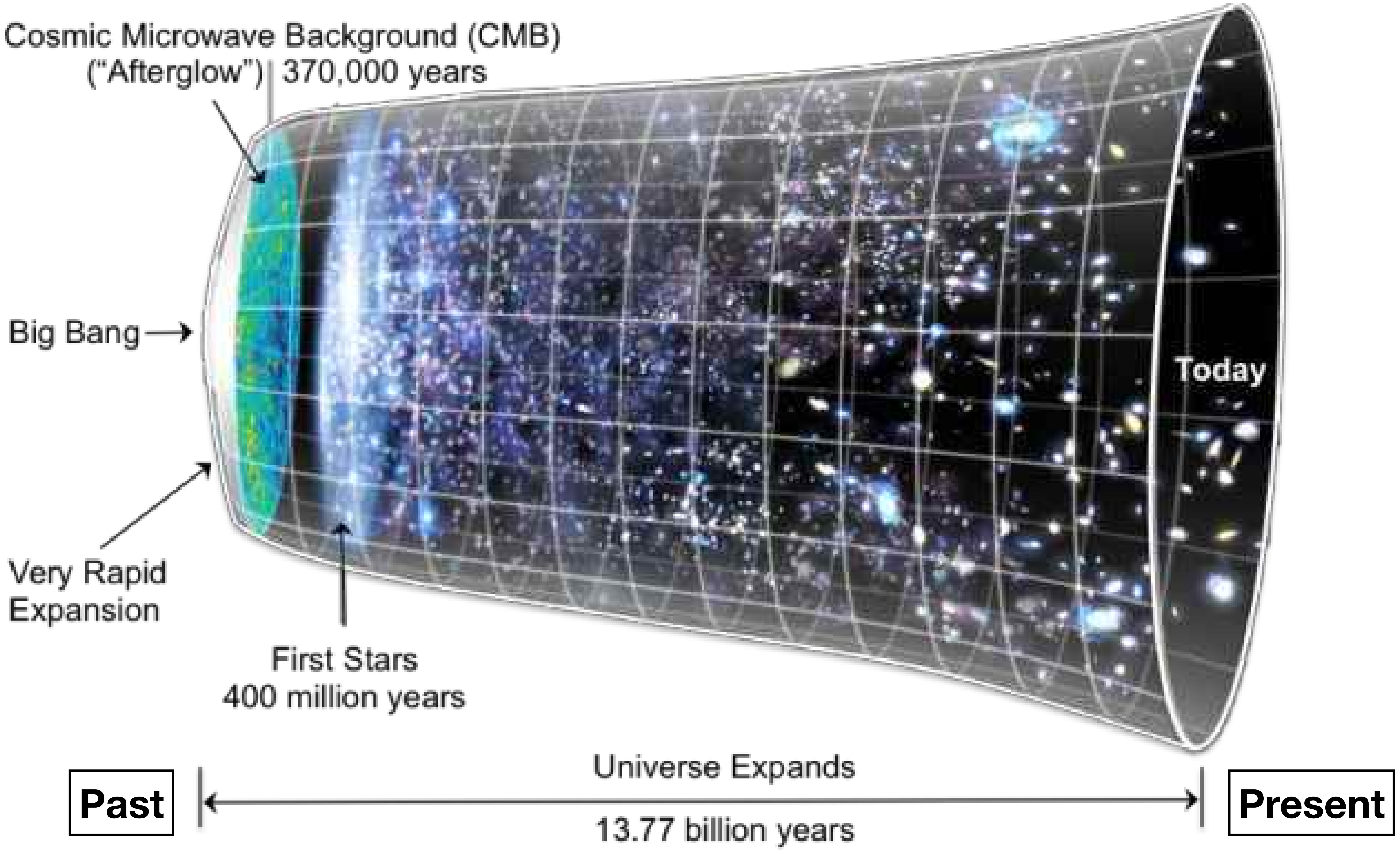
Deep Learning Application for Reconstruction of the Large-Scale Structure of the Universe

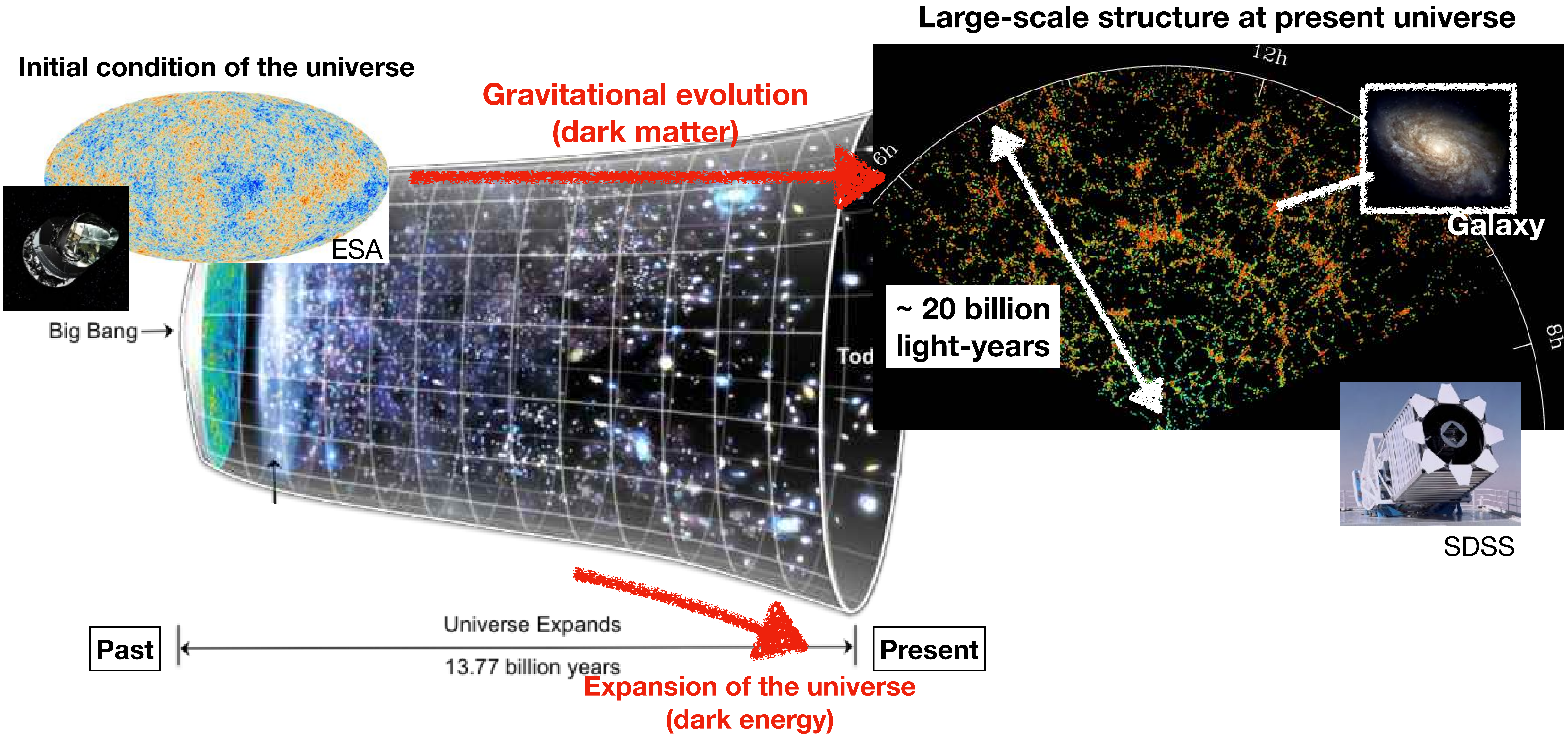
Kana Moriwaki (UTokyo UTAP/RESCEU)

ipi seminar
2022/10/19

Evolution of the Universe and the Large-Scale Structure of the Universe



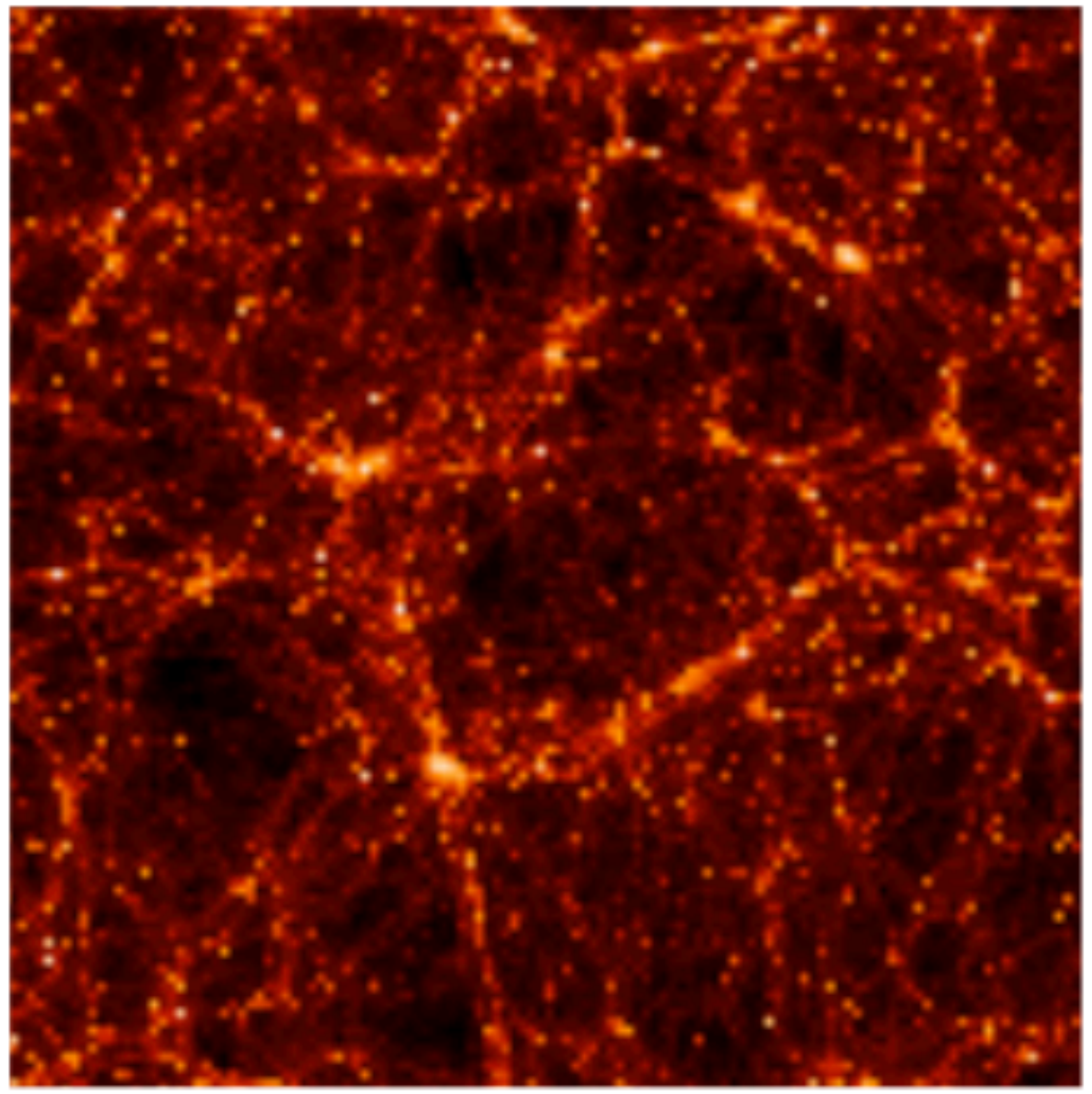
Evolution of the Universe and the Large-Scale Structure of the Universe



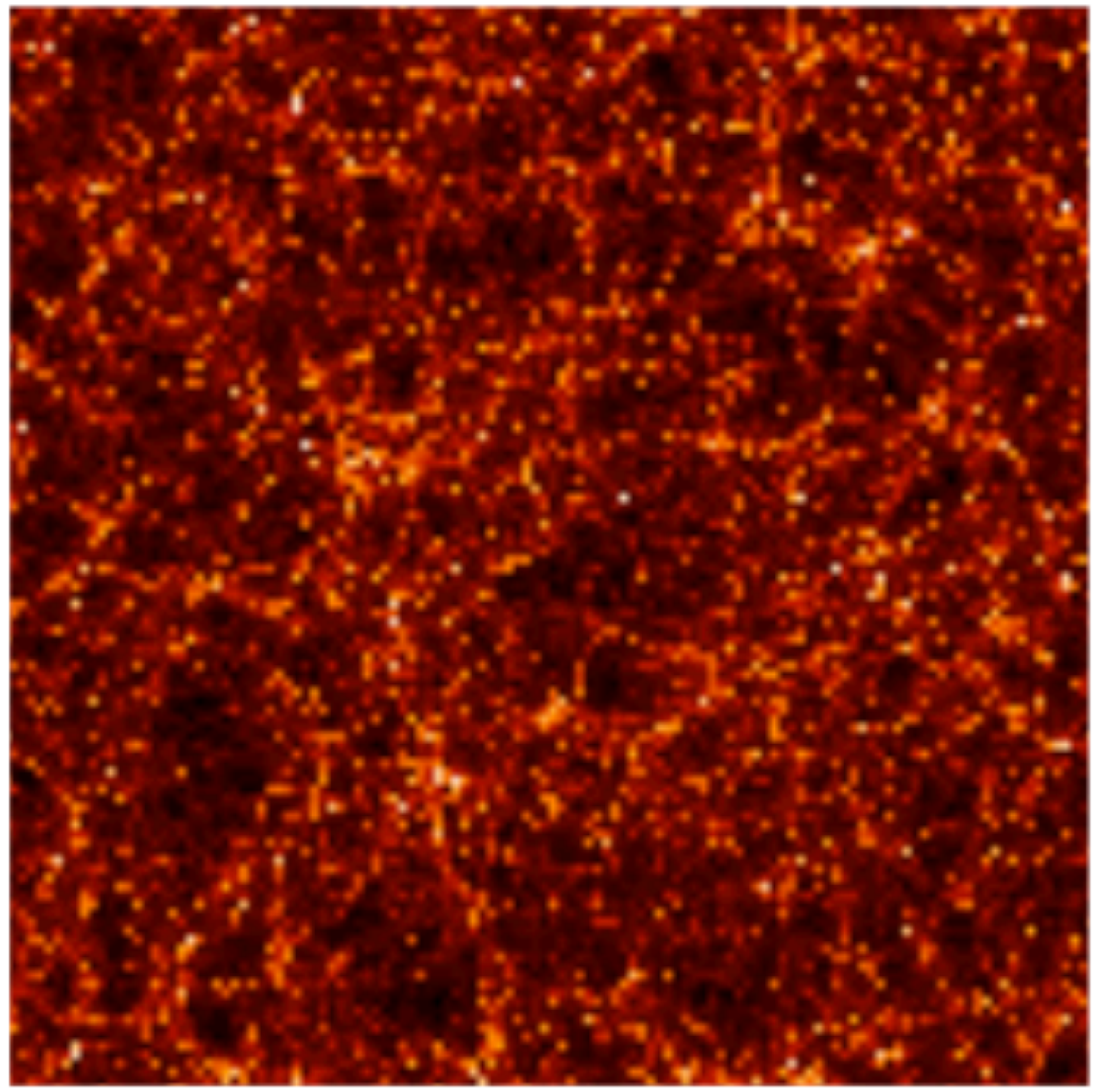
Evolution of the Universe and the Large-Scale Structure of the Universe

- The large-scale structure tells us about:
  - Contents of the universe including dark matter and dark energy
  - Initial condition of the universe
  - etc.

30% matter
70% dark energy



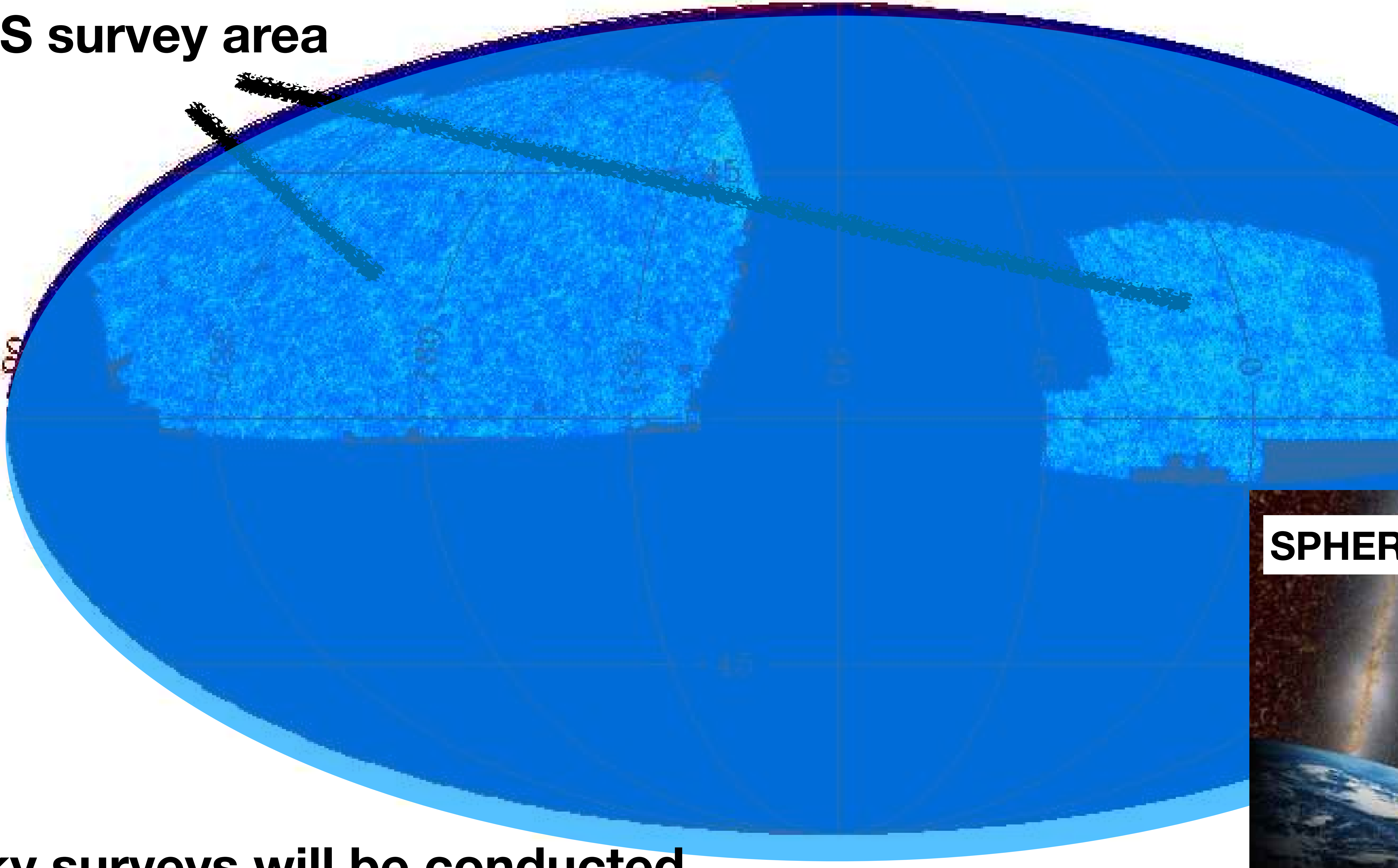
100% matter

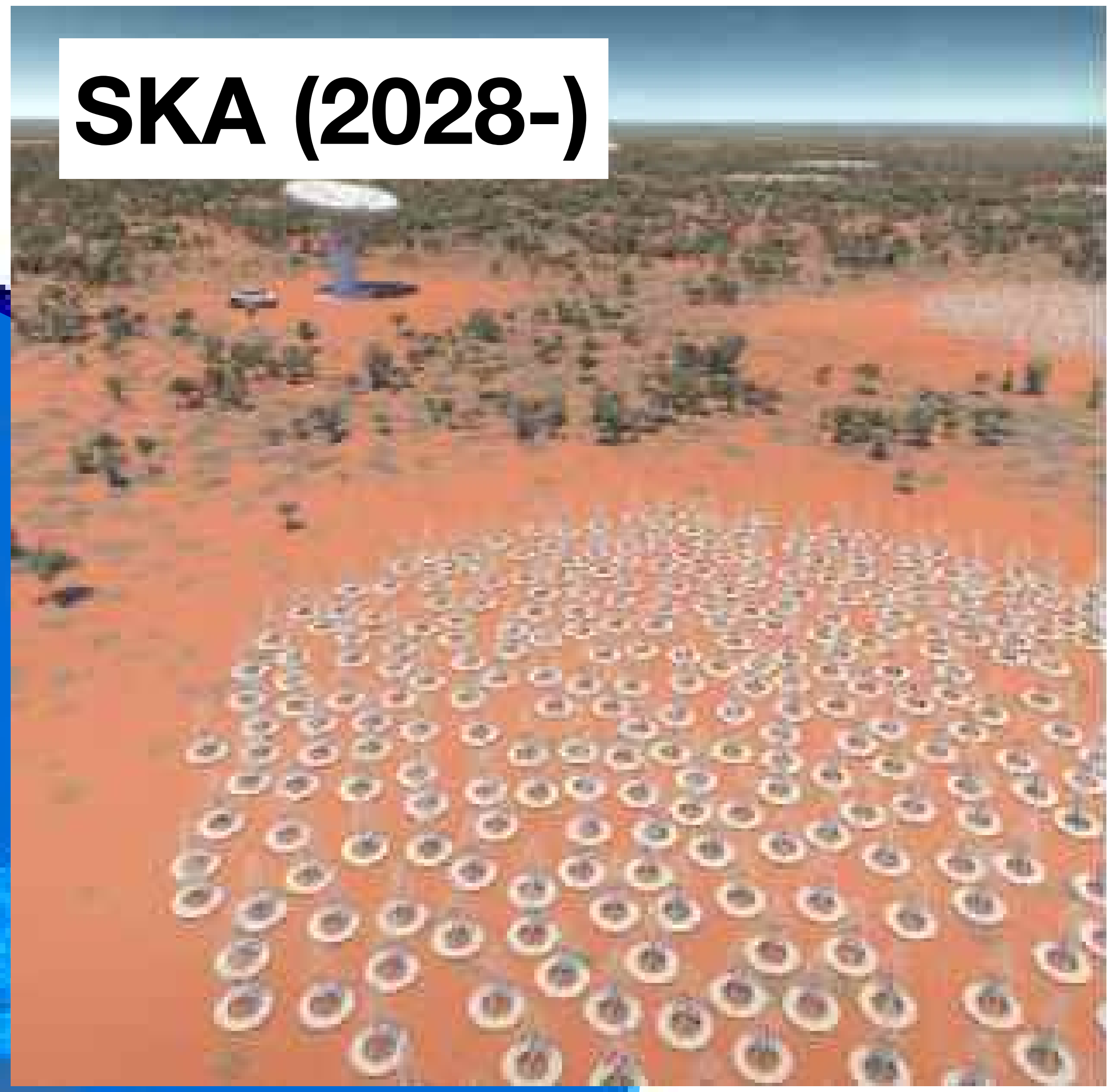


Huterer et al. 2013

Largest-volume data will be available soon!

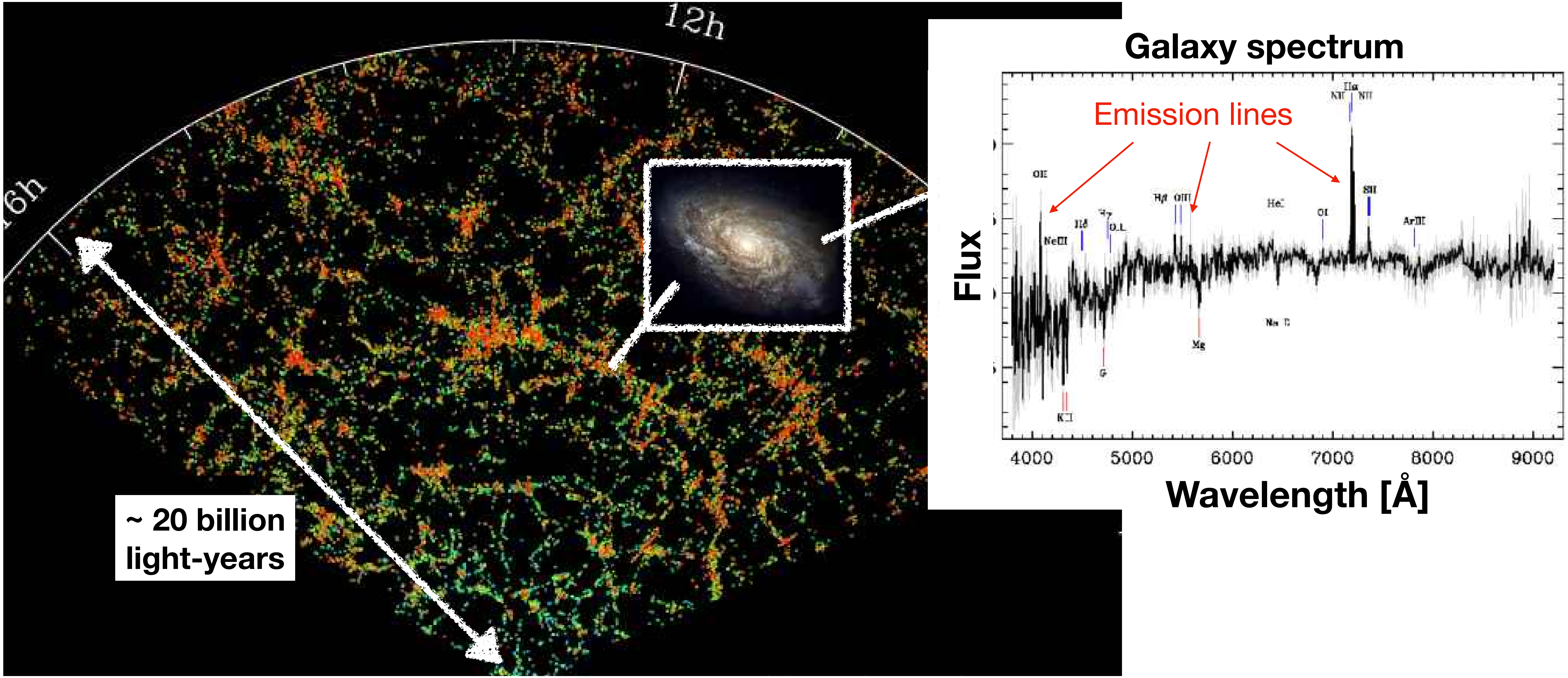
SDSS survey area





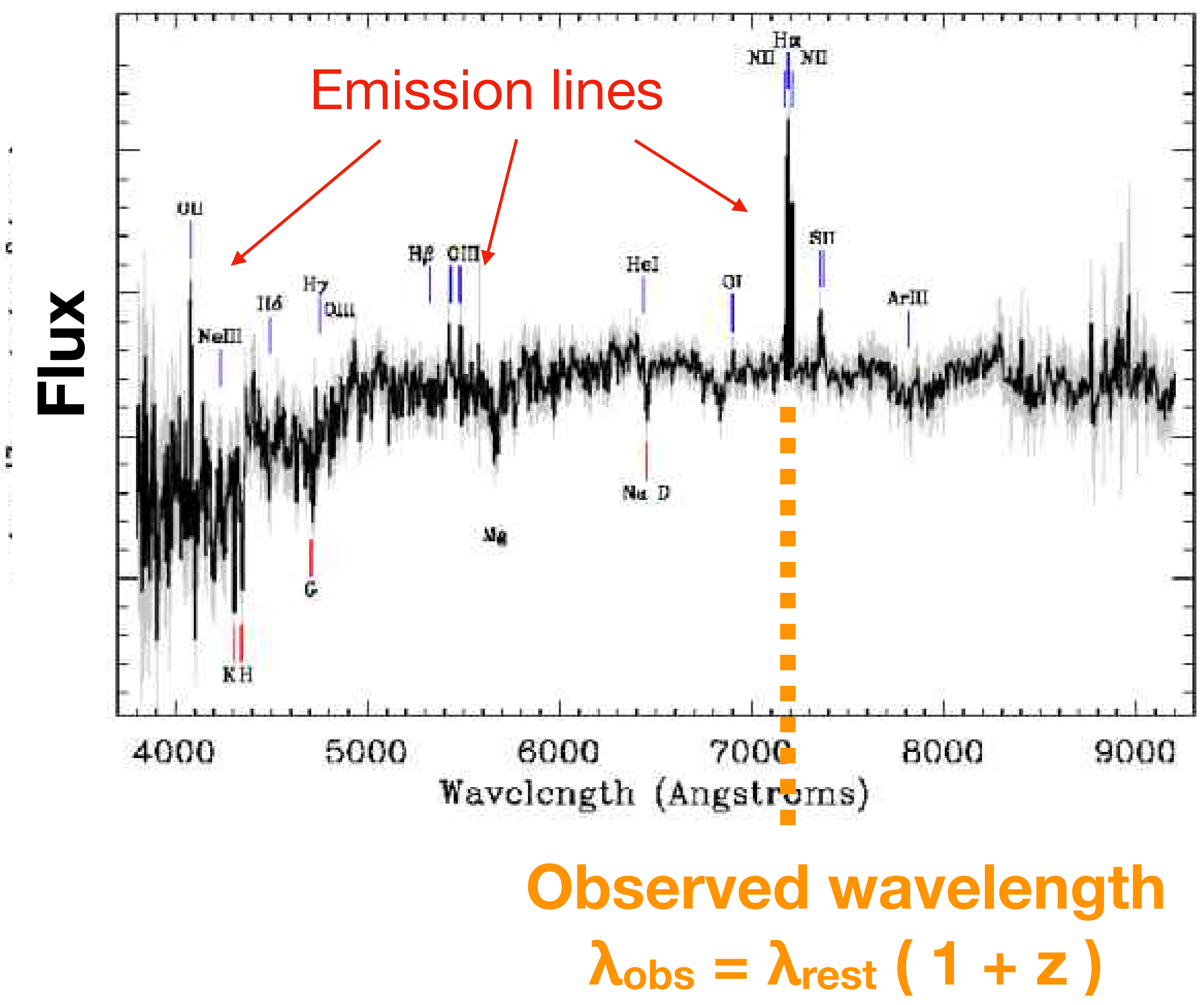
All-sky surveys will be conducted.

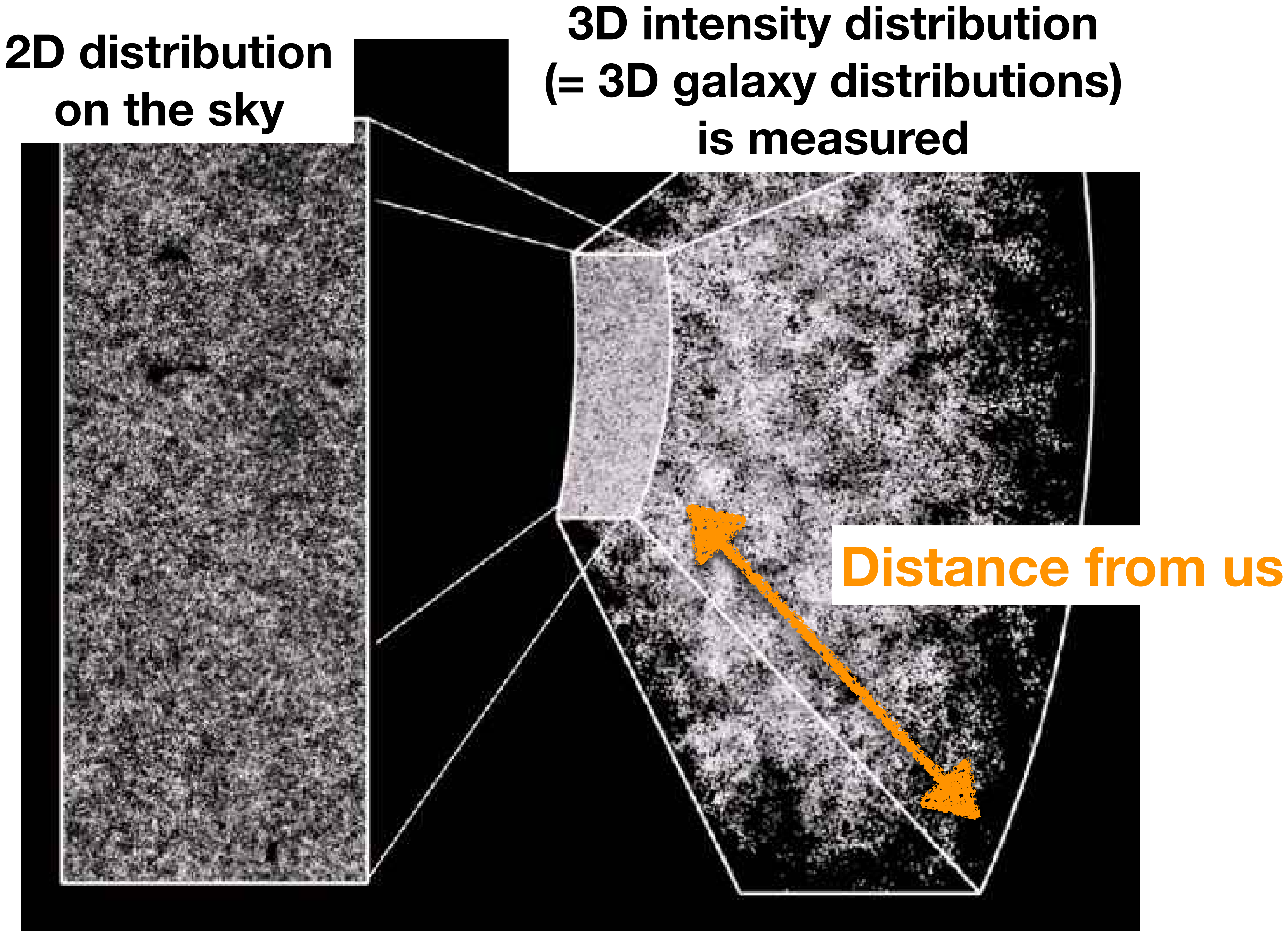
Spectroscopic Observations to Measure the Large-scale Distributions



Emission Line: a Key to Measure 3D Distributions

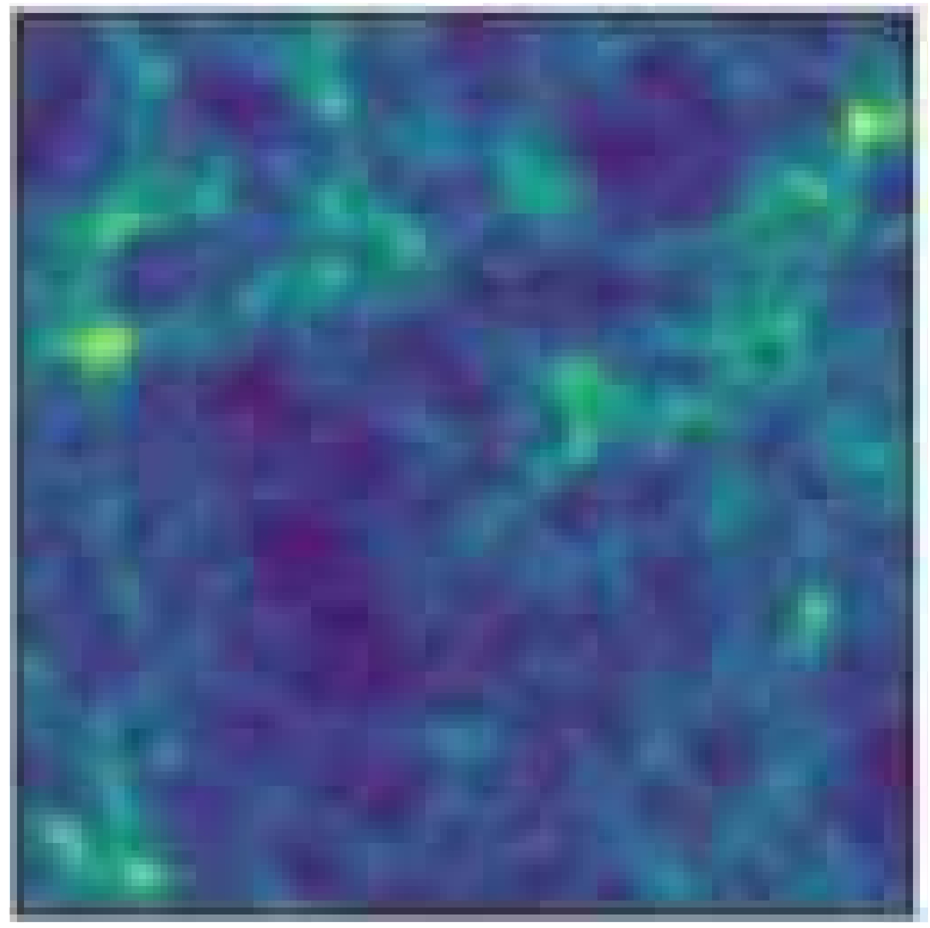
The observed wavelengths of emission lines are a measure of the distance.





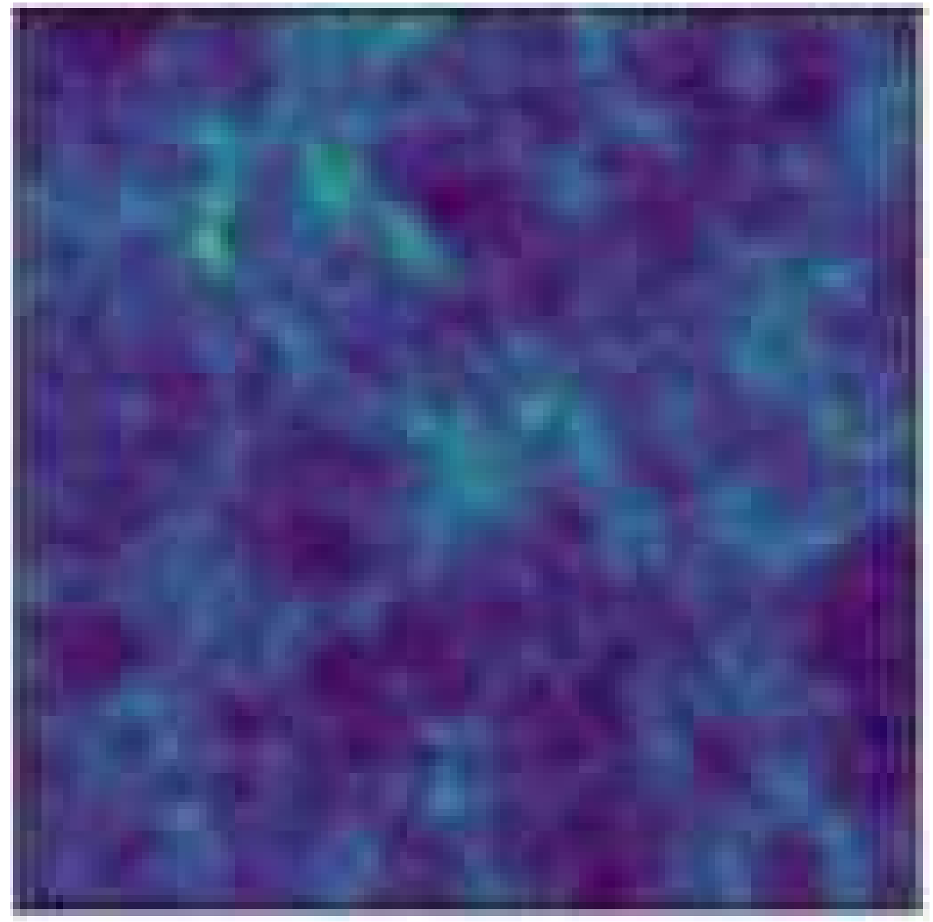
A Serious Problem: Contaminations and Noises

Hydrogen line signals from near galaxies



\+

Oxygen line signals from distant galaxies

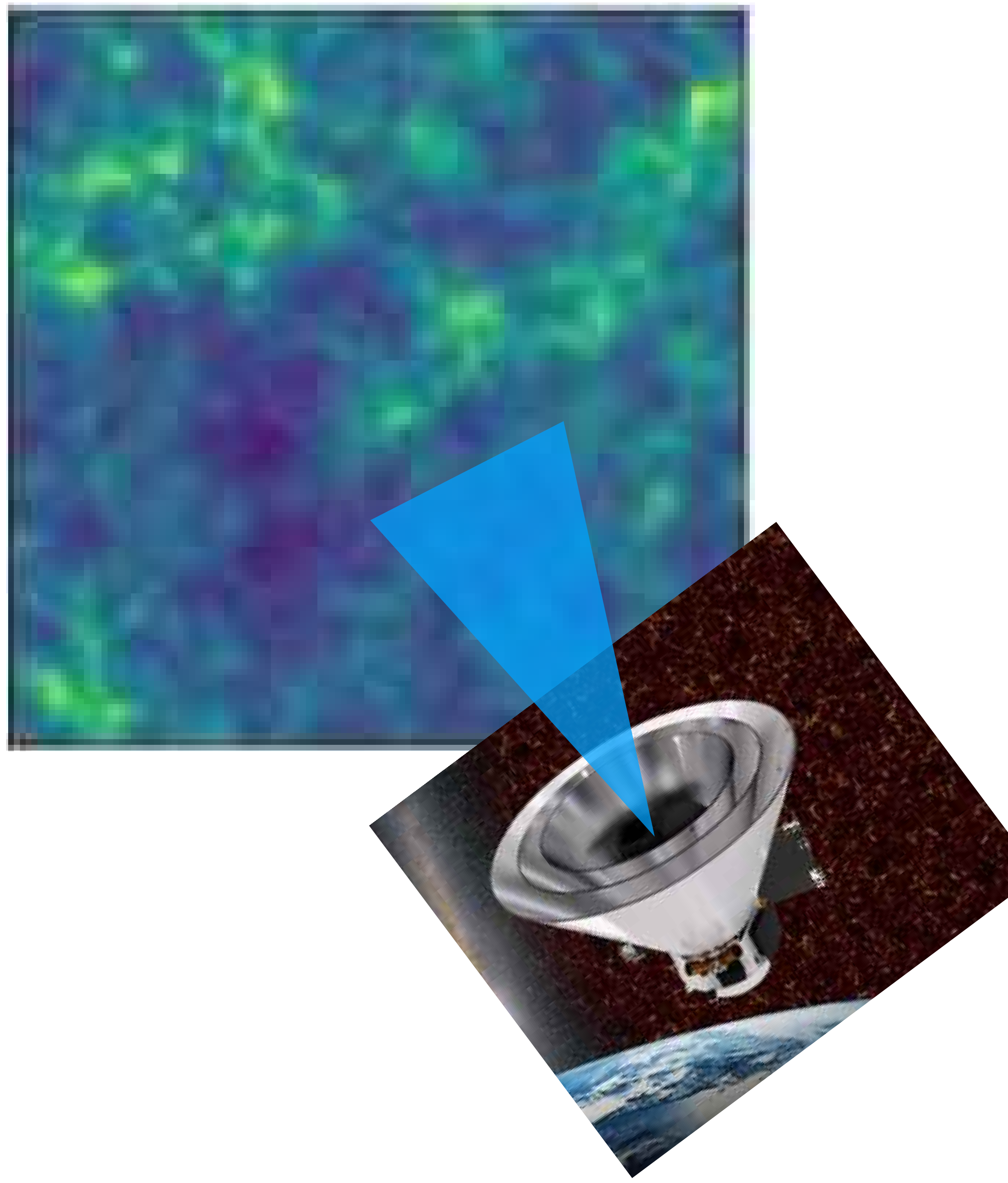


\+ **obs. noise**

Mixed up →

← **Our proposal: deep learning**

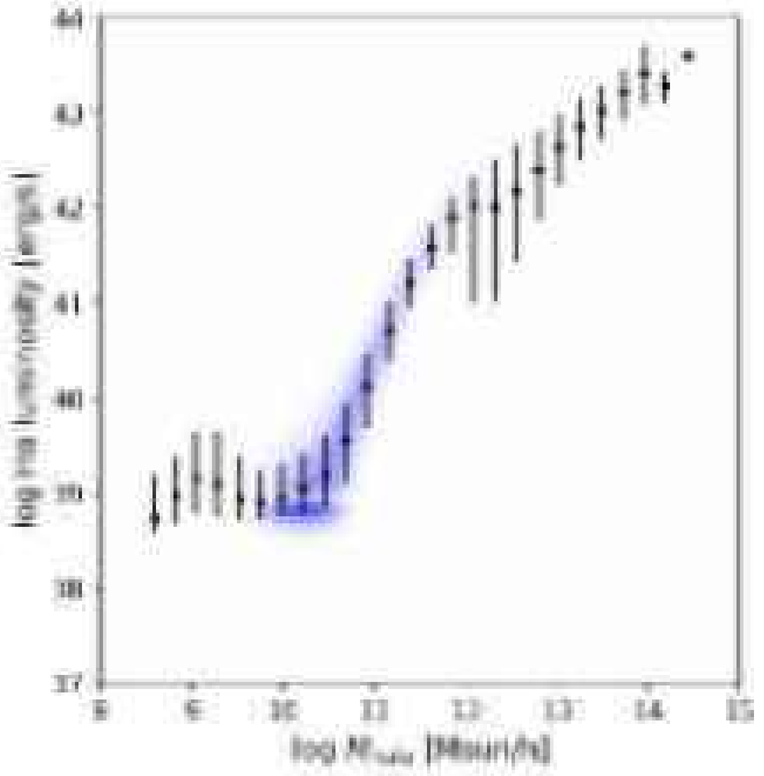
Observed data at wavelength λ_{obs}



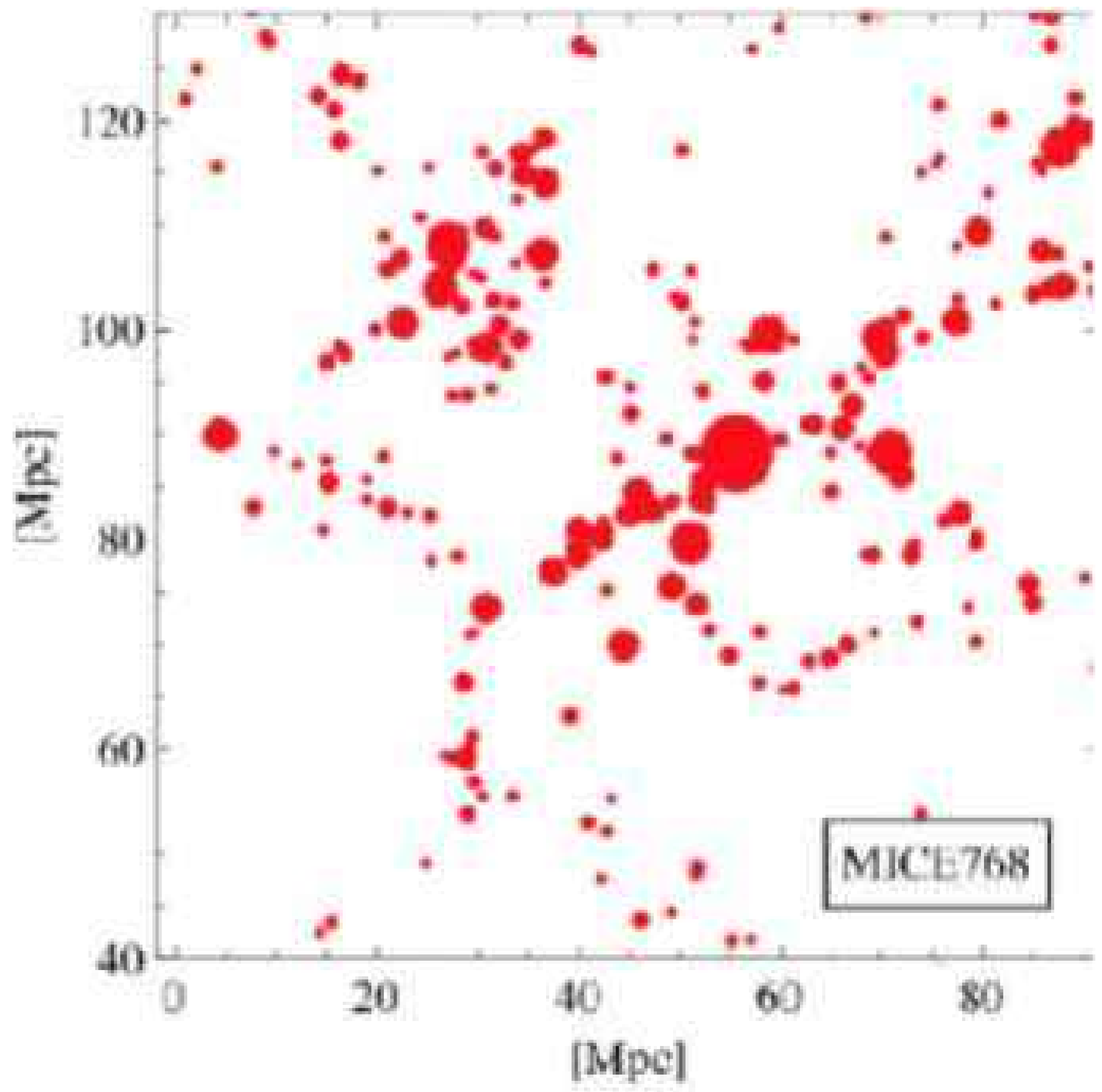
Train a Deep Learning Model with Mock Observational Data

Generate ~30,000 realistic mock observational maps using fast DM simulatino code + emission line model

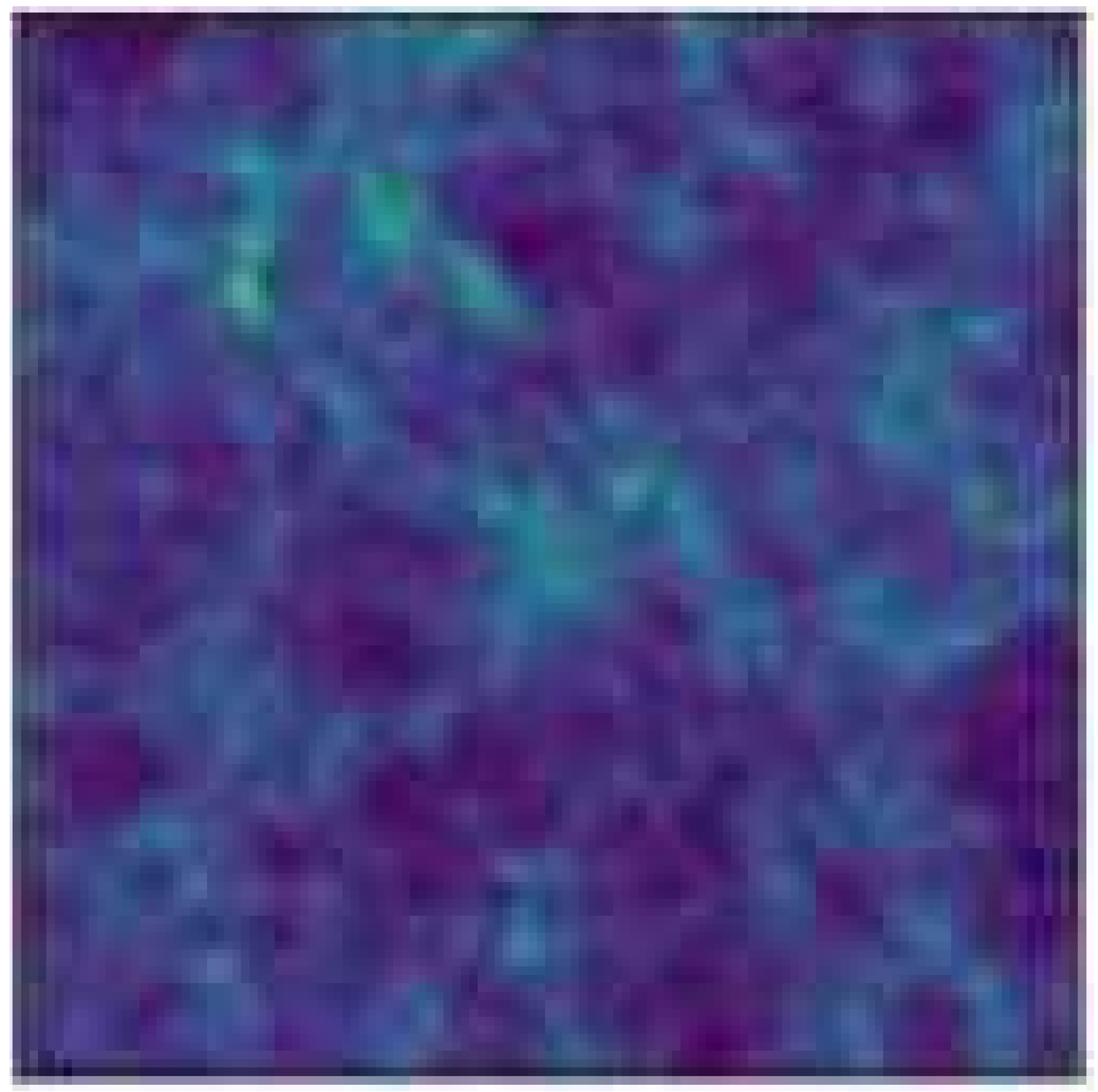
Emission line model (mass-to-luminosity relation)



DM halo distribution

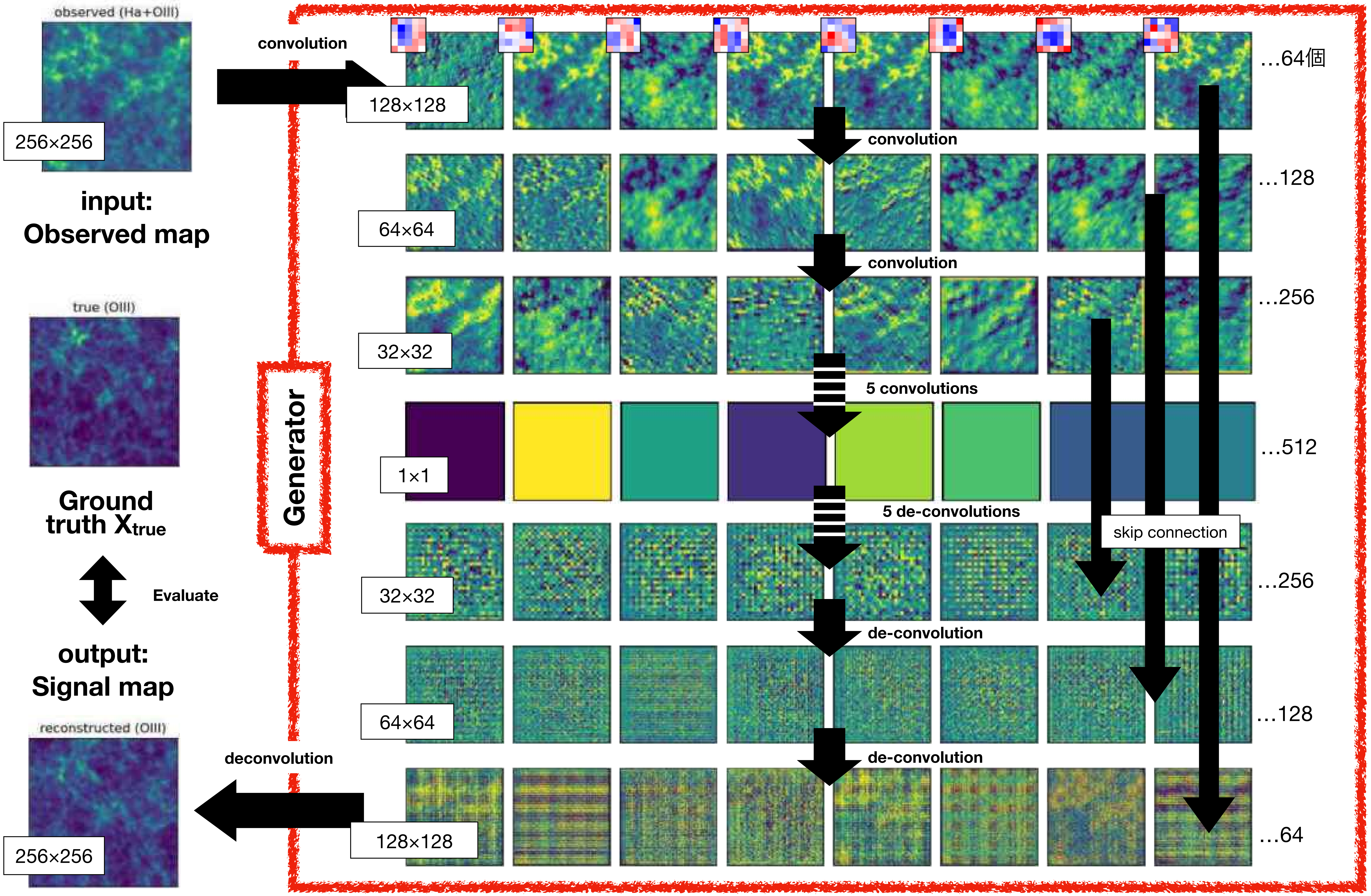


Mock data

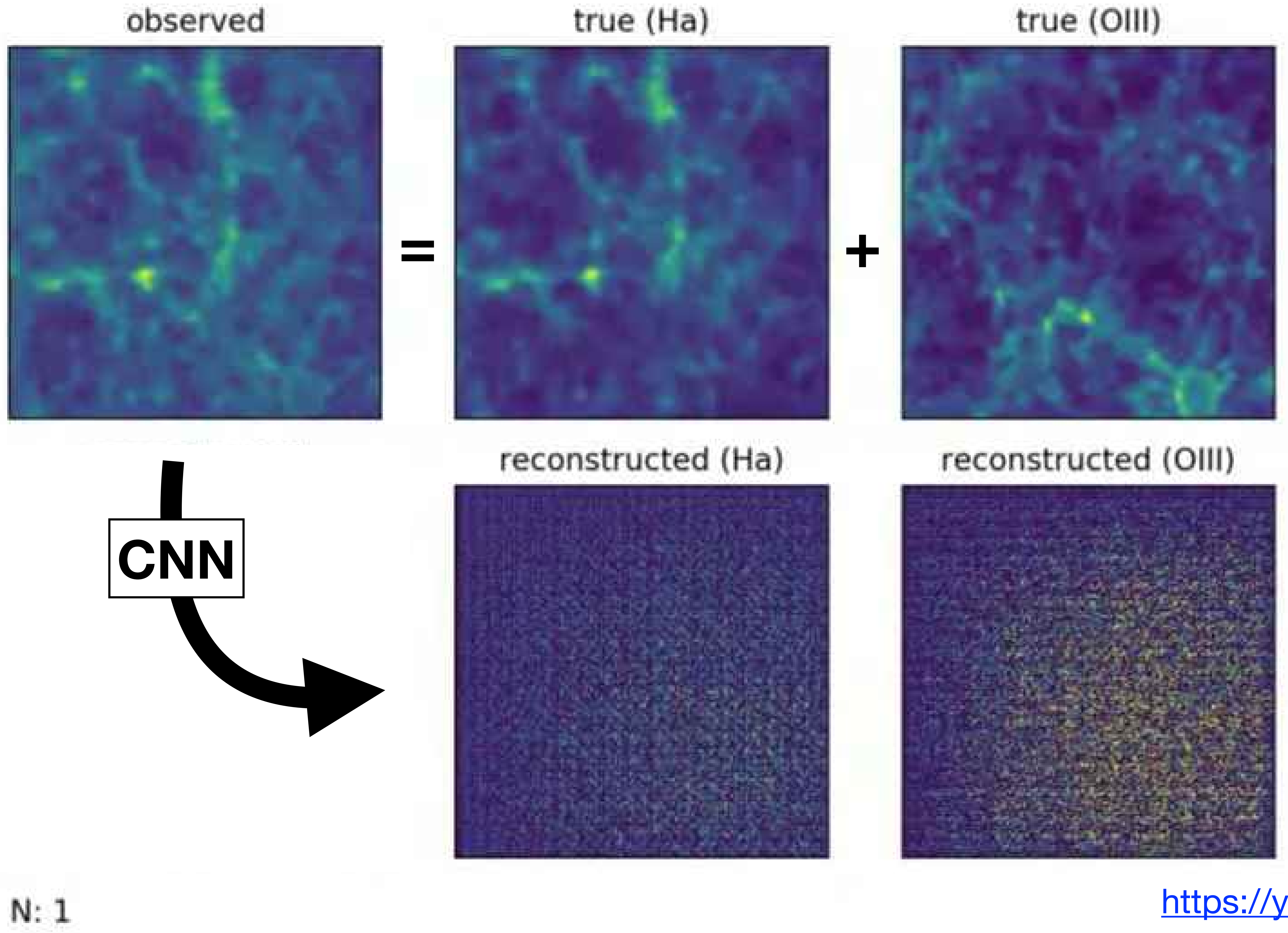


× 2 × 30,000
+ noise maps

e.g., pix2pix (Isola et al. 2016)



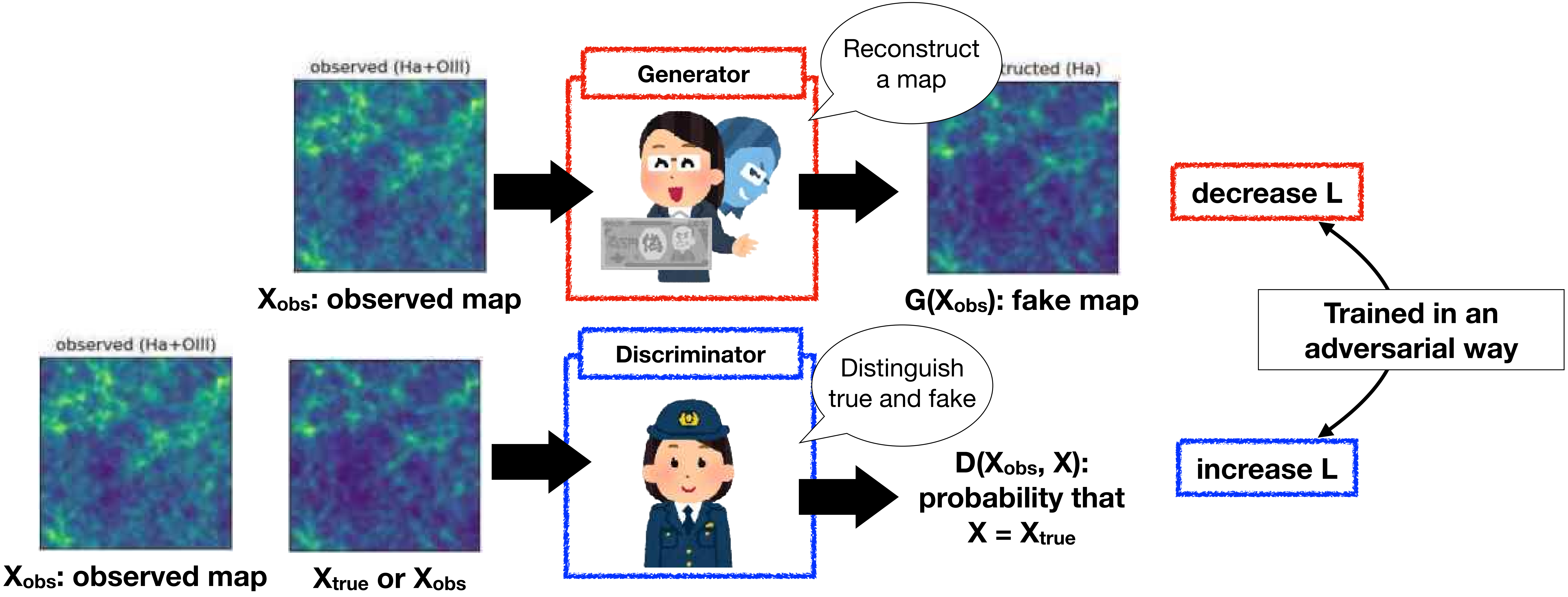
Machine Learns the Large-scale Structure…



https://youtu.be/J3c5Xk-5kT0

Conditional Generative Adversarial Network

- GAN: Generator and Discriminator are updated in an adversarial way.

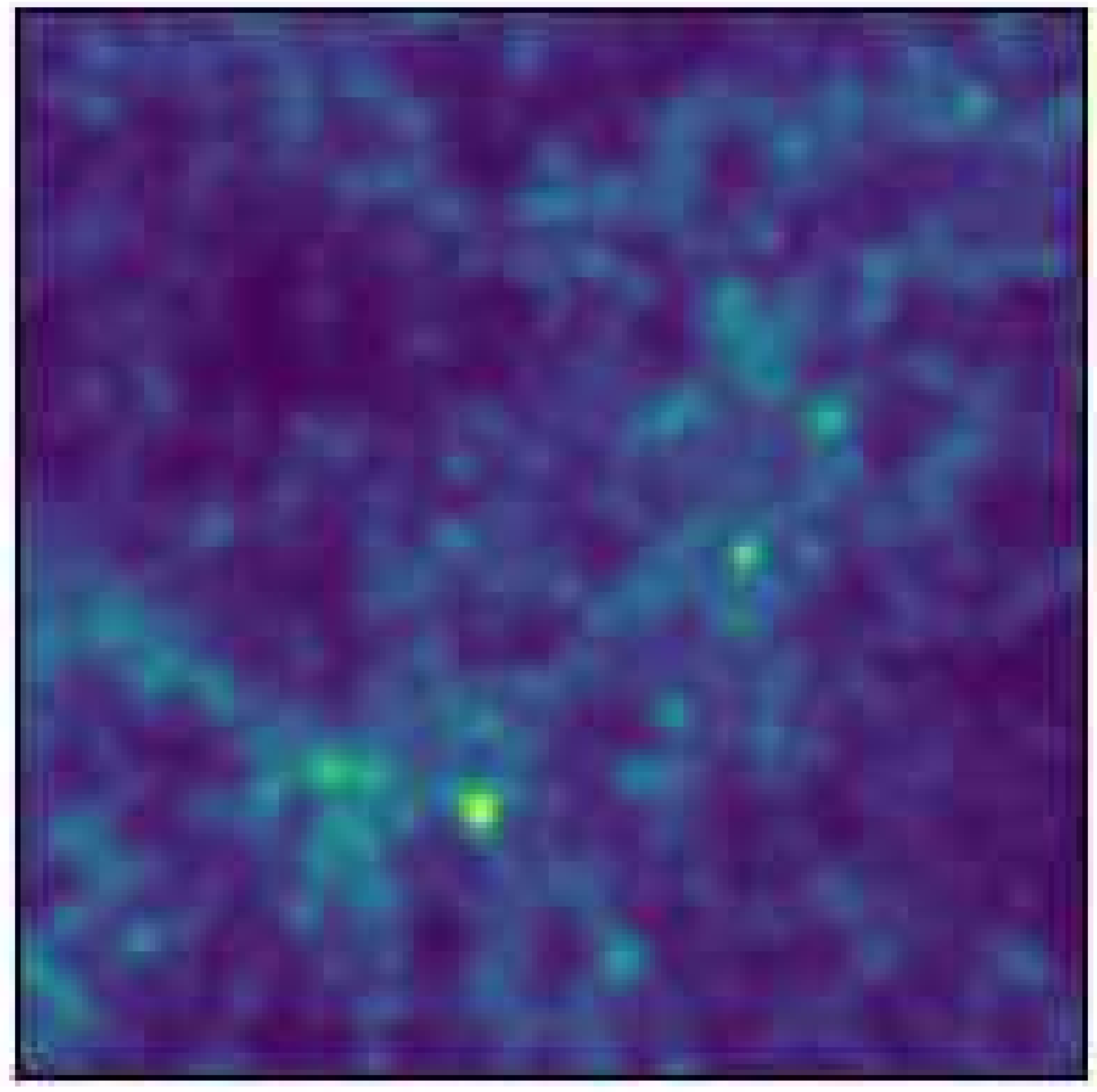


Loss function: $$L[G, D] = \log D(X_{\rm obs}, X_{\rm true}) + \log[1 - D(X_{\rm obs}, G(X_{\rm obs}))] + \lambda \langle |X_{\rm true} - G(X_{\rm obs})| \rangle$$

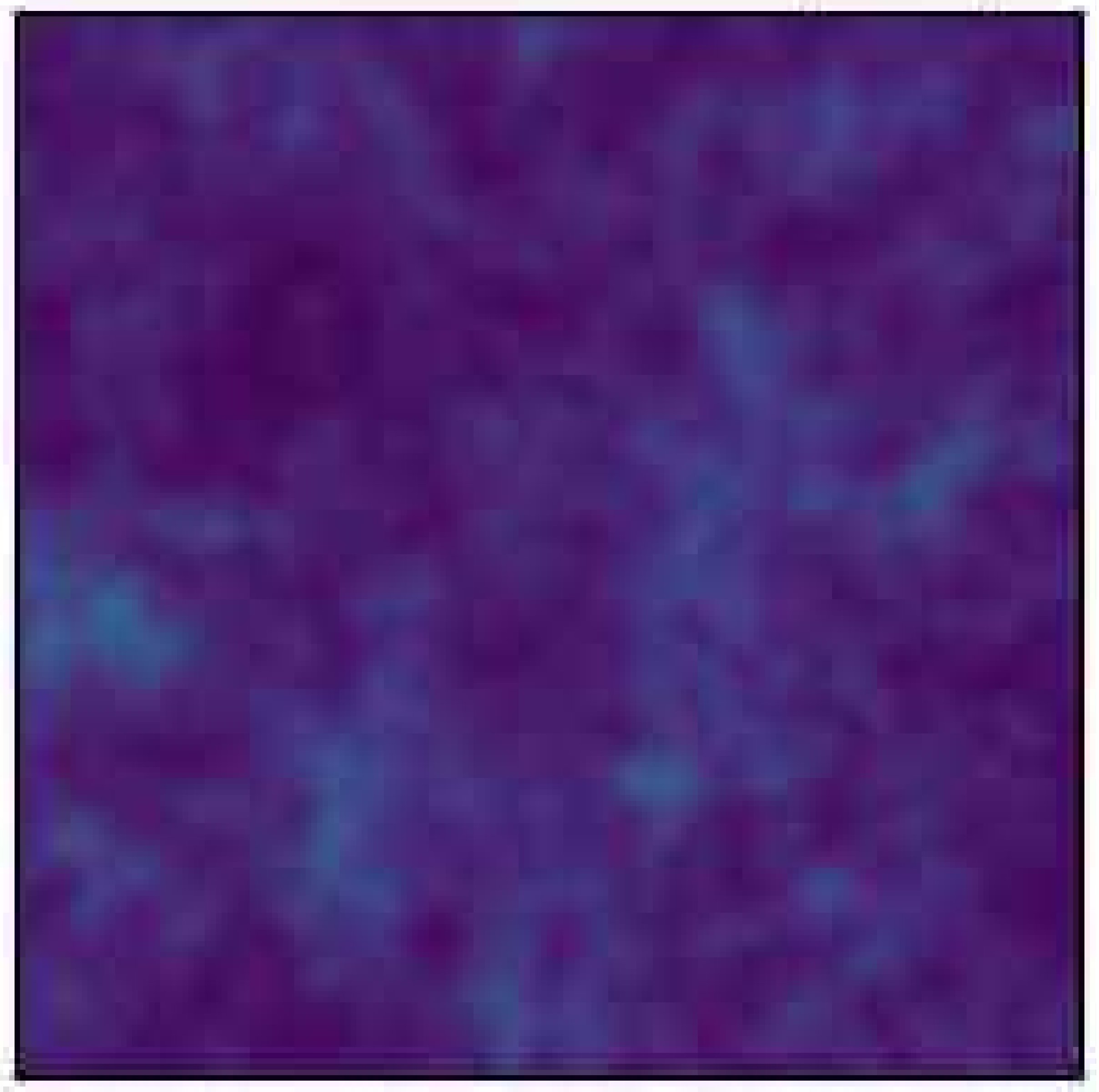
What if we do not use GAN?

The network tends to reproduce obscured images

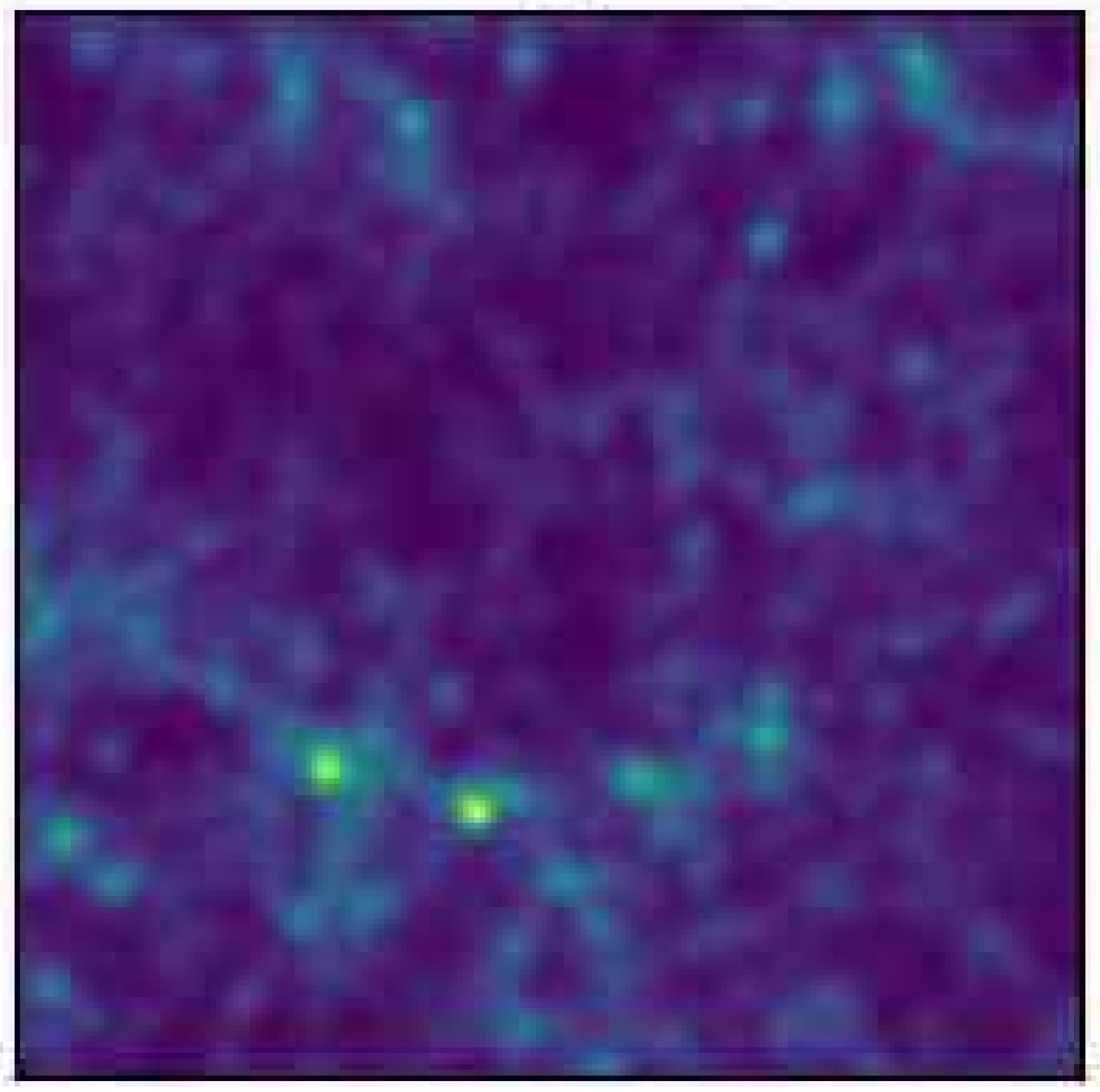
Observed (Line1+Line2)



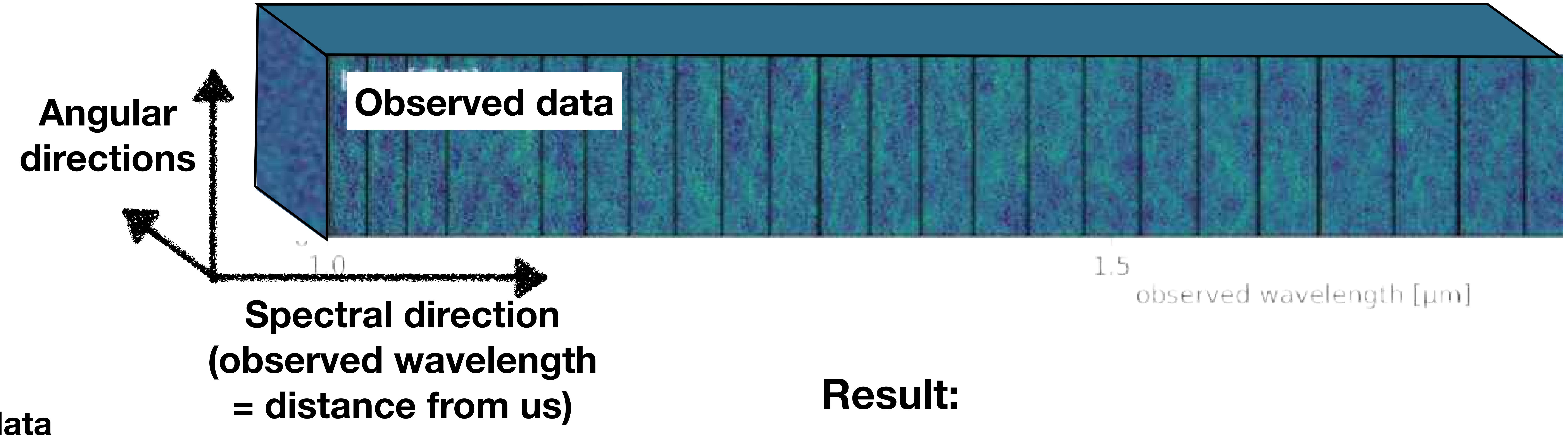
Reconstruct (Line2)

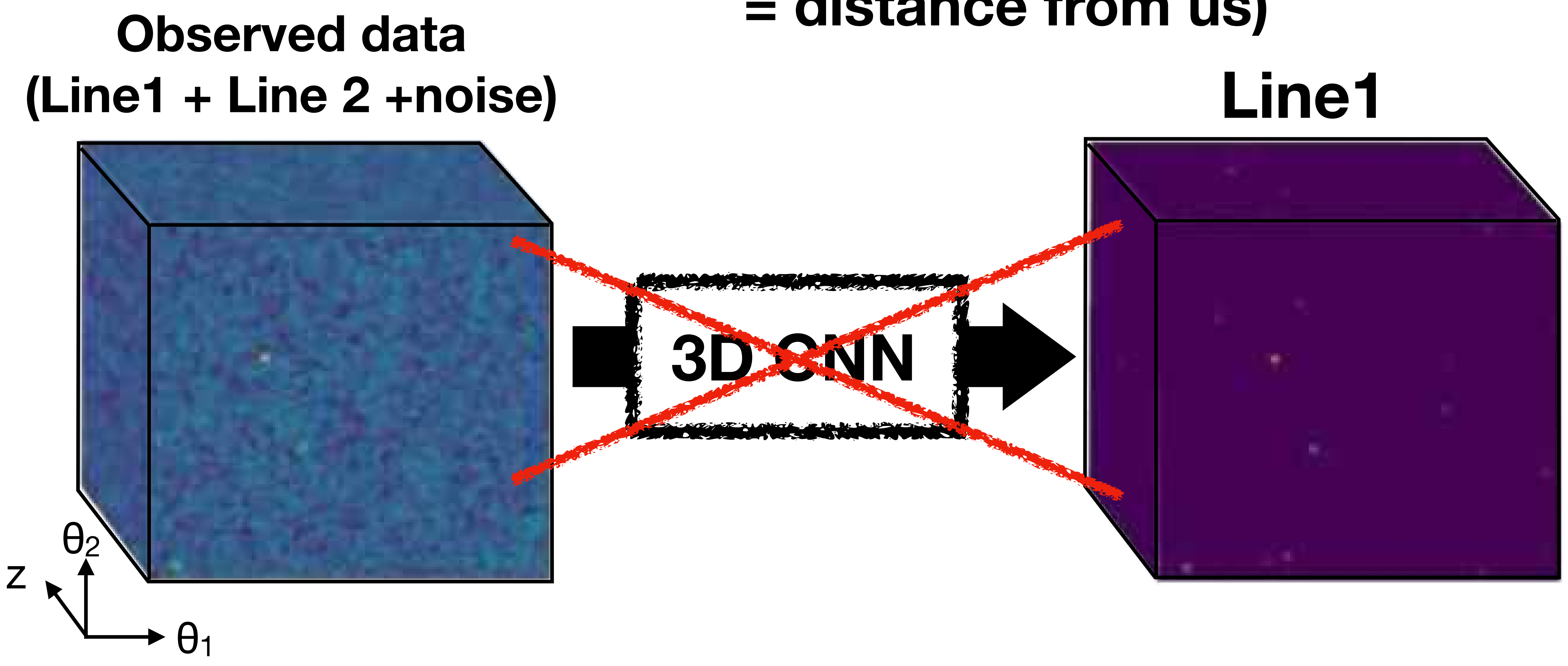


True (Line2)

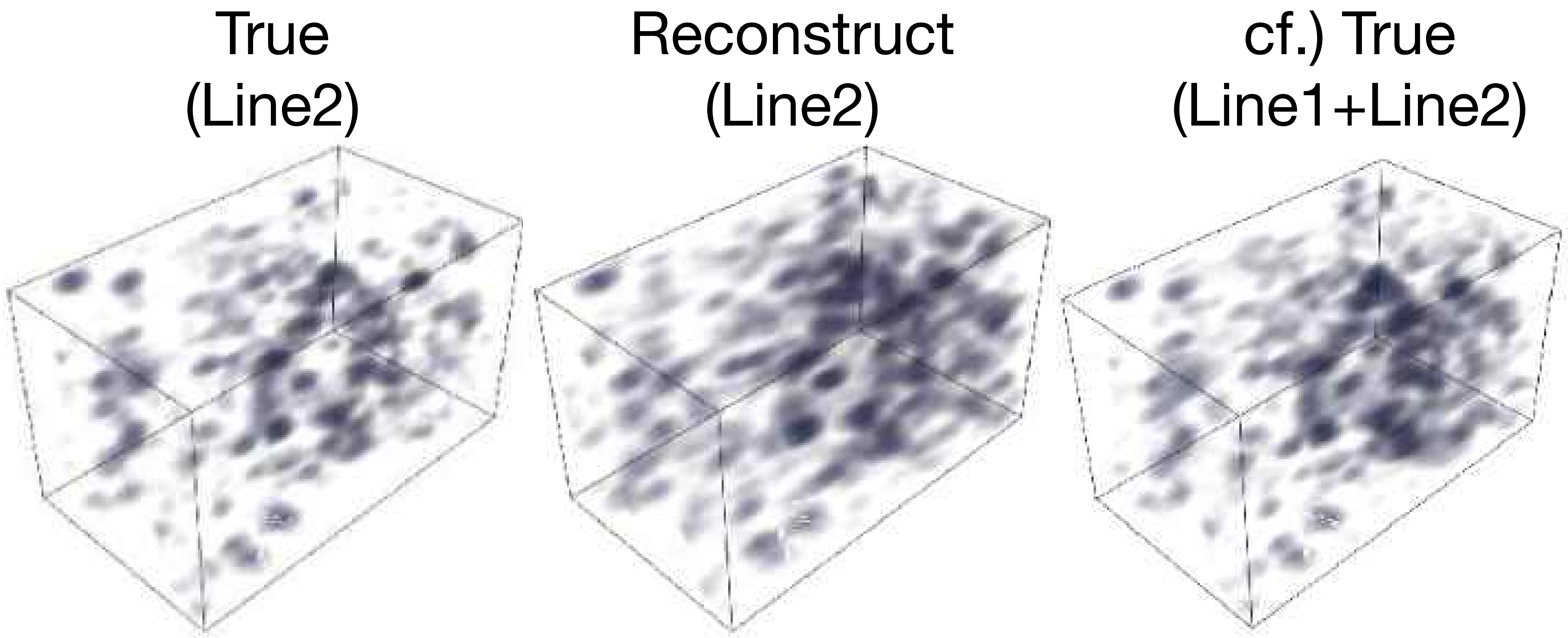


Reconstruction of 3D Maps



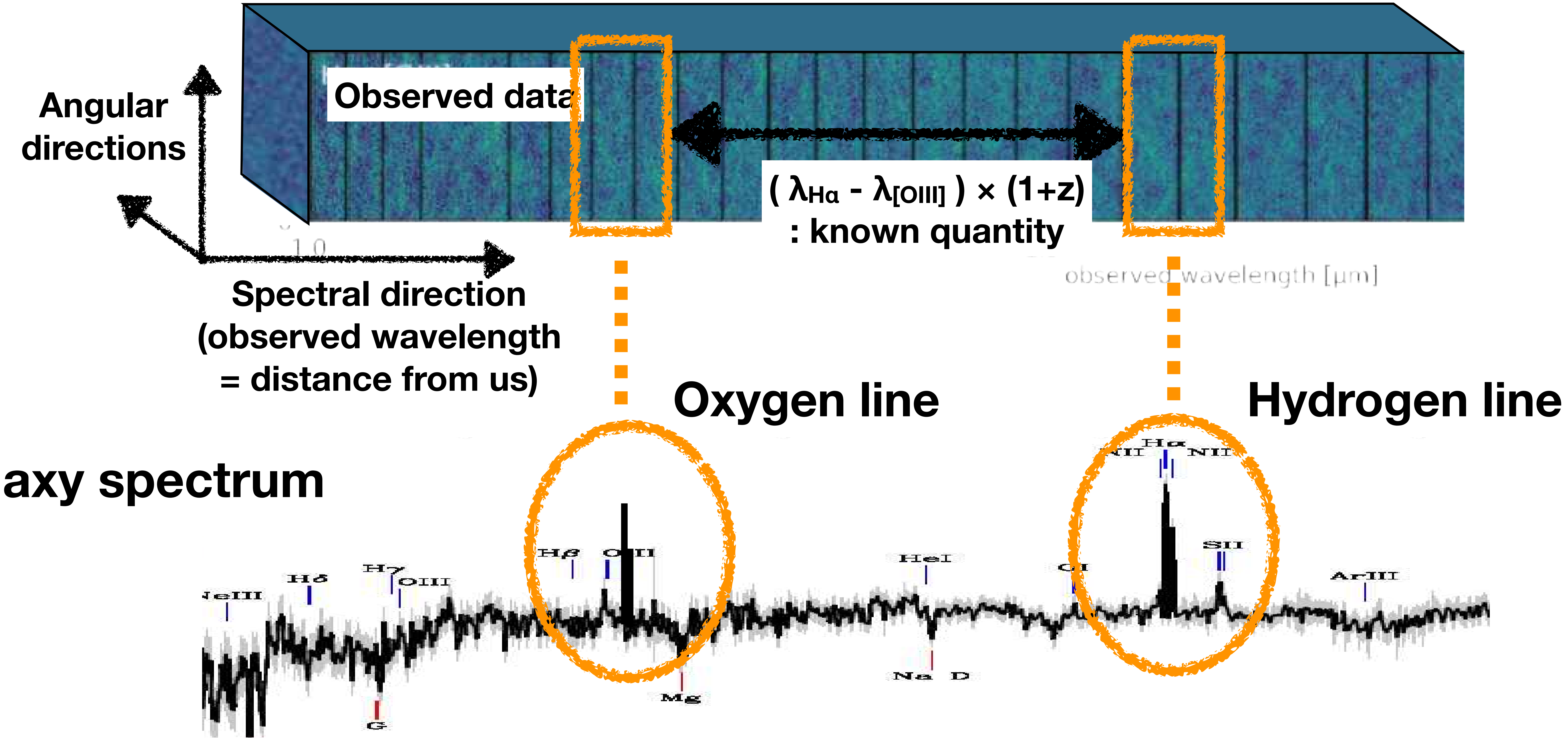


Result:



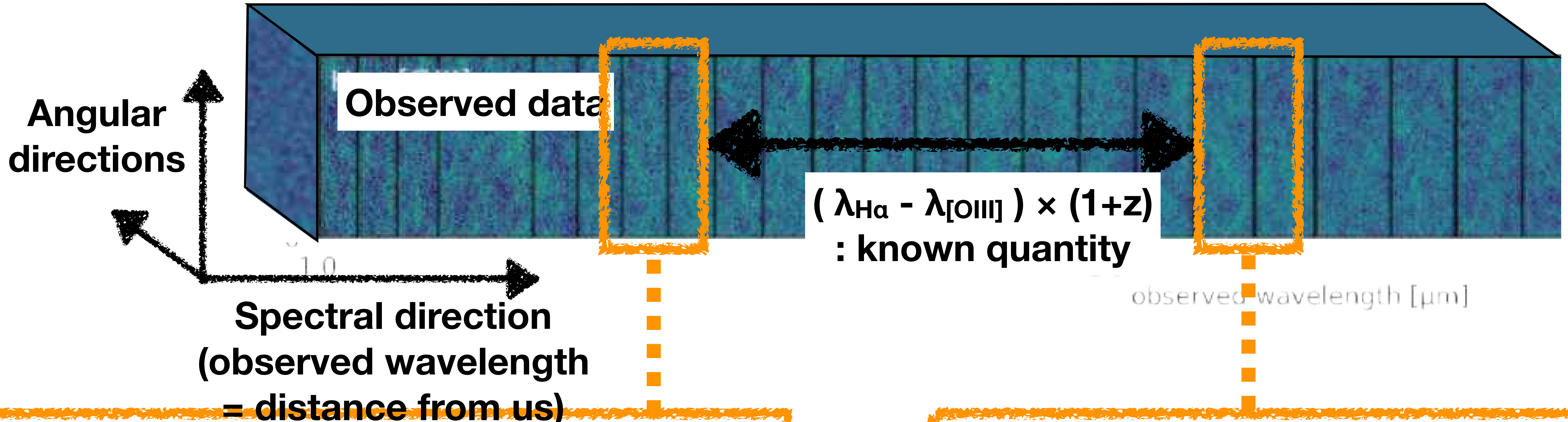
A simple network does not work well...

Pre-processing Input Data with *Physical informaiton*

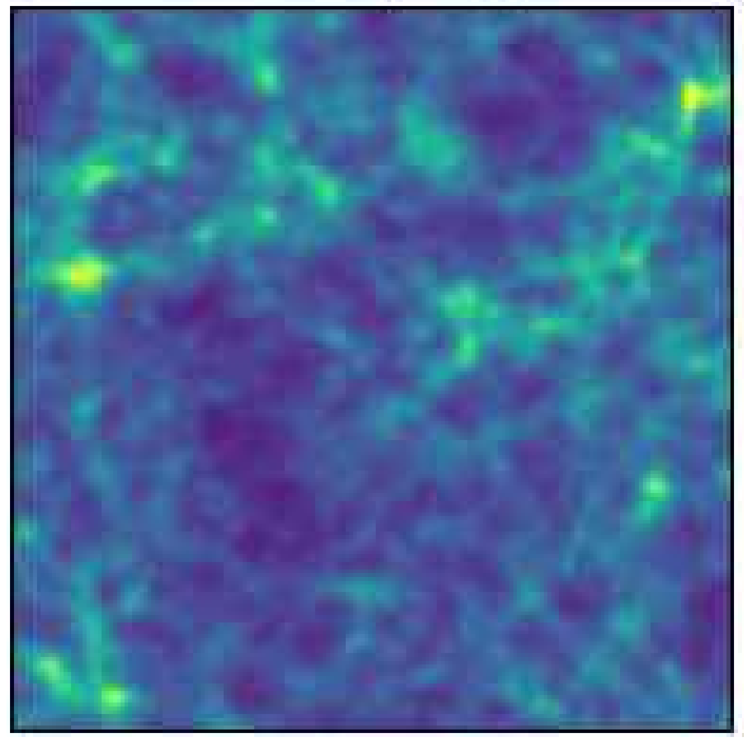


The contribution from the galaxy appears in two wavelengths

Pre-processing Input Data with *Physical informaiton*



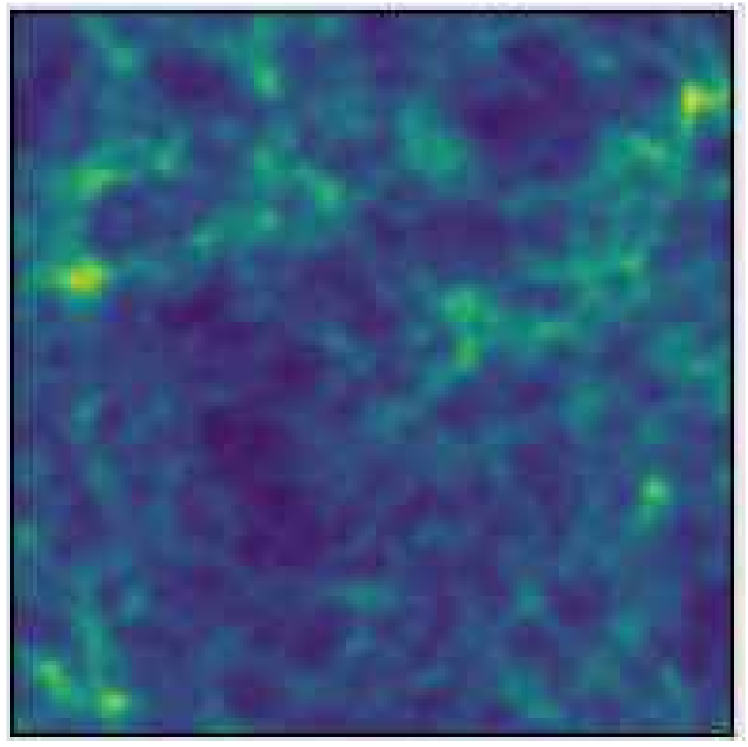
The structure traced by oxygen line



+ contaminations & obs. noise

(Almost) the same signals are buried in the two data cubes

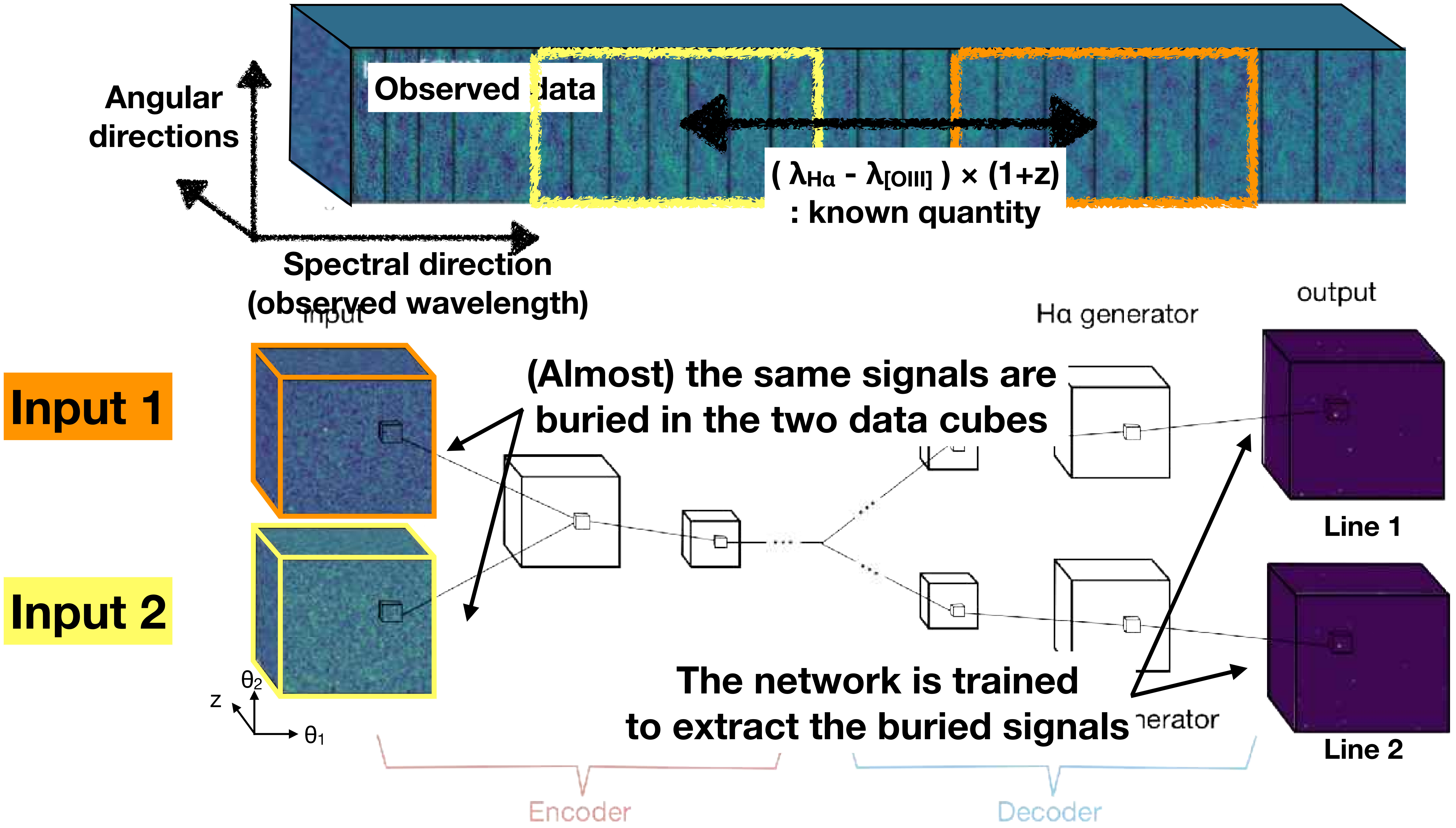
The structure traced by hydrogen line



+ contaminations & obs. noise

Pre-processing Input Data with *Physical informaiton*

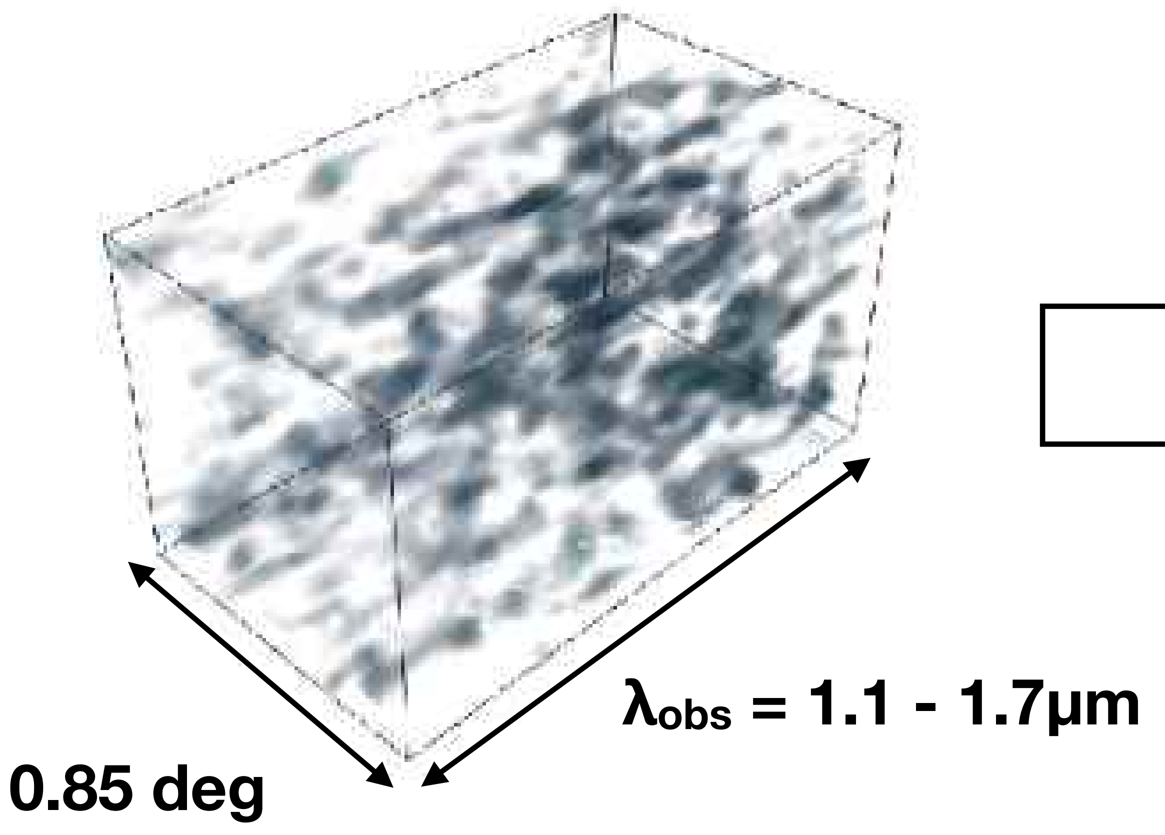
KM & Yoshida 2021

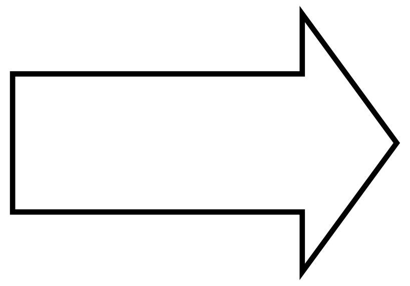


Reconstruction Result

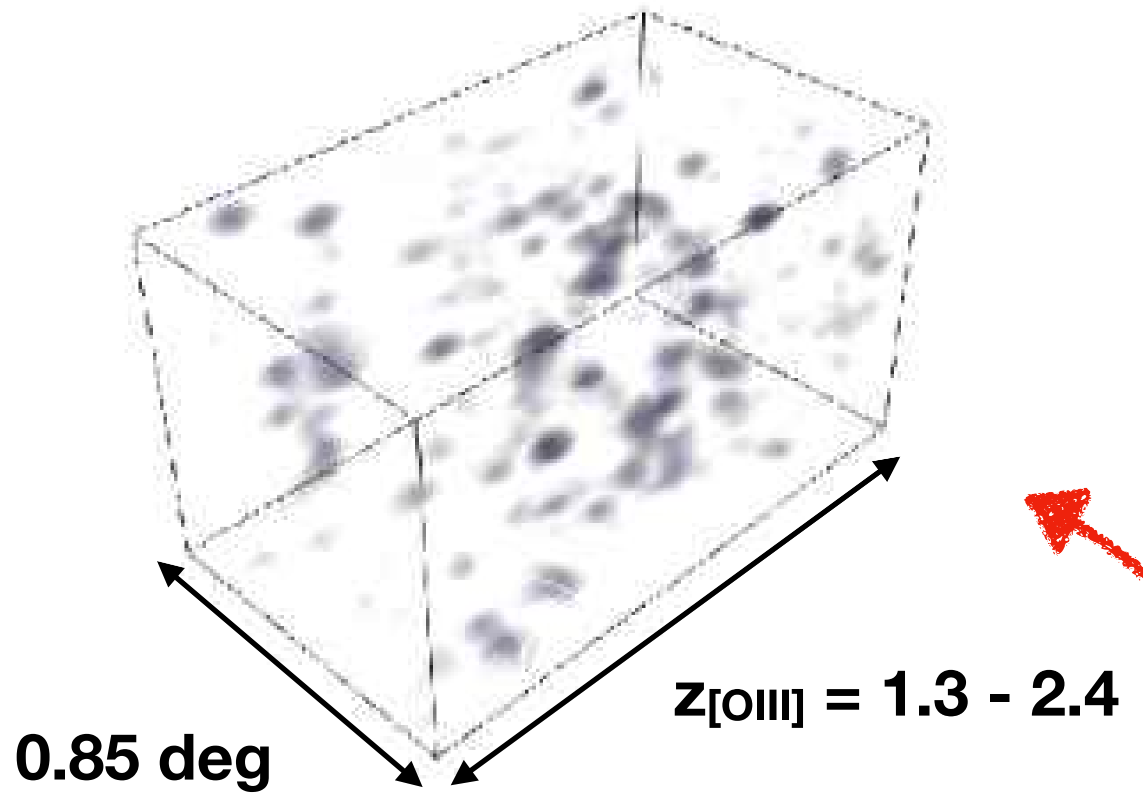
KM & Yoshida 2021

Observed (Hα+[OIII]+noise)

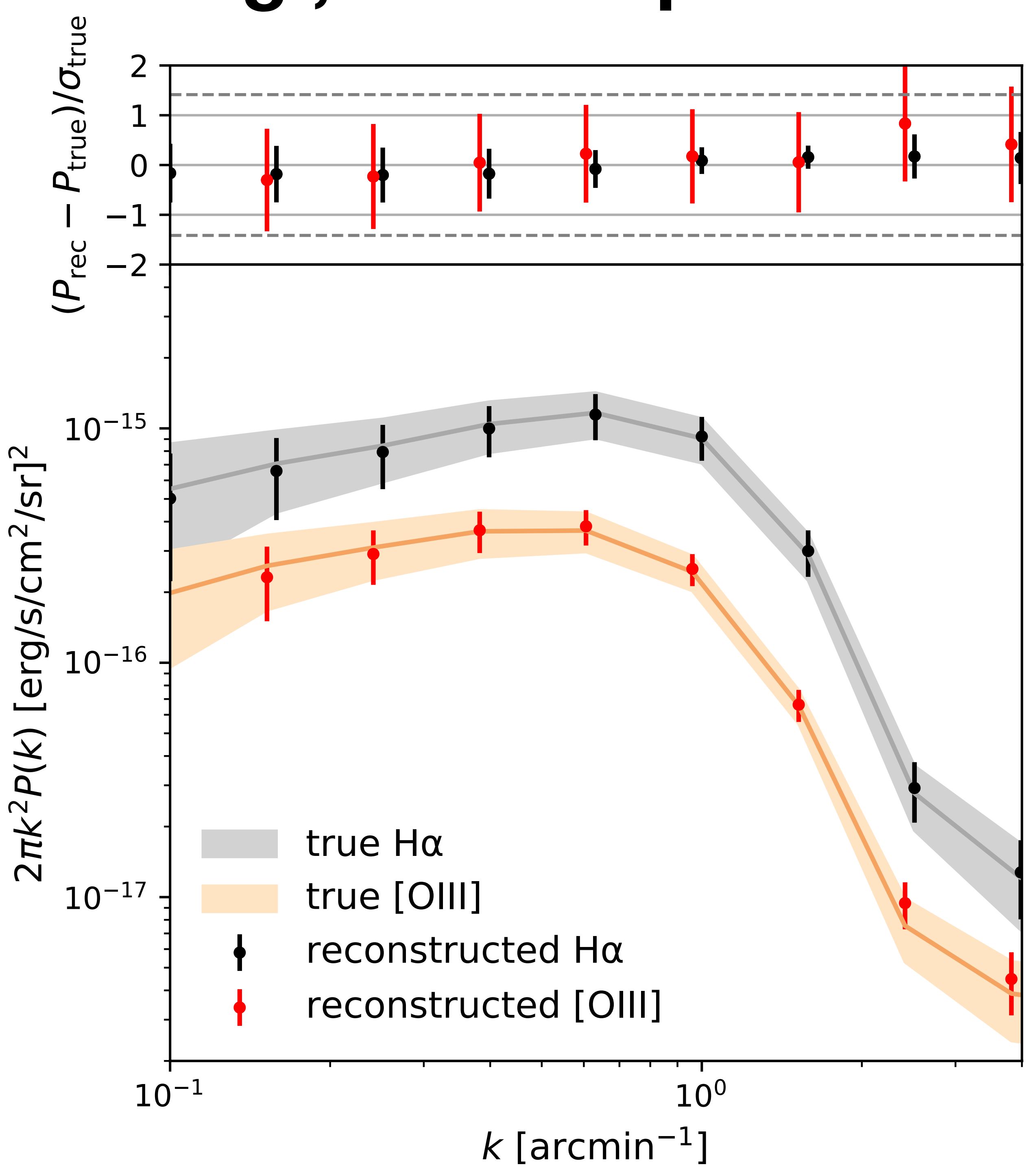




Reconstruct [OIII]



Statistics are also reproduced e.g., Power spectrum



Bright sources are properly reconstructed

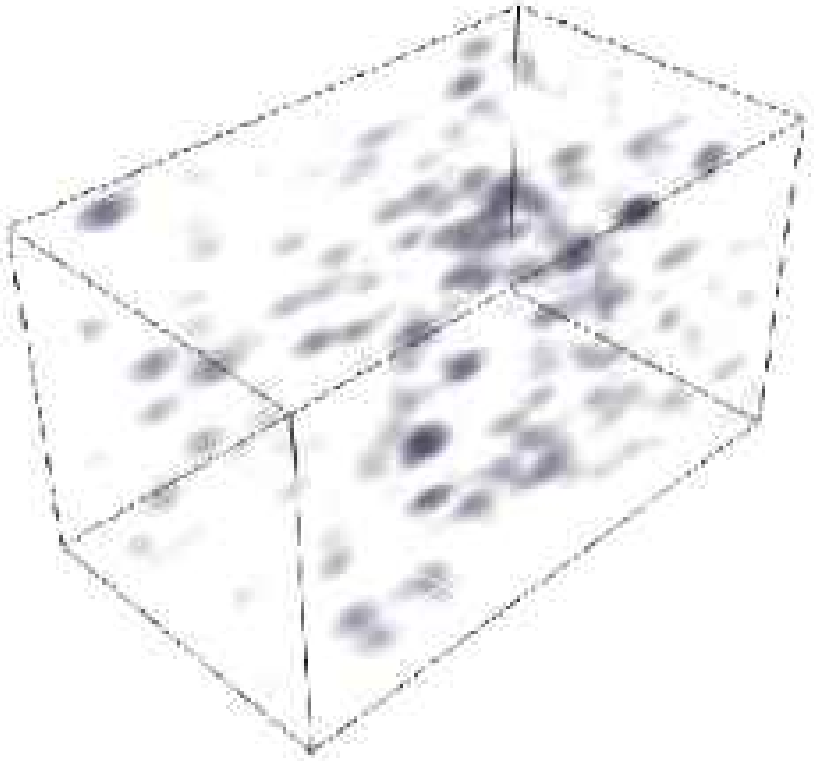
Reproducibility of bright peaks

Peak detectability of Hα and [OIII]

precision = 82%, 68%

recall = 80%, 77%

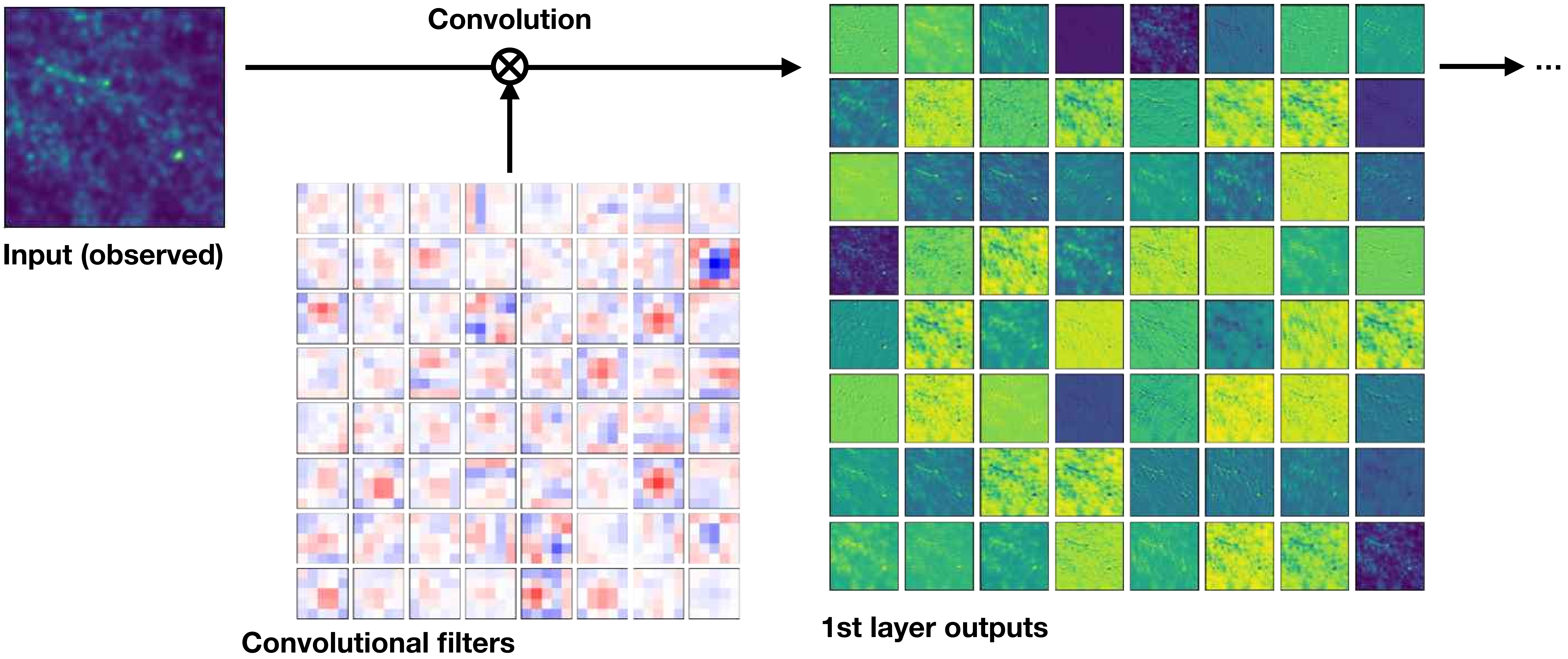
True [OIII]



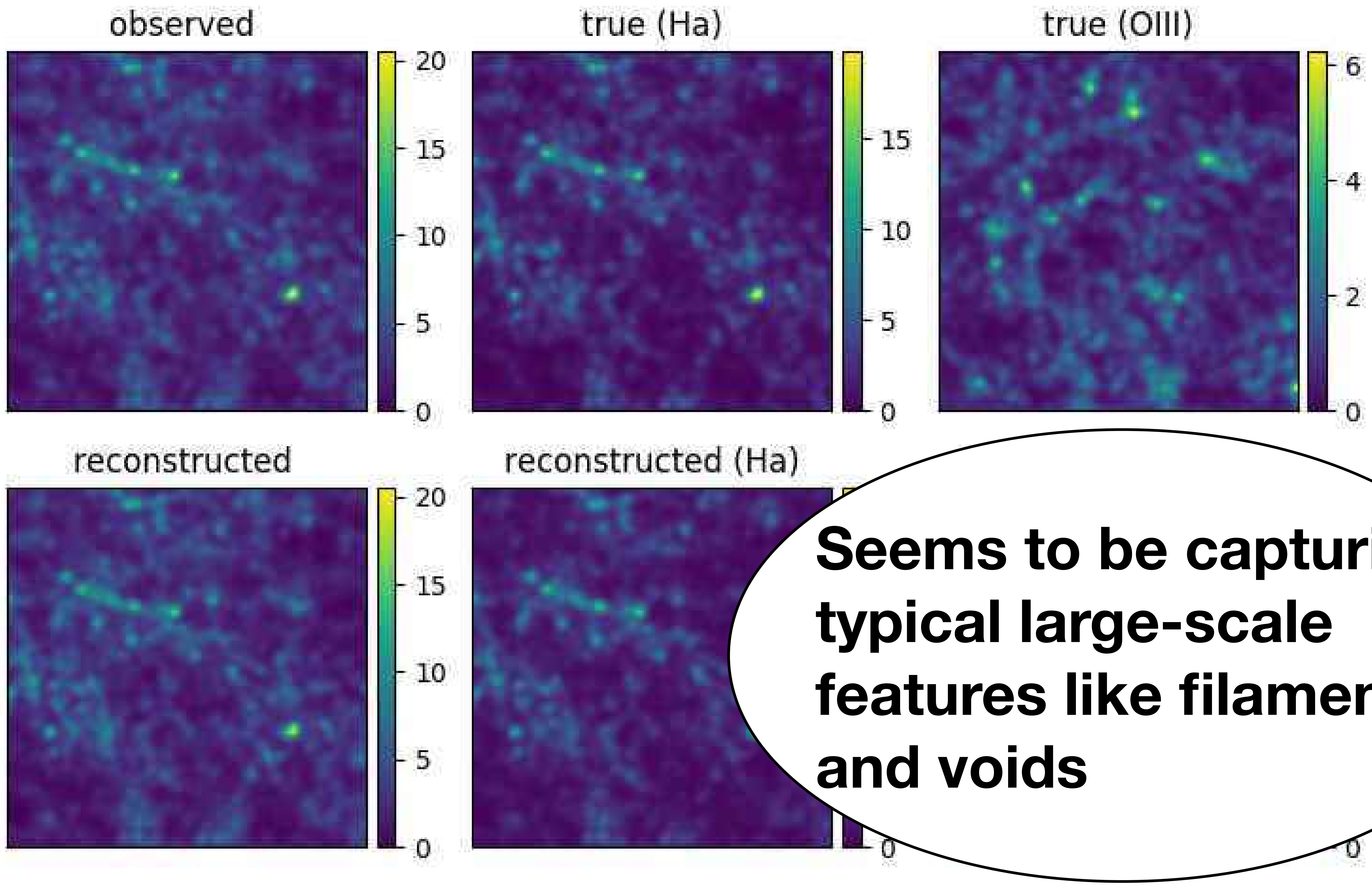
Reconstructed maps can be used for studying cosmology and galaxy formation and evolution

What Does the Machine Learns to Separate the Signals?

Let's have a look at the convolutional filters.

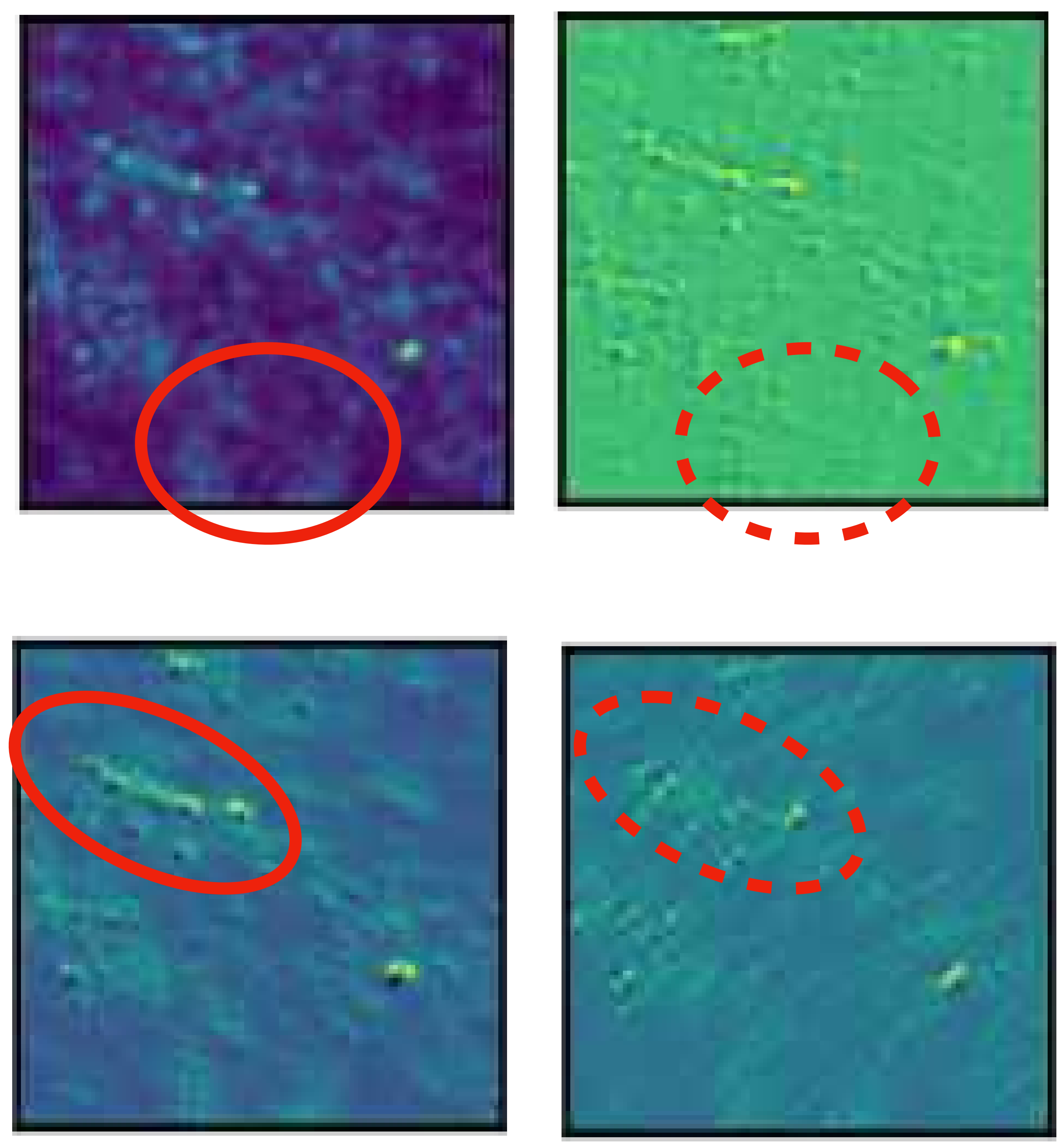


Filters in 2D Separation Models



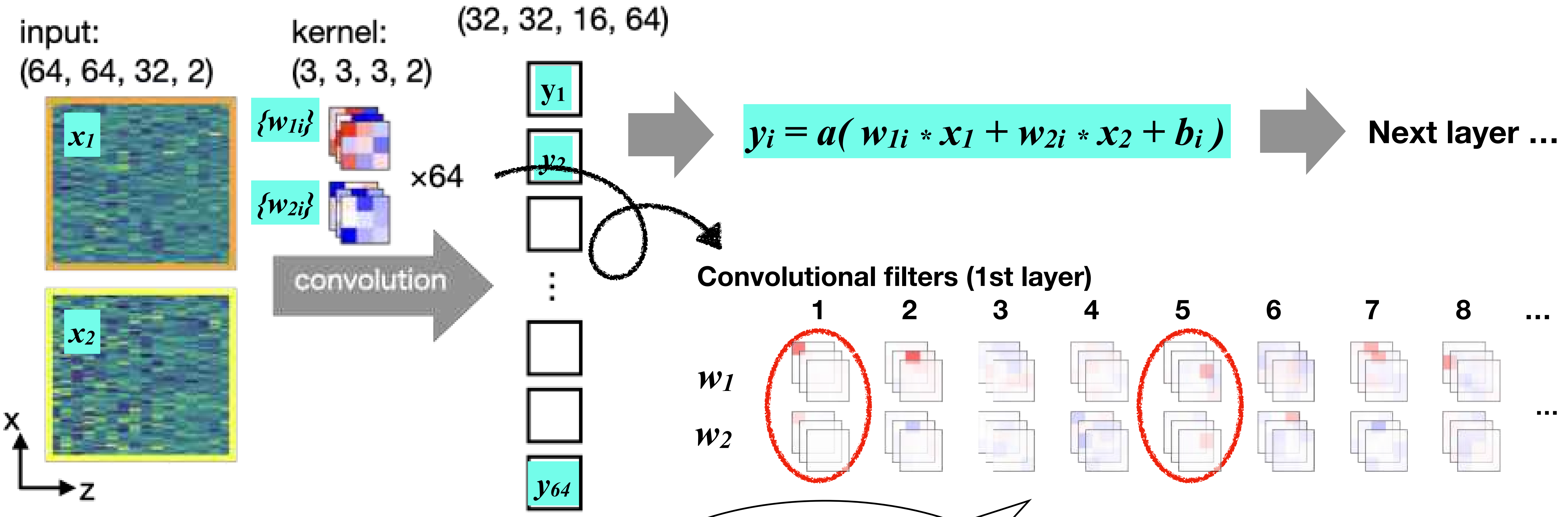
Seems to be capturing typical large-scale features like filaments and voids

Examples of the filters in the 1st layer



Structures at different distances have different features (e.g., scale length, bias).
→ The machine might distinguish signals from different distances by learning them.

Filters in 3D Network with Pre-processing



Some filters pick up synchronizing signals in two inputs

→ Our method with physical information is working effectively as expected!

Challenges in Application for Actual Observational Data

Can we trust the reconstructed maps?

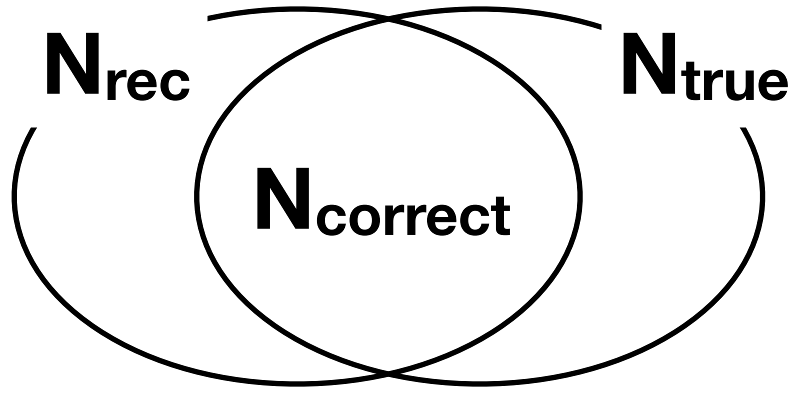
- **How precise is the reconstructed map? Is there a generation error?**
- **Is the model dependent on the assumption in the training model?**

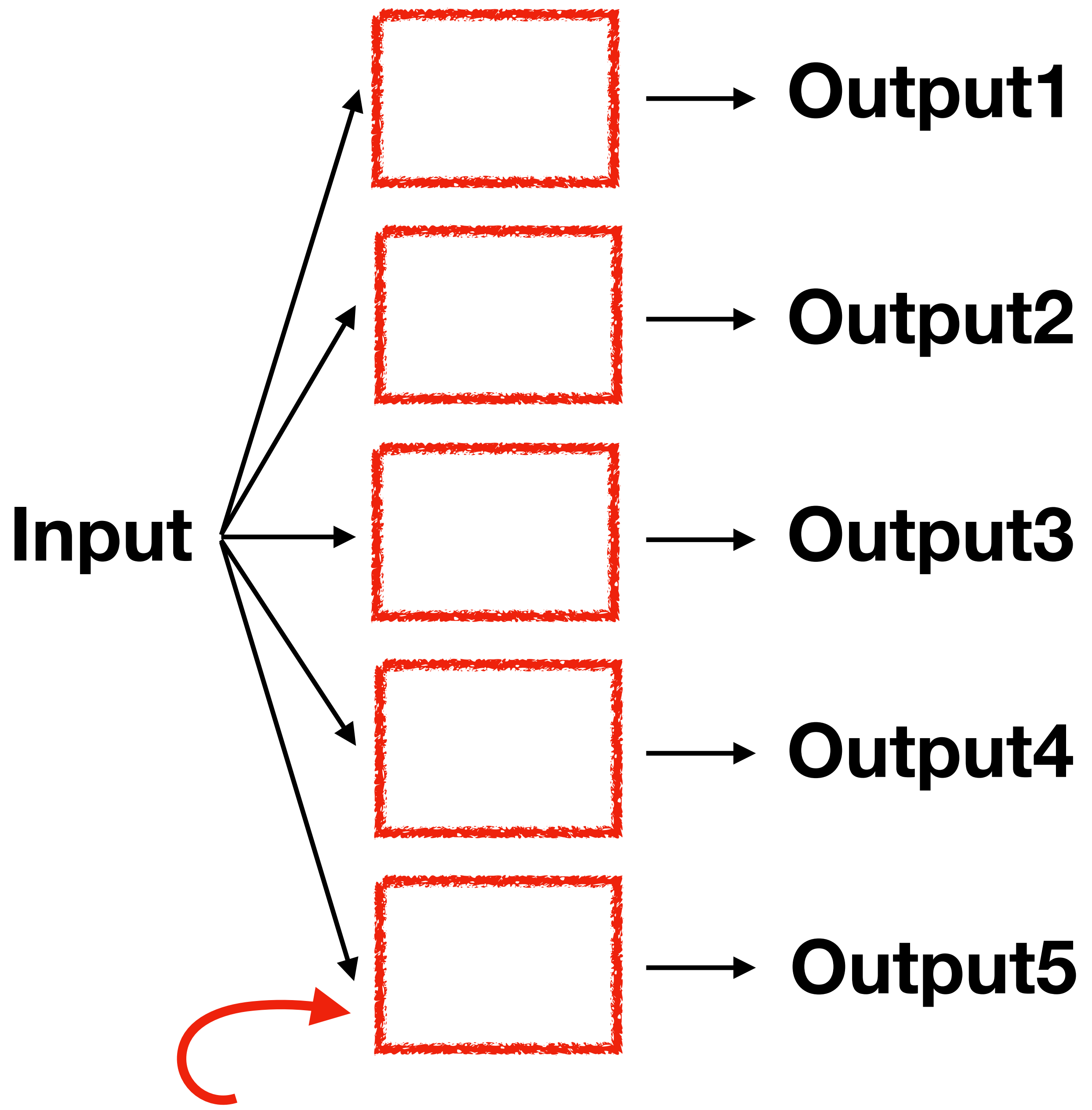
→ Evaluation of the generation error and the effects of the assumed model is important to extract cosmological information from future observational data.

How Precise Is the Reconstructed Map?

Detectability of > 3σ peaks

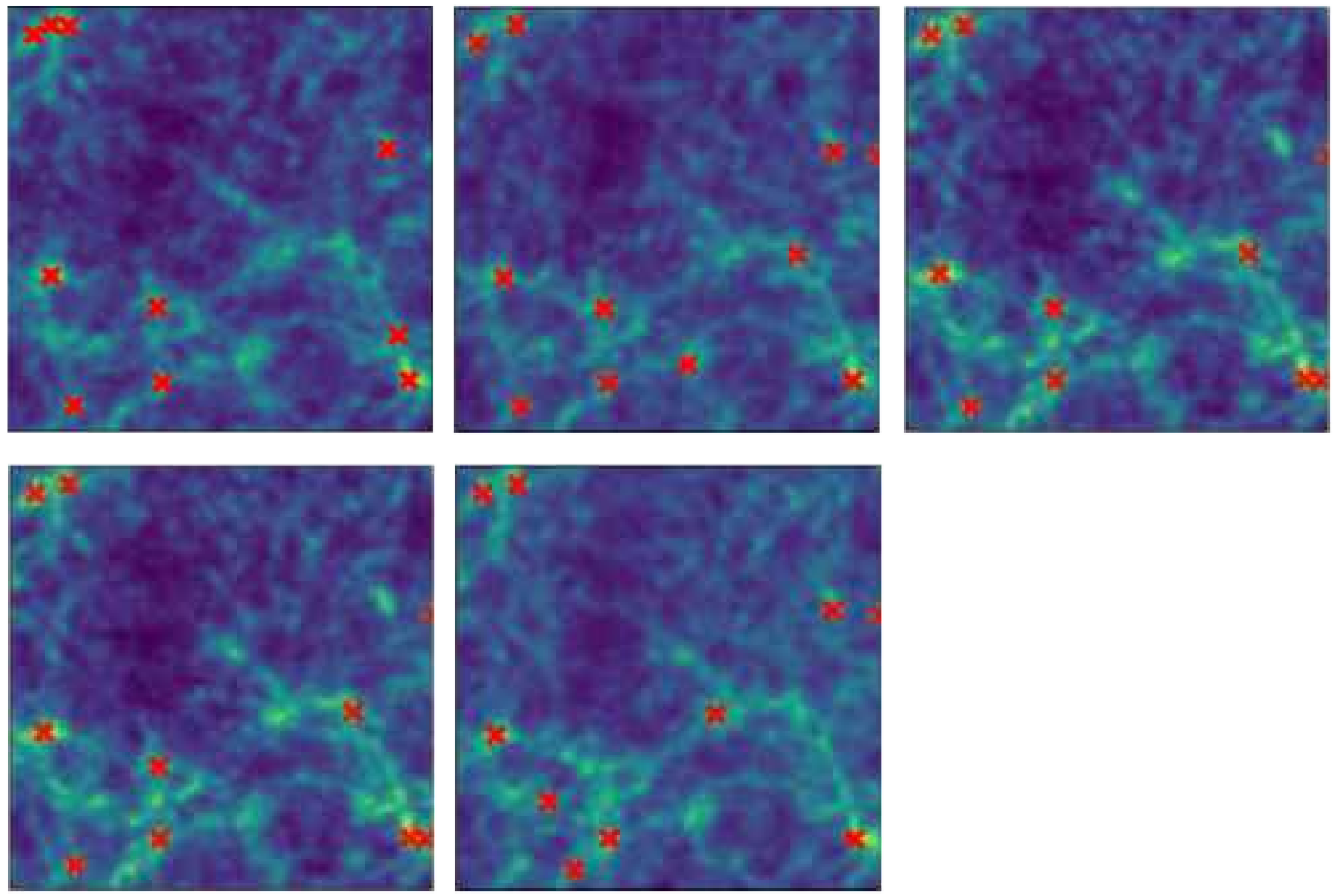
- Precision ($N_{correct}/N_{rec}$) of a machine: **76%**
- Precision when we *combine* five networks (bugging): **91%**



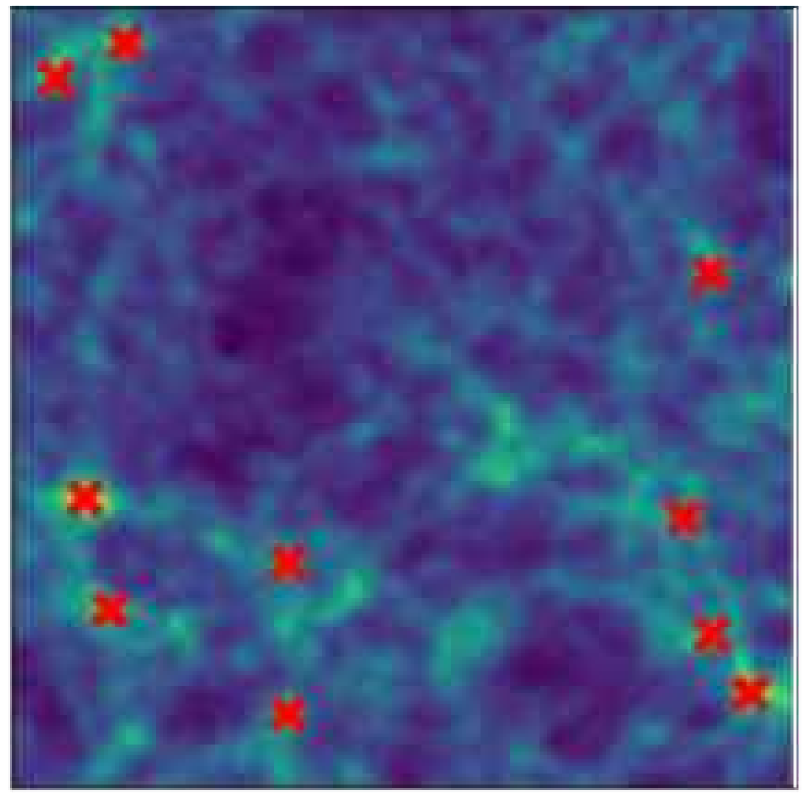


Machines trained on the different data and with different random feeds

outputs of 5 networks



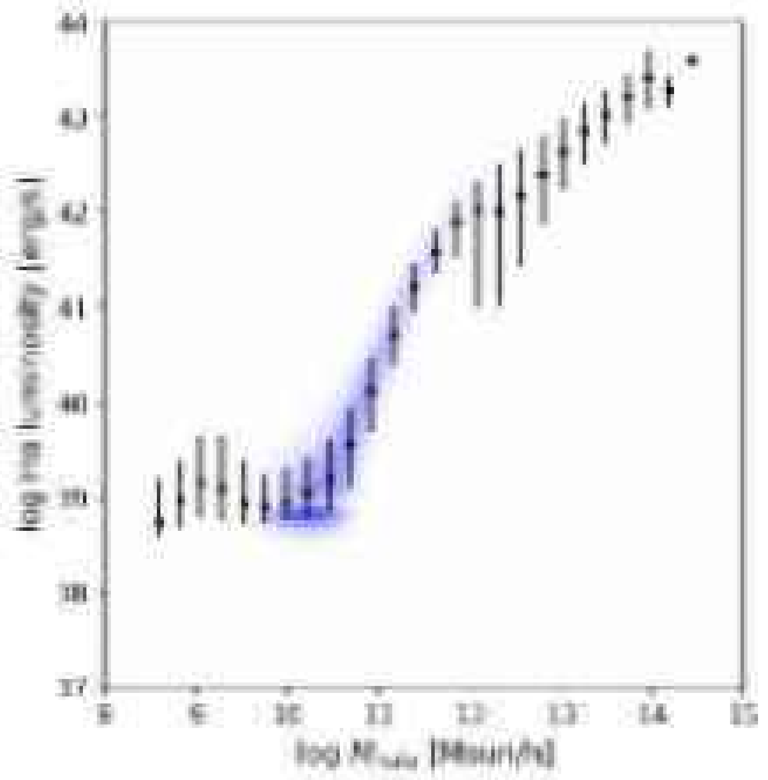
true distributions



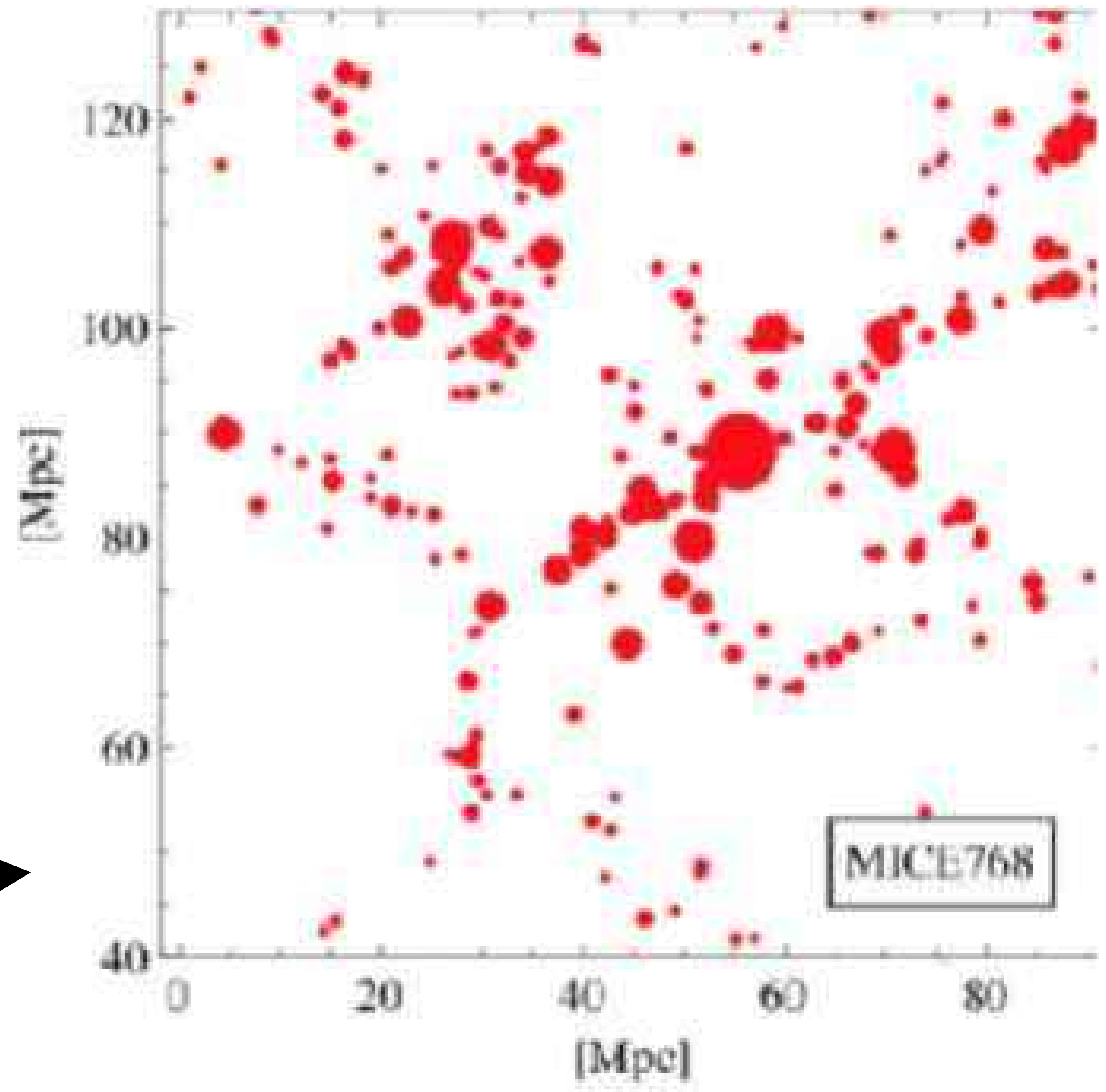
It is also possible to evaluate the generation error by combining multiple networks (e.g., taking the variance).

Does the Reconstruction Depend on the Assumed Model?

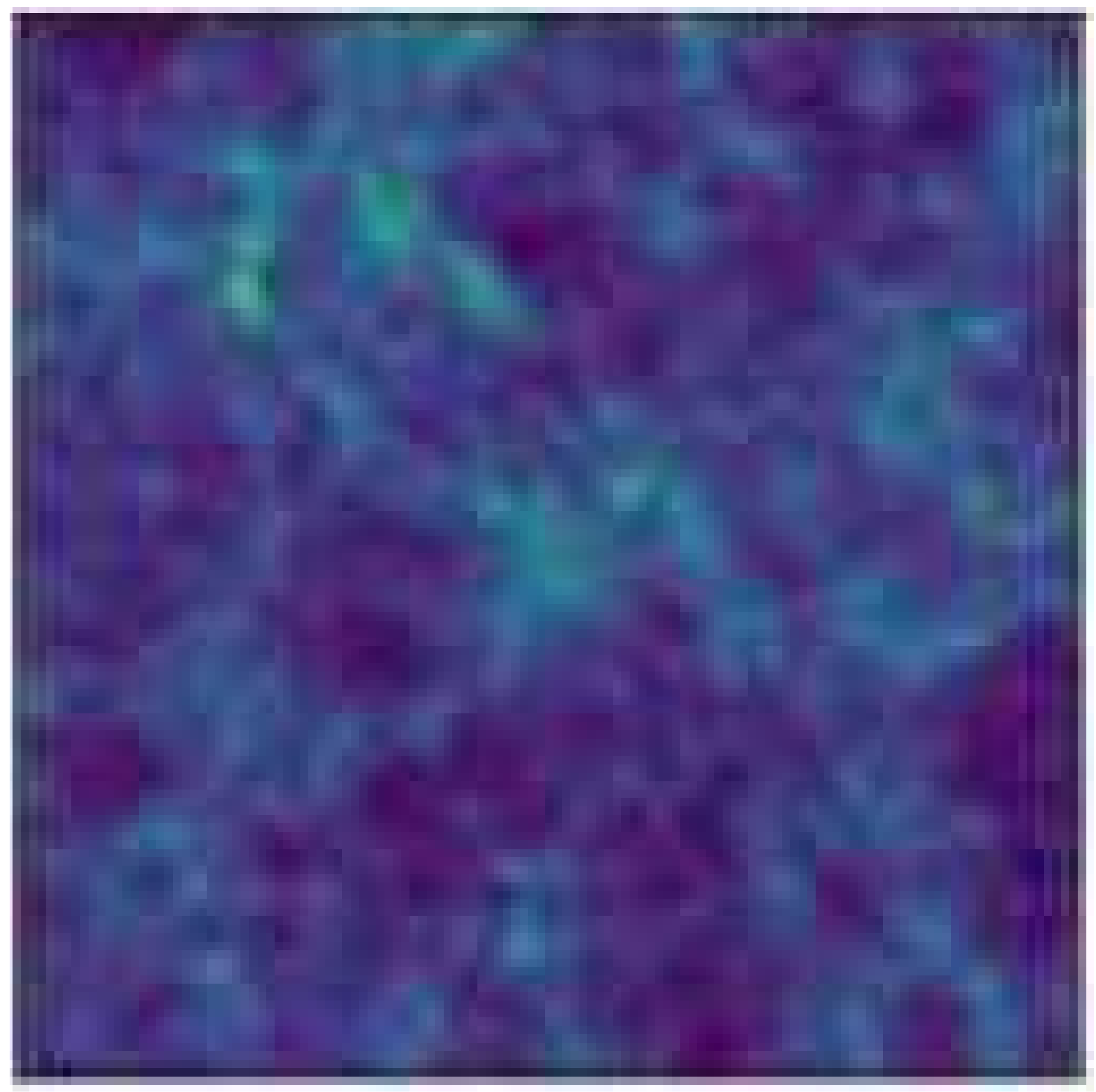
Emission line model:
(what we assume)



DM halo distribution



Mock data



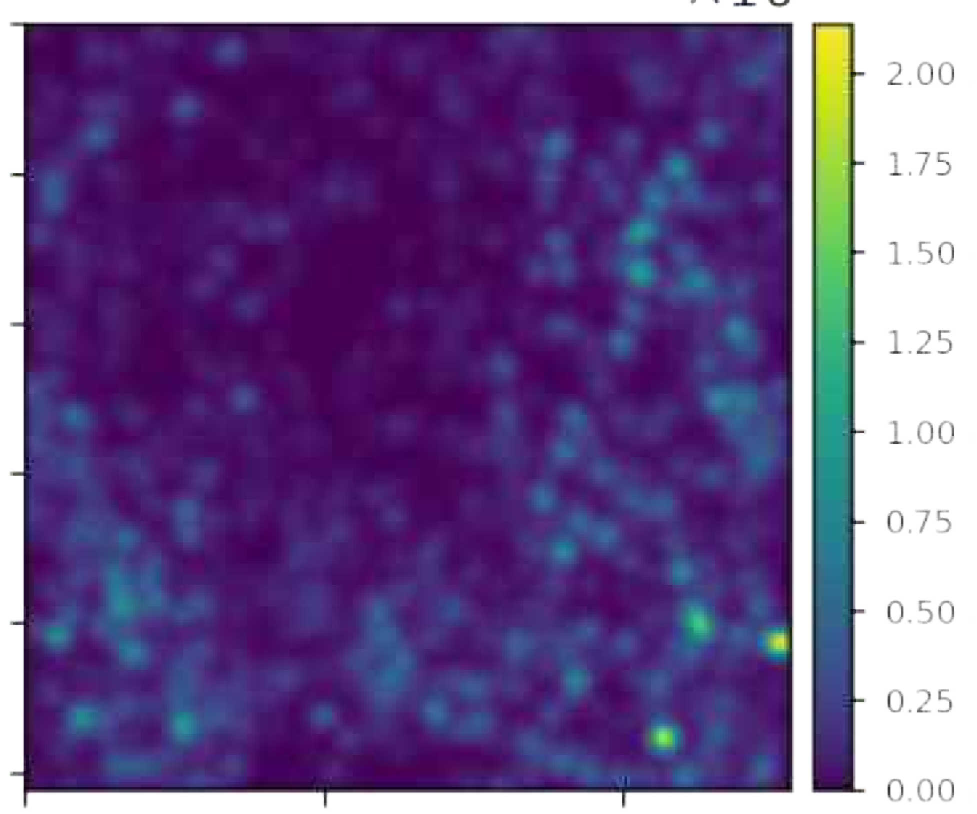
Cosmological information
(what we want to know from the observation)

What if the assumed line emission model in training data is wrong?

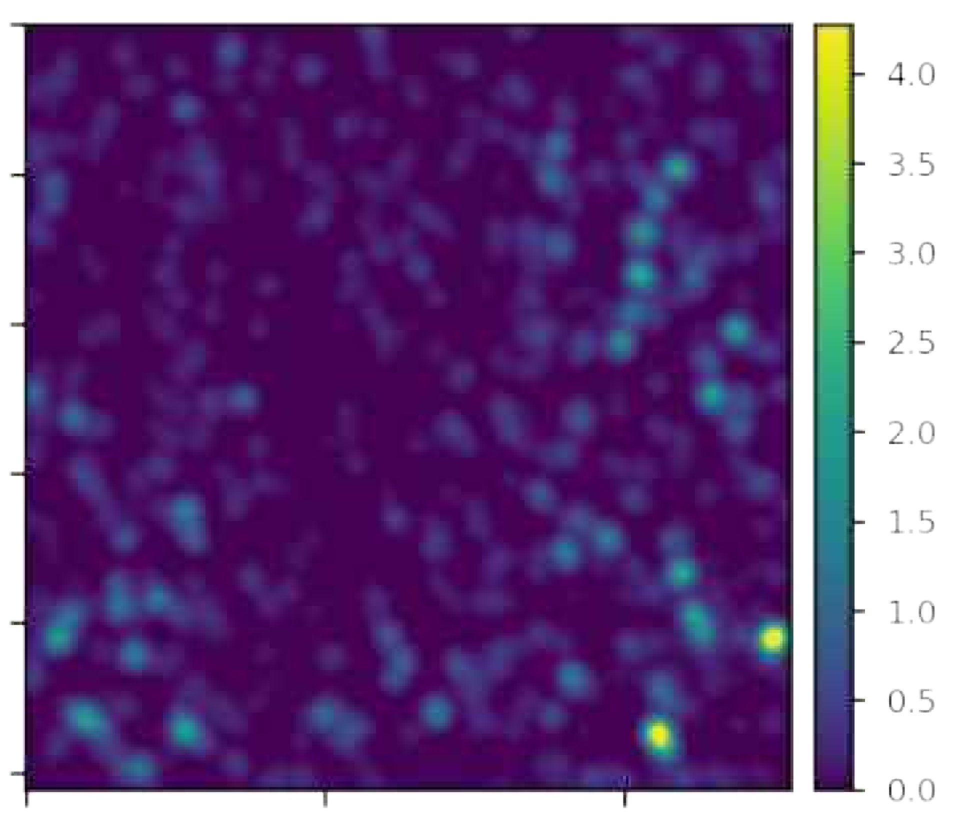
Test with Different Line Emission Models

Model 1 (×2 brighter intensity model)

True

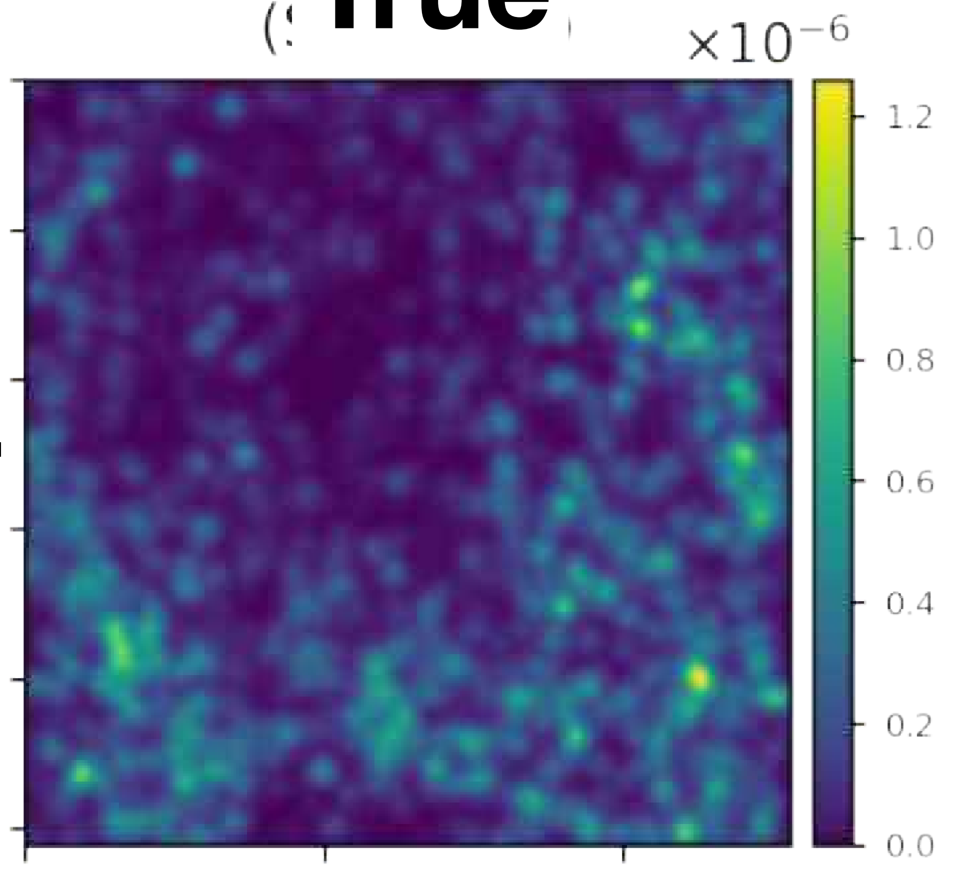


Reconstructed

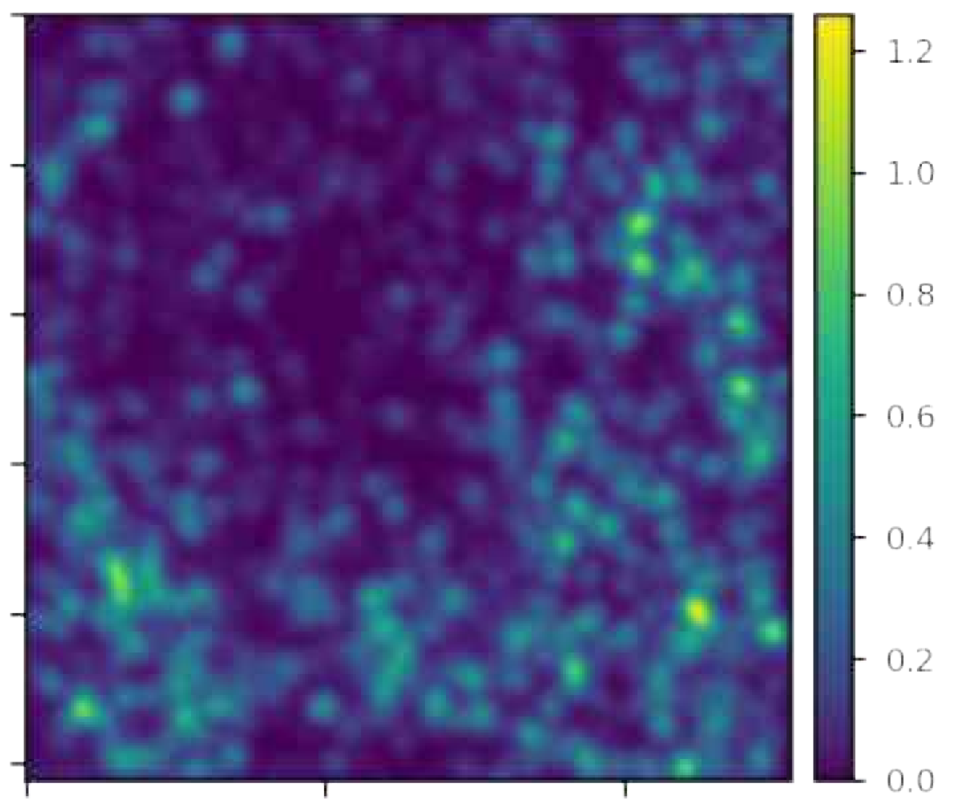


Model 2 (different mass-to-luminosity model)

True

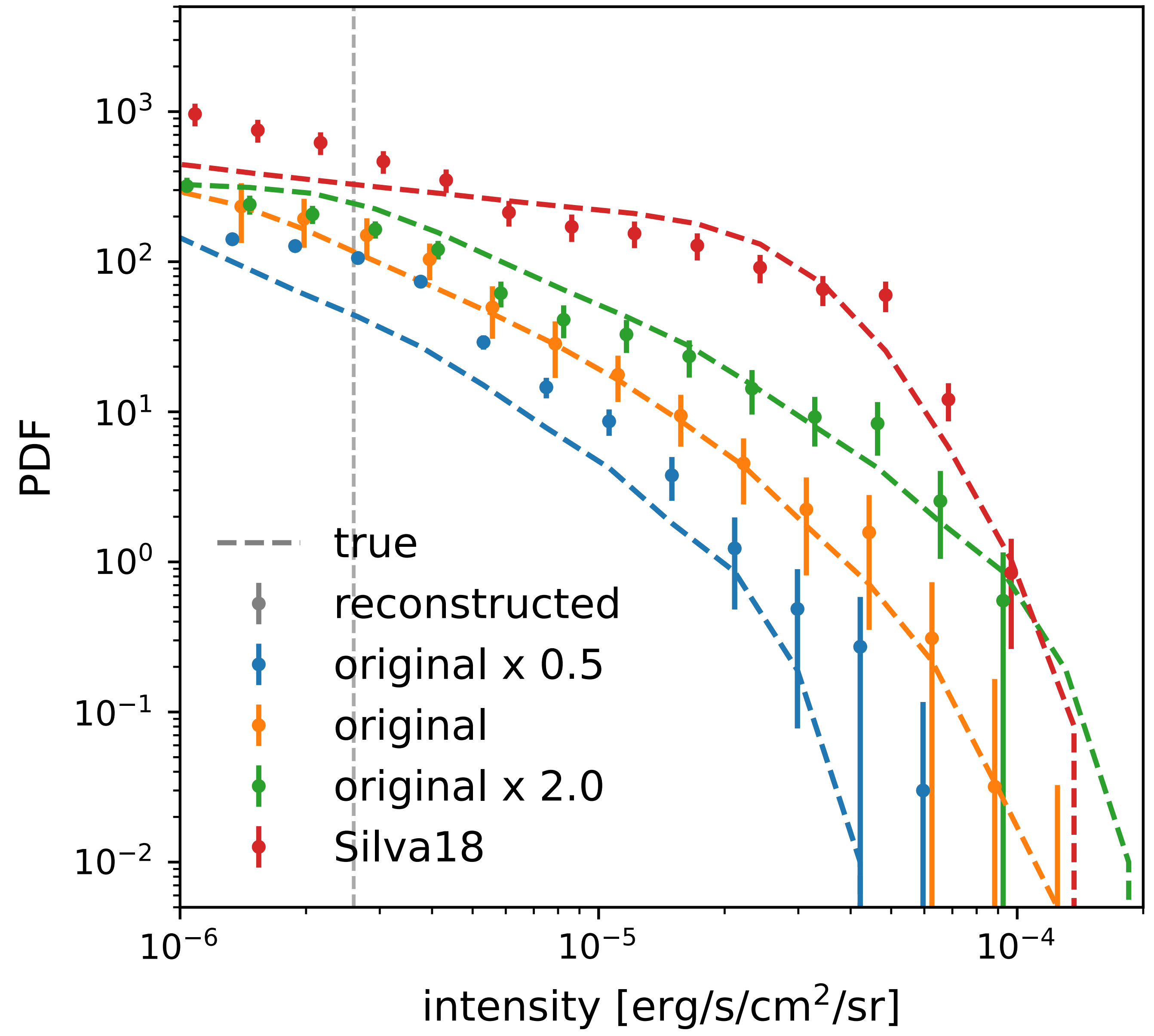


Reconstructed



KM+ 2021

Statistics as well as bright pixel positions are reproduced properly irrespective to the assumed models in test data



What about noise model? More different models? → Future study

Summary

- A generative adversarial network can be used to reconstruct the large-scale distributions of the universe from noisy observational maps.
- We can get good reproducibility by pre-processing the input data based on physical information.
- The machine learns the typical features in the large-scale structure as well as the synchronizing signals in two input data.
- Generation errors and the uncertainties in assumed models should be carefully evaluated in future actual use — combining multiple machines would be an important strategy.