

Radio Galaxy Zoo EMU: Harnessing Citizen Science and AI to Advance Open Science Catalogues

Eleni Vardoulaki¹, Hongming Tang², Micah Bowles³, Gary Segal^{4,5}, Soheb Mandhai⁶, Emma L. Alexander^{6,7}, Wendy Williams⁸, Yan Luo^{9,10}, Lawrence Rudnick¹¹, Andrew M. Hopkins¹², O. Ivy Wong^{13,14}, Stanislav S. Shabala¹⁵ and the RGZ EMU collaboration¹

¹IAASARS/National Observatory Athens, Hill of Nymfs, Athens 11810, Greece, ²Department of Physics, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China, ³Department of Astronomy, University of Oxford, Oxford, UK, ⁴School of Mathematics and Physics, University of Queensland, St Lucia, Brisbane, QLD 4072, Australia, ⁵CSIRO Space & Astronomy, P.O. Box 76, Epping, NSW 1710, Australia, ⁶Jodrell Bank Centre for Astrophysics, School of Physics and Astronomy, University of Manchester, Oxford Road, Manchester, M13 9PL, UK, ⁷School of Physics & Astronomy, University of Leeds, Leeds, LS2 9JT, UK, ⁸SKA Observatory, Jodrell Bank, Lower Whittington, Macclesfield, SK11 9FT, UK, ⁹Department of Astronomy, School of Physics, Peking University, Beijing 100871, China, ¹⁰Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing 100871, China, ¹¹Minnesota Institute for Astrophysics, University of Minnesota, 116 Church Street SE, Minneapolis, MN 55455, USA, ¹²School of Mathematical and Physical Sciences, 12 Wally's Walk, Macquarie University, NSW 2109, Australia, ¹³CSIRO Space and Astronomy, ATNF, POBox 1130, Bentley WA 6102, Australia, ¹⁴ICRAR-M468, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia, ¹⁵School of Physical Sciences, University of Tasmania, Private Bag 37, Hobart, Tasmania 7001, Australia

Abstract

Over the past decades, significant efforts have been devoted to developing sophisticated algorithms for automatically identifying and classifying radio sources in large surveys. However, even the most advanced methods face challenges in recognising complex radio structures and accurately associating radio emission with their host galaxies. Leveraging data from the ASKAP telescope and the Evolutionary Map of the Universe (EMU) survey, Radio Galaxy Zoo EMU (RGZ EMU) was created to generate high-quality radio source classifications for training deep learning models and cataloging millions of radio sources in the southern sky. By integrating novel machine learning techniques, including anomaly detection and natural language processing, our workflow actively engages citizen scientists to enhance classification accuracy. We present results from Phase I of the project and discuss how these data will contribute to improving open science catalogues like EMUCAT.

Keywords— citizen science, radio surveys, AI, open science

1 Introduction

The new generation of wide-field radio surveys, such as the Evolutionary Map of the Universe [Norris et al., 2011, 2021b, Hopkins et al., 2025, EMU], is transforming our ability to study active galactic nuclei (AGN), galaxy evolution, cosmic large-scale structure and rare astrophysical phenomena. Conducted with the Australian Square Kilometre Array Pathfinder (ASKAP), EMU will map the entire southern sky up to $+30^\circ$ declination. The survey is expected to detect tens of millions of radio sources, providing a legacy dataset of unprecedented scope.

Although automatic algorithms efficiently classify compact and unresolved objects, extended and morphologically complex sources remain a challenge [Mohan and Rafferty, 2015, Boyce et al., 2023a,b]. To address this, the Radio Galaxy Zoo EMU (RGZ EMU) project was officially launched in August 2024, after internal testing for more than two years, as a live citizen science initiative that integrates human pattern recognition with artificial intelligence (AI). The goal of RGZ EMU is to integrate its results into the open science ready catalogue for EMU, namely EMUCAT (Marvil et al., in prep.).

In the era of big data and all-sky surveys, citizen science offers a powerful means to support research by providing robust training samples for machine and deep learning algorithms. However, even with the efforts

of tens of thousands of volunteers, the task of classifying millions of radio sources would require centuries to complete if carried out manually. To overcome this challenge, RGZ EMU adopts a quasi-automated framework (see Fig. 1) that integrates the strengths of both citizen science and machine learning, enabling the efficient identification and classification of radio sources at the scale demanded by modern surveys [Tang et al., 2025].

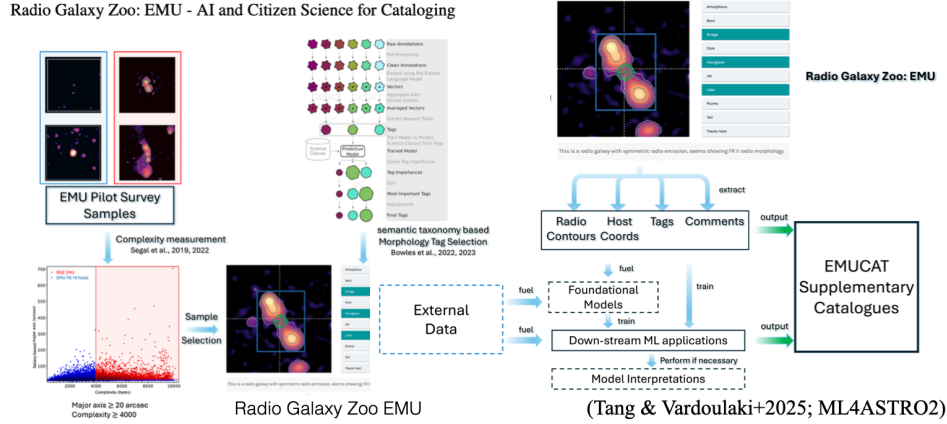


Figure 1: The RGZ EMU project framework: integrating citizen science with machine learning to classify extended radio sources in the EMU survey. Adopted from Tang et al. [2025].

2 Sample

The EMU survey, conducted with ASKAP, is designed to produce the benchmark radio atlas of the southern hemisphere. ASKAP comprises 36 12-metre antennas equipped with phased-array feeds, providing a 30 deg^2 instantaneous field of view. This wide coverage enables rapid surveys of the sky at $\sim \text{GHz}$, detecting not only AGN and star-forming galaxies but also unexpected sources such as Odd Radio Circles [Norris et al., 2021a, ORCs], relics, and cluster bridges.

Phase I of RGZ EMU uses data from the EMU Pilot Survey [Norris et al., 2021b, EMU-PS], which achieved an rms sensitivity of $25\text{--}30 \mu\text{Jy beam}^{-1}$ at a spatial resolution of $\sim 11\text{--}18 \text{ arcsec}$, and contains 287,555 radio components identified with the Selavy source finder [Whiting and Humphreys, 2012]. The radio data are complemented by optical observations from the Dark Energy Survey [Dark Energy Survey Collaboration, 2016, DES] and mid-infrared data from the AllWISE catalogue at $3.4 \mu\text{m}$ [Cutri et al., 2013]. To maximise scientific diversity, subjects were selected according to both morphological complexity [Segal et al., 2023] and angular size, the latter measured by the Selavy catalogue. The resulting Phase I dataset of 6,230 extended sources, spanning simpler to highly irregular morphologies, forms the input to the Zooniverse platform for citizen science classification.

3 Methodology

RGZ EMU employs a hybrid workflow (see Fig. 2) in which citizen science annotations will enhance AI pipelines. Sources with high image complexity are prioritised for classification. Following Segal et al. [2023], complexity is quantified using a coarse-grained Kolmogorov measure: sources with rich substructure compress less efficiently and preserve more information across scales. In practice, $6' \times 6'$ cutouts with higher complexity values are selected, ensuring that volunteers see the most challenging morphologies. To refine this selection, projected angular size (from Selavy) is combined with complexity, thereby targeting the most interesting sources for citizen scientists while leaving simple, compact sources to automated pipelines. Volunteers mark related radio components, identify host galaxies, and assign simple descriptive morphological tags [Rudnick, 2021]. The tags were derived using a semantic taxonomy [Bowles et al., 2022, 2023], which bridges descriptive language and machine-readable classification. An ontology of 22 morphological descriptors was refined into a set of ~ 10 tags for RGZ EMU. These tags (e.g. “hourglass”, “bent”, “traces host galaxy”) allow citizen scientists to classify complex structures intuitively, while creating a shared language that can be ingested by machine learning and linked to scientific categories such as active or star-forming galaxies.

As the project evolves, active learning loops will integrate citizen science results into training sets, allowing iterative improvement of AI pipelines. The outputs will feed into EMUCAT (Marvil et al. in prep.), producing a more complete and reliable science-ready catalogue.

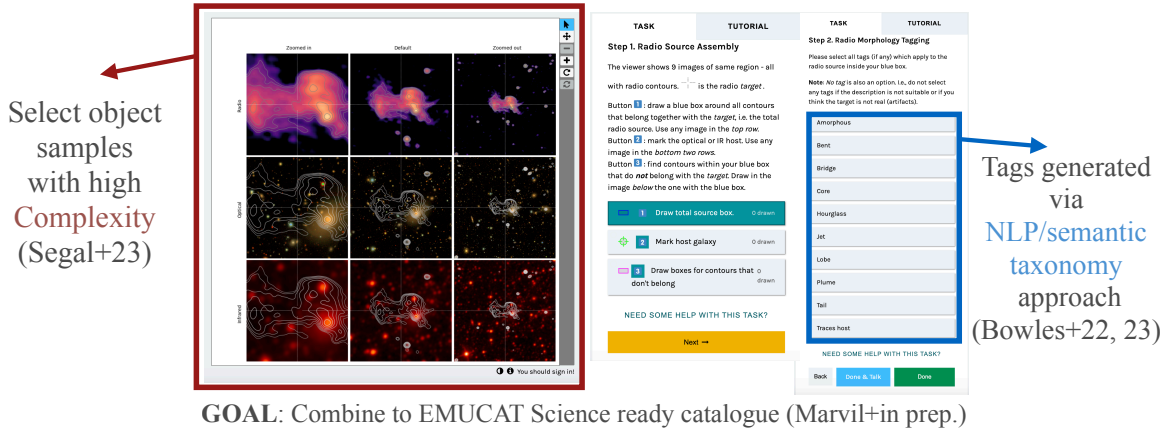


Figure 2: The RGZ EMU workflow asks citizen scientists to identify radio sources and their host galaxies, and to classify them using simple descriptive tags. These results will be feed into active learning pipelines, improving automated cataloguing of complex radio sources.

4 Discussion

Early results from over 2,500 volunteers and 97,000 classifications show that combining citizen science with AI improves the identification of complex morphologies compared to automated pipelines alone (Tang and Vardoulaki in prep.). Citizen scientists are particularly effective at recognising rare or ambiguous sources, and the project’s multilingual deployment (Greek, Chinese, Urdu, Japanese, with further languages in prep.) highlights its open and inclusive character. Scaling to millions of sources remains a challenge, especially in terms of server and storage access, but feasibility has already been demonstrated.

Beyond catalogue building, citizen science provides an open channel for discovery and ensures diversity in training data, reducing potential biases from a limited classifier pool. Preliminary Phase I results (Tang and Vardoulaki in prep.), consistent with earlier findings [Willett, 2016], indicate that once a sufficient number of independent classifications is obtained, the collective accuracy of volunteers approaches that of expert consensus. This supports the use of “golden samples” as benchmarks and feedback loops for machine learning, enabling the creation of high-quality catalogues and contributing to the development of trustworthy, interpretable AI methods.

5 Outlook

Over the next five years, RGZ EMU will expand to the full EMU survey, classifying around four million extended sources. The project provides a testbed for trustworthy AI in preparation for the SKA era and strengthens education and outreach through the RADIIIO program (PI: Vardoulaki, OAD-IAU funded), positioning RGZ EMU at the frontier of open science. Its combination of citizen science and machine learning exemplifies “hybrid intelligence”, where algorithms manage large-scale tasks while human expertise guides rare or complex cases. Links with surveys such as POSSUM, WALLABY, PEGASUS and Euclid, will enable multi-wavelength studies, offering deeper insight into galaxy evolution and feedback.

Through its training, education, and outreach initiatives, RGZ EMU lowers barriers to public participation and creates opportunities for students and early-career researchers, serving not only as a scientific tool but also as a model for inclusive, global science and education in the 21st century.

References

- Micah Bowles, Hongming Tang, Eleni Vardoulaki, and et al. A New Task: Deriving Semantic Class Targets for the Physical Sciences. *arXiv e-prints*, art. arXiv:2210.14760, October 2022. doi: 10.48550/arXiv.2210.14760.
- Micah Bowles, Hongming Tang, Eleni Vardoulaki, and et al. Radio galaxy zoo EMU: towards a semantic radio galaxy morphology taxonomy. *MNRAS*, 522(2):2584–2600, June 2023. doi: 10.1093/mnras/stad1021.
- M. M. Boyce, A. M. Hopkins, S. Riggi, and et al. Hydra I: An extensible multi-source-finder comparison and cataloguing tool. *PASA*, 40:e028, July 2023a. doi: 10.1017/pasa.2023.24.

- M. M. Boyce, A. M. Hopkins, S. Riggi, and et al. Hydra II: Characterisation of Aegean, Caesar, ProFound, PyBDSF, and Selavy source finders. *PASA*, 40:e027, July 2023b. doi: 10.1017/pasa.2023.29.
- R. M. Cutri, E. L. Wright, T. Conrow, and et al. Explanatory Supplement to the AllWISE Data Release Products. Explanatory Supplement to the AllWISE Data Release Products, by R. M. Cutri et al., November 2013.
- Dark Energy Survey Collaboration. The Dark Energy Survey: more than dark energy - an overview. *MNRAS*, 460(2):1270–1299, August 2016. doi: 10.1093/mnras/stw641.
- Andrew Hopkins, Anna Kapinska, Joshua Marvil, and et al. The Evolutionary Map of the Universe: A new radio atlas for the southern hemisphere sky. *PASA*, 42:e071, May 2025. doi: 10.1017/pasa.2025.10042.
- Niruj Mohan and David Rafferty. PyBDSF: Python Blob Detection and Source Finder. Astrophysics Source Code Library, record ascl:1502.007, February 2015.
- Ray P. Norris, A. M. Hopkins, J. Afonso, and et al. EMU: Evolutionary Map of the Universe. *PASA*, 28(3): 215–248, August 2011. doi: 10.1071/AS11021.
- Ray P. Norris, Evan Crawford, and Peter Macgregor. Odd Radio Circles and Their Environment. *Galaxies*, 9(4):83, October 2021a. doi: 10.3390/galaxies9040083.
- Ray P. Norris, Joshua Marvil, J. D. Collier, and et al. The Evolutionary Map of the Universe pilot survey. *PASA*, 38:e046, September 2021b. doi: 10.1017/pasa.2021.42.
- Lawrence Rudnick. Radio Galaxy Classification: #Tags, Not Boxes. *Galaxies*, 9(4):85, October 2021. doi: 10.3390/galaxies9040085.
- Gary Segal, David Parkinson, Ray Norris, and et al. Identifying anomalous radio sources in the Evolutionary Map of the Universe Pilot Survey using a complexity-based approach. *MNRAS*, 521(1):1429–1447, May 2023. doi: 10.1093/mnras/stad537.
- Hongming Tang, Eleni Vardoulaki, and RGZ EMU collaboration. Radio Galaxy Zoo: EMU – paving the way for EMU cataloging using AI and citizen science. *arXiv e-prints*, art. arXiv:2506.16138, June 2025. doi: 10.48550/arXiv.2506.16138.
- M. Whiting and B. Humphreys. Source-Finding for the Australian Square Kilometre Array Pathfinder. *PASA*, 29(3):371–381, August 2012. doi: 10.1071/AS12028.
- Kyle W. Willett. Radio Galaxy Zoo: host galaxies and radio morphologies for large surveys from visual inspection. *arXiv e-prints*, art. arXiv:1603.02645, March 2016. doi: 10.48550/arXiv.1603.02645.