Applications of Machine Learning in Astrophysics

Sultan Hassan (he/him)
Space Telescope Science Institute

Online Machine Learning Workshop
ICTP Physics Without Frontiers - Afghanistan
October 1, 2025

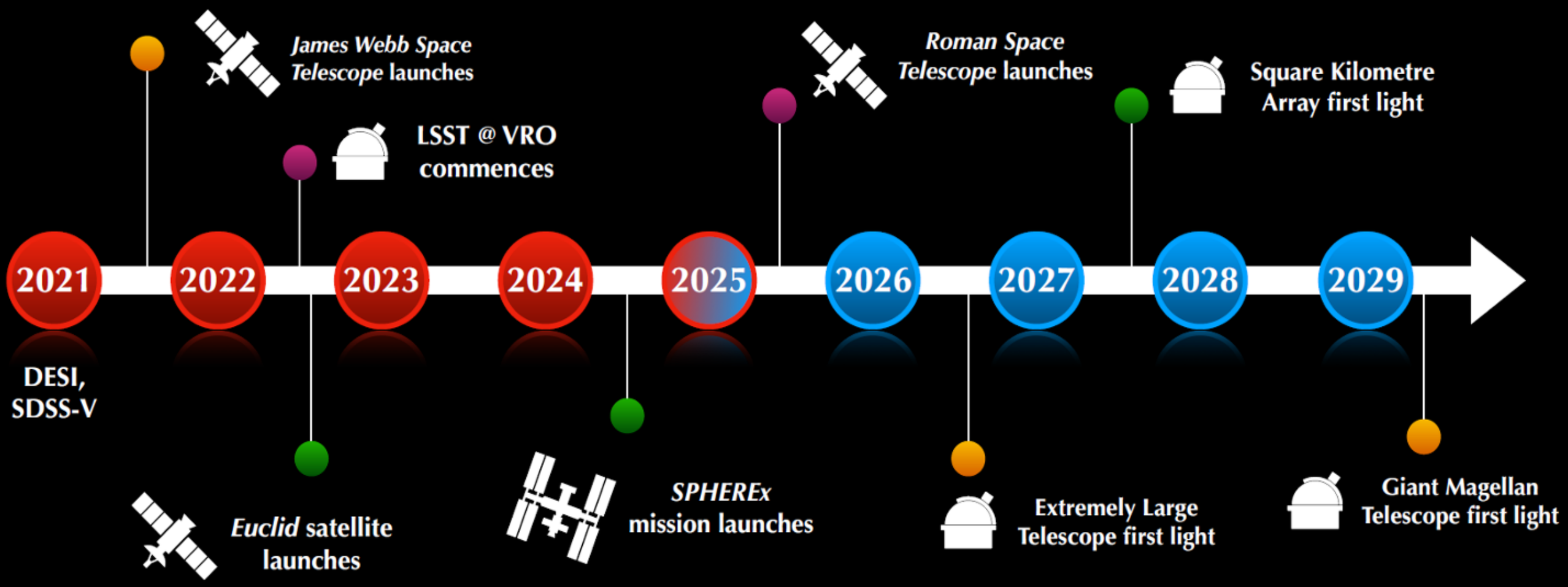
STScI | SPACE TELESCOPE SCIENCE INSTITUTE

Every dot is a galaxy!
Ho do we extract hidden information within these images?

credit:https://science.nasa.gov/asset/webb/webbs-first-deep-field-nircam-image/

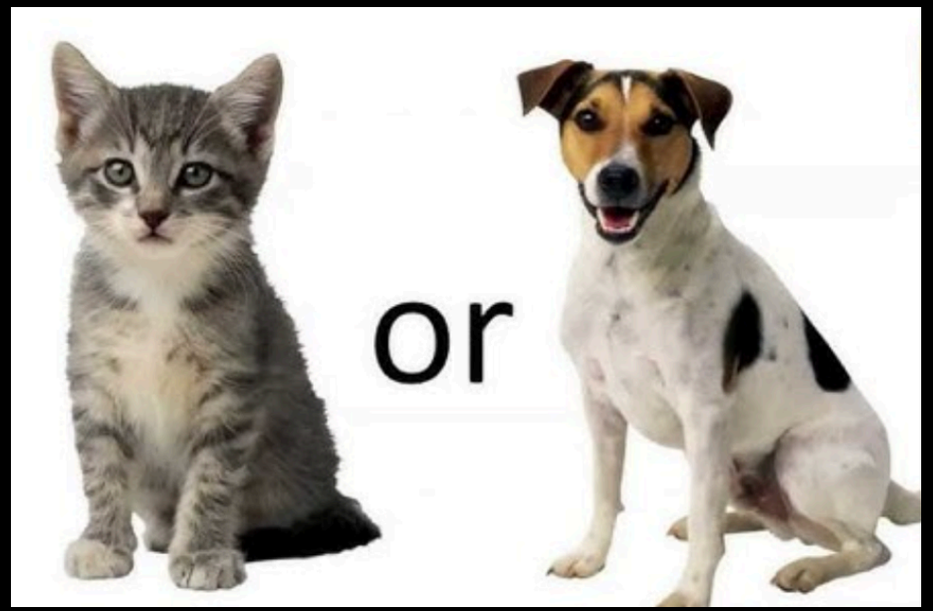
Why Machine Learning?

New generation of surveys
We are entering big data era



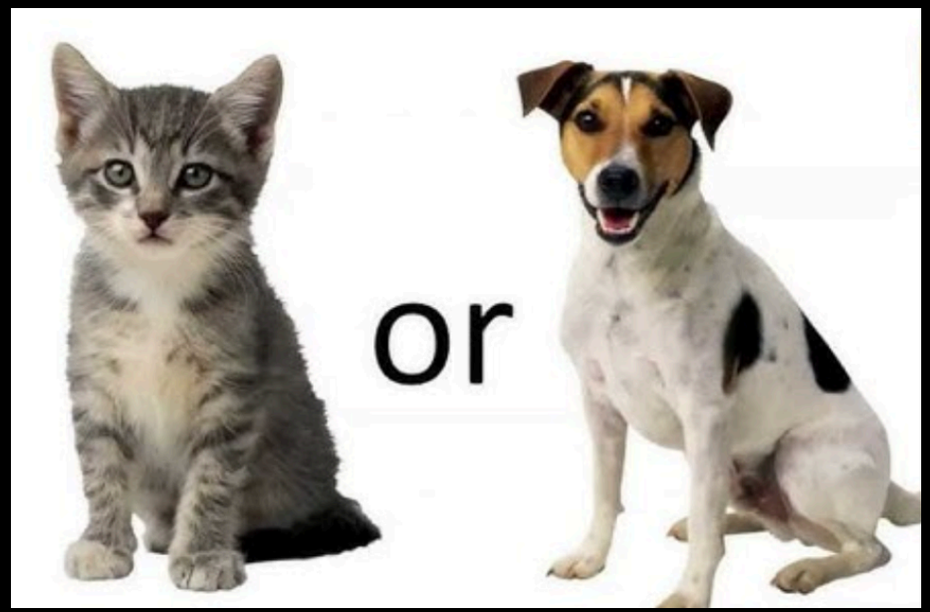
What is Machine Learning?

?



What is Machine Learning?

?



We show lots of data to learn from!

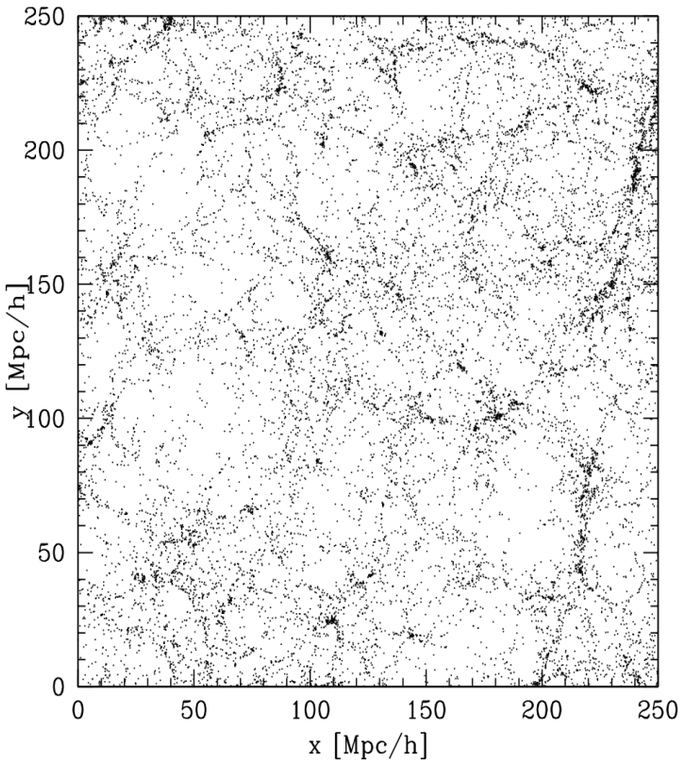
Same way for Astrophysics

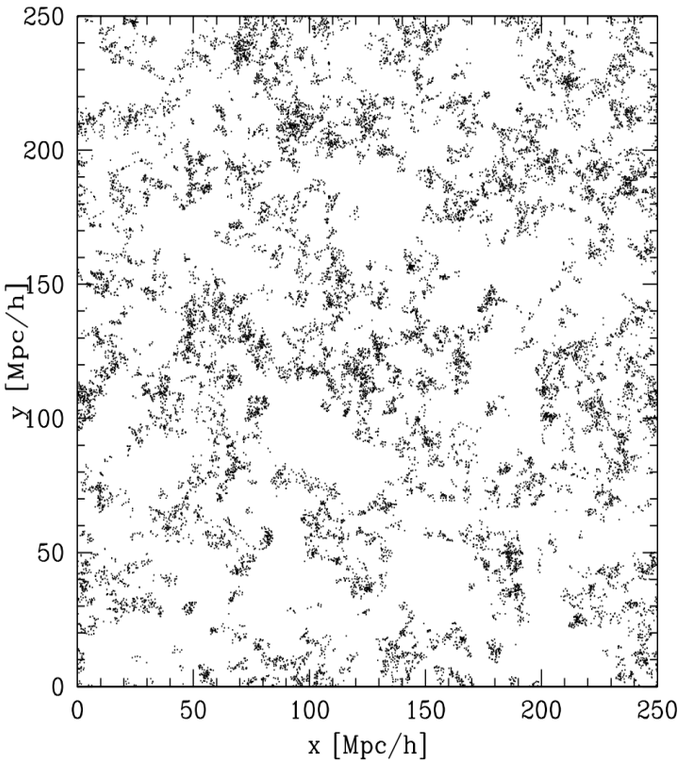
We show lots of galaxies or maps

Traditional methods: Data compression (summary statistics)

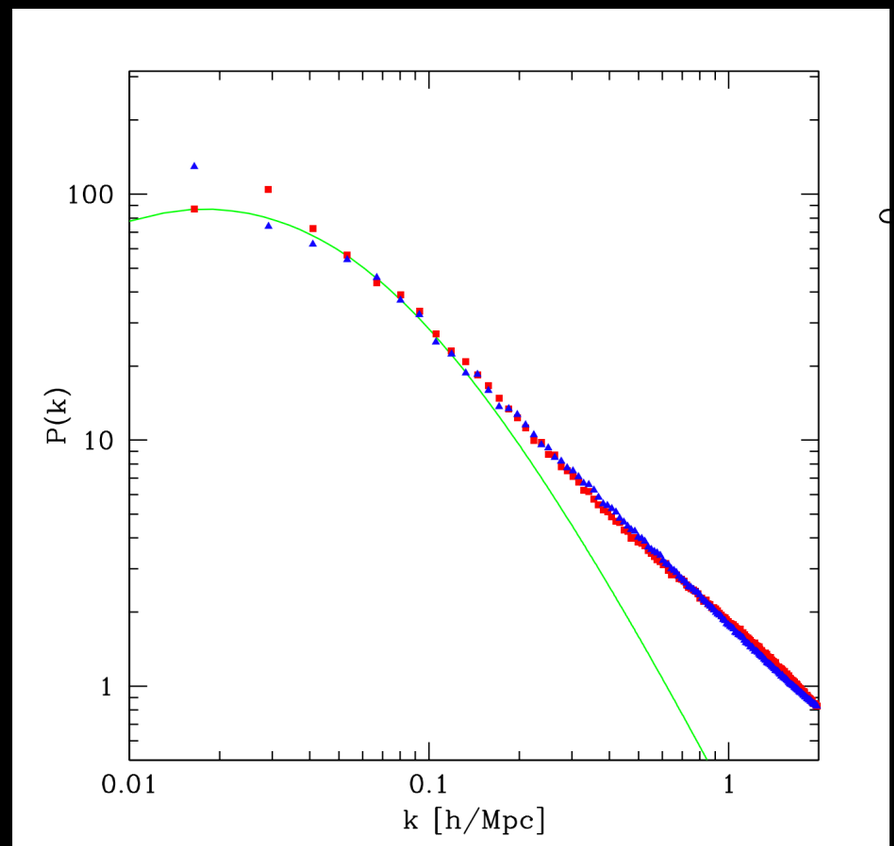
Is this a good way to describe these fields?

Different fields





Same power spectrum

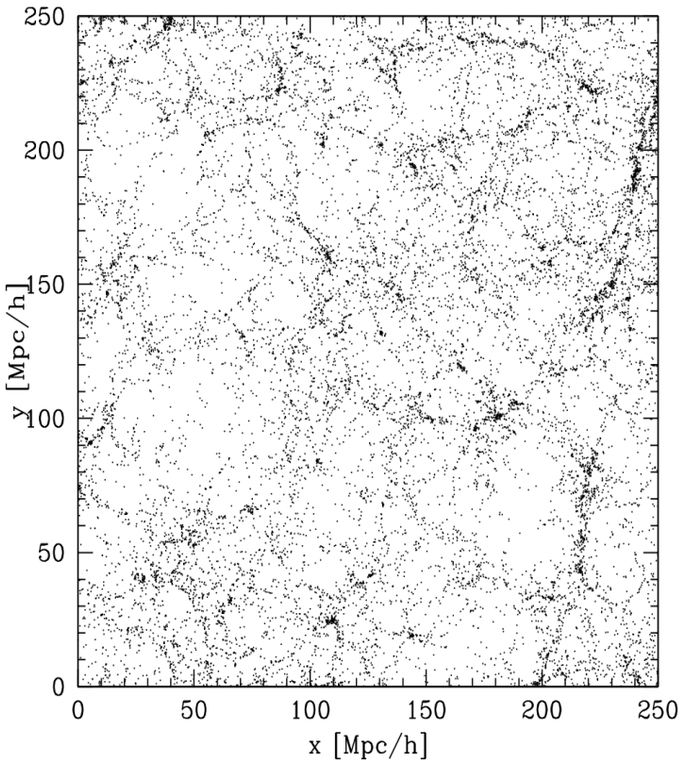


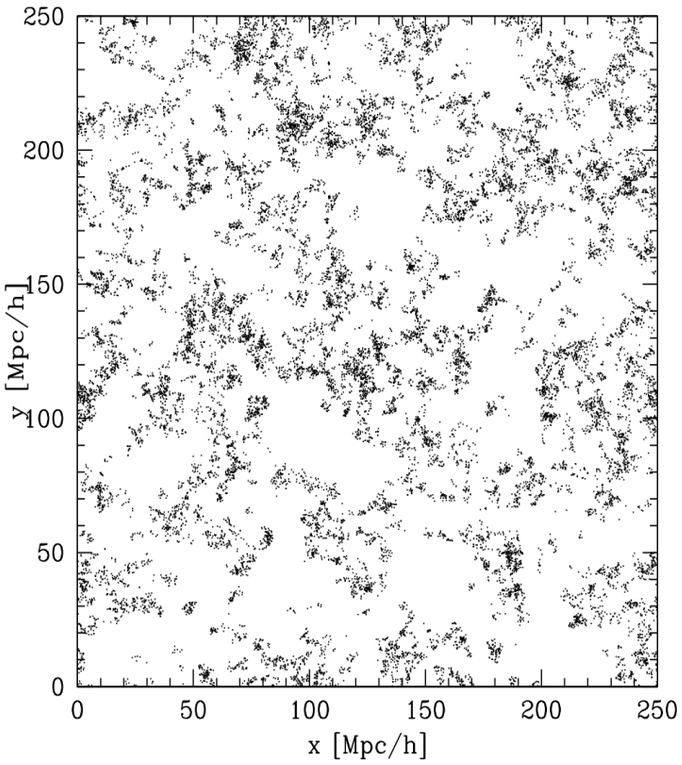
Sefusatti & Scoccimarro (2004)

Traditional methods: Data compression (summary statistics)

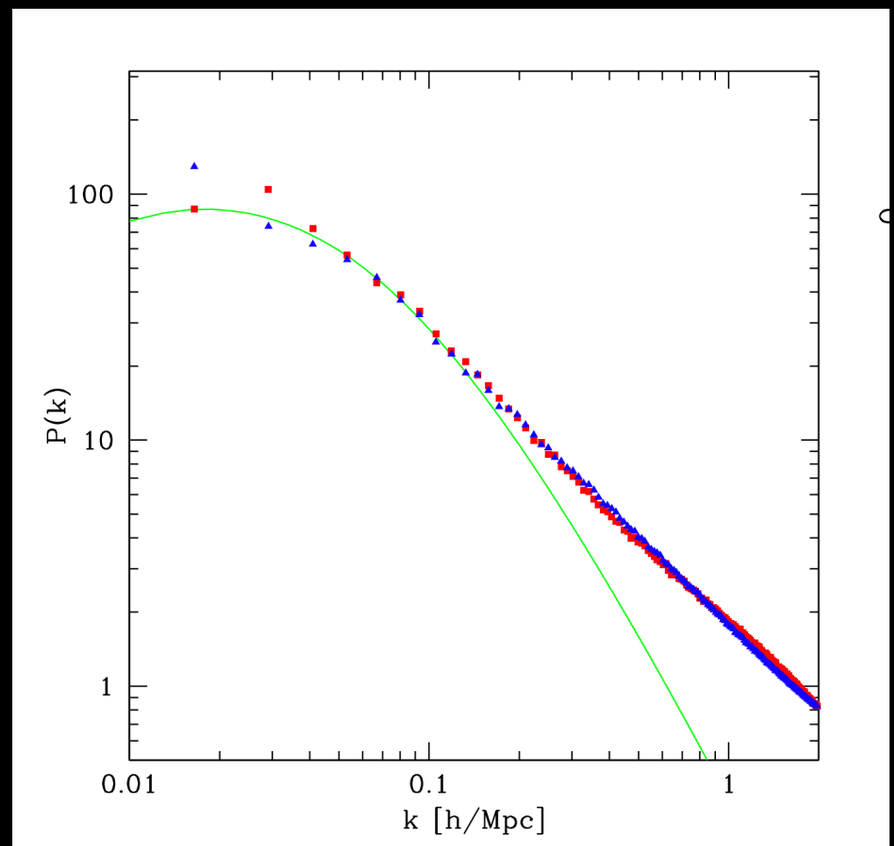
Is this a good way to describe these fields?

Different fields





Same power spectrum



We lose lots of information during data compression

Solution:
work directly with these large maps using machine learning

Sefusatti & Scoccimarro (2004)

Two main branches in Machine Learning..

$$p(\text{model} \mid \text{data}) \sim p(\text{data} \mid \text{model})\, p(\text{model})$$

Two main branches in Machine Learning..

Posterior Likelihood

$$p(\text{model} \mid \text{data}) \sim p(\text{data} \mid \text{model})\, p(\text{model})$$

Two main branches in Machine Learning..

Discriminative

Posterior

Likelihood

$$p(\text{model} \mid \text{data}) \sim p(\text{data} \mid \text{model})\, p(\text{model})$$

Two main branches in Machine Learning..

Discriminative

Posterior

Generative

Likelihood

$$p(\text{model} \mid \text{data}) \sim p(\text{data} \mid \text{model})\, p(\text{model})$$

Two main branches in Machine Learning..

Discriminative

Posterior

Generative

Likelihood

$$p(\text{model} \mid \text{data}) \sim p(\text{data} \mid \text{model})\, p(\text{model})$$

The current direction in ML is more towards generative models…

Two main branches in Machine Learning..

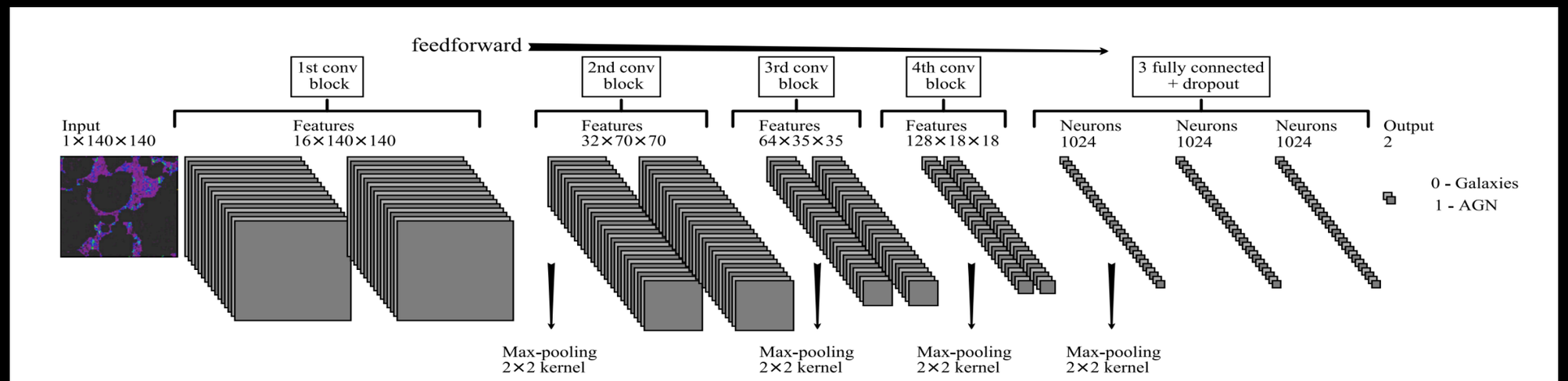
Discriminative

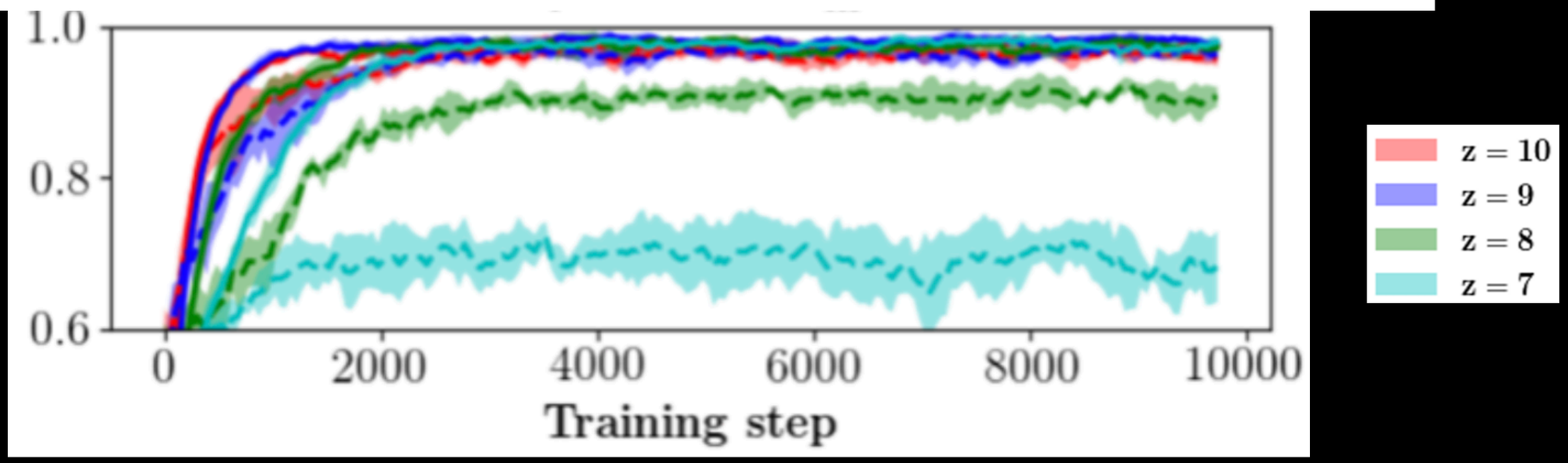
Posterior

$$p(\text{model} \mid \text{data}) \sim p(\text{data} \mid \text{model})\, p(\text{model})$$

Convolutional Neural Networks (CNNs) achieve the state-of-the-art performance in extracting information (e.g. recovering parameters) from fields.

$$p(\text{AGN or Galaxies} \mid \text{image})$$



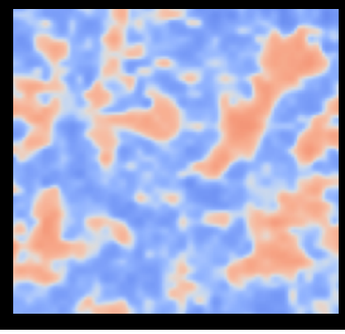


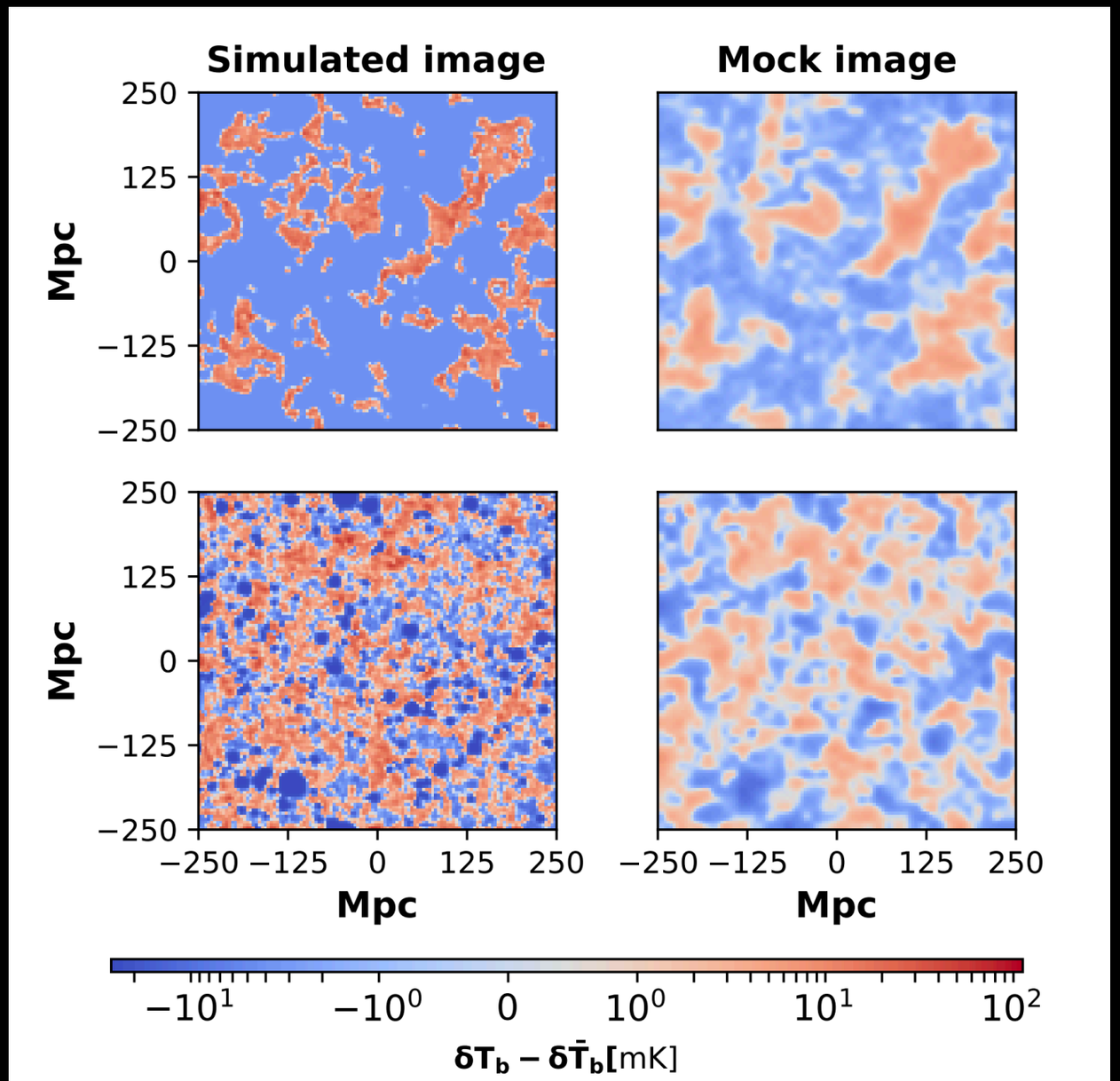
Hassan + 2019

Constraining the Reionization History with CNNs

Tumelo Mangena, SKA office
South Africa

$$p(x_{\mathrm{HI}} \mid \quad)$$



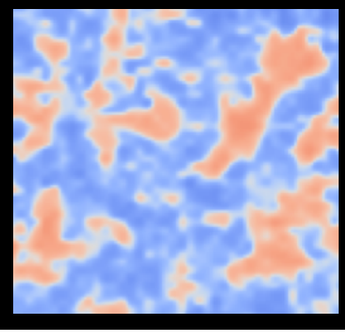


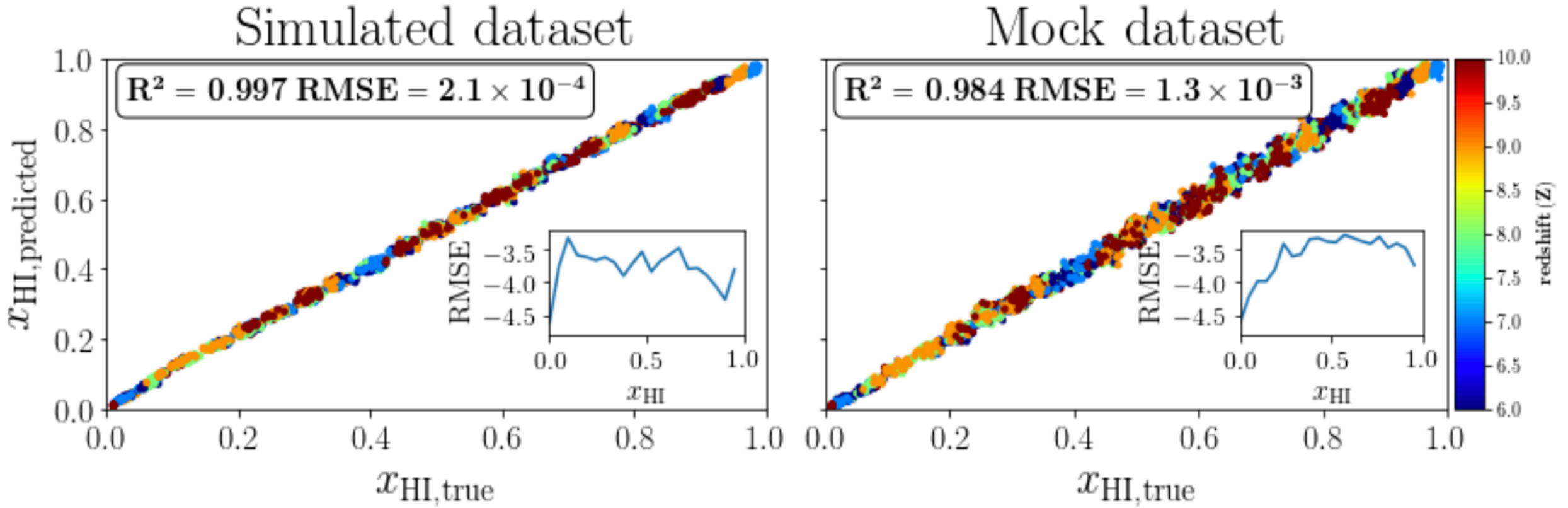
Mangena, Hassan et al. 2020

Constraining the Reionization History with CNNs

Tumelo Mangena, SKA office
South Africa

$$p(x_{\mathrm{HI}} \mid \quad)$$





Mangena, Hassan et al. 2020

Two main branches in Machine Learning..

Generative

Posterior

Likelihood

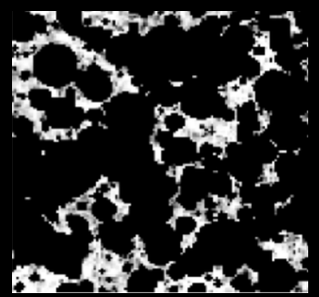
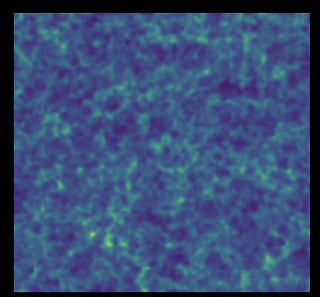
$$p(\text{model} \mid \text{data}) \sim p(\text{data} \mid \text{model})\, p(\text{model})$$

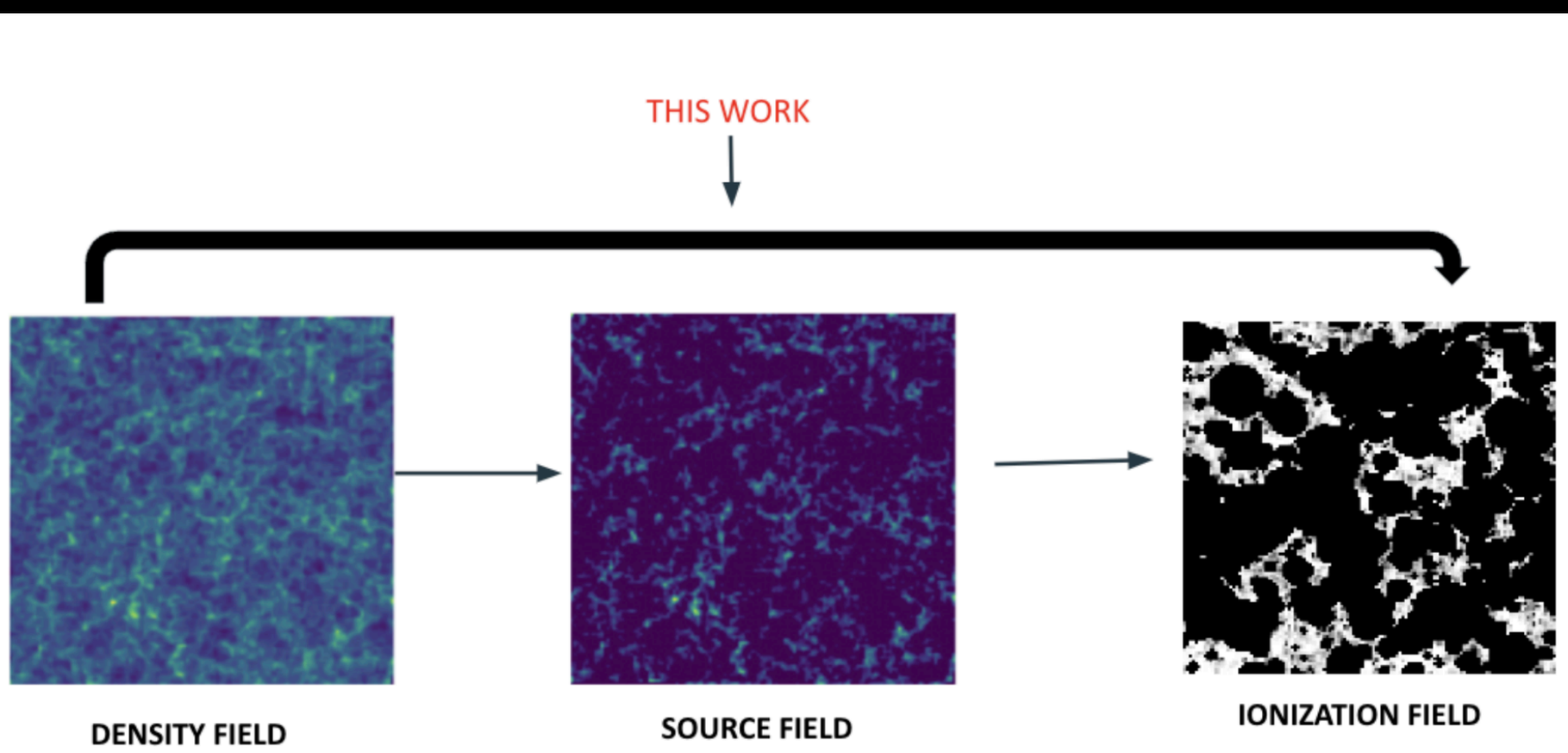
The current direction in ML is more towards generative models…

Emulating radiation transport on large scales using denoting U-Nets

(Masipa, Hassan, Santos, Contardo, Cho 2023, accepted to ICLR 2023, arXiv:2303.12065)

Mosima Masipa,
PhD student
University of the Western Cape, South Africa

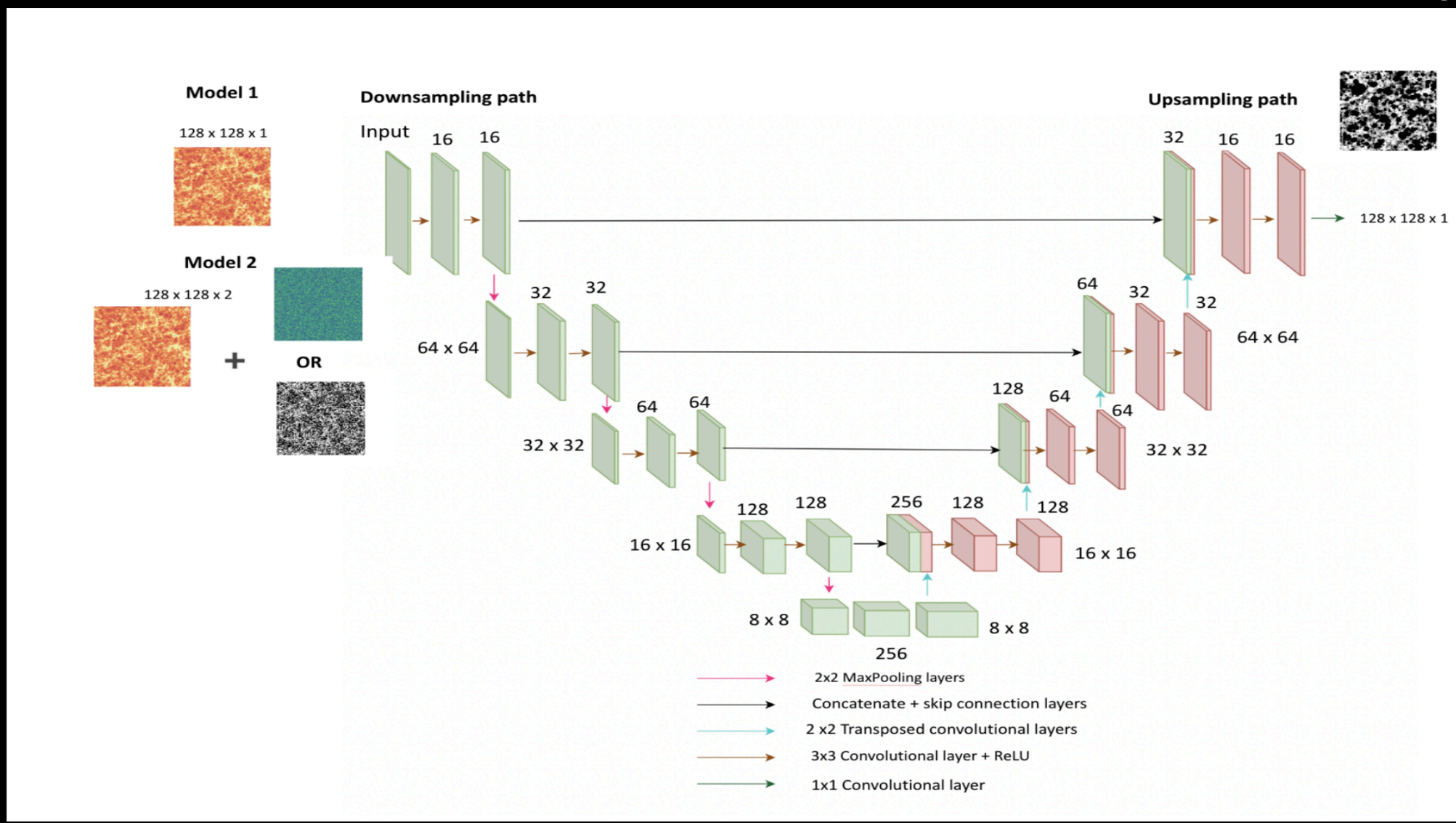
$p($  $|$  $)$



Emulating radiation transport on large scales using denoting U-Nets

(Masipa, Hassan, Santos, Contardo, Cho 2023, accepted to ICLR 2023, arXiv:2303.12065)

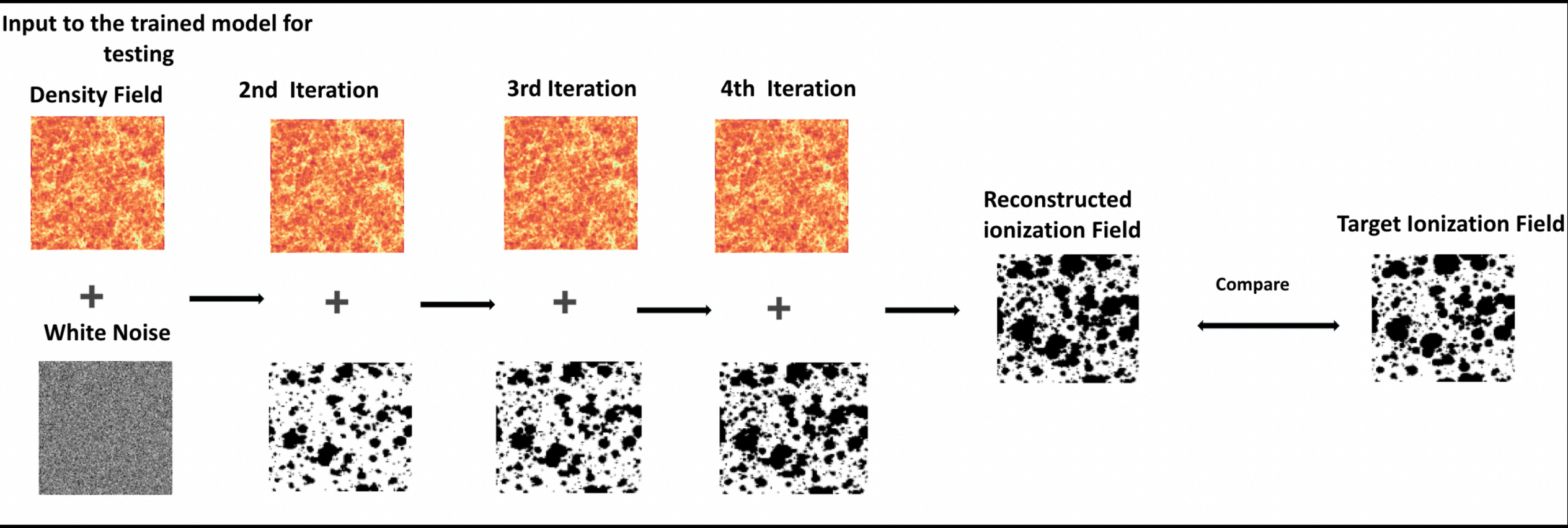
Mosima Masipa,
MSc student
University of the Western Cape, South Africa



Emulating radiation transport on large scales using denoting U-Nets

(Masipa, Hassan, Santos, Contardo, Cho 2023, accepted to ICLR 2023, arXiv:2303.12065)

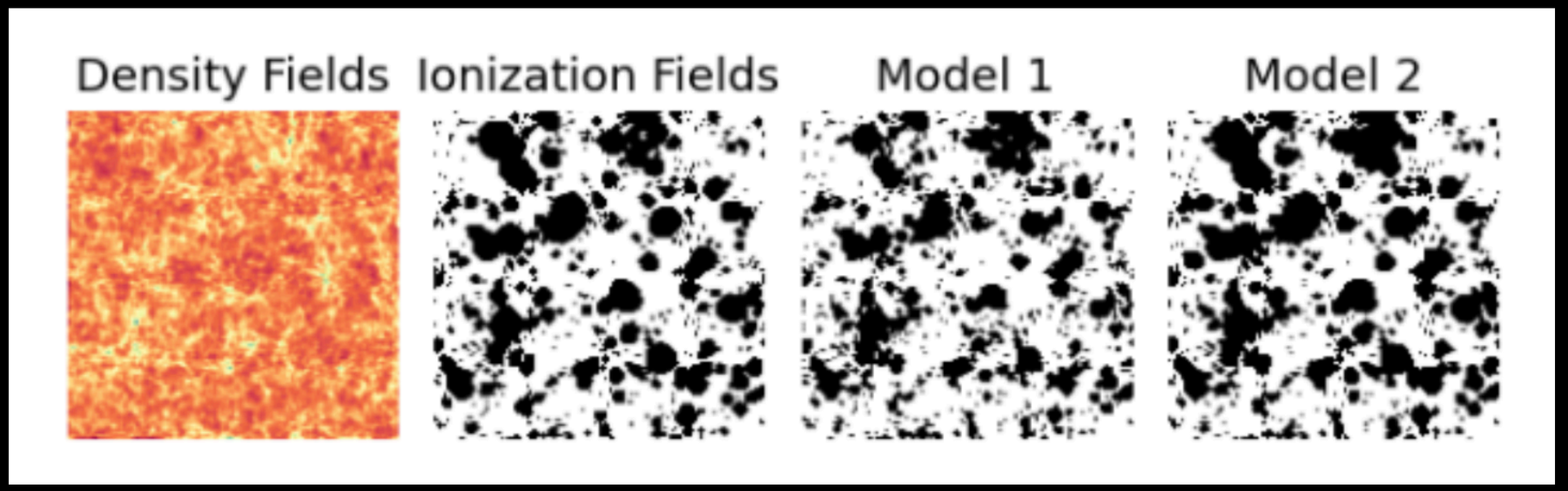
Mosima Masipa,
MSc student
University of the Western Cape, South Africa

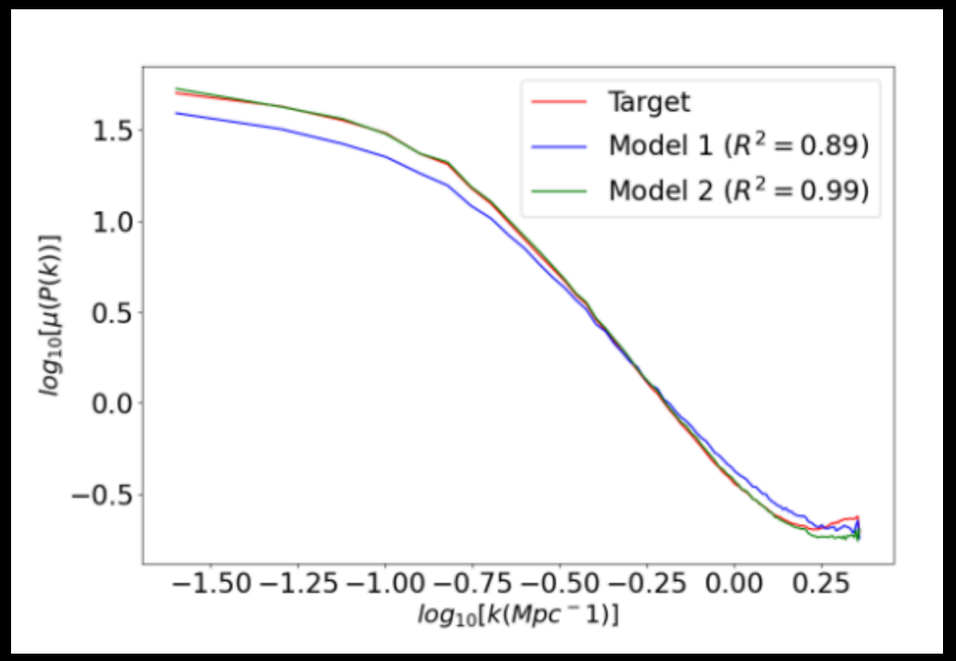


Emulating radiation transport on large scales using denoting U-Nets

(Masipa, Hassan, Santos, Contardo, Cho 2023, accepted to ICLR 2023, arXiv:2303.12065)

Mosima Masipa,
MSc student
University of the Western Cape, South Africa

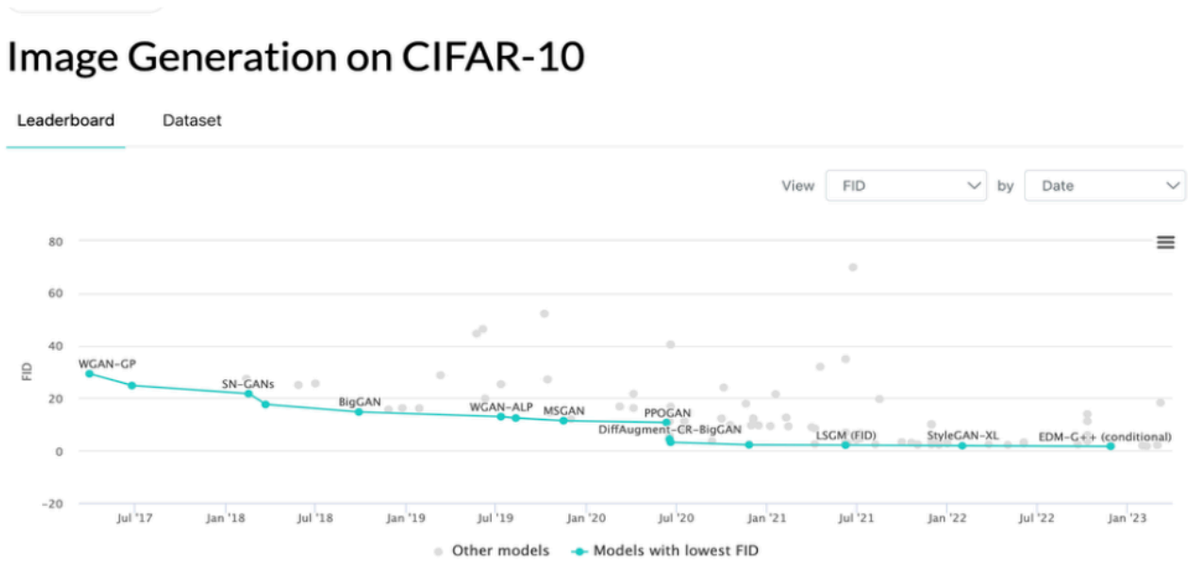




Why diffusion models?

On CIFAR-10 alone:

- Best ever model in image generation.
- 3 out the first 5.
- 5 out of the first 10.



Filter: GAN | Score-based | VAE | Flow-based | Diffusion | Transformer | PFGM | unconditional | ResNet | Optimal Transport | EBM | untagged

Edit Leaderboard

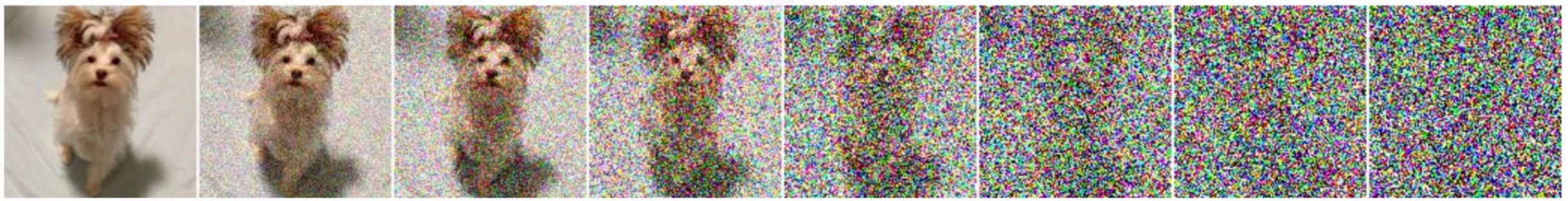
Rank	Model	FID	NFE	Inception score	bits/dimension	Paper	Code	Result	Year	Tags
1	EDM-G++ (conditional)	1.64	35			Refining Generative Process with Discriminator Guidance in Score-based Diffusion Models			2022	Diffusion, Score-based
2	PFGM++	1.74				PFGM++: Unlocking the Potential of Physics-Inspired Generative Models			2023	PFGM
3	EDM-G++ (unconditional)	1.77	35		2.55	Refining Generative Process with Discriminator Guidance in Score-based Diffusion Models			2022	Diffusion, Score-based
4	StyleGAN-XL	1.85				StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets			2022	GAN
5	STF (unconditional)	1.90				Stable Target Field for Reduced Variance Score Estimation in Diffusion Models			2023	Score-based

https://paperswithcode.com/sota/image-generation-on-cifar-10

Diffusion models

Forward SDE (data → noise)

$\mathbf{x}(0)$ —— $\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$ ——→ $\mathbf{x}(T)$



score function

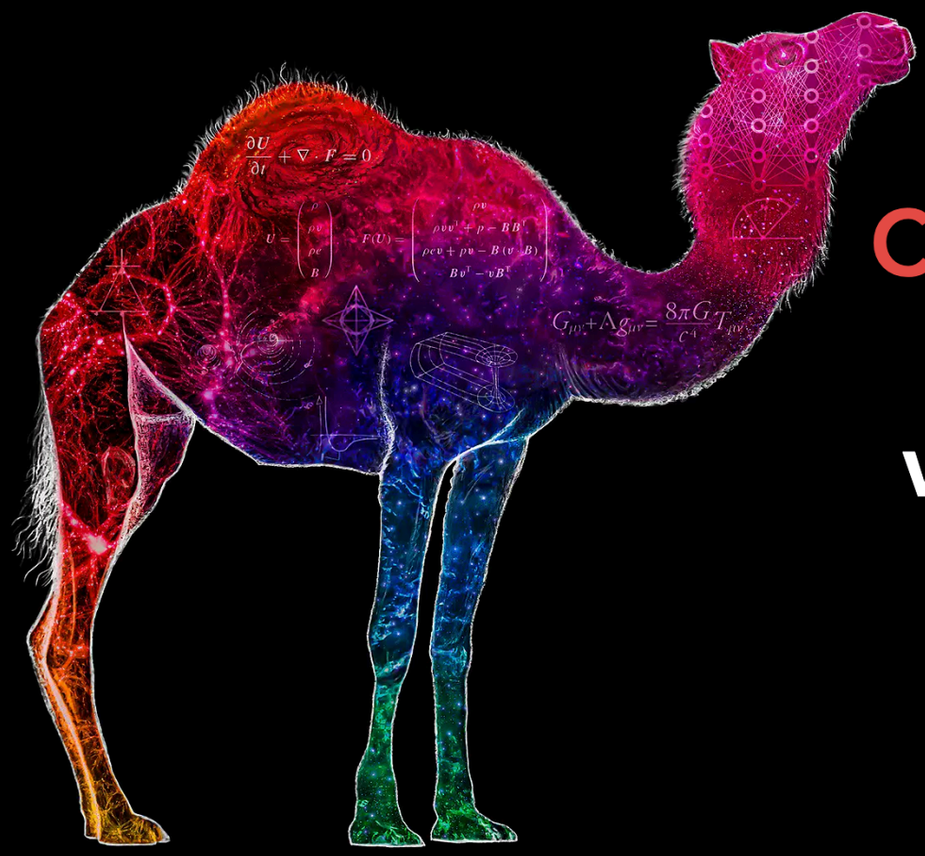
$\mathbf{x}(0)$ ←— $\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$ —— $\mathbf{x}(T)$

Reverse SDE (noise → data)

https://github.com/yang-song/score_sde_pytorch

Introducing the CAMELS dataset

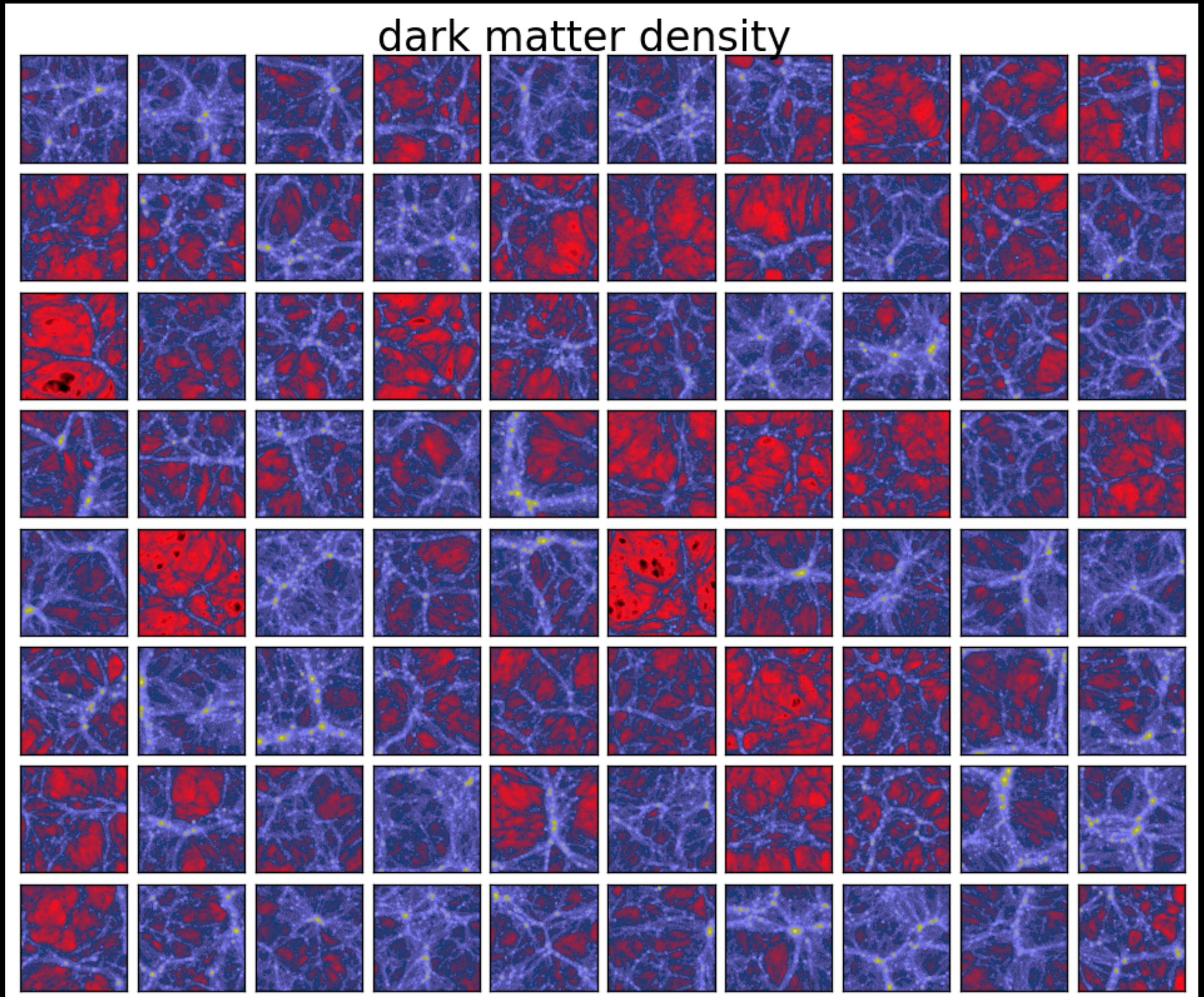
CAMELS



Cosmology and
Astrophysics
with Machin**E**
Learning
Simulations

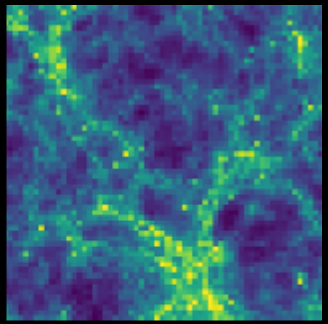
> 4,000 simulations run with *Simba* and *IllustrisTNG*
designed to train machine learning algorithms

https://www.camel-simulations.org/, Villaescusa-Navarro et al, incl. Hassan 2021



https://www.camel-simulations.org/, Villaescusa-Navarro et al, incl. Hassan 2021

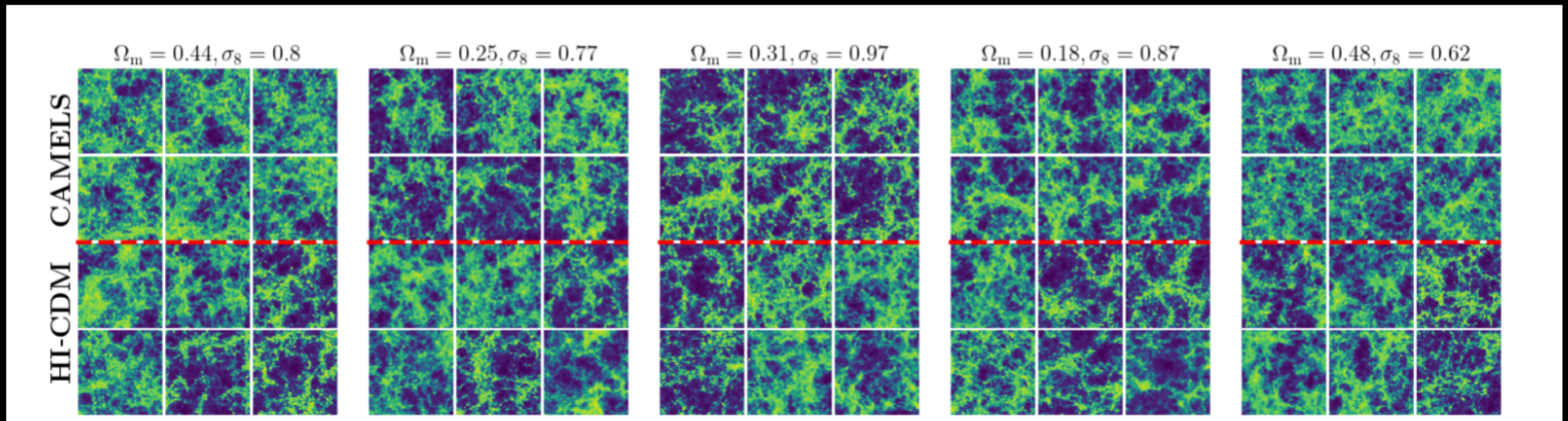
HI-CDM: Conditional diffusion model for high fidelity HI maps

$p($  $|\Omega_{\mathrm{m}}, \sigma_8)$

Hassan et al, in perp

HI-CDM: Conditional diffusion model for high fidelity HI maps

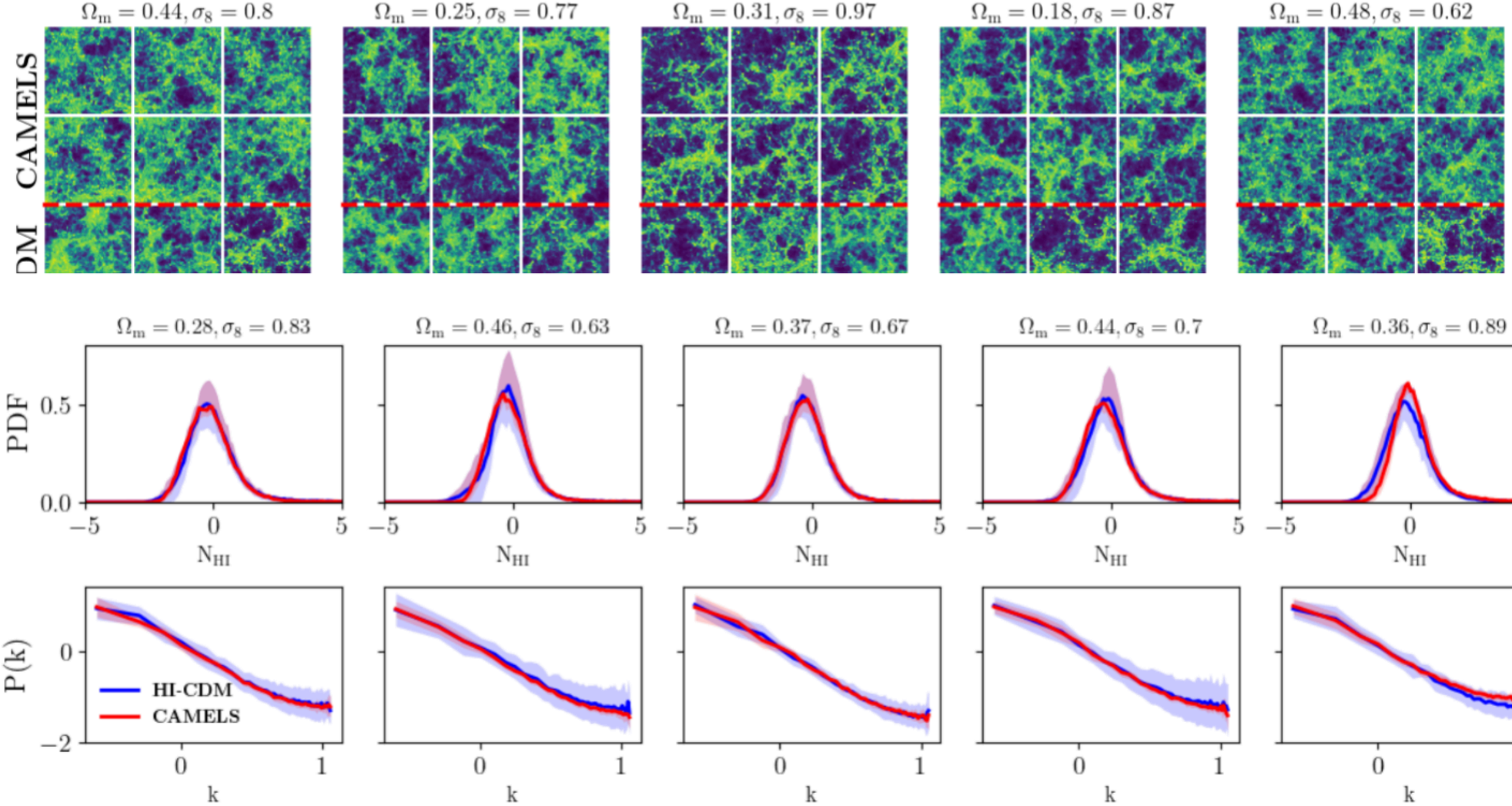
$$p(\quad | \Omega_{\rm m}, \sigma_8)$$



Hassan et al, in perp

HI-CDM: Conditional diffusion model for high fidelity HI maps

$$p(\quad | \Omega_{\rm m}, \sigma_8)$$

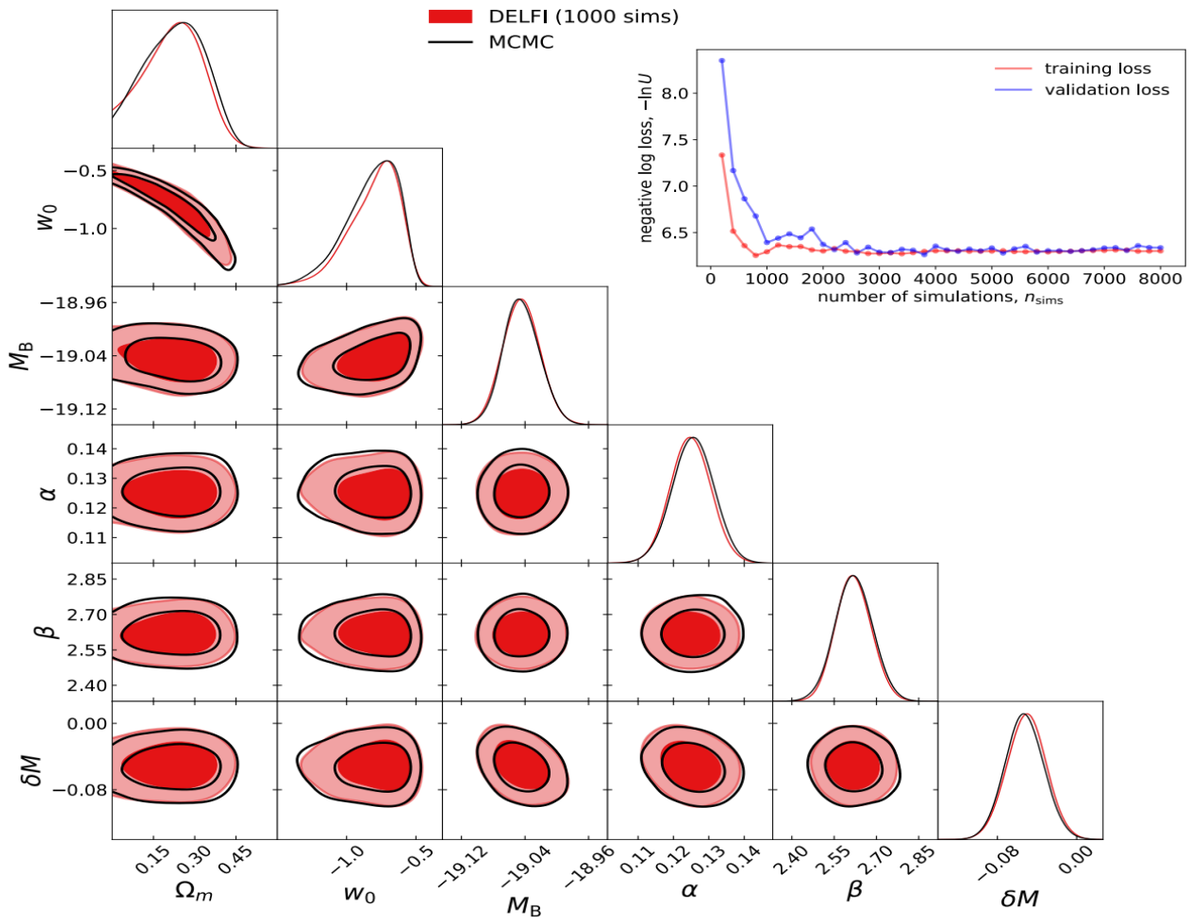


Hassan et al, in perp

Normalizing flows

Fake faces generated by NF model trained on CelebA

Likelihood free inference! (e.g. pydelfi)

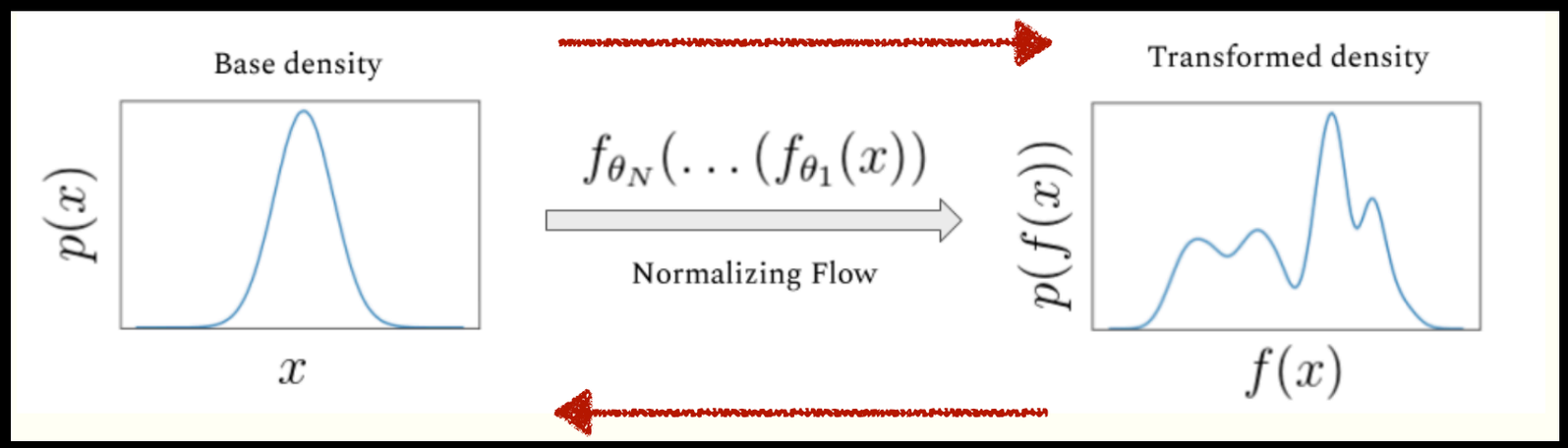


Normalizing flows

Converting *simple* distributions
to a more *complex* distributions
ALL with the change of variable formula!

$$p(x) = p(f^{-1}(x)) \left| \frac{\partial f^{-1}}{\partial x} \right|$$

Generate new examples



Papamakarios et al 2018
Dinh et al 2017

Inference

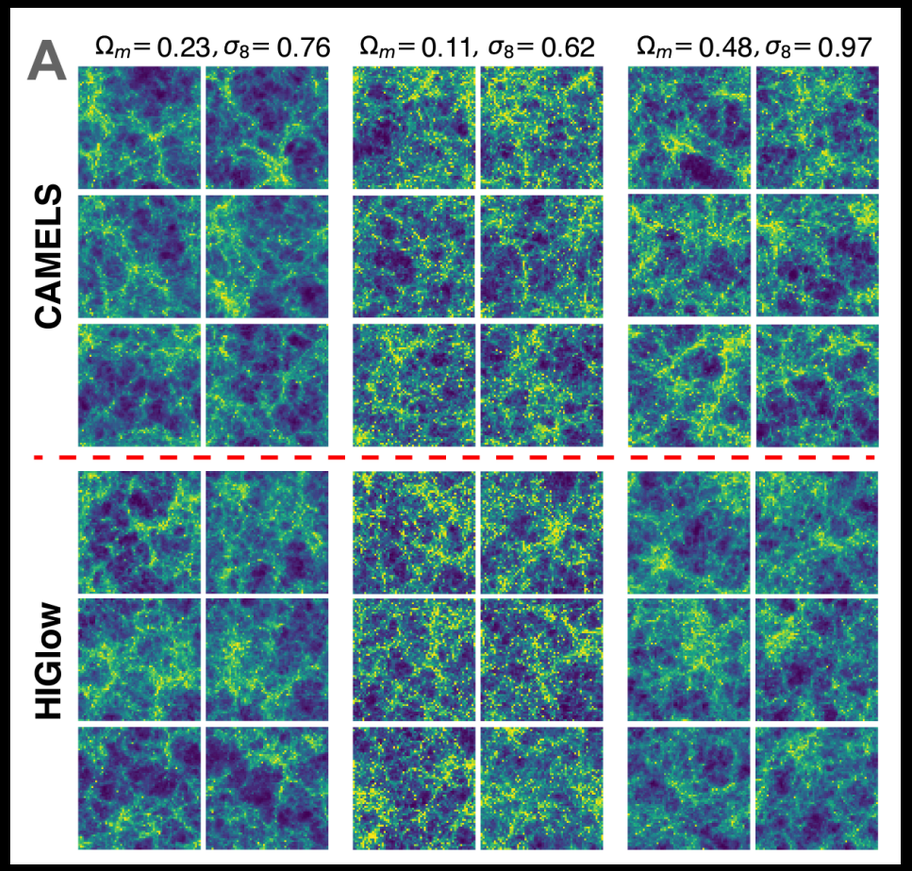
Image credit: https://siboehm.com/articles/19/normalizing-flow-network

HIGlow: Conditional Normalizing Flows for High-Fidelity HI Map Modeling

(Friedman & Hassan 2022, accepted to 2022 NeurIPS workshops, arXiv:2211.12724)

Roy Friedman, PhD student
Computer Science, The Hebrew University of Jerusalem

$$p(\ \ |\Omega_m, \sigma_8)$$

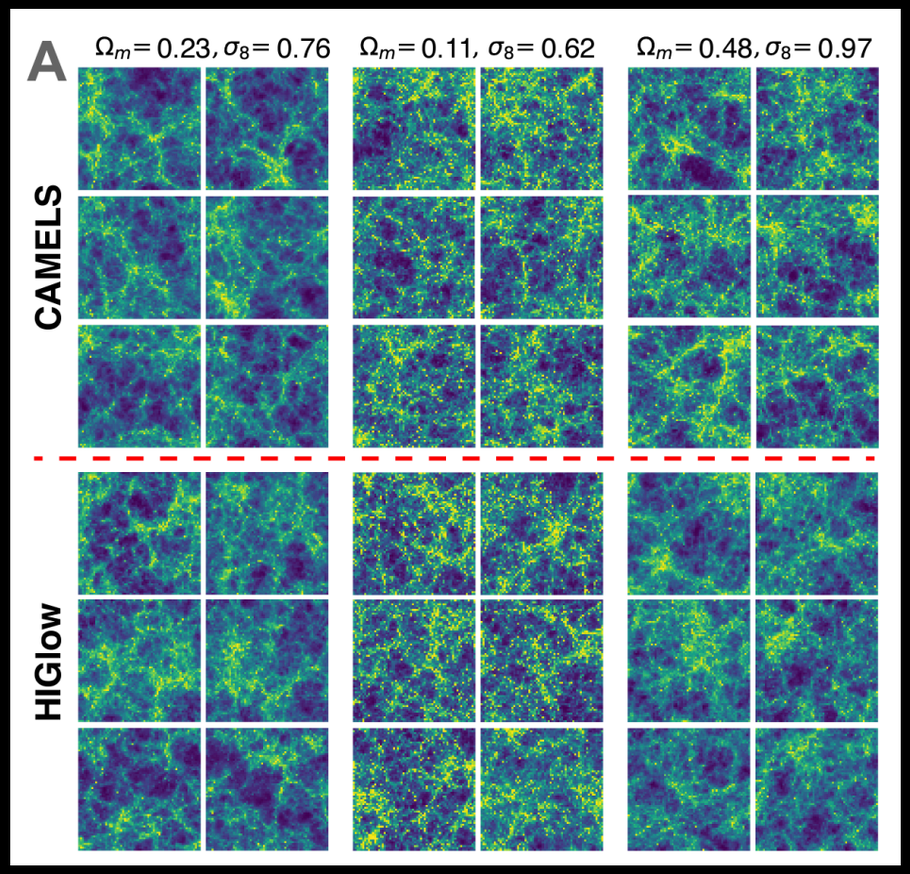


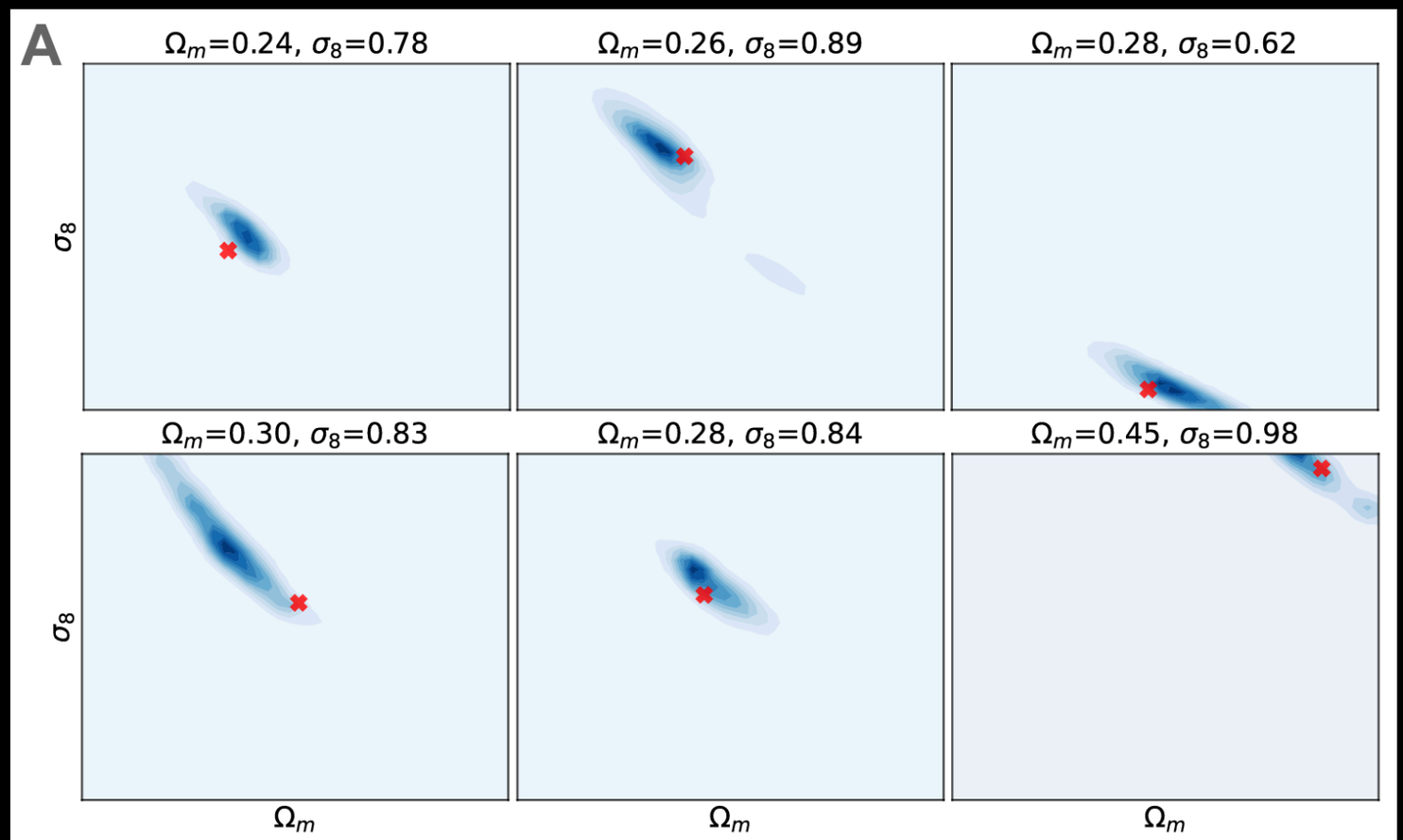
HIGlow: Conditional Normalizing Flows for High-Fidelity HI Map Modeling

(Friedman & Hassan 2022, accepted to 2022 NeurIPS workshops, arXiv:2211.12724)

Roy Friedman, PhD student
Computer Science, The Hebrew University of Jerusalem

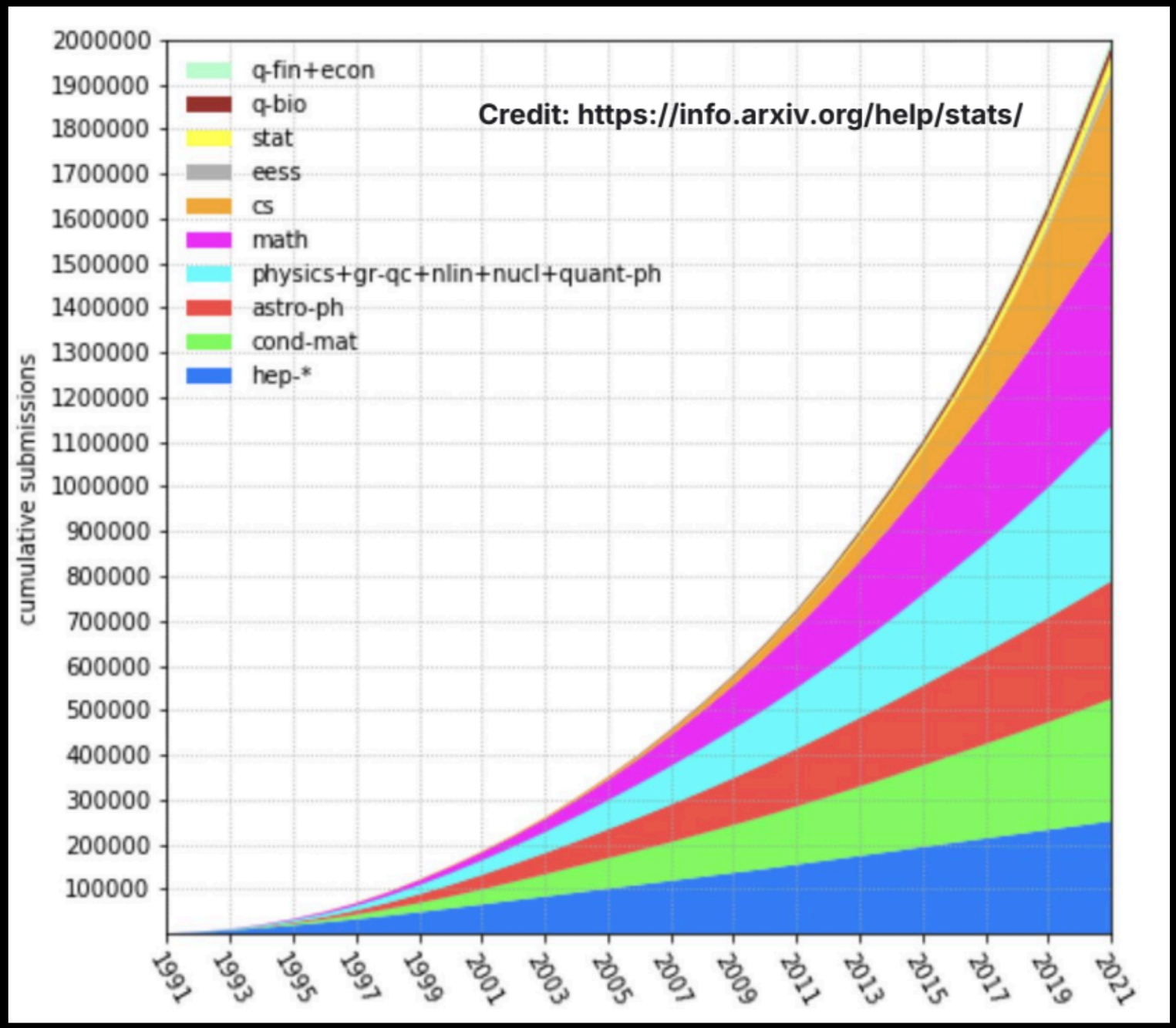
$$p(\ \ |\Omega_m, \sigma_8)$$





Application of Large Language Models in Astrophysics

Motivation: Conducting a literature search is becoming increasingly challenging due to the growing volume of new article submissions.



CosmoGemma Chatbot Pipeline

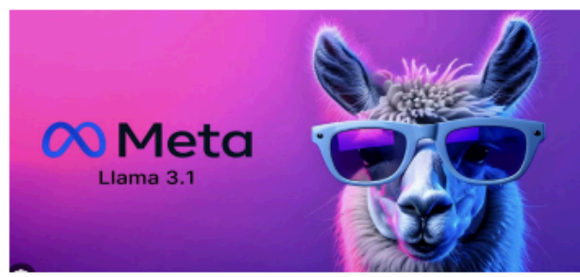
AI research assistant

1 - Dataset Generation

From Kaggle Datasets

arXiv > astro-ph.CO

LangChain X Ollama X



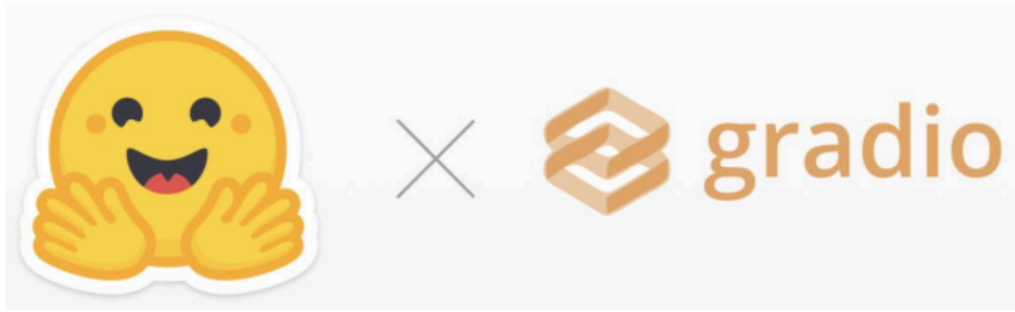
(a) **Select abstracts from 2018-2022 (~3.5k) for fine-tuning and from 2023 for testing (~1k)**

(b) **Generate Question-Answer Pairs given an abstract**

2 - Fine-tuning/Evaluating Gemma

Accuracy 95% on fine-tuning samples, 75% on testing samples

3 - Deployment



CosmoGemma is able to answer correctly multiple-choice questions and brainstorm ideas for projects.

Spaces | sultan-hassan/cosmology-expert | like 0 | Running | App | Files | Community | Settings

Hey I am CosmoGemma 👋 I can answer cosmology questions from astroph.CO research articles. Try me :)

Gemma_2b_en fine-tuned on QA pairs (3.5k) generated from Cosmology and Nongalactic Astrophysics articles (arXiv astro-ph.CO) from 2018-2022 and tested on QA pairs (1k) generated from 2023 articles, scoring over 75% accuracy.

input

You are an expert in general astrophysics. Your task is to answer the following multiple-choice question on astrophysics. The question is:

Question: What is the primary goal of calibrating subgrid feedback models in cosmological simulations?

A) To ensure that simulations accurately reproduce the observed properties of the interstellar medium.
B) To create a diverse range of galaxy morphologies in the simulations.
C) To achieve convergence in simulation results across different resolutions and box sizes.
D) To steer simulations towards producing a broadly realistic galaxy population that is consistent with key observational constraints.

Give only one answer, either A, B, C or D, but not more than one, and always give an answer.

Clear | Submit

output

D

input

Please list three ideas for interesting projects to work on to advance cosmology?

Clear | Submit

output

measuring Ω_M, Ω_Λ, and w_D from cosmic chronometers; tracking the expansion history of the universe with clusters and the intergalactic medium; making a precision measurement of the Hubble constant using weak lensing and galaxy clustering.

CosmoGemma is able to answer correctly multiple-choice questions and brainstorm ideas for projects.

Spaces | sultan-hassan/cosmology-expert | like 0 | Running | App | Files | Community | Settings

Hey I am CosmoGemma 👋 I can answer cosmology questions from astroph.CO research articles. Try me :)

Gemma_2b_en fine-tuned on QA pairs (3.5k) generated from Cosmology and Nongalactic Astrophysics articles (arXiv astro-ph.CO) from 2018-2022 and tested on QA pairs (1k) generated from 2023 articles, scoring over 75% accuracy.

input

You are an expert in general astrophysics. Your task is to answer the following multiple-choice question on astrophysics. The question is:

Question: What is the primary goal of calibrating subgrid feedback models in cosmological simulations?

A) To ensure that simulations accurately reproduce the observed properties of the interstellar medium.
B) To create a diverse range of galaxy morphologies in the simulations.
C) To achieve convergence in simulation results across different resolutions and box sizes.
D) To steer simulations towards producing a broadly realistic galaxy population that is consistent with key observational constraints.

Give only one answer, either A, B, C or D, but not more than one, and always give an answer.

Clear | Submit

output

D

input

Please list three ideas for interesting projects to work on to advance cosmology?

Clear | Submit

output

measuring Ω_M, Ω_Λ, and w_D from cosmic chronometers; tracking the expansion history of the universe with clusters and the intergalactic medium; making a precision measurement of the Hubble constant using weak lensing and galaxy clustering.

Apps links:

CosmoGemma

Github: https://github.com/sultan-hassan/CosmoGemma

App: https://huggingface.co/spaces/sultan-hassan/cosmology-expert

skimArxiv

App: https://huggingface.co/spaces/sultan-hassan/skimarXiv

Summary

Generative AI methods are transforming education, and research.

Learn more of AI tools…

Apply them to your research…

Adapt to the AI evolution…

But be careful, it can be addictive.

Lets connect:

email: sultanier@gmail.com

Linkedin: www.linkedin.com/in/sultan-h-hassan