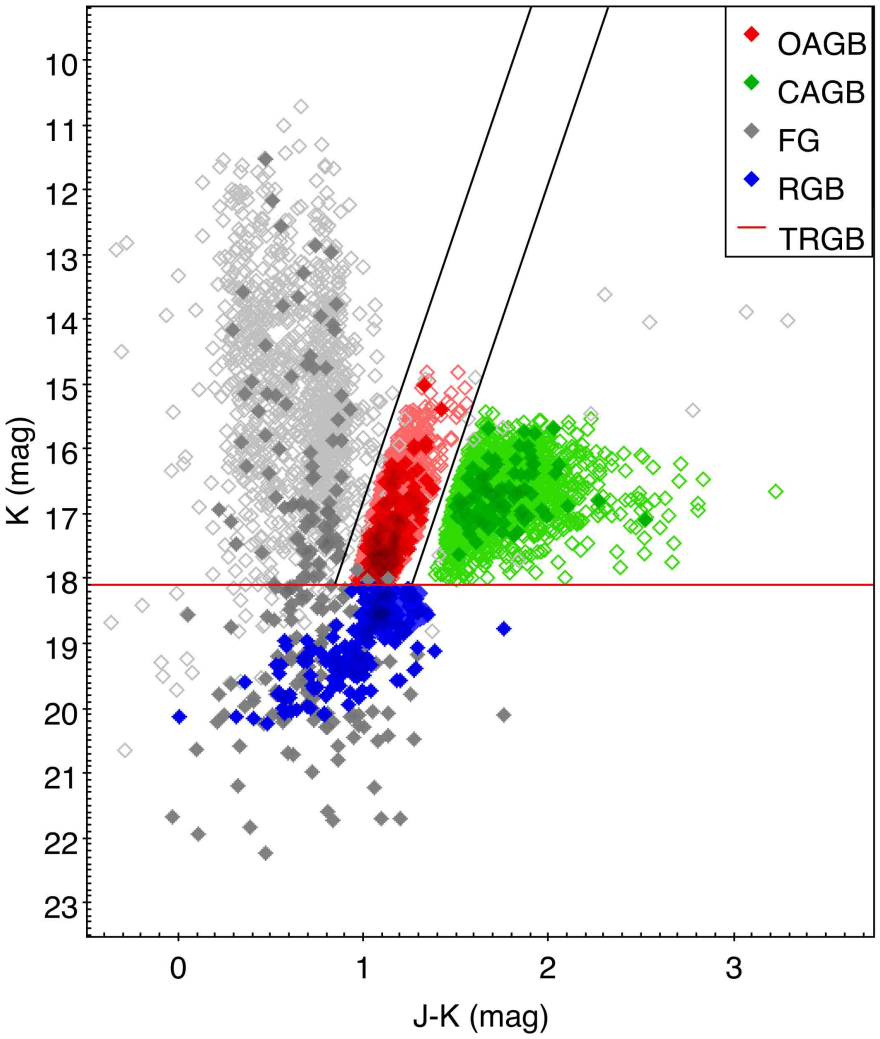

Figure 5.3: A CMD showing the four large classes, with the full set of data shown by open symbols and the down-sampled data by filled symbols. The parameter space for each class is well represented by the down-sampled data. The TRGB magnitude ($K_s = 18.11$ mag) and AGB colour-cuts adapted from Ren et al. (2021) are shown by the red and black lines respectively.

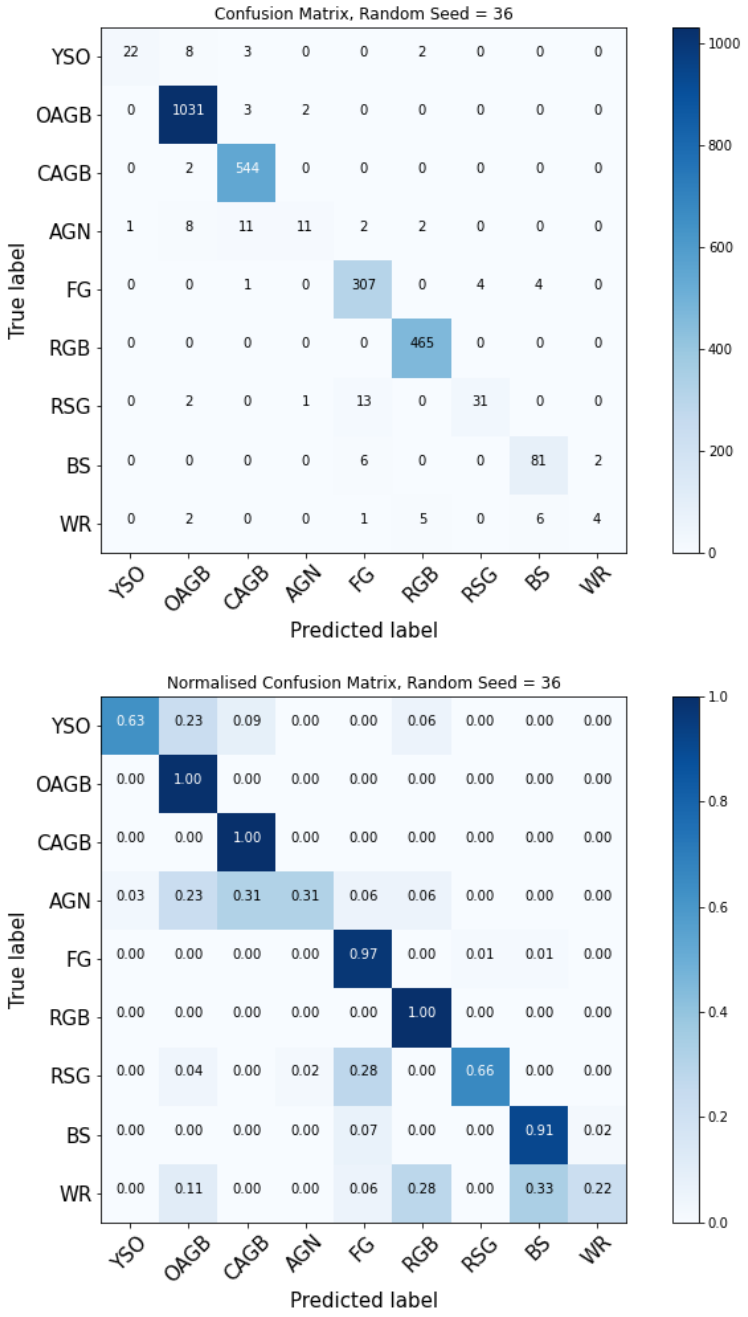

Figure 5.4: Non-normalised (top) and normalised (bottom) confusion matrices for an example PRF run with no class down-sampling (see text). The large classes achieve high accuracy, however for the smaller classes high levels of confusion are evident.

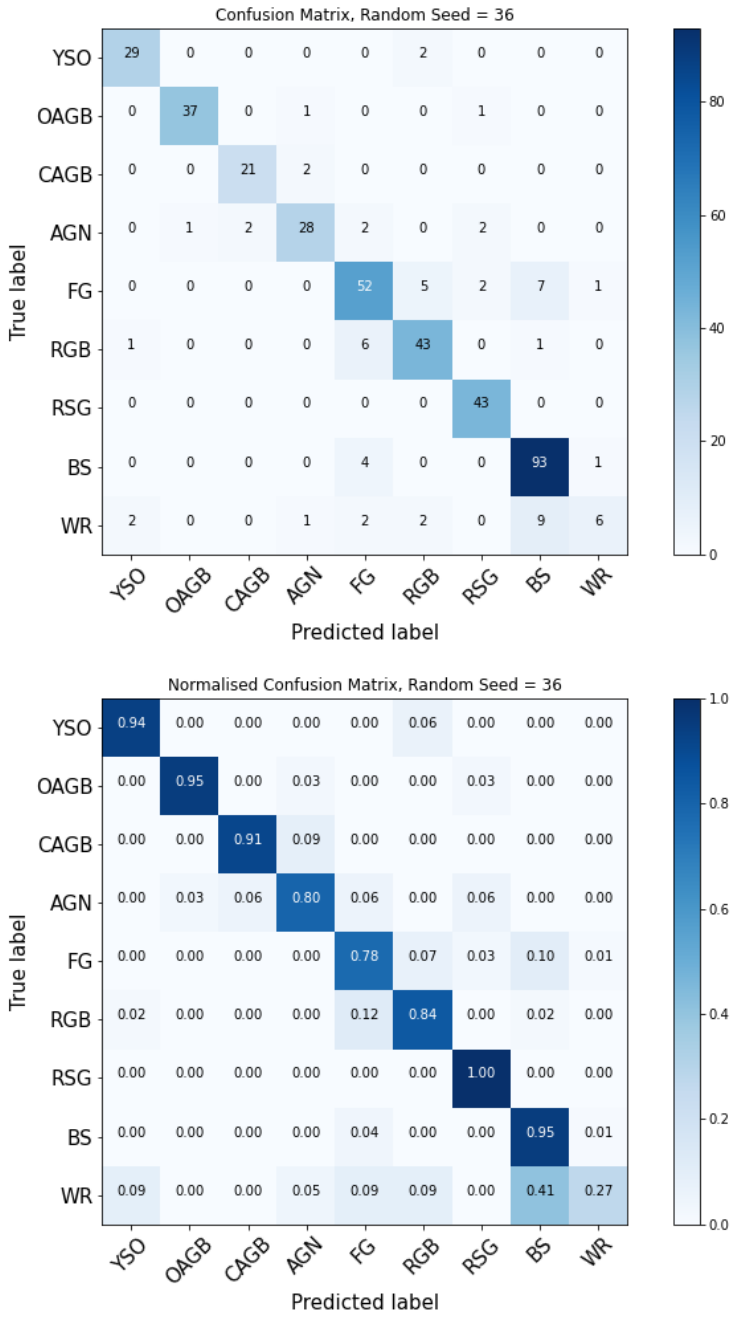

Figure 5.5: Non-normalised and normalised confusion matrices (respectively top and bottom) for the PRF run using the same random seed as those shown in Fig.5.4, but here with class down-sampling (see text). The misclassifications for the smaller classes are very effectively reduced.

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.

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. I use training set sources that after down-sampling are returned to the main catalogue to further investigate YSO contamination in the final classifier output (Sect. 5.4).

There are eight classified sources which have counterparts in the PN catalogue mentioned in Sect. 5.1 (see also Delgado-Inglada et al., 2020), within ~ 1 arc sec. Of these sources two are classified as YSOs, with the remainder three RGBs, one OAGB, one CAGB and one FG. Due to the larger separations (~ 1 arc sec), it is unclear whether these are the correct counterparts. To investigate this further, I compared the scaled brightnesses of LMC PN taken from Reid (2014) with the M 33 classified catalogue. Both YSOs coincident with PN have $K_s \sim 17.5$ mag equivalent to the most extreme LMC PN. Furthermore the majority of PN would be expected to lie below the magnitude range of the classified YSOs (see Fig. 5.7). Based on this comparison, contamination of YSOs by unidentified PN is not expected to be significant.

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 out-scoring the correct classification (see Fig. 5.5). 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.

RSGs are consistently correctly classified in many PRF runs (up to 100 per cent

as shown in Fig. 5.5). These sources occupy a region of parameter space distinct from other classes: RSGs are as bright as the upper vertical FG sequences but much redder. Coupled with above average far-IR brightnesses, this allows the PRF to easily distinguish between RSG and FG sources.

For the AGNs some confusion with the OAGB and CAGB classes is seen, 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., 2022a). A similar effect was also seen in AGN classifications behind NGC 6822 (Kinson et al., 2021).

5.4 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 consistently into the same class (i.e. $max(n_{\rm class}) = 100$), more specifically between 30 per cent in the central region affected by crowding and 42 per cent in the outer disk (see Sect. 5.5). Figure 5.6 shows the distribution of $n_{\rm class}$ values (see Sect. 4.2) for all non-training set sources. The two smaller peaks at $n_{\rm class} = 20$ and 80 arise from the five down-sampling runs. They occur when the 20 PRF runs for a particular down-sampled training set attribute $n_{\rm class} = 0$ or $n_{\rm class} = 20$ to a source, in a class distinct from the classification of the four other down-samplings sets. This is a consequence of high classifier confidence across multiple runs with the same training set data and highlights the importance of using multiple down-samplings to account for such stochastic effects (see Sect. 5.2).

The $n_{\rm class} = 100$ sources are included in my 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 5.2.

In Fig. 5.7, I 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,

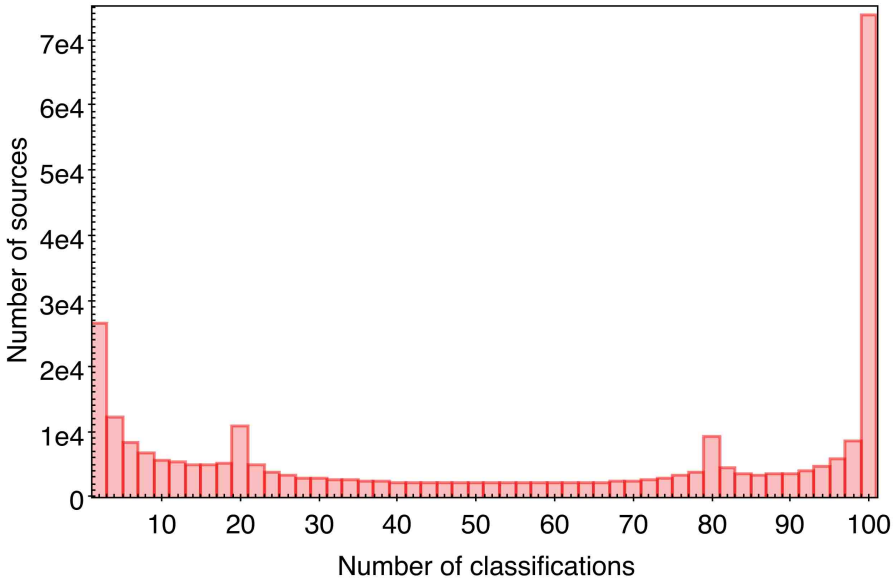

Figure 5.6: The distribution of the number of PRF classifications for each source across all classes. The most common classification is $n_{\rm class} = 100$. Smaller peaks at $n_{\rm class} = 20$ and 80 can be seen where sources with $n_{\rm class} = 0$ or 20 in a single down-sampling affect the overall distribution (see text). The very large peak at $n_{\rm class} = 0$ is omitted from the histogram for clarity.

Table 5.2: 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 Sect. 5.2).

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

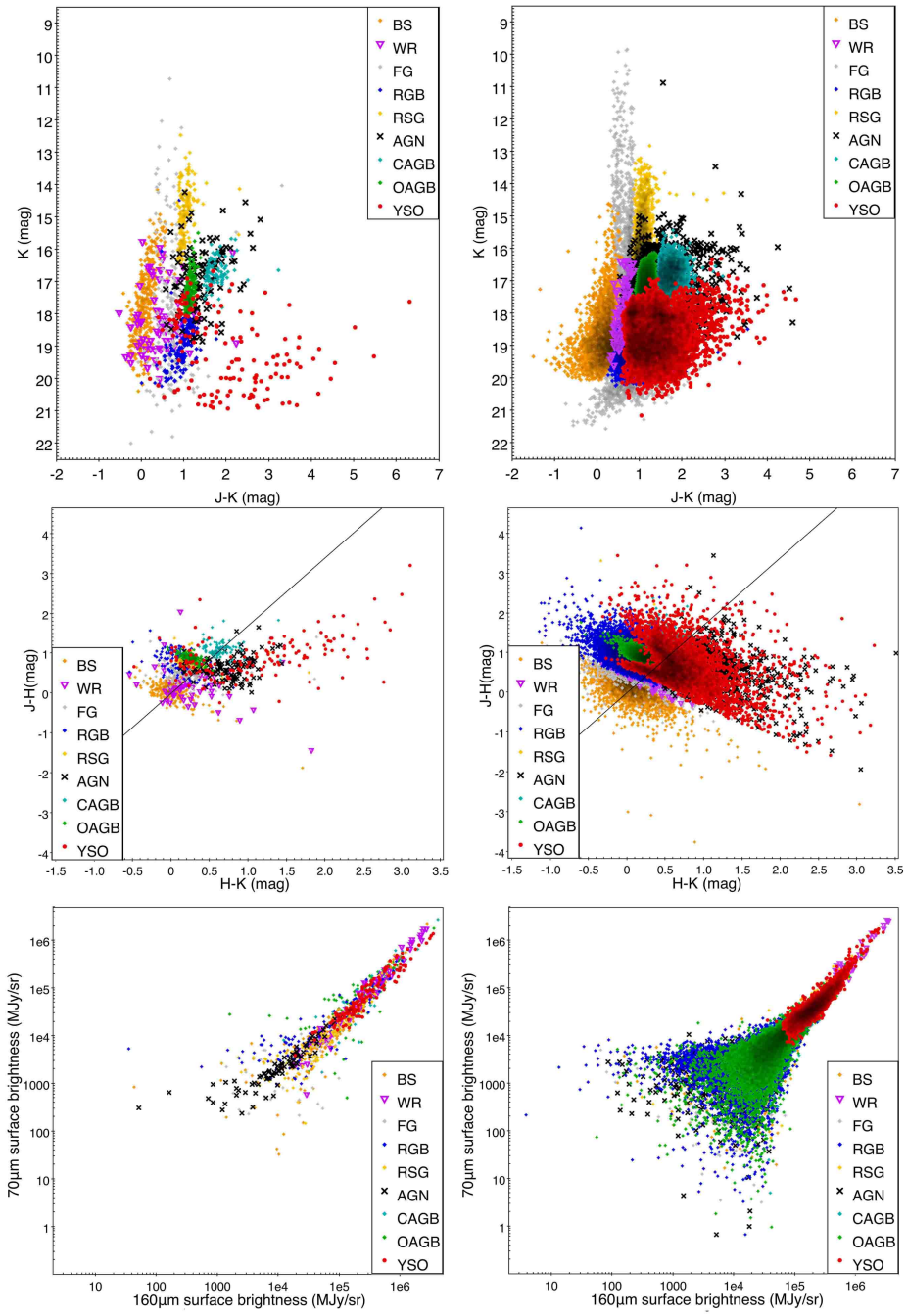

Figure 5.7: CMD, CCD and far-IR brightness plots of the training set sources (left) and for the classified sources (right). Colour-coding is given in the legend. The reddening line shown in the CCD plots is derived using the coefficients from Rieke & Lebofsky (1985).

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 fact that some training set YSOs can have J- and H-band magnitudes fainter than the near-IR catalogue's detection thresholds (see Sect. 3.2.1). This arises from practical considerations in the design of the near-IR observations, with shorter wavelength images not deep enough to characterise the redder sources, being these YSOs or AGBs. Therefore the faintest YSOs identified are not particularly red, and, as expected, no classified YSOs are found outside the colour and magnitude ranges described by the training set YSOs. Rather than a bias of the PRF algorithm this is instead evidence of the completeness issues in the near-IR catalogue.

Figure 5.7 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 Sect. 5.3.1.

I discussed potential YSO contamination in Sect. 5.3.1. 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. In total 655 sources are excluded from the RGB, FG, OAGB and CAGB training sets after down-sampling (see Sect. 5.2). These sources from the training set that are returned to the catalogue with known classification are used to provide an additional estimate of class contamination alongside the statistics provided by the confusion matrices (Sect. 5.3). Of these 655 sources, 358 (~ 55 per cent) are classified by the PRF (i.e. achieve $n_{\rm class} = 100$), 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 Sect. 5.3.1, none of the RGBs are misclassified, as YSO or any other class.

5.5 Spatial distributions

As noted in Sect. 3.2.1 sensitivity issues become apparent for $K_s > 19.2$ mag. The effects of source crowding 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. While crowding affects the identification for all classes, classes dominated by fainter sources are more severely affected.

In Figs. 5.8, 5.9 and 5.10 the spatial distributions of classified sources for each class are shown. Figures 5.11, 5.12 and 5.13 show the spatial distributions of classified sources in the central region of M 33, to highlight the complex environment. Some salient features of non-YSO distributions are briefly described, however a thorough discussion is beyond the scope of this project. The YSO distribution are discussed in Sect. 5.7.

AGN and FG sources are fairly evenly distributed across the field as expected. AGNs are less frequently 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 Sects. 5.3.1 and 5.4 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 disk of M 33 in agreement with the source density distributions previously reported (Javadi et al., 2015; Williams et al., 2021). 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.,

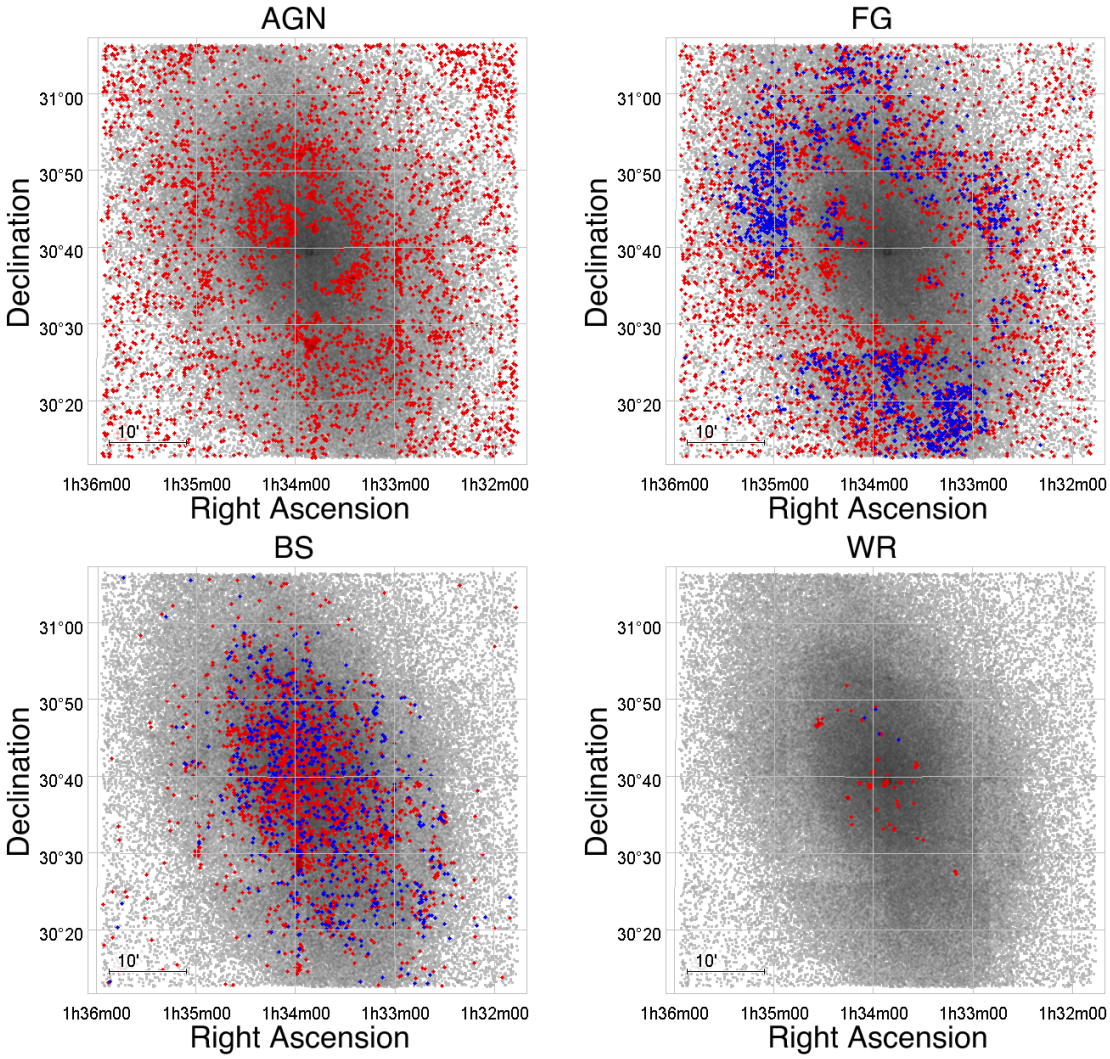

Figure 5.8: Spatial distributions for the AGN, FG, BS and WR classes. Sources with $K_s < 19.2$ mag and $K_s \geqslant 19.2$ mag are shown respectively in red and blue. The full catalogue is shown in the background.

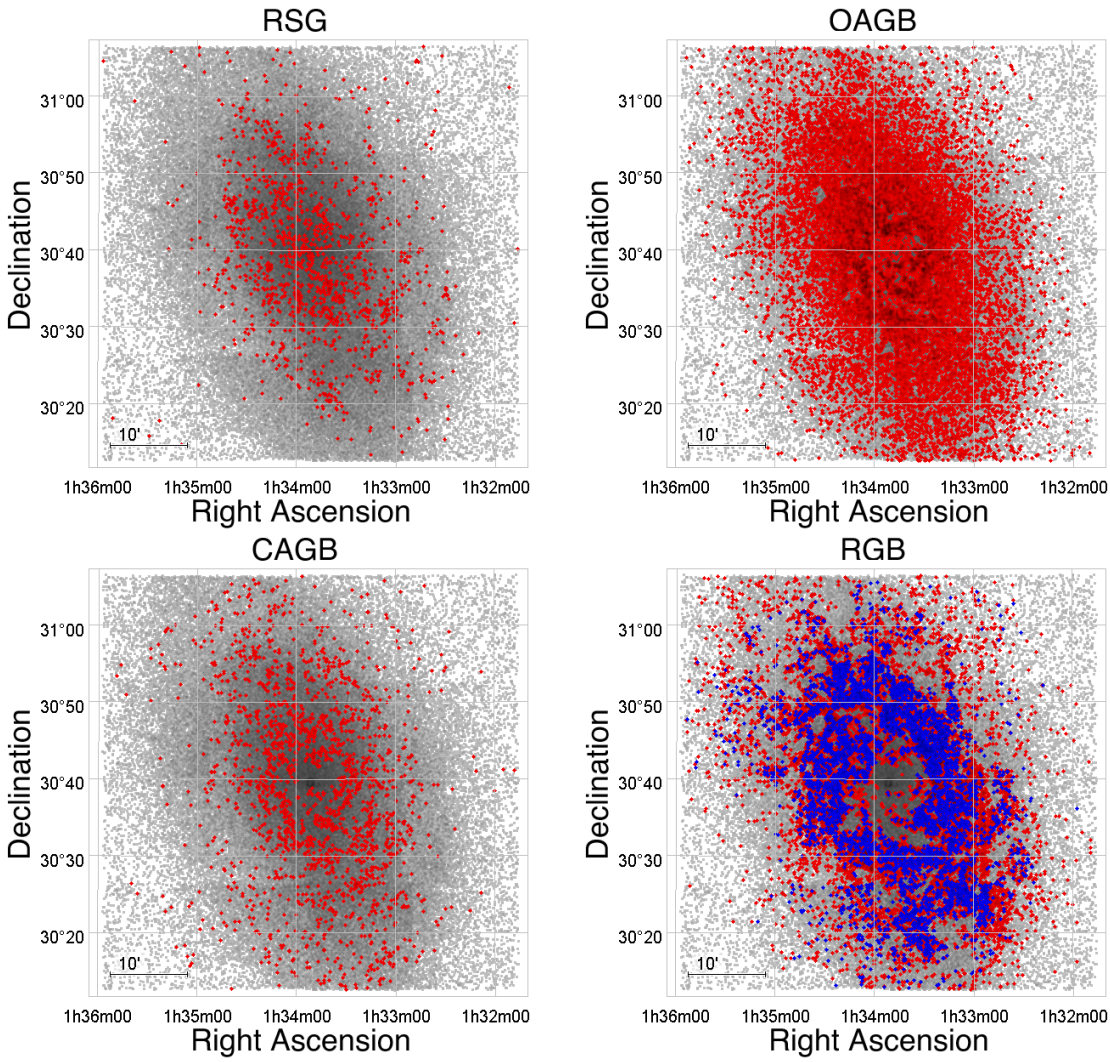

Figure 5.9: Spatial distributions for the RSG, OAGB, CAGB and RGB classes. Colour-coding as in Fig. 5.8.

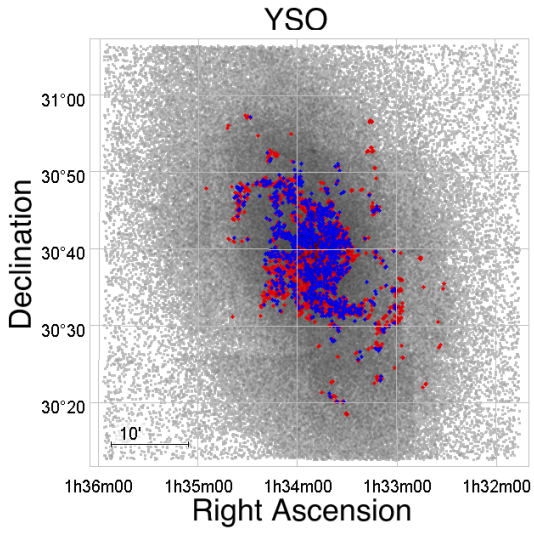

Figure 5.10: Spatial distributions for the YSO class. Colour-coding as in Fig. 5.8.

2021) is recovered. 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 my 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 figure 11) and Ren et al. (2021, see their figure 11). The WR source distribution, even though very sparse, loosely follows the distribution of YSOs (Sect. 5.7).

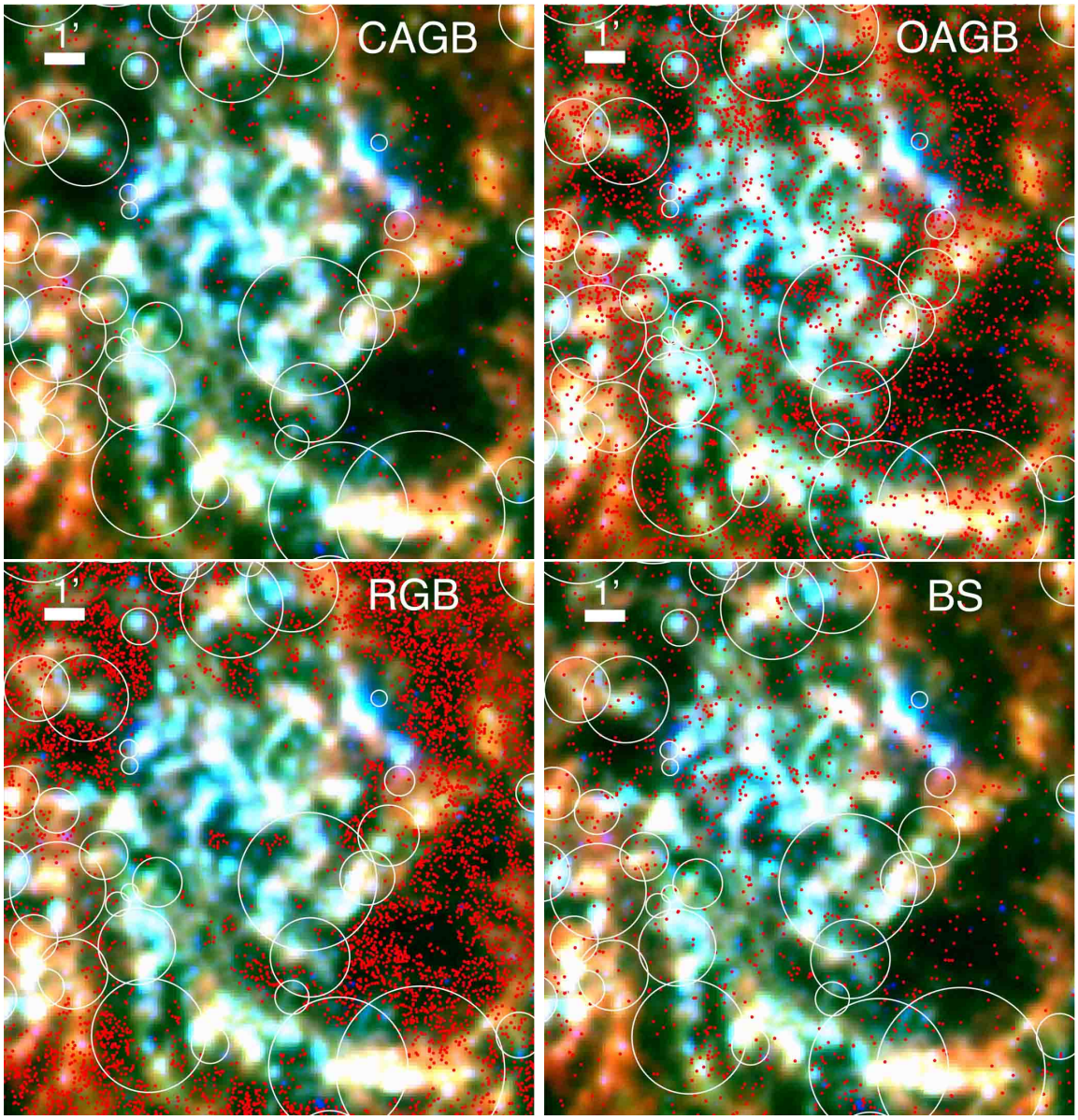

Figure 5.11: Spatial distributions of classified CAGB, OAGB, RGB and BS sources (red circles) in the central region of M 33, overlaid on an RGB image: VLA H I (red, Gratier et al., 2010b), 250 μm *Herschel*-SPIRE (green, Kramer et al., 2010), 24 μm *Spitzer*-MIPS (blue, Engelbracht et al., 2004). SFRs identified by the DBSCAN analysis (Sect. 5.6) are shown by the white circles.

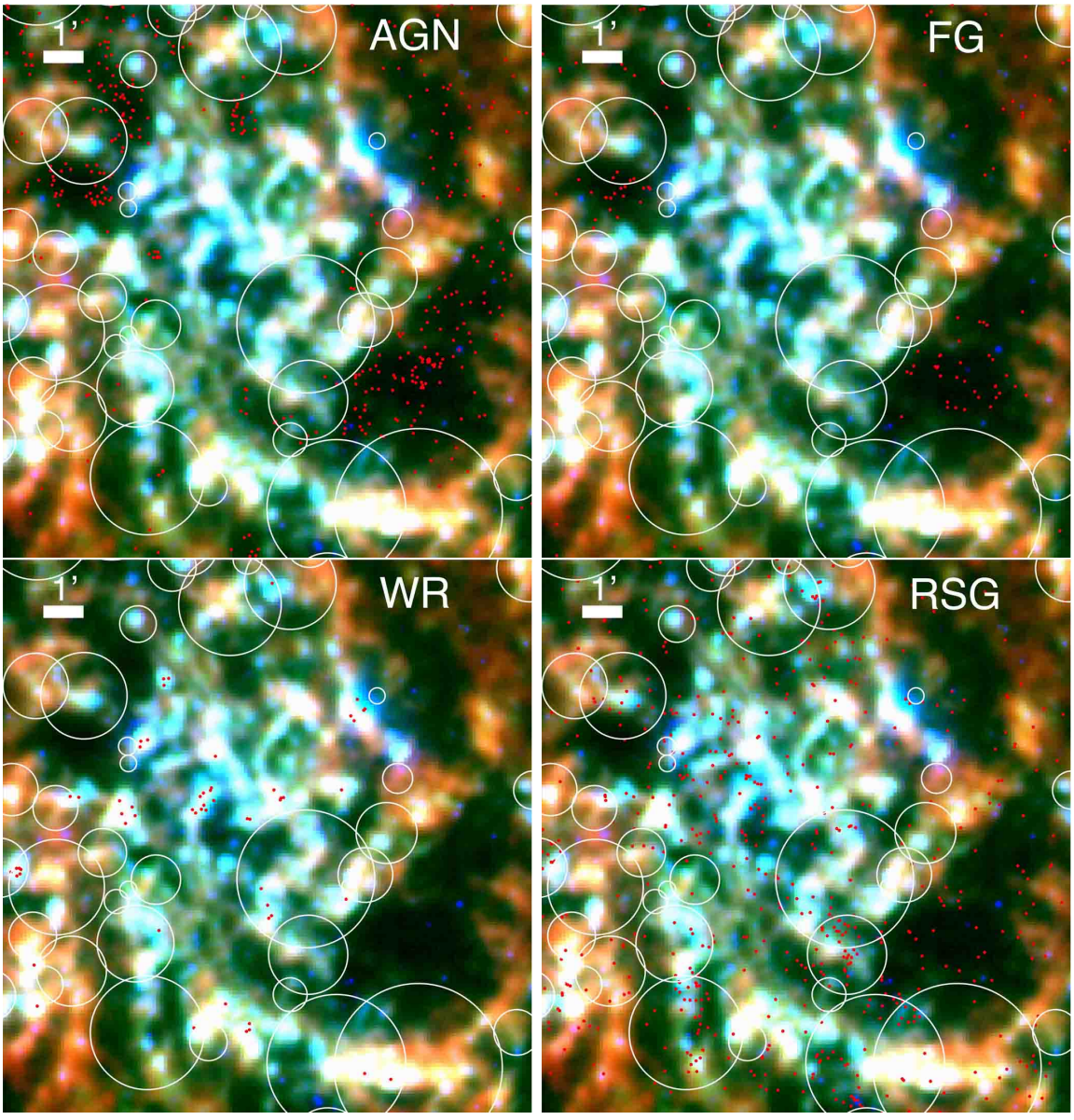

Figure 5.12: Spatial distributions of classified AGN, FG, WR and RSG sources. Images and symbols as in Fig. 5.11.

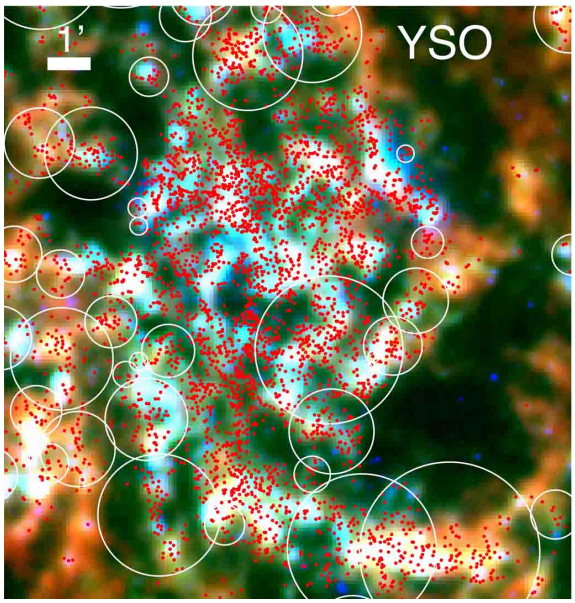

Figure 5.13: Spatial distributions of classified YSO sources.Images and symbols as in Fig. 5.11.

5.6 YSO distribution and clustering

The PRF identifies 4985 YSOs across the disk of M 33; their properties are listed in Table 5.3 and their distribution is shown in Fig. 5.14. As already discussed, the PRF classifies $\sim$30 to $\sim$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 (see Fig. 5.1). Arms I-N and I-S contain $\sim$300 YSOs each, with a similar total YSO mass (see Sect. 5.8 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.

I 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

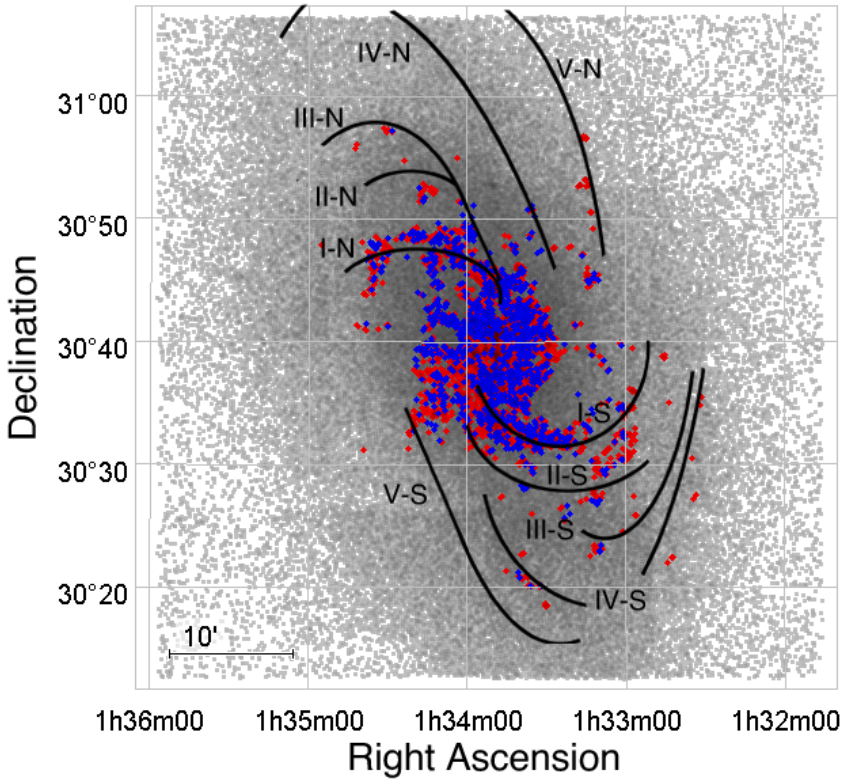

Figure 5.14: YSO distribution in M 33, with the spiral structure adapted from Humphreys & Sandage (1980) overlaid (colour-coding as in Fig. 5.8).

which finds density-based associations in spatial data. This process was performed using deprojected coordinates (see Appendix C, 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). I optimised 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 ($\epsilon = 0.1551$ kpc) identified 23 spatial associations or "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. 5.4). This process was repeated five times, after which the ϵ distance returned by the k-NN analysis effectively plateaued. This methodology of varying ϵ in successive runs was compared with Hierarchical-DBSCAN (HDBSCAN, McInnes et al., 2017), which employs a singular, automated ϵ selection method. Across the disk of M 33, HDBSCAN is significantly more biased towards larger cluster radii compared to DBSCAN. Furthermore HDBSCAN performs significantly worse in the outer regions of the disk allocating seemingly isolated YSOs at very large distances (~ 500 pc) to clusters, whereas using DBSCAN over multiple runs selecting decreasing ϵ values gives a tighter spatial association. These unrealistic larger cluster radii would negatively affect later analysis of SFRs (Sect. 5.7), and hence using DBSCAN with multiple ϵ values was selected as the preferable clustering algorithm. Overall, DBSCAN identifies 62 YSO clusters.

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.,

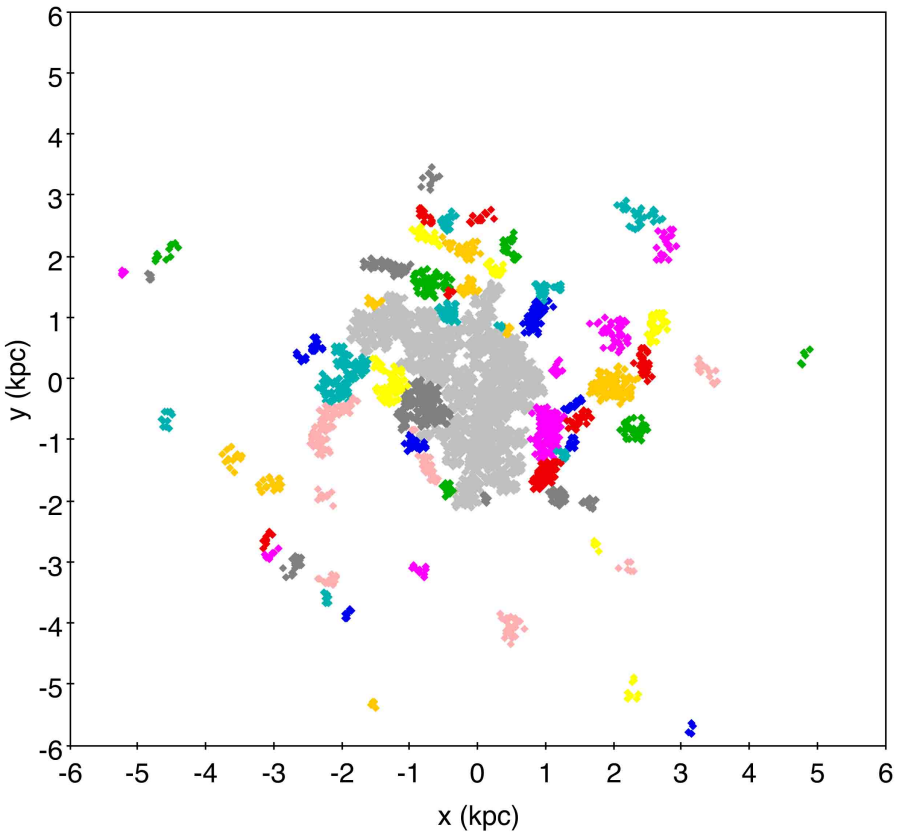

Figure 5.15: Clusters of YSOs identified by DBSCAN, displayed in deprojected coordinates. The central region (see text) without identified clusters is shown in light grey colour. This projection is rotated by 90 degree clockwise with respect to the sky coordinates shown in Fig. 5.14.

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 disk 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 distribution of GMC sizes in M 33 analysed by Corbelli et al. (2017); the relationship between SFRs and GMCs is discussed in Sect. 5.7.3. The SFR spatial distribution in deprojected coordinates is shown in Fig. 5.15. 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 Sect. 5.7). SFR properties are listed in Table 5.5.

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 5.5, 562 were unclustered and 2437 were left in the central dense "remnant" ($\sim 11.6 \times 10.4\,\mathrm{arcmin}^2$ or $2.8 \times 2.5\,\mathrm{kpc}^2$ in size, light grey in Fig. 5.15). 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 in the arms. As already discussed, fewer than expected RGB sources are identified in this region (see Figs. 5.9, 5.11 and Sect. 5.5). Given their expected centrally peaking distribution in the M 33 disk and strong overlap in colour-magnitude space with YSOs (see Fig. 5.7), RGBs are an important contaminant class (see Sect 5.3.1), even if no known RGBs are misclassified as YSOs by the PRF (Sect. 5.4). Nevertheless, assuming in extremis that all 811 YSOs overlapping the RGB region of the CMD space are contaminants, I estimate that at most 30 per cent of YSOs in this region could be wrongly classified as RGBs. I take this into account in the analysis in Sect. 5.8.

In Fig. 5.16 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 central region. Figure 5.17 shows the distribution of number of YSOs and radii for the 68 SFR regions identified with DBSCAN.

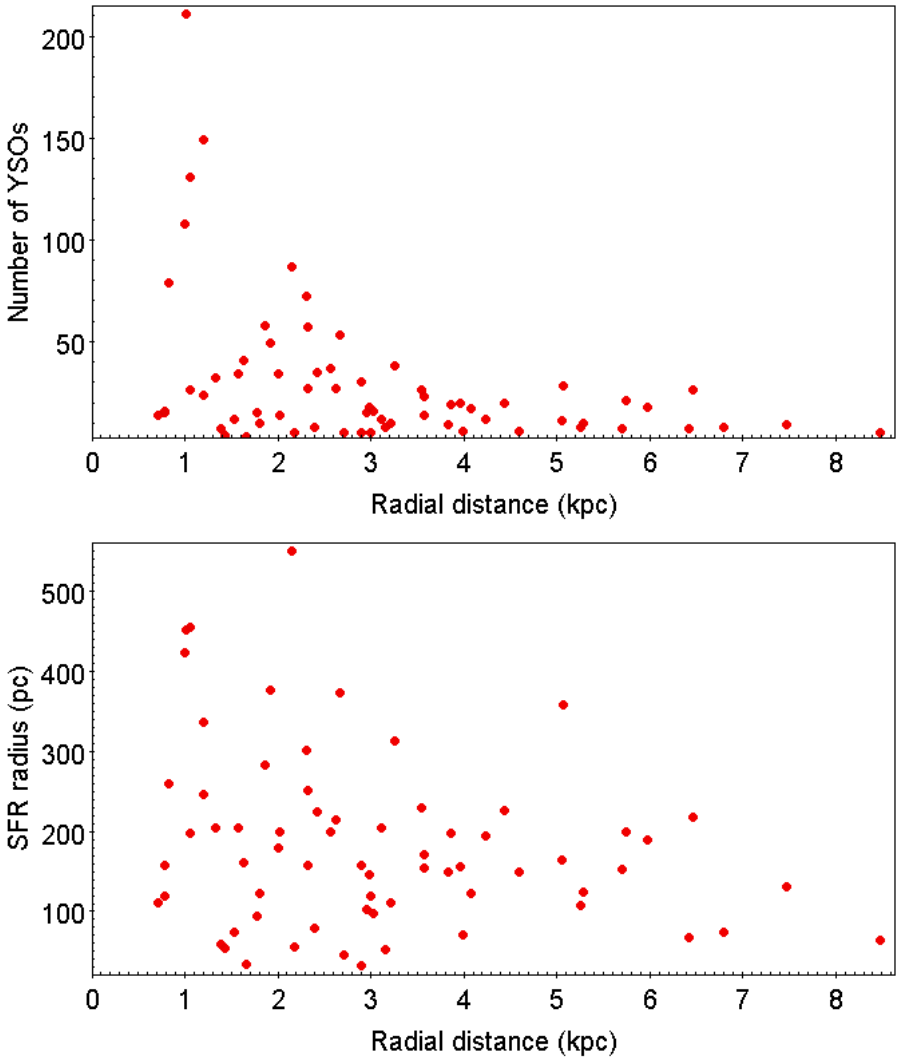

Figure 5.16: Number of YSOs (top) and radius (bottom) for each SFR identified by DBSCAN as a function of radial distance. A decrease in size with increasing distance from the centre is seen in both panels.

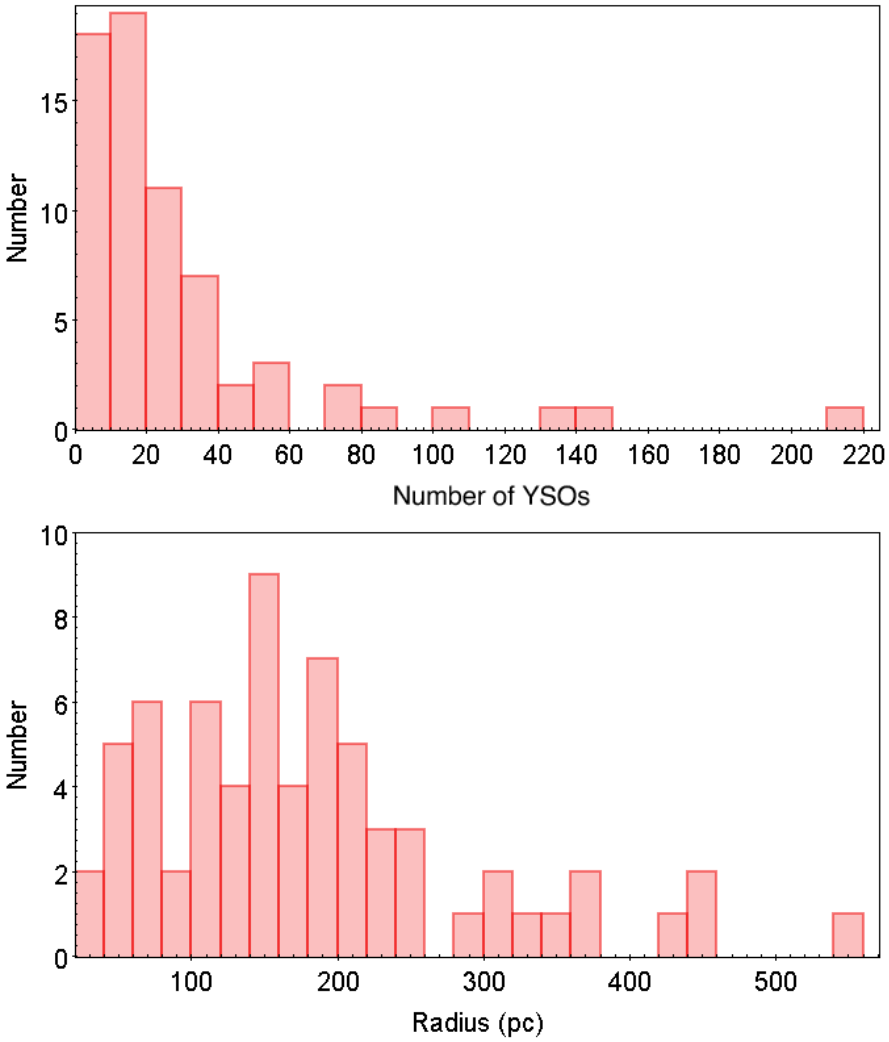

Figure 5.17: Histograms showing the distribution of number of YSOs and radii for the 68 YSO clusters identified with DBSCAN.

Table 5.3: 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 Sect. 5.8. A sample of the table is provided here, the full catalogue is available in the online material of Kinson et al. (2022).

RA (J2000)	Dec (J2000)	J	$J_{\rm err}$	H	$H_{\rm err}$	K	$K_{\rm err}$	SFR	mass
h:m:s	deg:m:s	mag	mag	mag	mag	mag	mag	ID	$M_\odot$
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

Table 5.4: 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

Table 5.5: Catalogue of SFRs in M 33 identified using DBSCAN. The evolution score is discussed in Sect. 5.7.2. A sample of the table is provided here, the full version is available in the online material of Kinson et al. (2022).

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	

5.7 The star forming regions in M33

In this section I 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.

5.7.1 SFR observed properties

Integrated optical to far-IR brightnesses can be used to characterise and probe the star formation 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 ionising radiation and winds of those stars to clear the surrounding, obscuring dust must have passed. Hence the ratio of Hα to 24μm 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 disk of M 33 (Moeller & Calzetti, 2022).

The ratio of far-IR emission 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-μm to 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 onto the main sequence (e.g. Lada & Lada, 2003; Portegies Zwart et al., 2010). Hence by comparing the ratios of Hα to 24μm and 250μm to 500μm ([Hα]/[24μm] and [250μm]/[500μm] 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 Sect. 3.4 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 5.5). In order to calibrate the properties and evolutionary status of SFRs in M 33 I used regions in NGC 6822 that have been well-characterised in the literature (Schruba et al., 2017; Jones et al., 2019; Kinson et al., 2021) as a benchmark. 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. I do not include in this analysis the smaller SFRs newly identified in Kinson et al. (2021) since they are significantly smaller and an established evolutionary sequence is not available for these regions.

Figure 5.18 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 (lower panel). Fig. 5.19 shows the ratio of [Hα]/[24μm] against [250μm]/[500μm]. The Hα and 24-μm brightnesses appear loosely correlated while the 250-μm and 500-μm brightnesses show a much tighter correlation (for M 33 SFRs, $r_{\mathrm{pearson}} \sim 0.24$ 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 Sect. 3.4, 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.

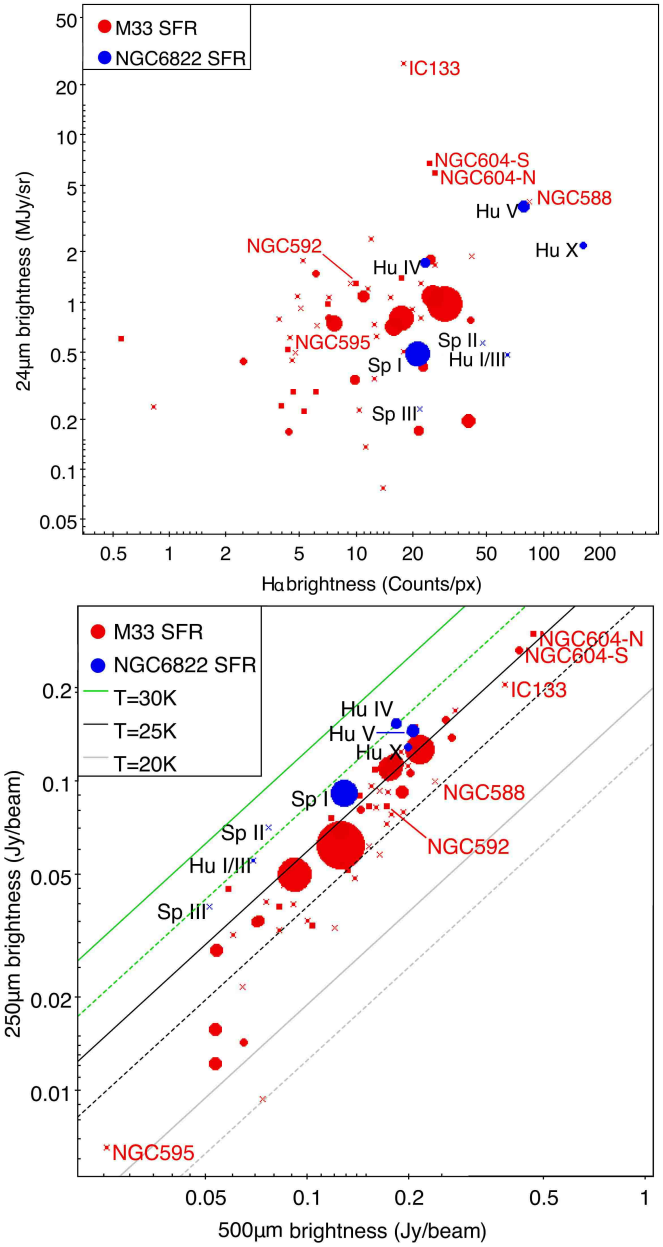

Figure 5.18: Photometric measurements for each SFR in M 33 and NGC 6822 (red and blue symbols respectively): Hα and 24μm (upper), 250 and 500μm (lower). The symbol size is proportional to the number of YSOs in each region (crosses mark particularly small regions); YSOs numbers for each SFR in M 33 and NGC 6822 are respectively from my analysis and from Kinson et al. (2021). In the lower 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).

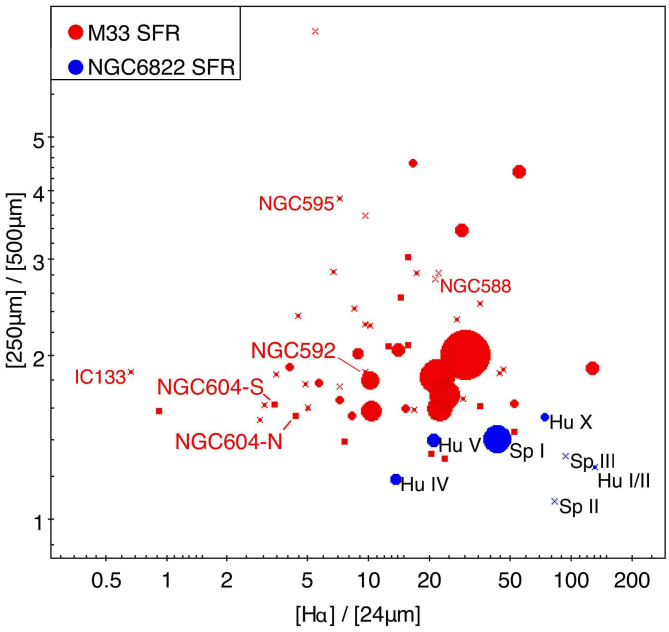

Figure 5.19: The ratio of photometric measurements [Hα]/[24μm] and [250μm]/[500μm] for each SFR in M 33 and NGC 6822. Symbol sizes and colours are as in Fig. 5.18.

The higher Hα brightnesses for SFRs in NGC 6822 may be a consequence of its lower metallicity ($\sim$0.2 $Z_{\odot}$, e.g. Skillman et al., 1989; Richer & McCall, 2007). At low metallicity, the interstellar medium (ISM) is more porous allowing for increased leakage of ionising 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. 5.18 (lower panel) the loci of theoretical modified blackbody emission, for dust temperatures 20, 25 and 30 K (colour-coded) and values of the dust emissivity index β ($\beta = 2$ and 1.5, solid and dashed lines respectively) are shown. 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 et al., 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 disk; for SFRs however, a value of $\beta = 2$ seems to be more appropriate (Braine et al., 2010; Tabatabaei et al., 2014). The position in Fig. 5.18 of the NGC 6822 SFRs is broadly consistent with those in M 33, with a slight offset to higher 250-μm values. This offset roughly corresponds to an increase in temperature of ~ 2 K or a variation in β of ~ 0.4. This offset could be due to the difference in dust properties: the ISM in NGC 6822 is dominated by smaller grain sizes 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 et al., 2022). With the available information it is not possible to postulate whether variations in temperature or β are more likely to be the cause of this offset.

The symbol sizes in Figs. 5.18 and 5.19 are proportional to the number of YSOs in the SFR; YSO numbers come from the DBSCAN analysis in Sect. 5.6 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 towards the ranges' averages (Fig. 5.18). 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. For instance 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 Figs. 5.16 and 5.17). 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).

5.7.2 SFR evolutionary status

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

The upper panel of Fig. 5.20 shows the rank sequence for the SFRs in NGC 6822. Using a combination of [Hα]/[24μm] 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 (see Sect. 4.9). This is broadly consistent with the position of the regions in Fig. 5.20: 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 this 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. 5.20), 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. 5.20). Some of the most prominent H II regions and SFR in M 33 are discussed further in Sect. 5.7.4.

In order to compare the evolution stage of SFRs in M 33 and NGC 6822, I convert the distance from the locus of rank parity in Fig. 5.20 into a measure of evolution, normalised to the number of sources in each sample. I call this the evolution score. A negative evolution score represents a less evolved, more embedded region in which the [250μm]/[500μm] ratio dominates over the [Hα]/[24μm] ratio. A positive value of

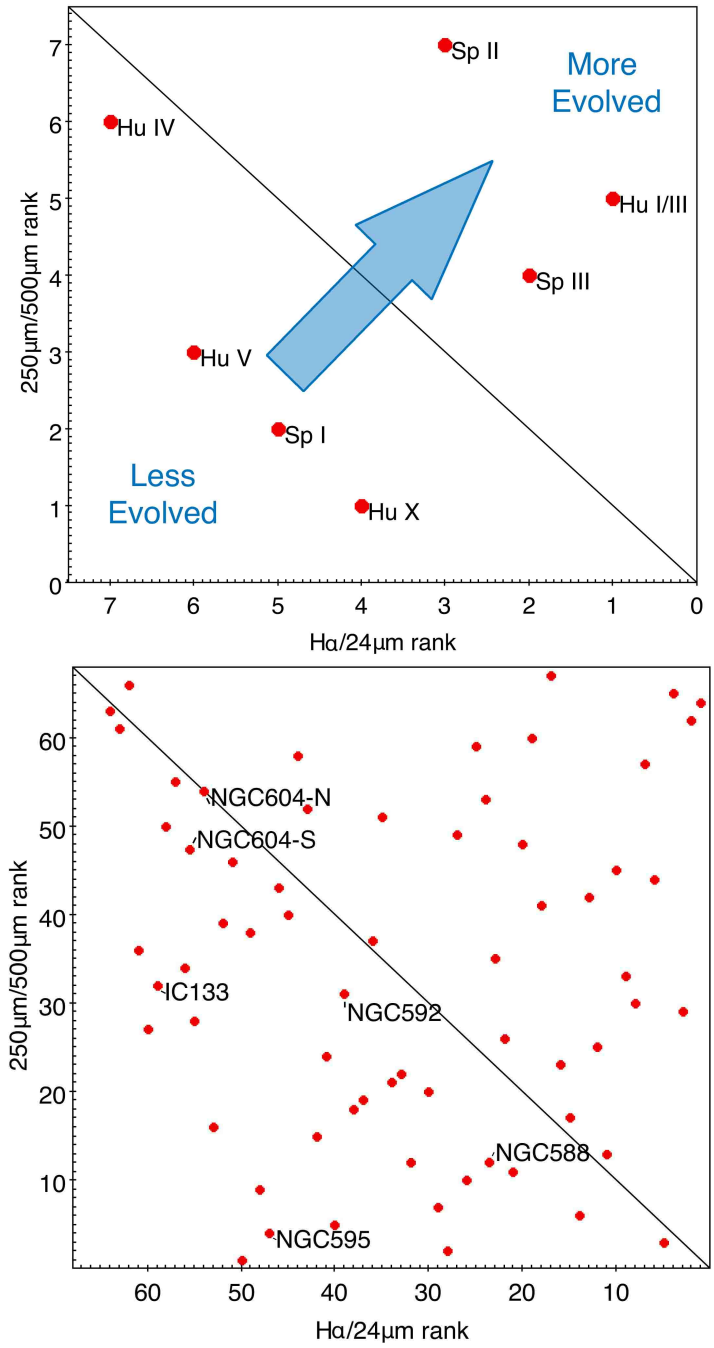

Figure 5.20: SFRs in NGC 6822 (upper) and M 33 (lower) shown by their relative ranks in the [Hα]/[24μm] and [250μm]/[500μm] ratios. The diagonal line indicates the locus of equal rank in both ratios. In the top panel the direction of SFR evolution is indicated by the arrow; significant SFRs are labelled (see text for more detail).

the normalised evolution score reflects a region in which the ISM is being cleared by bright young massive stars and neutral gas is ionising forming H II regions, allowing shorter-wavelength photons to freely propagate.

To characterise star formation activity across the disk of M 33 I investigate the relation between galactic location and evolution score. In Fig. 5.21 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 arms I-N and I-S; the least evolved SFRs mainly lie immediately surrounding the central region of the galaxy. In Fig. 5.22 I 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); outliers to this trend are IC 133 in arm V-N and NGC 588 which are discussed further in Sect. 5.7.4.

I compare the relation between the number of YSOs in a SFR to its evolution score in both M 33 and NGC 6822 in Fig. 5.23. 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 I used a 2-Dimensional KS test (Peacock, 1983; Fasano & Franceschini, 1987). I 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]/[24μm] ratio (Fig. 5.19) suggests that larger regions should correlate to higher evolution scores, this is in fact not seen in Fig. 5.23, with the exception of the very largest regions ($n_{\rm YSOs} \geqslant 50$); as discussed in Sect. 5.7.1 these regions likely result from the combination of multiple smaller SFRs.

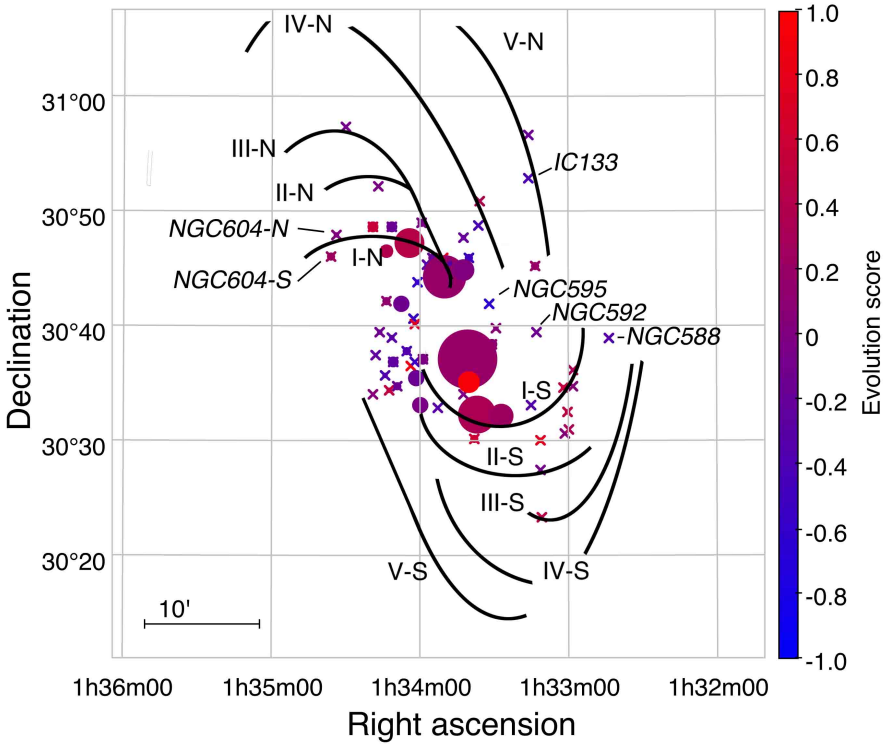

Figure 5.21: Galactic location of SFRs in M 33 shown with a schematic labelled spiral structure. Symbol size is proportional to the number of YSOs, colour shows the evolution score (the smallest regions are marked with a cross). The least evolved regions (purple hues) ring the centre of the galaxy with more evolved regions (red hues) located further out in the disk (see also Fig. 5.22). SFRs discussed in Sect. 5.7.4 are labelled.

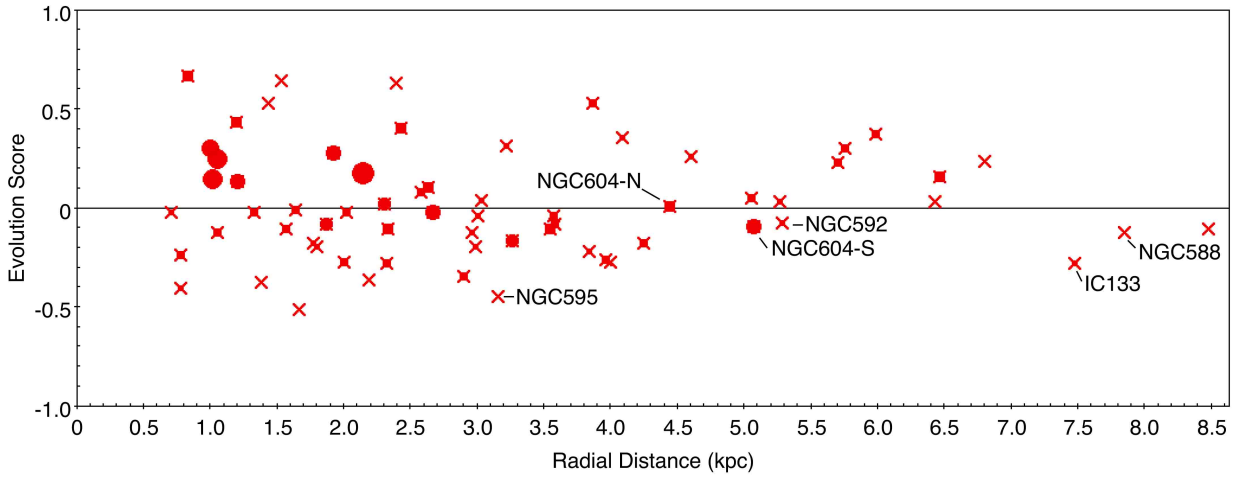

Figure 5.22: Normalised evolution score against radial distance for SFRs in M 33. Symbol size is proportional to the radius of each cluster. Counterparts to regions of star formation known in literature are labelled.

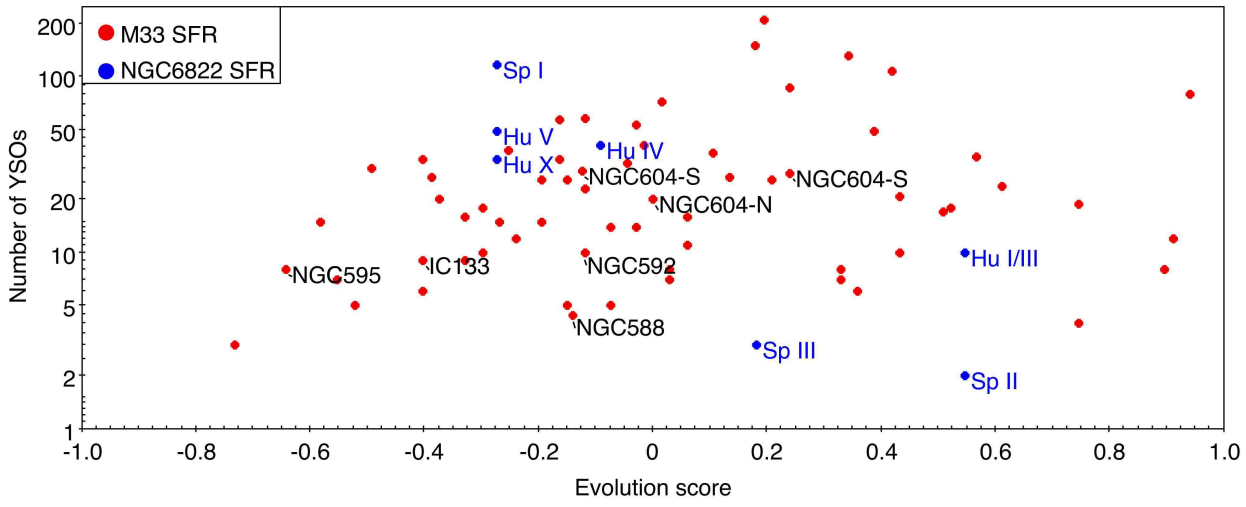

Figure 5.23: Number of YSOs against normalised evolution scores for SFRs in M 33 and NGC 6822. The number of YSOs for SFRs in M 33 and NGC 6822 are respectively from this analysis and from Kinson et al. (2021). There seems to be a slight tendency ($r_{\mathrm{pearson}} \sim 0.21$) for larger SFRs to appear more evolved in M 33.

5.7.3 SFRs in the context of GMCs

I checked the positions of the 68 SFRs identified in this 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. I find 17 type A, 16 type B, and 54 type C GMCs that have a positional overlap with 62 out of 68 SFRs ($\sim$ 91 per cent), using the SFR median radii provided in Table 5.5 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 et al., 2018) is required for a type C classification, most SFRs are indeed matched to this GMC type; furthermore as discussed in Sect. 5.8 my 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. I 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 several well studied GMCs along its extension as well as the prominent H II region NGC 604. I 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 figure 1 of Kondo et al. 2021), which correspond to SFRs 25 and 35. These regions in I-N were studied in detail with recent ALMA observations (Tokuda et al., 2020; Muraoka et al., 2020; Kondo et al., 2021). 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. NGC 604 is recovered in this analysis as two SFRs discussed further in Sect. 5.7.4.

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 I 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, see also Sect. 1.3.1). 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).

5.7.4 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 et al., 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).

Using GEMINI-NIRI photometry with excellent seeing conditions ($\sim$ 0.35 arcsec), Fariña et al. (2012) identified 68 massive YSOs in the central region of NGC 604 (see left panel of Fig. 5.24). Whilst all the YSOs identified by Fariña et al. (2012) are brighter than the near-IR catalogue sensitivity limits (see Sect. 3.2.1), 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 my analysis in this region of extremely bright ambient emission; the YSOs I identify in this 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. 5.24). The two SFRs (3 and 17 in Table 5.5), which I refer to as NGC 604-N and -S, contain 20 and 28 YSOs respectively. The separation of NGC 604 into two SFRs may be in part driven by the paucity of near-IR data described above. Alternatively 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 the SFRs identified in this project. I 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. I note however that my 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 $\sim 20\,\mathrm{km\,s^{-1}}$; 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 figure 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 et al., 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 eight YSOs are identified, is the second most lumi-

nous 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-μm and 250-μm emission (see Fig. 5.24).

As noted in Sect. 5.7.2 and Fig. 5.22, 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 et al., 1988; Greenhill et al., 1993) and OH maser (Staveley-Smith et al., 1987) emission. I 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 $\sim 0\farcs28$) 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 could 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. I assign NGC 592 an evolution score of −0.08, reflective of some indicators of youth.

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. 5.21. Only 4 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. 5.22).

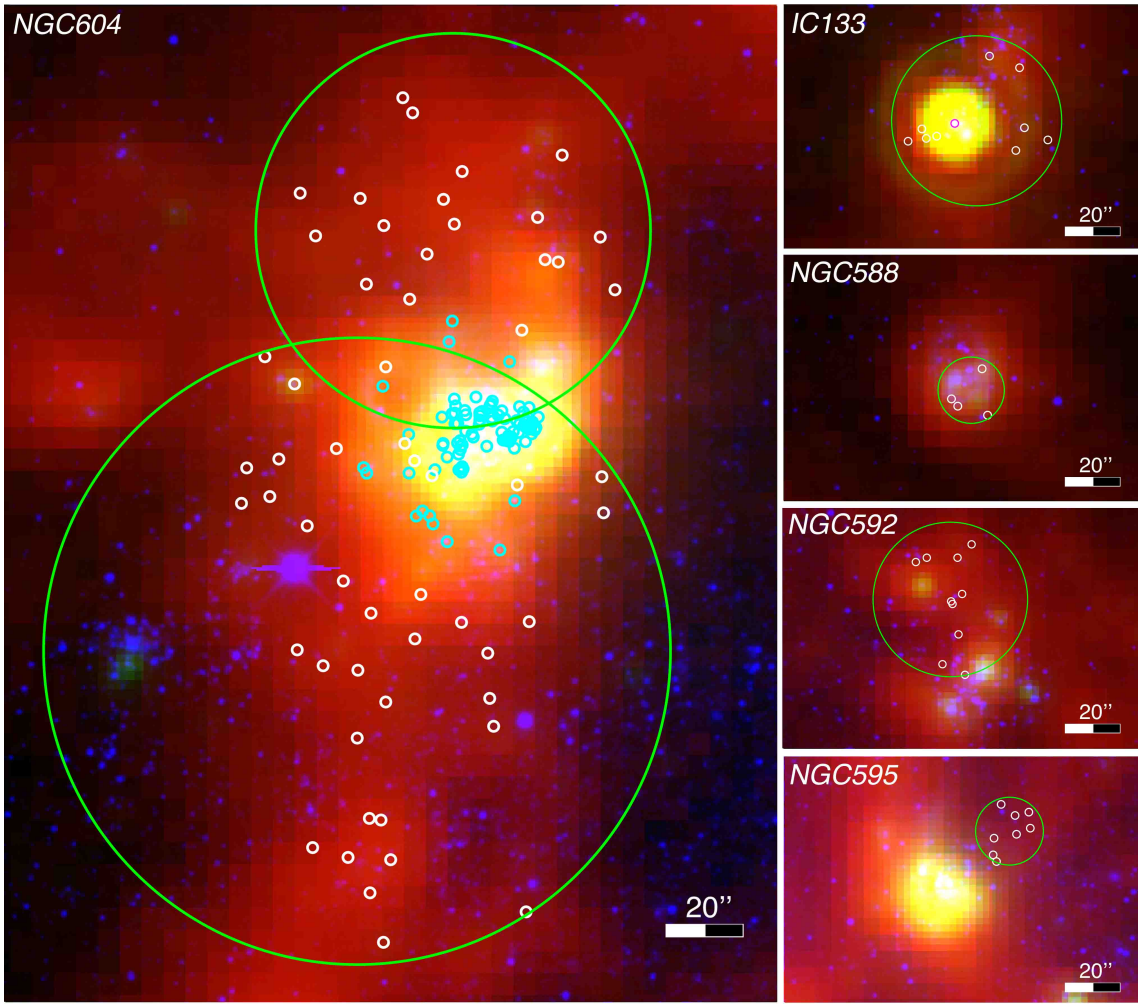

Figure 5.24: RGB image (250μm *Herschel*-SPIRE, 24μm *Spitzer*-MIPS, Hα respectively – see Sect. 3.4 for image details) of NGC 604, IC 133, NGC 588, NGC 592 and NGC 595. YSOs identified in this work are shown by white circles, the extent of each SFR is shown by the green circles, in NGC 604 cyan circles show YSOs identified in Fariña et al. (2012), in IC 133 the magenta circle shows the location of the maser counterpart (see text).

5.8 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 analyses 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 discussed below should be taken with some caution.

Since the YSOs identified in this 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). I 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 I thus obtained a mass estimate as described below. Due to the depth of the near-IR catalogue (see Sect. 3.2.1) my analysis is likely sensitive to only the most massive YSOs. Given these sources evolve rapidly onto the main sequence once they leave their embedded stages, I use only models in the grid corresponding to Stage 0/I YSOs. I 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; I 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. 5.25, with a total mass of $2.5 \times 10^4\,M_\odot$. Using the commonly adopted functional form for the IMF by Kroupa (2002), scaled to match the observed mass distribution, and integrated over the range $0.08 - 100\,M_\odot$, I estimate the total mass of YSOs in SFRs as $1.5 \times 10^5\,M_\odot$. Adopting a Stage 0/I lifetime of 0.2 Myr (e.g. Jones et al., 2019, and references therein) I estimate a star formation rate of $0.63\,M_\odot\,yr^{-1}$ in M 33's SFRs (green line in Fig. 5.25). Due to the effects of crowding, the lower PRF classification certainty, and potential contamination (see Sect. 5.6), I estimate the star formation rate separately for the unclustered YSOs in the central region: it ranges between $0.63\,M_\odot\,yr^{-1}$ in the case of maximum contamination by RGBs (see Sect. 5.6) and $0.95\,M_\odot\,yr^{-1}$ for no such contamination. This range is represented by the grey shaded region in Fig. 5.25. The lower value is consistent with that estimated for the SFRs throughout the disk of M 33. Considering all YSOs, the total star formation rate is $1.42 \pm 0.16\,M_\odot\,yr^{-1}$ (gold shaded region in Fig. 5.25).

There are numerous determinations of global star formation rates in the MW, as compiled in table 1 of Chomiuk & Povich (2011) for a range of methods (ionisation 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 $1.75 \pm 0.25\,M_\odot\,yr^{-1}$ (shown as the red line in Fig. 5.25). 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 ($M_{\rm gas} \sim 3 \times 10^9\,M_\odot$, Corbelli 2003) has a lower star formation rate than the MW ($M_{\rm gas} \sim 5 \times 10^{10}\,M_\odot$, Licquia & Newman 2015) as seen in Fig. 5.25.

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). My 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 $0.7\,M_\odot\,yr^{-1}$. My estimates for the star formation rate of M 33 are broadly consistent with these estimates, towards the upper end as expected from other galaxies.

5.9 M 33 summary

In this work, I identified and described the YSO population across the whole disk of the flocculent spiral galaxy M 33 for the first time. I 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 the PRF was applied 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 individual test runs). A total of 4985 YSOs were identified. The classifier cannot and should not classify outside the training parameter space. The WFCAM near-IR catalogues suffer from completeness issues for faint and very red

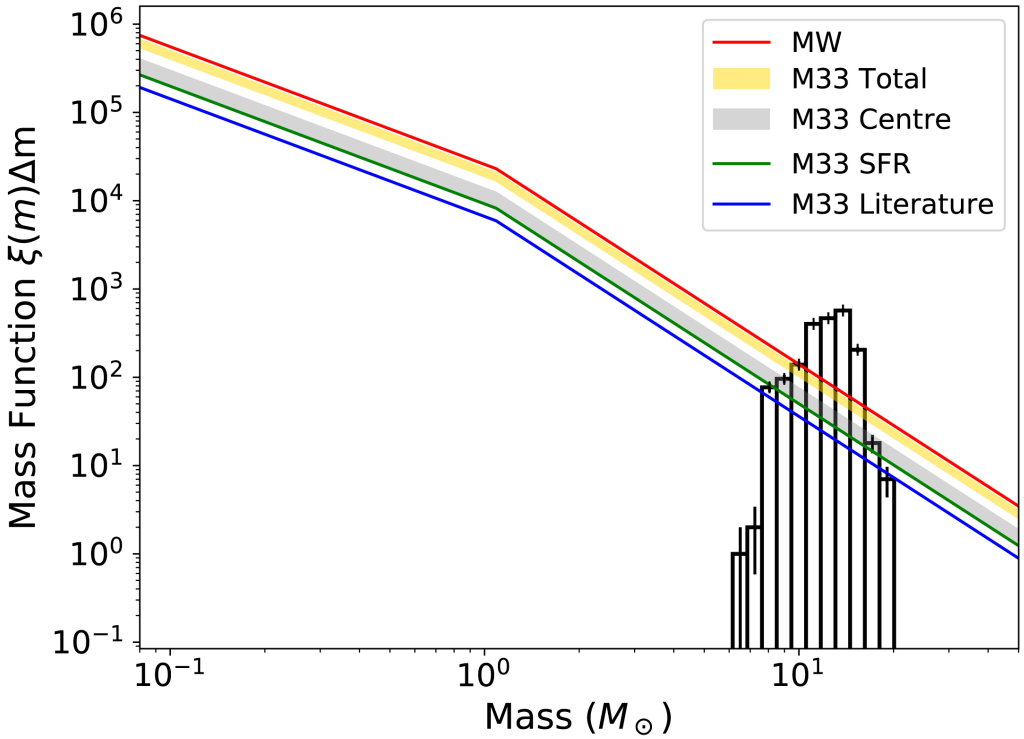

Figure 5.25: The mass distribution of the 1986 YSOs assigned to M 33 SFRs, with scaled Kroupa (2002) IMFs overlain, see text for details. Poisson errors are indicated for each histogram bin.

sources (see Sects. 3.2.1, 3.3 and 4.3.2); the classification of YSOs as red as the reddest training set sources is limited by the availability of such sources in the near-IR catalogue and not by the PRFs performance or any bias of the training set. This is a limitation of the catalogue completeness rather than an introduced classification bias.

A DBSCAN clustering analysis of the YSO population was used to identify 68 SFRs, mostly previously unknown, across the disk of M 33, containing 1986 YSOs. Most of these SFRs are located in the spiral arms. 2437 YSOs are found in the central $\sim 11.6 \times 10.4\,\mathrm{arcmin}^2$ region, that is too crowded for the clustering algorithm to be effective. The remainder 562 YSOs are seemingly isolated based on my analysis.

In total 62 out of the 68 identified 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. I 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]/[24μm] and [250μm]/[500μm] ratios was used to constrain the comparative evolutionary status of the M 33 SFRs, using regions in NGC 6822 as a benchmark sample. These ratios was converted into a common metric for ease of comparison. This evolution score was 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. The SFRs with the lowest evolution scores tend to lie just outside the central region of the galaxy. With the exception of well known SFRs (Sect. 5.7.4) those beyond ~ 4.5 kpc present as more evolved.

I 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 some evidence is seen in the 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.

Using model grids for Stage 0/I YSOs (Robitaille et al., 2006) I estimated 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, I 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 I 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).

I have for the first time identified massive YSOs and discussed their spatial distribution on galactic scales in a Local Group spiral galaxy, extending such analysis beyond the nearby star-forming dwarf galaxies (LMC, SMC and NGC 6822).

6 Conclusions, summary and prospects

In this thesis, a machine learning technique for the classification of point sources in archival near-IR data has been developed and applied in two Local Group galaxies, NGC 6822 and M 33 to identify massive YSOs. The method was validated in NGC 6822 before the technique was then adapted for application across M 33, classifying YSOs across the entire disk of a spiral galaxy for the first time.

In this final chapter I will discuss the methods developed in and results of this thesis in the context of the aims stated in Sect. 1.5. This is followed by a view to further studies which may soon be possible using the techniques and results of this project.

6.1 Machine learning techniques

Previous work to identify YSOs in photometric IR data has relied on a series of colour and magnitude cuts (e.g. Jones et al., 2019; Hirschauer et al., 2020). This approach considers colours and magnitudes in a piece-wise fashion, and assesses sources based on their positions relative to established cuts. This does not however take into account interdependancies and degeneracies between observable properties. Using machine learning it is possible to classify sources based on an ensemble of all colour and magnitude and measurement uncertainties.

A combination of 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, were used as features for both supervised and unsupervised machine learning methods. Unsupervised t-SNE maps were used to establish the need for all features included. Near-IR colours and magnitudes provide information on individual point source properties whilst far-IR measurements are used to contextualise each source. CO measurements were tested as a potential feature to separate YSOs (Sect. 4.8.2); whilst integrated CO brightness does appear to differentiate YSOs from other classes (Fig. 4.19), it was not adopted as a feature for two reasons. Firstly, peaks of CO brightness are co-spatial with the

brightest regions of the far-IR features in both galaxies to a high degree, secondly the spatial coverage of CO data for each galaxy is worse than for both far-IR wavelengths.

Supervised machine learning techniques such as those used in this project rely upon a set of data of known types on which the classifier is trained. The quality of these data is therefore of paramount importance to obtain good classifier performance. The ideal scenario would be to have a series of spectroscopically confirmed sources for each class within the same galaxy, however in practice training sources for certain stellar types were not available in each target galaxy (e.g. YSOs). In these cases training sources are drawn from the MC in which spectroscopically confirmed examples are available. Whilst these sources originate from environments of similar metallicity (see Table. 1.1) and were shown to be comparable with those in NGC 6822 and M 33 (Sects. 4.5 and 5.4), using training sources external to the galaxy being studied is a recognised limitation imposed by the available data.

Many supervised machine learning techniques are available; a series of tests were performed using a reduced number of classes and data from NGC 6822 to classify both with a PRF and classical RF (Sect. 2.2.3). An improvement was seen by using a PRF rather than an RF in classes with fewer available training sources such as YSOs, which rose from 89 per cent to 92 per cent. A benefit afforded by a PRF, not possible with an RF, is the ability to handle missing feature information. The near-IR catalogue for NGC 6822 contains several sources with one missing feature due to the way in which observations were performed and the near-IR catalogue constructed; therefore in NGC 6822 a PRF increases the number of sources available for classification from ~ 8000 to $\sim 12{,}000$ (see further details of the data in Chap. 3). Another key benefit of a PRF over a RF is that it deals with feature uncertainties therefore I selected a PRF over a classical RF as the primary classifier for this project.

In M 33 the PRF was used to classify 162,746 sources of which 66,378 are consistently assigned to the same class across a total of 100 PRF runs, with a median estimated accuracy of 86 per cent across all classes. For YSOs the estimated accuracy ranged from 62 to 97 per cent, with a median value of 82 per cent across all 100 runs. This wider range of accuracies is affected by a small number of runs (< 3) in which

classifier performance is affected by stochastic effects in training/test source collection. Using a greater number of PRF runs allowed me to implement multiple down samplings of large training set classes to mitigate the effect of large classes dominating the training data. Using multiple different down samplings also accounts for any additional stochasticity in training and test data selection with varying training sets. Train-test splitting is done on a global 75/25 per cent split rather than class-wise basis, leading to some unevenness in testing data class sizes. In M 33 additional steps were necessary to account for random down-sampling of especially large training set classes (Sect. 5.2). Having training set classes of sizes which are somewhat balanced against one another was seen to greatly improve classifier performance compared to trials in which few large classes dominated the training set (Sect. 5.2).

For every source a value $n_{\rm class}$ for each class is obtained: the number of runs a source is classified into that class. This $n_{\rm class}$ value allows me to assess the confidence for the object to belong to each particular class. As shown in Figs. 4.12, 4.13 and 5.7, target classes overlap with one another in feature parameter space. Correctly separating sources into the target classes is a non-trivial exercise precisely because of these intrinsic class overlaps. In these regions of overlapping classes it is likely that the PRF returned mixed classifications for many sources (i.e. $n_{\rm class}$ is split amongst multiple classes). By setting restrictions on $n_{\rm class}$ for inclusion in the subsequent astrophysics analysis cleanness of the identified samples can be achieved, with mixed $n_{\rm class}$ sources being treated as candidates. Running the classifier multiple times mimics a k-fold cross-validation approach without the removal of one feature per run (see Sect. 4.3).

To test the PRF classifications as well as the feature selection I used unsupervised t-SNE maps on sources with maximum $n_{\rm class}$ scores in NGC 6822 (see Sect. 4.7). Similar comparison of the PRF's classification in M 33 was not possible due to the much greater number of sources, making t-SNE calculation prohibitively computationally expensive. Nevertheless, in NGC 6822 t-SNE maps show a good separation of several target classes including Galactic foreground, RSG and both O- and C- AGB stars, but it is not able to identify YSOs as a distinct grouping.

6.2 YSO Identification

The PRF classifier sorts sources into eight target classes (YSO, OAGB, CAGB, RGB, RSG, MMS, FG and AGN) in NGC 6822 with an additional class (WR) included for M 33. In M 33 the BS class effectively replicates the MMS class for NGC 6822 but further includes other stellar types such as luminous blue variables. The PRF must assign a classification to each source, hence having multiple non-YSO classes decreases the probability of a false YSO classification and allows for comparison of spatial distributions for each class.

The PRF classifier was first applied in NGC 6822 in which massive YSOs have previously been identified in major SFRs. The PRF achieved an estimated accuracy of 91 per cent across all classes rising to 96 per cent for YSOs. A total of 324 YSOs ($n_{\rm class} = 20$) and candidates ($n_{\rm class} < 20$) are classified (Kinson et al., 2021). I confirm the nature of 125 out of 277 literature YSO candidates with sufficient feature information. Additionally 136 YSOs and 63 YSO candidates are classified for the first time in my analysis.

The spatial distributions of most stellar populations are essentially as expected, older populations are more dispersed with young populations more tightly structured (Sect. 4.6). RSG stars, that trace the recent star formation history, occupy the bar of NGC 6822, linking the more conspicuous SFRs. An extension of the bar to the South-East, into a region which has indicators of youth (e.g. De Blok & Walter, 2000) is seen in the RSG distribution, however no YSOs or candidates are classified there.

The identification of YSOs was used as a method of locating major regions of star formation. In NGC 6822 the recovery of known SFR provided confidence in the machine learning methodology. YSOs were identified in all known major star formation complexes in NGC 6822 (Hubble I/III, Hubble IV, Hubble V, Spitzer I, Spitzer II and Hubble X). Furthermore YSOs in smaller star formation sites were identified: the HII regions BHD 9/10 and 27, as well as new regions of star formation BHD 18 and a region to the North but physically distinct from Hubble IV, that I named Hubble IV−N. The detection of massive YSOs in new regions, especially in BHD 18 and Hubble IV−N, is

very suggestive of additional star formation occurring in the bar of NGC 6822 between the major previously known SFRs. The prospect of detecting further YSOs in the bar region below the mass sensitivity of my analysis ($M \sim 15\,M_\odot$) remains to be explored. A region of over-dense H I gas to the North-West of the bar (studied in Schruba et al., 2017), where star formation is possible, is beyond the coverage of the far-IR data and hence not investigated in this project.

I identified and described the YSO population across the whole disk of the flocculent spiral galaxy M 33 for the first time. The PRF classification technique which was successfully applied in NGC 6822 (Kinson et al., 2021) was adapted to better reflect the stellar populations in M 33. In M 33 the requirement was set that sources achieve $n_{\rm class} = 100$ in order to be carried forwards into the astrophysical analysis. This criterion was set more strictly than in NGC 6822 as there were no literature benchmarks against which YSO candidates could be assessed. A total of 4985 YSOs in M 33 were identified (Kinson et al., 2022).

Using model grids for Stage 0/I YSOs (Robitaille et al., 2006) I estimated the mass of each of the M 33 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 M 33 YSOs assigned to SFRs is $2.5 \times 10^4\,M_\odot$. Using a Stage 0/I lifetime of 0.2 Myr, I estimate a star formation rate of $0.63\,M_\odot\,yr^{-1}$ for the M 33 spiral arm SFRs. In the central region of M 33 I 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 (Sect. 5.4). These estimates give a total M 33 star formation rate of $1.42 \pm 0.16\,M_\odot\,yr^{-1}$ determined for the first time 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).

A DBSCAN clustering analysis of the M 33 YSO population was used to identify 68 SFRs, mostly previously unknown, across the disk of M 33, containing 1986 YSOs. The majority of these SFRs are located in the spiral arms. 2437 YSOs are found in the

central $\sim 11.6 \times 10.4\,\mathrm{arcmin}^2$ 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 68 identified 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. SFR counterparts are identified to the prominent H II regions IC 133, NGC 588, NGC 592, NGC 595 and NGC 604.

A novel approach combining [Hα]/[24μm] and [250μm]/[500μm] ratios was used to constrain the comparative evolutionary status of the M 33 SFRs, using major regions in NGC 6822 for which a literature evolutionary sequence exists as a benchmark sample. These ratios were converted into a common metric for ease of comparison (Sect. 5.7.2). This evolution score was 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 on a galaxy-wide scale. The largest SFRs, which lie at the base of the two primary spiral arms, are generally more evolved but do not have very large evolution scores; this may be an effect of merging multiple smaller regions with an averaging out of properties occurring. The least evolved SFRs mainly lie immediately outside the central region of the galaxy. At radii larger than ~ 4.5 kpc most SFRs appear more evolved with outliers to this tending to be well known SFRs (Sect. 5.7.2). This metric presents a useful strategy for comparing evolutionary stages of SFR within a galaxy, in isolated galaxies these scores could inform star formation theories such as SPSSF (Sect. 1.3).

6.3 Future studies

This project presents non-spatially biased surveys of the most massive YSOs across both NGC 6822 and M 33. Given the shape of the IMF (e.g. Kroupa, 2002) many more YSOs remain in these target galaxies to be identified below the sensitivity limits of the data used in this work.The next generation of near-IR capable observatories such as the *James Webb Space Telescope* (hereafter *JWST*), Extremely Large Telescope (formerly European-Extremely Large Telescope) and *Roman Space Telescope* (formerly known

as *WFIRST*) will transform the volume, sensitivity and resolution of available data. As noted in Sects. 4.8.2 and 5.8 the point sources classified in this thesis are likely the dominant source in an unresolved proto-cluster and future observations may be able to resolve the less massive and hence less luminous members of these clusters. The new generation of infrared instrumentation will aid in overcoming the issues of source crowding at the limits of WFCAM's resolution encountered in this project's study of the centre of M 33. *JWST* observations of the centre and arm I-S in M 33 are proposed (Lee et al., 2021; Rosolowsky et al., 2021), covering $\sim$2500 of the YSOs and $\sim$20 SFRs I identify. Bespoke observations aiming to identify fainter, less massive YSOs rather than catalogues resulting from an evolved star monitoring programme will mitigate the issues of sensitivity in the M 33 near-IR catalogues (Sect. 3.2.1). Observations extending to fainter magnitudes will also allow the identification of PN which were excluded from the analysis in this project as they lie below the sensitivity of the WFCAM near-IR data (see Sects. 4.2.7 and 5.1).

In order to survey YSOs across the entire disk of a spiral galaxy it was necessary to look outside the MW, with M 33 presenting the ideal first target due to its proximity and favourable inclination. The next generation of near-IR instrumentation will extend the range of galaxies in which similar study as in this project can be performed. These more distant galaxies include spirals of both flocculant, like M 33, and grand-design types. Two face on spiral galaxies in which the preliminary works for such study have been conducted are the multi-armed NGC 300 ($\sim$1.9 Mpc, Rizzi et al., 2006) and grand-design M 83 ($\sim$4.5 Mpc, Tully et al., 2016). *JWST*'s NIRCam instrument (Rieke et al., 2005) will be able to produce observations of both these galaxies with similar resolution to that of WFCAM in M 33. In NGC 300 a comparison of the relation of CO(1-0) and GMCs has been performed (Kruijssen et al., 2019), and sites of potential star formation identified (Semenov et al., 2021). *JWST* NIRCam observations across the disk of NGC 300 (Lee et al., 2021) and surveying for molecular gas in M 83 (Hernandez et al., 2021a) await implementation. Using this new instrumentation to locate YSOs and GMCs in relation to spiral structure in these galaxies as well as in M 33 will provide insight into models of QSS and DSF (see Sect. 1.3.1).

Whilst the Extreme Large Telescope's MICADO instrument (Davies et al., 2010) will be able to produce JHK_s observations, filter selection for the *Roman Space Telescope's* near-IR instrumentation has not yet been finalised (e.g. inclusion of a K_s filter Chary, 2021) and *JWST* uses an alternative filter system (Rieke et al., 2005, 2015). Therefore to conduct a study similar to this project using these instruments a training set in the corresponding filter set is necessary. Given that it may take some time for such observations to be conducted, especially when the increasing number of object types able to be included in classification is considered, alternative methods for training set construction may need to be considered. One approach to this could be using an off target field including SFRs to obtain training set sources of many classes which can then be used as a training set for wider classification. Finally it is important to consider inclusion of feature information from *JWST's* Mid-IR Instrument (MIRI, Rieke et al., 2015) which increases the wavelength range accessible at high resolution and offers a potential for greater discrimination power between stellar types.

Machine learning approaches, as I have demonstrated, offer an invaluable tool for disentangling and classifying large data sets. The analysis of large data sets is becoming an important factor in many settings outside of astronomy. Classification problems are prevalent in almost all data-driven fields, with the application of e.g. PRF or t-SNE maps offering valuable tools for such analysis. Whilst for very large data sets t-SNE mapping and other similar algorithms become prohibitively computationally expensive, for data sets on the order of 10^4 sources and ten or so features this is achievable on most current personal computers. As previously noted the performance of a supervised classifier is strongly linked to the quality of the training data, however labelled data is readily available in most potential academic or industrial applications and should not present a significant limitation to the implementation of a PRF.

Publications

Refereed

- Kinson D. A., Oliveira J. M., van Loon J. T., 2021, MNRAS, 507, 5106
- Kinson D. A., Oliveira J. M., van Loon J. T., 2022, MNRAS, 517, 140

A NGC 6822 Confusion Matrices

The full set of confusion matrices for all extended PRF runs are shown here. A representative example from a single PRF run for both normalised and un-normalised matrices was shown in Fig. 4.10. The accuracy scores returned by SKLEARN for these runs vary between 87 and 92 per cent with an average of 91 per cent.

By comparing between runs with different random seeds in the un-normalised confusion matrices, the variations arising from the random selection of training and test samples can be seen, e.g. in their raw numerical values (Fig. A.3). The strong diagonal seen in the normalised confusion matrices (Fig. A.6) is weaker in the un-normalised matrices (Fig. A.3) as a result of the different sizes of each target class. These un-normalised matrices do however show how many sources of each class are included in the testing data for each PRF run.

The normalised matrices allow for a better assessment of the success of the classifications, by evening out different class sizes. Fig. A.6 shows these normalised confusion matrices. A good recovery rate can be seen in all classes with the exception of RGB stars which have a significant level of confusion with Galactic contaminants due to how the FG training set is constructed (see Sect.4.4.2). The AGN class suffers from confusion with CAGB and FG classes in many runs. This is due in part to their similarities in near-IR colour, but also because of the limited number of available AGN training sources for a class with a large range of possible parameter space, as discussed in Sect. 4.3.1.

Figure A.1: Normalised confusion matrices for the 20 PRF runs using different random seeds to overcome any stochastic effects in train/test splitting.

Figure A.2: Cont.

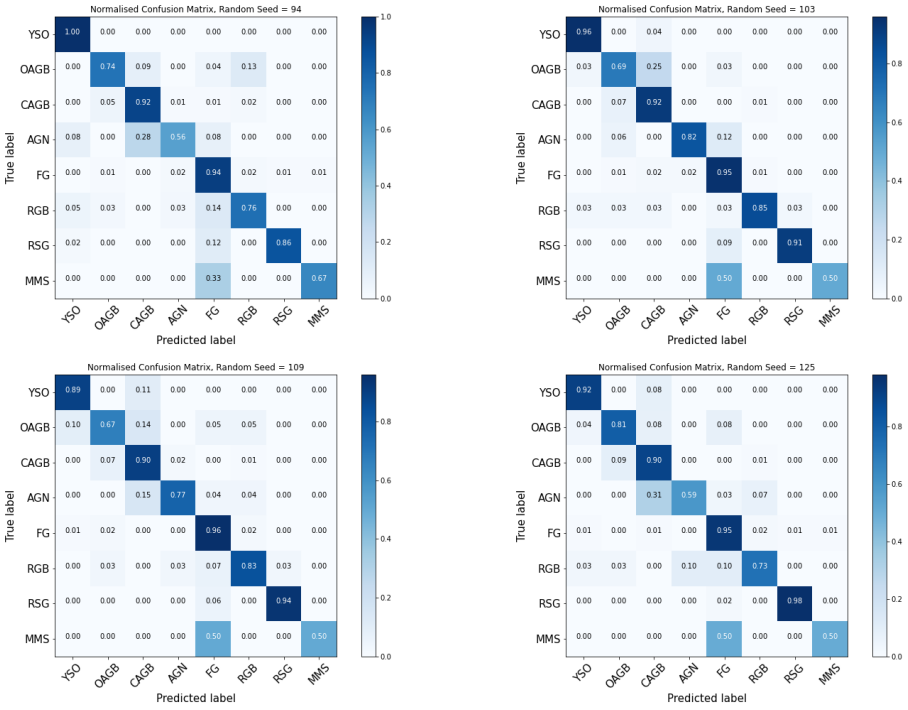

Figure A.3: Cont.

B Coordinate de-projection

In Sect. 5.6 I apply a DBSCAN clustering algorithm to the distribution of classified YSOs in deprojected coordinates. In order to transform from coordinates on the sky (α, δ) to a deprojected plane (x, y) I adapt standard spherical geometry techniques (as described e.g. Van der Marel & Cioni, 2001).

In what follows I adopt as the centre of M 33 the position $\alpha_0 = 01{:}33{:}50.89$, $\delta_0 = +30{:}39{:}36.63$ (Gaia Collaboration, 2020). For a point on the sky with right ascension and declination (α, δ) I define the angular coordinates with respect to the centre (α_0, δ_0) as (ρ, ϕ): ρ is the angular distance and ϕ is the position angle. Standard

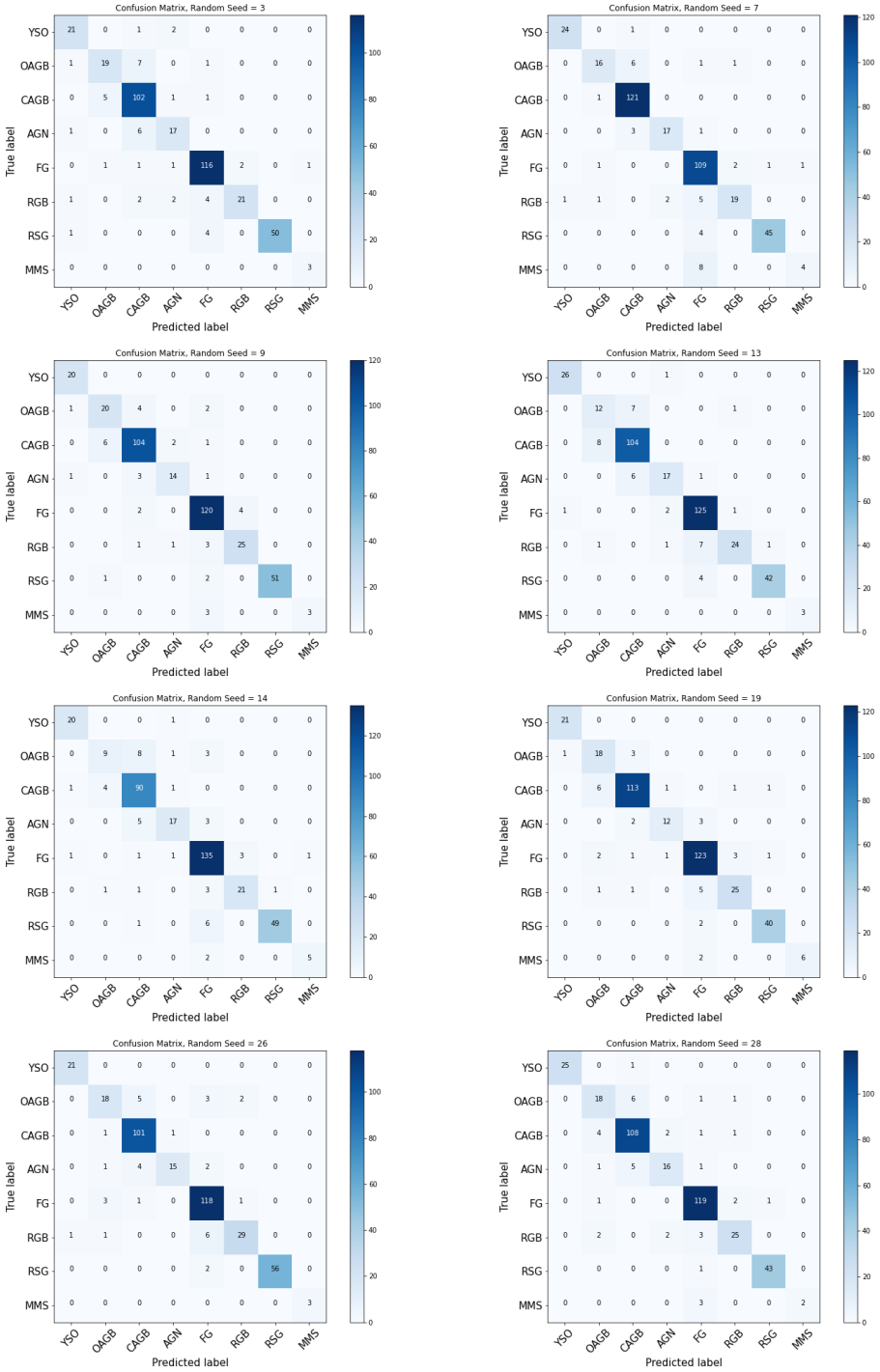

Figure A.4: Confusion matrices of the same runs shown in Fig.A.3.

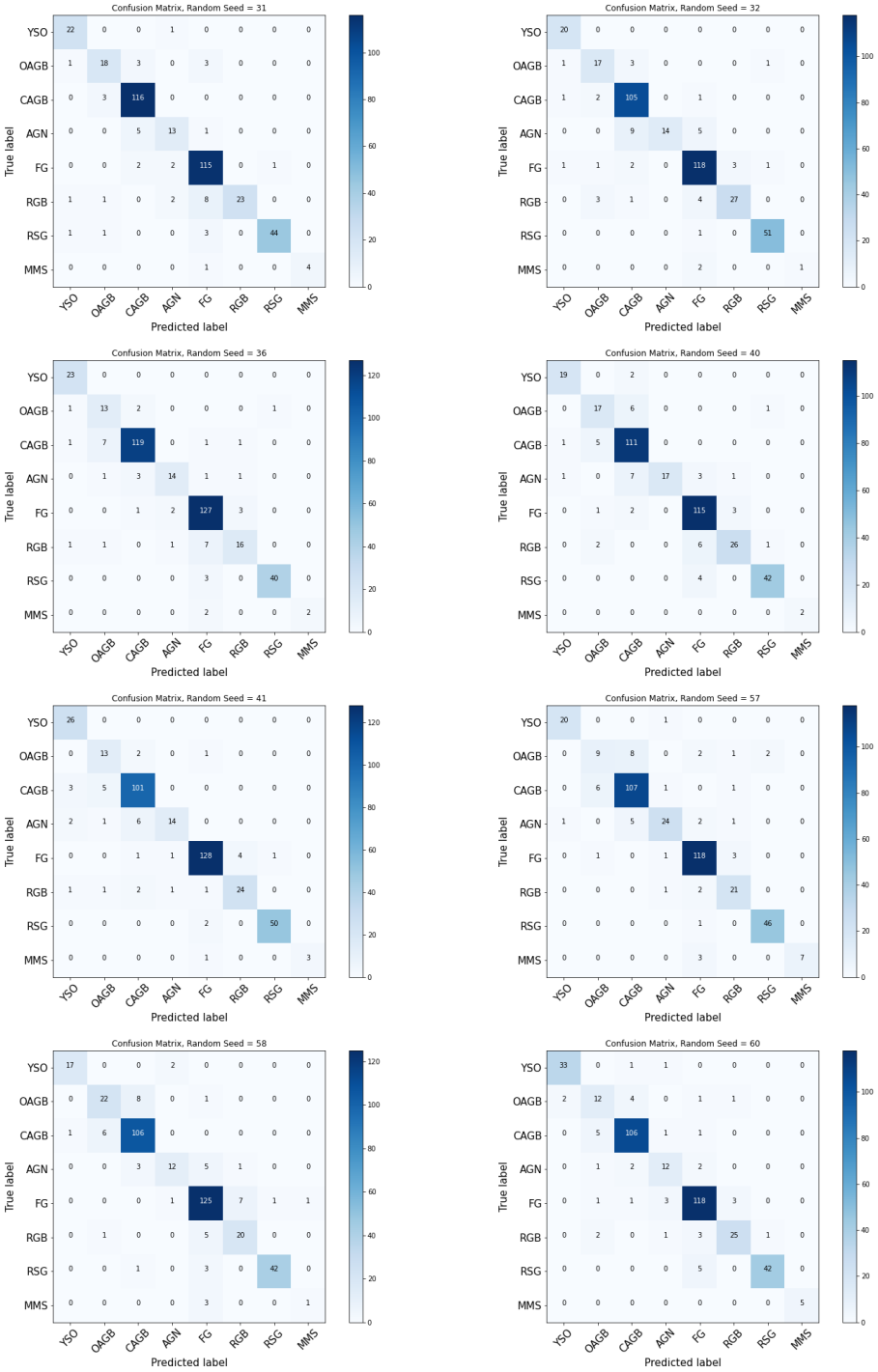

Figure A.5: Cont.

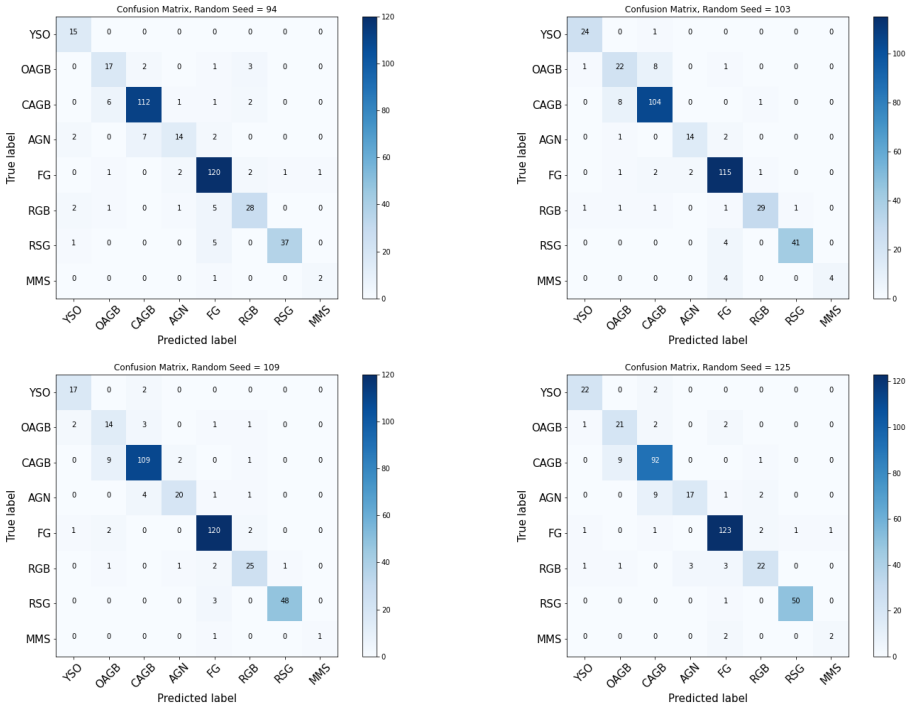

Figure A.6: Cont.

spherical trigonometry transformations lead to three equations:

$$\begin{aligned}
\cos\rho &= \cos\delta\,\cos\delta_0\,\cos(\alpha - \alpha_0) + \sin\delta\,\sin\delta_0 \\
\sin\rho\,\cos\phi &= -\cos\delta\,\sin(\alpha - \alpha_0) \\
\sin\rho\,\sin\phi &= \sin\delta\,\cos\delta_0 - \cos\delta\,\sin\delta_0\,\cos(\alpha - \alpha_0).
\end{aligned}$$

A Cartesian coordinate system $(x,\ y)$ is defined such that the x-axis is anti-parallel to the α-axis and y-axis is parallel to the δ-axis. For this purpose the thickness of the disk is ignored, i.e. $z = 0$. The transformations used are:

$$\begin{aligned}
x &= D\,\sin\rho\,\cos\phi \\
y &= D\,\sin\rho\,\sin\phi,
\end{aligned}$$

where D is the distance to M 33 (850 kpc, De Grijs & Bono, 2014). To correct

for M 33's position angle PA the (x, y) coordinates are rotated as follows:

$$x' = x \cos\theta + y \sin\theta$$
$$y' = -x \sin\theta + y \cos\theta.$$

The angle θ is defined with respect to the x-axis, while the PA is defined with respect to the y-axis. Therefore $\theta = PA + 90°$ (adopting $PA = 23°$, De Vaucouleurs et al. 1991).

The inclination i of M 33 disk with respect to the line of sight of the observer affects only the y'-axis. Therefore,

$$y'' = \frac{y'}{\cos i}.$$

Per Sect. 1.4.2 in this project I adopt $i = 54°$ (De Vaucouleurs et al., 1991). The Cartesian coordinate system (x', y'') is the deprojected coordinate system that is used in Sect. 5.6.

Bibliography

Alexeeva S., Zhao G., 2022, ApJ, 925, 76

Andre P., Ward-Thompson D., Barsony M., 1993, ApJ, 406, 122

Baba J., 2015, MNRAS, 454, 2954

Baba J., Morokuma-Matsui K., Egusa F., 2015, PASJ, 67, L4

Baba J., Morokuma-Matsui K., Miyamoto Y., Egusa F., Kuno N., 2016, MNRAS, 460, 2472

Bally J., 2008, in Reipurth B., ed., , Vol. 4, Handbook of Star Forming Regions, Volume I. 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

Bate M. R., 2012, MNRAS, 419, 3115

Bekki K., 2008, MNRAS, 390, L24

Bertin G., Lin C. C., 1996, Spiral structure in galaxies a density wave theory

Bertin G., Lin C. C., Lowe S. A., Thurstans R. P., 1989, ApJ, 338, 78

Besla G., 2015a, The Orbits of the Magellanic Clouds. Springer International Publishing, doi:10.1007/978-3-319-10614-4˙26

Besla G., 2015b, arXiv e-prints, p. arXiv:1511.03346

Beuther H., Churchwell E. B., McKee C. F., Tan J. C., 2007, in Reipurth B., Jewitt D., Keil K., eds, Protostars and Planets V. p. 165 (arXiv:astro-ph/0602012)

Bhandare A., Kuiper R., Henning T., Fendt C., Marleau G.-D., Kölligan A., 2018, A&A, 618, A95

Bianchi L., Scuderi S., Massey P., Romaniello M., 2001, AJ, 121, 2020

Bianchi L., Efremova B., Hodge P., Massey P., Olsen K. A. G., 2012, AJ, 143, 74

Blitz L., Rosolowsky E., 2006, ApJ, 650, 933

Blitz L., Fukui Y., Kawamura A., Leroy A., Mizuno N., Rosolowsky E., 2007, in Reipurth B., Jewitt D., Keil K., eds, Protostars and Planets V. p. 81 (arXiv:astro-ph/0602600)

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

Blum R. D., Barbosa C. L., Damineli A., Conti P. S., Ridgway S., 2004, ApJ, 617, 1167

Bonnell I. A., Bate M. R., 2006, MNRAS, 370, 488

Bonnell I. A., Bate M. R., Clarke C. J., Pringle J. E., 2001, MNRAS, 323, 785

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, doi:10.5281/zenodo.4044744, https://doi.org/10.5281/zenodo.4044744

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

Bresolin F., Stasińska G., Vílchez J. M., Simon J. D., Rosolowsky E., 2010, MNRAS, 404, 1679

Britavskiy N., et al., 2019, A&A, 624, A128

Brunthaler A., Henkel C., de Blok W. J. G., Reid M. J., Greenhill L. J., Falcke H., 2006, A&A, 457, 109

Burkhart B., 2018, ApJ, 863, 118

Burnham S. W., 1890, MNRAS, 51, 94

Cannon J. M., et al., 2006, ApJ, 652, 1170

Cannon J. M., et al., 2012, ApJ, 747, 122

Caratti o Garatti A., et al., 2017, Nature Physics, 13, 276

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

Casey C. M., Narayanan D., Cooray A., 2014, PhR, 541, 45

Castelli F., Kurucz R. L., 2003, in Piskunov N., Weiss W. W., Gray D. F., eds, Astronomical Society of the Pacific Vol. 210, Modelling of Stellar Atmospheres. p. A20 (arXiv:astro-ph/0405087)

Chandar R., Bianchi L., Ford H. C., 2000, AJ, 120, 3088

Chary R., 2021, in American Astronomical Society Meeting Abstracts. p. 416.02

Chen C. H. R., et al., 2010b, ApJ, 721, 1206

Chen C.-H. R., et al., 2010a, The Astrophysical Journal, 721, 1206

Chevance M., et al., 2020, MNRAS, 493, 2872

Chevance M., Krumholz M. R., McLeod A. F., Ostriker E. C., Rosolowsky E. W., Sternberg A., 2022, arXiv e-prints, p. arXiv:2203.09570

Chomiuk L., Povich M. S., 2011, AJ, 142, 197

Churchwell E., 2002, in Crowther P., ed., Astronomical Society of the Pacific Conference Series Vol. 267, Hot Star Workshop III: The Earliest Phases of Massive Star Birth. p. 3

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

Cioni M. R. L., Habing H. J., Israel F. P., 2000, A&A, 358, L9

Clark C. J. R., Roman-Duval J. C., Gordon K. D., Bot C., Smith M. W. L., 2021, arXiv e-prints, p. arXiv:2107.14302

Commerçon B., González M., Mignon-Risse R., Hennebelle P., Vaytet N., 2022, A&A, 658, A52

Cooper H. D. B., et al., 2013, MNRAS, 430, 1125

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

Cornu D., Montillaud J., 2020, arXiv e-prints, p. arXiv:2010.01601

Crowther P. A., Schnurr O., Hirschi R., Yusof N., Parker R. J., Goodwin S. P., Kassim H. A., 2010, MNRAS, 408, 731

Crowther P. A., et al., 2016, MNRAS, 458, 624

Crutcher R. M., 2012, ARA&A, 50, 29

Cyganowski C. J., Brogan C. L., Hunter T. R., Churchwell E., 2009, ApJ, 702, 1615

Dantona F., Mazzitelli I., 1985, ApJ, 296, 502

Davies R., et al., 2010, in McLean I. S., Ramsay S. K., Takami H., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 7735, Ground-based and Airborne Instrumentation for Astronomy III. p. 77352A (arXiv:1005.5009), doi:10.1117/12.856379

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 Blok W. J. G., Walter F., 2000, ApJL, 537, L95

De Blok W. J. G., Walter F., 2003, MNRAS, 341, L39

De Buizer J. M., 2003, MNRAS, 341, 277

De Grijs R., Bono G., 2014, AJ, 148, 17

De Grijs R., Bono G., 2015, AJ, 149, 179

De Vaucouleurs G., de Vaucouleurs A., Corwin Herold G. J., Buta R. J., Paturel G., Fouque P., 1991, Third Reference Catalogue of Bright Galaxies

De Villiers H. M., et al., 2015, MNRAS, 449, 119

Delgado-Inglada G., García-Rojas J., Stasińska G., Rechy-García J. S., 2020, MNRAS, 498, 5367

Dib S., Schmeja S., Parker R. J., 2018, MNRAS, 473, 849

Dimaratos A., Cormier D., Bigiel F., Madden S. C., 2015, A&A, 580, A135

Dobbs C., Baba J., 2014, PASA, 31, e035

Dobbs C. L., Bonnell I. A., 2008, MNRAS, 385, 1893

Dobbs C. L., Bonnell I. A., Clark P. C., 2005, MNRAS, 360, 2

Dobbs C. L., Burkert A., Pringle J. E., 2011, MNRAS, 417, 1318

Dobbs C. L., et al., 2014, in Beuther H., Klessen R. S., Dullemond C. P., Henning T., eds, Protostars and Planets VI. p. 3 (arXiv:1312.3223), doi:10.2458/azu˙uapress˙9780816531240-ch001

Dobbs C. L., Bending T. J. R., Pettitt A. R., Buckner A. S. M., Bate M. R., 2022, MNRAS,

Dobrzycki A., Macri L. M., Stanek K. Z., Groot P. J., 2003, AJ, 125, 1330

Drout M. R., Massey P., Meynet G., 2012, ApJ, 750, 97

Druard C., et al., 2014, A&A, 567, A118

Efremova B. V., et al., 2011, ApJ, 730, 88

Egan M. P., Price S. D., Kraemer K. E., 2003, in American Astronomical Society Meeting Abstracts. p. 57.08

Eldridge J. J., Relaño M., 2011, MNRAS, 411, 235

Elmegreen B. G., 2011, ApJ, 737, 10

Elmegreen D. M., et al., 2011, ApJ, 737, 32

Elsender D., Bate M. R., 2021, MNRAS, 508, 5279

Elson E. C., Kam S. Z., Chemin L., Carignan C., Jarrett T. H., 2019, MNRAS, 483, 931

Engargiola G., Plambeck R. L., Rosolowsky E., Blitz L., 2003, ApJS, 149, 343

Engelbracht C. W., MIPS Science Team SINGS Team 2004, in American Astronomical Society Meeting Abstracts #204. p. 33.11

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. 226–231

Evans N. W., Oh S., 2022, MNRAS, 512, 3846

Fariña C., Bosch G. L., Barbá R. H., 2012, AJ, 143, 43

Fasano G., Franceschini A., 1987, MNRAS, 225, 155

Felli M., Palagi F., 1995, MmSAI, 66, 709

Flesch E. W., 2021, MILLIQUAS - Million Quasars Catalog, Version 7.2, p. arXiv:2105.12985

Frank A., et al., 2014, in Beuther H., Klessen R. S., Dullemond C. P., Henning T., eds, Protostars and Planets VI. p. 451 (arXiv:1402.3553), doi:10.2458/azu˙uapress˙9780816531240-ch020

Fukui Y., Tsuge K., Sano H., Bekki K., Yozin C., Tachihara K., Inoue T., 2017, PASJ, 69, L5

Fukui Y., Ohno T., Tsuge K., Sano H., Tachihara K., 2020, arXiv e-prints, p. arXiv:2005.13750

Gaia Collaboration 2020, VizieR Online Data Catalog, p. I/350

Gaia Collaboration Brown A. G. A., Vallenari A., Prusti T., de Bruijne J. H. J., Babusiaux C., Biermann M., 2020, arXiv e-prints, p. arXiv:2012.01533

Galametz M., et al., 2010, A&A, 518, L55

Gallet F., Bouvier J., 2013, A&A, 556, A36

Geha M., et al., 2003, AJ, 125, 1

Gerola H., Seiden P. E., 1978, ApJ, 223, 129

Ginsburg A., Bally J., Goddi C., Plambeck R., Wright M., 2018, ApJ, 860, 119

Girardi L., Groenewegen M. A. T., Hatziminaoglou E., da Costa L., 2005, A&A, 436, 895

Girichidis P., Federrath C., Banerjee R., Klessen R. S., 2012, MNRAS, 420, 613

Gittins D. M., Clarke C. J., 2004, MNRAS, 349, 909

Goldreich P., Kwan J., 1974, ApJ, 189, 441

Goldreich P., Lynden-Bell D., 1965, MNRAS, 130, 125

Gordon K. D., et al., 2011, AJ, 142, 102

Gottesman S. T., Weliachew L., 1977, A&A, 61, 523

Gratier P., Braine J., Rodriguez-Fernandez N. J., Israel F. P., Schuster K. F., Brouillet N., Gardan E., 2010a, A&A, 512, A68

Gratier P., et al., 2010b, A&A, 522, A3

Gray M., 2012, Maser Sources in Astrophysics

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

Gruendl R. A., Chu Y.-H., 2009, ApJS, 184, 172

Gusev A. S., Shimanovskaya E. V., 2019, MNRAS, 488, 3045

Guszejnov D., Markey C., Offner S. S. R., Grudić M. Y., Faucher-Giguère C.-A., Rosen A. L., Hopkins P. F., 2022, MNRAS, 515, 167

Ha T., Li Y., Kounkel M., Xu S., Li H., Zheng Y., 2022, arXiv e-prints, p. arXiv:2205.00012

Heidmann J., 1983, Highlights of Astronomy, 6, 611

Henkel C., Hunt L. K., Izotov Y. I., 2022, Galaxies, 10, 11

Herbig G. H., 1960, ApJS, 4, 337

Hernandez S. S., et al., 2021a, Shining light on the CO-dark H2 gas in the heart of M83, JWST Proposal. Cycle 1, ID. #2219

Hernandez E. J., Srinivasan S., Marshall J., 2021b, in American Astronomical Society Meeting Abstracts. p. 541.06

Hester J. J., Desch S. J., 2005, in Krot A. N., Scott E. R. D., Reipurth B., eds, Astronomical Society of the Pacific Conference Series Vol. 341, Chondrites and the Protoplanetary Disk. p. 107 (`arXiv:astro-ph/0506190`)

Hildebrand R. H., 1983, QJRAS, 24, 267

Hilditch R. W., Howarth I. D., Harries T. J., 2005, MNRAS, 357, 304

Hirschauer A. S., Gray L., Meixner M., Jones O. C., Srinivasan S., Boyer M. L., Sargent B. A., 2020, ApJ, 892, 91

Hoare M. G., Kurtz S. E., Lizano S., Keto E., Hofner P., 2007, in Reipurth B., Jewitt D., Keil K., eds, Protostars and Planets V. p. 181 (`arXiv:astro-ph/0603560`)

Hodge P., Kennicutt R. C. J., Lee M. G., 1988, PASP, 100, 917

Hodgkin S. T., Irwin M. J., Hewett P. C., Warren S. J., 2009, MNRAS, 394, 675

Hollenbach D. J., Tielens A. G. G. M., 1997, ARA&A, 35, 179

Hollenbach D. J., Tielens A. G. G. M., 1999, Reviews of Modern Physics, 71, 173

Holmberg E., 1941, ApJ, 94, 385

Hony S., et al., 2011, A&A, 531, A137

Hotelling H., 1936, Biometrika, 28, 321

Hou L. G., Han J. L., 2014, A&A, 569, A125

Houck J. R., et al., 2004, ApJS, 154, 18

Hubble E. P., 1925, ApJ, 62, 409

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 Earl J. J., Lynds R., 1996, ApJ, 456, 174

Hunter D. A., Elmegreen B. G., Ludka B. C., 2010, AJ, 139, 447

Huxor A. P., Ferguson A. M. N., Veljanoski J., Mackey A. D., Tanvir N. R., 2013, MNRAS, 429, 1039

Ilee J. D., Cyganowski C. J., Brogan C. L., Hunter T. R., Forgan D. H., Haworth T. J., Clarke C. J., Harries T. J., 2018, ApJL, 869, L24

Inoue S., Yoshida N., 2018, MNRAS, 474, 3466

Ireland L. G., Zanni C., Matt S. P., Pantolmos G., 2021, ApJ, 906, 4

Israel F. P., Bontekoe T. R., Kester D. J. M., 1996, A&A, 308, 723

Ivanov V. D., et al., 2016, A&A, 588, A93

Janson M., Bonavita M., Klahr H., Lafrenière D., 2012, ApJ, 745, 4

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

Jeans J. H., 1902, Philosophical Transactions of the Royal Society of London Series A, 199, 1

Jeřábková T., Hasani Zonoozi A., Kroupa P., Beccari G., Yan Z., Vazdekis A., Zhang Z. Y., 2018, A&A, 620, A39

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

Jones O. C., Reiter M., Sanchez-Janssen R., Evans C. J., Robertson C. S., Meixner M., Ochsendorf B., 2022, MNRAS,

Kacharov N., Rejkuba M., Cioni M. R. L., 2012, A&A, 537, A108

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

Kaper L., Ellerbroek L. E., Ochsendorf B. B., Caballero Pouroutidou R. N., 2011, Astronomische Nachrichten, 332, 232

Kashi A., Soker N., 2010, ApJ, 723, 602

Kato D., et al., 2007, PASJ, 59, 615

Kennicutt Robert C. J., 1989, ApJ, 344, 685

Kennicutt R. C., Evans N. J., 2012, ARA&A, 50, 531

Kennicutt R. 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). pp 310–317, doi:10.1109/ICTAI.2007.46

Kim C.-G., Ostriker E. C., 2017, ApJ, 846, 133

Kinman T. D., Green J. R., Mahaffey C. T., 1979, PASP, 91, 749

Kinson D. A., Oliveira J. M., van Loon J. T., 2021, MNRAS, 507, 5106

Kinson D. A., Oliveira J. M., van Loon J. T., 2022, MNRAS, 517, 140

Kirby E. N., Cohen J. G., Guhathakurta P., Cheng L., Bullock J. S., Gallazzi A., 2013, ApJ, 779, 102

Kobulnicky H. A., et al., 2014, ApJS, 213, 34

Kondo H., et al., 2021, ApJ, 912, 66

Könyves V., et al., 2015, A&A, 584, A91

Kozłowski S., Kochanek C. S., Udalski A., 2011, ApJS, 194, 22

Kramer C., et al., 2010, A&A, 518, L67

Kroupa P., 2002, Science, 295, 82

Kruijssen J. M. D., et al., 2019, Nature, 569, 519

Krumholz M. R., 2015a, arXiv e-prints, p. arXiv:1511.03457

Krumholz M. R., 2015b, in Vink J. S., ed., Astrophysics and Space Science Library Vol. 412, Very Massive Stars in the Local Universe. p. 43 (arXiv:1403.3417), doi:10.1007/978-3-319-09596-7˙3

Krumholz M. R., Klein R. I., McKee C. F., 2011, ApJ, 740, 74

Kuiper G. P., Struve O., Strömgren B., 1937, ApJ, 86, 570

Kuiper R., Klahr H., Beuther H., Henning T., 2010, The Astrophysical Journal, 722, 1556

Kumamoto J., Noguchi M., 2016, ApJ, 822, 110

Lada C. J., 1987, in Peimbert M., Jugaku J., eds, Vol. 115, Star Forming Regions. p. 1

Lada C. J., Lada E. A., 2003, ARA&A, 41, 57

Lada C. J., Shu F. H., 1990, Science, 248, 564

Larson R. B., 1973, FCPh, 1, 1

Lazzarini M., et al., 2022, arXiv e-prints, p. arXiv:2206.11393

Lebouteiller V., Bernard-Salas J., Brandl B., Whelan D. G., Wu Y., Charmandaris V., Devost D., Houck J. R., 2008, ApJ, 680, 398

Lee W.-K., 2014, ApJ, 792, 122

Lee W.-K., Shu F. H., 2012, ApJ, 756, 45

Lee M. G., Freedman W. L., Madore B. F., 1993, ApJ, 417, 553

Lee J., et al., 2021, Embedded Star Formation in Nearby Galaxies: The Advent of Parsec Scale Studies beyond the Magellanic Clouds, JWST Proposal. Cycle 1, ID. #2130

Leisy P., Dennefeld M., Alard C., Guibert J., 1997, A&AS, 121, 407

Leisy P., Corradi R. L. M., Magrini L., Greimel R., Mampaso A., Dennefeld M., 2005, A&A, 436, 437

Leitherer C., 2020, Galaxies, 8, 13

Letarte B., Demers S., Battinelli P., Kunkel W. E., 2002, AJ, 123, 832

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

Louvet F., 2018, in Di Matteo P., Billebaud F., Herpin F., Lagarde N., Marquette J. B., Robin A., Venot O., eds, SF2A-2018: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics. p. Di

Lumsden S. L., Hoare M. G., Urquhart J. S., Oudmaijer R. D., Davies B., Mottram J. C., Cooper H. D. B., Moore T. J. T., 2013, ApJS, 208, 11

Lynden-Bell D., Pringle J. E., 1974, MNRAS, 168, 603

Ma J., 2001, Chinese Physics Letters, 18, 1420

Mac Low M.-M., 1999, ApJ, 524, 169

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., Villaver E., 2009, ApJ, 696, 729

Magrini L., Stanghellini L., Corbelli E., Galli D., Villaver E., 2010, A&A, 512, A63

Maitra C., Haberl F., Ivanov V. D., 2018, in 42nd COSPAR Scientific Assembly. pp E1.12–27–18

Maravelias G., Bonanos A. Z., Tramper F., de Wit S., Yang M., Bonfini P., 2022, arXiv e-prints, p. arXiv:2203.08125

Martínez-Galarza J. R., Hunter D., Groves B., Brandl B., 2012, ApJ, 761, 3

Martins F., Hillier D. J., Paumard T., Eisenhauer F., Ott T., Genzel R., 2008, A&A, 478, 219

Massey P., 1998, ApJ, 501, 153

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

Mateo M. L., 1998, ARA&A, 36, 435

Matsukoba R., Tanaka K. E. I., Omukai K., Vorobyov E. I., Hosokawa T., 2022, MNRAS, 515, 5506

Maud L. T., et al., 2019, A&A, 627, L6

McConnachie A. W., Higgs C. R., Thomas G. F., Venn K. A., Côté P., Battaglia G., Lewis G. F., 2021, MNRAS, 501, 2363

McInnes L., Healy J., Astels S., 2017, The Journal of Open Source Software, 2, 205

McKee C. F., Tan J. C., 2003, ApJ, 585, 850

Meixner M., et al., 2006, AJ, 132, 2268

Meixner M., et al., 2013, AJ, 146, 62

Meyer D. M. A., Vorobyov E. I., Kuiper R., Kley W., 2017, arXiv e-prints, p. arXiv:1710.02320

Miura R. E., et al., 2012, ApJ, 761, 37

Miville-Deschênes M.-A., Murray N., Lee E. J., 2017, ApJ, 834, 57

Mo H., van den Bosch F. C., White S., 2010, Galaxy Formation and Evolution

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

Mooney T. J., Solomon P. M., 1988, ApJL, 334, L51

More A. S., Rana D. P., 2017, in 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM). pp 72–78, doi:10.1109/ICISIM.2017.8122151

Mosteller F., Tukey J. W., 1968, in Lindzey G., Aronson E., eds, Handbook of Social Psychology, Vol. 2. Addison-Wesley

Mostoghiu R., Di Cintio A., Knebe A., Libeskind N. I., Minchev I., Brook C., 2018, MNRAS, 480, 4455

Mottram J. C., et al., 2011, ApJL, 730, L33

Mueller M. W., Arnett W. D., 1976, ApJ, 210, 670

Muraoka K., et al., 2020, ApJ, 903, 94

Navarete F., Damineli A., Barbosa C. L., Blum R. D., 2015, in Meynet G., Georgy C., Groh J., Stee P., eds, Vol. 307, New Windows on Massive Stars. pp 453–454, doi:10.1017/S1743921314007376

Neugent K. F., Massey P., 2011, ApJ, 733, 123

Neugent K. F., Levesque E. M., Massey P., Morrell N. I., Drout M. R., 2020, ApJ, 900, 118

Ochsendorf B. B., Brown A. G. A., Bally J., Tielens A. G. G. M., 2015, ApJ, 808, 111

Oey M. S., et al., 2018, ApJL, 867, L8

Ohno T., Fukui Y., Tsuge K., Sano H., Tachihara K., 2020, arXiv e-prints, p. arXiv:2006.02279

Oliva G. A., Kuiper R., 2020, A&A, 644, A41

Oliveira J. M., et al., 2009, ApJ, 707, 1269

Oliveira J. M., et al., 2011, MNRAS, 411, L36

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

Pecaut M. J., Mamajek E. E., 2013, ApJS, 208, 9

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

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 Nuclear Activity in Galaxies Across Cosmic Time. pp 335–338

Pennock C. M., et al., 2022a, MNRAS,

Pennock C. M., et al., 2022b, arXiv e-prints, p. arXiv:2207.12301

Peters T., Klessen R. S., Mac Low M.-M., Banerjee R., 2010, ApJ, 725, 134

Pietrzyński G., et al., 2013, Nature, 495, 76

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

Price-Whelan A. M., et al., 2018, The Astronomical Journal, 156, 123

Prole L. R., Clark P. C., Klessen R. S., Glover S. C. O., 2022, MNRAS, 510, 4019

Prusti T., Adorf H. M., Meurs E. J. A., 1992, A&A, 261, 685

Quirk A. C. N., et al., 2022, AJ, 163, 166

Rahmah N., Sitanggang I. S., 2016, IOP Conference Series: Earth and Environmental Science, 31, 012012

Ramírez-Galeano L., Ballesteros-Paredes J., Smith R. J., Camacho V., Zamora-Avilés M., 2022, MNRAS, 515, 2822

Reid W. A., 2014, MNRAS, 438, 2642

Reipurth B., Heathcote S., 1997, in Reipurth B., Bertout C., eds, Vol. 182, Herbig-Haro Flows and the Birth of Stars. pp 3–18

Reis I., Baron D., Shahaf S., 2019, AJ, 157, 16

Reiter M., Nayak O., Meixner M., Jones O., 2019, MNRAS, 483, 5211

Relaño M., Kennicutt Robert C. J., 2009, ApJ, 699, 1125

Rémy-Ruyer A., et al., 2015, A&A, 582, A121

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

Rieke M. J., Kelly D., Horner S., 2005, in Heaney J. B., Burriesci L. G., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 5904, Cryogenic Optical Systems and Instruments XI. pp 1–8, doi:10.1117/12.615554

Rieke G. H., et al., 2015, PASP, 127, 584

Rim P., Steinhardt C., Clark T., Diaconu A., Rusakov V., Sneppen A., 2022, in American Astronomical Society Meeting Abstracts. p. 241.38

Rizzi L., Bresolin F., Kudritzki R.-P., Gieren W., Pietrzyński G., 2006, ApJ, 638, 766

Roberts W. W., 1969, ApJ, 158, 123

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

Rosolowsky E., et al., 2021, The First Resolved View of Individual Star Formation Across a Spiral Arm, JWST Proposal. Cycle 1, ID. #2128

Rowe J. F., Richer H. B., Brewer J. P., Crabtree D. R., 2005, AJ, 129, 729

Ruffle P. M. E., et al., 2015, MNRAS, 451, 3504

Russeil D., 2003, A&A, 397, 133

Sana H., Gosset E., Nazé Y., Rauw G., Linder N., 2008, MNRAS, 386, 447

Sana H., et al., 2012, Science, 337, 444

Santiago C., Chaushev A., Sallum S., 2022, in American Astronomical Society Meeting Abstracts. p. 418.08

Schmidt M., 1959, ApJ, 129, 243

Schmidt T., et al., 2020, A&A, 641, A134

Schruba A., et al., 2017, ApJ, 835, 278

Schulz N. S., 2012, The Formation and Early Evolution of Stars, doi:10.1007/978-3-642-23926-7.

Seale J. P., Looney L. W., Chu Y.-H., Gruendl R. A., Brandl B., Chen C. H. R., Brandner W., Blake G. A., 2009, ApJ, 699, 150

Searle L., 1971, ApJ, 168, 327

Sellwood J. A., 2011, MNRAS, 410, 1637

Semczuk M., Łokas E. L., Salomon J.-B., Athanassoula E., D'Onghia E., 2018, ApJ, 864, 34

Semenov V. A., Kravtsov A. V., Gnedin N. Y., 2021, ApJ, 918, 13

Sewiło M., et al., 2013, ApJ, 778, 15

Shu F. H., Milione V., Gebel W., Yuan C., Goldsmith D. W., Roberts W. W., 1972, ApJ, 173, 557

Shu F. H., Adams F. C., Lizano S., 1987, ARA&A, 25, 23

Sibbons L. F., Ryan S. G., Cioni M. R. L., Irwin M., Napiwotzki R., 2012, A&A, 540, A135

Sibbons L. F., Ryan S. G., Napiwotzki R., Thompson G. P., 2015, A&A, 574, A102

Skillman E. D., Terlevich R., Melnick J., 1989, MNRAS, 240, 563

Skrutskie M. F., et al., 2006, AJ, 131, 1163

Spera M., Mapelli M., Jeffries R. D., 2016, MNRAS, 460, 317

Sridharan T. K., Beuther H., Schilke P., Menten K. M., Wyrowski F., 2002, ApJ, 566, 931

Staveley-Smith L., Cohen R. J., Chapman J. M., Pointon L., Unger S. W., 1987, MNRAS, 226, 689

Steinhardt C. L., Weaver J. R., Maxfield J., Davidzon I., Faisst A. L., Masters D., Schemel M., Toft S., 2020, ApJ, 891, 136

Stephens I. W., et al., 2017, ApJ, 834, 94

Stone J. M., Ostriker E. C., Gammie C. F., 1998, ApJL, 508, L99

Strömgren B., 1939, ApJ, 89, 526

Swan J., Cole A. A., Tolstoy E., Irwin M. J., 2016, MNRAS, 456, 4315

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. p. 149 (arXiv:1402.0919), doi:10.2458/azu˙uapress˙9780816531240-ch007

Tokuda K., et al., 2020, ApJ, 896, 36

Tolstoy E., Irwin M. J., Cole A. A., Pasquini L., Gilmozzi R., Gallagher J. S., 2001, MNRAS, 327, 918

Toomre A., 1977, ARA&A, 15, 437

Tosaki T., Miura R., Sawada T., Kuno N., Nakanishi K., Kohno K., Okumura S. K., Kawabe R., 2007, ApJL, 664, L27

Tsuge K., et al., 2019, ApJ, 871, 44

Tully R. B., 1974, ApJS, 27, 415

Tully R. B., Courtois H. M., Sorce J. G., 2016, AJ, 152, 50

Úbeda L., Drissen L., 2009, MNRAS, 394, 1847

Van Gelder M. L., et al., 2020, A&A, 636, A54

Van Loon J. T., 2008, MmSAI, 79, 412

Van Loon J. T., Sansom A. E., 2015, MNRAS, 453, 2341

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

Van der Maaten L., Hinton G., 2008, Journal of Machine Learning Research, 9, 2579

Van der Marel R. P., 2006, in Livio M., Brown T. M., eds, Vol. 17, The Local Group as an Astrophysical Laboratory. pp 47–71 (`arXiv:astro-ph/0404192`)

Van der Marel R. P., Cioni M.-R. L., 2001, AJ, 122, 1807

Verley S., Corbelli E., Giovanardi C., Hunt L. K., 2009, A&A, 493, 453

Vink J. S., 2018, A&A, 615, A119

Visser H. C. D., 1980, A&A, 88, 159

Volders L. M. J. S., Högbom J. A., 1961, BAN, 15, 307

Vorobyov E. I., 2009, ApJ, 704, 715

Wada K., 2008, ApJ, 675, 188

Wada K., Koda J., 2004, MNRAS, 349, 270

Wada K., Baba J., Saitoh T. R., 2011, ApJ, 735, 1

Wang Y., Gao J., Ren Y., Chen B., 2022, arXiv e-prints, p. arXiv:2204.05548

Ward-Thompson D., Whitworth A. P., 2015, An Introduction to Star Formation

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

Weidner C., Kroupa P., 2004, MNRAS, 348, 187

Weldrake D. T. F., de Blok W. J. G., Walter F., 2003, MNRAS, 340, 12

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, ApJL, 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

Wood D. O. S., Churchwell E., 1989, ApJ, 340, 265

Woodward P. R., 1975, ApJ, 195, 61

Wright N. J., Bouy H., Drew J. E., Sarro L. M., Bertin E., Cuillandre J.-C., Barrado D., 2016, MNRAS, 460, 2593

Xiang M., et al., 2018, ApJS, 237, 33

Zari E., Rix H. W., Frankel N., Xiang M., Poggio E., Drimmel R., Tkachenko A., 2021, A&A, 650, A112

Zelko I. A., Finkbeiner D. P., 2020, ApJ, 904, 38

Zhang Y., et al., 2022, arXiv e-prints, p. arXiv:2207.11320

Zivick P., et al., 2019, ApJ, 874, 78

Zombeck M. V., 2006, Handbook of Space Astronomy and Astrophysics, 3 edn. Cambridge University Press, doi:10.1017/CBO9780511536359

Zuckerman B., Palmer P., 1974, ARA&A, 12, 279