Identification of Protohalos with Deep Learning

Toka Alokda, Cristiano Porciani

Les Houches Dark Universe
24/07/2025

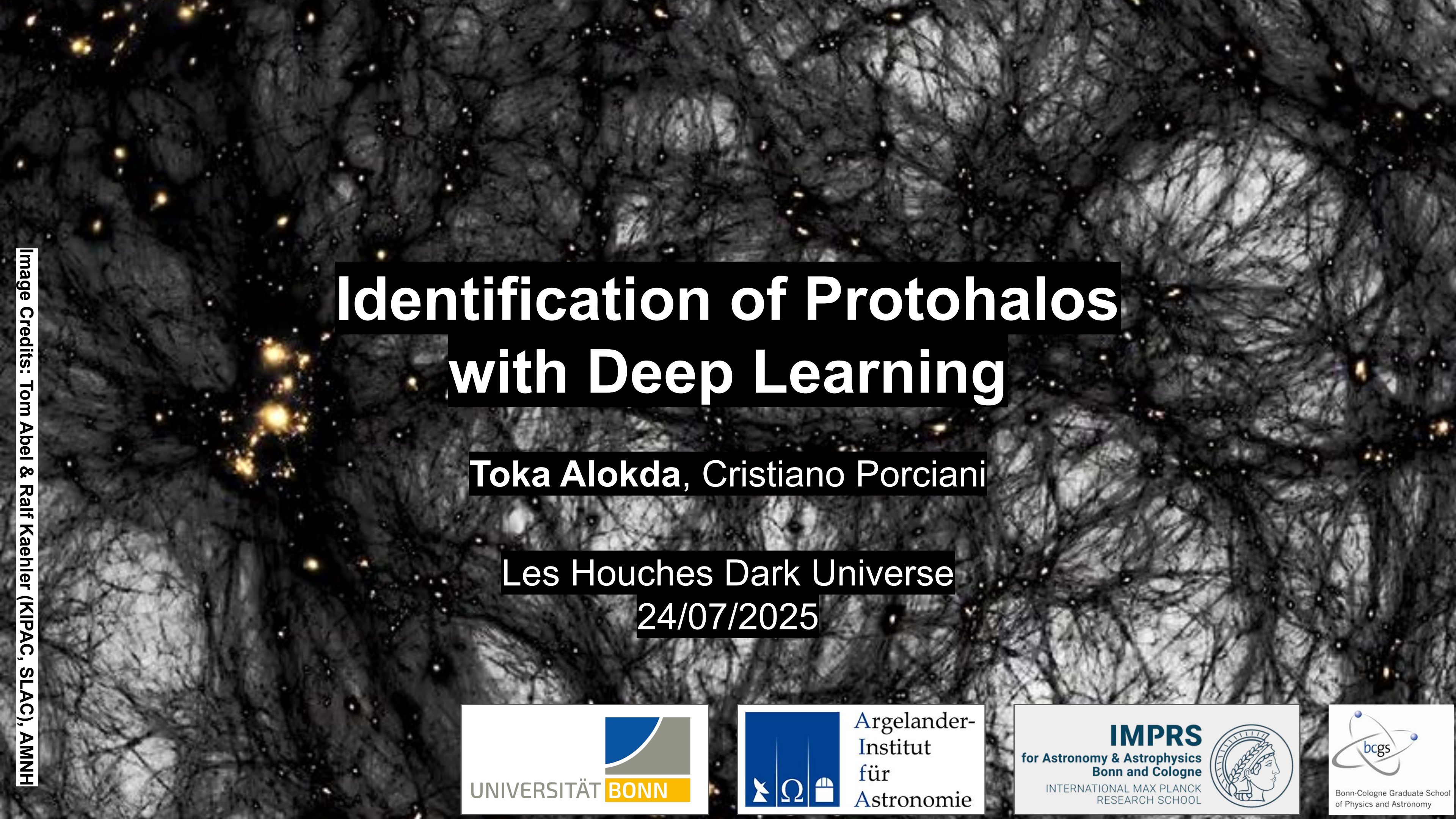
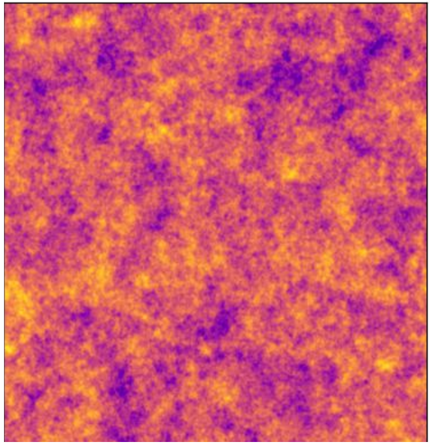


Image Credits: Tom Abel & Ralf Kaehler (KIPAC, SLAC), AMNH

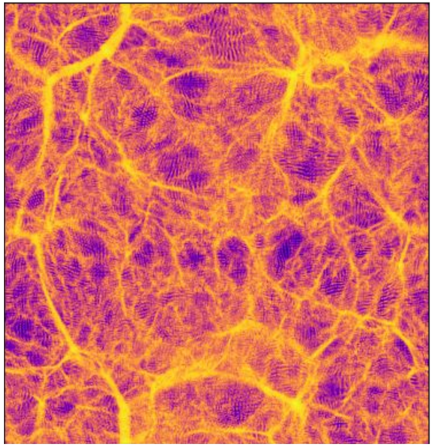
UNIVERSITÄT BONN
Argelander-Institut für Astronomie
IMPRS for Astronomy & Astrophysics Bonn and Cologne
INTERNATIONAL MAX PLANCK RESEARCH SCHOOL
bcgs
Bonn-Cologne Graduate School of Physics and Astronomy

Cosmological Structure Formation



z=99

Gravitational Collapse

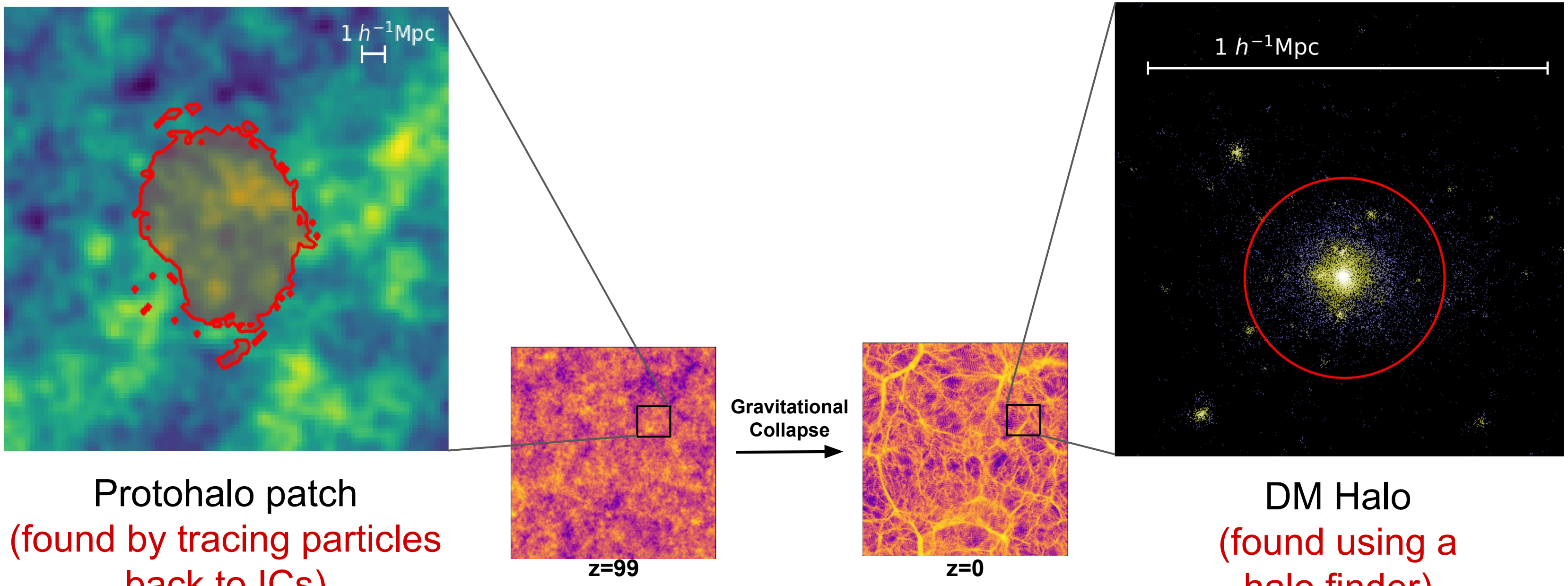


z=0

Highly non-linear

Best approximation: N-body simulations

Cosmological Structure Formation

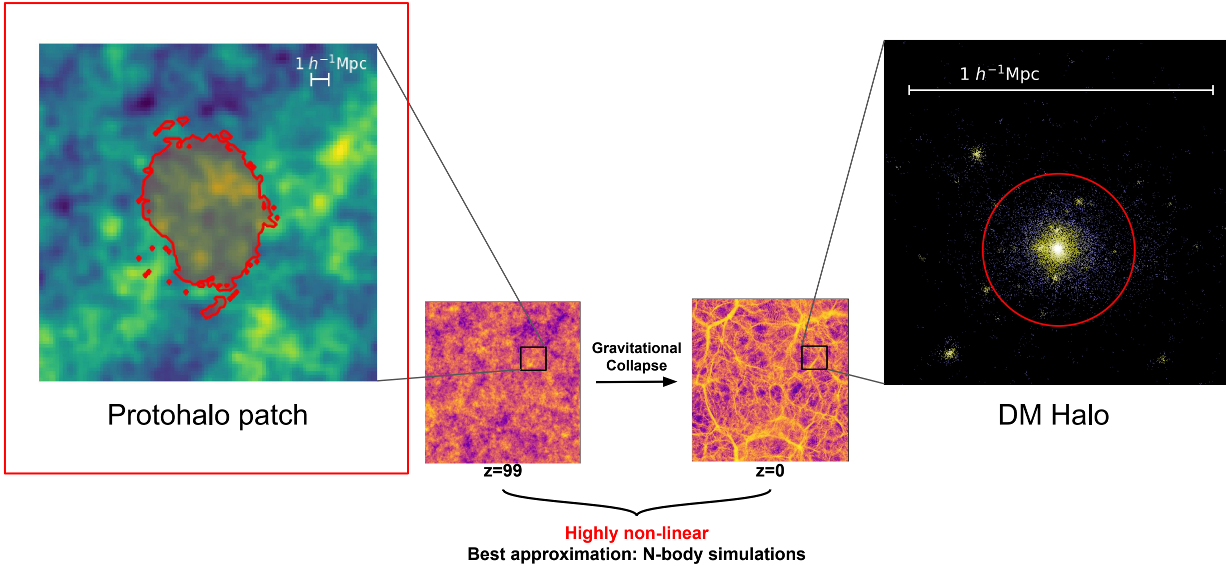


Protohalo patch
(found by tracing particles back to ICs)

DM Halo
(found using a halo finder)

Highly non-linear
Best approximation: N-body simulations

What do we know?



- Ellipsoidal regions (lumpy potato-shaped) 🥔
- Shape tends to align with the direction of max. compression of the local tidal field
- Fuzzy boundaries affected by halo definition

Question:
Can we accurately predict the formation of dark-matter halos from the initial conditions?

In Theory, one could…

- Use an ellipsoidal collapse model calibrated to simulations
(e.g. Sheth & Tormen ‘99)

Problem:
Does very poorly Predicting final halo mass for a random particle (S.D.M. White, 94’)

Question:
Can we accurately predict the formation of dark-matter halos from the initial conditions?

In Theory, one could…

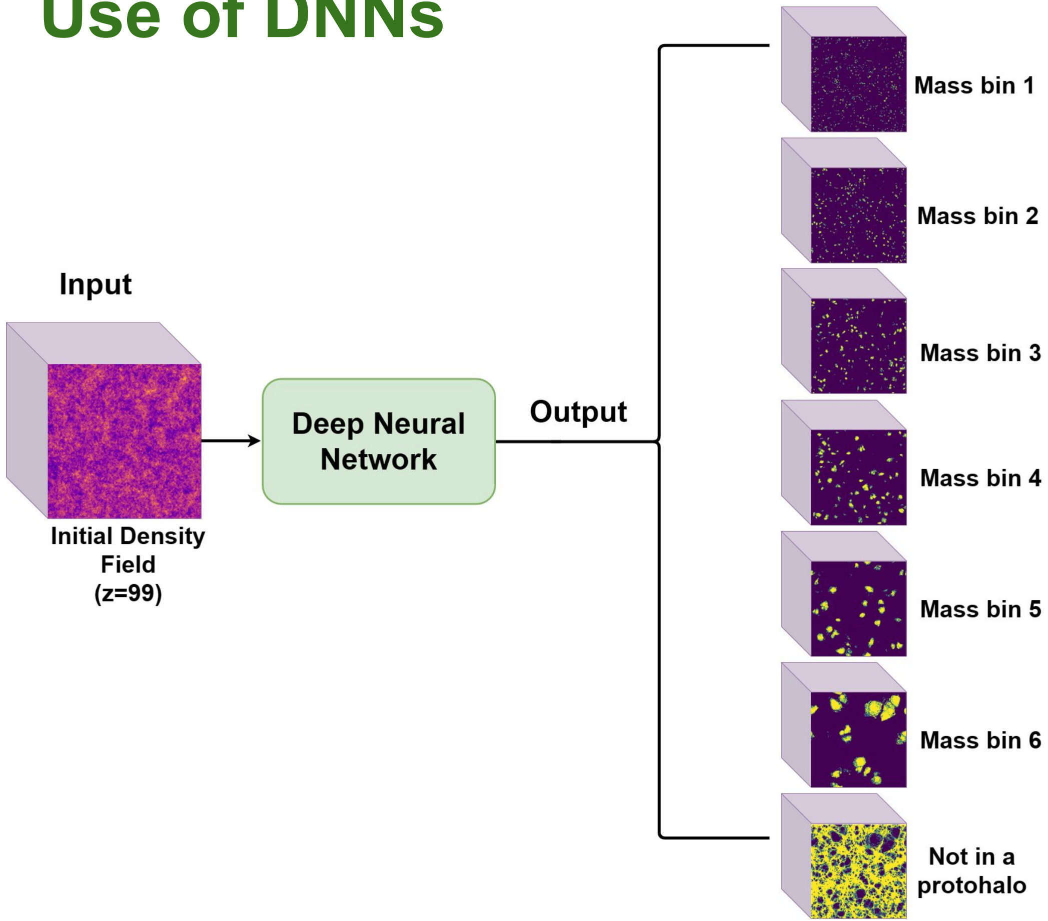
- Find the proto-halos (using e.g. peak-patch; grouping particles perturbatively…)
- Displace with LPT
- Calibrate to simulations a little bit for precision

Wonderful!

But, can we do more?

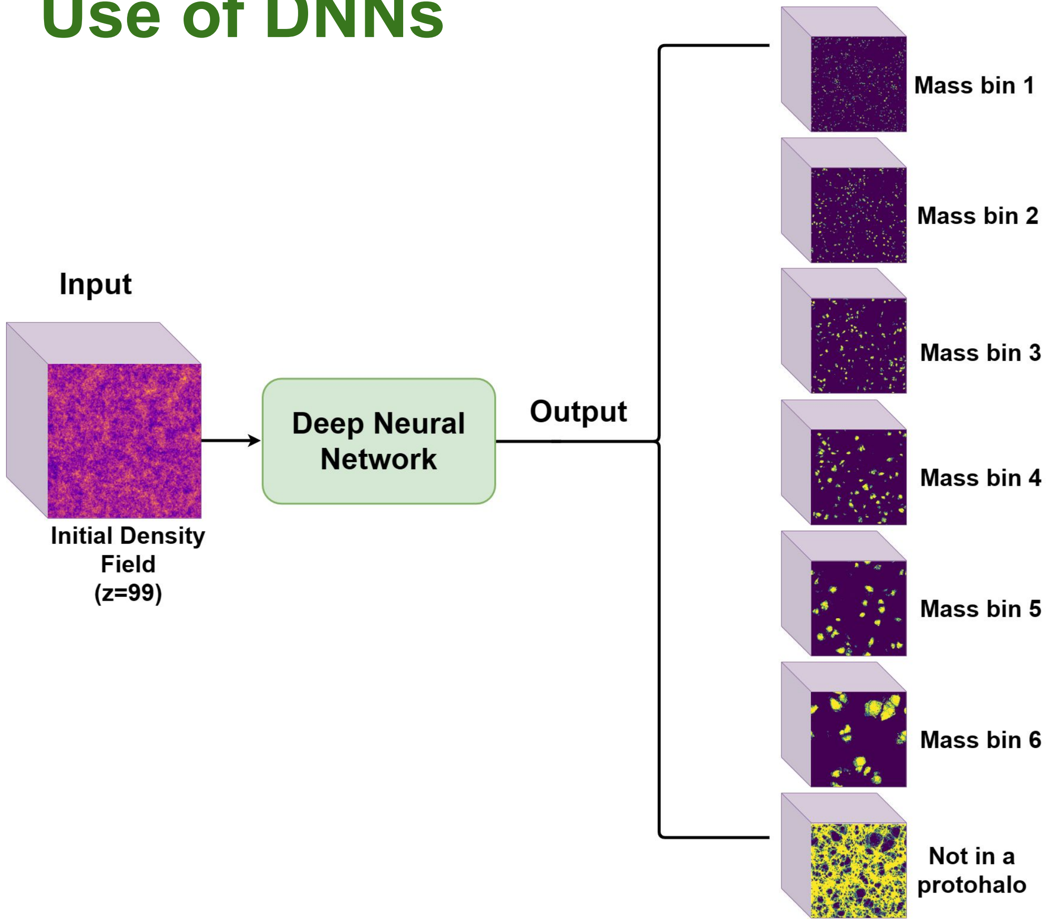
Some unresolved issues still– e.g. peakless halos (Ludlow & Porciani ‘11)

Use of DNNs



Input
Initial Density Field (z=99)
Deep Neural Network
Output
Mass bin 1
Mass bin 2
Mass bin 3
Mass bin 4
Mass bin 5
Mass bin 6
Not in a protohalo

Use of DNNs

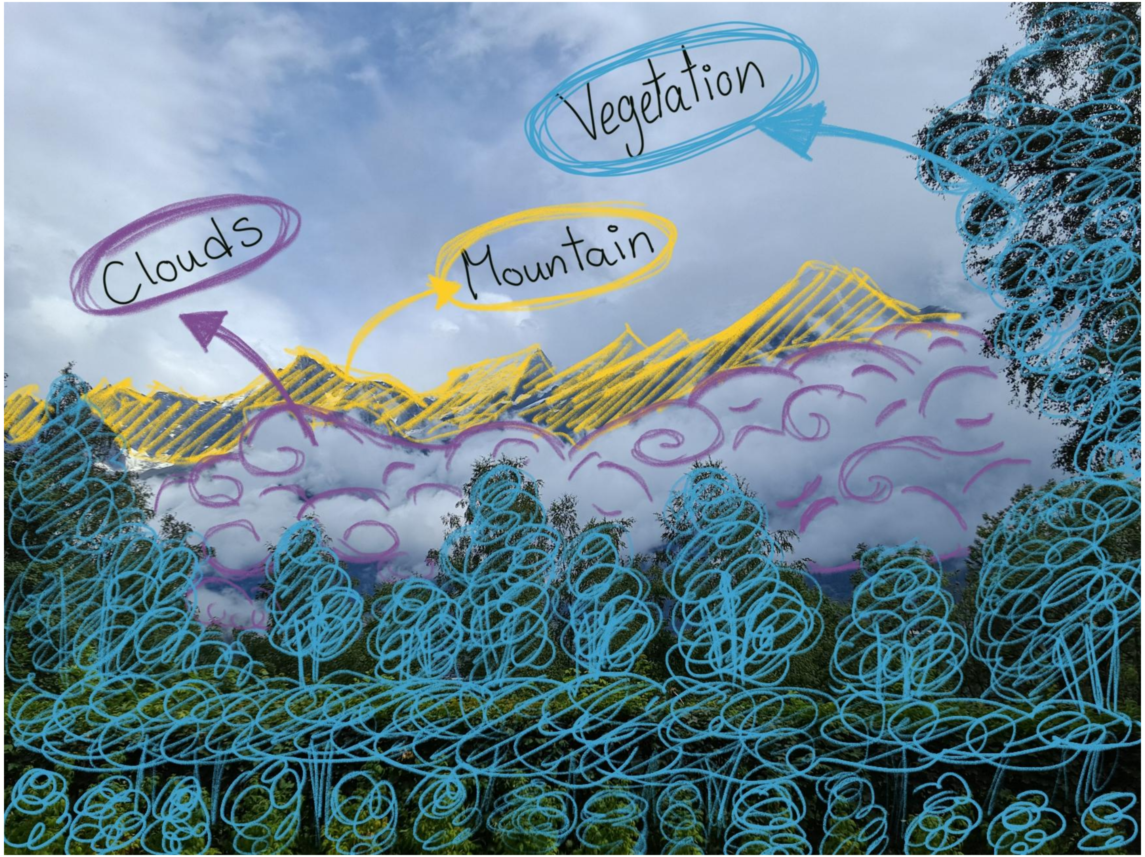


Possible Applications

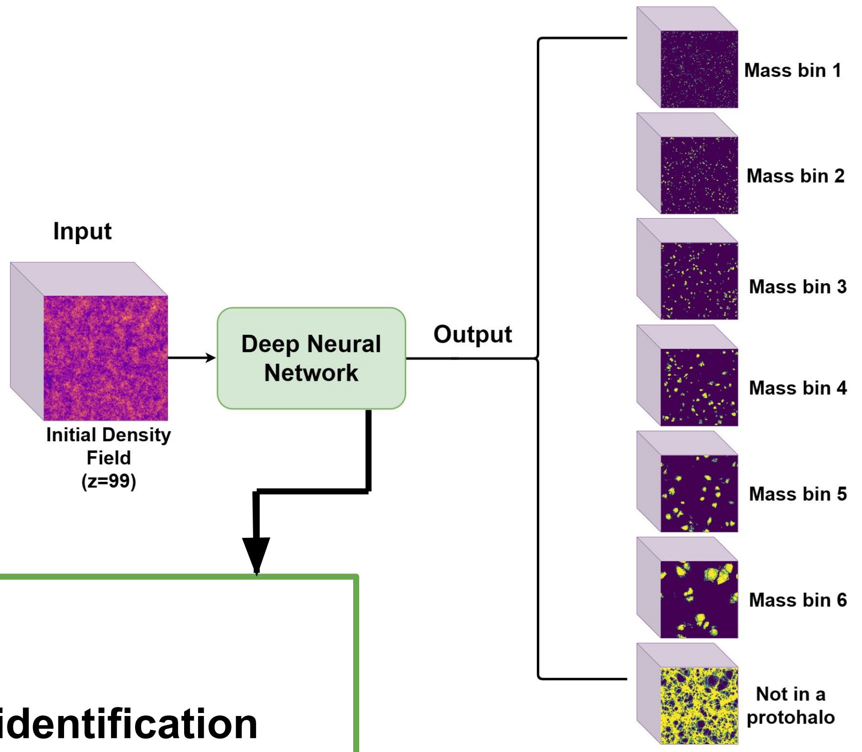
- Fast & accurate mock catalogs (e.g. Berger & Stein ‘19)
- Learning which regions could be interesting to zoom-in simulate
- (Informing priors + fast estimation of) Lagrangian bias
- Possible to get some insights on structure formation (e.g. Lucie-Smith+ ‘19 ‘24)

Semantic Segmentation

Semantic Segmentation relies on multi-scale features



Two approaches



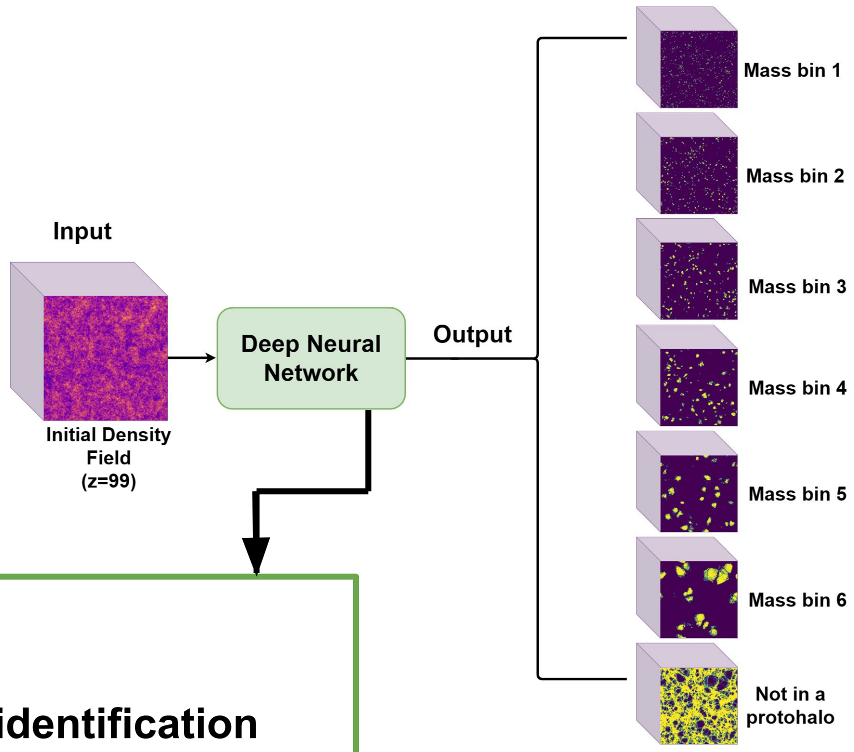
Two models 🐍 (coded up in Python)

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

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

Two approaches

"A Lucky Coincidence"
3D Medical images recieve a lot of attention from the **computer vision** community.

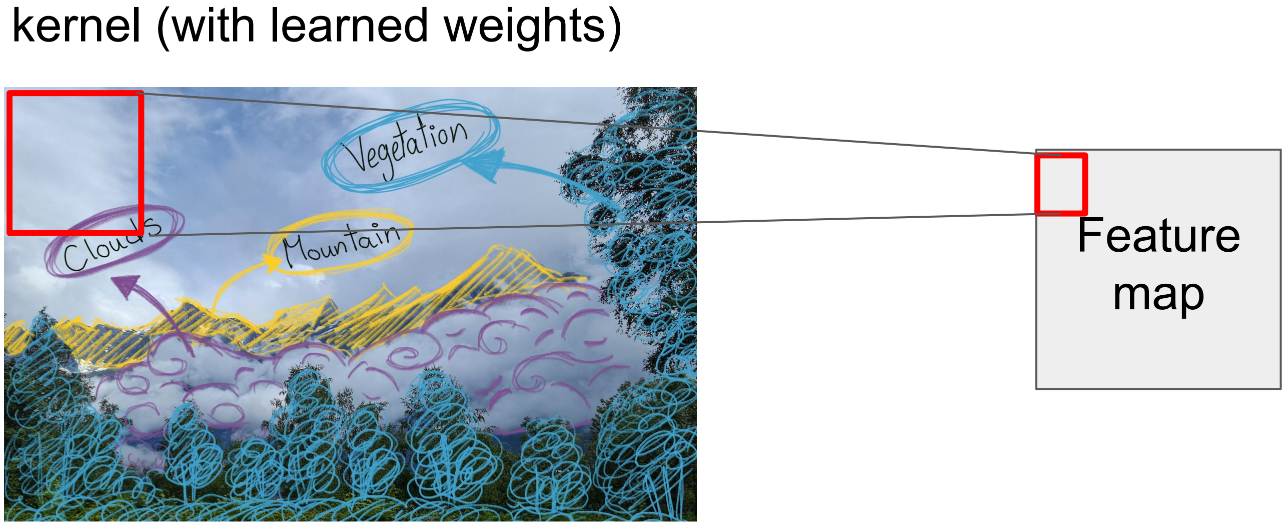


Two models 🐍 (coded up in Python)

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

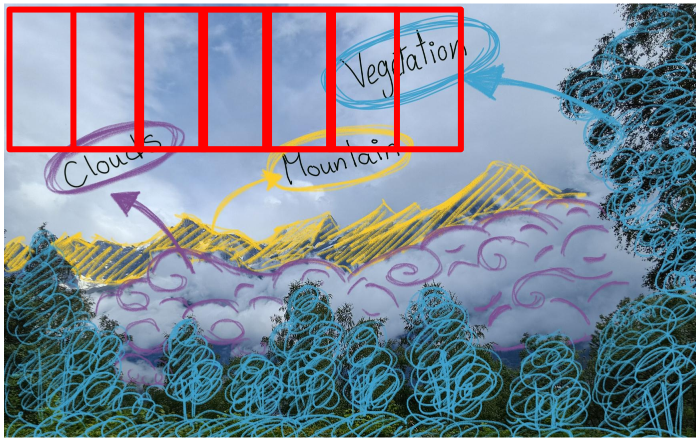
2. **VIPR: VIsion transformers for PRotohalo identification based on the UNETR (Hatamizadeh+ '21)**

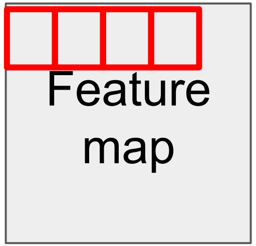
Convolutional Neural Networks (CNNs)



Convolutional Neural Networks (CNNs)

kernel (with learned weights)

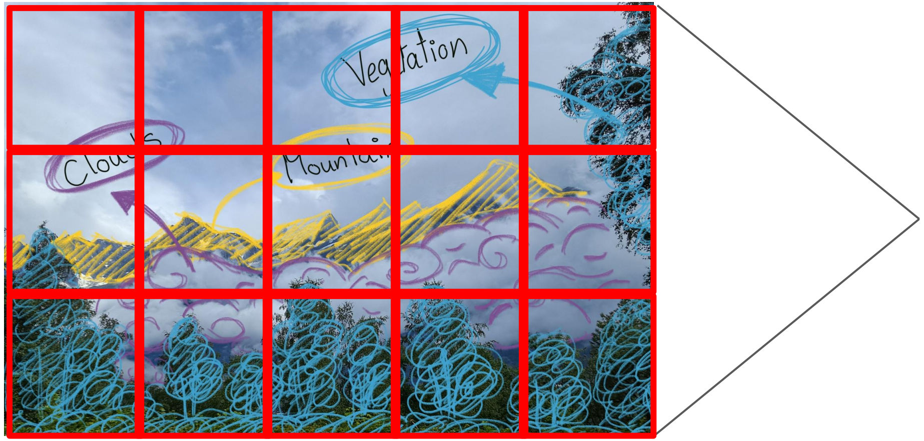




Feature map
Built bit by bit

Advantage: very efficient to train + quite accurate

Vision Transformers (ViTs) 🦅



Looks at all the bits at once
in context of each other
+
Assigns “attention” scores to each one based on its importance

Advantage: very accurate given enough training data
+ learns large-scale dependencies

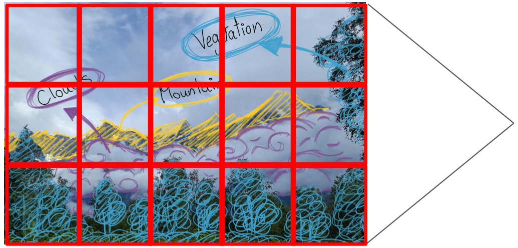
Vision Transformers (ViTs) 🦅

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of aidan@cs.tor
Łukasz Kaiser*
Illia Po illia.polosu

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com



Looks at all the bits at once
in context of each other
+
Assigns "attention" scores to each one based on its importance

Advantage: very accurate given enough training data
+ learns large-scale dependencies

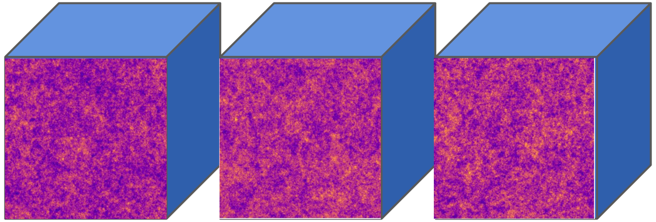
Training Data

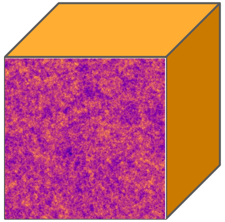
Four `GADGET` simulations

- $N = 512^3$
- $L = 100$ Mpc$/h$
- 2LPT initial conditions
- Run from z=99 to z=0
- AHF: min. 40 particles in R_{200}

Max. halo mass $\sim 10^{14.4} h^{-1} M_\odot$

Min. halo mass $= 10^{10.416} h^{-1} M_\odot$





Training

Testing

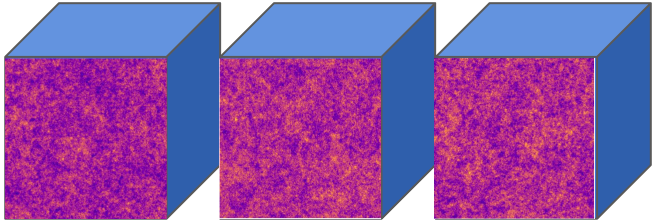
Training Data

Four `GADGET` simulations

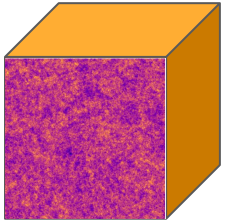
- $N = 512^3$
- $L = 100$ Mpc/h
- 2LPT initial conditions
- Run from z=99 to z=0
- AHF: min. 40 particles in R_{200}

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

→ **Generated 3,375 overlapping samples Of size 64^3 per simulation**



Training



Testing

Training Data

Four `GADGET` simulations

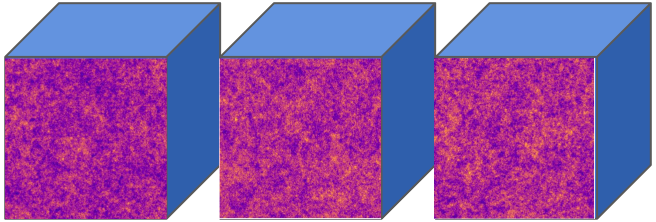
- $N = 512^3$
- $L = 100\ \mathrm{Mpc}/h$
- 2LPT initial conditions
- Run from z=99 to z=0
- AHF: min. 40 particles in R_{200}

→

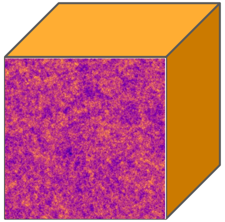
Created "output" volumes with 7 classes (6 mass bins + "no halo" class)

→

Generated 3,375 overlapping samples Of size 64^3 per simulation



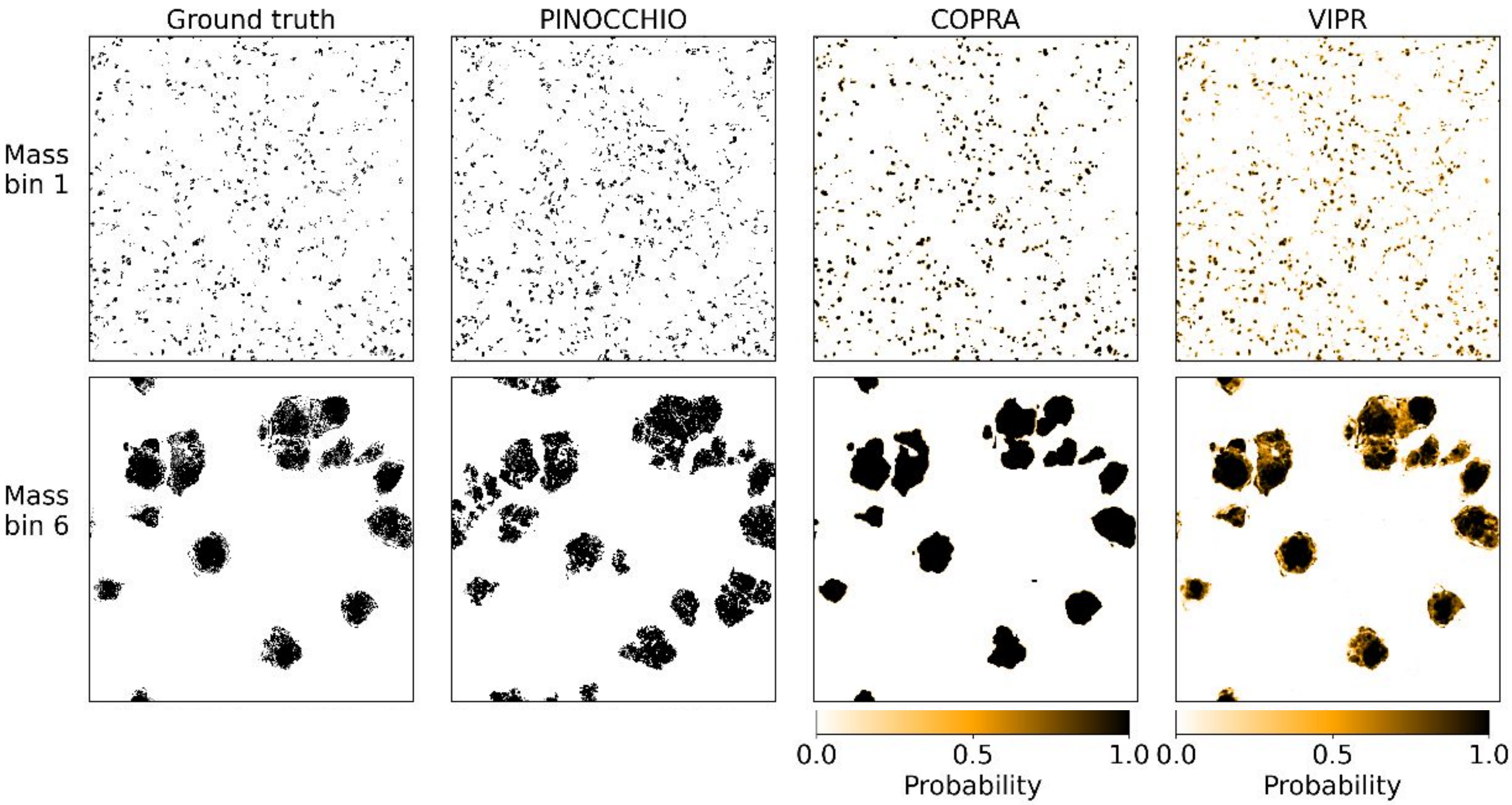
Training



Testing

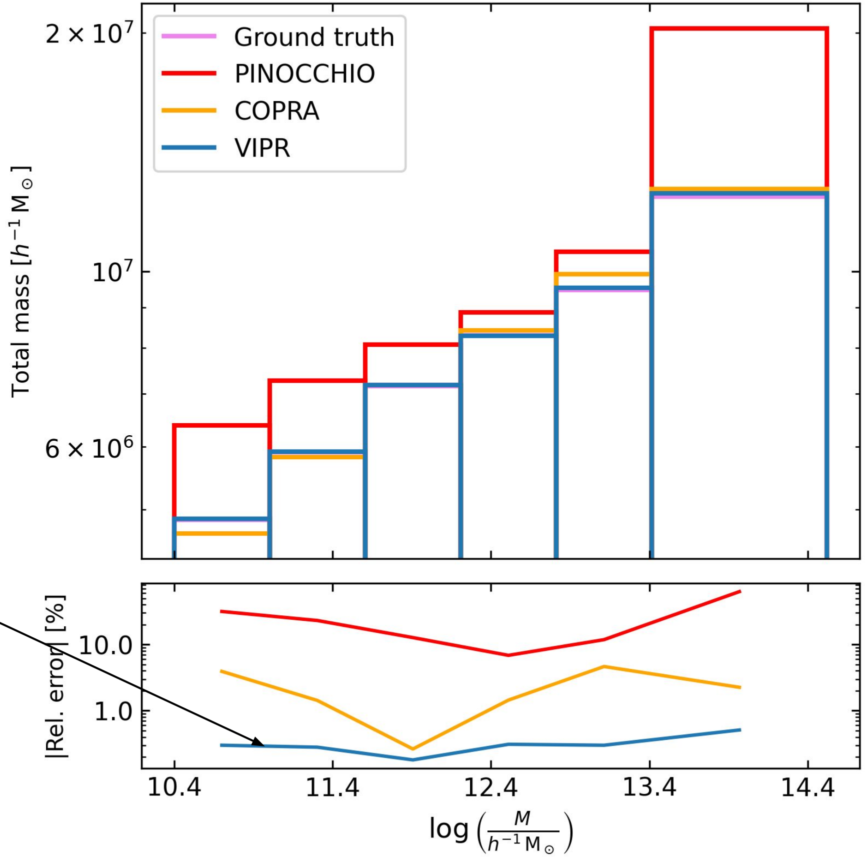
Also generated a catalog with 3LPT-based code **`PINOCCHIO`** (Monaco+ '02)

Predictions on the test simulation

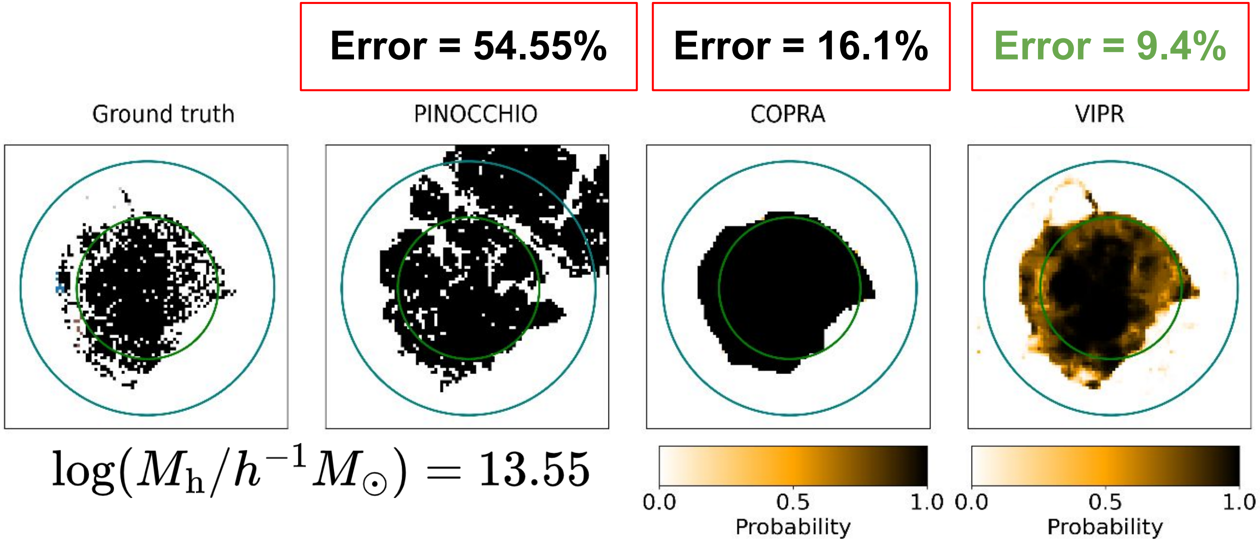


Predictions on the test simulation

Consistent <1% errors on box-level



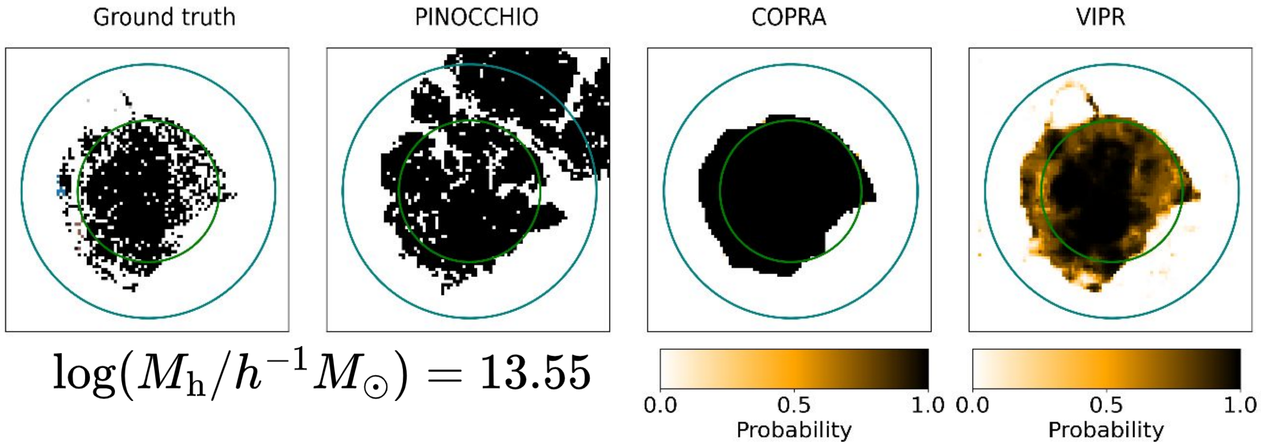
Object-level validation: halo mass



$$\log(M_{\rm h}/h^{-1}M_{\odot}) = 13.55$$

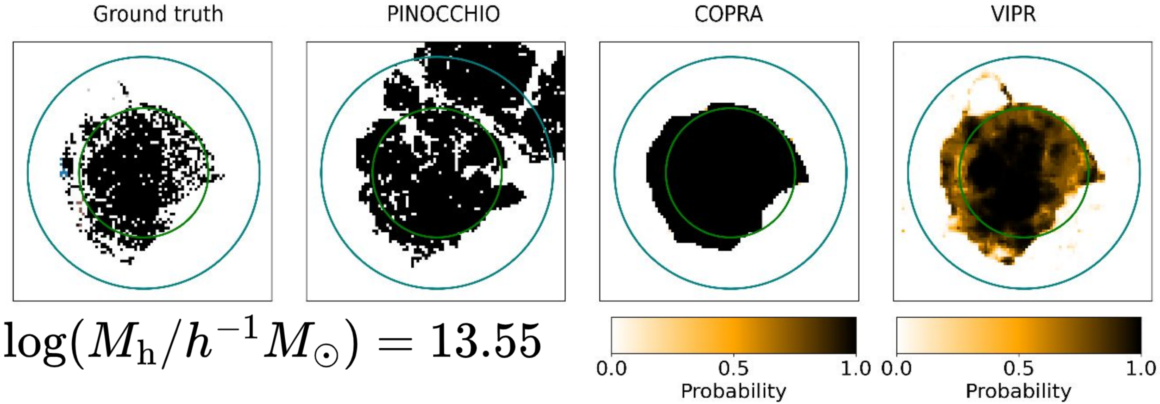
Object-level validation: halo mass

Error = 54.55% | Error = 16.1% | Error = 9.4%

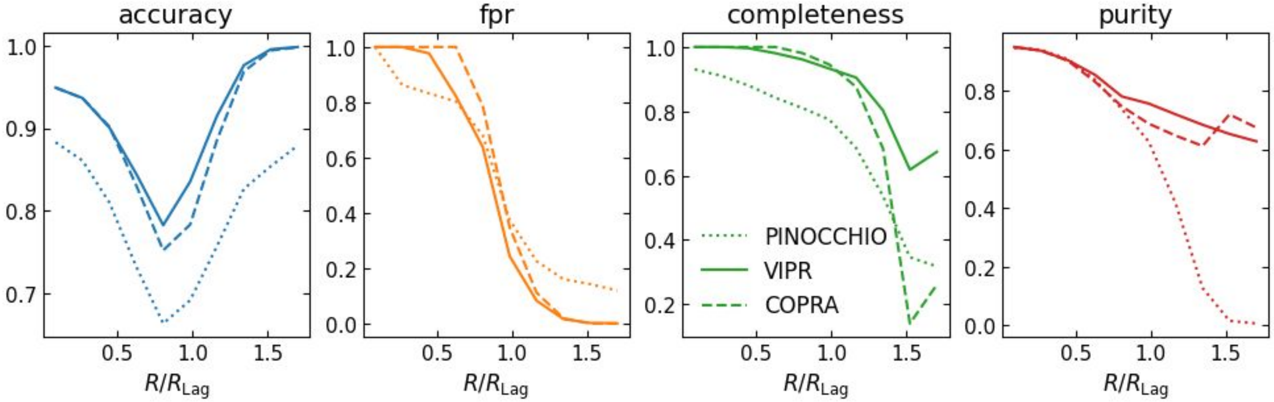


Goes to **<1%** when we average over many small protohalos of the same final mass

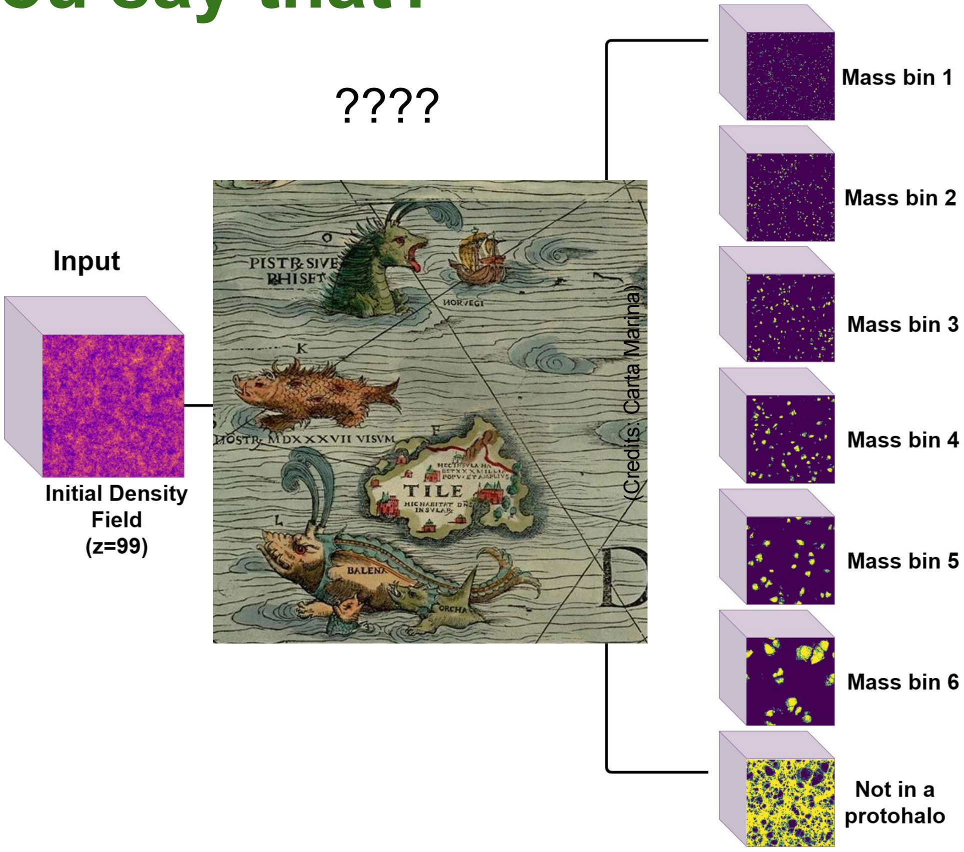
Object-level validation: classification quality



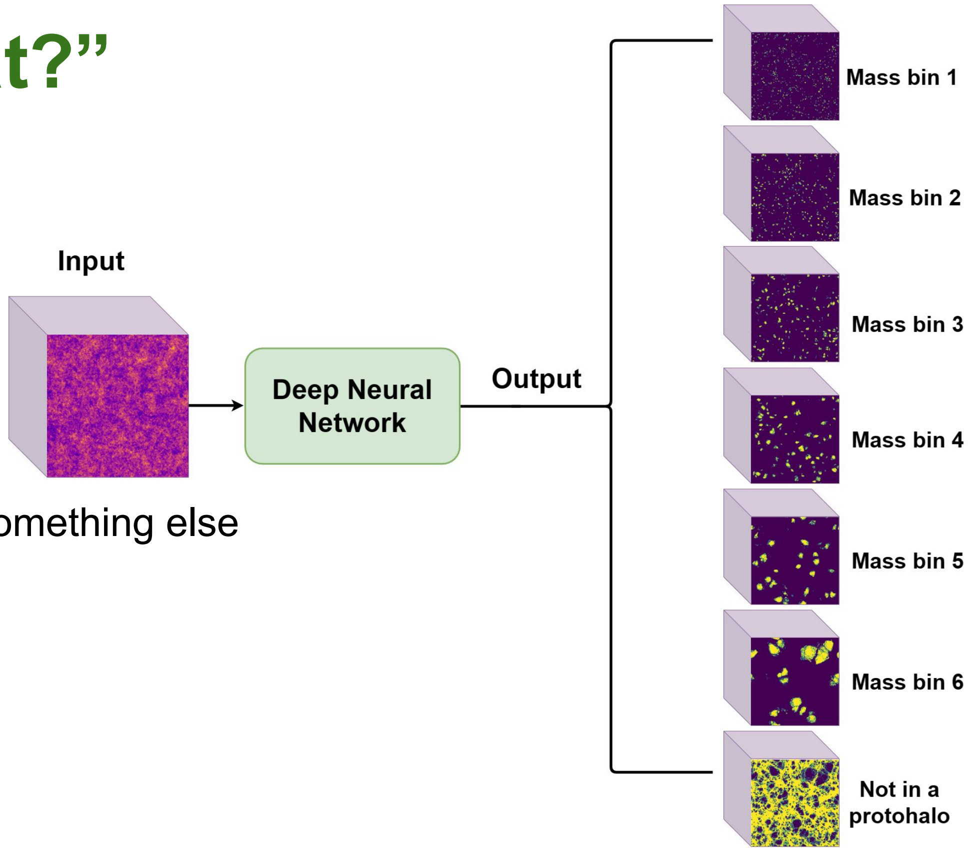
$$\log(M_\mathrm{h}/h^{-1}M_\odot) = 13.55$$



“Why did you say that?”



“Why did you say that?”



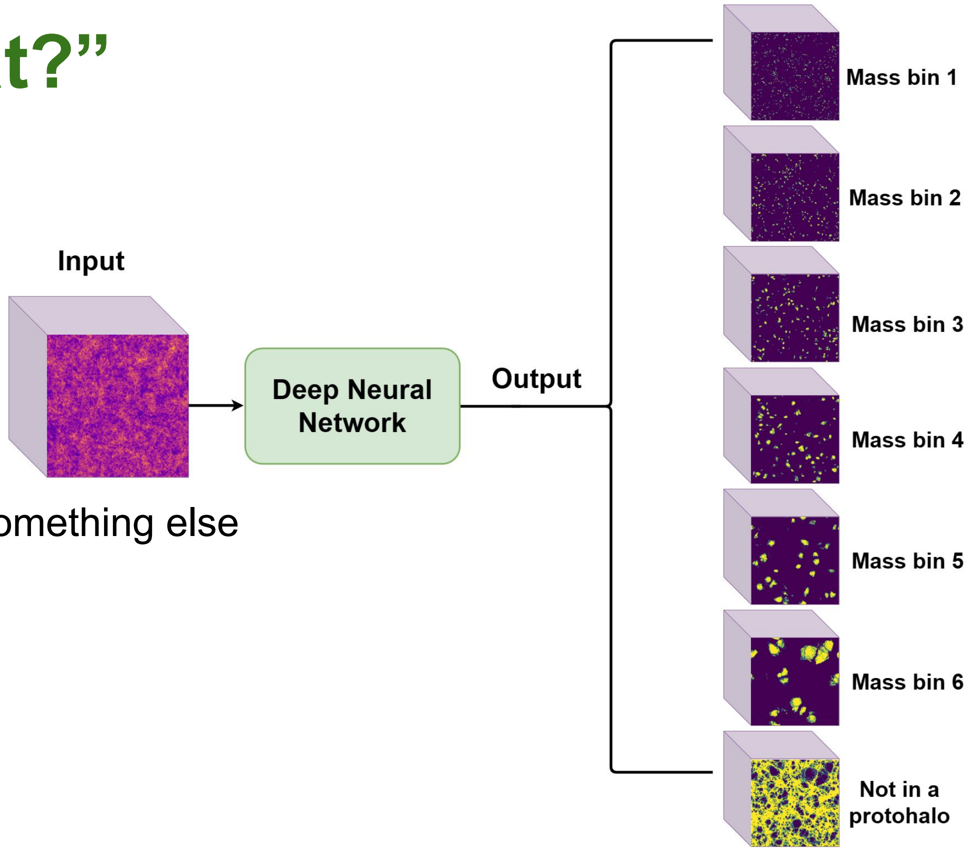
Something else

“Why did you say that?”

e.g. magnitude of the tidal tensor

$$T_{ij} = \frac{\partial^2\Phi}{\partial x_i \partial x_j} - \frac{1}{3}\nabla^2\Phi\,\delta_{ij}$$

$$T = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3}(T_{ij})^2}$$



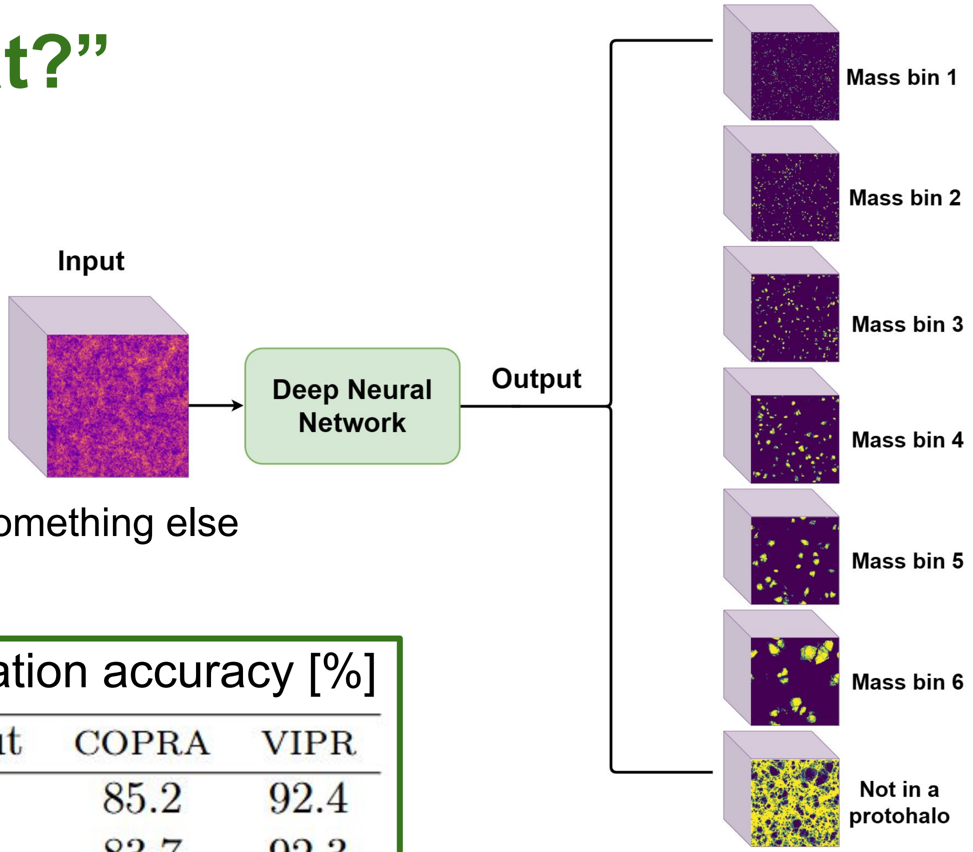
Something else

“Why did you say that?”

e.g. magnitude of the tidal tensor

$$T_{ij} = \frac{\partial^2 \Phi}{\partial x_i \partial x_j} - \frac{1}{3} \nabla^2 \Phi \, \delta_{ij}$$

$$T = \sqrt{\sum_{i=1}^{3} \sum_{j=1}^{3} (T_{ij})^2}$$

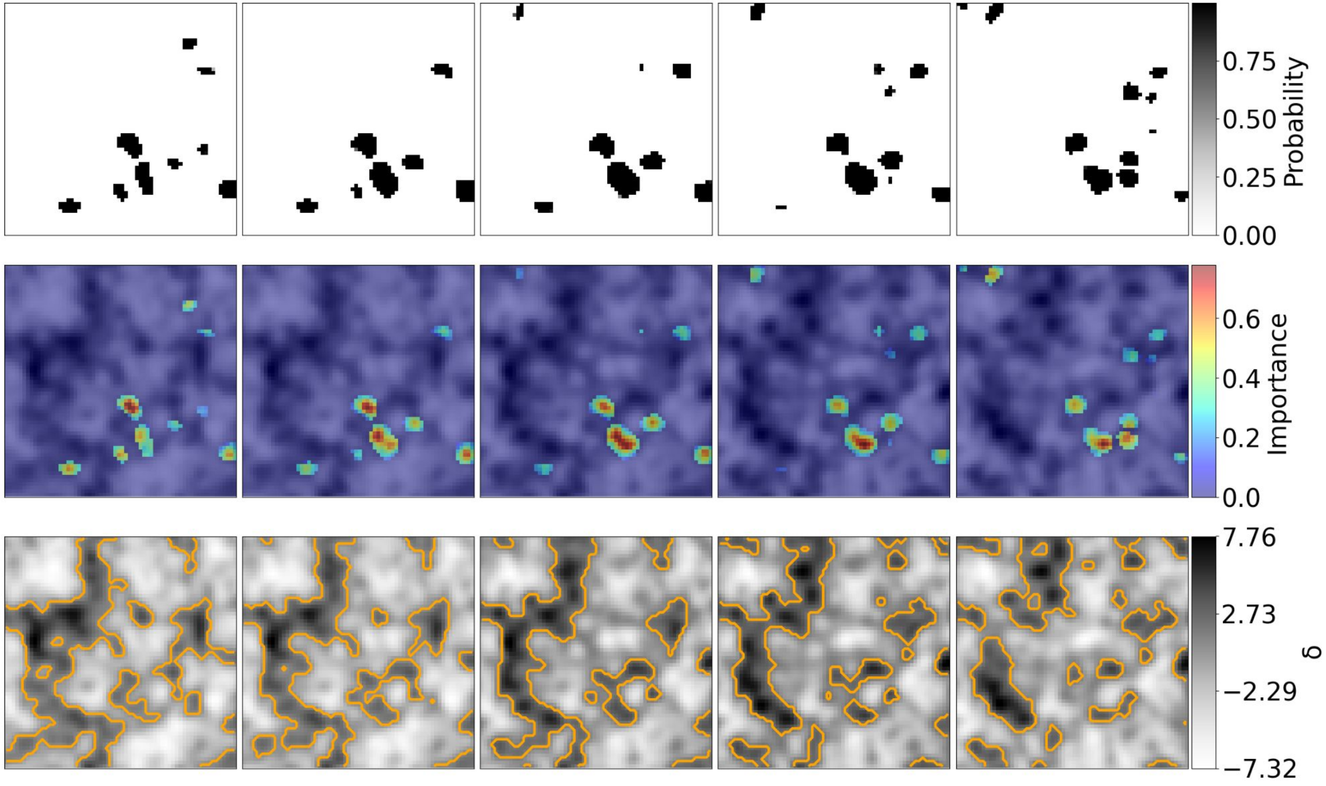


Something else

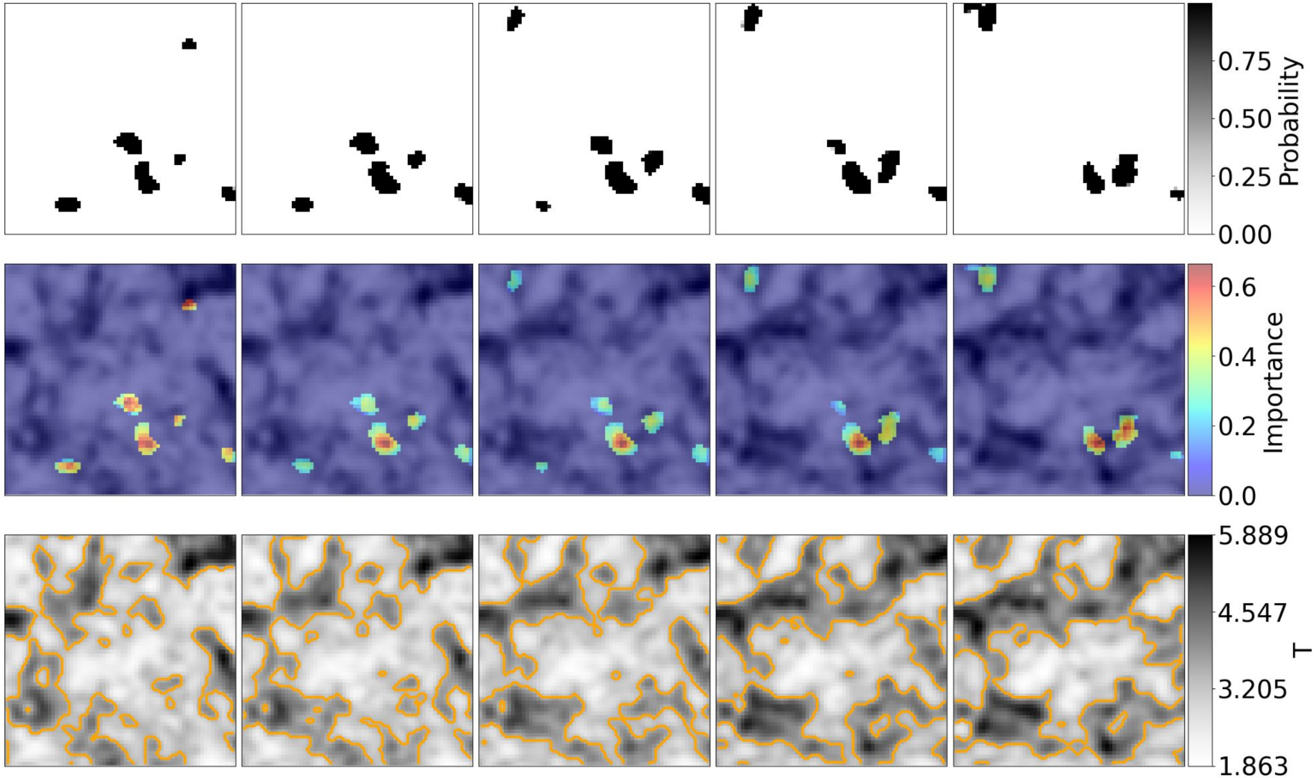
Validation accuracy [%]

Input	COPRA	VIPR
δ	85.2	92.4
T	83.7	92.3
$\delta \wedge T$	85.8	93.1

“Why did you say that?”
(Selvaraju+ ‘17) Grad-CAM



“Why did you say that?”
(Selvaraju+ ‘17) Grad-CAM



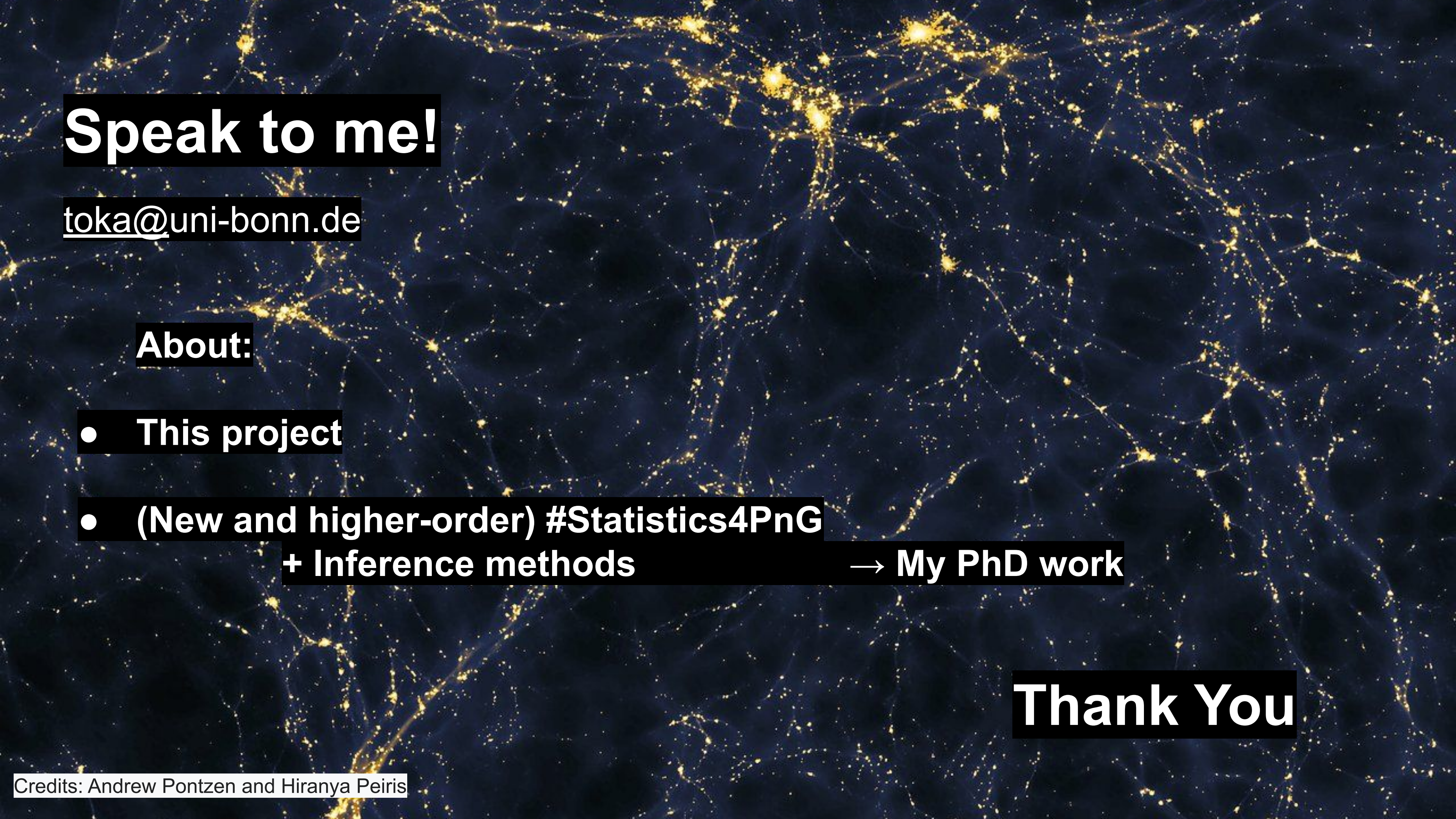
Summary and Conclusions

- **ViTs seem to learn much better than CNNs from Simulation cubes**
  1- It may be nice to invest in training them for useful tasks given their scalability.

  2- <1% error for the total mass, <10% error for individual halo masses.

- **Neural Networks seem like a promising alternative to standard tools (PINOCCHIO).**

- **May be possible to get some insights into what the model "looks" at, and which quantities play a bigger role in halo formation.**



Speak to me!

toka@uni-bonn.de

About:

- **This project**
- **(New and higher-order) #Statistics4PnG + Inference methods** → **My PhD work**

Thank You

Credits: Andrew Pontzen and Hiranya Peiris