Wide Area VISTA Extra-galactic Survey (WAVES): Unsupervised star-galaxy separation on the WAVES-Wide photometric input catalogue using UMAP and HDBSCAN

Todd L. Cook,[1⋆] Behnood Bandi,[1] Sam Philipsborn,[1] Jon Loveday,[1] Sabine Bellstedt,[2] Simon P. Driver,[2] Aaron S.G. Robotham,[2] Maciej Bilicki,[3] Gursharanjit Kaur,[3] Elmo Tempel,[4,5] Ivan Baldry, [6] Daniel Gruen, [7,8] Marcella Longhetti,[9] Angela Iovino,[9] Benne W. Holwerda, [10] and Ricardo Demarco[11]

[1] *Astronomy Centre, University of Sussex, Falmer, Brighton BN1 9QH, UK*
[2] *International Centre for Radio Astronomy Research (ICRAR), M468, University of Western Australia, Crawley, WA 6009, Australia*
[3] *Center for Theoretical Physics, Polish Academy of Sciences, al. Lotników 32/46, 02-668 Warsaw, Poland*
[4] *Tartu Observatory, University of Tartu, Observatooriumi 1, 61602 Tõravere, Estonia*
[5] *Estonian Academy of Sciences, Kohtu 6, 10130 Tallinn, Estonia*
[6] *Astrophysics Research Institute, Liverpool John Moores University, IC2, Liverpool Science Park, 146 Brownlow Hill, Liverpool, L3 5RF, UK*
[7] *University Observatory, Faculty of Physics, Ludwig-Maximilians-Universität München, Scheinerstr. 1, 81679 Munich, Germany*
[8] *Excellence Cluster ORIGINS, Boltzmannstr. 2, 85748 Garching, Germany*
[9] *INAF - Osservatorio Astronomico di Brera, via Brera 28, 20121 Milano - Italy*
[10] *Department of Physics and Astronomy, University of Louisville, Natural Science Building 102, Louisville, KY 40292, USA*
[11] *Institute of Astrophysics, Facultad de Ciencias Exactas, Universidad Andrés Bello, Sede Concepción, Talcahuano, Chile*

Background

- The classification of astronomical objects through their imaging is a key tool for astronomy.

- Spectroscopic surveys such as SDSS1(Sloan Digital Sky Survey), GAMA2(Galaxy And Mass Assembly) and the upcoming surveys using the 4MOST3 (4-metre Multi-Object Spectroscopic Telescope) instrument all require an input catalogue of selected targets.

- These targets are generated through the analysis of prior imaging, with the star-galaxy classification of the targets being a crucial step.

Background

- Modern-day star-galaxy separation techniques:

  Color-based, morphological-based and machine learning methods

- The DEVILS survey utilises NIR colours and surface brightness to filter stars from their target catalogue.
- The GAMA input catalogue utilises SDSS imaging, and classifies sources into stars versus galaxies using a combination of profile fitting and color separation.

Background

- **Machine learning on star-galaxy separation**
- Supervised machine learning: neural networks, random forest, support vector machines etc.).
  - Disadvantage: supervised machine learning requires a plethora of prior training data which are representative of the test data.
- Unsupervised machine learning： without the use of any training data or prior distributions,
  - Advantage: making use of all available features and finding sophisticated correlations in high dimensional space, without being potentially biased by any unrepresentative training data.

WAVES-Wide Survey Overview

- **Objective**: A spectroscopic survey focusing on the local Universe using the **4MOST instrument**.
- Observes ~14.8 million sources within a magnitude limit of ***Z*** **< 21.2**.
- Covers a total area of ~1,170 square degrees:

  North Region: Equatorial plane, spanning 157.25° –225.0° in Right Ascension.

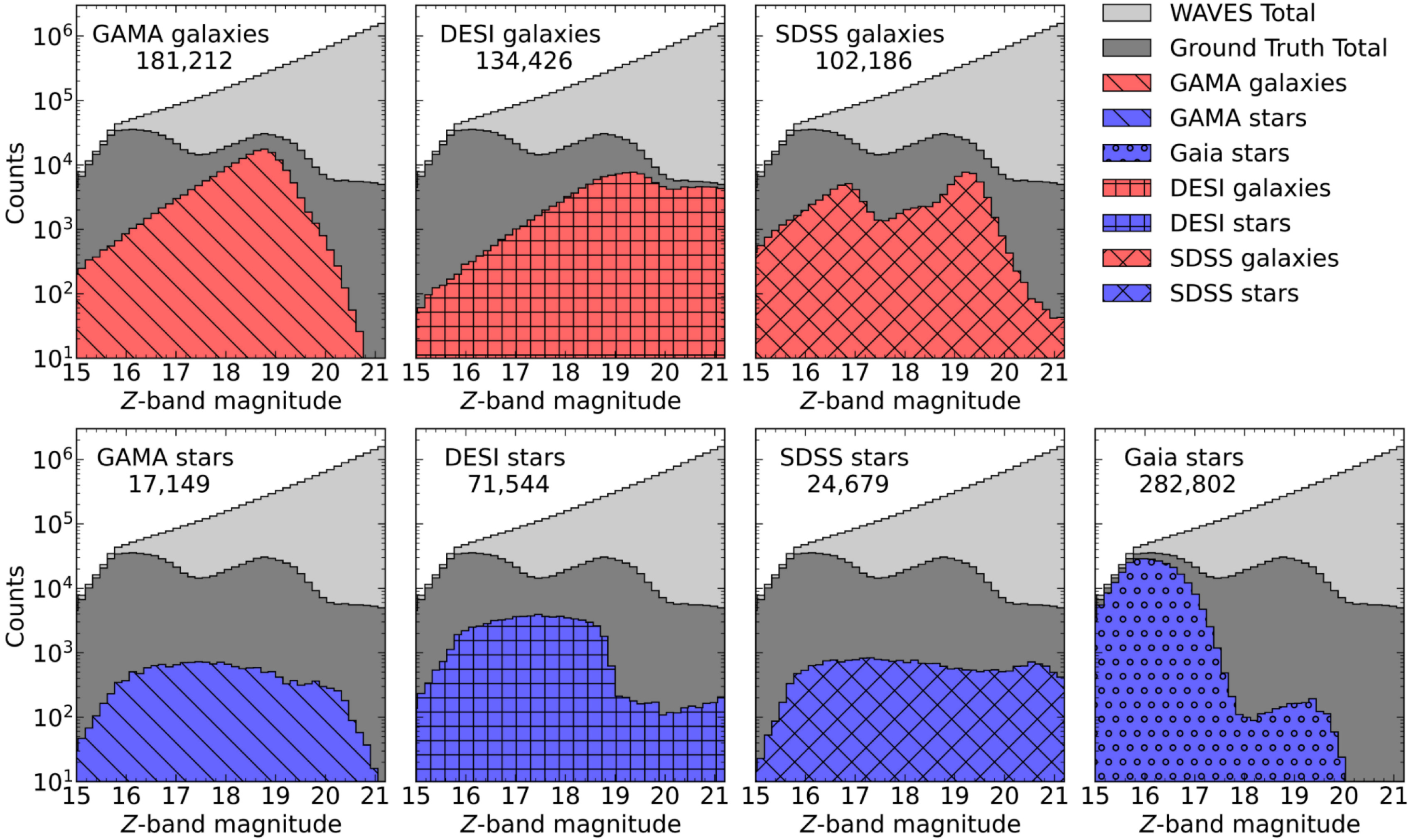
  South Region: Declination of -30° , spanning -30° –52.5° in Right Ascension.
- **Spectroscopic Coverage**:
  - 1.75 million low-resolution fibre hours.
  - Resolving power: R = 4000 - 7000 , wavelength range: 370 - 950 **nm**.
- **Photometric Redshift**: $z < 0.2$ for Wide fields and $z < 0.8$ for Deep fields.

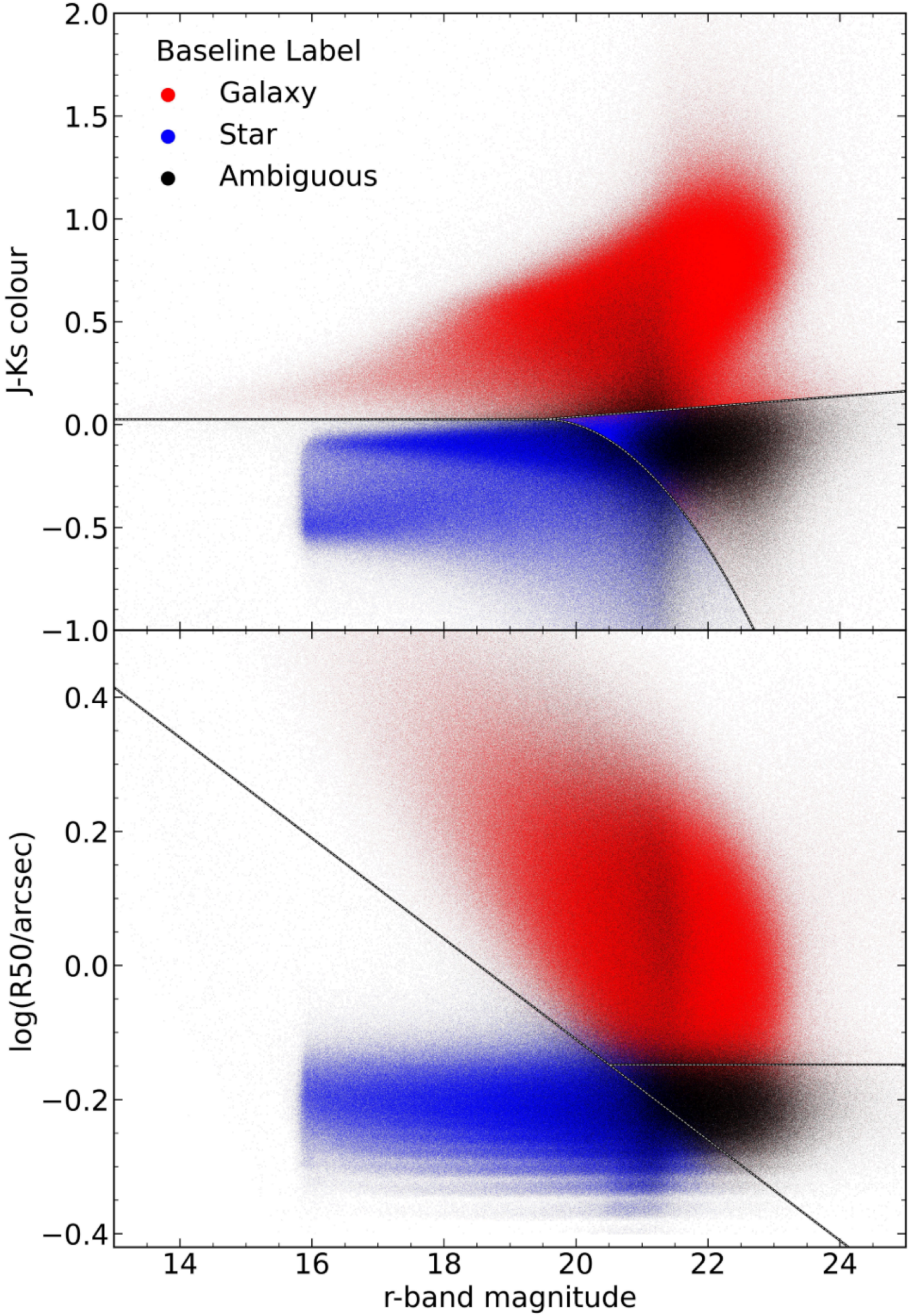
DATA

Input Catalogue Construction

- Data Sources:
  - VST KiDS: deep optical photometry in u, g, r, i bands.
  - VISTA VIKING: Near-infrared photometry in Z, Y, J, H, K_s bands.
    - The Planck E(B − V) extinction map is applied to the sources
- Source detection and characterization are carried out by the ProFound package
- The Planck E(B − V) extinction map is applied to the sources (Planck Collaboration et al. 2013), correcting their flux for Galactic dust absorption.
- Stars brighter than a Gaia G-band magnitude of 16.0 are removed, and all sources within a radius of $10^{1.6-0.15G}$ arcminutes of these bright stars are masked out because their flux can affect the estimate of other sources' fluxes in the photometry
- Finally, 14,802,032 sources within the $Z < 21.2$ magnitude limit that need to be classified.

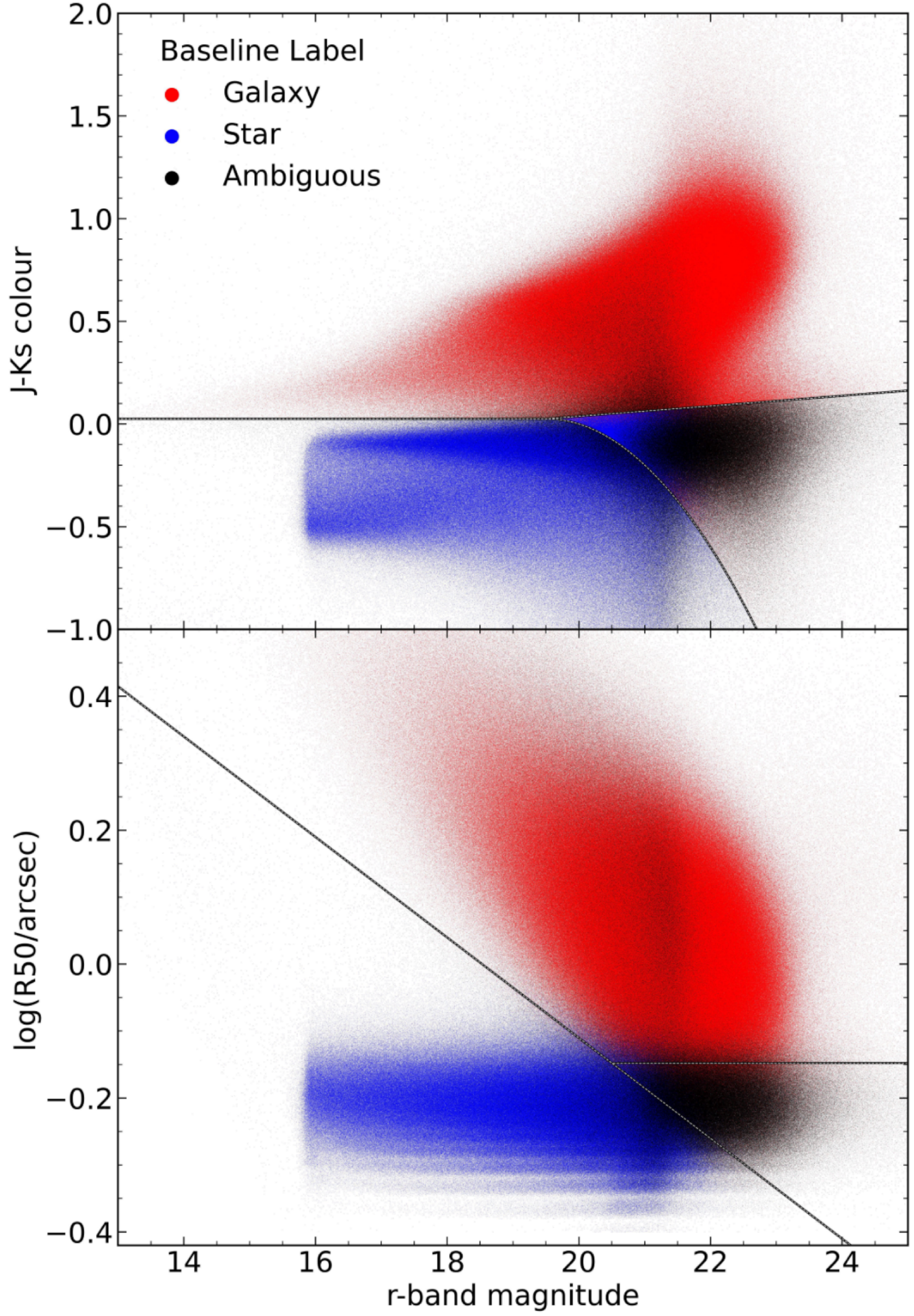
- 104,894 sources are observed across multiple surveys (GAMA, DESI, and SDSS) in the WAVES North region, with overlapping observations.
- The final ground-truth catalog comprises 370,248 stars and 338,402 galaxies, the figure illustrates how the number of sources with ground-truth labels compares to the total number of sources in the WAVES-Wide regions as a function of Z-band magnitude
- At the Z = 21.2 magnitude limit, only 0.27% of the sources have ground-truth labels





3. Star-Galaxy Separation

- The colours are derived from the ‘total’ magnitudes in different bands. This involves adding the total flux within the segment of each source, estimated on the detection bands, and then converting the flux to magnitudes. This is different from ‘colour’ magnitudes, which are derived using a fixed aperture across the multiple bands. For size information, the angular half-light radius *R*50 is used. This is the radius in arcseconds that contains half of the detection band flux within the segment.



3. Star-Galaxy Separation

- The baseline star-galaxy separation algorithm used by GAMA DR4 outlined in of Bellstedt et al. (2020), which uses very similar photometry. This algorithm utilises a combination of colour and size criteria.
- Lines are drawn to classify the sources into galaxy, star and ambiguous regions through the equations.

$$
\begin{aligned}
(J - K_s) &= 0.025, && \text{if } r < 19.5 \\
(J - K_s) &= 0.025 + 0.025(r - 19.5), && \text{if } r > 19.5 \\
(J - K_s) &= 0.025 - 0.1(r - 19.5)^2, && \text{if } r > 19.5,
\end{aligned}
\qquad (1)
$$

where the ambiguous region lies between the two lines beyond $r >$ 19.5 and:

$$
\begin{aligned}
\log(R_{50}) &= \Gamma + 0.05 - 0.075(r - 20.5), && \text{any } r \\
\log(R_{50}) &= \Gamma + 0.05, && \text{if } r > 20.5,
\end{aligned}
\qquad (2)
$$

where Γ is the median LOG10SEEING value, the log of the seeing in arcseconds in the detection band $(r + i + Z + Y)$. This value varies between -0.3 and -0.1 log(seeing/arcsec) depending on the tile.

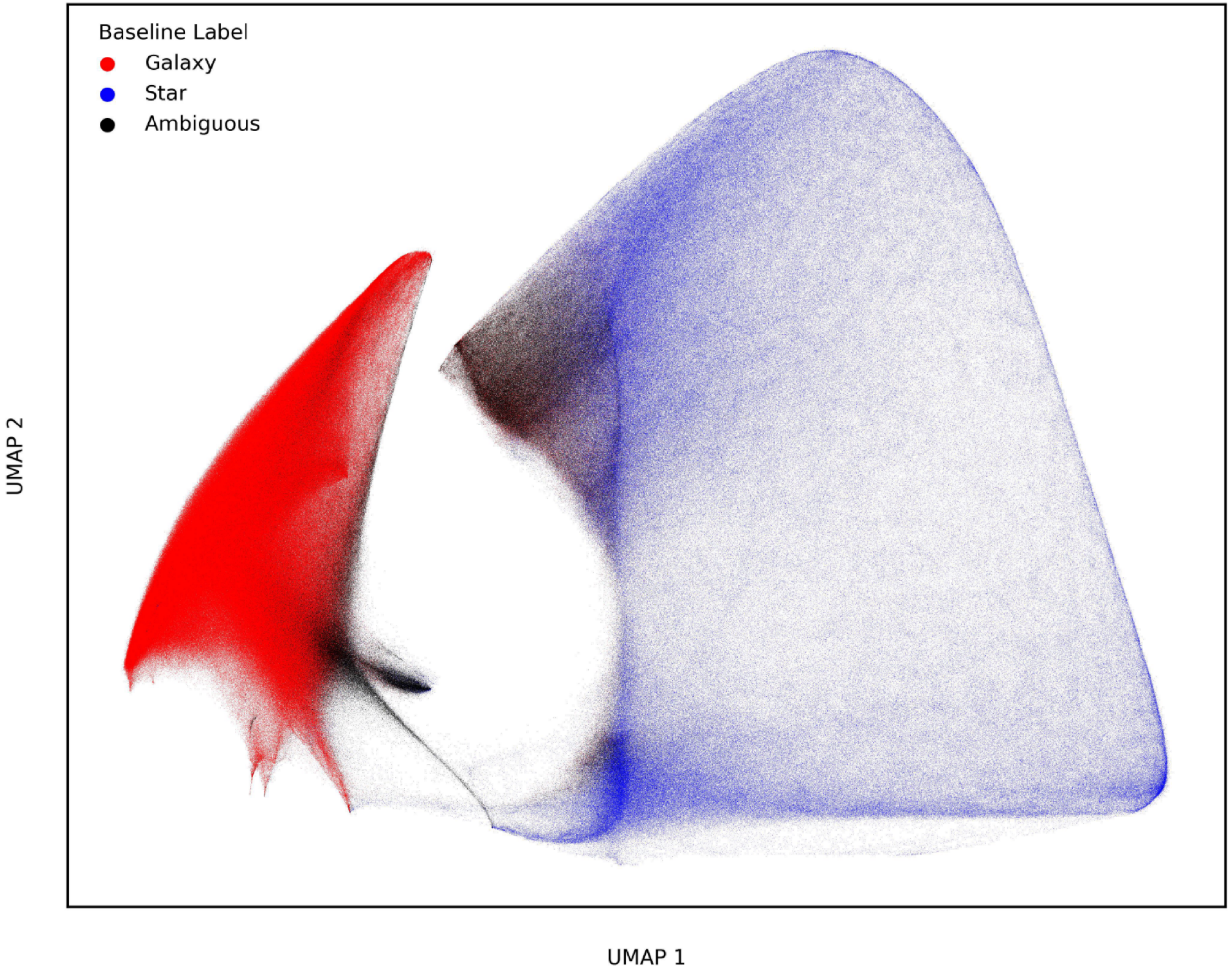
Preprocessing and Dimensionality Reduction

Preprocessing and dimensionality reduction are crucial for ensuring effective classification.

- Data cleaning：
  - Artefacts (sources with extremely unusual colors) are identified and removed
  - Sources missing data in any band are excluded since UMAP (the dimensionality reduction method) cannot handle missing values. This step eliminates 1.07% of the catalogue.
  - Sources with negative flux after sky subtraction are removed because magnitudes cannot be computed with negative flux. This step eliminates 0.73 % of the catalogue
- Feature formation
  - Includes magnitudes from 9 bands, 36 color combinations, logarithm of half-light radius (R_50), axial ratio, and astronomical seeing. A total of 48 features per source are used.
- Scaling
  - Data is scaled using **Z-score transformation** (mean of 0, unit variance) to ensure no feature dominates others, avoiding bias.

Dimensionality Reduction with UMAP

- **UMAP Methodology:**
  - Constructs a graph of *n*-nearest neighbors in high-dimensional space and projects it to a lower-dimensional embedding while preserving the graph structure.
- **UMAP Embeddings:**
  - In the WAVES-Wide South region, sources cluster into distinct nodes:
  - **Galaxies**: Left node.
  - **Stars**: Right node.
  - **Quasars (QSOs)**: Ambiguous sources densely populate areas attached to the galaxy node.
  - Blended sources (e.g., foreground stars contaminating background galaxies) appear between nodes.
- **Validation:**
  - The clustering in UMAP embeddings agrees with the baseline classification algorithm, indicating reliable performance.



Classification With HDBSCAN

- The overall classifications of the 14,802,032 sources is summarized in Table 1. If using the baseline classification scheme, wewould select sources classified as galaxy or ambiguous in order to ensure the required completeness. With UMAP/HDBSCAN (hereafter ‘cluster’) classification, we will simply select targets classified as galaxies.Moving to this new classifier will result in 1,672,758 fewer sources identified as galaxy or ambiguous, 11.3% of the catalogue.

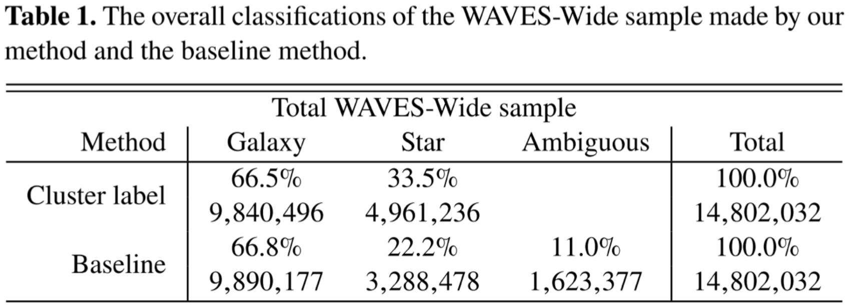
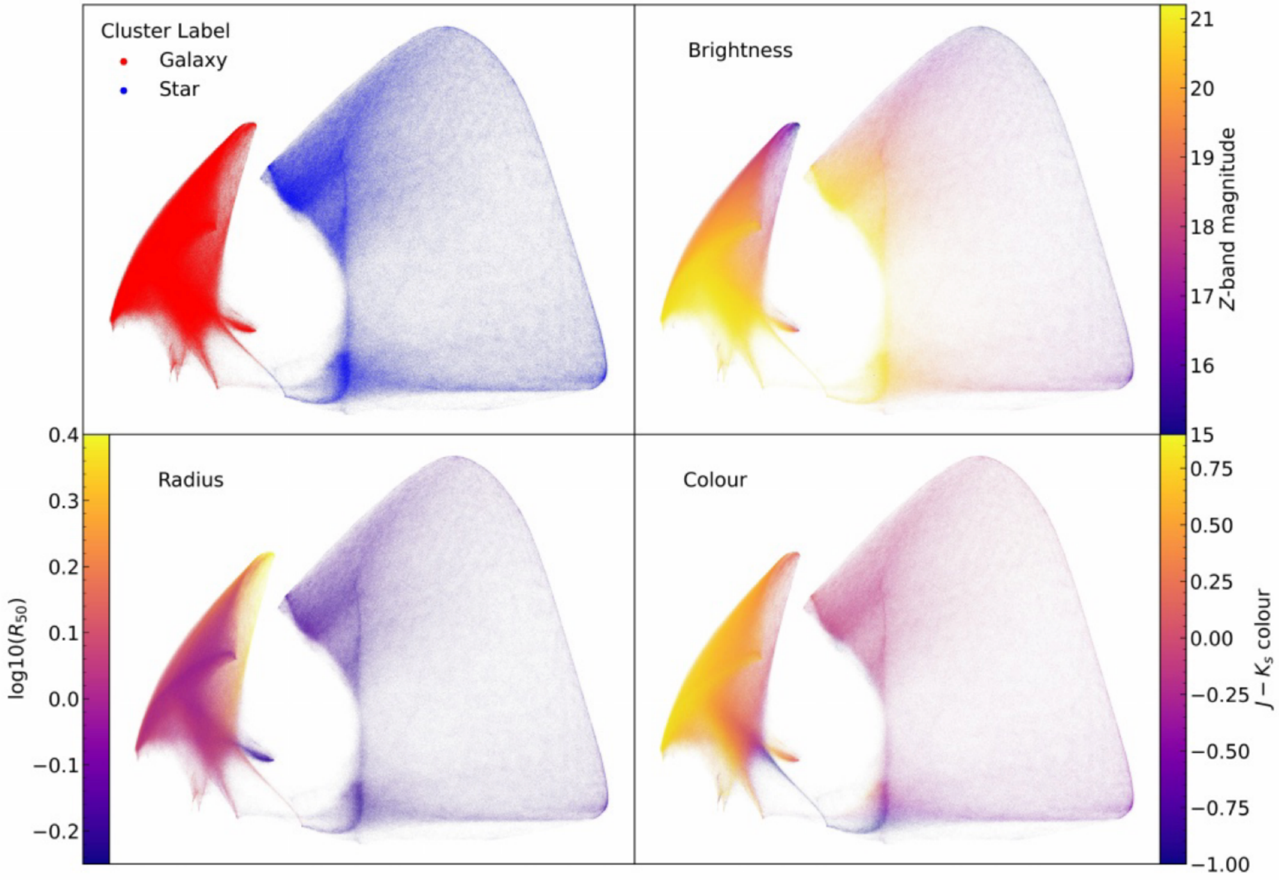


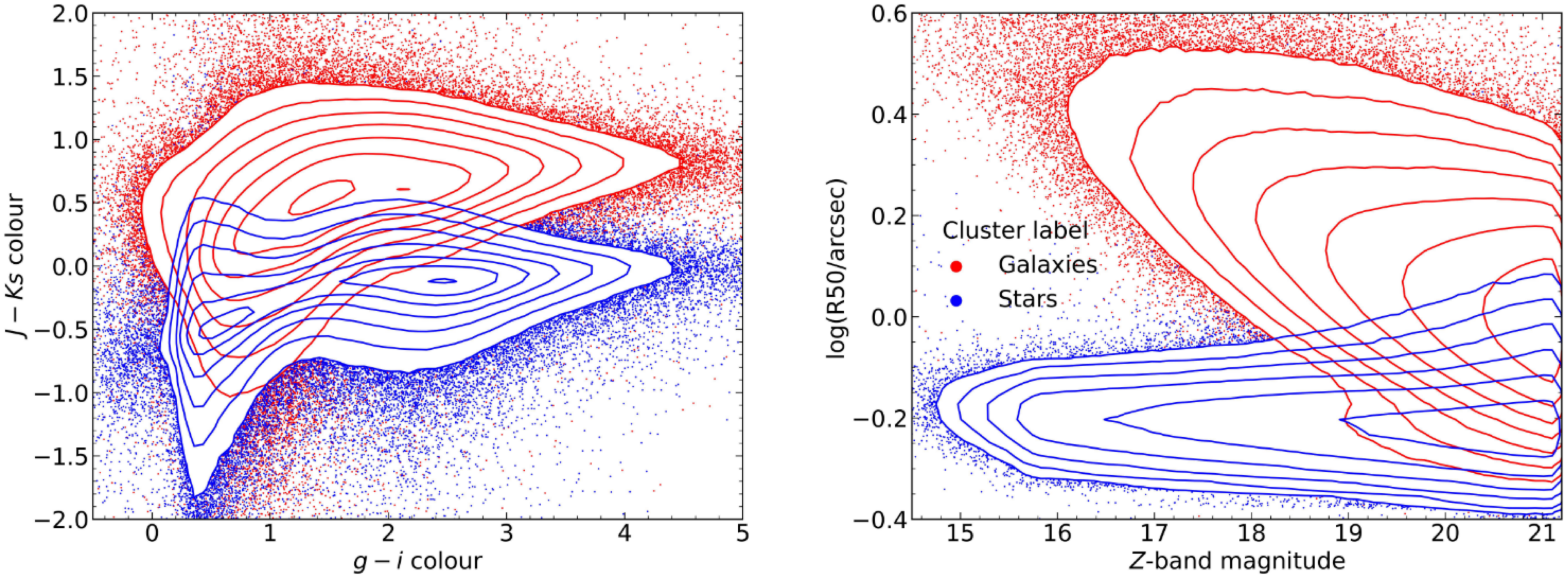
Table 1. The overall classifications of the WAVES-Wide sample made by our method and the baseline method.

	Total WAVES-Wide sample			
Method	Galaxy	Star	Ambiguous	Total
Cluster label	66.5% 9,840,496	33.5% 4,961,236		100.0% 14,802,032
Baseline	66.8% 9,890,177	22.2% 3,288,478	11.0% 1,623,377	100.0% 14,802,032

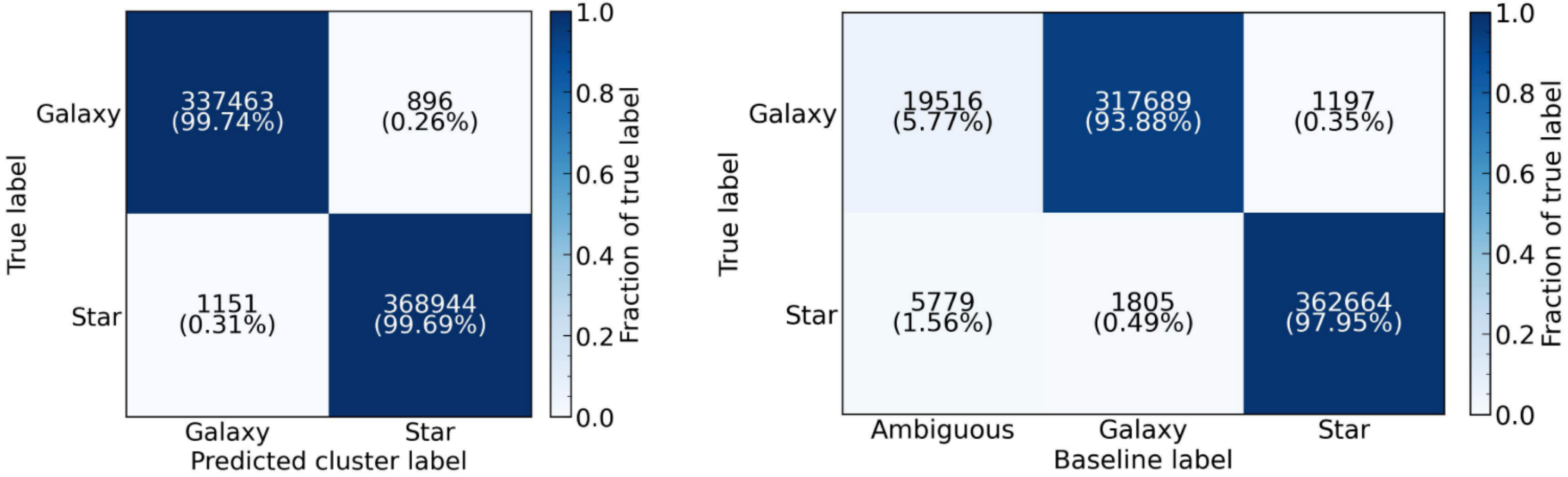


Classification Performance Using Ground Truth Labels

Observable properties of the sources labelled stars and galaxies by our classifier.
The left panel shows $J - Ks$ colour vs $g - i$ colour, and the right panel shows the log of half-light radius as a function of Z−band magnitude. Red and blue contours/points indicate galaxies and stars respectively. Contours are scaled logarithmically and the points show a random 10% of each population.



Classification Performance Using Ground Truth Labels



The confusion matrices generated by the ground-truth dataset. The left confusion matrix uses the labels generated by hdbscan, and the right confusion matrix uses the labels generated by the baseline algorithm described

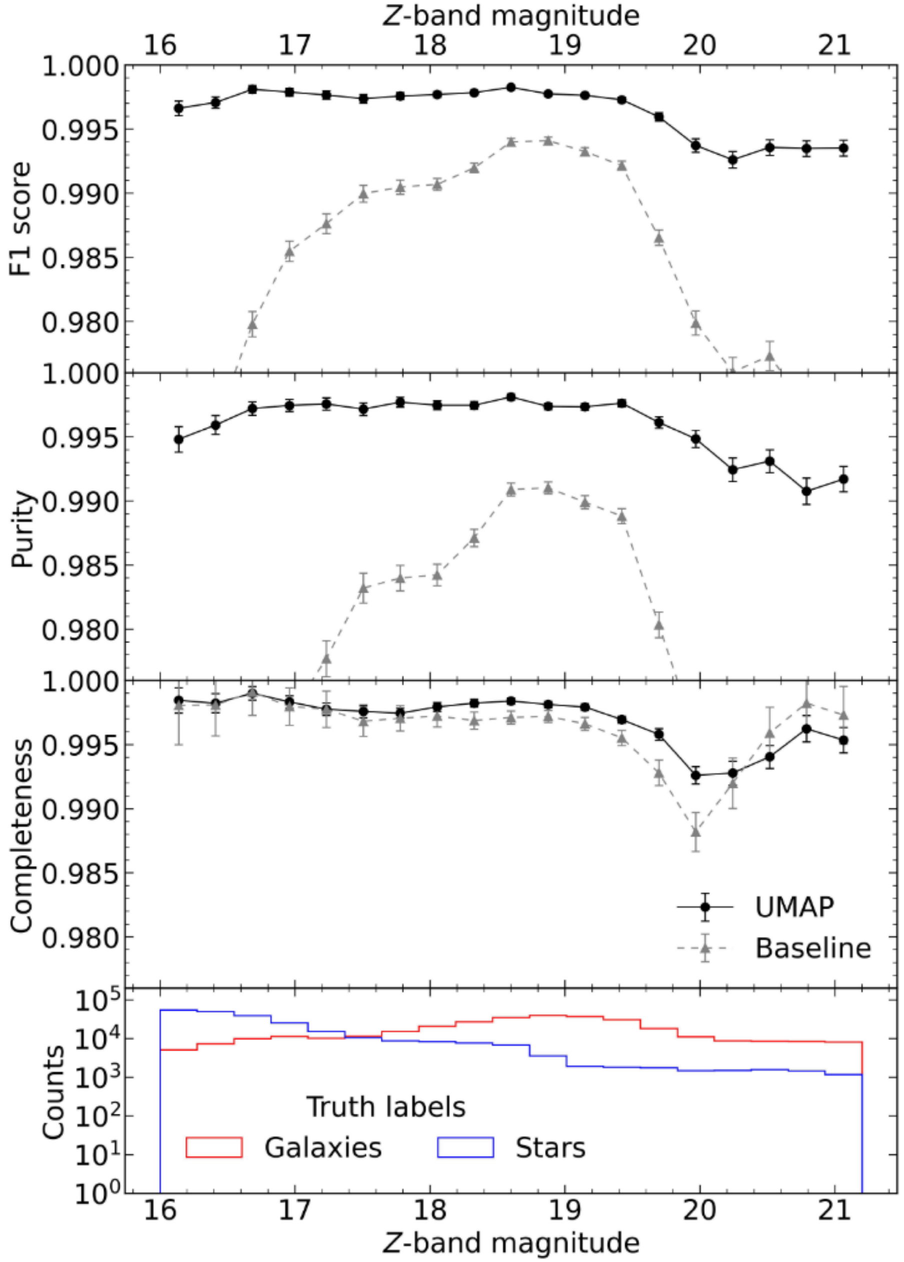


Table 2. Overall purity, completeness and F1 scores of the two methods, using all of the ground-truth labels.

Method	Purity	Completeness	F1
Cluster label	0.9966	0.9974	0.9970
Baseline[11]	0.9780	0.9964	0.9871

To quantify the fidelity of the method, we use the F1 score metric, which uses a harmonic mean of completeness and purity. In the context of identifying galaxies, we define a true positive (TP) as a true galaxy correctly identified, a false positive (FP) as a true star misclassified as a galaxy and a false negative (FN) as a true galaxy misclassified as a star. Purity is then defined as

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{3}$$

the fraction of true positives to total positives. Completeness is defined as

$$\mathrm{C} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{4}$$

the fraction of positive prediction to the total number of positives in the sample. These are combined to form the F1 score

$$\mathrm{F1} = 2\,\frac{\mathrm{P} \cdot \mathrm{C}}{\mathrm{P} + \mathrm{C}}, \tag{5}$$

SUMMARY

- We use unsupervised machine learning for star-galaxy separation of the WAVES-Wide input catalogue, and compare our results against a baseline method using a number of verification methods.
- We construct a catalogue of ground truth data for verification, formed from Gaia stars with high signal-to-noise parallax measurements, and stars and galaxies from GAMA, SDSS and DESI EDR spectroscopy. This gives us a sample of truth data even at faint magnitudes down to $Z < 21.2$.
- We utilise photometric features derived from the source-finding software ProFound, including KiDS and VIKING magnitudes, their colours, and sources' radii and axial ratios. A feature space is formed and reduced using UMAP, a non-linear dimensionality reduction algorithm, and then clustered into stars and galaxies using DBSCAN.
- Our method classifies 1,672,758 fewer sources as galaxy or ambiguous compared to the baseline method, or 11.3%, which is a associated with an approximate reduction of 70,000 fewer 4MOST fibre hours after photometric redshift cuts, and fewer suprious stars.

Thanks