SKACH

SKAO

DEEP LEARNING THE MAPPING BETWEEN SKA MOCKS AND HYDRODYNAMICAL SIMULATIONS

SKA research at the Centre for Artificial Intelligence ZHAW

12/01/2023

by

Philipp Denzel

SLIDES ON MY WEBSITE

https://phdenzel.github.io/

Link/QR code to the slides for later or to follow along

ABOUT MYSELF

- PhD in Physics from UZH @ ICS

ABOUT MYSELF

- PhD in Physics from UZH @ ICS
- **computational scientist & astrophysicist** by training, **machine learning** enthusiast
  - high-performance computing: dark-matter detection, star formation, etc.
  - gravitational lensing (e.g. see Denzel et al. 2021a or Denzel et al. 2021b)

ABOUT MYSELF

- PhD in Physics from UZH @ ICS
- **computational scientist & astrophysicist** by training, **machine learning** enthusiast
  - high-performance computing: dark-matter detection, star formation, etc.
  - gravitational lensing (e.g. see Denzel et al. 2021a or Denzel et al. 2021b)
- **machine learning engineer** @ CAI ZHAW
  - other research: AI certification for safety critical applications
  - main focus: **deep learning for SKA**
    - hydrodynamical simulations ↔ SKA mock observations

HYDRODYNAMICAL SIMULATIONS

- cosmological & astrophysical processes from first principle
- highly tuned on a vast range of scales
  - subgrid models for the processes that aren't resolved

HYDRODYNAMICAL SIMULATIONS

- cosmological & astrophysical processes from first principle
- highly tuned on a vast range of scales
    - subgrid models for the processes that aren't resolved
- large-scale structure of the Universe (dark-matter)

HYDRODYNAMICAL SIMULATIONS

- cosmological & astrophysical processes from first principle
- highly tuned on a vast range of scales
    - subgrid models for the processes that aren't resolved
- large-scale structure of the Universe (dark-matter)
- realistic galaxy models (baryons, radiation)

HYDRODYNAMICAL SIMULATIONS

- cosmological & astrophysical processes from first principle
- highly tuned on a vast range of scales
  - subgrid models for the processes that aren't resolved
- large-scale structure of the Universe (dark-matter)
- realistic galaxy models (baryons, radiation)
- latest simulations reach (almost) petabyte sizes → ideal for deep learning
  - IllustrisTNG, Simba, FIRE, EAGLE, and others

GENERATIVE DEEP LEARNING

- find parameters θ to approximate a data density (optionally conditioned on some information c)

$$P_\theta(x|c) \sim P_{\text{data}}(x|c)$$

GENERATIVE DEEP LEARNING

- find parameters θ to approximate a data density (optionally conditioned on some information c)

$$P_\theta(x \mid c) \sim P_{\text{data}}(x \mid c)$$

- in contrast to discriminative deep learning:
  - pattern recognition

GENERATIVE DEEP LEARNING

- find parameters θ to approximate a data density (optionally conditioned on some information c)

$$P_\theta(x \mid c) \sim P_{\text{data}}(x \mid c)$$

- in contrast to discriminative deep learning:
  - pattern recognition
- (inspired) creativity $\rightarrow$ much more ambitious

Latest successes

LDMs by Rombach et al. (2022), Google's Imagen, or OpenAI's DALLE-2

- new champions in semantic understanding
- generate images up to 1 Megapixel!

"A corgi's head depicted as an explosion of a nebula"

from Ramesh et al. (2022)

"A dolphin in an astronaut suit on saturn, artstation"

from Ramesh et al. (2022)

"Panda mad scientist mixing sparkling chemicals, artstation"

from Ramesh et al. (2022)

Approaches and objectives

- GANs: $\mathbb{E}_{x\sim p_{\text{data}}}[\log D_\theta(x)] + \mathbb{E}_{z\sim q(z)}[1 - \log D_\theta(G_\theta(z))]$
  - fast, high quality, implicit density, mode collapse

Approaches and objectives

- GANs: $\mathrm{E}_{x \sim p_{\mathrm{data}}}[\log D_\theta(x)] + \mathrm{E}_{z \sim q(z)}[1 - \log D_\theta(G_\theta(z))]$
  - fast, high quality, implicit density, mode collapse
- VAEs: $\log p(x) \geq \mathrm{E}_{z \sim q_\theta(z|x)}[\log p_\theta(x|z)] - D_{KL}\Big(q_\theta(z|x) \,||\, p(z)\Big)$
  - fast, regularized latent space, lower bound to LL, trade-offs: reconstruction ⇔ regularization

Approaches and objectives

- GANs: $\mathrm{E}_{x \sim p_{\mathrm{data}}}[\log D_\theta(x)] + \mathrm{E}_{z \sim q(z)}[1 - \log D_\theta(G_\theta(z))]$
  - fast, high quality, implicit density, mode collapse
- VAEs: $\log p(x) \geq \mathrm{E}_{z \sim q_\theta(z|x)}[\log p_\theta(x|z)] - D_{KL}\left(q_\theta(z|x) || p(z)\right)$
  - fast, regularized latent space, lower bound to LL, trade-offs: reconstruction ⇔ regularization
- Autoregressive models: $p(x) = \prod_i p_\theta(x_i | x_{<i})$
  - exact, good results, no latent representation, slow inference

Approaches and objectives

- GANs: $\mathrm{E}_{x \sim p_{\mathrm{data}}}[\log D_\theta(x)] + \mathrm{E}_{z \sim q(z)}[1 - \log D_\theta(G_\theta(z))]$
  - fast, high quality, implicit density, mode collapse
- VAEs: $\log p(x) \geq \mathrm{E}_{z \sim q_\theta(z|x)}[\log p_\theta(x|z)] - D_{KL}\left(q_\theta(z|x) || p(z)\right)$
  - fast, regularized latent space, lower bound to LL, trade-offs: reconstruction ⇔ regularization
- Autoregressive models: $p(x) = \prod_i p_\theta(x_i | x_{<i})$
  - exact, good results, no latent representation, slow inference
- Diffusion Models: $-\log p(x) \leq \mathrm{E}_q[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}]$
  - flexible, high fidelity, lower bound to LL

Approaches and objectives

- GANs: $\mathrm{E}_{x \sim p_{\mathrm{data}}}[\log D_\theta(x)] + \mathrm{E}_{z \sim q(z)}[1 - \log D_\theta(G_\theta(z))]$
  - fast, high quality, implicit density, mode collapse
- VAEs: $\log p(x) \geq \mathrm{E}_{z \sim q_\theta(z|x)}[\log p_\theta(x|z)] - D_{KL}\left(q_\theta(z|x) || p(z)\right)$
  - fast, regularized latent space, lower bound to LL, trade-offs: reconstruction ⇔ regularization
- Autoregressive models: $p(x) = \prod_i p_\theta(x_i | x_{<i})$
  - exact, good results, no latent representation, slow inference
- Diffusion Models: $-\log p(x) \leq \mathrm{E}_q[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}]$
  - flexible, high fidelity, lower bound to LL
- Normalizing flows: $p_\theta(x) = p(f_\theta(x)) \cdot J_{f_\theta^{-1}}(x)$
  - invertible, latent variable, exact likelihood, expensive in high-dimensional spaces

DATA

- publicly available:
  - runs: TNG50-1 / TNG100-1 / TNG300-1
  - box sizes: 50 Mpc / 100 Mpc / 300 Mpc

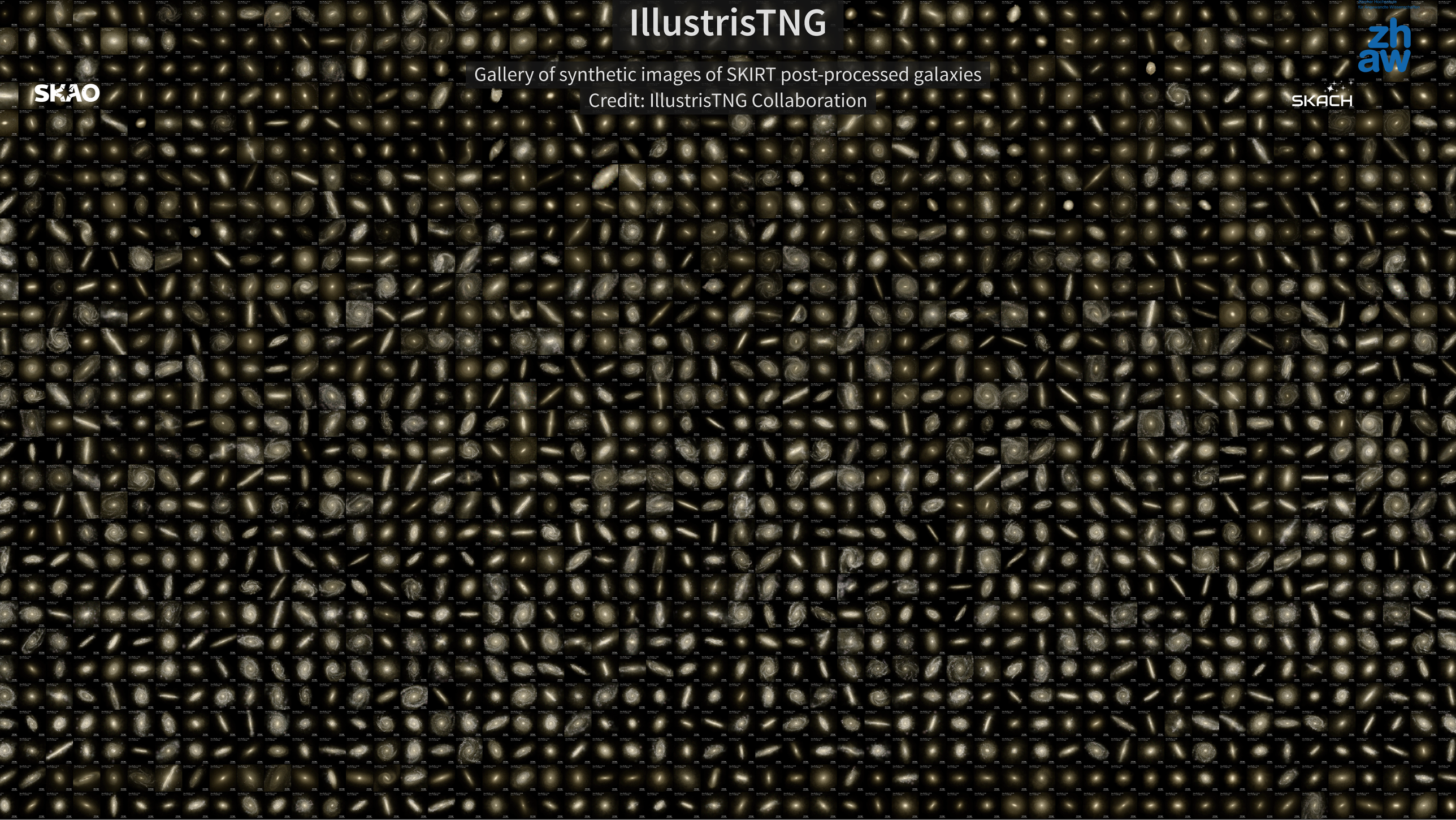
DATA

- publicly available:
  - runs: TNG50-1 / TNG100-1 / TNG300-1
  - box sizes: 50 Mpc / 100 Mpc / 300 Mpc
- TNG50-1:
  - $m_{DM} \sim 4 \times 10^5 \, M_\odot$ / $m_b \sim 8 \times 10^4 \, M_\odot$
  - 10^{10} cells/particles and around 10M "galaxies"

IllustrisTNG

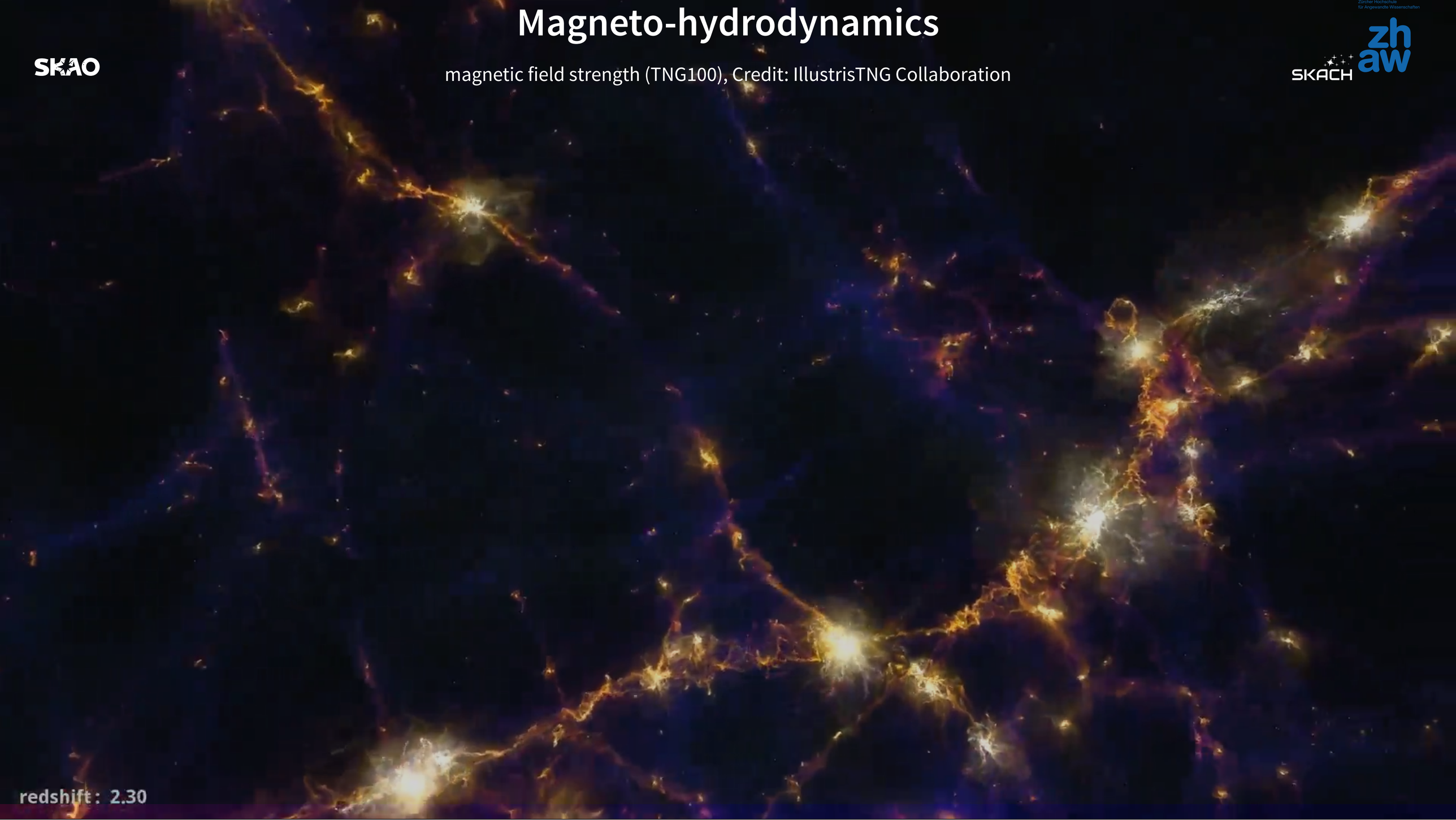
Gallery of synthetic images of SKIRT post-processed galaxies

Credit: IllustrisTNG Collaboration

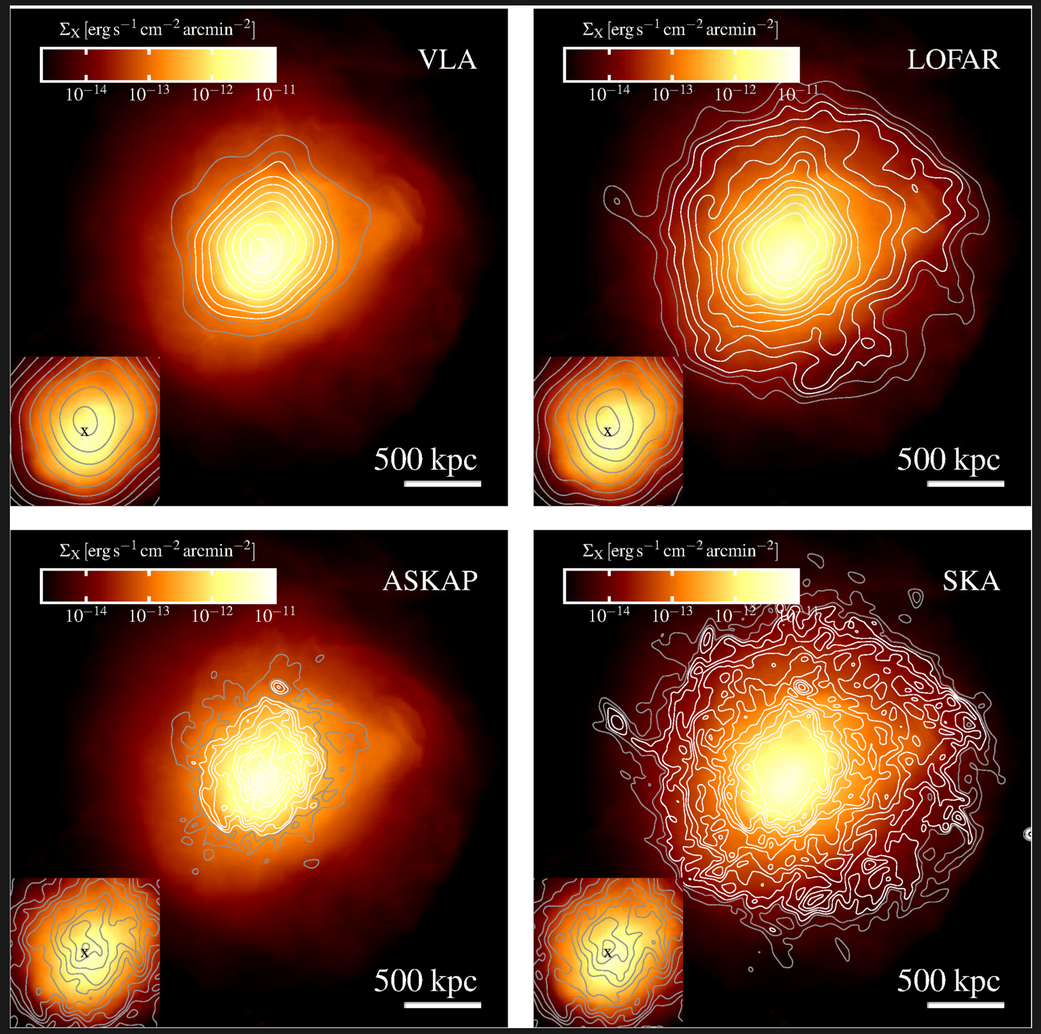


Magneto-hydrodynamics

magnetic field strength (TNG100), Credit: IllustrisTNG Collaboration



SKA mock observations



Estimated synchrotron emission of the most massive galaxy in TNG300 (at z=0) from Marinacci (2017); Credit: IllustrisTNG Collaboration

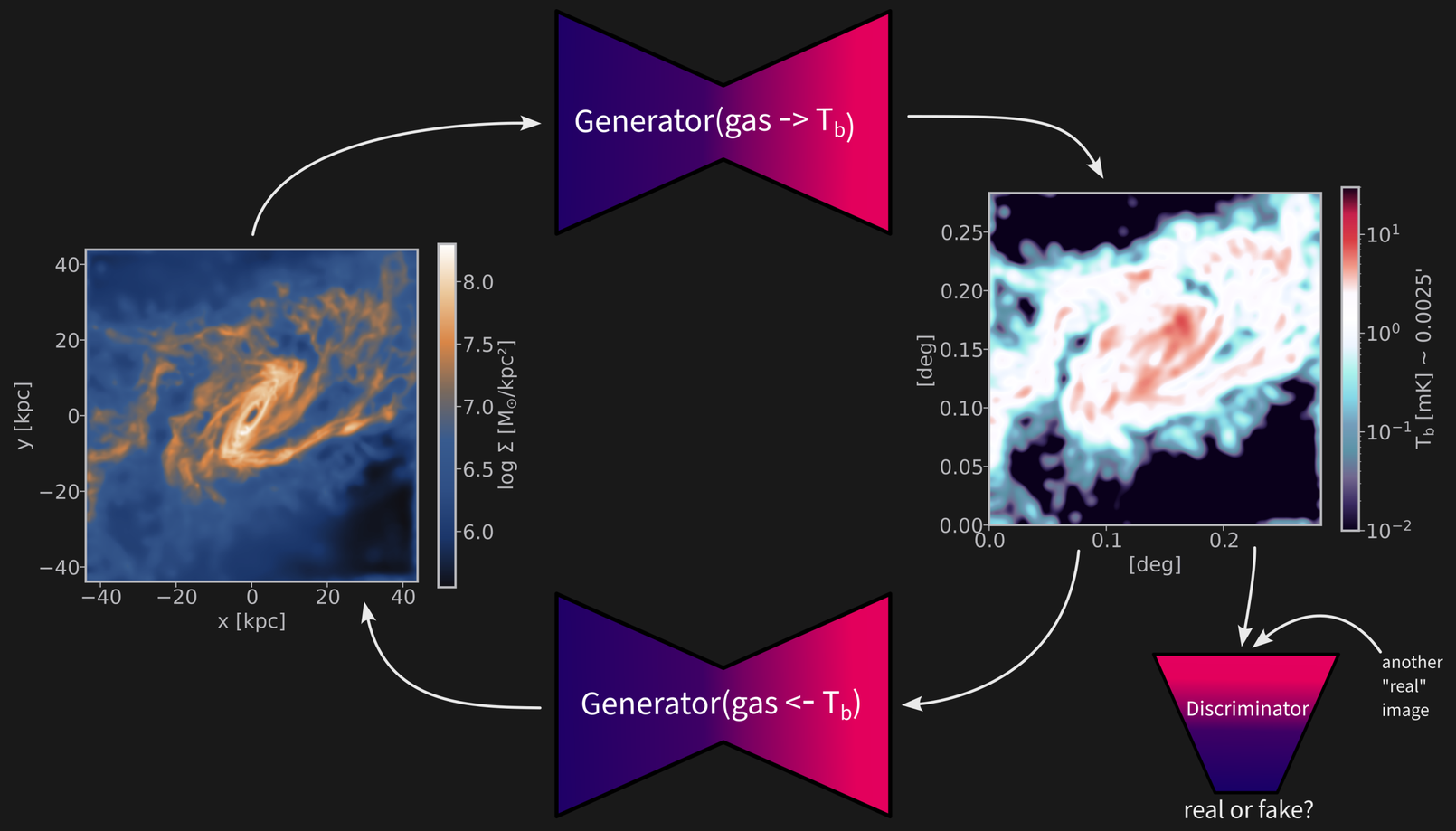
CYCLEGAN

Zhu et al. (2017)

- two generator - discriminator pairs
- learn the mapping from domain A ↔ B and vice versa

[Preliminary results]

- dataset: roughly 10'000 galaxies from TNG50-1
- brightness temperature of the gas $T_b(\mathbf{x}) = 189h \frac{H_0}{a^2 H(a)} \frac{\rho_{\mathrm{HI}}(\mathbf{x})}{\rho_c}\ \mathrm{mK}$



FUTURE PLANS

- include more physics
  - magnetic field strength
  - spectral models
  - noise
- actually simulate SKA instruments using OSKAR/Karabo
- try more types of generative deep learning models