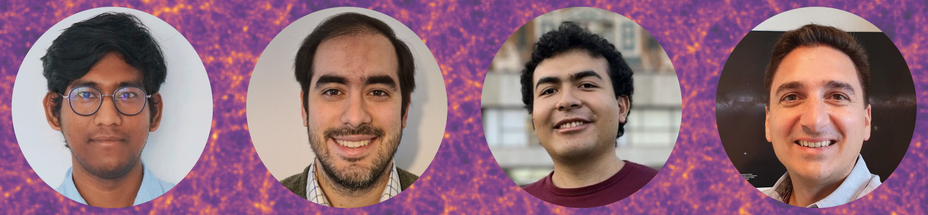
Application of Machine Learning to the Identification of Dark Matter Haloes

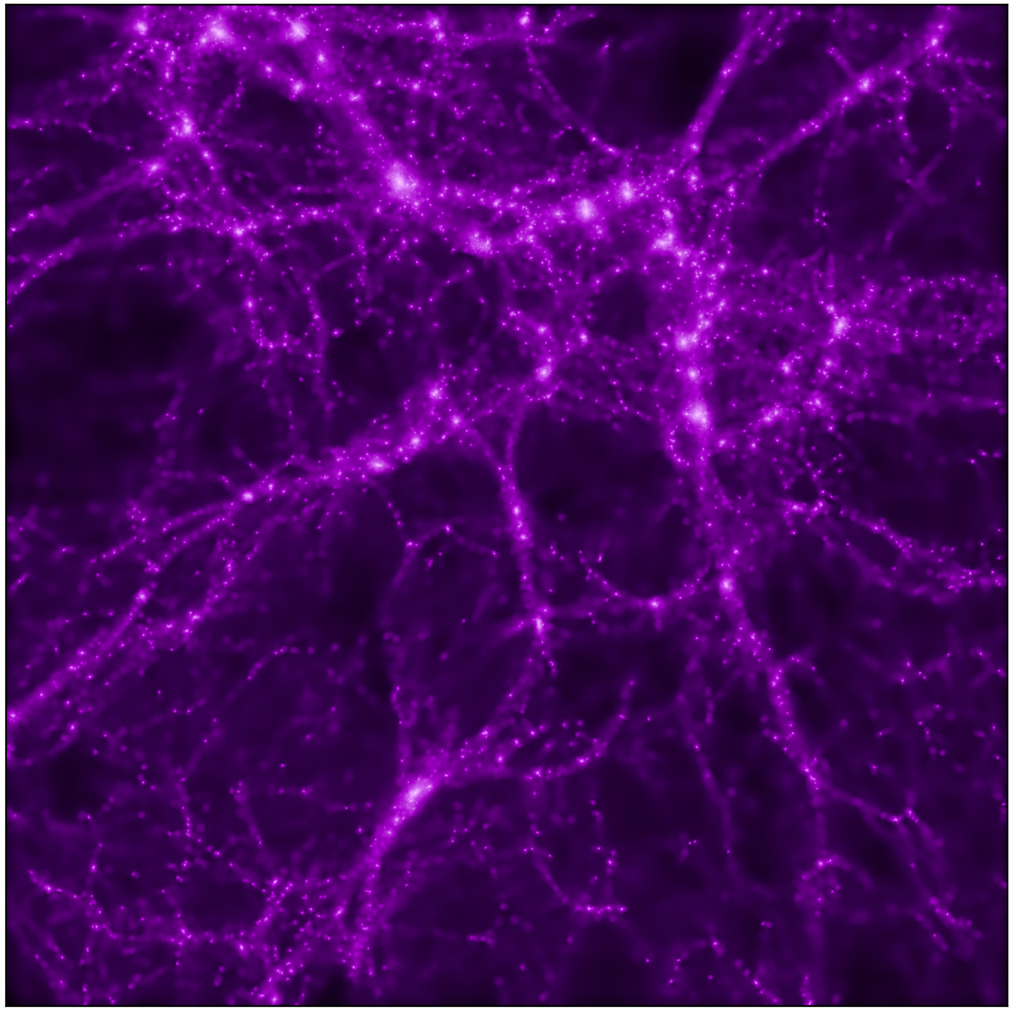
Soumadeep Maiti

Carlos Correa, Alejandro Pérez Fernández, Ariel Sánchez
MPE, OPINAS



Quantum Universe Attract.Workshop, 2025

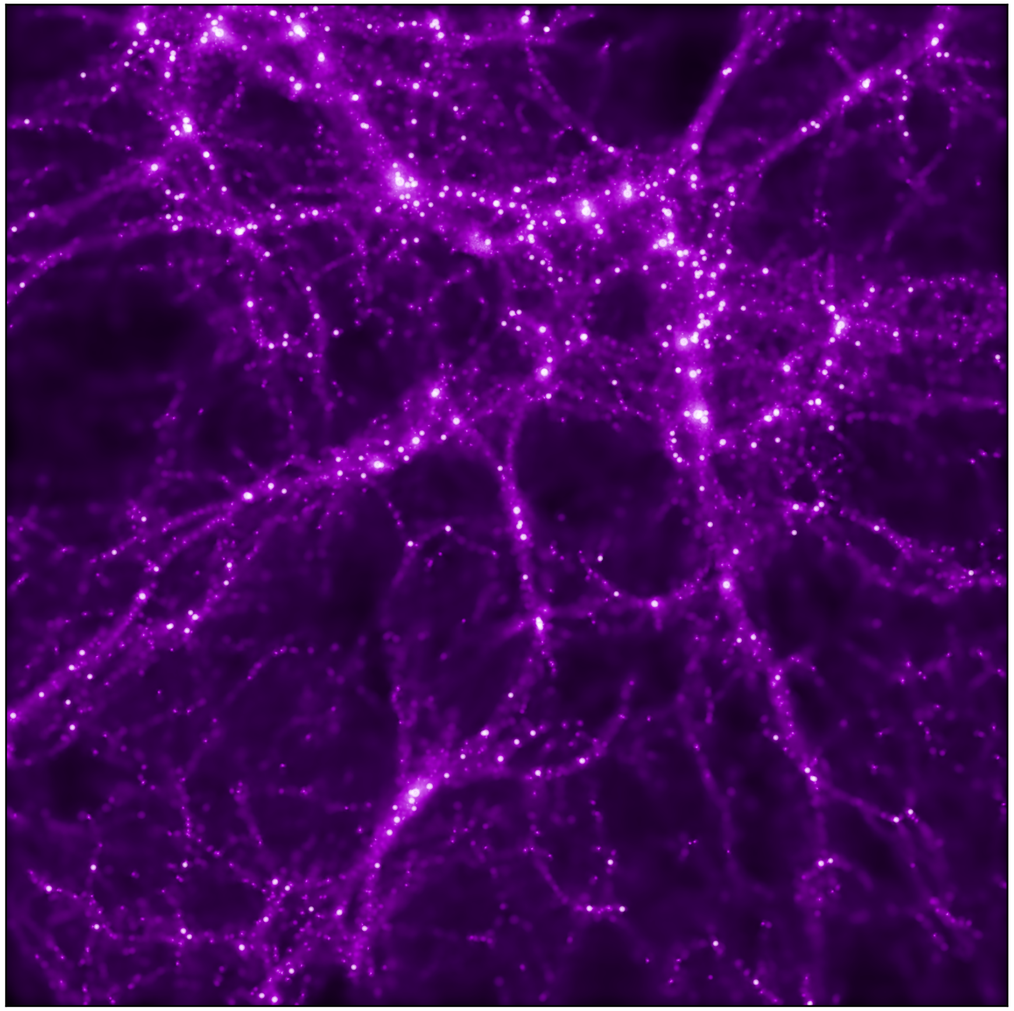
N-body simulations



Credits: Carlos Correa & Andrea Fiorilli

- As time progresses (redshift z decreases), structures collapse, and the **cosmic web** becomes more defined
- Theoretical frameworks are difficult to model at highly **non-linear scales**
- Theoretical approach: simulation-based inference (**SBI**)

N-body simulations



Credits: Carlos Correa & Andrea Fiorilli

- As time progresses (redshift z decreases), structures collapse, and the **cosmic web** becomes more defined
- Theoretical frameworks are difficult to model at highly **non-linear scales**
- Theoretical approach: simulation-based inference (**SBI**)
- **Dark matter haloes** appears at the highest peaks in the full density field

Why do we need a faster halo identification method?

- Current Halo Finders: **FOF** (Friends-of-Friends, Davis et al. 1985), **ROCKSTAR**(Robust Overdensity Calculation using K-Space Topologically Adaptive Refinement, Behroozi et al. 2013), etc.
- Increasing demand for **fast** and **mass** production of halo catalogue without sacrificing accuracy
- Halo identification in bigger simulations are computationally expensive

Halo Finders

↓

Time

↓

Machine Learning Algorithm

←

Particle Classifier (CNN)

+

Clustering Algorithm (FoF)

←

Dark Matter Haloes

Why do we need a faster halo identification method?

- Current Halo Finders: **FOF** (Friends-of-Friends, Davis et al. 1985), **ROCKSTAR**(Robust Overdensity Calculation using K-Space Topologically Adaptive Refinement, Behroozi et al. 2013), etc.
- Increasing demand for **fast** and **mass** production of halo catalogue without sacrificing accuracy
- Halo identification in bigger simulations are computationally expensive

Halo Finders

↓

Time ?

↓

Machine Learning Algorithm

← Particle Classifier (CNN)

+

Clustering Algorithm (FoF)

← Dark Matter Haloes

Why do we need a faster halo identification method?

- Current Halo Finders: **FOF** (Friends-of-Friends, Davis et al. 1985), **ROCKSTAR**(Robust Overdensity Calculation using K-Space Topologically Adaptive Refinement, Behroozi et al. 2013), etc.

- Increasing demand for **fast** and **mass** production of halo catalogue without sacrificing accuracy

- Halo identification in bigger simulations are computationally expensive

Halo Finders

↓

Time ?

↓

Machine Learning Algorithm

← Particle Classifier (CNN)

+

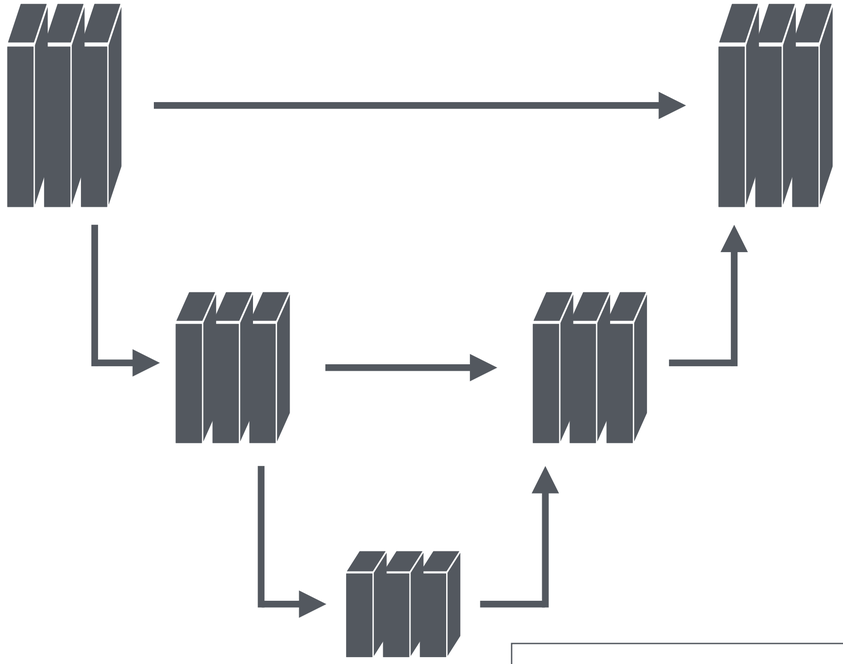
Clustering Algorithm (FoF)

← Dark Matter Haloes

Convolutional Neural Networks

An example: Image segmentation (V-Net)

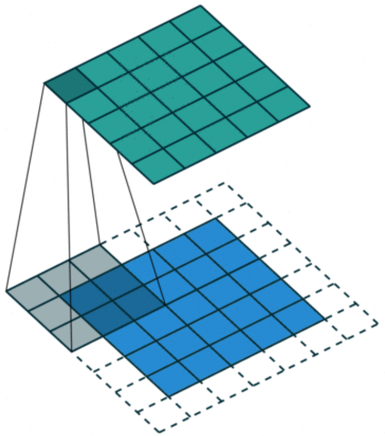
Input channels:
RGB



Output channels:
#classes

Target: true mask

Loss: cross entropy



concatenation

downsampling

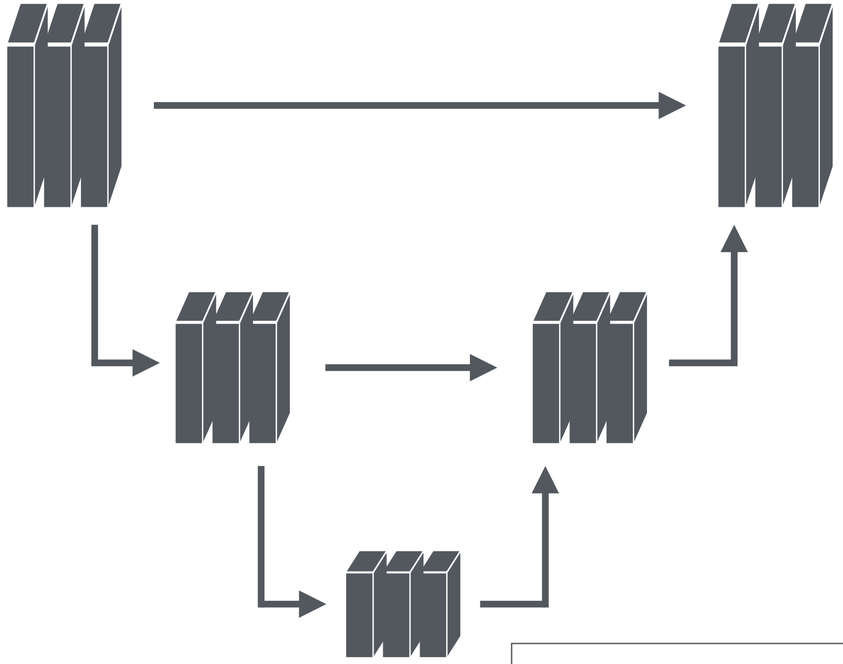
upsampling

- State-of-the-art architecture
- Applications: Medical imaging, video surveillance, pedestrian detection, …

Convolutional Neural Networks

An example: Image segmentation (V-Net)

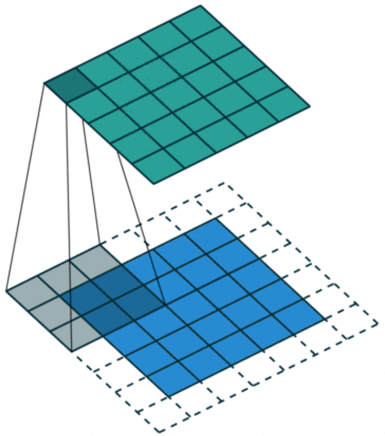
Input channels:
RGB



Output channels:
#classes

Target: true mask

Loss: cross entropy



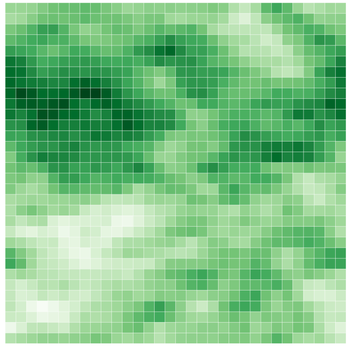
concatenation

downsampling

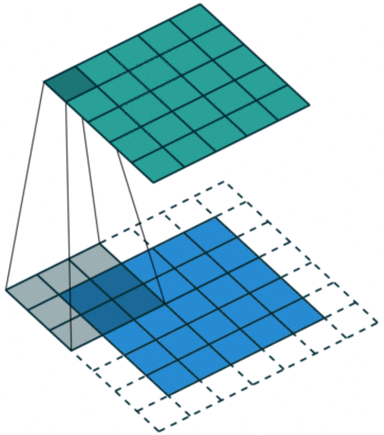
upsampling

- State-of-the-art architecture
- Applications: Medical imaging, video surveillance, pedestrian detection, …

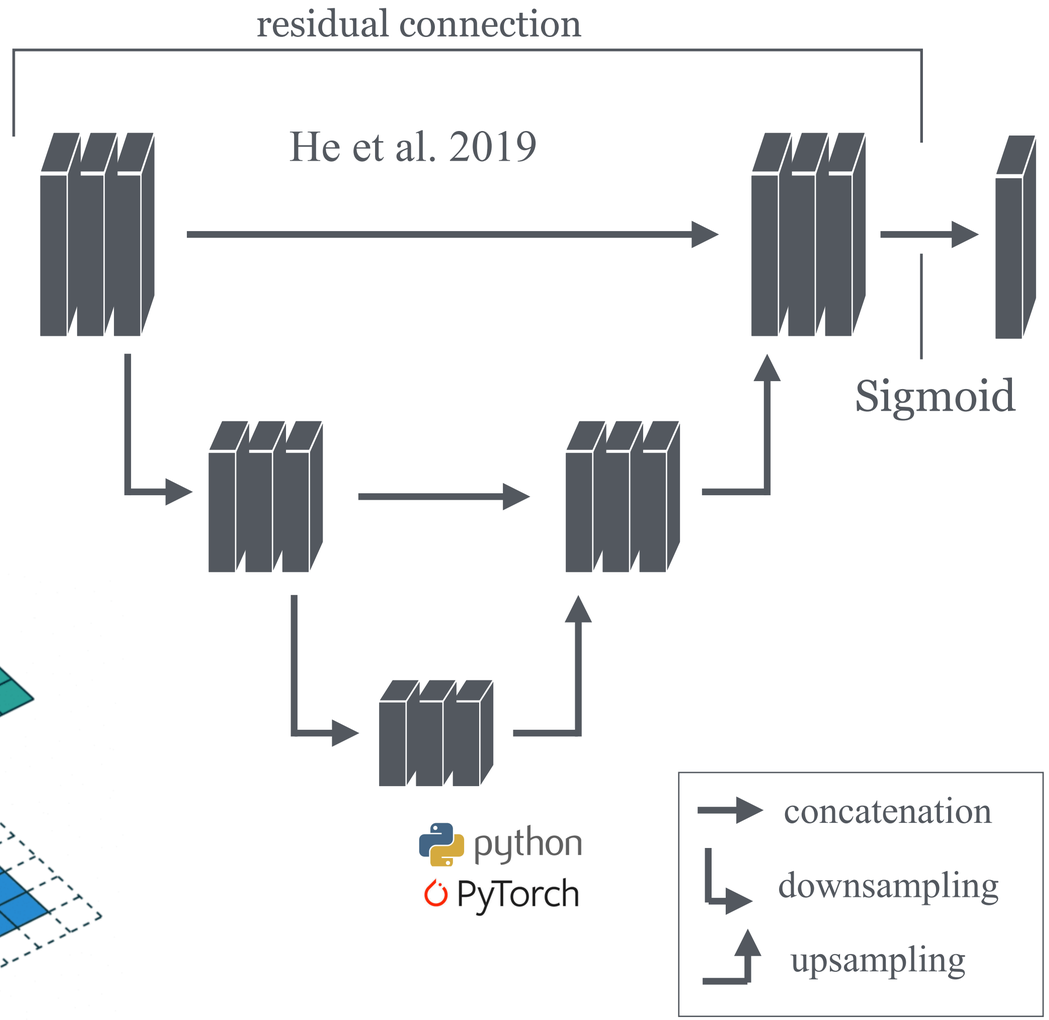
Learning the Dark Matter Particles



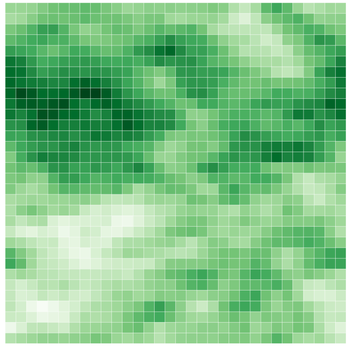
Input channels:
ψ_i, $\mathbf{v}_i$



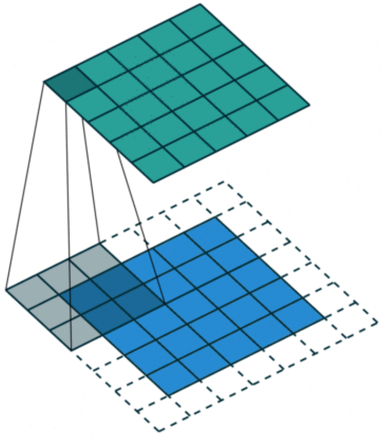
2D convolution



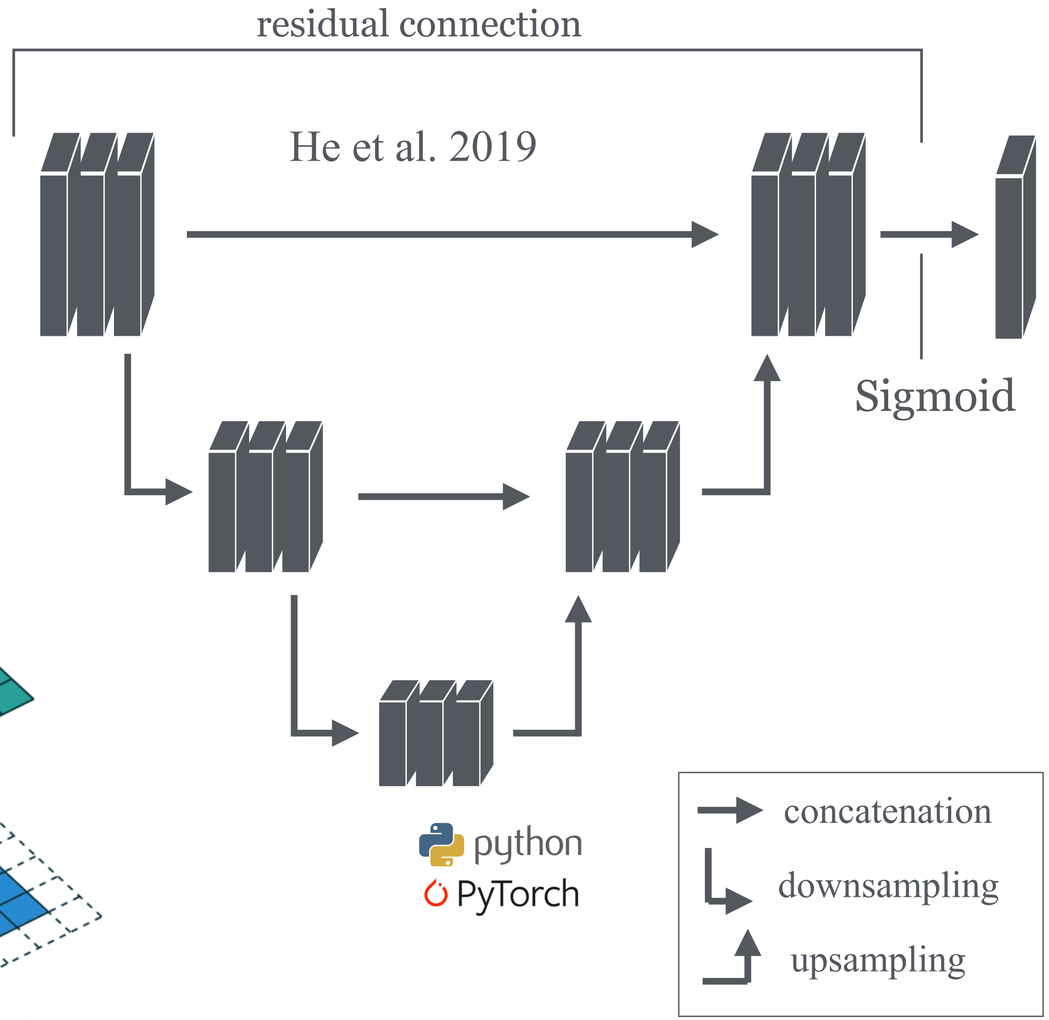
Learning the Dark Matter Particles



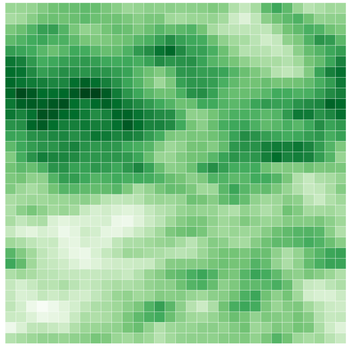
Input channels:
ψ_i, $\mathbf{v}_i$



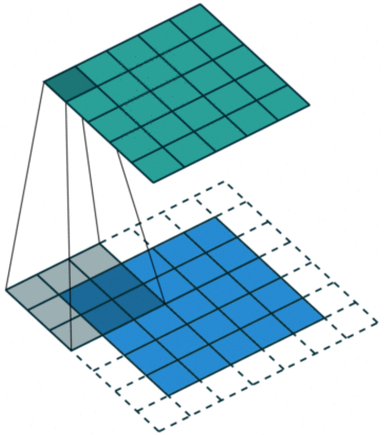
2D convolution



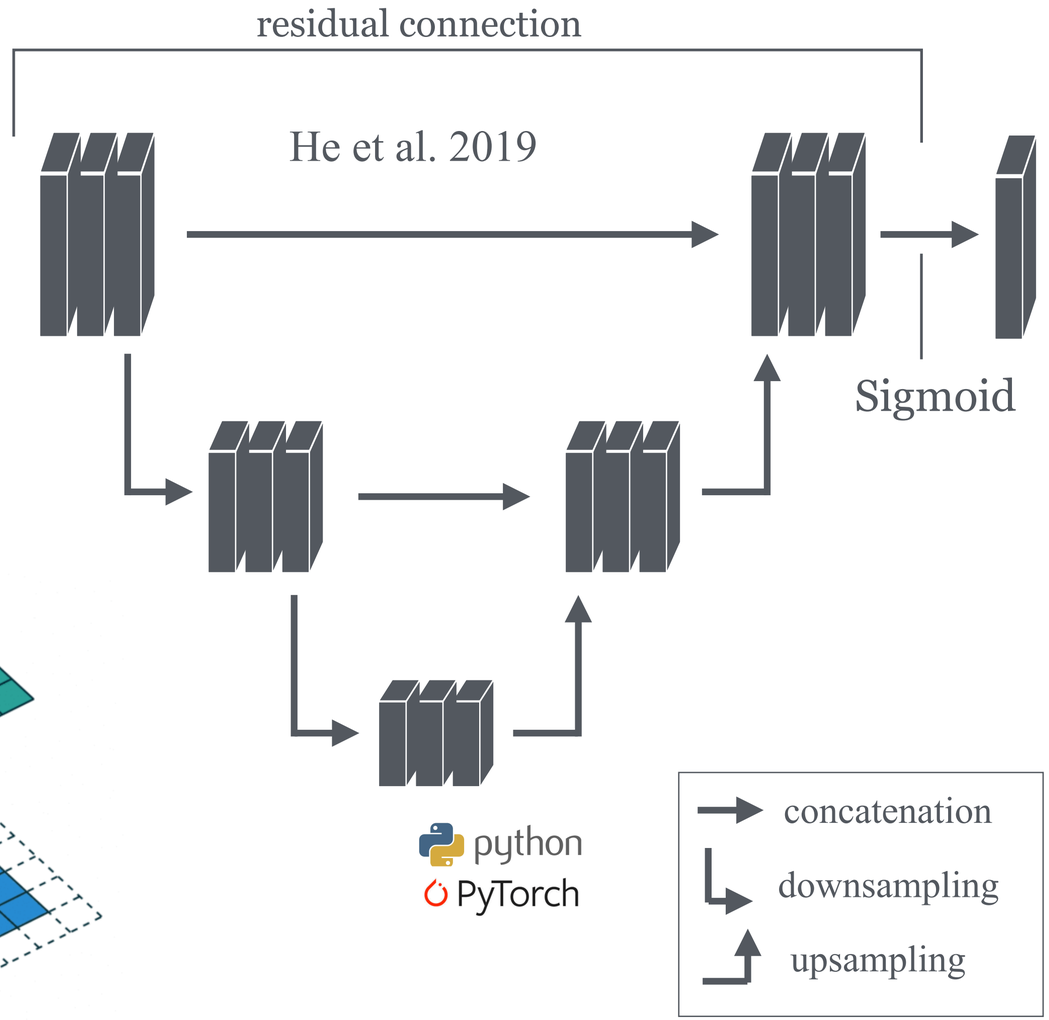
Learning the Dark Matter Particles



Input channels:
ψ_i, $\mathbf{v}_i$



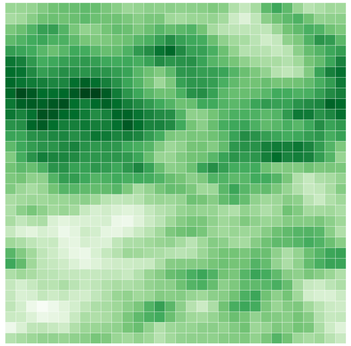
2D convolution



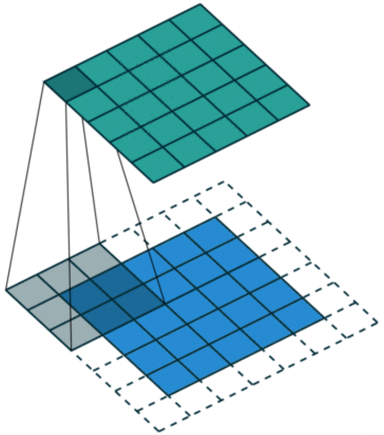
0	0	0	0	0	0	0	0	0	1
1	0	1	0	0	1	0	0	0	1
0	0	0	0	1	0	1	0	0	1
0	0	0	1	0	0	0	1	1	0
1	0	0	0	0	0	1	1	1	0
0	0	1	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1
0	1	0	0	0	1	0	0	0	1
1	0	0	1	0	0	0	1	0	0
0	1	0	0	0	0	0	0	0	0

Output channels:
1, 0

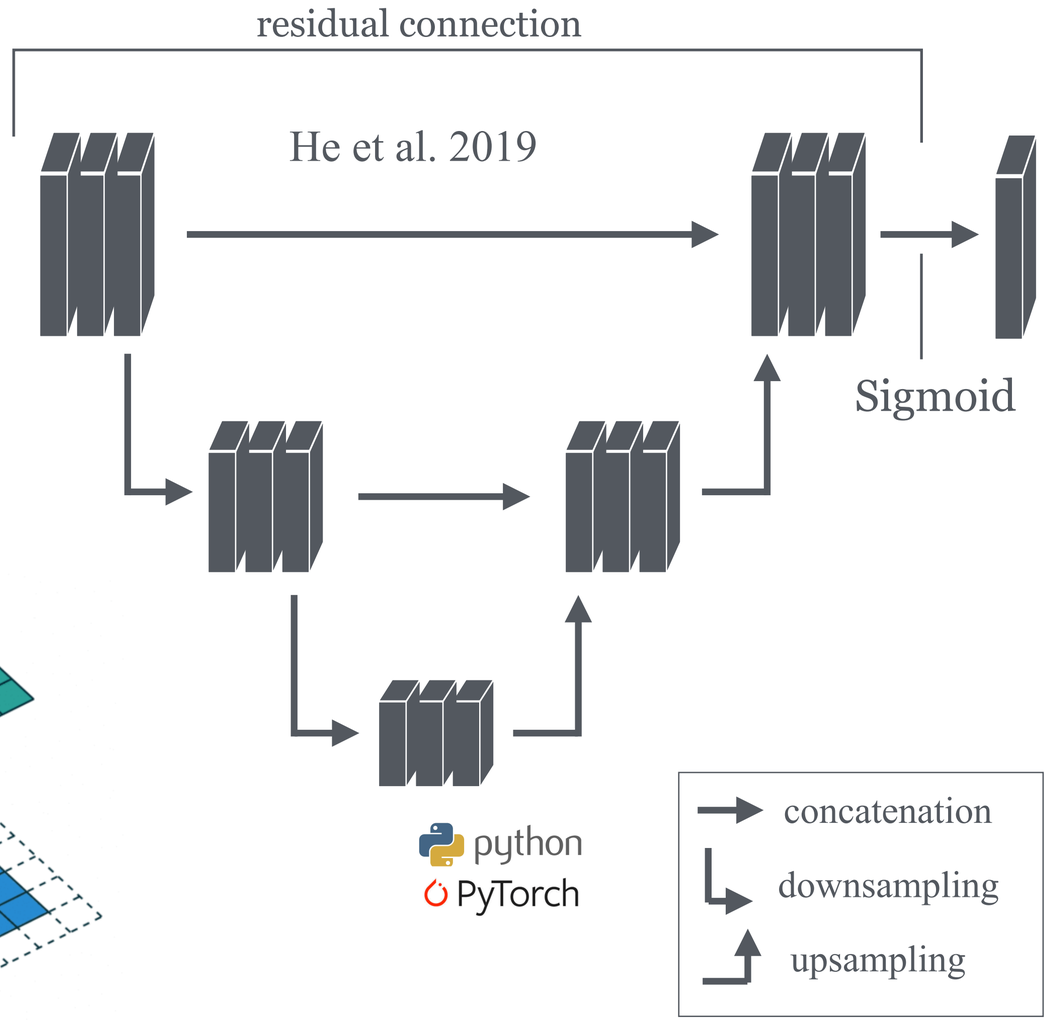
Learning the Dark Matter Particles



Input channels:
ψ_i, $\mathbf{v}_i$



2D convolution



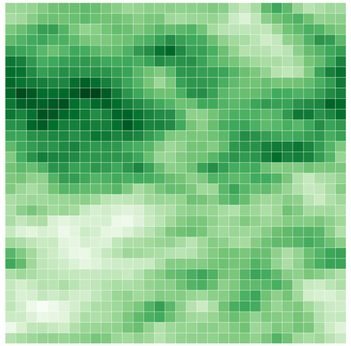
0	0	0	0	0	0	0	0	0	1
1	0	1	0	0	1	0	0	0	1
0	0	0	0	1	0	1	0	0	1
0	0	0	1	0	0	0	1	1	0
1	0	0	0	0	0	1	1	1	0
0	0	1	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1
0	1	0	0	0	1	0	0	0	1
1	0	0	1	0	0	0	1	0	0
0	1	0	0	0	0	0	0	0	0

Output channels:
1, 0

0	0	1	0	0	0	0	0	0	0
1	0	1	0	1	1	0	0	0	1
0	0	0	0	1	0	0	0	0	1
0	0	0	1	0	0	0	1	1	0
1	0	0	0	0	0	1	1	1	0
0	0	1	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1
0	1	0	0	0	1	0	0	0	1
1	0	0	1	0	0	0	1	0	0
0	1	0	0	0	0	0	0	0	0

Target: Rockstar Catalogue

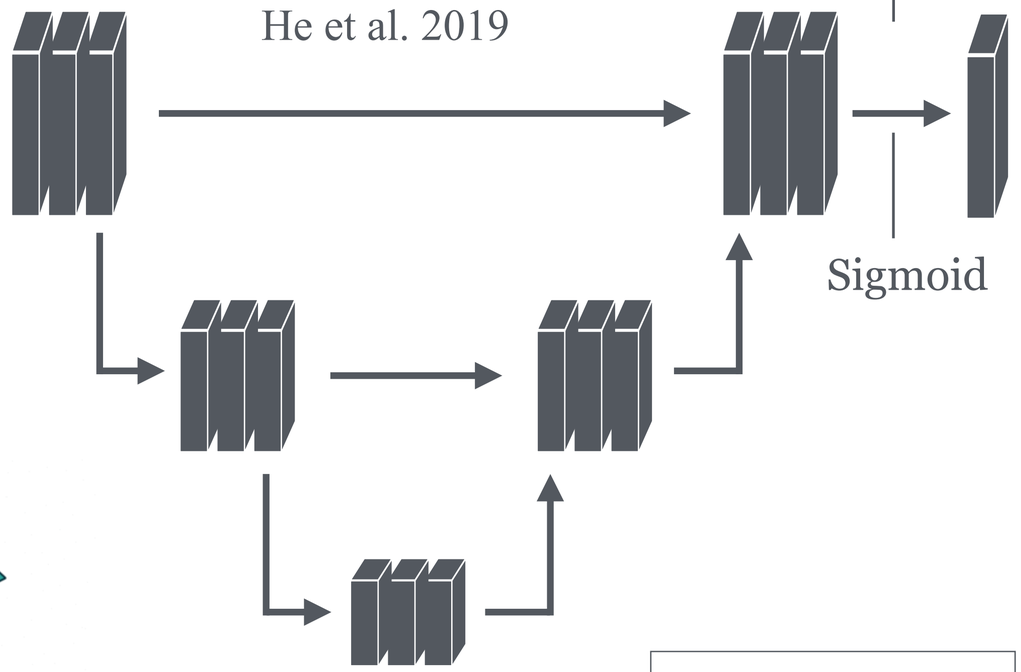
Learning the Dark Matter Particles



Input channels:
ψ_i, $\mathbf{v_i}$

residual connection

He et al. 2019



Sigmoid

python
PyTorch

concatenation
downsampling
upsampling

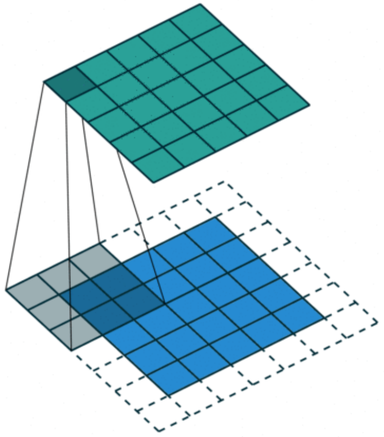
0	0	0	0	0	0	0	0	0	1
1	0	1	0	0	1	0	0	0	1
0	0	0	0	1	0	1	0	0	1
0	0	0	1	0	0	0	1	1	0
1	0	0	0	0	0	1	1	1	0
0	0	1	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1
0	1	0	0	0	1	0	0	0	1
1	0	0	1	0	0	0	1	0	0
0	1	0	0	0	0	0	0	0	0

Output channels:
1, 0

0	0	1	0	0	0	0	0	0	0
1	0	1	0	1	1	0	0	0	1
0	0	0	0	1	0	0	0	0	1
0	0	0	1	0	0	0	1	1	0
1	0	0	0	0	0	1	1	1	0
0	0	1	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1
0	1	0	0	0	1	0	0	0	1
1	0	0	1	0	0	0	1	0	0
0	1	0	0	0	0	0	0	0	0

Target: Rockstar Catalogue

Loss: Binary Cross Entropy



2D convolution

Building Dataset

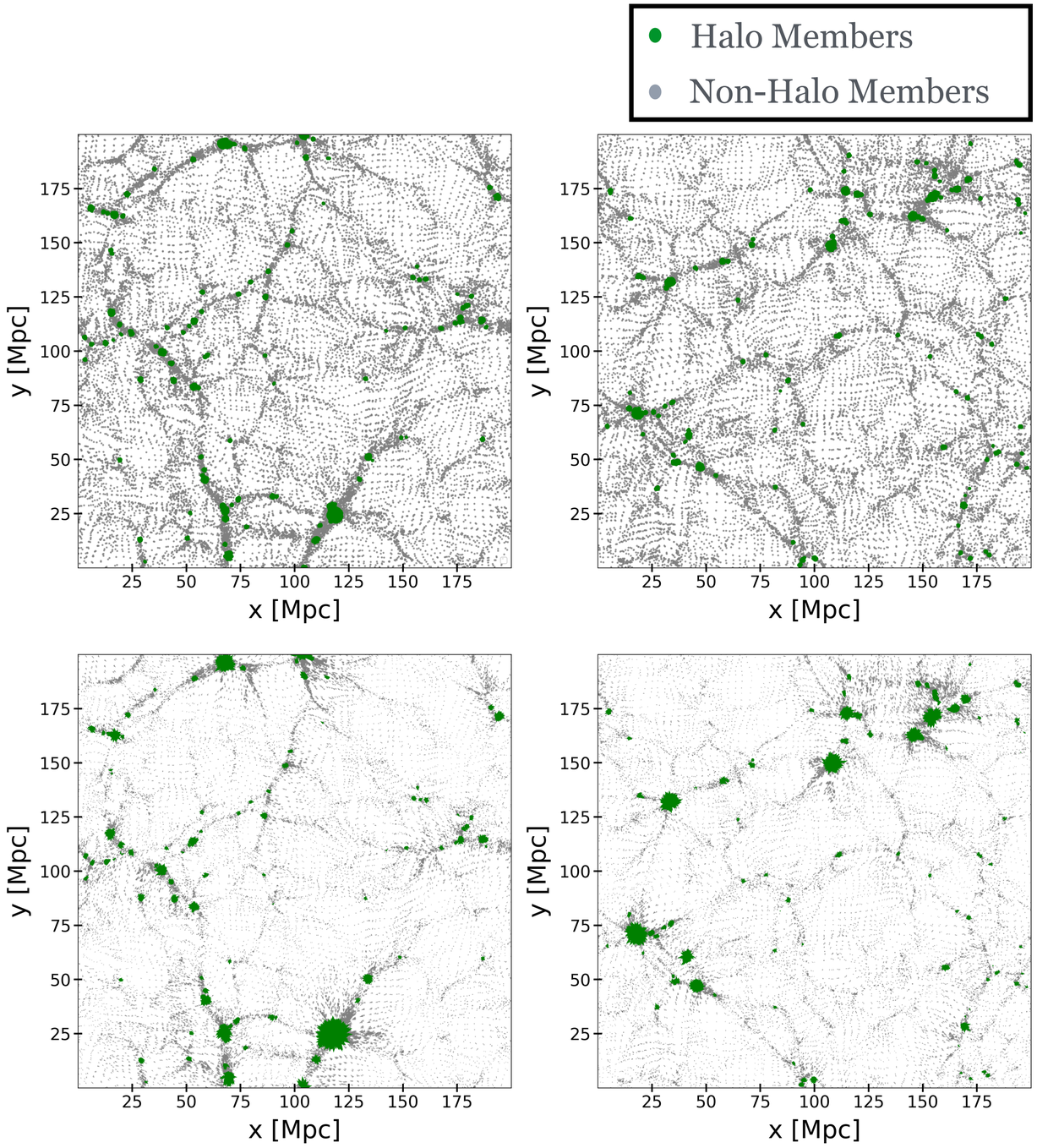
- Cosmology: Planck cosmology
- Simulation: **Set ref** using Gadget4 (Springel et al. 2021)
- Apply **Rockstar** on the Snapshots
- Label DM particles $(\mathbf{1}, \mathbf{0})$ based on the Halo Catalogue
- 500 realisations: Input($\overrightarrow{\psi_{\mathbf{i}}}, \overrightarrow{\mathbf{v}_{\mathbf{i}}}$)

Building Dataset

- Cosmology: Planck cosmology
- Simulation: **Set ref** using Gadget4 (Springel et al. 2021)
- Apply **Rockstar** on the Snapshots
- Label DM particles $(\mathbf{1}, \mathbf{0})$ based on the Halo Catalogue
- 500 realisations: Input($\vec{\psi_\mathbf{i}}$, $\vec{\mathbf{v}_\mathbf{i}}$)
  - Training: 350
  - Validation: 50
  - Testing: 100

Building Dataset

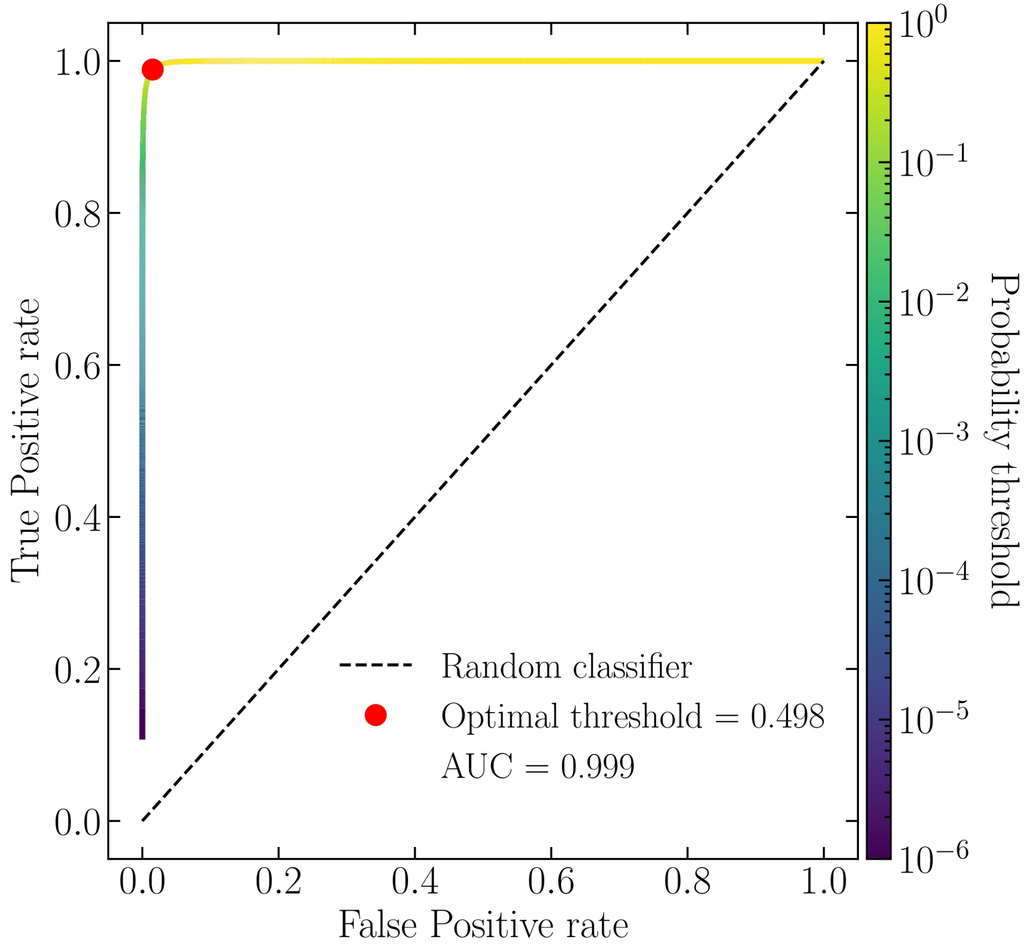
- Cosmology: Planck cosmology
- Simulation: **Set ref** using Gadget4 (Springel et al. 2021)
- Apply **Rockstar** on the Snapshots
- Label DM particles $(\mathbf{1}, \mathbf{0})$ based on the Halo Catalogue
- 500 realisations: Input($\vec{\psi_i}, \vec{\mathbf{v}_i}$)
  - Training: 350
  - Validation: 50
  - Testing: 100



$L200 - N128^3$

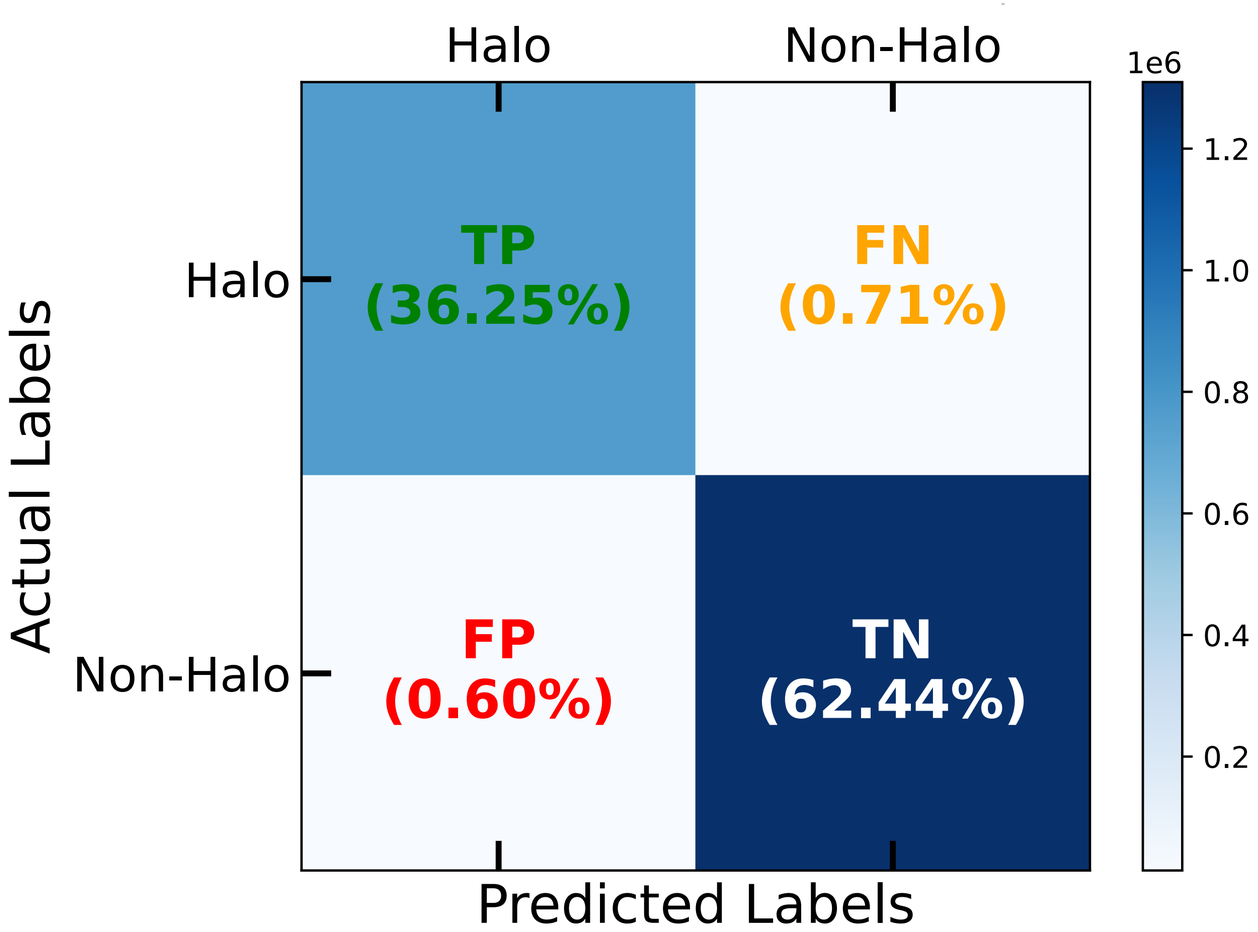
CNNs' Performance

CNNs' Performance



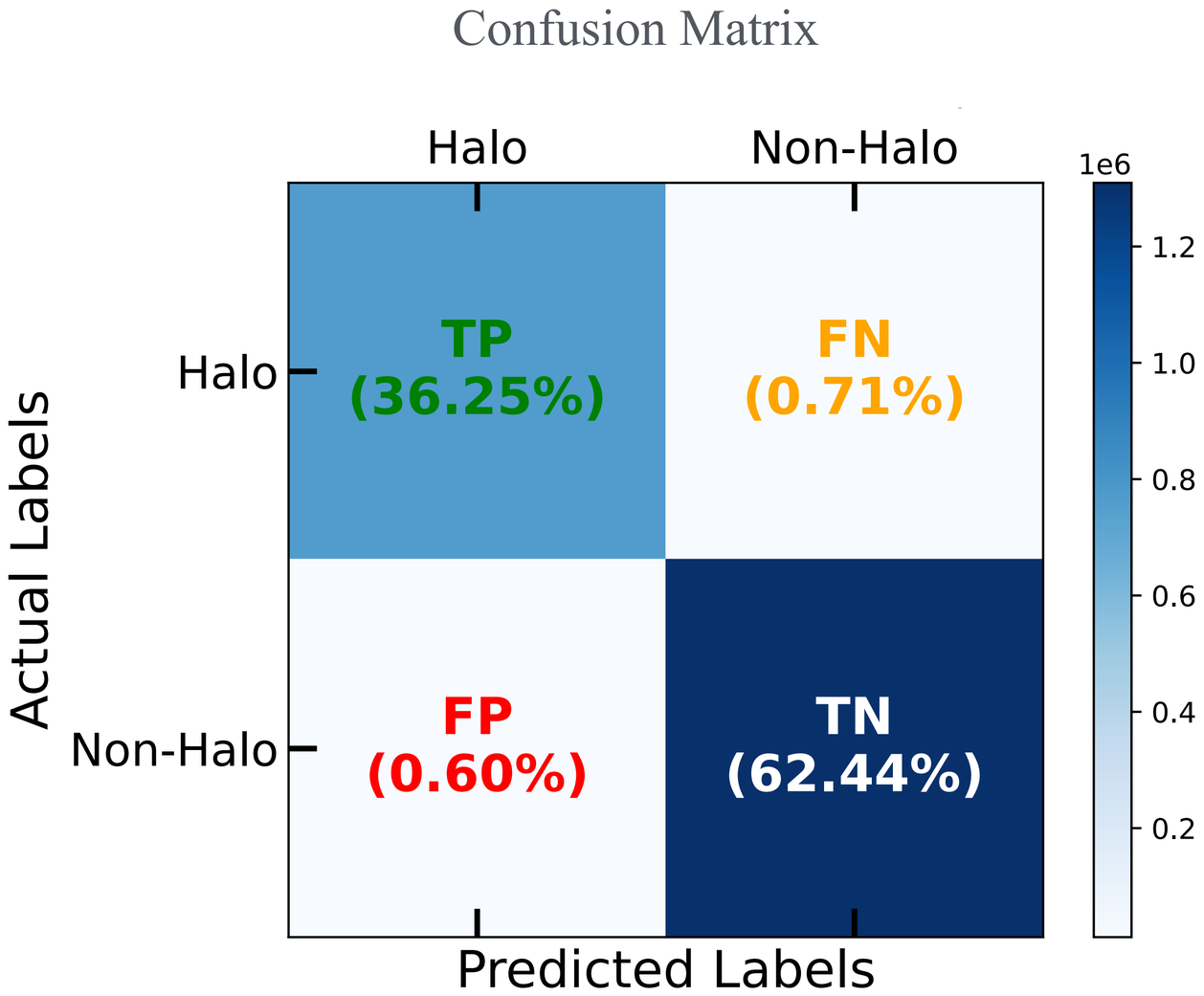
CNNs' Performance

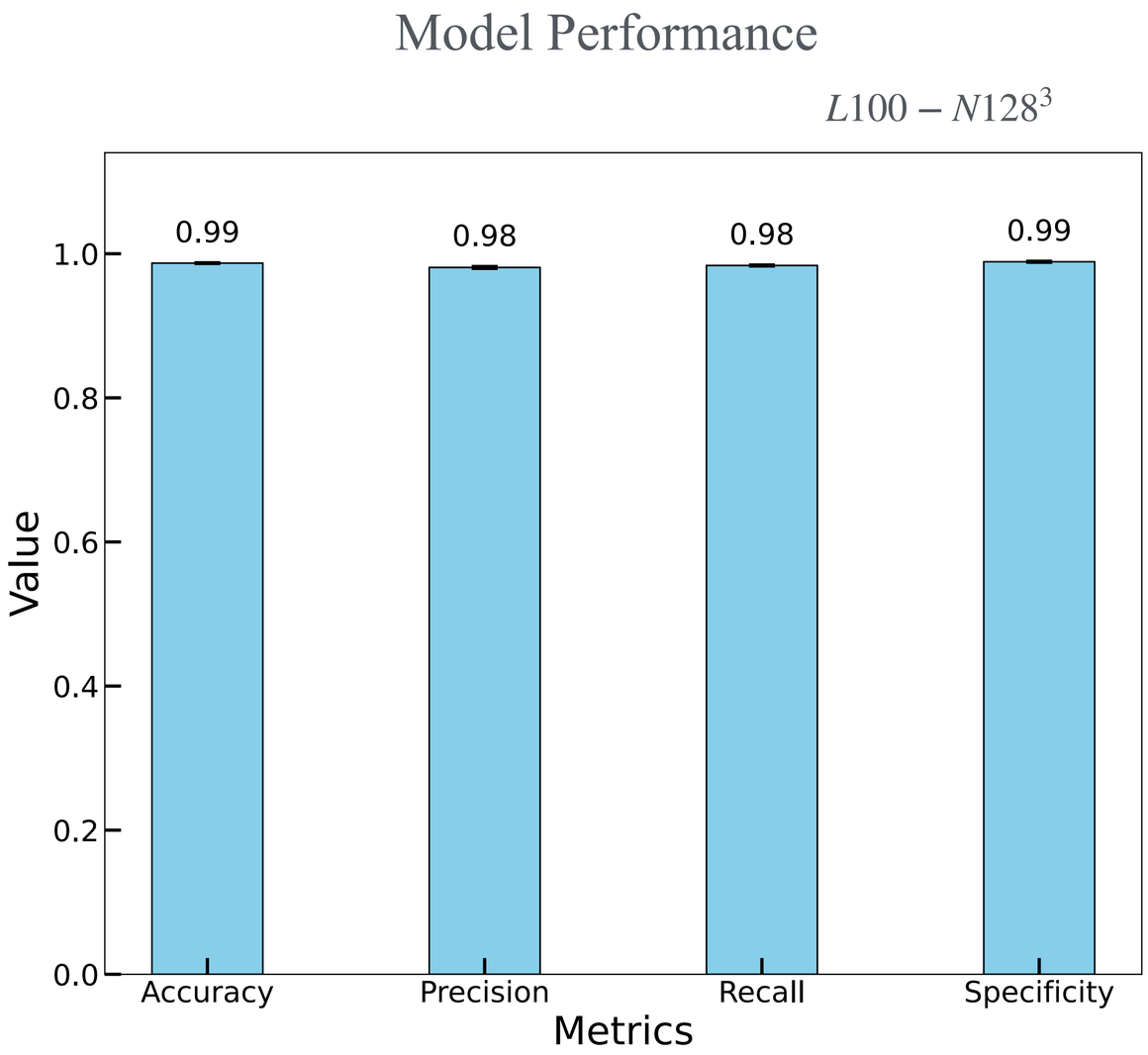
Confusion Matrix



- Correctly Classified: **98.69%**
- Misclassified: **1.31%**

CNNs' Performance





- Correctly Classified: **98.69%**
- Misclassified: **1.31%**

Trained Models

ROCKSTAR

Mass / Resolution	M_{200b}	M_{vir}
$L200 - N64^3$	✓	✓
$L400 - N128^3$	✓	✓
$L200 - N128^3$	✓	✓
$L100 - N128^3$	✓	✓

Resolution Increase

Trained Model on RAVEN GPUs (MPCDF)

Trained Models

ROCKSTAR

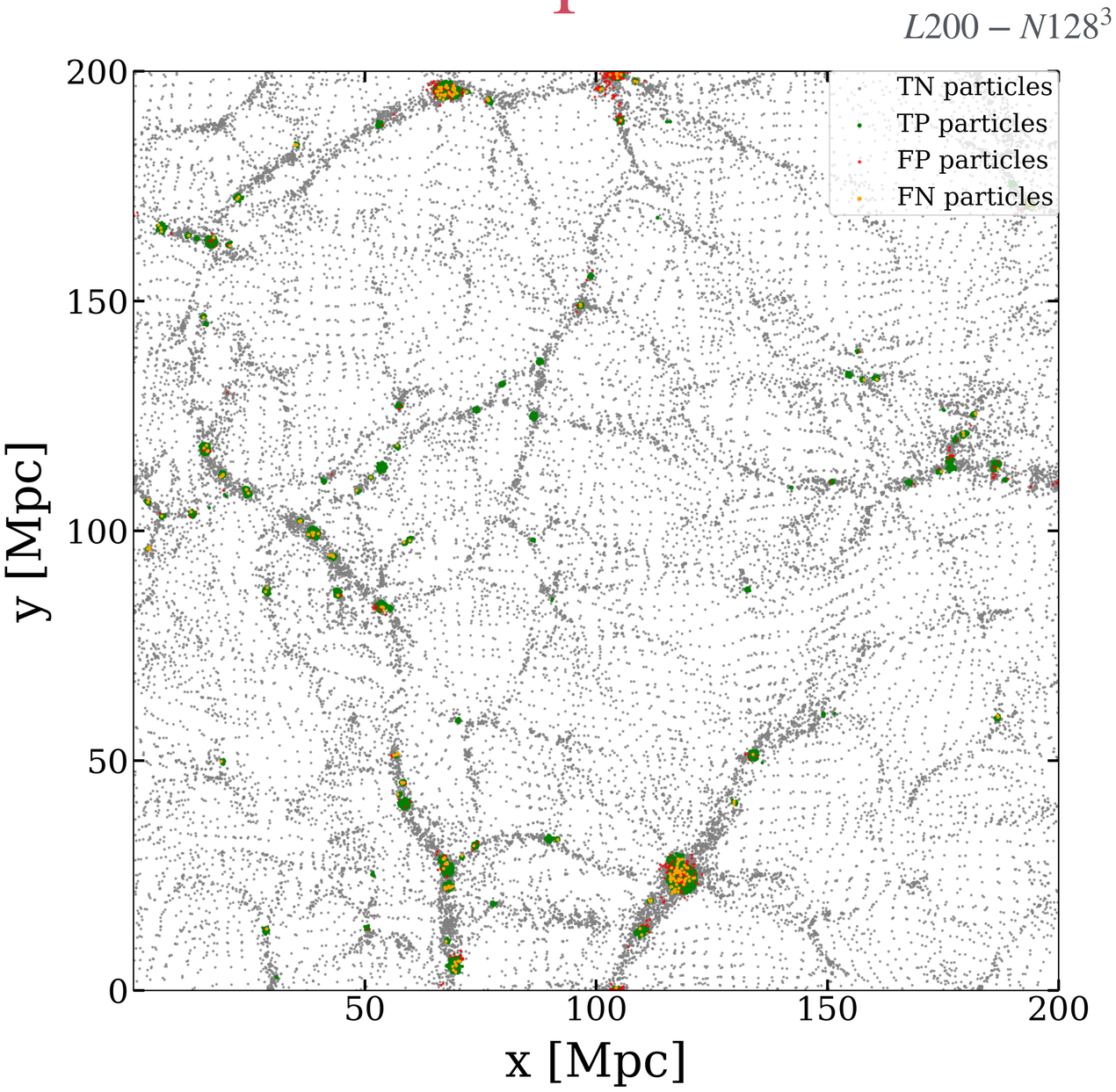
Resolution Increase

Mass / Resolution	M_{200b}	M_{vir}
$L200 - N64^3$	✓	✓
$L400 - N128^3$	✓	✓
$L200 - N128^3$	✓	✓
$L100 - N128^3$	✓	✓

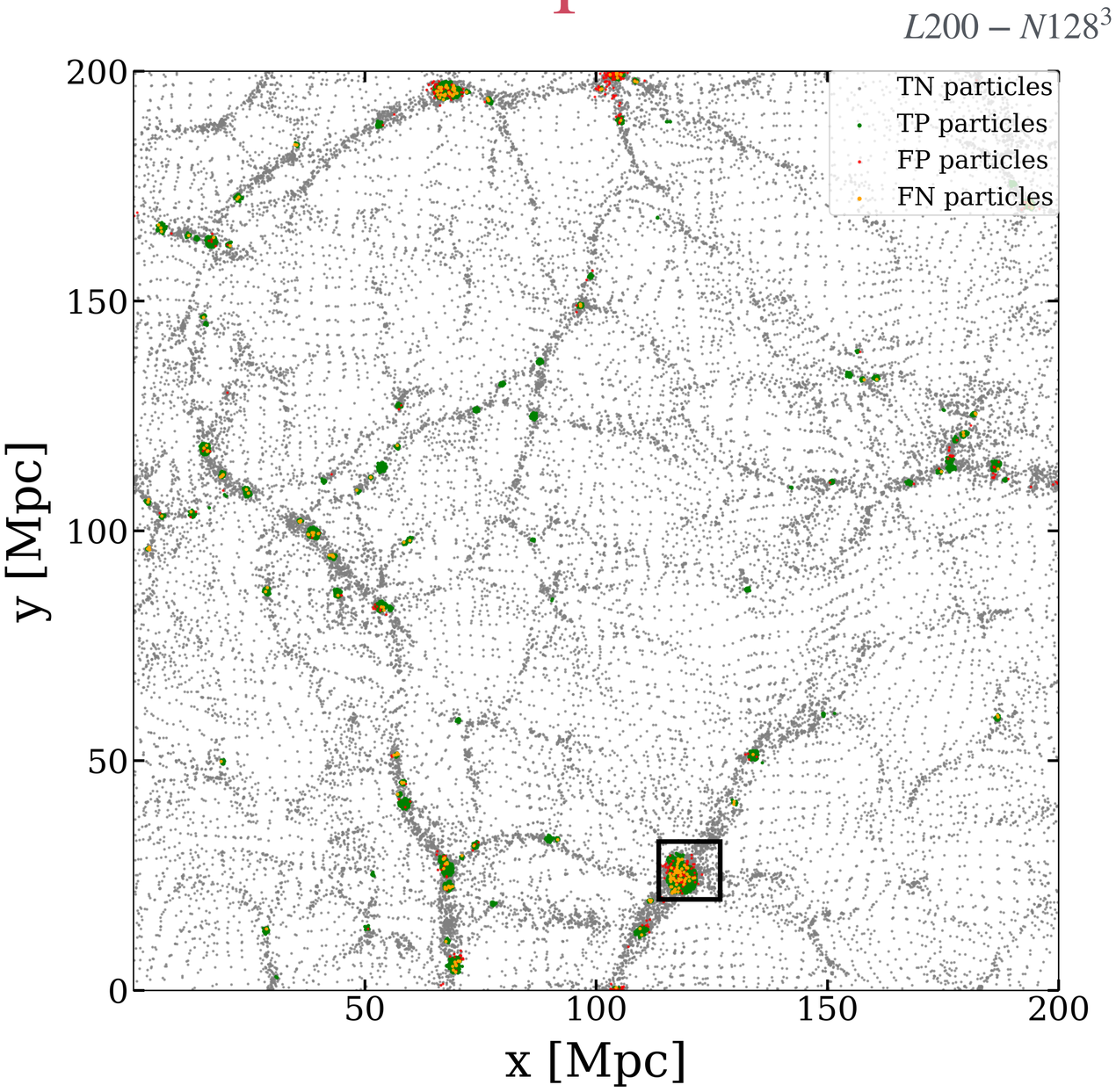
Trained Model on RAVEN GPUs (MPCDF)

Emulator Output

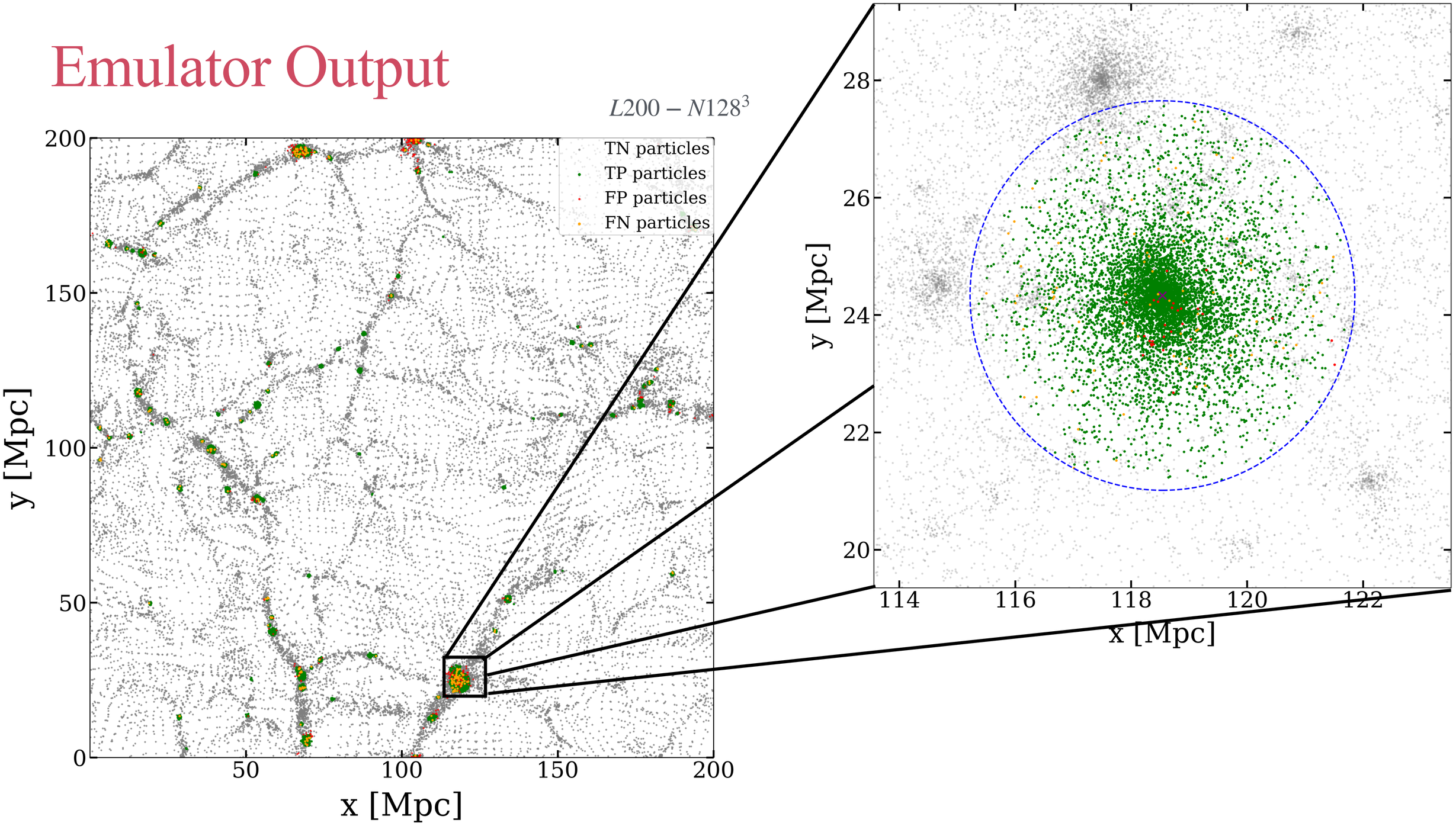
Emulator Output



Emulator Output



Emulator Output



Recovering Centres of Emulator Output

- Applied FoF on Emulator Identified Positive (TP+FP) Particles
- One-to-one correspondence of Rockstar and FoF Centers.

Recovering Centres of Emulator Output

- Applied FoF on Emulator Identified Positive (TP+FP) Particles

- One-to-one correspondence of Rockstar and FoF Centers.

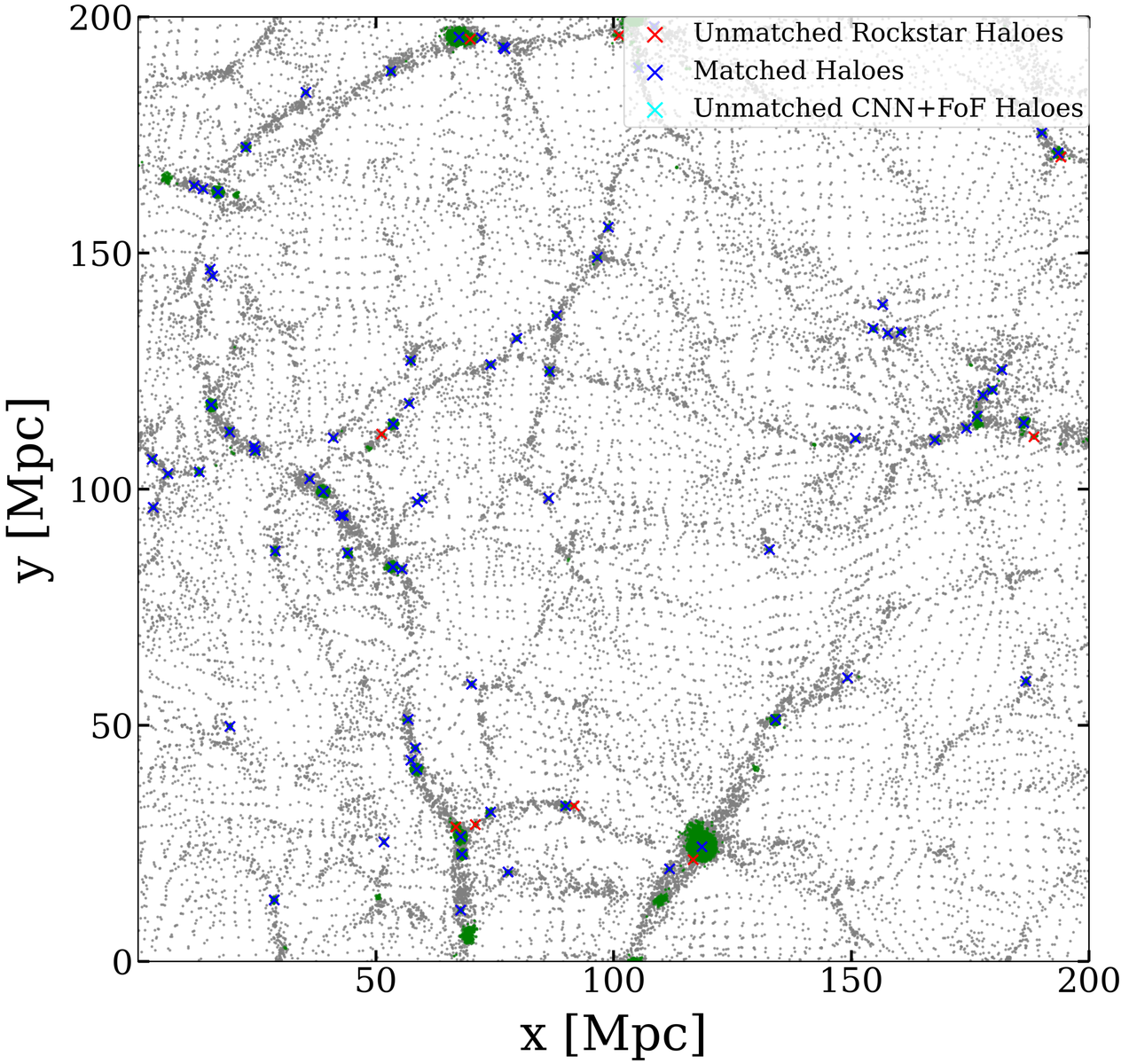
$$d_{FR} = \begin{cases} \leq R_{200b}, & \text{Matched Haloes} \\ > R_{200b}, & \text{Unmatched Haloes} \end{cases}$$

Recovering Centres of Emulator Output

$L200 - N128^3$

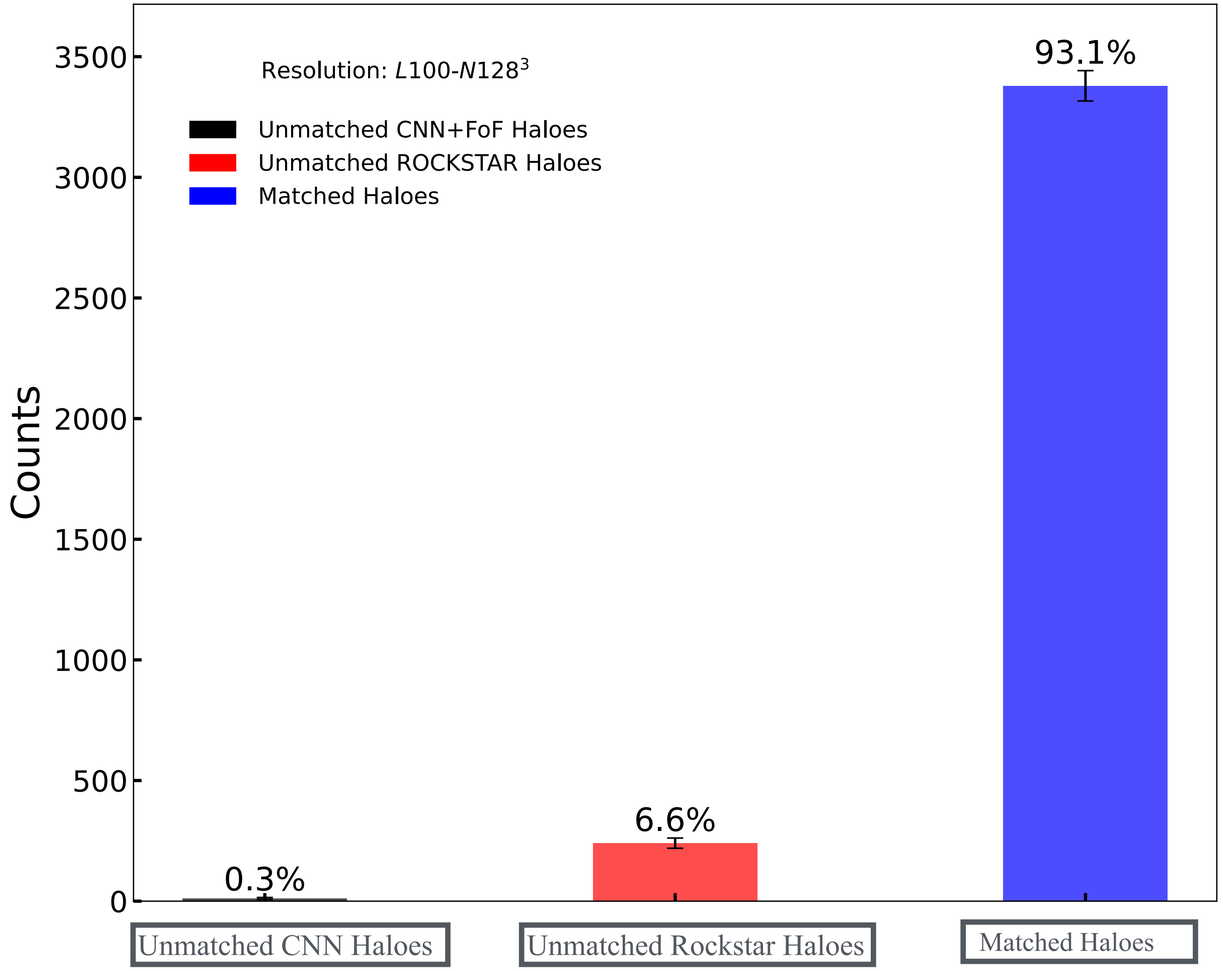
- Applied FoF on Emulator Identified Positive (TP+FP) Particles
- One-to-one correspondence of Rockstar and FoF Centers.

$$d_{FR} = \begin{cases} \leq R_{200b}, & \text{Matched Haloes} \\ > R_{200b}, & \text{Unmatched Haloes} \end{cases}$$

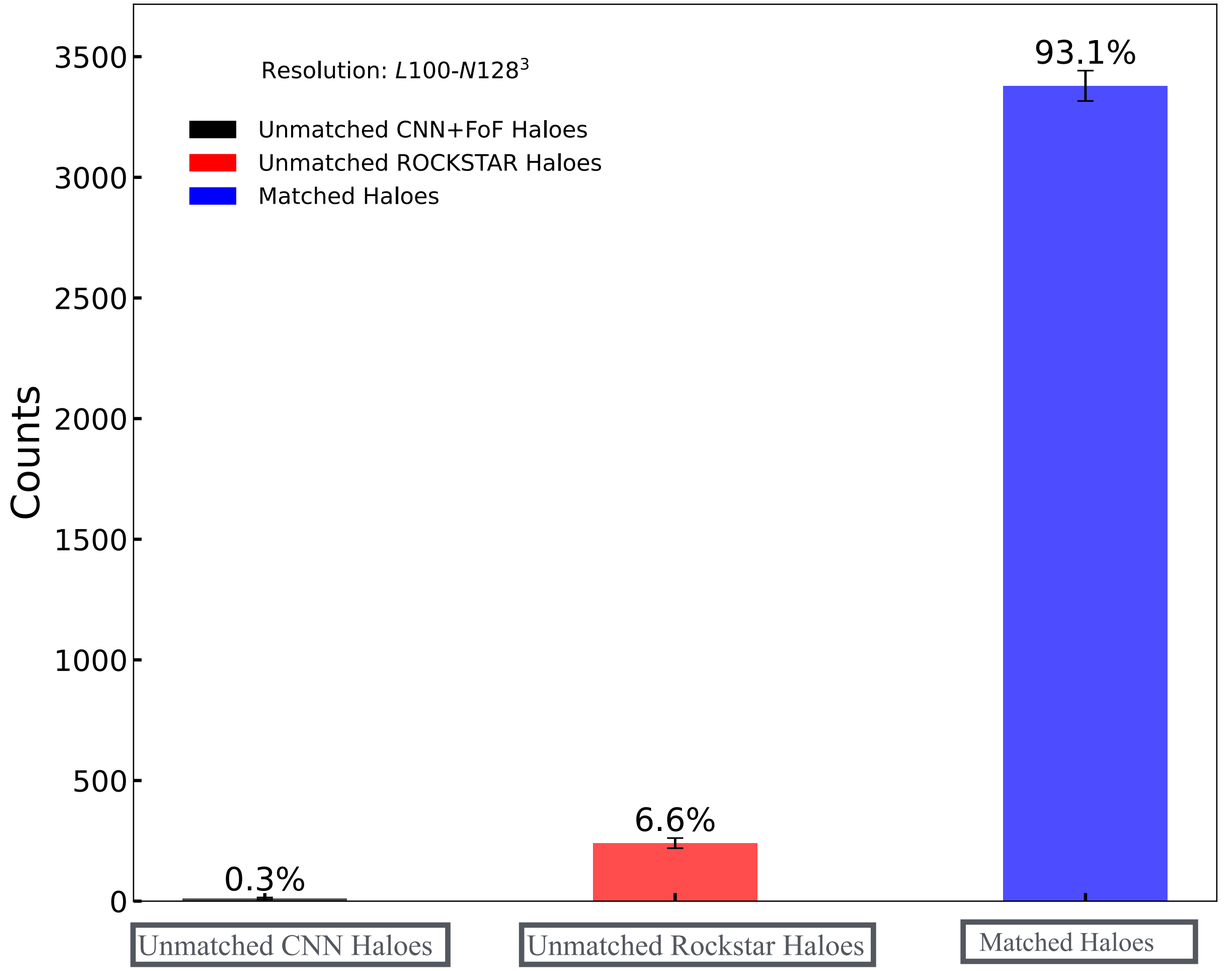


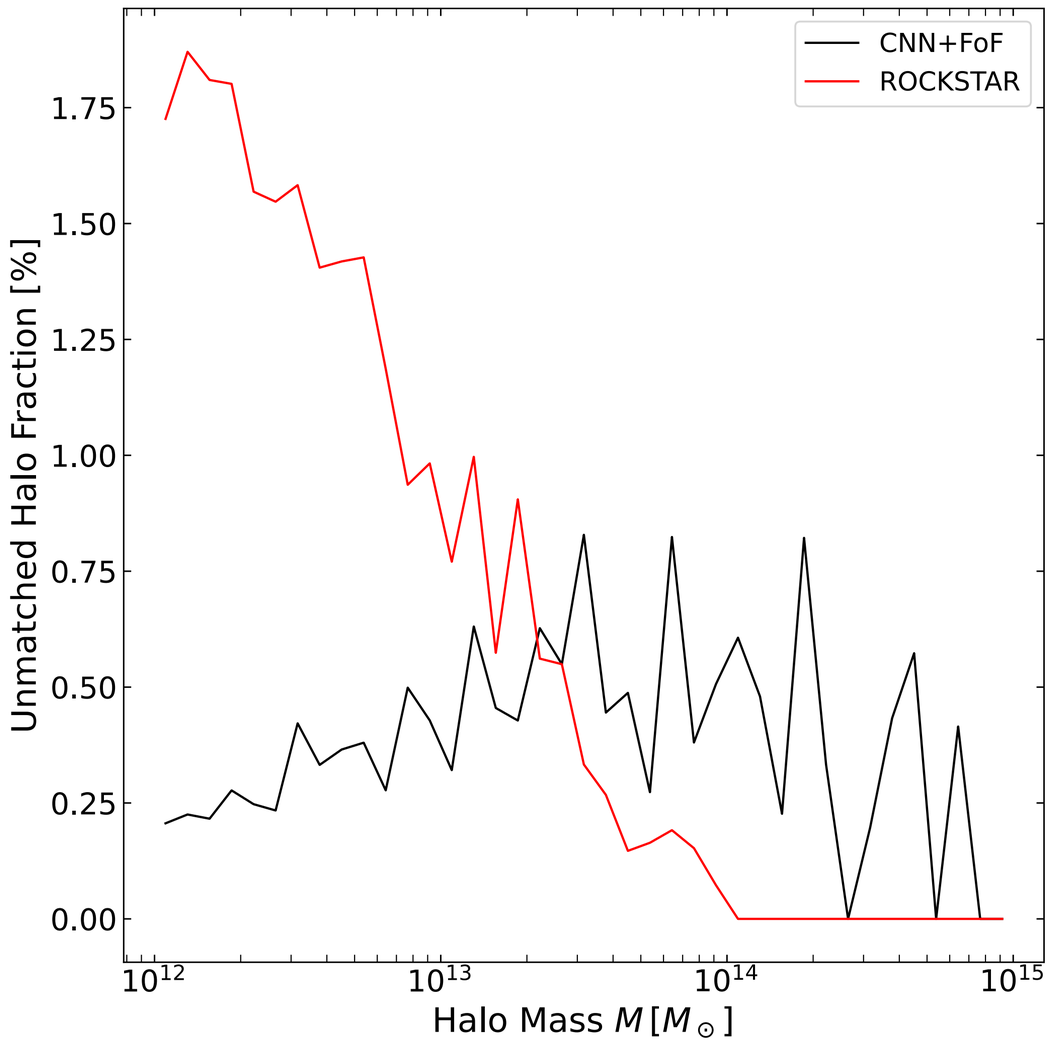
Halo Matchings

Halo Matchings

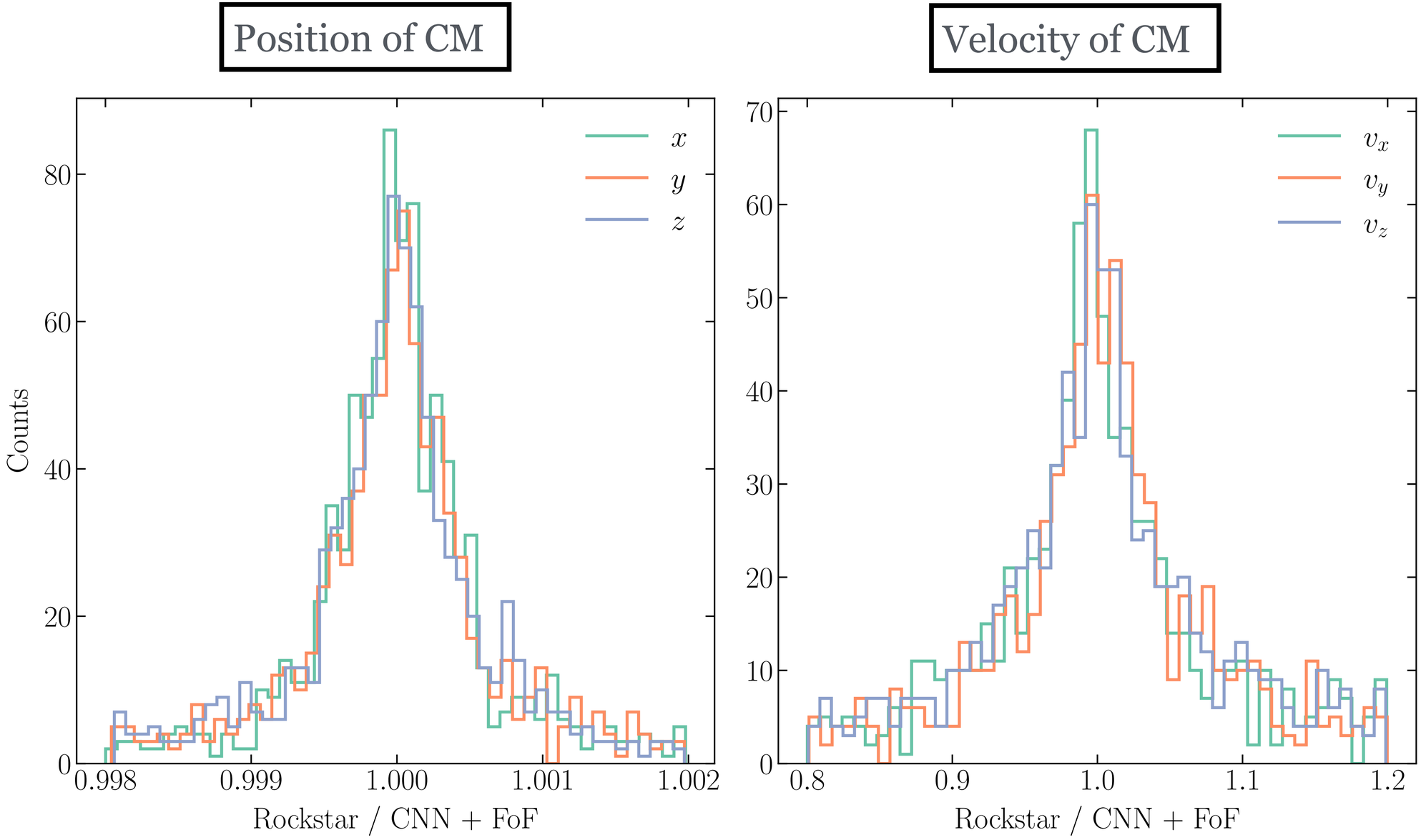


Halo Matchings





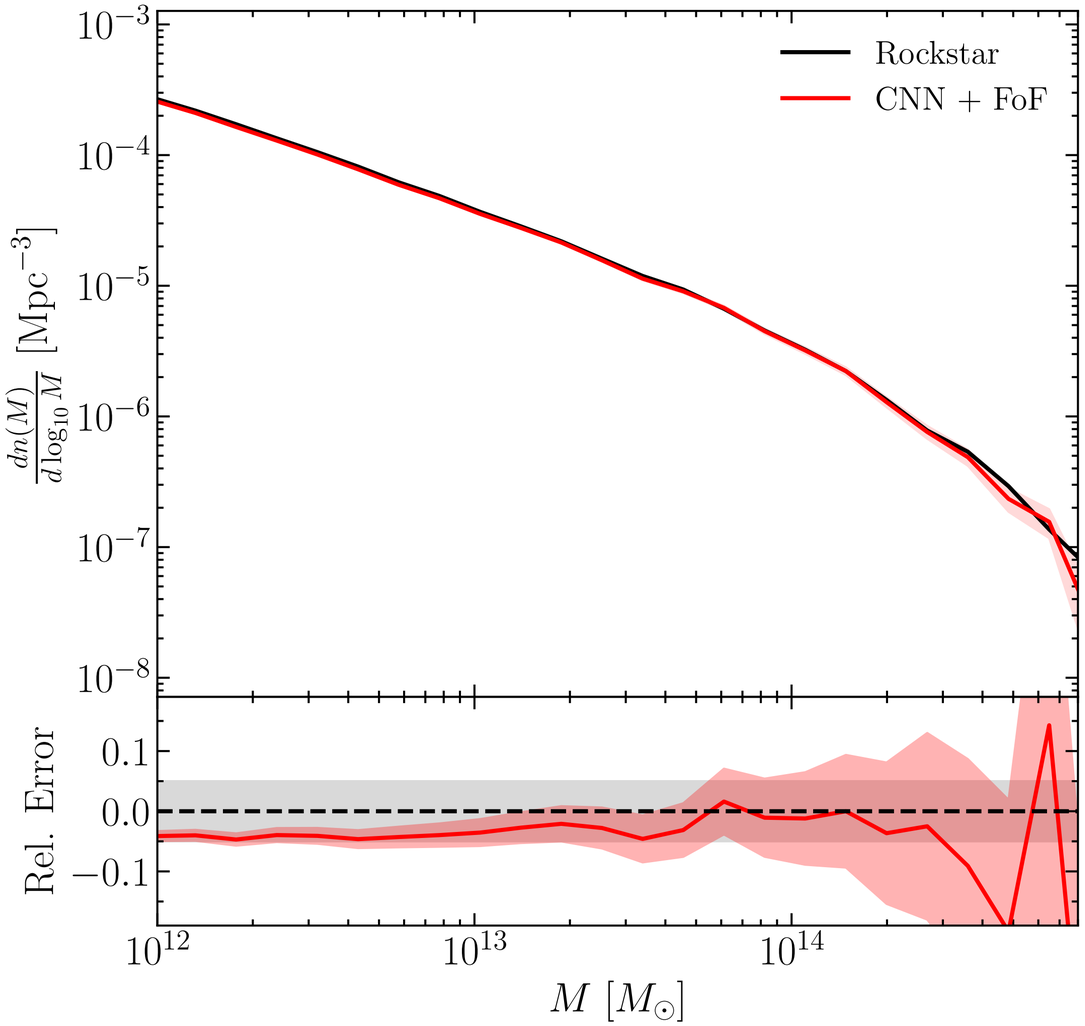
Position and Velocity Recovery



Clustering Statistics

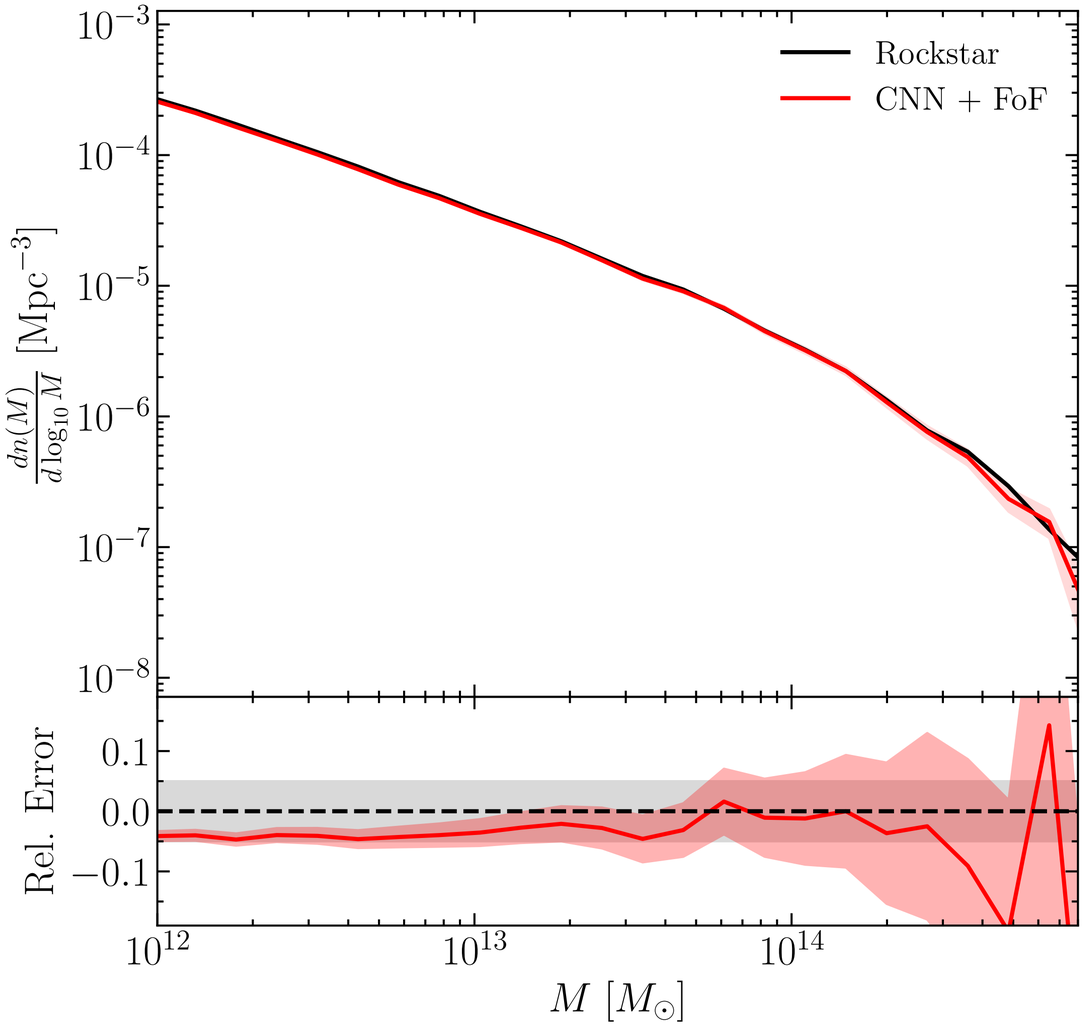
Clustering Statistics

Halo Mass Function

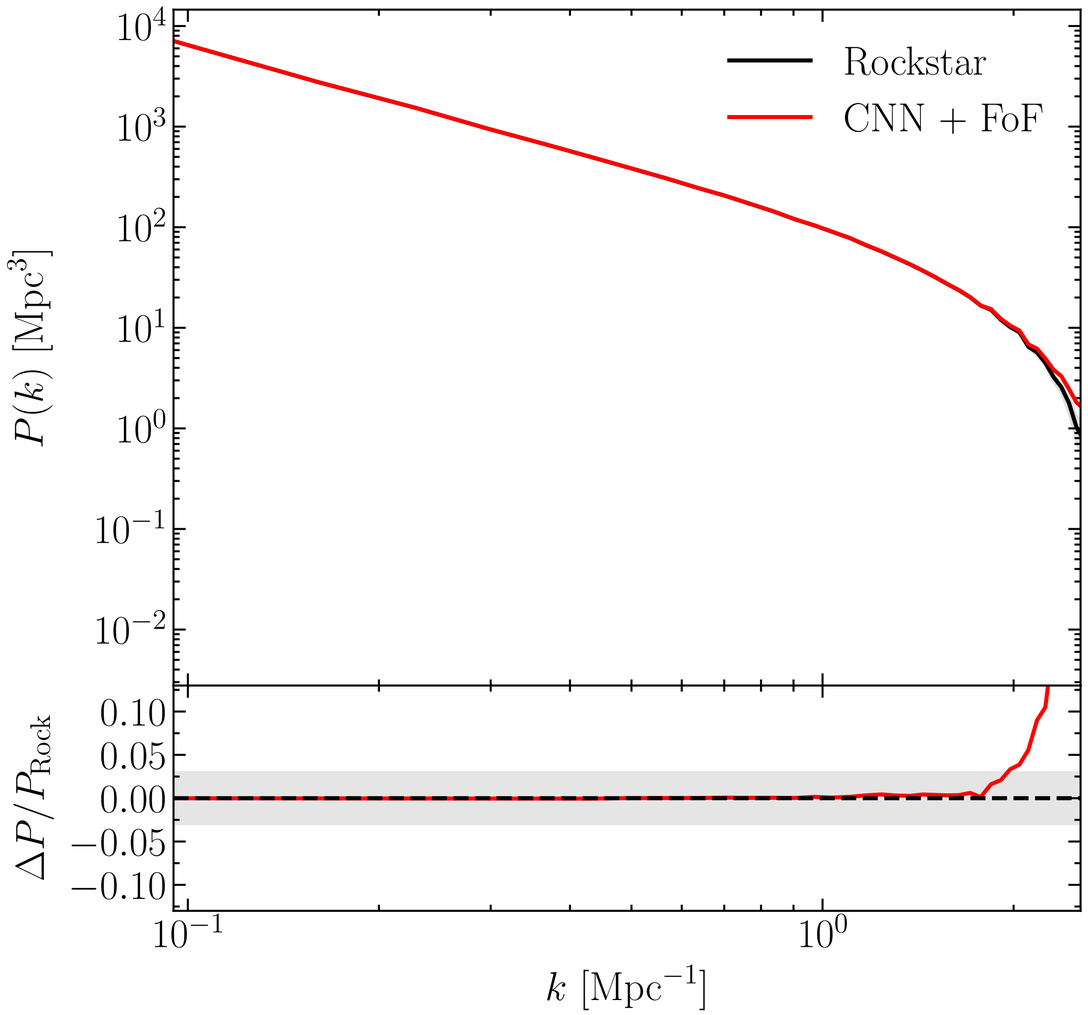


Clustering Statistics

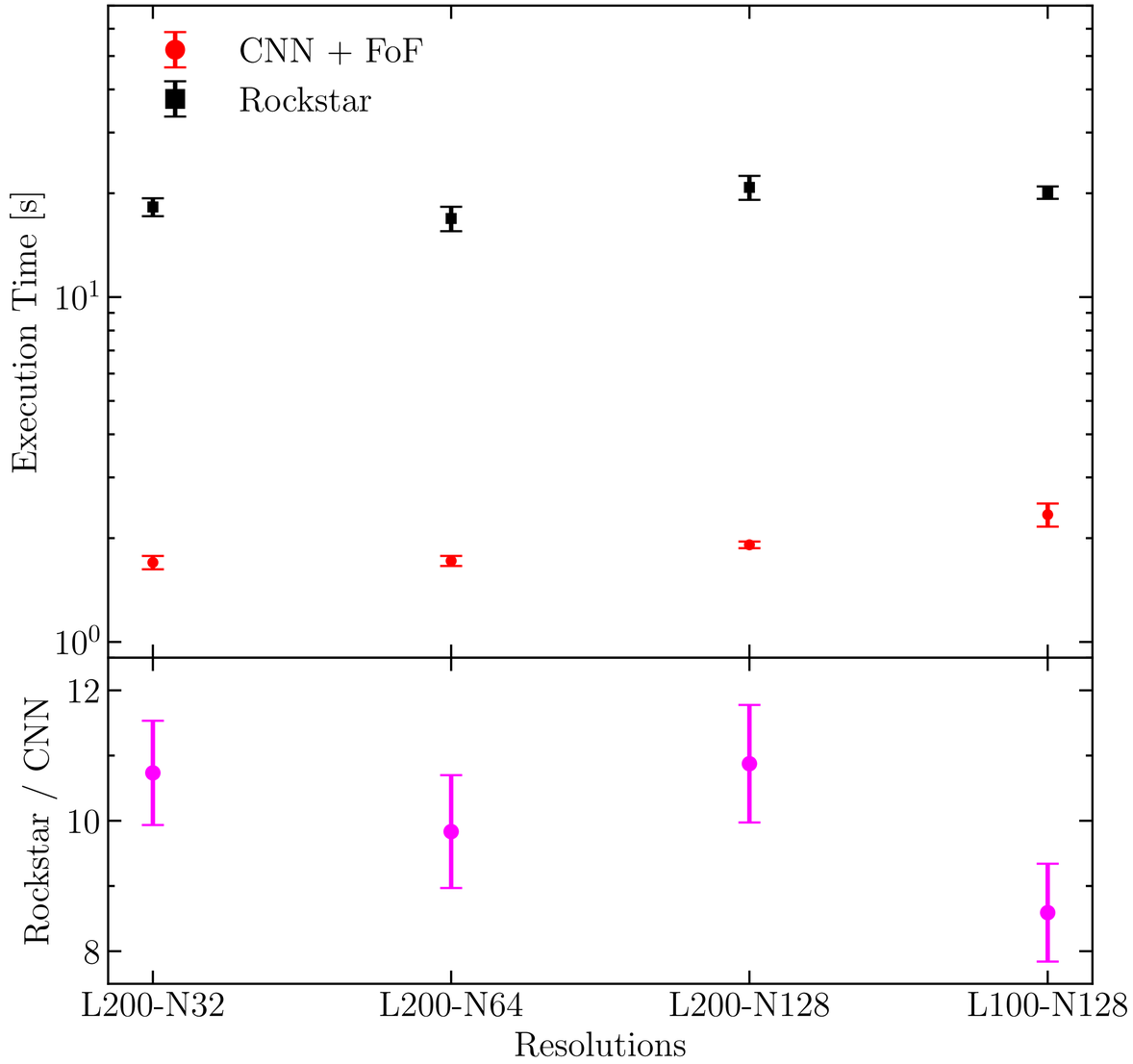
Halo Mass Function



Power Spectrum



Runtime Comparison



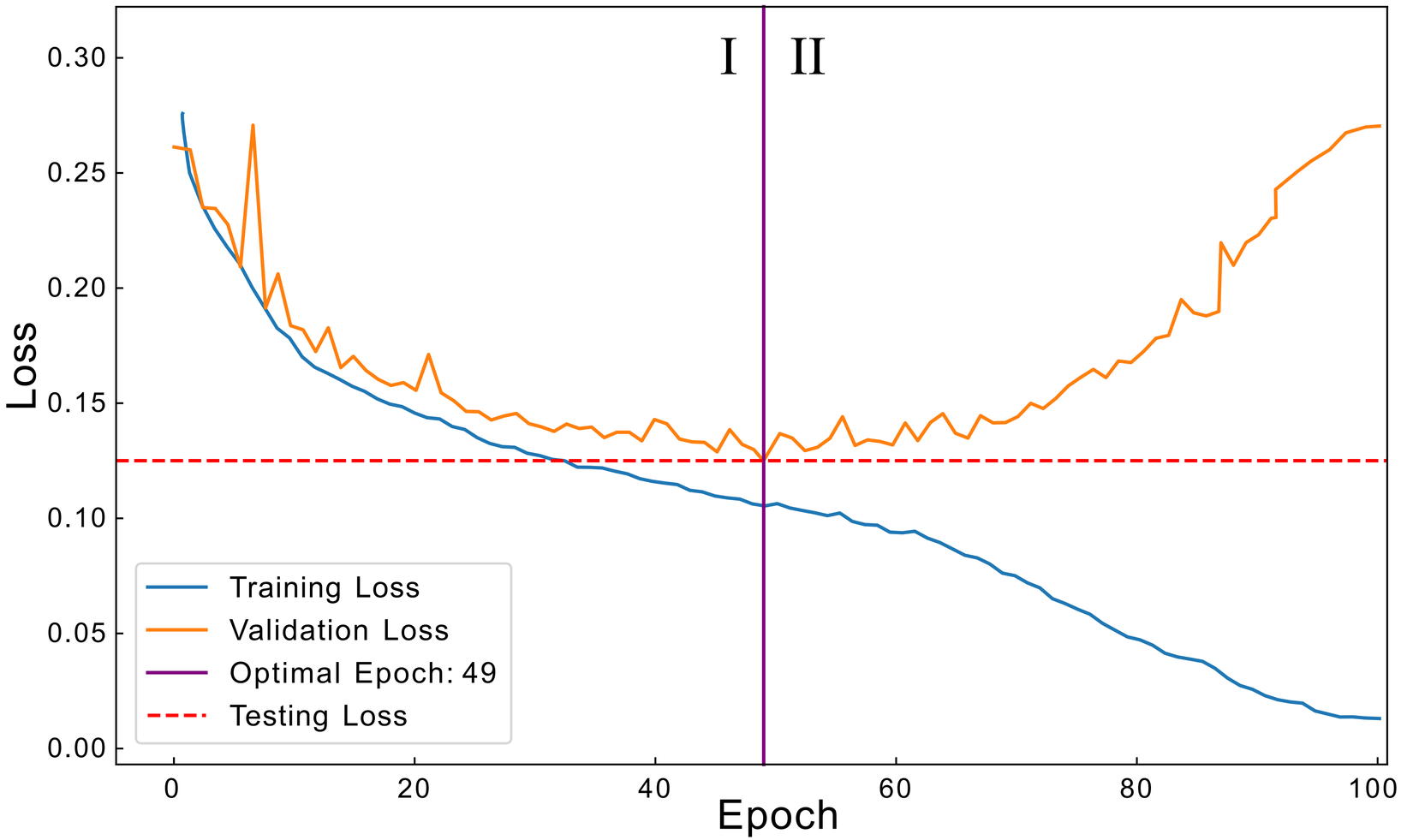
Conclusions

Final Remarks and Future work

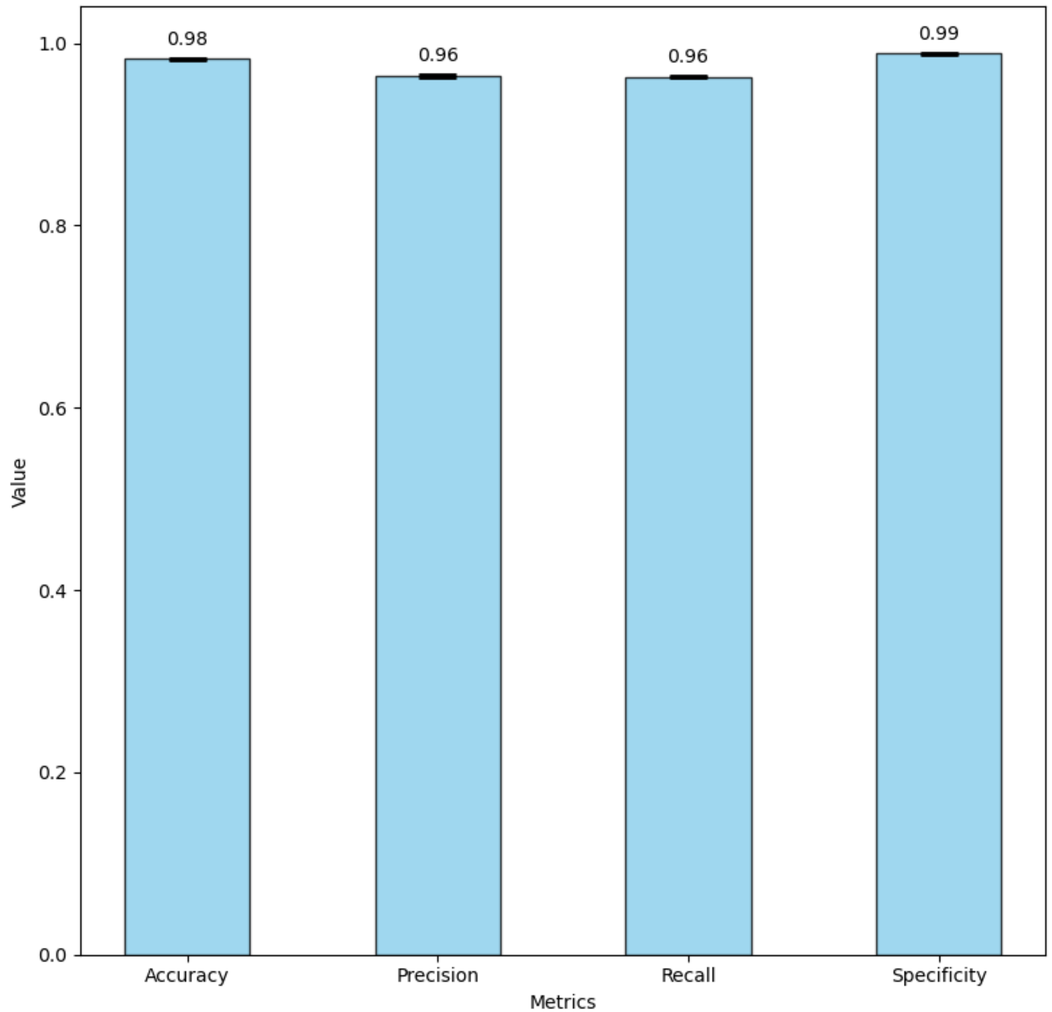
- **Massive** and **Fast** production of **halo catalogues** for cosmological analyses
- Halo Finder based on ML 2 steps: **classification** of particles with a CNN and **clustering algorithm** (FoF) to get the centers
- Our models:
  - **Resolutions**: $L200 - N128^3$, $L100 - N128^3$, etc.
  - **Mass definitions**: M_{200b}, M_{vir}
- **Classification step** works at **particle level** and it is cost-effective. Metrics show a very good performance (99% in accuracy)
- **Centre recovery step** is also working very well, as the matching analysis shows
- However it is still a bottleneck because of using FoF (work in progress $\longrightarrow$ extension of the CNN to also do this task)
- Apply our method to **bigger box** sizes. With the model of $L100 - N128^3$ we can work on $L1500 - N2048^3$ (work in progress)
- Cosmological analysis (work in progress)


Thank You

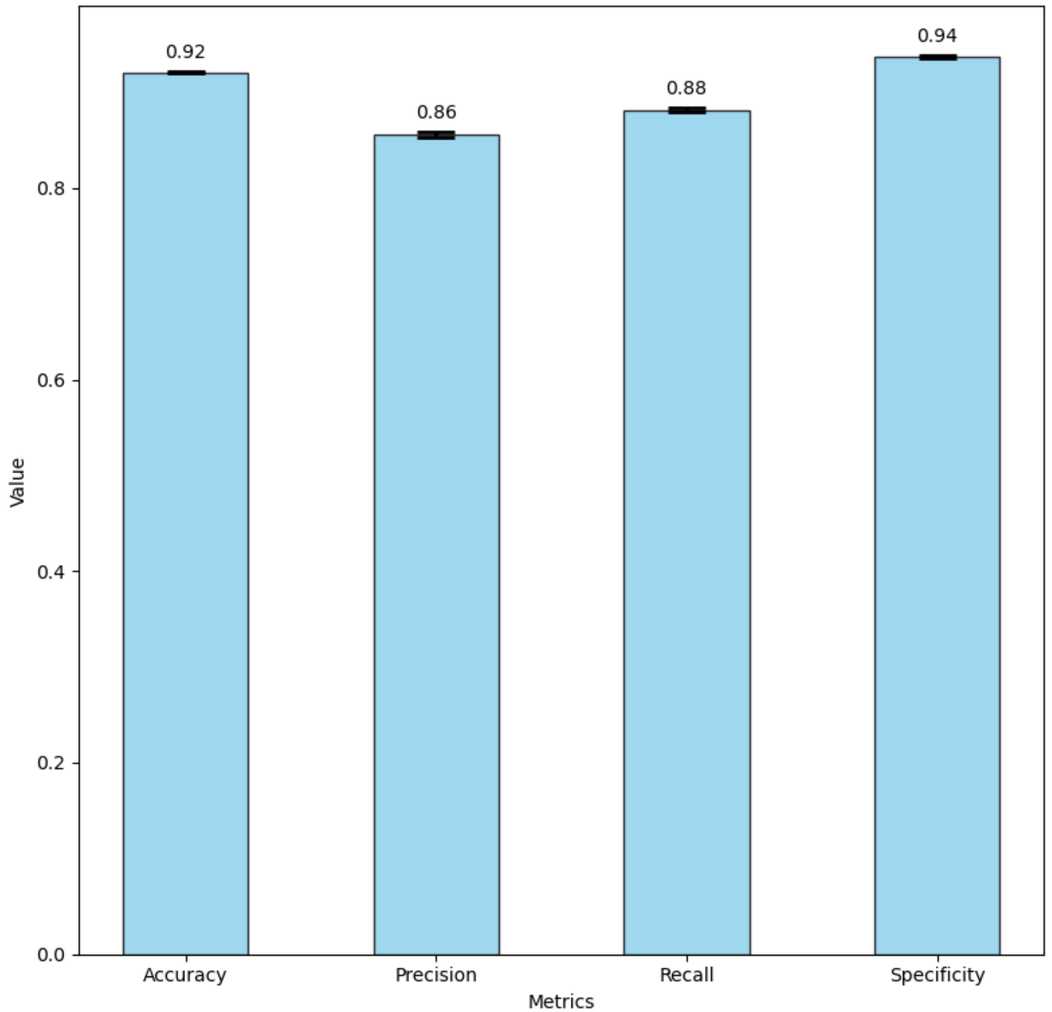
Backup

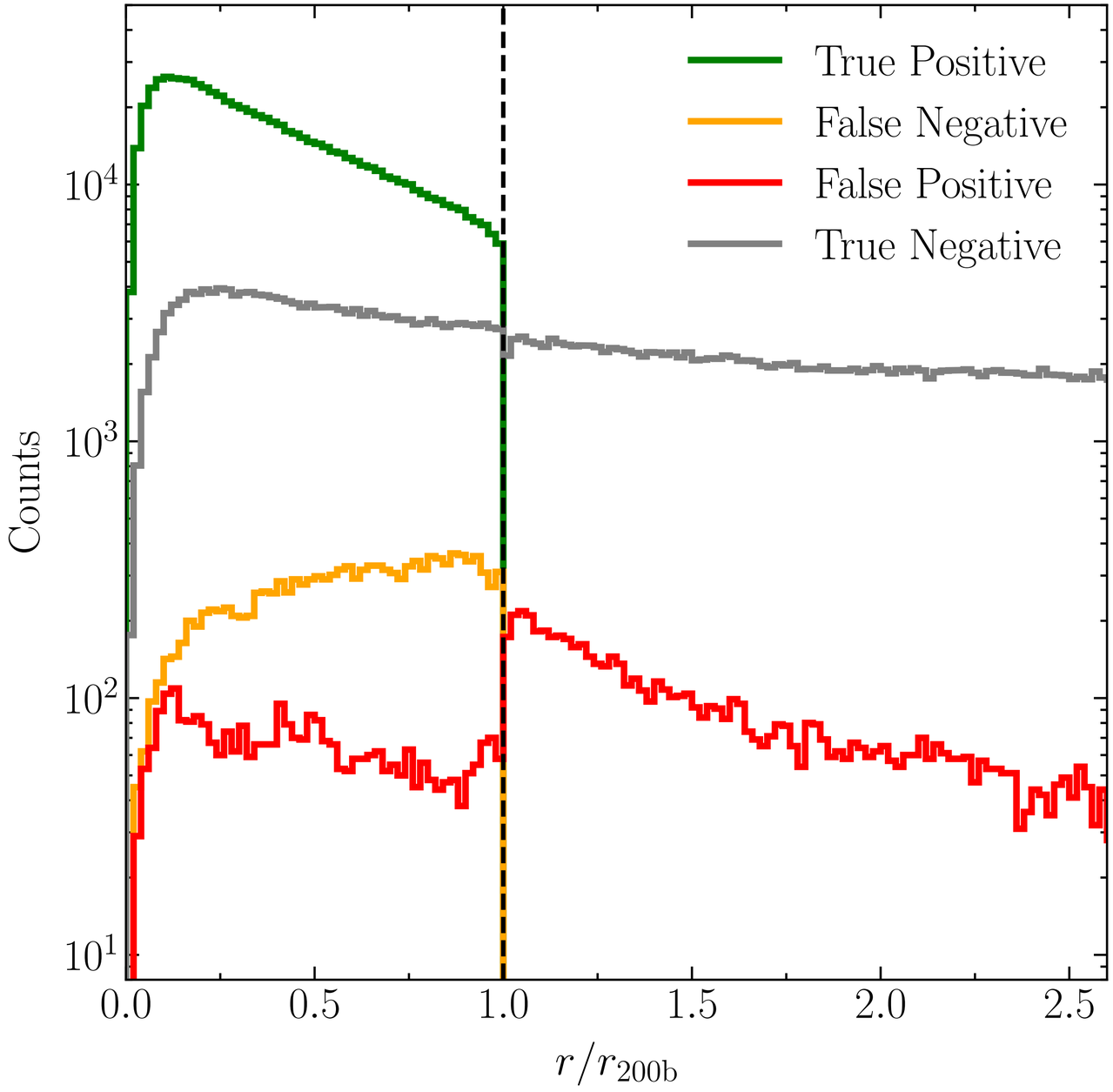

I
II
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Loss
0
20
40
60
80
100
Epoch
Training Loss
Validation Loss
Optimal Epoch: 49
Testing Loss

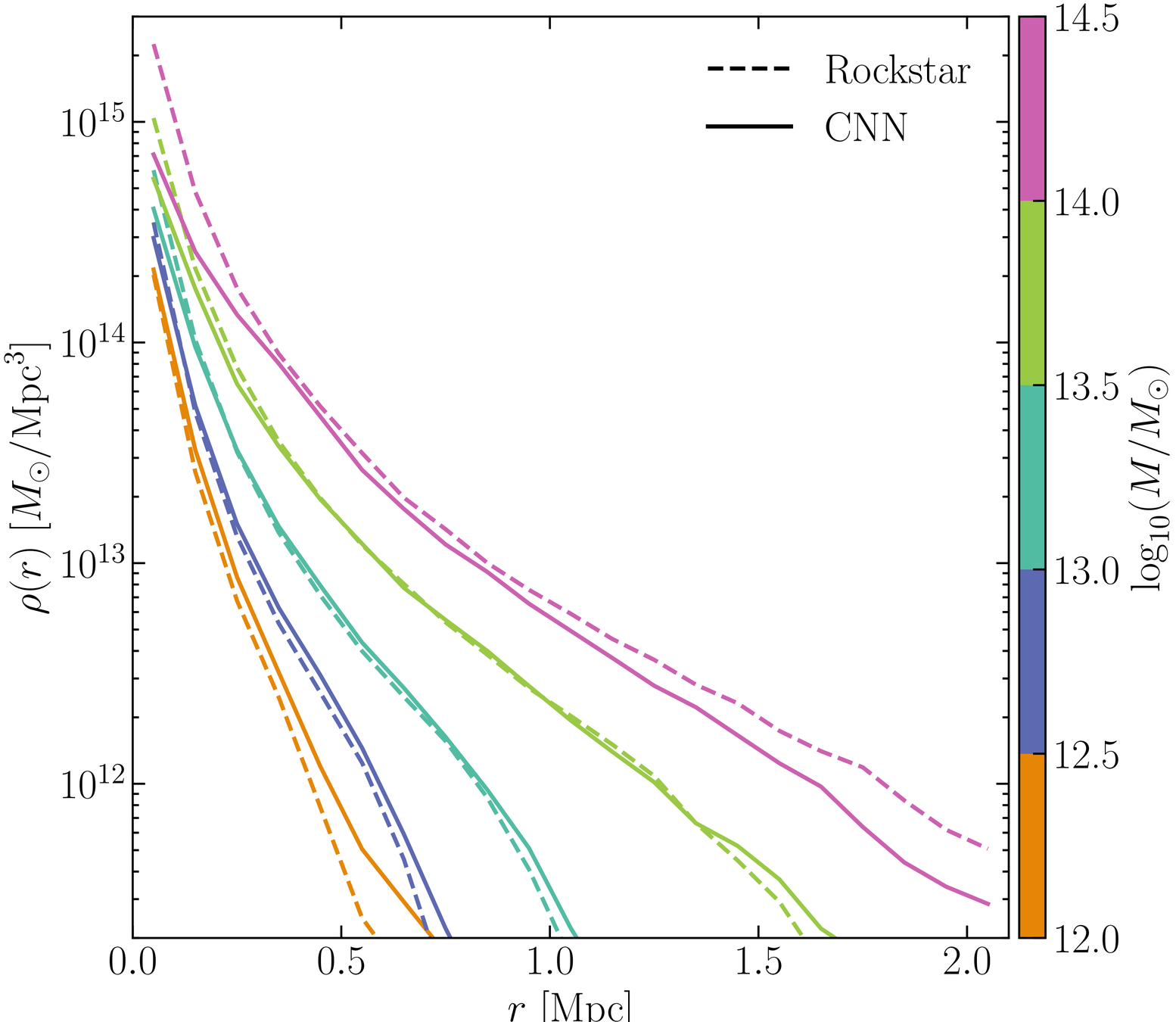
$\mathbf{M_{vir}}$



$\mathbf{M_{FoF}}$




True Positive
False Negative
False Positive
True Negative
Counts
10^4
10^3
10^2
10^1
0.0
0.5
1.0
1.5
2.0
2.5
r/r_{200b}


Rockstar
CNN
10^{15}
10^{14}
10^{13}
10^{12}
$\rho(r)$ $[M_\odot/\mathrm{Mpc}^3]$
0.0
0.5
1.0
1.5
2.0
r [Mpc]
14.5
14.0
13.5
13.0
12.5
12.0
$\log_{10}(M/M_\odot)$