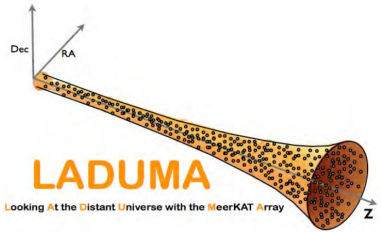
Automatic source finding in SKA precursor/pathfinder surveys with YOLO-CIANNA

Adrien ANTHORE[1], L. Chemin[1], D. Cornu[2]

[1] Observatoire Astronomique de Strasbourg
[2] LUX, Observatoire de Paris

CIANNA user workshop, Paris 2025

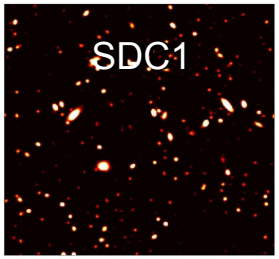


From simulation to observation

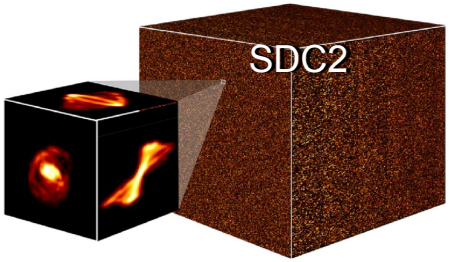
2D continuum emission

3D HI emission cube

Simulated



best score *a posteriori*
Cornu+24

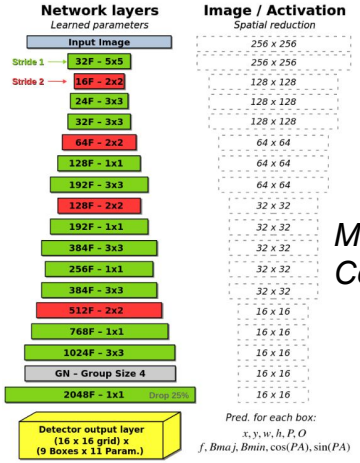


1st place in SDC2
Hartley+23, Cornu+25

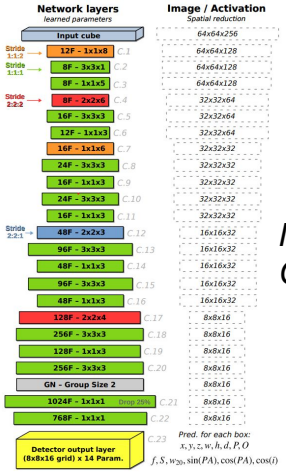
Analogue observations

Objective: deploy YOLO-CIANNA and adapt it to the observations

From simulation to observation



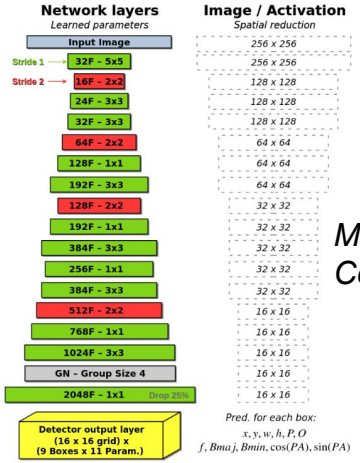
Model used for SDC1
Cornu+24



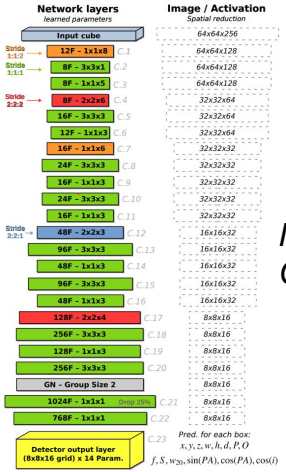
Model used for SDC2
Cornu+25

Objective: deploy YOLO-CIANNA and adapt it to the observations

From simulation to observation



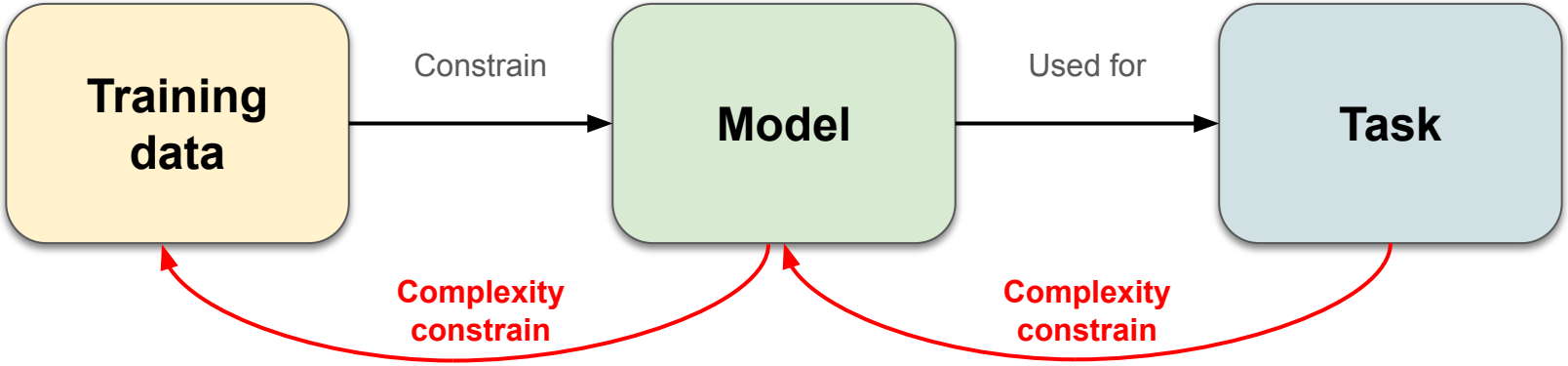
Model used for SDC1
Cornu+24



Model used for SDC2
Cornu+25

Objective: deploy YOLO-CIANNA and adapt it to the observations

Training dataset for supervised method



This training dataset must:

- Be **complete**: all specificities of the data must be represented (objects, effects, contexts, …)
- Have **pure labels**: labels should be as close as possible to the expectancy.

Not/wrongly labeled target:

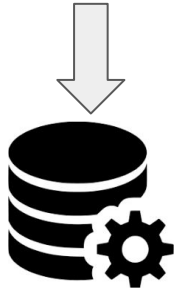
→If the network detects it, will **lower the probability of all the same kind of sources**

Labeled questionable target:

→If the network detects it, may **increase the probability of detecting noise**

How to train the model for observational data?

Observational data



Training dataset

Training

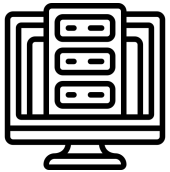
1st option: **using observational data**

Pros:
- Contain all instrumental effects and observational limitations

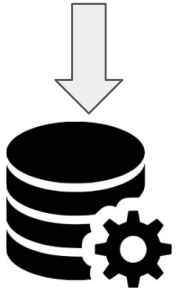
Cons:
- Limitation in examples
- Scarcity effects
- Difficult to label data

Because it requires a lot of exemple: labeling a **large portion of the survey makes ML useless**

How to train the model for observational data?



Simulated data



Training dataset

Training

2nd option: **using available simulated data**

Pros:

- As many examples as necessary
- Compensating for scarcity effects

Cons:

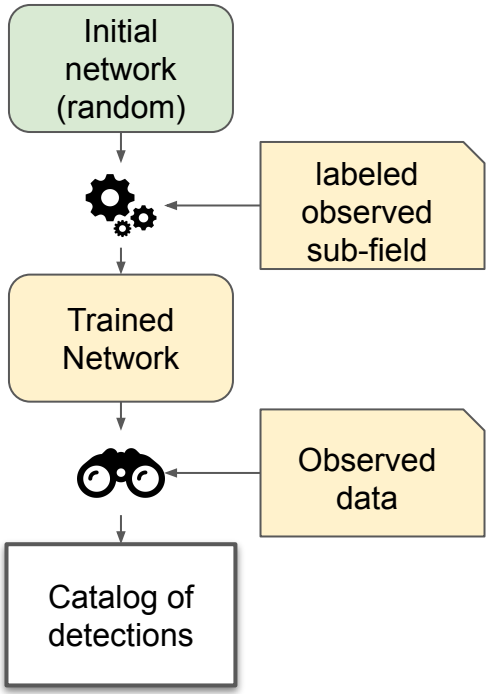
- Potentially biased or simplistic:

Physical model	Instrumental Model

In practice the two approaches can be combined

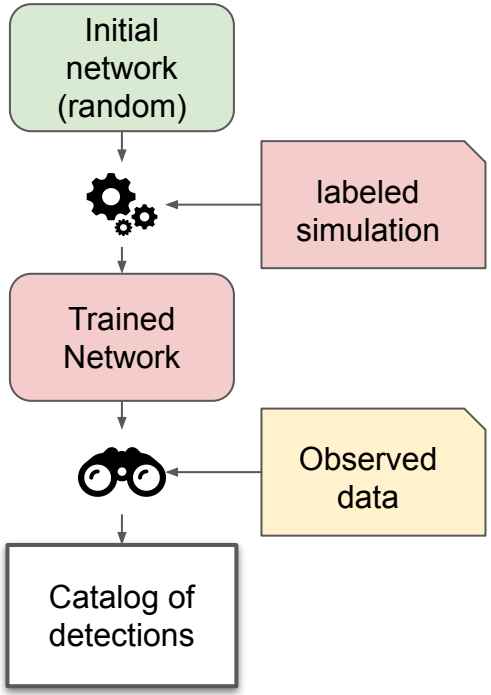
Approaches

Observed data only

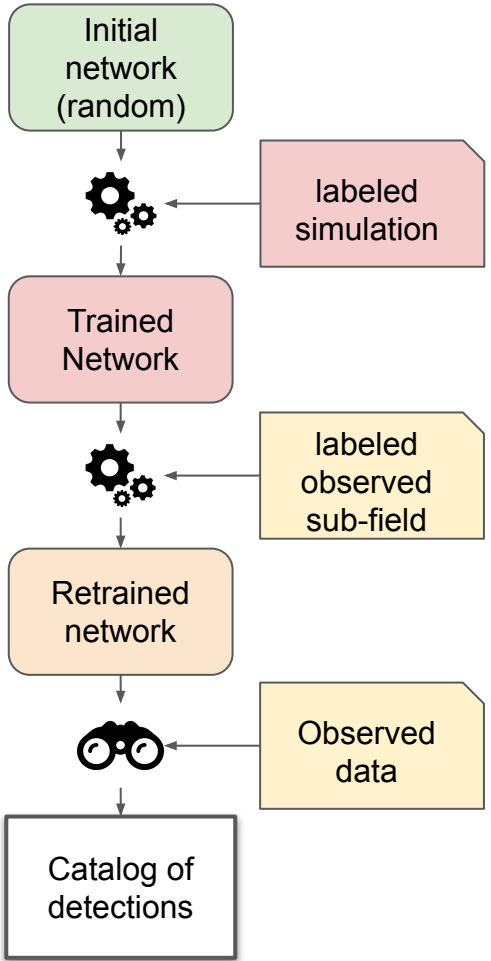


Training *Inference*

Direct application

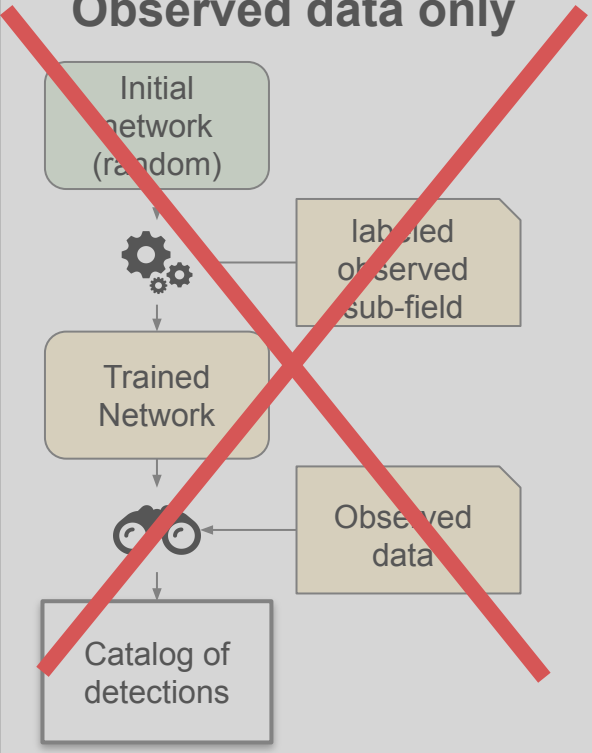


Transfer-Learning



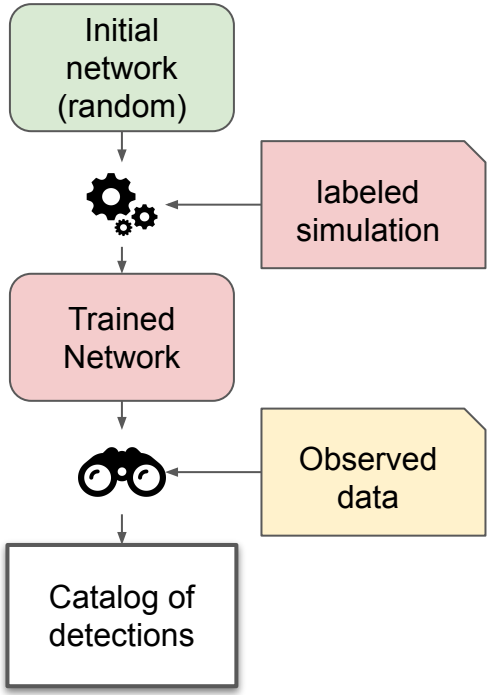
Approaches

Observed data only

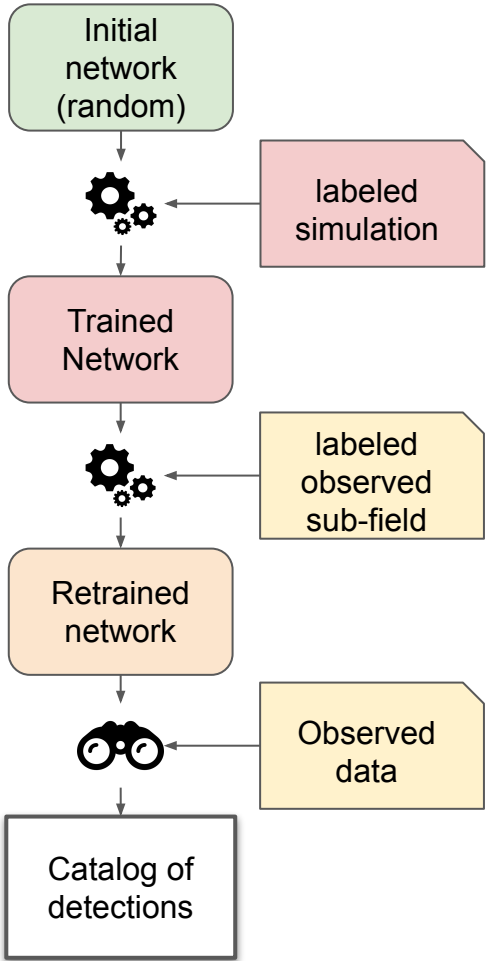


Training Inference

Direct application



Transfer-Learning

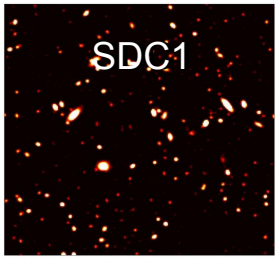


From simulation to observation

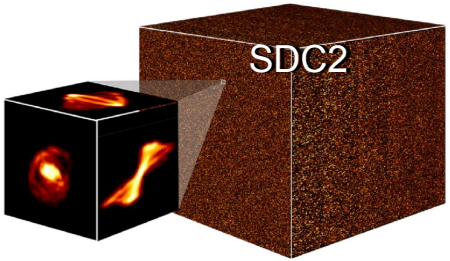
2D continuum emission

3D HI emission cube

Simulated



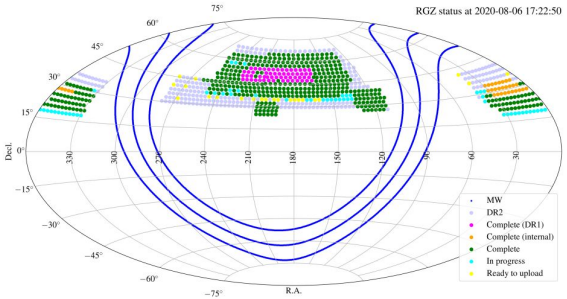
best score *a posteriori*
Cornu+24



1st place in SDC2
Hartley+23, Cornu+25

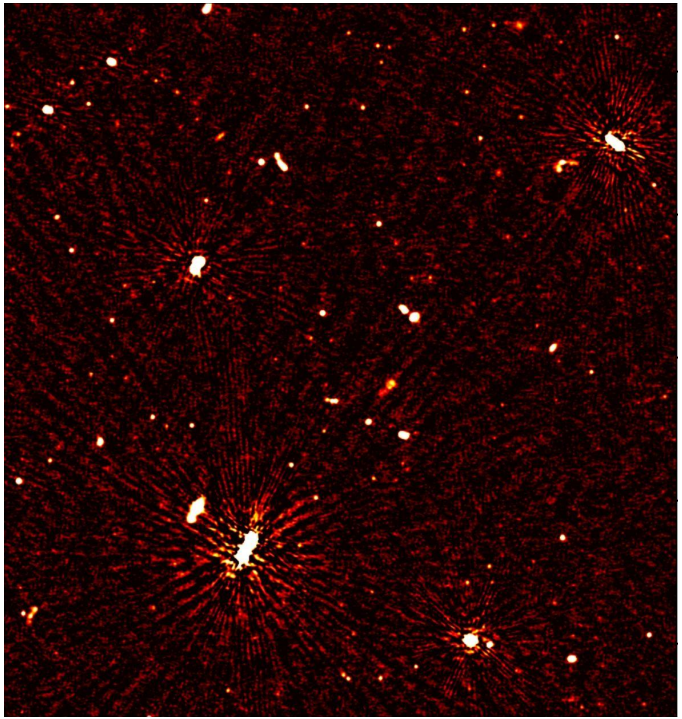
Analogue observations

LOw Frequency ARray (LOFAR)



LoTSS DR2 coverage

subfield of 25x25 arcmim²

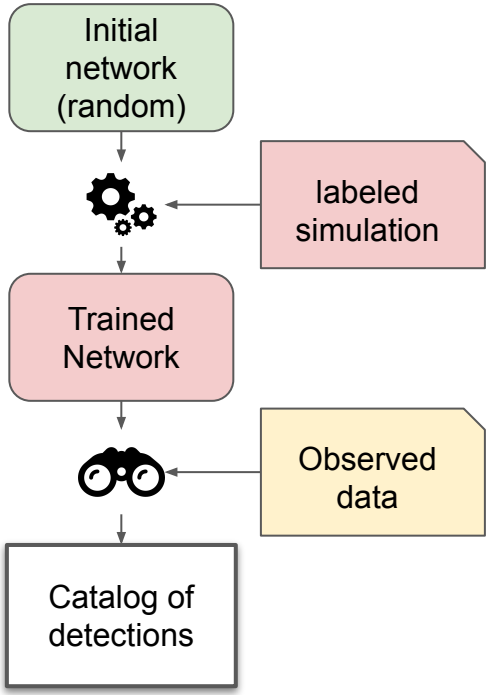


LOFAR Two meter Sky Survey (LoTSS) DR2:
Shimwell et al. 2022

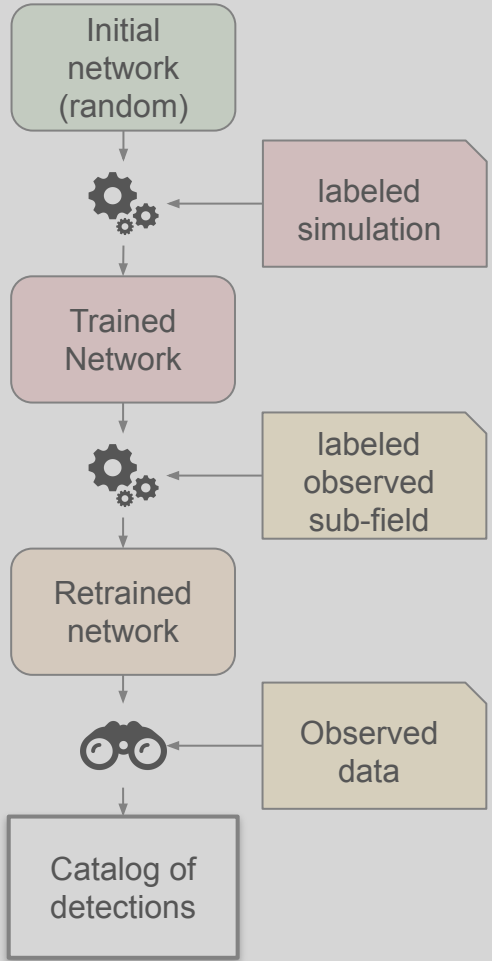
- Frequency: 120-168 MHz
- 27% of the Northern hemisphere
- **4,396,228 cataloged sources** from **PyBDSF**
- 841 mosaics

Approaches

Direct application



Transfer-Learning



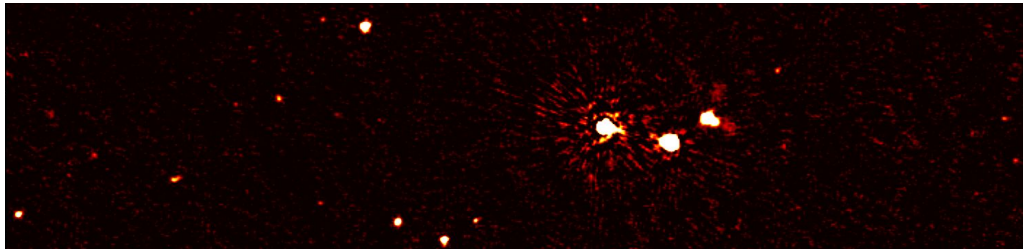
Inference

Training

Application of the method to LoTSS data

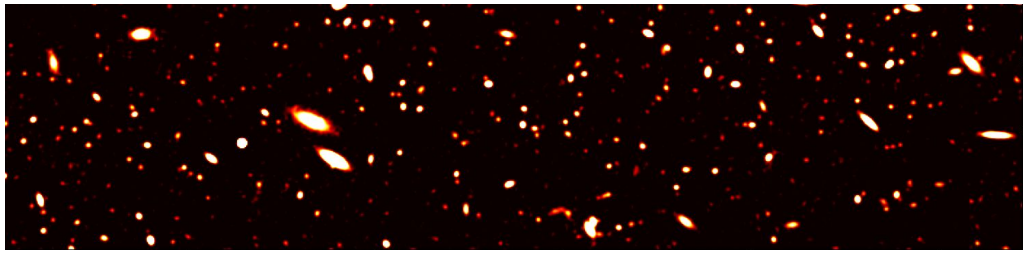
The inference data must match the training data: is it the case?

Similar?



LoTSS DR2
(Observational)
144 MHz

subfields of 150x700 pix



Simulated data
(SKAO SDC1)
560 MHz

Similarities:

- Same point-like sources
- Luminosity profiles
- Blending

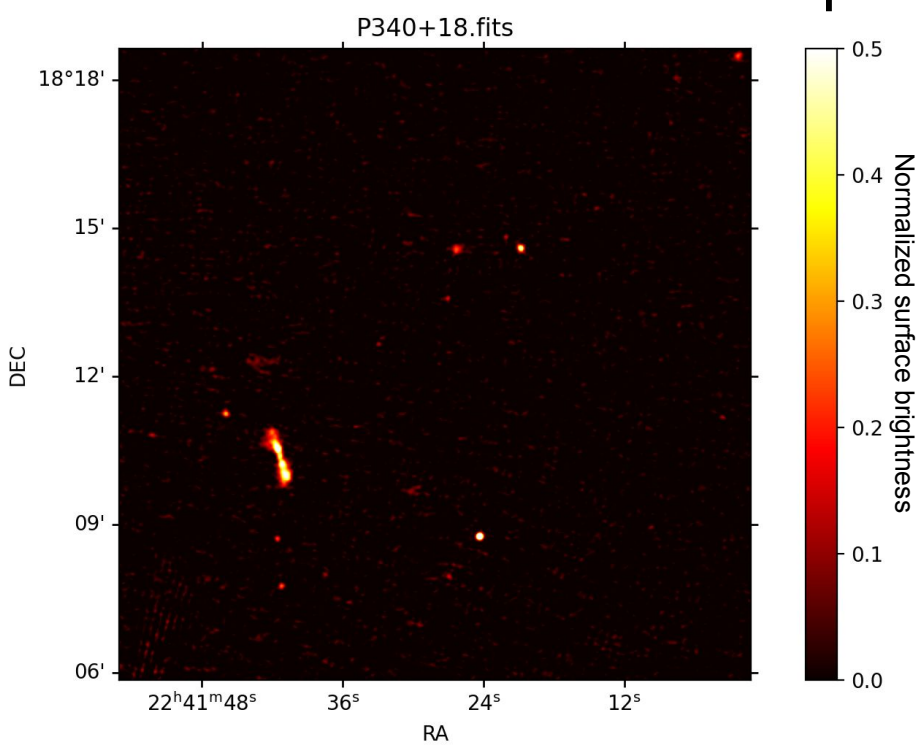
Dissimilarity:

- Resolution
- Pixels dynamics/Sensitivity
- Morphological diversity
- Instrumental specificities

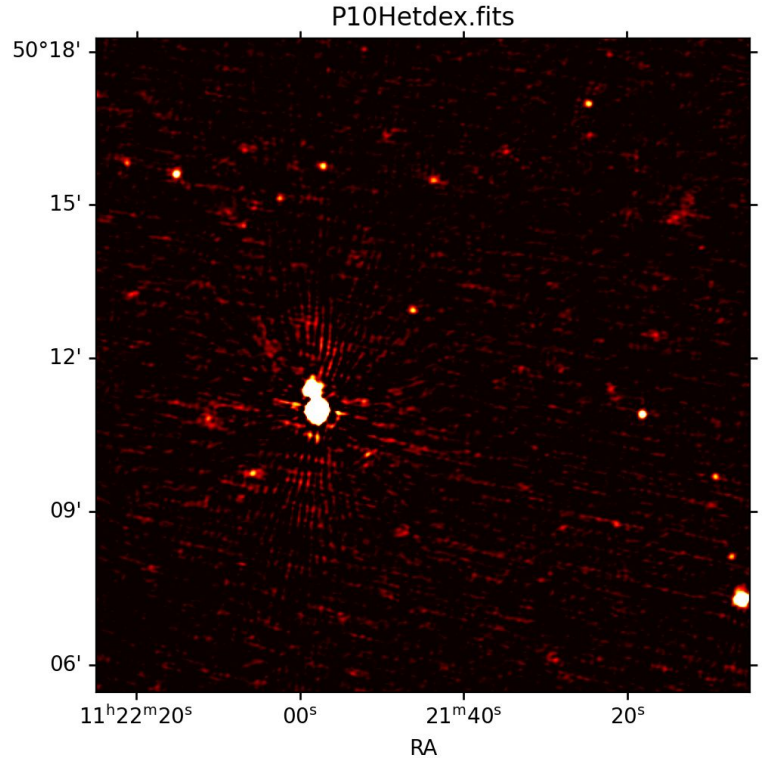
Similar enough, **but requiere**:

- Match **the pixel dynamics**
- Match **the sampling**

Exemple fields



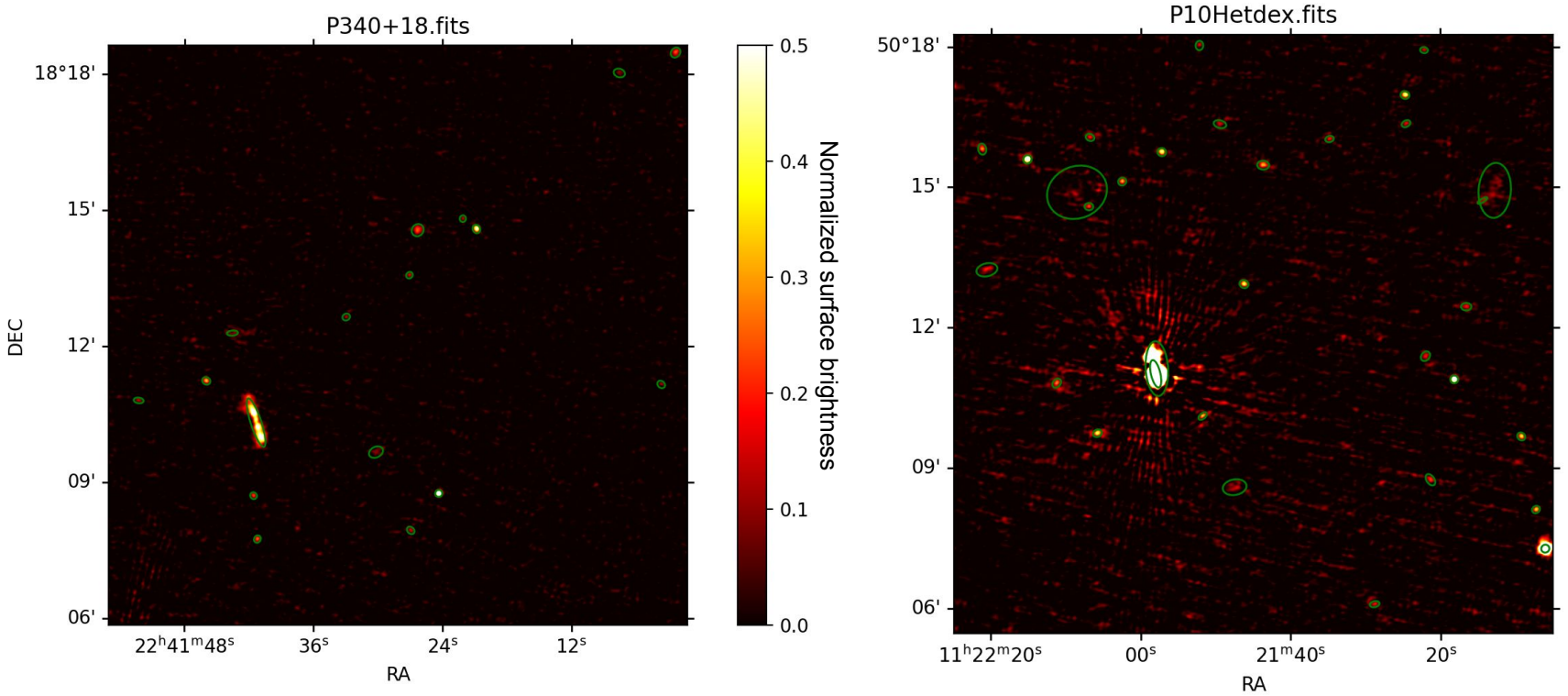
Field **without artifact**



Field **with artifacts**

Object of interest: Point-like sources; Extended sources; Artifacts around bright sources; Blending; Other artifacts

Evaluation on reference



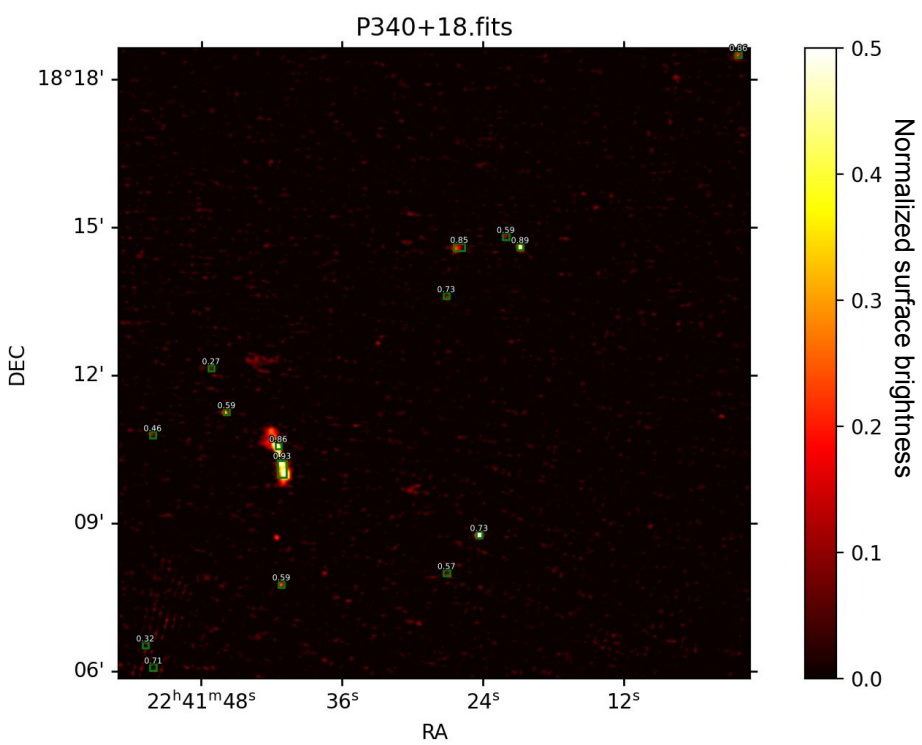
Reference: LoTSS DR2 (Shimwell et al. 2022)

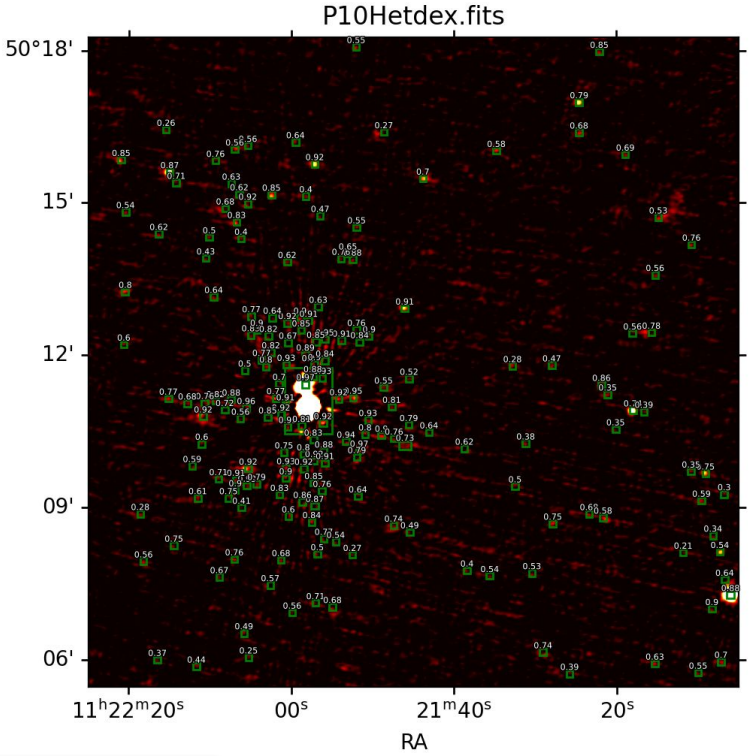
Methode: PyBDSF

Recall = *N_match* / *N_ref*

Precision = *N_match* / *N_test*

Results: “Direct” approach





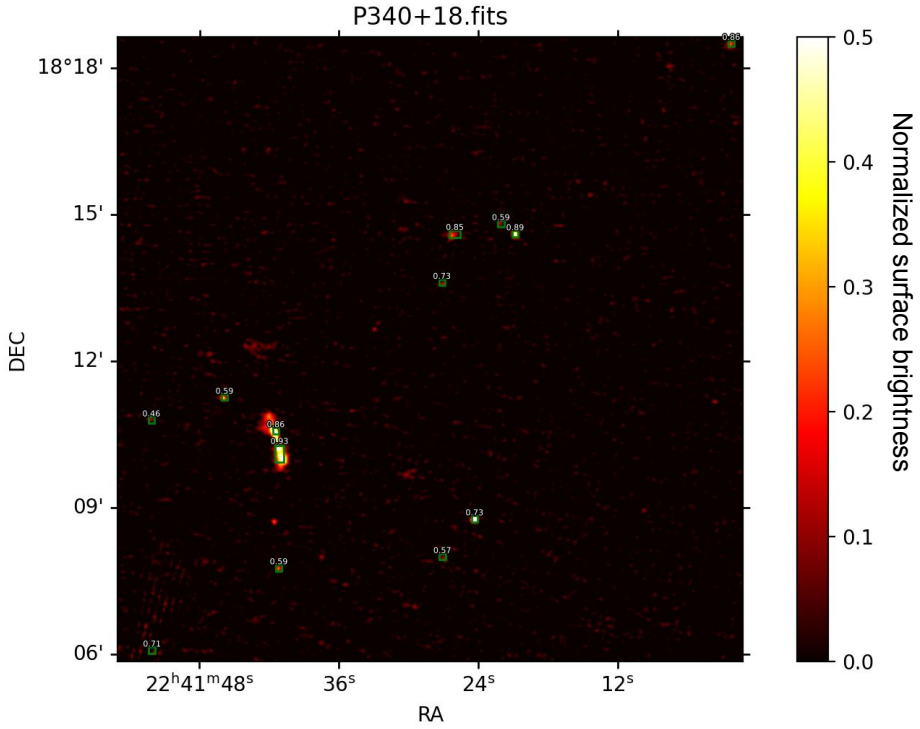
~50% Recall ; ~20% Precision

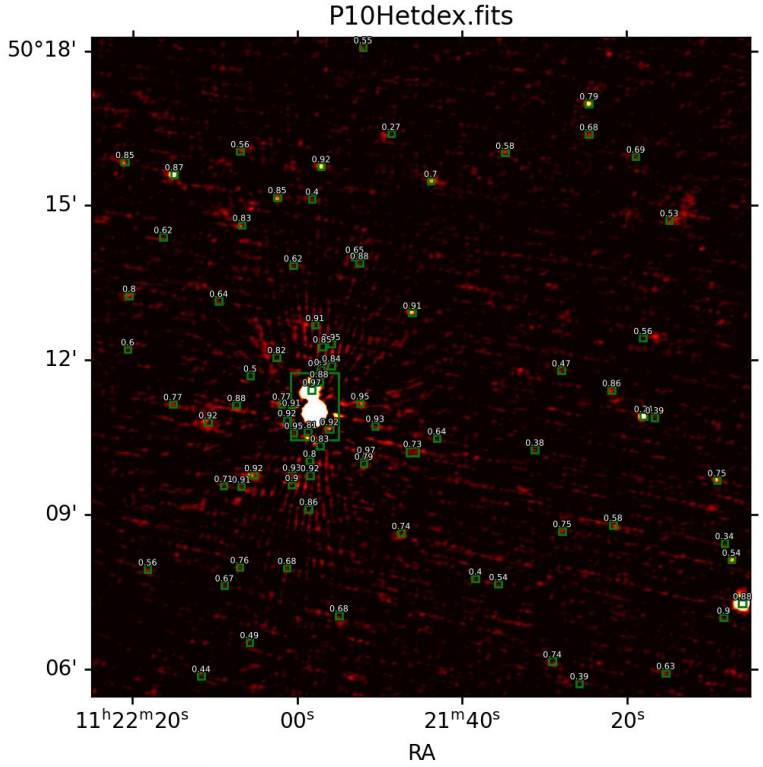
Too many false detections

XX

Detection and associated probability

Results: “Direct” approach + post-process





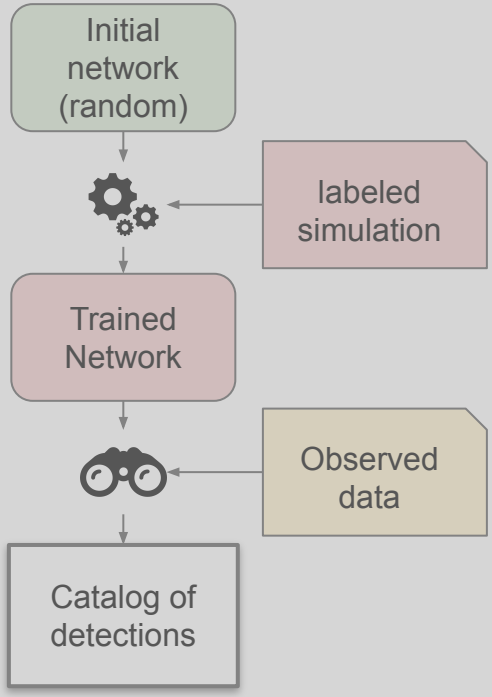
~45% Recall ; ~30% Precision

Less false detection
We reached the limit of improvement

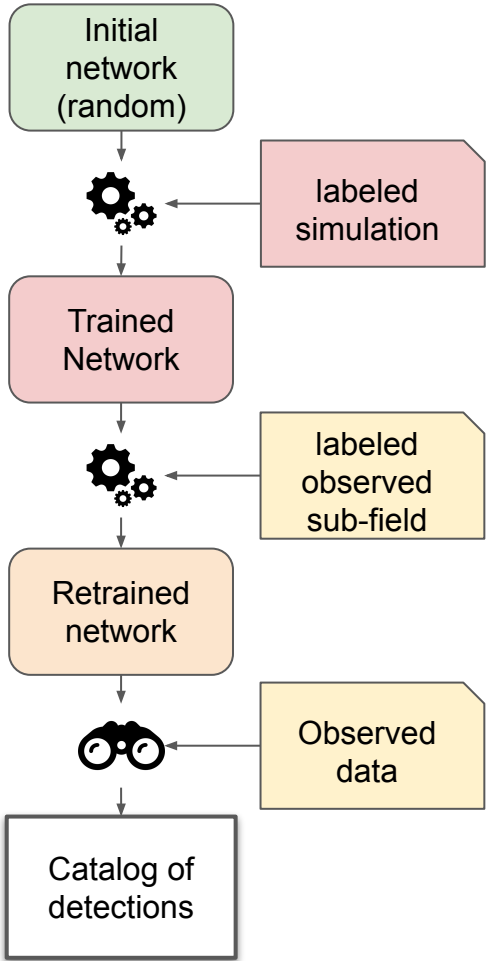
XX
Detection and associated probability

Approaches

Direct application



Transfer-Learning



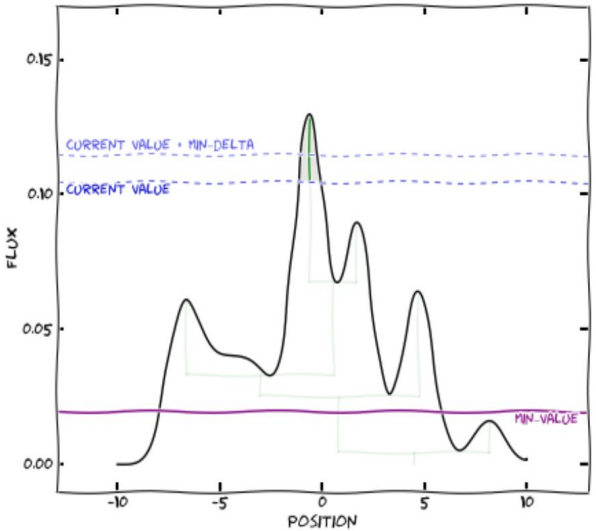
Inference

Training

Method for building the training set

Catalog optimized for ML detection ≠ existing catalogs today

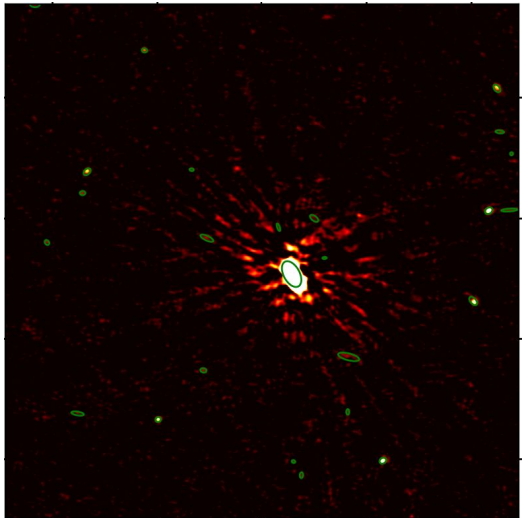
Choice of method: **Astrodendro**



Credit: https://dendrograms.readthedocs.io

Labeling process

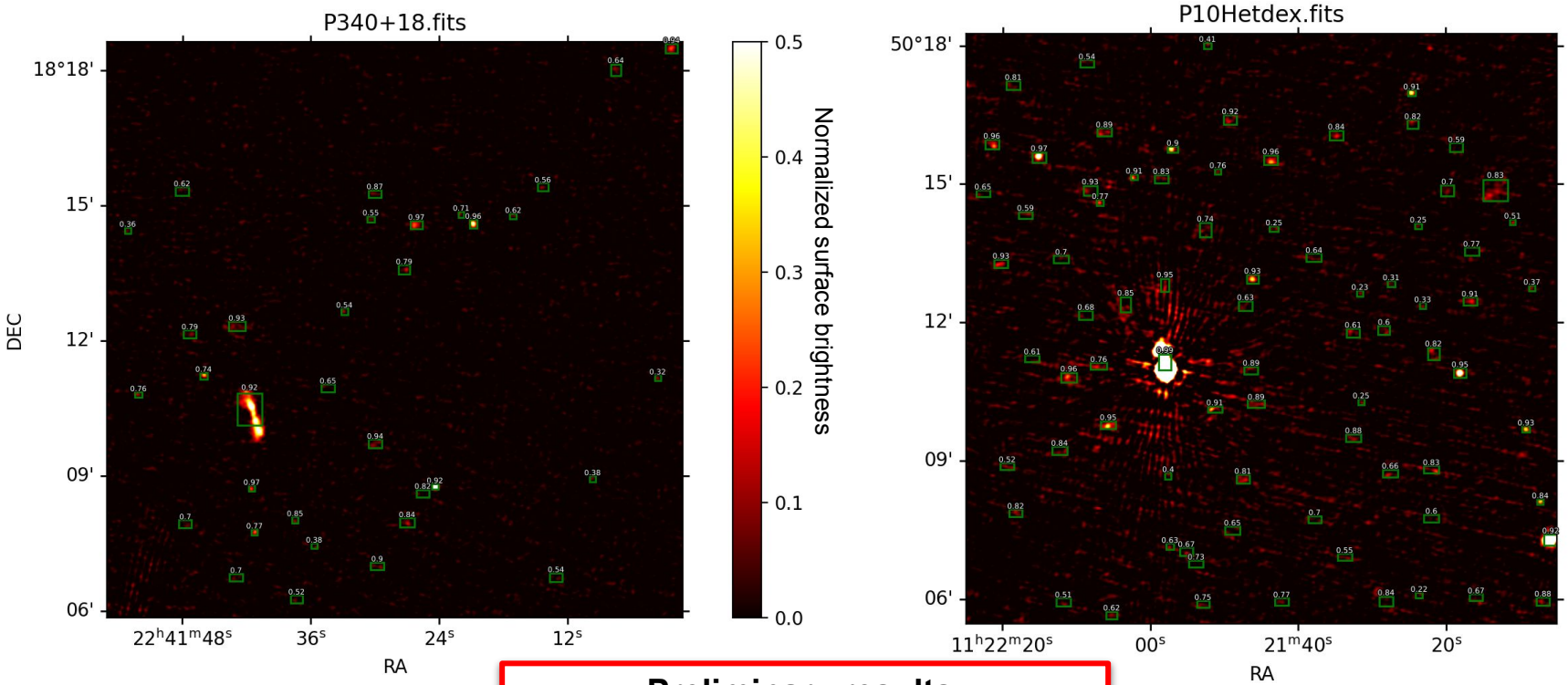
1) Detection in a subfield
2) Filtering after detection
3) IR / Optical counterparts



example of the training area
13x13 arcmin²

By combining classical detection method with counterparts:
We manage to construct a reliable training dataset

Preliminary results: “Transfer-Learning” approach



~90% Recall ; ~50% Precision

Preliminary results, possible enhancement:

- Improving training dataset
- Post-processing

XX

Detection and associated probability

LoTSS DR2 summary

Our **YOLO-CIANNA model** with LoTSS DR2 data**:**

- **Quick:** Training: not more than 8h;
  Inference ~ 5 sec/mosaics
  Inference and Training on Tesla V100 GPU
- **Complete:** High level of recall (w.r.t. LoTSS dr2)
  Best: ~90% Recall ; ~50% Precision

Still requires more exploration of the data, confirmations, new approach?, and also **technical exploration** (Network architectures, simulations, …) and more

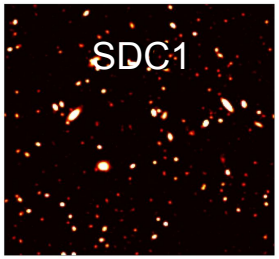
Adam ZARKA's PhD

From simulation to observation

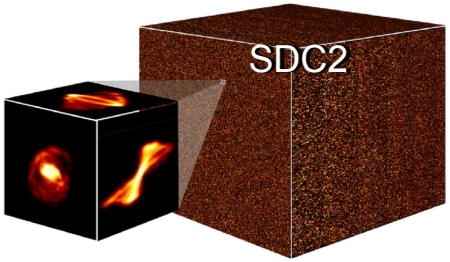
2D continuum emission

3D HI emission cube

Simulated



best score *a posteriori*
Cornu+24

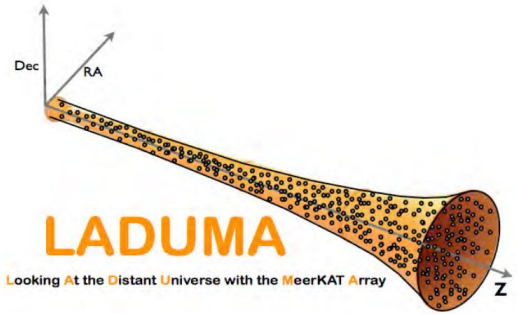


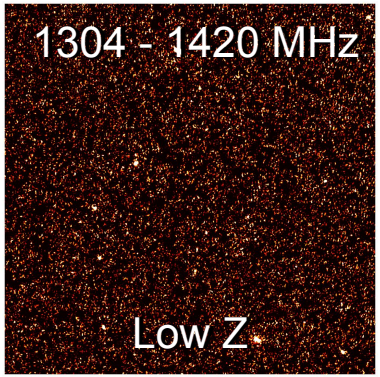
1st place in SDC2
Hartley+23, Cornu+25

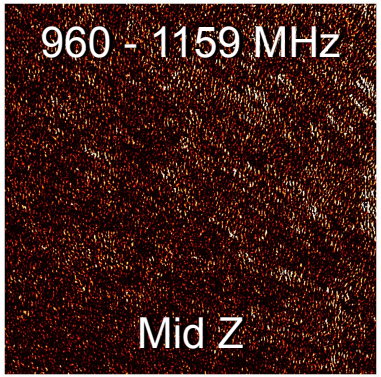
Analogue observations

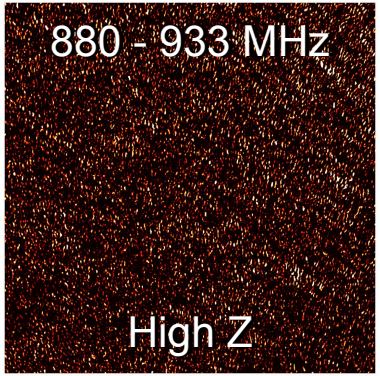
LADUMA

Looking At the Distant Universe with the MeerKAT Array







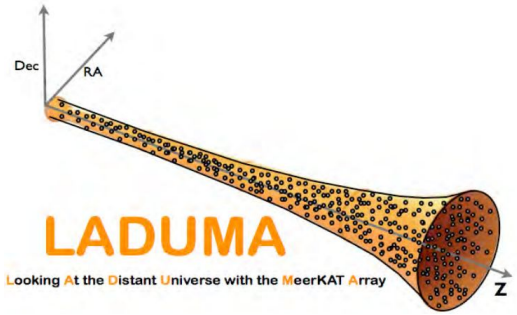


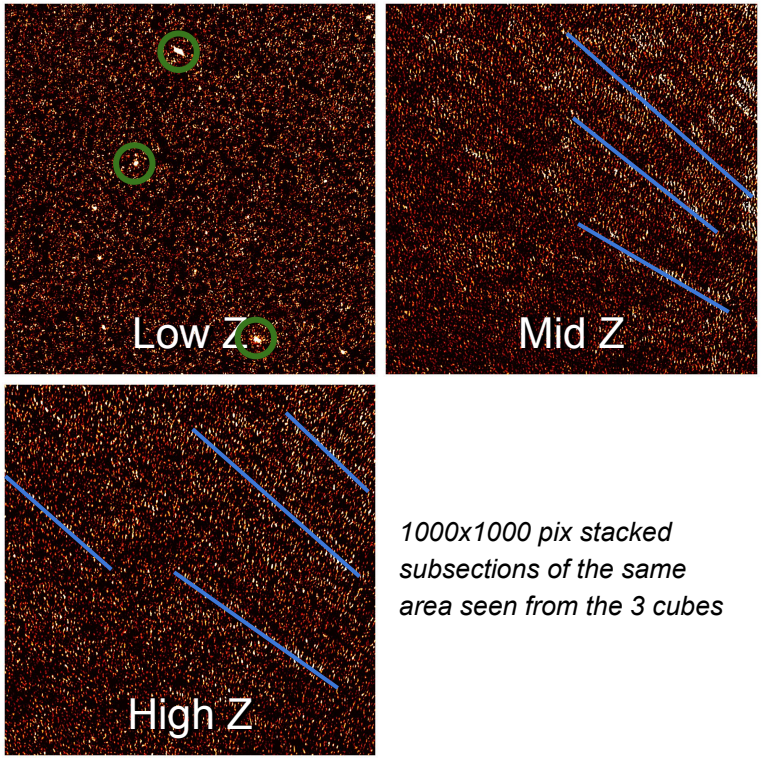
- 3 hyperspectral cubes (~310 GB)
- Frequencies from 880 to 1420 MHz
- ~ 1 sq. deg. coverage
- 243 sources detected by SoFiA (lowest redshift cube)

1000x1000 pix stacked subsections of the same area seen from the 3 cubes

LADUMA

Looking At the Distant Universe with the MeerKAT Array



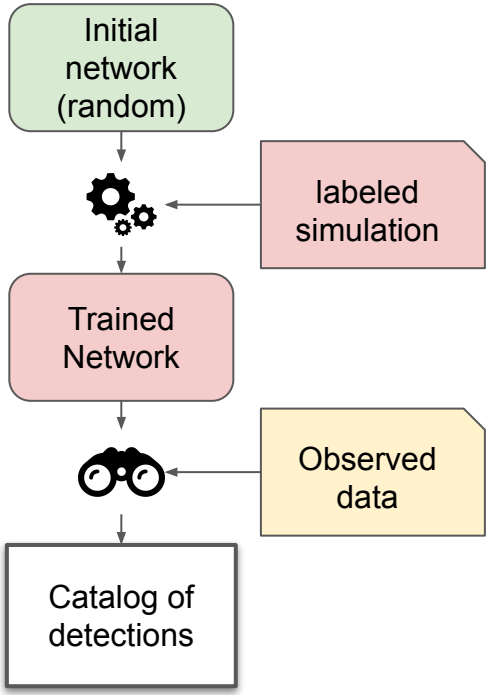


- 3 hyperspectral cubes (~310 GB)
- Frequencies from 880 to 1420 MHz
- ~ 1 sq. deg. coverage
- 243 sources detected by SoFiA (lowest redshift cube)

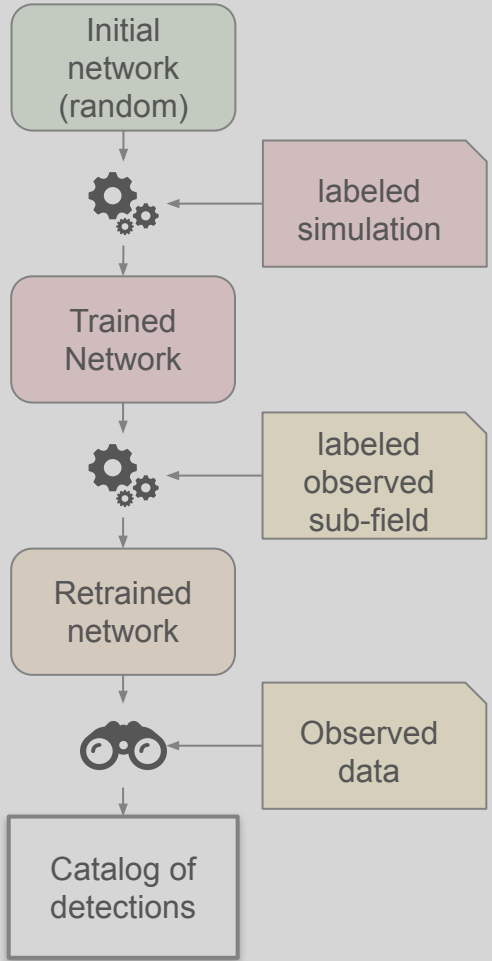
1000x1000 pix stacked subsections of the same area seen from the 3 cubes

Approaches

Direct application



Transfer-Learning



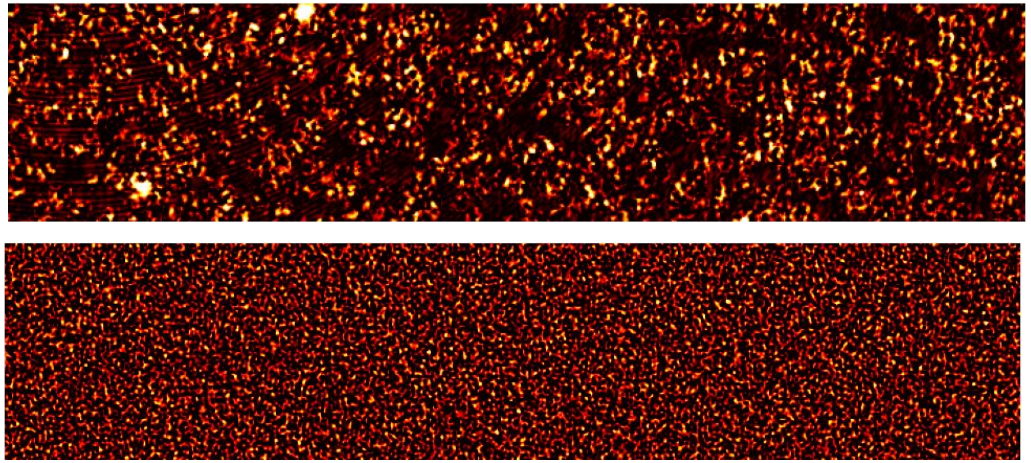
Inference

Training

Direct application to LADUMA data

The inference data must match the training data: is it the case?

Similar?



LADUMA data
low z cube
(Observational)

subfields of 150x700 pix

Idealized Simulated data
(SKAO SDC2)

Similarities:

- Same point-like sources
- Luminosity profiles

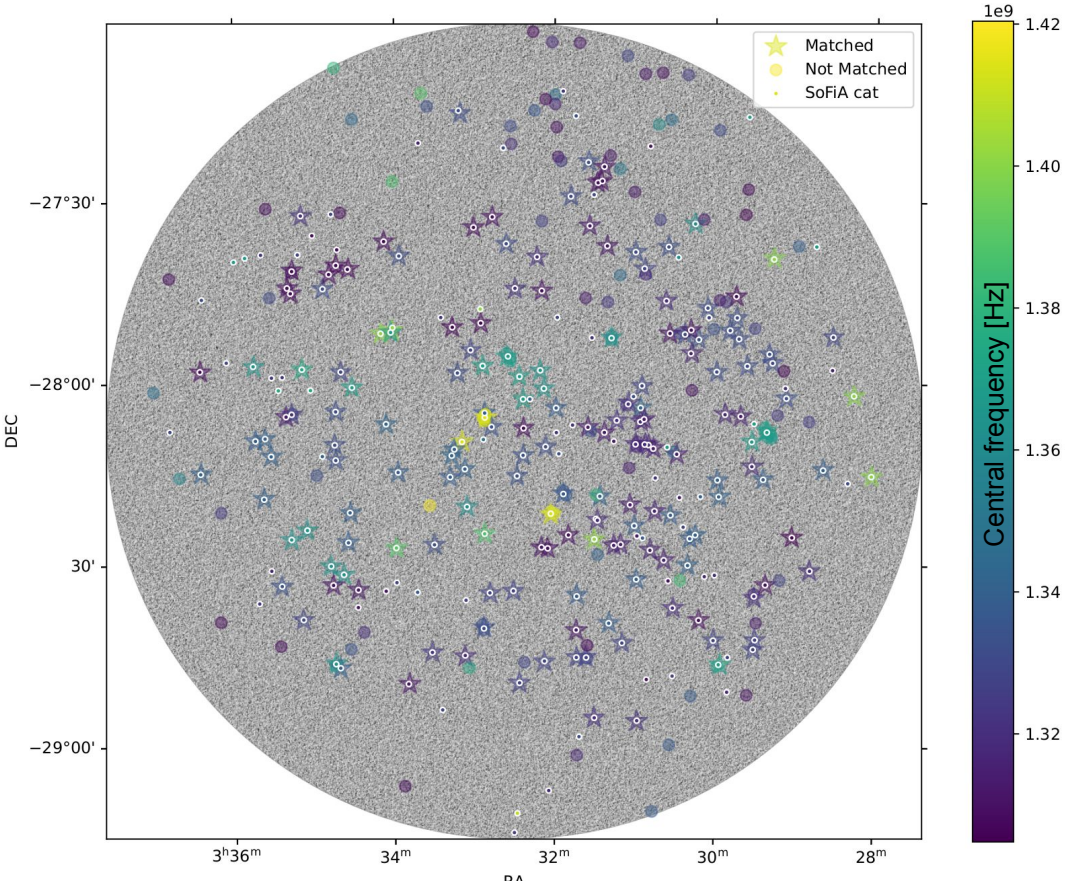
Dissimilarity:

- Resolution (space and freq)
- Pixels dynamics/Sensitivity
- Morphological diversity
- Instrumental specificities

Similar enough, **but requires**:

- Match **the pixel dynamics**
- Match **the sampling**

Preliminary results: Low z cube



Raw output: 354 detections
After (light) post-process: **283 detections**

	DR1 catalog* (243 sources)	**Not DR1 catalog***
Y-C model** (283 sources)	176	107
Not Y-C model**	67	

Recall ~ 73% ; Precision ~ 62%

Among the 107 unmatched detections: **80 are not at all close to DR1 catalog* detections**

*LADUMA Lowz catalog d.r. 1 derived from SoFiA and more

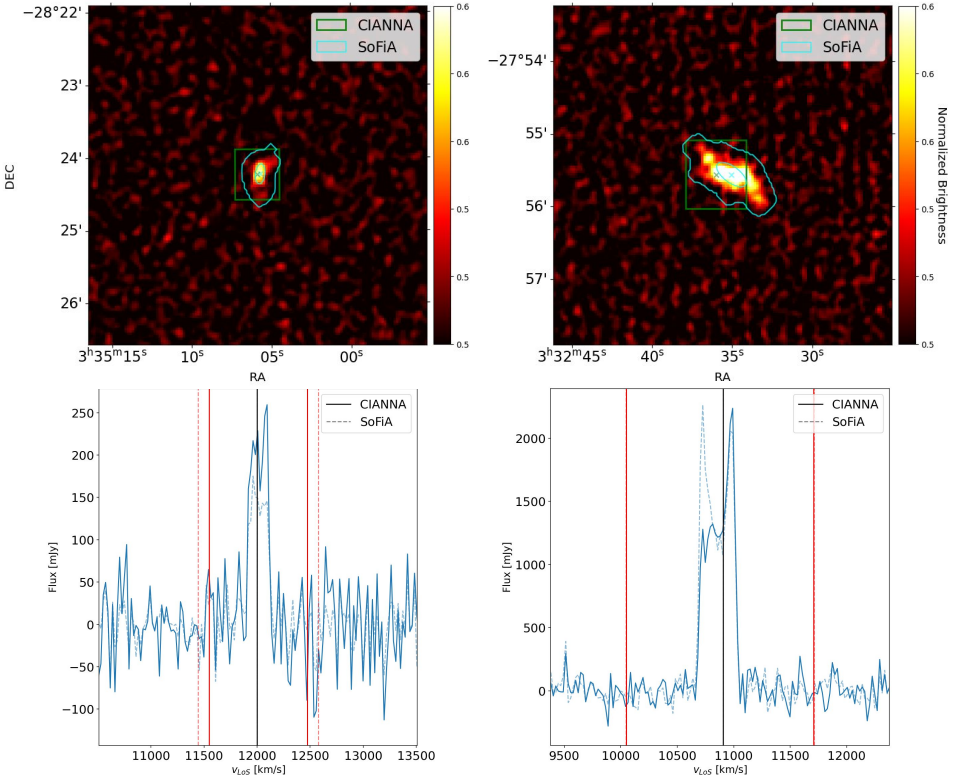
**Our YOLO-CIANNA model trained with a portion of SDC2 data smoothed in frequency

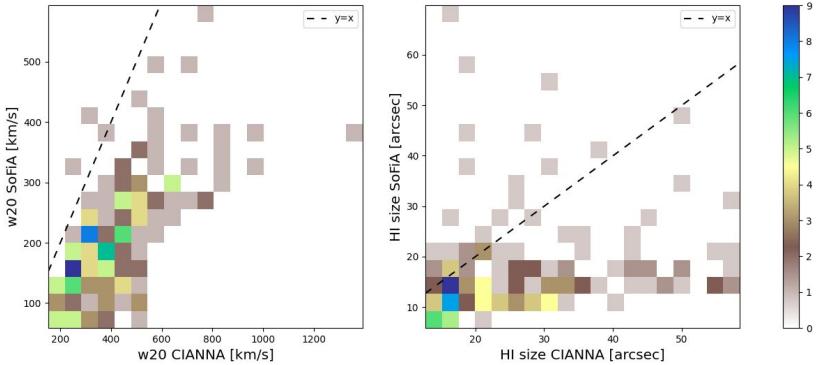
Recall = *N_match* / *N_ref*

Precision = *N_match* / *N_test*

Preliminary results: Low z cube

Our Y-C model matched detections



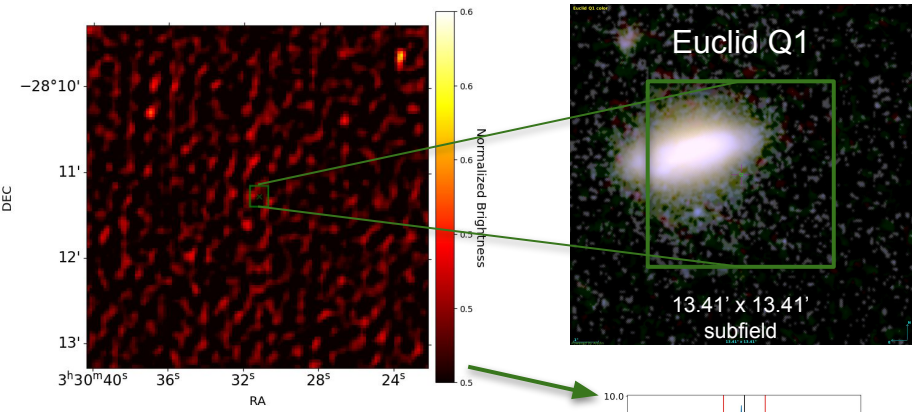


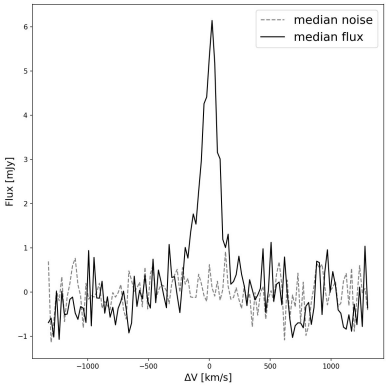
- Matched detections have **high confidence scores**
- Comparable detections for simple morphologies
- Difficult to characterize complex morphologies (training)
- Systematic difference: w20 and HI size (training)

Preliminary results: Low z cube

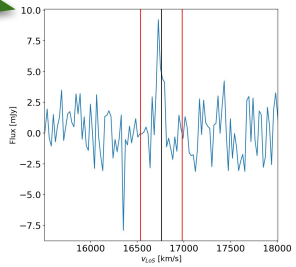
Our Y-C model candidates undetected in DR1

80 (total) **47 (after pruning) detecions**



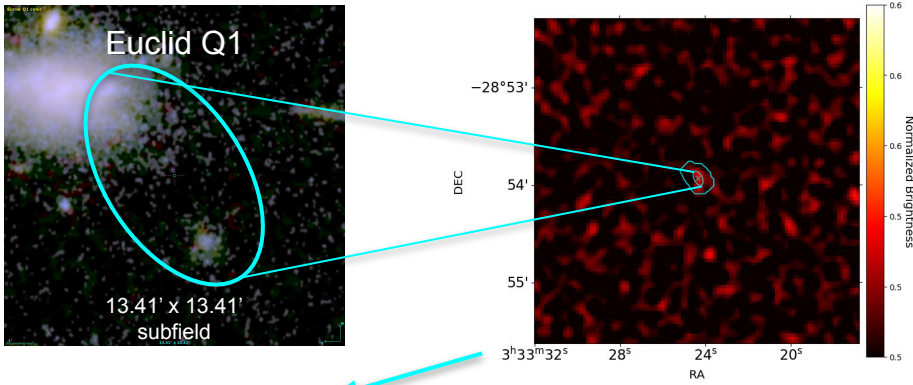


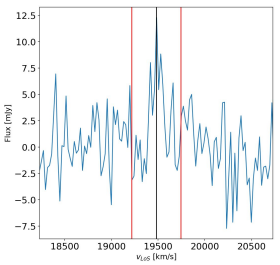
FWHM ~154 km/s

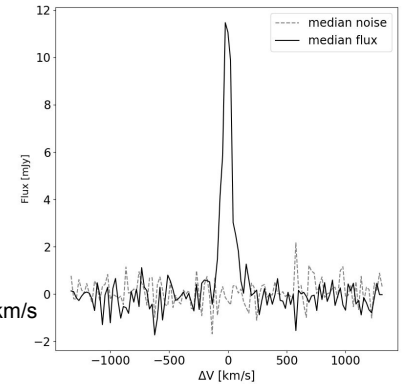


DR1 candidates undetected by our Y-C model

67 detecions





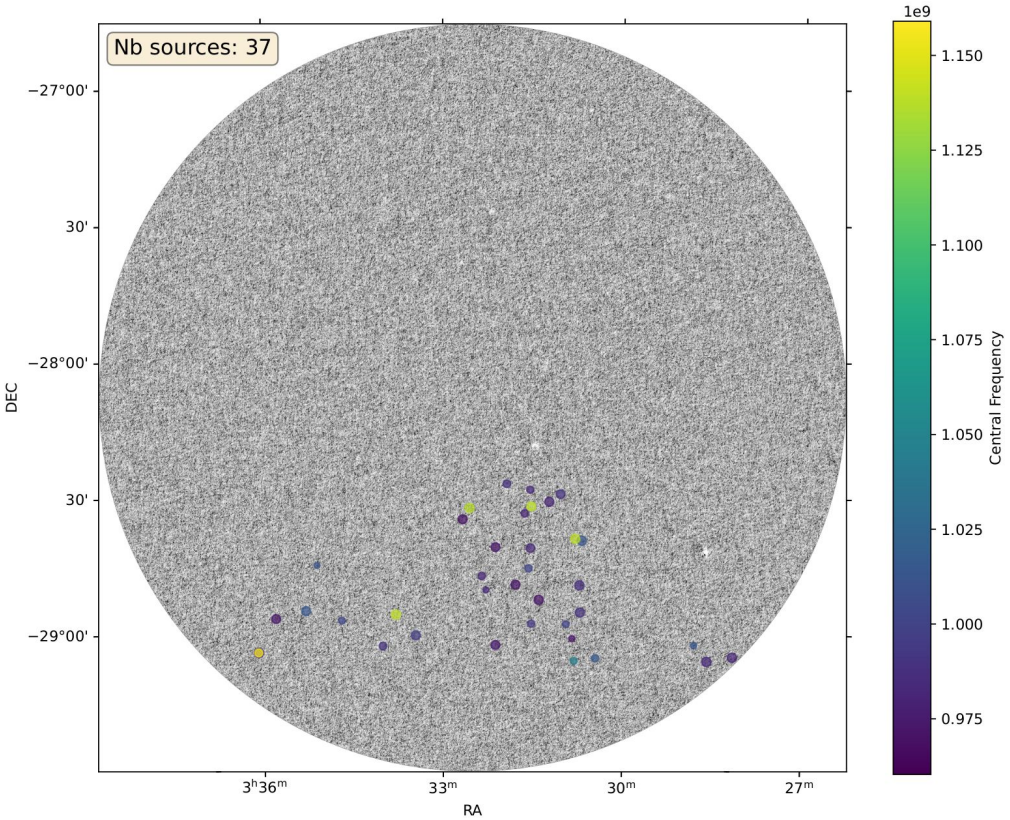


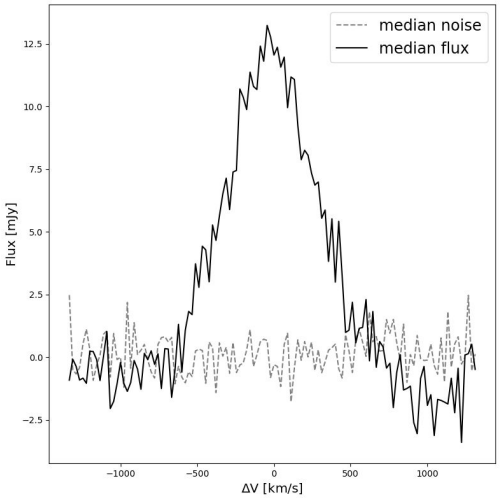
FWHM ~111 km/s

Both detections seems complementary

A. Anthore

Preliminary results: Mid z cube





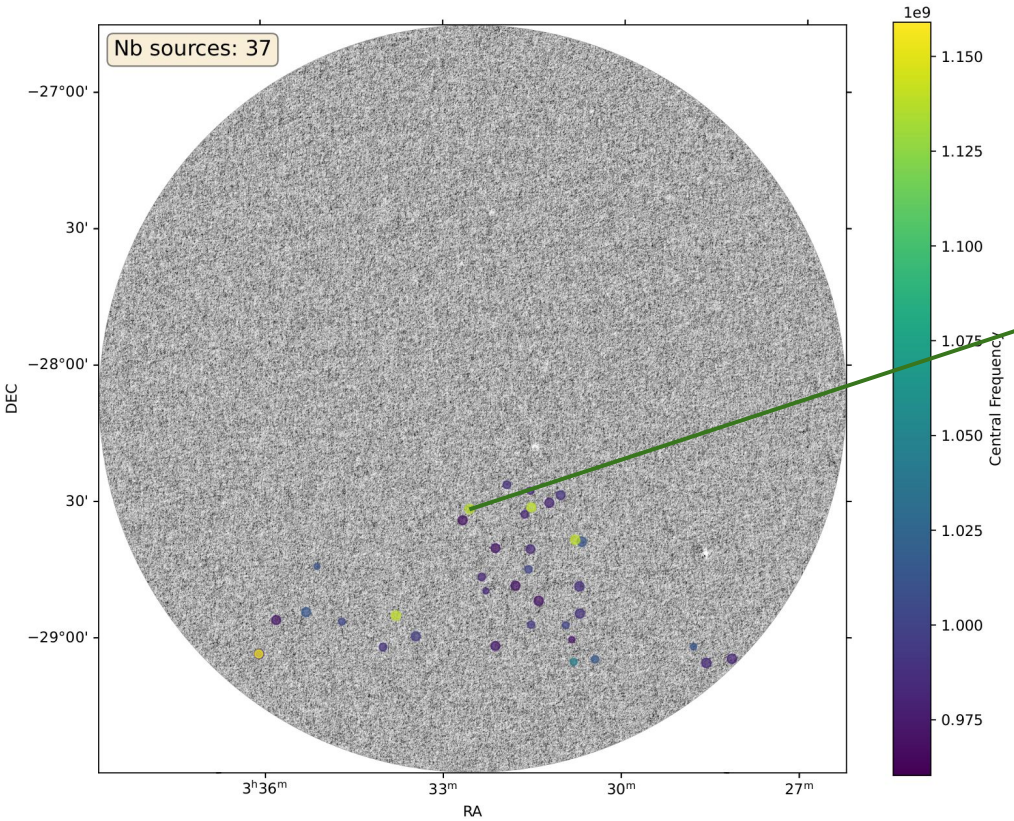
FWHM ~622 km/s

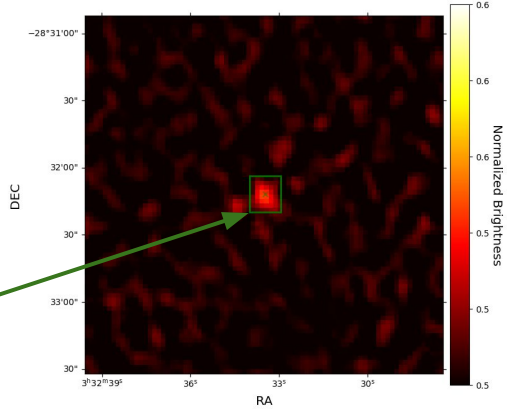
Restricted area (noise level + artifacts)

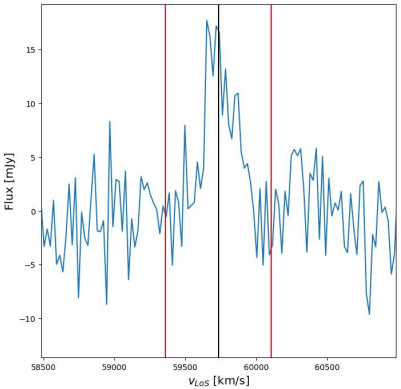
Wrong characterization (training, baseline, …)

Average confidence score is lower

Preliminary results: Mid z cube





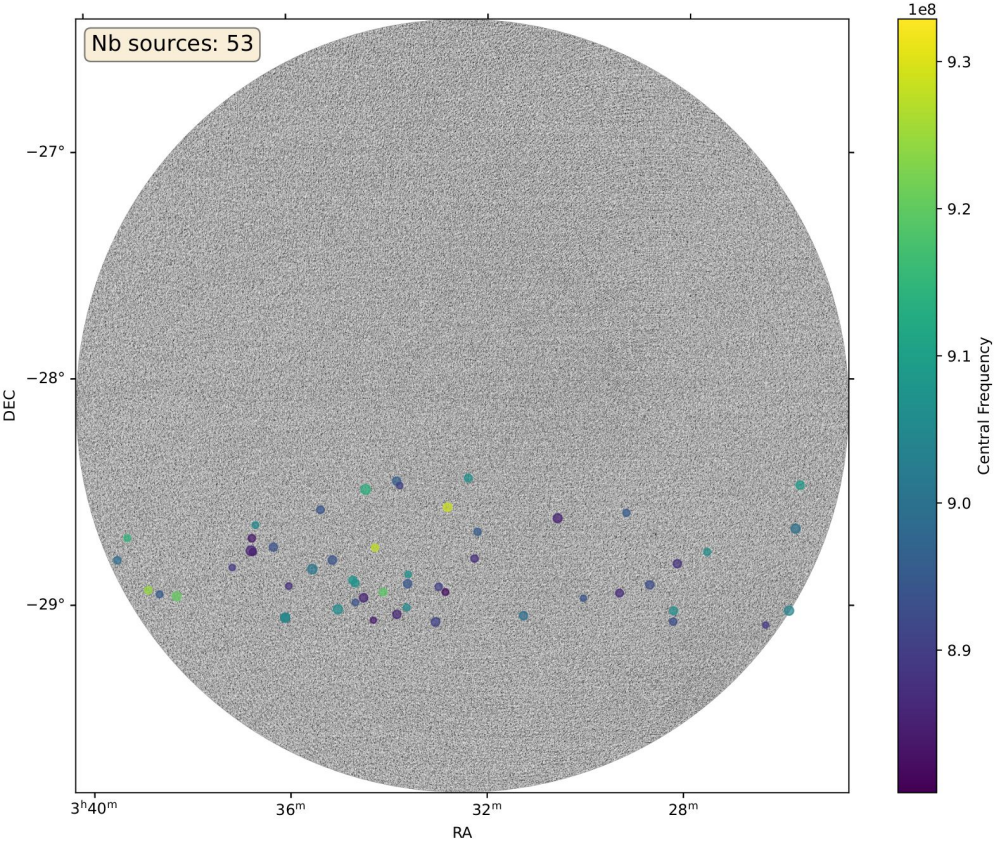


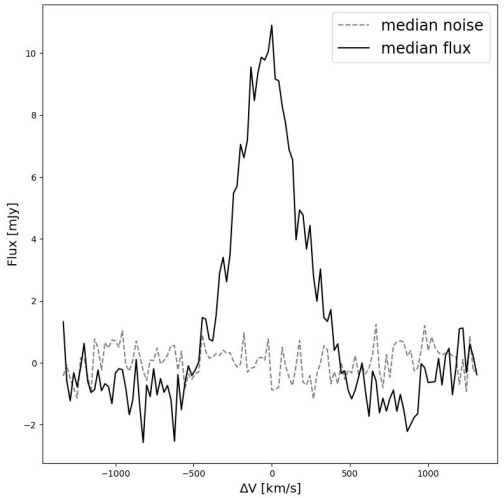
Restricted area (noise level + artifacts)

Wrong characterization (training, baseline, …)

Average confidence score is lower

Preliminary results: High z cube



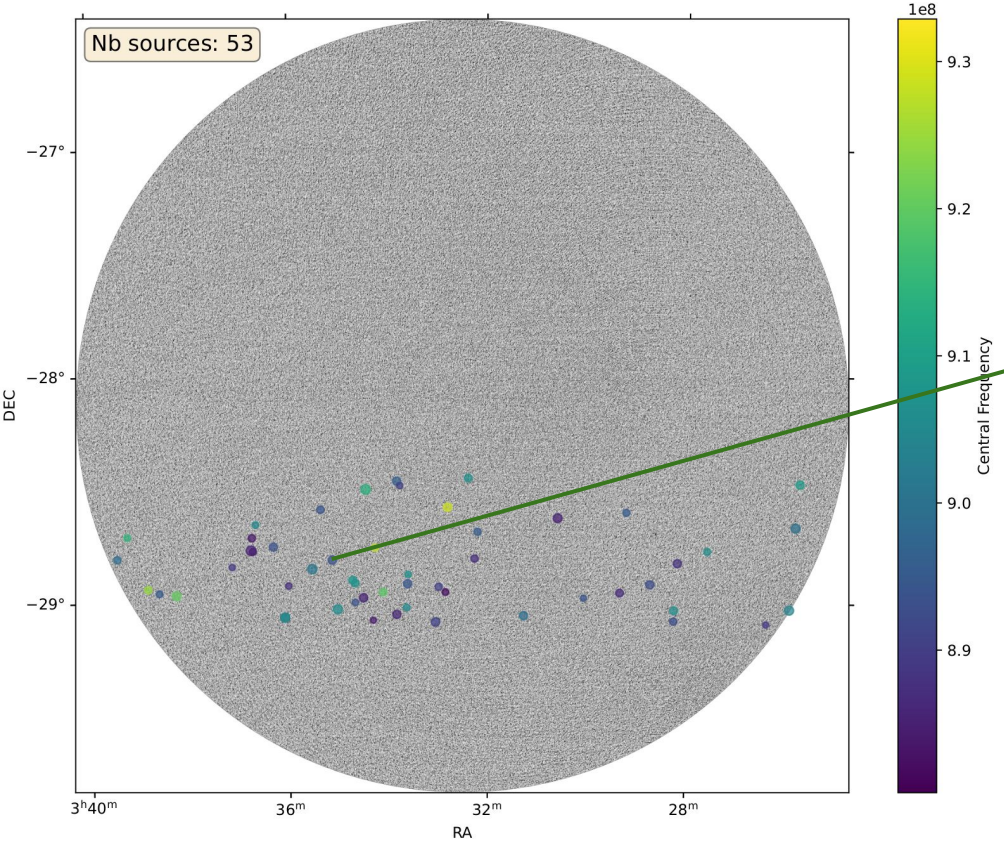


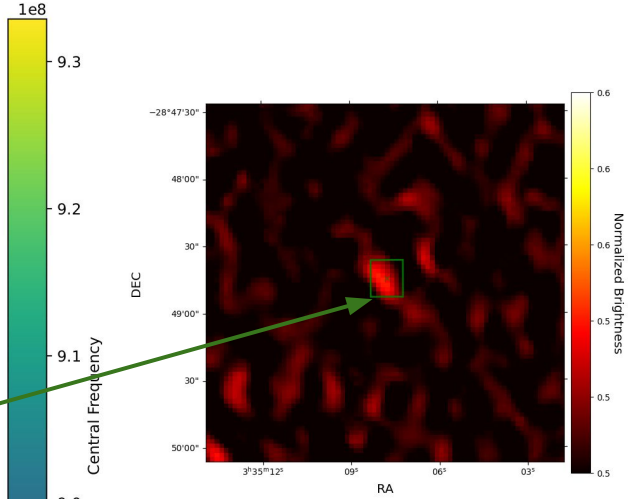
FWHM ~423 km/s

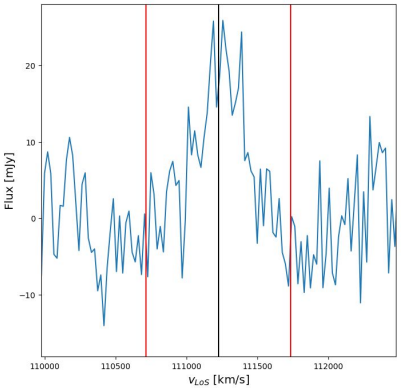
Same issues as the previous cube…

Average confidence score is even lower

Preliminary results: High z cube







Same issues as the previous cube…

Average confidence score is even lower

LADUMA DR1 summary

Our **YOLO-CIANNA model** with LADUMA DR 1.2 data**:**

- **Quick:** Training ~ 10h; Inference ~10 min/cubes
  With Nvidia H100 GPU (Jean-Zay)
- **Reliable:** Good agreement with DR1 detections
  Recall ~ 73% ; Precision ~ 62%
- **47 new candidates** at low z (w.r.t. Lowz catalog dr1)

Direct application of YOLO-CIANNA:

- Low z cube **90%**
- Mid z cube **60%**
- High z cube **60%**

Prospects:

- Better Mid/High z cube pre-processing **25%**
- Transfer-Learning approach **To be done**