Modeling halo and central galaxy orientations on the SO(3) manifold with score-based generative models

Yesukhei Jagvaral[*] **Rachel Mandelbaum**
McWilliams Center for Cosmology
NSF AI Planning Institute for Data-Driven Discovery in Physics
Department of Physics,
Carnegie Mellon University
Pittsburgh, PA 15213 USA

François Lanusse
AIM, CEA, CNRS
Université Paris-Saclay
Université Paris Cité
91191 Gif-sur-Yvette, France

Abstract

Upcoming cosmological weak lensing surveys are expected to constrain cosmological parameters with unprecedented precision. In preparation for these surveys, large simulations with realistic galaxy populations are required to test and validate analysis pipelines. However, these simulations are computationally very costly – and at the volumes and resolutions demanded by upcoming cosmological surveys, they are computationally infeasible. Here, we propose a Deep Generative Modeling approach to address the specific problem of emulating realistic 3D galaxy orientations in synthetic catalogs. For this purpose, we develop a novel Score-Based Diffusion Model specifically for the SO(3) manifold. The model accurately learns and reproduces correlated orientations of galaxies and dark matter halos that are statistically consistent with those of a reference high-resolution hydrodynamical simulation.

1 Introduction

Future wide-field astronomical imaging surveys, such as the Vera C. Rubin Observatory Legacy Survey of Space and Time[2], Roman Space Telescope[3] High Latitude Survey and Euclid[4] will provide precise constraints on cosmological parameters by imaging billions of galaxies. Deriving physical understanding from these data will require increasingly costly large-volume simulations with high resolution to test and validate analysis pipelines [DeRose et al, 2019, 2021, Korytov et al., 2019] and to constrain cosmology via Simulation-Based Inference [SBI; Jeffrey et al., 2021].

In this regard generative machine learning approaches represent an interesting avenue as they could serve as fast and robust emulators to greatly accelerate parts of the simulation pipelines. In particular, they could be used to populate realistic galaxies in large volume dark matter only simulations. Most machine learning methods in this line of research have been concerned with modeling scalar properties of galaxies, however in this work we are particularly interested in modeling the 3D orientations of galaxies and their host dark matter halos in simulations. These intrinsic orientations can indeed contaminate measurements of weak gravitational lensing in upcoming surveys and constitute a major source of systematic errors if not accounted for [Joachimi et al., 2015].

Diffusion models are flexible in their domains that the datal ives, we want to jointly model various properties that live on various different spaces/manifolds

[*] yjagvara@andrew.cmu.edu

[2] `https://www.lsst.org/`

[3] `https://roman.gsfc.nasa.gov/`

[4] `https://www.euclid-ec.org/`

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

Currently, score-based diffusion models represent the state-of-the-art in generative tasks such as: image, audio and molecules generation. [Hoogeboom et al., 2022]. Modeling distributions on the manifold of 3D rotations is however a non trivial task, and to address this problem we develop a new type of score-based diffusion model specifically for the SO(3) manifold, by extending the Euclidean framework introduced in Song et al. [2021]. We chose diffusion models due to their flexibility to model data that live on various different spaces (e.g. scalars and rotation matrices) compared normalizing flows and due to their stability compared to Generative Adversarial Networks. Based on these developments, we build a conditional generative model on SO(3) which allows us to sample from the posterior distribution of 3D orientations of galaxies and dark matter halo given information about their surrounding gravitational tidal field.

2 Related Work

Machine learning approaches have been adopted in astrophysics and cosmology in various contexts, including emulation methods, inference and forward modeling [Dvorkin et al., 2022]. In particular, deep generative models have been implemented in the works of Jagvaral et al. [2022] for generative modeling of correlated galaxy properties, such as shapes and orientations, with graph-based generative adversarial networks. Our work takes the next step to build generative models for various galaxy properties associated with galaxy and halo orientations (which are described by a non-Euclidean manifold) with score-based denoising diffusion models.

3 Score-Based Generative Model on SO(3)

Here we briefly outline our novel approach for modeling distributions on SO(3), heavily inspired by the diffusion framework developed in Song et al. [2021]. The idea behind diffusion models is to introduce a noising process that perturbs the data distribution until it reaches a nearly pure noise distribution. Consider the following Stochastic Differential Equation (SDE) on the SO(3) manifold:

$$\mathrm{d}X = \mathbf{f}(X,t)\,\mathrm{d}t + g(t)\,\mathrm{d}W, \tag{1}$$

where W is a Brownian process on SO(3), $\mathbf{f}(\cdot\,,t) : \mathrm{SO}(3) \to T_X\mathrm{SO}(3)$ is a drift term, and $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is a diffusion term. Given samples $X(0) \sim p_{\text{data}}$ from an empirical data distribution p_{data} at time $t = 0$, the marginal distribution of samples $X(t)$ evolved under this SDE at a subsequent time $t > 0$ will be denoted p_t, and will converge for large $t = T$ towards a given predetermined distribution p_T typically chosen to be easy to sample from. On SO(3), a natural choice for p_T is $\mathcal{U}_{SO(3)}$, the uniform distribution on SO(3).

The key realization of Song et al. [2021] is that under mild regularity conditions this noising process of the data process can be reversed, in particular through the following so-called probability flow Ordinary Differential Equation (ODE):

$$\mathrm{d}X = [\mathbf{f}(X,t) - g(t)^2 \nabla \log p_t(X)]\mathrm{d}t. \tag{2}$$

De Bortoli et al. [2022] recently extended this result to compact Riemannian manifolds, which include in particular SO(3). This deterministic process is entirely defined as soon as the *score function* $\nabla \log p_t(X) \in T_X\mathrm{SO}(3)$ is known, and running this ODE backward in time from samples $X(T) \sim p_T$ down to $t = 0$ will yield samples $X(0) \sim p_0 = p_{\text{data}}$. Training such a generative model will therefore boil down to estimating this score function with a neural network.

While these results are direct analogs of Euclidean diffusion models [as in Song et al., 2021], implementing similar models on SO(3) brings practical difficulties: Unlike in the Euclidean case where the Gaussian is a closed-form solution of heat diffusion (a key element in Euclidean SGMs), there is no closed-form solution on general Riemannian manifolds. Our contribution is to propose solutions to these issues in order to implement efficient score-based diffusion models on SO(3).

On SO(3), although the exact heat kernel is only available as an infinite series [Nikolayev and Savyolov, 1970], it can be robustly approximated in practice either by truncating this series or by using a closed form expression [Matthies et al., 1988], depending on the width of the kernel. It is used to define the so-called Isotropic Gaussian Distribution on SO(3), $\mathcal{IG}_{\mathrm{SO}(3)}(R,\epsilon)$ Nikolayev and Savyolov [1970], Matthies et al. [1988], Leach et al. [2022], where $R \in \mathrm{SO}(3)$ is a mean rotation matrix, and ϵ a scale parameter. $\mathcal{IG}_{\mathrm{SO}(3)}$ enjoys tractable likelihood evaluation and sampling, and most importantly is closed under convolution.

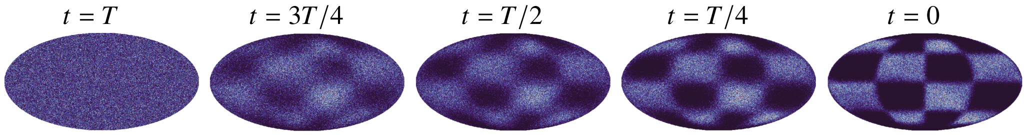

Figure 1: Learned synthetic density on SO(3). On the left, starting from uniform noise on the sphere at $t = T$, solving the ODE Equation 4 transports noise samples back into the target density at $t = 0$.

We can now define a noise kernel $p_\epsilon(X|X') = \mathcal{IG}_{\mathrm{SO(3)}}(X; X', \epsilon)$ which can be used to convolve the data distribution such that $p_\epsilon(X) = \int_{SO(3)} p_{\mathrm{data}}(X') p_\epsilon(X|X') \, \mathrm{d}X'$. For simplicity, we further make the following specific choice, for the diffusion SDE Equation 1: $\mathbf{f}(X, t) = 0$, $g(t) = \sqrt{\frac{\mathrm{d}\epsilon(t)}{\mathrm{d}t}}$ where $\epsilon(t)$ is a given noise schedule (e.g. $\epsilon(t) = t$). We then recover that convolving the data distribution with an $\mathcal{IG}_{\mathrm{SO(3)}}$ of scale $\epsilon(t)$ corresponds to the marginal distribution of the SDE at time t: $p_{\epsilon(t)} = p_t$.

This noise kernel allows us to use on SO(3) the usual Denoising Score-Matching loss at no extra complexity compared to the Euclidean case. To learn the score function we introduce a neural score estimator $s_\theta(X, \epsilon) : \mathrm{SO(3)} \times \mathbb{R}^{+\star} \longrightarrow \mathbb{R}^3$, which we train under the following loss:

$$\mathcal{L}_{DSM} = \mathbb{E}_{p_{\mathrm{data}}(X)} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)} \mathbb{E}_{p_{|\epsilon|}(\tilde{X}|X)} \left[|\epsilon| \;\; \| \, s_\theta(\tilde{X}, \epsilon) - \nabla \log p_{|\epsilon|}(\tilde{X}|X) \, \|_2^2 \right] \tag{3}$$

where we sample at training time random noise scales $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ similarly to Song and Ermon [2020]. The minimum of this loss will be achieved for $s_\theta(X, \epsilon) = \nabla \log p_\epsilon(X)$.

Once the score function is learned through $\mathcal{L}_{DSM}$, we can plug it in Equation 2, yielding the following sampling procedure given our specific choices for the SDE terms:

$$X_T \sim \mathcal{U}_{\mathrm{SO(3)}} \qquad ; \qquad \mathrm{d}X_t = -\frac{1}{2} \frac{\mathrm{d}\epsilon(t)}{\mathrm{d}t} s_\theta(X_t, \epsilon(t)) \, \mathrm{d}t \; . \tag{4}$$

We solve this ODE down to $t = 0$ to yield samples from the learned distribution. We illustrate this process in Figure 1. Note that this is a manifold-valued ODE, which we solve using Runge-Kutta-Munthe-Kaas (RK-MK) algorithms for ODEs on Lie Groups (and we direct the interested reader to Iserles et al. [2000] for a review). Finally, we note that this generative model can trivially be made conditional, by conditionning $s_\theta(X, t, y)$ on external information y during training and sampling.

4 Application: Emulating Galaxy Intrinsic Alignments in the Illustris-TNG simulations

Weak gravitational lensing occurs when light rays from distant galaxies get deflected due to the presence of massive objects along their trajectory [e.g., Bartelmann and Schneider, 2001]. By measuring the coherent shape distortions of ensembles of galaxies, we can study the lensing effect caused by the distribution of matter in the Universe, and thereby learn about dark energy [Kilbinger, 2015]. One important systematic to model when measuring lensing is the intrinsic alignments (IA) of galaxy shapes; IAs arise due to galaxies tending to point coherently towards other galaxies due to gravitational tidal effects, which mimics a coherent lensing effect [Troxel and Ishak, 2015]. For cosmological measurements, IA must be taken into account, which means that realistic models for it must be included in synthetic galaxy catalogs.

Cosmological Simulation We will explore the efficacy of our model using the hydrodynamical TNG100-1 run at $z = 0$ from the IllustrisTNG simulation suite [for more information, please refer to Nelson et al., 2018, Pillepich et al., 2018, Springel et al., 2018, Naiman et al., 2018, Marinacci et al., 2018, Nelson et al., 2019]. We employ a stellar mass threshold of $\log_{10}(M_*/M_\odot) \geq 9$ for all galaxies, using the stellar mass from their SUBFIND catalog, and select the central galaxies from each group for our analysis. The corresponding host dark matter halos were used to study halo alignments.

Results Throughout the section we refer to the sample generated from the diffusion model as the *SGM* sample, and the sample from TNG100 as the *TNG* sample. The inputs to the model are the

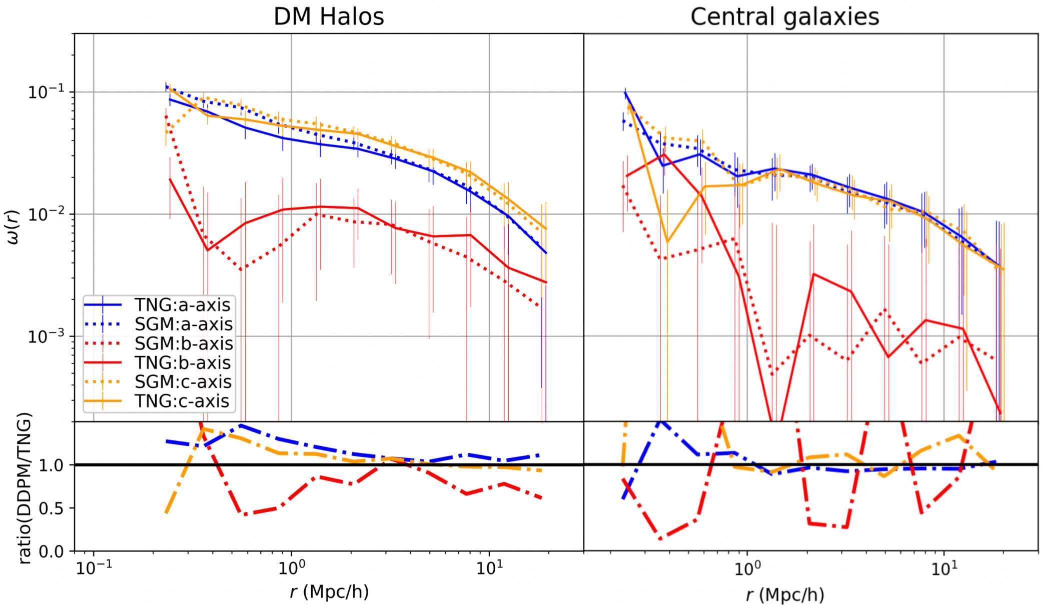

Figure 2: The two-point ED correlation function, $\omega(r)$, which captures the correlation between position and the axis direction, of all galaxy (right) and dark matter halo (left) axes with galaxy positions: the solid lines show the measured values from the TNG simulation, while the dashed lines show the generated values from the SGM. The top panels show measured $\omega(r)$ values, and the bottom panels show the ratio $\omega(r)$ from the SGM to that measured in TNG. The SGM curve was shifted by 5 per cent to the left for visual clarity. For the ellipsoid, we denote the major, intermediate, and minor axes as a, b, and c, respectively.

TNG100 gravitational tidal field (obtained from the 3D tidal tensor which carries some information about the alignment at large scales), and the outputs are the 3D orientations of halos and central galaxies: the model generates the orientations of halos and galaxies conditioned on the tidal field.

We test our model using the 3D orientation-position correlation function, $\omega(r)$, often referred to as the ED correlation. It captures the correlation between overdensity (galaxy positions) and orientations of the selected halo/galaxy axes (modeling the halos/galaxies as ellipsoids and selecting either the major, intermediate, or minor axis). Positive $\omega(r)$ values indicate that the selected halo/galaxy axis exhibits a coherent alignment towards the positions of nearby galaxies. The ED correlation functions for all three axes of the halos and galaxies are presented in Fig. 2. Here, the errorbars were calculated using the jackknife estimator. In general, the qualitative trend of ED as a function of 3D separation is captured by the SGM for both DM halos and central galaxies. For small scales (below $r \leq 1$ Mpc/h), there is a general deviation from the measured values, which may be explained by the highly complex non-linear processes that might not have been captured by the neural network. Quantitatively, for the major axes of both halos and central galaxies, the generated samples agree well with the simulation. For the intermediate axes of DM halos and central galaxies, the signal is very weak, though the SGM managed to captured the correlation with statistical consistency. However, for the minor axes, the SGM model slightly underestimates the correlation and overestimates it for central galaxies at small scales.

Overall, the SGM model can describe synthetic densities with high statistical correlations (as illustrated in Figure 1), and those with low statistical correlations, as shown in the case of galaxy/halo alignments. Regarding the limitations, the model did not capture the correlations at small scales to a good quantitative agreement, for which adding a graph-based layer may help [Jagvaral et al., 2022].

5 Conclusions

We have introduced a novel score-based generative model for the SO(3) manifold, and applied it in an astrophysical context to the modeling of the 3D orientation of galaxies and dark matter halos in the TNG100 hydrodynamical simulations. Predicting galaxy properties given a dark matter halo, or vice versa, is known as the galaxy-halo connection. Deep generative models show promise in tackling this high-dimensional multivariate problem. We have demonstrated that a smaller subset of the problem of modeling halo/galaxy orientation given the tidal field can be addressed with score-based denoising diffusion models. The diffusion model generates orientations that have statistical correlations consistent with those of the cosmological simulation, in addition to reproducing high-correlation synthetic densities on SO(3). In the future, we would like to extend this work by implementing a graph layer in order to fully capture the correlation at non-linear (small) scales and extend the number of halo and galaxy properties predicted by the model. Applying our model to a large volume cosmological simulation, to test the ability to model these alignments, will be highly useful for future weak lensing surveys.

Broader Impact

The proposed methodology of deep manifold learning on SO(3) will be practical in many disciplines outside of astrophysics/cosmology. For instance, in robotics the problem of estimating poses of objects is an intensely studied problem and our method provides a way of tackling this problem from a generative perspective with diffusion models. Additionally, in biochemistry it is often hard to find the optimal angle for molecular docking; with our proposed method biochemists could efficiently find the optimal angle that minimizes the potential energy. We do not believe that our work poses any negative societal impacts or ethics-related issues.

Acknowledgements

This work was supported in part by a grant from the Simons Foundation (Simons Investigator in Astrophysics, Award ID 620789) and by the NSF AI Institute: Physics of the Future, NSF PHY-2020295.

References

M. Bartelmann and P. Schneider. Weak gravitational lensing. *Physics Reports*, 340(4-5):291–472, January 2001. doi: 10.1016/S0370-1573(00)00082-X.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian Score-Based Generative Modeling, May 2022. URL `http://arxiv.org/abs/2202.02763`. arXiv:2202.02763 [cs, math, stat].

Joseph DeRose et al. The Buzzard Flock: Dark Energy Survey Synthetic Sky Catalogs. *arXiv e-prints*, art. arXiv:1901.02401, January 2019.

Joseph DeRose et al. Dark Energy Survey Year 3 results: cosmology from combined galaxy clustering and lensing – validation on cosmological simulations. *arXiv e-prints*, art. arXiv:2105.13547, May 2021.

Cora Dvorkin et al. Machine Learning and Cosmology. In *2022 Snowmass Summer Study*, 3 2022.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8867–8887. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/hoogeboom22a.html`.

Arieh Iserles, Hans Z. Munthe-Kaas, Syvert P. Nørsett, and Antonella Zanna. Lie-group methods. *Acta Numerica*, 9:215–365, 2000. doi: 10.1017/S0962492900002154.

Yesukhei Jagvaral, François Lanusse, Sukhdeep Singh, Rachel Mandelbaum, Siamak Ravanbakhsh, and Duncan Campbell. Galaxies and haloes on graph neural networks: Deep generative modelling scalar and vector quantities for intrinsic alignment. *Monthly Notices of the Royal Astronomical Society*, 516(2):2406–2419, 08 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac2083. URL `https://doi.org/10.1093/mnras/stac2083`.

Niall Jeffrey, Justin Alsing, and François Lanusse. Likelihood-free inference with neural compression of DES SV weak lensing map statistics. *Monthly Notices of the Royal Astronomical Society*, 501 (1):954–969, February 2021. doi: 10.1093/mnras/staa3594.

Benjamin Joachimi, Marcello Cacciato, Thomas D. Kitching, Adrienne Leonard, Rachel Mandelbaum, Björn Malte Schäfer, Cristóbal Sifón, Henk Hoekstra, Alina Kiessling, Donnacha Kirk, and Anais Rassat. Galaxy Alignments: An Overview. *Space Science Reviews*, 193(1-4):1–65, November 2015. doi: 10.1007/s11214-015-0177-4.

Martin Kilbinger. Cosmology with cosmic shear observations: a review. *Reports on Progress in Physics*, 78(8):086901, July 2015. doi: 10.1088/0034-4885/78/8/086901.

Danila Korytov, Andrew Hearin, Eve Kovacs, Patricia Larsen, Esteban Rangel, Joseph Hollowed, Andrew J. Benson, Katrin Heitmann, Yao-Yuan Mao, Anita Bahmanyar, Chihway Chang, Duncan Campbell, Joseph DeRose, Hal Finkel, Nicholas Frontiere, Eric Gawiser, Salman Habib, Benjamin Joachimi, François Lanusse, Nan Li, Rachel Mandelbaum, Christopher Morrison, Jeffrey A. Newman, Adrian Pope, Eli Rykoff, Melanie Simet, Chun-Hao To, Vinu Vikraman, Risa H. Wechsler, Martin White, and (The LSST Dark Energy Science Collaboration. CosmoDC2: A Synthetic Sky Catalog for Dark Energy Science with LSST. *The Astrophysical Journal*, 245(2):26, December 2019. doi: 10.3847/1538-4365/ab510c.

Adam Leach, Sebastian M Schmon, Matteo T Degiacomi, and Chris G Willcocks. Denoising diffusion probabilistic models on so (3) for rotational alignment. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022.

Federico Marinacci et al. First results from the IllustrisTNG simulations: radio haloes and magnetic fields. *Mon. Not. Roy. Astron. Soc.*, 480(4):5113–5139, 2018. doi: 10.1093/mnras/sty2206.

Siegfried Matthies, J. Muller, and G. W. Vinel. On the normal distribution in the orientation space. *Textures and Microstructures*, 10:77–96, 1988.

Jill P. Naiman, Annalisa Pillepich, Volker Springel, Enrico Ramirez-Ruiz, Paul Torrey, Mark Vogelsberger, Rüdiger Pakmor, Dylan Nelson, Federico Marinacci, Lars Hernquist, Rainer Weinberger, and Shy Genel. First results from the IllustrisTNG simulations: a tale of two elements - chemical evolution of magnesium and europium. *Monthly Notices of the Royal Astronomical Society*, 477(1):1206–1224, June 2018. doi: 10.1093/mnras/sty618.

Dylan Nelson, Volker Springel, Annalisa Pillepich, Vicente Rodriguez-Gomez, Paul Torrey, Shy Genel, Mark Vogelsberger, Ruediger Pakmor, Federico Marinacci, Rainer Weinberger, Luke Kelley, Mark Lovell, Benedikt Diemer, and Lars Hernquist. The IllustrisTNG simulations: public data release. *Computational Astrophysics and Cosmology*, 6(1):2, May 2019. doi: 10.1186/s40668-019-0028-x.

Dylan Nelson et al. First results from the IllustrisTNG simulations: the galaxy colour bimodality. *Mon. Not. Roy. Astron. Soc.*, 475(1):624–647, 2018. doi: 10.1093/mnras/stx3040.

Dmitry I Nikolayev and Tatjana I Savyolov. Normal distribution on the rotation group so (3). *Textures and Microstructures*, 29, 1970.

Annalisa Pillepich, Dylan Nelson, Lars Hernquist, Volker Springel, Rüdiger Pakmor, Paul Torrey, Rainer Weinberger, Shy Genel, Jill P. Naiman, Federico Marinacci, and Mark Vogelsberger. First results from the IllustrisTNG simulations: the stellar mass content of groups and clusters of galaxies. *Monthly Notices of the Royal Astronomical Society*, 475(1):648–675, March 2018. doi: 10.1093/mnras/stx3112.

Yang Song and Stefano Ermon. Improved Techniques for Training Score-Based Generative Models, October 2020. URL `http://arxiv.org/abs/2006.09011`. arXiv:2006.09011 [cs, stat].

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

Volker Springel et al. First results from the IllustrisTNG simulations: matter and galaxy clustering. *Mon. Not. Roy. Astron. Soc.*, 475(1):676–698, 2018. doi: 10.1093/mnras/stx3304.

M. A. Troxel and Mustapha Ishak. The intrinsic alignment of galaxies and its impact on weak gravitational lensing in an era of precision cosmology. *Physics Reports*, 558:1–59, February 2015. doi: 10.1016/j.physrep.2014.11.001.