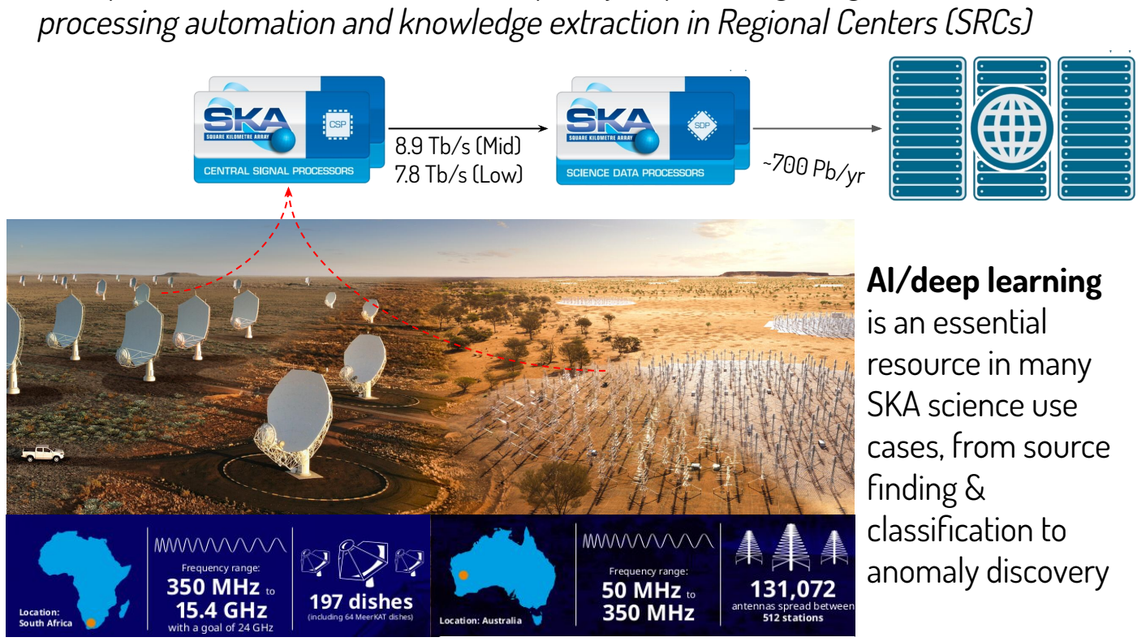
2º Forum della Ricerca Sperimentale e Tecnologica

SCIARADA : Self-supervised Contrastive learning for Inspection and Analysis of RAdio DAta in the SKA era

S. Riggi, *T. Cecconello*
INAF-OACT

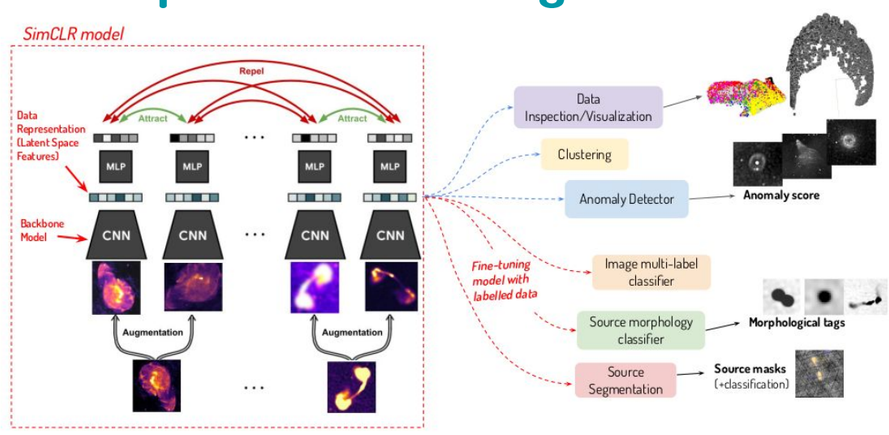
Context

SKA will be the largest radio telescope ever built
Its unprecedented data volume & complexity require a high degree of data processing automation and knowledge extraction in Regional Centers (SRCs)



AI/deep learning is an essential resource in many SKA science use cases, from source finding & classification to anomaly discovery

Self-Supervised Learning for radio



- **Supervised ML methods have limitations**
  - *Radio labelled datasets are often small & class-unbalanced*
  - *Labelling schema usually varying with the science case*
- **SSL methods learn from unlabelled data without supervision**
  - *Model & data representation learnt can then be used on small samples of labeled data for data inspection and different analysis tasks*

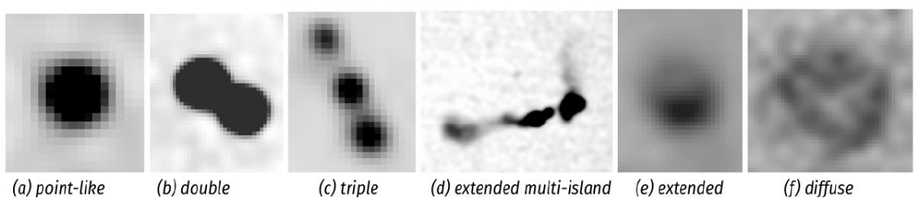
We have a huge amount of unlabelled radio survey data to exploit!

The SCIARADA Project

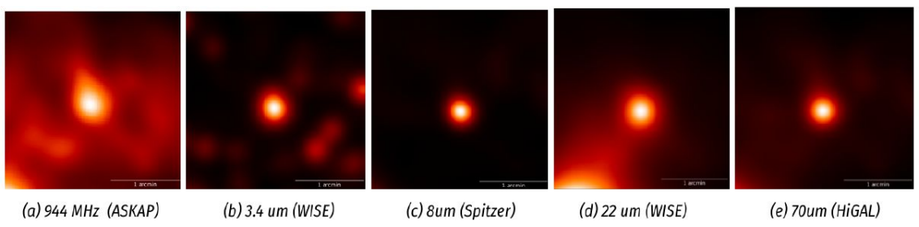
INAF Minigrant 2023

Tech goals

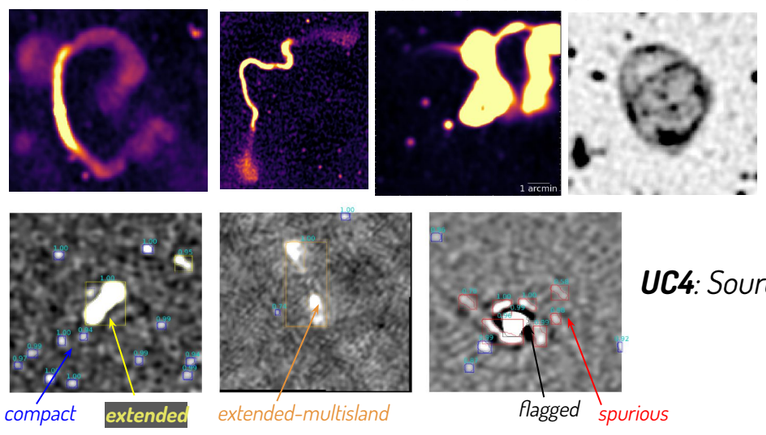
- ✓ *Training and evaluating SSL methods on selected analysis cases (source detection, classification, anomaly detection) using SKA precursor data (ASKAP, MeerKAT, etc);*
- ✓ *Delivering AI tools for SKA SRCs*



***UC1**: Source morphology classification*



***UC2:** Multi-wavelength source classification*



***UC3:** Peculiar object discovery*

***UC4**: Source detection*

Benefits for SKA science teams

- ✓ *produce advanced source catalogs and annotated datasets in shorter times;*
- ✓ *identify interesting, unexpected/peculiar objects, prioritizing follow-up studies;*
- ✓ *discover new Galactic sources, increasing their census.*

Work in-progress

- **Increasing size of pre-training radio datasets**
  - ✓ *Reached 1.5 M images with EMU main survey but curation strategies needed*
- **SSL model benchmark**
  - ✓ *All4One best, SimCLR/BYOL very close, no benefits with deeper networks (ResNet50)*
- **Evaluating vision-only & vision-language models (VLM) pretrained on web images**
  - ✓ *clear improvement over ImageNet by ~5%*
  - ✓ *ViT-based models (e.g. SigLIP) almost reaching performances of radio SSL models*
- **Fine-tuning small VLMs on radio image-instruction data**

INAF - Osservatorio Astrofisico di Catania
Via S. Sofia 78, 95123 Catania - Italy
+39 095 7332282
simone.riggi@inaf.it
simone.riggi
https://www.researchgate.net/profile/Simone-Riggi-2
http://www.linkedin.com/in/simoneriggi
https://github.com/simoneriggi
https://orcid.org/0000-0001-6368-8330

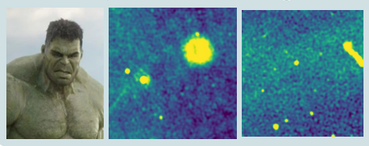
Results

SSL Model Pre-Training

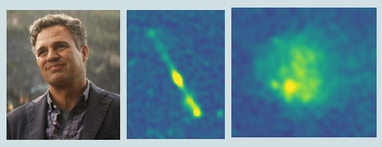
S. Riggi et al, PASA (2024)
https://arxiv.org/abs/2404.18462

Model: SimCLR (ResNet18)
- ✓ *Pre-processing:* 3-chans + minmax(0,1) + resize(224x224)
- ✓ *Augmentations:* crop, rotate, flip, blur, color jitter, random thresholding

Hulk datasets
random cutouts (256 x 256)
- ✓ SMGPS: *~178,000 images*
- ✓ EMU pilot: *~56,000 images*

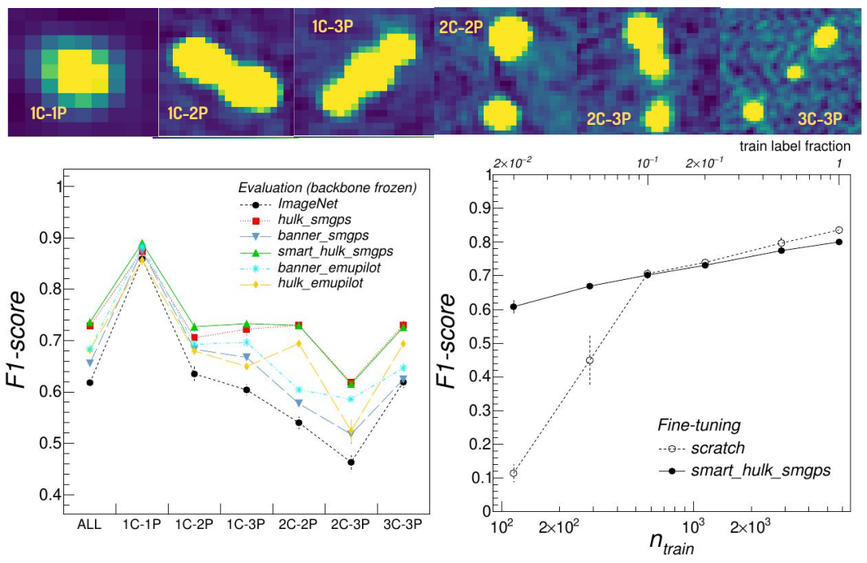


Banner datasets
cutouts centred and zoomed on extended sources
- ✓ SMGPS: *~17,000 images*
- ✓ EMU pilot: *~10,000 images*



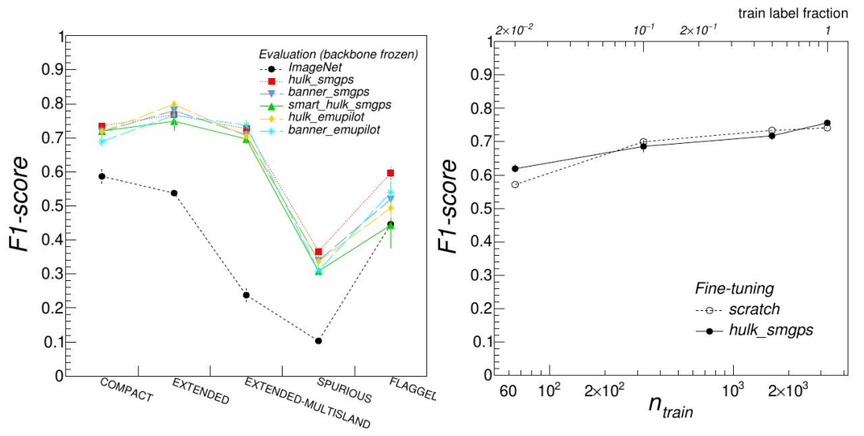
Source Morphology Classification

Method: *CNN classifier*
Dataset: *RGZ DR1*
Images/class: *1000 (train), 600 (test)*
Classes: *1C-1P, 1C-2P, 1C-3P, 2C-2P, 2C-3P, 3C-3P*
Surveys: *VLA FIRST*



Source Detection

Method: *Mask R-CNN*
Dataset: *rg-dataset*
Images: *~12,000*
Objects: *~38,000*
Classes: *SPURIOUS, COMPACT, EXTENDED, EXTENDED-MULTISLAND, FLAGGED*
Surveys: *ATCA Scorpio, VLA FIRST, ASKAP EMU*



Anomaly Detection

Method: *Isolation Forest*
Dataset: *hulk-emupilot SSL features*
Images: *~56,000*
Surveys: *ASKAP EMU pilot*

Features	Thr.	$\mathcal{R}$ (%)	$\mathcal{P}$ (%)	$\mathcal{P}_{pec+complex}$ (%)	$\mathcal{C}_{complex}$ (%)	$\mathcal{C}_{ordinary}$ (%)
top2	0.700	55.6	59.8	93.5	33.7	6.5
top5	0.750	61.2	63.0	93.0	30.0	7.0
top10	0.725	59.1	58.7	97.4	38.7	2.6
top15	0.660	57.7	58.7	95.5	36.8	4.5