Galaxies on Graph Neural Networks:

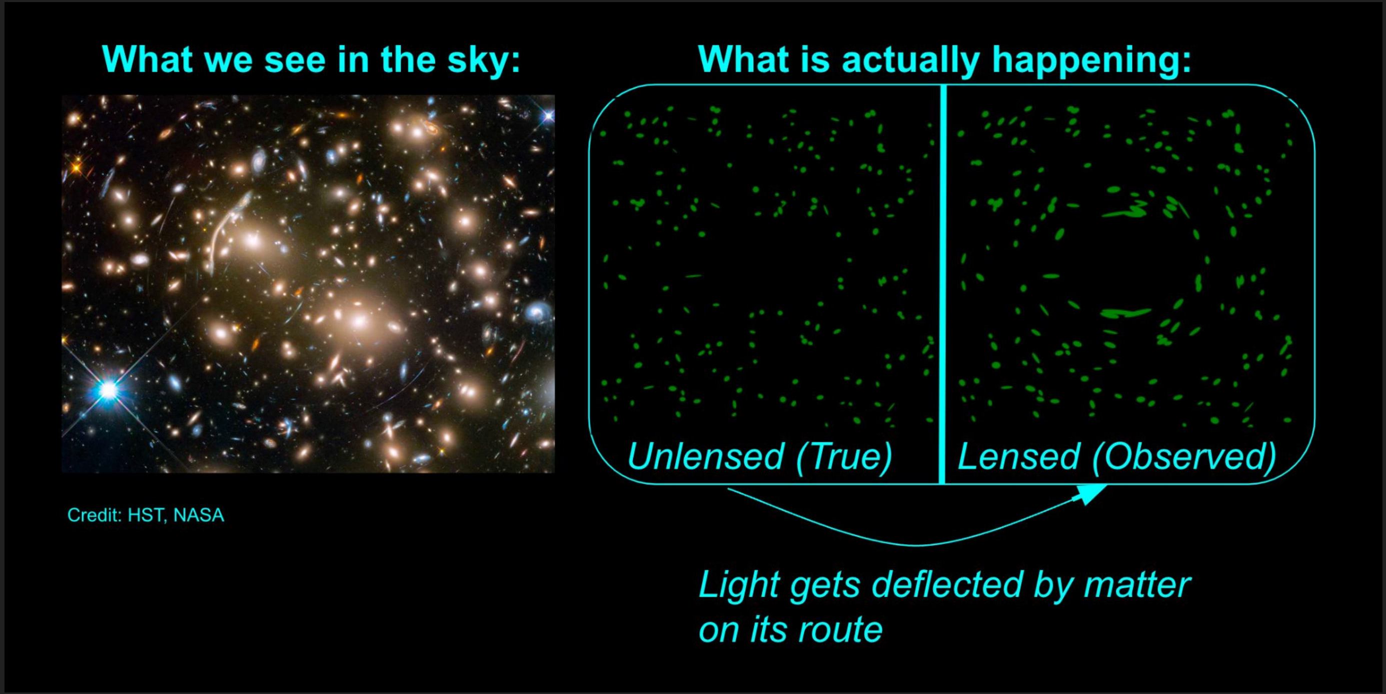
towards robust synthetic galaxy catalogs with deep generative models

Yesukhei Jagvaral (PhD student)

Carnegie Mellon University

Advised by:
Rachel Mandelbaum (CMU)
Francois Lanusse (Paris-Saclay/CCA Flatiron)

Introduction: Weak Lensing



- We measure the coherent shape distortions of m(b)illions of galaxies
- We can infer the matter content and the energy content of the Universe
- One important nuance: galaxies are not randomly oriented on the sky

Introduction: Intrinsic Alignment

Dark Energy equation of state parameters

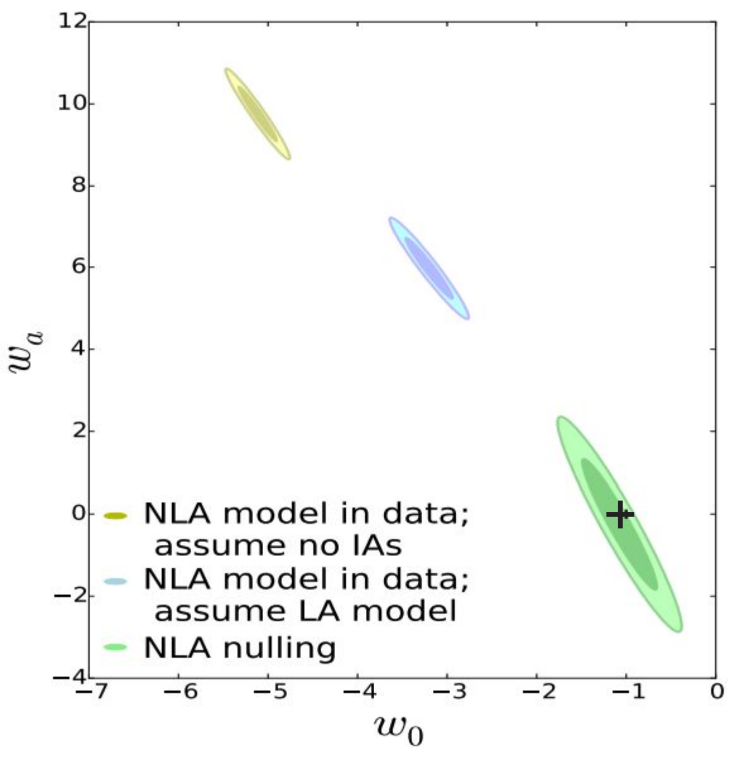
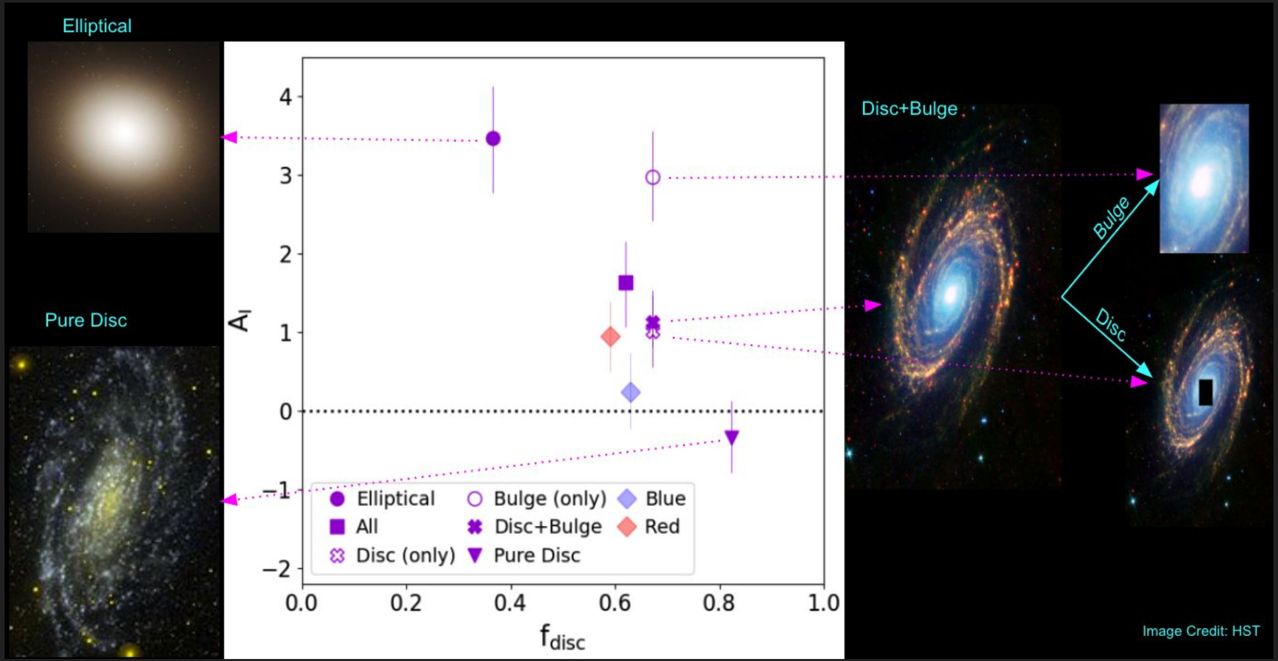


Image credit: Kirk et al, 2015

- Intrinsic Alignment (IA) is the tendency of galaxies to align with their neighboring galaxies and the underlying large scale structure.
- This effect can masquerade as a weak gravitational lensing signal
- Contributes to the systematic errors of weak lensing surveys.
- Need to develop good IA models and to test our ability to model/remove the IA signals using mock catalogs
- Need to include realistically complex IA in the catalogs

Intrinsic Alignment and morphology

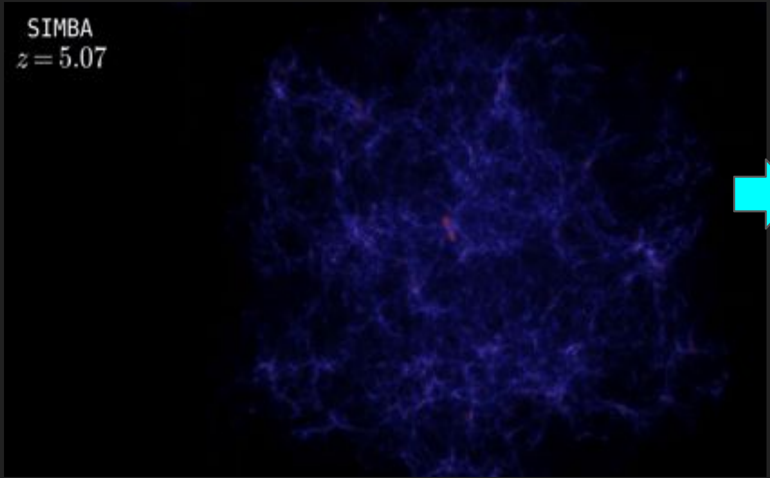


- We see a clear decreasing IA signal with increasing disky-ness
- Bulges do show statistically consistent signal with the Ellipticals
- Morphologically classified/decomposed samples reveal a much more complex picture compared to the traditional color split samples

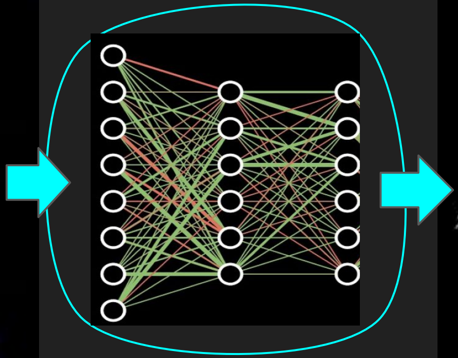
Jagvaral et al, 2022 MNRAS

Generating Galaxy Catalogs with Deep Learning

Gravity+Galaxy simulation

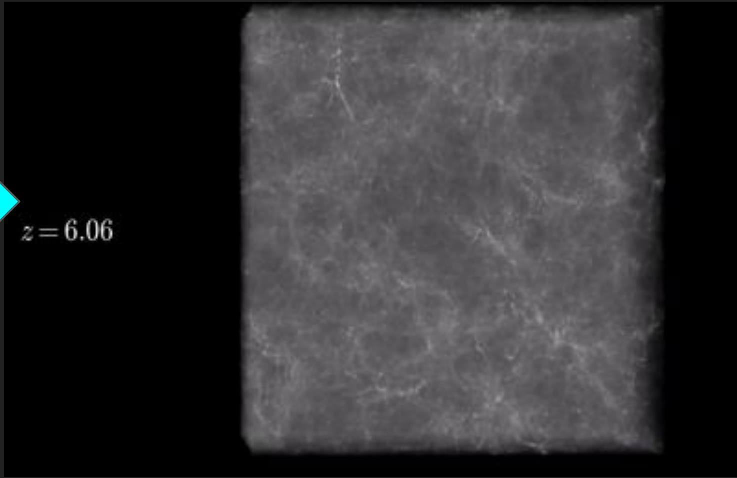


Deep Learning



Learn the “Galaxies”

Gravity-only simulation



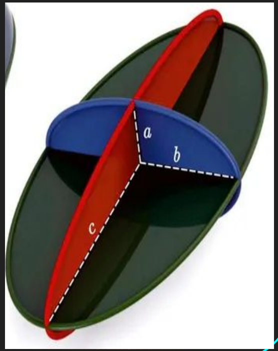
- Volume: Small
- Resolution: High
- Cost: High

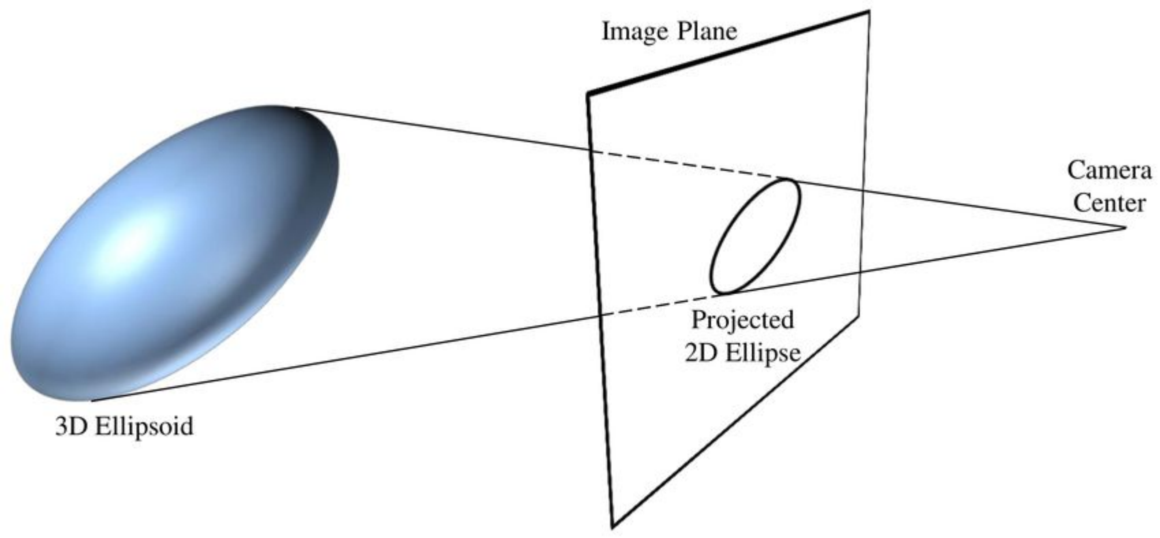
- Volume: Large
- Resolution: Low
- Cost: Low

Credit: 3Blue1Brown, camel-simulations.org

Galaxy orientations in the simulations

Galaxy orientations in 3D





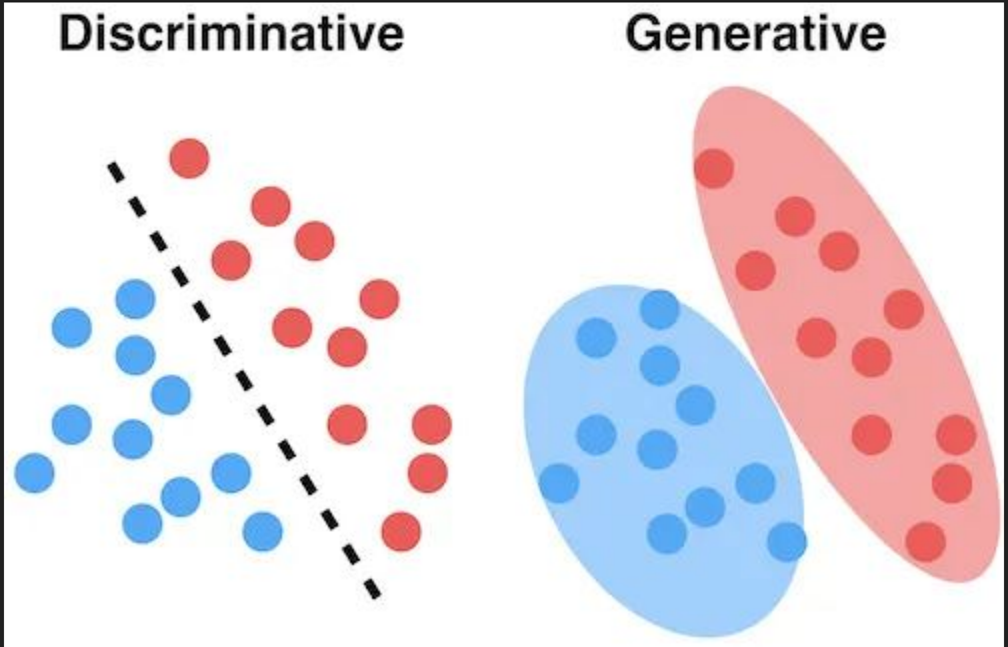
Sridhar et al, 2015

- Galaxy orientations as a function of their environment
- p(orientation | environment)
- Tools that are needed are deep generative models

Outline

1. What are Weak lensing and Intrinsic Alignment?
2. What are Generative models?
3. Our models: GANs ‘n’ Graphs
4. Our models: Diffusion Generative on SO(3) manifold

Introduction: Deep Generative models



- Can be sampled from
- Can be used as density estimators
- More generic than discriminative models
- Sensitive to outliers, can be used for anomaly detection

Image Source:
https://betterprogramming.pub/

Motivation: Why Generative Models?

Data

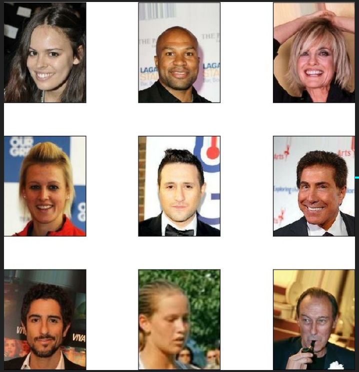
Condition

Emulator/Data simulator

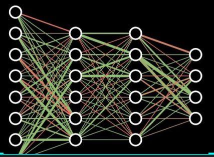
Probability

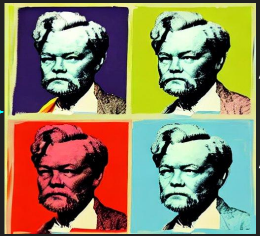
New data

Deep Prior



Deep generative model





Andrew Carnegie in the style of Andy Warhol

CelebA dataset, Stable Diffusion

Introduction: Deep Generative Models

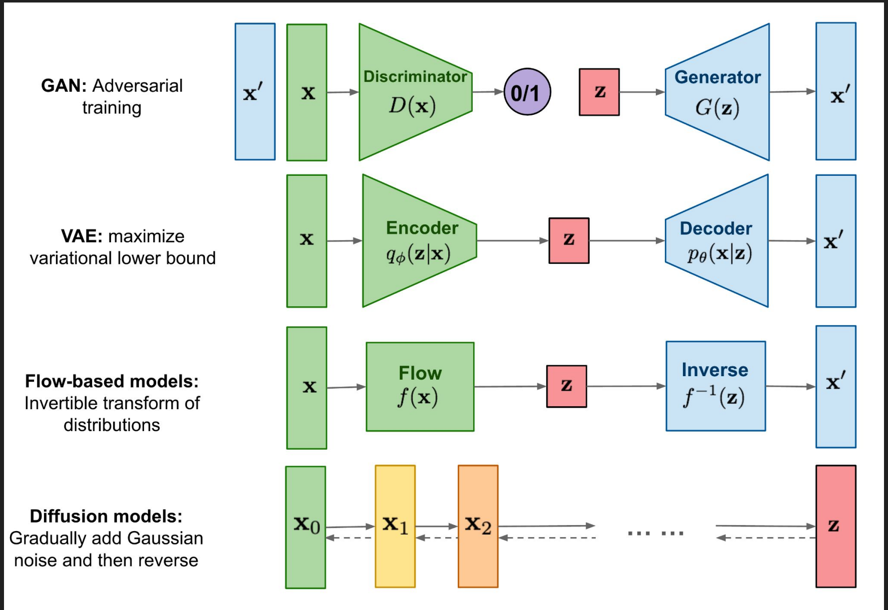
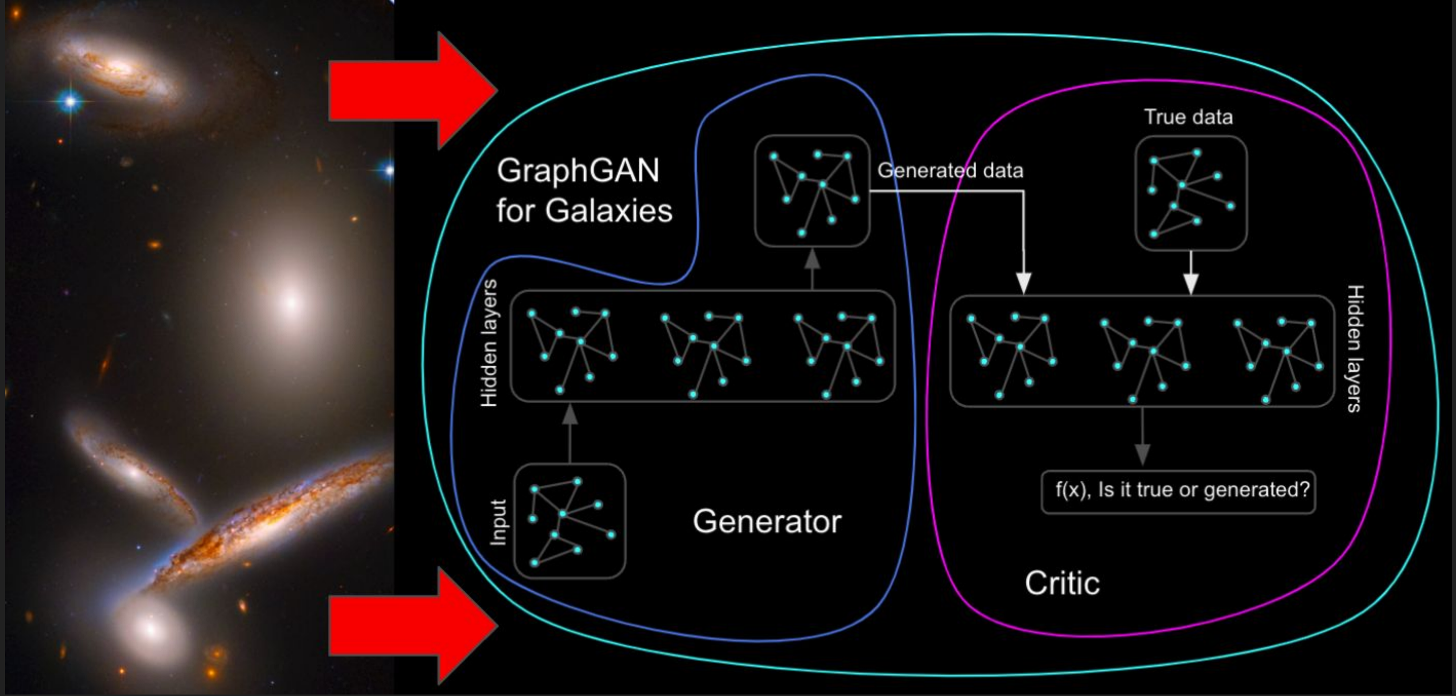


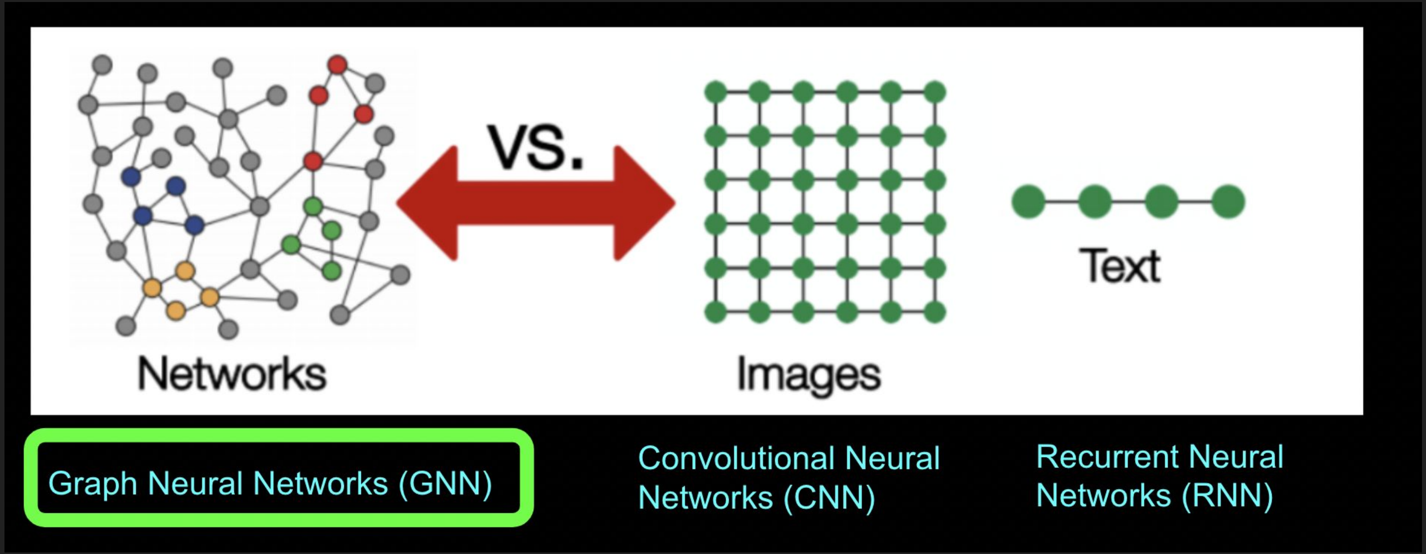
Image credit: Lilian Weng

- Current state of the art generative models
- Has shown great sample quality in images, audio etc...



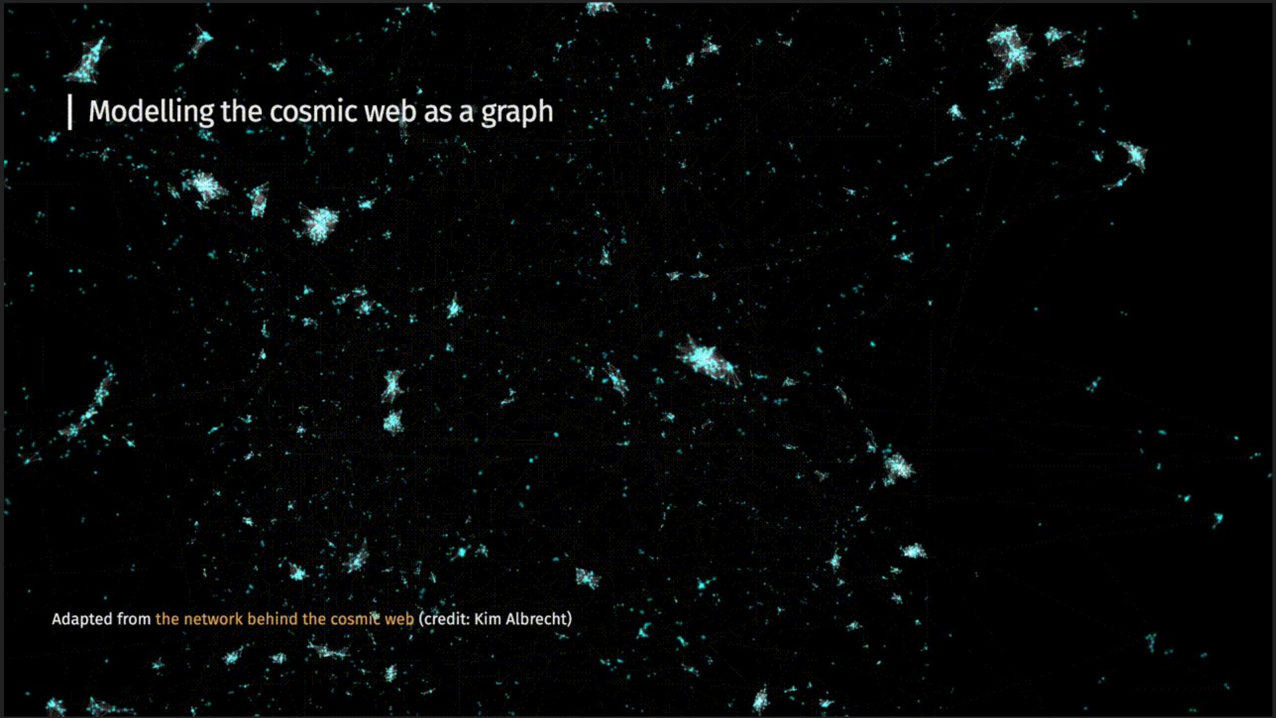
Part 1
GANs ‘n’ Graphs

What type of Neural Network do we need?



- CNNs are appropriate for grid-like data
- RNN are appropriate for time-series data
- Graphs are appropriate for sparsely distributed objects
- Also, Graphs are appropriate for capturing correlations among objects

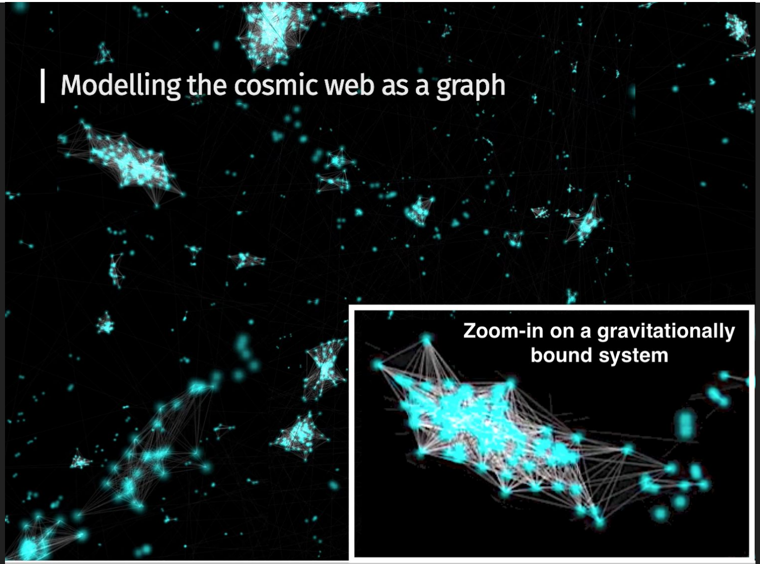
The Cosmic Web as a Set of Graphs



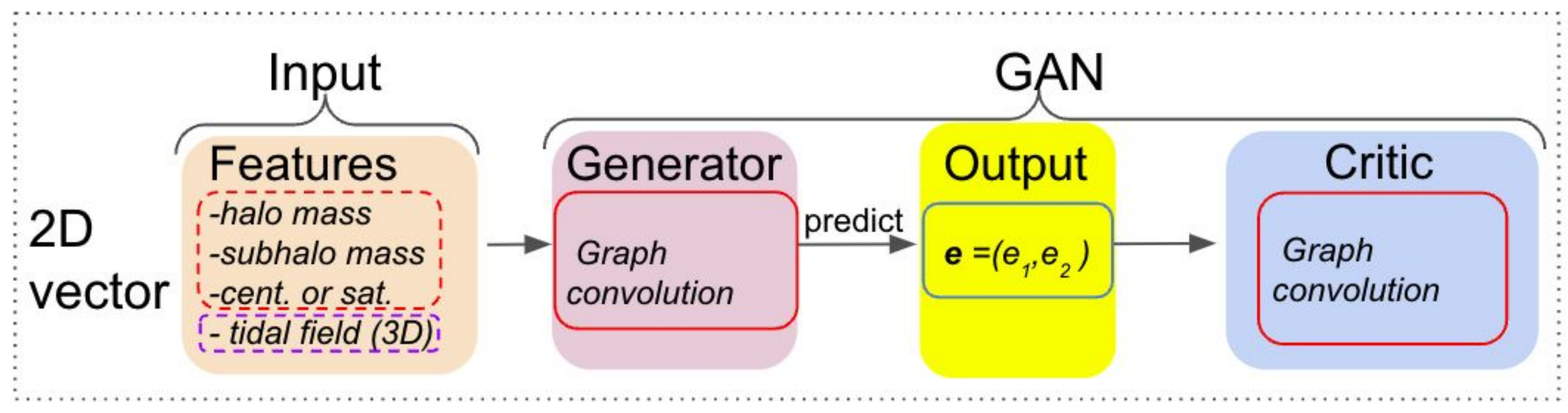
IllustrisTNG100 modeled as a set of Graphs

The Graph Convolution

$$y_i = b + \underbrace{\mathbf{W}_0 h_i}_{\text{self-connection}} + \underbrace{\sum_{m=1}^{M} \sum_{j \in \mathcal{N}_i} \overbrace{q_m(\mathbf{x}_i, \mathbf{x}_j)}^{\text{directional-distance}} w_{i,j} \, \mathbf{W}_m h_j}_{\text{average over neighbors}}$$

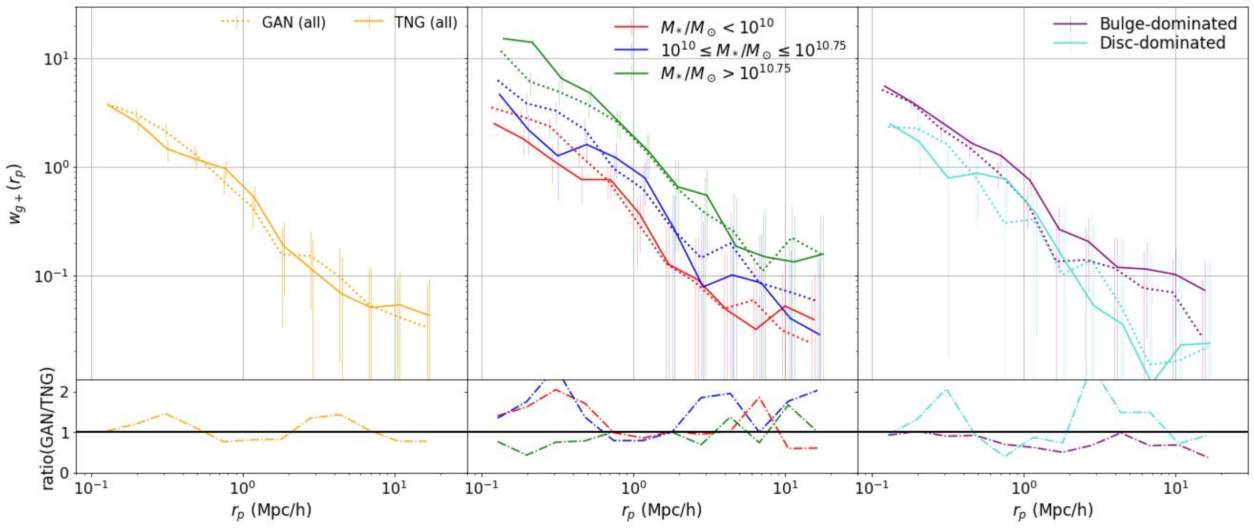


Results for the 2D position-shape correlations



- To the best of our knowledge, this is the first instance of a generative model on graphs in an astrophysical/cosmological context

Results for the 2D position-shape correlations

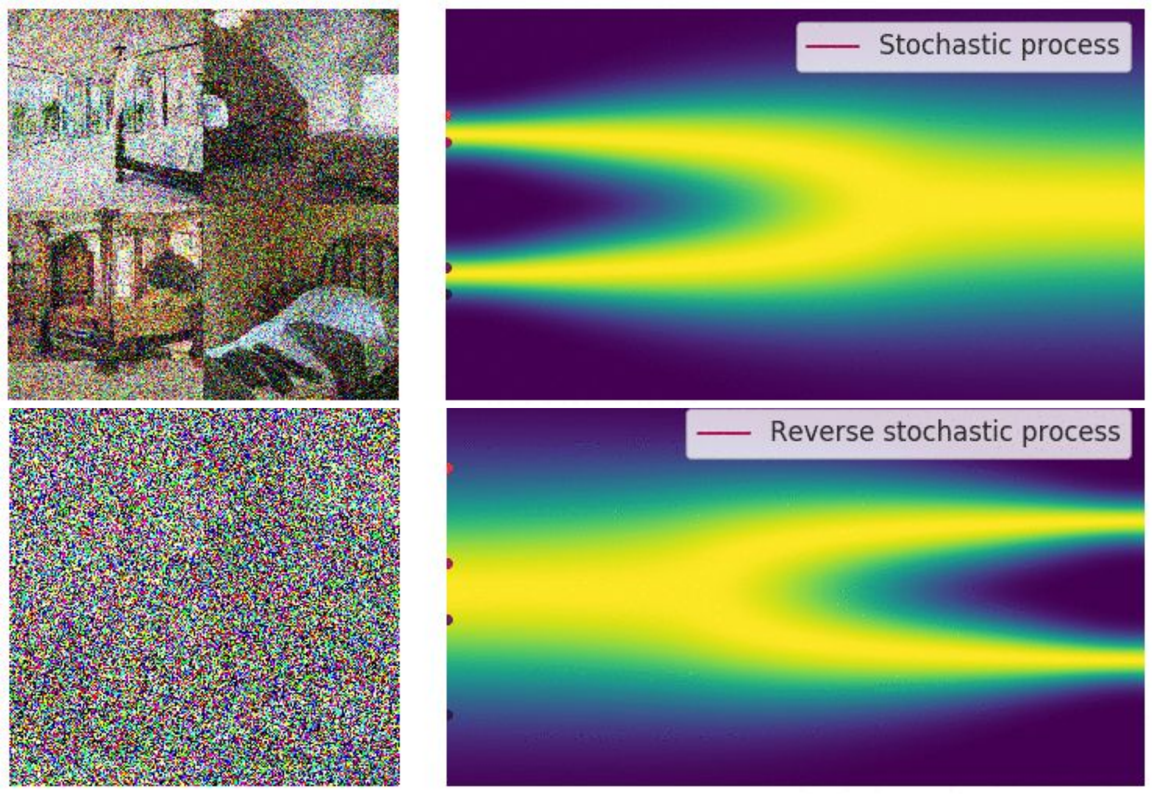


Projected two-point correlation functions $wg+$ of galaxy positions and the projected 2D ellipticities of all galaxies.

(higher values mean galaxies point towards neighboring ones more strongly on average)

- Good quantitative agreement between the model and the simulation
- To the best of our knowledge, this is the first instance of a generative model on graphs in an astrophysical/cosmological context

Full paper published in *MNRAS* as **Jagvaral et al, 2022**



Part 2
Diffusion

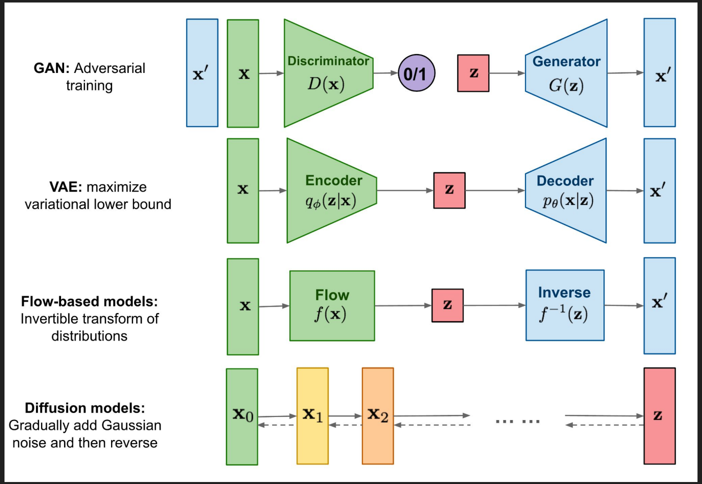
Image credit: NVIDIA

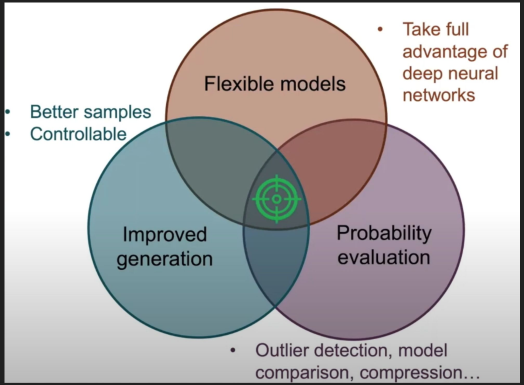
Challenges with Generative Models?

- High expressivity
- Mode coverage
- Tractable/computable
- *Slow sampling*

Diffusion models:

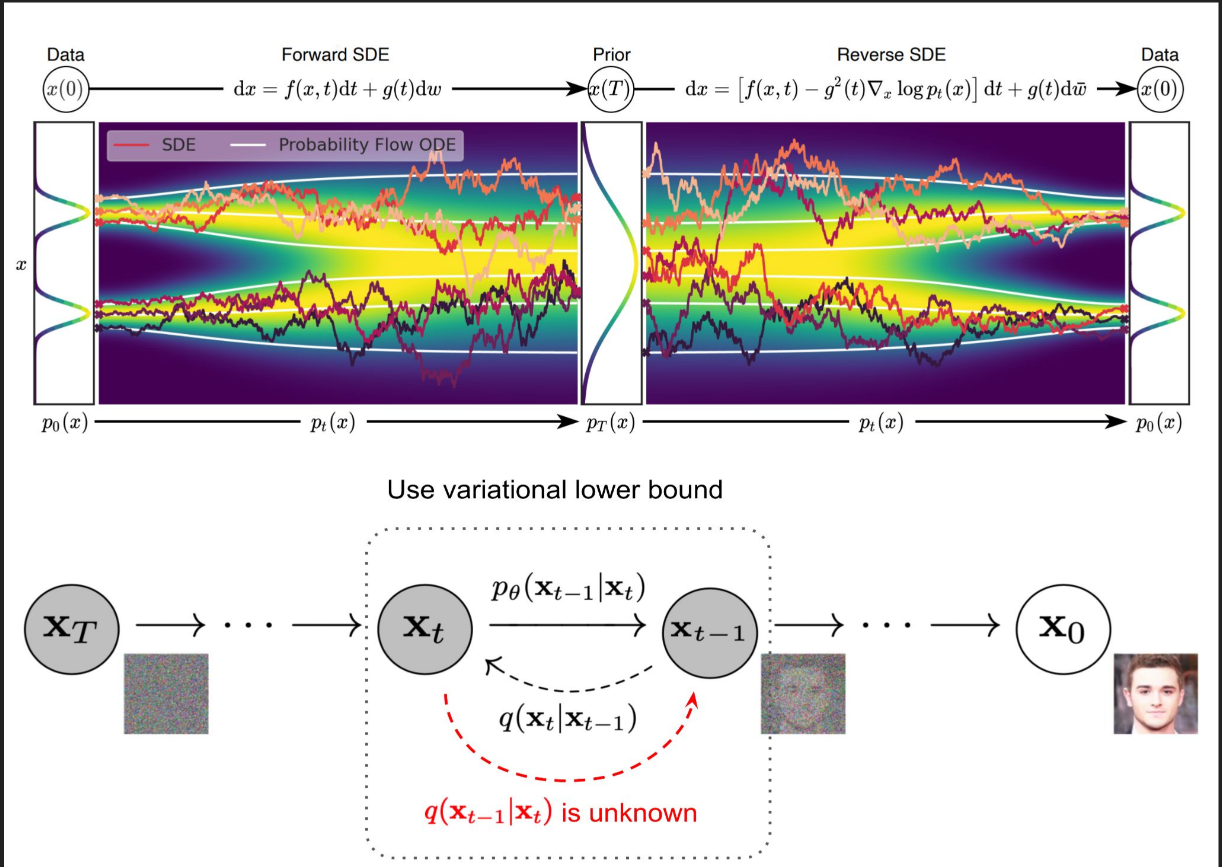
1. **Score-matching**
2. **Denoising diffusion**





Yang Song, 2022

Quick Intermission Summary



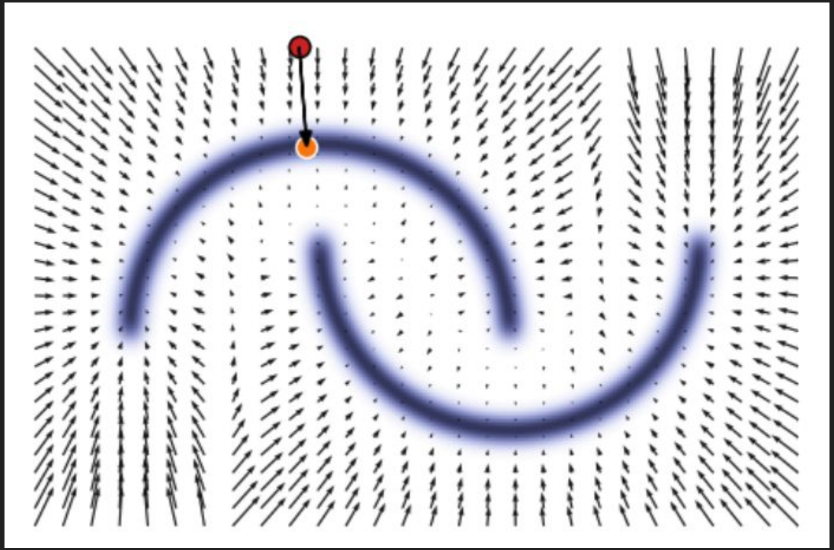
Ho et al, 2021

Proposal: Score function

Given a PDF **p(x),**

the (Stein) score function is defined as:

$$\nabla_x \log p(x)$$



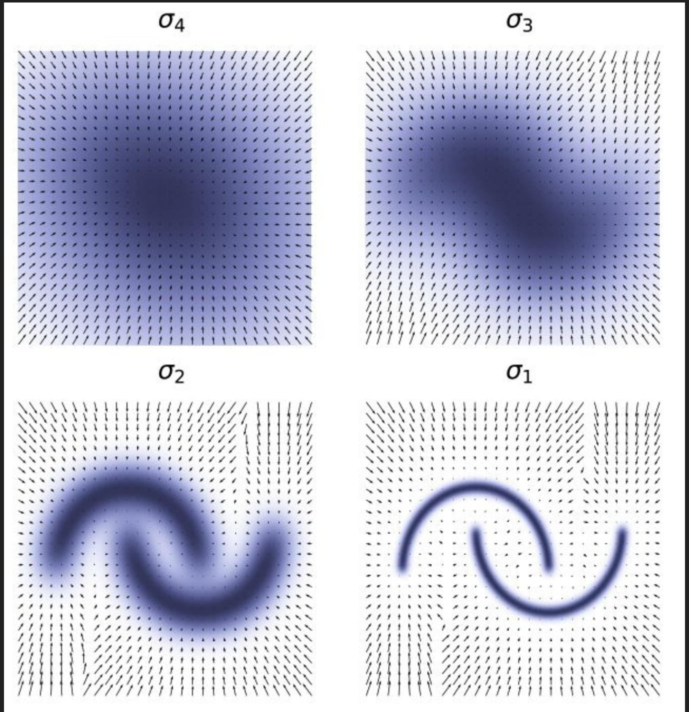
- Bypasses the costly normalization constant
- Accurate probability evaluations
- Controllable generation

Remy et al, 2021

Proposal: Score function

∇_x**log p(x)** approximate

with a neural network $\mathbf{r}_\theta$



- One problem original score field has bad coverage
- Use noising process to perturb the data

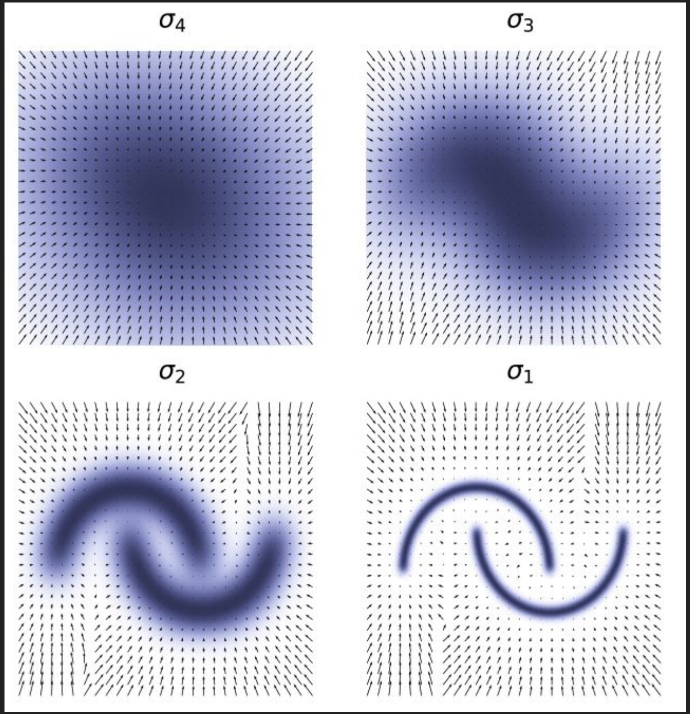
Remy et al, 2021

Proposal: Score function

∇_x**log p(x)** approximate

with a neural network $\mathbf{r}_\theta$

$$\mathcal{L}_{\mathrm{DAE}} = \mathbb{E}_{\boldsymbol{x}' \sim p_{\sigma^2}} \left[\|\boldsymbol{x} - \boldsymbol{r}_\theta(\boldsymbol{x}', \sigma)\|_2^2 \right].$$



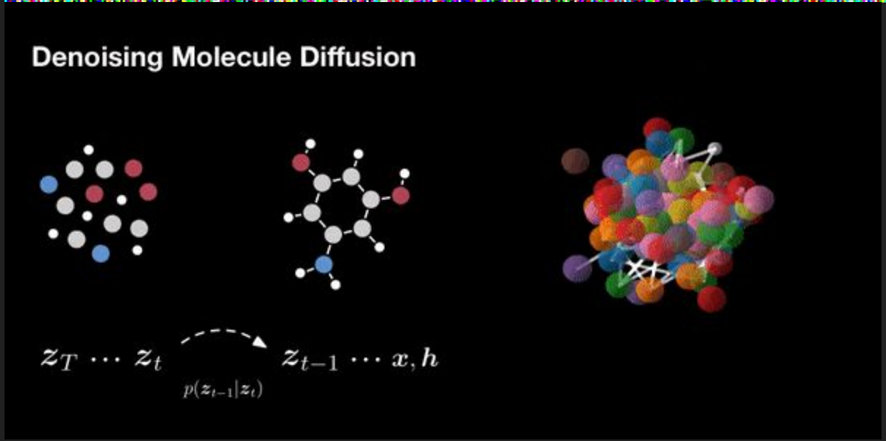
Remy et al, 2021

Score-based modeling in practice

So far the state-of-the-art for:

1. **Images**
2. **Video**
3. **Audio**
4. **Molecules**
5. ***Astrophysics?***

keras.io

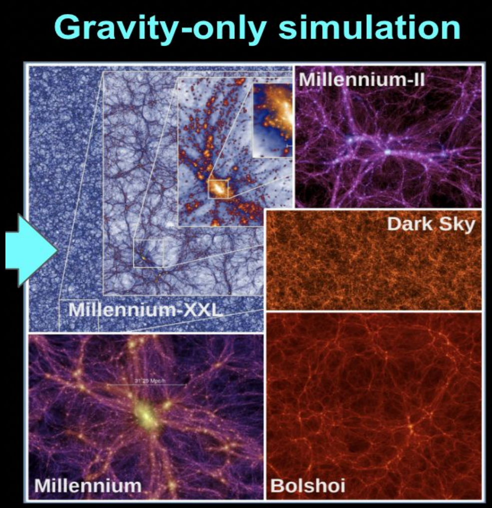


Hooge boom et al, 2021

Generating Galaxy Catalogs with Deep Learning

Traditionally:
Semi-analytic models
were used to
“paint” galaxies

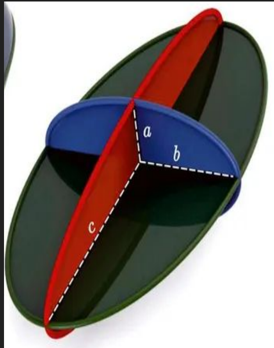
Credit: 3Blue1Brown, Vogelsberger et al 2020



- Mock catalogs are used to design/test analysis pipelines.
- Challenges:
    a. Galaxies are very expensive to simulate
    b. Large volumes with Galaxies are unreachable (Resolution vs Volume)
- We propose a Deep Generative model for making synthetic Galaxy Catalogs with realistic galaxies

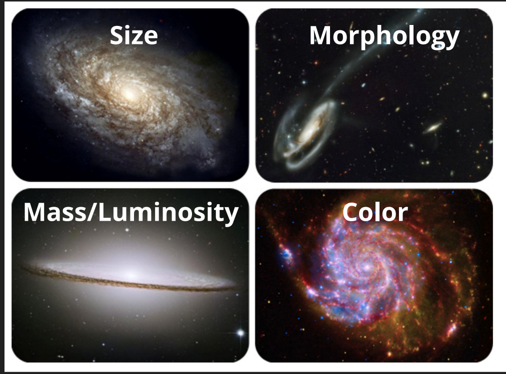
Galaxy properties

Galaxy orientations in 3D

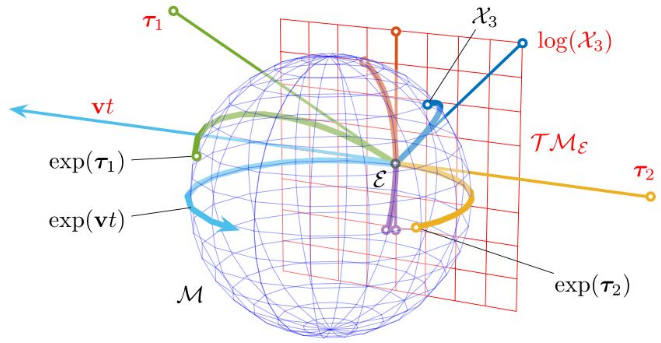


non-Euclidean manifold

Scalars



Euclidean manifold



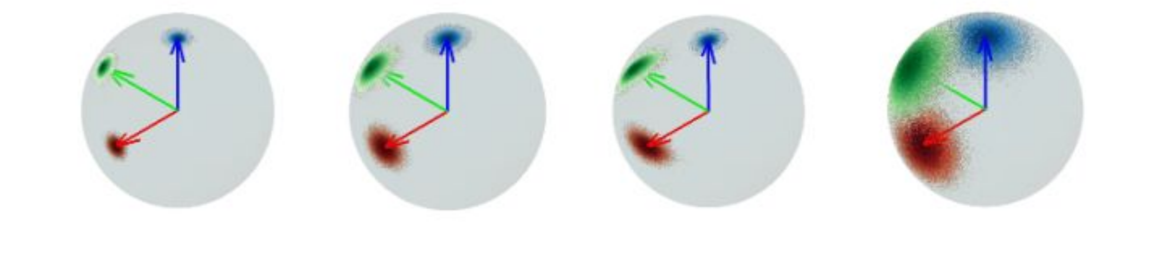
SO(3) - Special orthogonal group of 3D

- Constrained to the 4D hypersphere (Quaternion representation of rotations)

Credit: HST, forbes.com

Diffusion on SO(3): Training

- Diffusion Process on SO(3)



- Noise kernel

$$\mathcal{IG}_{\mathrm{SO(3)}}(\mathbf{x};\boldsymbol{\mu},\epsilon) = f_\epsilon\left(\arccos\left[2^{-1}\left(\mathrm{tr}(\boldsymbol{\mu}^T\mathbf{x}) - 1\right)\right]\right)$$

- Score of the noise kernel

$$\nabla_{X_i} \log p_\epsilon(\tilde{\mathbf{x}}|\mathbf{x}) = \frac{\mathrm{d}}{\mathrm{d}s} \log p_\epsilon(\tilde{\mathbf{x}} \exp(sX_i)|\mathbf{x})\Big|_{s=0}$$

- Denoising score matching

$$\mathcal{L}_{DSM} = \mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma_\epsilon^2)}\mathbb{E}_{p_{|\epsilon|}(\tilde{\mathbf{x}}|\mathbf{x})} \left[|\epsilon| \ \| s_\theta(\tilde{\mathbf{x}},\epsilon) - \nabla_X \log p_{|\epsilon|}(\tilde{\mathbf{x}}|\mathbf{x}) \|_2^2\right]$$

Credit: Gilitschenski et al, 2020

Diffusion on SO(3)

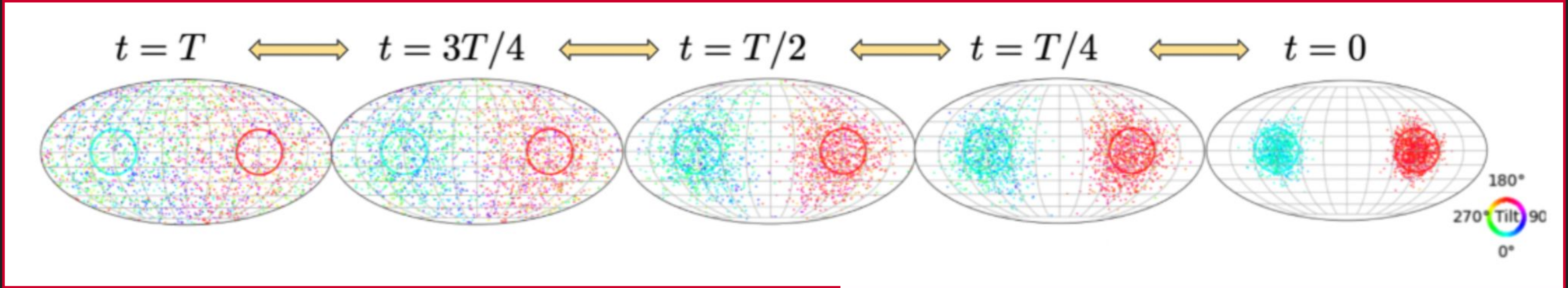
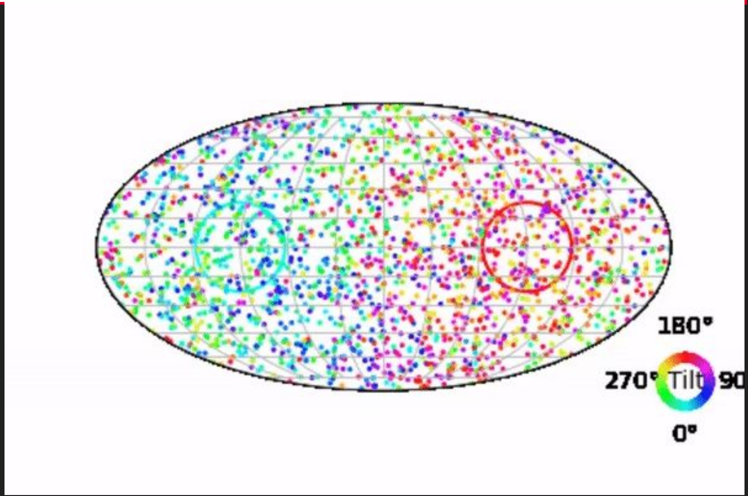


Illustration of reversible diffusion of a mixture of two Gaussian blobs on SO(3)



Diffusion on SO(3): Results

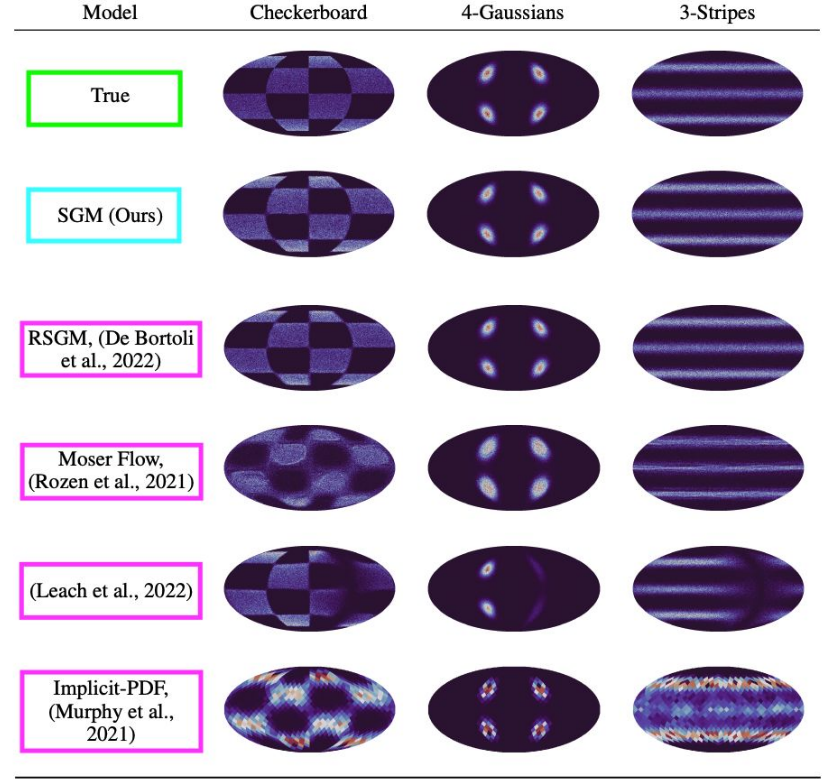


Figure 3: Density plot comparing samples from learned synthetic densities on SO(3). For visualization this density plot shows the distribution of canonical axes of sampled rotations projected on the sphere; the tilt around that axis is discarded.

Our model achieves the best results:
quantitatively and visually

Full paper at OpenReview.net
(submitted in *ICML* as **Jagvaral et al, 2023)**

Diffusion on SO(3): Results

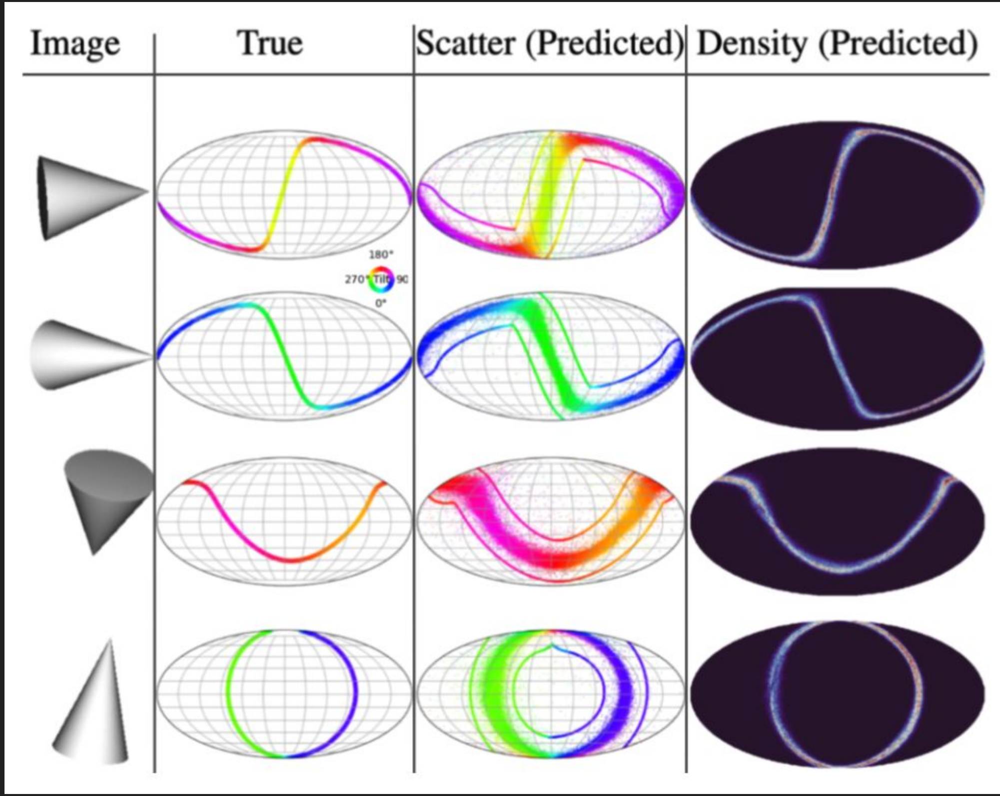
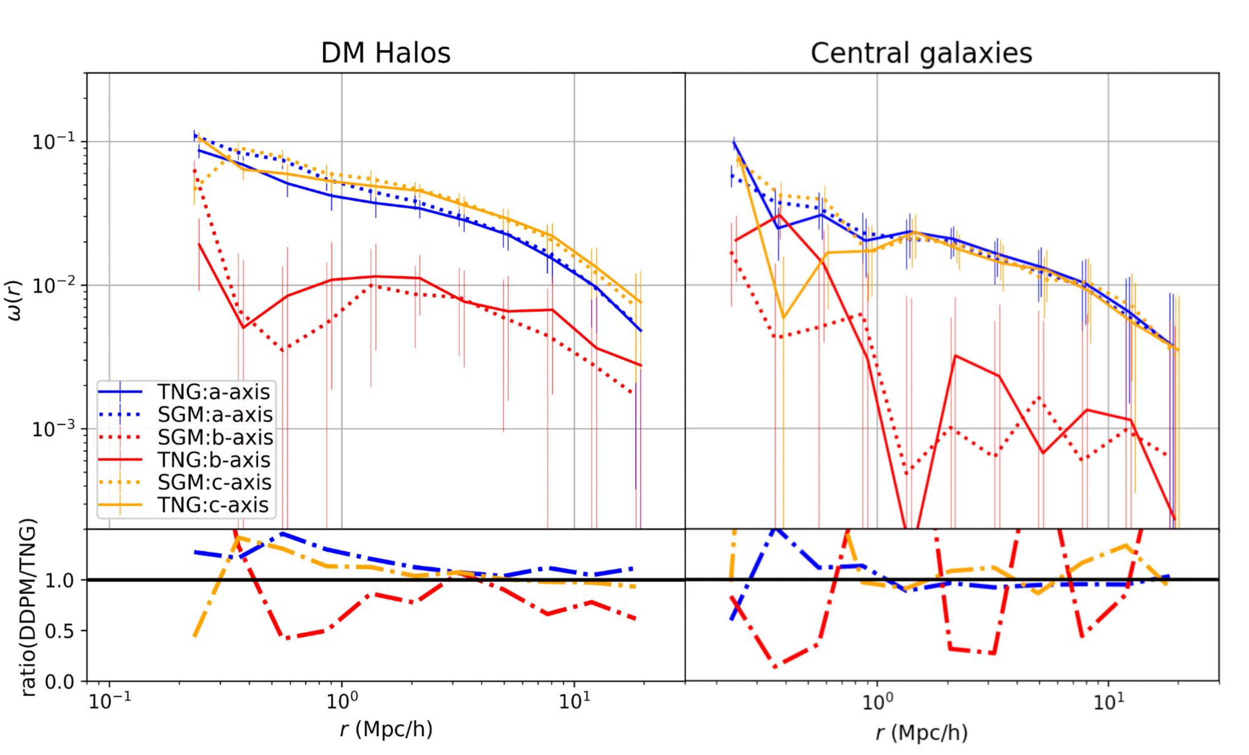


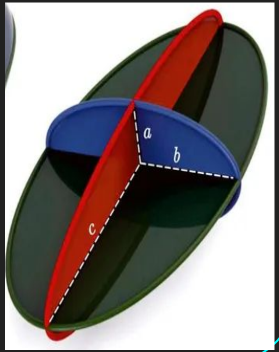
Illustration of pose estimation task in computer vision/robotics

Full paper at OpenReview.net
(submitted in *AAAI* as **Jagvaral et al, 2023)**

Diffusion on SO(3): Results



Galaxy orientations in 3D



Generative Diffusion on SO(3): Use cases

A Deep Generative Model For Galaxy Orientations:

- Need to implement Graphs for the non-linear regime
- Study distribution shifts

Computer Vision and Robotics

- Pose and orientation estimation in the real world

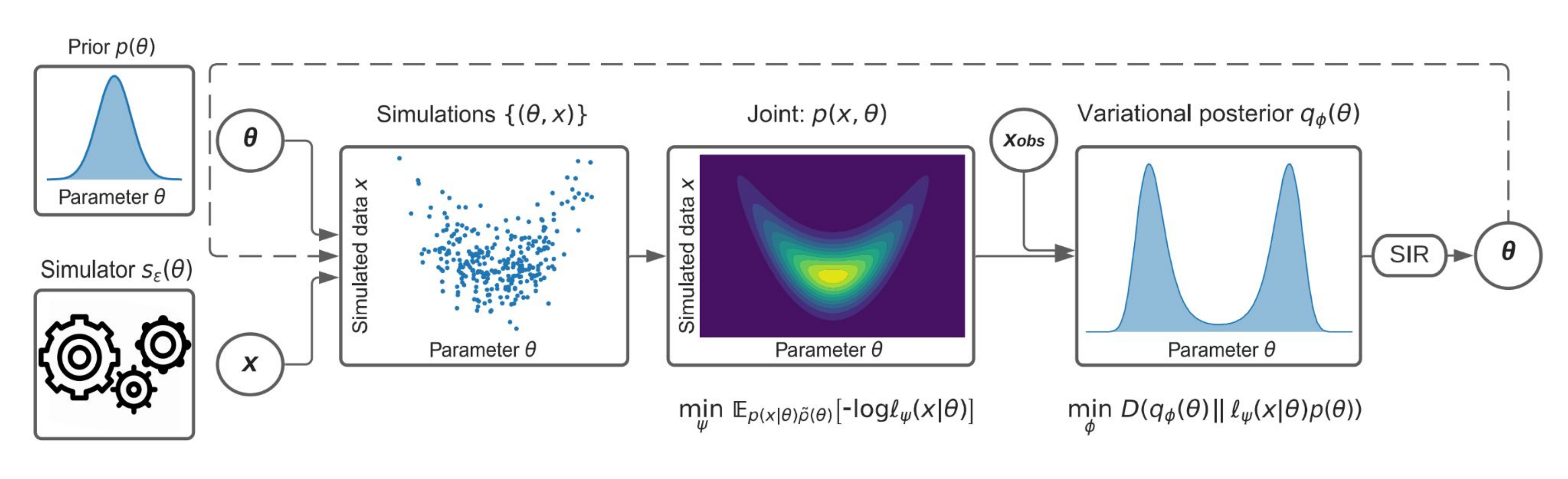
Simulations based Inference (Likelihood free) that has SO(3) symmetry

- Gravitational Wave Inference (locality)

Diffusion on SO(3): SBI Grav. Waves

Simulations based Inference (Likelihood free) that has SO(3) symmetry

- SBI, an emerging method to do inference
- Likelihood free
- Leverages advances in ML, brings in the shortcomings of ML

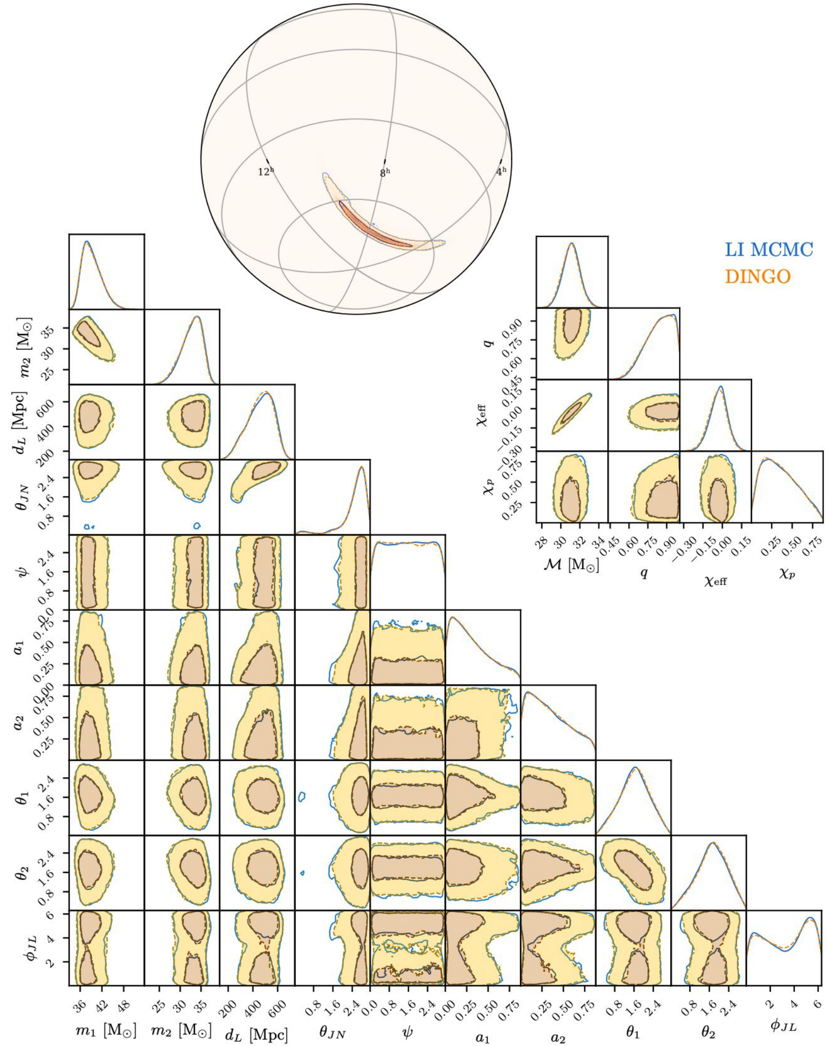


Glockler et al, 2022

SBI Grav. Waves

Simulations based Inference (Likelihood free) future of stats?

- DINGO based on Normalizing flows
- Diffusion models better in handling manifold valued data than Normalizing flows



Green & Gair, 2020

Our Contributions

A Deep Generative Model For Galaxy Orientations:

- We propose a GAN and Graph based approach, good quantitative and qualitative agreement with the baseline simulation
- We extend current SOTA diffusion onto the SO(3) manifold
- Achieve SOTA results on synthetic dataset
- Showed application in astrophysical context and computer vision
- Further work is needed to fully harness its power.
- Produce mock catalogs for DESC/LSST with Argonne group

Future work: Yes, many :), happy to chat