UNIVERSITÉ DE GENÈVE

Self-Supervised Learning for MeerKAT Images

E. Lastufka

M. Audard, O. Bait, M. Dessauges-Zavadsky, M. Drozdova, T. Holotyak, V. Kinakh, D. Piras, O. Taran, D. Schaerer, S. Voloshynovskiy

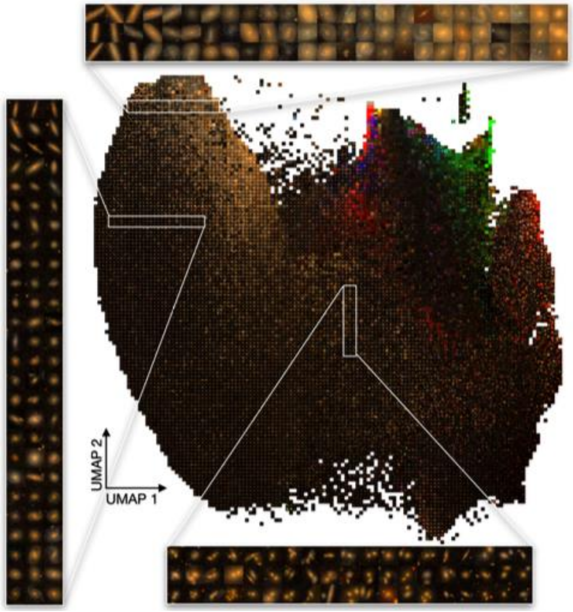
SKACH

Self-supervised learning is crucial for training foundation models

Train

Data (images)

SSL network



Hayat et al 2021

Evaluate

- Morphology classification
- Similarity search
- Redshift prediction
- Anomaly detection

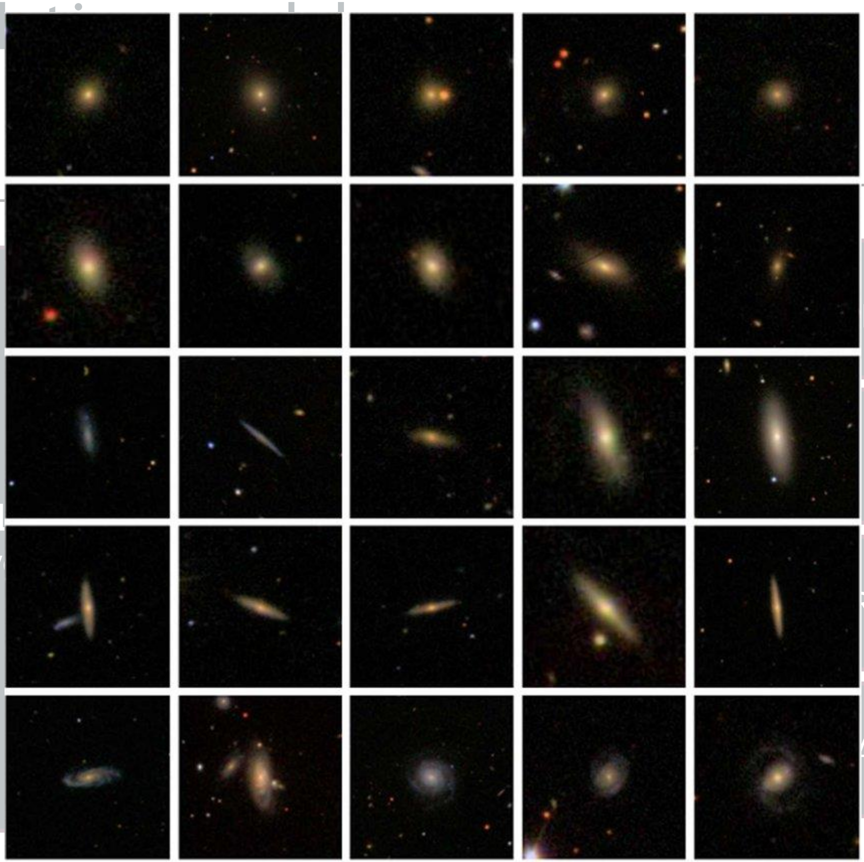
What is a found

Train

Data
(images)

SSI
netw

Galaxy Zoo SDSS (scales vary)



Evaluate

phology classification

Similarity search

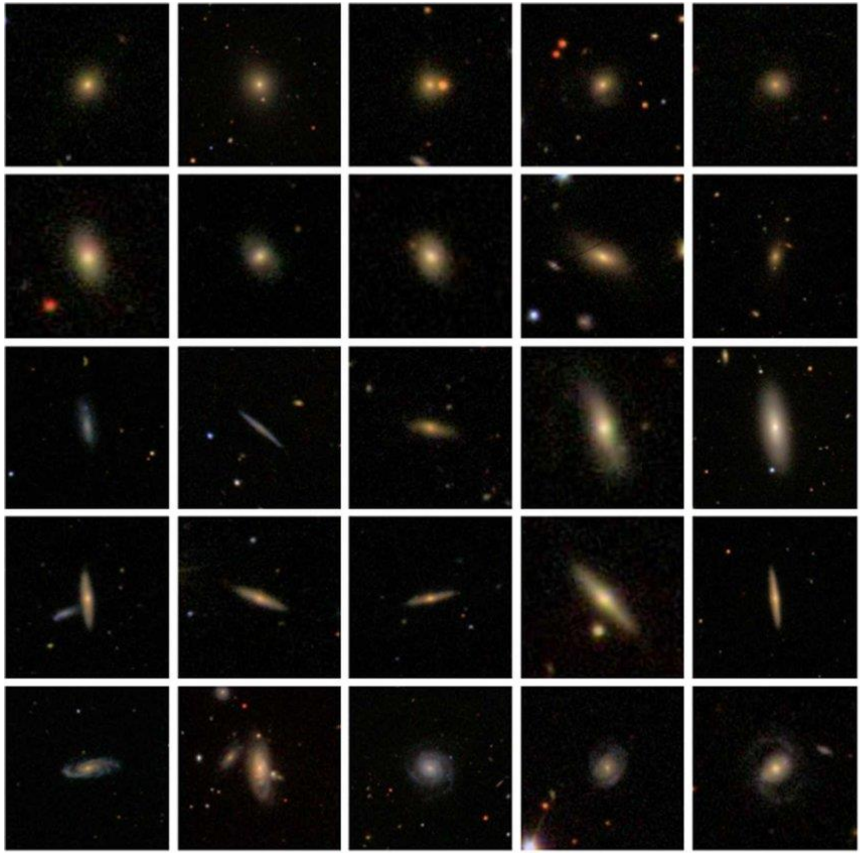
Redshift prediction

Anomaly detection

Current foundation models in astronomy

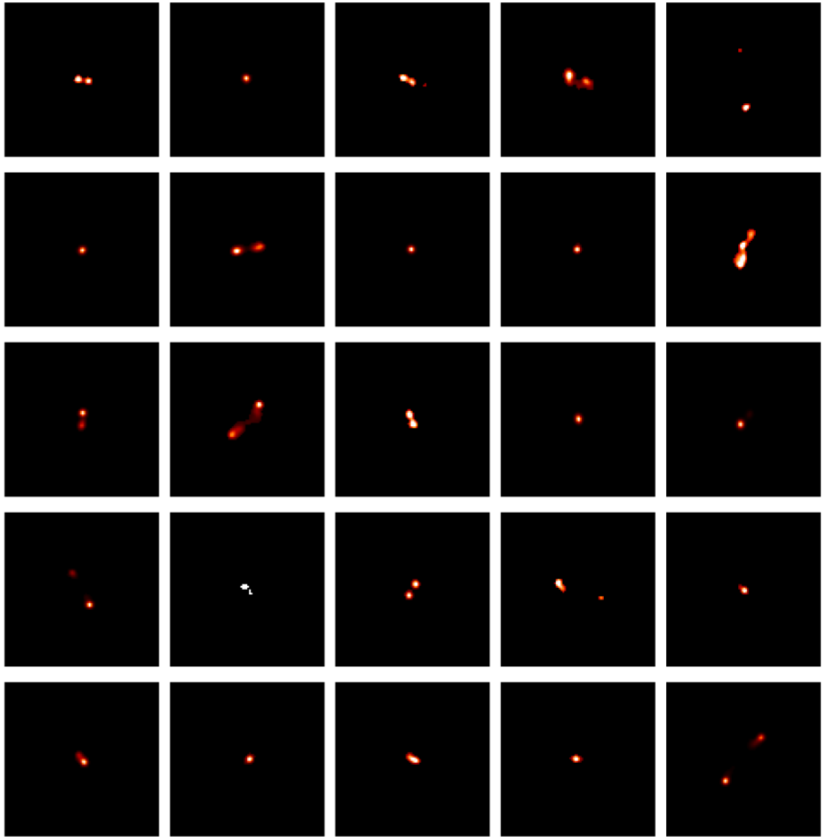
Regime	Work	Dataset	Network
Optical	Hayat et al 2021	Galaxy Zoo 2 (SDSS)	MoCo v2, Resnet50
Optical	Mohale & Lochner 2023	Galaxy Zoo DECaLS	BYOL, Resnet50
Radio	Slijepcevic et al 2023	Radio Galaxy Zoo	BYOL, Resnet18
Radio	Andrianomena & Tang 2023	Radio Galaxy Zoo	VDVAE, SimCLR BYOL, SimSiam Resnet34

Galaxy Zoo SDSS (scales vary)

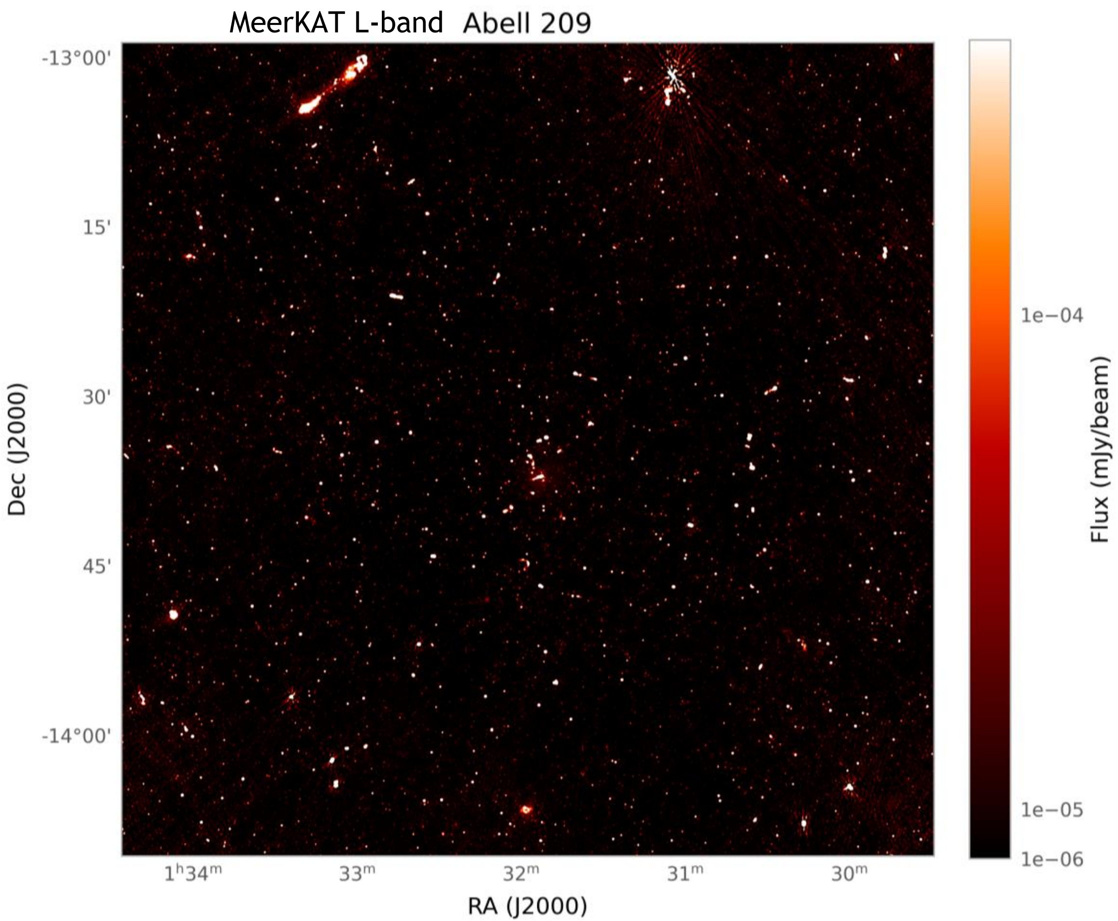


180”

Radio Galaxy Zoo (VLA)

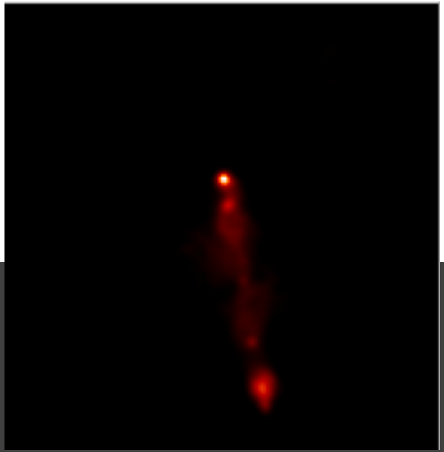


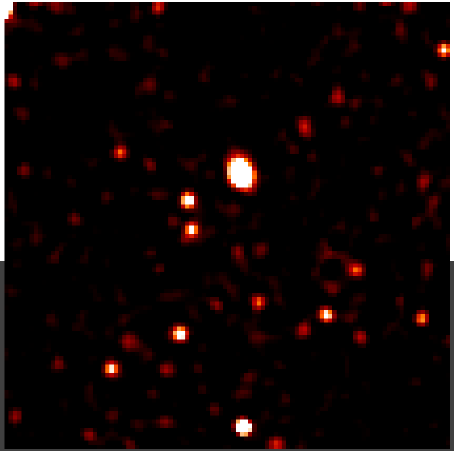
UNIVERSITÉ DE GENÈVE



MeerKAT L-band Abell 209
-13°00'
15'
30'
45'
-14°00'
Dec (J2000)
1h34m
33m
32m
31m
30m
RA (J2000)
1e−04
1e−05
1e−06
Flux (mJy/beam)

Can we train multi-purpose foundation models with pipeline data?





MGCLS
MeerKAT Galaxy Cluster Legacy Survey

- MeerKAT observations of 115 galactic clusters
- Wide-field 1.2 degree images
  - ~20,000 crops of pixel size 256 x 256
- Continuum CLEAN images

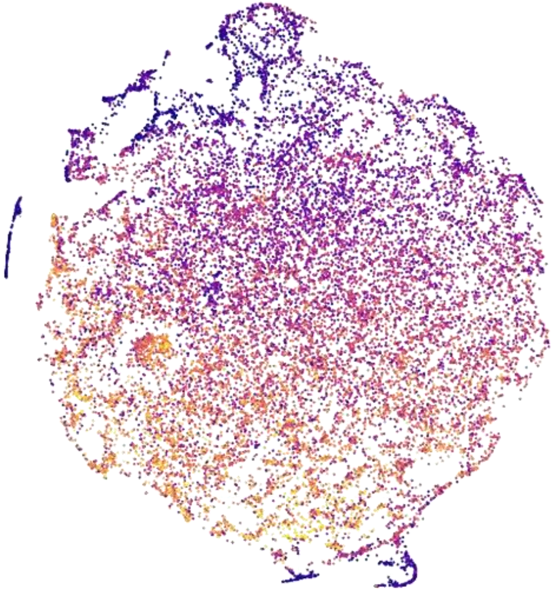
Knowledge distillation with no labels: DINO

Can SSL create representations that characterize source-rich MeerKAT data?

Train

Data (continuum CLEAN images)

SSL network (DINO)



Evaluate

Source catalogs

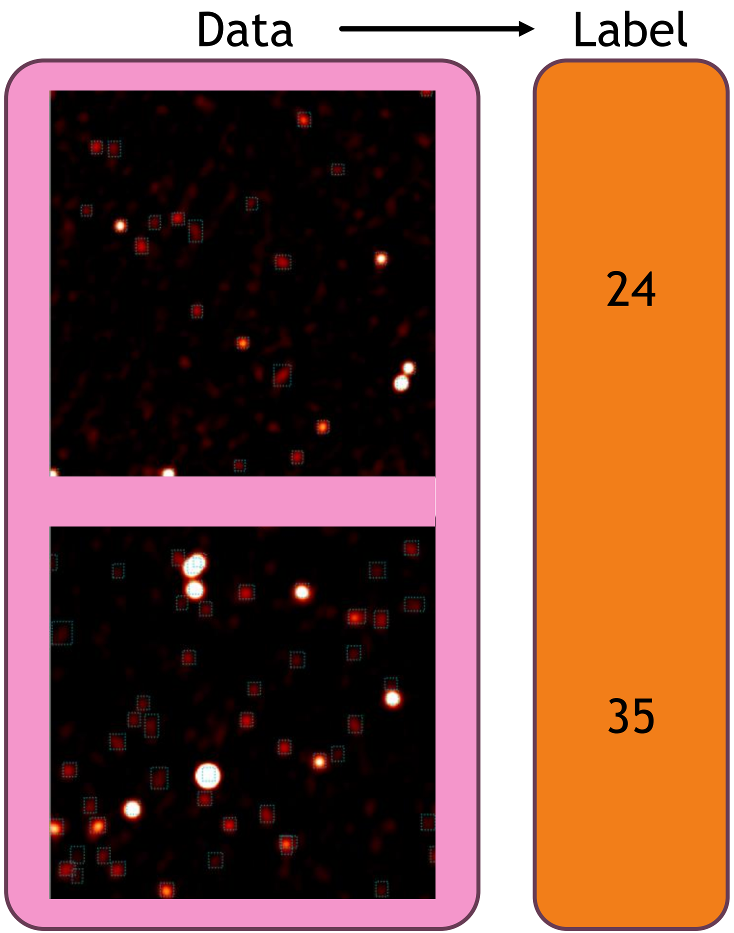
Compact source count

MIGHTEE and MiraBest hand-labelled

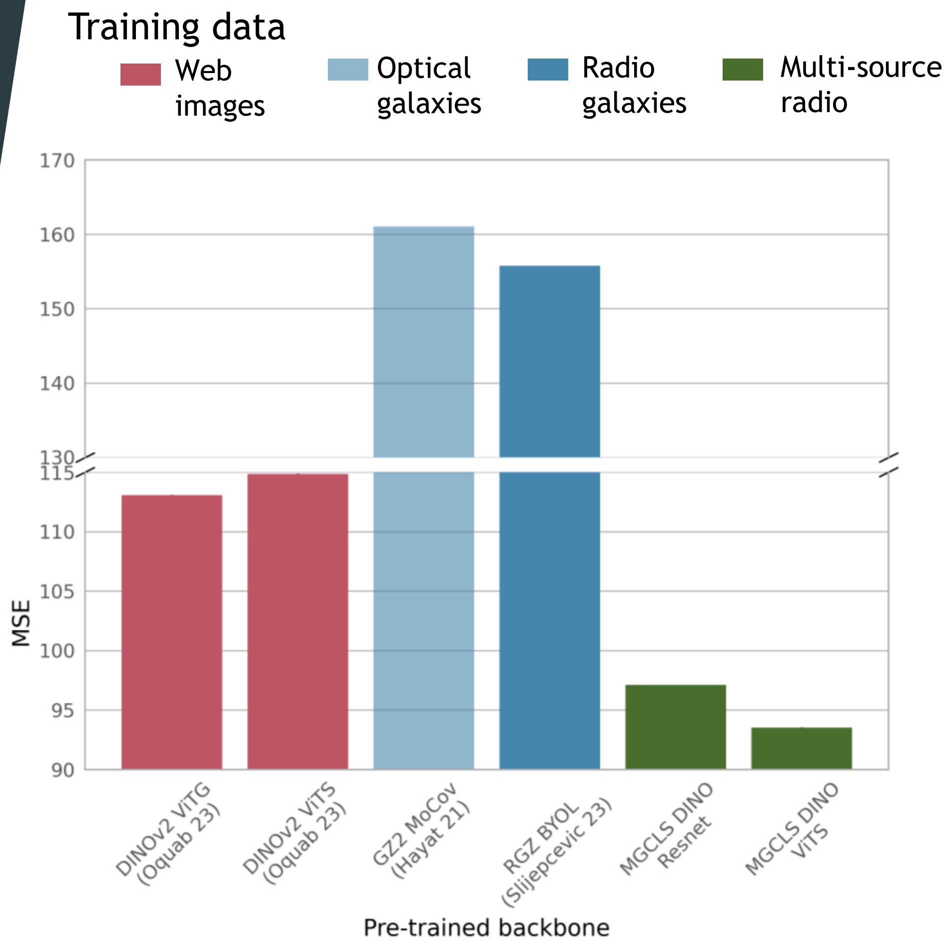
morphology classification

Evaluation: compact source count prediction

- Labels are from compact source catalog (pyBDSF)
- Task: predict source count with image as input
- Metric: mean squared error (MSE), the lower the better

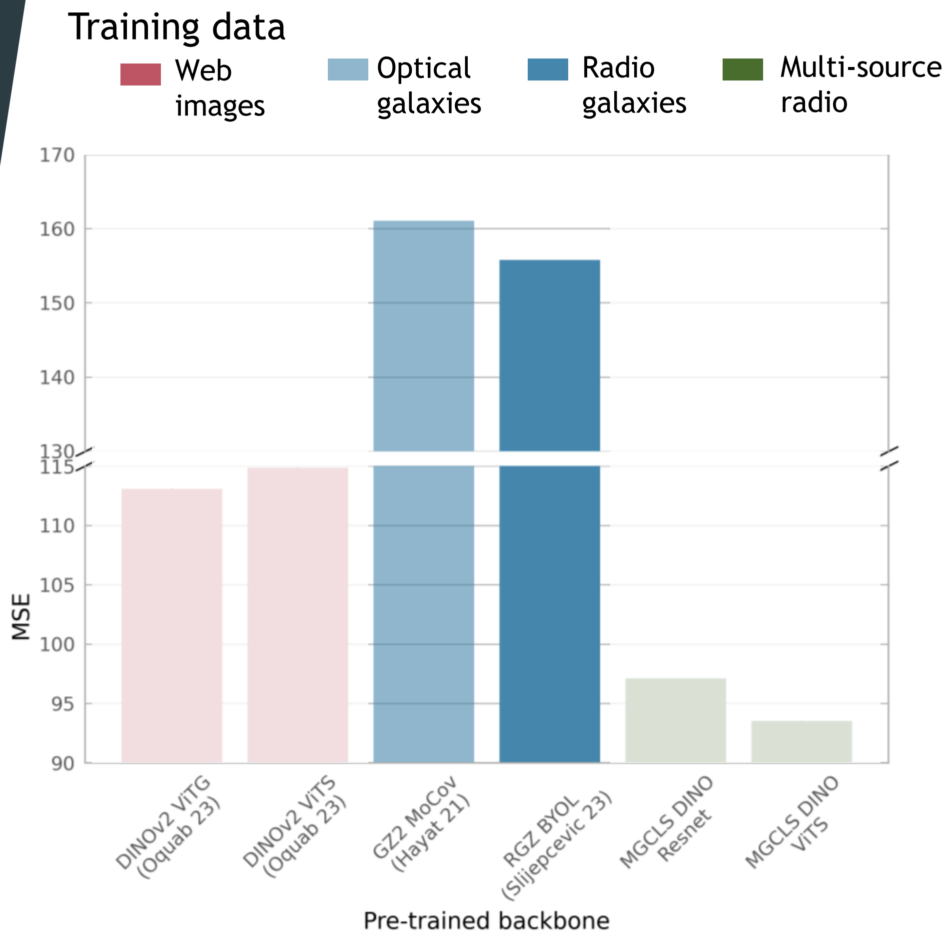


Evaluation: compact source count prediction



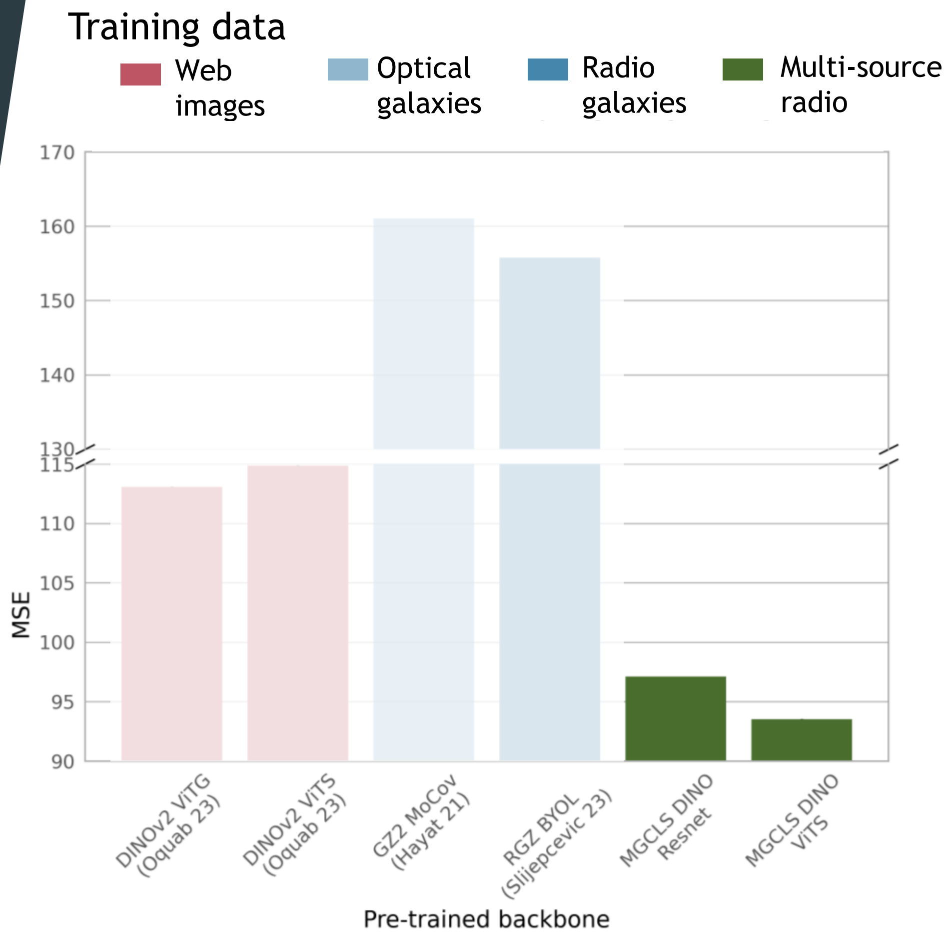
Evaluation: compact source count prediction

- Backbones trained on curated, single-galaxy images cannot perform this task!



Evaluation: compact source count prediction

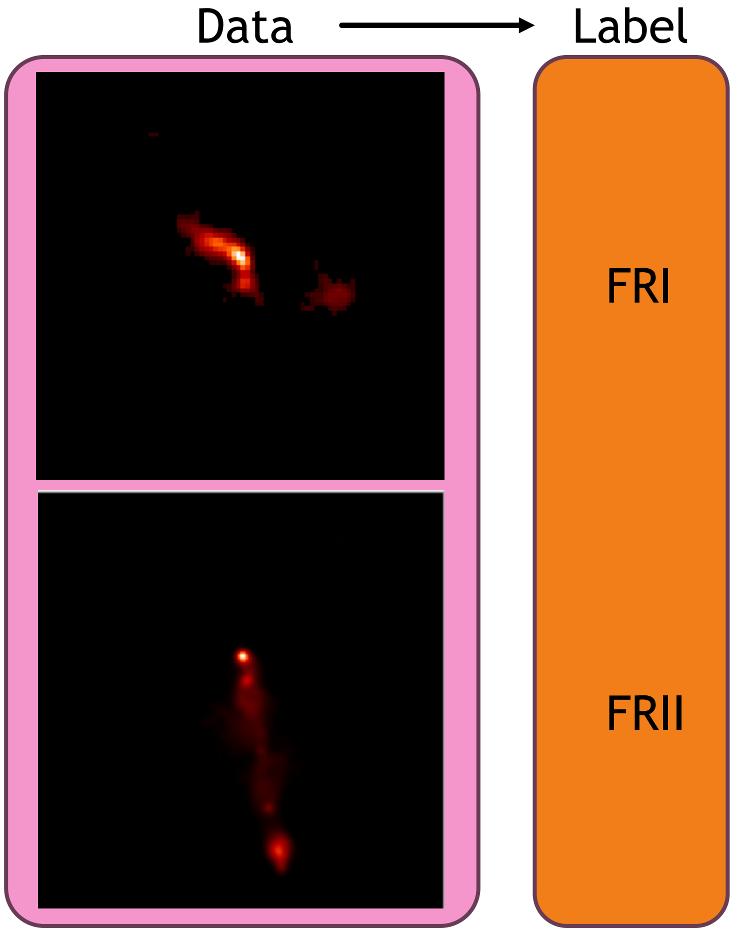
- Domain-specific training data is best for this task



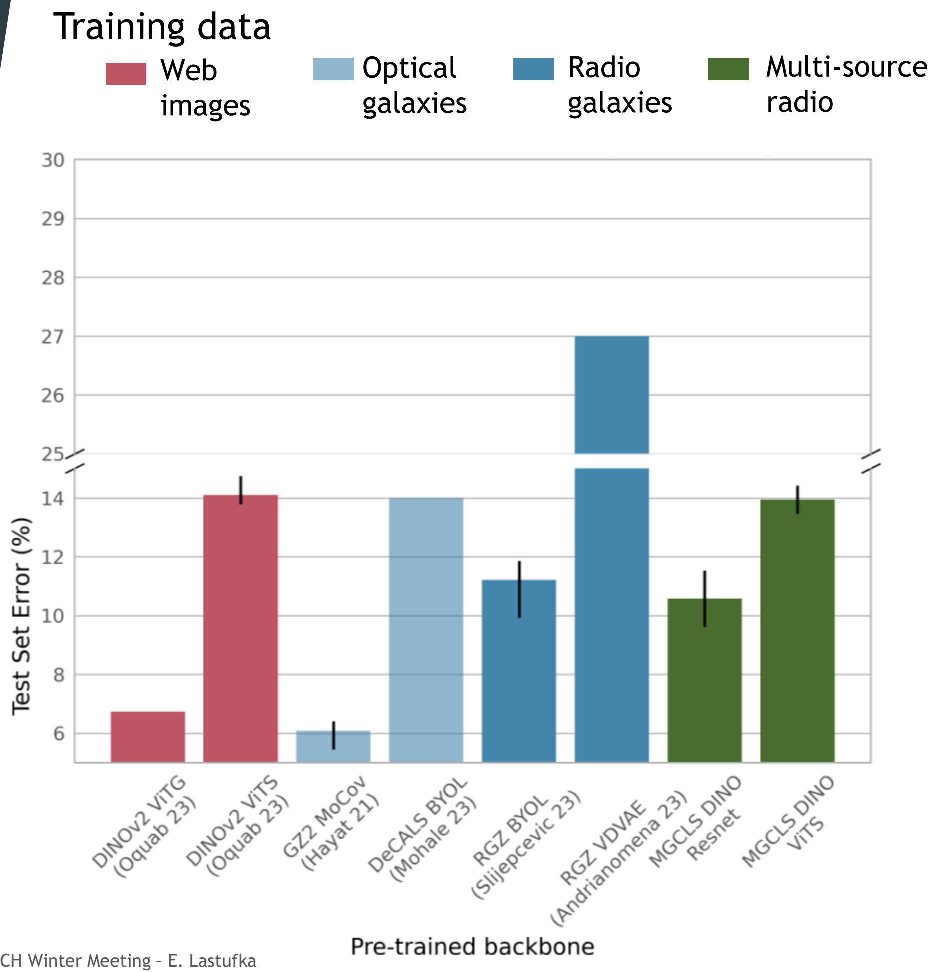
Can the network do more than one thing?

Evaluation: FRI/FRII galaxy morphology classification

- Public MiraBest dataset from VLA images
- Metric: Test set error (1-accuracy), lower is better

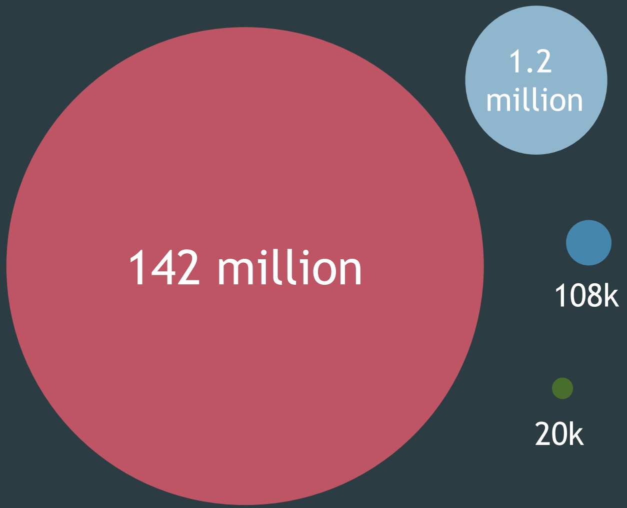


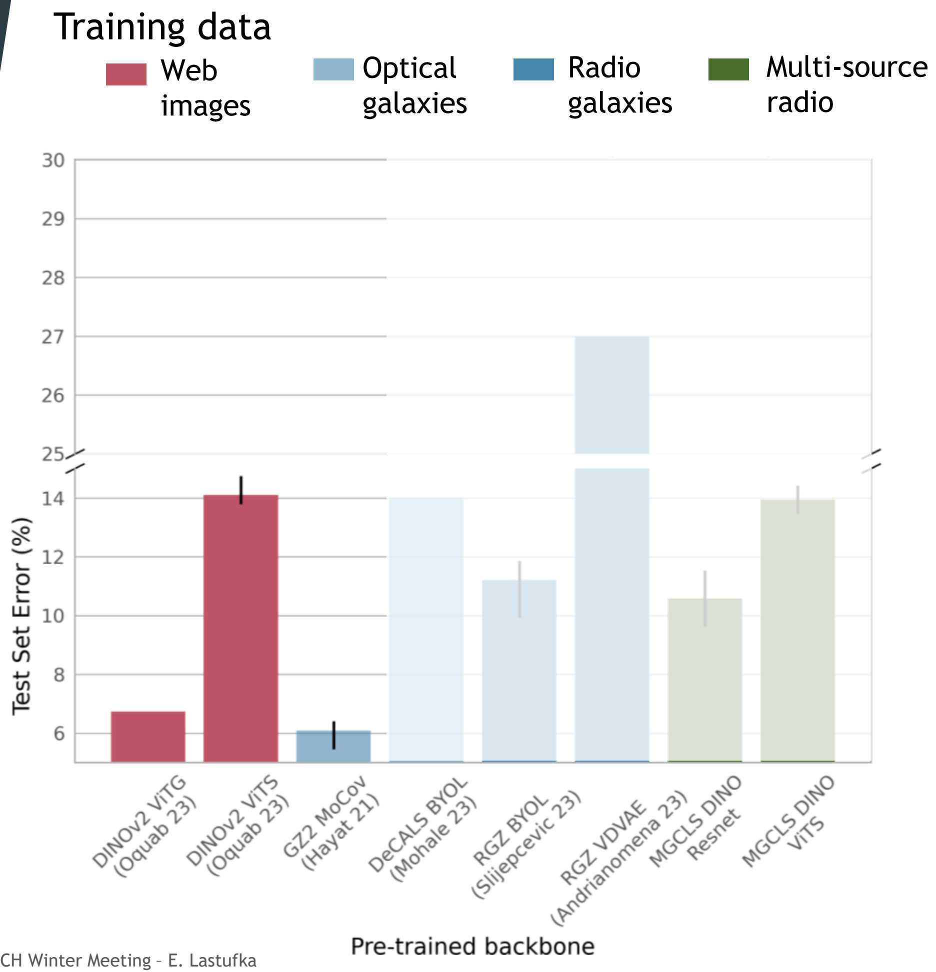
Evaluation: FRI/FRII galaxy morphology classification



Evaluation: FRI/FRII galaxy morphology classification

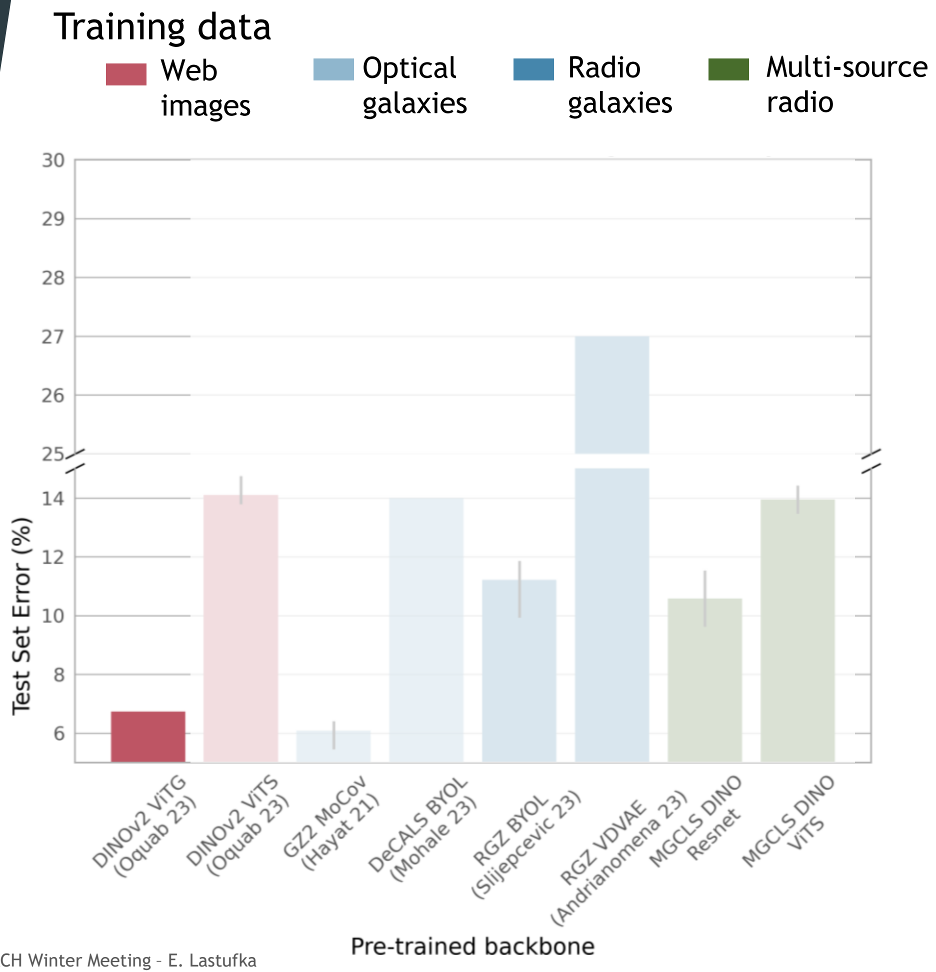
- Big training datasets gives good performance!





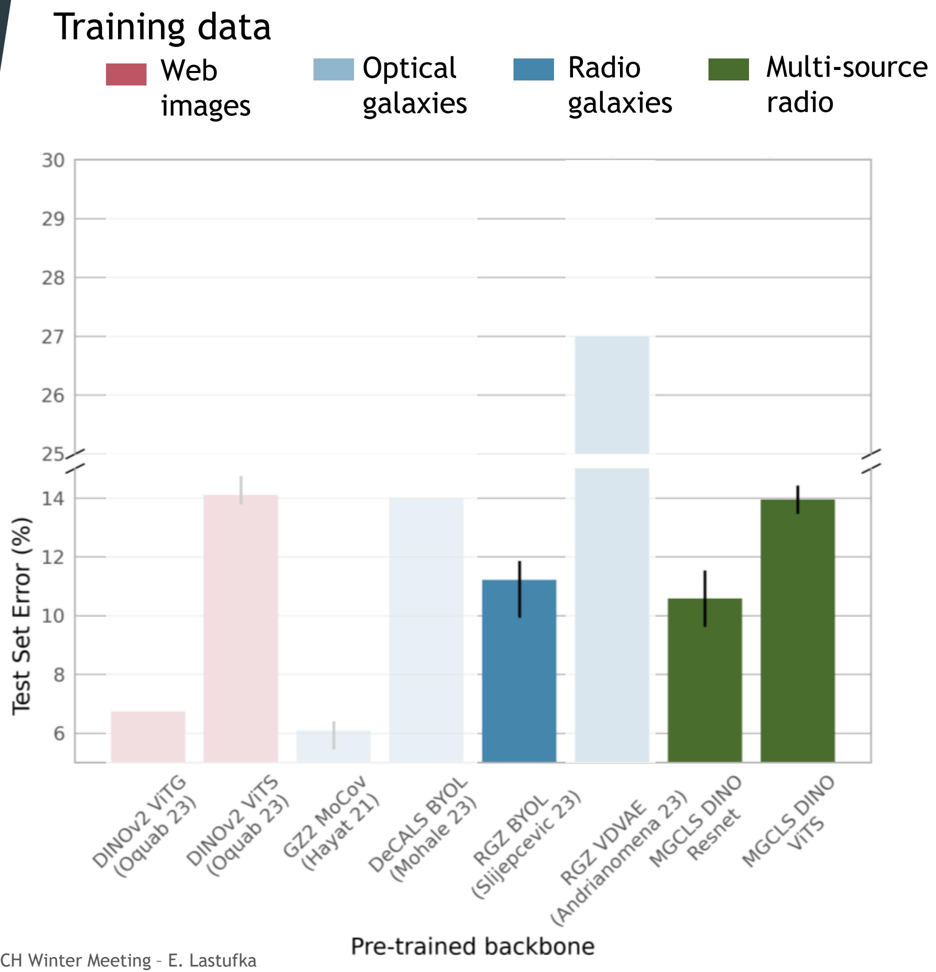
Evaluation: FRI/FRII galaxy morphology classification

- Big model makes up for lack of domain-specific training data?

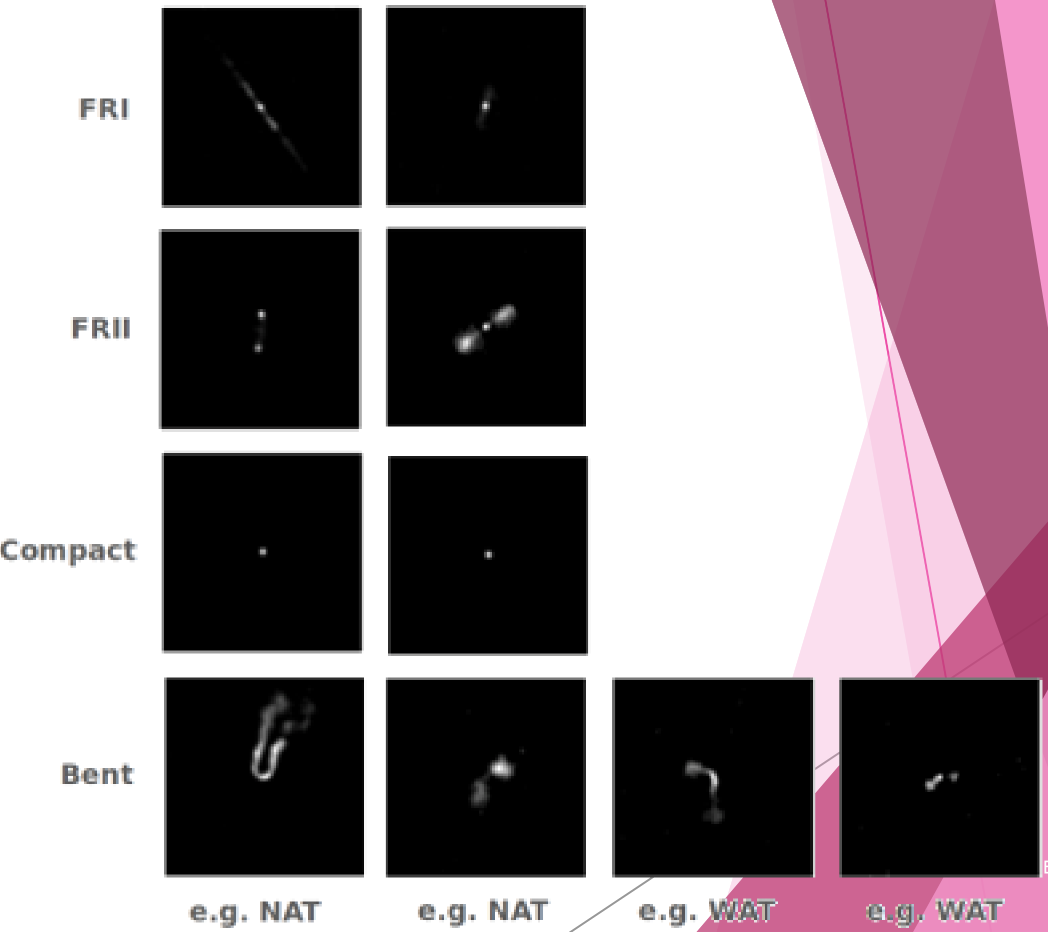


Evaluation: FRI/FRII galaxy morphology classification

- Our model trained on source-rich data can perform as well as academic state-of-the-art trained on highly curated data

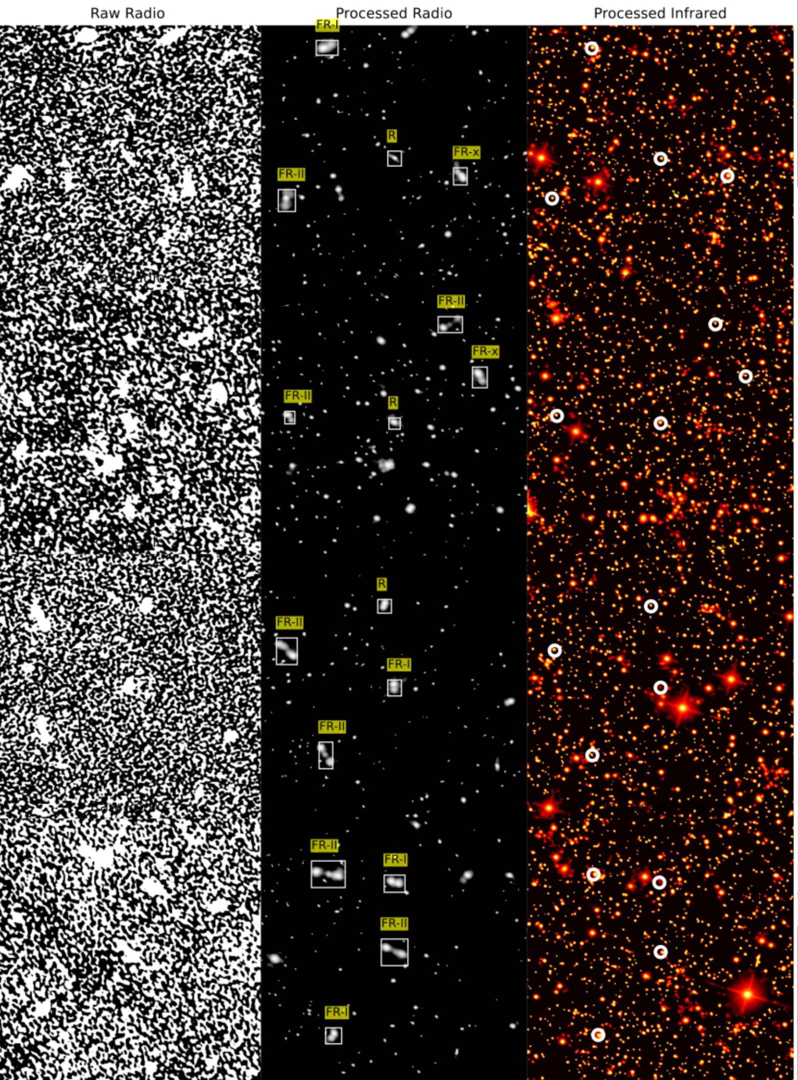


Can the network do complex things?



Multi-class morphology classification

- RadioGalaxyDataset
  - VLA images of four morphology classes



Source detection and segmentation

- RadioGalaxyNET
  - ASKAP + WISE radio/infrared dataset for object detection

Future foundation models in astronomy

Training data can come straight from the data processing pipeline!

A mixture of source-rich and single-source?

Evaluated on multiple high-complexity tasks

Extra slides

Evaluation: multi-class galaxy morphology classification