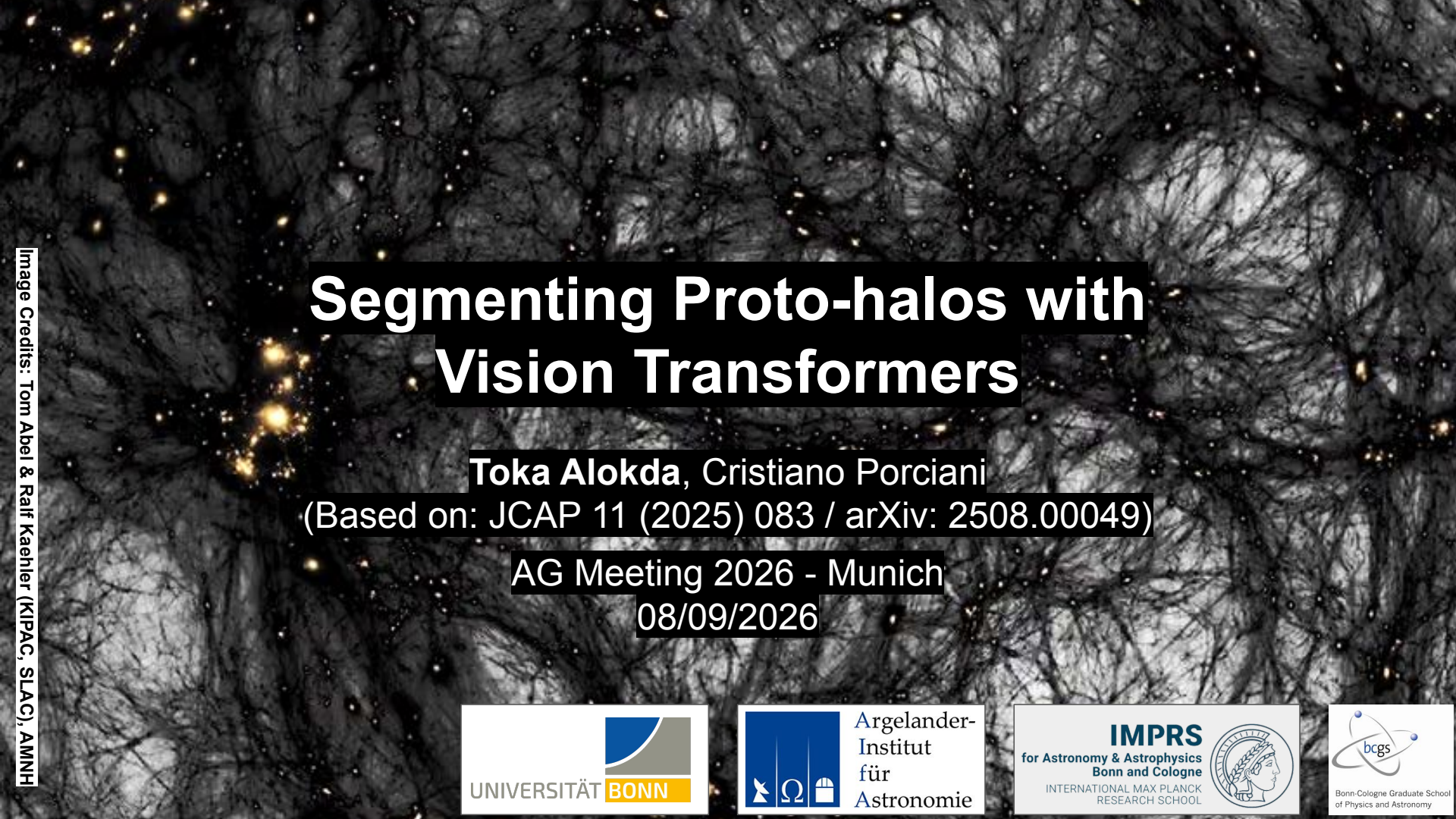


Segmenting Proto-halos with Vision Transformers

Toka Alokda, Cristiano Porciani

(Based on: JCAP 11 (2025) 083 / arXiv: 2508.00049)

AG Meeting 2026 - Munich

08/09/2026

Image Credits: Tom Abel & Ralf Kaehler (KIPAC, SLAC), AMNH

UNIVERSITÄT BONN

Argelander-Institut für Astronomie

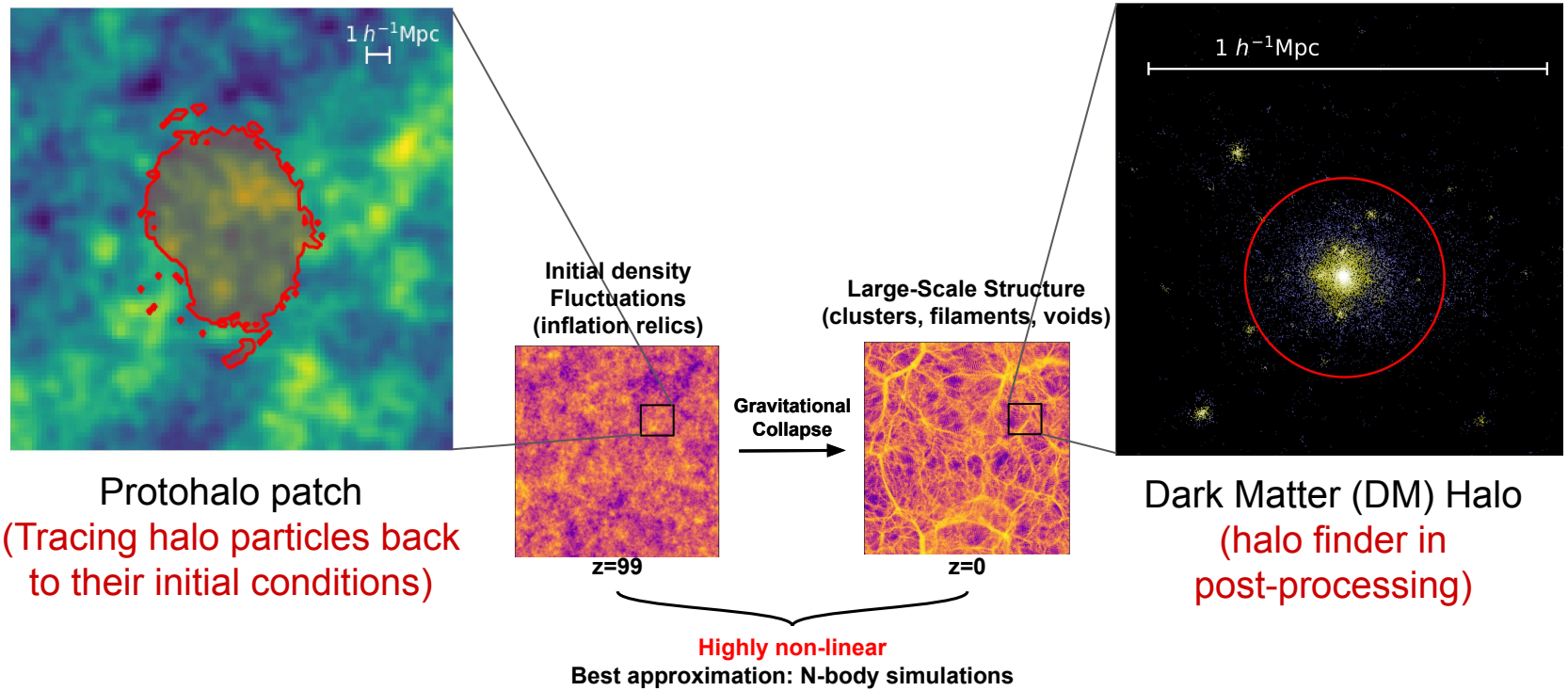
IMPRS for Astronomy & Astrophysics Bonn and Cologne

INTERNATIONAL MAX PLANCK RESEARCH SCHOOL

bcgs

Bonn-Cologne Graduate School of Physics and Astronomy

Cosmological Structure Formation



Protohalo patch
(Tracing halo particles back to their initial conditions)

Dark Matter (DM) Halo
(halo finder in post-processing)

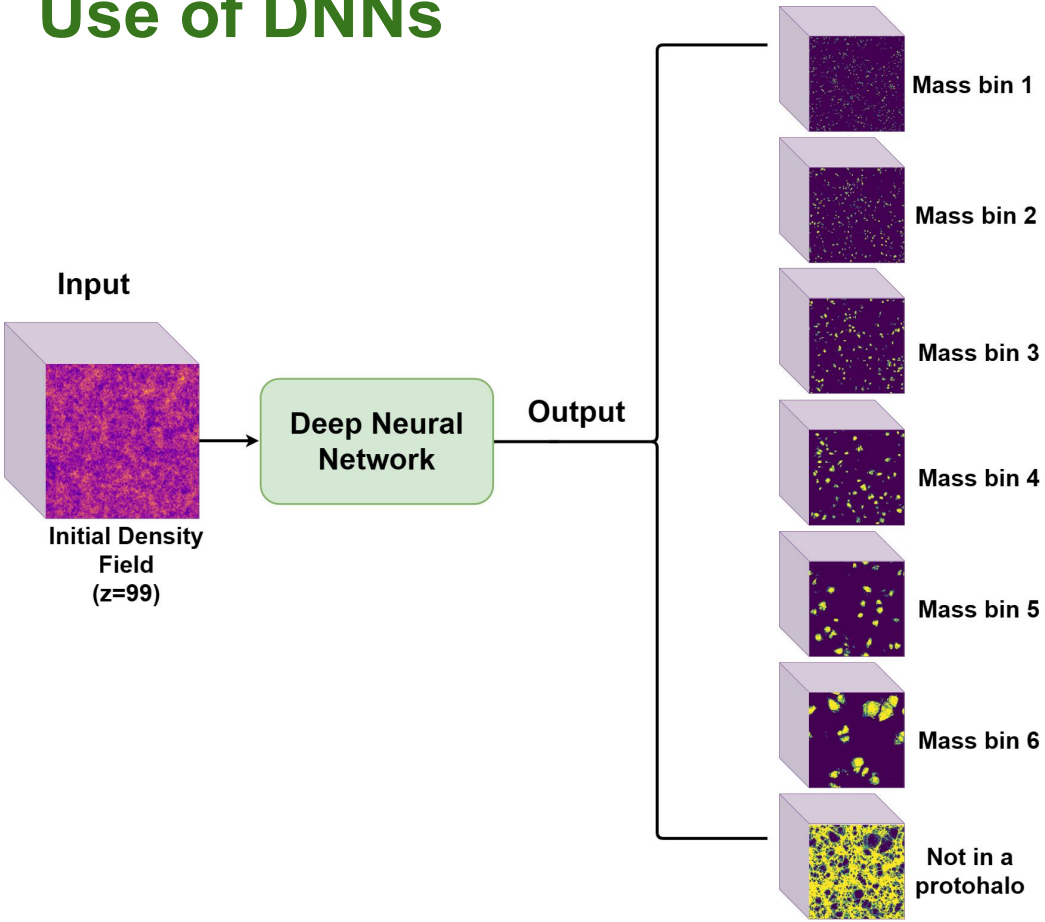
Question:
Does a function
***f* (initial conditions) = late-time cosmic structure**
exist?

Possible solution: Deep Neural Networks (DNNs)

Universal approximation theorem (UAT):

*"feedforward neural networks with **at least one hidden layer** can approximate any **continuous, real-valued function** to any desired degree of precision, provided they have sufficient neurons and a non-polynomial activation function"* **(e.g. Hornik+ '89)**

Use of DNNs



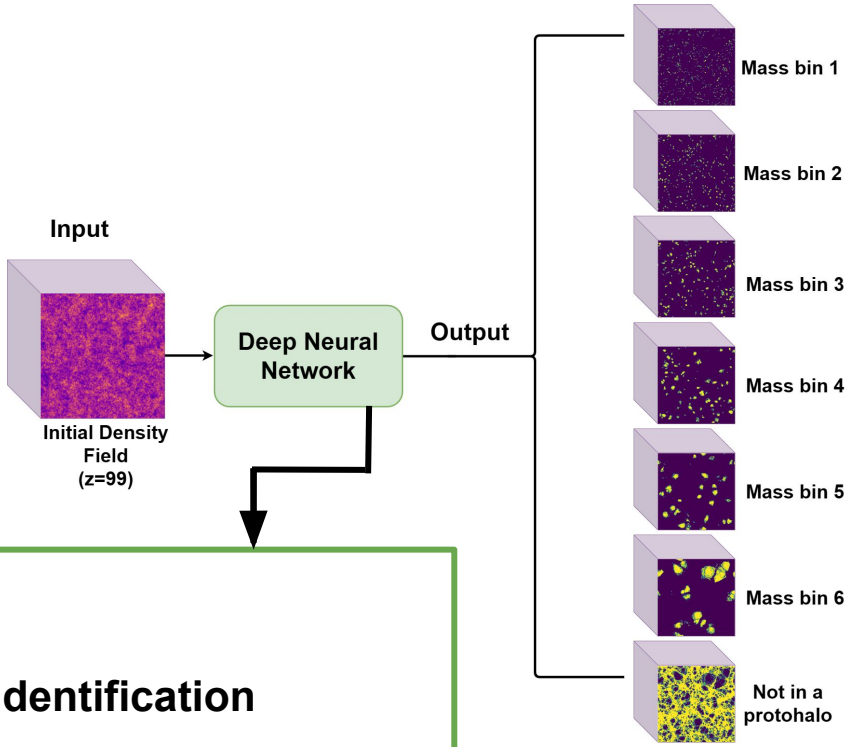
Such trained models can make a good base for downstream tasks like:

- **Fast & accurate mock (galaxy/halo) catalogs**
  To boost galaxy redshift survey data analysis
  (e.g. Berger & Stein ‘19)

- **Learning which regions could be interesting to zoom-in simulate**
  To reap all the benefits with a fraction of the costs

… and more.

Two flavors of DNNs



Two models 🐍 (coded up in Python)

(CNN encoder and decoder)

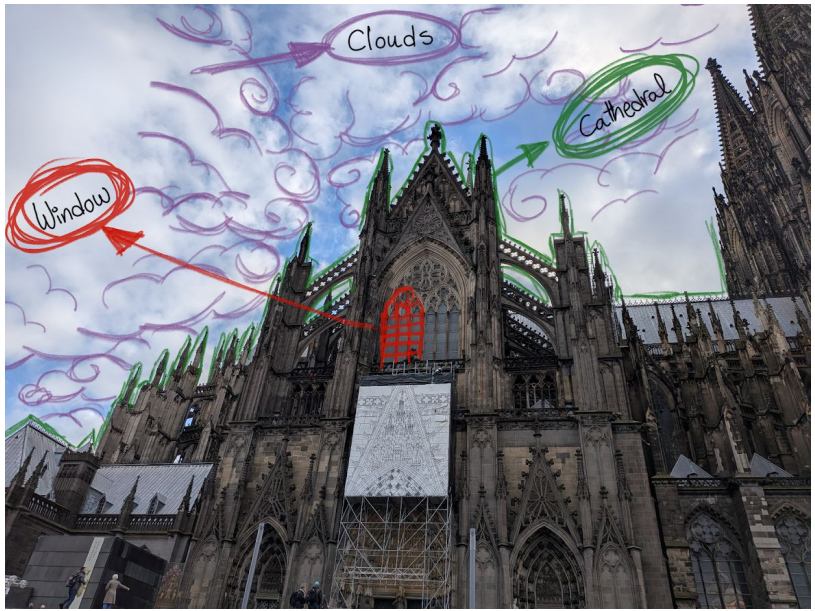
1. **COPRA: COnvolutions for PRoto-hAlo identification based on the V-net (Milletari+ '17)**

(ViT encoder + CNN decoder)

2. **VIPR: VIsion transformers for PRotohalo identification based on the UNETR (U-net Transformer, Hatamizadeh+ '21)**

Why vision transformers (ViTs)?

Semantic Segmentation uses multi-scale features



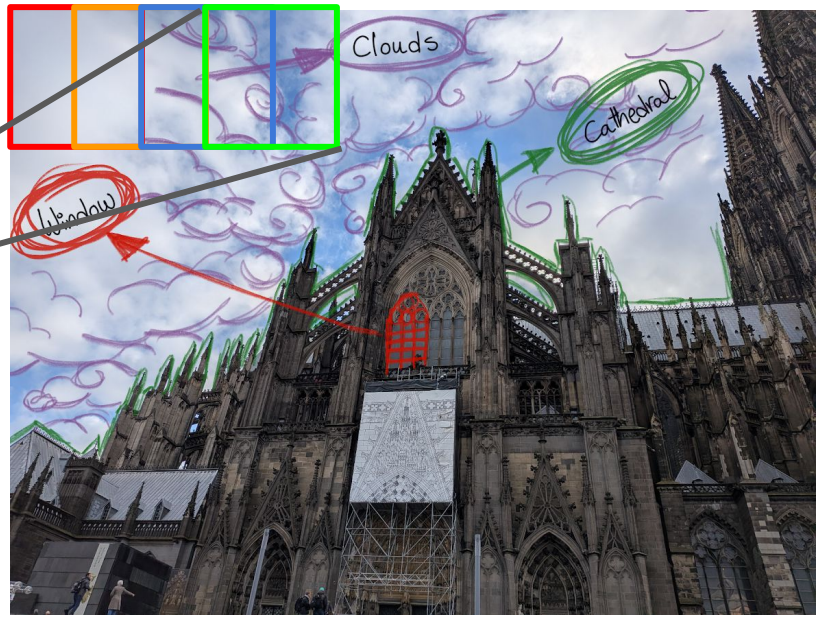
Why vision transformers (ViTs)?

Semantic Segmentation uses multi-scale features

CNN
Kernel
(learned weights)

Feature
map

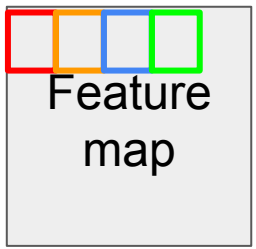
Feature map is
Built **bit by bit**



Why vision transformers (ViTs)?

Semantic Segmentation uses multi-scale features

CNN
Kernel
(learned weights)



Feature map is
Built **bit by bit**

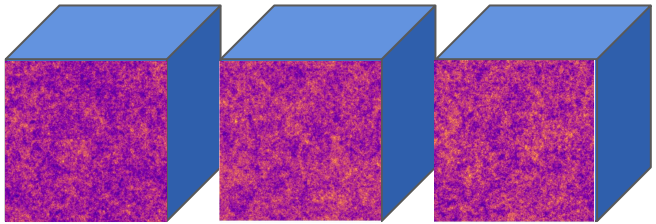
ViT
Sees the **full picture**
+
Assigns **“attention”** scores to patches, based on their relevance to the output

Training Data

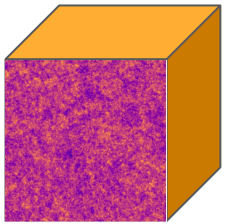
Four GADGET-2 simulations
- $L = 100$ Mpc/h

Created "output" volumes with 7 output classes (6 mass bins + "no halo" class)

Generated 3,375 overlapping samples per simulation box (via slashing them into mini-cubes of ~12.5 Mpc/h)



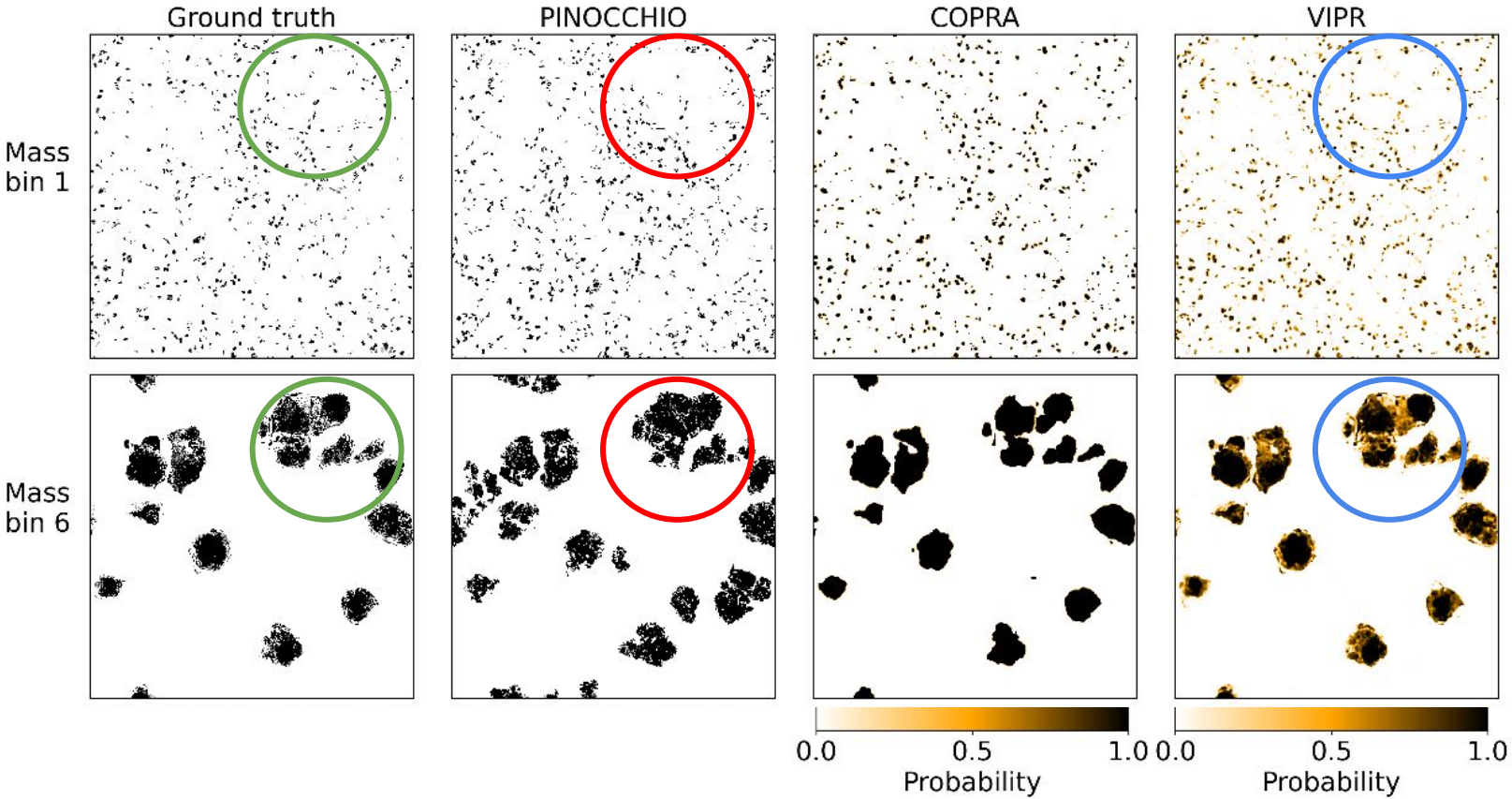
Training



Testing

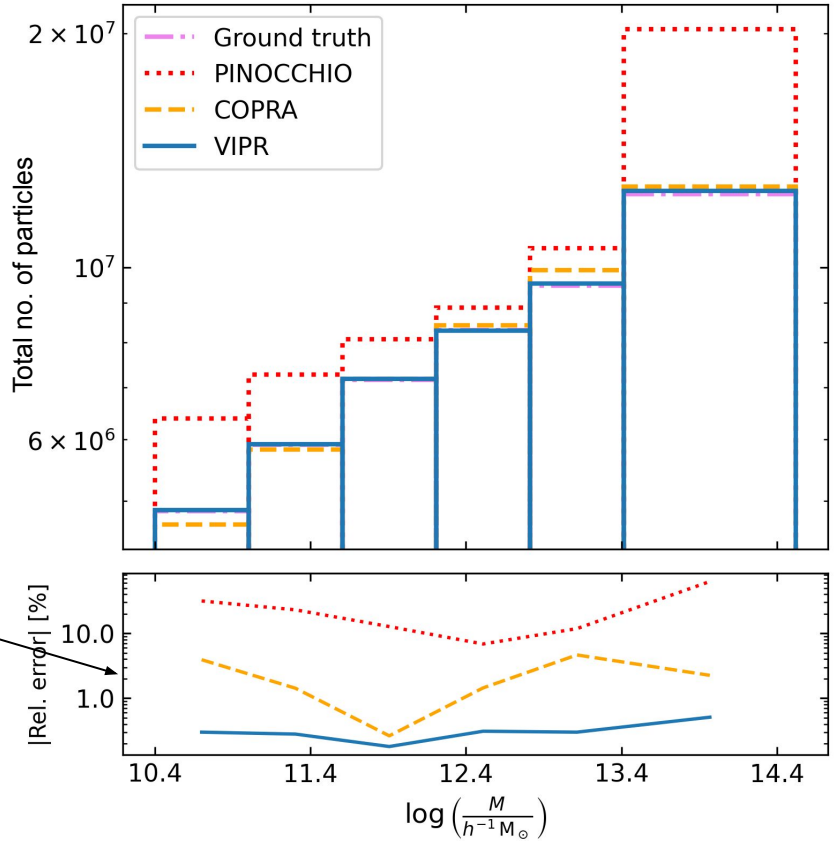
Goal: compare ML predictions, N-body simulations, and Theory-based predictions (PINOCCHIO)

Predictions on the test simulation - box level

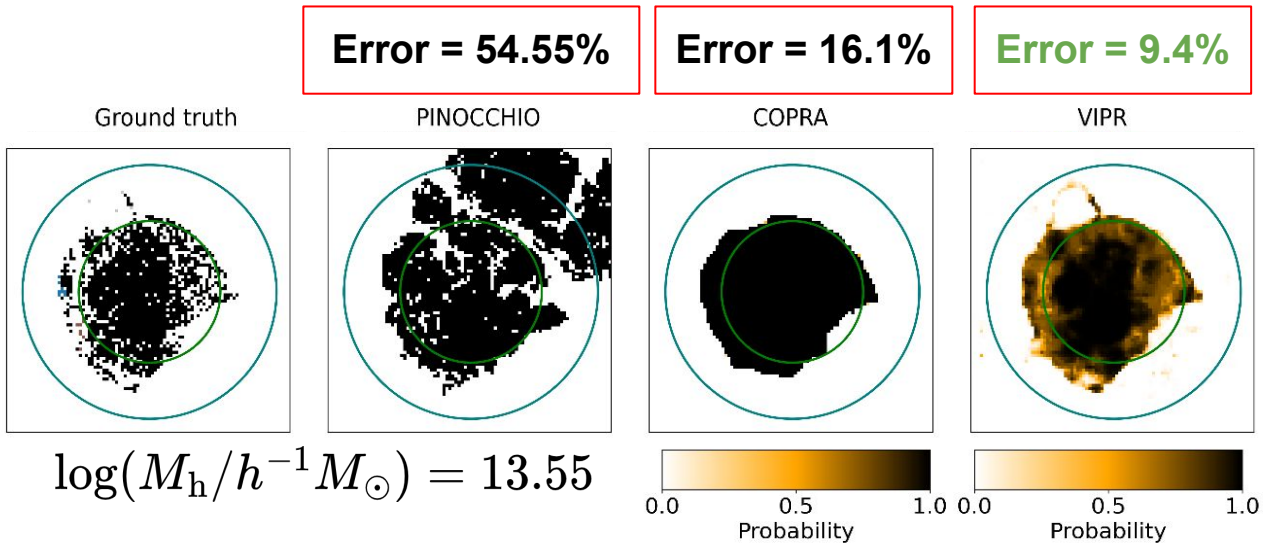


Predictions on the test simulation - box level

Consistent <1% errors on box-level

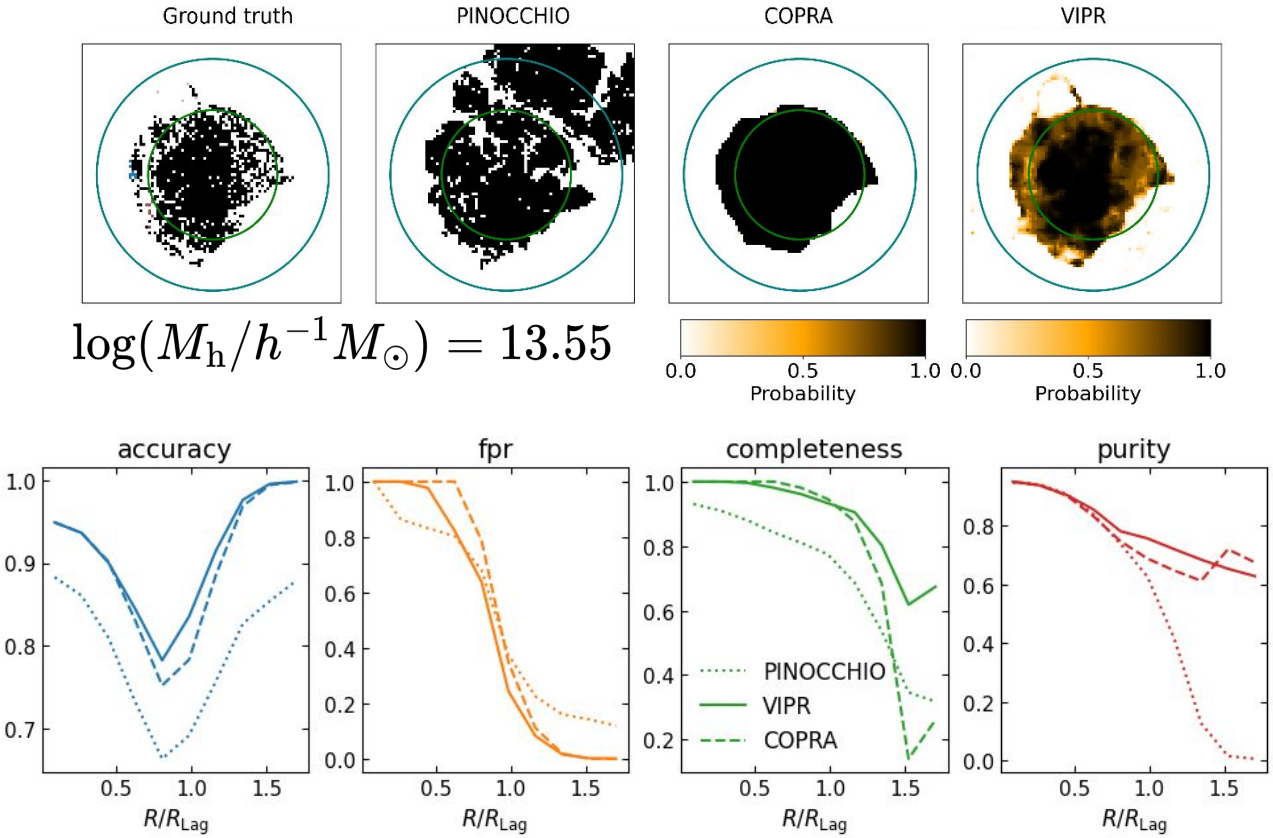


Object-level validation: halo mass



Goes to **<1%** when we average over many small protohalos of the same final mass

Object-level validation: classification quality



In Summary

ViTs seem to encode the input features and their spatial dependencies substantially better than CNNs

Given the scalability of ViT-encoders (similarly to the widely-adopted CNNs), this shows great promise for giving them more *attention* when building AI-based scientific software.

Want to learn more?

Read the paper!