# Identification of Anomalous E+A Galaxies in GAMA Using an Isolation Forest

Kieran Broadbelt[1], Kevin Pimbblet[1,2], and Daniel J. Farrow[1,2]
[1]E.A. Milne Centre, Faculty of Science and Engineering, University of Hull, Cottingham Road, Hull HU6 7RX, UK
[2]Centre of Excellence for Data Science, Artificial Intelligence and Modelling (DAIM), University of Hull, Cottingham Road, Hull HU6 7RX, UK

Accepted XXX. Received YYY; in original form ZZZ



**ABSTRACT**

We implement an outlier detection model, an Isolation Forest (iForest), to uncover anomalous objects in the Galaxy and Mass Assembly Fourth Data Release (GAMA DR4). The iForest algorithm is an unsupervised Machine Learning (ML) technique. The data we use is spectroscopic and photometric data from GAMA DR4, which compiles information for over 300,000 objects. We select two samples of galaxies to isolate, high signal-to-noise galaxies, to analyse the iForest's robustness, and E+A galaxies, to study the extremes of their population. We result in six sub-samples of spectroscopic, photometric and combined data isolations, finding 101 anomalous objects, 50% of which have not been identified as outliers in other works. We also find a number of fringing errors and false emission lines, displaying the iForest's potential in detecting these errors. We find anomalous E+A galaxies - that although selected in a 'normal' manner using low [OII] and strong H$\delta$ absorption - are still star-forming, with strong H$\alpha$ emission. We propose two solutions to how these E+A galaxies are still star-forming but also question if these galaxies can be truly classified as E+A galaxies. We suggest that small-scale interactions of gas poor objects cause small star bursts, but the radiative pressure when high mass star form, expels the accreting material quicker than it can be accreted. We also suggest that the Jeans limit in our anomalous E+A galaxies is so low that it is simply not possible for O and B class stars to form, but it does not entirely prevent star formation.

**Key words:** galaxies: general – galaxies: peculiar – methods: data analysis – methods: statistical

## 1 INTRODUCTION

As galaxy data expands rapidly with large surveys such as GAMA (Driver et al. 2022), Rubin LSST (Ivezić et al. 2019), DESI (Adame et al. 2024), Euclid (Euclid Collaboration et al. 2024), we are left with an immense wealth of information to process. A current example of this wealth of data, GAMA DR4 (Driver et al. 2022), compiles over 300,000 spectra and results in 253,144 reliable galaxy redshifts. Furthermore, the Sloan Digital Sky Survey (SDSS) has also provided the community with multi-coloured images of more than three million objects (York et al. 2000). These immense datasets have led to a wide range of automated tools that can characterise and classify galaxies (Baron & Poznanski 2016; Lochner et al. 2016; Clarke et al. 2020; Reza 2021; Chang et al. 2021). But, we still need more tools to probe for novel objects that can provide new information. ML algorithms are already in use to identify pre-defined objects like supernovae, irregular galaxies, active galactic nuclei (AGN), etc. (Lochner et al. 2016; Reza 2021; Chang et al. 2021) and in use to classify previously unseen data into known classes (Clarke et al. 2020). These algorithms, although effective at identifying known classes of object, do not extract the novel, anomalous objects we believe can provide new insights for astronomers.

Because of the size of astronomical datasets it is impossible for researchers to physically view and study all of the stars, galaxies, artifacts and data that is appearing. One such answer for this problem comes from outlier detection models (Liu et al. 2008; Goel & Montgomery 2015; Baron & Poznanski 2016; Margalef-Bentabol et al. 2020). Outlier detection will reveal new information that can provide insight into current questions that astronomers have. Baron & Poznanski (2016) (hereafter BP16) uses an unsupervised Random Forest (RF) algorithm on the over two million spectra in SDSS and find 400 anomalous galaxies that are further analysed and studied. Some of these objects are AGN galaxies, post-starburst galaxies, extreme starformers and more. Their unsupervised ML algorithm identifies a large number of objects that are not found through other ML techniques. BP16 also find errors in the SDSS pipeline, wrong classifications, 'bad' spectra, and unusually broad [OIII] emission lines.

The main population of unusual galaxies that we aim to study are post-starburst galaxies. These galaxies are an important link between the star-forming spiral galaxies and their evolution to quiescent elliptical/S0 galaxies (Dressler & Gunn 1982; Couch & Sharples 1987; Wilkinson et al. 2017). Post-starburst galaxies can also help in the understanding of how the environment influences the evolution of galaxies (Dressler & Gunn 1983; Zabludoff et al. 1996). One type of post-starburst galaxy in particular, is thought to be a transitionary phase between star-forming spirals and quiescent elliptical galaxies, namely, E+A galaxies. E+A galaxies were identified by Dressler & Gunn (1983) and have elliptical morphology whilst having high populations of A-class stars. These E+A galaxies display spectra with no [OII] emission but have deep Balmer (H$\delta$) absorption lines. Strong [OII] is an indicator of ongoing star formation whilst the deep Balmer lines is a sign of young, recently formed, A-class stars

(Wilkinson et al. 2017). There is a variance however on how these E+A galaxies are defined, with some observational work selecting low [OII] emission and strong H$\delta$ absorption (Poggianti et al. 2009; Vergani et al. 2010; Wilkinson et al. 2017). Other authors however, use a lack of H$\alpha$ emission in their selection (Hogg et al. 2006; Goto 2007; Wilkinson et al. 2017; Chen et al. 2019; Greene et al. 2021) to find 'pure' E+As. Wilkinson et al. (2017) compares these selection techniques, as does Greene et al. (2021), and more detail about how our selection is performed will be discussed in 3.2. The results of Wilkinson et al. (2017) shows that the cuts of low [OII] and strong H$\delta$ find a population of green valley discs, a middle step between star-forming spirals and quiescent elliptical galaxies. The additional cut on low H$\alpha$ emission shows mostly early-type red galaxies and are classified as 'pure' E+As.

In this work, an iForest (Liu et al. 2008) algorithm will be used on the spectroscopic and photometric data of our samples. The iForest is a novel approach at outlier detection that specifically targets outlying instances rather than profiling the normal instances. We will test the iForest on a sample of high S/N galaxies by applying a cut on the S/N measure. The S/N sample will be a secondary sample to the E+A sample and will consist of nearly 10,000 galaxies from GAMA that have S/N$\geq$ 8.

## 2 DATA

We take our data from GAMA DR4 (Driver et al. 2022) and the necessary data management units (DMU) required for analysis (Liske et al. 2015; Gordon et al. 2017; Bellstedt et al. 2020). As mentioned in section 1, GAMA DR4 measured over 250,000 redshifts, and, in combination with earlier surveys, results in over 300,000 spectra across five sky regions (Driver et al. 2022). GAMA also catalogues a large number of DMUs that contain a wide variety of additional galaxy properties. The DMUs we use are: SPECLINESSFRV05 (Gordon et al. 2017) that compiles the line flux and equivalent width (EW) measures for the GAMA spectra; SPECCATV27 (Liske et al. 2015) that contains the spectra images, fibre data and other observation information; APMATCHEDCATV06 (Liske et al. 2015) that contains photometric data for the SDSS $urgiz$ and VIKING $ZYJHK$ surveys; and GKVINPUTCATV02 (Bellstedt et al. 2020) that contains the photometric data in $FUV$, $NUV$, $ugri$, $ZYJHK$, $W1$ and $W2$ bands. For the spectroscopic analysis, we isolate only the EW measures, of which there are 51 features. For the photometric isolation, we isolate on the features of flux, magnitude and kcorrected colours, the calculations of which are discussed below, resulting in 326 features. The exact waves lengths, fitting procedures and other information about these values can be found in Gordon et al. (2017) for the spectroscopic measures and Liske et al. (2015) and Bellstedt et al. (2020) for the photometric measures. For our analysis we have directly taken values from the aforementioned DMUs with no additional scaling or normalization.

We utilise colour in our photometric isolations but there is no direct colour DMU in GAMA DR4. We take the magnitude values from APMATCHEDCATV06 and use the Python module KCORRECT[1] to scale the magnitudes to $z = 0$. We then calculate the colours for the SDSS $ugriz$ bands and the VIKING $ZYJHK$ bands and input them back in to the feature list alongside the magnitude and flux values for future isolation.

With outlier detection in mind, we introduce a quality selection on the redshift quality of nQ $\geq$ 3 which reduces the data to approximately 230,000 spectra. The GAMA DMU SPECLINESSFR uses the value nQ to represent redshift quality. This quality selection is to ensure our outliers are not simply noisy or messy data and it confirms that the redshift is secure (Driver et al. 2022). Gordon et al. (2017) also introduce 'dummy' values in to the spectroscopic measures. These dummy values occur where there were either: i) not enough pixels to perform a fit, ii) in the case of direct summation lines, they are in magnitudes and so logarithm of the flux returns an invalid value or iii) the flux value for molecular lines is always set at -99999.0, (Gordon et al. 2017). These dummy values would corrupt the iForest algorithm, so they are replaced with the mean of the non-dummy values in the column. This could cause false flags or incorrect outliers but we believe it to be better than simply removing all objects with dummy values.

The image data used for later morphological analysis comes from the SDSS survey, extracted via the ASTROQUERY Python package (Ginsburg et al. 2019). The SDSS survey contains approximately 3 million multi-coloured images (York et al. 2000) for use in analysis. We use these images alongside a second Python package STATMORPH (Rodriquez-Gomez et al. 2019) that has been designed and tested for calculating morphometric parameters.

## 3 METHODOLOGY

Here we will introduce in more detail the iForest being utilised in section 3.1 and two main samples we are applying it to, the E+A sample and the S/N sample. Lastly, we will discuss the morphometric parameters that will be used to support the classification of our anomalous galaxies in section 3.4.

### 3.1 Isolation Forest

Anomalies are defined as data patterns that have different characteristics from normal instances (Liu et al. 2008). The detection of these anomalies can be of significant importance when discovering novel data that can be probed for new information. In section 1, most of the existing model-based approaches to outlier detection will construct a profile of the standard instances within the data and will then identify instances that do not fit that profile. Notable examples in astronomy outlier detection include the use of Deep Generative Networks (Margalef-Bentabol et al. 2020), statistical and machine learning techniques (Goel & Montgomery 2015), and RF algorithms (BP16). Many of these existing methods are constrained to low dimensional data because of their high computational complexity (Liu et al. 2008; Tahi & Hadi 2019). This issue is something that recent literature has been attempting to overcome (Liu et al. 2018; Kamalov & Leung 2020).

Liu et al. (2008) propose a different model-based method that explicitly isolates outlying data, rather than profiling normal instances. This is achieved by utilising two quantitative properties of anomalous data: i) anomalies will be fewer than normal instances and ii) they have attribute-values that are especially different from those of normal instances (Liu et al. 2008). What this means is that anomalies by definition are going to be few and different, meaning, they are more susceptible to isolation. Using this idea, a tree structure is constructed that isolates every instance in the dataset, called an Isolation Tree (iTree). Because of the fact anomalies are more susceptible to isolation, they are more likely to be found at the base of the iTree structure, whereas normal instances would be found in the deeper 'branches'. Due to the outlying data lying close to the root of the

[1] https://kcorrect.readthedocs.io/en/stable/index.html

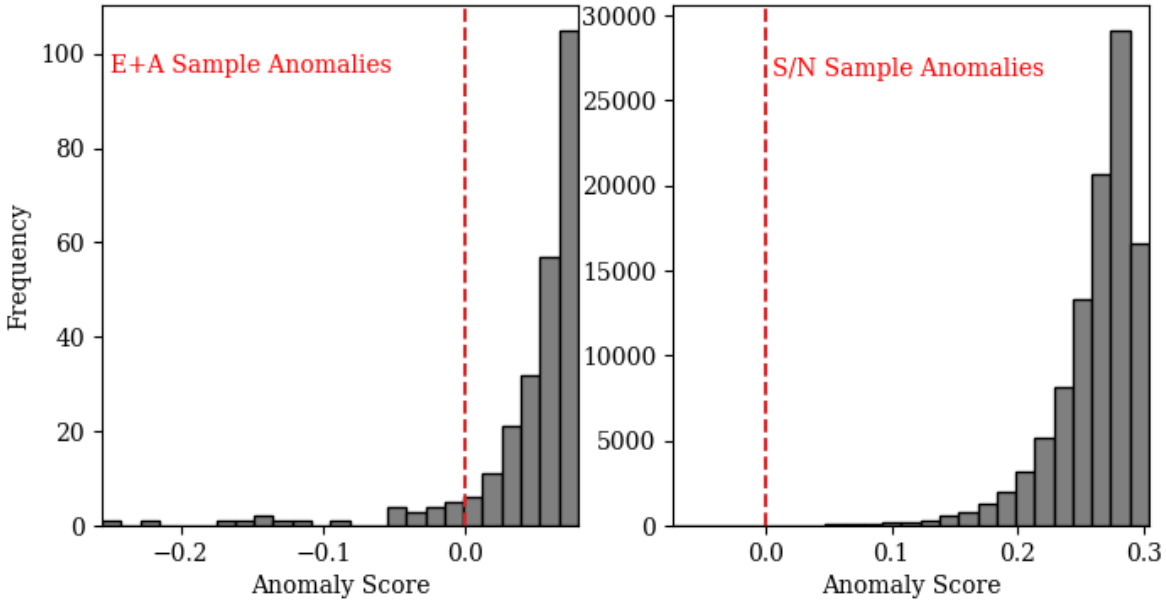


**Figure 1.** Anomaly scores of the two samples. Left: High S/N sample; consisting of $\sim 10^5$ objects, the red dashed line delimits where anomalies are, below 0 indicates anomaly, above indicates nominal. Right: E+A sample; consisting of 287 objects, with the same delimiting line as the S/N plot.

tree, the largest part of the tree, that isolates the normal instances, does not need to be generated. This allows the model to build partial models and to exploit sub-sampling (Liu et al. 2008).

Sub-sampling is a necessary step in the process of the iForest method. Large sample sizes can reduce the iTree's ability to confidently identify anomalies and thus the iForest as a whole becomes less effective. This is due to swamping and masking. Swamping is when the normal instances are too similar to the outlying instances, and masking is where the existence of too many outliers hides their own presence (Liu et al. 2008). Sub-sampling overcomes these issues by creating smaller, partial selections of the data. This leads to smaller iTrees that can better isolate data, as well as each iTree being able to specialise as each sub-sample would contain different sets of anomalies or even no anomalies. Furthermore, this allows the algorithm to be robust to overfitting and avoids the need for expensive distance calculations with larger iTrees (Ishida et al. 2021).

The iForest in this work is trained on two key samples. The samples are E+A galaxies (see section 3.2) and high S/N galaxies (see section 3.3). The data given to the iForest includes both spectroscopic data (equivalent width (EW) measures) and photometric data (colour, mag and fluxes). Three isolations are run on the galaxy selections. The first iForest is given purely the spectroscopic data, the second iForest is given the photometric data and the third is provided with both spectroscopic and photometric data. The iForest will use a sub-sample sizes of 256 instances to manage this data which is the standard value in Liu et al. (2008). Figure 1 displays the resultant anomaly scores of the spectroscopic isolations of both samples. As stated, a score below zero (the red dashed line) indicates an anomalous object and we selected to isolate 25 anomalies from each sample.

### 3.2 E+A Galaxies

E+A galaxies are a strong tool for investigating the processes of galaxy evolution. They are an important link between star-forming spirals and quiescent ellipticals (Dressler & Gunn 1982; Couch & Sharples 1987; Wilkinson et al. 2017). We discussed the many selection methods in section 1 and there are merits and pitfalls of each method. More recent work have stated the necessity of limiting H$\alpha$ (Chen et al. 2019; Greene et al. 2021). H$\alpha$ is an indicator of star formation and although this selection can remove dusty star formers, it can remove E+A galaxies that have H$\alpha$ because of AGN activity (Wilkinson et al. 2017). This AGN possibility is something we come across in our later analysis of our results (section 5.2). We also wish to further study the selection method and the homogeneity of such selection methods.

Wilkinson et al. (2017) studies three selections in depth in their work. They have an additional look at solely strong H$\delta$ Balmer absorption to further analyse the selection process. The master sample of H$\delta$ strong galaxies uses an $EW_{H\delta} > 3$Å. From this sample, a further two selections are made, both have low [OII] emission using $EW_{[OII]} > -2.5$Å called the 'E+A' sample. The third selection has a final cut of low H$\alpha$, $EW_{H\alpha} > -3$Å called the 'pure E+A' sample. Wilkinson et al. (2017) concludes that the H$\delta$ strong sample commonly have late-type morphology and are often star formers. The E+A sample have a mix of elliptical and spiral morphologies. Lastly, the pure E+As are most commonly red elliptical galaxies and have a higher fraction in denser environments compared to the H$\delta$ strong and E+A galaxies. Wilkinson et al. (2017) states that these results suggest an evolutionary sequence from blue-disc galaxies to quiescent red elliptical galaxies.

We want to study the evolution of the E+As into the quiescent early types and as such, choose to use the E+A selection (Wilkinson et al. 2017). As such we use H$\delta$ strong as $EW_{H\delta} < -3$Å, and low [OII] absorption as $EW_{[OII]} < 2.5$Å. We change the signs to fit

with the GAMA conventions for denoting emission and absorption lines. This selection keeps the $H\alpha$ strong galaxies in the sample and will allow for future analysis of AGN activity in the galaxy. AGN-driven feedback is thought to be a possible mechanism for quenching galaxies e.g. (Hopkins et al. 2008; Wilkinson et al. 2017).

We apply additional S/N cuts to the EW measures. This is to ensure that we are not simply finding noisy galaxies that might muddy the results. The S/N cut is performed as follows,

$$\mathrm{S/N} = \frac{|\mathrm{EW}|}{\mathrm{EW_{err}}} \geq 3. \tag{1}$$

Through our selection method for E+A galaxies we create a sample of 287 galaxies. This sample constitutes ∼1% of the GAMA DR4 total dataset, a typical amount when compared to the local Universe (Greene et al. 2021).

### 3.3 High S/N Galaxies

The signal to noise cut is designed to remove the noisiest data from the sample and leave only high S/N galaxies. We implement a simple S/N > 8 cut to the full 230,000 spectra in the sample, resulting in a sample of $\sim 10^5$. Similar to the E+A galaxy isolation, we perform the iForest with three inputs of data, spectroscopic, photometric and a combination of both. This sample is secondary to the E+A sample and is here as test of the iForest's ability to isolate anomalies in a larger sample.

### 3.4 Morphometric Parameters

We will be using 8 morphometric parameters. Concentration, asymmetry and smoothness ($CAS$ Conselice 2003; Bershady et al. 2000), Gini and $M_{20}$ (Abraham et al. 2003; Lotz et al. 2004) are used extensively in optical images of galaxies and are a versatile tool in quantising the shape, size and other morphological qualities of galaxies. We will also include a less tested but interesting set of morphometry parameters: multimode, intensity and deviation ($MID$ Freeman et al. 2013; Peth et al. 2015). We utilise statmorph (Rodriquez-Gomez et al. 2019), a commonly used Python tool, to compute these morphometrics using the standard definitions from the above mentioned papers.

#### 3.4.1 *Concentration-Asymmetry-Smoothness (CAS)*

CAS is a commonly used space that was developed by Bershady et al. (2000) and Conselice (2003). Concentration ($C$) measures the ratio of light within an inner aperture to the light within an outer aperture. The definition in Bershady et al. (2000) states,

$$C = 5\log_{10}\left(\frac{r_{80}}{r_{20}}\right), \tag{2}$$

where $r_{20}$ and $r_{80}$ are the radii of the circular apertures that contain 20% and 80% of the total flux respectively. We take the total flux to be the flux contained within 1.5 times the Petrosian radius ($r_p$), the standard definition from Conselice (2003). The centre for these apertures is determined via the minimisation of asymmetry which we will discuss below. High $C$ values indicate a bright central bulge region which is a main feature of spiral galaxies. The circular apertures about the centre however, make this measure inconsistent if the bright structures exist outside this central region, e.g multiple nuclei. Typical values for discs are $C > 3.5$ and for large ellipitcals, $C = 1-3$. Irregular galaxies span the entire spectra (Conselice 2003)

Asymmetry ($A$) is defined as the number computed when a galaxy is rotated 180° about its centre and then subtracted from the original galaxy image (Conselice 2003):

$$A = \sum_{ij} \frac{|I(i,j) - I_{180}(i,j)|}{|I(i,j)|} - B_{\mathrm{asymm}}, \tag{3}$$

where, $I$ is the galaxy's image and $I_{180}$ is the rotated image, $i$ and $j$ describe the pixel positions on a 2D image. $B_{\mathrm{asymm}}$ is an estimate of the contribution the sky background has on the asymmetry measure. This means $A$ can be negative if the background asymmetry is large. Due to the noise correction, it is unreliable to compute $A$ for low S/N images, this should not be an issue in our data as we have set lower limits to S/N before calculations. Like $C$, $A$ is calculated within $1.5r_p$ about the galactic centre, this centre is determined by minimising $A$. This minimisation is calculated by moving the centre of rotation about a grid at the centre of the image. A high asymmetry value indicates an asymmetric galaxy, typical of irregular and merging galaxies. Smooth, elliptical light profiles will result in a lower $A$ value due to their rotational symmetry.

Smoothness ($S$) is the final parameter developed by Conselice (2003), utilised to quantify the degree of small-scale structure. The image of the galaxy is smoothed and then subtracted from the original image:

$$S = \sum_{ij} \frac{|I(i,j) - I_{\mathrm{smooth}}(i,j)|}{|I(i,j)|} - B_{\mathrm{smooth}}. \tag{4}$$

Here, $I_{\mathrm{smooth}}$ is the smoothed galaxy image and $B_{\mathrm{smooth}}$ is an estimate of the contribution the sky background has on the smoothness measure. In this work we use a gaussian smoothing with a standard deviation of $0.25r_p$. $S$ is typically an indicator of recent star-formation as the small-scale structures are often where star-formation is occurring in late-type spirals. Smooth ellipticals that are no longer star forming will have a low $S$ value (Conselice 2003). Like $A$, $S$ is not as effective when applied to poorly resolved galaxies such as those at high redshifts.

#### 3.4.2 *Gini and $M_{20}$*

Abraham et al. (2003) and Lotz et al. (2004) introduce the $GM_{20}$ space as an alternate measure to $CAS$. Their aim was to create parameters that are free from the centre selection processes and are more susceptive to merger and interaction signatures.

The Gini coefficient ($G$) is a statistical measure based on the Lorenz curve, the rank-ordered cumulative distribution function of a population's wealth. A Gini value of 0 indicates perfect equality (all pixels have an equal fraction of the flux) and a Gini value of 1 indicates perfect inequality (a single pixel contains all the flux). This measure is repurposed as a distribution measure of the galaxy's pixel values (Abraham et al. 2003; Lotz et al. 2004). For a discrete population, $G$ can be defined as the mean of the absolute difference between all pixel values. Lotz et al. (2004) sort the pixel values into increasing order to allow for more efficient computation and use the definition,

$$G = \frac{1}{|X|n(n-1)} \sum_{i}^{n} (2i - n - 1)|X_i|, \tag{5}$$

where $X$ is the mean over all pixel flux values $X_i$, $n$ is the total number of pixels in the galaxy, $i$ is the pixel index. For the majority of local galaxies, $G$ correlates with $C$ and will increase with the ratio of light in the central component of the galaxy. Abraham et al. (2003) finds that $G$ strongly correlates with surface brightness and $C$ when studying 930 galaxies in the SDSS Early Data Release.

$C$ and $G$ deviate however because $G$ is found independently from the spatial distribution of flux. Strong $G$ values can be found when bright pixels exist outside of the central component, where as strong $C$ values typically denote only a bright central component (Conselice 2003; Lotz et al. 2004). When computing $G$, care must be given to the background sky. $G$ must be calculated from only the pixels belonging to the galaxy in order to be a true and accurate measurement of the galaxy's $G$ coefficient (Lotz et al. 2004).

The second-order moment of the brightest 20% of the flux ($M_{20}$) is another measure of flux distribution. We first define the second order moment of a pixel $i$ as in Lotz et al. (2004),

$$M_i = I_i[(x - x_c)^2 + (y - y_c)^2]. \tag{6}$$

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 is then given by:

$$M_{\rm tot} = \sum_i^n M_i. \tag{7}$$

Lotz et al. (2004) uses the relative contribution to the second order moment of the pixels that contain 20% of the total flux after sorting the list of pixels by descending intensity, giving us:

$$M_{20} = \log_{10}\left(\frac{\sum_i M_i}{M_{\rm tot}}\right) \text{ while } \sum_i I_i < 0.2 I_{\rm tot}. \tag{8}$$

Here $I_{\rm tot}$ is the total flux of the galaxy pixels. $M_{20}$ is sensitive to bright regions in the outer regions of discs and a larger value for $M_{20}$ indicates star-forming outer regions or strongly interacting discs. Furthermore, $M_{20}$ is more sensitive to multiple nuclei than $C$ as it does not use apertures in its implementation and its centre point is a free parameter.

### 3.4.3 *Multimode-Intensity-Deiviation (MID)*

The MID statistics (Freeman et al. 2013; Peth et al. 2015) were first introduced as an alternative to the $CAS$ and $GM_{20}$ parameters. Their aim is to be more sensitive to recent mergers than the prior morphometrics and to overcome the difficulties in identifying post-merger morphologies. They also aimed to overcome the degradation of $CAS$ and $GM_{20}$ that comes with increasing redshift. These statistics have not been tested as extensively as the $CAS$ and $GM_{20}$ measures, especially using hydronamical simulations (see discussion in Rodriquez-Gomez et al. 2019; Holwerda et al. 2025).

The multimode $M$ statistic measures the ratio between the areas of the two most 'prominent' clumps within a galaxy. This has the implicit assumption that a well resolved galaxy will have at least two well-defined clumps (Holwerda et al. 2025). The bright regions are found using a threshold method where $q_l$ represents the normalized flux value and $l$ percent of pixel fluxes are less than $q_l$. This results in a binary image $g_{i,j}$ where 1 represents fluxes larger than $q_l$ and 0 represents fluxes less than $q_l$:

$$g_{i,j} = \begin{cases} 1 & f_{i,j} \geq q_l \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

The number of pixels in contiguous groups of pixels with value 1 are then sorted in descending order by area. The two largest groups $A_{l,(2)}$ and $A_{l,(1)}$ define an area ratio $R_l$ (Peth et al. 2015):

$$R_l = \frac{A_{l,(2)}}{A_{l,(1)}}. \tag{10}$$

The $M$ statistic is then the maximum value of $R_l$. Values of $M$ that approach 1 represent multiple nuclei, while values near 0 are single nuclei systems (Peth et al. 2015). The original work of Freeman et al. (2013) applies an additional factor of $A_{l,(2)}$ to limit the affects of hot pixels, but STATMORPH utilises the above method from Peth et al. (2015) so the measure is size independent.

The intensity statistic ($I$) measures the ratio between the two brightest subregions of the galaxy. To calculate $I$, the galaxy image must first be smoothed slightly using a Gaussian kernal with $\sigma = 1$ pixel. The image is then partioned into pixel groups according to the watershed algorithm. This algorithm states that each distinct subregion consists of all the pixels such that their maximum gradient paths lead to the same local maximum. The surrounding eight pixels of every pixel are inspected and the path of maximal intensity increase is followed until a local maximum is reached. These maximums are the local brightness maximums of the pixels flux (Rodriquez-Gomez et al. 2019). Once the pixel groups are defined, their summed intensities are sorted into descending order: $I_1, I_2, etc$. Intensity is then defined as (Freeman et al. 2013):

$$I = \frac{I_2}{I_1}. \tag{11}$$

Lastly, Deviation ($D$) measures the distance between the intensity centroid $(x_{\rm c}, y_{\rm c})$, calculated for the pixels identified by the $MID$ segmentation map, and the brightest peak found during computation of the $I$ statistic $(x_{I_1}, y_{I_1}$ Freeman et al. 2013). The intensity centroid is defined as,

$$(x_{\rm c}, y_{\rm c}) = \left(\frac{1}{n_{\rm seg}}\sum_i\sum_j i f_{i,j}, \frac{1}{n_{\rm seg}}\sum_i\sum_j j f_{i,j}\right). \tag{12}$$

with the summation being overall $n_{\rm seg}$ pixels within the segmentation map. The distance between the two centroids will be affected by the absolute size of the galaxy and as such a normalisation is applied using $\sqrt{n_{\rm seg}/\pi}$. The $D$ statistic is the defined as:

$$D = \sqrt{\frac{\pi}{n_{\rm seg}}}\sqrt{(x_{\rm c} - x_{I_1})^2 + (y_{\rm c} - y_{I_1})^2}. \tag{13}$$

### 3.4.4 *Automated Morphology Measuring*

We supply STATMORPH with cutout images extracted from SDSS using ASTROQUERY (Ginsburg et al. 2019). The initial cutout size of each image is a 50px by 50px square, which provides 20 by 20 arcsecond image. We then apply a threshold mask to find the objects in the cutout and reduce the cutout size of the galaxy to $1.5r_p$ of the galaxies flux. By supplying these images to STATMORPH, it can extract the full morphometric space and give us our 8 parameters and more (see in Rodriquez-Gomez et al. 2019). Figure 2 shows a corner plot of the most commonly used morphometric parameters, C,A,S,G & $M_{20}$.

## 4 RESULTS

We calibrate the iForest to find the top 25 most anomalous galaxies from each sample and dataset. We chose 25 because it have a feasible sample size to analyse, while still maintaining a statistically useful sample count. These isolations results in a total of 150 anomalous objects identified from the GAMA DR4 data. In total, we find that there are 103 unique objects due to duplicate anomalies identified across the spectroscopic, photometric and combined isolations. We

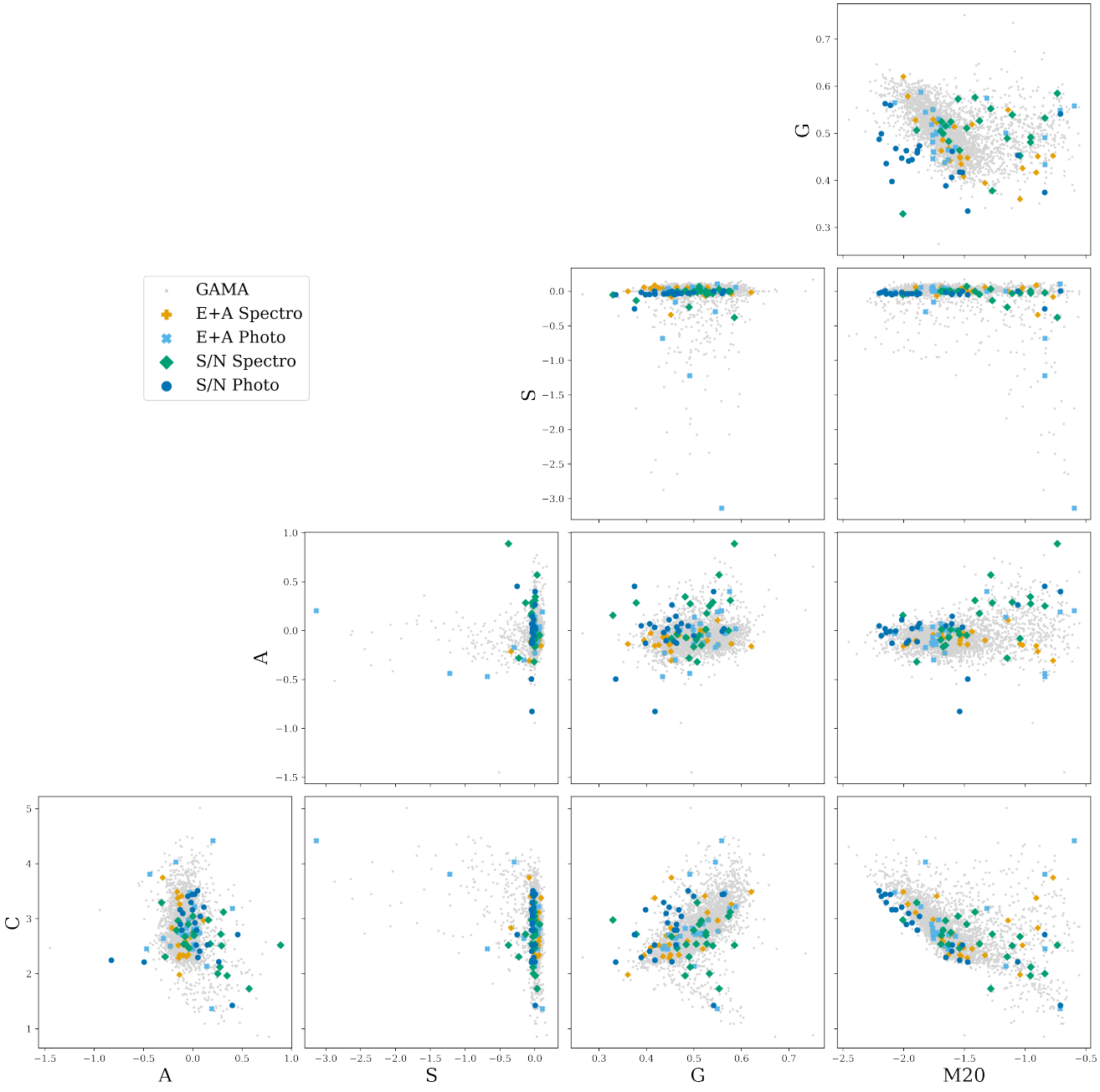


**Figure 2.** A corner plot of the most common 5 morphometric measures, concentration, asymmetry, smoothness, Gini & $M_{20}$. Orange pluses are E+A spectroscopic anomalies, light blue crosses are the E+A photometric anomalies, green diamonds are S/N spectroscopic anomalies and navy blue circles are S/N photometric anomalies. We can see from this corner plot that a significant portion of the anomalous objects, have 'normal' morphologies, in that, they lie within the bulk of the other objects in GAMA. There are some extreme outliers morphologically such as in asymmetry and smoothness.

find that two of those 103 objects are falsely identified as galaxies by the GAMA pipeline and are in fact shredded larger galaxies, this can be seen in figure 3. The remaining 101 galaxies are confidently identified as galaxies and are further inspected. Of the 101, we find that 13 have 'bad spectra', consisting of pipeline issues, fringing errors, data reduction errors and false emission lines. This shows the iForest's ability to robustly extract data errors in a given sample without training. We display examples of two types of bad spectra in figure 4, namely the fringing error and the false emission lines.

We find 54 unique objects from the S/N spectroscopic, photometric and combined samples, 2 of which are the falsely identified shredded galaxies, resulting in 52 confirmed galaxies. 20 objects are explicitly isolated from the spectroscopic data, 22 are isolated from the photometric sample and/or the combined sample and 10 are found exclusively in the combined spectro/photo sample (see in table A2).

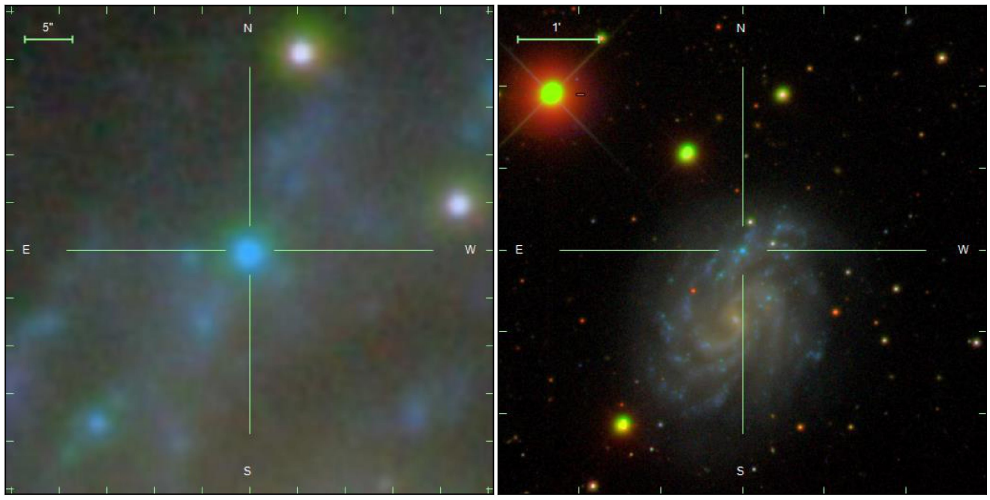


**Figure 3.** Example shredded galaxy. Left: The falsely identified galaxy that is actually a bright sub-structure in a larger galaxy, right.

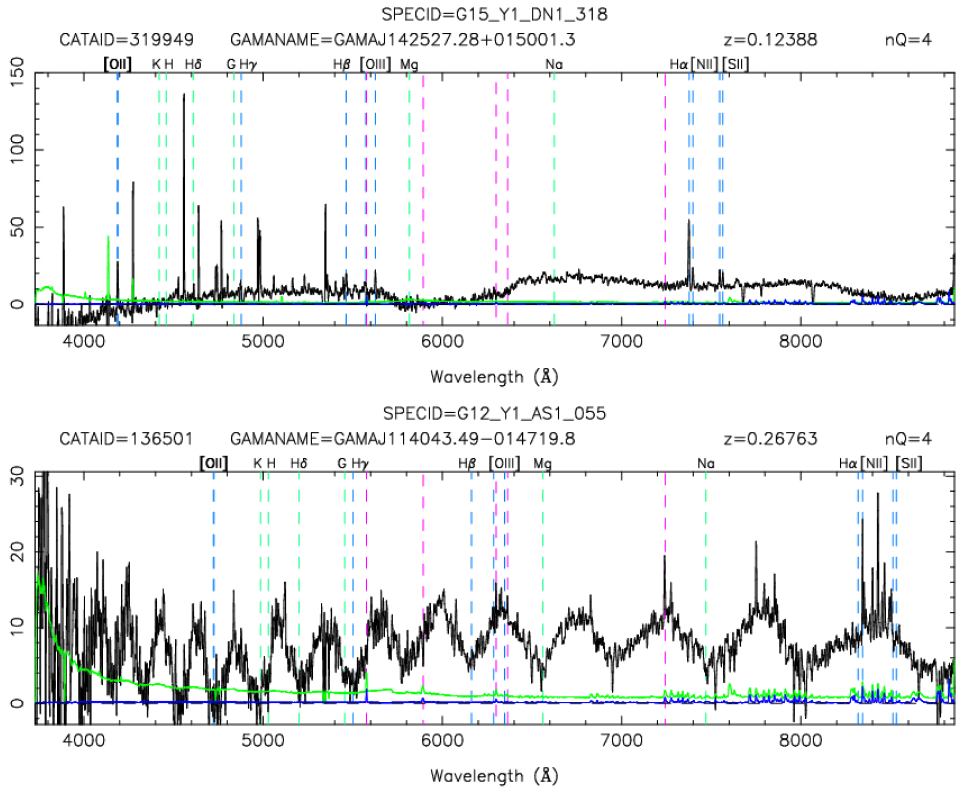


**Figure 4.** Two types of 'bad spectra' found by the iForest. Top: false emission line spectra; displaying peaks in areas that are not valid emission lines. This is caused due to poor flat fielding that was an issue with early AAOmega that has since been fixed (Croom, priv. comm.). Bottom: data reduction error; this presents as a sinusoidal pattern in the spectra line.

These galaxies encompass a wide range of astronomical phenomena such as extreme emission line galaxies (EELG), a subset of EELGs known as 'Green Beans', AGN hosts and red spiral galaxies. We find that 13 galaxies are EELGs, selected as having either $EW_{[OIII]}$ or $EW_{H\alpha}$ > 300Å (Lumbreras-Calle et al. 2022) and an example spectra of one of these EELGs is displayed in figure 5. 9 of the identified EELGs are not cited by any works targeting EELGs. One of the EELGs identified is the aforementioned 'Green Bean' galaxy (see in Prescott & Sanderson 2019). We use BPT and WHAN diagrams seen in figure 6 and identify 4 possible AGN hosts, one of which being the Green Bean galaxy. The largest sample of anomalous galaxies are the red spiral galaxies, consisting of 18 out of the 52 anomalies found from the S/N samples. This value equates to approximately 35% of the anomalies in this sample. This leaves 18 galaxies that

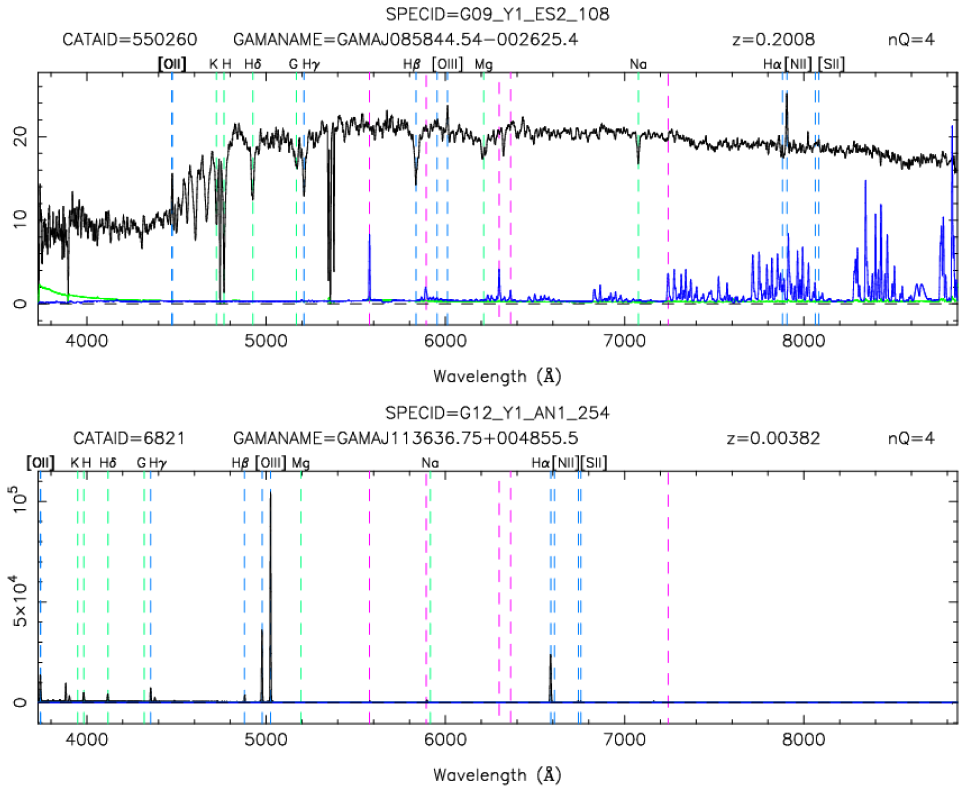


**Figure 5.** Two types of anomalous spectra found by the iForest. Top: strong H$\alpha$ E+A galaxy extracted from GAMA DR4. This galaxy has a $EW_{H\alpha} \approx 21$Å, $EW_{[OII]} \approx 2$Åand $EW_{H\delta} \approx -3$Å. Bottom: EELG extracted from GAMA DR4. This galaxy has a $EW_{[OIII]} \approx 546$Å.

are not immediately considered scientifically important galaxies and from study of their spectra we see that many have spectral lines that lie below the continuum.

We choose to isolate ∼10% of the E+A sample, resulting in 25 objects from each sample, 75 objects total. We find that 25 anomalous objects are isolated from the spectroscopic data and 24 objects are isolated from the photometric data (see in table A1). We note that no unique galaxies are found exclusively in the combined isolation. From the 49 objects isolated from the E+A sample, we find that there are 29 objects with abnormally large H$\alpha$ emission lines ($EW_{H\alpha} > 10$Å). 20 of these galaxies come from the spectroscopic isolation, and 9 come from the photometric isolation. We expected that the anomalous E+A galaxies isolated from the spectroscopic data would have strong H$\alpha$ but we also detect strong emission in some of the E+A galaxies found from the photometric isolation. The remaining population of objects isolated do not appear to be scientifically interesting, giving credence to the notion that just because objects are statistically rare, does not mean they are automatically interesting. Of the 16 remaining photometric anomalies, they all appear visually typical for E+A galaxies, being red in colour and elliptical in shape.

## 5 DISCUSSION

Using the 8 morphometric parameters discussed in section 3.4, we can quantise the morphologies of our anomalies. Figure 2 displays a corner plot of the parameters, concentration, asymmetry, smoothness, gini and M20. The plot indicates that the anomalous galaxies are typically normal in morphology. There are however some extremes, particularly within the smoothness and M20 regimes. We find the most anomalous galaxies by morphology, to do this we take the C-A plot and calculate the population density of the points near by, shown in Fig. 7. From the population densities we show the 16 galaxies that have the lowest population density, and therefore are the most morphologically anomalous, in figure A1.

The quantised parameters can be utilised to check if E+A galaxies follow the typical S0/elliptical shape. Using a Gini-$M_{20}$ plot and region lines discussed in Lotz et al. (2008) we roughly group our isolated objects into three regions, merger candidates, Sb/Sc/Irr candidates and E/S0/Sa candidates. From Fig. 8 we can see that the approximately 75% of our isolated galaxies lie within the Sb/Sc/Irr region, this agrees with visual inspection as many of our isolated galaxies have irregular morphology. We can see that only 3 isolated E+A galaxies lie within the E/S0/Sa region. Since E/S0/Sa is the typical morphology of E+A galaxies (Dressler & Gunn 1983), this again supports that the iForest algorithm has worked to find the weirdest E+A galaxies. We also find around about 20% of the E+A galaxies are merger candidates, which could be the cause of their recent starbursts.

We also utilise a Python tool SHAP to analyse the impact value of each feature in the datasets. SHapley Additive exPlanations (SHAP) is a game theoretic approach to explain the output of any machine learning model (Lundberg et al. 2020). It connects optimal credit allocation with local explanations using the classic Shapley values from game theory to quantise the 'impact' each feature has on the model. From figure 9 we can see that TiO2 and TiO1 has a strong impact on defining our anomalies, this line of TiO1/TiO2 suggests

large populations of low mass stars. We can also see the forbidden line of [OIII] having a strong impact in the S/N sample. We are surprised that H$\alpha$ is not in the top 10 most impactful features when studying the E+A spectroscopic sample, but, due the amount of features included in the isolation, each individual feature has a rather small impact compared to the overall dataset. Similarly, when considering the impact of the photometric features in figure 10, each individual feature has a small impact due to the volume of data. We do note however that the u band photometry plays the largest role when designating anomalies for both the high S/N sample and the E+A sample.

### 5.1 High S/N Anomalies

13 EELGs are isolated from the S/N sample, with either $EW_{[OIII]}$ or $EW_{H\alpha} > 300$Å. EELGs are a rare population of galaxies that have been studied extensively over the past decade (Maseda et al. 2014; Smit et al. 2015; Lumbreras-Calle et al. 2022; Llerna et al. 2024). These galaxies are especially rare in the local Universe, but their number density increases with increasing redshift (Smit et al. 2015). Typically, local EELGs have low stellar masses, $M_* < 10^9 M_\odot$. The EELGs we have found have stellar masses ranging from $10^8 \rightarrow 10^{11} M_\odot$ and have redshift values $0.01 < z < 0.3$. The most common lines that we find EELGs in are [OIII]$\lambda$4959Å, $\lambda$5007Å (here after as [OIII]) and H$\alpha$ in the rest-frame optical (Iglesias-Páramo, J. et al. 2022; Llerna et al. 2024). We note that all 13 of the EELGs come from the spectroscopic isolation and find that 9 of these EELGs have not been identified by other works. One EELG that has been identified in other works is the 'Green Bean' galaxy (Prescott & Sanderson 2019). Green Beans are a similar galaxy to 'Green Pea' galaxies, first noted in (Cardamone et al. 2009) and observed by volunteers in the Galaxy Zoo project. Their appearance is like that of its namesake, small elliptical and green in colour. This green colour is due to the strong optical emission lines of [OIII] and they display some of the highest specific star formation rates (sSFR) in the local Universe (Cardamone et al. 2009; Amorín et al. 2015). The Green Beans are similar to Green Peas, being also sites of extremely high sSFR but they are larger and less compact than the Green Peas. Green Beans also have morphology that resembles that of Type 2 AGN, further supported by spectroscopic data indicating AGN like emission ratios resembling Type 2 AGN.

We isolate 18 red spirals, most commonly isolated from the photometric isolation. These galaxies were initially observed by Masters et al. (2010) and they are though to make up ~6% of all late-type spirals. Further study showed that red spirals are more common with stellar masses above $10^{10}$M$\odot$ and become less common with decreasing stellar mass (Maseda et al. 2014). By evaluating the stellar masses of the red spirals we have isolated, we find that 16 have stellar masses $> 10^{10}$M$\odot$. This population of red spirals makes up approximately 35% of isolated high S/N galaxies which is higher than we expected. Using the Python library SHAP, we can visualise which features have the most impact when designating anomalies. From the photometric analysis shown in figure 10, we can see that the u band colours and magnitudes are the most impactful features for designating anomalous objects. This tells us that the u band photometry is the most unique feature for this anomalous high S/N galaxies from the photometric isolation.

### 5.2 Anomalous E+A Galaxies

The most interesting outcome of the E+A isolations is the strong H$\alpha$ galaxies. We find that 29 of the E+A anomalies have have strong H$\alpha$ emission ($EW_{H\alpha} > 5$Å), 20 from the spectroscopic isolation and 9 from the photometric isolation. This strong H$\alpha$ is typically suggest of current star-formation (Kennicutt 1998), it could also be attributed to AGN activity (Wilkinson et al. 2017; Pawlik et al. 2018; Greene et al. 2021) or could be a sign of dust obscuration (Goto et al. 2003; Pawlik et al. 2018).

AGNs can be identified by broad emission lines, most commonly H$\alpha$ as well as NII (Baldwin et al. 1981; Kewley et al. 2006). By studying the spectra of these star-forming E+A galaxies, we initially find that 9 of the anomalies display $EW_{[OIII]} > 10$Å and as such, we broadly flag these galaxies as possible AGN hosts. Furthermore, we examine BPT and WHAN plots, shown in figure 6, and we can see that there are only 2 possible AGN candidates. Checking the CATAIDs of these 2 objects, we see they are also in the same sample of strong [OIII]. Looking again at the BPT-WHAN plot, we note that ≈ 85% of the anomalous E+A galaxies lie within the star-forming delimiting region. Typically we would expect E+A galaxies to lie within the passive region (Greene et al. 2021). Greene et al. (2021) argues that the E+A galaxies found within this region are still undergoing inside out quenching and as such, they are uncertain as to whether or not they can be classed as E+A galaxies. Pawlik et al. (2018) however, argues that these strong H$\alpha$ galaxies are a part of the evolutionary process of star-bursting galaxy to quenched elliptical and designates them as ePSBs. After analysing the BPT-WHAN plots and the spectra of our anomalous E+A galaxies, we conclude that 20 galaxies have strong H$\alpha$ and do not exhibit possibilities of AGN hosts.

Goto et al. (2003) studies the presence of strong H$\alpha$ in E+A galaxies, compiling a sample of approximately 3000 E+A galaxies. These E+A galaxies have strong Balmer absorption and show little to no [OII]. Of the 3000 galaxies, Goto et al. (2003) find the 52% have non-insignificant H$\alpha$ emission and they suggest that these are dusty star-formers polluting the data set. Pawlik et al. (2018) also delimits the previous ePSBs into a sub group of dPSBs, or dusty star-formers. These dPSBs are not, in-fact, post starburst galaxies but have been labelled such for convenience and are found by analysing the ratio between H$\alpha$ and H$\beta$. If $H\alpha/H\beta > 2.87$ then Pawlik et al. (2018) states they are dust obscured. We check our 20 E+A for dust obscuration and note that 7 of the objects are dust obscured according to the $H\alpha/H\beta$ ratio, resulting in 13 E+A galaxies that are star-forming but not dust obscured. We want to further check for dust obscuration and analyse the colour-magnitude diagrams of the $g$ and $r$ bands. Using the definition by Shearman & Pimbblet (2014) note than none of the galaxies are anomalously red. Further study of the optical $g, r$ and infra-red $i$ bands agrees that there are no more dusty star-formers.

The spectra of these 13 star-forming E+A galaxies (SFE+A) suggests that they are not forming high-mass stars ($M \geq 8M_\odot$) like O and B class, but are still efficiently forming lower mass stars ($M \leq 8M_\odot$, i.e. A,F,G,K & M class stars). The lack of [OII] and [OIII] forbidden lines but the presence of strong H$\alpha$ supports the notion that star-formation is happening, but massive, hot stars are not being produced.

### 5.3 Small-scale Interactions

The formation of low-mass stars is a well understood and observed process (Larson 1969; Shu et al. 1987; Swift & Welch 2008) and they are believed to accrete much of their mass before nuclear burning can begin (Shu et al. 1987). High-mass stars however, have Kelvin-Helmholtz timescales that are less that their dynamical time, meaning the process of nuclear burning begins before all of their mass has been accreted (McNally 1964; Bodenheimer & Sweigart 1968; Larson 1969). This nuclear burning would cause an immense radia-

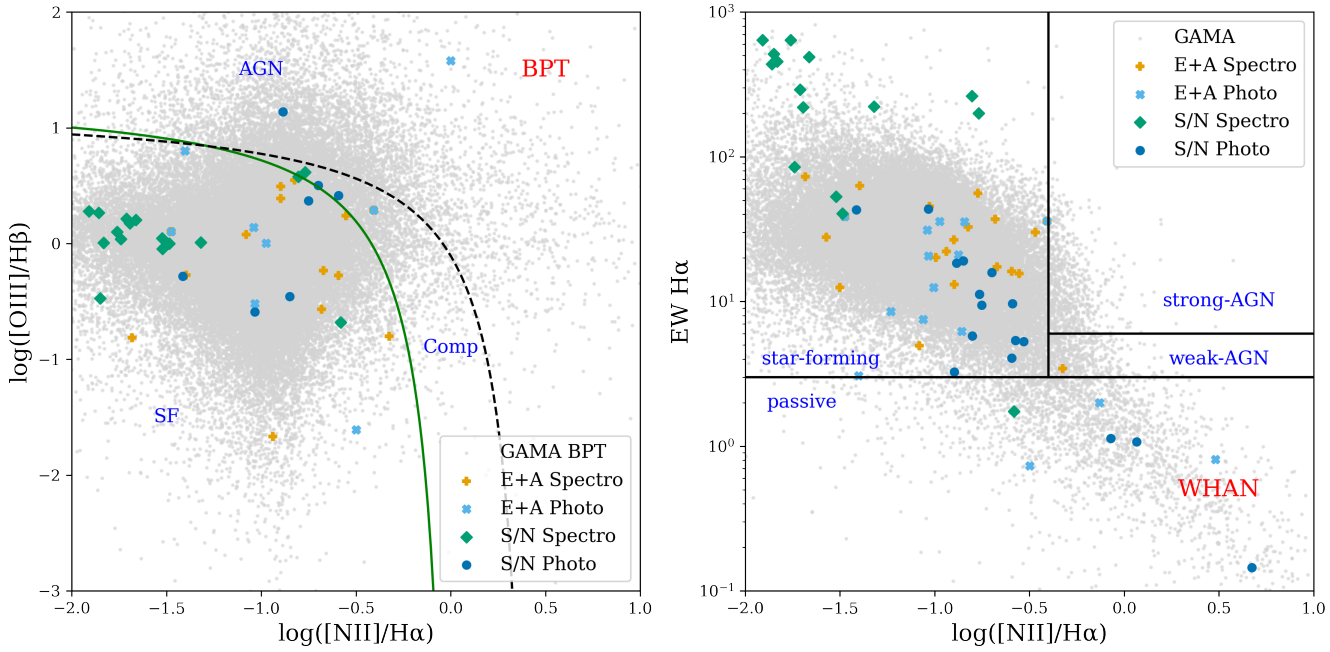


**Figure 6.** Left: BPT diagram; The black dashed line is the extreme starburst classifier line and the green solid line is the pure star formation line from Kewley et al. (2006). Right: WHAN diagram; the lines delimit the spectral classes defined in Cid Fernandes et al. (2010). The key is the same as in figure 2. We see 3 objects are possible AGN hosts from both BPT and WHAN diagrams. We also see that ∼85% of the E+A galaxies lie within the star-forming region in the WHAN plot when they should lie in the passive region.

tive pressure that repels the still accreting mass, suggesting a much faster accretion rate is needed for higher mass stars than lower mass stars (Yorke & Sonnhalter 2002; Keto & Wood 2006; Peters et al. 2010; Kuiper et al. 2011). Observations of high-mass objects (Keto & Wood 2006), as well as simulated scenarios (Haemmerle et al. 2015), shows that large accretion disks and rapid accretion rates can form high-mass stars in dense environments. We note that all of the identified SF E+A objects are low-mass and diffuse objects and suggest that recent small-scale interactions has caused a burst in star-formation that has not formed O or B class stars. We propose that this is due to the diffuse, gas poor galaxies not being able to accrete the necessary material for high mass stars before the radiative pressure expels the accreting material.

We analyse the morphology of our anomalous E+A galaxies using the parameters discussed in section 3.4. In particular we use $G$, $M_{20}$ and $A$ to find merger signatures. Initial visual inspection showed that that all of the anomalous E+A galaxies are small, faint objects, with 80% being in cluster environments. We use the $G-M_{20}$ plot shown in figure 8 to quantise each galaxy morphology. Only 8 anomalous E+A galaxies lie within the merger region of the plot, with a further 3 lying in the E/S0/Sa region and the remainder lying in the Sb/Sc/Irr region. We note that typical E+A galaxies should fall in the E/S0/Sa region, having elliptical morphologies. The 3 spectroscopically isolated E+A that lie in the merger region are all part of the SFE+A population, but remaining 10 SFE+As all fall in the Sb/Sc/Irr region. This suggests that perhaps they are not merging, but it does support that they are undergoing minor interactions causing the irregular morphology.

### 5.4 Low Jeans Limits

Steinhardt (2024) proposed the idea that galaxies can still be star-forming, even though they appear by colour to be quiescent. This theory arose due to two issues in local large galaxies, namely; i) most observed massive blue star-forming galaxies that are presumed to be nearing the end of their stellar mass growth are ∼1dex less massive than local massive quiescent galaxies, and ii) the quiescent galaxies can not have gained this stellar mass from interactions or mergers since their mass ratios between stellar mass and black hole mass ($M_*/M_{BH}$) are also ∼1dex less massive than expected (Steinhardt 2024). M87 is taken as an example of a local massive quiescent galaxy. The stellar mass of M87 is $M_* = 10^{12.3} M_\odot$ and the black hole mass is $M_{BH} = 10^{9.8} M_\odot$. This stellar mass is ∼1dex larger than the largest known star-forming galaxies, suggesting that its stellar mass must have grown after its supposed turnoff. Furthermore, the $M_*/M_{BH}$ ratio is much higher than theoretical work predicts (Kormendy & Ho 2013). Steinhardt (2024) proposes then, that much of the final stellar mass is formed not during its blue phase but instead during its red phase, coining these galaxies, red star forming galaxies (RSFGs). Our proposed SFE+A galaxies could be a part of this population of ongoing star-forming galaxies as Steinhardt (2024)'s proposed RSFGS would appear, spectroscopically, very similar to E+A galaxies, with strong H$\delta$, low [OII] and non-negligible H$\alpha$.

Steinhardt (2024) provides a theory as to why these galaxies are not forming large mass stars, stating that the Jeans mass is so low that it is not possible for more massive stars to form. For a typical galactic molecular cloud, the speed of sound is ∼0.2 km/s (Krumholz & McKee 2008), and has a density of ∼100 gcm$^{-3}$, giving a Jeans mass of $10 M_\odot^2$. As the temperature of the cloud drops and metallicity increases, the Jeans mass will also decrease. At a certain temperature and metallicity, the Jeans mass will be so low as to be unable to form O and B class stars. This would not completely prevent star formation as theoretically, AFGKM class stars could still efficiently form.

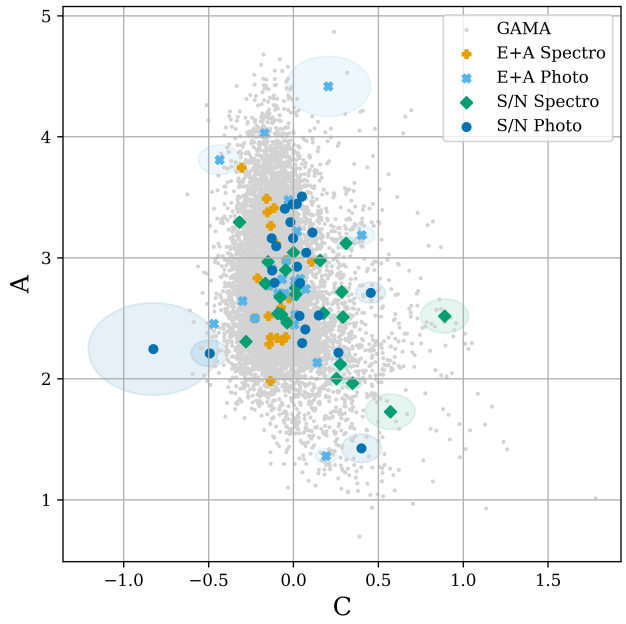


**Figure 7.** Concentration-Asymmetry plot used to find the population density of each object in the $C - A$ space. We find the distance to the nearest 10th neighbour to identify the anomalous objects with the most extreme morphology. The 9 galaxies with the lowest densities (most different) are displayed in figure A1.

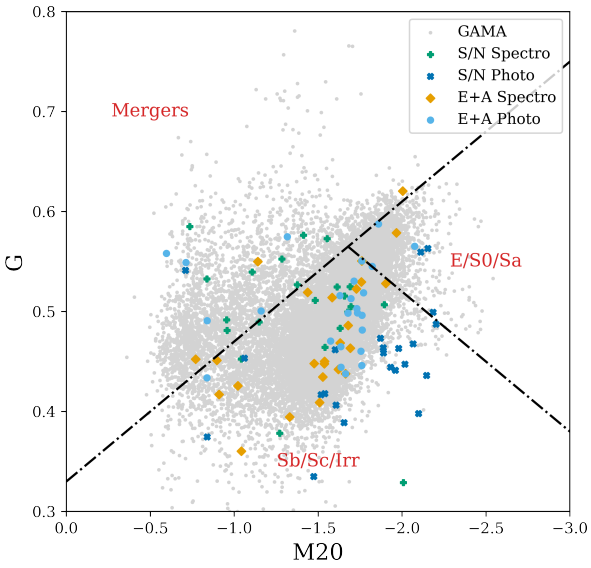


**Figure 8.** $G$-$M_{20}$ plot with delimiting lines from (Lotz et al. 2008). The key for the points are the same as in figure 2. We can see that the many of the E+A galaxies lie within the Sb/Sc/Irr region, where they should lie in the E/S0/Sa region. This suggests that something has disturbed the morphology of the E+A galaxies and could hint at a reason for the E+A galaxy's starburst.

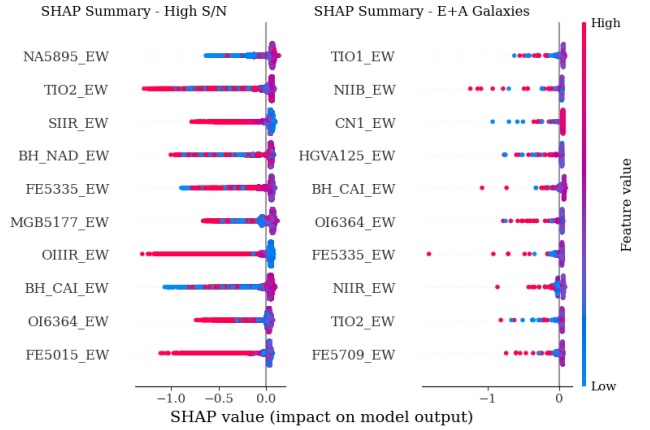


**Figure 9.** A summary plot created using the SHAP Python tool. The plot shows the 10 most impactful features from the spectroscopic data set in descending order. Left: High S/N sample, we can see that some of the most impactful features are sodium and titanium, typical of small class stars like K and M. We also see [OIII] we is typical of extreme star-formers or AGN hosts. Right: E+A sample, we can again see titanium and iron, we also see nitrogen, hinting at possible AGN hosts.

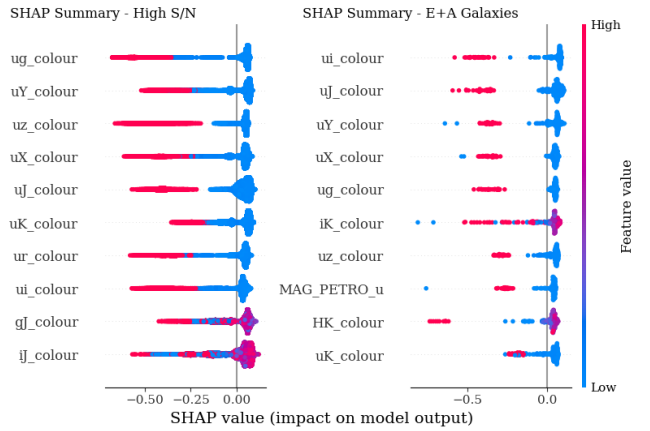


**Figure 10.** A summary plot created using the SHAP Python tool. The plot shows the 10 most impactful features from the photometric data set in descending order. Left: High S/N sample. Right: E+A sample. We see in both samples that the u band colours have the greatest impact on the anomaly scores.

### 5.5 Data Reduction Errors

We also discuss the errors that the iForest algorithm has uncovered. All of these objects have been isolated from the spectroscopic datasets and we will discuss the reasons behind this shortly. We have identified 2 objects that are shredded galaxies, where a larger galaxy is falsely segmented into smaller galaxies. An example of the shredded galaxy is displayed in figure 3 and it was identified by the iForest as it is an extremely bright star-forming sub-structure in a larger galaxy. The iForest found 13 'bad' spectra in the dataset, including 9 data reduction errors (example shown in figure 4) and 4 false emission line spectra (examples shown in figure 4). The iForest detects the sinusodial spectra due to errors in the EW measures which are flagged as unusual only in the spectroscopic data. The false emission lines

do not effect the EW measures directly but they are instead detected as the far left side of the spectra dips below the continuum, causing errors in some of the EW measures. We study these false emission lines and find that they do not fall on any specific spectral lines. We analyse them further and find that all of these false emission line were observed on the same date (5th March 2008) but originate in different fibres, ruling out fibre issues. We plot the coordinates of each object and find no obvious correlation, making us rule out a sky phenomena such as a comet or other object. Though communications with S. Croom, we conclude that the errors are due to poor flat fielding and sky subtraction (Croom, priv. comm.). At the blue end of the spectra, flat fielding leads to zero divisions when normalising and could cause the line. Towards the red end of the spectra, flat fielding could be caused by fibre fringing which was an issue in early AAOmega that has since been fixed (Sharp et al. 2006).

## 6 CONCLUSIONS

We have utilised an iForest algorithm to isolate anomalous galaxies in two core samples, E+A galaxies, selected in a traditional manner, and high S/N galaxies. We supplied the unsupervised model with spectroscopic and photometric data for $\approx 10^5$ objects and we find several rare phenomena and a number of odd morphologies.

As highlighted in BP16 and their outlier detection model, the unsupervised iForest in the work also allows us to automatically identify outliers from the data, without having to choose explicit cuts. The only free parameter in the model is the outlier fraction, all other model parameters are learned automatically through the data and feature names.

From an input of 287 E+A galaxies, we isolate 49 unique anomalous galaxies, we chose to extract 10% of the sample per dataset to produce a statistically important sample size. We find 25 E+A anomalies from the spectroscopic data and 24 anomalies from the photometric data. From an input of $\approx 10^5$ S/N> 8 galaxies, we isolate 52 unique anomalous galaxies, with 20 anomalies isolated from the spectroscopic data, 22 from photometric data and 10 from the combined spectroscopic/photometric data. Studying the 101 galaxies we find unusual emission lines, data reduction errors, EELGS, red spirals and possible star-forming E+A galaxies.

We chose to select our E+A galaxies in the same manner as (Dressler & Gunn 1983) – with strong Balmer absorption and low [OII]. Through the spectroscopic isolation, we find that a significant portion of the E+A galaxies display extreme H$\alpha$ emission. We initially theorised that this H$\alpha$ emission could be caused by AGN activity. Visual inspection and a study of the [OIII] EW highlighted that of the 29 strong H$\alpha$ E+As, 9 have strong [OIII] lines or show signs of an AGN. This result indicated that it was unlikely to be AGN activity leading to the strong H$\alpha$. For further confirmation, we used BPT and WHAN (Baldwin et al. 1981; Cid Fernandes et al. 2010) plots and find that there are very few AGN hosting E+A galaxies. We propose then, that these E+A galaxies are still star-forming, and this star-formation is what produces the H$\alpha$ lines.

Goto et al. (2003) suggests that these galaxies are dusty star-formers and the [OII] and [OIII] forbidden lines are obscured by dust. Using the definition by Shearman & Pimbblet (2014) we study the colour-mag ratios of the $g$ and $r$ bands and note than none of the galaxies are anomalously red. Study of the optical $g, r$ and infra-red $i$ bands suggests that these are not dusty star-formers so we use the method from Pawlik et al. (2018). Pawlik et al. (2018) uses the ratio of H$\alpha$/H$\beta > 2.87$ to confirm dust obscuration. We find that 7 of the E+A galaxies are dust obscured and are a separate population to the strong [OIII] population. This results in a total of 13 SFE+A galaxies that are not dusty, nor are they possible AGN hosts.

Strong H$\alpha$ but no [OII] suggests that although star-formation is occurring, there are no high mass stars forming that can ionise the surrounding gas. If no high mass stars are forming, we provide two solutions: i) Small-scale interactions have perturbed the gas poor galaxies, resulting in some star-formation, but the radiative pressure of accreting massive stars prevents them from being formed, or ii) The E+A galaxies the iForest has identified are still star-forming efficiently but are unable to form O and B class stars due to higher metallicity and low temperatures.

This work shows that to select 'pure' E+A galaxies, it is a necessity to select on H$\alpha$ as well as H$\delta$ and [OII]. We can see, from the anomalies within the E+A selection method we chose, that the sample is polluted with dusty star-formers, possible AGN hosts and RSFGs. We also show the robustness of the iForest algorithm in detecting anomalies with no prior learning. We find over 50% of the anomalies have not be identified in other works but we note that not all of the objects are scientifically interesting just because they are flagged as anomalies. This work could be expanded upon by utilising Active Anomaly Detection (AAD) to utilise expert input to tailor the models output. We can also test the hypotheses by studying the extended structure of the E+A galaxies to analyse how the galaxies might be quenching and interaction. We will probe this extended structure via image stacking and finding the averaged surface brightness profiles and the Sérsic profile that best fits it.

## ACKNOWLEDGEMENTS

The anonymous reviewer is thanked for providing valuable feedback and revisions on the manuscript.

We thank C. Conselice for providing useful insights in to the analysis of our data.

We would also like to thank S. Croom for their assistance and correspondence that answered some of our unknowns.

KB would also wish to thank H. Green for assisting in classifying the visual morphology and for proof readings.

We would like to thank J. Liske who created all of the spectra shown in the paper figures, see in (Liske et al. 2015) for more detail. All of the PNG images were downloaded from https://www.gama-survey.org/.

This research made extensive use of IPYTHON (Pérez & Granger 2007) and the following PYTHON packages: astropy[2], a community-developed core Python package and an ecosystem of tools and resources for astronomy (Astropy Collaboration et al. 2013, 2018, 2022); astroquery[3], a coordinated package for astropy (Ginsburg et al. 2019); photutils[4], a Python library that provides commonly-used tools and key functionality for detecting and performing photometry of astronomical sources (Bradley et al. 2024); scikit-learn[5] (Pedregosa et al. 2012); shap[6], a game theoretic approach to explain the output of any machine learning model (Lundberg et al. 2020); and statmorph[7], an affiliated packed of astropy for calculating non-parametric morphological diagnostics of galaxy images (e.g., $G$-$M_{20}$

[2] www.astropy.org/
[3] astroquery.readthedocs.io/en/latest/
[4] https://photutils.readthedocs.io/en/stable/
[5] www.scikit-learn.org/
[6] https://shap.readthedocs.io/en/latest/
[7] statmorph.readthedocs.io/en/latest/overview.html

and *CAS* statistics), as well as fitting 2D Sérsic profiles (Rodriquez-Gomez et al. 2019).

## DATA AVAILABILITY

This work made extensive use of the GAMA DR4 data. GAMA is a joint European-Australasian project based around a spectroscopic campaign using the Anglo-Australian Telescope. The GAMA input catalogue is based on data taken from the Sloan Digital Sky Survey and the UKIRT Infrared Deep Sky Survey. Complementary imaging of the GAMA regions is being obtained by a number of independent survey programmes including GALEX MIS, VST KiDS, VISTA VIKING, WISE, Herschel-ATLAS, GMRT and ASKAP providing UV to radio coverage. GAMA is funded by the STFC (UK), the ARC (Australia), the AAO, and the participating institutions. The GAMA website is https://www.gama-survey.org/. We make specific use of SPECLINESSFR (Gordon et al. 2017), which compiles a significant amount of spectroscopic data and EW measures. We further use GKVSCIENCECAT (Bellstedt et al. 2020), which compiles the main survey selection including redshifts and photometry data. Lastly, we utilise SPECCAT and APMATCHEDPHOTOM (Liske et al. 2015) for extracting the spectra plots and for further photometry measures. We made extensive use of th GAMA Single Object Viewer[8].

## REFERENCES


Abraham R. G., van der Bergh S., Nair P., 2003, ApJ, 588, 218
Adame A. G., et al., 2024, A&A, 168, 58
Amorín R., et al., 2015, A&A, 578, A105
Astropy Collaboration et al., 2013, A&A, 558, A33
Astropy Collaboration et al., 2018, AJ, 156, 123
Astropy Collaboration et al., 2022, ApJ, 935, 167
Baldwin J. A., Phillips M. M., Terlevich R., 1981, PASP, 93, 5
Baron D., Poznanski D., 2016, MNRAS, 465, 4530
Bellstedt S., et al., 2020, MNRAS, 496, 3235
Bershady M. A., Jangren A., Conselice C. J., 2000, AJ, 119, 2645
Bodenheimer P., Sweigart A., 1968, ApJ, 152, 515
Bradley L., et al., 2024, Zenodo
Cardamone C., et al., 2009, MNRAS, 399, 1191
Chang Y.-Y., Hsieh B., Wang W.-H., Lin Y.-T., Lim C.-F., Toba Y., Zhong Y., Chang S.-Y., 2021, ApJ, 920, 68
Chen Y.-M., et al., 2019, MNRAS, 489, 5709
Cid Fernandes r., Stasinska G., Schlickmann M. S., Mateus A., Vale Asari N., Schoenell W., Sodre L., 2010, MNRAS, 403, 1036
Clarke A. O., Scaife A. M. M., Greenhalgh R., Griguta V., 2020, A&A, 639, A84
Conselice C., 2003, ApJS, 147, 1
Couch W. J., Sharples R. M., 1987, MNRAS, 229, 423
Dressler A., Gunn J. E., 1982, ApJ, 263, 533
Dressler A., Gunn J. E., 1983, ApJ, 270, 7
Driver S., et al., 2022, MNRAS, 513, 439
Euclid Collaboration et al., 2024, arXiv e-prints, p. arXiv:2405.13491
Freeman P. E., Izbicki R., Lee A. B., Newman J. A., Conselice C. J., Koekemoer A. M., Lotz J. M., Mozena M., 2013, MNRAS, 434, 282
Ginsburg A., et al., 2019, AJ, 157, 98
Goel A., Montgomery M., 2015, in IAU General Assembly. p. 2255500
Gordon Y., et al., 2017, MNRAS, 465, 2671
Goto T., 2007, MNRAS, 377, 1222
Goto T., et al., 2003, PASJ, 55, 771
Greene O. A., Anderson M. R., Marinelli M., Holley-Bockelmann K., Campbel L. E. P., Liu C. T., 2021, ApJ, 910, 162
Haemmerle L., Eggenberger P., Meynet G., Maeder A., Charbonnel C., 2015, A&A, 585, A65
Hogg D. W., Masjedi M., Berlind A. A., R B. M., Quintero A. D., J B., 2006, ApJ, 650, 763
Holwerda B. W., et al., 2025, PASA, 42, e028
Hopkins P. F., Hernquist L., Cox T. J., Kereš D., 2008, ApJS, 175, 356
Iglesias-Páramo, J. et al., 2022, A&A, 665, A95
Ishida E. E. O., et al., 2021, A&A, 650, A195
Ivezić Ž., et al., 2019, ApJ, 873, 111
Kamalov F., Leung H. H., 2020, JIKM, 19, 2040013
Kennicutt Jr. R. C., 1998, ARA&A, 36, 189
Keto E., Wood K., 2006, ApJ, 637, 850
Kewley L. J., Groves B., Kauffmann G., Heckman T., 2006, MNRAS, 372, 961
Kormendy J., Ho L. C., 2013, ARA&A, 51, 511
Krumholz M. R., McKee C. F., 2008, Nature, 451, 1082
Kuiper R., Klahr H., Beuther H., Henning T., 2011, ApJ, 732, 11
Larson R. B., 1969, MNRAS, 145, 271
Liske J., et al., 2015, MNRAS, 452, 2087
Liu F. T., Ting K. M., Zhou Z.-H., 2008, in 2008 Eighth IEE International Conference on Data Mining. pp 413–422, doi:10.1109/ICDM.2008.17
Liu H., Li X., Li J., Zhang S., 2018, IEEE Trans. Syst. Man. Cybern., 48, 2451
Llerna M., et al., 2024, A&A, 691, A59
Lochner M., McEwen J. D., Peiris H. V., Lahav O., Winter M. K., 2016, ApJS, 225, 31
Lotz J., Primack J., Madau P., 2004, AJ, 128, 163
Lotz J. M., et al., 2008, ApJ, 672, 177
Lumbreras-Calle A., et al., 2022, A&A, 668, A60
Lundberg S. M., et al., 2020, Nature Machine Intelligence, 2, 2522
Margalef-Bentabol B., Huertas-Compant M., Charnock T., Margalef-Bentabol C., Bernardi M., Dubois Y., Storey-Fisher K., Zanisi L., 2020, MNRAS, 496, 2346
Maseda M. C., et al., 2014, ApJ, 791, 17
Masters K. L., et al., 2010, MNRAS, 405, 783
McNally D., 1964, ApJ, 140, 1088
Pawlik M. M., et al., 2018, MNRAS, 477, 1708
Pedregosa F., et al., 2012, J. Mach. Learn. Res., 12, 2825
Pérez F., Granger B. E., 2007, Computing in Science and Engineering, 9, 21
Peters T., Banerjee R., Klessen R. S., Mac Low M. M., Galvan-Madrid R nd Keto E. R., 2010, ApJ, 711, 1017
Peth M. A., et al., 2015, MNRAS, 458, 963
Poggianti B. M., et al., 2009, ApJ, 693, 112
Prescott M. K. M., Sanderson K. N., 2019, ApJ, 885, 40
Reza M., 2021, Astron. Comput., 37, 100492
Rodriquez-Gomez V., et al., 2019, MNRAS, 483, 4140
Sharp R., et al., 2006, SPIE, 6269E, 14
Shearman O., Pimbblet K. A., 2014, PASA, 31, 38
Shu F. H., Adams F. C., Lizano S., 1987, ARA&A, 25, 23
Smit R nd Bouwens R. J., et al., 2015, ApJ, 801, 122
Steinhardt C. L., 2024, arXiv e-prints, p. arXiv:2402.03423
Swift J. J., Welch W. J., 2008, ApJS, 174, 202
Tahi A., Hadi A. S., 2019, ACM Comput. Surv., 52, 1
Vergani D., et al., 2010, A&A, 509, A42
Wilkinson C., Pimbblet K., Stott J., 2017, MNRAS, 472, 1447
York D. G., et al., 2000, AJ, 120, 1579
Yorke H. W., Sonnhalter C., 2002, ApJ, 569, 846
Zabludoff A. I., Zaritsky D., Lin H., Tucker D., Hasimoto Y., Shectman S. A., Oemler A., Kirshner R. P., 1996, ApJ, 466, 104


[8] https://www.gama-survey.org/dr4/tools/sov.php

## APPENDIX A: SUMMARY

In this appendix we summarise all of the outlying galaxies in tables displaying their coordinates, classification and the sample they have been isolated from and the data used in said isolation. The galaxy classification has been allocated through visual inspection by KB,

while also taking into account the designation in the SDSS navigator, and a secondary inspection by H. Green was performed to confirm the classification.

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

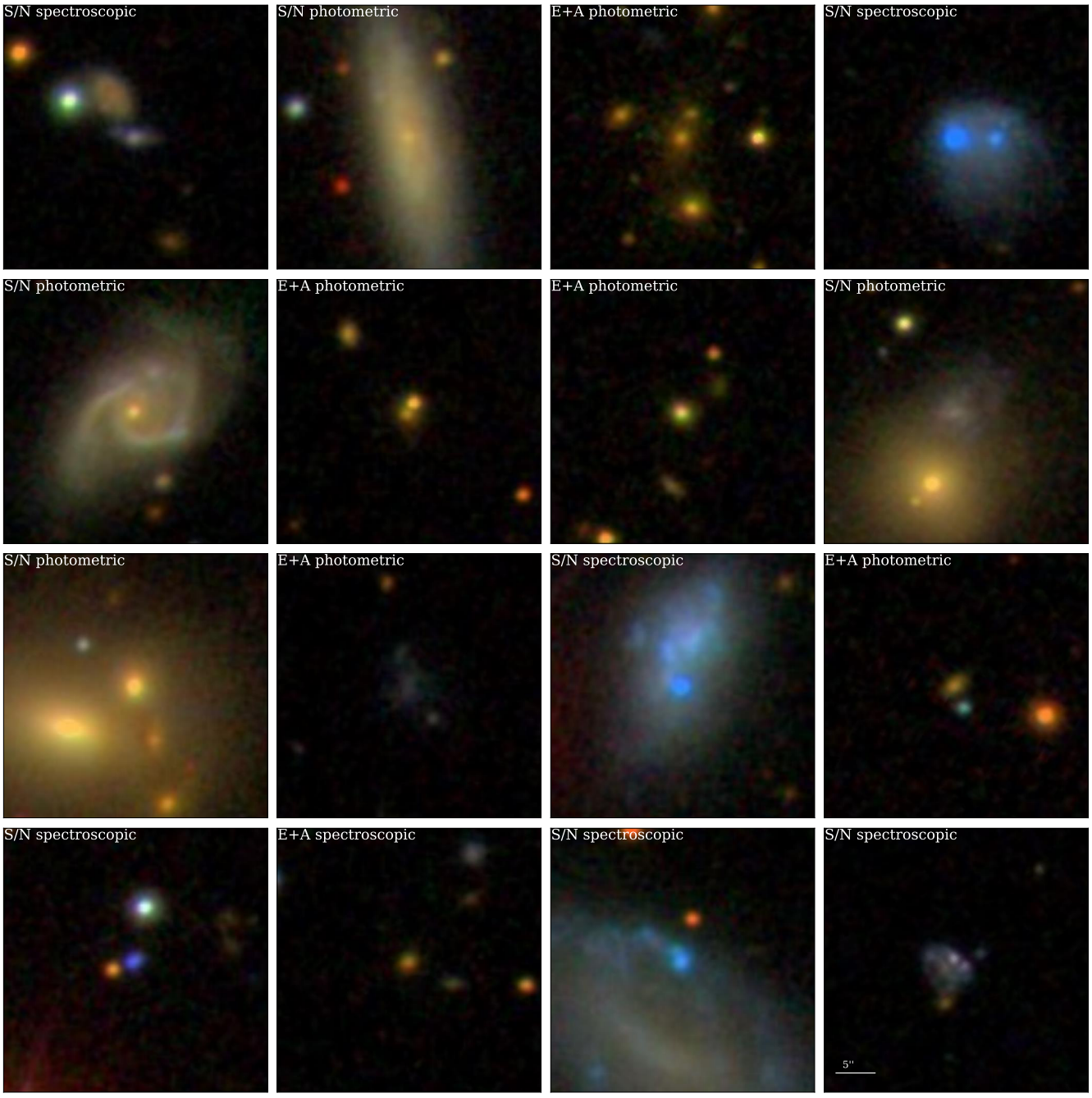


**Figure A1.** Shown above are the most anomalous 16 galaxies found from the CAS selection. Anomalies in this case are defined by plotting the CAS parameters and then finding the point density of the nearest 10 neighbours of the iForest galaxies. Shown in the top corner of each image is the sample the galaxy has been selected from. E+A and S/N shows the data selecting process and spectroscopic, photometric or combination shows the data provided to the iForest.

**Table A1.** The anomalous objects found from the E+A isolations are highlighted in the table. We display the object name from GAMA where possible and where not possible an alternate name. We further display the RA and DEC coordinates in deg. The class column indicates the type of object, S = Spiral, S0, I = Irregular, M = Multiple Nuclei, E = Elliptical. The isolation data column indicates from which isolation the object has been selected from, S = Spectroscopic, P = Photometric and C = Combination. Where duplicate objects are found, we indicate all samples the object is isolated from.

| Galaxy Name | RA (deg) | DEC (deg) | Class | Isolation Data |
|---|---|---|---|---|
| GAMA 623866 | 140.63771 | 0.81636 | S0 | S + P + C |
| DESI J216.5852+01.6198 | 216.58525 | 1.61986 | S | S + C |
| GAMA 250176 | 214.18704 | 1.97408 | I | S + C |
| GAMA 319942 | 216.30633 | 1.84812 | S0 | S + C |
| GAMA 599349 | 131.04029 | 0.22654 | S0 | S + C |
| GAMA 098625 | 179.03608 | 0.9419 | S0 | S + C |
| SDSS J142046.10+004641.8 | 215.19208 | 0.77829 | M | S + C |
| SDSS J120756.28-020301.3 | 181.98454 | -2.05038 | I | S |
| SDSS J120825.77-020134.1 | 182.10737 | -2.02615 | S0 | S |
| GAMA 136501 | 175.18121 | -1.78885 | I | S |
| GAMA 185670 | 181.19825 | -1.5217 | I | S |
| GAMA 272241 | 178.43688 | 1.43316 | S0 | S |
| GAMA 298897 | 222.12246 | 1.19695 | E | S |
| GAMA 319949 | 216.36367 | 1.83371 | M | S |
| GAMA 512767 | 219.90375 | -1.06983 | S0 | S |
| GAMA 535853 | 179.51933 | -0.85615 | I | S |
| GAMA 536399 | 182.154007 | -0.9574 | M | S |
| GAMA 560613 | 180.7675 | -0.59587 | S0 | S |
| GAMA 084615 | 178.84121 | 0.47832 | I | S |
| GAMA 092370 | 216.0315 | 0.53364 | S | S |
| GAMA 093959 | 223.16904 | 0.56275 | I | S |
| SDSS J090230.75+003137.8 | 135.62817 | 0.5272 | M | S |
| SDSS J090248.47-001215.5 | 135.70196 | -0.20432 | S | S |
| SDSS J141533.72-005913.5 | 213.89054 | -0.9871 | E | S |
| SDSS J141546.02-010922.1 | 213.94179 | -1.15615 | S0 | S |
| GAMA 130823 | 178.82567 | -2.19445 | S0 | P + C |
| GAMA 203840 | 135.44063 | -0.29678 | M | P + C |
| GAMA 215981 | 136.03192 | 0.52556 | E | P + C |
| GAMA 302687 | 138.82412 | 1.43235 | M | P + C |
| GAMA 302729 | 139.07225 | 1.47696 | E | P + C |
| GAMA 324714 | 136.92746 | 1.72548 | S | P + C |
| GAMA 382668 | 137.91567 | 1.94651 | S0 | P + C |
| GAMA 387344 | 135.94775 | 2.33613 | E | P + C |
| GAMA 049759 | 222.42671 | -0.70777 | S0 | P + C |
| GAMA 056197 | 184.84033 | -0.25884 | E | P + C |
| GAMA 007568 | 178.24217 | 0.77447 | E | P + C |
| GAMA 007758 | 179.35896 | 0.66938 | E | P + C |
| SDSS J091504.74+024105.1 | 138.64029 | 1.73369 | M | P + C |
| SDSS J092233.04+004858.8 | 133.21933 | -1.61439 | E | P + C |
| SDSS J115518.16-021140.0 | 177.07012 | 0.64766 | E | P + C |
| SDSS J141124.12+024258.4 | 212.8505 | 2.71623 | I | P + C |
| WHL J085252.6-013652 | 139.83333 | -1.93921 | S0 | P + C |
| WISEA J091919.99-015621.2 | 138.76975 | 2.68476 | E | P + C |
| GAMA 208796 | 130.18838 | 0.03409 | S | P |
| GAMA 278848 | 133.949997 | 0.81399 | E | P |
| GAMA 302685 | 138.82104 | 1.37455 | M | P |
| GAMA 550260 | 134.68558 | -0.44041 | E | P |
| GAMA 601598 | 140.64658 | 0.36883 | S | P |
| GAMA 694909 | 184.97717 | 0.58512 | S0 | P |

**Table A2.** The anomalous objects found from the S/N isolations are highlighted in the table. We display the object name from GAMA where possible and where not possible an alternate name. We further display the RA and DEC coordinates in deg. The class column indicates the type of object, S = Spiral, S0, I = Irregular, M = Multiple Nuclei, E = Elliptical and F = False flagged objects (not galaxies). The isolation data column indicates from which isolation the object has been selected from, S = Spectroscopic, P = Photometric and C = Combination. Where duplicate objects are found, we indicate all samples the object is isolated from.

| Galaxy Name | RA (deg) | DEC (deg) | Class | Isolation Data |
|---|---|---|---|---|
| DESI J212.9710-01.9480 | 212.97275 | -0.05066 | S | C |
| GAMA 345754 | 131.11858 | 2.06396 | I | C |
| GAMA 521949 | 131.96658 | 2.92294 | M | C |
| SDSS J114838.04-014557.9 | 177.15942 | -0.23458 | F | C |
| SDSS J141504.11+021706.9 | 213.76712 | 2.28525 | S | C |
| SDSS J144612.77+021834.7 | 221.55358 | 2.31047 | S | C |
| WISEA J084735.34-010236.2 | 131.90192 | -0.95406 | S | C |
| WISEA J085902.01-021322.3 | 134.75604 | -1.77504 | S | C |
| WISEA J091248.93+024451.4 | 138.20425 | 1.97904 | S | C |
| WISEA J092026.68-025859.9 | 140.11092 | -1.01865 | S0 | C |
| SDSS J142439.90+024120.4 | 216.16625 | 2.68903 | S | P |
| GAMA 345447 | 129.48546 | 2.02111 | S | P |
| GAMA 365248 | 129.02217 | 1.89409 | S | P |
| GAMA 375560 | 129.87421 | 1.23344 | E | P |
| GAMA 380578 | 129.24721 | 1.79204 | S0 | P |
| PGC1 0051957 GROUP | 218.11875 | 0.29401 | S | P |
| SDSS J114006.78-010944.4 | 175.02667 | -0.83774 | E | P |
| SDSS J140929.04-020544.0 | 212.37017 | -1.9064 | S | P |
| WISEA J113902.88-014736.5 | 174.76383 | -0.206 | S | P |
| GALEXASC J090337.86-025224.8 | 135.90387 | -1.13016 | I | P + C |
| GAMA 085416 | 182.36933 | 0.53327 | S | P + C |
| GAMA 214184 | 129.00642 | 0.43808 | S | P + C |
| GAMA 278390 | 130.94883 | 0.83844 | S | P + C |
| GAMA 372571 | 136.46367 | 1.13359 | S0 | P + C |
| GAMA 517279 | 131.68037 | 2.53922 | M | P + C |
| GAMA J141735.55+020311.9 | 214.39804 | 2.05322 | M | P + C |
| SDSS-C4 1366 | 223.39608 | 0.01039 | S | P + C |
| WISEA J090121.80-020859.2 | 135.34375 | -1.85325 | S | P + C |
| WISEA J090330.87-025354.8 | 135.87854 | -1.10327 | S | P + C |
| WISEA J090400.77-015455.3 | 136.00417 | -0.08876 | S | P + C |
| WISEA J120843.40-012715.9 | 182.18421 | -0.54084 | E | P + C |
| WISEA J143104.87-010449.7 | 217.76737 | -0.9191 | S | P + C |
| SDSS J115133.34-022221.9 | 177.88896 | -2.37276 | M | S |
| DESI J217.4525-01.1694 | 217.45254 | -1.16935 | M | S |
| GAMA 006821 | 174.15312 | 0.81543 | M | S |
| GAMA 136473 | 175.11621 | -1.63778 | I | S |
| GAMA 137625 | 178.93579 | -1.79443 | E | S |
| GAMA 210224 | 137.08458 | 0.19551 | I | S |
| GAMA 376183 | 132.31346 | 1.48846 | S0 | S |
| GAMA 418979 | 138.35871 | 2.74863 | I | S |
| GAMA 691332 | 184.44154 | -0.78311 | E | S |
| SDSS J084152.37+025336.2 | 130.46825 | 2.8934 | M | S |
| SDSS J085218.91-010458.8 | 133.07883 | -1.08304 | S0 | S |
| SDSS J115036.14-003410.2 | 177.65058 | -0.56951 | M | S |
| SDSS J121713.09+013747.4 | 184.30454 | 1.62984 | S0 | S |
| SDSS J141440.82-003826.5 | 213.67008 | -0.64071 | S0 | S |
| SDSS J142223.42-002225.3 | 215.59762 | -0.37369 | F | S |
| SDSS J142435.47-014638.2 | 216.14779 | -1.7773 | E | S |
| SDSS J144148.48+004128.1 | 220.45208 | 0.69113 | F | S |
| SDSS J145107.11+023125.0 | 222.77963 | 2.52362 | S0 | S |
| WISEA J085226.85-010241.9 | 133.110992 | -1.04479 | M | S |
| WISEA J085229.19-011710.0 | 133.12171 | -1.28604 | E | S |
| WISEA J090937.59-010912.2 | 137.40617 | -1.15371 | M | S |
| WISEA J091004.95-012910.3 | 137.52046 | -1.4863 | S0 | S |