Publications of the Astronomical Society of Australia (2025), **42**, e028, 18 pages
doi:10.1017/pasa.2025.5

CAMBRIDGE UNIVERSITY PRESS

Research Article

WALLABY Pilot Survey: kNN identification of perturbed galaxies through H 1 morphometrics

Benne Willem Holwerda[1], Helga Dénes[2], Jonghwan Rhee[3], Denis Leahy[4], Bärbel Silvia Koribalski[5,6], Niankun Yu[7,8], Nathan Deg[9], T. Westmeier[3], Karen Lee-Waddell[10], Yago Ascasibar[11,12], Manasvee Saraf[3,13,14], Xuchen Lin[15], Barbara Catinella[16,14] and Kelley Hess[17]

[1]Department of Physics and Astronomy, University of Louisville, Louisville, KY, USA, [2]School of Physical Sciences and Nanotechnology, Yachay Tech University, Urcuquí, Ecuador, [3]International Centre for Radio Astronomy Research (ICRAR), University of Western Australia, Crawley, WA, Australia, [4]Department of Physics and Astronomy, University of Calgary, Calgary, AB, Canada, [5]Australia Telescope National Facility, CSIRO, Space and Astronomy, Parkes, NSW, Australia, [6]School of Science, Western Sydney University, Penrith, NSW, Australia, [7]National Astronomical Observatories, Chinese Academy of Sciences, Beijing, People's Republic of China, [8]Key Laboratory of Radio Astronomy and Technology, Chinese Academy of Sciences, Beijing, People's Republic of China, [9]Department of Physics, Engineering Physics, and Astronomy, Queen's University, Kingston, ON, Canada, [10]Australian SKA Regional Centre, Perth, Australia, [11]Departamento de Física Teórica, Universidad Autónoma de Madrid (UAM), Madrid, Spain, [12]Centro de Investigación Avanzada en Física Fundamental (CIAFF-UAM), Madrid, Spain, [13]Australia Telescope National Facility, CSIRO, Space and Astronomy, Bentley, WA, Australia, [14]ARC Centre of Excellence for All-Sky Astrophysics in 3 Dimensions (ASTRO 3D), Sydney, Australia , [15]Department of Astronomy, School of Physics, Peking University, Beijing, People's Republic of China, [16]International Centre for Radio Astronomy Research, The University of Western Australia, Crawley, WA, Australia and [17]Chalmers University of Technology, Onsala Space Observatory, Göteborg, Sweden

Abstract

Galaxy morphology in stellar light can be described by a series of 'non-parametric' or 'morphometric' parameters, such as concentration-asymmetry-smoothness, Gini, M_{20}, and Sérsic fit. These parameters can be applied to column density maps of atomic hydrogen (H 1). The H 1 distribution is susceptible to perturbations by environmental effects, for example, intergalactic medium pressure and tidal interactions. Therefore, H 1 morphology can potentially identify galaxies undergoing ram-pressure stripping or tidal interactions. We explore three fields in the WALLABY Pilot H 1 survey and identify perturbed galaxies based on a k-nearest neighbour (kNN) algorithm using an H 1 morphometric feature space. For training, we used labelled galaxies in the combined NGC 4808 and NGC 4636 fields with six H 1 morphometrics to train and test a kNN classifier. The kNN classification is proficient in classifying perturbed galaxies with all metrics – accuracy, precision, and recall – at 70–80%. By using the kNN method to identify perturbed galaxies in the deployment field, the NGC 5044 mosaic, we find that in most regards, the scaling relations of perturbed and unperturbed galaxies have similar distribution in the scaling relations of stellar mass versus star formation rate and the Baryonic Tully–Fisher relation, but the H 1 and stellar mass relation flatter than of the unperturbed galaxies. Our results for NGC 5044 provide a prediction for future studies on the fraction of galaxies undergoing interaction in this catalogue and to build a training sample to classify such galaxies in the full WALLABY survey.

Keywords: Galaxies: evolution; galaxies galaxies: interactions; galaxies galaxies: ISM; galaxies galaxies: structure; galaxies

(Received 7 August 2024; revised 3 January 2025; accepted 10 January 2025)

1. Introduction

The atomic gas (H 1) disc extends well beyond the stellar disc of spiral galaxies at the same surface density (e.g. Bosma 1978; Begeman 1989; Meurer et al. 1996; Meurer, Staveley-Smith, & Killeen 1998; Swaters et al. 2002; Noordermeer et al. 2005; Walter et al. 2008; Boomsma et al. 2008; Elson, de Blok, & Kraan-Korteweg 2011; Heald et al. 2011b; Heald et al. 2011a; Zschaechner et al. 2011; de Blok et al. 2008; Koribalski et al. 2018; de Blok et al. 2020) for examples and discussions on H 1 discs). For comparison, see Trujillo, Chamba, & Knapen (2020), Chamba, Trujillo, & Knapen (2022) for a discussion on the defined edge of stellar discs. The outer regions of these discs are sensitive to ram-pressure stripping by the intergalactic medium (IGM, Wang et al. 2021; Reynolds et al. 2021; Reynolds et al. 2022). A lopsided appearance of the outer H 1 disc (Jog & Combes 2009; van Eymeren et al. 2011a; van Eymeren et al. 2011b; Koribalski et al. 2018) or an asymmetry (Giese et al. 2016; Reynolds et al. 2020) can be attributed to tidal interactions (Jog & Combes 2009; Koribalski & López-Sánchez 2009), ram-pressure stripping (Moore & Gottesman 1998; Westmeier, Koribalski, & Braun 2013; Hess et al. 2022), a lopsided dark matter halo (Jog 2002), ongoing mergers, or a combination of these. The H 1 in the outer part of disc galaxies is much more sensitive to gravitational (tidal) interaction as well as pressure interactions with the group or cluster medium than the stellar component (e.g. Hibbard et al. 2001). As galaxies are pre-processed in groups, one of the first signs of tidal

Corresponding author: Benne Willem Holwerda, Email: benne.holwerda@louisville.edu.

Cite this article: Holwerda BW, Dénes H, Rhee J, Leahy D, Koribalski BS, Yu N, Deg N, Westmeier T, Lee-Waddell K, Ascasibar Y, Saraf M, Lin X, Catinella B and Hess K. (2025) WALLABY Pilot Survey: kNN identification of perturbed galaxies through H 1 morphometrics. *Publications of the Astronomical Society of Australia* **42**, e028, 1–18. https://doi.org/10.1017/pasa.2025.5

interactions will be the changes in their gas discs. It is likely that the H 1 asymmetry is caused by either tidal interaction, or ram-pressure stripping, or both (Yu et al. 2022; Watts et al. 2021). But some internal perturbation could affect the H 1 distribution in a similar way, such as AGN feedback (e.g. Villaescusa-Navarro et al. 2016; Morganti 2017) or stellar feedback (e.g. Ashley et al. 2017), or mergers (Zuo et al. 2022). As with warps in the H 1 disc or stellar disc truncations, multiple mechanisms, both internal and external, could be responsible.

Parameterisation of H 1 disc appearance is different from stellar parametrisation because the H 1 disc is based on line emission and therefore has a much lower dynamic range: high density H 1 would become molecular hydrogen while low density H 1 is difficult to detect and lack of self-shielding would result in transition to ionised hydrogen. The area covered by an H 1 disc is larger, but the spatial resolution is typically an order of magnitude lower due to the much larger H 1 beam (or PSF) compared to optical imaging. The morphometric parameter space is one used extensively in ultraviolet/optical images of galaxies: the C-A-S (Conselice 2003), Gini-M_{20} (Lotz, Primack, & Madau 2004), DIM (Rodriguez-Gomez et al. 2019), and Sérsic profile (Sérsic 1968).

Here, we apply the galaxy morphometrics originally developed for stellar discs which were applied with some success on H 1 data in the past (Holwerda et al. 2011c; Holwerda et al. 2011d; Holwerda et al. 2011a; Holwerda et al. 2011b; Holwerda et al. 2011e; Holwerda, Pirzkal, & Heiner 2012; Giese et al. 2016; Reynolds et al. 2020; Reynolds et al. 2023; Deg et al. 2023; Holwerda et al. 2023) but on often heterogeneous data. For example Giese et al. (2016) pointed out that these H 1 morphometrics depend strongly on the signal-to-noise ratio of each object, complicating their use across surveys or with varying s/n. Reynolds et al. (2020) illustrated the challenge to compare morphometrics, specifically asymmetry, across different H 1 surveys. The optimal application is therefore within a single survey and a well-documented implementation, that is, STATMORPH implementation of these morphometrics (Rodriguez-Gomez et al. 2019). Here we use STATMORPH, a python-based tool to compute the most commonly used galaxy morphometrics on already segmented images. This tool is public and uses the commonly used definitions of each morphology parameter and fits a single Sérsic profile to the light distribution. It was developed for ultraviolet/optical/near-infrared imaging but translates well to H 1 images (Holwerda et al. 2023).

H 1 morphometrics are a potential feature space for machine learning algorithms. One could classify if galaxies are undergoing ram-pressure stripping, tidal interactions, or even ongoing mergers based on their position in the H 1 morphology space. The caveat is that a sufficient training set has to be available. Ideally, the training set spans the input feature space and all the possible use cases. Our goal here is to examine how well one can train a simple classifier based on the H 1 catalogue of a single field of galaxies observed by Widefield ASKAP L-band Legacy All-sky Blind surveY (WALLABY, Koribalski 2012; Koribalski et al. 2020) and generalise the results to other groups. The H 1 morphometric space is familiar but it remains unclear which morphometrics are most useful to identify H 1 perturbations. Our goals break down into how well one can get an interaction fraction in a given group, that is, a population characteristic and how well one can identify individual galaxies as undergoing a disturbance, be it tidal or ram-pressure stripping.

The WALLABY (Koribalski 2012; Koribalski et al. 2020) is an interferometric H 1 survey carried out with the Australian Square Kilometer Array Pathfinder (ASKAP, Johnston et al. 2008; Hotan et al. 2021), which provides an ideal laboratory for H 1 morphometrics. The survey is of uniform image quality and will cover a large fraction of the sky and the local Universe. In the future the WALLABY pipeline will create higher resolution postage stamps for pre-selected galaxies but we use the present pipeline products here.

Deep, high-resolution, and uniform H 1 maps, across S/N, resolution, and sensitivity, for a large number of galaxies allow us to compare across environments. The WALLABY pilot survey (Westmeier et al. 2022; Kim et al. 2023; Courtois et al. 2023; Grundy et al. 2023) has observed several groups and clusters of galaxies. Here, we use the data of three fields centred on groups of galaxies: NGC 4636, NGC 4808, and NGC 5044 to analyse the effects of environmental effects on H 1 morphology. One of these groups, NGC 4636, has been examined in detail by Lin et al. (2023) and assessed for signs of ram-pressure stripping, tidal effects, and mergers. Their labels extend into the NGC 4808 field as well. We will use labelling in these fields as our training/testing sample and the sources in the remaining mosaic on NGC 5044 as the application sample.

Throughout we use the Planck (2015) cosmology ($H_0 = 67.74$ km/s/Mpc, $\Omega_0 = 0.3075$, Planck Collaboration et al. 2016). We adopt a Chabrier initial mass function (IMF, Chabrier 2003) to uniformly derive SFRs and stellar masses. The paper is organised as follows: Section 2 describes briefly the WALLABY pilot survey and other data products used, Section 4 details the definitions of the morphometric parameters used, Section 5 introduces the machine learning algorithm used and the input considerations, Section 6 shows the results of the k-nearest neighbour (kNN) classification effort, Section 7 discusses these results in context of future uses, and Section 8 are our conclusions.

2. WALLABY data

The WALLABY survey (Koribalski 2012; Koribalski et al. 2020) is an all-sky H I survey carried out with wide-field Phased Array Feeds (PAFs) on the Australian Square Kilometre Array Pathfinder (ASKAP, Johnston et al. 2008; Hotan et al. 2021). ASKAP consists of 36 × 12-m telescopes forming a 6-km diameter interferometer. The PAFs are used to from 36 overlapping beams and together deliver a field-of-view of ∼30 square degrees with a resolution of 30" and 4 km s^{-1}. Before the start of full survey operations, a number of fields were observed in early science and pilot survey programmes (Serra et al. 2015b; Lee-Waddell et al. 2019; Kleiner et al. 2019; For et al. 2019; For et al. 2021).

The Phase 2 WALLABY pilot survey is described in detail in Westmeier et al. (2022). The pilot data was made available to the collaboration for initial science projects. This includes the single tile on NGC 4808, NGC 4636, and Vela fields and the 4-tile mosaic in the direction of the NGC 5044 group. Thanks to improvements in data quality and source finding with SoFiA (Serra et al. 2015a; Westmeier et al. 2021), the total number of H 1 detections is higher in the final pilot data. The WALLABY data of these three groups are described in detail in Murugeshan et al. (2024, *submitted*). The ASKAP interferometric WALLABY survey has a beam size of ∼30" and an rms of 1.6 mJy beam^{-1} for a velocity resolution of 4 km/s (Koribalski et al. 2020).

https://doi.org/10.1017/pasa.2025.5

Table 1. Basic properties of the three galaxy group WALLABY fields analysed here.

Field	RA (deg)	DEC (deg)	Group distance (Mpc)	cz km/s	No. ASKAP fields #	No. objects Full SoFiA catalogue	Resolved ($D < 60$ Mpc)
Virgo							
NGC 4636	190.7084	2.6880	16.2	919	1	231	48
NGC 4808	193.953958	4.304111	16.0	760	1	147	89
NGC 5044	198.849875	−16.385528	45.7		4	1326	258

2.1 Virgo (NGC 4636 and NGC 4808 fields)

WALLABY's Phase 2 pilot programme observed two close fields, each centred on one of two Virgo groups, NGC 4636 (Lin et al. 2023) and NGC 4808 (Murugeshan et al. submitted). NGC 4636 is a relatively close group at a distance of 16.2 Mpc (Kourkchi & Tully 2017) and a radius of 0.61 Mpc, based on ROSAT X-ray measurements (Reiprich & Böhringer 2002). Two galaxies, NGC 6156 (in the Norma Field) and NGC 4632 (in this field), were studied in detail in Deg et al. (2023) as they show a polar ring structure in H 1. Lin et al. (2023) presents a catalogue of galaxies around the N4636 group centre with redshift measurements from several H 1 and optical catalogues. Lin et al. (2023) note that of the 19 galaxies detected by WALLABY belonging to this group, six galaxies are resolved enough for detailed moment-0 map study. They present flags for different types of interaction based on the combined WALLABY-FAST data, which include objects in the NGC 4808 field. This is the basis for our training sample (see Section 5.1). The second WALLABY pilot field is centred around NGC 4808 group. The Tully–Fisher (T-F) distances in this field are presented in Courtois et al. (2023). This group is similarly close, at approximately ∼16 Mpc.

2.2 NGC 5044 mosaic

The third field, a mosaic of four fields centred on the NGC 5044 group, has been studied across wavelengths before in the optical, x-ray, and H 1 observations (Ferguson & Sandage 1990; Ferguson & Sandage 1991; Tamura et al. 2003; Buote et al. 2003; Buote, Brighenti, & Mathews 2004; Osmond & Ponman 2004; McKay et al. 2004; Forbes et al. 2006).

The WALLABY internal release on this field (DR3) covers 120 deg^2 of the NGC 5044 four-tile mosaic across a 21 cm H 1 line red shift range of $cz \sim 500$ to 25 400 km/s ($z < 0.085$), which uses the full RFI-free bandwidth available to WALLABY. NGC 5044 DR3 includes 1 326 detections. The resulting catalogue is richest of the three in source counts with a large number of sources well behind the nearby group around NGC 5044 (Fig. 1).

Fig. 1 shows the distribution of H 1 detections in all three catalogues. A majority of sources in the three fields is *not* associated with the groups themselves. In the NGC 5044 field especially, several groups and clusters can be identified in the background. There are full samples of these fields for which we compute H 1 morphometrics and assign stellar mass and star formation rates.

3. Stellar mass and star formation rates

To ensure uniform stellar mass and star formation rates over the three datasets, we adopt the WISE photometry derived stellar mass and star formation rates expressions as described in Jarrett et al. (2011), Jarrett et al. (2013), Cluver et al. (2014). For stellar mass, we use their eq. 2 with an absolute solar magnitude $M_{W1} = 3.24$ (W1, Vega), and for the SFR their equation (5), with $M_{W3} = 3.24$ (W3, Vega). Solar luminosities are from Willmer (2018) and for each object, we search the ALLWISE catalogue accessible through IPAC. Stellar mass-to-light ratios are based on the W1MPRO – W2MPRO colour, the stellar mass is derived from the above mass-to-light ratio, the W1MPRO and the distance derived from the H 1 redshift (Fig. 1). Star formation is based on the W3MPRO for all galaxies and the equation (5) in Cluver et al. (2014). Like all single colour-based mass-to-light ratios and single-band star formation indicators, these estimates are approximate with a greater degree of uncertainty than dedicated Spectral Energy Distribution modelling results (cf de los Reyes et al. 2024). In a similar vein, the distance derived from the H 1 redshift may be influenced by peculiar motion within the group. We opted for the H 1-redshift derived distances and the WISE derived stellar mass and star formation primarily because they are available for all three groups with a uniform level of quality.

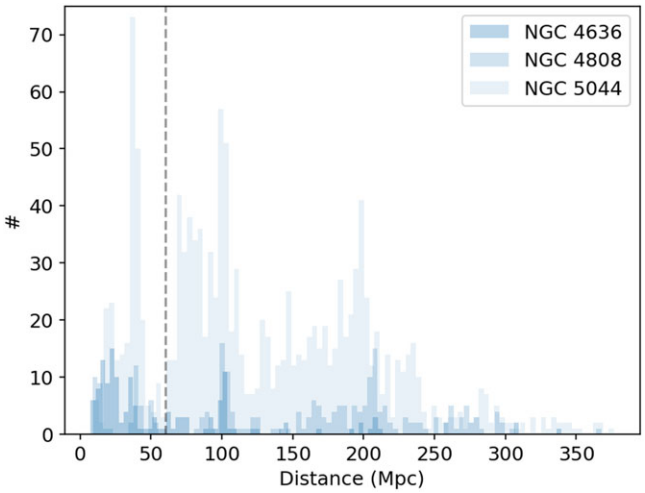

Figure 1. Distribution of distance for galaxies in the three WALLABY fields, centred on NGC 4808, NGC 4646, and NGC 5044. The vertical dashed line is the 60 Mpc cutoff for selection for the training sample in NGC 4808 and the application samples.

4. Morphometrics

One observational approach to characterise galaxy appearances is to derive morphometric parameters[a]: unitless parameters that do not depend on a preconceived idea about the shape of the profile and are invariant with distance. These morphometric parameters

[a]Sometimes called 'non-parametric' as these do not assume a Gaussian distribution of pixel values.

https://doi.org/10.1017/pasa.2025.5

(Davenport 2015) can then be used to classify galaxies along the Hubble Tuning fork or to identify mergers in a population of galaxies (Pearson, Li, & Dye 2019; Holwerda et al. 2023).

The morphometric parameters considered here are Concentration, Asymmetry and Smoothness (CAS) from Conselice (2003), M_{20} and Gini from Lotz et al. (2004), and the Multimode-Intensity-Deviation (MID) parameters from Peth et al. (2016). We use the STATMORPH package described in Rodriguez-Gomez et al. (2019) to compute the morphometrics.

We utilise a Gaussian smoothing kernel with a 1 pixel FWHM (6") for the H 1 implementations of STATMORPH. This choice is not critical for most computed morphometrics except for the Sérsic profile fit in STATMORPH and the Intensity and Smoothness morphometric parameters (see Sections 4.1 and 4.3). H 1 profiles are typically not well described with such a Sérsic profile (cf Leroy et al. 2008; Bigiel, Leroy, & Walter 2011; Wang et al. 2014; Swaters et al. 2002; Reynolds et al. 2023). We did not anticipate the use of Smoothness or Intensity because of this additional tuning parameter (see Section 5.2). The central position of the galaxy (x_c, y_c) is re-computed by STATMORPH, and the segmentation map is the SoFiA 3D mask (Westmeier et al. 2021) with the frequency axis collapsed.

4.1 Concentration-Asymmetry-Smoothness (CAS) morphometrics

CAS refers to the commonly used Concentration-Asymmetry-Smoothness space (Conselice 2003) for stellar morphological analysis of distant galaxies. Concentration of the light, symmetry around the centre and smoothness is an indication of substructure.

Concentration is defined by Bershady, Jangren, & Conselice (2000) as:

$$C = 5\ \log\left(r_{80}/r_{20}\right) \tag{1}$$

with r_f as the radius containing percentage f of the light of the galaxy (see definitions of r_f in Bertin & Arnouts 1996; Holwerda 2005). In the optical regime (i.e. stellar component), typical values for the concentration index are $C = 2 - 3$ for discs, $C > 3.5$ for massive ellipticals, while peculiars span the entire range (Conselice 2003).

The asymmetry is defined as the level of *point*-, (or rotational-) symmetry around the centre of the galaxy (Abraham et al. 1994; Conselice 2003):

$$A = \frac{\Sigma_{i,j}|I(i,j) - I_{180}(i,j)|}{\Sigma_{i,j}|I(i,j)|} - A_{bgr}, \tag{2}$$

where $I(i,j)$ is the value of the pixel at the position $[i,j]$ in the image, and $I_{180}(i,j)$ is the pixel at position $[i,j]$ in the galaxy's image, after it was rotated 180° around the centre of the galaxy. A_{bgr} is an estimate of the contribution of the background to this value. This is the definition, without the background contribution, used in Holwerda et al. (2012) for H 1 as line emission does not have a clear background contribution to asymmetry. Because we use the postage stamps extracted by SoFiA for the calculation, we use the definition of asymmetry without the background computation.

In the STATMORPH implementation, the asymmetry is calculated in the inner 1.5 Petrosian[b] radii (typical size of the stellar disc), the background asymmetry is subtracted, and A is minimised by moving the centre of rotation. Note that the maximum value for the asymmetry is 2 (all pixels off-centre) and can be negative if the background asymmetry value is large. We note that we do not subtract a background when using the moment-0 H 1 maps as these are extracted from the field using a 3D source mask. A background subtraction makes more sense with continuum emission where a substantial contribution to morphometrics can be expected (i.e. optical or ultraviolet emission) as opposed to line emission maps as is the case here. Moreover, the subtraction has already happened in the radio continuum subtraction that was applied to the data-cube prior to H 1 line extraction. In our case, background subtraction is a separate step in the data reduction process. To obtain a background asymmetry contribution, one would have to combine continuum subtraction, source extraction, and asymmetry computation. Reynolds et al. (2020) compute this background component using an empty Section of the H 1 cube with the same shape as the mask. This was more useful for their comparison between different H 1 surveys. Here, the background contribution would be dominated by the H 1 mask shape but STATMORPH expects to compute it based on a sky background just outside the aperture.

Asymmetry in H 1 maps or profiles has shown a lot of promise in recent studies to identify perturbed or disrupted disc galaxies (e.g. Reynolds et al. 2020; Glowacki et al. 2022; Watts et al. 2023; Holwerda et al. 2023).

Inspired by the 'unsharp masking' technique (Malin 1978), Smoothness is defined by Takamiya (1999) and Conselice (2003) as:

$$S = \frac{\Sigma_{i,j}|I(i,j) - I_S(i,j)|}{\Sigma_{i,j}|I(i,j)|} \tag{3}$$

where $I_S(i,j)$ is the same pixel in a smoothed image. What type of smoothing is used has changed over the years. Often a fixed Gaussian smoothing kernel is chosen for simplicity.

The fact that this Smoothness has another input parameter in the form of the size of the smoothing kernel, makes it highly 'tunable', meaning one gets out exactly what the parameter was optimised for. It is very difficult to reliably compare between catalogues and especially samples over different distances. The kernel employed here is a Gaussian with a width of 2.5 pixels in the moment0 map. This is less than the beam size of the instrument in question. Thanks to the lower dynamical range in H 1 maps, one does not expect the high-contrast areas such as HII regions in star-forming galaxies. The smoothing kernel choice is therefore a conservative choice (low amount of smoothing) for the Smoothness parameter. The Smoothness parameter is expected to be less useful in H 1 than in optical or ultraviolet imaging.

4.2 Gini and M_{20}

Abraham, van den Bergh, & Nair (2003) and Lotz et al. (2004) introduce the Gini parameter to quantify the distribution of flux over the pixels in an image. They use the following definition:

$$G = \frac{1}{\bar{I}n(n-1)}\Sigma_i(2i - n - 1)I_i, \tag{4}$$

[b]The Petrosian radius is one of several definitions to automatically assign a size and aperture to inherently fuzzy galaxies. For a comprehensive treatment on them, see Graham et al. (2005), Graham & Driver (2005). A different size measure of R1 (1 $M_\odot/kpc^2$ similar to those proposed by Trujillo et al. 2020; Chamba et al. 2022) may make more sense for H 1

https://doi.org/10.1017/pasa.2025.5

I_i is the value of pixel i in an ordered list of the pixels, n is the number of pixels in the image, and $\bar{I}$ is the mean pixel value in the image.

The Gini parameter is an indication of equality in a distribution (initially an economic indicator Gini 1912; Yitzhaki 1991), with $G = 0$ the perfect equality (all pixels have the same fraction of the flux) and $G = 1$ perfect inequality (all the intensity is in a single pixel). Its behaviour is therefore in between that of a structural measure and concentration. Gini appears quite sturdy as it does not require the centre of the object to be computed. It remains relatively unchanged, even when the object is lensed (Florian, Li, & Gladders 2016), and it is popular for this reason. However, it depends strongly on the image's signal-to-noise (Lisker 2008); noise forces the inclusion of a lot of low-signal pixels, throwing off the entire distribution. This issue is not noisy data but how it typically affects image segmentation. In essence, noise can add pixels with no fraction of the flux in them, artificially increasing the Gini value. However, with a less concentrated radial profile and choices of segmentation already made by SoFiA, Gini is a good fit for H 1 maps.

Lotz et al. (2004) also introduced a way to parameterise the extent of the light in a galaxy image. They define the spatial second order moment as the product of the intensity with the square of the projected distance to the centre of the galaxy. This gives more weight to emission further out in the disc. It is sensitive to substructures such as spiral arms and star-forming regions but insensitive to whether these are distributed symmetrically or not. The second order moment of a pixel i is defined as:

$$M_i = I_i \times [(x - x_c)^2 + (y - y_c)^2], \tag{5}$$

where $[x, y]$ is the position of a pixel with intensity value I_i in the image and $[x_c, y_c]$ is the central pixel position of the galaxy in the image.

The total second order moment of the image is given by:

$$M_{tot} = \Sigma_i M_i = \Sigma I_i[(x_i - x_c)^2 + (y_i - y_c)^2]. \tag{6}$$

Lotz et al. (2004) use the relative contribution of the brightest 20% of the pixels to the second order moment as a measure of disturbance of a galaxy after sorting the list of pixels by intensity (I_i):

$$M_{20} = \log\left(\frac{\Sigma_i M_i}{M_{tot}}\right), \text{ for } \Sigma_i I_i < 0.2 I_{tot}. \tag{7}$$

The M_{20} parameter is sensitive to bright regions in the outskirts of discs and higher values can be expected in galaxy images (in the optical and UV) with star-forming outer regions as well as those images of strongly interacting discs. Due to a lack of high contrast clumps at higher radii, the M_{20} parameter is not expected to show as much of a range in H 1 compared to star formation dominated wavelengths where it was first employed.

4.3 Multimode–Intensity–Deviation (MID) morphometrics

The MID morphometrics (Freeman et al. 2013; Peth et al. 2016) were introduced as an alternative to the Gini–M20 and CAS morphometrics to be more sensitive to recent mergers. However, these new morphometrics have not been tested as extensively as the Gini–M20 and CAS statistics, especially using hydrodynamic simulations (Lotz et al. 2008; Lotz et al. 2010; Lotz et al. 2011; Bignone et al. 2017), see also the discussion in the implementation in STATMORPH (Rodriguez-Gomez et al. 2019). In the case of H 1 data for the Hydra cluster, these parameters did not contribute new information (Holwerda et al. 2023).

The multimode statistic (M) measures the ratio between the areas of the two most 'prominent' clumps within a galaxy. The implicit assumption is that the galaxy is well resolved and has at least two well-defined clumps. Its calculation mostly consists in finding such substructures. First, all pixels within the MID segmentation map are sorted by brightness. Then, for a given quintile q (between 0 and 1), the set of all pixels with flux values above the qth quintile will generally consist of n groups of contiguous pixels, which are sorted by area (largest first). Finally, M is defined as the quintile q that maximises the area ratio between the two largest groups (Peth et al. 2016):

$$M = \max\left(\frac{A_{q,2}}{A_{q,1}}\right) \tag{8}$$

where $A_{q,1}$ is the largest quintile and $A_{q,2}$ is the second-to-largest quintile area.

The intensity statistic (I) measures the ratio between the two brightest subregions of a galaxy. To calculate it, the galaxy image is first slightly smoothed using a Gaussian kernel with $\sigma = 1$ pixel. Then, the image is partitioned into pixel groups according to the watershed algorithm: each distinct subregion consists of all the pixels such that their maximum gradient paths lead to the same local maximum. Once the pixel groups are defined, their summed intensities are sorted into descending order: I1, I2, etc. The intensity statistic is then defined as Freeman et al. (2013):

$$I = \frac{I_2}{I_1} \tag{9}$$

The same issue that can be raised for M can be raised here. There is a built-in assumption of resolved structure, and that this structure has not fractured the segmentation map into separate catalogue entries.

The deviation statistic (D) measures the distance between the image centroid, (x_c, y_c), calculated for the pixels identified by the MID segmentation map, and the brightest peak found during the computation of the I statistic, (x_I, y_I). This distance is normalised by $\sqrt{n_{seg}/\pi}$, where n_{seg} is the number of pixels in the segmentation map, which represents an approximate galaxy 'radius' Freeman et al. (2013):

$$D = \sqrt{\frac{\pi}{n_{seg}}}\sqrt{(x_c - x_I)^2 + (y_c - y_I)^2} \tag{10}$$

This is a metric that can be calculated from Source Extractor (Bertin & Arnouts 1996; Holwerda 2005) output by using the image centroid and the peak location, albeit again that this makes assumptions on the number of substructures in the galaxy image.

4.4 Patchiness

A recent addition to the morphometric parameter space is a 'patchiness' parameter (Fetherolf et al. 2023) defines as:

$$P = -\log_{10}\left\{\Pi_i^N \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{(X_i - \bar{X_w})^2}{2\sigma_i^2}\right]\right\} \tag{11}$$

where N_i is the number of pixels, X_i is the value of pixel i. The Gaussian probability that all the pixels equal the weighted average is lower when the image is 'patchier'. Here, X_w is the weighted (or not) average of the distribution of pixels that make up the object and σ_i is the pixel uncertainty. The benefits are that this

https://doi.org/10.1017/pasa.2025.5

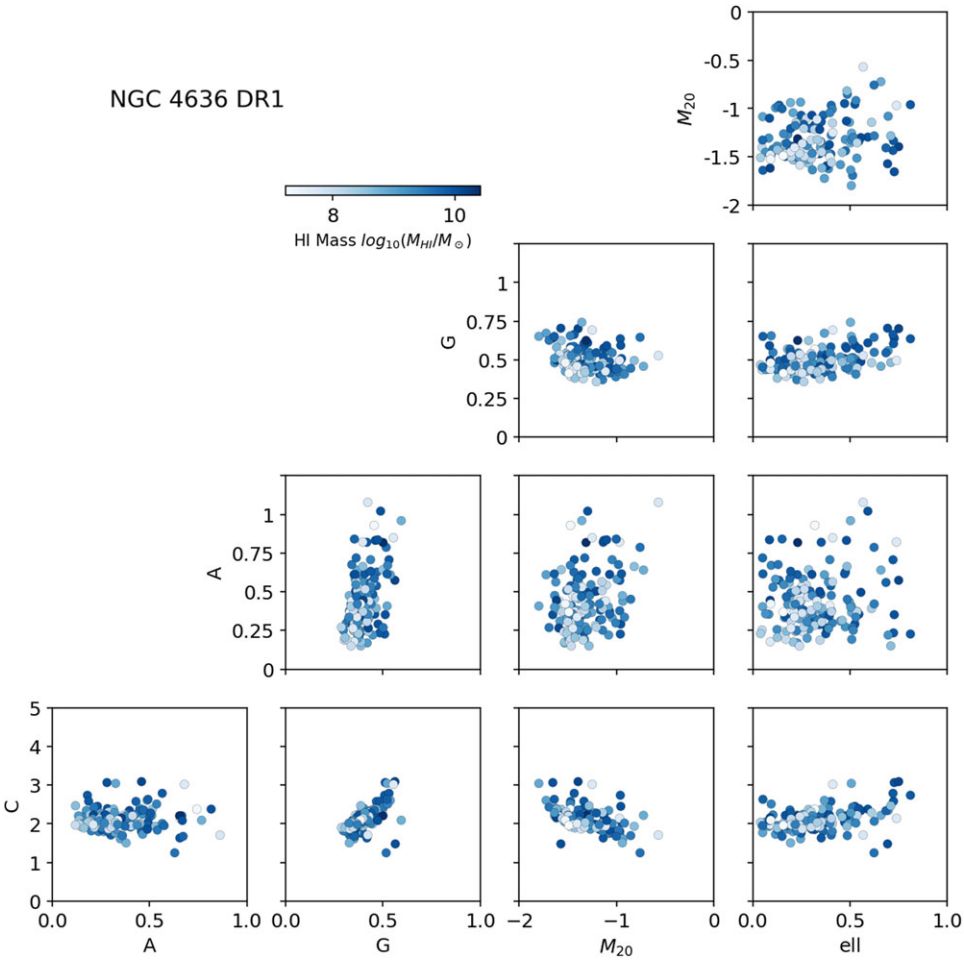

Figure 2. A corner plot of the H 1 morphometrics of galaxies in the NGC 4636 pointing based on the SoFiA segmentation maps.

measure is sensitive to deviations above and below the average. It is also notable that this parameter, like the Gini parameter, does not depend on the central position, unlike M_{20} or Asymmetry, which relies on a bright subset of pixels and the object's central position. Fetherolf et al. (2023) use their parameter for Voronoi tesselations of their objects and not individual pixels. However, we implement a pixel-based definition here. The implementation in Fetherolf et al. (2023) focused on their reddening maps, that is, the dust distribution. Therefore, this seemed a likely H 1 morphometric. Upon implementation however, it became clear that the values computed from SoFiA maps are often infinite. For completeness, we include the values in our final catalogue, but not use it in the kNN training below.

4.5 Sérsic profile

The final step by STATMORPH is to fit a single Sérsic profile with the effective radius (r_{50}) and index (n) to the pixel collection constituting each object. This is not the optimal description of the H 1 disc which is usually described with a $R_{1M_\odot}$, the radius where the profile reaches 1 $M_\odot/pc^2$ in H 1 mass. However, this alternate morphometric, very commonly used in optical studies, is available for use here, and we include the Sérsic index for consideration.

4.6 STATMORPH

Calculating catalogues for these WALLABY data is straightforward for the cutouts provided by the WALLABY data-release. One can run through all the entries made by the SoFiA source detection and run STATMORPH (Rodriguez-Gomez et al. 2019). These catalogues are our starting point for the machine learning approach described in the following sections.

Part of the parameter space of H 1 morphometrics are presented in Figs. 2 through 4, colour-coded by the inferred H 1 mass from the SoFiA catalogue. The H 1 masses are derived after application of the H 1 flux correction as described in Westmeier et al. (2022). This flux correction is a critical step to match the H 1 size-mass relation. Concentration, Asymmetry, Gini and M_{20} are the most commonly used parameters.

These are full morphometric catalogues for each field, that is, all galaxies at all redshifts. This approach gives us a sense of the range of values expected. For the subsequent analysis, we apply a cut of $D < 60$ Mpc to select galaxies at mostly similar distances (similar to the samples in Reynolds et al. 2020; Holwerda et al. 2011d). This distance cutoff ensures the larger features in the H 1 discs are included in the morphometric calculation; WALLABY's spatial resolution of 30" $\simeq$ 10 kpc at this distance. We intentionally do not select known group members because eventually we

https://doi.org/10.1017/pasa.2025.5

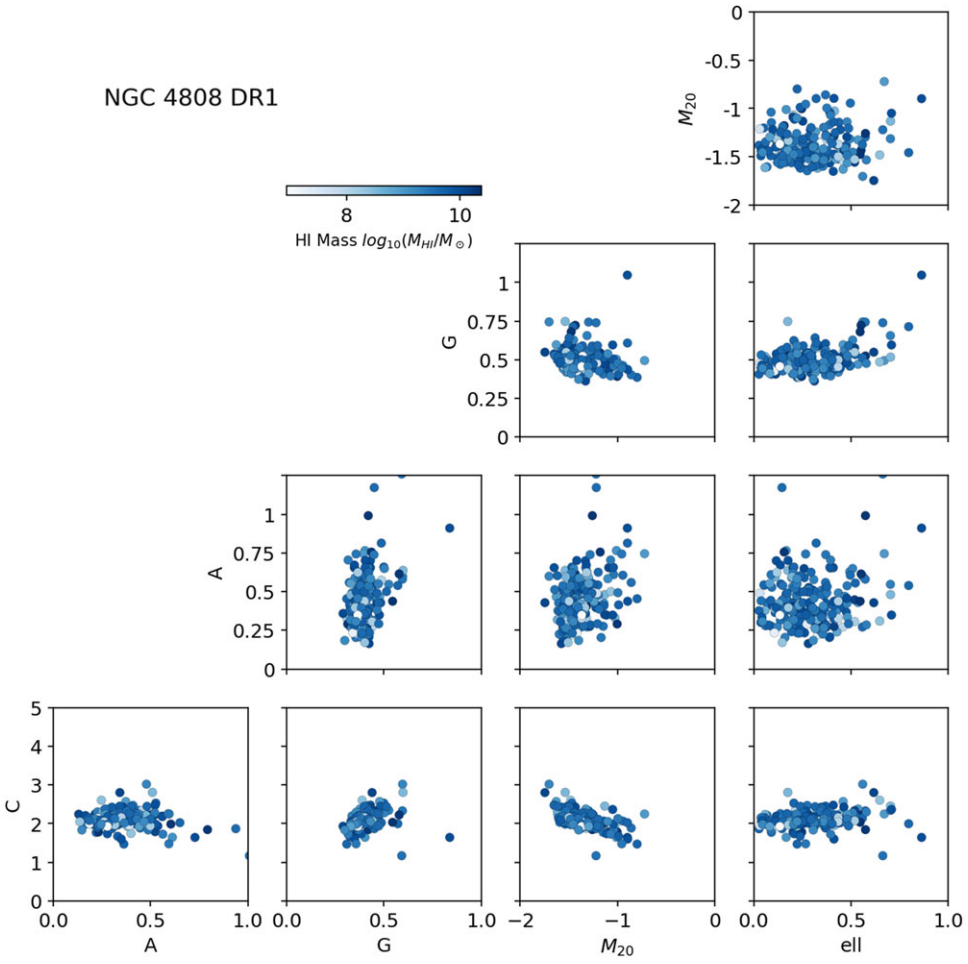

Figure 3. A corner plot of the H 1 morphometrics of galaxies in the NGC 4808 pointing based on the SoFiA segmentation maps.

hope to apply this technique on WALLABY blindly, without prior knowledge of group membership.

Because we do not have full intuition which morphometrics are the optimal feature space to train a machine learning algorithm on –even after the initial work in Holwerda et al. (2023) – we start with the full morphometric space provided by STATMORPH. We do know that Smoothness and Intensity are likely too dependent on the smoothing kernel to be of use in this lower spatial resolution data (see Section 4). This in a way is limiting since there could be other morphometrics much better suited for the identification of perturbed H 1 discs. It could be possible to define entirely new ones, perhaps including kinematic information as well (cf Deg et al. 2023). For now, we adopt the morphometric space provided by STATMORPH with our addition of Patchiness.

Figs. 2 through 4 show corner plots of the most commonly used morphometrics (modelled after the corner plot in Scarlata et al. 2007). There are some correlations between Concentration and Gini or Concentration and M_{20} evident, something noted by Conselice (2008) and Lotz et al. (2008). This morphometric space is not an orthogonal one, especially not with lower resolution data. An orthogonal space would be the easiest to train a machine learning algorithm on and engineer the feature space. Thanks to a large body of work applying these morphometrics to data from ultraviolet through radio wavelengths, the morphometric space is a familiar one to astronomy.

5. Machine learning

Our approach to these data-sets is to use the objects in the two Virgo fields (NGC 4808 and NGC 4636) as the training set. We have a series of labels for this set from Lin et al. (2023) which can be converted to a simplified flag. Trained on the training sample, we can then exploit first how well classification works (train and test) and then deploy the classifier on the other galaxies in and near these three groups.

5.1 Training sample

To construct a training sample, we require WALLABY H 1 morphometrics and a label. For the labelling, we use the sample from Lin et al. (2023) who classified galaxies in this field using FAST and WALLABY information. We crosscorrelated the catalogue of Lin et al. (2023) with both the NGC 4636 and NGC 4808 fields using an arcsecond. We found an overlap with the 63 sources from Lin et al. (2023) with the WALLABY catalogues of 21 and 15 sources in the NGC 4636 and NGC 4808 fields, respectively. We impose

https://doi.org/10.1017/pasa.2025.5

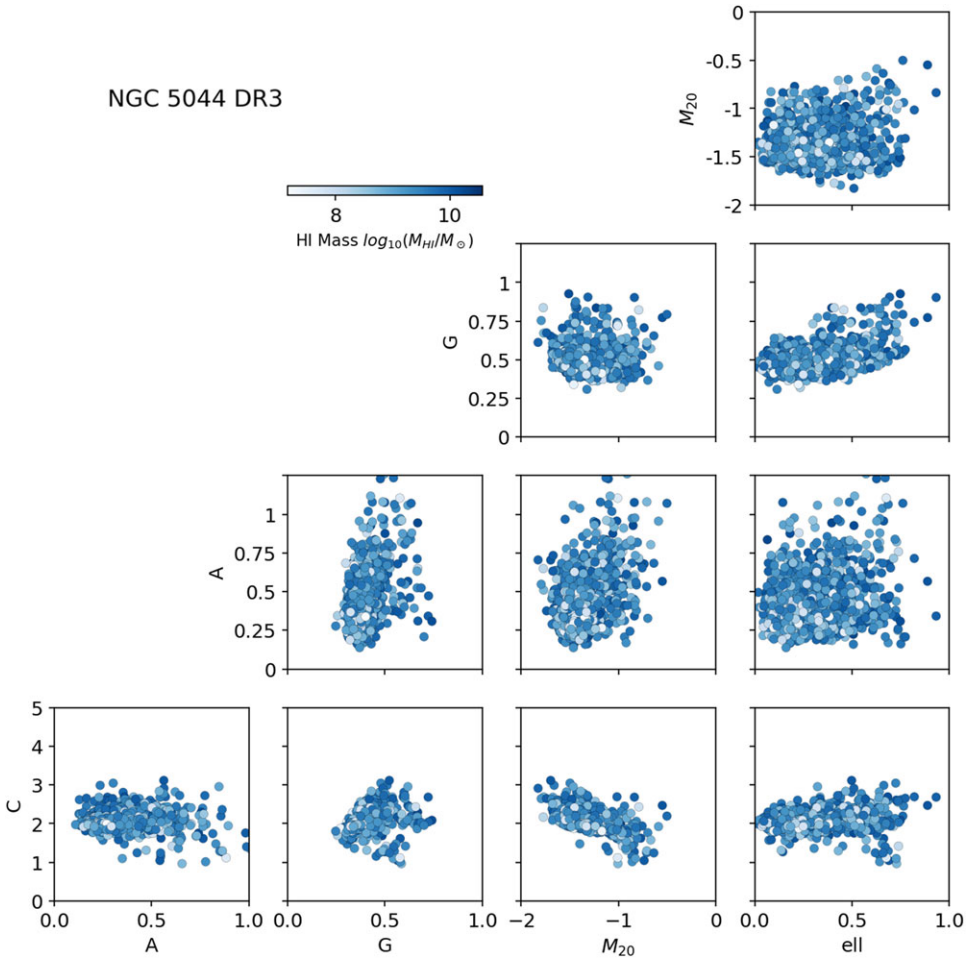

Figure 4. A corner plot of the H 1 morphometrics of galaxies in the NGC 5044 pointing based on the SoFiA segmentation maps.

our distance limit of 60 Mpc, arriving at a training sample with Lin et al. (2023) labels of 29 sources and 57 WALLABY sources without a label but within that distance and the area of the catalogue (the green circle in Fig. 5). These unlabelled WALLABY sources are considered to be 'non-perturbed'. Combined, these form our training sample.

The final training sample is 29 WALLABY sources with some sort of perturbance and 57 WALLABY sources without the perturbed label. This is a reasonably size and balanced training/test sample which can be complemented using SMOTE[c] (Synthetic Minority Oversampling Technique, Kegelmeyer 2002) to fully balance the training sample. The galaxies outside the green circle in Fig. 5 as well as all the objects in the NGC 5044 catalogue are our 'deployment' sample: the sources the trained algorithm will be deployed on for independent classification.

Fig. 6 shows the H 1 morphometric feature space with the label from Lin et al. (2023) for the galaxies that are undergoing ram-pressure stripping (flag=1), a tidal interaction (flag=2), or a gravitational merger (flag=3). The perturbed sample is spread throughout the full H 1 morphometric space, preempting any possibility to simple cuts in parameter space to separate the two labels. As noted above, the morphometric feature space is degenerate.

With a limited size training sample, a feature set that is degenerate and no good preset hyperparameter for the ML algorithm (the number of neighbours in this case), we will explore the feature engineering and hyperparameter settings, first separately and then combined.

For the metrics on performance, we will use precision, recall, and F1. Starting with True Positive (TP), True Negative (TN), False Positives (FP), and False Negatives (FN), precision is defined as: $precision = \frac{TP}{TP+FP}$ and recall as: $recall = \frac{TP}{TP+FN}$. F1 is a combination of these: $F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall}$.

5.2 Feature engineering

Because of the size of the training set, we must be extra careful to select a feature space from the H 1 morphometrics available. Because the undisturbed and perturbed galaxies lie well mixed in the H 1 feature space, kNN or a random forest (RF) make the most sense to test on this feature space. Each iteration, before we train on this set, we apply SMOTE to balance and then the built-in STANDARDSCALER in SK-LEARN to whiten (normalise) the data.

First, we examine how many features we will need. Naively, one would use the full H 1 morphometric space but there is a point

[c]A common technique to balance a training set by re-sampling the under-represented label.

https://doi.org/10.1017/pasa.2025.5

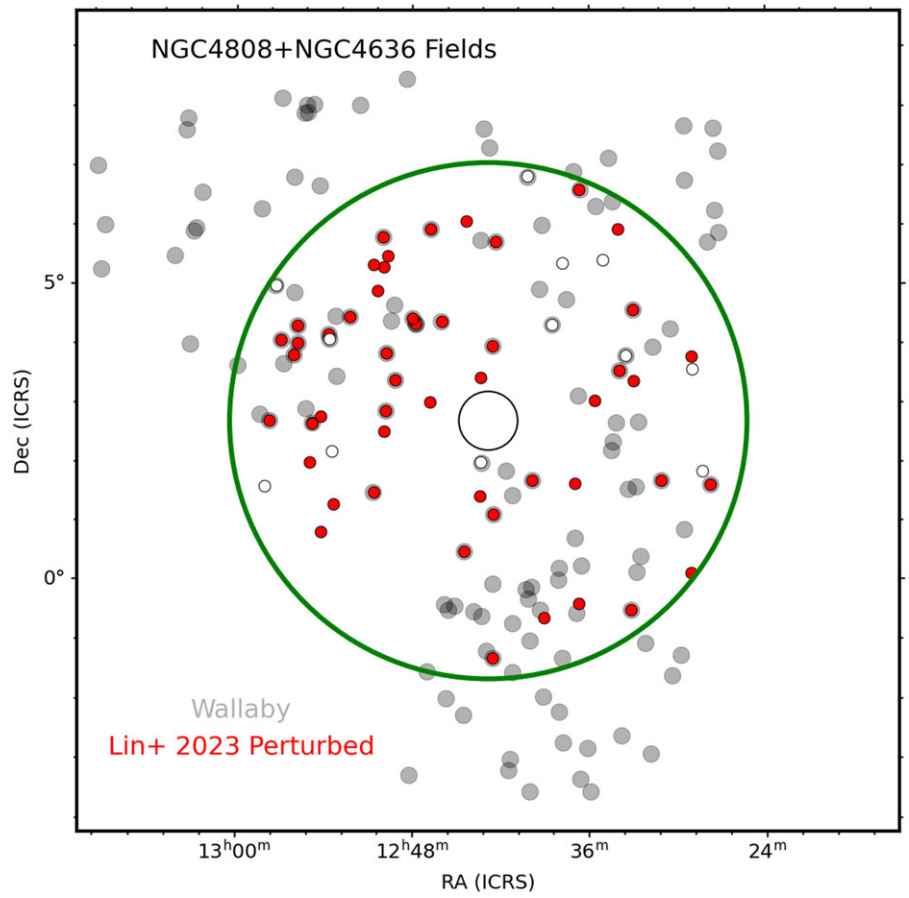

Figure 5. The WALLABY detections for both the NGC 4808 and NGC 4636 fields within (60 Mpc) in grey. Superimposed is the catalogue from Lin et al. (2023) with the perturbed (red circles) and unperturbed (white circles). Not every source in Lin et al. (2023) has a counterpart in the two WALLABY catalogues but a sufficient number is available for training. Because the Lin et al. (2023) catalogue is based on different data, we select all the sources within the green circle to be used as the WALLABY training sample with those without a Lin et al. (2023) classification deemed 'unperturbed'.

of limited return as this is a known degenerate parameter space (see Scarlata et al. 2007). At some point, one would no longer provide new information, just artificially weigh in on features already provided in another format. If we use the built-in function SELECTKBEST in SKLEARN, and ask for the 6 highest performing features for the interaction label, we arrive at Concentration, Gini, M_{20}, Multimode, Deviation, and Sérsic index (n). This validates our initial suspicion that Smoothness and Intensity do not hold much additional information in this data and are too dependent on the smoothing kernel size.

5.3 Hyperparameter optimisation

Fig. 7 shows the number of neighbours used and the different metrics. There is a notable interSection at $k = 2$ and $k = 6$ when using the full parameter space. Here, the metrics are very similar, while at $k = 1,3,4$, and 5, the trade-off between recall and precision is overly skewed in favour of recall. This is not clearly reflected in $F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$. Ideally we would keep the number of neighbours low since we are dealing with a small training set. The high number of neighbours ($k = 6$) would average over a large fraction of the training set every time. A single neighbour suffers from high variance in the classification and affect reliability, essentially over-fitting.

We examine the kNN mean and variance of all the metrics by running multiple iterations with a number of features, randomly selected and a setting for the hyper-parameter (k), the number of neighbours (Fig. 8).

Optimisation of both the hyperparameter (k, number of neighbors) and the feature space, specifically how many features to use, depends on which metric is considered more valuable. Does one want high precision (accurate classifications) or a high recall (reliable classifications) and the F1 metric is meant to reflect a balance between the two. Historically, for merger statistics using morphometrics and other techniques, a high precision was valued since the merger fraction was the aim of the study. However, with more detailed individual galaxy studies, recall may be of higher value for observational follow-up. We therefore aim to strike a balance.

To map out the balance between precision and recall here, we map both the mean value and variance as a function of the number of neighbours (k) and the number of features in Fig. 8 for each metric. The key here is not that the number of features is increased but that which are used is chosen randomly. So the training set does not automatically start with concentration and moves on from there. The mean value for a combination of neighbours and features tells us how well the kNN algorithm is performing but the variance for that combination (the right side panels in Fig. 8)

https://doi.org/10.1017/pasa.2025.5

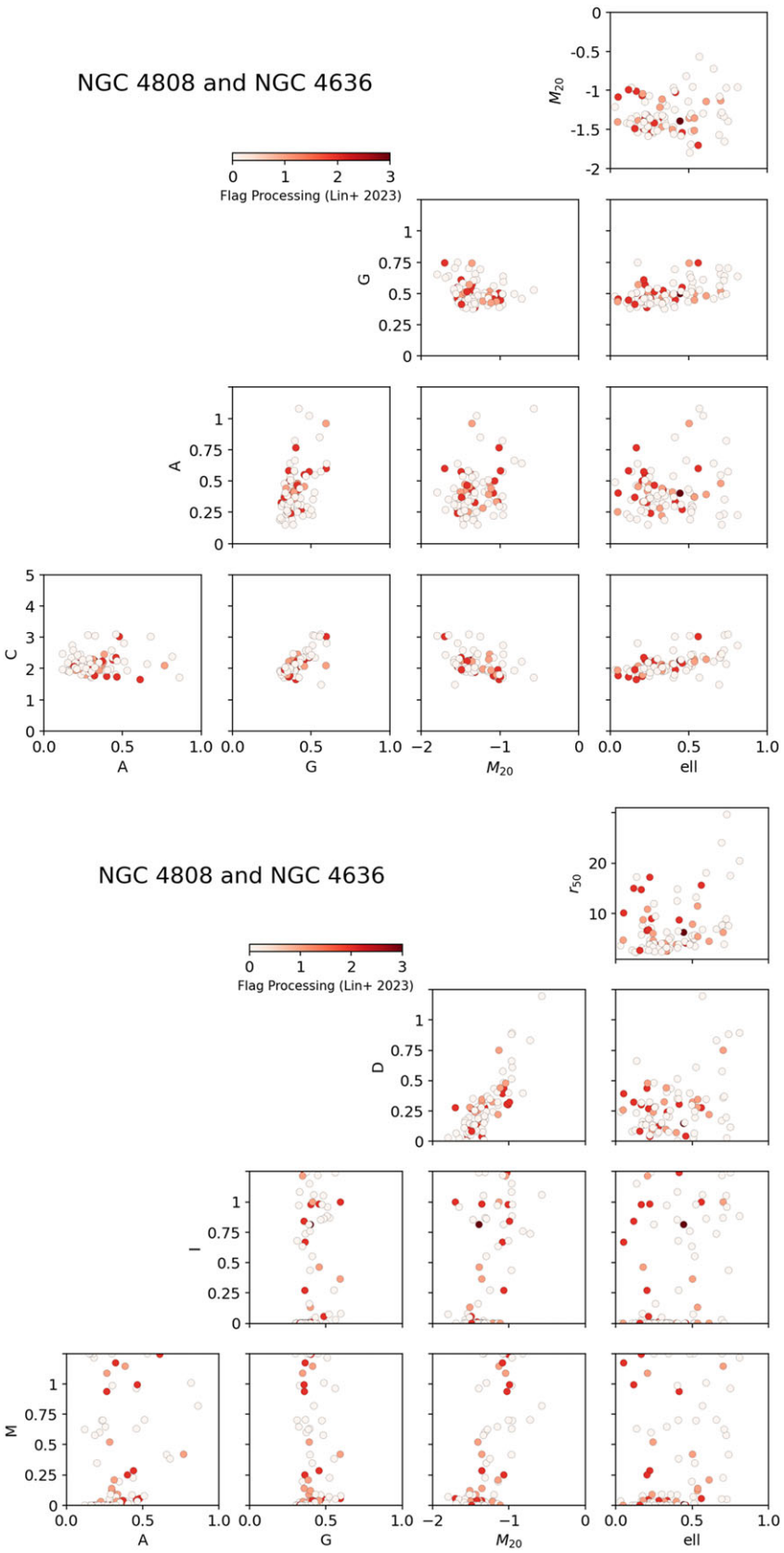

Figure 6. The processing flag for WALLABY objects according to Lin et al. (2023): ram-pressure (1), tidal interaction (2) or merger (3). We train kNN to distinguish between an undisturbed (0) and processing (1) label which includes all three here (1-3).

https://doi.org/10.1017/pasa.2025.5

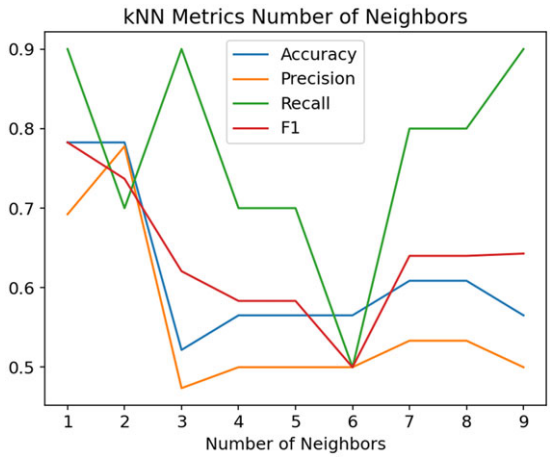

Figure 7. Hyperparameter choice for a set feature space with metrics as a function of the number of neighbours (k). This is the performance for the *full* of morphometric space.

informs us how reliable that performance is. This is missing in a simpler diagnostic plot such as Figs. 7 or 9 which concentrate on just one aspect.

Given the size of the training sample and feature space (large but not orthogonal), we opt for $k = 2$ neighbours and more than 4 features for optimal performance. This is partly motivated by Figs. 7 and 9 but validated when inspecting Fig. 8 for low variance in performance. We select those listed in Table 2 based on the experience with Hydra, Fig. 8, and which parameters are reported with high F1 scores and close precision/recall scores.

To illustrate the importance of the choice of feature space, Fig. 9 shows the metrics for the six features selected by SELECTKBEST in SKLEARN. Interestingly, this feature space performs better than the full morphometric space, and it is more consistent with metrics. The choice of $k = 2$ is still well motivated as Recall and the other metrics diverge at $k = 3$ and higher.

Similarly, one can argue which six morphometrics are preferred. For example, asymmetry is better understood and more widely adopted than the MID parameters. This could be an argument to include asymmetry instead of the multimode parameter. To ascertain the effectiveness of the kNN on this data-set, we evaluate the average of a series training-test runs, where the training/test sample split is 80%. We do this ten times. This approach is very similar to bootstrapping a simple fit. Fig. 10 shows the average confusion matrix for the test sample after training on 80% of total sample. The average metrics of this configuration (the features in Table 2 and $k = 2$) are listed in Table 3. Thanks to the repeat in kNN training/test instances, the metrics also come with a standard deviation around the mean performance. These mean performance metrics are proficient for a simple machine learning algorithm.

However, for application to other data-sets, it can be beneficial to use the entire labelled sample as the training sample. If we do this, the metrics become those in Table 4 and the confusion matrix in Fig. 11. Performance is quite good considering the size of the training sample. We will now employ this kNN (trained on the full labelled sample) on the other catalogue, the one for the NGC 5044 mosaic.

5.4 Biases

There remains the possibility of biases applying a training sample on a new data-set. The objects in the NGC 5044 mosaic are biased towards greater distances, the signal-to-noise in the different data-cubes varies due to RFI or other factors, etc. The parameterisation of morphology through the above morphometrics is meant to be mostly invariant to small changes. Aside from familiarity, this is a prominent reason to convert to morphometrics first before attempting a machine learning algorithm. Our cut in distance to just those galaxies closer than 60 Mpc is also meant to remove biases in the training and application set (e.g. there are many sources in the wide field behind the NGC 5044 group that would skew our results). Fig. 12 shows the position of the galaxies in each field closer than 60 Mpc with the kNN classification marked. That said, small differences and thus biases between data-sets may well be present. Based on the distribution of sources in the parameter space, we estimate the issue to be small. However, moving from one sample to another with a (small) fundamental difference is a known issue in machine learning known as 'transfer learning'.

5.5 Application on NGC 5044

We cut down the samples to only those galaxies below 60 Mpc for the training sample in the NGC 4808/4636 fields. The NGC 5044 field is a little further away on average but richer and well within this distance limit. The rationale for the distance limit is that it generously includes all labelled galaxies while removing the majority of unresolved background objects (Fig. 1). The resulting sample is 258 galaxies (within $D < 60$ Mpc) for the NGC 5044 field. In the comparisons in scaling relations, we will compare the training sample scaling relation to this deployment sample of NGC 5044 mosaic.

6. Results

6.1 Fraction of perturbed galaxies

Table 5 lists the fraction of the galaxies below 60 Mpc. in each of the three groups that the kNN trained on the Lin et al. (2023) classifications as perturbed somehow (i.e. ram-pressure stripping, tidally disturbed or merging). We compare these to 'perturbed' criteria from the literature, similar to Holwerda et al. (2011d).

The fraction of perturbed galaxies in the catalogue of Lin et al. (2023) is slightly lower than what we find using kNN. In general, the kNN finds a similar fraction of galaxies perturbed in each field. Based on the metrics listed as listed in Table 4, one would expect these fractions to be accurate to within a few percentage points. The difference of ∼1% in NGC 4808 is therefore illustrative of what the uncertainty should be.

Previous uses of morphometrics used a simple criterion to separate perturbed from unperturbed galaxies. Holwerda et al. (2011d) reviews these in the context of their use on H 1 surveys. H 1 morphology is expected to be perturbed earlier and longer during a gravitational interaction. Table 5 lists the fractions of galaxies that meet the various criteria as well. It is notable that a the basic asymmetry criterion ($A > 0.35$) identifies a similar percentage as the kNN classifier. Once compared however, that Asymmetry criterion is biased towards false negatives. Since in the past, the goal of morphometric identification of mergers

https://doi.org/10.1017/pasa.2025.5

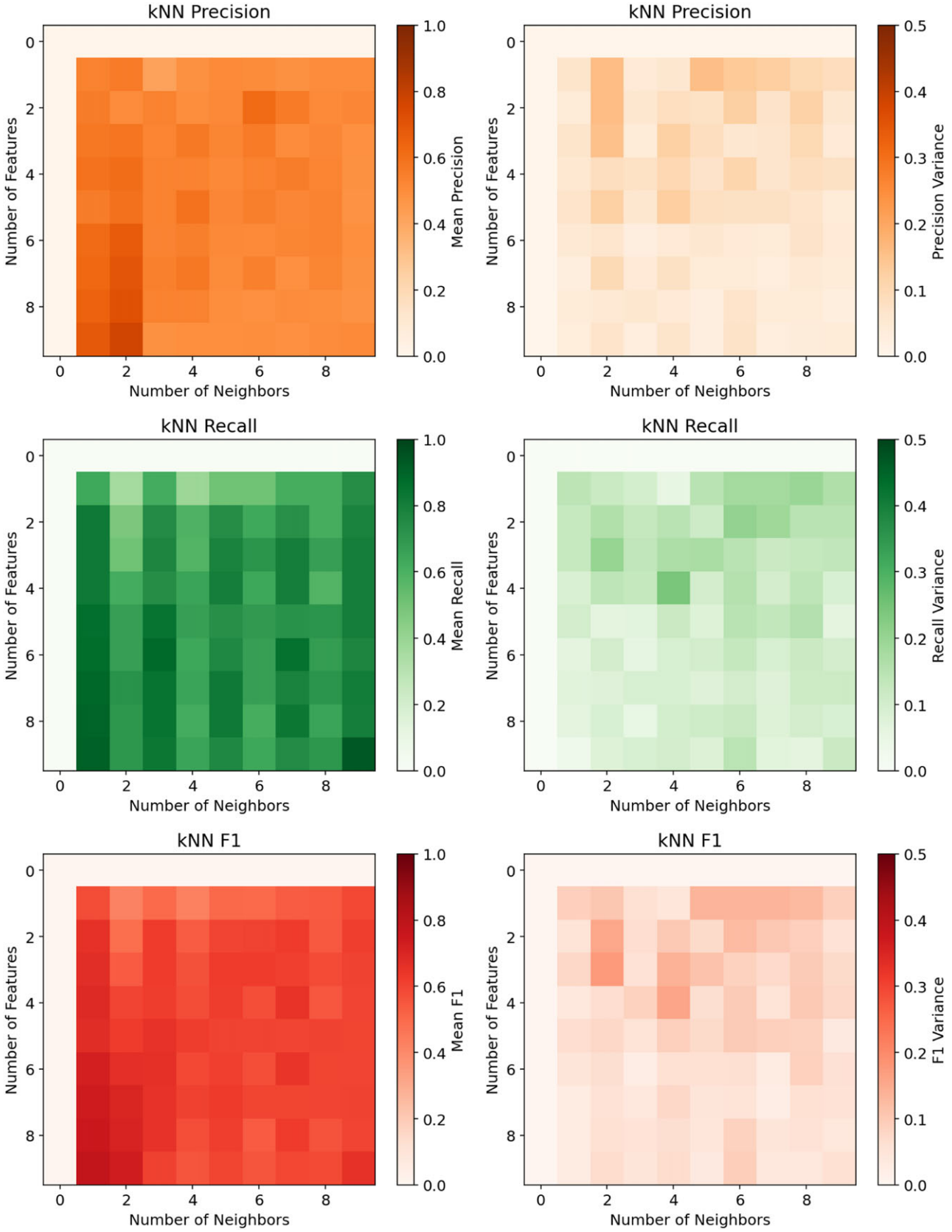

Figure 8. The mean (left row) and variance (right row) map of the precision, recall and F1. Mean and variance are determined by drawing a random set of features in the H 1 morphometric space and running the kNN on it. Variance tends to be high for $k = 1$ or $n = 2$ features.

https://doi.org/10.1017/pasa.2025.5

Table 2. The features selected for the final iteration of the kNN.

Feature	Abbreviation
Concentration	C
Asymmetry	A
Gini	G
M_{20}	M_{20}
Deviation	D
Sérsic index	n

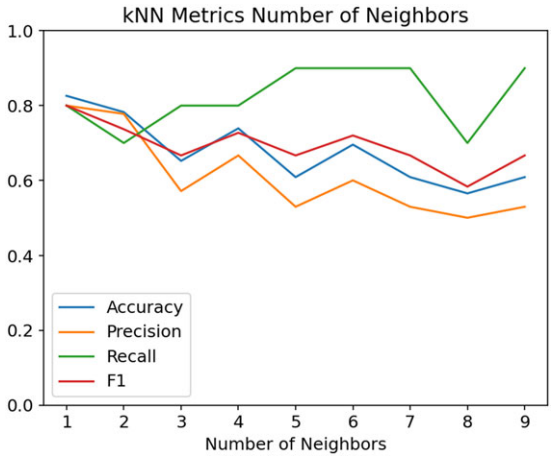

Figure 9. Hyperparameter choice for a set feature space with metrics as a function of the number of neighbours (k). This is for the optimal set of features in Table 2.

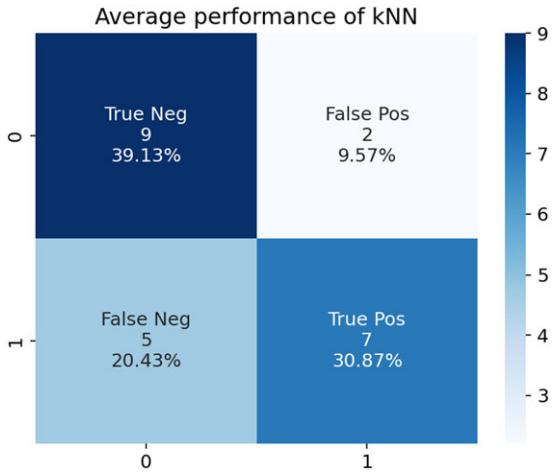

Figure 10. The average confusion matrix for the kNN ($k = 2$, trained on subsamples of 80%, tested on the remaining 20% shown here) with the optimised feature space listed in Table 2 for all the members of the NGC 4636 and NGC 4636 groups. We repeated the training/test ten times and these are the averages of all ten split-train-test iterations.

was to identify the merger fractions at different epochs or environments, the kNN approach works certainly well enough on a population.

Table 3. The performance metrics of the WALLABY training catalogue ($D < 60$ Mpc within the green circle in Fig. 5) split into subsections using the features listed in Table 2. By iterating ten times over this sample and splitting off 20% for testing, these are the mean and variance of the kNN performance.

Accuracy	73.04 ± 6.96
Precision	79.58 ± 9.03
Recall	66.67 ± 13.71
F1.	71.34 ± 4.01

Table 4. The performance metrics of in the full WALLABY training catalogue ($D <$ 60 Mpc within the green circle in Fig. 5) using the features listed in Table 2.

Accuracy	86%
Precision	88%
Recall	70%
F1	78%

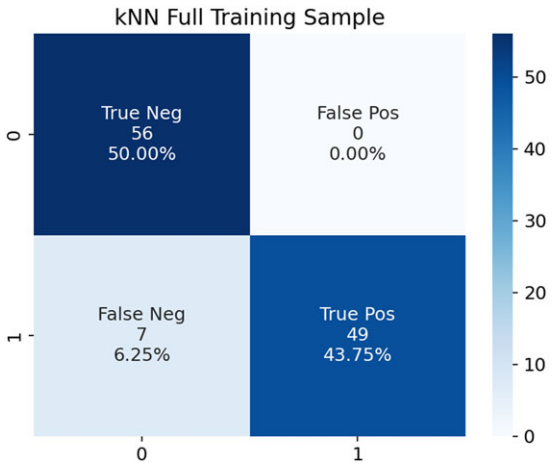

Figure 11. The confusion matrix for the kNN ($k = 2$, trained on a subsample of 80%) with the optimised feature space listed in Table 2 for all the objects in the combined catalogue of the NGC 4808 and NGC 4636 fields ($D < 60$ Mpc within the green circle in Fig. 5).

6.2 Galaxy scaling relations

One application of a ML classifier is to rapidly classify galaxies to then examine the galaxy scaling relations for those galaxies undergoing some interaction to those that are not. Here we examine three: the star-forming galaxy main sequence, the H 1 and stellar mass relation, and the Baryonic Tully-Fisher relation. We also looked at the H 1 size-mass relation but there is little difference between galaxies marked perturbed and not. The lack of an H 1 size-mass relation can be attributed to the still relatively low spatial resolution of the WALLABY pilot observations, expected to improve, and relatively simple size measures.

https://doi.org/10.1017/pasa.2025.5

Table 5. The fraction of galaxies that were perturbed as reported by Lin et al. (2023) and the kNN trained on NGC 4808+4636 (WALLABY training sample). For comparison, the morphometric criteria for merging or perturbed galaxies from Conselice (2003), Lotz et al. (2004, (2008), and Holwerda et al. (2011d) are listed as well.

	Perturbed percentage			
Criterion	NGC 4808	NGC 4636	NGC 5044	Reference
Lin+ (2023)	33.33	–	–	Lin et al. (2023)
WALLABY training sample	34.88	–	–	
kNN	21.35	33.33	21.71	This work
$A > 0.38$	35.42	26.97	27.91	Conselice (2003)
$G > -0.115 \times M_{20} + 0.384$	6.25	4.49	4.26	Lotz et al. (2004)
$G < -0.15 \times M_{20} + 0.33$	6.25	6.74	4.65	Lotz et al. (2008)
$A < -0.2 \times M_{20} + 0.25$	95.83	87.64	88.37	Holwerda et al. (2011d)
$C > -5 \times M_{20} + 3.$	0.00	0.00	0.00	Holwerda et al. (2011d)

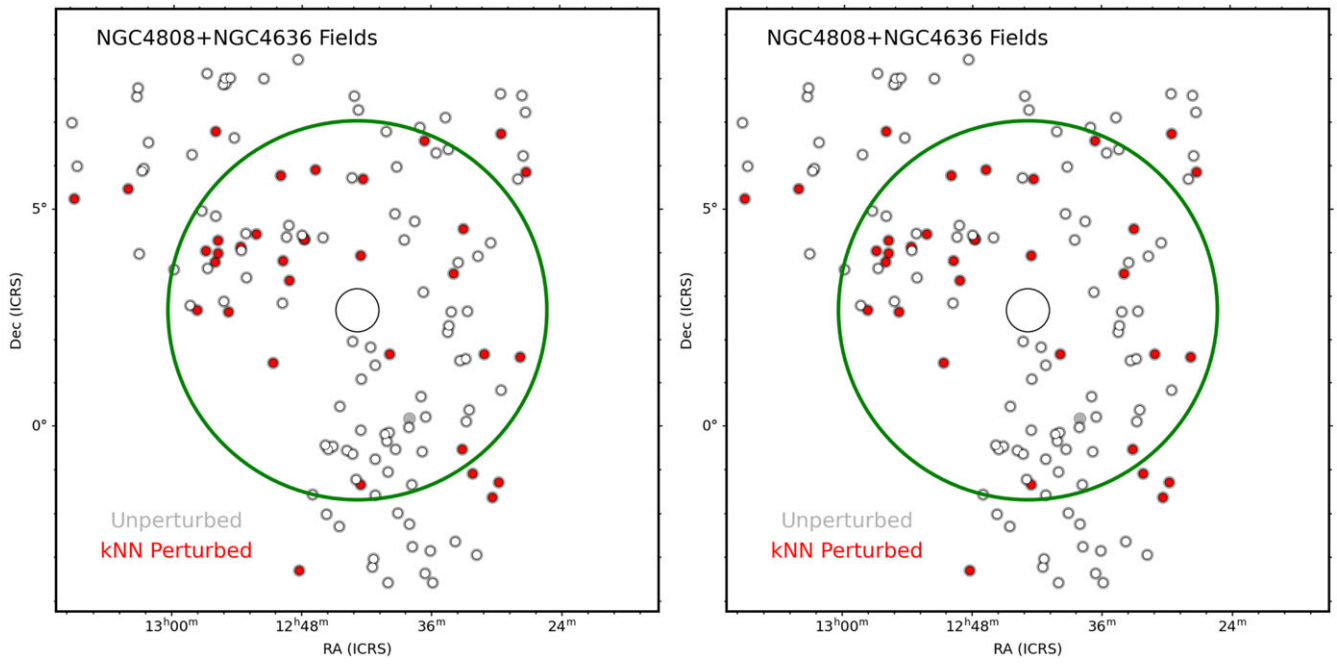

Figure 12. The kNN labelling in both the training sample (left) and the deployment field, NGC 5044. Compare to the labels in Fig. 5.

6.2.1 Star-forming main sequence

The star-forming galaxies main sequence (e.g. Noeske et al. 2007) is an important relation between the stellar mass of galaxies and their (relative) growth rate.

Fig. 13 shows the stellar mass and star formation relation for the WALLABY training sample and the deployment data in the NGC 5044 mosaic. The kNN-identified perturbed galaxies are mixed in with the main sequence of star-forming galaxies. A linear fit to the stellar mass and star formation relation for these galaxies, all of whom are on the star-forming main sequence, is essentially the same for perturbed and non-perturbed sets (Table 6). We note there is a normalisation difference between the training and deployment sample for the SFR estimate from WISE. It is unclear if this is a distance effect, or additional flux in WISE W3 due to Galactic Cirrus. The training and NGC5044 samples show the same slope and intercept within their respective bootstrap errors (Table 6).

There is no functional difference in the slope and intercepts between perturbed and unperturbed galaxies. There is a difference between training and deployment samples but that is to be expected when moving to a sample with a difference mass range.

6.2.2 Stellar and HI mass

Fig. 14 shows the stellar mass, as derived from the WISE W1 flux, and the H 1 mass from the WALLABY catalogue for the training sample and the deployment data of the NGC 5044 mosaic. We note that a quantitative comparison with existing relations (e.g. Catinella et al. 2018) is not done here because stellar mass estimates are based on catalogue photometry on a single filter from WISE. Qualitatively, the correlations for this relation are similar

https://doi.org/10.1017/pasa.2025.5

Table 6. The linear fits to the stellar mass and star formation relation for the training sample and the deployment sample of NGC 5044 for all the galaxies in the sample, the unperturbed and perturbed ones. Qualitatively the fits are similar but the deployment fits have lower slopes and higher intercepts than the training sample.

	Training		NGC 5044	
Sample	Slope	Intercept	Slope	Intercept
All	0.56 ± 0.11	-5.83 ± 0.83	0.87 ± 0.03	-6.38 ± 0.28
Unperturbed	0.60 ± 0.48	-6.40 ± 3.27	0.85 ± 0.04	-6.15 ± 0.33
Perturbed	0.55 ± 0.12	-5.66 ± 0.88	0.94 ± 0.07	-6.97 ± 0.54

for training and deployment samples; unperturbed galaxy relation has a higher slope, the perturbed ones a lower slope than the whole sample fit. Table 7 quantifies this with bootstrapped errors. We note here that the training set skews a little lower mass than the deployment sample.

6.2.3 Baryonic Tully–Fisher relation

Fig. 15 shows the Baryonic Tully–Fisher relation for all three groups combined. We used the WISE W1 based stellar mass and a factor 1.33 to convert the H 1 mass into a total gas mass including Helium. The H 1 velocity is the W50 measurement corrected for inclination from the SoFiA measurements:

$$V_{HI} = \frac{w50}{\sqrt{1 - \left(\frac{b}{a}\right)^2}} \quad (12)$$

The slope and the intercept were fit with a standard linear regression. Because the uncertainty in the Baryonic mass is under-estimated with the formal uncertainties, we estimate the variance in the slope and intercept using a bootstrapping of the fits. The BTF linear fit through all the galaxies and those labelled

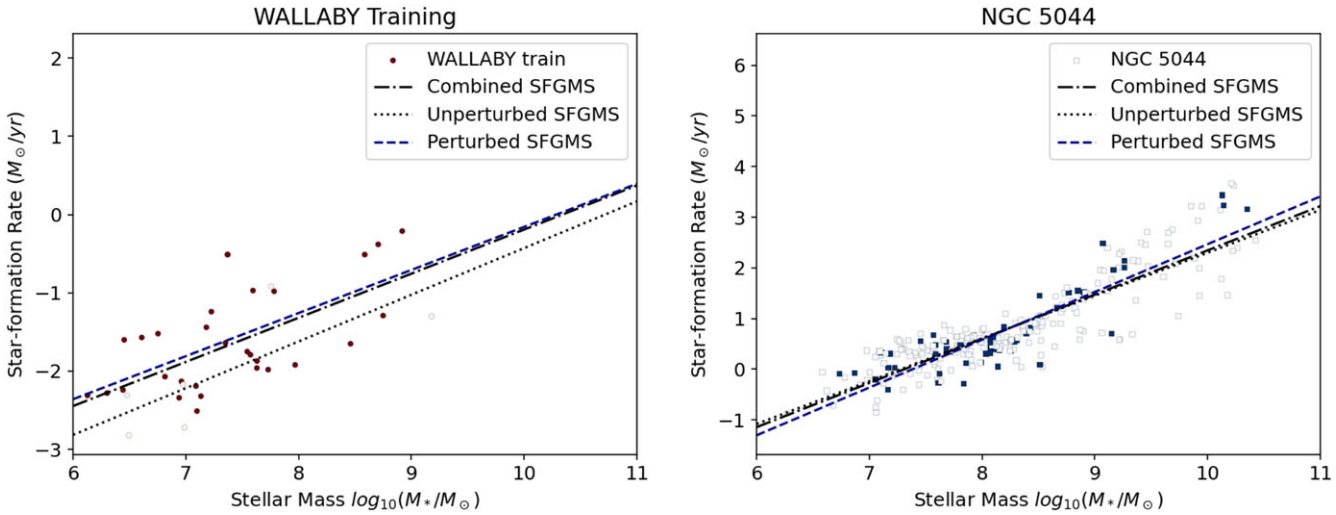

Figure 13. The stellar mass and star formation relation for the WALLABY training sample (left) and the NGC 5044 deployment sample (right). Qualitatively, the results are similar for the star-forming galaxy main sequence: similar slopes for all three populations, unperturbed, perturbed, and all galaxies, but there are quantitative differences in the SFGMS slope and interecept between the training and the deployement samples.

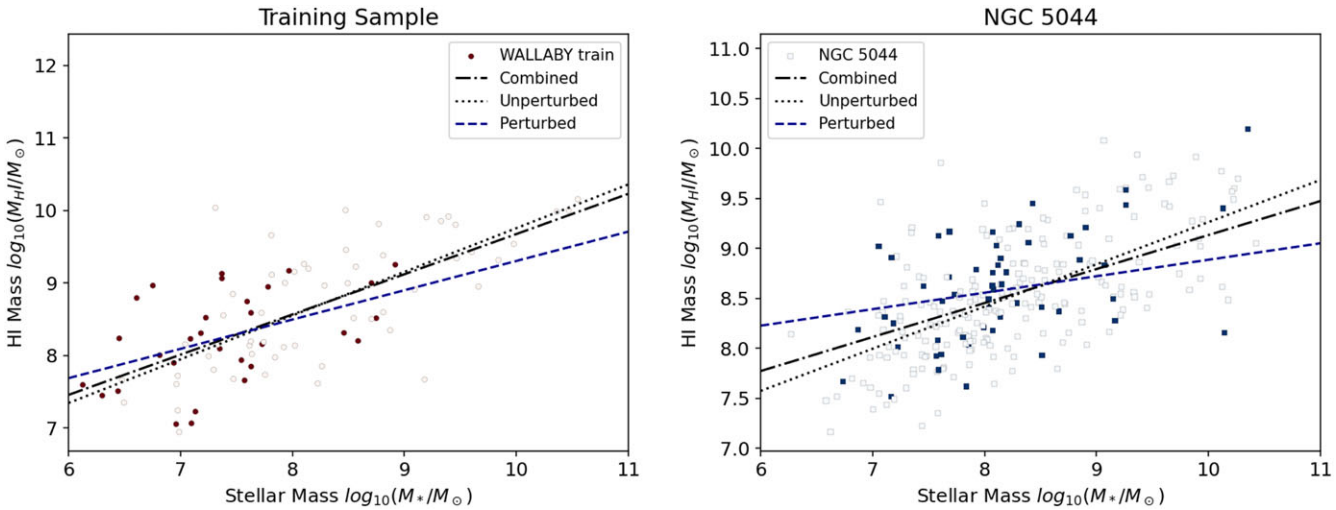

Figure 14. The stellar mass and H 1 mass relation for the training sample (left) and the deployment sample, the NGC 5044 mosaic. Both the combined and the unperturbed samples show very similar fits and the galaxies indicated as perturbed in the training sample as well as in the NGC 5044 mosaic both show less H 1 mass for a given stellar mass.

https://doi.org/10.1017/pasa.2025.5

Table 7. The linear fits to the stellar and H 1 mass relation for the training sample and the deployment sample of NGC 5044 for all the galaxies in the sample, the unperturbed and perturbed ones. Qualitatively the fits are similar but the deployment fits have lower slopes and higher intercepts than the training sample.

	Training		NGC 5044	
Sample	Slope	Intercept	Slope	Intercept
All	0.55 ± 0.05	4.15 ± 0.45	0.34 ± 0.07	5.73 ± 0.57
Unperturbed	0.60 ± 0.07	3.76 ± 0.57	0.42 ± 0.04	5.03 ± 0.32
Perturbed	0.40 ± 0.11	5.26 ± 0.85	0.17 ± 0.16	7.19 ± 1.29

Table 8. The linear fits to the Baryonic Tully–Fisher relation for the training sample and the deployment sample of NGC 5044 for all the galaxies in the sample, the unperturbed and perturbed ones.

	Training		NGC 5044	
Sample	Slope	Intercept	Slope	Intercept
All:	1.82 ± 0.14	-1.32 ± 0.76	1.47 ± 0.10	0.72 ± 0.56
Unperturbed:	1.96 ± 0.15	-1.95 ± 0.84	1.45 ± 0.11	0.83 ± 0.60
Perturbed:	1.13 ± 0.26	2.25 ± 1.46	1.54 ± 0.30	0.39 ± 1.64

unperturbed are very similar but the kNN-identified perturbed population shows a flatter BTF relation. The uncertainties reported in Table 8 are a standard deviation. The discrepancy is therefore significant.

The measurements for individual galaxies can have some uncertainties due to model fits. For example, the conversion from WISE W1 flux to a stellar mass and the correction for inclination using the SoFiA measured major and minor axes. Especially in recently perturbed galaxies, this axis ratio may not be indicative of the disc's inherent inclination.

7. Discussion

7.1 kNN performance

In this paper, we considered only the kNN classification on H 1 morphometrics, following the experiences in Holwerda et al. (2023). Generally speaking, the perturbed and unperturbed samples are well-mixed in H 1 morphometric space without a clean separation in this multidimensional space. This is reflected in the low number of neighbours which would optimise classification metrics (Fig. 7). Once the feature space is optimised, kNN behaviour is much more reasonable but still performs best with two neighbours (see Fig. 9).

Overall the kNN classification is proficient with acceptable precision and recall (Table 4). We saw a similar performance in Holwerda et al. (2023) for the Hydra cluster. The expectation is that it will identify most of the perturbed galaxies in a sample from their H 1 morphometrics this way with still some sizeable contamination, that is, the sample will be mostly complete but somewhat contaminated. This would reduce the number of galaxies that would need to be inspected visually significantly.

7.2 Galaxy scaling relations

The galaxy scaling relations for these samples are rudimentary. Stellar mass and star formation estimates are based on WISE photometry alone. For an in-depth discussion on the scaling relations for these galaxies, we refer the reader to Deg et al. (in preparation). Our aim here was to determine if there were substantial differences between the perturbed and unperturbed marked samples and how this translated from training set to deployment set. In the case of the BTF relation, there is a marked difference in the scaling relation for the training set but this disappears for the deployment. In the stellar and H 1 mass relation, there is a flatter relation for the perturbed subsample in both the training and deployment sample.

If a scaling relation trend holds with the transition from training sample to deployment, does it build confidence in the observed effect? We note that between training and deployment sample, there is a difference in distance and thus resolution (the objects in the NGC 4808 and 4636 fields are closer). And that the training sample is still fairly small for training purposes. The fact that the BTF is functionally identical in NGC 5044 irrespective of class, could be attributed to these issues of distance and training sample size. But the persistence of flatter relation between stellar and H 1 has more the appearance of an inherent

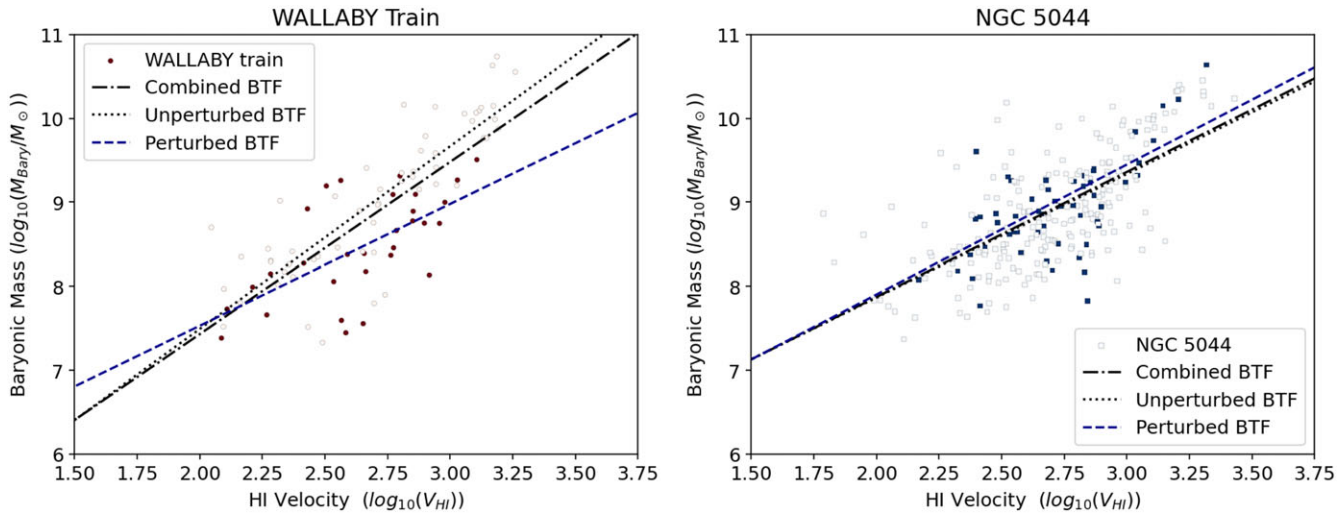

Figure 15. The Baryonic Tully–Fisher relation for the WALLABY training sample (left) and the deployment sample on NGC 5044 (right). The velocity is computed according to equation 12 with the unit in m/s.

https://doi.org/10.1017/pasa.2025.5

difference between perturbed looking galaxies and unperturbed appearing ones.

8. Conclusions

We present a H 1 morphometrics catalogue for three WALLABY fields (centred on NGC4636, 4808, and 5044) observed as part of the early pilot WALLABY observations with ASKAP.

The NGC 5044 mosaic shows the greatest diversity of H 1 morphologies and hence morphometrics. This is the richest catalogue with a substantial number of unresolved detections well beyond the central object's distance.

The NGC 4636 field has been studied in detail in Lin et al. (2023) using a mix of WALLABY and FAST data to identify those galaxies which are undergoing ram-pressure stripping and gravitational interactions such as tidal interactions or full mergers. All these are anecdotally already known to cause mild to severe changes in the H 1 morphology. Using their flags for the three phenomena (ram-pressure stripping, tidal interaction, and mergers) as an all-encompassing 'perturbed' label for both the NGC 4636 and the neighbouring NGC 4808 field, we trained a nearest neighbours algorithm, using 2 neighbours and 6 features in the morphometrics space. The training sample is small but optimised like this performs reasonably well (Table 4), minimising variance as much as practical. Exactly which 6 features remains somewhat undetermined as kNN performs well with different combinations but the six in Table 2 are our choice for this paper.

The kNN classifier, even trained on a relatively small training sample performs well in the identification of the perturbed population, enough to identify the fraction of galaxies affected and identify individual galaxies with reasonable confidence. It is a marked improvement on a simple selection criterion based on one or two H 1 morphometrics, both in stability of the identified fraction as accuracy.

Applying this kNN classifier on the objects in the three WALLABY fields within a distance of 60 Mpc, we find 'perturbed' populations in all three, mixed with the unperturbed population in most of the galaxy characteristics. The star-forming main sequence, to which most of these galaxies belong, is functionally the same for the perturbed and non-perturbed populations. The fact that both populations are well mixed-together in position points to short time-scale effects, that is, localised ones for the source of the perturbance, not throughout the field.

We construct scaling relations for training and deployment samples using WISE W1 and W3 fluxes as proxies for stellar mass and star formation rate and the SoFiA output. These are somewhat less precise as the scaling relations in Deg et al. (in preparation.) but are only to be used to compare between the 'perturbed' and 'unperturbed' classes. The perturbed population does have a lower lower H 1 mass compared to the stellar mass. The other scaling relations are indistinguishable from each other. We note that the Baryonic Tully–Fisher relation for the training sample shows a difference, while the deployment sample, the NGC 5044 field objects, does not, likely a result of the (still) low number statistics in the training sample.

Our main result is a prediction for a study similar to that of Lin et al. (2023): a list of candidate perturbed galaxies in the NGC 5044 mosaic. Once a similar study is conducted, a training sample for full deployment on all of WALLABY H 1 morphometrics will be available.

Acknowledgement. This scientific work uses data obtained from Inyarrimanha Ilgari Bundara/the Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamaji People as the Traditional Owners and native title holders of the Observatory site. CSIRO's ASKAP radio telescope is part of the Australia Telescope National Facility (https://ror.org/05qajvd42). Operation of ASKAP is funded by the Australian Government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Research Centre. Establishment of ASKAP, Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory and the Pawsey Supercomputing Research Centre are initiatives of the Australian Government, with support from the Government of Western Australia and the Science and Industry Endowment Fund.

Parts of this research were supported by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013.

This research made use of Astropy, a community-developed core Python package for Astronomy (Astropy Collaboration et al. 2013; Astropy Collaboration et al. 2018).

Funding statement. N.K.Y. acknowledges the China Postdoctoral Science Foundation (2022M723175, GZB20230766).

Data availability statement. All ASKAP data products are publicly available in the CSIRO ASKAP Science Data Archive (CASDA[d]).

ALL WALLABY PDR1 data is publicly available at WALLABY PDR1. The kinematic modelling proto-pipeline is available at WKAPP code. The H 1 morphometric catalogues are available with this paper. The specific morphometric and kNN analysis scripts are available upon request.

References

Abraham, R. G., Valdes, F., Yee, H. K. C., & van den Bergh, S. 1994, ApJ, 432, 75
Abraham, R. G., van den Bergh, S., & Nair, P. 2003, ApJ, 588, 218
Ashley, T., Simpson, C. E., Elmegreen, B. G., Johnson, M., & Pokhrel, N. R. 2017, ArXiv e-prints AJ, 153(3):132, March 2017, DOI: 10.3847/1538-3881/aa5ca7
Astropy Collaboration, et al. 2013, A&A, 558, A33
Astropy Collaboration, et al. 2018, AJ, 156, 123
Begeman, K. G. 1989, A&A, 223, 47
Bershady, M. A., Jangren, A., & Conselice, C. J. 2000, AJ, 119, 2645
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bigiel, F., Leroy, A., & Walter, F. 2011, in Computational Star Formation, Vol. 270, ed. J. Alves, B. G. Elmegreen, J. M. Girart, & V. Trimble, AJ, 327–334
Bignone, L. A., et al. 2017, MNRAS, 465, 1106
Boomsma, R., Oosterloo, T. A., Fraternali, F., van der Hulst, J. M., & Sancisi, R. 2008, A&A, 490, 555
Bosma, A. 1978, PhD thesis, University of Groningen, Netherlands
Buote, D. A., Brighenti, F., & Mathews, W. G. 2004, ApJ, 607, L91
Buote, D. A., Lewis, A. D., Brighenti, F., & Mathews, W. G. 2003, ApJ, 595, 151
Catinella, B., et al. 2018, MNRAS, 476, 875
Chabrier, G. 2003, PASP, 115, 763
Chamba, N., Trujillo, I., & Knapen, J. H. 2022, A&A, 667, A87
Cluver, M. E., et al. 2014, ApJ, 782, 90
Conselice, C. J. 2003, ApJS, 147, 1
Conselice, C. J. 2008, in Astronomical Society of the Pacific Conference Series, Vol. 390, Pathways Through an Eclectic Universe, ed. J. H. Knapen, T. J. Mahoney, & A. Vazdekis, 403
Courtois, H. M., et al. 2023, MNRAS, 519, 4589
Davenport, J. R. A. 2015, I really want to find an astronomical application for morphometrics, https://twitter.com/jradavenport/status/571064841344917504

[d]https://data.csiro.au/.

de Blok, W. J. G., et al. 2008, AJ, 136, 2648
de Blok, W. J. G., et al. 2020, A&A, 643, A147
de los Reyes, M. A. C., et al. 2024, arXiv e-prints, arXiv:2409.03959
Deg, N., et al. 2023, MNRAS, 523, 4340
Elson, E. C., de Blok, W. J. G., & Kraan-Korteweg, R. C. 2011, MNRAS, 415, 323
Ferguson, H. C., & Sandage, A. 1990, AJ, 100, 1
Ferguson, H. C., & Sandage, A. 1991, AJ, 101, 765
Fetherolf, T., et al. 2023, MNRAS, 518, 4214
Florian, M. K., Li, N., & Gladders, M. D. 2016, ApJ, 832, 168
For, B. Q., et al. 2019, MNRAS, 489, 5723
For, B. Q., et al. 2021, MNRAS, 507, 2300
Forbes, D. A., et al. 2006, PASA, 23, 38
Freeman, P. E., et al. 2013, MNRAS, 434, 282
Giese, N., van der Hulst, T., Serra, P., & Oosterloo, T. 2016, MNRAS, 461, 1656
Gini, C. 1912
Glowacki, M., et al. 2022, MNRAS, 517, 1282
Graham, A. W., & Driver, S. P. 2005, PASA, 22, 118
Graham, A. W., et al. 2005, AJ, 130, 1535
Grundy, J. A., et al. 2023, PASA, 40, e012
Heald, G., et al. 2011a, in IAU Symposium, Vol. 277, IAU Symposium, ed. C. Carignan, F. Combes, & K. C. Freeman, 59, DOI: 10.1017/S1743921311022460
Heald, G., et al. 2011b, A&A, 526, A118
Hess, K. M., et al. 2022, A&A, 668, A184
Hibbard, J. E., van Gorkom, J. H., Rupen, M. P., & Schiminovich, D. 2001, in Astronomical Society of the Pacific Conference Series, Vol. 240, Gas and Galaxy Evolution, ed. J. E. Hibbard, M. Rupen, & J. H. van Gorkom, 657, DOI: 10.48550/arXiv.astro-ph/0110667
Holwerda, B. W. 2005, astro-ph/0512139
Holwerda, B. W., et al. 2011a, MNRAS, 416, 2426
Holwerda, B. W., et al. 2011b, MNRAS, 416, 2437
Holwerda, B. W., et al. 2011c, MNRAS, 416, 2401
Holwerda, B. W., et al. 2011d, MNRAS, 416, 2415
Holwerda, B. W., Pirzkal, N., de Blok, W. J. G., & van Driel, W. 2011e, MNRAS, 416, 2447
Holwerda, B. W., Pirzkal, N., & Heiner, J. S. 2012, MNRAS, 427, 3159
Holwerda, B. W., et al. 2023, arXiv e-prints, arXiv:2302.07963
Hotan, A. W., et al. 2021, PASA, 38, e009
Jarrett, T. H., et al. 2011, ApJ, 735, 112
Jarrett, T. H., et al. 2013, 145, 6, DOI: 10.1088/0004-6256/145/1/6
Jog, C. J. 2002, A&A, 391, 471
Jog, C. J., & Combes, F. 2009, PhyR, 471, 75
Johnston, S., et al. 2008, ExpAs, 22, 151
Kegelmeyer, N. V. C. K. W. B. L. O. H. W. P. 2002, JAIR, 16, 321
Kim, S.-J., et al. 2023, MNRAS, 519, 318
Kleiner, D., et al. 2019, MNRAS, 488, 5352
Koribalski, B. S. 2012, PASA, 29, 359
Koribalski, B. S., & López-Sánchez, Á. R. 2009, MNRAS, 400, 1749
Koribalski, B. S., et al. 2018, MNRAS, 478, 1611
Koribalski, B. S., et al. 2020, Ap&SS, 365, 118
Kourkchi, E., & Tully, R. B. 2017, ApJ, 843, 16
Lee-Waddell, K., et al. 2019, MNRAS, 487, 5248
Leroy, A. K., et al. 2008, AJ, 136, 2782
Lin, X., et al. 2023, ApJ, 956, 148
Lisker, T. 2008, ApJS, 179, 319
Lotz, J. M., et al. 2011, ApJ, 742, 103
Lotz, J. M., Jonsson, P., Cox, T. J., & Primack, J. R. 2008, MNRAS, 391, 1137
Lotz, J. M., Jonsson, P., Cox, T. J., & Primack, J. R. 2010, MNRAS, 404, 590
Lotz, J. M., Primack, J., & Madau, P. 2004, AJ, 128, 163
Malin, D. F. 1978, Nature, 276, 591
McKay, N. P. F., et al. 2004, MNRAS, 352, 1121
Meurer, G. R., Carignan, C., Beaulieu, S. F., & Freeman, K. C. 1996, AJ, 111, 1551
Meurer, G. R., Staveley-Smith, L., & Killeen, N. E. B. 1998, MNRAS, 300, 705
Moore, E. M., & Gottesman, S. T. 1998, MNRAS, 294, 353
Morganti, R. 2017, NatAs, 1, 596
Noeske, K. G., et al. 2007, ApJ, 660, L43
Noordermeer, E., van der Hulst, J. M., Sancisi, R., Swaters, R. A., & van Albada, T. S. 2005, A&A, 442, 137
Osmond, J. P. F., & Ponman, T. J. 2004, MNRAS, 350, 1511
Pearson, J., Li, N., & Dye, S. 2019, MNRAS, 488, 991
Peth, M. A., et al. 2016, MNRAS, 458, 963
Planck Collaboration, et al. 2016, A&A, 596, A100
Reiprich, T. H., & Böhringer, H. 2002, ApJ, 567, 716
Reynolds, T. N., Westmeier, T., Staveley-Smith, L., Chauhan, G., & Lagos, C. D. P. 2020, MNRAS, 493, 5089
Reynolds, T. N., et al. 2021, MNRAS, 505, 1891
Reynolds, T. N., et al. 2022, MNRAS, 510, 1716
Reynolds, T. N., et al. 2023, PASA, 40, e032
Rodriguez-Gomez, V., et al. 2019, MNRAS, 483, 4140
Scarlata, C., et al. 2007, ApJS, 172, 406
Serra, P., et al. 2015a, MNRAS, 452, 2680
Serra, P., et al. 2015b, MNRAS, 448, 1922
Sérsic, J. L. 1968, Atlas de Galaxias Australes, ed. J. L. Sérsic
Swaters, R. A., van Albada, T. S., van der Hulst, J. M., & Sancisi, R. 2002, A&A, 390, 829
Takamiya, M. 1999, ApJS, 122, 109
Tamura, T., Kaastra, J. S., Makishima, K., & Takahashi, I. 2003, A&A, 399, 497
Trujillo, I., Chamba, N., & Knapen, J. H. 2020, MNRAS, 493, 87
van Eymeren, J., Jütte, E., Jog, C. J., Stein, Y., & Dettmar, R. J. 2011a, A&A, 530, A29
van Eymeren, J., Jütte, E., Jog, C. J., Stein, Y., & Dettmar, R. J. 2011b, A&A, 530, A30
Villaescusa-Navarro, F., et al. 2016, MNRAS, 456, 3553
Walter, F., et al. 2008, AJ, 136, 2563
Wang, J., et al. 2021, ApJ, 915, 70
Wang, Y., et al. 2014, MNRAS, 440, 3100
Watts, A. B., Catinella, B., Cortese, L., Power, C., & Ellison, S. L. 2021, MNRAS, 504, 1989
Watts, A. B., et al. 2023, MNRAS, 519, 1452
Westmeier, T., Koribalski, B. S., & Braun, R. 2013, MNRAS, 434, 3511
Westmeier, T., et al. 2021, 506(3):3962–3976, September 2021, DOI: 10.1093/mnras/stab1881 arXiv: 2106.15789
Westmeier, T., et al. 2022, PASA, 39, e058
Willmer, C. N. A. 2018, ApJS, 236, 47
Yitzhaki, S. 1991, ASA, 9, 235
Yu, N., Ho, L. C., Wang, J., & Li, H. 2022, ApJS, 261, 21
Zschaechner, L. K., Rand, R. J., Heald, G. H., Gentile, G., & Kamphuis, P. 2011, ApJ, 740, 35
Zuo, P., Ho, L. C., Wang, J., Yu, N., & Shangguan, J. 2022, ApJ, 929, 15