2027 PhD Project Proposal

Cover page

1. Project Title

   Unsupervised Machine Learning for Multiwavelength Data with Foundation Models

2. Research Area

   Science

3. Academic Level

   PhD

4. Primary Supervisor's Details
   a. **Full name:** Dr. Michelle Lochner
   b. **Email address:** mlochner@uwc.ac.za
   c. **University:** University of the Western Cape

5. Abstract

Upcoming telescopes such as the Square Kilometre Array (SKA) and the Vera C. Rubin Observatory (LSST) are producing datasets of unprecedented size. It is becoming quickly apparent that it is not feasible to run bespoke machine learning algorithms on every dataset, and traditional approaches often rely on time-consuming human labelling. Foundation models–large, general-purpose machine learning algorithms trained on vast quantities of data to solve diverse problems–hold a lot of promise for handling this "data deluge".

This project proposes to gather a large quantity of unlabelled data, focusing on MeerKAT, to train and adapt foundation models for radio astronomy. We will leverage existing labelled data from surveys like EMU, LoTSS and MGCLS to act as benchmarks for applications using the foundation models. We will cross-match these radio datasets with existing optical data from DECALS and eventually new LSST releases. The student will adapt existing foundation models, such as DINO, AstroCLIP or AION, for this combined radio and optical dataset. These models will then be tested on downstream tasks including classification, clustering and anomaly detection paving the way for large scale data mining in the SKA era.

Details of research project

1. Scientific Merit

Machine learning is critical to handling the data deluge from modern telescopes. However, to train robust algorithms usually requires a large amount of labelled data, which is expensive to obtain. Foundation models do not require labelled data for training, allowing us to utilise the vast archives of data available from MeerKAT, ASKAP and LOFAR.

We will gather data into a large collection that will in itself be extremely scientifically valuable as well as useful for training machine learning. While foundation models rely on unlabelled data, we will use small labelled datasets for benchmarking. Furthermore, cross-matching with DECALS and eventually LSST will be critical for full scientific exploitation. Currently, no one has researched how to combine foundation models from each wavelength effectively. As the supervisor is a South African Principal Investigator for LSST, the student will have access to early data, placing them at the forefront of multiwavelength discovery.

Finally, this project addresses the skills shortage in the "fourth industrial revolution" in South Africa. It will develop critical local data science skills and pave the way for use of these techniques with the SKA.

2. Feasibility and Methods

Data Availability:

We will gather a large dataset of radio images from MALS, MeerKLASS, LoTSS, and EMU. These datasets are either already public or, in the case of MeerKLASS, access has been secured through the PI, Prof. Mario Santos. Optical data will be sourced from DECALS and LSST (via the supervisor's PI status). Computing resources are available through UWC's membership of IDIA (the Ilifu supercomputer).

Methodology:

The student will start by collecting the unlabelled and labelled radio images and cross-matching them with optical data. We will focus on simple positional cross-matching as we are exclusively interested in morphology. They will then develop classification and anomaly detection benchmarks using the labelled subsets.

The core work involves running or adapting existing foundation models such as the Radio Galaxy Zoo model, DINO, AstroCLIP or AION. The student will evaluate the performance of these models on classification, clustering and anomaly detection, developing insights into how foundation models behave when combining optical and radio data. If additional labels are needed for benchmarking, we may make use of citizen science platforms.

Downstream applications include the discovery of new classes of objects with enhanced information from the multiwavelength view, the creation of morphology classification catalogues, and potential improvements to tasks like redshift estimation.

3. Skills development

This project will develop high-level skills including:

- Radio and optical image analysis.
- Machine learning skills (specifically foundation models and unsupervised learning).
- Programming skills (Python, PyTorch, Astronomaly).
- Engagement with large international collaborations (MeerKLASS and LSST).

4. Availability of data and resources

The project relies on public data or data where access is already secured.

Radio: Public surveys (LoTSS, EMU) and MeerKLASS (via Mario Santos).

Optical: DECALS (public) and LSST (via Lochner).

Resources: Access to Ilifu supercomputer through UWC's membership of IDIA; Office space at UWC; Engagement with the Astro group at UWC.

5. Milestones and Timeline

Year 1: Collate and process data including potential additional source-finding. Cross match with DECALS and LSST where possible.

Year 2: Apply existing foundation models to both datasets, combining features from optical and radio data.

Year 3: Benchmark foundation models against classification, clustering and anomaly detection tasks. Publish paper and write up thesis.

6. Student Requirements

Good programming skills are critical for this project, primarily in python. Experience with machine learning is advantageous, but not essential. Experience working with either radio or optical images is advantageous.

References

- https://arxiv.org/abs/2305.16127 (Slijepcevic et al., 2023)
- https://arxiv.org/abs/2310.03024 (AstroCLIP: Parker et al., 2023)
- https://arxiv.org/abs/2508.10104 (DINOv3: Siméoni et al., 2025)
- https://arxiv.org/abs/2010.11202 (Astronomaly: M. Lochner & B. Bassett, 2020)
- https://arxiv.org/abs/2311.14157 (Mohale & Lochner, 2024)
- https://arxiv.org/abs/2102.08414 (Walmsley et al., 2021 - Zoobot)

Gmail **Michelle Lochner <dr.michelle.lochner@gmail.com>**

Confirmation of access to MeerKLASS data

Mario Santos <mariogrs@gmail.com> 17 February 2026 at 09:35
To: Michelle Lochner <dr.michelle.lochner@gmail.com>

Hi Michelle,

I can confirm that I am aware of the proposal and the student will be provided access to the necessary MeerKLASS data.

Kind regards,
Mario

On Tue, 17 Feb 2026 at 09:26, Michelle Lochner <dr.michelle.lochner@gmail.com> wrote:

> Hi Mario
>
> As we discussed, can you please confirm by email that I and my potential new PhD student in 2027 (for whom I am putting in a SARAO proposal) would have access to MeerKLASS continuum images for a project focusing on training multiwavelength foundation models. We will both of course join MeerKLASS and engage with the collaboration, especially with Suman who has been leading the processing of the image data.
>
> Cheers
> Michelle