

# Galaxy detection with deep learning in radio data, prospects for SKA

**David Cornu**

*LUX, Paris Observatory – Univ. PSL*

**LAM Seminar – 21 Sept. 2026**



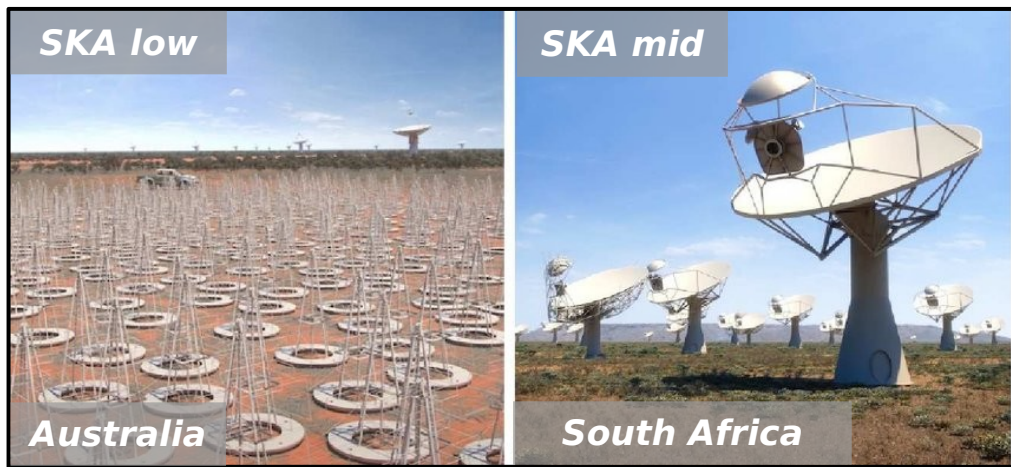
Observatoire  
de Paris



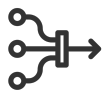
UNIVERSITÉ  
PARIS  
SCIENCES  
& LETTRES

PR[AI]RIE  
Paris School

# The Square Kilometer Array



## New big data reference in astronomy



**~1 TB / s** data rate  
(entering SDP centers)

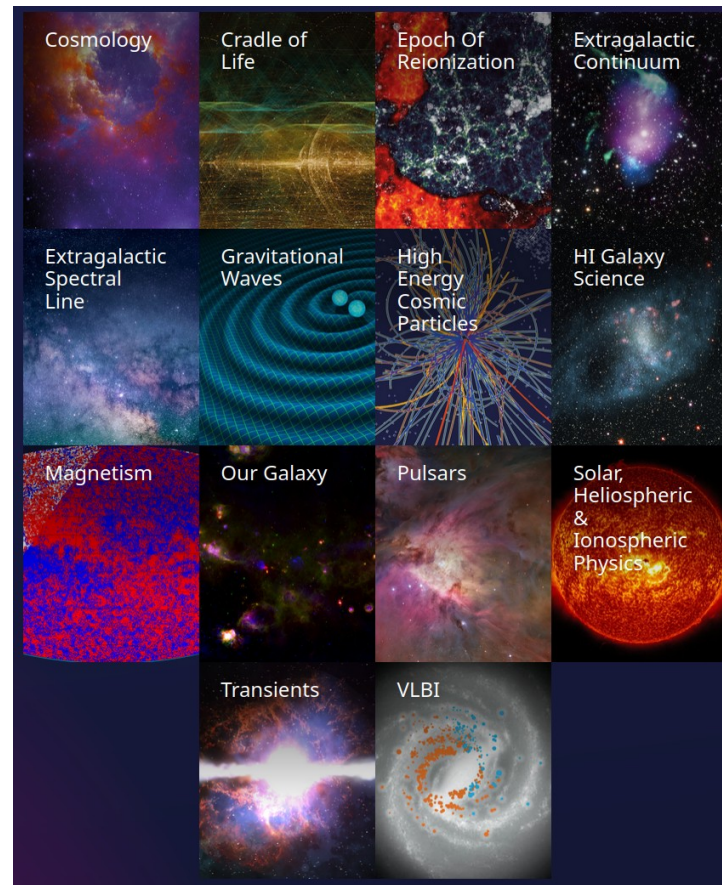


**~250 PB / year** stored

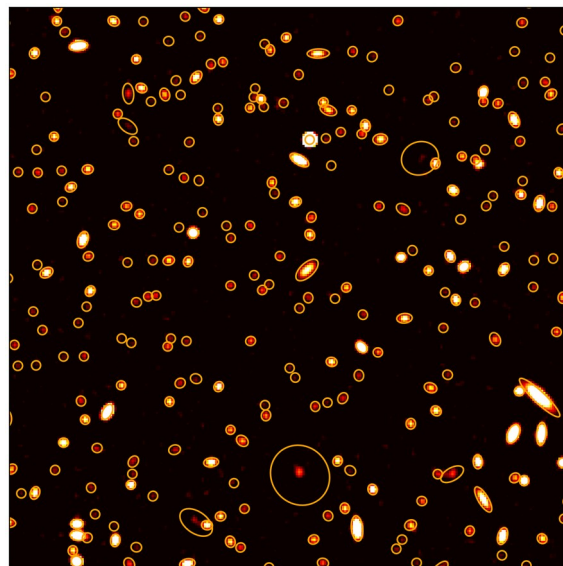
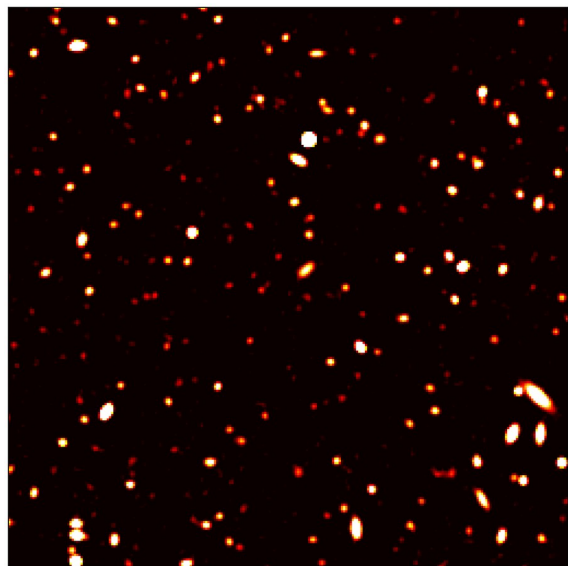
→ Most data pipelines don't scale well enough ...

Also true for other modern instruments → **Euclid, JWST, Vera Rubin, ...**

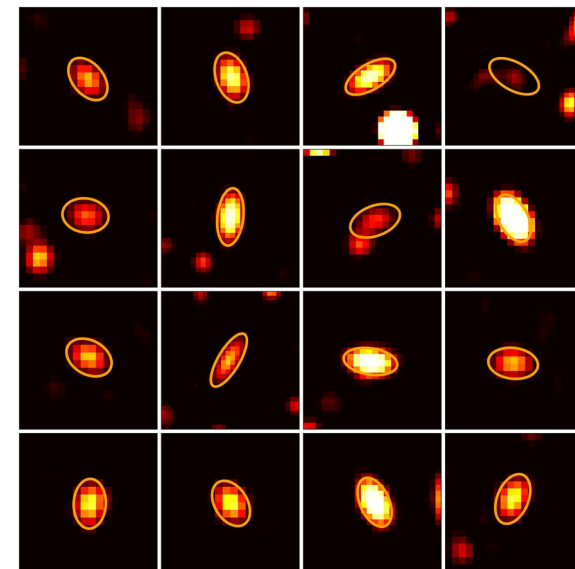
## First scientific observations ~2028



## Survey (images)



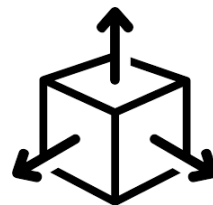
## Catalogs (lists)



*\*Simulated radio data (SDC1) at 560 MHz*

**Automated detection tools :**

*PyBDSF, SFIND, SourceExtractor,  
BLOBCAT, Aegan, Selavy, SoFiA, ...*

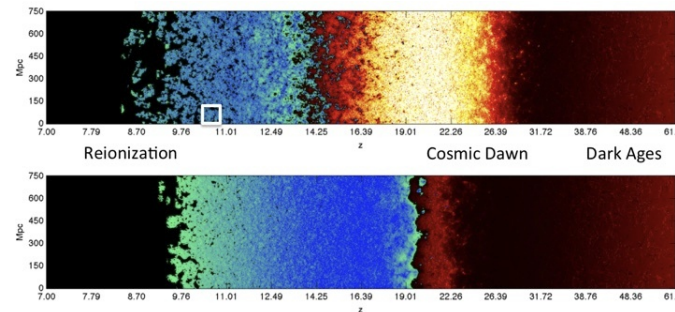
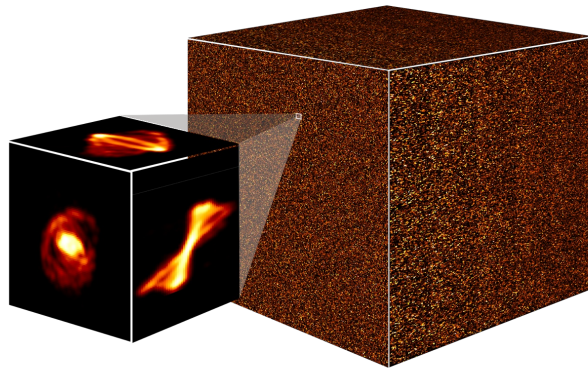
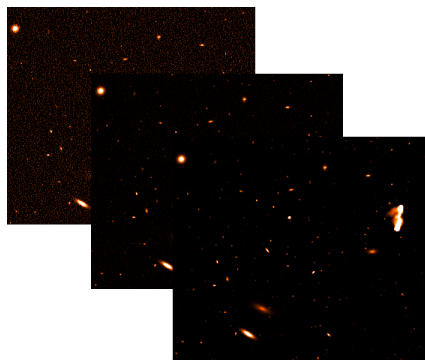
**Problems of classical tools:**

- 1 - Bad scalability with image size
- 2 - Expect simple source morphology

**Simulated datasets** that should resemble typical SKA data products

→ prepare astronomers, and stimulate the creation of new data analysis pipeline

### **Source (galaxy) detection and characterization**



**SDC1:** Continuum 2D images  
3 integration times x 3 bands

**Each image = 4 GB**

*From Dec 2018 to April 2019*

**SDC2:** Hyperspectral cube  
of HI emission

**Full cube = 1 TB**

*From Feb 2021 to July 2021*

**SDC3:** 21 cm emission  
Visibility and Image

**Full size ~ 7 TB**

*EoR Focused, 2023-2025*

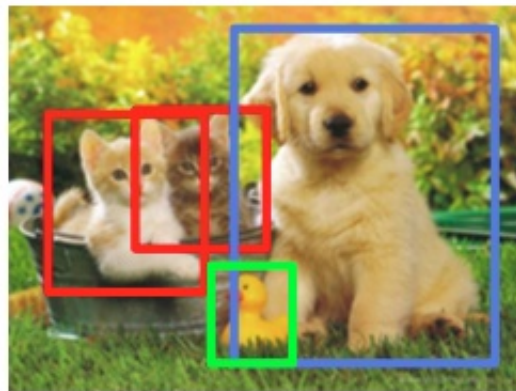
**(Obs.Paris) MINERVA team's objective** → create a **DL object detector** applicable to SDC1 & 2

**Classification**

CAT

**Classification  
+ Localization**

CAT

**Object Detection**

CAT, DOG, DUCK

**Instance  
Segmentation**

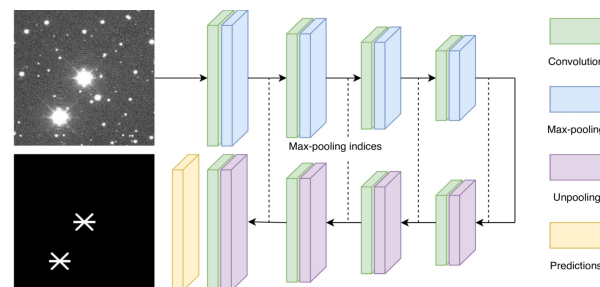
CAT, DOG, DUCK

**Bounding box** = smallest box that contains all pixels belonging to an object

## Segmentation

**Methods:** U-Nets and variants

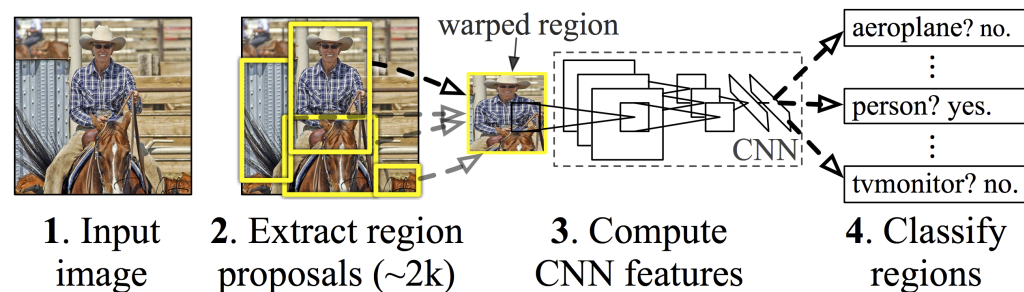
**Pros:** produce segmentation maps, provides usable latent space, ...



## Region-based

**Methods :** R-CNN (Fast and Faster), SPP-net, Mask R-CNN, ...

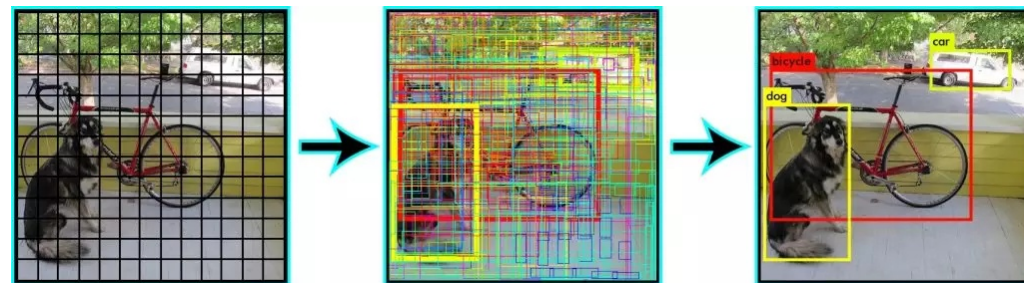
**Pros:** Best accuracy, ...



## Regression-based

**Methods :** SSD (Single Shot Detector), YOLO (You Only Look Once), ...

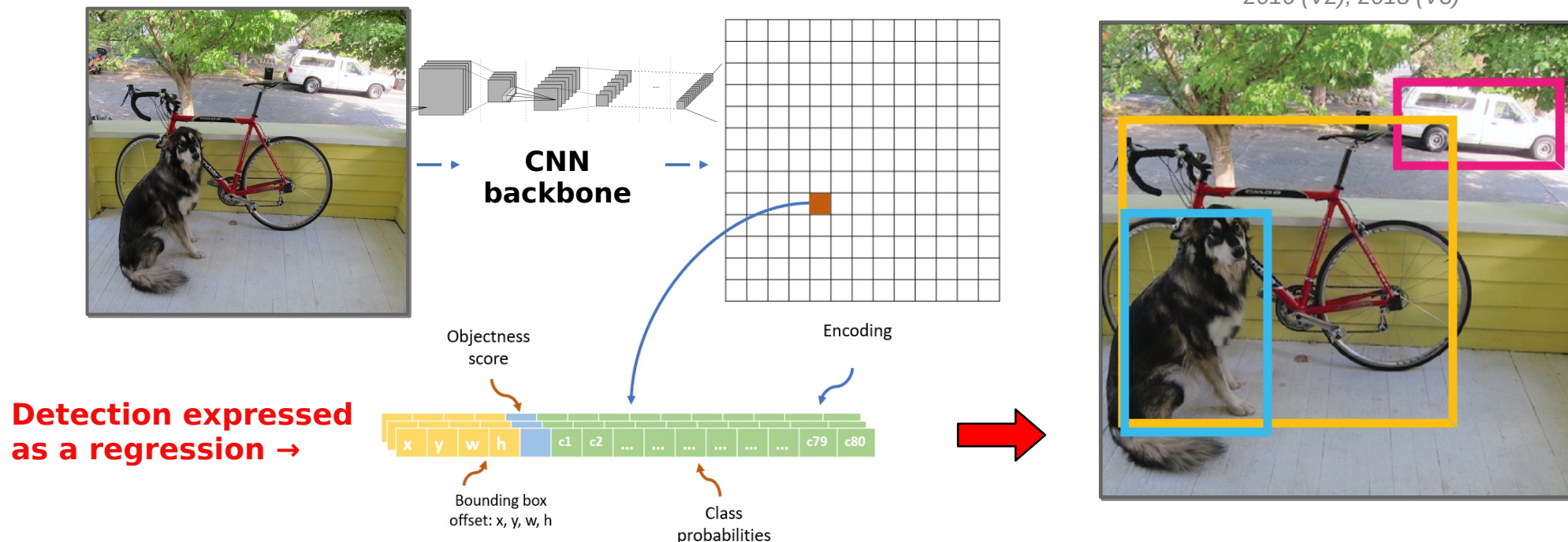
**Pros:** Very Fast, straightforward architecture,...



# You Only Look Once (YOLO)

## Supervised regression-based deep learning object detector

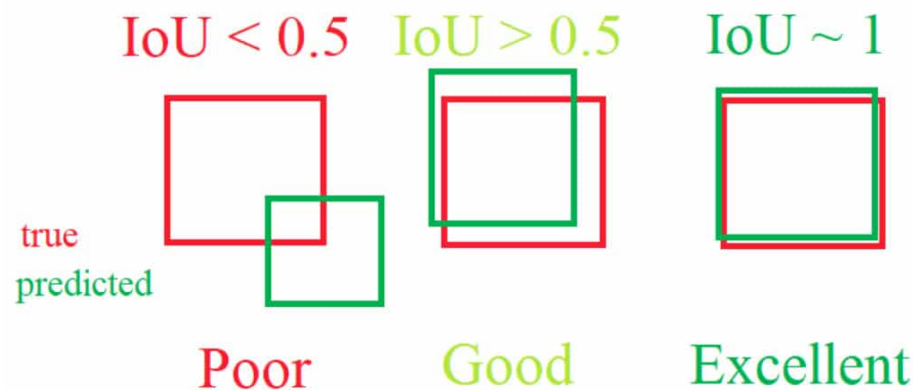
Redmon et al. 2015 (V1),  
2016 (V2), 2018 (V3)



**Supervised approach → Learn from a list of images in which objects of interest have already been identified (labeled images)**

## How to estimate box « correctness » ?

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

For each box the network will predict an “objectness” score defined as:

$$Pr(\text{object}) * IOU(b, \text{object}) = \sigma(t_o)$$

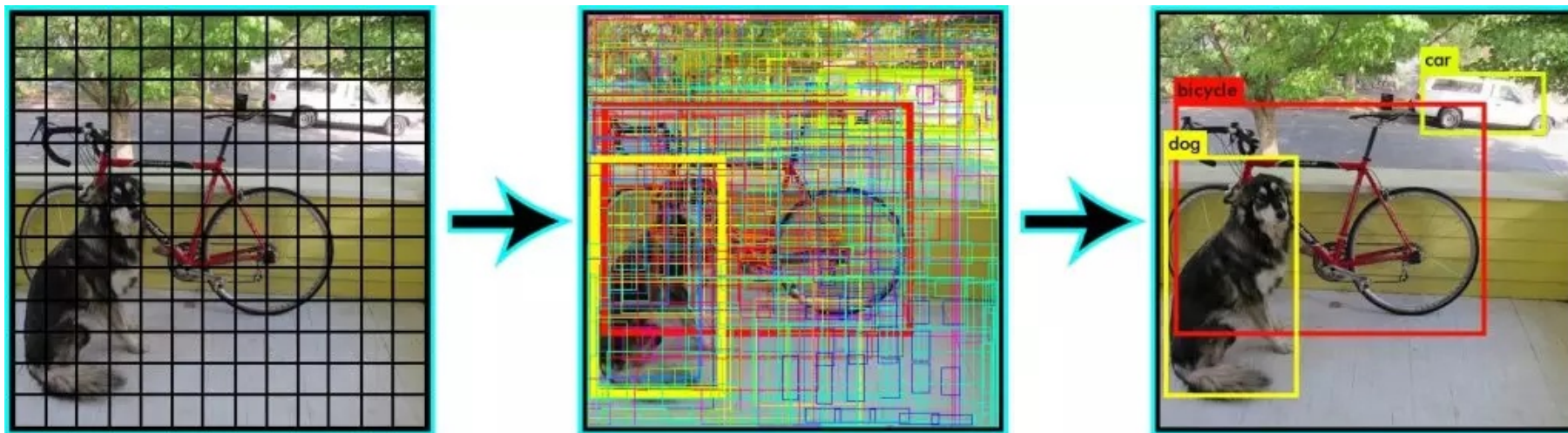
**For each training image:**

**1) Forward:** Use the current model to predict boxes

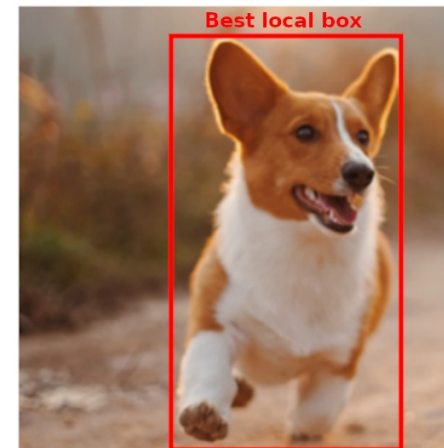
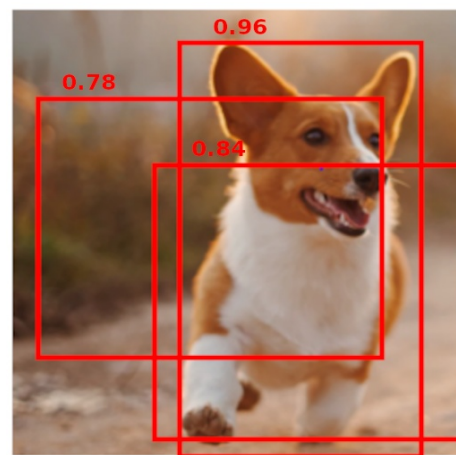
**2) Assignment phase:** Search the “best” predicted box for each target box (mostly IoU based)

**3) Add each detection unit contribution to the loss and update the model:**

$$\begin{aligned}
 \mathcal{L} = & \sum_{i=0}^{N_g} \sum_{j=0}^{N_b} \mathbb{1}_{ij}^{\text{match}} \left( \begin{aligned} & \lambda_{\text{pos}} \left[ (o_{ij}^x - \hat{o}_{ij}^x)^2 + (o_{ij}^y - \hat{o}_{ij}^y)^2 \right] \\ & + \lambda_{\text{size}} \left[ (o_{ij}^w - \hat{o}_{ij}^w)^2 + (o_{ij}^h - \hat{o}_{ij}^h)^2 \right] \\ & + \lambda_{\text{obj}} \left( o_{ij} - \text{IoU}_{ij} \right)^2 \\ & + \lambda_{\text{class}} \sum_k^{N_C} \left( -\hat{C}_{ij}^k \log(C_{ij}^k) \right) \end{aligned} \right) \quad \text{--- A. Best matches: update all parameters toward the target and increase the Obj.} \\
 & + \sum_{i=0}^{N_g} \sum_{j=0}^{N_b} \mathbb{1}_{ij}^{\text{void}} \lambda_{\text{void}}^j \lambda_{\text{obj}} (o_{ij} - 0)^2 \quad \text{--- B. Good (IoU > 0.5) but not best units are ignored} \\
 & \quad \quad \quad \text{--- C. No match / background: only lower the Obj.}
 \end{aligned}$$



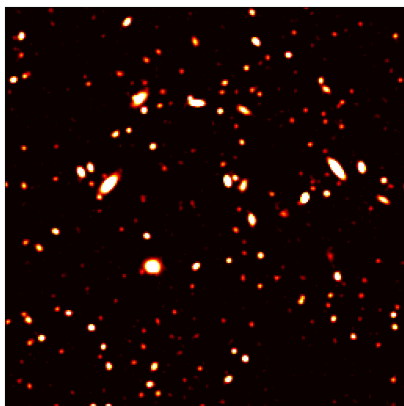
- 1) Most probable boxes are kept based on an **objectness threshold**
- 2) Removes multiple detections of single objects using **Non Max Supression**



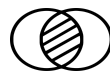
## Astronomical images and objects do not fulfill classical implicit biases



≠



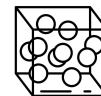
No sharp edges



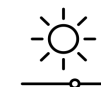
Source blending



Disjoint objects



Extreme density  
(high or low)



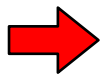
Large dynamic



Noise



Redmon et al. 2016+



Cornu et al. 2024

Complete rework for  
astronomical applications

+ Sub-pixel  
positioning

+ Resilient to  
high density

+ Accurate flux  
quantization

+ On-the-fly  
characterization

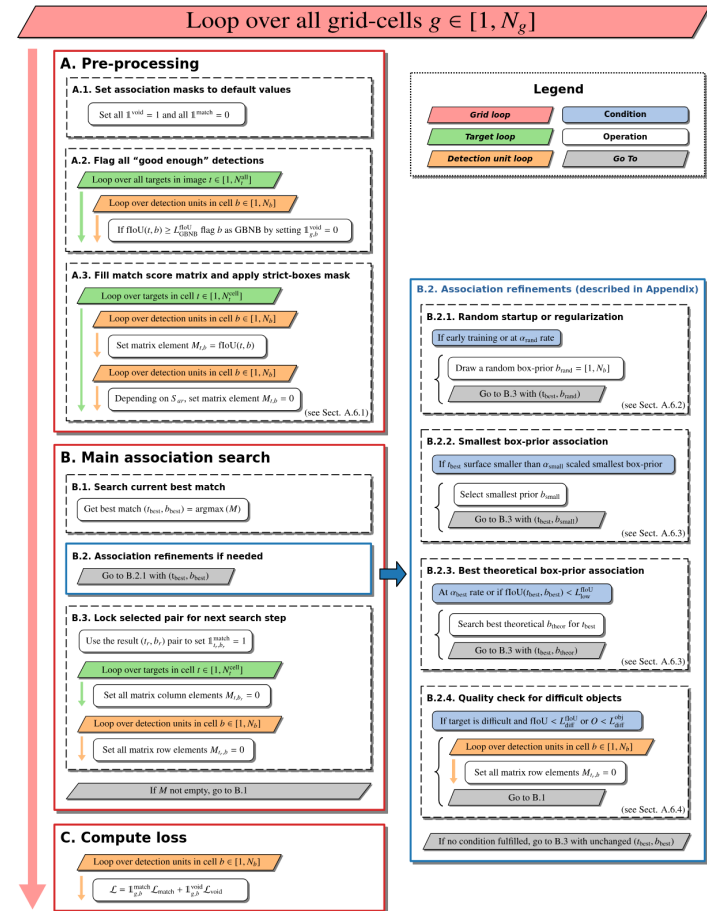
+ 3D detection  
(end-to-end 3D conv.)

Details are in **Cornu et al. 2024** (42 pages paper, 11 pages of method)

*Reworked Loss function*

$$\begin{aligned}
 \mathcal{L} = & \sum_{i=0}^{N_g} \sum_{j=0}^{N_b} \mathbb{1}_{ij}^{\text{match}} \left( \lambda_{\text{pos}} \left[ (o_{ij}^x - \hat{o}_{ij}^x)^2 + (o_{ij}^y - \hat{o}_{ij}^y)^2 \right] \right. \\
 & + \lambda_{\text{size}} \left[ (o_{ij}^w - \hat{o}_{ij}^w)^2 + (o_{ij}^h - \hat{o}_{ij}^h)^2 \right] \\
 & + \lambda_{\text{class}} \mathbb{1}_{ij}^C \sum_k^{N_C} \left( -\hat{C}_{ij}^k \log(C_{ij}^k) \right) \\
 & + \lambda_{\text{param}} \mathbb{1}_{ij}^P \sum_k^{N_p} \gamma^k \left( p_{ij}^k - \hat{p}_{ij}^k \right)^2 \\
 & + \lambda_{\text{prob}} \mathbb{1}_{ij}^P \left( P_{ij} - 1 \right)^2 \\
 & \left. + \lambda_{\text{obj}} \mathbb{1}_{ij}^O \left( O_{ij} - \text{fIoU}_{ij} \right)^2 \right) \\
 & + \sum_{i=0}^{N_g} \sum_{j=0}^{N_b} \mathbb{1}_{ij}^{\text{void}} \lambda_{\text{void}}^j \left( \lambda_{\text{prob}} \left( P_{ij} - 0 \right)^2 + \lambda_{\text{obj}} \left( O_{ij} - 0 \right)^2 \right).
 \end{aligned}$$

*New assignment function*



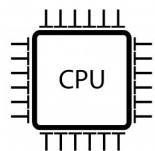


*Convolutional Interactive Artificial Neural Networks by/for Astrophysicists*

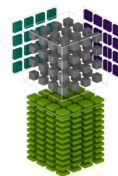
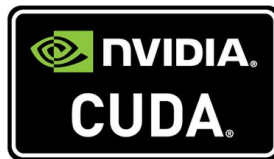
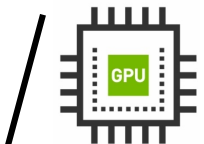
**General purpose ANN development framework**

**BUT** specialized for **astronomical applications**

*Development team*



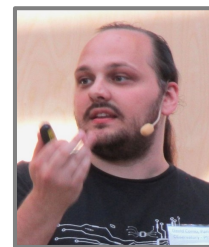
OpenMP  
OpenBLAS



High level  
interface



Large hardware compatibility



D. Cornu



G. Sainton  
(IR IA)



A. Doussot  
(IR HPC)

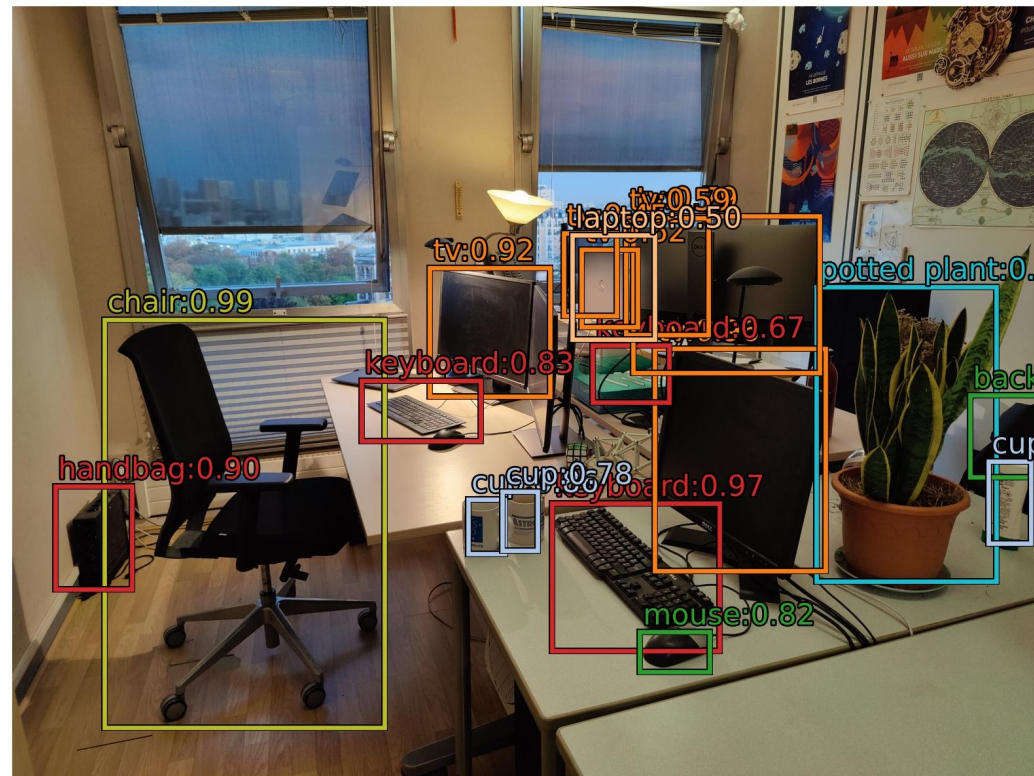
[github.com/Deyht/CIANNA](https://github.com/Deyht/CIANNA)  
or  
[cianna.obspm.fr](https://cianna.obspm.fr)

*Archived on  
Zenodo and ASCL*



→ include our custom **YOLO-CIANNA** detector

Model pre-trained on **ImageNet** and then trained on the **COCO** datasets independently.



**Data:** Simulated continuum image

- 5.5 square degree area (pixel size 0.6")
- 560 MHz, 1000h integration time
- **4GB image (32 768 pixels square)**

**The challenge:**

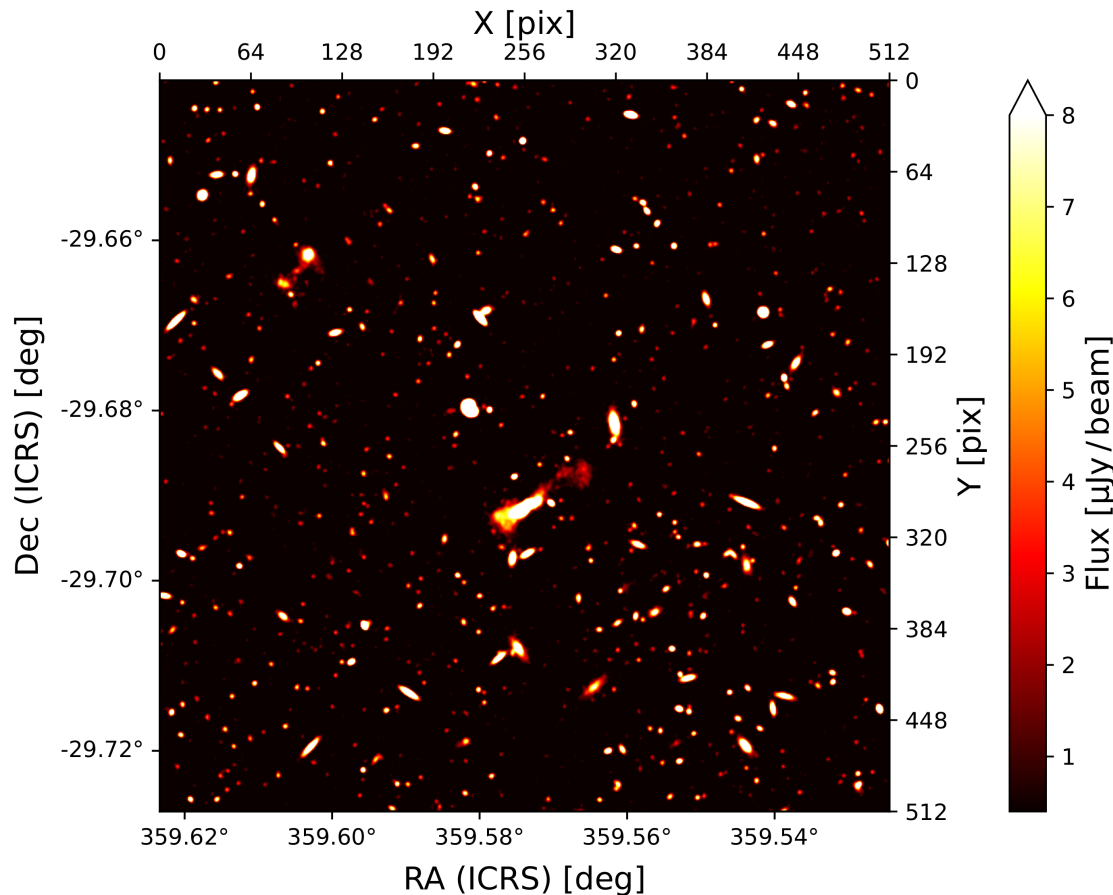
1. Find the sources (RA, Dec)
2. Characterize each source:  
→ (Flux, Bmaj, Bmin, PA, ...)

Training labels provided for a subpart of the image (5% of the surface, ~34 000 sources).

→ **SDC1 summary paper, Bonaldi et al. 2021**

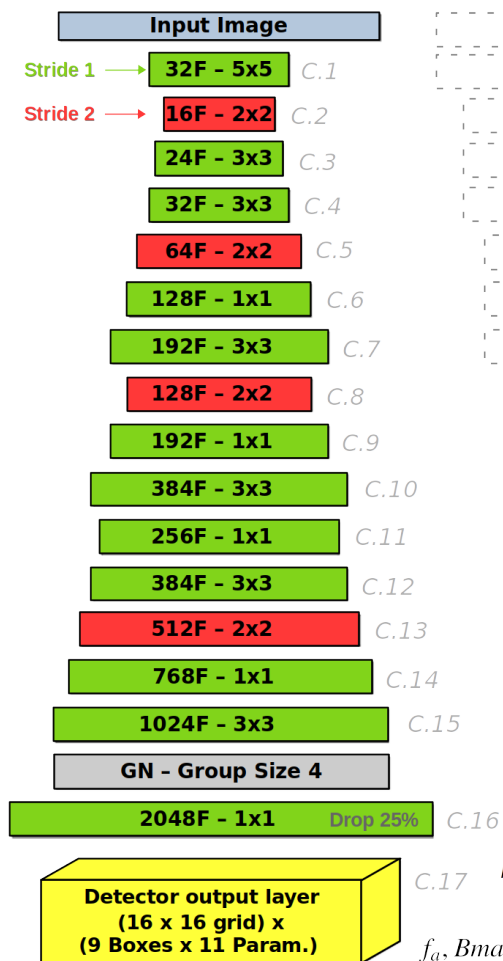
SKA SDC1 took place early 2020. Data from the challenge are still freely accessible on the dedicated [web-page](#).

*Example 512<sup>2</sup> sub-field (~5')*



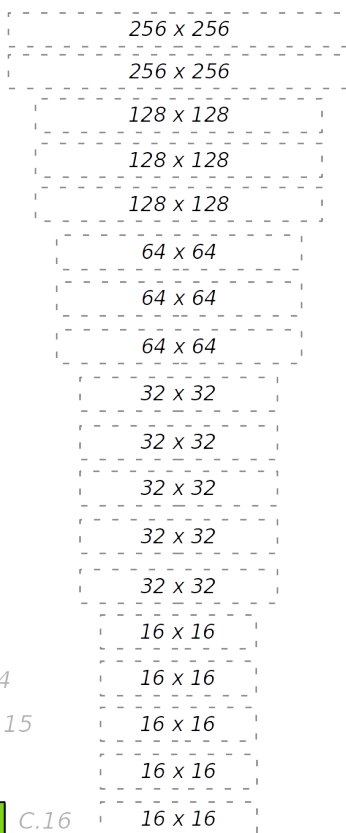
## Network layers

Learned parameters



## Image / Activation

Spatial reduction



Pred. for each box:  
 $x, y, w, h, P, O$   
 $f_a, Bmaj, Bmin, \cos(PA), \sin(PA)$

## Architecture:

- 17 conv. layers** → **~13 Million parameters (~50MB)**  
→ +8% in score compared to the classical darknet19 backbone
- 9 box priors** ranging from 10 to 32 pixels
- 5 additional source param.:** Flux, Bmaj, Bmin, cos(PA), sin(PA)

Training the network using  CIANNA

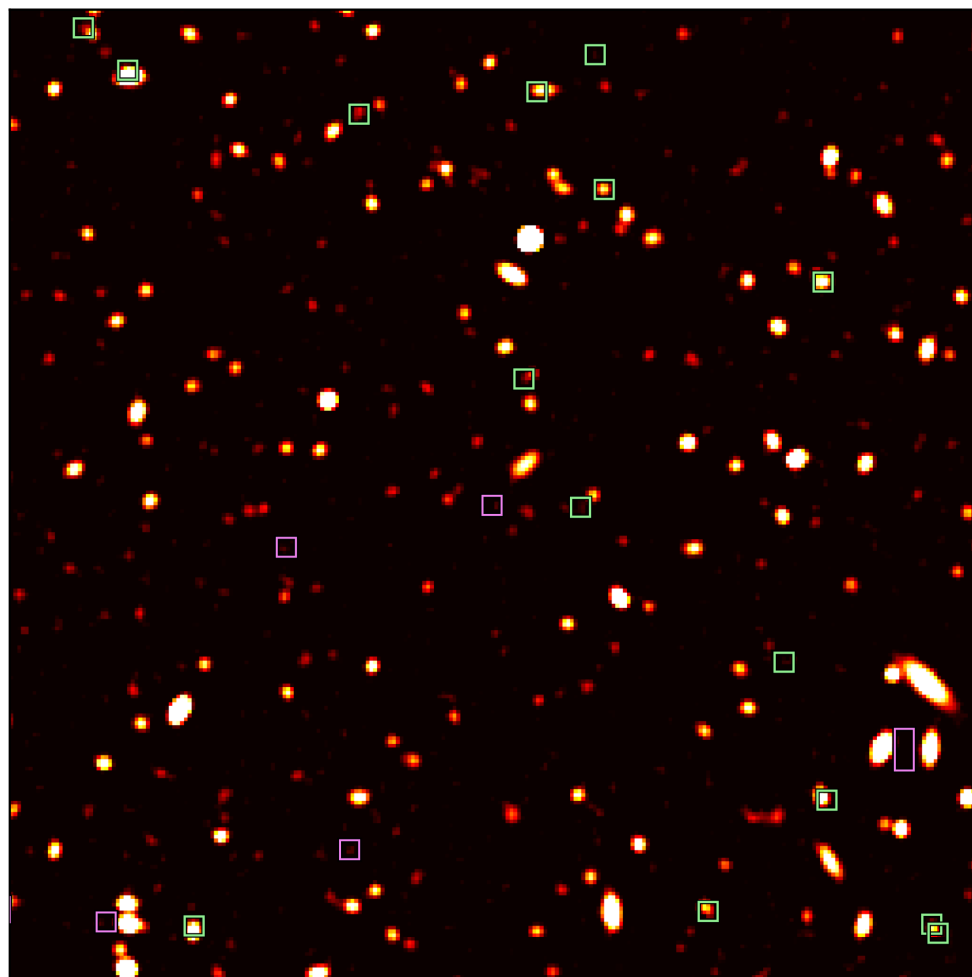
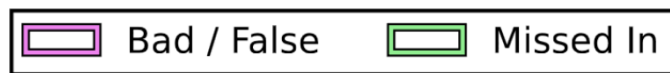
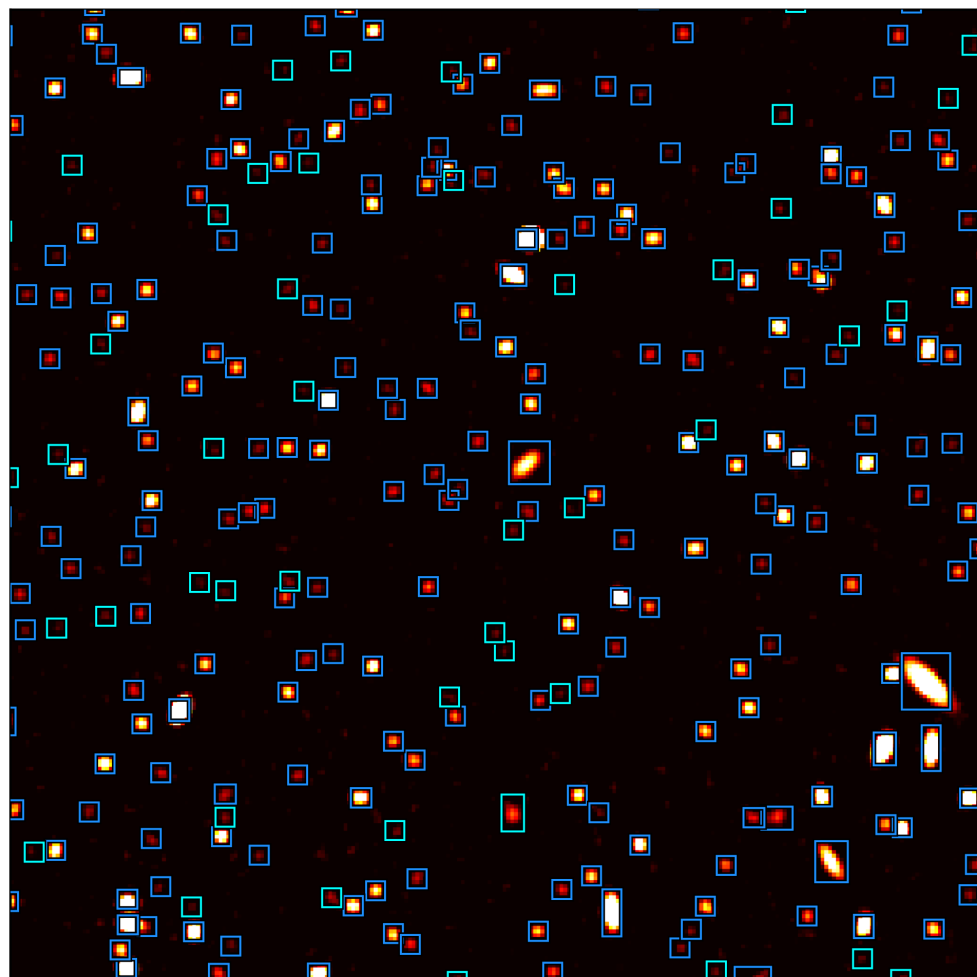
- 256x256** cutouts are randomly selected in the training area (54 MB)
- ~ 34000 sources** in the selected training catalog
- Augmented based on cutout position and flips

Training takes **~ 4 hours** on a single modern GPU (RTX 4090)

## Inference:

- The full SKA SDC1 image is split in 512x512 regions with an overlap of 32 pixels, → **~4500 images**
- Overlapping regions filtered with a dedicated NMS

The full inference in FP16-TC takes **~8 sec → 130 Mpix/s**



Based on **Bonaldi et al. 2021** + submitted catalogs

### Challenge metric

→ **Score = f (completeness, purity, characterization)**

*Multi-D match criteria*

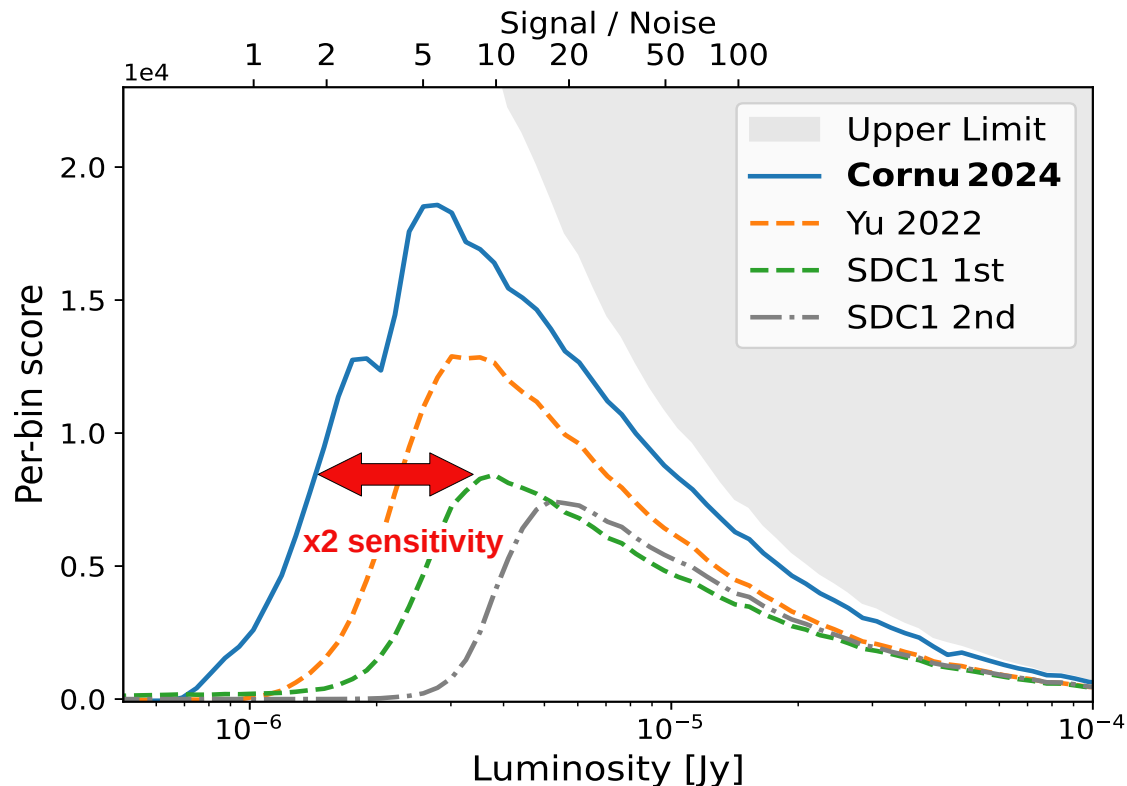
$$E_{tot} = \sqrt{E_{pos}^2 + E_{size}^2 + E_{flux}^2}, \text{ with}$$

$$E_{pos} \propto \sqrt{(x - \hat{x})^2 + (y - \hat{y})^2 / \hat{S}'},$$

$$E_{size} \propto |S - \hat{S}| / \hat{S}', \text{ and}$$

$$E_{flux} \propto |f - \hat{f}| / \hat{f},$$

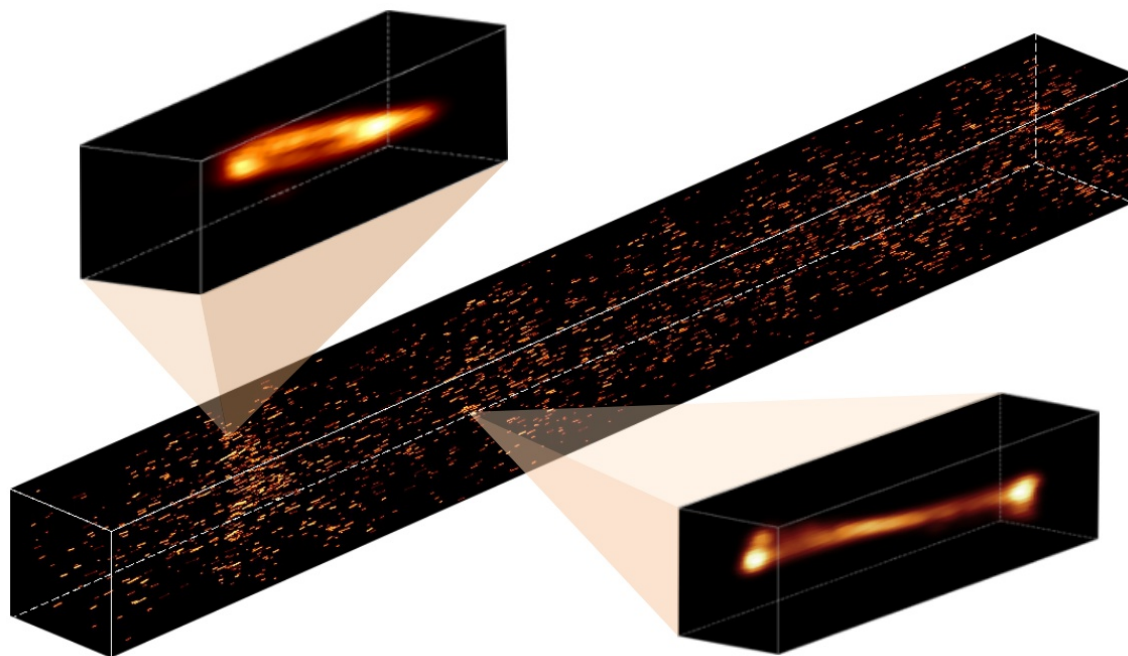
$$s_i = \frac{1}{7} \sum_j s_i^j \quad M_s = \sum_i^{N_{match}} s_i - N_{false}$$



→ **SDC1 is still a very interesting dataset for source detection pipeline development !**

**YOLO-CIANNA SDC1 → Cornu et al. 2024**

# SCIENCE DATA CHALLENGE 2



**Data:** a 3D cube of simulated HI emission

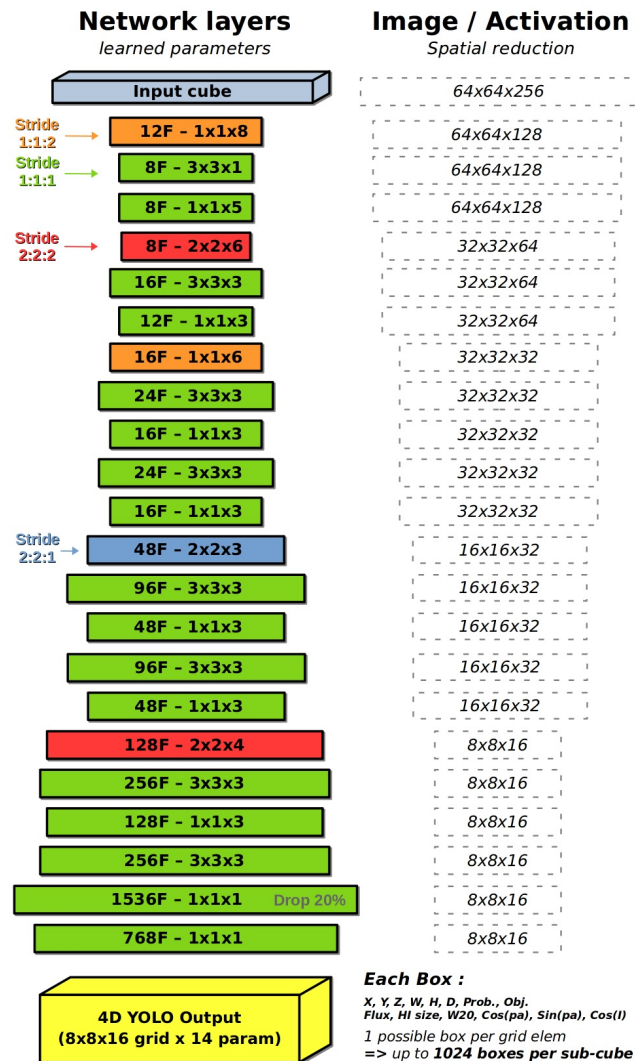
- 20 square degree area
- 950 to 1150 MHz frequency (30KHz res;  $z = 0.235\text{--}0.495$ )
- **2000h integration time**
- **Near 1 TB cube (5851 x 5851 x 6668)**

## The challenge:

1. Find the sources (RA, Dec, Freq)
2. Characterize each source:  
→ Flux, HI size, line width, PA, Inclination

→ **SDC2 summary paper, Hartley et al. 2023**

SKA SDC2 took place in 2021. Data from the challenge are still freely accessible on the dedicated [web-page](#).



## YOLO parameters:

- **Generalized to 3D detection**
- **23 layers** ~ 4 Million parameters
- 1 single box prior per grid element
- **6 additional source param.**

## Training the network using CIANNA

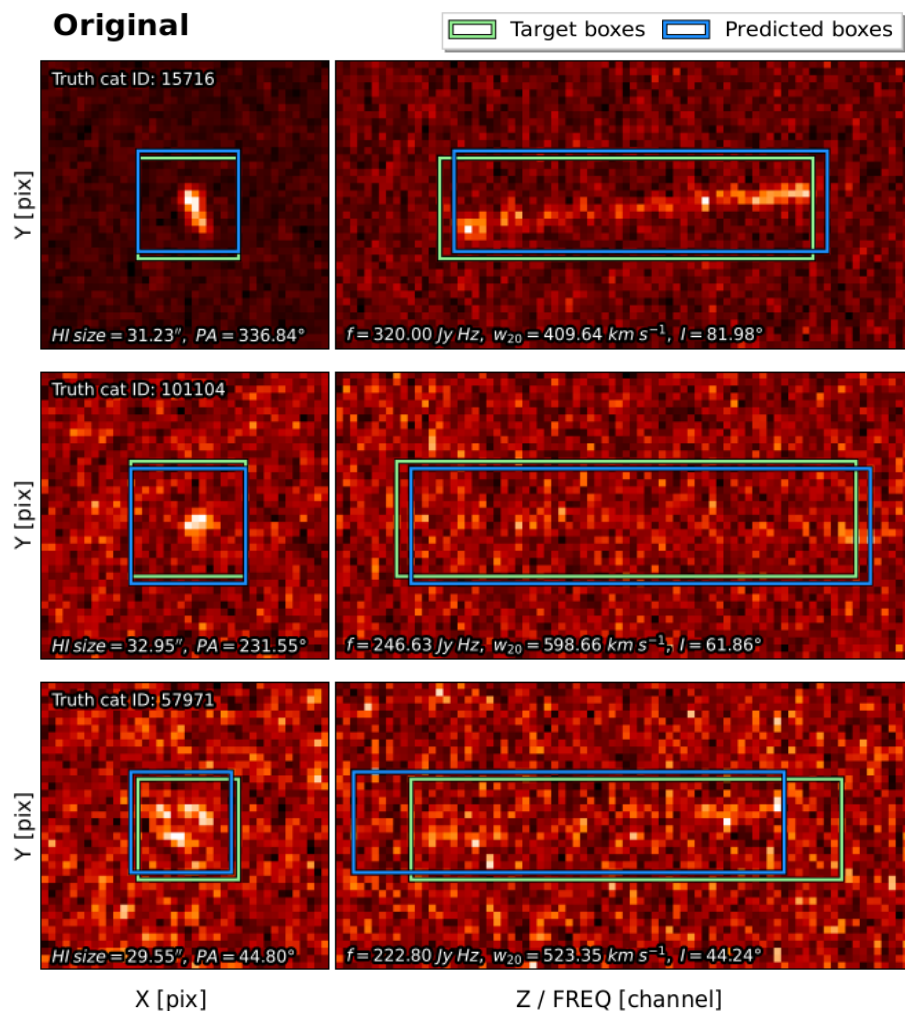
- **64x64x256** cubes are randomly selected in the training area (40 GB)
- Around **2000 sources** in the selected train catalog
- Data are augmented using shifting and flips

Training takes **~12 hours** on a single modern GPU (RTX 4090)  
→ already good results after 6-8h

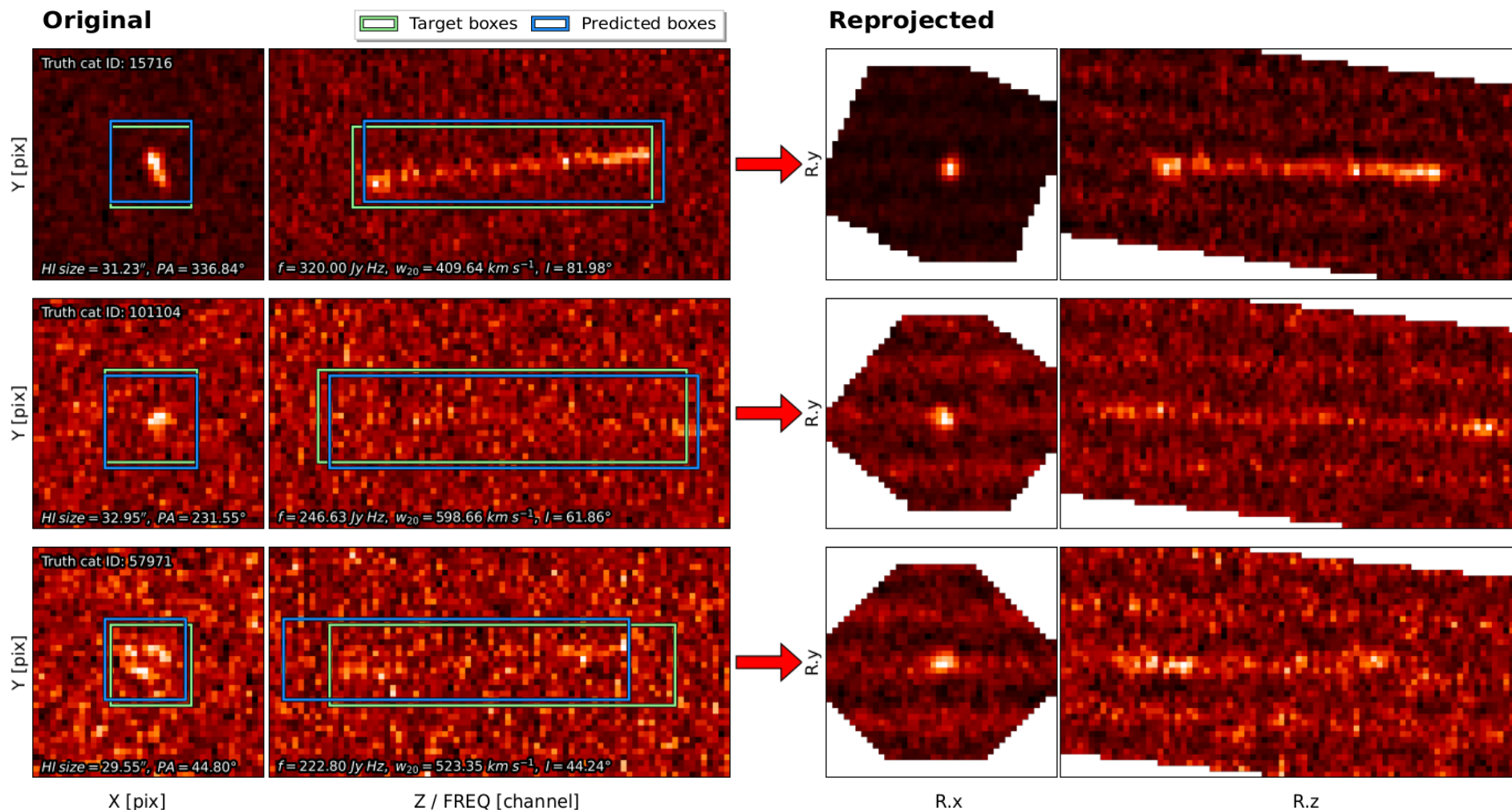
## Inference:

- The full SKA SDC2 1 TB cube is split in regions with large overlaps
- Box in overlapping regions are filtered with a dedicated NMS

**The full cube prediction takes ~30 min** (dominated by data loading time) **using a single RTX 4090 (260 sub-cubes per s.)**



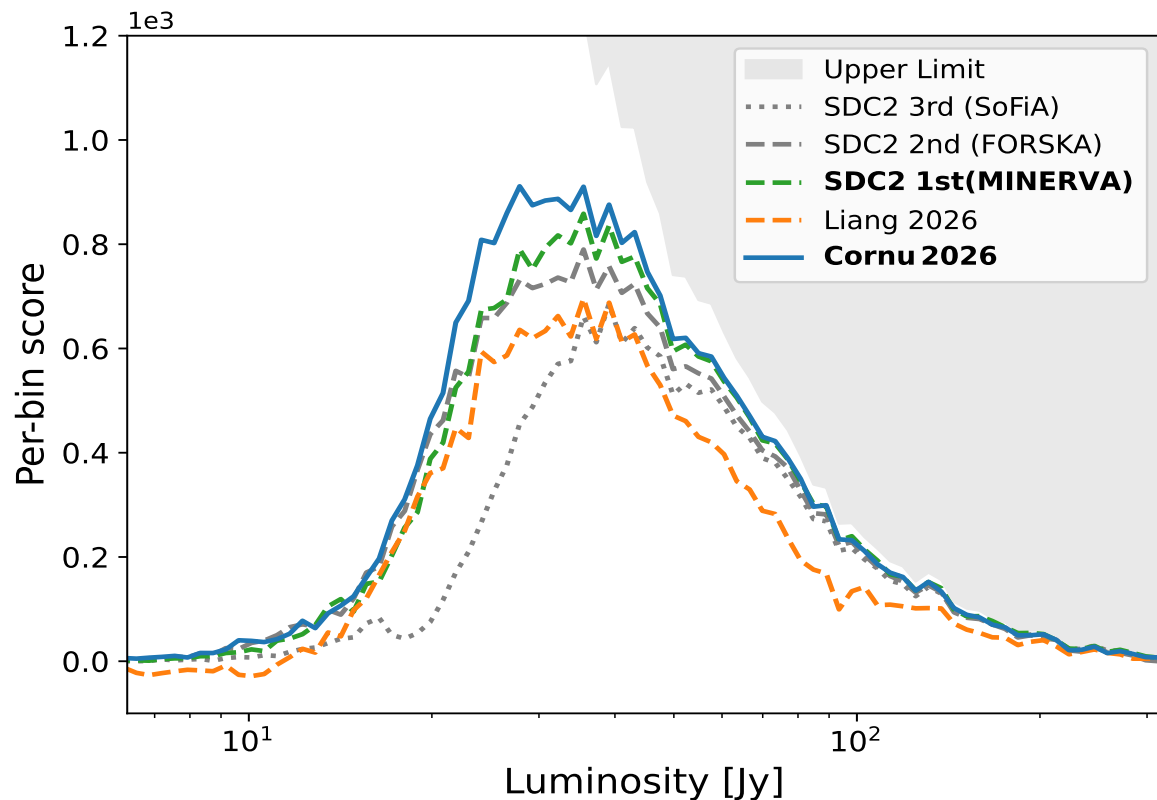
*Images are based on 40x40x120 cutouts centered on a source. Signal is averaged over the source dimension in the projected axis.*



Images are based on  $40 \times 40 \times 120$  cutouts centered on a source. Signal is averaged over the source dimension in the projected axis.

Based on **Hartley et al. 2023** + submitted catalogs

|                      | Team name           | Score        |
|----------------------|---------------------|--------------|
| <b>ML</b>            | <b>MINERVA New*</b> | <b>25453</b> |
| <hr/>                |                     |              |
| <b>ML</b>            | <b>MINERVA</b>      | <b>23254</b> |
| <b>ML + SoFiA</b>    | FORSKA-Sweden       | 22489        |
| <b>SoFiA</b>         | Team SoFiA          | 16822        |
| <b>SoFiA</b>         | NAOC-Tianlai        | 14416        |
| <b>SoFiA</b>         | HI-FRIENDS          | 13903        |
| <b>Wavelets + ML</b> | EPFL                | 8515         |
| <b>SoFiA</b>         | Spardha             | 5615         |
| <b>SoFiA</b>         | Starmech            | 2096         |
| <b>ML</b>            | JLRAT               | 1080         |
| <b>Wavelets + ML</b> | Coin                | -2           |
| <b>ML</b>            | HIRAXers            | -2           |
| <b>Other</b>         | SHAO                | -471         |

**YOLO-CIANN 3D SDC2 → Cornu et al. 2026** (Our SDC2 trained models are openly distributed)

# Generalizing to surveys from SKA precursors

Multiple groups are already at work, developing ML pipelines for several instruments

## LOFAR



## ASKAP



## MeerKAT



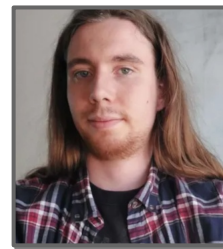
Adam Zarka  
(LUX, Obs. Paris)



One PhD on generalizing  
to **continuum surveys**  
(e.g., LoTSS, RACS)



One PhD on generalizing  
to **HI surveys**  
(e.g., LADUMA, WALLABY)



Adrien Anthore  
(ObAS)

## How to create training samples for existing surveys?

Example of calibration artifacts  
not in the SDC simulations



### Simulations

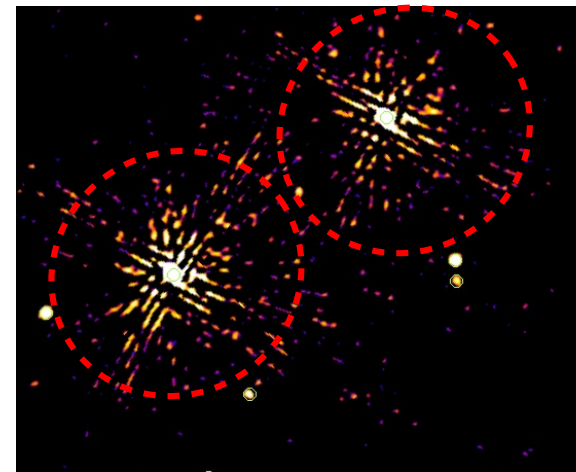
- + Perfect labels
- + Infinite sample
- Simplified representation of physical processes and Instrumental effects

VS



### Observations

- + Exhaustive data (objects, instrumental effects, processing pipeline)
- No label → external method required (inherit biases and incompleteness)



*LoTSS DR2 (LOFAR) sub field*

### Standard approach :

- Pre-training on simulated data  
(Large statistic, perfect labels)
- Fine tuning on labeled observations  
(Complete and representative)

### New approach :

- Self supervised pre-training directly on observational data
- Fine tuning on simulations or labeled data



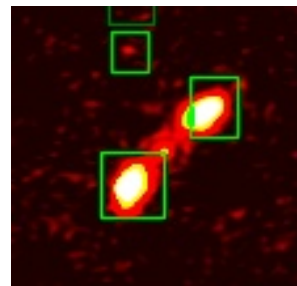
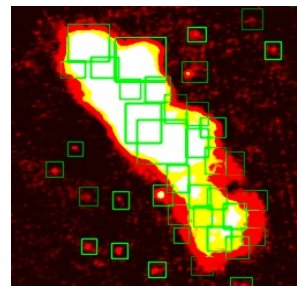
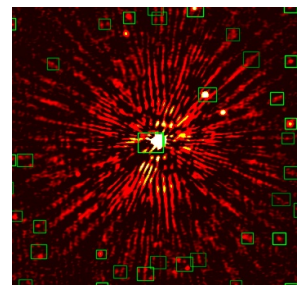
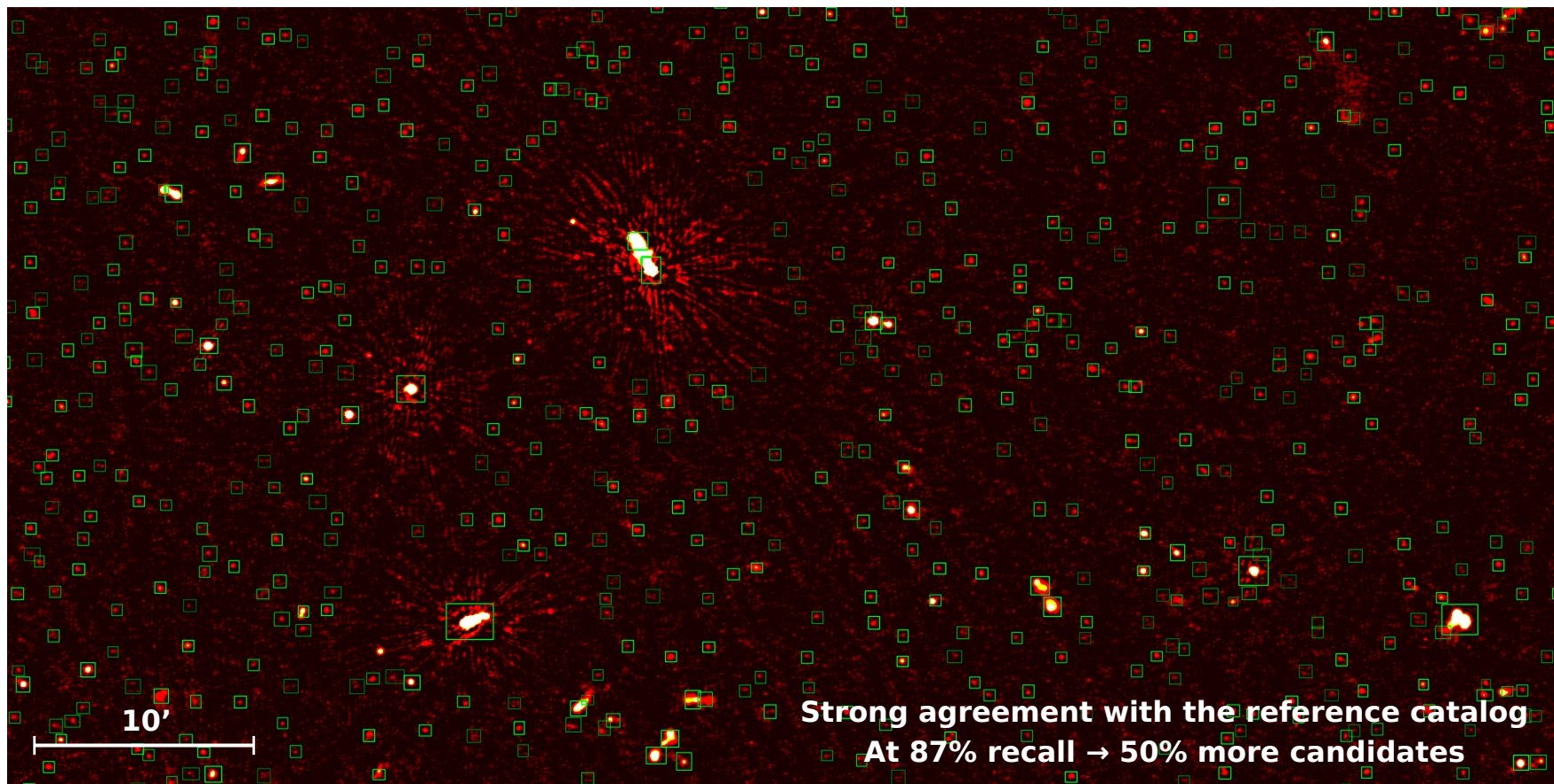
Adam Zarka  
(LUX, Obs.Paris)

## PRELIMINARY: LoTSS DR2 generalization

26

**YOLO-CIANNA trained on synthetic LoTSS-DR2 images (~144 MHz)**

→ see poster: <https://ska-meudon2026.sciencesconf.org/file/1396793>



**Strong agreement with the reference catalog**  
**At 87% recall → 50% more candidates**



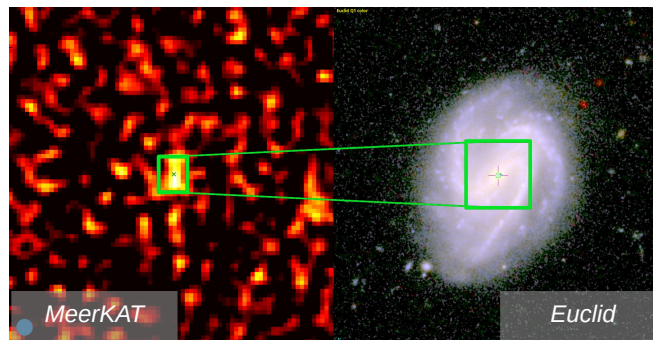
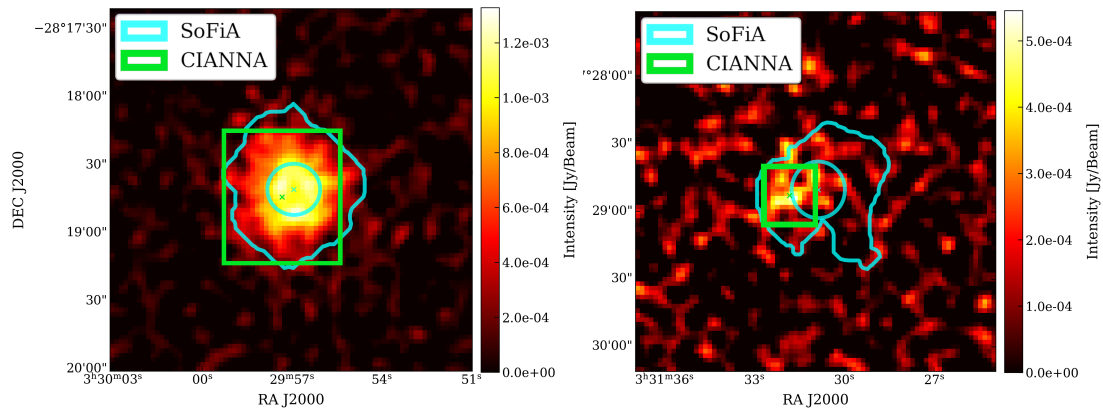
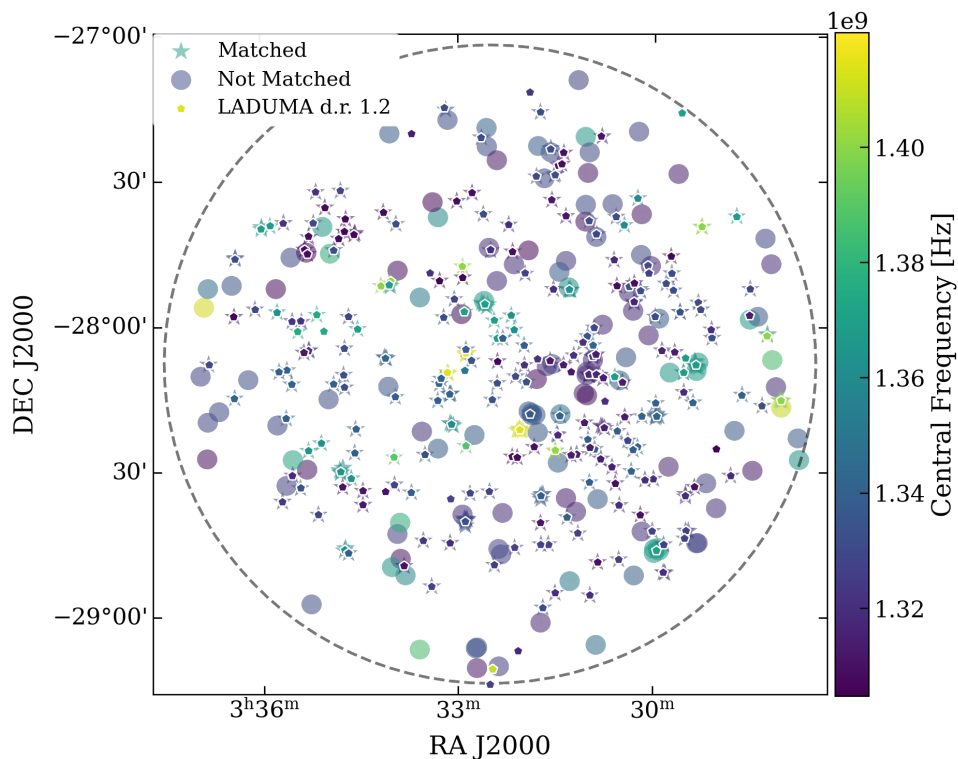
Adrien Anthore  
(ObAS)

# PRELIMINARY: LADUMA generalization

27

**SDC2 trained model applied on a LADUMA Cube (1420-1303 MHz)**

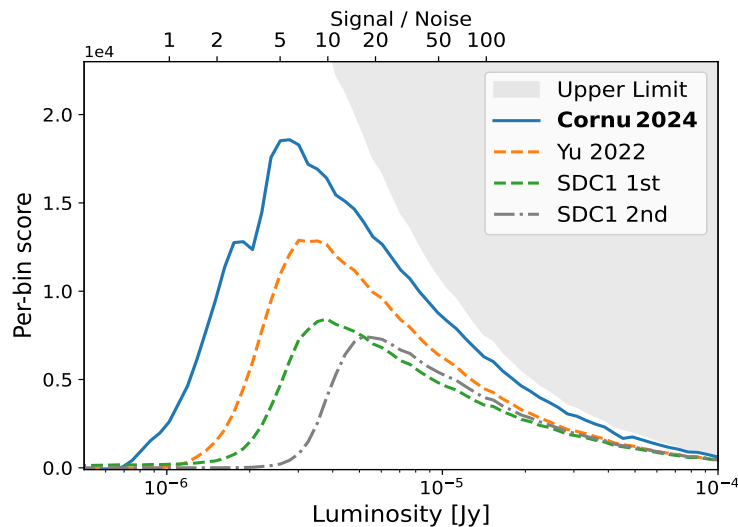
→ see poster: <https://ska-meudon2026.sciencesconf.org/file/1396768>



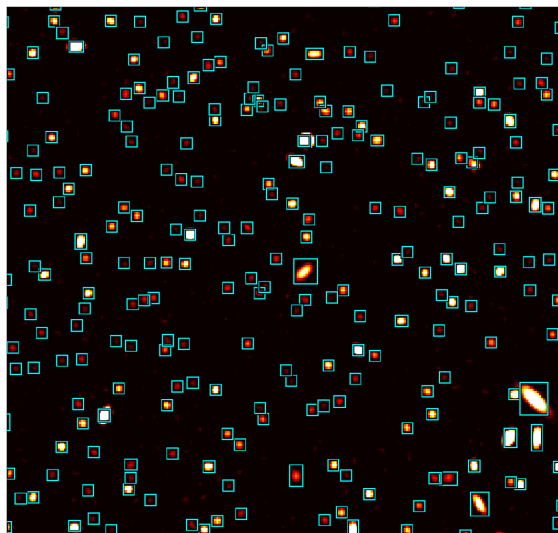
**50+ new HI  
candidates  
with matches  
(over ~300 sources)**

# Summary / Conclusion

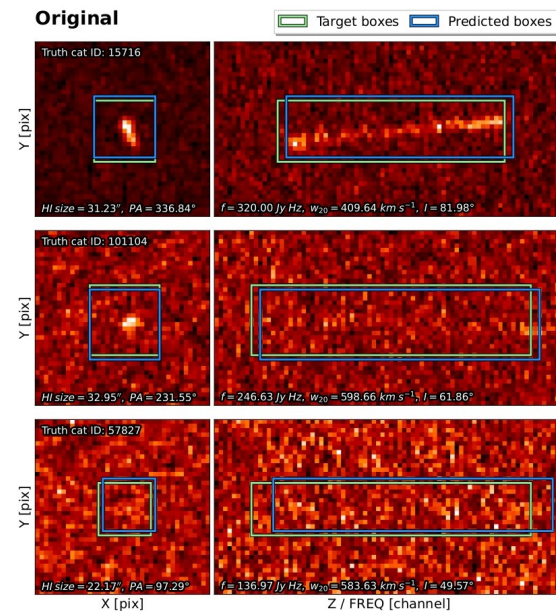
- Large modern radio surveys require advanced detection methods
- The **YOLO-CIANNA** method already achieves **SOTA results on simulated SKA data**



*SDC1 score distribution*



*SDC1 detection examples*



*SDC2 detection examples*

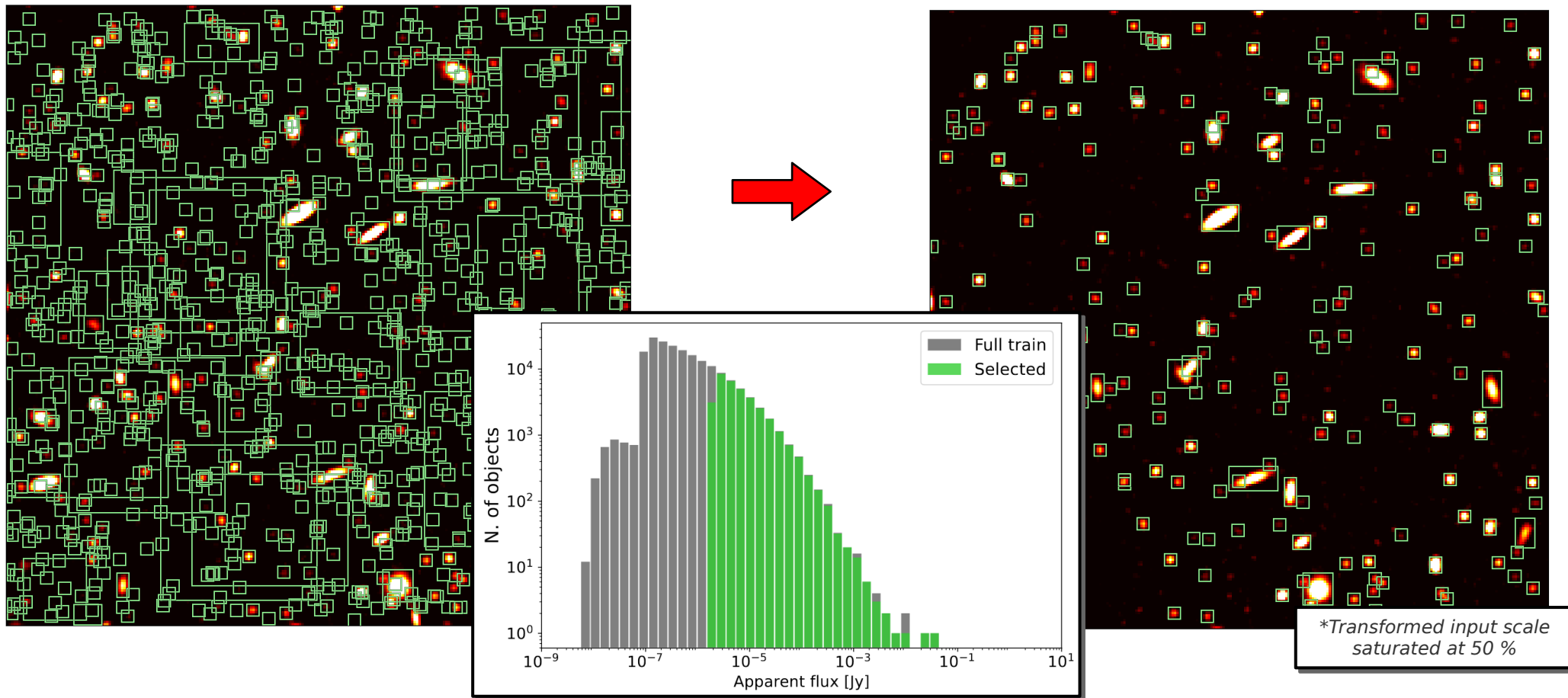
- Generalization of the method to actual observed surveys is ongoing, and results are promising
- New approaches based on **Self-supervised pre-training** are being explored to reduce the impact of labelisation biases and of the lack of realism of the observations



**Backup slides**

## SDC1 Training data selection function

- CNN must not be given the task to detect “impossible / invisible” sources !
- Selection based on surface brightness → only ~10% of the labeled catalog remains



# SDC1 Full image detection distribution

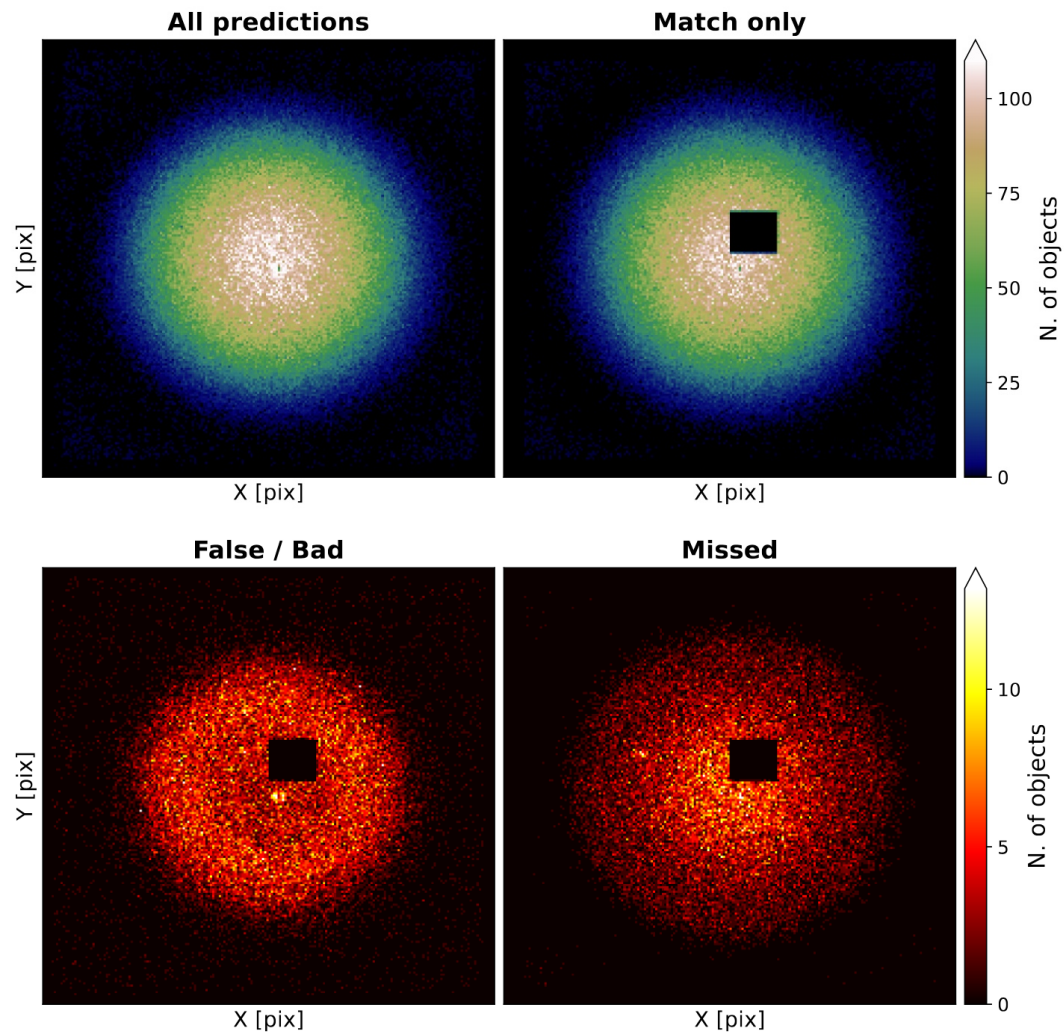
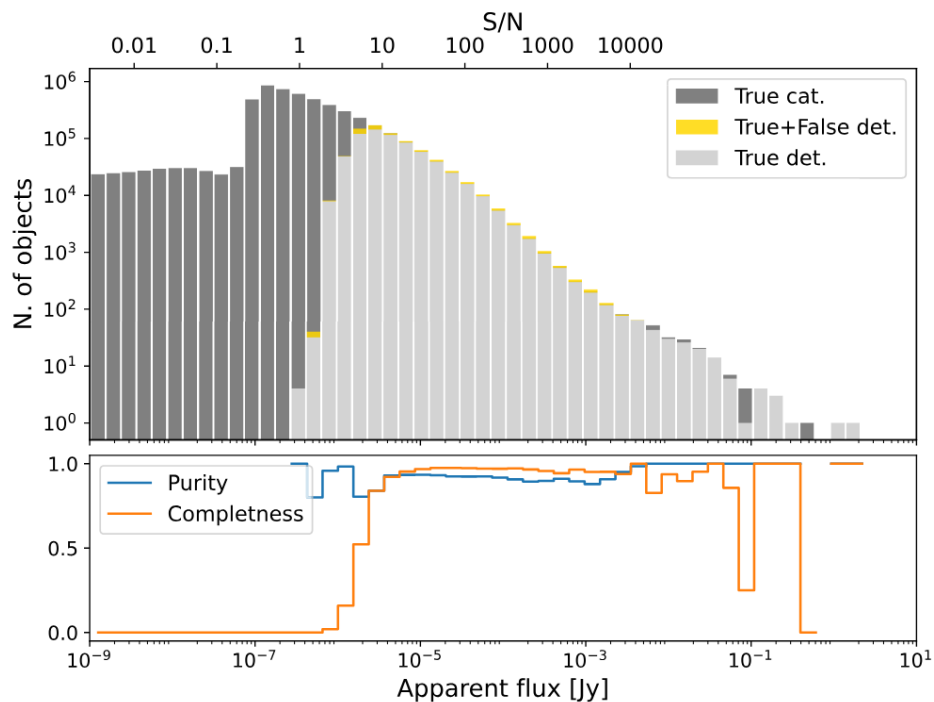
## Scorer multi-D match criteria

$$E_{tot} = \sqrt{E_{pos}^2 + E_{size}^2 + E_{flux}^2}, \text{ with}$$

$$E_{pos} \propto \sqrt{(x - \hat{x})^2 + (y - \hat{y})^2 / \hat{S}'},$$

$$E_{size} \propto |S - \hat{S}| / \hat{S}', \text{ and}$$

$$E_{flux} \propto |f - \hat{f}| / \hat{f},$$



# SDC1 Results comparison

Based on Bonaldi et al. 2021 + submitted catalogs

$$s_i = \frac{1}{7} \sum_j^7 s_i^j \quad M_s = \sum_i^{N_{match}} s_i - N_{false}$$

Average per-source score

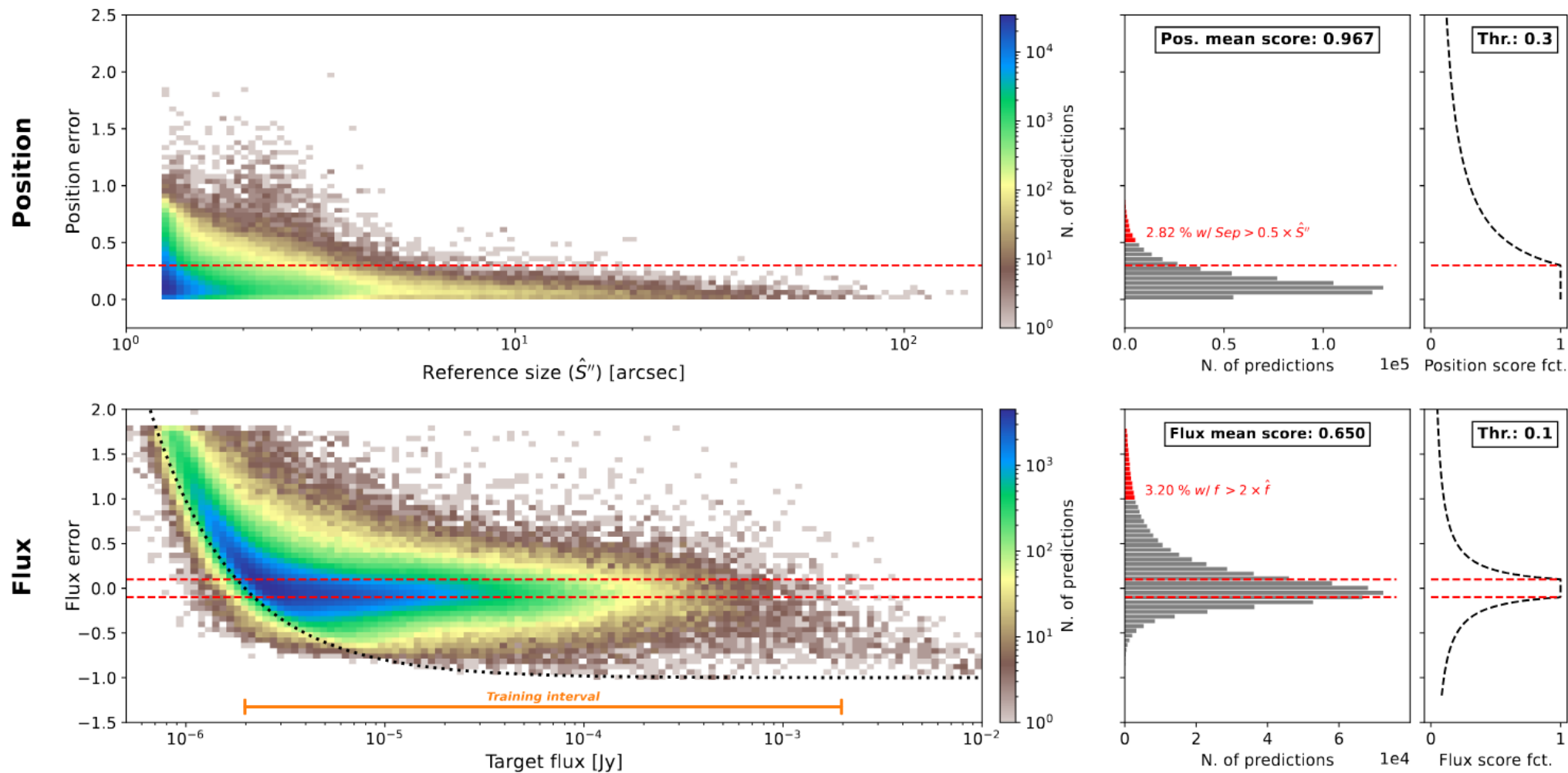
|                  | Team (method)                              | $M_s$ (Score) | $N_{det}$ | $N_{match}$ | $N_{false}$ | $N_{bad} \in N_{false}$ | Purity        | $\bar{s}$ |                          |
|------------------|--------------------------------------------|---------------|-----------|-------------|-------------|-------------------------|---------------|-----------|--------------------------|
|                  | <i>Post-challenge results</i>              |               |           |             |             |                         |               |           |                          |
| <b>ML (CNN)</b>  | MINERVA (YOLO-CIANNA)                      | <b>480450</b> | 724480    | 680000      | 44480       | 16839                   | 93.86%        | 0.7719    | } After challenge ending |
|                  | ↪ <i>purity-focus thresholds</i>           | 418434        | 541542    | 536412      | 5130        | 2506                    | <b>99.06%</b> | 0.7896    |                          |
| <b>ML (CNN)</b>  | JLRAT2 (JSFM2)                             | <b>298201</b> | 502146    | 484212      | 17934       | 2274                    | 96.43%        | 0.6529    |                          |
|                  | <i>Original challenge results</i>          |               |           |             |             |                         |               |           |                          |
| <b>Classical</b> | Engage-SKA (PROFOUND)                      | <b>200939</b> | 421992    | 418384      | 3608        | 2677                    | 99.15%        | 0.4889    | } Original leaderboard   |
| <b>Classical</b> | Shanghai (multiple methods)                | <b>158841</b> | 292646    | 291553      | 1093        | 698                     | 99.63%        | 0.5486    |                          |
| <b>ML (CNN)</b>  | ICRAR (CLARAN)                             | <b>142784</b> | 279898    | 259806      | 20092       | 6875                    | 92.82%        | 0.6269    |                          |
|                  | <b>7 other teams, one other ML attempt</b> | ...           | ...       | ...         | ...         | ...                     | ...           | ...       |                          |

→ SDC1 is still a very interesting dataset for source detection pipeline development !

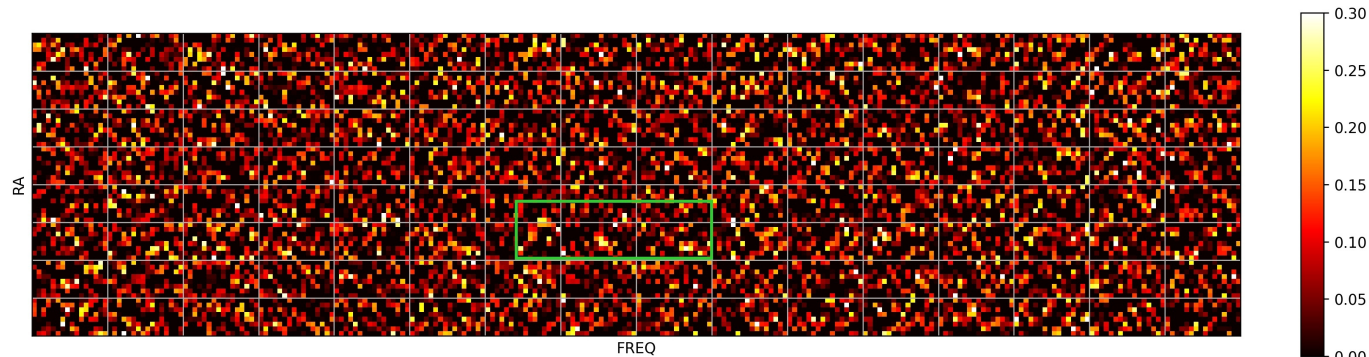
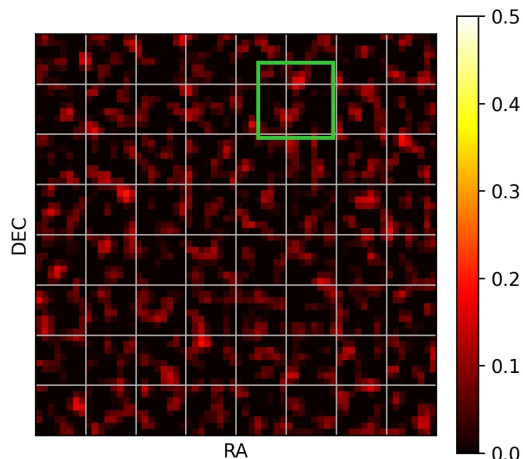
**YOLO-CIANNA SDC1 → Cornu et al. 2024**

## SDC1 Source characterization

$$s_i^j = \min\left(1, \frac{T^j}{E_i^j}\right), \quad s_i = \frac{1}{7} \sum_j s_i^j$$



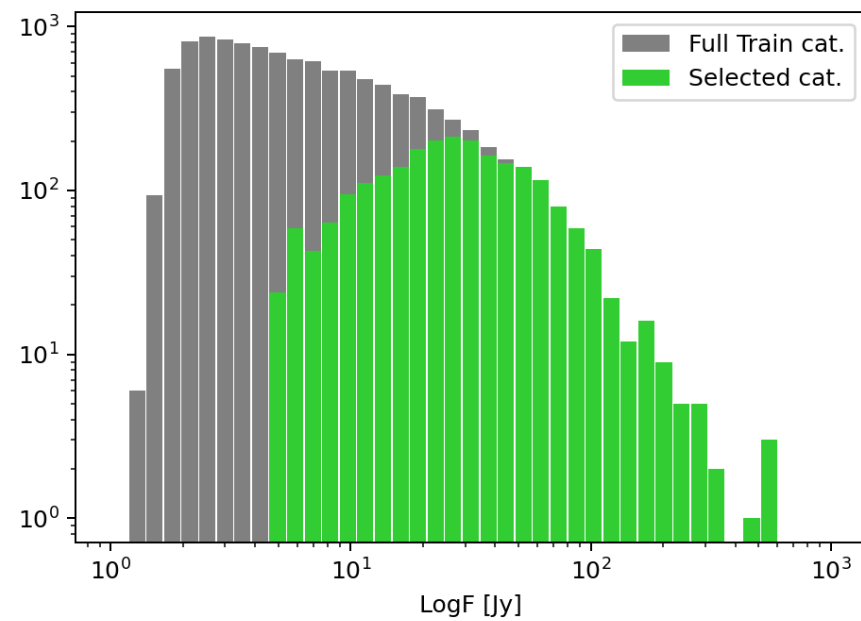
# SDC2 Selection function difficulties



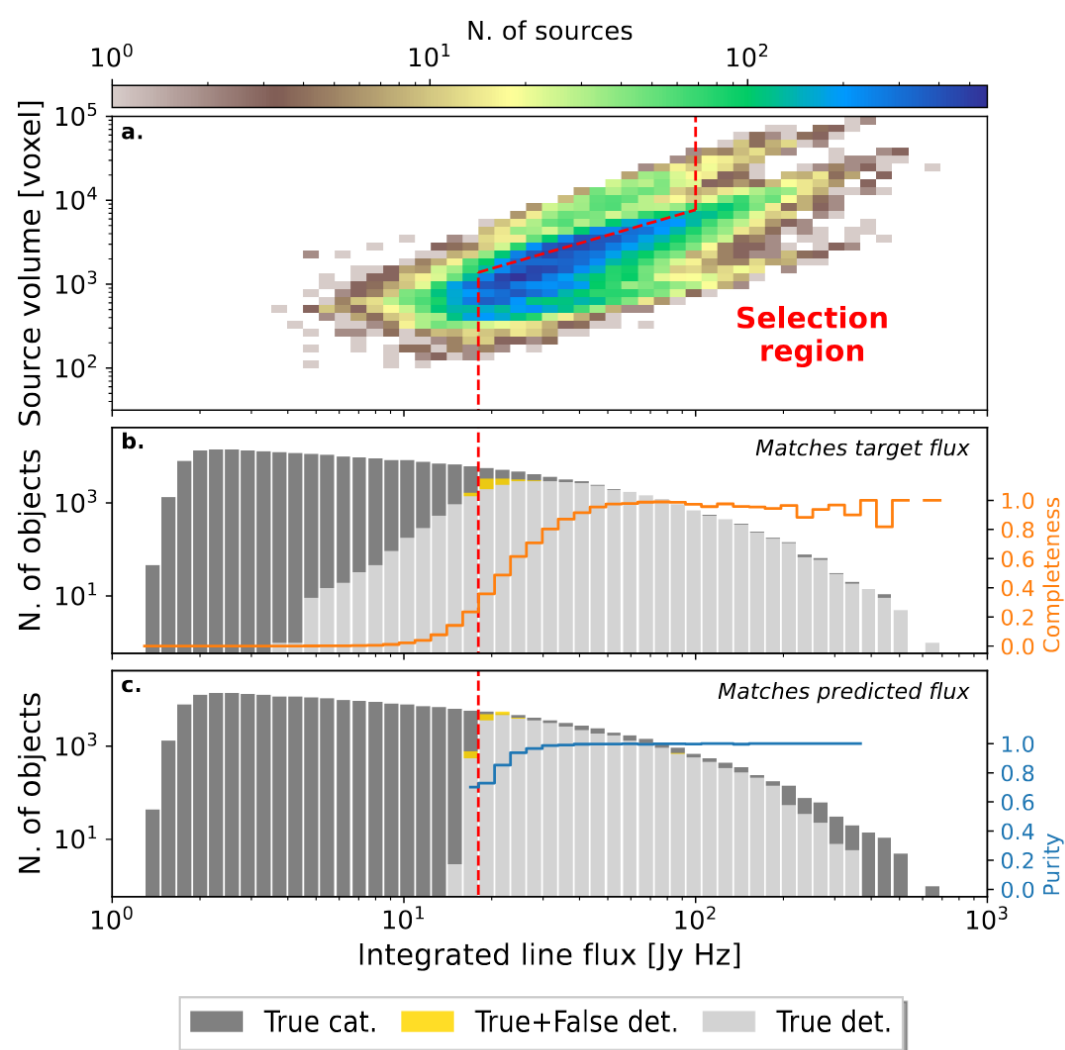
**Selection function based on brightness or SNR are not sufficient to fully represent the noisy 3D information.**

→ **“Bootstrap learning”** (~active learning):

After a first training, **un-selected true sources** with **high predicted objectness** can be re-injected in the training sample.



## SDC2 Full cube matches and statistics



$$D_{\text{tot}} = \sqrt{D_{\text{pos}}^2 + D_{\text{freq}}^2 + D_{\text{flux}}^2 + D_{\text{size}}^2 + D_{\text{width}}^2}, \text{ with}$$

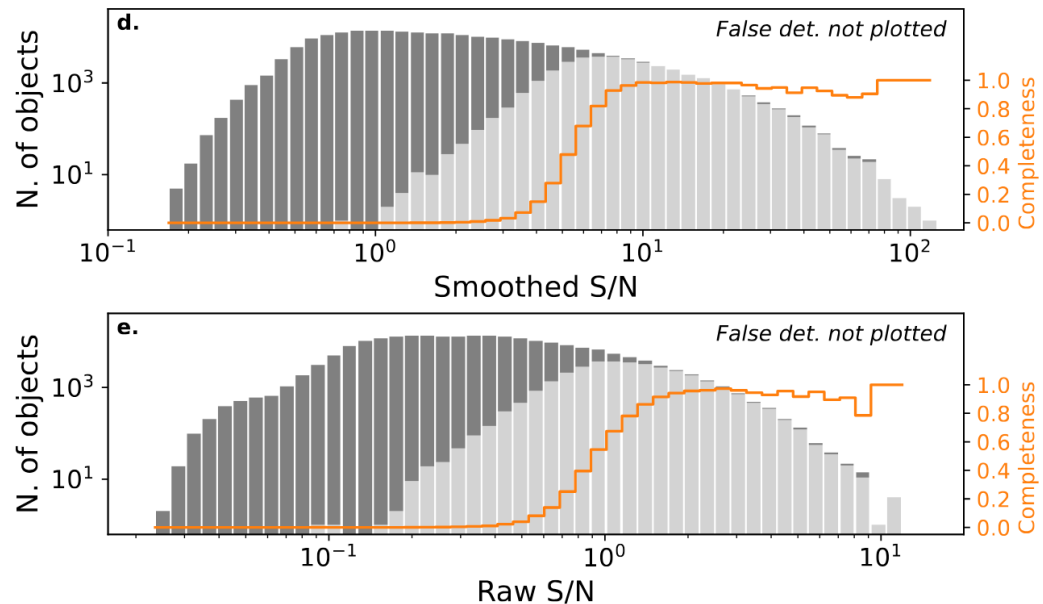
$$D_{\text{pos}} = \sqrt{(x - \hat{x})^2 + (y - \hat{y})^2 / \hat{S}'},$$

$$D_{\text{freq}} = |\nu - \hat{\nu}| / \hat{w}_{20, \text{Hz}},$$

$$D_{\text{flux}} = |f - \hat{f}| / \hat{f},$$

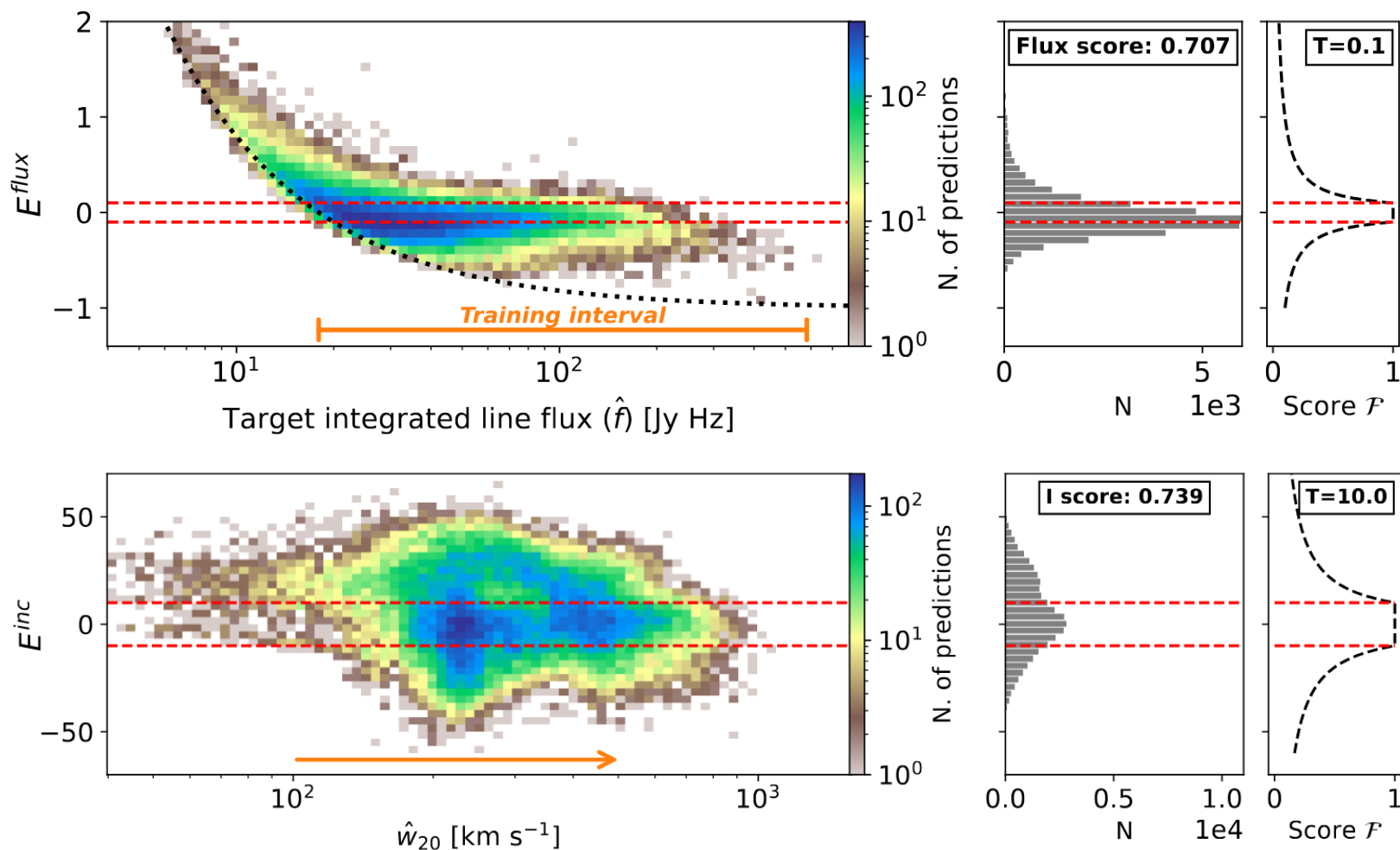
$$D_{\text{size}} = |S - \hat{S}| / \hat{S}', \text{ and}$$

$$D_{\text{width}} = |w_{20} - \hat{w}_{20}| / \hat{w}_{20},$$



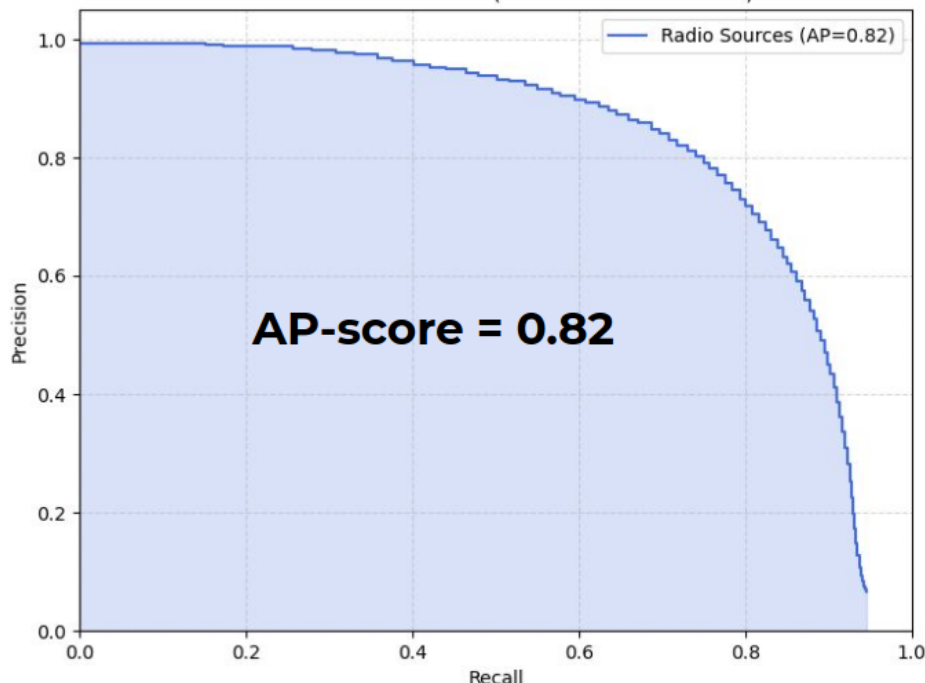
## SDC2 Source characterization

$$s_i^j = \min\left(1, \frac{T^j}{E_i^j}\right), \quad s_i = \frac{1}{7} \sum_j s_i^j$$





Precision-Recall Curve (YOLO vs Shimwell Cat)



## Comparison with the **Shimwell+2022 raw radio component catalog of LoTSS-DR2**

- Catalog computed using **PyBDSF** (Mohan & Rafferty 2015)
  - Refined through source association/deblending, and **cross-identification** with optical/infrared
- ~90% precision at a recall of 60%

However, the precision is mostly impacted by our detector retrieving ~1.5 times more sources per sky surface unit, performing remarkably for low SNR sources. It is also mainly impacted by false detections in ionospheric artefacts.

### Conclusion :

- Already robust pipeline for source detection
- Very good agreement with automatic detection methods for “easy detections”
- Shows a higher performances in the management of “harder detections” (low SNR sources, artifacts management, various source morphologies...)