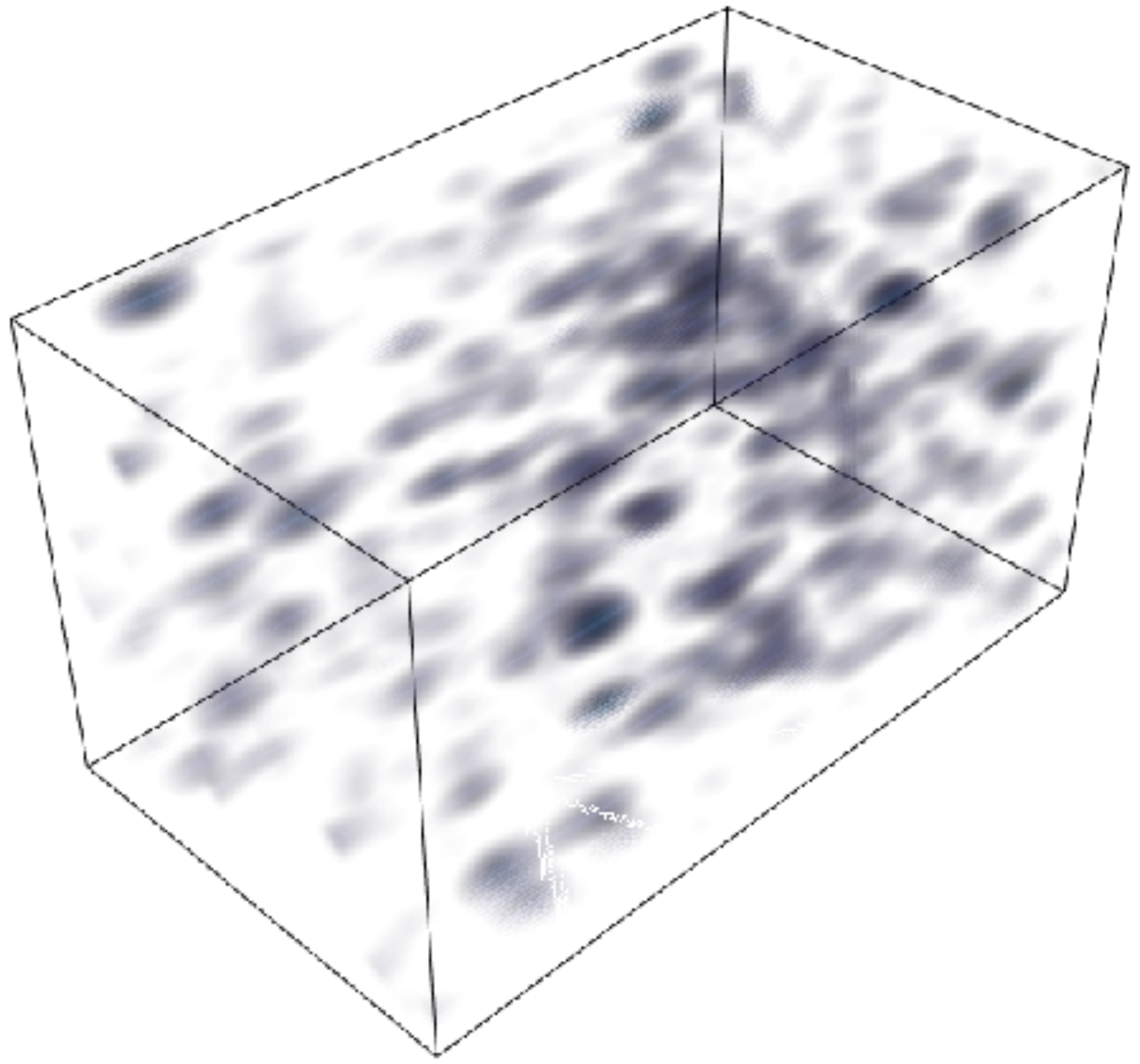
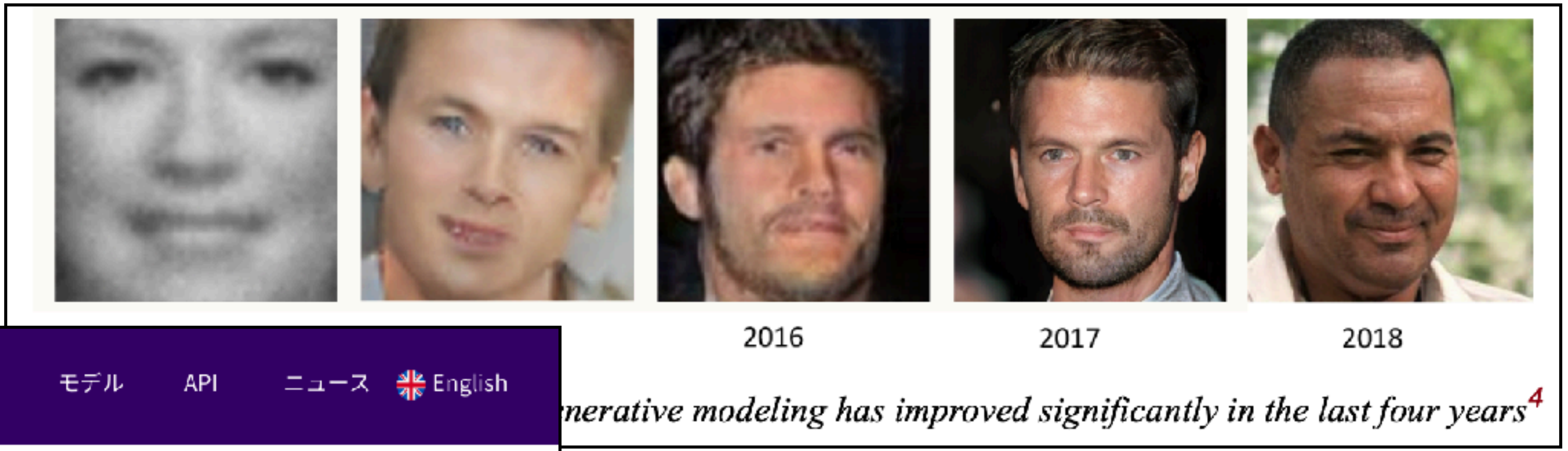
Generative models for large-scale structures

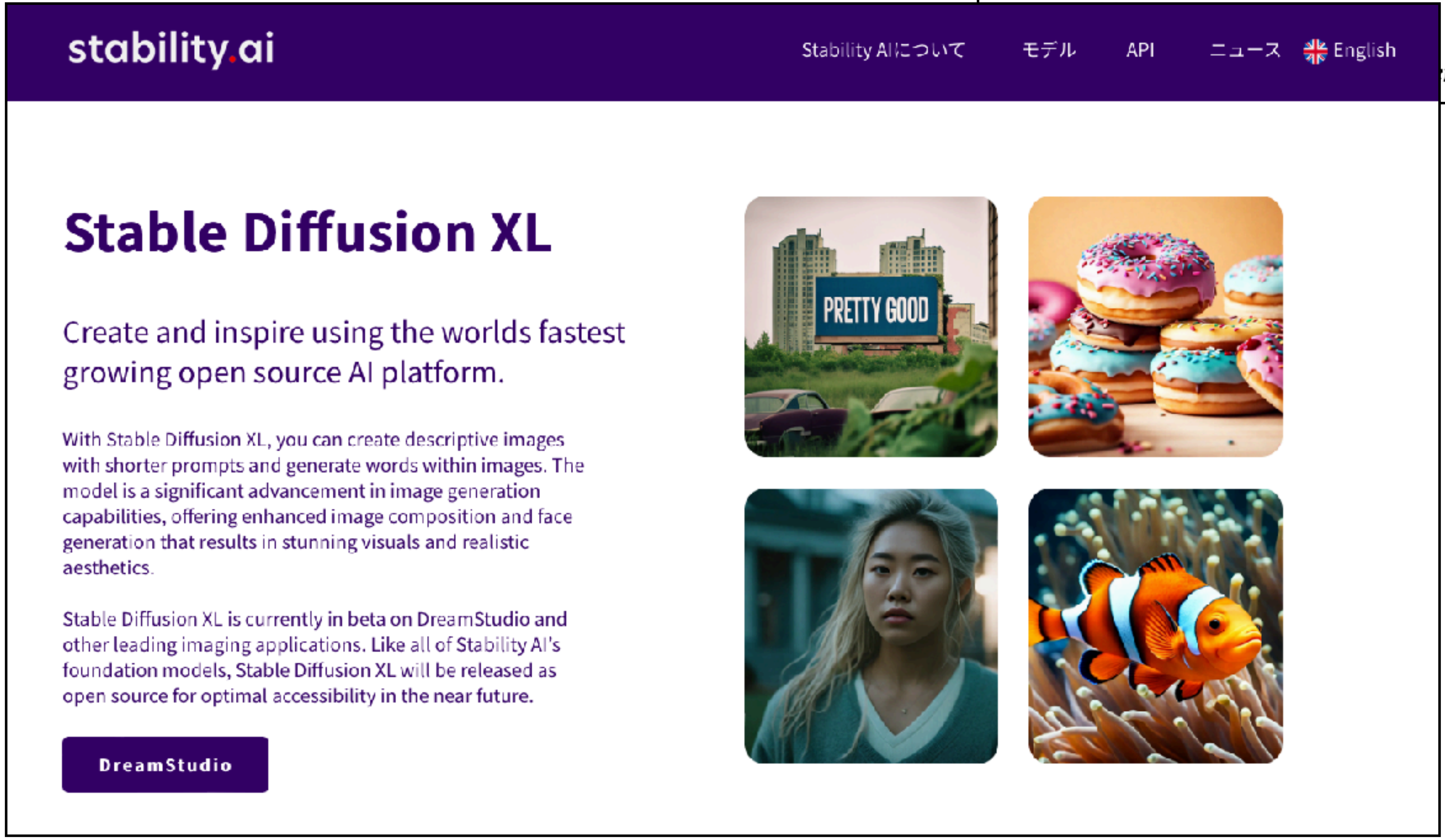
Kana Moriwaki (The University of Tokyo)



Physics in the AI Era
24 - 27 Sep. 2024

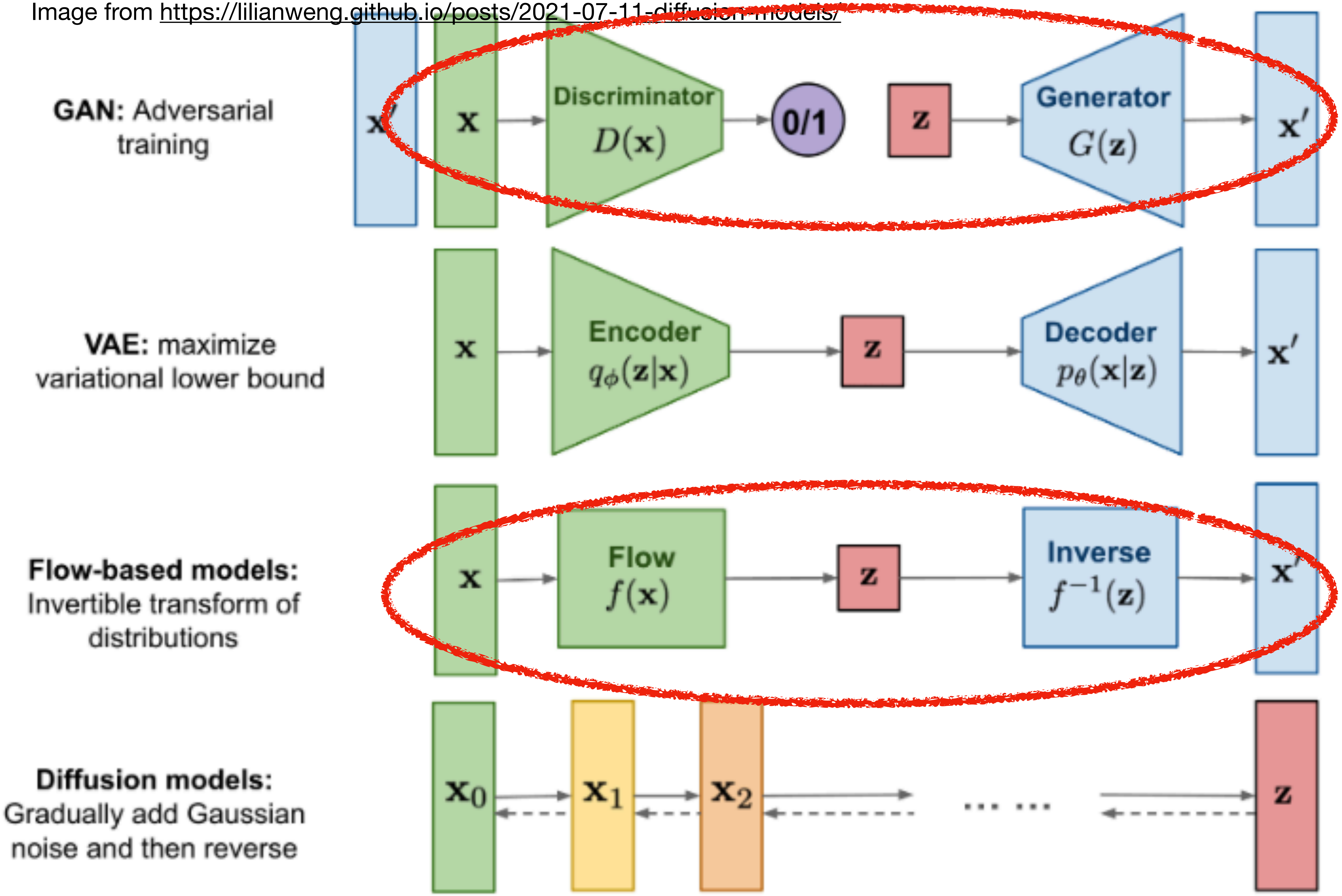
Introduction: generative models





There are several different generative models

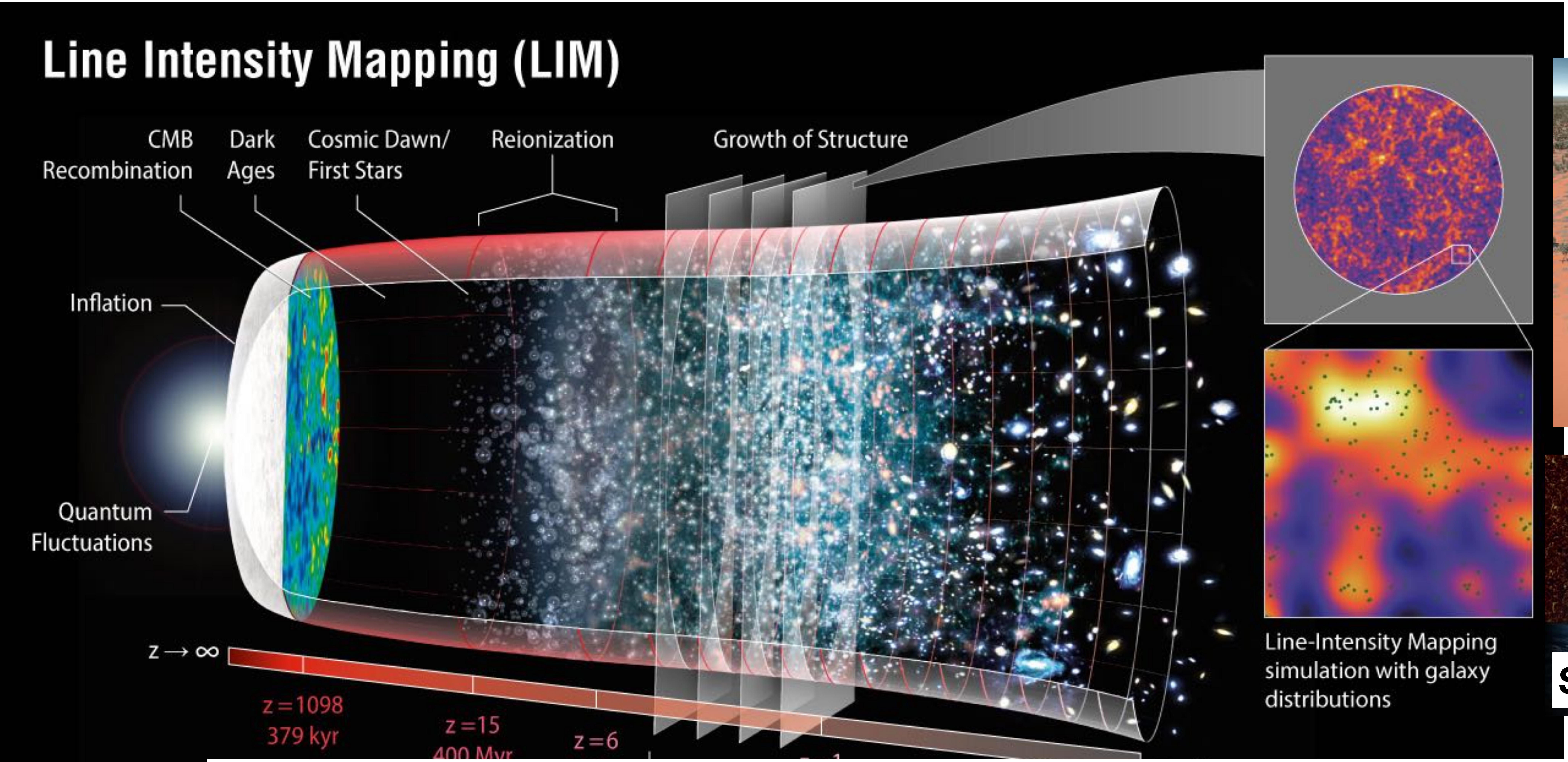
Image from https://lilianweng.github.io/posts/2021-07-11-diffusion-models/



First topic: GAN for line intensity maps ([Moriwaki et al. 2021](#))

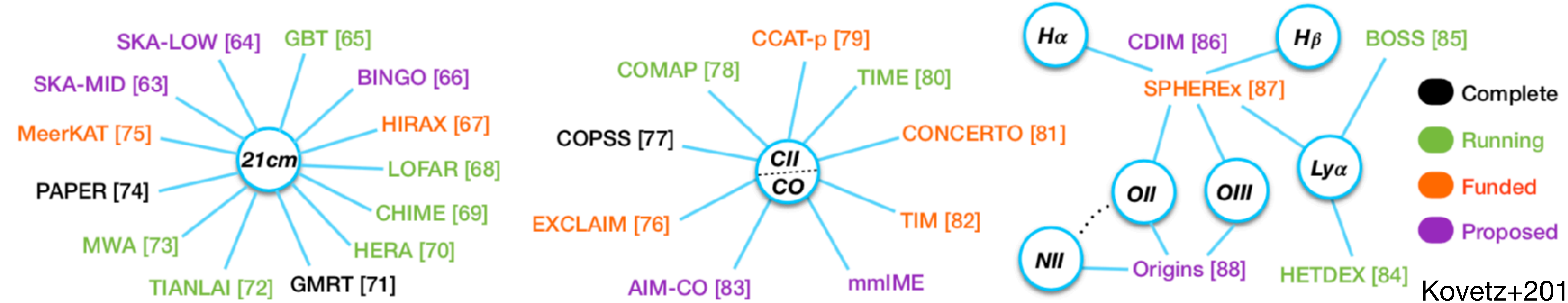
Second topic: NF for EoR parameter inference (Moriwaki in prep)

Conditional GAN for signal reconstruction from line intensity maps


Line Intensity Mapping (LIM)
CMB
Recombination
Dark
Ages
Cosmic Dawn/
First Stars
Reionization
Growth of Structure
Inflation
Quantum
Fluctuations
z → ∞
z = 1098
379 kyr
z = 15
z = 6
Line-Intensity Mapping
simulation with galaxy
distributions

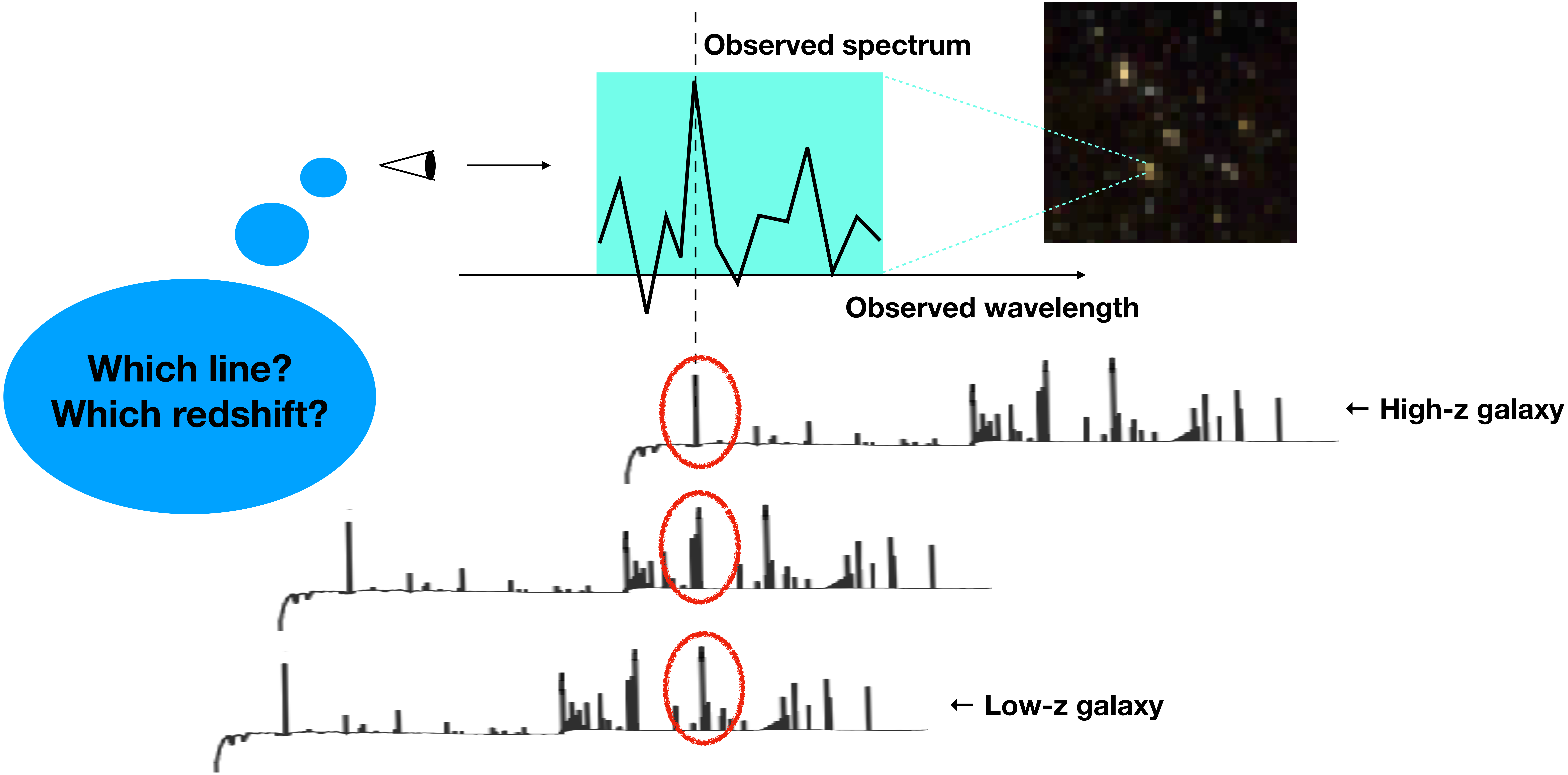

SKA (2028-)


SPHEREx (2025-)


SKA-LOW [64]
GBT [65]
SKA-MID [63]
BINGO [66]
MeerKAT [75]
HIRAX [67]
21cm
LOFAR [68]
PAPER [74]
CHIME [69]
MWA [73]
HERA [70]
TIANLAI [72]
GMRT [71]
CCAT-p [79]
COMAP [78]
TIME [80]
COPSS [77]
CII
CO
CONCERTO [81]
EXCLAIM [76]
TIM [82]
AIM-CO [83]
mmIME
Hα
CDIM [86]
Hβ
BOSS [85]
SPHEREx [87]
OII
OIII
Lyα
NII
Origins [88]
HETDEX [84]
Complete
Running
Funded
Proposed

Kovetz+2019

A Problem: Line Confusion



A Problem: Line Confusion

Example:

observed data at λ = 1.5μm | **Hα intensity at z = 1.3** | **[OIII] intensity at z = 2.0**

SPHEREx

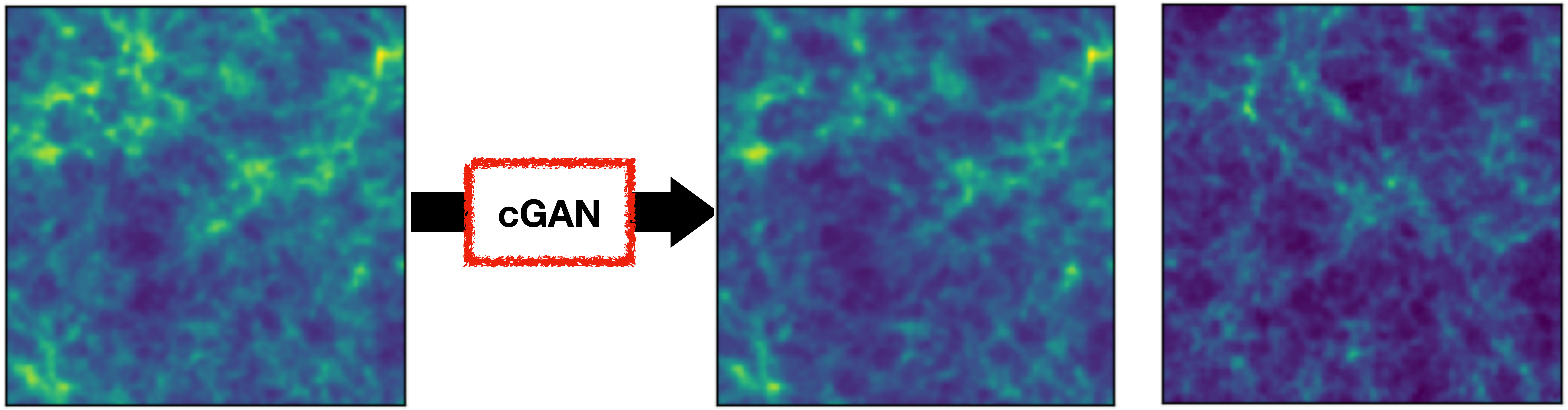
Can we separate different signals on pixel-by-pixel basis?

Our solution: deep learning

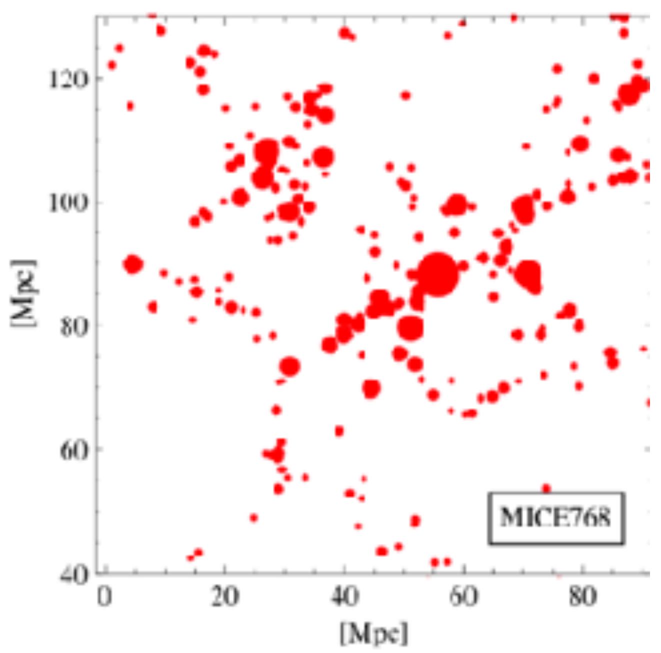
observed data
at λ = 1.5μm

Hα intensity
at z = 1.3

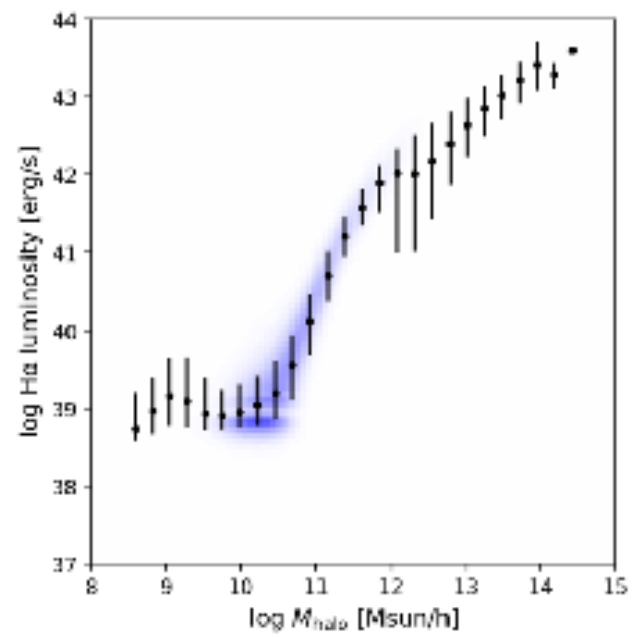
[OIII] intensity
at z = 2.0



halo catalogs



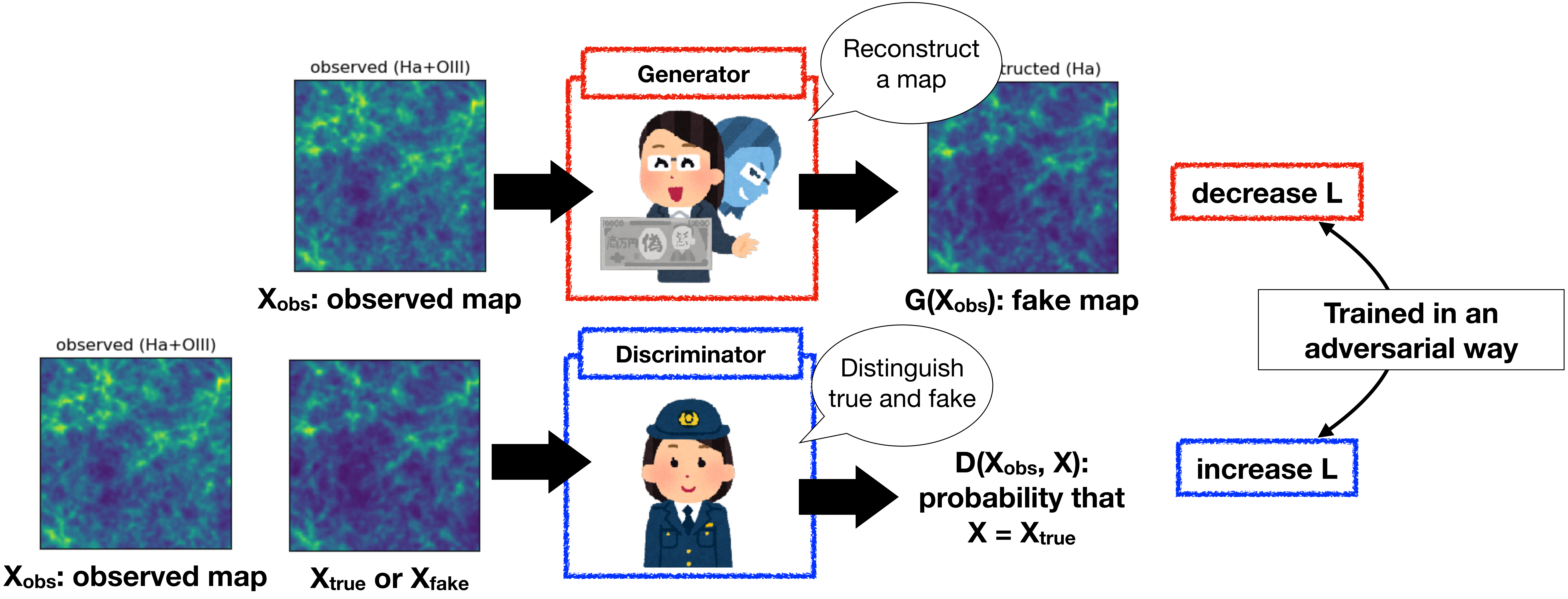
M-L relation



- Generate ~30,000 mock observational maps using halo catalog generation code PINOCCHIO (Monaco+13) and halo mass vs line luminosity relation derived from a hydrodynamics simulation.

Conditional generative adversarial network

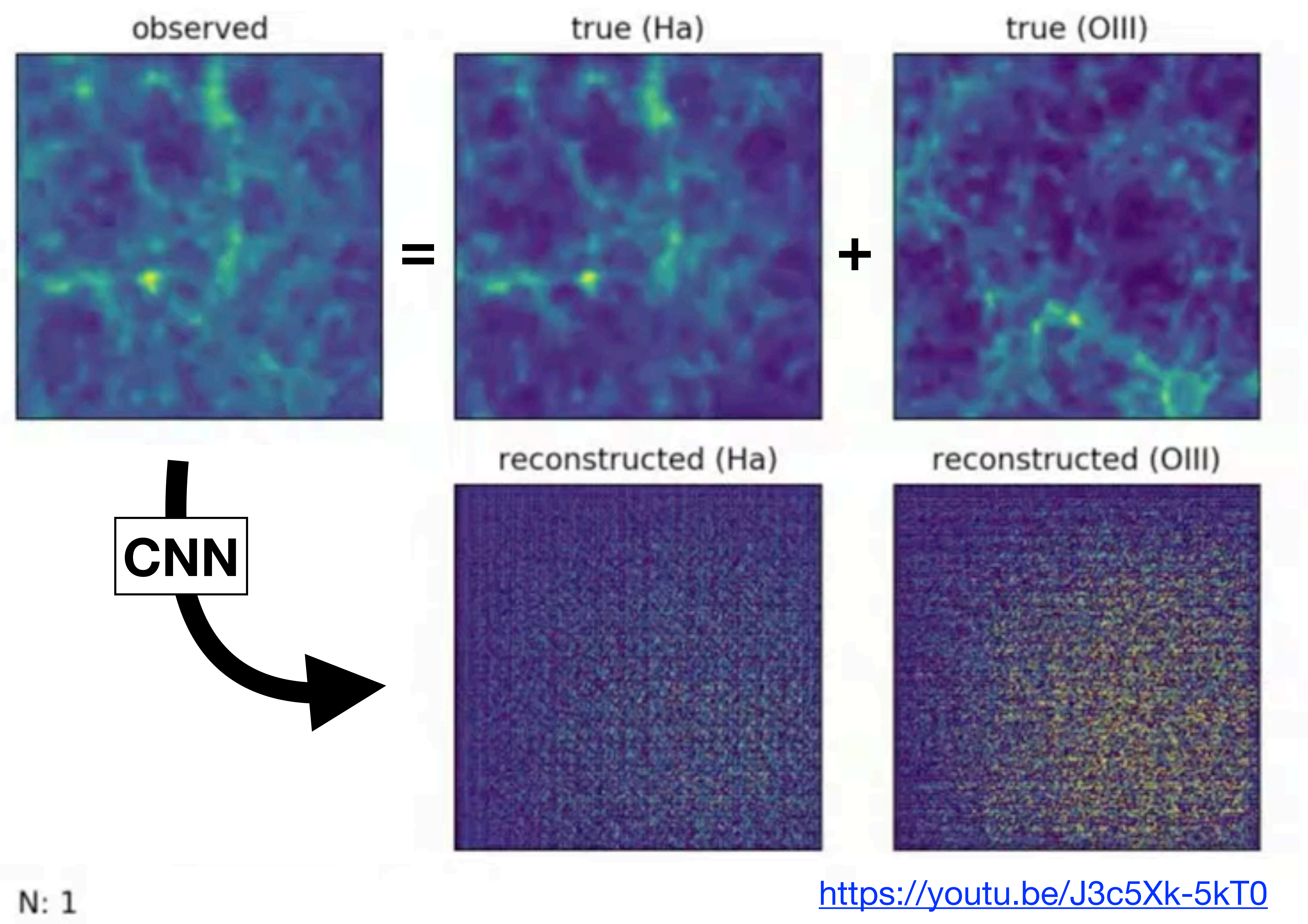
- GAN: Generator and Discriminator are updated in an adversarial way.



Loss function:

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

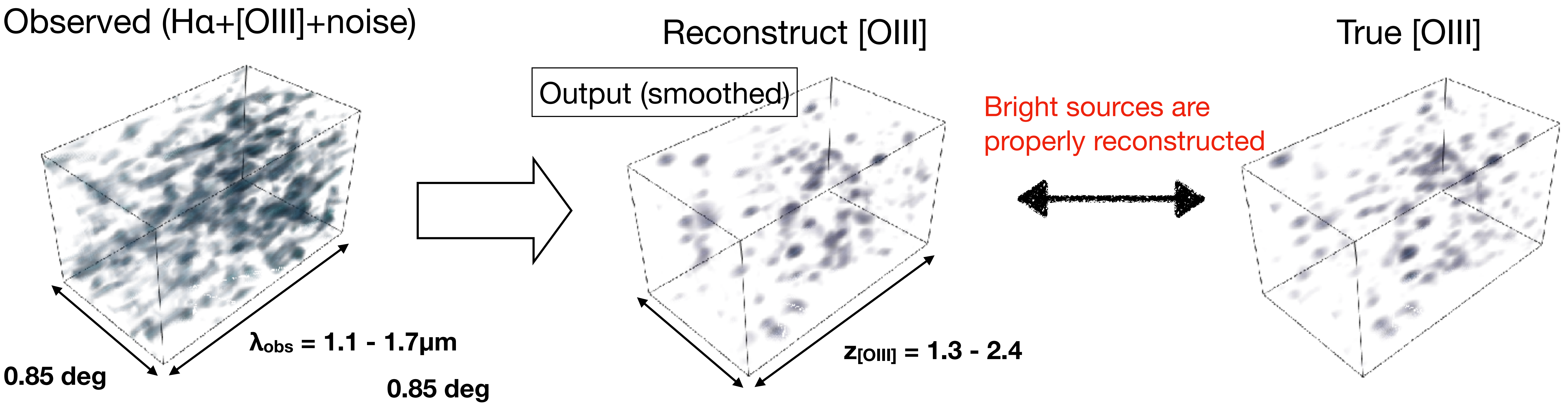
Reconstruction of the Large-Scale Distributions of Emission-Line Galaxies



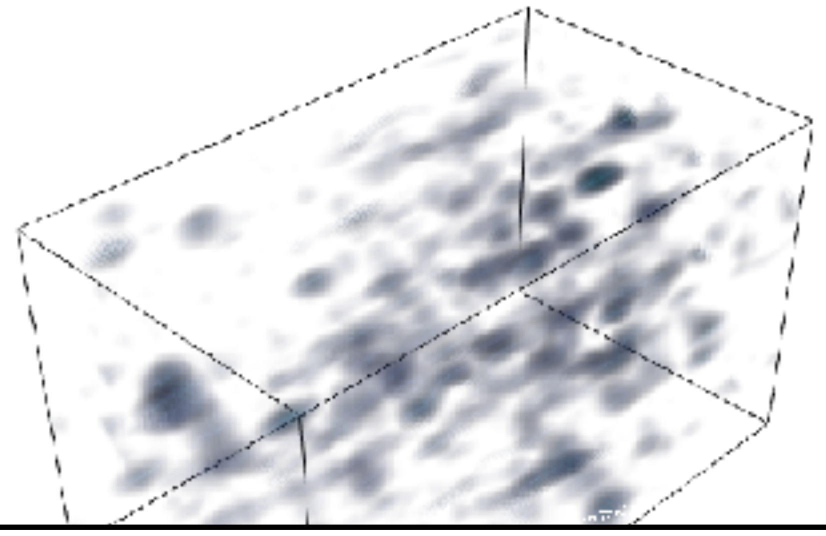
https://youtu.be/J3c5Xk-5kT0

Result: reconstruction of 3D data cube

KM & Yoshida, ApJ, 923, L7



cf) True Hα+[OIII]



We are planning to apply this method to the future LIM observations.

Reproducibility improved!!

Peak detectability of Hα and [OIII]

precision = 82%, 68% (cf. 2D: 76%, 27%)

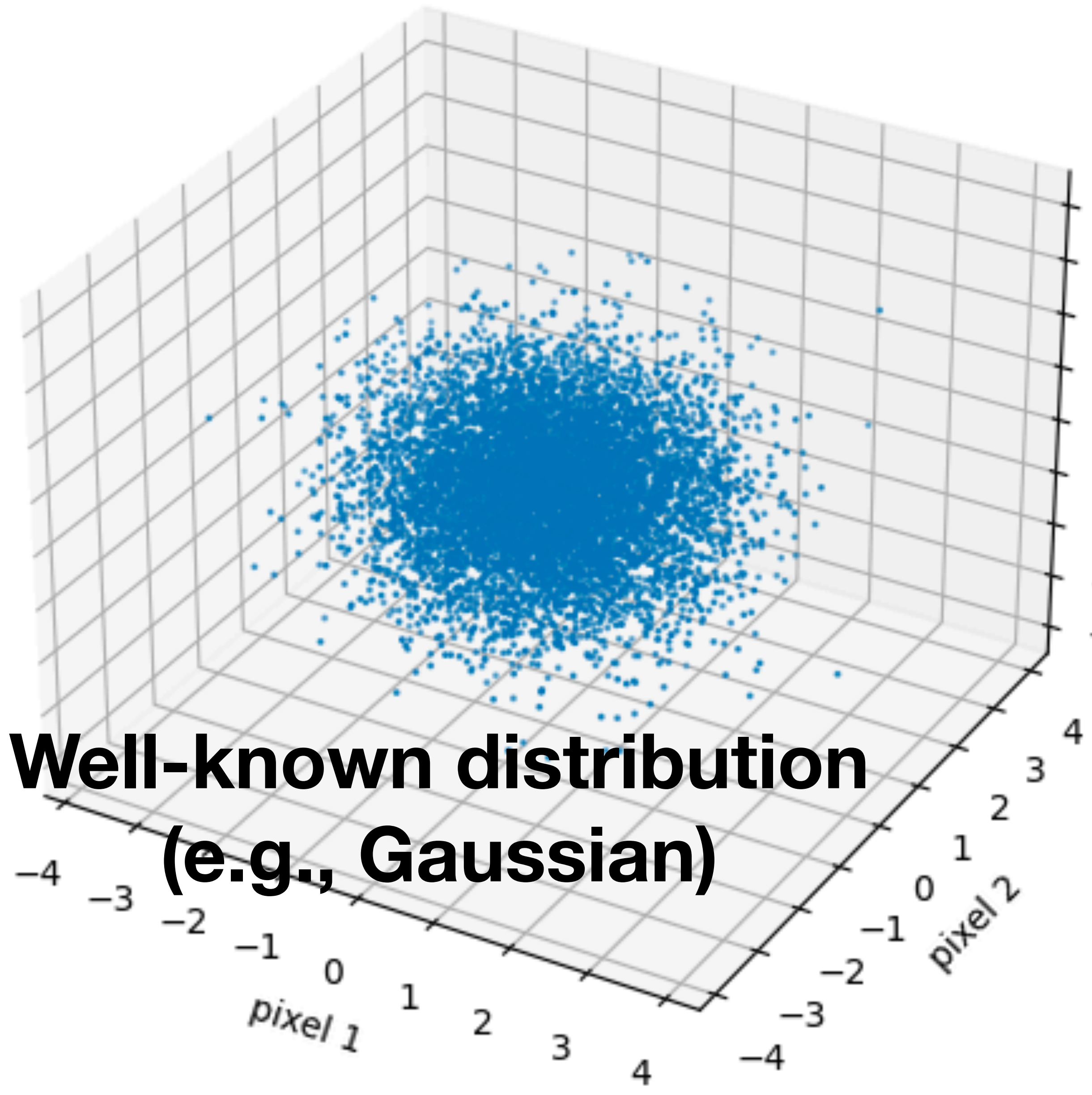
recall = 80%, 77% (cf. 2D: 74%, 29%)

Normalizing flow
for EoR parameter inference

Data space ⇔ Latent space

Flow-based models transfer data into latent space with invertible functions.

latent space



Well-known distribution (e.g., Gaussian)

f(x)

f^{-1}(z)

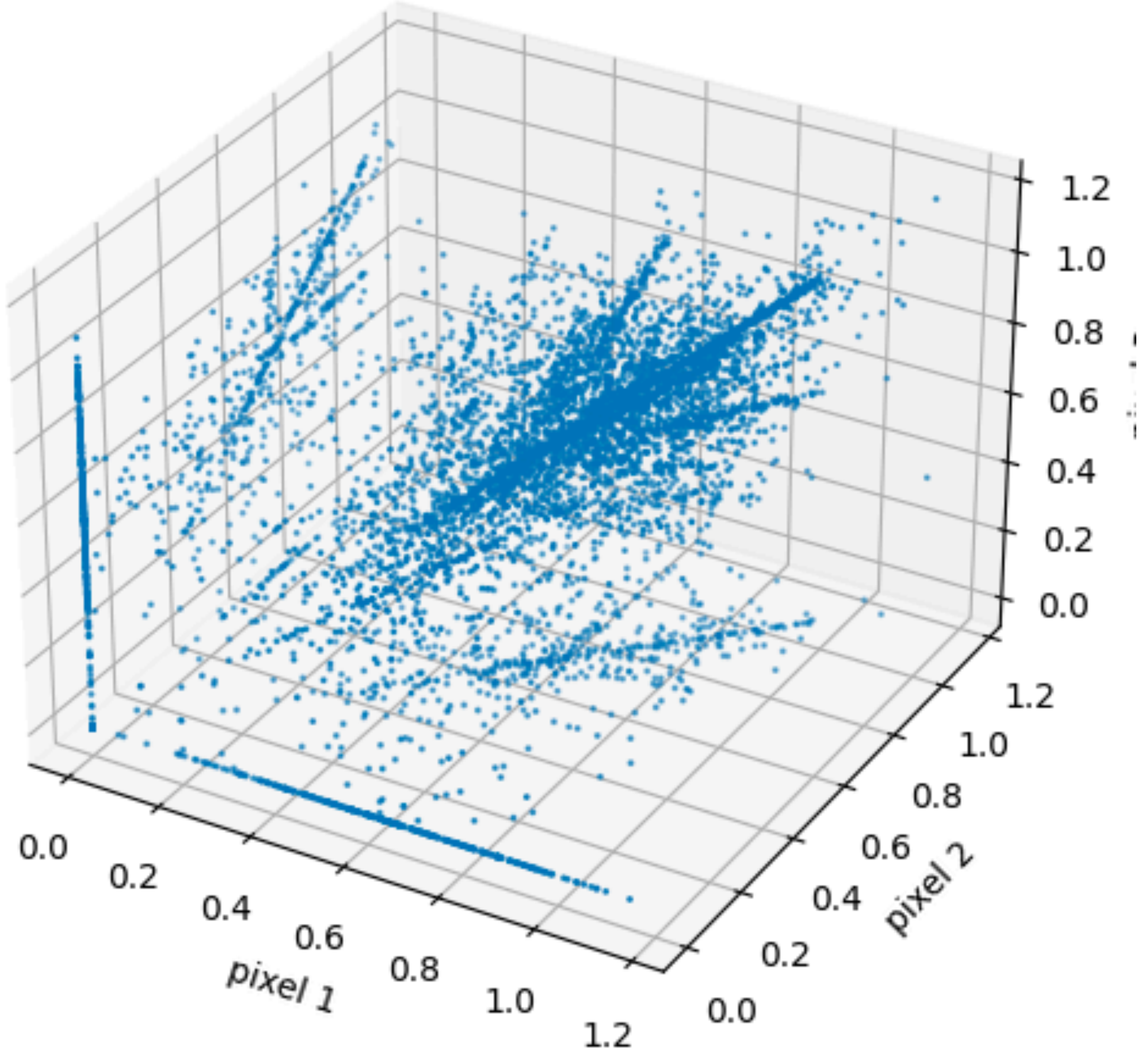
f: invertible

e.g., $f(x) = \alpha x \ (\alpha \neq 0)$

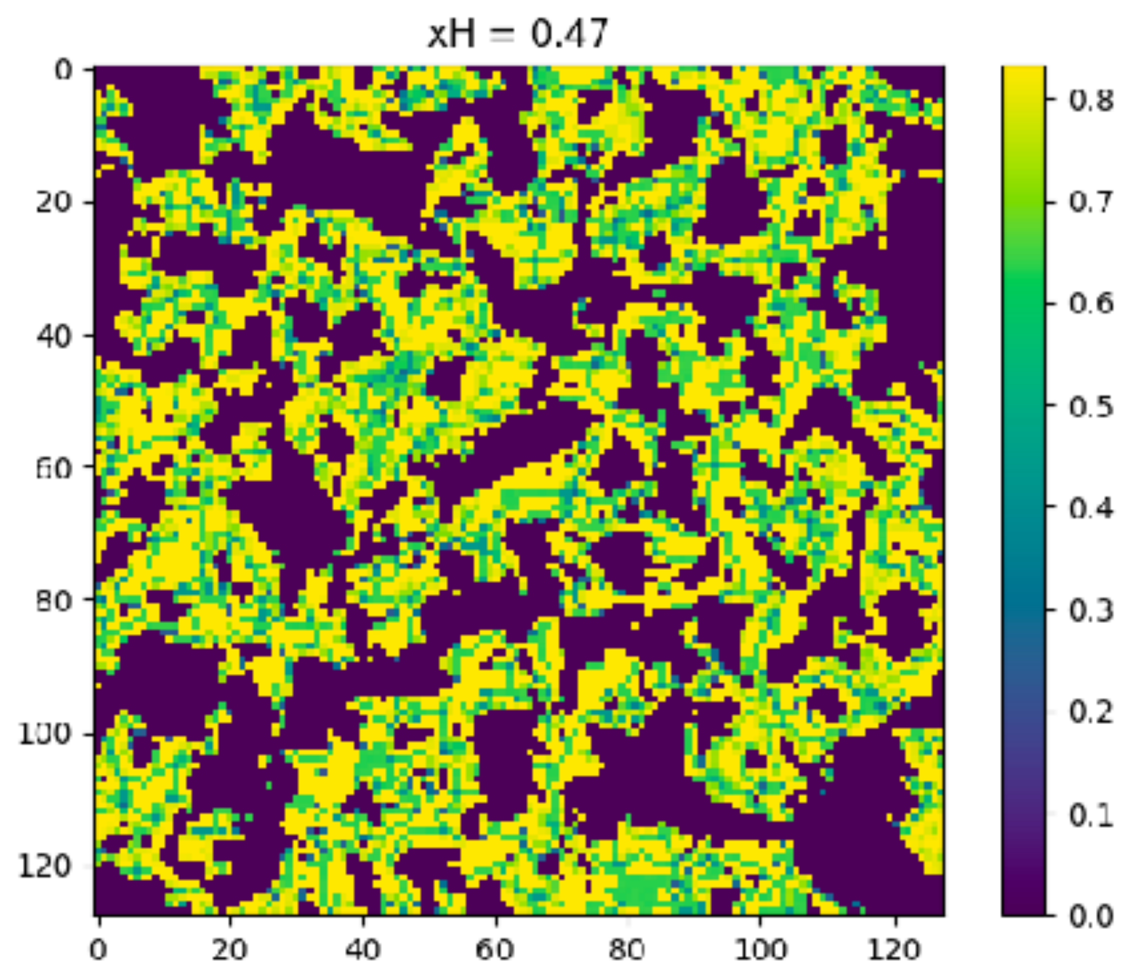
$$f(x) = \begin{pmatrix} 3 & 2 \\ 0 & 4 \end{pmatrix} x$$

$f:$ permutation

data space:
256×256 dimension



← distribution of pixel values of 21cm map

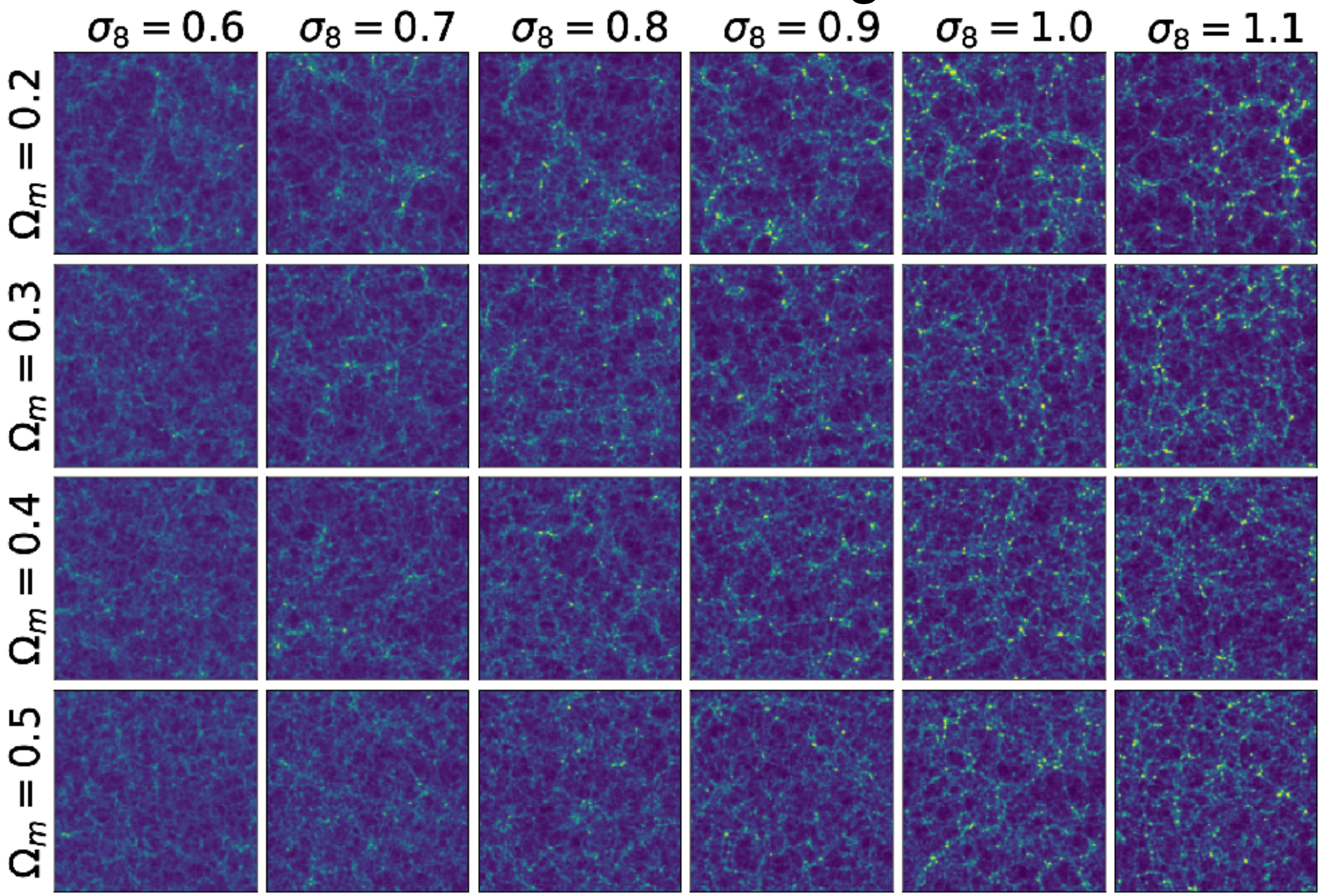


- **When generating a new data, one can sample a random point in the latent space and apply "f^{-1}"**
- **Invertible functions can be used to infer the probability density of the data.**

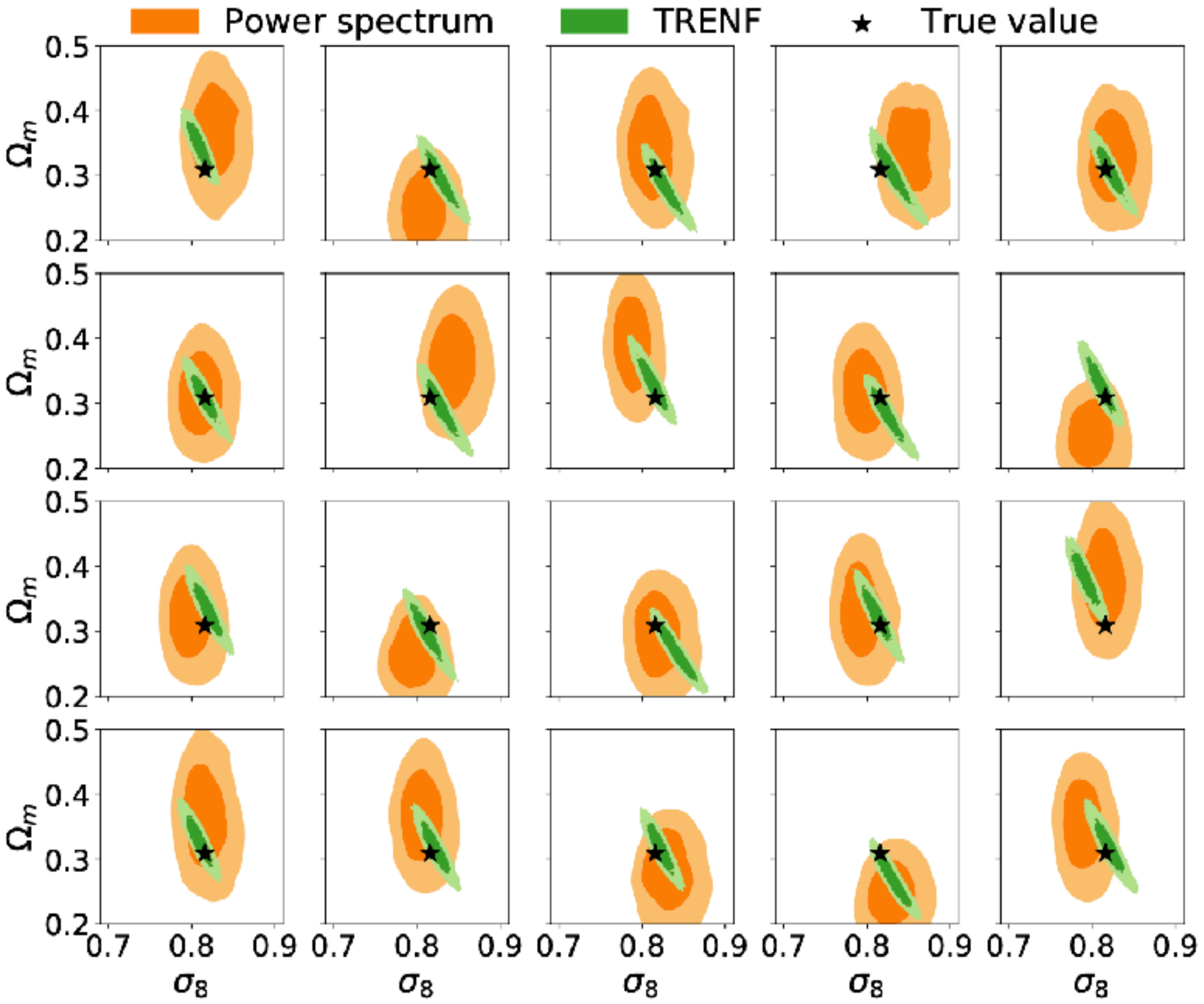
Flow-based models for astrophysics

Dai+2022

Generated LSS images



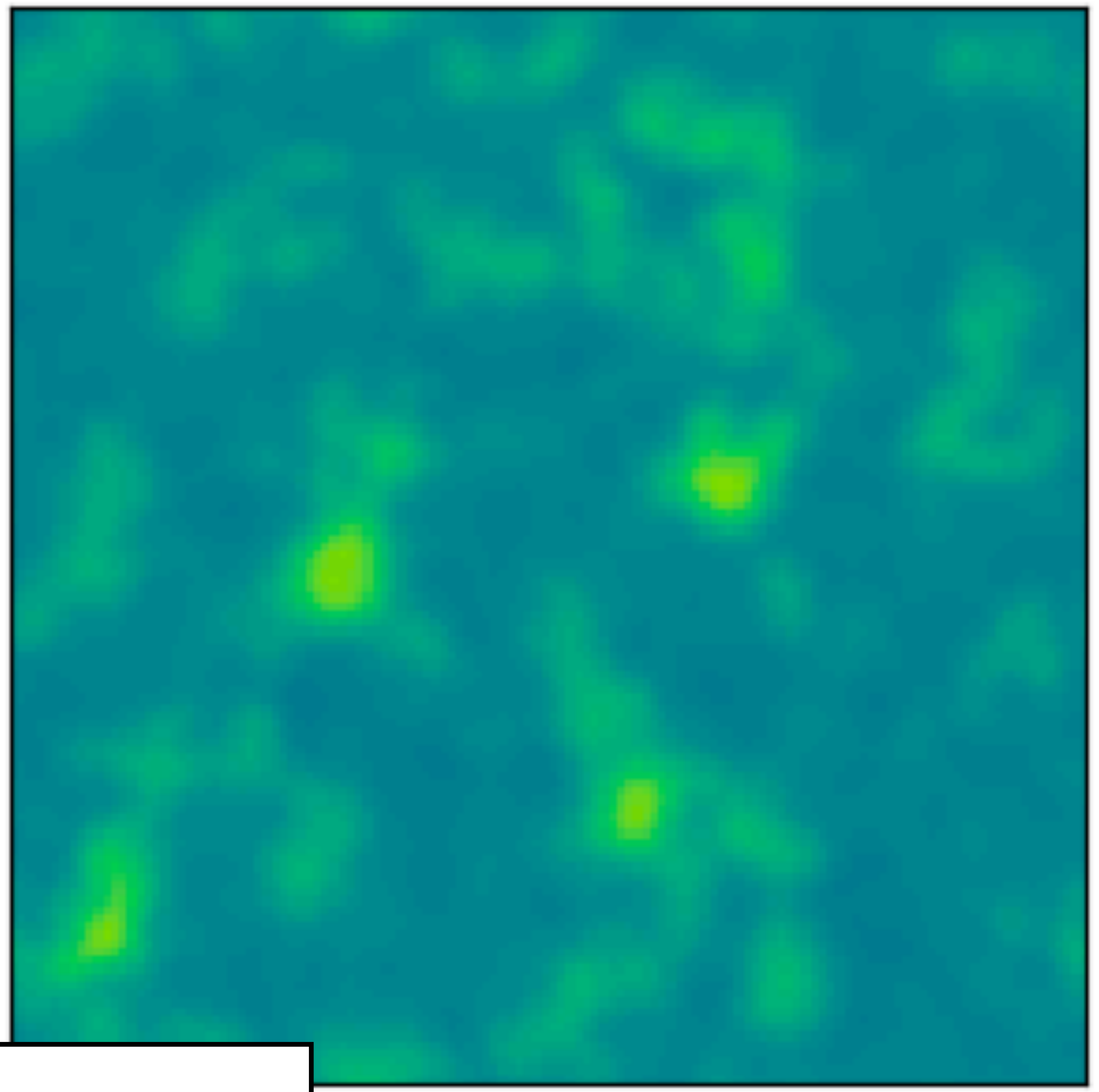
Cosmological parameter inference



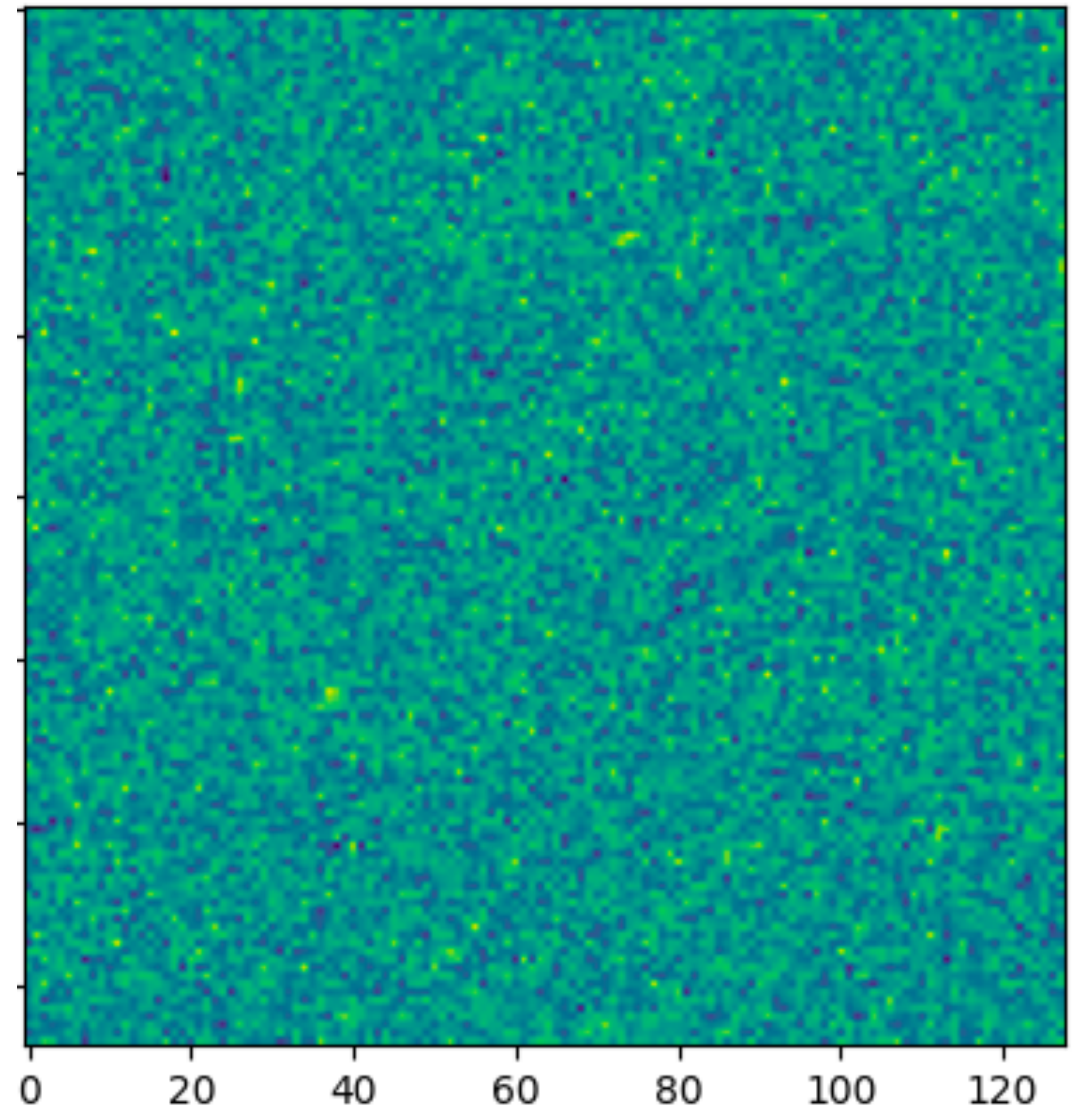
Dai+2022 use Translation and Rotation Equivariant Normalizing Flow (TRENF)
See also Hassan et al. (2022) for application of flow-based model to post reionization 21cm maps

Let us use NF for EoR 21cm map!

Training data
21cm map (21cmFAST)
+ noise + smoothing



Latent variable



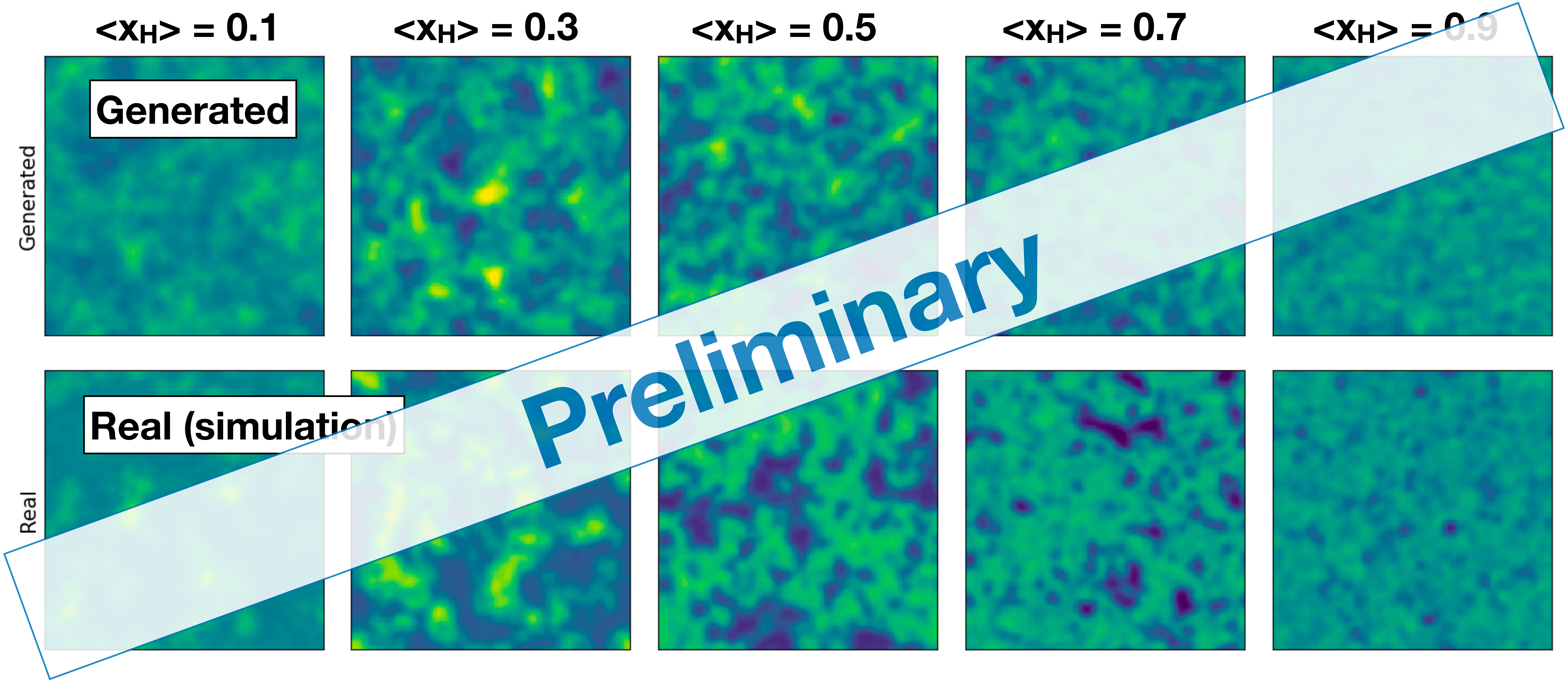
Transformation parameters
$T_1, T_2, \ldots, T_N$

NN

neutral fraction

Training data:
21cmFAST with different parameters of f^*. α^*, f_{esc}, α_{esc}, t^*, M_{turn}, L_X, E_0, α_X

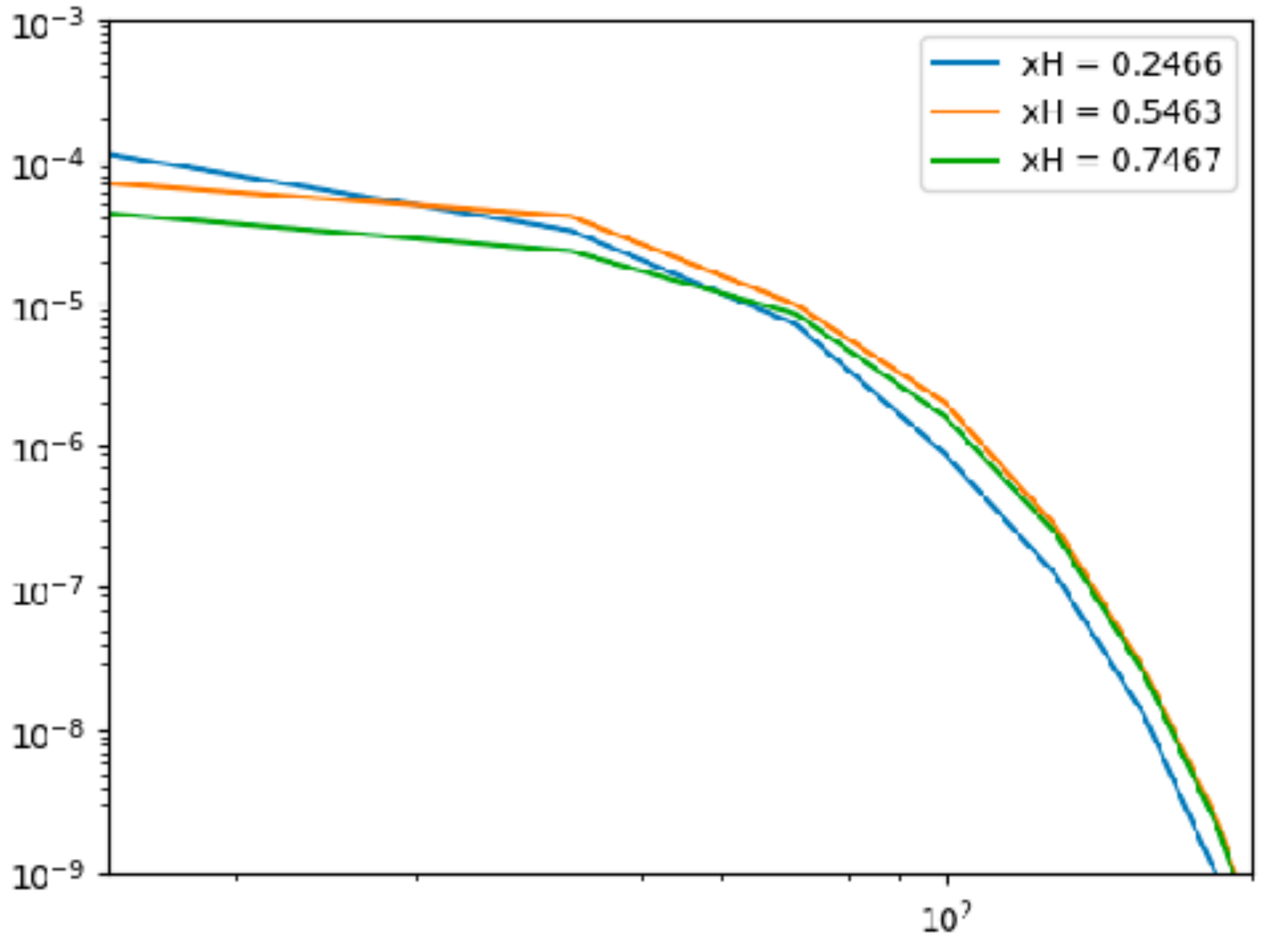
Result: generated images



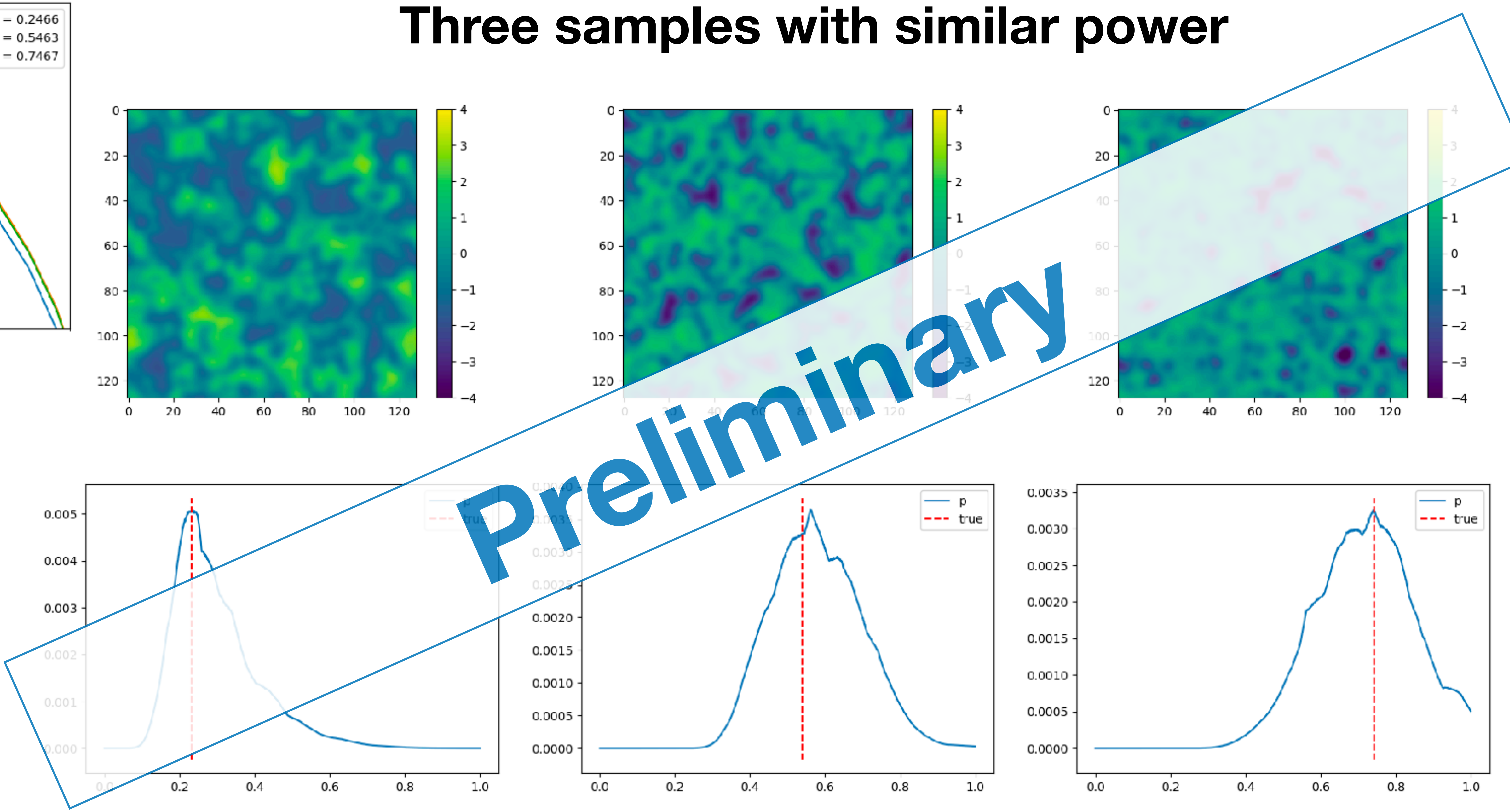
Result: parameter inference

$x_H = 0.1$ $x_H = 0.3$ $x_H = 0.5$ $x_H = 0.7$ $x_H = 0.9$

Is it learning something beyond power spectrum?

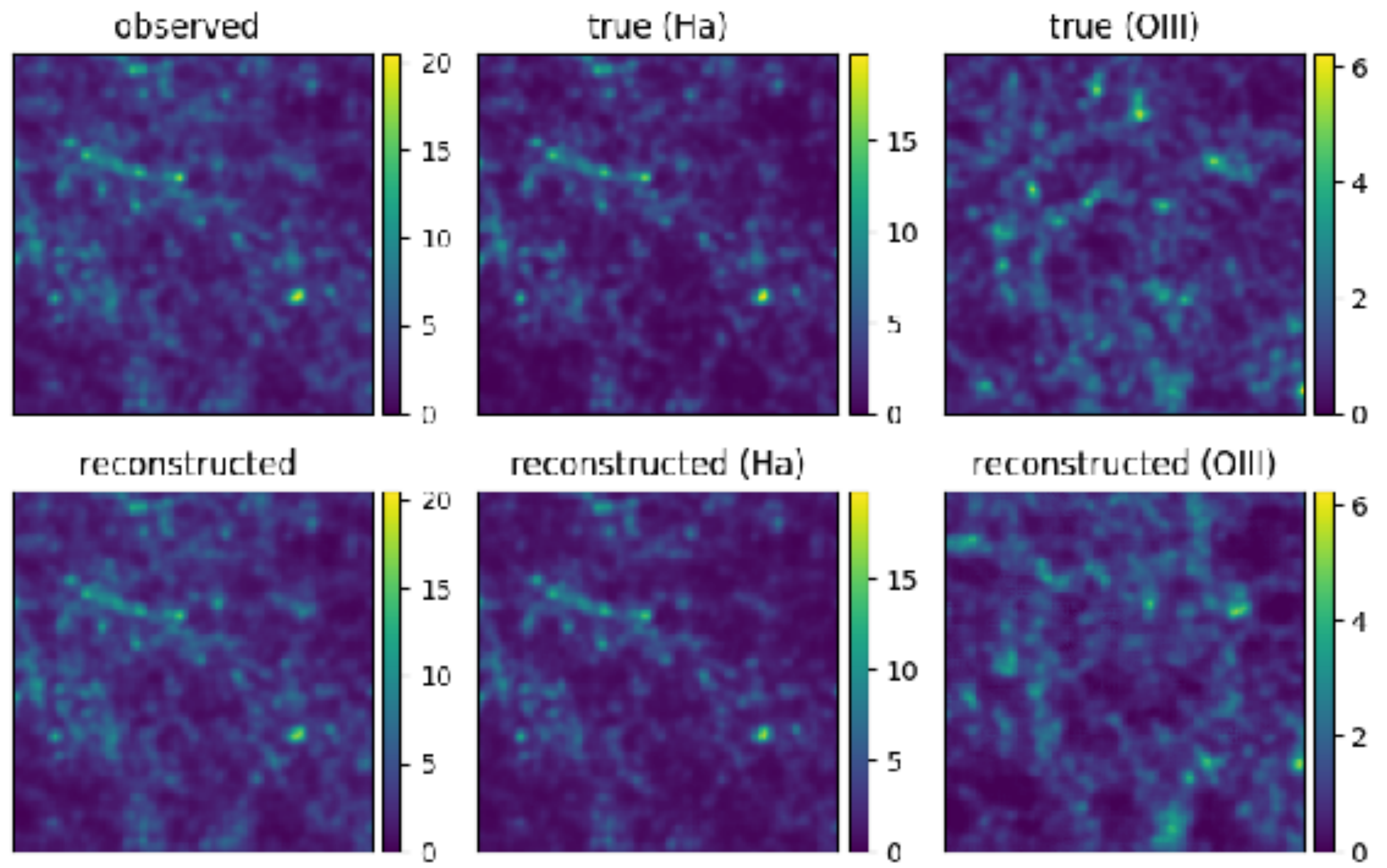


Three samples with similar power

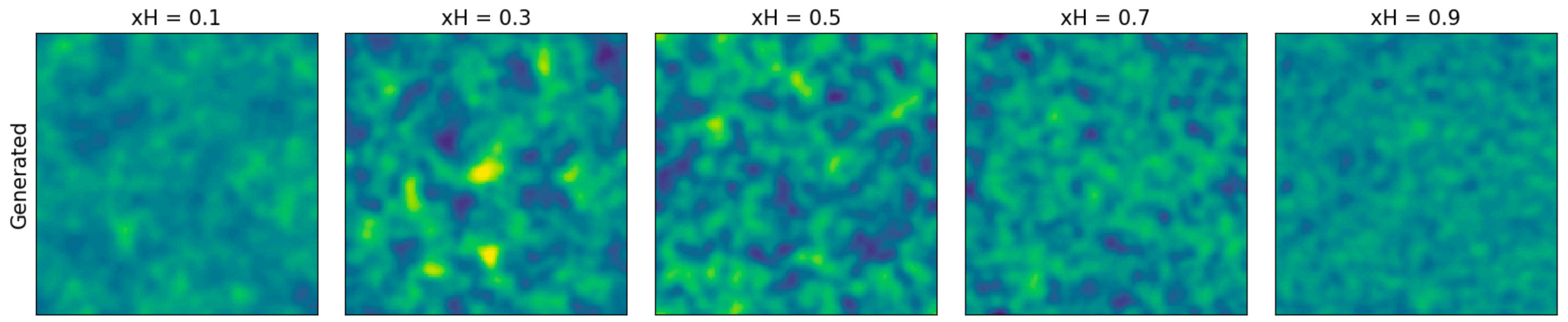


Summary

- **cGAN**
  - For line interlopers in non-21cm LIM observations
  - It could be useful when studying galaxy LIM × 21cm cross-correlation at the EoR



- **Normalizing flow**
  - For EoR parameter inference with 21cm maps
  - Compared to the other CNN-based methods, it could be more robust and flexible (need to be checked)



- **General Issues**
  - **Choice of training data (which simulation to use?)**
  - **What is pros and cons in using ML?**